Data Analysis and Exploration of Asteroids that Have Been Detected Close to Earth

Nutless Neanderthals

2024-04-28

Introduction

Our data is data on asteroids detected close to Earth from Nasa. It was retrieved from Kaggle, and the creator of the dataset took the data from NASA's Center for Near Earth Object Studies webiste. There are 40 variables in the dataset and 4687 observations, including multiple continuous variables that give physical descriptors of the asteroid (e.g. miss distance, diameter) as well as two categorical variables (near-miss date and hazardous classification). We are especially interested in the hazardous classification, which is an indication given by Nasa to predict whether an asteroid is likely to collide with earth. According to the Center for Near Earth Object Studies, an asteroid is considered 'hazardous' if "In other words, asteroids that can't get any closer to the Earth (i.e., MOID) than 0.05 au (roughly 7,480,000 km or 4,650,000 mi) or are smaller than about 140 m (~500 ft) in diameter (i.e., H = 22.0 with assumed albedo of 14%) are not considered PHAs." While this means that size of the asteroid is liekly a clear correlate for whether an asteroid is hazardous or not, we are still interested about whether other features of the asteroid correlate with its state of hazard. Another thing we are interested in is studying correlates for relative velocity- do asteroids of different sizes have different velocity?

Data Overview and Cleaning

We chose to study eight specific variables:

Hazardous: either True or False, tells us if a asteroid is a threat or not

Eccentricity: a scalar that measures the deviation of an orbit from a perfect circle. Ranges from 0 (a perfect circle) to 1. A more eccentric orbit means the path of the asteroid varies more significantly in distance from the sun

Estimated Diameter in KM(Max): the maximum estimated diameter of the asteroid in kilometers.

Absolute Magnitude: measure of the intrinsic brightness of an asteroid. It is defined as the apparent brightness an object would have if it were located exactly one astronomical unit (AU) from both the sun and the observer. Lower values mean the object is brighter.

Relative Velocity in KM/Sec: indicates the velocity of the asteroid relative to Earth, measured in kilometers per second. It represents the speed at which the asteroid is moving relative to Earth.

Miss Distance: the distance by which an asteroid passes near Earth, usually measured in kilometers. It represents the closest point in the asteroid's orbit to Earth. The dataset contains 4 different columns corresponding to different units for Miss Distance; our analysis uses the Miss Distance measured in kilometers.

Est diameter in KM(min): This represents the minimum estimated diameter of the asteroid in kilometers. Like the maximum diameter, it's an estimation based on the asteroid's observed brightness.

Close Approach Date: A continuous variable that represents the exact day the asteroid was detected close to Earth.

```
## [1] "Variable Names"
```

```
##  [1] "Neo.Reference.ID"             "Name"
##  [3] "Absolute.Magnitude"           "Est.Dia.in.KM.min."
##  [5] "Est.Dia.in.KM.max."           "Est.Dia.in.M.min."
##  [7] "Est.Dia.in.M.max."            "Est.Dia.in.Miles.min."
##  [9] "Est.Dia.in.Miles.max."        "Est.Dia.in.Feet.min."
## [11] "Est.Dia.in.Feet.max."         "Close.Approach.Date"
## [13] "Epoch.Date.Close.Approach"    "Relative.Velocity.km.per.sec"
## [15] "Relative.Velocity.km.per.hr"  "Miles.per.hour"
## [17] "Miss.Dist..Astronomical."     "Miss.Dist..lunar."
## [19] "Miss.Dist..kilometers."       "Miss.Dist..miles."
## [21] "Orbiting.Body"                "Orbit.ID"
## [23] "Orbit.Determination.Date"     "Orbit.Uncertainity"
## [25] "Minimum.Orbit.Intersection"   "Jupiter.Tisserand.Invariant"
## [27] "Epoch.Osculation"             "Eccentricity"
## [29] "Semi.Major.Axis"              "Inclination"
## [31] "Asc.Node.Longitude"           "Orbital.Period"
## [33] "Perihelion.Distance"          "Perihelion.Arg"
## [35] "Aphelion.Dist"                "Perihelion.Time"
## [37] "Mean.Anomaly"                 "Mean.Motion"
## [39] "Equinox"                      "Hazardous"
```

```
## [1] "Dataset Dimensions"
```

```
## [1] 4687   40
```

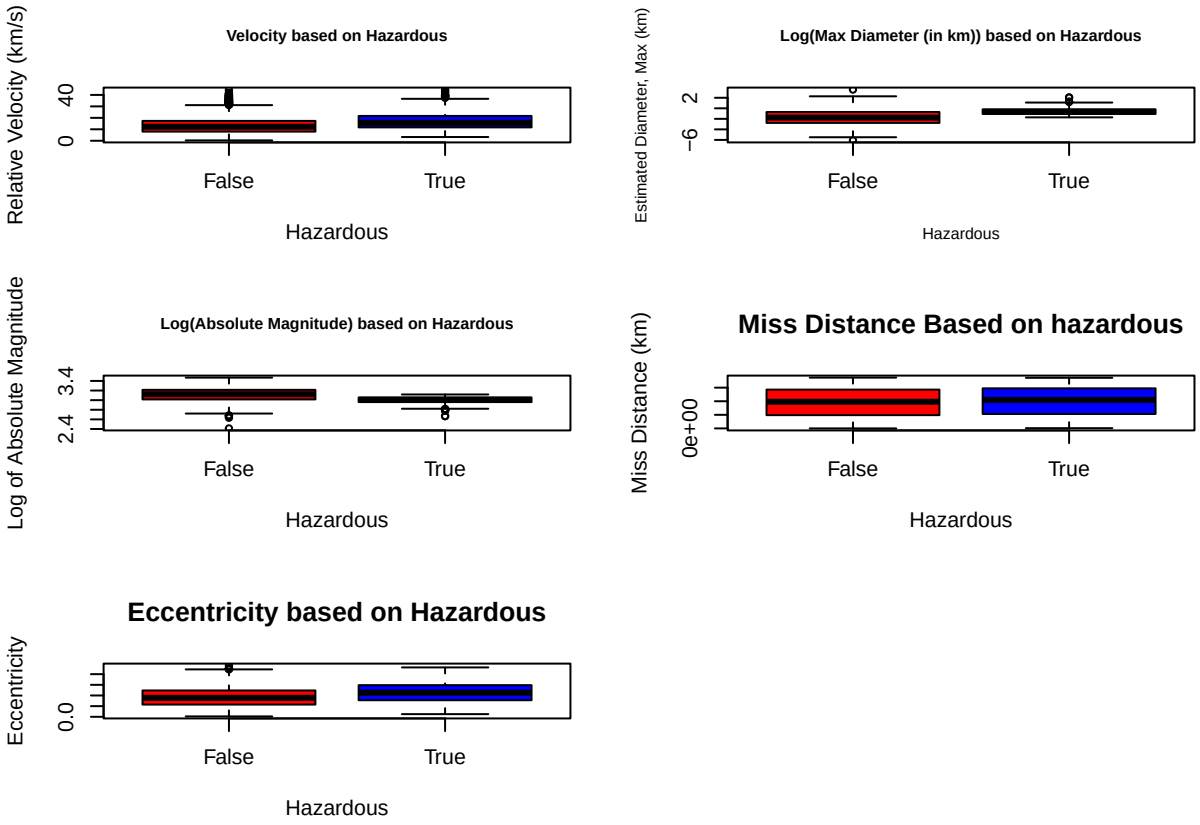
Data Cleaning

Our primary data cleaning revolved around changing our second categorical variable, date of close approach, into several larger categories each representing a 5 year-span. We hoped to do this to find correlates between date of close approach and other statistics.

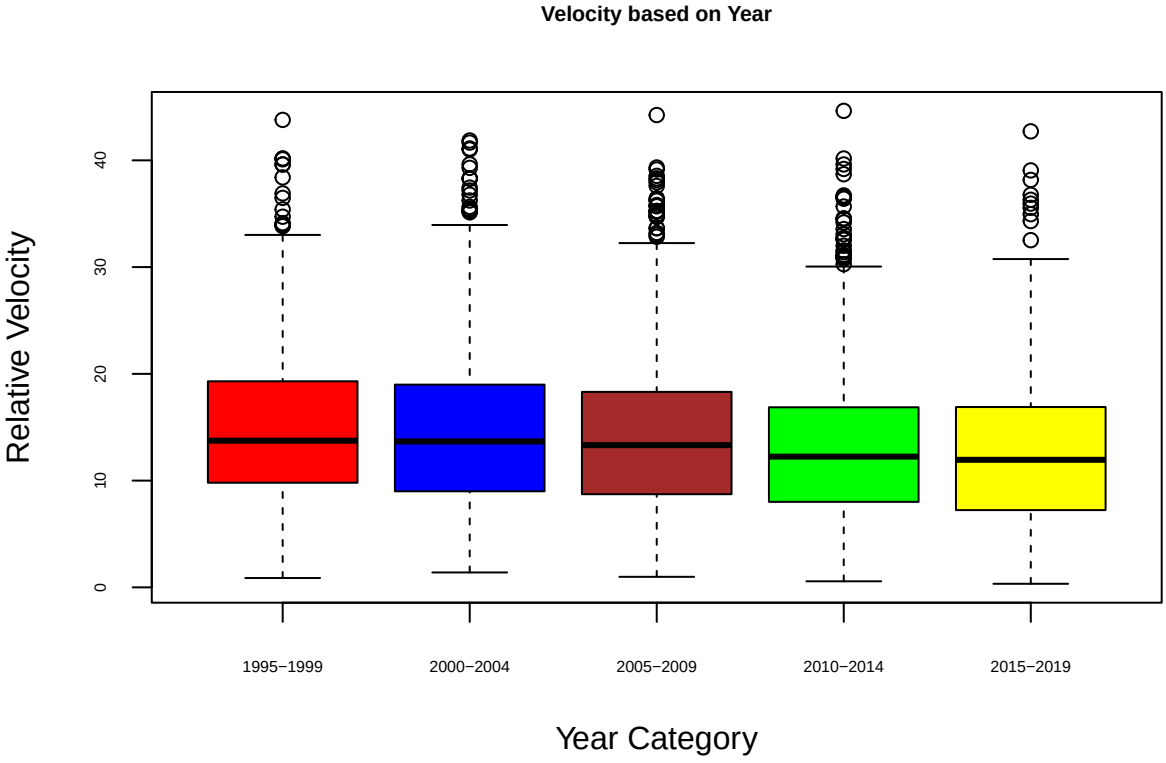
Graphics

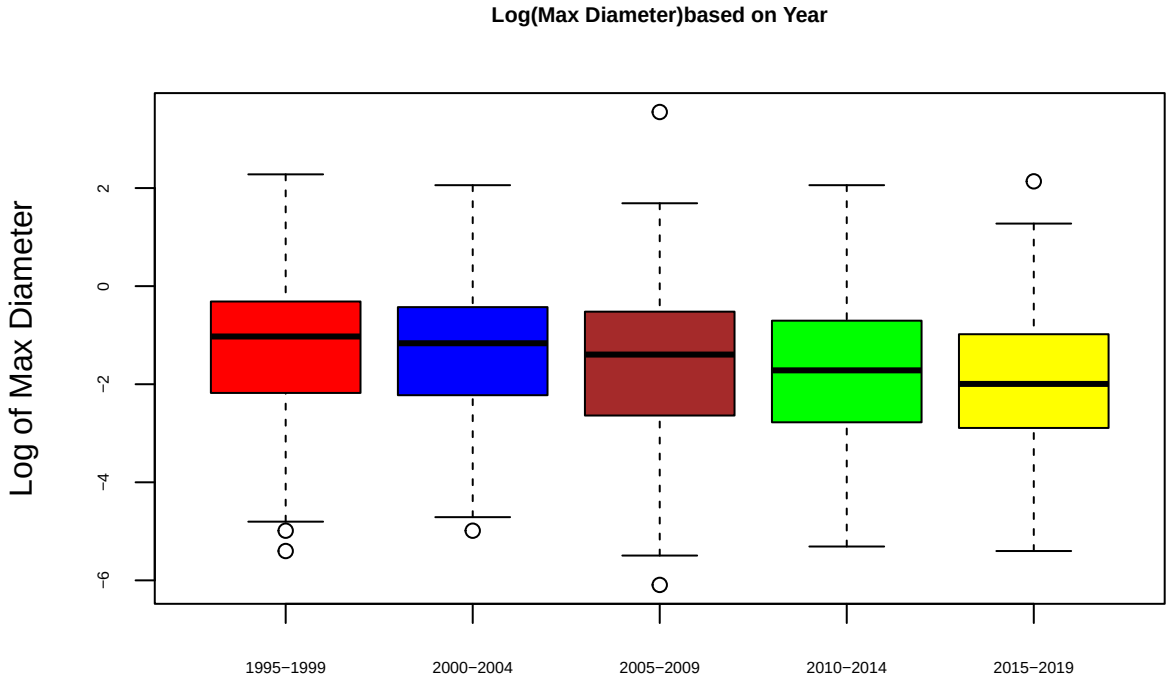
Boxplots

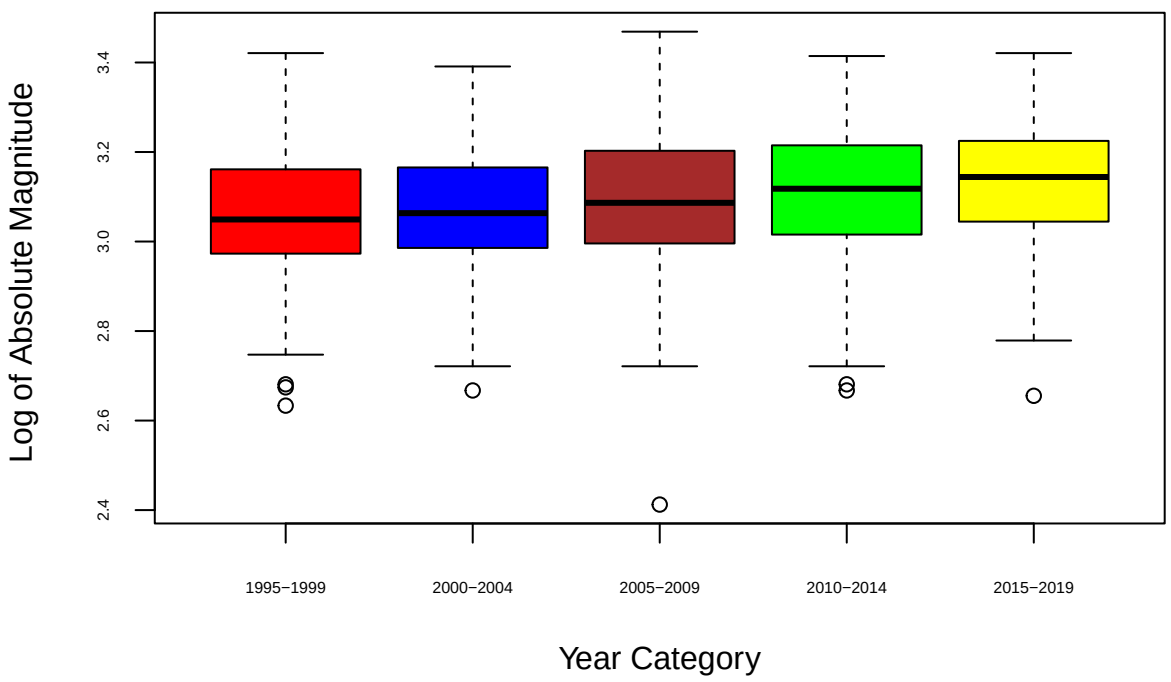
We wanted to find correlations for other variables by hazard and by close approach year.

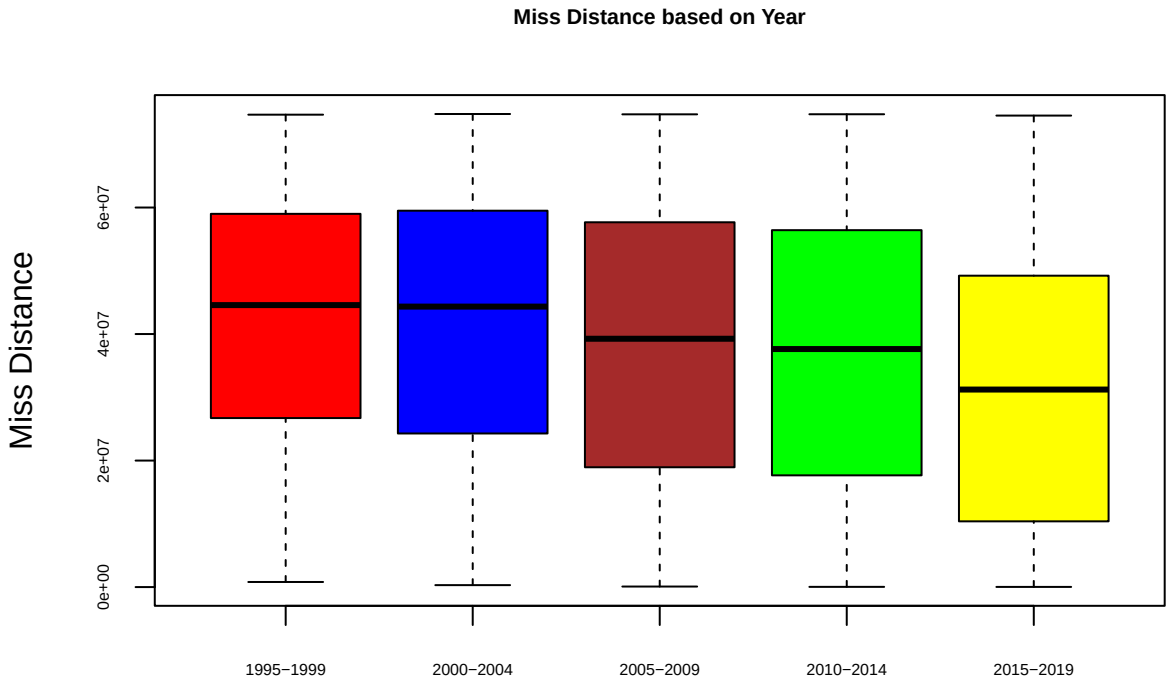


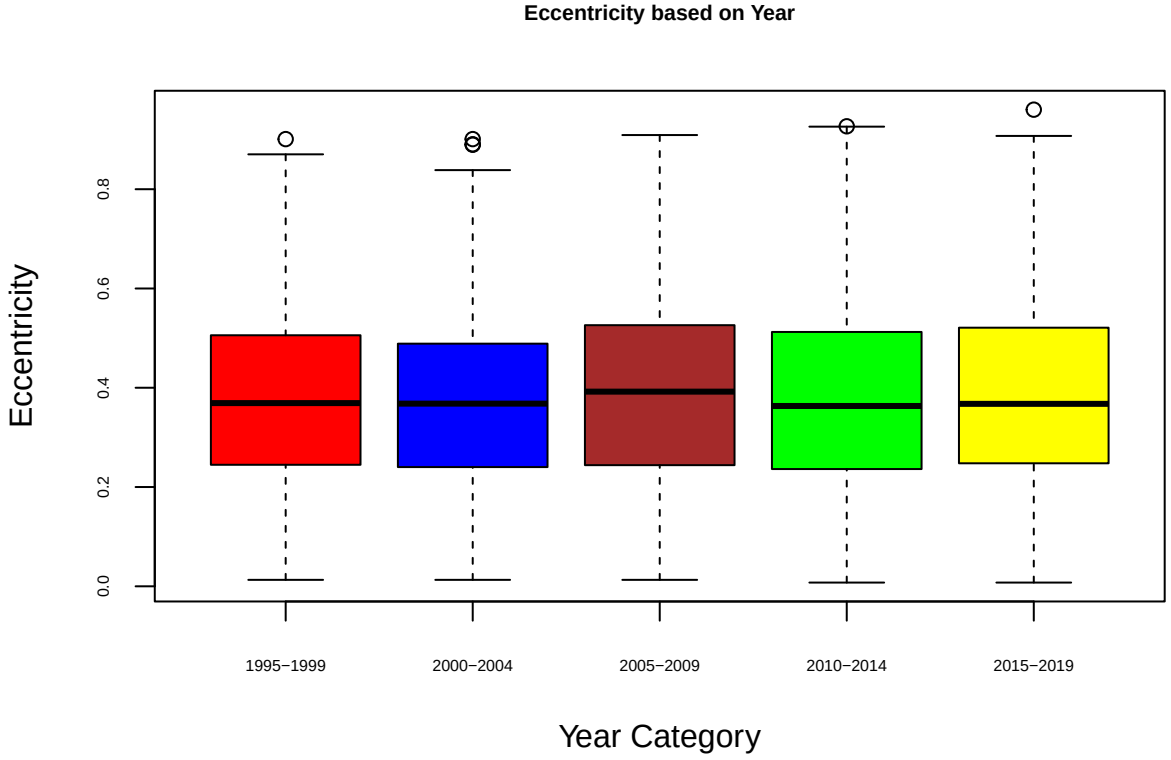
There is a clear correlation between hazardous and the variables relative velocity, log(max estimated diameter), log(absolute magnitude), and eccentricity. There doesn't seem to be much of a difference between hazardous and non-hazardous asteroids and miss distance.




Log(Max Diameter)based on Year
Log of Max Diameter
2
0
−2
−4
−6
1995−1999
2000−2004
2005−2009
2010−2014
2015−2019
Year Category


Log(Absolute Magnitude) based on Year
Log of Absolute Magnitude
3.4
3.2
3.0
2.8
2.6
2.4
1995−1999
2000−2004
2005−2009
2010−2014
2015−2019
Year Category

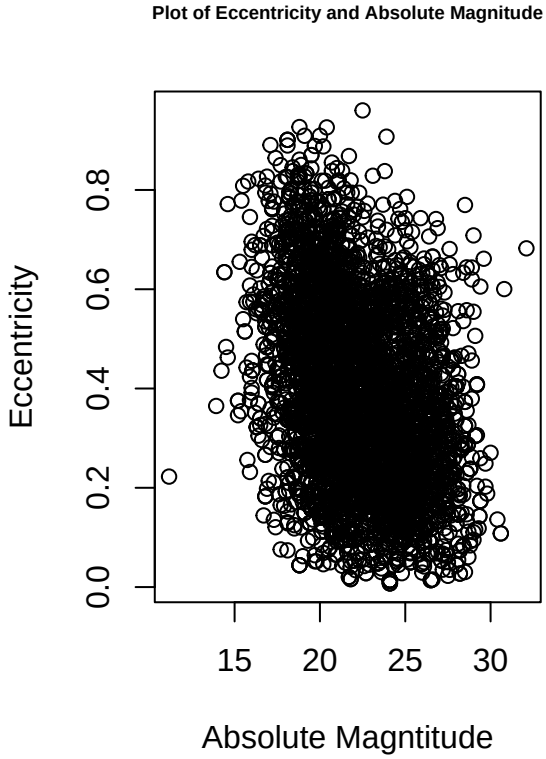


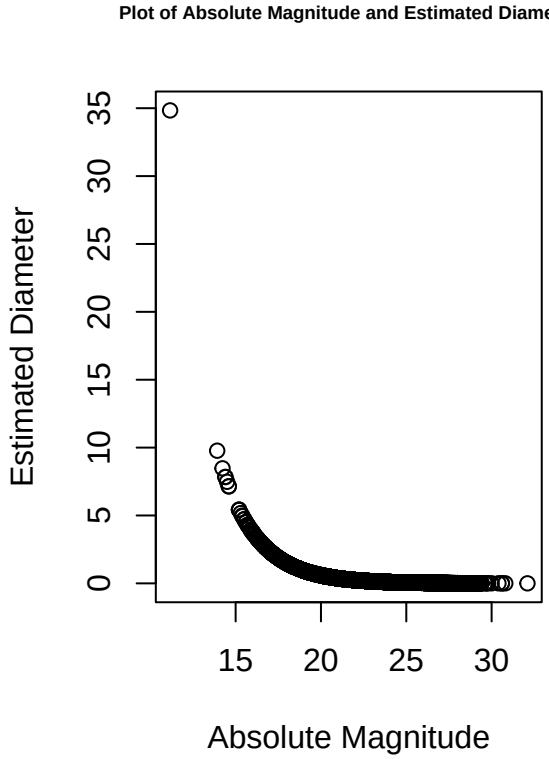


There does not seem to be clear trends, though miss distance does decrease as time goes on, magnitude increases as time goes on, and velocity decreases slightly as time goes on.

Scatterplots

We were interested in how absolute magnitude which we predicted would be an important defining feature of asteroids would correlate with eccentricity and diameter.

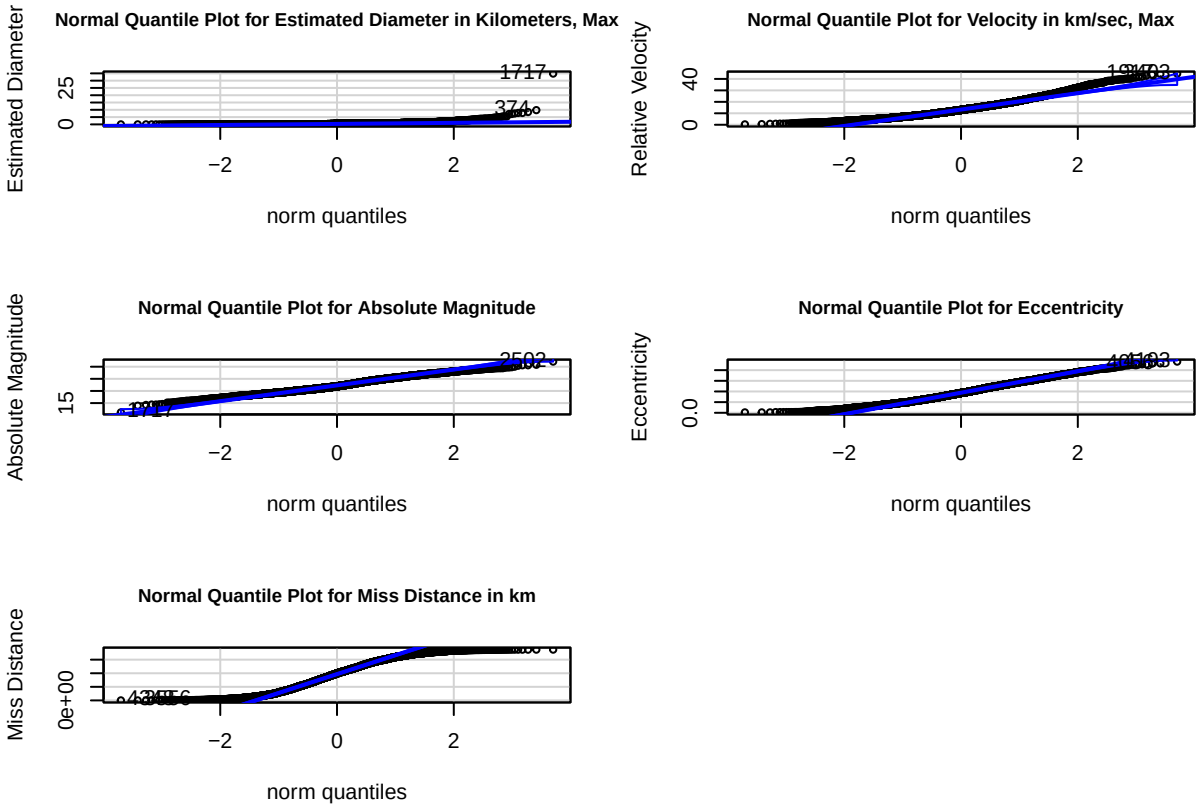




There seems to be a positive trend between absolute magnitude and estimated diameter in kilometer max but less so absolute magnitude and eccentricity.

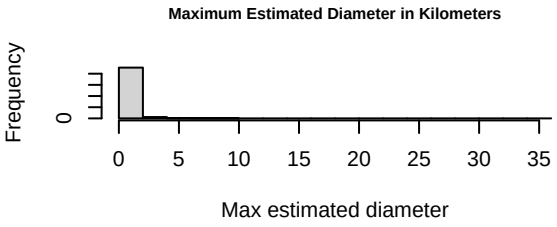
QQPlots

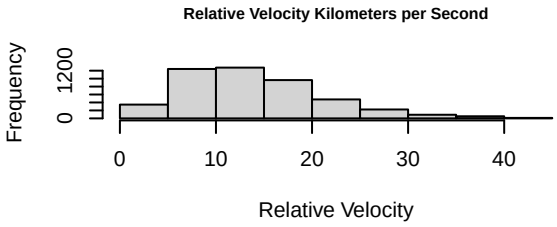
```
## [1] 1717  374
## [1] 3403 1917
## [1] 1717 2502
## [1] 4193 4056
## [1] 4348 3556
```

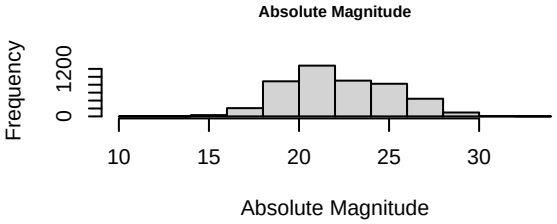


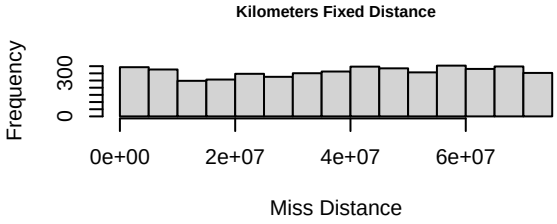
Histograms

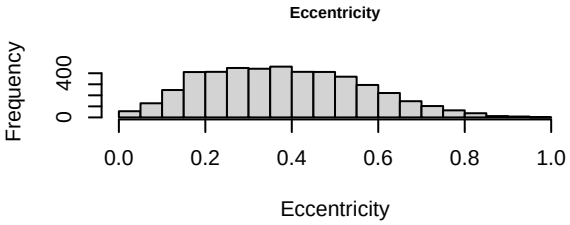
We wanted to see the overall shape of each variable- most appeared to be skewed right based on the QQPlots.







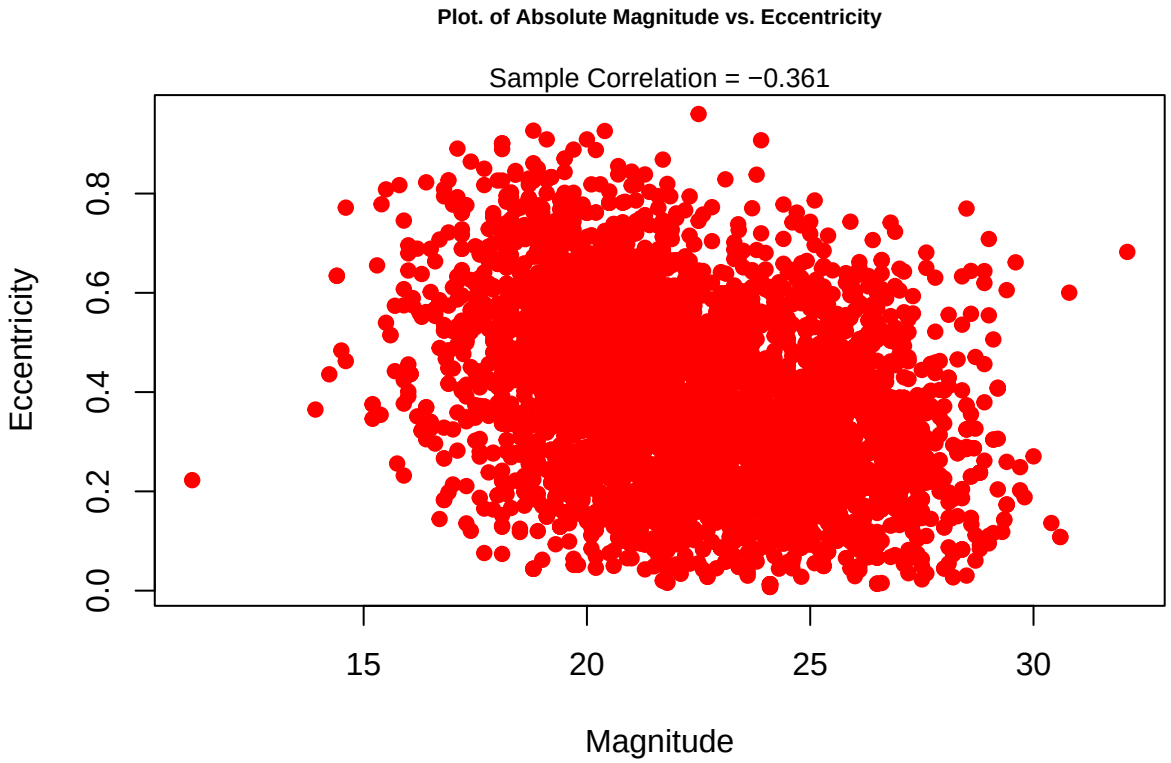


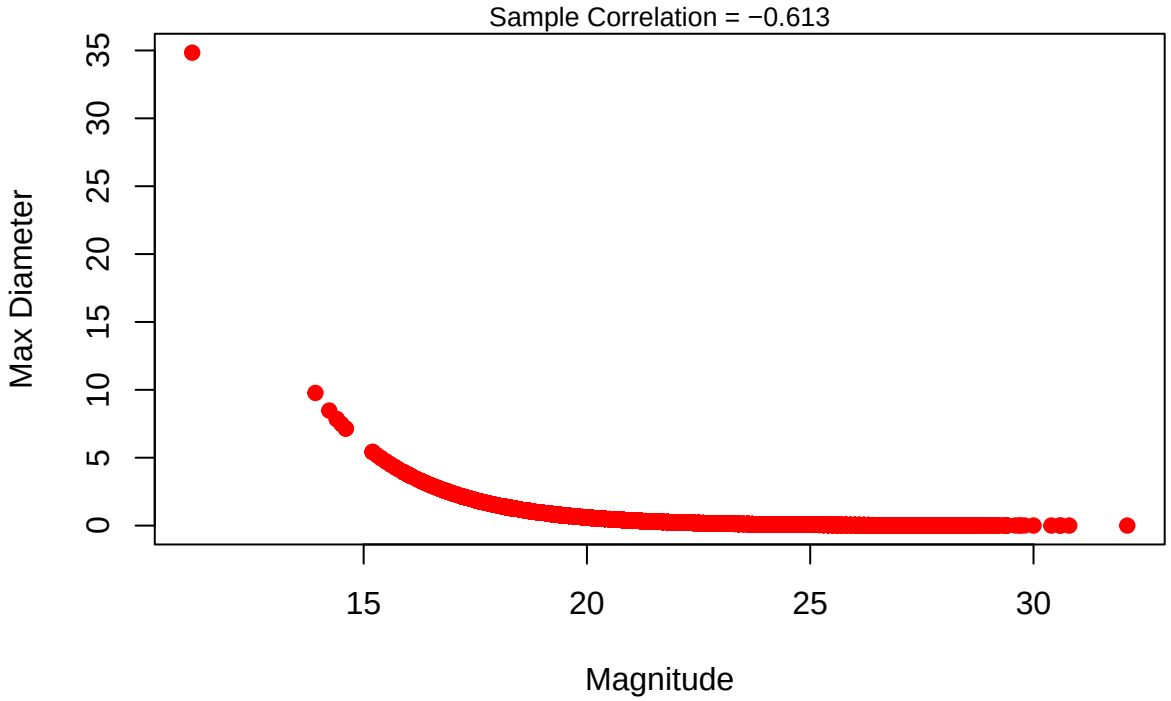


Basic Tests

Correlation

We wanted to see how Absolute Magnitude correlated with variables like eccentricity and estimated diameter and wanted to confirm our suspicions in our previous scatterplot.





The data proves us right, as there is high correlation between estimated diameter and absolute magnitude(-0.6) and lower correlation between absolute magnitude and eccentricity(-0.3). Both correlations are negative, meaning that as absolute magnitude increases, max estimated diameter and eccentricity decrease.

T Test

```
##
##  Welch Two Sample t-test
##
## data:  nasa$Absolute.Magnitude by nasa$Hazardous
## t = 37.707, df = 2345.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group False and group True is not equal to 0
## 95 percent confidence interval:
##  2.426599 2.692842
## sample estimates:
## mean in group False  mean in group True
##            22.68019            20.12047
```

Because we get a p-value of less than 0.05, we have statistically significant evidence to reject the null hypothesis that the difference in absolute magnitude between hazardous and non-hazardous asteroids is zero. Moreover, we know there is a statistically significant difference because 0 is not in the confidence interval.

Bootstrap

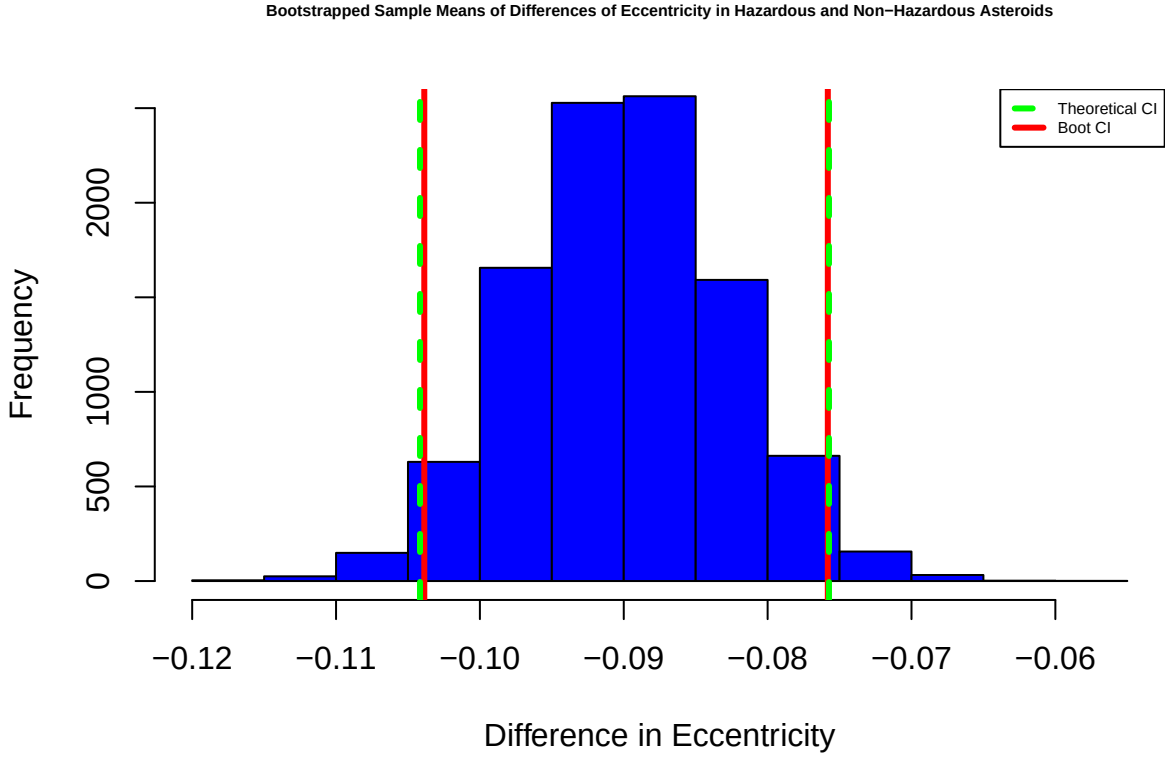
Theoretical interval for means of differences in eccentricity in hazardous and non-hazardous asteroids

```
## [1] -0.10416463 -0.07573473
## attr(,"conf.level")
## [1] 0.95
```

Bootstrap interval for means of differences in eccentricity in hazardous and non-hazardous asteroid

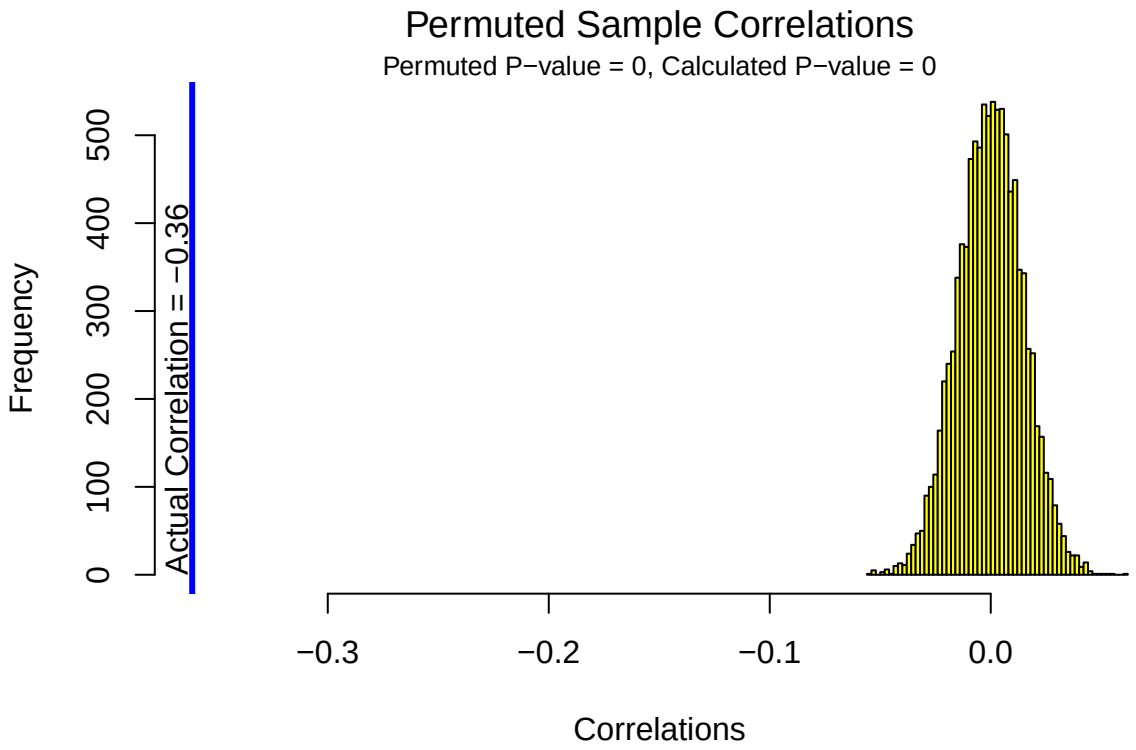
```
##        2.5%       97.5%
## -0.10386538 -0.07581717
```

```
##  2.5% 97.5%
## -0.10 -0.08
```



First, to interpret the theoretical confidence interval, 0 is not within the 95 percent confidence interval of difference in eccentricity between hazardous and non-hazardous asteroids, so we have statistically significant evidence that the difference in eccentricity between hazardous and non-hazardous asteroids is not zero. Second, the 95 percent bootstrap confidence interval gives us a similar result as 0 is not within that confidence interval. There is not much discrepancy between the bootstrapped CI and theoretical CI, though both bounds of the bootstrapped CI are slightly smaller.

Permutation

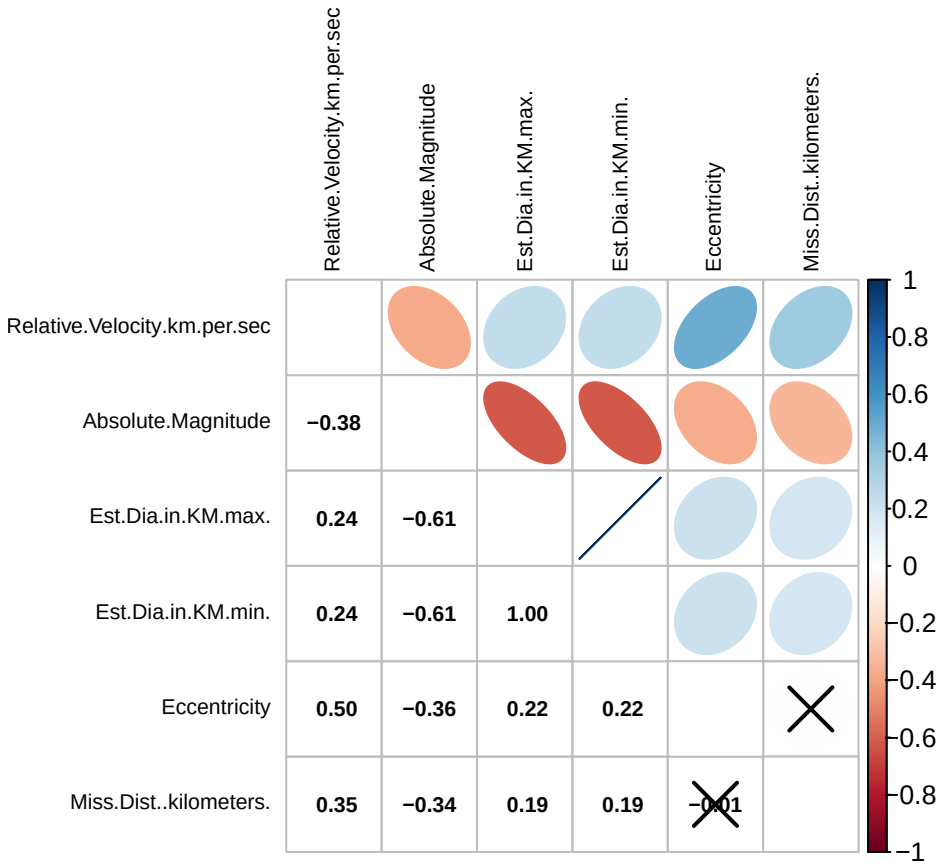


We observe a relatively weak negative correlation of -0.36. There was not a clear linear relationship between the variables (and there was no evidence of the relationship being exponential/polynomial) but we wanted to calculate the correlation to make sure. The permutation test gives a p-value of 0 so we conclude that there is a statistically significant non-zero correlation between eccentricity and absolute magnitude.

Multiple Regression

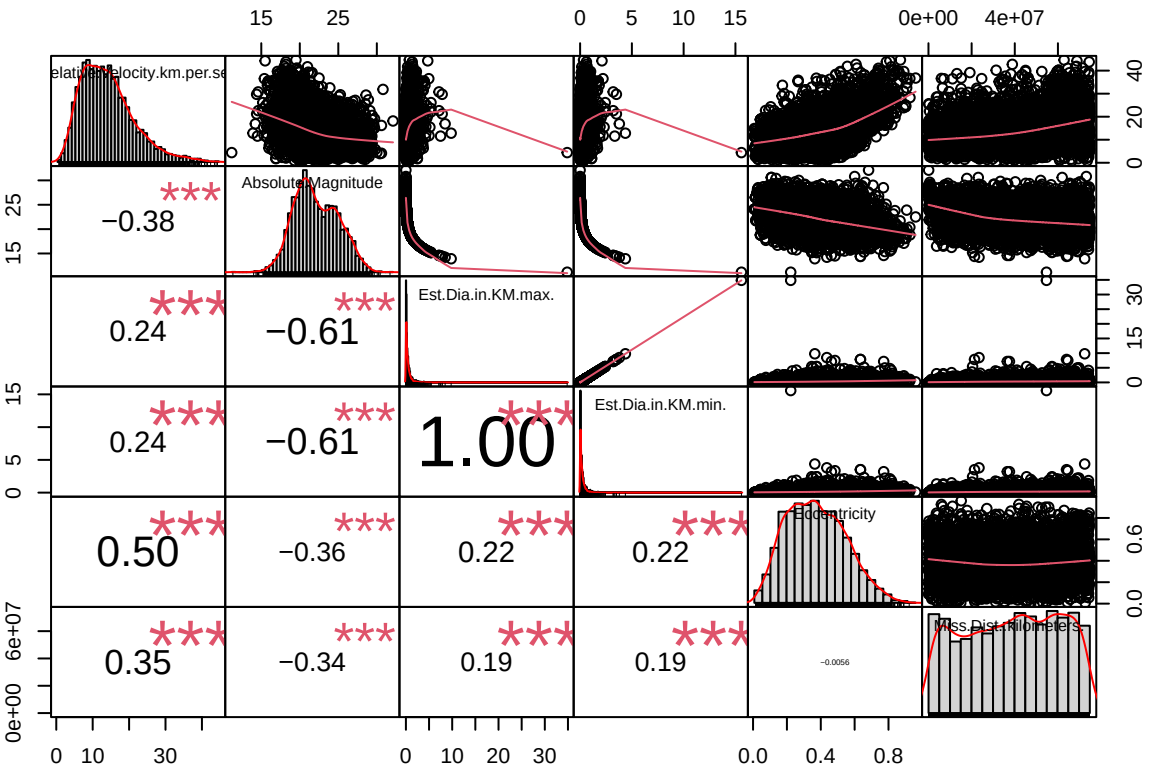
Introduction

The continuous variable we decided to predict with our multiple regression was relative velocity.. The variables we chose to use were relative velocity, estimated diameter(both maximum and minimum), eccentricity, miss distance. First, we decided to check out correlations between the variables and discovered an issue of colinearity between max and min estimated diameter, so we took min estimated diameter out of the dataset. Then, we fit a best subsets regression and used the Bayesian Information Criteria- we were especially interested in BIC because it penalizes overfitting, which is an issue we wanted to avoid. After fitting, we made residual plots, which showed some evidence of heteroskadisticity. Hence, we decided to make a boxcox transformation and refit relative velocity. While our model final is not perfectly linear, the fit vs. studentized residual plots shows less evidence of heteroskadisticity.



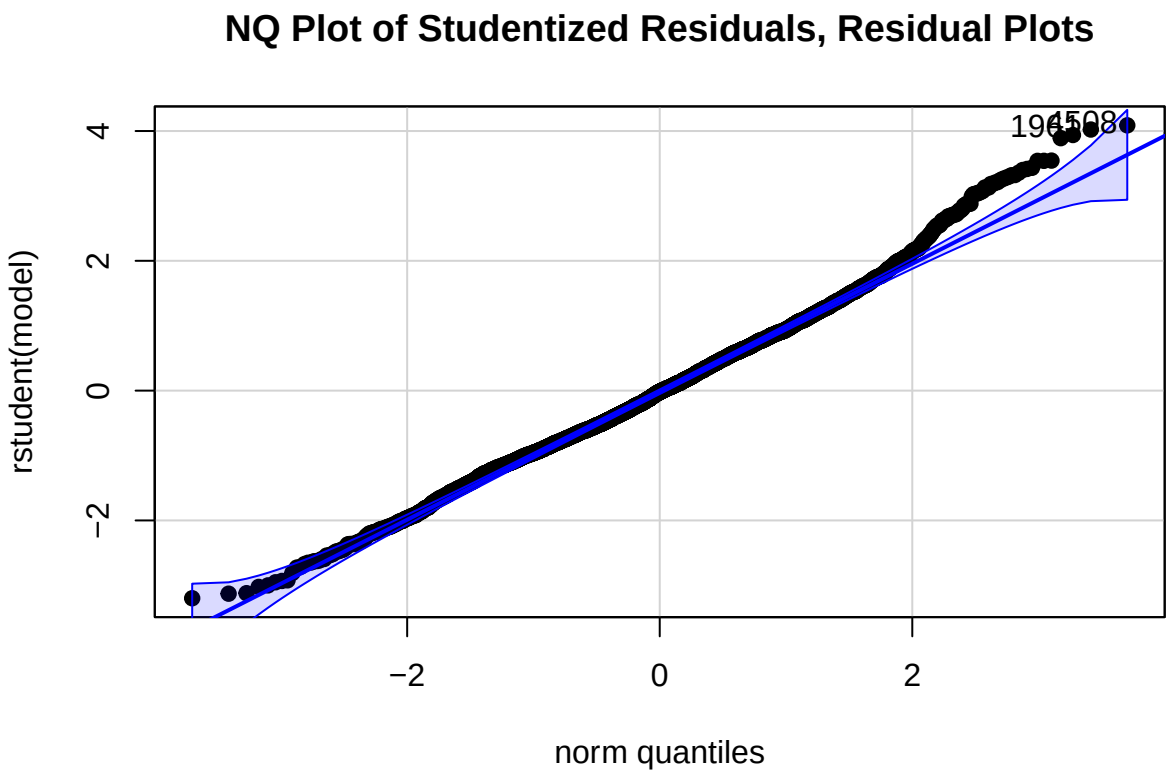

```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```

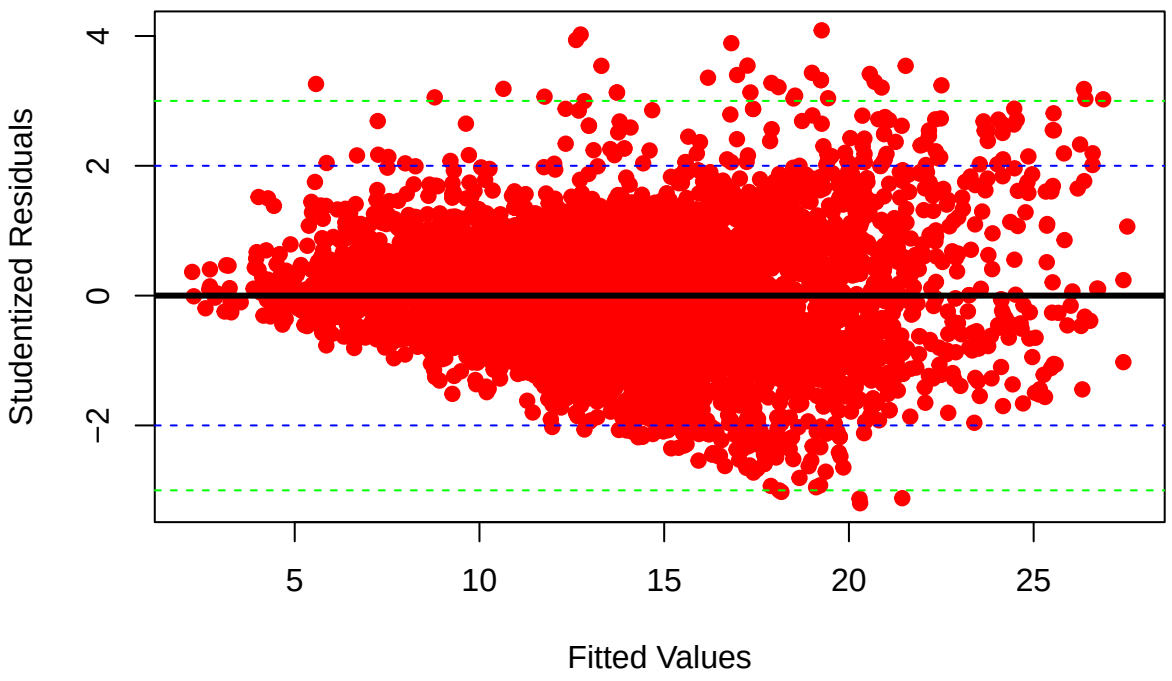
```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```



There is a clear colinearity between Est.Dia.in.KM(max) and Est.Dia.in.KM(min).

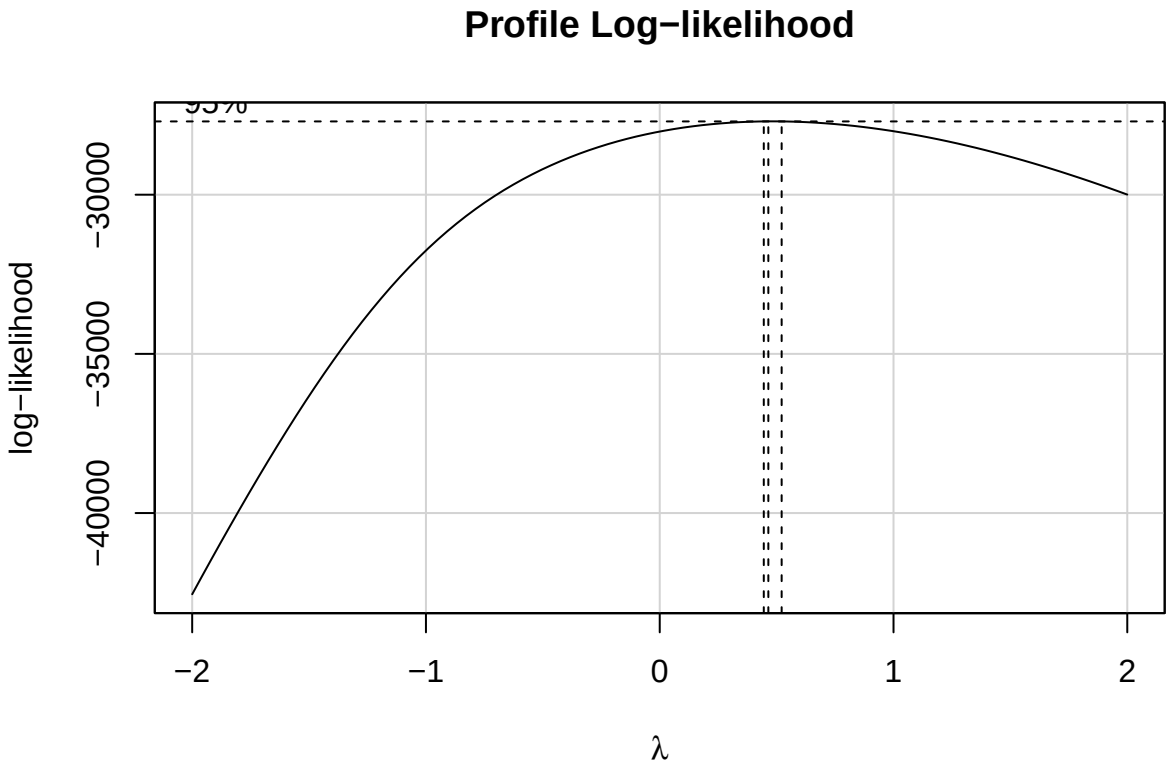
```
## [1] "Variables in best subsets model"

##   (Intercept) Absolute.Magnitude Est.Dia.in.KM.max. Eccentricity
## 1        TRUE              FALSE              FALSE         TRUE
## 2        TRUE              FALSE              FALSE         TRUE
## 3        TRUE               TRUE              FALSE         TRUE
## 4        TRUE               TRUE               TRUE         TRUE
##   Miss.Dist..kilometers.
## 1                  FALSE
## 2                   TRUE
## 3                   TRUE
## 4                   TRUE

## [1] "Variables that minimize BIC"

## [1] "Absolute.Magnitude"     "Eccentricity"           "Miss.Dist..kilometers."
```





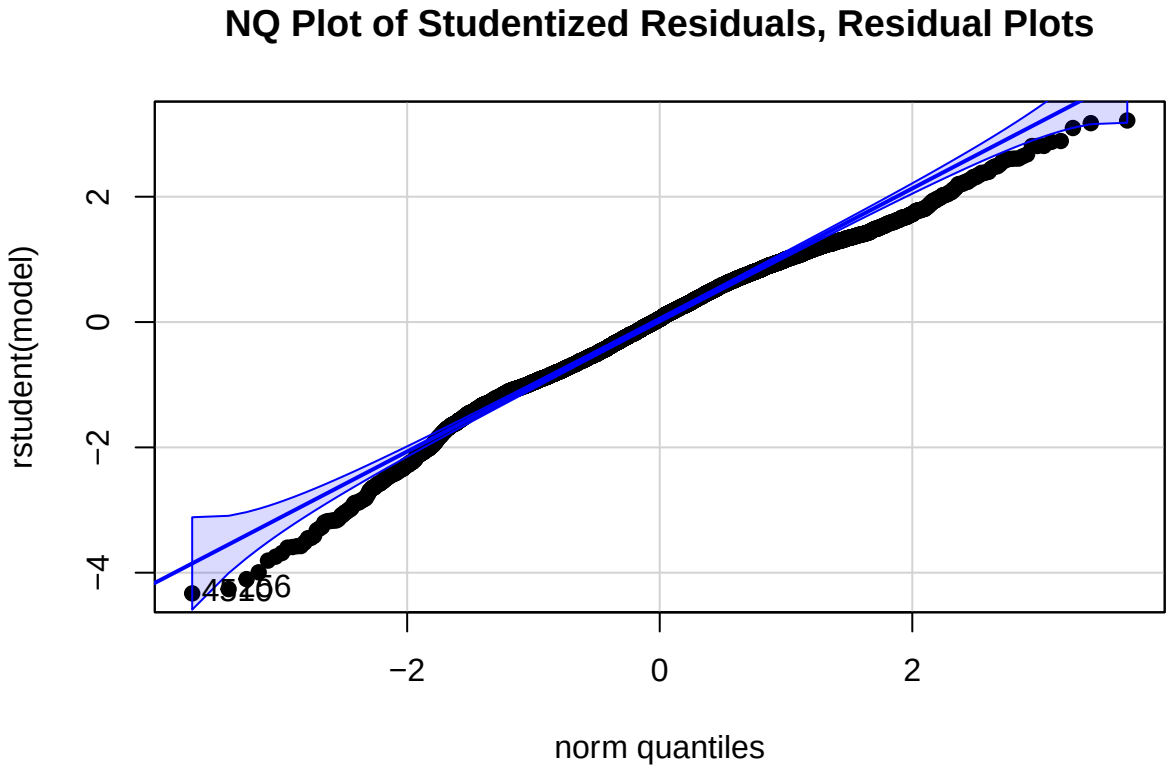
The NQ Plot is not fully linear and there is some evidence of heteroskadisticity in the Fits vs. Studentized residuals, so we will try a box-cox transformation.

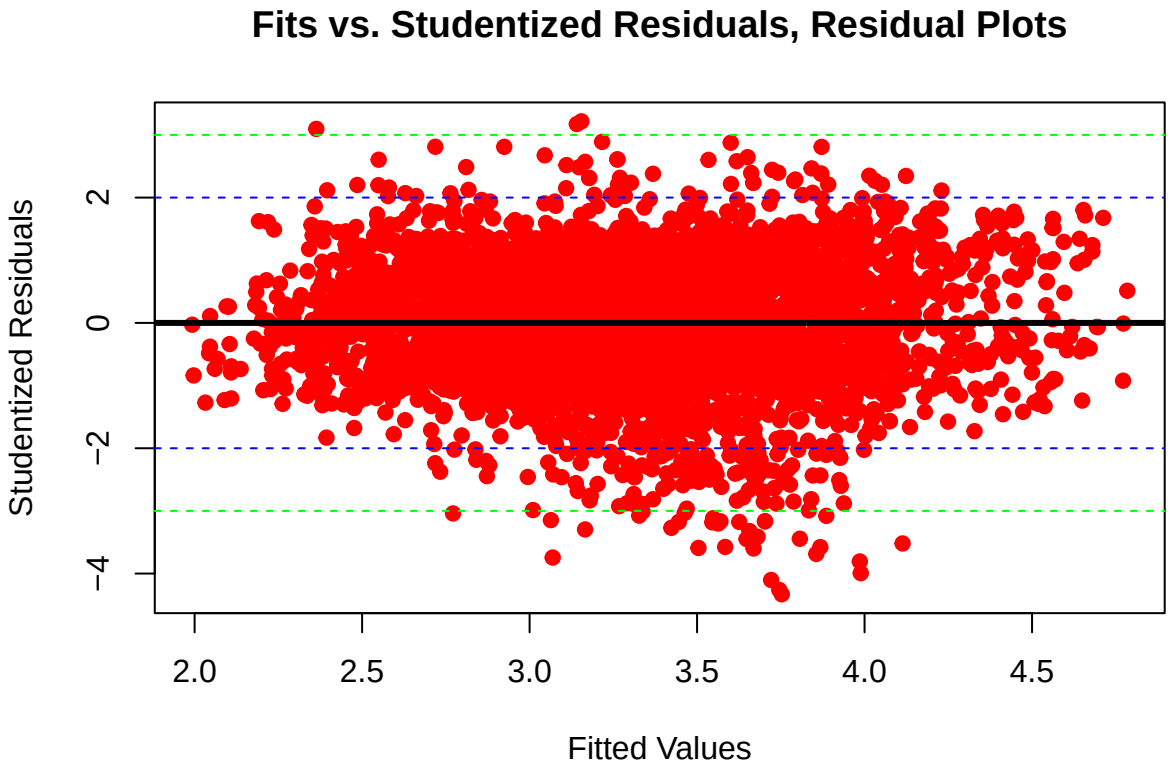
```
## [1] "BoxCox Transformation"
```




```
## [1] 0.4646465
```

```
## [1] "New, post-box cox transformation residual plots"
```





This new NQ plot is not fully linear(with curvature at lower and higher norm quantities), but there is less evidence of heteroskadisticity in the fit vs. studentized residuals plot. All in all, with the final model containing absolute magnitude, eccentricity, and miss distance in kilometers, all three factors have a p-value of lower than 0.05, meaning they are statistically significant predictors for relative velocity. Coefficients for absolute magnitude are negative, meaning it is inversely related to velocity, while those for eccentricity and miss distance are positive, meaning they are directly relatedd with velocity. Our adjusted r-squared value is 0.38, meaning 38% of the variability in velocity can be explained by our model.

Logistic Regression

We chose a logistic binary regression for Hazardous predicted by absolute magnitude.

```
##
## Call:
## glm(formula = logistichazard2 ~ Absolute.Magnitude, family = binomial,
##     data = nasa)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -6.5018     0.3849  -16.89   <2e-16 ***
## Absolute.Magnitude   0.3824     0.0186   20.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4138.3  on 4686  degrees of freedom
## Residual deviance: 3580.5  on 4685  degrees of freedom
## AIC: 3584.5
##
```

```
## Number of Fisher Scoring iterations: 5
```

```
## [1] "Odds Ratio:"
```

```
## Waiting for profiling to be done...
```

```
##                              OR      2.5 %  97.5 %         p
## (Intercept)         0.00150073 0.00069937  0.0032 < 2.2e-16 ***
## Absolute.Magnitude 1.46581456 1.41405743  1.5211 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for absolute magnitude is less than 0.05, meaning that we have statistically significant evidence that absolute magnitude is a predictor for logistichazard. The odds ratio for absolute magnitude is 1.47. indicating that the odds of an asteroid being hazardous increases for each one unit increase in absolute magnitude. 1 is not within 95 percent confidence interval, meaning we have statistical significane that the odds do increase of an asteroid being hazardous as absolute magnitude increases.

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: logistichazard2
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev
## NULL                                 4686     4138.3
## Absolute.Magnitude  1   557.78      4685     3580.5
```

```
## [1] "P-value for Deviance Test of Entire Model"
```

```
## [1] 0
```

Furthermore, the p-value for deviance test is 0.00, meaning that we can reject the null hypotehsis and have statistically significant evidence that the model has significant predictors, which is absolute magnitude.

Conclusion

In this exploration, we first identified variables could correlate with Hazardous. Then, we created boxplots that compared continuous variables of Hazardous and non-hazardous asteroids and by the close approach date. We also attempted to find correlations amongst correlations via scatterplots, correlation tests, and permutation correlation tests, though it seemed like very few were actually correlated with one another(rather than the obvious, like absolute magnitude correlating with diameter). We used t-tests to prove differences in eccentricities and magnitude by whether an asteroid is classified as 'Hazardous', and created a multiple linear regression model for velocity. Finally,we created a binary logistic regression model predicting hazard-status based on absolute magnitude. Our main conclusion for this project was that while there were not strong correlations between the continuous variables, there were clear differences between the cotninuous variables by different categories(e.g. no correlation between eccentricity and absolute magnitude, but Hazardous is a good predictor of both variables). This makes sense, as the data was originally designed to predict Hazard status, so NASA specifically tested features of the asteroids that could impact Hazard. Our project confirms this.