Group 5: Project Proposal

Team Members:
Nate Marohl
Mike Nutile
Andrew Seaman

Option 1: SF Crime Reports [01/01/2018 - 11/11/2020] vs SF Real-Estate pricing [11/11/2020]

For this analysis our team will utilize two datasets sourcing data from public Police reports and private real-estate data. Our objective will be to investigate the correlation between crime and real-estate pricing.

In order to accomplish this we will have to merge the datasets into a common table; clean the police data; and perform a correlation analysis between the cleaned police data vs. housing zip-codes/city names to see if there are any impacts, between area, crime, types of crime, housing prices etc.

https://www.redfin.com/city/17151/CA/San-Francisco
https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783/data

Option 2: Election Tweets:

This dataset has approximately 2M records, half of them #donaldtrump and half of them #joebiden. We can analyze the top emoji's being used (if this is possible) for each hashtag. We can see if certain emoji's are more correlated to #donaldtrump and #jorbiden. We could even do a time series analysis showing how emojis have changed over time. If we are unable to analyze emojis, we can check if certain words are correlated with either #donaldtrump or #joebiden.

Another option we may pursue is testing if tweets are created by twitter bots. When trying to determine if a tweet was created by a bot we can look at IP address correlations, time based correlations of tweets, account creation date, and account activity. If we are unable to perform analysis on twitter bots we will stick with analyzing emojis.

https://www.kaggle.com/manchunhui/us-election-2020-tweets?select=hashtag_joebiden.csv

Option 3: Used Car Market Analysis between USA & Ukraine

For this analysis our team will utilize two datasets sourcing data from both the US and Ukraine used car markets. Our objective will be to discover if there are any similarities or differences between the two markets in terms of car preference, the amount asking for each type of car by location, and the year of the vehicles.

In order to accomplish this, our team will have to merge the datasets into a common table and correlate different named columns.

https://www.kaggle.com/doaaalsenani/usa-cers-dataset
https://www.kaggle.com/antfarol/car-sale-advertisements