**Assignment 1 LaTeX Template**

**Note**: The images in this template are to indicate where to include your plots. You may find your plots should be larger in size than these template images in order for them to be easily read.

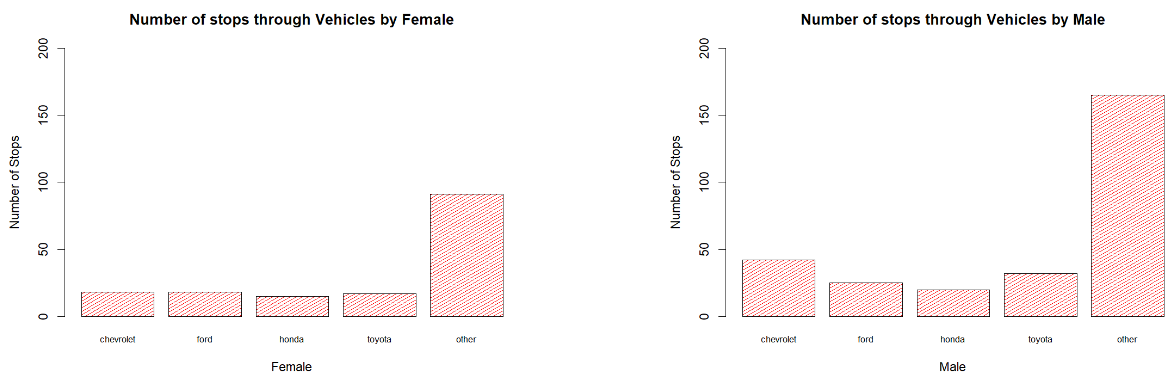
## Analysis 1

**1a**: My ID number is [20995558]. I will be analyzing the `vehicle.make` variate for Chicago.

**1b**: Note: Only one table is required, make sure you delete the other!

| Make | Female Frequency (%) | Male Frequency (%) |
|---:|---:|---:|
| Chevrolet | 18 (11.3%) | 42 (14.8%) |
| Ford | 18 (11.3%) | 25 (8.8%) |
| Honda | 15 (9.4%) | 20 (7.0%) |
| Toyota | 17 (10.7%) | 32 (11.3%) |
| Other | 91 (57.2%) | 165 (58.1%) |

**1c**: Barplots of my chosen variate:




(a) `subject.sex` = 'female' (b) `subject.sex` = 'male'

**1d**: The distributions of `vehicle.make` for subjects identified as female and male are somewhat similar. For subjects identified as female, we can see that the number of traffic stops for vehicles like Chevrolet, Ford, Honda, and Toyota is relatively consistent, whereas other vehicles exhibit a notably higher number of stops compared to these aforementioned brands. For subjects identified as male, it is evident that there is a greater number of traffic stops involving vehicles such as Chevrolet, Ford, Toyota, and other vehicles, as compared to the female demographic.

**1e**:

| | | `subject.sex` | |
|---|---:|---:|---:|
| | | Female | Male |
| City | Chicago | 159 | 284 |
| | San Francisco | 119 | 297 |

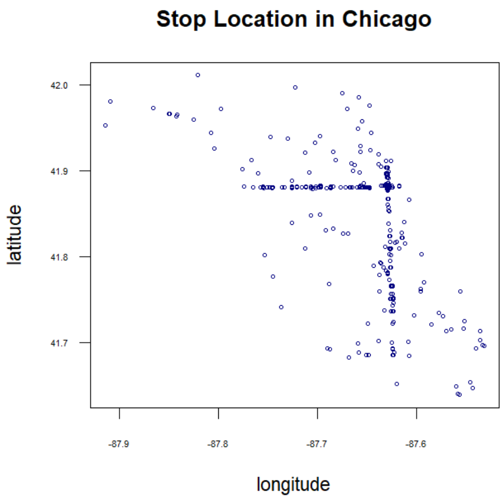
**1f**: The proportion of traffic stops in Chicago which were of subjects identified as female was 0.359. The proportion of traffic stops in San Francisco which were of subjects identified as female was 0.286.

**1g**: The relative risk is calculated by dividing the proportion in Chicago by that in San Francisco (0.359/0.286). This gives a relative risk of 1.255.

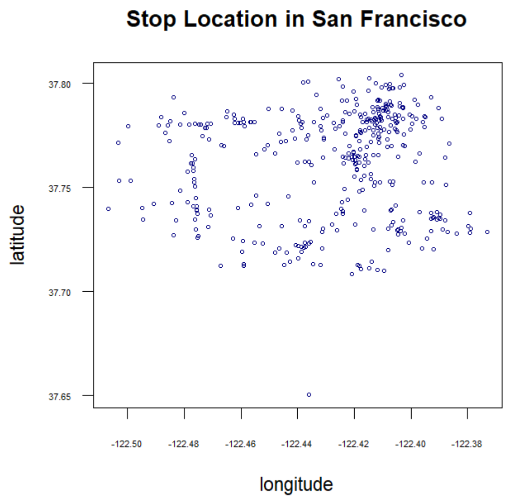
### Analysis 2

**2a**: My ID number is [20995558].

**2b**: Scatterplots of `lat` and `lng` for each city:




(a) Chicago




(b) San Francisco

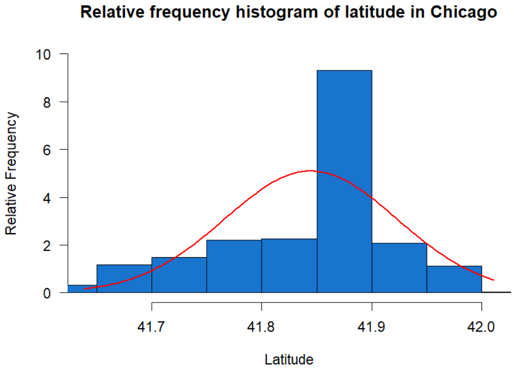
**2c**: The sample correlation between latitude and longitude for Chicago is -0.584. This suggests that there is a moderate negative linear relationship between latitude and longitude in Chicago, meaning that when the stop location moves northward (higher latitude) in Chicago, there is a moderate tendency for the corresponding longitude values to be farther west.

The sample correlation between latitude and longitude for San Francisco is 0.117. This suggests that there is a weak positive linear relationship between latitude and longitude in San Francisco. This means that when the stop location moves northward (higher latitude) in San Francisco, there is a slight tendency for the corresponding longitude values to be farther east.

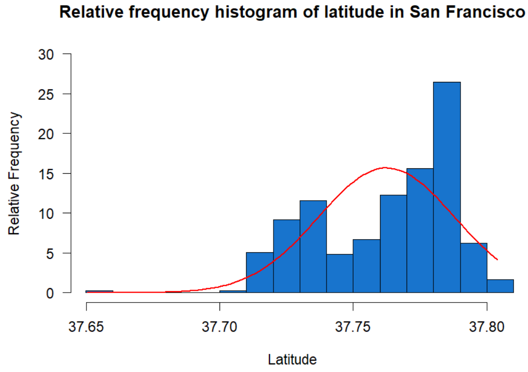
**2d**:

| Sample statistic | Chicago | San Francisco |
|---|---|---|
| Mean | 41.845 | 37.762 |
| Median | 41.880 | 37.770 |
| SD | 0.078 | 0.025 |
| Skewness | -0.728 | -0.644 |
| Kurtosis | 2.826 | 2.714 |

**2e**: Relative frequency histograms of `lat` for each city with superimposed probability density function curves:

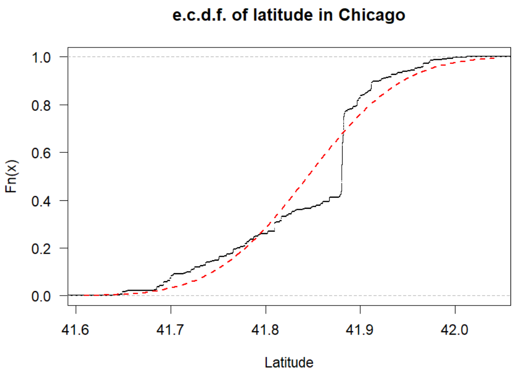



(a) Chicago

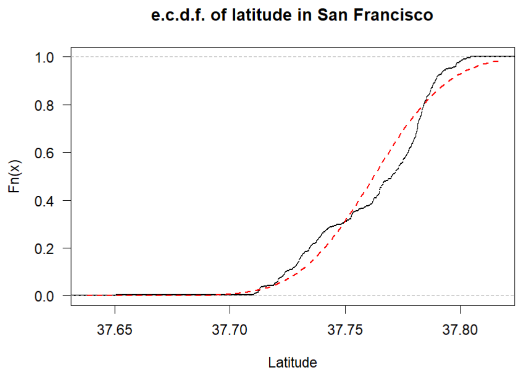



(b) San Francisco

**2f**: Empirical cumulative distribution function plots of `lat` for each city with superimposed cumulative distribution function curves:




(a) Chicago




(b) San Francisco

**2g**: **Chicago**: Based on the plot in Analysis 2e, we can see that there is a high volume of stops between latitude 41.85 and 41.9, while for data generated from a Gaussian distribution, we would expect to see a distinctive bell-shaped curve, with the highest relative frequencies centered at the mean, followed by a gradual decrease as one moves away from this central point in either direction. In the specific context of the provided diagram, several key statistical indicators align closely with Gaussian characteristics. Notably, both the mean and median are in close proximity, indicating a balanced distribution. The value of skewness, approaching 0, signifies a relative absence of skew, reinforcing the symmetry of the data. Additionally, the kurtosis value of 2.826, while slightly deviating from the ideal "bell-shaped" value of 3, still demonstrates a resemblance to Gaussian characteristics. Furthermore, the diagram effectively illustrates the presence of a central peak, which seems to be centered around the mean latitude of 41.845. This central peak offers compelling evidence of conformity to Gaussian characteristics. The smooth transition of the curve, both below and above the latitude of 41.854, supports the notion of a continuous and gradual decrease in relative frequencies. The relative frequency distribution, especially around the mean, showcases a remarkably smooth and consistent pattern, suggesting a certain degree of regularity within the data. Overall, the Gaussian model appears to fit reasonably well for the latitude data in Chicago. The presence of a central peak, the observed symmetry in the distribution, and the general bell-shaped pattern in the histogram, combined with the statistical indicators, collectively support the model's appropriateness for describing the data.

**San Francisco**: Based on the plot in Analysis 2e, we can see that there is a high volume of stops between latitude 37.75 and 37.8 while for data generated from a Gaussian distribution, we would expect to see a distinctive bell-shaped curve, with the highest relative frequencies centered at the mean, followed by a gradual decrease as one moves away from this central point in either direction. The proximity of the mean and median values suggests a well-balanced distribution of the data. Additionally, the skewness value, close to 0, indicates a relative absence of skewness, reinforcing the data's symmetry. The kurtosis value, approximating 3, aligns with the typical "bell-shaped" expectation, further supporting Gaussian-like characteristics. Meanwhile, the histogram reveals similar patterns in two latitude ranges: between 37.71 to 37.73 and 37.74 to 37.78. This repetition underscores the existence of distinct regions or zones within the latitude distribution where traffic stops tend to cluster, possibly due to specific geographic or environmental factors. Overall, the Gaussian model seems to fit reasonably well, as the presence of a central peak and certain symmetry align with Gaussian expectations.