IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chavone Garza
Aug 12, 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

In this capstone project, our goal is to predict whether the SpaceX Falcon 9 first stage will successfully land. By determining the likelihood of a successful landing, we can estimate the cost of a launch. This prediction will be achieved using various machine learning classification algorithms.

Our methodology involves:

- Data Collection
- Data Wrangling and Preprocessing
- Exploratory Data Analysis
- Data Visualization
- Machine Learning Prediction.

Throughout our investigation, we identified features of rocket launches that correlate with successful or failed outcomes. Ultimately, we conclude that the Decision Tree algorithm may be the most suitable for this problem.

Introduction

The primary objective of this capstone project is to predict the success of the Falcon 9 first-stage landing. SpaceX emphasizes its capability to reuse the first stage of its rockets, highlighting this on its website where they advertise launch costs at $62 million, compared to other providers that charge upwards of $165 million. These significant savings are largely due to the reusability of the first stage. By predicting the likelihood of a successful landing, we can estimate the cost of a launch, which is crucial for alternative companies looking to compete with SpaceX for rocket launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
    - Data was gathered using two approaches: retrieving data from the SpaceX API and web scraping launch data from a Wikipedia page.
- Perform data wrangling
    - data – by filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Build classification models to predict landing outcomes. Tune and evaluate these models to identify the best performing model and optimal parameters.

Data Collection

- Two forms of data collection was used for this presentation.
- First data was collected using SpaceX-API
- Next data was collected by web scraping—using SpaceX Wikipedia page.

Data Collection – SpaceX API

1. Request data from the SpaceX API using a GET request
2. Using JSON file, normalize data into data frame using dataX = response.json()
3.
3. With customized functions, extract specific data columns
4. Create new dictionary from data
5. From new dictionary, create data frame using Pandas
6. Filter data frame to include Falcon 9 launches only
7. With calculated .mean(), replace missing data for Payload Mass
8. Export data to CSV file

Github URL:

SpaceX Data Collection- API

Data Collection - Scraping

1. Request Falcon 9 data from Wikipedia page
2. Create a Beautiful Soup using HTML response
3. From the HTML table header, extract columns names
4. Parse through tables to collect data
5. Create a dictionary from data using Pandas
6. Create a data frame using dictionary
7. Export data to CSV file

Github URL:

Data Collection, Webscraping

Data Wrangling

1. Calculate number of launches from each launch site using .value_counts
2. Calculate number and occurrence of each orbit
3. Calculate number and occurrence of mission outcome of the orbits
4. Create landing outcome label from Outcome column using df['Class'] = df['Outcome']

Github URL:

[SpaceX Data Wrangling]

EDA with Data Visualization

To better visualize the relationship between the data use

- Scatter plot, sns.catplot
  - Used to analyze the relationship between FlightNumber vs Payload Mass
- Bar chart, sns.barplot
  - Used to analyze the relationship between success rate and orbit

To visualize the trend for the data

- Line chart, sns.lineplot
  - Used to visualize launch success yearly trend

Github URL:
Falcon 9 Landing prediction with Data Visualization

EDA with SQL

- Used SELECT DISTINCT to find unique launch sites
- Displayed 5 launch sites beginning with 'CCA' using LIKE LIMIT 5
- Used SUM() for the total payload mass, launched by NASA (CRS)
- Used AVG() to display average payload mass carried by booster version F9 v1.1
- Used min(Date) to find first successful landing outcome
- Listed total successful and failure mission outcome using Count() GROUP BY ""
- Ranked landing outcomes using BETWEEN AND

Github URL:
[EDA with SQL, peer reviewed notebook assignment]

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

- Added marker objects to display all launch sites on a map
  - Used circles to cluster groups
- Along with the successful and failed launches for each.
  - Assigned color green to successful outcome and red to failed.
- Used line objects to calculate and visualize the distances between each launch site and its nearby areas.
  - Nearby area coordinates found using mouse position

Github URL: *Location with Folium*

Build a Dashboard with Plotly Dash

Added dropdown List with Launch Sites that allow users to:

- Select all launch sites or a certain launch site Pie Chart Showing Successful Launches
- See successful and unsuccessful launches as a percent of the total Slider of Payload Mass Range
- Select payload mass range Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version
- See the correlation between Payload and Launch Success

Github URL:
Dashboard with Plotly Dash

Predictive Analysis (Classification)

For classification model

- Create a class columns using Numpy Arrays

Transform data

- Standardized that data using StandardScaler.fit

Train and Test data

- Train, Test, split data using train_test _split

Apply SVM, Decision Trees, K-Nearest Neighbours and Logistic Regression.

Identify accuracy using confusion matrix

Github URL
SpaceX launch Dashboard

Results

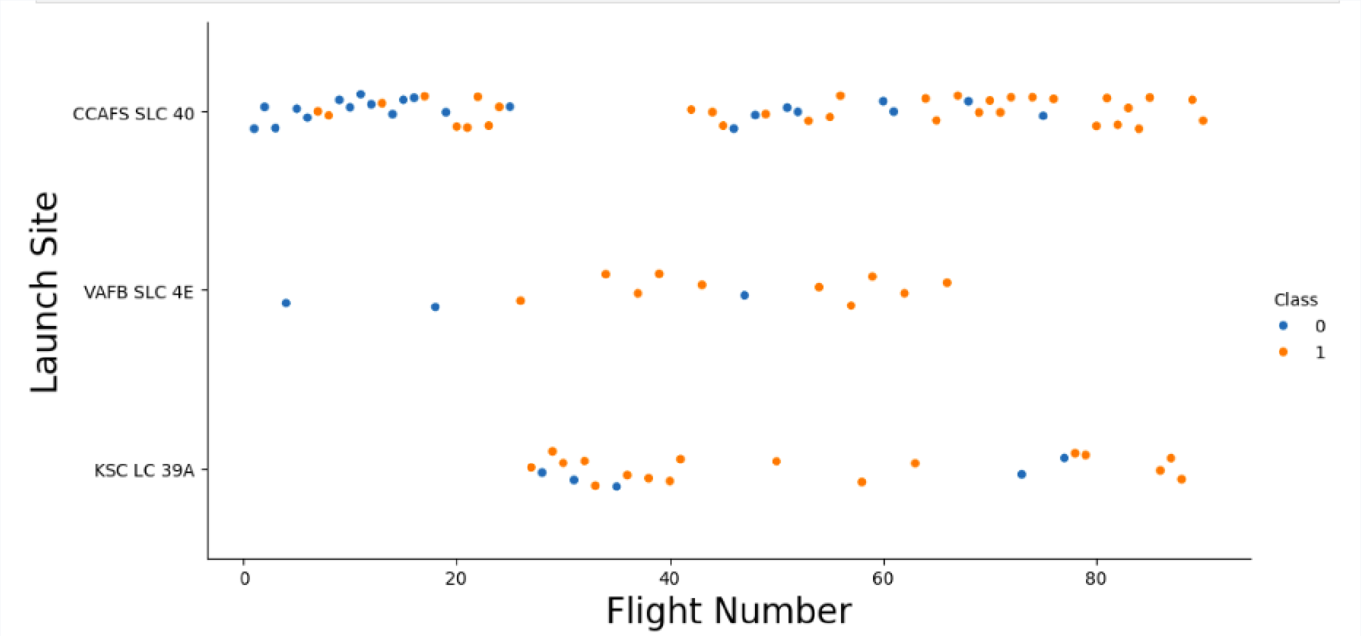
- The launch had a rate of 66.66% success
- The launch sites were located close to coasts
- The decision tree was the best predictive model, predicting the correct outcome 94% of the time

Section 2

Insights drawn from EDA

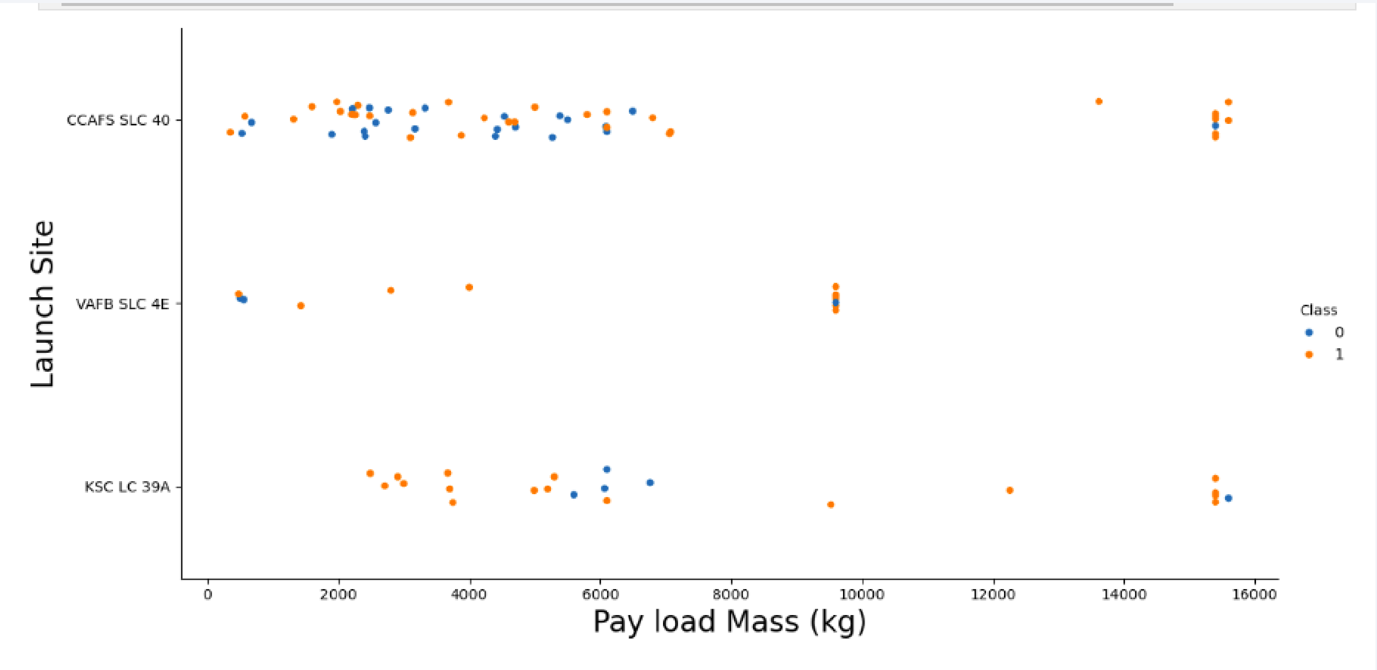
Flight Number vs. Launch Site

- As shown in this plot, the earlier launch missions were less successful than the later launches
- Blue dot represent fail
- Red dots represent success



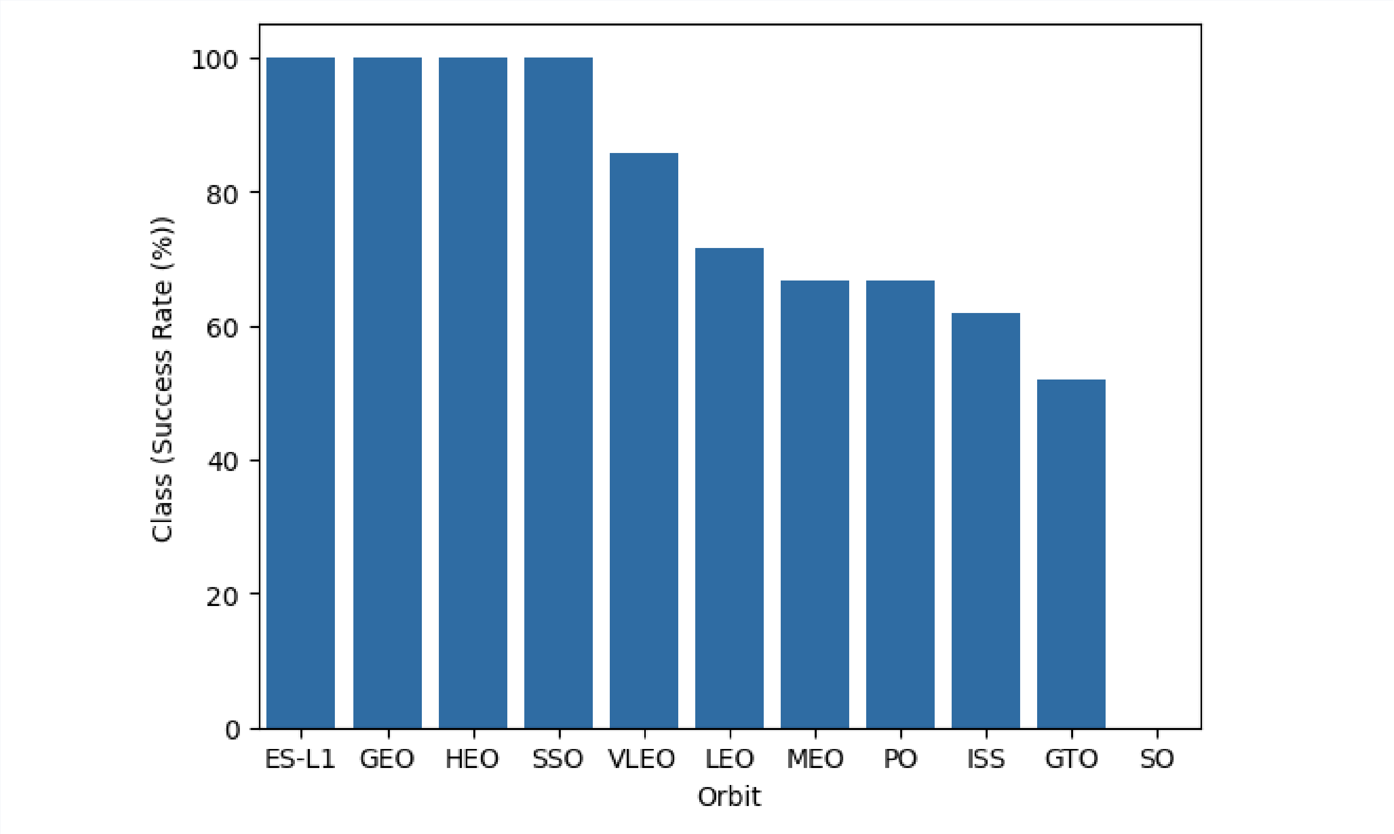
Payload vs. Launch Site

- The difference between successful and failed launches aren't a metric to use for lower Payloads.
- Payloads above 10k seem to have a higher success rate.
- The blue dots represent failed
- The Red dots represent successful



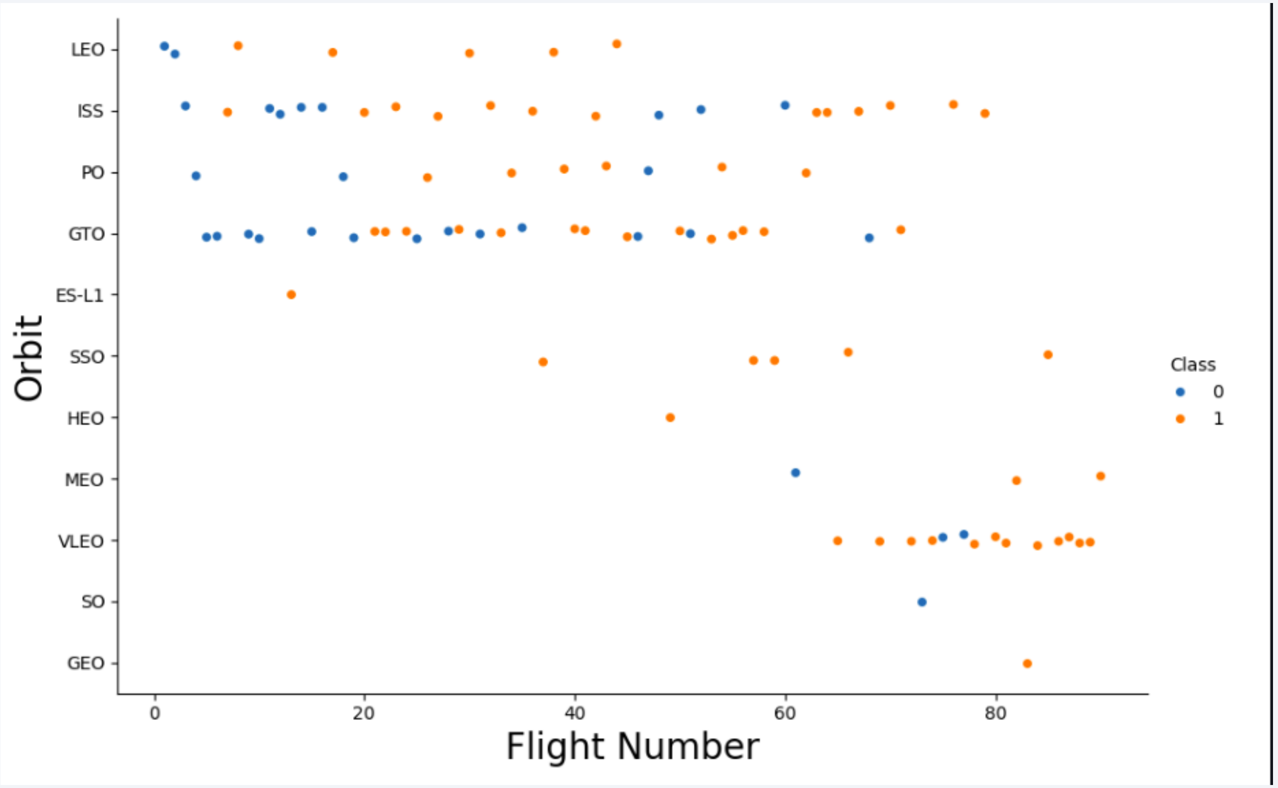
Success Rate vs. Orbit Type

- ES-L, GEO, HEO & SSO have a higher mission success with 100% rate.
- VLEO has 85% success
- LEO, MEO, PO, ISS, and GTO success rate is between 50% and 80%.
- SO failed all at 0%



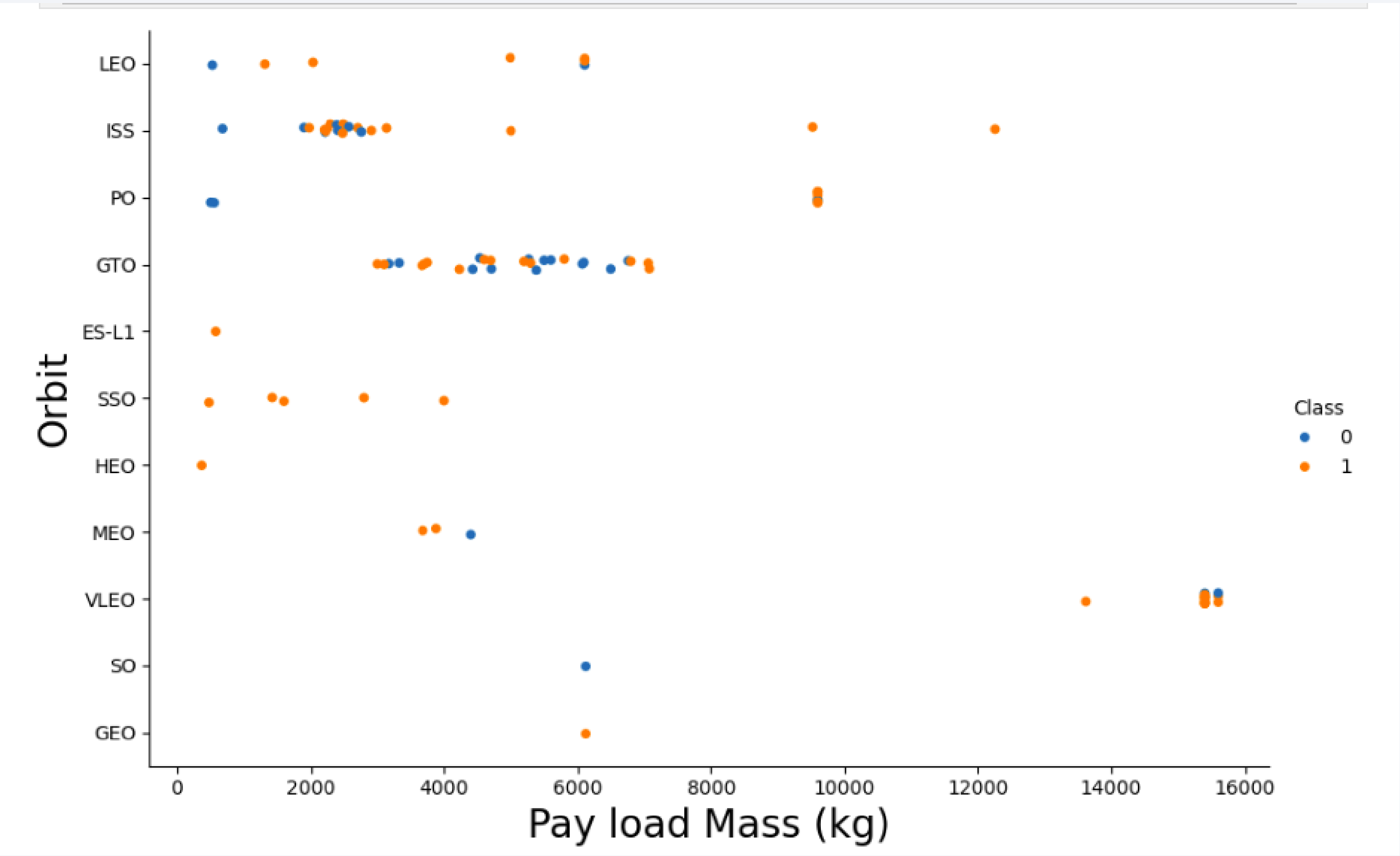
Flight Number vs. Orbit Type

- Blue dot represent successful launch, Orange dot represent failed.
- More Orbits generally shows a higher success rate
- ES-L1, GEO and SO do not follow the trend



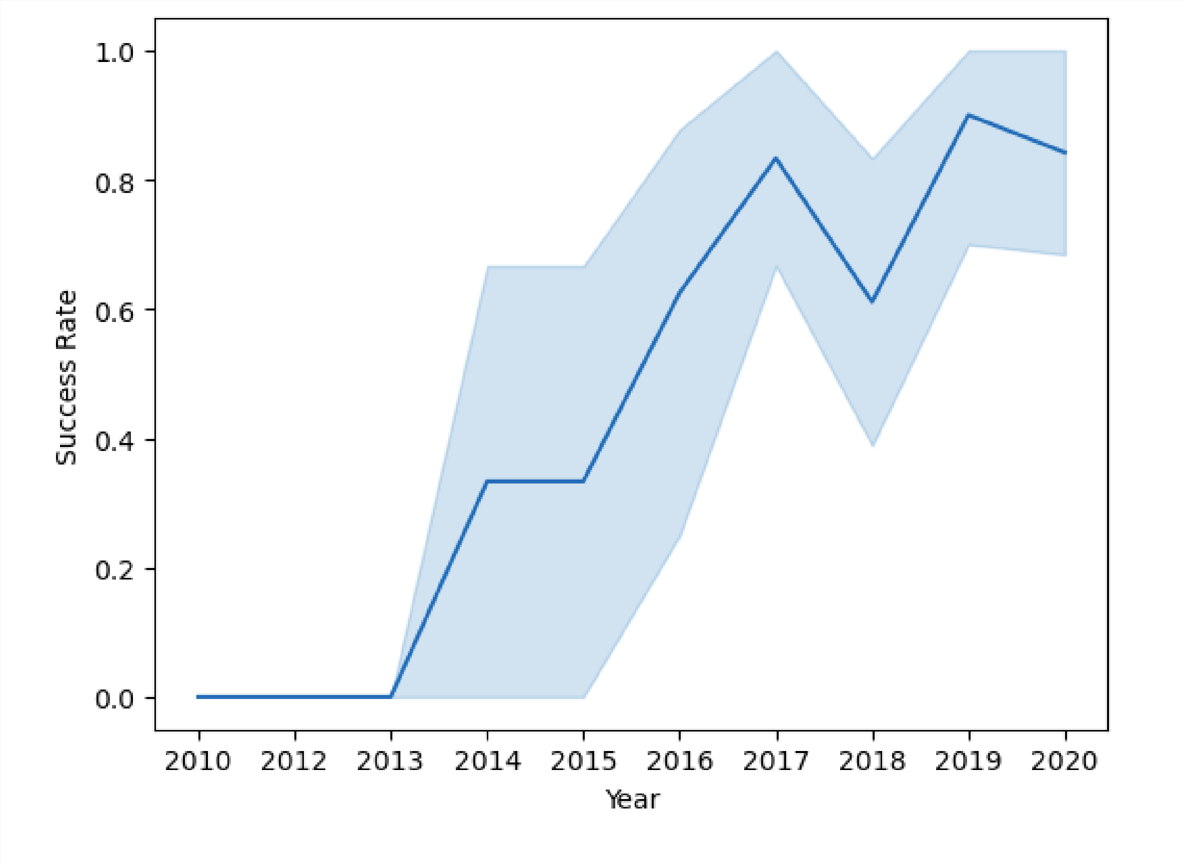
Payload vs. Orbit Type

- The blue dots represent failure, the orange represent success
- Higher payloads generally had a higher success rate
- ES-L, SSO and HEO all buck the trend.



Launch Success Yearly Trend

- The launches became successful over time
- 2018 bucks the trend.



All Launch Site Names

Using the SELECT DISTINCT command to obtain the launch sites, four launch sites was observed:

1. CCAFS LC-40
2. VAFB SLC-4E
3. KSC LC-39A
4. CCAFS SLC-40

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

Launch Site Names Begin with 'CCA'

- Five launch sites from CCA using From Like cause, 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	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

Total Payload Mass

Used the sum() clause to calculate Total Payload mass for NASA

4]:

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- The avg() and Where were used to give the average payload for the F9 v1.1

Payload Mass Kgs	Customer	Booster_Version
2534.6666666666665	MDA	F9 v1.1 B1003

First Successful Ground Landing Date

Using the min(date), and WHERE for success ground pad, the first successful ground landing was obtained as 01/05/2017.

```
: [('01-05-2017', 'Success (ground pad)')]
```

Successful Drone Ship Landing with Payload between 4000 and 6000

Using the BETWEEN AND clause to obtain successful drone ship landing between 4000 and 6000. There was four booster returned:

1. F9 FT B1022
2. F9 FT B1026
3. F9 FT B1021.2
4. F9 FT B1031.2

]:

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

Total Number of Successful and Failure Mission Outcomes

Used count() to obtain total successful and failed, filtered using GROUPBY Mission Outcomes.

- 99 successful misions
- 1 in flight failure
- 1 successful with unclear payload status

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

Boosters Carried Maximum Payload

Using a subquery max() function, 12 missions were shown to carry maximum payload:

- 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

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

Used Date & LIKE subquery to obtain the drone ship failures for 2015, shown to be in January and April

```
%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_NAME, \
    LANDING__OUTCOME AS LANDING__OUTCOME, \
    BOOSTER_VERSION AS BOOSTER_VERSION, \
    LAUNCH_SITE AS LAUNCH_SITE \
    FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND "DATE" LIKE '%2015%'
```

month_name	landing__outcome	booster_version	launch_site
JANUARY	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
APRIL	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Used SELECT DATE Count() with Between 6/1/2010 AND 3/20/2017 to filter the date with subquery GROUPBY DATE and ORDERBY as desc for landing outcome in descending order.

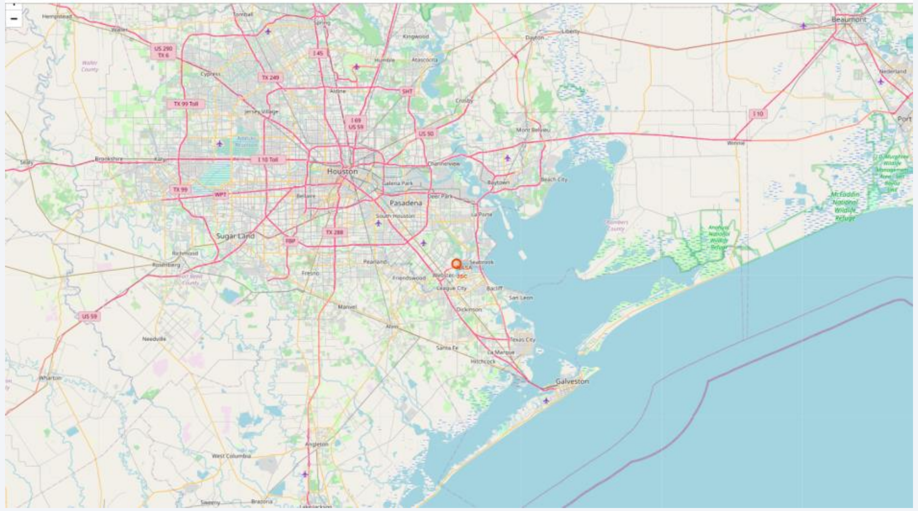
DATE	COUNT
2015-12-22	1
2016-04-08	1
2016-05-06	1
2016-05-27	1
2016-07-18	1
2016-08-14	1
2017-01-14	1
2017-02-19	1

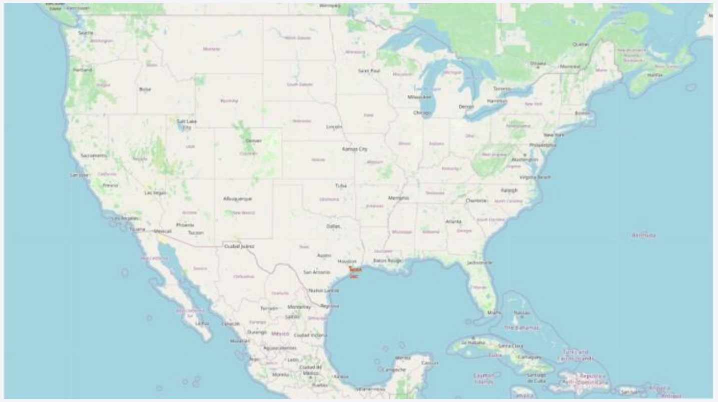
Section 3

Launch Sites Proximities Analysis

Starting Launch Site

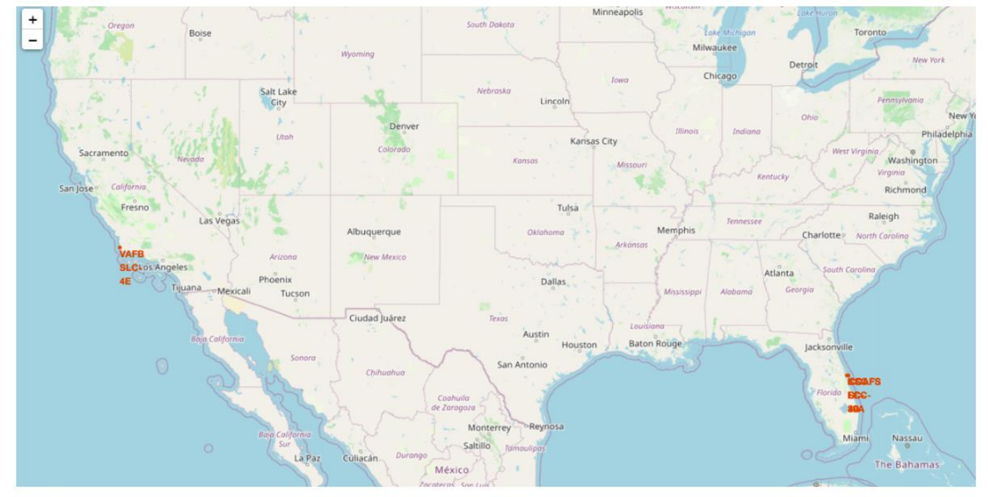
- NASA Johnson Space Center is the starting launch site, located near HoustonTexas off the coastline.
- Used folium.circle to highlight the launch site

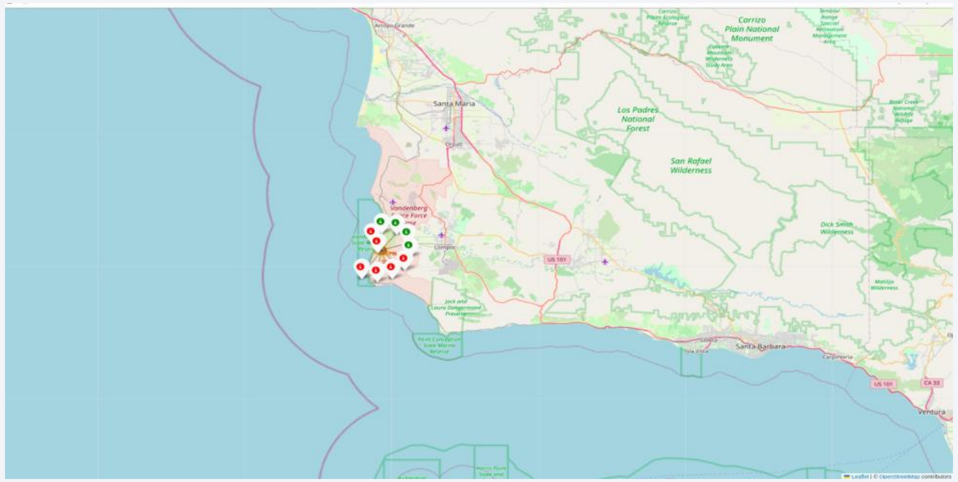


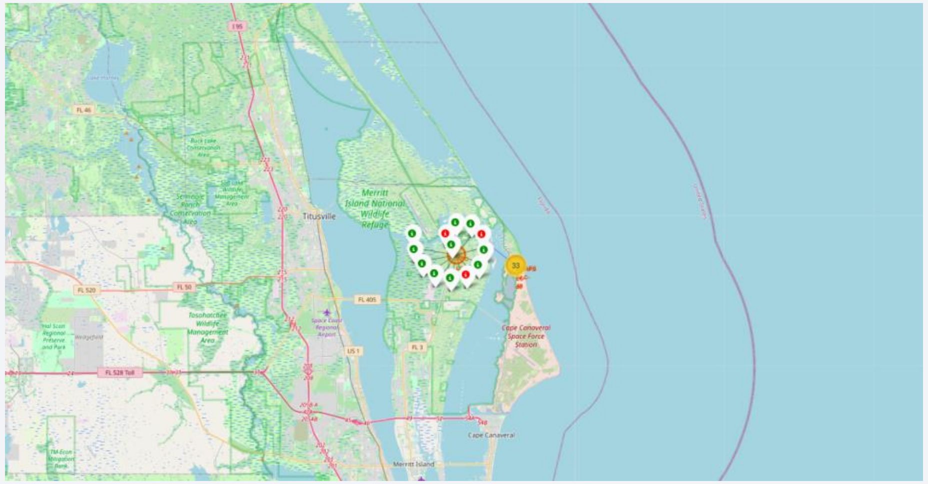


Bicoastal launch sites

- Launch sites located off the coast of California and Florida
- Green dots indicate successful launch
- Red dots indicate launch failures

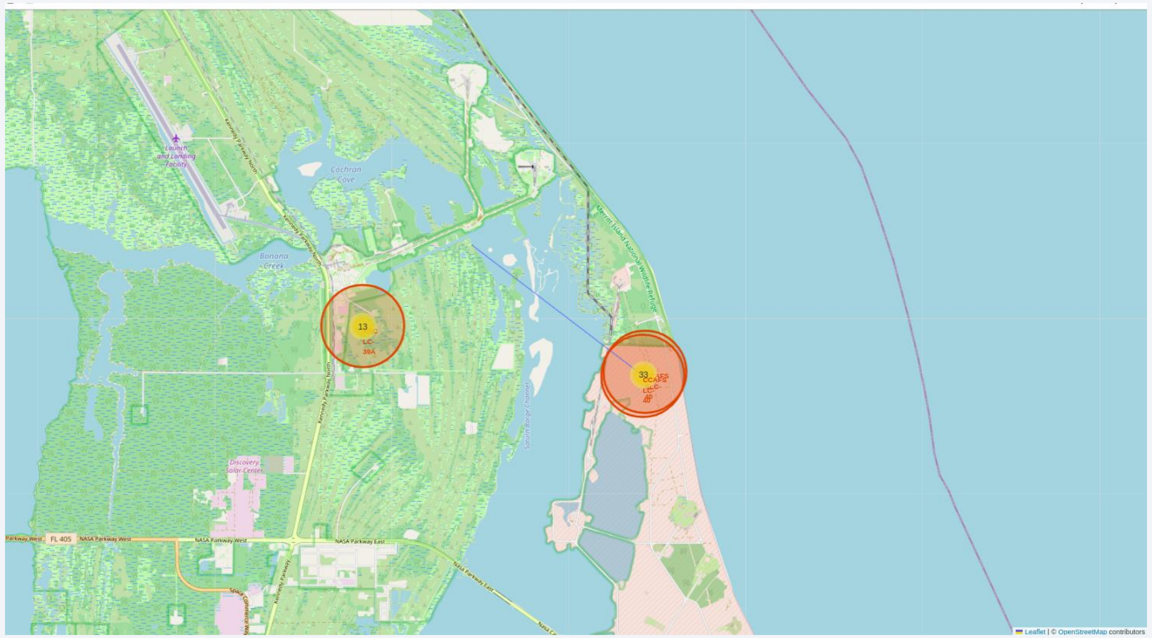


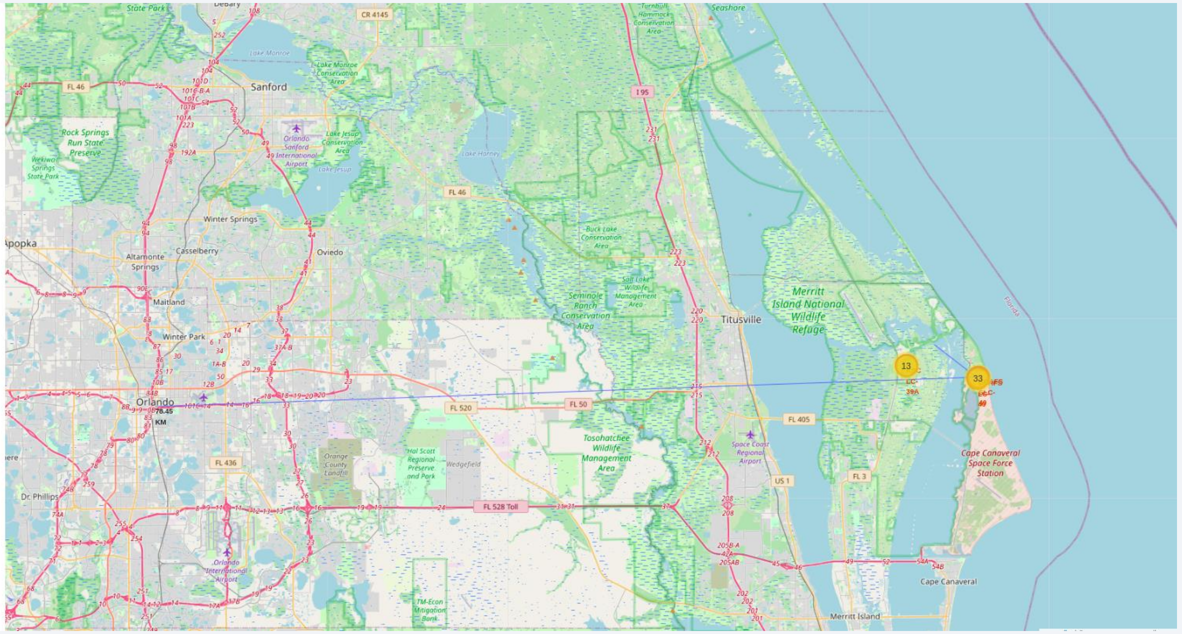




Proximity and distance for CCAFS SLC-40

- Launch sites are closer to the equator and coasts.
- This distance is further from the city center of Orlando, FL.





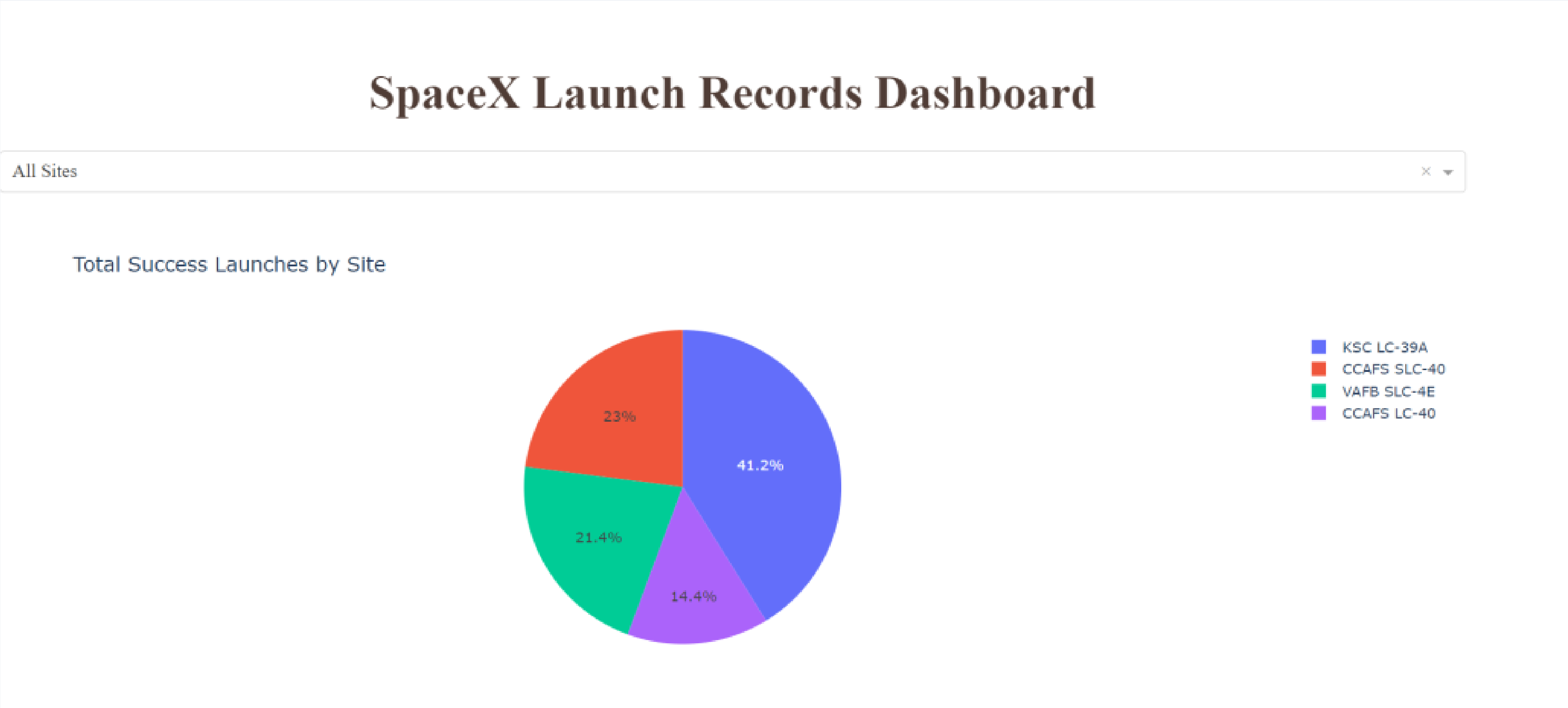
Section 4

Build a Dashboard with Plotly Dash

SpaceX Success by Launch Site

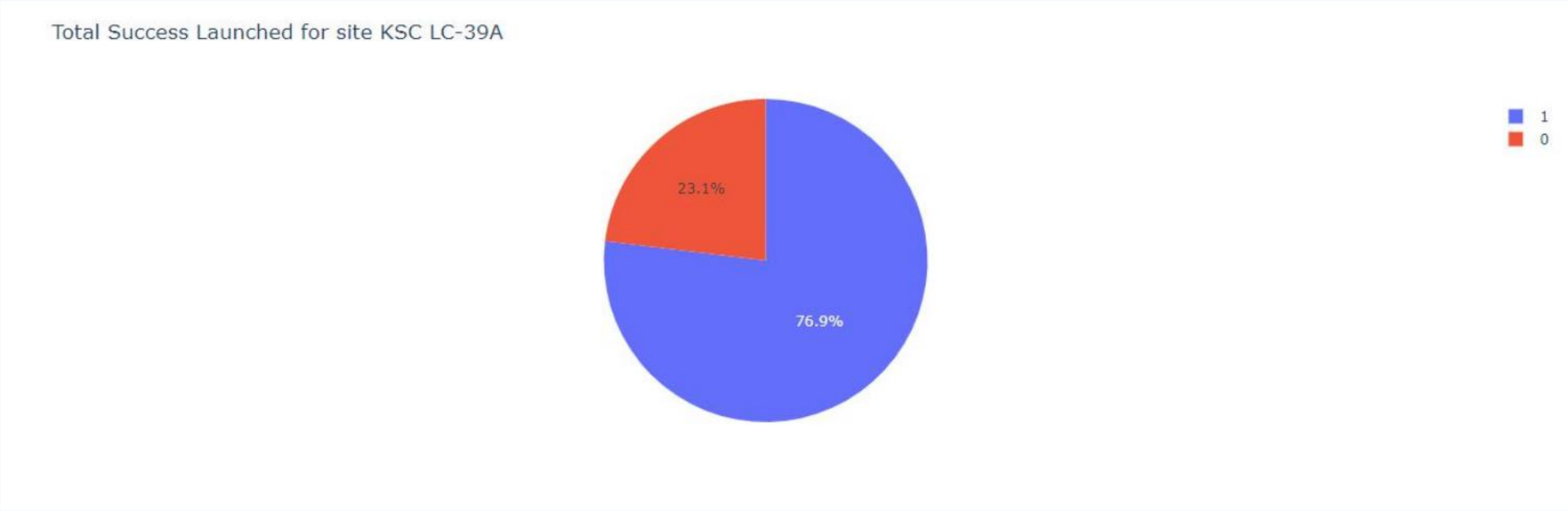
The pie chart shows the percentage of success rate by site.

- KSC LC-39A has highest success rate at 41.2%
- CCAFS SLC-40 has 23% success rate
- VAFB SLC-4E has 21.4% success rate
- CCAFS LC-40 has the lowest at14.4% success rate



SpaceX Highest Success Rate by Launch Site

- • The KSLC-39A launch site has the highest success rate with 76.9%.



Payload vs Class for Outcome Success

- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- V1.1 has the highest success rate in both weight ranges

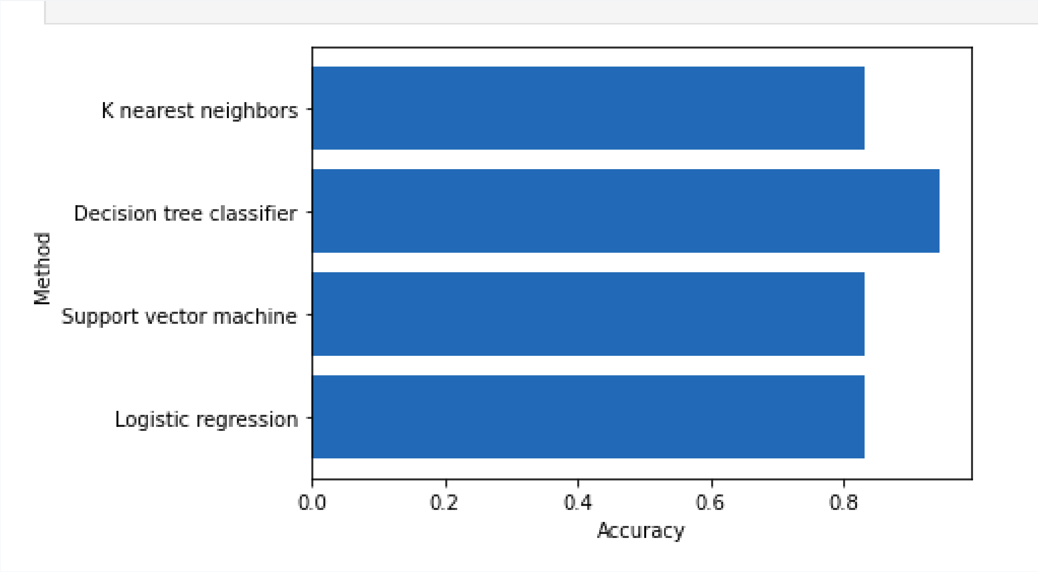
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Four models using classification method are:

1. Logistic Regression (LR),
2. Support Vector Machine (SVM),
3. Decision Tree Classifier
4. K nearest neighbor (KNN)



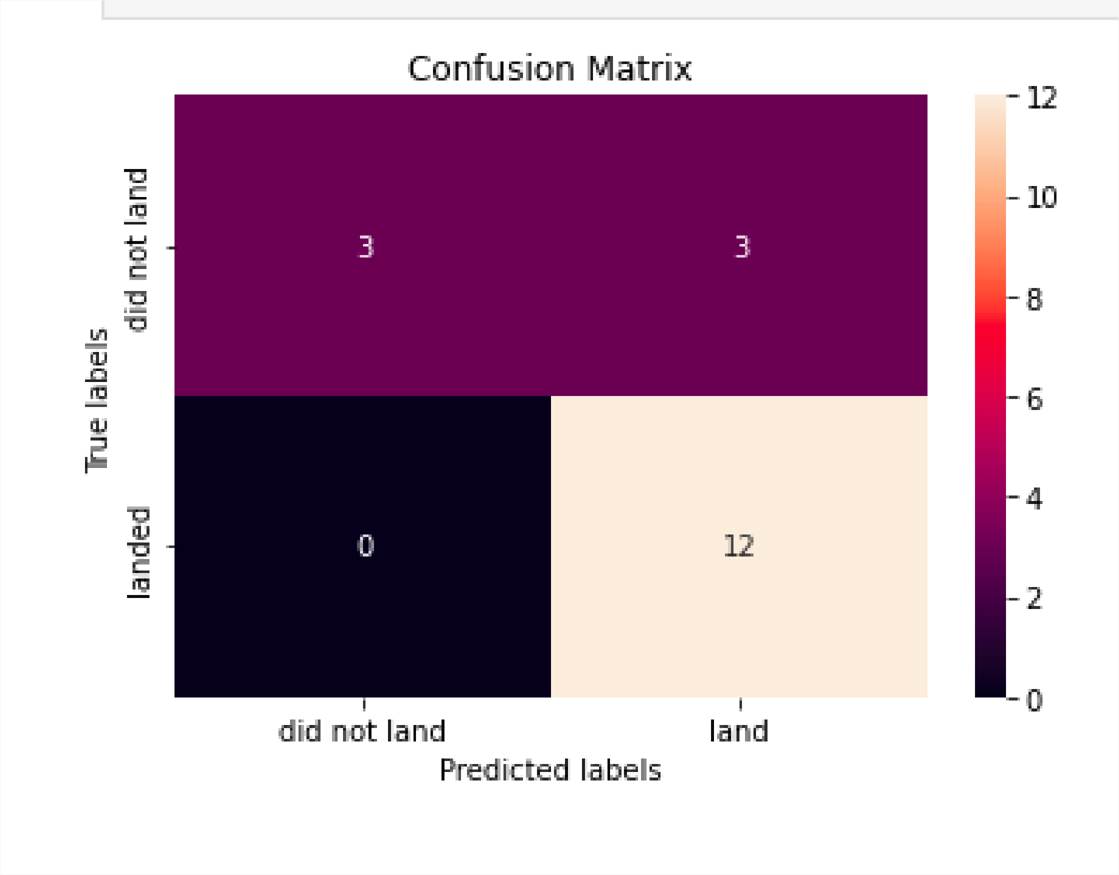
	method	accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.944444
3	K nearest neighbors	0.833333

Decision Tree Classifier is the best method with a 94.44% accuracy.

Confusion Matrix

A confusion matrix summarizes the performance of a classification algorithm

- With 18 predictions, the model predicted 15 successful landings and 0 failures.
- The model was correct for 12 successful landing (true positive) but misidentified 3 (false positive).
- The model predicted 3 successful landings, though there were 0 (false negative).
- This confusion matrix is fairly accurate to predict successful landings.



Conclusions

- SSO, HEO, GEO, and ES-L1 had the highest success in landing with 100% success.
- Payload is not a good indicator of successful landings.
- Most launch sites are near the coast and closer to the equator.
- Over time, with more launches, the success rate went up.
- The Decision Tree Classifier was the most useful method with an accuracy of 94.44%.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!