**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

Obraz zawierający tekst

Opis wygenerowany automatycznie

**Abstract**

Contents

[1 Introduction 4](#_Toc161852038)

[1.1 Research Question 4](#_Toc161852039)

[2 Methodology 4](#_Toc161852040)

[2.1 Data Gathering 6](#_Toc161852041)

[2.2 Cloud Functions 7](#_Toc161852042)

[2.3 ETL Pipeline 7](#_Toc161852043)

[2.4 Data Mart 7](#_Toc161852044)

[3 Results 7](#_Toc161852045)

[4 Discussion 7](#_Toc161852046)

[5 Conclusion 7](#_Toc161852047)

[6 Bibliography 8](#_Toc161852048)

# Introduction

## Research Question

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

Kind Regards,

[27 million passengers on Lisbon Metro - The Portugal News](https://www.theportugalnews.com/news/2024-03-17/27-million-passengers-on-lisbon-metro/86999)

["Regularity of Public Transport Usage" by Stefan Foell, Santi Phithakkitnukoon et al. (usf.edu)](https://digitalcommons.usf.edu/jpt/vol19/iss4/10/)

[Portal do INE](https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0007963&contexto=bd&selTab=tab2)

[Carris Metropolitana quer ultrapassar 14 milhões de passageiros por mês – Observador](https://observador.pt/2023/12/30/carris-metropolitana-quer-ultrapassar-14-milhoes-de-passageiros-por-mes/)

Passageiros transportados keyword.

# 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.

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.

A table with numbers and symbols

Description automatically generated

*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 (Brown & Michael, 2016)

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 |

## 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.

## 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.

A screenshot of a computer

Description automatically generated

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

A diagram of a diagram

Description automatically generated

## 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 espaco 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 espaco 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

A close-up of a document

Description automatically generated

This data model will be used to calculate the WSM model in Looker Studio and visualize it.

# Results

dada

# Discussion y Conclusions

# Bibliography

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 Defi nitive Guide to Dimensional Modeling, 3ed.* Indianapolis: John Wiley & Sons, Inc.