# Week4_Lab

Adam Peetz

MSDS660 Week 4 Discussion

Regis University

Dr. Siripun Sanguansintukul

November 10th 2022

## Week4 Discussion

```
library(tidyverse)
library(data.table)
library(Hmisc)
library(ggpubr)

no_nas = function(x){
  return(sum(!is.na(x)))
}

data <- read_csv("wildlife.csv",show_col_types = FALSE)

data<-as.data.table(data)

names(data) <- tolower(names(data)) #convert column names to lower case
```

## Discussion Activity:

a. Focus on the collisions with one and two engine planes.
b. Remove outliers using one the methods shown in the demo.
c. Did you decide if you want to impute values? Tell us what you decided on.
d. Compute the average and standard deviation of the speed variable.

```
summary(data$num_engs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   1.000   2.000   2.000   2.042   2.000   4.000    4192
```

```
#select rows where num_eng has a value
num_eng_df <- data[!is.na(data$num_engs), ]
```

```
print("Distance Summary")
```

```
## [1] "Distance Summary"
```

```
summary(num_eng_df$distance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.000   0.000   0.441   0.000  50.000    1050
```

```
print("Speed Summary")
```

```
## [1] "Speed Summary"
```

```
summary(num_eng_df$speed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     0.0   120.0   140.0   142.8   150.0   354.0    2853
```

# Imputing Values

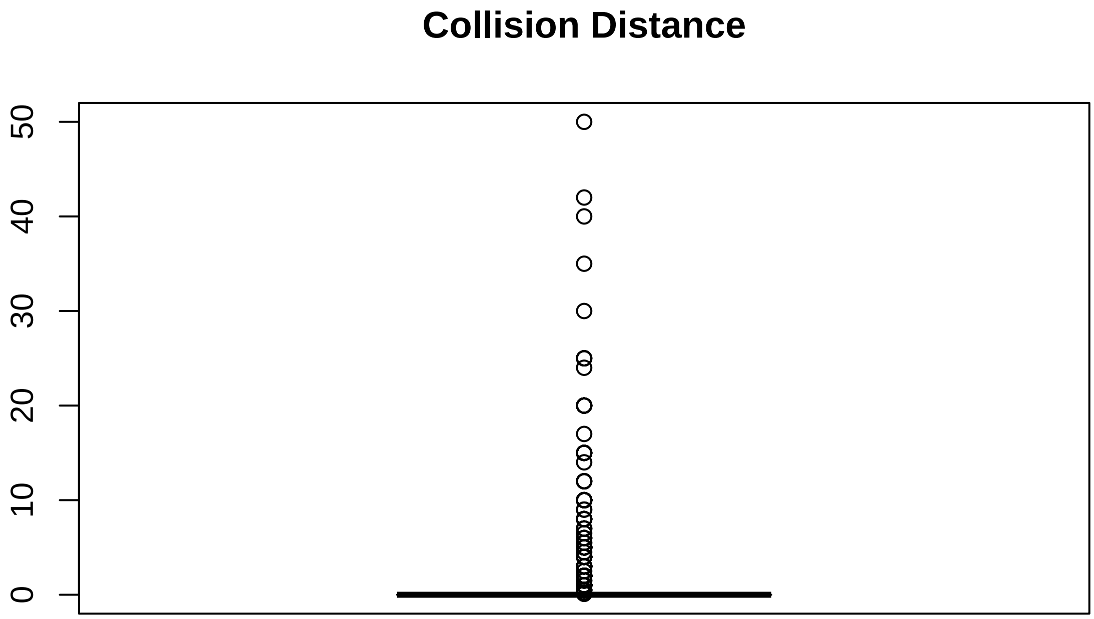
There are 1050 missing distance metrics and 2853 missing speeds. Imputing this many values really skews the standard deviation of the feature. An example of this is shown in boxplots below. I would prefer not to impute this many values but it has been done as an academic proof of work. All of the values for num_eng = 1 speeds are missing and have to be imputed which injects intolerable amounts of bias into values for that feature.

# Detecting Outliers

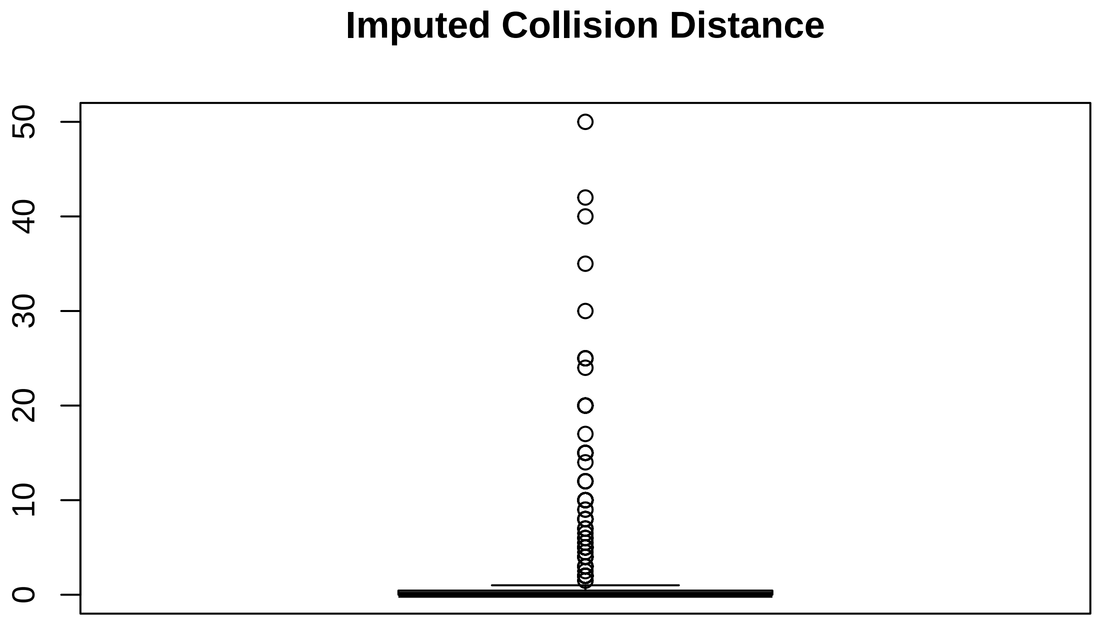
Outliers in this data set are not problematic. All of the values are within the range of possible distances and speeds for an aircraft birdstrike and are likely to be ground truth speeds and distances. I have removed them from the data set using quantile clipping as an academic proof of work but left them in later calculations for confidence intervals.

```
num_eng_df$imputed_distance <- impute(num_eng_df$distance, mean)
num_eng_df$imputed_speed <- impute(num_eng_df$speed, mean)
```

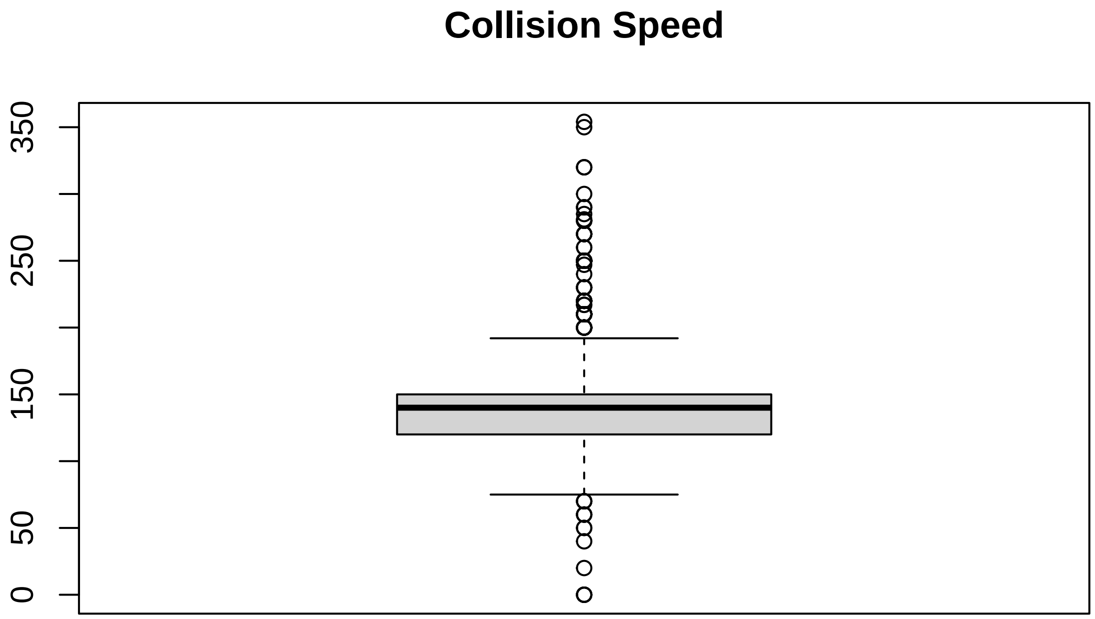
```
boxplot(num_eng_df$distance, main="Collision Distance")
```



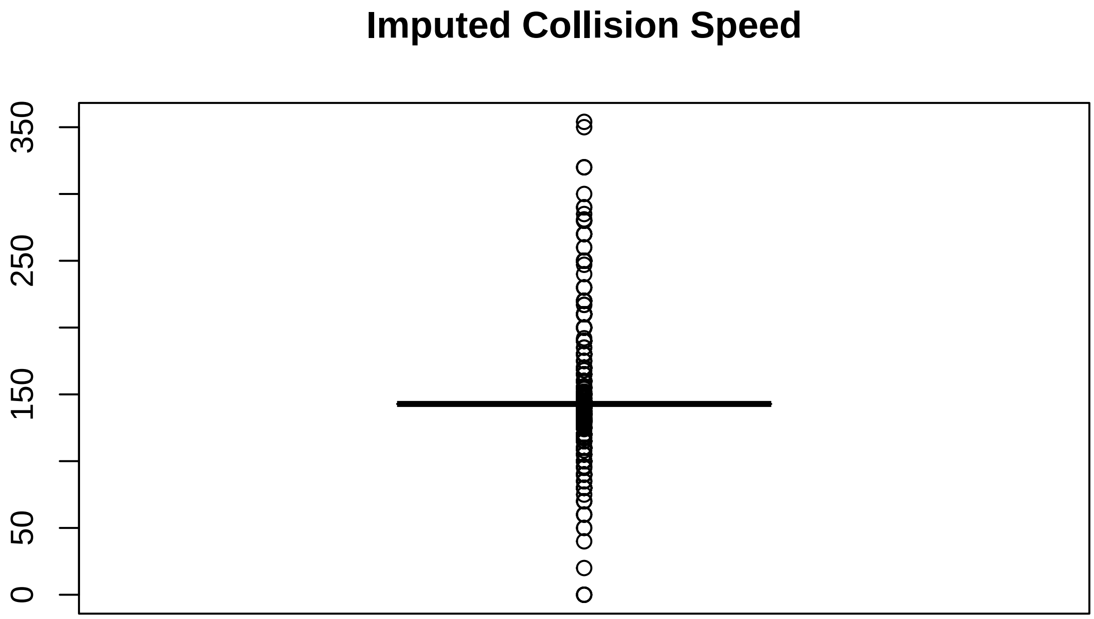
```
boxplot(num_eng_df$imputed_distance, main="Imputed Collision Distance")
```



```
boxplot(num_eng_df$speed, main="Collision Speed")
```



```r
boxplot(num_eng_df$imputed_speed, main="Imputed Collision Speed")
```



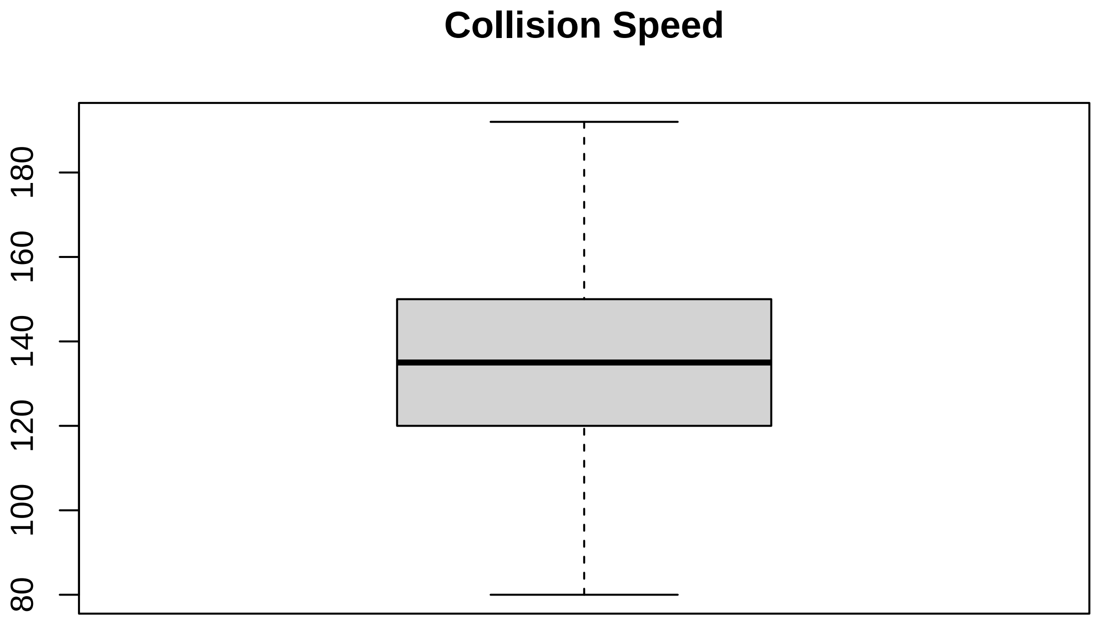
## Removing outliers with Quantile Clipping

```
Q <- quantile(num_eng_df$speed, probs=c(.25, .75), na.rm = TRUE)
iqr <- IQR(num_eng_df$speed, na.rm = TRUE)

up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range

cleaned_df_1 <- subset(num_eng_df, num_eng_df$speed > (Q[1] - 1.5*iqr) & num_eng_df$speed < (Q[2]+1.5*iq

boxplot(cleaned_df_1$speed, main="Collision Speed")
```



## Calculating mean and standard dev

```
#subest data to aircraft eng 1 and 2 only
aircraft_by_eng = num_eng_df[num_eng_df$num_engs  %in% c(1,2), ]

# calculate mean
print("Mean of aircraft speed")
```

```
## [1] "Mean of aircraft speed"
```

```
tapply(aircraft_by_eng$imputed_speed, aircraft_by_eng$num_engs, mean, na.rm=TRUE)
```

```
##        1        2
## 142.8299 142.7881
```

```
# calculate standard deviation
print("Standard deviation of aircraft speed")
```

```
## [1] "Standard deviation of aircraft speed"
```

```
tapply(aircraft_by_eng$imputed_speed, aircraft_by_eng$num_engs, sd, na.rm=TRUE)
```

```
##        1        2
##  0.00000 22.88663
```

# Compute a 95 percent confidence interval for the difference in mean speed at collision between one-engine and two-engine airplanes.

The mean for single engine planes is imputed from the mean of the dataset. All imputed values are 142.8299. This results in a flat confidence interval where the expected mean is 142.8299

There is a bit more variation in the speeds for 2 engine planes. This creates a wider confidence interval where the mean is between 142.0865 and 143.4897

```
#subset to single engine planes
single_df = num_eng_df[num_eng_df$num_engs  %in% c(1), ]

#calculate confidence interval
xbar = mean(single_df$imputed_speed, na.rm=TRUE)
se_xbar = sd(single_df$imputed_speed, na.rm=TRUE)/sqrt(no_nas(single_df$imputed_speed))
lower = xbar - qt(0.975, df = no_nas(single_df$imputed_speed)-1)*se_xbar
upper = xbar + qt(0.975, df = no_nas(single_df$imputed_speed)-1)*se_xbar
c(lower, upper)
```

```
## [1] 142.8299 142.8299
```

```
#subset to two engine planes
two_df = num_eng_df[num_eng_df$num_engs  %in% c(2), ]

#calculate confidence interval
xbar = mean(two_df$imputed_speed, na.rm=TRUE)
se_xbar = sd(two_df$imputed_speed, na.rm=TRUE)/sqrt(no_nas(two_df$imputed_speed))
lower = xbar - qt(0.975, df = no_nas(two_df$imputed_speed)-1)*se_xbar
upper = xbar + qt(0.975, df = no_nas(two_df$imputed_speed)-1)*se_xbar
c(lower, upper)
```

```
## [1] 142.0865 143.4897
```

# Conduct a one sample t-test for the average speed of all bird-airplane collisions

a. What is the conclusion of the one sample t-test?

The mean speed is between 142.0874 mph and 143.4889 mph as calculated by the confidence interval. Alternate hypothesis true mean is not equal to 0. P-value of 0.00000000000000022 is below a 5% significance threshold. Alternate hypothesis is proved, true mean is not equal to 0

```
# perform t.test
t.test(aircraft_by_eng$imputed_speed, mu=0, alternative = "two.sided")   ## gives the 95 percent CI as
```

```
##
##  One Sample t-test
##
## data:  aircraft_by_eng$imputed_speed
## t = 399.49, df = 4094, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  142.0874 143.4889
## sample estimates:
## mean of x
##  142.7882
```