A non-profit scientific paper
for obtaining the degree “Master of Science”
at the Technical University of Berlin

Data based logistics for loss reduced food value networks

Submitted by:
Lennard Ernst-August Heuer
Matriculation number: 377934
Room 302, 40, Chungmu-ro
6beon-gil, Jung-gu,
Daejeon, 35041, Rep. of Korea

Supervisor:
Dr. Julia Kleineidam

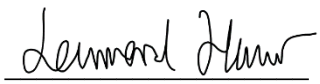
Date of submission: April 17, 2023

Affidavit

I hereby declare under penalty of perjury that I have prepared this thesis independently and without unauthorized outside assistance, that I have not used any sources or aids other than those indicated, and that I have marked the passages taken verbatim or in substance from the sources used as such.

Eidesstattliche Erklärung (Affidavit's German version):

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne unerlaubte fremde Hilfe angefertigt, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und die benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.



Daejeon, April 17, 2023

Acknowledgment

I would like to express my profound gratitude to my supervisor, Dr. Julia Kleineidam, for graciously accepting me as a thesis writer for the assigned topic. Her supervision has enriched my understanding of scientific research and her concise and honest feedback has played a crucial role in shaping my work.

Additionally, I would like to extend my appreciation to the entire Department of Supply Chain Management at the Technical University of Berlin. In particular, I would like to thank M.Sc. Stephanie Ihlenburg for sparking my interest in supply chain management through her interactive introductory course. I am also grateful to M.Sc. Benno Gerlach, whose course Supply Chain Analytics helped me enhance my data science skills and prepared me to write this thesis.

Content

Table of Figures

List of Tables

List of Abbreviations

AI .....Artificial Intelligence
FAO .....Food Agricultural Organization
FL.....Food loss(es)
FLW .....Food loss and waste
FW .....Food waste
N&C.....Nuts and cacao beans
O&P .....Oilseeds and pulses
R&T ..... Roots and tubers
SC..... Supply chain
SSA .....Sub-Saharan Africa

Abstrakt (In German)

Die vorliegende Masterarbeit beschäftigt sich mit der Forschungsfrage, welche logistische Maßnahmen am effektivsten gegen Lebensmittelverschwendung enlang von Logistiketten in Subsahara Afrika eingesetzt werden können.

Um geeignete logistische Gegenmaßnahmen zu finden, wurden nicht konkrete Gegenmaßnahmen gegeneinander abgewogen, sondern nur übergeordnete Handlungsfelder gegen Lebensmittelverluste nach Kleineidam (2020, p. 10) in Betracht gezogen.

Um die genannte Forschungsfrage zu beantworten, mussten zunächst Schwerpunkte der Lebensmittelverschwendung entlang von Logistikketten in Subsahara Afrika identifiziert werden. Danach wurden weitere in den Daten erhältliche Informationen identifiziert und quantifiziert, die Rückschluss auf Gründe für Lebensmittelverluste bieten und wurden daraufhin mit korrespondierenden Handlungsfeldern nach Kleineidam (2020, p. 10) verknüpft.

Die Arbeit bedient sich einer grundlegenden Restrukturierung der Datenbank „Food Loss Index“, veröffentlichet von der Welternährungsorganisation der UN (FAO), und verwendet darüberhinaus explorative Methoden der Datenanalyse, multiple lineare Regression und die Auswertung von menschlicher Sprache im Freitextformat.

Ergebnisse und Handlungsempfehlungen der Datenanalyse unterschieden sich je nach Kombination aus Lebensmittelkategorie und Logistikkettenstufe. Inhaltlich plädieren die Ergebnisse im Fall von mehreren Kombinationen aus Lebensmittelkategorien und Logistikstufen für bessere Kollaboration zwischen Akteuren der Lebensmittelkette.

Mehrwert der Abschlussarbeit liegt vorrangig in der Schaffung eines Entscheidungsassistenten, der auch in Zukunft mit neuen Daten einsetzbar ist.

1. Introduction

The United Nations projects a substantial increase in Sub-Saharan Africa's (SSA's) total population, with the number expected to grow from approximately 1.2 billion in 2022 to 1.401 billion by 2030 and 2.094 billion by 2050 (United Nations 2022, p. 5). This population growth will undoubtedly lead to higher food consumption. In addition to the population increase, geopolitical conflicts and climate change exacerbate the food supply situation in SSA, as described below.

The Ukrainian war led to a considerable general rise in food prices, which could potentially worsen the food supply in African countries (Glauben et al. 2022, p. 6). In a blog entry of August 2022, The World Bank highlighted that the Ukrainian war heavily impacts food security and food prices in African countries reliant on cereal and fertilizer imports from Russia and Ukraine (Abay et al. 2022).

Climate change also contributes to food scarcity in SSA by causing water shortages, recurrent droughts, and increased water demand for crop irrigation. This threatens food, water, and energy supplies in the region. Furthermore, some areas may experience soil degradation, rendering them unsuitable for agriculture and leading to increased competition for the remaining arable land (Niang et al. 2014 cited by (Mpandeli 2018, p. 7)).

Moreover, climate change negatively impacts multiple stages of the food supply chain (SC), such as harvesting, drying, and storage (Stathers et al. 2013, p. 14). In addition, food production is inherently resource-intensive, utilizing 20% of the world's landmass, 70% of global water withdrawals, and 32% of the global energy production (Spang et al. 2019, p. 118). Hence, climate change poses a significant risk to the worldwide food supply, and it is probable that climate change is even exacerbated by the resource-intensive nature of food cultivation, encompassing both produced food and food loss.

In developing countries, food loss and waste (FLW) primarily occur post-harvest, during processing, packaging, and distribution SC stages, while FLW at the consumer stage constitutes a smaller fraction of FLW (HLPE 2014, p. 27). Consequently, examining the relationship between logistics and FLW in developing countries is particularly promising, as logistical measures could potentially yield significant FLW reductions in these regions.

The World Bank (2011) estimates that between 2005 and 2007, annual economic losses due to FLW in SSA amounted to approximately $4 billion. This figure equals the total

value of all food donations made to SSA countries from 1998 to 2008. It is also roughly equivalent to what was spent annually on food imports in SSA during the same period (World Bank 2011, 14). Even though the data in these figures might not be up-to-date, they still indicate a considerable potential for reducing food donations to SSA through FLW reduction in the region.

The role of improved logistics in contributing to a reduction in FLW has been explored by different scholars. Michael Blakeney's 2019 book "Food Loss and Food Waste" provides a structured overview of approaches to reducing FLW along the food SC (Blakeney 2019, pp. 138–168). Additionally, the Routledge's 2020 "Handbook of Food Waste" presents an extensive range of innovative ways to reduce FLW, including logistical countermeasures aimed at minimizing FLW (Reynolds 2020, pp. 443–454).

Gustavsson et al. (2011) have previously delineated four different macro-regions of the world, illustrating the proportion of FLW occurring across various SC stages and food categories in these four macro-regions. However, based on their own definition, their modeled SC encompasses only two stages when only considering the processes from harvesting to the point just before retail (Gustavsson et al. 2011, pp. 4–9). The data published by the FAO in the so-called "Food Loss Index" (FLI) offer a more precise resolution in terms of SC stages, as they include many additional SC stages, depending on the entities who input the information. Consequently, the available data allow for a more detailed examination of the extent to which logistics along SCs potentially influences FLW.

However, the state of research on FLW is uneven across different world regions. Hadi et al. (2020), in a literature review on the topic of FLW, conclude that FLW research is predominantly focused on developed countries, with less research being conducted in developing countries (Hadi et al. 2020, p. 19).

Kleineidam (2020) suggested, as a research question yet to be investigated, to explore "which fields of action [against food losses have] a particularly strong impact on network partners in developing countries" (Kleineidam 2020, pp. 16–17), which is the primary research question of this paper. In this context, this master's thesis could make a valuable contribution to addressing an unanswered question in FLW research. The geographical scope of the thesis focuses exclusively on SSA, a macro region commonly regarded as overall still underdeveloped.

To support this statement with figures from the entire African continent, in a press release of 2021, The World Bank stated that "in 2019, Africa accounted for [only] 2.8% of world trade […] while in 2019, 478 million [African] people lived in extreme poverty" (United Nations Conference on Trade and Development 2021). Based on these plain numbers, it is evident that logistical measures against FLW in SSA need to meet the specific requirements found on the African continent. Particularly, it is desirable to implement logistical measures that are both low-cost and impactful.

The goal of this thesis is, therefore, to identify logistical measures that would achieve a significant impact in terms of food loss FLW reduction at an acceptable cost. To maintain a manageable complexity of this thesis, only fields of action, each field encompassing multiple logistical countermeasures, and their suitability to reduce FLW were examined. Kleineidam (2020, p. 10) already presented a comprehensive compilation of fields of action against FLW, which will be used within this thesis.

In order to answer the primary research question, namely the question of how FLW can be addressed most effectively by the use of logistical measures, several sub-questions need to be addressed:

1. What are the current hotspots of FLW along SCs in SSA? Identifying these hotspots will help target the most significant losses for intervention.
2. How can the causes of FLW be determined from the data? Gaining insight into the underlying factors contributing to FLW is essential for devising effective solutions that target the actual causes of FLW rather than merely mitigating some of its effects.
3. What methodologies can be employed to relate the identified causes of FLW to logistical countermeasures against them? Establishing these connections will facilitate the implementation of targeted strategies to reduce FLW.

Additionally, if possible, the differences among various SSA countries should be considered, as these disparities may impact the implementation and effectiveness of proposed solutions.

In order to address the primary research question and its associated sub-questions, it is essential to first establish a solid foundation for the readers' comprehension of the paper, ensuring that they possess a thorough understanding of the relevant terminology and scientific background knowledge employed in the data analysis.

The data analysis within this thesis is structured by a well-established project cycle specific to data science projects. Throughout the execution of the project cycle, all secondary research questions are answered, as the data undergoes enrichment, restructuring, exploration, and integration into a decision model. The results are then computed collectively for specific combinations of food categories and SC stages. Afterward, the generalization of the results and their contribution to the scientific knowledge base are discussed. Finally, the data-driven decision assistant is critically evaluated, and recommendations are offered for enhancing future data science research on the FLI database.

2. Theory

This chapter provides essential definitions, and foundational knowledge is explained to enable the reader to comprehend the remainder of this thesis.

2.1 Technical terminology and classifier

The main underlying dataset used for the data analysis in this thesis is the FLI, published by the FAO. Within this thesis, each row of the FLI dataset is named a "data point", each column is also referred to as a "feature". A subset of the FLI dataset that only contains only data of a specific combination of a food category and a SC stage will be referred to as a "food loss spot" (FL spot). In this thesis, references to "CCxx" denote references to a code chunk with the corresponding number xx within the R Markdown script, found in Appendix H.

2.2 General understanding of the term "logistics"

In this thesis, which aims to mitigate FL through logistical measures, it is crucial to clearly define the term "logistics" in order to establish the scope of the measures considered. Larson and Halldorsson (2004) assert that there are essentially four approaches to understanding the interrelation between "logistics" and "supply chain management" (Larson and Halldorsson 2004, pp. 18–21). Owing to the varied interpretations of these terms, it is necessary to specify the approach adopted in this paper. This thesis employs the "re-labeling" approach, suggesting that "logistics" and "supply chain management" share the same meaning and can be used interchangeably. This perspective is one of the four beforementioned views on the interrelation between the two terms (Larson and Halldorsson 2004, pp. 18–21).

To clarify the term "supply chain management", this thesis uses the definition provided by Mentzer et al. (2001). They describe it as "the systemic, strategic coordination of the traditional business functions and the tactics across these business functions within a particular company and across businesses within the supply chain, for the purposes of improving the long-term performance of the individual companies and the SC as a whole" (Mentzer et al. 2001, pp. 17–18).

Consequently, within this thesis, SC management and logistics are, firstly, regarded as interchangeable terms, and secondly, they surpass mere operational activities, such as transportation and warehousing, by incorporating a more comprehensive business

outlook. This perspective includes cross-functional integration, leveraging business intelligence, and strategic collaboration, as emphasized by Mentzer et al. (2001, pp. 17-18).

2.3 Scope of the supply chain investigated

Defining the scope of the specific segment within the SC examined in this thesis is crucial for cropping the data in a manner that aligns with the primary research question. As suggested by Östergren et al. (2014), the initiation point of each food SC is when the food product is prepared for collection. Examples include eggs that have been laid, crops and fruits that have been harvested, or animals that have been slaughtered, such as red meat or fish (Östergren et al. 2014, p. 20). The same authors define the end of each food SC as the point in time "when food is a) eaten or consumed or b) removed from the food SC" (Östergren et al. 2014, p. 21).

In this thesis, the scope of research is exclusively focused on FLW occurring between the harvest and pre-retail stages along food SCs in SSA. This decision was made based on the findings of (HLPE 2014, p. 27), which indicated that FLW in SSA predominantly occur upstream within the SC. By concentrating on these stages, this thesis aims to identify the most significant issues contributing to FL in this macro region and to show effective ways to mitigate them.

The analysis begins with an examination of FLW at the harvest stage of the SC and proceeds up to, but not encompassing, the retail stage. Hence, the retail stage, as well as the consumer stage—which is further downstream compared to the retail stage—are beyond the scope of this thesis.

This approach diverges from the preceding definition of the starting point of each food SC, as proposed by Östergren et al. (2014, p. 20), as it includes the harvest stage. However, this inclusion is deliberate and justifiable, as it allows for the consideration of potential opportunities to improve harvest processes through logistical interventions. Examples of such improvements, as proposed in this thesis, may involve optimizing routing solutions during harvest or incorporating the harvest stage into subsequent activities such as storage, transport, and processing. The cut off just before the retail stage was decided for in accordance with the view of the (FAO 2023b) mentioned above, that stated that the retail stage already belongs to the area of FW. These ideas will partly be adopted in Chapter 4.4.

2.4 Defining the terms "food loss" and "food waste"

As the term "food loss" is frequently used within this thesis, it is important to establish a common ground for its meaning. According to the FAO's view on the definition of "food loss" (FL) and "food waste" (FW), FL refers to the "decrease in edible food mass at the production, post-harvest, and processing stages of the food chain, mostly in developing countries. FW refers to discarding of edible food at the retail and consumer levels, mostly in developed countries" (FAO 2023b).

Following this definition, FL therefore encompasses any losses occurring along the SC up to, but not including, the retail stage, while all other losses further down the SC are categorized as FW. In this paper, the denotation FL adheres to this definition. Consequently, when the term "food loss and waste" (FLW) is used within this thesis, the losses discussed are, in part, beyond the purview of this thesis, given that FW occurs downstream along the SC, whereas this thesis exclusively investigates the upstream stages of the SC.

2.5 Fields of action against food losses along supply chains in Sub–Saharan Africa

The primary objective of this thesis is to ascertain effective logistical countermeasures against FL along SCs in SSA. To ensure manageable complexity, the logistical countermeasures are not presented as explicit measures, but as 13 overarching fields of action against FL, following the approach proposed by (Kleineidam 2020, p. 10). This decision was made because formulating recommendations tailored to certain stakeholders of a SC, by giving them advice on the level of concrete countermeasures would necessitate a thorough evaluation of their current conditions and additional information concerning the causes of losses. However, such an in-depth analysis falls beyond the scope of this thesis.

Each of these 13 fields of action represents a cluster of potential countermeasures to prevent FL (Kleineidam 2020, p. 1). These fields of action are transparency, quality management, packaging management, transport optimization, warehouse management, network structure, regulation, financing opportunities, physical characteristics, shelf-life optimization, network cooperation, and mindfulness (Kleineidam 2020, p. 10). The elucidation of the 13 fields of action, as proposed by Kleineidam (2020, p. 10) is available in Appendix B for further reference.

2.6 Classification of the thesis' approach to mitigating food losses

In pursuit of the thesis' objective to identify the most efficient ways to reduce FL in SSA, a fundamental question must be addressed: The extent to which potential prevention measures should focus on addressing the cause of FL instead of merely alleviating its symptoms.

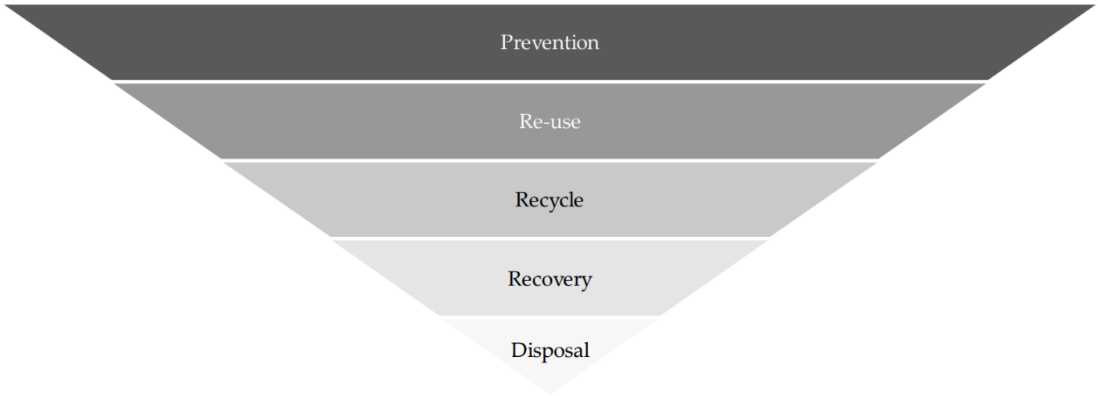


Figure 1: Food-use-not-waste hierarchy (Kleineidam 2020, p. 4 based on Papargyropoulou et al. 2014, p. 3)

Figure 1 portrays the various levels at which FLW can be addressed and the negative impacts of FLW can be mitigated (Kleineidam 2020, p. 4). However, for the purpose of this thesis and to simplify the analysis, within this thesis, only the upper-most level, the level of prevention, is considered in mitigating FL. Other levels at which FLW can be addressed will be briefly discussed in the outlook provided in Chapter 6.

A further reason for focusing only on the prevention level is that the 13 fields of actions against FL, described by Kleineidam (2020, p. 10), exclusively consider countermeasures that lead to the prevention of FL (Kleineidam 2020, p. 1) and not those that merely alleviate the symptoms of FL, such as repurposing spoiled food. As this thesis will later employ these fields of action to propose suitable countermeasures, it solely concentrates on addressing FL at the prevention level. In addition, although the dataset mentions the term "quality" eight times (CC23), this implies that the respective food has likely been reduced in quality but not completely lost. However, this phenomenon is ignored, and each value of FL [%] designated in these data points is treated as non-retrievable FL in order to maintain a reasonable level of complexity throughout the data analysis.

2.7 Geographical scope and scope of time

The classification of African countries considered as part of SSA in this thesis is based on the World Bank's established definition and categorization of SSA nations (United

Nations Statistics Division 2023). Alongside the geographical limitation, it is necessary to define the time period throughout which data points are included in the data analysis.

Upon accessing the FAO's website during the writing of this thesis, the default year range was set from 2000 to 2021 (FAO 2023a). This range was accepted as an implicit suggestion and appeared to be an appropriate balance between data up-to-dateness and data abundance. Consequently, the data analysis contains information ranging from 2000 to 2021.

The FLI dataset that was used in this thesis was downloaded on January 10, 2023, and comprises 18,509 data points within the defined SC section and the geographical and temporal scope.

2.8 Target audience of this thesis

A crucial aspect of addressing is identifying the intended target audience for this thesis. The target audience will benefit from the research findings based on their location within the various system levels discussed in this section. Levels of stakeholders encompass the micro level (focusing on a single company), the meso level (adopting a SC perspective involving multiple actors), and the macro level (considering a nation's economy), as outlined by (Vanhamäki et al. 2019, p. 33) in Figure 2 on the topic of circular economies.

Figure 2: System levels of possible target audience of the thesis
(Vanhamäki et al. 2019, p. 33 based on Manskinen 2016)

The target audience of this thesis includes stakeholders spanning all three dimensions. At the micro level, wherein a single company operates within a specific SC stage and deals with a particular food category, the findings may encourage them to rethink their current FL reduction strategies in daily operations. The catalyst for rethinking may emerge from the suggested relevance of fields of action against FL, as determined by the decision assistant. At the meso and macro levels, stakeholders operate within local ecosystems, industrial networks, and major administrative units, likely span multiple food categories and SC stages. For these stakeholders, this research offers invaluable insights by not only suggesting specific fields of action for combating FL for a particular FL spot, but also by identifying FL hotspots and directing efforts toward effective FL mitigation through prioritizing combinations of food categories and SC stages that are the most prospective to start combating. The two aforementioned groups of stakeholders can therefore derive multifaceted advantages from the study's findings.

In summary, the thesis is particularly beneficial at higher levels of the topic of FL in SSA. It synthesizes data from the FLI database and generates a comprehensive overview of FL in SSA, their causes, and suggests ways to combat FL effectively.

2.9 Scientific classification

This thesis is positioned within the realm of applied sciences. It constitutes a quantitative research endeavor, as it encompasses quantitatively measurable data and addresses a real-world issue by employing scientific methodologies.

Theoretical research, as Edgar and Manz (2017) define it, "is a logical exploration of a system of beliefs and assumptions [...and…] is valuable in understanding the bounds, edge cases, and emergent behaviors of a system" (Edgar and Manz 2017, p. 71).

In contrast to theoretical research, "empirical research is based on observed and measured phenomena and derives knowledge from actual experience rather than from theory or belief" (La Salle University 2022).

This thesis works with real-world data or modelled data based on the real world of FL, and, considering the definitions provided, it can therefore be classified as empirical research rather than theoretical research. However, it should be noted that the researcher conducting the data analysis did not personally collect the data.

3. Methodology

To ensure a structured data analysis of the FLI dataset and avoid overlooking critical aspects of it, an established model specifically designed for data science projects has been employed. The primary model utilized throughout this thesis is "The Data Science Process" by (Schutt and O'Neil 2013, pp. 41–43), which is supplemented in parts by the s-SCRISP-A-Cycle by (Herden 2019, pp. 219–234). The relatively generic model for data science projects, named "The Data Science Process", designed by (Schutt and O'Neil 2013, pp. 41–43), is shown in Figure 3.

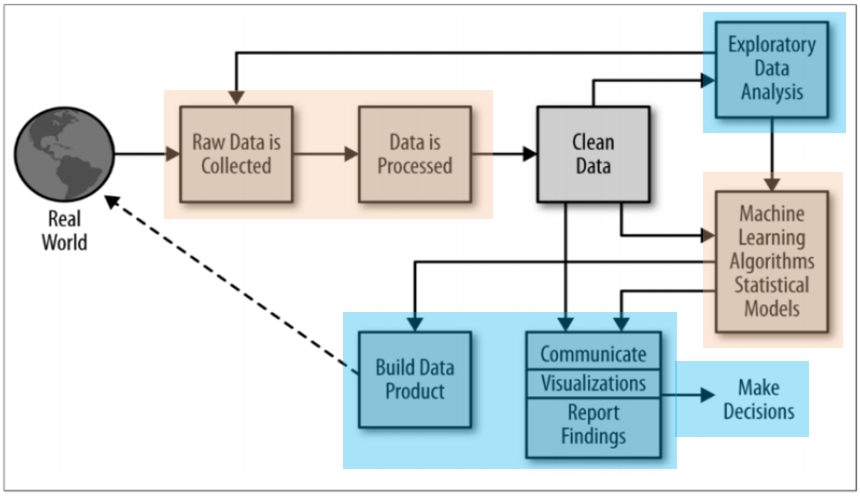


Figure 3: "The Data Science Process" according to Schutt and O'Neil (2013, p. 41), with the methodological phases highlighted in orange and all further phases in blue

In their book "Doing Data Science", Schutt and O'Neil (2013) propose a project cycle for conducting a data science project, which they refer to as "The Data Science Process". This project cycle consists of several phases, represented by the building blocks in Figure 3. Their book also offers an overview of the objectives and common challenges associated with most of the respective phases (Schutt and O'Neil 2013, pp. 17–250).

As per Google Scholar statistics from April 2, 2023, the referenced book has accumulated 413 citations (Google Scholar 2023), indicating its well-established status within the scientific community. The book's structure is based on the "The Data Science Process" project cycle, which lends credibility to the cycle's establishment and trustworthiness of it as a model. Furthermore, as illustrated in Figure 3, "The Data Science Process" by Schutt and O'Neil (2013, p. 41) demonstrates that apparently, none of the phases displayed is exclusively tailored for business use. This aspect contributes to the model's notably generic nature, as it is not explicitly restricted to academic research projects or

business-oriented projects alone.

Within this thesis, activities from each phase of Schutt and O'Neil's (2013) project cycle “The Data Science Process” are selectively complemented by activities incorporated in the s-SCRISP-A-cycle presented by Herden (2019, pp. 219–234). Although s-SCRISP-A’s documentation is more detailed, it is primarily tailored for business use rather than academic research projects. This is evident by the inclusion of “business understanding” as one of its phases (Herden 2019, p. 219). Furthermore, the project cycle's phases encompass numerous subordinate tasks that are predominantly business-focused and exhibit a high level of detail. Moreover, the inclusion of this vast range of subordinate tasks outlined by (Herden 2019, pp. 219–234) would substantially extend the page count of this thesis, exceeding its intended scope.

This chapter offers an overview of the methodology employed throughout the data analysis. The iterative nature of data analyses in general, is emphasized by the feedback loop from exploratory data analysis to raw data collection, as illustrated in Figure 3 for “The Data Science Process“ (Schutt and O'Neil 2013, p. 43). The s-SCRISP-A-cycle’s overview scheme (Herden 2019, p. 219) also recommends multiple feedback loops.

In accordance with this observation, this thesis was developed through an iterative approach featuring multiple feedback loops, such as returning to data processing following exploratory data analysis, as illustrated in Figure 3. Therefore, while forward references to other sections are generally not considered good scientific practice, occasional references to Chapter 4, in which the data exploration is conducted, are necessary to justify the methods chosen in this chapter.

Furthermore, as discussed in this chapter and Chapter 3, “The Data Science Process“ by Schutt and O'Neil (2013, pp. 41-43), as well as the s-SCRISP-A-Cycle according to Herden (2019, p. 219), inherently incorporate elements of describing the methodology employed for reaching the desired results of the data analysis.

The present dilemma involves the customary inclusion of a methodology chapter in scientific work, while the methodology is addressed at different points in both project cycles that were used to structure the data analysis. To avoid redundancy, Chapter 3 deals with the methodology and highlights those phases of “The Data Science Process“ by Schutt and O'Neil (2013, pp. 41-43) that focus on methodology (see Figure 3). These are the phases: data collection, data processing, machine learning, and statistical methods.

The subsequent three phases are individually delineated in separate chapters, building on the methodological phases. One exception to this is that exploratory data analysis is built on the phases of data collection and data processing but not on the block of machine learning algorithms and statistical models (Figure 3). Instead, it establishes the foundation for the named phase. Although this is somewhat out of order, the above-describe structure is the only way to comply with the thesis requirement for a separate chapter on methodology.

A separate phase of data cleaning is not necessary within this thesis for reasons explained in Chapter 3.2.2 and Chapter 4.

For a comprehensive replication of the data analysis, refer to the R Markdown script provided in Appendix H, which encompasses the code and corresponding comments. The decision support system constructed, also referred to as a data product, is the R Markdown script itself, as it performs the necessary operations on the dataset.

3.1 Collecting the raw data

The "Data Science Process," as outlined by Schutt and O'Neil (2013), commences with the collection of raw data derived from the real world, as illustrated in Figure 3.

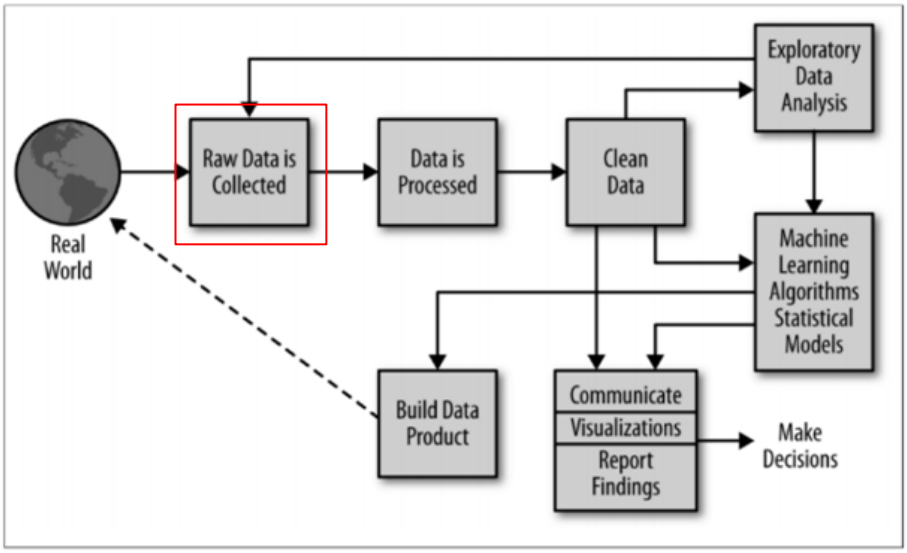


Figure 3: "The Data Science Process" according to (Schutt and O'Neil 2013, p. 41)

3.1.1 Data sources

Schutt and O'Neil (2013, p. 43) pointed out that there are various data sources to consider, such as "emails, logs, medical records, surveys, blood drawn, Olympic records, […] and web pages". Consequently, it is important to describe the data source utilized in this

project.

The present thesis aimed to collect raw data on FL and logistics performance in SSA. To this end, two sources of data were used: the FLI, published by the FAO, which was then enriched by the "Logistics Performance Index" (LPI), provided by the World Bank. According to the FAO, the FLI database comprises approximately 29,000 data points from various sources, including scientific journals, governmental releases, and national and international organizations (FAO 2023a).

The FLI dataset used in this thesis was downloaded on January 10, 2023, but it is important to consider that new data may be added to the database frequently.

The LPI is a benchmarking tool that compares logistics performance among countries, incorporating six indicators that cover customs, infrastructure, international shipments, logistics competence, tracking and tracing, and timeliness (World Bank 2023a). For this thesis, the aggregated LPI from 2012 to 2018 was selected, as it features a broader coverage of countries than the most recent 2018 data. The aggregation from 2012 to 2018 includes and combines data from 2012, 2014, 2016, and 2018, applying a weighting to prioritize the latest data (World Bank 2023a).

3.1.2 Data accessibility and handleability

As referenced in Chapter 3.1.1, the FLI data were enriched by merging them with the aggregated LPI dataset, which contains data on the logistics performance of countries aggregated throughout the years 2012-2018. At the current stage, there is no need for additional data collection beyond obtaining the LPI data.

The s-SCRISP-A-cycle emphasizes the need for "documentation of physical locations of data collection, their accessibility, required tools and infrastructure needed to handle data, and evaluation of the data integrity of the source the data is extracted from as well as evaluation of the ability to automate data extraction" (Herden 2019, p. 225).

The FLI dataset, released by the FAO, and the LPI, released by the World Bank, are freely accessible in CSV and Excel formats, respectively, and can be read by standard computer software. As these datasets are downloaded from the official websites of reputable international organizations, their reliability can be considered as high. While automated data extraction may be feasible in principle, it is not deemed essential, considering that the data necessitates only sporadic updates. This is associated with the updating of new FLI data, as demonstrated in Section 4.3.3, and also because the issue pertains to FL, and

recommendations for countermeasures possess a more long-term character, not being subject to daily or weekly fluctuations.

3.1.3 Inventorization and additional data collection

The s-SCRISP-A-cycle mentions the "examination and inventorization of data and data sources to build understanding on availability of data, accessibility of data sources, additional data collection needs, and data acquisition needs from third parties" (Herden 2019, p. 225).

As outlined in Chapter 2.3, the focus of this thesis is on SSA as its geographical scope. The availability of FLI and LPI data for the respective countries within this region is illustrated in Figure 4.

It is worth noting that some geographically small countries lack data in at least one category (the FLI or the LPI) but are not visible on the map due to their size. A detailed enumeration of data availability is included in Appendix A. As the FLI dataset is the main dataset that was merely enriched with the FLI dataset, data points belonging to all countries that lack FLI data were removed from the dataset. In the cases where only LPI data was missing, the data points were kept and only later temporarily dropped from the dataset for the purpose of conducting a multiple regression by the use of LPI data.

There is only a small number of nations on the map that show a lack of data, and these are not among the most populous. In summary, the data coverage for the countries within the scope is generally comprehensive. This observation is crucial for the subsequent discussion of results, as well as the generalization and specialization of findings that follow the data analysis.

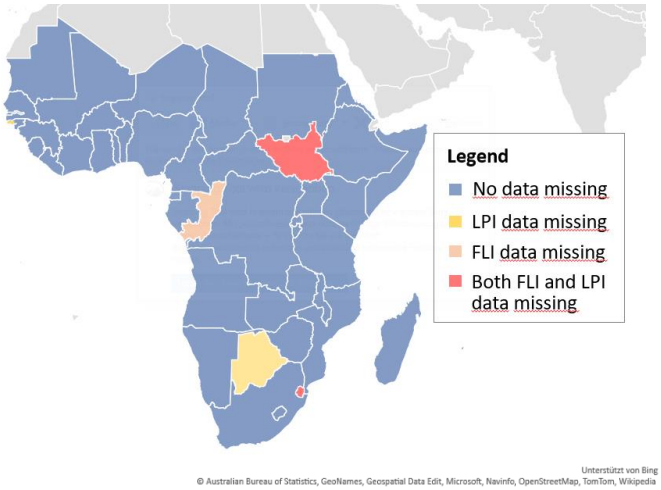


Figure 4: Overview of all Sub-Saharan African countries and their data availability (own figure)

An inventorization of data in terms of an overview of the columns of the merged dataset will later be provided in Chapter 4.2.2.

3.2 Data processing

After the raw data is collected, they are further processed, as seen in Figure 3. This subchapter's agenda will discuss the tasks that were suggested by O'Neil (2013) and Herden (2019) and later carried out.

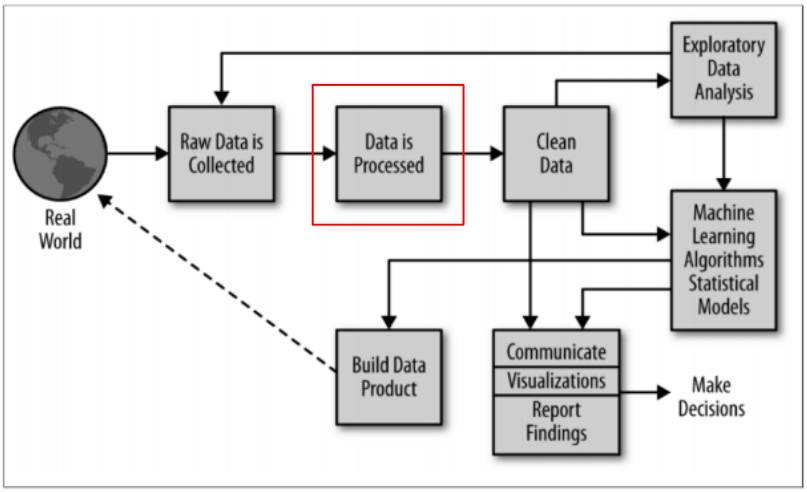


Figure 3: "The Data Science Process" according to Schutt and O'Neil (2013, p. 41)

3.2.1 General approach to data processing

O'Neil (2013) enumerated the following data processing techniques: "pipelining, web-scraping, cleaning, munging, joining and wrangling" (Schutt and O'Neil 2013, p. 43).

As for this data analysis, the data will be handled in a pipe-lined approach, making it not a one-time solution but a program that can be deployed even in coming times when data might become more abundant, with minimal effort to maintain the solution. This factor is crucial for ensuring the long-term value of the thesis.

3.2.2 Adapting the data to the supply chain scope of the thesis

Adapting the dataset to the specified SC scope was not explicitly delineated by Herden (2019) or Schutt and O'Neil (2013). Nevertheless, it is evident that after theoretically defining the geographical scope, temporal scope, and the SC scope, the data must be adapted accordingly. This adaptation is executed in the present subchapter.

The temporal scope didn't need to be particularly addressed in this section, as the data was initially obtained within the appropriate time frame (2000 - 2021). In section CC07-

CC09, the entire FLI dataset was then adjusted to align with the geographical scope of SSA, utilizing the World Bank's definition of SSA (World Bank 2023b).

As stated in Chapter 2.3, this thesis sets clear boundaries regarding which data along the SC should be included. It encompasses data from harvest to retail, excluding the retail SC stage itself. The remaining data must be omitted to comply with the defined system boundaries. The identification of all unique SC stages present in the dataset, after it has been cropped to the geographical scope and temporal scope, can be found in CC12, which are: farm, transport, storage, harvest, whole supply chain, export, wholesale, retail, processing, trader, post-harvest, market, distribution, households, and NA (not available).

Harvest, processing, transport, and storage are in the scope defined and will be adopted. The stages wholesale, retail, trader, market, and households of the SC must be excluded since they represent SC stages on a retail level or further downstream.

Given the uncertainty surrounding the precise interpretation of the term "export" in relation to concrete activities in this context, data points with a SC stage defined as export were removed from the dataset.

One of the major challenges at this point of the data analysis is to deal with different resolutions of data on FL on various SC stages: whole supply chain, post-harvest, and distribution encapsulate more than one SC stages of all SC stages that were included in the dataset. Hence, the original dataset contains data of numerous SC resolutions.

Within this thesis, is crucial to have a clear understanding of the exact point at the SC where FL occurs. Without this precise knowledge, it becomes challenging to devise targeted strategies to mitigate FL effectively. Data points belonging to these SC stages could become useful only if they accurately summarize the exact scope of the SC chosen within this thesis, as they might then be used for checking or controlling purposes.

Straka (2019) explains: "distribution can be understood as the subsystem of logistics where elements are the means of storing and packing […]" (Straka 2019, p. 9).
However, this statement leaves some ambiguity regarding the inclusion of retail stages, and it certainly does not encompass the harvest SC stage. The term "post-harvest losses" does exclude the harvest stage of the SC by definition. The expression "whole supply chain" implies that the entire SC is taken into account, encompassing the FL up to the retail SC stage and further downstream.

None of these three summarizing SC stages accurately represented the specific

accumulation of SC stages within the SC scope considered in this thesis, and hence, they could not serve as effective checking or controlling SC stages to verify the sum of the other, more distinct, SC stages. Consequently, the respective data points were removed from the dataset.

3.2.3 Data construction

Instead of the term "data processing", the s-SCRISP-A-cycle uses the term data preparation that, among others, encompasses the step of construction of data. Further, Herden (2019) describes the construction of data as "the identification of relevant features in the dataset and the creation of features on the basis of the available data" (Herden 2019, p. 227).

Overview of features of the dataset

CC20 shows an overview of the features of the present dataset, consisting of FLI and LPI data.

Key features of the dataset, deemed particularly relevant to addressing the primary and secondary research questions, include: "commodity", " FL [%] ", "food SC", "cause of loss", and "country". These four features are particularly important, as together they provide insight into the question of which type of food was wasted where, at which SC stage, and why.

The features "URL" and "reference" can provide with meaningful information on the origin of data that could potentially be used to determine whether two data points originate from the same source.

Other less important features may be: "loss_quantity", which was only filled in in 0.07% percent of the cases (CC37). Also, there is "treatment" feature, for which there information on experiments.

Features with no apparent relevance to the thesis's objective are "region" and "FL [%] original", which, however, will not be removed throughout this data analysis, as they might still become use full throughout the course of the data analysis. Regarding the LPI dataset, features associated with countries' rankings concerning their overall LPI score or LPI indicators were excluded from the analysis. This decision was made due to the presence of data from other nations within the dataset, which complicates the interpretation of score relationships between SSA countries. Additionally, utilizing

ranking values mathematically proves challenging, as it is difficult to determine the significance of the difference between rankings. For instance, it is unclear whether a #1 ranking is 33 times superior to a #33 ranking. Consequently, solely the indicator scores or overall LPI scores provide a meaningful basis for comparison among different SSA countries.

An important observation from the overview of the dataset features is that the dataset appears to be pre-cleaned. This can be deduced from the presence of the "FL [%]" and "FL [%] original" columns. The "FL [%] original" column likely contains loss values on a character scale, while the "FL [%]" column has them in a numerical format. This suggests that the contributors initially provided the loss values in various formats, possibly including the percentage sign. The FAO may have subsequently processed these values and converted them into a standardized, numerical format for ease of analysis.

Adding the feature "index"

In CC03, immediately after downloading the dataset, an "index" feature was added to maintain a consistent numbering system. This index allows for the unique identification of each data point, regardless of any subsequent operation performed on the dataset.

Adding the feature "food category"

The addition of the feature "food category" is essential for analyzing data associated with commodity types while maintaining a manageable level of complexity. Choosing the appropriate number of food categories necessitates a trade-off decision: If the data points within a particular category are diminished, the validity of the results for those categories may be compromised. On the other hand, when food categories encompass highly varied commodities, it becomes difficult to apply them collectively to a single outcome and obtain meaningful results from the analyses conducted on them.

Xue and Liu (2019) have limited the number of food groups and conducted a high-level FLW data analysis using 10 categories of food. The suggested food categories are cereals, roots and tubers (R&T), oilseeds and pulses (O&P), fruits, vegetables, meat, fish and seafood, dairy products, eggs, and "others or not specified" (Xue and Liu 2019, p. 5). The adoption of these 10 food categories, as proposed by Xue and Liu (2019), appears to strike an optimal balance between acquiring meaningful information based on the collective use of food categories that are relatively homogenous internally, while preserving simplicity in data visualization, and maintaining reasonably large sample sizes and, thus, statistical

power.

Appendix C demonstrates how all data points within the scope of this thesis were allocated to the 10 food categories recommended by Xue and Liu (2019), using their corresponding CPC codes. CPC, or Commodity Product Category, is a classification system for products developed by the United Nations, which has published an extensive register of CPC codes and their meanings in relation to commodities (United Nations Statistics Division 2023).

Overall, Appendix C indicates that matching the appropriate CPC codes with the 10 food categories outlined by Xue and Liu (2019) has been relatively straightforward, with minimal influence from researcher bias. In instances where commodities could not be directly assigned to a food category using the key list of commodities and CPC codes as per the (United Nations Statistics Division 2023) these unclear assignments were justified and explained in CC08-CC10.

Reconstruction of supply chain stages

The construction of an adapted food SC is essential to ensure it aligns with the defined scope and the overall objective of this thesis.

In the previous subchapter, the relevant SC stages were selected, was the SC scope was defined. As a result, only the SC stages harvest, farm, transport, processing and storage were retained, which can be seen in Figure 5.

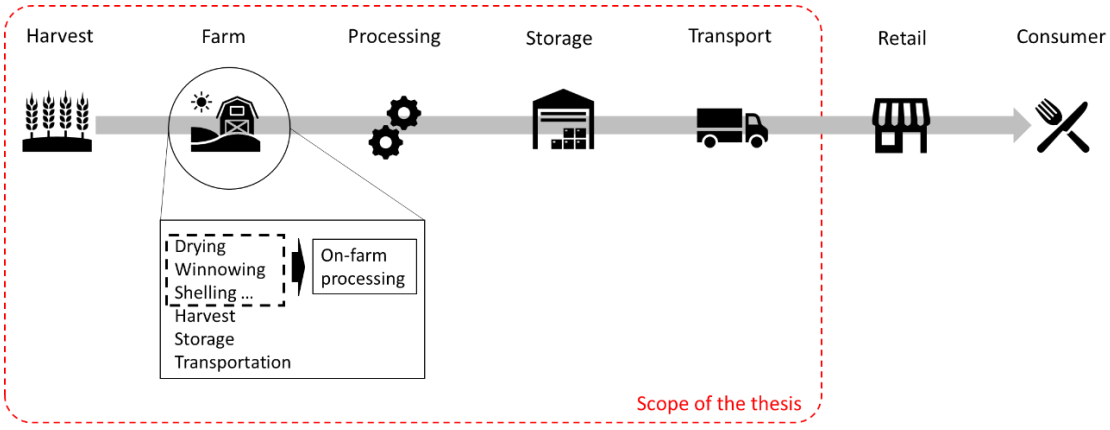


Figure 5: SC stages and scope of the thesis after stage selection and before reconstructing SC stages (own figure)

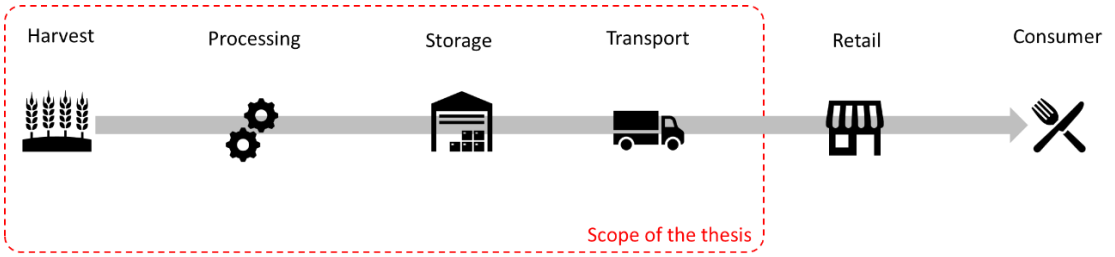


Figure 6: SC stages and scope of the thesis after reconstructing SC stages (own figure)

As established in Chapter 2.2, this thesis' analysis begins with the harvest level of the food SC which is included in the data analysis. The farm's sub-SC stages involve various activities, detailed in CC12, that can be categorized into four groups: harvest, on-farm processing, storage, and transportation. While the SC stage of transport appears only once both in Figure 5 and in Figure 6, specifically between storage and retail, these depictions of the SC are rather meant to serve illustrative purposes. Transportation could indeed also occur multiple times along the SC chain, and in the reconstructed dataset, the SC stage of transport indeed accounts for multiple possible transports along the SC chain.

A significant challenge when working with this data is addressing its multileveled nature. As discussed in the previous chapter, Chapter 4.2.1, the dataset presents varying resolutions of the SC stages, with some stages encompassing multiple SC stages. Additionally, the SC stage farm itself, as illustrated in Figure 5, comprises the previously mentioned substages, which are also found outside the SC stage farm.

In order to address the multi-level nature of data caused by the farm SC stage, a radical approach was adopted. This method involved dissolving the farm SC stage entirely and redistributing its data points to the corresponding SC stages, according to entries in the activity column. Two primary reasons justify the necessity of completely dissolving the farm SC stage.

Firstly, the farm SC stage represents a physical location, whereas the other four SC stages denote activities rather than physical locations. By dissolving the farm SC stage, the SC design is transformed into an activity-driven model instead of a facility-driven one. To answer the primary research question, how FL can be most effectively combated in SSA using logistical measures, it is crucial to group data by activities. This approach may enable more meaningful and actionable recommendations for improvements on activities, rather than concentrating on FL reduction within a single facility where various activities take place. From a broader perspective focused on the facility, the exact processes might not be evident, making it more challenging to provide meaningful recommendations on

actions that cause FL.

Secondly, as demonstrated in CC12, the activities of data points belonging to the farm SC stage and of those belonging to the processing SC stage exhibit a significant similarity. It is possible that much of the data from the processing stage of the SC actually pertains to FL during food processing on the farm. This overlap of activities might also apply to the farm and storage and potentially to farm and transport.

Consequently, it is challenging, if not impossible, to distinctly differentiate between the farm and the other four SC stages due to their overlapping nature. This issue might arise from the individuals responsible for inputting the data, who may have employed different approached to indicating SC stages, resulting in ambiguity and overlapping data.

In summary, processing, transport, storage, and harvest occur both as SC stages under their respective SC stage labels and as activities within the farm stage of the SC. Consequently, it is crucial to handle the farm data effectively and without redundancy.

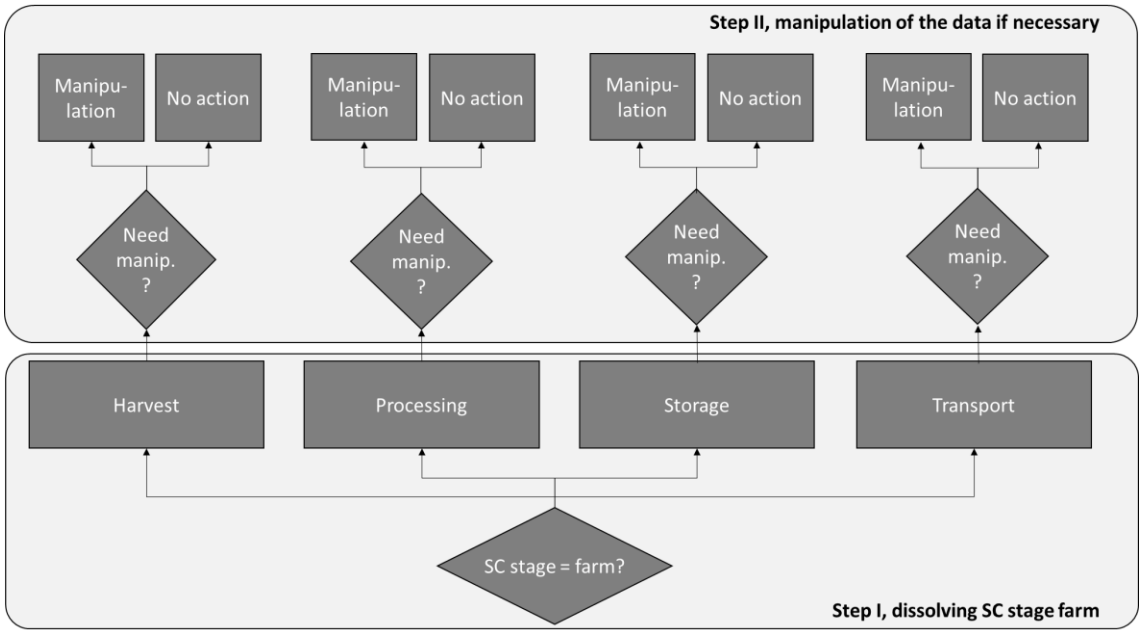


Figure 7: Dissolving the farm SC stage and further manipulation of data points (own figure)

In response, the farm data underwent reconstruction. Figure 7 provides an overview of the reconstruction approach in step I and data manipulation in step II, both of which will be further elaborated in this excerpt.

Initially, data points associated with the farm SC stage were redistributed among the other four SC stages, leveraging the information found in the activity column to determine their suitable allocation. For instance, if a data point with a SC stage attribute labeled as farm

contains information in the activity column such as "threshing", the data point is then reassigned to the processing SC stage. This reassignment is achieved by altering the SC stage attribute from farm to processing in the dataset for the specific data point.

If no information is given on the activity, the data points were dropped from the dataset. Dissolving the entire SC stage farm was necessary for several reasons. Firstly, attributing suitable countermeasures to this stage would be highly uncertain, given the diverse tasks involved and the ambiguity regarding which activity specific countermeasures should apply to. Second, organizing data based on activities enhances the interpretability of the results, facilitating a more comprehensive understanding of whether the outcomes are logical in terms of retrospective sanity checks. Finally, retaining the farm SC stage would result in double counting, as numerous activities at the farm and processing stages were discovered to be identical and might refer to the same activity at the same location caused by individuals inputting the data on their own assessment.

The downside of this approach, however, is that the information regarding whether an activity, such as transport, actually took place on the farm goes lost, as the SC stage of the reassigned data is no longer labeled as "farm". Nevertheless, this limitation can be justified, as discussed in the following: In CC12, activities on the farm and activities during processing are compared. It is observed that these activities were, overall, relatively similar. It can, therefore - not only for processing - be assumed that the individuals inputting the information were given a certain degree of freedom in specifying wether they input FL data as denoting the activity as a SC stage or whether they indicate the farm stage of the SC and optionally add the activity to the activity column, among other options to choose from. For now, it can be assumed that much of the FL assigned to SC stages, like storage and processing actually takes place on the farm, with individuals merely entering the activity as SC stage rather than the location.

In the following step, after assigning the farm data points to their appropriate SC stages, a sophisticated approach was employed to deal with the problem of false computations of averages where summarization would be needed. To illustrate the problem, a fictive example is used:

> Assuming that there are two data points that report on FL of oranges in Kenya, in the year 2007, at the SC stage of storage, contributed by the same individual, but they just differ in their cause of loss or their activity. One data point's entry in the cause of loss column is denoted as "Loss due to moisture", and the other data point's entry in the cause of loss column is described as, "Loss due to rodents". If the data points were contributed by the same author, averaging the FL [%] values would be inappropriate since the correct approach would be to sum the FL [%] values across these individual causes of losses. In reality, these two sources of losses would not exclude one another but would occur simultaneously, accumulating on top of each other.

Ensuring that the data originated from the same author was done by grouping the features by the features region, reference, and URL, as they might identify the entity who has contributed the data to the dataset. Additionally, the treatment column was considered and added to the features that were grouped by. This is not to trace back the data points to a single contributor, but to avoid a false adding of data in the case of repetitive experiments where different treatments were tested and the FL [%] was noted. It is straightforward that in such cases the FL [%] values should not be added up together.

This approach offers a more accurate representation of the total FL occurring at the "farm" stage, taking into account the cumulative impact of all associated activities. The process was executed following the logic described in CC13-CC19. The adding of FL [%] values was then performed using the mutate function in R, which facilitates data manipulation of a specific column while retaining the original data points and their values in any other columns but the mutated one. Consequently, even after this step II, all data points remain consistent in terms of their overall number and information given in each of their columns, with the exception of the FL [%]. The use of the mutate function was mindfully selected, since a simple aggregation of these data, by the means of a pivot table, would have meant that the number of data points would be reduced and certain data, for example information in the cause of loss column, would have gone lost.

3.4 Natural language processing and multiple regression

As discussed in the presentation of relevant features, the features commodity, SC stage, and food category can be utilized to identify and visualize hotspots of FL across food categories and SC stages. This addresses **secondary research question no. 1**, which

seeks to determine the hotspots of FL across food categories and SC stages in SSA. The data exploration and retrieval of results will be conducted in Chapter 4, using a heatmap and a stacked bar chart. These hotspots can serve as starting points for improvement, provided that the effort required to mitigate these losses is reasonably low.

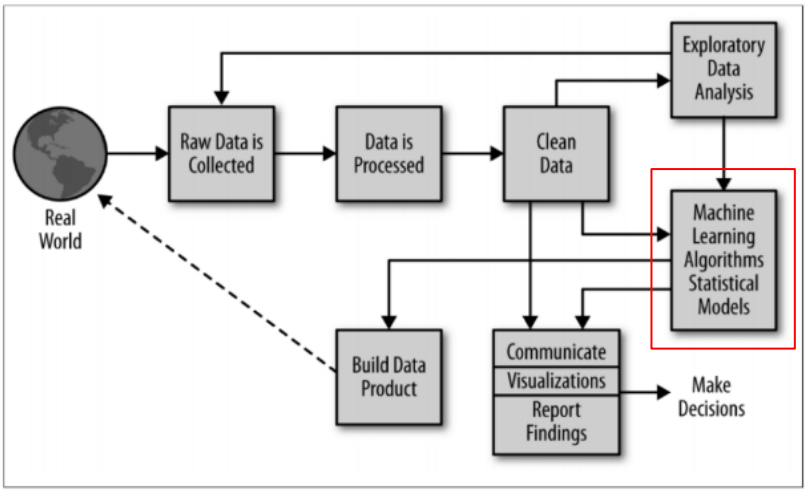


Figure 3: "The Data Science Process" according to Schutt and O'Neil (2013, p. 41)

In this phase of the data analysis, the second secondary research question is being addressed, which focuses on how inferences can be made regarding the causes of FL using only the present dataset, without initiating an extensive literature analysis. Within the dataset, two sources of information may aid in addressing the **secondary research questions no. 2 and no. 3**, which specifically ask how the causes of losses for FL can be identified and how these causes can be related to fields of action against FL. These two sources of information are the cause of loss column and the six LPI indicators, which may suggest the most decisive logistical determinants in reducing FL. To extract information from these two data sources, basic NLP techniques and statistical methods and were employed for the intended purpose.

As Schutt and O'Neil (2013) don't name specific sub-phases or tasks associated with this phase, only elements from Herden (2019, pp. 219–234) were used to structure this phase of the data analysis. Herden (2019) mentions the selection of modeling techniques, which includes "communication of the chosen technique to stakeholders including their pros and cons" (Herden 2019, p. 228). This chapter was consequently structured into three sections, each dedicated to a specific method employed. The sections include the selection of the technique, which also encompasses the advantages of and reasons for choosing this particular method; the technical implementation, where the application of the method is

discussed in detail; and the limitations, where the drawbacks or “cons“ of the method are outlined.

As a final note, the naming of the phase was changed from “Machine learning algorithms and statistical models“ to “Natural language processing and multiple regression“ to better align the heading with the content of this chapter.

3.4.1 Natural language processing – selection of the technique

Natural language processing (NLP) applies to unstructured human text data with the aim of retrieving information from it (Institute of Electrical and Electronics Engineers 2011, p. 1). Feng et al. (2021) further explain that NLP is a subfield of artificial intelligence (AI) and linguistics, aimed at enabling computers to comprehend human language statements or words (Feng et al. 2021, p. 1). In NLP, a computer takes natural language input and transforms it into natural language output (Chopra et al. 2013, p. 1).
In this thesis, “Question Answering”, one of the major tasks in NLP (Chopra et al. 2013, p. 133), is utilized to develop a decision support model that, among other functions, aligns information in the causes of loss column of the LPI dataset with suitable fields of action as per Kleineidam (2020, p. 10). This approach facilitates the identification of the most promising fields of action for addressing FL.

However, due to the diversity of the data in the cause of loss column (as later discussed in Chapter 4.3.7), simply counting word occurrences is deemed impractical, and the need to incorporate world knowledge into the decision-making is recognized. A method to integrate world knowledge into the task has to be found, with the ability to assign the causes of losses preferably in an automated way, as 247 entries in the cause of loss column (CC48) is a significant number, and new data is likely to be added frequently. Conducting expert consultations on a frequent basis as new data emerges is not practical, considering the initial objective of creating a permanent solution. An automated method for assigning causes of losses to fields of action needed to be devised, which is why NLP was chosen. Another objective of using an existing NLP model for classifying cause of loss data was to minimize researcher bias and reduce the effort required to assign a set of causes of loss to potential fields of action. After an unsuccessful attempt with Google's BERT model for the assignment task, the decision was made to automate the assignment process using an API to OpenAI's program ChatGPT.

ChatGPT is a Chatbot that was created to conversationally engage with users by taking

user input and generating answers that resemble human answers. The underlying technology is the GTP (Generative Pretrained Transformer) language model, which was developed by OpenAI (Aydın and Karaarslan 2022, p. 22).

The ChatGPT API under the ChaptGPT 3.5 Version of January 9, 2023, was used in a Python environment inside the R Markdown script to automate the assignment task of entries in the cause of loss column to fields of actions. The query automatically submitted to ChatGPT for each piece of information in the cause of loss column is displayed below, with the {cause of loss} representing the input information on the cause of loss directly sourced from the FLI.

```
"What's the best way to reduce food loss if the cause of loss is {cause_of_loss}?

Possible fields of action for counter measures would be: '-1- Transparency' which is described as 'Increase of transparency within a company as well as between companies of a network', '-2- Quality management', which is described as [...]

Please only print out the highly relevant answers and keep the numbers inside the delimiters and return the result only as a numerical list in square brackets. And don't print out any explanation of the decisions taken, print out nothing but the list."
```

Figure 8: Input query for the NLP assignment task (own figure)

The output was then automatically stored as an evaluation matrix, as programmed.

The objective of deploying the ChatGPT API was to replace human expert consultation while offering expert-level answers. However, the challenge in accomplishing this involved selecting an AI tool that possesses general reasoning capabilities to examine information based on their actual meaning. Moreover, the AI needed to apply causality to determine which countermeasures would lead to a reduction or elimination of FL caused by the given cause of loss. To lend credibility to the results of the NLP assignment task, it is crucial to assess ChatGPT's capability to effectively manage this assignment task. This evaluation is conducted as follows:

What's the best way to reduce food loss if the cause of loss is insect pest damage?

[...]

Please keep the numbers in the delimiters and only name the highly relevant field of actions and return the answer with explanations that are as precise as possible.

The most relevant fields of action to reduce food loss caused by insect pest damage are:

-2- Quality management: Improving the quality control measures to detect and prevent insect pest damage in the early stages. This can be done by regular monitoring of storage and transport conditions, using appropriate insecticides, and implementing hygiene measures to prevent insect infestations.

-3- Packaging management: Improving packaging management during transport, storage, and distribution processes can prevent insect pest damage. This can include the use of insect-resistant packaging materials, proper sealing, and storage in insect-proof containers.

-5- Warehouse management: Improving warehouse management using suitable storage equipment, storage strategies, and adapted layout planning can prevent insect pest damage. This can include proper ventilation, temperature control, and regular cleaning.

-10- Shelf-life optimization: Process adaptations that allow the shelf life of the products to be taken into account in decision making can help to prevent insect pest damage. This can include implementing proper storage and handling practices, using appropriate packaging, and ensuring timely distribution.

By focusing on these key areas, companies can prevent insect pest damage and reduce food losses, which will improve the sustainability and profitability of their businesses while also contributing to food security and reducing food waste.

Figure 9: Answer to exemplary input query under ChatGPT 3.5 Janaury 9, 2023-version, result retrieved on January 30, 2023

In Figure 9, the presented query and its corresponding answer serve to demonstrate the AI's capability for addressing the given assignment task, with the inserted cause of loss being an authentic entry from the cause of loss column. The response appears reasonable. As noted, quality management can typically tackle the root causes of FL resulting from insect pest damage, as it encompasses a holistic approach aimed at enhancing quality. Adequate packaging management, as depicted by the text generator, can contribute to safeguarding the package contents from external factors, such as insects. Given that insect pest damage often transpires when food is stored in a single location for a prolonged duration and under unsuitable storage conditions, the proposition of warehouse management as a crucial field of action in this instance is sensible as well. Nonetheless,

the reference to shelf-life optimization is not immediately apparent, as insect pest damage can occur irrespective of the duration between harvest and consumption. Conversely, an extended period of time allows for increased opportunities for insects to locate the food. Moreover, many food products become more appealing to insects as they ripen, due to factors such as the softening of their skin and changes in odor. For instance, these alterations make the products more susceptible to insect infestation. Although the responses are not exceptionally detailed, they generally appear sensible from the perspective of the thesis author.

Furthermore, during the examination of the cause of loss column and subsequent visualization of results, sanity checks are conducted on the go. For instance, it would be somewhat unusual or questionable if the NLP analysis for a specific food category and the transport stage of the SC would not evaluate the field of action "transport management" as relevant in the vast majority of cases.

3.4.2 Natural language processing – technical implementation

The objective of this part of the thesis is to retrieve meaning from the cause of loss column and relate its information to fields of action relevant to mitigating the causes of losses for a specific FL spot. The proposed metric for evaluation is the frequency with which the AI system identifies a particular field of action as relevant for reducing FL caused by the identified factors. This will be expressed as a proportion to the total number of instances the AI was queried for information for a certain FL spot.

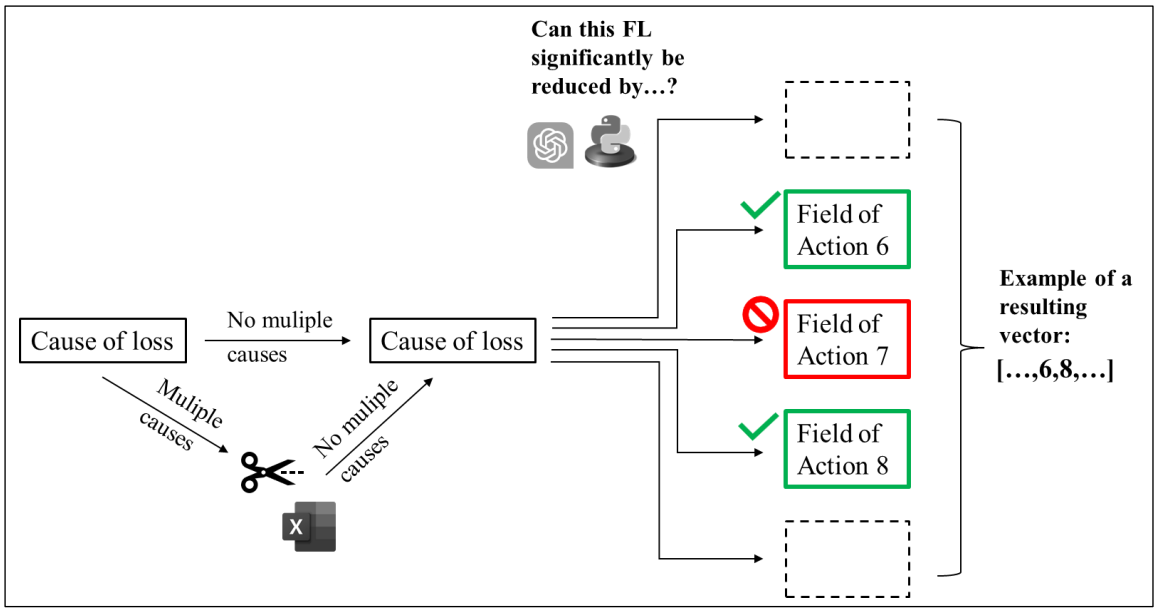


Figure 10: Process of matching causes of losses with fields of action (own figure)

The methodology employed in this study to extract meaning from the cause of loss column (see Figure 10) first involved reducing the causes of losses to a single cause of loss per data point through manual curation. This modification affected only a relatively small number of data points, as the majority of data points already contained a single cause of loss] already or none. The reduction was necessitated by the AI's inability to manage the assignment task with multiple causes of losses during the analysis period. Consequently, the outcomes were returned as a vector comprising the field of action as numbers from one to 13 that were classified as relevant by the AI. The resulting vectors, as displayed in Figure 10, were then transformed into an evaluation matrix containing numbers from one to 13, with each number representing a specific field of action. In the resulting evaluation matrix, each row corresponds to a data point containing information in the cause of loss column, while the columns are filled with the relevant fields of action, and any gaps are filled with zeros. The automated communication of information with ChatGPT through an API, as well as the transformation of individual vectors into the evaluation matrix, were implemented using a Python embedding within the R script.

The evaluation matrix only needed to be computed once; through indexing, it was made possible to retrospectively select a subset of data that exclusively contains information for one FL spot. The FL spots were then assessed as a whole, and it was investigated how many times a particular field of action, represented by a number, was considered relevant for all data points containing information on the cause of loss within the selected data subset, the FL spot. However, later, only those data subsets were utilized that included at least 10 entries in the cause of loss column to minimize the impact of randomness later. The eligibility of the FL spots for the NLP analysis is shown in Figure 11.

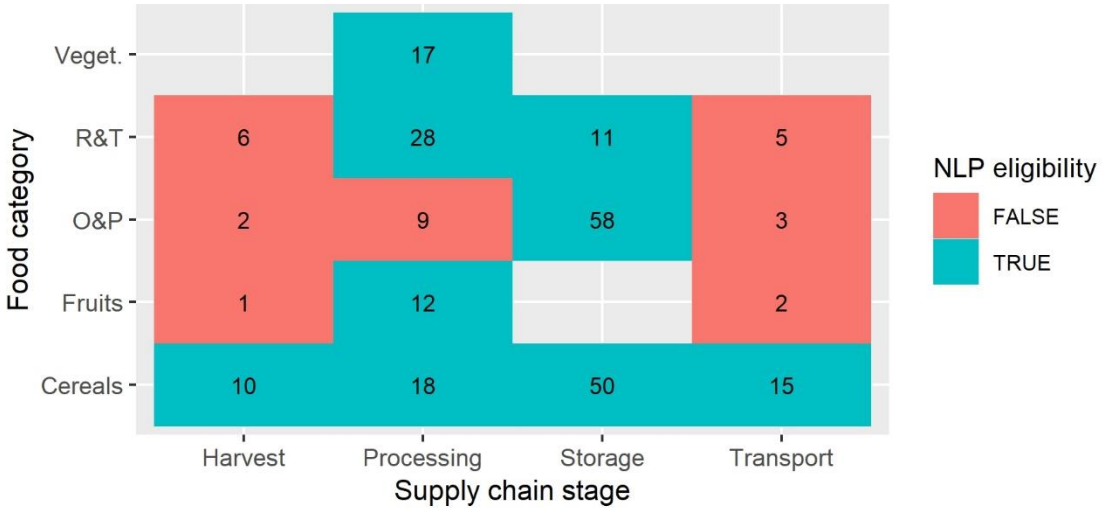


Figure 11: Sample size eligibility for NLP analysis of FL spots (own figure, CC68)

As depicted in Figure 11, all FL spots featuring cereals as a food category qualify for the NLP analysis. Additionally, numerous individual FL spots have more than 10 entries in the cause of loss column, making them suitable for the NLP analysis as well. Setting a minimum of 10 entries for each FL spot to be analyzed in the NLP analysis strikes a reasonable balance between minimizing potential randomness in results due to a small sample size and the objective of conducting the NLP analysis on as many FL spots as possible to obtain comprehensive results across multiple food categories and SC stages. In total, the NLP analysis will provide recommendations on fields of action for nine FL spots.

Moving to the topic of visualization, when examining Figure 12 and selecting data segments corresponding to FL spots, the proposed relevance of fields of action for mitigating the FL spots' causes of losses, are displayed in a bar plot. As mentioned previously, ChatGPT, based on GPT 3.5, version of January 9, 2023, was employed to match the causes of losses with the fields of action identified by Kleineidam (2020, p. 10). The red stripes represent the results when all data points of the dataset adapted to the scope of this thesis containing information in the cause of loss column are assessed collectively.

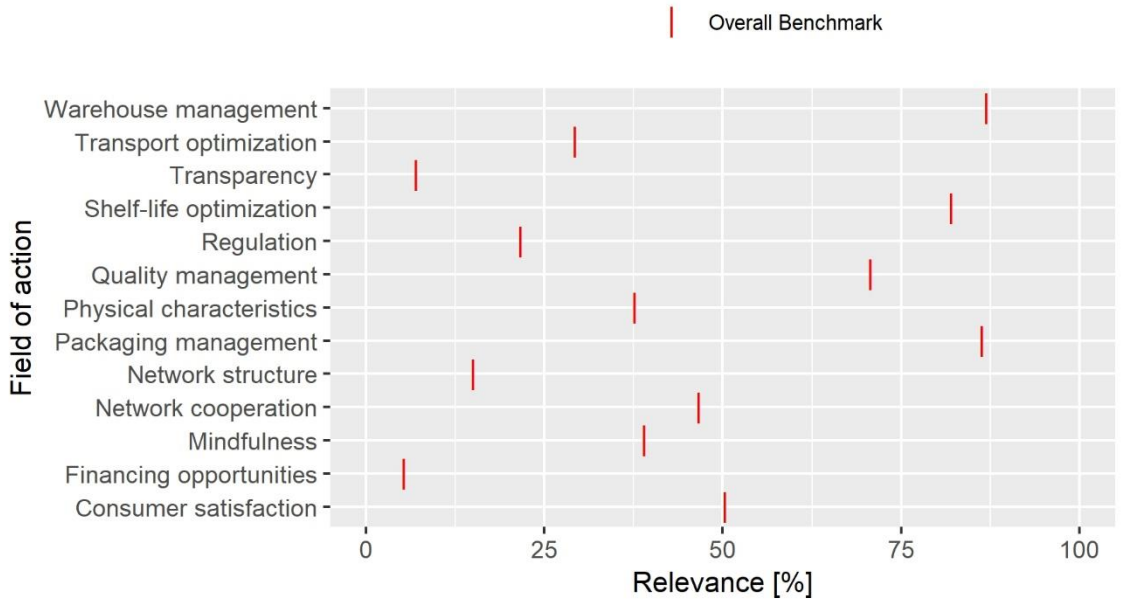


Figure 12: Visualization of benchmarks of the NLP analysis (own figure)

As observed in Figure 12, by examining the red stripes, the AI tended to rate certain fields of action significantly more frequently as relevant than others. This poses a challenge to determining the *actual* relevance of fields of action as proposed by the AI, which was addressed by incorporating the previously described red benchmarks into the evaluation and always comparing the results of computed relevance of fields of action for a given FL spot with the results that would be obtained when evaluating all cause of loss entries of the entire dataset (adapted to the scope of the thesis) altogether, translating to the beforementioned red striped, the "overall benchmarks". It is crucial to interpret the results in light of the overall benchmark comparison, as the AI system's naming of fields of action is, as mentioned, highly unequal. This discrepancy can probably be attributed to the nature of the fields of action themselves: some fields of action represent more general solutions and may therefore be mentioned more frequently than other, more specific, fields of action, which may naturally be named less often as they address a relatively specific issue.

Given the relatively limited number of 247 entries in the cause of loss column (CC48), it is tempting to examine the individual cause of loss entries. Particularly at FL spots where cause of loss data is more abundant, directly examining the cause of loss entries becomes impractical (up to 58 data points, see Figure 12). This size makes it challenging to scrutinize individual entries in the cause of loss column and determine which fields of action might be particularly relevant for all data points as a whole. Thus, for the sake of a consistent procedure across all FL spots, the information in the cause of loss column is

evaluated solely based on the outcome of the NLP analysis for each FL spot. Furthermore, with anticipated data updates, the possibility of more abundant data in the future is not only desirable but prospective (see Chapter 4.3.3).

3.4.3 Natural language processing – limitations

The employed approach was subject to the researcher's influence through prompt engineering. Prompt engineering refers to the process of modifying a query to facilitate a pre-trained model in accurately identifying target information (Yong et al. 2022, p. 5). Consequently, the wording of the input query can significantly impact the output of such a pre-trained model, which, in this case, was performed by a single author, thereby introducing researcher bias.

The selection of a minimal requirement of 10 cause of loss entries for each FL to be eligible for the NLP analysis was based on the researcher's discretion, potentially introducing researcher bias into the thesis.

Additionally, the crucial scientific principle of reproducibility is somewhat limited in this thesis. The ChatGPT API, powered by ChatGPT 3.5 based on the version from January 9, 2023, was utilized. However, this version has since been updated, and AI systems are known to undergo continuous improvements, potentially yielding slightly different results for the same task when computed now. However, it is important to note that even in a hypothetical scenario where ten experts in the field of FL collaborated to assign 247 causes of losses to the 13 fields of actions, achieving a consensus for each entry might still result in varying assignments if the process were repeated, especially if other experts in the same domain were involved. Therefore, even expert consultations might not guarantee perfect reproducibility in terms of obtaining exactly the same results.

A further limitation of the NLP analysis was that in evaluating the results of the assignment task, each cause of loss entry were assumed to be entirely equally important and therefore treated equally weighted. This simplified approach overlooked the possibility that a single individual may have contributed a disproportionately large number of cause of loss entries, thereby introducing skewness. Furthermore, the entries and their corresponding assignment task outcomes were not weighted based on the actual quantity of lost food. The latter could not be realized because only 0.07% of data points inside the considered scope actually contained quantity data on FL (CC37).

3.4.4 Multiple regression – selection of the technique

The idea behind applying multiple regression to certain features of the dataset is to evaluate whether there are significant correlations between the indicators of the LPI, each one representing different aspects of logistics and FL.

Per definition, according to Moore et al. (2006), "multiple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables, whose values are known, to predict the value of the single dependent variable. As the outcome of a regression analysis, each predictor value is weighed, the weights denoting their relative contribution to the predicted variable" (Moore et al. 2006, p. 235).

As previously mentioned in Chapter 4.4.2, there are certain FL spots in the scatter plot depicted in Figure 28, which cautiously suggest that, as for some FL spots, improved overall logistics of a country, represented by the overall LPI score, correlates with lower FL [%]. The multiple regression expands on this finding by utilizing the LPI indicators instead of the overall LPI score. Given that the primary objective of this thesis is to mitigate FL using logistical approaches, it is crucial to investigate which *aspects* of logistics should be enhanced to achieve a reduction in FL.

Before carrying out the regression analysis, several critical questions must be addressed, such as determining the data on which the regression will be performed. A trade-off needs to be considered between enhancing statistical power by using a larger sample size and preserving comparability across groups of data points.

The chosen trade-off is to limit the data analysis and its sample sizes to a combination of a food category and an SC stage, which, as previously mentioned, is referred to as an FL spot within this thesis. This approach is also advantageous as it aligns with the scope of the input data that were collectively analyzed within the NLP analysis. This alignment enables the synthesis of the NLP analyses' results and the multiple regression analyses' outcomes.

Another crucial task is to carefully select which predicting variables to include in the regression model, as incorporating irrelevant predictors may lead to a decreased accuracy in predicting important predictors. More complex models are particularly prone to this

phenomenon, which makes it essential to conduct sanity checks when building a model, based on background knowledge and reasonable expectations of the regression outcome (The Pennsylvania State University 2018).

Customs, as an LPI indicator, is commonly associated with the export of products and, therefore, excluded from the regression analysis since not all food products are exported. International shipments, as an LPI indicator, pertains to the export of goods as well and is consequently excluded from the regression analysis for the same reason. Infrastructure, logistics competence, and tracking and tracing appear to be relevant for the purpose of this thesis. Timeliness seems pertinent as well, however, it can be argued that it is heavily influenced by, or even a direct outcome of, the LPI indicators infrastructure, logistics competence, and tracking and tracing, and was thus omitted from the regression analysis. Therefore, the three variables used to predict FL [%] for different FL spots are infrastructure, logistics competence, and tracking and tracing.

3.4.5 Multiple regression – technical implementation

The independent variables in this study are the three, formerly six, indicators of the LPI, each represented by distinct features in the present dataset. The dependent variable is FL [%]. The predictor variables are numeric, ranging from 1.69 as the lowest LPI indicator value in the category of infrastructure to 3.56 in the category of tracking and tracing (CC25). Consequently, as all variables of the regression are numeric and on the same scale, that probably ranges between 0 and 5, the regression parameters can be determined directly without the necessity for normalization across them.

Upon closer inspection, data points that would be duplicates to the multiple regression must be identified and eliminated (CC64). These critical duplicates, initially distinct, underwent manipulation in Chapter 3.2.3, resulting in different values in the FL [%] column being equalized by summation. Nonetheless, before conducting the regression analysis, it is essential to reduce the manipulated data points into a single representative data point for each group of data points that were manipulated collectively.

This is because, by definition, they would inherently contain identical LPI indicator data, as they correspond to the same country. Moreover, they share the same FL [%] data, as they were formerly manipulated together. Consequently, incorporating them into the regression analysis would result in the contribution of non-independent data points to the analysis, as

they inherently possess the same values for both the predictor and outcome variables (CC64).

To ensure the trustworthiness of the regressions' results, certain measures must be implemented. As one of such measures, a minimum sample size requirement is established. However, there is disagreement among scholars regarding the appropriate sample size for multiple regression. Harrell (2015) synthesized multiple studies, suggesting that a 1:10 to 1:20 ratio of data points to predictor variables was sufficient (Harrell 2015, p. 72), while Austin and Steyerberg (2015) recommend a minimum ratio of only 1:2 (Austin and Steyerberg 2015, p. 636).

In this study, a 1:10 ratio was selected, necessitating a minimum sample size of 30 data points to determine the three regression coefficients. These regression coefficients are then compared with each other, as to observe which LPI indicator shows the strongest negative correlation with FL [%], which may suggest that this particular LPI indicator represents an aspect of logistics that would be especially crucial for reducing FL [%].

The modelled estimates data, dominating the FLI in terms of overall numbers of data points, may not be deemed sufficiently reliable for comparing countries with one another, a topic that is further discussed in Chapter 4.3. Consequently, modelled data was excluded from the regression analysis. This measure was not necessary in the case of the NLP analysis, since there is no modelled data that holds information on the cause of loss (CC68).

In Figure 13, FL spots with a sufficient sample size of 30 or more data points, excluding all modelled data, are presented.

Figure 13: Sample size eligibility for NLP analysis of FL spots (own figure, CC65)

In summary, as depicted in Figure 13, the majority of FL spots do not meet the requisite sample size threshold of 30 data points. However, three FL spots belonging to cereals and the FL spot of O&P on the storage stage of the SC meet the requisite.

The regressions, based on data of the different FL spots, were computed in R using the "summary" function that provides comprehensive information about the regression parameters and the values of the F-test and the t-tests.

The "Multiple Linear Regression R Guide" by (Stenroos and Dzubak 2018) provides a detailed methodology for conducting multiple regression in R, which this thesis follows. However, this study solely focuses on examining the model's output in R, using R's "summary" function and does not delve into more advanced steps of multiple linear regression, such as transformation, post-model reduction, or multicollinearity analysis, mentioned by (Stenroos and Dzubak 2018).

In analyzing the model's outcome, this thesis considers the regression coefficients, the F-test, and the T-test. The F-test is used to determine the existence of any relationship between the predictor variables and the response variable (Stenroos and Dzubak 2018), with an alpha level of 0.05 for rejecting the null hypothesis (no relation), since an alpha level of 0.05 is traditionally being used as a threshold in many scientific fields (Miller and Ulrich 2019, p. 1). If a linear regression model for a specific FL spot *as a whole* passes the F-test, the t-test is conducted on individual predictor variables using the same alpha level of 0.05. The t-test assesses the likely existence of a relation between *single* predictor *variables* and the response variable (Stenroos and Dzubak 2018). Both the F-test's result and the t-tests' results can be immediately read from the R console, as it outputs the result of the "summary" function.

In summary, this thesis recognizes potential limitations in the regression analysis and implements necessary precautions to ensure the trustworthiness of the results. A three-layered approach was employed to ensure the reliability of the regression analysis, and only if all three layers are passed, the data subset for the respective FL spot's regression results is considered to be trustworthy and presented in the thesis. The three layers consist of a sample size requirement of 30 data points, the F-test, and t-tests.

Figure 14 illustrates the visualization of the results obtained from the regression analyses. The figure is intentionally left blank without the values, as it is meant to demonstrate the method of the visualization of the regression analyses' results.

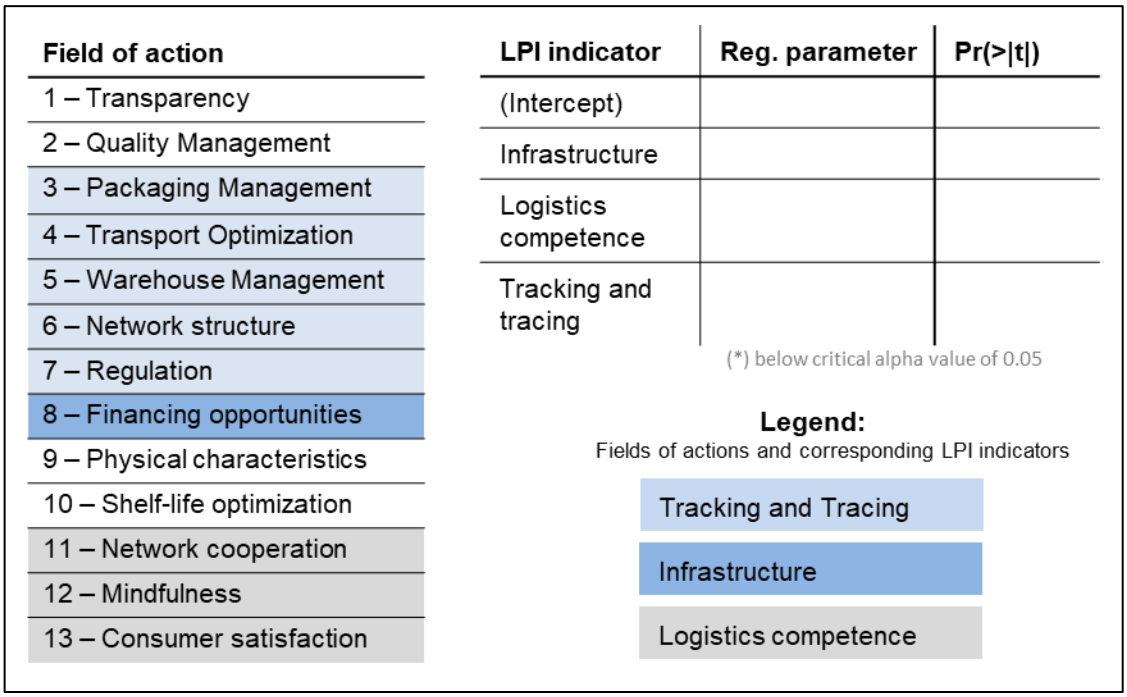


Figure 14: Visualization concept of the regression analysis (own figure)

After presenting the results of the multiple regression analysis, including regression parameters and t-test values, **secondary research question no. 3** arises, which is about how to connect the identified causes of losses, in this instance, the three LPI indicators, with the field of actions.

The alignment of the LPI indicators with the fields of action was primarily done based on the conceptual framework proposed by (Kleineidam 2020, p. 9) in the same research paper. In that study, the fields of action were classified into four distinct groups, namely: people, framework conditions, processes, and physical characteristics.

The various colors in the table on the left side of Figure 14 represent the corresponding LPI indicators that are closely related to each field of action. To begin with the most evident association, the category of "people" corresponds to logistics competence, as enhancing logistics competence empowers individuals to perform more effectively in logistics-related tasks. The tracking and tracing of products would likely improve SC processes, as stakeholders would be better connected, and weak points could be identified retrospectively. Therefore, those fields of action that belong to the "people category" as per Kleineidam (2020, p. 9) were matched with the logistics competence as the LPI indicator, those fields of action that belong to the category of "processes" as per Kleineidam (2020, p. 9) were matched with the LPI indicator tracking and tracing.

Regarding infrastructure, the LPI broadly defines it as "the quality of trade and transport-

related infrastructure (e.g., ports, railroads, roads, information technology) in a country” (World Bank 2023a). Among the 13 fields of action outlined by Kleineidam (2020, p. 10), financing is the primary mean through which infrastructure can be enhanced. When enhancing the logistical framework conditions, financing plays a crucial role in empowering communities or even countries at a larger scale to develop various aspects of infrastructure, including roads, railways, and digital infrastructure. Therefore, financing, as a field of action, was matched with infrastructure as an LPI indicator.

Moreover, it should be noted that for several fields of action, according to Kleineidam (2020, p. 10) there is no clear correspondence to a LPI indicator.

3.4.6 Multiple regression – limitations

A major challenge in conducting a regression analysis is dealing with non-linearity. Kotchoni (2018) posited that a linearity assumption may erroneously falsify a theory when the true relationship of interest is non-linear. Identifying the appropriate functional form for a non-linear relationship can be difficult (Kotchoni 2018, p. 2). To rephrase the problem, the relation between LPI indicators and the FL [%] at FL spots might not the linear but have a different relation. The linear regression would still only test for linearity.

Additionally, it is important to note that correlation does not imply causality. This concept was outlined by Schutt and O'Neil (2013), who described that confounding variables, which are variables that influence both the predictor variable(s) and the dependent variable, can lead to misleading conclusions when inferring causality from correlation (Schutt and O'Neil 2013, pp. 274–279).

Additionally, the assignment of LPI indicators to the respective fields of losses was based on the overall categorization of fields of action by Kleineidam (2020, p. 9). However, it also incorporated a certain degree of personal decision-making, as it required a certain degree of interpretation to assign the FLI indicators to the categories of fields of action, as per Kleineidam (2020, p. 10).

4 Exploratory data analysis

This chapter presents the exploration of the dataset, which has been merged from the FLI and LPI datasets and tailored to the scope of the thesis. As depicted in Figure 3, the exploratory data analysis (EDA) constitutes a fundamental building block upon which the two subsequent phases rely, utilizing data that has been preprocessed and cleansed.

A thorough understanding of the properties and shortcomings of the features is crucial for developing the decision assistant focused on reducing FL along SC stages in SSA. Schutt and O'Neil (2013) emphasize the significance of thoroughly understanding data and identifying patterns before constructing data products in the context of EDA. They advise against the immediate development of a model and propose a meticulous examination of the data's attributes. This method fosters a comprehensive intuition of the data, consequently enabling data scientists to devise more effective models in subsequent stages (Schutt and O'Neil 2013, pp. 35–37).

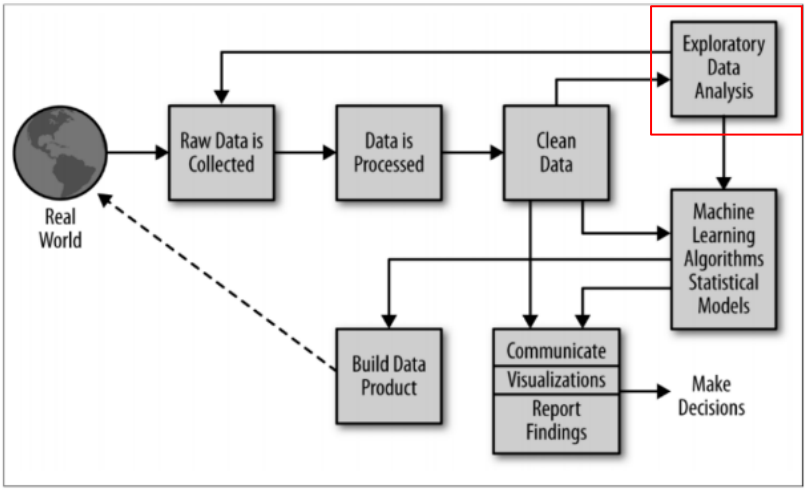


Figure 3: "The Data Science Process" according to (Schutt and O'Neil 2013, p. 41)

O'Neil (2013) highlights that the tasks, associated with the EDA, are to check the data's scale and format, finding missing data and outliers, summarization of the data and investigate comparisons between distributions (Schutt and O'Neil 2013, p. 36).

Notably, in the data exploration, the emphasis will be on summarizing the data in a manner that enables appropriate visualizations, assisting in the exploration of properties of selected features. Additionally, there is a focus on comparisons between distributions, specifically between manifestations of features and their interactions.

In Chapter 3.2.2, it was determined that the data had likely been pre-cleaned by the FAO.

As a result, the examination of the data's scale and the search for missing data and outliers serve as a means to verify the quality of the pre-cleaning process. Consequently, this finding renders the implementation of a separate "data cleaning" phase unnecessary within this thesis.

Additionally, it should be noted that within this Chapter 4.1, the dataset used had already been cropped according to the geographical scope, the time scope, and the scope of SC, as defined in Chapter 2.4 and Chapter 2.7.

4.1 Examining the data’s scale and format

As detailed in CC20, the dataset's scale, and format at this stage of the analysis can be observed. The following feature scales were modified in accordance with the nature and meaning of the features:

• The scale of the loss quantity column was changed from character to numeric, as the quantities are numbers and must therefore be represented as numeric values.

• The scale of the food supply stage column was changed from character to factor, as these values represent a distinct range of manifestations, not an open-ended range.

• The scale of the sample size column was changed from character to numeric, as the sample size refers to a number and must be represented as numeric values.

Numerous functions in R are sensitive to the scale of the data. Therefore, adjusting the scale for each feature is essential for further data analysis in R. Regarding the format, no peculiarities were found. The dataset uses dots as separators for large numbers and commas to denote decimal places, adhering to the American style. This consistent formatting is applied throughout this thesis and the R Markdown script.

4.2 Identifying missing data and outliers

According to CC22, there is no more missing data in the LFI columns, only in the column on the position 21 to 28, which are all columns depicting data of the LPI. Furthermore, CC25 indicates and explains that there are no outliers when considering the numerical data only. The rationale for outlier detection that was implemented focused on the following considerations:

M49 code and CPC code are codes and can assume many different values (non-determinable if outlier). The range of years aligns with the defined scope of 2000-2021,

and FL [%] entries are neither negative nor above 100%, which would be theoretically impossible. The index, starting with index number 8, might surprise some observers, but it is acceptable because the index was added in CC03 right after downloading the dataset to maintain a consistent numbering system, ensuring that data points can always be identified regardless of the operations performed on them. The retrieved results are already the outcomes after cropping the dataset to the scope of the thesis, which means data points with the index value 1-7 have already been removed.

4.3 Data summarization

Summarizing the data is crucial when dealing with a large dataset. Visualization techniques such as bar diagrams, heat maps, scatterplots, and box plots can effectively condense the data, presenting its characteristics in a more comprehensible manner.

4.3.1 Method of data collection

According to CC26, the adapted dataset for the scope of the thesis comprises 18,509 data points. To gain an initial understanding of the data's nature, examining the methods employed for data collection, which constitute its origin, is valuable.

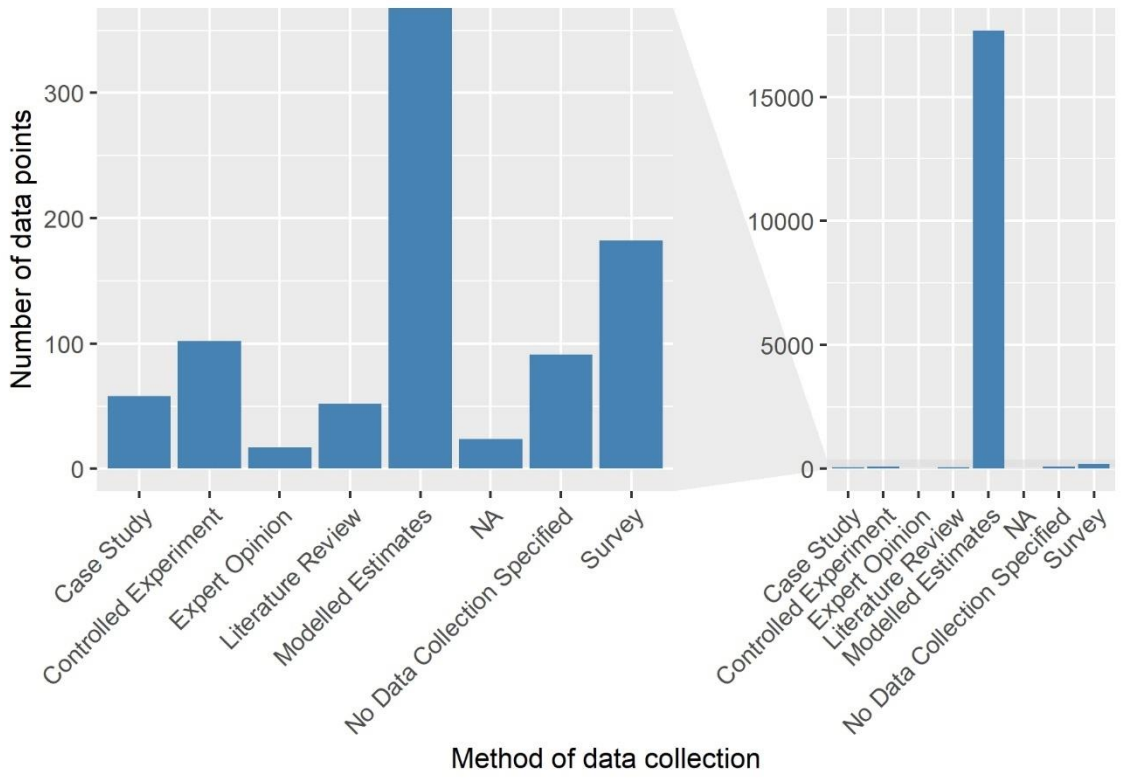


Figure 15: Method of data collection, number of occurrences (own figure)

Figure 15 illustrates the distribution of data points in the dataset across all provided methods of data collection. It is apparent that the number of modelled estimates surpasses the counts of data points generated by all other data collection techniques. Additionally, data derived from surveys, FAO questionnaires, and controlled experiments contribute a substantial number of data points, however, at a different level compared to the modelled estimates.

In Figure 16, Xue and Liu (2019) highlight differences in aspects such as time, cost, accuracy, objectivity, and reliability among various data collection methods. They created the present table, which offers valuable insights into the performance of each method of data collection used for the quantification of FLW.

	Method	Symbol	Time	Cost	Accuracy	Objectivity	Reliability
Direct measurement	Weighing	W	●●●	●●●	●●●	●●●	●●●
	Garbage collection	G	●●●	●●●	●●●	●●●	●●●
	Surveys	S	●●	●●	●●	●●	●●
	Diaries	D	●●●	●●	●●	●●	●●
	Records	R	●	●	●●	●●	●●
	Observation	O	●	●	●	●	●
Indirect measurement	Modeling	M	●●	●	●	●●	●
	Food balance	F	●	●	●●	●●●	●●
	Use of proxy data	P	●	●	●●	●●●	●●
	Use of literature data	L	●	●	●●	●●●	●

Figure 16: Advantages and disadvantages of different methods for FLW quantification (Xue and Liu 2019, p. 14)

Figure 16 demonstrates that various data collection methods show notable differences in different metrics of data quality and ease of retrieval. It is therefore essential to comprehend the composition of the methods of data collection employed in the FLI dataset. Examining the prevalence of these data collection methods within the FLI dataset will offer a more in-depth understanding of the overall data quality and guide any required adjustments or considerations for subsequent analyses.

With regards to modelled estimates data, Figure 16 indicates that Xue and Liu (2019) rate the factor of time as "moderate". However, it is straightforward that the time invested may vary per data point depending on the size of the dataset, as a model must be established once and can then be applied to a dataset of any size. Given that, according to Figure 16, modelled estimates data is generally low in accuracy and reliability and moderately good in objectivity, it is critical to examine the proportion of modelled estimates within the entire dataset, as demonstrated in Figure 16, as well as their prevalence across FL spots.

A potential danger is that modelled data, which can be classified as "low-quality" data

points within the dataset based on two out of three data quality metrics in Figure 16, may overshadow "higher-quality" data points concerning the metrics presented in Figure 16. It is crucial to be aware of this phenomenon. When constructing a decision support model, this consideration can be addressed by filtering the data before running models on it, among other possible measures.

4.3.2 In-depth analysis of modelled data in the Food Loss Index

Understanding the predominance of modelled estimates requires an investigation into their origins. A significant portion of the modelled estimates (99.8%) originates from the APHLIS database (CC32). Overall, the APHLIS database contributes 97.0% (CC32) of all data points to the dataset used in this thesis, while Xue and Liu (2019, p. 14) generally assert that modelled data has shortcomings in accuracy and reliability. Consequently, the characteristics and trustworthiness of data from the APHLIS database must be explored.

The APHLIS database provides estimates of post-harvest cereal FL in SSA, and it is supported by a network of local agriculturalists who provide country-specific data (Hodges et al. 2014, p. viii). The real-word postharvest FL data that underlies the APHLIS estimates were recorded between 1970 and 2014, with significant portions in the late 80s, early 90s, and late 2000s (Hodges et al. 2014, p. viii).

Deduction and induction are two fundamental mechanisms of scientific enlightenment. Deduction involves drawing conclusions from specific premises, while induction aims to create theories or generalizations through observation (University of Leipzig 2020/2021). This thesis, which seeks to develop a decision-support model based on real-world data, is classified as inductive research. However, the reliability of inductive conclusions may be compromised when a model relies on data from another model, such as in the present case.

Although the modelled estimates on the APHLIS database are originally based on real-world data, these real-world data are not only relatively outdated but also highly limited in quantity, and for certain SC stages, data is not even available (Hodges et al. 2014, p. viii). Consequently, the data from the AHPLIS database must be handled with caution, and conclusions drawn from this data should be considered in light of the nature modelled data in general and the characteristics of the APHLIS database.

4.3.3 Data availability throughout the years

Figure 17 displays how many data points were contributed to the dataset across the years

2000-2021. Understanding the composition of the data points with respect to their years of recording is essential. It can reveal whether the data is relatively up-to-date or if it is rather outdated, leaning towards the past side of the defined time scope. Additionally, examining patterns over time may allow for adjustments or considerations when building the decision assistant, as described in this section.

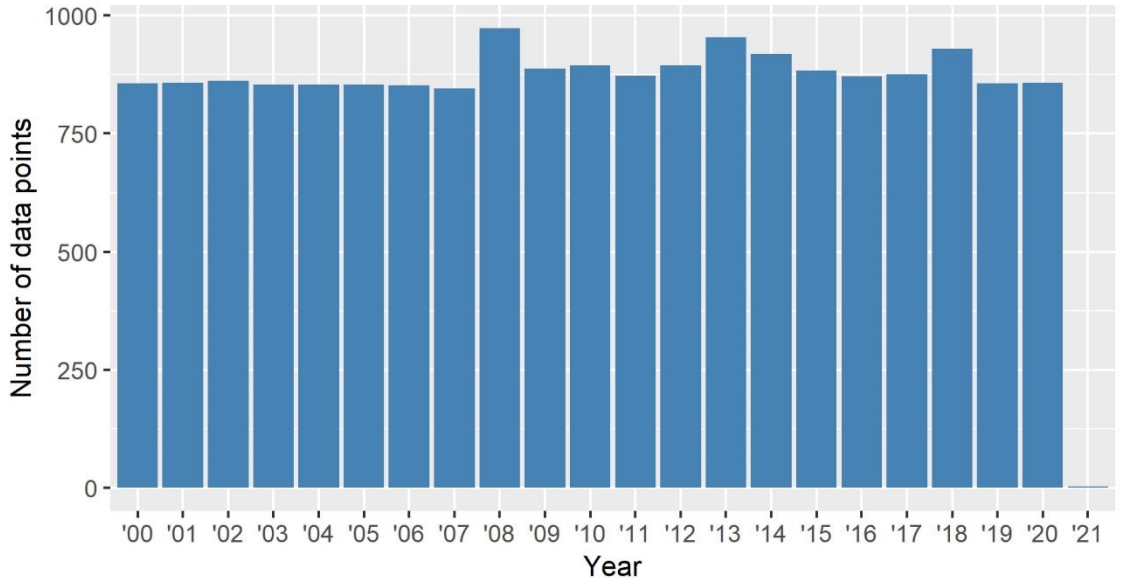


Figure 17: Data availability across years (own figure, CC35)

Upon initial examination of Figure 17, the number of data points contributed each year seems relatively stable, hovering around 400 data points annually. However, minor fluctuations are observed, with small peaks in 2008, 2013, including 2013's adjacent years, as well as in 2018.

As for 2021, the available data is comparably low. This could be attributed to the FLI dataset not being updated yet, or possibly due to data processing and validation procedures undertaken by the FAO, which might require an extended period of time. Upon initial inspection, the data seems to be fairly uniformly distributed throughout the given time frame. However, upon closer examination, it becomes evident that the high prevalence modelled estimates data points has not been taken into consideration.

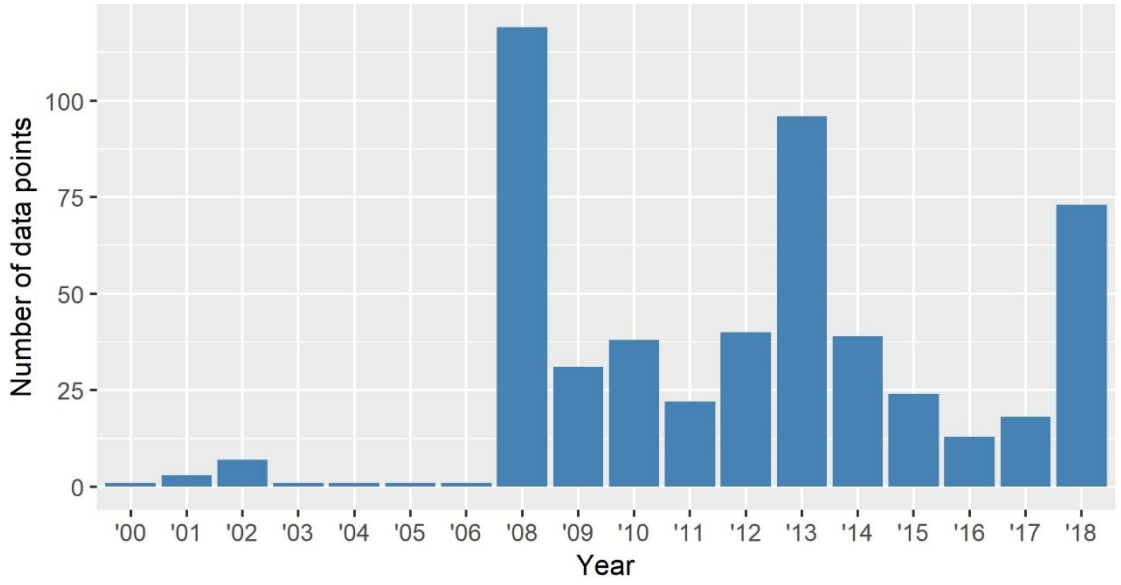


Figure 18: Data availability across years, excluding modelled-estimates data (own figure, CC36)

Figure 18, which displays the number of data points contributed each year, excluding modelled estimates, shows a significant increase in the number of observations from 2006 onwards, peaking in 2008 and 2013. In contrast, the period from 2000 to 2005 exhibits a limited number of observations. Notably, towards the end of the timeline, data occurrences seem to surge once again. The future development of this trend remains uncertain. However, as data levels have not returned to pre-2008 levels, it may suggest that FL research in SSA has garnered more public attention or importance compared to the times before 2008, or this increase could be attributed to specific recording reasons within the FLI database. For instance, the FAO may have only begun to actively work on the FLI from 2008 onwards, thus impacting the observed data trends. Interestingly, there is a distinctive peak in data contribution every five years from 2008 to 2018, specifically in 2008, 2013, and 2018. There is likely a reason for this pattern. It is possible that a large portion of data is processed in batches with a five-year time span.

Intriguingly, the dataset does not encompass FL data beyond 2018, indicating a possible delay in updating the FL data within the FLI dataset.

The key insights from this analysis suggest that, for constructing the decision assistant, it is not crucial to have data continuously updated in real-time, such as through web scraping. This is attributable to the fact that data updates from the FAO are not instantaneous. As a result, it is adequate to periodically examine the FLI database for new information, subsequently integrating this data into the decision assistant. Furthermore, there is no indication of a decline to pre-2008 levels, implying that a substantial quantity

of new data will persistently be accessible. This emphasizes the importance of designing the decision assistant as a model that can be continuously supplied with new data, even in the future, thereby establishing it as a long-lasting solution rather than a one-time resolution.

Peaks in data contribution are indeed significant and could be further investigated, as adapting the maintenance of the decision assistant model to this pattern and presenting results every five years could prove beneficial. Nonetheless, an in-depth investigation of this phenomenon was not undertaken in this thesis to preserve a manageable scope and length.

4.3.4 Data availability across food categories and supply chain stages

Figures 19 illustrates the data availability across food categories and SC stages. Upon analysis of that figure, it becomes apparent that the two food categories N&C and diary are considerably underrepresented. In fact, none of these food categories have data linked to all SC stages included in the analysis, making hotspot calculations unattainable due to the absence of comparative stages.

In conclusion, it is reasonable to decide that the two aforementioned food categories, dairy and N&C, are not appropriate for inclusion in the analysis and that the respective data points should be dropped from the dataset.

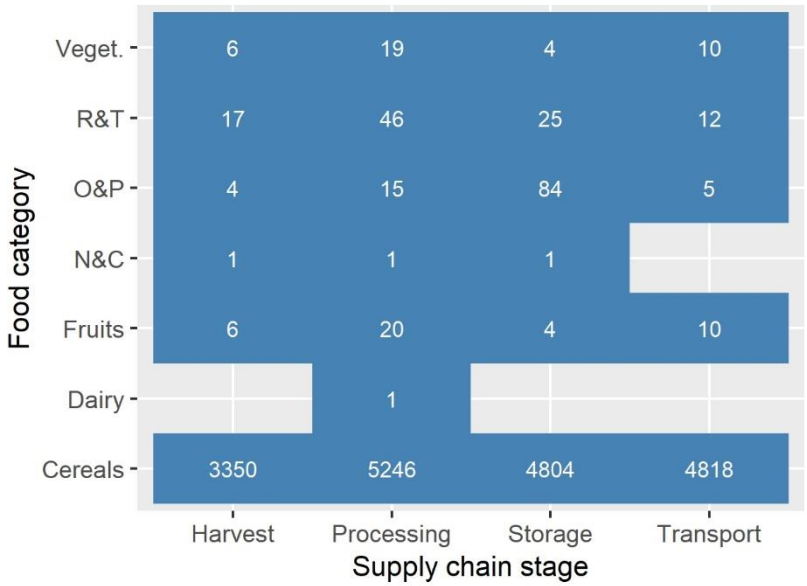


Figure 19: Map of data availability across SC stages and food categories (own figure, CC41)

Another insight gleaned from the data is that the food category cereals consistently exhibits a relatively large number of data points for each SC stage when compared to

other food categories. This even holds true when excluding modelled estimates (CC43). Therefore, the results of the decision assistant that will be computed for the food category cereals may be more meaningful and robust compared to the other food categories. Nevertheless, also the four remaining food categories will be considered in the analysis.

Altogether, the results presented in Figure 19 reveal that, despite the initially mentioned total number of data points (18,507) in this thesis' dataset, the amount of data for most FL spots is relatively small, even when including modelled data. This observation is particularly striking, considering that the data was collected over an extensive period of 22 years. To further clarify this statement, using an example from Figure 19, although the data was collected over a period of 22 years, there are only four data points for fruits at the storage stage of the SC. This remarkably low data coverage is in clear contrast to the overall number of 18,507 data points, which might initially give a deceptive impression of high data availability.

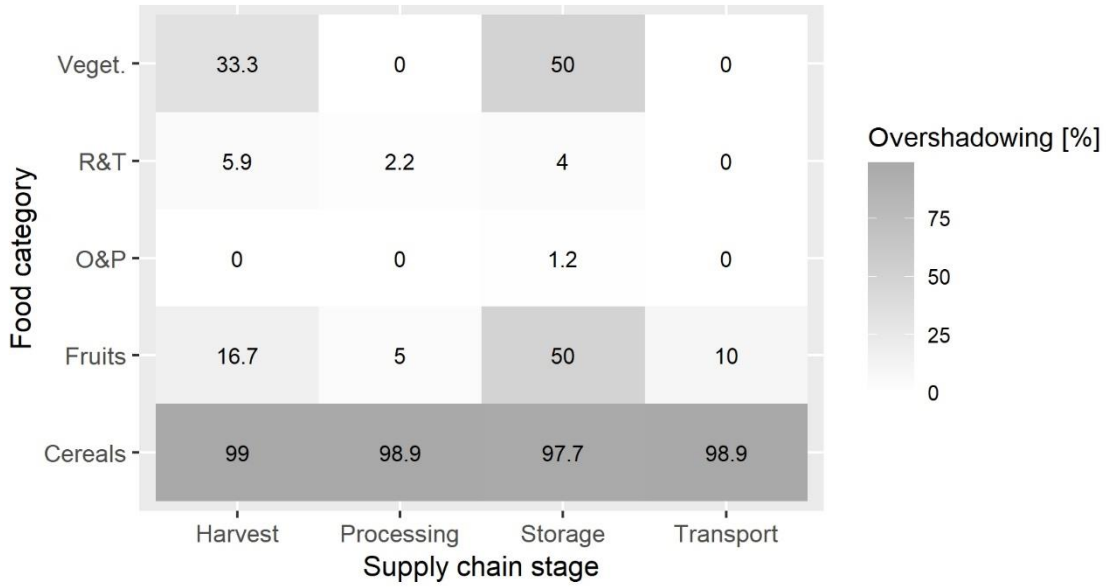


Figure 20: Prevalence of modelled data across food categories and supply chain stages (own figure, CC45)

Figure 20 presents the prevalence of modelled estimates for FL spots in proportion to their overall data availability. Since most data points in this dataset are modelled estimates and there are concerns about their data quality (see Chapter 4.3.1), these modelled estimates may "overshadow" the rest of the data, impacting the overall data quality in a highly negative way, which is presented in Figure 20. When observing Figure 20, it becomes evident that significant "overshadowing" does not occur uniformly at every FL spot. Instead, it primarily affects certain spots in a pronounced manner, predominantly for the food category of cereals. Consequently, the cereals data is predominantly

composed of modelled data, while this phenomenon is less prominent in all other food categories.

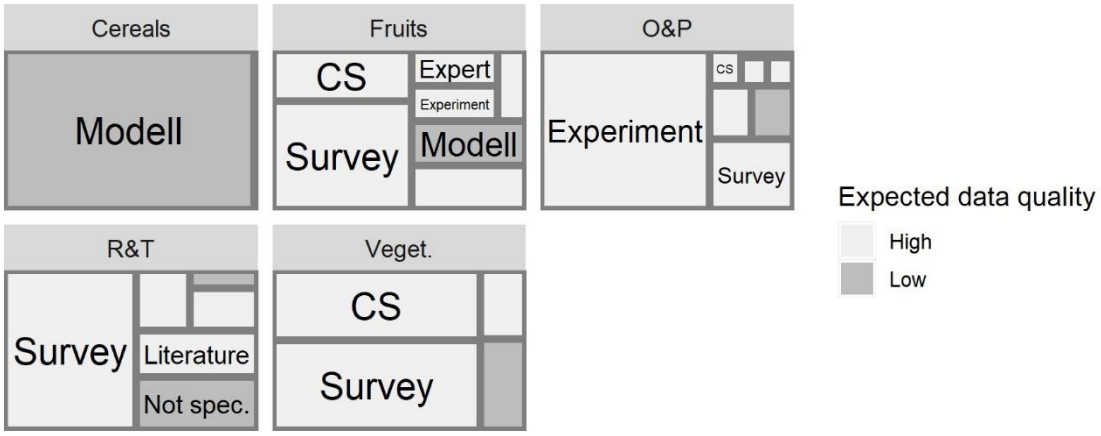


Figure 21: Two tree maps of expected data quality across food categories and SC stages (own figure, CC57)

Another data quality issue is detected in Figure 21. The fact that most of the O&P data is derived from experiments could potentially distort further data analysis, especially in conjunction with other data. This is because experiments often test for a specific aspect of FL, comparing different treatments against FL under the same framework conditions. This produces experimental results that are highly comparable to each other but may not accurately represent the true scale of FL. Therefore the results data analysis performed on O&P data should be interpreted especially cautiously in light of the method of data collection.

In Figure 21, it was assumed that, based on Figure 16, the data quality of modelled estimates is generally expected to be "low", while acknowledging that the term data quality is quite broad. Additionally, in cases where no method of data collection was mentioned, a pessimistic assumption of "low" data quality was made.

In summary, acknowledging and addressing the influence of modelled data on the overall analysis can help provide a more nuanced understanding of the FL patterns and inform targeted mitigation strategies. Broadly, it raises concerns if there is only one entity that contributes the vast majority of data in this dataset. Investigating the origin of the data in Chapter 4.3.1 was necessary to include only the data that will likely produce reliable results in the decision assistant.

In this thesis, it was assumed that the modelled data provide a rough estimation of FL magnitudes and can be utilized for the FL hotspot analysis. However, when identifying

the causes of losses from the data using multiple regression, the decision was made not to rely on these estimates. This is because, when performing the regression later in Chapter 5, the results are compared on a country-by-country basis, and the modelled estimates may not be accurate enough to represent the distinct differences in FL situations in individual countries. Furthermore, relying broadly on these modelled estimates for creating models would evidently contradict the principle of inductive research.

4.3.5 Data availability across food categories and countries

The heatmap illustrating data availability across countries and food categories in Figure 22 is a crucial tool in comprehending the degree to which the data points in the present dataset can be linked to individual countries, divided into food categories. This is of great significance, as it helps to identify any potential biases resulting from over-reliance on data from a particular country, respectively, from a group of particular countries. Such biases must be considered during hotspot analysis and during the construction of the decision model to combat FL, as a single country, respectively a single group of countries may not represent the entire SSA region adequately. Drawing inspiration from the FAO webpage's current dashboard on the FLI (FAO 2023a), a reconstruction of the dashboard’s heatmap is warranted in Figure 22.

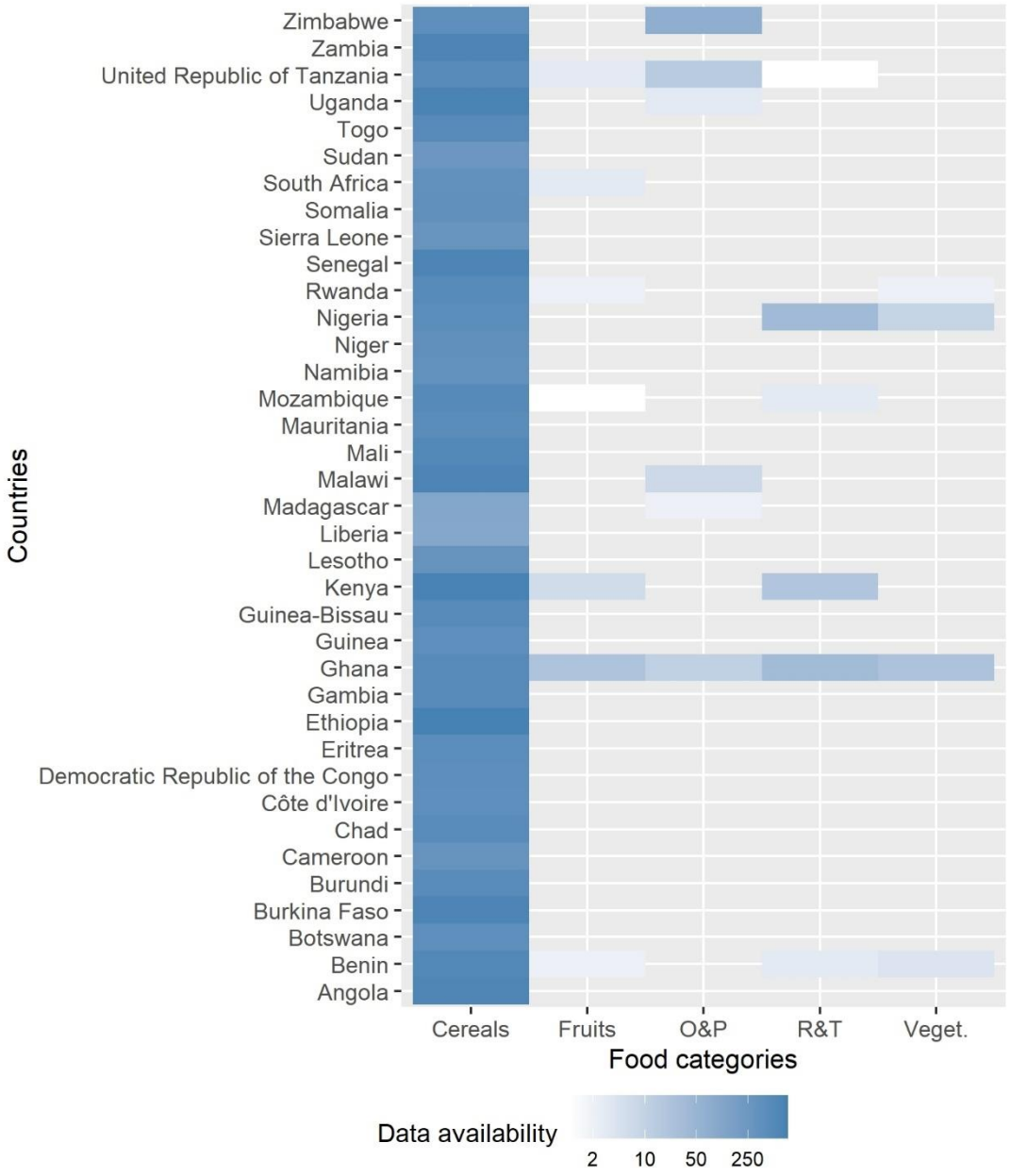


Figure 22: Data availability across food categories and countries (own figure, CC46)

The map displayed in Figure 22 shows the data availability for all SSA countries within the scope of this thesis, with a logarithmic scale used to detect changes among low values of data availability. Ghana stands out as an exception to the overall lacking data availability across countries, with relatively abundant data available for all food categories considered.

In general, there is no single country or apparent group of countries contributing data in such a substantial manner that it would heavily skew the overall results. Although potential clusters of countries disproportionately contributing to the dataset are not immediately apparent, it can be observed that some countries provide significantly more

data than others, particularly in food categories other than cereals. Given the importance of the relation between countries and FL [%] values, an in-depth analysis will be conducted on this topic in the final part of this chapter, which focuses on comparisons between distributions.

4.3.6 Hotspot analysis

Having established the sources and composition of the data, it is crucial to identify the combinations of food categories and SC stages where significant FL [%] accumulations occur. Gaining insight into such FL hotspots across food categories and SC stages is essential for providing recommendations on how to maximize the effectiveness of FL mitigation efforts so as to mitigate FL where they occur the most.

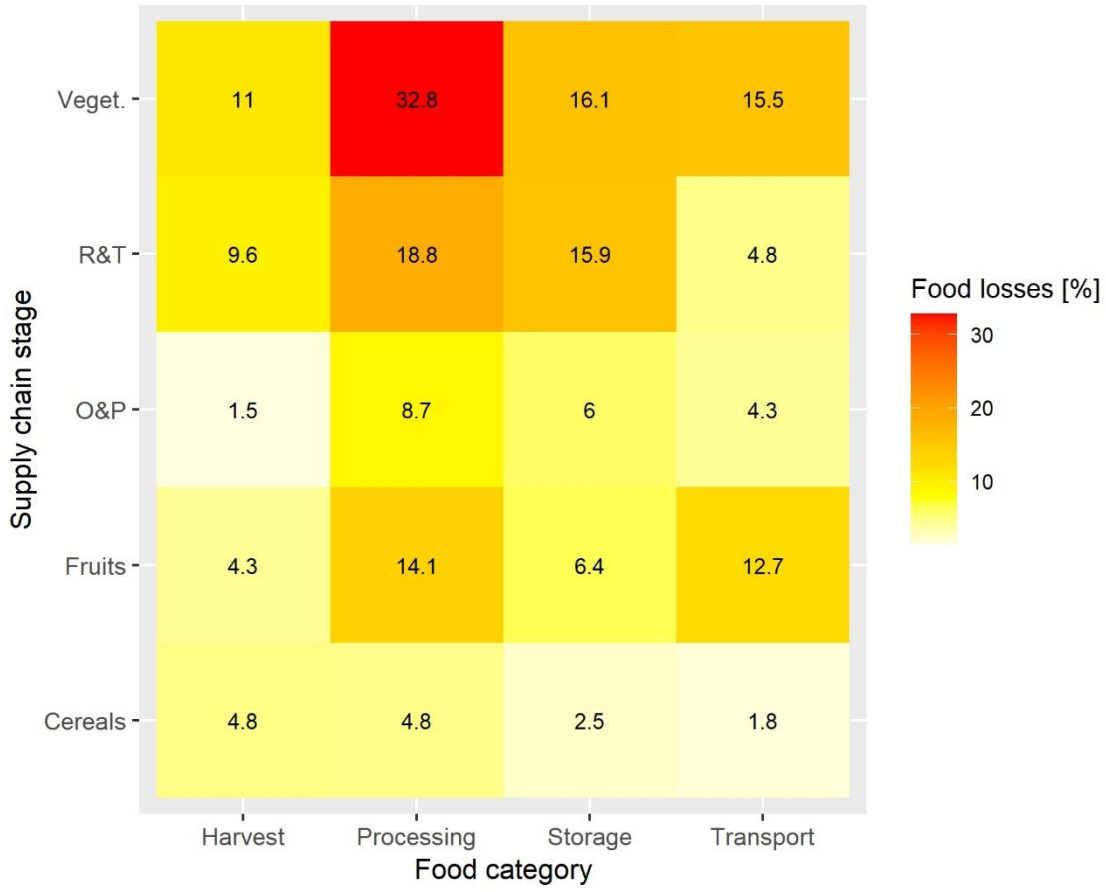


Figure 23: Heatmap of FL across food categories and SC stages (own figure, CC51)

As a result of that, Figure 23 displays that cereals consistently exhibit relatively low FL [%] across all SC stages. In comparison, fruits display increased susceptibility to FL, with aggregated FL values being significantly higher across all SC stages, except for the harvest stage, where losses are marginally lower than those of cereals. Notably, FL losses of fruits during processing (14.1%) and transport (12.7%) stages considerably surpass

those of cereals. O&P exhibit FL patterns similar to fruits across SC stages, however, at lower levels. R&T present moderately high harvest losses (9.6%) and relatively elevated losses during processing (18.8%) and storage (15.9%) stages, while losses during transport are significantly lower. Conversely, vegetables display the highest propensity for FL [%] at every SC stage, with a remarkable loss value of FL [%] during processing.

In summary, the processing stage consistently contributes to substantial losses across all food categories, which is potentially due to the diverse range of activities involved in the processing of food (CC12). FL values at other SC stages vary significantly depending on the specific food categories examined.

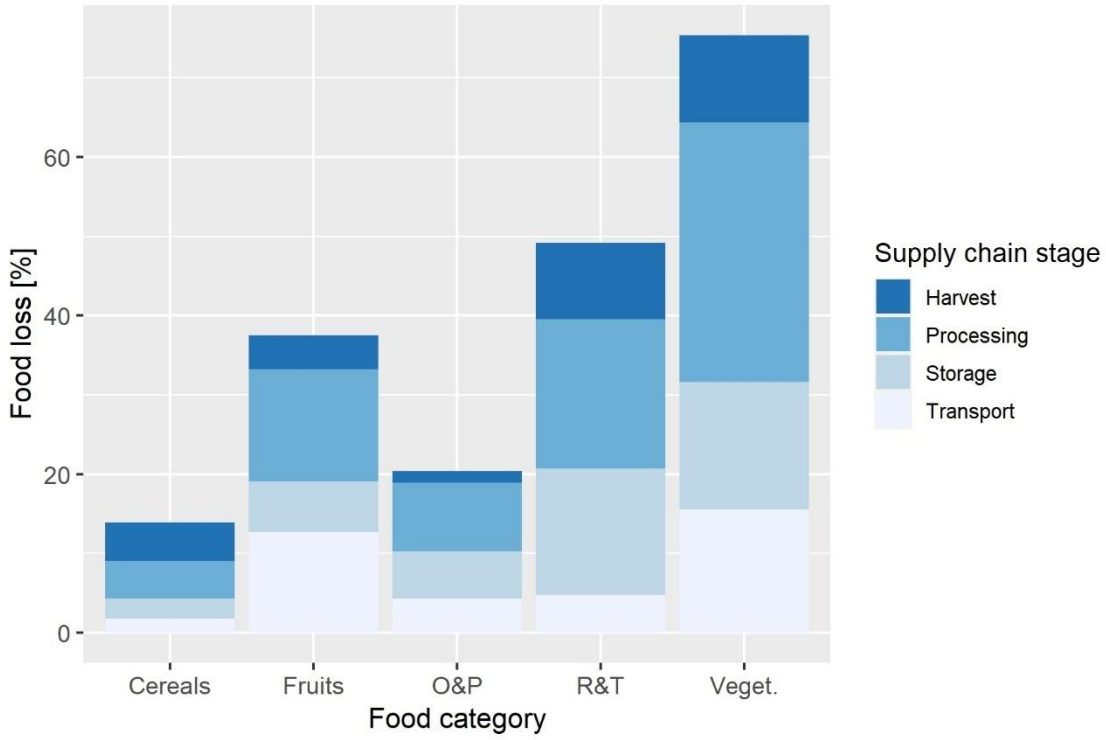


Figure 24: FL across food categories and SC stages, stacked bar chart (own figure, CC52)

Figure 24 presents an alternative perspective on the same information. When examining the overall FL for each food group, it is evident that vegetables experience the highest losses of 75%, followed by R&T at around 50%, fruits at approximately 35%, O&P at about 25%, and finally, cereals at roughly 15%. It is crucial to note that these percentages should probably not be interpreted in a cumulative manner in the sense of a definitive quantitative statement. Instead, they should be understood as a means to establish comparisons within and across food categories.

For example, it would be incorrect to conclude that only 25% of the initial ready-for-

harvest vegetables reach the retail stage. To arrive at a conclusion of how much FL there was in total across all SC stages examined, the FL [%] values would need to be multiplied by each other.

> This issue is illustrated using a fictive example:
>
> If one individual reports a 50% loss of oranges in Kenya in 2007 at the SC stage harvest and another individual inputs a 50% loss of oranges in Kenya in 2007 during transport, assuming there exist no other SC stages, then 25% of the ready-to-be harvested oranges would reach the retail stage, rather than 0%. This is because the FL [%] values would need to be multiplied instead of added.

Appendix E presents a hotspot analysis for the same dataset in the wide-data format. Although the assumptions for the wide-data format in Appendix E are somewhat unrealistic and the outcome of this section is not utilized for the remaining parts of the thesis, the wide-data format analysis still indicates that the data for vegetables has been collected in a highly selective manner. This suggests that, for the food category of vegetables, data on a particular commodity, country and year, inputted by one single entity was often available only for a small number of SC stages simultaneously rather than for all SC stages collectively. This pattern might be indicative of the strong influence of the phenomena described by (Affognon et al. 2015, p. 62), wherein FL research in SSA is conducted primarily where losses are anticipated, potentially leading to overestimation of FL values. This issue will be discussed in greater detail later in Chapter 5.2.1 of this thesis.

In response to secondary research question no. 1

As a result of Figure 23, the food categories, ordered by their susceptibility to FL, are: vegetables, fruits, R&T, O&P, and cereals.

Attention should, therefore, especially be paid to processing, and further evaluation should be undertaken to determine the extent to which processing FL can actually be avoided, particularly in consideration of losses inedible parts that potentially have occurred to a large extend while processing and might have been erroneously classified as FL by individuals contributing data to the FLI. It is important to consider whether all researchers and those individuals contributing the data share the same definition of FL when interpreting these findings.

FL spots with significantly high FL [%] values are: vegetables/processing, R&T/processing, R&T/storage, fruits/processing, and fruits/transport. However, for all these instances, it is crucial to compare the identified hotspots with the data abundance underlying the computations in Figure 19 and Figure 20 to assess the robustness of the results.

4.3.7 Exploring the entries in the cause of loss column

The data in the cause of loss column is particularly appealing because, unlike using LPI indicators that which may only provide an approximate indication of the underlying factors contributing to FL, the entries in the cause of loss column offer direct insights into the causes leading to FL. Nonetheless, the data presented in the cause of loss column is encoded in natural language, necessitating preliminary processing of this textual information. This contrasts with numerical data, which can be readily utilized for computations, such as calculating averages, without requiring additional manipulation. The information in the cause of loss column can be classified as unstructured data by the definition of Weglarz (2004), who describes unstructured data as any type of data without a conceptual definition, requiring human intervention to make it machine-readable (Weglarz 2004, p. 2).

The unstructured data in the "cause of loss" column of the dataset is written in natural language, making it challenging to work with. Nevertheless, it is an invaluable source of information, as it explicitly identifies the underlying cause of the respective FL, allowing for the development of targeted strategies to mitigate them. While the FL [%] contained in each data point is mandatory information, only 247 out of the data points contain data on the causes of losses (CC48), which corresponds to a coverage of the cause of loss data of ca. 1.4%, which implies that filling in the respective data is optional for those people contributing information to the FLI. Given that the data points are optional, requiring manual and activie filling in, when someone fills in this field, it is reasonable to assume that domain knowledge is involved.

Figure 25: Word cloud of entries in cause of loss column (own figure, CC48)

Figure 25 presents a word cloud of all entries in the cause of loss column that are included in the scope of the analysis, the font size corresponding to the frequency of the words mentioned, while words that don't convey actual meaning, such as "the" were removed. Although the word cloud does hardly allow for quantitative evaluations, it enables the viewer to get an overview of the content written down inside the column.

The diversity of reasons for FL, as illustrated in Figure 25, presents a significant challenge for the classification and analysis of the content in the cause of loss column. According to Figure 25, the causes of loss can be attributed to both biological and mechanical factors, making it difficult to establish a simple classification system. The flexibility granted to the column editors, upon examination of the actual dataset, appears to have lacked a prescribed format, which consequently results in the observed variability in the data.

Additionally, the inclusion of commodity names within the entries further complicates the analysis. Employing a simple algorithm to cluster the entries based on semantic similarities of words might lead to inaccurate results. For example, such an algorithm could potentially identify a false similarity between "rotten tomatoes" and "squeezed tomatoes" simply because the word "tomato" appears in both entries.

Given these complexities, it is essential to utilize a natural language model with extensive world-knowledge capable of detecting the meaning of groups of words, that is not merely able of counting word occurrences.

The presence of the term "quality" in mid-size may suggest that quality issues play a role, and the data might not be solely focused on losses but also encompass instances where food quality was reduced rather than entirely lost.

4.4 Comparison between distributions – exploring the role of the development stages in the magnitude of food losses

According to the World Bank (2022), SSA "is composed of low, lower-middle, upper-middle, and high-income countries, 22 of which are fragile or conflict-affected. Africa also has 13 small states, characterized by a small population, limited human capital, and a confined land area" (World Bank 2022). While SSA countries may share certain commonalities on a global scale, it is therefore essential to recognize its regional diversity as stated above.

Hence, it is recommended to not only identify hotspots by examining combinations of food categories and SC stages, but also to examine FL in the light of additional characteristics. A conceivable factor, potentially influencing FL and the availability of FL data, is the correlation between a country's level of development and its FL at certain FL spots, as they might be correlated. While a direct measure of wealth or general development stage is not present in the dataset, the overall LPI can be used as a proxy in this thesis. It is assumed that more developed countries exhibit higher overall LPI scores, meaning a better overall logistics performance, compared to less developed countries.

The primary objective of this subchapter is therefore to examine the relationship between development stages and wealth and FL in SSA. If a plausible association between the development stages of countries and FL is established, it may be beneficial to cluster countries accordingly and add this cluster group as a feature to the dataset. When results are later retrieved from the decision assistant, they can then be tailored to the specific development stage of different SSA countries.

In Chapter 4.3, the primary focus was on data availability, single-feature summarizations, and the use of FL [%] as a performance variable based on food category and supply chain stage features. This chapter, however, aims to explore the interaction between two performance variables, the overall LPI score and FL [%], where the term "performance

variable" implies here that certain manifestations of the variable can be considered superior to others. This investigation seeks to understand the interplay between these two performance variables.

4.4.1 Development stages of SSA countries and data availability

In Chapter 4.3.5, Figure 22 demonstrated that although no single country strongly dominates the dataset in terms of data points, there is substantial variation in the representation of countries, which could be attributed to their different stages of development. Prior to investigating the interplay between overall LPI score and FL [%], and potentially dividing the dataset into smaller subsets based on overall LPI score before feeding it into a decision assistant, it is crucial to assess data availability. Robust results can only be obtained when the countries contributing to the FLI database exhibit a certain degree of diversity with regard to their overall LPI scores.

As mentioned in Chapter 1 by (Hadi et al. 2020, p. 19), the majority of FL research has been conducted in industrialized countries, which is evident, as they possess greater financial resources for such studies. Prior to further dividing the dataset into country groups by wealth, it is essential to examine whether the same "principle" applies to the inner-African countries as well. Examining the pattern of data availability across countries is necessary before computing FL across these groups to gain insight into whether these groups contain enough data points for further robust analyses and the further splitting of them into food categories and SC stages.

Category I: The 33.33% of SSA countries in the dataset with the lowest overall LPI score, are termed "low performance countries".

Category II: SSA countries in the dataset whose overall LPI score ranges between 33.33% and 66.66% among the overall LPI scores of all SSA countries, are termed "medium performance countries".

Category III: The 33.33% of SSA countries in the dataset with the highest overall LPI score, are termed "low performance countries".

Each country performance group encompasses 13 countries (CC61). In the low-performance country group, there are 4838 data points, while the medium-performance country group contains 6604 data points. Lastly, the high-performance country group includes 6768 data points (CC62).

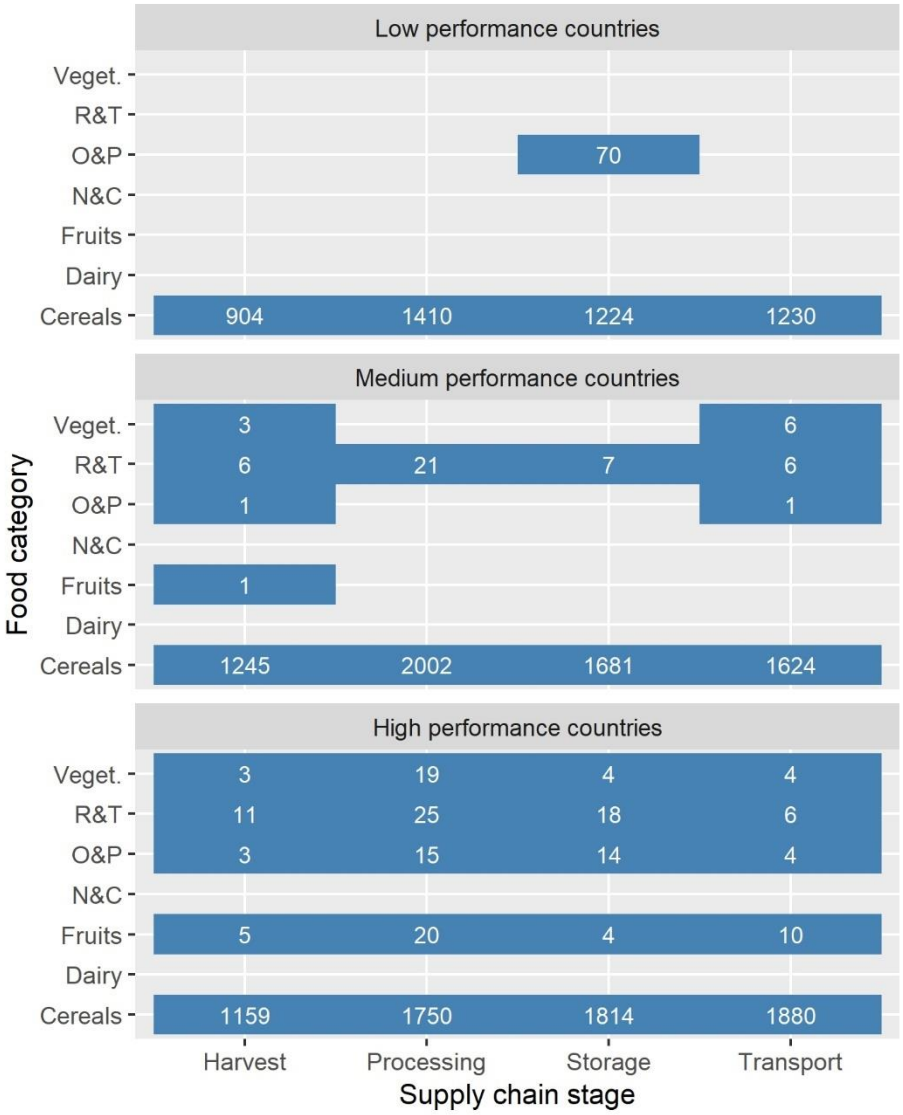


Figure 26: Data availability in occurrences of data points across "country performance groups" (own figure, CC62)

As demonstrated in CC61, each CPG comprises 13 countries. Figure 26 depicts the rough pattern that the "better" the CPG, the more comprising the quantity of available data. The CPG "low performance countries" only contains one combination of food category and SC stage, excluding cereals, for which data is available. Consequently, utilizing this subset of "low performance countries", which is highly limited in data coverage, would not be advantageous. In contrast to the "low performance country" group, the "medium performance country" group presents a larger number of data points with comprehensible data coverage, however still showing noticeable gaps. The "high performance countries" surpass both aforementioned groups in terms of data coverage, as all combinations of food categories and SC stages are covered, with a higher volume of data. Apart from the cereals data, which is predominantly composed of modelled data, the dataset is more abundant for every FL spot, except for R&T/processing and vegetables/transport.

In conclusion, the data availability has been found to be skewed towards countries with high overall LPI scores, which can be associated with superior logistics performance. As the combined data from “medium” and “low performance countries” is insufficient to cover all food categories and SC stages (see Figure 26), grouping the data according to overall LPI scores prior to feeding it into the decision assistant is not recommended. Subsequent analyses will be conducted on the dataset representing the entire SSA region regardless of the CPG to ensure a sufficient amount of data for obtaining robust results.

4.4.2 Interplay of LPI score and food losses

Although “low performance countries” have very limited data availability, and “medium performance countries” countries display patchy data coverage, which hinders a comprehensive clustering throughout the entire data analysis process, including decision assistance, it is still possible to conduct a visual examination of FL across various FL spots based on overall LPI scores. This analysis can provide valuable insights into the relationships between FL, logistics performance, and their interplay across different food categories and SC stages.

The scatter plot analysis depicted in Figure 27 supplies an approximate evaluation of how improvements in a country's logistics performance may influence FL reduction. Representing correlations through a scatter plot and engaging in a subsequent discussion seems more less cumbersome within the context of this thesis compared to calculating multiple regression models for multiple FL spot which show enough data points and diversity in overall LPI scores. This assessment may help address the question of whether significant improvements in a country's overall logistics performance could yield tangible effects on FL at specific FL spots.

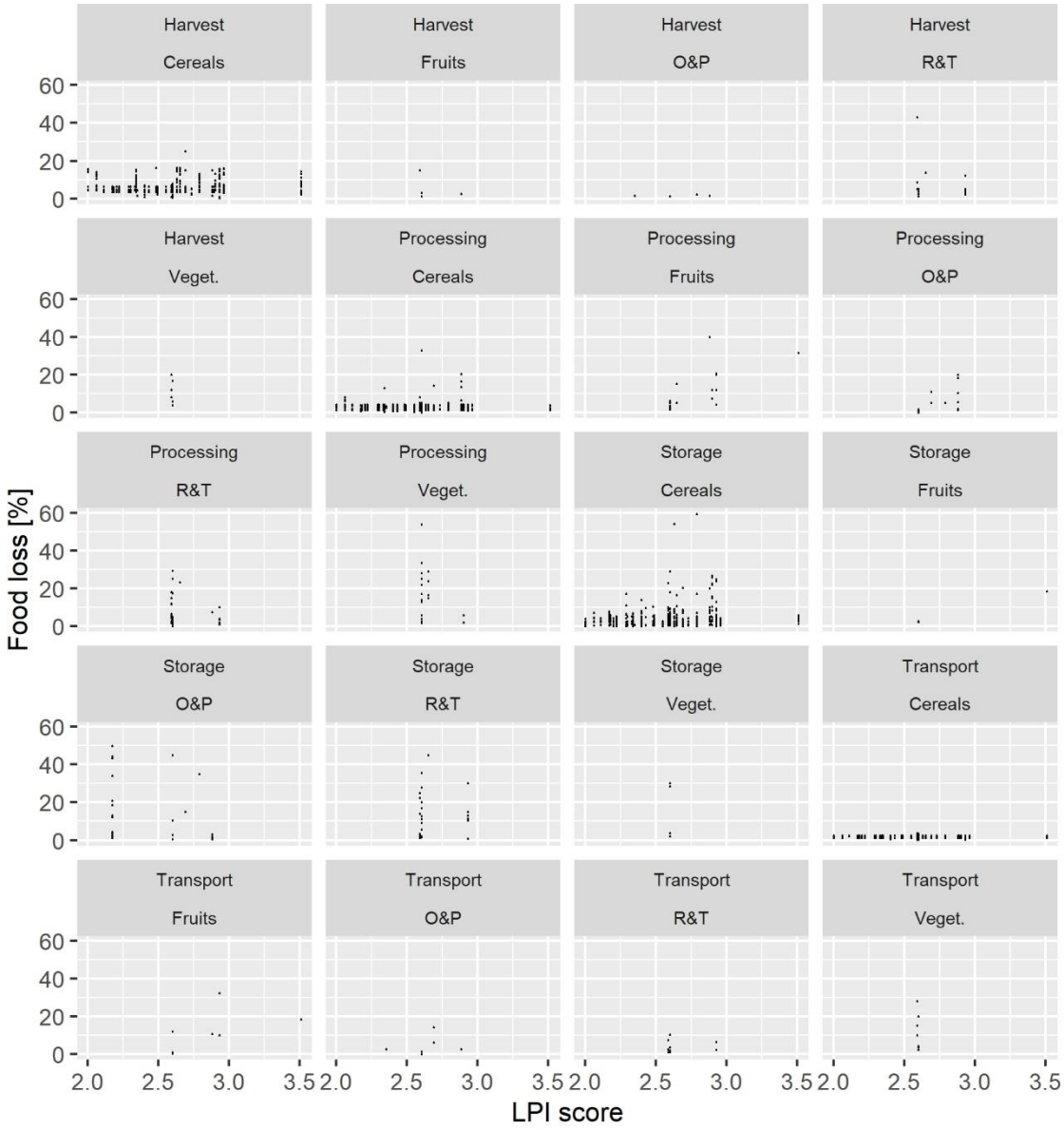


Figure 27: Correlation of overall LPI scores and FL [%] for FL spots (own figure, CC58)

When scrutinizing Figure 27, no definitive and omni-applicable relationship between development stage of a country, as approximated by the overall LPI score, and loss percentage for all FL spots is evident. The scatterplots results for the FL spots vegetables/farm, R&T/processing, O&P/storage, as well as vegetables/processing, suggest that countries with better logistical conditions show lower FL than countries with poorer logistical conditions. Conversely, the FL spots cereals/storage, fruits/processing and O&P/processing seem to indicate that higher LPI scores correspond to higher FL.

Although the data might imply that a higher LPI correlates with reduced FL, it is not possible to draw a definitive conclusion based on the scatterplot. This limitation is partly due to insufficient data for numerous combinations of food category and SC stage.

The computation of the scatter plot was essential for gaining insights into the potential interrelation between logistics performance and FL reduction. Although the results are not conclusive across all FL spots, they provide a foundation for further research into the phenomenon and an investigation of whether individual LPI indicators may have a decisive effect on FL. In general, it can be stated that, based on Figure 27, better overall logistics performance or a higher development level of a country does not automatically guarantee low FL.

This finding weakly suggests that reducing FL across FL spots might not entirely depend on overall logistics performance or development stages, but rather on more far-reaching factors that extend beyond overall logistics performance. Addressing the primary research question might require looking beyond just improving "logistics" countrywide and expecting significant reductions in FL in return. It is essential to comprehend the FL phenomena within specific countries for respective FL spots and develop customized countermeasures that align with local framework conditions. This highlights the necessity for action recommendations provided by a decision assistant that is as tailored as possible. The findings of this subchapter strengthen the position to create a decision assistant that, even though clustering according to countries was not feasible, offers recommendations tailored to the specific FL spots. This approach prevents the generation of overly generic recommendation, such as the need for "improved logistics", by considering FL in SSA as a whole.

5 Result communication - visualizations, report findings, and making decisions

In this section, the final results, as computed by the decisions assistant, for distinct FL spots are presented while maintaining a realistic perspective and providing transparency regarding the capabilities and limitations of the decision assistant. Figure 4 illustrates the current position in the project cycle.

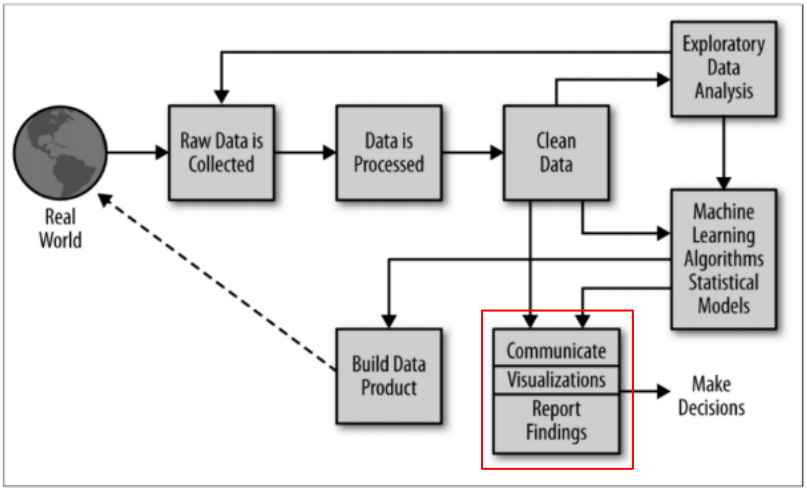


Figure 4: "The Data Science Process" according to (Schutt and O'Neil 2013, p. 41)

Schutt and O'Neil (2013) do not provide explicit guidance on the subtasks of this phase, and Herden's (2019, p. 230) descriptions are primarily business-focused at this point. Therefore, the three key elements of the phase, namely visualizations, report findings, and decision-making, each of these elements can be regarded as a method for communicating the results of the data analysis, are used to structure this chapter.

It is crucial to consider the unique characteristics of each individual FL spot, as the underlying conditions and reasons for FL may vary depending on the specific FL spot in question. Consequently, the visualization of the results, their interpretation, and action recommendations presented in this subchapter will be customized for each FL spot. The three steps of result communication include visualizing the results, interpreting the findings, and providing action recommendations. These three steps, always divided into three paragraphs, ensure a comprehensive understanding of the data and facilitate informed decision-making based on the analysis.

Eventually, this chapter will feature a discussion on the relative attractiveness of mitigating FL at the analyzed FL spots. Utilizing a "low-hanging-fruit" matrix, the most

promising FL spots for mitigation efforts will be identified.

The World Bank in association with the China Development Bank (2017) posited that Africa possesses leapfrogging potential across various industries, as long as proper governance structures are implemented. One prominent example is the mobile money sector, which was initially cultivated in Africa and subsequently expanded to other developing regions worldwide. Consequently, Africa represents a compelling testing ground for the adoption and adaptation of emerging technologies (World Bank, China Development Bank 2017, p. 159) Leapfrogging can be understood as a phenomenon where new entities, instead of following the technology path of their predecessors, invest in emerging technologies and rapidly catch up with technology leaders. Analogous to a leapfrog, this strategy allows latecomers to potentially surpass forerunners by capitalizing on advanced technologies, thus enabling them to compete effectively in the market (Lee et al. 2021, p. 123 based on Perez and Soete 1988).

Therefore, the action recommendations within this thesis will aim to present a rough and overarching strategy by utilizing and combining the most relevant fields of action compared to the overall benchmark, and these high-level recommendations will, in parts, be focused on fostering (digital) leapfrogging opportunities.

5.1 Report findings for cereals/harvest

Figure 28 presents the outcome of the NLP analysis for the FL spot cereals/harvest. The results were derived based on an overall number of 10 entries in the cause of loss column, which is important to note as different numbers of entries can indicate the robustness of the results, with more entries reducing randomness. Considering that the minimum requirement of ten entries has been barely met, the outcomes of the NLP analysis for this particular FL spot should be approached with caution. While the number of entries might not be explicitly addressed for the NLP analyses' results in subsequent FL spots, it remains critical to interpret all visualizations with an awareness that this figure can impact the findings.

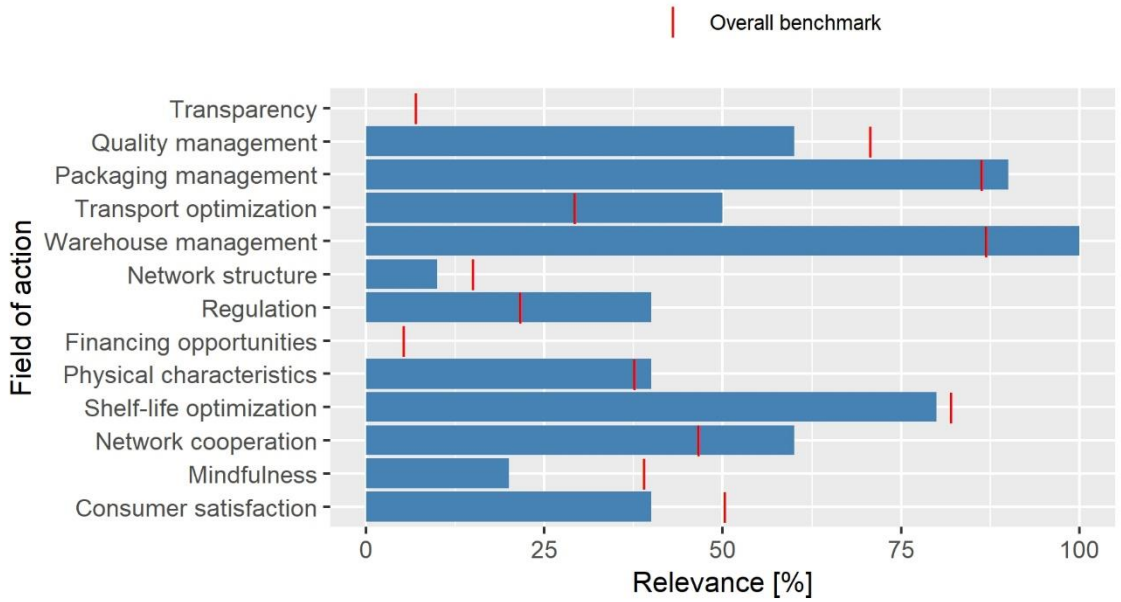


Figure 28: Result of the cause of loss analysis for the FL spot cereals/harvest, n = 10 (own figure, CC82)

As for the NLP analysis (Figure 28), the AI-computed results revealed that warehouse management, regulation, transport, and network cooperation were rated significantly above their overall benchmarks in terms of relevance. Conversely, the relevance of quality management, mindfulness, and consumer satisfaction was rated notably lower compared to their overall benchmarks.

Field of action
1 - Transparency
2 – Quality Management
3 – Packaging Management
4 – Transport Optimization
5 – Warehouse Management
6 – Network structure
7 – Regulation
8 – Financing opportunities
9 – Physical characteristics
10 – Shelf-life optimization
11 – Network cooperation
12 – Mindfulness
13 – Consumer satisfaction

LPI indicator	Reg. parameter	Pr(>|t|)
(Intercept)	-9.09	0.42
Infractructure	6.00	0.66
Logistics competence	25.44	0.04 (*)
Tracking and tracing	-25.09	0.01 (*)

(*) below critical alpha value of 0.05

Legend:
Fields of actions and corresponding LPI indicators

Tracking and Tracing

Infrastructure

Logistics competence

Figure 29: Results of the LPI indicator analysis for the FL spot cereals-harvest (own figure, CC67)

Figure 29 presents the outcome of the multiple regression analysis on the same FL spot.

It is the only FL spot that passed both the sample size requirement and the F-test and some t-tests and is therefore displayed here in the results. In the LPI indicator regression analysis, two regression coefficients passed the t-test, which are logistics competence and tracking and tracing. However, infrastructure did not pass the t-test, indicating that its effect on FL cannot be confidently established for the given subset of data. According to the regression analysis, increasing the tracking and tracing indicator values would lead to a higher reduction of FL [%] compared to enhancing the value of logistics competence if inferring causality from correlation. It is important to note that the absolute values of the regression parameters should not be emphasized in this context. They might erroneously suggest that improving logistics competence would lead to increased FL [%], as the regression model is designed to identify any regression parameters that best describe the distribution of data points. The model does not account for the fact that, in reality, all values of regression coefficients would likely need to be negative, as improved logistics would theoretically contribute to reduced FL. The primary focus should be on comparing the values of regression coefficients relative to each other. Hence, it can be concluded that for cereals at the harvest SC stage, if causality is inferred from correlation, enhancing tracking and tracing is more crucial for reducing FL than improving logistics competence.

As for the NLP analysis, the high relevance of warehouse management as a prospective logistical countermeasure may be explained by the assumption that many goods are stored in the field upon harvest and left without immediate proper storage. Furthermore, transport optimization is rated relatively high, which is understandable considering the importance of promptly transporting harvested cereals away from the field. Network cooperation also emerges as a relatively crucial factor, as it combines warehousing and transportation, tasks that may be fulfilled by different actors. As for the multiple regression analysis, the high importance of tracking and tracing, as determined by the regression analysis, may suggest that for this FL spot, it is particularly crucial to establish comprehensive oversight of food commodities along the SC, including the initial step of harvest. Consequently, this could help in mitigating inadequate communication among the entities involved. The NLP analysis highlights the relevance of network cooperation, suggesting that the harvesting of cereals should be seamlessly integrated into the overarching SC, encompassing both storage and transport. This observation implies a potential need to equip farmers with proper storage facilities directly at the harvest location. Given that a single farmer may not independently undertake all tasks and might

collaborate with other entities, such as those responsible for collecting cereals or other farmers sharing storage facilities, network cooperation would be a vital aspect of the process. The LPI indicator analysis aligns with this finding, as its results emphasize the importance of tracking and tracing, which can be a pledge that more information-sharing mechanisms should be in place to enable communication among stakeholders.

In summary, both the NLP analysis and the LPI indicator analysis indicate that integrated approaches, which combine various stakeholders such as farmers, transport companies, and processing entities, under the goal of FL reduction would be beneficial. This could be facilitated, for instance, through the development of dedicated mobile applications that connect these different parties. By ensuring that stakeholders are informed about farmers' upcoming harvest plans well in advance, they can make the necessary preparations ahead of time. Consequently, this can lead to enhanced productivity, more efficient collaboration among the involved parties, and possibly to FL reduction.

5.2 Report findings for cereals/processing

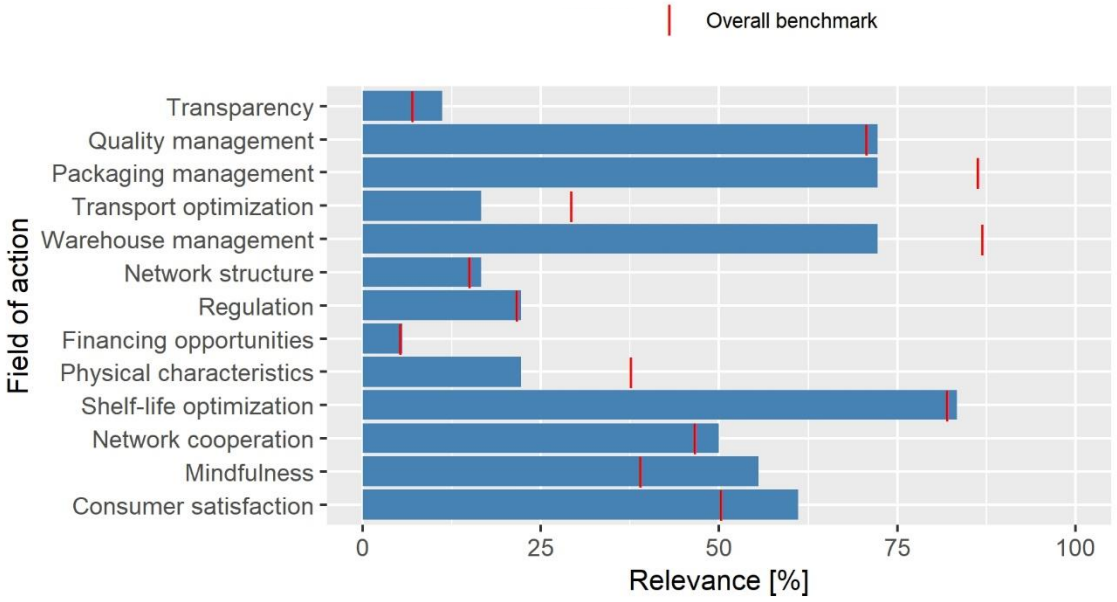


Figure 30: Result of the NLP analysis for the FL spot cereals/processing, n = 18 (own figure, CC83)

Key observations from Figure 30 include the following: Generally, the fields of action tend to be named less frequently compared to the overall benchmark. Only transparency, mindfulness, and consumer satisfaction are rated to be significantly more relevant compared to the overall benchmark. In contrast, packaging management, transport optimization, warehouse management, physical characteristics, are rated significantly less relevant compared to the overall benchmark.

The importance of mindfulness is understandable, as employees are usually involved in the production process and should therefore be skilled. This is especially relevant for cereals, as they are probably often not sold immediately but processed beforehand using milling or other techniques to make them usable for cooking and preservation. Consumer satisfaction is also a sensible consideration, as consumer preferences can be taken into account during the processing stage in various ways, including the shaping of the physical characteristics of products, which helps avoid the production of goods that do not sell or do not sell quickly enough, which could perish before their expiration date.

Possible action recommendations include providing education and training to employees through workshops and educational videos, making them aware of the problem of FLW and the role that they play to mitigate FLW at their workplace. Additionally, consumer satisfaction may be enhanced by improving the understanding of their preferences and buying patterns in order to process cereals in ways that align with consumer preferences and purchase timing. For this purpose, current retail data from retail outlets could be utilized and integrated into decision-making.

5.3 Report findings for cereals/storage

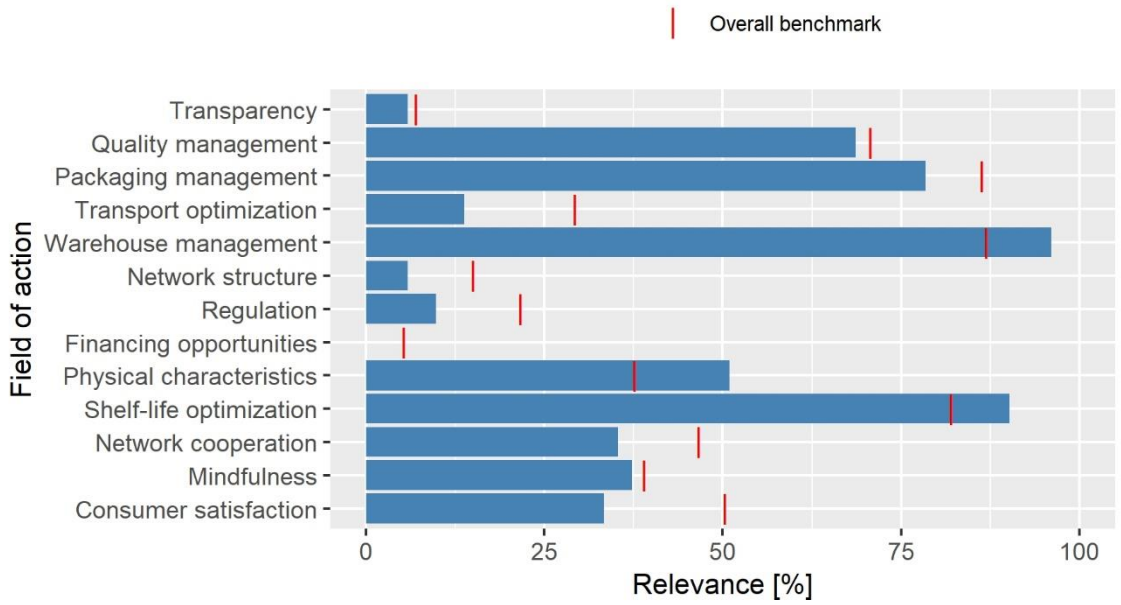


Figure 31: Result of the NLP analysis for the FL spot cereals/storage, n = 51 (own figure, CC84)

The results of the NLP analysis for the FL spot cereals/storage, as displayed in Figure 31, is that warehouse management is of paramount importance. Also, the fields of actions physical characteristics and shelf-life optimization clearly exceed that benchmark that was set. In comparison to the overall benchmark, transport optimization is decidedly less

relevant.

The fact that warehouse management clearly surpasses the defined benchmark has been expectable given the strong association between storage and warehouse management, which can be regarded as a sanity check for the NLP analysis at this point. The two other crucial aspects in this context are the consideration of physical characteristics and shelf-life optimization. The first of the two refers to the fact that the physical attributes of a product should be taken into account when designing appropriate SC processes (Kleineidam 2020, p. 10). Both fields of action emphasize the importance of considering the product when making decisions regarding storage, which may imply that currently cereals are often stored in a way that their characteristics are not taken into account, therefore storing them without making conscious and informed decisions on turnover rates, temperature, humidity, and other factors. On the other hand, factors such as regulation, network collaboration, and consumer satisfaction seem to be less pertinent compared to the relevance calculated as the overall benchmark. This suggests that the primary causes of losses are not directly linked to misalignment within a network of stakeholders or the overlooking of consumer preferences. Instead, these losses occur due to individual companies and their internal processes. Furthermore, in comparison to the overall benchmark, transport optimization is decidedly less relevant. This has also been expectable since, although transportation within a warehouse is a secondary function; at the storage stage of the SC, transportation may still hold some importance, although not as much as in the external environment.

As discussed, it is essential to prioritize the incorporation of shelf-life considerations into the storage of cereals. Evaluating whether the respective personnel have the necessary knowledge to make informed decisions on the given subject is essential. A potential solution might involve implementing a tracking system that keeps track of each bag of grain, so that by continuously monitoring these bags, there can be issuing alerts and the system could help prevent excessive storage duration, thus reducing potential losses while, at the same time, allowing to retrieval of statistical information of storage of cereals. Also, given the importance of shelf-life optimization in combination with warehouse management (storage), there should be an enhancing of the understanding of different types of cereals could potentially lead to improved storage practices. Furthermore, as part of warehouse management, it might be valuable to explore if distinct storage facilities or containers are necessary for diverse cereal varieties to maintain

optimal storage conditions, or if varying further storage strategies could be beneficial for different types of commodities in terms of shelf-life oriented aspects of storage strategies such as warehouse order picking and warehouse compartmentalization.

5.4 Report findings for cereals/transport

As for the NLP result of the FL spot cereals/transport, it has been observed that packaging management, transport optimization, regulation, financing opportunities, network corporation, and consumer satisfaction clearly surpass the overall benchmark in terms of relevance (Figure 32).

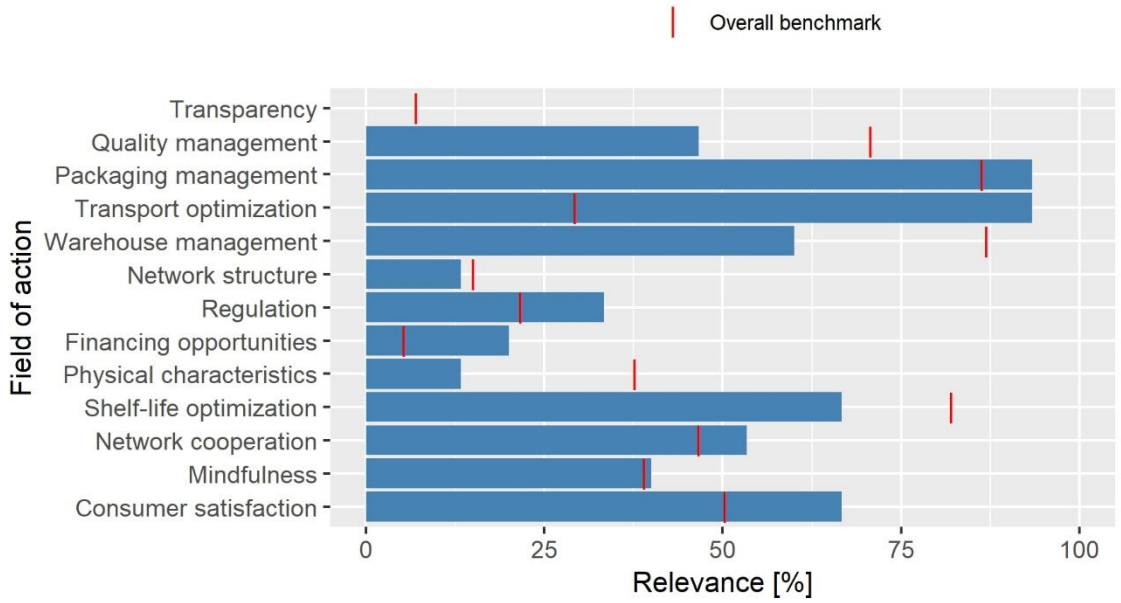


Figure 32: Results of the NLP analysis for the FL spot cereals/transport, n = 15 (own figure, CC85)

The enormous calculated relevance of transport optimization in the context of transporting cereals validates the NLP model's accuracy. Two other noteworthy aspects include the relatively high relevance of regulation and financing opportunities, the latter indicating a demand for enhanced infrastructure, which could be achieved through investments in both private property and public transportation routes. These investments may also pertain to the acquisition of transportation equipment, such as trucks, and their internal components, such as grids or bins. As mentioned, the calculated relevance of regulation exceeds the overall benchmark considerably, suggesting potential current challenges concerning road or rail regulations. One plausible explanation could be varying regulations across (small) African countries, which may hinder cross-border trade. In contrast, quality management, warehouse management, physical characteristics and shelf-life optimization demonstrate a significantly lower computed relevance

compared to their overall benchmarks.

As an action recommendation, improving packaging and transport management is of considerable importance, particularly as they can be means to alleviate challenges posed by the costly and time-consuming process of constructing new or enhancing existing infrastructure. In contrast to large-scale, collective, and nationwide endeavors, such as the building of new public infrastructure, individual companies can invest in transportation vehicles and transport equipment while concurrently improving packaging techniques. These efforts may potentially protect products from adverse external conditions, such as uneven road surfaces, to a certain extent. This is particularly crucial given the relatively low LPI score for infrastructure in the SSA region as a whole. By analyzing available LPI data, it becomes evident that the infrastructure indicator for the selected countries within the SSA region is generally markedly lower in comparison to the global average, as demonstrated in Appendix G. Given the substantial importance of transport optimization relative to the benchmark, it is imperative to also concentrate on improving transport management itself. As previously mentioned, altering the infrastructure may pose challenges due to the potential need for significant investments. The specific approach to enhance transport optimization would depend on individual circumstances, yet there could be a general strategy involving improved handling, network collaboration through collective and data-driven truck utilization and understanding of a network partners' business behavior to coordinate actions along SCs, and comprehending consumer demands to predict required deliveries. Potential solutions might include deliveries during early morning or night hours to circumvent traffic congestion, using real-time data of traffic conditions in metropolitan cities. On a higher administrative level, simplifying and harmonizing transport regulations, akin to the European Union model, might prove beneficial.

5.5 Report findings for fruits/processing

In the context of fruits/processing (see Figure 33), the calculated relevance of various fields of action is generally lower than their overall benchmark, barring a few exceptions. Notably, the computed relevance of packaging management significantly surpasses the overall benchmark. In contrast, the relevance of fields such as quality management, regulation, network cooperation, mindfulness, and consumer satisfaction is considerably lower compared to their overall benchmarks.

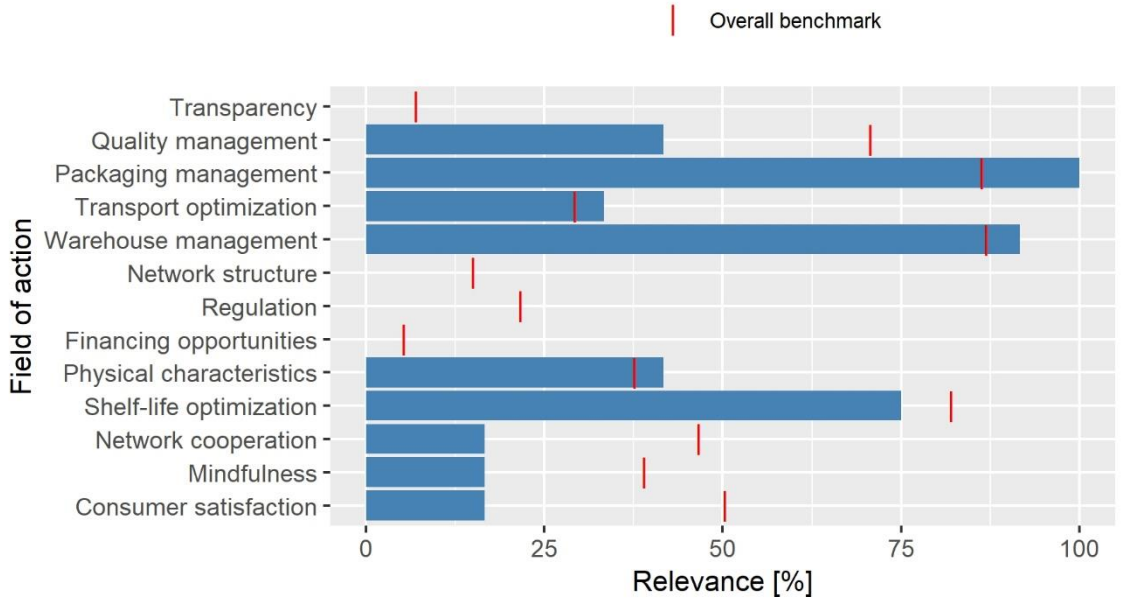


Figure 33: Results of the NLP analysis for the FL spot fruits/processing, n = 12 (own figure, CC86)

When critically evaluating the NLP analysis results for this FL spot, it can be observed that the computed relevance of four fields of action is equal to 0%, and the identical value of computed relevance is present for three additional fields of action. This similarity in computed relevance values across multiple fields of action raises suspicion. If the computed relevance values were indeed similar for many fields of action, this would suggest that the underlying causes of losses for the assignment task may have been relatively similar, potentially involving the identical causes of losses, inputted by the same individual. Given the small sample size of only 12 entries, this scenario lowers the trustworthiness of the computed results for this FL spot.

Compared to the food category of cereals, it can be hypothesized that many fruits enter the market without undergoing extensive processing prior to sale, as many fruits that can just be consumed when cleaned might typically not undergo procedures like milling, pressing, baking, frying, although some drying may occur. The high relevance of

packaging management suggests that the cause of loss column may encompass a relatively large number of direct or indirect packaging issues. This may be because there are fewer processing steps to consider apart from cleaning the commodities, which may lead to the remaining causes of losses dealing with packaging. The frequent naming of the respective field of action could also indicate that current efforts in this area may not be effective. This assumption concerning the elevated importance of packaging in fruit processing is further supported by the observation that the computed relevance of quality management and mindfulness are notably lower than the overall benchmark. Both fields of action might be essential to physical processing, such as cutting or applying heat, which encompasses processing not related to packaging. Consequently, the lower relevance attributed to these two fields of action further affirms the suggested significance of packaging as the primary factor contributing to FL for that FL spot. Moreover, the delicate nature of many fruits might contribute to the heightened emphasis on packaging management. If packaging is not implemented adequately, fruits are therefore at a higher risk of spoilage along the SC during activities such as transportation, stacking, and handling, compared to many other types of food commodities.

As an action recommendation that appears to be especially striking, innovative solutions must be sought to enhance the packaging of fruits. However, waste management in SSA still poses challenges that hinder countries' progress, and trash is usually not recycled (Debrah et al. 2022, p. 11). The challenges of waste management should be considered when developing packaging solutions. Potential packaging innovations could be tailored specifically for the African market, such as biodegradable materials that, at the same time, can withstand local weather conditions, including heat and humidity. This approach could mitigate the environmental impact of using packages while still addressing the need for adequate packaging in the region. Innovative packaging solutions may be developed in the form of lightweight, single-use materials, potentially allowing consumers to retain the packaging for personal use. Moreover, a circular management system for load carriers could be established, ensuring the reuse of packages by implementing smartphone scanning technology for identification purposes of the load carriers. Regardless of the approach, it is essential that packaging adequately safeguards the fruits from external influences throughout the supply chain (SC). In cases where there might be a shortage of modern trucks equipped with appropriate bins and grids, reinforced packaging can serve as an interim solution.

5.6 Report findings for O&P/storage

Upon examining Figure 34, which displays the computed relevance for the FL spot of O&P/storage, it is evident that nearly every field of action has been assigned greater relevance than the overall benchmark, with only a few exceptions. These are transparency, transport optimization, and mindfulness.

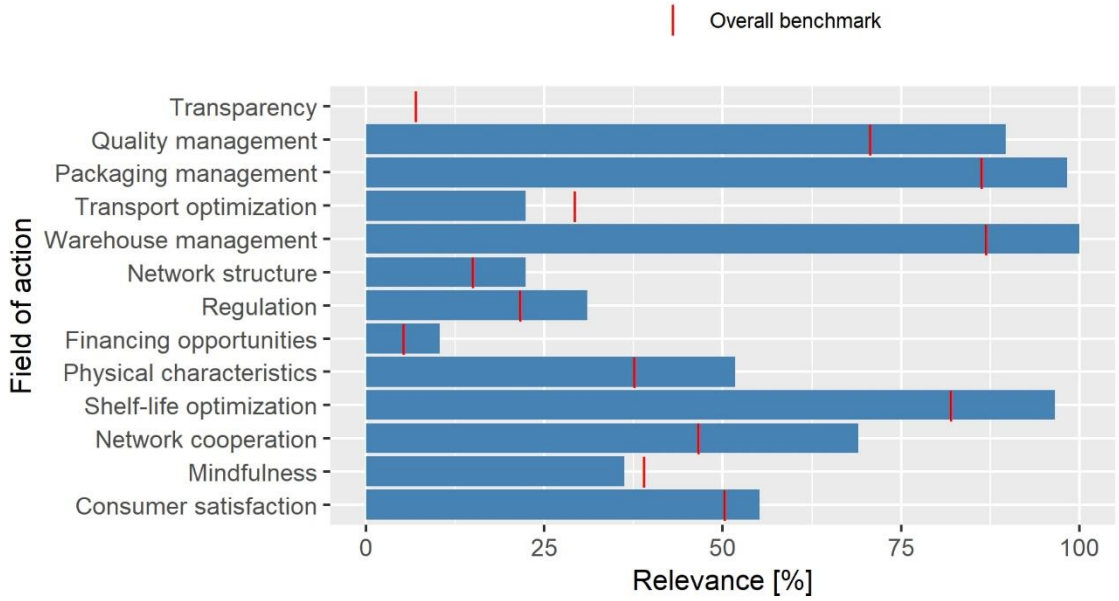


Figure 34: Results of the NLP analysis for the FL spot O&P/storage, n = 58 (own figure, CC87)

The analysis of fields of action presents a complex picture, with 10 fields of action rated above the overall benchmark. This complexity suggests that a single, all-encompassing solution may not exist to address the challenges faced when storing O&P in SSA. In particular, network cooperation significantly exceeds its overall benchmark, highlighting the importance of improved communication among stakeholders. The lower relevance of transport optimization is understandable and serves to validate the results, as this analysis focuses on the storage level, where transportation might not play a significant role. Instead, transportation likely plays a more substantial role outside the storage facility. In the context of O&P, it can be assumed that its SC is relatively intricate and staged, perhaps one more similar to the SC of cereals, as opposed to the one of fruits that is assumed to be usually rather simple. Oilseeds may contribute to a larger value chain, as they undergo processing to extract oil, which is then utilized in the production of various goods. Similarly, pulses may require preservation, such as canning or conservation, for extended storage and future use.

In light of these assumptions, improved collaboration and integration among stakeholders

in the O&P sector may result in enhanced overall efficiency and reduced FL. This suggests that adopting an industry-wide approach, mirroring the structure of a large corporation could be particularly advantageous for this sector. Given a high mobile penetration rate in SSA (World Bank 2023c), leveraging digital networking through smartphones could prove advantageous in enhancing SC visibility by providing real-time data updates and generating alerts in case of organizational misalignments. Mapping the SC with data in this context can be particularly beneficial for identifying causes of losses and the responsible entities, thereby enabling the rectification of errors. Furthermore, these FL information can be integrated into companies' utility functions, incentivizing behavior that minimizes storage losses of O&P.

5.7 Report findings for R&T/processing

The pattern of the relevance of each field of action in Figure 35, which shows the results of the NLP analysis for the FL spot, R&T/processing, reveals a general alignment with the overall benchmarks, barring a few notable exceptions. Quality management and consumer satisfaction display significantly higher ratings compared to the benchmark, whereas shelf-life optimization scores moderately above the overall benchmark.
On the other hand, physical characteristics exhibits considerably lower scores compared to the respective benchmark.

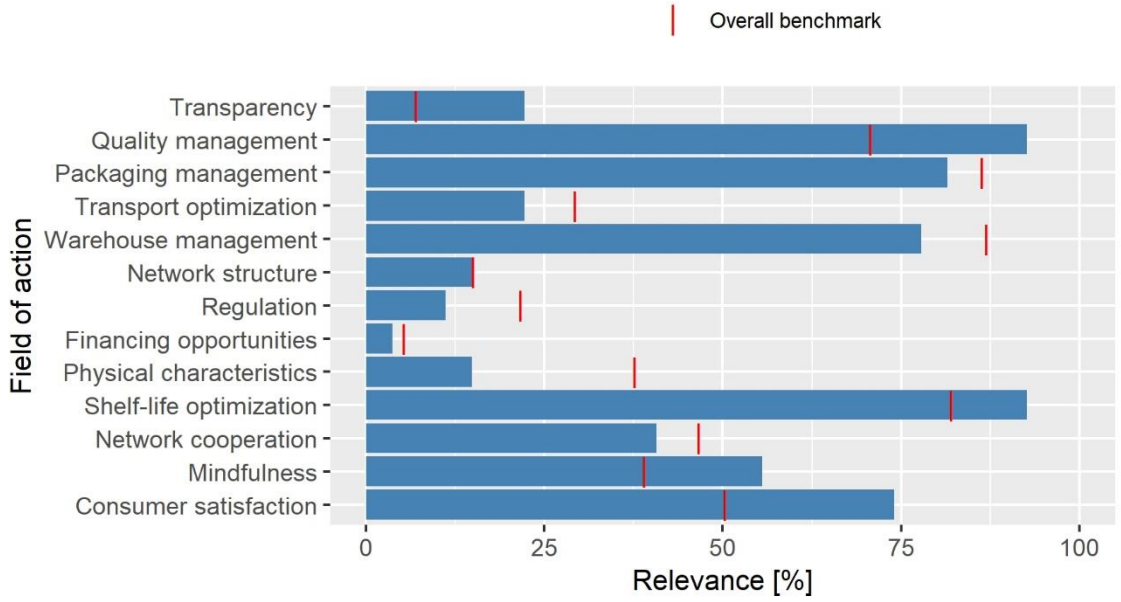


Figure 35: Results of the NLP analysis for the FL spot R&T/processing, n = 27 (own figure, CC88)

The pattern of the relevance of each field of action in Figure 38, which shows the results of the NLP analysis for the FL spot, R&T/processing, reveals a general alignment with the overall benchmarks, barring a few notable exceptions. Quality management and consumer satisfaction display significantly higher ratings compared to the benchmark, whereas shelf-life optimization scores moderately above the overall benchmark.
On the other hand, physical characteristics exhibits considerably lower scores compared to the respective benchmark.

The denoted high relevance of quality management indicates that this field of action is particularly relevant for mitigating FL at the respective FL spot, and it is suggesting that a sophisticated quality management system may currently be lacking. As for this and for all other FL spots involving processing, consumer satisfaction may be viewed with some skepticism, as the AI may have assigned it this way because any spoilage of food, leading to quality reduction or total loss, results in the food product losing value in the consumers' eyes without the problem being in the sphere of the consumer. At the same time, the high relevance of consumer satisfaction can be comprehensible, since the processing SC stage, unlike the other three SC stages, allows for heavy modifications of the product, adapting it to the consumers' needs.

To address the identified areas of improvement, it is recommended that stakeholders concentrate on enhancing quality management, which, for example, can be achieved by implementing smartphone applications that deal with quality management and by employee training seminars at processing sites in SSA. At this time, various online quality educational programs and certifications exist. The implementation of a quality management system specifically designed for SSA's R&T market at the SC stage of processing could be considered. Lastly, optimizing batching practices with a focus on product shelf-life considerations can help minimize spoilage and waste along the SC.

5.8 Report findings for R&T/storage

The NLP analysis (see Figure 36) shows that only warehouse management, shelf-life optimization, and mindfulness moderately exceed their overall benchmarks, while regulation and network cooperation fall significantly below the values of their overall benchmarks. These findings draw parallels with the NLP analysis' results for the FL spot fruits/processing, where both warehouse management and packaging management surpass the overall benchmarks and network cooperation is deemed less relevant compared to their overall benchmarks.

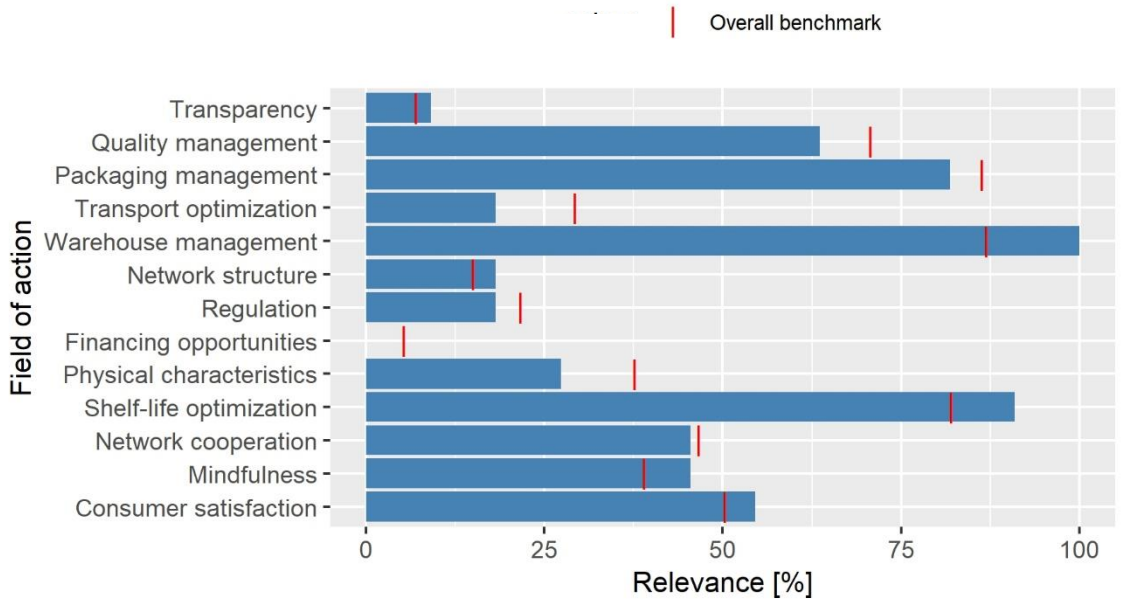


Figure 36: Results of the NLP analysis for the FL spot R&T/storage, n = 11 (own figure, CC89)

The high value of relevance in warehouse management, again, serves as a validation of the reliability of the automated assignment powered by AI system, as it is evident that warehouse management and storage are closely interrelated. Furthermore, it is crucial to consider the shelf life of products during storage to ensure goods are not stored beyond their capacity of remaining unspoiled. Interestingly, mindfulness is also rated more relevant than the benchmark, suggesting that in the case of roots and tubers storage, many losses could be prevented if employees were more attentive or better trained with regards to the proper storage of R&T. After all, a relevance rating of mindfulness significantly moderately its benchmark should indicate that the many of the respective losses of R&T arose from human mistakes.

Action recommendations include enhancing warehouse management through smartphone applications to optimize storage and logistics processes. As R&T are probably stored longer than certain types of highly perishable fruits or vegetables, since their physical characteristics probably allow for pro-longed storage, maintaining an overview of inventory and storage duration of R&T becomes more challenging, thus making such digital logistics applications valuable in these domains. As stated earlier, employee mindfulness can be increased through educational videos and workshops, fostering awareness of their responsibilities. Workers should understand that their actions can determine the edibility of commodities or even contribute to potential toxicity. Emphasizing the importance of employee mindfulness may be a favorable approach in

SSA, where hiring and training personnel might be more cost-effective than purchasing sophisticated machinery or upgrading infrastructure.

5.9 Report findings for vegetables/processing

In the case of the processing of vegetables, warehouse management, packaging management, and consumer satisfaction are assigned with greater relevance compared to the overall benchmarks. On the other hand, regulation, shelf-life optimization, and network cooperation are assigned with much lower relevance compared to their overall benchmarks (see Figure 37).

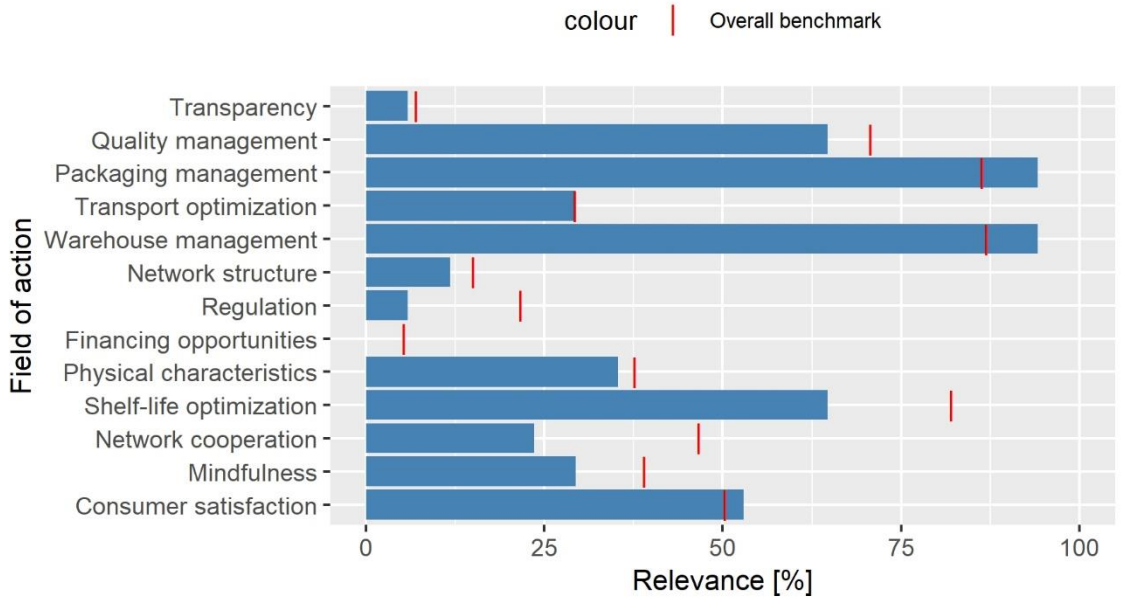


Figure 37: Result of the NLP analysis for the FL spot vegetables/processing, n = 17 (own figure, CC90)

The high rating of relevance of packaging management suggests that vegetables are often packaged at the site where they are processed, emphasizing the importance of processors considering downstream aspects of the SC. In contrast to the outcomes of processing of O&P and R&T, network cooperation scores much lower, which might indicate that companies that process vegetables typically do not collaborate or require collaboration with downstream processing sites, as they are already outputting final products. However, collaboration supposedly exists between the processing company and transportation companies. Additionally, consumer satisfaction is quite remarkable. Understanding the consumers well, including their tolerances, preferences for ripeness, and demand for quantities, is crucial to ensure that vegetables are processed appropriately and at the right time. However, as earlier touched upon, the high relevance of consumer satisfaction may have been computed because the processing made the food become spoiled in a way that

it can no longer be sold to the end consumer. Generally, the results of the NLP analysis for the FL spot vegetables/processing are overall relatively similar to those of the FL spot fruits/processing.

As an action recommendation, companies processing vegetables may contemplate adopting a more far-sighted approach, concentrating not only on accurate processing but also on enhancing the probability of the product reaching the end consumer unspoiled. In essence, these actors should be cognizant of the importance of proper packaging and comprehend how their packaging methods influence the potential spoilage of vegetables. Even though they may not witness the spoilage of vegetables directly, it may occur further downstream of the SC, and they should be aware of this danger. The relatively high relevance of warehouse management can be addressed by adopting more holistic organizational approaches within enterprises. Although the primary task of a processing company is to process vegetables, it is essential not to neglect other aspects of business operations. Ensuring all aspects are managed effectively allows for products to be sold at reasonable prices, subsequently generating revenues.

5.10 Prioritization of combating food losses across combinations of food categories and supply chain stages

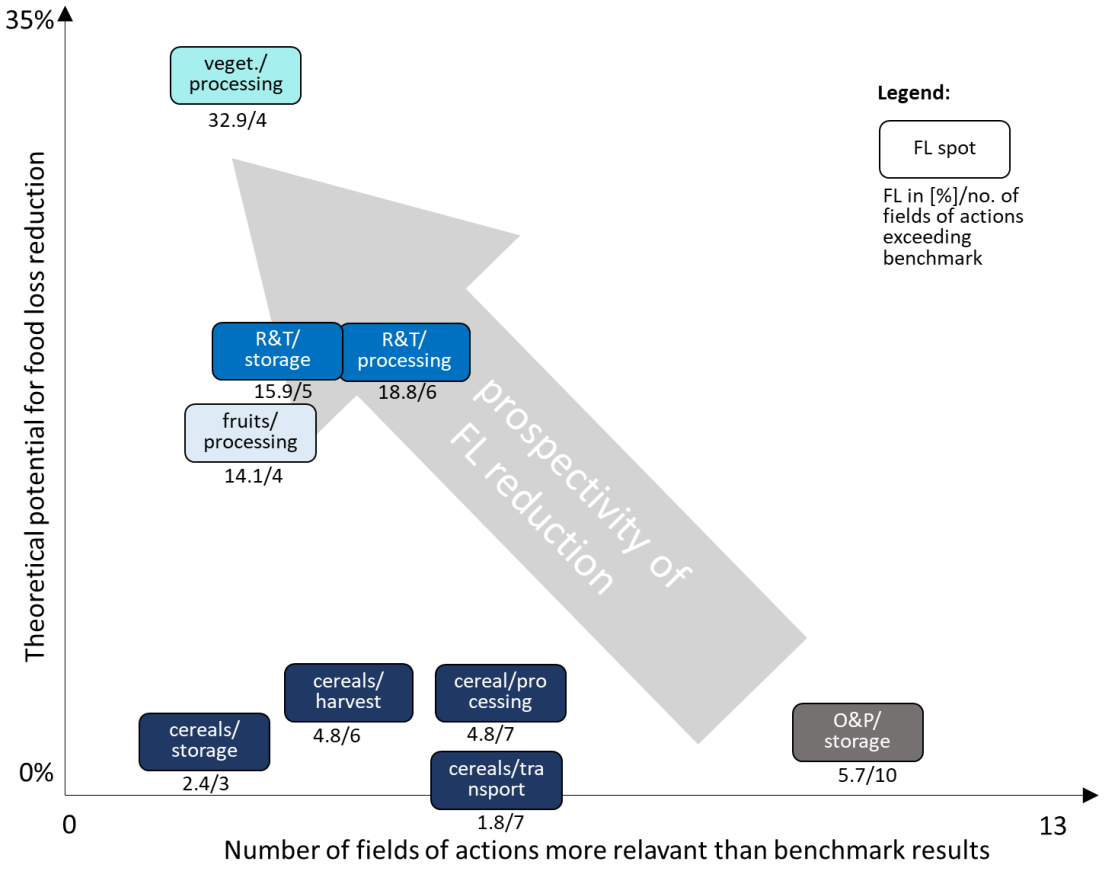


Figure 38: "Low hanging fruits" matrix of combating FL for FL spots (own figure)

In an effort to synthesize the overall picture of FL, their causes, and mitigation strategies for each FL spot, Figure 38 displays the FL spots examined in this chapter. The graph in Figure 38 does not provide a highly detailed representation of the metrics but aims to deliver a general overview. The x-axis signifies the number of fields of action scoring above the benchmark in the cause of loss analysis, serving as a rough proxy for the complexity of achieving substantial improvements at the respective spots. Intuitively, addressing FL issues at a spot with numerous relevant fields of action would likely demand greater effort and time for improvements. The y-axis represents the computed average losses at each FL spot, demonstrating the location of FL spots concerning the estimated effort required to achieve significant FL reductions and the achievable FL reduction. The arrow indicates the portion of the graph displaying FL spots where, according to the underlying assumptions, relatively immediate gains in FL reduction can potentially be attained in terms of reduced effort and time and significant FL reduction.

It is recommended to mitigate FL on vegetable/processing first, followed by R&T-storage and R&T/processing, and then fruit/processing. It is also worth noting that the problems arising during O&P/storage are complex to the extent that, compared to other spots, these areas might not offer immediate gains.

Cereal FL, depending on their SC stage, would be relatively easy to moderate in difficulty to mitigate.

As an additional statement and justification for the low-hanging fruit matrix being entirely based on the NLP analysis, it is crucial to explain the shortcomings of the regression analysis. According to the processed, selected, and computed data, there is no statistically robust relationship between the LPI indicators and the FL at different FL spots, with the exception from the FL spot of cereals at the SC stage of harvest.

6 Data product

While the current analysis offers a comprehensive understanding of the characteristics of FL in SSA and suggests ways to mitigate them, it can be assumed that new data will be frequently added to the FLI database in the future. As mentioned in Chapter 4.3.3, data influx has been fluctuating since 2008, but at a higher level compared to the period prior to 2008, with no definite indication of this trend ceasing. This increase in data availability served as a primary motivation for developing the code, primarily written in R, using a pipelined approach. This design allows for the seamless integration of new data without necessitating significant alterations to the codebase.

As Schutt and O'Neil (2013) don't explicitly name subtasks that need to be handled related to building a data product, in that context, Herden (2019) points out the importance of documentation of knowledge around the project and maintenance in the sense that there must be regular performance tests on the analysis solution and necessary adjustments to the analysis solution need to be taken. Also, future potential on technical improvement should be considered (Herden 2019, p. 234).

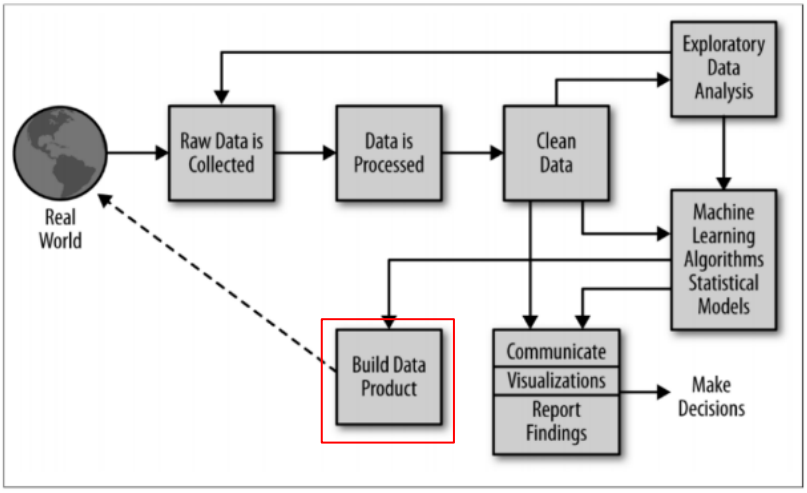


Figure 4: "The Data Science Process" according to Schutt and O'Neil (2013, p. 41)

6.1 Documentation of project knowledge

The code for this project, predominantly written in R, can be found in the Appendix H. The thesis and all associated code and Excel spread sheets are freely accessible via the following GitHub repository:

https://github.com/Lennard-Heuer/Data-based-logistics-for-loss-reduced-food-value-networks.git

To execute the code, it is necessary to install R, RStudio, Python, and the Anaconda package manager. The code provided is commented, which helps in understanding the primary concepts behind the programming. Further, it is important to note that the output of plots within the R Markdown in Appendix H is displayed on a provisional level. Their appearance is properly formatted only when they are exported as PNG files.

6.2 Maintenance of the data product

1. CPC codes and country names should be consistently updated to accommodate changes due to geopolitical events.
2. There may be a need to assign new commodities to food categories if new types of commodities are included in the dataset. Furthermore, reassignment of data points and additional reconstruction of SC stages might become necessary if new types of SC stages and SC activities are introduced to the dataset or if the format of the FAO questionnaire undergoes significant changes.
3. As of March 25, 2023, the API found in the Python library, used in January 2023, still functions. However, in order to avoid frequency capping, it is advisable to switch to the official API, released by OpenAI on March 1, 2023, for integrating ChatGPT into apps and products (Open AI 3/1/2023), which was not used within this thesis, as it was not released yet by the time of working on the project.
4. In the rare event when ChatGPT encounters multiple causes of losses within a single cell in the cause of loss column, manual editing in Excel was necessary in January 2023 under ChatGPT 3.5 and may still be necessary in the future. This step was carried out manually in Excel, with the relevant document available on the GitHub repository. In cases of enumerations, the cause of loss mentioned first was always selected for further analysis. It is possible that subsequent versions of ChatGPT 3.5 may be able to handle this task. This should be monitored to determine if future versions of ChatGPT can automate this process, thereby eliminating the need for manual intervention.
5. The aggregated LPI, that was used and merged with the FLI dataset, encompasses data representing logistics performance from 2012 to 2018. Before 2018, the data was updated every two years (World Bank 2023a). If the absence of LPI updates for 2020 and 2022 is primarily due to the COVID-19 pandemic and its diverse repercussions, it is essential to monitor the availability of new LPI data and incorporate it into the database for future computations.

6.3 Future potential on improvement

To effectively visualize the results, a web app could be developed using the Shiny package in R. Shiny enables the creation of interactive web applications directly from R (Posit 2023). The advantage of constructing an interactive web app lies in its accessibility for users with diverse backgrounds, allowing them to easily access the data analysis results through a browser.

Furthermore, employing a web app would enable users to apply specific limitations to the data, such as displaying only information from East Africa within the SSA region. Chapter 6.2 will provide recommendations for enhancing data collection and addressing additional general concerns, with a focus on content rather than just the technical implementation.

5. Discussion

This chapter explores the generalizability of the findings to the global context, as well as their relevance in offering guidance for individual SSA countries. Moreover, it sheds light on the ways, in which the data analysis enhances the current scientific knowledge base.

5.1 Generalizability of the findings to the global context and individual Sub-Saharan African countries

When considering the extrapolation of the results to a global scale, it is crucial to recall the statement made by HLPE (2014) in Chapter 1, which highlights that in developed countries, FLW primarily occurs at the consumer level and it tends to occur upstream along the SC in developing countries (HLPE 2014, p. 27). Taking into account the general conditions and economic hardships in SSA, as described by the (United Nations Conference on Trade and Development 2021) in Chapter 1, along with the overall inadequate state of infrastructure in SSA (Appendix G), SSA can be regarded as a distinct macro-region in the world. Consequently, the results of this thesis might not be easily generalizable to the entire world, as the framework conditions in other world regions are likely to be significantly different compared to SSA. However, the results may be comparable or applicable to other countries outside of SSA with roughly similar framework conditions. In summary, caution should be exercised when applying these findings in different regional contexts.

(World Bank 2022) has emphasized the diversity among SSA countries and the attempts to cluster SSA countries in this thesis proved to be unsuccessful due to insufficient amounts of data, particularly for countries with comparatively poor logistics performance. Chapter 4.3.5 revealed significant disparities in data availability across these countries, depending on their overall LPI score. The applicability of the results may therefore vary, depending on the specific SSA country. As the results were derived by mainly using data from "high performance countries" which are probably at the same time also rather the wealthier countries according to SSA standards, the findings of this data analysis may be more applicable to the relatively wealthier countries within the SSA region.

If more data becomes available in the future, a more nuanced data analysis on distinct regions in SSA may become possible. For example, the dataset could not only by divided into CPGs, but also into SSA's four major regions: East, Central, South, and West Africa (World Bank 2023b).

5.2 Implications for science and practical applications

In this section, the implications of the thesis' results for the scientific knowledge base will be discussed, focusing on areas where the findings confirm existing knowledge or present contradictions. Additionally, the impact of the thesis' results on perspectives beyond the scientific community will be examined.

5.2.1 Comparison with current knowledge base about FL in SSA

To ensure the validity and reliability of the results obtained from the hotspot analysis of FL across various food categories and SC stages in Chapter 4.3.6, a careful evaluation of the data is necessary. Comparing the findings of the hotspot analysis with the current scientific knowledge base in the field of FL research serves as a valuable sanity check for the results. Additionally, these comparisons contribute to the scientific knowledge base by either confirming or contradicting existing knowledge, ultimately leading to a more comprehensive understanding of the subject.

Gustavsson et al. (2011) reported that, in the case of SSA, FL of fruits and vegetables were relatively higher compared FL of other food categories such as cereals, roots and tubers, oilseeds and pulses, meat, and fish and seafood. In order, starting with the lowest FLW magnitudes as presented by Gustavsson et al. (2011), the categories are as follows: cereals, R&T, and O&P and Fruits and Vegetables (Gustavsson et al. 2011, pp. 6–7). This pattern was also observed in the data analysis of this thesis (see Figure 24).

One notable observation is the substantial difference in FL [%] between fruits and vegetables (Figure 24). It was expected that, when combined, their losses would surpass those of R&T, which they did, as R&T's overall FL [%] is only marginally higher than the overall FL of fruits and significantly lower than the overall FL of vegetables. However, the large discrepancy in the overall FL [%] of vegetables and fruits across all SC stages is puzzling. This observation may indicate a potential bias in the data because Gustavsson et al. (2011, pp. 6–7) used vegetables and fruits together as one group, implying that they believed the two food categories behaved somewhat similarly in their FL magnitudes.

Furthermore, the exceedingly high figures computed for vegetables lend some support to an assertion made by Affognon et al. (2015), who argued that FL estimates in SSA are often overestimated due to a tendency to focus on areas where high losses are anticipated (Affognon et al. 2015, p. 62).

It is also noteworthy that the analysis partially aligns with the conclusions drawn by Barrett (2016) in their literature review, which pinpointed the farm [SC stage] as the primary hotspot for post-harvest FL in SSA (Barrett 2016, p. 21). Upon examining Figure 23, the heatmap of FL across food categories and SC stages, it is evident that processing accounts for a substantial portion of overall FL for each food category. For fruits, O&P, and vegetables, processing is the SC stage with the highest percentage of FL. Significant FL are also observed at the harvest stage for all food categories except O&P. Although it is not possible to definitively distinguish between on-farm and beyond-farm data for the processing SC stage, the available information seems to challenge the assertion made by Sheahan and Barrett (2017) that most FL occurs beyond the farm. However, this contradiction is not certain, as the processing SC stage can be somewhat ambiguous, with activities occurring both on and off the farm.

Furthermore, Sheahan and Barrett (2017) state that as a results of most FL occurring on-farm, most attempts to combat post-harvest losses were directed on-farm, mainly around storage. In contrast to that, there are only a few evaluations of techniques against FL apart from storage, in particular beyond the farm SC stage (Sheahan and Barrett 2017, p. 10). Although a comprehensive analysis of the distribution of techniques against FL was not conducted in this thesis, the available data suggest that significant losses occur beyond the farm (Figure 23). Therefore, efforts to mitigate FL should also encompass SC stages beyond the farm. The only potential exception is cereals, where FL beyond the farm is considerably lower than on-farm losses, indicating that off-farm mitigation may be less critical in the case of cereals.

Spang et al. (2019) highlight the existence of fundamental gaps in FLW data for SSA, both in terms of food categories and SC stages (Spang et al. 2019, p. 146). This observation was confirmed at various points during the data analysis process, posing challenges when working with the data. Figure 19, in particular, revealed substantial data gaps across food categories and SC stages. As a result, Chapter 4.3.4 determined that it was necessary to exclude the dairy and N&C food categories from the dataset.

5.2.2 Current state of data retrieval and data visualization

The FAO's FLI database currently hosts an interactive dashboard, encompassing data-driven visualizations from the FLI, which display three distinct plots related to various SC stages and nations (FAO 2023a). Contrarily, the data analysis performed in this study incorporates food category differentiation, a feature absent from the dashboard.

Furthermore, the FAO's dashboard does not distinguish between modelled and non-modelled data, rendering it vulnerable to the "overshadowing effect". Additionally, it overlooks the multi-level nature of the data and the issue of double counting. As a result, these visualizations may be partially misleading and unsuitable for quantitative applications.

5.3 Further value added by the thesis

In this section, the contributions of this thesis to the scientific community will be evaluated. Considering the global and interdisciplinary nature of the research question, this analysis encompasses not only the academic domain but also the contributions made to organizations that provided data for the dataset, namely the FAO and the World Bank.

5.3.1 A novel methodology developed and applied

In addition to the recommendations on which fields of action to employ for combating FL at various FL spots, the methodology itself employed in this thesis represents a significant contribution to the current research on addressing FL in SSA, as it suggest a novel methodology to answering the research question how FL in SSA can be effectively mitigated by logistical measures. The proposed approach encompasses data reconstruction to ensure compatibility with the study's objectives, as well as sensibility. It also included a hotspot analysis, the application of multiple regression and NLP to investigate on causes of FL and then to match existing data with logistical countermeasures, and the finally prioritize FL mitigation efforts across food categories and SC stages. Additionally, a recommendation for clustering countries within SSA based on their overall LPI score has been put forth to feed to decision assistant with these clustered data, as to yield results that more closely represent individual nations. While the current data availability does not allow for regression analyses across all FL spots, the methodology presented remains applicable in the event of future data expansion.

5.3.2 Revealing flaws of the FLI database

Furthermore, the extensive exploratory data analysis performed in this study has revealed critical pitfalls and inherent phenomena related to the FLI dataset. These insights were essential for constructing and selecting suitable data for the hotspot analysis, the subsequent NLP and regression analyses, and their interpretation. In essence, this research has contributed to a more profound comprehension of the dataset's composition and attributes, with a particular focus on its flaws.

The flaws of the dataset include insufficient amounts of data across combinations of SC stages and food categories (see Figure 19), unequal expected data quality across different food categories (see Figure 20 & Figure 21), the problem of multi-leveled SC stages and the issue of double counting of FL values (Chapter 3.2.3), among other issues. However, the insights gained from this thesis can help inform future research, improve the understanding of FL data in SSA and provide a role model in analyzing the dataset and give recommendations on how to improve the data quality in the dataset when retrieving new data. Overall, it also shows the significance of a careful approach to working with this data and the need for data preparation prior to running analyses on them.

5.3.3 Improving data quality in the data collection by giving recommendations to the FAO

The predominance of data originating from the APHLIS database may compromise the quality of the overall dataset if no appropriate countermeasures are taken. In the case of particular food categories, such as cereals, computing averages across groups, food categories, and SC stages might yield results similar to those obtained by using data of the APHLIS database alone, causing a significant "overshadowing" of the remaining data (CC44). This is a significant issue that requires attention.

In addressing the first issue, the "overshadowing" of data that is expected be of high quality by data that is expected to be of lower quality, solutions include restricting the quantity of data points that can be submitted by a single entity, implementing rigorous evaluations of data quality, source and completeness regarding entries in all of its columns prior to integration into the FLI dataset. Another approach would be the establishment of a separate database for modelled data only. This supplementary database can offer approximate estimates in situations where reliable results from alternative data collection methods are unattainable. Consequently, this additional database would be consulted only in specific cases when non-modelled data is scarce, serving as a provisional solution.

Moreover, embracing a more structured data collection methodology, as opposed to allowing free-text input, could be beneficial. Respondents might be directed to choose from a limited, predefined set of causes of loss or SC stages. Implementing stricter data collection forms could streamline the current multi-layered and intricate SC stages by providing options at a single level and facilitating information processing in the cause of loss column to circumvent free text input. While the more stringent format may deter some respondents from contributing data, the resulting dataset would likely show greater

homogeneity, thus facilitating analysis.

Additionally, it is crucial to investigate on techniques for quantifying FL in SSA. According to numbers from the World Bank, in 2021 there existed 93 mobile cellular subscriptions per 100 people in SSA (World Bank 2023c). Although this number does not reflect the number of unique users having a cellular subscription, it suggests a high level of mobile penetration and connectivity via mobile phones.

Therefore, traditional questionnaires could be replaced with mobile-based data collection, allowing for more frequent and accessible data retrieval. Directly engaging farmers in the data collection process could further enhance data accuracy. Additionally, establishing an association with members spanning the African food industry, who submit FL data and receive financial compensation, might be a viable approach to improving data quality and availability.

As a general note, data preparation is a critical stage in data analysis, primarily due to the significant impact of input data quality on the quality of the resultant findings of the actual analysis (Sattler and Schallehn 2002, p. 1). Oliveira et al. (2005) referred to this phenomena as the "garbage in, garbage out" principle (Oliveira et al. 2005, p. 1). Data preparation, with subtasks such as selection, transformation, and cleaning of data, can be very cost-intensive. Often, 50-70% of the effort in data analysis is dedicated to data preparation, with the remainder focused on the actual analysis (Sattler and Schallehn 2002, p. 1). This thesis has significantly attributed to this aspect by reconstructing the data and carefully defining the scope of data inclusion in alignment with the primary research question's objectives prior to the actual data analysis.

The considerable effort devoted to data reconstruction in this thesis, along with the beforementioned statement by Sattler and Schallehn (2002, p. 1) raises concerns about the reliability of FAO's dashboard (FAO 2023a) within the FAO database. This dashboard was developed without documented data reconstruction, aside from the apparent data cleaning already performed by the FAO. It is plausible that the entire dataset was simply incorporated and converted into graphs for the dashboard's creation.

Consequently, it would be advisable for the FAO to undertake data preparation in terms of reconstruction, if not already done so, prior to releasing new updates to the database and the respective dashboard. This responsibility should lie with the FAO, given their extensive knowledge of the dataset, enabling them to sensibly prepare the data.

Concurrently, the FAO ought to disclose the originally collected data to enable verification of the appropriateness of the data preparation process.

6. Outlook

In this final chapter, a critical assessment of the thesis, its structure, and methodologies is provided. Furthermore, future research needs emerging from this thesis will be described.

6.1 Critical evaluation

The following enumeration presents potential points of critique against this thesis, along with justifications for the chosen methods and approaches. In accordance with the two project cycles used to structure this thesis, the critical evaluation for the NLP analysis and the multiple regression have already been formulated in Chapter 3.4.3, respectively in Chapter 3.4.6. Further points of critique are:

(1) As described in Chapter 2.7, data from the years 2000-2021 was included in the thesis. One could argue that data dating back to 2000 might be outdated and should thus be excluded. However, this choice represents a reasonable compromise between between data up-to-date-ness and availability.

(2) The SC stage farm was reconstructed with the aim of making it more suitable for analysis, thus enabling to obtain meaningful results. During the reconstruction, data points with a SC stage of 'farm' were reassigned to other SC stages based on the information in the activity column. This assignment, documented in CC13-CC15, introduces a certain degree of researcher bias, which was unavoidable.

(3) The clustering of countries according to their LPI (Logistics Performance Index) reveals that the group of countries with the worst overall LPI scores contributed the least data to the dataset. One might argue that the 13 "Low performance countries" are small SSA countries, and therefore, they do not contribute much data to the FLI. However, this argument is refuted, as the 13 "Low performance countries" also include populous nations such as the Democratic Republic of Congo, Angola, and Zimbabwe.

(4) The prioritization of mitigation measures against FL for FL spots using the "Low-hanging-fruit" matrix can be criticized in several ways: First, the FL for FL spots were not weighted by their quantities. Second, the measure of difficulty in mitigating FL, by counting the number of fields of action whose relevance was computed as higher compared to their benchmarks, is imprecise. Achieving improvements in one field of action may not be comparable to achieving improvement in another field of action in terms of cost and effort. Thirdly, just because there exists a certain percentage of FL for

a specific FL spot, that does not imply that the FL can be easily reduced to zero FL. The reduction in FL that can be achieved depends on external circumstances.

These three points present valid critiques of the applied methodology. However, weighting by quantities was not feasible (see Chapter 3.4.3.), and in the other two cases, the best available proxy was used.

6.2 Derived need for more future research

In this section it is outlined, what data-based research on FL in SSA should be focusing on to create more value to the quest in reducing FL along SCs in SSA.
This chapter focuses on the methodology used to create better decision assistance for combating FL in SSA, rather than on the specific countermeasure for combating FL in the region. This approach is consistent with the high-level perspective taken in this thesis, as providing detailed advice to individual stakeholders along the supply SC would require additional extensive background information on each specific case. Therefore, the emphasis is on developing a methodology that can be applied more broadly to support decision-making processes related to FL in SSA. This chapter highlights the importance of improving data collection processes, given that Chapter 4.3.2 and Chapter 4.3.4 have demonstrated that insufficient data availability, and probably data quality, are significant issues.

6.2.1 Need for more abundant data in consideration of future development of food losses in Sub-Saharan Africa

In future times, data-driven research on FL in SSA should take changing food patterns into account. According to Figure 39, which displays the data availability excluding modelled data, an over proportionally big part of data points reports on data of the food category of cereals. This holds true even when subtracting the modelled data from the overall dataset (CC44).

Spang et al. (2019) describe that food patterns in SSA are subject to dynamic changes. Furthermore they assert that long-term changes in food production, processing, SC logistics, and retail operations, as well as changes in consumer diets and habits, will resolve some issues of FLW while simultaneously introducing new challenges (Spang et al. 2019, p. 146).

The fact that currently, the food category of cereals makes of the largest part of data

among all food categories, regardless of whether modelled data is included or not (CC44) may lead to the assumption that cereals, or to put int differently starch-based nutrition, plays a major role for most people in SSA.

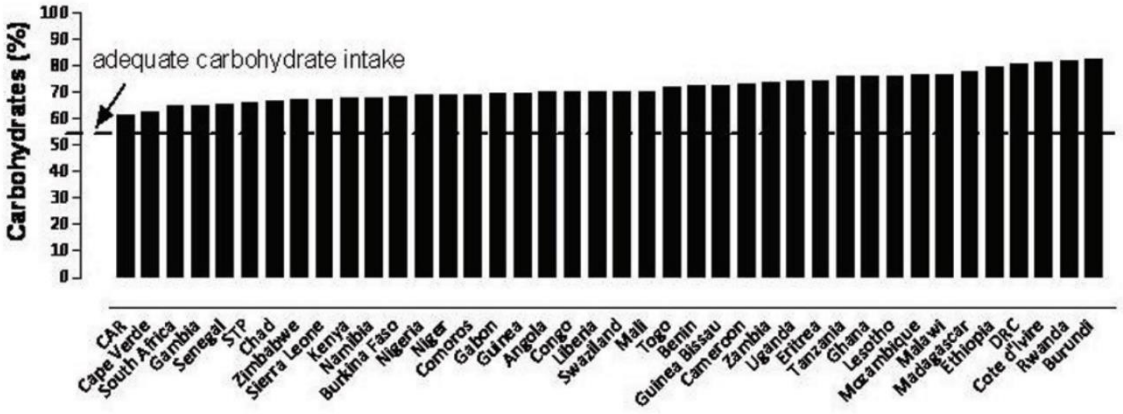


Figure 39: Intake of carbohydrates in SSA countries (Abrahams et al. 2011, 7)

Abrahams et al. (2011, p. 7), in Figure 39, support this assumption by illustrating the proportion of people's overall caloric intake constituted by carbohydrates compared to their overall intake of calories in a wide range of SSA countries. Notably, all considered countries exhibit values significantly above the recommended percentage of approximately 55%. The observation that South Africa, a relatively developed country in SSA, appears closer to the suggested 55% line on the left-hand side of the chart may indicate that people living in wealthier countries in SSA tend to have a less cereal (carbohydrate) - dominated diet.

Spang et al. (2019, p. 146) highlight a trend among SSA countries transitioning from a starch (carbohydrate) - based diet to a diet more similar to that of people living in developed countries.

If the current nutrition pattern in SSA is indeed heavily starch-dominated, future trends may lean towards more diversified diets. For example, increased demand for meat could lead to more complex SCs compared to those for cereals, as meat handling requires proper cooling and a stricter contamination prevention. Such dynamic shifts in food patterns and SCs necessitate timely data updates to the FLI.

Relatedly, this thesis utilized data from a 22-year time window. However, given the potential for rapid changes in dietary patterns in SSA, it may already be essential to consider only data no older than, for instance, five years. To evaluate these numbers in a statistically robust manner, more comprehensive data would be required, including a

more balanced distribution of information across all relevant food categories and SC stages.

6.2.2 Food quality reductions

This thesis adopted a relatively binary perspective on the topic of FL, classifying food as either lost or not lost. Consequently, the thesis did not provide a more nuanced understanding of FL and quality reductions.

Regarding FL and food quality reductions, Sheahan and Barrett (2012) concluded that in contrast to food quantity losses, food quality losses are most probably more widespread than commonly assumed. They added, that there have been significant efforts to measure the extent of food quantity losses and explore new ways to address the issue of quantity losses; however, until today, there has been no comparable effort to investigate on the distribution of food quality reductions in SSA (Sheahan and Barrett 2016, p. 21).

Future data science research on this topic could yield a more nuanced comprehension of FL and quality reductions. The overall feasibility of considering quality when analyzing the FLI data is apparent, as it is mentioned 8 times in total in the cause of loss column (CC23).

According to the World Food Program's conception, food poisoning is another major issue in SSA (Costa 2014, p. 5). This observation further underscores the importance of considering food quality losses, as the boundaries between critically poor food quality and acceptable food quality may be ambiguous and not clearly delineated. It may even be possible that some effects of consuming food with significantly reduced quality only manifest in later stages of one's life.

Additionally, in the food-use-not-waste-hierarchy in Figure 1, Kleineidam (2020, p. 4, based on Papargyropoulou et al. 2014) demonstrated that there are various levels to addressing the issue of FLW with the goal of preventing disposal, as it incorporates options such as re-use and recycle. This indicates that not only is the phenomenon of FL nuanced in term of quantity and quality losses, but so are the countermeasures that can be taken to mitigate it.

An ideal decision assistant should, therefore, on the one hand consider information regarding quality degradation and also provide recommendations that encompass not just prevention strategies but also mitigation strategies beyond prevention.

According to Sheahan and Barrett (2017), reduction of post-harvest FL in SSA come at a high price. Meanwhile the entire benefit of such interventions remains uncertain. Efforts that have been made for the reduction off postharvest losses turn out to be not so cost effective in achieving the objectives. Therefore the mere reduction may not be the perfect way to tackle the issue. (Sheahan and Barrett 2017, p. 10). Consequently, this assertion indicates that, under specific circumstances, it may be more pragmatic to recognize when food is no longer suitable for human consumption and repurpose it according to the different levels of the food-use-not-waste-hierarchy by Kleineidam (2020, p. 4, based on Papargyropoulou et al. 2014). Additionally, it might be reasonable to accept a certain degree of unavoidable losses if the efforts to mitigate and to repurpose them are excessively burdensome. Under certain circumstances, allocating financial resources in SSA to enhance agricultural productivity could potentially yield more cost-effective and ecological results, rather than focusing on an unrealistic and excessively value-driven pursuit of reducing FL in an impractical manner.

In summary, addressing the challenge of ensuring food security in SSA necessitates a comprehensive approach. FL prevention represents merely one component of this broader strategy, which may also encompass repurposing food no longer suitable for human consumption and improving agricultural productivity.

6.3 Demand for further research

In this section, the focus is on identifying key areas for data-based research on FL in SSA to generate greater value in generating knowledge for reducing FL along SCs in the region. This chapter does not emphasize specific methods for combating FL in SSA, but rather the methodology for creating better decision data-driven assistance, consistent with the high-level approach adopted throughout this entire thesis, which does not primarily target giving specific advice to individual stakeholders along SCs in SSA.

6.3.1 Integration of results into scoring model

Initially, this thesis aimed to perform a combined evaluation of all computed results using a scoring process incorporating various criteria. A recommendation for fields of action according to Kleineidam (2020, p. 10), might have been based on input values such as prioritization scores derived from the hotspot analysis, which take into account the quantities of FL, along with resulting scores from the regression analysis and scores gained through the NLP analysis. These three scores could then be combined or calculated

together. However, this approach was ultimately abandoned due to the inconsistent data richness across food categories and SC stages, which precluded the possibility of performing multiple regression on numerous FL spots. Additionally, the anticipated imprecision of the results from the hotspot analysis, cause of loss analysis, and the LPI indicator regression analysis rendered the initial plan unfeasible. As it can create a false impression of accuracy if imprecise data is simply assigned a score and combined, making it appear as if it is mathematically correct.

Consequently, the adopted approach for the hotspot, the regression, and NLP analyses, and their combination, was to utilize the available data as effectively as possible while being transparent on the capabilities and limitations of the methods. In the future, with increased data availability and improved clarity regarding SC entries and redundancies, employing a multiplication of scores or a quantitative weighting of importance based on the hotspot analysis may become a more viable approach.

6.3.2 Examples among countries

Ghana serves as a notable example among other countries in terms of comprehensive data contribution (see Figure 22). It is essential to investigate the reasons behind the country's ability to contribute a significant number of data points to the dataset. This is particularly important, as decision-makers require adequate data to make informed decisions. Even the most robust methodology cannot fully compensate for the limitations of a small-scale dataset, which may be susceptible to random influences.

Furthermore, beyond deriving insights into the causes of losses and potential mitigation strategies from the data, FL data can also prove valuable for stakeholders in tracking performance over time. By monitoring FL throughout a specific period, stakeholders can evaluate the success of implemented strategies or identify setbacks that may occur.

6.3.3 Clustering of Sub-Saharan African countries

Additionally, as data abundance may increase in the future, it would be beneficial to explore ways to group countries by development stages or other factors. Research should be conducted to determine the most effective approach to cluster countries, either by FL metrics, geographical characteristics, or other relevant criteria. If clustered by FL metrics, the clusters could be interpreted, possibly providing insights into factors that might influence FL in specific ways. On the other hand, geographical clustering could reveal particular strengths of a major region in SSA, allowing other regions to learn from and

follow the example. Furthermore, clustering based on other characteristics might uncover hidden patterns and relationships that can inform targeted strategies to address FL in SSA. Thus, improved data availability and clustering methods can contribute to a better understanding of regional differences and potential strategies to address FL in SSA.

6.3.4 Recommender system on the level of concrete countermeasures.

Additionally, future research could investigate on the potential for developing a recommender system that suggests a set of specific countermeasures against FL, ideally providing an explanation for the recommendations and potentially offering successful case studies where similar measures have been employed and resulted in substantial FL reductions. This approach would delve deeper than the current level of fields of action, which are summaries of clusters of concrete countermeasures.

To build such a recommender system, utilizing scientific databases and web-crawling information could be valuable. For instance, once the reasons for FL are identified, this information could be used as a search query on platforms like Web of Science. Another option might involve compiling a comprehensive database of all possible countermeasures and case studies. This approach would enable a recommendation system to suggest suitable and concrete countermeasures for a specific case without web-crawling, effectively filtering relevant information to address the FL challenges faced by stakeholders, based on their framework conditions.

Publication bibliography

Abay, Kibrom A.; Yonzan, Nishant; Kurdi, Sikandra; Tafere, Kibrom (2022): Africa might have dodged a bullet, but systemic warnings abound for poverty reduction efforts on the continent. The World Bank. Available online at https://blogs.worldbank.org/developmenttalk/africa-might-have-dodged-bullet-systemic-warnings-abound-poverty-reduction-efforts, checked on 2/23/2023.

Abrahams, Zulfa; Mchiza, Zandile; Steyn, Nelia P. (2011): Diet and mortality rates in Sub-Saharan Africa: Stages in the nutrition transition. In *BMC Public Health* 11 (1), p. 801. DOI: 10.1186/1471-2458-11-801.

Affognon, Hippolyte; Mutungi, Christopher; Sanginga, Pascal; Borgemeister, Christian (2015): Unpacking Postharvest Losses in Sub-Saharan Africa: A Meta-Analysis. In *World Development* 66, pp. 49–68. DOI: 10.1016/j.worlddev.2014.08.002.

Austin, Peter C.; Steyerberg, Ewout W. (2015): The number of subjects per variable required in linear regression analyses. In *Journal of clinical epidemiology* 68 (6), pp. 627–636. DOI: 10.1016/j.jclinepi.2014.12.014.

Aydın, Ömer; Karaarslan, Enis (2022): OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare. In *Emerging Computer Technologies* 2, pp. 22–31. Available online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4308687.

Blakeney, Michael (2019): Food Loss and Food Waste: Causes and Solutions. With assistance of Letizia Gianformaggio. Cheltenham, UK: Edward Elgar Publishing.

Chopra, Abhimanyu; Prashar, Abhinav; Sain, Chandresh (2013): Natural Language Processing. In *International Journal of Technology Enhancements and Emerging Enigneering Research* 1, Article 12. Available online at https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=eeace1d14e266a5cd44fe781a874c662928602fd, checked on 2/26/2023.

Costa, Simon J. (2014): Reducing Food Losses in Sub-Sahran Africa. improving Post-Harvest Management and Storage Technologies of Smallholder Farmers. UN World Food Programme. Kampala, Uganda. Available online at https://documents.wfp.org/stellent/groups/public/documents/special_initiatives/WFP265205.pdf, checked on 2/26/2023.

Debrah, Justice Kofi; Teye, Godfred Kwesi; Dinis, Maria Alzira Pimenta (2022): Barriers

and Challenges to Waste Management Hindering the Circular Economy in Sub-Saharan Africa. In *Urban Science* 6 (3), p. 57. DOI: 10.3390/urbansci6030057.

Edgar, Thomas W.; Manz, David O. (Eds.) (2017): Research Methods for Cyber Security: Syngress. Available online at https://www.sciencedirect.com/book/9780128053492/research-methods-for-cyber-security#book-info, checked on 4/12/2023.

FAO (2023a): Food Loss and Waste Database. Edited by FAO. Available online at https://www.fao.org/platform-food-loss-waste/flw-data/en/, checked on 4/3/2022.

FAO (2023b): Food Wastage Footprint. Available online at https://www.fao.org/nr/sustainability/food-loss-and-waste/en/#:~:text=Food%20loss%20refers%20to%20the,levels%2C%20mostly%20in%20developed%20countries., checked on 1/26/2023.

Feng, Steven; Gangal, Varun; Wei, Jason; Chandar, Sarath; Vosoughi, Soroush; Mitamura, Teruko; Hovy, Eduard (2021): A Survey of Data Augmentation Approaches for NLP. Available online at https://arxiv.org/pdf/2105.03075.pdf, checked on 2/26/2023.

Glauben et al., Thomas (2022): Der Ukrainekrieg offenbart angespannte Versorgungslagen auf Weltagrarmärkten: Gefordert sind wettbewerblich agierende Der Ukrainekrieg offenbart angespannte Versorgungslagen auf Weltagrarmärkten. Gefordert sind wettbewerblich agierende Gefordert sind globale Handelsstrukturen zur Krisenbewältigung¹. Edited by Leibniz Institute of Agricultural Development in Transition Economies (IAMO), Halle (Saale). Available online at https://www.econstor.eu/bitstream/10419/253633/1/1800025394.pdf, checked on 3/30/2023.

Google Scholar (Ed.) (2023): Google Scholar search for "Doing Data Science". Available online at https://scholar.google.de/scholar?hl=de&as_sdt=0%2C5&q=the+data+science+process+o+neil&btnG=, checked on 4/2/2023.

Gustavsson, Jenny; Cederberg, Christel; Sonesson, Ulf; van Otterdijk, Robert, Meybeck, Alexandre (2011): Global food losses and food waste. Extent, causes and prevention. Edited by FAO. Swedish Institute for Food and Biotechnology (SIK), FAO. Düsseldorf, Germany. Available online at https://www.fao.org/3/i2697e/i2697e.pdf.

Hadi, Suryadi; Rombe, Elimawaty; Vesakha, Gatha; Mustamin, Mustamin (2020): Food Loss and Food Waste: A Literature Review in 2009-2018. In *International Journal of Psychosocial Rehabilitation* 24, pp. 910–925. DOI: 10.37200/IJPR/V24I3/PR200842.

Harrell, Frank E. (2015): Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis. 2nd ed. New York, NY: Springer (Springer eBook Collection Mathematics and Statistics). Available online at https://warin.ca/ressources/books/2015_Book_RegressionModelingStrategies.pdf, checked on 3/30/2023.

Herden, Tino T. (2019): Managing Supply Chain Analytics. Guiding organizations to execute Analytics initiatives in Logistics and Supply Chain Management. Dissertation. Technical University of Berlin, Berlin. Available online at https://api-depositonce.tu-berlin.de/server/api/core/bitstreams/d7f0fc46-e5a9-4265-937b-ab307e00077b/content, checked on 12/16/2022.

HLPE (2014): Food losses and waste in the context of sustainable food systems. A report by the High Level Panel of Experts on Food Security and Nutrition of the Committee on World Food Security. FAO. Rome. Available online at https://www.fao.org/3/i3901e/i3901e.pdf#page=60&zoom=100,81,646, checked on 3/30/2023.

Hodges, Rick; Bernard, Marc; Rembold, Felix (2014): APHLIS - Postharvest cereal losses in Sub-Saharan Africa, their estimation, assessment and reduction. Available online at file:///C:/Users/User/Downloads/lbna26897enn.pdf, checked on 3/1/2023.

Institute of Electrical and Electronics Engineers (2011): 2011 5th International Conference on Application of Information and Communication Technologies (AICT 2011). Baku, Azerbaijan, 12 - 14 October 2011. 2011 5th International Conference on Application of Information and Communication Technologies (AICT). Baku, Azerbaijan, 10/12/2011 - 10/14/2011. Piscataway, NJ: IEEE.

Kleineidam, Julia (2020): Fields of Action for Designing Measures to Avoid Food Losses in Logistics Networks. In *Sustainability* 12. DOI: 10.3390/su12156093.

Kotchoni, Rachidi (2018): Detecting and Measuring Nonlinearity. In *Econometrics* 6, p. 37. DOI: 10.3390/econometrics6030037.

La Salle University (2022): Qualitative and Quantitative Research: What is "Empirical

Research"? Available online at https://library.lasalle.edu/c.php?g=225780&p=3112085#:~:text=Empirical%20research%20is%20based%20on,than%20from%20theory%20or%20belief, updated on 8/1/2022, checked on 4/12/2023.

Larson, Paul; Halldorsson, Arni (2004): Logistics Versus Supply Chain Management: An International Survey. In *International Journal of Logistics-research and Applications - INT J LOGIST-RES APPL* 7, pp. 17–31. DOI: 10.1080/13675560310001619240.

Lee, Jeong-Dong; Lee, Keun; Meissner, Dirk; Radosevic, Slavo; Vonortas, Nicholas (2021): Challenge of Technology and Economic Catch-up in Emerging Economies, JD Lee et al. eds. Ch 1. "Technology Upgrading and Economic Catch-up".

Mentzer, John; Dewitt, William; Keebler, James; Min, Soonhong; Nix, Nancy; Smith, Carlo; Zacharia, Zach (2001): Defining Supply Chain Management. In *Journal of Business Logistics* 22. DOI: 10.1002/j.2158-1592.2001.tb00001.x.

Miller, Jeff; Ulrich, Rolf (2019): The quest for an optimal alpha. In *PloS one* 14 (1), e0208631. DOI: 10.1371/journal.pone.0208631.

Moore, Andrew W.; Anderson, Brigham; Das, Kaustav; Wong, Weng-Keen (2006): CHAPTER 15 - Combining Multiple Signals for Biosurveillance. In Michael M. Wagner, Andrew W. Moore, Ron M. Aryel (Eds.): Handbook of Biosurveillance. Burlington: Academic Press, pp. 235–242. Available online at https://www.sciencedirect.com/science/article/pii/B978012369378550017X, checked on 3/30/2023.

Mpandeli, Sylvester et. al. (2018): Climate Change Adaptation through the Water-Energy-Food Nexus in Southern Africa. In *International Journal of Environmental Research and Public Health* 15 (2306). Available online at https://www.mdpi.com/1660-4601/15/10/2306, checked on 3/30/2022.

Oliveira, Paulo; Rodrigues, Fátima; Rangel Henriques, Pedro (2005): A Formal Definition of Data Quality Problems. In *Proceedings of the 2005 International Conference on Information Quality, ICIQ 2005*. Available online at https://www.researchgate.net/publication/220918803_A_Formal_Definition_of_Data_Quality_Problems, checked on 3/30/2023.

Open AI (3/1/2023): Introducing ChatGPT and Whisper APIs. Available online at

https://openai.com/blog/introducing-chatgpt-and-whisper-apis, checked on 1/27/2023.

Östergren, Karin; Gustavsson, Jenny; Bos-Brouwers, Hilke; Timmermans, Toine; Hansen, Ole-Jørgen; Møller, Hanne et al. (2014): FUSIONS definitional framework for food waste. Full report. Available online at https://www.eu-fusions.org/phocadownload/Publications/FUSIONS%20Definitional%20Framework%20for%20Food%20Waste%202014.pdf, checked on 12/5/2022.

Posit (Ed.) (2023): Welcome to Shiny. Available online at https://shiny.rstudio.com/tutorial/written-tutorial/lesson1/, checked on 3/22/2023.

Reynolds, Christian (Ed.) (2020): Routledge Handbook of Food Waste. With assistance of Scott Lougheed, Charlotte Spring. London: Routledge.

Sattler, Kai-Uwe; Schallehn, Eike (2002): A Data Preparation Framework based on a Multidatabase Language. Available online at https://www.researchgate.net/publication/2521657_A_Data_Preparation_Framework_based_on_a_Multidatabase_Language, checked on 3/30/2023.

Schutt, Rachel; O'Neil, Cathy (2013): Doing Data Science. Straight Talk From the Frontline. 1st ed. Sebastopol: O'Reilly.

Sheahan, Megan; Barrett, Christopher (2017): Food loss and waste in Sub-Saharan Africa: A critical review. In *Food Policy* 70, pp. 1–12. DOI: 10.1016/j.foodpol.2017.03.012.

Sheahan, Megan; Barrett, Christopher B. (2016): Food loss and waste in Sub-Saharan Africa: A critical review. Cornell University, 210B Warren Hall, Ithaca, USA. Dyson School of Applied Economics and Management. Available online at http://barrett.dyson.cornell.edu/files/papers/Sheahan%20Barrett_post-harvest%20losses_FULL%20TEXT.pdf, checked on 2/26/2023.

Spang, Edward S.; Moreno, Laura C.; Pace, Sara A.; Achmon, Yigal; Donis-Gonzalez, Irwin; Gosliner, Wendi A. et al. (2019): Food Loss and Waste: Measurement, Drivers, and Solutions. In *Annual Review of Environment and Resources* 44 (1), pp. 117–156. DOI: 10.1146/annurev-environ-101718-033228.

Stathers, Tanya; Richard, Lamboll; Brighton, M. Mvumi (2013): Post-harvest agriculture in a changing climate. Journal-Artikel. University of Greenwich, United Kingdom; University of Zimbabwe. Natural Resources Institute (NRI). Available online at

https://www.rural21.com/fileadmin/downloads/2013/en-01/rural2013_01-S12-14.pdf, checked on 3/30/2022.

Stenroos, Monty; Dzubak, Jocob (2018): Multiple Linear Regression R Guide. Edited by RStudio. Available online at https://rstudio-pubs-static.s3.amazonaws.com/385153_a5244c2b65a844b8803494f10c88e495.html, checked on 3/21/2023.

Straka, Martin (2019): Distribution and Supply Logistics: Cambridge Scholars Publishing. Available online at https://www.cambridgescholars.com/resources/pdfs/978-1-5275-3607-4-sample.pdf, checked on 2/26/2023.

The Pennsylvania State University (2018): Applied Regression Analysis. Other Regression Pitfalls. Available online at https://online.stat.psu.edu/stat462/node/185/, checked on 3/22/2023.

United Nations (2022): World Population World Population Prospects 2022. Summary of Results. Edited by United Nations. New York. Available online at https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf, checked on 3/30/2023.

United Nations Conference on Trade and Development (2021): Economic Development in Africa Report 2021. Reaping the potential benefits of the African Continental Free Trade Area for inclusive growth. Available online at https://unctad.org/press-material/facts-and-figures-7, checked on 1/28/2023.

United Nations Statistics Division (2023): Classification on Economic Statistics. CPC - Version 2.1. United Nations Statistics Division. Available online at https://unstats.un.org/unsd/classifications/Econ/search, updated on 2023, checked on 1/28/2023.

University of Leipzig (2020/2021): Deduktion und Induktion. Available online at https://home.uni-leipzig.de/methodenportal/deduktion_induktion/#:~:text=W%C3%A4hrend%20es%20bei%20deduktiven%20Verfahren,Befunden%20eine%20Theorie%20zu%20erstellen., updated on 2020/2021, checked on 3/1/2023.

Vanhamäki, Susanna; Medkova, Katerina; Malamakis, Apostolos; Kontogianni, Stamatia; Marišová, Eleonóra; Huisman Dellago, David; Moussiopoulos, Nicolas (2019):

Bio-based circular economy in European national and regional strategies. In *International Journal of Sustainable Development and Planning* 14, pp. 31–43. DOI: 10.2495/SDP-V14-N1-31-43.

Weglarz, Geoffrey (2004): Two Worlds of Data – Unstructured and Structured. In *DM Review*. Available online at https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6eb99aa78e572146aca31225ec9353c12d797070, checked on 4/5/2023.

World Bank (2011): Missing Food : The Case of Postharvest Grain Losses in Sub-Saharan Africa. Washington, DC. Available online at https://openknowledge.worldbank.org/bitstream/handle/10986/2824/603710SR0White0W110Missing0Food0web.pdf?sequence=1&isAllowed=y, checked on 3/30/2023.

World Bank (2022): The World Bank in Africa. Available online at https://www.worldbank.org/en/region/afr/overview, updated on 10/14/2022, checked on 3/21/2023.

World Bank (2023a): Aggregated LPI 2012-2018. Available online at https://lpi.worldbank.org/international/aggregated-ranking, checked on 11/5/2022.

World Bank (2023b): FOCUS: Sub-Saharan Africa. Available online at https://openknowledge.worldbank.org/pages/focus-sub-saharan-africa, checked on 4/17/2023.

World Bank (Ed.) (2023c): Mobile cellular subscriptions (per 100 people) - Sub-Saharan Africa. Available online at https://data.worldbank.org/indicator/IT.CEL.SETS.P2?locations=ZG, checked on 4/16/2023.

World Bank, China Development Bank (2017): Leapfrog: The Key to Africa's Development? From Constraints to Investments Opoortunities. Dakar. Available online at https://documents1.worldbank.org/curated/en/121581505973379739/pdf/Leapfrogging-the-key-to-Africas-development-from-constraints-to-investment-opportunities.pdf#page=147&zoom=100,0,0, checked on 4/12/2023.

Xue, Li; Liu, Gang (2019): Introduction to global food losses and food waste. In Charis M. Galanakis (Ed.): Saving Food: Academic Press, pp. 1–31. Available online at

https://www.sciencedirect.com/science/article/pii/B9780128153574000018, checked on 1/28/2023.

Yong, Gunwoo; Jeon, Kahyun; Gil, Daeyoung; Lee, Ghang (2022): Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. In *Computer-Aided Civil and Infrastructure Engineering*, pp. 1–19. DOI: 10.1111/mice.12954.

Appendix:

Appendix A - Problems of integrating country names and data availability of countries

Table 1: Integrating data availability of different countries (own figure, CC04-CC07)

Name of Country	**FLI data is missing**	**LPI data is missing**	**Both FLI and LPI data is missing**	**Conflict of names**
Botswana		x		
Cabo Verde			x	
Central African Republic				x
Cote d'Ivoire:				x
Congo (NOT Dem. Rep. of Congo):	x			
Democratic Republic of the Congo				x
Estwatini			x	
Gambia		x		x
Sao Tome and Principe				x
Seychelles		x		
South Sudan			x	
Tanzania				x

Appendix B – Fields of Action according to Kleineidam (2020, p. 10)

Table 2: Field of actions against FL (Kleineidam 2020, p. 10)

Field of Action	Description
Transparency	Increase of transparency within a company as well as between companies of a network
Quality management	Improvement of quality management for early detection of weaknesses
Packaging management	Improvement of packaging management during transport and storage processes as well as for distribution to the end customer
Transport optimization	Improvement of transport management with regard to route planning, loading of vehicles, and coordination of vehicles
Warehouse management	Improvement of warehouse management using suitable storage equipment, storage strategies, and adapted layout planning
Network structure	Improvement of the network structure using strategic network planning and location management
Regulation	Adapted regulations by the administration to support companies in reducing FL as required
Financing opportunities	Providing appropriate financial support from the administration to weaker network partners
Physical characteristics	Adaptation of processes to consider special physical requirements of the products, including temperature, pressure sensitivity, and air composition
Shelf-life optimization	Process adaptations that allow the shelf life of the products to be taken into account in decision making
Network cooperation	Improving cooperation within networks, including information sharing and efforts to develop comprehensive measures against FL
Mindfulness	Promoting awareness among employees at all levels in companies of the relevance of the problem of FL in everyday life
Consumer satisfaction	Adaptation of internal processes with the aim of meeting specific customer requirements

Appendix C – Assignment of data points

Table 3: Assignment of commodities to food groups based on (United Nations Statistics Division 2023)

Categorization in Data Analysis by Xue and Liu (2019)	**Corresponding cpc-code(s)** - and names of cpc-categories, if different from first column	**Occurrences in Dataset**
Cereal and cereal products	011	1830
Roots and tubers	015 ("Edible roots and tubers with high starch or inulin content")	135
Oilseeds and pulses	014 ("Oilseeds and oleaginous fruits"); 017 ("Pulses (dried leguminous vegetables)");	128
Fruits	013 ("Fruits and nuts"), excluding 0137 ("Nuts"), 21421 ("Groundnuts, shelled")	90
Vegetables	012	97
Meat	211	3
Fish and seafood	04 ("Fish and other fishing products")	-
Dairy products	22 ("Dairy products and egg products"), excluding 223 ("Eggs, in shell, preserved or cooked")	3
Eggs	223 ("Eggs, in shell, preserved or cooked")	6
Others or not specified	-	7
Nuts & Cacao beans	01640 ("Cocoa beans")	5

Appendix D – Assignment activities on the farm stage of the supply chain to alternative supply chain stages

Table 4: Assignment of entries at the "farm stage" to the other SC stages based on the entries in the activist column

Activity of processing SC stage	**SC Stage assigned to**
Harvest	Harvest
Harvisting, Storage	Harvest
Harvesting	Harvest
Transportation	Transport
Storage	Storage
Transportation	Storage
Farm, Handling, Storage	Storage
Stacking	Storage
Handling, storage	Storage
Handling	Storage
Lifting	Storage
Shelling, Threshing	Processing
Shelling	Processing
Winnowing	Processing
Drying	Processing
Sorting	Processing
Grading, Sorting	Processing
Threshing	Processing
Assembling, Farm	Processing
Bagging	Processing
Shelling	Processing
Milling	Processing
Drying, Farm	Processing

Appendix E – Heatmap of data availibility across countreis and SC stages

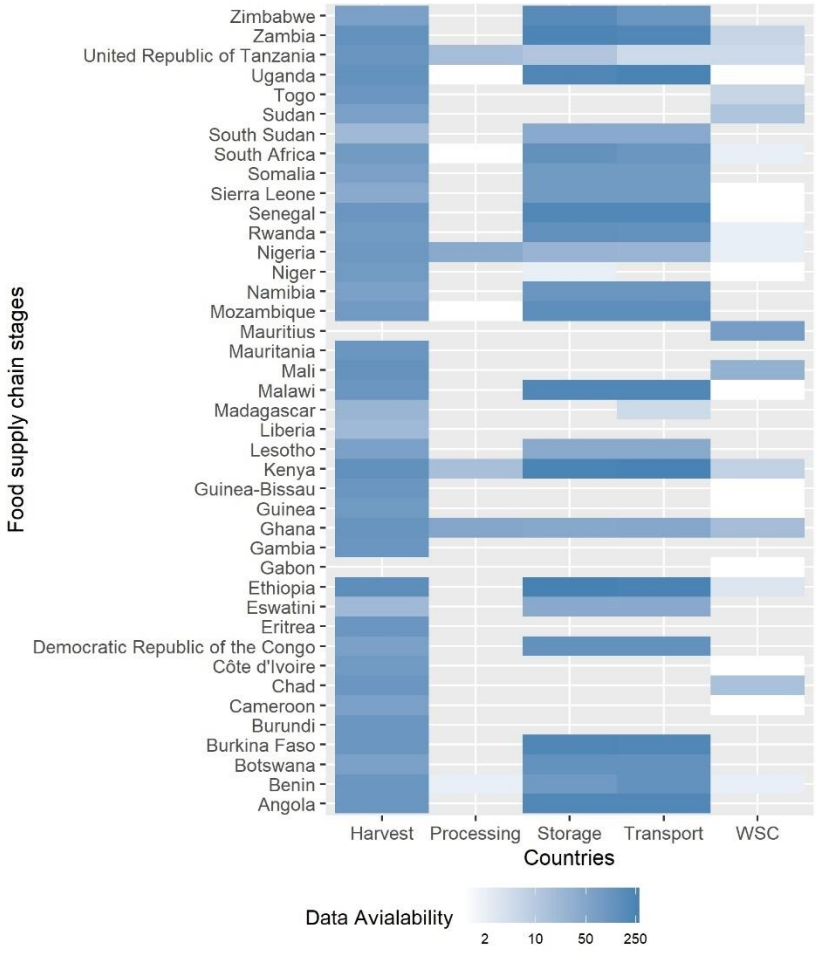


Figure 40: Heatmap of data availability across food supply stages and countries

Appendix F: Wide data analysis

A critical consideration in data analysis is the choice between handling the present dataset in its long or wide format. In this case, long format data is the original state of the dataset after data its retrieval, without any transformations applied. In the long format data, each entry of food loss [%] is represented by a single row, which leads to the fact that one row (data point) can always only refer to one single SC stage. In contrast, the wide format data makes each data point represent an entire SC, including all stages considered within the scope of the thesis. The FL in [%] are then represented as individual columns within the dataset, one separate columns for each SC stage, consequently increasing the number of columns.

The advantage of the wide format data over the long format data lies in its ability to reduce the potential bias that results from focusing mainly on those FL spots along the SC where significant losses are anticipated. The wide format includes all SC stages for every combination of country, commodity, and year, treating the data in a way that assumes complete data for each SC stage for every country, commodity, and year. The downside of this approach is that it may result in data gaps, and in cases where data of FL on a particular SC stage is missing, it could lead to the erroneous assumption that no food loss has occurred, when filling these gaps with zeros.

The findings in Figure 41 highlight that the data points belonging to the food category of vegetables have originally been collected in a highly selective manner, since its overall FL values are significantly lower compared to the stacked bar chart in the long data format in Chapter 4.3.6.

Table 5: Exemplary excerpt of the dataset in long format (own figure)

Country	**Commodity**	**Year**	**SC-Stage**	**FL [%]**
Country A	Commodity A	Year A	Stage U	Percentage X
Country A	Commodity A	Year A	Stage V	Percentage Y
Country A	Commodity A	Year A	Stage W	Percentage Z

Table 6: Exemplary excerpt of the dataset in wide format (own figure)

Country	**Commodity**	**Year**	**Food loss on SC-Stage U**	**Food loss on SC-Stage V**	**Food loss on SC-Stage W**
Country A	Commodity A	Year A	Percentage X	Percentage Y	Percentage Z

Figure 41: FL across food categories in the wide data format, stacked bar chart

Appendix G – Comparison of LPI indicator "infrastructure" of SSA countries and countries world-wide

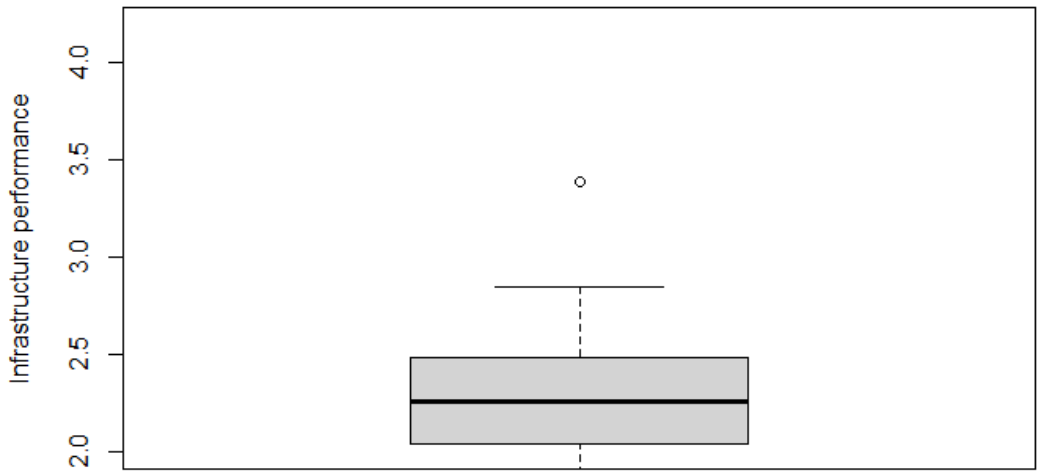


Figure 42: Boxplot of overall LPI scores of all countries included in the FLI (own fiure, CC90)

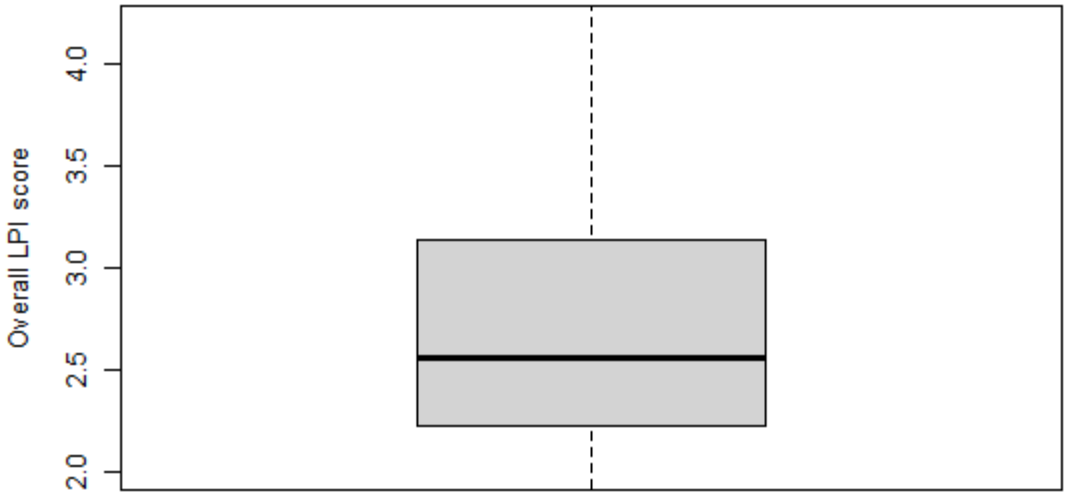


Figure 43: Boxplot of overall LPI scores of all countries included in the FLI (own figure, CC90)

Appendix H – Coding of Project Data

Coding of Project Data

Lennard E.-A. Heuer

2022-17-04

-0- General Information

- CC represents "Code Chunk"
- The plots that are shown in this knitted RMarkdown document should only indicate to which plot certain parts of coding actually belong. The plots are not formatted according for the best visual performance in RMarkdown, but they were optimized for exporting them to a png-file stored on the computer.

-1- Loading Data and Packages (data colleciton)

CC01 Loading the packages (not part of the compiling, but packages have to be installed prior to running the RMarkdown script)

```
install.packages("readxl")
install.packages("writexl")
install.packages("tidyverse")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("ggforce")
install.packages("plotly")
install.packages("stringr")
install.packages("gghighlight")
install.packages("plots")
install.packages("wordcloud")
install.packages("RColorBrewer")
install.packages("wordcloud2")
install.packages("tm")
install.packages("treemapify")
install.packages("reticulate")
```

CC02 Loading several standart libraries

```
library(readxl)
library(writexl)
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyv
erse 1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   1.0.0
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.5.0
```

```
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ──────────────────────────────────────── tidyverse_c
onflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()

library(ggplot2)
library(dplyr)
```

CC03 Loading of all data relevant for the thesis

```
## New names:
## • `Customs` -> `Customs...4`
## • `Customs` -> `Customs...5`
## • `Infrastructure` -> `Infrastructure...6`
## • `Infrastructure` -> `Infrastructure...7`
## • `International shipments` -> `International shipments...8`
## • `International shipments` -> `International shipments...9`
## • `Logistics competence` -> `Logistics competence...10`
## • `Logistics competence` -> `Logistics competence...11`
## • `Tracking & tracing` -> `Tracking & tracing...12`
## • `Tracking & tracing` -> `Tracking & tracing...13`
## • `Timeliness` -> `Timeliness...14`
## • `Timeliness` -> `Timeliness...15`
```

-2- Data Processing

CC04 Merging the data

```
# For merging the data, the column names across the data frames
# "SSA_m49_codes_plain" and "m49"
colnames(SSA_m49_codes_plain) <- c("m49_code")
m49 <- m49[,1:2]
colnames(m49) <- c("m49_code", "Country")

# First merge between the project data (PD) and the m49-codes of the c
ountries
# The scope of the research
PD <- merge(PD, SSA_m49_codes_plain, "m49_code")

# Second merge between PD and the data of the Logistics Performance
# Indicator, aggregated over the years 2012-2018
# This is problematical, since the naming of SSA countries varies
# by source. Thus, a harmonization of names has to be achieved first.

# Then, compare the names of the SSA countries in the name-code key da
ta frame
# and the names of the African countries in the LPI data frame.

Key_SSA_Countries_m49_code <- merge(m49, SSA_m49_codes_plain, "m49_cod
e")

Key_No_SSA_Countries_m49_code <-
```

```
  subset(m49, !m49$m49_code %in% SSA_m49_codes_plain$m49_code)

# First sorting out of countries not are not of those countries that a
re
# definetily out of the scope. Hence, the might be in the geographical
scope or
# there is a problem of missing data or naming of countries
LPI_Agg_12_18_FS <- subset(LPI_Agg_12_18, !LPI_Agg_12_18$Country %in%
Key_No_SSA_Countries_m49_code$Country)

# These countries definitely fall inside the scope
LPI_Agg_12_18_DI <- subset(LPI_Agg_12_18, LPI_Agg_12_18$Country %in% K
ey_SSA_Countries_m49_code$Country)

# As for these countries it is uncertain whether they fall into the sc
ope or
# not. They must be further investigated on.
To_be_cleared <- subset(LPI_Agg_12_18_FS, !LPI_Agg_12_18_FS$Country %i
n% LPI_Agg_12_18_DI$Country)

# These countries are missing in the LPI
Missing_SSA_Countries_in_LPI = subset(Key_SSA_Countries_m49_code, !Key
_SSA_Countries_m49_code$Country %in% LPI_Agg_12_18_DI$Country)

To_be_cleared <- To_be_cleared[order(To_be_cleared$Country),]
Missing_SSA_Countries_in_LPI <- Missing_SSA_Countries_in_LPI[order(Mis
sing_SSA_Countries_in_LPI$Country),]

# As for the countries in the Missing_SSA_Countries_in_LPI-table, it i
s certian
# that they are SSA countreis but they don't apppear to be in the FLI
when
# comparing them to the SSA-key-list. Howeve, since they are SSA count
ries one
# would expect them to be in the FLI data base.
# As for the table To_be_cleared, it is just known that they are not d
efinitely
# inside the scope of this thesis. Therefore, they could be inside or
outside
# the scope of this thesis.
# What now follows is a comparison of the two tables.
To_be_cleared

## # A tibble: 26 × 15
##    Country       LPI R…¹ LPI S…² Custo…³ Custo…⁴ Infra…⁵ Infra…⁶ In
ter…⁷ Inter…⁸
##    <chr>           <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
<dbl>   <dbl>
##  1 Bahamas, The       90    2.65      59    2.72      84    2.56
100    2.66
##  2 Bolivia           136    2.36     134    2.24     138    2.16
122    2.48
##  3 Brunei             73    2.78      61    2.7       77    2.59
84    2.74
```

```
##  4 C.A.R.            150    2.26     117    2.35     135    2.17
150    2.25
##  5 Congo, Dem. …     143    2.33     135    2.23     152    2.04
149    2.26
##  6 Congo, Rep.       133    2.38     151    2.07     141    2.12
107    2.58
##  7 Cote d'Ivoire      66    2.89      68    2.66      69    2.67
58    2.96
##  8 Czech Republ…      26    3.62      26    3.34      29    3.38
12    3.65
##  9 Egypt, Arab …      60    2.95      65    2.67      55    2.91
59    2.94
## 10 Gambia, The       138    2.34     149    2.08     161    1.9
91    2.68
## # … with 16 more rows, 6 more variables: `Logistics competence...10
` <dbl>,
## #   `Logistics competence...11` <dbl>, `Tracking & tracing...12` <d
bl>,
## #   `Tracking & tracing...13` <dbl>, Timeliness...14 <dbl>,
## #   Timeliness...15 <dbl>, and abbreviated variable names ¹`LPI Ran
k`,
## #   ²`LPI Score`, ³Customs...4, ⁴Customs...5, ⁵Infrastructure...6,
## #   ⁶Infrastructure...7, ⁷`International shipments...8`,
## #   ⁸`International shipments...9`

Missing_SSA_Countries_in_LPI

##    m49_code                          Country
## 5       132                       Cabo Verde
## 6       140        Central African Republic
## 9       178                            Congo
## 19      384                    Côte d'Ivoire
## 10      180 Democratic Republic of the Congo
## 43      748                         Eswatini
## 16      270                           Gambia
## 34      678           Sao Tome and Principe
## 36      690                        Seychelles
## 41      728                       South Sudan
## 46      834     United Republic of Tanzania
```

Comparison of the two tables:

Cabo Verde – Not part of LPI Central African Republic – "C.A.R." (rename) Congo – "Congo, Rep." (rename) Côte d'Ivoire – "Cote d'Ivoire" (rename) Democratic Republic of the Congo – "Congo, Dem. Rep." (rename) Eswatini - Not part of LPI Gambia – "Gambia, The" (rename) Sao Tome and Principe - "São Tomé and Príncipe" (rename) Seychelles - Not part of LPI South Sudan - Not part of LPI United Republic of Tanzania - "Tanzania" (rename)

CC05 Rename of country names in the LPI

```
LPI_Agg_12_18$Country <- gsub("C.A.R.", "Central African Republic",
                              LPI_Agg_12_18$Country)
LPI_Agg_12_18$Country <- gsub("Cote d'Ivoire", "Côte d'Ivoire",
                              LPI_Agg_12_18$Country)
```

```
LPI_Agg_12_18$Country <- gsub("Congo, Dem. Rep."
                  , "Democratic Republic of the Congo", LPI_Agg_12_18$
Country)
LPI_Agg_12_18$Country <- gsub("Gambia, The", "Gambia", LPI_Agg_12_18$C
ountry)
LPI_Agg_12_18$Country <- gsub("São Tomé and Príncipe",
                              "Sao Tome and Principe", LPI_Agg_12_18$C
ountry)
LPI_Agg_12_18$Country <- gsub("Tanzania", "United Republic of Tanzania
", LPI_Agg_12_18$Country)
```

Remark: It is obvious that LPI data from certain countries is missing. This issue will be touched upon later.

CC06 Limiting the project data to the geographical scope of the thesis

```
# Create subset of LPI data in which all entries are definitely inside
the scope of the thesis.
LPI_Agg_12_18_DI <- subset(LPI_Agg_12_18, LPI_Agg_12_18$Country %in% K
ey_SSA_Countries_m49_code$Country)

# Joining the LPI and m49-data
hand_over_to_PD <- full_join(Key_SSA_Countries_m49_code,
                             LPI_Agg_12_18_DI, by = "Country")
# Countries not inside the FLI
subset(hand_over_to_PD, is.na(`LPI Score`))

##    m49_code     Country LPI Rank LPI Score Customs...4 Customs...5
## 5       132  Cabo Verde       NA        NA          NA          NA
## 9       178       Congo       NA        NA          NA          NA
## 36      690  Seychelles       NA        NA          NA          NA
## 41      728 South Sudan       NA        NA          NA          NA
## 43      748    Eswatini       NA        NA          NA          NA
##    Infrastructure...6 Infrastructure...7 International shipments...
8
## 5                  NA                 NA                          N
A
## 9                  NA                 NA                          N
A
## 36                 NA                 NA                          N
A
## 41                 NA                 NA                          N
A
## 43                 NA                 NA                          N
A
##    International shipments...9 Logistics competence...10
## 5                           NA                         NA
## 9                           NA                         NA
## 36                          NA                         NA
## 41                          NA                         NA
## 43                          NA                         NA
##    Logistics competence...11 Tracking & tracing...12 Tracking & tra
cing...13
## 5                         NA                      NA
NA
```

```
## 9                          NA                         NA
NA
## 36                         NA                         NA
NA
## 41                         NA                         NA
NA
## 43                         NA                         NA
NA
##    Timeliness...14 Timeliness...15
## 5               NA              NA
## 9               NA              NA
## 36              NA              NA
## 41              NA              NA
## 43              NA              NA

# The countries above are not part of the FLI
PD <- full_join(PD, hand_over_to_PD, by = "m49_code")
```

CC07 Dropping and renaming columns

```
# The ranking data of the indicator categories can be dropped
PD <- PD[, -which(names(PD) %in% c("Customs...4", "Infrastructure...6"
,
                                   "International shipments...8", "Log
istics competence...10", "Tracking & tracing...12", "Timeliness...14")
)]

# It is tedious to deal with spaced column names in R.
# Therefore, spaces are replaced by underlines here.
# In order no to confuse the program "&" is replaced by
# an "and"
names(PD) <- str_replace_all(names(PD), c(" " = "_"))
names(PD) <- str_replace_all(names(PD), c("&" = "and"))

# Renaming of certain columns
PD <- PD  %>% rename(
    LPI_rank = LPI_Rank,
    LPI_score = LPI_Score,
    customs = Customs...5,
    infrastructure = Infrastructure...7,
    international_shipments = International_shipments...9,
    logistics_competence = Logistics_competence...11,
    tracking_tracing = Tracking_and_tracing...13,
    timeliness = Timeliness...15
)
```

-2- Constructing the data

CC08 First categorization of food into food categories

```
PD <- PD %>%
  add_column(food_category = NA)

#For the sake of filtering
PD$cpc_code <- as.character(PD$cpc_code)
```

```
PD$food_category <- as.character(PD$food_category)

# Categories easy to code
for (i in 1:length(PD$cpc_code)) {
  if (grepl("^011", PD$cpc_code[i])) {
    PD$food_category[i] =  "Cereals"
  } else if (grepl("^015", PD$cpc_code[i])) {
                PD$food_category[i] =  "R&T"
  } else if (grepl("^014", PD$cpc_code[i]) |
             (grepl("^017", PD$cpc_code[i]))){
                PD$food_category[i] = "O&P"
  } else if (grepl("^013", PD$cpc_code[i]) &
             !(grepl("^0137", PD$cpc_code[i]))) {
                PD$food_category[i] = "Fruits"
  } else if (grepl("^012", PD$cpc_code[i])) {
                PD$food_category[i] =  "Veget."
  } else if (grepl("^211", PD$cpc_code[i])) {
                PD$food_category[i] =  "Meat"
  } else if (grepl("^04", PD$cpc_code[i])) {
                PD$food_category[i] =  "Fish"
  } else if (grepl("^22", PD$cpc_code[i]) &
            (!grepl("^223", PD$cpc_code[i]))) {
                PD$food_category[i] =  "Dairy"
  } else if (grepl("^223", PD$cpc_code[i])) {
                PD$food_category[i] =  "Eggs"
  } else {
                PD$food_category[i] =  "Others/NA"
  }
}

unique(PD$food_category)

## [1] "Cereals"   "Fruits"    "R&T"       "Veget."    "Others/NA" "O&
P"
## [7] "Meat"      "Dairy"

# Observations that are not assigned to a food category yet
count(PD, food_category == "Others/NA")

##   food_category == "Others/NA"     n
## 1                        FALSE 18943
## 2                         TRUE    70

obs_without_f_catgory = filter(PD, food_category == "Others/NA")
print(sort(unique(obs_without_f_catgory$cpc_code)))

##  [1] "01640"    "01651"    "01652"    "02111"    "02211"    "0231"
##  [7] "21421"    "23110"    "23120.09" "23161.02" "23170.01"
```

There are 10 cpc_codes that weren't assigned to a food category yet.
https://unstats.un.org/unsd/classifications/Econ/Structure

I (Cacao beans): 01640 - Cocoa beans

II (Capsicum): 01651 - Pepper (Piper spp.), raw 01652 - Chillies and peppers, dry (Capsicum spp., Pimenta spp.), raw

III (Live animals): 02111 - Cattle (live animals)

IV (Milk): 02211 - Raw milk of cattle

V (Hen Eggs): 0231 - Hen eggs in shell, fresh

VI (Nuts): 21421 - Groundnuts, shelled

VII (Grain mill products): 23110 - Wheat and meslin flour 23120.09 - Other cereal flours 23161.02 - Rice, semi- or wholly milled 23170.01 - Other vegetable flours and meals

CC09 Assign remaining food categories

```
for (i in 1:length(PD$cpc_code)) {
         if (grepl("^01640", PD$cpc_code[i]) |
            (grepl("^21421", PD$cpc_code[i]))){
                PD$food_category[i] =  "N&C"
  } else if ((grepl("^01651", PD$cpc_code[i])) |
             (grepl("^01652", PD$cpc_code[i]))){
                PD$food_category[i] = "Veget."
  } else if (grepl("^02211", PD$cpc_code[i])) {
                PD$food_category[i] = "Dairy"
  } else if (grepl("^0231", PD$cpc_code[i])) {
                PD$food_category[i] = "Eggs"
  } else if ((grepl("^23110", PD$cpc_code[i])) |
            (grepl("^23120", PD$cpc_code[i])) |
            (grepl("^23161", PD$cpc_code[i])) |
            (grepl("^23170", PD$cpc_code[i]))){
                PD$food_category[i] = "Cereals"
  }
}
Cattle = filter(PD, cpc_code == "02111")
count(Cattle)

##   n
## 1 1

print(Cattle)

##   m49_code country region cpc_code commodity year loss_percentage
## 1      288   Ghana           02111    Cattle 2011              28
##   loss_percentage_original loss_quantity activity food_supply_stage
treatment
## 1                       28                              Post-harvest
##   cause_of_loss sample_size method_data_collection   reference
## 1                                Literature Review Egyir, 2011
##
url
## 1 http://www.mod.gov.tr/Lists/RecentPublications/Attachments/120/Re
ducing%20Postharvest%20Losses%20in%20the%20OIC%20Member%20Countries.pd
f
##   notes index Country LPI_rank LPI_score customs infrastructure
```

```
## 1        6586   Ghana      101       2.6    2.41           2.46
##   international_shipments logistics_competence tracking_tracing tim
eliness
## 1                    2.63                 2.51            2.58
2.95
##   food_category
## 1     Others/NA

# There is only one cattle observation. Since the respective food_supp
ly_stage
# was named "post-harvest", it can be assumed that meat of cattle is a
ctually
# meant. Therefore, the data point will be assigned to the food catego
ry meat.
for (i in 1:length(PD$cpc_code)) {
         if (grepl("^02111", PD$cpc_code[i])) {
                PD$food_category[i] =  "Meat"
    }
}
```

CC10 Checking the food categories again for unassigned data points

```
subset(PD, food_category == "Others/NA")

##       m49_code country region cpc_code commodity year loss_percenta
ge
## 19007      132    <NA>   <NA>     <NA>      <NA>   NA
NA
## 19008      140    <NA>   <NA>     <NA>      <NA>   NA
NA
## 19009      174    <NA>   <NA>     <NA>      <NA>   NA
NA
## 19010      178    <NA>   <NA>     <NA>      <NA>   NA
NA
## 19011      226    <NA>   <NA>     <NA>      <NA>   NA
NA
## 19012      678    <NA>   <NA>     <NA>      <NA>   NA
NA
## 19013      690    <NA>   <NA>     <NA>      <NA>   NA
NA
##       loss_percentage_original loss_quantity activity food_supply_s
tage
## 19007                     <NA>          <NA>     <NA>
<NA>
## 19008                     <NA>          <NA>     <NA>
<NA>
## 19009                     <NA>          <NA>     <NA>
<NA>
## 19010                     <NA>          <NA>     <NA>
<NA>
## 19011                     <NA>          <NA>     <NA>
<NA>
## 19012                     <NA>          <NA>     <NA>
<NA>
## 19013                     <NA>          <NA>     <NA>
```

```
<NA>
##       treatment cause_of_loss sample_size method_data_collection re
ference  url
## 19007      <NA>          <NA>        <NA>                    <NA>
<NA> <NA>
## 19008      <NA>          <NA>        <NA>                    <NA>
<NA> <NA>
## 19009      <NA>          <NA>        <NA>                    <NA>
<NA> <NA>
## 19010      <NA>          <NA>        <NA>                    <NA>
<NA> <NA>
## 19011      <NA>          <NA>        <NA>                    <NA>
<NA> <NA>
## 19012      <NA>          <NA>        <NA>                    <NA>
<NA> <NA>
## 19013      <NA>          <NA>        <NA>                    <NA>
<NA> <NA>
##       notes index                  Country LPI_rank LPI_score custo
ms
## 19007  <NA>    NA               Cabo Verde       NA        NA
NA
## 19008  <NA>    NA Central African Republic      150      2.26    2.
35
## 19009  <NA>    NA                  Comoros      114      2.51    2.
58
## 19010  <NA>    NA                    Congo       NA        NA
NA
## 19011  <NA>    NA        Equatorial Guinea      156      2.21    1.
99
## 19012  <NA>    NA    Sao Tome and Principe      105      2.56    2.
52
## 19013  <NA>    NA               Seychelles       NA        NA
NA
##       infrastructure international_shipments logistics_competence
## 19007             NA                      NA                    NA
## 19008           2.17                    2.25                  2.13
## 19009           2.27                    2.47                  2.32
## 19010             NA                      NA                    NA
## 19011           1.82                    2.46                  2.11
## 19012           2.30                    2.44                  2.55
## 19013             NA                      NA                    NA
##       tracking_tracing timeliness food_category
## 19007               NA         NA     Others/NA
## 19008             2.21       2.46     Others/NA
## 19009             2.67       2.74     Others/NA
## 19010               NA         NA     Others/NA
## 19011             2.14       2.66     Others/NA
## 19012             2.66       2.90     Others/NA
## 19013               NA         NA     Others/NA

PD = subset(PD, food_category != "Others/NA")
```

CC11 Creating a back-up-version of the dataset after first processing of the dataset

```
PD_orig. <- PD
PD <- PD_orig.
```

There is no more data that is not assigned to a food category yet.

CC12 Retrieve list of all SC stages and take a look at activities on the SC stage "farm"

```
# These are all SC stages within the dataset
unique(PD$food_supply_stage)

##  [1] "Farm"               "Transport"          "Storage"
##  [4] "Harvest"            "Whole supply chain" "Export"
##  [7] "Wholesale"          "Retail"             "Processing"
## [10] ""                   "Trader"             "Post-harvest"
## [13] "Market"             "Distribution"       "Households"

# These are all activities that occur among data point belonging to th
e
# SC stage "farm".
unique((filter(PD, PD$food_supply_stage == "Farm"))$activity)

##  [1] "Shelling, Threshing"     "Storage"
##  [3] "Winnowing"               "Transportation"
##  [5] "Drying"                  ""
##  [7] "Sorting"                 "Grading, Sorting"
##  [9] "Stacking"                "Threshing"
## [11] "Farm"                    "Assembling, Farm"
## [13] "Farm, Handling, Storage" "Bagging"
## [15] "Handling"                "Harvesting"
## [17] "Shelling"                "Lifting"
## [19] "Handling, Storage"       "Harvesting, Storage"
## [21] "Milling"                 "Drying, Farm"

# Comparison of the activities on the farm SC stages with the activiti
es on the
# processing SC stage
unique((filter(PD, PD$food_supply_stage == "Processing"))$activity)

##  [1] ""                       "Sifting"               "Processing"
##  [4] "Grating"                "Grading, Sorting"      "Drying"
##  [7] "Peeling"                "Roasting"              "Dewatering"
## [10] "Packaging"              "Packaging, Processing" "Bagging, Pack
aging"
## [13] "Milling"                "Preservation"          "Ripening"
## [16] "Processing, Storage"
```

There are many equal and similar entries of data points belonging to the farm SC stage and data points belonging to the processing SC stage.

CC13 Breaking up the SC stage of farm I

```
PD$food_supply_stage <- as.character(PD$food_supply_stage)

PD$method_data_collection <- as.factor(PD$method_data_collection)
summary(PD$method_data_collection)
```

```
##
##                                                  41
##                                          Case Study
##                                                 188
##                                              Census
##                                                   1
##                                 Controlled Experiment
##                                                 114
##                                       Expert Opinion
##                                                  30
## FAO's annual Agriculture Production Questionnaires
##                                                 147
##                                    Literature Review
##                                                  79
##                                  Modelled Estimates
##                                               17994
##                         No Data Collection Specified
##                                                 138
##                                              Survey
##                                                 274

# Using the feature activity as an indicator, which SC stage to transf
er the
# data points to.

for (i in 1:length(PD$cpc_code)) {
  if(PD$food_supply_stage[i] == "Farm" &
    (PD$activity[i] == "Harvest" | PD$activity[i] == "Harvesting, Stor
age" |
     PD$activity[i] == "Harvesting")) {
      PD$food_supply_stage[i] = "Harvest"
  }
}

for (i in 1:length(PD$cpc_code)) {
  if(PD$food_supply_stage[i] == "Farm" & (PD$activity[i] == "Transport
ation")){
      PD$food_supply_stage[i] = "Transport"
  }
}

for (i in 1:length(PD$cpc_code)) {
  if(PD$food_supply_stage[i] == "Farm" & (PD$activity[i] == "Storage"
|
     PD$activity[i] == "Farm, Handling, Storage" |
     PD$activity[i] == "Stacking" | PD$activity[i] =="Handling, Storag
e" |
     PD$activity[i] == "Handling" | PD$activity[i] == "Lifting")) {
      PD$food_supply_stage[i] = "Storage"
  }
}

subset(PD, activity == "Lifting")
```

```
##      m49_code country region cpc_code                     commodity
year
## 8632      454  Malawi   Zulu     0142 Groundnuts, excluding shelled
2015
## 8633      454  Malawi   Zulu     0142 Groundnuts, excluding shelled
2015
##      loss_percentage loss_percentage_original loss_quantity activit
y
## 8632               6                        6               Liftin
g
## 8633               6                        6               Liftin
g
##      food_supply_stage treatment     cause_of_loss   sample_size
## 8632           Storage           Method Of Lifting 20 households
## 8633           Storage           Method Of Lifting 20 households
##            method_data_collection reference
## 8632 No Data Collection Specified
## 8633
##
url
## 8632 https://www.dropbox.com/sh/oii6pjz81e4634x/AABvz_fGQYEC3d4Urzj
olEHDa?dl=0
## 8633 https://www.dropbox.com/sh/oii6pjz81e4634x/AABvz_fGQYEC3d4Urzj
olEHDa?dl=0
##      notes index Country LPI_rank LPI_score customs infrastructure
## 8632       12180  Malawi       84      2.69    2.58           2.56
## 8633       12181  Malawi       84      2.69    2.58           2.56
##      international_shipments logistics_competence tracking_tracing
timeliness
## 8632                    2.61                 2.76            2.65
2.99
## 8633                    2.61                 2.76            2.65
2.99
##      food_category
## 8632           O&P
## 8633           O&P
```

CC14 The activities that are not assigned yet are listed below

```
unique((filter(PD, PD$food_supply_stage == "Farm"))$activity)
```

```
##  [1] "Shelling, Threshing" "Winnowing"           "Drying"
##  [4] ""                    "Sorting"             "Grading, Sorting"
##  [7] "Threshing"           "Farm"                "Assembling, Farm"
## [10] "Bagging"             "Shelling"            "Milling"
## [13] "Drying, Farm"
```

CC15 Breaking up the SC stage of farm I

```
for (i in 1:length(PD$cpc_code)) {
  if(PD$food_supply_stage[i] == "Farm" &
     (PD$activity[i] == "Shelling, Threshing" | PD$activity[i] == "Win
nowing" |
      PD$activity[i] == "Drying" | PD$activity[i] == "Sorting" |
      PD$activity[i] =="Grading, Sorting" | PD$activity[i] == "Threshi
```

```
ng" |
      PD$activity[i] == "Assembling, Farm" | PD$activity[i] == "Baggin
g" |
      PD$activity[i] == "Shelling" | PD$activity[i] == "Milling" |
      PD$activity[i] == "Drying, Farm")) {
        PD$food_supply_stage[i] = "Processing"
  }
}
```

CC16 After constructing and before cleaning, a copy is taken for the use later on.

```
PD_copy <- PD
```

According to the definition of the scope, only supply chain stages until (exclusively) the retail stage are considered. Therefore all data supply chain stages of retail and further downstream need to be dropped. These are export, households, trader, wholesale, market and retail. Blank spaces and whole supply chain data needs to be dropped as well for it is of no use for the case. In the case of distribution, past-harvest, and whole suppply chain they are excludid because it is not clear which stage along the supply chain they are actually referring to.

CC17 Cleaning the food_supply_stage column

```
# Filter out the SC stages that may be relevant for the scope of the t
hesis
PD = subset(PD, !(food_supply_stage %in% c('Distribution','Export',
            'Households', 'Trader', 'Wholesale', 'Market', 'Post-harve
st',
            'Retail', 'Whole supply chain', '')), drop=FALSE)
```

CC18 Quick check-up on remaining farm data and deletion of the same

```
head(filter(PD, food_supply_stage == "Farm"))
```

```
##   m49_code country           region cpc_code
commodity
## 1      204   Benin                     0112                    Mai
ze (corn)
## 2      204   Benin Parakou (Borgou)    01316 Mangoes, guavas and ma
ngosteens
## 3      204   Benin                  01520.01                 Cassa
va, fresh
## 4      204   Benin                     01323
Oranges
## 5      204   Benin                     01234
Tomatoes
## 6      204   Benin                     01234
Tomatoes
##   year loss_percentage loss_percentage_original loss_quantity activ
ity
## 1 2013            27.0                      27%
## 2 2006            50.0                       50
## 3 2013             8.5                      8.5
## 4 2010            10.0                   5.0-15
## 5 2010            53.0                       53
```

```
## 6 2010            26.0                      23-29
##   food_supply_stage treatment                          cause_of_loss s
ample_size
## 1              Farm
## 2              Farm           Measured In May; Due To Fruit Flies 3
000 fruits
## 3              Farm
## 4              Farm
## 5              Farm
## 6              Farm
##         method_data_collection                                    r
eference
## 1            Literature Review                             . Hodge
s (2012)
## 2        Controlled Experiment Vayssieres Korie Coulibaly Temple an
d Boueyi
## 3 No Data Collection Specified                  (Mutungi and Affogn
on 2013)
## 4                   Case Study
Kitinoja
## 5
Kitinoja
## 6                   Case Study
Kitinoja
#
#
url
## 1
http://biblio.iita.org/documents/U14ArtAbassPostharvestInthomDev.pdf-6
08a86dd987926565841fccb38f388ce.pdf
## 2 https://www.researchgate.net/publication/46103208_The_mango_tree_
in_central_and_northern_Benin_damage_caused_by_fruit_flies_Diptera_Tep
hritidae_and_computation_of_economic_injury_level
## 3                                          http://www.mod.go
v.tr/Lists/RecentPublications/Attachments/120/Reducing%20Postharvest%2
0Losses%20in%20the%20OIC%20Member%20Countries.pdf
## 4
http://www.sciencedirect.com/science/article/pii/S0305750X14002307
## 5
http://www.sciencedirect.com/science/article/pii/S0305750X14002307
## 6
http://ucanr.edu/datastoreFiles/234-1847.pdf
##
notes
## 1 . Hodges (2012) estimated post-harvest loss of grains in  Tanzani
a as 22% (excluding field loss) and 27% for Benin Republic.    Adebayo
B. Abass a, *, Gabriel Ndunguru a, Peter Mamiro b, Bamidele Alenkhe c,
Nicholas Mlingi a, Mateete Bekunda a; U14ArtAbassPostharvestInthomDev.
pdf-608a86dd987926565841fccb38f388ce.pdf
##
2
##
3
## 4
```

```
## 5
##
6
##   index Country LPI_rank LPI_score customs infrastructure
## 1  2676   Benin       93      2.65    2.48           2.45
## 2  2922   Benin       93      2.65    2.48           2.45
## 3  2679   Benin       93      2.65    2.48           2.45
## 4  2780   Benin       93      2.65    2.48           2.45
## 5  2777   Benin       93      2.65    2.48           2.45
## 6  2766   Benin       93      2.65    2.48           2.45
##   international_shipments logistics_competence tracking_tracing tim
eliness
## 1                    2.66                  2.5            2.58
3.17
## 2                    2.66                  2.5            2.58
3.17
## 3                    2.66                  2.5            2.58
3.17
## 4                    2.66                  2.5            2.58
3.17
## 5                    2.66                  2.5            2.58
3.17
## 6                    2.66                  2.5            2.58
3.17
##   food_category
## 1       Cereals
## 2        Fruits
## 3           R&T
## 4        Fruits
## 5        Veget.
## 6        Veget.

# There are still 96 rows that are unassigned. They are dropped from t
he
# dataset, as it is not clear which SC stage they are referring to bec
ause they
# do either contain no data on the activity or they contain activity d
ata that
# cannot unambiguously be related to any supply chain stage.

# As it is not clear, which SC stage they would come close to, the dec
ision was
# made to drop them from the dataset used in this data analysis.
PD <- filter(PD, food_supply_stage != "Farm")

# Setting levels of the remaining SC stages is only important for the
order by
# which the SC stages are shown later on in the plots.
PD$food_supply_stage <- factor(PD$food_supply_stage, levels=c("Harvest
",
  "Processing", "Transport", "Storage"))
```

The following approach is sophisticated and implies that if in comparing of two data points country, commodity, year, method_data_collection, food_suppy_stage,

region, treatment, reference, and URL are the same, while the combination of cause_of_loss or activity differ, the values of loss percentage are added, while all other columns stay the same.

CC19 Adding up values in the column of loss_percentage

```
PD$gsize = 1
PD$gID = 0

PD_dc_comp <- PD %>%
  group_by(country, commodity, year, method_data_collection,
           food_supply_stage, region, treatment, reference, url) %>%
           distinct(cause_of_loss, activity, .keep_all = TRUE) %>%
           ungroup()

# The table "gap" is the data container of data points that don't need
to be
# stacked as they occur without another data point sharing the same co
mbination
# of features discussed above.
gap = anti_join(PD, PD_dc_comp)

## Joining, by = c("m49_code", "country", "region", "cpc_code", "commo
dity",
## "year", "loss_percentage", "loss_percentage_original", "loss_quanti
ty",
## "activity", "food_supply_stage", "treatment", "cause_of_loss", "sam
ple_size",
## "method_data_collection", "reference", "url", "notes", "index", "Co
untry",
## "LPI_rank", "LPI_score", "customs", "infrastructure",
## "international_shipments", "logistics_competence", "tracking_tracin
g",
## "timeliness", "food_category", "gsize", "gID")

PD_dc <- PD %>%
    group_by(country, commodity, year, method_data_collection,
             food_supply_stage, region, treatment, reference, url) %>%
             distinct(cause_of_loss, activity, .keep_all = TRUE) %>%
             mutate(loss_percentage = sum(loss_percentage),
                    gID = cur_group_id(), gsize = n()) %>%
                    ungroup()

PD <- rbind(PD_dc, gap)
```

-3- EDA and Data Cleaning

CC20 - View on the structure of the data

```
str(PD)

## tibble [18,509 × 31] (S3: tbl_df/tbl/data.frame)
##  $ m49_code                : int [1:18509] 24 24 24 24 24 24 24 24
24 24 ...
```

```
##  $ country                 : chr [1:18509] "Angola" "Angola" "Angol
a" "Angola" ...
##  $ region                  : chr [1:18509] "" "" "" "" ...
##  $ cpc_code                : chr [1:18509] "0114" "0114" "0118" "01
14" ...
##  $ commodity               : chr [1:18509] "Sorghum" "Sorghum" "Mil
let" "Sorghum" ...
##  $ year                    : int [1:18509] 2019 2019 2019 2020 2019
2019 2008 2019 2008 2019 ...
##  $ loss_percentage         : num [1:18509] 3.6 1 1.3 3.66 1.3 5.39
1.25 1.3 5.43 2.5 ...
##  $ loss_percentage_original: chr [1:18509] "3.6" "1" "1.3" "3.66" .
..
##  $ loss_quantity           : chr [1:18509] "" "" "" "" ...
##  $ activity                : chr [1:18509] "Shelling, Threshing" "T
ransportation" "Storage" "Storage" ...
##  $ food_supply_stage       : Factor w/ 4 levels "Harvest","Processi
ng",..: 2 3 4 4 4 2 3 4 2 3 ...
##  $ treatment               : chr [1:18509] "" "" "" "" ...
##  $ cause_of_loss           : chr [1:18509] "" "" "" "" ...
##  $ sample_size             : chr [1:18509] "" "" "" "" ...
##  $ method_data_collection  : Factor w/ 10 levels "","Case Study",..
: 8 8 8 8 8 8 8 8 8 8 ...
##  $ reference               : chr [1:18509] "" "" "" "" ...
##  $ url                     : chr [1:18509] "https://www.aphlis.net/
en/page/20/data-tables#/datatables?tab=value_chain&metric=prc" "https:
//www.aphlis.net/en/page/20/data-tables#/datatables?tab=value_chain&me
tric=prc" "https://www.aphlis.net/en/page/20/data-tables#/datatables?t
ab=value_chain&metric=prc" "https://www.aphlis.net/en/page/20/data-tab
les#/datatables?tab=value_chain&metric=prc" ...
##  $ notes                   : chr [1:18509] "" "" "" "" ...
##  $ index                   : int [1:18509] 4739 4740 4746 4708 4734
4736 5099 4741 5098 4744 ...
##  $ Country                 : chr [1:18509] "Angola" "Angola" "Angol
a" "Angola" ...
##  $ LPI_rank                : num [1:18509] 160 160 160 160 160 160
160 160 160 160 ...
##  $ LPI_score               : num [1:18509] 2.18 2.18 2.18 2.18 2.18
2.18 2.18 2.18 2.18 2.18 ...
##  $ customs                 : num [1:18509] 1.79 1.79 1.79 1.79 1.79
1.79 1.79 1.79 1.79 1.79 ...
##  $ infrastructure          : num [1:18509] 2.01 2.01 2.01 2.01 2.01
2.01 2.01 2.01 2.01 2.01 ...
##  $ international_shipments : num [1:18509] 2.33 2.33 2.33 2.33 2.33
2.33 2.33 2.33 2.33 2.33 ...
##  $ logistics_competence    : num [1:18509] 2.13 2.13 2.13 2.13 2.13
2.13 2.13 2.13 2.13 2.13 ...
##  $ tracking_tracing        : num [1:18509] 2.14 2.14 2.14 2.14 2.14
2.14 2.14 2.14 2.14 2.14 ...
##  $ timeliness              : num [1:18509] 2.65 2.65 2.65 2.65 2.65
2.65 2.65 2.65 2.65 2.65 ...
##  $ food_category           : chr [1:18509] "Cereals" "Cereals" "Cer
eals" "Cereals" ...
##  $ gsize                   : num [1:18509] 1 1 1 1 1 2 1 1 2 1 ...
```

```
##  $ gID                     : num [1:18509] 330 331 164 336 248 246
203 332 202 163 ...
```

CC21 Changing data types

```
# The changes of data types is necessary for computations later in the
course
# of the data analysis.
PD$cpc_code <- as.numeric(PD$cpc_code)
PD$Country <- as.factor(PD$Country)
PD$country <- as.factor(PD$country)
PD$food_category <- as.factor(PD$food_category)
```

Outliers, missing data (and range)

CC22 Find missing data in interesting columns

```
print ("Row and Col positions of NA values")

## [1] "Row and Col positions of NA values"

summary(which(is.na(PD), arr.ind=TRUE))

##       row             col
##  Min.   :12532   Min.   :21.00
##  1st Qu.:12605   1st Qu.:22.75
##  Median :12931   Median :24.50
##  Mean   :14216   Mean   :24.50
##  3rd Qu.:17832   3rd Qu.:26.25
##  Max.   :17905   Max.   :28.00
```

As column 21 to 28 contain LPI data only, NAs are only found within the LPI data.

CC23 Searching for evidence of quality issues in contract to pure food losses

```
print ("Quality issues")

## [1] "Quality issues"

# Capitalized quatlity
cache <- str_count(PD$cause_of_loss, "Quality")
sum(cache)

## [1] 2

# Not capitlized quality
cache <- str_count(PD$cause_of_loss, "quality")
sum(cache)

## [1] 6
```

CC24 Subsetting of data containing LPI information

```
# The reason the data was kept in the first place was that the data co
ntains
# valuable data depicting the food losses itself.
# As not to do interfere with research in coming code chunks, the data
is
```

```
# subsetted
PD_with_LPI <- subset(PD, !is.na(LPI_score))
PD_with_LPI$food_supply_stage <- as.character(PD_with_LPI$food_supply_
stage)
```

Quality issues are cleary mentioned in the paper at several occasions

CC25 Ranges for all numeric data

```
PD_numeric <- PD[,unlist(lapply(PD, is.numeric))]
data.frame(min=sapply(PD_numeric,min),max=sapply(PD_numeric,max))

##                               min      max
## m49_code                    24.00   894.00
## cpc_code                   111.00 23170.01
## year                      2000.00  2021.00
## loss_percentage              0.01    79.35
## index                        8.00 27773.00
## LPI_rank                       NA       NA
## LPI_score                      NA       NA
## customs                        NA       NA
## infrastructure                 NA       NA
## international_shipments        NA       NA
## logistics_competence           NA       NA
## tracking_tracing               NA       NA
## timeliness                     NA       NA
## gsize                        1.00     6.00
## gID                          0.00 13502.00

PD_with_LPI_numeric <- PD_with_LPI[,unlist(lapply(PD_with_LPI, is.nume
ric))]
data.frame(min=sapply(PD_with_LPI_numeric,min),
           max=sapply(PD_with_LPI_numeric,max))

##                               min      max
## m49_code                    24.00   894.00
## cpc_code                   111.00 23170.01
## year                      2000.00  2021.00
## loss_percentage              0.01    79.35
## index                        8.00 27773.00
## LPI_rank                    29.00   167.00
## LPI_score                    2.00     3.51
## customs                      1.79     3.29
## infrastructure               1.69     3.39
## international_shipments      2.12     3.53
## logistics_competence         1.96     3.42
## tracking_tracing             1.94     3.56
## timeliness                   2.18     3.85
## gsize                        1.00     6.00
## gID                          0.00 13502.00
```

There is no apparent outlier when merely looking at the output of the code chunk above.

One additional observation: The majority of data cleaning has already been done by the FAO, for example in as such as loss percentage_original has been transport in the so-called column loss_column by removing %-signs, computing averages when a range was filled in and transforming decimal data into the 100% scale, without adding the %-sign.

Despite the first cleaning carried out by the FAO, ranges should be tested, as to not overlook outliers. There are no outliers of food loss.

CC26 - Count rows of the data set

```
nrow(PD)

## [1] 18509
```

CC27 - Ranges of selected columns

```
unique(PD$food_supply_stage)

## [1] Processing Transport  Storage    Harvest
## Levels: Harvest Processing Transport Storage
```

CC28 Compare the two columns Country and country

```
setdiff(PD$Country, PD$country)

## character(0)
```

It can be noted that both columns are exactly identical.

CC29 Eliminate the superflous Country-column

```
PD <- select(PD, -c("Country"))
```

-3- Summarization of the data

CC30 Conversion function for saving plots as png-files

```
ready_conversion <- function(x) {
  x +
  theme(axis.text = element_text(size = 9)) +
  # geom_text(aes(label = n), size = 3.5) +
  theme(axis.title = element_text(size = 10.5)) +
  theme(legend.title = element_text(size = 10)) +
  theme(legend.text = element_text(size = 8))
}
```

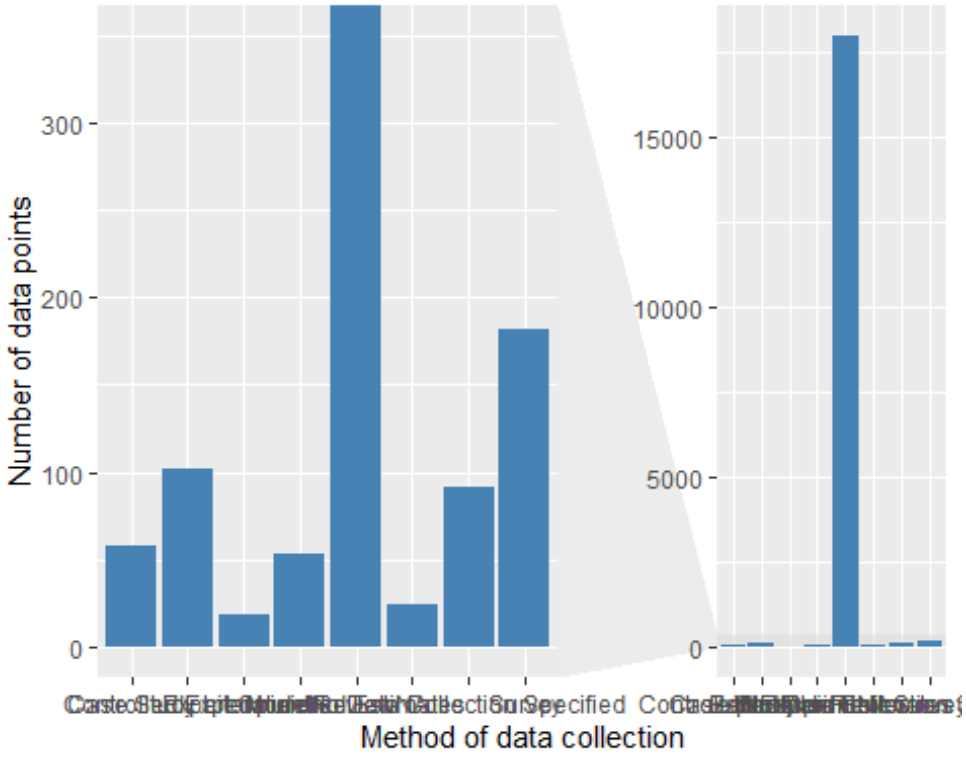
CC31 Occurences of methods of data collection

```
library(ggforce)
# Prior to creating the plot the labeling lenght for the labeling:
# "FAO's annual Agriculture Production Questionnaires" has to be reduc
ed
# in length to a reasonable size

PD$method_data_collection <- as.character(PD$method_data_collection)
```

```
for (i in 1:length(PD$cpc_code)) {
  if (PD$method_data_collection[i] == "") {
    PD$method_data_collection[i] =  "NA"
  }
}

unique(PD$method_data_collection)
```

```
## [1] "Modelled Estimates"           "NA"
## [3] "No Data Collection Specified" "Survey"
## [5] "Case Study"                   "Controlled Experiment"
## [7] "Literature Review"            "Expert Opinion"
```

```
# JPEG device
g <- ggplot(data = PD) +
  aes(x = method_data_collection) +
  geom_bar(stat="count", fill = "steelblue") +
  facet_zoom(ylim = c(0, 350)) +
  ylab("Number of data points") +
  xlab("Method of data collection")
g
```




```
jpeg("Occ_methods.png", quality = 100, width = 15, height = 10.5 ,  units = "cm",
     res= 300)
ready_conversion(g) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
dev.off()
```

```
## png
##   2
```

CC32 A closer look at the modelled estimates data

```
PD_modelled <- subset(PD, method_data_collection == "Modelled Estimate
s")

# Total number of modelled estimates within the dataset inside the sco
pe of
# the thesis
count(PD_modelled)
```

```
## # A tibble: 1 × 1
##       n
##   <int>
## 1 17981
```

```
# Unique data in the modelled estimates
unique(PD_modelled$url)
```

```
## [1] "https://www.aphlis.net/en/page/20/data-tables#/datatables?tab=
value_chain&metric=prc"
## [2] ""
```

```
# Taking subsets of the modelled estimates data according to whether t
hey
# derived from the APHLIS-database
Aphilis_true <- subset(PD_modelled, url == "https://www.aphlis.net/en/
page/20/data-tables#/datatables?tab=value_chain&metric=prc")
Aphilis_false <- subset(PD_modelled, url == "")

# Share of Aphlis data in relation to all modelled estimates
(count(Aphilis_true))/(count(Aphilis_false)+count(Aphilis_true))
```

```
##           n
## 1 0.9986096
```

```
# Share of Aphlis data in relation to whole dataset within the scope o
f
# the thesis
(count(Aphilis_true))/(count(PD))
```

```
##           n
## 1 0.9701226
```

CC33 Food categories across APHLIS data

```
unique(Aphilis_true$food_category)
```

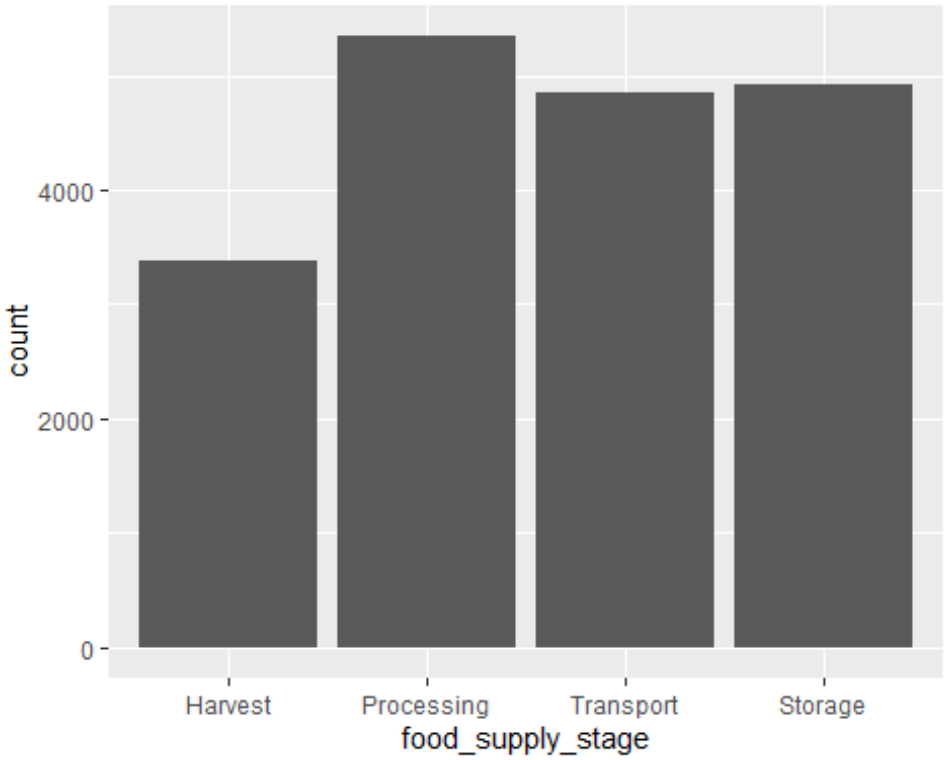
```
## [1] Cereals
## Levels: Cereals Dairy Fruits N&C O&P R&T Veget.
```

All Aphlis data consists of cereal data only.

Impact of both food supply chain stage and food category in food loss percentage is visualized by the means of a heat map, whose fill-scale is logarythmic. A logarythmic fill-scale was chosen do to the divergent nature of the data.

CC34 Data occurences across food supply stages

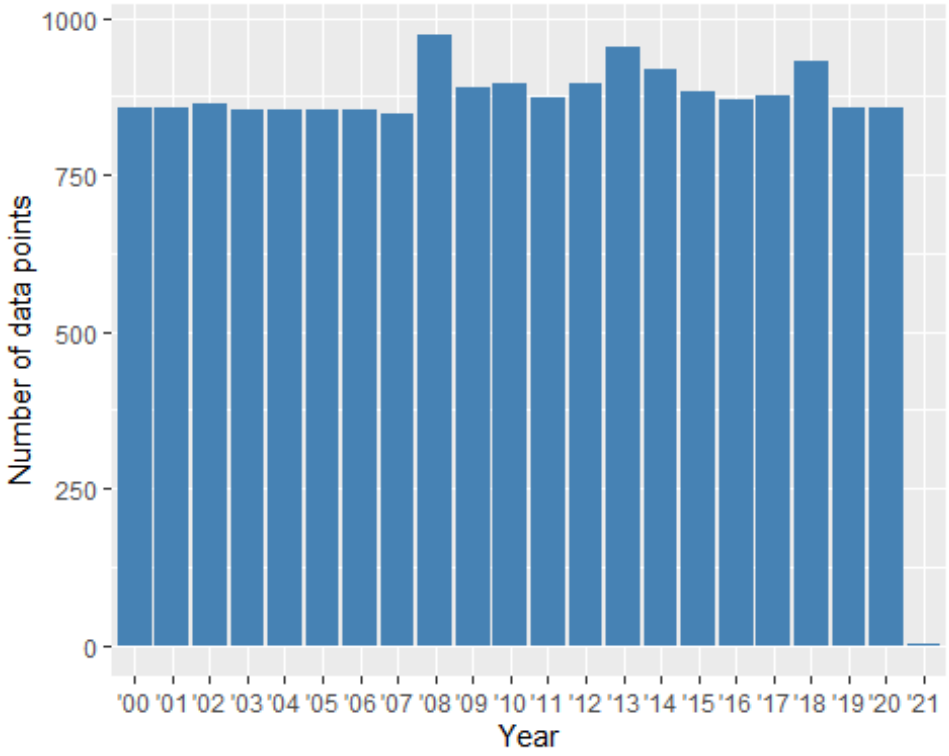
```
g <- ggplot(data = PD) +
  aes(x = food_supply_stage) + geom_bar(stat="count")

g
```




```
jpeg("Occ_SC_stages.png", quality = 100, width = 15, height = 13 ,
     units = "cm", res= 300)
ready_conversion(g)
dev.off()

## png
##   2
```

CC35 Data occurrences across years

```
PD$year <- as.character(PD$year)
substrRight <- function(x, n){
  substr(x, nchar(x)-n+1, nchar(x))
}

PD$years <- substrRight(PD$year, 2)
PD$years <- str_glue("'{PD$years}")
PD$years <- as.factor(PD$years)

p<-ggplot(data=PD, aes(x=years)) +
  geom_bar(stat="count", fill = "steelblue") +
  ylab("Number of data points") +
  xlab("Year")
p
```



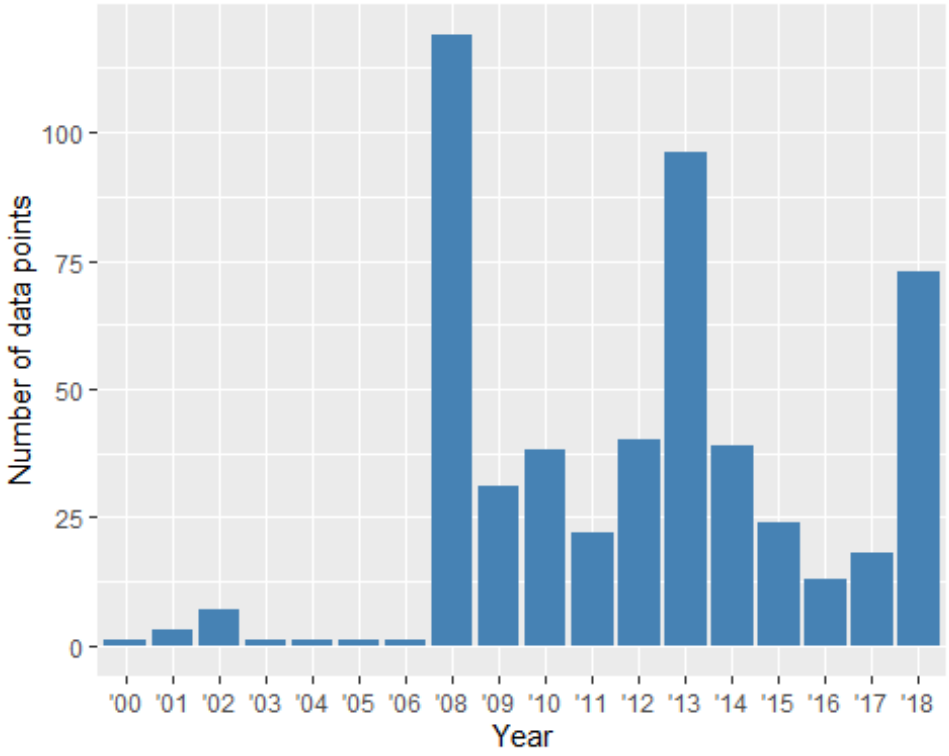

```
# Capture the plot:
jpeg("Occ_years.png", quality = 100, width = 15, height = 8 ,
     units = "cm", res= 300)
ready_conversion(p)
dev.off()

## png
##   2

# https://stackoverflow.com/questions/7963898/extracting-the-last-n-characters-from-a-string-in-r
```

At the first glance it looks like the data points are relatively evenly spread throughout the years.

Now a view on the same data, excluding the modelled data:

CC36 Data occurences across years 2000-2021, without modelled data

```
PD_no_Model<- PD %>%
  filter(PD$method_data_collection != "Modelled Estimates")

g<-ggplot(data=PD_no_Model, aes(x=years)) +
  geom_bar(stat="count", fill = "steelblue") +
  ylab("Number of data points") +
  xlab("Year")
g
```




```
jpeg("Occ_years_no_model.png", quality = 100, width = 15, height = 8 ,
     units = "cm", res= 300)
ready_conversion(g)
dev.off()

## png
##   2
```

A significantly reduced number of data points is retreived when excluding the modelling data.

CC37 Checking entreis in the loss quantity column

```
df_lq <- summary(!PD$loss_quantity == "")
df_lq <- df_lq %>%
  as.array() %>%
  as.data.frame()

df_lq <- df_lq[2:3,]
df_lq$Freq <- as.numeric(df_lq$Freq)
df_lq$Freq <- (df_lq$Freq/nrow(PD))
df_lq

##    Var1        Freq
## 2 FALSE 0.999297639
## 3  TRUE 0.000702361
```

The column only contains a small number of loss_quantity entries. Threrefore, the coumn should not be taken into account for the further analysis since they cannot be used for weighting.

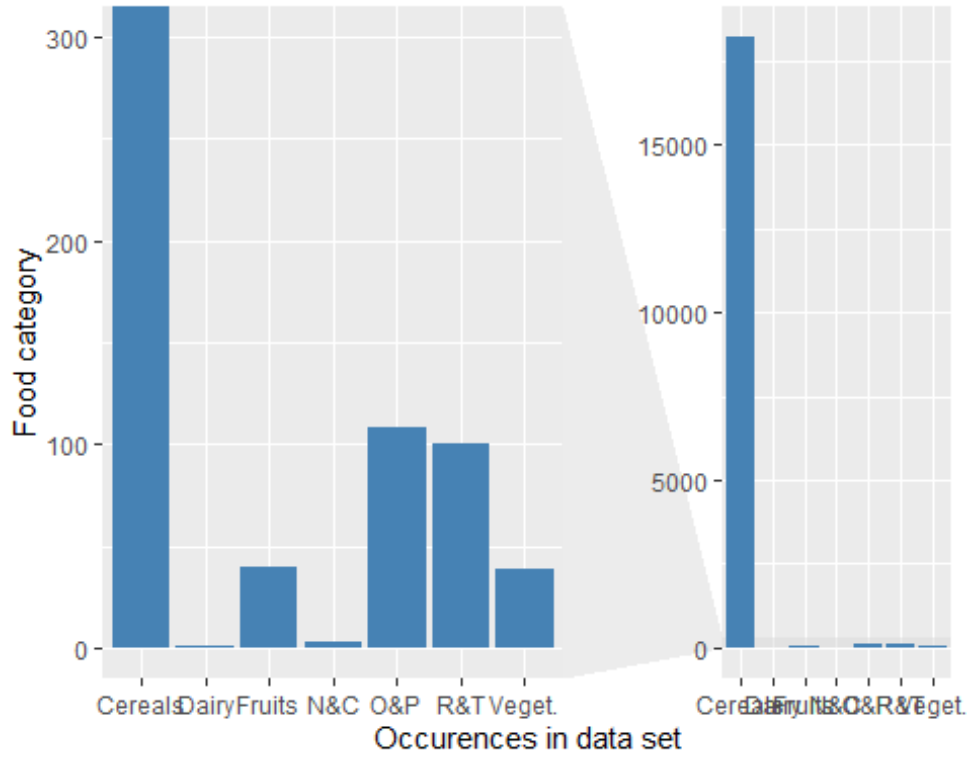
CC38 Checking entries in the region column

```
df_lq <- summary(!PD$region == "")
df_lq <- df_lq %>%
  as.array() %>%
  as.data.frame()

df_lq <- df_lq[2:3,]
df_lq$Freq <- as.numeric(df_lq$Freq)
df_lq$Freq <- (df_lq$Freq/nrow(PD))
df_lq

##    Var1       Freq
## 2 FALSE 0.98978875
## 3  TRUE 0.01021125
```

The column only contains a small number of 'region'-entries. Therefore, the column should not be taken into account for the further analysis since they cannot be used for weighting.

CC39 Data occurences across food categories

```
g <- ggplot(data = PD) +
  aes(x = food_category) +
  geom_bar(stat="count", fill = "steelblue") +
  facet_zoom(ylim = c(0, 300)) +
  ylab("Food category") +
  xlab("Occurences in data set")
g
```




```
jpeg("Occ_FC.png", quality = 100, width = 15, height = 8 ,
     units = "cm", res= 300)
  ready_conversion(g) + scale_x_discrete(guide = guide_axis(n.dodge=3)
```

```
)
dev.off()

## png
##   2
```

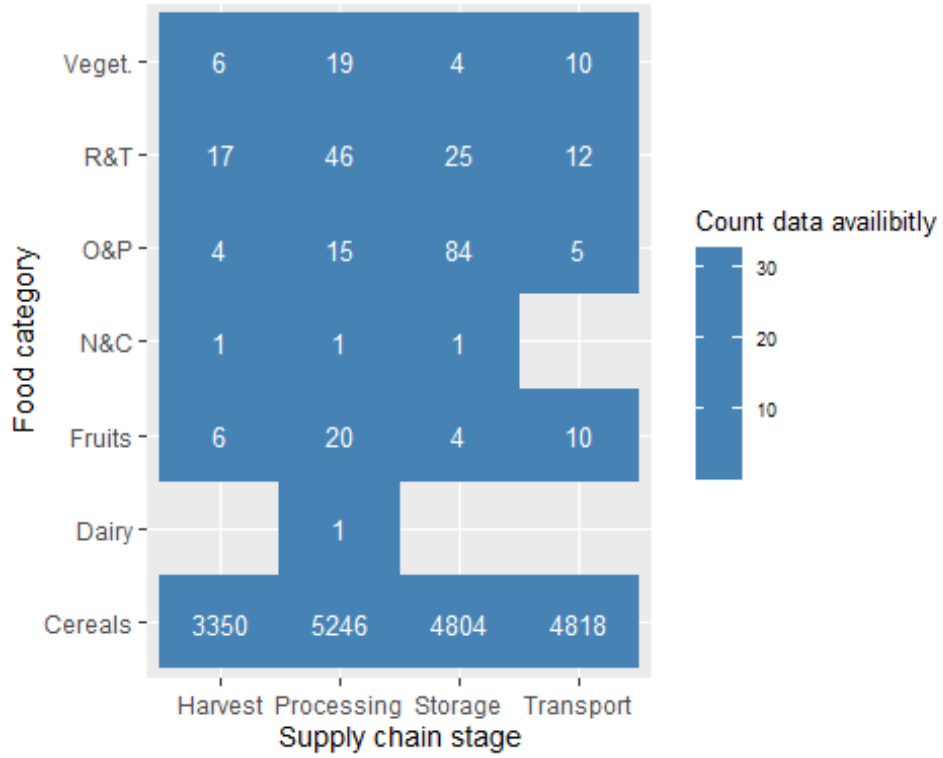
CC40 Ordering the SC stages

```
PD$food_supply_stage <- factor(PD$food_supply_stage, levels=c("Harvest
",
  "Processing", "Storage", "Transport"))
```

CC41 Data availability across food categories and SC stages

```
library(plotly)

##
## Attache Paket: 'plotly'

## Das folgende Objekt ist maskiert 'package:ggplot2':
##
##     last_plot

## Das folgende Objekt ist maskiert 'package:stats':
##
##     filter

## Das folgende Objekt ist maskiert 'package:graphics':
##
##     layout

library(stringr)
library(gghighlight)

cache1 <- PD %>%
  group_by(food_supply_stage, food_category) %>%
  summarise_each(funs(mean))

cache1$food_supply_stage <- as.factor(cache1$food_supply_stage)

x <- PD %>%
  count(food_supply_stage, food_category)

cache1 <- merge(cache1,x)
cache1_copy_for_regr <- cache1

cache2 <- ggplot(cache1, aes(x=cache1$food_supply_stage,
                             drop = F, y=cache1$food_category, drop =
F,
                             fill=cache1$loss_percentage, showscale =
F,
                             cex(2.5))) +
  geom_tile() +
  scale_fill_gradient(low="steelblue", high="steelblue",
                      name="Count data availibitly") +
                      xlab("Supply chain stage") +
```

```
                      ylab("Food category") +
                      scale_x_discrete(drop=FALSE) +
                      scale_y_discrete(drop=FALSE)

cache3 <- cache2 + theme(axis.text = element_text(size = 9)) +
  geom_text(aes(label = n), size = 3.5, color = "white") +
  theme(axis.title = element_text(size = 10.5)) +
  theme(legend.title = element_text(size = 10)) +
  theme(legend.text = element_text(size = 7))
cache3
```



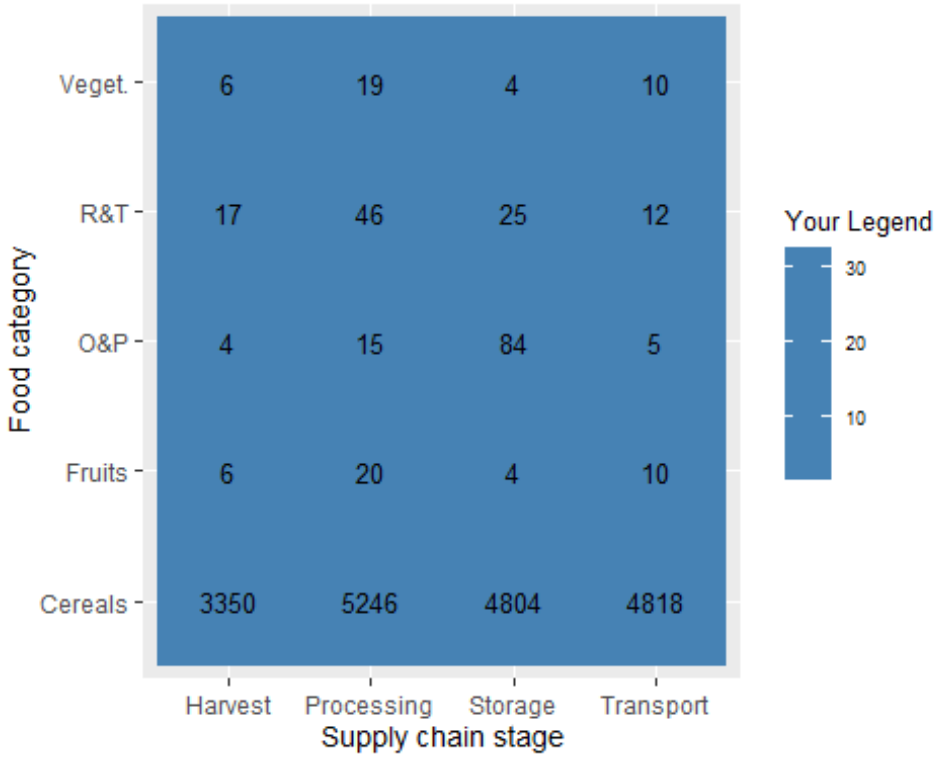

```
jpeg("SC_FC_Occ_all_prior.png", quality = 100, width = 15, height = 8
,
     units = "cm", res= 300)
ready_conversion(cache2) +
  geom_text(aes(label = n), size = 2.9, color = "white")
dev.off()

## png
##   2
```

CC42 Deleting the food categories dairy, eggs, meat and nuts & cacao beans due to insufficient data

```
PD = subset(PD, !(food_category %in% c('Dairy', 'Eggs', 'Meat', 'N&C')
),
            drop=TRUE)
```

CC43 Data availability across food categories and SC stages without the deleted food categoriees

```
cache1 <- PD %>%
  group_by(food_supply_stage, food_category) %>%
  summarise_each(funs(mean))

cache1$food_supply_stage <- as.factor(cache1$food_supply_stage)

x <- PD %>%
  count(food_supply_stage, food_category)

cache1 <- merge(cache1,x)
cache1_copy_for_regr <- cache1

cache2 <- ggplot(cache1, aes(x=cache1$food_supply_stage, drop = F,
                             y=cache1$food_category, drop = F,
                             fill=cache1$loss_percentage, showscale =
F,
                             cex(2.5))) +
                             geom_tile() +
                             scale_fill_gradient(low="steelblue",
                                                 high="steelblue",
                                                 name="Your Legend") +
                             xlab("Supply chain stage") +
                             ylab("Food category") +
                             scale_x_discrete(drop=TRUE) +
                             scale_y_discrete(drop=TRUE)

cache3 <- cache2 + theme(axis.text = element_text(size = 9)) +
                   geom_text(aes(label = n), size = 3.5) +
                   theme(axis.title = element_text(size = 10.5)) +
                   theme(legend.title = element_text(size = 10)) +
                   theme(legend.text = element_text(size = 7))
cache3
```

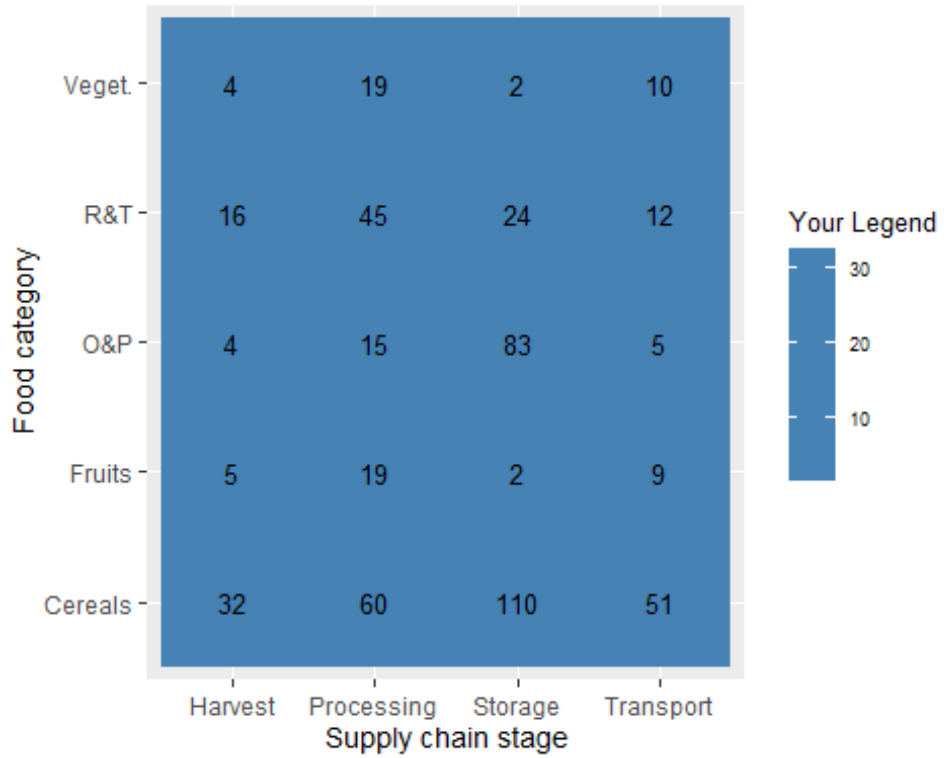



```
jpeg("SC_FC_Occ_all_after.png", quality = 100, width = 15, height = 8
,
     units = "cm", res= 300)
ready_conversion(cache2) +
  geom_text(aes(label = n), size = 2.5)
dev.off()

## png
##   2
```

CC44 Data availability across food categories and SC stages, no modelled data

```
cache1 <- PD %>%
  filter(method_data_collection != "Modelled Estimates") %>%
  group_by(food_supply_stage, food_category) %>%
  summarise_each(funs(mean))

cache1$food_supply_stage <- as.factor(cache1$food_supply_stage)

x <- PD %>%
  filter(method_data_collection != "Modelled Estimates")  %>%
  count(food_supply_stage, food_category)

cache1 <- merge(cache1,x)
cache1_copy_for_regr <- cache1

cache2 <- ggplot(cache1, aes(x=cache1$food_supply_stage, drop = F,
                             y=cache1$food_category, drop = F,
                             fill=cache1$loss_percentage, showscale =
F,
                             cex(2.5))) +
```

```
                              geom_tile() +
                              scale_fill_gradient(low="steelblue",
                                                  high="steelblue",
                                                  name="Your Legend") +
  xlab("Supply chain stage") +
  ylab("Food category") +
  scale_x_discrete(drop=TRUE) +
  scale_y_discrete(drop=TRUE)+
  theme(axis.text = element_text(size = 9)) +
  geom_text(aes(label = n), size = 3.5) +
  theme(axis.title = element_text(size = 10.5)) +
  theme(legend.title = element_text(size = 10)) +
  theme(legend.text = element_text(size = 7))

cache2
```

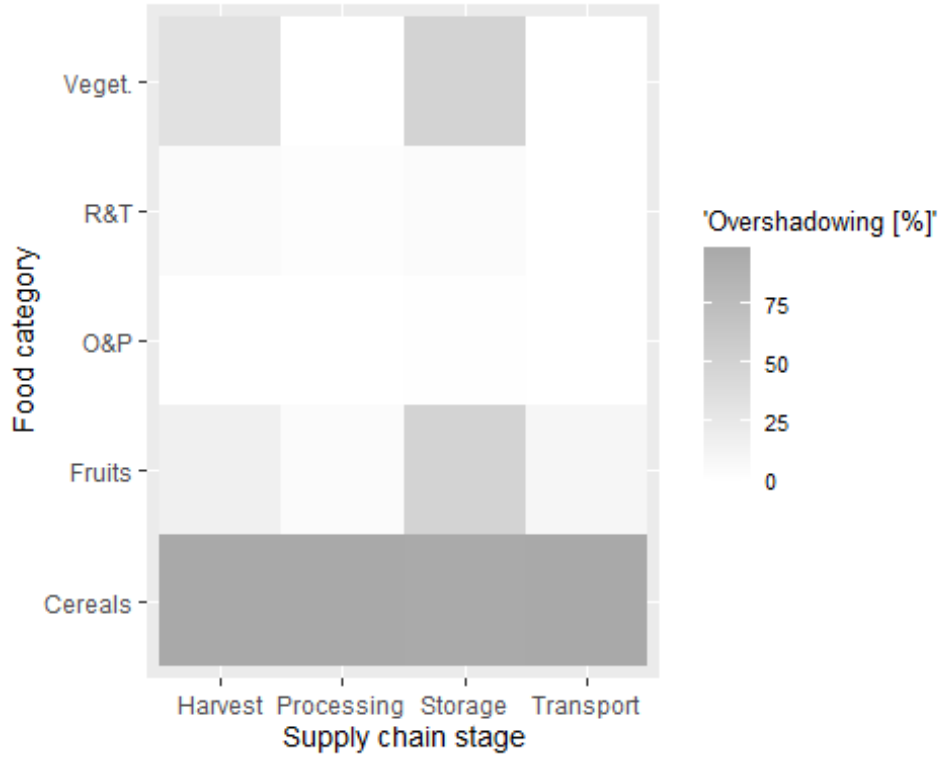



```
jpeg("years_points_Data_AV_heatmap2.png", quality = 100, width = 15, h
eight = 8,
     units = "cm", res= 300)
cache2
dev.off()

## png
##   2
```

CC45 Measuring the overshodowing of data

```
sub_all <- select(PD, food_supply_stage, food_category)
sub_all$occ = 0

cache0 <- sub_all %>%
```

```
  group_by(food_supply_stage, food_category) %>%
  summarize(occ = n())

## `summarise()` has grouped output by 'food_supply_stage'. You can ov
erride using
## the `.groups` argument.

cache0$food_supply_stage <- as.factor(cache0$food_supply_stage)

sub_no_mod <- PD %>% filter(method_data_collection != "Modelled Estima
tes") %>%
  select(food_supply_stage, food_category)
sub_no_mod$occ_no_mod = 0

cache1 <- sub_no_mod %>%
  group_by(food_supply_stage, food_category) %>%
  summarize(occ_no_mod = n())

## `summarise()` has grouped output by 'food_supply_stage'. You can ov
erride using
## the `.groups` argument.

cache1$food_supply_stage <- as.factor(cache1$food_supply_stage)

merged <- merge(cache0,cache1, by = c("food_supply_stage", "food_categ
ory"),
                all = TRUE)

for(i in 1:length(merged$food_supply_stage)){
  if(is.na(merged$occ_no_mod[i])){
    merged$occ_no_mod[i] = 0
  }
}

merged$shares = round((1-(merged$occ_no_mod/merged$occ)), 3)*100

cache2 <- ggplot(merged, aes(x=merged$food_supply_stage, drop = F,
                             y=merged$food_category, drop = F,
                             fill=merged$shares, showscale = F, cex(2.
5))) +
  geom_tile() +
  scale_fill_gradient(low="white", high="darkgrey", name="'Overshadowi
ng [%]'") +
  xlab("Supply chain stage") +
  ylab("Food category") +
  scale_x_discrete(drop=TRUE) +
  scale_y_discrete(drop=TRUE)+
  theme(axis.text = element_text(size = 9)) +
  theme(axis.title = element_text(size = 10.5)) +
  theme(legend.title = element_text(size = 10)) +
  theme(legend.text = element_text(size = 8))

cache2
```

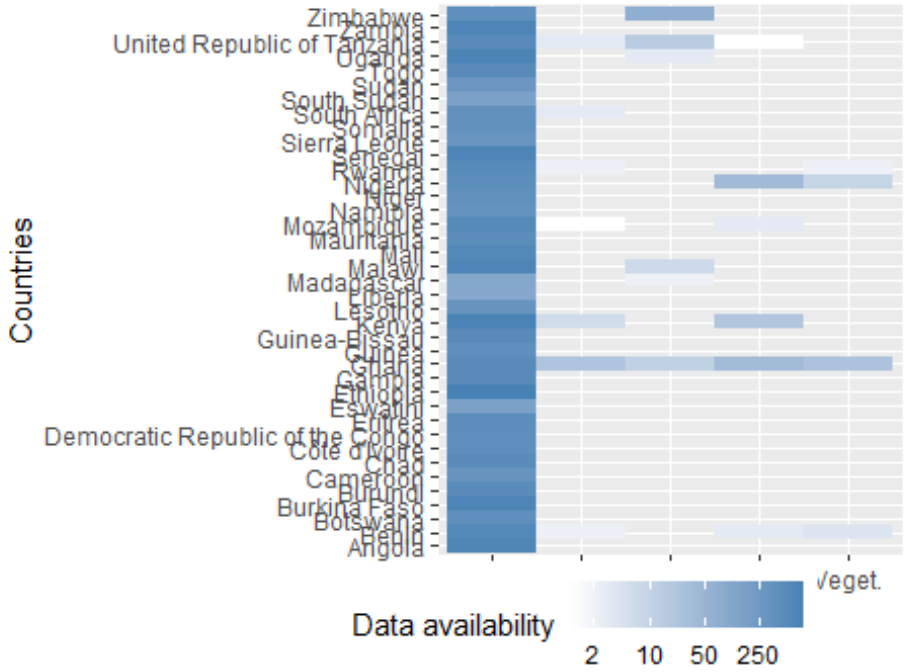



```
jpeg("overshadowing.png", quality = 100, width = 15, height = 8 ,
     units = "cm", res= 300)
cache2 +
  geom_text(aes(label = shares), size = 2.9)
dev.off()

## png
##   2
```

CC46 Heat map of available data across countries and food categories

```
library(gplots)

## Warning: Paket 'gplots' wurde unter R Version 4.2.3 erstellt

##
## Attache Paket: 'gplots'

## Das folgende Objekt ist maskiert 'package:stats':
##
##     lowess

PD$dummy <- rep(1, length(PD$m49_code))

u <- data.frame(PD$country, PD$dummy)

cache1 <- PD[c("country", "dummy", "food_category")] %>%
  group_by(country, food_category) %>%
  summarize(dummy = sum(dummy))
```

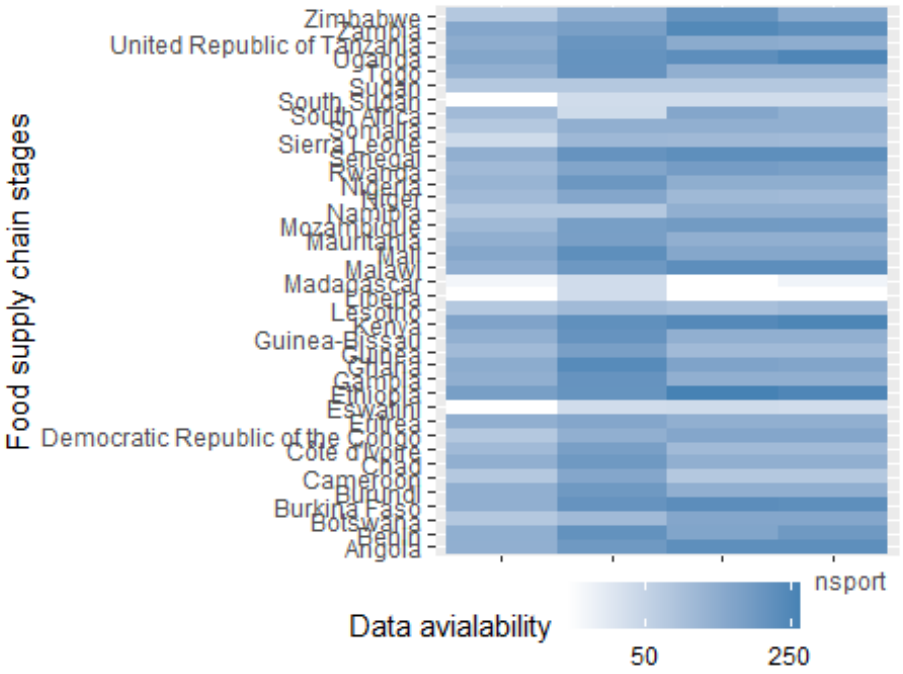
```
## `summarise()` has grouped output by 'country'. You can override usi
ng the
## `.groups` argument.

cache2 <- ggplot(cache1, aes(x=cache1$country,
                             y=cache1$food_category,
                             fill=cache1$dummy)) +
                             geom_tile() +
  scale_fill_gradient(low="white", high="steelblue",
                      name="Data availability")
cache2 <- cache2 + coord_flip()

my_breaks = c(2, 10, 50, 250)
cache2 <- cache2 + scale_fill_gradient(name = "Data availability",
                                       trans = "log", low="white",
                                       high="steelblue", breaks = my_b
reaks,
                                       labels = my_breaks) +
  ylab("Food categories") +
  xlab("Countries") +
  theme(legend.position=c(.3, -.125),
        plot.margin=grid::unit(c(0.5,0.5,1.5,0.5), "cm"),
        legend.direction ="horizontal")

## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scal
e.

cache2
```



```
jpeg("Data_AV_heatmap_countries_FC.png", quality = 100, width = 15, he
ight = 17,
     units = "cm", res = 300)
ready_conversion(cache2)
dev.off()

## png
##   2
```

CC47 Heat map of available data of countries and supply chain stages (Appendix)

```
# Part of the appendix due to lack of relevance
PD$dummy <- rep(1, length(PD$m49_code))

u <- data.frame(PD$country, PD$dummy)

cache1 <- PD[c("country", "dummy", "food_supply_stage")] %>%
  group_by(country, food_supply_stage) %>%
  summarize(dummy = sum(dummy))

## `summarise()` has grouped output by 'country'. You can override usi
ng the
## `.groups` argument.

cache2 <- ggplot(cache1, aes(x=cache1$country,
                             y=cache1$food_supply_stage,
                             fill=cache1$dummy)) +
                             geom_tile() +
  scale_fill_gradient(low="white", high="steelblue",
                      name="Data avialability")
cache2 <- cache2 + coord_flip()

my_breaks = c(2, 10, 50, 250)
cache2 <- cache2 + scale_fill_gradient(name = "Data avialability",
                                       trans = "log", low="white",
                                       high="steelblue", breaks = my_b
reaks,
                                       labels = my_breaks) + ylab("Cou
ntries") +
  xlab("Food supply chain stages") +
  theme(legend.position=c(.3, -.125),
        plot.margin=grid::unit(c(0.5,0.5,1.5,0.5), "cm"),
        legend.direction ="horizontal")

## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scal
e.

cache2
```




```
jpeg("Data_AV_heatmap_countries_SC_stages.png", quality = 100, width =
15,
     height = 17, units = "cm", res = 300)
ready_conversion(cache2)
dev.off()

## png
##   2
```

CC48 Creating a word cloud for the cause of loss column

```
library(wordcloud)

## Lade nötiges Paket: RColorBrewer

##
## Attache Paket: 'wordcloud'

## Das folgende Objekt ist maskiert 'package:gplots':
##
##     textplot

library(RColorBrewer)
library(wordcloud2)
library(tm)

## Lade nötiges Paket: NLP

##
## Attache Paket: 'NLP'
```

```
## Das folgende Objekt ist maskiert 'package:ggplot2':
##
##     annotate

# Overview of the availability of cause of loss data within the scope
of the
# thesis.
PD_with_comments_within_scope <- subset(PD, cause_of_loss != "")
# Number of entries in the cause of loss column in the inner scope of
the
# thesis.
nrow(PD_with_comments_within_scope)

## [1] 248

# Comparison to the overall number of entires in the defined scope of
# the thesis.
nrow(PD_with_comments_within_scope)/nrow(PD)

## [1] 0.01340178

# For the sake of fast computation only data points to no empty data i
n the
# column "cause of loss" will be included in this sub-data set.
PD_with_comments <- subset(PD, cause_of_loss != "")

# For the word cloud exactly the subset was taken that was later used
for the
# cause of loss analysis

# Create a vector containing only the text
text <- PD_with_comments$cause_of_loss
# Create a corpus
docs <- Corpus(VectorSource(text))

docs <- docs %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeWords, stopwords("english"))

dtm <- TermDocumentMatrix(docs)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix),decreasing=TRUE)
df <- data.frame(word = names(words),freq=words)

# Seed for reproducibility
set.seed(1234)

wordcloud <- wordcloud(words = df$word, freq = df$freq, min.freq = 1,
          max.words=80, random.order=FALSE, rot.per=0.35)
```


```
jpeg("wordcloud.png", quality = 100, width = 15, height = 11, units =
"cm",
     res = 300)
wordcloud

## NULL

dev.off()

## png
##   2

# Saving the word cloud as a png-file did unfortunately not work out.
# Picture had to be directly copied from RStudio
# All code for word cloud taken from: https://towardsdatascience.com/c
reate-a-word-cloud-with-r-bde3e7422e8a
```

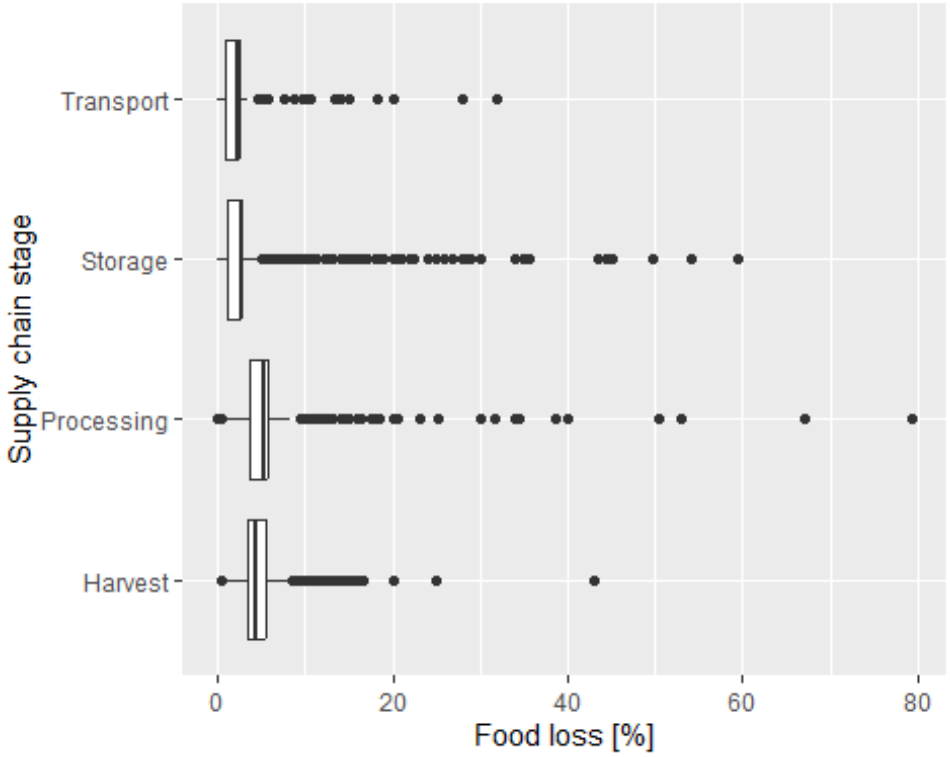
CC49 Overview of entries in the columm "cause of Loss":

```
overall_entries <- length(PD_orig.$cause_of_loss)
df_CoL_entries <- filter(PD_orig., cause_of_loss == "")
no_CoL_entries <- length(df_CoL_entries$cause_of_loss)
# Necessary for computing, since it wasn't saved as integer before
overall_entries = as.integer(overall_entries)
no_CoL_entries = as.integer(no_CoL_entries)
round(overall_entries/no_CoL_entries,3)

## [1] 1.022
```

Only about 1.03% of the data point contained with information in the cause of loss column.

CC50 Overview of food losses by supply chain stages:

```
g <- ggplot(PD, aes(x = loss_percentage , y = food_supply_stage))+
  geom_boxplot()+
  xlab("Food loss [%]")+
  ylab("Supply chain stage")
g
```



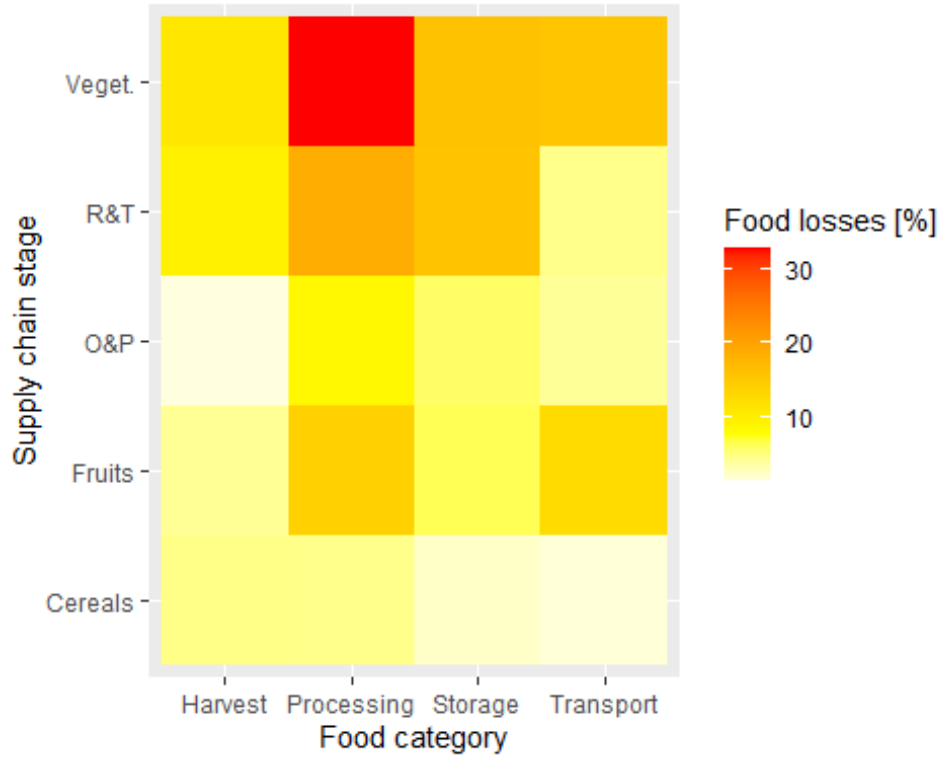

```
jpeg("boxplot_SC_losses.png", quality = 100, width = 1000, height = 800)
ready_conversion(g)
dev.off()

## png
##   2
```

For an overview of the food losses across food categories it is important to compute the overall food loss adding up the food losses occurring at all SC stages.

CC51 Hotspot analysis I: creating a heat map across supply chain stages and food categories

```
cache1.0 <- PD %>%
  group_by(food_supply_stage, food_category) %>%
  summarise_each(funs(mean))

x = PD %>%
  count(food_supply_stage, food_category)

cache1.0 <- merge(cache1.0,x)
```

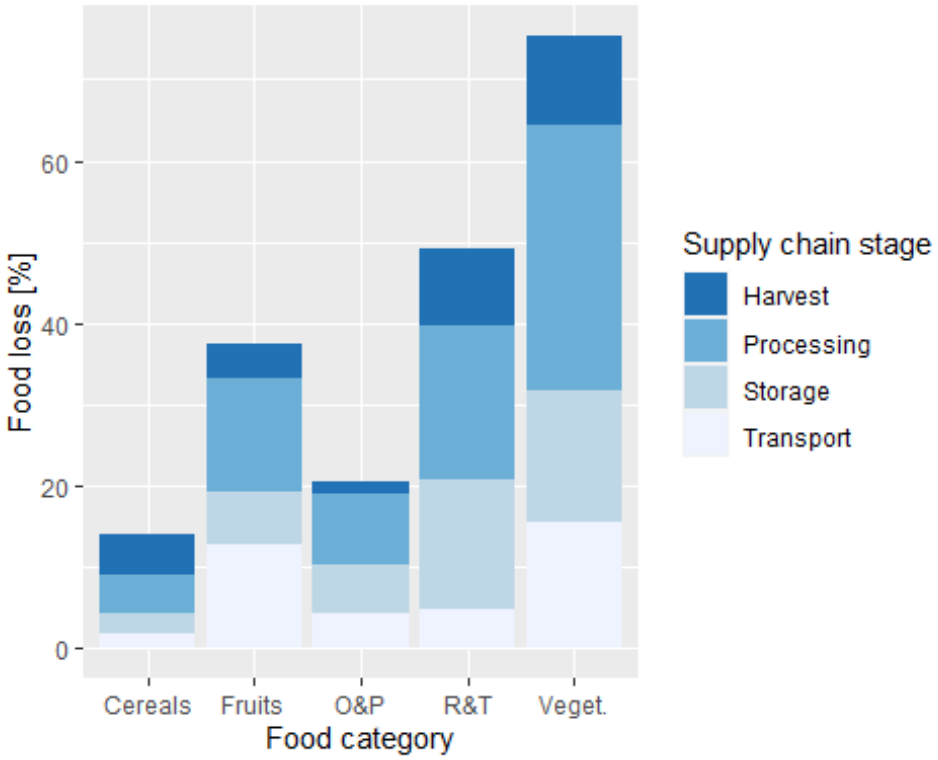
```
cache2 <- ggplot(cache1.0, aes(x=cache1.0$food_supply_stage,
                               y=cache1.0$food_category,
                               fill=cache1.0$loss_percentage)) +
  geom_tile() +
  scale_fill_gradientn(colours = c("red", "yellow", "lightyellow"),
                       values = c(1, 0.2, 0), name="Food losses [%]")
+
  xlab("Food category") +
  ylab("Supply chain stage")

cache2
```




```
cache3 <- cache2 +
  theme(axis.text = element_text(size = 18)) +
  geom_text(aes(label = round(loss_percentage, 1)), size = 8) +
  theme(axis.title = element_text(size = 24)) +
  theme(legend.title = element_text(size = 18)) +
  theme(legend.text = element_text(size = 18)) +
  geom_text(aes(label = round(loss_percentage, 1)))

jpeg("FL_data_heatmap_red_white.png", quality = 100, width = 15, heigh
t = 12,
     units = "cm", res = 300)
ready_conversion(cache2) +
  geom_text(aes(label = round(loss_percentage, 1)), size = 2.9)
dev.off()

## png
##   2
```

CC52 Hotspot analysis II: creating stacked bar chart

```
cache1.0$food_supply_stage <- as.character(cache1.0$food_supply_stage)

cache1_no_doubling <- subset(cache1.0, !food_supply_stage %in%
                               c("Distribution", "WSC", "Post-harvest"))

unique(cache1_no_doubling$food_supply_stage)

## [1] "Harvest"    "Processing" "Storage"    "Transport"

plot <- ggplot(cache1_no_doubling, aes(fill=food_supply_stage,
                                       y=loss_percentage, x=food_categ
ory)) +
  geom_bar(position="stack", stat = "identity") +
  xlab("Food category") +
  ylab("Food loss [%]") +
  scale_fill_brewer(palette = "Blues", name = "Supply chain stage", di
rection = -1)

plot
```




```
jpeg("FL_data_stacked_bar_chart.png", quality = 100, width = 15, heigh
t = 10,
     units = "cm", res = 300)
ready_conversion(plot) + theme(legend.key.size = unit(0.5, 'cm'))
dev.off()

## png
##   2
```

CC53 Discussion: Long Data and Wide Data and it's implications: Long Format Section (Appendix)

```
close_up_veget._processing <- subset(PD, food_category == "Veget." &
                                       food_supply_stage == "Processing"
)
close_up_veget._processing [,7:13]
```

```
## # A tibble: 19 × 7
##    loss_percentage loss_percentage_ori…¹ loss_…² activ…³ food_…⁴ tr
eat…⁵ cause…⁶
##              <dbl> <chr>                 <chr>   <chr>   <fct>   <c
hr>   <chr>
##  1           16.4  9.7 - 23%             ""      Sorting Proces… ""
"lack …
##  2           53    24                    ""      Gradin… Proces… ""
"Decay"
##  3           53    29                    ""      Gradin… Proces… ""
"Mecha…
##  4           15    15                    ""      Gradin… Proces… ""
"Mecha…
##  5            5.81 1.4 - 2.09            ""      Assemb… Proces… ""
""
##  6            5.81 2.67-5.46             ""      Gradin… Proces… ""
""
##  7           34    28                    ""      Gradin… Proces… ""
"Mecha…
##  8           38.7  2.8                   ""      Gradin… Proces… ""
"Decay"
##  9           67    54                    ""      Gradin… Proces… ""
"Mecha…
## 10           38.7  13.9                  ""      Gradin… Proces… ""
"Physi…
## 11           34    6                     ""      Gradin… Proces… ""
"Decay"
## 12           38.7  22                    ""      Gradin… Proces… ""
"Mecha…
## 13           50.5  17                    ""      Gradin… Proces… ""
"Decay"
## 14           13.1  12.9 - 13.3%          ""      Sorting Proces… ""
"lack …
## 15           67    13                    ""      Gradin… Proces… ""
"Decay"
## 16           50.5  33.5                  ""      Gradin… Proces… ""
"Mecha…
## 17           25.1  25.1                  ""      Gradin… Proces… ""
"Physi…
## 18            8    2                     ""      Gradin… Proces… ""
"Mecha…
## 19            8    6                     ""      Gradin… Proces… ""
"Decay"
## # … with abbreviated variable names ¹loss_percentage_original, ²los
s_quantity,
## #   ³activity, ⁴food_supply_stage, ⁵treatment, ⁶cause_of_loss
```

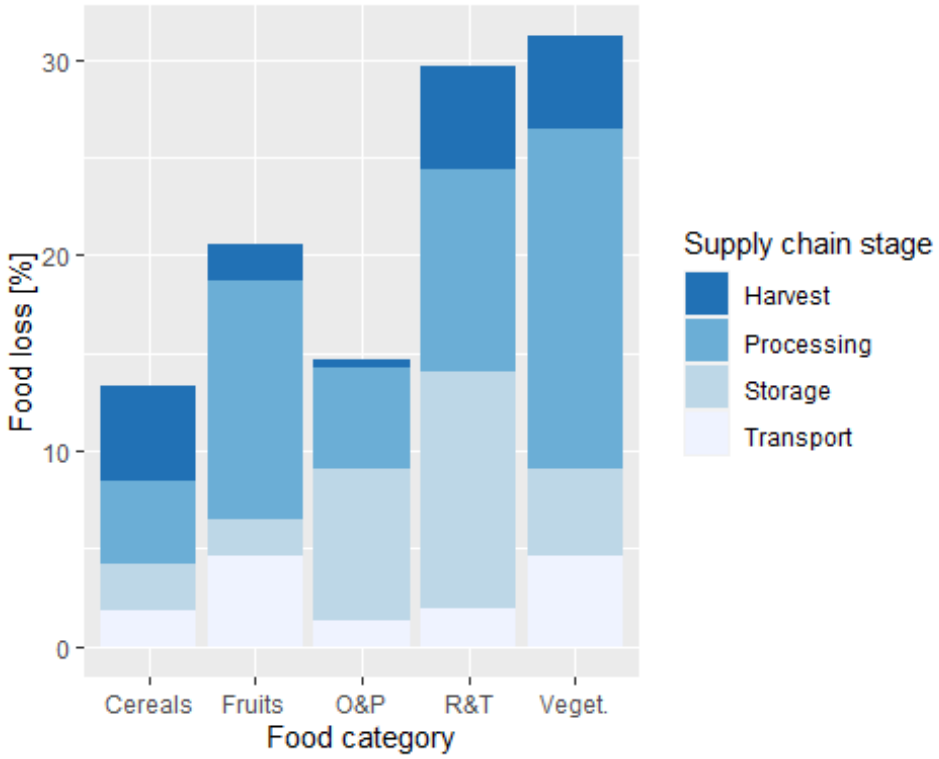
CC54 Long format data to wide format data (Appendix)

```
library(tidyr)
# Simpley selecting of columns
PD_subset <- select(PD, c("m49_code", "commodity", "year", "food_suppl
y_stage", "loss_percentage", "customs", "infrastructure", "internation
al_shipments", "logistics_competence", "tracking_tracing", "timeliness
", "food_category", "food_supply_stage"))

# Certain information has to be aggregated for each supply chain stage
PD_subset <- PD_subset %>% group_by(m49_code, commodity, year,
                                    food_supply_stage, food_category)
%>%
  summarise(loss_percentage=
              mean(loss_percentage),
              customs=mean(customs),
              infrastructure= mean(infrastructure),
              international_shipments=mean(international_shipments),
              logistics_competence=mean(logistics_competence),
              tracking_tracing=mean(tracking_tracing),
              timeliness=mean(timeliness),
              .groups = 'drop') %>%
  as.data.frame()

# LPI data actually not necessary for identify, but they shall be kept
.
PD_wide_format <- pivot_wider(
  PD_subset,
  id_cols = c(m49_code, commodity, year, customs, infrastructure,
              international_shipments, logistics_competence, tracking_
tracing,
              timeliness, food_category),
  names_from = food_supply_stage,
  values_from = loss_percentage,
  names_prefix = "FL.in.percent_"
)
# FL.in.percent_ is a remnant of the countries with missing FLI or mis
sing
# LPI data
```

55 Creating a stacked bar plot for the data in wide format

```
PD_wide_format <- PD_wide_format %>%
  mutate_at(c(11:14), ~replace(.,is.na(.),0))
names(PD_wide_format) <- str_replace_all(names(PD_wide_format), c(" "
= "_"))

cache1_wf <- PD_wide_format %>%
  group_by(food_category) %>%
  summarise(n=n(),
              Harvest = mean(FL.in.percent_Harvest, na.rm = T),
              Storage = mean(FL.in.percent_Storage, na.rm = T),
              Transport = mean(FL.in.percent_Transport, na.rm = T),
              Processing = mean(FL.in.percent_Processing,
                                na.rm = T))
```

```
cache2_wf <- cache1_wf %>%
  gather(., supply_chain_stage, food_loss, 3:6, -n)

ggplot(cache2_wf, aes(fill=supply_chain_stage, y=food_loss, x=food_cat
egory)) +
  geom_bar(position="stack", stat = "identity") +
  xlab("Food category") +
  ylab("Supply chain stage") +
  xlab("Food category") +
  ylab("Food loss [%]") +
  scale_fill_brewer(palette = "Blues", name = "Supply chain stage",
                    direction = -1)
```



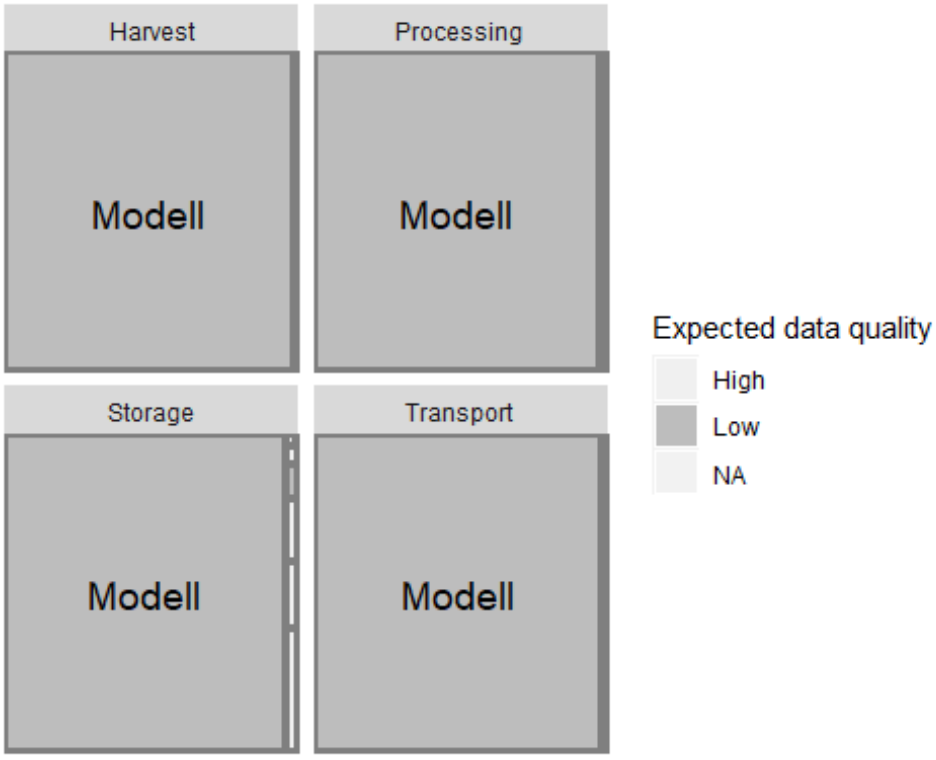
CC56 - Treemap data quality for supply chain stages, incl. facet wrapt

```
library(treemap)
library(treemapify)
library(ggplot2)

PD_v2 = PD

PD_v2$food_supply_stage <- as.character(PD_v2$food_supply_stage)

# Renaming of methods of data collection to decrease complexity on the
chart
PD_v2$method_data_collection <-
  recode_factor(PD_v2$method_data_collection, `Modelled Estimates` = "
Modell",
  `No Data Collection Specified` = "Not spec.",
  `FAO's annual Agriculture Production Questionnaires` = "FAO Quest.",
  `Literature Review` = "Literature",
```

```
  `Controlled Experiment` = "Experiment",
  `Case Study` = "CS", `Expert Opinion` = "Expert")

unique(PD_v2$method_data_collection)

## [1] Modell     NA         Not spec.  Survey     CS         Experime
nt Literature
## [8] Expert
## Levels: Modell Not spec. Literature Experiment CS Expert NA Survey

PD_copy2 <- PD_v2 %>%
  drop_na(c("infrastructure", "logistics_competence", "tracking_tracin
g",
            "food_supply_stage")) %>%
  count(food_supply_stage, method_data_collection)

PD_copy2 <- subset(PD_copy2, )

PD_copy2$q_score = NA

PD_copy2$method_data_collection <- as.character(PD_copy2$method_data_c
ollection)

PD_copy2 = subset(PD_copy2, (food_supply_stage %in% c("Farm", "Transpo
rt",
                                                      "Storage", "Harv
est",
                                                      "Whole supply ch
ain",
                                                      "Processing",
                                                      "Post-harvest",
                                                      "Distribution"))
,
                  drop=FALSE)

for(j in 1:length(PD_copy2$n)) {

if(PD_copy2$method_data_collection[j] == "Modell") {PD_copy2$q_score[j
] = 2}
if(PD_copy2$method_data_collection[j] == "Not spec.") {PD_copy2$q_scor
e[j] = 2}
if(PD_copy2$method_data_collection[j] == "FAO Quest.") {PD_copy2$q_sco
re[j] = 1}
if(PD_copy2$method_data_collection[j] == "") {PD_copy2$q_score[j] = 1}
if(PD_copy2$method_data_collection[j] == "Literature") {PD_copy2$q_sco
re[j] = 1}
if(PD_copy2$method_data_collection[j] == "Experiment") {PD_copy2$q_sco
re[j] = 1}
if(PD_copy2$method_data_collection[j] == "CS") {PD_copy2$q_score[j] =
1}
if(PD_copy2$method_data_collection[j] == "Survey") {PD_copy2$q_score[j
] = 1}
if(PD_copy2$method_data_collection[j] == "Census") {PD_copy2$q_score[j
] = 1}
```

```
if(PD_copy2$method_data_collection[j] == "Expert") {PD_copy2$q_score[j
] = 1}
}

cache1 <- PD_copy2

cache1$q_score <- as.factor(cache1$q_score)
cache1$food_supply_stage <- as.character(cache1$food_supply_stage)
cache1$method_data_collection <- as.character(cache1$method_data_colle
ction)
cache1$n <- as.character(cache1$n)
cache1$q_score <- as.character(cache1$q_score)

# Start of building the model and return the outputs

new_row1 = c("dummy", "dummy", "0", "1")
new_row2 = c("dummy", "dummy", "0", "2")

cache1 <- rbind(cache1, new_row1, new_row2)

cache1$food_supply_stage <- as.factor(cache1$food_supply_stage)
cache1$method_data_collection <- as.factor(cache1$method_data_collecti
on)
cache1$n <- as.integer(cache1$n)
cache1$q_score <- as.factor(cache1$q_score)

cache1 %>% drop_na(method_data_collection)

## # A tibble: 34 × 4
##    food_supply_stage method_data_collection     n q_score
##    <fct>             <fct>                  <int> <fct>
##  1 Harvest           Modell                  3280 2
##  2 Harvest           Not spec.                  9 2
##  3 Harvest           Literature                 2 1
##  4 Harvest           Experiment                 6 1
##  5 Harvest           CS                         6 1
##  6 Harvest           Expert                     5 1
##  7 Harvest           NA                         3 <NA>
##  8 Harvest           Survey                    30 1
##  9 Processing        Modell                  5104 2
## 10 Processing        Not spec.                 29 2
## # … with 24 more rows

cache1 %>% drop_na(q_score)

## # A tibble: 30 × 4
##    food_supply_stage method_data_collection     n q_score
##    <fct>             <fct>                  <int> <fct>
##  1 Harvest           Modell                  3280 2
##  2 Harvest           Not spec.                  9 2
##  3 Harvest           Literature                 2 1
##  4 Harvest           Experiment                 6 1
##  5 Harvest           CS                         6 1
##  6 Harvest           Expert                     5 1
```

```
##  7 Harvest           Survey                    30 1
##  8 Processing        Modell                  5104 2
##  9 Processing        Not spec.                 29 2
## 10 Processing        Literature                 6 1
## # … with 20 more rows

cache1$q_score <- as.factor(cache1$q_score)

cache1 <- filter(cache1, food_supply_stage != "dummy")

show <- ggplot(cache1, aes(area = n, fill = q_score,
                           label = method_data_collection,
                           subgroup = method_data_collection)) +
  facet_wrap(~ food_supply_stage) +
  geom_treemap() +
  geom_treemap_text(colour = "black", place = "centre",
                    grow = FALSE, alpha = 1, size = 14) +
  geom_treemap_subgroup_border() +
  scale_fill_brewer(palette = c("Greys"), direction = 1,
                    labels = c("High", "Low")) +
  labs(fill = "Expected data quality")

print(show)
```



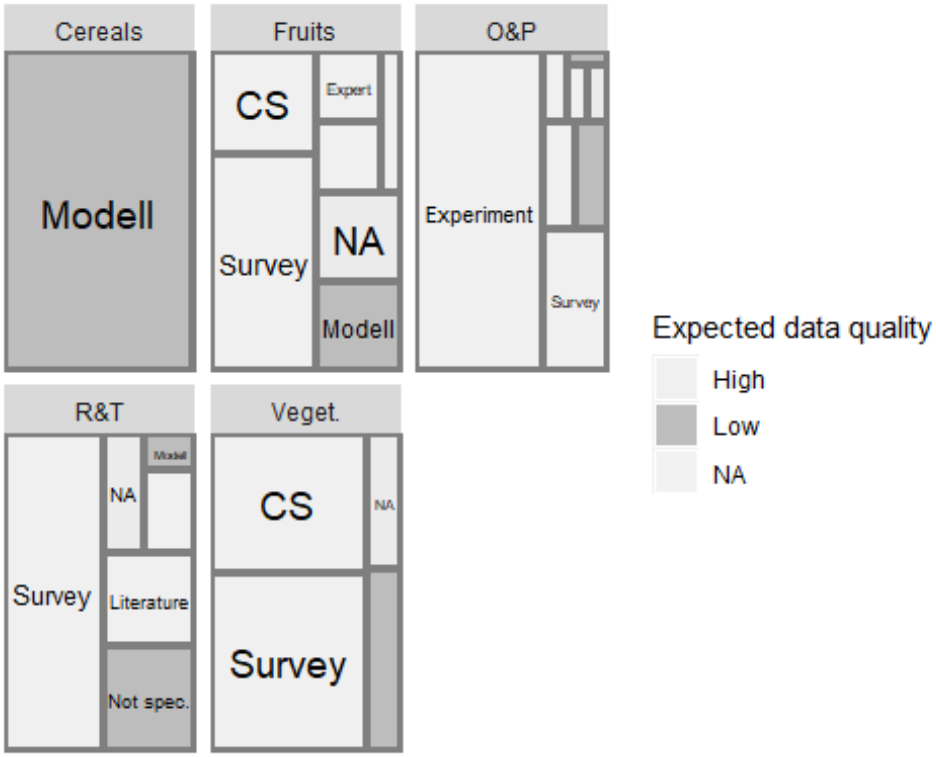

```
jpeg("Tree_part1.png", quality = 100, width = 15, height = 6,
     units = "cm", res = 300)
ready_conversion(show) + theme(legend.key.size = unit(0.5, 'cm'))
dev.off()

## png
##   2
```

CC57 Treemap data quality for food categories, incl. facet wrap

```
# Renaming of methods of data collection: Not spec = "No Data Collecti
on
# Specified", FAO quest = "FAO's ann. Questionaires", Literature = "
# Literature Review", Experiment = "Controlled Experiment", CS = "Case
Study",
# Expert = "Expert Opinion"

library(treemap)

PD_v2 = PD

PD_v2$food_supply_stage <- as.character(PD_v2$food_supply_stage)

PD_v2$method_data_collection <-
  recode_factor(PD_v2$method_data_collection,
                `Modelled Estimates` = "Modell",
                `No Data Collection Specified` =
                  "Not spec.", `FAO's annual Agriculture Production Qu
estionnaires` =
                  "FAO Quest.", `Literature Review` = "Literature", `C
ontrolled Experiment` =
                  "Experiment", `Case Study` = "CS", `Expert Opinion`
= "Expert")

unique(PD_v2$method_data_collection)

## [1] Modell     NA         Not spec.  Survey     CS         Experime
nt Literature
## [8] Expert
## Levels: Modell Not spec. Literature Experiment CS Expert NA Survey

PD_copy2 <- PD_v2 %>%
  drop_na(c("infrastructure", "logistics_competence", "tracking_tracin
g",
            "food_category", "method_data_collection")) %>%
  count(food_category, method_data_collection)

PD_copy2$q_score = NA

PD_copy2$method_data_collection <- as.character(PD_copy2$method_data_c
ollection)


for(j in 1:length(PD_copy2$n)) {
  if(PD_copy2$method_data_collection[j] == "Modell") {PD_copy2$q_score
[j] = 2}
  if(PD_copy2$method_data_collection[j] == "Not spec.") {PD_copy2$q_sc
ore[j] = 2}
  if(PD_copy2$method_data_collection[j] == "FAO Quest.") {PD_copy2$q_s
core[j] = 1}
  if(PD_copy2$method_data_collection[j] == "") {PD_copy2$q_score[j] =
1}
```

```
  if(PD_copy2$method_data_collection[j] == "Literature") {PD_copy2$q_s
core[j] = 1}
  if(PD_copy2$method_data_collection[j] == "Experiment") {PD_copy2$q_s
core[j] = 1}
  if(PD_copy2$method_data_collection[j] == "CS") {PD_copy2$q_score[j]
= 1}
  if(PD_copy2$method_data_collection[j] == "Survey") {PD_copy2$q_score
[j] = 1}
  if(PD_copy2$method_data_collection[j] == "Census") {PD_copy2$q_score
[j] = 1}
  if(PD_copy2$method_data_collection[j] == "Expert") {PD_copy2$q_score
[j] = 1}
}

cache1 <- PD_copy2

cache1$q_score <- as.factor(cache1$q_score)
cache1$food_category <- as.character(cache1$food_category)
cache1$method_data_collection <- as.character(cache1$method_data_colle
ction)
cache1$n <- as.character(cache1$n)
cache1$q_score <- as.character(cache1$q_score)

# Start of building the model and return the outputs

new_row1 = c("dummy", "dummy", "0", "1")
new_row2 = c("dummy", "dummy", "0", "2")

cache1 <- rbind(cache1, new_row1, new_row2)

cache1$food_category <- as.factor(cache1$food_category)
cache1$method_data_collection <- as.factor(cache1$method_data_collecti
on)
cache1$n <- as.integer(cache1$n)
cache1$q_score <- as.factor(cache1$q_score)

cache1 %>% drop_na(method_data_collection)

## # A tibble: 35 × 4
##    food_category method_data_collection     n q_score
##    <fct>         <fct>                  <int> <fct>
##  1 Cereals       Modell                 17671 2
##  2 Cereals       Not spec.                 69 2
##  3 Cereals       Literature                30 1
##  4 Cereals       Experiment                20 1
##  5 Cereals       CS                        34 1
##  6 Cereals       Expert                    12 1
##  7 Cereals       NA                         6 <NA>
##  8 Cereals       Survey                    81 1
##  9 Fruits        Modell                     5 2
## 10 Fruits        Literature                 2 1
## # … with 25 more rows

cache1 %>% drop_na(q_score)
```

```
## # A tibble: 30 × 4
##    food_category method_data_collection     n q_score
##    <fct>         <fct>                  <int> <fct>
##  1 Cereals       Modell                 17671 2
##  2 Cereals       Not spec.                 69 2
##  3 Cereals       Literature                30 1
##  4 Cereals       Experiment                20 1
##  5 Cereals       CS                        34 1
##  6 Cereals       Expert                    12 1
##  7 Cereals       Survey                    81 1
##  8 Fruits        Modell                     5 2
##  9 Fruits        Literature                 2 1
## 10 Fruits        Experiment                 3 1
## # … with 20 more rows

cache1$q_score <- as.factor(cache1$q_score)

cache1 <- filter(cache1, food_category != "dummy")

show <- ggplot(cache1, aes(area = n, fill = q_score,
                           label = method_data_collection,
                           subgroup = method_data_collection)) +
  facet_wrap(~ food_category) +
  geom_treemap() +
  geom_treemap_text(colour = "black", place = "centre",
                    grow = FALSE, alpha = 1, size = 14) +
  geom_treemap_subgroup_border() +
  scale_fill_brewer(palette = c("Greys"), direction = 1,
                    labels = c("High", "Low")) +
  labs(fill = "Expected data quality")
print(show)
```



```
jpeg("Tree_part2.png", quality = 100, width = 15, height = 6,
     units = "cm", res = 300)
ready_conversion(show) + theme(legend.key.size = unit(0.5, 'cm'))
dev.off()

## png
##   2
```

-4- Pairwise relationship between variables

Consideration of different development stages/logistics performance of countires

CC58 Scatterplot of LPI_score vs. food_loss

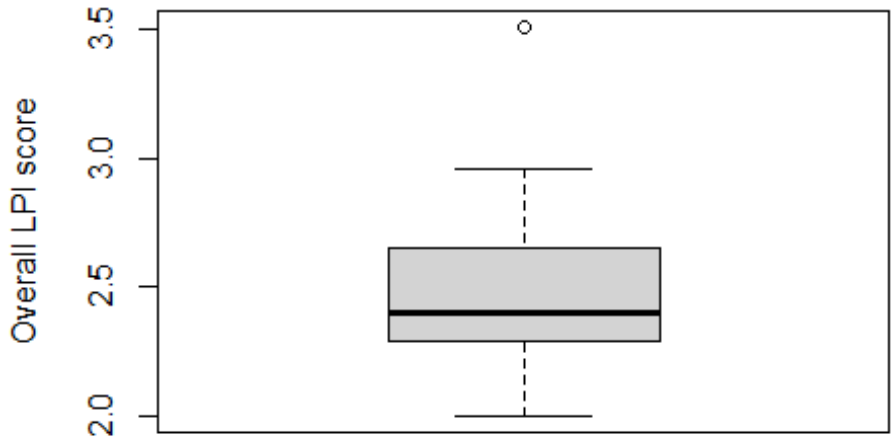
```
cache <- ggplot(PD, aes(x=LPI_score, y=loss_percentage)) +
  facet_wrap(~ food_supply_stage +
               food_category, ncol = 4) +
  xlab("LPI score") +
  ylab("Food loss [%]") +
    theme(strip.text = element_text(size = 7))

cache
```


```
jpeg("years_points_Data_heatmap_general.png", quality = 100, width = 1
5,
     height = 16, units = "cm", res = 300)
ready_conversion(cache) + geom_point(size=0.05)

## Warning: Removed 295 rows containing missing values (`geom_point()`
).

dev.off()
```

```
## png
##   2
```

CC59 Preparation of country segmentation into three classes

```
box_plot_LPI <- PD[c("country", "LPI_score")] %>%
  group_by(country) %>%
  summarize(LPI_score = mean(LPI_score))
box_plot_LPI

## # A tibble: 39 × 2
##    country                          LPI_score
##    <fct>                                <dbl>
##  1 Angola                                2.18
##  2 Benin                                 2.65
##  3 Botswana                              2.96
##  4 Burkina Faso                          2.63
##  5 Burundi                               2.22
##  6 Cameroon                              2.43
##  7 Chad                                  2.34
##  8 Côte d'Ivoire                         2.89
##  9 Democratic Republic of the Congo      2.33
## 10 Eritrea                               2.11
## # … with 29 more rows

boxplot(box_plot_LPI$LPI_score, ylab ="Overall LPI score")
```




```
# Note, there is one outlier, with the LPI Value of 3.5
box_plot_LPI <- box_plot_LPI[order(box_plot_LPI$LPI_score),]
# HDI could be tested here also
quantiles = quantile(box_plot_LPI$LPI_score, na.rm = T, probs = c(.333
,.666))
```

```
limitgroup1 = as.numeric(quantiles[1])
limitgroup2 = as.numeric(quantiles[2])

# Create empty column, named CPG, which means
# "Country Performance Group"
PD <- PD %>%
  add_column(CPG = NA)

PD <- PD %>%
  drop_na(LPI_score)

# Remove NAs fors!!!
for (i in 1:length(PD$LPI_score)) {
  if (PD$LPI_score[i] <= limitgroup1) {
    PD$CPG[i] = "Low performance countries"
  } else if (PD$LPI_score[i] <= limitgroup2) {
    PD$CPG[i] = "Medium performance countries"
  } else if (PD$LPI_score[i] <= 999) {
    PD$CPG[i] = "High performance countries"
  }
  else if (PD$LPI_score[i] == NA) {
    PD$CPG[i] = "None"
  }
}

PD$CPG <- as.factor(PD$CPG)
summary(PD$CPG)
```

```
##   High performance countries    Low performance countries
##                         6768                         4838
## Medium performance countries
##                         6604
```

CC60 Further exploration, outlier box plot of LPI-values

```
unique((subset(PD, PD$LPI_score > 3.3))$country)
```

```
## [1] South Africa
## 39 Levels: Angola Benin Botswana Burkina Faso Burundi Cameroon ...
Zimbabwe
```

The outstanding LPI_score belongs to South Africa

CC61 Overview of countries in the country performance groups

```
# Low Performance
unique((subset(PD, PD$CPG == "Low performance countries"))$country)
```

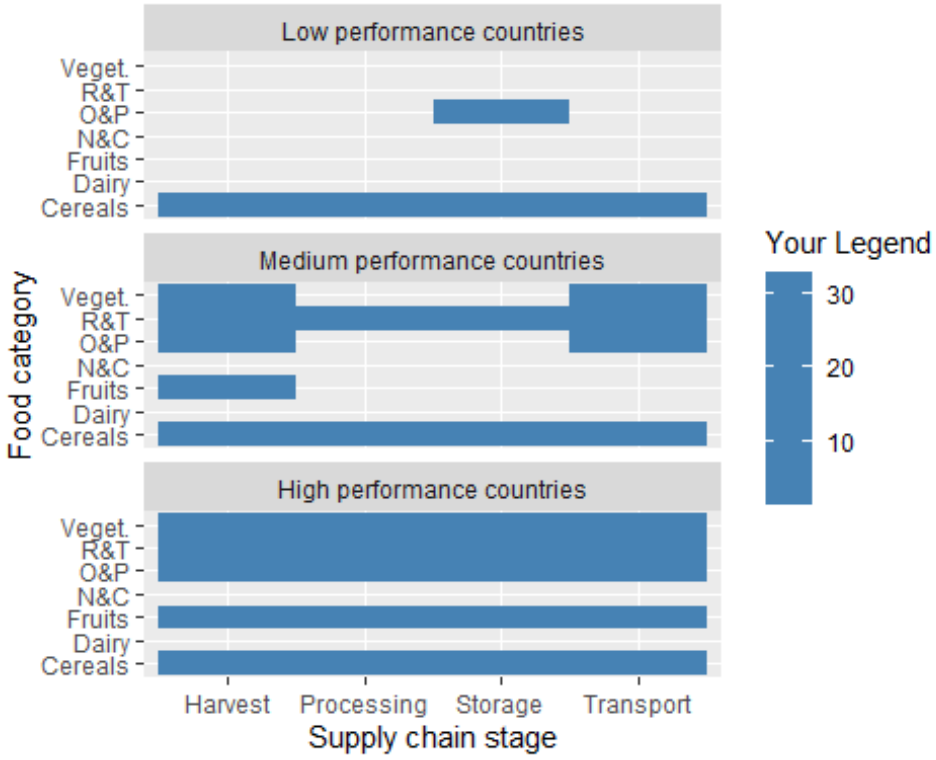
```
##  [1] Angola                          Burundi
##  [3] Democratic Republic of the Congo Eritrea
##  [5] Guinea                          Lesotho
##  [7] Liberia                         Mauritania
##  [9] Niger                           Sierra Leone
## [11] Somalia                         Zimbabwe
```

```
## 39 Levels: Angola Benin Botswana Burkina Faso Burundi Cameroon ...
Zimbabwe

# Medium Performance
unique((subset(PD, PD$CPG == "Medium performance countries"))$country)

##  [1] Cameroon      Chad          Ethiopia      Gambia        Madaga
scar
##  [6] Mali          Mozambique    Nigeria       Guinea-Bissau Senega
l
## [11] Sudan         Togo          Zambia
## 39 Levels: Angola Benin Botswana Burkina Faso Burundi Cameroon ...
Zimbabwe

# High Performance
unique((subset(PD, PD$CPG == "High performance countries"))$country)

##  [1] Botswana                    Benin
##  [3] Ghana                       Côte d'Ivoire
##  [5] Kenya                       Malawi
##  [7] Namibia                     Rwanda
##  [9] South Africa                Uganda
## [11] United Republic of Tanzania Burkina Faso
## 39 Levels: Angola Benin Botswana Burkina Faso Burundi Cameroon ...
Zimbabwe

unique((subset(PD, PD$CPG == "None"))$country)

## factor(0)
## 39 Levels: Angola Benin Botswana Burkina Faso Burundi Cameroon ...
Zimbabwe
```

CC62 Number of occurences inside the cells

```
# Re-order the CPG-labels
PD$CPG = factor(PD$CPG, levels = c("Low performance countries", "Mediu
m performance countries", "High performance countries"))
  cache1 <- PD %>%
    group_by(food_supply_stage, food_category, CPG) %>%
    summarise(mean_loss = mean(loss_percentage, na.rm = T))

## `summarise()` has grouped output by 'food_supply_stage', 'food_cate
gory'. You
## can override using the `.groups` argument.

  colnames(cache1) <- c('food_supply_stage', 'food_category', 'CPG',
                        'mean_loss')


  x = PD %>%
    count(food_supply_stage, food_category, CPG)
  cache1 <- merge(cache1,x)

  cache2 <- ggplot(cache1, aes(x=cache1$food_supply_stage, drop = F,
                               y=cache1$food_category, drop = F,
                               fill=cache1$mean_loss)) +
    geom_tile() +
```

```
    scale_fill_gradient(low="steelblue", high="steelblue", name="Your
Legend") +
    xlab("Supply chain stage") +
    ylab("Food category") +
    scale_x_discrete(drop=FALSE) +
    scale_y_discrete(drop=FALSE) +
    facet_wrap(~CPG, ncol = 1)
cache2
```




```
jpeg("DA_CPG.png", quality = 100, width = 15, height = 15, units = "cm
",
     res = 300)
ready_conversion(cache2) +
geom_text(aes(label = n), size = 3, color = "white")
dev.off()

## png
##   2

# Examining group sizes:

# High performance countries:
nrow(PD %>% filter(CPG == "Low performance countries"))

## [1] 4838

# High performance countries:
nrow(PD %>% filter(CPG == "Medium performance countries"))

## [1] 6604
```

```
# High performance countries:
nrow(PD %>% filter(CPG == "High performance countries"))

## [1] 6768
```

Immediate observation: As for the observations of countries with a low LPI Index there as an apparent lack of data for many FL spots belongign cereals data, which overwhelmingly originated from a single source, a model. Probably, there is lacking food loss research taking place in these countries, therefore there aren't many more observations on other food categories and the related supply chain stages.

CC63 Converting food categories back to categorical values

```
PD$food_category <- as.factor(PD$food_category)
```

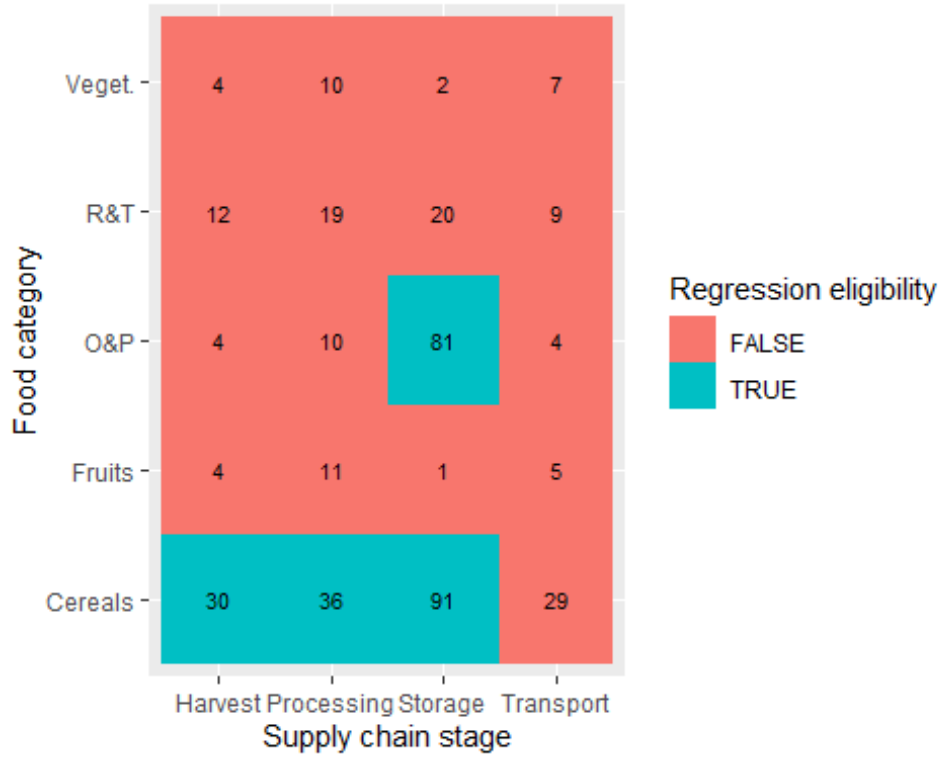
CC64 Preparing LPI regression analysis

```
# For the regression anaylsis, it is important to only consider one
# representative of each stacked set of data points for the regression
# eligibility.

Combi_selector <- function(f_category, sc_stage){
x0 <- subset(PD_with_LPI, gID == 0)
xelse <- subset(PD_with_LPI, gID != 0)
xelse <- xelse[!duplicated(xelse$gID),]
x <- rbind(x0, xelse)

x <- x %>%
  filter((method_data_collection != "Modelled Estimates") &
         (method_data_collection != "")) %>%
  filter(food_supply_stage == sc_stage) %>%
  filter(food_category == f_category)
xe <- str_glue('{f_category}_{sc_stage}')
assign(xe, x, envir = globalenv())
}

# Now selecting the data of the single FL spots
Combi_selector("Cereals", "Harvest")
Combi_selector("Cereals", "Processing")
Combi_selector("Cereals", "Storage")
Combi_selector("Cereals", "Transport")

Combi_selector("Fruits", "Harvest")
Combi_selector("Fruits", "Processing")
Combi_selector("Fruits", "Storage")
Combi_selector("Fruits", "Transport")

Combi_selector("O&P", "Harvest")
Combi_selector("O&P", "Processing")
Combi_selector("O&P", "Storage")
Combi_selector("O&P", "Transport")

Combi_selector("R&T", "Harvest")
```

```
Combi_selector("R&T", "Processing")
Combi_selector("R&T", "Storage")
Combi_selector("R&T", "Transport")

Combi_selector("Veget.", "Harvest")
Combi_selector("Veget.", "Processing")
Combi_selector("Veget.", "Storage")
Combi_selector("Veget.", "Transport")
```

CC65 Sample size and regression eligibility

```
cache1 <- PD_with_LPI %>%
  group_by(food_supply_stage, food_category) %>%
  filter((method_data_collection != "Modelled Estimates") &
         (method_data_collection != "")) %>%
  filter((food_category != "Dairy") &
         (method_data_collection != "N&C")) %>%
  summarise_each(funs(mean))

cache1$food_supply_stage <- as.factor(cache1$food_supply_stage)

x0 <- subset(PD_with_LPI, gID == 0)
xelse <- subset(PD_with_LPI, gID != 0)
xelse <- xelse[!duplicated(xelse$gID),]
x <- rbind(x0, xelse)

x <- x  %>%
  filter((method_data_collection != "Modelled Estimates") &
         (method_data_collection != "")) %>%
  count(food_supply_stage, food_category)

cache1 <- merge(cache1,x)
cache1_copy_for_regr <- cache1

cache1_copy_for_regr$RegrIncl = (cache1_copy_for_regr$n > 29)

cache2 <- ggplot(cache1_copy_for_regr, aes(x=food_supply_stage, y=food
_category,
                                           drop = F, fill=RegrIncl,
                                           showscale = F)) +
  geom_tile() +
  xlab("Supply chain stage") +
  ylab("Food category") +
  scale_y_discrete(drop=TRUE) +
  scale_fill_discrete(name = "Regression eligibility")

cache2 + geom_text(aes(label = n), size = 3)
```




```
jpeg("Reg_Elig.png", quality = 100, width = 15, height = 7,
     units = "cm", res = 300)
ready_conversion(cache2) +
geom_text(aes(label = n), size = 3)
dev.off()

## png
##   2
```

The combinations of food categories and supply chain stages eligible (>30 data points) for the regression analysis are:

cereals/harvest cereals/processing cereals/storage

O&P/storage

CC66 Building the regression function and a close-up look on the countries and thier occurrences

```
RegrFunction <- function(x){
# Dividing the whole data into sub sets
  name <- deparse(substitute(x))
  x <- x %>% drop_na(c("infrastructure", "logistics_competence",
                       "tracking_tracing")) %>%
    filter(method_data_collection != "Modelled Estimates")

# Start of building the model and return the outputs
  model_outcome <- summary(lm(loss_percentage ~
                              infrastructure+logistics_competence+
                              tracking_tracing, data = x))
# Naming the models
  ye <- str_glue('MO_{name}')
```

```
  assign(ye, model_outcome, envir=globalenv())
}

# Showing the countries that contribute data to the respective regress
ion
# analysis and how many data points are associated with them
close_up <- function(spot){
spot <- filter(spot, (spot$method_data_collection != "Modelled Estimat
es") |
                 (spot$method_data_collection != ""))
spot$country <- as.factor(spot$country)
spot <- spot[!is.na(spot$country), ]
summary(spot$country)
}
```

CC67 Running the regression function subsets of data that show eligible sample size, selected in CC65

```
RegrFunction(Cereals_Harvest)
close_up(Cereals_Harvest)
```

```
##                           Angola                            Benin
##                                0                                0
##                         Botswana                     Burkina Faso
##                                0                                0
##                          Burundi                         Cameroon
##                                0                                0
##                             Chad                    Côte d'Ivoire
##                                0                                0
## Democratic Republic of the Congo                          Eritrea
##                                0                                0
##                         Eswatini                         Ethiopia
##                                0                                2
##                           Gambia                            Ghana
##                                0                               12
##                           Guinea                    Guinea-Bissau
##                                0                                0
##                            Kenya                          Lesotho
##                                3                                0
##                          Liberia                       Madagascar
##                                0                                3
##                           Malawi                             Mali
##                                2                                0
##                       Mauritania                       Mozambique
##                                0                                0
##                          Namibia                            Niger
##                                0                                0
##                          Nigeria                           Rwanda
##                                3                                0
##                          Senegal                     Sierra Leone
##                                0                                1
##                          Somalia                     South Africa
##                                0                                0
##                      South Sudan                            Sudan
```

```
##                                0                                0
##                             Togo                           Uganda
##                                0                                1
##      United Republic of Tanzania                           Zambia
##                                3                                0
##                         Zimbabwe
##                                0
```

```
MO_Cereals_Harvest
```

```
##
## Call:
## lm(formula = loss_percentage ~ infrastructure + logistics_competenc
e +
##     tracking_tracing, data = x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.972 -3.085 -1.070  2.162 14.486
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -13.438     11.792  -1.140   0.2649
## infrastructure          9.067     14.143   0.641   0.5271
## logistics_competence   25.797     11.977   2.154   0.0407 *
## tracking_tracing      -26.588      9.737  -2.731   0.0112 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.888 on 26 degrees of freedom
## Multiple R-squared:  0.289,  Adjusted R-squared:  0.207
## F-statistic: 3.523 on 3 and 26 DF,  p-value: 0.02888
```

```
RegrFunction(Cereals_Processing)
close_up(Cereals_Processing)
```

```
##                            Angola                            Benin
##                                 0                                0
##                          Botswana                     Burkina Faso
##                                 0                                0
##                           Burundi                         Cameroon
##                                 0                                0
##                              Chad                      Côte d'Ivoire
##                                 0                                0
## Democratic Republic of the Congo                          Eritrea
##                                 0                                0
##                          Eswatini                         Ethiopia
##                                 0                                1
##                            Gambia                            Ghana
##                                 1                               19
##                            Guinea                      Guinea-Bissau
##                                 0                                0
##                             Kenya                          Lesotho
##                                 3                                0
##                           Liberia                       Madagascar
```

```
##                                 0                                 0
##                            Malawi                              Mali
##                                 1                                 0
##                        Mauritania                        Mozambique
##                                 0                                 0
##                           Namibia                             Niger
##                                 0                                 0
##                           Nigeria                            Rwanda
##                                 4                                 0
##                           Senegal                      Sierra Leone
##                                 0                                 2
##                           Somalia                      South Africa
##                                 0                                 0
##                       South Sudan                             Sudan
##                                 0                                 0
##                              Togo                            Uganda
##                                 0                                 1
##      United Republic of Tanzania                            Zambia
##                                 4                                 0
##                          Zimbabwe
##                                 0

MO_Cereals_Processing

##
## Call:
## lm(formula = loss_percentage ~ infrastructure + logistics_competenc
e +
##     tracking_tracing, data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2154 -4.8829 -2.1406  0.4413 28.2846
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -2.361     18.085  -0.131    0.897
## infrastructure          -1.713     18.026  -0.095    0.925
## logistics_competence     2.255     22.466   0.100    0.921
## tracking_tracing         2.771     16.182   0.171    0.865
##
## Residual standard error: 8.096 on 32 degrees of freedom
## Multiple R-squared:  0.01013,    Adjusted R-squared:  -0.08267
## F-statistic: 0.1091 on 3 and 32 DF,  p-value: 0.9541

RegrFunction(Cereals_Storage)
close_up(Cereals_Storage)

##                            Angola                             Benin
##                                 0                                 3
##                          Botswana                      Burkina Faso
##                                 0                                 8
##                           Burundi                          Cameroon
##                                 0                                 0
##                              Chad                     Côte d'Ivoire
```

```
##                                0                                0
## Democratic Republic of the Congo                          Eritrea
##                                0                                0
##                         Eswatini                         Ethiopia
##                                0                                6
##                          Gambia                            Ghana
##                                0                               21
##                          Guinea                      Guinea-Bissau
##                                0                                0
##                           Kenya                          Lesotho
##                              14                                0
##                         Liberia                       Madagascar
##                                0                                0
##                          Malawi                             Mali
##                                5                                0
##                      Mauritania                       Mozambique
##                                0                                0
##                         Namibia                            Niger
##                                0                                1
##                         Nigeria                           Rwanda
##                              11                                6
##                         Senegal                     Sierra Leone
##                                0                                1
##                         Somalia                     South Africa
##                                0                                0
##                     South Sudan                            Sudan
##                                0                                0
##                            Togo                           Uganda
##                                0                                9
##      United Republic of Tanzania                           Zambia
##                                4                                2
##                         Zimbabwe
##                                0

MO_Cereals_Storage

##
## Call:
## lm(formula = loss_percentage ~ infrastructure + logistics_competenc
e +
##     tracking_tracing, data = x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.913 -6.105 -3.406  1.624 50.455
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -8.584     21.191  -0.405    0.686
## infrastructure         -1.251     17.033  -0.073    0.942
## logistics_competence    6.427     14.453   0.445    0.658
## tracking_tracing        1.235     10.217   0.121    0.904
##
## Residual standard error: 10.39 on 87 degrees of freedom
```

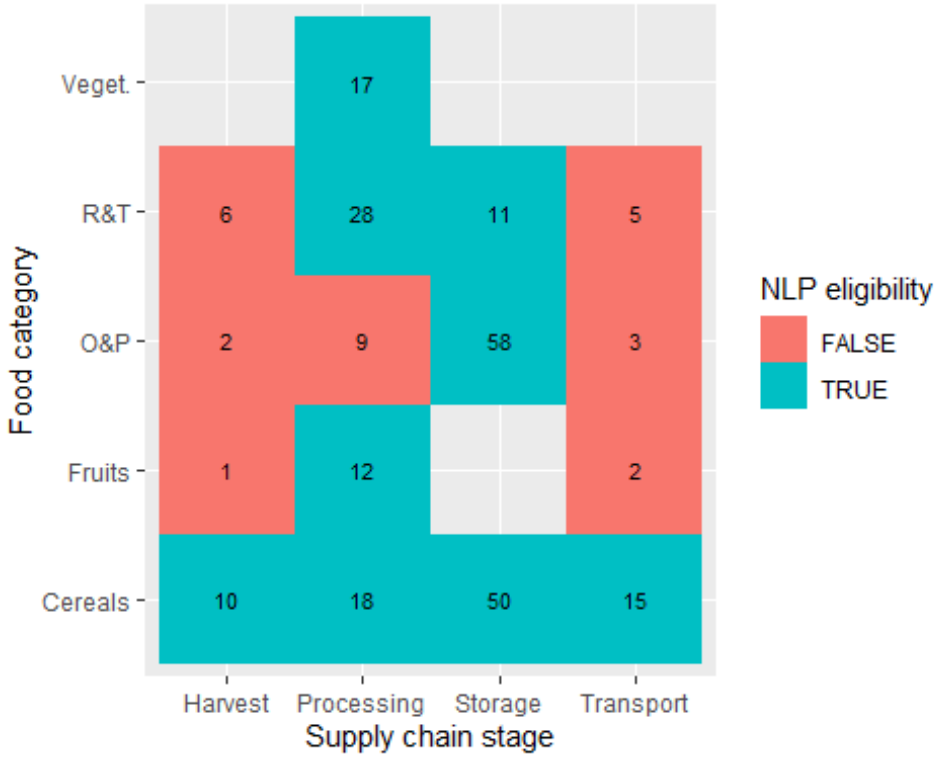
```
## Multiple R-squared:  0.01422,    Adjusted R-squared:  -0.01978
## F-statistic: 0.4182 on 3 and 87 DF,  p-value: 0.7404

RegrFunction(`O&P_Storage`)
close_up(`O&P_Storage`)

##                             Angola                              Benin
##                                  0                                  0
##                           Botswana                       Burkina Faso
##                                  0                                  0
##                            Burundi                           Cameroon
##                                  0                                  0
##                               Chad                      Côte d'Ivoire
##                                  0                                  0
## Democratic Republic of the Congo                            Eritrea
##                                  0                                  0
##                           Eswatini                           Ethiopia
##                                  0                                  0
##                             Gambia                              Ghana
##                                  0                                  3
##                             Guinea                      Guinea-Bissau
##                                  0                                  0
##                              Kenya                            Lesotho
##                                  0                                  0
##                            Liberia                         Madagascar
##                                  0                                  0
##                             Malawi                               Mali
##                                  1                                  0
##                         Mauritania                         Mozambique
##                                  0                                  0
##                            Namibia                              Niger
##                                  0                                  0
##                            Nigeria                             Rwanda
##                                  0                                  0
##                            Senegal                       Sierra Leone
##                                  0                                  0
##                            Somalia                       South Africa
##                                  0                                  0
##                         South Sudan                              Sudan
##                                  0                                  0
##                               Togo                             Uganda
##                                  0                                  1
##      United Republic of Tanzania                             Zambia
##                                  6                                  0
##                           Zimbabwe
##                                 70

`MO_O&P_Storage`

##
## Call:
## lm(formula = loss_percentage ~ infrastructure + logistics_competenc
e +
##     tracking_tracing, data = x)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.518 -4.090 -3.690 -2.590 44.310
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -37.41      36.71  -1.019    0.311
## infrastructure         -55.01      90.47  -0.608    0.545
## logistics_competence    54.54      67.27   0.811    0.420
## tracking_tracing        15.24     117.75   0.129    0.897
##
## Residual standard error: 11.2 on 77 degrees of freedom
## Multiple R-squared:  0.02493,    Adjusted R-squared:  -0.01306
## F-statistic: 0.6561 on 3 and 77 DF,  p-value: 0.5816
```

Retriving meaning from cause of loss data by the means of NLP:
Classification/Causalitiy-Classification for decision support model:

CC68 Sample size and NLP eligibility

```
# Entries in the cause of loss column
nrow(PD)

## [1] 18210

# Entries in the cause of loss column, without modeled estimates
nrow(PD %>%
  filter(cause_of_loss != ""))

## [1] 247

PD_cache <- PD %>%
  filter(cause_of_loss != "")  %>%
  group_by(food_supply_stage, food_category) %>%
  count(.)

nrow(PD %>%
  filter(cause_of_loss != ""))

## [1] 247

PD_cache$RegrIncl = (PD_cache$n > 9)

cache2 <- ggplot(PD_cache, aes(x=food_supply_stage, y=food_category,
                                          drop = F, fill=RegrIncl,
                                          showscale = F)) +
  geom_tile() +
  xlab("Supply chain stage") +
  ylab("Food category") +
  scale_y_discrete(drop=TRUE) +
  scale_fill_discrete(name = "NLP eligibility")

cache2 + geom_text(aes(label = n), size = 3)
```




```
jpeg("Reg_Elig_NLP.png", quality = 100, width = 15, height = 7,
     units = "cm", res = 300)
ready_conversion(cache2) +
geom_text(aes(label = n), size = 3)
dev.off()

## png
##   2
```

CC69 Direct hand-over of a data frame from to python can be avoided by storing it as a csv-file, then opening the csv-file again.

```
# Since the SC stages Households, wholesale, market and retail are out
of the
# project's scope.
PD_with_comments = subset(PD_copy, !(food_supply_stage %in% c('Household
s',
                                                              'Wholesa
le',
                                                              'Market'
,
                                                              'Retail'
)),
                          drop=FALSE)

# For the sake of fast compoutation only data points to no empty data
in the
# column "cause of loss" will be included in this sub-data set.
PD_with_comments <- subset(PD_with_comments, cause_of_loss != "")
PD_with_comments <- PD_with_comments %>% select(cause_of_loss, index)
```

```
#Adding an Index-Column to identify the columns later on
PD_with_comments <- PD_with_comments %>% mutate(id = row_number())
write.csv(x = PD_with_comments, file = "PD_with_comments_all_SSA_all_S
C.csv")

# Number of entries in the cause of loss column
length(PD_with_comments$index)

## [1] 301
```

Classification task including cuasality:

CC70 Prepare for use of Python

```
library(reticulate)
# It's necessary to use the anaconda shell as a package manager
use_python("C:/Users/User/anaconda3", required = T)
```

Authentication via session token, then using an API for automated use of ChatGPT. When running the script, probably a new token has to be generated, which can be accessed when logged in on ChatPGT via F12 -> Application -> session token

CC71 Authentification via sesseion token

```
# For the sake of a seamless computation of all code chunks, the code
chunks in
# python were set to eval = False, so that they are not run by default
.
# The instructions for using the ChatGPT API and its implementation in
Python
# were derived from a YouTube video by (1littlecoder 2022)
# https://www.youtube.com/watch?v=S3okwVkxDgA

from pyChatGPT import ChatGPT
session_token = 'eyJhbGciOiJkaXIiLCJlbmMiOiJBMjU2R0NNIn0..DWL2ozYHLSZV
MTv6.a3R7TuIrsnWJ0EN2glb9BDIgY5-FaueJmFSUg7C61D4lRvzQkXjUiBFdwjw1F4jq7
3Io3uw7DRRTgZf8AqGwMbRO5qJz_BKec64ly3sIA6tStjV90v8K9_JRMepN56wPqSSa-hF
KV9DgV9LLHU0X2NbsSc_ZoL_2CptEtVvz9u9zuPJzz_hUn-YR689zmfR2SjBiOlc7PNtvp
oogjG5HSLt3J5mHkvTBDQ25lgHewRsI1Q-7X8LpVIMsGdRyhntGYFI5sovQ6u_cT2BWOBH
CiI_-_PhoG-UJoppYf3xbDoxAMw_kiQ23Z-iW99VOX9Ci5l3NInhkSXh4BCavTQ2lzTbRg
LF_csvlKY9gJvq-HSmj2RC6ZHfHAIM22CjjGi0hPzNl9BfVHHa8xMLvX7wM-OJ3Sl-xarU
5XgvgnTeVezksXXpTH3X8QuSBHTq6XOK7OSBOXrn8qk4BoybJxGoVBKTxekbsBpAp24iYI
QYNiiLvF_3Tb2iysx3ZryR8uyKwb9ML9aCQRh0E95cJVJzCwJXPK1lxuvrmq4da7bE4AaI
NEs6446YfxLfd3KV8Dw9mPznFAvbcIvsx3CLxjMAtsgnv0ZSxOpzvl76czAX6aG4SKeTrZ
9ko5BQxBxaK4MxOpV-UZNPZI9vFx9ZuHlNc_MSv_BY-jabmJsZJo48nGiWAeQsc3ikeS44
gvFAZXrj166t5mlaTn5GfzZQ2723xphLRi5x3zf-ktneDXhAAq5hChC1lczY_VMdJfdVw5
2EfVnzE9BjXSdgdGkfzROVurH_JbNnOERI93sU97_eY1Da5hwGiLsBAtaNUpV5KJLk2_Dw
r34C7-kl_oxEJesfEYwKOEmaeUR3tiE3K6ysBNBYpTu1yh0Yup7b9wLtgwPx_XjZN3f0Ob
1AcC_SHqQ84fm7zXq61NUovzJHX9oGF9ztKtMtP1q3MRcfZYoUeol9tb3IwkTJUw3S-UDr
E96CjwO5Psi-d8903VPZpyInFpg5_J-4byEJdyP_cKfizKjKeF6OwDGwK7oAzT5r0XAeRk
r0Nt-Kbi_JjmByIRZ9QCpNV-zqFV6-Sorxq2sQAm_K6nP2PudqyTH_tDVPUkvWpIJzQHt6
gFYQuvKXwVSJaWsDVbHxVydW3NvW60sgmlmY6BcnrY9bsUqwQuJkGGuS9oZmQuo_PBLgoM
TYgwXobar4QRusN2yNHRblnvvNvxunpdQqYnfjkS44k_tkI1rA_IwCXIyKLDu-tkba3jEL
s6vJWqBKTsi4yNJClIyuCQGS9jaU8F2gfbOfB2rq4vVkp90Fasm72d-IeCcU3yg5MX0THS
vGf5b5s5dwH52tzsDo7WA1Wxuc5rwvgfuOPJ-_lYZWYc6NRHNu0E5Y205Jsf4EPaNSlmaP
```

```
IUxgxxV97aGQdLheouTZg2NijIaKC102IjEy250v445_7PtXVWpb8UeHUFsXcNA7sx_0QO
LE29VoLERC-p8Tr0bnTGGuUiyuGb7RaEuC7-kEbEBhAnqbiovcxf4E8Ukl8WQ6rRIwriwW
WFCLERGxD2U_WPxmmOrfUrrRwzbpLaFuHnZPamTZk_sLYYcHfZAQ6kCmL9GInm84k_JmyC
VB0AI7C5snP6fhaFzI73jkI5A7ZRDj3EICJG-rYZm3PhGJxm9Up7qzCHuLD1Qcszsoo4Kr
H9IOKRWUoYulRAus9lomOHAfusvCV9-wRPX4ROvhUddn635S5F-GACByg0uJyl66rDJZ1M
V3rIOJ0kCxRH1Co4CBMsAeleY7mZp83689WhUGkbsVl1QY6EZlHnbujKIXRbScP3d8HJM3
3Lzur41uo7grIW2JpfYfKSVDRjZ43wAg7eEWnE6bU0fBBubAFzqXWLLKi_148vVQnUKLMi
2Dy1NfJ4ILjkVnwPvnnROS9Vk8QNuvs7ArBsOJlr4pehO9cVP8A4UFKkhrrG8OmatmLjtB
Gto3Uhykx4T5jGP66vz4U4qNmZ7U-1t15Z14O6mbAHynMGHws8ZTmpM6ugJNyliIIh2Pnk
nfNzS_JxASSspLabl2tGzVjKrQAKtDS7505bQRDu4_vxmwExYPl5OiHZACxDMCsn81GQsC
V44irsr3AuJ0IiWY_YeiwlYL0ywfJdAFhJ2AVhcgJ7NbLj5IM7heyPfczM6mgNzS2TeLie
GK9NgtuFbq3wKaVJOUM00INFEuHVIwh-1U2D5ef3FBHSHL0JPxX2_QZ_JO2IuadnwJmVdu
VWGBPmH31Zs7HowhsSSyQ2VyVqemgJDbIZBgq2nc8OwcugEyTQ4qZL4Ck6sVW5JfR58T0g
kU8SoYqi0a6KsK7HNw_ybwdGg7xobVGCdQOxOSu_gNOdY3GTYvG7Mrt2P6vs9gDQQOOijl
znNbxtJ6V0IXR7YTuqhS3KR96cs7ALPhjdFvz1AAyplEDJdd2rJBv5coqqDPd6O5ifN8GO
FOaDRQ25NaOyRev-_h_HNb1nA938BAI2-tvAKM_Wdw1qi07OYWp_YYhVVLMctLPR-n-M58
Ov8H3V7J5vXA1j0X5Ekwt152m1iNiv-krUfuMxq3-ZnzMaKDxUKiLacLfCwYcNwhBpF_Ta
c4qIaQaf1qxD-x38f06Ea4iZlwbc6xdn2DwIzHM-cK-nBLf3G_G4f9_nTu7hzH7jxJZ-6S
ZrtrIVHBVhpk9VmJx1qneYDzqa-fR26-pS05XQyR1pMQNZ7QhNO1d9U4.XaaehEP23hT56
v64MzdTLA'
api = ChatGPT(session_token)
```

Package for ChatGPT-API only exists for Python. Therefore, the python programming environment inside the RMarkdown had to be used.

One major problem doing the assigment task with the API was that in the case of the few entries in which multiple causes of losses were given, the classification failed. Therefore, these entries needed to be cleaned in a way that they would only contain one cause of loss. As to find a way to achieve this, always the first cause of loss was chosen and other causes of loss thereafter were deleted. The result of the cleaned cause of loss table can be retrieved from the GitHub Repository.

CC72 IMport libraries and initialize matrix to store the results in

```
# For the sake of a seamless computation of all code chunks, the code
chunks in
# python were set to eval = False, so that they are not run by default
.

import pandas as pd
import numpy as np
import json

PD_with_comments_edi = pd.read_excel('CoLs_Edited.xlsx',
sheet_name='CoLs Edited')

# v0 is a dummy vector, which is used to to initialize the
# data frame "vector_collect"
vectorcollect = np.array([0,0,0,0,0,0,0,0,0,0,0,0,0,0])
```

CC73 ChatGPT query

```
# Data of data retrieval: in the night from the 29th to the 30th Jan,
acc.
```

```
# to UTS +1. In reality, the following chunk had to be executed severa
l times
# since ChatGPT only allows for about 60 inquiries per hour to prevent
excessive
# use of its server capacities coming from the same IP address. After
about
# 60 data retrivals, I had to wait for another hour.

  i = (vector_collect.ndim-1)
  while i <(len(PD_with_comments_edi)-1):
  cachy = PD_with_comments_edi.loc[i, "cause_of_loss_edit"]
  resp = api.send_message(f"What's the best way to reduce food loss if
the cause of loss is {cachy} ? Possible fields of action for counter m
easures would be: '-1- Tranparency' which is described as 'Increase of
transparency within a company as well as between companies of a networ
k', '-2- Quality management', which is described as 'Improvement of qu
ality management for early detection of weaknesses'. '-3- Packaging ma
nagment' which is described as Improvement of packaging management dur
ing transport and storage processes as well as for distribution to the
end customer, loading of vehicles, and coordination of vehicles', '-3-
Financial opportunities' which is described as 'Providing appropriate
financial support from the administration to weaker network partners',
'-4- Transport optimization' which is described as 'Improvement of tra
nsport management with regard to route planning, loading of vehicles,
and coordination of vehicles', '-5- Warehouse management' which is des
cribed as 'Improvement of warehouse management using suitable storage
equipment, storage strategies, and adapted layout planning'. '-6- Netw
ork structure' which is described as 'Improvement of the network struc
ture using strategic network planning and location management', '-7- R
egulation' which is described as 'Adapted regulations by the administr
ation to support companies in reducing food losses as required', '-8-
Financing opportunities' which is described as 'Providing appropriate
financial support from the administration to weaker network partners',
-9- Physical characterictics' which is described as 'Adaptation of pro
cesses to consider special physical requirements of the products, incl
uding temperature, pressure sensitivity, and air composition, '-10- Sh
elf-life optimization' which is described as 'Process adaptations that
allow the shelf life of the products to be taken into account in decis
ion making, '-11- Network cooperation' which is described as 'Improvin
g cooperation within networks, including information sharing and effor
ts to develop comprehensive measures against food losses', '-12- Mindf
ulness' which is described as 'Promoting awareness among employees at
all levels in companies of the relevance of the problem of food losses
in everyday life', '-13- Consumer satisfaction' which is described as
Adaptation of internal processes with the aim of meeting specific cust
omer requirements. Please only print out the highly relevant answers a
nd keep the numbers inside the delimiters and return the result only a
s a numerical list in square brackets. And don't print out any explana
tion of the decisions taken, print out nothing but the list.")
  print(resp['message'])
  api.reset_conversation()
  chachy2 = resp['message']
  chachy2 = chachy2.rstrip(chachy2[-1])
  vector1 = chachy2
```

```
  vector1 = json.loads(vector1)
  vector1 = np.array(vector1)
  vector1.resize((1,13), refcheck = False)
  vector1 = np.append(i, vector1)
  vector_collect = np.row_stack( (vector_collect, vector1) )
  i += 1

# Important to note: The term "Prio" is misleading, since the the fiel
ds of
# actions that werde deemed to be relevant were not sorted by relevanc
e. It was
# first planned to include prioritization, but later rejected.
data_frame_saved = pd.DataFrame(vector_collect, columns=['Entry number
',
'1st Prio','2nd Prio','3rd Prio','4th Prio','5th Prio','6th Prio','7th
Prio',
'8th Prio','9th Prio','10th Prio','11th Prio','12th Prio','13th Prio']
)

data_frame_saved.to_csv('GPTv1.csv', index=False)
```

CC74 Breaking the outcome down again into SC stages

```
Edi_col <- read_csv("PD_with_comments_all_SSA_all_SC.csv")

## New names:
## Rows: 301 Columns: 4
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter:
"," chr
## (1): cause_of_loss dbl (3): ...1, index, id
## ℹ Use `spec()` to retrieve the full column specification for this d
ata. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet t
his message.
## • `` -> `...1`

GPT_outcome <- read_csv("GPTv1.csv")

## Rows: 301 Columns: 14
## ── Column specification ──────────────────────────────────────────
──────────
## Delimiter: ","
## dbl (14): Entry number, 1st Prio, 2nd Prio, 3rd Prio, 4th Prio, 5th
Prio, 6t...
##
## ℹ Use `spec()` to retrieve the full column specification for this d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
this message.

# The first row is the initialization row containing nothing but rows,
which has
# to be deleted again now.
GPT_outcome <- GPT_outcome[2:301,]
```

```
GPT_outcome$`Entry number` <- GPT_outcome$`Entry number`+1

colnames(GPT_outcome) <- c("id", "p1", "p2", "p3", "p4","p5", "p6", "p
7", "p8",
                           "p9", "p10", "p11", "p12", "p13")
```

```
#{r} # Replace row 45 Manually (mistake spotted later)
GPT_outcome[45,] <- data.frame(45, 2, 3, 5, 9, 10, 0, 0, 0, 0, 0, 0,
0, 0) #
```

CC75 Matching the data frame outputted by ChatGPT in Python and the the data frame of data that was fed into the API

```
combined_df <- merge(Edi_col, GPT_outcome, by = "id")
matcher <- select(PD_copy, c("food_supply_stage", "index", "food_categ
ory",
                             "activity"))
combined_df <- merge(combined_df, matcher, by = "index")
# Quick security check that indices match with one another
which(!(combined_df$index %in% PD_with_comments$index))

## integer(0)
```

CC76 Preparing subsets of data for natural language processing (NLP)

```
# Modeled data did not have to be dropped as modelled data would usual
ly not
# contain information on the cause of loss anyway due to the nature of
modelled
# data.
Combi_selector <- function(f_category, sc_stage){
x <- combined_df %>%
  filter(food_supply_stage == sc_stage) %>%
  filter(food_category == f_category)

xe <- str_glue('{f_category}_{sc_stage}')
assign(xe, x, envir = globalenv())
}

# Creating subsets for all combinations for food categories and SC sta
ges
# Cereals
Combi_selector("Cereals", "Harvest")
Combi_selector("Cereals", "Processing")
Combi_selector("Cereals", "Storage")
Combi_selector("Cereals", "Transport")
# Fruits
Combi_selector("Fruits", "Harvest")
Combi_selector("Fruits", "Processing")
Combi_selector("Fruits", "Storage")
Combi_selector("Fruits", "Transport")
# O&P
Combi_selector("O&P", "Harvest")
Combi_selector("O&P", "Processing")
Combi_selector("O&P", "Storage")
```

```
Combi_selector("O&P", "Transport")
# R&T
Combi_selector("R&T", "Harvest")
Combi_selector("R&T", "Processing")
Combi_selector("R&T", "Storage")
Combi_selector("R&T", "Transport")
# Vegetables
Combi_selector("Veget.", "Harvest")
Combi_selector("Veget.", "Processing")
Combi_selector("Veget.", "Storage")
Combi_selector("Veget.", "Transport")
```

Building of a scoring model:

CC77 Cut NLP analysis results into combinations of food category and supply chain stages

```
# Creating the sub sets of data and the respective models
xe <- data.frame()
wert <- data.frame()

gptresults <- function(hotspot_data){
# Dividing the whole data into sub sets
  matcher <- select(hotspot_data, c("index"))
  combined_df <- merge(combined_df, matcher, by = "index")
  wert <- deparse(substitute(hotspot_data))
  x <- combined_df
  xe <- str_glue('GPT_{wert}')
  assign(xe, x, envir = globalenv())
}
```

CC78 Compute NLP analysis reusults for all combinations of food categories and SC stages

```
gptresults(Cereals_Harvest)
gptresults(Cereals_Processing)
gptresults(Cereals_Storage)
gptresults(Cereals_Transport)

gptresults(Fruits_Harvest)
gptresults(Fruits_Processing)
gptresults(Fruits_Storage)
gptresults(Fruits_Transport)


gptresults(`O&P_Harvest`)
gptresults(`O&P_Processing`)
gptresults(`O&P_Storage`)
gptresults(`O&P_Transport`)

gptresults(`R&T_Harvest`)
gptresults(`R&T_Processing`)
gptresults(`R&T_Storage`)
gptresults(`R&T_Transport`)
```

```
gptresults(Veget._Harvest)
gptresults(Veget._Processing)
gptresults(Veget._Storage)
gptresults(Veget._Transport)
```

CC79 Creating a benchmark with overall numbers of all cause of loss entries

```
table_and_column_showerall <- function(g, occ_supply_stage){
myvector <- data.frame()
for(i in 1:13) {
v <- print(sum(g[,5:17] == i))
myvector[i,1] <- v
}
indexcol <- as.data.frame(c(1:13))
myvector <- cbind(indexcol, myvector)
myvector$percent = NA
for(i in 1:length(myvector$`c(1:13)`)){
myvector$percent[i] = (myvector$V1[i]/occ_supply_stage)
}

colnames(myvector) <- c("Field of action", "Occurences",
                        "Percentage of all occurences")

myvector[1,1] <- "Transparency"
myvector[2,1] <- "Quality management"
myvector[3,1] <- "Packaging management"
myvector[4,1] <- "Transport optimization"
myvector[5,1] <- "Warehouse management"
myvector[6,1] <- "Network structure"
myvector[7,1] <- "Regulation"
myvector[8,1] <- "Financing opportunities"
myvector[9,1] <- "Physical characteristics"
myvector[10,1] <- "Shelf-life optimization"
myvector[11,1] <- "Network cooperation"
myvector[12,1] <- "Mindfulness"
myvector[13,1] <- "Consumer satisfaction"

assign("myvectorall", myvector, envir = globalenv())
}
table_and_column_showerall(combined_df, length(combined_df[,1]))
```

```
## [1] 21
## [1] 212
## [1] 259
## [1] 88
## [1] 261
## [1] 45
## [1] 65
## [1] 16
## [1] 113
## [1] 246
## [1] 140
```

```
## [1] 117
## [1] 151
```

CC80 Creating a a visualization of the results of the cause of loss analysis

```
table_and_column_shower <- function(g, occ_supply_stage){
myvector <- data.frame()
for(i in 1:13) {
v <- print(sum(g[,5:17] == i))
myvector[i,1] <- v
}
indexcol <- as.data.frame(c(1:13))
myvector <- cbind(indexcol, myvector)
myvector$percent = NA

# Column and bar chart integrated
for(i in 1:length(myvector$`c(1:13)`)){
myvector$percent[i] = (myvector$V1[i]/occ_supply_stage)
}

colnames(myvector) <- c("Field of action", "Occurences",
                        "Percentage of all occurences")

myvector[1,1] <- "Transparency"
myvector[2,1] <- "Quality management"
myvector[3,1] <- "Packaging management"
myvector[4,1] <- "Transport optimization"
myvector[5,1] <- "Warehouse management"
myvector[6,1] <- "Network structure"
myvector[7,1] <- "Regulation"
myvector[8,1] <- "Financing opportunities"
myvector[9,1] <- "Physical characteristics"
myvector[10,1] <- "Shelf-life optimization"
myvector[11,1] <- "Network cooperation"
myvector[12,1] <- "Mindfulness"
myvector[13,1] <- "Consumer satisfaction"

myvector <- bind_cols(myvector, myvectorall)
colnames(myvector) <- c("Field of action", "Occurences Hot Spot",
                        "Percentage_of_all_occurences", "Field of acti
onA",
                        "Occurences Hot SpotA",
                        "overall")
myvector$`Field of action` <- factor(myvector$`Field of action`, level
s =
    c("Consumer satisfaction", "Mindfulness", "Network cooperation",
      "Shelf-life optimization", "Physical characteristics", "Financin
g opportunities","Regulation", "Network structure",  "Warehouse manage
ment",
      "Transport optimization", "Packaging management", "Quality manag
ement", "Transparency"))
assign("myvector", myvector, envir = globalenv())
}
```

```
ttob <- function(myvector){
# Transforming the numbers from decimals to percentages
myvector$Percentage_of_all_occurences =
  (myvector$Percentage_of_all_occurences)*100
myvector$overall = (myvector$overall)*100
g <- ggplot(myvector, aes(`Field of action`)) +
  geom_col(aes(y=`Percentage_of_all_occurences`), fill="steelblue") +
  geom_point(aes(y=overall, color="overall"), size=4, shape=124) +
  scale_color_manual(values=c(overall="red"),
                     labels=c(overall= "Overall benchmark")) +
  coord_flip()+
  guides(size=FALSE) +
  theme(legend.box="horizontal",
        legend.key=element_blank(), legend.title=element_blank(),
        legend.position="top") + scale_y_continuous("Relevance [%]",
                                                    limits = c(0,100))
assign("g", g, envir = globalenv())
}
```

CC81 Checking the eligibility of combinations for the cause of loss analysis

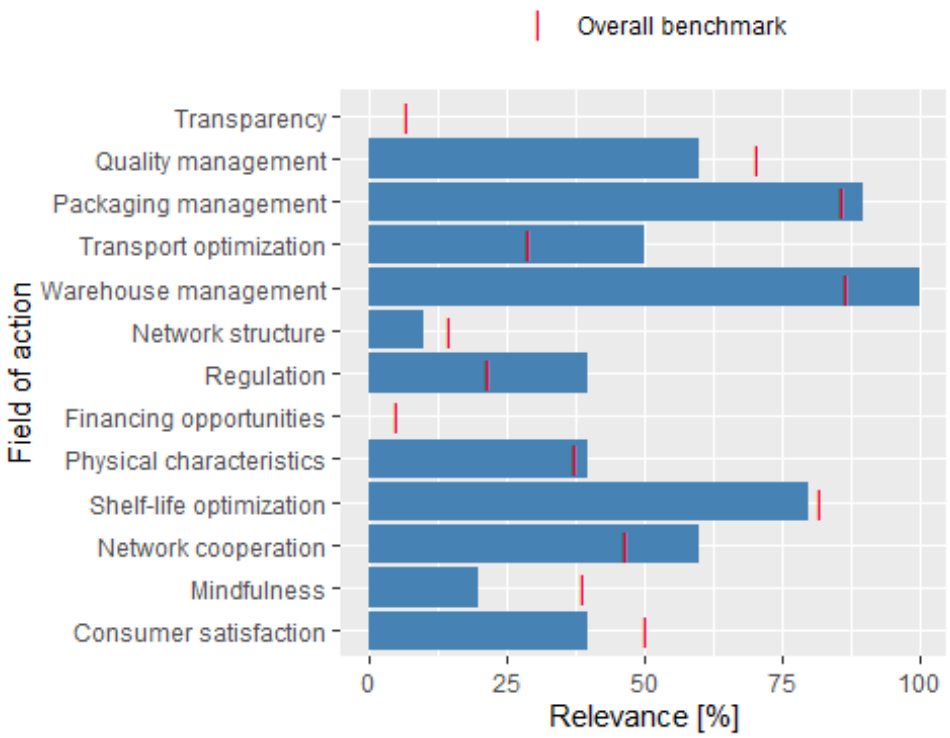
```
nrow(GPT_Cereals_Harvest) # eligible

## [1] 10

nrow(GPT_Cereals_Processing) # eligible

## [1] 18

nrow(GPT_Cereals_Transport) # eligible

## [1] 15

nrow(GPT_Cereals_Storage) # eligible

## [1] 51

nrow(GPT_Fruits_Harvest)

## [1] 1

nrow(GPT_Fruits_Processing) # eligible

## [1] 12

nrow(GPT_Fruits_Transport)

## [1] 2

nrow(GPT_Fruits_Storage)

## [1] 0

nrow(`GPT_O&P_Harvest`)

## [1] 2

nrow(`GPT_O&P_Processing`)
```

```
## [1] 9

nrow(`GPT_O&P_Transport`)

## [1] 3

nrow(`GPT_O&P_Storage`) # eligible

## [1] 58

nrow(`GPT_R&T_Harvest`)

## [1] 6

nrow(`GPT_R&T_Processing`) # eligible

## [1] 27

nrow(`GPT_R&T_Transport`)

## [1] 5

nrow(`GPT_R&T_Storage`) # eligible

## [1] 11

nrow(GPT_Veget._Harvest)

## [1] 0

nrow(GPT_Veget._Processing) # eligible

## [1] 17

nrow(GPT_Veget._Transport)

## [1] 0

nrow(GPT_Veget._Storage)

## [1] 0
```

CC82 Cause of loss analysis for combination of cereals-harvest

```
table_and_column_shower(GPT_Cereals_Harvest, length(GPT_Cereals_Harves
t[,1]))

## [1] 0
## [1] 6
## [1] 9
## [1] 5
## [1] 10
## [1] 1
## [1] 4
## [1] 0
## [1] 4
## [1] 8
## [1] 6
```

```
## [1] 2
## [1] 4

## New names:
## • `Field of action` -> `Field of action...1`
## • `Occurences` -> `Occurences...2`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.3`
## • `Field of action` -> `Field of action...4`
## • `Occurences` -> `Occurences...5`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.6`

ttob(myvector)

## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Us
e "none" instead as
## of ggplot2 3.3.4.

g
```




```
jpeg("GPTF_Cereals_Harvest.png", quality = 100, width = 15, height = 8
.5,
     units = "cm", res = 300)
ready_conversion(g)
dev.off()

## png
##   2
```

```
# Calculation the n (numbers of data points) in the sub data set.
# They all contain data in the cause of loss column.
nrow(GPT_Cereals_Harvest)

## [1] 10
```

CC83 Cause of loss analysis for combination of cereals-processing

```
table_and_column_shower(GPT_Cereals_Processing,
                        length(GPT_Cereals_Processing[,1]))

## [1] 2
## [1] 13
## [1] 13
## [1] 3
## [1] 13
## [1] 3
## [1] 4
## [1] 1
## [1] 4
## [1] 15
## [1] 9
## [1] 10
## [1] 11

## New names:
## • `Field of action` -> `Field of action...1`
## • `Occurences` -> `Occurences...2`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.3`
## • `Field of action` -> `Field of action...4`
## • `Occurences` -> `Occurences...5`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.6`

ttob(myvector)
g
```


```
jpeg("GPTF_Cereals_Processing.png", quality = 100, width = 15, height
= 8.5,
     units = "cm", res = 300)
ready_conversion(g)
dev.off()

## png
##   2

# Calculation the n (numbers of data points) in the sub data set.
# They all contain data in the cause of loss column.
```

CC84 Cause of loss analysis for combination of cereals-storage

```
table_and_column_shower(GPT_Cereals_Storage, length(GPT_Cereals_Storag
e[,1]))

## [1] 3
## [1] 35
## [1] 40
## [1] 7
## [1] 49
## [1] 3
## [1] 5
## [1] 0
## [1] 26
## [1] 46
## [1] 18
## [1] 19
## [1] 17
```

```
## New names:
## • `Field of action` -> `Field of action...1`
## • `Occurences` -> `Occurences...2`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.3`
## • `Field of action` -> `Field of action...4`
## • `Occurences` -> `Occurences...5`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.6`

ttob(myvector)
g
```


```
jpeg("GPTF_Cereals_Storage.png", quality = 100, width = 15, height = 8
.5,
     units = "cm", res = 300)
ready_conversion(g)
dev.off()

## png
##   2

# Calculation the n (numbers of data points) in the sub data set.
# They all contain data in the cause of loss column.
nrow(GPT_Cereals_Storage)

## [1] 51
```

CC85 Cause of loss analysis for combination of cereals-transport

```
table_and_column_shower(GPT_Cereals_Transport,
                        length(GPT_Cereals_Transport[,1]))
```

```
## [1] 0
## [1] 7
## [1] 14
## [1] 14
## [1] 9
## [1] 2
## [1] 5
## [1] 3
## [1] 2
## [1] 10
## [1] 8
## [1] 6
## [1] 10

## New names:
## • `Field of action` -> `Field of action...1`
## • `Occurences` -> `Occurences...2`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.3`
## • `Field of action` -> `Field of action...4`
## • `Occurences` -> `Occurences...5`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.6`

ttob(myvector)
g
```


```
jpeg("GPTF_Cereals_Transport.png", quality = 100, width = 15, height =
8.5,
     units = "cm", res = 300)
```

```
ready_conversion(g)
dev.off()

## png
##   2

# Calculation the n (numbers of data points) in the sub data set.
# They all contain data in the cause of loss column.
nrow(GPT_Cereals_Transport)

## [1] 15
```

CC86 Cause of loss analysis for combination of fruits-processing

```
table_and_column_shower(GPT_Fruits_Processing,
                        length(GPT_Fruits_Processing[,1]))

## [1] 0
## [1] 5
## [1] 12
## [1] 4
## [1] 11
## [1] 0
## [1] 0
## [1] 0
## [1] 5
## [1] 9
## [1] 2
## [1] 2
## [1] 2

## New names:
## • `Field of action` -> `Field of action...1`
## • `Occurences` -> `Occurences...2`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.3`
## • `Field of action` -> `Field of action...4`
## • `Occurences` -> `Occurences...5`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.6`

ttob(myvector)
g
```


```
jpeg("GPTF_Fruits_Processing.png", quality = 100, width = 15, height =
8.5,
     units = "cm", res = 300)
ready_conversion(g)
dev.off()

## png
##   2

# Calculation the n (numbers of data points) in the sub data set.
# They all contain data in the cause of loss column.
nrow(GPT_Fruits_Processing)

## [1] 12
```

CC87 Cause of loss analysis for combination of O&P-storage

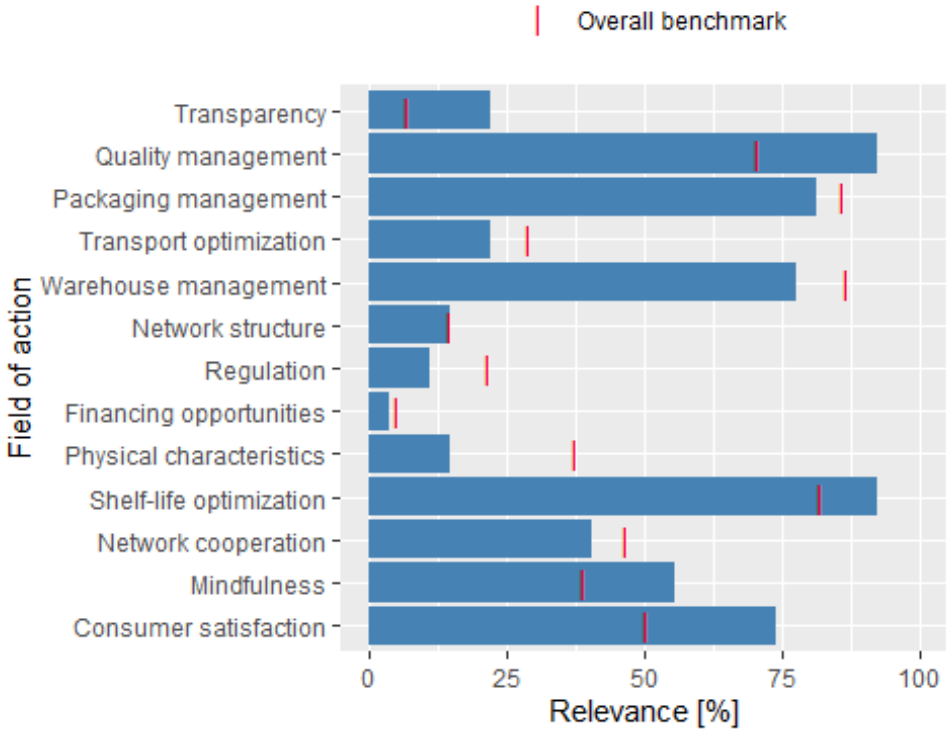
```
table_and_column_shower(`GPT_O&P_Storage`, length(`GPT_O&P_Storage`[,1
]))

## [1] 0
## [1] 52
## [1] 57
## [1] 13
## [1] 58
## [1] 13
## [1] 18
## [1] 6
## [1] 30
## [1] 56
## [1] 40
```

```
## [1] 21
## [1] 32

## New names:
## • `Field of action` -> `Field of action...1`
## • `Occurences` -> `Occurences...2`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.3`
## • `Field of action` -> `Field of action...4`
## • `Occurences` -> `Occurences...5`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.6`

ttob(myvector)
g
```


```
jpeg("GPTF_O&P_Storage.png", quality = 100, width = 15, height = 8.5,
     units = "cm", res = 300)
ready_conversion(g)
dev.off()

## png
##   2

# Calculation the n (numbers of data points) in the sub data set.
# They all contain data in the cause of loss column.
nrow(`GPT_O&P_Storage`)

## [1] 58
```

CC88 Cause of loss analysis for combination of R&T-processing

```
table_and_column_shower(`GPT_R&T_Processing`, length(`GPT_R&T_Processi
ng`[,1]))

## [1] 6
## [1] 25
## [1] 22
## [1] 6
## [1] 21
## [1] 4
## [1] 3
## [1] 1
## [1] 4
## [1] 25
## [1] 11
## [1] 15
## [1] 20

## New names:
## • `Field of action` -> `Field of action...1`
## • `Occurences` -> `Occurences...2`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.3`
## • `Field of action` -> `Field of action...4`
## • `Occurences` -> `Occurences...5`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.6`

ttob(myvector)
g
```



```
jpeg("GPTF_R&T_Processing.png", quality = 100, width = 15, height = 8.
5,
     units = "cm", res = 300)
ready_conversion(g)
dev.off()

## png
##   2

# Calculation the n (numbers of data points) in the sub data set.
# They all contain data in the cause of loss column.
nrow(`GPT_R&T_Processing`)

## [1] 27
```

CC89 Cause of loss analysis for combination of R&T-storage

```
table_and_column_shower(`GPT_R&T_Storage`, length(`GPT_R&T_Storage`[,1
]))

## [1] 1
## [1] 7
## [1] 9
## [1] 2
## [1] 11
## [1] 2
## [1] 2
## [1] 0
## [1] 3
## [1] 10
## [1] 5
## [1] 5
## [1] 6

## New names:
## • `Field of action` -> `Field of action...1`
## • `Occurences` -> `Occurences...2`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.3`
## • `Field of action` -> `Field of action...4`
## • `Occurences` -> `Occurences...5`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.6`

ttob(myvector)
g
```


```
jpeg("GPTF_R&T_Storage.png", quality = 100, width = 15, height = 8.5,
     units = "cm", res = 300)
ready_conversion(g)
dev.off()

## png
##   2

# Calculation the n (numbers of data points) in the sub data set.
# They all contain data in the cause of loss column.
nrow(`GPT_R&T_Storage`)

## [1] 11
```

CC90 Cause of loss analysis for combination of Veget.-processing

```
table_and_column_shower(GPT_Veget._Processing,
                        length(GPT_Veget._Processing[,1]))

## [1] 1
## [1] 11
## [1] 16
## [1] 5
## [1] 16
## [1] 2
## [1] 1
## [1] 0
## [1] 6
## [1] 11
## [1] 4
## [1] 5
## [1] 9
```

```
## New names:
## • `Field of action` -> `Field of action...1`
## • `Occurences` -> `Occurences...2`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.3`
## • `Field of action` -> `Field of action...4`
## • `Occurences` -> `Occurences...5`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.6`

ttob(myvector)
g
```


```
jpeg("GPTF_Veget._Processing.png", quality = 100, width = 15, height =
8.5,
     units = "cm", res = 300)
ready_conversion(g)
dev.off()

## png
##   2

# Calculation the n (numbers of data points) in the sub data set.
# They all contain data in the cause of loss column.
nrow(GPT_Veget._Processing)

## [1] 17
```

CC91 Template for visualization of NLP analysis' results

```
table_and_column_shower(GPT_Veget._Storage,
                        length(GPT_Veget._Storage[,1]))
```

```
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
```

```
## New names:
## • `Field of action` -> `Field of action...1`
## • `Occurences` -> `Occurences...2`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.3`
## • `Field of action` -> `Field of action...4`
## • `Occurences` -> `Occurences...5`
## • `Percentage of all occurences` -> `Percentage of all occurences..
.6`
```

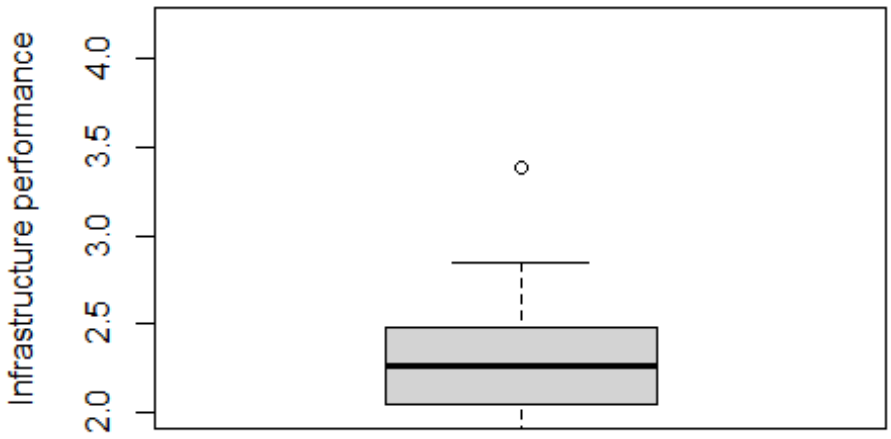
```
ttob(myvector)
g
```

```
## Warning: Removed 13 rows containing missing values (`position_stack
()`).
```

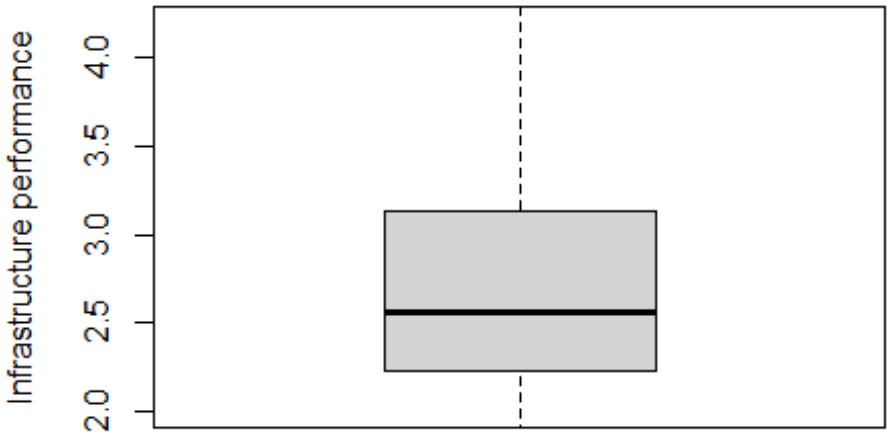
```
jpeg("GPTF_NLP_result_template.png", quality = 100, width = 15, height
= 8.5,
     units = "cm", res = 300)
ready_conversion(g)

## Warning: Removed 13 rows containing missing values (`position_stack
()`).

dev.off()

## png
##   2

# Calculation the n (numbers of data points) in the sub data set.
# They all contain data in the cause of loss column.
nrow(GPT_Veget._Storage)

## [1] 0
```

CC92 Comparing overall LPI scores in SSA with overall LPI scores word-wide

```
# Box plot of SSA only
boxplot(PD$infrastructure, ylab ="Infrastructure performance", ylim =
c(2,4.2))
```




```
# Now compared to the whole world:
boxplot(LPI_Agg_12_18$`Infrastructure...7`, ylab ="Infrastructure perf
ormance", ylim = c(2,4.2))
```


Infrastructure performance
2.0
2.5
3.0
3.5
4.0