Decision Support
Institut für Wirtschaftsinformatik

Bachelor 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

Ich, Aymen Ben Aicha (5031286), bestätige hiermit, dass ich mir der durch die Corona-Pandemie eingeschränkten Möglichkeiten bei der Bearbeitung der Abschlussarbeit (Literaturversorgung etc.) bewusst bin und trotz der gegebenen Umstände die Bachelorarbeit Extraktion von Informationsquellen aus Mobilitätsdaten regulär anfertigen kann und anmelden möchte.

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Bachelorarbeit selbstständig verfasst sowie alle benutzten Quellen und Hilfsmittel vollständig angegeben habe und dass die Arbeit nicht bereits als Prüfungsarbeit vorgelegen hat.

Braunschweig, August 14, 2022

Acknowledgment

I would like to acknowledge and give my warmest thanks to my supervisor, Mr. Felix Spühler, who made this work possible. He was always there for me, whenever I needed help to finish my thesis. His guidance and advice carried me through all the stages of finishing this project.

I would also like to give special thanks to my parents, relatives and friends as a whole for their continuous support and understanding when undertaking research and writing the thesis.

This thesis is dedicated to you, the ones who believed in me and kept trusting in me even at times of difficulties.

Contents

List of Figures

Abbreviations

- VRP: Vehicle Routing Problem
- CSV: Comma-separated values
- CRISP-DM: CRoss Industry Standard Process for Data Mining
- ETL: Extract, Transform and Load.
- NYC TLC : New York City Taxi and Limousine Commission
- API: Application Programming Interface
- PUDay: Pick up day
- PUhour: Pick up hour
- DODay: Drop-off day
- DOHour: Drop-off hour
- UI: User Interface

1. Introduction

In a competitive market, each platform of consumer delivery service tries to get an edge over its competitors by offering a transparent view to their customers, such as a live tracking option, a real-time update, and a precise delivery prediction. Improving customer satisfaction is the most challenging step to gaining consumers' trust since it is always related to predictions based on gained data from the past. It is not only about the client's perspective but also the business itself; it should dig deep into available data to opt for the most profitable strategies regarding preferred work areas, temporal preferences, or other available preferences.

The data that is generated through this kind of business activities are defined as mobility data, to manage them, as an unprecedented amount of data generated by different transportation means, requires setting up robust analysis and mining infrastructure in order to help companies operating in different related fields in planning and decision-making. Those requirements are to be satisfied whenever the user, either government in city planning or private firms in goods or person transport. For those reasons, the collected data and information created during those services will be saved in a ready-to-use condition. This process requires more effort to have operational templates, making the analysis and the extraction process automated and easy to use.

Where to start picking up clients, when or what are the most profitable trips, and areas for New York City Taxi and Limousine Commission are parameters to be considered before even sharing real-time information with the client. Therefore, having suitable datasets to conduct those analyses to overcome questions related to profitability requires a coherent dataset that gives highlights about a chosen area or period. Companies with businesses related to delivery or pick-ups, such as taxi firms, restaurant meal delivery, or packages delivery, are still trying to work on their algorithms to have better route planning and more precise arrival time estimation, increasing customer experience. At the same time, lower accuracy rates may lead to dissatisfaction. Since those firms are becoming world leaders, such as Uber, Just eat, or even Amazon services, the urge to have some automated template to pre-process available data, discover its insides, and extract smaller dataset that satisfies the algorithm's needs, becomes more prominent. Thus we put our energy into speeding up this process through data mining.

While using Data Mining models, precisely by satisfying the CRISP-DM model as an example, the main goal will be to provide instructions for generating suitable datasets that will be used in solving problems related to VRP-Models.
The suggested process will be to work on the given dataset, retaining consistent data to generate smaller and more meaningful datasets that could be directly used depending on the business wishes. The aim will be to translate some potential business needs into data mining tasks, therefore smaller datasets, by suggesting appropriate data transformations and data mining techniques. After validating the dataset, the challenge will be to work on different parameters and define some possible use cases. We propose a template as a final solution, which could select only relevant parameters with their wished values in different scenarios.

2. Foundations

To enable a meaningful discussion during this thesis, we aim to mention the crucial techniques, terms and notions.

2.1. Data-Mining

Data mining [1] is a very well-known term in data science, although it can also be referred to as knowledge extraction, information discovery, or even data pattern processing. It is collecting different kinds of data to find valuable patterns and results. Specific algorithms should be applied during the data mining process that could help us extract these patterns. It could be summed up into two primary purposes:

1. Describe the target data set.
2. Predict outcomes through the use of machine learning algorithms.

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

2.2. Mobility Data

Mobility data [2] are data generated by activity, events, or transactions using digitally-enabled mobility services or devices. It mainly reveals information related to the spatial location of the performed activity, such as longitude and latitude, which could be collected through smartphones or mobility vehicles. It also contains temporal details, relating the spatial to temporal element. 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. Vehicle Routing Problem

The Vehicle Routing Problem [3] is a combinatorial optimization and integer programming problem which belong to the operational research problem, and it generalizes the Traveling Salesman Problem(TSP). Each organization needs to determine which orders (food or package delivery, pickups of passengers) should be serviced by each route and in what order the places should be visited. It is considered a primary goal to distribute the orders and minimize the optimal operating cost for the fleet of vehicles. The VRP is not also valid for route calculating but can solve more specific problems because numerous parameters could interfere, such as matching vehicle capacities with order quantiles, assigning the right driver at the right time, giving equal breaks to drivers, and pairing orders, so the matching ones in terms of directions go on the exact vehicle.
VRP dates back to 1959, when Dantzig and Ramser set the mathematical programming formulation and algorithmic approach to solve gasoline delivery problems to service stations. In 1964,

Clarke and Wright improved on Dantzig and Ramser's approach using an effective greedy algorithm called the savings algorithm.
VRP is an NP-hard problem that can be exactly solved only for small instances of the problem.
A few of the most common objectives of VRP are:

- Minimize the number of used vehicles to serve orders.
- Reduce transportation costs based on chosen travel paths, as well as the fixed costs associated with using cars and hiring drivers.
- Maximize profits.

2.4. Extract, Load and Transform

ETL [4], which stands for extract, transform and load, is a data integration process and helps in:

- First, extracting the needed data from the sources.
- Cleansing the data to improve its quality and consistency.
- Applying the wished transformations on the source data to generate the expected output.
- Loading final data into target database.

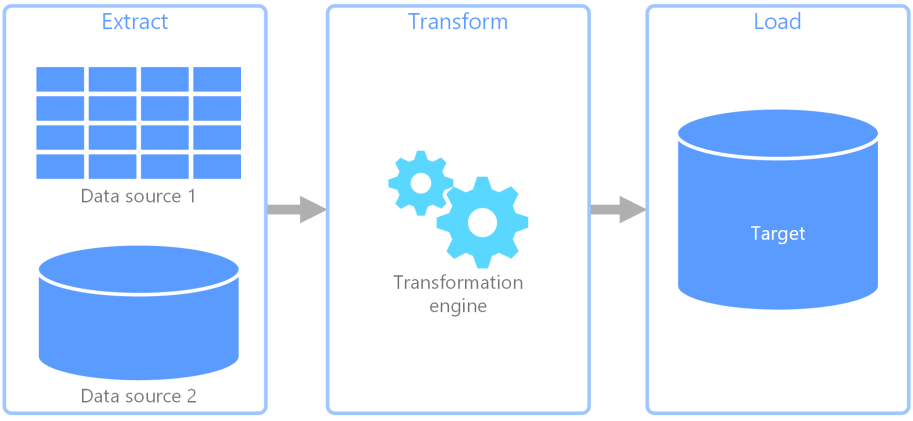

Figure 2.1.: ETL Process[5]

3. Methodology

This chapter introduces the methodology behind the work plan. First, we define the used model and then carry on with the theoretical basis behind the solution.

3.1. CRoss Industry Standard Process for Data Mining

CRoss Industry Standard Process [6] for Data Mining, known as CRISP-DM, is an open standard process model describing common approaches data mining experts use.

1. As a methodology, it includes a 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 deployed solutions push into new business questions, which will lead to revisiting previous steps in the same project or new knowledge for upcoming projects.

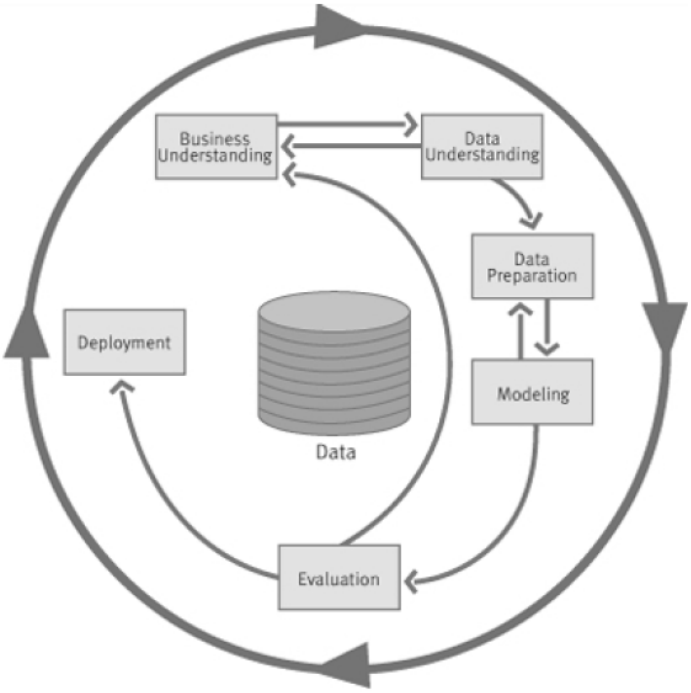

Figure 3.1.: CRISP-DM: The Process Model [7]

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

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

3.1.1. Business Understanding

This initial phase describes what the customer wants to accomplish from a business perspective by receiving his requirements and converting all possible and accessible client preferences into data mining goals. Then, after having a less vague vision, we will create a preliminary plan to achieve the settled goals.

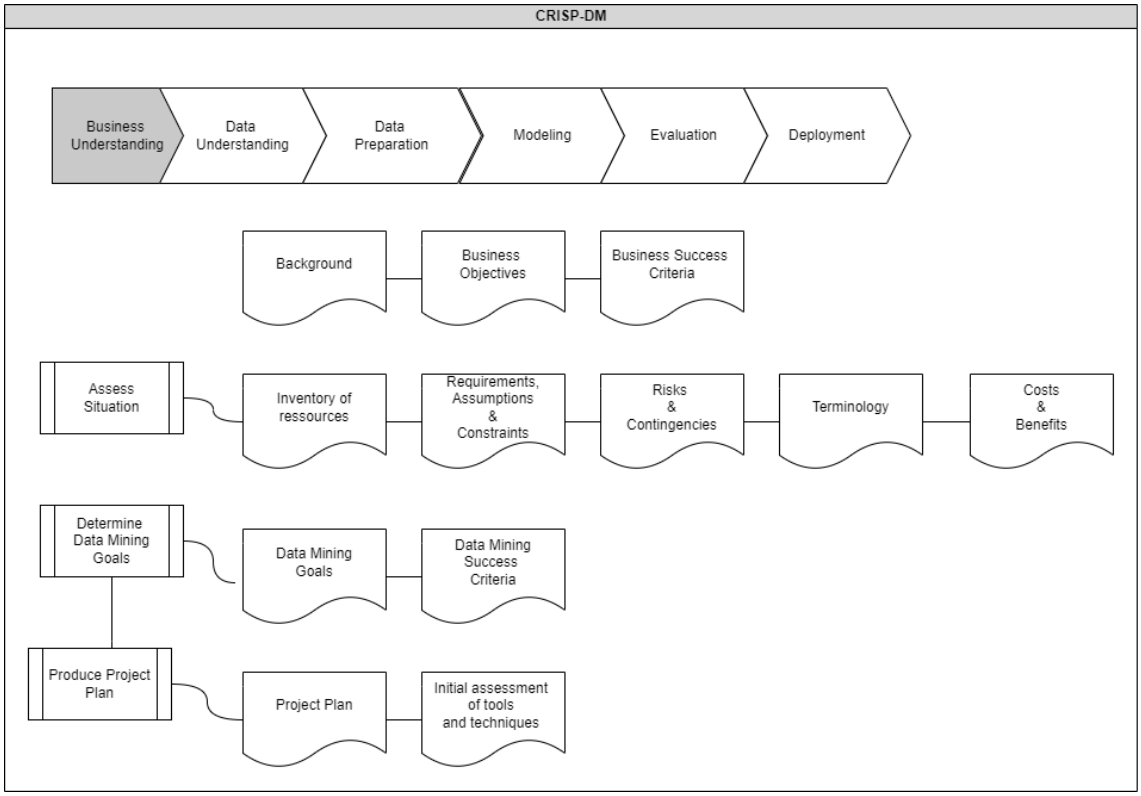

Figure 3.2.: CRISP-DM: Business Understanding [8]

3.1.2. Data Understanding

The Data understanding phase begins with collecting data in its raw form, getting familiar with it, and rating data in terms of quality. This step will help gather attractive information as first insights into the data and detect interesting subsets from the original data.
There is a close link between Business Understanding and Data Understanding. Therefore, to have a precise formulation of the data mining problem and a clear project plan, those two steps should be conducted extensively since they play a role in the upcoming process.

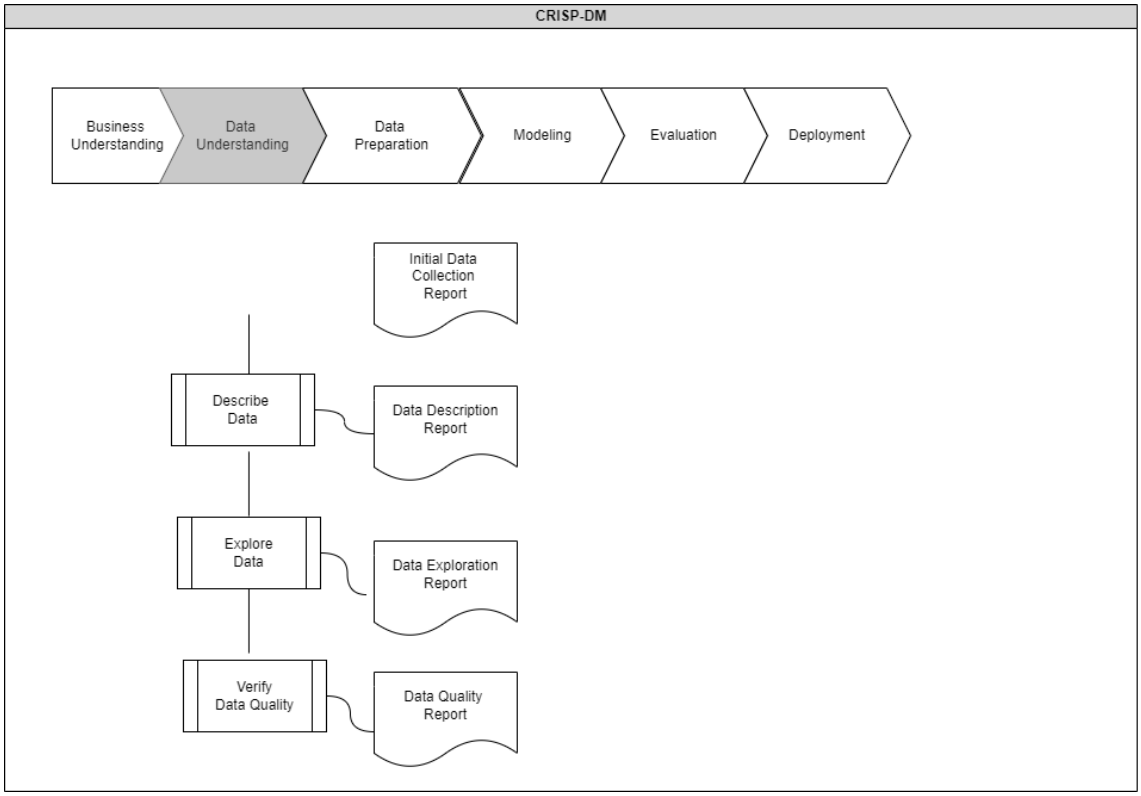

Figure 3.3.: CRISP-DM: Data Understanding [8]

There is a close link between Business Understanding and Data Understanding.To have a clear formulation of the data mining problem and to have a clear project plan, those two steps should be conducted extensively since they play a role in upcoming process.

3.1.3. Data Preparation

This step aims to construct the final data set from raw data, which will be the input for the next steps. It could be implemented multiple times in the following steps and not in any pre-defined order, whenever the user has more knowledge about his data.
At the first loop of the data construction, the primary results will be considered as a draft, containing our perception of the next steps.
After knowing more what we need, the construction could be re-launched to match the advance.

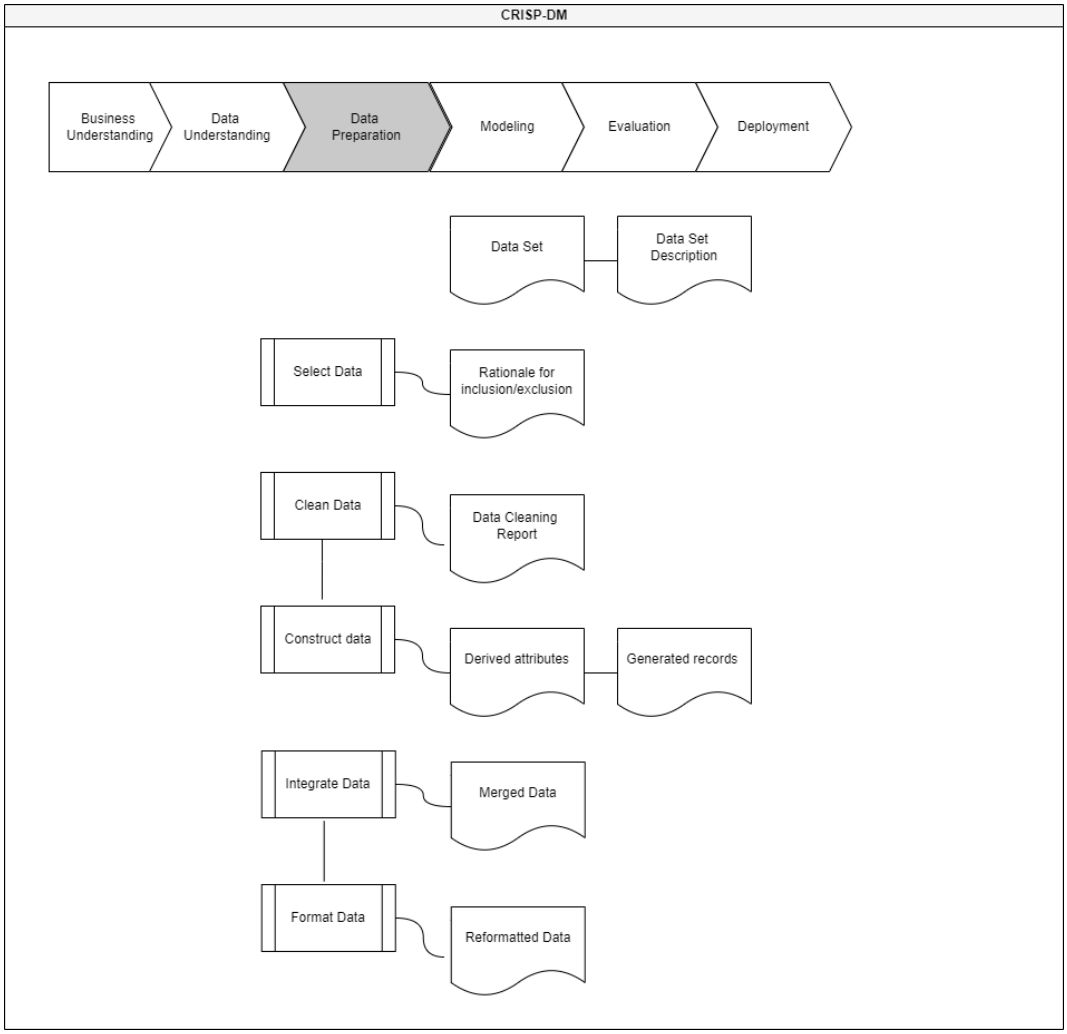

Figure 3.4.: CRISP-DM: Data preparation [8]

3.1.4. Modeling

In the Modeling stage, the main goal is to select the proper modeling technique to solve data mining problems. Parameters are to be chosen wisely to have optimal values. Furthermore, a quick step back into the data pre-processing stage is highly recommended.
Modeling in our work consists in creating data set extraction technique. For it we will define in the next Chapter 3.2, the model we need to present the solution, build it and explain it in the application part.
There is a close link between Data Preparation and Modeling. Typically, data problems could only be highlighted, or some ideas appear related to data construction when starting the modeling stage.

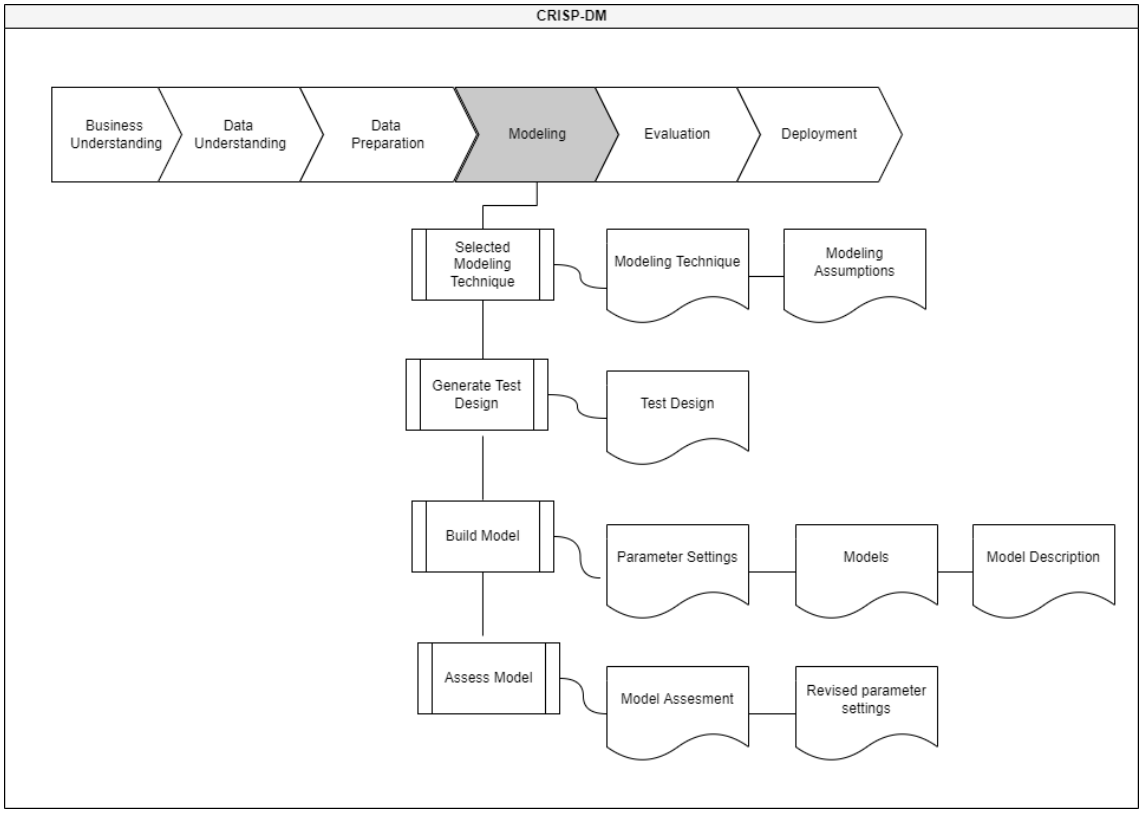

Figure 3.5.: CRISP-DM: Modeling [8]

3.1.5. Further Steps: Evaluation and Deployment

Evaluation and Deployment are two further steps belonging to the CRISP-DM model. The following diagram represents the steps as well as the output of each of the steps.

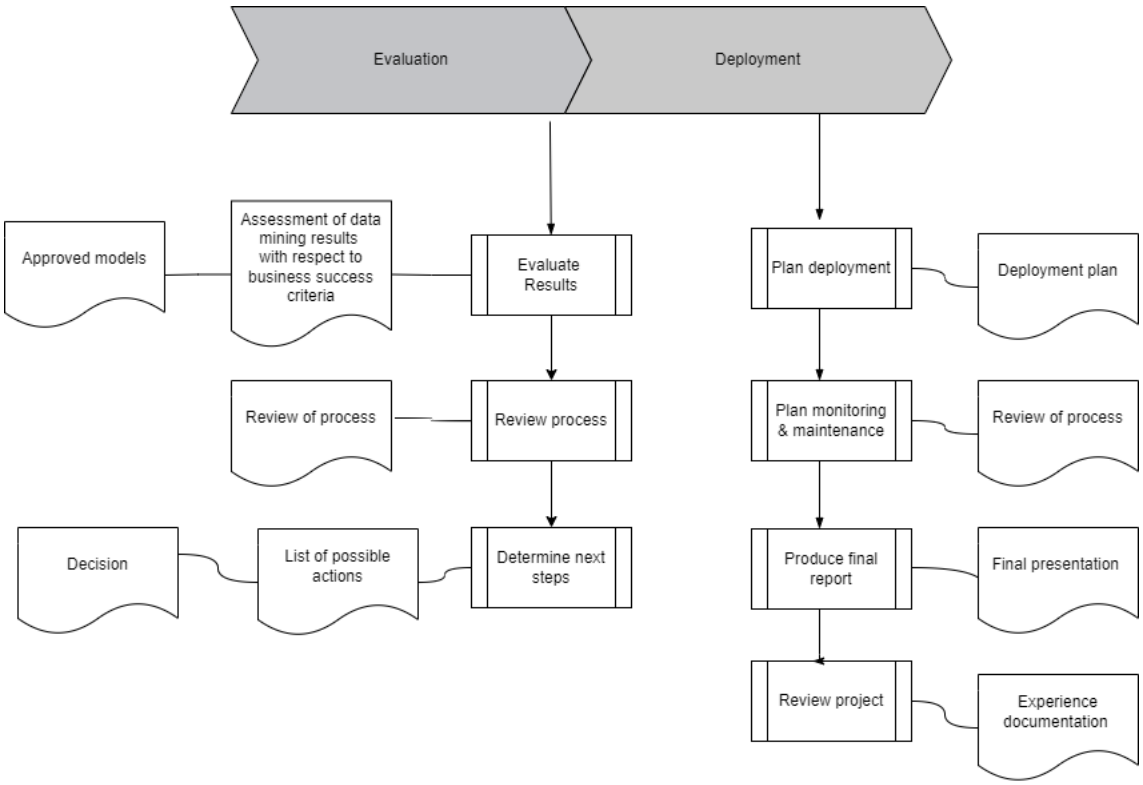

Figure 3.6.: CRISP-DM: Evaluation and Deployment [8]

At the Evaluation stage, the model is already built. Nevertheless, before deploying the model, this step exists to rate if the model meets the expected business objectives and if there is a critical objective that is not yet tackled. As a result of this step, a decision should be made on whether the whole process could reach the settled data mining goals.

Finally, the process ends with the Deployment step, where the goal is not only to have more knowledge from the data but also to make the results useable for users. In the simplest case, this step will require generating reports, or at the other extreme, it will require having a whole implementation strategy. Generally, the customer carries on this step after having valid instructions from the analyst.

3.2. Dataset Reduction Techniques

This section introduces the final solution's main parameters, goals, and output, a general definition of the function, and the logic behind the created template.

- Goal: After preprocessing data, this phase reduces the cleaned and smaller datasets, matching the user's preferences after considering different conducted analyses. It allows the developer to carry on the next steps while having better conditions to enhance the output's accuracy or optimality. The whole process will look like this diagram:

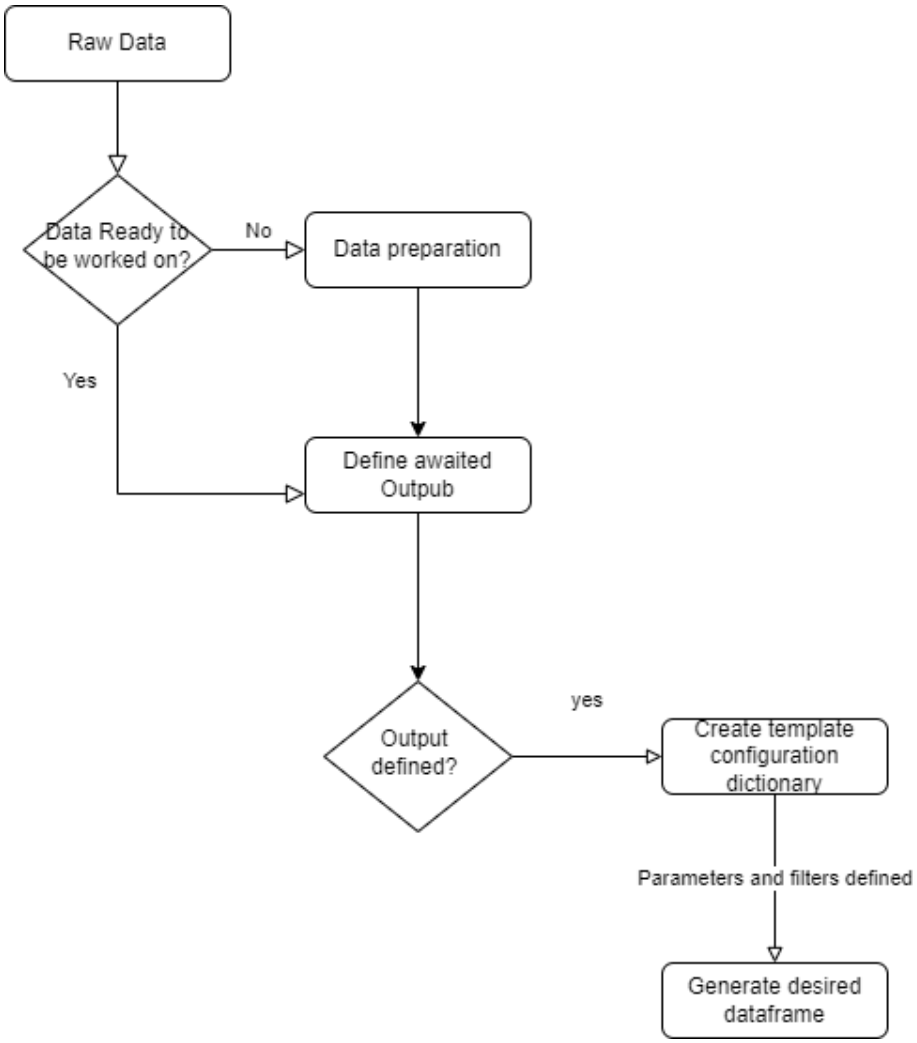

Figure 3.7.: Flowchart: Solution work plan

- Example of possible applications: For VRP-Problems, for example, it makes more sense to consider a precise temporal basis or spatial preference since every selected item has its results, and the conclusion is whether to opt for location X at time Y or location Z at time Y. That is why we are introducing this user-friendly template, considered a querying method. Therefore, the theoretical foundation of the work should be introduced, manifesting how the template should be filled and could be considered while querying. This dictionary should contain different column names as well as the parameters to be worked on.

- Querying: The main process consists in querying in the dataset. Thus, the user should select the wished output in terms of :
  - Columns: What columns does the user want to have in his resulting data frame?
  - Length: Is it possible to select a data frame's length by limiting one of its column entries?
  - General filter: The following preferences are pre-defined: It is possible to show the n first appearing trips or the n last ones. Furthermore, we could select a random number of trips to be given as an output, without having a logical order, but only on a random basis.
  - Filtering specific rows: Each row could be either just selected with all its possible values or filtered. With this, we have the following possible criteria: Selecting the maximum or minimum recorded values for a specific row and a specific value by choosing from equal values or intervals.
  - Original index: It is also possible to save the index of the selected columns in the original data frame; it is a true/false parameter to be chosen.
  - Rules: The queries are to be executed in the order the user tipped them. Furthermore, only the second input line will be considered if one column is filtered twice.
    When entering an interval, it should be in the form of a list while executing an equal or interval command.
    If the user wants to have a column without specifications, the column's name should be given, followed by null as a value.
  - Saving Results: The results are to be saved as .csv files

4. Experiments

The previous chapter, defined the theortical frame of our work. This chapter starts with the experimenting part, first with an execution of the CRISP-DM, then with the suggested solution.

4.1. Execution of CRISP-DM

4.1.1. Business Understanding

Determine Business objectives: Understand, from a business perspective, which the client is, his expectations from work, and generally his settled goals. All important factors that can influence the project's outcome should be uncovered in this step.

a) Output: Background
The New York City Taxi and Limousine Commission (TLC) [9] 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 and avoid congested paths, they 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 the client's different questions.
For example, wishes and inquiries should be categorized; they could be related to producing a better customer experience.
Furthermore, prediction is considered a chapter, and questions concerning preferred pick-up times, drop-offs, or the expected length of a trip should be answered, and results should be optimized.
Nonetheless, inquiries related to rentability, business growth, and getting over competitors are to be considered as business objectives, and the solution should work on resolving them.

c) Output: Business success criteria
The main goal is to extract appropriate data sets suitable for the upcoming work.
Here a cleaned data set is expected as an output, containing the wished original and created attributes, which are helpful in implementing VRP-Algorithms.
Moreover, it is expected to define clusters in a practical way, where they are ready for implementation.

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

a) Output: Inventory of resources
   * Available data: We deal with Yellow Taxi Trip Records (available from 01.2009)
   * 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
   * Programming languages: Use of python and Jupyter Notebooks

b) Output: Requirements, Assumptions, and constraints
   * Competition schedule: Gantt-Diagram
   * Working process should be on CRISP-DM oriented
   * Data mining steps should be respected (data cleaning, data merging, and further steps)
   * 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 arise. To avoid this problem, we will try to select ideal data (such as data for one year and one type of taxi) to avoid all possible outliers and extremes as well as incompatible attributes definition.
   * Since we do not have a direct communication channel between the data provider and the business itself, we will assume that the asked questions are generally related to VRP problems and that their answers will remain under standardized fulfillment of data mining goals as well as CRISP-DM models.

Determine Data-Mining Goals: Project expected outputs of the project while defining general success criteria.

a) Output: Data-Mining Goals
   We extract the best from the data set by eliminating outliers and missing values. For a better data understanding, we await to define different data relations; relations between passenger count and the fare amount, the impact of a specific location on rides, or even the identification of 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 data sets (clustering)

4.1.2. Data Understanding

Collect initial data: 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 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 unique app token.
    For further references, check link in [10]
  - Available data (see 2.2.1.2 Inventory of resources)
* Selection of data: Selecting Yellow Taxi Trip Records in the year 2021. This limited selection helps in discovering more insides and interesting shapes of the data we wish to work on

Describe data: Examination of the surface proprieties of acquired data. The data has originally

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
Fare_amount	The time-and-distance fare calculated by the meter.

Figure 4.1.: Data Dictionary: Declaration of Main Attributes

18 columns, having non-null values and being defined as objects. Therefore, we will categorize it into the following classifications:

1. IDs
2. date times
3. different amounts related to the trip
4. categorical variables

After allocating the suitable data types for each attribute, pursuing the analysis process will be easier since we are now sure what the numerical and categorical values are. Nonetheless, this step will be redone after in data preparing step, but doing it at the first stage will help in the discovery, analysis, and preprocessing path.

	passenger_count	trip_distance	fare_amount
count	1000000.000000	1000000.000000	1000000.000000
mean	1.415547	2.650791	11.123815
std	1.063357	3.522187	12.732147
min	0.000000	0.000000	-250.500000
25%	1.000000	0.990000	6.000000
50%	1.000000	1.620000	8.000000
75%	1.000000	2.800000	12.000000
max	8.000000	427.700000	6960.500000

Figure 4.2.: Dataset Attributes and Main Related Facts

This table illustrates the count of available rows, mean, and max and demonstrates the standard deviation and the data's 1/4, 1/2, and 3/4 distribution.
At first glance, we could easily see that the fare amount and its corresponding amount contain some false inputs since it manifests some negative payment values, which could not be accurate. Trip distance or passenger count of 0 is some suspicious values to be highlighted and considered. To dig deeper into those analyses, we will conduct further specific analyses for each of the needed columns. Explore Data: Establishment of graphics and reports as well as queries to directly tackle data mining goals
In this step, we will work on different attributes and their relations between each other in order to discover some data inside.

1. Correlation analysis:
   This Diagram, in the next page, shows correlations between different attributes and will be used later to compare the first raw data to processed results.

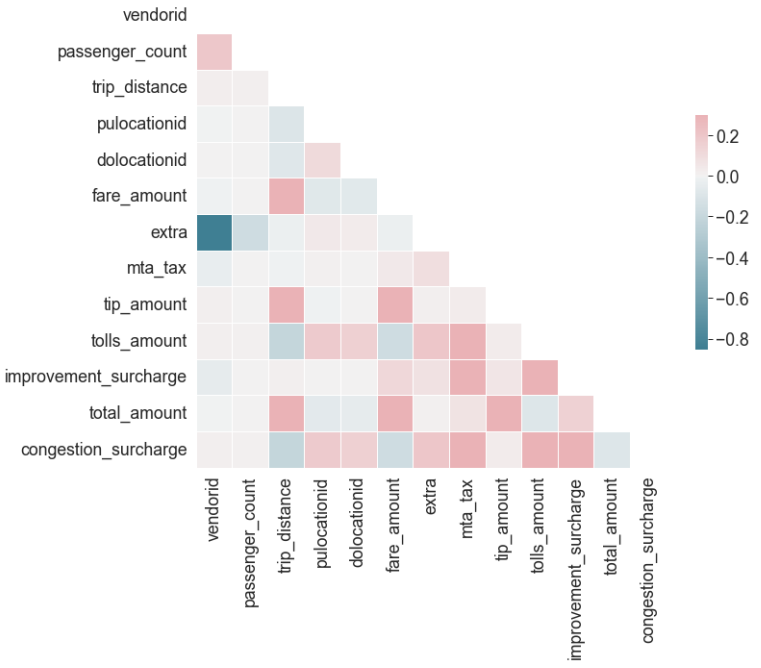

Figure 4.3.: Attributes Correlations

2. Temporal analysis:
   First, we will consider classifying timestamps into days to get a general view of 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

The table above shows a nearly equal distribution in matter of days. Since all values range between 112022 and 168374, it makes more sense to consider other time slots. Thus, we will define four-time intervals.

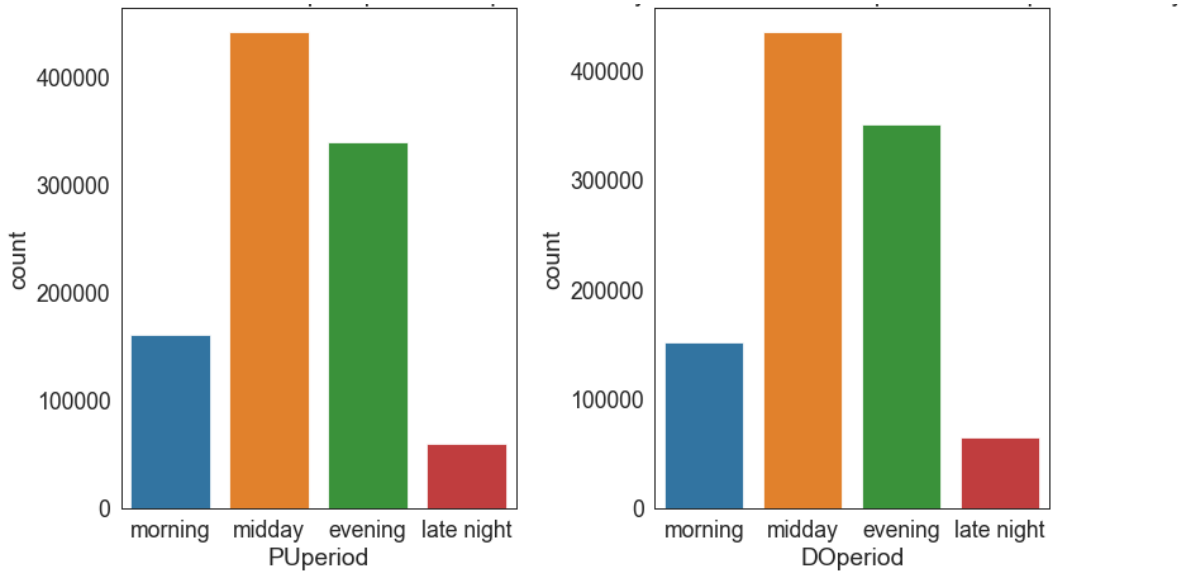

Figure 4.4.: Time Distribution Based on the Part of the Day

The attractive output is that 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.

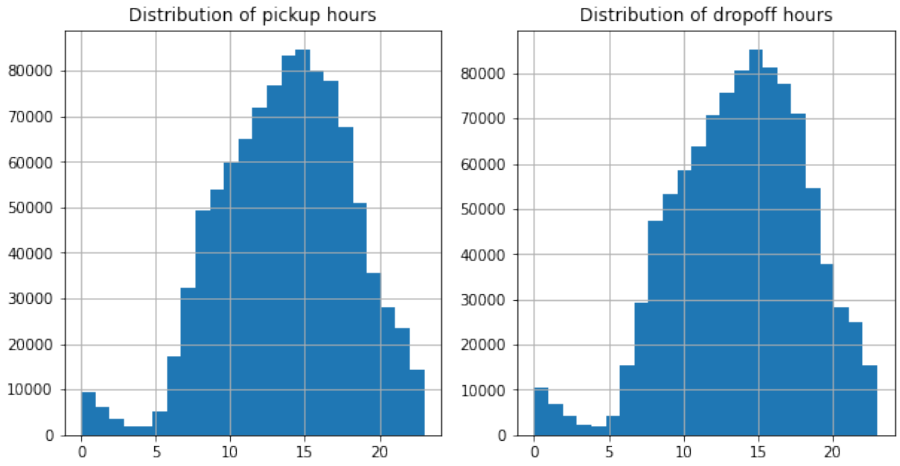

Figure 4.5.: Hourly Distribution

The hourly distribution confirms the statement related to Figure 4.4 that starting from 10 PM, the need for night rides continuously decreases until 5 AM. It will therefore increase to achieve its peak at 4 PM.

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

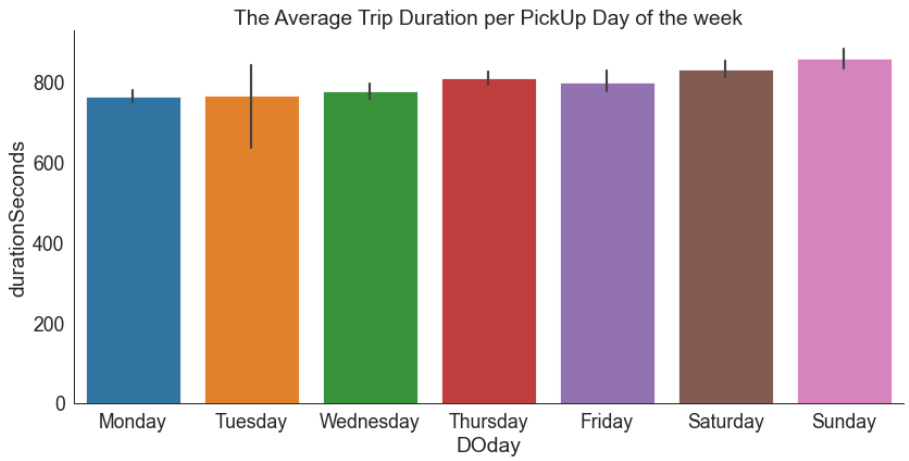

Figure 4.6.: The Average Trip Duration per Pickup Day of the Week

The graph denotes the average duration of a trip for each day of the week, in seconds. The error bars provide some indication of the uncertainty around that estimate. Figure 4.5 and Figure 4.7 shows longer trips on weekends, but the difference is not huge, so an hourly distribution makes more sense (Figure 4.8)

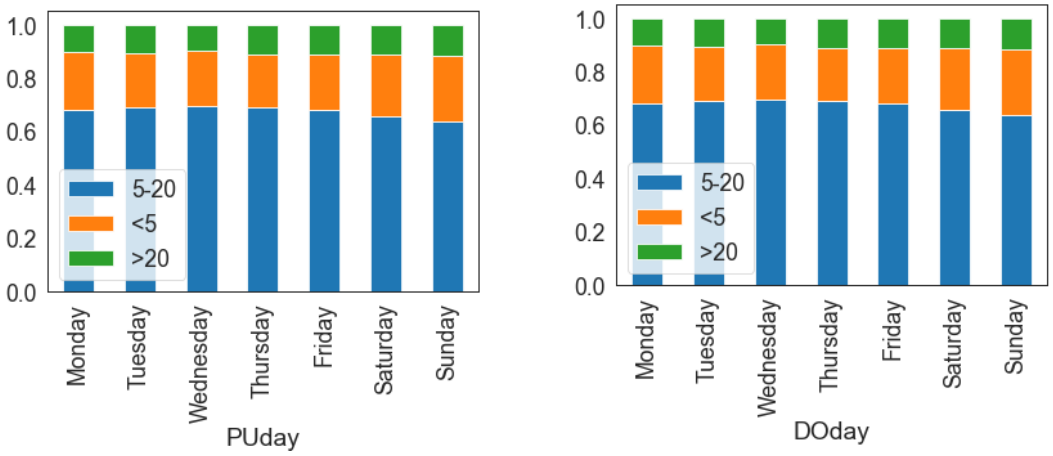

Figure 4.7.: Relationship Between Trip Duration Intervals and Pickup/Drop-off Day

From the previous Diagram, trips lasting 5 to 20 minutes are the more attractive, regardless of the day, and the longer lasting trips are on weekends, with a relatively small difference. Not only that, but on Sundays, the demand for short-lasting trips is also higher.

Figures from 4.4 to 4.8 intensify this statement:
Weekends are more attractive days for trips, especially the longest ones. Furthermore, late-night rides starting from 22 o'clock are the most favorable for a taxi ride.
To defend this statement, the following concluding chart represents pickup/drop-off hours and the trip's duration:

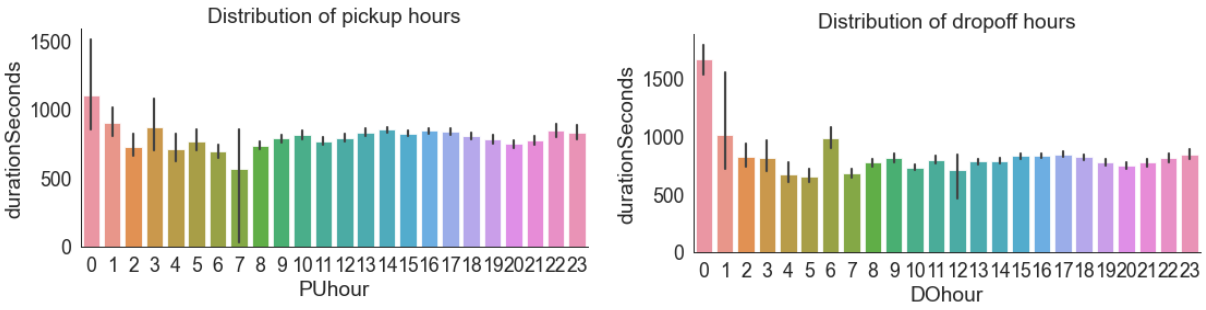

Figure 4.8.: Duration's Hourly Distribution

4. Geospatial analysis:
   The Geospatial analysis is a visual manifestation of gathered data that contains geographic coordinates.
   To better understand the available geospatial data, we merged our initial dataset with other different datasets, such as taxi lookup.csv dataset, to obtain the related Borough and zone to each coordinate.

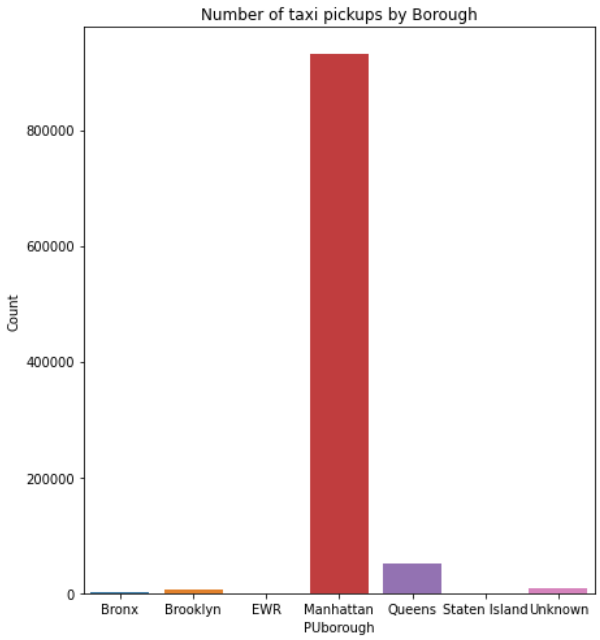

Figure 4.9.: NYC Pick-up Boroughs

Figures 4.9 and 4.10 highlight the distribution of pickups and drop-offs depending on the five boroughs of New York City and the Newark Liberty International Airport(EWR). Some given data are here classified under unknown. The reason could be unspecific geographical data. In a nutshell, The attractive pickups, as well as drop-offs, are mainly in Manhattan, highly far from other places, which were already marked in red on our map.

Furthermore, the geopandas library is to be imported in order to plot geospatial visualization to compare pickup and drop-off amounts from each zone. If we want to have a general view of the pickup and drop-off heatmaps, we will have the following map view:

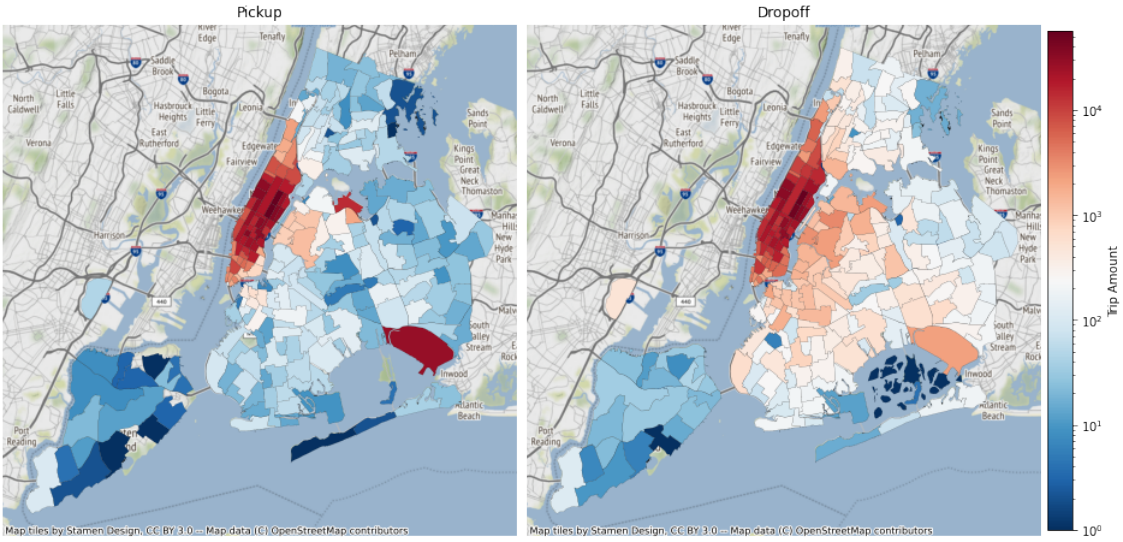

Figure 4.10.: NYC Heat Map

In the following steps, template users need to have simply understandable graphics, manifesting the hotspots that should be considered. In this case, Manhattan and JFK - John F. Kennedy International Airport are considered hotspots for pick-ups and drop-offs. Through this geospatial analysis, we could rapidly identify where to start our business and what datasets should be filtered and taken into consideration.

5. Relation between Passenger Count and Trip Distance
   In order to make the data cleaning step faster, further analyses are to be conducted, which puts some outliers into the spotlight.
   For this part, we will choose some key attributes and carry on with distribution and count analysis in order to identify possible data to be deleted in advanced steps.

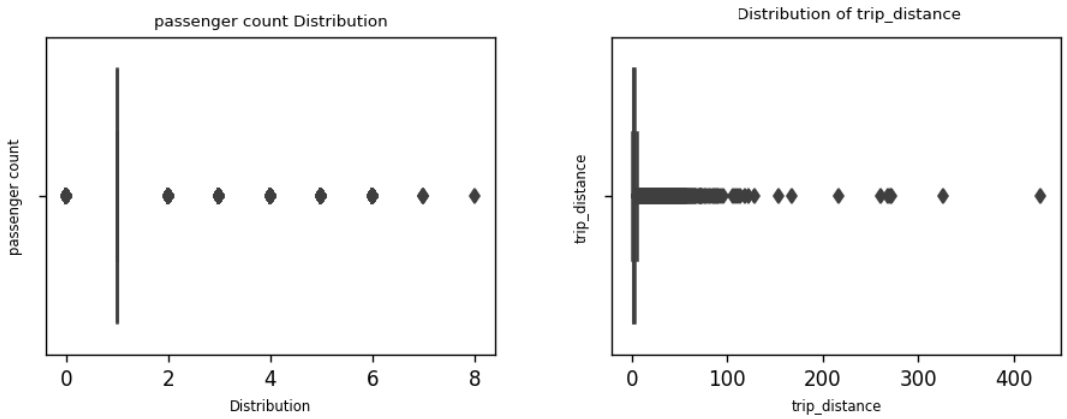

Figure 4.11.: Distribution of Passenger Count and Trip Distance

6. relation trip distance and trip duration

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

Figure 4.12.: Different Manifestation of the Relation Between Passenger Count and the Couple Trip Distance & Trip Duration

Here, the dark green is a manifestation of the highest means, length, or standard deviation when fixing the passenger count as a parameter.
For seven 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 mostly appear are passengers going alone for different trips, causing the highest standard deviation at the trip's duration level.

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

* Z-Score [10] (also called a standard score) shows how far a data point is from the mean. However, it more technically measures 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) Concerning fare amount:
     0.02 % of the data is placed on above three standard deviations from the mean

  b) In relation with the trip distance:
     approximately 0.02 % is also in the same area.

  c) Regarding passenger count:
     approximately 0.05 % belongs to the same area

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

4.1.3. Data Preparation

The whole Data Analyst work will be mainly done in the upcoming steps.

Data Cleaning

Improve the quality of data due to the previously conducted analysis and work. For example, the data cleaning report could be established by detecting which issues are outstanding and what effect these can have on the project's outcome.

1. Data cleansing:
   Records with implausible values or errors are removed based on the conducted analysis related to data distribution and common sense to ensure data's correctness and avoid all possible derivations.

   - Passenger count is one passenger or more, but it should not exceed six riders since it is legally not allowed.
   - The 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.

Construct Data

- Duration of trips is due to 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 a three-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

At the end of the two steps, Data Cleaning and Data Construction, the dataset will exactly have this shape after each step.

Rows	Columns
1000000	18
1000000	8
978886	8
963503	8
957517	8
894571	8
889660	10
889660	16
889660	18

The table demonstrates the development of the dataset, as it went from being 1M*18 columns to being reduced in terms of rows but also having newly appearing columns after deleting the irrelevant ones.

4.1.4. Integrated and Formatted Data

The final shape is 889600 columns * 18 rows. The result of this intensive work is saved as a .csv file, having a size of 127 KB.

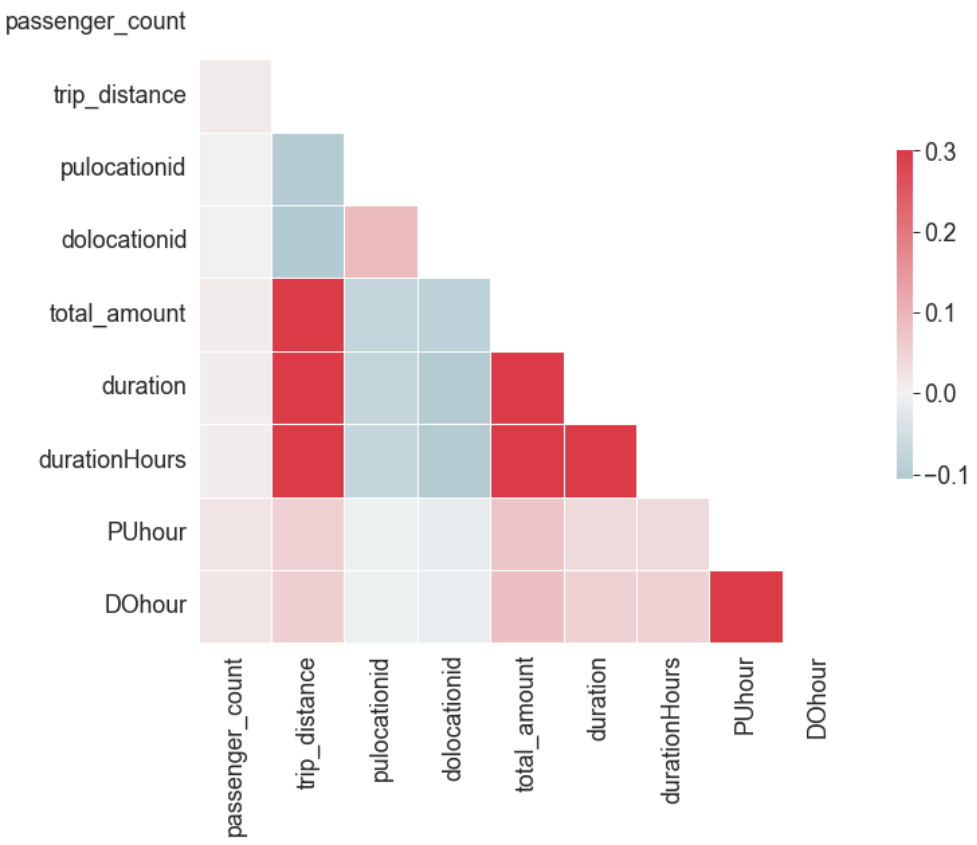

Figure 4.13.: Data Correlation After Treatment

As a matter of fact, the columns extracted from each other will always be highly correlated. Nevertheless, here we observe a higher correlation between the total amount and duration on the one hand and the trip distance on the other. PUlocation and DOlocation are also related to each other, as well as PUhour and DOhour. The total trip amount is related to pick-up and drop-off time and location and affects the trip's total amount.

To start in the next steps, a cleaned and well-preprocessed dataset was saved as .csv file, containing only the necessary columns and a dataset that went through all the preprocessing stages.

4.2. Results

First, we will define the template, in a flow chart, by presenting the different possible .

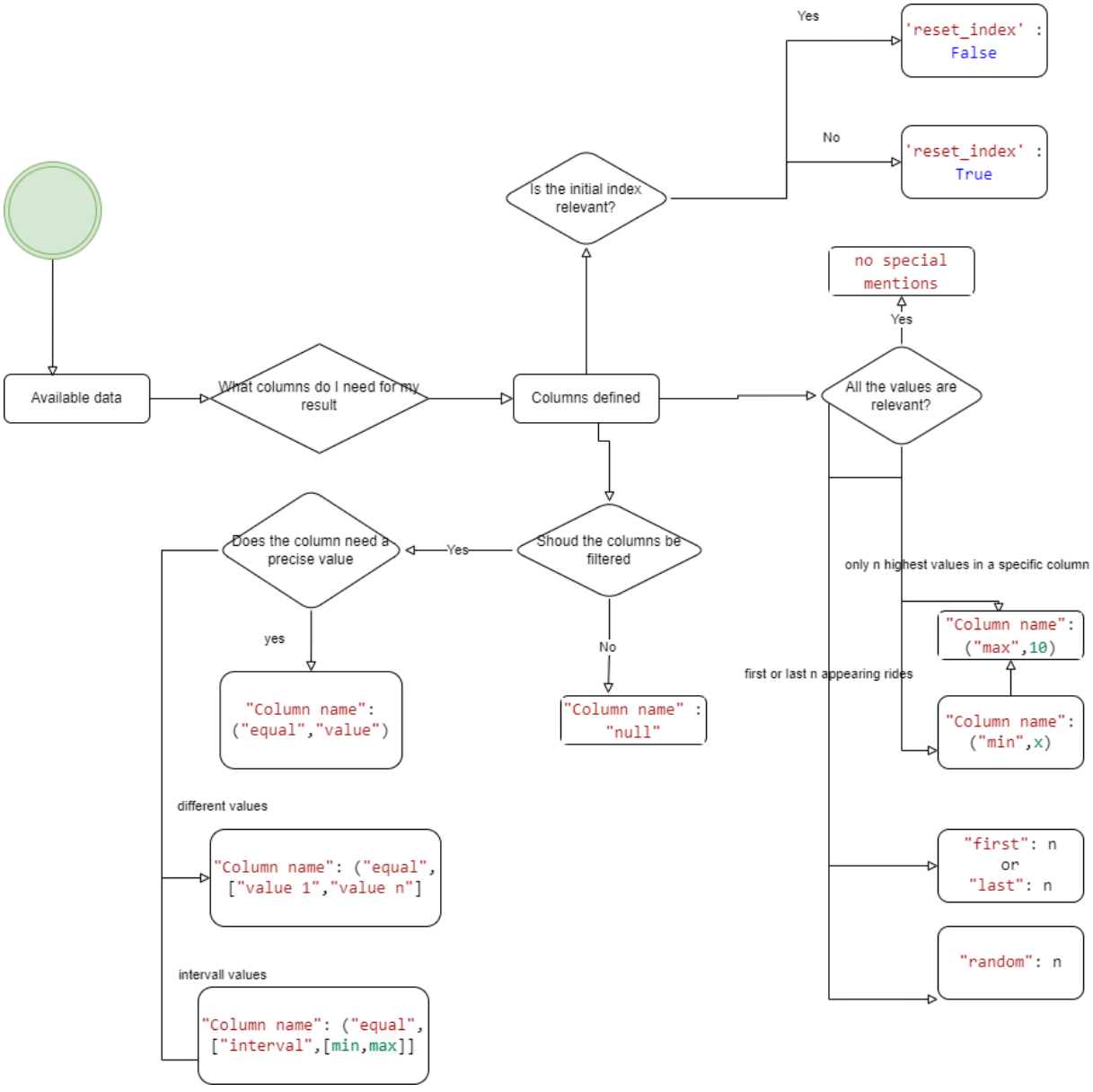

Figure 4.14.: Flow-Chart: Reduction Logic

In this stage, some examples of the results of this function are to be shown and explained for different possible scenarios.

- Use Case 1:

```
config = {
    'columns' : {"PUborough" : ("equal","Queens"),
                 "trip_distance":"null",
                 "random" : 12},
    'reset_index' : False
}
```

Figure 4.15.: Query of the First Use Case

The figure could be translated into the following:
Select Random 12 rides that started from Queens and show all related trip distances while maintaining the original entry index on the primary dataset.
The result of the previous query will be the upcoming data frame, revealing 12 rides in Queens and random trip distances. For this data frame, we could not conclude since it is only an example of the query.

	index	PUborough	trip_distance
0	848277	Queens	6.53
1	464937	Queens	17.95
2	802449	Queens	23.00
3	607688	Queens	7.80
4	735118	Queens	23.36
5	763	Queens	19.48
6	877279	Queens	3.91
7	828292	Queens	2.30
8	887024	Queens	9.77
9	741245	Queens	10.54
10	884628	Queens	17.70
11	687420	Queens	8.26

Figure 4.16.: Data F rame as a Result of the First Query

- Use Case 2: The user needs to have rides that start on Saturday or Friday since he concluded that weekends are more attractive for rides. The selected pick-up borough is not essential in this case since we will fetch the importance of trip distance by selecting a low distance interval from 1 to 3 km while having the biggest ten appearing passenger count.

```
config2 = {
    'columns' : {"PUDay" : ("equal",["Saturday","Friday"]),
                 "PUborough" : "null",
                 "trip_distance" : ("interval",[1,3]),
                 "passenger_count": ("max",10)
                },
     'reset_index' : True
}
```

Figure 4.17.: Query of the Second Use Case

The data frame matches the expected length and criteria and gives us an insider about that kind of preferred trip. Moreover, they are mainly performed in Manhattan, which could be taken as a first conclusion related to trips with six passengers.

	PUDay	PUborough	trip_distance	passenger_count
0	Saturday	Manhattan	1.51	6
1	Saturday	Manhattan	1.84	6
2	Saturday	Manhattan	1.16	6
3	Friday	Manhattan	2.08	6
4	Friday	Manhattan	1.49	6
5	Saturday	Manhattan	1.50	6
6	Saturday	Manhattan	1.33	6
7	Saturday	Manhattan	2.42	6
8	Friday	Manhattan	2.10	6
9	Saturday	Manhattan	2.34	6

Figure 4.18.: Data frame as a Result of the Second Query

- Use Case 3: After discovering the different conducted analyses in previous phases, we want to summarize the highlights of the different steps in one query. We will only consider pickups performed Sunday at midday in Manhattan. The passenger count should be one, and the trip distance should be more than 0 and less than 100 km. The resulting query will be:

```
config3 = {
    'columns' : {"PUborough" : ("equal","Manhattan"),
                 "PUDay": ("equal","Sunday"),

                 "PUperiod":("equal","midday"),
                 "trip_distance" : ("interval",[1,100]),
                 "passenger_count": ("equal",1)
                 },
     'reset_index' : True
}
```

Figure 4.19.: Query of the Third Use Case

4.3. Extensions

To make the user's work more accessible, a feature could be added to extend the data frame in order to have a better correlations diagram; that is why we introduced another extended correlation analysis, so the end user could identify some exciting facts about the dataset and extract the needed dataset for his application. An extended dataset is only a helpful tool generated by getting dummies from pandas. It will be applied on the following columns 'PUDay','DODay','PUhour','DOhour','PUperiod','DOperiod', 'PUborough', 'DOborough'. Thus, this dataset will not be saved since it does not belong to the final results.

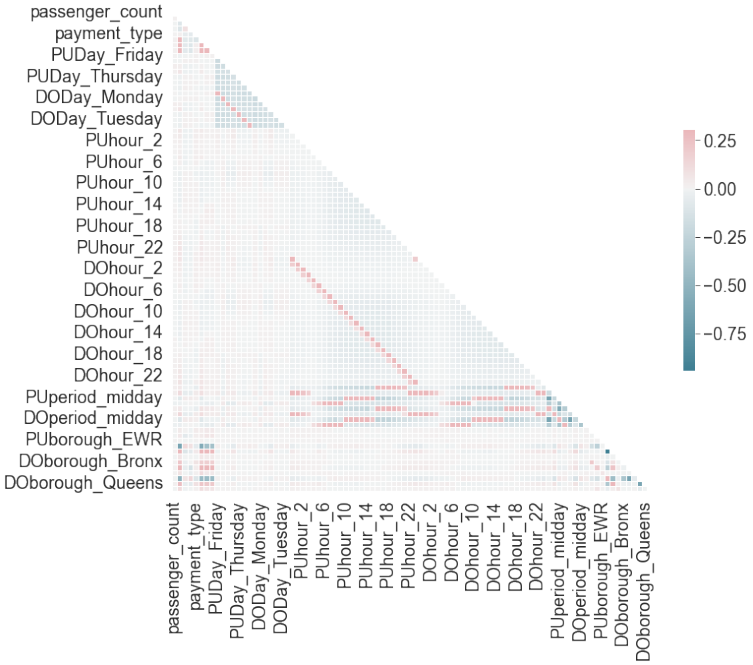

Figure 4.20.: Extended Correlation analysis

4.4. Discussion

During this work, we introduced a data mining model named CRISP-DM, which we used to prepare the data for our advanced work. Then, in the 3.2 part, we presented a simplified function as a solution to generate, after defining preferences, the wished datasets. Finally, in chapter 4, we proceeded with experimenting, running, and concluding what the used method did bring us.

After defining what the function needs as input, a data frame, and a dictionary containing the preferences, we performed the work by defining some use cases to try different functionalities. Furthermore, for the second use case in part 4.2, there was a conclusion that the trip is mainly conducted in Manhattan for those parameters. To argue this statement, we conducted a query on the initial data frame and found that for the selected parameters, we got a data frame of 3161 rows, of which 3125 rows took place in Manhattan, which defends the statement.

As for the limitation part, Since the available dataset was an open source dataset, the first step of CRISP-DM was hard to implement since the business need has two faces; what the data provider is expecting and what we expect from this work, in a business view. That is why the results are more useful for our institute and still have immense potential for the data provider. Testing and Deployment as phases of CRISP-DM could not be implemented since the direct application of the dataset is not available, so the CRISP-DM was only valuable for business understanding and data pre-processing. The proposed solution is based on a few assumptions, which gave a huge space for creativity and scenarios definition, which could have led to some superficialities or more than expected deep data analysis.

Furthermore, the suggested solution also has some limits, as it is designed to treat only one query per column.

Related to ETL-Process, we extracted data here using an API and transformed it from raw form into valuable datasets, but what could be better done is foremost the automatization of the process by creating a UI or a drag and drop tool. It will also be more helpful to define an entity for saving data(data warehouse or other target databases), so the work could be more labeled, and the results could be ready for other potential uses, like dashboarding or analytics.

5. Conclusion

Dealing with significant data frames was always challenging, especially when the goal was creating some automation in terms of data analysis and data extraction. Within this work, we presented a template as a method to facilitate the extraction of meaningful datasets while letting the user define his preferences in querying. Due to unclarities and the enormous amount of information, getting through different data mining steps was necessary to deduce some business insights to surmount blurry visions. To surmount these challenges, we started with implementing the CRISP-DM model to end up with a cleaned and ready-to-use data frame. This step is necessary for any data mining step, and it pushes the users to revise and ameliorate the output of previous steps whenever new knowledge is made available.

It can be seen from the presented work that before the application of selected data mining methods, comprehensive data collection and a systematic data pre-processing process must take place.
This step ensures a current, in-depth understanding of the data and that the data mining processes work more effectively due to the increased data quality. Furthermore, data collection and pre-processing include statistical analysis that makes handling missing and inconsistent values easier.

By having an adequate dataset, the primary workload will be put into algorithms and training of models, we found it more convenient to start reducing the size of the available data in the data cleaning step, but not only the size but the workload put into treating meaningless data should be minimized. For this reason, we have proposed a method facilitating the generation of small datasets only related to the user's wishes and are not limited to some pre-defined forms. Furthermore, the dictionaries containing filter configurations are user-friendly and can be easily customized. The main challenge was to have a clear view of what the next user is expecting from the pre-processing phase and to surmount this point, and we have proposed a general template, customizable to avoid conflicts or limitations of choices.
Our solution remains accurate when business grows, data changes, and sources differ, to the best of our knowledge and experiments. An additional improvement in datasets quality is achieved by selecting relevant facts from conducted analysis, like in the presented examples.

The work dealt with some open-source databases, maintaining high quality from the source. As a potential work, we should consider a quicker way to pre-process the real-time data by creating visualization dashboards, working with drag and drop, so it will not be problematic how many attributes were addressed.
The process should be carried on by implementing suitable VRP-Algorithms on the generated databases, and after that, the user could better define attributes and relations. It is a closed circle, so whenever the implementation of the upcoming steps related to optimization problems needs new attributes or sees potential in any relationship between facts, the dataset to be extracted for the following work will consider this knowledge and start with it as a user preference.

A. Storage Medium

With this thesis is a USB-Stick delivered, containing a copy of the thesis and the written code during the work.

Bibliography

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. 1996.

[2] NACTO POLICY 2019 MANAGING MOBILITY DATA(10.08.2022) https://nacto.org/wp-content/uploads/2019/05/NACTO_IMLA_Managing-Mobility-Data.pdf

[3] LIONG CHOONG YEUN, WAN ROSMANIRA ISMAIL, KHAIRUDDIN OMAR & MOURAD ZIROUR. VEHICLE ROUTING PROBLEM: MODELS AND SOLUTIONS. 2008.

[4] A Survey of Extract-Transform-Load Technology, Article in International Journal of Data Warehousing and Mining · July 2009. Panos Vassiliadis, University of Ioannina, Greece

[5] Microsoft ETL Documentation (14.08.2022) https://docs.microsoft.com/de-de/azure/architecture/data-guide/relational-data/etl

[6] IBM SPSS Modeler CRISP-DM Guide, IBM Corporation 1994, 2011. (01.08.2022) https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

[7] CRISP-DM Documentation (01.08.2022) http://www.crisp-dm.org

[8] Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler). CRISP-DM 1.0 Step-by-step data mining guide

[9] Ahmad, Naveed, et al. The Home Away From Home: An Analysis Of The Lodging Industry In 2015. American Journal of Entrepreneurship, vol. 9, no. 1, Addleton Academic Publishers, June 2016, p. 60.

[10] SOCRATA APP Documentation (08.08.2022) `https://dev.socrata.com/foundry/data.cityofnewyork.us/t29m-gskq`

[11] Statistics How To;(02.08.2022) https://www.statisticshowto.com/probability-and-statistics/z-score/#Whatisazscore

[] Greater N.Y. Taxi Assn. v New York City Taxi & Limousine Commn., 42 Misc 3d 324, reversed. (04.08.2022) https://www.nycourts.gov/REPORTER/3dseries/2014/2014_04156.html

[] Business Analytics, CRISP-DM, ISMLL, University of Hildesheim, Germany, Tomas Horvath