W200 - Project 2 - March 2021

Estrella Ndrianasy, Viswanathan Thiagarajan, Kumar Narayanan

Repo: https://github.com/UC-Berkeley-I-School/Project2_Narayanan_Ndrianasy_Thiagarajan

Introduction:

This document is a collection of analysis of the datasets and questions considered for Project 2. The following proposal options and datasets are submitted to Professor Gunnar Kleeman for approval. Awaiting further instructions from the instructor regarding the approved option/dataset for refinement of the proposal with additional initial plots and tables.

Analysis: Car Accident Analysis in 2014

Dataset:

The car crash data set (**CrashReport2014.csv)** has the following properties.
Link: https://www.kaggle.com/qcarver/crashes-2014-csv

- Size: 128.61MB
- 292019 row and 80 columns

- Out of the 80 columns the interesting ones are:
  - ROUTE: The Highway or the road where accident happened
  - YEAR, MONTH, DAY, HOUR, and DAY_O_ WEEK
  - NUM_VEH, INJURIES, FATALITIES, COLL_TYPE
  - WEATHER, LIGHTING
  - SURF_COND, RD_DEFECT, RD_FEATURE, TRAF_CNTRL
  - DRIVER_1, VEH1_TYPE, VEH1_MANUV, VEH1_EVENT1, VEH1_LOC1
  - XCOORD, YCOORD, INTERSEC, WorkZone, WorkZoneTy, WorkersPre, ExceedSpee, CellPhoneU

- **Proposed questions:**
  - What time of the day (based on LIGHTING) most accidents occur?
  - Is there a particular day of the week (based on DAY_O_WEEK) when accidents peak?
  - What are the different types of accidents?
  - What weather condition caused most accidents?
  - Is there a correlation between location (X/Y Coordinates), WorkZone, Speed limit exceeded, Cell Phone Usage and accidents?
  - What type of vehicle causes most accidents?
  - Are there correlations between factors such as presence of Intersection, Work Zone etc. to accidents?

Analysis: Wildfires in the US: Causes and Development

Dataset:
Subsample (50,000 random observations) of 1.88 million US fires from 1990s to mid 2010s. The data is combined with historical weather data at specific longitude and latitude, historical vegetation data, and a fire remoteness metric.

Source: U.S. Wildfire Data on Kaggle
File: FW_Veg_Rem_Combined.csv
Specifications: 42 variables, 55366 observations

Selected columns of interest: fire size, fire class, stated cause description, latitude, longitude, discovery clean date, containment clean date, fire magnitude, weatherfile, temperature, wind, humidity, precipitation, remoteness.

Proposed questions:

- Risk factors in wildfires: likelihood to occur
  - Sample analysis on an area's potential to experience a wildfire based on environmental components
- Speed of development and containment of wildfires
  - Analysis based on environmental parameters and past history in the area
- Study on wildfire duration from discovery to put out time
  - Time duration of wildfires based on impacted environment
- Link between wildfires and population density and/or relationship to distance with closest populated area
  - Determine the relationship between remoteness and chances in wildfires occurrences

Analysis: Deaths in US Jails (2008 to 2019)

Reuters journalists filed public records requests to gain death data from 2008 to 2019 in the nation's biggest jails. Reuters examined every large jail in the United States, those with 750 or more inmates. The data covers 523 jails or jail systems in the US.

Dataset: Dying Inside: The data behind @Reuters investigation of US jail deaths

File: all_deaths.csv (https://graphics.thomsonreuters.com/data/jails/Allstatesinsurvey.zip)

7571 entries and 22 columns
Columns of interest: jail, county, state, date_of_death, date_incarcerated, cause_short, dob, race, gender, custody_status

File: all_jails.csv (https://graphics.thomsonreuters.com/data/jails/Allstatesinsurvey.zip)

523 entries and 16 columns
Columns of interest: jail, county, state, the columns related to number of deaths by cause for each year from 2008 to 2019 (84 columns), the columns related to the average daily jail population (12 columns), columns related to the healthcare provider at these jails each year - private vs public (12 columns)

GDP data: GDP per capita by state between 2013 and 2017
Bea-gdp-by-state.csv: https://www.kaggle.com/solorzano/gdp-per-capita-in-us-states

60 entries and 7 columns
Columns of interest: State level GDP per capita data between 2013 and 2017 (6 columns)

Questions that are of interest:

1. What is the trend of death rate of inmates across the large jails from 2008 to 2019
    a. In the US
    b. In the states
    c. In the counties within a state
2. What is the major cause of death across the years and how each cause either increased or decreased when compared to one another
3. The percentage of US overall vs state level deaths
    a. By race and gender
    b. By incarceration type (under trial vs convicted)
4. Correlation between health care providers at these jails (public vs private healthcare) and the death rates. Is there a trend that we can observe based on the type of provider?
    a. How are the trends between the private healthcare providers? (Is there a provider who is better than others?)
5. How are the death rates in the states correlated with the GDP per capita of these states between 2013 and 2017? (For e.g. Does higher per capita mean a lower death rate?)
6. How are the states doing in terms of per capita vs death rate (for e.g. a state with higher per capita is doing worse in terms of death rate than a state with lower per capita)