# Preliminary Report

Annika Lin, Hannah Norman, & Maddie Pfister

10 February 2023

## Description of Data

Our dataset is composed of quantitative metrics that NASA collects on asteroids. The original dataset contains 40 variables and 4687 observations across 3692 asteroids (i.e., 995 of the observations are on previously observed asteroids). It's worth noting that some of the variables denote the same measurement values but in different units. After getting rid of such columns as well as those for sample identifiers, variables lacking variability, and variables missing many values, we end up with 20 usable candidate predictors (see Appendix A for proposed list). Furthermore, we see the potential to add additional predictors, such as a quantitative Estimated Diameter Range variable that summarizes the range between the Estimated Min Diameter and Estimated Max Diameter variables.

From our preliminary data analysis, we believe that the data contains many small asteroids and a few large asteroids. An informal, supplemental numerical/graphical analysis of the candidate predictor for Estimated Max Diameter suggests the presence of extreme outliers (see Appendix B). We intend to remove these observations from our data so as to avoid any undue skewing within our future analyses.

The data contains a boolean variable for *whether an asteroid is considered hazardous.* This is our proposed response variable.

## Source of Data

We found this dataset through Kaggle [link]. The data is originally from the NASA JPL Asteroid team, and it was obtained through NeoWs [link] (Near Earth Object Web Service), which is a RESTful API.

## Question of Interest

*Which attributes are the best predictors as to whether an asteroid is hazardous or not?*

## Numerical Summary of Response Variable

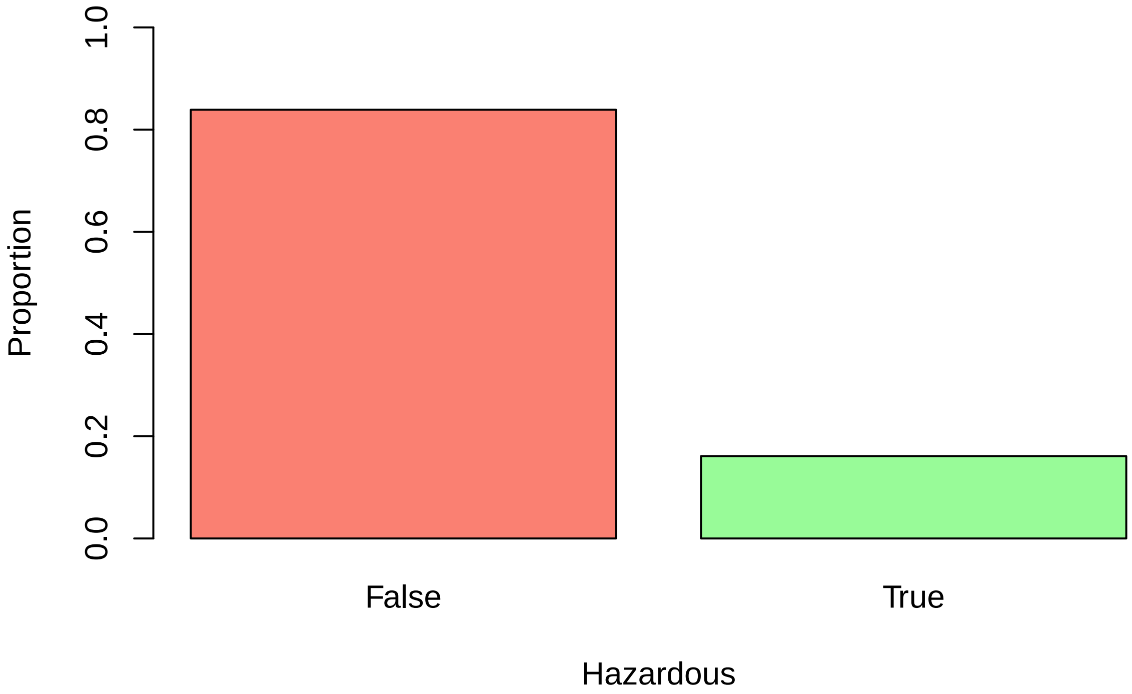
```r
nasa <- read.csv("nasa.csv")
prop.hazardous <- prop.table(table(nasa$Hazardous))
prop.hazardous
```

```
##
##     False      True
## 0.8389162 0.1610838
```

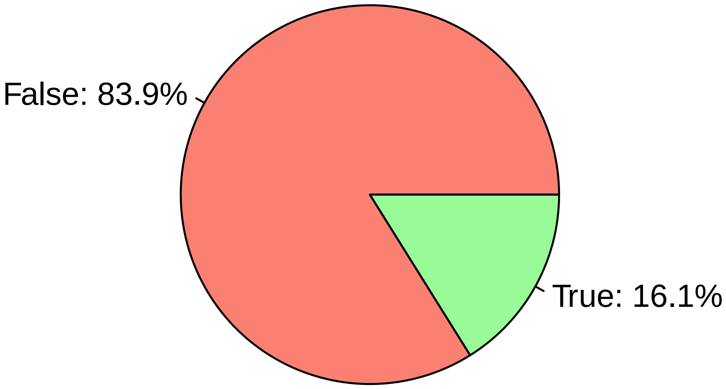
About 16.1% of asteroid observations were categorized as hazardous, whereas 83.9% were categorized as non-hazardous.

### Graphical Summaries of Response Variable

```
barplot(prop.hazardous, xlab="Hazardous", ylab="Proportion", ylim=c(0, 1.0),
        col=c("salmon", "palegreen"))
```



```
count.hazardous <- table(nasa$Hazardous)
lbls <- paste(levels(as.factor(nasa$Hazardous)), ": ",
              round(prop.hazardous,3)*100, "%", sep="")
pie(count.hazardous, labels=lbls, col=c("salmon", "palegreen"))
```



# Appendix A

List of proposed candidate predictors, excluding asteroid IDs, non-varying variables, and variables missing many values. We have yet to decide on units of measurement for those variables marked with an asterisk (*).

1. Absolute Magnitude
2. Estimated Diameter (min)*
3. Estimated Diameter (max)*
4. Relative Velocity*
5. Miss Distance*
6. Orbit ID
7. Orbit Uncertainty
8. Minimum Orbit Intersection
9. Jupiter Tisserand Invariant
10. Eccentricity
11. Semi Major Axis
12. Inclination
13. Ascending Node Longitude
14. Orbital Period
15. Perihelion Distance
16. Perihelion Argument
17. Aphelion Distance
18. Perihelion Time
19. Mean Anomaly
20. Mean Motion

# Appendix B

Numerical and graphical summaries of Estimated Max Diameter variable illustrating the presence of extreme outliers that risk skewing future analyses if not removed.

```
summary(nasa$Est.Dia.in.Miles.max.)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##  0.001404  0.046493  0.153954  0.284283  0.352688 21.646663
```

```
boxplot(nasa$Est.Dia.in.Miles.max.)
```