# Introduction

This report is an exploratory investigation into an *astronauts* dataset, about astronauts and their missions, and a *missions* dataset, which contains data about missions and the rockets involved. The data in csv form and a text file provided to contextualize the columns are in this report’s folder. The purpose of this report is to inform a new space agency of any initial insights gleaned from this data. Unless stated otherwise, I assume high data quality (i.e. good Completeness, Consistency, Conformity, Accuracy, Integrity and Timeliness). I conducted this analysis in a Jupyter Notebook, also in this report’s folder, using the *pandas*, *numpy,* *seaborn* and *matplotlib* libraries.

# Analysis

## Missions Data

In my opinion, for a new space agency looking to develop a business case, the two main metrics of interest in the *missions* dataset are *Cost* and *Status Mission.* It is important that missions have the least risk of failure, whilst also minimizing price. This data also includes *Company Name*, and so can provide meaningful insights into which companies can be used as a case study when considering the business case or a market segment the new space agency is interested in.

When analysing *Status Mission*, I have assumed that *Success/Failure* are binary by assigning *Success 1, Failure* and *Prelaunch Failure 0*, and *Partial Failure 0.5*. Clearly, this is a large assumption as missions can certainly be more successful/fail harder than others. Binary success data also obscures relationships with other variables.

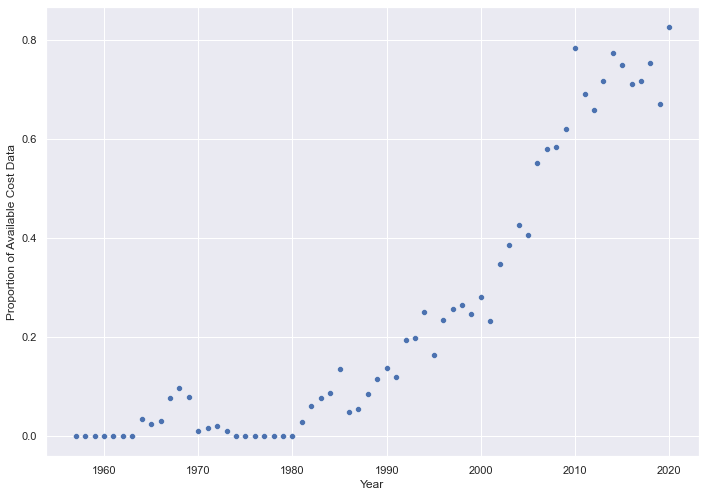
An issue with our data is that out of *4324* rows, we are missing *3360* *Cost* values. **Figure 1** shows that *Cost* data is clearly missing non-randomly, so removing *NaN*s will lead to bias, which we should remember in our analysis. For example, before *2005*, most *Cost* data was missing. Our analysis assumes that costs have been adjusted for inflation/purchasing power parity.

Figure : Proportion of Available Cost Data by Year

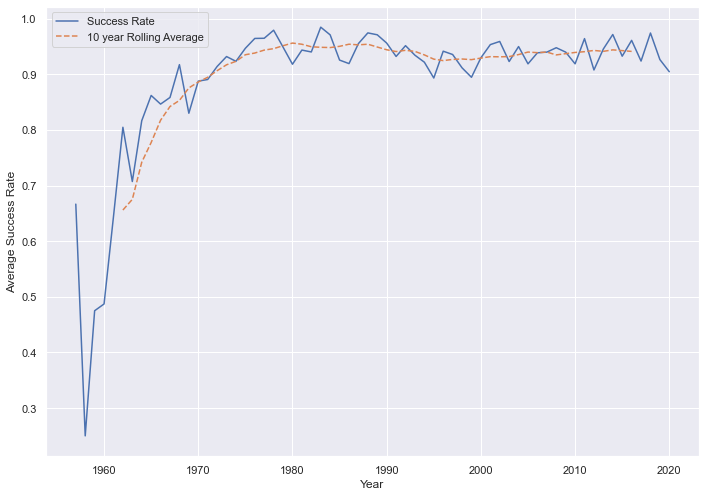
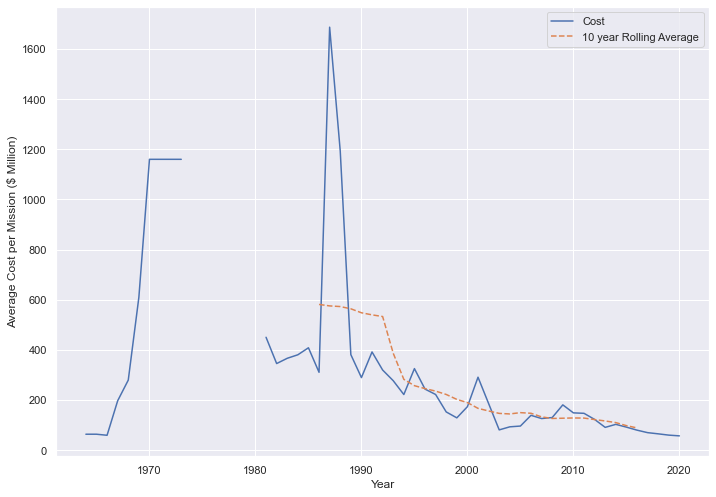


Figure : Average Cost per Mission ($ Million) and Average Success Rate by Year; Year by Year and 10 Year Rolling Average

**Figure 2** shows that average cost per mission has been greatly decreasing, the average decreasing from *$591.1M* in *1964-1973*, to *$88.3M* in *2011-2020*. The success rate of missions has increased greatly, the average increasing from *65.6%* in *1957-1966*, to *94.1%* in *2011-2020*. Clearly, there have been great advancements in space technology over the years.

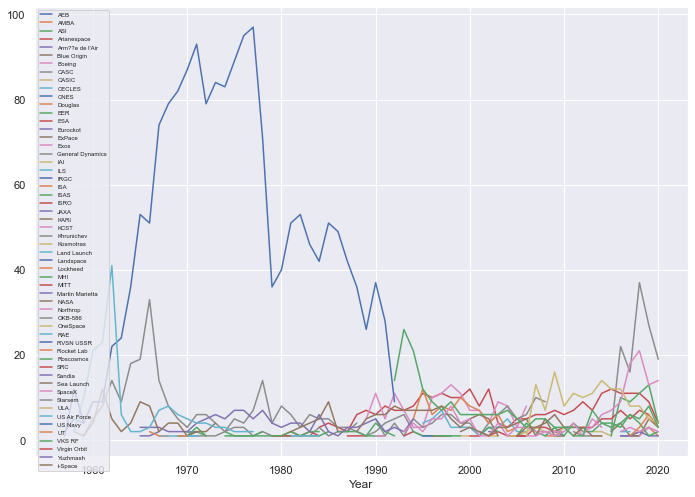


Figure : Count of Mission Undertaken by Year, Grouped by Company

**Figure 3** shows that the companies in the industry have greatly changed over years, with *RVSN USSR* being responsible for the bulk of the missions before *1990*, with *97* missions at its peak in *1977*, running its last mission in *1998*. In *2020* however, *SpaceX* ran the most missions, at *14*.

## Missions Data: Active Rockets

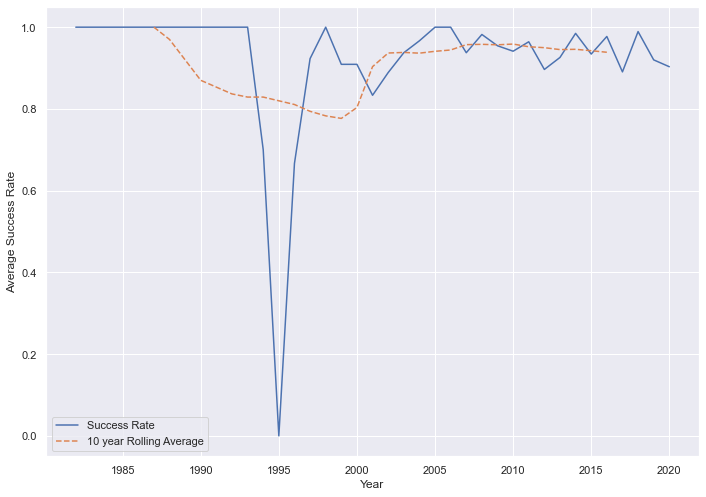
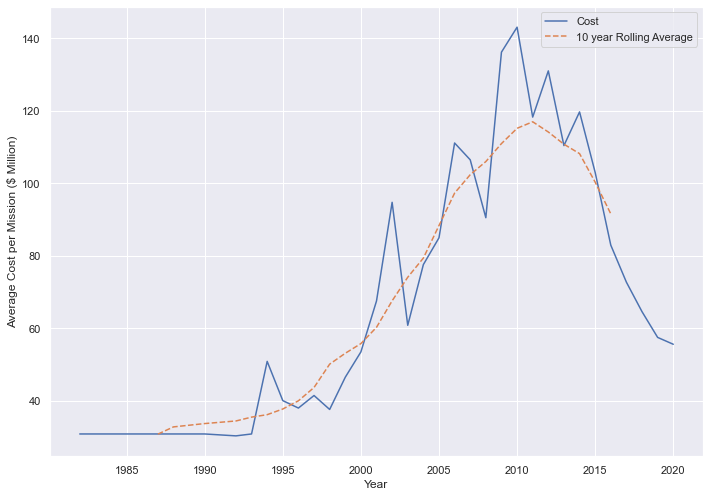
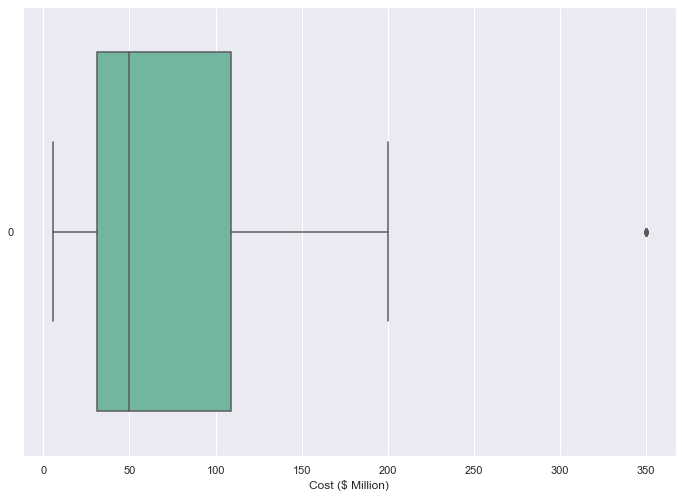


Figure : Average Cost per Mission ($ Million) and Average Success Rate by Year, Out of Active Rockets; Year by Year and 10 Year Rolling Average

Because of technological advancements and changes to the companies in the market, for analysis purposes it may be more meaningful to consider only companies with active rockets, as they are would-be competitors and so have comparable costs/success rates. For the rest of our *missions* data analysis, we consider missions with column *Status Rocket* equal to *StatusActive* only.

**Figure 4** shows the rolling average of average cost of active rockets peaked in *2011* at *$117.0M* and now is declining. The average cost in *2020* is *$55.6M*. Average success rates have largely remained stable, the average being *94.0%.*

**Figure 5** investigates for *Cost* outliers. There appears to be one possible outlier of *350*. Upon further investigation, there are *11* missions with active rockets with a cost of *350*, *10* of which are performed by company ULA, so it is unlikely these values are outliers/anomalous, and so I decided not to remove them.

Figure : Boxplot of Cost With Outliers Displayed

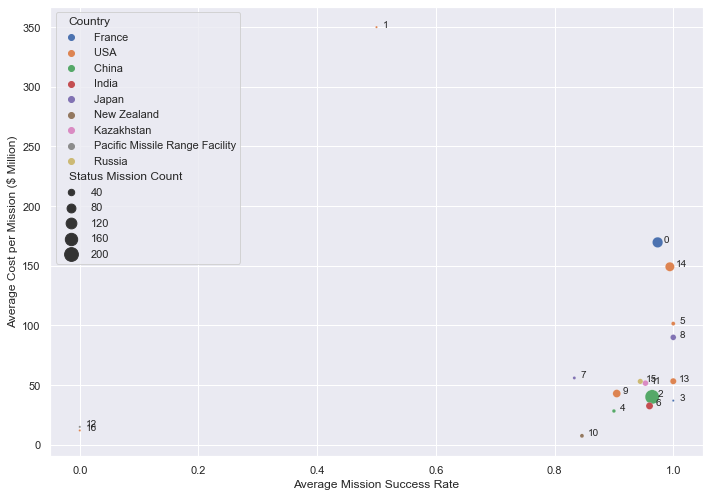


Figure : Average Cost per Mission by Average Success Rate by Company Name; Companies with Active Rockets and Non-Zero Cost Values; Size Denotes Number of Missions; Colour Denotes Country

**Figure 6** can be used to help us identify interesting companies that could be investigated further, with the bottom right of the graph being the ideal. Companies *12* and *16* (*Virgin Orbit* and *Sandia*) both have had one failed mission – perhaps they would provide an interesting case study on a method/technology that should be avoided. They also spent lower than average, *$12.0M* and *$15.0M* respectively – far below the average cost spent in 2020. Company *10* (*Rocket Lab)* on the other hand has even lower costs – averaging *$7.5M* per mission – but a success rate of *84.6%*. If the new space agency is looking to compete in the low-cost market and are willing to trade success rate, then perhaps this is a company they can investigate.

In contrast, companies *0* and *14* (*Arianespace* and *ULA*) are experienced competitors, with *114* and *87* missions respectively, and with high success rates of *97.4%* and *99.4%.* However, they also spent far above even the peak average cost at *$169.6M* and *$149.2M*. If a higher success rate is what our new space agency is aiming for, investigating these companies further may provide insights.

## Astronauts Data

Given that training and selection of astronauts are likely a large part of space agency costs, it is likely that it is preferable for astronauts to go on multiple missions, with a high number of hours spent on each mission. Therefore, I will investigate *Hours\_Mission*, *Eva\_Mission* as well as astronauts’ traits, including age at selection, numbers years spent between selection and mission (which I’ve defined as experience), and whether they are military or civilian, and how these have changed over time. These insights can help guide the creation of the selection process, and the training process.

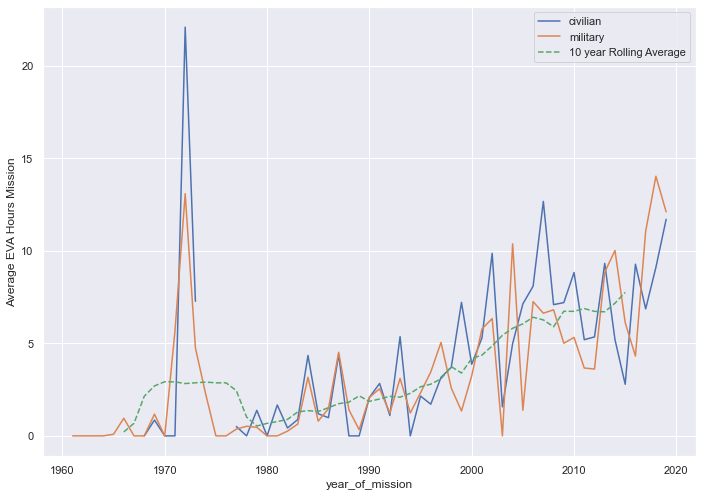
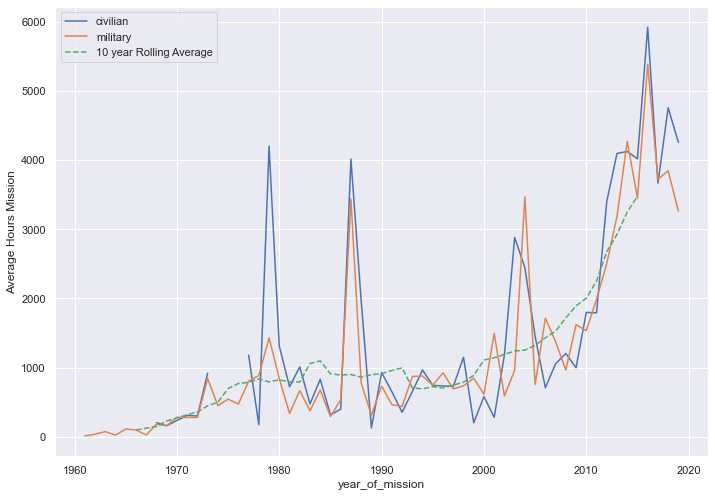


Figure : Average Hours Mission and Average EVA Hours Mission by Year; Year by Year Split by Civilian/Military; 10 Year Rolling Average

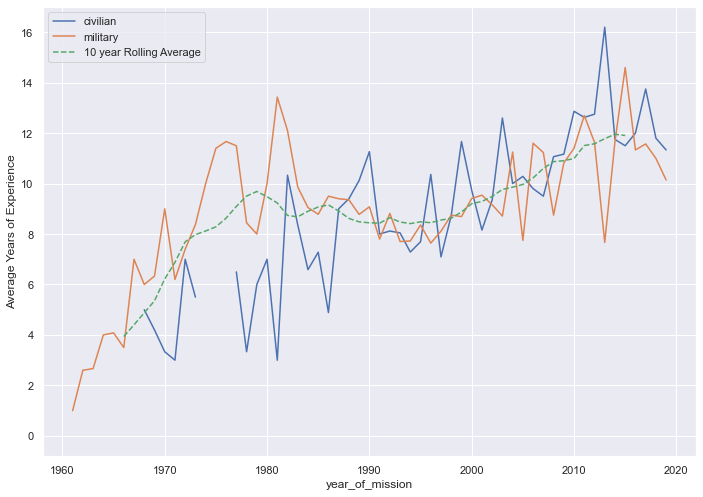
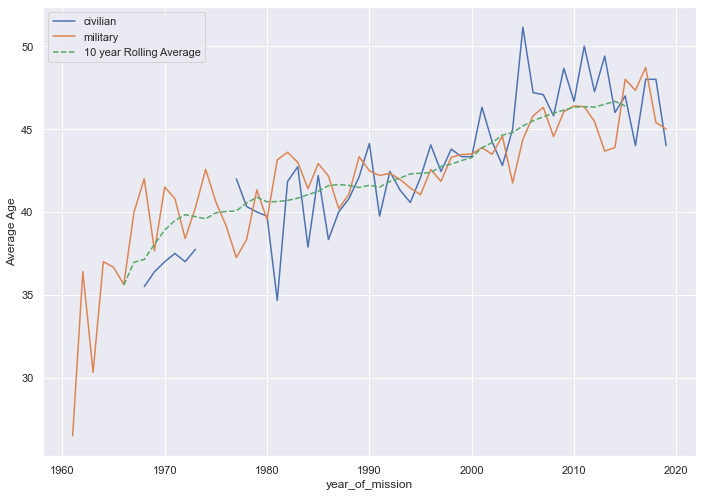


Figure : Average Age and Average Years of Experience During Mission by Year; Year by Year Split by Civilian/Military; 10 Year Rolling Average

**Figure 7** shows that the average hours per mission have increased from *97.8h* in *1961-70* to *3475.5h* in *2010-2019*, indicating that expectations of astronaut mission length have greatly increased with time – an important factor when selecting an astronaut. Likewise, the average of EVA hours per mission have increased from *0.21h* in *1961-1970* to *7.75h* in *2010-2019.* So perhaps EVA training for astronauts is becoming increasingly important.

**Figure 8** shows that the average age of an astronaut at time of mission has also increased, from *35.6* years in *1961-70* to *46.4* years in *2010-2019*. This may be related to increased average experience of an astronaut at time of mission, which has risen from *3.9* years in *1961-70* to *11.9* years in *2010-2019.* This indicates that a space company may have to be prepared to train an astronaut for many years (however this does not consider astronauts that have gone on multiple missions).

Overall, there does not seem an immediate distinction between civilian and military astronauts – an idea which can be used in the selection process.

# Further Thoughts

I decided to focus my report on the *missions* data, as I believed that it would provide the most valuable business insights for a space agency, as it contained *Cost* and *Success* data. I could have gone further examining differences in successful/failed missions by investigating binary classification models. I could have also dealt with missing cost data differently – by removing the non-randomly missing *NaN*s, I would have introduced bias. Imputation could be investigated as an option. I could have investigated the *astronauts* data further, considering why astronauts go on multiple missions, and other astronaut traits such as *sex* or *Country*.

Although I considered these two datasets separately, there also could be additional insights if a relationship can be established, either through additional data or a method – perhaps matching *‘shuttle’* in the astronaut dataset against *‘detail’* in the mission dataset. An assessment suggested this would take a lot of manual work, so best left out of the scope of the initial findings. Had this relationship been available in the data, we could have investigated the astronauts onboard successful/failed missions and investigate the relationship between success and astronaut traits. A good question would be whether astronaut experience and mission success correlate.

# Conclusion

The purpose of this report was to do an initial exploration into the *missions* and *astronauts* datasets, with a view of gathering valuable insights to inform a potential new space agency. From our analysis, we have profiled and identified several competing agencies with active rockets which can provide the basis for further investigation (e.g. *RocketLab* or *ULA*) which can then be used to build our business case. We have also investigated several key characteristics of astronauts and how those have changed over the years – for instance, the average age of an astronaut at time of mission over the past 10 years is *46.4* years, with *11.9* years of training. This will help inform and define astronaut selection and training processes.