



Data Exploration in Python

April 1, 2022

Client Data Scientist Takehome Assessment

Munji Kahalah

Contents

1 Astronauts Dataset	3
1.1 Basic Stats on Our Dataset	3
1.2 Citizens Conducting Space Mission	3
1.3 Time Spent on Missions	4
1.4 Women Pioneers in Space	4
1.5 Distribution of Missions Initiated	5
1.6 Occupation Involved in Missions	5
2 Missions Dataset	6
2.1 Missing Data	6
2.2 Most Popular Days to Launch Missions	7
3 Conclusions	7

Overview

As a data scientist we must understand each dataset we're given before we start building ML models. We need to analyze the datasets by summarizing each features characteristics and relationships, one way to do this is with visualizations. Visualizations are a powerful tool we can use such as scatter-plots, box-plots, heat maps, etc. Using Jupyter Notebook we can clean our data, perform statistical modeling, data visualization, etc. In addition, we must have *clean* data that our ML models can understand. In today's overview we will be performing exploratory data analysis on two datasets about space missions:


- one containing information about astronauts and their missions,
- and the second with some information about the missions themselves.

After conducting EDA on these datasets we found intriguing results that can help share a story to any space agency that would like to understand where each country stands on space missions.

1 Astronauts Dataset

1.1 Basic Stats on Our Dataset

Examining our astronauts dataset we run a useful function by pandas `.describe()`, which provides the count, mean, standard deviation, minimum and maximum values of the data. In figure 1 we get an idea of averages on each numerical category. Looking at the first column we see the average astronaut was born in 1951, oldest was born in 1921, and youngest was in 1983. Most astronauts will conduct about 3 missions in their career and spend 1050 hours (43.75 days) per mission. We can also examine upper or lower percentiles of each respective column.



	year_of_birth	year_of_selection	mission_number	total_number_of_missions	year_of_mission	hours_mission	total_hrs_sum	field21	eva_hrs_mission
count	1277.000000	1277.000000	1277.000000	1277.000000	1277.000000	1277.000000	1277.000000	1277.000000	1277.000000
mean	1951.683634	1985.58888	1.992169	2.982772	1994.597494	1050.883984	2968.341410	0.628818	3.661287
std	11.435117	12.21917	1.145361	1.400745	12.583237	1714.791959	4214.715104	1.165753	7.287245
min	1921.000000	1959.00000	1.000000	1.000000	1961.000000	0.000000	0.610000	0.000000	0.000000
25%	1944.000000	1978.00000	1.000000	2.000000	1986.000000	190.030000	482.000000	0.000000	0.000000
50%	1952.000000	1987.00000	2.000000	3.000000	1995.000000	261.000000	932.000000	0.000000	0.000000
75%	1959.000000	1995.00000	3.000000	4.000000	2003.000000	382.000000	4264.000000	1.000000	4.720000
max	1983.000000	2018.00000	7.000000	7.000000	2019.000000	10505.000000	21083.520000	7.000000	89.130000

Figure 1: Describe function on Numerical Data

1.2 Citizens Conducting Space Mission

From the data we see that 40 countries have participated in space missions, with 195 countries total that means only 20% have the capabilities to conduct such missions. In figure 2 we see the number of missions conducted by each nationality with USA dominating this sector. Second in our plot is Russia, we can clearly see that this massive surge between the two countries was initiated by the great Space Race back in the Cold War.

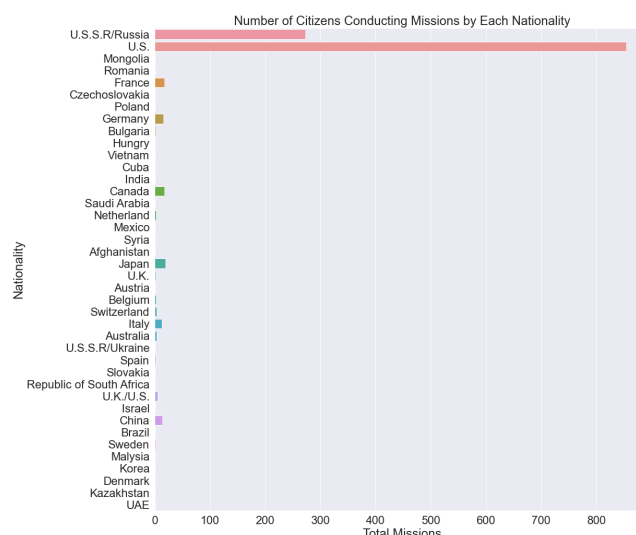


Figure 2: Number of Citizens Who Have Gone on Missions

Furthermore in the era of the Space Race (1955-1975) Russia launched 97 missions while the USA launched 169 missions.

1.3 Time Spent on Missions

Now we dive into which astronauts have spent the most time on missions, in figure 3 we analyze each nationality and their astronauts who have spent 6 months or more on missions. Most points fall between the range of 10,000 hours and under 30 of extravehicular hours. With some outliers going above and beyond 17,500 hours on some mission. The Russian, Gennady Padalka, spent a world record of 21,000 hours in space.

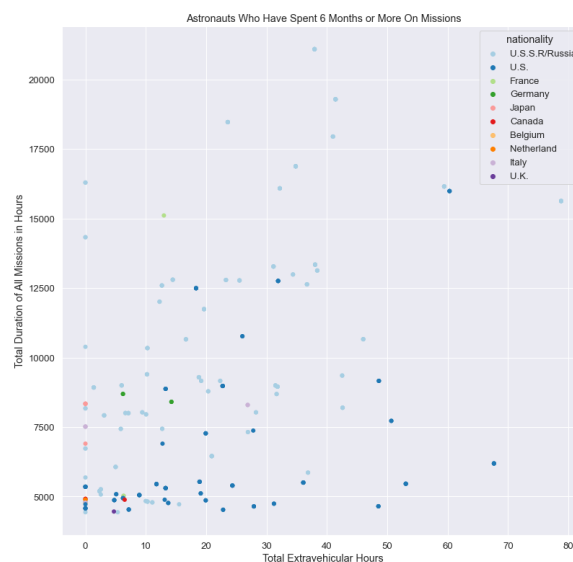


Figure 3: Amount of Time Astronauts Spend in Space

1.4 Women Pioneers in Space

Based off of our data we see that only 11% of women have participated in space missions, with a total of 143 women. Figure 4 shows us that USA comes in first place for involving women in missions, with Russia in second place again.

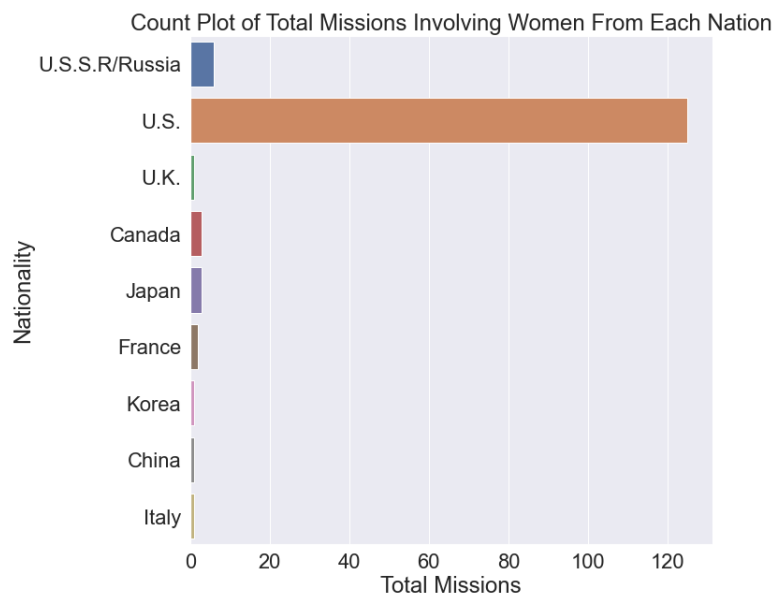


Figure 4: Women Involved in Space Missions

1.5 Distribution of Missions Initiated

Most missions have taken place from the late 1980's to mid 2010's, but the peak was late 90's and early 2000's. Figure 5 shows the frequency distribution of when missions actually took place with the late 90's having over 100 missions.

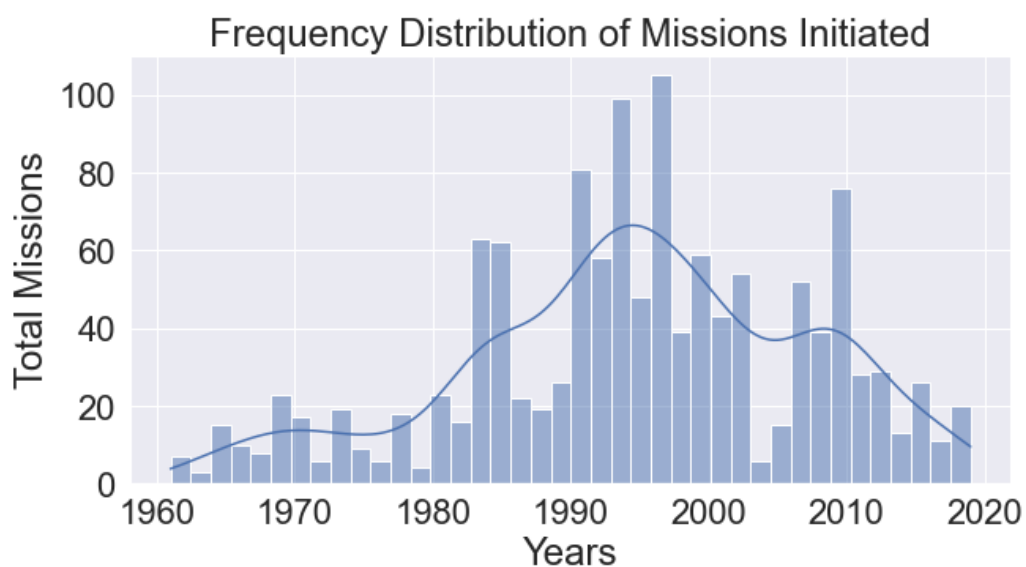


Figure 5: Missions Initiated

1.6 Occupation Involved in Missions

Each astronaut come from different backgrounds, figure 6 illustrates that Russia has utilized it's position of commanders the most to be involved in space missions, while small countries such as Malaysia and Korea were space tourist.

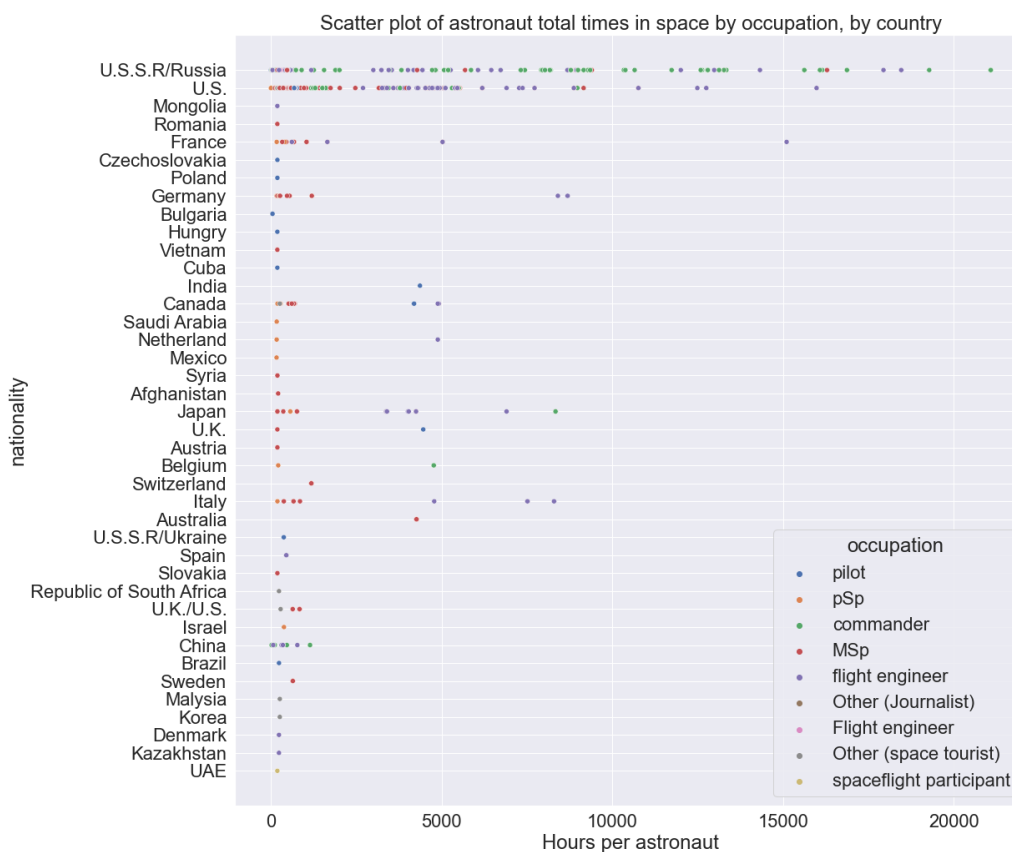
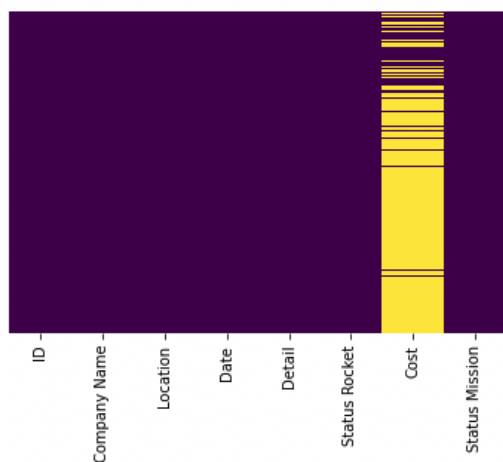


Figure 6: Occupation Involved in Missions

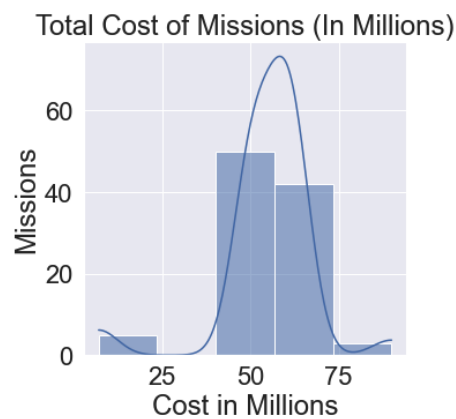
2 Missions Dataset

2.1 Missing Data

Looking at each mission we see that companies are notorious for not reporting their cost, with 77% values missing. I filtered our dataset between USA and Russia and found that Russia didn't report 96% of its cost, while USA didn't account for 63%. I filtered our data to examine SpaceX and imputed values for missing data to get the average cost. Figure 7 shows us the missing data and SpaceX mission costs.



(a) Missing data



(b) SpaceX Cost of missions

Figure 7: Missing Data and SpaceX Cost

2.2 Most Popular Days to Launch Missions

For fun I decided to figure out what days were most popular to launch for each company. Although Friday was high up on the list, I was a bit surprised to see Wednesday take the lead with 821 launch days. Figure 8 conveys our findings with Sunday being the least for space launches.

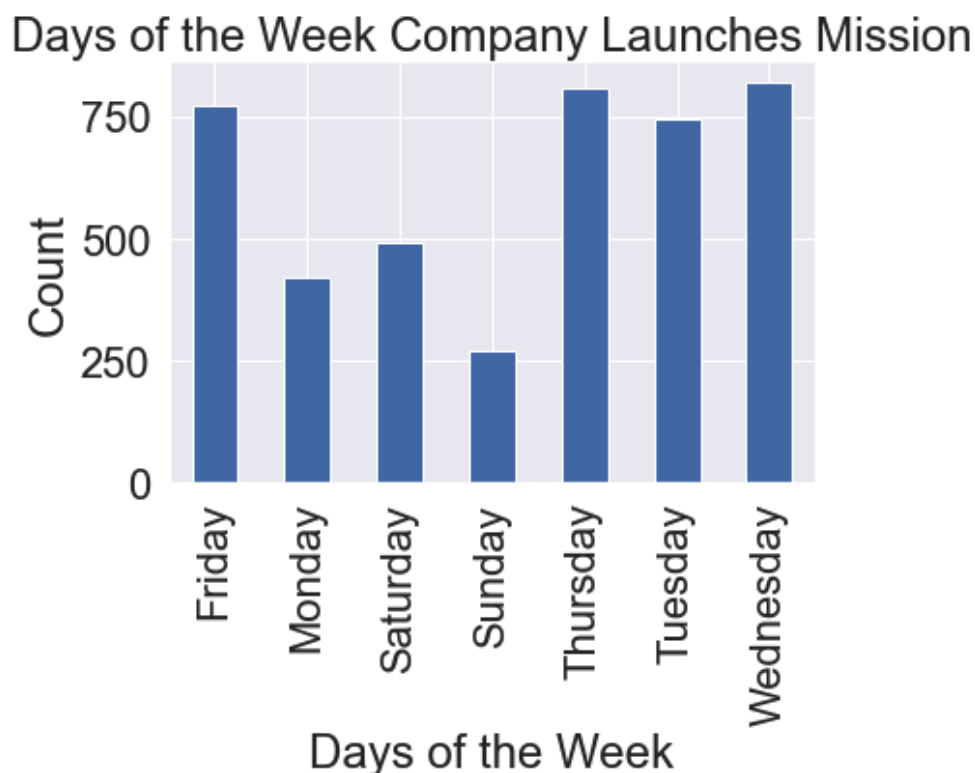


Figure 8: Popular Days for Launch

3 Conclusions

In my Jupyter Notebooks you'll find more data that I chose not to share in my report, I decided to convey what I found to be the most interesting and useful for analysis. I would have liked to include my heat-map of my correlation matrix but we didn't account for categorical features, I would have started one hot encoding and then run a script to test which features are causing possible over-fitting. With this dataset I would have made ML models for how long an astronaut can expect to spend time in space, while with the missions dataset I would've made a ML model to show success rate in missions. My next steps would have been to provide scaling, feature engineering, and feature selection.