Stat 231 Assignment 1
MingMing Z.

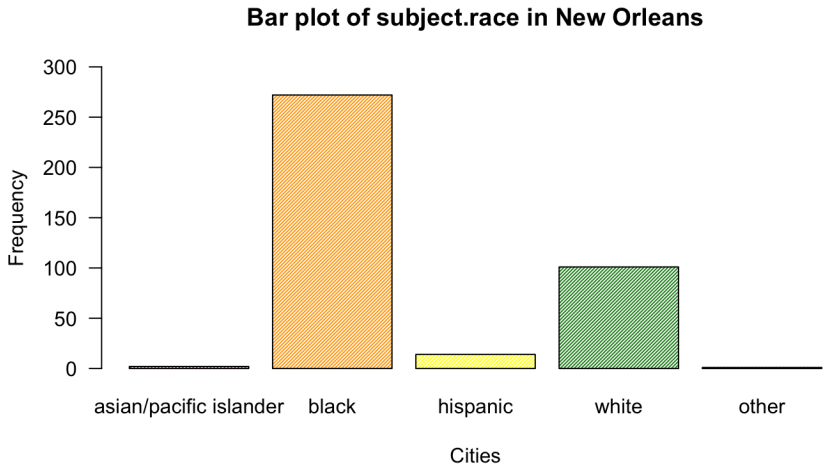
Analysis 1

1a: My ID number is 21058539. I will be analysing the `subject.race` variate.

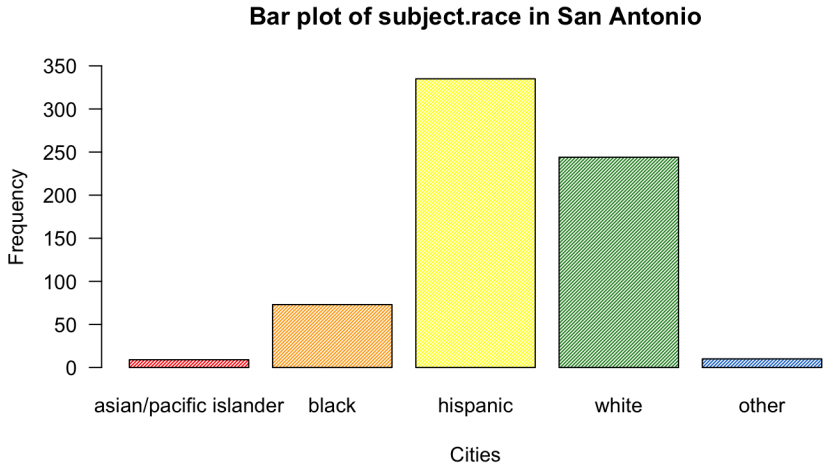
1b:

Race	New Orleans Frequency (%)	San Antonio Frequency (%)
Asian/Pacific Islander	2(0.5%)	9(1.3%)
Black	272(69.7%)	73(10.9%)
Hispanic	14(3.6%)	335(49.9%)
White	101(25.9%)	244(36.4%)
Other	1(0.3%)	10(1.5%)

1c: Barplots of my chosen variate:



(a) New Orleans



(b) San Antonio

1d: The distributions of `subject.race` for subjects stopped in San Antonio compared to New Orleans are not very similar. For subjects stopped in San Antonio, we can see that Hispanics (around 50%) are the majority, while for subjects stopped in New Orleans, we can see that Black is the majority, which accounts for about 70% while Hispanics only account for 3.6%.

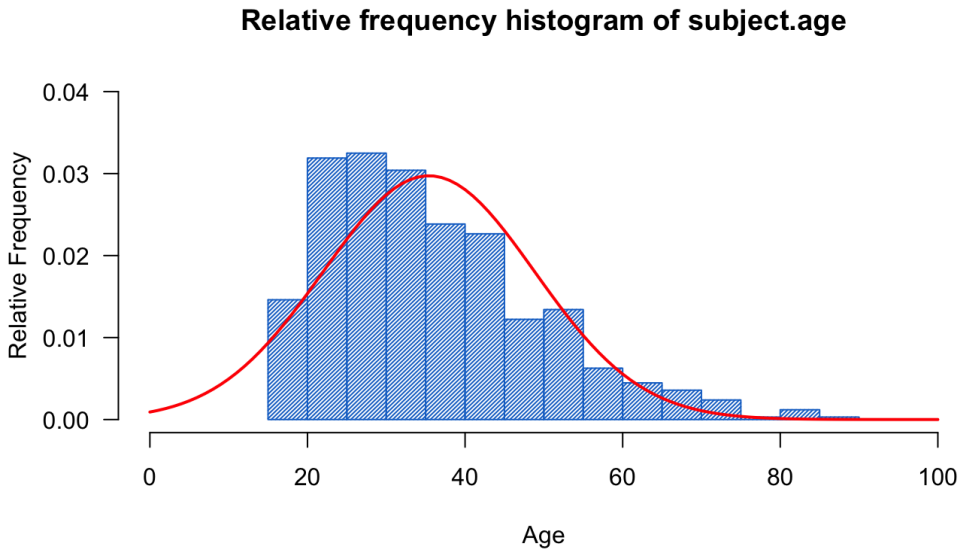
Analysis 2

2a: My ID number is 21058539. I will be analysing the data for San Antonio.

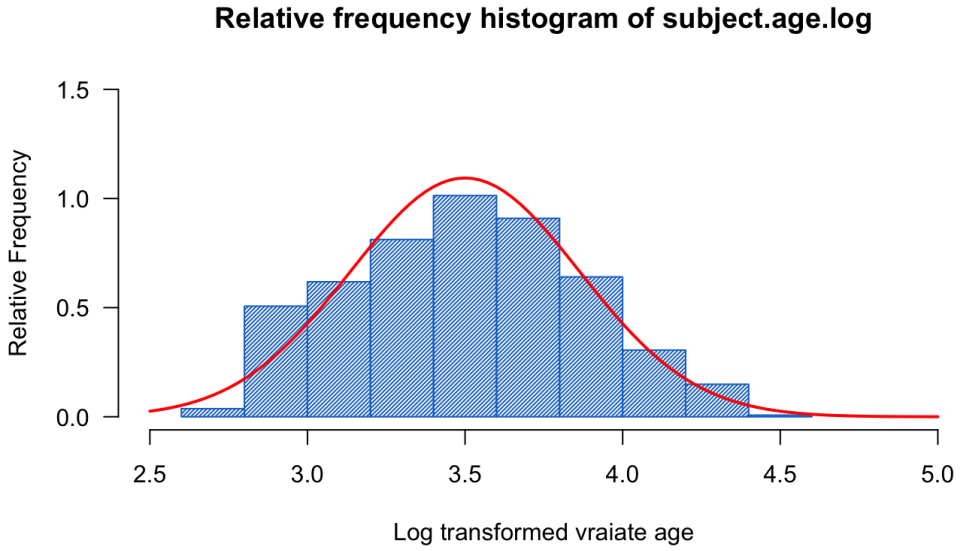
2b:

Sample statistic	`subject.age`	`subject.age.log`
Mean	35.411	3.500
Median	33.000	3.497
SD	13.418	0.365
Skewness	0.943	0.175
Kurtosis	3.648	2.318

2c: Relative frequency histograms of `subject.age` and `subject.age.log` with superimposed probability density function curves:

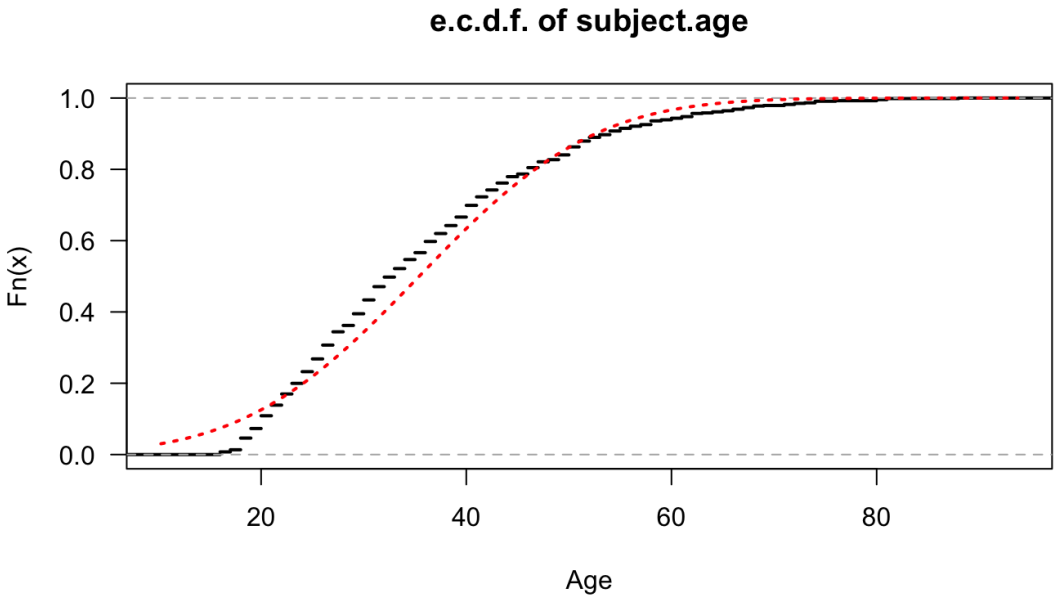


(a) `subject.age`



(b) `subject.age.log`

2d: Empirical cumulative distribution function plots of `subject.age` and `subject.age.log` with superimposed cumulative distribution function curves:



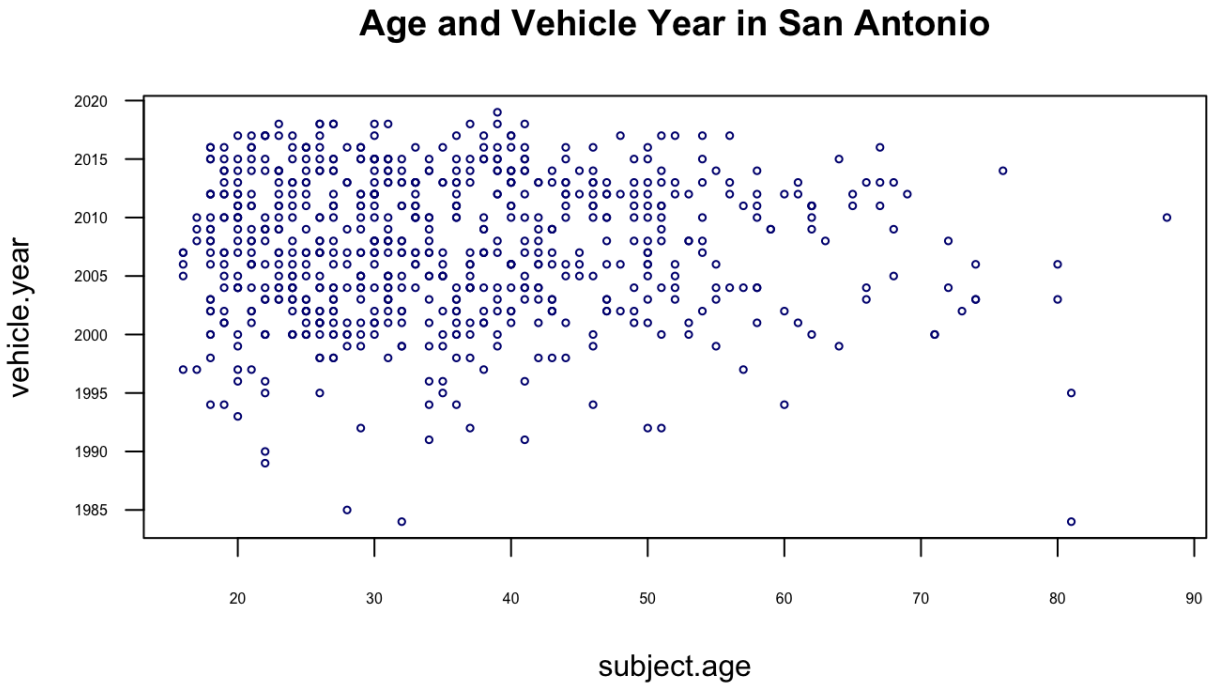
(a) `subject.age`

(b) `subject.age.log`

2e: `subject.age`: Based on the plot in Analysis 2c, we can see that there is a right skew while for data generated from a Gaussian distribution we would expect to see the density function curve to be more centered. To be more specific, the skewness of 0.94 is only slightly close to 0. Additionally, the sample kurtosis is 3.65, which is relatively close to 3 but still implies more weight in the center. Overall, even though there's a bell shape to the superimposed PDF curve, the Gaussian model would not fit as well as `subject.age.log`, though I can see the argument of Gaussian model being acceptable as the sample mean and median is relatively close, and 95.4% of points fall within two sample standard deviations of the sample mean.

`subject.age.log`: Based on the plot in Analysis 2c, we can see that the superimposed PDF curve follows a bell shape with sample data skewness of 0.18, which is reasonably close to 0. For data generated from a Gaussian distribution we would expect to see something similar. Comparing the plots in 2d, we can also see the superimposed curve fits the ECDF of `subject.age.log` much more closely than `subject.age`. The mean and median of the sample data is also extremely close. The kurtosis being 2.32 is relatively close to 3 but there is still more weight to the tails of the curve compared to a standard Gaussian model. Overall, the Gaussian model would fit reasonably well.

2f: Scatterplot of `subject.age` and `vehicle.year` for your chosen city:



2g: The sample correlation between subject age and vehicle year is -0.000567 (3.s.f). This extremely small negative value that is close to 0 suggests no correlation between the pairs of variates.

Analysis 3

3a. My student number is 21058539 and I like the flapjack (also called a pancake) octopus!

3b. Around 17%

3c. 2019 to 1984 (Range = 35)

3d. For San Antonio, modal colour is `other`. For New Orleans, it's `other` as well. If we don't count `other`, it would be `white` for both.

3e. The median vehicle colour is `other`. Although, it would depend on how the categorical colour data is ordered.

Appendix: R Code

```
# setting up
mydata <- read.csv("~/Desktop/School/Stat 231/stat231f24dataset21058539.csv")
dim(mydata)
colnames(mydata)
sum(is.na(mydata))

# beginning of analysis
# 1b.

" testing out vehicle.make:
vehtable <- table(mydata$vehicle.make, mydata$city)
vehtable
prop.table(vehtable, 2) "

mydata$subject.race <- factor(mydata$subject.race,
    levels = c( "asian/pacific islander", "black", "hispanic", "white", "other"))
racetable <- table(mydata$subject.race, mydata$city)
racetable
prop.table(racetable, 2)

# 1c.
sanan <- subset(mydata, city=="sa")
newor <- subset(mydata, city== "no")
barplot(table(sanan$subject.race), xlab = "Cities", ylab = "Frequency", las = 1,
    main = "Bar plot of subject.race in San Antonio",
    col = c("red", "orange", "yellow", "forestgreen",
    "dodgerblue3"), ylim = c(0, 350), density = 50)
barplot(table(newor$subject.race), xlab = "Cities", ylab = "Frequency", las = 1,
    main = "Bar plot of subject.race in New Orleans",
    col = c("red", "orange", "yellow", "forestgreen",
    "dodgerblue3"), ylim = c(0, 300), density = 50)

# 2b.
mydata$subject.age.log <- log(mydata$subject.age)
SanAntonio <- subset(mydata, city=="sa")
skewness <- function(x) {
  (sum((x - mean(x))^3)/length(x))/(sum((x - mean(x))^2)/length(x))^(3/2)
  }
kurtosis <- function(x) {
  (sum((x - mean(x))^4)/length(x))/(sum((x - mean(x))^2)/length(x))^2
}

mean(SanAntonio$subject.age)
median(SanAntonio$subject.age)
sd(SanAntonio$subject.age)
skewness(SanAntonio$subject.age)
kurtosis(SanAntonio$subject.age)
```

```
mean(SanAntonio$subject.age.log)
median(SanAntonio$subject.age.log)
sd(SanAntonio$subject.age.log)
skewness(SanAntonio$subject.age.log)
kurtosis(SanAntonio$subject.age.log)

# 2c.
library(MASS)
truehist(SanAntonio$subject.age, xlab = "Age",
         ylab = "Relative Frequency", main =
         "Relative frequency histogram of subject.age",
         xlim = c(0, 100), ylim = c(0, 0.04), las = 1,
         col = "dodgerblue3", density = 50)
curve(dnorm(x, mean(SanAntonio$subject.age), sd(SanAntonio$subject.age)),
      col = "red", add = TRUE, lwd = 2)
truehist(SanAntonio$subject.age.log, xlab = "Log transformed vraiate age",
         ylab = "Relative Frequency", main =
         "Relative frequency histogram of subject.age.log",
         xlim = c(2.5, 5), ylim = c(0, 1.5), las = 1,
         col = "dodgerblue3", density = 50)
curve(dnorm(x, mean(SanAntonio$subject.age.log), sd(SanAntonio$subject.age.log)),
      col = "red", add = TRUE, lwd = 2)

#2d.
plot(ecdf(SanAntonio$subject.age), xlab = "Age",
     main = "e.c.d.f. of subject.age", las = 1, lwd = 2, pch = NA)
curve(pnorm(x, mean(SanAntonio$subject.age), sd(SanAntonio$subject.age)),
      col = "red", add = TRUE, lwd = 2, lty = 3)
plot(ecdf(SanAntonio$subject.age.log), xlab = "Log transformed vraiate age",
     main = "e.c.d.f. of subject.age.log", las = 1, lwd = 2, pch = NA)
curve(pnorm(x, mean(SanAntonio$subject.age.log), sd(SanAntonio$subject.age.log)),
      col = "red", add = TRUE, lwd = 2, lty = 3)

#2e.
agem = mean(SanAntonio$subject.age)
agesd = sd(SanAntonio$subject.age)
age_lower <- (agem - 2*agesd - agem) / agesd
age_upper <- (agem + 2*agesd - agem) / agesd
pnorm(age_upper) - pnorm(age_lower)

agelm = mean(SanAntonio$subject.age.log)
agelsd = sd(SanAntonio$subject.age.log)
agel_lower <- (agelm - 2*agelsd - agelm) / agelsd
agel_upper <- (agelm + 2*agelsd - agelm) / agelsd
pnorm(agel_upper) - pnorm(agel_lower)
```

```
#2f.
plot(SanAntonio$subject.age, SanAntonio$vehicle.year,
     xlab = "subject.age", ylab = "vehicle.year",
     main = "Age and Vehicle Year in San Antonio", pch = 1, cex = 0.5, col = "navy",
     las = 1, cex.axis = 0.5)

#2g.
cor(SanAntonio$subject.age, SanAntonio$vehicle.year)

#3b.
vehtable <- table(mydata$vehicle.make, mydata$city)
prop.table(vehtable, 2)

#3c.
range(SanAntonio$vehicle.year)

#3d.
table(SanAntonio$vehicle.colour)
table(newor$vehicle.colour)

#3e.
median(SanAntonio$vehicle.colour)
```