Bachelor's Thesis

Extracting information models from mobility data

Aymen Ben Aicha

Technical University of Braunschweig
Business Information Systems
Department: Decision Support
Prof. Dr. Dirk C. Mattfeld

Supervisors:
M. Sc. Felix Spühler
First examiner: Prof. Dr. Dirk C. Mattfeld
Second examiner: Prof. Dr. Thomas S. Spengler

Statement of Originality

This thesis has been performed independently with the support of my supervisor/s. To the best of the author's knowledge, this thesis contains no material previously published or written by another person except where due reference is made in the text.

Braunschweig, June 20, 2022

_____

Abstract

Hier steht ein Text auf Deutsch.
Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Zusammenfassung

This is an english text.
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Contents

List of Tables

1. Introduction

Die Einleitung erklärt den Kontext der eigenen Arbeit und führt zur Fragestellung hin, die bearbeitet wurde. Es sollte klar werden, in welchem Bereich die Arbeit verfasst wurde und warum sie relevant ist. Im Gegensatz zum Abstract wird die Arbeit hier nicht zusammengefasst. Am Ende der Einleitung kann der Aufbau der restlichen Arbeit erläutert werden.

1.1. Description of the thema

Unterkapitel sollten ein abgeschlossenes Thema behandeln. Einzelne Unterkapitel in einem Kapitel sind zu vermeiden, also z. B. in Kapitel 2 das Unterkapitel 2.1, aber kein weiteres Unterkapitel. In diesem Fall ist es besser entweder den Inhalt von 2.1 direkt in Kapitel 2 zu schreiben, oder falls 2 und 2.1 thematisch zu weit voneinander entfernt sind, aus Unterkapitel 2.1 ein eigenes Kapitel 3 zu machen.

1.2. Motivation

Unterkapitel sollten ein abgeschlossenes Thema behandeln. Einzelne Unterkapitel in einem Kapitel sind zu vermeiden, also z. B. in Kapitel 2 das Unterkapitel 2.1, aber kein weiteres Unterkapitel. In diesem Fall ist es besser entweder den Inhalt von 2.1 direkt in Kapitel 2 zu schreiben, oder falls 2 und 2.1 thematisch zu weit voneinander entfernt sind, aus Unterkapitel 2.1 ein eigenes Kapitel 3 zu machen.

1.3. Objectives

Unterkapitel sollten ein abgeschlossenes Thema behandeln. Einzelne Unterkapitel in einem Kapitel sind zu vermeiden, also z. B. in Kapitel 2 das Unterkapitel 2.1, aber kein weiteres Unterkapitel. In diesem Fall ist es besser entweder den Inhalt von 2.1 direkt in Kapitel 2 zu schreiben, oder falls 2 und 2.1 thematisch zu weit voneinander entfernt sind, aus Unterkapitel 2.1 ein eigenes Kapitel 3 zu machen.

1.4. Structure of the work

Unterkapitel sollten ein abgeschlossenes Thema behandeln. Einzelne Unterkapitel in einem Kapitel sind zu vermeiden, also z. B. in Kapitel 2 das Unterkapitel 2.1, aber kein weiteres Unterkapitel. In diesem Fall ist es besser entweder den Inhalt von 2.1 direkt in Kapitel 2 zu schreiben, oder falls 2 und 2.1 thematisch zu weit voneinander entfernt sind, aus Unterkapitel 2.1 ein eigenes Kapitel 3 zu machen.

1.5. Methodical Approach

Unterkapitel sollten ein abgeschlossenes Thema behandeln. Einzelne Unterkapitel in einem Kapitel sind zu vermeiden, also z. B. in Kapitel 2 das Unterkapitel 2.1, aber kein weiteres Unterkapitel. In diesem Fall ist es besser entweder den Inhalt von 2.1 direkt in Kapitel 2 zu schreiben, oder falls 2 und 2.1 thematisch zu weit voneinander entfernt sind, aus Unterkapitel 2.1 ein eigenes Kapitel 3 zu machen.

2. Methodology

Now we are going to fetch the section focusing Technologies which are used in dealing with huge dataset and same keywords that are relevant to the work.

2.1. Data-Mining

It is the process of uncovering patterns and other valuable information from large datasets. It could have two main purposes:

1. Describe the target dataset.
2. Predict outcomes through the use of machine leaning algorithms.

The first part is highly relevant to our work, since we aim to extract meaningful data from large datasets.

2.2. Mobility data

This term describes data generated by activity, events or transactions using digitally-enabled mobility services or devices. It mainly reveals data related to spatial location of the performed activity such as longitude and latitude, which could be collected through smartphones or mobility vehicles at defined intervals. Other characteristics that could be provided could be related to timing(start/end of an activity), categorical data such as payments or orders and different IDs relevant to the case.

2.3. CRoss Industry Standard Process for Data Mining

CRoss Industry Standard Process for Data Mining known as CRISP-DM is an open standard process model that describes common approaches used by data mining experts.

1. As a methodology, it includes description of the typical phases of a project, tasks involved with each phase and an explanation of the relationships between tasks.
2. as a process model, it provides an overview of the data mining life cycle.

This model is highly flexible and can be easily customized. It contains six steps. The outer circle of the upcoming figure symbolizes the cyclical nature of data mining itself. The learned lessons during this process and from deployed solutions push into new business questions, which will lead either to revisiting previous steps in the same project, or to new knowledge for upcoming projects.

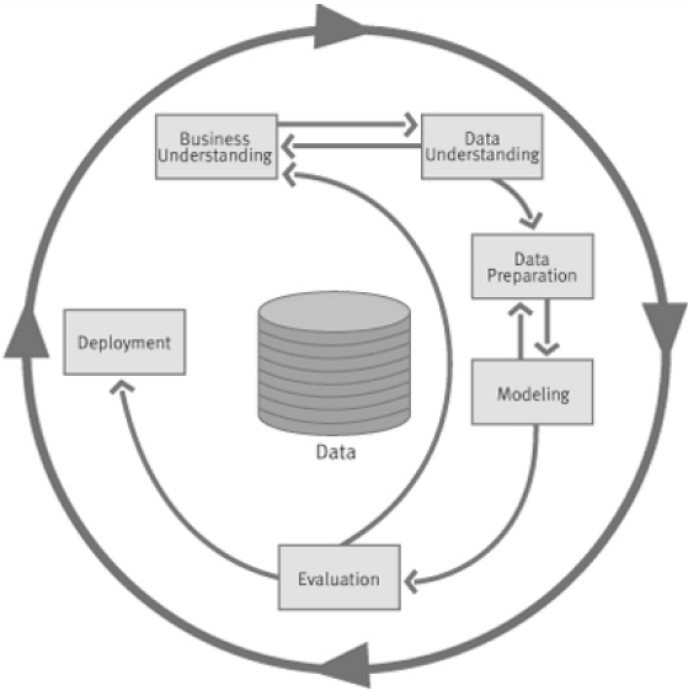

Figure 2.1.: CRISP-DM: The process model

The upcoming sections are mainly about the different steps of the model, definition of each step as well as its tasks and outputs.A generic task is the general level for tasks which should be:

1. complete: covering the whole data mining process as all possible applications
2. stable: valid for yet unforeseen techniques

2.3.1. Business understanding

This initial phase consists in describing what the customer really wants to accomplish from a business view, by receiving his requirements and converting all possible and accessible clients preferences into data mining goals. After having a less vague vision, we will proceed to creating a preliminary plan to achieve the settled goals.

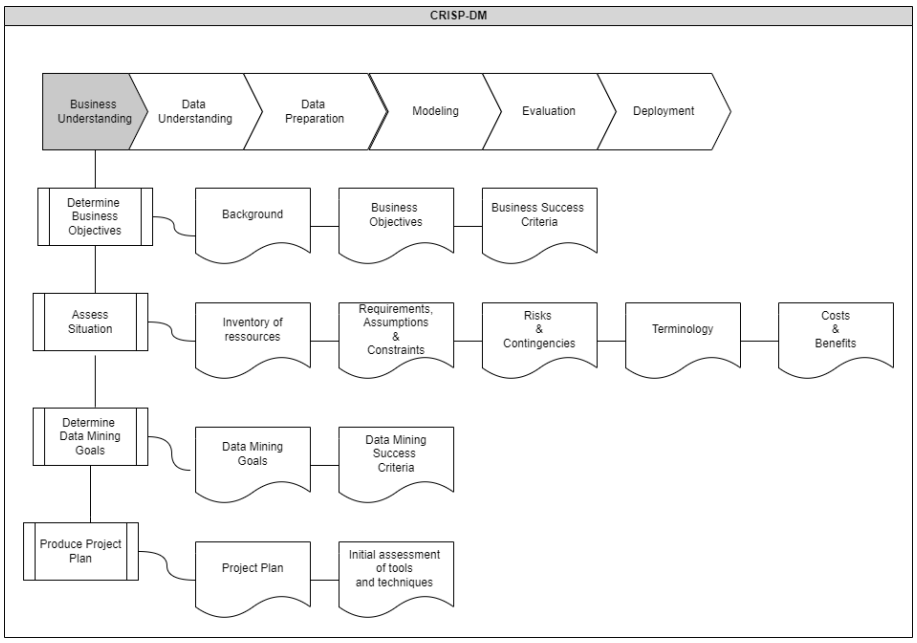

Figure 2.2.: CRISP-DM: Business Understanding

2.3.1.1. Task: Determine Business objectives

Goal: Understand, from a business perspective, what who the client is, his expectations from the work and generally his settled goals. All important factors, that can influence the outcome of the project should be uncovered in this step.

a) Output: Background
The New York City Taxi and Limousine Commission (TLC) licenses and 55 regulates taxis, rental vehicles, commuter vans and transit vehicles. Since this company, and others like it, are always looking to reduce waiting times, as well as avoid congested paths, they therefore face problems that could be categorized under the vehicle routing problem (VRP). To facilitate their daily life, we will process their open-source data by conducting a data mining process, aiming to extract smaller databases, which could be used depending on the desired use case.

b) Output: Business Objectives
The aim will be to answer client's major questions:

- How does pickup/drop-off time & area and number of orders affect the rentability of the firm?

- How to produce a better customer experience?
- Is the predicted pickup drop out and selected paths, after using clustered treated data, better than concurrent?

c) Output: Business success criteria

- Small flexible datasets, allowing a quick data understanding
- Clearly defined clusters.
- Datasets cleaned containing the wished attributes, ready to be used in VRP-Algorithms.

2.3.1.2. Task: Assess situation

Goal: Investigation of the available resources; data, software programs and other factors, that affect directly the project

a) Output: Inventory of resources

* Available data: We deal with 4 datasets, related to:
  i. Yellow Taxi Trip Records (available from 01.2009)
  ii. Green Taxi Trip Records (available from 08.2013)
  iii. For-Hire Vehicle Trip Records(available from 01.2016)
  iv. High Volume For-Hire Vehicle Trip Records (available from 02.2019)
* Taxi Zone Maps and Lookup Tables
  i. Taxi Zone Lookup Table (CSV): contains a list of TLC taxi zone location IDs, location names and corresponding boroughs of each zone.
  ii. Taxi Zone Shapefile (CSV): contains geographic information of each taxi zone
  iii. Taxi Zone Map Bronx (JPG)
  iv. Taxi Zone Map Brooklyn (JPG)
  v. Taxi Zone Map Manhattan (JPG)
  vi. Taxi Zone Map Queens (JPG)
  vii. Taxi Zone Map Staten Island (JPG)
* Programming languages: Use of python and Jupyter Notebooks

b) Output: Requirements, Assumptions and constraints

* Competition schedule: Gantt-Digramm
* Working process should be on CRISP-DM oriented
* Data mining steps should be respected (data cleaning, data merging etc)
* Legal & security issues: use of open-source data

c) Output: Risks and Contingencies

* If the provided data is manipulated or classified as poor, problems related to data understanding can come into question. To avoid this kind of problems, we will try to select ideal data (such as a data of one year and one type of taxis) so we avoid all possible outliers and extremes as well as incompatible attributes definition.
* Since we don't have a direct communication channel to the data provider as well as the business itself, we will be assuming that the asked questions are generally to VRP problems, and that their answers will remain under standardized fulfillment of data mining goals as well as CRISP-DM models.

2.3.1.3. Task: Determine Data-Mining Goals

Goal: Projecting expected outputs of the project, while defining general success criteria.

a) Output: Data-Mining Goals
   * Define different data relations(relation between number of travelers and fare amount, geographic location effect etc...)
   * Identify Rush hours and Hotspots

b) Output: Data-Mining success criteria
   * Data understanding and creation of templates for a quick data analysis
   * Scenario development and therefore extraction of correspondent datasets (clustering)

2.3.1.4. Task: Produce project plan

Goal: Listing the detailed project plan, including milestones and risks, in a Gantt-Diagramm.

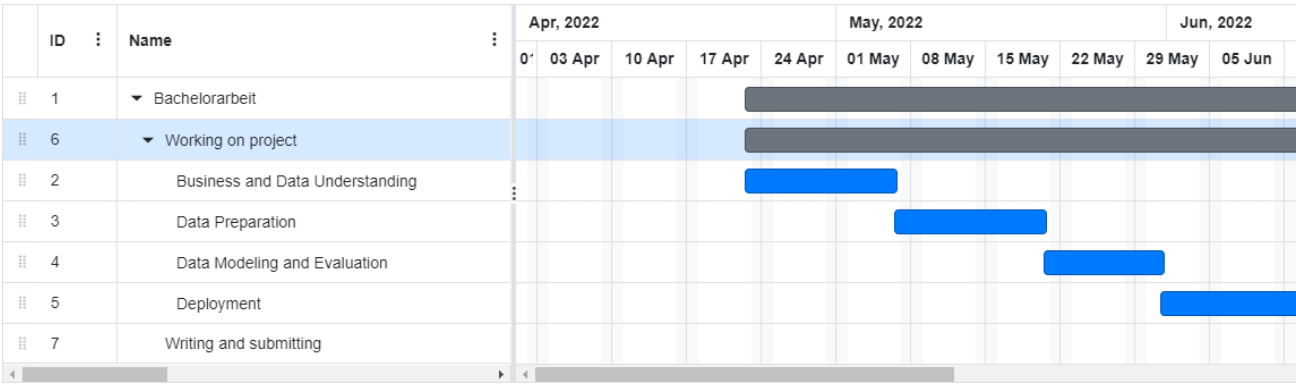

Figure 2.3.: Gantt-Diagramm: Project plan

2.3.2. Data understanding

The Data understanding phase begins with collecting data in its raw form, getting familiar with it, rating data in terms of quality. The nutshell of this step, will be to gather some attractive information and subsets from the original data.

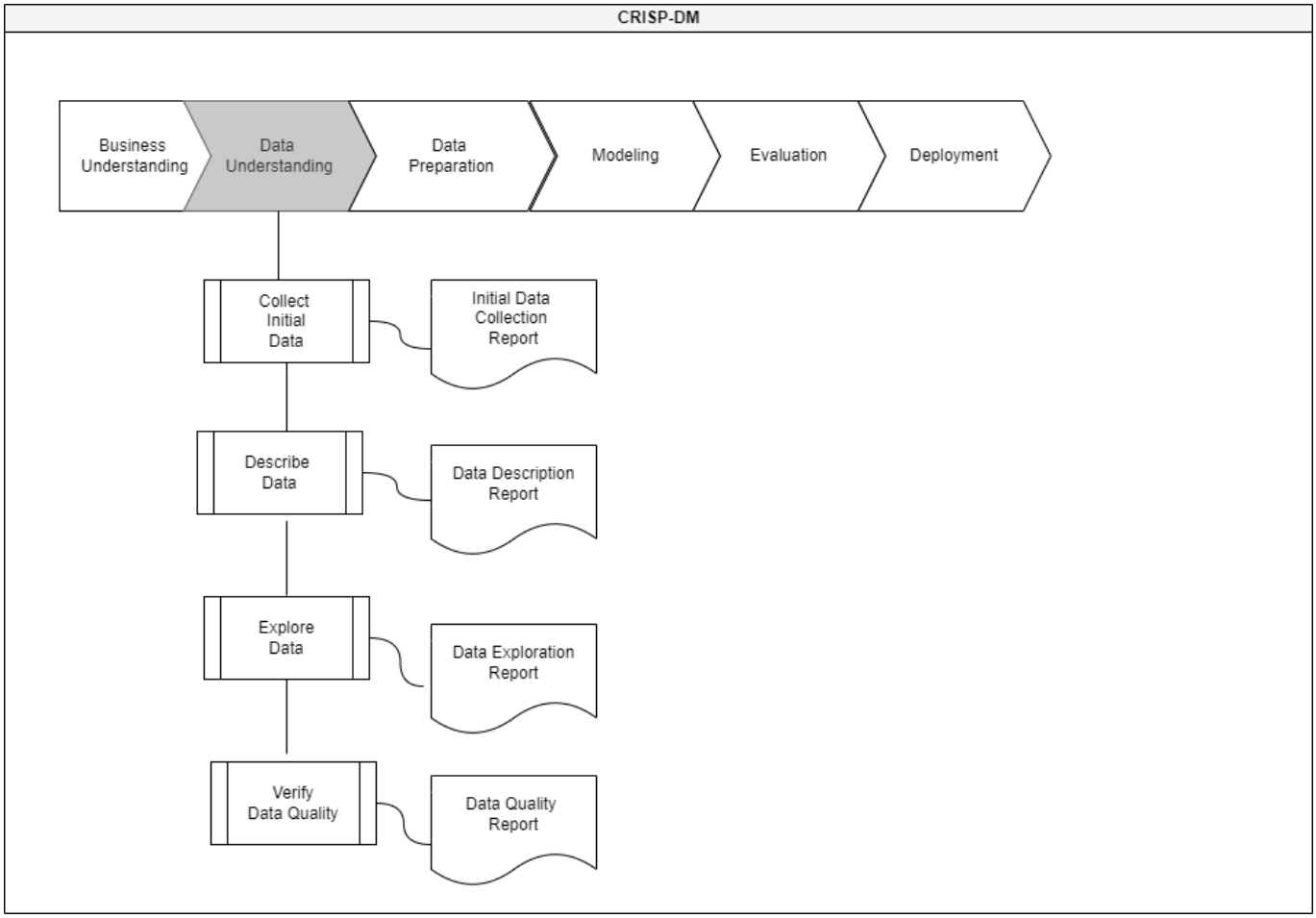

Figure 2.4.: CRISP-DM: Data Understanding

2.3.2.1. Task: Collect initial data

Goal: Loading data from project resources and then the first steps of data preprocessing.

a) Output: Initial data collection report

* Getting data: To avoid extra work and memory consumption, we will work with API-Queries to avoid waste of time. Here we will define what are the important libraries to be imported and how it could be done.

- SODA API: The Socrata Open Data API (SODA) provides programmatic access to datasets, including the ability to filter, query, and aggregate data.
- SOCRATA APP Token: All requests should include an app token that identifies the application, and each application should have its own unique app token. For further references, see `https://dev.socrata.com/foundry/data.cityofnewyork.us/t29m-gskq`
- Available data (see 2.2.1.2 Inventory of resources)

* Selection of data: Here we will be limiting our work into the data related to Yellow Taxi Trip Records in year 2021, so we could take out the most of this data.

2.3.2.2. Task: Describe data

Goal: Examination of the surface proprieties of acquired data.

a) Volumetric analysis of data:

Data Dictionary – Yellow Taxi Trip Records May 1, 2018 Page 1 of 1

This data dictionary describes yellow taxi trip data. For a dictionary describing green taxi data, or a map of the TLC Taxi Zones, please visit http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

Field Name	Description
VendorID	A code indicating the TPEP provider that provided the record. **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.**
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
RateCodeID	The final rate code in effect at the end of the trip. **1= Standard rate** **2=JFK** **3=Newark** **4=Nassau or Westchester** **5=Negotiated fare** **6=Group ride**

Figure 2.5.: Data dictionary

Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. **Y= store and forward trip** **N= not a store and forward trip**
Payment_type	A numeric code signifying how the passenger paid for the trip. **1= Credit card** **2= Cash** **3= No charge** **4= Dispute** **5= Unknown** **6= Voided trip**
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.

Figure 2.6.: Data dictionary

The data has originally 18 columns, having non-null values and being defined as objects. We will categorize it into the following classifications:

a) IDs: (vendorid, pulocationid, dolocationid, ratecodeid)

b) datetimes(tpep_pickup_datetime, tpep_dropoff_datetime)

c) different amounts related to the trip(fareamout, extra, mta tax, tip amount, tolls amount, improvement surcharge, total amount, congestion surcharge)

d) categorical variables (store and fwd flag, payment type)

In the next step, we will assign the correspondent data type to the attributes.
After assigning the correspondent data type to its attributes, the attributes will have this form after modifications:
dtypes: datetime64(2), float64(9), int64(4), object(3).

To describe the data, the following table presents some insides from the data; its range and some significant indexes.

This table shows us that some data has a nonsense, for example a negative fare amount or a passenger count of 0 people.This table shows us that some data has a nonsense, for example a negative fare amount or a passenger count of 0 people.This table shows us that some data has a nonsense, for example a negative fare amount or a passenger count of 0 people. This table shows us that some data has a nonsense, for example a negative fare amount or a passenger count of 0 people.

	vendorid	passenger_count	trip_distance	pulocationid	dolocationid	fare_amount	extra
count	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000
mean	1.678007	1.415547	2.650791	166.082827	163.277403	11.123815	0.969759
std	0.467241	1.063357	3.522187	66.616367	71.495143	12.732147	1.220303
min	1.000000	0.000000	0.000000	1.000000	1.000000	-250.500000	-5.500000
25%	1.000000	1.000000	0.990000	132.000000	107.000000	6.000000	0.000000
50%	2.000000	1.000000	1.620000	162.000000	162.000000	8.000000	0.500000
75%	2.000000	1.000000	2.800000	236.000000	236.000000	12.000000	2.500000
max	2.000000	8.000000	427.700000	265.000000	265.000000	6960.500000	7.000000

mta_tax	tip_amount	tolls_amount	improvement_surcharge	total_amount	congestion_surcharge
1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000
0.492801	1.943094	2.231518	0.296702	16.535093	2.231518
0.078289	2.752124	0.808159	0.043860	14.803632	0.808159
-0.500000	-100.000000	-2.500000	-0.300000	-252.300000	-2.500000
0.500000	0.000000	2.500000	0.300000	10.560000	2.500000
0.500000	1.850000	2.500000	0.300000	13.500000	2.500000
0.500000	2.700000	2.500000	0.300000	17.850000	2.500000
0.500000	1140.440000	2.500000	0.300000	7661.280000	2.500000

Figure 2.7.: Dataset key values

2.3.2.3. Task: Explore Data

Goal: Establishment of graphics and reports as well as queries to directly tackle data mining goals

In this step, we will work on different attributes, their relations between each other in order to discover some data insides.

1. Correlation analysis:
   This diagram shows correlations between different attributes, and it will be used later on to compare first raw data to processed results.

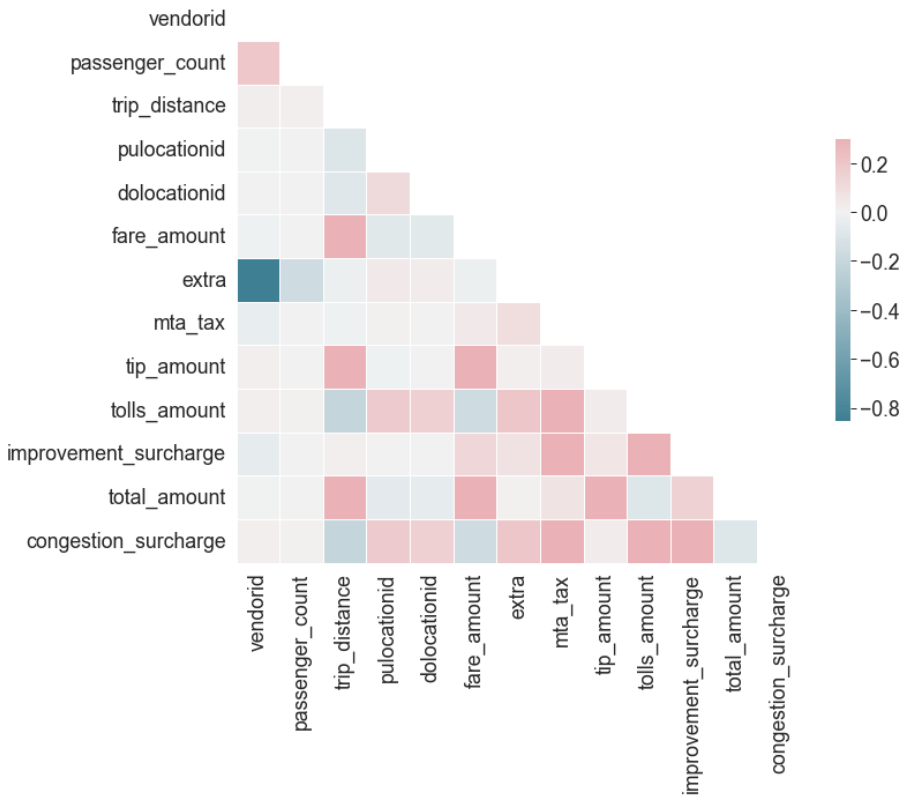

Figure 2.8.: Attributes correlations

2. Temporal analysis:
   The available pickup and drop off datetimes are written in the form of YYYYMMDD HH:MM:SS.
   First we will consider classifying timestamps into days, so we get a general view about the distribution.

Day	Pickup Day	Drop-off day
Monday	153016	153213
Tuesday	136467	136479
Wednesday	138784	138801
Thursday	145051	144926
Friday	168621	168374
Saturday	146171	146185
Sunday	111890	112022

Hereby we observe a nearly equal distribution in matter of days, and it would make more sense to consider other time slots.

Thus, we will define 4 time intervals, The graphic above, after defining four time slots;

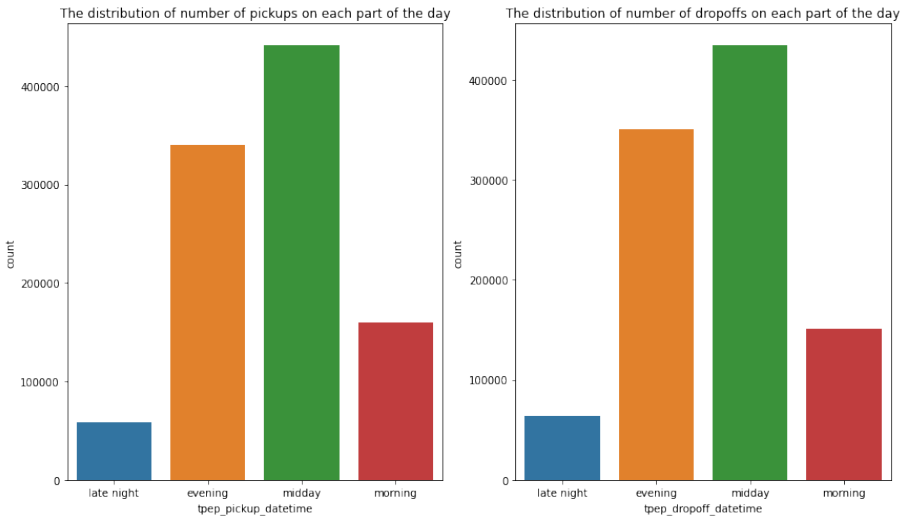

Figure 2.9.: Data set key values

morning (4 hrs to 10 hrs), midday (10 hrs to 16 hrs), evening (16 hrs to 22 hrs) and late night (22 hrs to 4 hrs).
The attractive output, that a less demand is observable for morning and late night trips. To defend this statement, we will consider an hourly distribution of the trips in the following graph.

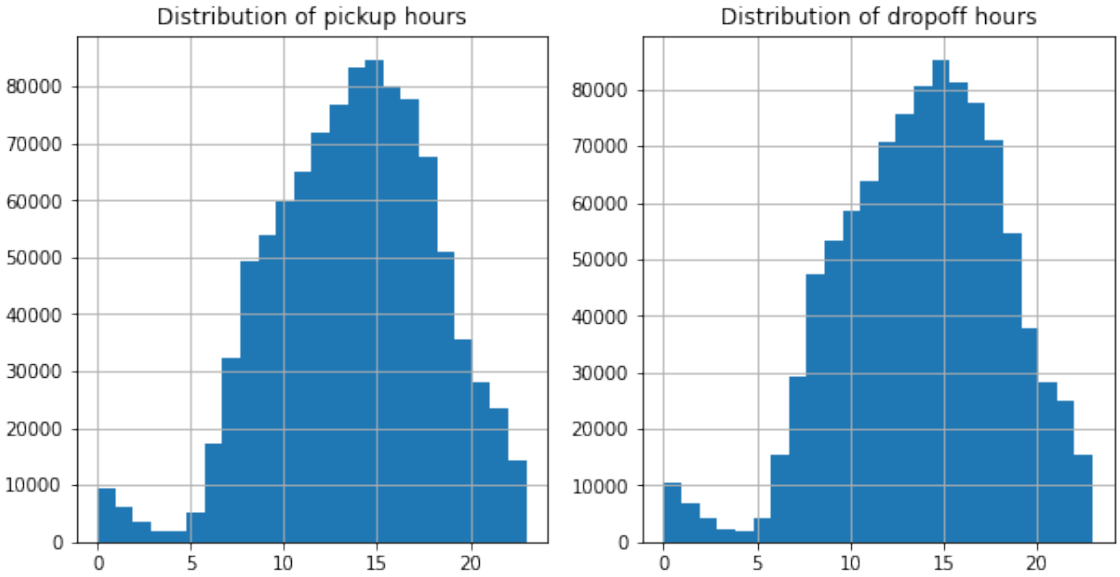

Figure 2.10.: hourly distribution

The hourly distribution confirms the statement related to Figure 2.8, that starting from 10 PM, the need for night rides continuously decreasing until 5 AM. It will therefore increase to achieve its peak at 16 PM.

3. Duration analysis:
   The duration of each trip is calculated based on the difference between pick-ups and drop-offs.
   In the following Diagram. the relation between pick-up/drop-off day and the trip's duration will be taken into consideration.

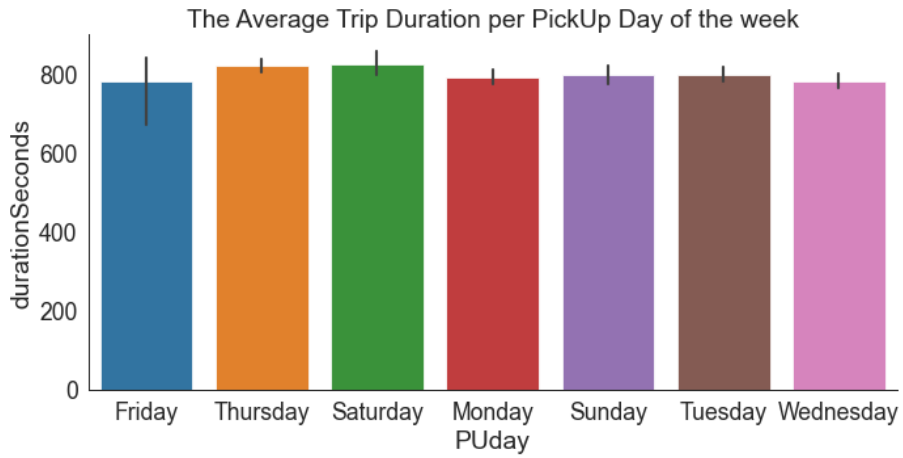

Figure 2.11.: The Average Trip Duration per Pickup Day of the week

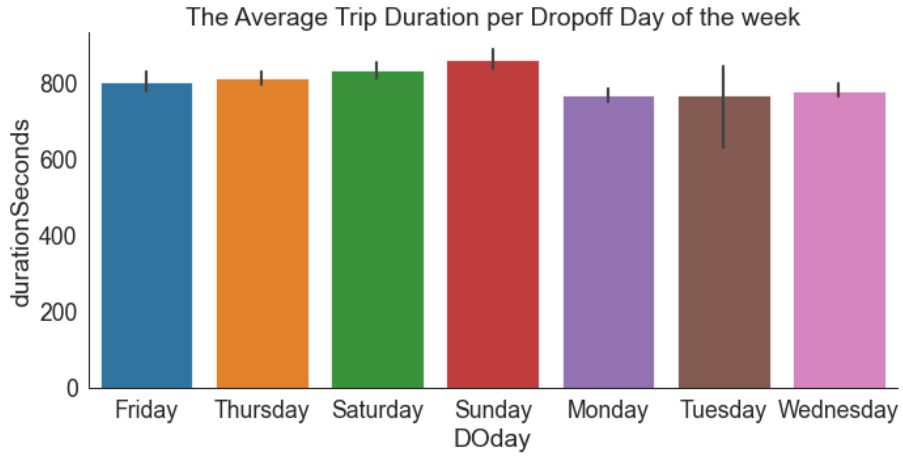

Figure 2.12.: The Average Trip Duration per Drop-off Day of the week

The graphs denote the average estimate of a trip for each day of the week. The error bars provide some indication of the uncertainty around that estimate
To go deeper into our duration/timing analysis, the same principle will be maintained to produce charts highlighting trips lengths in relation with hourly distribution

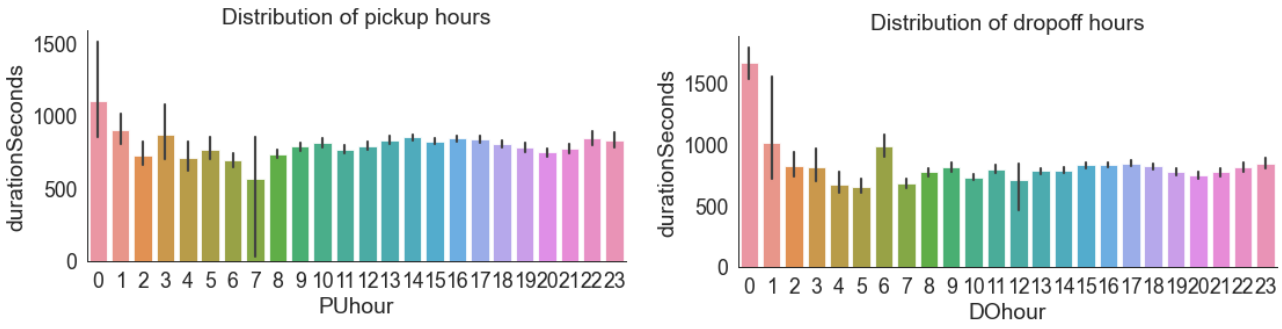

Figure 2.13.: Duration's hourly distribution

All generated charts push us into the following conclusion, which matches analysis done with temporal analysis. To take with, is that weekends are the most attractive days for long-lasting trips, and that late nights starting from 22 o'clock, are the most favorable for the trips. To defend this statement, the following concluding chart is available:

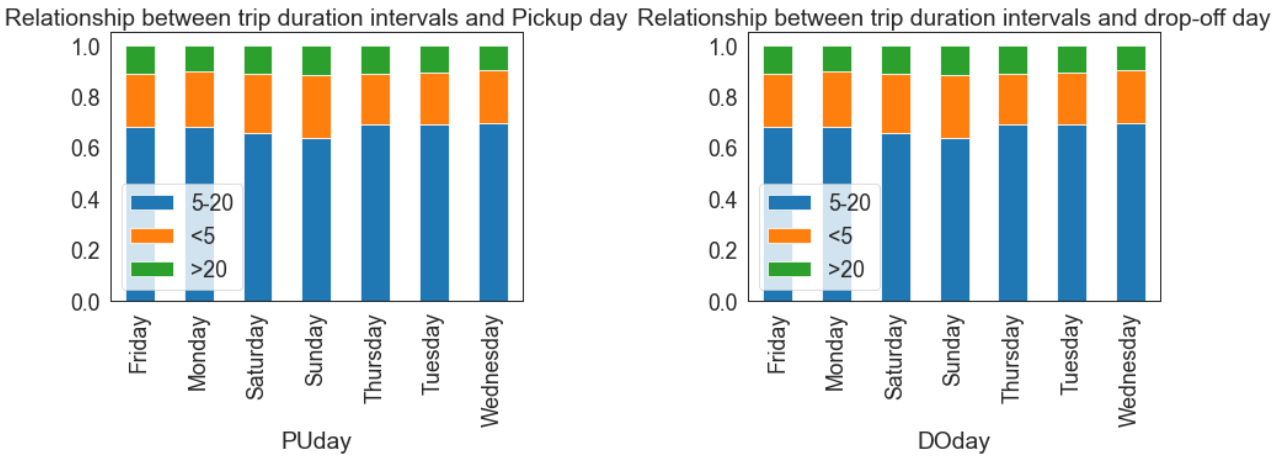

Figure 2.14.: Relationship between trip duration intervals and Pickup/drop-off day

It is clearly seen, that trip lasting 5 to 20 minutes are the more attractive, regardless of the day, and that the longer lasting trips are on weekends, with relatively small difference. Not only that, but on Sundays the demand on short lasting trips is also higher.

4. Geospatial analysis:
   If we want to have a general view about the pickup and drop-off heatmaps, we will have the following view:

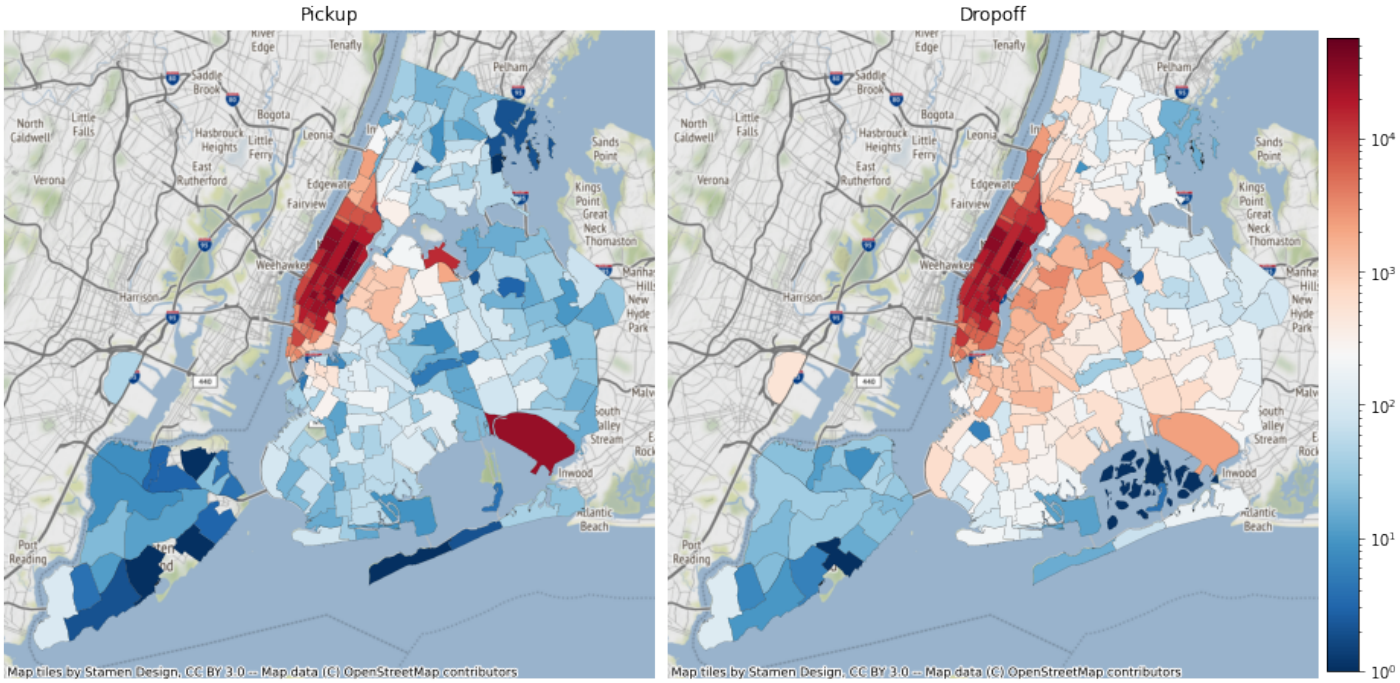

Figure 2.15.: correlation between continuous attributes

To better understand the following map, we need to have a step back, and look and a bigger scale, and we will try to identify the heat zones, colorized in red in our map.

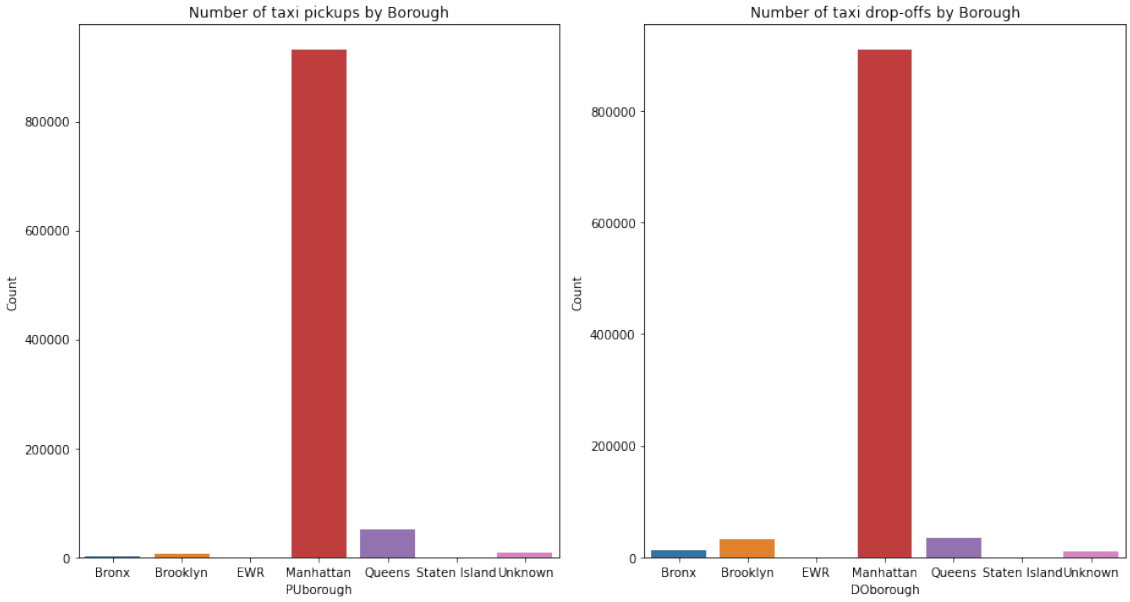

Figure 2.16.: Borough distribution of drop-offs and pickups

The attractive pickups as well as drop-offs are mainly in Manhattan, highly far from other places, which was already marked in red in our map.

5. Passenger and trip distance

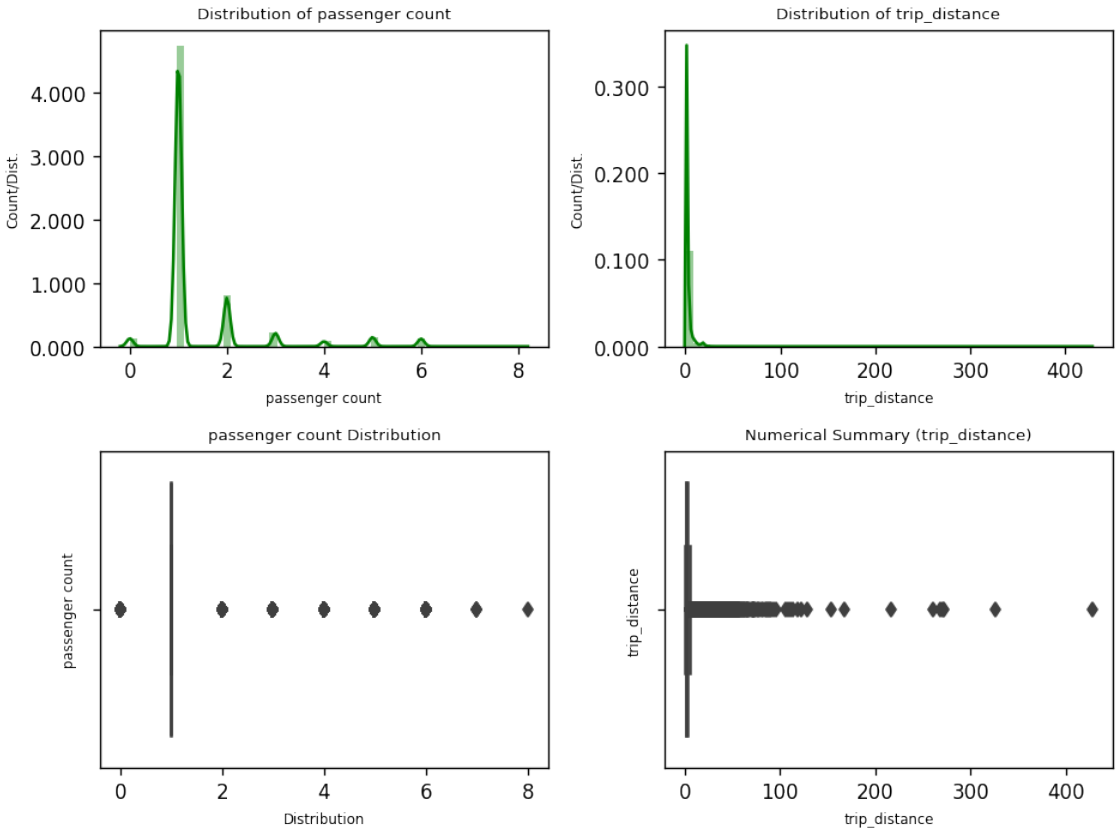

Figure 2.17.: Dataset key values

Potentially more analysis/charts after better understanding data relations

6. relation trip distance and trip duration

	mean		len		std	
	trip_distance	trip_duration	trip_distance	trip_duration	trip_distance	trip_duration
passenger_count						
0	2.596416	639.036098	21109	21109	4.049786	548.608006
1	2.604687	767.190109	757518	757518	3.456806	10060.707954
2	2.850129	853.577426	128529	128529	3.771701	3897.557229
3	2.807030	927.068583	35213	35213	3.603606	4634.557306
4	2.920736	915.283589	13107	13107	3.794537	4532.842098
5	2.757648	1274.938175	24246	24246	3.547016	7208.718941
6	2.590537	925.338874	20273	20273	3.282768	4905.403388
7	13.917500	1126.500000	4	4	10.246357	766.798322
8	0.000000	318.000000	1	1	nan	nan

Figure 2.18.: Different manifestation of the relation between passenger count and the couple trip distance & trip duration

For 7 passengers, the mean of the trip's distance will be at its highest in our sample, thus a higher mean of trip duration, that lasts approximately 19.5 minutes for a 13.9Km ride. The highest mean of a trip's duration is at 21 Minutes and 15 seconds. The values that are mostly appearing, are passengers going alone for different trips, causing the highest standard deviation, at the trip's duration level.

2.3.2.4. Task: Verify data quality

Goal: Examination of the data quality by addressing some questions:

* Is there any null or missed values?
  The considered dataset, does not contain null values, is complete; having values for every attribute, have up to 260 location IDs, a big range of timestamps, up to 3423 different trip distances as well as 5873 trip amounts.

vendorid	2	vendorid	0
tpep_pickup_datetime	711581	tpep_pickup_datetime	0
tpep_dropoff_datetime	711760	tpep_dropoff_datetime	0
passenger_count	9	passenger_count	0
trip_distance	3423	trip_distance	0
ratecodeid	7	ratecodeid	0
store_and_fwd_flag	2	store_and_fwd_flag	0
pulocationid	255	pulocationid	0
dolocationid	260	dolocationid	0
payment_type	4	payment_type	0
fare_amount	1149	fare_amount	0
extra	19	extra	0
mta_tax	3	mta_tax	0
tip_amount	1922	tip_amount	0
tolls_amount	4	tolls_amount	0
improvement_surcharge	3	improvement_surcharge	0
total_amount	5873	total_amount	0
congestion_surcharge	4	congestion_surcharge	0
dtype: int64		dtype: int64	

Figure 2.19.: Data Quality

* Z-Score (also called a standard score) gives an idea of how far from the mean a data point is. But more technically, its a measure of how many standard deviations below or above the population mean a raw score is.
  If we consider a Z more than +3 standard deviation units away from the mean, since we did not find data on the left side of 0, thereby we will have the following output:
  a) in relation with fare amount:
     0.02 % of the data is placed on above 3 standard deviations from the mean
  b) in relation with trip distance:
     approximately 0.02 % is also in the same area.
  c) regarding passenger count:
     approximately 0.05 % belongs to the same area

2.4. Data preparation

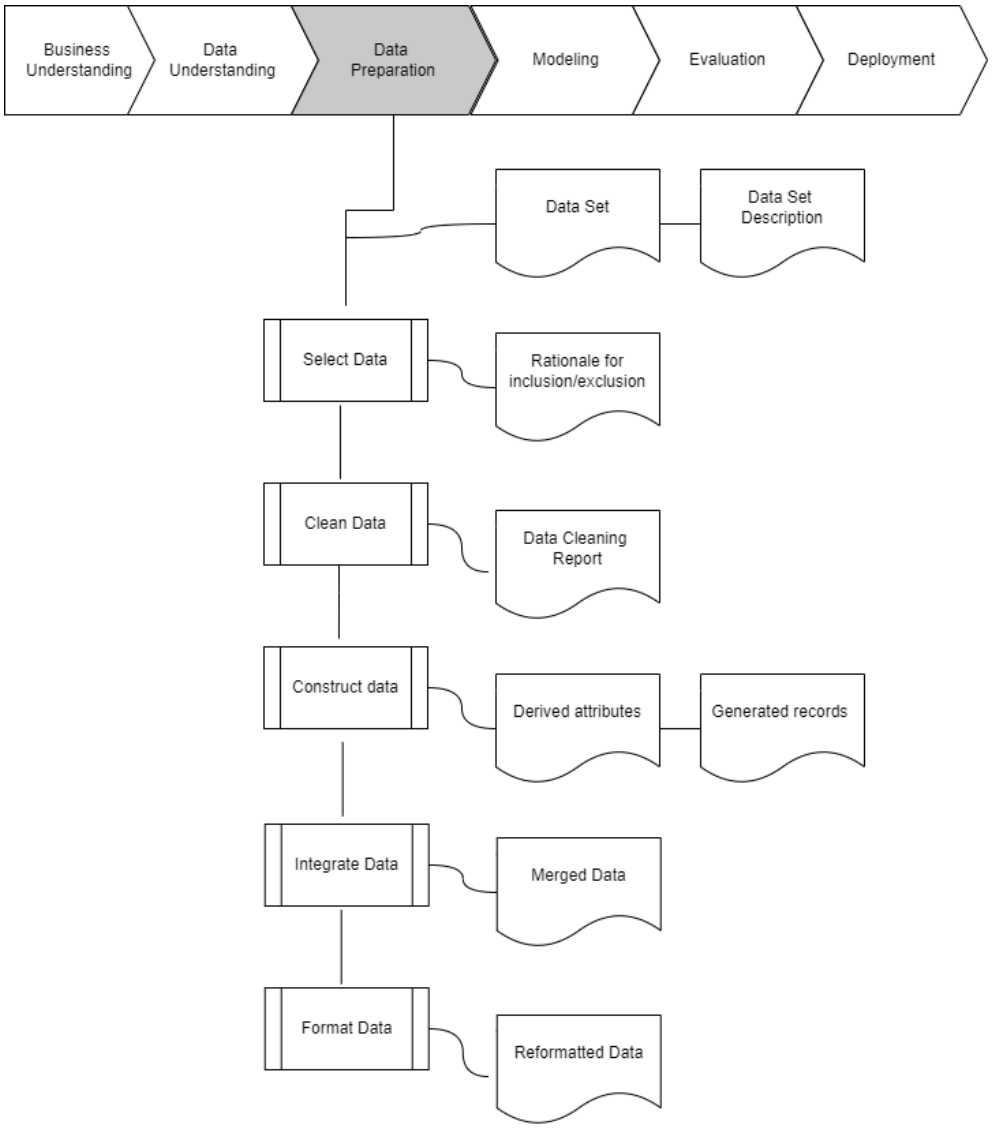

Figure 2.20.: CRISP-DM: Data preparation

In this step, we will only use the data that we already put into test (task no.1 will not be described here), try to clean it and make it ready to be used.

2.4.1. Data cleaning:

- Dropping columns:
  First, we will drop the following columns since we don't them and are irrelevant for the work. "ratecodeid","store and fwd flag","extra","mta tax","tolls amount","improvement surcharge","congestion surcharge","fare amount","tip amount"

2. Data cleansing:
   Records with implausible values or errors are removed based of the conducted analysis related to data distribution, common sense to ensure data's correctness and avoid all possible derivations.
   - Passenger count is one passenger or more, but it shouldn't exceed 6 riders since it's legally not allowed
   - Trip distance should be greater than 0 km but also less than 100 km to avoid outliers.
   - The trip's total amount should be strictly more than 2.5 USD and less than 100 USD in our example.
   - Pick-up/Drop-off location ID should be within the range of [1;263]
   - trip duration should be more than a minute but also less than 4 hours.

   fyi: here we will introduce the development of the dataset after removing each of the items mentioned above

2.4.2. Construct Data:

Since we will dig deep into our data, we will need to create some useful data, extracted from available columns. As a part of data construction, we will use the Feature engineering, which is a process of transforming raw data into useful features to get the most out of the data. We will identify:

- Duration of trips as a result of the difference between pick-up and drop-off time. The results will be in seconds, and after that the duration interval is to be defined. Here we consider 3 time interval of the trip's duration; less than 5 minutes, from 5 to 2s 0 and more than 20 minutes.
- Months, day of the week, hour and the period of the related pick-up and drop-off
- Borough of each pick-up and drop-off location

2.4.3. Integrate and format data:

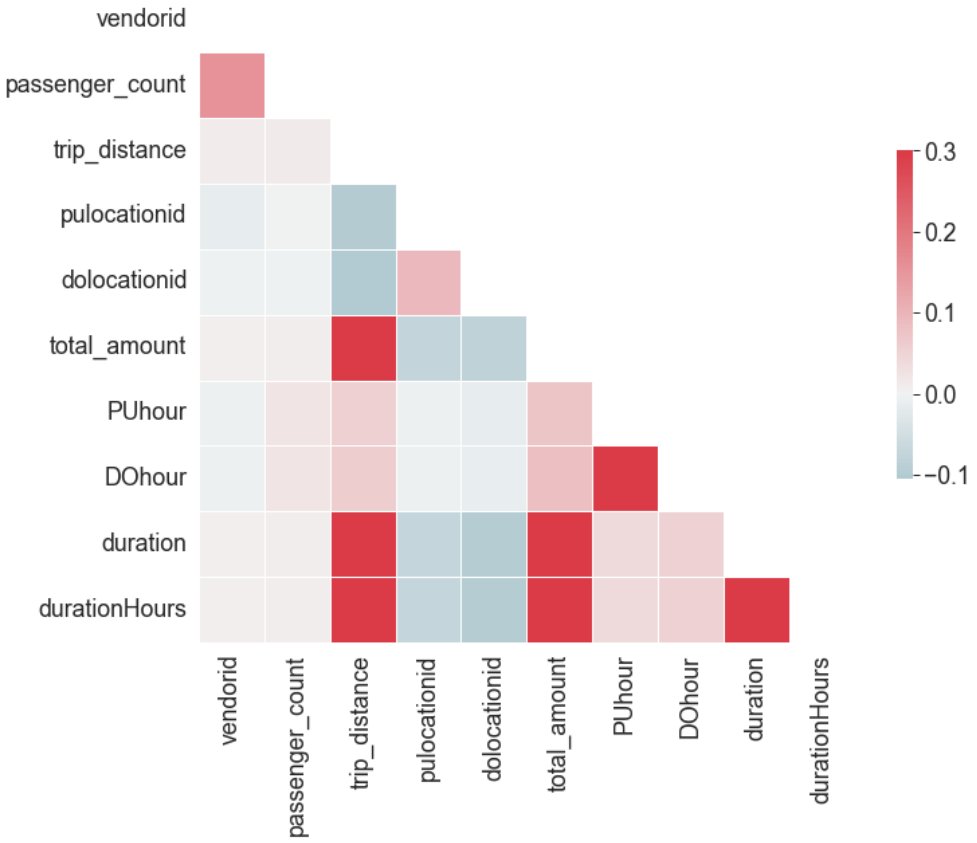

Figure 2.21.: Data correlation after treatments

It is obvious that data, extracted from each other, will always be highly correlated. But here we observe a higher correlation than before between total amount and duration from one hand, and the trip distance from the other hand. PUlocation and DOlocation are also related to each other, as well as PUhour and DOhour. The total trip's amount is related to pickup and drop-off time and location, and affects the trip's total amount.

2.5. Modeling

In der Informatik sind die häufigsten Grafiken entweder Diagramme oder Plots. Beide Arten von Grafiken lassen sich gut als Vektorgrafiken erstellen und einbinden. Der Vorteil von Vektor- gegenüber Pixelgrafiken ist, dass beliebig weit in eine Grafik hereingezoomt werden kann, ohne dass sie unscharf wird. Zudem benötigen Vektorgrafiken meistens weniger Speicherplatz.

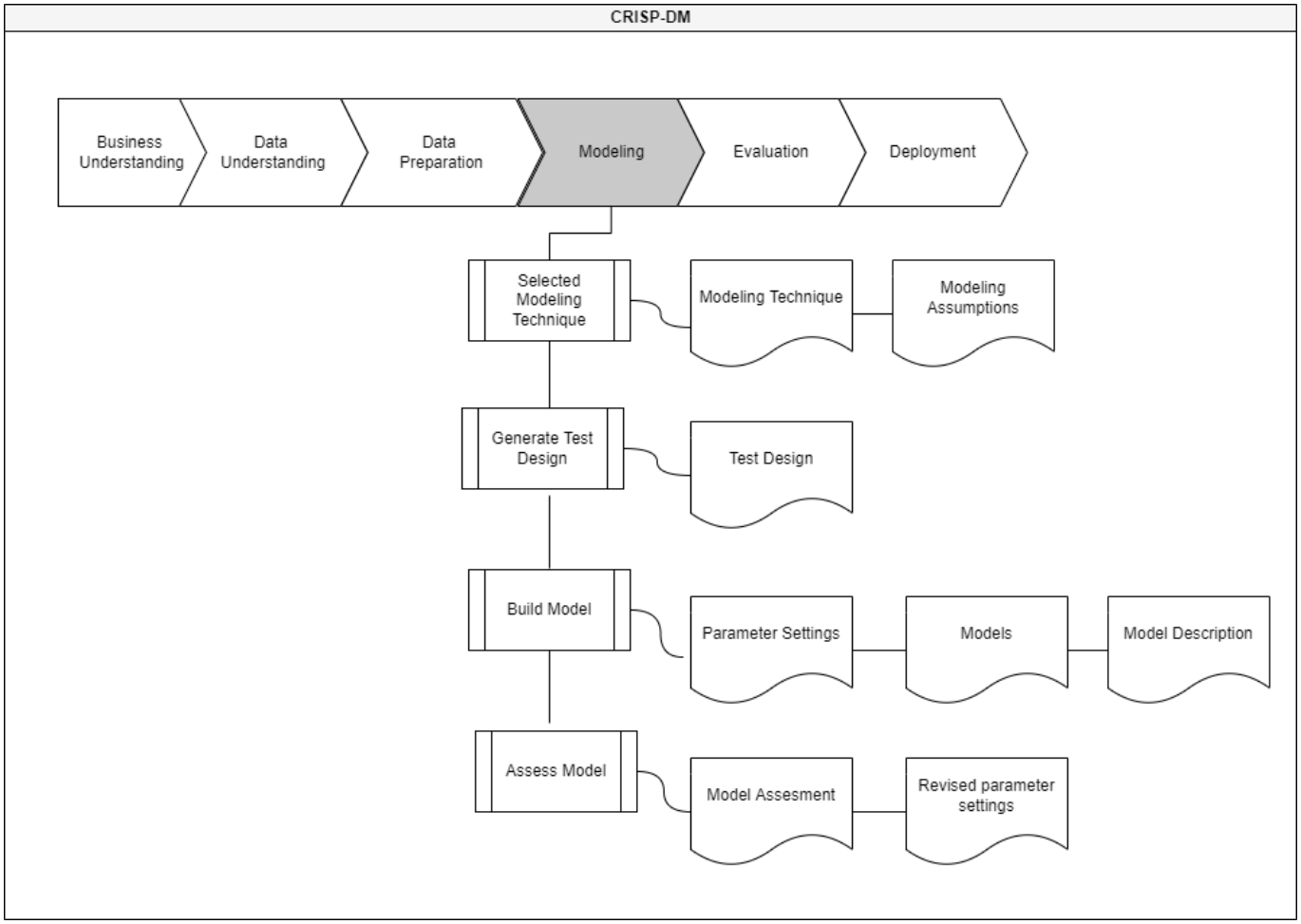

Figure 2.22.: CRISP-DM: Modeling

2.6. Evaluation

In der Informatik sind die häufigsten Grafiken entweder Diagramme oder Plots. Beide Arten von Grafiken lassen sich gut als Vektorgrafiken erstellen und einbinden. Der Vorteil von Vektor- gegenüber Pixelgrafiken ist, dass beliebig weit in eine Grafik hereingezoomt werden kann, ohne dass sie unscharf wird. Zudem benötigen Vektorgrafiken meistens weniger Speicherplatz.

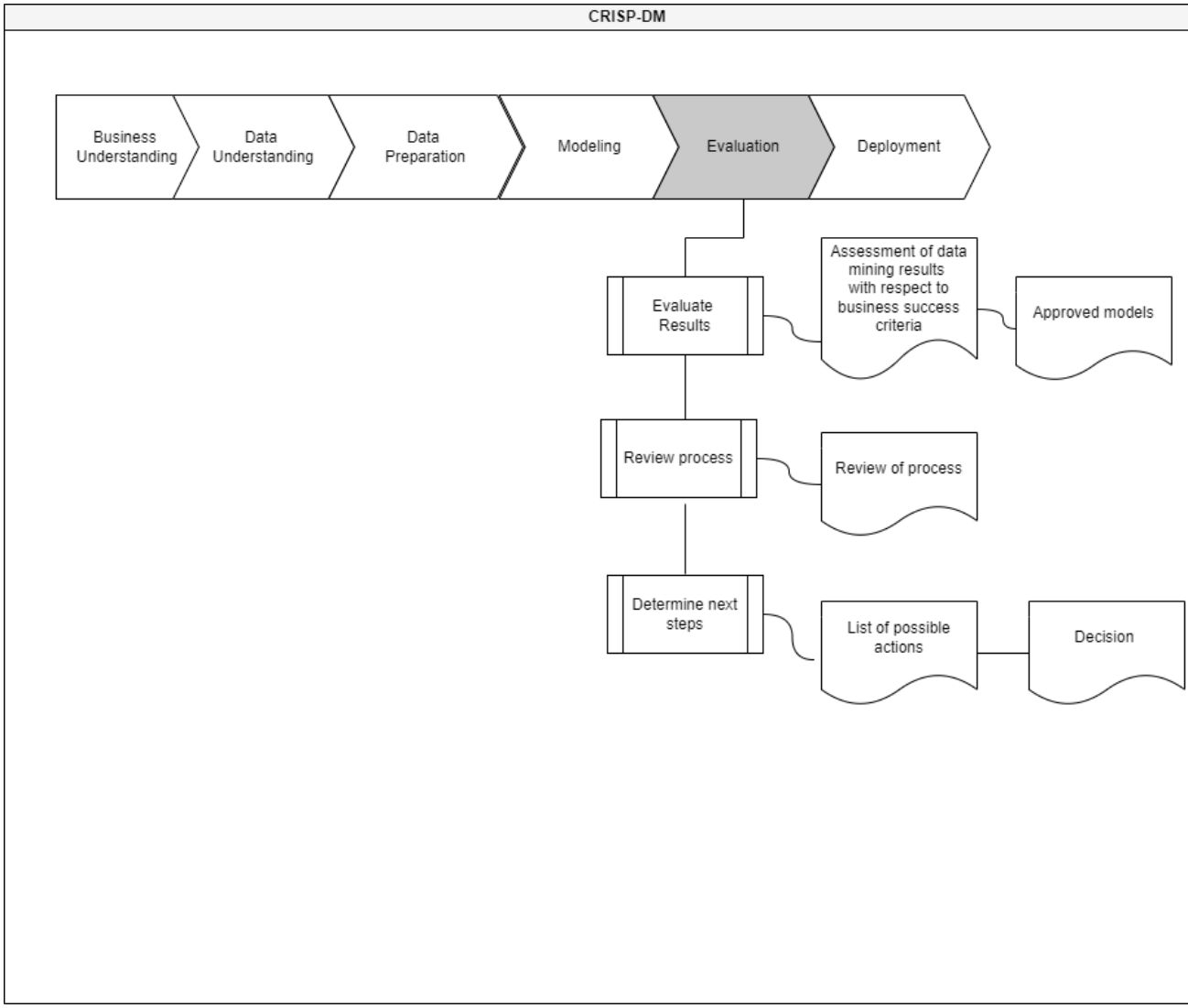

Figure 2.23.: CRISP-DM: Evaluation

Grafiken können direkt in LaTeX mit dem TikZ Paket tikz erstellt werden. Die Verwendung ist etwas gewöhnungsbedürftig, da Grafiken mit Code beschrieben werden, bietet aber viele Freiheiten. Außerdem werden die so erstellten Grafiken direkt in LaTeX gerendert und verwenden die selbe Schriftart wie im Text und eine konsistente Schriftgröße im gesamten Dokument. Ein weiteres beliebtes Programm zum Erstellen von Vektorgrafiken ist Inkscape inkscape. Zudem bieten viele Programme die Möglichkeit eine Grafik z. B. als PDF zu exportieren, was in LaTeX als Vektorgrafik eingebunden werden kann.

Pixelgrafiken lassen sich nicht immer vermeiden, z. B. wenn eine Foto in die Arbeit eingebunden werden soll. In diesem Fall sollte darauf geachtet werden, dass die Grafik über eine ausreichende Auflösung verfügt. Eine Auflösung von 300 dpi ist ein guter Richtwert, um beim Drucken ein gutes Ergebnis zu erhalten.

Abbildungen ?? und ?? zeigen beide den Aufbau des IEEE Floating Point Formats. Abbidlung ?? ist eine Pixelgrafik, während Abbildung ?? mit TikZ erstellt wurde. Der Unterschied wird beim hereinzoomen deutlich.

3. Experiments and Results

3.1. Tables

Tabellen können in LaTeX direkt erstellt werden. Tabelle 3.1 zeigt ein Beispiel dafür. Einfache Tabellen lassen sich schnell erstellen, bei komplizierteren Tabellen ist es manchmal einfacher zusätzliche Pakete zu verwenden. Mit dem Paket multirow multirow können z. B. einfacher Tabellen erstellt werden, bei denen einzelne Zeilen oder Spalten zusammengefasst sind.

Parameter	binary16	binary32	binary64	binary128
k, storage width in bits	16	32	64	128
w, exponent field width in bits	5	8	11	15
t, significand field width in bits	10	23	52	112
emax, maximum exponent e	15	127	1023	16383
bias, $E - e$	15	127	1023	16383

Table 3.1.: IEEE 754-2019 Floating Point Formate als Beispiel für das Einbinden einer Tabelle.

3.2. Source code

Um Quellcode in die Arbeit einzubinden, können in LaTeX Listings verwendet werden. Es gibt für populäre Sprachen vorgefertigte Umgebungen, welche die Syntax farblich hervorheben. Quellcode sollte eingebunden werden, wenn eine konkrete Implementierung in einer Sprache erläutert wird. Für die Erklärung eines Algorithmus ist es oft übersichtlicher ein Schaubild oder Pseudocode zu verwenden. Es sollten nur kurze Codeabschnitte eingebunden werden, die für den Leser einfach nachvollziehbar sind und nur den für die Erklärung relevanten Code enthalten. Längere Codeabschnitte können im Anhang stehen. Der komplette Code, der für die Arbeit geschrieben wurde, sollte in einem Repository abgelegt werden.

Listing ?? zeigt ein Beispiel für ein Codelisting in der Programmiersprache C. Algorithmus ?? zeigt einen Routing Algorithmus als Pseudocode. Der Code wurde mit dem Paket algorithm2e algorithm2e erstellt.

3.3. Litteratur

Ein Literaturverzeichnis kann z. B. mit dem Bibtex Paket bibtex erstellt werden. Dazu wird für jede Quelle ein Eintrag in der Datei `references.bib` angelegt. An der passenden Stelle im Text können diese Einträge mit dem `\cite{}` Befehl zitiert werden. Für jede Quelle die zitiert wird, legt LaTeX im Literaturverzeichnis einen Eintrag an.

Beschreibungen der Quellen im Bibtex-Format müssen meistens nicht selbst erstellt werden, sondern können direkt bei vielen Verlagen und Bibliotheken direkt generiert werden. Wis-

senschaftliche Softwareprojekte geben ebenfalls oft auf ihrer Website eine Beschreibung im Bibtex-Format an.

3.4. Abbreviations

Für jede verwendete Abkürzung kann ein Eintrag in der Datei `acronyms.tex` angelegt werden. Wenn diese Abkürzung im Text zum ersten Mal auftaucht, sollte der Begriff ausgeschrieben werden mit der Abkürzung in Klammern dahinter. Bei weiteren Vorkommen im Text kann dann die eigentliche Abkürzung verwendet werden. In gibt es dafür spezielle Befehle. Beispiel für ausgeschriebene Abkürzung

A. Glossar

A.1. Abkürzungen

A.2. Symbole

B. Speichermedium

Hier gehört eine Tabelle des Inhalts deines beigefügten Speichermediums (SD-Karte/USB-Stick) hin. Ggf. müssen Kommentare/Erklärungen dazu geschrieben werden.