***Project #1 – Team Pillow – Project Narrative***

*Jamie R, Chris, Melissa M, Rahul, Andy*

Intro – Why CRASH data?

Team Pillow, during its initial brainstorming session, chose to investigate factors that contribute to the frequency of motor vehicle accidents and resulting injuries and fatalities in the United States. We selected this project subject based on: general group interest in the topic, a wide range of brainstormed questions (and thus potential contributing variables) to research further, and the known availability of large and consistent data sets.

What VARIABLES did our team analyze?

Our group came up with multiple questions regarding possible variables that might impact the frequency of accident fatalities. Inquiries fell into three general categories: *Who? When? Where?*

# Who

*Which (if any) characteristics of the* ***PERSON(S)*** *involved impact accident/fatality frequency?*

* + **Impaired/Distracted Driving**
    - Do the majority of traffic fatalities involve an impaired or distracted driver?
    - What causes the most fatalities: driving impaired, driving distracted, or a combination?

# When

*In what ways (if any) does the* ***TIMING*** *of the accident impact fatality frequency?*

* + **Time of day** (e.g. 4:00-5:00 pm local)
    - Does the frequency of accidents and/or fatalities increase during rush hour (or correlate to congestion or # of vehicles present on the roads?
  + **Day of week** (e.g. Sunday)
    - Does the frequency of accidents and/or fatalities increase on the weekends (Fri-Sun)?
  + **Time of year** (e.g. January)
    - Does the frequency of accident/fatalities increase during the winter months.

# Where

*What (if any) attributes of the* ***LOCATION*** *of occurrence impact accident/fatality frequency?*

* + **Geographic Location**
    - Do urban areas have more accidents/fatalities than rural areas?
    - Do areas with seasonal changes (e.g. winter) have more accidents/fatalities due to the change in driving conditions? Or do those areas have fewer due to drivers being more cautious/avoiding the roads?
  + **Roadway Type**
    - Does the type of road (residential vs. freeway) impact the number of accident/fatatlities?
  + **Intersection Type**
    - Do intersections increase the likelihood of traffic fatalities? If so, are stop lights versus four-way stops more dangerous?

Data Exploration

We initially did a broader exploration of the types of files and the data attributes available in each. We started with the full FARS/GES set of data files (totaling 100+). We landed on the auxiliary files, largely for the reason they were created; they include more meaningful variables that have already been derived using known classifications. Once we settled on the FARS auxiliary files, we ended up dividing those three files for even deeper dive exploration.

The Data – FARS & GES

We chose two primary data sources, the **Fatality Analysis Reporting System (FARS)** distributed by the *National Highway and Traffic Safety Administration (NHTSA)* and the **General Estimates System** **(GES)** put out by the *National Automotive Sampling System (NASS).* These sources were standardized in 2011 in terms of data nomenclature and terminology in 2011, which allows the data sets to be more easily compiled, queried, and compared.

# FARS/GES Data Set

* The entire FARS data set includes national motor vehicle crashes resulting in at least one human fatality within 30 days of the date of the accident.
* The GES data set is much more inclusive, including non-traffic, non-fatal crashes in addition to those included in the FARS data set. It is a random sample, however, and only represents about x% of the total annual accidents that occur.
* This data is collected/compiled from various state-level documents: police reports, death certificates, state vehicle registration, medical examiner reports, state driver licensing files, state highway department, and others.

# FARS/GES Auxilary Data

* Since 2010, FARS has put out three auxiliary data sets in the categories of: accident, vehicle, and person. ***These are the data sets we used for the majority of our analysis.***
* The auxiliary file contain 60 FARS variables and 40 GES variables, along with new variables derived using *National Center for Super Computing Applications (NCSA)* analytical data classifications. Examples are “speeding-related” and “distracted driving.”
* We chose to look at 2015 primarily, as that was the most recent year in the Data World library where we initially pulled the files. We also pulled 2017 data later on, and did some analysis on that year’s data set. We did not see any significant difference in variable relationship between 2015 and 2017, so chose not to switch everything over to 2017.
* We did extensive exploration of all three auxiliary data sets, primarily focusing on the FARS (or fatality-specific) data. Below is a description of the information types housed in each file:
  + **Person** – data primarily concerning the human attributes of those involved in the crash (particularly those who die as a result): age, race, seat belt use, injury/fatality type, whether fatality was driver, passenger, pedestrian, bicyclist
  + **Vehicle** – data primarily concerning the attributes of the motor vehicles involved in the crash: impact point, license status, vehicle type, damage measures, etc.
  + **Accident** – data primarily concerning all other crash attributes including: date/time info, involvement of impaired/distracted/drowsy driver, crash type, road/intersection type, road conditions

Data Clean-up

Pandas was used in Jupyter Notebook to clean-up the data set and make it more accessible. We reviewed raw data file column headers (data attributes); used FARS/GES analytical user manuals to create conversion code to map column headers to meaningful business headings (example: *Column header <A\_DRDRO>* was changed to *<Drowsy\_Driver>.*

We also reviewed the raw file column values for clarity; again we used the Analytical User Manuals code descriptions to create conversion chart to map raw coded variables to meaningful business attribute set; imported raw files to Jupyter Notebook, used “map function” in pandas to execute column variable cleaning. One example would be changing the column <Weekday> value set of numeric values <1-7> to using key-value pairs <1=Sunday, 2=Monday, etc>.

We also combined some data objects in order to form more limited data sets. For example, we have a visual of every state and the frequency of fatalities. During clean-up, the data point “region” was derived using the “state” data column. For example: MN, ND, SD, et. Al = North region.

During analysis, we used the CityPy library to pull in “nearest city” to a crash based on latitude/longitude values. This was then used in conjunction with the Google Maps API to create a visual showing the frequency of crashes via a heat map.