IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yazeed Alregaiey
December 30, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through SpaceX REST API and Web Scraping.
  - Data Wrangling.
  - Exploratory Data Analysis with Data Visualization and SQL.
  - Predictions done using Machine learning.
- Summary of all results
  - Exploratory Data Analysis results.
  - Interactive maps with Folium and dashboard with Plotly.
  - Machine learning Prediction results.

Introduction

- Project background and context:
  - The goal of this research is to forecast the likelihood that the Falcon 9 first stage will successfully land. According to SpaceX's website, the Falcon 9 rocket launch cost 62 million dollars. Other service providers might cost up to 165 million dollars each. The price difference is explained by SpaceX's ability to reuse the first stage. We can calculate the cost of a launch by estimating the probability that the stage will land. If another firm wants to compete with SpaceX for a rocket launch, this knowledge may be useful.
- Problems you want to find answers
  - Determining each one of the elements that impact the landing outcome.
  - What are the primary criteria of a successful or unsuccessful landing?
  - The optimal conditions need to improve the likelihood of a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping
- Perform data wrangling
  - One-Hot-Encoding to categorical features
- 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

- We collected the data using two techniques. The first was through the SpaceX REST API, which began with a get request. The response content was then decoded as json and converted into a Dataframe using json normalize (). The data was then cleaned.

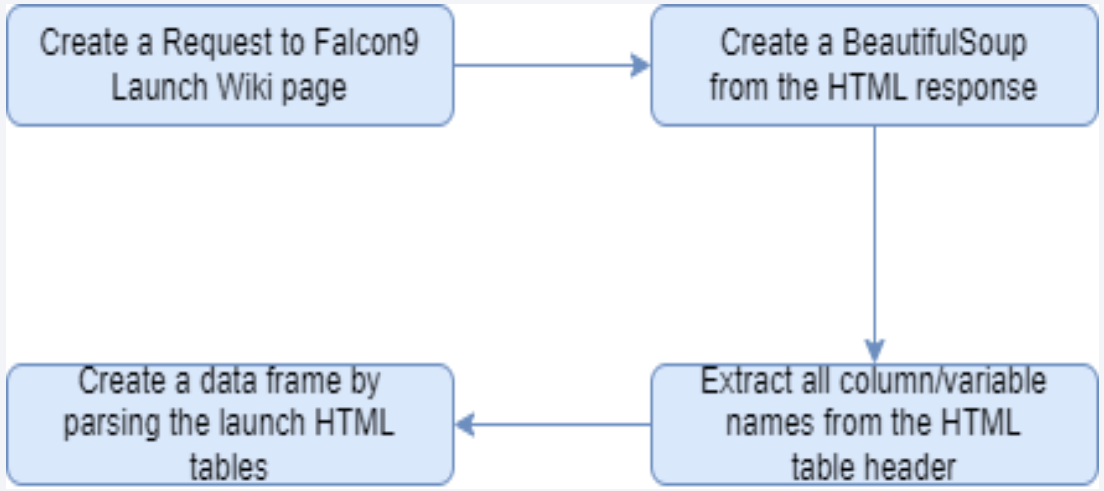
- The second way is web scraping, in which we utilized the BeautifulSoup package to extract the launch records as an HTML table from the Wikipedia page, then parse and transform the table to a pandas Dataframe.

Data Collection – SpaceX API

- Using the get request API, obtain rocket launch data.
- To convert a json output to a dataframe, we used the json_normalize function.
- https://github.com/yazeed1998/IBM-Applied-Data-Science-Capstone-project/blob/d09c1a18cdd865217b0875caad63269465ccb805/Data%20Collection%20API.ipynb

Data Collection - Scraping

- We used BeautifulSoup to scrape the Falcon 9 launch wiki page.
- We processed the table and transformed it to a pandas dataframe.
- https://github.com/yazeed1998/IBM-Applied-Data-Science-Capstone-project/blob/d09c1a18cdd865217b0875caad63269465ccb805/Data%20Collection%20Web%20Scraping.ipynb



Data Wrangling

- Exploratory data analysis was done to establish the training labels.
- We calculated the number of launches at each location as well as the frequency and number of orbits.
- We exported the findings to a csv file and constructed a landing outcome label from the outcome column.
- https://github.com/yazeed1998/IBM-Applied-Data-Science-Capstone-project/blob/1d19f59d4c340a980694cbfef338fa9b0311a7d0/Data%20Wrangling%20SpaceX.ipynb

EDA with Data Visualization

- We analyzed the data by showing the association between:
  - flight number and launch site.
  - payload and launch site.
  - success rate of each orbit type.
  - flight number and orbit type.
  - yearly trend in launch success.
- https://github.com/yazeed1998/IBM-Applied-Data-Science-Capstone-project/blob/c37e48951db2b7608b65b09ff2accecf1bc81006/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- We used SQL queries to collect and analyze data from the dataset and the queries are:
  - The names of the distinct launch locations of the space mission are displayed.
  - Show 5 records where the launch locations start with the string 'CCA'.
  - Display the total payload mass carried by NASA-launched rockets (CRS).
  - Display the average payload mass carried by the F9 v1.1 booster variant.
  - List the date of the first successful landing outcome in the ground pad.
  - List the names of boosters that have been successful in drone ships with payload masses more than 4000 but less than 6000.
  - Total the number of successful and unsuccessful mission results.
  - List the names of the booster versions that have transported the most payload mass.
  - List the records for the months in 2015 that will exhibit the month names, failure landing outcomes in drone ship, booster versions, and launch site.
  - Rank the number of successful landing outcomes in descending order between the dates 04 06 2010 and 20 03 2017.
  - https://github.com/yazeed1998/IBM-Applied-Data-Science-Capstone-project/blob/c37e48951db2b7608b65b09ff2accecf1bc81006/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- We identified all launch locations and added map elements such as markers, circles, and lines to the folium map to indicate the success or failure of launches for each site.
- We discovered which launch locations had a relatively high success rate using color labeled marker clusters.
- We estimated the distances between launch sites and their surroundings.

- https://github.com/yazeed1998/IBM-Applied-Data-Science-Capstone-project/blob/c37e48951db2b7608b65b09ff2accecf1bc81006/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- We created pie charts displaying the overall number of launches by specific sites.

- We created a scatter graph to highlight the association between Outcome and Payload Mass (Kg) for each booster version.

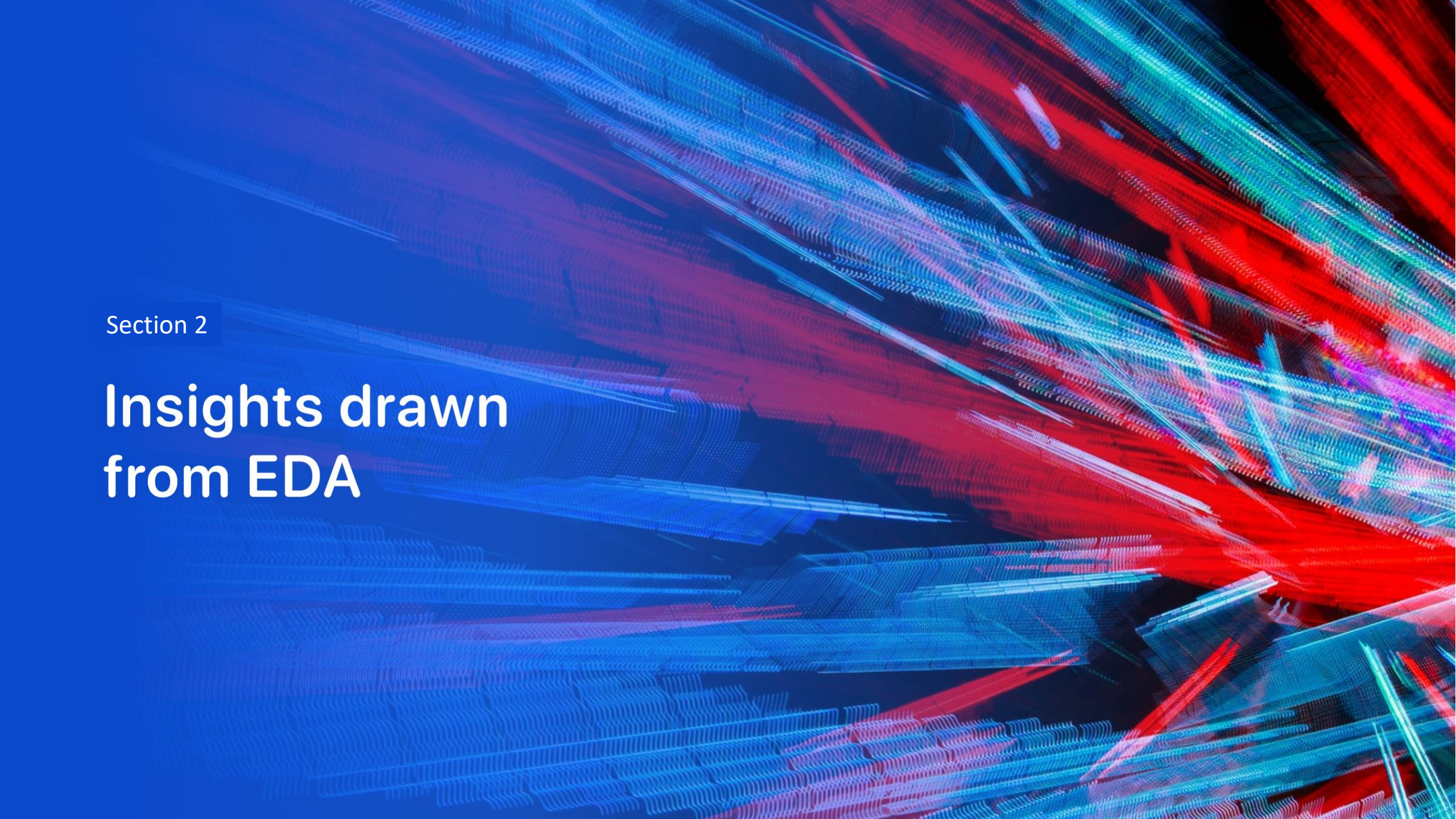
- https://github.com/yazeed1998/IBM-Applied-Data-Science-Capstone-project/blob/c37e48951db2b7608b65b09ff2accecf1bc81006/spacex_dash_app.py

Predictive Analysis (Classification)

- We imported the data with NumPy and Pandas, converted it, then split it into training and testing sets.
- We utilized accuracy as our model's measure and increased it through feature engineering and algorithm tuning.
- Using GridSearchCV, we created various machine learning models and tuned various hyperparameters.
- We discovered the most effective categorization model to be decision tree.
- https://github.com/yazeed1998/IBM-Applied-Data-Science-Capstone-project/blob/c37e48951db2b7608b65b09ff2accecf1bc81006/Machine%20Learning%20Prediction.ipynb

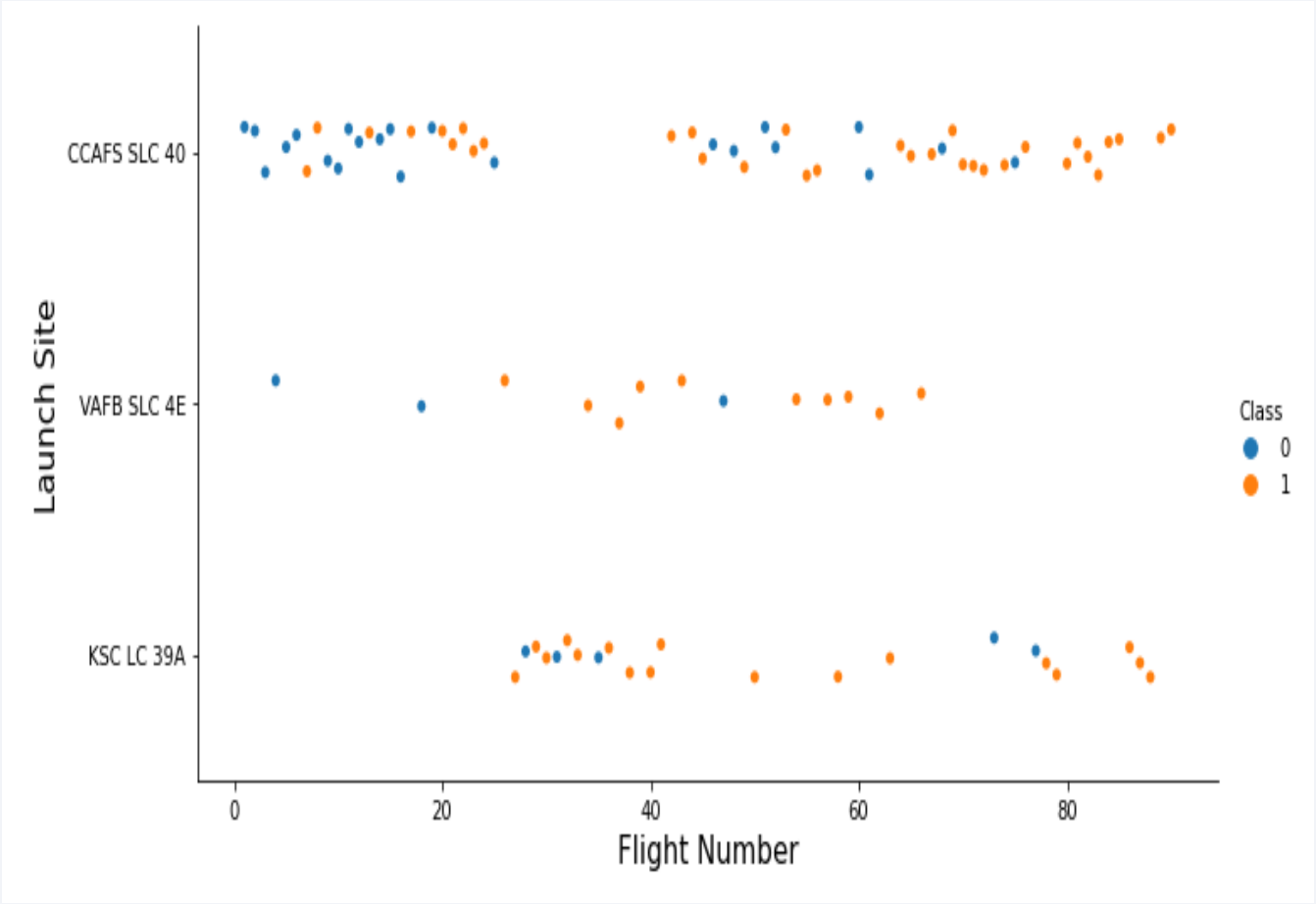
Results

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


Section 2
Insights drawn from EDA

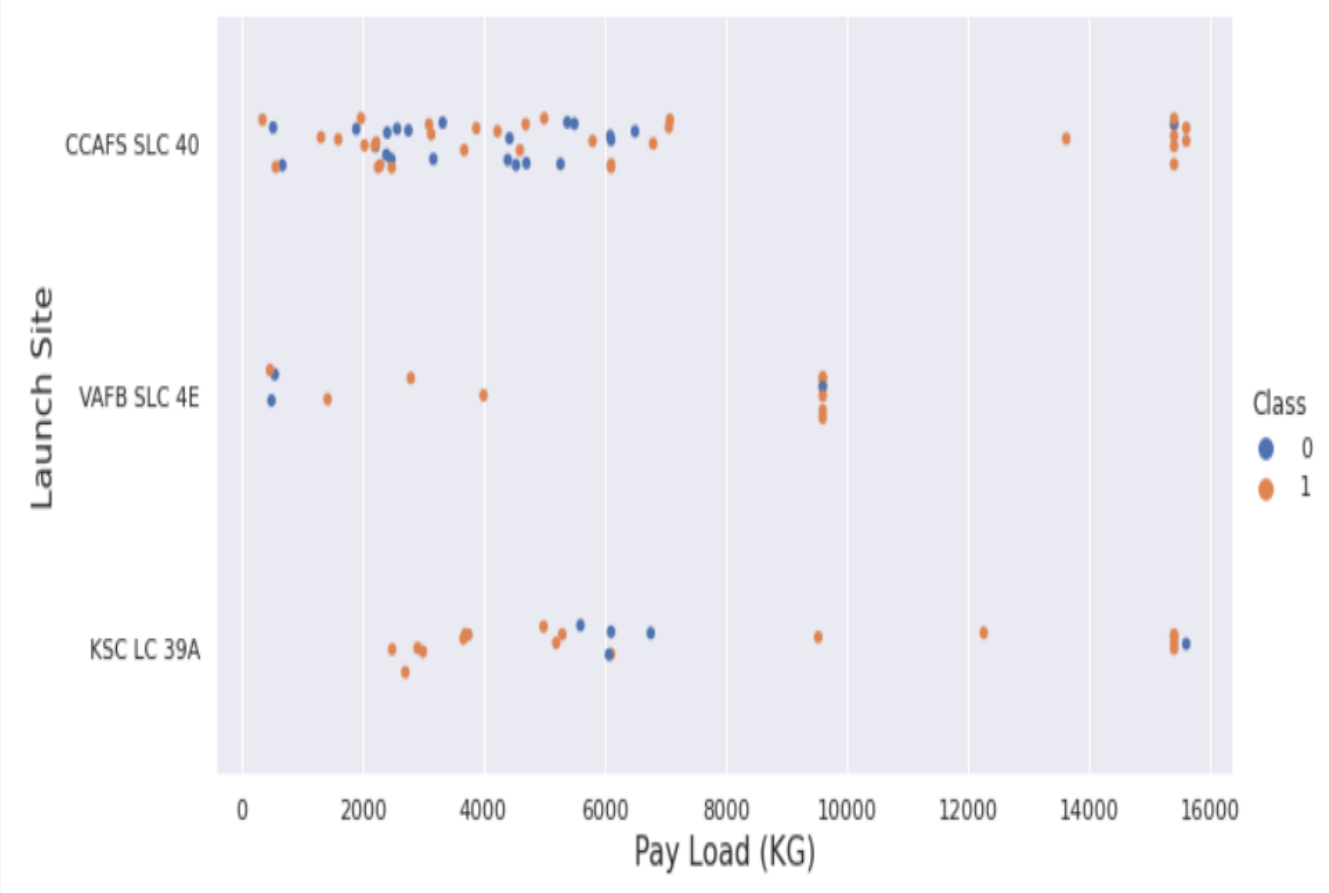
Flight Number vs. Launch Site

- The plot shows that the majority of the testing was performed in CCAFS SLC 40.
- There were several unsuccessful landings in the early phases of their testing, but they improved over time.



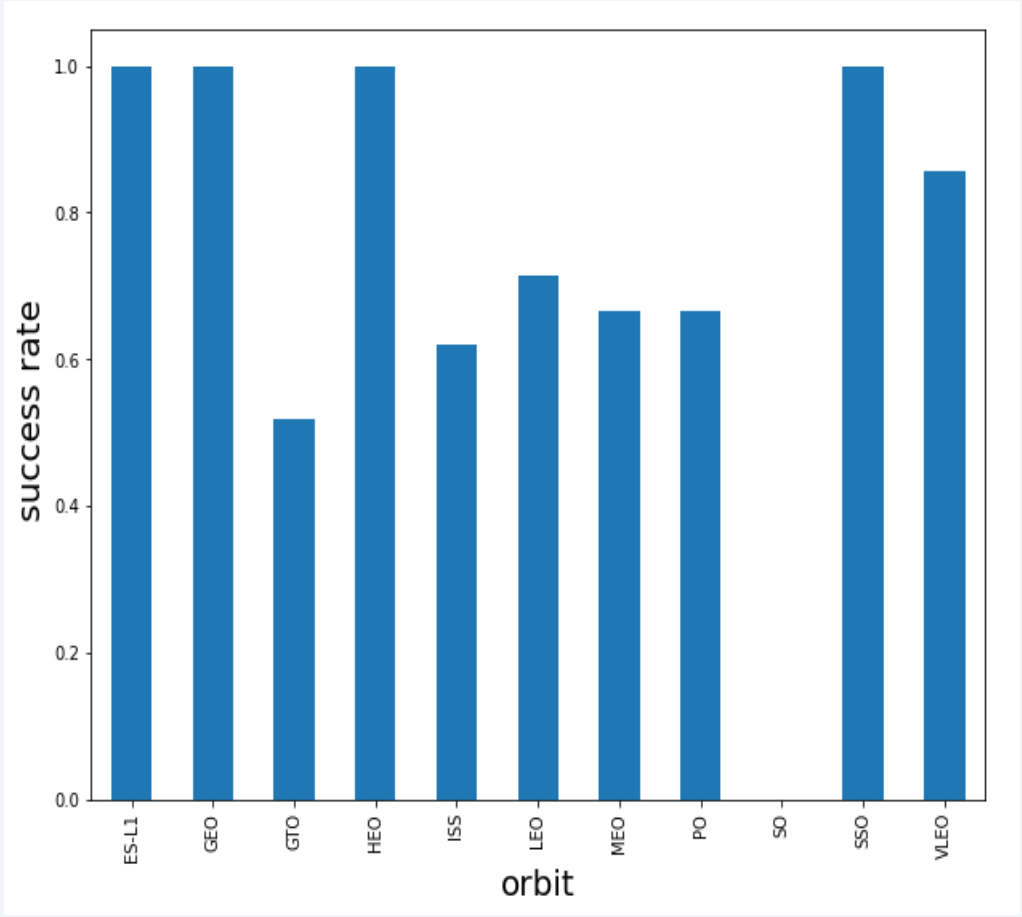
Payload vs. Launch Site

- Payloads weighing more than 8000 kg appear to have a better success rate than payloads weighing less than 8000 kg.



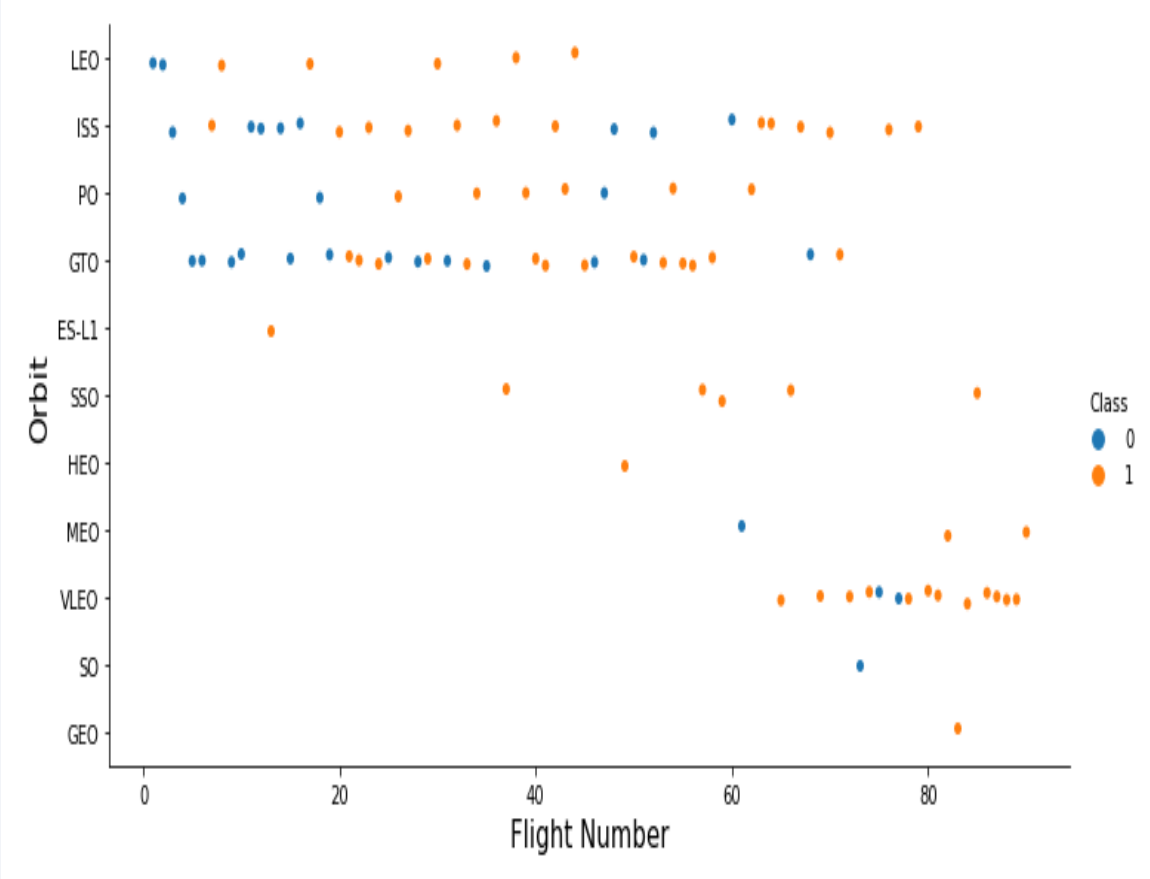
Success Rate vs. Orbit Type

- This graph demonstrated the possibilities of orbits influencing landing results, since certain orbits had a 100% success rate, such as SSO, HEO, GEO, and ES-L1.



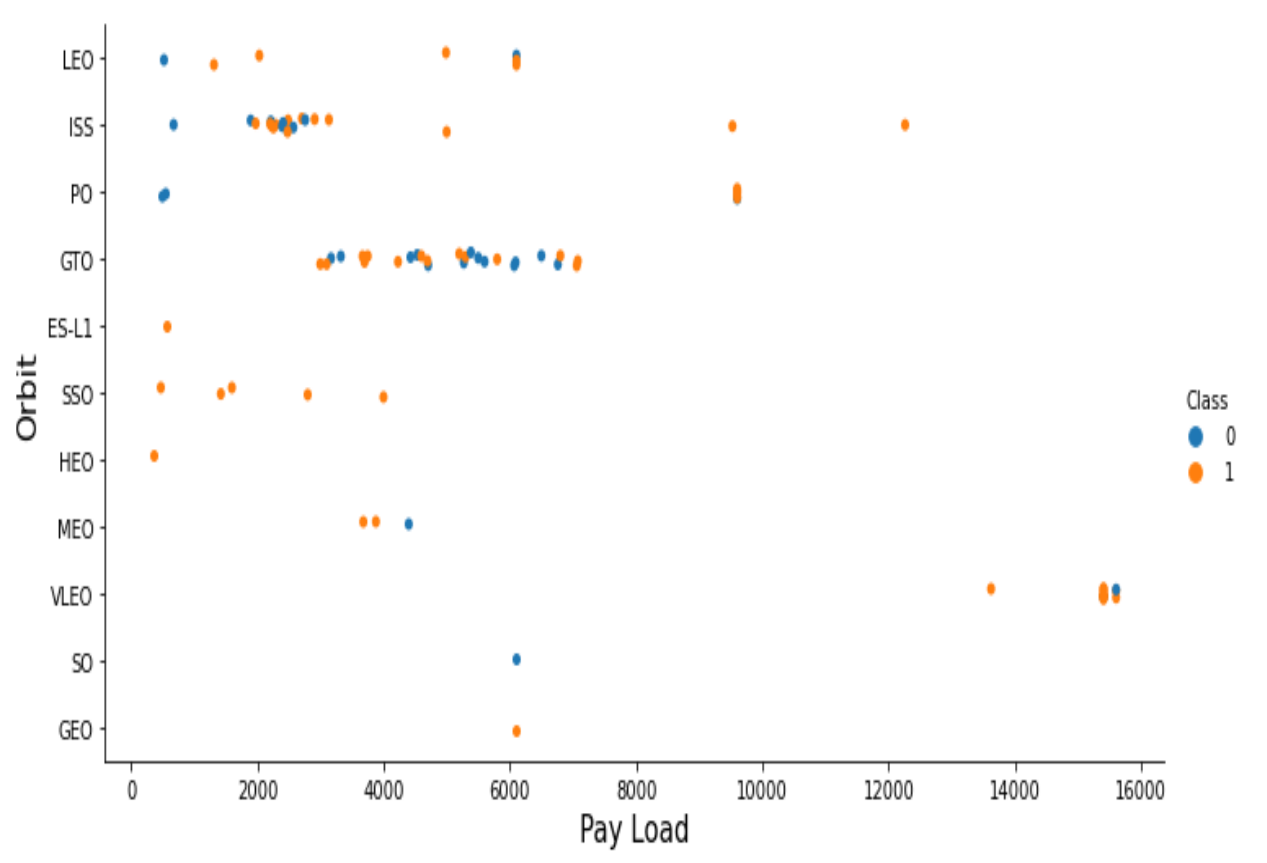
Flight Number vs. Orbit Type

- This scatter plot demonstrates that the overall success rate improved with time, resulting in more successful landings in newer orbits such as VLEO.



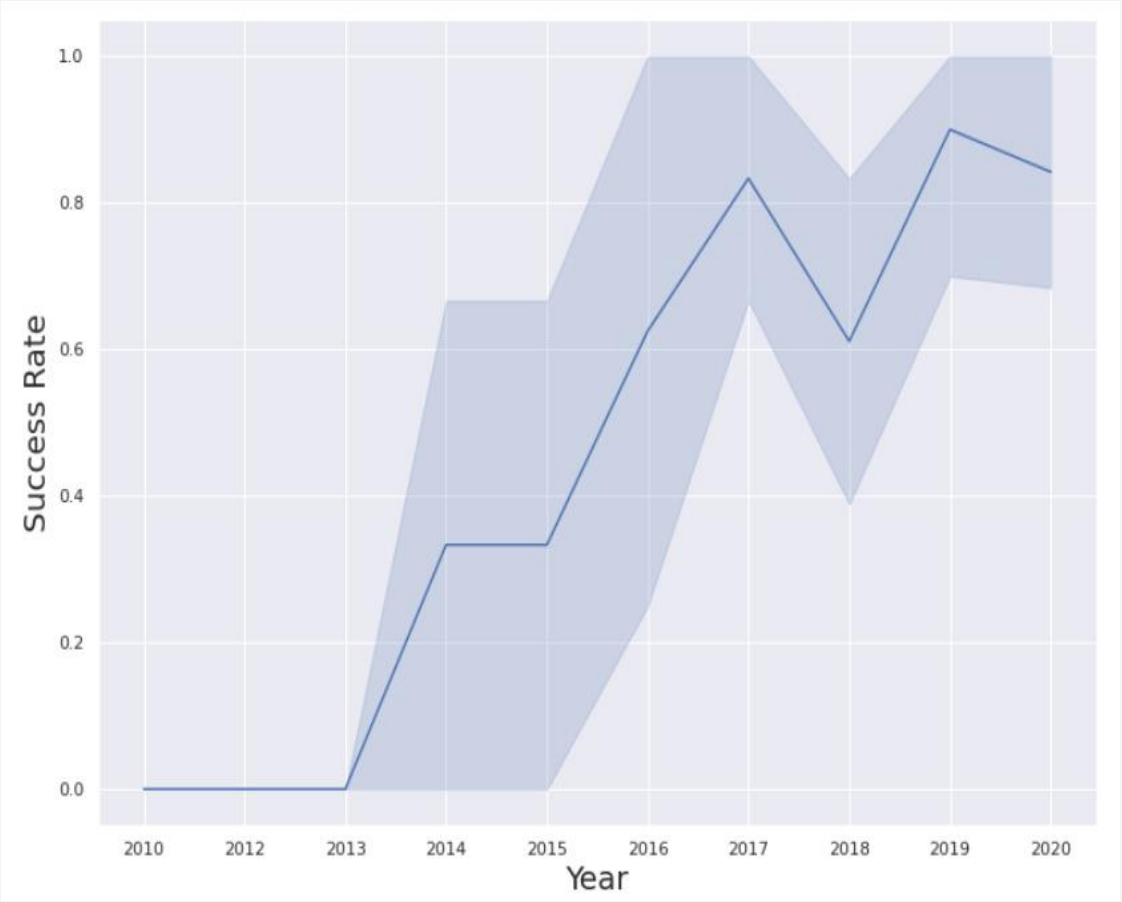
Payload vs. Orbit Type

- The weight of the payloads can have a significant impact on the success rate of launches in specific orbits.
- For example, MEO orbit will have a greater success rate with lower payloads, whereas orbit PO will have a higher chance of succeeding with heavier payloads.



Launch Success Yearly Trend

- This data clearly demonstrated a rising trajectory from 2013 to 2020.
- If this trajectory continues, the success rate will rise steadily.



All Launch Site Names

- The list indicates that we only have four launch sites.
- we were able to gather this data using the keyword distinct.

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX ORDER BY 1;
```

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- By utilizing the LIKE function to identify launch sites that begin with CCA and the LIMIT function to only display the first five records, we were able to obtain this result.

```
%sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

- We were able to obtain the sum of the payload by using the function SUM, and we filtered the data with the LIKE Function to obtain just the customer NASA (CRS).

```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEX WHERE CUSTOMER LIKE 'NASA (CRS)';
```

total_payload
45596

Average Payload Mass by F9 v1.1

- We were able to obtain the payload's average by utilizing the function AVG and filtering the data using the = Function to obtain just the Booster version F9 v1.1.

```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1';
```

avg_payload
2928

First Successful Ground Landing Date

- We obtained the date of the first successful landing by utilizing the function MIN and filtering the data using the = Function to obtain just the Landing result Success (ground pad).

```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEX WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

first_success_gp
2015/12/22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the function DISTINCT to retrieve only the booster version without duplicates, then the function BETWEEN to get payloads ranging from 400 to 6000.

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone ship)';
```

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

Total Number of Successful and Failure Mission Outcomes

- We obtained the number of successful and failed landings by using the COUNT function, and we classified them by mission outcome using GROUP BY.

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEX GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

mission_outcome	qty
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Using a subquery in the WHERE condition and the MAX() method, we found the booster that carried the max payload.

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX) ORDER BY BOOSTER_VERSION;
```

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

2015 Launch Records

- For the year 2015, we utilized a combination of the WHERE, LIKE, AND, and BETWEEN conditions to filter for unsuccessful landing outcomes in drone ships, booster versions, and launch site names.

```
sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE LIKE '%2015%';
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- This query provides the number of landing result records with the COUNT clause and is between the dates 04/06/2010 and 20/03/2017 with the BETWEEN clause.
- We aggregate them using the GROUP BY clause based on the landing outcome, then ORDER BY DESC displays the results in decreasing order.

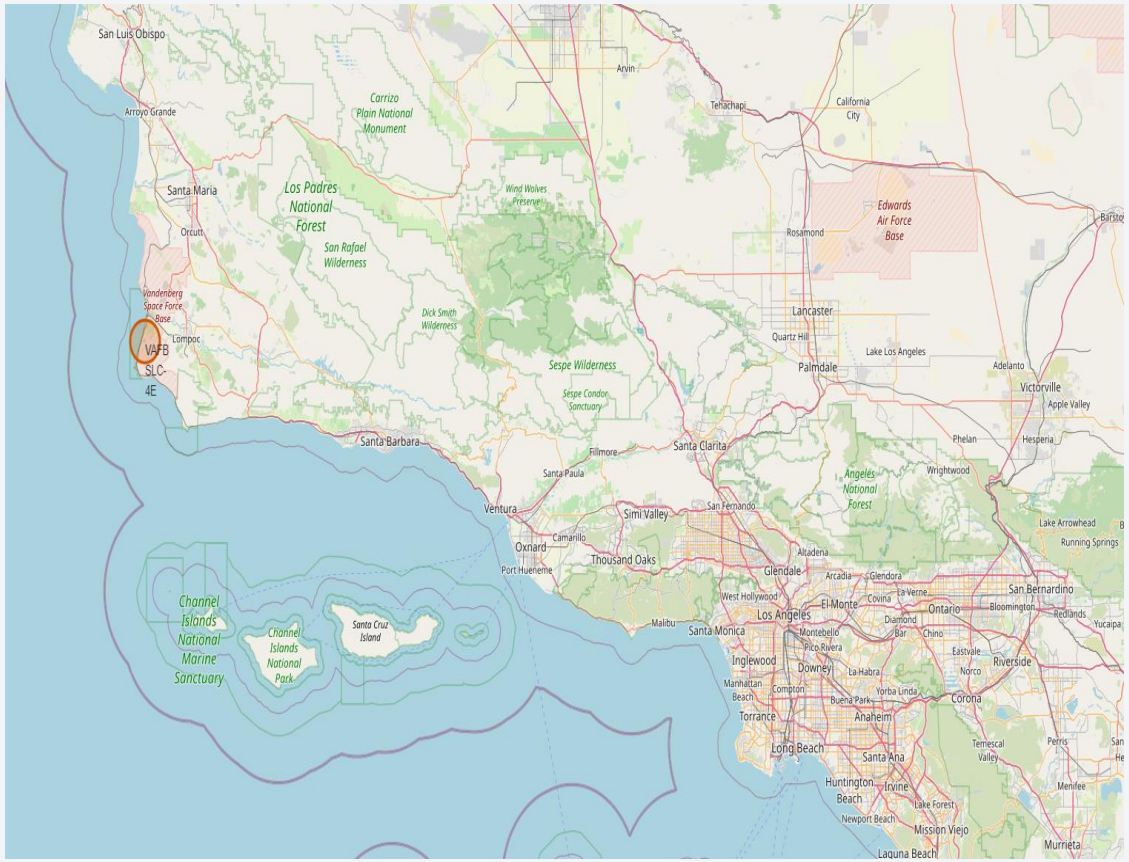
```
sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY QTY DESC;
```

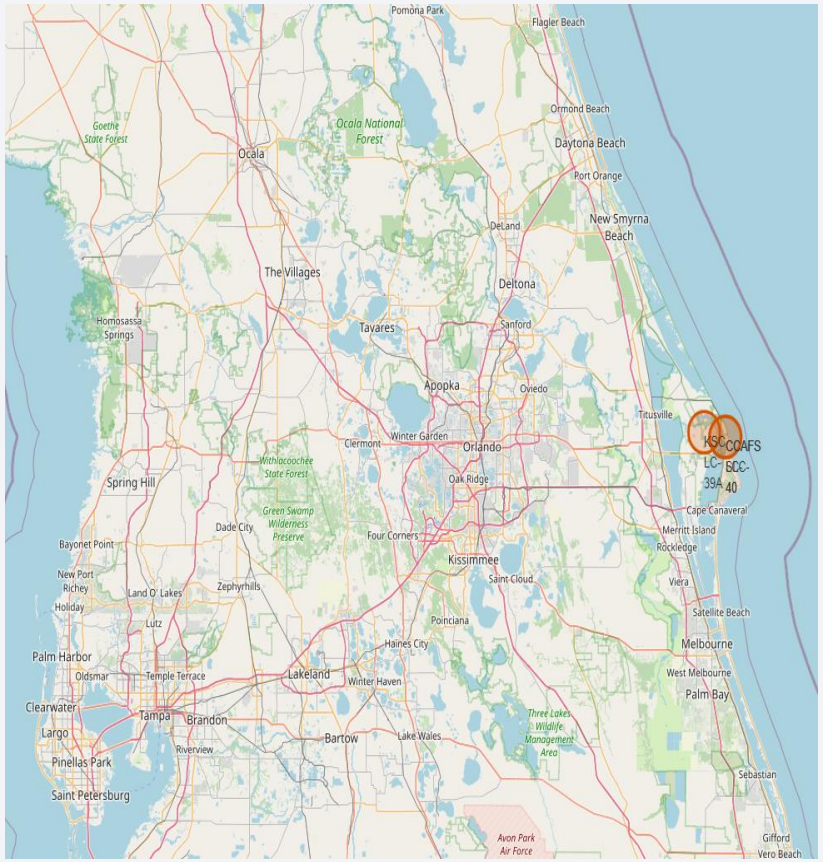
landing__outcome	qty
No attempt	9
Failure (drone ship)	5
Success (drone ship)	4
Controlled (ocean)	3
Failure (parachute)	2
Success (ground pad)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

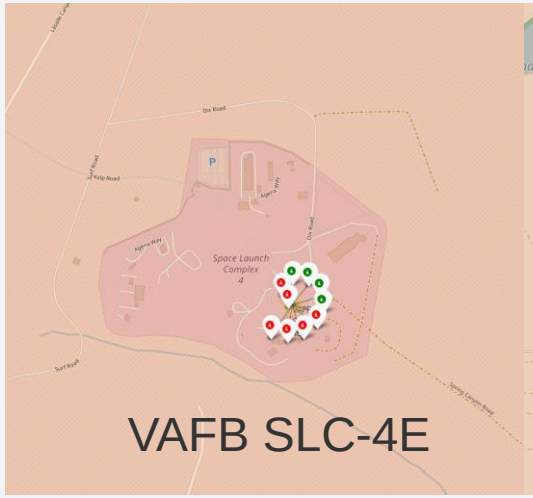
Launch Site's

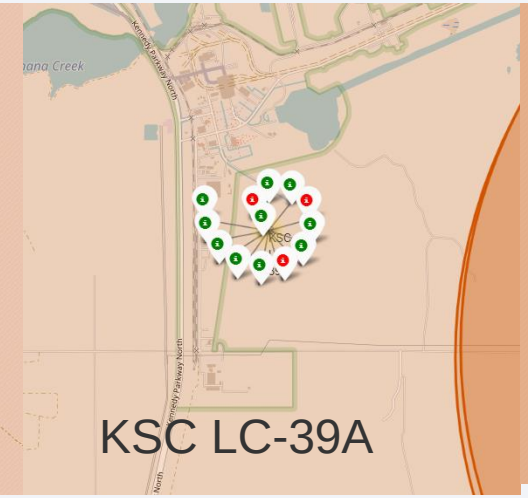


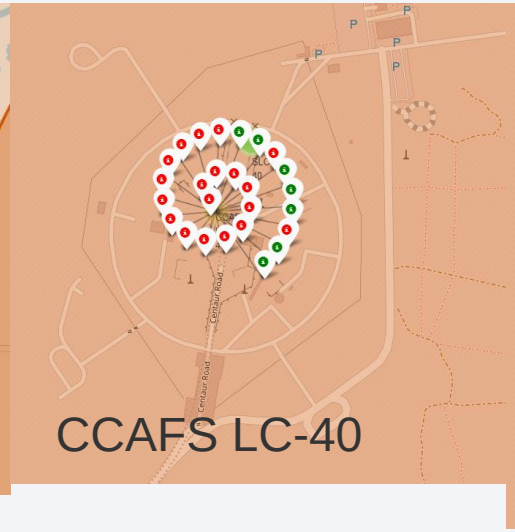


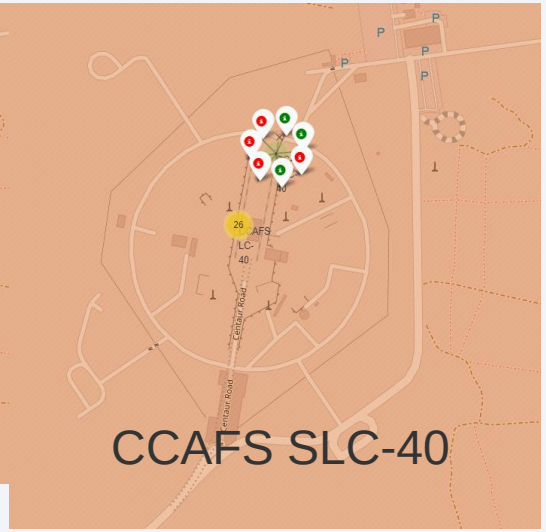
We can see that all of SpaceX's launch sites are located near the coast for safety reasons.

Failed and Successful launches









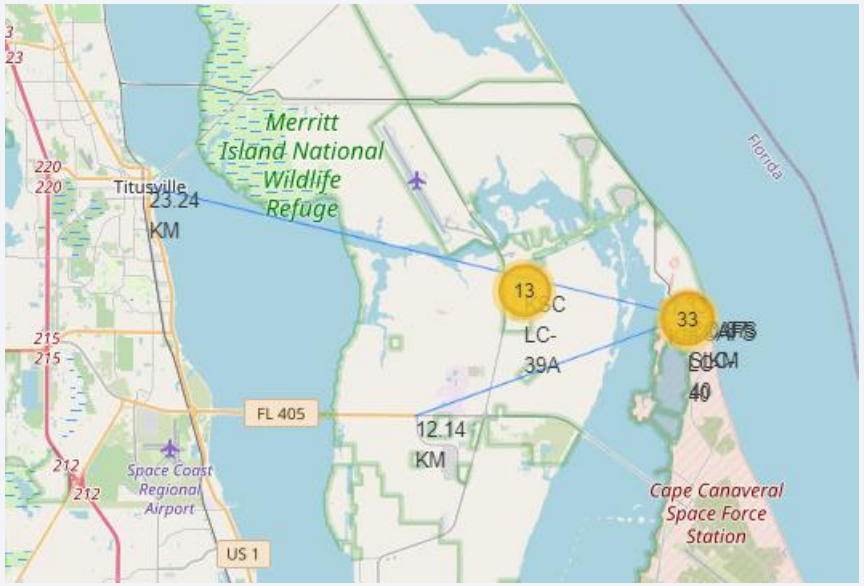
California launch

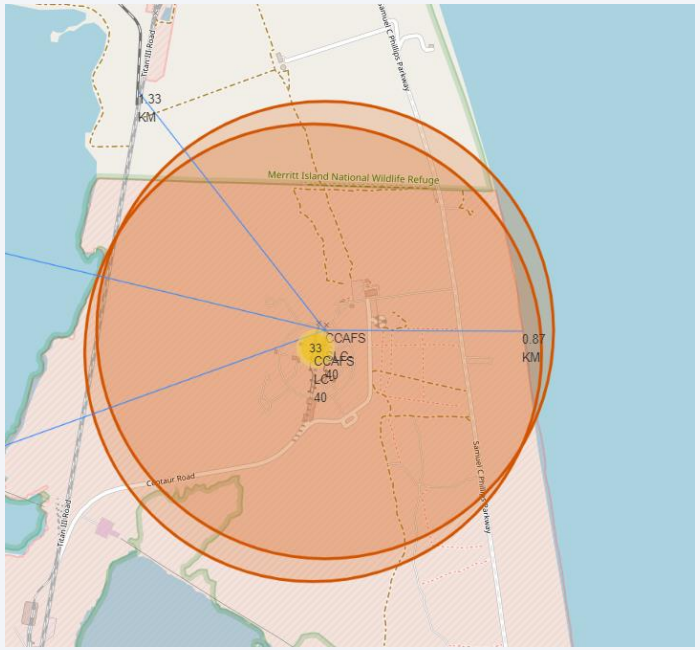
Florida launch

The green marker indicates successful launches, whereas the red marker represents failed launches.

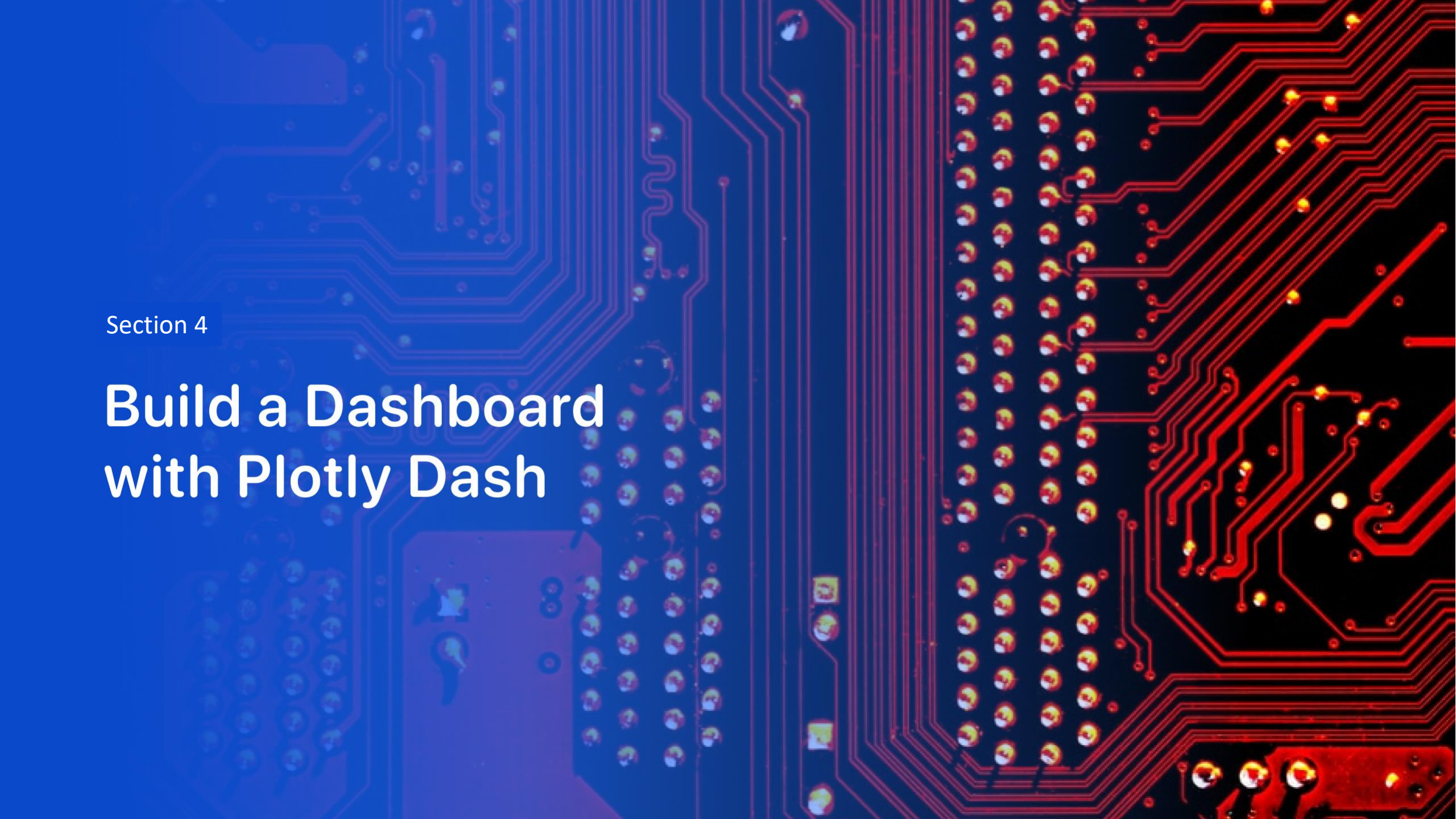
Based on this, the majority of launches take place in Florida, and the largest number of successful launches take place in KSC LC-39A.

Launch Site Distance to public spaces

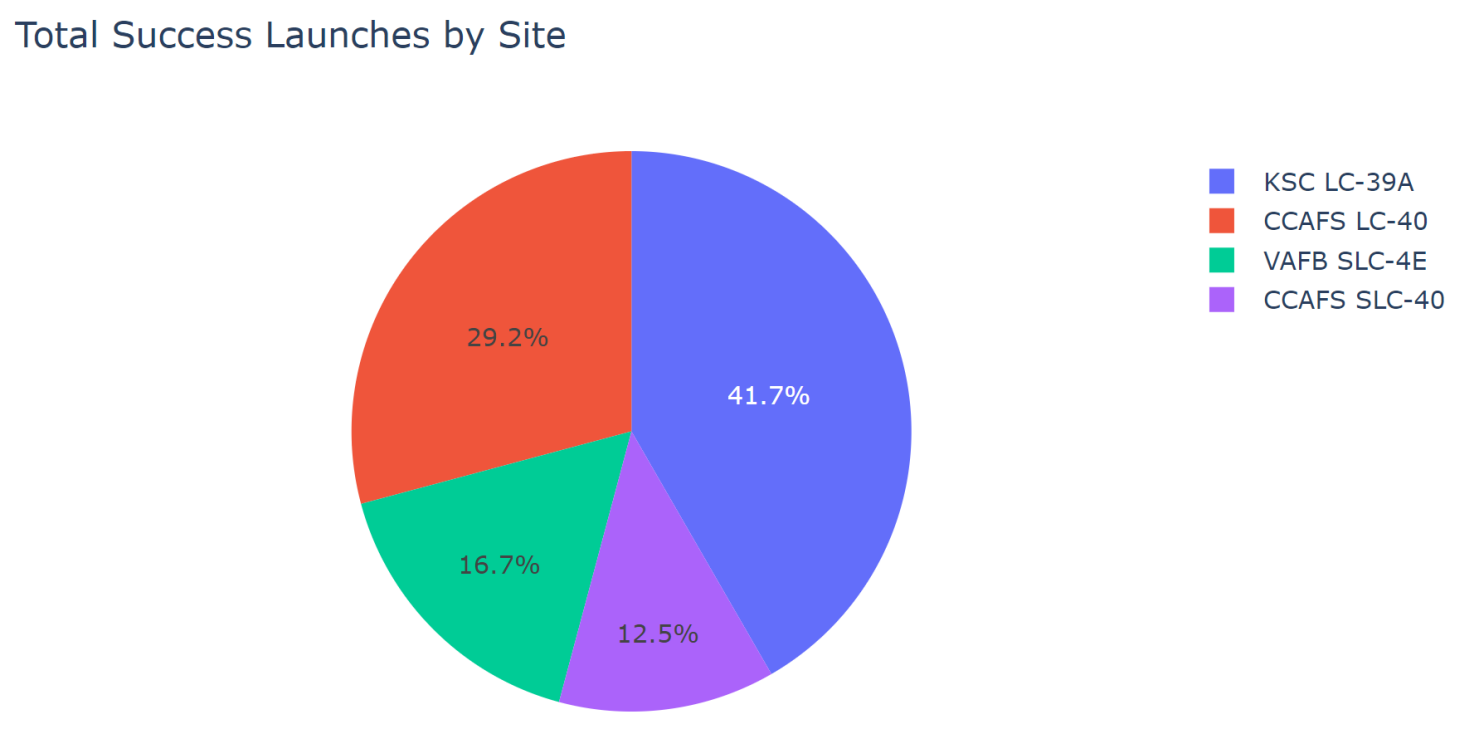




The map shows that the launch location is close to a nearby town, and that there is a railroad nearby that is most likely to deliver rocket parts, as well as a highway utilized by SpaceX employees.

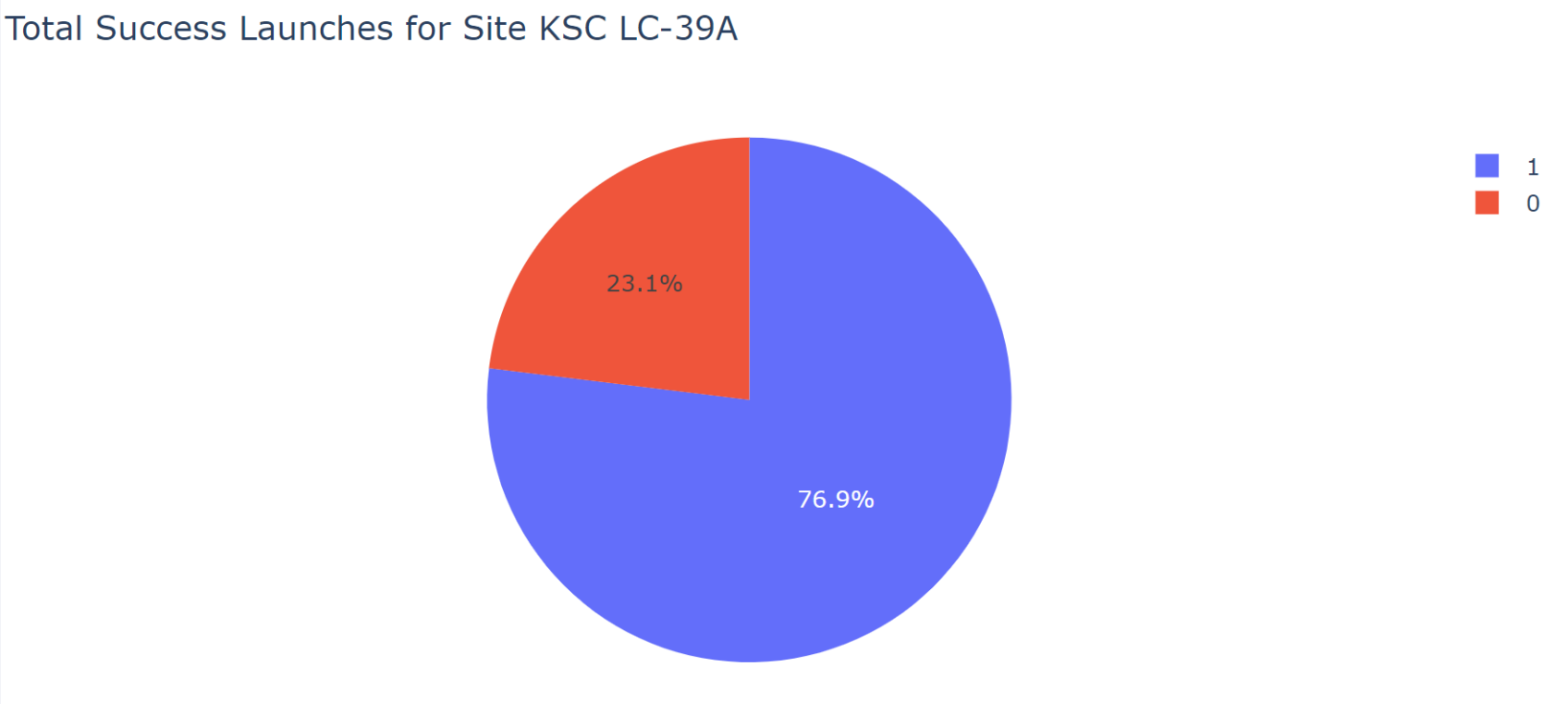

Section 4
Build a Dashboard with Plotly Dash

Successful launches for all sites



The graph shows that the most successful location is KSC LC-39A.

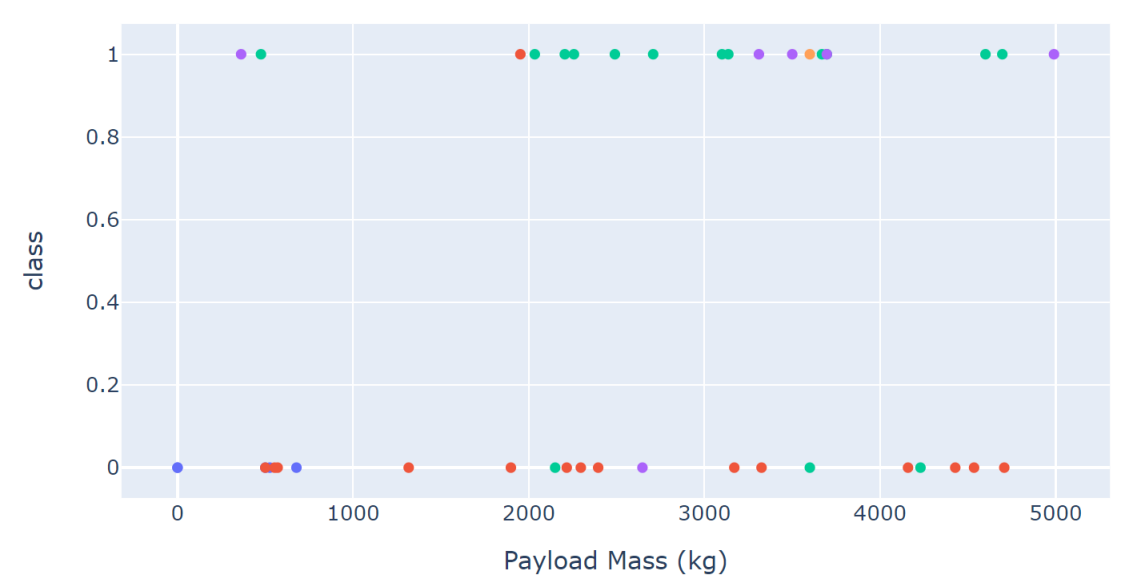
The highest launch-success ratio



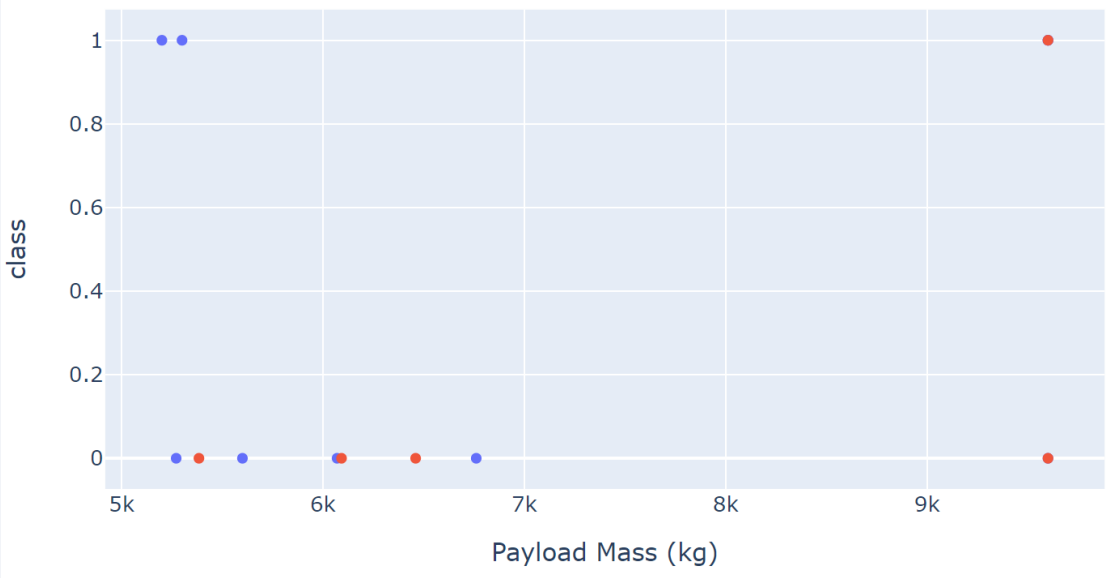
KSC LC -39A has the highest success rate of all sites at 76.9%.

Payload vs. Launch Outcome scatter plot for all sites

Light weight payload



Heavy weight payload



We can observe that the most effective payloads are those between 2000 and 4000 kg.

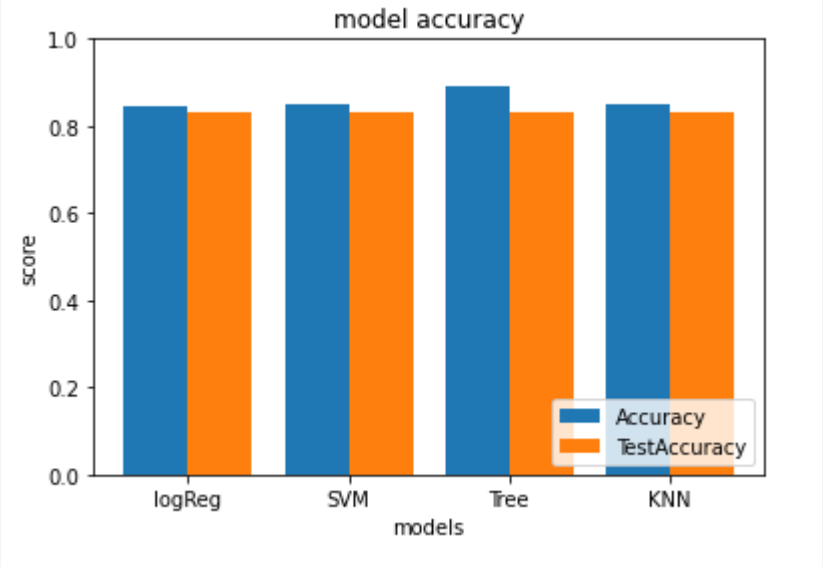
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- We used a bar chart to represent the models, and as we can see, virtually all of them performed similarly in Test Accuracy and Accuracy.

- Based on the accuracy, we may conclude that the Decision Tree produced the best results.

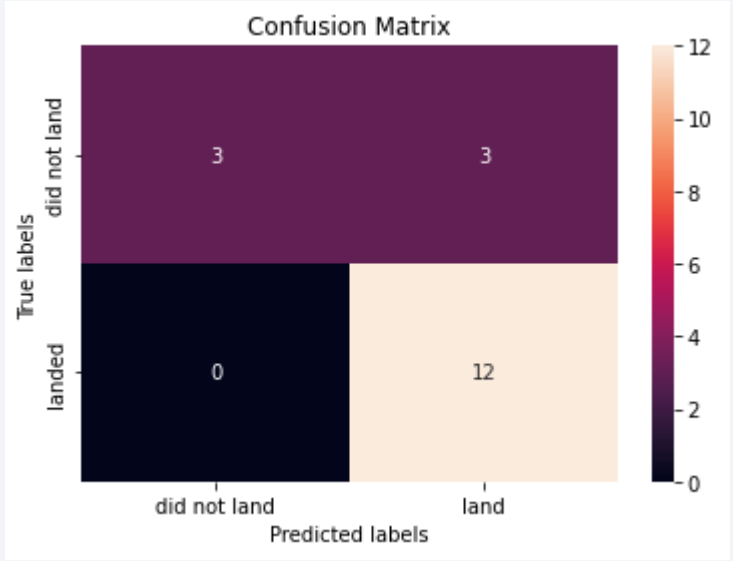


	Accuracy	TestAccuarcy
logReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.88929	0.83333
KNN	0.84821	0.83333

Confusion Matrix

- The confusion matrix for the decision tree classifier demonstrates that the classifier can differentiate between the various classes.

- The main issue is false positives.
  That is the classifier considers an unsuccessful landing to be a successful landing.



Conclusions

- KSC LC-39A has had the most successful launches of any site, with 76.9% success rate.
- The best preforming algorithm with this problem is the Decision tree classifier.
- Payloads weighing between 200 and 4000 KG performed the best of all payloads.
- Depending on the orbits, payload mass might be a factor to consider for mission success. Some orbits need either a low or hefty payload mass.
- GEO, HEO, SSO, and ES L1 orbits have the highest success rates.

Thank you!