IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Angela Suriyakumaran
18/12/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using SpaceX API
  - SpaceX data collection using web scraping methods
  - SpaceX data wrangling
  - SpaceX exploratory data analysis using SQL
  - SpaceX EDA data visualisation using Pandas and Matplotlib
  - SpaceX launch site locations using Folium and Plotly Dash
  - SpaceX machine learning prediction
- Summary of all results
  - Results from EDA
  - Interactive visuals and dashboards
  - Predictive analysis

Introduction

- Project background and context

  SpaceX 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

  In this project, we will predict if the Falcon 9 first stage will land successfully using pre-existing data on the previous rocket launches from the API and its website.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

- Data was first collected using the SpaceX API by creating a get request. We do this by first defining a series of helper function to extract information using ID numbers.
- Next, the launch data is parsed using a GET request to make the JSON results more consistent. This data was then filtered to only include Falcon 9 launches.
- I also performed web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled `List of Falcon 9 and Falcon Heavy launches` with the use of BeautifulSoup and request libraries. The data is parsed and converted into a Pandas

Data Collection – SpaceX API

- Data collected via SpaceX API using a GET request. The data is then parsed as a JSON result and converted into the Pandas dataframe.

- The GitHub URL of the completed SpaceX API calls notebook: https://github.com/asuriy/Data-Science-Capstone-SpaceX-Launch-Prediction/blob/main/Code/1-spacex-data-collection-api.ipynb

Import libraries and define auxiliary functions

↓

Request and parse the SpaceX launch data using GET request

↓

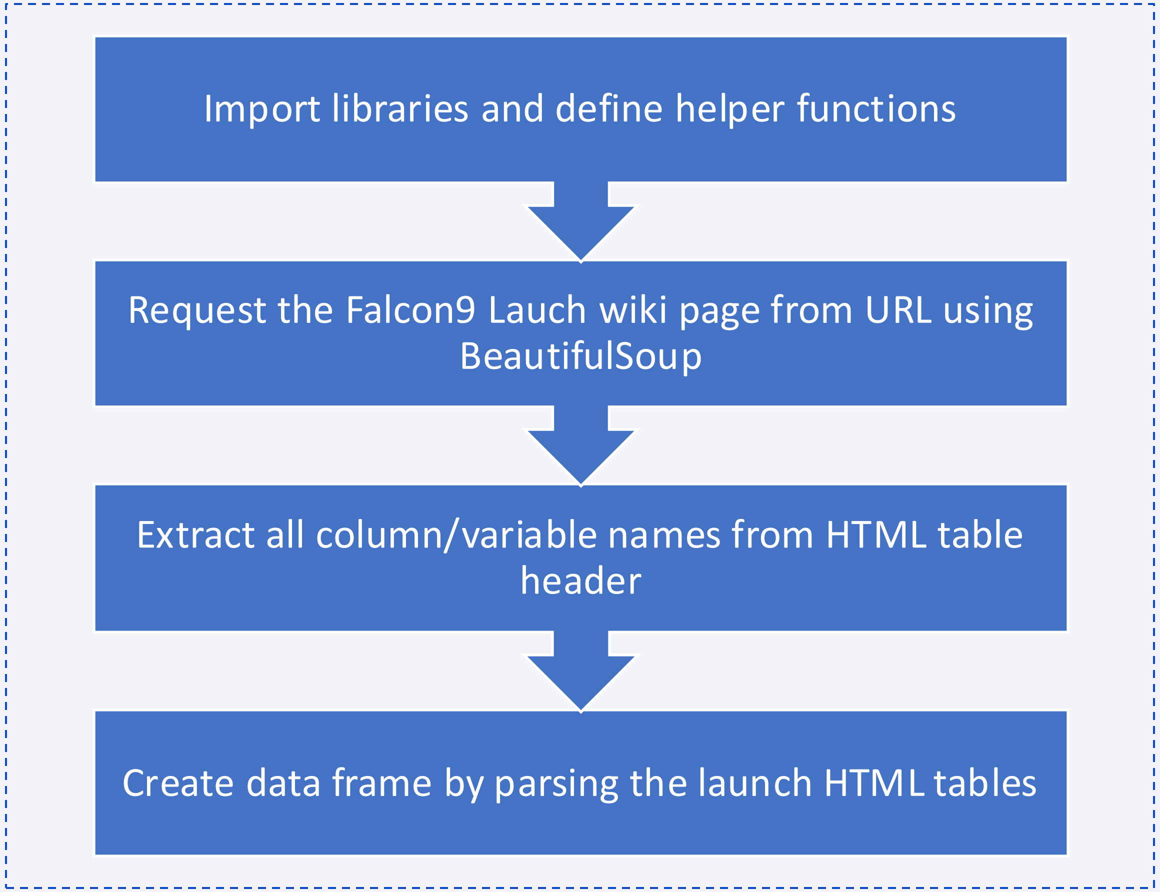
Filter dataframe to only include Falcon 9 launches

↓

Deal with missing values

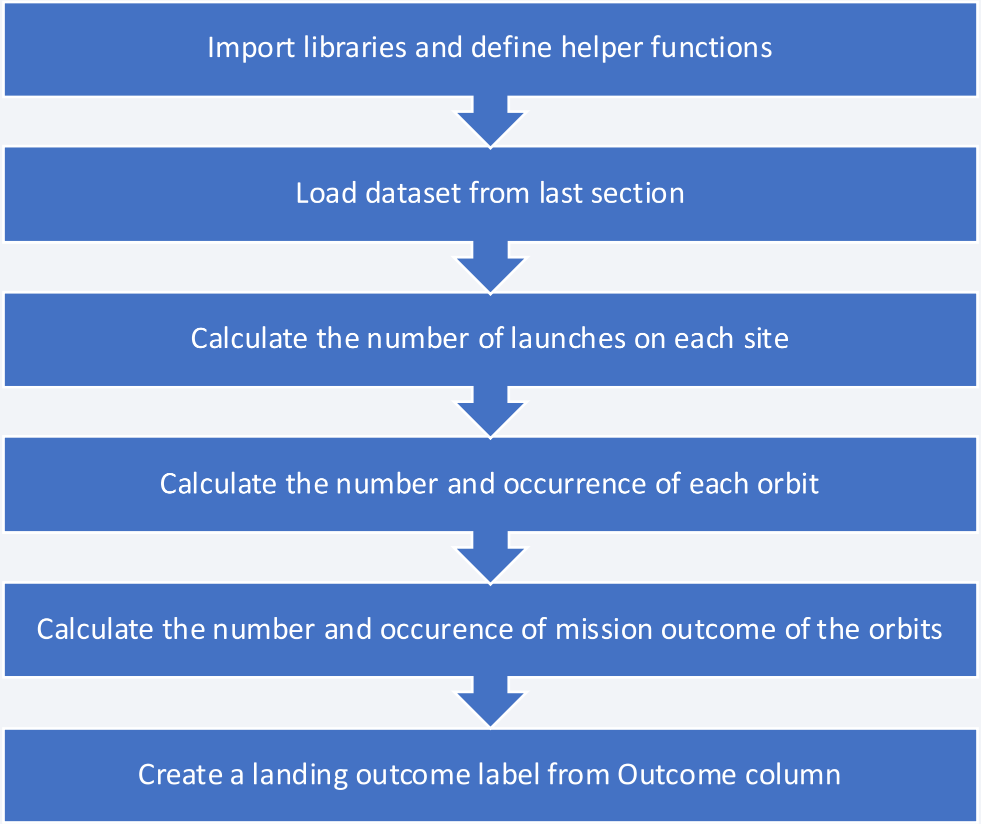
Data Collection - Scraping

- GitHub URL of the completed web scraping notebook: https://github.com/asuriy/Data-Science-Capstone-SpaceX-Launch-Prediction/blob/main/Code/2-webscraping.ipynb



Data Wrangling

- GitHub URL of your completed data wrangling related notebook: https://github.com/asuriy/Data-Science-Capstone-SpaceX-Launch-Prediction/blob/main/Code/3-spacex-data-wrangling.ipynb



EDA with SQL

- GitHub URL of completed EDA with SQL notebook: https://github.com/asuriy/Data-Science-Capstone-SpaceX-Launch-Prediction/blob/main/Code/4-eda-sql-coursera_sqllite.ipynb
- List of SQL queries:

```
%sql SELECT distinct(LAUNCH_SITE) FROM SPACEXTBL
```

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

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

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)'
```

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

```
%sql SELECT COUNT(MISSION_OUTCOME) FROM SPACEXTBL WHERE MISSION_OUTCOME = 'Success' OR MISSION_OUTCOME = 'Failure'
```

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

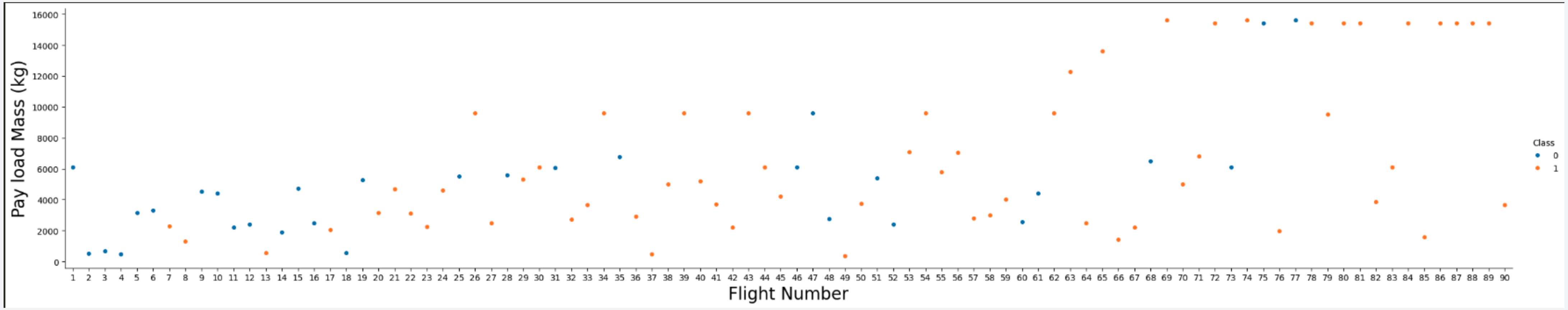
```
%sql SELECT SUBSTR(Date,6,2) AS Month, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND SUBSTR(Date,0,5) = '2015'
```

```
%sql SELECT LANDING_OUTCOME, COUNT(*) AS TOTAL FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)' OR LANDING_OUTCOME = 'Failure (drone ship)' AND Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY TOTAL DESC;
```

EDA with Data Visualization

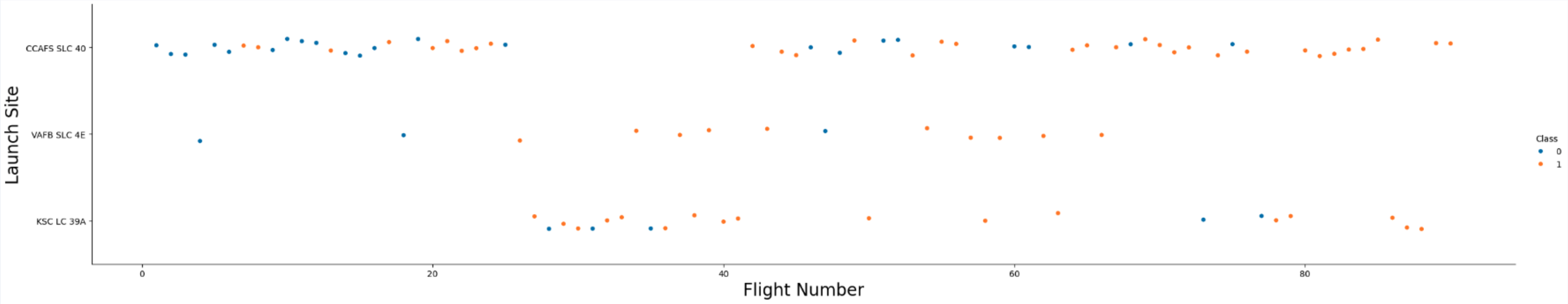
- GitHub URL of your completed EDA with data visualization notebook: https://github.com/asuriy/Data-Science-Capstone-SpaceX-Launch-Prediction/blob/main/Code/5-eda-data-visualisation.ipynb

- Plot to see how flight number and payload affect success/failure

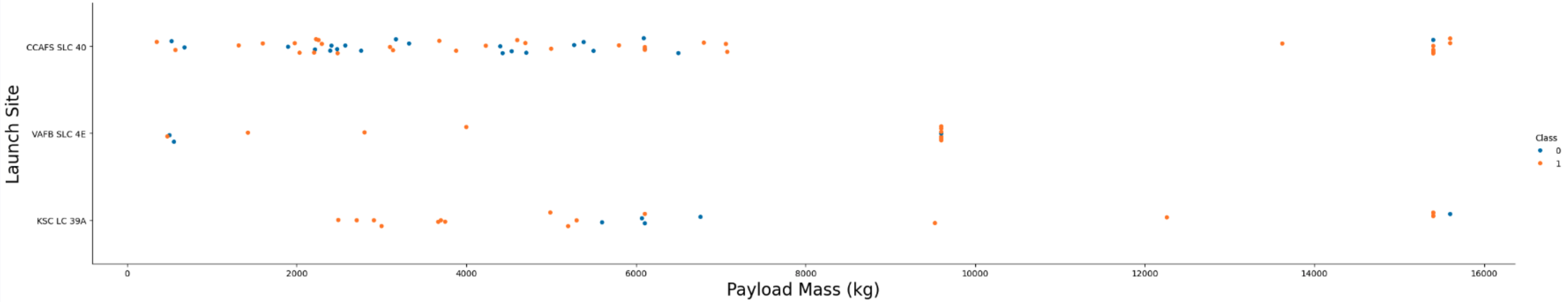


EDA with Data Visualization

- Flight number relationship with launch site and success/failure

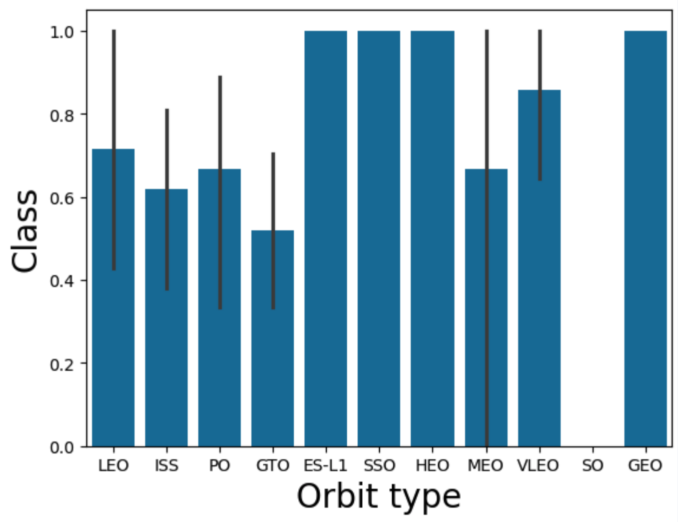


- Payload mass relationship with launch site and success/failure

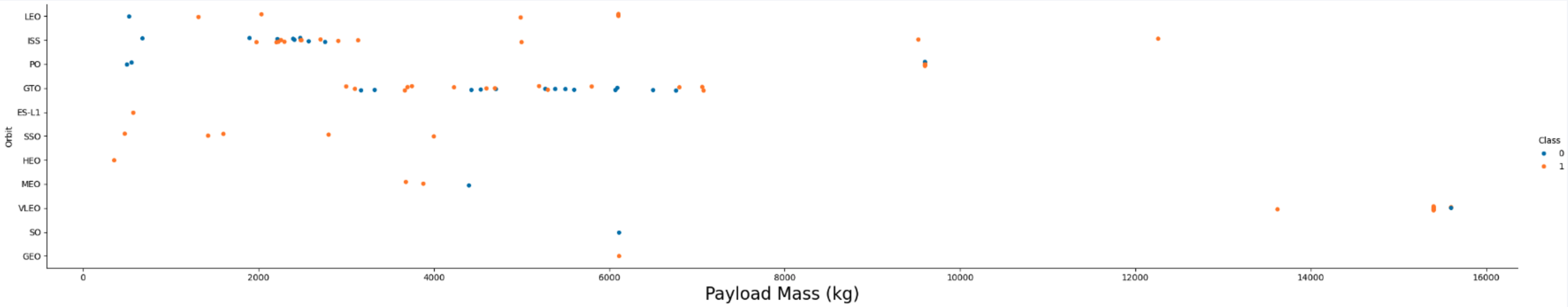


EDA with Data Visualization

- Success/failure of each orbit type

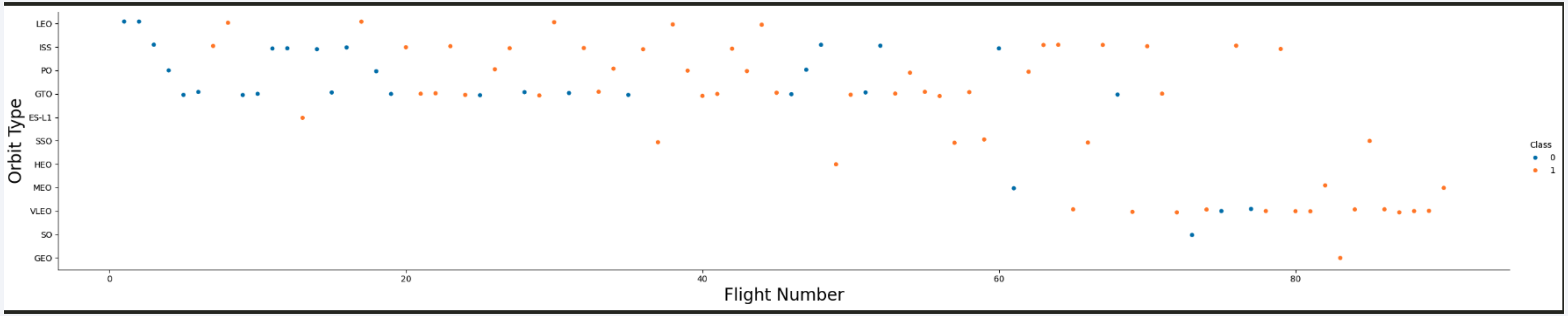


- Payload mass relationship with orbit type and success/failure

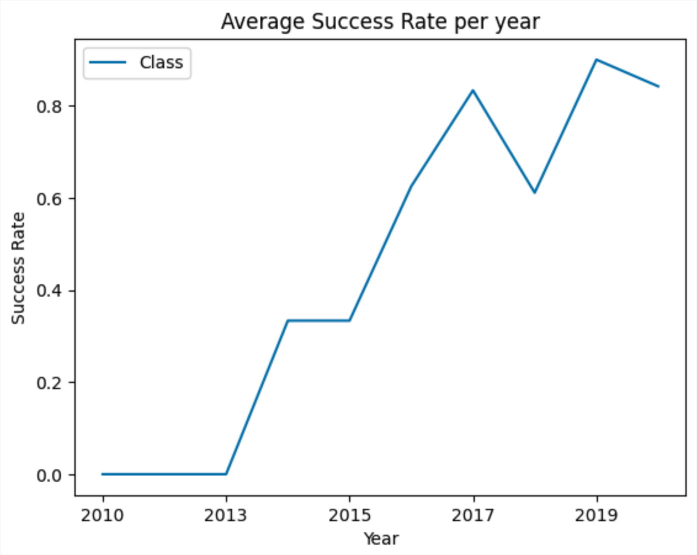


EDA with Data Visualization

- Flight number relationship with orbit type and success/failure



- Average success rate per year



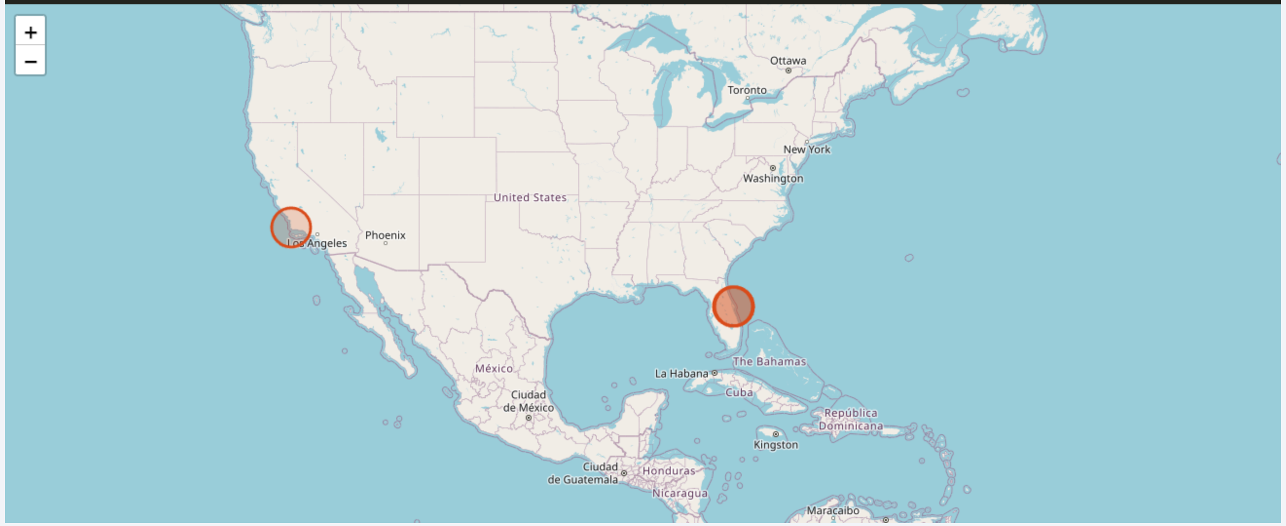
Build an Interactive Map with Folium

- GitHub URL of your completed interactive map with Folium map: https://github.com/asuriy/Data-Science-Capstone-SpaceX-Launch-Prediction/blob/main/Code/7-launch-site-location.ipynb

- NASA launch sites mapped using Circle markers:

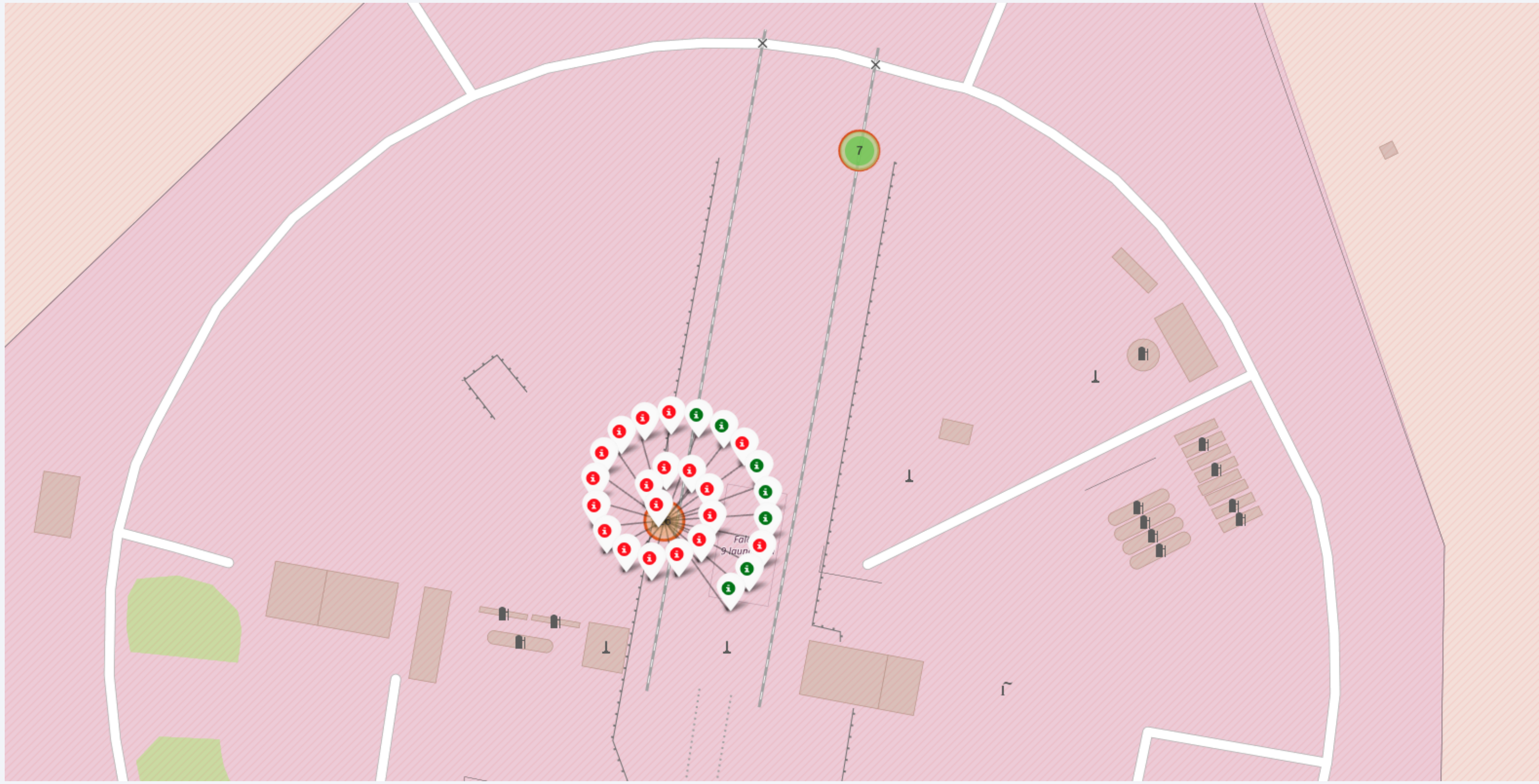
```python
# Initial the map
site_map = folium.Map(location=nasa_coordinate, zoom_start=5)
# For each launch site, add a Circle object based on its coordinate (Lat, Long) values. In addition, add Launch site name as a popup label
for lat, lng, label in zip(launch_sites_df['Lat'], launch_sites_df['Long'], launch_sites_df['Launch Site']):
    folium.vector_layers.CircleMarker(
        [lat,lng],
        radius=20,
        color='#d35400',
        fill=True,
        popup=label
    ).add_to(site_map)

site_map
```



Build an Interactive Map with Folium

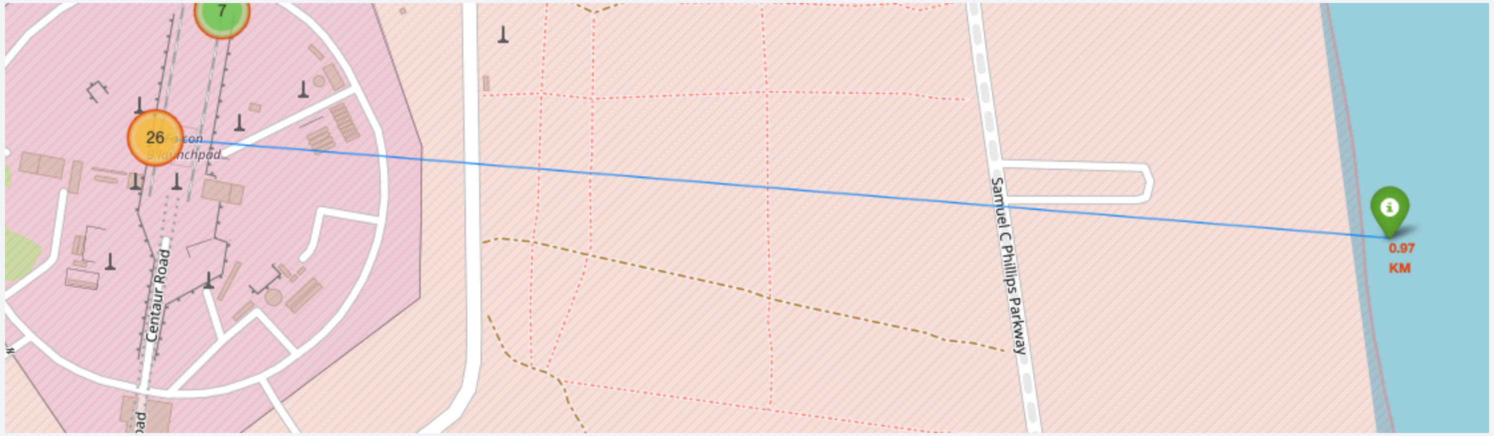
- Launch attempts using markers (green for successful and red for failure):



Build an Interactive Map with Folium

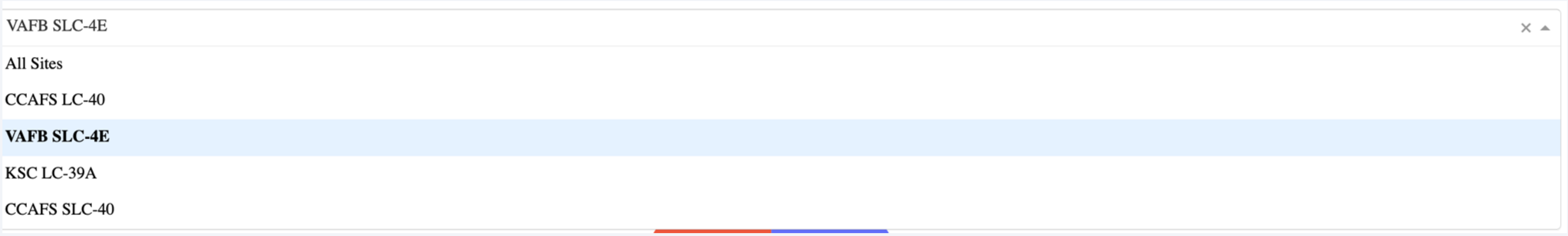
- Add formatting to show mouse pointer:

- Distance from highway to launch site (also did city and railway):



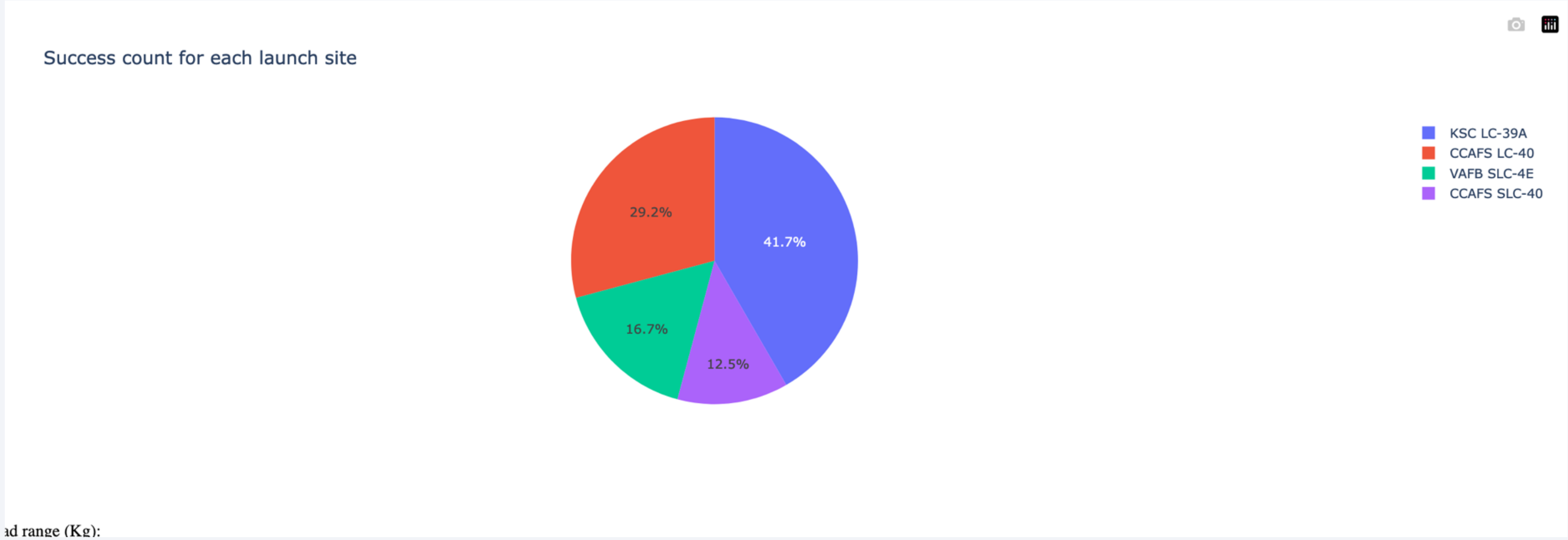
Build a Dashboard with Plotly Dash

- GitHub URL of completed Plotly Dash lab: https://github.com/asuriy/Data-Science-Capstone-SpaceX-Launch-Prediction/blob/main/Code/8-spacex-dash-app.py
- Option to view all sites or singular launch site success and failure rate:



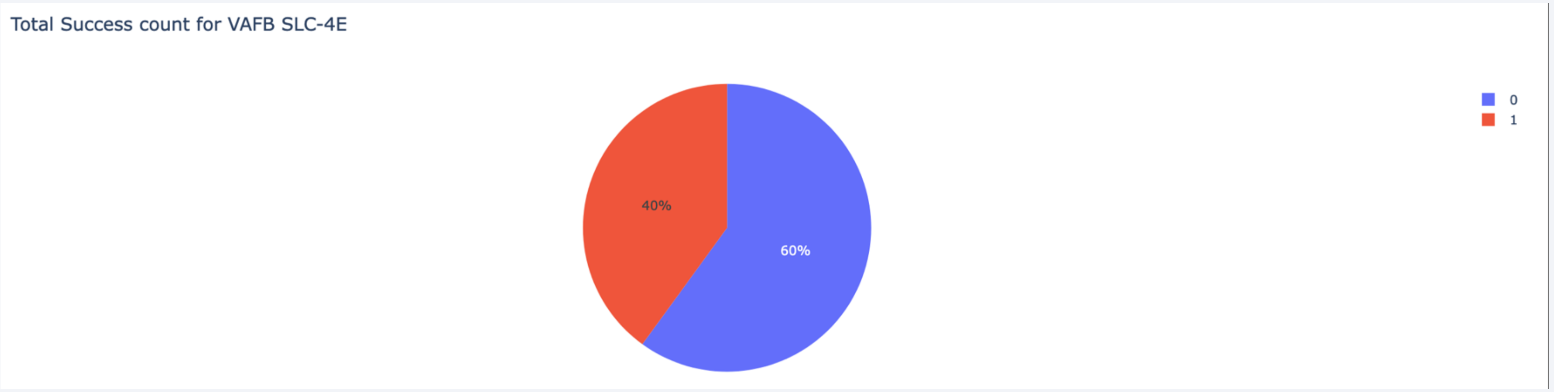
Build a Dashboard with Plotly Dash

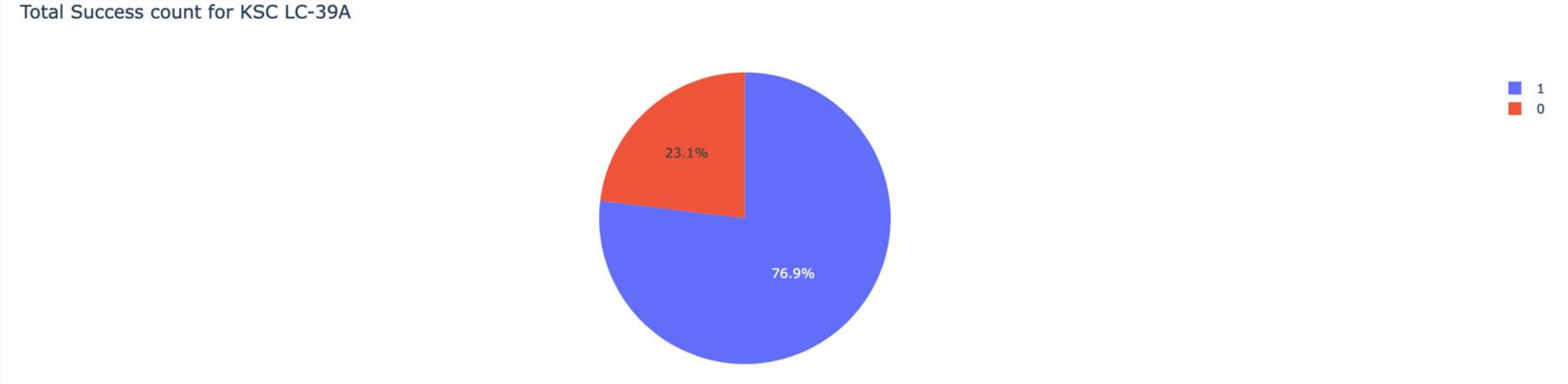
- Pie chart showing distribution of success for each launch site:



Build a Dashboard with Plotly Dash

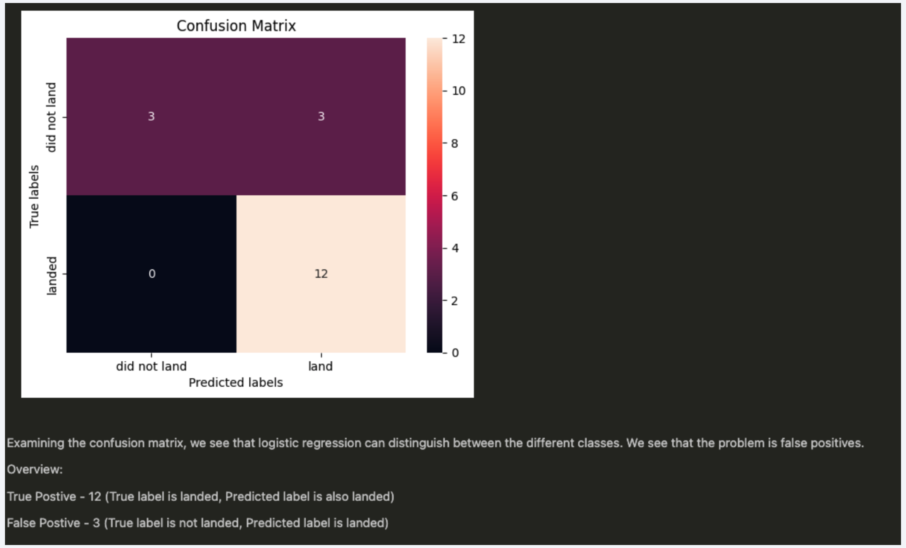
- Dropdown allows us to choose an individual site:

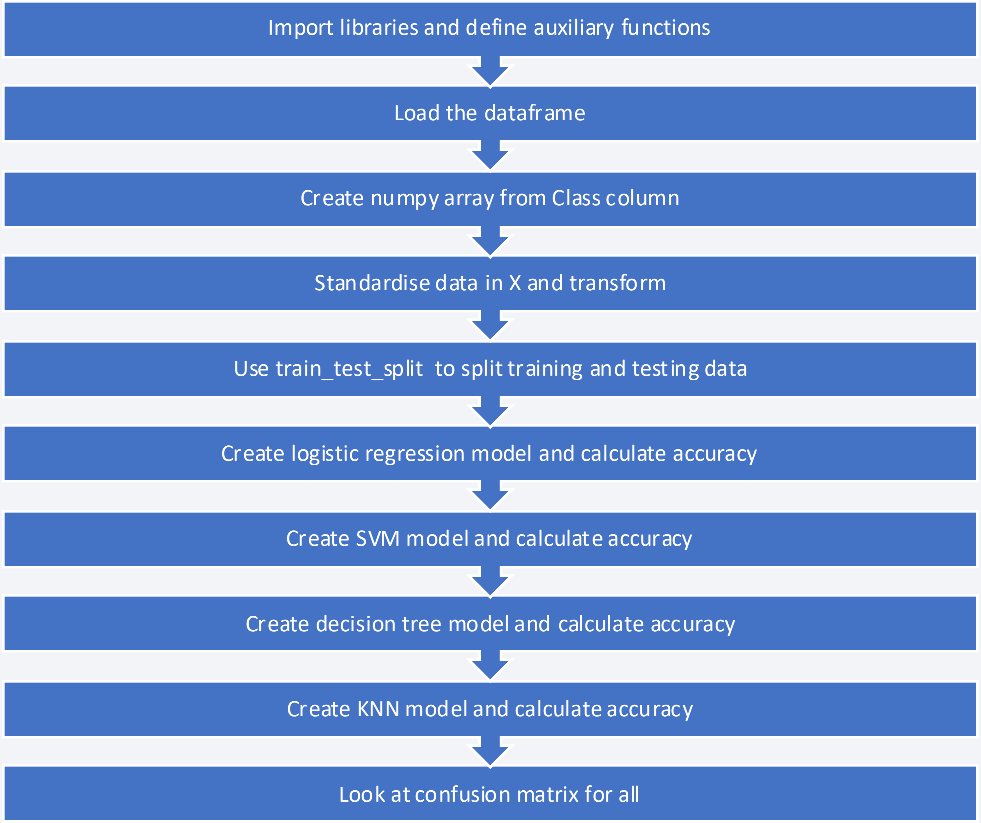




Predictive Analysis (Classification)

- GitHub URL of completed predictive analysis lab: https://github.com/asuriy/Data-Science-Capstone-SpaceX-Launch-Prediction/blob/main/Code/9-spacex-machine-learning-prediction.ipynb
- KNN, SVM and logistic models tied for the highest accuracy score. These results need to be taken with caution due to a low test sample count of 18.
- Example of confusion matrix:





Results

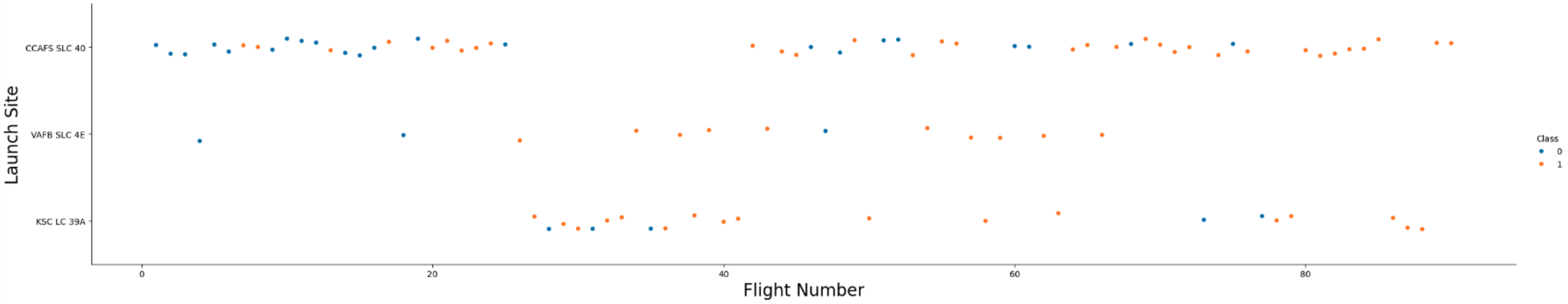
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

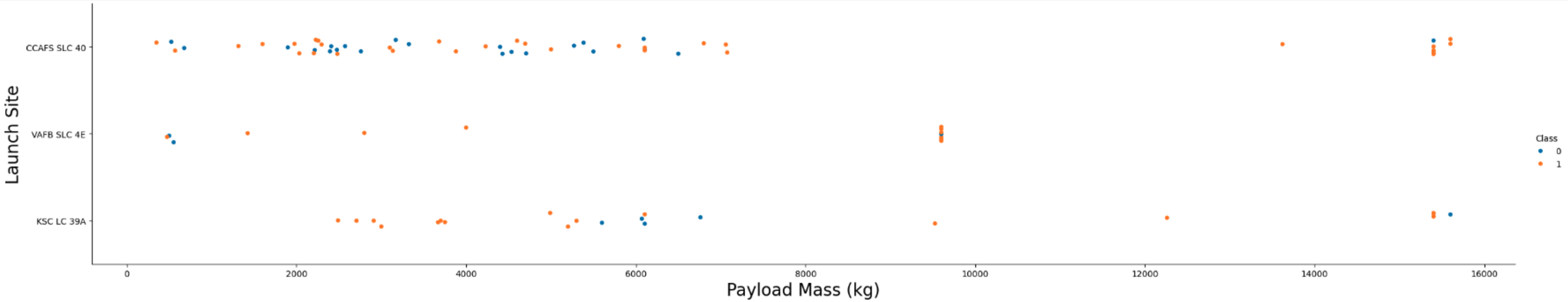
- A scatter plot of Flight Number vs. Launch Site



- KSC LC 39A has a higher success rate, with CCAFS showing a lower proportion of succession. Increase in flight number seems to increase success rate.

Payload vs. Launch Site

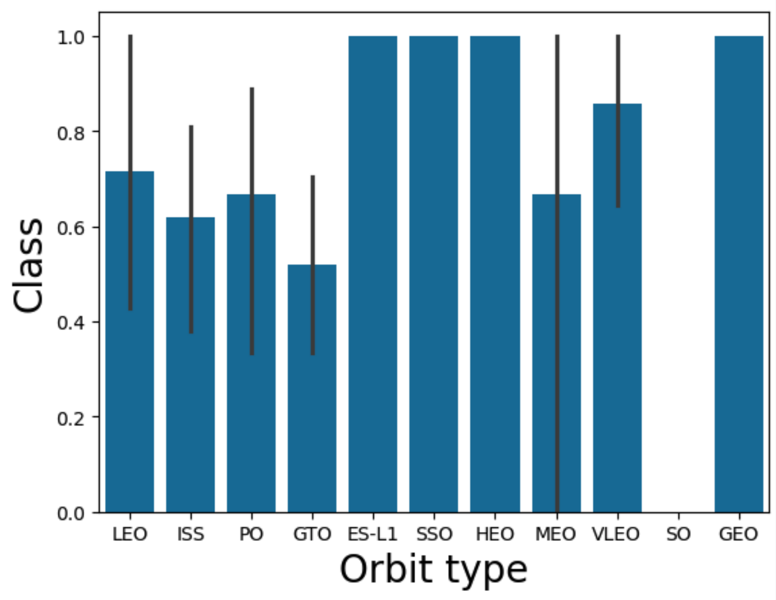
- A scatter plot of Payload vs. Launch Site



- KSC LC 39A has 100% success rate for payload mass between 2000 kg and 5000 kg. CCAFS launch site has variable success rates for payloads under 7000 kg.

Success Rate vs. Orbit Type

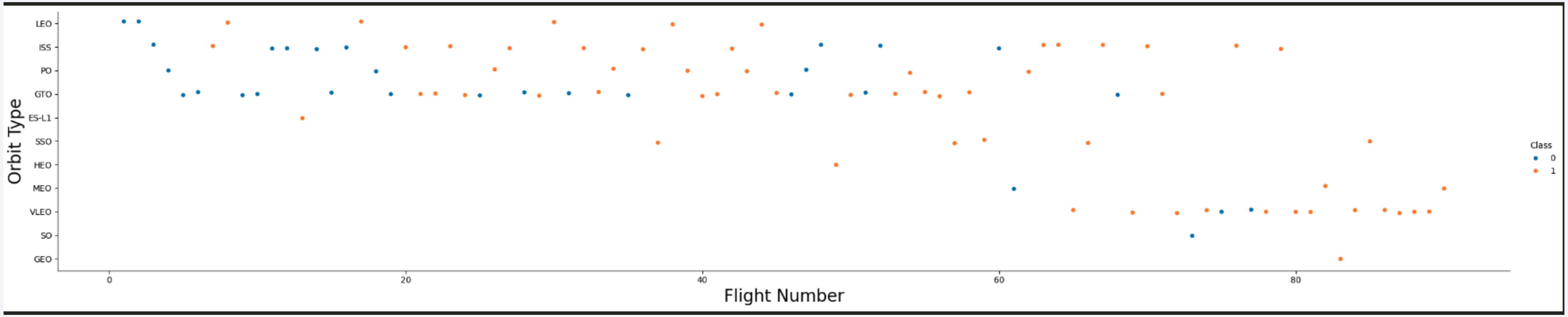
- A bar chart for the success rate of each orbit type



- ES-L1, SSO, HEO and GEO orbit types were the most successful with 100%. SO was the least successful with 0%.

Flight Number vs. Orbit Type

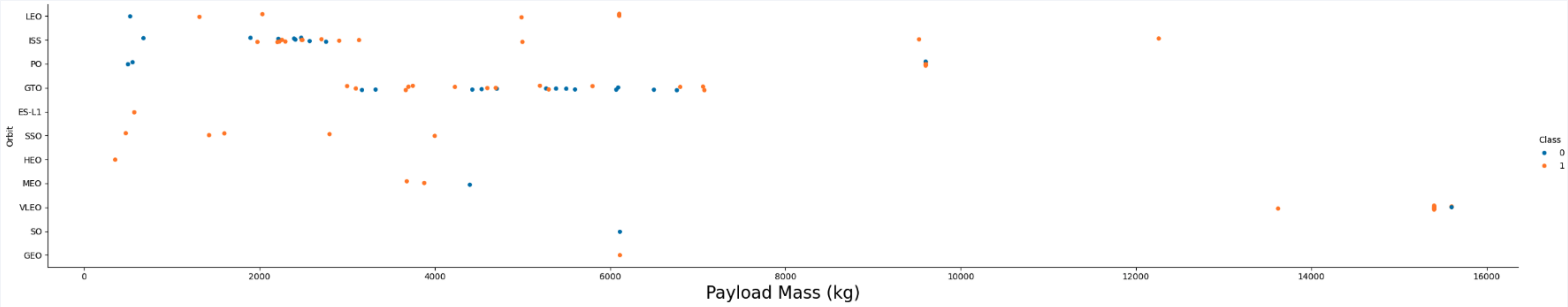
- A scatter point of Flight number vs. Orbit type



- Scatter point confirms the orbit type success rates in the bar chart. Increase in flight numbers, increases success rate.

Payload vs. Orbit Type

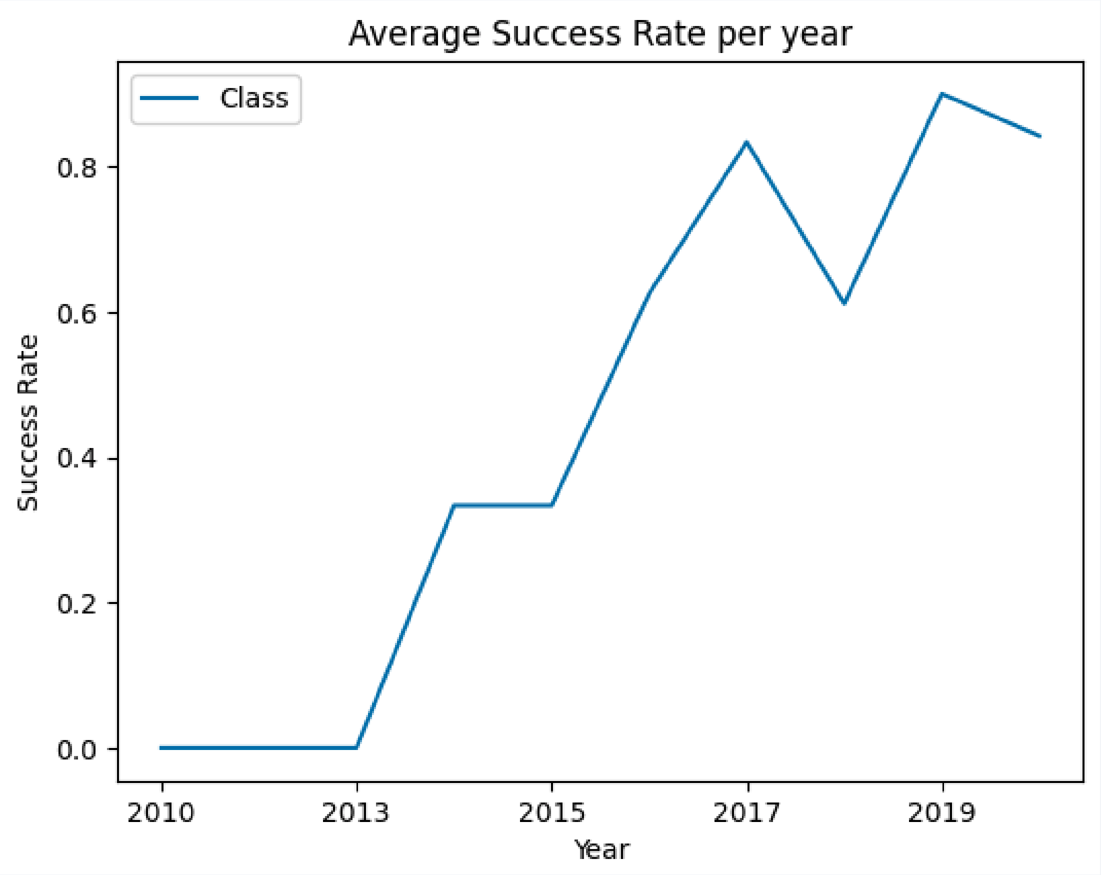
- A scatter point of payload vs. orbit type



- ES-L1, SSO and HEO have 100 % success rate in the lower payload mass.

Launch Success Yearly Trend

- A line chart of yearly average success rate:
- Steady climb of success rate up to 2017 and then in 2018, it momentarily dipped before the success rate increases. 2020 sees a dip in success rate.



All Launch Site Names

- Find the names of the unique launch sites

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

- All the distinct launch sites from the SPACEXTBL.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites 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

- Shows the first 5 results of site ‘CCA’ in the table in chronological order; the mission outcome looks to be all successes with variable landing outcomes.

Total Payload Mass

- Total payload mass = 45596 kg
- This query was specific to the NASA (CRS) team.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1 = 2928.4 kg

First Successful Ground Landing Date

- First date was 2015-12-22, the launches started in 2010 so this means it took five years of developing for a successful landing outcome on the ground.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

```
 * 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

- Calculate the total number of successful and failure mission outcomes = 98

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

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

- 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

Landing_Outcome	TOTAL
Success (ground pad)	9
Failure (drone ship)	5

Section 3

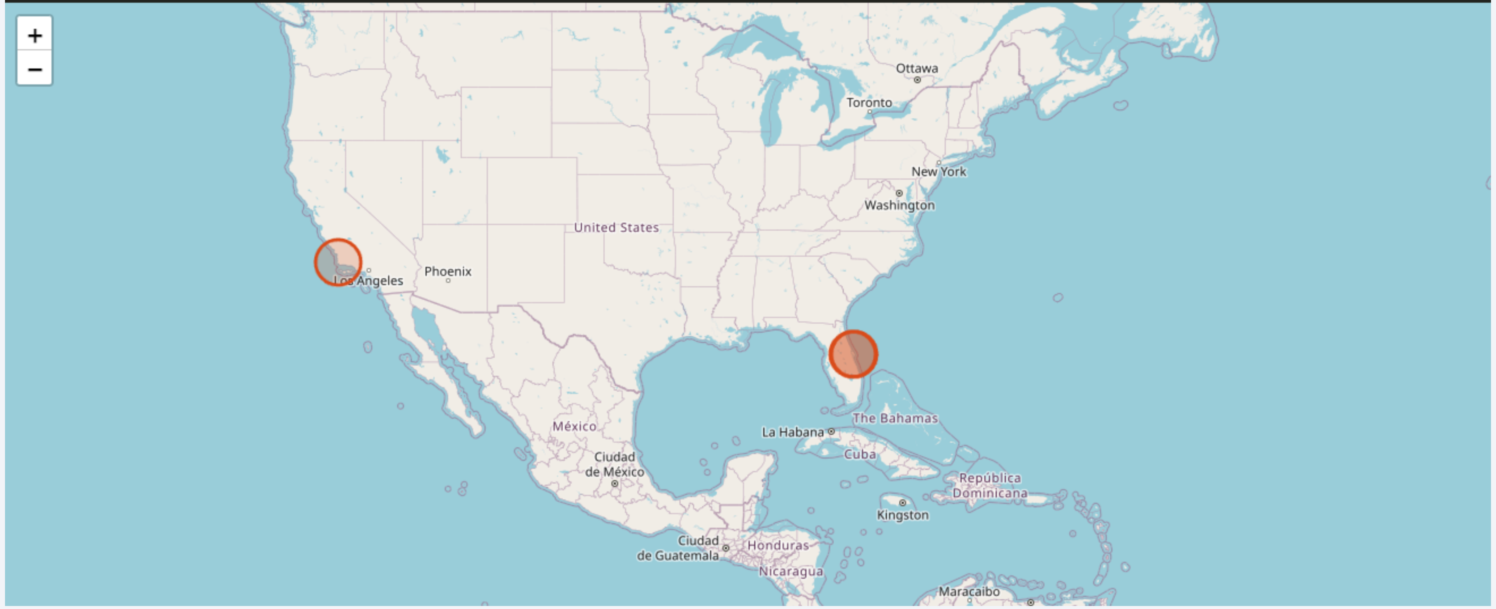
Launch Sites Proximities Analysis

All launch sites

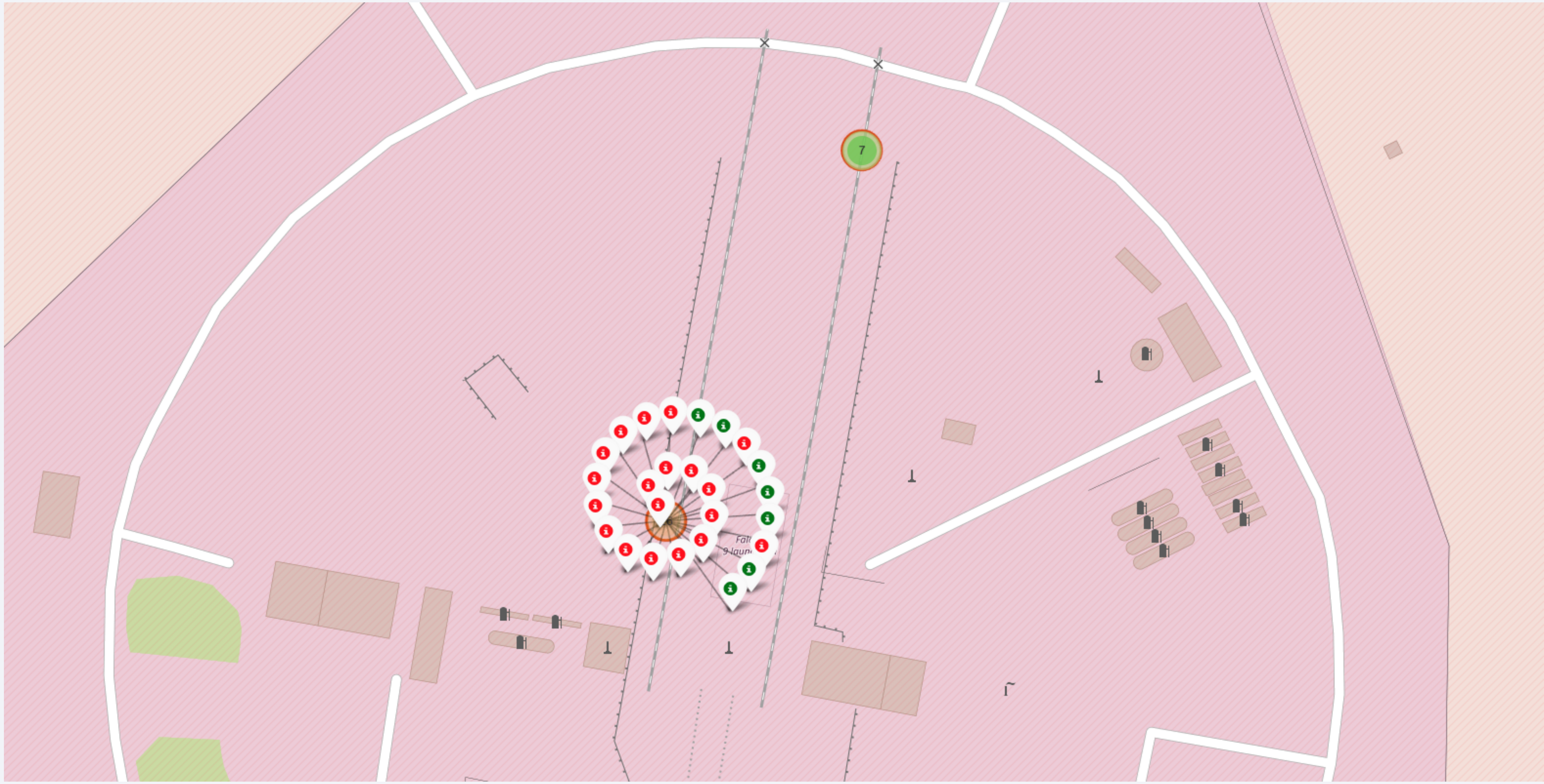
- Launch sites are mainly based near the coast.

```
# Initial the map
site_map = folium.Map(location=nasa_coordinate, zoom_start=5)
# For each launch site, add a Circle object based on its coordinate (Lat, Long) values. In addition, add Launch site name as a popup label
for lat, lng, label in zip(launch_sites_df['Lat'], launch_sites_df['Long'], launch_sites_df['Launch Site']):
    folium.vector_layers.CircleMarker(
        [lat,lng],
        radius=20,
        color='#d35400',
        fill=True,
        popup=label
    ).add_to(site_map)

site_map
```

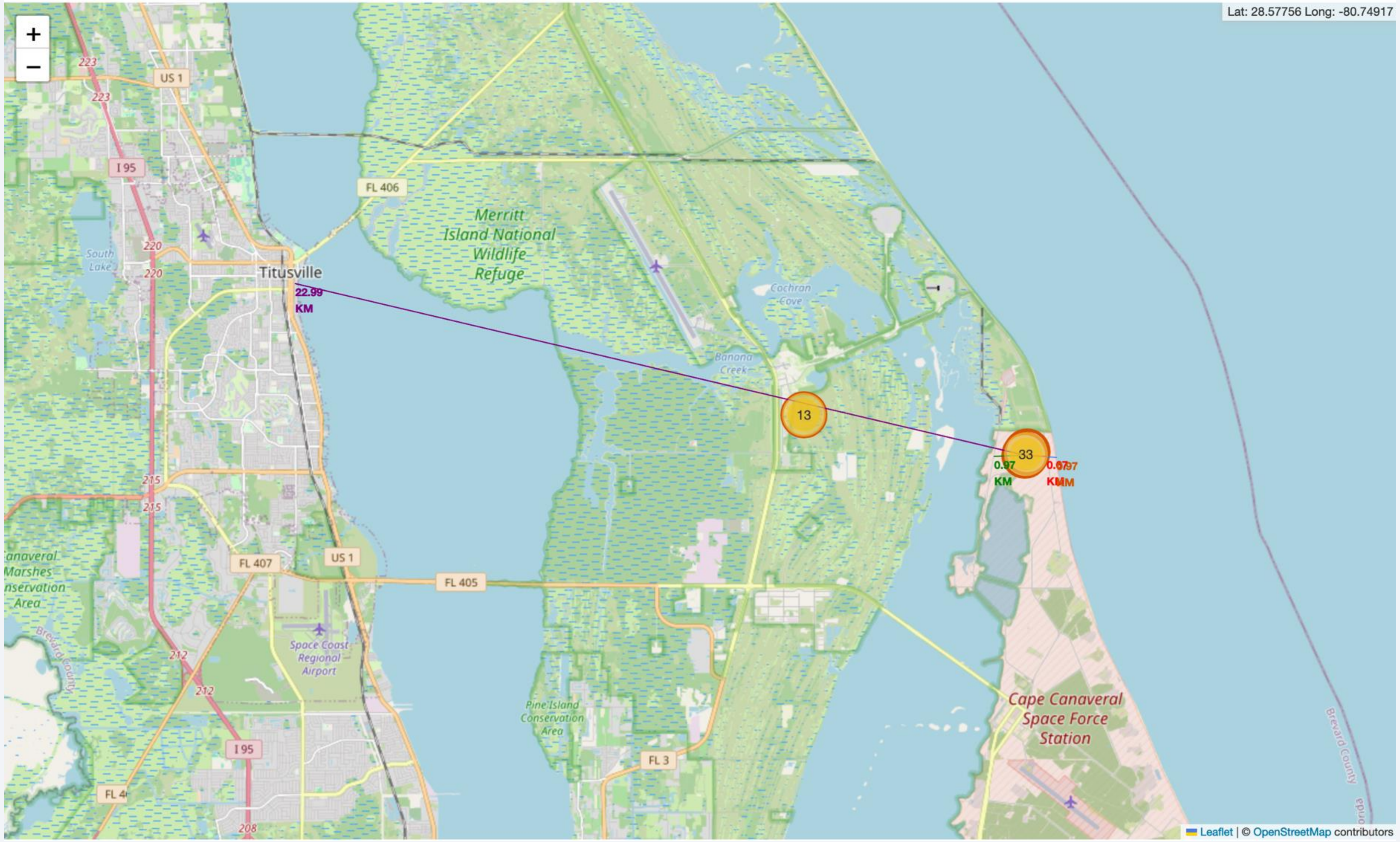


Launch outcomes



Launch site proximity highway, railway and city

- Close to coast, rail and highway but further from the city.

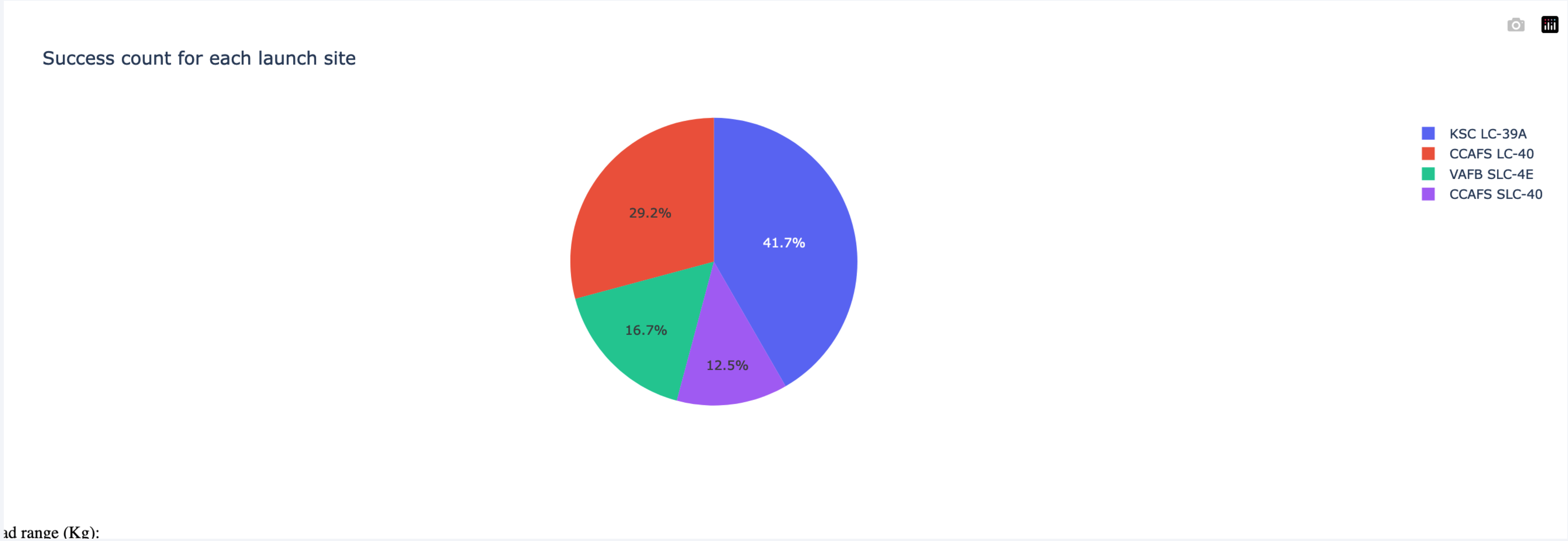


Section 4

Build a Dashboard with Plotly Dash

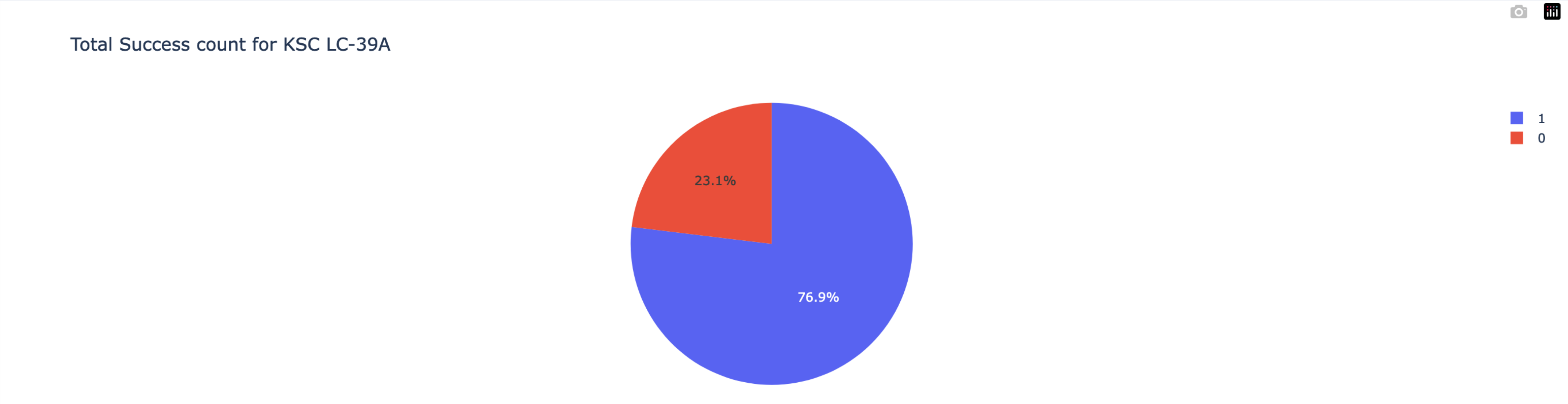
Pie chart of all launch site success rates

- KSC has the highest success rate and CCAFS has the lowest overall.



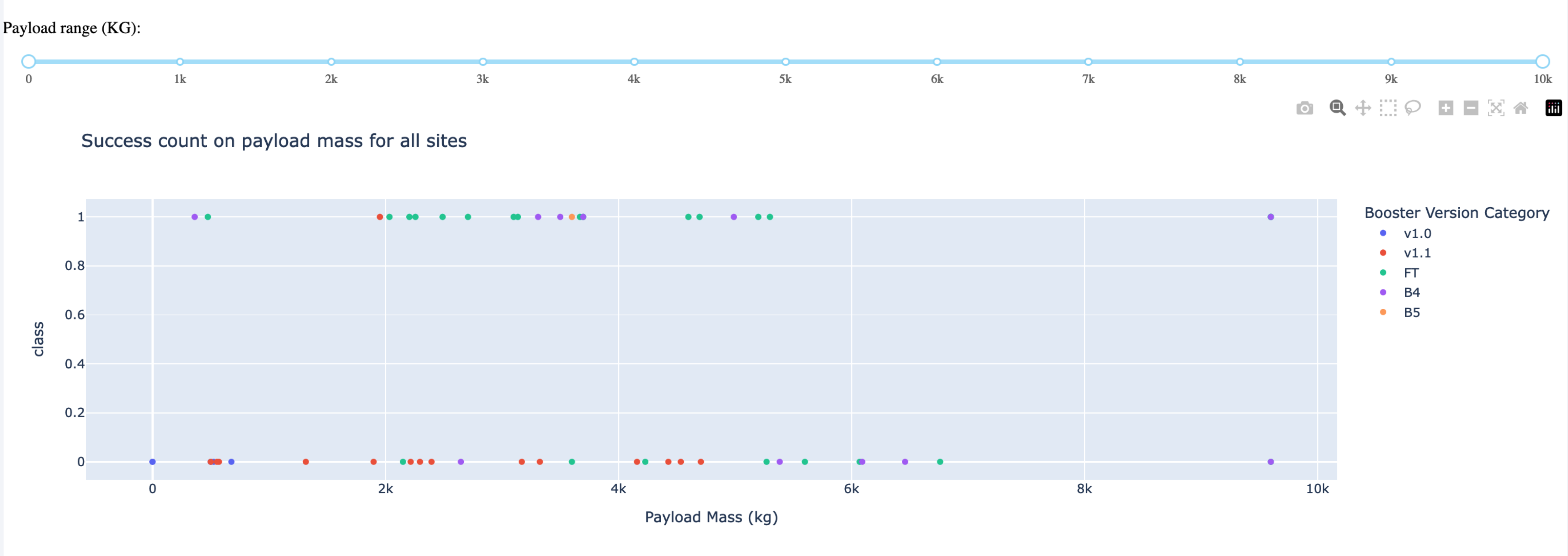
Pie chart with highest launch site success ratio

- KSC has the highest launch site ratio with 76.9 % success rate.



Payload vs Launch Outcome scatter plot

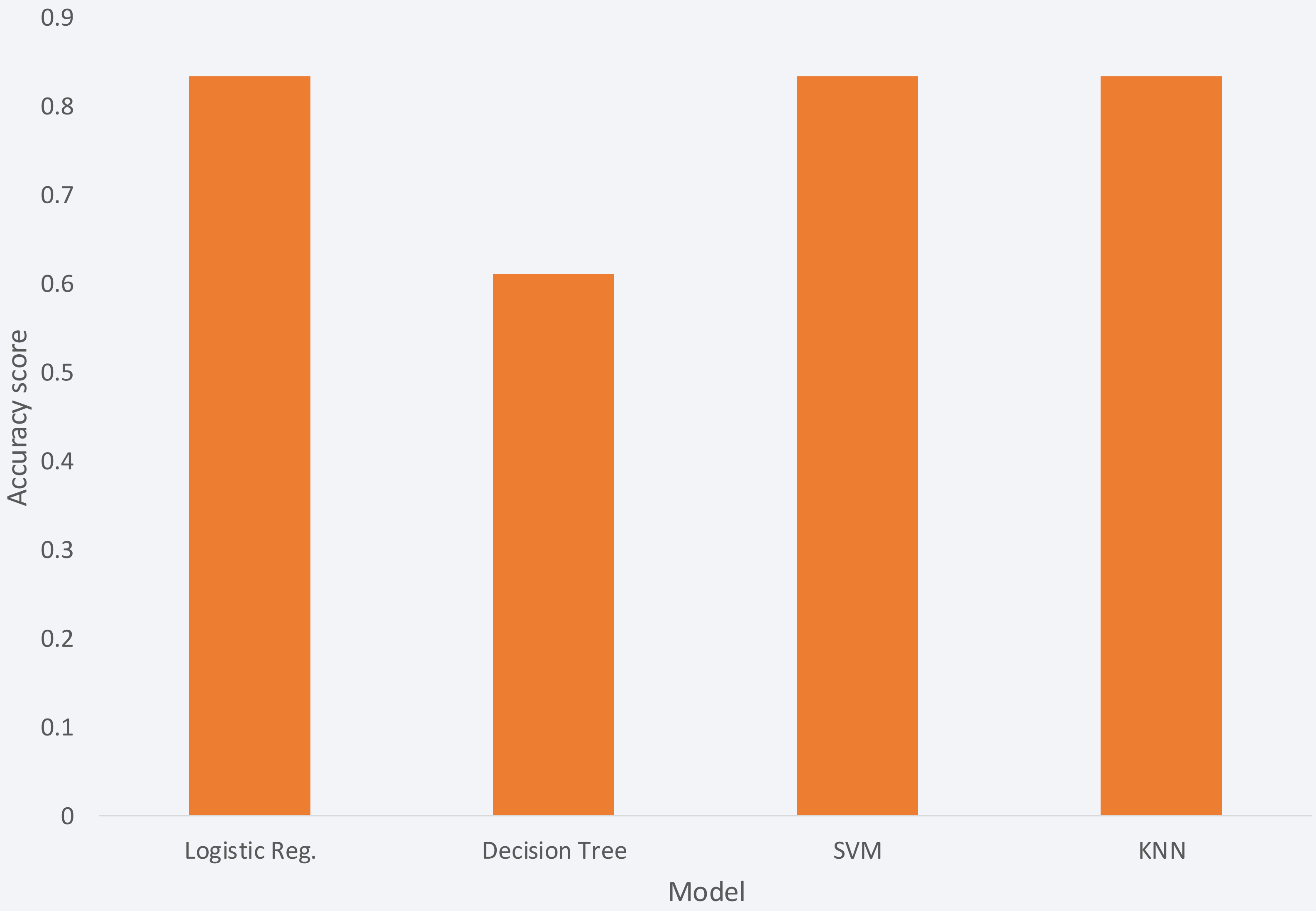
- FT booster version has the most success, with a payload between 2000 and 6000 kg.



Section 5

Predictive Analysis (Classification)

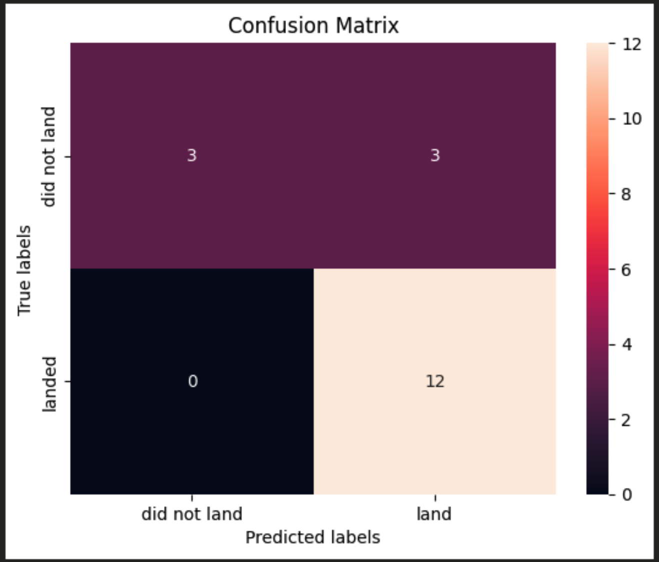
Classification Accuracy



Confusion Matrix

- Confusion matrix of one of the best performing models (KNN and SVM gave the same output):

```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.

Overview:

True Postive - 12 (True label is landed, Predicted label is also landed)

False Postive - 3 (True label is not landed, Predicted label is landed)

Conclusions

- SVM, KNN and logistic regression models performed similarly at 83 %.
- The launch sites are close to the coast.
- Launch success increases over time with more experience and knowledge of rocket builds.
- KSC launch site has the highest success rate.
- ES-L1, GEO, HEO and SSO have 100 % success with orbit type and SO has 0 % success.
- Overall, higher payload mass correlates with higher success rate.

Thank you!

Appendix

- Full GitHub repo: https://github.com/asuriy/Data-Science-Capstone-SpaceX-Launch-Prediction/tree/main