ISTA 116: Lab Assigment #2 (50 pts)

Kyle R Almryde

Problem 1: Categorical Data (25 pts)

a. (3 pts)

```
> table(License)
License
  AL   AS   AZ   CA   CN   CO   CT   DC   DE   FL   GA   GM   IL   IN   KS   KT   LA
   2    1    4   12   14    4  214   53  243  164   40    1   10    8    2    1    2
  LS   MA   MD   ME   MI   MN   MO   MS   NC   ND   NE   NH   NJ   NV   NY   OH   OK
   1  185  702   12   11    4    1    1  137    2    1   16 2267    2 1106   12    1
  PA   RI   SC   TN   TX    U   VA   VT   WA   WI   WV
 539   24   45    5   16  131  505   19    4    5    7
```

b. (3 pts)

```
> rev(sort(round(((table(License)/2267)*100), digit = 2)))
License
   NJ    NY    MD    PA    VA    DE    CT    MA    FL    NC     U    DC    SC    GA    RI    VT
100.00 48.79 30.97 23.78 22.28 10.72  9.44  8.16  7.23  6.04  5.78  2.34  1.99  1.76  1.06  0.84
   TX    NH    CN    OH    ME    CA    MI    IL    IN    WV    WI    TN    WA    MN    CO    AZ
 0.71  0.71  0.62  0.53  0.53  0.53  0.49  0.44  0.35  0.31  0.22  0.22  0.18  0.18  0.18  0.18
   NV    ND    LA    KS    AL    OK    NE    MS    MO    LS    KT    GM    AS
 0.09  0.09  0.09  0.09  0.09  0.04  0.04  0.04  0.04  0.04  0.04  0.04  0.04
>
```

c. (3 pts)

```
> sort(round(((table(License)/2267)*100), digit = 2))
License
   AS    GM    KT    LS    MO    MS    NE    OK    AL    KS    LA    ND
 0.04  0.04  0.04  0.04  0.04  0.04  0.04  0.04  0.09  0.09  0.09  0.09
   NV    AZ    CO    MN    WA    TN    WI    WV    IN    IL    MI    CA
 0.09  0.18  0.18  0.18  0.18  0.22  0.22  0.31  0.35  0.44  0.49  0.53
   ME    OH    CN    NH    TX    VT    RI    GA    SC    DC     U    NC
 0.53  0.53  0.62  0.71  0.71  0.84  1.06  1.76  1.99  2.34  5.78  6.04
   FL    MA    CT    DE    VA    PA    MD    NY    NJ
 7.23  8.16  9.44 10.72 22.28 23.78 30.97 48.79 100.00
```

d. (4 pts)

```
> pie(round(((table(License)/2267)*100), digit = 2), main = "% of tickets by State License", radius = 1)
```

They all look the same, which is to say not very well. Pie charts (to put it bluntly) suck at displaying information, particularly when there is this much information (46 categories all together). It is difficult to judge the relative area visually without some sort of linear measurement.

e. (4 pts)

I might try to label the actual percentage that each slice of the pie equates to, that way we have a numerical measurement that provides us with some information about the pie chart is trying to display. We could also use colors to provide more visual information, or we could just not use a pie chart.

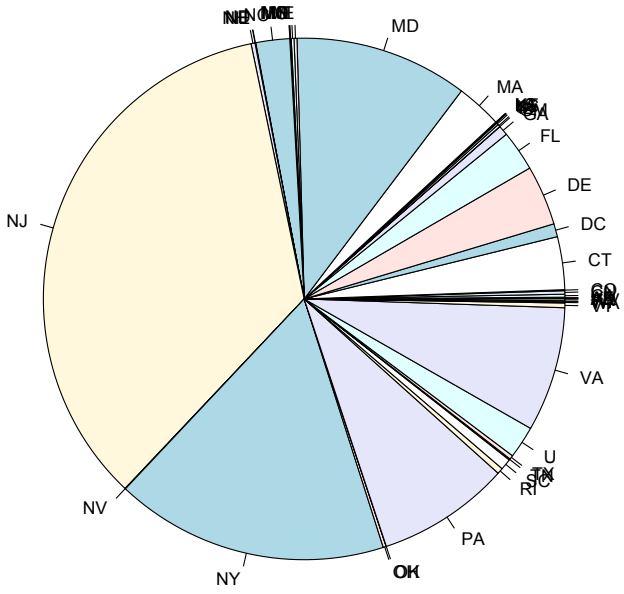
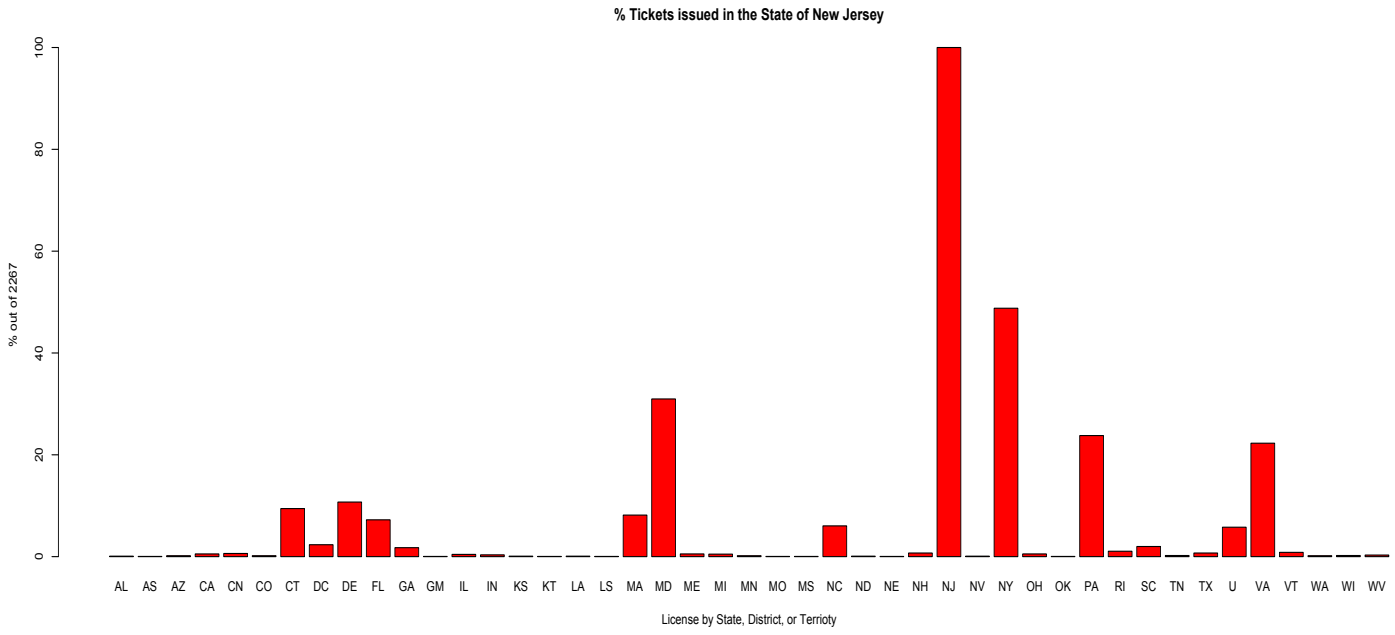


Figure 1: My (sucky) Pie chart

f. (4 pts)

```
barplot(round(((table(License)/2267)*100), digit = 2), main = "%
Tickets issued in the State of New Jersey",  xlab = "License by
State, District, or Terrioty", ylab = "% out of 2267", col = "Red")
```

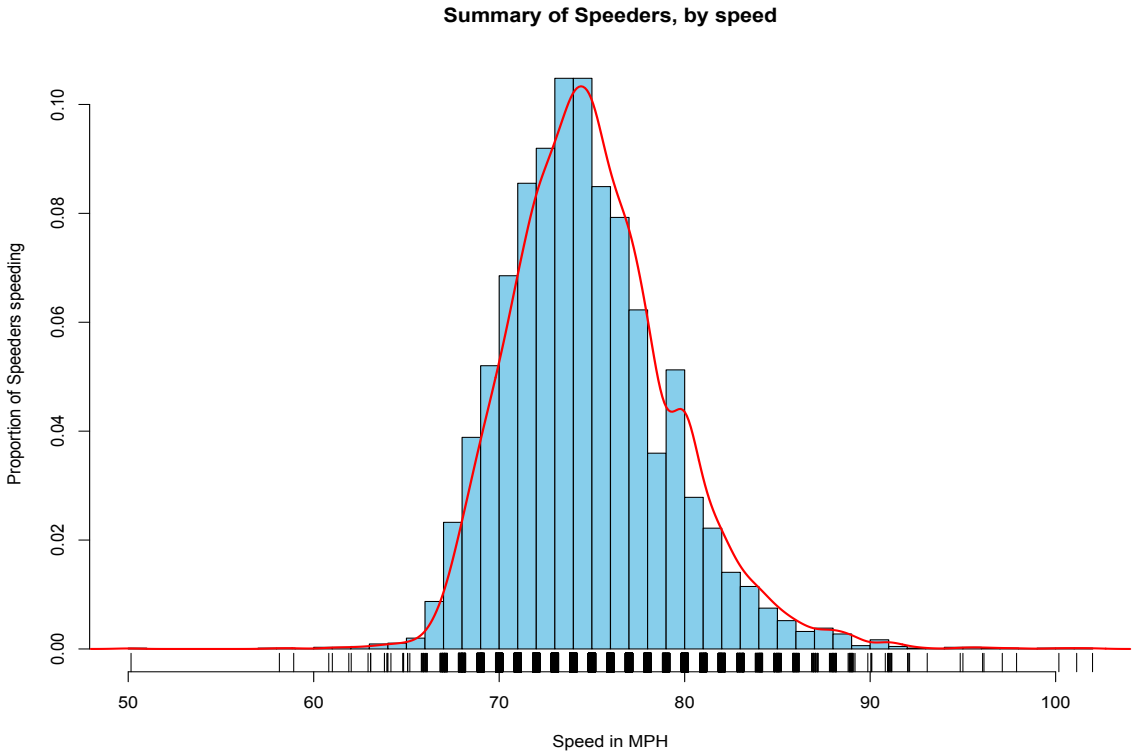


g. (4 pts)
The state of NJ has significantly more tickets than any other state, district, or territory combined. The main factor that would account for this would be that the number of NJ drivers is far greater than any drivers of any state. Considering this is a record of tickets issued in the state of NJ, this is not surprising.

Problem 2: Numeric Data (25 pts)

a. (4 pts)

```
hist(Speed, breaks= 45, main = "Summary of Speeders, by speed", xlab
= "Speed in MPH", ylab = "Proportion of Speeders speeding", col =
"Sky Blue", prob=TRUE); lines(density(Speed), col = "Red", lwd = 2);
rug(jitter(Speed))
```



b. (4 pts)
There is a drop within the distribution at around 78-79mphs, and then a "spike" in the distribution at about 80mphs. This suggests to me, that more people are being caught speeding at 80mphs than at 78-79mph, because people tend to drive around even number or proportional numbers (I do anyway). It is also likely that a speeding ticket is cheaper under 80mphs, so its possible traffic officers are "waiting" for speeders to reach 80mphs in order to hit them with a higher fine.

c. (2 pts)

```
Fast=(1.61*Speed)
```

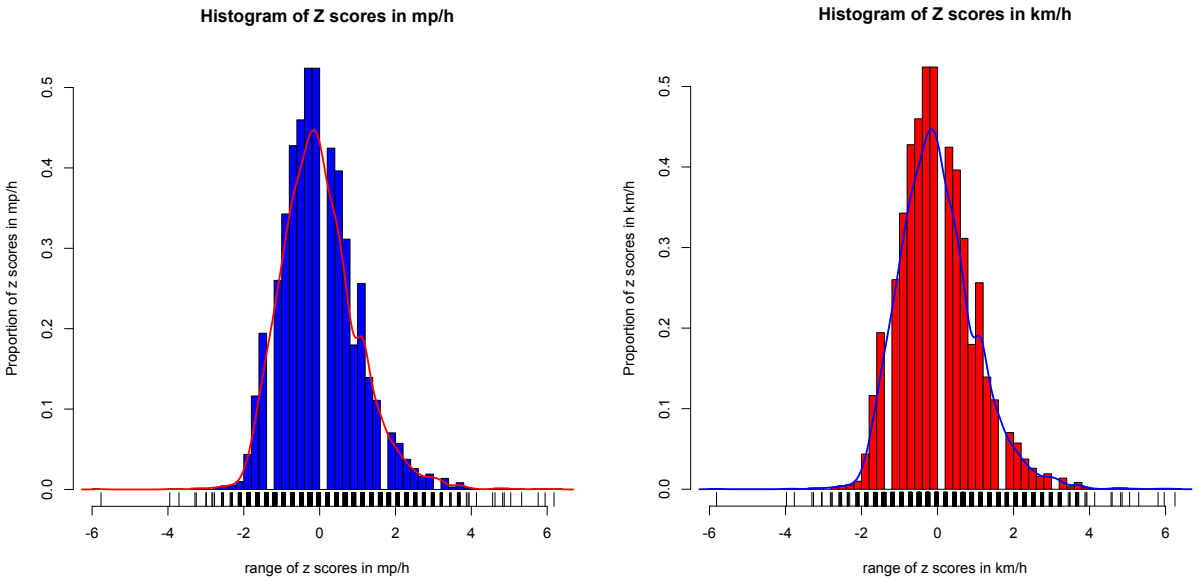
d.

```
mean(Fast)… 120.9094; sd(Fast)… 6.963298
mean(Speed) 75.09899; sd(Speed)… 4.32503
```

The mean and standard deviation of the speed in km/h are simply the multiple of 1.61. The values are certainly different, they 1.61 times greater then the mean and standard deviation of the Speed in mp/h.

e. (5 pts)

```
zSpeed=(Speed-(mean(Speed)))/(sd(Speed))
zFast=(Fast-(mean(Fast)))/(sd(Fast))
```



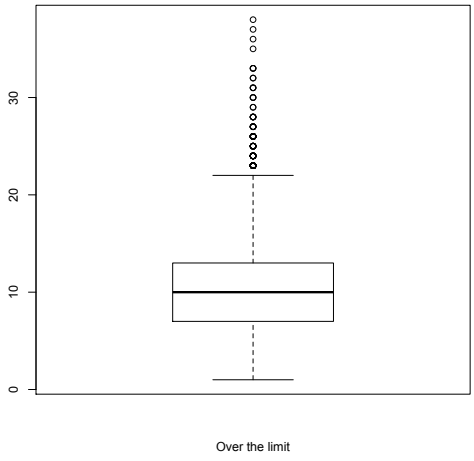

```
hist(zSpeed, prob=T, main="Histogram of Z scores in mp/h", xlab="range of z scores in
mp/h", ylab="Proportion of z scores in mp/h", col="blue", breaks=43);
lines(density(zSpeed), col="Red", lwd=2); rug(jitter(zSpeed))
```

```
hist(zFast, prob=T, main="Histogram of Z scores in km/h", xlab="range of z scores in km/h",
ylab="Proportion of z scores in km/h", col="Red", breaks=43); lines(density(zFast),
col="Blue", lwd=2); rug(jitter(zFast))
```

*They are identical. This is not surprising since a z score is **standard score**, which indicates how many standard deviations a singular point of data is above or below the mean. Because the Speed in mp/h and km/p represents the same thing, we see no difference in z scores and by proxy, the histograms.*

f. (5 pts)

```
fivenum(Overlimit)… 1  7 10 13 38
boxplot(Overlimit, xlab="Over the limit")
```



The box is displaying the *Q1-Q3,* with Q1=1, the mean being 10, Q3=13, and the majority of the data falling within the whiskers, in a of range approximaltey between1-23. There are a few outliers extending beyond 22mp/h over the limit, though those numbers drop off significantly, and are thus represented as outlier points on the graph.