Winning Space Race with Data Science

Ashwiin Nedun

08 September 2022

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Build an Interactive Map with Folium
  - Build 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

- Space X promotes Falcon 9 rocket launch for a fee of 62 million dollars. At the same time, their competitors are charging more than 165 million dollars for each launch. The significant amount of savings coming from Space X is due to its ability to reuse the first stage.
- We can calculate the launch cost if we can predict whether the first stage will land. This information can be utilized if a competing firm wants to compete with SpaceX for a rocket launch contract.
- The goal is to eventually determine whether or not the Space X Falcon 9 first stage can land successfully.

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - Using Space X API and Web Scraping from List of Falcon 9 and Falcon Heavy Launches
- **Perform data wrangling**
  - Replacing NaN values with mean for the specific columns
  - Creating a landing outcome label showing the booster did or did not land successfully
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models (SVM, Classification Trees, and Logistic Regression)**

Data Collection

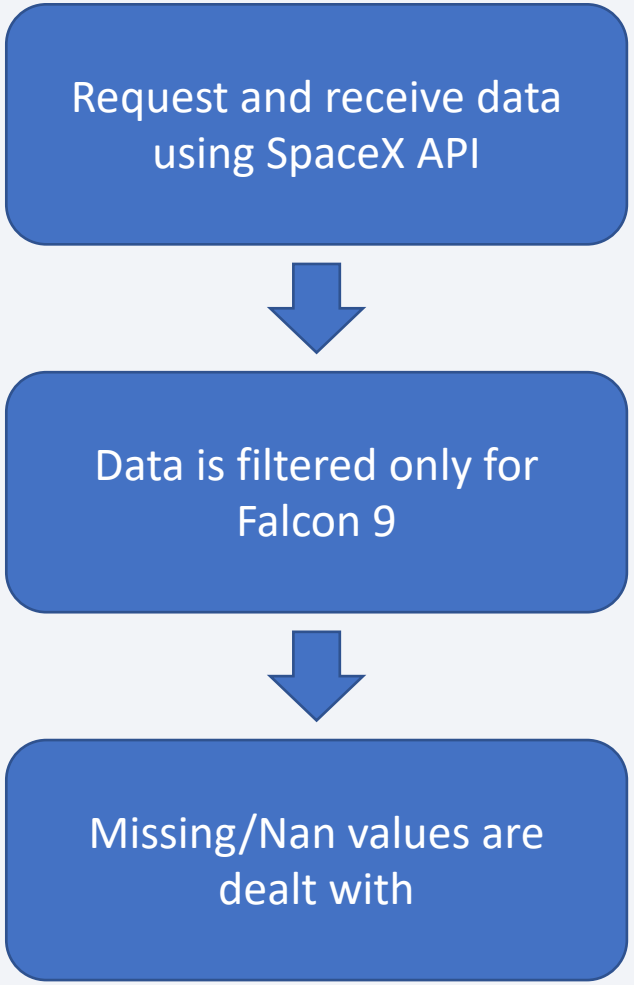
- The technique of collecting data involves web scraping data from a table on the SpaceX, Falcon 9 and Falcon Heavy Launches Records page on Wikipedia in addition to making API queries to the SpaceX API.
- API collected was from (https://api.spacexdata.com/v4/rockets/) and web scraping was done on (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- SpaceX provides a public API via which data may be retrieved and utilized. Using this API according to the accompanying flowchart, data is then persisted.

- The link to the notebook is :

  Data Collection - API.ipynb

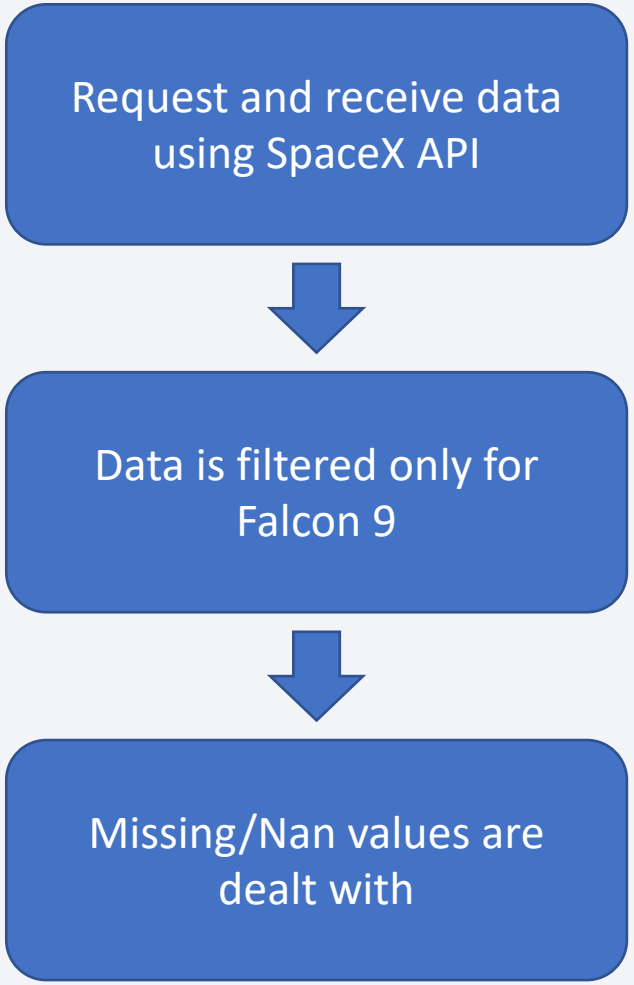


Data Collection - Scraping

- Wikipedia also has information on SpaceX's launches. According to the flowchart, data are obtained from Wikipedia and then persisted.

- The link to the notebook is :

  Data Collection - Web Scraping.ipynb



Data Wrangling

- The dataset was initially analyzed using Exploratory Data Analysis (EDA).
- Summaries of launches by site, orbits, and mission outcomes per orbit type were then determined.
- Finally, the outcome label for the landing page was generated using the Outcome column.

- The link to the notebook is Data Wrangling.ipynb

EDA with Data Visualization

- **Scatter charts** were produced to visualize the relationships between:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Orbit Type and Flight Number
  - Payload and Orbit Type
- Scatter charts are useful to observe relationships, or correlations, between two numeric variables.
- A **bar chart** was produced to visualize the relationship between: Success Rate and Orbit Type
- Bar charts are used to compare a numerical value to a categorical variable. Horizontal or vertical bar charts can be used, depending on the size of the data.
- **Line charts** were produced to visualize the relationships between: Success Rate and Year
- Line charts contain numerical values on both axes, to show the change of a variable over time.

EDA with SQL

- The following SQL queries were 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 achieved.
  - 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.
  - Failed landing_outcomes in drone ship, their booster versions, and launch site names for in 2015
- The link to the notebook is Exploratory Data Analysis - SQL.ipynb

Build an Interactive Map with Folium

- Folium Maps made use of markers, circles, lines, and clusters of markers.
- Markers indicate points like launch sites
- Circles represent highlighted regions surrounding specified coordinates, such as NASA Johnson Space Center.
- The distance between two coordinates is shown by lines.
- Clusters of markers represent groups of occurrences in each coordinate, such as launches at a launch site.

- The link to the notebook is Exploratory Data Analysis - Data Visualization.ipynb

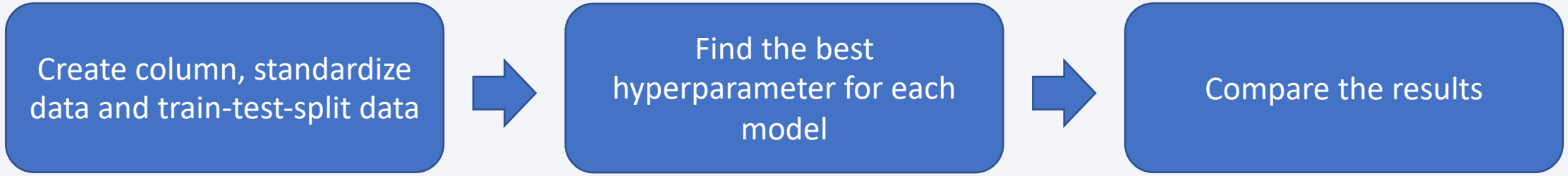
Build a Dashboard with Plotly Dash

- In Plotly Dashboard:
  - Pie charts and Scatter charts were used to show the total successful launches count for all sites and the correlation between payload and launch success, respectively.
  - Dropdown list and slider were used to enable launch site selection and payload range section, respectively

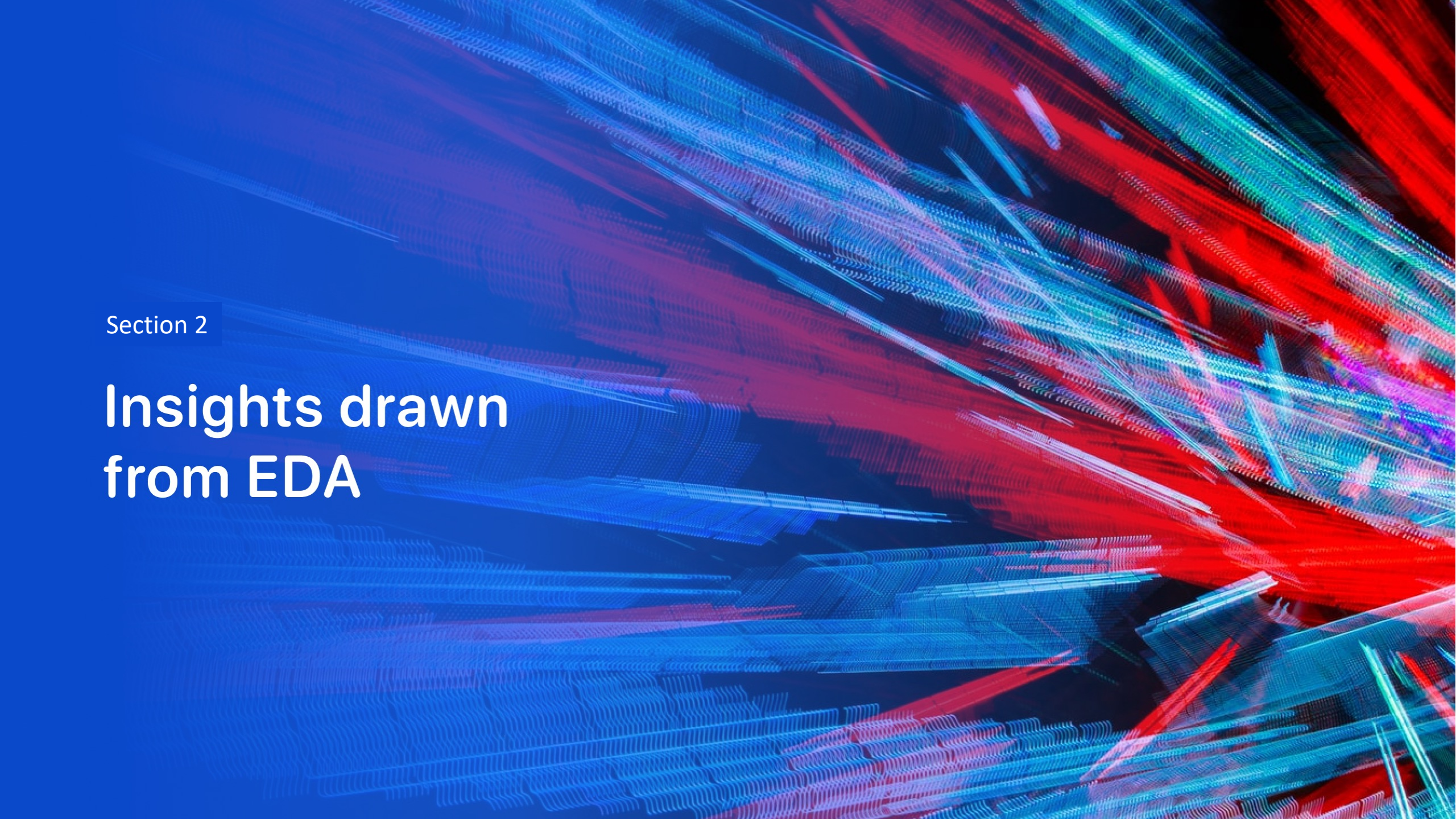
- The link to the notebook is spacex_dash_app.py

Predictive Analysis (Classification)

- A total of 4 classification models were used for comparison. They were logistic regression, support vector machine, decision tree and k nearest neighbors.

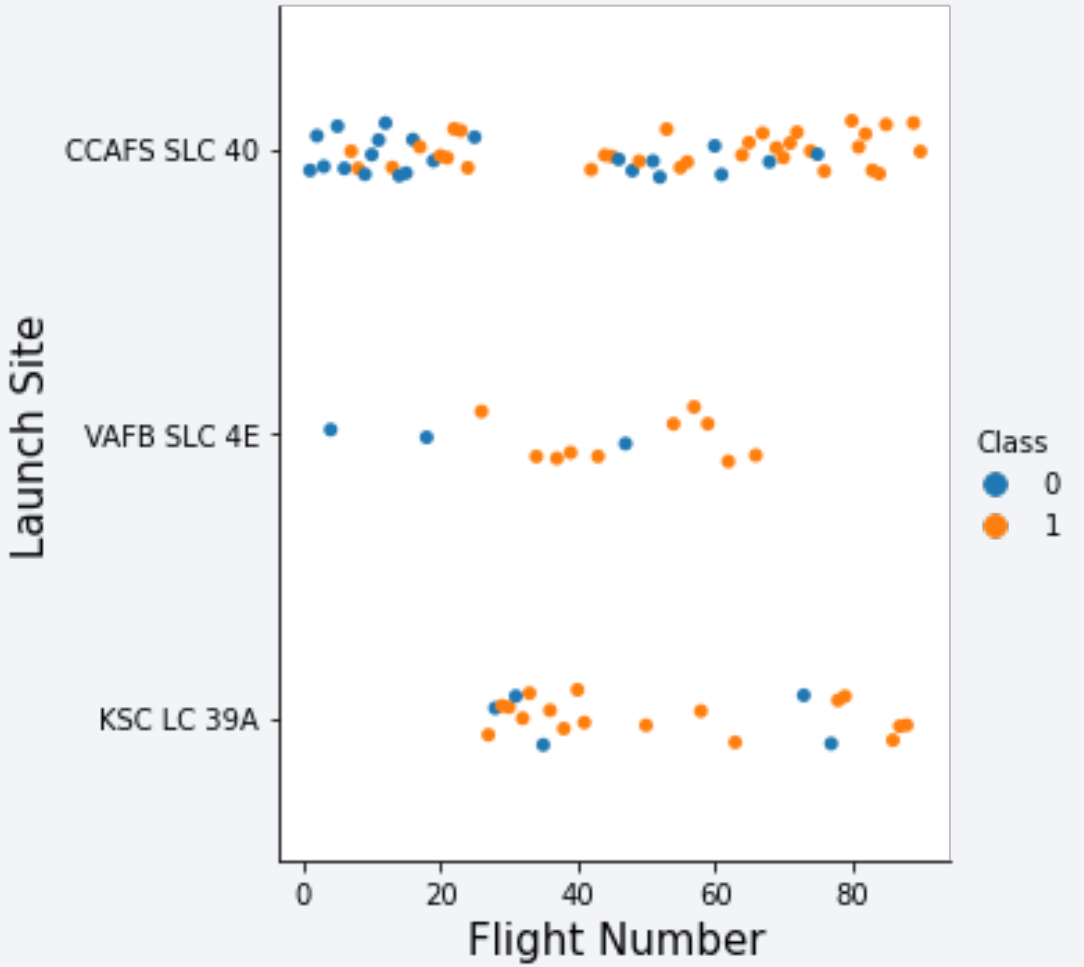


- The link to the notebook is Predictive Analysis (Classification).ipynb


Section 2
Insights drawn from EDA

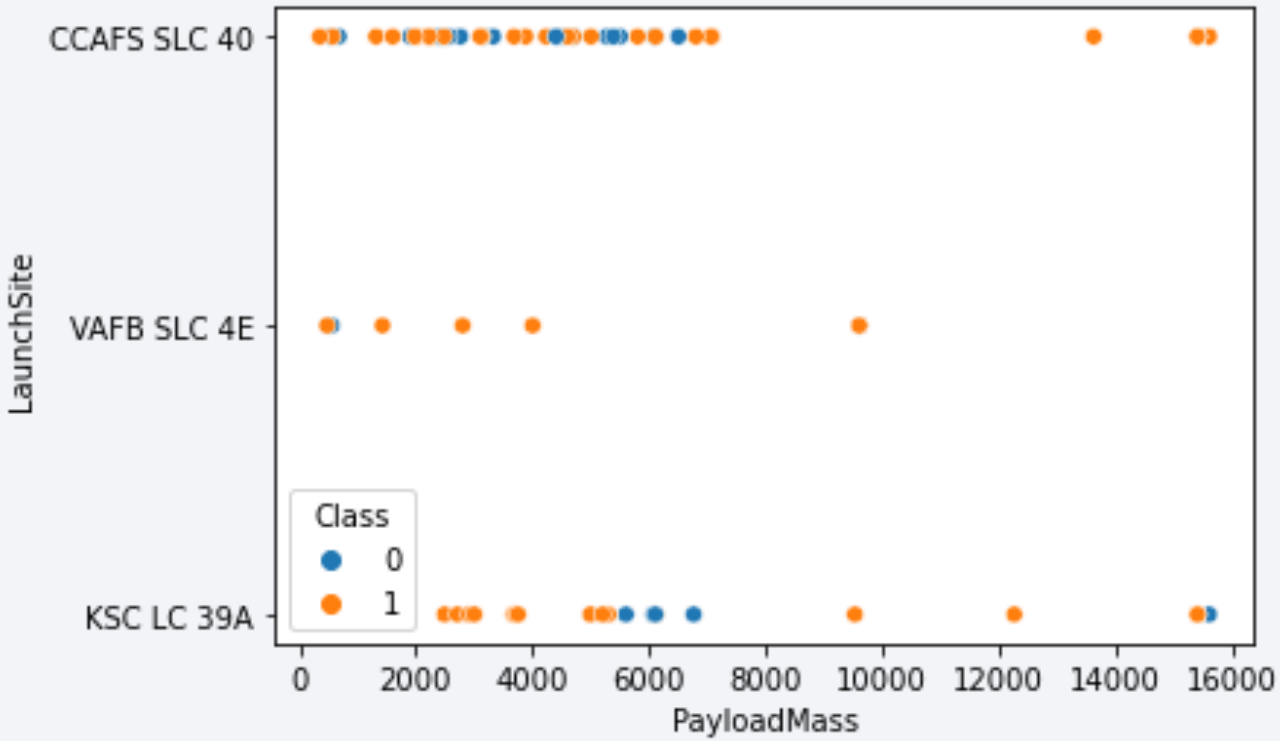
Flight Number vs. Launch Site

- Class 0 (blue) represents an unsuccessful launch while Class 1 (orange) represents a successful launch.
- We can see a trend showing that the increased number of flights achieves an increased success rate.



Payload vs. Launch Site

- Class 0 (blue) represents an unsuccessful launch while Class 1 (orange) represents a successful launch.
- There is no clear correlation between payload mass and success rate for a given launch site.



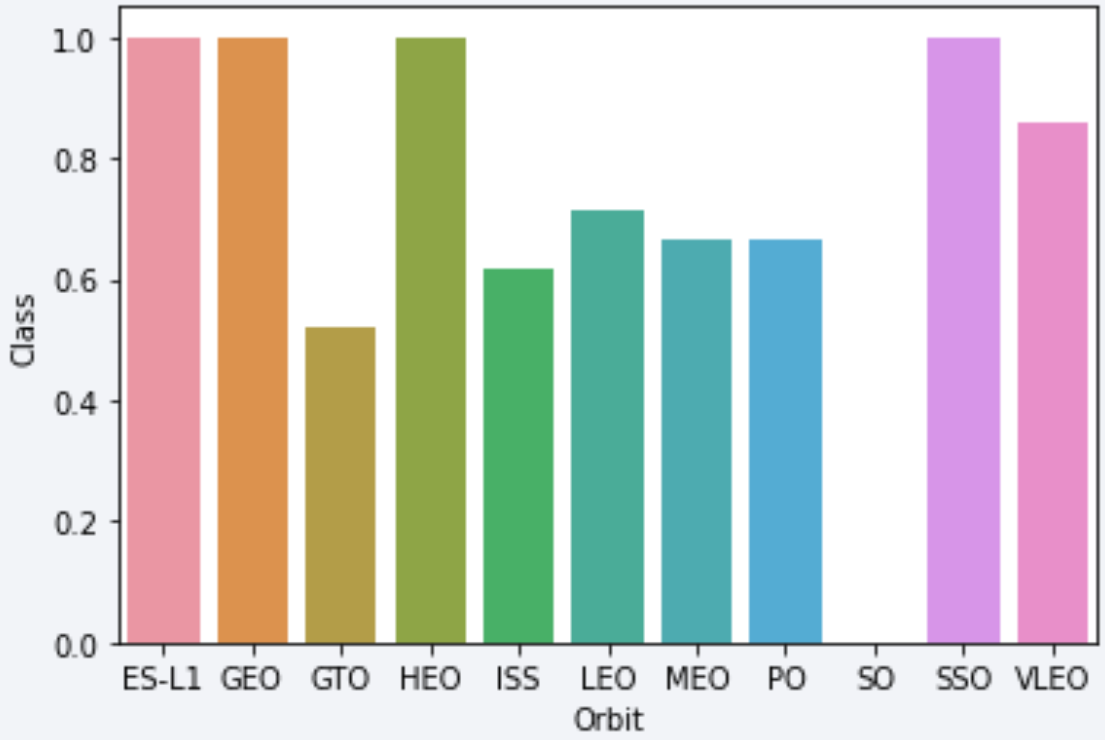
Success Rate vs. Orbit Type

We can observe a 100% success rate for the following orbit:

- ES-L1
- GEO
- SSO
- HEO

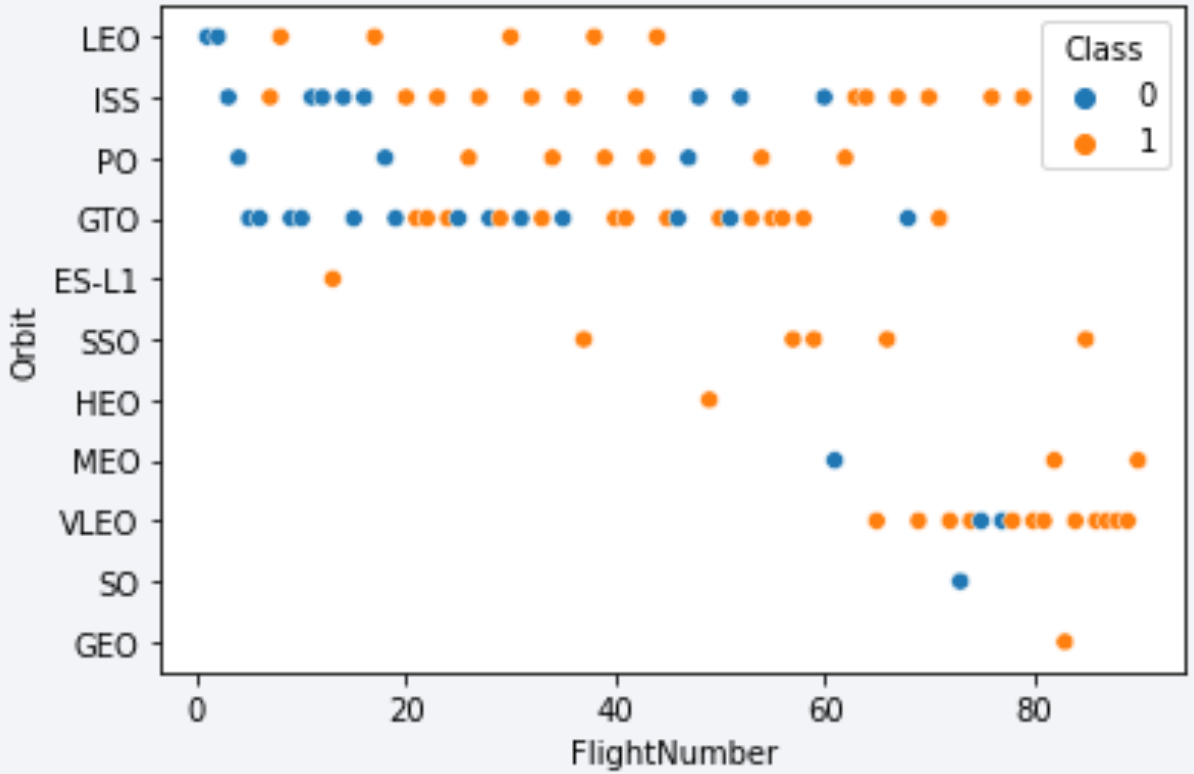
The lowest success rate is 0% for the following orbit:

- SO



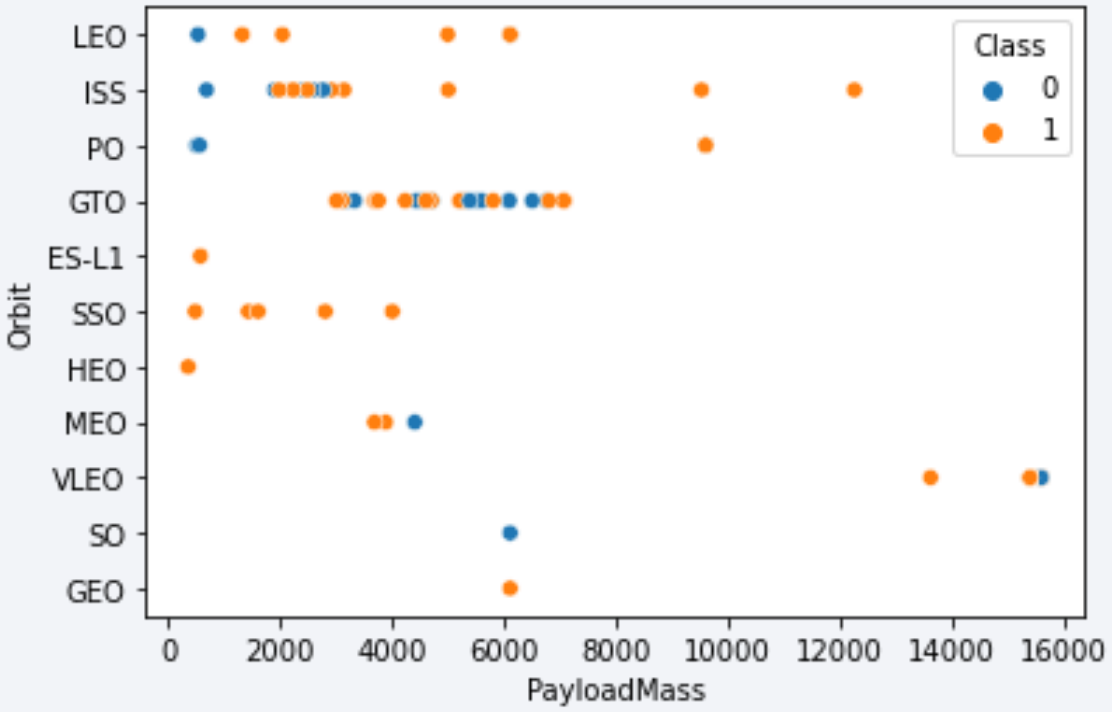
Flight Number vs. Orbit Type

- Class 0 (blue) represents an unsuccessful launch while Class 1 (orange) represents a successful launch.
- SSO has a 100% successful flight
- There are a trend of success rate increases as Flight number increase. These can be remarkably observed in LEO and VLEO orbits.



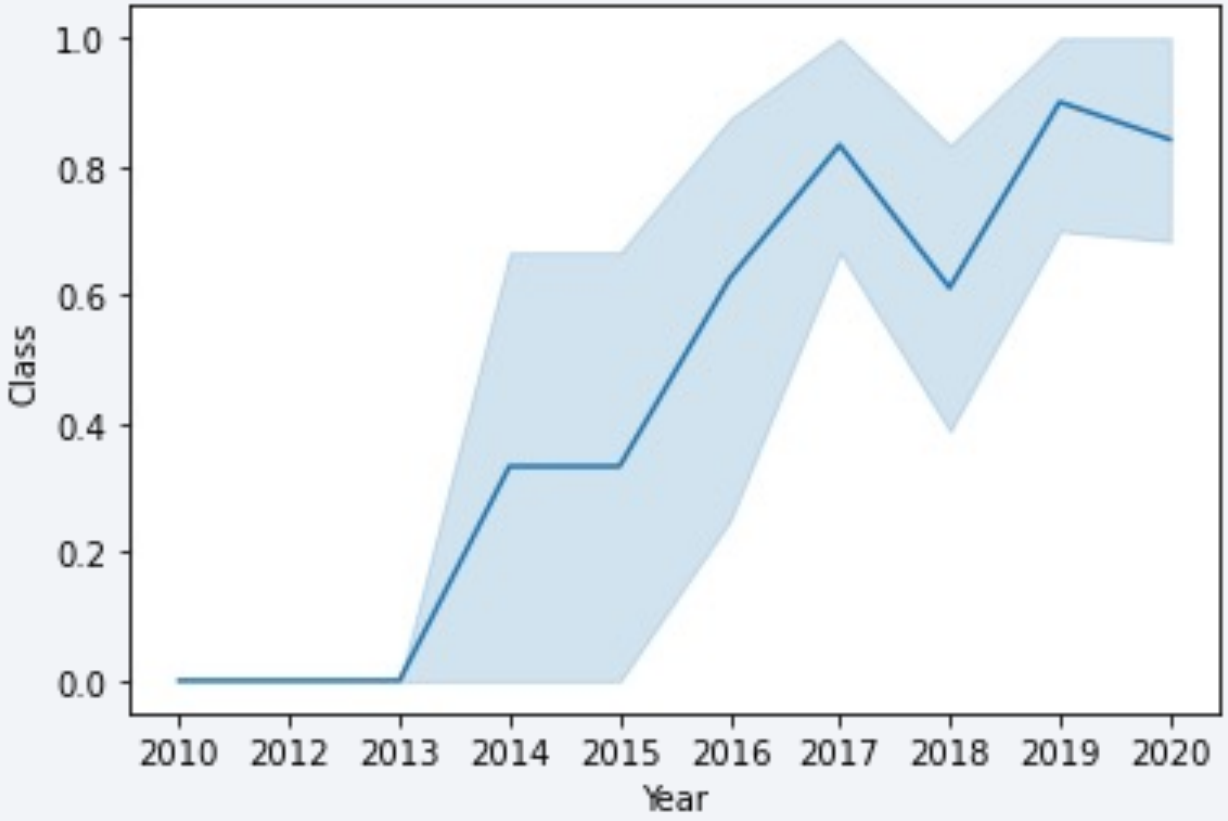
Payload vs. Orbit Type

- Class 0 (blue) represents an unsuccessful launch while Class 1 (orange) represents a successful launch.
- Uncertain is the correlation between GTO payload mass and success rate.
- With heavier payloads, the positive landing success percentage is higher for LEO and ISS.
- There are few launches to the orbits SO and GEO



Launch Success Yearly Trend

- The success rate began to increase in 2013 and continued until 2020
- From 2010 through 2013, no landings were accomplished (as the success rate is 0).
- Show the screenshot of the scatter plot with explanations



All Launch Site Names

- In the launch_site column of the SpaceX database, only unique values are presented when the SQL UNIQUE clause is applied to the query.
- There are four unique launch sites:
  - CCAFS LC-40,
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

```
%%sql
SELECT UNIQUE(Launch_Site)
FROM SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

- In the launch_site column of the SpaceX database, only unique values are presented when the SQL UNIQUE clause is applied to the query.
- LIMIT 5 retrieves just 5 records, whereas LIKE with the wildcard 'CCA%' retrieves string values beginning with 'CCA'.

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

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- In the launch_site column of the SpaceX database, only unique values are presented when the SQL UNIQUE clause is applied to the query.
- The SUM keyword is used to determine the total of the LAUNCH column.
- The WHERE keyword (and its accompanying SUM) restricts the results to boosters from NASA exclusively (CRS).

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
AS TOTAL_PAYLOAD_MASS
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)'
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- The AVG keyword is used to determine the average value of PAYLOAD_MASS__KG_ column.
- The WHERE keyword (and its accompanying AVG) restricts the calculation to Booster_Version = ‘F9 v1.1’

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
AS AVERAGE_PAYLOAD_MASS
FROM SPACEXTBL
WHERE Booster_Version = 'F9 v1.1'
```

average_payload_mass
2928

First Successful Ground Landing Date

- The MIN keyword is used to determine the earliest date of DATE column.
- The WHERE keyword (and its accompanying AVG) calculates for those Landing_Outcome is Success(ground pad).

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

first_successful_ground_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The WHERE keyword filter the data set to find those Landing_Outcome is Success(ground pad).
- The AND keyword (and its accompanying WHERE) adds an additional condition to only select the range from 4000 to 6000 in PAYLOAD_MASS__KG_ by using the BETWEEN keyword

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE (Landing__Outcome = 'Success
(drone ship)') AND (PAYLOAD_MASS__KG_
BETWEEN 4000 AND 6000);
```

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

Total Number of Successful and Failure Mission Outcomes

- The COUNT keyword is used to calculate the total number of mission outcomes.
- The GROUP BY keyword groups the results by the types of mission outcomes.

```
%%sql
SELECT MISSION_OUTCOME,
COUNT(MISSION_OUTCOME) AS
TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

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

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT(BOOSTER_VERSION)
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_)
FROM SPACEXTBL);
```

- By using a subquery, the SELECT statement within the brackets finds the maximum payload, and this value is used in the WHERE condition.
- The DISTINCT keyword is then used to retrieve only distinct booster versions.

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

2015 Launch Records

- The WHERE keyword is used to filter the results for only failed landing outcomes.
- The AND keyword (accompanying WHERE) filters the selection for the year of 2015.

```
%%sql
SELECT Landing__Outcome ,
Booster_Version , Launch_Site
FROM SPACEXTBL
WHERE Landing__Outcome = 'Failure
(drone ship)' AND YEAR(Date) = 2015;
```

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

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

```
%%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;
```

- The WHERE keyword filters the results to be calculated in the date range of ‘2010-06-04’ to ‘2017-03-20’.
- The results are then grouped and ordered, using the keywords GROUP BY and ORDER BY, respectively, where DESC keyword is used to specify the descending order.

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

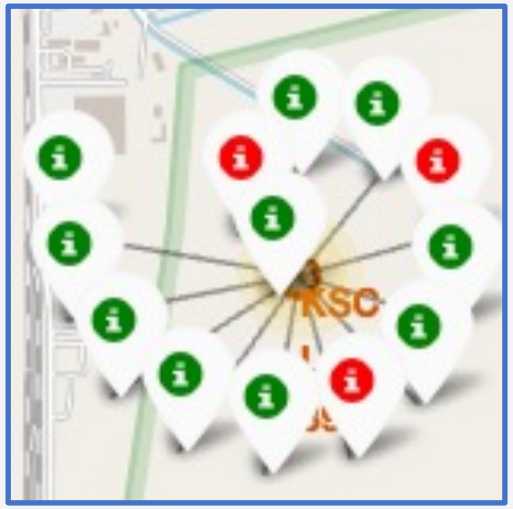

Section 3
Launch Sites
Proximities Analysis

All Launch Sites Locations

- This map shows all the launch site locations, and we can see that they are all located near the coastline.
- The sites set up near the coastline are most probably for the public's safety in scenarios where unexpected accidents occur.

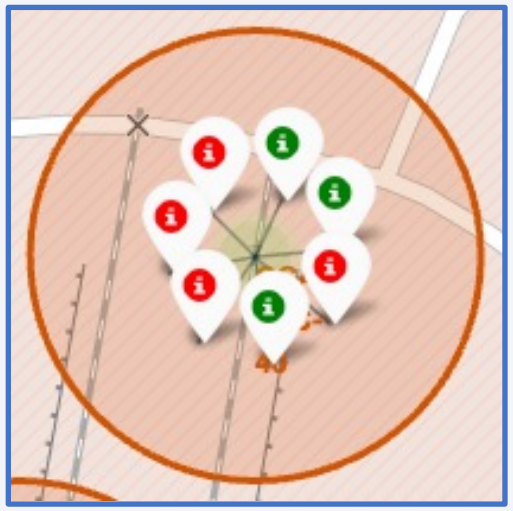
SUCCESS/FAILED LAUNCHES FOR EACH SITE

VAFB SLC-4E



KSC LC-39A

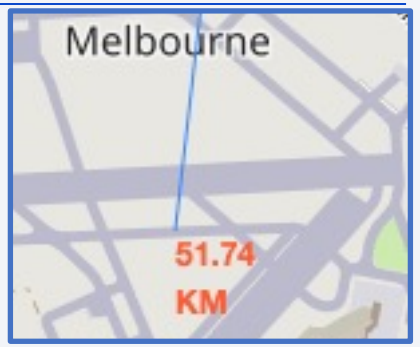
CCAFS LC-40



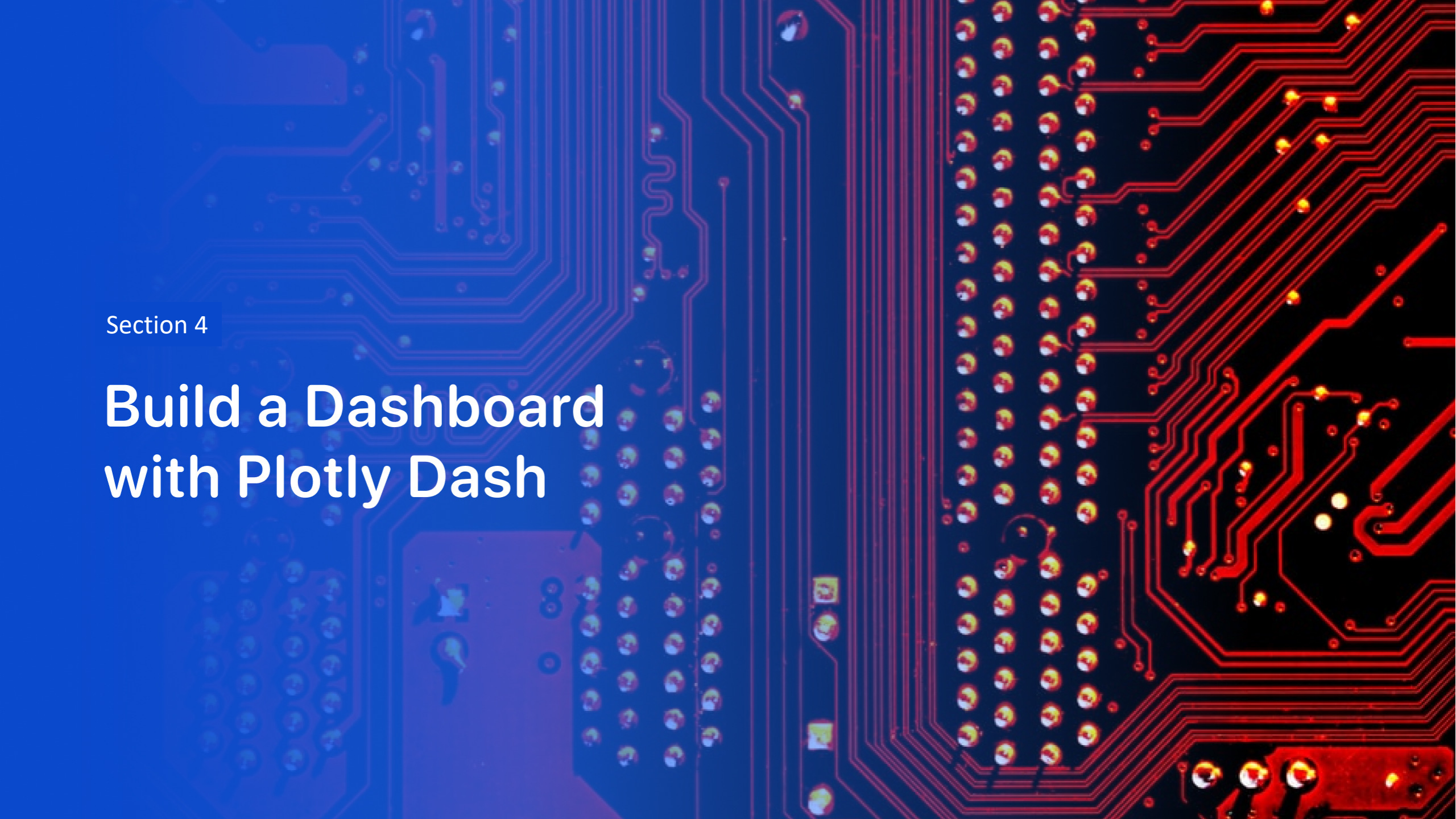
CCAFS SLC-40

- Launches have been grouped into clusters, and annotated with green icons for successful launches, and red icons for failed launches.

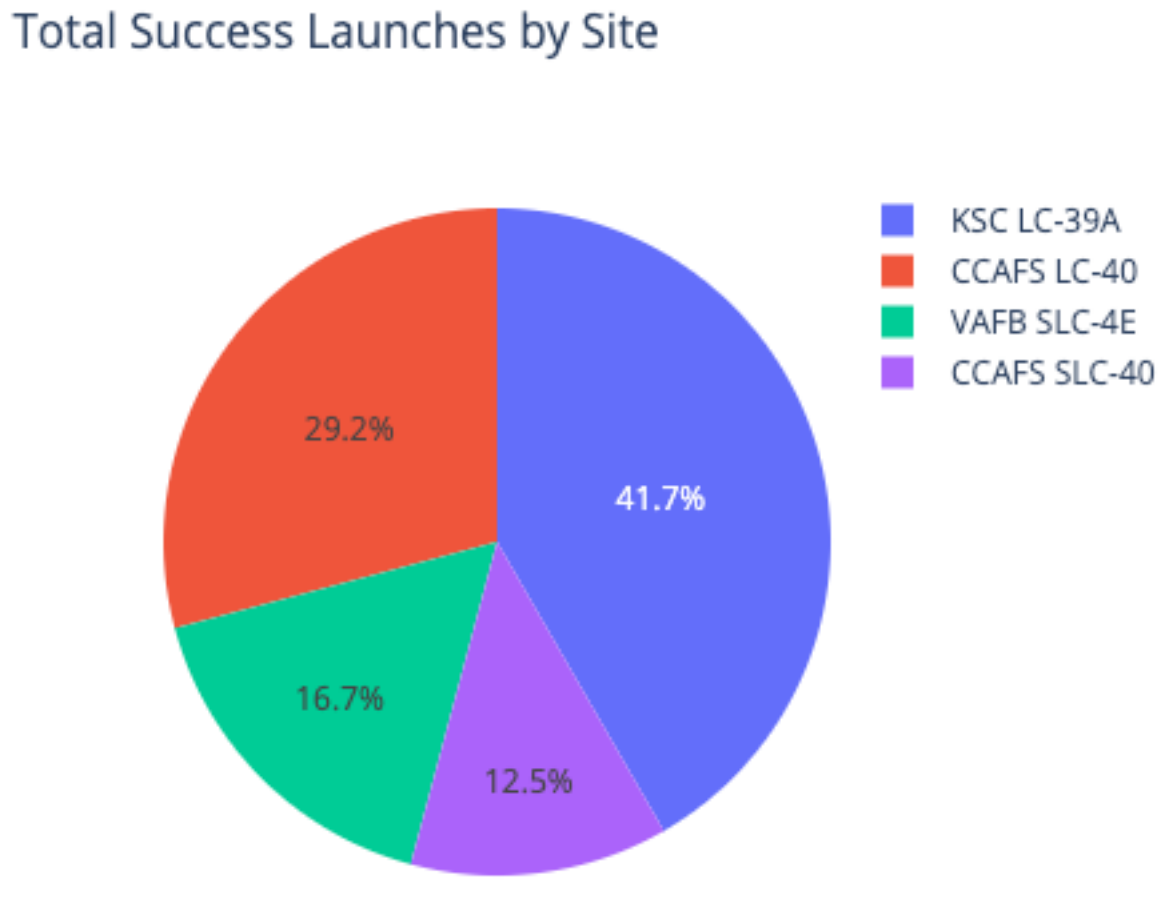
Proximities of Launch Sites



- Are launch sites in close proximity to railways? **YES**.
- The coastline is only 0.87 km due East. Are launch sites in close proximity to highways? **YES**.
- The nearest highway is only 0.59km away. Are launch sites in close proximity to railways? **YES**.
- The nearest railway is only 1.29 km away. Do launch sites keep certain distance away from cities? **YES**.
- The nearest city is 51.74 km away. Do launch sites keep certain distance away from cities? **YES**

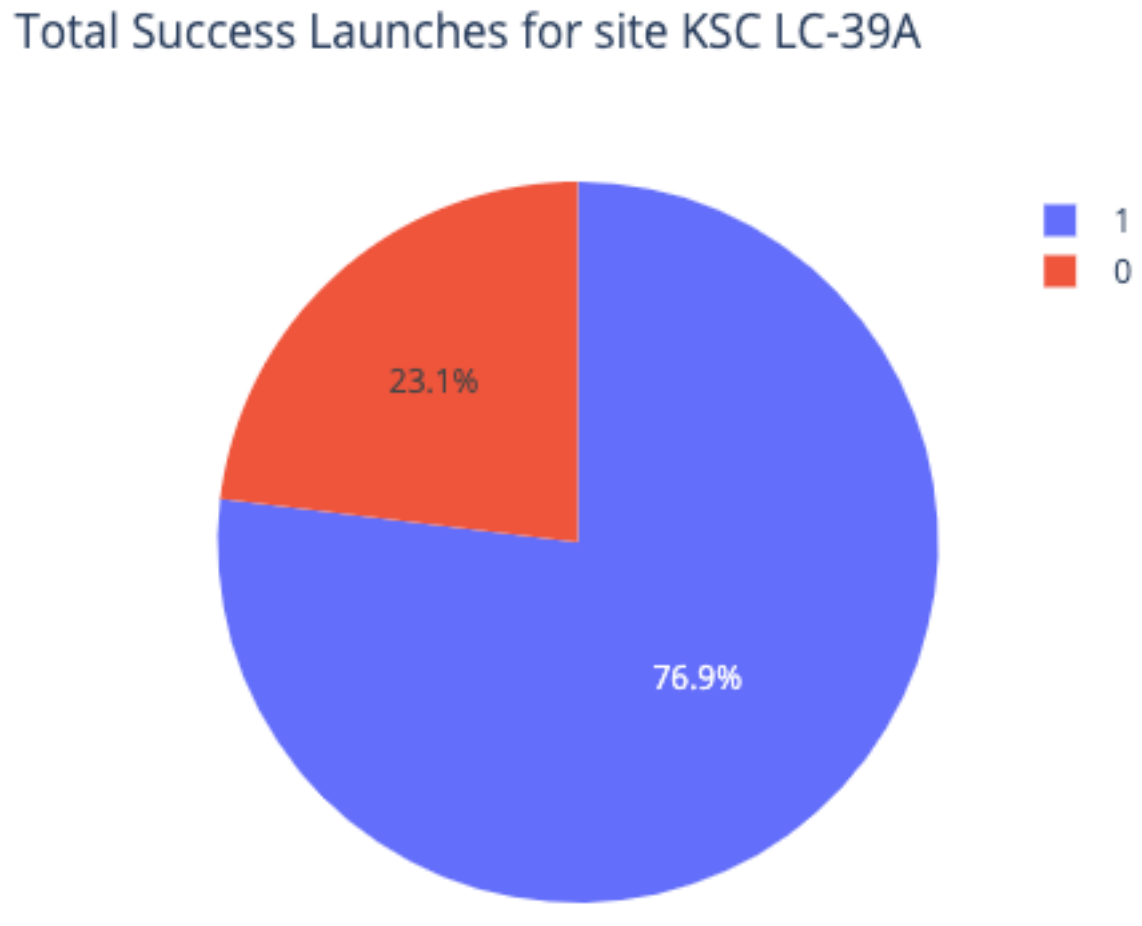

Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



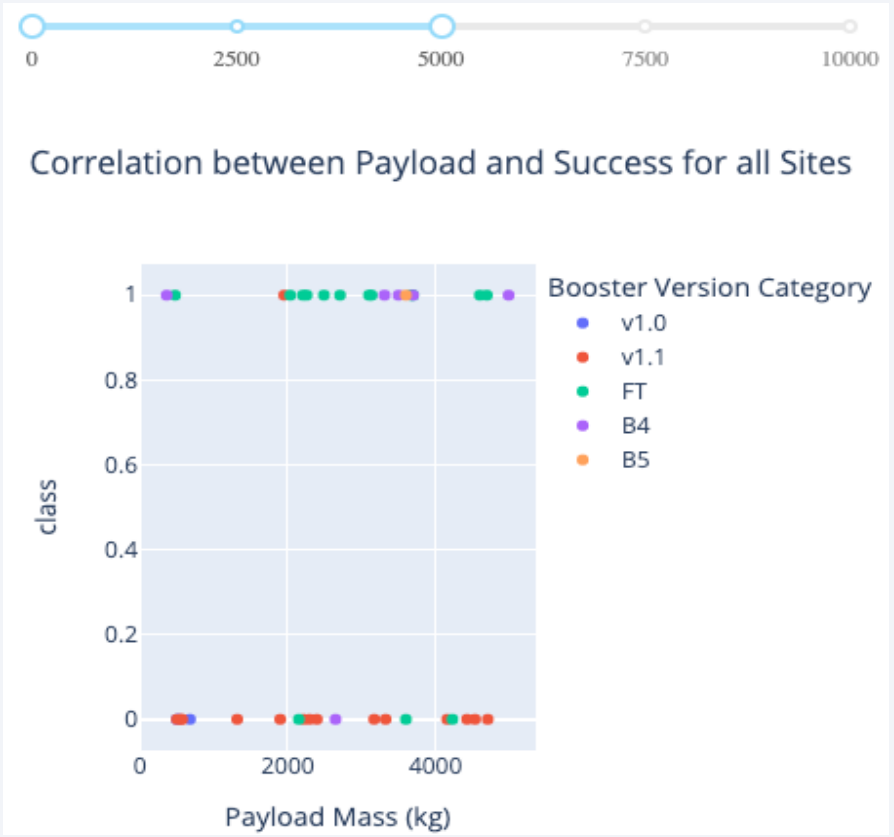
The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches.

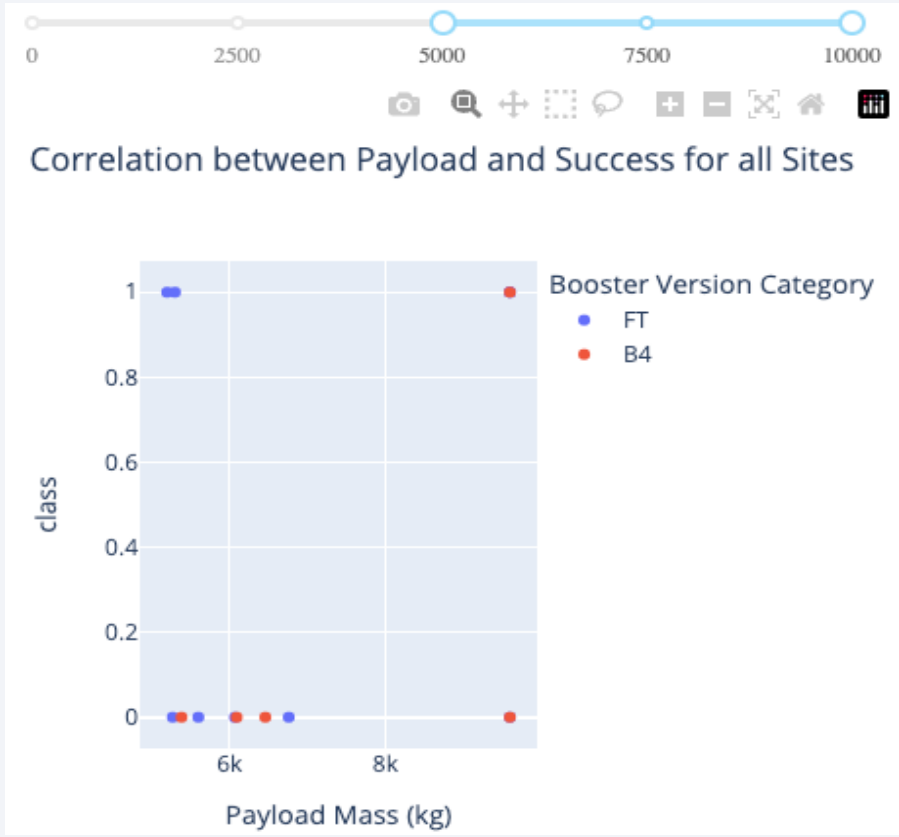
<Dashboard Screenshot 2>



The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches.

<Dashboard Screenshot 3>



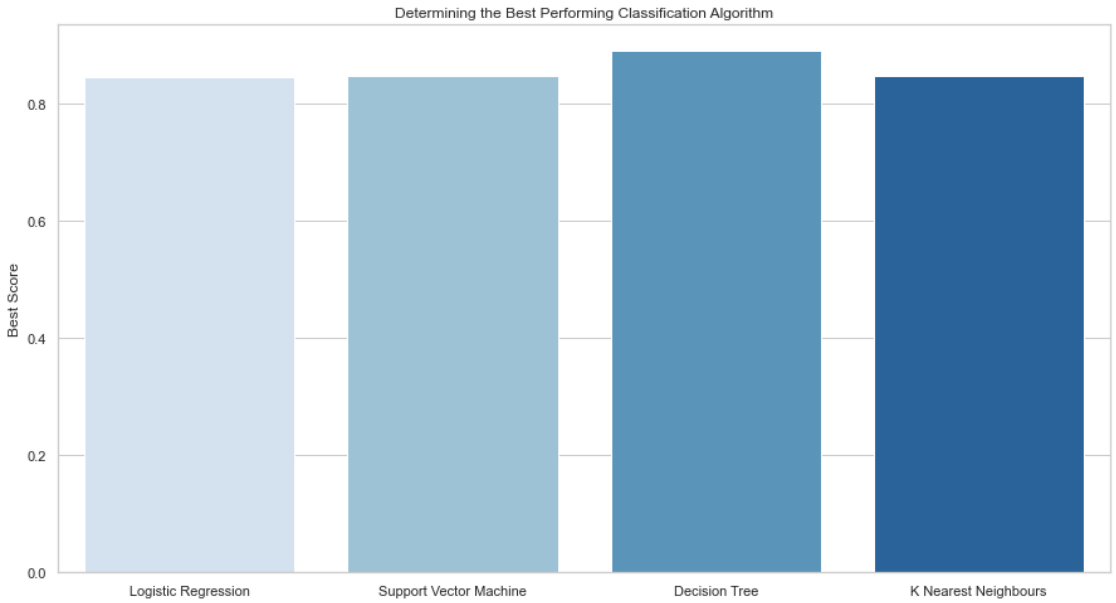


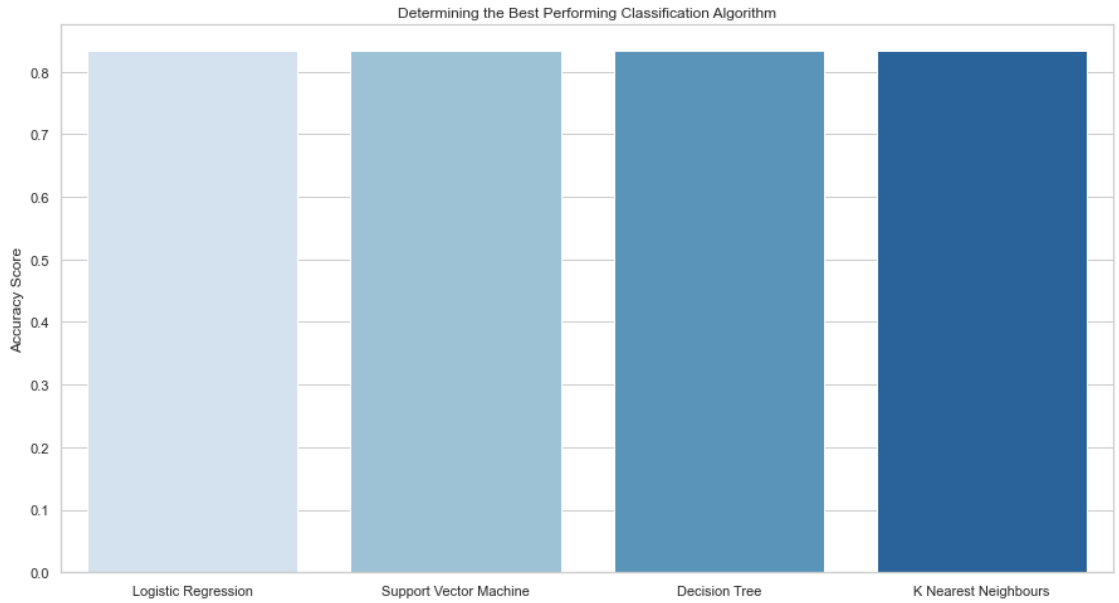
- Low weighted payloads (0-5000 kg) have a higher launch success rate (class 1) than heavy payloads (5000-10000 kg).

Section 5

Predictive Analysis (Classification)

Classification Accuracy

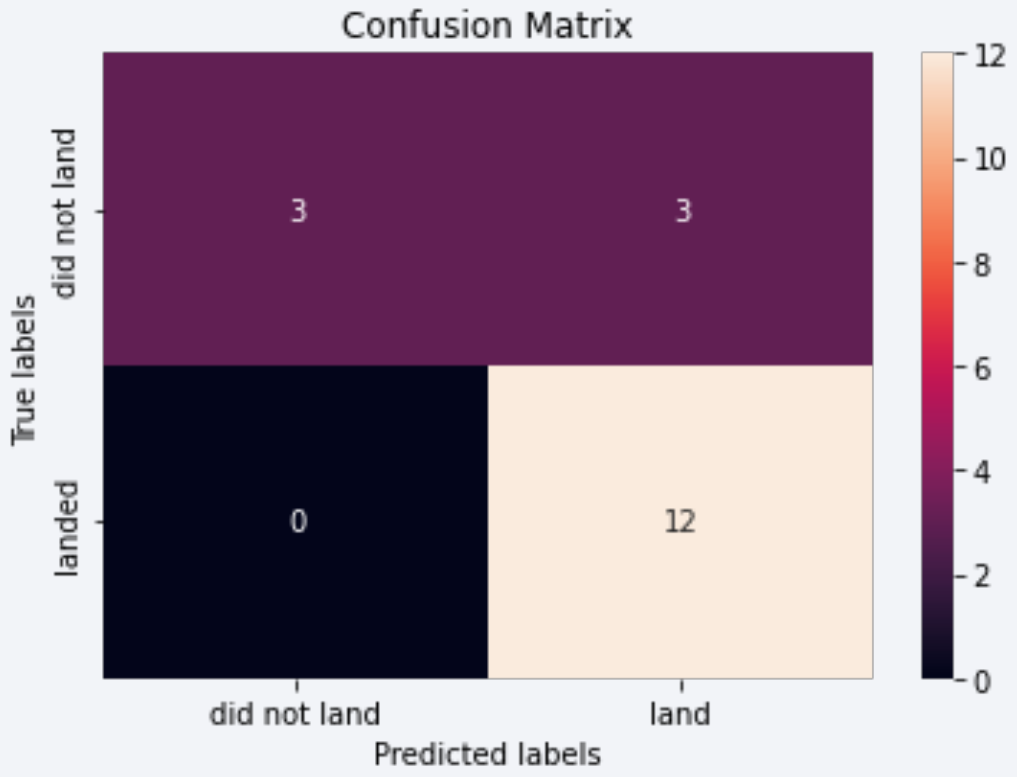




- We can observe the accuracy score all at the same at 83.33%.
- Due to a small test size the accuracy score would contribute to be the same. Thus, more data would be required to determine an optimal model

	Algorithm	Accuracy Score	Best Score
0	Logistic Regression	0.833333	0.846429
1	Support Vector Machine	0.833333	0.848214
2	Decision Tree	0.833333	0.891071
3	K Nearest Neighbours	0.833333	0.848214

Confusion Matrix



- The confusion matrix resulted the same for all the models as they were conducted on the same test set.
- The confusion matrix, which shows 3 out of 18 total results classified incorrectly (a false positive, shown in the top-right corner).
- The remaining 15 have been classified correctly:
  - 3 did not land
  - 12 did land.

Conclusions

- KSLC-39A has the maximum number of successful launches and the maximum rate of success across all sites.
- Orbit types ES-L1, GEO, HEO, and SSO, have the highest (100%) success rate.
- In this data set, all models have the same accuracy (83.33%). However, it appears that further data are required to establish the ideal model because of the limited data size.
- The launch site is close to railways, highways, and coastlines but far from cities.
- The launch success rate of small payloads is better than that of heavier payloads.

Appendix

- GitHub URL - https://github.com/Ashwiin/IBM-Data-Science-Capstone-Project

Thank you!