Documentation for P1:
Traffic data was taken from a database in kaggle (https://www.kaggle.com/nhtsa/2015-traffic-fatalities/data). From this database, we examined the spreadsheets accidents.csv and persons.csv. Population data was taken from a file, “People.csv”, provided in the 2/21 lecture.

The main file we used from the, accidents.csv, included among several other variables time(year, month, day, hour, minute), location(2 digit state code, county GSA code, city GSA code, highway route, latitude, longitude), and whether those involved in the accident had drank alcohol. There were many other variables used in the traffic dataset that were not fully explored regarding the circumstances of accidents.
The population file, People.csv, had demographics data for the United States.The file provided the location(FIPS code, state, county name) and demographics(population, population changes, education, ethnicities, home ownership, age, and gender) of the counties.
Specific variables we examined are described as follows:
Accidents.csv

"STATE", "COUNTY"

These variables represent the state and county where the accident occured. State and county were encoded as two and three digit codes respectively according to the GSA encoding.

"DAY", "MONTH", "DAY_WEEK", "HOUR", "MINUTE"

These variables represent the time when the accident occured.

"DRUNK_DR"

This variable represents the intoxicated status of the drivers when the accident occured, where a 0 shows no alcohol involved and a 1 shows police reported alcohol involvement or a positive alcohol test result.

People.csv

"FIPS"

This variable represents the identifier for the county as encoded in a 5 digit Fips code.

"TotalPopEst2015"

This variable represents the forecasted population for the designated county in the year 2015.

The accidents data was only by each incident rather than total number and populations were only provided by county so reorganization of the data was needed.
For map data, the accidents data was filtered to only include drunk incidents based on “DRUNK_DR” values. Then the GSA data from accidents was combined to FIPS codes. GSA state codes were prepended to the county GSA codes. County GSA codes less than 3 digits long were prepended with zeros. From here, the number of times a FIPS code, and a state GSA code came up in the accident dataset was tallied in test3 and test4 datasets.

The population dataset was integrated with the drunk accidents by dividing the number of drunk drivers from the FIPS code by the population of the FIPS code.

Persons.csv

"PER_TYPE"

This variable indicated the kind of person from the car involved that was described in the row of the dataset. A value of 1 indicated a driver.

"DRINKING"

This variable reflects whether the person matched to in this row had been determined to be consuming alcohol. Allowed values were: no, yes, not reported, unknown.

"RACE"

This variable reflects the race of the deceased, only in incidents where the person died.

"SEX"

Identifies the sex of the person involved in the crash. Possible values were male, female, not reported, and unknown.

"AGE"

Identifies the age of the person involved in the crash at the time of the accident in years. Possible values range from below 1 year, up to 120 years old, not known, not reported.

The persons data was only by each person involved in an accident rather than total number of people involved in accidents. Selection was done to insure that the accidents involved were fatal to a driver. These drivers were tallied according to each race, age blocks of 15 to 35, 36 to 65 and 66 to 120 (boundaries included) , and by sex. Then the total number was counted and the number of drunk drivers was counted.

For map data, the accidents data was filtered to only include drunk incidents based on "DRUNK_DR" values. Then the GSA data from accidents was combined to FIPS codes. GSA state codes were prepended to the county GSA codes. County GSA codes less than 3 digits long were prepended with zeros. From here, the number of times a FIPS code, and a state GSA code came up in the accident dataset was tallied in test3 and test4 datasets.

A shapefile, us.json was use for maps. It was provided in the 2/21 lecture.
We used the state property in the map to color each state according to its properties.

We had originally chosen to evaluate datasets by county and by day. However, due to the high degree of granularity in this data, we wound up moving our scale to state and month to make the data more readable and interpretable. The natural inclination of our group was to look at the dimensions of when and where these accidents happen. After observing a mapping of where accidents occured closely mirrored their populations, a normalization for population was added.

A description of the mapping from data to visual elements.
Describe the scales you used, such as position, color, or shape.

Time Plot

Scales xScaleBox and yScaleBox which map along the x and y axis respectively. xScaleBox scales the date (from January 1st to December 31st)to a pixel coordinate. Axis are drawn using lines across the range of both scales. Some benchmark values are provided based on the dataset.

The daily accident frequency is plotted as small slightly transparent circles. The max and min frequencies are plotted in a darker red circles. Text that state the date these values occurred on was also added to accompany these darker red circles. The remaining are in pale red circles.

For each month's average data, a box plot was placed using the x scale at the middle date of its month along with a text label at the top of the plot. These box plots were designed to show, in descending order: months and each box shows the maximum, 3rd quarter value, median, 1st quarter value, and minimum. The max was drawn as a small horizontal line with a vertical line reaching down to the 3rd quarter. The 3rd quarter, median, and 1st quarter were drawn as two pale blue rectangles placed on top of eachother with visible edges. The top edge was the third quarter, the median was the middle edges, the bottom edge was the 1st quarter. The min was drawn as a small horizontal line with a vertical line reaching up to the 1st quarter.

A legend is drawn for the figure to show what these figures represent. It shows one dot indicates the number of accidents in a day. It also shows that the box plot represents the maximum, 3rd quarter, median, 1st quarter, and minimum number of accidents daily for the month.

Number of accidents per person map

The number of accidents by state was translated into a map of the united states. A scale of color from white to blue, percentScale, was used to show the number of accidents per state divided by population in millions. The graph scale is somewhat unintuitive, with an overall domain from the min and the maximum. It has linear scaling between its min and its 1st quartile, its 1st quartile and median, its median and 3rd quarter, and its 3rd quarter and its maximum. This scaling was done because the values followed a somewhat normal distribution and when linear scaling was used, many states were indistinguishable from each other.

Relative occurrence of drunkenness in accidents fatal to the driver by race.

The number of accidents where a driver died from the person spreadsheet was totaled by race. We determined the subset of these accidents in which a driver was drunk and aggregated these by race as well. By dividing these drunken accidents by the total, we created a relative occurrence of drunkenness for these kind of accidents for each race. We plotted these values on a bar graph. The races were ordered by rank in this bar graph, with the highest values on top. Ages were listed in increasing order. Males and females were listed in that order. Each bar has the value at the end of its bar's length. Different shades of blue were used for race, sex, and gender to help differentiate these groups more intuitively. Bar length is determined by a

linear scale. This allowed us to plot the graph without using an axis for the frequency. Each demographic is labeled at the beginning of the bar for which it represents.

C).

Our project tells us about fatal drunk driving accidents. It invites a familiarity with how these types of accidents can be expected to occur, mainly in terms of by who, where, and when they occur.
The yearly view of fatal drunk driving crashes, labeled “When”, shows when the most fatal drunk driving accidents occur in the United States. Our graph shows that the months have much more internal variety than they do between themselves. However, it is also visible to the viewers that the accidents peak in summer months and are lower in winters.

The map view of fatal drunk driving crashes, labeled “Where”, shows where the most fatal drunk driving accidents occur in the United States. The map shows that the there are regional trends in drunk driving levels that are gradual and extend beyond state levels. Most of the states are similar to the color of their neighbors. This level of coloring can help to show cultural trends and blocks in the United States. The lowest rates of drunk driving tend to occur in the Southwest, whereas the highest tend to occur in the midwest. Notice that Illinois starkly contrasts its neighbors.

The bar graph view of drunk driving accidents fatal to the driver, labeled “Who”, shows where the relative occurrence of drunkenness in accidents fatal to the driver. This provides data on how many of drivers that died were drunk when they were in an accident for several different demographics. This bar graph shows that there are huge differences in which groups are likely to die behind the wheel while drunk. It shows the large differences by gender and age. We see that non-elderly men are behind many drunk driving incidents. We can also see that Koreans and American Indians are groups that have a high occurrence of drunkenness.