IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Hussain Alibrahim>
<29-11-2022>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using API and Webscraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA) using SQL and Data Visualization
  - Interactive Visual Analytics and Dashboards using Folium and Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis (EDA) results
  - Interactive Analytics Map and Dashboard
  - Predictive Analysis results

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. In this project, we will predict if the Falcon 9 first stage will land successfully.
- Problems you want to find answers
  - Determine the price of each launch
  - Determine if the first stage will land successfully
  - Determine/predict if SpaceX will reuse the first stage

Section 1

Methodology

Methodology

- Data collection methodology:
  - API: SpaceX REST API
  - Web Scraping: Falcon 9 Launch Wiki page
- Perform data wrangling
  - Filtering the data frame to only include Falcon 9 launches
  - Dealing with missing values
  - Creating a landing outcome label from Outcome column
- 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

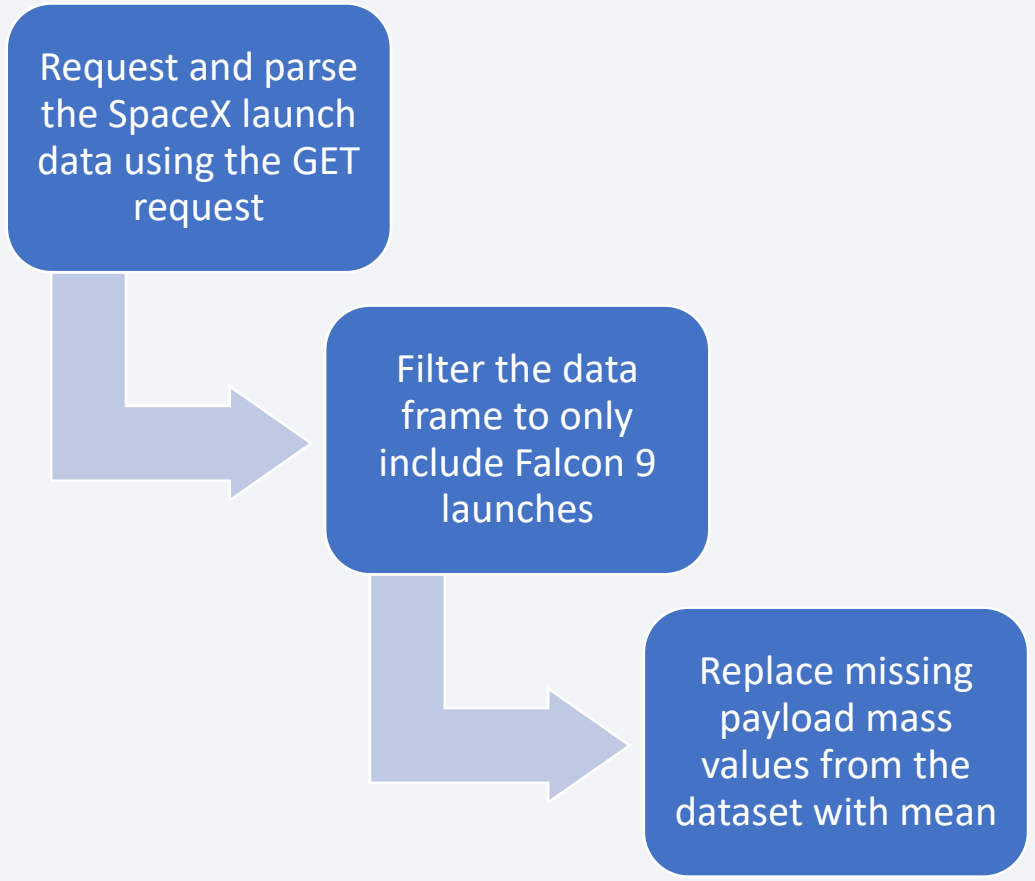
Data Collection

- Making a request to the SpaceX API with rocket, payloads, launchpad, and cores columns to extract information using identification numbers in the launch data: (https://api.spacexdata.com/v4/)
- Performing web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches: (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- Data collection with SpaceX REST calls using key phrases and flowcharts:

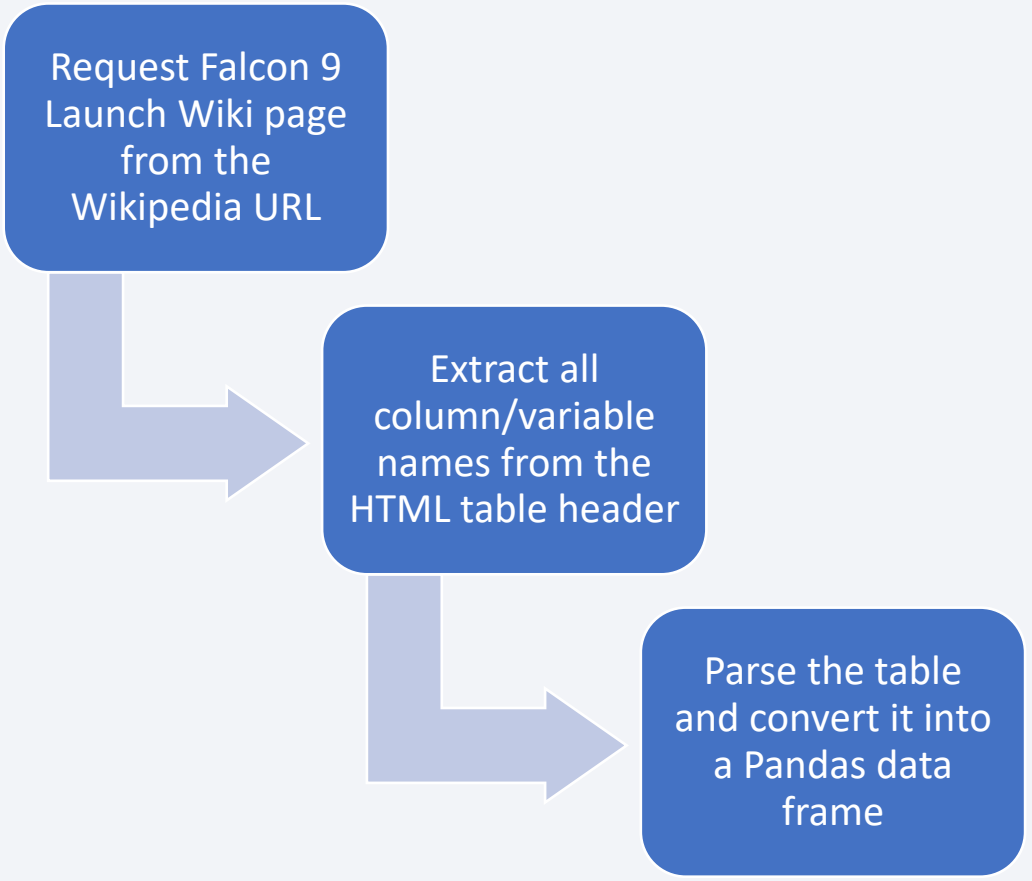
- GitHub URL of the completed SpaceX API calls notebook:
  - https://github.com/alibrahim83/Applied_Data_Science_Capstone/blob/main/spacex-data-collection-api.ipynb



Data Collection - Scraping

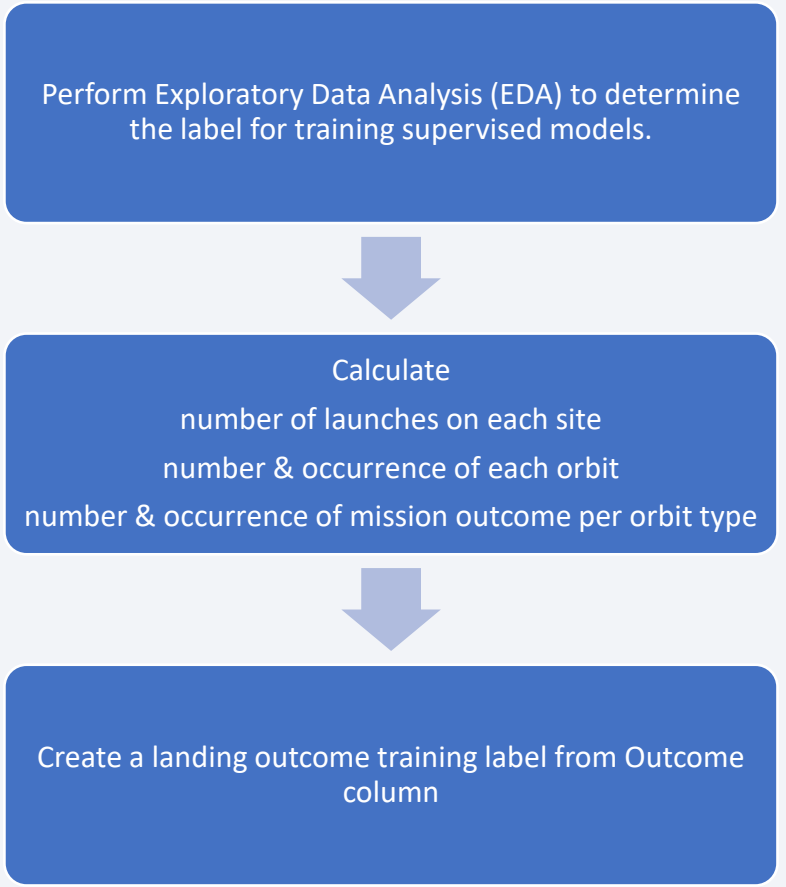
- Web scraping process using key phrases and flowcharts:

- GitHub URL of the completed web scraping notebook:
  - https://github.com/alibrahim83/Applied_Data_Science_Capstone/blob/main/webscraping.ipynb



Data Wrangling

- We performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- There are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example:
  - True Ocean - False Ocean - True RTLS - False RTLS - True ASDS - False ASDS
- We converted those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- GitHub URL of the completed data wrangling related notebooks:
  - https://github.com/alibrahim83/Applied_Data_Science_Capstone/blob/main/spacex-data_wrangling.ipynb



EDA with Data Visualization

- Scatter Plot
  - Displays values pertaining to typically two variables against each other to determine if any correlation between the two variables exists
    - Flight Number vs. Payload - Flight Number vs. Launch Site - Payload vs. Launch Site - Flight Number vs. Orbit Type - Payload vs. Orbit type
- Bar Chat
  - Compare the values of a variable at a given point in time
    - Success Rate vs. Orbit Type
- Line Plot
  - Visualizing the continuous dataset over a period of time
    - Success Rate vs. Year
- GitHub URL of the completed EDA with data visualization notebook:
  - https://github.com/alibrahim83/Applied_Data_Science_Capstone/blob/main/eda-datavisualization.ipynb

EDA with SQL

- SQL queries performed:
  - 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
  - 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 between the date 2010-06-04 and 2017-03-20, in descending order
- GitHub URL of the completed EDA with SQL notebook:
  - https://github.com/alibrahim83/Applied_Data_Science_Capstone/blob/main/eda-sql.ipynb

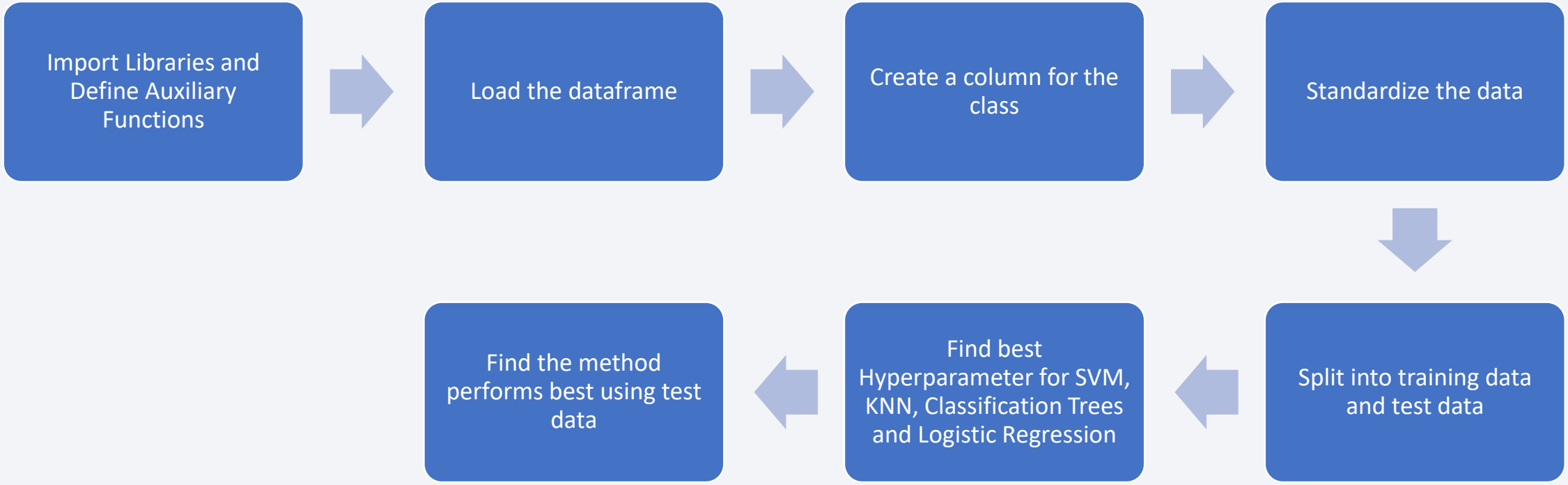
Build an Interactive Map with Folium

- Create a folium Map object, with an initial center location to be NASA Johnson Space Center at Houston, Texas
- Add an orange circle at NASA Johnson Space Center's and each launch site based on their coordinates (Lat, Long) with a popup label showing their names
- Mark the success (green) / failed (red) launches for each site on the map with marker cluster
- Calculate the distances between a launch site to its proximities and draw a line between them
- We added those objects to find some geographical patterns about launch sites and see which sites have high success rates
- GitHub URL of the completed interactive map with Folium map
  - https://github.com/alibrahim83/Applied_Data_Science_Capstone/blob/main/launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to the dashboard:
  - Dropdown list to enable Launch Site selection
  - Pie chart to show the total successful launches count for all sites
    - If a specific launch site was selected, show the Success vs. Failed counts for the site
  - Slider to select payload range
  - Scatter chart to show the correlation between payload and launch success
- GitHub URL of the completed Plotly Dash lab:
  - https://github.com/alibrahim83/Applied_Data_Science_Capstone/blob/main/spacex_dash_app.py

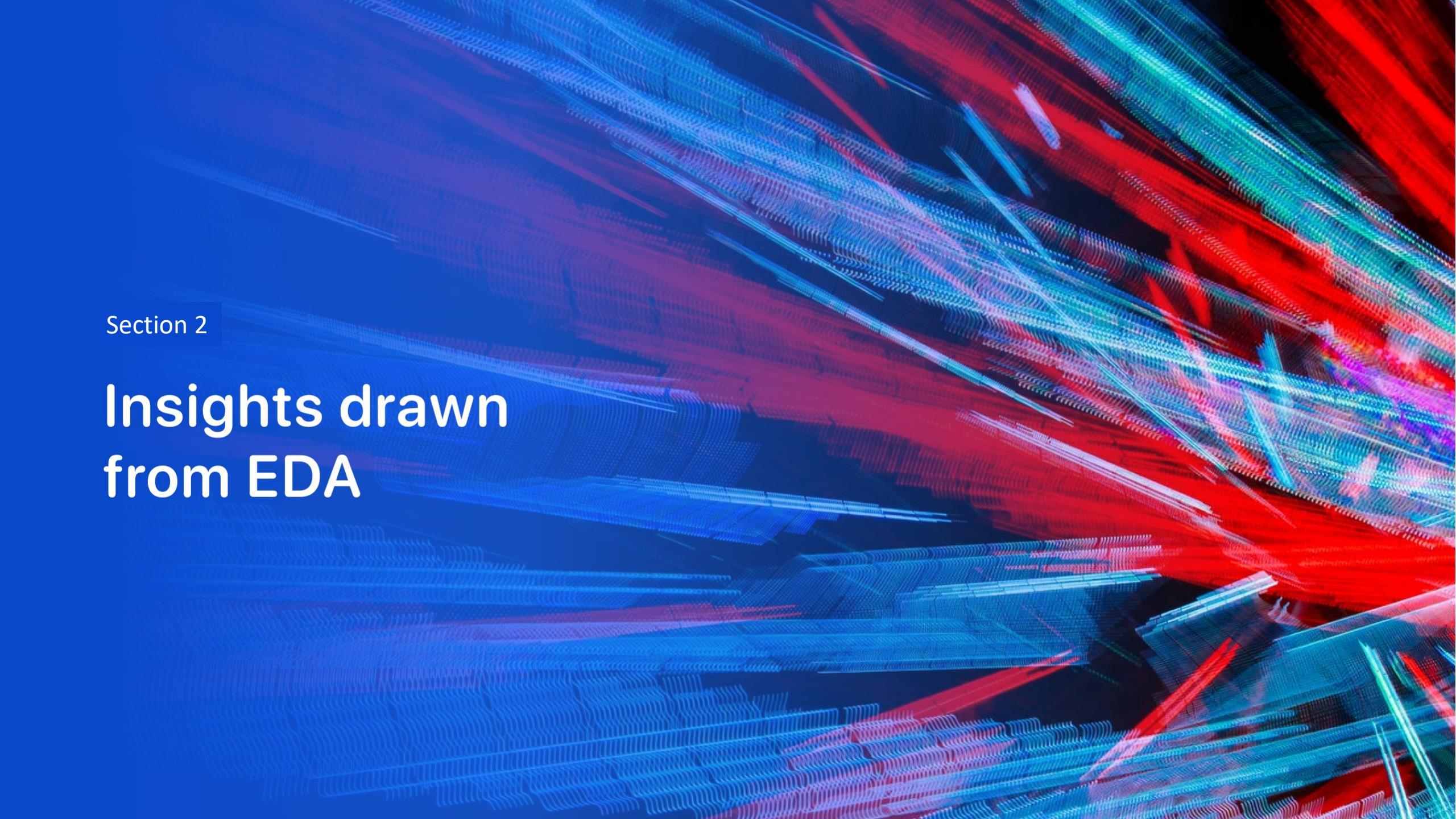
Predictive Analysis (Classification)



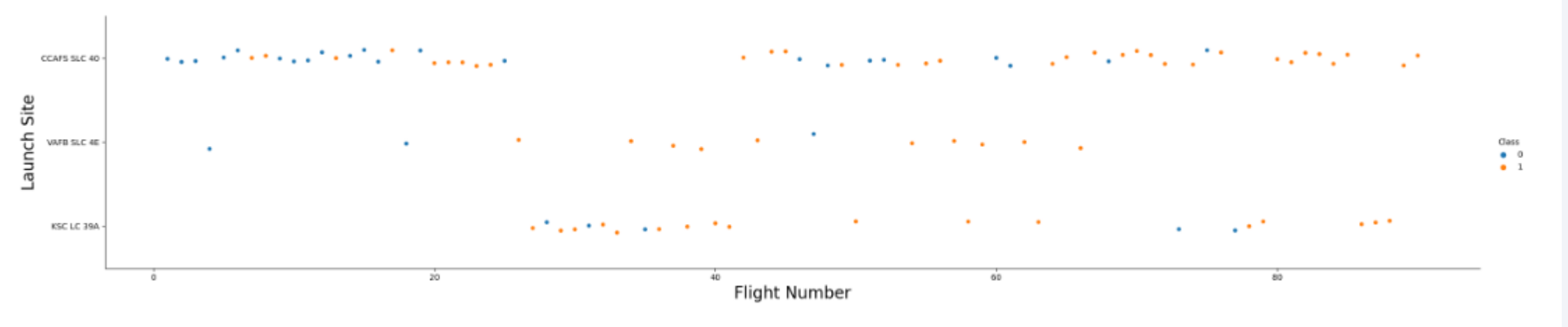
- GitHub URL of the completed predictive analysis lab:
  - https://github.com/alibrahim83/Applied_Data_Science_Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

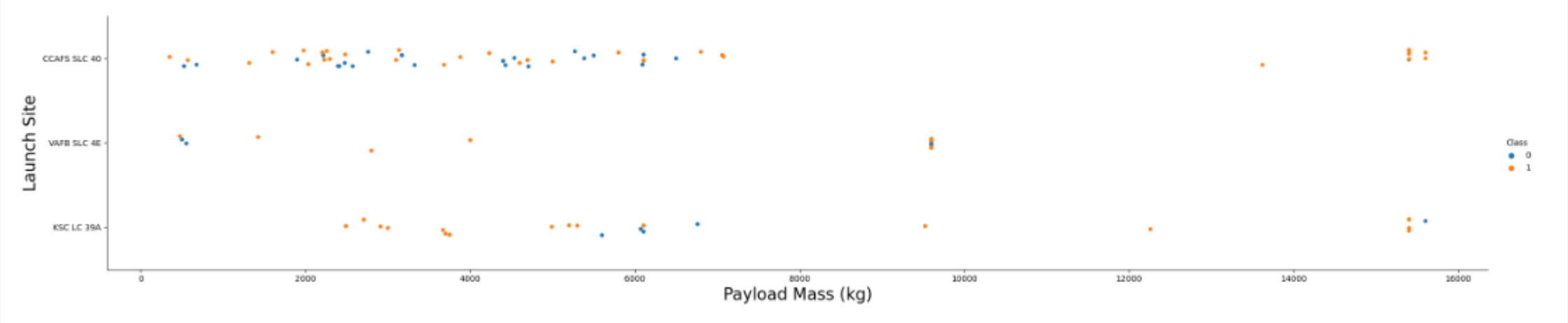

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



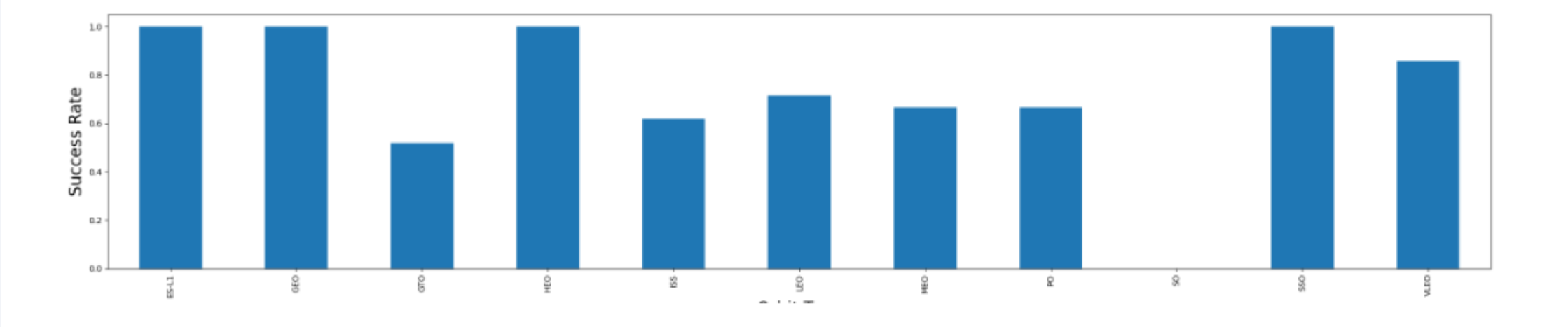
- Success rate is increasing for each site over time
- CCAF5 SLC 40 site has the most launches flights

Payload vs. Launch Site



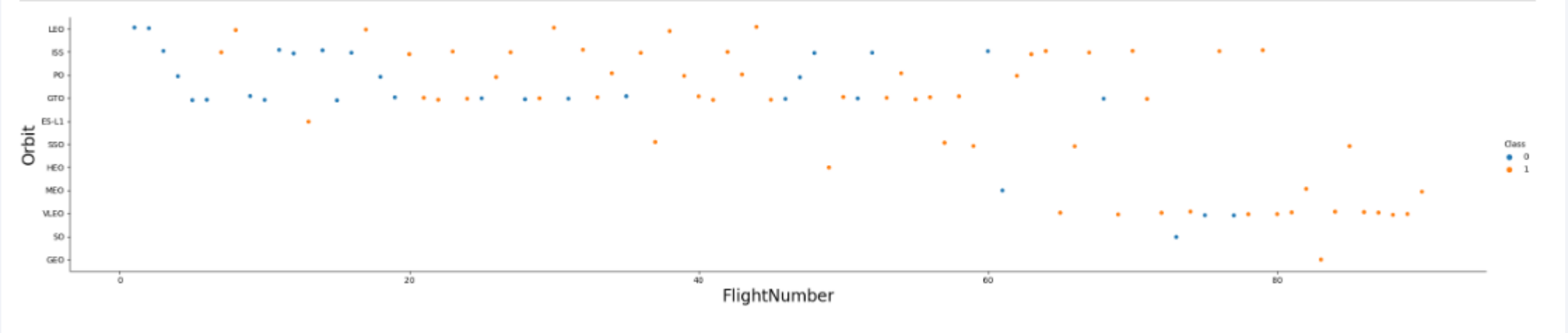
- VAFB-SLC launch site has no rockets launched for heavy payload mass (greater than 10000)
- CCAFS SLC 40 and KSC LC 39A seem to be preferred for heavier payloads

Success Rate vs. Orbit Type



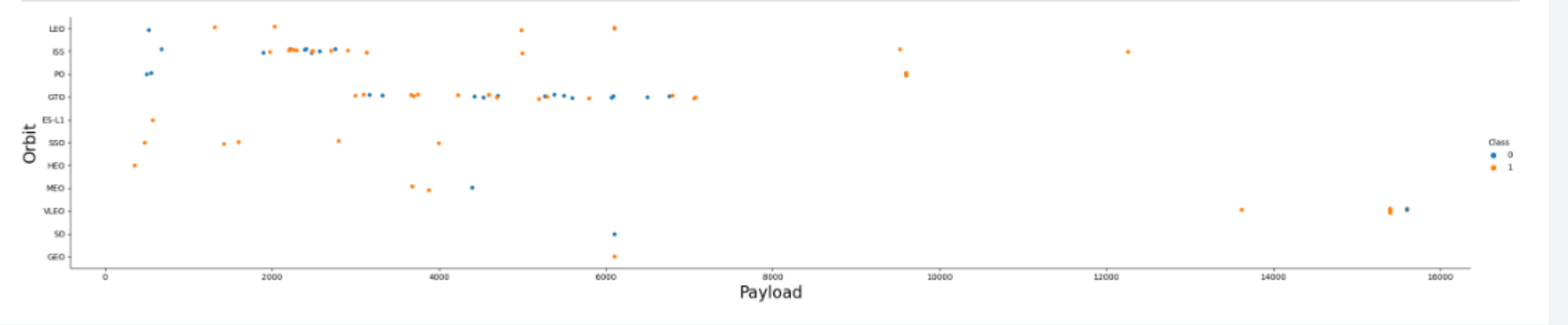
- ES-L1, GEO, HEO and SSO orbit had 100% success rate
- SO orbit had 0% success rate

Flight Number vs. Orbit Type



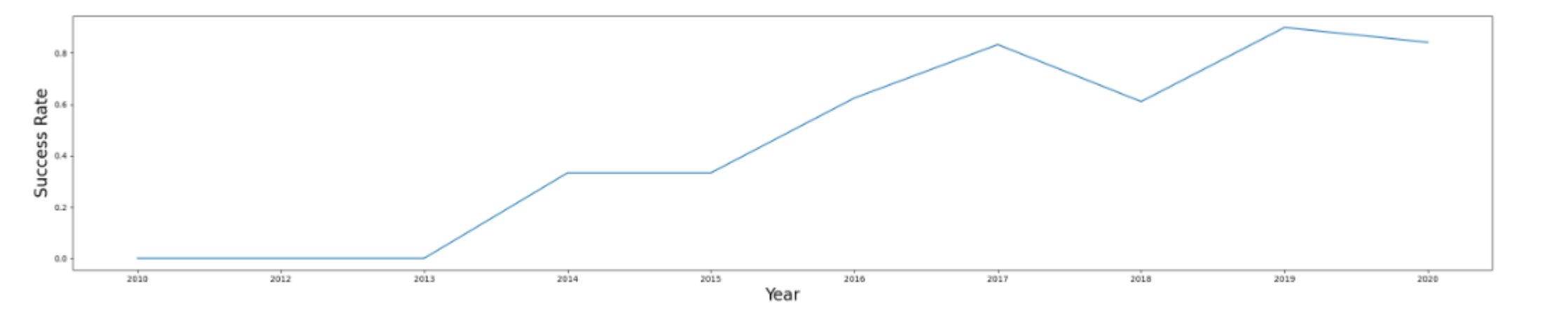
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.

Launch Success Yearly Trend



- Success rate since 2013 kept increasing till 2020

All Launch Site Names

In [30]:
```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

```
 * ibm_db_sa://shg46120:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[30]:

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

- Display the names of the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

In [31]:

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

```
 * ibm_db_sa://shg46120:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[31]:

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

- Display 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

In [32]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://shg46120:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[32]:

1
45596

- Display the total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

In [33]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1%';
```

```
 * ibm_db_sa://shg46120:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[33]:

1
2534

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

First Successful Ground Landing Date

In [34]:

```
%%sql
SELECT MIN(Date)
FROM SPACEXTBL
WHERE Landing__Outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://shg46120:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[34]:

1
2015-12-22

- List the date when the first successful landing outcome in ground pad was acheived

Successful Drone Ship Landing with Payload between 4000 and 6000

In [35]:
```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)' AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

```
 * ibm_db_sa://shg46120:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[35]:

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

In [36]:
```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://shg46120:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[36]:

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

- List the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

In [37]:
```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * ibm_db_sa://shg46120:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[37]:

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

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

2015 Launch Records

In [38]:
```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing__Outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://shg46120:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[38]:

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

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

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

In [39]:
```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

```
 * ibm_db_sa://shg46120:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

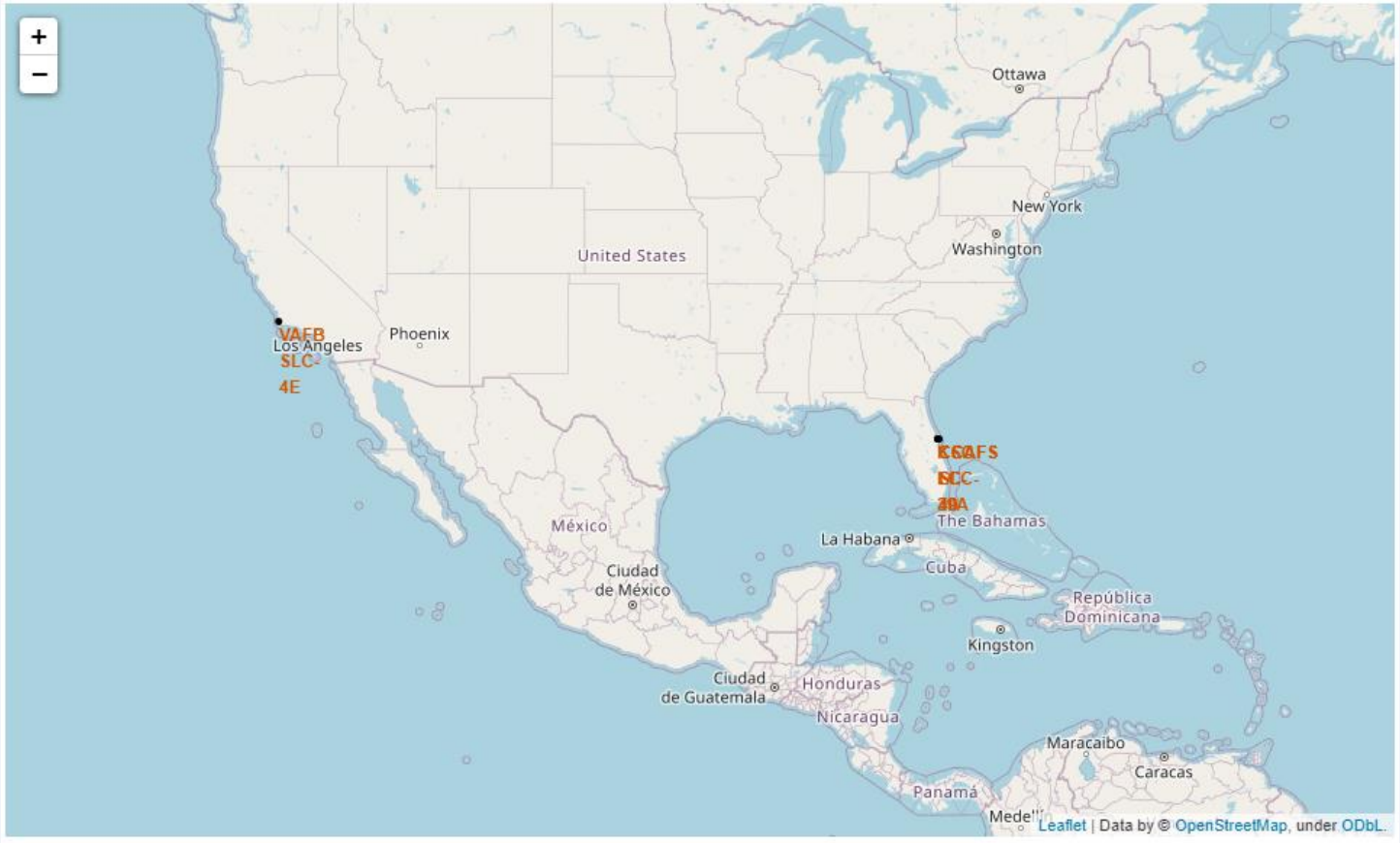
Out[39]:

landing__outcome	total_number
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

- 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

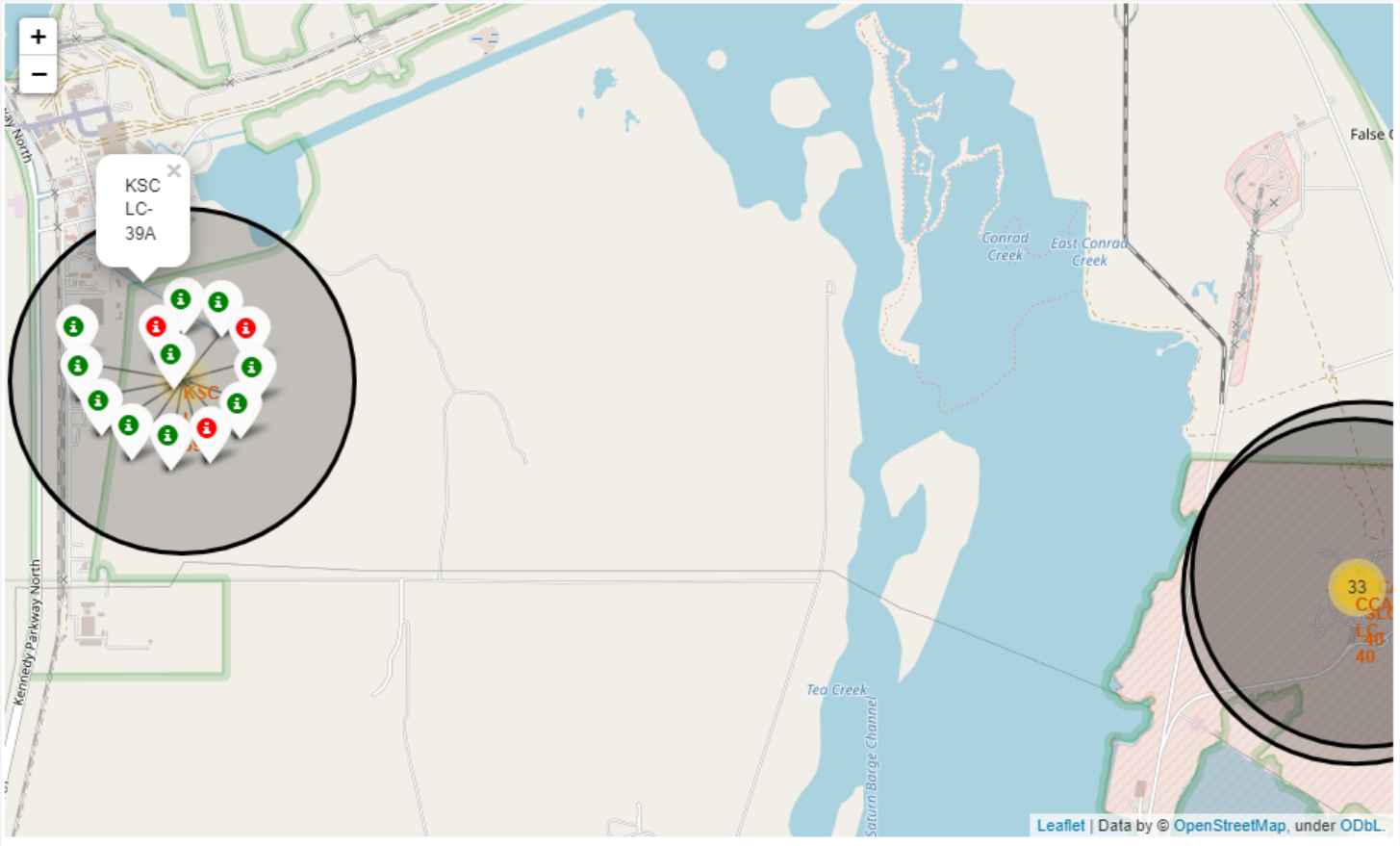

Section 3
Launch Sites
Proximities Analysis

All Launch Sites' Location Markers



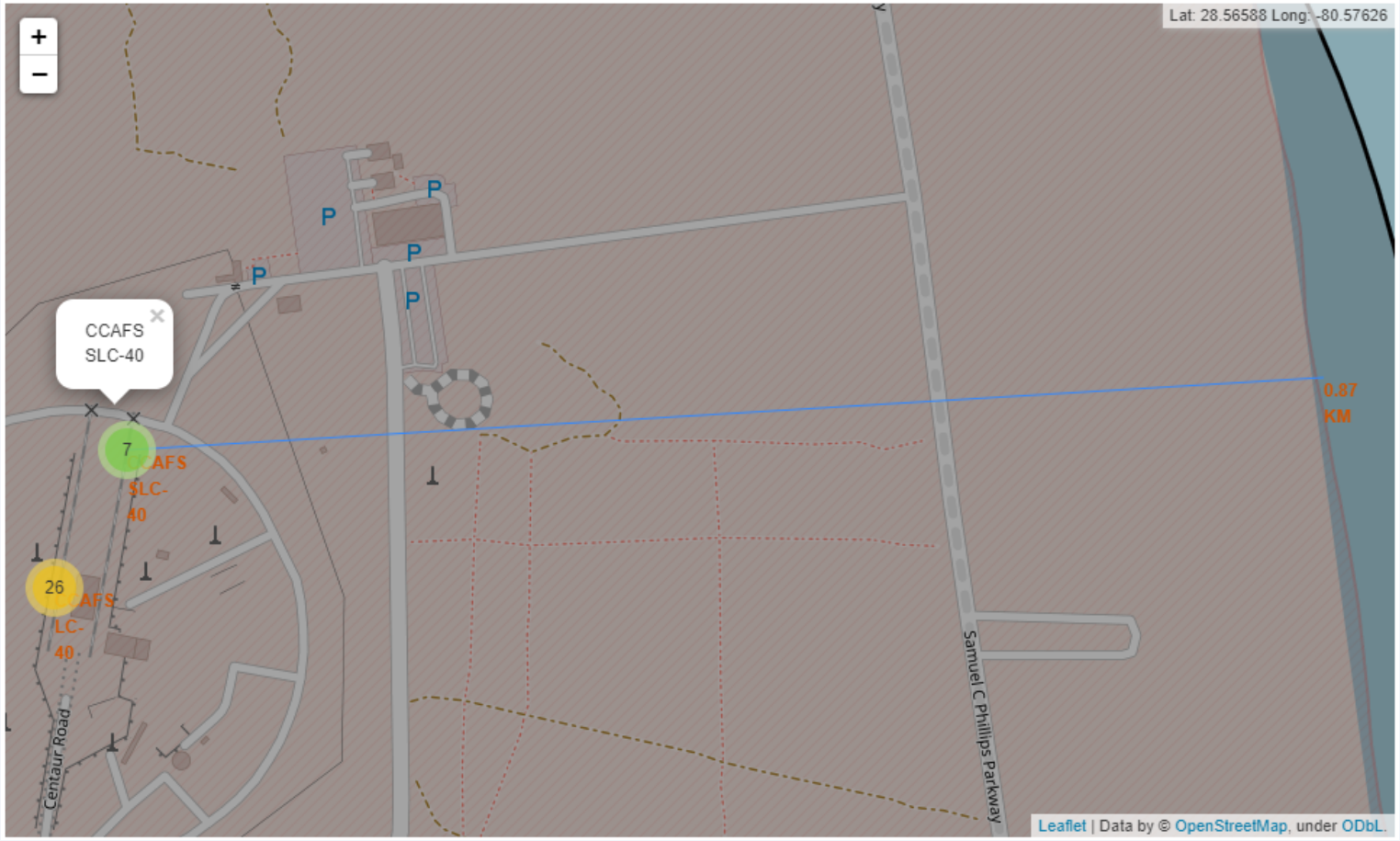
- All launch sites in proximity to the Equator line and in very close proximity to the coast

Color-Labeled Launch Outcomes

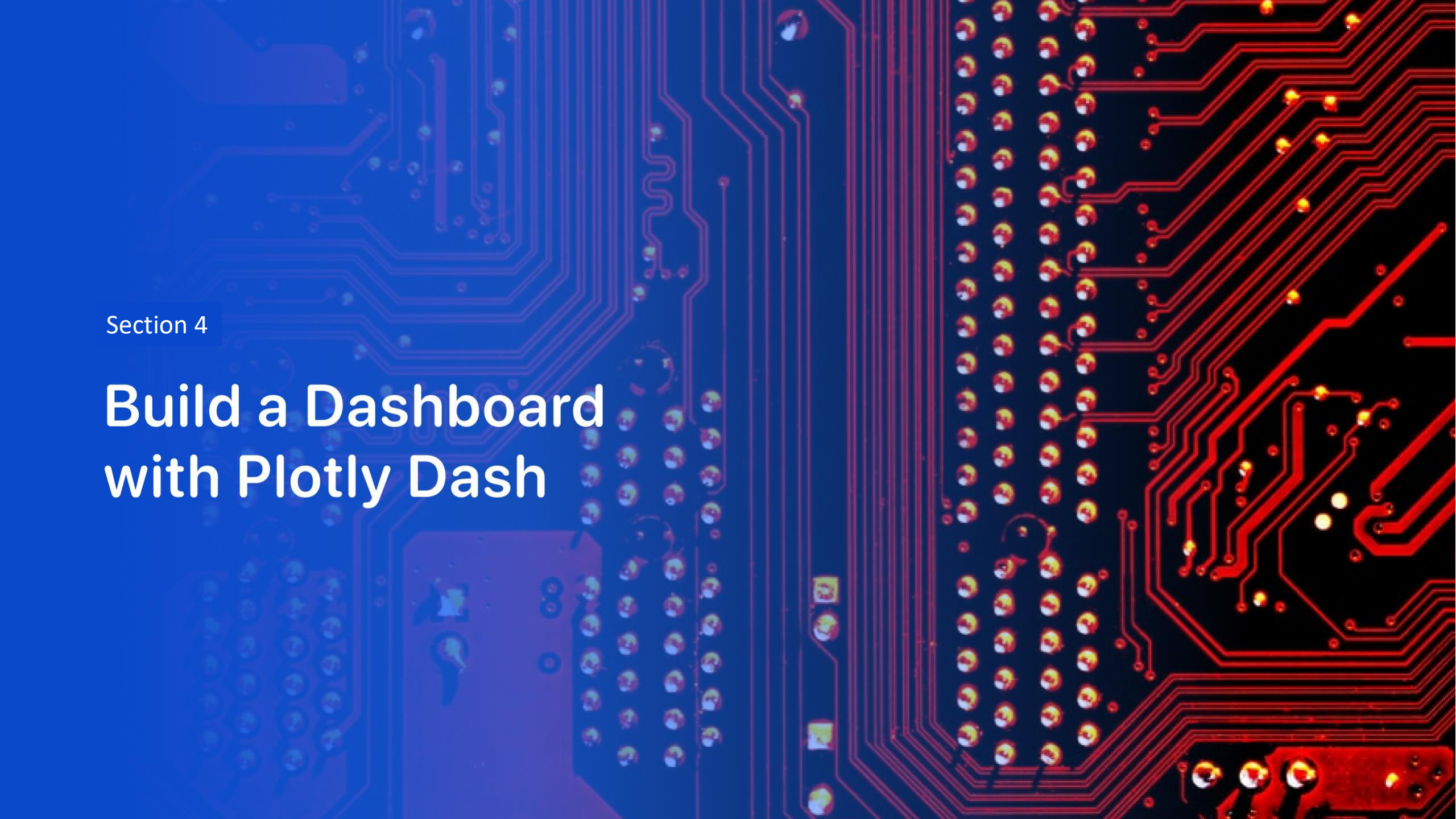


- From the color-labeled markers in marker clusters, we should be able to easily identify which launch sites have relatively high success rates, namely KSC LC-39A.

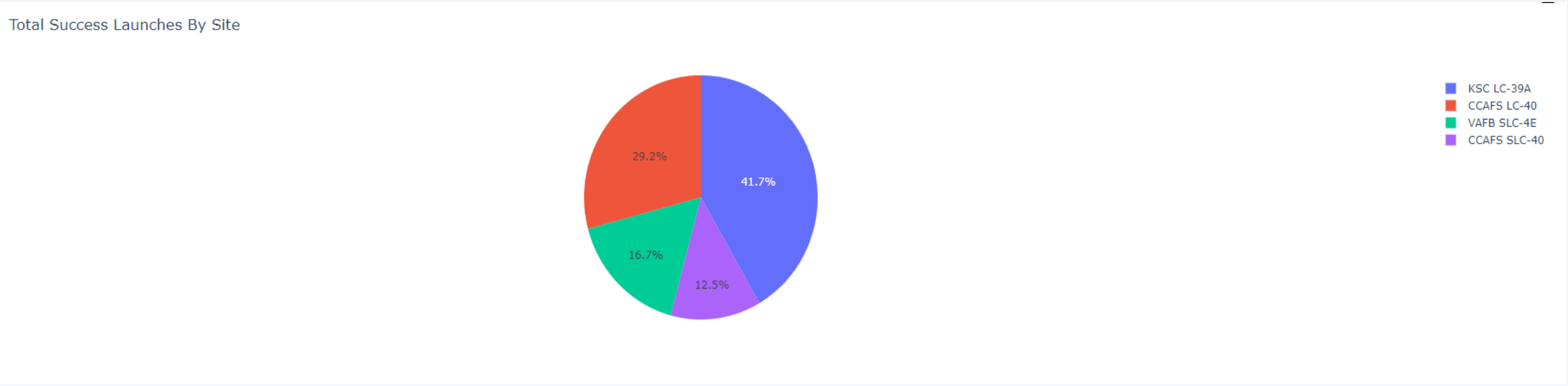
Launch Site to its Proximities



- After we plot distance lines to the proximities, we can see launch sites near coastline, railways, highways and cities; for example, CCAFS SLC-40 site is very close to coastline (0.87 km)

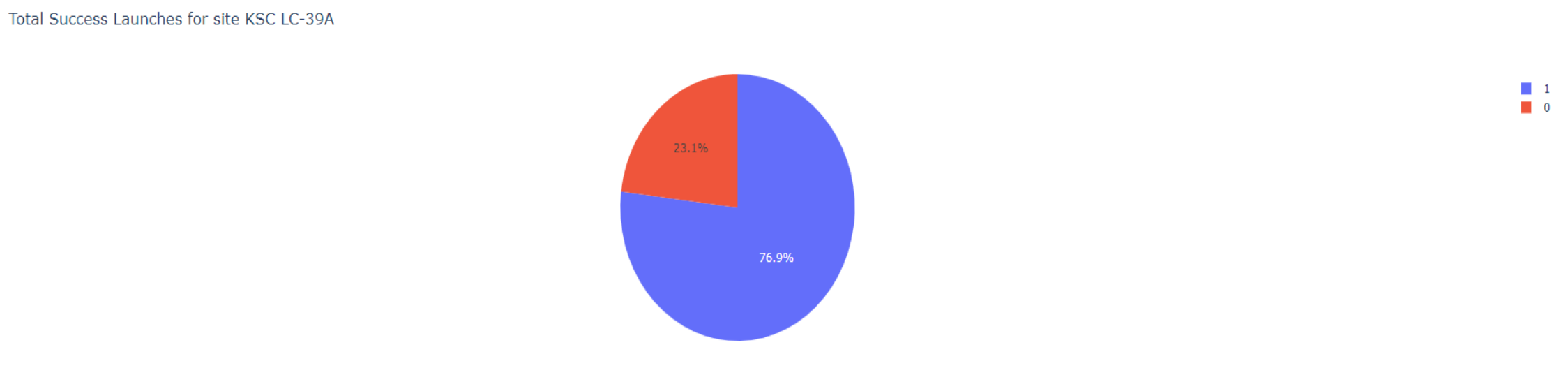

Section 4
Build a Dashboard with Plotly Dash

Success Launches by Site



- Total successful launches count for all sites
- KSC LC-39A site has the most successful launches (41.7%)

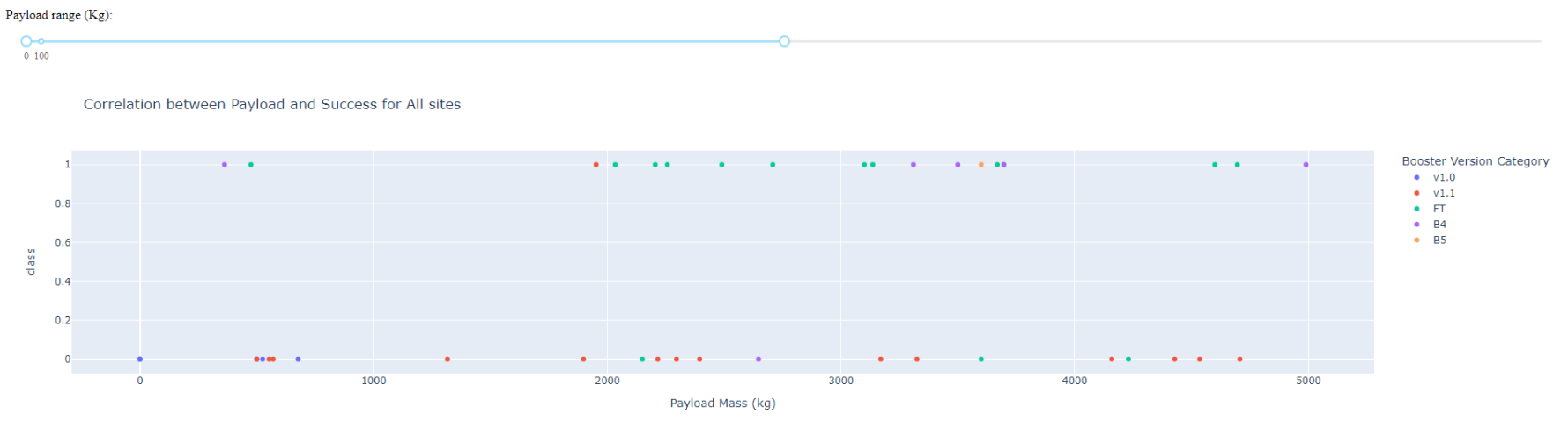
Success Launches for Site KSC LC-39A



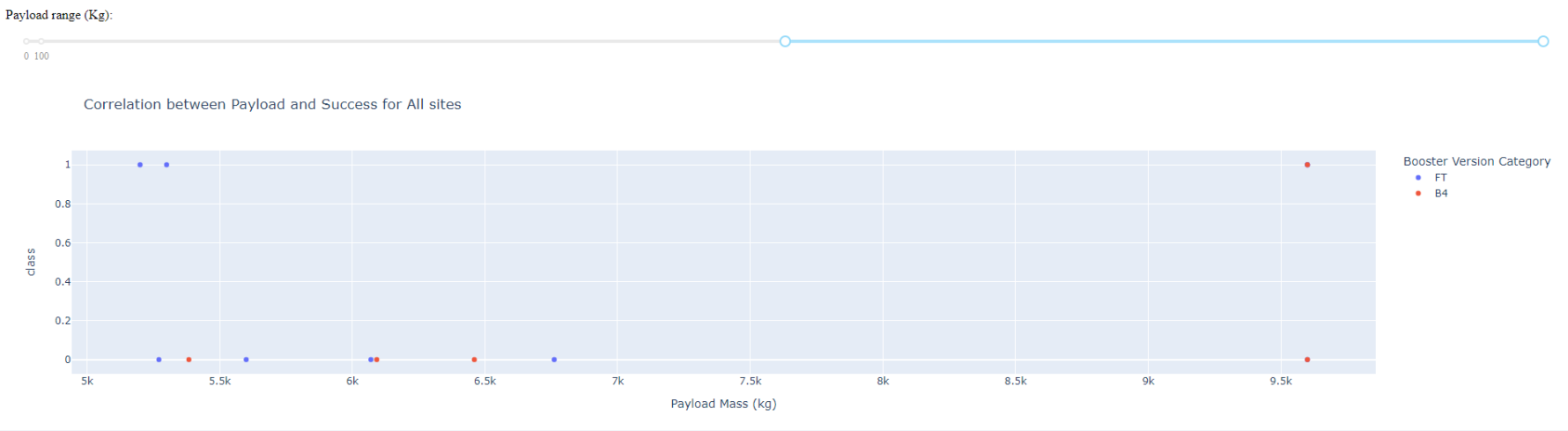
- KSC LC-39A site has 76.9% total success of all launches

Payload Mass vs. Launch Outcome

- Payloads < 5,300 kg had highest booster landing success rate



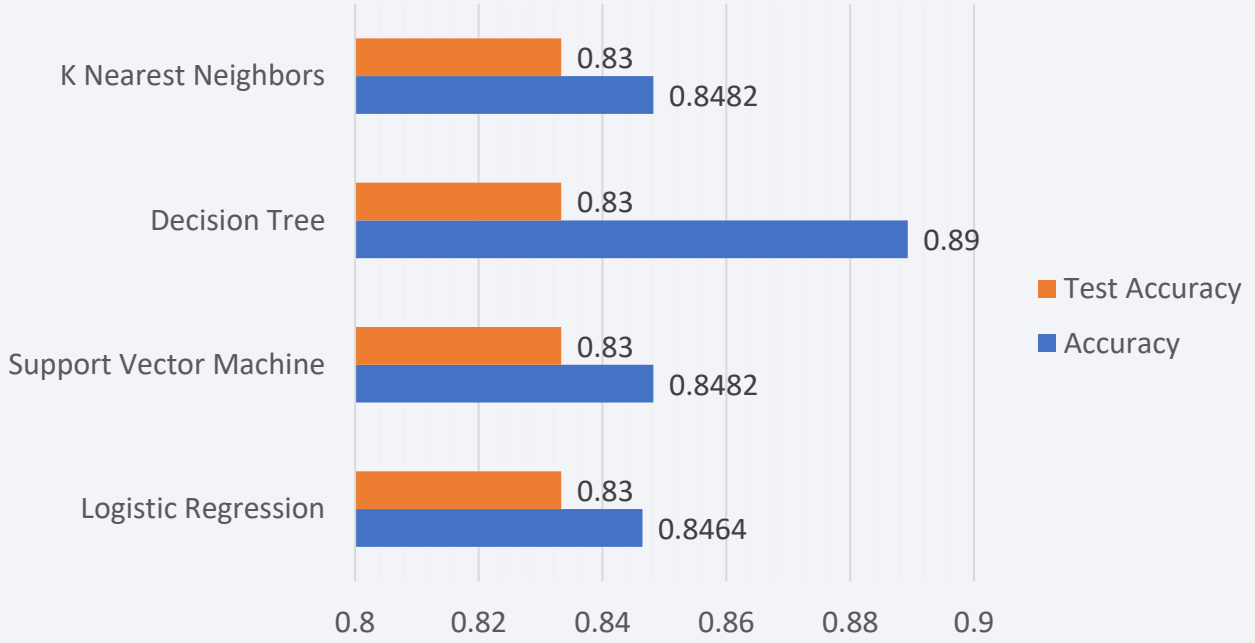
- Payloads > 5,300 kg had lowest booster landing success rate



Section 5

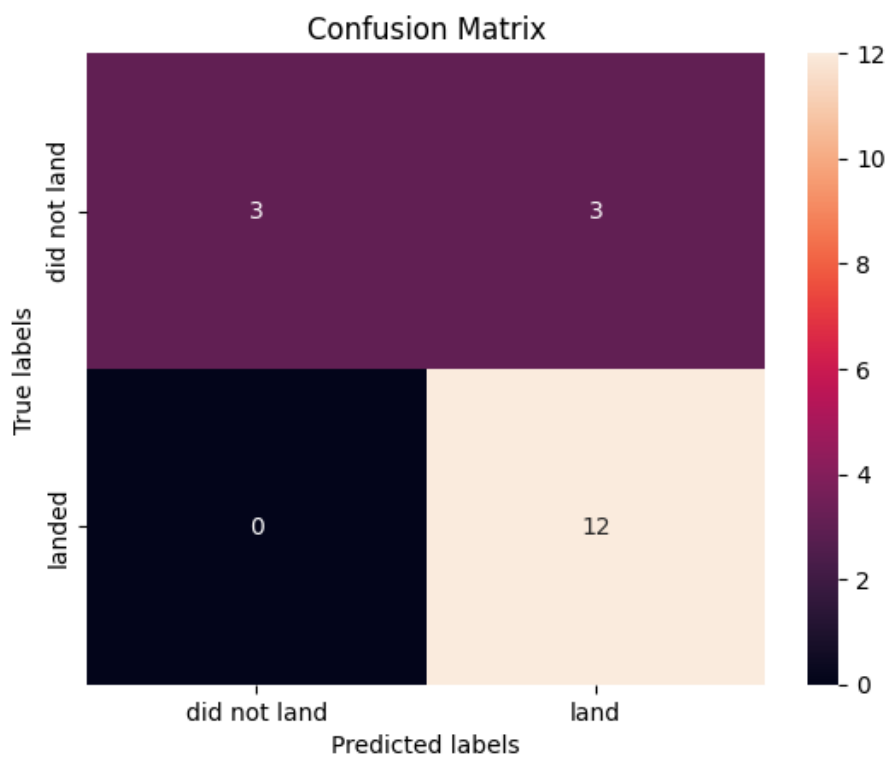
Predictive Analysis (Classification)

Classification Accuracy



- Decision tree model has the highest accuracy score (88.93%)
- All four models has the same test accuracy score (83.33%)

Confusion Matrix



- Confusion matrix for all models are the same
- The major problem is false positives as all models incorrectly predicting the 1st stage booster to land in 3 out of 18 samples in the test data set

Conclusions

- The best launch site is KSC LC-39A
- Launches with a low payload mass show better results than launches with a larger payload mass
- All launch sites are in proximity to the Equator line and in very close proximity to the coast
- The success rate of launches improve over the years
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate
- Decision Tree Classifier is the best algorithm for this dataset

Appendix

- Notebooks:
  - Launch sites locations analysis with folium notebook didn't show on Github correctly, so I used nbviewer: https://nbviewer.org/github/alibrahim83/Applied_Data_Science_Capstone/blob/main/launch_site_location.ipynb
- Acknowledgments:
  - Thanks to all instructors at IBM Skills Network for creating the courses and the materials

Thank you!