MTA Exploratory Data Analysis Project Proposal

By: Mashari Almuhanna

Introduction

a lot of people in New York using The Metropolitan Transportation subways as main way for transportation. The MTA says that it is serving a population of 15.3 million people across a 5,000-square-mile travel area surrounding New York City through Long Island, southeastern New York State, and Connecticut. So it has a huge amount of traffic daily. So from that traffic the advertising companies will not miss that so when you enter any subway station you will notice there are many and huge billboards every where in the station.

Problem

The problem here that these advertisement methods are old and not compatible with the traffic and what are they interested on.

So, I will analyze data to investigate if there are some patterns from the visitors to the MTA stations. To have better billboards and to have compatible advertisements to the people and to know what the value of the billboard and how many people is will see it.

Dataset

I Will using the datasets from The Metropolitan Authority website ((http://web.mta.info/developers/turnstile.html) and they are publish every week their turnstiles data it has 8 columns.

- C/A: Control Area
- UNIT: Remote Unit for a station
- SCP: Subunit Channel Position represents a specific address for a device
- STATION: Represents the station name the device is located at
- LINE NAME: Represents all train lines that can be boarded at this station
- DIVISION: Represents the Line originally the station belonged to
- DATE: Represents the date in (MM-DD-YY) format
- TIME: Represents the time (hh:mm:ss) for a scheduled audit event
- DESC: Represent the "REGULAR" scheduled audit event
- ENTRIES: The cumulative entry register value for a device
- EXITS: The cumulative exit register value for a device

I will add some features to the data using columns from above to have better analysis. The features for now will be:

- ENTRIES_NUM : The number of entry register value for a device
- EXITS_NUM: The number of exit register value for a device
- Traffic : this will be the traffic for a device(ENTRIES_NUM+EXITS_NUM)

Tools

- Python3
- Jupyter notebook
- Pandas
- Numpy
- Seaborn
- Matplotlib