



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Christopher Cable  
April 12, 2023



# Outline

---

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

# Executive Summary

---

In order to predict whether SpaceX would successfully land the first stage of their next launch, data was collected from the SpaceX API and Wikipedia about SpaceX's previous launches. Exploratory data analysis was performed using various data visualization techniques to identify related variables that had a strong correlation with the success or failure of the first stage landing. Using these insights, target features were identified and used to compile a cleaned dataset. This dataset was then used to train and test various machine learning models in order to establish which model was the most accurate at predicting the success of SpaceX's landing attempts. It was determined that a decision tree classifier was most appropriate for this task.

# Introduction

---

The industry of commercial space launches is a highly competitive emerging field. While the typical cost of a commercial space launch can be upwards of \$165 million, SpaceX (a leading competitor in this field) offers space launches using its Falcon 9 rocket at a price as low as \$65 million. These cost savings come from SpaceX's ability to land and reuse the first stage of their Falcon 9 rockets. Therefore, in order to determine the appropriate contract bid when competing against SpaceX for a commercial launch, we must first predict if SpaceX will land and reuse their launch vehicle. We will accomplish this by creating a predictive model trained on publicly available data from SpaceX's previous missions.





Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection methodology:
  - Data was collected using both the SpaceX API and web scraping techniques to pull data from SpaceX's own publicly available data and from online articles.
- Perform data wrangling
  - Data was processed by preliminary exploratory analysis. Once correlations were discovered and established the irrelevant data was removed and the remaining data was given proper labels.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

# Data Collection

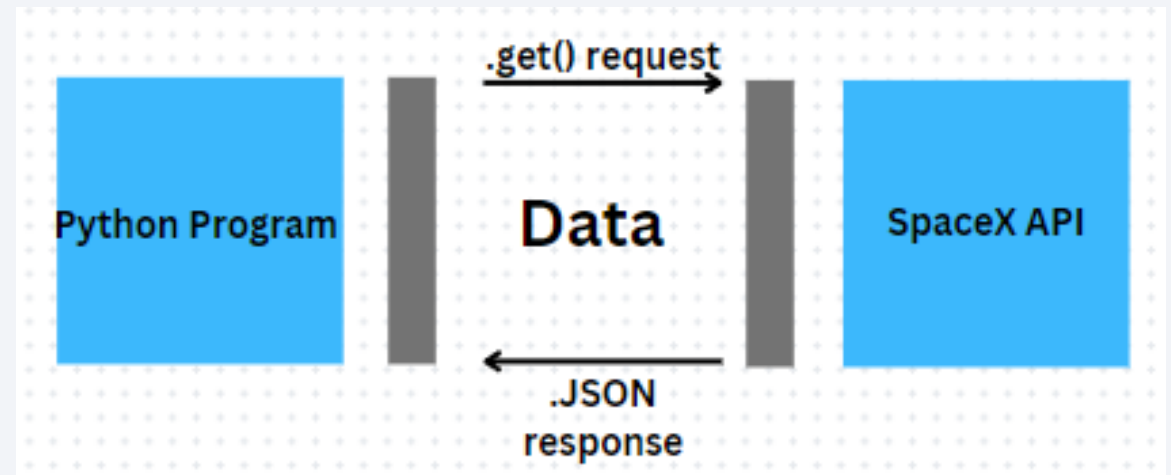
---

During the data collection phase, public data was pulled both directly from SpaceX and from web articles cataloging their previous launch history. Using both an official SpaceX API, as well as basic web scraping techniques, the raw data was gathered into two starting dataframes, before being processed and refined for use in predictive models.

# Data Collection – SpaceX API

---

- Using a `.get()` request, the SpaceX API was called to retrieve data from the SpaceX online database in JSON format. It was then normalized and converted into a dataframe.
- API = “<https://api.spacexdata.com/v4/launches/past>”
- GitHub URL:  
<https://github.com/cbcable/Capstone-Project/blob/e0b9cd2fce0e67726b59ea319e1d7796970ead66/Part%201%20Data%20Collection%20API%20Lab.ipynb>

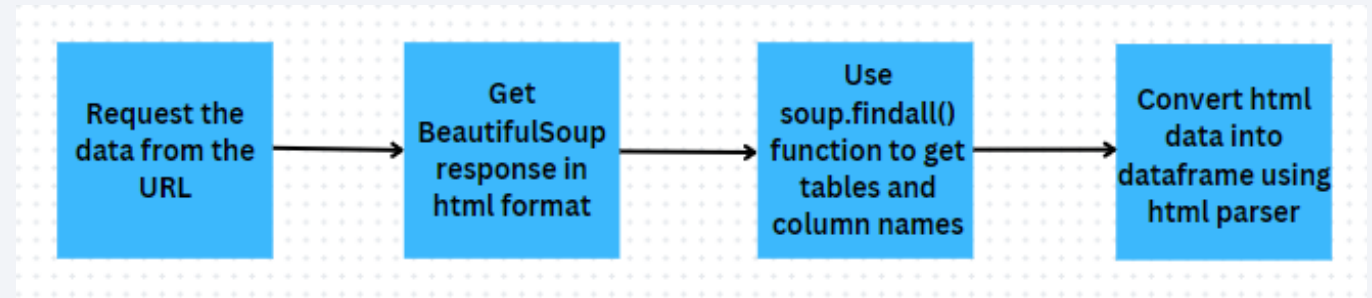




# Data Collection - Scraping

---

- Using BeautifulSoup, data was scraped from the Wikipedia page covering the Falcon 9 and Falcon Heavy launches in tabular format.
- [https://github.com/cbcable/Capstone-Project/blob/6afe218e19f8fc4fe62dde2ad6eb893e8d9ce435/Part%202\\_%20Data%20Web scraping.ipynb](https://github.com/cbcable/Capstone-Project/blob/6afe218e19f8fc4fe62dde2ad6eb893e8d9ce435/Part%202_%20Data%20Web scraping.ipynb)



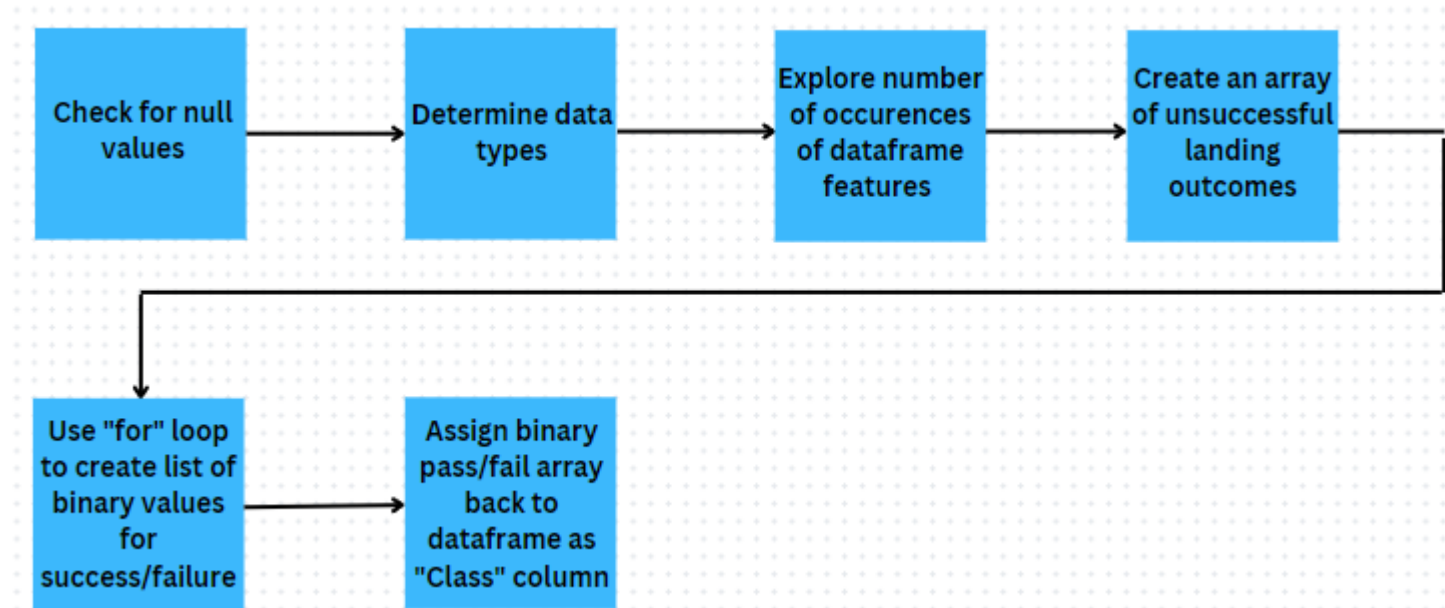
# Data Wrangling

---

The first step in processing the data was to check for null values and determine the data types of each column. We then explore the occurrence counts of various features, such as orbit type and total successful landing outcomes. We then create an array of unsuccessful landing outcomes, which we use to check against the dataframe using a “for” loop to create a list of binary values where a successful landing is equal to 1 and an unsuccessful landing is equal to 0. This list was then appended to the dataframe under the column key “Class”.

# Data Wrangling

- [https://github.com/cbcable/Capstone-Project/blob/6afe218e19f8fc4fe62dde2ad6eb893e8d9ce435/Part%2003\\_%20Data%20Wrangling.ipynb](https://github.com/cbcable/Capstone-Project/blob/6afe218e19f8fc4fe62dde2ad6eb893e8d9ce435/Part%2003_%20Data%20Wrangling.ipynb)



# EDA with Data Visualization

---

During the data visualization process, scatter plots and bar charts were employed to visualize relationships in the data. Scatter plots were used to visualize the correlation between launch site and flight number, between launch site and payload mass, and between payload mass and flight number. A bar chart was then used to display the average success rate for each orbit type, which revealed the top four orbit types with the most successful launches.

[https://github.com/cbcable/Capstone-Project/blob/54c53aaf0a8ef75fd405b022b9597b5783af3564/Part%205\\_%20EDA%20with%20Pandas%20and%20Matplotlib.ipynb](https://github.com/cbcable/Capstone-Project/blob/54c53aaf0a8ef75fd405b022b9597b5783af3564/Part%205_%20EDA%20with%20Pandas%20and%20Matplotlib.ipynb)

# EDA with SQL

---

SQL queries were used to explore the data. These queries were as follows:

- `Select distinct(LAUNCH_SITE) from CAPSTONE.SPACEX;`
- `Select * from CAPSTONE.SPACEX where LAUNCH_SITE like '%CCA%' limit 5;`
- `Select sum(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD_MAS_KG from CAPSTONE.SPACEX where CAPSTONE.SPACEX.CUSTOMER like '%NASA (CRS)%';`
- `Select avg(PAYLOAD_MASS__KG_) as AVERAGE_MASS from CAPSTONE.SPACEX where CAPSTONE.SPACEX.BOOSTER_VERSION like '%F9 v1.1%';`
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose



# EDA with SQL

- Select BOOSTER\_VERSION, PAYLOAD\_MASS\_\_KG\_, "Landing\_Outcome" from CAPSTONE.SPACEX where "Landing\_Outcome" like '%Success (drone ship)%' and PAYLOAD\_MASS\_\_KG\_ between 4000 and 6000;
- Select count(MISSION\_OUTCOME) as TOTAL\_MISSIONS from CAPSTONE.SPACEX;
- Select BOOSTER\_VERSION from CAPSTONE.SPACEX where PAYLOAD\_MASS\_\_KG\_ = (select max(PAYLOAD\_MASS\_\_KG\_) from CAPSTONE.SPACEX);
- Select BOOSTER\_VERSION, LAUNCH\_SITE, "Landing\_Outcome", DATE from CAPSTONE.SPACEX where "Landing\_Outcome" like '%Failure (drone ship)%' and DATE like '%2015%';
- Select distinct("Landing\_Outcome") as OUTCOMES, count("Landing\_Outcome") as NUM\_OCCURENCES from CAPSTONE.SPACEX where DATE between '2010-06-04' and '2017-03-20';

# EDA with SQL

GitHub Link

[https://github.com/cbcable/Capstone-Project/blob/54c53aaf0a8ef75fd405b022b9597b5783af3564/Part%204\\_%20Exploratory%20Analysis%20Using%20SQL%20\(1\).ipynb](https://github.com/cbcable/Capstone-Project/blob/54c53aaf0a8ef75fd405b022b9597b5783af3564/Part%204_%20Exploratory%20Analysis%20Using%20SQL%20(1).ipynb)

# Build an Interactive Map with Folium

---

Using Folium, maps were generated to display the launch sites. Map markers were added to each launch site to show the successful and failed launches for each locations, which were then grouped into marker clusters. Next, the distance from one launch site to the nearest coast, town, highway, and rail line was calculated and displayed using a line. This created a simple-to-use interactive map that visualized the success rate of each launch site, its geographic location, and the relative distance from major landmarks. This allowed for a better understanding of the possible relationship between a launch site's geographic location and proximity to various transportation resources, and its success rate.

- [https://github.com/cbcable/Capstone-Project/blob/54c53aaf0a8ef75fd405b022b9597b5783af3564/Part%206\\_%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb](https://github.com/cbcable/Capstone-Project/blob/54c53aaf0a8ef75fd405b022b9597b5783af3564/Part%206_%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb)

# Build a Dashboard with Plotly Dash

---

Using Plotly Dash, I created a dashboard to graphically compare launch factors. First, I used a pie chart to show the successful launch rate of each launch site, both together and individually. Next, the dashboard included a scatter plot that depicted the successful and failed launches for each booster version at a given payload, to see which boosters performed best at a certain lift mass. A data slicer was added in the form of a slider bar to select the payload range the user wished to examine, which filtered the data and displayed only data point that fell inside the specified payload range.

- <https://github.com/cbcable/Capstone-Project/blob/a5d8c4b13d46a81266d0231856088cce03d95186/Part%207%20Dash%20Interactivity.py>

# Predictive Analysis (Classification)

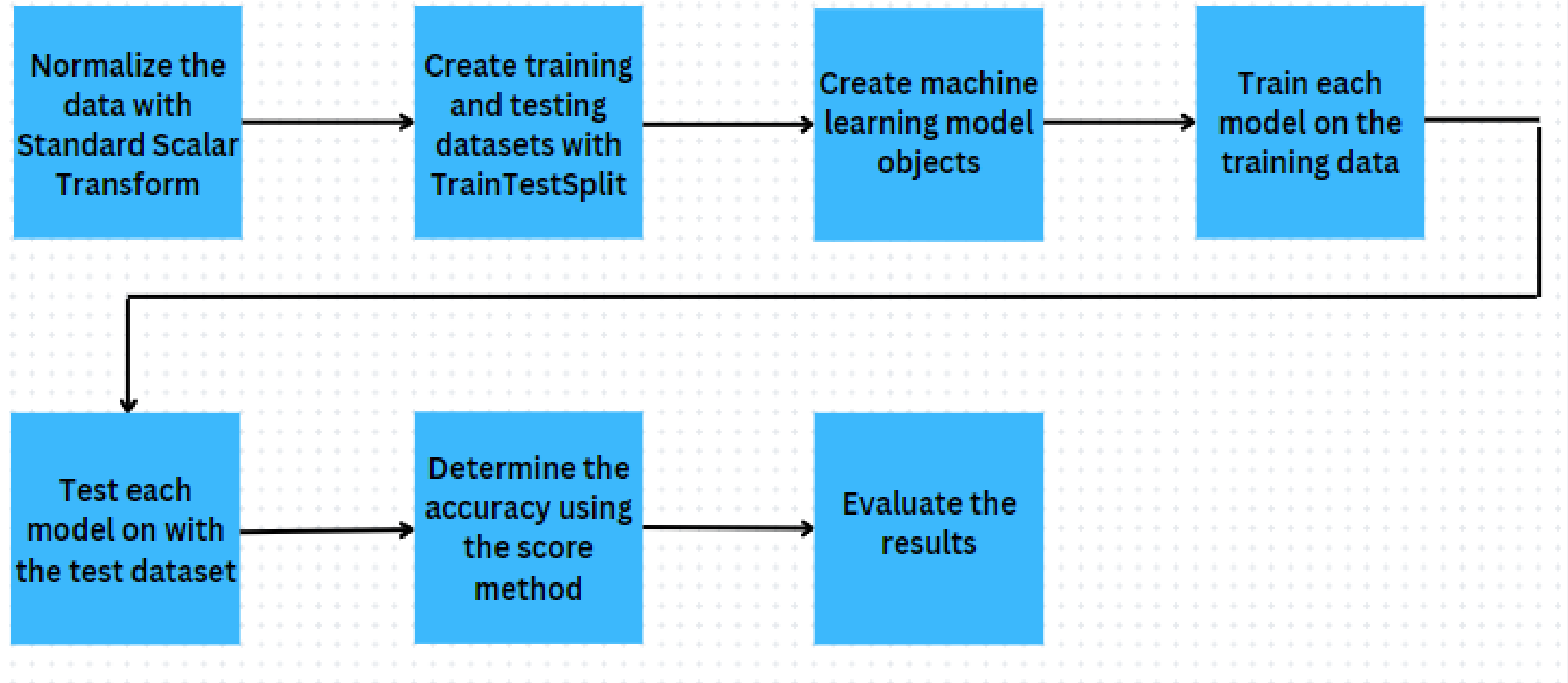
---

In order to build our model, we first normalized the data using a Standard Scalar transform, then divided the resulting dataset into training and testing groups using TrainTestSplit. Next, we created four different machine learning models (Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbor) and trained each model on the training data. We then tested each model using the test set, and evaluated the accuracy of the results using the score method. Based on the results of each model's accuracy score, it was determined that the Decision Tree Classifier was the most accurate model for this application.

- [https://github.com/cbcable/Capstone-Project/blob/a5d8c4b13d46a81266d0231856088cce03d95186/Part%208\\_%20Machine%20Learning%20Prediction%20Lab.ipynb](https://github.com/cbcable/Capstone-Project/blob/a5d8c4b13d46a81266d0231856088cce03d95186/Part%208_%20Machine%20Learning%20Prediction%20Lab.ipynb)



# Predictive Analysis (Classification)





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 red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

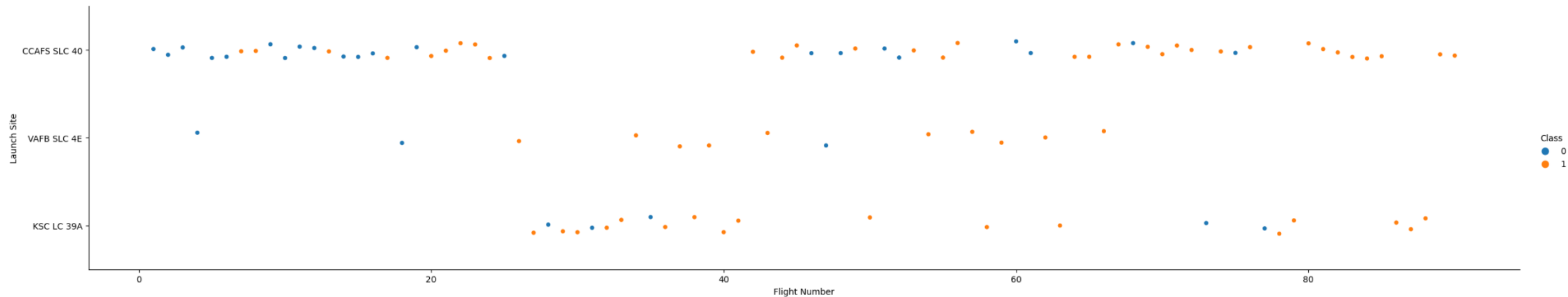
Section 2

# Insights drawn from EDA



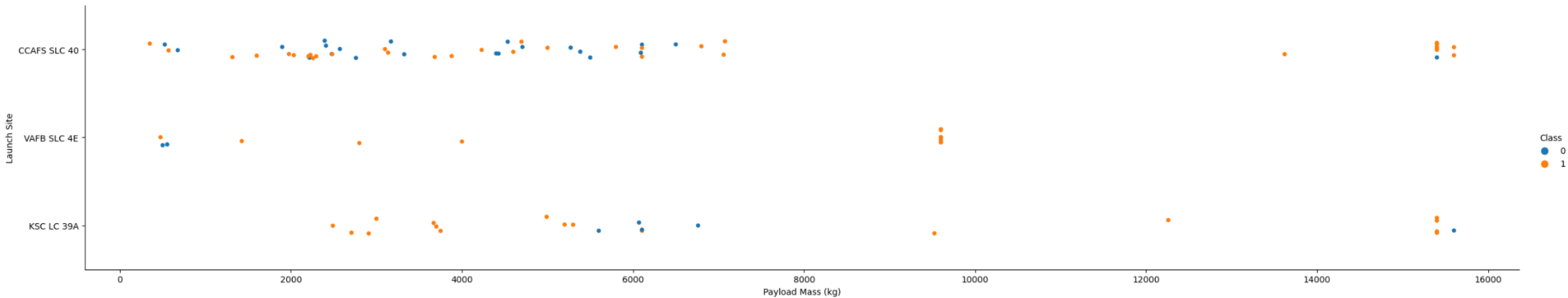
# Flight Number vs. Launch Site

As shown in the scatterplot below, the success rate of each launch site increases as the flight number increases, indicating a strong correlation that can be used in modeling. This can be attributed to the refining of the landing technology over time with each successive attempt. That said, KSC LC-39A has the highest success rate for all launches at that location.



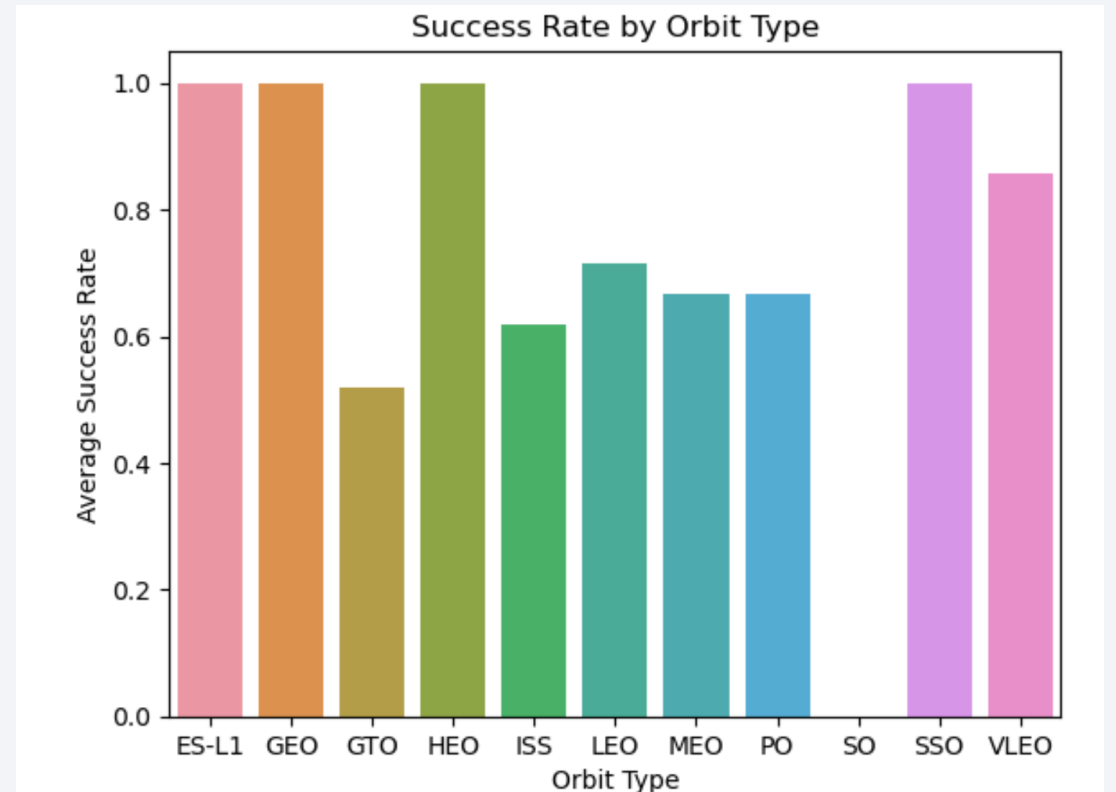
# Payload vs. Launch Site

As seen in the graph, VAFB SLC-4E has not been used for any payloads greater than 10,000kg, and CCAFS SLC was heavily utilized for payloads under 8000kg.



# Success Rate vs. Orbit Type

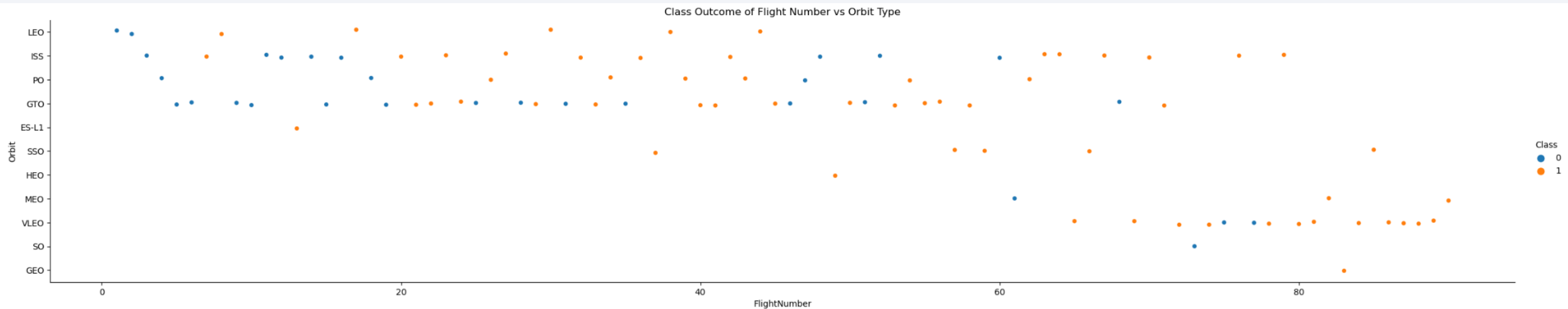
As seen, launches to different orbits have different success rates. The orbits with the highest rates of success are ES-L1, GEO, HEO, and SSO with nearly 100% average success rate. By contrast, the least successful orbit to launch to has been GTO, with an average success rate of about 50%.





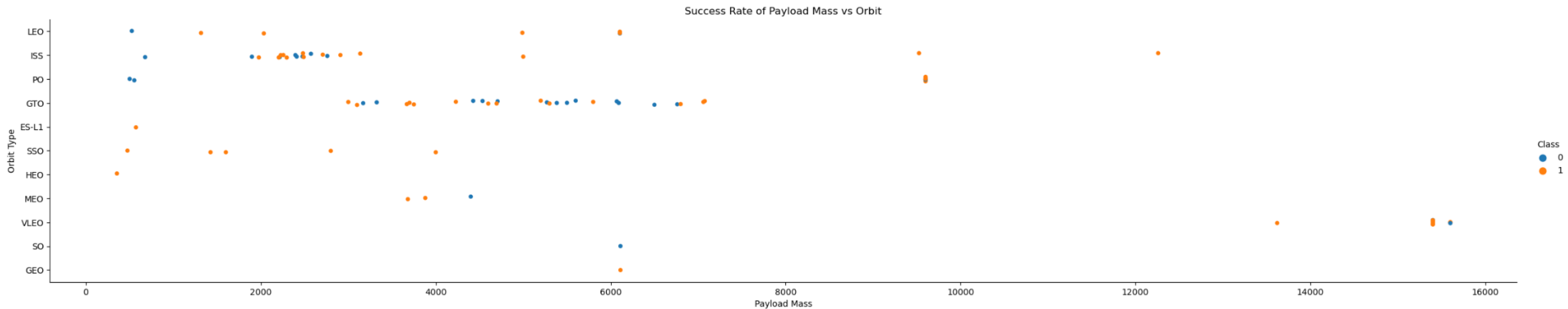
# Flight Number vs. Orbit Type

As seen in the scatterplot below, orbits such as GEO, SO, VLEO, and MEO were not used for early launches. Rather, the predominant orbits of early launches were GTO and ISS. The data shows a mix of successful and failed launches for each orbit type, even in later flight numbers, which does not indicate any strong correlation emerging in this comparison.



# Payload vs. Orbit Type

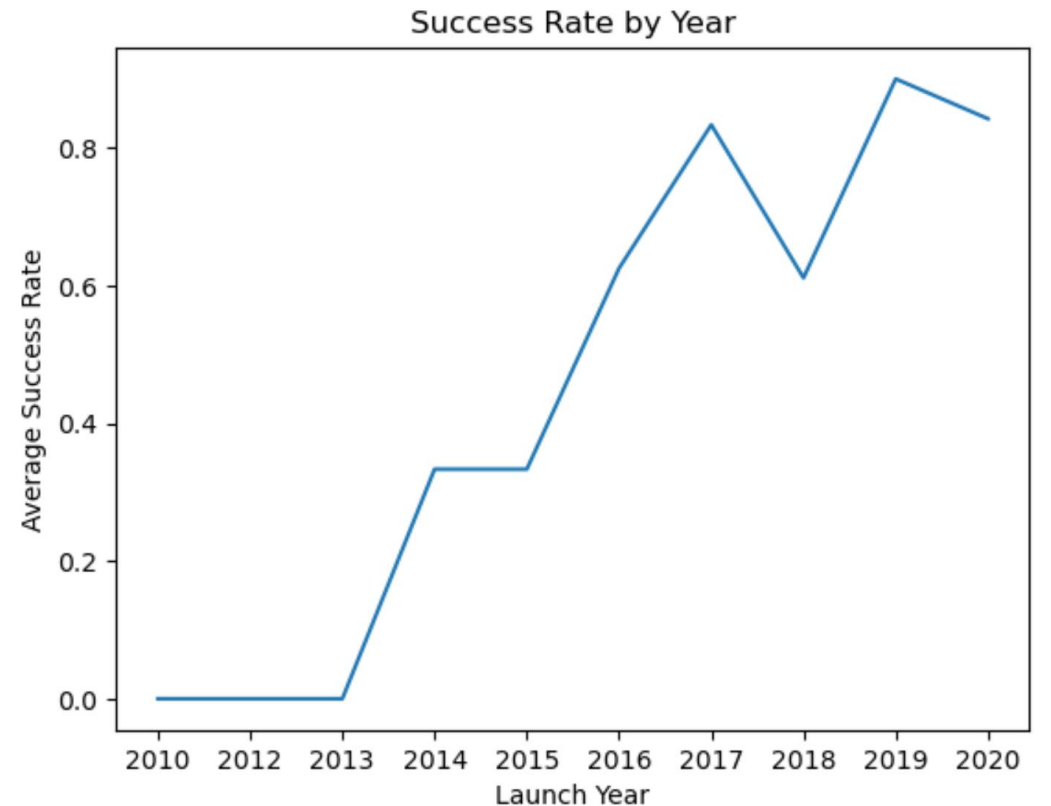
Here we see that the success rate with heavy payloads is greater for Polar, LEO, and ISS orbits. However, with a GTO orbit the data is a mix of positive and negative outcomes, showing that payload mass is not universally correlated with orbit type.



# Launch Success Yearly Trend

---

Here we see that the calculated average yearly success rate increases over time. We noticed this trend in earlier analyses, now we can plot it directly. Therefore, the launch year and the average success rate have a clear positive correlation.



# All Launch Site Names

---

Using the following SQL query, we were able to find the names of all launch sites:

```
select distinct(LAUNCH_SITE) from CAPSTONE.SPACEX;
```

This returns the following list of distinct launch site names:

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

# Launch Site Names Begin with 'CCA'

Using the above launch site names, we are able to query the database to filter for records where the launch site name includes 'CCA' (which will return values for CCAFS LC and SLC).

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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



# Total Payload Mass

---

In order to find the total payload mass carried by NASA, for example, we can use the following query:

```
%%sql
select sum(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD_MASS_KG from CAPSTONE.SPACEX
where CAPSTONE.SPACEX.CUSTOMER like '%NASA (CRS)%';
```

This returns the following calculated value, which represent the total cumulative payload mass (in kg) launched by NASA:

total_payload_mass_kg
-----------------------

48213
-------

# Average Payload Mass by F9 v1.1

---

To calculate the average payload mass carried by the Falcon 9 v1.1, we use the following SQL query:

```
%%sql
select avg(PAYLOAD_MASS__KG_) as AVERAGE_MASS from CAPSTONE.SPACEX
where CAPSTONE.SPACEX.BOOSTER_VERSION like '%F9 v1.1%';
```

This filters the database for records where the booster version includes the label F9 v1.1, and averages the payload mass of those records to return the following value:

**average\_mass**

2534

# First Successful Ground Landing Date

---

Next, we want to find the first occurrence of a successful ground landing. We use the following query to filter the data records for entries where the landing outcome is 'Success (ground pad)', then use a "min" function on the date column to find the first instance:

```
%%sql  
select min(DATE) as FIRST_GROUND_SUCCESS from CAPSTONE.SPACEX where "Landing _Outcome" like '%Success (ground pad)%';
```

This returns the following date:

first_ground_success
2015-12-22

## Successful Drone Ship Landing with Payload between 4000 and 6000

---

Now we want to find the names of the booster versions that have successfully launched payloads between 4000 and 6000 kilograms. We use the following query to filter the data for results matching our criteria:

```
%%sql
select BOOSTER_VERSION, PAYLOAD_MASS__KG_, "Landing _Outcome" from CAPSTONE.SPACEX
where "Landing _Outcome" like '%Success (drone ship)%'
and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

This returns the following list of booster versions:

booster_version	payload_mass_kg_	Landing _Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)


# Total Number of Successful and Failure Mission Outcomes

---

To check the total number of successful and failed mission outcomes, we filter the Mission Outcomes for successful missions and count the number of occurrences, as follows:

```
select count(MISSION_OUTCOME) from CAPSTONE.SPACEX where MISSION_OUTCOME like '%Success%';
```

This query informs us that of the total number of missions flown, 100 of them were successful flights.

Result set 1	
	Filter table
1	
100	

# Boosters Carried Maximum Payload

---

To find the booster versions that have been used to carry the maximum payload mass, we employ a subquery in the following manner:

```
%%sql
select BOOSTER_VERSION from CAPSTONE.SPACEX
where PAYLOAD_MASS_KG_ = (select max(PAYLOAD_MASS_KG_) from CAPSTONE.SPACEX);
```

This applies a filter to the data to return values from the booster version column where the payload mass is equal to the maximum value. The query returns the following list of booster versions, which can be seen on the next slide.

# Boosters Carried Maximum Payload

## 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



# 2015 Launch Records

---

In order to pull the launch records from 2015, we use two concurrent similarity filters in our SQL query, as follows:

```
%%sql
select BOOSTER_VERSION, LAUNCH_SITE, "Landing _Outcome", DATE from CAPSTONE.SPACEX
where "Landing _Outcome" like '%Failure (drone ship)%'
and DATE like '%2015%'
```

This returns a list of entries where the Landing \_Outcome and DATE columns contain the specified strings.

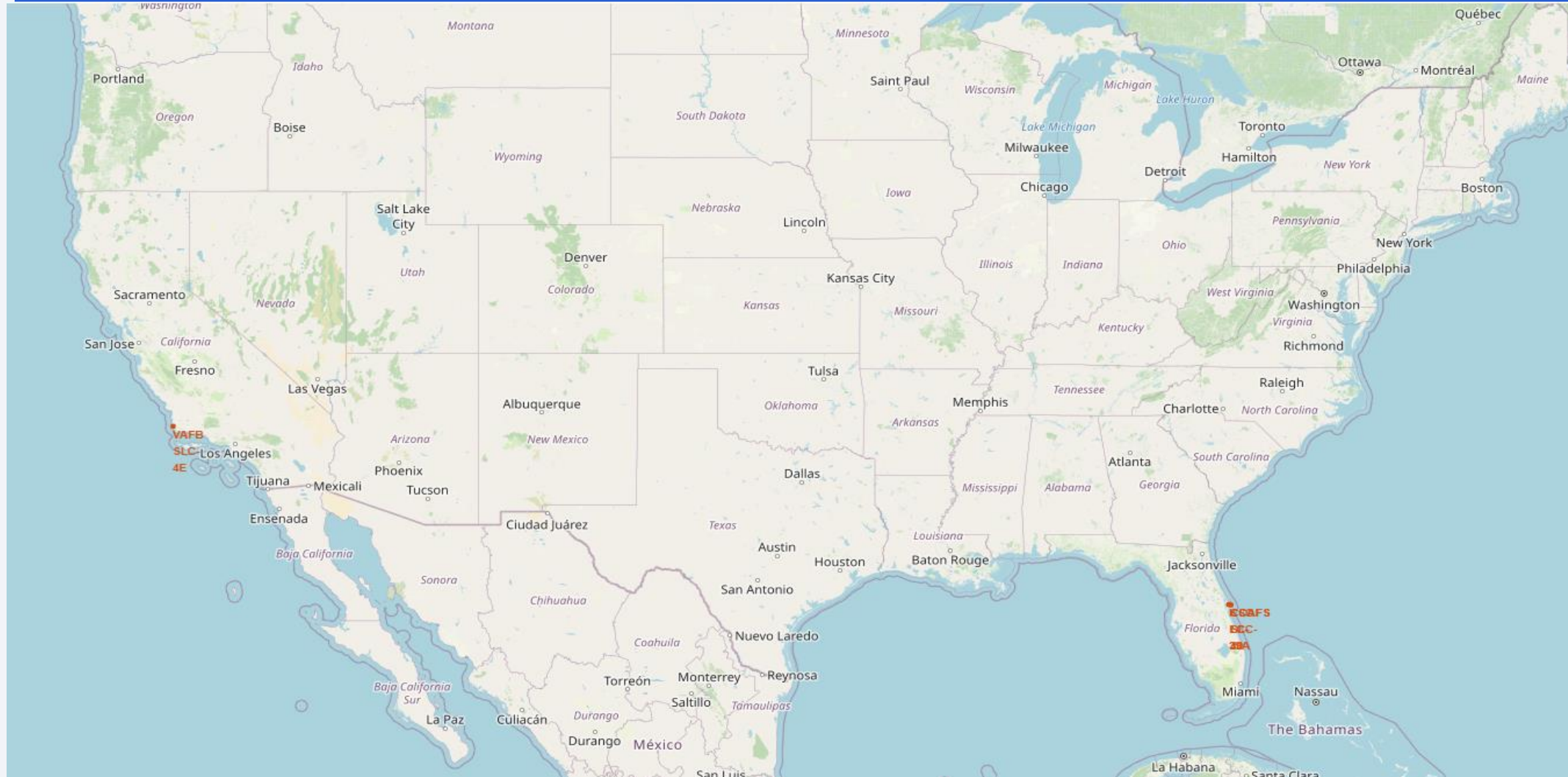
booster_version	launch_site	Landing _Outcome	DATE
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2015-01-10
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	2015-04-14

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

# Launch Sites Proximities Analysis

# Global Launch Site Locations



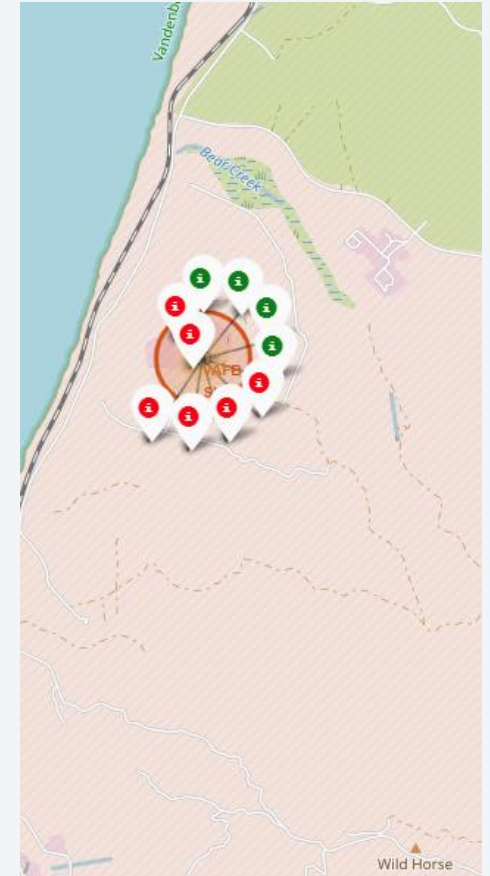
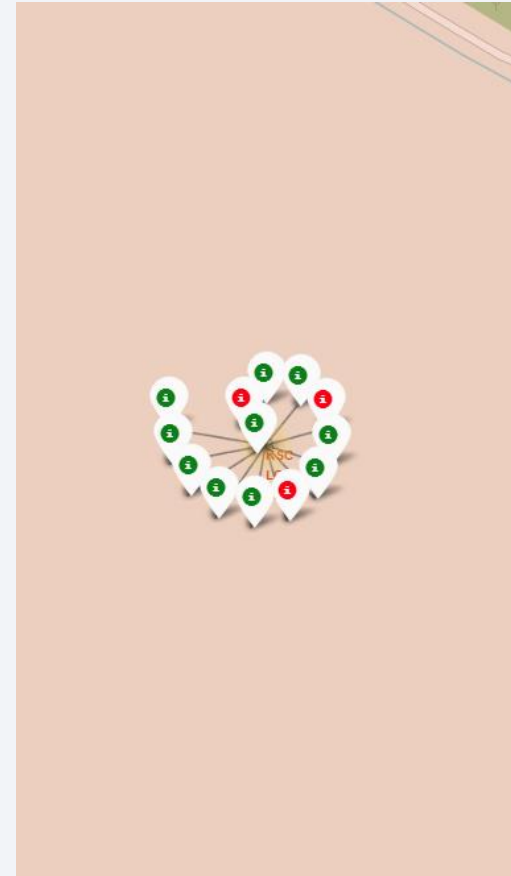
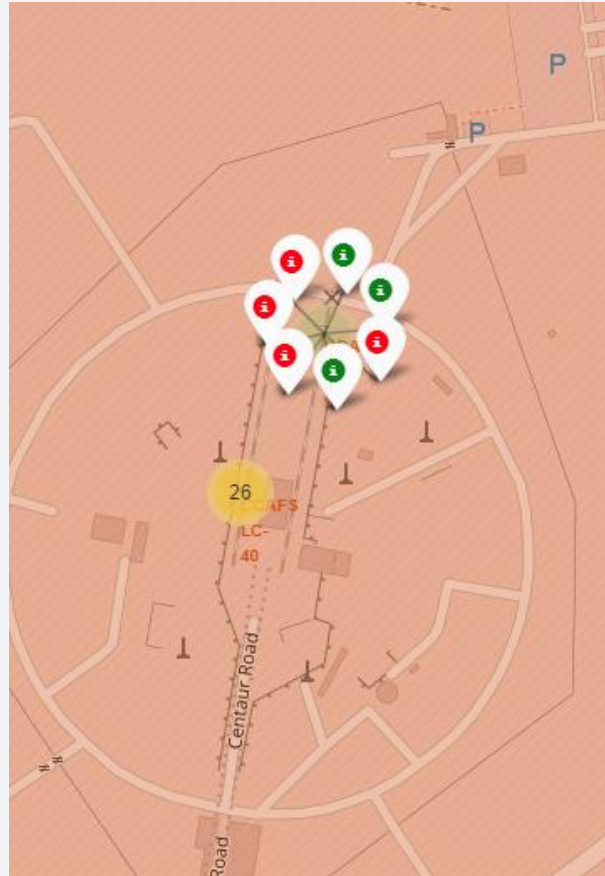
# Global Launch Site Locations

---

As seen in the map above, the launch sites are located at the extreme East and West ends of the country, on the coastline. This provides easy access by sea for drone ship recovery as well as transportation to and from the launch sites.



# Color Coded Launch Records by Site



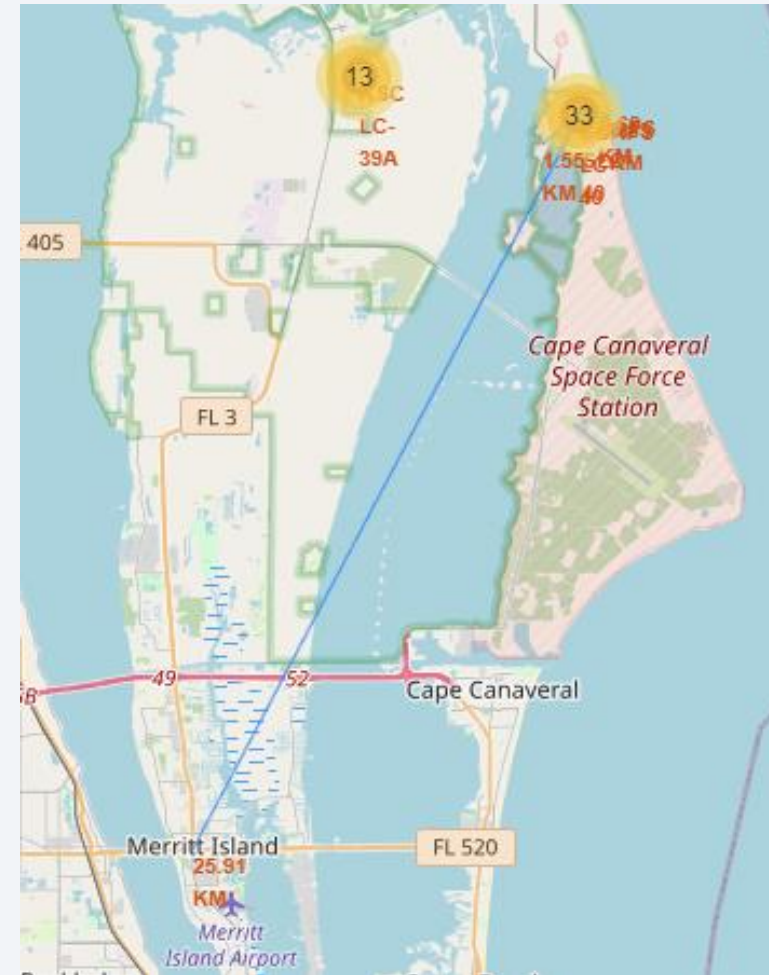
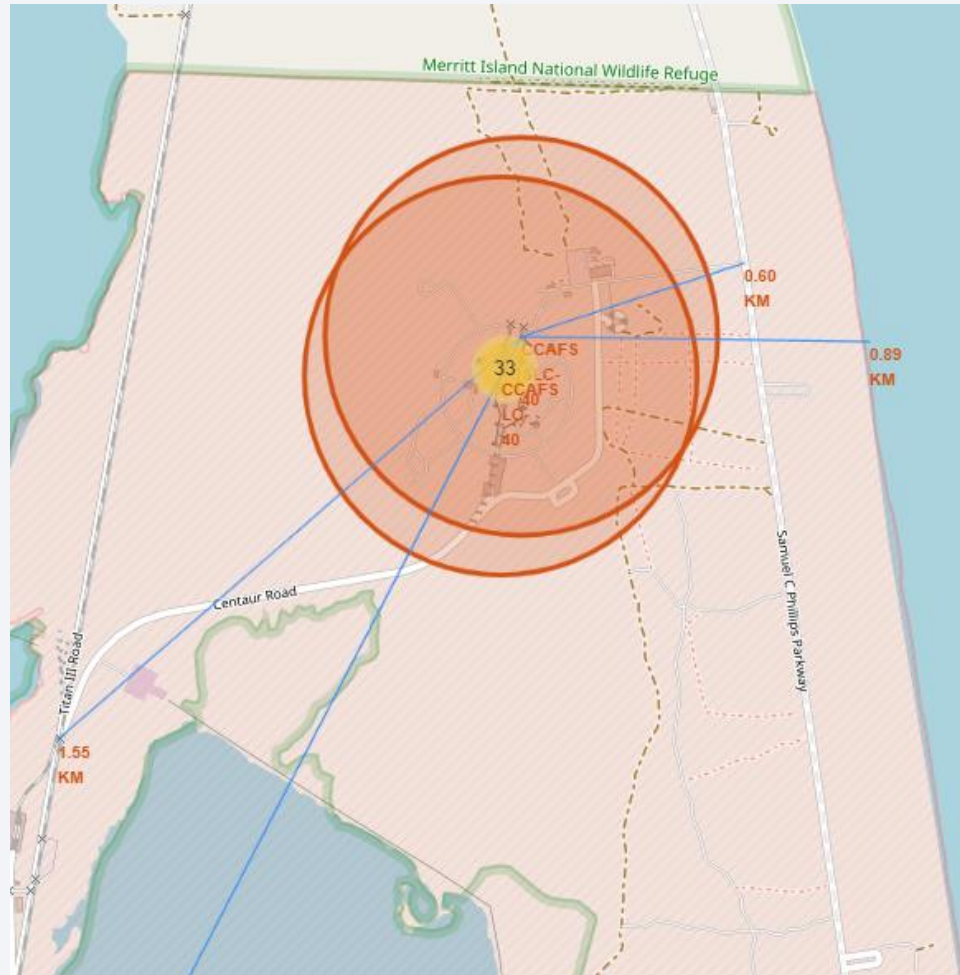
# Color Coded Launch Records by Site

---

Here we have a visual representation of the successful and failed launches conducted at each launch site, as represented by red and green icons clustered at each location. This gives a quick, at-a-glance representation of the relative success rates of each launch site.



# Distances From Major Landmarks



# Distances From Major Landmarks

---

As seen in the map images above, launch sites are typically located very close to major access point, such as roads, rail lines, and sea. The above example, Cape Canaveral, is also located only 25km from the nearest town of Merritt Island, which increases the access to labor and material imports accessible to the launch site.

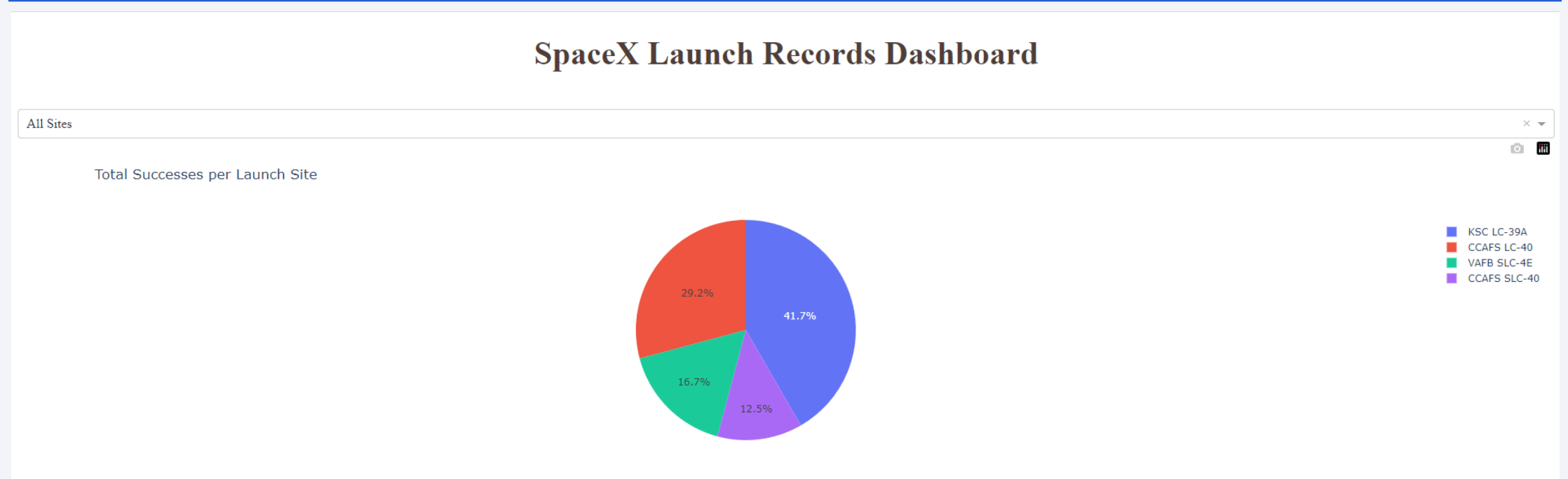


Section 4

# Build a Dashboard with Plotly Dash

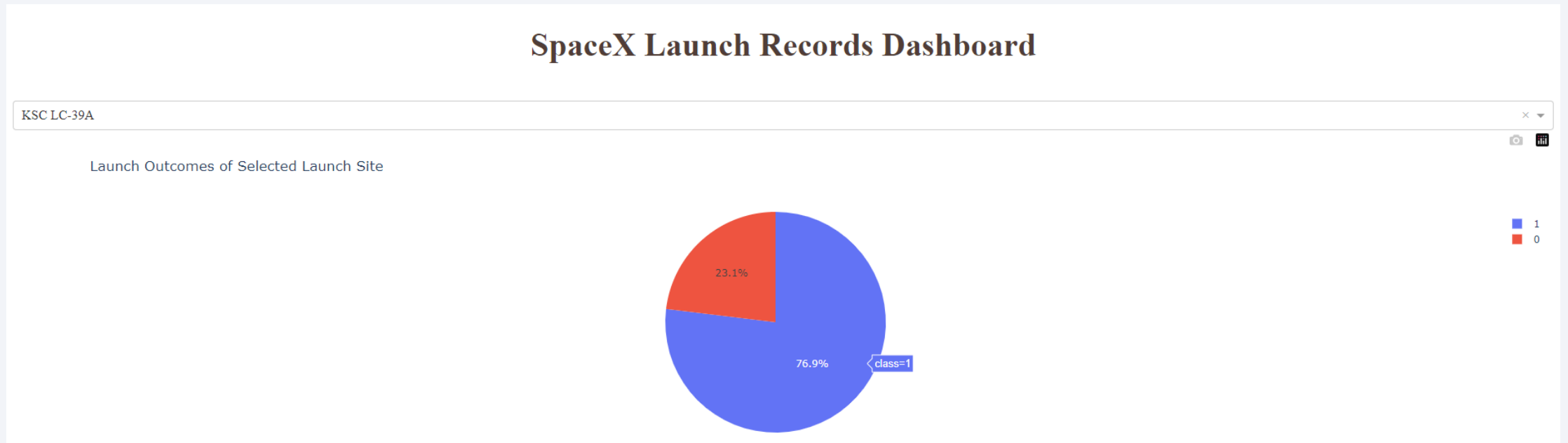


# Total Launch Successes for All Sites



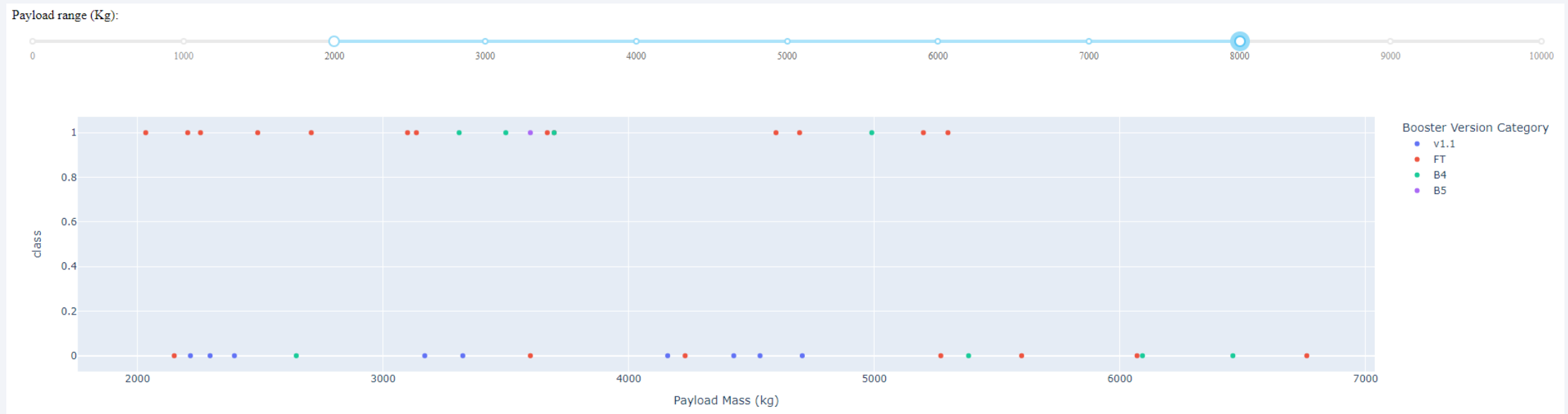
Here we see that at the full payload range KSC LC-39A has the highest success rate of all four launch sites. Launches coming from that location are 12.5% more likely to be successful than the those from the next most successful site, which is CCAFS LC-40.

# Highest Launch Success Ratio



On further inspection, we see that KSC LC-39A have an overall success rate of 76.9%, compared to the next highest total success rate of 64.7%.

# Payload vs. Launch Outcome



Here we see that in the payload range from 2000kg – 8000kg, the Falcon 9 FT booster has the highest success rate of all four booster versions.

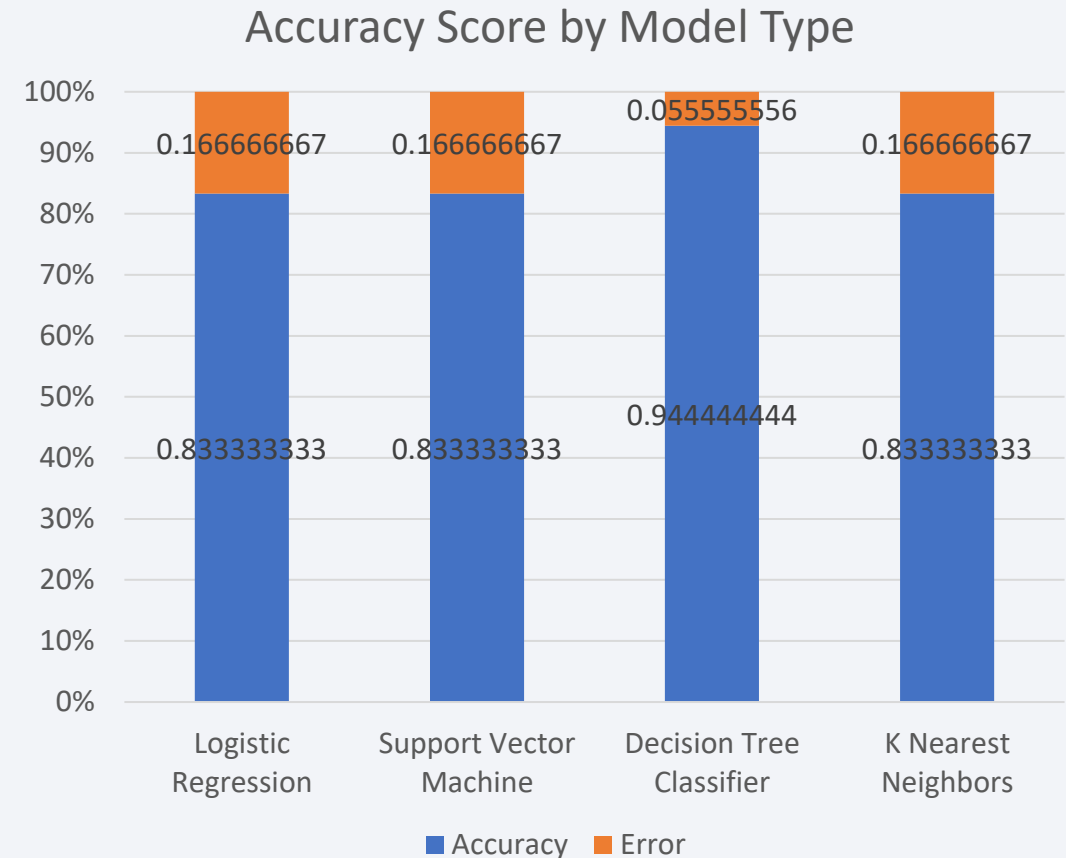


Section 5

# Predictive Analysis (Classification)

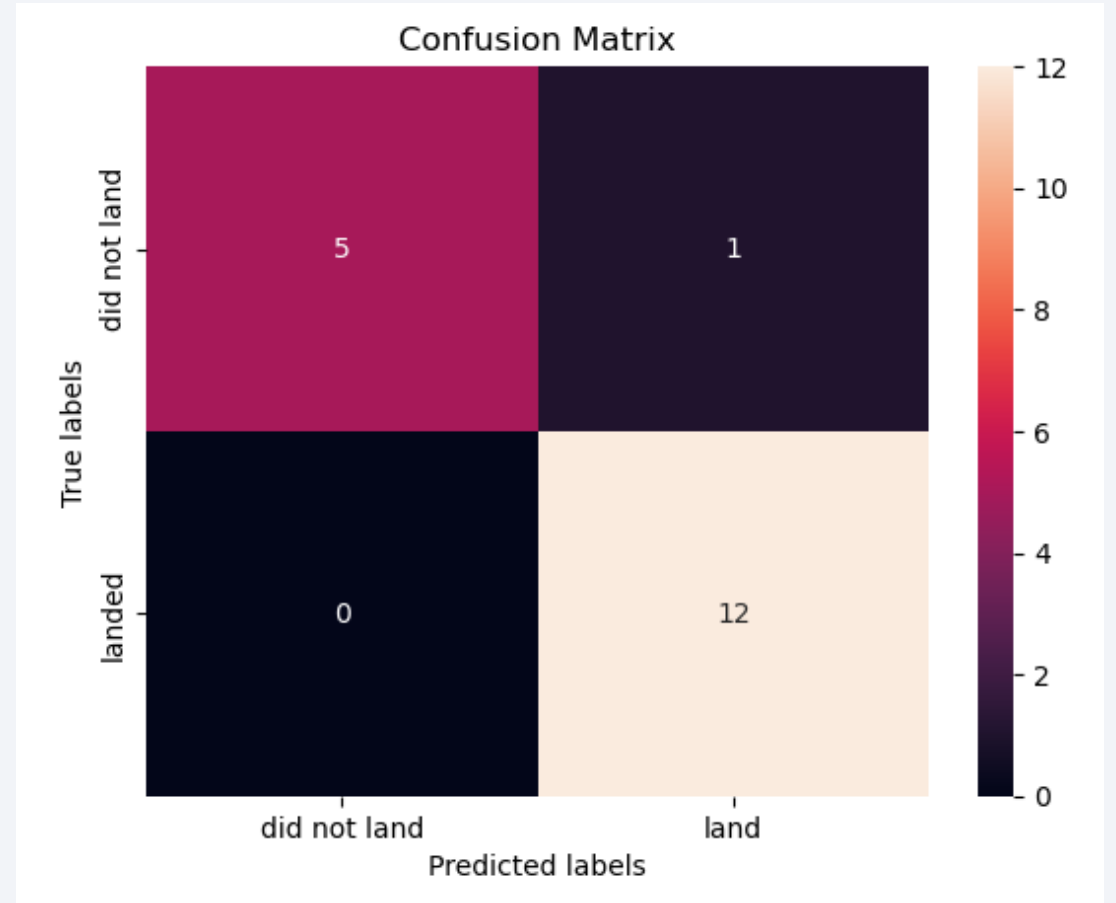
# Classification Accuracy

After training and testing each of the four model types, we calculated the accuracy score and determined that the decision tree classifier model was the most accurate choice, with nearly 11% more accurate predictions than the other three models.



# Confusion Matrix

As seen in the confusion matrix on the right, the decision tree classifier accurately predicted nearly all cases, with the exception of one false positive.



# Conclusions

---

- A decision tree classifier is the best model for predicting whether SpaceX will land the booster from their next launch.
- Key factors in this prediction will include payload mass, booster version, target orbit, number of previous booster reuses, etc.
- Planned launch site will also need to be known ahead of time.

Thank you!

