Assignment 2 LaTeXTemplate

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

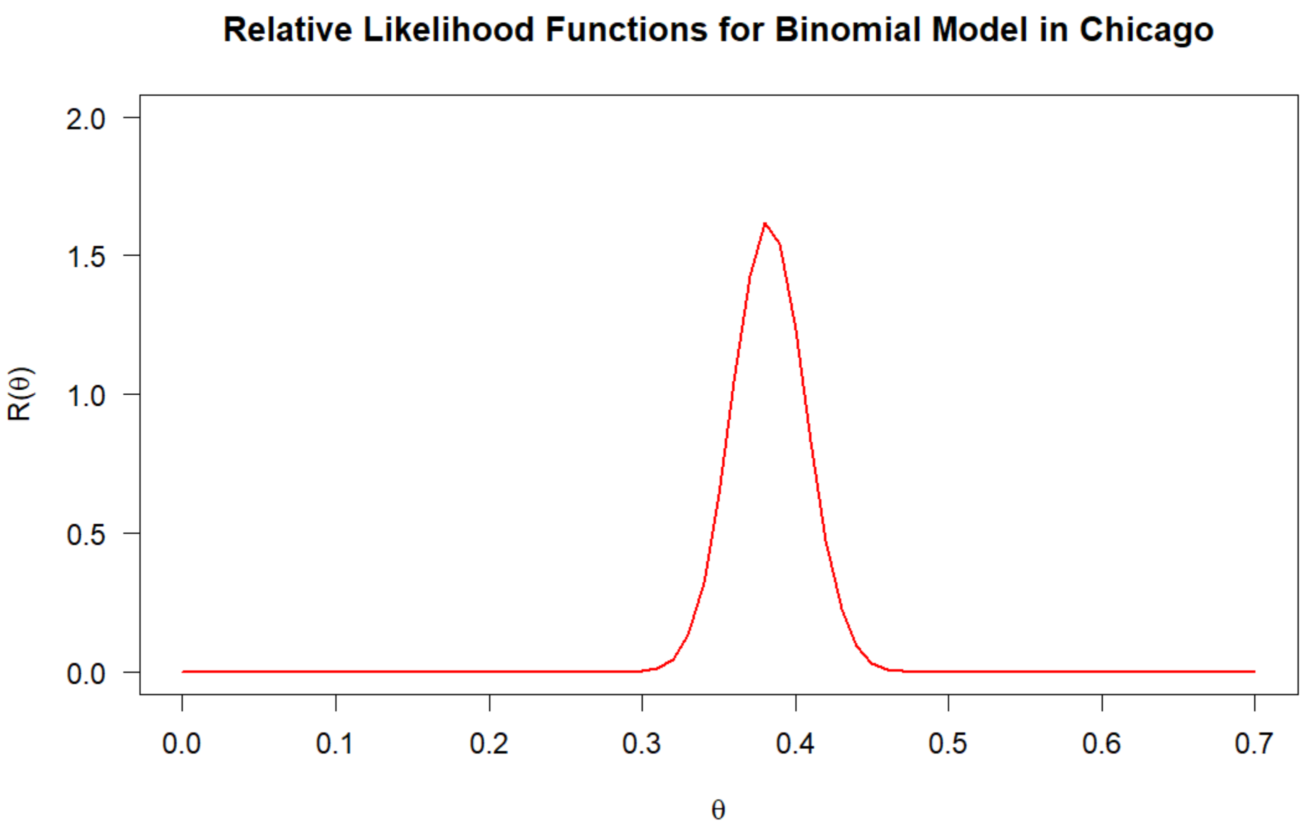
1a: My ID number is 20995558.

1b: I do have concerns about study error in this study. This is because despite the extensive and comprehensive data collection by the Stanford Open Policing Project, potential biases or discrepancies in how different police departments record or report data might exist, and there's no mention of external verification of the data. Additionally, the vast scope of the project, facing numerous law enforcement agencies, might introduce variations in data accuracy and consistency.

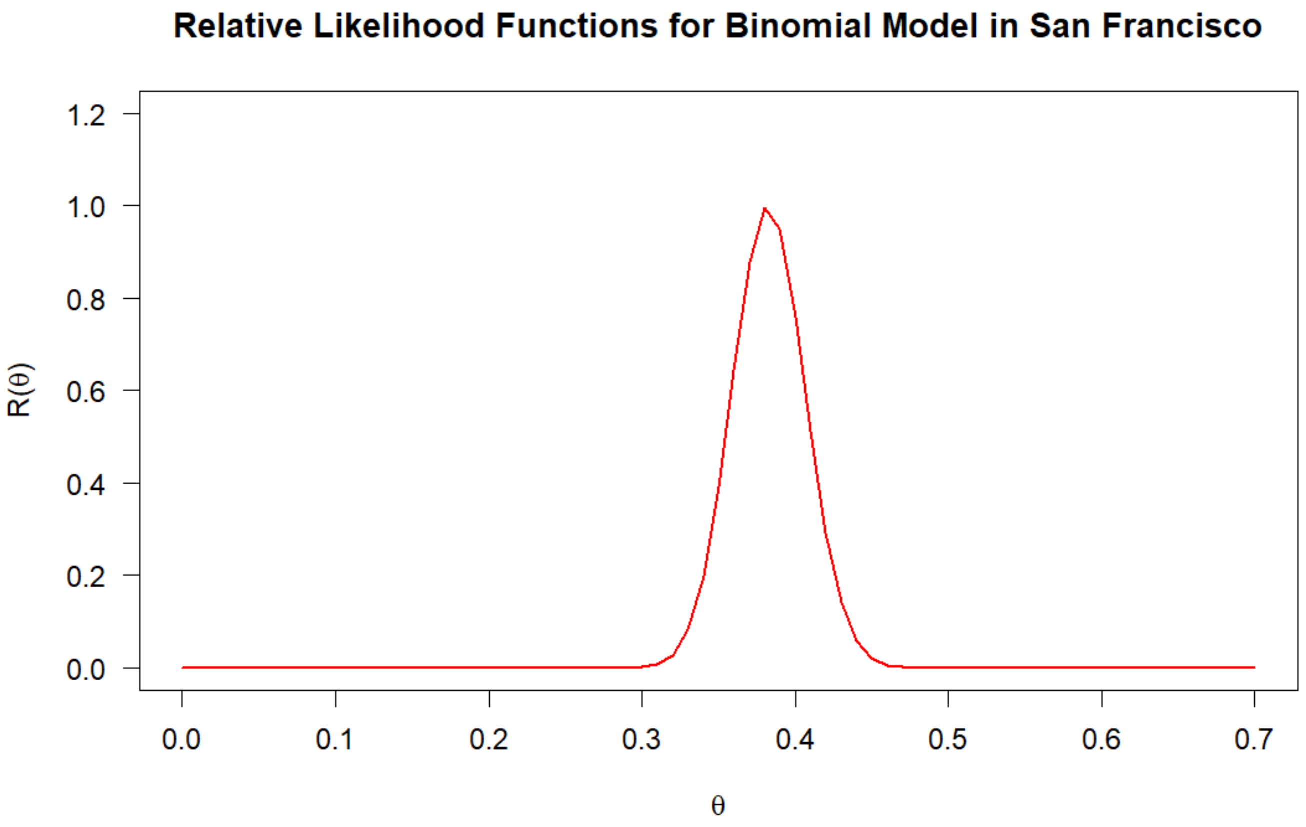
1c: The maximum likelihood estimate for Chicago is 0.3589165, while for San Francisco it is 0.3822115. These were calculated by dividing the number of female stops by the total traffic stops in each city after filtering the dataset for the respective city.

```
chicago <- subset(mydata, subset = (city == "chicago"))
female_data <- subset(chicago, subject.sex == "female")
Fc <- nrow(female_data)
n <- nrow(chicago)
thetahat_c <- Fc / n
thetahat_c
```

1d: Relative likelihood function plots:



(a) Chicago



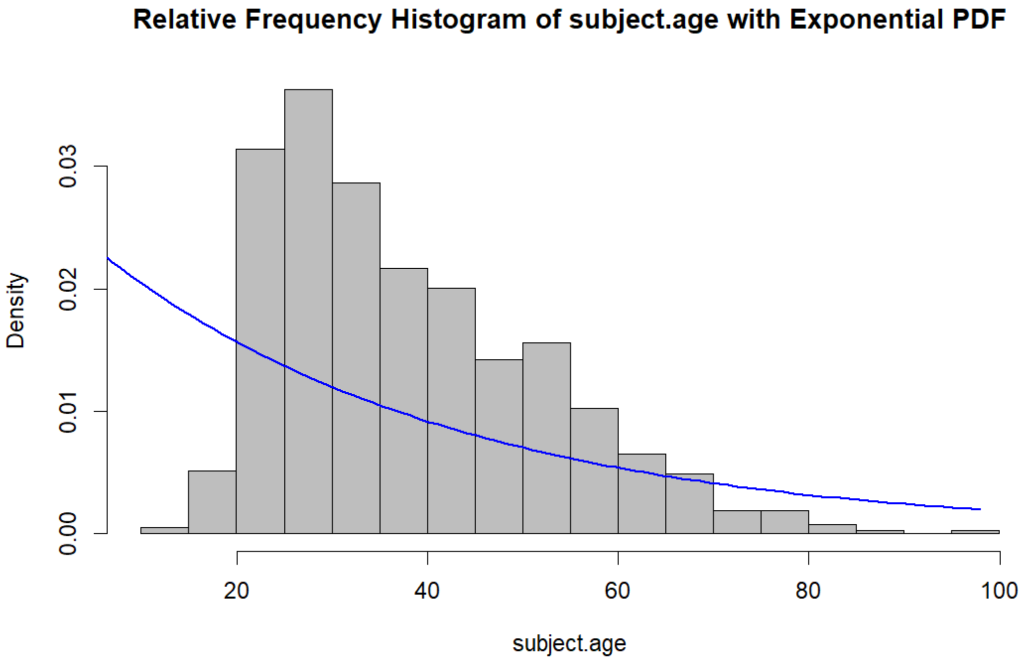
(b) San Francisco

Analysis 2

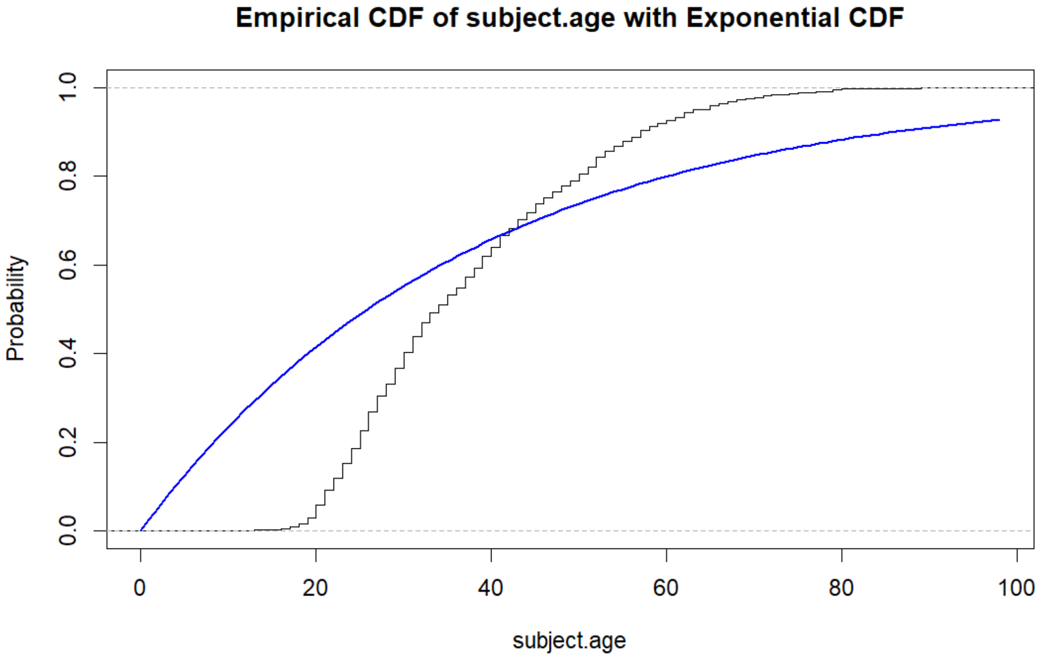
2a: My ID number is 20995558.

2b: The sample size, mean, median, and standard deviation are, respectively, 859, 37.380, 34, 14.074.

2c: Relative frequency histogram:

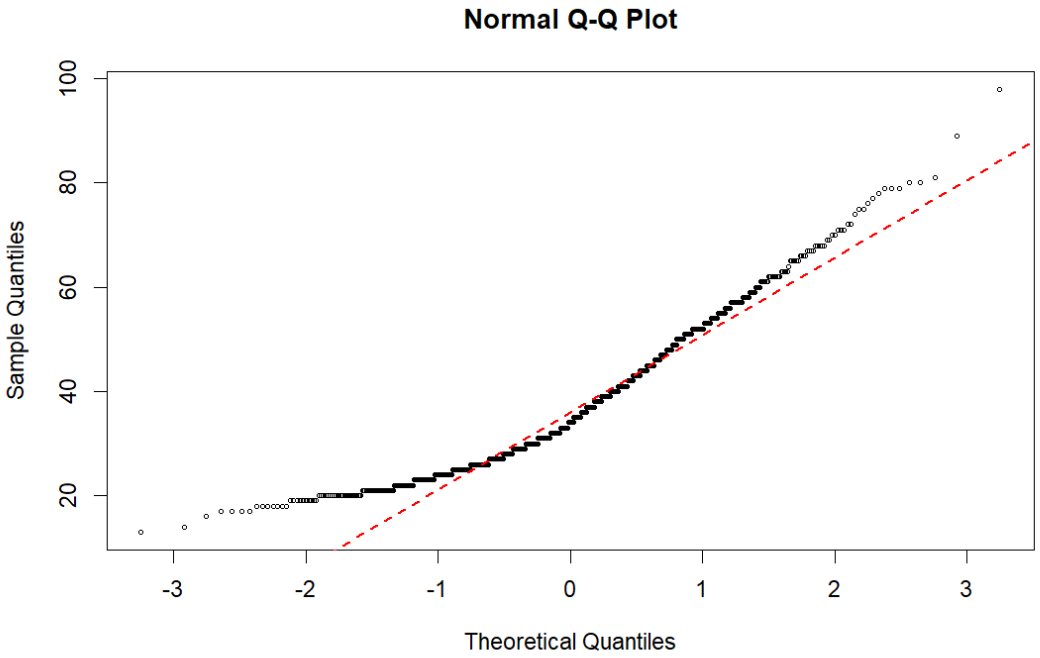


2d: Empirical cumulative distribution function plot:



2e: Based on the plot in Analysis 2d, we can see the empirical CDF for the sample data of the subject.age increases gradually at the beginning, significantly sharp in the mid-range, and then levels off towards the tail end, while for data generated from an Exponential distribution we anticipate seeing a steady, continuously increasing curve from the beginning to the end. Specifically, the empirical CDF starts increasing at a slower pace for ages 0 to 20, contrasts sharply about the ages 20 and 45, and then aligns more closely, though not perfectly, with the moderate rise of the Exponential distribution for the elder age groups 60 to 100. Overall, while there are some similarities, the Exponential model might not fully fit the age distribution reflected in the dataset aged from 20 to 45.

2f: Q-Q plot:



2g: Based on the Q-Q plot, the plot shows a moderate linear trend along the theoretical quantiles, especially in the central portion, which suggests that the middle values of subject.age aligns well with a normal distribution. However, in the lower tail and upper tail, the black data points are above the red dashed line, which means the data's lower tail is lighter than what would be expected under a standard normal distribution. Overall, although the middle portion of the subject.age distribution seems to align with a normal distribution, there are noticeable differences from ideal normality throughout the distribution.

Analysis 3

3a: My ID number is 20995558.

3b: I do have concerns about sample error in this study. This is because While the Stanford Open Policing Project conducts an expansive and thorough data collection, there may be inherent biases or inconsistencies in how different police departments record or report their data. The absence of an external verification process further underscores potential data reliability issues. The wide range of law enforcement agencies involved could also lead to variations in the accuracy and consistency of the data.

3c: The maximum likelihood estimate is 37.380. This was found by the maximum likelihood estimate (MLE) of lambda is simply the sample mean of the data.

```
mle <- mean(mydata$subject.age)
mle
```

3d: The maximum likelihood estimate of the probability a randomly chosen traffic stop subject is 30 years old or younger is 0.128. This was found by computing the cumulative distribution function (CDF) of a Poisson distribution with an lambda value of 37.380 up to 30 using the ppois R function.

```
mle <- mean(mydata$subject.age)
probability_30_or_younger <- ppois(30, lambda = mle)
probability_30_or_younger
```

3e: $R(39) = 1.83131\text{e-}13$. Based on this, we can say that the value is very close to zero and indicates that a lambda value of 39 is highly implausible based on the sample data.

```
PoisRLF <- function(theta, n, thetahat) {
  exp(n * thetahat * log(theta/thetahat) + n * (thetahat -
                                                  theta))
}
n <- length(mydata$subject.age)
thetahat <- mean(mydata$subject.age)
R_39_value <- PoisRLF(39, n, thetahat)
R_39_value
```

Analysis 4

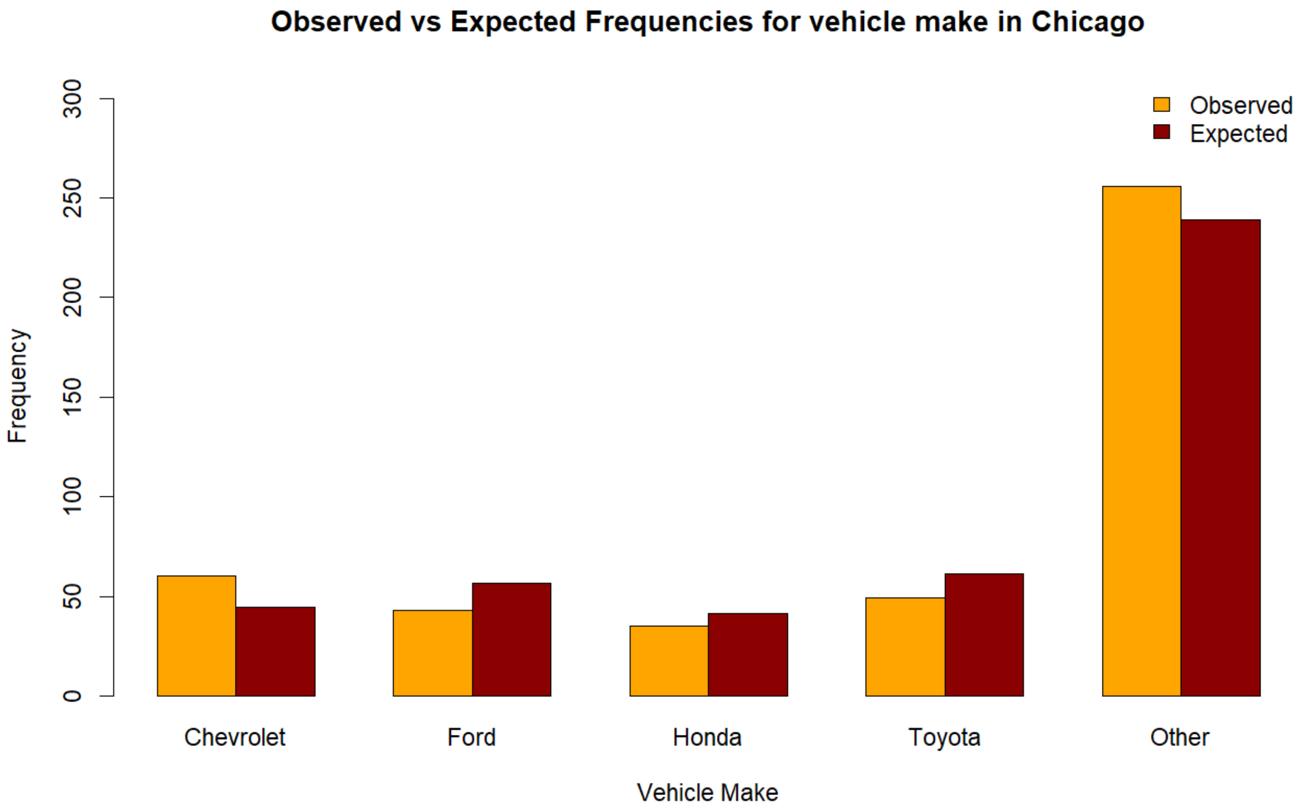
4a: My ID number is 20995558. I will be analyzing the `vehicle.make` variate for Chicago.

4b: I do have concerns about measurement error in the `vehicle.make` variate. This is because the table has a significant portion of the data that falls under a very broad and unspecific "Other" category, about 256 out of 443. This substantial count hints at the possibility of multiple vehicle makes being inaccurately categorized or potential errors in categorization, which could cause inaccuracies regarding specific vehicle makes in the dataset.

4c: Table of observed and expected frequencies (reminder: your Report should only contain one such table for your chosen variate):

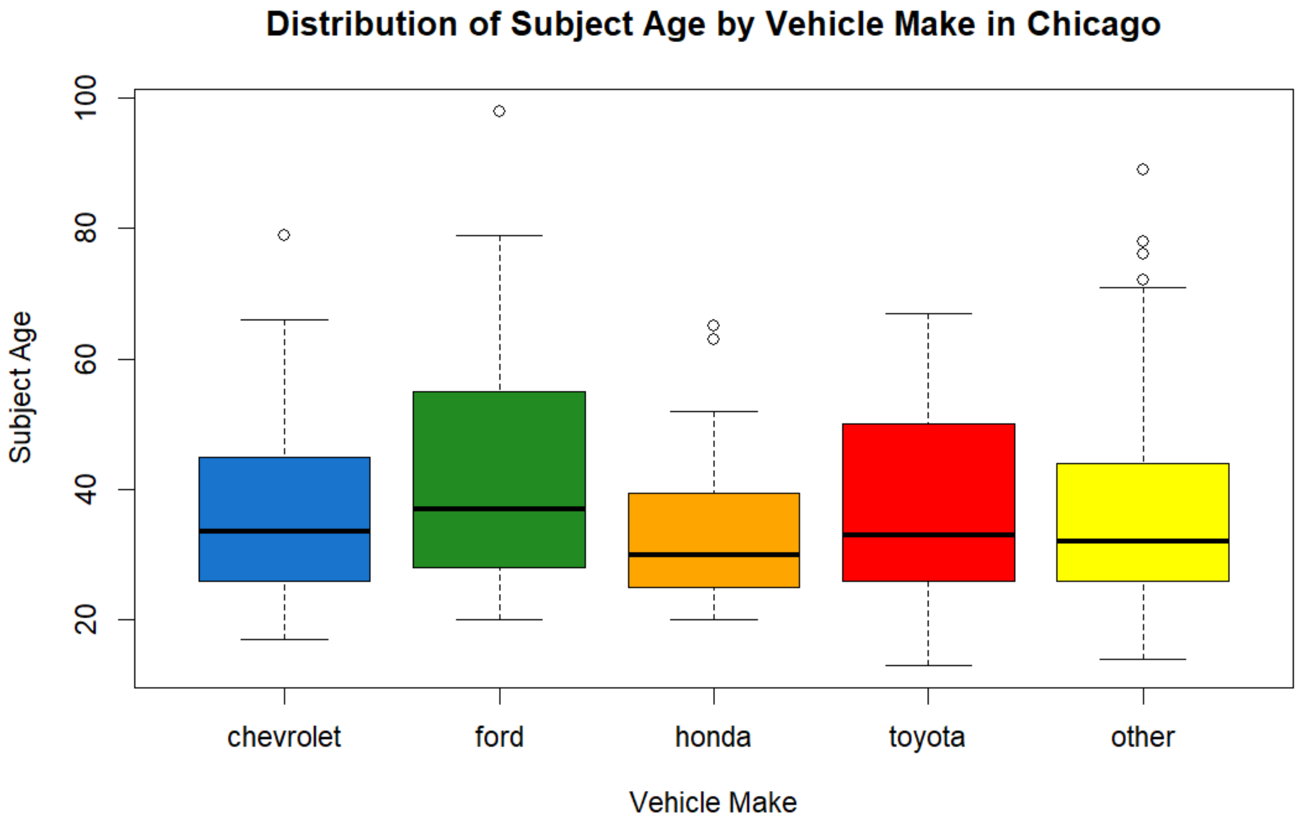
Make	Observed Frequency	Expected Frequency
Chevrolet	60	44.743
Ford	43	56.704
Honda	35	41.199
Toyota	49	61.134
Other	256	239.220

4d: Grouped barplot of observed and expected frequencies:



4e: Based on Analyses 4c and 4d, it's evident that certain vehicle makes, namely Ford, Honda, and Toyota, have observed frequencies that are marginally lower than the expected frequencies. In contrast, Chevrolet and the aggregated "Other" category demonstrate observed frequencies surpassing their respective expected values. Overall, while there are discrepancies between the observed and expected frequencies for some vehicle makes, the observed data largely align with the expected frequencies.

4f: Boxplot of `subject.age`:



4g: Based on the results of Analysis 4f, we observe that `subject.age` does appear to be similar across the categories of `vehicle.make`. In particular, the median ages for Chevrolet, Honda, Toyota and 'other' category seem to be clustered around the 30 to 35 age range. Additionally, the 'Ford' category has a slightly higher median age compared to the rest. The presence of upper outliers, especially pronounced in the Ford, Chevrolet, Honda and 'other' categories, indicates that there are individuals who are significantly older than the typical age group for those vehicle makes.