IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Muhammad Babar
22 September, 2021

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- **Summary of all results**
  - Exploratory data analysis results
  - - Interactive analytics demo in screenshots
  - - Predictive analysis results

Introduction

- **Project background and context**

The commercial space age is here, companies are making space travel affordable for everyone. the most successful is SpaceX. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first

- **Project Scope**

Space Y would like to compete with Space X. To determine the price of each launch, this is done by gathering information about Space X and creating dashboards.

Also determine if the first stage will land successfully, by training a machine learning model and use public information to predict if SpaceX will reuse the first stage.

- **Questions**
  - What are the variables influencing the rocket landing?
  - What is the relationship of the variables with successful landing?
  - What are the best conditions to ensure highest successful landing?

Methodology

Executive Summary

- **Data collection methodology:**
  - SpaceX Rest API
  - Web Scrapping
- **Perform data wrangling**
  - One Hot Encoding data fields and dropping irrelevant columns
- **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

Section 1

Methodology

Data Collection

- **SpaceX REST API:**
  - Data is gathered from the SpaceX REST API.
  - This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/
- **Web Scraping:**
  - To collect Falcon 9 historical launch records from a Wikipedia page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

1. Get a response from API.
2. Convert response to a .json file.
3. Apply custom functions to gather data.
4. Create a dictionary and convert it to data frame.
5. Filter the data frame for Falcon 9 and convert to csv.

GitHub URL of the Notebook

Data Collection - Scraping

1. Get a response from HTML.
2. Create a BeautifulSoup Object.
3. Find tables and create column names.
4. Create a dictionary and append data to keys.
5. Convert dictionary to data frame and convert to csv.

GitHub URL of the Notebook

Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully.
- Sometimes a landing was attempted but failed due to an accident.
- Converted those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

- Perform Exploratory Data Analysis (Process)
  - Number of launches at each site
  - Number of occurrence of each orbit
  - Number of occurrence, mission outcome per orbit
  - Create a landing outcome label from Outcome
  - Success rate for every landing calculation
  - Export data as .csv

GitHub URL of the Notebook

EDA with Data Visualization

- **Scatter Graphs:**
  - Flight Number against Payload Mass
  - Flight Number against Launch Site
  - Payload against Launch Site
  - Orbit against Flight Number
  - Payload against Orbit
  - Orbit against Payload Mass

Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation . Scatter plots usually consist of a large body of data.

- **Bar Graph:**
  - Mean against Orbit

A bar diagram makes it easy to compare sets of data between different groups at a glance. The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes. Bar charts can also show big changes in data over time.

- **Line Graph:**
  - Success Rate against Year

Line graphs are useful in that they show data variables and trends very clearly and can help to make predictions about the results of data not yet recorded.

GitHub URL of the Notebook

EDA with SQL

- SQL queries you performed:
  - Unique launch sites in the space mission.
  - 5 records where launch sites begin with the string ‘CAA’.
  - Total payload mass carried by boosters launched by NASA (CRS).
  - Average payload mass carried by booster version F9 v1.1.
  - Date of the first successful landing outcome in ground pad.
  - Booster Version 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.
  - Booster versions which have carried the maximum payload mass in descending order.
  - Successful landing outcomes in drone ship, their booster versions and launch site names for the year 2015.
  - Ranking the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

GitHub URL of the Notebook

Build an Interactive Map with Folium

- To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site. We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()
- Using a formula we calculated the distance from the Launch Site to various locations to find various trends about what is around the Launch Site to measure patterns. Lines are drawn on the map to measure distance to the location.

GitHub URL of the Notebook

Build a Dashboard with Plotly Dash

- Graphs
  - Pie Chart showing the total launches by a certain site/all sites.
  - Display relative proportions of multiple classes of data.
  - Size of the circle can be made proportional to the total quantity it represents.
- Scatter Graph showing the relationship with Outcome and Payload Mass (Kg) for the different Booster Versions
  - It shows the relationship between two variables.
  - It is the best method to show you a non-linear pattern.
  - The range of data flow, i.e. maximum and minimum value, can be determined.
  - Observation and reading are straightforward.

GitHub URL of the Notebook

Predictive Analysis (Classification)

- **BUILDING MODEL**
  - Load our dataset into NumPy and Pandas and transform data
  - Split our data into training and test data sets
  - Decide which type of machine learning algorithms we want to use
  - Set our parameters and algorithms to GridSearchCV
  - Fit our datasets into the GridSearchCV objects and train our dataset.
- **EVALUATING MODEL**
  - Check accuracy for each model
  - Get tuned hyperparameters for each type of algorithm
  - Plot Confusion Matrix
- **IMPROVING MODEL**
  - Feature Engineering
  - Algorithm Tuning
- **The model with the best accuracy score wins the best performing model**

GitHub URL of the Notebook

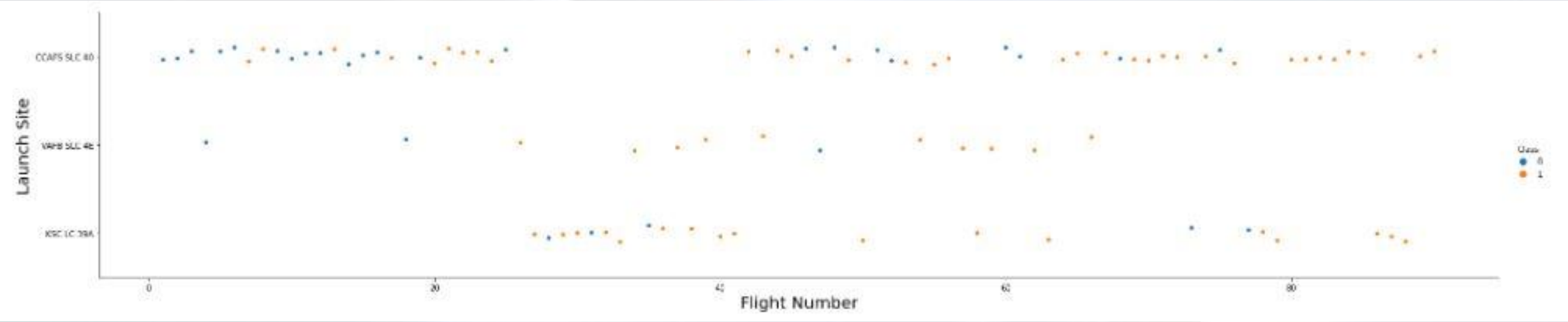
Results

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

Section 2

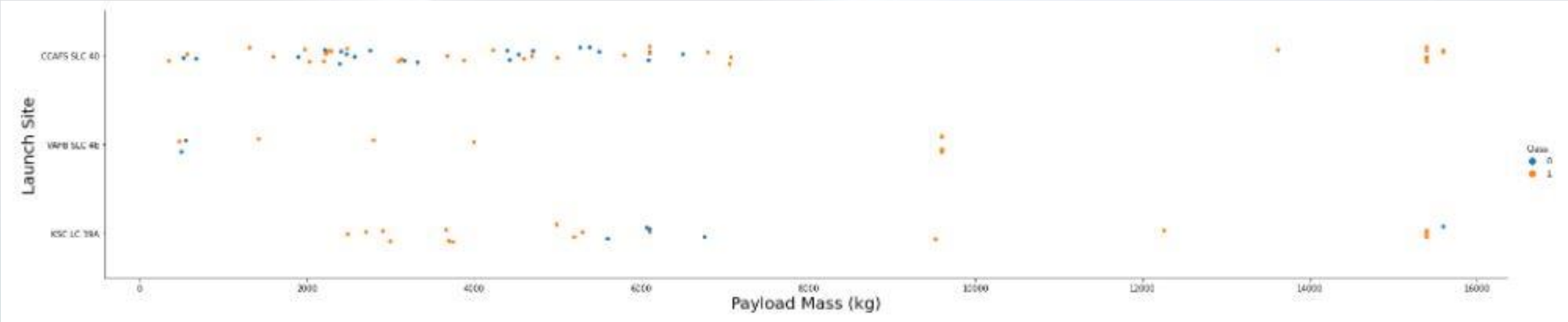
Insights drawn from EDA

Flight Number vs. Launch Site



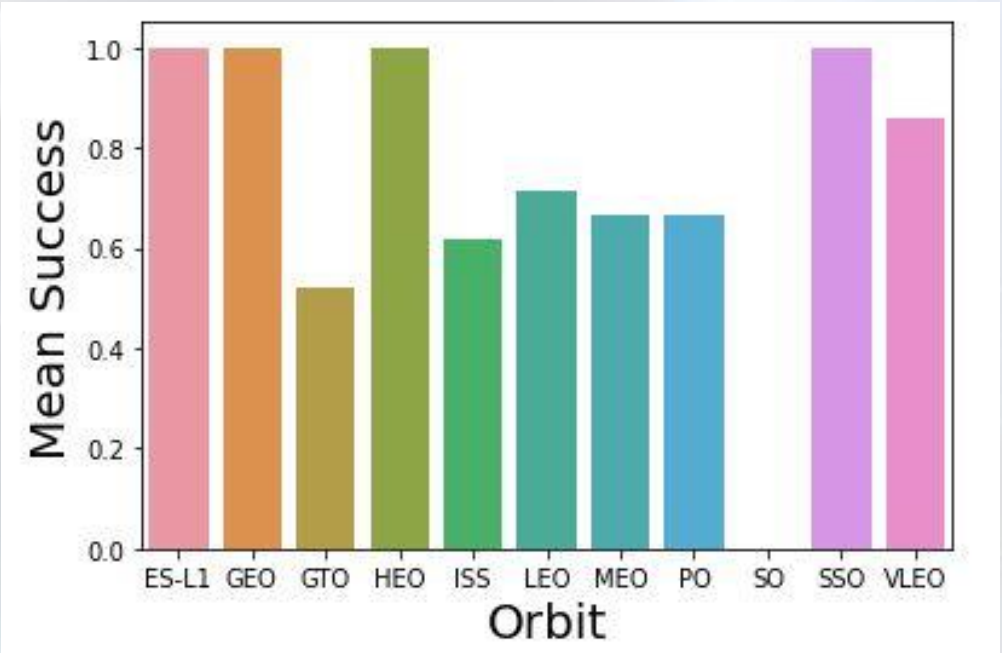
Success rate is increasing as Flight Number increases. The most successful Launch Site is KSC LC-39A. The highest number of Flights is from Launch Site CCAFS SLC 40.

Payload vs. Launch Site



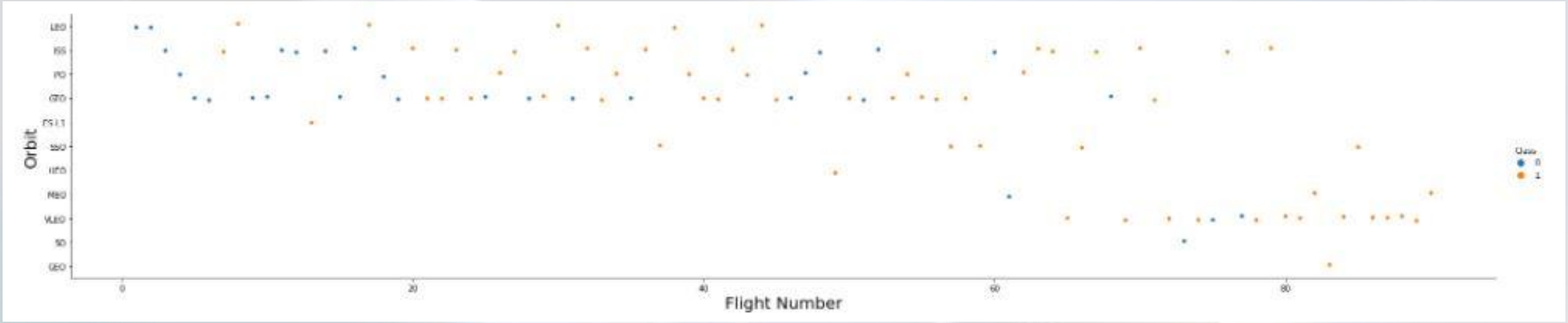
Very few launches with payload more than 8000 and significantly high change of success can be observed.

Success Rate vs. Orbit Type



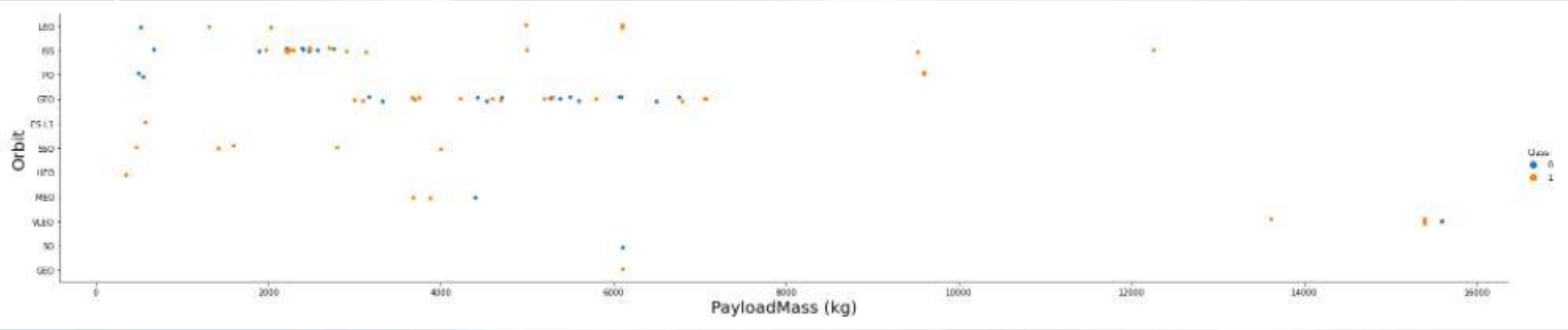
Orbit ES-L1, GEO, HEO and SSO have highest success rate. Where as GTO has the lowest success rate.

Flight Number vs. Orbit Type



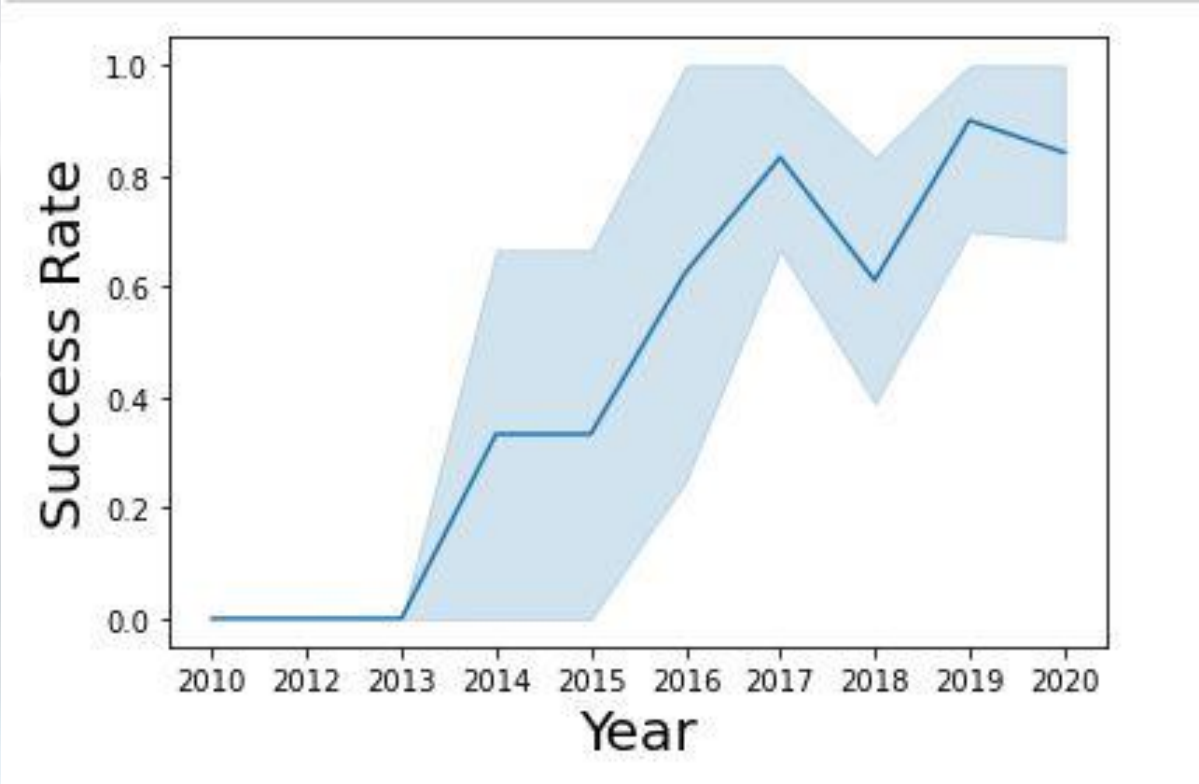
Orbit LEO has success rate related to increasing number of Flights. Orbit GTO has the highest number of Flights.

Payload vs. Orbit Type



Orbit has an inverse relationship with PayloadMass. However, Orbit LEO, ISS and PO seems to be having a direct relationship with increase in Payload.

Launch Success Yearly Trend



Success rate has been increasing since the year 2013. The rates dipped in the year 2018 before being the highest in the year 2019.

All Launch Site Names

Task 1 ¶

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

```
%%sql
SELECT DISTINCT(Launch_Site) FROM Spacex
```

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

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

Using the word **DISTINCT** in the query means that it will only show Unique values in the **Launch_Site** column from table **SpaceX**

Launch Site Names Begin with 'CCA'

Task 2

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

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

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

index	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
0	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	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)
2	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
3	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Using **LIMIT** in the query means that it will only show **5** records from table **SpaceX** and **LIKE** keyword has a wild card with the words '**CAA%**' the percentage in the end suggests that the **Launch_Site** name must start with CAA.

Total Payload Mass

Task 3

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM Spacex
WHERE Customer = "NASA (CRS)";
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Using the function **SUM** summates the total in the column **PAYLOAD_MASS_KG_**. Whereas, the **WHERE** clause filters the dataset to only perform calculations when **Customer** **is** **NASA (CRS).**

Average Payload Mass by F9 v1.1

Task 4

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM Spacex
WHERE Customer = "NASA (CRS)";
```

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

AVG(PAYLOAD_MASS__KG_)
2279.8

Using the function **AVG** averages the values in the column **PAYLOAD_MASS_KG_**. Whereas, the **WHERE** clause filters the dataset to only perform calculations when **Customer** is **NASA (CRS).**

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
%%sql
SELECT Min(Date) FROM Spacex
WHERE "Landing _Outcome" LIKE 'Success (ground pad)'
```

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

Min(Date)
01-05-2017

Using the function **MIN** gets the minimum value. Where the **WHERE** clause filters the dataset to only perform calculations when **Landing_Outcome** is **Success (ground pad).**

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

```
%%sql
SELECT Booster_Version FROM Spacex
WHERE "Landing _Outcome" = "Success (drone ship)" AND
PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000;
```

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

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

Select only **Booster_Version**, **WHERE** the dataset **Landing_Outcome** is **Success (drone ship)** and **PAYLOAD_MASS_KG_** is between **4000 - 6000**

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%%sql
SELECT Mission_Outcome, Count(Mission_Outcome) FROM Spacex
GROUP BY Mission_Outcome;
```

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

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

Select **Mission_Outcome** and **Count** the number of **Mission_Outcome**. **GROUP BY** ensures summing of Count values.

Boosters Carried Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
SELECT Booster_Version, max(PAYLOAD_MASS__KG_) as PayloadMass FROM Spacex
GROUP BY Booster_Version ORDER BY PayloadMass DESC;
```

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

Booster_Version	PayloadMass
F9 B5 B1060.3	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1056.4	15600
F9 B5 B1051.6	15600

Select **Booster_Version** and **Max(PAYLOAD_MASS_KG_)** only.
GROUP BY ensures distinct **Booster_Version** and **ORDER BY (DESC)** sets the order of the results.

2015 Launch Records

Task 9

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

```
%%sql
SELECT Booster_Version, Launch_Site FROM Spacex
WHERE "Landing _Outcome" LIKE "%Failure (drone ship)%" and
"Date" LIKE "%2015%"
```

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

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

Select **Booster_Version** and **Launch_Site**, **WHERE** the dataset **Landing_Outcome** is **Failed (drone ship)** and **Date** has the word **2015** to ensure the year.

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

Task 10

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

```
%%sql
SELECT "Landing _Outcome", Count("Landing _Outcome") as Counts  FROM Spacex
WHERE "DATE" >= "04-06-2010" AND "DATE" <= "20-03-2010"
GROUP BY "Landing _Outcome"
ORDER BY Counts desc
```

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

Landing _Outcome	Counts
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

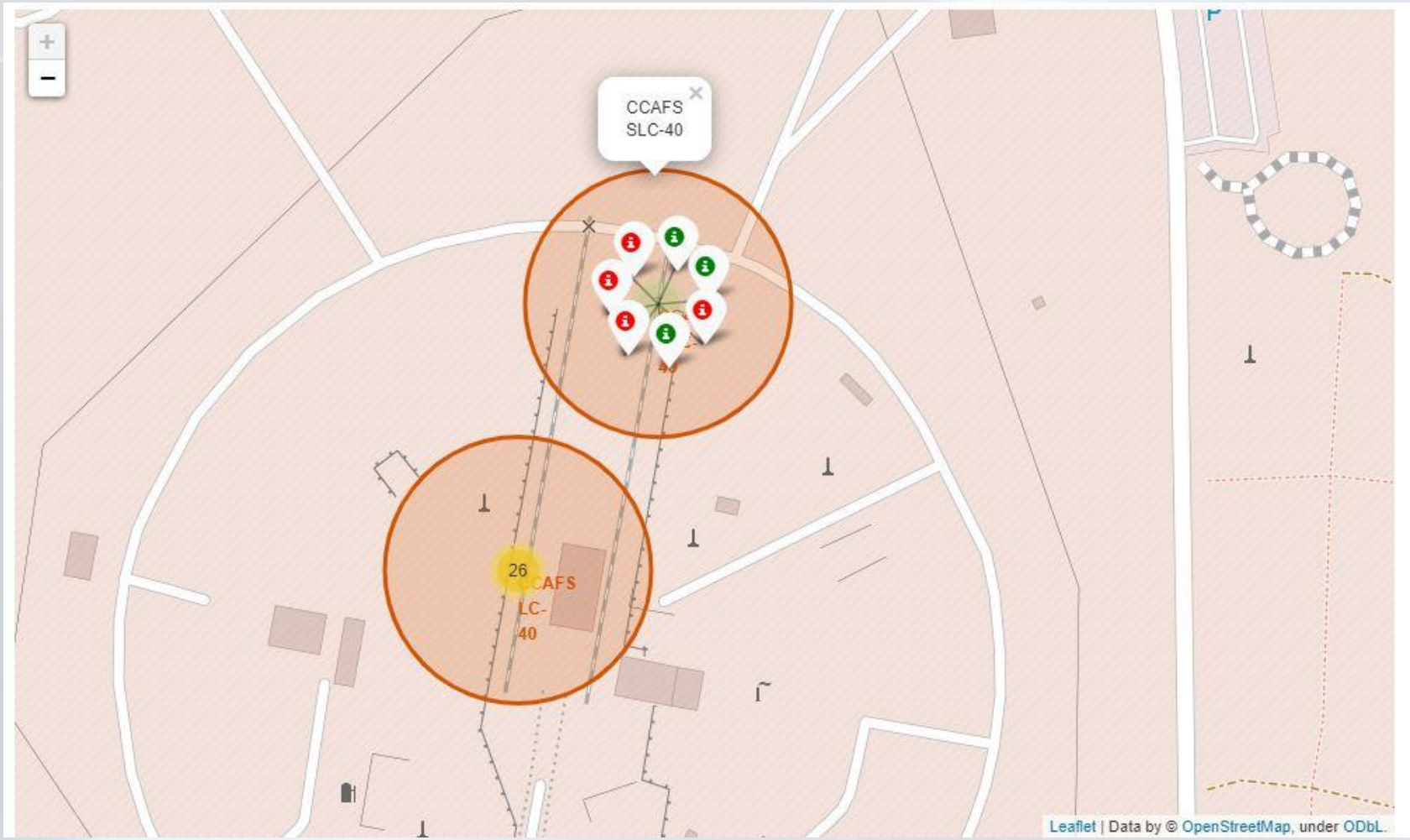
Section 4

Launch Sites Proximities Analysis

All Launch Site for Falcon 9

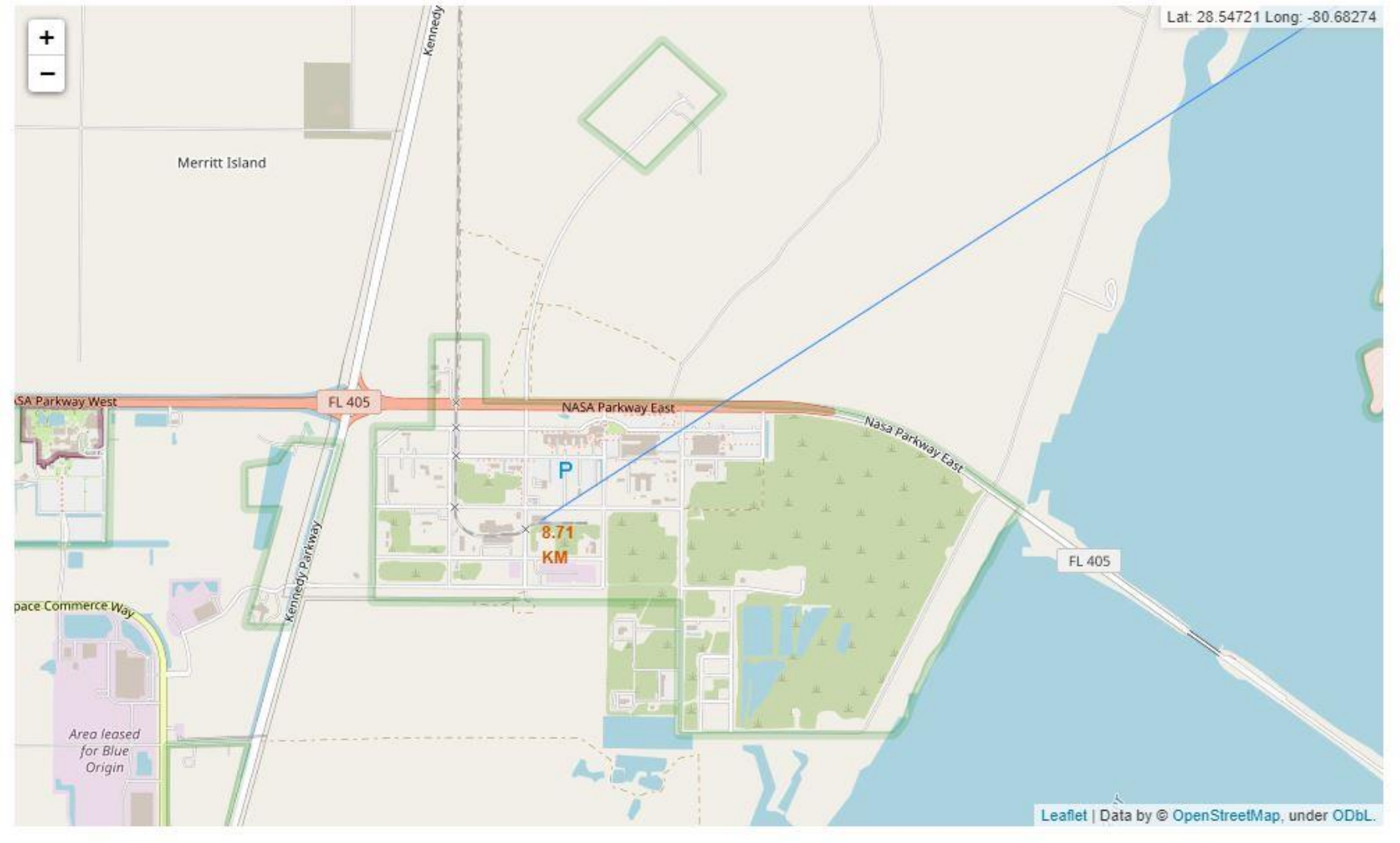
We can see that the SpaceX launch sites are in the United States of America coasts.

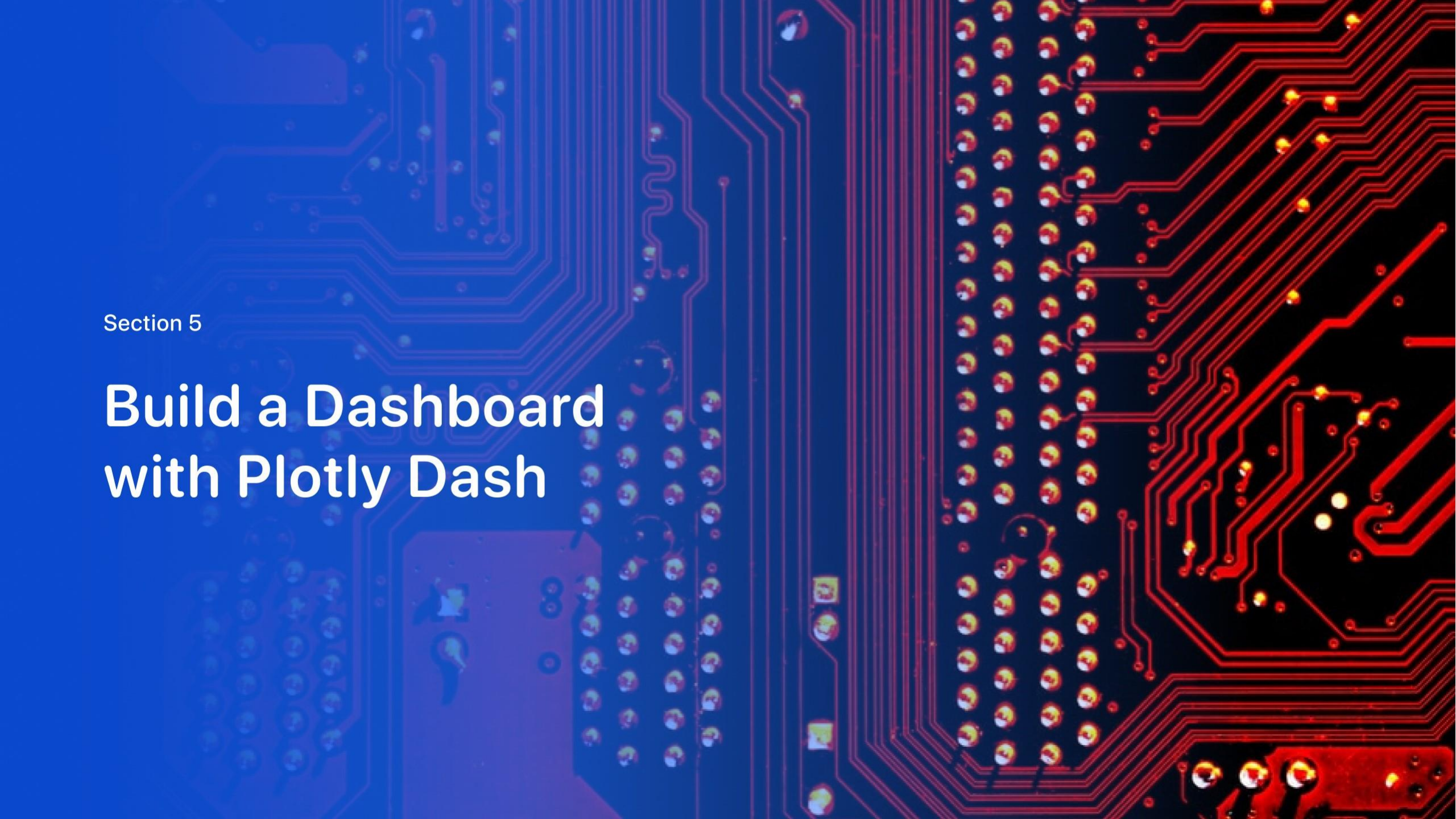
Colour Labelled Markers for Sites



Green Marker shows successful Launches and **Red Marker** shows Failures. The **Yellow Marker** shows the number of launches.

Launch Site distance to Railway Station

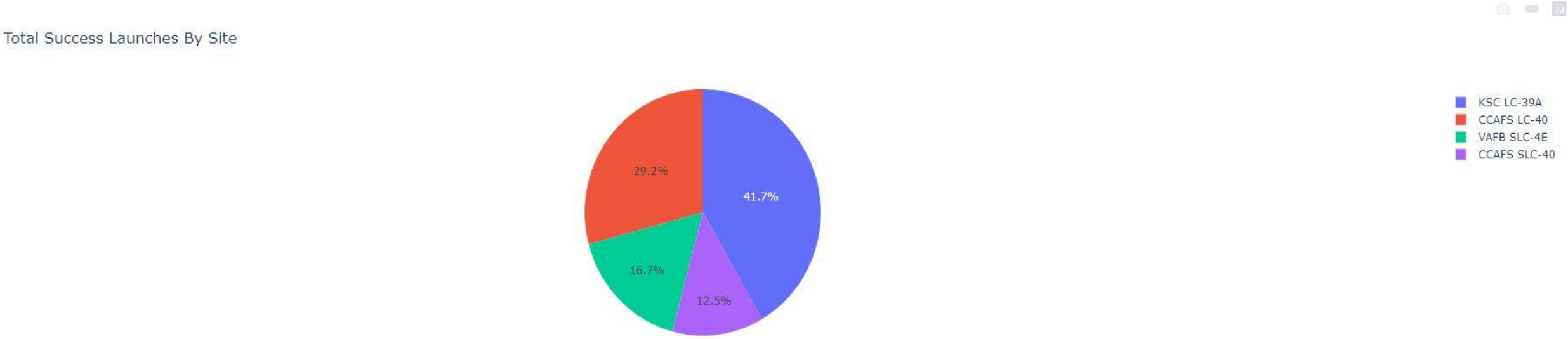




Section 5

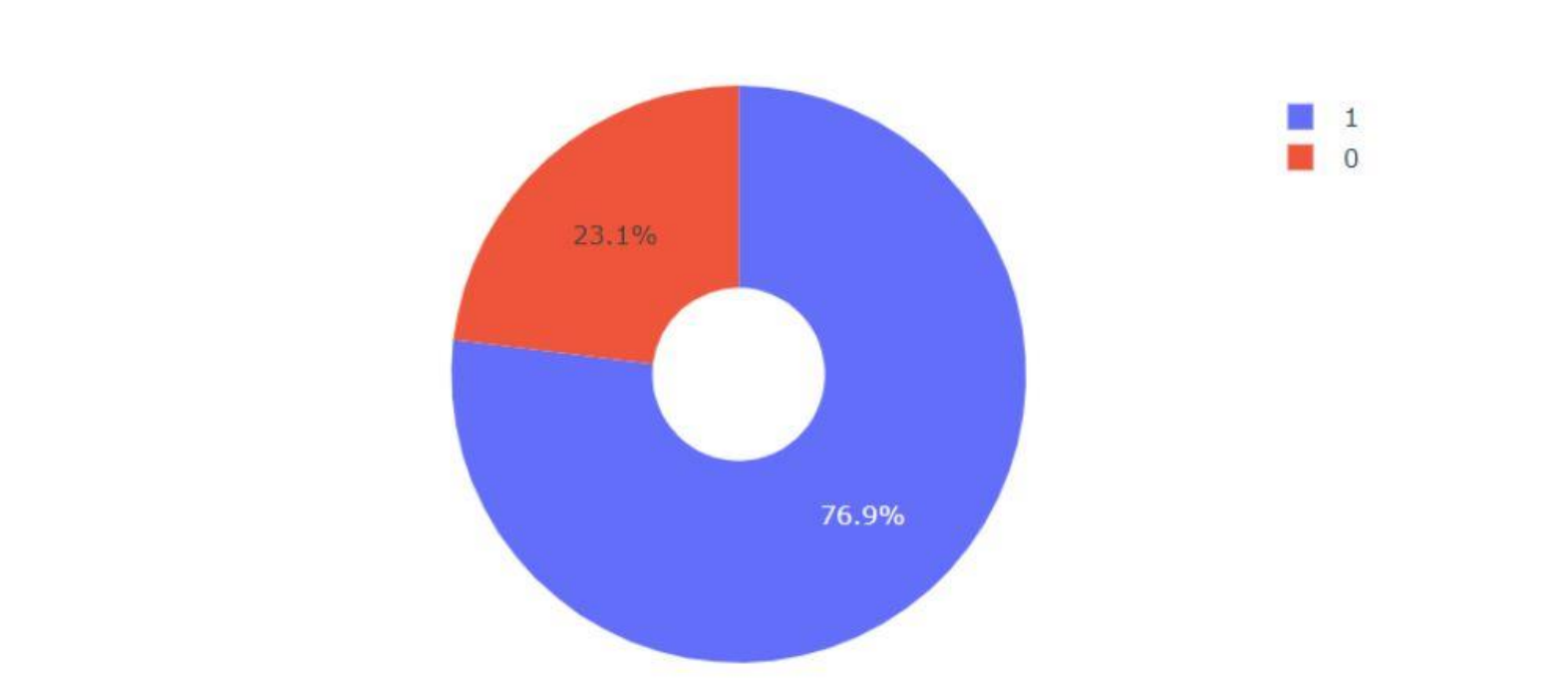
Build a Dashboard with Plotly Dash

Success percentage by Launch Site



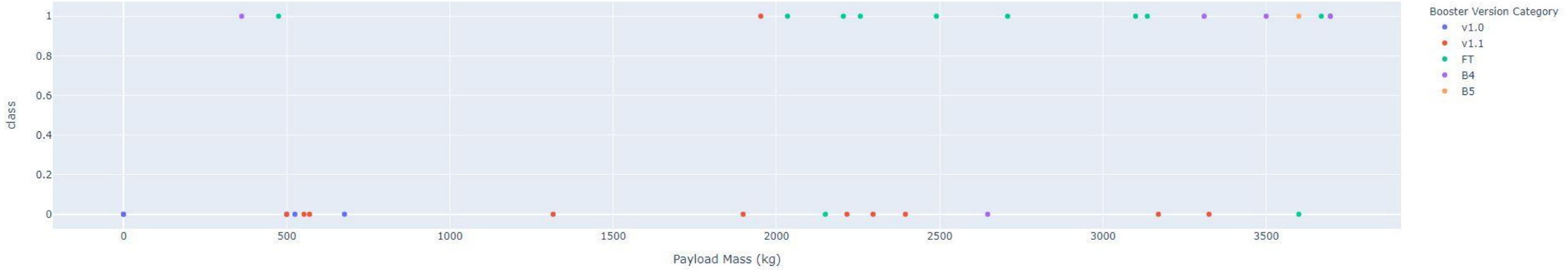
We can see that KSC LC-39A had the most successful launches from all the sites. A particular Launch Site can be selected on the right side to show results on for that particular Site.

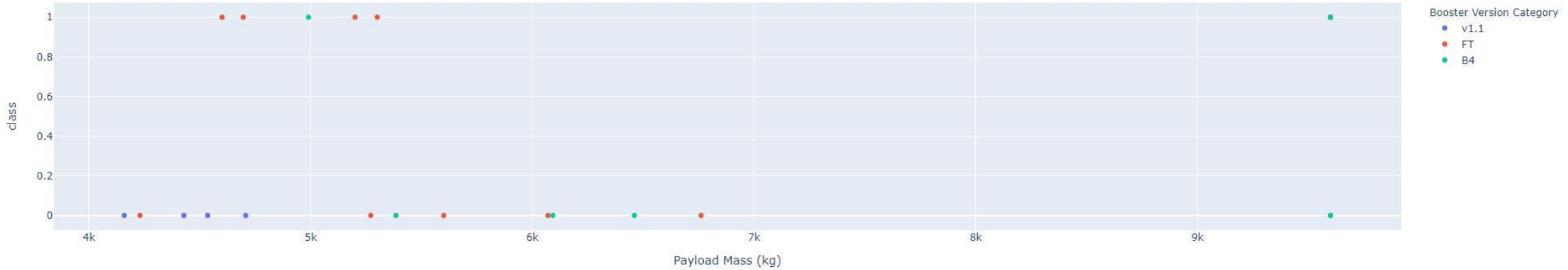
Success Ratio for KSC LC-39A



KSC LC-39A achieved a success rate of 76.9% while the failure rate was 23.1%

Payload against Launch Outcome





It can be observed that the success rate for low weighted payloads is higher than the heavy weighted payloads

Section 6

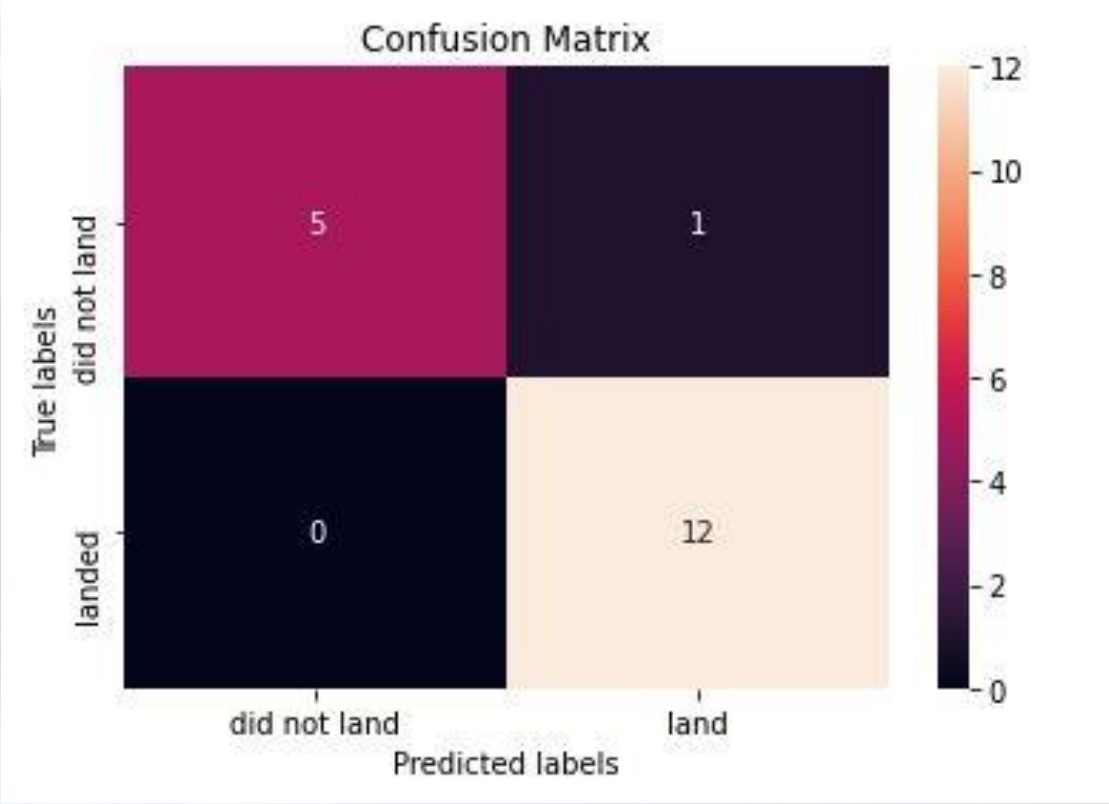
Predictive Analysis (Classification)

Classification Accuracy

	Algorithm	Accuracy
0	SVM	0.833333
1	Tree	0.944444
2	KNN	0.833333

After selecting the best hyperparameters for the decision tree classifier using the validation data, achieved 94.44% accuracy on the test data.

Confusion Matrix



Examining the confusion matrix, we see that Tree can distinguish between the different classes. We see that the major problem is false negatives.

Conclusions

- The Tree Classifier Algorithm is the best for Machine Learning for this dataset
- High weighted payloads perform better than the lower payloads
- The success rates for launches is positively related to time
- We can see that KSC LC-39A had the most successful launches from all the sites
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

Thank you!