
# EDA project Final Report

IKEA Advertisement campaign

# IKEA Advertisement campaign

## Abstract:

The goal of this project was to help IKEA to do Advertisement campaign based on the analysis of MTA data and we want to know which stations are the busiest in order to distribute IKEA sleeping capsules to those stations and which days are the busiest also which time slot of the day is more busiest.

## Design:

The data is provided by MTA website and the data represents the daily entries and exists of NYC subways and I did the analysis to answer three main questions

- What are the top stations on traffic to do the campaign on those stations?
- Is weekday or weekend better to do the campaign on?
- What is the best time of the day to do the campaign?

## DATA:

The dataset contains 2722581 turnstile readings for different time slots (4 hours )with 11 features and the combination of the Features [C/A,UNIT,SCP,STATION,LINENAME,DIVISION] represent a unique reading for one turnstile in a station and the entries & exists represent the cumulative entries & exists for a station.

# Algorithms:

## Data gathering:

I web scraped the dataset from the MTA website using python code then I stored the dataset in a database called MTA.db then I put the data in a table called MTA.

I selected the period June to Sep 2021 of the MTA data because I wanted to study the current behavior of the ridership's.

## Data assessing:

I assessed the data using two methods:

Visually: by looking to the dataset and try to understand the data and it's pattern

Programmatically: by using pandas methods like .info() and doing masks to understand the data more clearly.

## Data Cleaning:

I checked for missing values but there wasn't and I found 29 duplicated turnstile readings and I decided to drop them since the have the type 'recover'.

When I calculated the hourly entries from the cumulative entries I faced three issues:

- Sometimes the counter of the entries resets so I checked for that and I returned 0 in that case
- When I used the diff function to calculate the entries some of returned values wore negative so I return the absolute value
- I faced a lot of outliers after calculating the hourly entires so I decided to remove them using the IQR by taking the values above 25% and under 75% from the entries

## Tools:

- Technologies: SQL, SQLite ,Python, Jupyter notebook
- Libraries: Numpy, Pandas, Matplotlib, Seaborn

## Communication: some of screenshots of my presentation