MTA TURNSTILE- DATA ANALYSIS

PROJECT PRPOSAL

Written by: Wafaa Alharbi

Introduction

October is the Breast Cancer Awareness Month, Unfortunately, breast cancer touches 1 in 8 women (12.4%) over the course of her life time according to statistics from www.breastcancer.org.

In US for women, breast cancer death rates are higher than those for any other cancer, except lung cancer

the Estimated new cases in 2021 is 284,200 for all states in us and 17,540 cases in NewYork since the beginning of the year.

Glamified brand is proud to debut 2 new eyeshadow palettes that will support 3 organizations that do outstanding work for breast cancer education, research, awareness, and direct patient support.

From launch on 09/01/2021 at 3PM CST through the end of November, $2 from each sale of the palettes (up to $200,000) will benefit The Breast Cancer Research Foundation, Living Beyond Breast Cancer, and The National Breast Cancer Foundation.

With your support, we are determined to meet our total donation goal of $200,00, Whereas the company is interested in harnessing the power of data analytics to optimize the effectiveness of street team work which is crucial to boost the sales and help women in their fight of Breast Cancer.The street team will be placed at the entrances to subway stations to market the new products.

Data Description

- This project based on the following datasets:

MTA Turnstile dataset located at web.mta.info/developers/turnstile.html

United states census located at data.census.gov

- Data was used in this project for a period of more than 3 months from SEP to NOV 2020.
- I used this close period to get a quick result in this period because these period after the corona pandemic, and the behaviour of people similar to the period of this year 2021 so we can produce approximate results.
- Number of rows = Approx. 208K row/week
- Number of columns = 15 columns

I will add more features(columns) which are the following:

• Turnstile_location = which is a combination of C/A + unit + SCP can be used to locate the near

by places around the turnstile on google map

• entries_num = which is the number of entries for the station timestamp observed by taking the difference of the cumulative entries and the previous one.

• exits_num = which is the number of entries for the station timestamp observed by taking the difference of the cumulative exits and the previous one.

Tools Description

To carry out the project and explore the data, I will be using:

Technologies:

- Jupyter notebook
- SQLite
- Tabluea

Libraries:

- Pandas
- Numpy
- SQLAlchemy
- Matplotlib
- seaborn