Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection Through APIs and Web Scraping
  - Data Cleansing and Wrangling
  - Data Visualization, Exploration and Interpretation
  - Machine Learning Predictions
- Summary of all results
  - Exploratory Data Analysis Result
  - Interactive Analytics in Screenshots
  - Predictive Analytics Result

Introduction

SpaceX provides orbital delivery services through their Falcon 9 rockets. These services cost about $62 million while other providers cost up to $165 million. Much of the savings from SpaceX is their creative design that allows them to reuse the first stage of the rocket. As such, if we can determine the probability of the first stage landing, we can predict the cost for a launch. This information may be used to bid against SpaceX. The goal of this project is to understand probability factors and identify the success of the SpaceX business model.

- Problems for which we want to find answers
  - What factors determine if the rocket will land successfully?
  - How do different features of the data interact that determine a success rate?
  - What operating conditions provide a successful landing rate?

Section 1

Methodology

Methodology

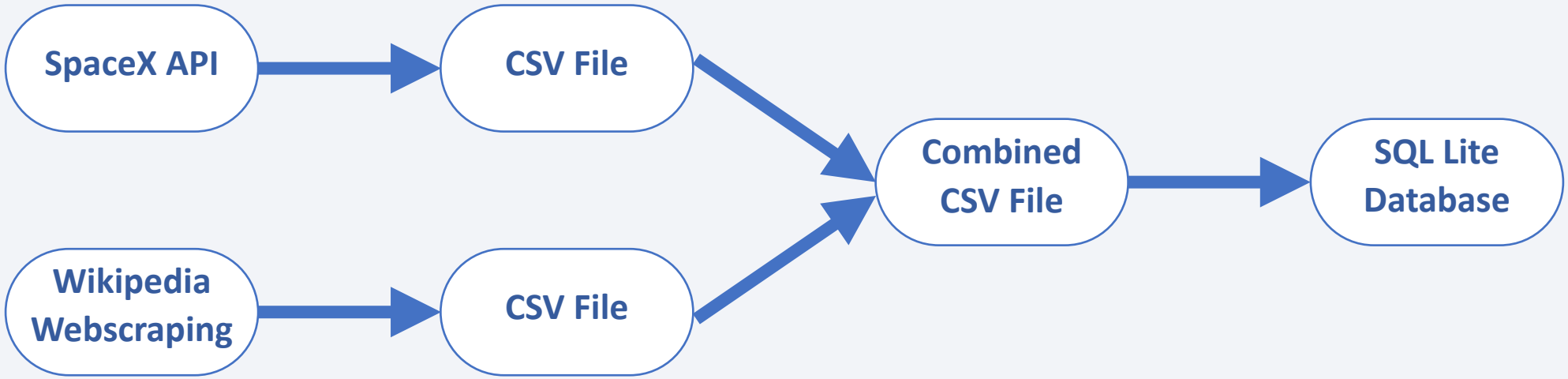
Executive Summary

- Data collection methodology:
  - Data was collected through an API to SpaceX data, through scraping Wikipedia.
- Perform data wrangling
  - Missing data was synthesized through mean values, impertinent data was pruned, all data was normalized and formatted as needed.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - SpaceX Data was loaded into a SQL Lite database from which SQL Alchemy was used to query and gain quick KPIs.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Data was visualized through the use of Plotly Dash to demonstrate the number of successful missions and the ratio of success for each launch pad.
- Perform predictive analysis using classification models
  - Data was standardized and split into testing and training data. Nearest Neighbor, Support Vector Machine, Classification Trees and Logistic Regression models were created and tested for the best hyper parameters.

Data Collection

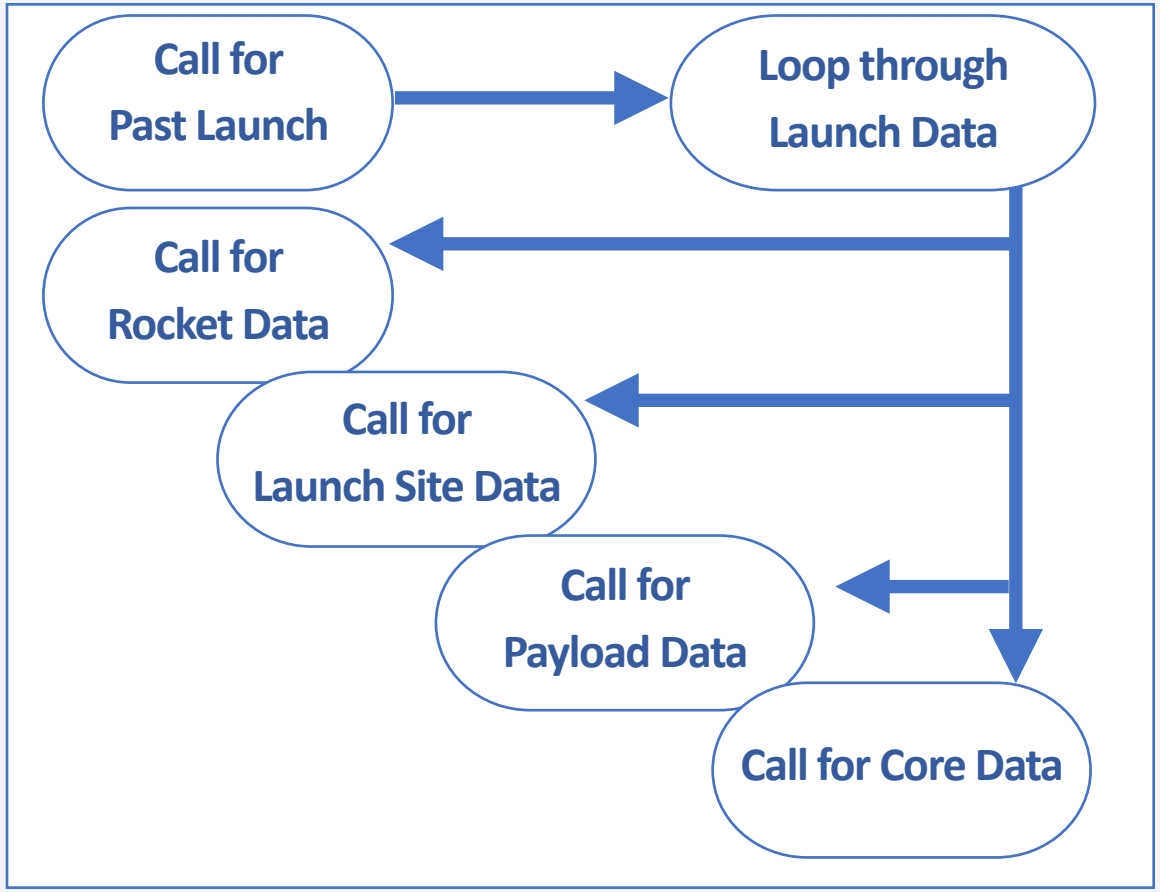
Data was collected through API calls and web scraping.

After some basic cleaning, it was stored in csv files and a SQL Lite database for analysis.



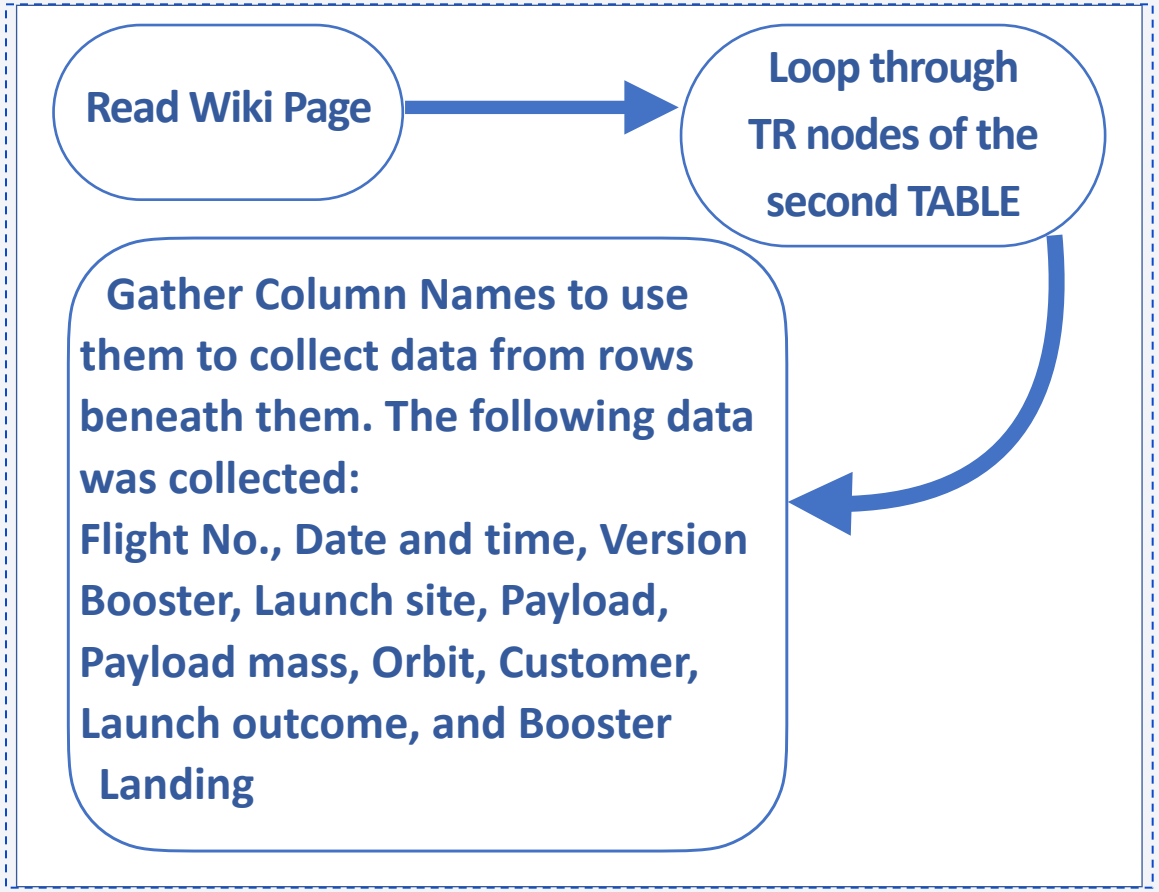
Data Collection – SpaceX API

- The flowchart to the right demonstrates the API calling process used to collect data through SpaceX's API
- Completed API calls can be found in the following notebook: https://github.com/Paurian/Coursera-Work/blob/main/Applied Data Science Capstone/Module 1/jupyter-labs-spacex-data-collection-api.ipynb



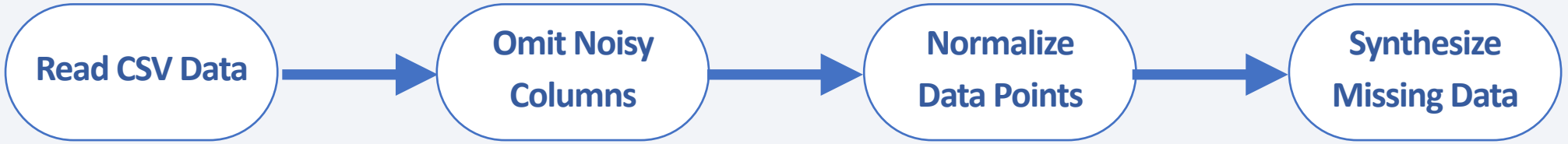
Data Collection - Scraping

- Webscraping is performed using Beautiful Soup. The entire web page is read, then nodes are pulled and parsed for data
- Completed Scraping calls can be found in the following notebook: https://github.com/Paurian/Coursera-Work/blob/main/Applied%20Data%20Science%20Capstone/Module%201/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data Wrangling involved finding noise such as the GTO orbit data and omitting that from the remaining analysis. It also involved creating a basic set of outcomes and assigning landing classes based on these outcomes.
- Classes are given a numeric, binary representation of successful landings where 0 is used for failures and 1 is used for successes.
- In situations where the landing success data is missing, the average success rate was used.
- Results were then stored to a CSV for future analysis.



EDA with Data Visualization

- Exploring the data for anomalies and successes, and the relationships of successful landing rates with later flights (indicating experience gains), payload masses, launch sites and orbits help to understand if there are any correlations between successes and other factors.
- Scatter Plots were used to identify pay loads to experience (higher flight numbers) and do show a correlation where greater mass is sent to orbit as experience increases. A cross-reference of launch sites and pay loads from the flight number also indicate that some sites are better equipped for heavier launches than others.
- A bar graph to show the relationship between success rate to orbit helps to indicate that the outermost orbit, GEO, has a greater success rate than smaller orbits like the MEO and LEO range.
- When Line Graphed across time, the success rate truly improves with slight degradation, attributed to some issues at the KSC LC 39A Launch Site to the SO orbit in 2018.
- The full EDA and Data Visualization can be viewed at https://github.com/Paurian/Coursera-Work/blob/main/Applied%20Data%20Science%20Capstone/Module%202/eda_data_viz.ipynb

EDA with SQL

- SQL Queries Performed Include:
  - Creating and populating a table with the CSV data.
  - Listing the different launch sites, focusing in on the CCA tier launch sites
  - Identifying the total payload mass contracted from NASA CRS
  - Identifying the average payload mass from Booster F9 v1.1 and across all F9 Boosters
  - Identifying the date of the first successful pad landing
  - Listing the boosters with success in drone ship landings and a moderate payload
  - The total number of successful and failed missions (based on the scrubbed data)
  - The dates, landing pads, boosters and launch sites of failed drone ship landings (not the same as failed missions)
  - The overall success and failures of each landing outcome and their locations.
- These findings can be reviewed at: https://github.com/Paurian/Coursera-Work/blob/main/Applied%20Data%20Science%20Capstone/Module%202/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium was used to circle and label the locations of NASA and the land-based launch sites.
- In addition to these locations, the number of successes and failures that originated from each launch site were placed in group markers so that by zooming into these regions and clicking on the launch site, they encircle the location with indicators on success (green) and failure (red).
- Finally, the distance from one of the closest east-coast launch pads was measured to the shore line with a line drawn from the general site to the ocean. Rocket launches pose a high risk of injury and by performing them closer to the ocean where there is less population we reduce these risks. Additionally, being close to ocean ports allows large equipment to be brought to the launch pad more effectively.
- This study may be viewed at: https://github.com/Paurian/Coursera-Work/blob/main/Applied%20Data%20Science%20Capstone/Module%203/lab_jupyter_launch_site_location%20(online%20version).ipynb
- Note that maps might not draw due to browser and/or web-based security and the notebook might need to be downloaded and ran locally to fully view its content.

Build a Dashboard with Plotly Dash

- Plotly Dash was used to generate a web-based dashboard for interactive visualizations and analysis with an intent to find interesting correlations between launch sites, payload and booster correlations.
- Variables: Launch Sites and Payloads
- Visuals:
  - Pie Chart displays Success Rates when All Launch Sites are selected and displays Failures and Successes for individual sites that are selected.
  - Scatter Plot Displaying the success (class 1) and failure (class 0) for payloads with booster versions identified by legend color. This can be filtered using the Launch Site and the weight slider.
- Source can be found at https://github.com/Paurian/Coursera-Work/blob/main/Applied%20Data%20Science%20Capstone/Module%203/spacex-dash-app.py

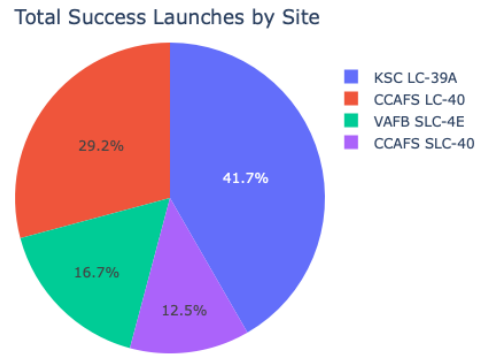
Predictive Analysis (Classification)

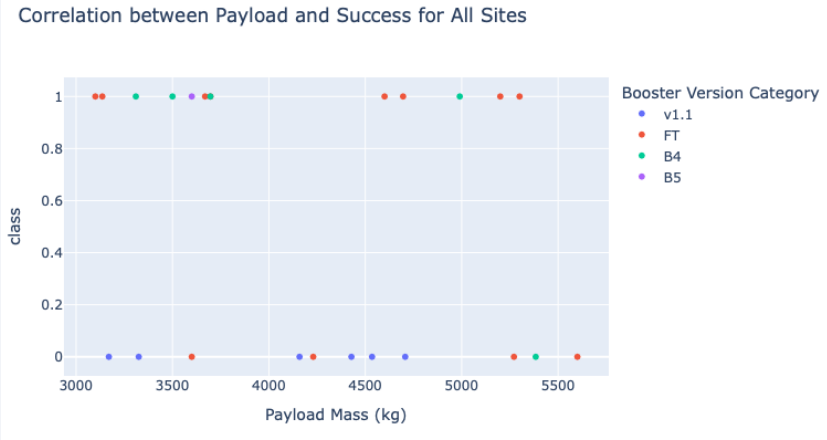
- Models were created to predict first stage landing success. Different models were built and tested to identify the best model for these predictions.
- Classification of successful landings was used as the target (Y) and all other variables were normalized to numbers then transformed through a standard scaler algorithm.
- The independent factors (X) were then fit to the overall scale so that all factors would be treated equally in the training and discovery.
- Training and splitting was at a 20% test withholding (80% given to train) with a mild random state of 2. This resulted in a modest 18 test samples.
- The models trained and tested are: Logistic Regression, Support Vector Machine, Decision Tree, and Nearest Neighbor.

- The full set of tests may be found at https://github.com/Paurian/Coursera-Work/blob/main/Applied%20Data%20Science%20Capstone/Module%204/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results
  - There is a positive correlation between experience and success rates, higher orbits and success rates.
- Interactive analytics demo in screenshots



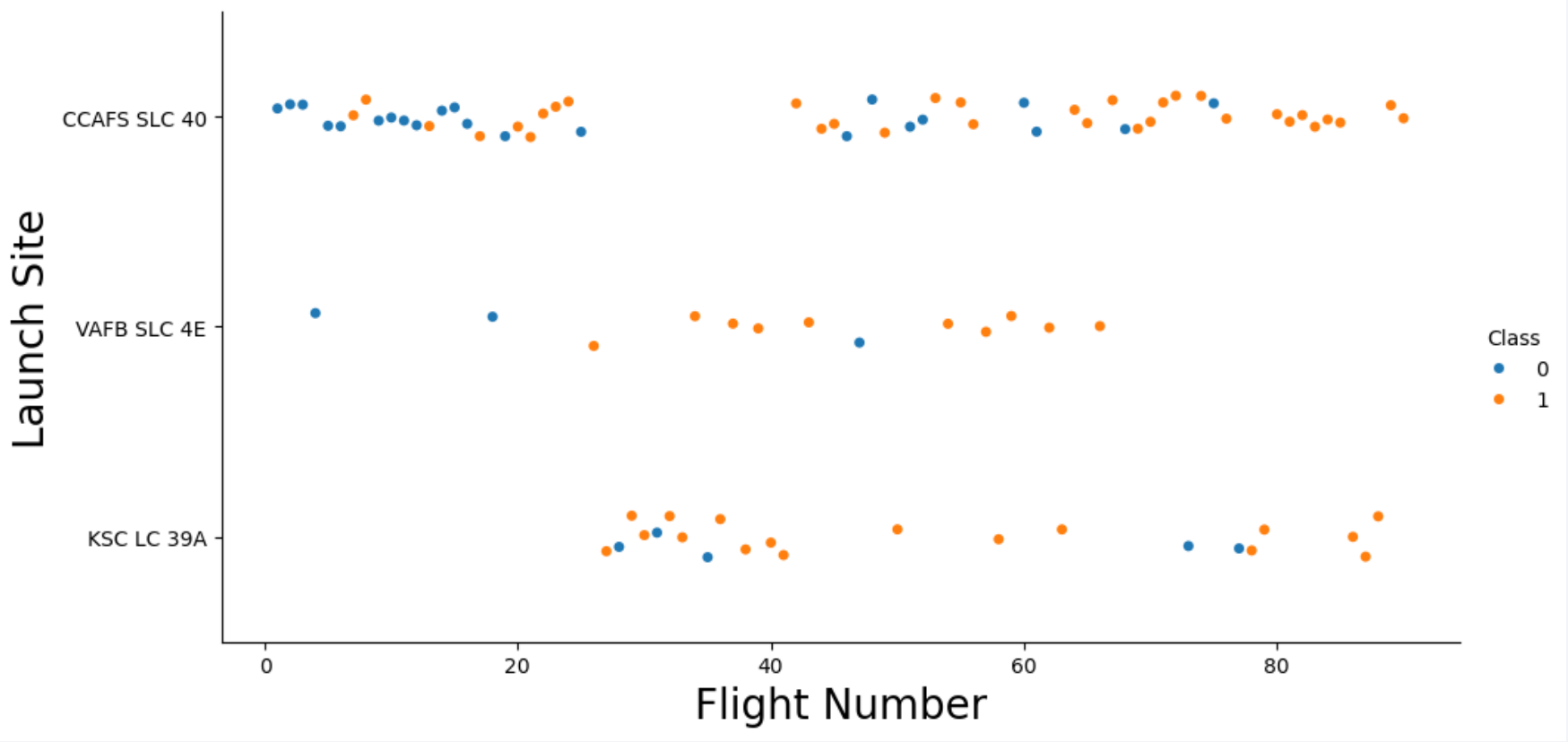


- Predictive analysis results
  - Decision Tree performed just slightly better than the other models

Section 2

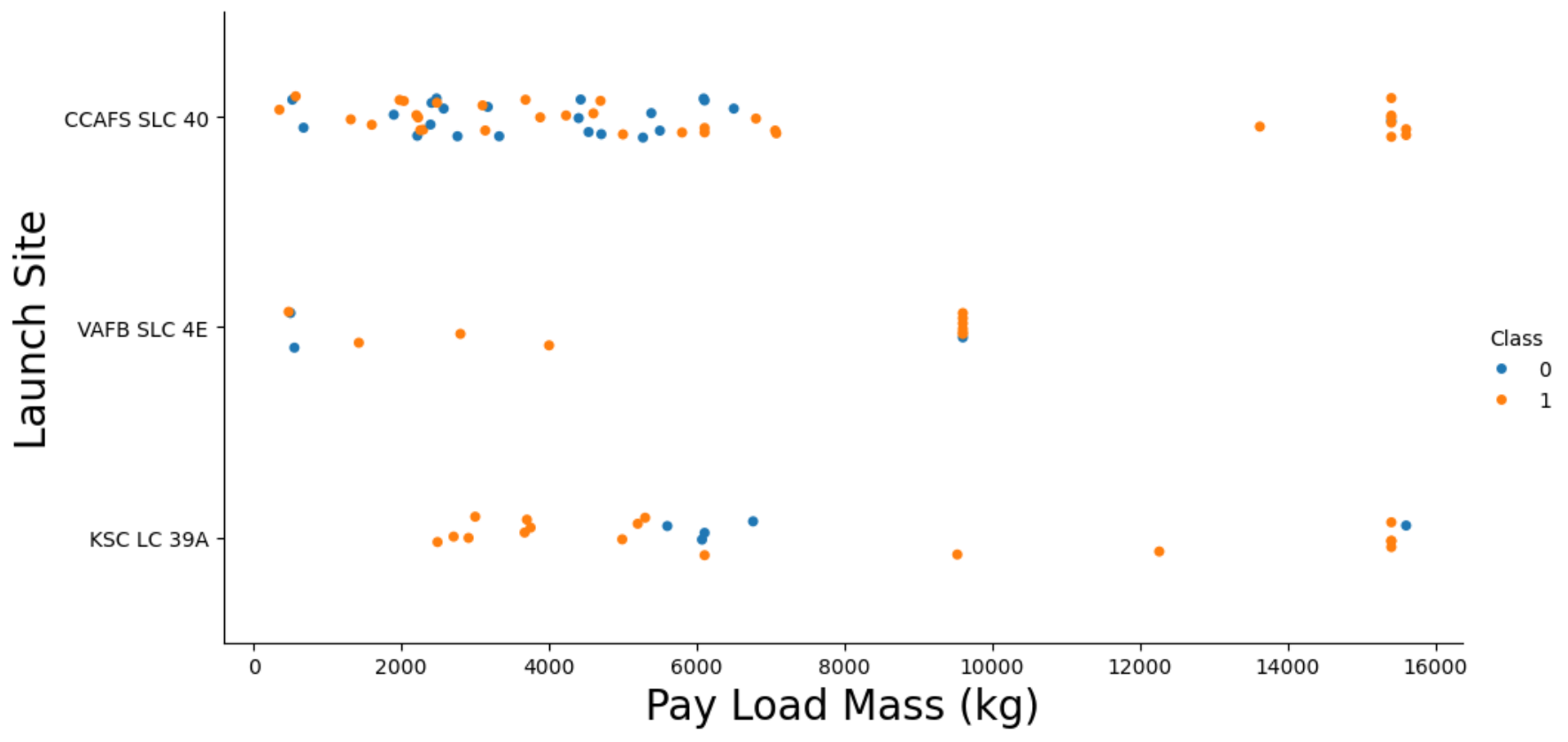
Insights drawn from EDA

Flight Number vs. Launch Site



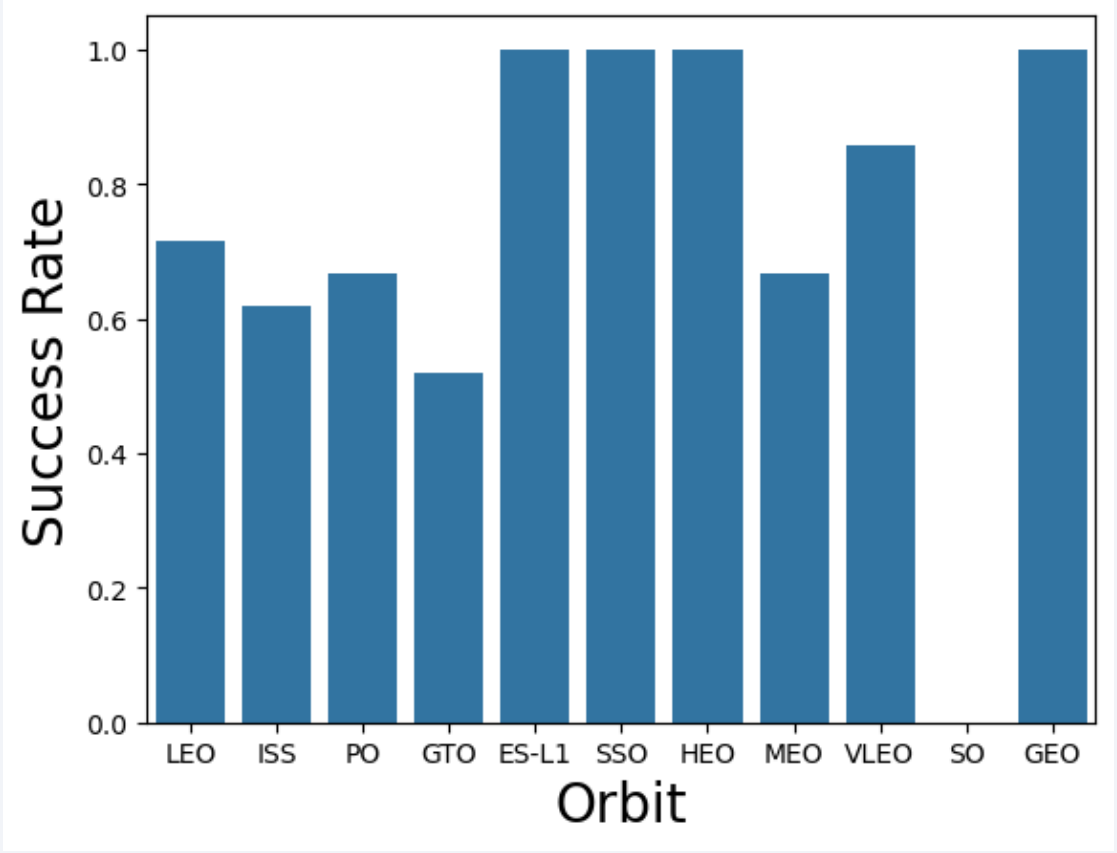
Success improves in later flights regardless of launch site. This could be a general indicator of adjusting for variances in each site (e.g. launching in familiar fields) or it could indicate that as SpaceX experience rises, its success rate does as well. There does not appear to be a correlation between higher success at any specific launch site.

Payload vs. Launch Site



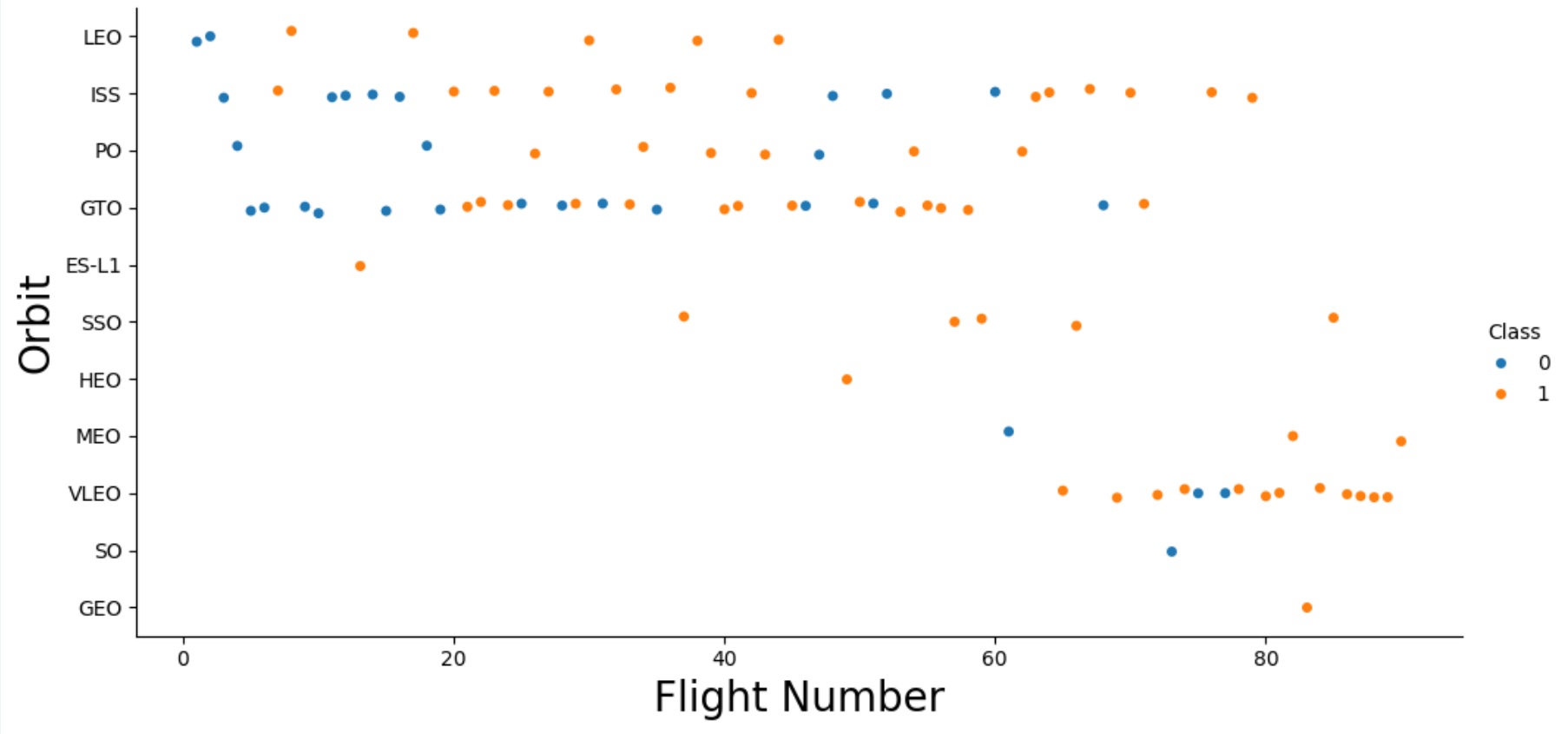
Pay Load Mass seems to indicate success rates with higher mass perhaps stabilizing the rocket during detachment. Lower mass seems more successful at the KSC launch site, which is slightly more inland than the others. Perhaps this could be weather related.

Success Rate vs. Orbit Type



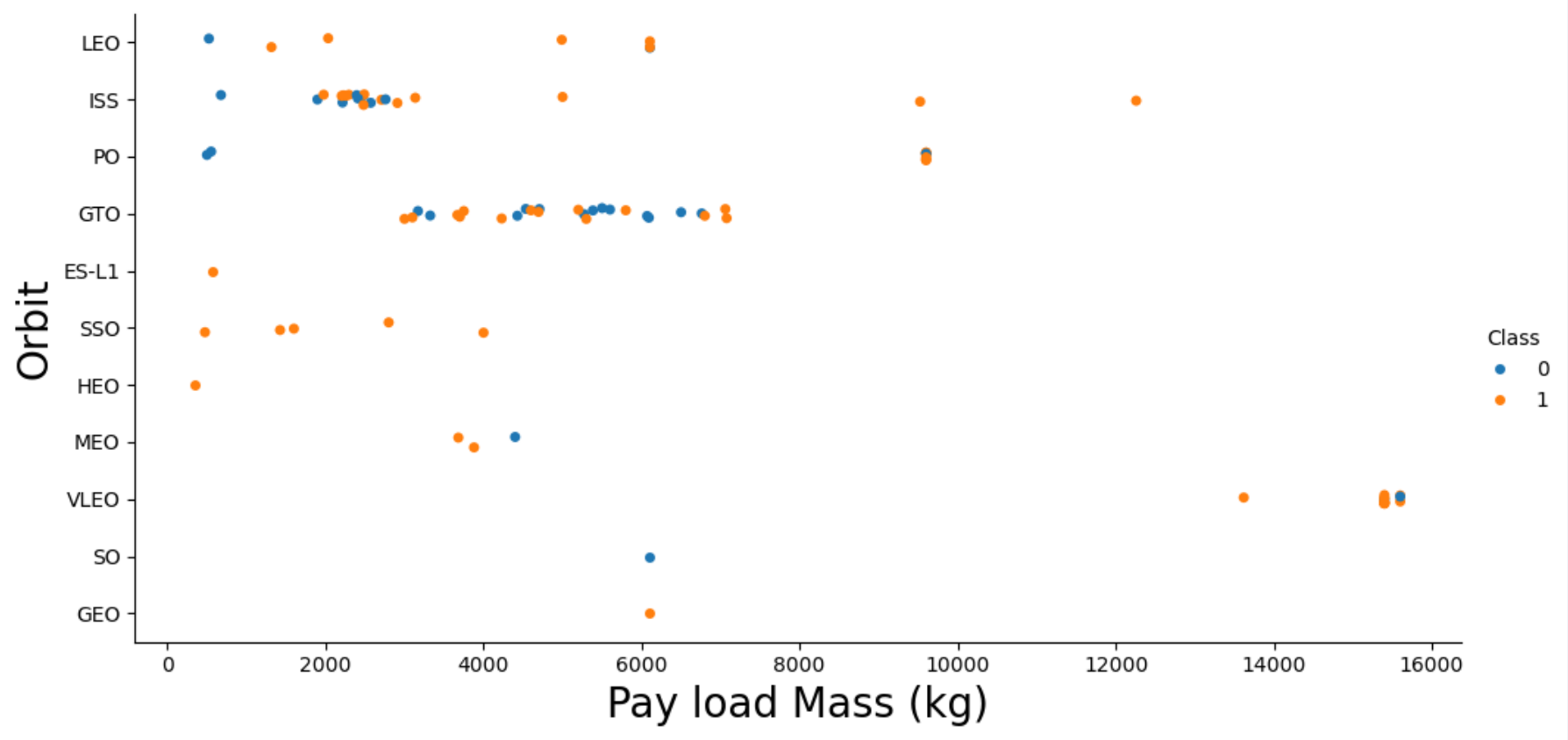
The high HEO and GEO orbits have the greatest combined success rate while lower orbits like LEO, ISS, VLEO and MEO have moderate to low success rates overall.

Flight Number vs. Orbit Type



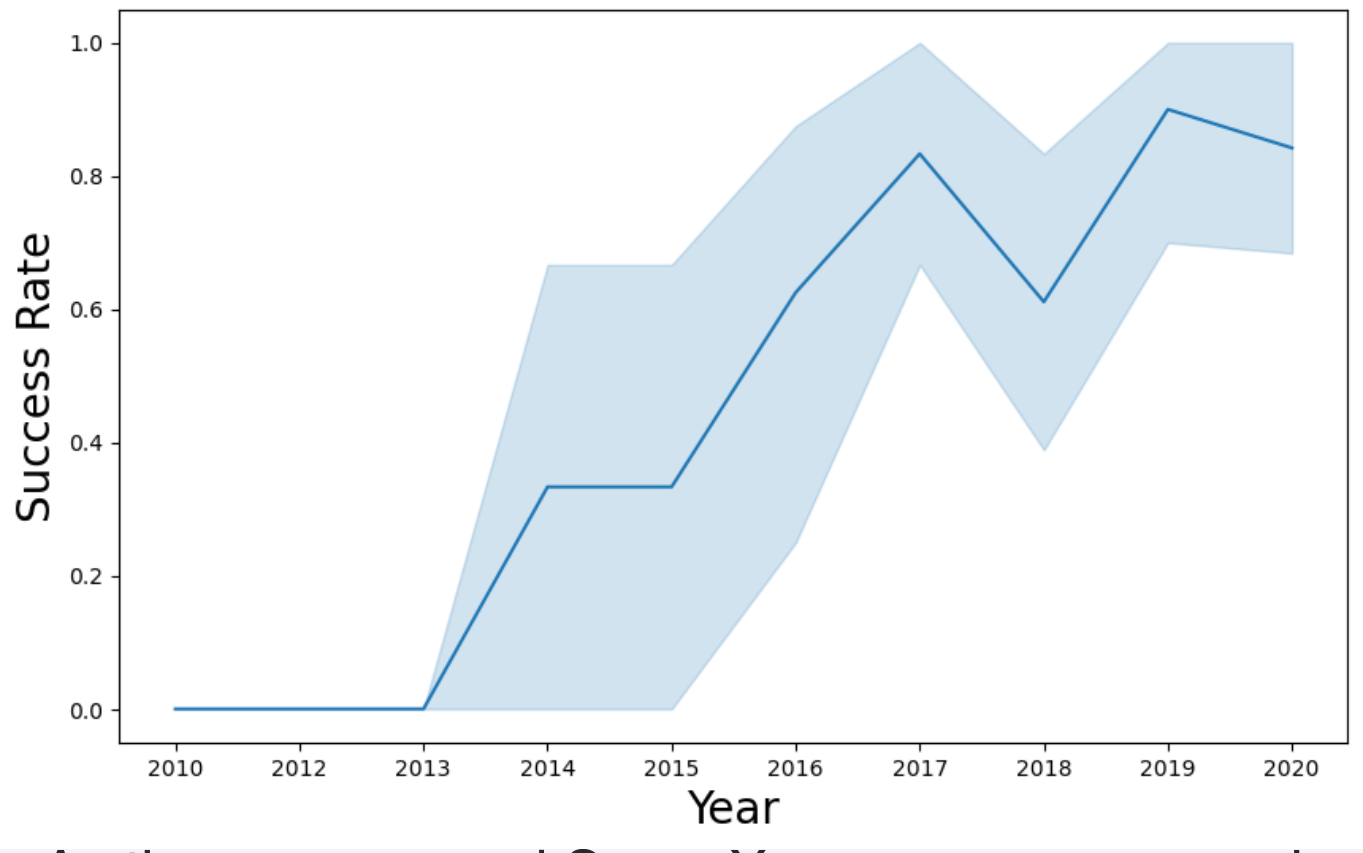
As the number of flights increase, riskier missions are attempted such as high and very low earth orbits. VLEOs require much higher precision and resolution to achieve.

Payload vs. Orbit Type



Heavy payloads tend to be for very low earth orbits and for ISS accommodations. Moderate payloads are for satellite delivery in most orbital ranges and light deliveries for general low orbits and, again, for the ISS.

Launch Success Yearly Trend



As time passes and SpaceX garners more experience, its success rate steadily climbs. Failures in 2018 are from new experimental equipment and wear on its first stage rockets.

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

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

Querying for only distinct (unique) sites from our data.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster Version	Launch Site	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

Querying for any launch sites that begin with CCA (i.e. Launch_Site LIKE ‘CCA%’)

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE CUSTOMER = 'NASA (CRS)';
```

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

Total_Payload_Mass
45596

The total payload mass from NASA CRS (cargo to resupply the international space station) totals to 45,596 Kilos. That's over 50 tons of supplies!

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
# If literal, then:
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

Avg_Payload_Mass
2928.4

```
# But if the question means to get the average of all Booster Versions of type F9 v1.1 ... which can have differe
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

Avg_Payload_Mass
2534.6666666666665

The average payload of the F9 v1.1 is 2928 Kilos. Note, however, that there are variants of the F9 v1.1 that reduce the average payload when added to the calculation.

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Mission_Outcome = 'Success' AND Landing_Outcome LIKE '%ground%';
```

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

MIN(Date)
2015-12-22

The first successful landing outcome for ground pads was on December 22nd, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Mission_Outcome = 'Success' AND
Landing_Outcome LIKE '%drone ship%' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

Booster_Version
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

There are 5 different boosters that have successful landings with mid-size payloads. These are F9 FT: B1020, B1022, B1026, B1021.2 and B1031.2

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT SUBSTRING(TRIM(Mission_Outcome), 0, 8) AS Outcome, COUNT(1) AS Occurrences FROM SPACEXTABLE GROUP BY
```

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

Outcome	Occurrences
Failure	1
Success	100

Note that mission outcomes are not the same as landing outcomes. For example, SpaceX doesn't always require the successful landing of their first stage because they don't intend to reuse it. But they still measure landing outcomes.

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = ( SELECT
MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE ) ORDER BY Booster_Version;
```

This results in the following F9 B5 booster *versions* that have all carried the highest payload:

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

This reveals that there is a consistency in improvements and modifications in repeated missions and that there isn't a stagnancy in SpaceX with improving even the common and mundane missions.

2015 Launch Records

```
# %sql SELECT * FROM SPACEXTABLE;
%sql SELECT \
    CASE SUBSTR(Date, 6, 2) WHEN '01' THEN 'Jan' WHEN '02' THEN 'Feb' WHEN '03' THEN 'Mar' WHEN '04' THEN 'Apr' W
    Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE SUBSTR(Date, 0, 5)='2015' AND Landing_Ou
```

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

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

Interesting that in 2015 there were two failed attempts, each happening early in the year and both happening at the start of the first quarter with the exact same booster.

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

```
%sql SELECT Landing_Outcome, COUNT(1) AS Occurrences FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND
'2017-03-20' GROUP BY Landing_Outcome ORDER BY Occurrences DESC;
```

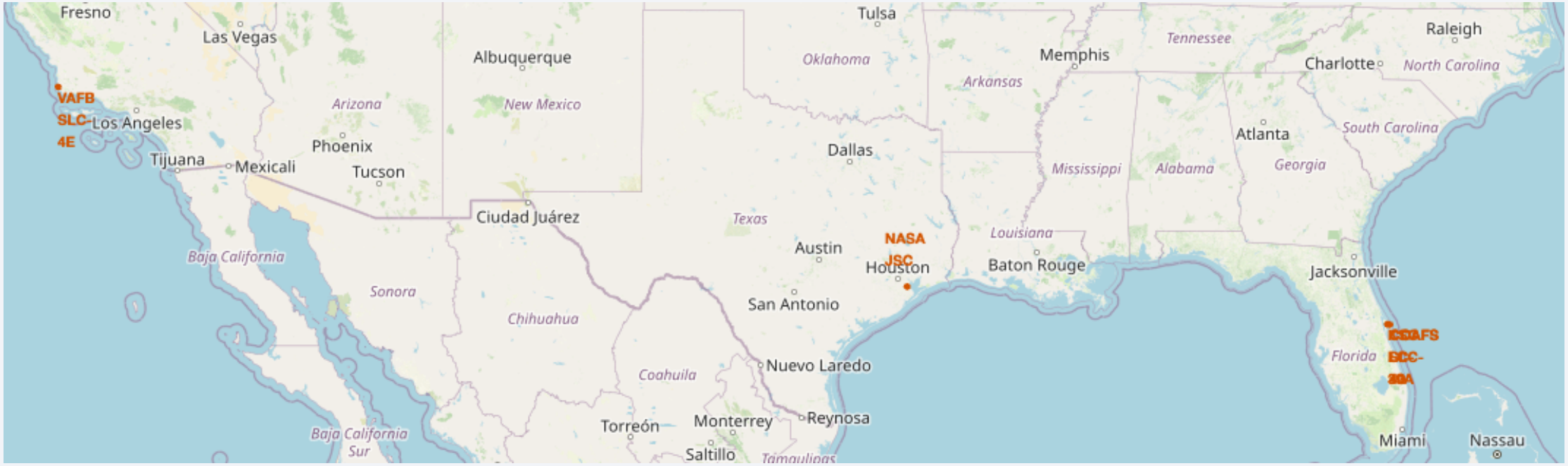
Landing_Outcome	Occurrences
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Drone Ship landings pose the greatest risk, but also the greatest improvement factor. Dependency on parachutes should probably be avoided.

Section 3

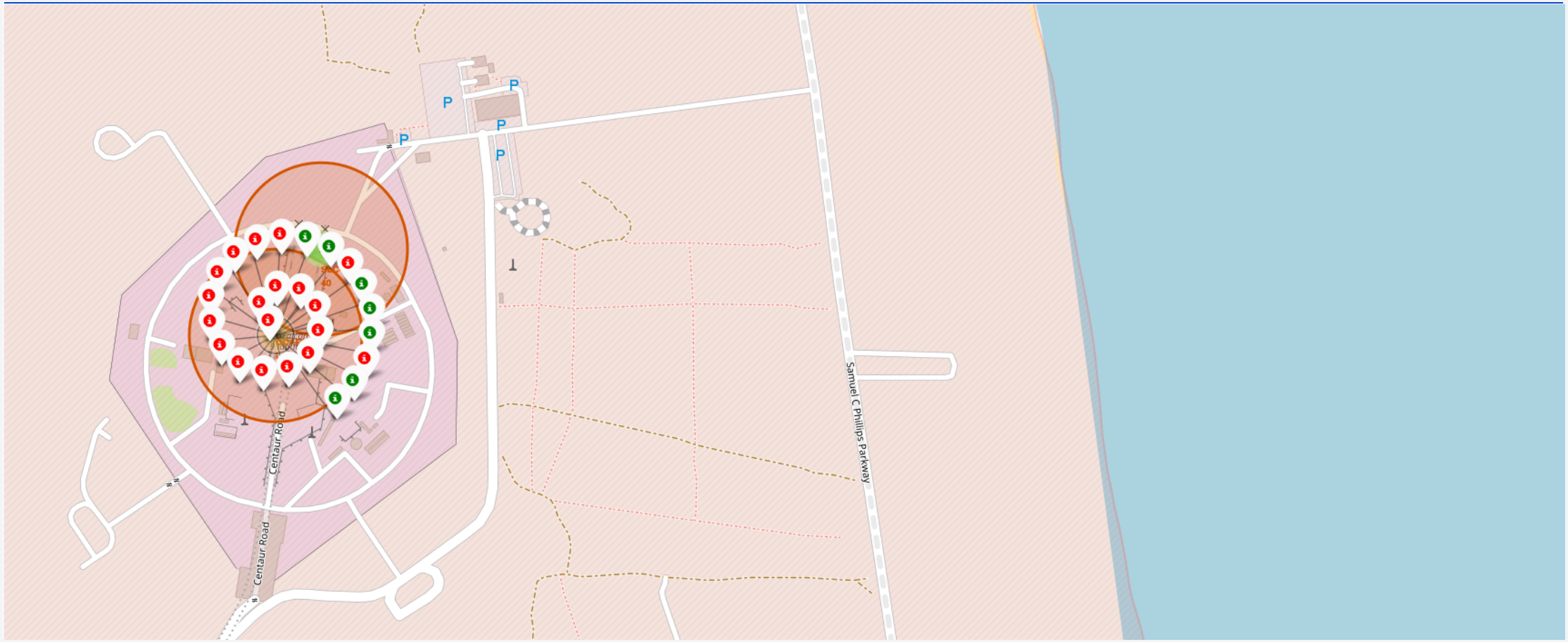
Launch Sites Proximities Analysis

All Launch Sites (Relative to NASA)



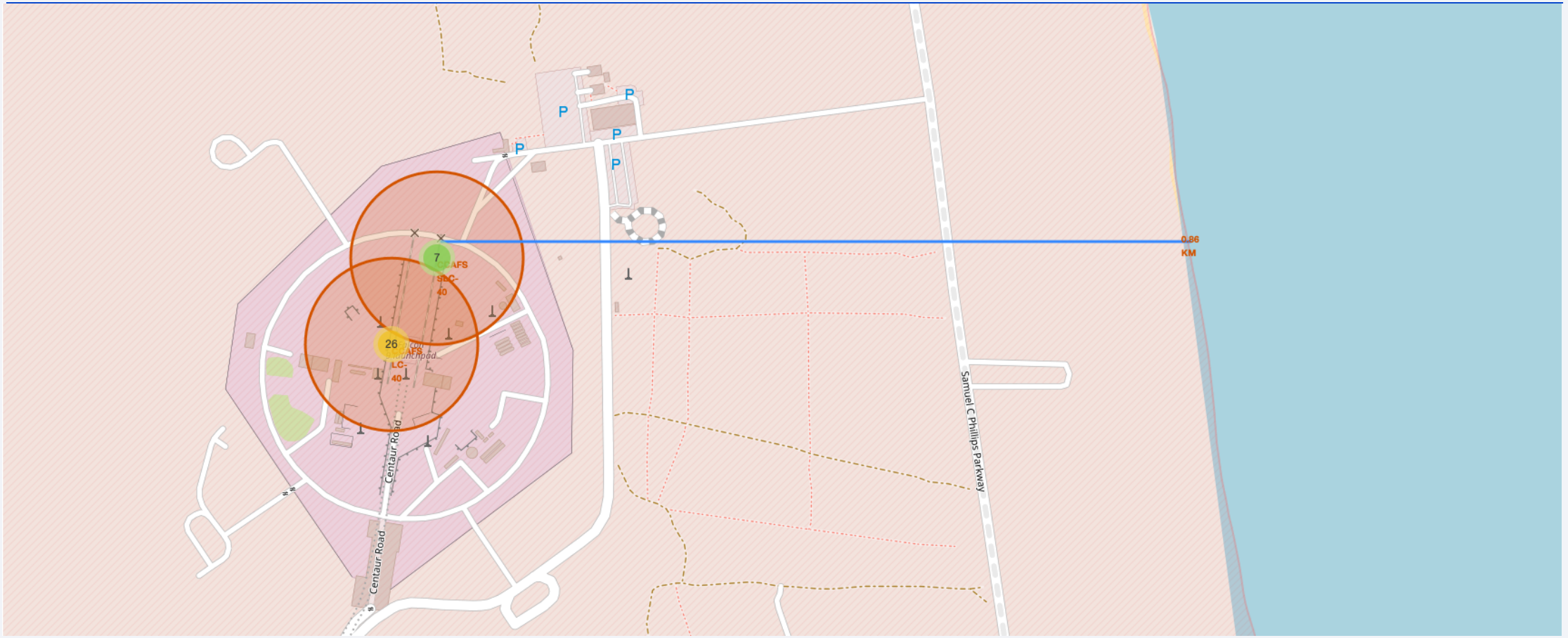
One of the benefits of SpaceX is that its control centers and launch pads span different coastlines while NASA's launch station is more inland. This allows SpaceX to experiment with riskier launch conditions and to learn how to accommodate them for successful missions in the future.

Launch Outcomes

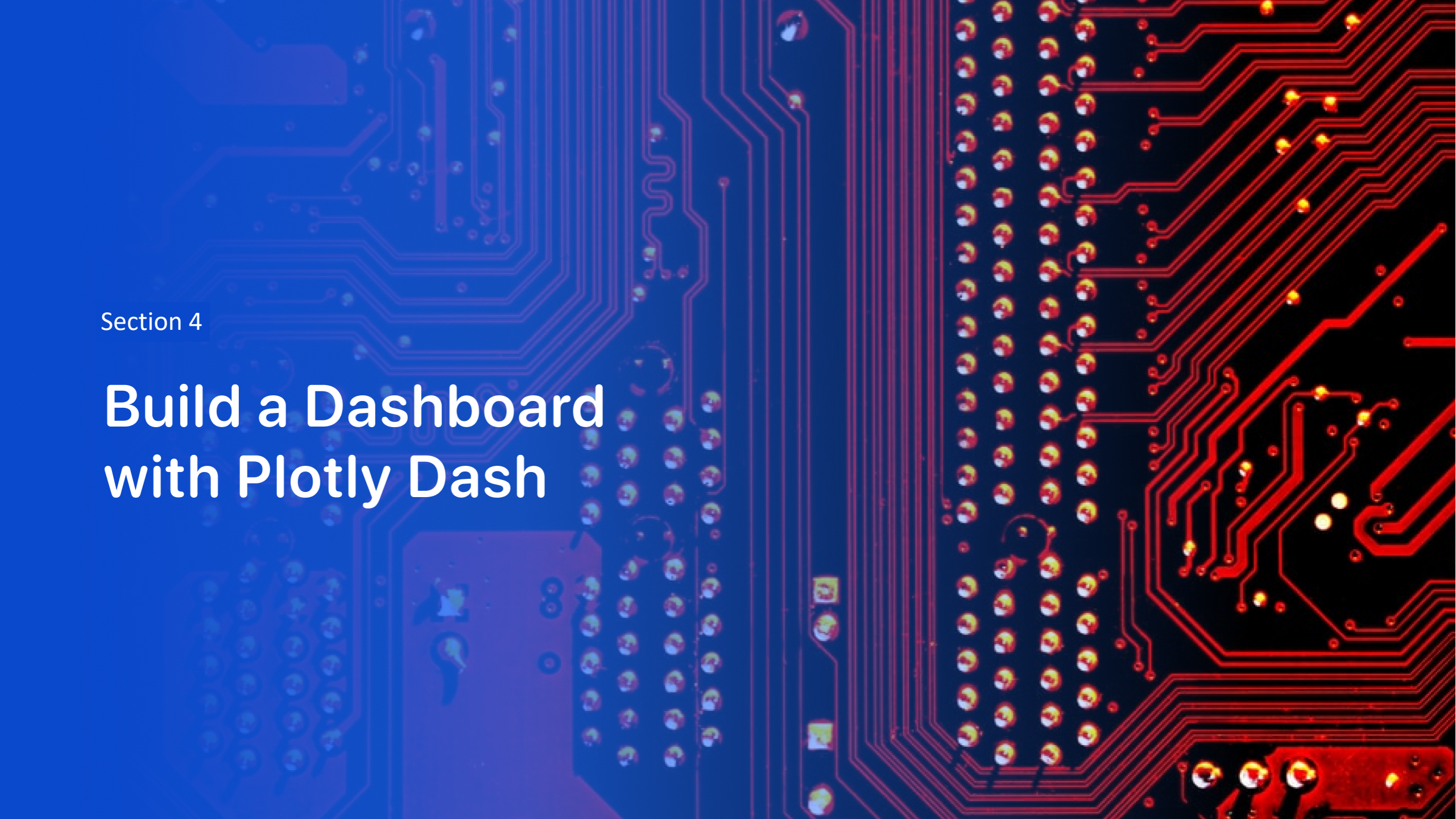


Launch Outcomes from one of the larger pads indicates an early succession of failures with successful outcomes following.

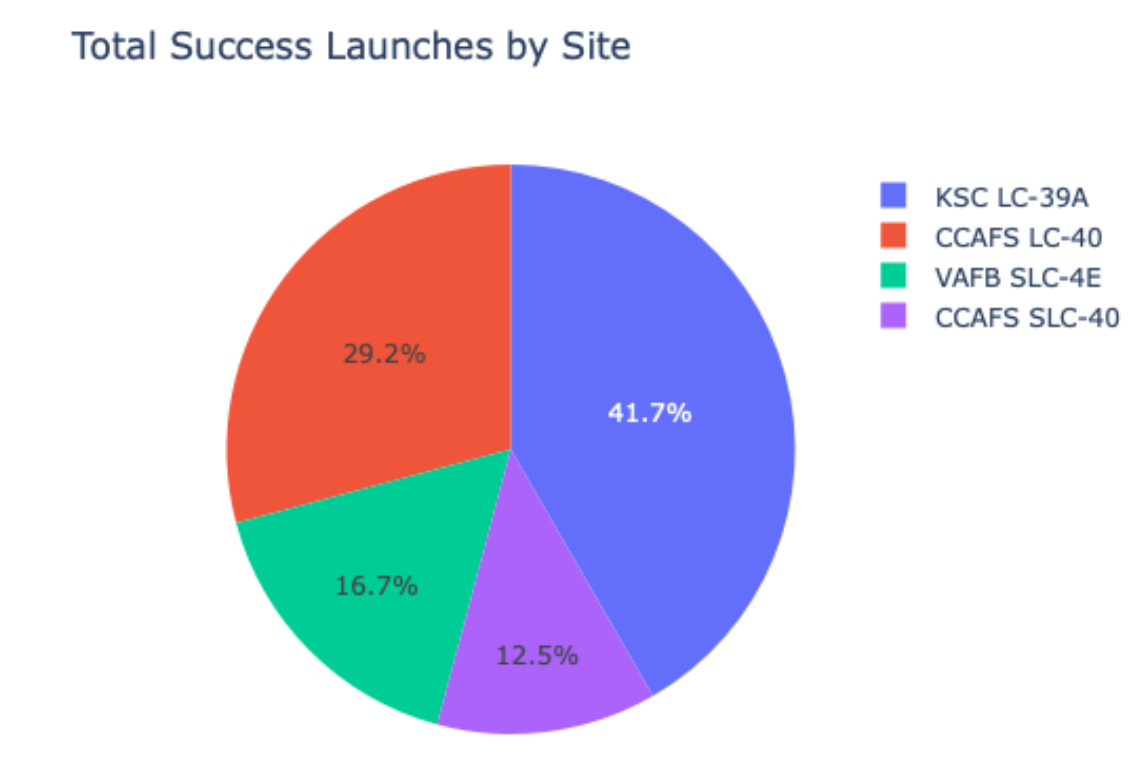
Proximity to the Ocean



The close proximity to the ocean (under 1 kilometer) indicates the tendency for rocket launches to occur where they're expected to have the least impact on injuries of person or property.

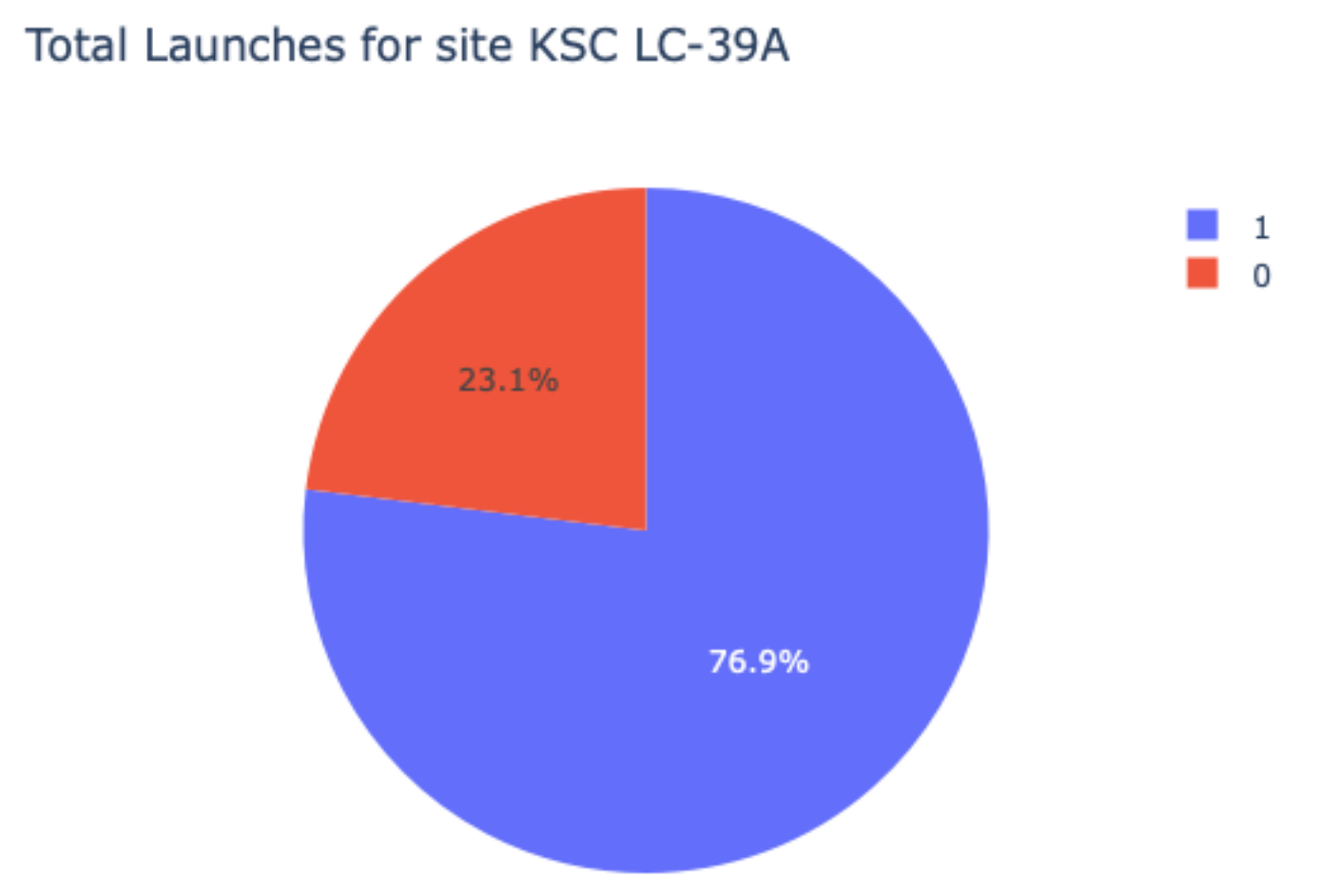

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



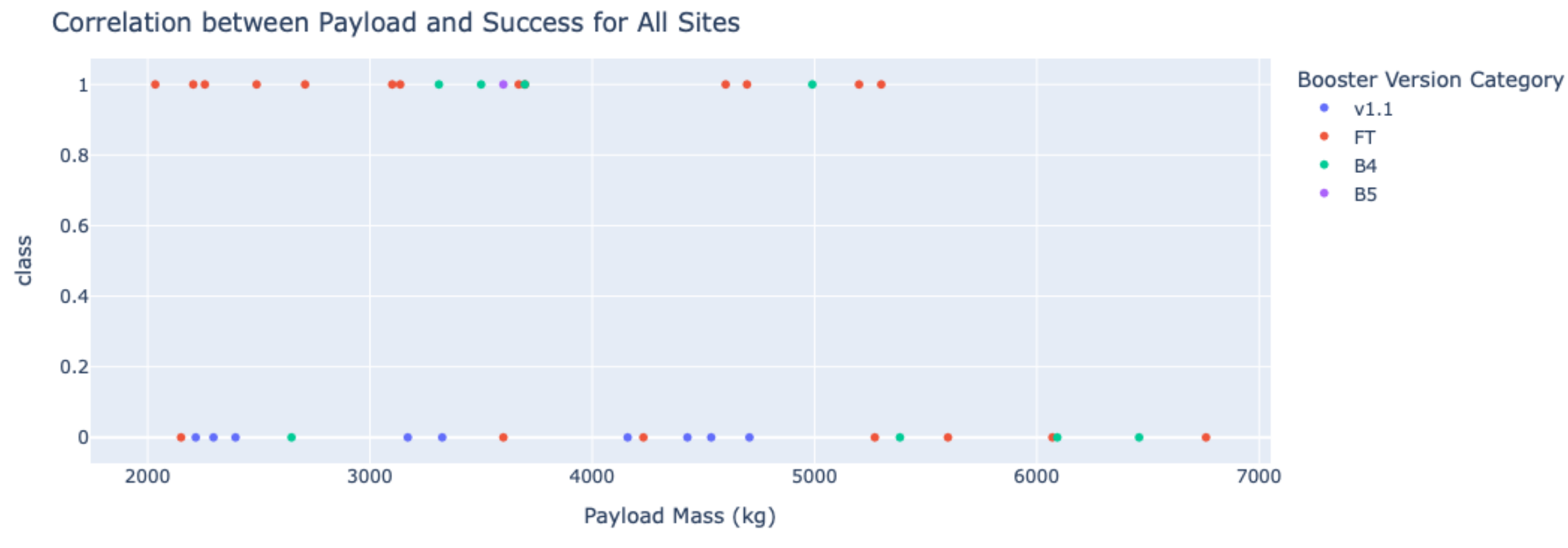
A vast majority of the successful launches have occurred at the KSC LC-39A (the Kennedy Space Center). This sits about 8 miles from the shore line.

Highest Successful Rated Launch Site



With nearly 77% success rate, the Kennedy Space Center has the highest success rate.

Lower and Upper Payloads and Their Successes

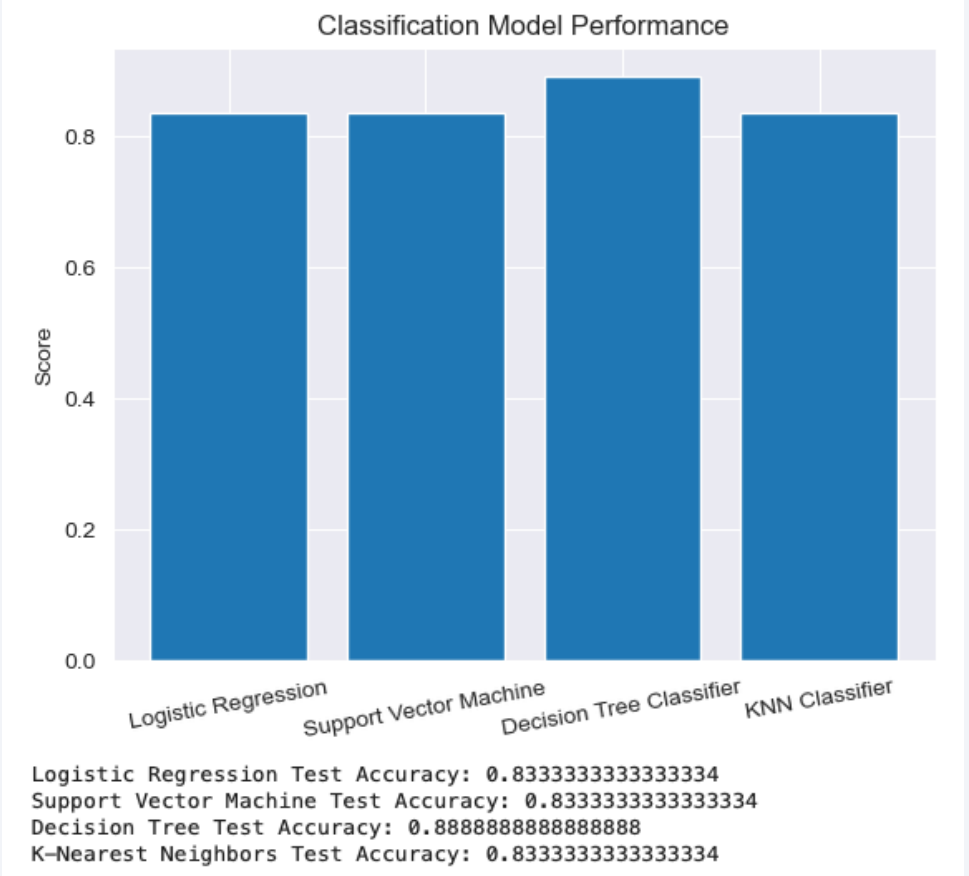


Note that booster version v1.1 has all failures. When this is removed from our set, there are 18 successes to 11 failures, about a 62% success rate.

Section 5

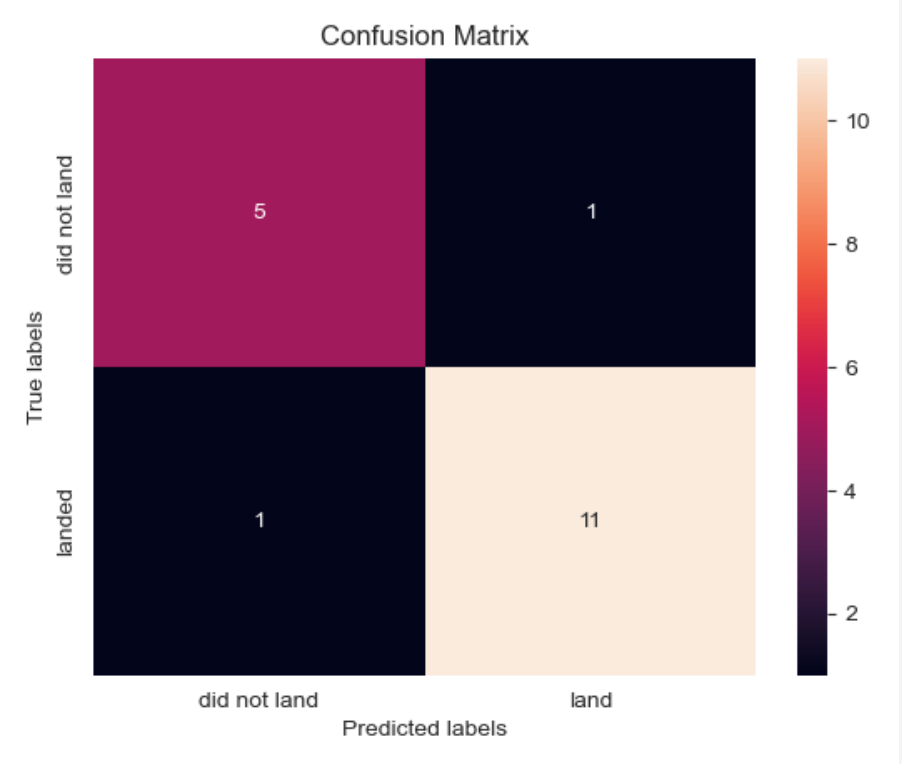
Predictive Analysis (Classification)

Classification Accuracy



Decision Trees have just a slightly higher accuracy at 89% compared to 83% of the other models.

Confusion Matrix



Because it has only 1 false negative this model provides a better prediction rate. We are most concerned with low contradictions. Other models went as hight as 3.

Conclusions

- The highest indicator of success is experience and time in the industry
- Although SpaceX had about an 85% landing success rate in 2017 (the stopping point of our data) it now has over a 99% success rate, supporting the evidence that experience is key.
- The ever-increasing version numbers of their booster rockets show that SpaceX is dedicated to achieving a perfect success rate, if at all possible.
- Their experience is also demonstrated in the complex nature of some of their missions like very low earth orbits (VLEO) and high GEO orbits.
- Some launch locations might provide a more stable result. Weight might also play a role in their success.
- To best compete with SpaceX, take on lower cargo missions launched to mid-level orbits from KSC.

Appendix

This project and a bulk of its resources were provided by IBM through the Coursera Data Scientist master class.

Thank you!