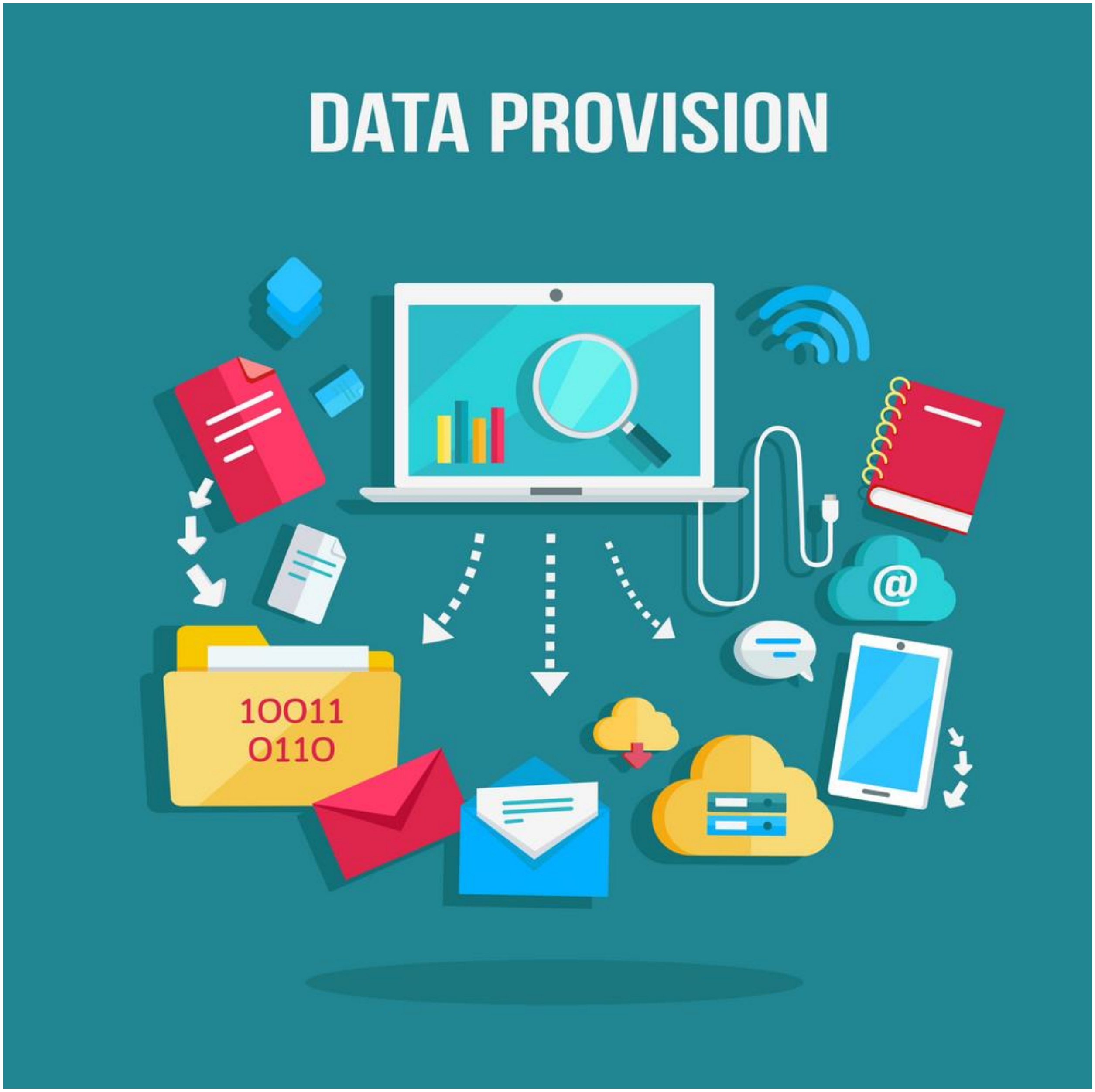

Class: 19-5-2022 **Mentor:**

AI44 Martijn Lamers

Members Pythonatic:

Eva Bijker David van Rijthoven

Antonia Dineva Alex Svetoslavov

Introduction

This document serves as purpose of giving insight into how the data provided for this project can be accessed and stored. It also includes overview of the data requirements and what from data exactly will be used for the solution. Furthermore, you can find information regarding how the end solution will look like, how it can be implemented and accessed.

Table of Contents

Data requirements

Stakeholders

The project will go through several phases. They will be marked by meeting with the client after which feedback will be collected and considered. The stakeholders are everyone that is involved in the project either as a developer, customer, or user. Those roles and their representatives are outlined below.

- The client is Ordina company. Our project tutor is Mr. Lamers. They are the people who will be reported to and keep informed about the project development. They are also the people whose feedback will be collected before the final product is presented and ready to be used by the end user.
- The end users are Ordina company employees. They are the target audience and the one who will use the product.
- Pythonatic is the group working on this project to deliver the product to meet the stakeholders' needs in the end. Based on the planning outlined in this document they will work to deliver the agreed upon product.
- NS/the train companies are also an indirect stakeholder of our project. This is because if we can accurately predict which trains are going to have a delay then they would also want to investigate it so they can improve their services
- We also get out data from "Onofficieel archief reisinformatie Nederlands Openbaar Vervoer" so that also makes them on indirect stakeholder.
- We also are going to get some holidays related data from personeelsnet and overzicht-feestadage.nl so they are also an indirect stakeholder.

Identify required data elements

To be able to make predictions a date is required, and the one the client advice was about is from 2016. Because there are a lot of train data per day at least one month of data will be used, but it would be complete if the other months of the year are also used. That way potential seasonal patterns will be identified. Furthermore, its useful to have information regarding the departure train station and the destination station. The planned departure time/date and the actual departure time/date is required. The delay will be calculated based on that. The train number, train company, the type of train and also an indication of the route it is following (route number or something with all the stops on that route), could be used. It would also be nice if holiday days are noted and also the weekday and an indicator if its in a rush-hour or not.

Identify Candidate data sources

The train data is from source provided by the client: https://trein.fwrite.org/idx/DVS.html. This is an unofficial archive travel information Dutch public transport. They got this information from API's from the NS and NDOV Loket. This website is from Adriaan van Natijne and is giving the licences it needs to function legally. The data there matches the accuracy quality criteria mentioned in the beginning of the document.

The other information this information about the school vacations and holidays/free days. Unfortunately, such dataset was not found, but below some similar ones are listed:

- https://www.personeelsnet.nl/bericht/rechten-en-plichten-vakantie-en-vrije-dagen-2015-2016
- http://www.overzicht-feestdagen.nl/feestdagen/2016/alle-feestdagen.html

and because it seems to be in tables in the website it will probably also be not too difficult to extract it. A possible way to do that in the following link: https://stackoverflow.com/questions/10556048/how-to-extract-tables-from-websites-in-python.

Data quality criteria

Data quality is pivotal when is comes to evaluating whether the data we have can serve its purpose in the given context. The quality of the data used in this project was considered from several quality characteristic described below:

Accuracy – The way to access the data in this characteristic is measuring if the information contained reflects a real-world situation. Given that the data is from unofficial archive regarding Dutch public transport it contains real life data of train travel, and it was recorded accurately.

Completeness – This characteristic addressed if the data is complete and comprehensive. As for the purpose of the project – measuring train delay and its probability we have enough data for train departure and arrival. From the data available the project group can use it to predict whether there will be train delay and how long will that delay be.

Reliability – That characteristic is regarding if the data provided does not contradict information from another source. The data that will be used in the project is from (finish ….)

Relevance - The data that we used should be relevant for the purpose of this project. Since the main questions are, Is there train delay? ‘’ and how long will the delay be? ‘’, the data provided is relevant and can serve answer those questions.

Timeliness – That characteristic refers to whether the data is timely, up to date. The data used is collected and reorder from 2016. The train information is still relevant as the train numbers, types and routes are still the same nowadays.

Overview of the data requirements

First the compulsory elements required:

- Planned departure date/time – ideally, all the months of 2016 will be used, but only one month is also fine to start with. (a combination of date and time)
- Actual departure date/time – sometimes the actual departure is different than the planned one. (a combination of date and time)
- Station name – The name of the station the train is departing from
- Train company – The train company the train is from
- Train number – the number of the train to check if it's the same train.
- Train destination planned – the planned destination of the train ride
- Train destination actual - the actual destination of the train (sometimes it's different, because of delays, problems on the track or another reason)
- Delay – the delay in seconds

Secondly additional features that could be used potentially:

- Route number- the number of the route it is riding on (which then is defined somewhere so we could look up the next stop)
- Planned next stop – the name of the planned next station where the train is going to stop
- Actual next stop – the name of the actual next station where the train stopped
- Number of stops – the total number of stops that a certain train ride is going to stop
- Holiday or not – a Boolean if that specific day is a holiday or not
- Weekday – the day of the week
- Time group – the time group that the train is in which can be something like, early morning, morning rush-hour, in between rush-hours, afternoon rush-hour and late evening.
- Reason for delay – the reason why the delay is happening

Data collection

Determine what information you want to collect

In the data requirements it is mentioned what will be collected. Finding this data will not be time consuming as the train data source is provided by the client. As mentioned, the focus is 2016, because this is requirement set by Ordina. Potentially data from other year will also be used to so see if there are changes etc. However, that could result in confusing for the models. For example, Covid-19 made less people go to work and so impact the amount of trains and so delays.

To broaden the data set other data sources may be used to look for the dates of the holidays for instance. That is not a focus so no more than a day will be spend researching that. Finding additional data will be done by google search.

Where to store your retrieved data

After cleaning the data, there will be 10 gigabytes of data. It will be stored as csv on a USB-drive and on Microsoft Teams files. In case of problems with the internet connection, the data can be extracted from the USB-drive. Every team member from Pythonatic has a copy of the data on their laptop to work with.

Ideally each team member would have fast enough laptop so that the data can just run from the laptops (the laptops would then also need enough storage space). Another option would be to have an online server that is easily accessible, but then to be able to run still a fast laptop is needed.

If the client would want to run the notebooks he/she would also need the download the dataset on their laptop. If it was on a online server access will be granted to them. In the end of the project the client will get insights in the model(s) created in the form of a presentation, where Pythonatic will explain how they could use it and what value it gives to them. Pythonatic will also give them the notebooks and all the other necessary files they would need to run it.

Use a (traceable) version and naming system

There will not be a traceable version and naming system. Microsoft Teams will be used by Pythonatic for documents and uploading of other files. Teams allows multiple people to work on the same document at the same time. The notebooks and other files will be named is such a way that is known which ones are most up to date and we will also upload and keep previous versions.

Determine how and how often you want to retrieve (or reload) the data

Because the data used for this challenge is from 2016, all the daily data must be retrieved at once. The data from 2016 will not change, so it not necessary to reload the data because it will not change.

Data extraction

The dataset that was provided by Ordina contains data in xml format. Therefore, it has to be converted to the xml code into readable data frame, that can be used in the notebooks.

To extract the data, the elements from the xml code need to be retrieved and converted into variables. This is done by using xml converting libraries. Then the variables and their data can be put in arrays and convert to data frame.

However, this will be for one day only, since one csv file, containing the xml data, holds data for one day. Therefore, the extraction code needs to be put in a function and call it for each day of the year, parsing the csv file for each day.

Obviously, with such a huge amount of data, it is difficult to extract the data from our laptops. For that reason, a server is needed to run python scripts for each month, extracting the data for each day in the month. After each script, a data frame will be created containing all the days in the month, combined in one data frame with much less file size (since we are selecting only the features that are relevant to us).

At the end, the data frames for each month will be combined in one data frame and stored in a USB flash drive, so that is accessible the data easily.

After Pythonatic can access this data on the laptop where it will be cleaned. The following rules to clean the data will be applied:

- The data and time data should be in readable form for us and python to read.
- The delay should be in the same measurement, we decided that that measurement would be seconds (so it's a number)
- We should add groups for the different time groups and weekdays

Overall, the following steps will be used for every month:

1. Load the data
2. Delete the index column
3. Put the dates in readable format for python (dates here is a combination of a date and a time)
4. Also put the dates into 2 additional columns, one with only the date and one with only the time
5. Delete the columns we are not interested in
6. Put the delay in a delta format

At the end all those months will be combined with the concat function, which will basically add the months to each other vertically (so it will create more rows).

After the first presentation with the client their feedback was start small so the focus in on January. Because of this Pythonatic decided to do the EDA for January, while doing the EDA the data was cleaned further by making the duration a number in seconds, calculation the max number of stops and adding the weekday and time groups.

Conclusion

The solution that Pythonatic offers is a Jupiter notebook. The notebook will contain prediction models of train delay based on the data provided by Ordina and can be accessed by the employees. The notebook will come with accompanying document to serve as a guide on how the notebook should be used.

References

KleinMedia.nl. (n.d.). *Wanneer vallen de feestdagen 2016*. Overzicht-Feestdagen.Nl. Retrieved May 31, 2022, from http://www.overzicht-feestdagen.nl/feestdagen/2016/alle-feestdagen.html

Onofficieel archief reisinformatie Nederlands Openbaar Vervoer. (2022, April 8). Onofficieel archief reisinformatie Nederlands Openbaar Vervoer. Retrieved May 31, 2022, from https://trein.fwrite.org/idx.html

Personeelsnet. (2014, February 26). *Rechten en plichten vakantie en vrije dagen 2015 - 2016*. Retrieved May 31, 2022, from https://www.personeelsnet.nl/bericht/rechten-en-plichten-vakantie-en-vrije-dagen-2015-2016

Research Papers Risk Assessment for Scientific Data, https://datascience.codata.org/articles/10.5334/dsj-2020-010/