IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jeongeun Son
04-26-2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis Result
  - Interactive Analytics in Screenshots
  - Predictive Analytics Result

Introduction

- Project background and context

  SpaceX promotes the launch of its Falcon 9 rocket on its website at a price of $62 million, which is considerably lower than the prices of other providers who charge $165 million or more. This is mainly because SpaceX can reuse the first stage of the rocket. To estimate the cost of a launch, it is important to predict if the first stage can be reused. As a data scientist working for a new rocket company named Space Y, our goal is to determine the price of each launch using the data of SpaceX available online.

- Problems you want to find answers

  - Identifying factors that determine the successful landing of the rocket
  - The interaction amongst various features that determine its success rate.
  - The best conditions needed to be in place to ensure a successful landing program

Section 1

Methodology

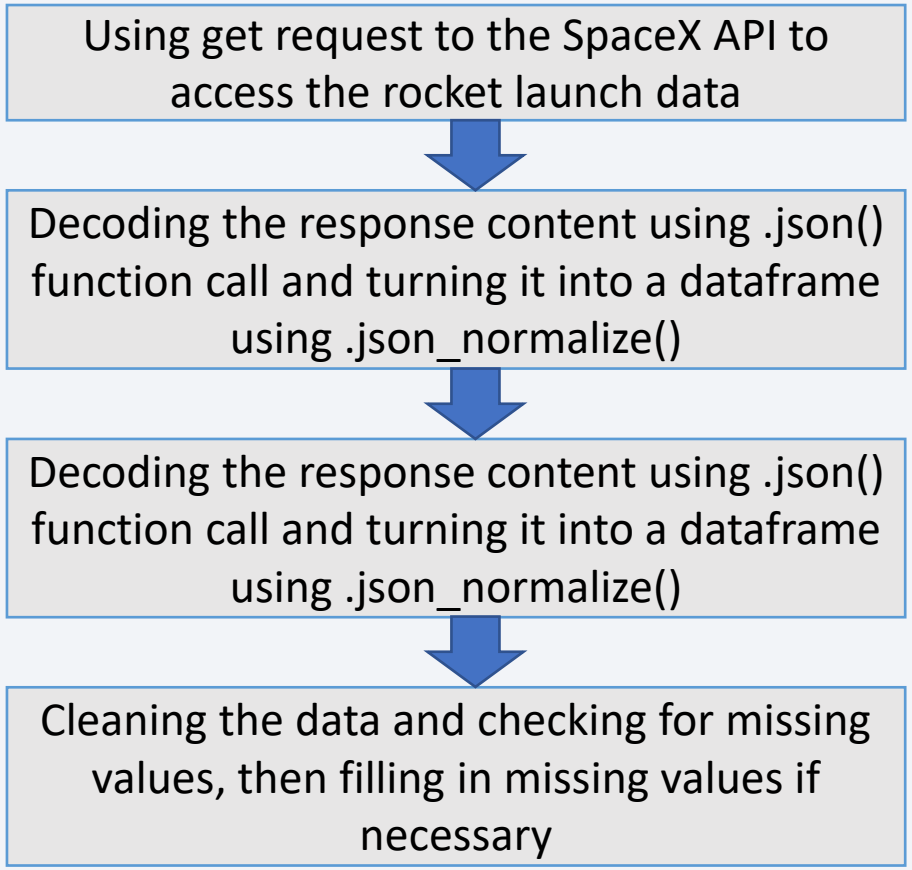
Methodology

Executive Summary

- Data collection methodology:
  - Data on the rocket launch was collected via SpaceX API and web scrapping from Wikipedia.
- Perform data wrangling
  - Data was processed using one-hot encoding to deal with categorial data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models were built, tuned, and evaluated to ensure the best results

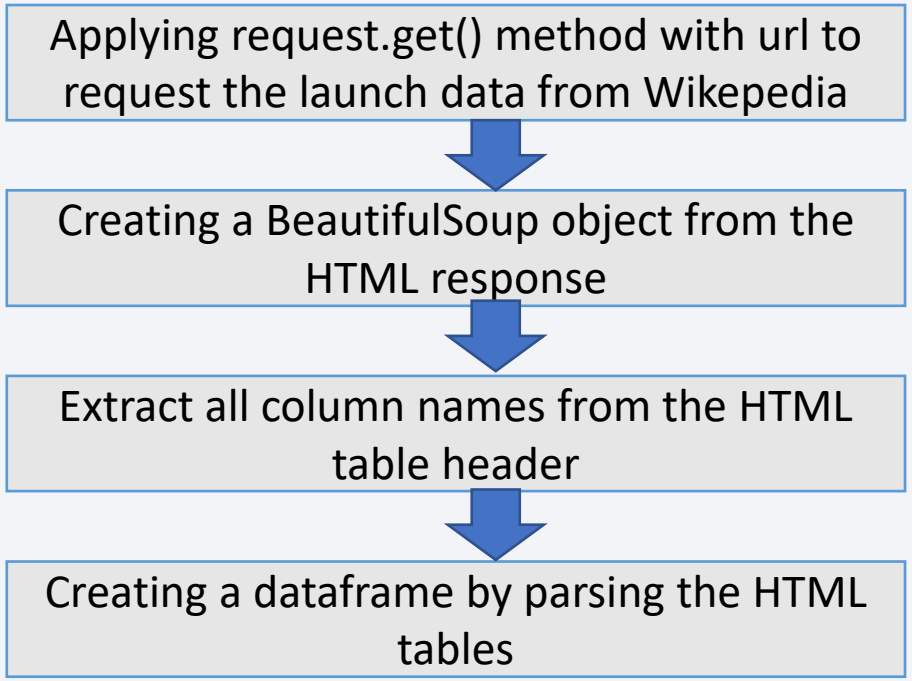
Data Collection

- SpaceX API



GitHub URL: https://github.com/sonje0113/IBM-Applied-Data-Science-Capstone/blob/main/Data_Collection__SpaceX_API.ipynb

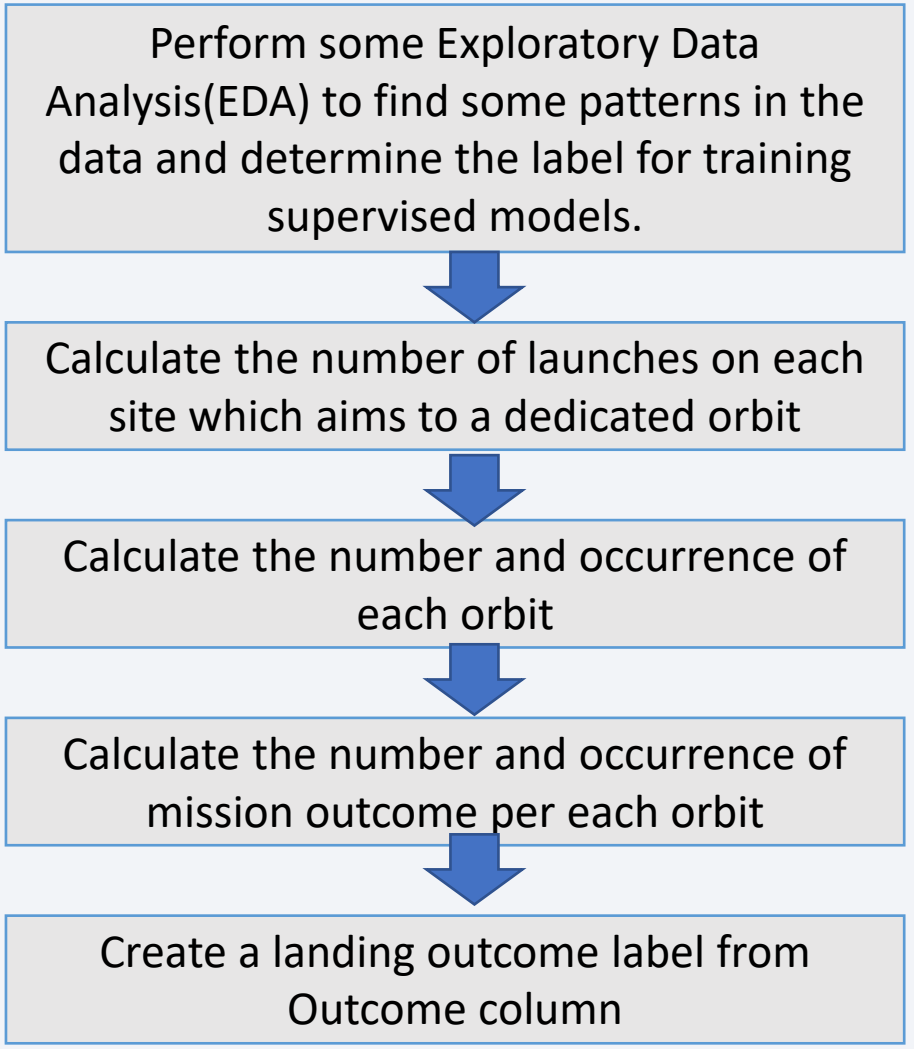
- Web scraping



GitHub URL: https://github.com/sonje0113/IBM-Applied-Data-Science-Capstone/blob/main/Data_Collection__SpaceX_Web_Scarping.ipynb

Data Wrangling

- Perform exploratory Data Analysis and determine Training Labels



GitHub URL: https://github.com/sonje0113/IBM-Applied-Data-Science-Capstone/blob/main/Data_Wrangling__SpaceX.ipynb

EDA with Data Visualization

- Perform exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib to predict if the Falcon 9 first stage will land successfully.
- Scatter plots, a bar chart, and a linear plot were generated.
  - Scatter plots for Flight Number vs. Payload Mass/Flight Number vs. Launch Site/ Payload Mass vs. Launch Site/ Flight Number vs. Orbit/ Payload Mass vs. Orbit.
  - Bar chat for Success Rate vs. Orbit type
  - Line plot for Success Rate per year to see the launch success yearly trend.

GitHub URL: https://github.com/sonje0113/IBM-Applied-Data-Science-Capstone/blob/main/EDA_Data_Visualization__SpaceX_DV.ipynb

EDA with SQL

Performed SQL queries are as follows.

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

GitHub URL: https://github.com/sonje0113/IBM-Applied-Data-Science-Capstone/blob/main/EDA_SQL__SpaceX_SQL.ipynb

Build an Interactive Map with Folium

To investigate the launch success rate depending on the location and proximities of a launch site, i.e., the initial position of rocket trajectories, an interactive map with Folium was generated.

- Using the coordinate of the latitude and longitude, each launch site was marked, where map objects such as markers, circles, popup labels, and lines were used.
- To indicate the success or failure of launches for each site on the Folium map, colored markers of success (Green) and failure (Red) launches were added using MarkerCluster().
- The distance between a launch site to its proximities like railways, highways, and coastlines was calculated and added with lines on the map.

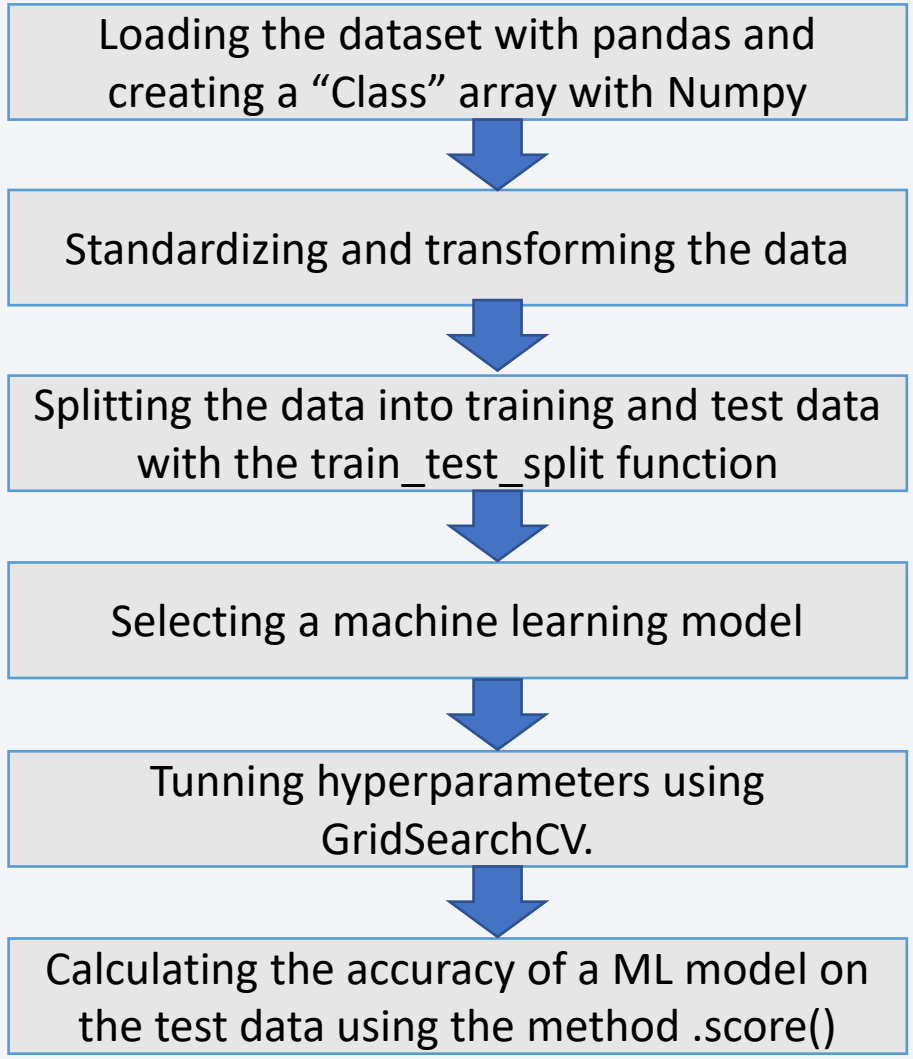
GitHub URL: https://github.com/sonje0113/IBM-Applied-Data-Science-Capstone/blob/main/Interactive_Visual_Analytics__SpaceX_Folium.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard with Plotly Dash was built with Launch Sites dropdown list and slider of Payload range, which allows users to explore the information they need by adjusting them.
- Once Launch Site is selected (all sites/a specific site) on the dashboard, a pie chart is plotted showing the total successful launches for all sites and the total counts of success and failed launches for a certain site.
- Once Payload range is adjusted on the dashboard, a scatter chart is plotted to show the relationship between Payload and Outcome for different booster versions.

GitHub URL: https://github.com/sonje0113/IBM-Applied-Data-Science-Capstone/blob/main/Interactive_Dashboard__SpaceX_Plotly_Dash.py

Predictive Analysis (Classification)

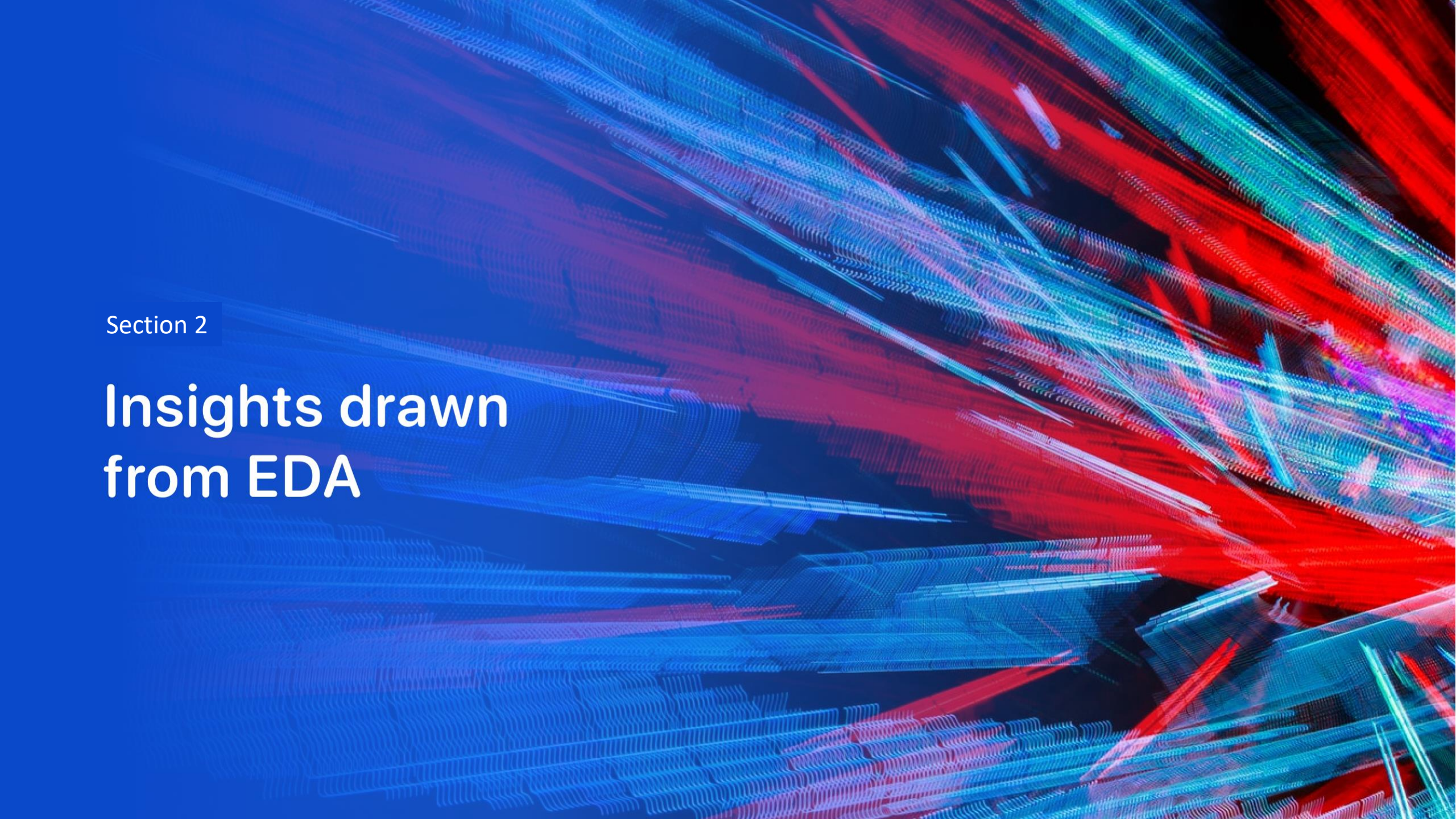


- The following ML models were built.
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K Nearest Neighbor
- Such approaches were compared in terms of accuracy calculated by the method .score() to find the best performing classification model.

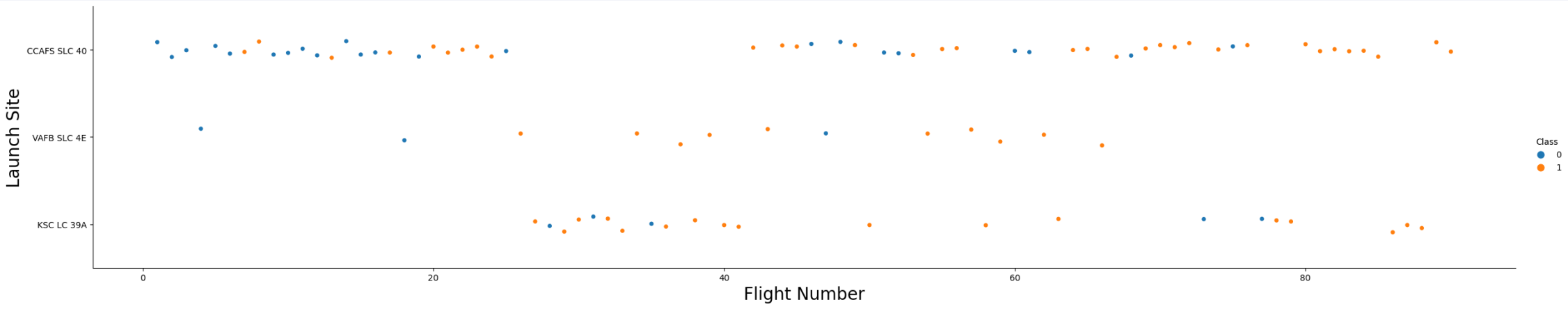
GitHub URL: https://github.com/sonje0113/IBM-Applied-Data-Science-Capstone/blob/main/ML_Prediction__SpaceX_Classification.ipynb

Results

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

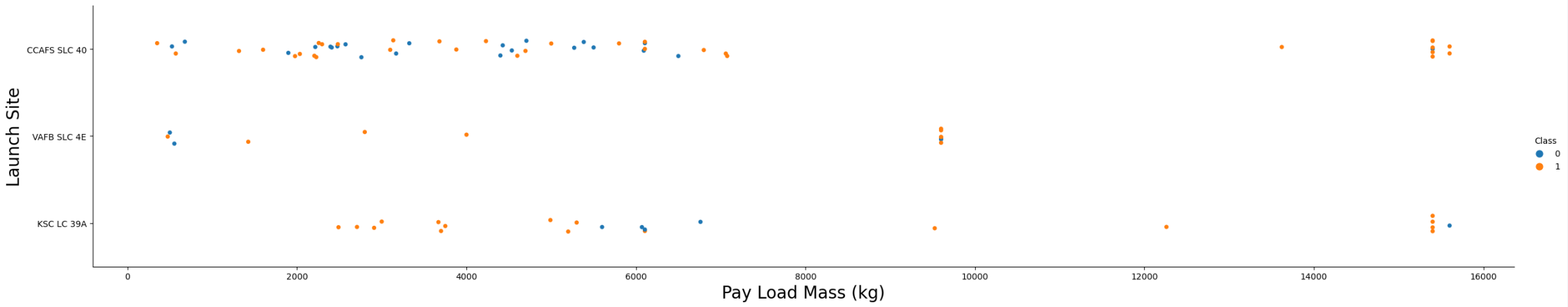

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



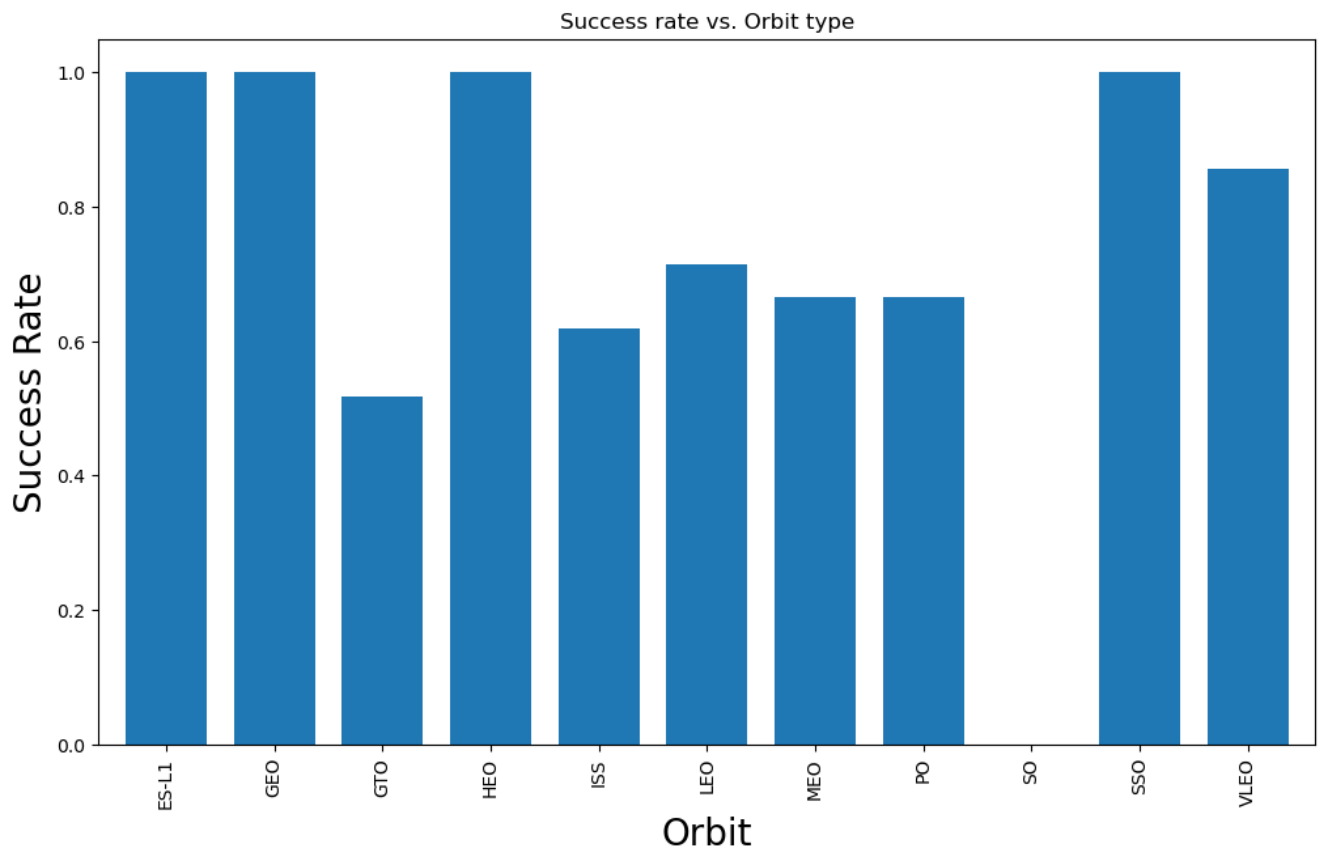
- When the flight number is less than 20, the success rate is significantly low (see blue markers), but it shows an increasing success rate as the flight number increases.

Payload vs. Launch Site



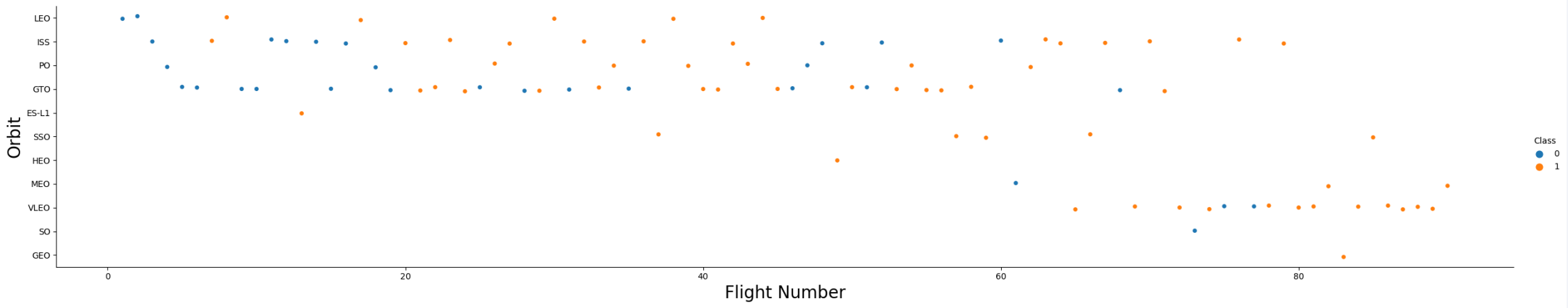
- The larger the payload mass, the higher the success rate for the rocket launches.
- CCAFS SLC 40 and KSC LC 39A show the highest success rate at ~15000 kg of payload mass, while WAFB SLC 4E has the highest success rate at ~9500 kg.

Success Rate vs. Orbit Type



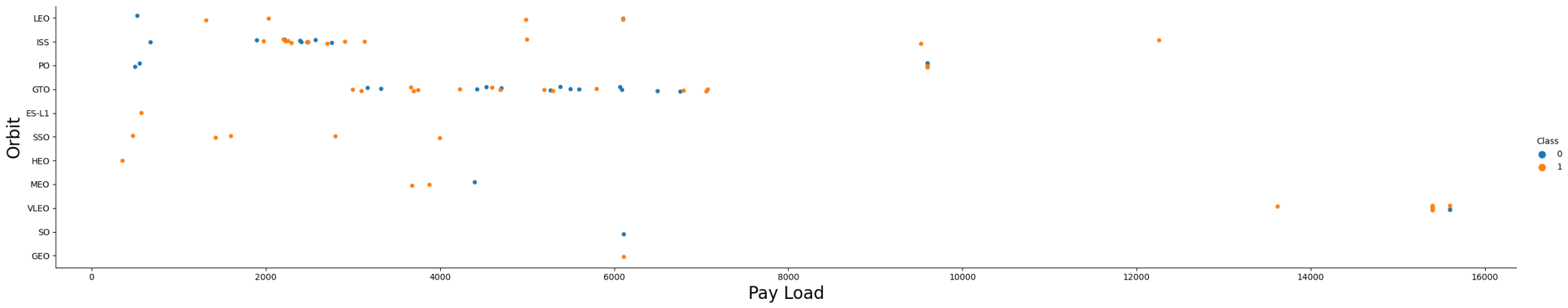
- Four orbits, ES-L1, GEO, HEO, and SSO, have a success rate close to 100%, while SO orbit shows zero success rate.

Flight Number vs. Orbit Type



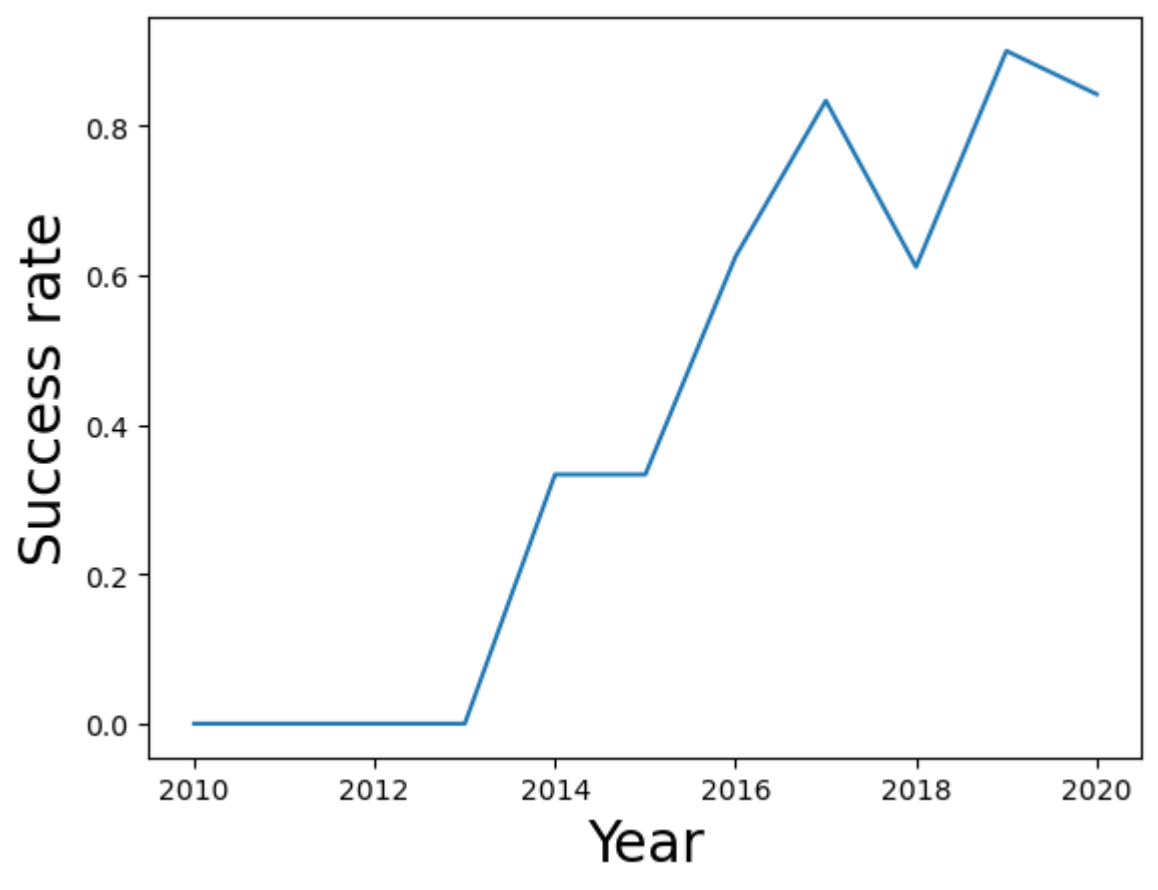
- The LEO orbit, located on top of the plot, shows the correlation between the success rate and the number of flights, while there are no clear patterns for other orbits with the number of flight.

Payload vs. Orbit Type



- The orbit SSO shows that the payload does not affect the successful landing.
- The successful landing of orbit GTO is negatively affected by the payload, while the payload has a positive effect on the successful landing of orbit LEO.

Launch Success Yearly Trend



- As seen, it was found that the success rate has increased since 2013 until 2020.

All Launch Site Names

```
In [5]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
 * ibm_db_sa://shb33137:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[5]:

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

- The SELECT DISTINCT statement is used to find the names of the unique launch sites

Launch Site Names Begin with 'CCA'

```
In [6]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE '%CCA%' LIMIT 5
```

```
 * ibm_db_sa://shb33137:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[6]:

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

- 5 records where launch sites begin with the string 'CCA' were displayed using the statement above.

Total Payload Mass

```
In [9]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
 * ibm_db_sa://shb33137:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
Out[9]: total_payload_mass
        45596
```

- The total payload mass carried by boosters launched by NASA (CRS) was displayed using the statement above.

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [16]:
```
%sql SELECT AVG(payload_mass__kg_) AS AVG_PM FROM SPACEXTBL WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://shb33137:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[16]:

avg_pm
2928

- The average payload mass carried by booster version F9 v1.1was displayed using the statement above.

First Successful Ground Landing Date

```
In [22]: %sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_LANDING FROM SPACEXTBL WHERE landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://shb33137:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
Out[22]: first_successful_landing
         2015-12-22
```

- The date when the first successful landing outcome in ground pad was achieved was listed using the statement above.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [26]:
```
%sql SELECT booster_version FROM SPACEXTBL WHERE landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

```
 * ibm_db_sa://shb33137:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[26]:

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

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 was listed using the statement above.

Total Number of Successful and Failure Mission Outcomes

In [47]:
```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS Total FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://shb33137:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[47]:

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

- The total number of successful and failure mission outcomes was listed using the statement above.

Boosters Carried Maximum Payload

```
In [51]: %sql SELECT booster_version, payload_mass__kg_ FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL DATE

 * ibm_db_sa://shb33137:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[51]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- The names of the booster_versions which have carried the maximum payload mass. was listed using the statement above.

2015 Launch Records

```
In [54]: %sql SELECT landing__outcome, booster_version, launch_site FROM SPACEXTBL WHERE landing__outcome LIKE '%Failure%' and YEAR(DATE) = 2015
 * ibm_db_sa://shb33137:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[54]:

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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 was listed using the statement above.

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

```
In [81]: %sql SELECT landing__outcome, COUNT(landing__outcome) AS TOTAL FROM SPACEXTBL  WHERE DATE BETWEEN '2010-06-04' AND
 * ibm_db_sa://shb33137:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3042
6/bludb
Done.
```

Out[81]:

landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

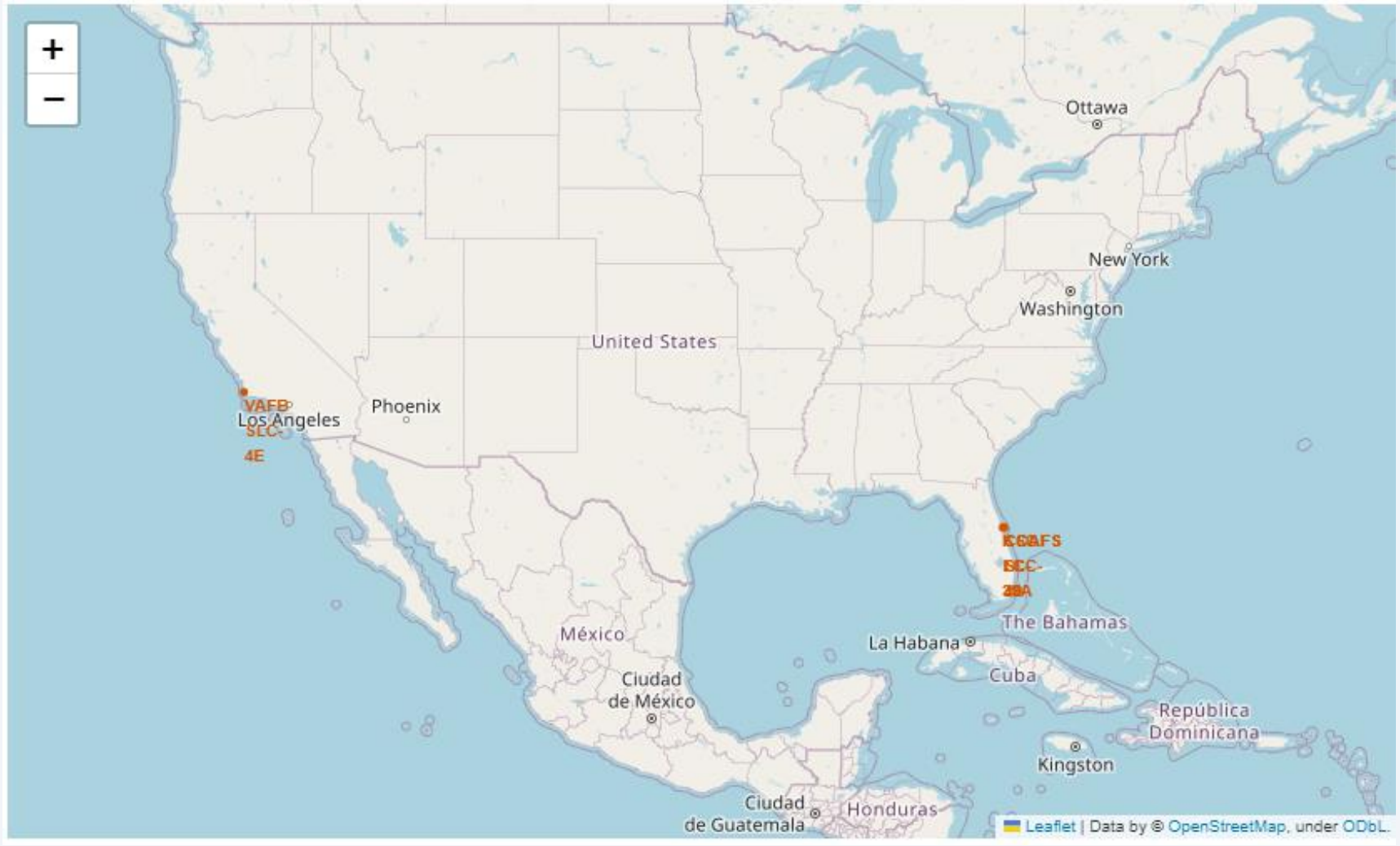
%sql SELECT landing__outcome, COUNT(landing__outcome) AS TOTAL FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing__outcome ORDER BY TOTAL DESC

- Using the statement above, we ranked 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

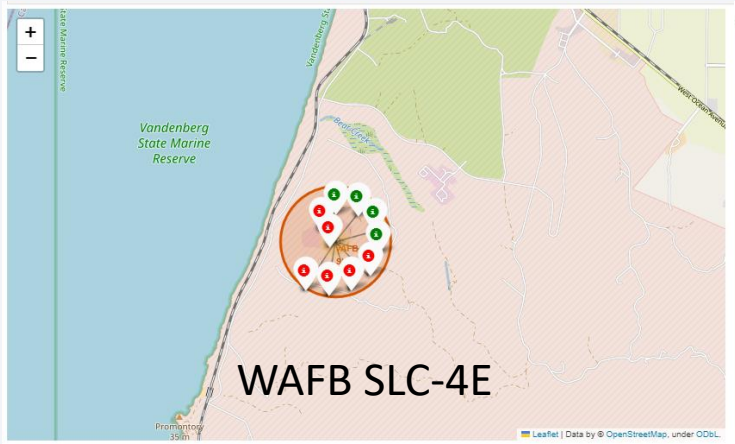
Section 3

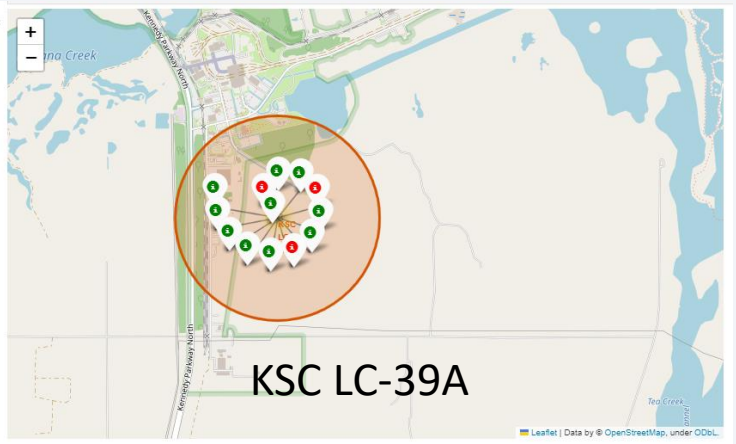
Launch Sites Proximities Analysis

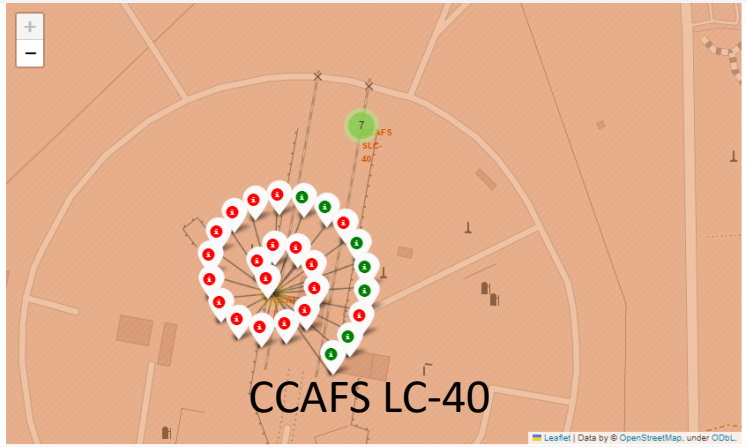
All launch site location with global map markers

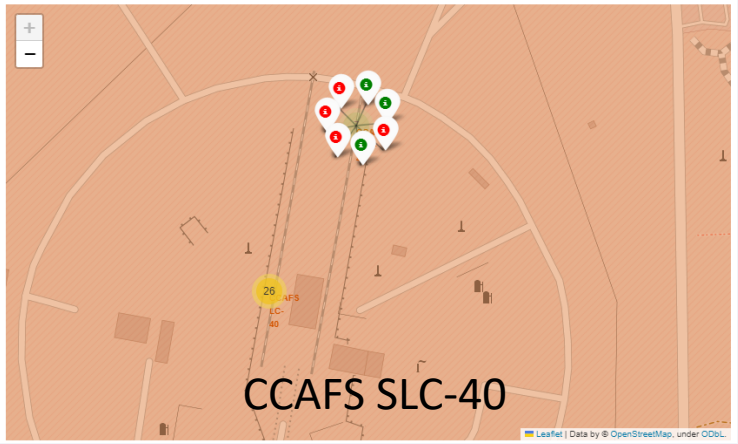


Color-labeled markers to indicate launch outcomes



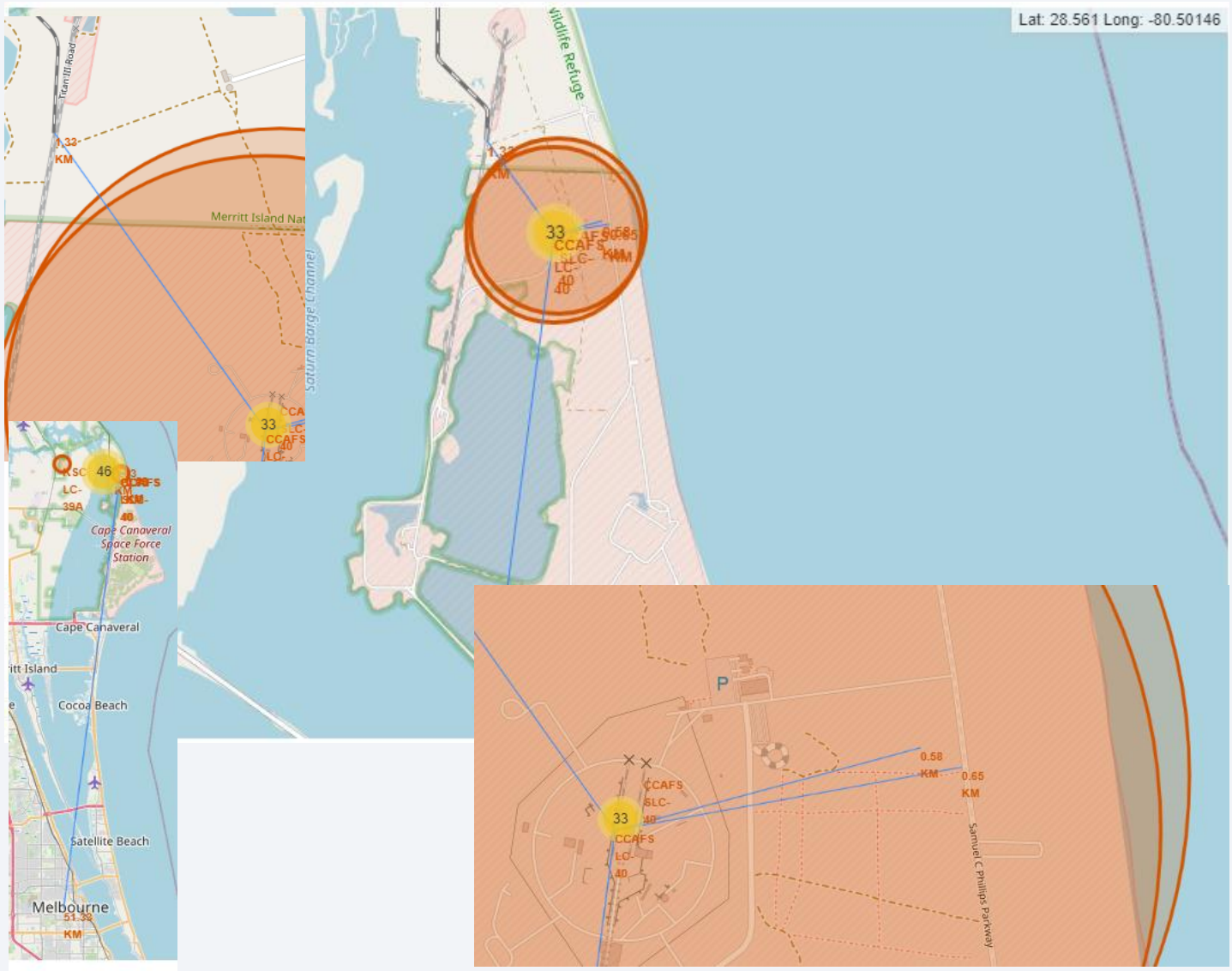




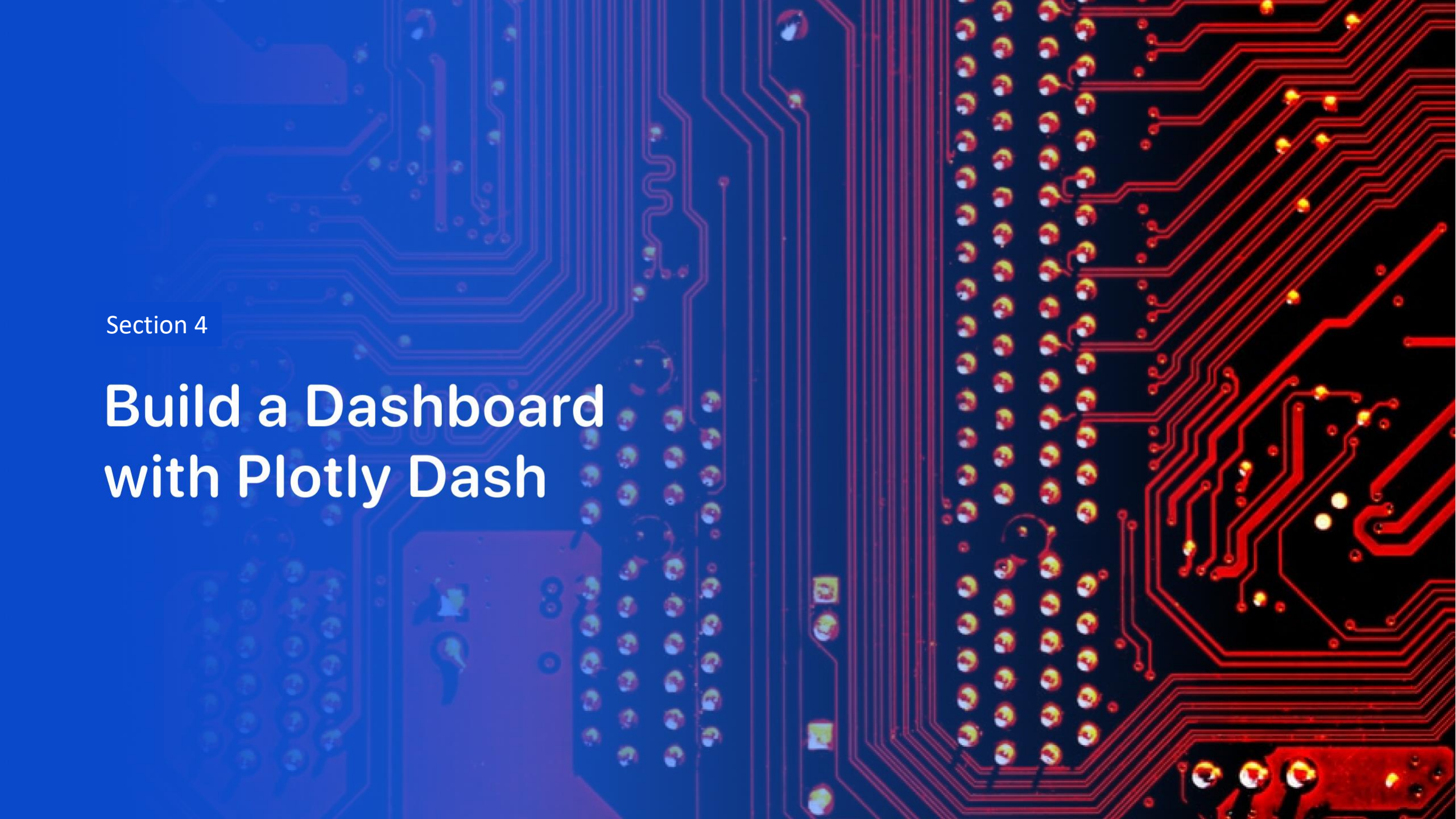


- Green Marker: Successful launch
- Red Marker: Failure launch
- High Success Rate: KSC LC-39A
- Low Success Rate: CCAFS LC-40

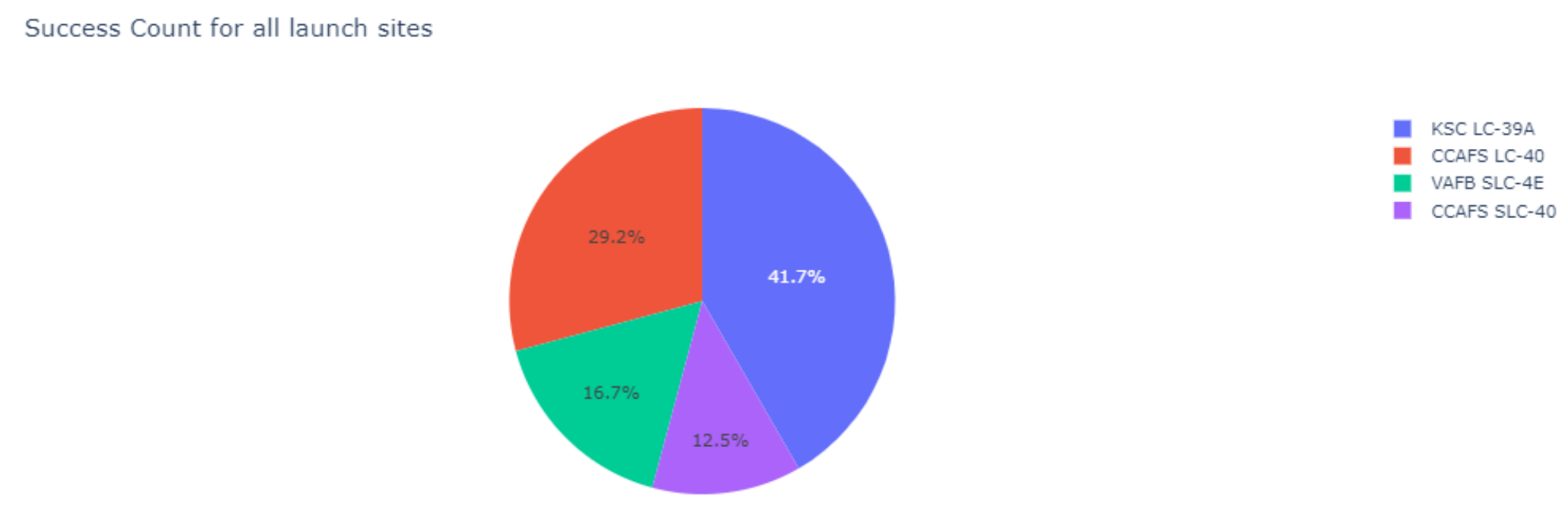
Distance from the launch site to its proximities



- Distance lines to the proximities were plotted to check if launch sites are in close proximity to railways, highways, coastline, and cities.

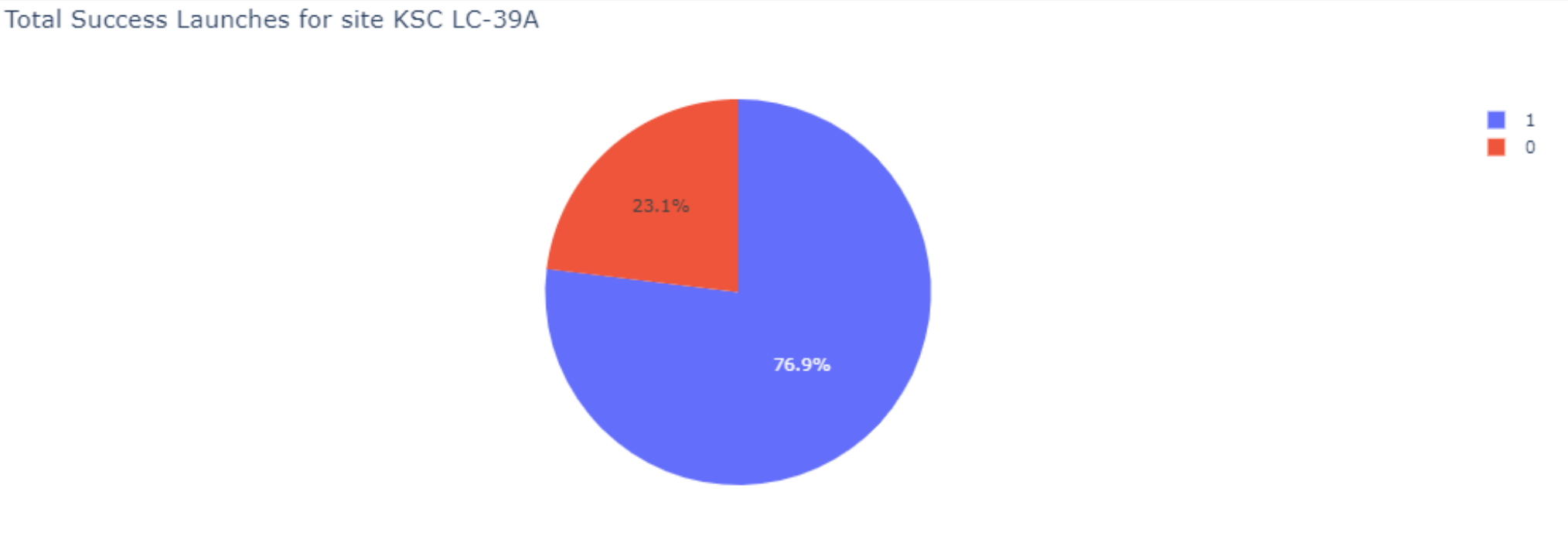

Section 4
Build a Dashboard with Plotly Dash

Pie chart – Success count for all launch site



- KSC LC-39A has the most successful launches among the four launch sites.

Pie chart – Success launches for KSC LC-39A site



- KSC LC-39A achieved 76.9% of successful launches and 23.1 % of failed launches

Scatter plot – Payload Mass vs. Launch Outcome

- The success rate was found to be high, between ~2000 kg and ~5500 kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Method: .score(X_test, Y_test)

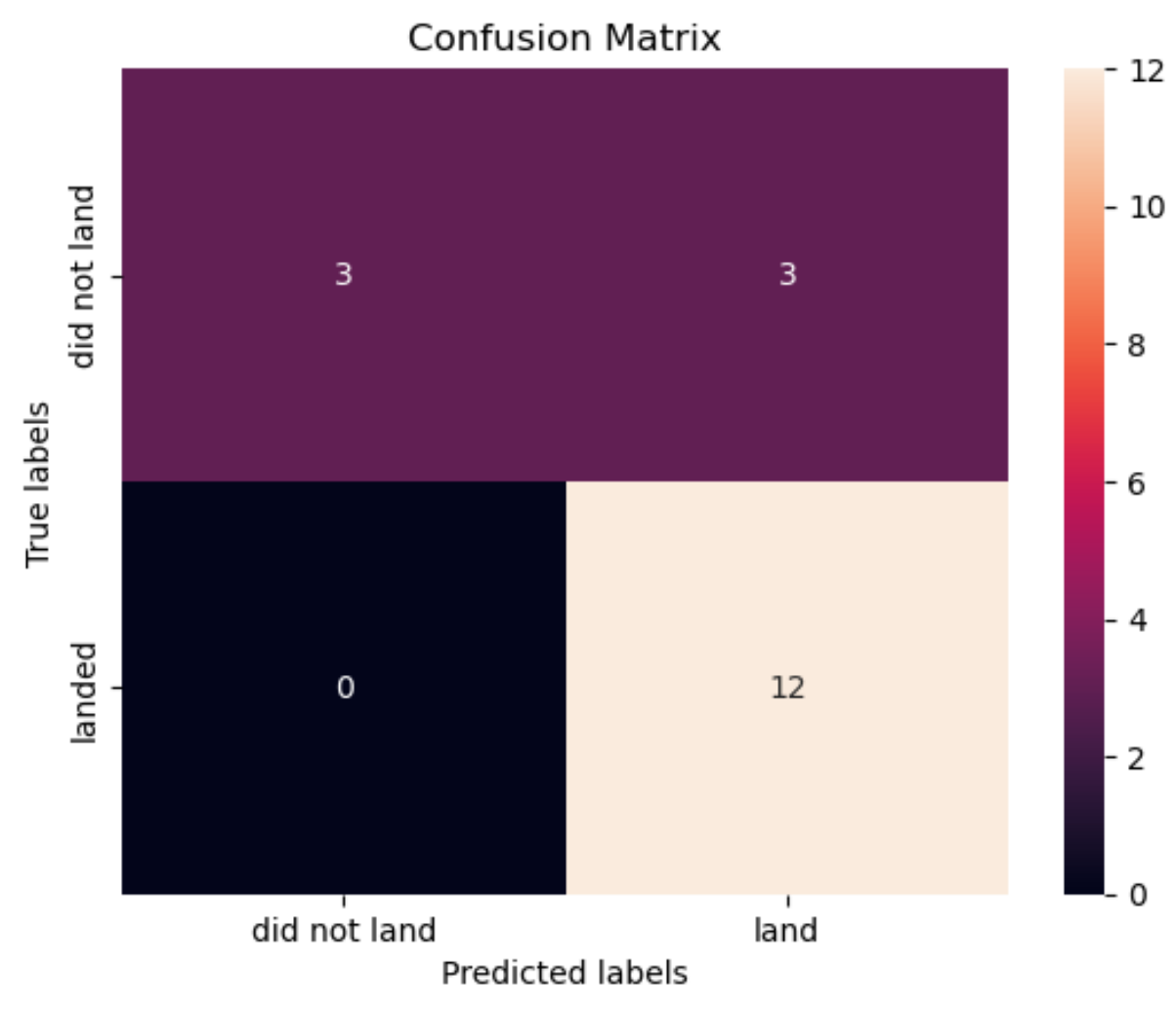
	ML Method	Accuracy
0	Logistic Regression	0.833333
1	Support Vector Machine	0.833333
2	Decision Tree	0.833333
3	K Nearest Neighbour	0.833333

Method: .best_score_

	ML Method	Accuracy
0	Logistic Regression	0.846429
1	Support Vector Machine	0.848214
2	Decision Tree	0.885714
3	K Nearest Neighbour	0.848214

- Calculate the accuracy using two methods, .score() and .best_score_
- It was found that all ML models have comparable accuracy, but the decision tree approach has the highest classification accuracy.

Confusion Matrix



- The confusion matrix for the decision tree classifier is shown, where the classifier can distinguish between the different classes.
- False positives were found in this classifier, which is the major problem.

Conclusions

- As the flight number increases at a launch site, the success rate increases.
- The low-weighted payloads show better performance than the heavy payload.
- The success rate for SpaceX increases from the year 2013 to 2020.
- KSC LC-39A shows the most successful launches from all the sites (76.9%)
- Four orbits, ES-L1, GEO, HEO, and SSO, have a success rate close to 100%.
- The decision tree classifier is the best machine-learning approach for this dataset.

Thank you!