SDAIA T5 bootcamp

Exploratory Data Analysis (EDA)

Course Project Introduction

The Metropolitan Transportation Authority (MTA)

Writeup

Basma Ghazi Abdullah Alduaiji

29/09/2021

Contents

Abstract 3

Design 3

Data 3

Algorithms 3

Tools 4

Communication 4

Conclusion 4

Abstract

Because of the lack of Metropolitan Transportation Authority (MTA) transport project and the extreme growth of population within New York City, the crowded train stations and the long waiting time and lack of arrival time accuracy in some cases respectively are crucial issues the consultant company will recommend solutions for it.

The intrinsic goal of this project is to study and understand the MTA data set provided by the Metropolitan Transportation Authority then design, plot representative graphs. After that, the analysis provides a solution that will enable the population of New York City and tourists to easily access the MTA stations and increase the satisfaction levels of New York City people and MTA passengers.

Design

As a well-known consultant company working in data science, which provides consulting and solutions throughout analysing data. The last week, the company has received an email from a government institution. The institution works on improving New York City. The email contains a proposal to provide a solution to organize the overwhelmed stations specially during rush hours and solve the other mentioned issues, to reach the required satisfaction levels using data since.

The data set in this project is taken from the MTA website. This data set presents the complete status of the metro in New York City. However, data analysis is used to know the most crowded stations to expand stations space as a short-term solution and add extra stations close to these stations as a long-term solution.

Data

The dataset contains over two million rows with eleven features for each provided by the MTA. Moreover, the columns are station names, lines names, number of entries and exits, date and time, number of entries and exits. Nearly a turnstile of the individual features could be grouped into more general category parts such as control area, unit and subunit channel position.

Algorithms

The steps to analyze the MTA data set are gathering data from the MTA data set within the needed months. However, the months of this project are May, June and

July, exploring data by using all the functions like info and describe ext. Then cleaning data by removing null values and duplicates as well as, remove any data mistakes. After that, plot the graphs using seaborn and matplotlib modules from python. Finally, connect with the SQLalchemy database and write some representative queries.

Tools

- Numpy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting
- Python word cloud to find most repeated word in the data set
- Python Date and Time to deal with date and time columns
- Python SQLalchemy to connect with database

Communication

Moreover, the code and graphs will be available on my GitHub BasmaGAlduaiji/MTA-project (github.com).

Conclusion

The aim of the company project was the analyse and understand of MTA data set. First, the company has started by gathering data. Then, exploring of data. Afterwards, the company focused on the clean of data by removing duplicates and null values using pandas. In particular, the company depicted the plot of the main graphs of our data set. Then connect with the database and write some queries. As a result, the company provides a solution for the short term to expand the station area and increase the number of stations near to the busiest ones as long-term solution.