CV Homework 1

Code

Ramesh Kanakala

Step 1

Hide

```
#exploring Traffic instead of Boston
library(MASS)
#loading data set
data("Traffic")
str(Traffic)
```

The Traffic data set contains 184 observations describing the 92 days of 1961 as well as 1962 with 4 attributes: traffic accident countar, day, limit, and y, the traffic accident count for that day. All of the attributes are qualitative as they are numerical however ‘limit’, an indicator whether there was a speed limit (no or traffic accident counts), is categorical. The ‘traffic accident countar’ variable should perhaps be a factor as well, as there are only two traffic accident countars, 1961 and 1962. Overall, quite a small data set that seems to originally be used to assess the effect of a speed limit on the motorway accident rate.

2

Hide

```
#display the first few rows
head(Traffic)
#display the last 2 rows
tail(Traffic, n = 2)
#display row 5
Traffic[5,]
#display the first few rows of column 1 by combining head() and indexing
head(Traffic)[1]
#display the variable names
names(Traffic)
```

3

Hide

```
#mean, median, range of the traffic accident count column
mean(Traffic$y)
median(Traffic$y)
range(Traffic$y)
```

4

Hide

```
#histogram of the y column
hist(Traffic$y, main = "Traffic Accident Counts", xlab = "Traffic Accident Counts", )
```

This histogram shows that the traffic accident counts are somewhat right-skewed; the mean is greater than the median. The median would be a better measure of the center of this distribution and observing that a majority of the graph is less than 25 means that the traffic accident counts are generally lower on most days.

5

Hide

```
#correlation between traffic accident count and the day of year
cor(Traffic$day, Traffic$y)
```

The correlation between traffic accident count and median house value is positive meaning that as the days go by, the traffic accident count increases. However, I would say the correlation's magnitude isn't close to 1 and much more close to 0; it is very small meaning an increase in day is most likely not a deciding factor of traffic accident count.

6

Hide

```
#plot showing the traffic accident count vs day of year
plot(Traffic$day, Traffic$y, pch=20, col="blue",
main="Day of year vs. Traffic Accident Count", xlab="Day of Year", ylab="Traffic Accident Count"
)
#correlation between these two variables (only two numerical variables so performed this again)
cor(Traffic$day, Traffic$y)
```

The graph between rooms per dwelling and the median housing value has an upward trend and closely grouped points (ignoring a few outliers) suggesting a strong positive correlation. The correlation value confirms this, being positive and quite close to 1; as the number of rooms of a dwelling increases, the median housing value increases as well most of the time.

7

Hide

```
is.factor(Traffic$year)
#plot of median housing value and chas
plot(Traffic$year, Traffic$y, pch=20, col="blue",
main="Year vs. Traffic Accident Count", xlab="Year", ylab="Traffic Accident Count")
Traffic$year <- as.factor(Traffic$year)
#plot of median housing value and chas as factor
plot(Traffic$year, Traffic$y, pch=20, col="blue",
main="Year vs. Traffic Accident Count", xlab="Year", ylab="Traffic Accident Count")
```

The first plot is a scatterplot attempt to show a relationship using coordinates between 1961 and 1962, though there arent any, wheareas the second is a boxplot that is more focused on the distribution, mostly the interquartile range, of the median traffic accident count for a year value of 1961 and 1962. Looking at the boxplot can lead us to

believe that 1961 has a higher median traffic accident count.

8

Hide

```
#rad variable
Traffic$y
summary(Traffic$y)
unique(Traffic$y)
sum(Traffic$y<20)
#percentage of days where 20 or less accident count
sum(Traffic$y<20)/sum(Traffic$y>0)
```

Taking a look at all the values at once allows us to see what kind of values ‘y’ holds discrete; here it is numerical. Observing the summary() function tells us the variable ranges between 7 and 49, it’s 25th and 75th percentile, as well as it’s mean and median, 21.55 and 20 respectively. The unique() function shows counts that are the only values used once in this column. Using the sum() function helps us see that 88 days have a count less than 24 and dividing that by the number of all observations shows us this is a percentage of 47.82609% of all days.

9

Hide

```
lessthanmed <- Traffic$y
#far is true of y < 20
for(i in 1:length(lessthanmed)) {
  if(lessthanmed[i]<20) {
    lessthanmed[i] <- TRUE
  } else {
    lessthanmed[i] <- FALSE
  }
}
lessthanmed <- as.factor(lessthanmed)
#plot of accessibility of 24 vs 1-23 and median housing value
plot(Traffic$year, lessthanmed, pch=20, col="blue", main="Year vs. Traffic Count < 20", xlab="Ye
ar", ylab="Traffic Count < 20 (1 is true)")
```

The graph between year and boolean traffic count under 20 shows the distribution of traffic count under 20 for each; as 1 means there were less than 20 accidents, the median, it seems that 1962 had days with a higher median accident count.

10

Hide

```
#y
summary(Traffic$year)
#rm
summary(Traffic$day)
#lstat
summary(Traffic$limit)
#medv
summary(Traffic$y)
#neighborhood with highest median housing value
which.max(Traffic$y)
Traffic[132, c(1, 2, 3, 4)]
#plotting if there was a limit against accident count
plot(Traffic$limit, Traffic$y, pch=20, col="blue",
main="Limit vs. Traffic Accident Count", xlab="Limit", ylab="Traffic Accident Count")
```

Day 132, the one with the highest traffic accident count, 49, higher than it's average, 21.55, had no speed limit and was of year 1962. I wanted to explore the limit factor a little more as it seems to be an importact condition in regulated speed and by extension possible accidents. We see a higher median and interquartile range with no speed limit; with no limit, the days seem to be prone to more traffic accidents.