IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Soumyajit Chowdhury
21st April, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection, from public SpaceX API and SpaceX Wikipedia page.
  - Data wrangling, including adding an indicator for successful landings.
  - EDA, via SQL queries and various visualizations and data summaries.
  - Explored and analyzed further, using Folium to generate interactive maps.
  - Interactive dashboard developed with Plotly Dash.
  - Machine learning models trained to predict successful landings through classification techniques.
- Summary of all results
  - The resulting models all produced similar results, with an accuracy rate of -94.44% when tested with a test data set. The models tended to over predict successful landings. Training the model with more data could lead to improved accuracy.

Introduction

- Project background and context

SpaceX offers Falcon 9 rocket launches for 62 million dollars, significantly less than other providers who charge upward of 165 million dollars. One reason for the low cost is the ability to reuse the first stage of the rocket, which provides the initial thrust to get the spacecraft off the ground and into space. If the first stage lands successfully, SpaceX can reuse it on future launches, significantly reducing the cost of the launch. However, if the first stage doesn't land successfully, a new first stage needs to be built, which increases the cost. This information can help an alternate company estimate the cost of a competitive bid.

- Problems you want to find answers
  - What are the factors that affect the probability of a rocket successfully landing?
  - What are the optimal conditions required for a successful rocket landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX REST API & WEB SCRAPING SPACEX Wikipedia Page
- Perform data wrangling
  - Removing Irrelevant fields & Imputing NULL Values
  - Using One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

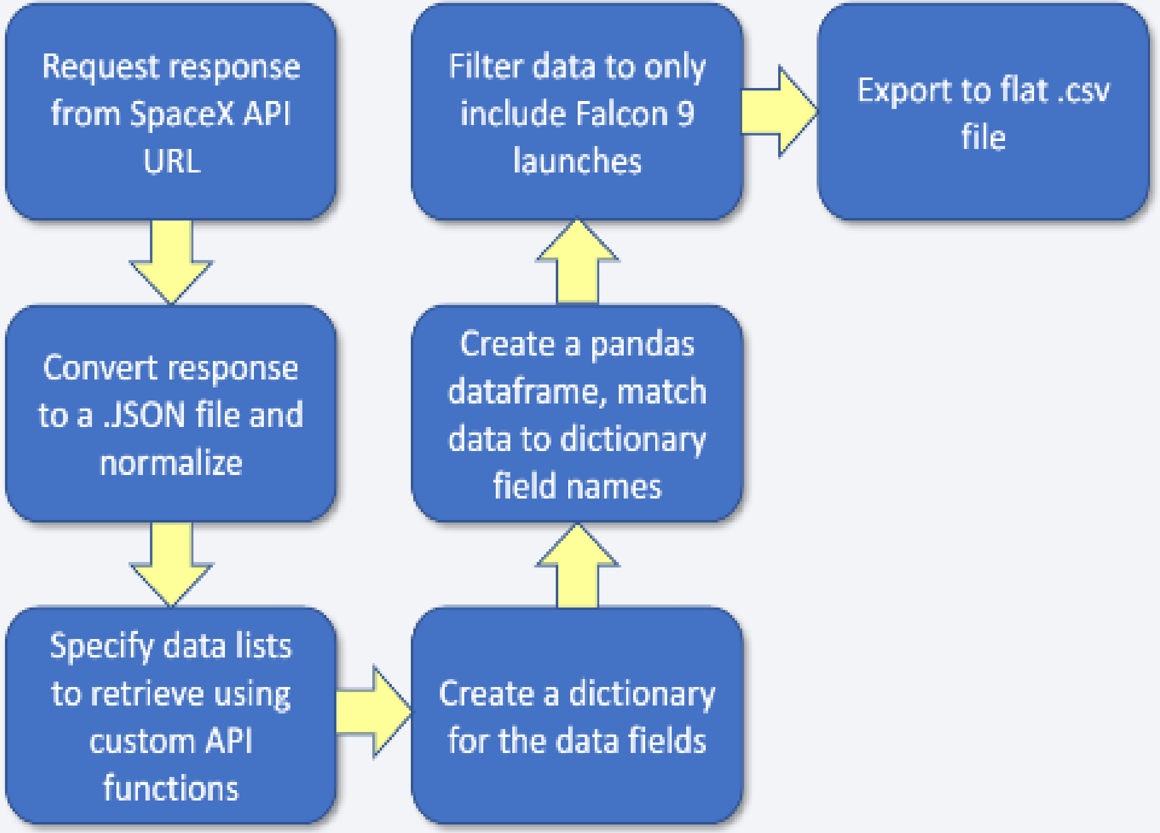
Data Collection

- SPACEX REST API
  - The data includes information about payload, landing outcome & other launches & landing specifications in the form of .JSON file. Data is normalized into a flat .CSV file.
- WEB SCRAPING
  - Scraping the Wikipedia page the data is collected using Beautiful Soup. At last normalized into a flat .CSV file.

Data Collection – SpaceX API

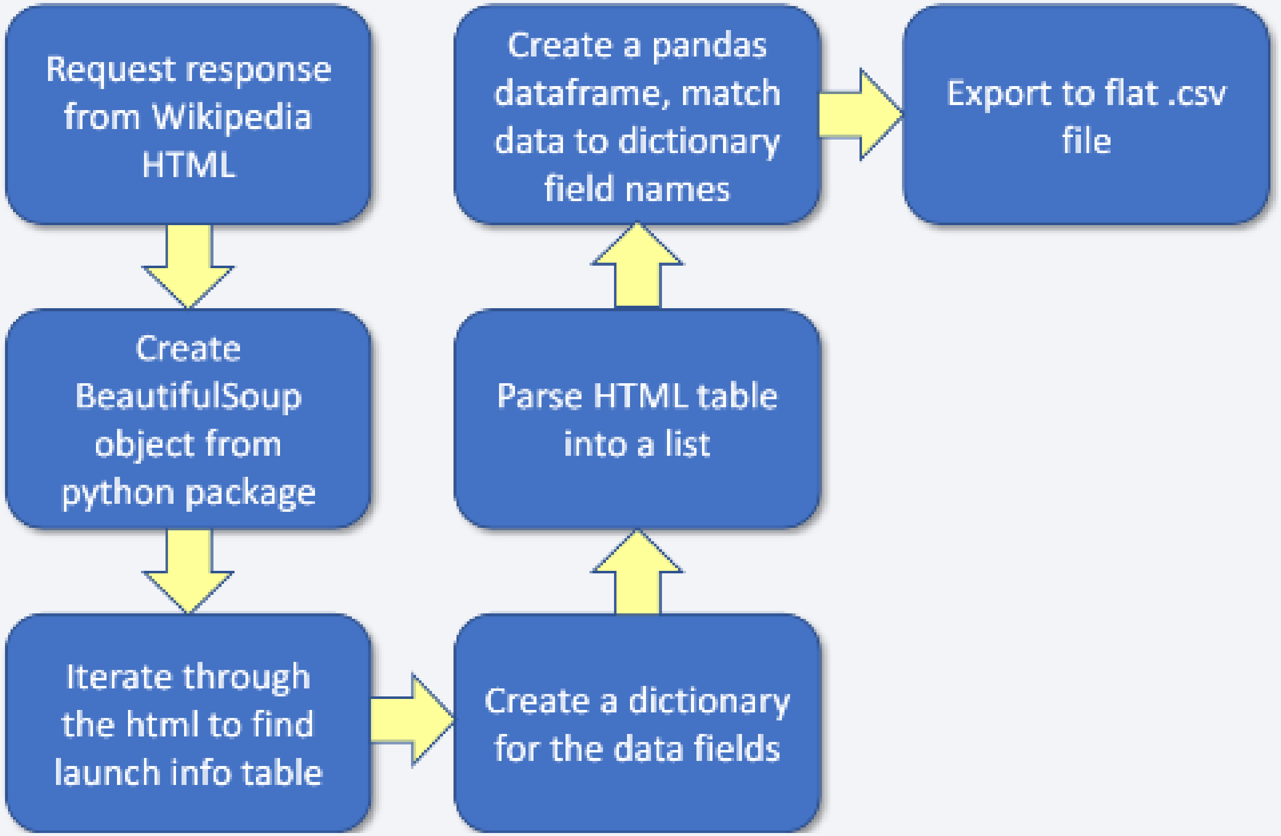
Flowchart

[IBM-Data-Science/Data Collection API .ipynb at main · soumojeet/IBM-Data-Science (github.com)]



Data Collection - Scraping

Flowchart



[IBM-Data-Science/Data Collection with Web Scraping.ipynb at main · soumojeet/IBM-Data-Science (github.com)]

Data Wrangling

- Two Components: 'Mission Outcome' & 'Landing Outcome'
- Successful Landing = 1 & Unsuccessful Landing = 0
- Value Mapping:
  - Outcomes 'True ASDS', 'True RTLS' & 'True Ocean' - set Class 1
  - Outcomes 'None None', 'False ASDS', 'None ASDS', 'False Ocean', 'False RTLS' - set Class 0

IBM-Data-Science/Data Wrangling.ipynb at main · soumojeet/IBM-Data-Science (github.com)

EDA with Data Visualization

- Relationship between variables 'Flight Number', 'Payload Mass', 'Launch Site', 'Orbit', 'Class' & 'Year'.

- Scatter Plots:
  - Flight number Vs Payload Mass
  - Flight Number Vs Launch Site
  - Payload Vs Launch Site
  - Orbit Vs Flight Number
  - Payload Vs Orbit Type
  - Orbit Vs Payload Mass
- Bar Charts
  - Mean Vs Orbit
- Line Charts
  - Success Rate Vs Year

[IBM-Data-Science/EDA with Data Visualization.ipynb at main · soumojeet/IBM-Data-Science (github.com)]

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was acheive.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- 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.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

[IBM-Data-Science/EDA with SQL.ipynb at main · soumojeet/IBM-Data-Science (github.com)]()

Build an Interactive Map with Folium

- Visualize the launch data onto an interactive map. Using latitude & longitude coordinates of each launch site, we add markers at the launch site. We can calculate distance to key locations on the map and mark a line on the map to visualize. E.g. Distance to nearest Railway, Highway, Coast & City.

[IBM-Data-Science/Interactive Visual Analytics with Folium.ipynb at main · soumojeet/IBM-Data-Science (github.com)]

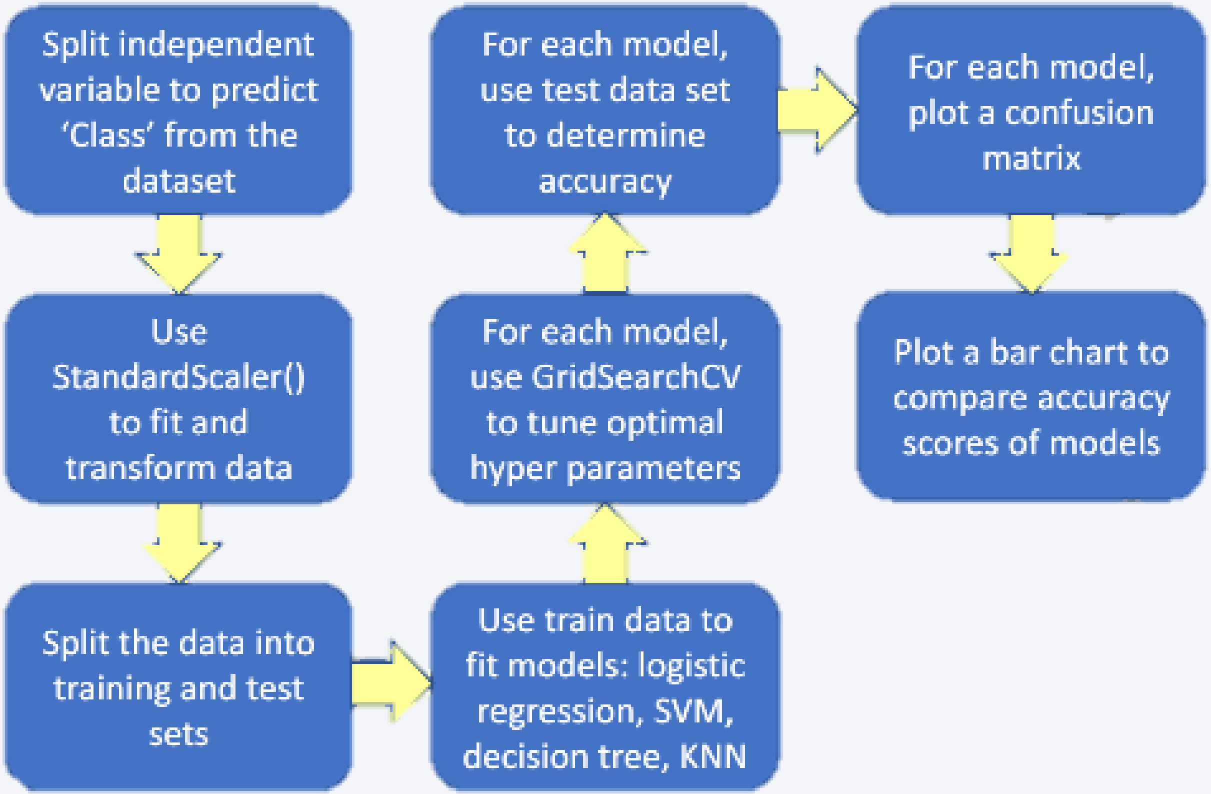
Build a Dashboard with Plotly Dash

- Dashboard including Pie Chart & Scatter Plot
- Interactive Charts to Visualize Site Success Rate, Distribution of Successful Landings Across All Launch Sites Or Distribution of Successful Landings for Specific Individual Launch Sites
- Interactive Plots to Visualize Success Varies Dependent on Payload Mass & Booster Version Category

[IBM-Data-Science/Build an Interactive Dashboard with Ploty Dash.py at main · soumojeet/IBM-Data-Science (github.com)]

Predictive Analysis (Classification)

Flowchart



[IBM-Data-Science/Machine Learning Prediction lab.ipynb at main · soumojeet/IBM-Data-Science (github.com)]

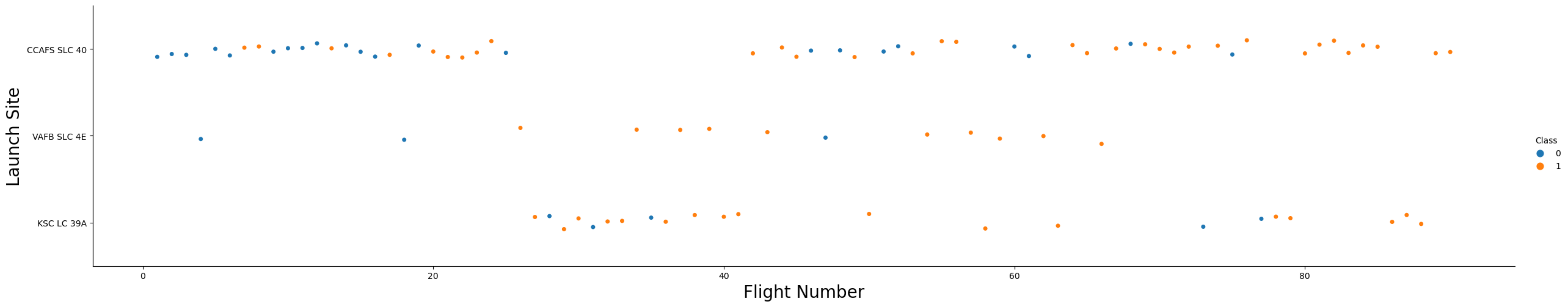
Results

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

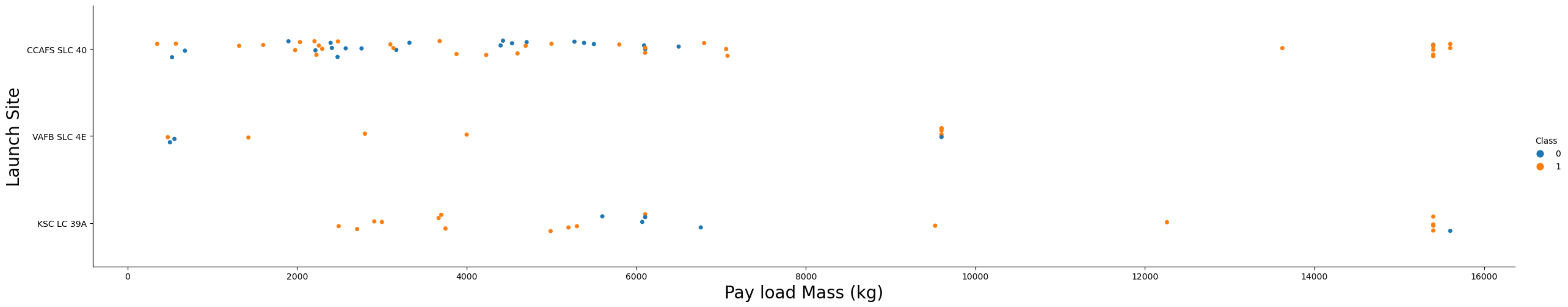
Section 2

Insights drawn from EDA

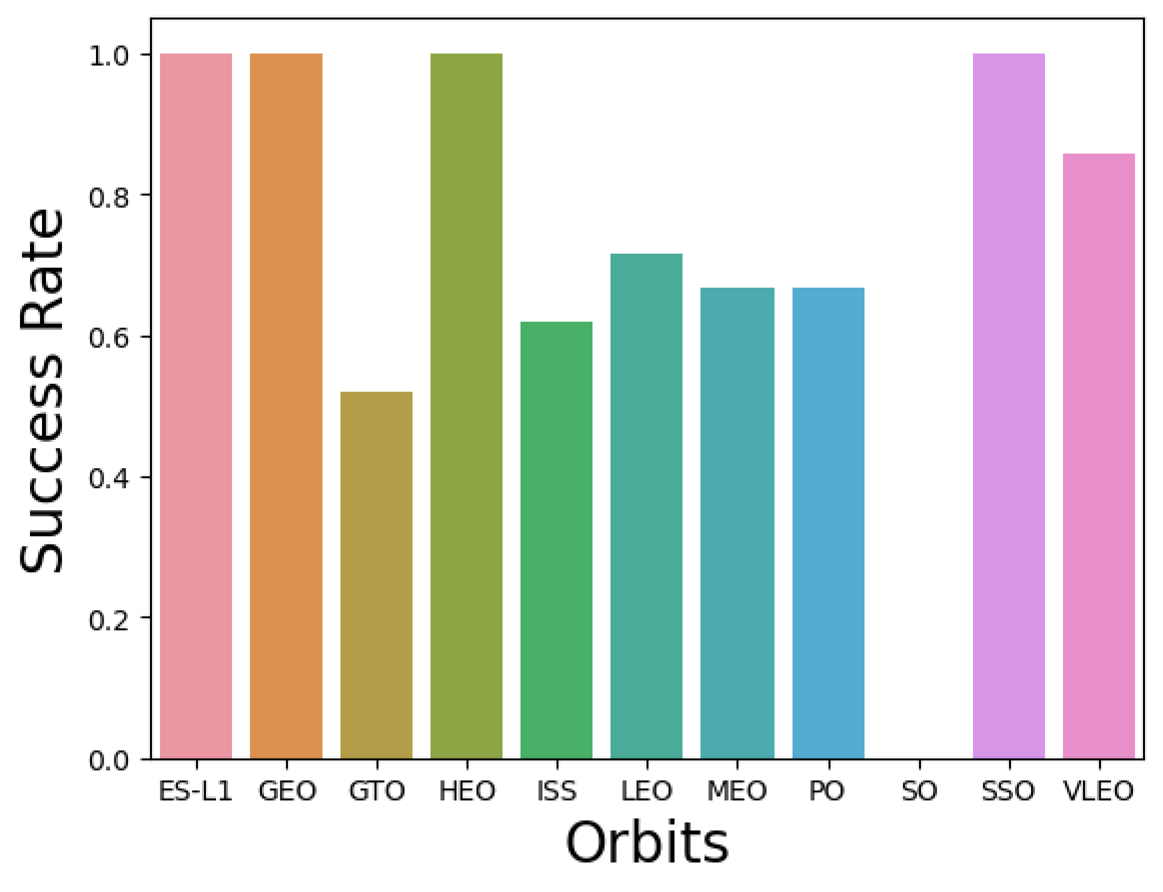
Flight Number vs. Launch Site



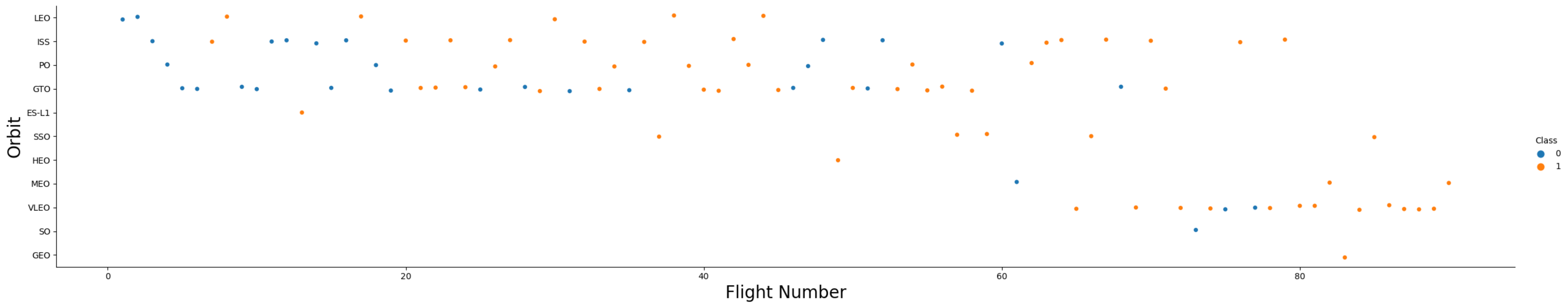
Payload vs. Launch Site



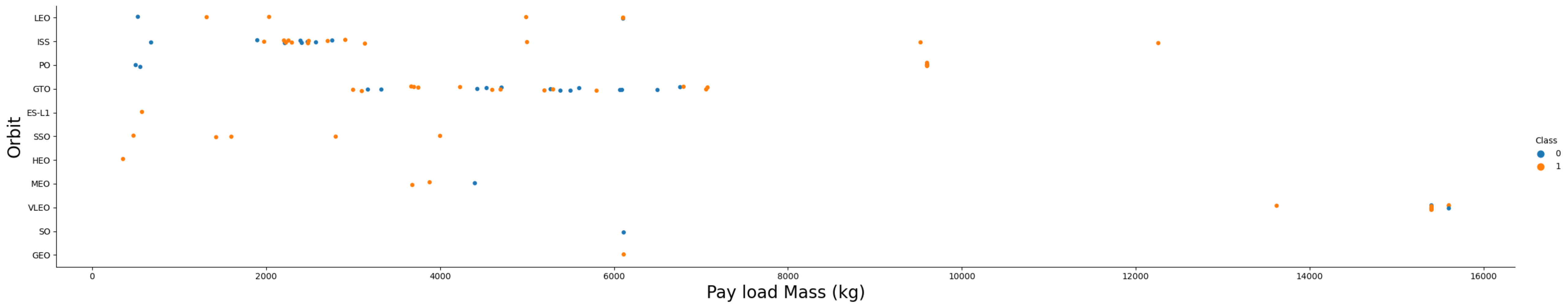
Success Rate vs. Orbit Type



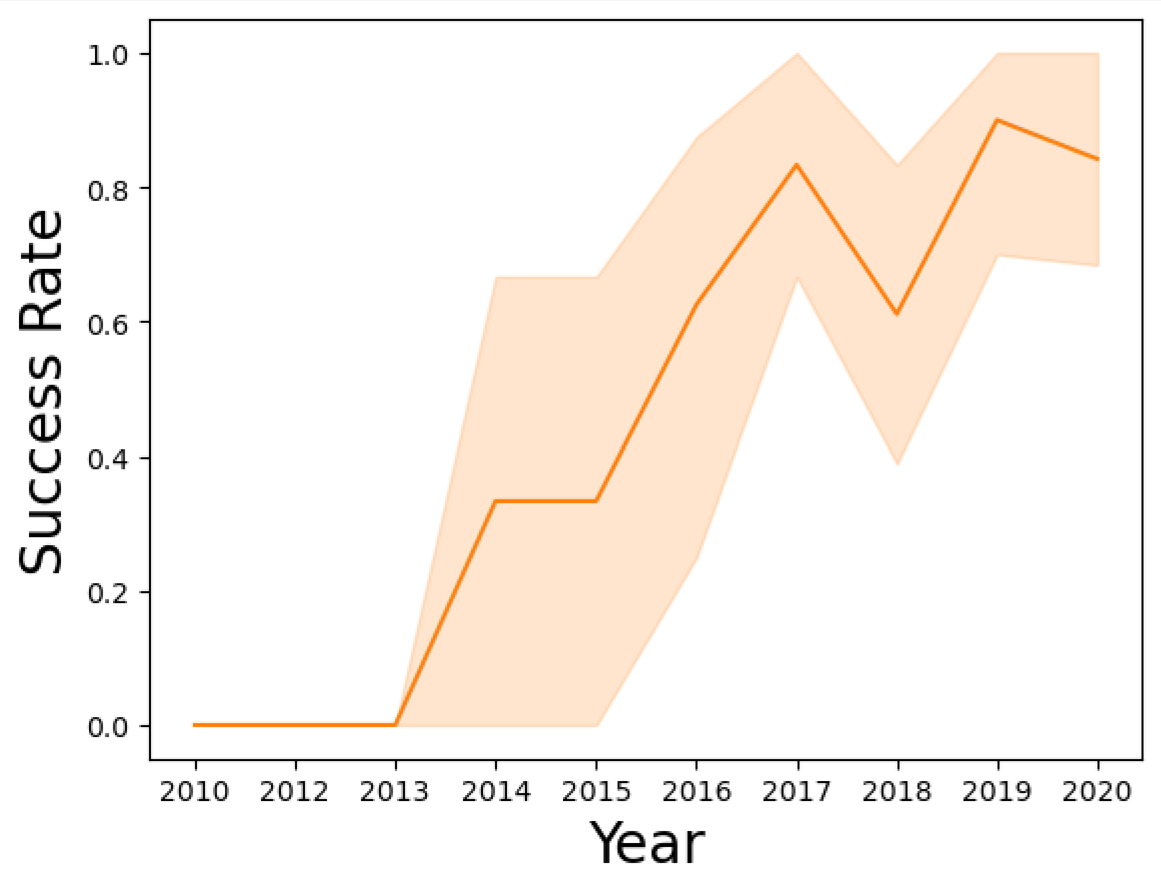
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

```
%%sql
SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE Customer LIKE 'NASA (CRS)'
```

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

TOTAL_PAYLOAD
45596

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1'
```

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

AVERAGE_PAYLOAD
2928.4

First Successful Ground Landing Date

```
%%sql
SELECT MIN(Date) as FIRST_SUCCESS_LAUNCH FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success (ground pad)'
```

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

FIRST_SUCCESS_LAUNCH
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome LIKE 'Success (drone ship)'
```

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

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

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT Mission_Outcome, COUNT(*) AS NO_OF_OUTCOME FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

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

```
%%sql
SELECT substr(Date, 4, 2) AS MONTH, Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTBL WHERE substr(Date,7,4) LIKE '2015' AND Landing_Outcome LIKE 'Failure (drone ship)'
```

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

MONTH	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
%%sql
SELECT Landing_outcome, COUNT(*) AS TOTAL FROM SPACEXTBL
WHERE Date BETWEEN '04-06-2010' AND '20-03-2017' AND Landing_Outcome LIKE '%Success%'
GROUP BY Landing_Outcome
ORDER BY COUNT(*) DESC
```

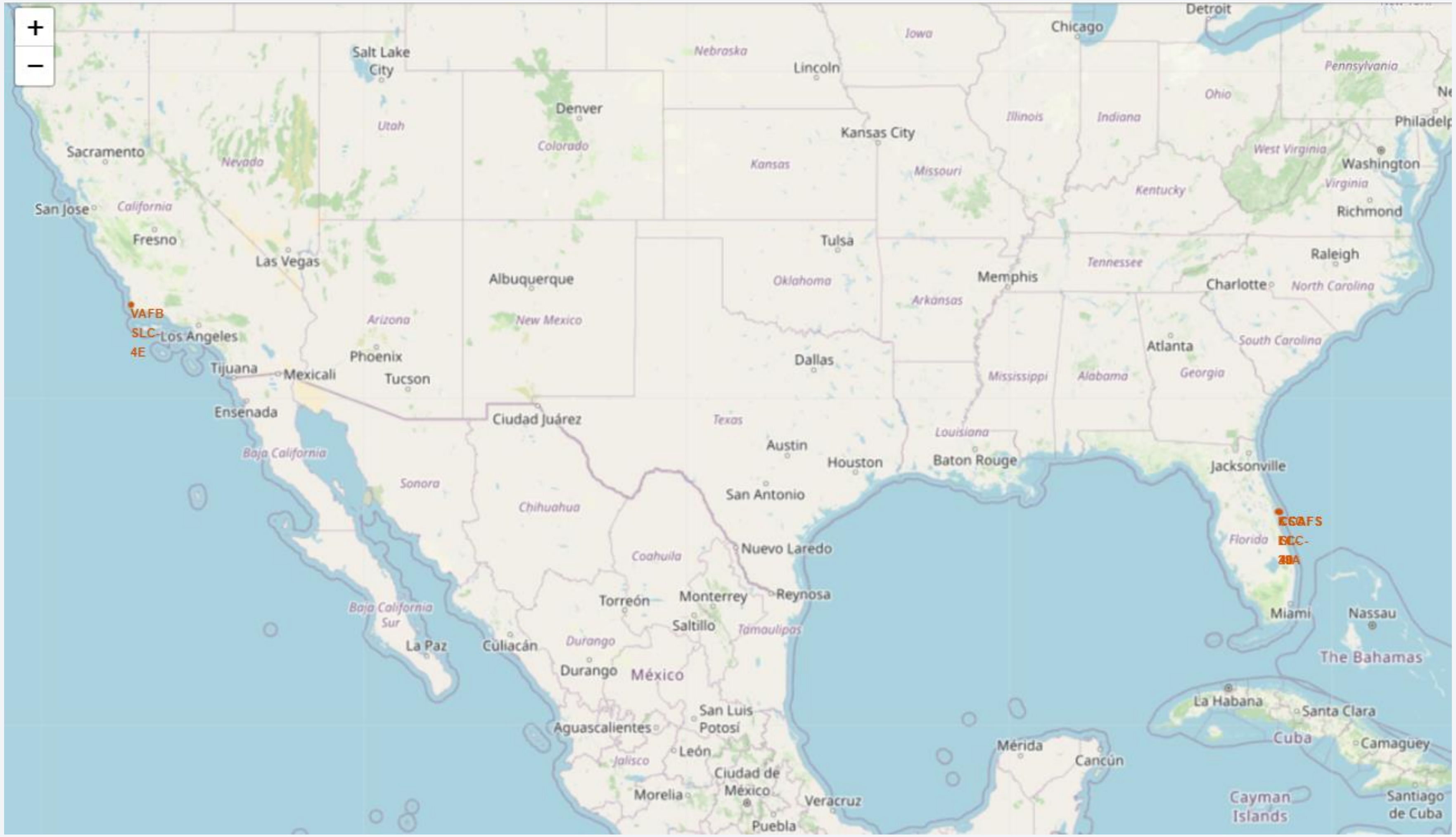
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	TOTAL
Success	20
Success (drone ship)	8
Success (ground pad)	6

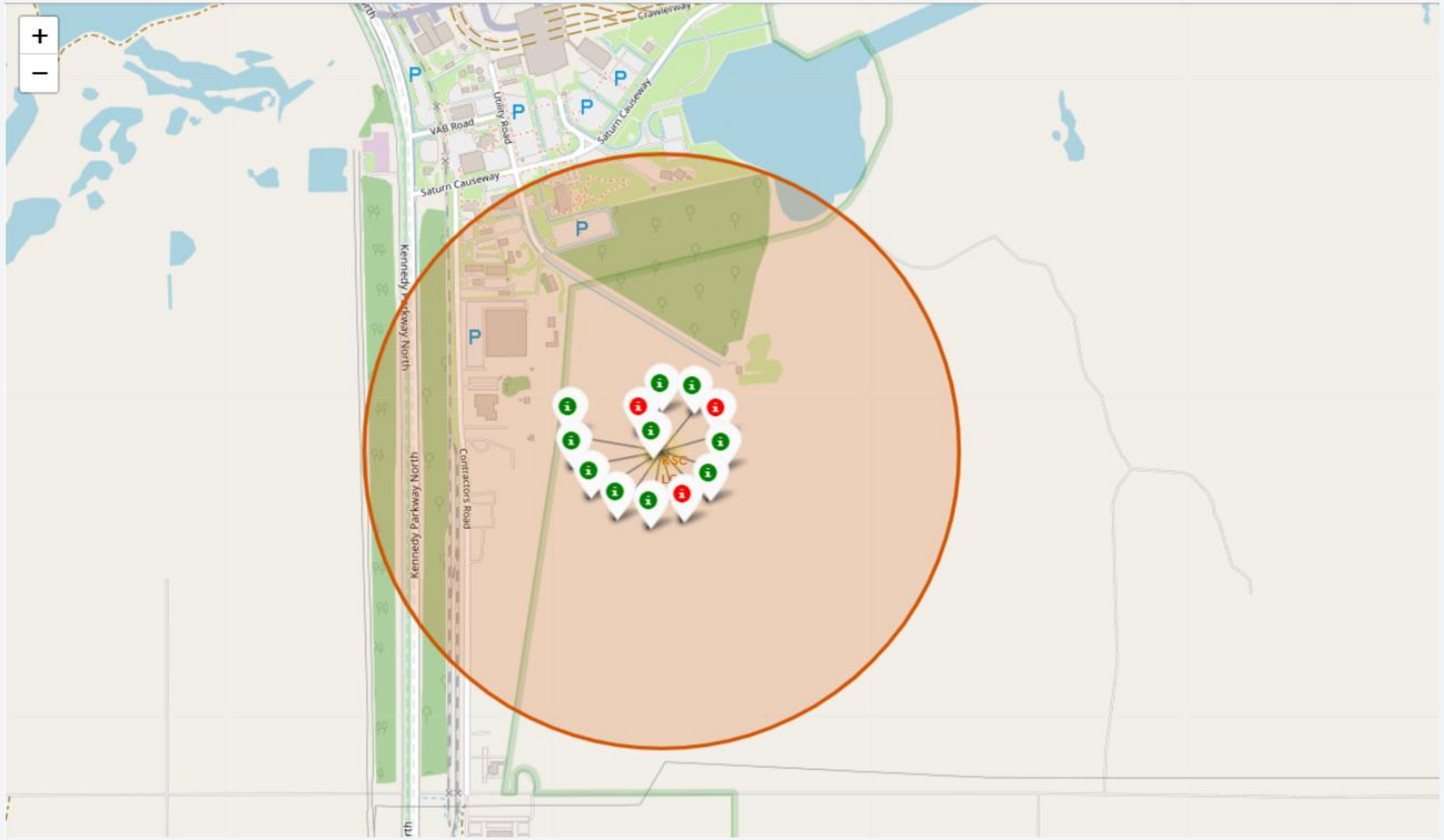
Section 3

Launch Sites Proximities Analysis

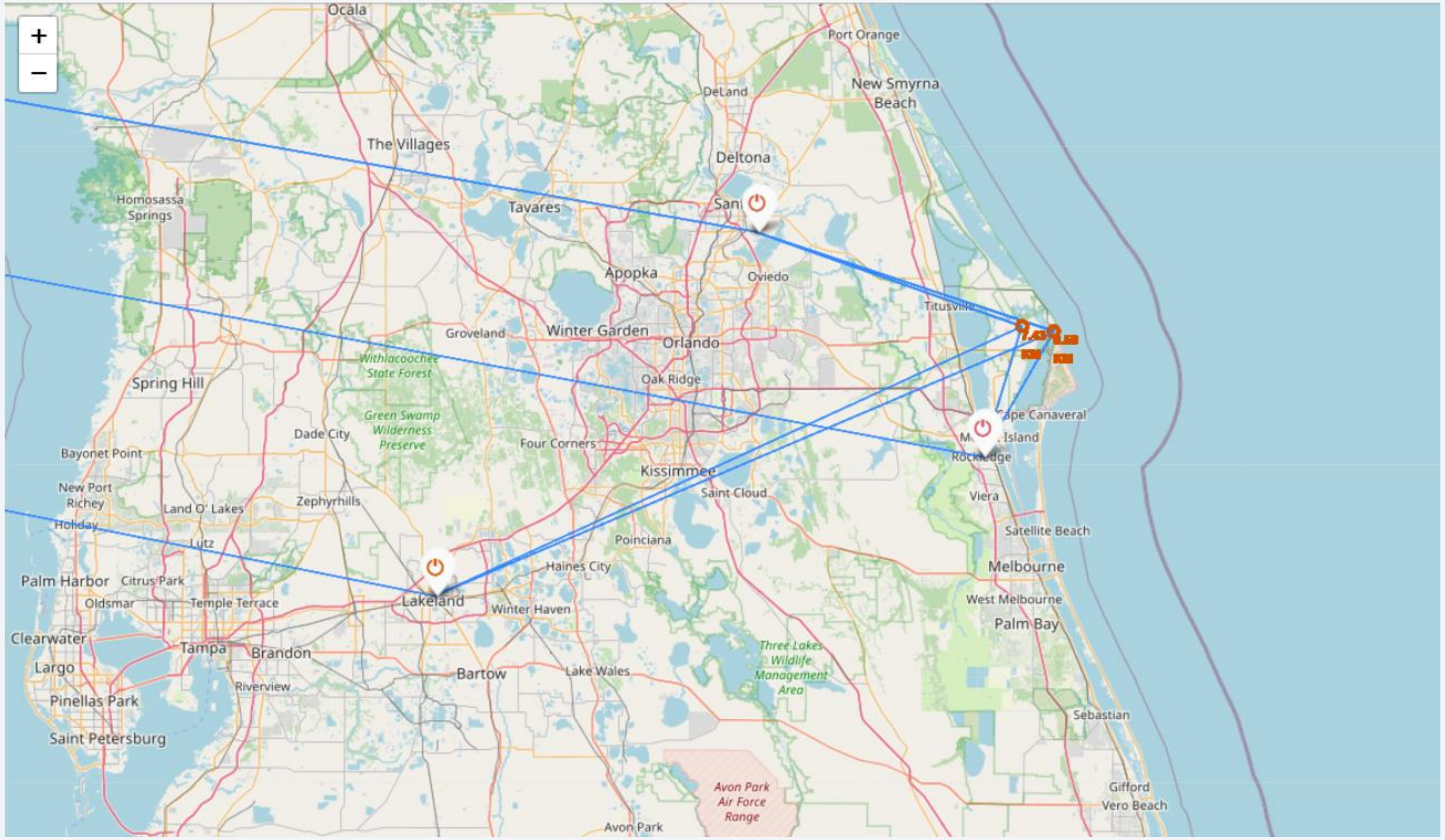
<Folium Map Screenshot 1>



<Folium Map Screenshot 2>



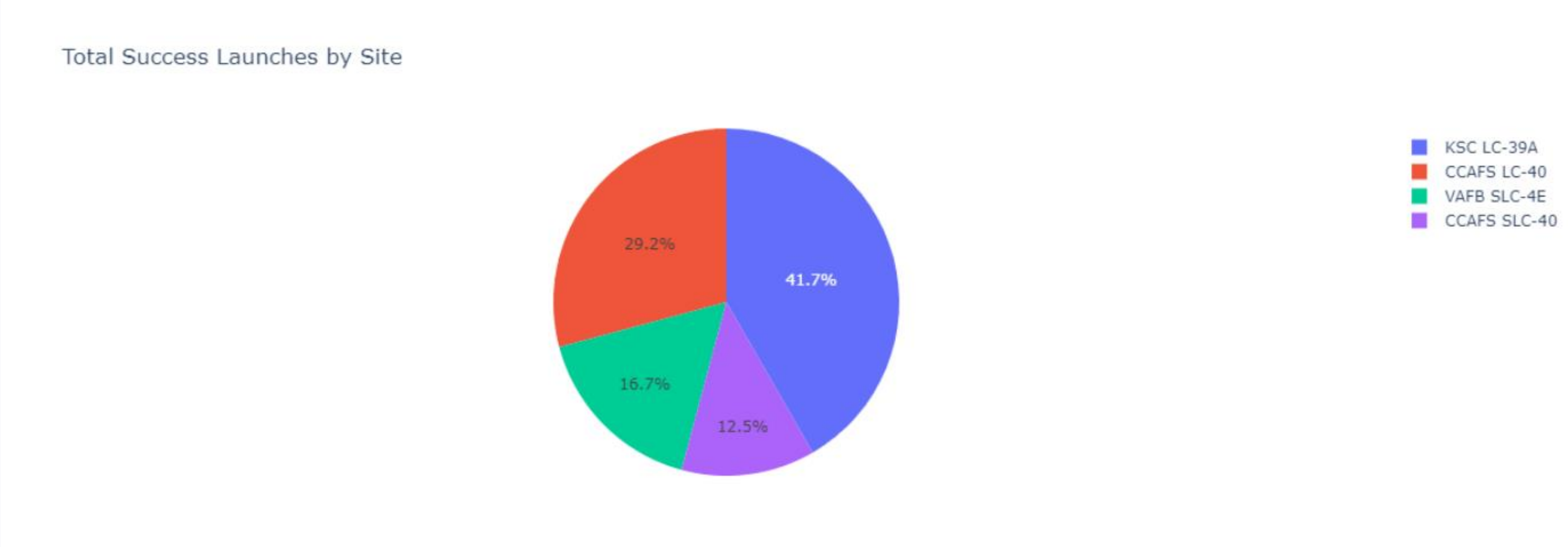
<Folium Map Screenshot 3>



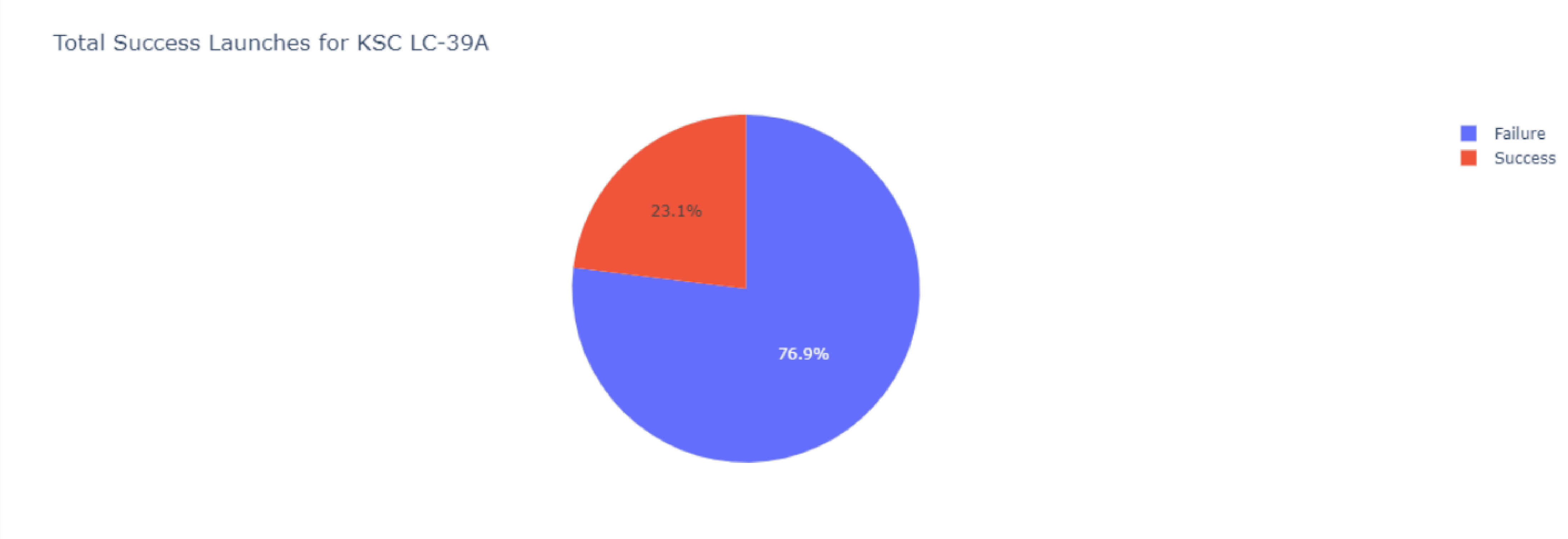
Section 4

Build a Dashboard with Plotly Dash

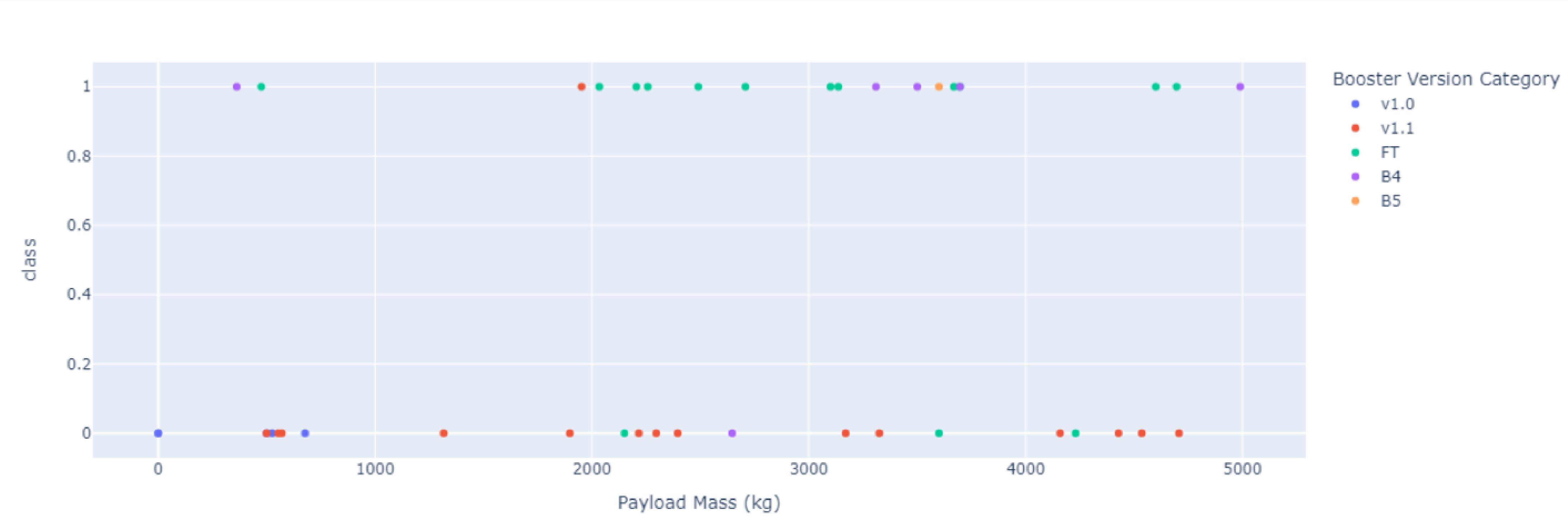
<Dashboard Screenshot 1>



<Dashboard Screenshot 2>



<Dashboard Screenshot 3>

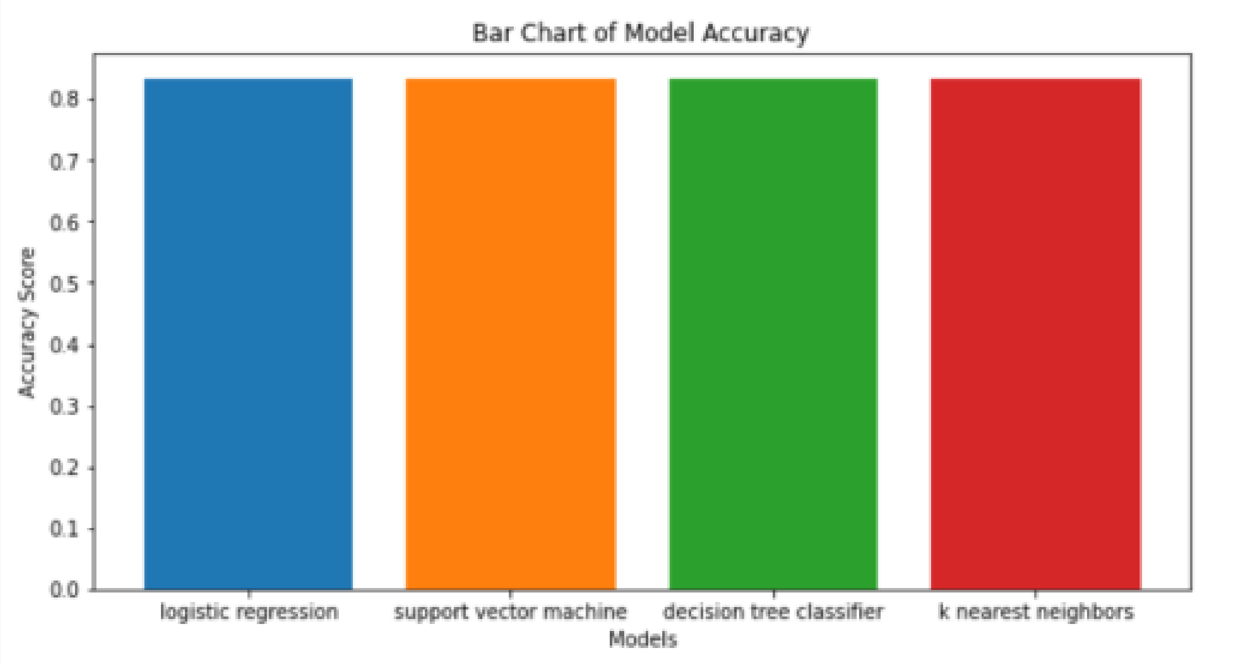


Section 5

Predictive Analysis (Classification)

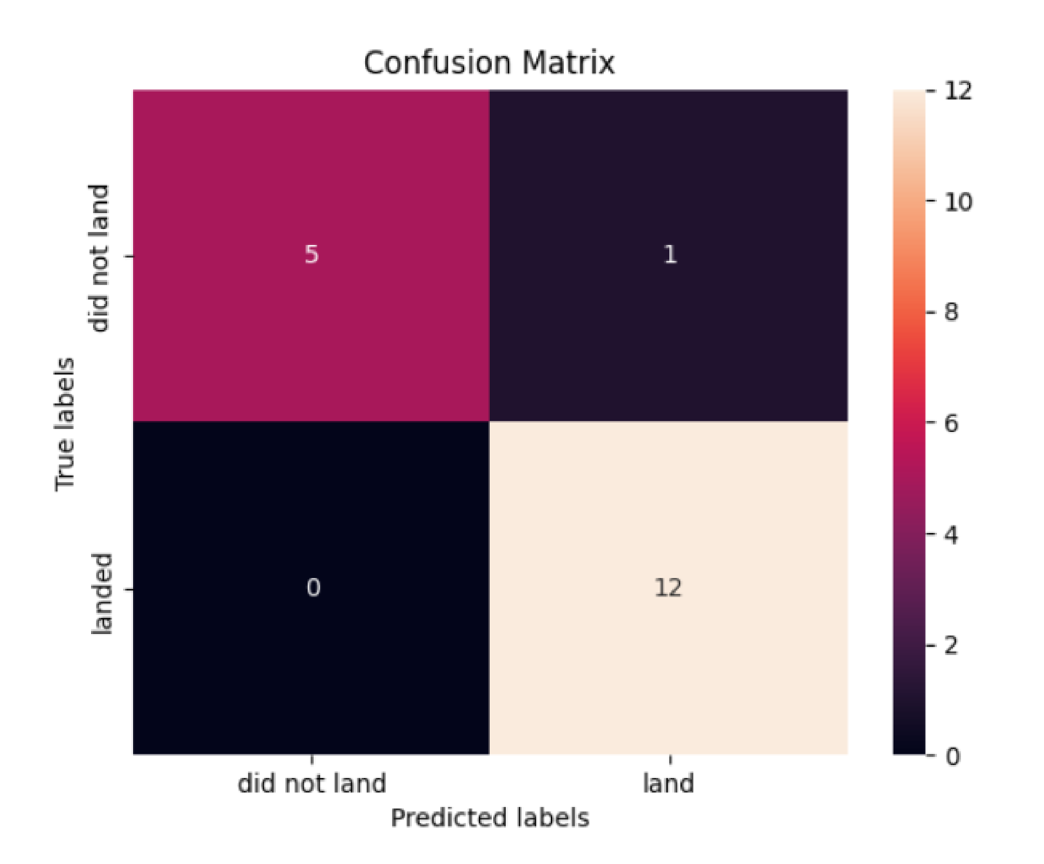
Classification Accuracy

- All model produce the same accuracy against the test data set ~94.44



Confusion Matrix

- All model generated the same confusion matrix



Conclusions

- Goal to develop a machine learning model to predict if stage 1 will successfully land for a given launch
- All predicted successful landings with ~94.44 accuracy for some test data. The models tend to predict successful landings, the models could be improved by using more data.

Appendix

[Github Link]

Thank you!