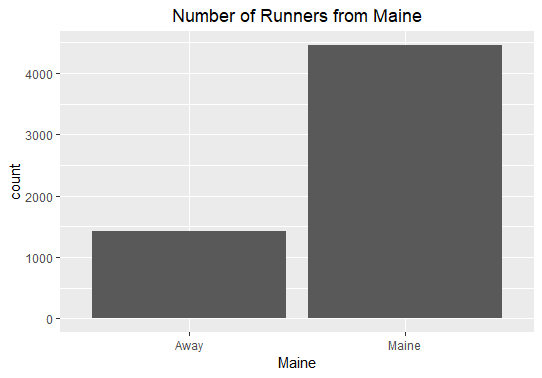
**Mini Project # 2**

**Harrison Jansma**

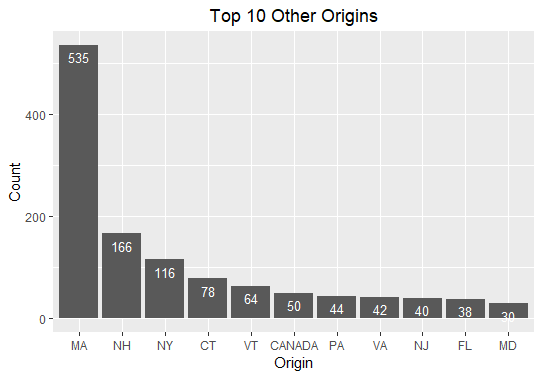
**2/26/2019**

**Section 1**

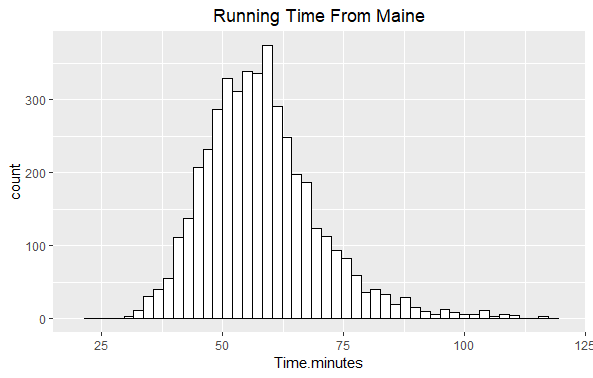
1.a. From the bar plot below, we see that most runners came from within Maine. Of the 5875 runners, 76% of were labelled as from Maine.

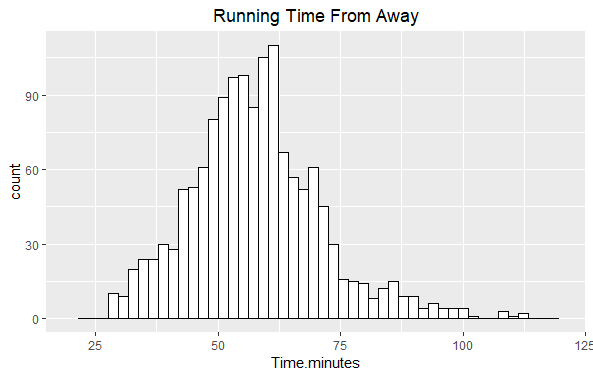


Of the runners that were not from Maine, most came from a few select states. A majority of which came from Massachusetts.



1.b. Below are the histograms for run time in minutes from the group of runners from Maine and from Away.



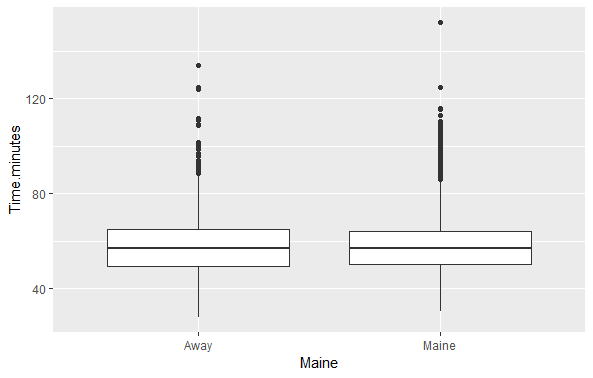


It seems that the distribution of run times from Non-Mainers has higher variance than the run times from within Maine. In the histogram for Away runners, you can see much fatter tails (especially for the left tail).

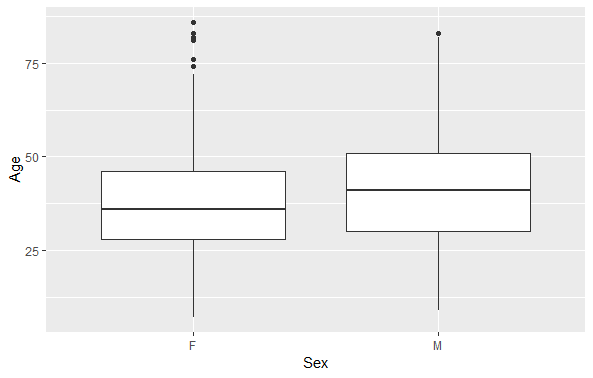
Upon looking at the summary statistics for both populations, we see that the runtime of the Away category does have a higher standard deviation and IQR. We also see that both populations have a rightward skew. (median<mean)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Std Dev** | **Range** | **Median** | **IQR** |
| Maine | 58.20 | 12.19 | 30.6-152.2 | 57.03 | 14.24 |
| Away | 57.82 | 13.84 | 27.8-133.7 | 56.92 | 15.67 |

1.c. From the boxplot we can see that the whiskers and box of the Away group are spread a little bit farther than the Maine group. We also can see that the both groups have quite a few upper outliers, showing that the data set is right skewed.



1.d. We can see from the following boxplot that the distribution of age for male runners has a higher centering than the distribution of female runners. Their variances look to be roughly proportional.

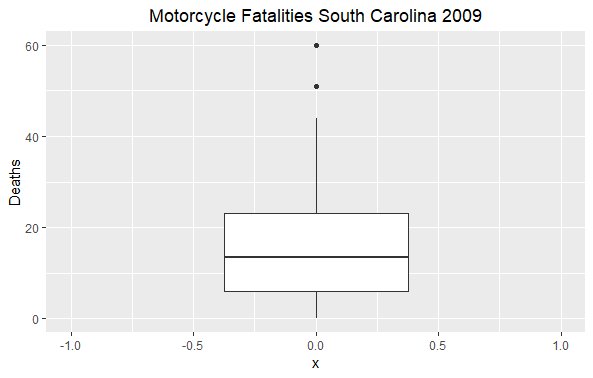


With these summary statistics, we see that both the mean and median for Age of the male population is higher than the female population. We also see that the distribution of age for male runners has a higher spread. Both variance and IQR are higher than the female counterpart.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Std Dev** | **Range** | **Median** | **IQR** |
| F | 37.24 | 12.27 | 7-86 | 36 | 18 |
| M | 40.45 | 13.99 | 9-83 | 41 | 21 |

2. Let’s examine the distribution of motorcycle deaths by county in South Carolina for the year of 2009.

Below is a boxplot of the data. We see that the data is centered around roughly 17 deaths per county. There is also a good deal of variance in the number of deaths per county, the upper whisker extends all the way to around 44 deaths, while the dataset ranges from a minimum of 0 deaths to 60 deaths in a single county.



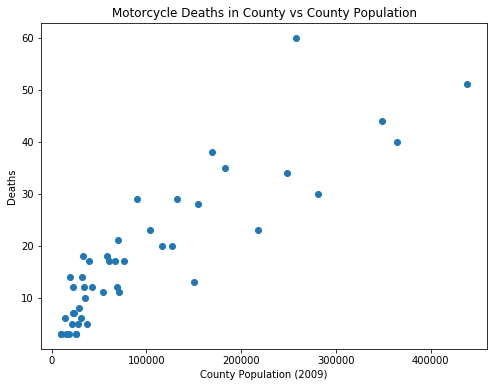
For a more precise description of the dataset, the following summary statistics can supplement the boxplot above. We were correct in identifying that the average county in South Carolina faced 17 motorcycle deaths in the year 2009. The full dataset ranges from 0 to 60, and is significantly right-skewed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Std. Dev** | **Range** | **Median** | **IQR** |
| Fatalities | 17.02 | 13.81 | 0-60 | 13.5 | 17 |

There are three potential upward outliers in the dataset. Each has been identified because their fatality count exceeded the mean by 1.5\*IQR. These counties probably have higher fatality rates because they each have a higher population density than other counties in the dataset.

|  |  |
| --- | --- |
| **County** | **Deaths** |
| Charleston | 44 |
| Greenville | 51 |
| Horry | 60 |

I pulled US census data from 2009 and merged it with the original data. Below is a scatter plot of Deaths vs Population for each SC county. We see a clear positive correlation, and notice that the three counties listed above each have exceptionally high populations compared to other counties in the dataset.



library(tidyverse)

df = read\_csv("C://users//Harrison//Desktop//MiniProject2//roadrace.csv")

names(df)[c(7,8,9,10,11,12)] = c("State.Country","TIme.seconds","MilePace.seconds","From.USA","Maine","Time.minutes")

#barplot of runners from Maine

ggplot(data=df)+

geom\_bar(mapping=aes(x=Maine))+

ggtitle("Number of Runners from Maine")+

theme(plot.title = element\_text(hjust = 0.5))

allRunners=length(df$Maine)

mainers= sum(df$Maine=="Maine")

awayers= allRunners-mainers

percentMainers=mainers/allRunners

#Number of runners from Maine:

mainers

#Number of runners from Away:

awayers

#Percent of all runners from Maine

percentMainers

places=count(df,vars=State.Country,sort=TRUE)

places$percent=round(places$n / allRunners,digits=3)

#Top 10 runner origins after Maine

places[2:12,]

#barplot:top 10 runners from outside Maine

ggplot(data=places[2:12,], aes(x=reorder(vars,-n), y=n)) +

geom\_bar(stat="identity")+

geom\_text(aes(label=n), vjust=1.6, color="white", size=3.5)+

ggtitle("Top 10 Other Origins")+

theme(plot.title = element\_text(hjust = 0.5))+

labs(x="Origin",y="Count")

#subframes of Maine and NonMaine

dfMaine=df[df$Maine=="Maine"]

dfNonMaine=df[df$Maine=="Away"]

#Histogram of Mainers

ggplot(dfMaine, aes(x=Time.minutes))+geom\_histogram(bins=50,color="black",fill="white")+

xlim(20,120)+

theme(plot.title = element\_text(hjust = 0.5))+

ggtitle("Running Time From Maine")

#Histogram of nonMainers (Away)

ggplot(dfNonMaine, aes(x=Time.minutes))+geom\_histogram(bins=50,color="black",fill="white")+

xlim(20,120)+

theme(plot.title = element\_text(hjust = 0.5))+

ggtitle("Running Time From Away (10K)")

#Summary stats for Mainers

mean(dfMaine$Time.minutes)

sd(dfMaine$Time.minutes)

range(dfMaine$Time.minutes)

median(dfMaine$Time.minutes)

IQR(dfMaine$Time.minutes)

#Summary stats for Awayers

mean(dfNonMaine$Time.minutes)

sd(dfNonMaine$Time.minutes)

range(dfNonMaine$Time.minutes)

median(dfNonMaine$Time.minutes)

IQR(dfNonMaine$Time.minutes)

#Boxplot for Runtime in Maine and NonMaine

ggplot(df, aes(x=Maine, y=Time.minutes)) +

geom\_boxplot()

#remove NA values from sex

df = na.omit(df)

#Boxplot for Age and Gender

ggplot(df, aes(x=Sex, y=Age)) +

geom\_boxplot()

#subframes for male and female subpops

dfF=df[df$Sex=="F",]

dfM=df[df$Sex=="M",]

#Summary stats for Female

mean(dfF$Age)

sd(dfF$Age)

range(dfF$Age)

median(dfF$Age)

IQR(dfF$Age)

#Summary stats for Male

mean(dfM$Age)

sd(dfM$Age)

range(dfM$Age)

median(dfM$Age)

IQR(dfM$Age)

cyclesdf = read\_csv("C://users//Harrison//Desktop//MiniProject2//motorcycle.csv")

names(cyclesdf)=c("County","Deaths")

#Histogram of Deaths

ggplot(cyclesdf, aes(x=Deaths))+geom\_histogram(bins=50,color="black",fill="white")+

theme(plot.title = element\_text(hjust = 0.5))+

ggtitle("Motorcycle Fatalities South Carolina 2009")

#barplot:Fatalities by county

ggplot(data=cyclesdf, aes(x=reorder(County,-Deaths), y=Deaths)) +

geom\_bar(stat="identity")+

ggtitle("Fatalities by County")+

theme(plot.title = element\_text(hjust = 0.5))+

theme(axis.text.x = element\_text(angle = 90, hjust = 1))+

labs(x="County",y="Deaths")

#Boxplot for Fatalities

ggplot(cyclesdf, aes(x=0,y=Deaths)) +

xlim(-1,1)+

theme(plot.title = element\_text(hjust = 0.5))+

ggtitle("Motorcycle Fatalities South Carolina 2009")+

geom\_boxplot()

#Summary stats for Fatalities

mean(cyclesdf$Deaths)

sd(cyclesdf$Deaths)

range(cyclesdf$Deaths)

median(cyclesdf$Deaths)

IQR(cyclesdf$Deaths)

outliers = cyclesdf[cyclesdf$Deaths>mean(cyclesdf$Deaths)+1.5\*IQR(cyclesdf$Deaths),]

outliers

popdf = read.table(file="C://users//Harrison//Desktop//MiniProject2//population.txt",sep='\t',header=TRUE)

popdf=popdf[1:46,c(1,11)]

names(popdf)=c("County","Population.2009")

popdf$County=toupper(popdf$County)

cyclesdf = merge(cyclesdf,popdf,by="County")

as.numeric(levels(cyclesdf$Population.2009))[cyclesdf$Population.2009]

ggplot(cyclesdf,aes(x=Population.2009,y=Deaths))+

geom\_point()+

theme(plot.title = element\_text(hjust = 0.5))+

theme(axis.text.x = element\_text(angle = 90))+

ggtitle("Motorcycle Fatalities vs Population")