IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Colleen Cowgill
8/20/25

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary

Summary of Analytic Approaches

- Data collection via API and web scraping
- Data wrangling with Python to facilitate EDA
- Exploratory Data Analysis (EDA) with SQL
- EDA with visualizations
  - Interactive visual analytics mapping with Folium
  - Interactive dashboard with Plotly Dash
- Comparison of machine learning classification models and prediction accuracy

Summary of Findings

- Success rates of landings for all sites generally improved over time across attributes
- High payload masses appeared to be related to fewer landing failures, although there were also fewer instances for comparison
- Launch site KSC LC-39 and Orbit Types ES-L1, SSO, and GEO demonstrated the highest success rates
- A decision tree classification machine learning model performed the best compared to other models (such as KNN) according to confusion matrices

Introduction

Project Background: According to the SpaceX website, the cost of a Falcon 9 rocket launch is approximately $62 million, whereas comparable launch services from other providers exceed $165 million per mission. The principal source of this cost savings is SpaceX's ability to reuse the rocket's first stage, reducing overall expenditures. Consequently, determining whether the first stage is successfully recovered is a key indicator of launch cost. This information could help alternative providers seeking to develop competitive bids against SpaceX in the commercial launch market.

Questions to Explore:

- What is the relationship between available variables and landing outcomes?
- What are the variables that most influence the outcome of a landing?
- What machine learning approach can we use to best predict landing outcomes?

Section 1

Methodology

Methodology

Collected data:

- Retrieved and consolidated data with SpaceX API
- Collected supplementary data with web scraping (Wikipedia)

Wrangled data

- Performed data wrangling, examining initial patterns and determining labels for training supervised models by examining orbits and outcomes

Performed exploratory data analysis (EDA)

- Examined variable relationships and derived descriptive statistics from aggregated data using SQL queries

Implemented interactive visual analytics

- Generated bar charts and scatterplots to examine attribute relationships and outcomes

Used Folium and Plotly Dash to create interactive maps and dashboards

- Mapped all launch sites and calculated distances to relevant geographic locations, enabling interactive data exploration
- Marked successes and failures via color-coded map indicators

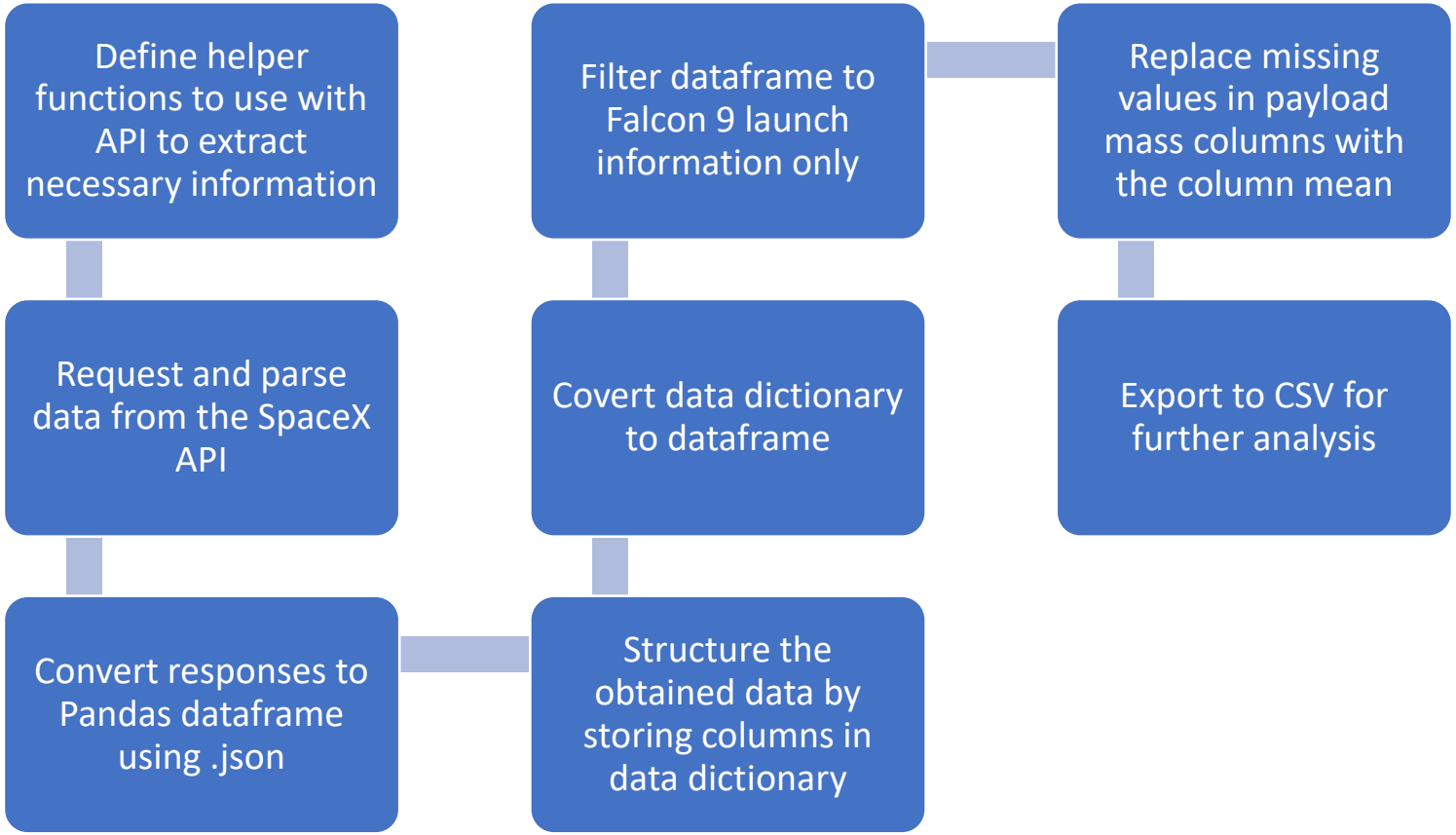
Conducted predictive analysis

- Used classification techniques to perform predictive analysis
- Built, evaluated, and compared multiple predictive classification models

Data Collection

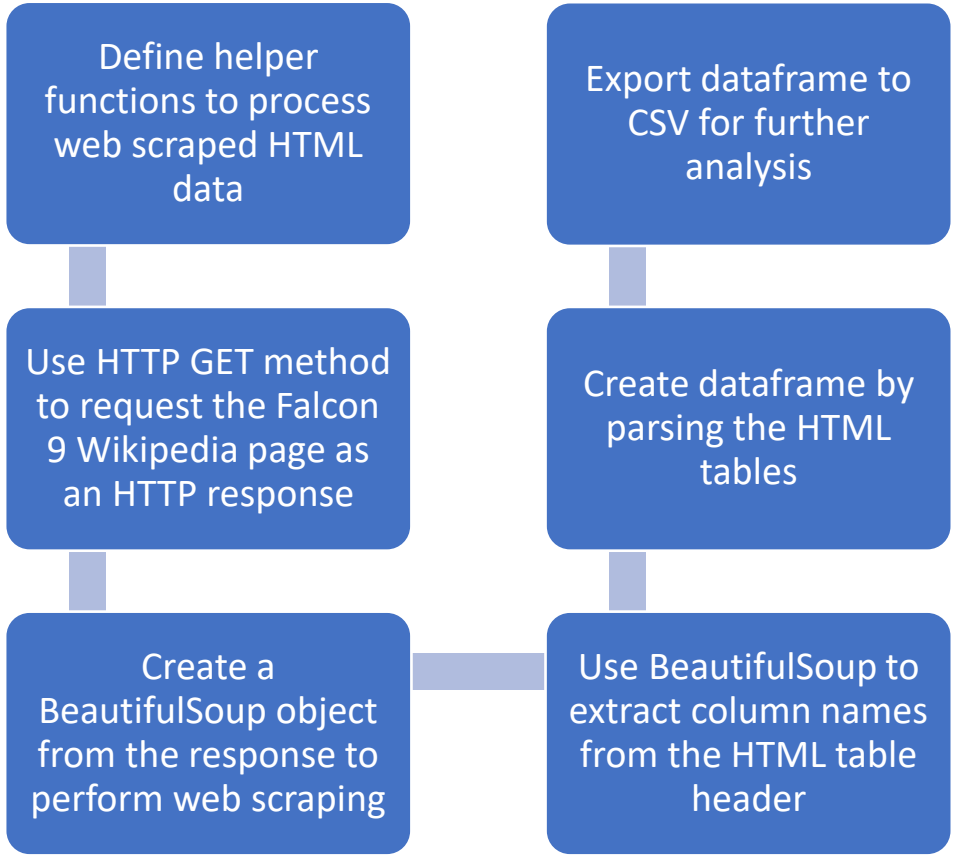
- **Methodology:** Data for various launch attributes was collected using a combination of API requests from the SpaceX API and web scraping SpaceX's Wikipedia entry. Using both methods yielded more complete information about rocket launches for comprehensive analyses.
- **Some data attributes derived with SpaceX API:** Booster version, payload mass, orbit type, launch site longitude and latitude, launch outcome, flight numbers, whether grid fins were used, how many launches reused first stages, the core serial number, etc.
- **Some data attributes derived with web scraping Wikipedia**: Booster version, flight number, orbit type, launch outcome, booster landing outcome, type of payload, payload mass, customer name, date and time of launches, etc.

Data Collection – SpaceX API



- GitHub link to the completed SpaceX API calls notebook

Data Collection – Web Scraping



- GitHub link to the completed web scraping notebook

Data Wrangling

- **Context:** The dataset contained information demonstrating that boosters occasionally do not successfully land due to an accident. For example, the dataset attribute “True Ocean” means a successful landing in a specific region of the ocean while “False Ocean” means there was not a successful landing in a specific region of the ocean.
- **Data Wrangling:** The methodology involved converting these outcomes into Training Labels with “1” indicating the booster successfully landed and “0” indicating the booster did not land successfully.
- GitHub link to completed data wrangling notebook

EDA with Data Visualization

Charts created for EDA Visualizations

- **Bar Chart:** Success Rate vs. Orbit Type
- **Scatterplot:** Flight Number vs. Launch Site on Succes Rate
- **Scatterplot:** Payload vs. Launch Site on Success Rate
- **Scatterplot:** Flight Number vs. Orbit Type on Success Rate
- **Scatterplot:** Payload vs Orbit Type on Success Rate
- **Line Chart:** Landing Success Rate Yearly Trend

Scatter plots demonstrate the relationship among attributes, which could indicate what features to include in machine learning models if strong relationships are suspected.

Bar charts display comparisons among categorical variables on some value.

Line charts depicted any trends in data over time

GitHub link to completed EDA with data visualization notebook

EDA with SQL Queries

SQL queries were scripted and executed to display the following information:

- Launch site names
- 5 records of launch sites beginning with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date of first successful landing outcome in ground pad
- Names of boosters with successful outcomes that have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed outcomes
- Names of the boosters that have carried the maximum payload masses
- Failed landing outcomes of various booster versions, launch sites, and their month of occurrence in year 2015
- Ranked count of different landing outcomes between the 2010-06-04 and 2017-03-20

GitHub link to completed EDA with SQL notebook

Build an Interactive Map with Folium

Locations of Launch Sites:

- Markers with popup and text labels of all launch sites using latitude and longitude coordinates to show their geographical locations

Markers of the Launch Outcomes :

- Markers of successful (green) and failed (red) launches for launch sites

Distances Between a Launch Site and a Key Location

- Line showing shortest distance between the Launch Site CCAFS-SLC 40 and the nearest coastline

GitHub link to completed interactive map with Folium map

Build a Dashboard with Plotly Dash

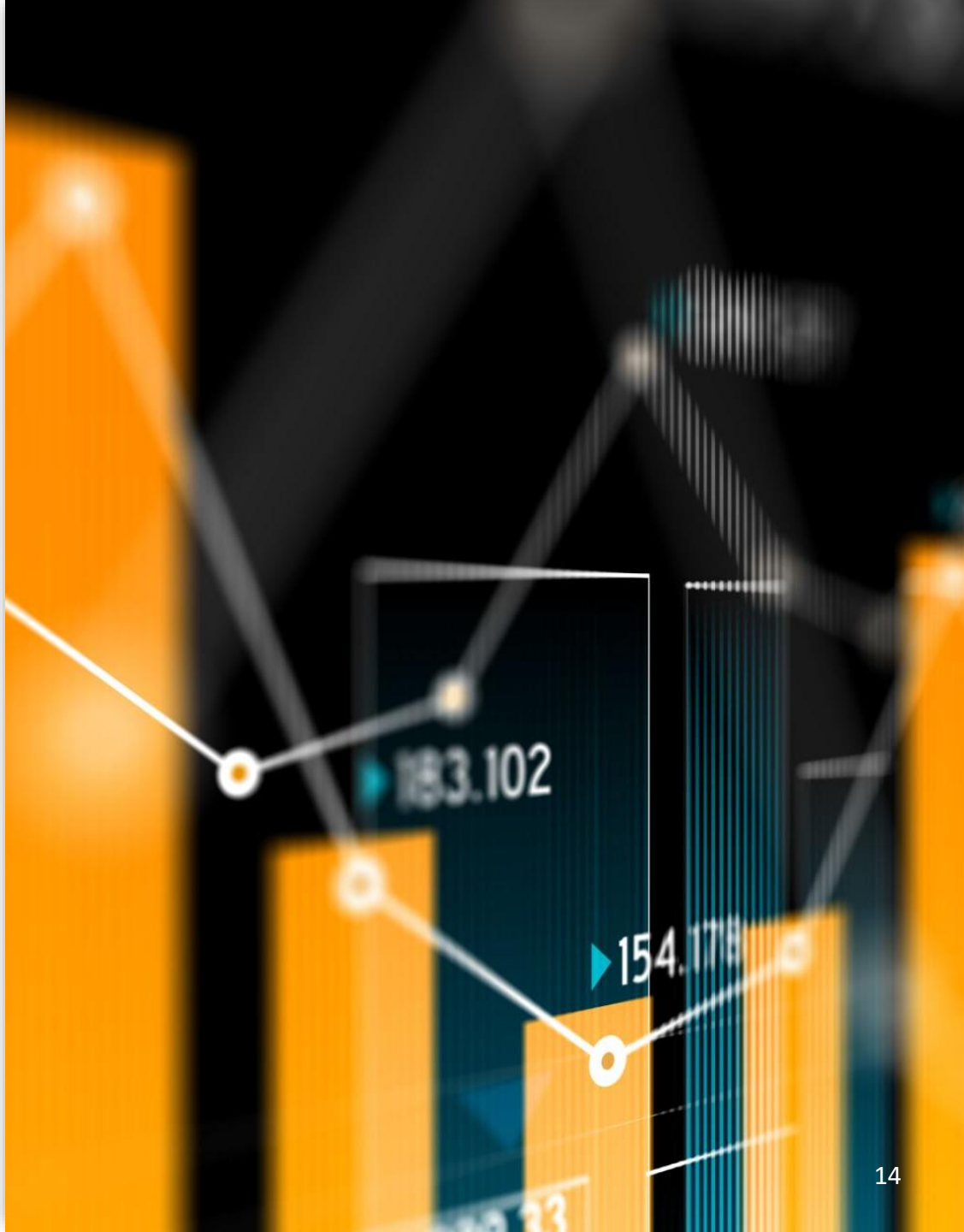
Launch Sites Dropdown: A dropdown list feature was added to the dashboard to enable viewing outcome results by specific launch site for comparison

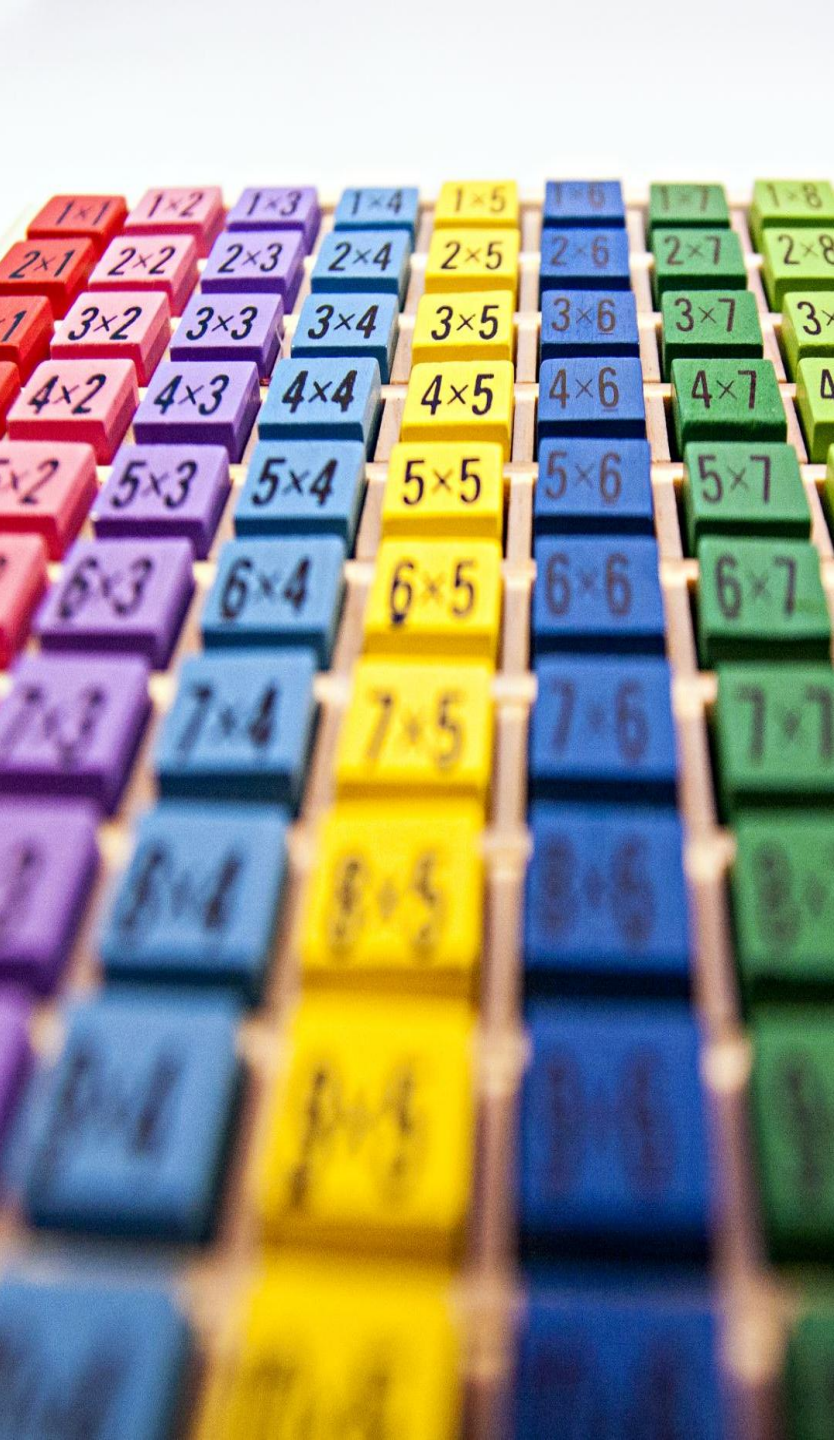
Interactive Pie Chart of Landing Failure/Success Proportions: The dashboard pie chart shows the total successful and failed landings proportion for all launch sites or for a specific site depending on the dropdown menu selection.

Slider for Payload Mass: A slide was added to the dashboard to allow a selection of certain payload mass ranges when viewing landing outcomes by booster version.

Scatterplot of Outcomes by Booster Version: The scatterplots show the relationship between booster version and landing outcomes, customizable by payload mass range as an additional attribute for comparison.

GitHub link to completed Plotly Dash code





Predictive Analysis (Classification Machine Learning Models)

- Created a NumPy array for landing outcomes and standardized the data
- Split data into training and test sets
- Created model objects and GridSearch objects to determine the best model parameters across each model techniques
- Fit the models to obtain basic performance metrics
- Compared metrics of classification methods to evaluate performance differences
- Created and compared confusion matrices for all models
- Compared accuracy metrics for all models
- Determined best performing classification technique using all available information

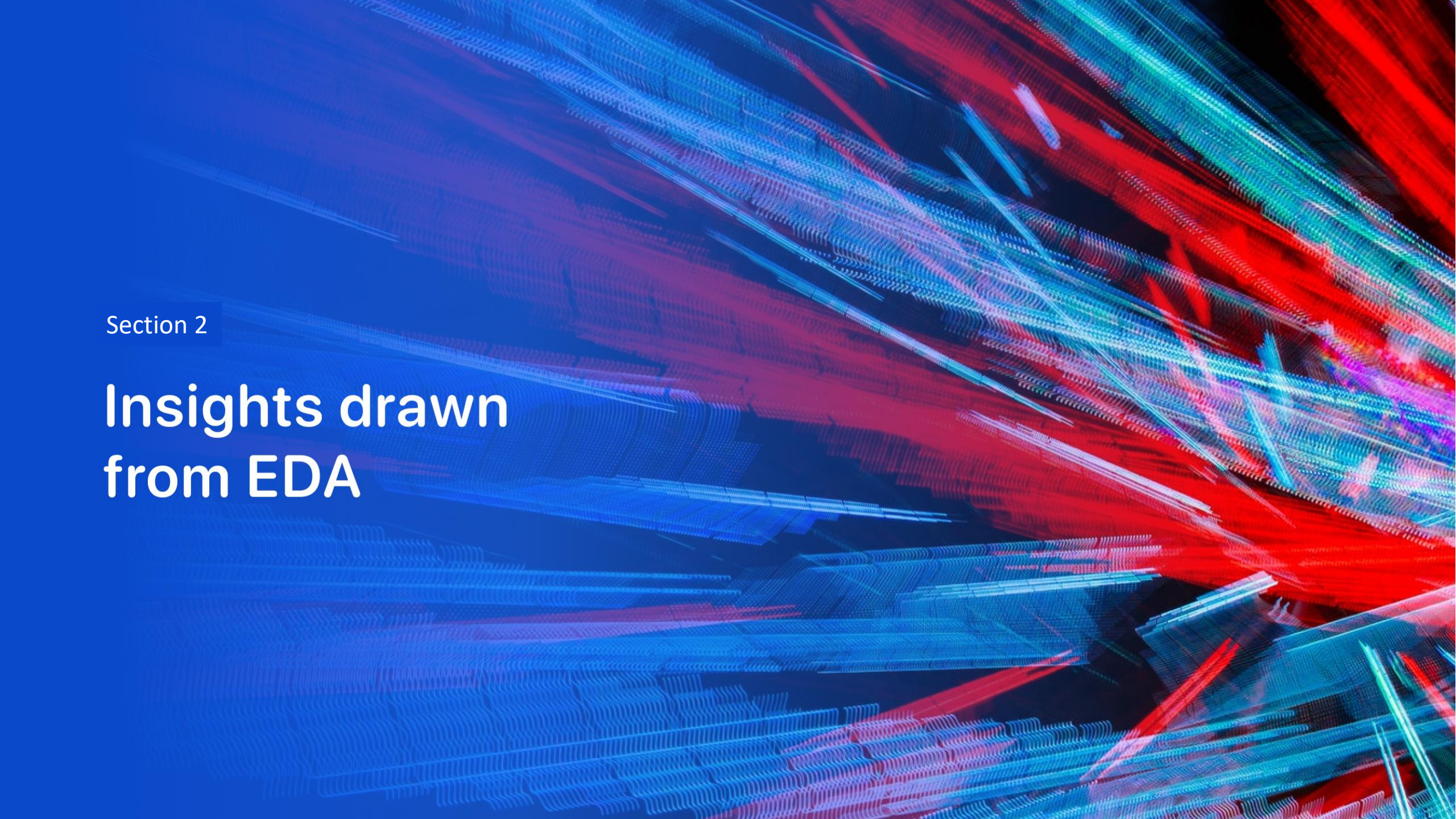
[GitHub link to completed predictive analysis lab]

Results Contents

EDA Using Visualizations and SQL Queries

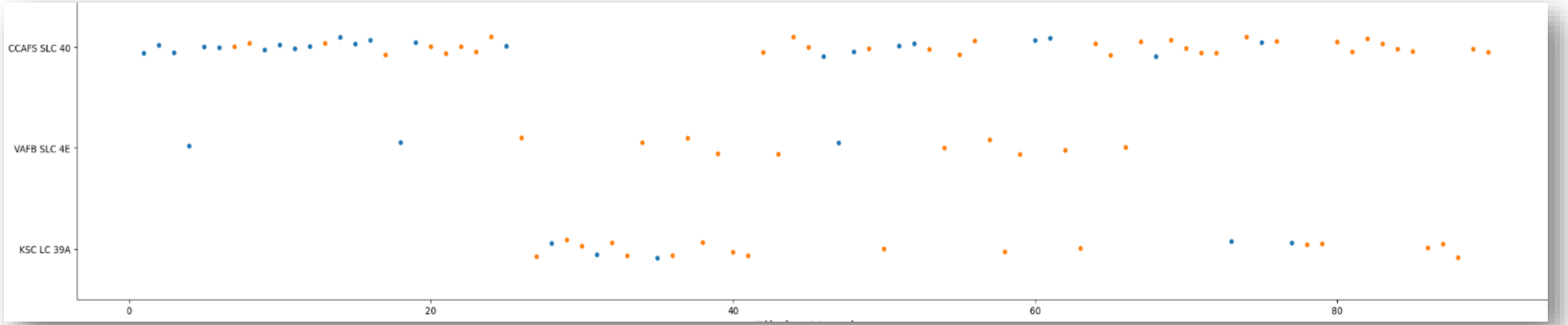
Screenshot Samples of Interactive Map and Dashboard Analytics

Predictive Analytics Comparing Classification Machine Learning Models


Section 2
Insights drawn from EDA

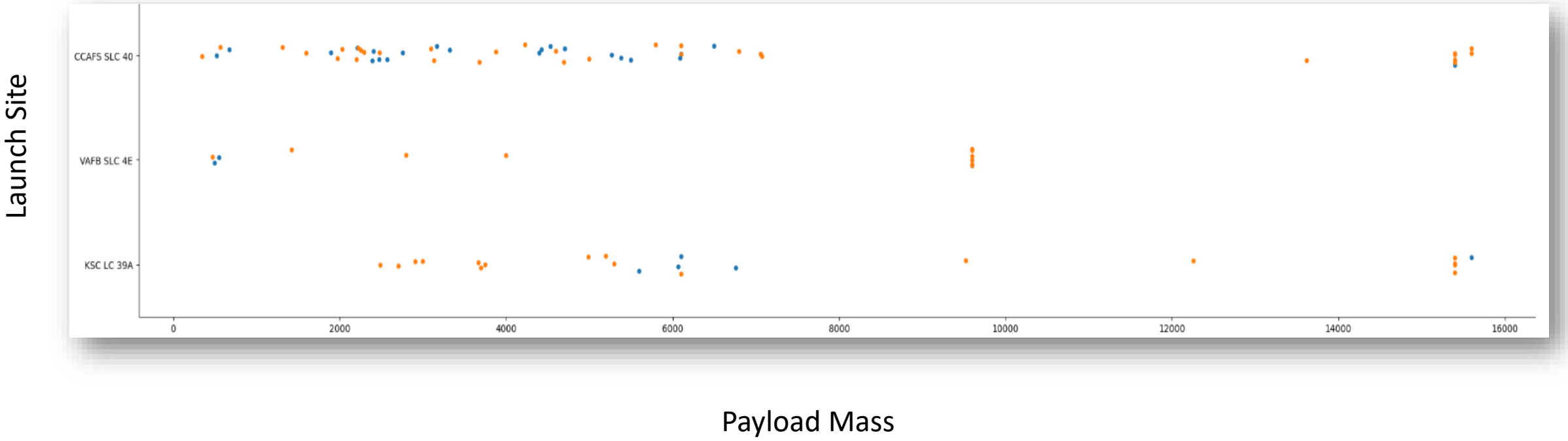
Flight Number vs. Launch Site

Below shows the scatterplot of landing successes by the flight number vs launch site. Successes are indicated by orange dots while failures are blue dots. Landing success across all launch sites appear to improve with increased flight number.



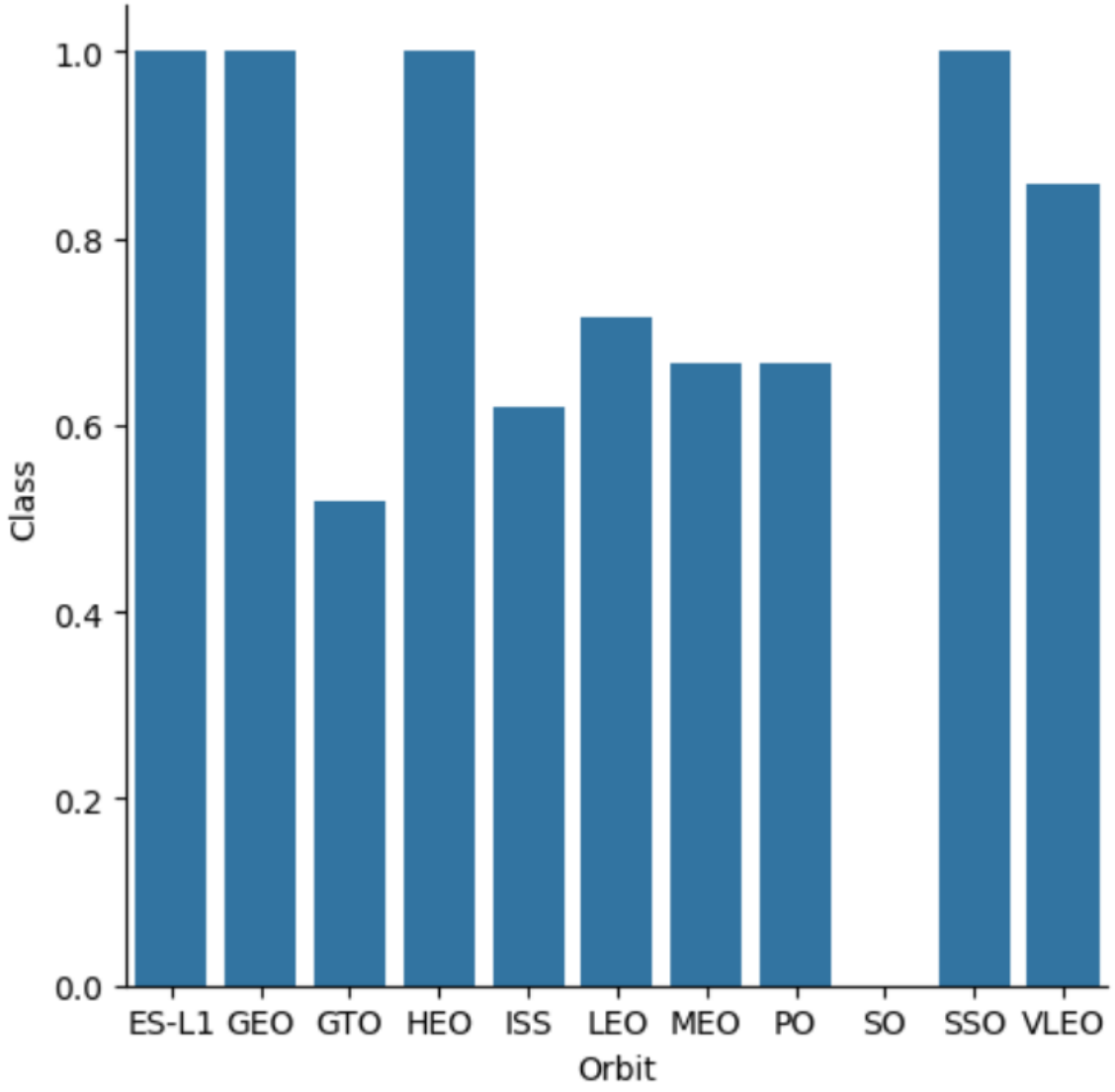
Payload vs. Launch Site

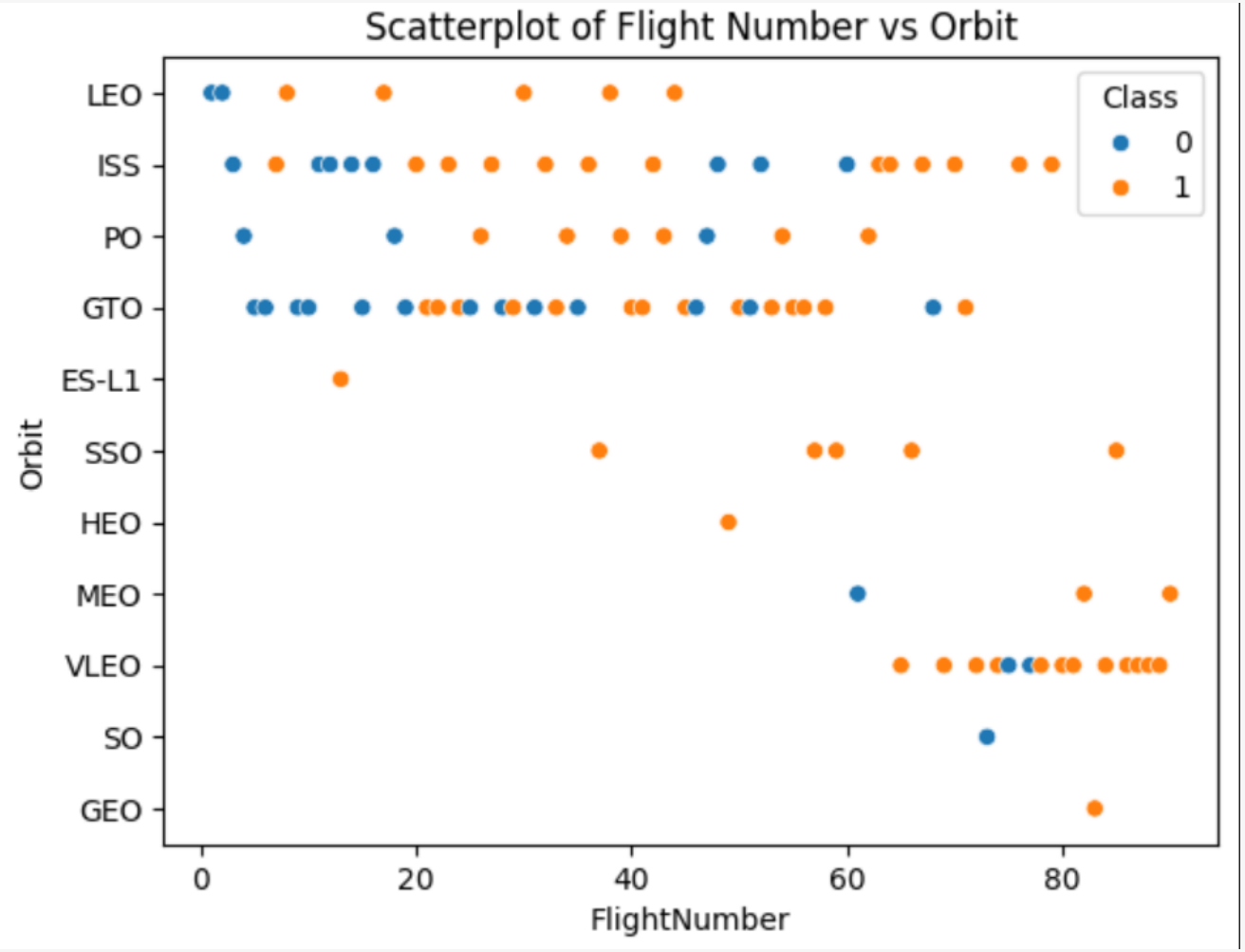
Below shows the scatterplot of landing successes by payload mass vs launch site. Successes are indicated by orange dots while failures are blue dots. Launch site VAFB does not show any launches over a payload mass of 10,000 kg, whereas CCFAS and KSC do. These two launch sites seem to have fewer failures at these very high payloads.



Success Rate vs. Orbit Type

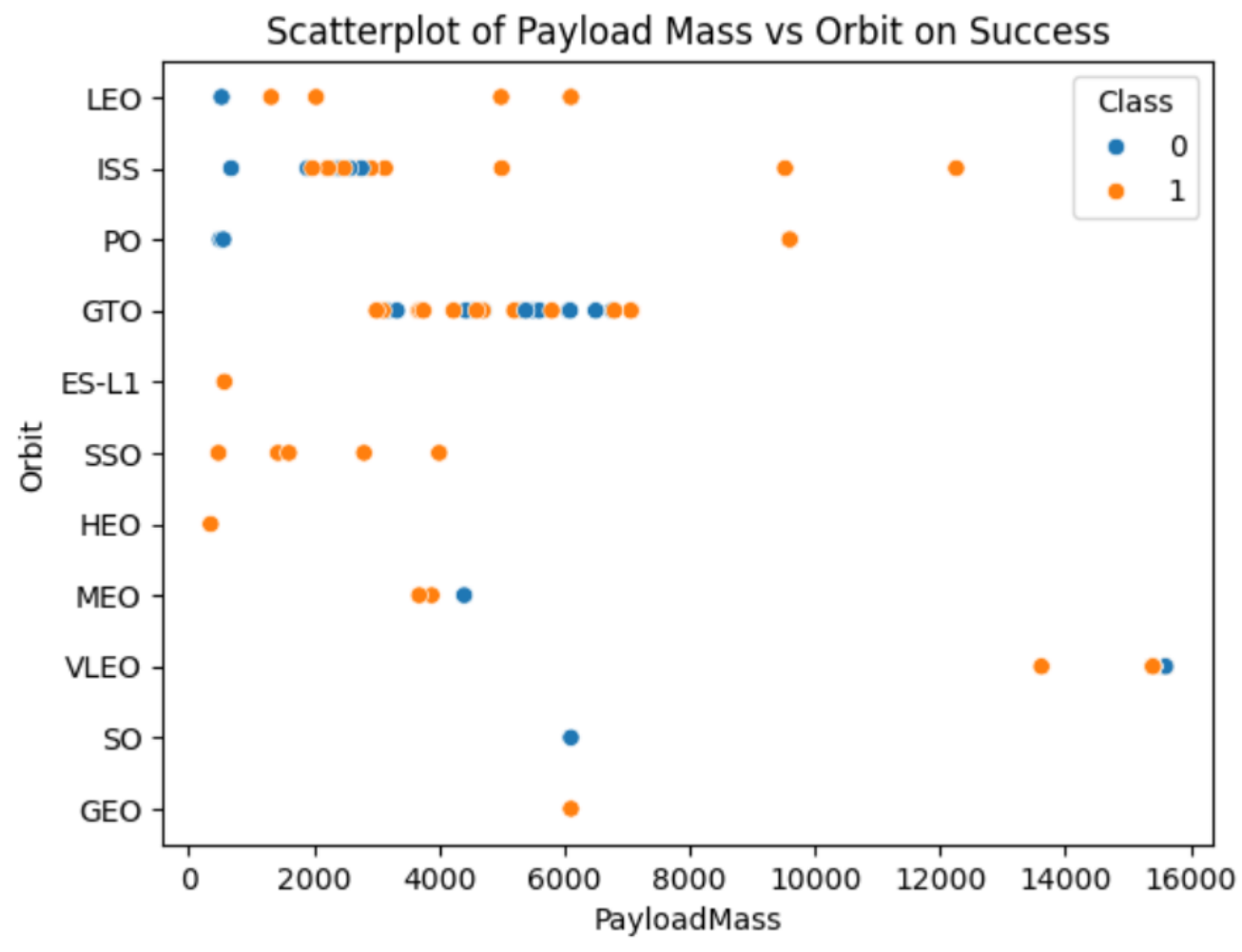
- To the left is a bar chart displaying the landing success rate by each orbit type.
- GTO has the lowest success rate, while ES-L1, GEO, HEO, and SSO have the highest launch success rates.
- The next graphs may provide more insight into the nature of these relationships.





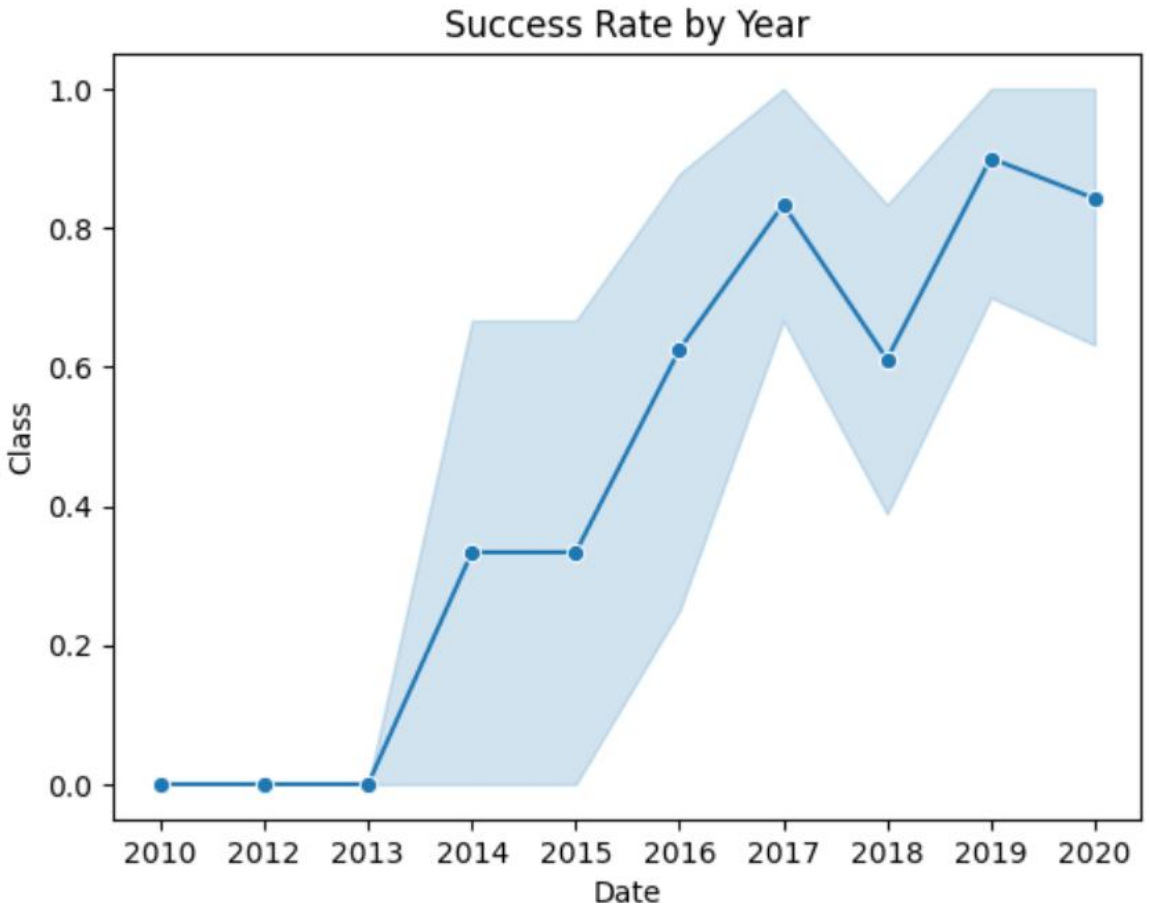
Flight Number vs. Orbit Type

- The left shows a scatter plot relating flight number and orbit type and indicating landing outcomes. Orange dots indicate successes while blue dots indicate failures.
- There appears to be a relationship between flight number and success rate, such that higher flight numbers have fewer failures. It appears that VLEO was used for later flights and has a higher success rate than, for example, GTO and ISS, which were used for many earlier flights.



Payload vs. Orbit Type

- The left shows a scatter plot of landing outcomes by rocket payload mass vs. orbit type. Orange dots indicate successes while blue dots indicate failures.
- Although the data is highly variable, there appear to be fewer landing failures at higher payload masses (although there are also fewer launches), and GTO appears to have a more unpredictable success rate. Once again, GTO shows the lowest success rate of all orbit types.



Launch Success Yearly Trend

- The left displays a line chart of yearly average success rate for launches, given by the “Class” attribute. Higher “Class” numbers indicate more landing successes.
- Clearly, the average success rate increases by year.

All Launch Site Names

[15]:
```
%%sql
SELECT DISTINCT "Launch_Site"
FROM SPACEXTABLE;
```

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

[15]:

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

SQL query to display all Launch Site names

Launch Site Names Begin with 'CCA'

[11]:
```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5;
```

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

[11]:

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

SQL query to display information for 5 Launch Sites beginning with the string ‘CCA’

Total Payload Mass

```
[27]: %%sql
      SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload
      FROM SPACEXTABLE
      WHERE Customer = 'NASA (CRS)';
```

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

[27]:

total_payload
45596

SQL query showing the total payload mass carried by launches for customer NASA

Result: 45,596 kg

Average Payload Mass by F9 v1.1

[29]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload
FROM SPACEXTABLE
WHERE "Booster_Version" = 'F9 v1.1';
```

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

[29]:

avg_payload
2928.4

SQL query showing the average payload mass carried by Booster Version F9 v1.1

Result: 2928.40 kg

First Successful Ground Landing Date

[30]:
```
%%sql
SELECT MIN("Date") AS mindate
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success';
```

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

[30]:

mindate
2018-07-22

SQL query to display the date of the earliest successful landing outcome

Result: 07/22/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

[33]:
```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (drone ship)'
  AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

[33]:

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

SQL query to display the Booster Versions carrying a payload between 4,000 kg and 6,000 kg with successful landings

Total Number of Successful and Failure Mission Outcomes

[35]:
```
%%sql

SELECT COUNT(*) AS sum_mission
FROM SPACEXTABLE
WHERE "Mission_Outcome" IN ('Success', 'Failure');
```

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

[35]:

sum_mission
98

SQL query giving the total number of mission successes and failures.

Result: 98

Boosters Carrying Maximum Payload

```
36]: %%sql
SELECT DISTINCT "Booster_Version"
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE
);
```

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

36]:

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

SQL query to show all Booster Versions that carried the maximum payload

2015 Launch Records

[37]:
```
%%sql

SELECT
     substr(Date, 6, 2) AS Month,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE substr(Date, 1, 4) = '2015'
  AND "Landing_Outcome" LIKE 'Failure (drone ship)%';
```

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

[37]:

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

SQL query showing the month number, Booster Version, and Launch Site of failed landings in 2015

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

[38]:
```
%%sql
SELECT
    "Landing_Outcome",
    COUNT(*) AS outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY outcome_Count DESC;
```

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

[38]:

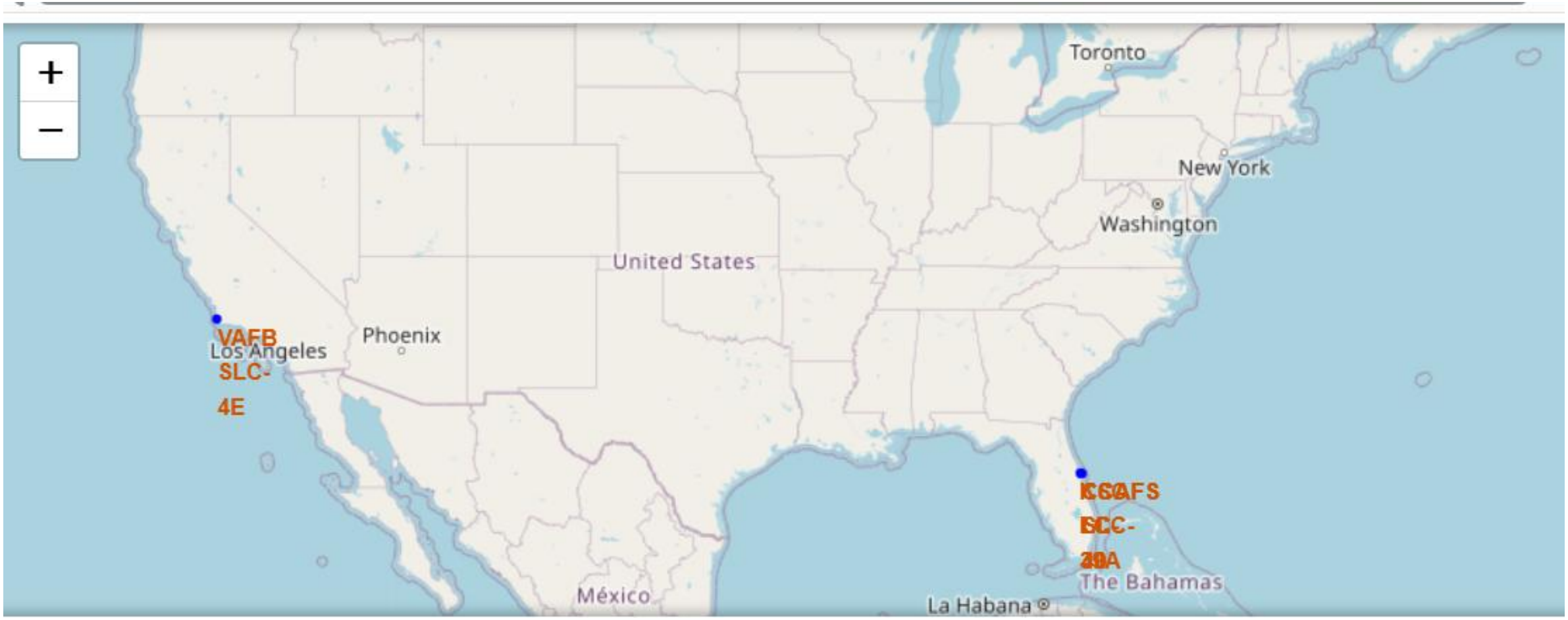
Landing_Outcome	outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

SQL query to display the numbers of types of landings outcomes between 6/4/2010 and 3/20/2017, ranked

Section 3

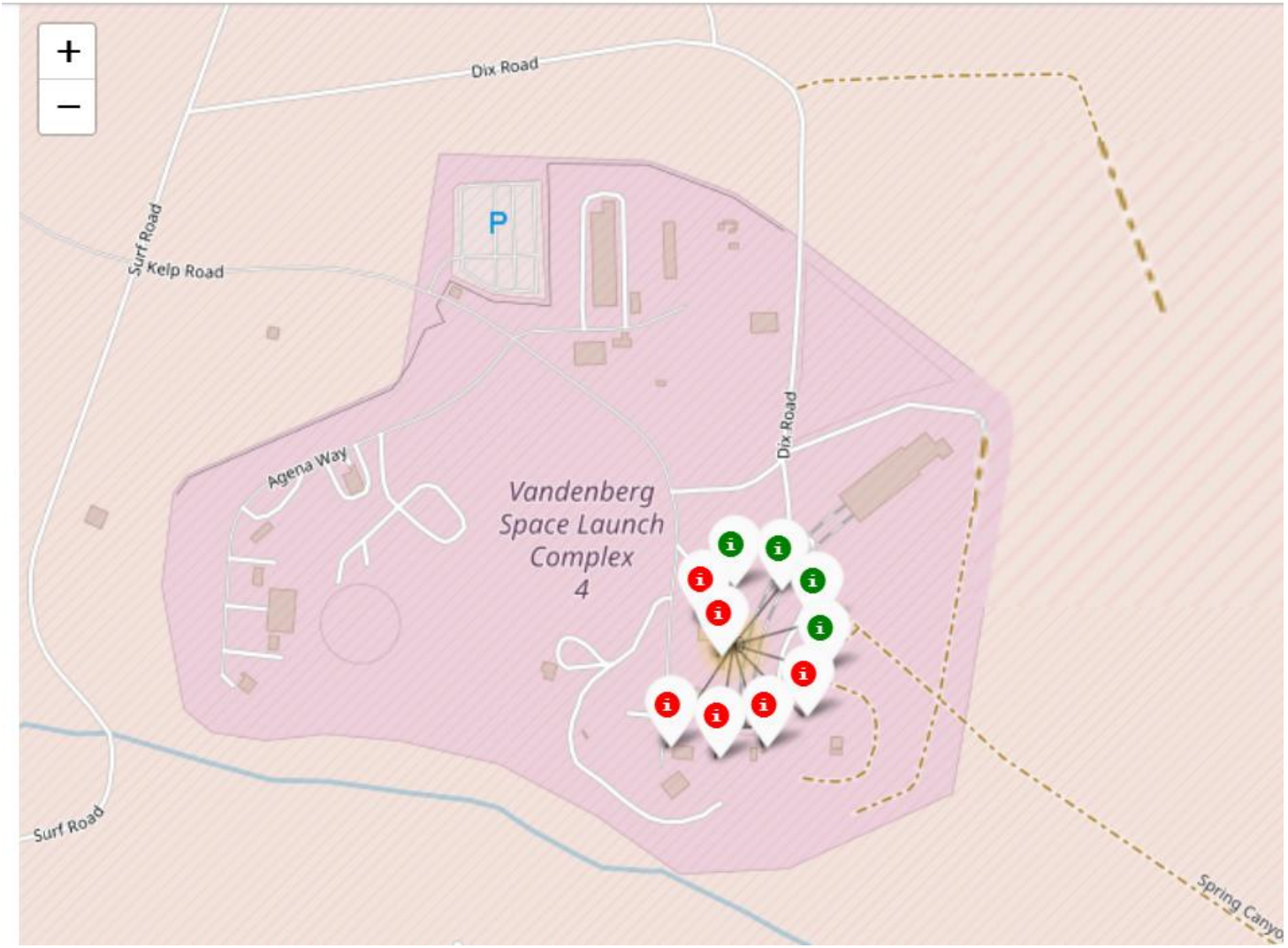
Launch Sites Proximities Analysis

Full Map of all SpaceX Launch Sites



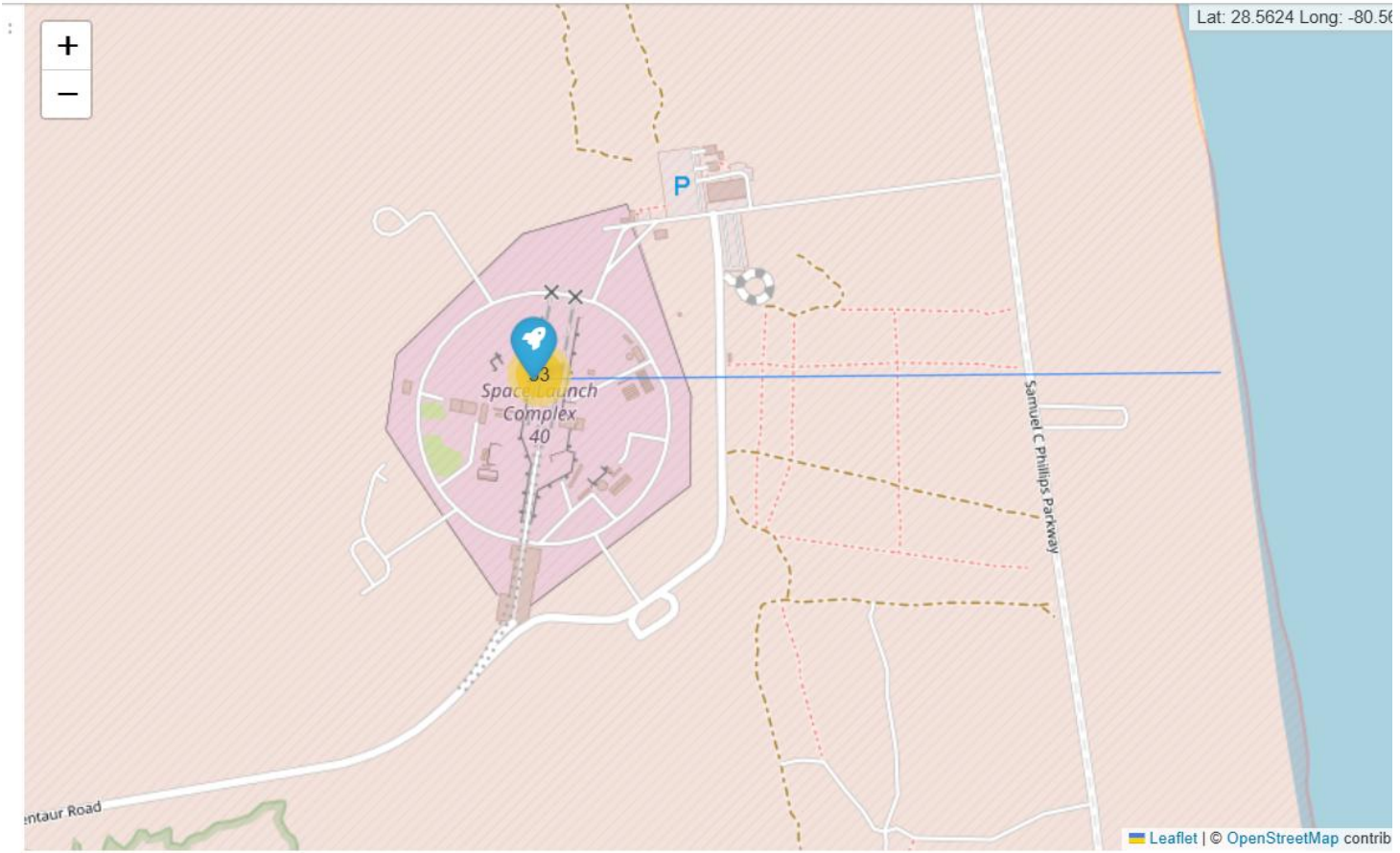
Folium Map Screenshot of Color-Coded Launch Outcomes

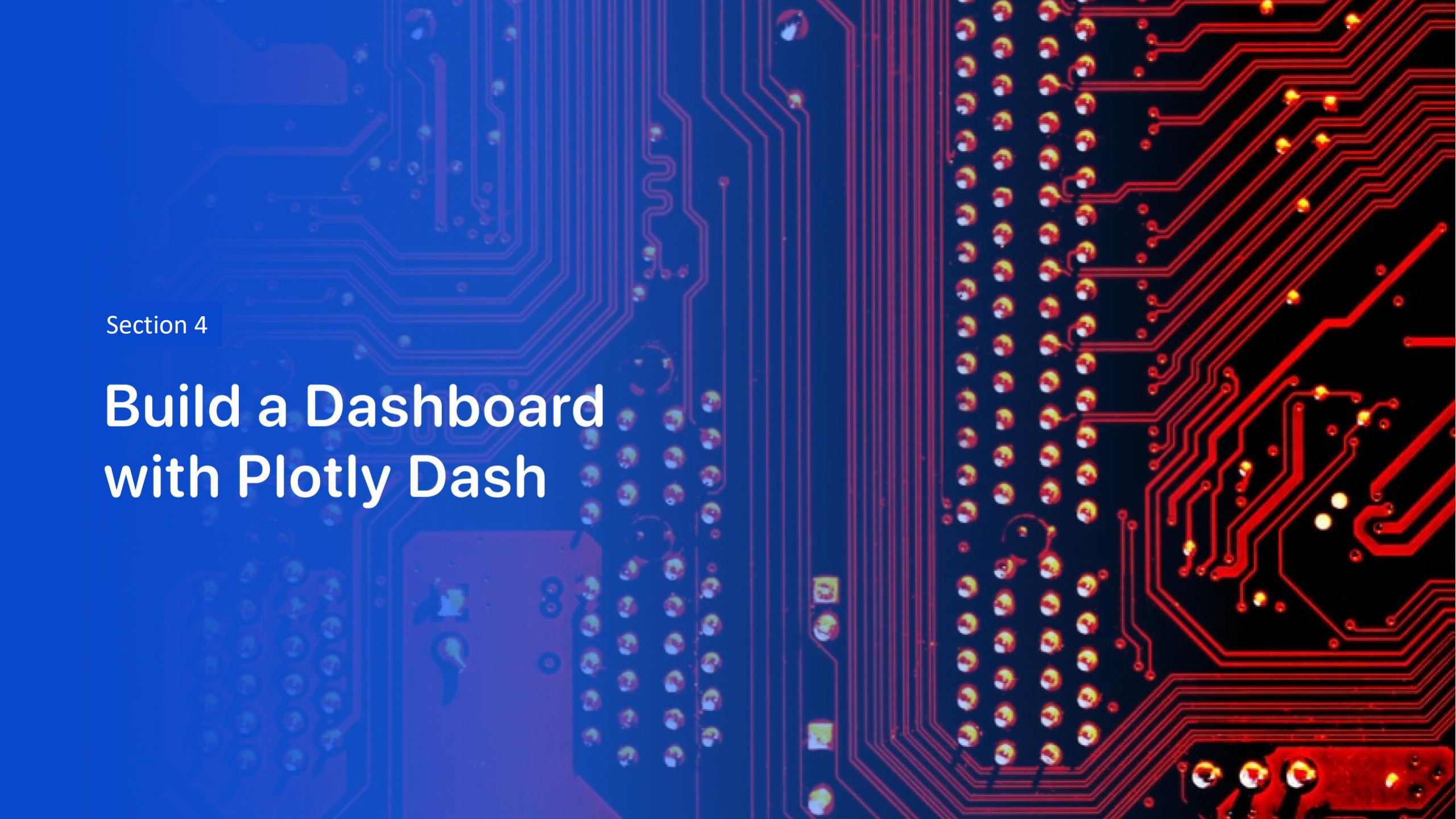
- The right shows an example of the Vandenberg Space Launch Complex 4 launch site with map markers colored green to indicate successful landings and red for failed landings.
- The Folium map shows that this launch site has comparatively high failure rate



Folium Map of Launch Site Distance to Ocean

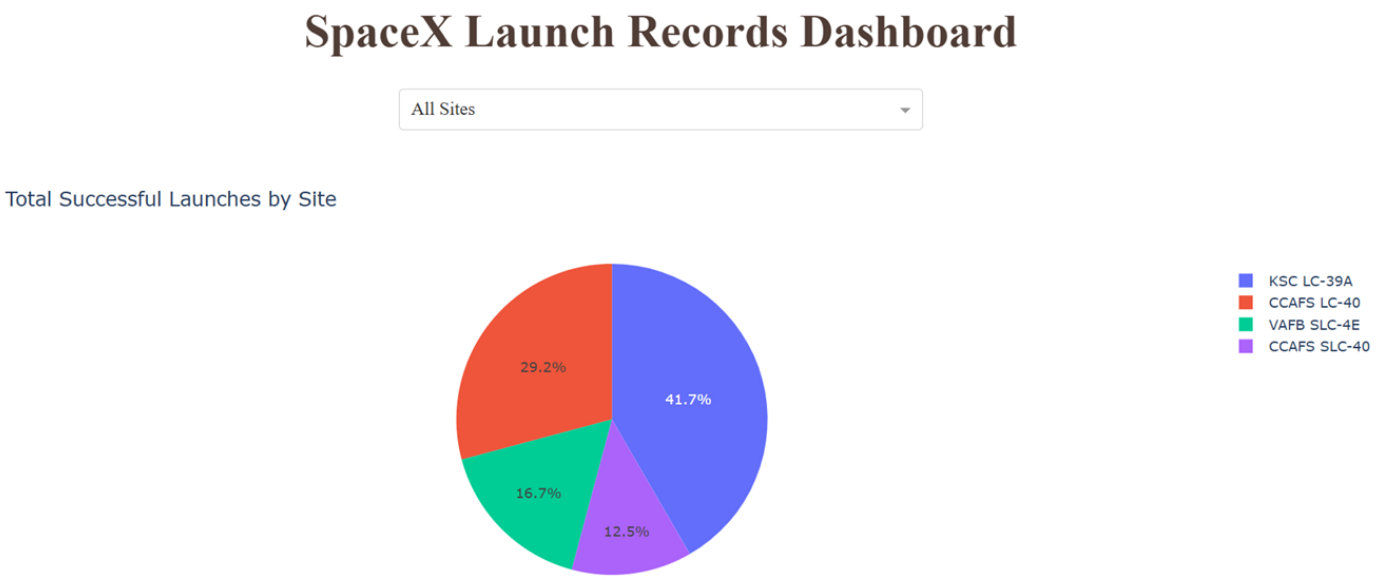
- The added Folium map feature shows a direct line from the launch site to a geographic feature of interest (in this case, the coastline).




Section 4
Build a Dashboard with Plotly Dash

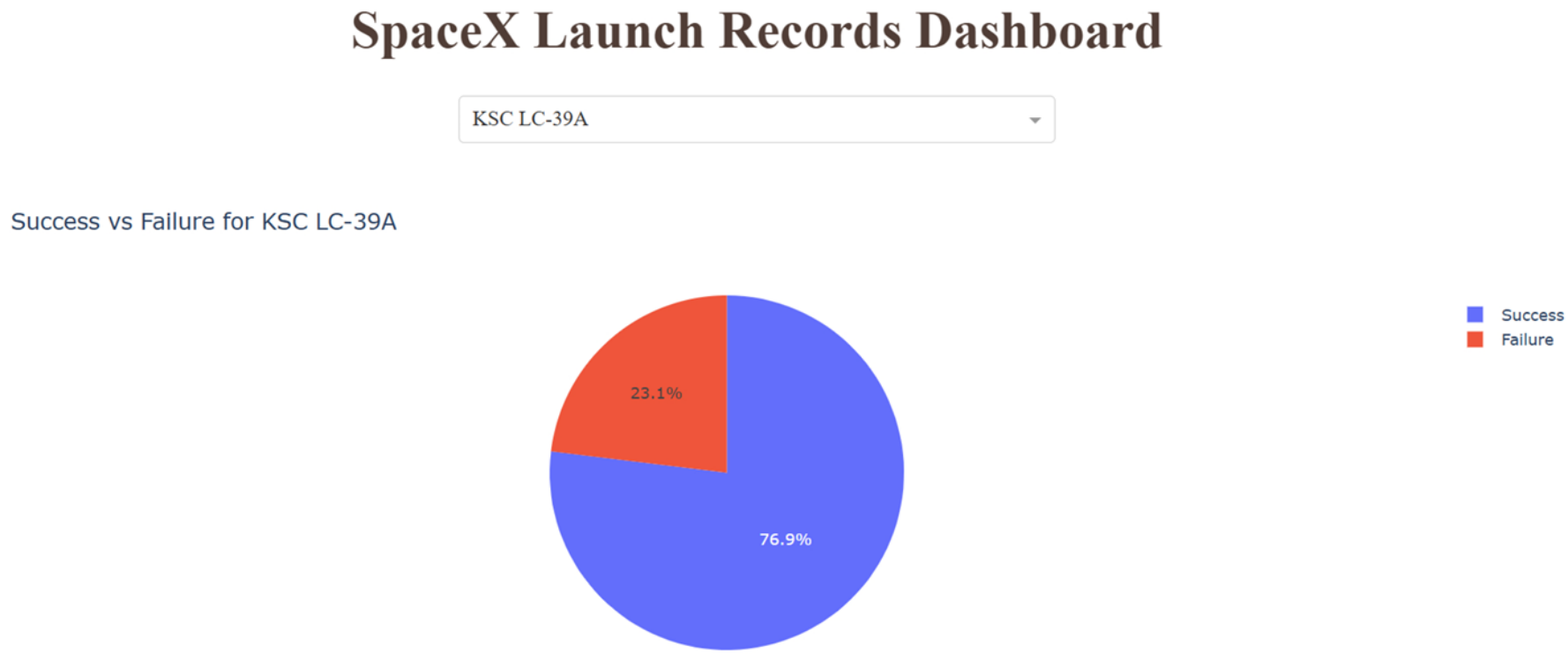
Plotly Dash SpaceX Launch Records Dashboard: Launch Successes for All Launch Sites

Displayed right is a screenshot of the interactive Plotly Dash dashboard developed to give the counts of landing successes by launch site. KSC-LC-39A shows the highest proportion of successful launches of all aggregated launch site success counts.



Plotly Dash SpaceX Launch Records Dashboard: Launch Site with Highest Success Rate

Below displays another screenshot of the interactive dashboard when the drop-down menu filters for the launch site “KSC LC-39A” – the launch site with the highest success rate. About ¾ of all landings were successful at this site.



Plotly Dash: Landing Outcomes by Booster Version and Payload Mass

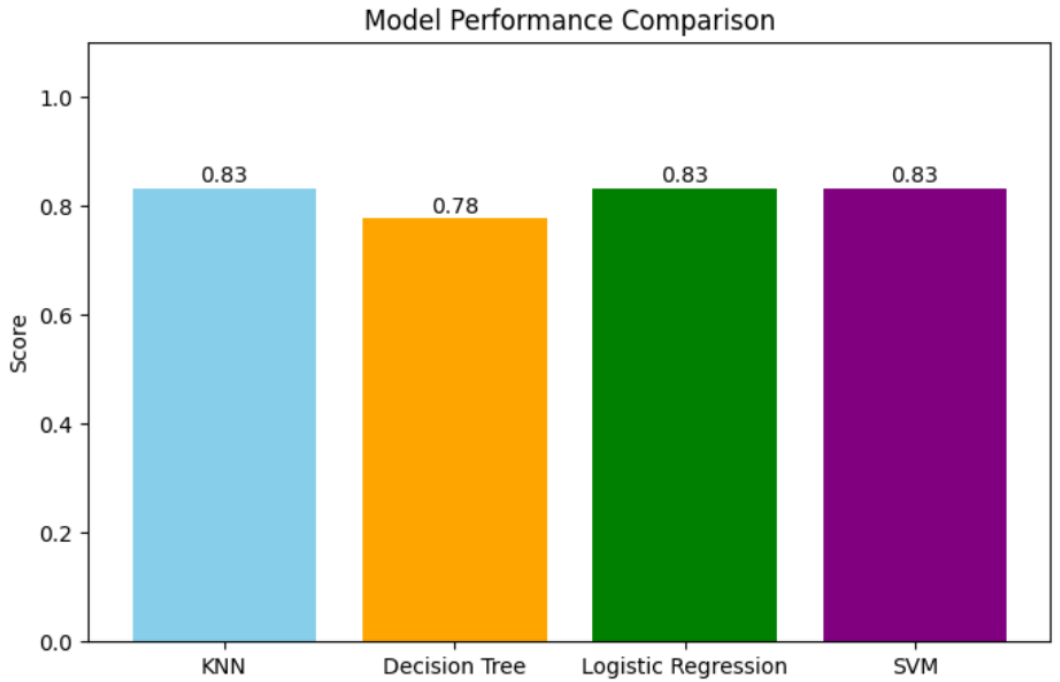
- Displayed right are screenshots of payload mass vs. launch outcome scatterplot for all sites with different payload mass ranges selected in the slider.
- With successes coded as “1” and failures coded as “0”, the payload range with the higher number of successes of these two examples appears to be the 2,500 kg-10,000 kg range. The booster version with the highest landing success rate is FT.

Section 5

Predictive Analysis (Classification)

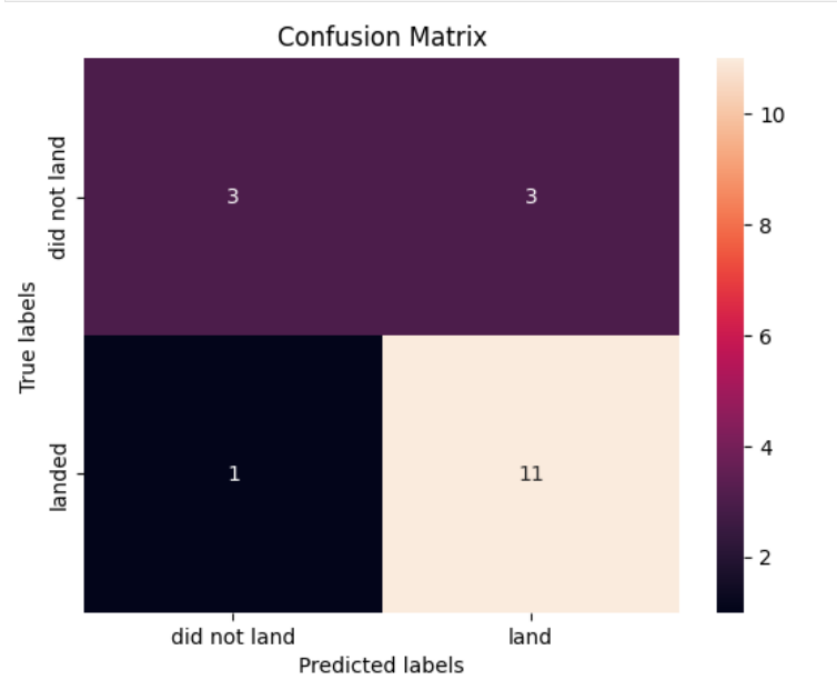
Classification Accuracy

- Of the classification machine learning models compared, KNN, Logistic Regression, and SVM all have the same accuracy metrics, with the Decision Tree Classifier method showing the lowest accuracy metric.



Confusion Matrix

- Despite having the lowest calculated accuracy (77.77), the decision tree classifier performs best on the test data according to the confusion matrix. This is because accuracy calculations ignore class balance issues.



Primary Conclusions

Different orbit types, launch sites, and payload masses show different success rates, but this may be attributable to the third variable of time and experience

Landing success rates show consistent upward trend across all factors, reflecting improvements with experience and technical developments

Classification machine learning models performed predictions of landing outcomes with fair accuracy, although a decision tree model performed the best according to confusion matrices

Appendix

- This link below directs to the GitHub repository containing all Jupyter notebooks with complete code used to perform these analyses. The repo also contains many relevant screenshots at full size for inspection.

Thank you!