IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rehan S. Jayakar
14-Jul-2023

Outline

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

Abstract

This project looks to discover whether it can be predicted if the first stage of the SpaceX Falcon 9 rockets will land successfully, which will thereby allow for determining the cost of a launch.

Through the various techniques, it is possible to collect data, perform data wrangling, investigate the data through exploratory analysis and visualise these results through different techniques as well as perform interactive visual analytics through maps to explore the launch sites with visual aids adding to the level of detail. Relationship between different variables were found.

Finally, through machine learning techniques, it is found possible to perform predictive analysis using classification models for successful and failed launches of the Falcon 9 rockets with a high level of prediction accuracy.

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

- EDA Results
- Interactive analytics
- Predictive analysis

Introduction

Project background and context

- SpaceX advertises the Falcon 9 rocket launches on its website, with a cost of $62m; other providers cost upwards of $165m each, much of the savings is because SpaceX can reuse the first stage. Therefore, if one can determine if the first stage of the SpaceX Falcon 9 will land successfully, one 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. This is the ambition of the project.

Section 1

Methodology

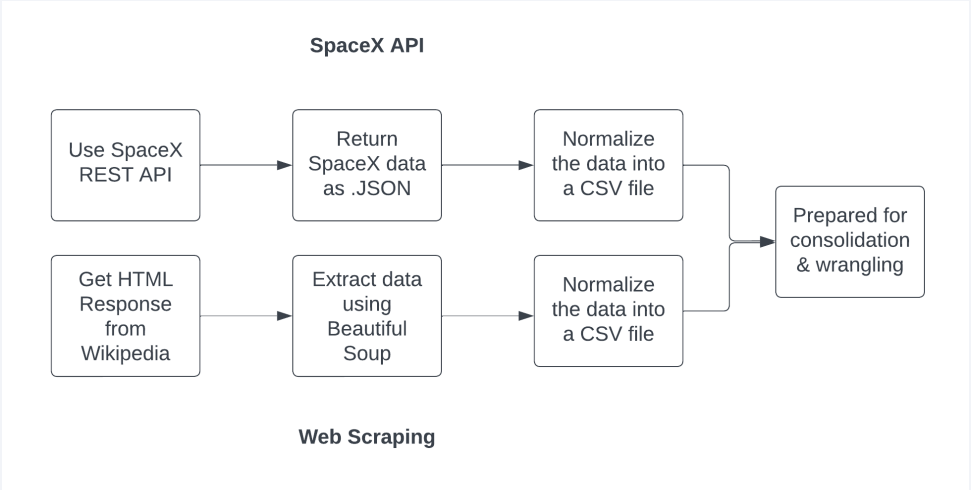
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and Data Cleaning of null values and 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
  - LR, KNN, SVM, DT Models have been built, tuned and evaluated for the best classifier.

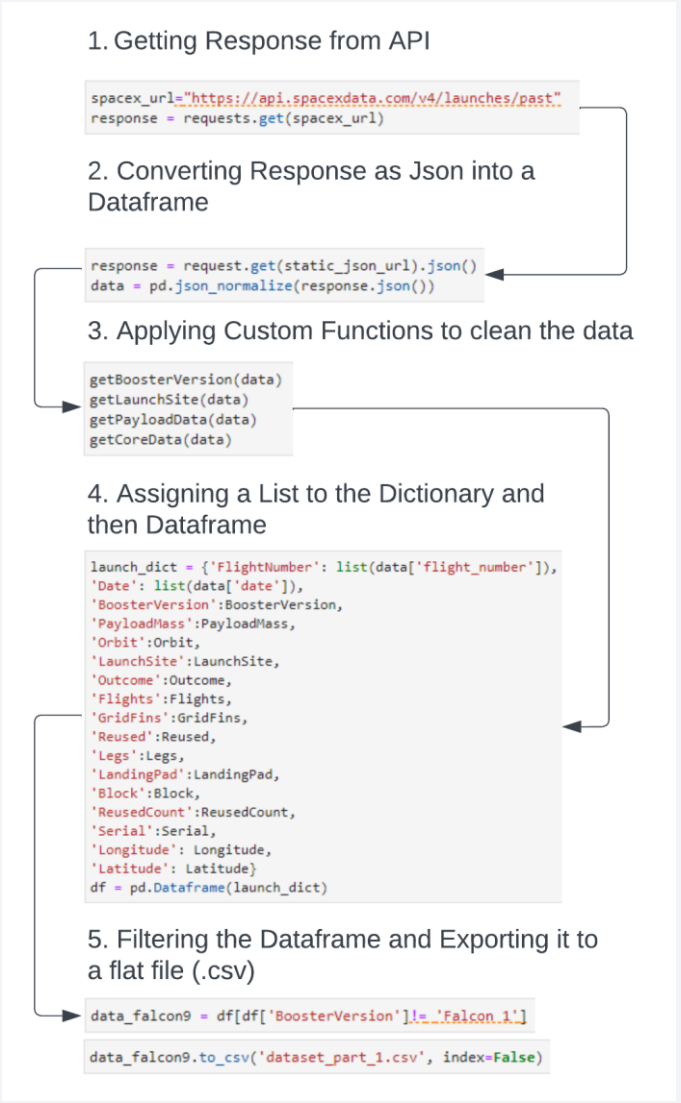
Data Collection

- How the datasets were collected:
  - SpaceX launch data that is gathered from SpaceX REST API.
  - This API gives data about launches, including information about the rockets used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.
  - Another popular data source for obtaining Falcon9 data is web scraping Wikipedia using BeautifulSoup.



Data Collection – SpaceX API

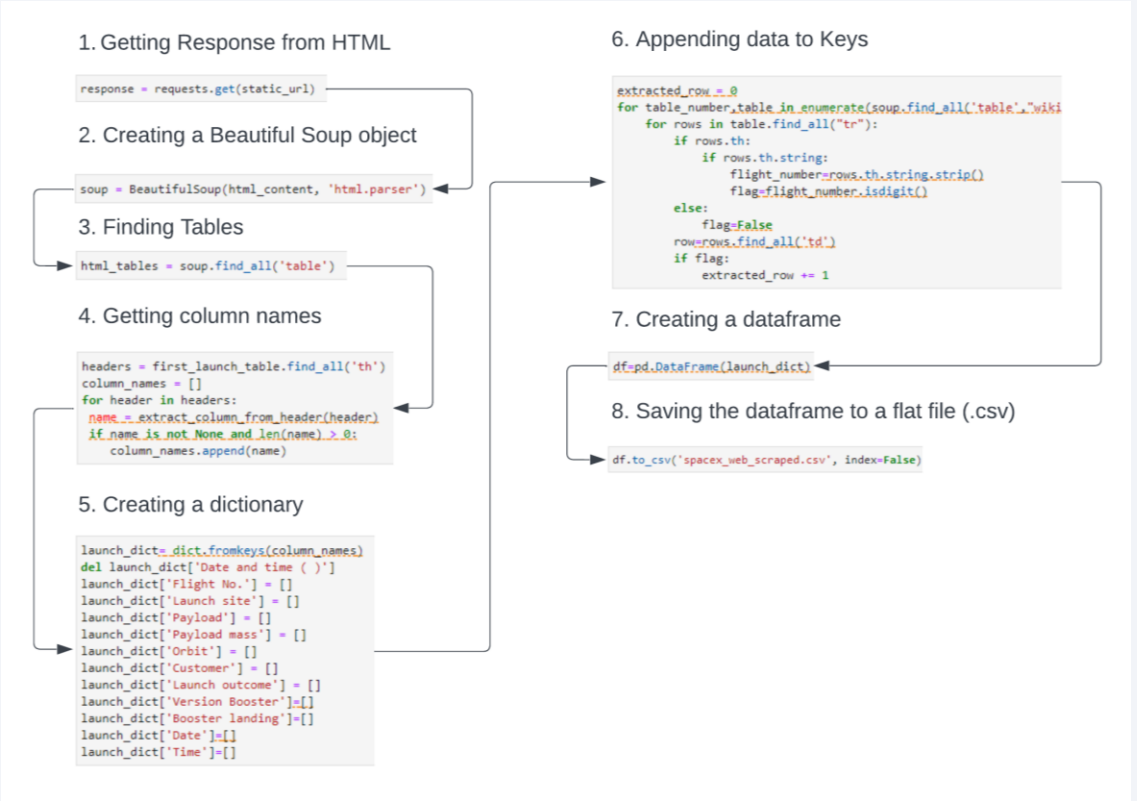
- Data collection with SpaceX REST Calls
- The flowchart visualises the data collection process involved when accessing the SpaceX API all the way to creating a CSV file to store the data as a data frame.



https://github.com/RehanSJ/IBM-Skills-Network-Labs/blob/Module-10-Applied-Data-Science-with-Capstone/Module%2010%20Week%201%20jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Web Scraping

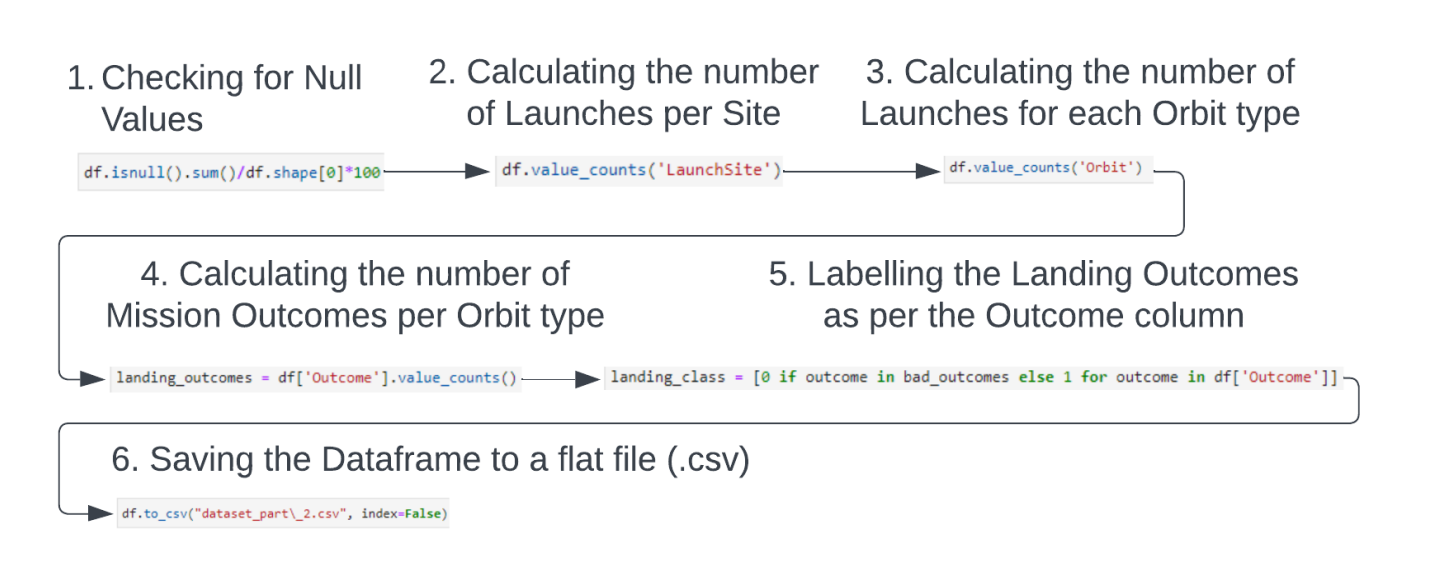
- Data collection with Web Scraping from Wikipedia
- The flowchart used illustrates the data collection process involved when accessing the data from Wikipedia all the way to creating a CSV file to store the data as a data frame.



https://github.com/RehanSJ/IBM-Skills-Network-Labs/blob/Module-10-Applied-Data-Science-with-Capstone/Module%2010%20Week%201%20Webscraping.ipynb

Data Wrangling

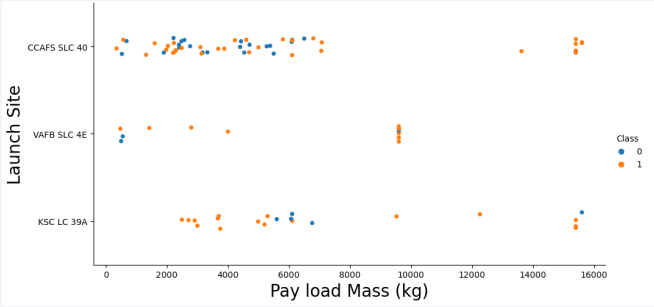
- The flowchart visualizes the data wrangling process from checking for null values to creating a label to identify successful and failure launches from the Mission Outcome data as a binary format – 1, if successful, and 0, if unsuccessful – and then saving that data frame to a flat file.

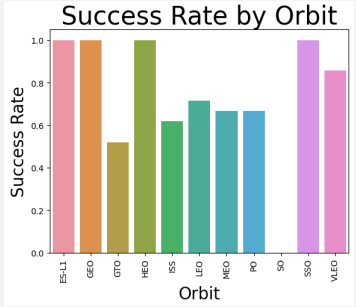


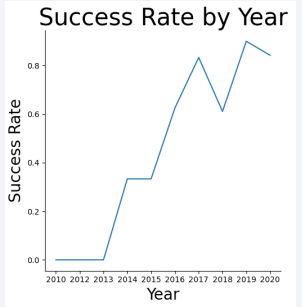
https://github.com/RehanSJ/IBM-Skills-Network-Labs/blob/Module-10-Applied-Data-Science-with-Capstone/Module%2010%20Week%201%20Data%20Wrangling.jupyterlite.ipynb

EDA with Data Visualization

- Various types of charts are used to investigate the potential existence of relationships between different variables in the data set.
- The variable combinations being investigated are as follows:
  - Flight number and payload mass, flight number and launch site, payload and launch site, success rate and orbit type, flight number and orbit type, payload and orbit type, success rate and year. Below here are some examples of the types of charts used for the data.







https://github.com/RehanSJ/IBM-Skills-Network-Labs/blob/Module-10-Applied-Data-Science-with-Capstone/Module%2010%20Week%202%20EDA%20Data%20Visualization.ipynb.jupyterlite.ipynb

EDA with SQL

SQL queries performed:

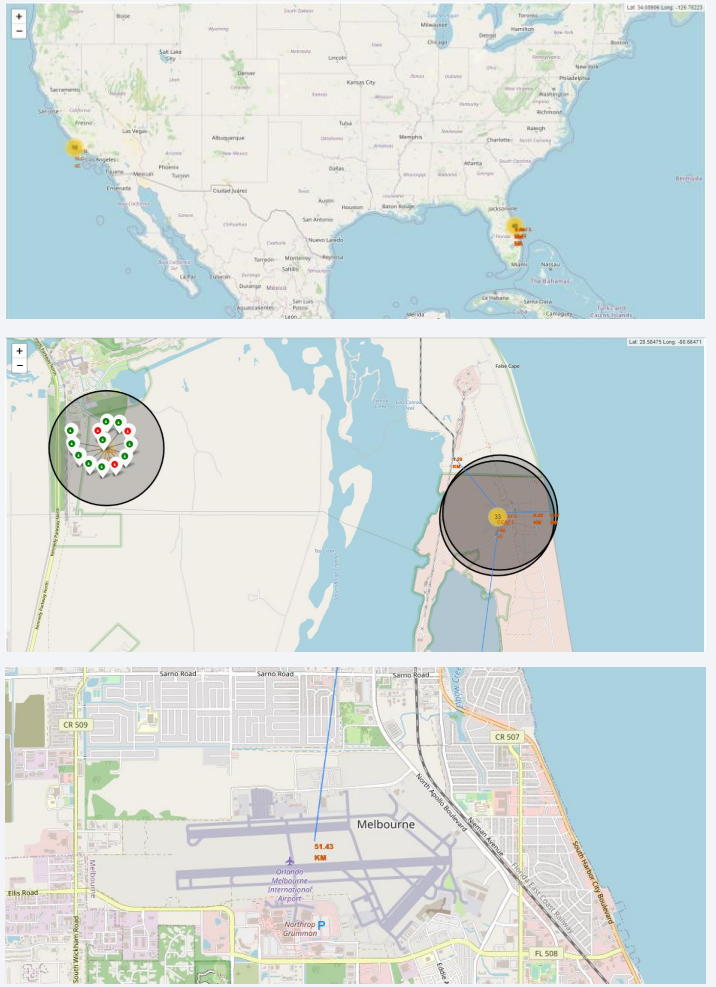
- Displaying the names of the unique launch sites in the space mission.
- Displaying 5 records of launch sates beginning with the string 'CCA'.
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Displaying average payload mass carried by booster version F9 v1.1.
- Listing the date where the successful landing outcome in the ground pad was achieved.
- Listing the names of the boosters with success in the drone ship and a payload mass between 4000kg and 6000kg.
- Listing the total number of successful and failure mission outcomes.
- Listing the names of booster versions which carried maximum payload mass.
- Listing the records which will display the month names, failure landing outcomes in drone ship, booster versions, and launch site for the months in year 2015.
- Ranking the count of successful landing outcomes between 4th Jun 2010 and 20th Mar 2017 in descending order.

https://github.com/RehanSJ/IBM-Skills-Network-Labs/blob/Module-10-Applied-Data-Science-with-Capstone/Module%2010%20Week%202%20SQL%20Exploratory%20Analysis.ipynb

Build an Interactive Map with Folium

The various map objects added:

- The name of the launch site as a title is shown in first image .
- Circles and Markers for all launch sites to easily find them.
- The number of each launches per site is labelled by the Marker. The Markers are also either red, to indicate a failed launch, or green, to indicate a successful launch.
- Blue lines were also added to point out the distances to the nearest key features in the map, such as the coastline, highway, railroad, and city, in addition to the calculated distances to these features which are labelled.

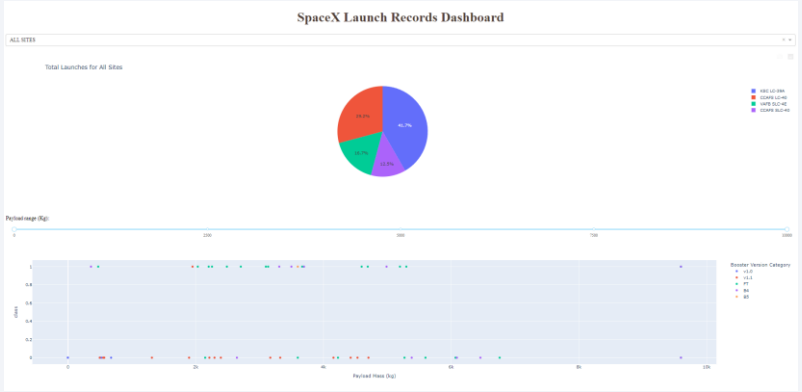


https://github.com/RehanSJ/IBM-Skills-Network-Labs/blob/Module-10-Applied-Data-Science-with-Capstone/Module%2010%20Week%203%20Interactive%20Visual%20Analytics%20with%20Folium.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

The plots/graphs and interactions added to the dashboard:

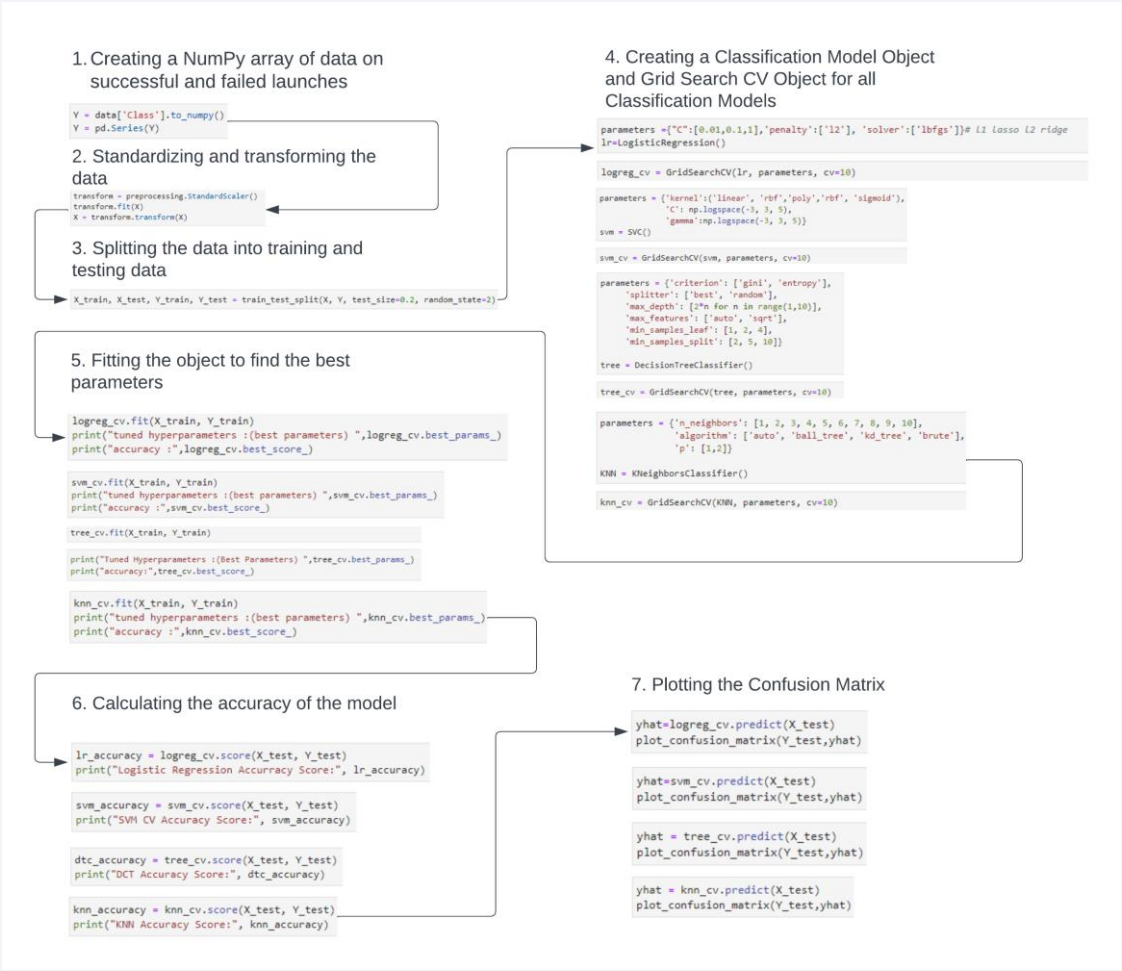
- A dropdown menu, to select either 'All' the launch sites or each one individually when visualizing the data.
- A pie chart to represent the ratio of launches all sites or success to failure for each site individually.
- A range slider to determine the selected Payload Mass amount for the scatter graph.
- A scatter graph to illustrate the successful and failed launches based on the payload mass data selected prior for either all the sites or each site individually.



https://github.com/RehanSJ/IBM-Skills-Network-Labs/blob/Module-10-Applied-Data-Science-with-Capstone/Module%2010%20Week%203%20Dashboard%20Application%20Plotly%20Dash.ipynb

Predictive Analysis (Classification)

- The flowchart visually encapsulates the process adopted to build, evaluate, improve, and identify the best performing classification model.
- This covers the process starting from creating a NumPy array of data for successful and failed mission outcomes all the way to plotting the confusion matrices after the prediction accuracies of the model are calculated.



https://github.com/RehanSJ/IBM-Skills-Network-Labs/blob/Module-10-Applied-Data-Science-with-Capstone/Module%2010%20Week%204%20Machine_Learning_Prediction.jupyterlite.ipynb

Results

Exploratory data analysis results:

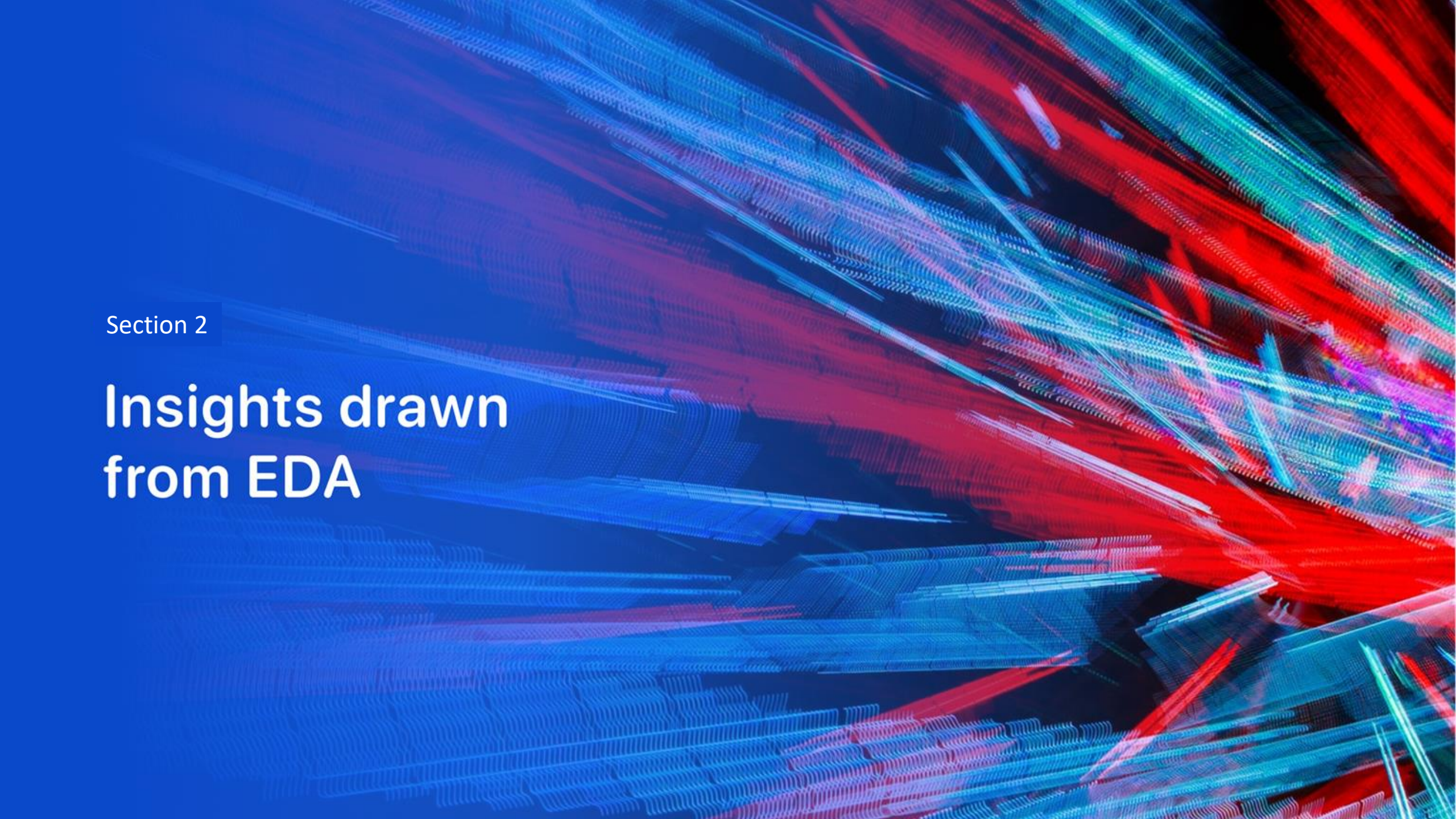
- Heavy weighted payloads tend to perform better than lighter payloads.
- The success rates for SpaceX is positively correlated with the time in years spent on the launches.
- KSC LC 39A had the most successful launches from all sites
- Orbits GEO, HEO, SSO, and ES L1 have the best success rate.
- FT is relatively the most successful booster version.

Interactive Analytics results:

- Launch sites selected are relatively close to the coastline, but further from railways, highways, and especially cities.

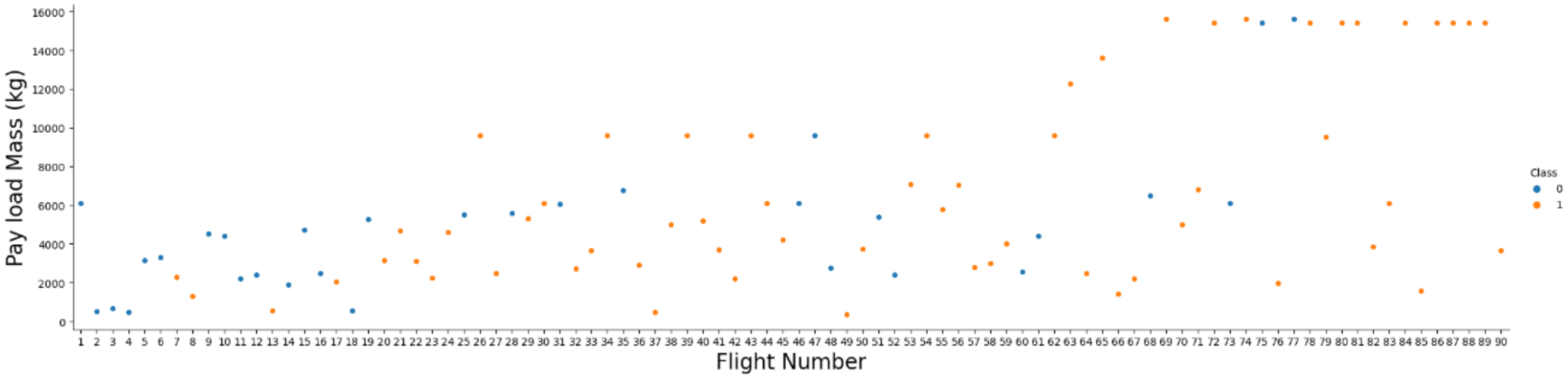
Predictive analysis results:

- The SVM, KNN, and Logistic Regression models are optimal with a prediction accuracy of 83.33%.


Section 2
Insights drawn
from EDA

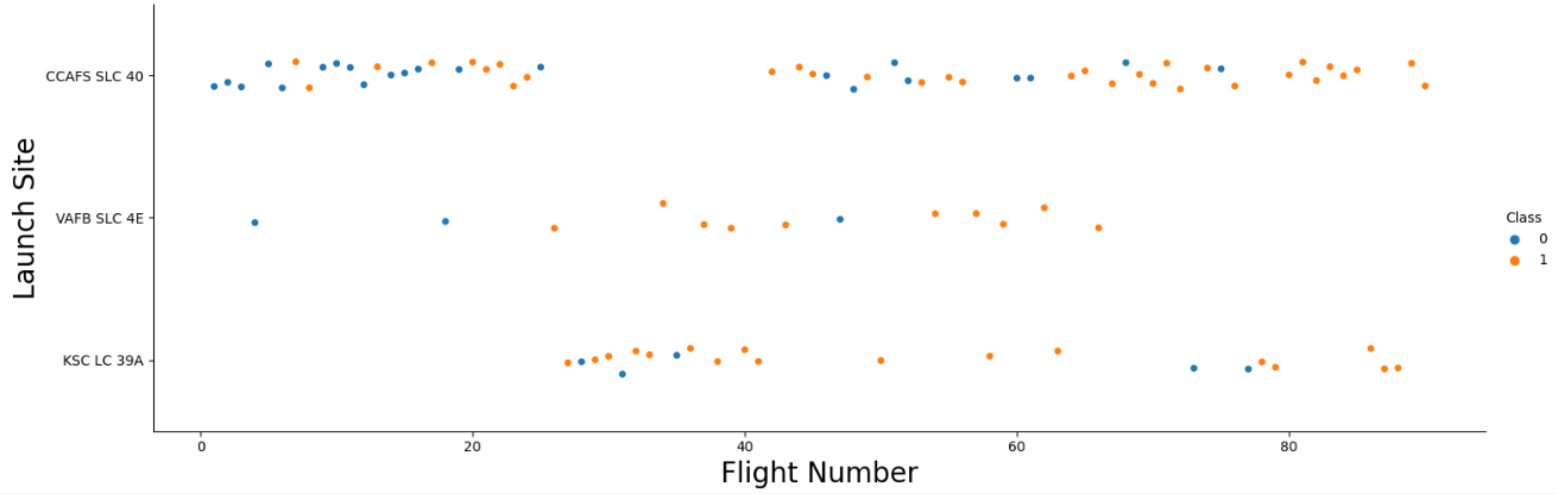
Flight Number vs. Launch Site

- The scatter plot visualizes the relationship between Flight Number and Payload mass.
- We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.



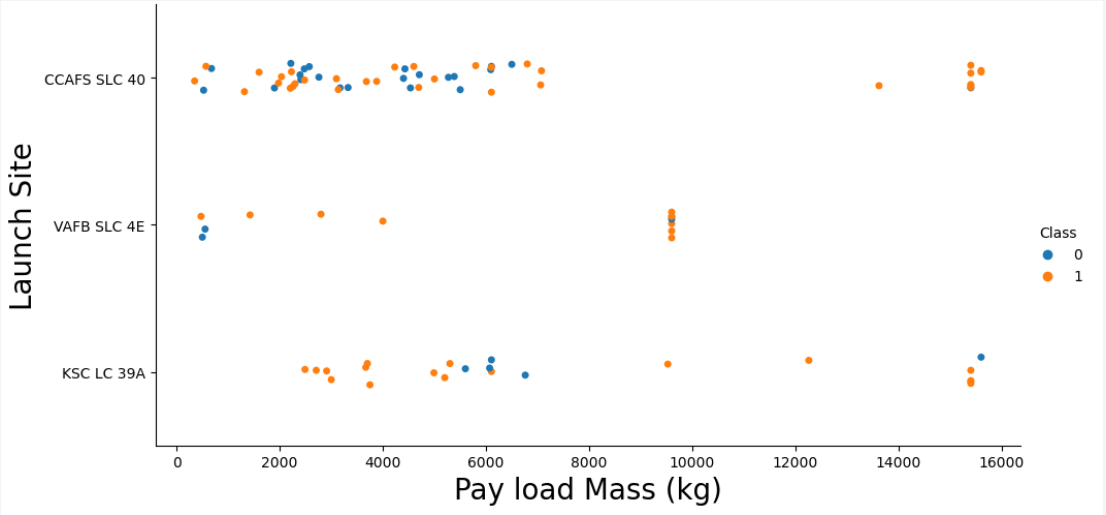
Flight Number vs. Launch Site

- The scatter plot visualizes the relationship between Flight Number and Launch Site.
- CCAFS LC-40 (CCA) has a success rate of 60 %, while KSC LC-39A (KSC) and VAFB SLC 4E (VAF) have a success rate of 77%.
- The number of launches from CCA are significantly greater than the other launch sites.
- Typically for all sites, there is greater probability for success as flight number rises.



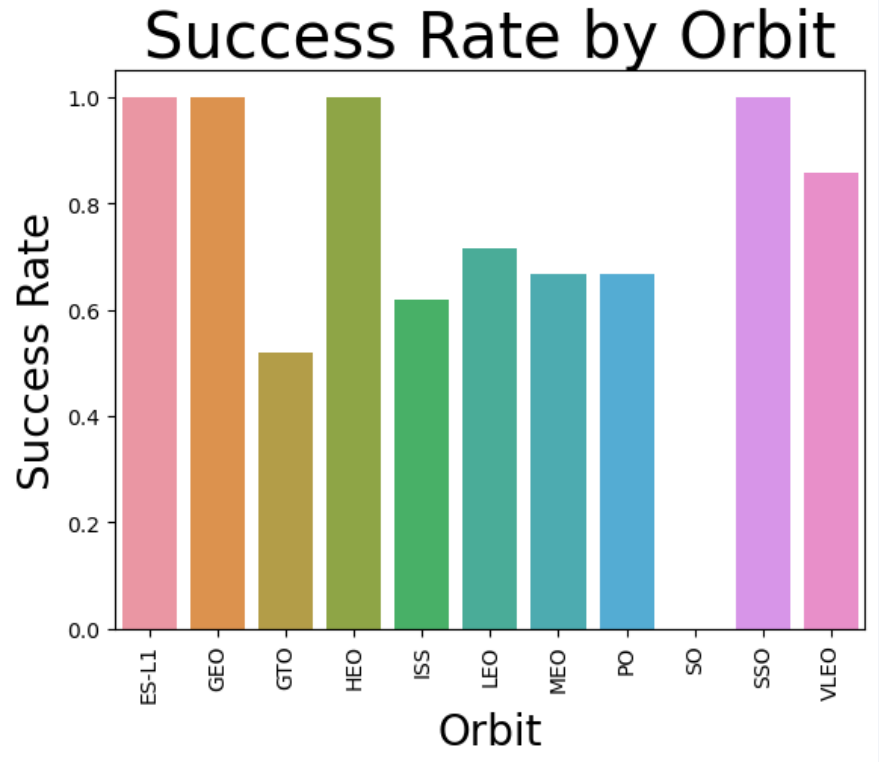
Payload vs. Launch Site

- The scatter plot illustrates the relationship between the Launch Site and Payload Mass.
- For the VAF launch site, there are no rockets launched for a heavy payload mass (>10,000kg).
- Most of the lighter payload mass records are present with CCA but, tend to be more successful with KSC.
- CCA is more successful with heavier payloads.



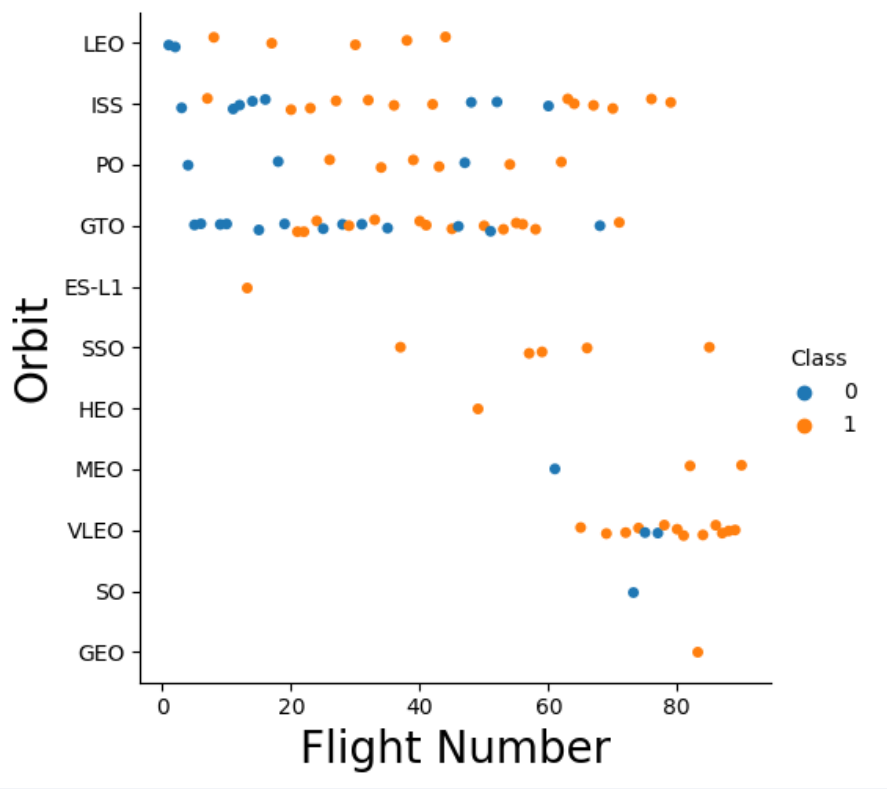
Success Rate vs. Orbit Type

- The bar chart visualizes the success rate of each orbit type
- ES-L1, GEO, HEO, and SSO are the most success at a rate of 100%.
- The others range between 50% to 70%, except for VLEO which nears 80%.



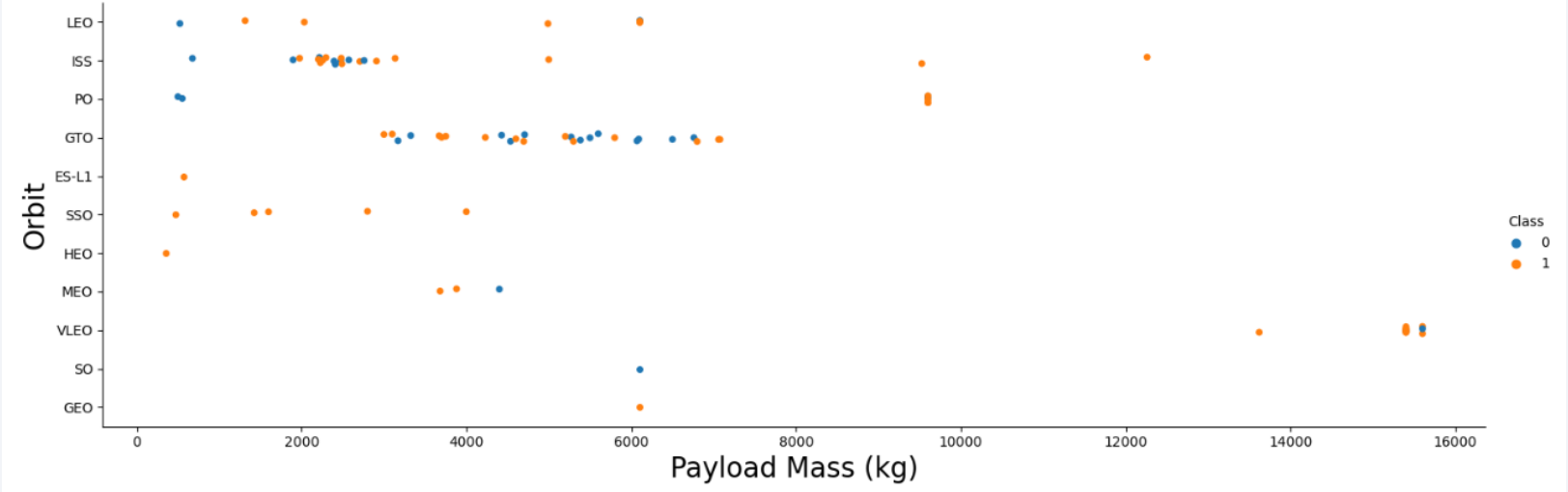
Flight Number vs. Orbit Type

- The scatter plot illustrates whether a relationship exists between Flight number and Orbit type.
- The LEO orbit's 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.
- Later flight numbers tend to cluster at VLEO making it the more commonly used orbit type in recent years.



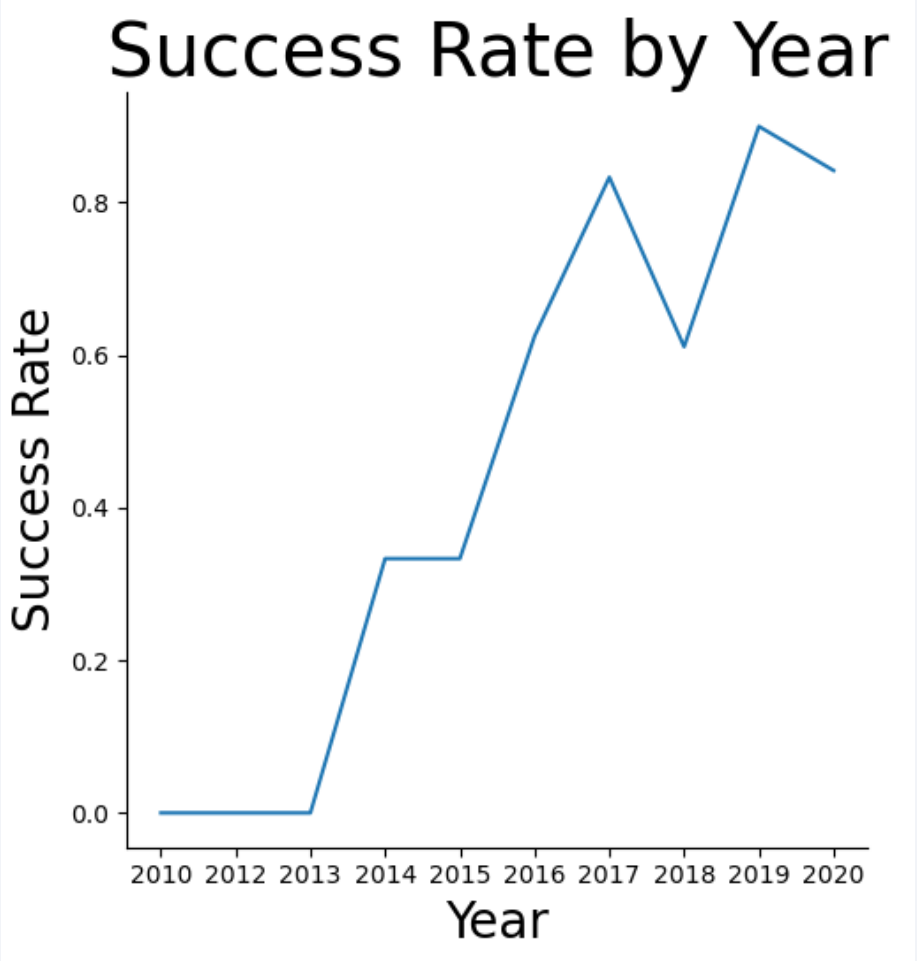
Payload vs. Orbit Type

- The scatter plot visualizes the spread of the launches based on payload mass and orbit type and is coloured based on mission outcome.
- With heavier payloads, the successful/positive landing rate are more common for Polar, LEO and ISS. These are positive correlations.
- However, for GTO we cannot distinguish a relationship due to many positive and negative results clustered.
- There are clusters for ISS at 2500 kg, GTO between 2500 kg and 7000kg, and VLEO at 15,000 kg.



Launch Success Yearly Trend

- The line chart visualizes the yearly average success rate of the launches
- The launch success rate has increased overall since 2013 to just over 83%. It had also flattened from 2019.



All Launch Site Names

- This query is used to find the names of the unique launch sites. This presents four unique launch site names, if we exclude ‘None’.

```
%sql SELECT DISTINCT Launch_Site AS UniqueLaunchSite FROM SPACEXTBL
```

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

UniqueLaunchSite
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

- This query is used to find 5 records where launch sites begin with ‘CCA’. The output gives the results for data with respect to Data, Time, Booster Version, Launch Site, Payload, Payload Mass, Orbit, Customer, Mission Outcome, and Landing Outcome.

```
%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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- This query is used to calculate the total payload carried by boosters from NASA. The outcome is a total payload mass of 45596 kg.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS 'TotalPayloadMass(KG)' FROM SPACEXTBL WHERE CUSTOMER IS 'NASA (CRS)'
```

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

TotalPayloadMass(KG)
45596.0

Average Payload Mass by F9 v1.1

- This query is used to calculate the average payload mass carried by booster version F9 v1.1. The result is an average of 2928.4 kg.

```
%sql SELECT avg(PAYLOAD_MASS__KG_) AS 'AVGPayloadMass(KG)' FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

AVGPayloadMass(KG)
2928.4

First Successful Ground Landing Date

- This query is used to find the date of the first successful landing outcome on ground pad. The outcome is 22nd Dec 2015.

```
%sql SELECT Date FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success (ground pad)%' ORDER BY Date DESC LIMIT 1
```

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

Date
22/12/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query is used to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000kg but less than 6000kg. Four booster versions are listed in this range.

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME LIKE 'Success (drone ship)%' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

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 query is used calculate the total number of successful and failure mission outcomes. 100 out of 101 total mission outcomes are total successes.

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS TotalCount FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%'
```

```
%sql SELECT SUM(CASE WHEN Mission_Outcome LIKE '%Success%' THEN 1 ELSE 0 END) AS TOTALSUCCESS, SUM(CASE WHEN Mission_Outcome LIKE '%Failure%' THEN 1 ELSE 0 END) AS TOTALFAILURE FROM SPACEXTBL
```

TOTALSUCCESS	TOTALFAILURE
100	1

Boosters Carried Maximum Payload

- The query is used to list the names of the booster which have carried the maximum payload mass.

```
%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

2015 Launch Records

- The query is used to list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

```
%%sql
SELECT substr(Date, 4, 2) AS month,
    Landing_Outcome AS failure_landing_outcomes,
    Booster_Version,
    Launch_Site
FROM
    SPACEXTBL
WHERE
    substr(Date, 7, 4) LIKE '%2015%'
    AND landing_outcome LIKE '%Failure%'
    AND booster_version IS NOT NULL
    AND launch_site IS NOT NULL;
```

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

month	failure_landing_outcomes	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The query is used to 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.
- “Success” as a landing outcomes had the highest count.

```
%%sql
SELECT landing_outcome, COUNT(*) AS count
FROM SPACEXTBL
WHERE date >= '04/06/2010' AND date <= '20/03/2017'
GROUP BY landing_outcome
ORDER BY count DESC;
```

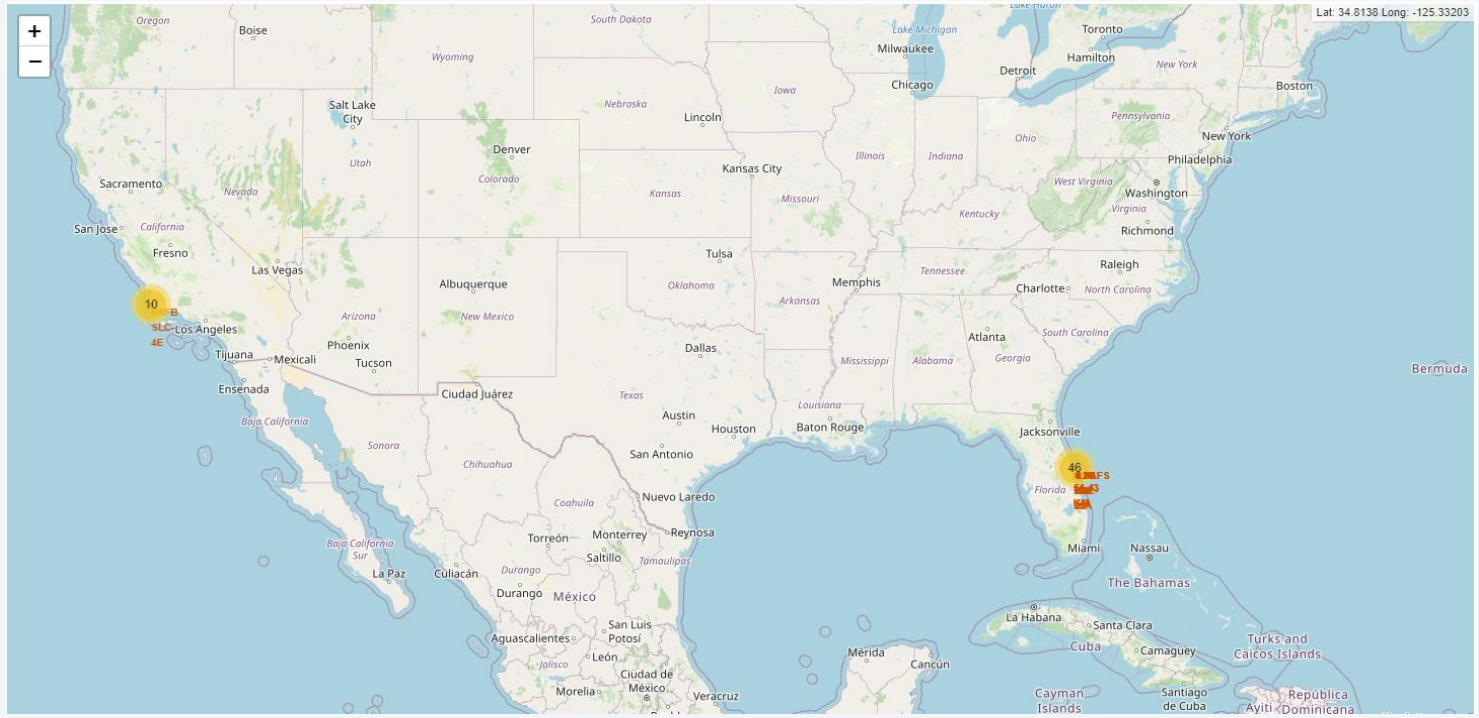
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1


Section 3
Launch Sites
Proximities Analysis

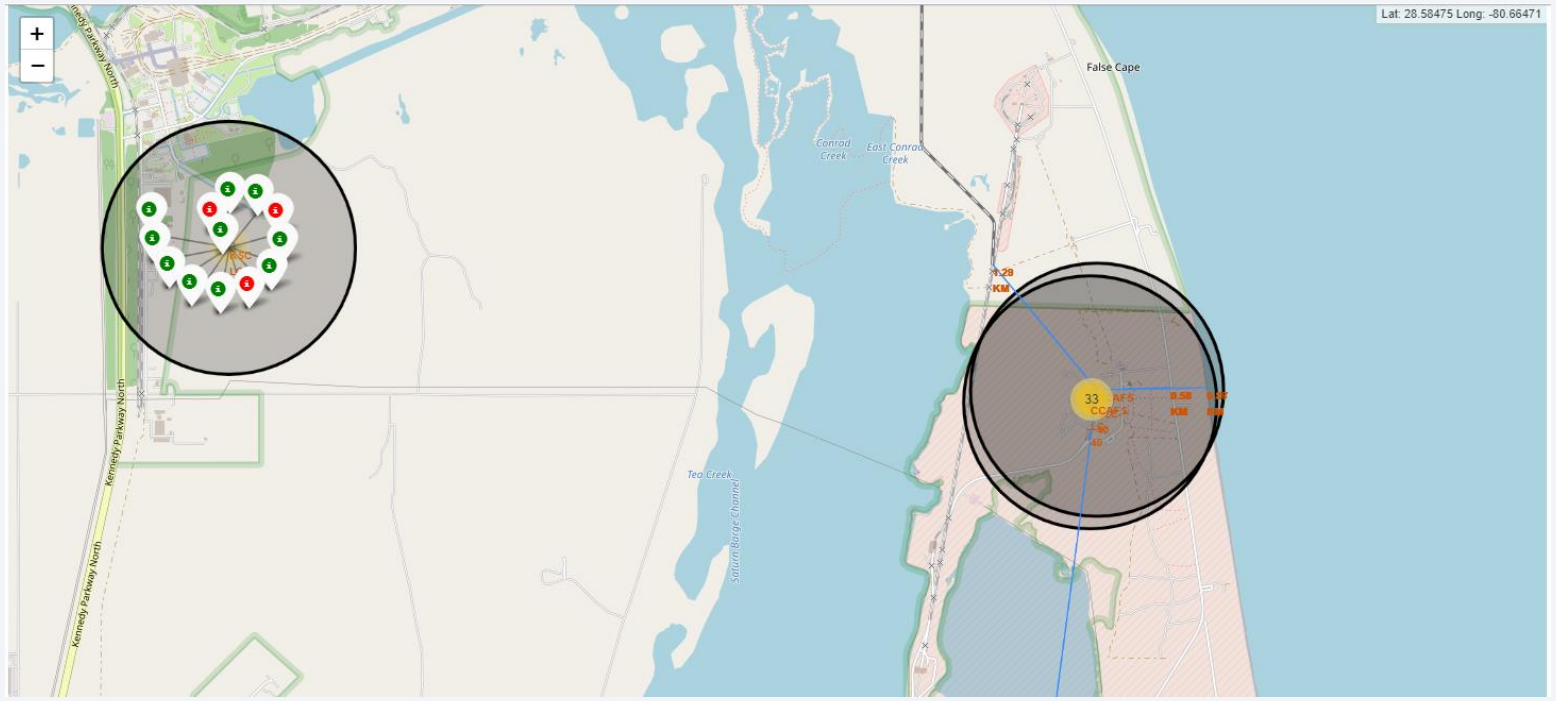
All Launch Sites on the Map

- The name of the launch site as a title is shown in the Folium Map image
- Circles and Markers are used for each launch site to easily identify them.



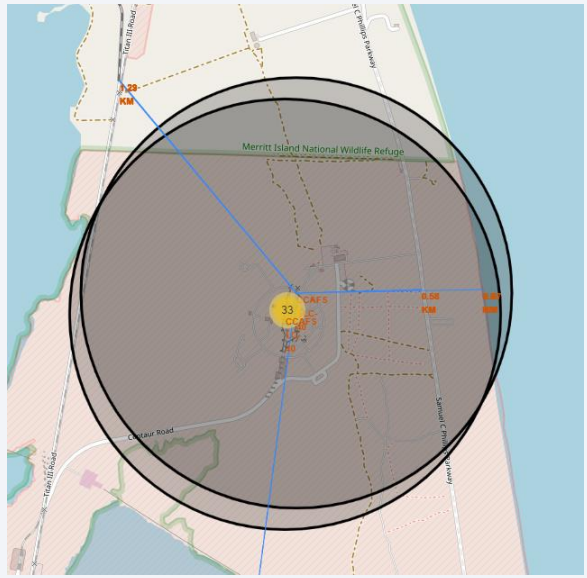
Successful/Failed Landing Markers

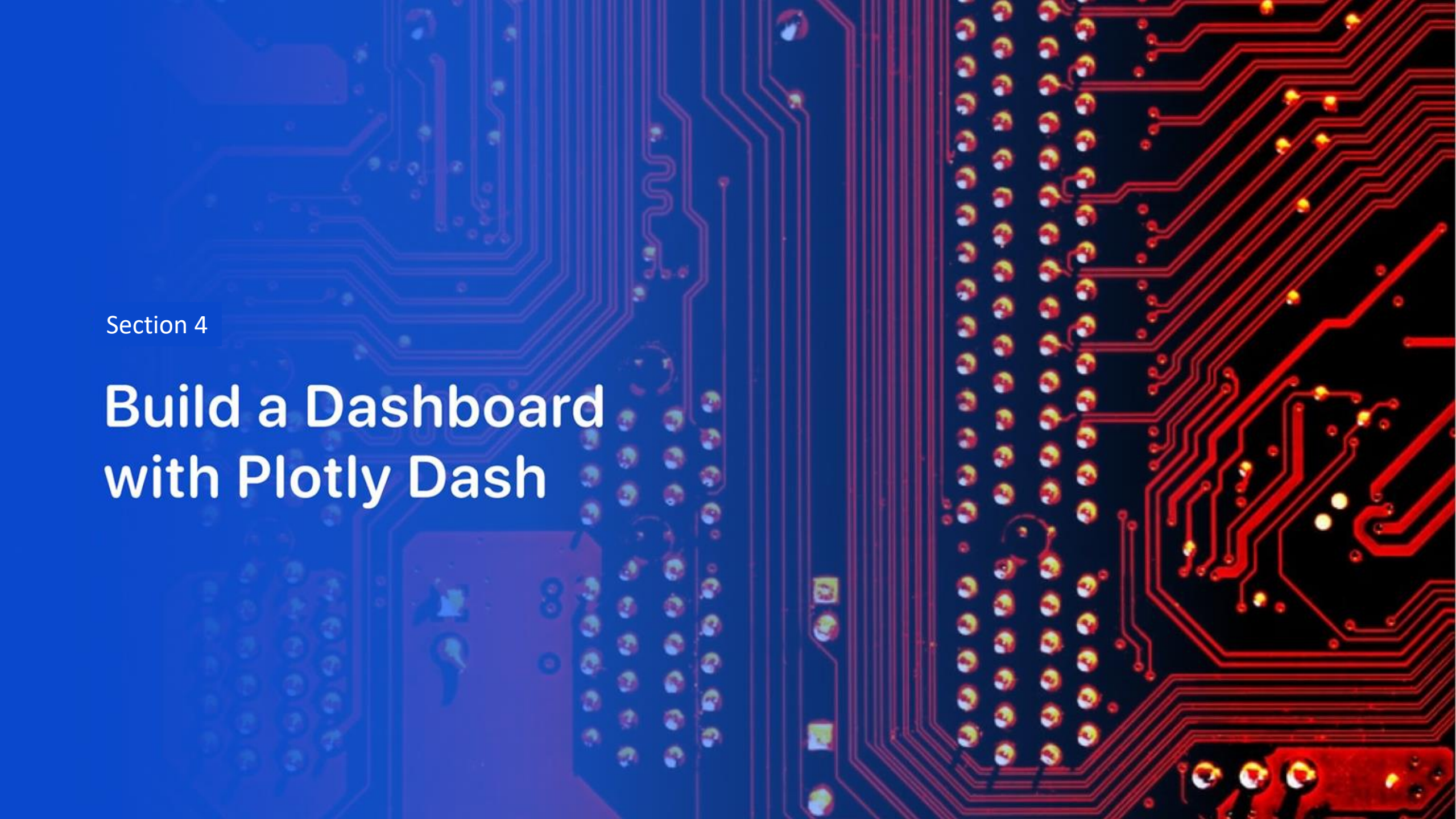
- The number of each launches per site is labelled by the Markers as shown. Also, the Markers are either red, to indicate a failed launch, or green, to indicate a successful launch.



Distances between launch sites to its proximities

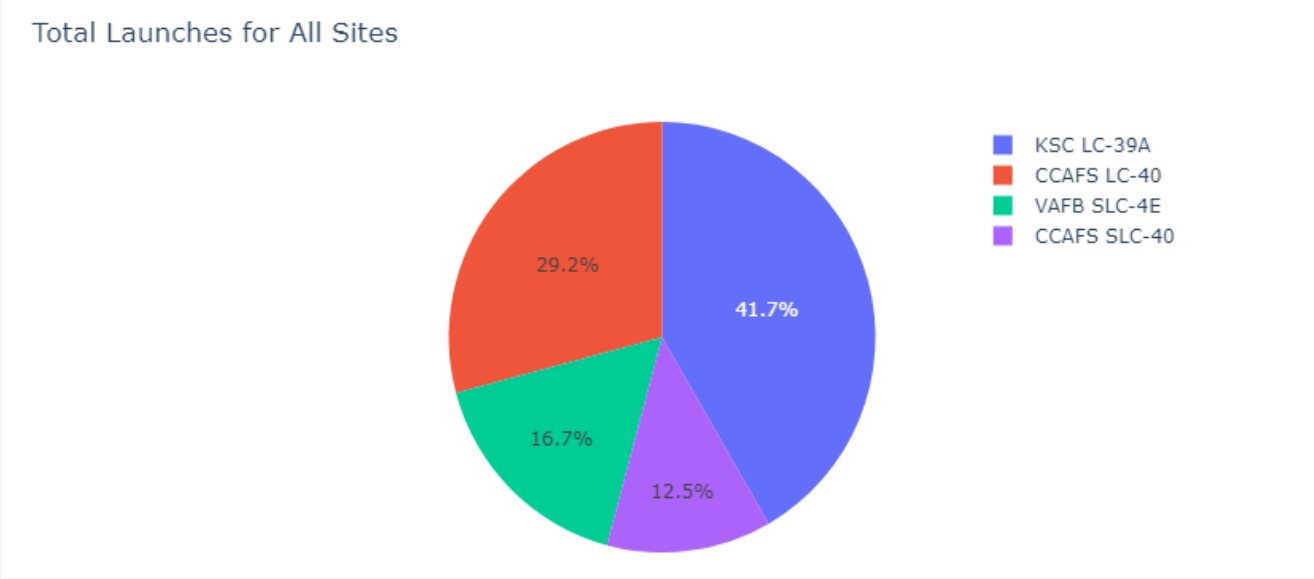
- The images show the distances and line that travel from the selected launch site to its closest proximities such as railway, highway, coastline, with distance calculated and displayed.
- The launch site used to measure the distances to the various proximities is CCAFS SLC-40.




Section 4
Build a Dashboard with Plotly Dash

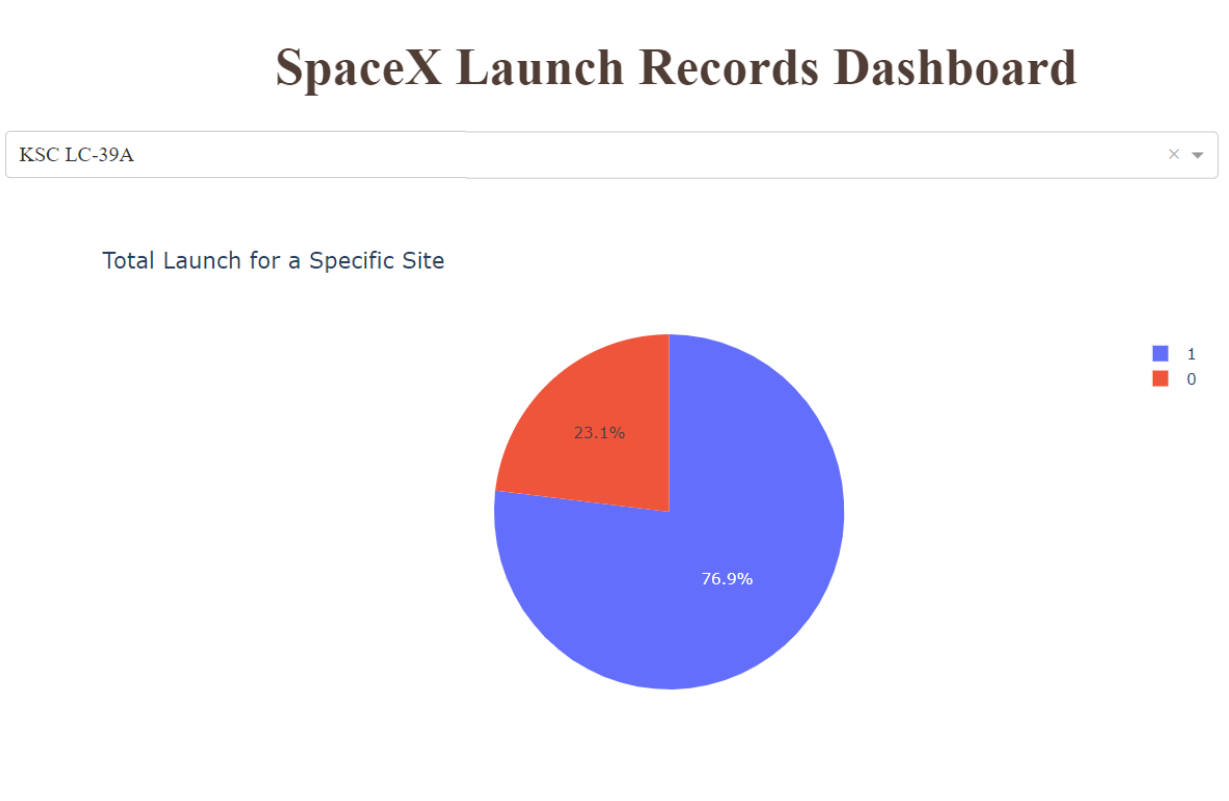
Total Launches by All Launch Sites

- The pie chart shows the total launches by all launch sites. It is found that KSC LC-39A had the highest value relative to the other launch sites with the number of launches at 41.7% of the total.



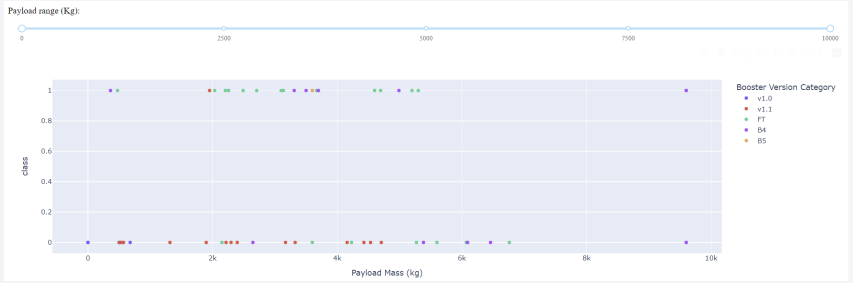
Launch Site with Highest Success Rate

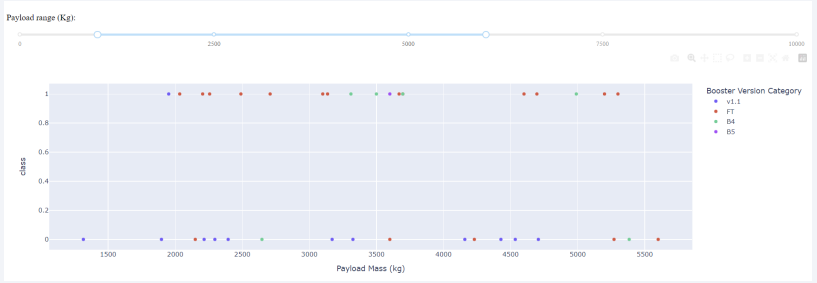
- The pie chart the illustrates the success rate of the launch site with the highest number of successful landings.
- KSC LC-39A achieved a 76.9% success rate of landings leaving a failure rate of 23.1%. This was the highest success rate of all launch sites.



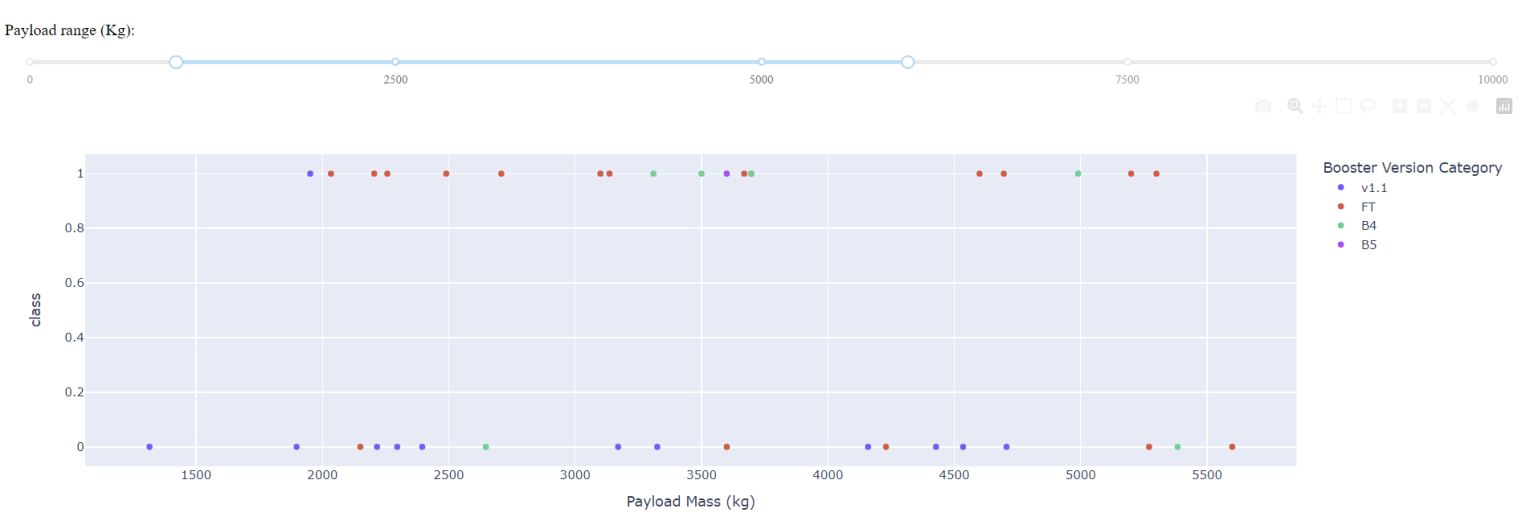
Payload Mass vs Launch Outcome (“Class”)

- The images – which are also shown in the next slide – visualize the spread of the data for all launch sites based on their success or failure (“Class”) and the payload mass (kg) based on their booster version.
- The payload mass range was also selected as between 1000kg and 6000kg to focus on the more successful landings that are found more frequently within that range.
- Typically, **more successful landings** found are at the **lower levels of payload mass**, but more specifically between **2000kg and 4000kg** (or arguably even up to 5500kg).
- **FT** is most frequently recorded as a **successful booster version** within this range.





Payload Mass vs Launch Outcome (“Class”)



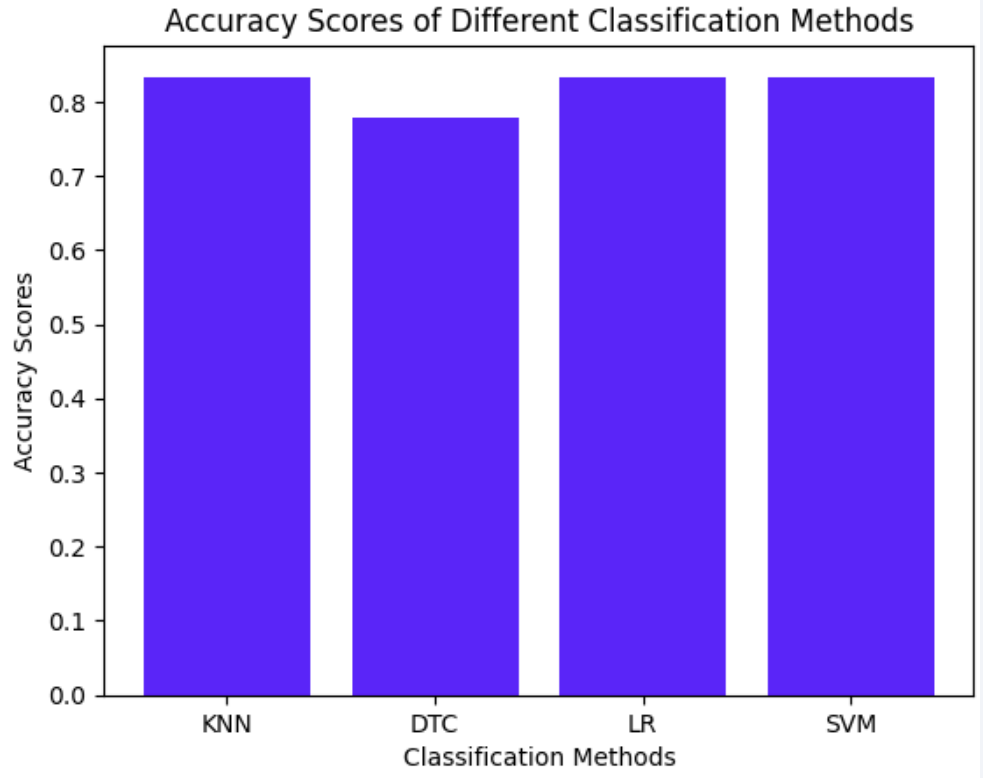
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Model accuracy for all built classification models

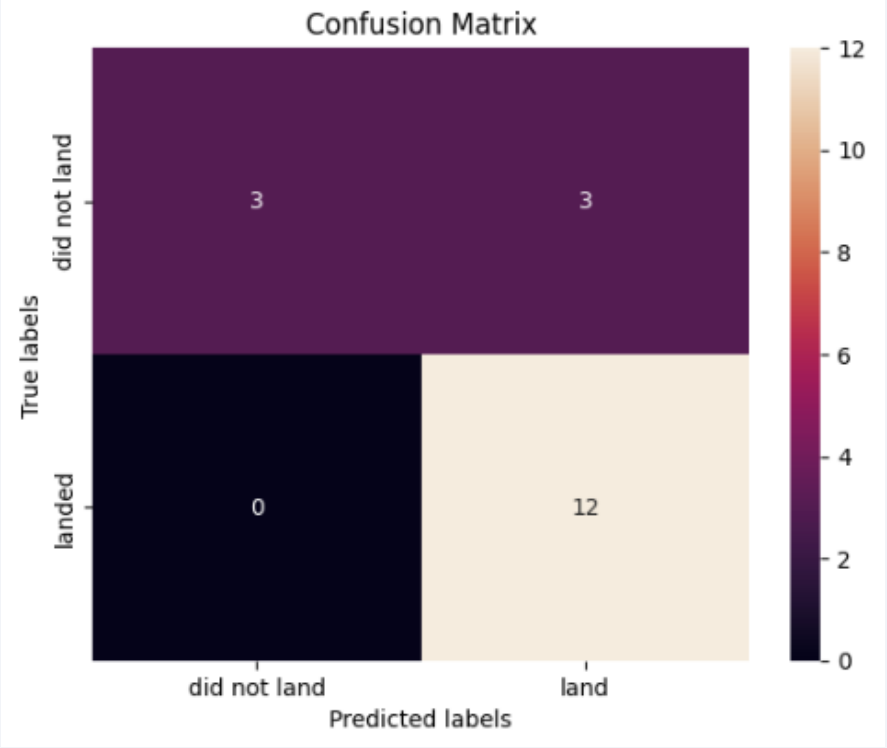
- As we can see in the bar chart, K-Nearest Numbers, Logistic Regression, and Support Vector Machine all equally had the highest accuracy score of 83.33%.
- Decision Tree Classifier had the lowest score at 77.77%.



Confusion Matrix

- KNN, SVM and Logistic Regression were the best performing model with an accuracy score of 83.33%
- From the confusion matrix we can identify that:

1. Of the True labels that did not land, the algorithm correctly predicted three of them, but mistook another three that did not land as ones that's did land – a **False Positive**, or **Type I Error**.
2. Of the True labels that did land, there were no incorrect predictions made where the algorithm assumed that those records didn't land. Twelve records were correctly predicted to land as per the predicted labels.



Conclusions

Exploratory data analysis conclusion:

- Through EDA, various discoveries are made regarding the variables that are investigated. For example, Heavy weighted payloads tend to perform better than lighter payloads; The success rates for SpaceX is positively correlated with the time in years spent on the launches; KSC LC 39A had the most successful launches from all sites; Orbits GEO, HEO, SSO, and ES L1 have the best success rate; FT is relatively the most successful booster version.

Interactive Analytics conclusion:

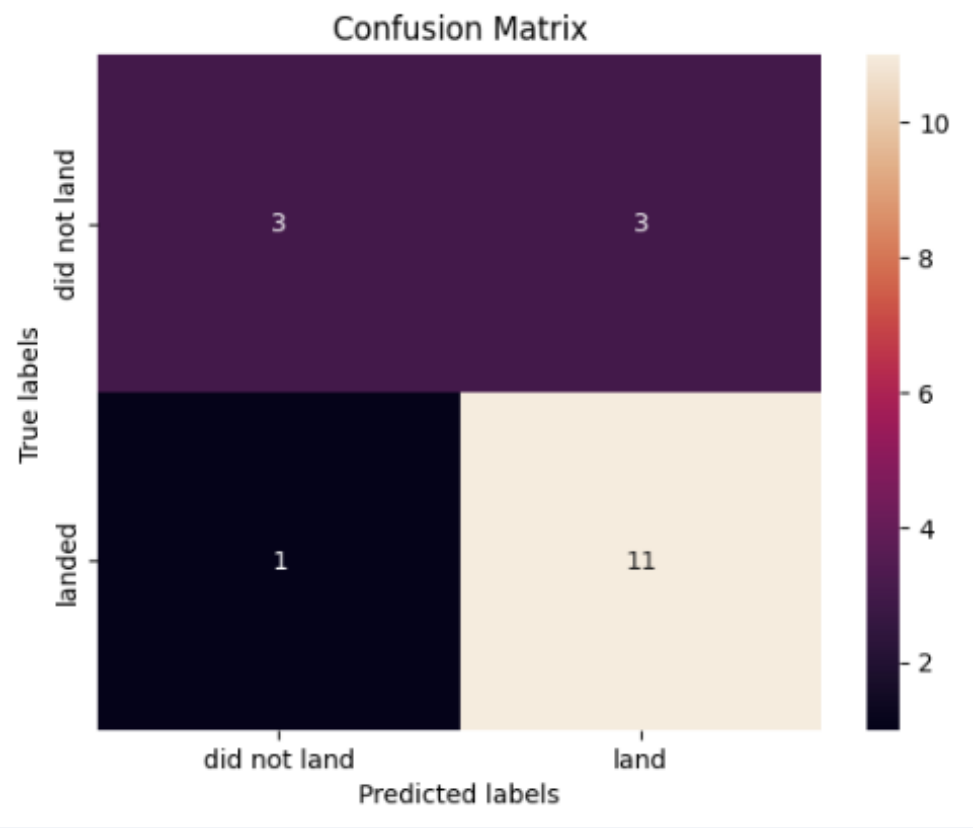
- Through Folium it is discovered that launch sites selected are relatively close to the coastline, but further from railways, highways, and especially cities. Through Plotly Dash, various relationships between variables are visualised to easily identify how they related with respect to different launch sites.

Predictive analysis results:

- Through predictive analysis, four classification models are developed to predict landing outcomes. Of these models, SVM, KNN, and Logistic Regression models are optimal with respect to prediction accuracy for the data set.

Appendix

- This is the Confusion Matrix of the Decision Tree Classifier. The only difference between this classification model and the others is the false negative, or Type II, error we find included in this one as the model predicted one count of a launch not landing when it, in fact, did land based on the True label. This caused the prediction accuracy of the model to fall.



Thank you!