



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alexander Paul
8 April, 2025



Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
 - SpaceX Launch Data Collection
 - Data wrangling
 - Exploratory Data Analysis (EDA) using SQL
 - EDA with Data Visualization
 - Interactive Visual Analytics with Folium
 - Interactive Dashboard with with Plotly Dash
 - Predictive Analysis (ML Classification to predict successful rocket launches)
- Summary of all results
 - EDA identified the most successful launch sites, and key factors associated with successful launches
 - EDA found that the launch success rate generally increased from 2013 – 2020
 - Predictive analysis resulted in a ML model that may predict future launch success with 83.33% accuracy

Introduction

- Project background and context:

SpaceX is the world's most successful and prevalent space launch provider, servicing both private companies and national space programs. It provides the most affordable rocket launches via Falcon 9 rockets, with a cost of 62 million dollars, while other providers cost upward of 165 million dollars. Much of the savings SpaceX can provide is because they can reuse the first stage of the Falcon9 rocket by landing it after it has delivered the payload to space. Therefore, if we can predict whether the first stage of a SpaceX launch will land, we can determine the cost of the launch and potentially out bid SpaceX for their business. Based on public information of SpaceX Falcon9 launches, we will determine factors that associate with successful stage one landings and develop a machine learning (ML) model to predict if SpaceX will land and reuse the first stage of future rocket launches.

- Key questions to answer:

- How do factors such as launch site, payload mass, the number of launches, and the target orbit affect stage one landing success?
- What is average yearly rate and trend of successful launches?
- What is the best ML model for predicting stage one reuse?

Section 1

Methodology

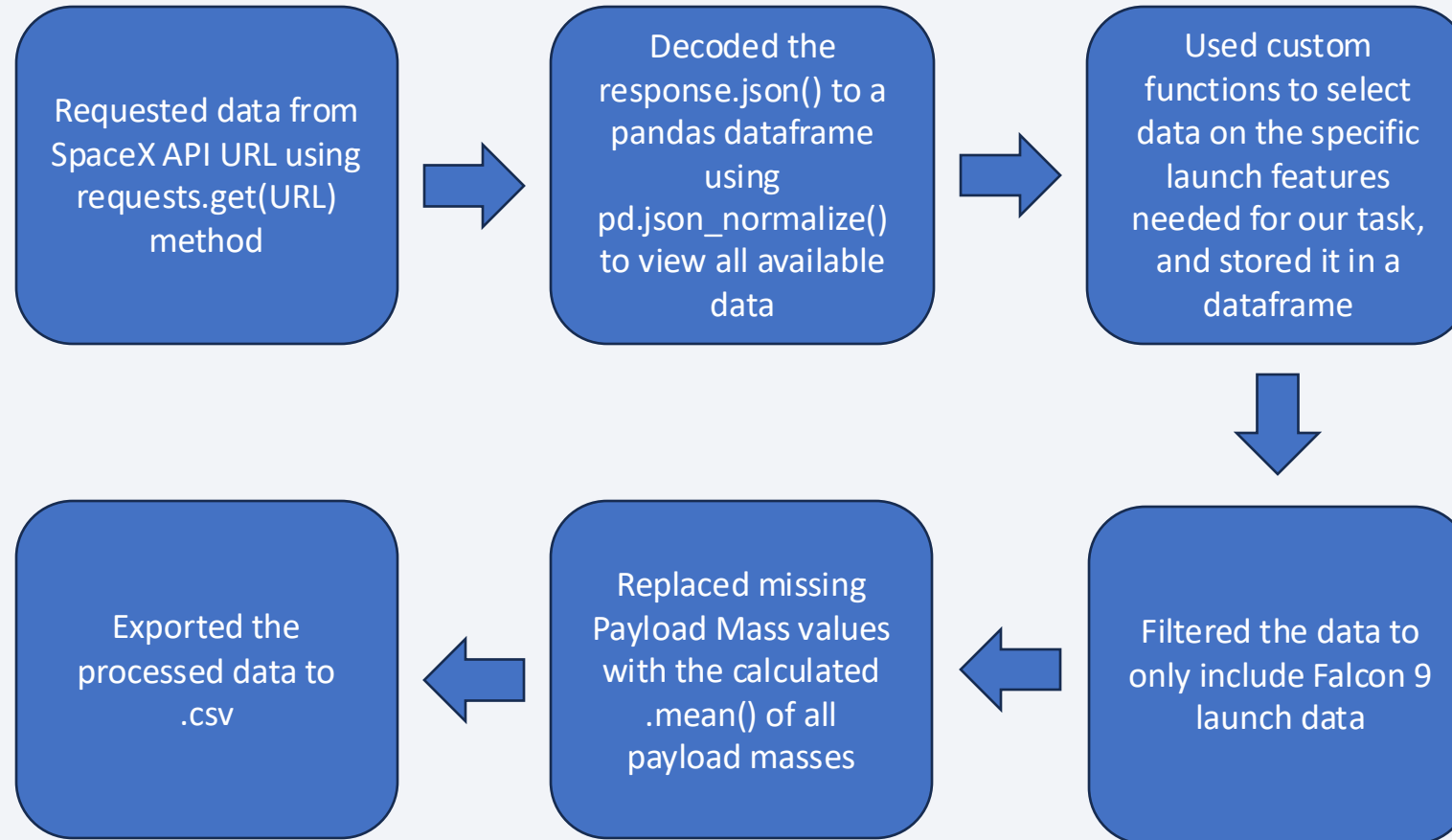
Methodology: Executive Summary

- Data collection methodology:
 - Data were collected via the SpaceX REST API
 - Data were collected via web scraping from SpaceX Wikipedia
- Performed data wrangling
 - Data were filtered to focus on Falcon9 launches
 - Missing values were dealt with
 - Categorical variables were One-Hot encoded
 - Engineered a feature to represent successful and unsuccessful stage one landings
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
 - Developed pipelines to test several ML classification models, and optimized them using GridSearchCV on training and testing splits of the data

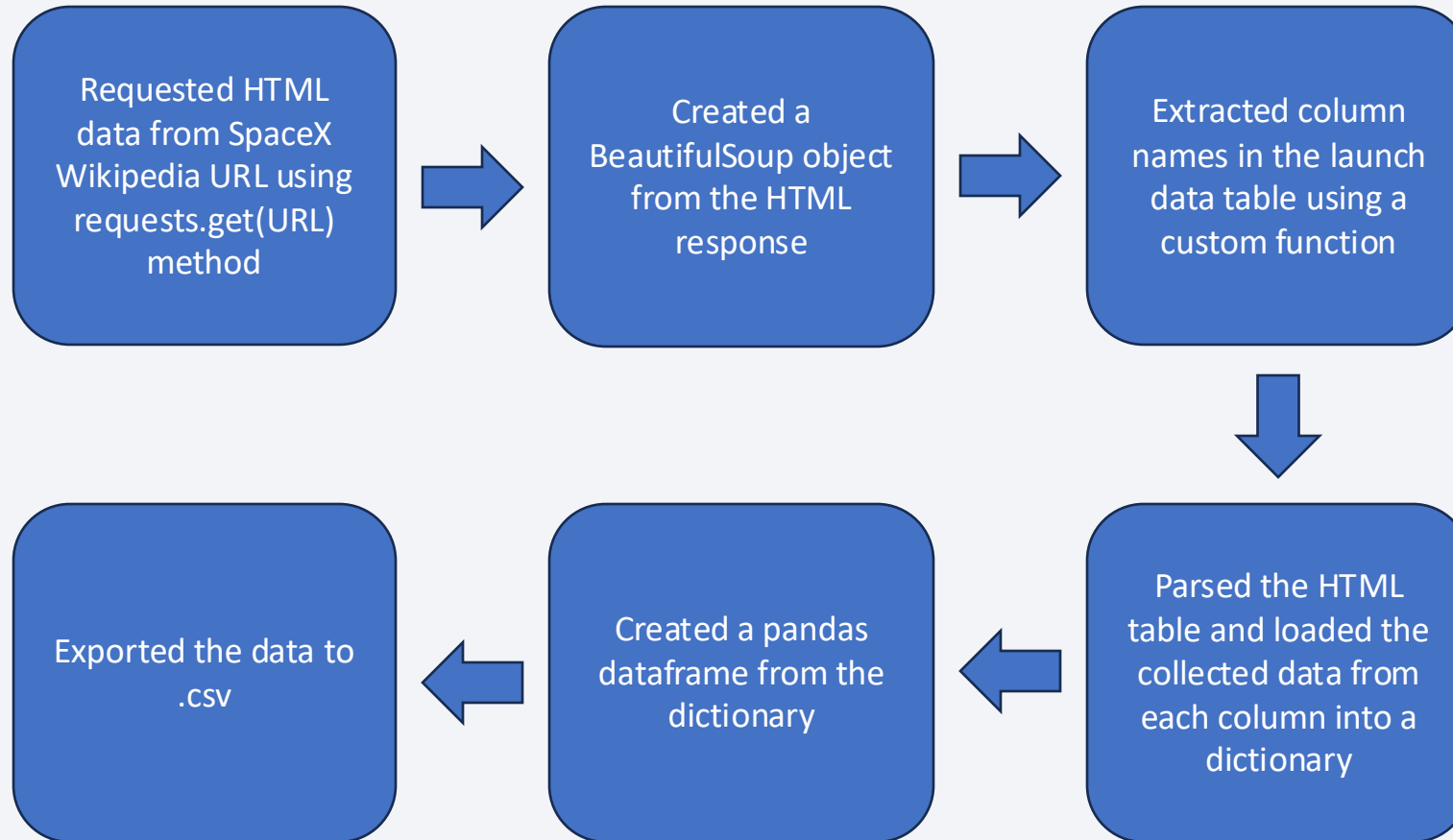
Data Collection

- Data were collected from two sources and filtered and aggregated to form a comprehensive data set on historical SpaceX Falcon 9 rocket launches.
- We gathered data from:
 - the SpaceX REST API
 - a table on the SpaceX Wikipedia page using Web Scraping

Data Collection – SpaceX API



Data Collection - Web Scraping

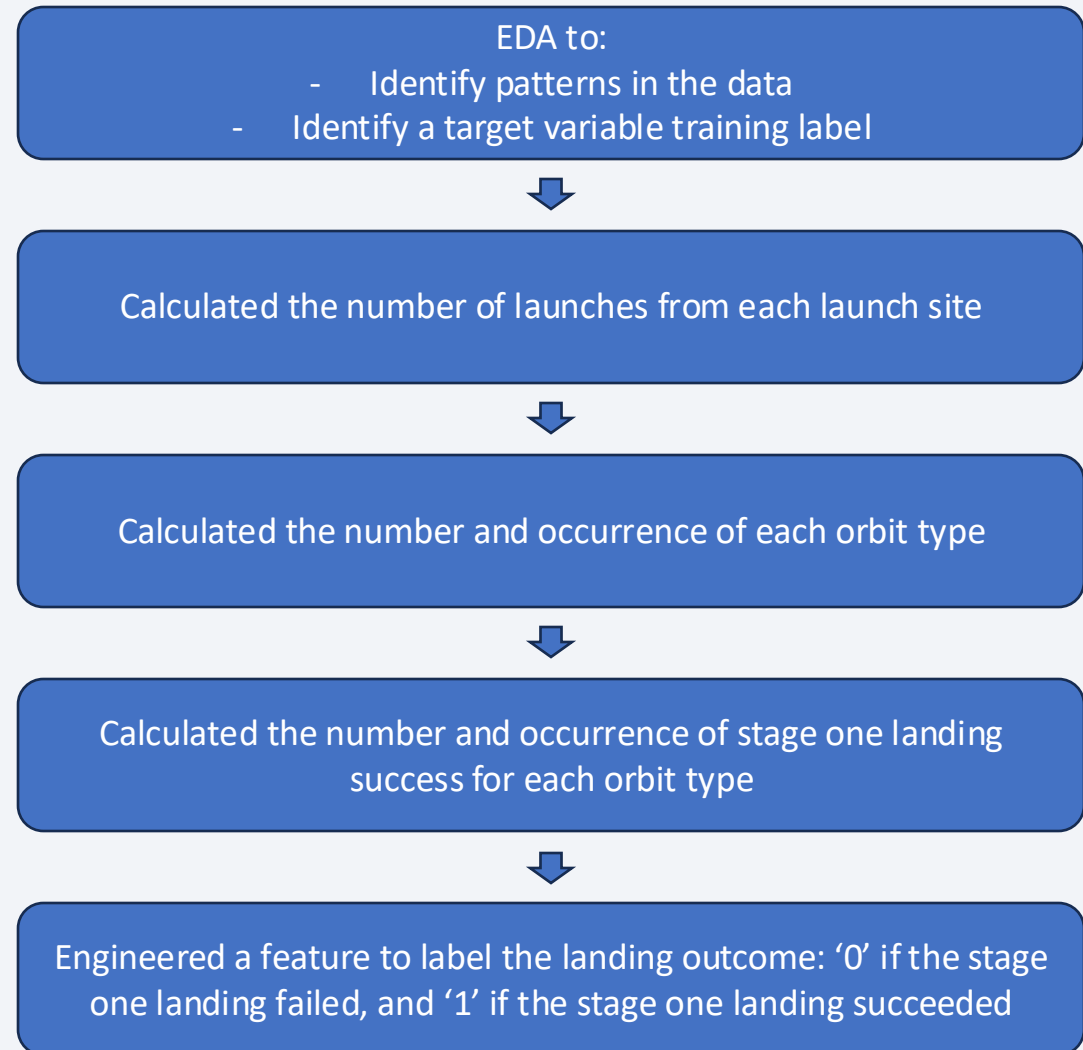


Data Wrangling

We performed preliminary exploratory data analysis (EDA) to identify basic patterns in the data and to identify a target variable for training supervised ML classification models for stage one landing success.

The main finding was that the collected SpaceX launch data contains several stage one landing success designations according to the landing pad type. For our purposes, we want to train a binary classifier on landing success or failure and do not need to include the landing pad type. We thus engineered a new feature to label the landing outcome: '0' if the stage one landing failed for any reason, and '1' if the stage one landing succeeded.

[GitHub Link: Data Wrangling](#)



EDA with Data Visualization

- We plotted the following charts:
 - Flight number vs Payload mass (scatter plot)
 - Flight number vs Launch site (scatter plot)
 - Payload mass vs Launch site (scatter plot)
 - Success rate by Orbit type (bar chart)
 - Flight number by Orbit type (scatter plot)
 - Payload Mass by Orbit type (scatter plot)
 - Success rate by year (line chart)

The scatter plots were used to visually check for potential correlations between the plotted variables. These could indicate key factors to use in the predictive classification ML model.

The bar chart was used to visually identify the orbit types with most and least successful landings.

The line chart of the total stage one landing success rate per year allows visualization of both the yearly trend and variability in stage one landing success.

EDA with SQL

- The data were queried using SQL to:
 - Display the unique launch site names
 - Display the data for launch sites beginning with the string 'CCA'
 - Display the total payload mass carried by boosters launched by NASA (CRS)
 - Display the average payload mass carried by booster version F9 v1.1
 - List the date when the first successful landing outcome in ground pad was achieved
 - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
 - List the total number of successful and failure mission outcomes
 - List all the booster versions that have carried the maximum payload mass
 - List the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
 - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

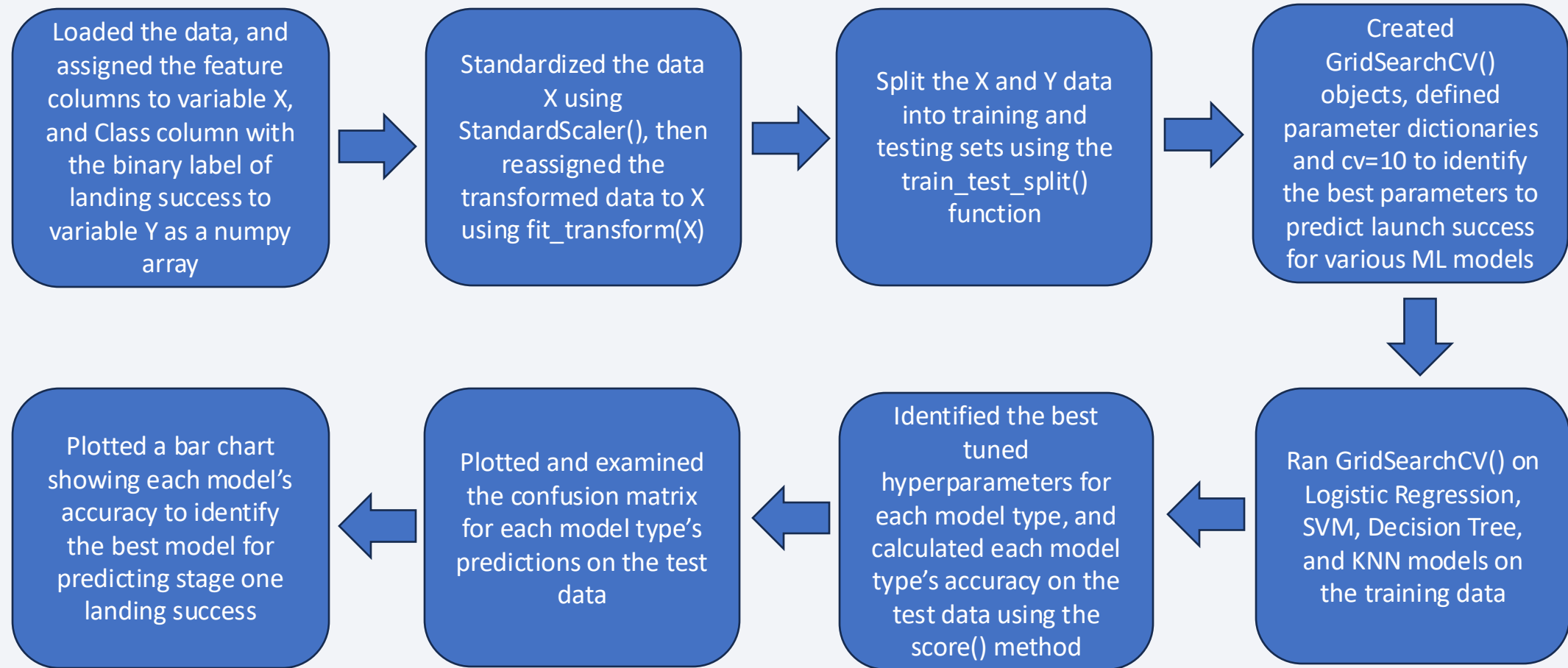
Build an Interactive Map with Folium

- We added the following objects to an interactive map:
 - Markers of launch sites:
 - Circle marker with a pop-up label of NASA Johnson Space Center using its latitude and longitude coordinates. This was done as a starting example.
 - Circle markers with pop-up labels for all SpaceX Falcon9 launch sites using their latitude and longitude coordinates. These were marked to show the launch sites locations and their proximities to various geographical features.
 - Colored markers to indicate the launches and their outcomes for each launch site:
 - Used MarkerCluster to add markers for each launch colored by their outcome with green for success, and red for failure. This created an easy way to visually distinguish the most and least successful launch sites.
 - Distance measures between geographical features and an example Launch site:
 - Added lines with measured distances between the CCAFS LC-40 launch site and the nearest highway, coast, city, and railway. This was done to show that launch sites are typically close to highways, railways, and coasts, but farther from cities.

Build a Dashboard with Plotly Dash

- We added the following interactions and plots to a Dashboard:
 - Launch site drop down list, that allows a user to select to see data from all launch sites, or from a specific launch site.
 - Pie chart showing the proportion and counts of launch success of the selected launch site(s). This allows users to see which launch sites are the most and least successful.
 - A scatter plot of Payload mass vs Success rate, colored for the different booster versions. This allows users to visualize possible correlation between payload mass and payload mass for the different booster versions of the selected launch site(s).
 - An interactive slider allowing users to select a payload mass range. This allows users to check if certain ranges of payload associate with launch success.

Predictive Analysis (Classification)



Results

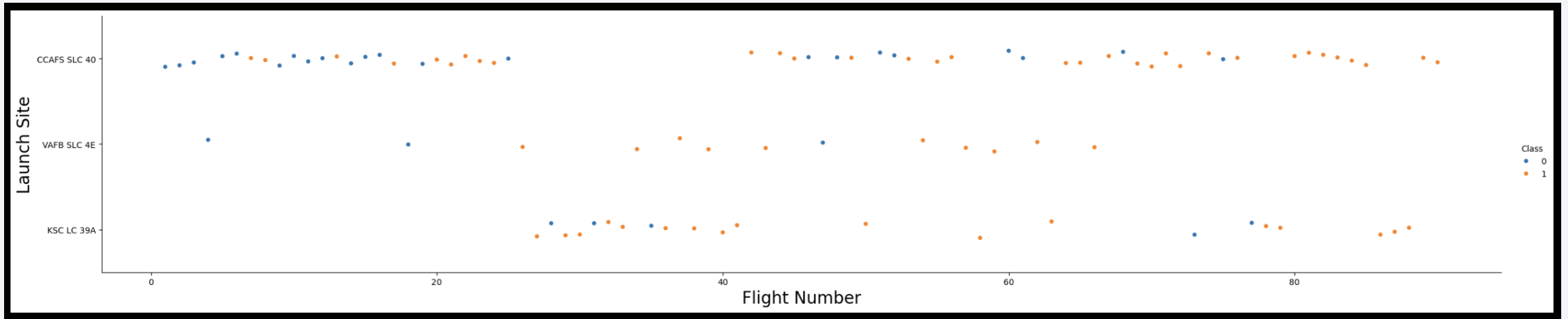
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of blue and red, creating a sense of motion or data flow. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is high-tech and digital.

Section 2

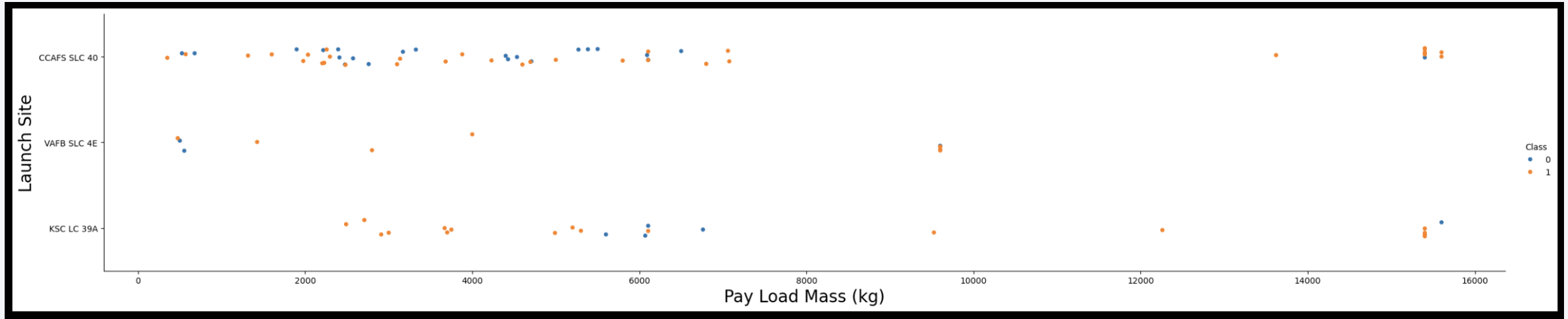
Insights drawn from EDA

Flight Number vs. Launch Site



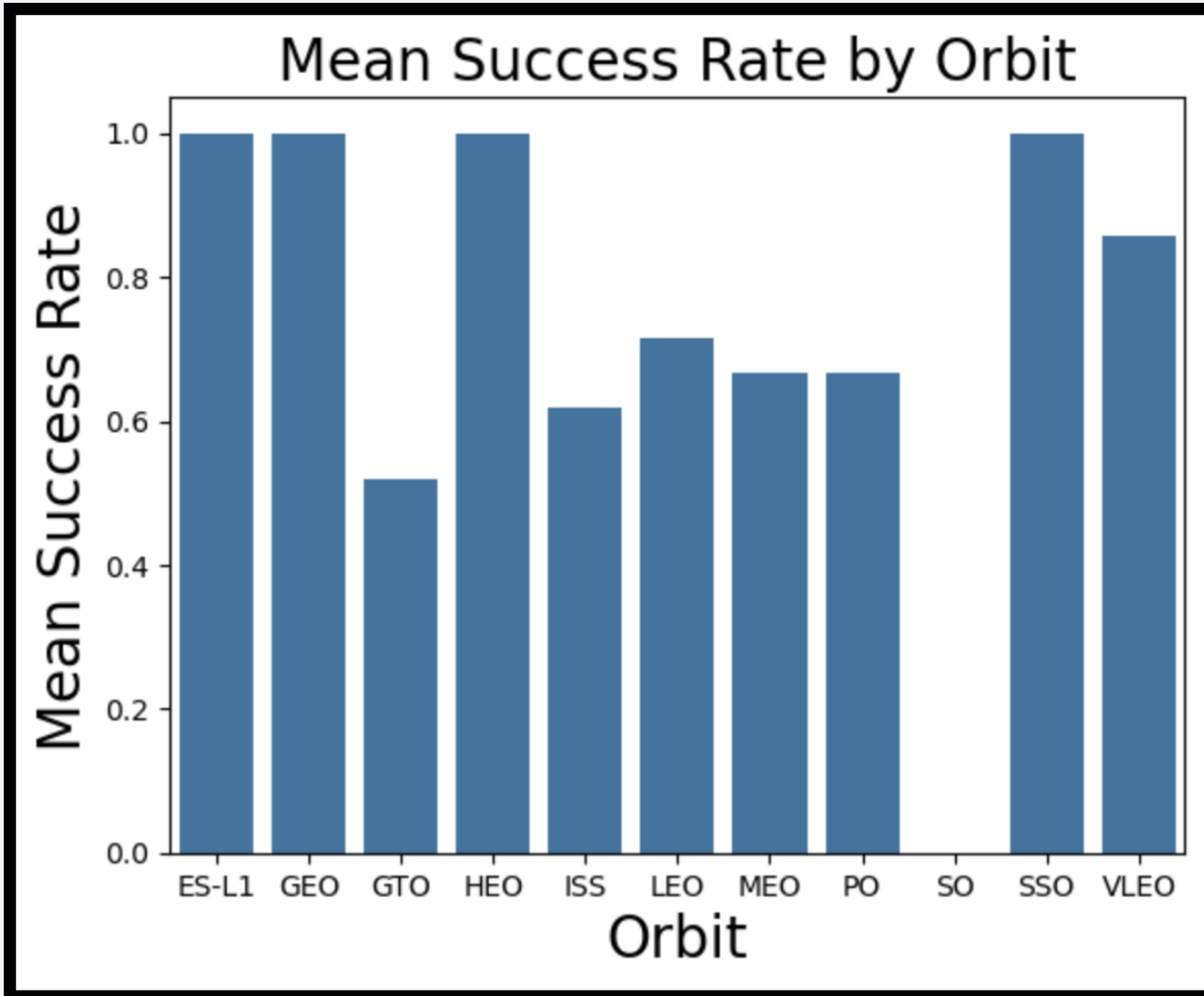
- Insights from the plot:
 - The first launches at each site tend to have failed (blue dots), then as flight number increases the prevalence of successful launches (orange dots) increases at each launch site
 - Site CCAFS SLC 40 has the most launches, but the lowest success rate
 - Sites VAFB SLC 4E, and KSC LC 39A have fewer launches, but higher success rates
 - The trends indicate that SpaceX increased the stage one landing success of its Falcon 9 rocket launches for every launch site over time

Payload vs. Launch Site



- Insights from the plot:
 - Each launch site tends to have higher success with higher payload masses.
 - Nearly all launches (across all launch sites) with greater than 7000 kg payload mass were successful
 - Site CCAFS SLC 40 has the lowest success rate for launches with between 0 and 8000 kg of payload mass
 - Site VAFB SLC 4E has near perfect launch success rate from 1000 to 10000 kg of payload mass
 - Site KSC LC 39A has perfect launch success for 2000 to 4000 kg payload mass, low success for 5000 to 7000 kg payload mass, and near perfect success for 7000 to 16000 kg payload mass

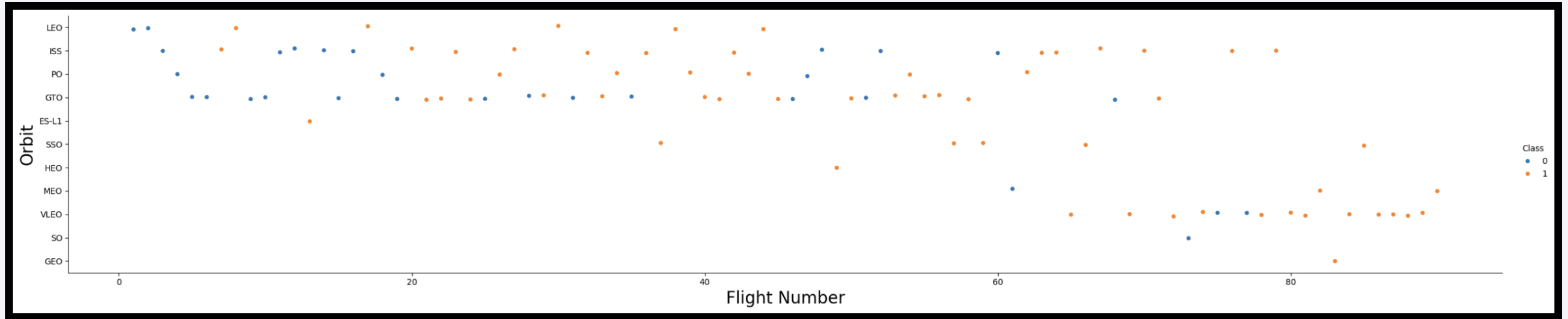
Success Rate vs. Orbit Type



- Insights from the plot:

- Launches to orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
- Launches to orbits GTO, ISS, LEO, MEO, PO, and VLEO have success rate between 50% and 85%
- Launches to orbit SO have 0% success

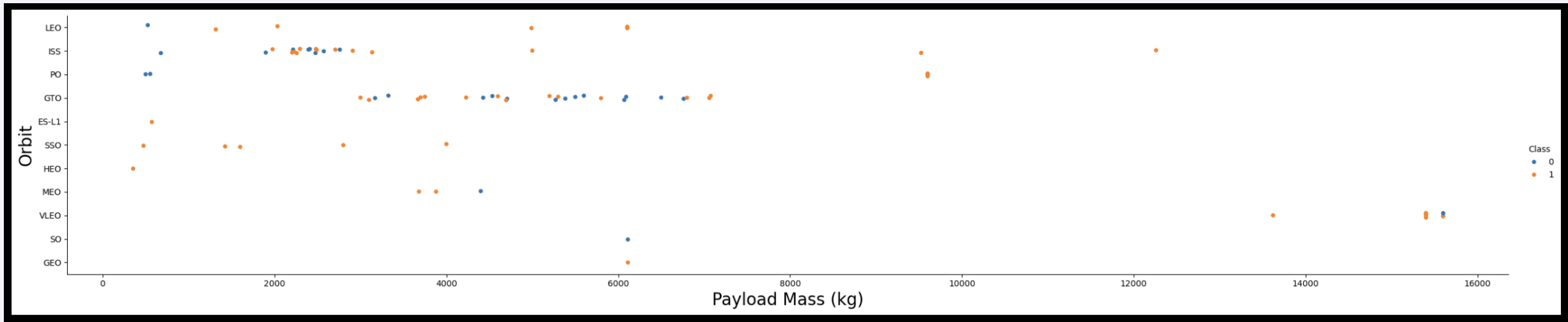
Flight Number vs. Orbit Type



- Insights from the plot:

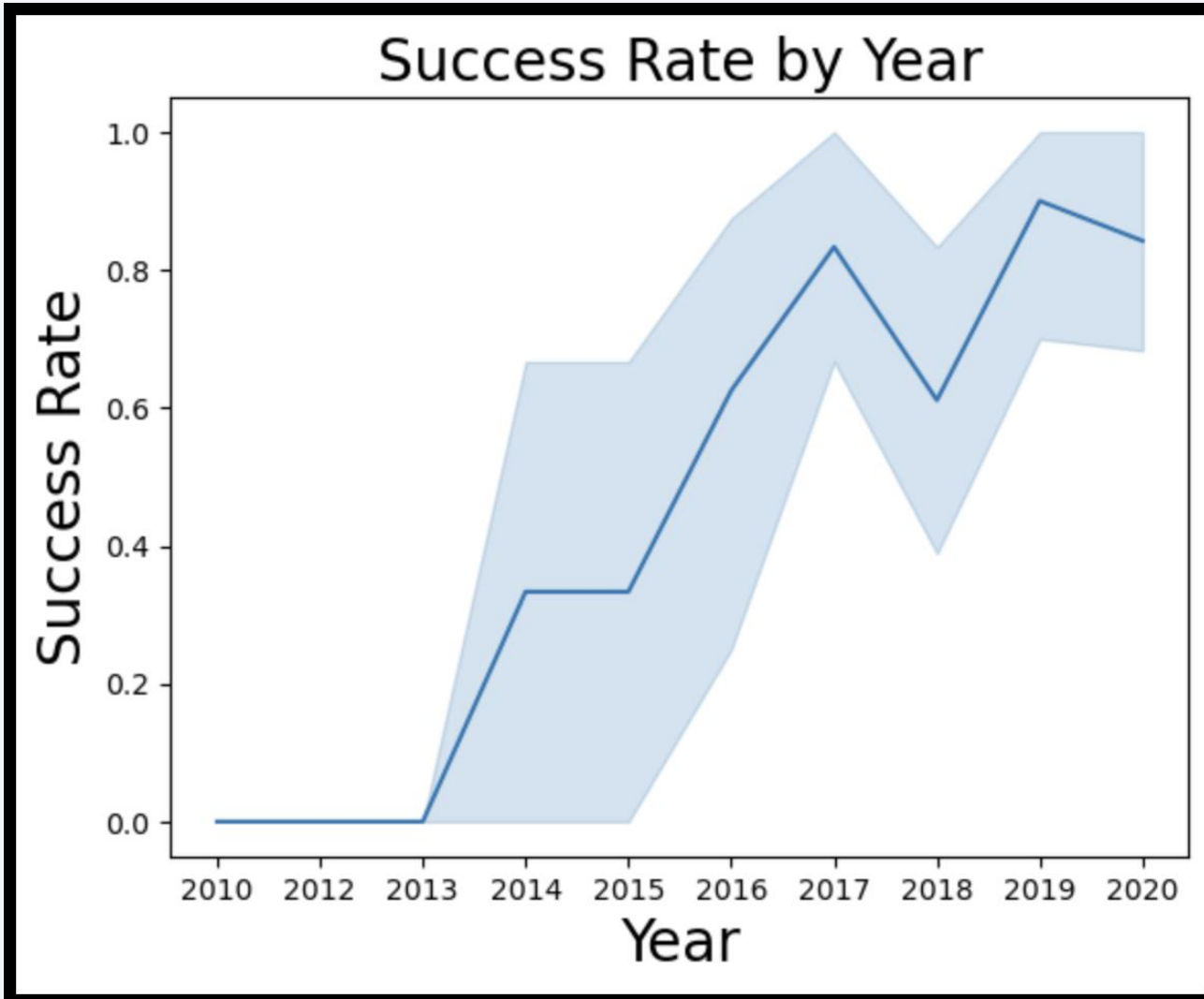
- The orbits that have the most launches are those with success rates ranging between 50% and 85% (GTO, ISS, LEO, MEO, PO, and VLEO). We may conclude that if enough launches are made to any orbit the success rate may fall between 50% and 85%.
- The orbits that have 100% success rate (ES-L1, GEO, HEO, and SSO) have only 1 to 5 launches. Thus, if the number of launches to these orbits increased, we may expect the success rate to decrease to within the 50% to 85% success rate range.
- Orbit SO, which has 0% success rate, only had 1 launch. Thus, if the number of launches to SO increased we may expect the success rate to increase to within the 50% to 85% success rate range.

Payload vs. Orbit Type



- Insights from the plot:
 - Launch success to LEO, ISS, PO, and VLEO orbits appears to increase with higher payload masses.
 - Launch success to GTO and MEO orbits appears to decrease with higher payload masses.
 - Launch success to SSO orbit is perfect over a range of 0 to 4000 kg of payload mass.

Launch Success Yearly Trend



- Insights from the plot:
 - Overall, the success rate has tended to increase from 2013 to 2020 from 0% to about 85% success.
 - The variability of launch success has also decreased with time.
 - No launches between 2010 and 2013 were successful.
 - From 2013 to 2017, the success rate increased, on average, despite a period of no change from 2014 to 2015.
 - From 2017 to 2018 the success rate dipped, then increased 2018 to 2019, and slightly decreased from 2019 to 2020.

All Launch Site Names

```
1 #use sql to display the names of the unique launch sites
2 %sql select distinct "Launch_Site" from SPACEXTABLE;
```

[12] ✓ 0.0s Python

... * [sqlite:///my_data1.db](#)
Done.

...

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- This query selects and displays the names of the unique launch sites in dataset

Launch Site Names Begin with 'CCA'

```
1 # use sql to Display 5 records where launch sites begin with the string 'CCA'
2 %sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5;
```

[13] ✓ 0.0s Python

... * [sqlite:///my_data1.db](#)

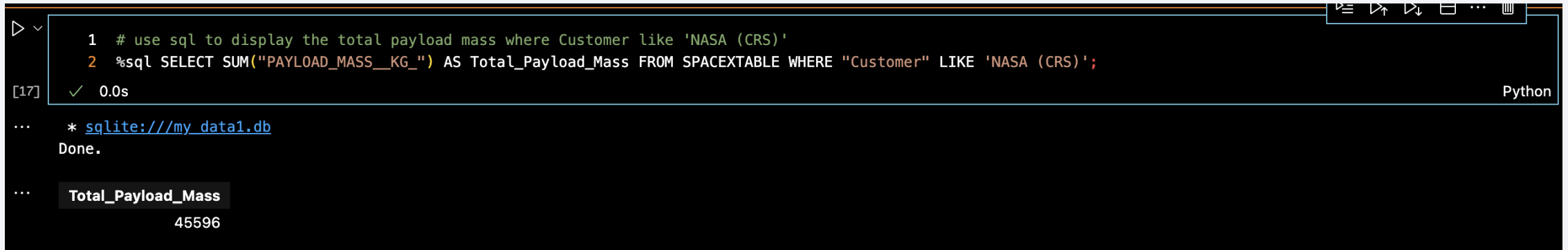
Done.

...

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- This query selects and displays five data records for launch sites with names that begin with the string 'CCA'

Total Payload Mass



A screenshot of a Jupyter Notebook interface. The top part shows a code cell with two lines of SQL code. The first line is a comment: `# use sql to display the total payload mass where Customer like 'NASA (CRS)'`. The second line is the SQL query: `%sql SELECT SUM("PAYLOAD_MASS_KG") AS Total_Payload_Mass FROM SPACEXTABLE WHERE "Customer" LIKE 'NASA (CRS)';`. Below the code cell, the output shows a green checkmark, a success message, and a execution time of 0.0s. The output also includes a prompt `* sqlite:///my_data1.db` and the word `Done.`. At the bottom, a table with one column `Total_Payload_Mass` and one row containing the value `45596` is displayed. The Jupyter Notebook interface includes standard icons for running, stepping through, and saving code, as well as a 'Python' label in the bottom right corner of the code cell.

```
1 # use sql to display the total payload mass where Customer like 'NASA (CRS)'
```

```
2 %sql SELECT SUM("PAYLOAD_MASS_KG") AS Total_Payload_Mass FROM SPACEXTABLE WHERE "Customer" LIKE 'NASA (CRS)';
```

[17] ✓ 0.0s Python

... * [sqlite:///my_data1.db](#)

Done.

Total_Payload_Mass
45596

- This query calculates and displays the total payload mass for launches ordered by the customer NASA (CRS)

Average Payload Mass by F9 v1.1

```
[18] 1 # use sql to Display average payload mass carried by booster version F9 v1.1
      2 %sql SELECT AVG("PAYLOAD_MASS_KG") AS Average_Payload_Mass FROM SPACEXTABLE WHERE "Booster_Version" LIKE 'F9 v1.1%';
✓ 0.0s Python

... * sqlite:///my_data1.db
Done.

... Average_Payload_Mass
2534.6666666666665
```

- This query calculates and displays the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
1 #use sql min function to list the date when the first successful landing outcome in ground pad was achieved
2 %sql SELECT MIN("Date") AS First_Successful_Landing FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (ground pad)';
```

[21] ✓ 0.0s Python

... * [sqlite:///my_data1.db](#)
Done.

... **First_Successful_Landing**
2015-12-22

- This query finds the date of the first successful landing outcome on a ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

```
1 #List the names of the "Booster_Version" where the "Landing_Outcome" is "Success (drone ship)" and "PAYLOAD_MASS_KG_" is greater than 4000 but less than 6000
2 %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS_KG_" > 4000 AND "PAYLOAD_MASS_KG_" < 6000
3
4
```

[22] ✓ 0.0s Python

... * [sqlite:///my_data1.db](#)
Done.

... **Booster_Version**

F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- This query lists the names of booster versions which have successfully landed on a drone ship and had payload mass between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

```
1 %sql select "Mission_Outcome", count(*) as Total from SPACEXTABLE group by "Mission_Outcome" order by Total desc;
```

[49] ✓ 0.0s Python

... * [sqlite:///my_data1.db](#)
Done.

...

Mission_Outcome	Total
Success	98
Success (payload status unclear)	1
Success	1
Failure (in flight)	1

- This query calculates and displays the total number of successful and failed mission outcomes

Boosters Carried Maximum Payload

```
1 #List all the booster_versions that have carried the maximum payload mass. Use a subquery.
2 %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS_KG_" = (SELECT MAX("PAYLOAD_MASS_KG_") FROM SPACEXTABLE);
```

[51] ✓ 0.0s Python

... * [sqlite:///my_data1.db](#)
Done.

... **Booster_Version**

F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- This query lists the names of the booster versions which have carried the maximum payload mass

2015 Launch Records

```
1 #use sqllite to list the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2
2 %sql SELECT strftime('%m', "Date") AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Failure (drone ship'

[31] ✓ 0.0s Python
```

... * [sqlite:///my_data1.db](#)
Done.

...

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- This query lists the failed landing outcomes in drone ship, their booster versions, and launch site for the months (01 = January, and 04 = April) in 2015

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

```
1 #Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
2 %sql SELECT "Landing_Outcome", COUNT("Landing_Outcome") AS Count FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORD
```

[32] ✓ 0.0s Python

... * [sqlite:///my_data1.db](#)
Done.

...

Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

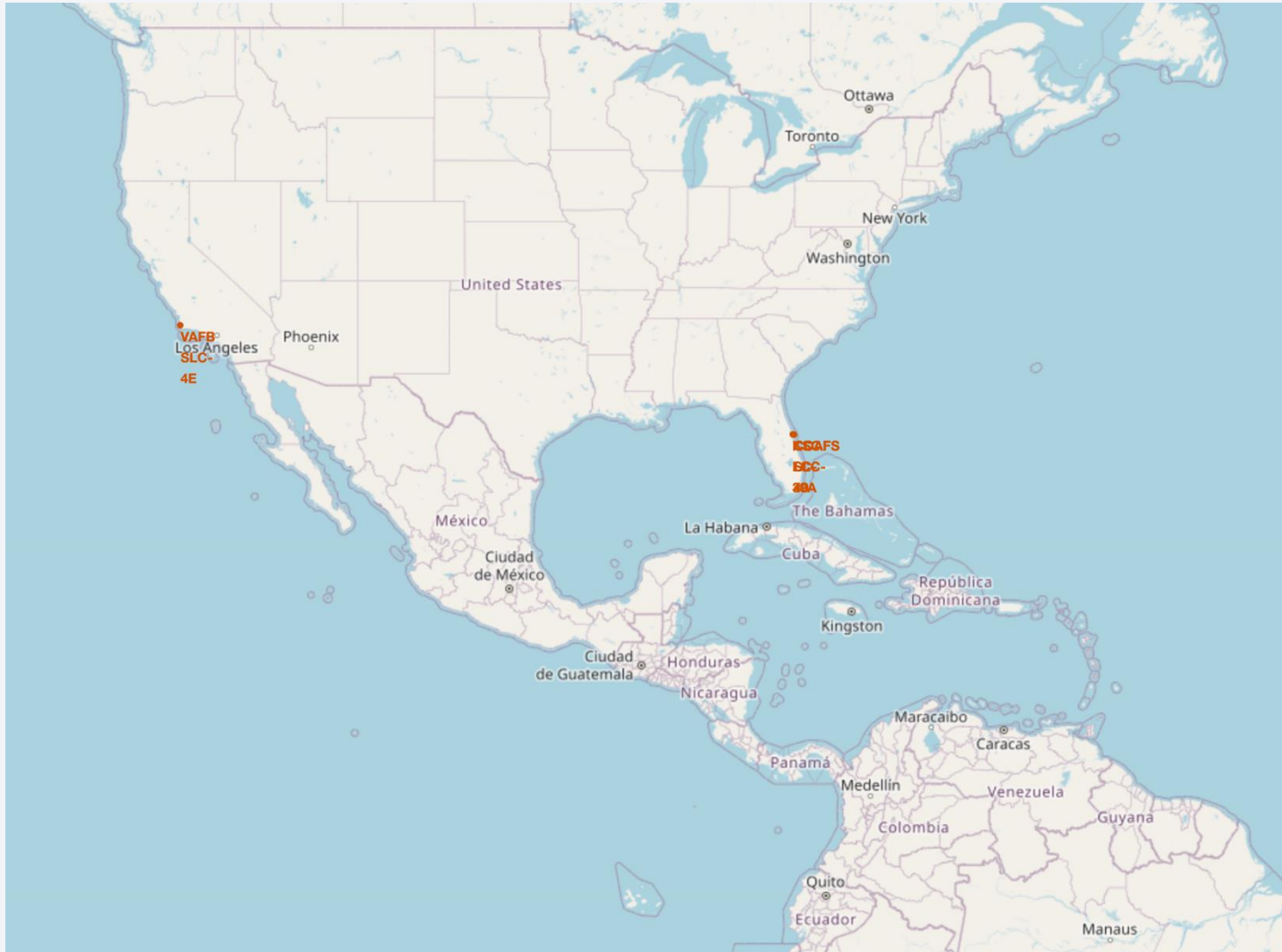
- This query ranks the count of landing outcomes for launches between the dates 2010-06-04 and 2017-03-20, in descending order

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a solid blue background on the left and a satellite photograph of Earth on the right. The Earth's surface is dark, with numerous bright yellow and orange lights representing cities and urban areas. The horizon of the Earth is visible as a thin, curved line separating the dark surface from the deep blue of space.

Section 3

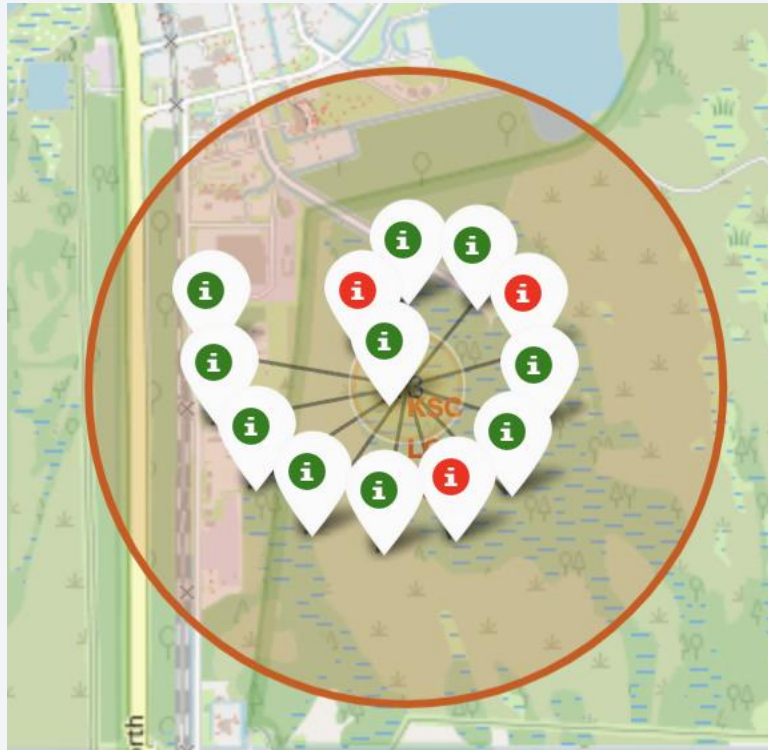
Launch Sites Proximities Analysis

Falcon9 rocket launch site locations

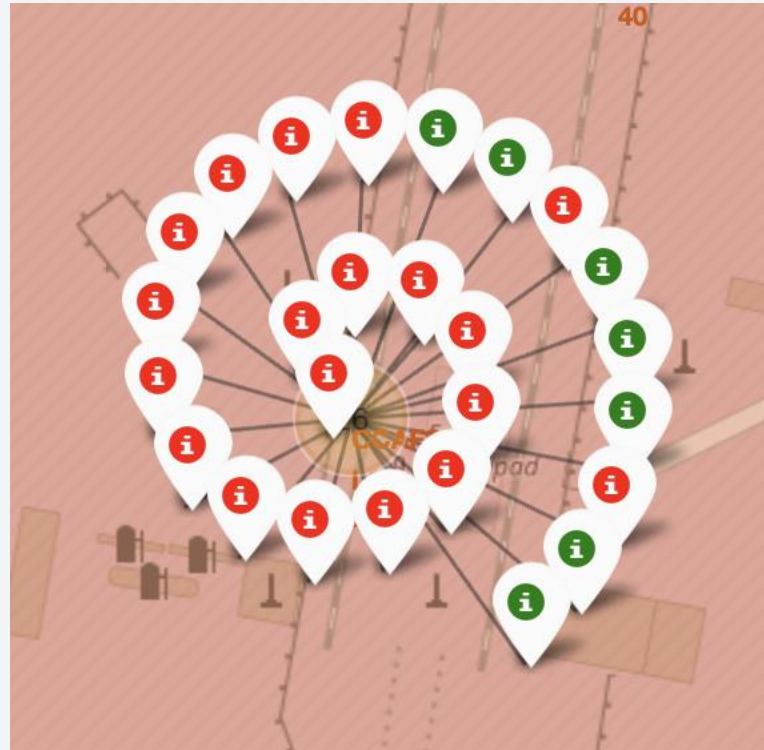


- Insights from the map:
 - The launch sites are located in the southern United States. This places the launch sites close to the equator, where the rotational speed of the earth has the greatest impact on the launched rocket's orbital velocity via inertia.
 - The launch sites are placed close to coastlines. Rockets are launched over the oceans such that any failed rocket launch or explosion during flight is unlikely to occur near, or spread debris onto, human population centers and infrastructure.

Color-labeled launch outcomes on the map



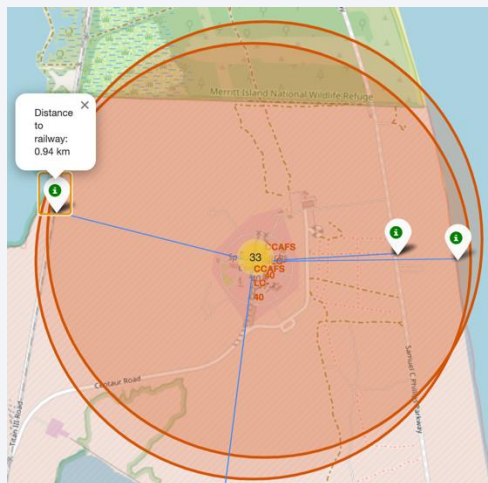
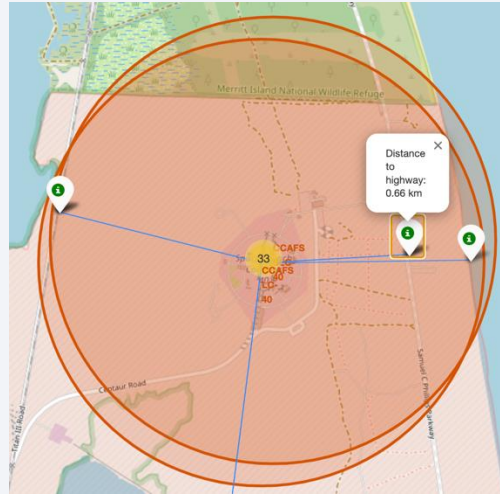
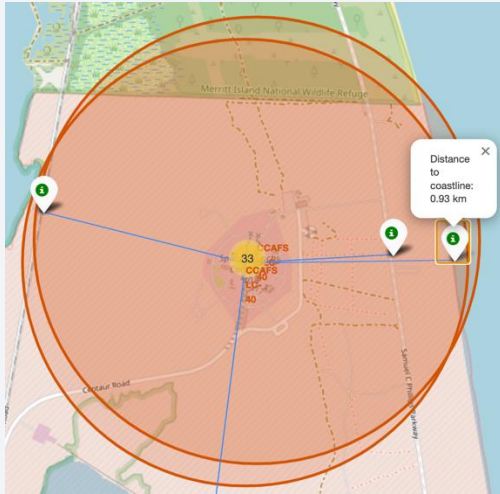
Site KSC LC-39A



Site CCAFS LC-40

- Insights from the map:
 - The color-labeled launch outcome markers allow easy identification of the success rate of each launch site.
 - **Green** = success
 - **Red** = failure
 - Shown here, launch site KSC LC-39A has a high success rate, while launch site CCAFS LC-40 has a low success rate.

Proximity of Launch Site CCAFS LC-40 to coast, city, and transportation lines



- Insights from the map:
- Using CCAFS LC-40 as an example launch site, we can see that it is
 - Less than 1km from the nearest coastline, highway, and railway
 - More than 19km away from the nearest city (Cape Canaveral)
- This places the launch site at safe distance from human population centers, while keeping it closely connected to transportation lines and the coast over which rockets are launched.



Section 4

Build a Dashboard with Plotly Dash

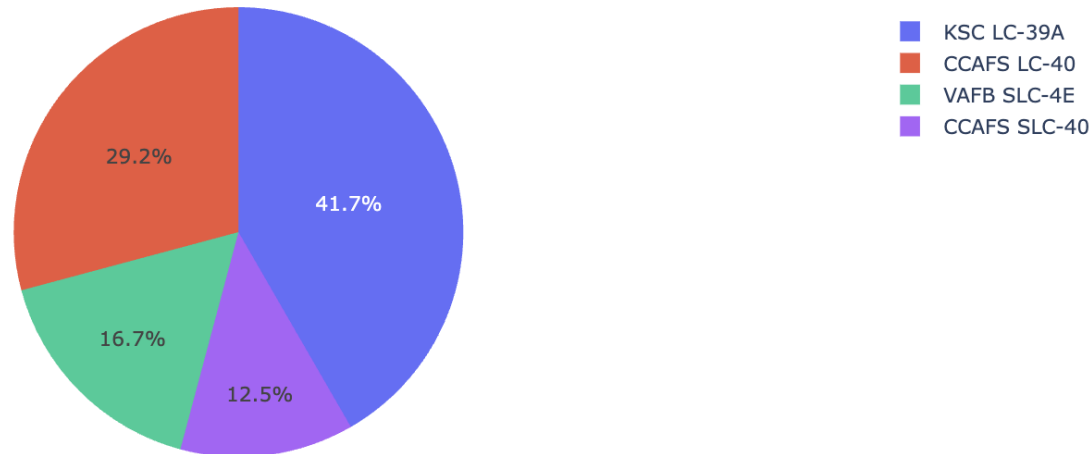
Dashboard: Launch Success of All Sites

SpaceX Launch Records Dashboard

All Sites



Total Successful Launches By Site



- Insights:
 - With the dropdown selector on “All Sites” we see in the pie chart the proportion of total successful launches by launch site
 - Most successful launches are from site KSC LC-39A (41.7%)
 - Then CCAFS LC-40, VAFB SLC-4E, and CCAFS SLC-40 contribute 29.2%, 16.7%, and 12.5% of successful launches, respectively.

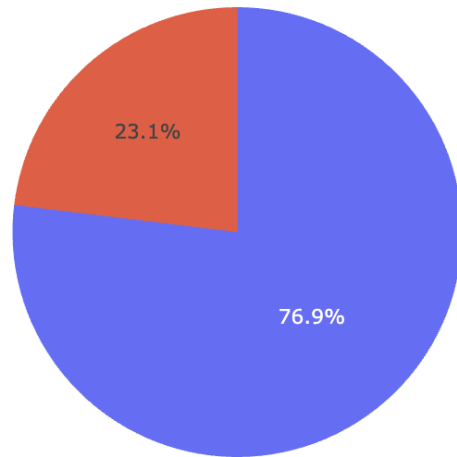
Dashboard: Launch site with highest success ratio

SpaceX Launch Records Dashboard

KSC LC-39A



Success vs. Failed Launches for KSC LC-39A



- Insights:

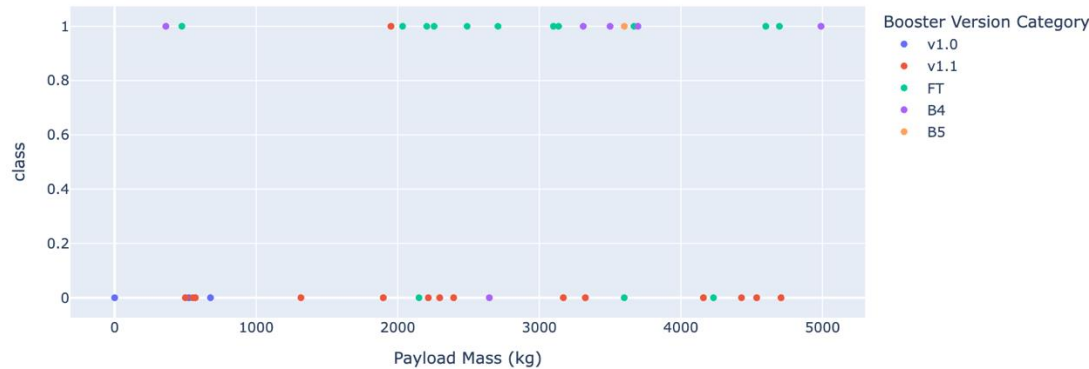
- Launch site KSC LC-39A has the highest ratio of successful launches 76.9% success.
- 23.1% of launches from this site end in failure.

Dashboard: Payload Mass vs Launch Outcome for all sites

Payload range (Kg):



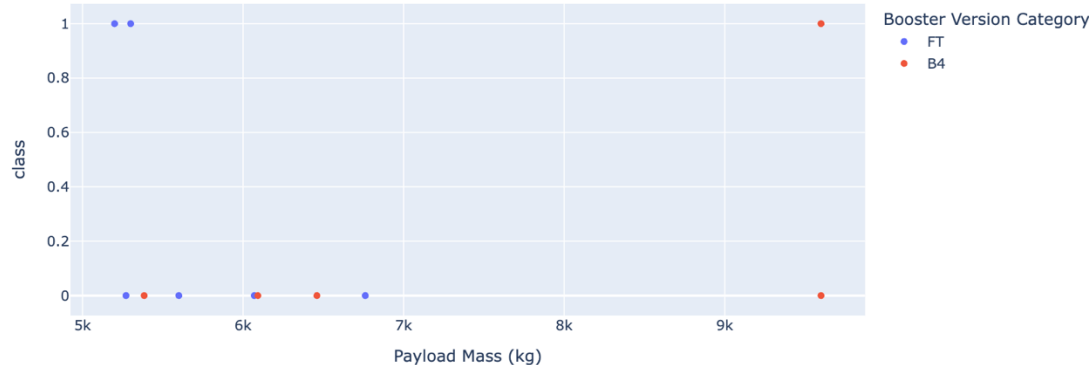
Correlation between Payload and Launch Success (All Sites)



Payload range (Kg):



Correlation between Payload and Launch Success (All Sites)



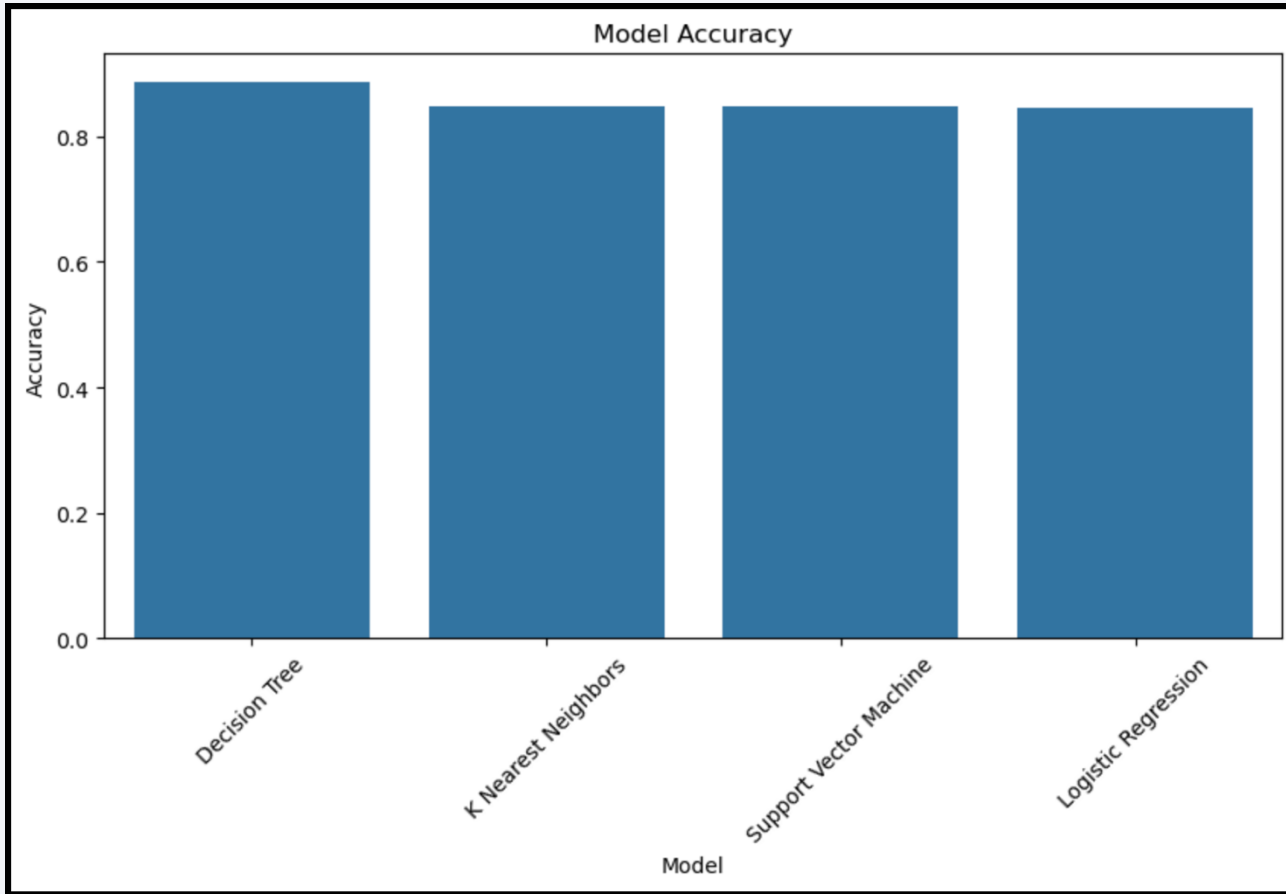
- Insights:

- Comparing the charts, most successful launches have payload masses between 1900kg and 5500kg
- We can also see that booster versions FT (green dots) and B4 (purple dots) have the highest success rates
- The booster versions v1.0 (blue dots) and v1.1 (red dots) have the lowest success rates

Section 5

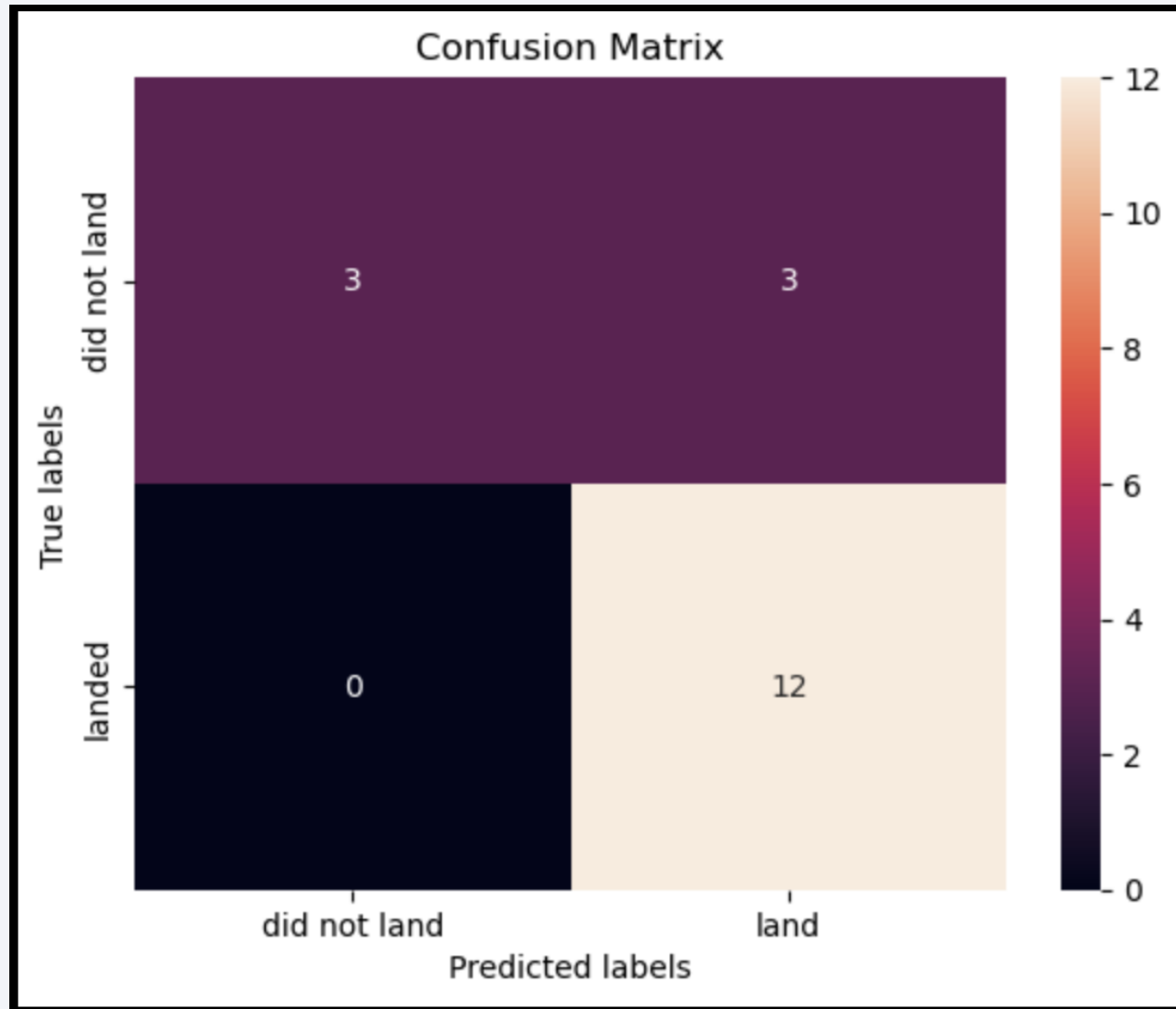
Predictive Analysis (Classification)

Classification Accuracy



- Insights from the bar chart:
 - Comparing the accuracy of the different ML models, we can see that the KNN, SVM and LogReg ML models perform similarly, while the Decision Tree ML model has the highest accuracy.
 - The accuracy of the Decision Tree is close to 90%.

Confusion Matrix



- Insights from the confusion matrix of the Decision Tree model:
 - The Decision Tree has no false negative errors (Predicted 'did not land', Actual 'landed')
 - The Decision Tree however suffers from false positive errors (Predicted 'landed', Actual 'did not land')

Conclusions

- The Decision Tree model is the best for predicting launch success on this dataset, however it suffers from false positives.
- The success rate of SpaceX Falcon9 rocket launches has increased over the years.
- The launch sites analyzed have different success rates, with site KSC LC-39A having the highest launch success rate.
- Launch sites are located close to coastlines, the equator, and major transportation lines, but far from human population centers.
- Most successful launches have lower range of payload mass, though very high payload mass launches have high success rates.
- The majority of launches go to orbits that have success between 50% and 85%.

Appendix

All Python code, SQL queries, charts, and data sets within Jupyter notebooks created during this project can be accessed in the GitHub repository “IBM_Data_Sci_Capstone_Project” at the following URL:

https://github.com/ayexander/IBM_Data_Sci_Capstone_Project

Thank you!

