

LISBOA SCHOOL OF ECONOMICS & MANAGEMENT

Individual Project

**Navigating Lisbon: Unveiling the Best Carris Espaço
Navegante through Data Exploration**

Academic year: 2023/2024

Course name: Big Data Tools and Analytics

Daniel Enriquez 60369



**Lisbon School
of Economics
& Management**
Universidade de Lisboa



Abstract

This paper aims to identify the optimal Carris Espaço Navegante within the Lisbon Metropolitan Area. Leveraging Carris API and Google Reviews API through Google Cloud Platform services, the author constructed an ELT Pipeline to gather and process data. A Weighted Sum Model incorporating four pivotal variables—expected wait time during rush hour, weekly operating hours, number of bus stops, and average Google Reviews rating—was employed. Results indicate that Alcochete emerges as the top espaço navegante. It is noteworthy that the findings are dynamic, as data processing occurs continuously, albeit specified intervals. Access to the real-time results dashboard on Looker Studio is provided via the link at the bottom of this abstract. Furthermore, the code utilized for this study is available on GitHub. The author is prepared to furnish the complete dataset, comprising approximately one million rows, upon request.

Keywords: Carris Espaço Navegante, ELT Pipeline, Google Cloud Platform, Weighted Sum Model, Lisbon Metropolitan Area.

GitHub Repository: [danee593/Carris-Espacos-navegante \(github.com\)](https://github.com/danee593/Carris-Espacos-navegante)

Looker Studio Dashboard: <https://lookerstudio.google.com/reporting/3de23040-8ada-4b8a-91fb-1b17bb608d75>

Contents

1	Introduction	5
1.1	Research Question.....	5
2	Methodology.....	6
2.1	Data Gathering	8
2.2	ETL Pipeline.....	8
2.3	Data Warehouse.....	10
3	Results	12
4	Conclusions and Limitations	13
5	Bibliography	15

Figures

Figure 1. Unix-cron job set up for cloud function.	9
Figure 2. ELT Process	10
Figure 3. Logical Data Model Diagram of Carris Espaço Navegante Data Mart.....	12

Tables

Table 1. WSM matrix.....	7
Table 2. Criterion and weights to decide the best Carris Espaço Navegante.....	8
Table 3. Top 10 Best Espaço Navegante in Lisbon Metropolitan Area.....	13
Table 4. Top 3 Worst Espaço Navegante in Lisbon Metropolitan Area.	13

Equations

Equation 1. Weighted Sum Model.....	7
--	---

1 Introduction

The number of passengers using the public transportation system in Lisbon has been increasing over the last years. Only the Metro of Lisbon has transported 27 million passengers between January and February of 2024, this represents a 3% increase compared to 2019 (pre-pandemic), during the same period, and 7% increase compared to 2023, same period (TPN/Lusa, 2024).

Not only the Lisbon Metro has increased, as Agência Lusa (2023) points out in an interview to Carris Metropolitana, in the Lisbon Metropolitan Area, excluding Barreiro, Cascais and Lisbon municipalities, the number of passengers is close to 13.5 million in October and November, and Carris expects to exceed 14 million passengers per month.

These numbers are big and increasing. Given this raise in passengers it's expected that at least some of them are or will become recurrent users of the public transportation system, the cheapest and most convenient way for a recurrent user to access to the public transportation system is with a card, the “navegante® personalizado” card. That allows users to buy a monthly pass.

There are multiple ways to access this card: online, customer service points in particular Metro stations, “Ponto navegantes” or in the “Espaço Navegante”. The fastest way to get the card is online, but if a person does not have a metro station nearby or wants to get this card on the same day the most convenient way is attending a Espaço Navegante, which is a customer service point. Lines in these customer service points can be large and the reviews in Google Maps are usually not the best. This is the reason why the author poses the following research question.

1.1 Research Question

What's the best Carris Espaço Navegante in the Lisbon Metropolitan Area?

2 Methodology

The research question poses a choice between different alternatives, when such type of problems is faced decision theory offers a concise, simple, and effective way to make decisions, or chose between the available options (Hansson, 2005). There are two main methods for decision-making Mono-criterion methods and multi-criterion methods. The difference between them is the number of criteria or variables involved in the decision. This paper will focus on Multiple-Criteria Analysis (MCA).

To make decisions is important to differentiate the key elements involved in decision-making. As Dean (2022) suggests, the key elements are option, objective, and criterion. Options are the different alternatives that can come up for a given decision, in the case of this research paper the options are the various Carris customer service points in the Lisbon Metropolitan Area. Objective is what we try to accomplish with this decision, in this paper the objective is to select the “best”, our most convenient customer service point. Finally, the criterion are indications of the performance of the given options. In this paper the criterion will be certain metrics to evaluate the best or most convenient.

There are many methods for MCA, including some formal methods such as linear programming, goal programming, among others (Dean, 2022). However, for this research the simplest and most intuitive form of MCA will be used given its simplicity and practicality, the weighted sum model (WSM).

Equation 1 summarizes the WSM. Where A is the weighted sum model score for the i -th option, w is the weight and a is the value for the j -th criterion. This results in a list of values that can be comparable and ordered, then the “best” customer service point will be the one with the highest WSM score among the options.

Equation 1.

Weighted Sum Model.

$$A_i^{WSM-score} = \sum_{j=1}^n w_j a_{ij} \text{ for } i = 1, \dots, n$$

This equation can also be represented in a matrix, as seen in table 1. In the left can be seen the options or alternatives, in the upper region in bold text the criterion, in the lowest region the weights, and finally in the right and in bold text the WSM score.

Table 1.

WSM matrix.

Criteria <i>AT alternative</i>	1	2	3	$I_{a,b,c}$
<i>a</i>	2	3	2	2.8
<i>b</i>	3	4	4	3.4
<i>c</i>	5	2	2	3.8
$w_{1,2,3}$	<i>0.61</i>	<i>0.28</i>	<i>0.11</i>	

Note. Adapted from Remotely designed appropriate technology for emergency disaster response in Nepal [table], by Brown & Michael, 2016, Procedia Engineering 159 275 – 283. CC BY-NC-ND 4.0 DEED

In table 2 is a summary of the criteria and weights that will be used for the WSM model. The expected wait time in rush hour will have a higher weight than the rest of the criteria, the best customer service point should have low queue lines or fast service to be the best and this is the most important criterion. The number of opening hours per week and number of bus stops weighs 0.15 each, ideally the best customer service point should be open for many hours and have many busses stops to get there. Finally, the average rating on Google Reviews weighs 0.25, this number is high but not so much that it overshadows other criteria, reviews are useful, but Google Reviews are not verified, so everyone can complain about a customer service point

without even going there or can rate it lower based on personal perceptions of the overall public transportation system in Lisbon.

Table 2.

Criterion and weights to decide the best Carris Espaço Navegante

Criterion	Weight
Expected wait time in rush hour	0.45
Number of opening hours per week	0.15
Number of bus stops	0.15
Average rating on Google Reviews	0.25

2.1 Data Gathering

For this research secondary data sources will be used to find the best Carris Espaço Navegante in the Lisbon Metropolitan Area. The design of this research is documentary because secondary data will be collected and analyzed.

The data is collected from the Carris API (2024), and from Google Reviews API (2024). Google provides historical data, whereas Carris does not. Carris only provides real-time data, therefore, this paper proposes to store the data collected from both APIs in a data lake to have the raw information and then create a data warehouse to store the cleaned datasets. Google Cloud Platform (GCP) provides an environment to set up the data collection, lake storage, data cleaning, warehouse storage and data analytics.

2.2 ETL Pipeline

Google reviews API will be used once per customer service point, and this data will be stored in a csv file that will later be incorporated into the ELT pipeline, there won't be any set up to keep ingesting data from Google reviews. On the other hand, for the customer service

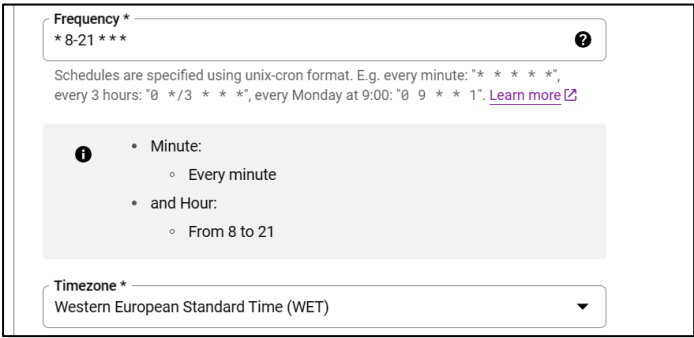
points there is no historical data, therefore a system must be developed to continuously ingest and store this information.

To continuously ingest and store this information the microservice Cloud Functions will be used, it consists of a python script that queries the Carris API every minute, transforms the Json file into a dataframe and then inserts this dataframe to a Big Query table, the dataframe must match the schema of this Big Query table.

To set up this cloud function to run every minute a Unix-Cron job must be programmed on GCP, this cloud function will run every minute to match the opening hours of the least restrictive customer service point, this is the “Costa da Caparica” customer service point that is open every day from 8:00 until 21:00 interruptedly, as seen in figure 1.

Figure 1.

Unix-cron job set up for cloud function.



Frequency *

* * * * *

Schedules are specified using unix-cron format. E.g. every minute: "* * * * *", every 3 hours: "0 */3 * * *", every Monday at 9:00: "0 9 * * * 1". [Learn more](#)

- Minute:
 - Every minute
- and Hour:
 - From 8 to 21

Timezone *

Western European Standard Time (WET)

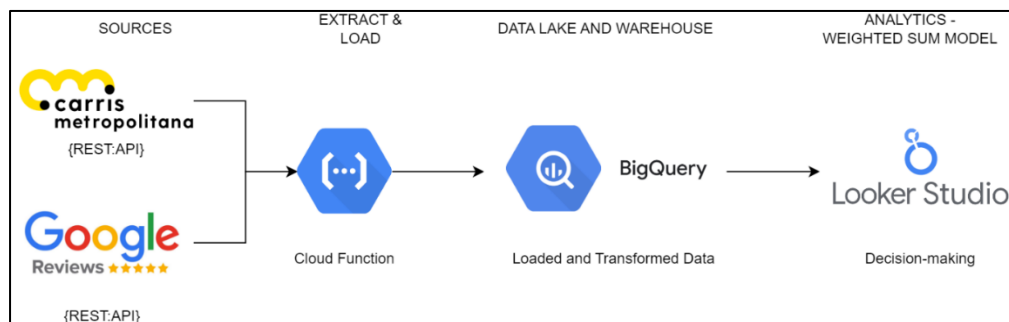
Once the cloud function is configured and running and the csv file containing the Google Reviews data has been uploaded to a Big Query table the ingestion to the data lake is concluded. This data is in raw format, most of the columns are in string format and some cleansing needs to be done before moving this data into another database where the Warehouse will be located.

The whole ELT process can be seen in figure 2. The data sources are Carris API and Google Reviews API, the extraction and loading phase is done with a cloud function and a Unix-cron job. Then this data is stored within Big Query in a data lake. This raw data is then

processed within Big Query with scheduled SQL queries to clean the data and move it to the Warehouse database, after this process the data can be analyzed within Looker Studio with the WSM to answer the research question. The Data Warehouse proposed for this research will be analyzed in the next section.

Figure 2.

ELT Process



2.3 Data Warehouse

The Data Warehouse proposed in this paper will follow the star schema proposed by (Kimball & Ross, 2013) for Enterprise Data Warehouse (EDW). This EDW consists of a single DataMart that consists of two dimensions, espaço navegante and time, two fact tables, customers queue and customer reviews, the logical data model diagram can be seen in figure 3.

The dimension espaço navegante consists of an id that serves as primary key, name of the customer service point, location in coordinates, phone, address, postal code, municipality, district, shift, number of hours open per week and number of bus stops. This dimension has an interesting property, the number of hours open per week and number of bus stops might vary with time, making it a slowly changing dimension. This will be considered a type 1 and be overwritten if changes (Kimball & Ross, 2013).

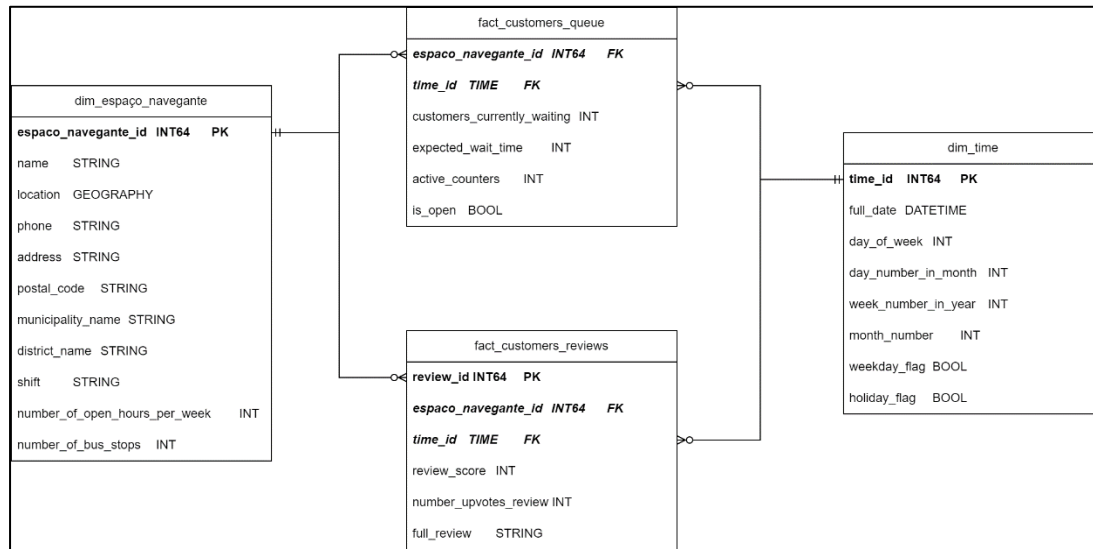
The next dimension is time. This is a straightforward dimension, it contains time id as primary key, full date as a timestamp, note that the granularity of this table is per minute, the rest of the attributes of this table are mainly for convenience of reporting to not calculate new attributes on the BI tool.

The fact table customers queue has espaço navegante id and time id, both columns compose the primary key and foreign keys of this table, it is connected to both tables in a one-to-many relationship. It's important to note that the granularity of this table is one row per customer service point and minute, as this matches the data extraction process of the ELT pipeline. The measures of this table are number of customers currently waiting in the line, expected wait time, number of active counters, is open is a flag attribute that informs if at that timestamp that customer service point is open or closed.

Finally, the fact customers reviews table has espaço navegante id and time id, both columns compose the primary key and foreign keys of this table, it is connected to both tables in a one-to-many relationship. The granularity of this table does not match the other fact table customer's queue. This entails an important consideration, measures between both tables can not be comparable, that's one of the reasons why the weight given to Average rating on Google Reviews criterion is low in the WSM. The other important consideration is that the dimension time that has a granularity of minute won't work with this table, therefore a view of time dimension will be created to match the granularity of this table. The important metric in this table is the review score.

Figure 3.

Logical Data Model Diagram of Carris Espaço Navegante Data Mart



3 Results

Results are published to the public in a dashboard on Looker Studio, results might be changing daily, since the pipeline will continue to ingest data every minute. The results presented in this paper are until March 23, 2024.

Table 3 shows the top 10 best Espaço Navegante in the Lisbon Metropolitan Area. Notice that in this table the WSM score is presented in the first column and the values without normalization can be seen from column 3 to 6. It's evident that since the largest weight was given to the expected wait time in rush hour the best places will be where you wait the least amount of time during a rush hour. It's interesting to note also that the only 2 espaços navegante with 5 stars on Google Reviews have a low wait time during rush hour but higher number of bus stops compared to the “best”. This suggests that people might be more likely to give a positive review if you wait less time during a rush hour, however that correlation goes beyond the scope of this report.

Table 3.

Top 10 Best Espaço Navegante in Lisbon Metropolitan Area.

Name	WSM Score ▾	Expected wait time in rush hour	Average rating on Google Reviews	Number of opening hours per week	Number of bus stops
Espaço navegante Alcochete	0.82	1	4.8	55	4
Espaço navegante Brejos de Azeitão	0.58	2	4.58	55	3
Espaço navegante Palmela	0.56	4	5	55	10
Espaço navegante Mafra	0.49	5	5	55	6
Espaço navegante Bucelas	0.45	4	4.14	55	4
Espaço navegante Setúbal	0.44	17	2.79	91	12
Espaço navegante Costa da Caparica	0.44	9	3.91	91	4
Espaço navegante Loures	0.42	15	4.47	73.5	4
Espaço navegante Alverca	0.41	6	3.57	55	6
Espaço navegante Vila Franca de Xira	0.4	5	3.5	55	4

In table 4, the worst espaços navegantes in Lisbon Metropolitan Area are shown. Given the weights that were selected for each variable it's evident that the customer service point where customers are expected to wait more in rush hours will get a lower score, but it's interesting that some of the worst customer service points also score lower on Google Reviews, and have less number of opening hours per week than the maximum number of opening hours per week among all of the espaços navegante.

Table 4.

Top 3 Worst Espaço Navegante in Lisbon Metropolitan Area.

Name	WSM Score ▲	Expected wait time in rush hour	Average rating on Google Reviews	Number of opening hours per week	Number of bus stops
Espaço navegante Sacavém	0.31	24	2.5	40	8
Espaço navegante Oeiras	0.33	13	2.75	55	3
Espaço navegante Amadora (Estação)	0.35	53	1.78	55	14

4 Conclusions and Limitations

The best espaço navegante in the Lisbon Metropolitan Area is Alcochete, in this customer service point customers are expected to wait the least in rush hours, has a good average rating on Google Reviews, it's open more than 40 hours per week but the number of bus stops is low compared to other espaços navegante. Given this limitation, if a person wants to sacrifice some minutes of waiting time but gain in the number of bus stops Palmela might

be a good option. The average wait time during rush hour is only 3 minutes more, the average rating on Google Reviews is better than Alcochete the number of hours opened per week is the same, but it has 6 more bus stops than Alcochete.

The limitations found in this research paper are:

- Time span discrepancy: Google Reviews cover a longer time span compared to the Carris table, making direct comparisons non-specific due to differing granularity. Some reviews go as far as 2017. The customer service might have varied significantly between 2017 and 2024.
- Review verification: Reviews are not verified, leading to a broader range of complaints beyond customer service issues or waiting times.
- Data quality: In the Carris (2024) documentation, states that the data is real-time, but some exploratory analysis suggests that is it not the case with the *espaços navegante* table. In a given hour even if the number of customers waiting is high there is practically no variability in an entire hour, if the data is indeed real-time the variability occurring within an hour should be higher but the number of customers waiting or the expected wait time is almost the same number in the entire hour, changing towards the beginning or end of that hour and staying the same for almost 55 minutes to 60 minutes.
- Model complexity and subjectiveness: WSM is a powerful yet simple decision model, however the weights are subjective, the criteria considered in this paper might not be the same for every person, some people might value more the reviews, or some might value more the number of buses stops even if you must sacrifice waiting times.

5 Bibliography

- Agência Lusa. (2023, 12 30). *Carris Metropolitana quer ultrapassar 14 milhões de passageiros por mês*. Retrieved from Observador:
<https://observador.pt/2023/12/30/carris-metropolitana-quer-ultrapassar-14-milhoes-de-passageiros-por-mes/>
- Brown, A., & Michael, A. (2016). Remotely designed appropriate technology for emergency disaster . *Procedia Engineering* 159, 275 – 283.
- Carris Metropolitana. (2024). *GitHub*. Retrieved from Carris Metropolitana API (Beta):
<https://github.com/carrismetropolitana/api>
- Dean, M. (2022). *A Practical Guide to Multi-Criteria Analysis*. Retrieved from ResearchGate: 10.13140/RG.2.2.15007.02722.
- Google LLC. (2024). *Business Profile APIs*. Retrieved from Work with review data:
<https://developers.google.com/my-business/content/review-data>
- Hansson, S. (2005). *Decision theory: A brief introduction*. Stockholm: Royal Institute of Technology (KTH).
- Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3ed*. Indianapolis: John Wiley & Sons, Inc.
- TPN/Lusa. (2024, 03 17). *27 million passengers on Lisbon Metro*. Retrieved from The Portugal News: <https://www.theportugalnews.com/news/2024-03-17/27-million-passengers-on-lisbon-metro/86999>