An Explorotory Analysis of Asteroid Impacts

Jonathan Shea, Albert Magpoc, Mina Fahim

March 26th, 2018

## Introduction

This analysis uses data collected by NASA on potential asteroid impacts with earth, and orbital properties of those asteroids. It was collected using the Sentry automated collision monitoring system managed by NASA, which scans the known asteroids to determine their probability and threat of impact with earth. The data contains information on what asteroids have potential impacts with earth, and how dangerous these impacts would be. The orbits data set has information relating to the orbital characteristics of specific asteroids. The data was downloaded from Kagle.com at <https://www.kaggle.com/nasa/asteroid-impacts>.

This analysis is designed to show the impact characteristics of the different classifications of asteroids, and how these impacts change over the next 100 years.

1. Which class of asteroid poses the greatest risk?
2. What is the distance to earth of dangerous v. non-dangerous asteroids?
3. How do the number of potential impacts change over time?

…

## 1. Data Import and Cleaning

#### Data Importing

Data was imported using the built in csv reading function.

library(tidyverse)

## ── Attaching packages ───────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 2.2.1 ✔ purrr 0.2.4  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.4  
## ✔ tidyr 0.8.0 ✔ stringr 1.2.0  
## ✔ readr 1.1.1 ✔ forcats 0.2.0

## Warning: package 'ggplot2' was built under R version 3.3.2

## Warning: package 'readr' was built under R version 3.3.2

## Warning: package 'purrr' was built under R version 3.3.2

## Warning: package 'dplyr' was built under R version 3.3.2

## Warning: package 'stringr' was built under R version 3.3.2

## Warning: package 'forcats' was built under R version 3.3.2

## ── Conflicts ──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

impacts <- read.csv("impacts.csv")  
orbits <- read.csv("orbits.csv", stringsAsFactors = FALSE)  
  
head(impacts, 1)

## Object.Name Period.Start Period.End Possible.Impacts  
## 1 2006 WP1 2017 2017 1  
## Cumulative.Impact.Probability Asteroid.Velocity Asteroid.Magnitude  
## 1 5.2e-09 17.77 28.3  
## Asteroid.Diameter..km. Cumulative.Palermo.Scale Maximum.Palermo.Scale  
## 1 0.007 -8.31 -8.31  
## Maximum.Torino.Scale  
## 1 0

head(orbits, 1)

## Object.Name Object.Classification Epoch..TDB. Orbit.Axis..AU.  
## 1 433 Eros Amor Asteroid 57800 1.4579  
## Orbit.Eccentricity Orbit.Inclination..deg. Perihelion.Argument..deg.  
## 1 0.2226 10.8277 178.805  
## Node.Longitude..deg. Mean.Anomoly..deg. Perihelion.Distance..AU.  
## 1 304.3265 319.3111 1.1335  
## Aphelion.Distance..AU. Orbital.Period..yr.  
## 1 1.78 1.76  
## Minimum.Orbit.Intersection.Distance..AU. Orbital.Reference  
## 1 0.1492 598  
## Asteroid.Magnitude  
## 1 11.16

It was noticed that the Maximum Torino Scale value for all asteroid impacts was 0. After some research, it was determined that this was in fact not a mistake, and that the torino scale is meant to predict serious threats to humans, and thus far there have been no serious asteroid impacts discovered. The Palermo scale is more sensitive to lesser threats, and thus has a wide variety of values in the data set.

One of the first steps we took to make working with the data easier was to change the column names to be shorter and easier to type. The magnitude parameter in the orbits data set was interpreted as a string, so we converted that to a numeric type using mutate().

impacts <- rename(  
 impacts,  
   
 name = Object.Name,  
 start = Period.Start,  
 end = Period.End,  
 impact = Possible.Impacts,  
 cum\_impact = Cumulative.Impact.Probability,  
 velocity = Asteroid.Velocity,  
 magnitude = Asteroid.Magnitude,  
 diameter = Asteroid.Diameter..km.,  
 cum\_pal = Cumulative.Palermo.Scale,  
 max\_pal = Maximum.Palermo.Scale,  
 max\_tor = Maximum.Torino.Scale  
 )

orbits <- orbits%>%  
 rename(  
 name = Object.Name,  
 class = Object.Classification,  
 epoch = Epoch..TDB.,  
 axis = Orbit.Axis..AU.,  
 eccen = Orbit.Eccentricity,  
 inc = Orbit.Inclination..deg.,  
 ph\_arg = Perihelion.Argument..deg.,  
 long = Node.Longitude..deg.,  
 mean = Mean.Anomoly..deg.,  
 ph\_dist = Perihelion.Distance..AU.,  
 ah\_dist = Aphelion.Distance..AU.,  
 dist = Orbital.Period..yr.,  
 min\_int\_dist = Minimum.Orbit.Intersection.Distance..AU.,  
 ref = Orbital.Reference,  
 mag = Asteroid.Magnitude  
 )  
  
orbits <- suppressWarnings(orbits %>% mutate(mag=as.numeric(mag)))

The two data sets had slightly varying names. The name column for the orbit data set sometimes contained a insignificant number, with the true name in parenthesis. To clean this, we split the data set into two sets, for whether or not it has this useless number in the name.

The name also contains the year of observation in it. We split this into a seperate column using separate, and then merge orbits and impacts into one asteroids dataset, and drop the null rows. Another dataset, containing all asteroid objects, was created, We replaced the missing impact column with 0 for asteroids with no potential impact.

Finally, any entries without a class were removed since that is the basis of our analysis.

orbits\_no\_num <- filter(orbits, !grepl("\\(", name)) %>% separate(name, c("year", "name"))

## Warning: Expected 2 pieces. Additional pieces discarded in 3 rows [26, 49,  
## 60].

orbits\_num <- filter(orbits, grepl("\\(", name)) %>% separate(name, c("number", "name"), "\\(") %>% mutate(name = gsub("\\)", "", name)) %>% select(name:mag) %>% separate(name, c("year", "name"))

## Warning: Expected 2 pieces. Additional pieces discarded in 2 rows [413,  
## 15481].

orbits\_named <- rbind(orbits\_no\_num, orbits\_num)  
impacts\_mod <- impacts %>% separate(name, c("year", "name"))

## Warning: Expected 2 pieces. Additional pieces discarded in 5 rows [331,  
## 607, 681, 682, 683].

asteroids\_raw <- full\_join(impacts\_mod, orbits\_named)

## Joining, by = c("year", "name")

asteroid\_impacts <- asteroids\_raw %>% drop\_na()  
  
asteroids <- asteroids\_raw %>% mutate(impact = ifelse(is.na(impact), 0, impact)) %>% filter(!is.na(class))  
head(asteroids, 1)

## year name start end impact cum\_impact velocity magnitude diameter  
## 1 2006 WP1 2017 2017 1 5.2e-09 17.77 28.3 0.007  
## cum\_pal max\_pal max\_tor class epoch axis eccen inc  
## 1 -8.31 -8.31 0 Apollo Asteroid 54058 1.7082 0.6071 5.8995  
## ph\_arg long mean ph\_dist ah\_dist dist min\_int\_dist ref mag  
## 1 98.1906 234.2977 22.0794 0.6711 2.75 2.23 7e-04 5 28.2

head(asteroid\_impacts, 1)

## year name start end impact cum\_impact velocity magnitude diameter  
## 1 2006 WP1 2017 2017 1 5.2e-09 17.77 28.3 0.007  
## cum\_pal max\_pal max\_tor class epoch axis eccen inc  
## 1 -8.31 -8.31 0 Apollo Asteroid 54058 1.7082 0.6071 5.8995  
## ph\_arg long mean ph\_dist ah\_dist dist min\_int\_dist ref mag  
## 1 98.1906 234.2977 22.0794 0.6711 2.75 2.23 7e-04 5 28.2

## 2. Data Analysis

The only major data transformation that was required was to determine the approximate number of potential impacts per year. The given data only provides the starting and ending year, and how many impacts may occur over that timeframe. To better visualize the number of impacts over time, we split the rows into additional rows with key-value pairs of year:approx\_number\_of\_impacts. The approximate number of impacts was calculated by impacts / (end-start)+1.

This was accomplished by: 1. Creating a new column comprised of the sequence of years from start year to end year. 2. Using unnest() to split each individual year into another row 3. Creating a new column to represent the approximate number of impacts to occur in that specific year

expand\_years <- function(start, end){  
 result <- paste(seq(start, end), collapse=',')  
 return(result)  
}  
yearly\_impacts <- asteroid\_impacts %>% rowwise() %>% mutate(year\_key = expand\_years(start, end)) %>% mutate(year\_key = strsplit(year\_key, ",")) %>% unnest(year\_key) %>% mutate(impacts\_per\_year = impact / ((end-start)+1)) %>% mutate(year\_key = as.numeric(year\_key))  
head(yearly\_impacts, 5)

## # A tibble: 5 x 28  
## year name start end impact cum\_impact velocity magnitude diameter  
## <chr> <chr> <int> <int> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 2006 WP1 2017 2017 1 0.00000000520 17.8 28.3 0.00700  
## 2 2013 YB 2017 2046 23 0.0000760 8.98 31.4 0.00200  
## 3 2013 YB 2017 2046 23 0.0000760 8.98 31.4 0.00200  
## 4 2013 YB 2017 2046 23 0.0000760 8.98 31.4 0.00200  
## 5 2013 YB 2017 2046 23 0.0000760 8.98 31.4 0.00200  
## # ... with 19 more variables: cum\_pal <dbl>, max\_pal <dbl>, max\_tor <fct>,  
## # class <chr>, epoch <int>, axis <dbl>, eccen <dbl>, inc <dbl>,  
## # ph\_arg <dbl>, long <dbl>, mean <dbl>, ph\_dist <dbl>, ah\_dist <dbl>,  
## # dist <dbl>, min\_int\_dist <dbl>, ref <int>, mag <dbl>, year\_key <dbl>,  
## # impacts\_per\_year <dbl>

### Question 1: Which class of asteroid poses the greatest risk?

The data transformation for this question revolved around shifting the data to positive values so it could be graphed using a logarithmic-scale axis. The maximum palermo scale value is negative for some entries, so we added a constant value to each row so that it was always positive.

This question was addressed using violin plots to show the distribution of impacts and risk for each class of asteroid. We can see most of the asteroid classes follow an hourglass shape for the number of impacts, however the Apollo and Amor asteroid is heavily comprised of low-number of impact asteroids, while the Apollo Hazard and Aten asteroids have a large number of high-impact asteroids. The palermo scale for these classes shows that all classes except the Apollo hazard asteroids are centered around the same value, with the Amor asteroids having a heavier distribution towards lower palermo values. The Apollo Hazard asteroids on the other hand are almost all towards the highest end of the values.

From this, we can clearly see that the Apollo Hazard asteroids are the most commonly dangerous asteroids.

asteroids <- asteroids %>% mutate(max\_pal = max\_pal+11.01)  
asteroids %>% group\_by(class) %>% filter(n() > 3) %>% ungroup() %>% ggplot() + geom\_violin(mapping=aes(x=class, y=impact, fill=class)) + scale\_y\_log10() + theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + labs(title="Distribution of Number of Impacts", x="Asteroid Class", y="Potential Impacts")

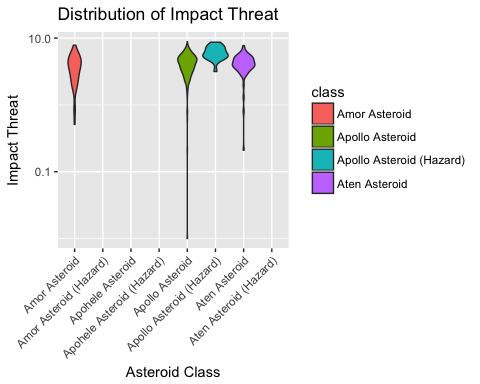
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 14956 rows containing non-finite values (stat\_ydensity).



asteroids %>% group\_by(class) %>% filter(n() > 3) %>% ungroup() %>% ggplot() + geom\_violin(mapping=aes(x=class, y=max\_pal, fill=class)) + scale\_y\_log10() + theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + labs(title="Distribution of Impact Threat", x="Asteroid Class", y="Impact Threat")

## Warning: Removed 14956 rows containing non-finite values (stat\_ydensity).



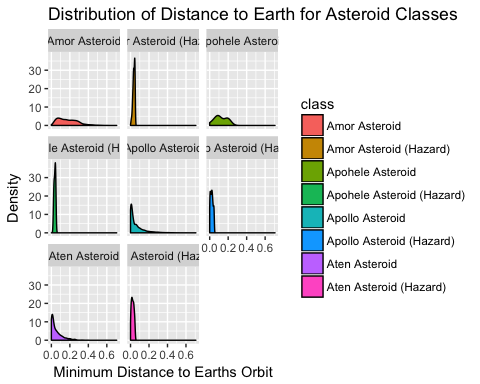
### Question 2: What is the distance to earth of dangerous v. non-dangerous asteroids?

This question required no additional data transformation.

The distribution of the minimal orbit intersection distance, or how close the asteroid gets to our orbit, can shed insight into how the asteroids are classified.

From the density plots, we can tell that the asteroids classified as hazardous must have to fall within a specific distance of our orbit, as all the asteroids are below a given value. This does not mean that all close asteroids are determined to be hazardous however, as the other classes have asteroids within that range.

asteroids %>% ggplot() + geom\_density(mapping=aes(x=min\_int\_dist, fill=class), alpha=1) + facet\_wrap(~class) + labs(title="Distribution of Distance to Earth for Asteroid Classes", x="Minimum Distance to Earths Orbit", y="Density")



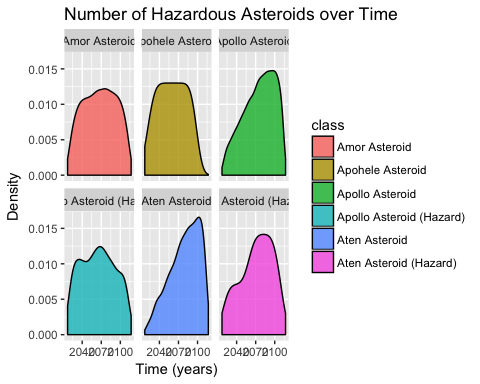
### Question 3: How do the number of potential impacts change over time?

This question required a large deal of data transformation which was covered in Part 2: Data Analysis.

For this question, we looked at how the number of hazardous asteroids changed over time for each class, as well as approxamitely how many impacts each asteroid class was expected to have over time.

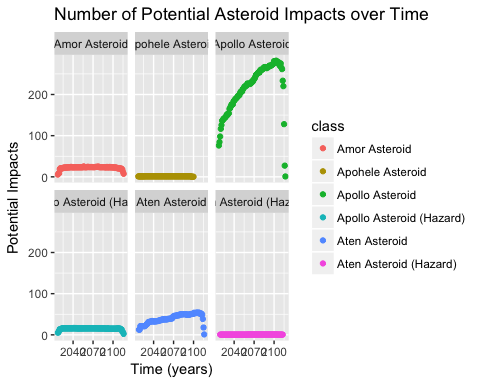
Based on the distribution of hazardous asteroids for each class over time, we see that the Aten and Apollo asteroids grow over time, whereas the other classes follow a more normal distribution. The normal distribution would be expected due to the nature of having a fixed minimum and maximum year for analysis.

yearly\_impacts %>% filter(year\_key < 2150) %>% ggplot() + geom\_density(mapping=aes(x=year\_key, fill=class), alpha=0.8) + facet\_wrap(~class) + labs(title="Number of Hazardous Asteroids over Time", x="Time (years)", y="Density")



When looking at how the number of actual number of potential impacts over time, we see that again the Apollo and Aten asteroids gradually increase over time, while the other asteroid classes remain at a fixed value over time.

sample <- yearly\_impacts %>% group\_by(year\_key, class) %>% summarise(total\_impacts = sum(impacts\_per\_year))  
sample %>% filter(year\_key < 2150) %>% ggplot() + geom\_point(mapping=aes(x=year\_key, y=total\_impacts, color=class)) + facet\_wrap(~class) + labs(title="Number of Potential Asteroid Impacts over Time", x="Time (years)", y="Potential Impacts")



…

## 3. Summary

The main things learned from this analysis is that asteroids in the Apollo class are the most dangerous to earth, however thus far there have not been any significant threats discovered. The number of potential impacts is increasing however, but only for the classes with low Palermo scales, which means it is likely not a threat. The biggest limitation of this analysis is that the number of impacts per year was estimated based on the provided data, and may differ from the simulations run at NASA. This means that the number of impacts over time may not be completely accurate. Data from prior to 2017 would be beneficial for analyzing the distribution of hazardous asteroids over time, to better determine whether the number plateus at a certain point or continues to increase.

## References

Possible Asteroid Impacts with Earth - NASA <https://www.kaggle.com/nasa/asteroid-impacts>