How Much of A Risk?

A LOOK INTO ASTEROIDS AND THEIR THREAT TO EARTH

Aretha Kassegnin | MAT 342 Explorations in Data Science| 11-28

Introduction

This report is going to inform you about asteroids and their risk to the plant we all call Earth. This data includes their mass, how fast they are traveling, the eccentricity of their path, and whether they are hazardous to Earth. This report uses this data to inform you of that important question.

PART 1

The target variable we are using called Hazardous and we are going to see how it relates to the five predictor variables.

1. Mean Motion

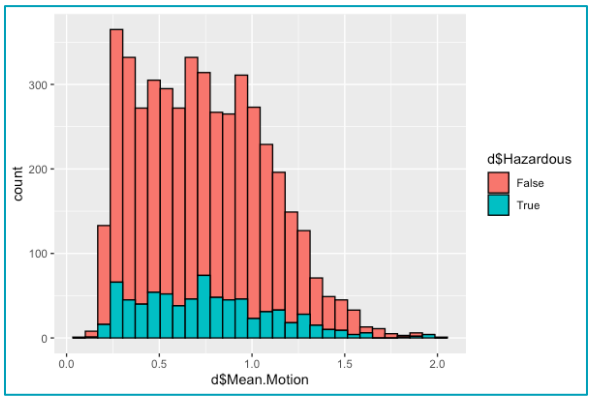


Figure 1.1: Hazardous vs Mean Motion-Stacked

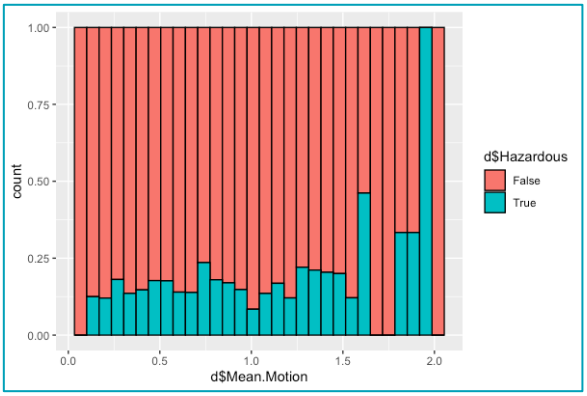


Figure 1.2: Hazardous vs Mean Motion- Fill

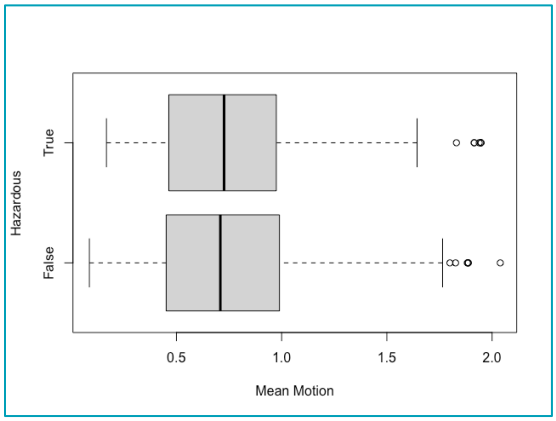


Figure 1.3 Boxplot of Mean Motion vs Hazardous

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.90282	-0.83167	-0.07383	0.00000	0.71923	3.79642

Figure 1.4 5-Number Summary of Mean Motion (ZScore)

The first target variable is Mean Motion compared to the Hazardous variable, looking at figure 1.3 we can see for that the variable Mean Motion has a few high outliers which is about 3.5 standard deviations away from the mean, there are no low outliers. We can see that the data for true and false are both equal in the amount of data in each of them we can see from figure 1.1 that the data is skewed to the right and that Mean Motion

variable compared to hazardous there are more false entries then true. We can see this from figure 1.1, 1.2 and 1.3 but the data is spread out evenly throughout the graph. As for my understanding to what is happening, I believe that the data for false and true is too similar to compare what is going on.

2. Absolute Magnitude

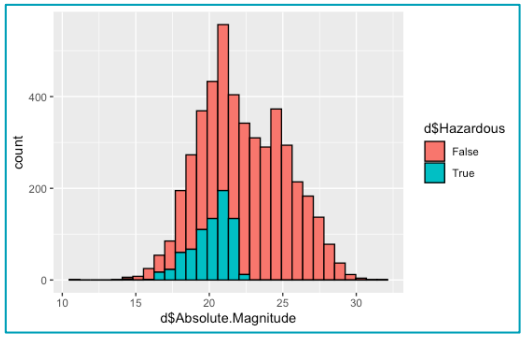


Figure2.1 : Hazardous vs Absolute Magnitude Stacked

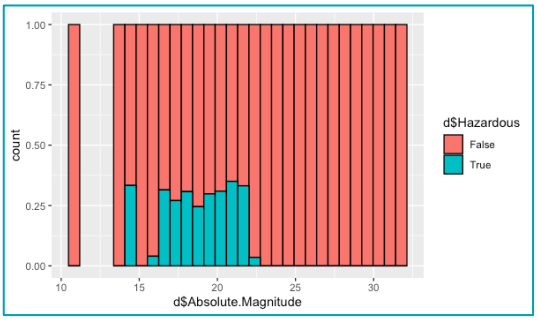


Figure 2.2 Same as 2.1 Filled

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.8423 -0.7499 -0.1272  0.0000  0.7721  3.4010
```

Figure 2.3 5 Number Summary of Absolute Magnitude (Z-Score)

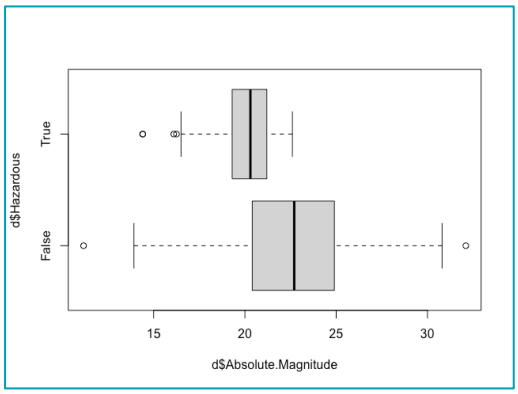


Figure 2.4 Boxplot of Hazardous vs Absolute Magnitude

The second variable was Absolute Magnitude and as we can see from figure 2.3 there are some outliers we see that some data are high , there is a data point 3.4 standard deviations from the mean and a low outlier which is -3.8 standard deviations from the mean we see that there ae more false data then true looking at figure 2.4 and when you look at the distribution of the data we see its mostly symmetric. Looking at the spread I see that in figure 1.2 that after about 24 there are no more true values show in the graphs and that goes for figure 2.2 as well. But looking at figure 2.4 we can see why the boxplot for True shows that its smaller and the IQR is less than the IQR for true. The boxplot for Ture is skewed left while False looks like it's not skewed at all.

3. Est DIA in KM min

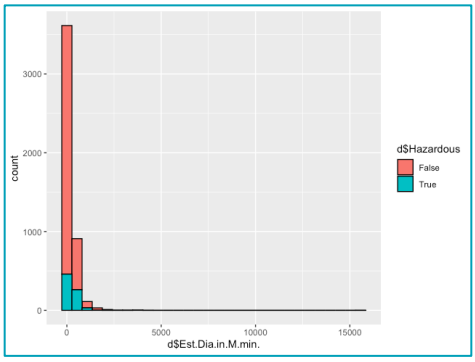


Figure 3.1: Hazardours vs Est DIA in KM min- (stacked)

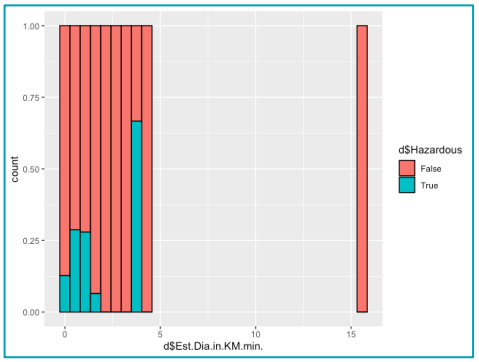


Figure 3.2: Same as 3.1 but filled

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	0	0	0

Figure 3.3: 5 number summary of (Zscore)

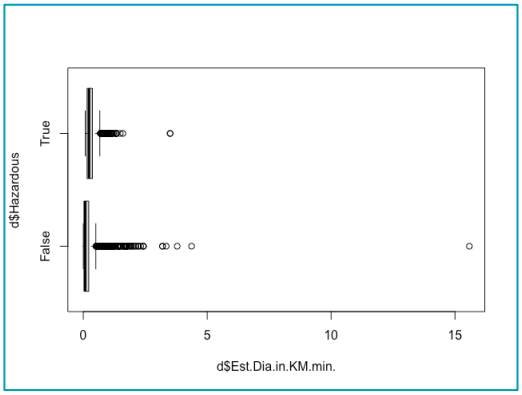


Figure 3.4 Boxplot of Hazardours vs Est DIA in KM min

As you can see from our third variable there is no relationship between hazardous vs estimated distance in KM/min. we can see that figure 3.3 is all zeros and the figure 3.4 we can barely see anything. Even the histograms (figure 3.2 and 3.2) don't really make sense. Even though there are a few outliers shown on figure I know that we will not be seeing this variable in our CART model.

4. Epoch Date Close Approach

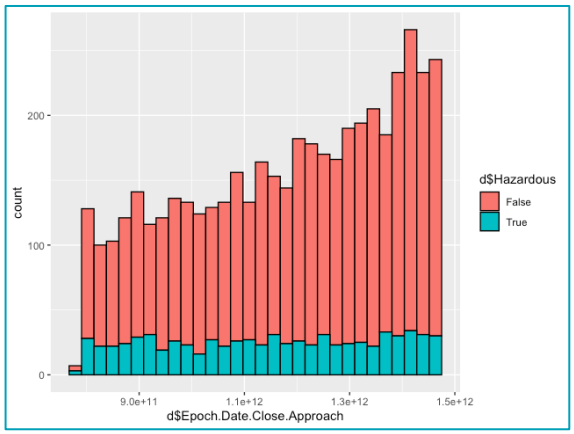


Figure 4.1 Haz. vs Epoch Date Close Approach(stacked).

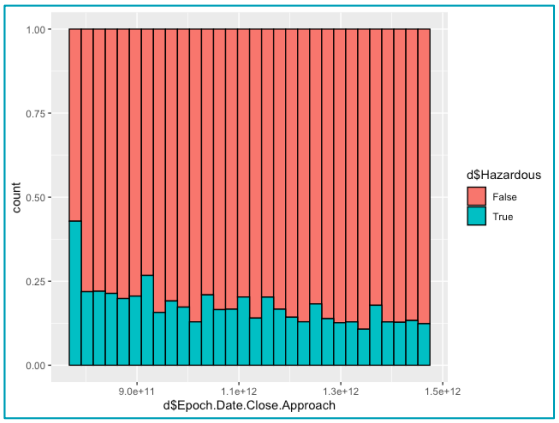


Figure 4.2 Same as 4.1 -filled

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.9729	-0.8292	0.1170	0.0000	0.8866	1.4809

Figure4.3: 5 number summary of (Zscore)

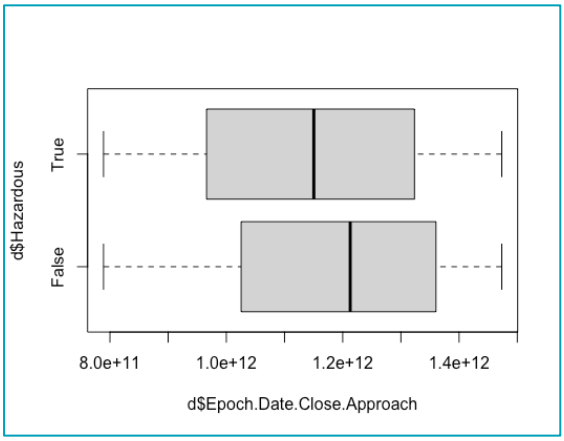
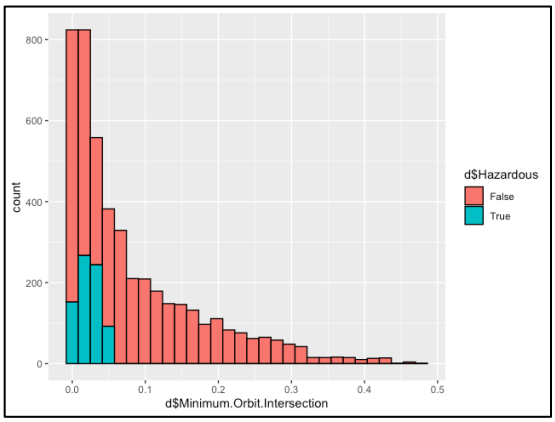


Figure 4.4 Boxplot of Haz. Vs Epoch Date Close Approach

Working with our fourth variable "Epoch Date Close Approach" we can see the relationship between the variables clearly. We see from figure 4.1 that the data is skewed to the right and that there are more false data then true. This is shown in figure 4.2 which shows the histogram filled. Looking at the 5 number summary of the zscore of the variable we see that are no outliers because everything most of the data points are near the mean. Lasty we can look at the boxplot figure 4.4 shows the distribution of the Hazardous compared to Epoch Date Close Approach. Looking at True boxplot we see that the data is evenly spread out and for false the same but IQRs are not similar we can see that True IQR is less then False' IQR.

5. Minimum Orbit Intersection



6. Figure 5.1 Haz vs Minimum Orbit Intersection

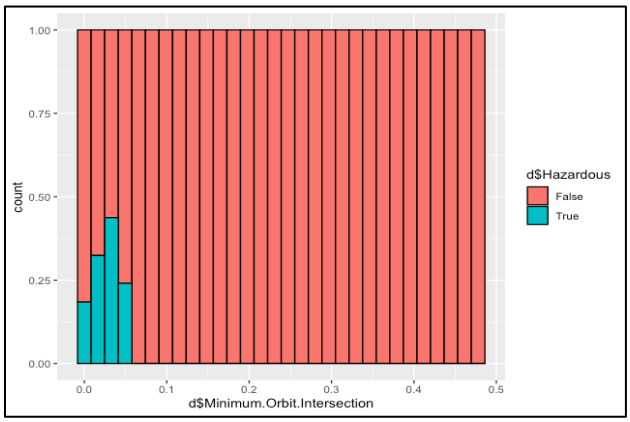


Figure 5.2 Same as 5.1 but fill

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.9116	-0.7501	-0.3871	0.0000	0.4571	4.3806

Figure 5.3 :5 number summary of Z-score

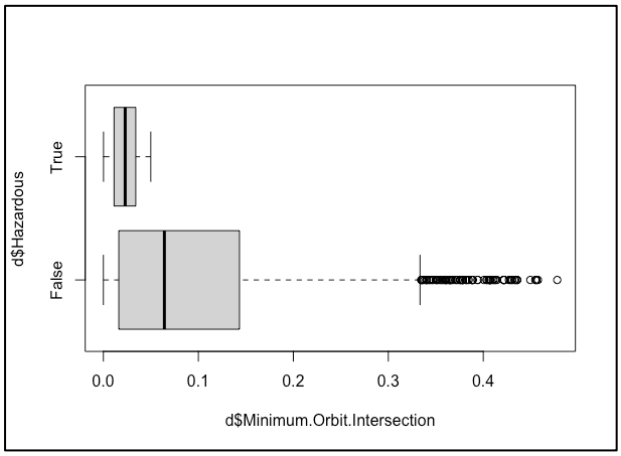


Figure 5.4 Boxplot of Hazardous vs Minimum Orbit Intersection

The last variable being compared to Hazardous is called “Minimum Orbit Intersection”. We see from figure 5.1 that the data is skewed to the right with most data being from the false section. We also see that most of the data the Hazardous false “overpowers” the data. It’s shown in more detail in figure 5.2 where we see true not just in the bottom left and everything else is false. Figure 5.3 shows that there are high outliers about 4 standard deviations away from the mean but there are no low outliers. Lasty the boxplot comparing the two variables show this as well as, looking that the True boxplot is very small and looks like all the data is evenly spread out. For the false boxplot I see that it is skewed to the right and there are not of outliers that are high.

PART 02

I believe that the bassline model will be the variable Minimum Orbit Intersection and model the accuracy of the baseline is 83.87% . This is because when you make a table of the data that has the greatest value

Part 03

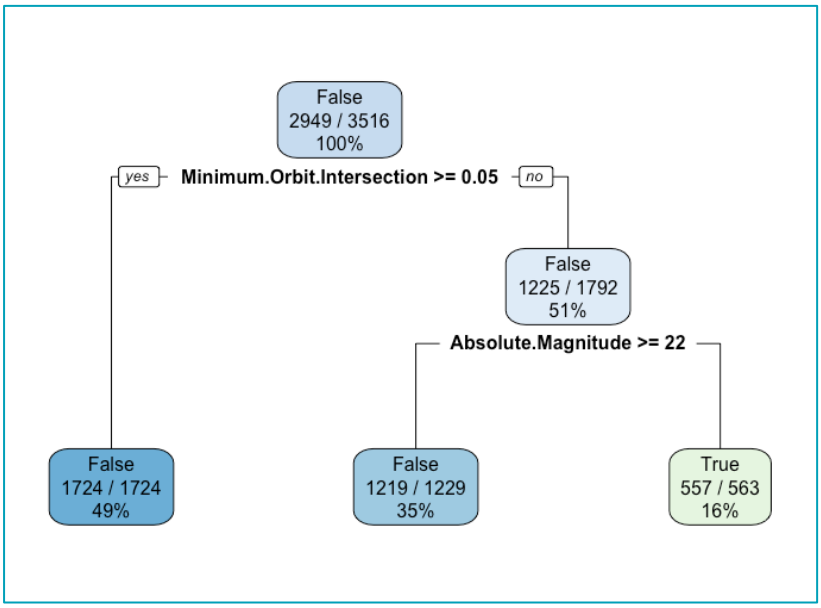


Figure 6.1 CART model based training on the data given

	cart01.pred	
	False	True
False	2943	6
True	10	557

Figure 6.2 Table of Accuracy of training data

The accuracy of this model is (2943+557)/3516 = 0.9954. Since this is greater then the baseline this means the model is good and working correctly for the training data.

```
n= 3516

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 3516 567 False (0.838737201 0.161262799)
  2) Minimum.Orbit.Intersection>=0.04976555 1724   0 False (1.000000000 0.000000000) *
  3) Minimum.Orbit.Intersection< 0.04976555 1792 567 False (0.683593750 0.316406250)
    6) Absolute.Magnitude>=22.05 1229  10 False (0.991863303 0.008136697) *
    7) Absolute.Magnitude< 22.05 563   6 True (0.010657194 0.989342806) *
```

Figure 6.3 Shows what the CART model did

	cart01.pred.test	
	False	True
False	981	2
True	3	185

Figure 6.4 Table of Accuracy of Testing data

Figure 6.4 shows the CART model to the testing data and the accuracy is (981+185)/1171= 0.0957. Again this is greater than the baseline model. Now we are going to see the accuracy of the CART model by doing k-fold cross-validation on our CART model.

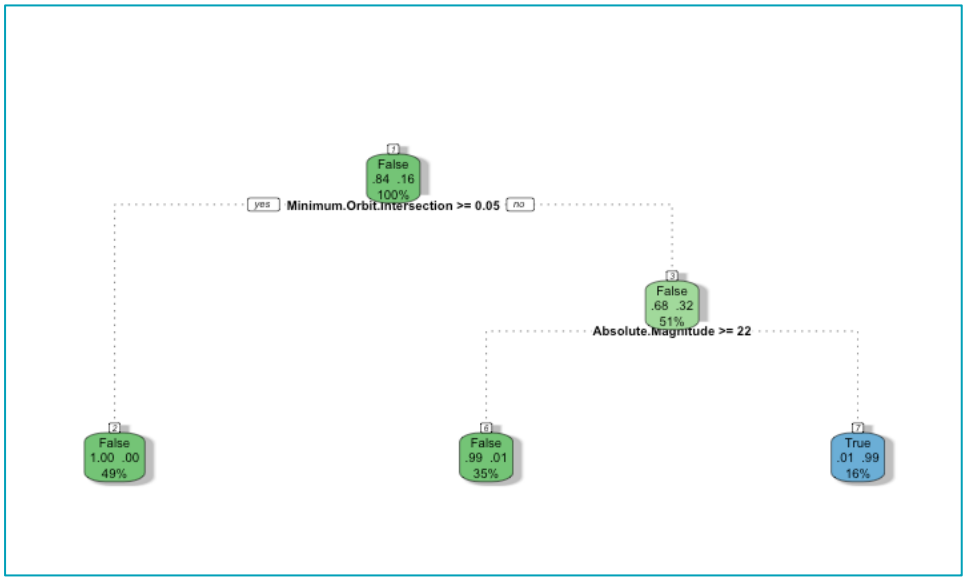


Figure 6.5 k-fold validation on CART model

	pred.kfold.train	
	False	True
False	2943	6
True	10	557

Figure 6.6 Table of Accuracy of K-fold on Training

	pred.kfold	
	False	True
False	981	2
True	3	185

Figure 6.7 Table of Accuracy with Testing

From Figure 6.6 we can see the accuracy is 99.54. We can do the same thing the testing dataset and we can see from Figure 6.7 that our CART model is accuracy predicted 99.57 correctly.

Final Thoughts

Overall the CART model performed well but this means that they might be some overfitting more research must be done with different datasets.

Appendix (The Code)

```
d <- nasa

d <- d[,-c(1, 2, 12, 21:24, 39) ]

d$Hazardous <- as.factor(d$Hazardous)


#data prep

is.na(d)

n <- dim(d)[1]

#adding index

d$Index <-c(1:n)

library(caret)


# Explore how the following predictor variables(x axis) are related to the target variable Hazardous(y axis):

#a. Mean.Motion

#b. Absolute.Magnitude

#c. Est.Dia.in.KM.min.

#d. Epoch.Date.Close.Approach

#e. Minimum.Orbit.Intersection

summary(d$Hazardous)

#a

summary(d$Mean.Motion)

boxplot(d$Mean.Motion)

sd_mean.m <- sd(d$Mean.Motion)

mean_mean.m <- mean(d$Mean.Motion)

z_score_mean.motion <- (d$Mean.Motion - mean_mean.m)/sd_mean.m

summary(z_score_mean.motion)

a <-  table(d$Mean.Motion, d$Hazardous)

boxplot(d$Mean.Motion ~ d$Hazardous, horizontal = TRUE, xlab = "Mean Motion", ylab = "Hazardous")
```

```
par(mfrow=c(1,1))


ggplot(d, aes(d$Mean.Motion))+
  geom_histogram(aes(fill = d$Hazardous),
                 color = "black", position = "stack")


ggplot(d, aes(d$Mean.Motion))+
  geom_histogram(aes(fill = d$Hazardous),
                 color = "black", position = "fill")
#b
summary(d$Absolute.Magnitude)
boxplot(d$Absolute.Magnitude)
boxplot(d$Absolute.Magnitude~ d$Hazardous, horizontal = TRUE)
sd_abs.mag <-sd(d$Absolute.Magnitude)
mean_abs.mag <- mean(d$Absolute.Magnitude)
z_score_abs.mag <- (d$Absolute.Magnitude- mean_abs.mag)/ sd_abs.mag
summary(z_score_abs.mag)


ggplot(d, aes(d$Absolute.Magnitude))+
  geom_histogram(aes(fill = d$Hazardous),
                 color = "black", position = "stack")
ggplot(d, aes(d$Absolute.Magnitude))+
  geom_histogram(aes(fill = d$Hazardous),
                 color = "black", position = "fill")


#c #this variable will not show up in the Cart Model because its too small
summary(d$Est.Dia.in.KM.min.)
boxplot(d$Est.Dia.in.KM.min.~d$Hazardous, horizontal = TRUE)
```

```
sd_est.dia <- sd(d$Est.Dia.in.KM.min.)
mean_est.dia <- (d$Est.Dia.in.KM.min.)
z_score_est.dia <- (d$Est.Dia.in.KM.min. - mean_est.dia)/ sd_est.dia
summary(z_score_est.dia)

ggplot(d, aes(d$Est.Dia.in.M.min.))+
  geom_histogram(aes(fill = d$Hazardous),
                 color = "black", position = "stack")
ggplot(d, aes(d$Est.Dia.in.KM.min.))+
  geom_histogram(aes(fill = d$Hazardous),
                 color = "black", position = "fill")

#d
summary(d$Epoch.Date.Close.Approach)
boxplot(d$Epoch.Date.Close.Approach ~ d$Hazardous, horizontal = TRUE)
sd_epoch <- sd(d$Epoch.Date.Close.Approach)
mean_epoch <- mean(d$Epoch.Date.Close.Approach)
z_score_epoch <- (d$Epoch.Date.Close.Approach- mean_epoch)/ sd_epoch
summary(z_score_epoch)
ggplot(d, aes(d$Epoch.Date.Close.Approach))+
  geom_histogram(aes(fill = d$Hazardous),
                 color = "black", position = "stack")
ggplot(d, aes(d$Epoch.Date.Close.Approach))+
  geom_histogram(aes(fill = d$Hazardous),
                 color = "black", position = "fill")
#e
summary(d$Minimum.Orbit.Intersection)
boxplot(d$Minimum.Orbit.Intersection)
```

```
boxplot(d$Minimum.Orbit.Intersection~ d$Hazardous, horizontal = TRUE)
sd_min <- sd(d$Minimum.Orbit.Intersection)
mean_min <- mean(d$Minimum.Orbit.Intersection)
z_score_min <-(d$Minimum.Orbit.Intersection- mean_min)/ sd_min
summary(z_score_min)
ggplot(d, aes(d$Minimum.Orbit.Intersection))+
  geom_histogram(aes(fill = d$Hazardous),
                 color = "black", position = "stack")
ggplot(d, aes(d$Minimum.Orbit.Intersection))+
  geom_histogram(aes(fill = d$Hazardous),
                 color = "black", position = "fill")
#Part 2
head(d)
set.seed(325)
inTrain <-createDataPartition( y= d$Hazardous,
                               p= .75,
                               list = FALSE)
d.train <- d[ inTrain ,  ]
d.test <- d[ -inTrain ,  ]
dim(d)
dim(d.test)
dim(d.train)
table(dim(d.test))
#a. Mean.Motion
#b. Absolute.Magnitude
#c. Est.Dia.in.KM.min.
#d. Epoch.Date.Close.Approach
#e. Minimum.Orbit.Intersection
```

```
head(d.train)
d$Mean.Motion
hist(d$Mean.Anomaly)
par(mfrow=c(1,2))
hist(d.train$Mean.Motion)
hist(d.test$Mean.Motion)
summary(d.train$Mean.Motion)
summary(d.test$Mean.Motion)

# the variable hazar. is

d.train$trainortest <-
  rep("train", nrow(d.train))
d.test$trainortest <-
  rep("test", nrow(d.test))

d.all <- rbind(d.train, d.test)

head(d.all)
par(mfrow=c(1,2))
hist(d.train$Mean.Motion)
hist(d.test$Mean.Motion)
par(mfrow=c(1,1))
boxplot(Mean.Motion ~ as.factor(trainortest),
        data = d.all)
kruskal.test(Mean.Motion ~ as.factor(trainortest),
             data = d.all)$p.value
```

```
hist(d$Absolute.Magnitude)
par(mfrow=c(1,2))
hist(d.train$Absolute.Magnitude)
hist(d.test$Absolute.Magnitude)
par(mfrow=c(1,1))
summary(d.test$Absolute.Magnitude)
summary(d.train$Absolute.Magnitude)
boxplot(Absolute.Magnitude ~ as.factor(trainortest),
        data = d.all)
kruskal.test(Absolute.Magnitude~ as.factor(trainortest),
             data = d.all)$p.value

hist(d$Est.Dia.in.KM.min.)
par(mfrow=c(1,2))
hist(d.test$Est.Dia.in.KM.min.)
hist(d.train$Est.Dia.in.KM.min.)
par(mfrow=c(1,1))
boxplot(Est.Dia.in.KM.min. ~ as.factor(trainortest),
        data = d.all)
kruskal.test(Est.Dia.in.KM.min.~ as.factor(trainortest),
             data = d.all)$p.value

hist(d$Epoch.Date.Close.Approach)
par(mfrow=c(1,2))
hist(d.train$Epoch.Date.Close.Approach)
hist (d.test$Epoch.Date.Close.Approach)
par(mfrow=c(1,1))
```

```
boxplot(Epoch.Date.Close.Approach ~ as.factor(trainortest),
        data = d.all)
kruskal.test(Epoch.Date.Close.Approach~ as.factor(trainortest),
             data = d.all)$p.value


hist(d$Minimum.Orbit.Intersection)
hist(d.train$Minimum.Orbit.Intersection)
hist(d.test$Minimum.Orbit.Intersection)
boxplot(Minimum.Orbit.Intersection ~ as.factor(trainortest),
        data = d.all)
kruskal.test(Minimum.Orbit.Intersection~ as.factor(trainortest),
             data = d.all)$p.value
#Part 3
library(psych)
library(plyr)
library(ggplot2)
library(rpart)
library(rpart.plot)




carto1 <- rpart(Hazardous ~ .,
                data = d.train,
                method = "class")
carto1
rpart.plot(carto1, type = 2, extra =102)
```

```
cart.train <- predict(carto1, newdata = d.train, type ="class")

cart.pred_test<- predict(carto1, newdata = d.test, type ="class")

summary(carto1)

table(d.train$Hazardous, cart_seen)

table (d.train$Hazardous, cart.pred_test)
```