Winning Space Race with Data Science

Jordan Prevatt
30/09/2022

Outline

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

Executive Summary

To complete this project, the methodologies used were:

- Data Collection via API and Web Scraping
- Data Wrangling
- Exploratory Data Analysis (EDA) using SQL, Pandas and Matplotlib
- Interactive Visual Analytics and Dashboard with Folium and Plotly Dash
- Predictive Analysis (Classification)

The results obtained in this project include:

- SpaceX's Falcon 9 Historical Flight Data
- Insights into Falcon 9's Historical Successes and Failures
- Predictions of the Success of Falcon 9's First Stage Landing

Introduction

SpaceX is currently perhaps the most successful rocket company. This is due to their relatively inexpensive rocket launches which are made possible by the company's reuse of the first stage. Thus, the success of the first stage's landing is the determining factor for the cost of the launch.

As a new rocket company, SpaceY, and a competitor to SpaceX, this project aimed to determine the price of each launch using public information gathered about SpaceX and determine whether SpaceX would reuse the first stage.

Section 1

Methodology

Methodology

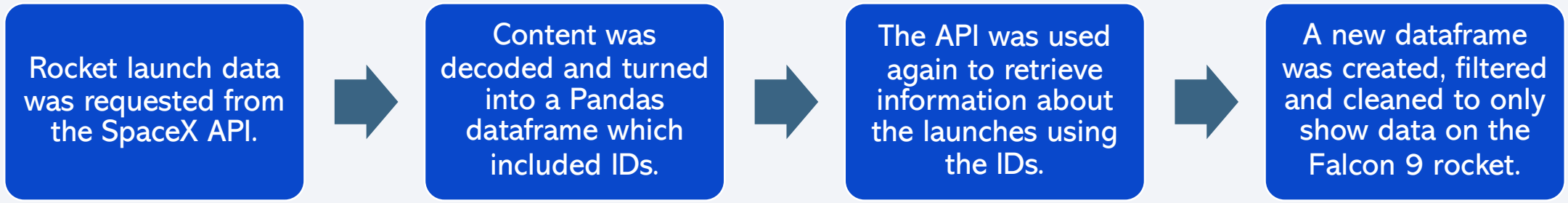
Executive Summary

- Data collection methodology:
  - Data was collected in two ways:
    1. Making a get request to the SpaceX API
    2. Web scraping historical launch records from a Wikipedia page
- Perform data wrangling
  - Data was processed by dealing with missing values, converting categorical data to numerical, and one hot encoding.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models were built, tuned, and evaluated

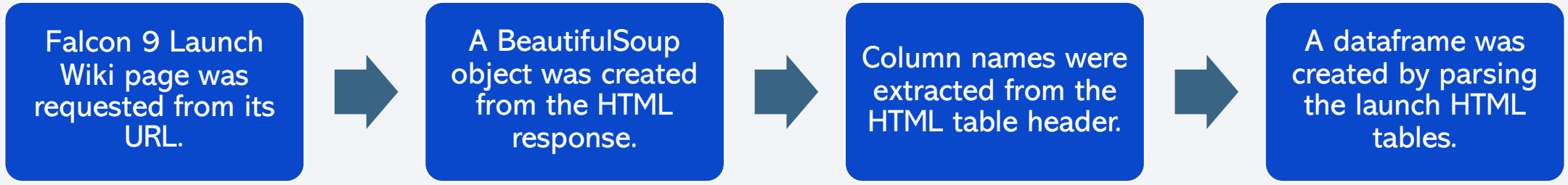
Data Collection

- The Datasets were collected in two ways:

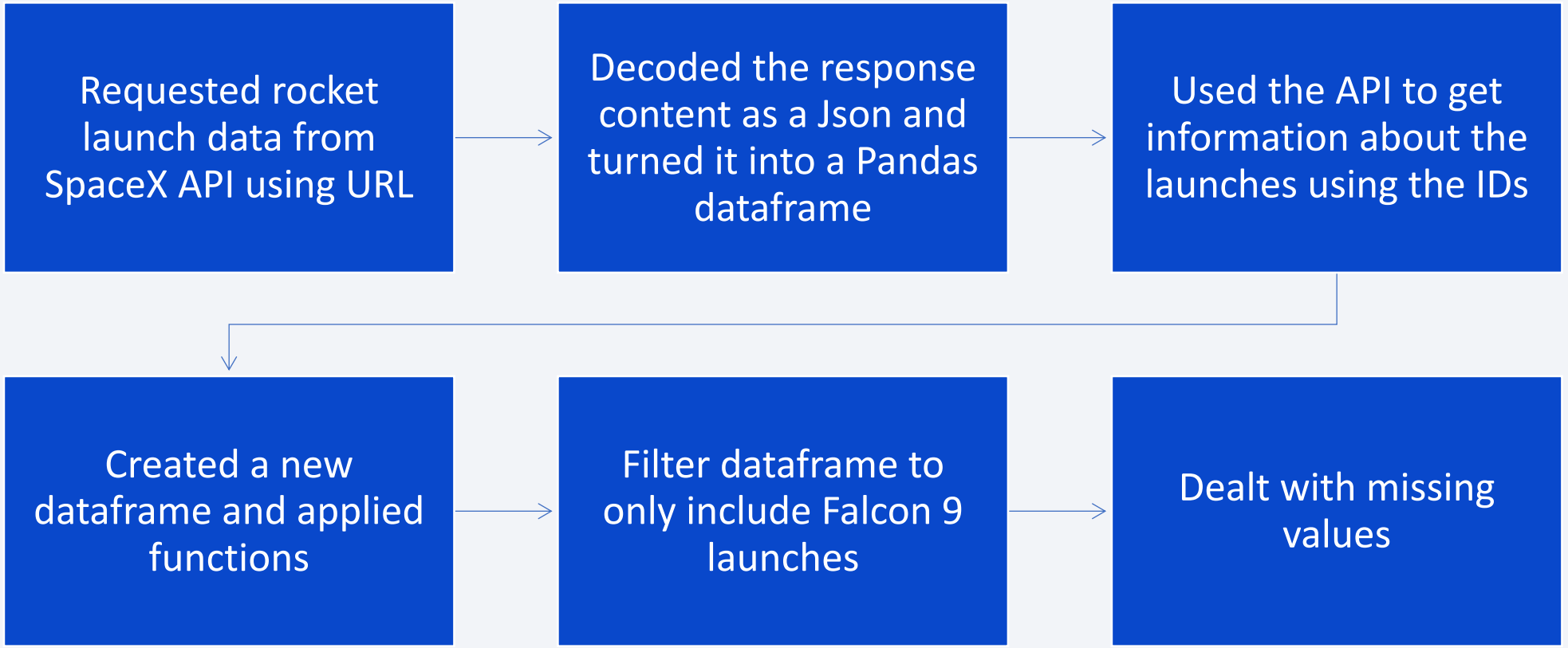
1. SpaceX API



2. Web Scraping

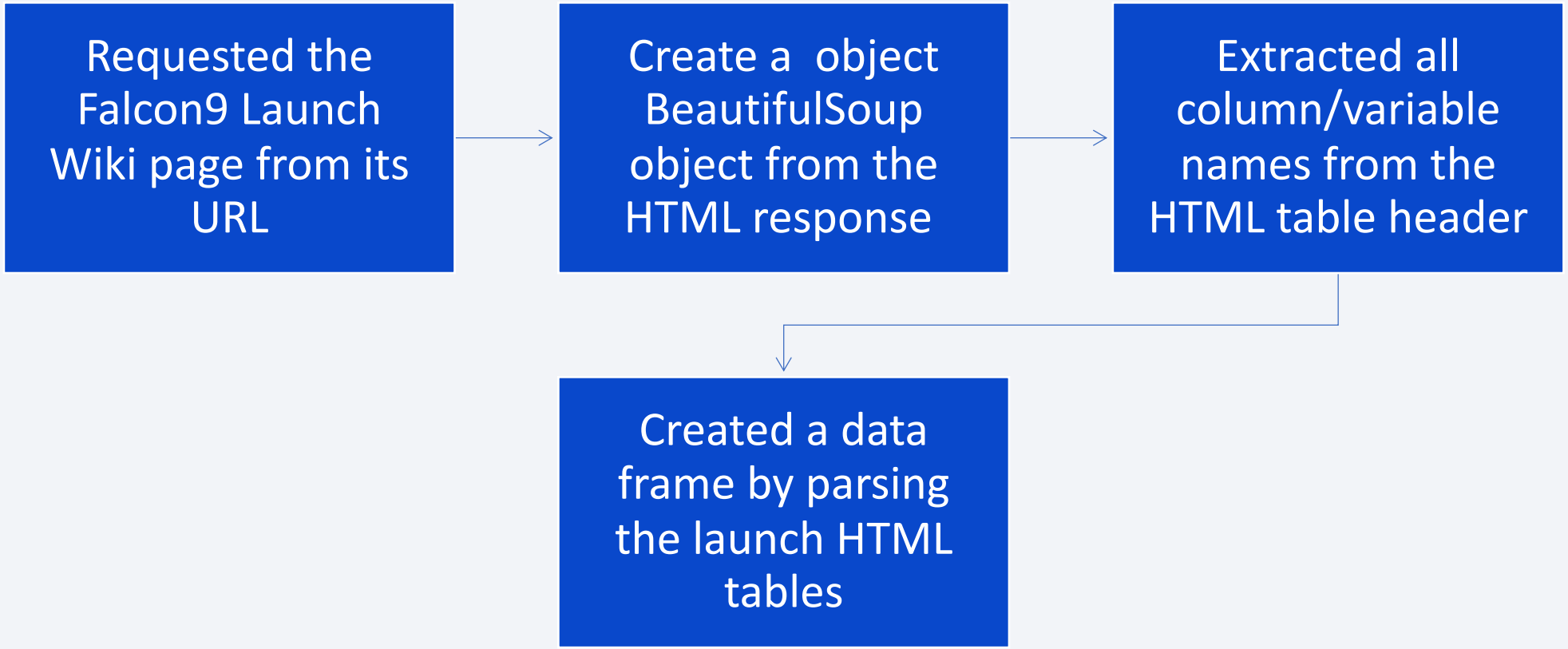


Data Collection – SpaceX API



https://github.com/jordanprev27/IBM_Applied_Data_Science_Capstone/blob/9212527c9c35a1e44e184093c1829ed538ef1427/Data_Collection_API.ipynb

Data Collection - Scraping



https://github.com/jordanprev27/IBM_Applied_Data_Science_Capstone/blob/9212527c9c35a1e44e184093c1829ed538ef1427/Webscraping.ipynb

Data Wrangling

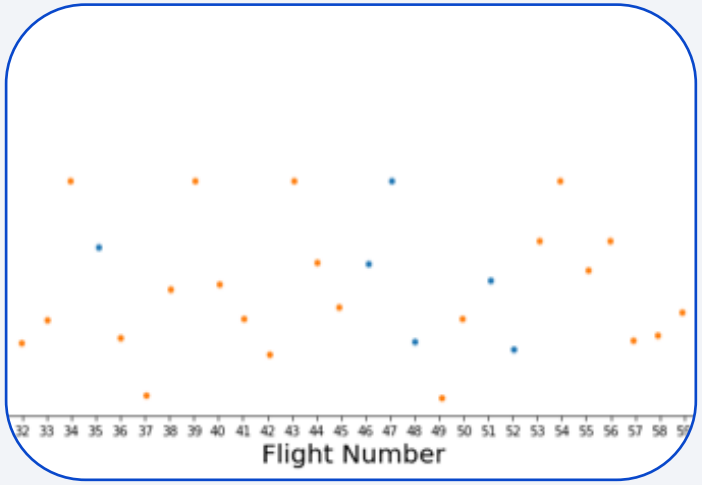
Missing values and categorical columns were identified.

A set of missions with unsuccessful outcomes was created.

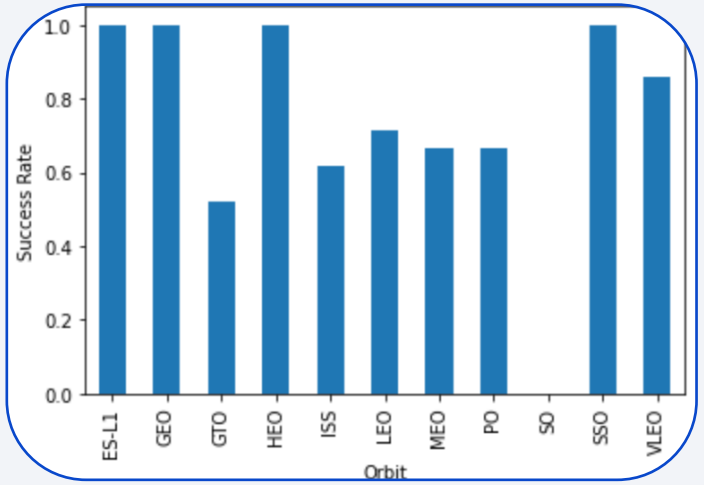
Landing outcome labels were created for the Outcome column using one hot encoding.

https://github.com/jordanprev27/IBM_Applied_Data_Science_Capstone/blob/930bd664078d31e6f92202a447e1600db3c4eacb/Data_Wrangling.ipynb

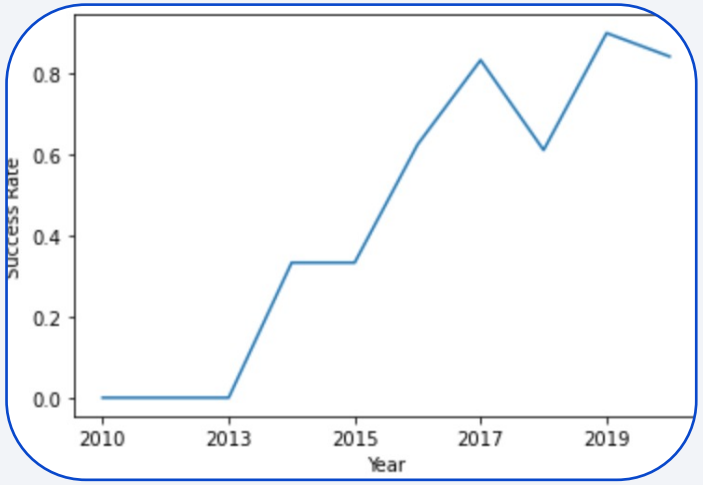
EDA with Data Visualization



Scatter Plots were plotted to visualize relationships between variables.



A bar chart was plotted to visualize the relationship between numerical and categorical data.



A line chart was plotted to visualize trends in numerical data over time.

https://github.com/jordanprev27/IBM_Applied_Data_Science_Capstone/blob/930bd664078d31e6f92202a447e1600db3c4eacb/EDA_Visualisation.ipynb

EDA with SQL

The SQL queries performed include:

- Displaying the names of unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015
- Ranking the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

https://github.com/jordanprev27/IBM_Applied_Data_Science_Capstone/blob/930bd664078d31e6f92202a447e1600db3c4eacb/EDA_SQL.ipynb

Build an Interactive Map with Folium

- Each launch site's location was added on a map using circles and markers with text labels to highlight the area.
- Success and failed launches for each site on the map were marked using marker clusters with green and red markers to simplify the map as many markers had similar coordinates.
- The distances from the nearest coast and a parkway to one of the launch sites were displayed using a marker and lines were drawn to connect the launch site to the points.

https://github.com/jordanprev27/IBM_Applied_Data_Science_Capstone/blob/930bd664078d31e6f92202a447e1600db3c4eacb/Interactive_Visual_Analytics_and_Dashboard.ipynb

Build a Dashboard with Plotly Dash

- A launch site dropdown menu was added to allow selection of different launch sites.
- A pie chart visualizing launch success counts based on the selected site was rendered by adding a callback function.
- A range slider was added to select different payload ranges.
- A scatter chart visualizing the correlation between payload and success based on the selected site was rendered by adding a callback function.

https://github.com/jordanprev27/IBM_Applied_Data_Science_Capstone/blob/930bd664078d31e6f92202a447e1600db3c4eacb/spacex_dash_app.py

Predictive Analysis (Classification)

Building

- Variable Y was assigned from the Class column and X was loaded from a CSV file then standardized.
- The data X and Y was split into training and testing data.
- The training data was used to fit the models

Evaluating

- The accuracy of the model on test data was calculated.
- The confusion matrix was created.

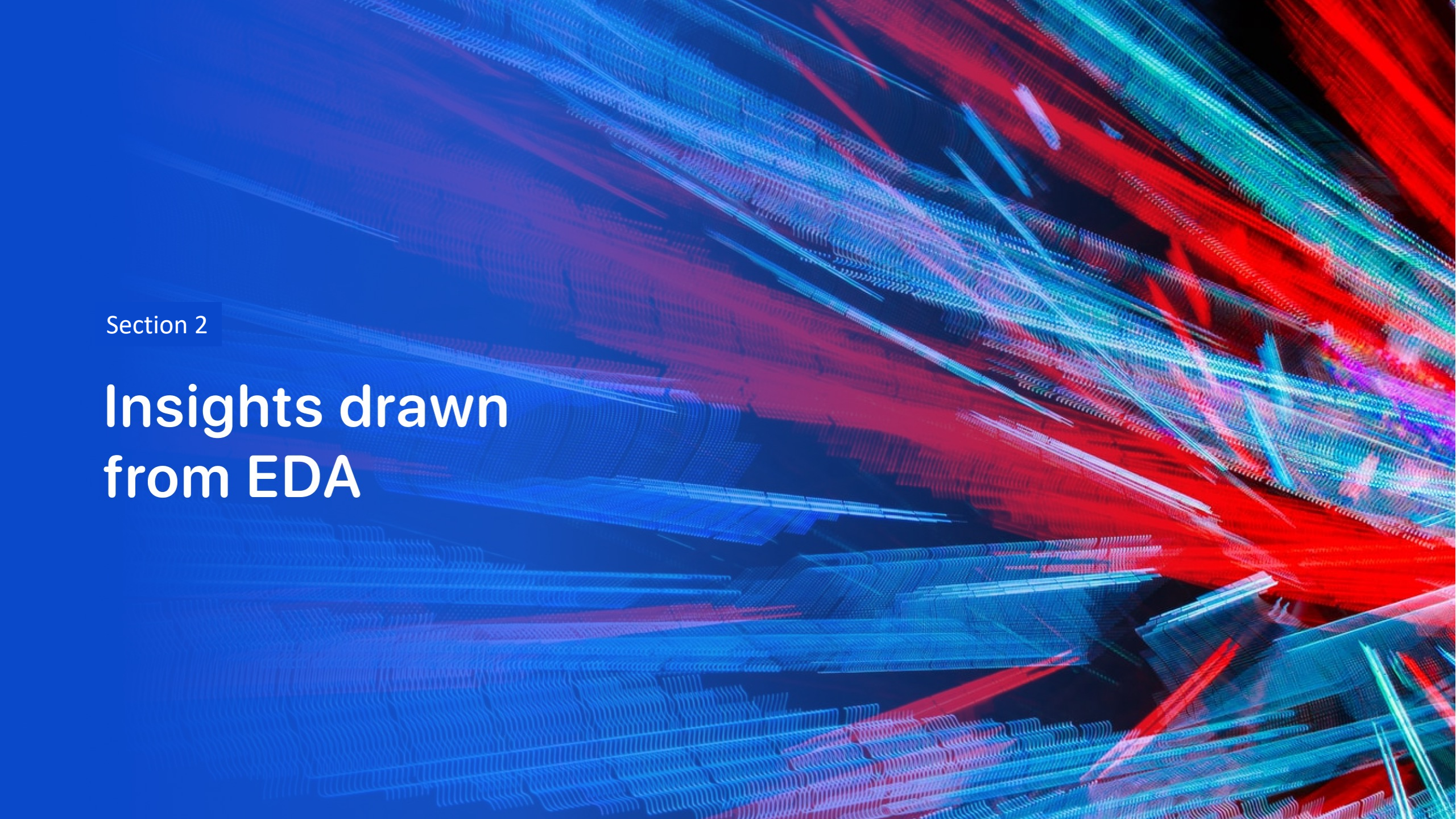
Finding the Best

- The accuracy for each model was evaluated to find the one with the best performance.

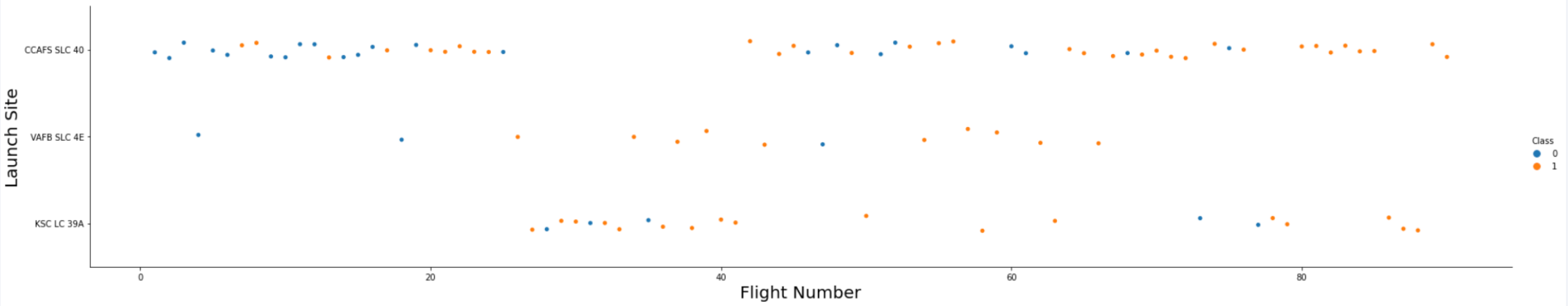
https://github.com/jordanprev27/IBM_Applied_Data_Science_Capstone/blob/930bd664078d31e6f92202a447e1600db3c4eacb/Machine_Learning.ipynb

Results

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

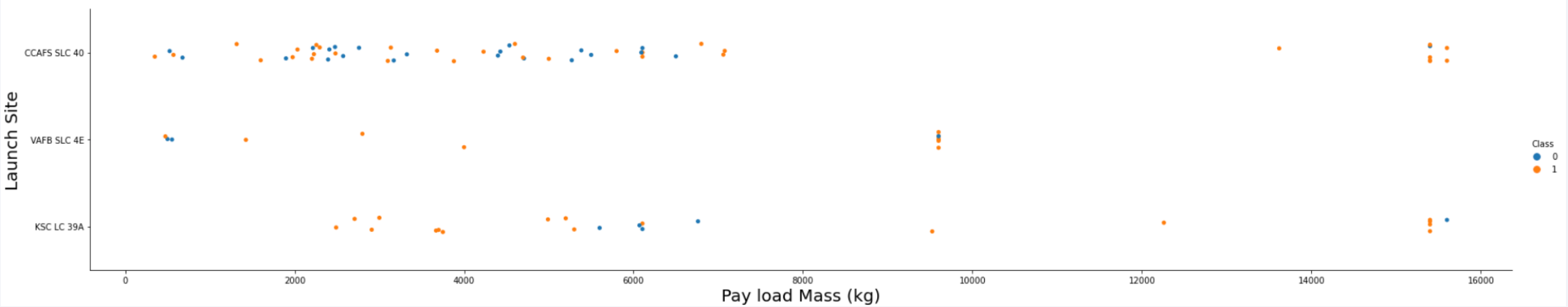

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



Launch site CCAFS SLC 40 has the most flights while VAFB SLC 4E has the least. The success rates vary between the launch sites.

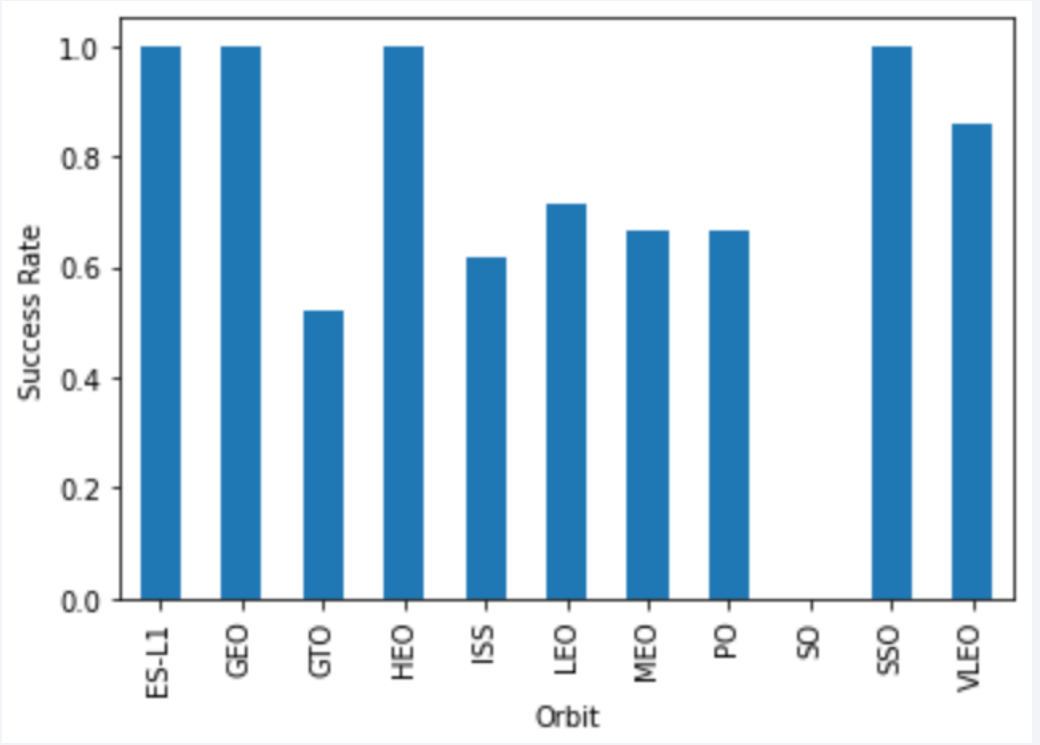
Payload vs. Launch Site



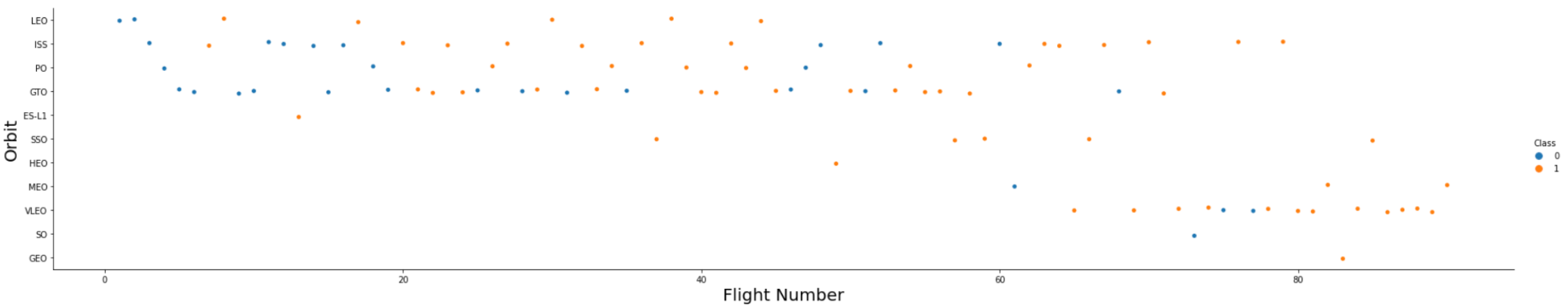
There are no rockets launched from VAFB SLC 4E with a payload mass greater than 10,000kg.

Success Rate vs. Orbit Type

The ES-L1, GEO, HEO, and SSO orbits have the highest success rates followed by VLEO and LEO.

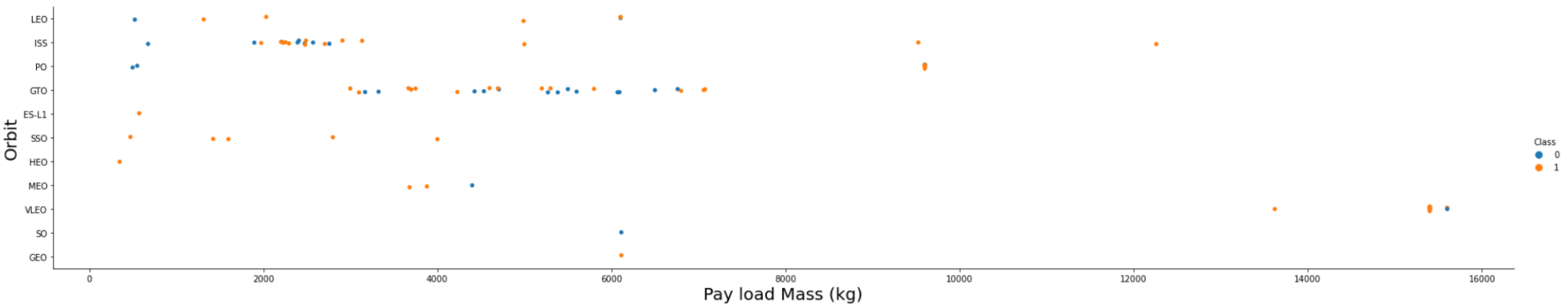


Flight Number vs. Orbit Type



Success is related to the number of flights in the LEO orbit. No relationship can be determined for the GTO orbit.

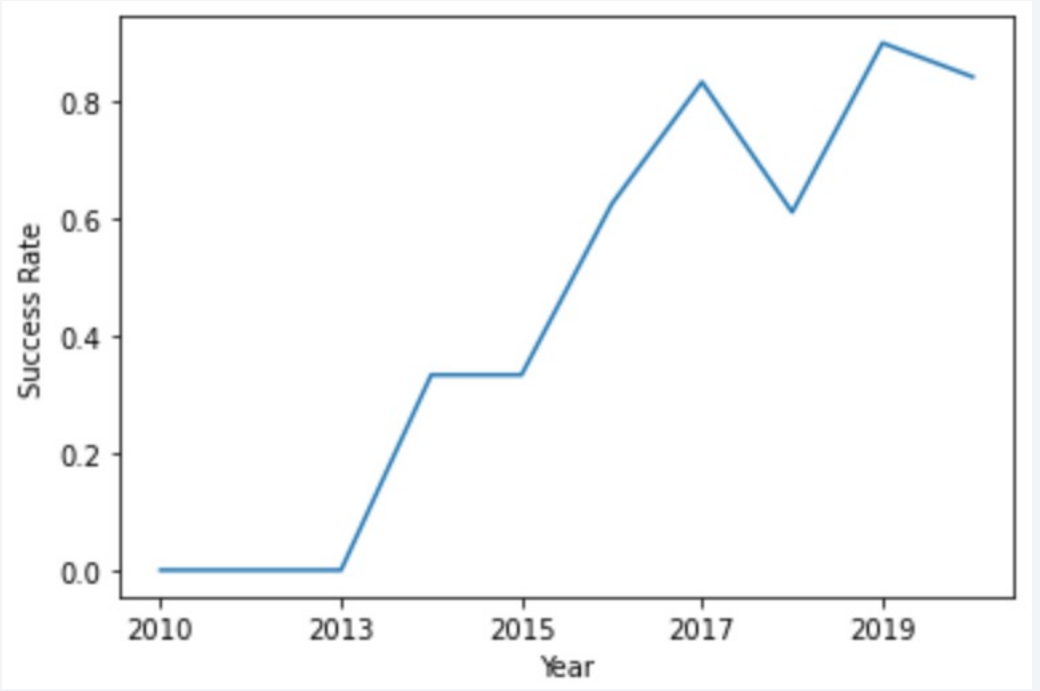
Payload vs. Orbit Type



The ISS, PO and VLEO orbit have had successes in landing heavy payloads. However, the GTO orbit has had mixed results.

Launch Success Yearly Trend

The success rate has increased beginning in 2013 and continuing up to 2020.



All Launch Site Names

The unique launch sites are:

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

```
%%sql
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

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

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

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

These are the first 5 records with launch sites beginning with `CCA`

Total Payload Mass

The total payload carried by boosters from NASA is 45,596kg.

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

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2,928.4kg.

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad was 01-05-2017.

```
%%sql
SELECT MIN(DATE) FROM SPACEXTBL
WHERE "LANDING _OUTCOME" = 'Success (ground pad)';
```

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

MIN(DATE)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.

```
%%sql
SELECT BOOSTER_VERSION, "LANDING _OUTCOME" FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
AND "LANDING _OUTCOME" = 'Success (drone ship)';
```

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

Booster_Version	Landing _Outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT (SELECT COUNT(*) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%') AS "No of Successes",
(SELECT COUNT(*) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Failure%') AS "No of Failures";
```

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

No of Successes	No of Failures
100	1

The total number of successful and failure mission outcomes are 100 and 1 respectively.

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT 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

Listed are the names of the booster which have carried the maximum payload mass.

2015 Launch Records

```
%%sql
SELECT SUBSTR(Date,4,2) AS Month, "LANDING _OUTCOME", BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL
WHERE DATE LIKE '%2015'
AND "LANDING _OUTCOME" = 'Failure (drone ship)';
```

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

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

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are listed as shown above.

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

The rank of 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 is Success, Success (drone ship), then Success (ground pad).

```
%%sql
SELECT "LANDING _OUTCOME", COUNT(*) FROM SPACEXTBL
WHERE "LANDING _OUTCOME" LIKE 'Success%'
AND DATE BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY "LANDING _OUTCOME"
ORDER BY COUNT(*) DESC ;
```

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

Landing _Outcome	COUNT(*)
Success	20
Success (drone ship)	8
Success (ground pad)	6

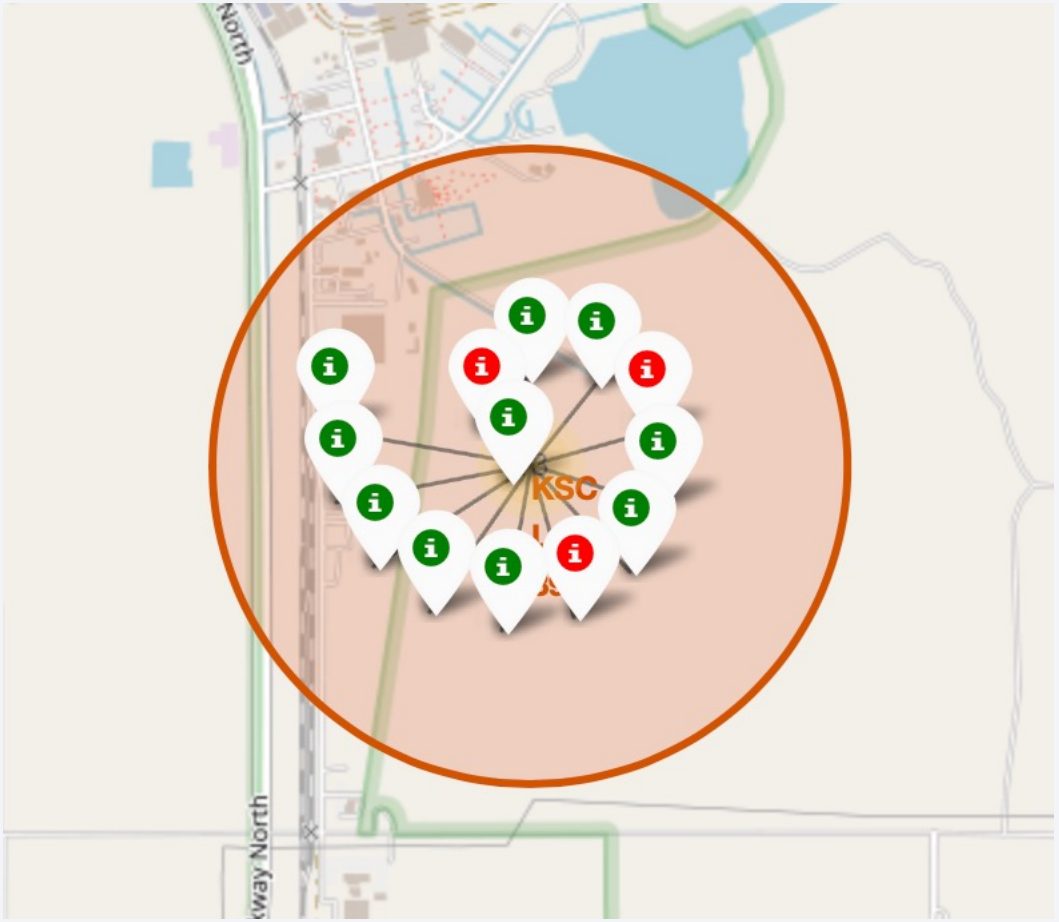

Section 3
Launch Sites
Proximities Analysis

All Launch Sites on Map

The generated folium map includes all launch sites' location markers on a global map.

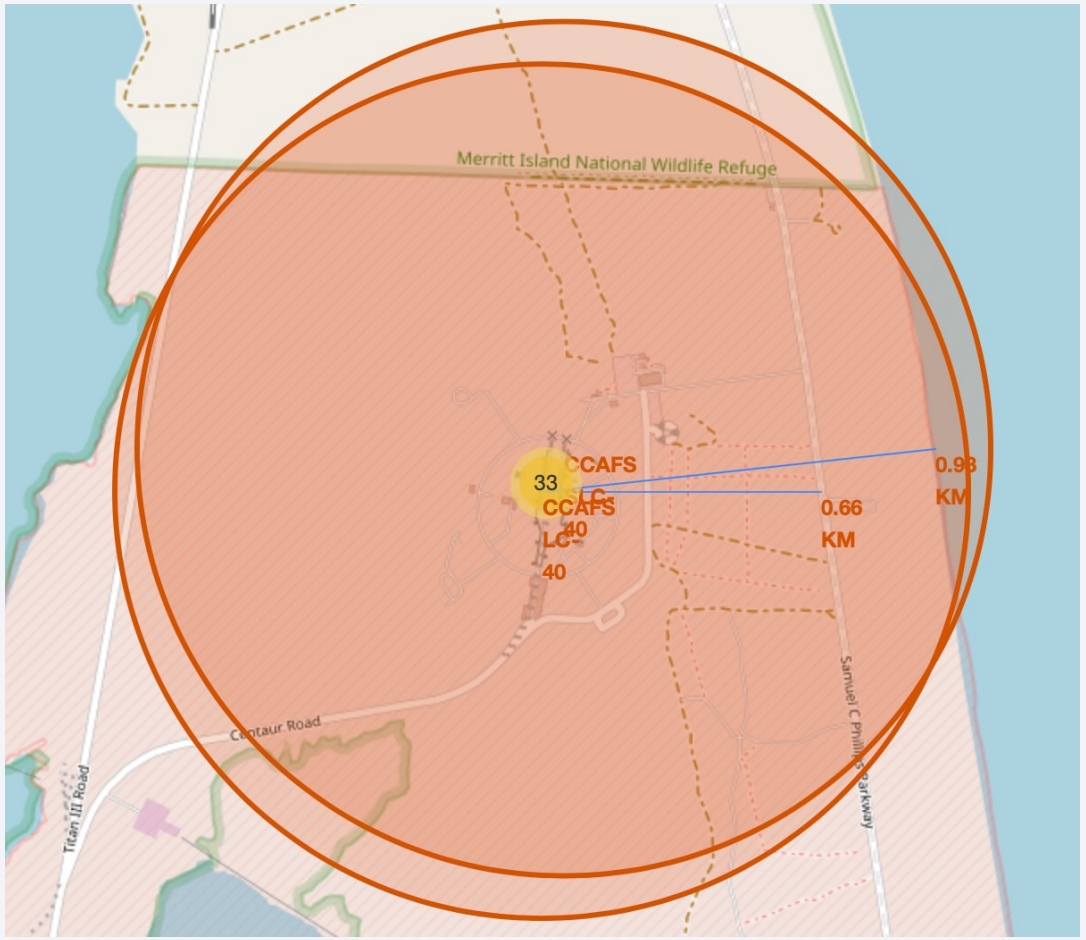
Launch Outcomes at KSC LC-39A

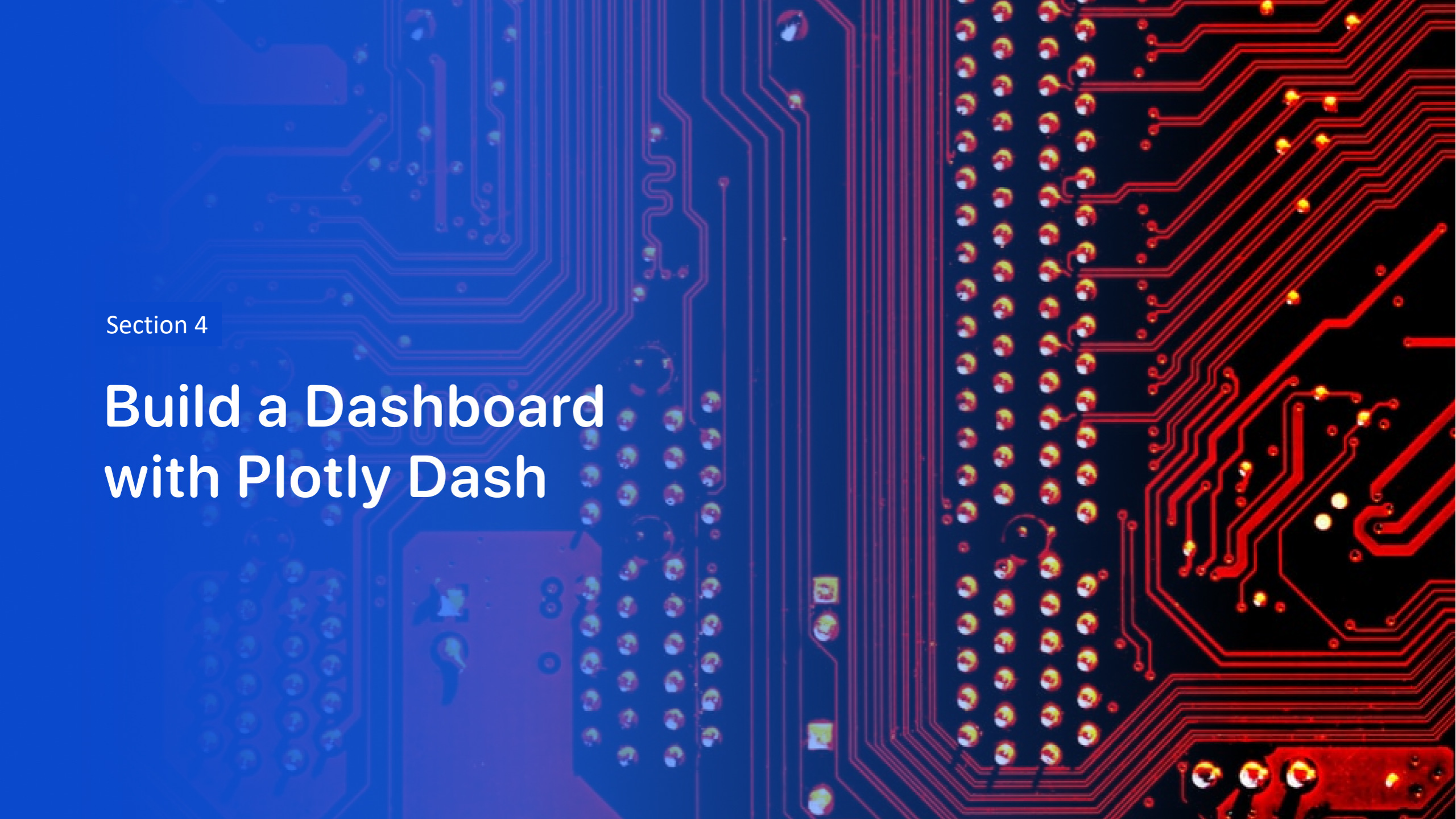
The folium map shows the color-labeled launch outcomes at KSC LC-39A which appears to have a high rate of success.



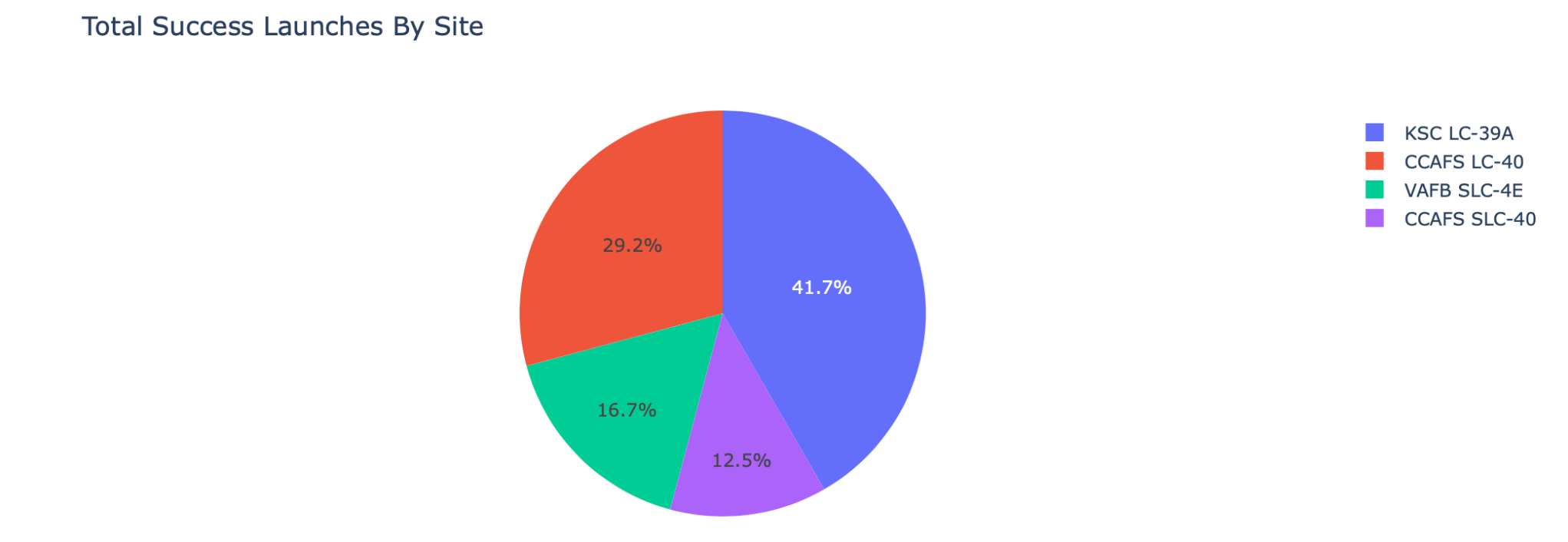
Distances Between CCAFS LC-40 to Its Proximities

The generated folium map shows the launch site CCAFS LC-40 and its proximities such as a parkway and the coastline with distance calculated and displayed.



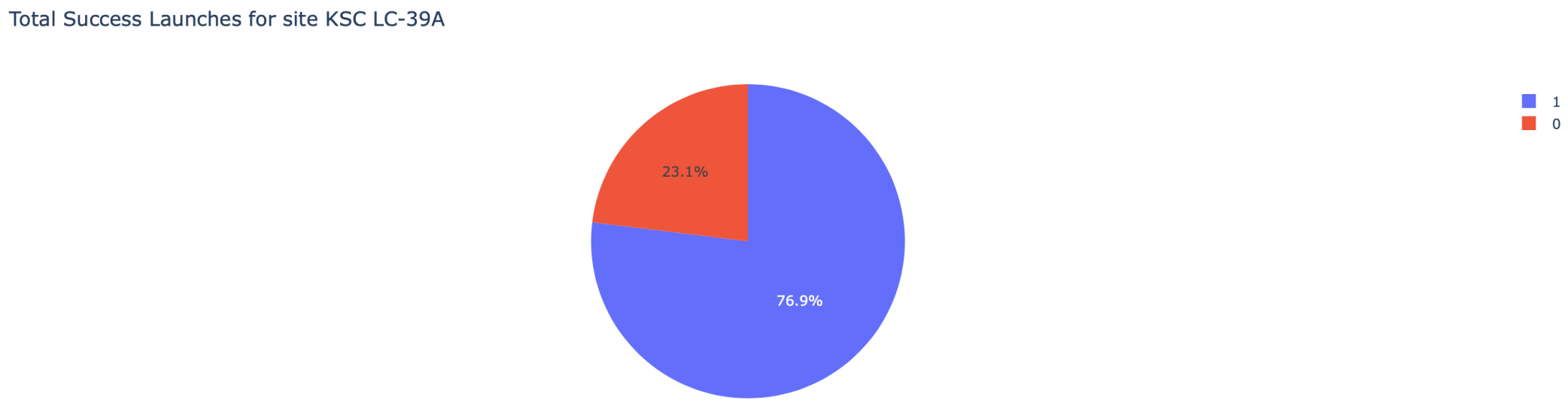

Section 4
Build a Dashboard with Plotly Dash

Launch Success for All Sites



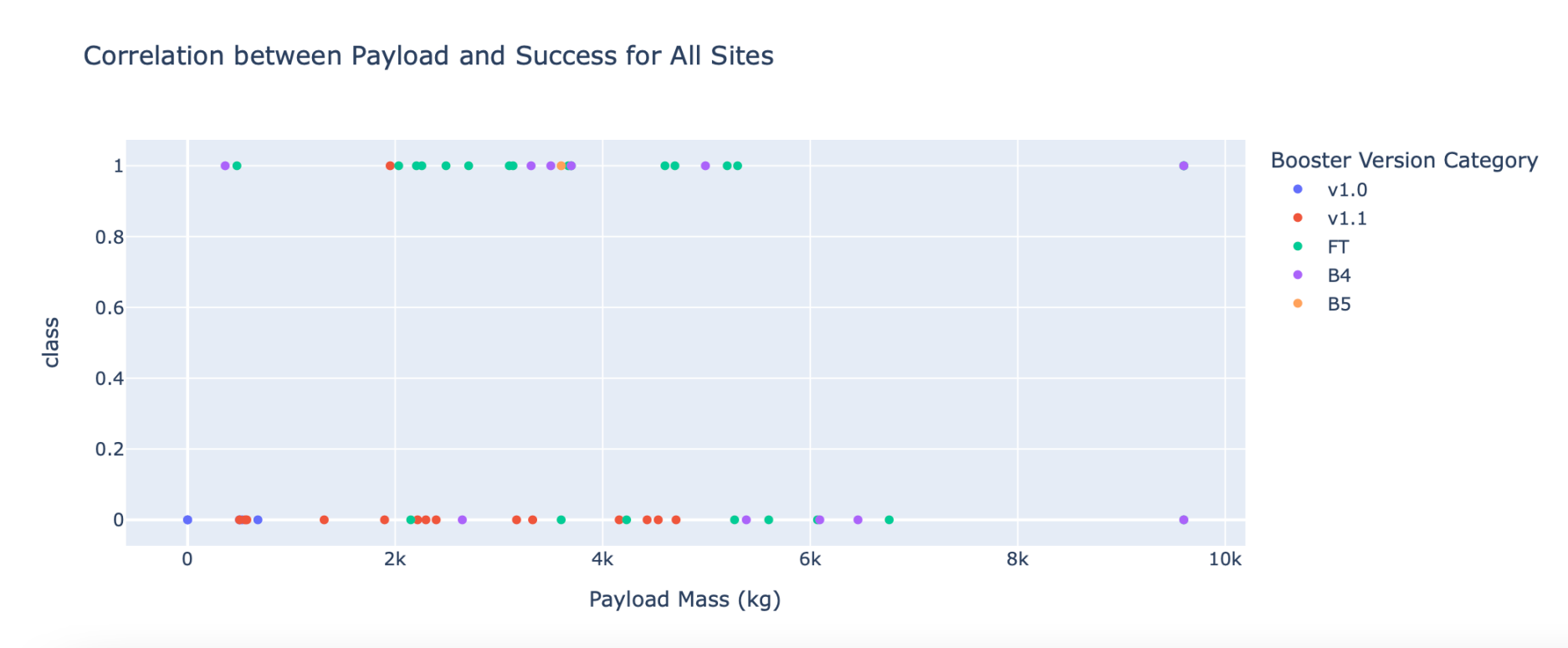
Launches from KSC LC-39A are the most successful making up the majority of the total successes. On the other hand, launches from CCAFS SLC-40 are the least successful.

Launch Success for KSC LC-39A



The success rate for KSC LC-39A is 76.9%

Payload vs. Launch Outcome



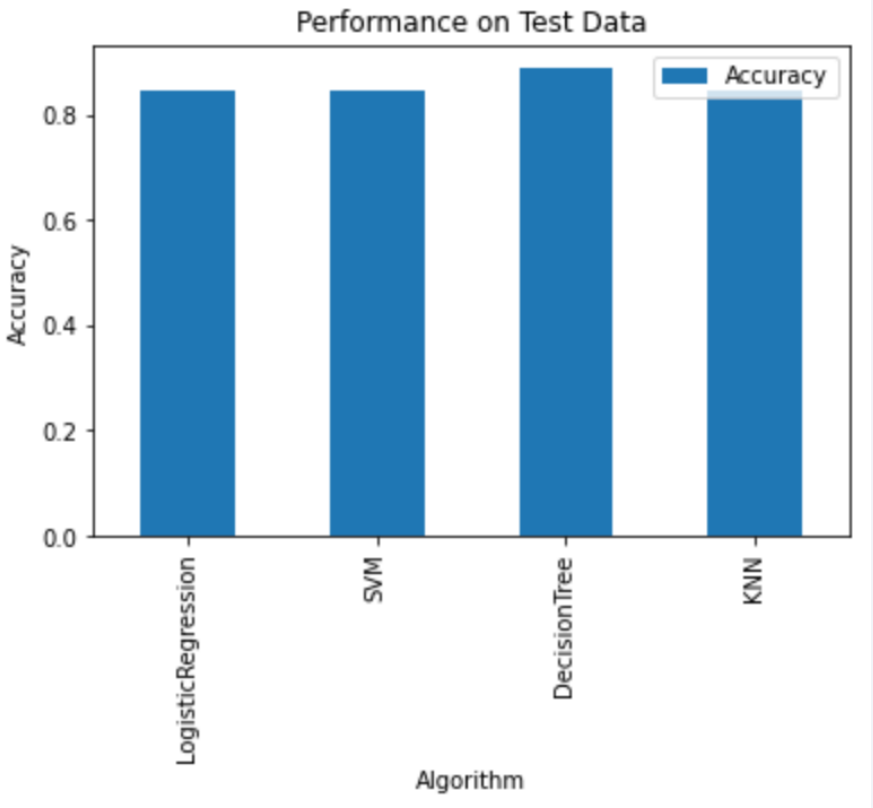
The payload range of 2,000 to 4,000kg and the booster version category of FT have the largest success rate.

Section 5

Predictive Analysis (Classification)

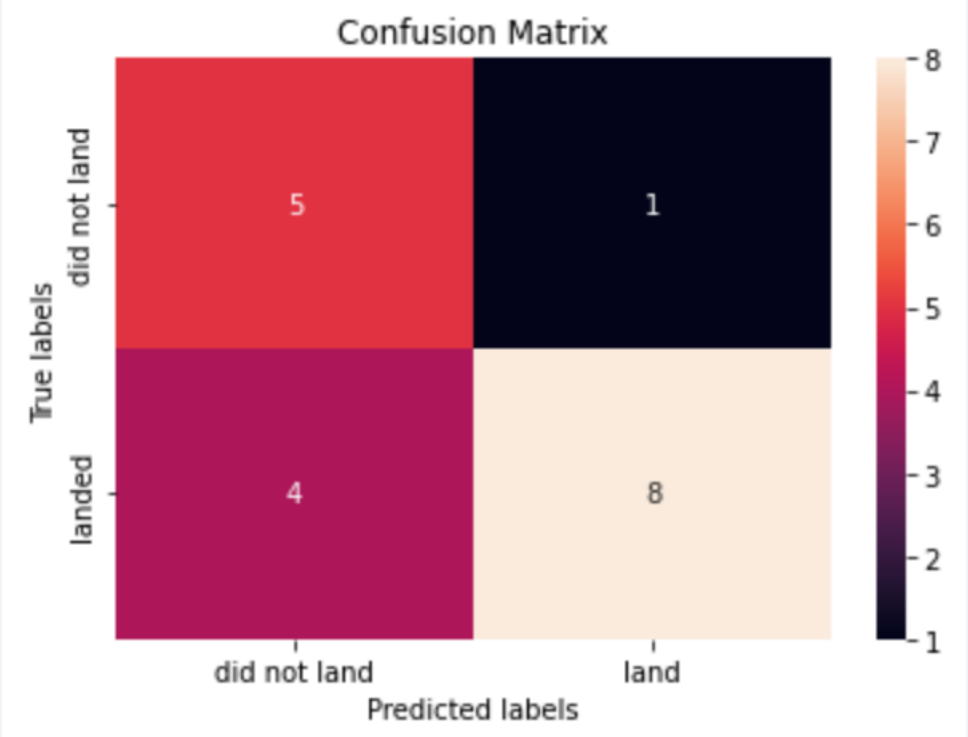
Classification Accuracy

The Decision Tree model has the highest classification accuracy.



Confusion Matrix

The confusion matrix of the Decision Tree model shows that of the predictions, 5 were true positives, 1 was a false negative, 4 were false positives, and 8 were true negatives.



Conclusions

- SpaceX Falcon 9 historical launch data was obtained using an API and Web Scraping.
- Most of flights come from CCAFS SLC-40.
- The success rate of launches is increasing.
- KSC LC-39A has the highest rate of success.
- The payload range of 2,000 to 4,000kg and the booster version category of FT have the largest success rate.
- The Decision Tree model had the highest classification accuracy

Thank you!