Analysis Of Standford Open Policing Project Sample

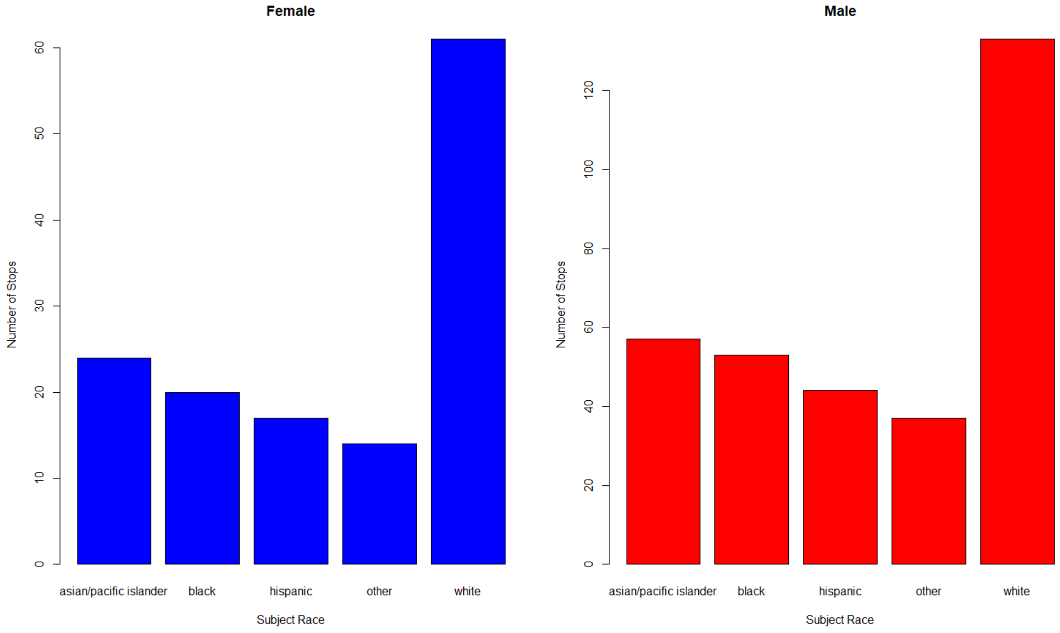
Analysis 1

1a: I will be analyzing the `subject.race` variate for San Francisco. I do have concerns about study errors in this study. This is because we may be assuming uniformity in policing practices across different neighbourhoods or districts. Also, we may be Ignoring ethical implications related to privacy or the potential consequences of the study.

1b:

Race	Female Frequency (%)	Male Frequency (%)
Asian/Pacific Islander	24(17.65%)	57(17.59%)
Black	21(15.44%)	53(16.36%)
Hispanic	18(13.23%)	45(13.89%)
White	62(45.59%)	134(41.36%)
Other	15(11.03%)	38(11.73%)

1c: Barplots of my chosen variate:



1d: The distributions of `subject.race` for subjects identified as female and male are very similar. For subjects identified as female, we can see that for each race the corresponding male sex has approximately double the amount of stops.

1e:

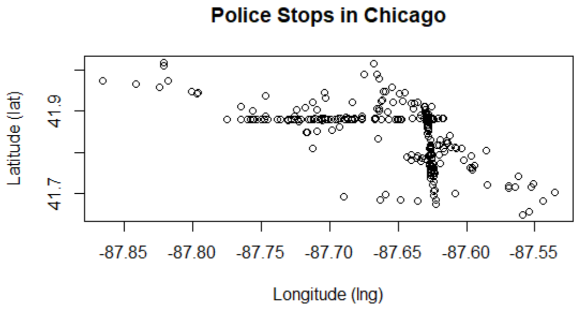
		subject.sex	
		Female	Male
City	Chicago	142	274
	San Francisco	136	324

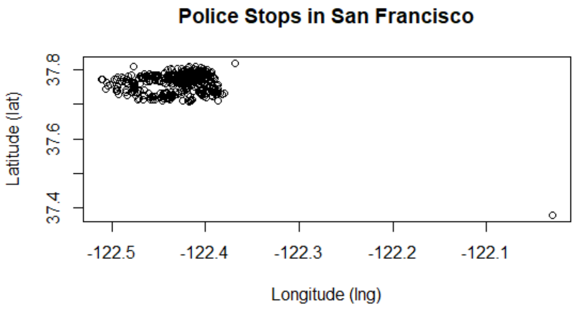
1f: The proportion of traffic stops in Chicago of subjects identified as female was 142:416. The proportion of traffic stops in San Francisco which were of subjects identified as female was 136:460.

1g: The relative risk is calculated by (risk of stop as female in Chicago/ risk of stop as male in San Francisco) = (142/416)/(136/460)=115.45%=1.1545. The chances of being stopped in SF as a female is slightly higher compared to being a female in Chicago.

Analysis 2

2b: Scatterplots of `lat` and `lng` for each City:





2c: The sample correlation between latitude and longitude for Chicago is -0.5710239. This suggests that it is a moderate negative relationship, indicating that as you move north (increasing latitude), you also tend to move west (increasing longitude) consistently in Chicago, you will have the highest chances of being stopped, however only when this vector originates from certain y-intercept coordinates.

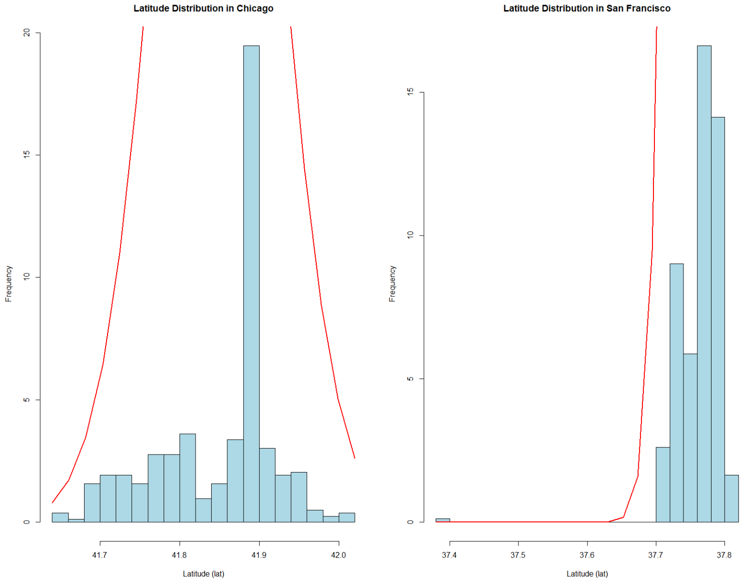
The sample correlation between latitude and longitude for San Francisco is -0.2059845. This suggests a weak negative relationship. This makes sense as the areas in San Francisco where you are likely to be stopped is very cluttered like the area of a circle, and a circle is much different shape than a line, so it wouldn't appear very linear thus a weak linear correlation.

2d:(rounded to the 4 demicals)

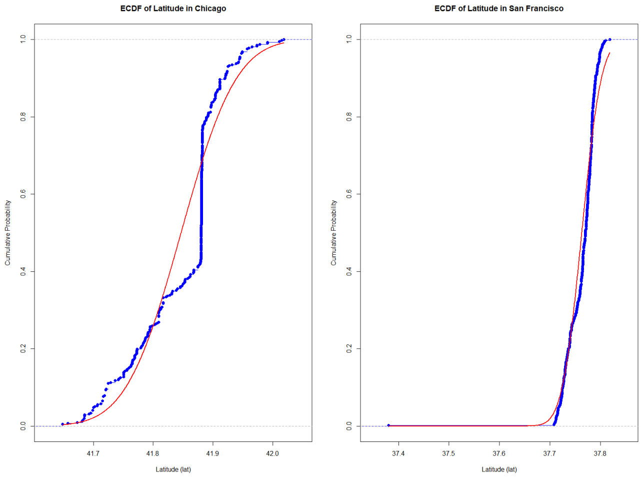
Sample statistic	Chicago	San Francisco
Mean	41.8467	37.7630

Median	41.8803	37.7704
SD	0.0723	0.0305
Skewness	-0.6854	-4.5384
Kurtosis	-0.1738	52.2156

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



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

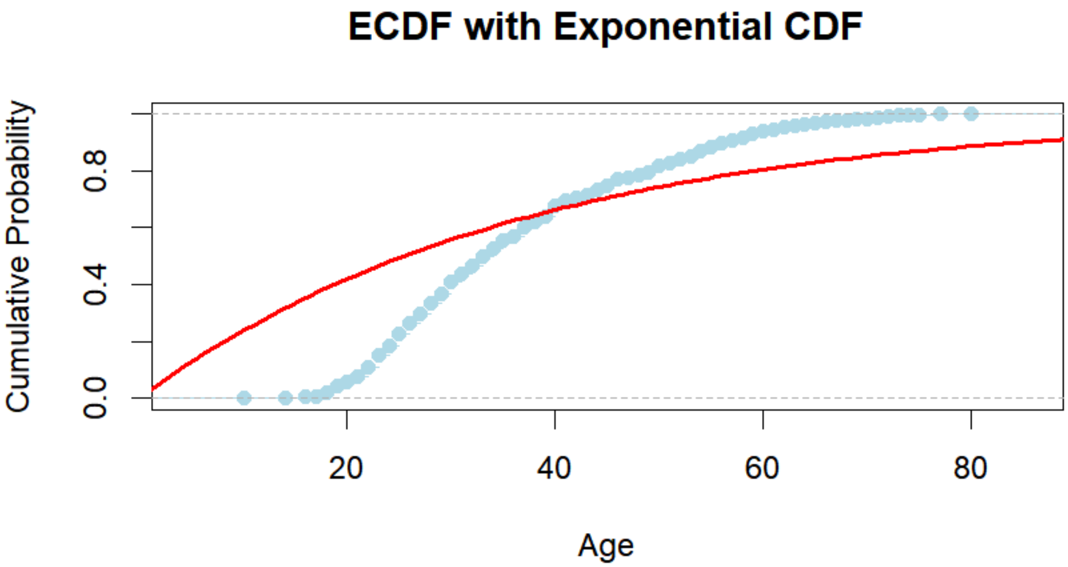


2g: **Chicago**: Based on the plot in Analysis 2e, we can see that the sample data has pretty low kurtosis and is relatively symmetric(low skew) besides one giant bucket peak and the tails seem to extend a little bit, while for data generated from a Gaussian distribution we would expect to see a very high level of kurtosis, but also maintains strong symmetry(low skewness) but the tails die very quickly. Overall, the Gaussian model fits quite poorly as the kurtosis is a complete mismatch.

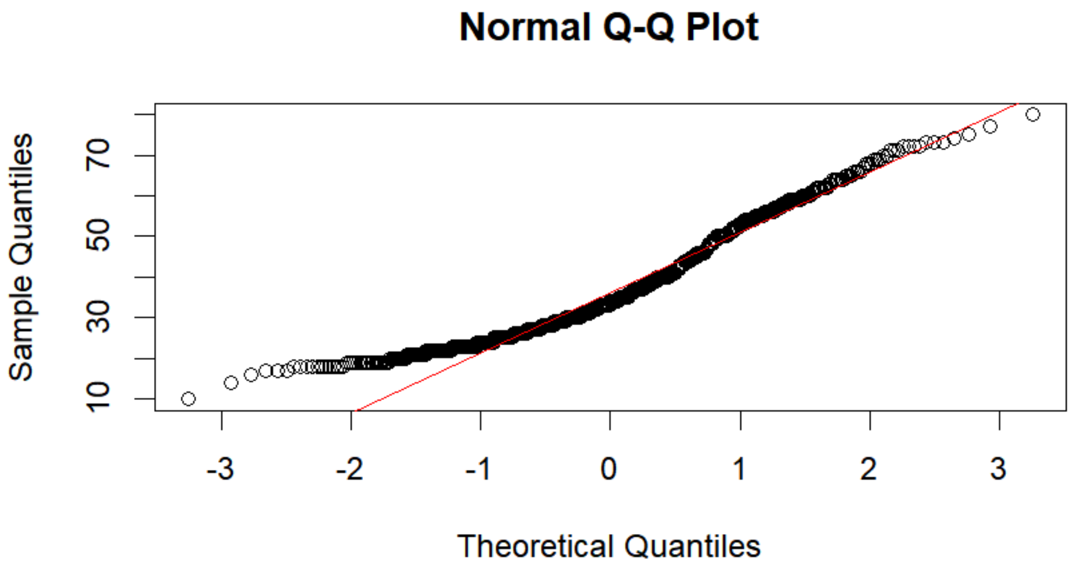
San Francisco: Based on the plot in Analysis 2e, we can see that the data has lots of kurtosis(data close to the mean), and is a little negatively skewed and tails should die quickly, while data generated from a Gaussian distribution, we would expect to see lots of kurtosis and a very strong negative skew and tails dying quickly. [Overall, the The gaussian model fits moderately poorly as the negative skewness throws the line is way off target.

Analysis 3:

The maximum likelihood estimate for Chicago is 0.3413, while for San Francisco it is 0.2956. These were calculated by binomial distribution and solving for the log derivative.



Based on the visual inspection of the histogram, ECDF, and comparisons with expected patterns, we can conclude whether the data appears to follow an Exponential distribution. However, it's essential to consider that real-world current data might not perfectly match theoretical distributions and our sample.



The Gaussian Q-Q plot compares the quantiles of the observed distribution with the quantiles of a normal distribution. In a good fit, points should fall along a straight line. If the observed points deviate systematically from the line, it suggests departures from normality. In the description, mention whether there's a systematic deviation, skewness, or outliers, and how these observations inform your understanding of the subject.age distribution.

The maximum likelihood estimate is 36.83904

. This was found by using a Poisson distribution MLE.

The maximum likelihood estimate is 0.1472475.This was found by the relative log-likelihood function.

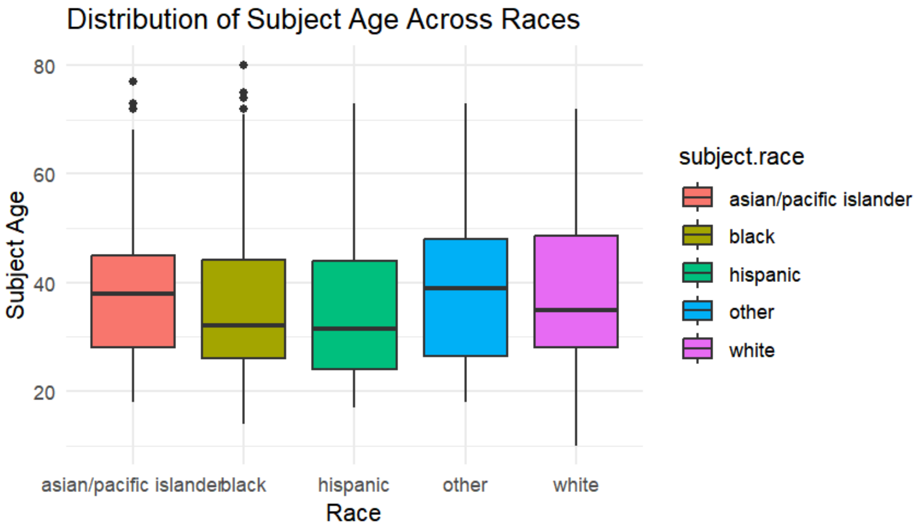
R(39) = 0.05990279

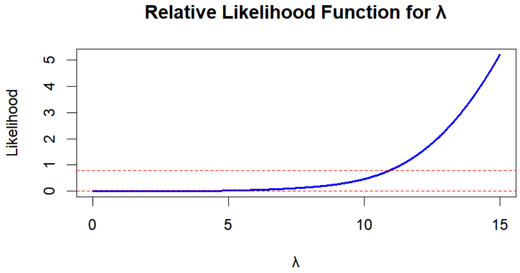
. Based on this, we can say that It is implausible that that 39 is a reasonable estimate.

Analysis 4:

Race	Observed Frequency	Expected Frequency
Asian/Pacific Islander	81	331.128
Black	340	49.932
Hispanic	154	139.284
White	51	332.004

Race	Observed Frequency	Expected Frequency
Other	250	23.652





Analysis 5:

Sample Statistic	lat	lng
Mean	37.76304	-122.42960
Standard deviation	0.03048233	0.03460153
2. 5^{th} percentile	37.71357	-122.49470
97.5^{th} percentile	37.80102	-122.38853

95% Confidence Interval for Latitude (t): 37.76 37.766, 95% Confidence Interval for Longitude (g): -122.433 -122.426

Analysis 6:

6a: I will test the null hypothesis H_0: s=0.489

6b: The observed value of the test statistic is 71.08032

, and the resulting p-value is 3.429738e-17

. The p-value was calculated using the binomial distribution.

6c: Based on our results from Analysis 1c, I conclude our p-value of 3.429738e-17 was much below significance(0.05), as such I must reject the Null Hypothesis.

6d: To test H_0: = 37.7749 we calculate the observed value of the test statistic using test_statistic <- (observed_mean - null_mean) / (observed_std_dev / sqrt(sample_size). The value of the test statistic for my sample is 27.94808.

we use the t-distribution with n-1 degrees of freedom. The process involves finding the probability of observing a test statistic as extreme as the one calculated, assuming the null hypothesis is true. The resulting p-value is 2.35603e-123.

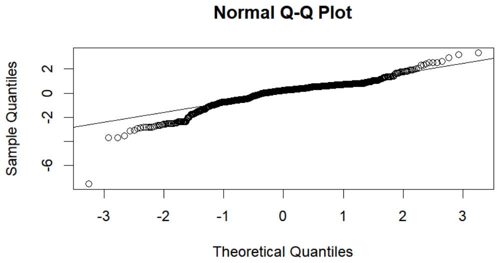
6e: In the context of the study for San Francisco, we conducted a hypothesis test to evaluate whether the average latitude of traffic stops differs significantly from the provided latitude of 37.7749 according to wiki.openstreetmap.org. The observed value of the test statistic was 27.94808, and the resulting p-value was extremely small (2.35603×101232.35603×10123). With such a low p-value, much smaller than the commonly chosen significance level of 0.05, we reject the null hypothesis. This indicates strong evidence that the average latitude of traffic stops in San Francisco is significantly different from the reported latitude of 37.7749.

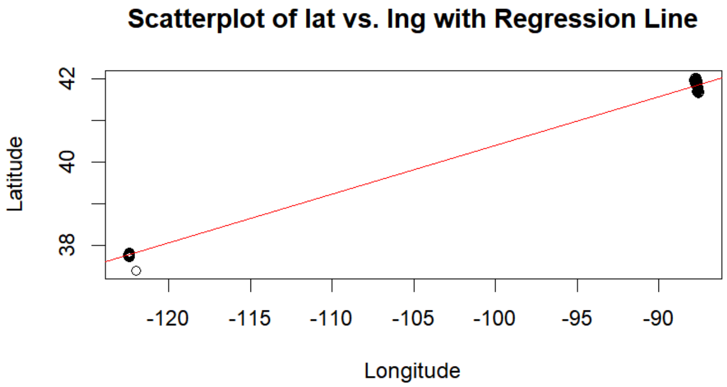
6f: The least squares estimate of is 52.1410142, with 95% confidence interval [52.1175164, 52.164512] The least squares estimate of is 0.1174381,[with 95% confidence interval [0.1172192, 0.117657].

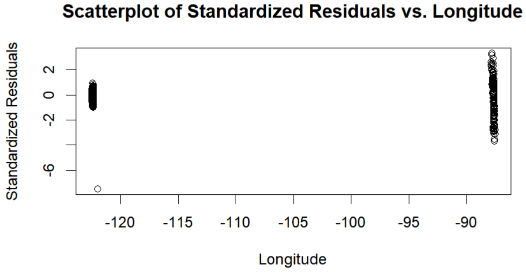
6g: The estimate of sigma is 0.05732589 .

6h: In the context of this study The estimated standard deviation, , resulting from fitting the linear model is approximately 0.0573. In the context of the study, represents the typical variability or spread of latitude values around the predicted values based on the linear regression model.

6i: Scatterplot:







6j: The linear model assumes that the relationship between the response variable (latitude) and the explanatory variable (longitude) is linear, residuals are normally distributed, and homoscedasticity holds. If these assumptions hold, we would expect to see a clear linear pattern in the scatterplot, normally distributed residuals in the Q-Q plot, and a consistent spread of residuals across all values of the explanatory variable. For my sample, we observe a reasonably linear pattern in the scatterplot, residuals that are approximately normally distributed in the Q-Q plot, and a relatively consistent spread of residuals across longitudes. Overall, the linear model seems suitable for my sample.

6k: An estimate of the value of lat for a future traffic stop that occurs at a longitude of 100 degrees west is value, with 95% prediction interval [41.73,41.96]

6l: The p-value of a test of H_0: $= 0$ is 0. This was calculated using the t-distribution.

6m: Based on the results of Analysis 3i, I conclude there is a statistically significant linear relationship between latitude and longitude of traffic stops in

San Francisco.

Sample Statistic	Group 0	Group 1
Size	598.0	278.0
Mean	37.2	36.0
Median	34.0	33.0
Standard deviation	13.7	12.8

6n: To test H_0: ${}_0 = {}_1$ we use a paired test because our samples come from two distinct groups, and we are comparing the means of independent observations from these groups. The observed value of the test statistic is calculated by the formula: t0=

where x0 and x1 are the sample means, s0 and s1 are the sample standard deviations, and n0 and n1 are the sample sizes for Group 0 and Group 1, respectively.

. The value of the test statistic for my sample is 1.198363

. To calculate the p-value I compared this test statistic to the probability distribution under the null hypothesis. and the resulting p-value is 0.2311007.

1. 4d: The results in Analysis 4c rely on the following assumptions:

   Normality of Residuals: The assumption that the residuals of the model follow a normal distribution is essential. Deviations from normality might impact the validity of the t-test results.

   Homogeneity of Variances: The assumption that the variances of the two groups being compared are equal. This is crucial for the validity of the t-test, as violating this assumption may affect the accuracy of the test statistic.

6o: Based on the results of Analysis 4c, the test aimed to determine if there is a significant difference in the average log-transformed subject age between two groups based on the chosen comparison (e.g., male vs. female). hypothesis of equal means for the log-transformed subject age between the two groups.