IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daleffe Santiago
February, 2024

Outline

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

Executive Summary

- **Summary of methodologies:**
  - Data collection using API and Web Scraping
  - Data wrangling and preprocessing
  - EDA with Pandas and SQL
  - Visualization using Matplotlib and Seaborn
  - Interactive maps and dashboards with Folium, Plotly and Dash
- **Summary of all results**
  - EDA
  - Geospatial analysis
  - Predictive analysis

Introduction

- **Project background and context**
  - In this project we are predicting if Falcon 9 first stage, from SpaceX, will land successfully. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - Features that predict if the rocket will land successfully.
  - The relationship between the rocket and the features that affect the landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We collected the data using the SpaceX REST API and Web Scrapping from their Wikipedia page.
- Perform data wrangling
  - Handling missing values and creating training 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

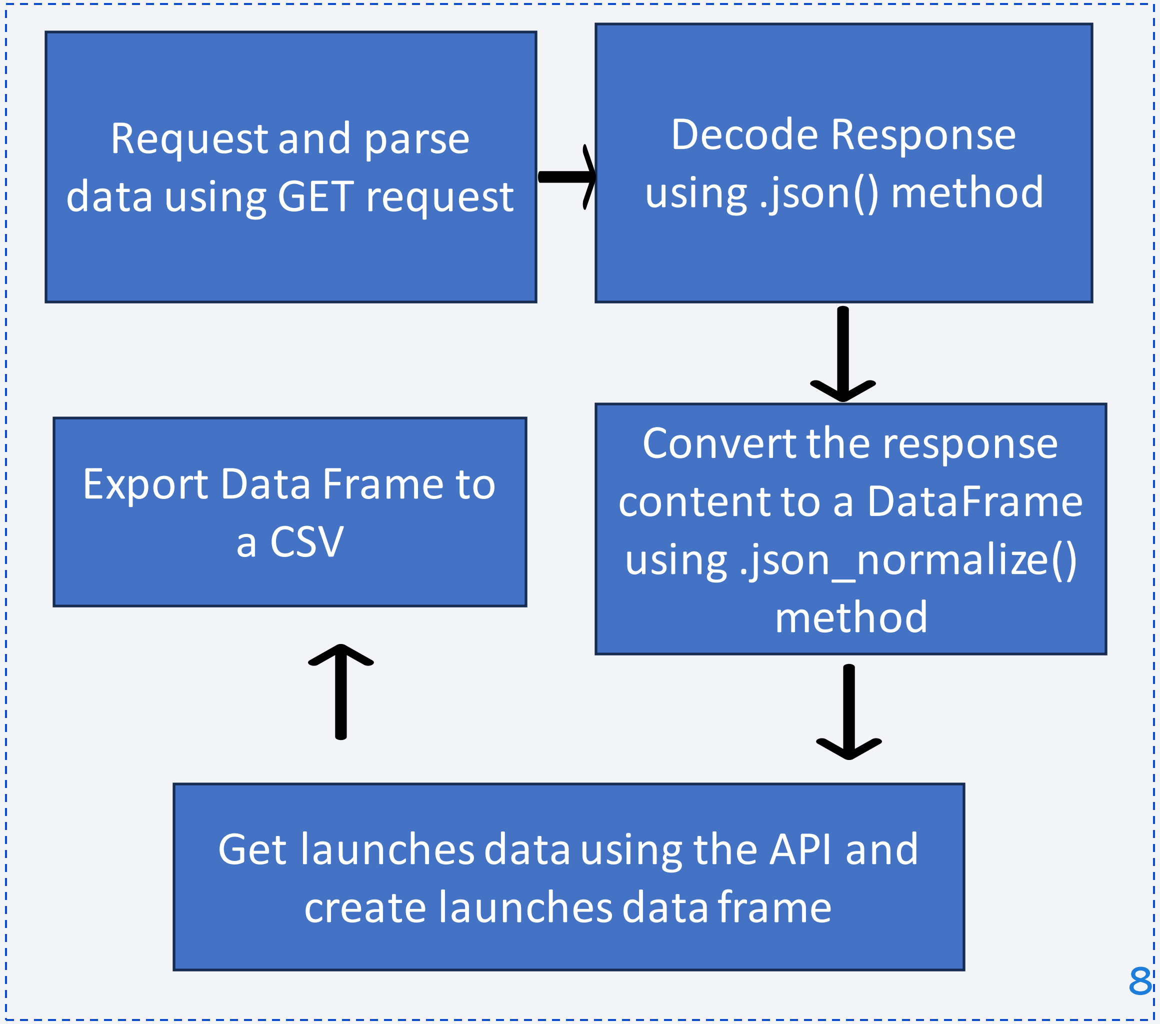
Data Collection

- Describe how data sets were collected.
  - We made requests to the SpaceX REST API and then cleaned the requested data. We have requested and parsed the SpaceX data using the GET request, and then decoded the response content as a JSON using .json() and turned it into a Pandas DataFrame using .json_normalize(). The dataframe was filtered to only include Falcon 9 launches.
- You need to present your data collection process use key phrases and flowcharts
  - We have also scraped Falcon 9 launch records HTML table from Wikipedia using BeautifulSoup. We then parsed the table and converted it into a Pandas DataFrame.

Data Collection – SpaceX API

- With this flowchart we can see the data collection process with the API.

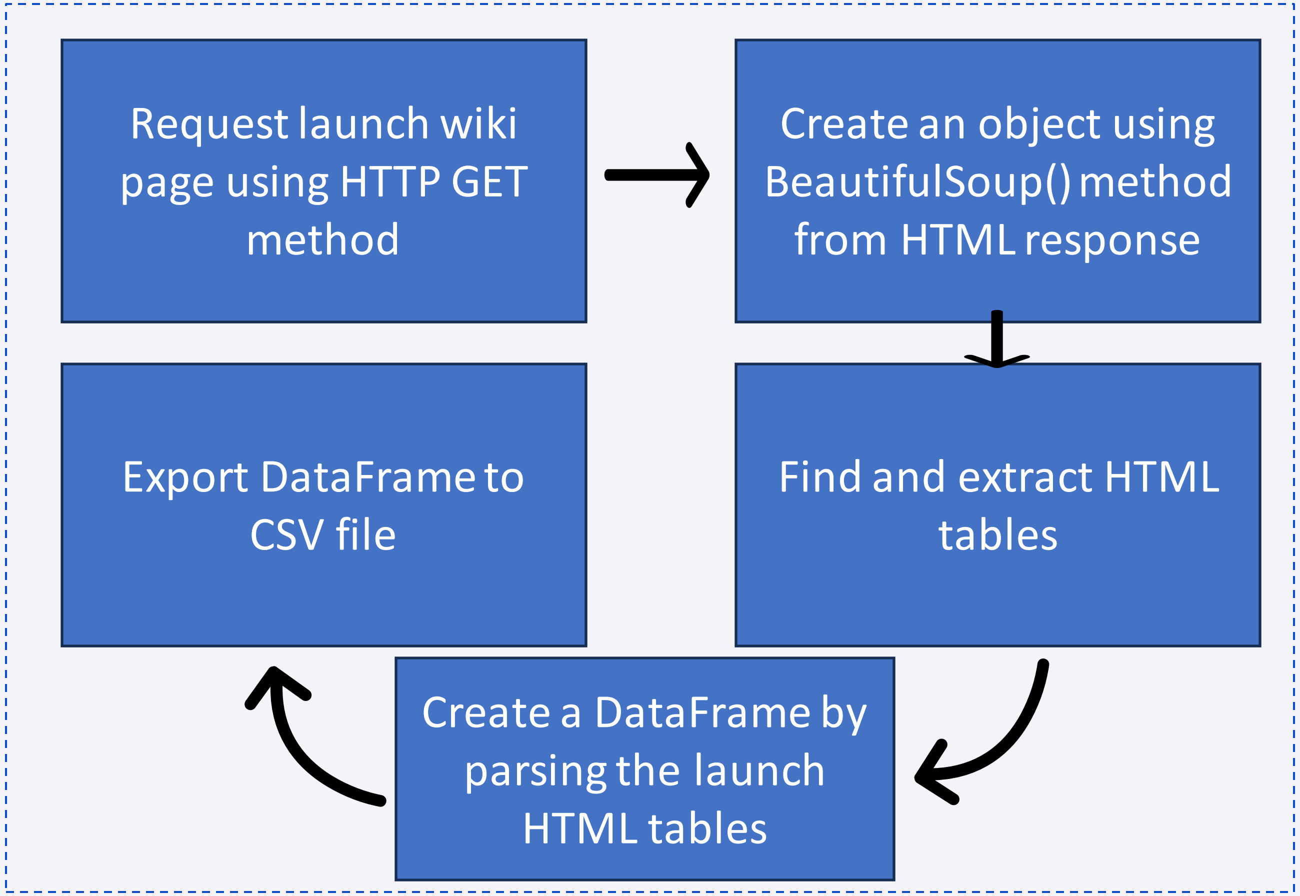
https://github.com/Poggerv2/DS-Capstone-Proyect-/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- With this flowchart we can see the Web Scrapping process with BeautifulSoup() method.

https://github.com/Poggerv2/DS-Capstone-Proyect-/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- After reading the data into a Pandas DataFrame, the following data wrangling processes were performed.
    1. Identifying the calculating % of missing values.
    2. Identifying which columns are numerical and which are categorical.
    3. Calculating the number of launches on each site, the number and occurrence of each orbit and mission outcome of the orbits.
    4. Creating a landing outcome label from outcome column.
    5. Exporting the final data to a csv file.

https://github.com/Poggerv2/DS-Capstone-Proyect-/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Data were explored by visualizing the relationship between:

  1. Flight number and launch site.
  2. Payload and launch site.
  3. Success rate and orbit type.
  4. Flight number and orbit type.
  5. Payload and orbit type.

  We have also visualized the rocket launch success yearly trend.

  Finally, we have performed feature engineering and created dummy variables of categorical columns and converted all columns to float64, to prepare data for machine learning modeling.

  https://github.com/Poggerv2/DS-Capstone-Proyect-/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- We performed the following SQL queries to explore and understand the data:

1. Display the names of the unique launch sites in the space mission.
2. Display 5 records where launch sites begin with the string ‘CCA’.
3. Display the total payload mass carried by boosters launched by NASA (CRS).
4. Display average payload mass carried by booster version F9 v1.1.
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
7. List the total number of successful and failure mission outcomes.
8. List the names of the booster_versions which have carried the maximum payload mass.
9. List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

https://github.com/Poggerv2/DS-Capstone-Proyect-/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We have created and added the following map objects to the Folium map: folium.Circle, MarkerCluster object, folium.Marker, MousePosition, and folium.PolyLine.
- These objects were added to the map to highlight specific areas, mark the success/failed launches for each site of the map, calculate the distances between a launch site to its proximities, and to draw a line between a launch site and a selected point.

https://github.com/Poggerv2/DS-Capstone-Proyect-/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The following were added to the dashboad:

  1. Interactions : Launch site drop-down input component, and a range slider to select payload.
  2. Pie chart and a scatter point plot to visualize total success by site and correlation between payload and success for all sites.
  3. Two callback functions to render pie chart and scatter plot based on the selected site dropdown.

https://github.com/Poggerv2/DS-Capstone-Proyect-/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- After loading the dataframe, we have performed the following:

1. Create a NumPy array from the column Class in the DataFrame using the method to_numpy().
2. Normalize the features DataFrame using StandardScaler().
3. Split the datasets into training and testing sets.
4. Create objects for the classification algorithms (logistic regression, SVM, decision tree, and KNN) and also create a GridSearchCV object for each of them.
5. Train each model using training data.
6. Calculate the accuracy of each model using the test data.
7. Plot confusion matrix for each model.
8. Compare models based on their accuracy to select best performing one.

https://github.com/Poggerv2/DS-Capstone-Proyect-/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

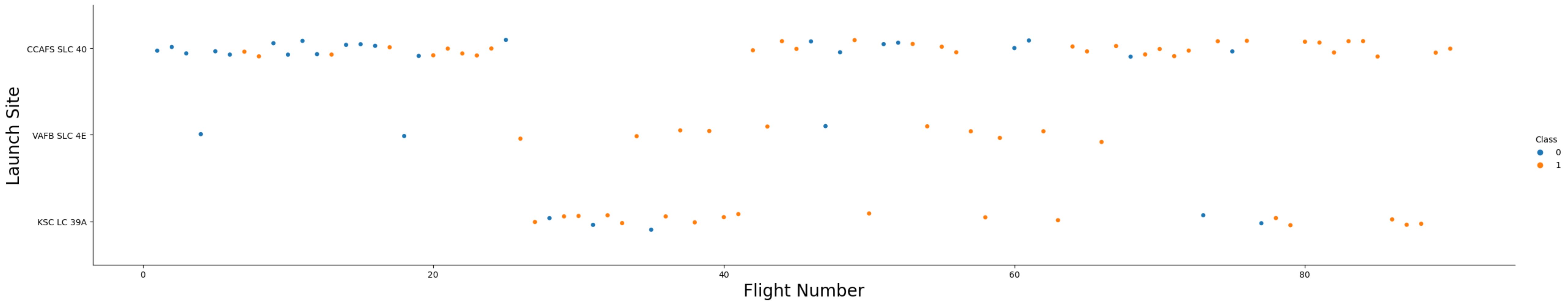
Results

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

Section 2

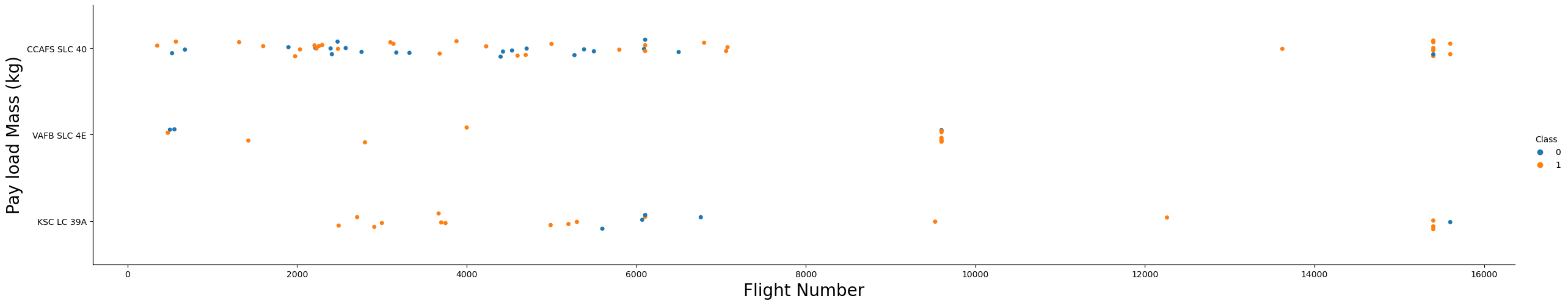
Insights drawn from EDA

Flight Number vs. Launch Site



- As the number of flights increase, the success rate for launch sites increase

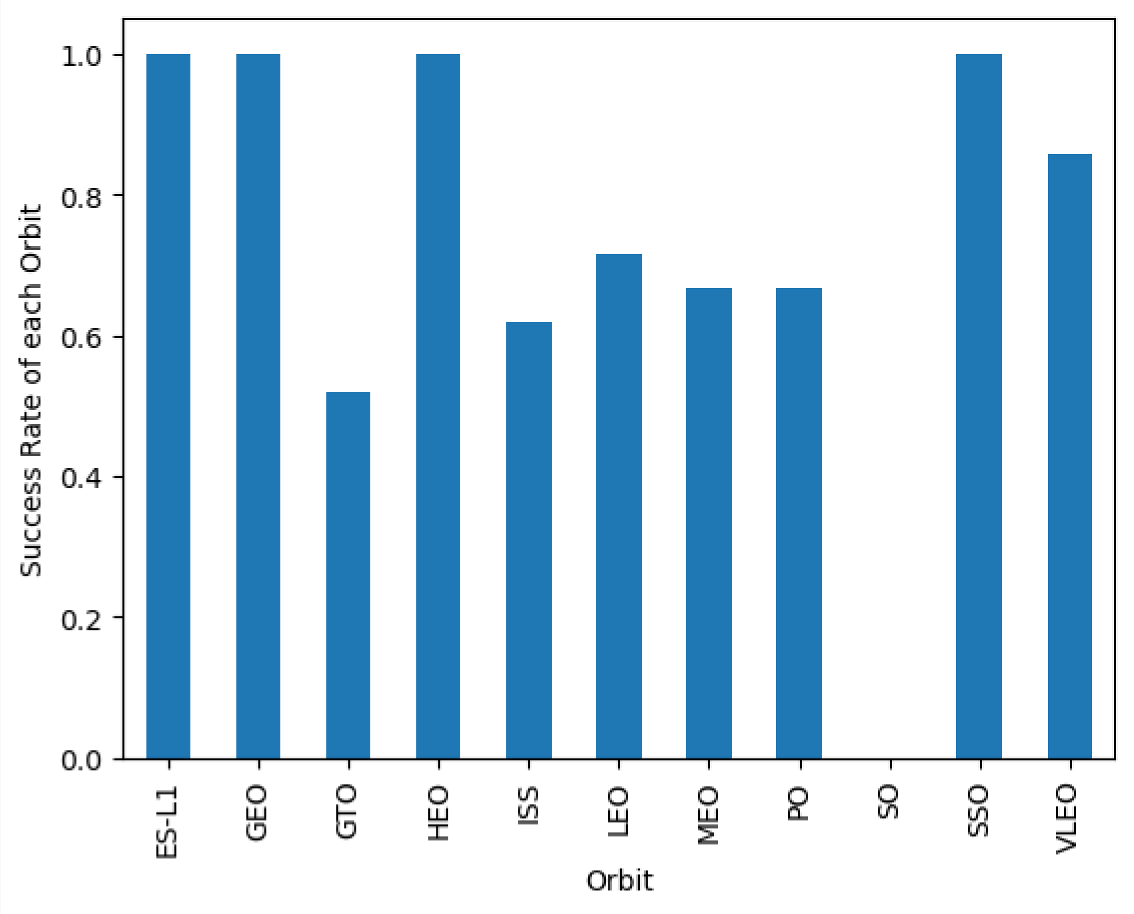
Payload vs. Launch Site



*The plot is labeled wrong, but the x and y values are good.

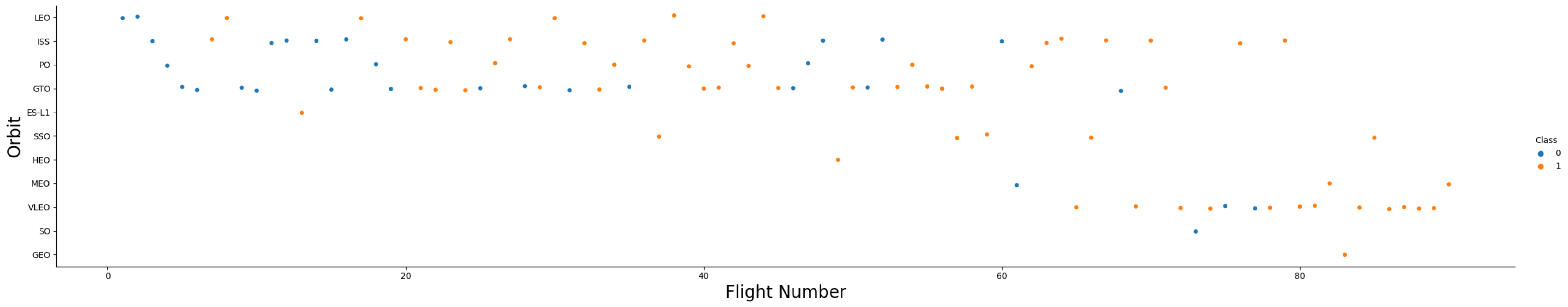
- Launch site success rate increases with payload mass. For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



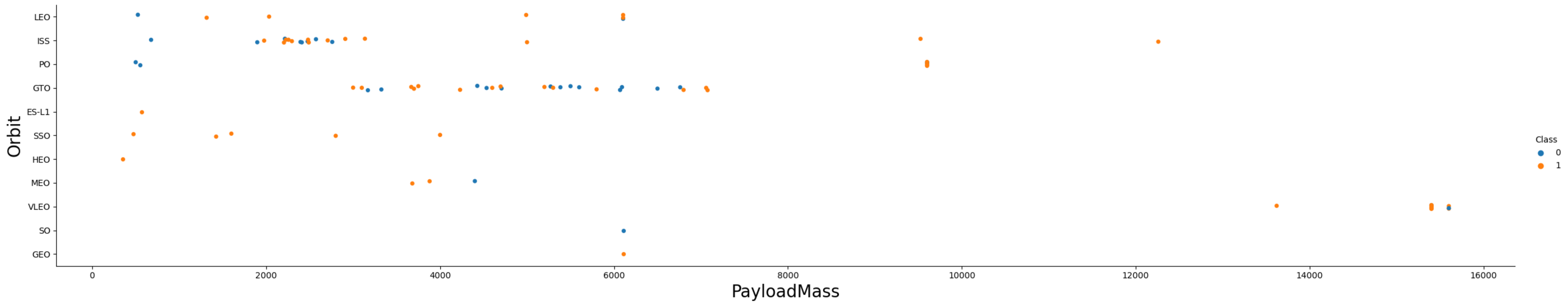
- The bar plot shows that ES-L1, GEO, HEO, and SSO has the highest success rate compared to other orbits. So has no success rate :(

Flight Number vs. Orbit Type



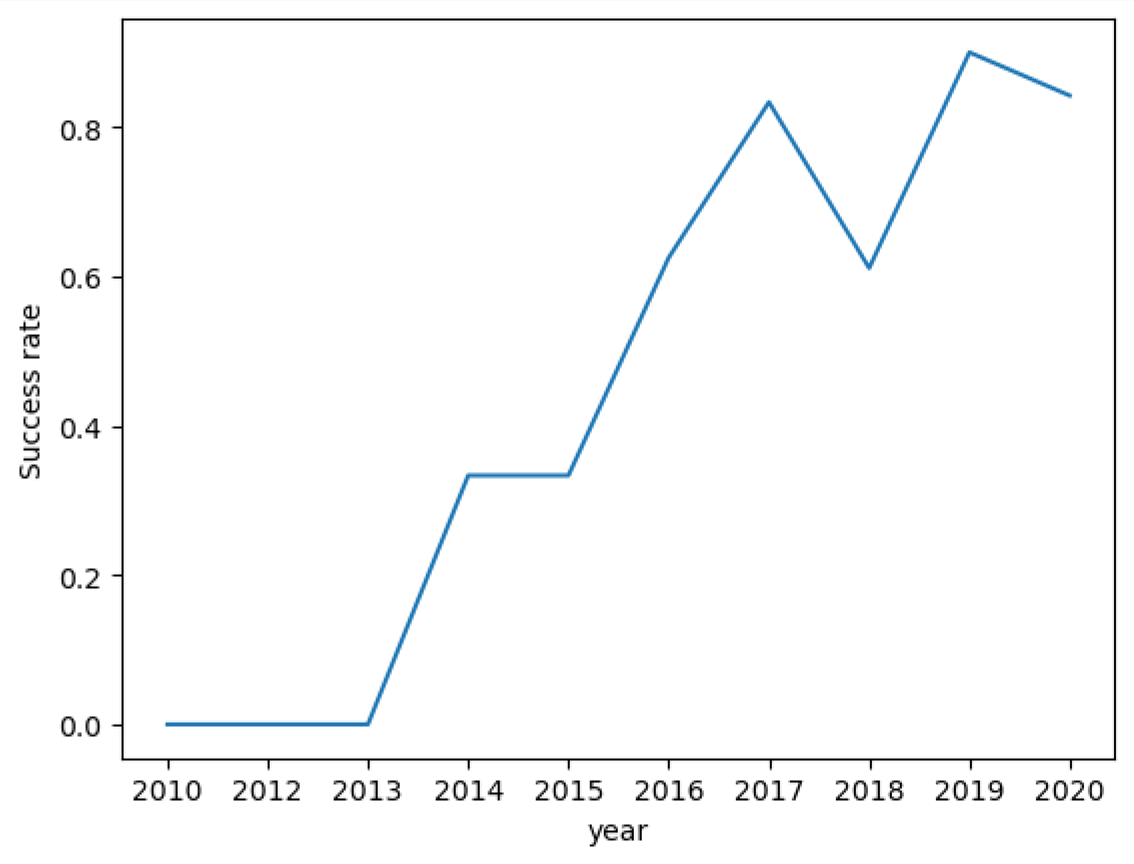
- In the low Earth orbit (LEO), there seems to be a correlation between success and the frequency of flights, whereas in the geostationary transfer orbit (GTO), there appears to be no such correlation between flight frequency and success.

Payload vs. Orbit Type



- When dealing with heavy payloads, the likelihood of successful landings is higher for missions to Polar, Low Earth Orbit (LEO), and the International Space Station (ISS). However, distinguishing between successful and unsuccessful landings is more challenging for missions to Geostationary Transfer Orbit (GTO), as both outcomes are present.

Launch Success Yearly Trend



- The line graph illustrates a consistent rise in success rates from 2013 to 2017, with a stable period in 2014. Following 2015, the trend continued upwards.

All Launch Site Names

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

- By employing DISTINCT in the query, duplicate site names are eliminated, resulting in the retrieval of the distinct four launch sites.

Launch Site Names Begin with 'CCA'

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

- The table is filtered to only show launches that occurred from the CCAFS LC-40 launch site.

Total Payload Mass

SUM(PAYLOAD_MASS__KG_)
45596

- We are selecting the sum of the "PAYLOAD MASS KG" column from the "SPACEXTBL" table, where the "CUSTOMER" column is equal to "NASA (CRS)". The result of the query is 45596.

Average Payload Mass by F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2534.666666666665

- We are selecting the average payload mass from a table named SPACEXTBL where the booster version is like '%F9 v1.1%'. The result of the query is 2534.67 kg.

First Successful Ground Landing Date

MIN(DATE)
2015-12-22

- MIN was used to select the first successful ground landing date, which is December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- The query shows four successful booster versions: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.

Total Number of Successful and Failure Mission Outcomes

SUCCESS	FAILURE
100	1

- The query retrieves the number of successful and failed missions from a table named SPACEXTBL. The result shows that there were 100 successful missions and 1 failed mission.

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

- The table shows the Booster Versions with the highest "PAYLOAD MASS KG" by comparing each mass to the overall maximum within the same table.

2015 Launch Records

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- The table shows the 2015 Launch Records. It also shows month of launch, booster version and launch site.

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

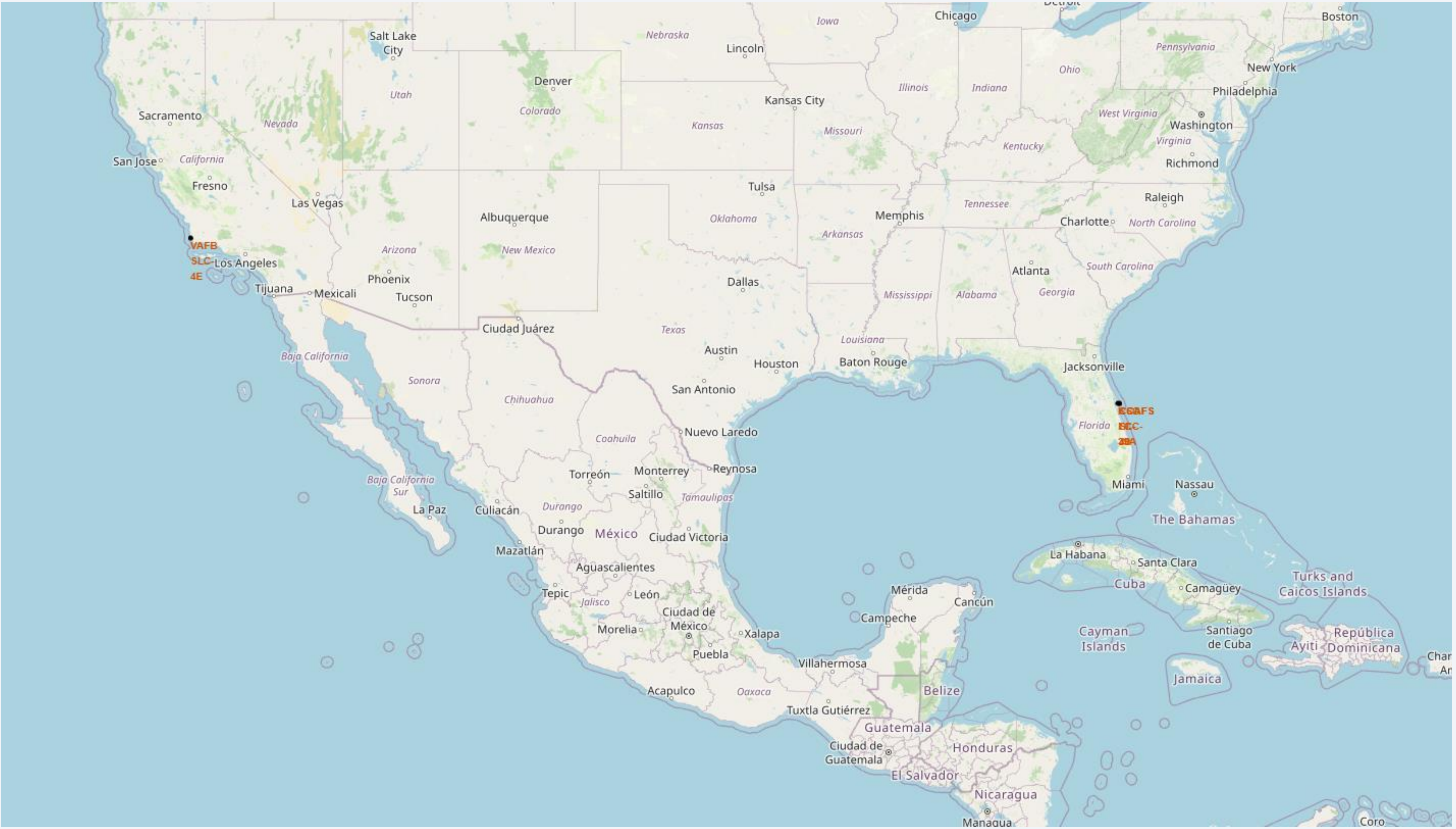
Date	Landing_Outcome	count_
2016-04-08	Success (drone ship)	5
2015-12-22	Success (ground pad)	3
2015-06-28	Precluded (drone ship)	1
2015-01-10	Failure (drone ship)	5
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2012-05-22	No attempt	10
2010-06-04	Failure (parachute)	2

- The table shows the Date of landing outcomes (2010-2017), landing outcome, and counts of landing outcomes showed as count_.
- The most common landing outcome is "No attempt", with 10 outcomes. This means that there were 10 times when a rocket did not attempt to land during the specified time period.

Section 3

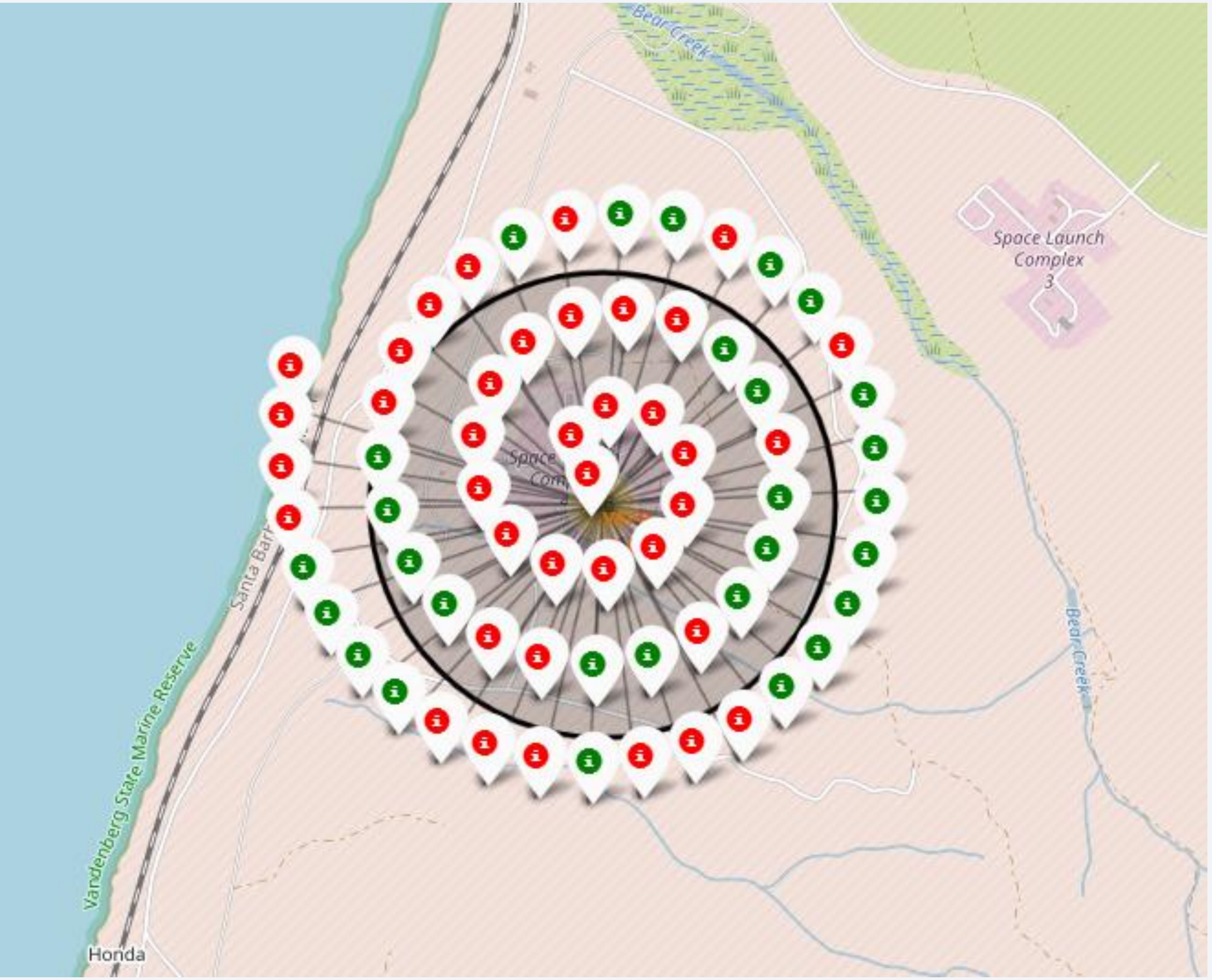
Launch Sites Proximities Analysis

SpaceX Launch Sites



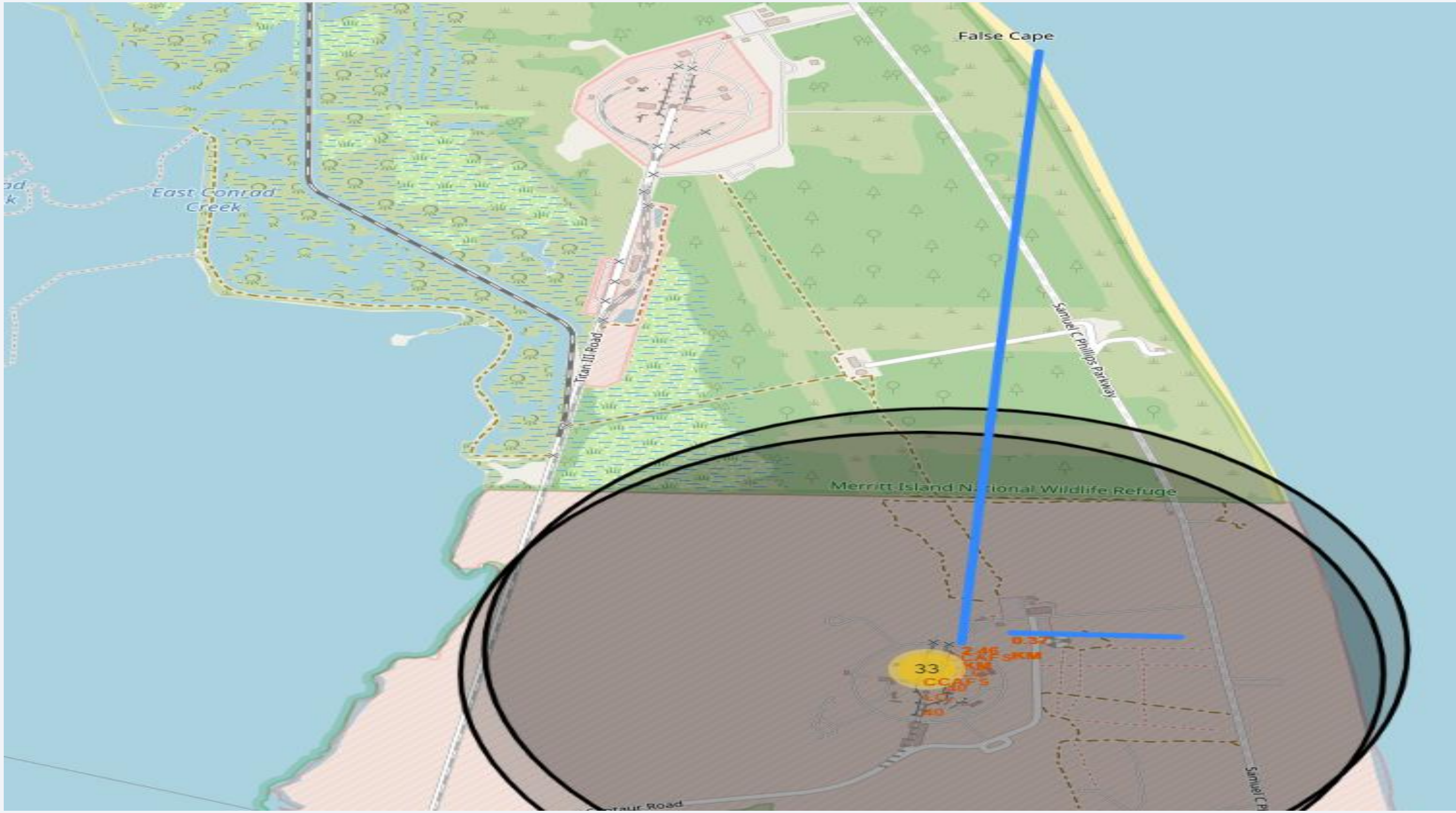
- The map shows that SpaceX launch sites are located in Florida and California, USA.

Success/Failed Launches for VAFB SLC-4E Site



- The green ones indicates what launches where successful

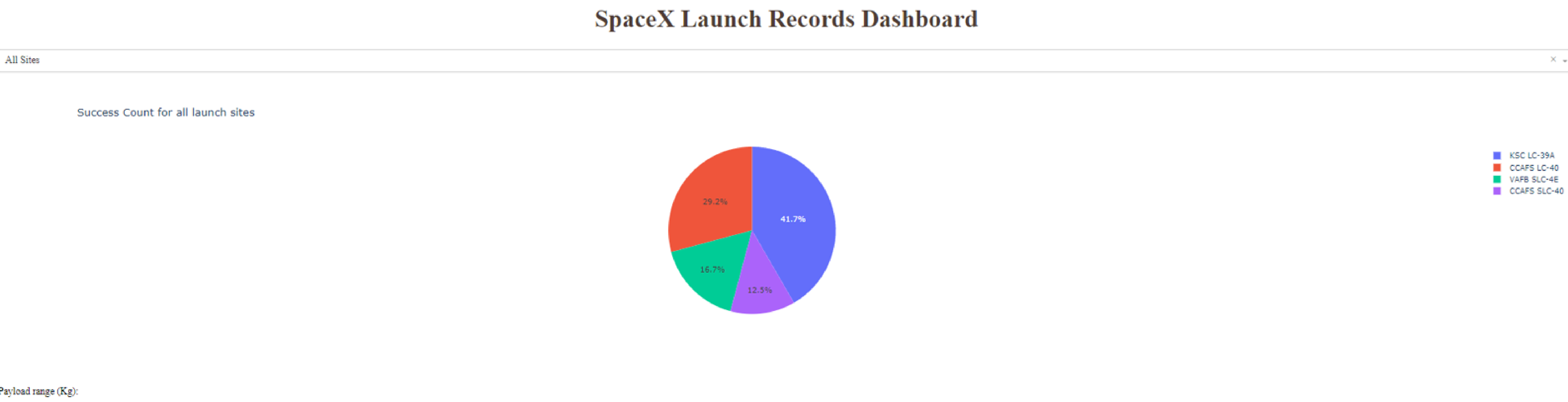
Distances between CCAFS SLC-40 and its Proximities



Section 4

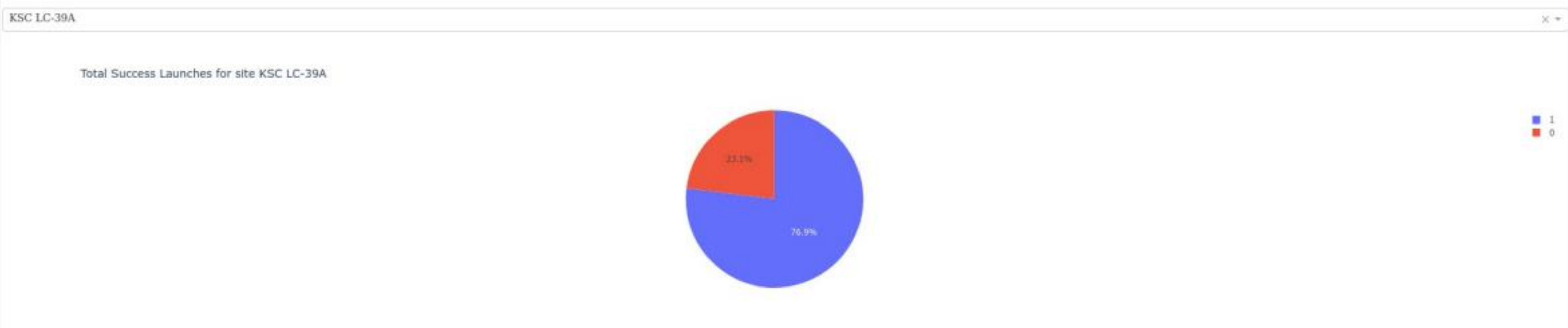
Build a Dashboard with Plotly Dash

Total Success Launches by Site



- We can see from the pie chart that KSC LC-39A has be best success launches compared to the other three sites.

<Dashboard Screenshot 1>



- KSC LC-39A achieved 76.9% success launches.

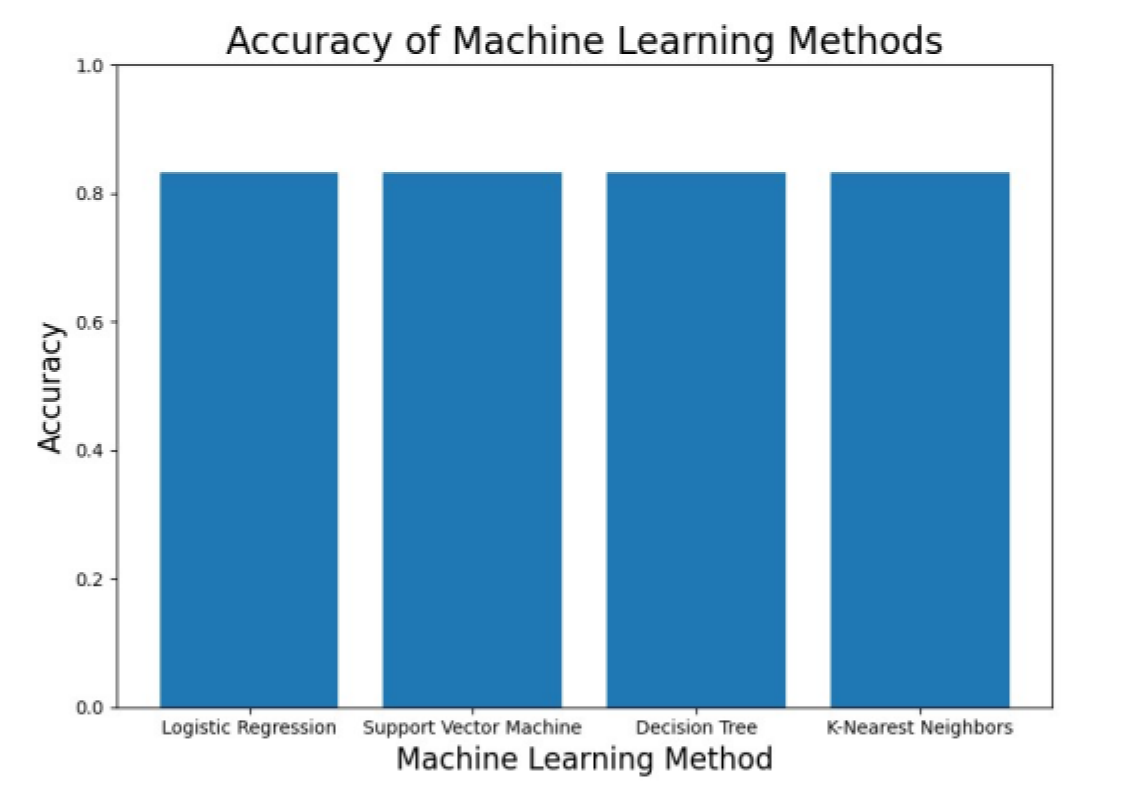
<Dashboard Screenshot 3>

Section 5

Predictive Analysis (Classification)

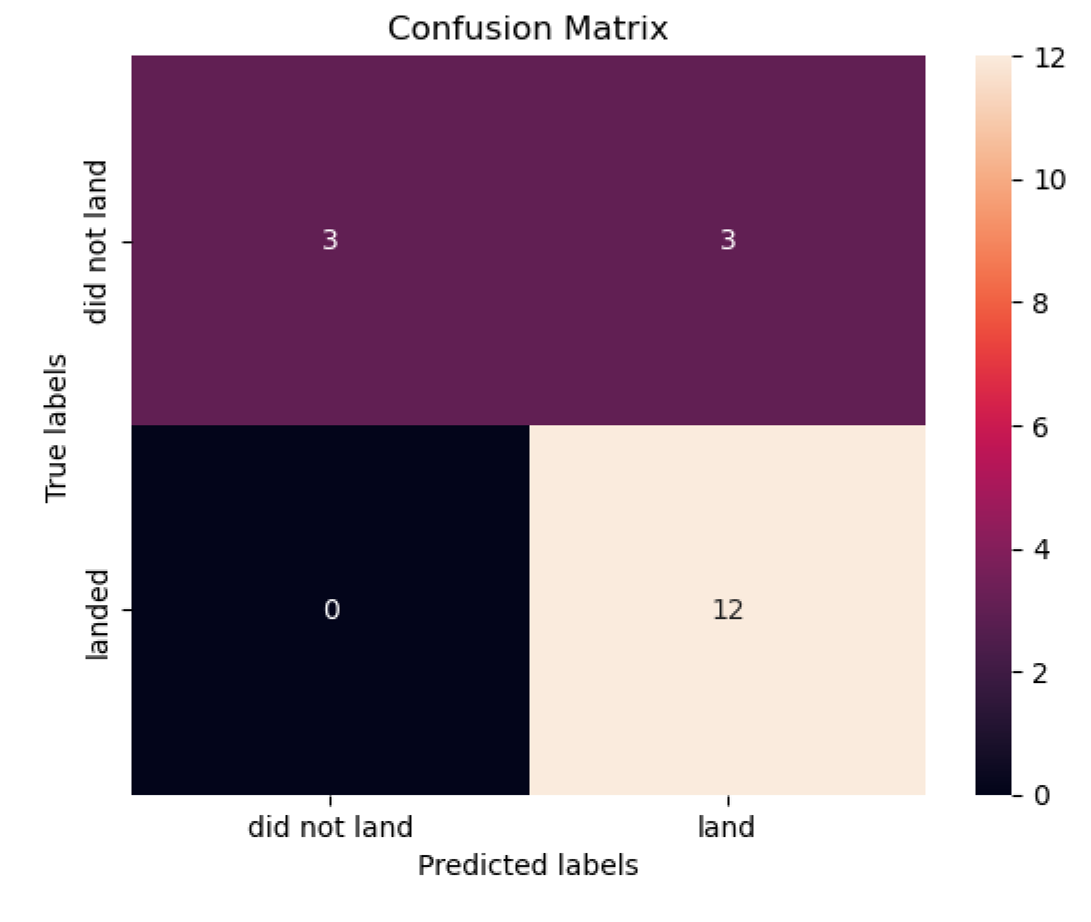
Classification Accuracy

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision Tree method: 0.8333333333333334
Accuracy for K nearest neighbor method: 0.8333333333333334
```



- We can see from the bar chart that all four models achieved the same accuracy. This could be due to the small data.

Confusion Matrix



- The confusion matrices for all four models (logistic regression, SVM, decision tree, and KNN) are the same. One problem that all models have in common is false positives.

Conclusions

- The success of SpaceX's Falcon 9 first stage landing is affected by a number of factors, including payload mass, orbit type, and launch site.
- Orbits with highest success rates are ES-L1, GEO, HEO, and SSO.
- KSC LC-39A launch site has be best success launches compared to the other three sites.
- All four classification models trained (logistic regression, SVM, decision tree, and KNN) achieved the same accuracy score (0.833).

Appendix

- Python apps used:
- NumPy
- Pandas
- Scikit-Learn
- Matplotlib
- Seaborn
- Plotly
- Dash
- Folium
- BeautifulSoup

Thank you!