IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Akinbanji Akinyera
06-15-2023

Outline

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

Executive Summary

- Summary of all methodologies
  - Dats Collection via API and Web Scraping
  - Explanatory Data Analysis with SQL and Data Visualization
  - Data Wrangling
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive maps and Dashboard Analysis results
  - Predictive Analysis results

Section 1

Methodology

Introduction

- Project background and context
  - The objective of the project is to determine if the first stage of the SpaceX Falcon 9 rocket will land successfully. This is required by company SpaceY (imaginary) that wants to compete with SpaceX. SpaceX advertises Falcon 9 rocket launches on its website with a cost of $62M. Other providers cost upward of $165M each. Much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

- Problems you want to find answers
  - To determine the price of each launch.
  - To determine if SpaceX will reuse the first stage.

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX data was extracted from two sources:
    - From (https://api.spacexdata.com/v4/launches/past) via SpaceX Rest
    - From (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922) via Webscraping
- Perform data wrangling
  - A landing outcome label called landing_class was created using Outcome column to distinguish between successful landing outcome and unsuccessful one.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
  - Data was standardized, spilt into train and test data and models were developed based on four different types of classification models. The best model was chosen based on accuracy.

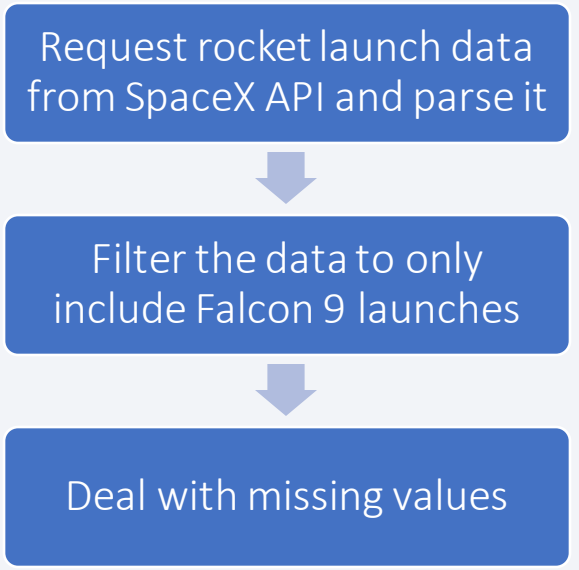
Data Collection

- Data was collected from (https://api.spacexdata.com/v4/launches/past) using SpaceX Rest API and from (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922) using Web scraping.

Data Collection – SpaceX API

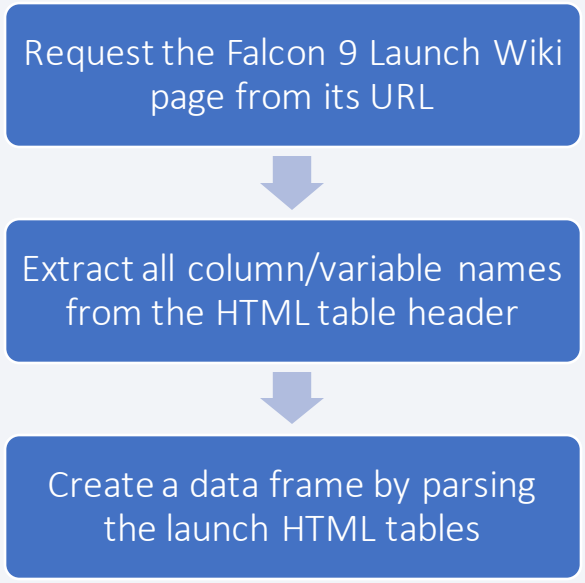
- SpaceX launch data was obtained using API. The API is based on the flowchart on the right side.

- Source code: https://github.com/Beejay7621/IBM-Applied-Data-Science-Capstone/blob/main/Data_Collection_With_API.ipynb



Data Collection - Scraping

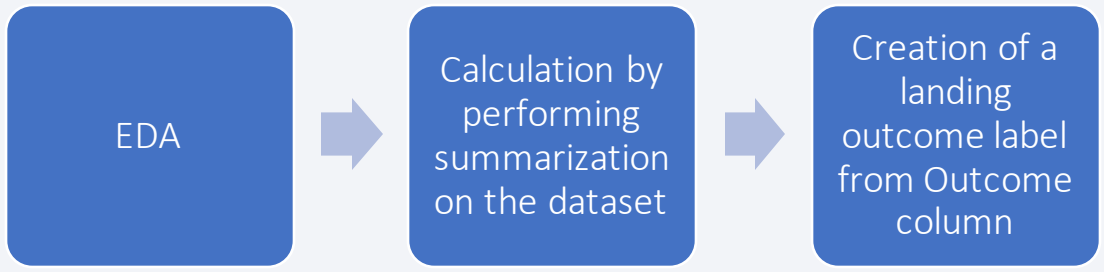
- SpaceX launch data was also obtained from Wikipedia via Web Scraping. The process is based on the flow chart on the right side.



- Source code: https://github.com/Beejay7621/IBM-Applied-Data-Science-Capstone/blob/main/Date_Collection-Web_Scraping.ipynb

Data Wrangling

- Exploratory Data Analysis was performed on the dataset initially.
- Then the number of launches on each site, the number and occurrence of each orbit, the number and occurrence of mission outcome per orbit type were calculated.
- Finally, a landing outcome label called Class was created from Outcome column.

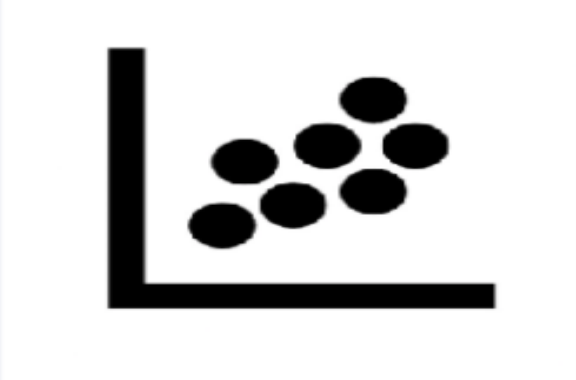


Data Wrangling

- Source code: https://github.com/Beejay7621/IBM-Applied-Data-Science-Capstone/blob/main/SpaceX_Data_Wrangling.ipynb

EDA with Data Visualization

- Scatter plots were used to visualize the relationships between Pay Load Mass and Flight Number, Launch Site and Flight Number, Launch Site and Pay Load Mass, Orbit and Flight Number, Orbit and Pay Load Mass. A scatter plot can be represented by the diagram below.



EDA with Data Visualization

- A bar chart was plotted to visualize the relationship between Success rate and Orbit type. A bar chart can be represented by the diagram below.

EDA with Data Visualization

- A bar chart was plotted to visualize the relationship between Success rate and Year. A line chart can be represented by the diagram below.

EDA with Data Visualization

- Source code: https://github.com/Beejay7621/IBM-Applied-Data-Science-Capstone/blob/main/EDA_Visualization.ipynb

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 the average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes.
  - List the names of the booster versions which have carried the maximum payload mass
  - List the records which will display the month names failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015

EDA with SQL

- Rank the most successful landing outcomes between the dates 04-06-2010 and 20-03-2017 in descending order

Source code: https://github.com/Beejay7621/IBM-Applied-Data-Science-Capstone/blob/main/EDA_sqlite.ipynb

Build an Interactive Map with Folium

- Folium maps were created by making use of markers, circles, lines and marker clusters.
  - Markers represent successful and unsuccessful landing. Green indicates successful landing while red indicates unsuccessful landing.
  - Circles represents coordinates like NASA Johnson Space center and the launch sites.
  - A line is used to indicate distance between two coordinates.
  - Marker clusters contain group of events in each coordinate. For example, launches in a launch site.
- Source code: https://github.com/Beejay7621/IBM-Applied-Data-Science-Capstone/blob/main/Interactive_Visual_Analytics_With_Folium.ipynb

Build a Dashboard with Plotly Dash

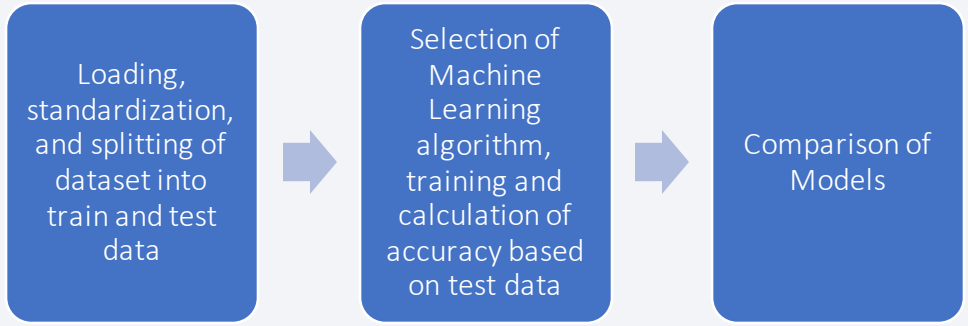
- The following interactions and plots/graphs were used to visualize the data.
  - A drop down menu which allows a user to choose a launch site or all sites.
  - A range slider which allows a user to choose a range of payload mass.
  - A pie chart to show the percentage of successful launches for all sites and percentage of successful and unsuccessful launches for each site.
  - A scatter plot which shows the relationship between payload mass and mission outcome (class).

Source code: https://github.com/Beejay7621/IBM-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Predictive Analysis was carried out through these processes:

- The dataset was loaded, standardized, split into train and test data.
- Machine learning algorithm was selected and parameters were set for each algorithm to GridSearchCV. The model was then trained using the training dataset.
- Models were compared based on accuracy and the best model was eventually chosen.



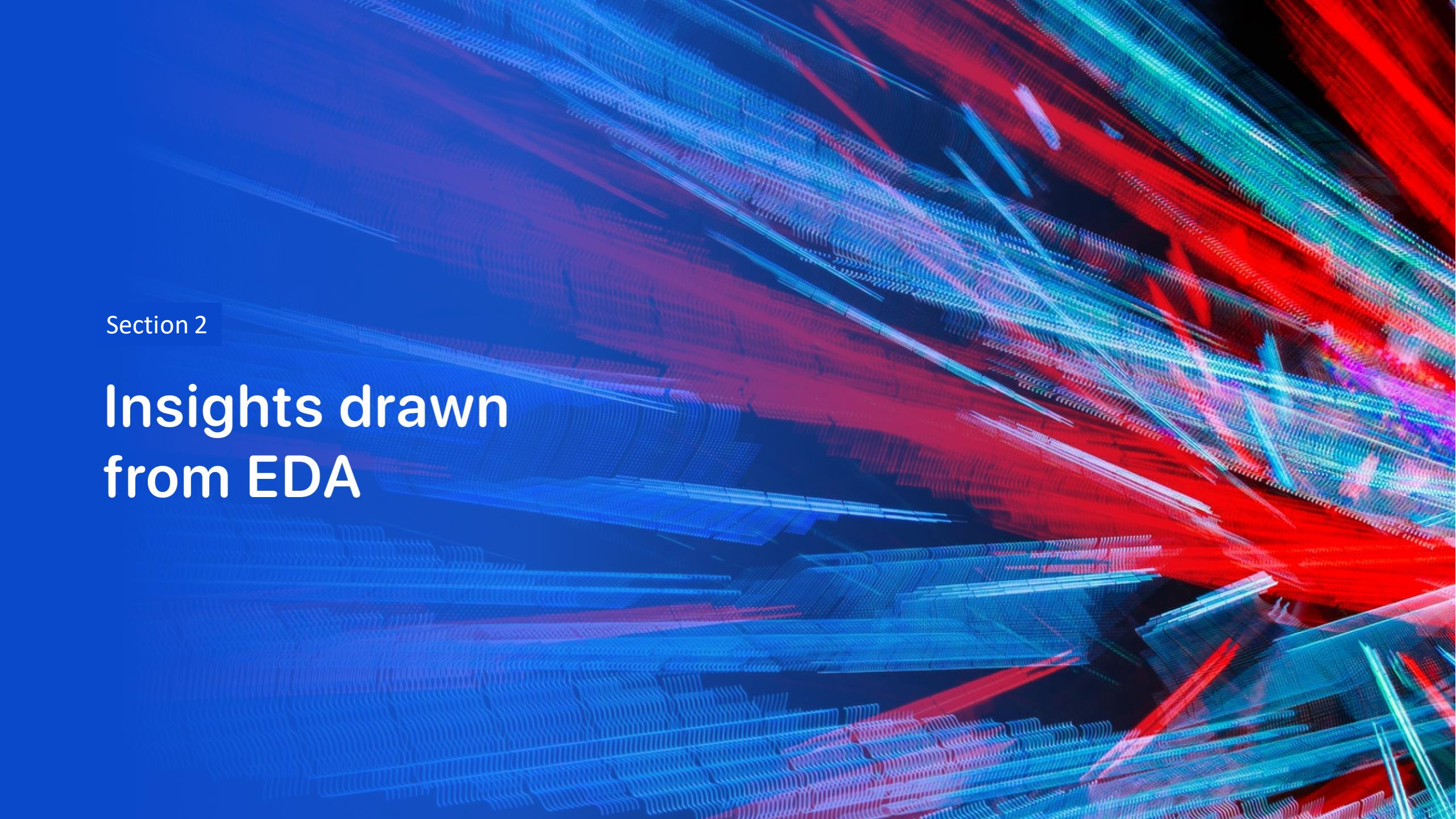
Source code: https://github.com/Beejay7621/IBM-Applied-Data-Science-Capstone/blob/main/Spacex_Machine_Learning_Prediction.ipynb

Results

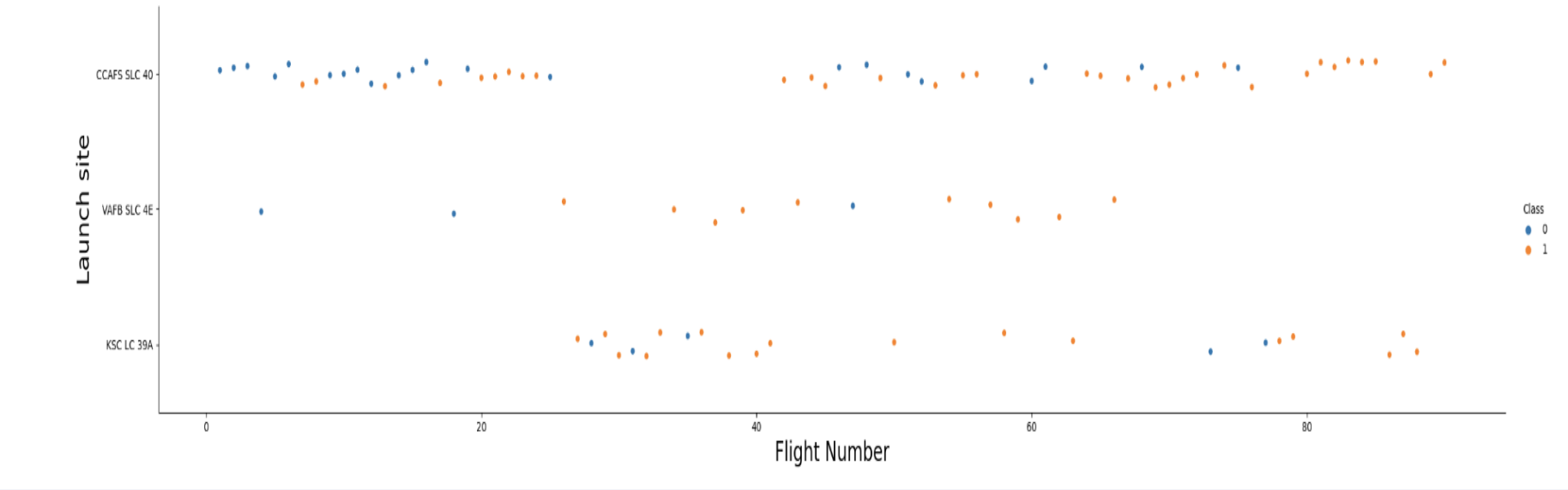
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



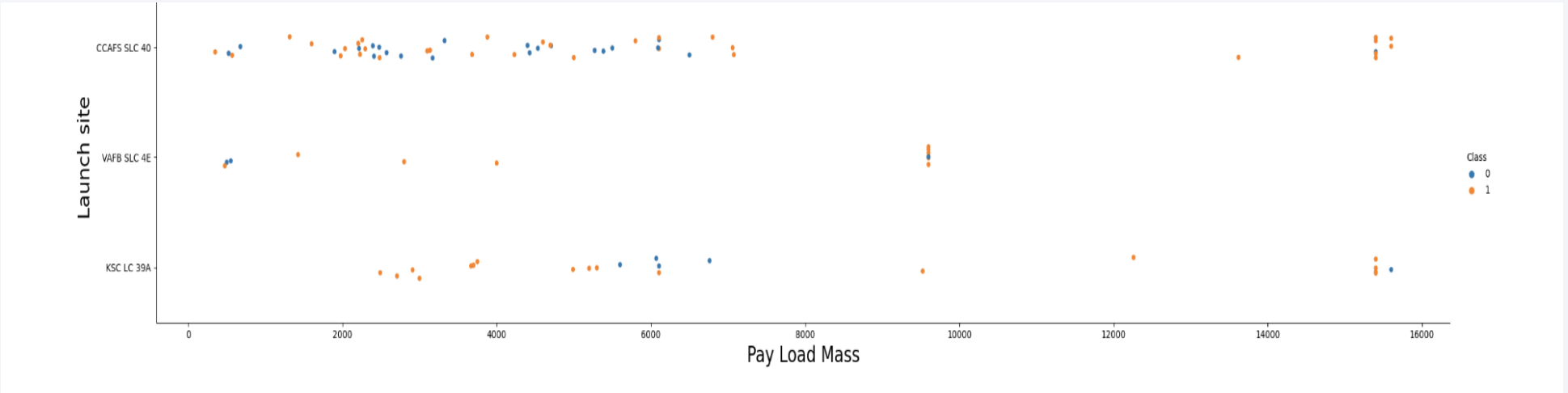
Flight Number vs. Launch Site



From the chart above, it can be deduced that:

- Success rate increases as Flight number increases.
- CCAFS SLC-40 has the highest number of launches and most successful recent launches.

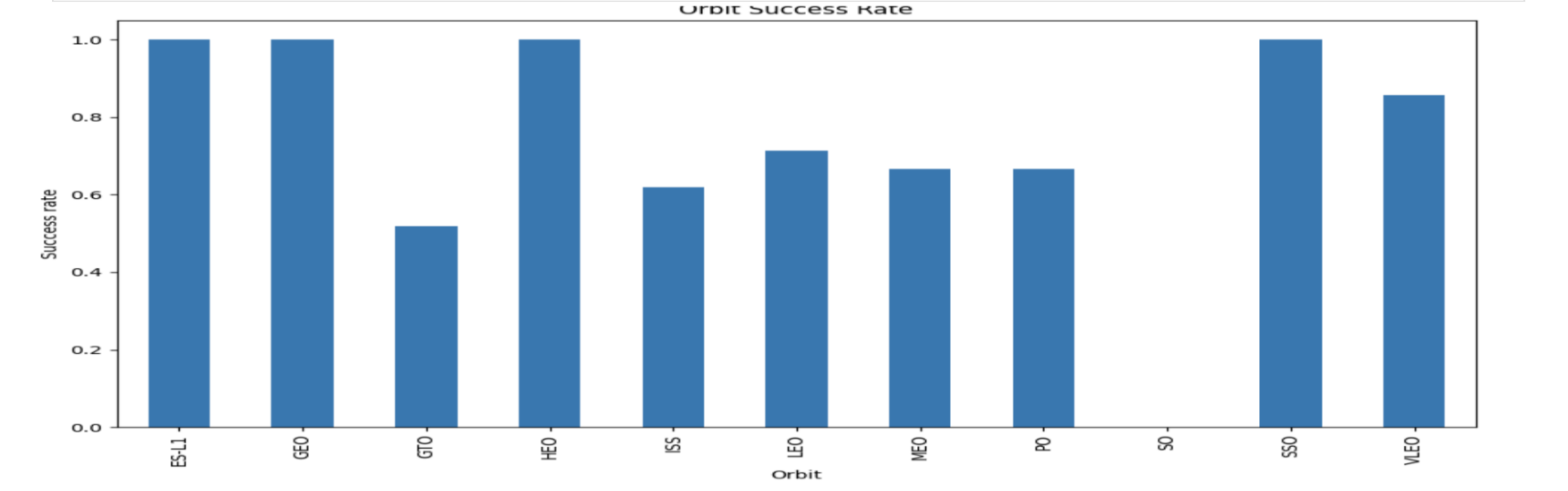
Payload vs. Launch Site



From the chart above, it can be deduced that:

- From payload greater than 8000kg, all the launch sites have relatively higher successful launches.
- For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass greater than 10000kg.

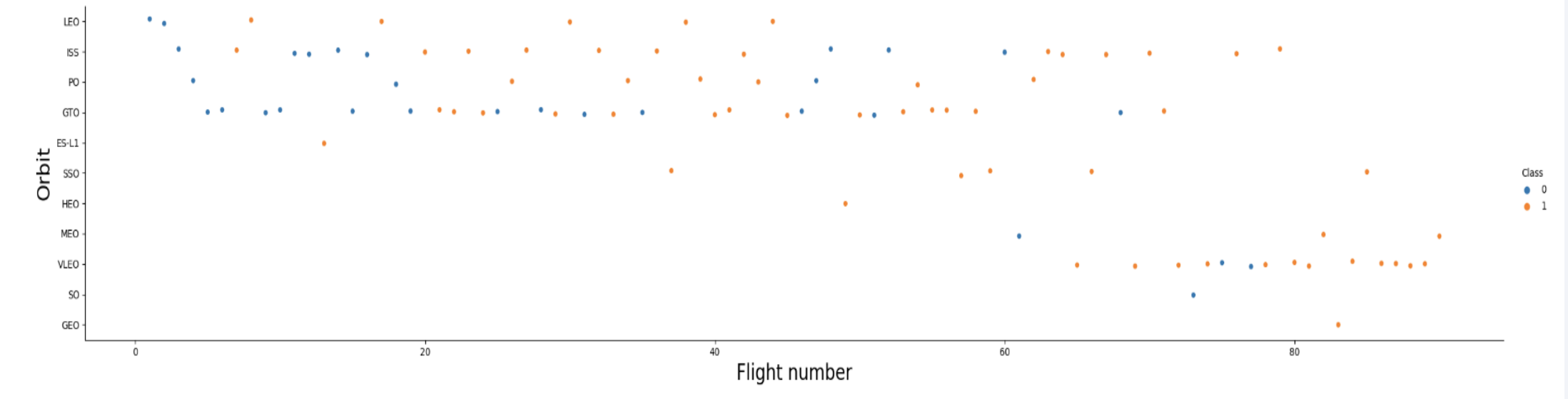
Success Rate vs. Orbit Type



From the chart above, it can be deduced that:

- ES-L1, GEO, HEO and SSO have the highest success rate.

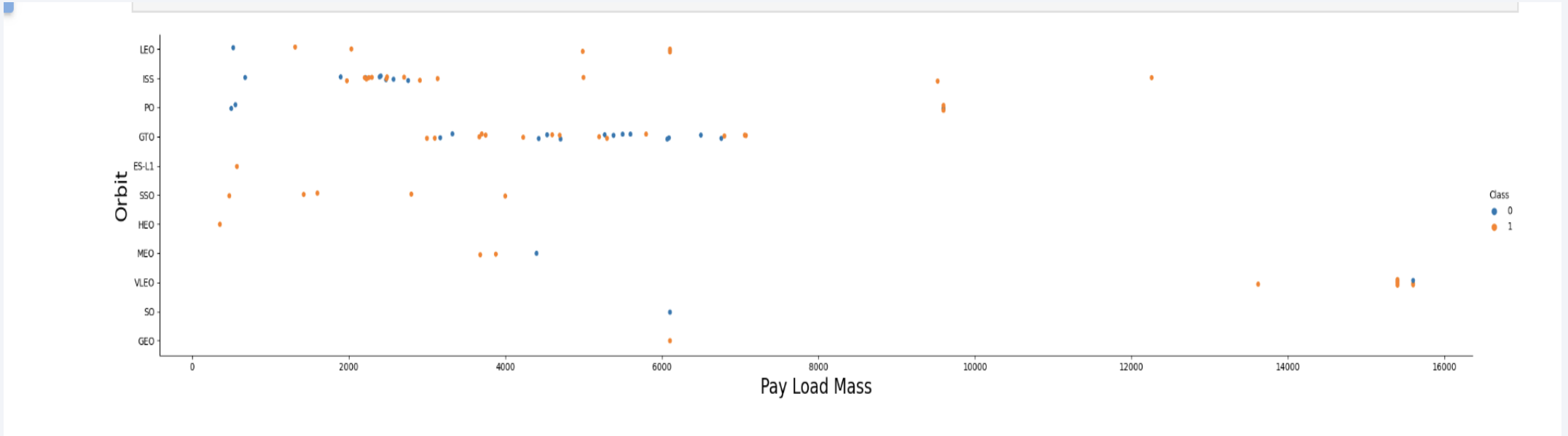
Flight Number vs. Orbit Type



From the chart above, it can be deduced that:

- For the LEO orbit, success appears related to the number of flights. On the other hand, there seems to be no relationship for the GTO orbit.

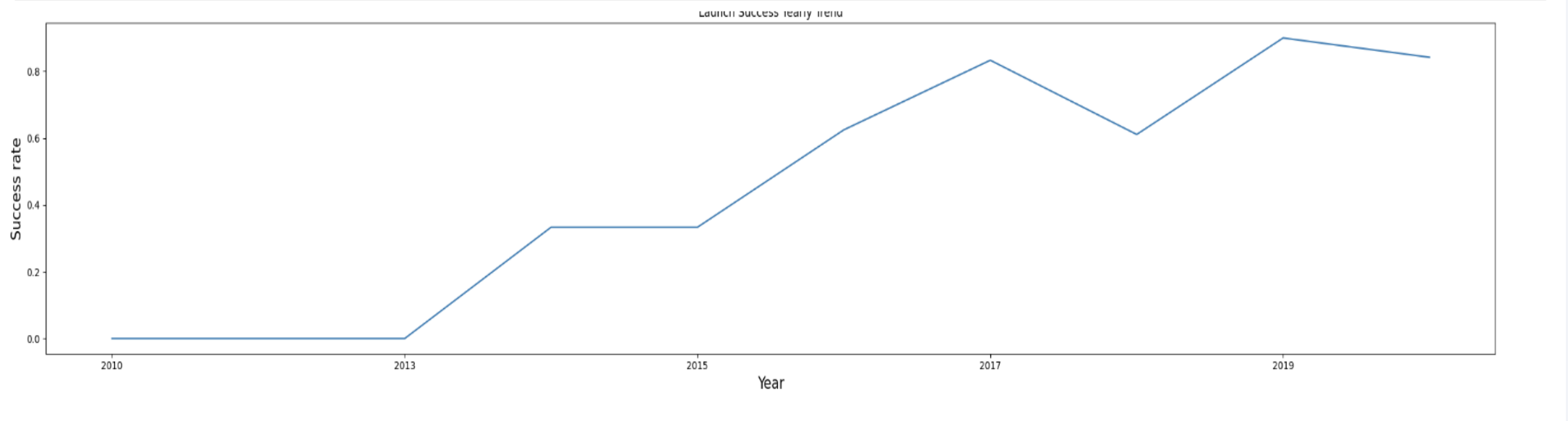
Payload vs. Orbit Type



From the chart above, it can be deduced that:

- The success rate improves with heavier payload mass for Orbit LEO and ISS.

Launch Success Yearly Trend



From the chart above, it can be deduced that:

- The success rate keeps increasing from 2013 to 2020.

All Launch Site Names

According to the data from the dataset, there are four launch sites:

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

These are obtained by selecting distinct values of "Launch_Site" from the dataset.

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with 'CCA':

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landin _Outcom
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failu (parachut
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	Failu (parachut
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

The table above displays 5 records from the dataset where launch sites begin with letters 'CCA'

Total Payload Mass

Total payload mass by boosters launched by NASA (CRS):

TOTAL_PAYLOAD_MASS
45596

The table above displays the total payload mass carried by customer NASA (CRS)

Average Payload Mass by F9 v1.1

Average payload mass by booster version F9 v1.1:

AVERAGE_PAYLOAD_MASS
2534.6666666666665

The table above displays the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad:

Earlliest_date
2015-12-22

The table above displays the date when first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters that have successfully landed on drone ship and with payload mass greater than 4000 but less than 6000.

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

The table above displays the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

SUCCESS	FAILURE
100	1

The table above displays the number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

The names of boosters which have carried the maximum payload mass:

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

The table above displays the names of the booster version which have carried the maximum payload mass

2015 Launch Records

The failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015:

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 table above displays the month, failure landing outcome in drone ship, booster versions, launch site for the year 2015.

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

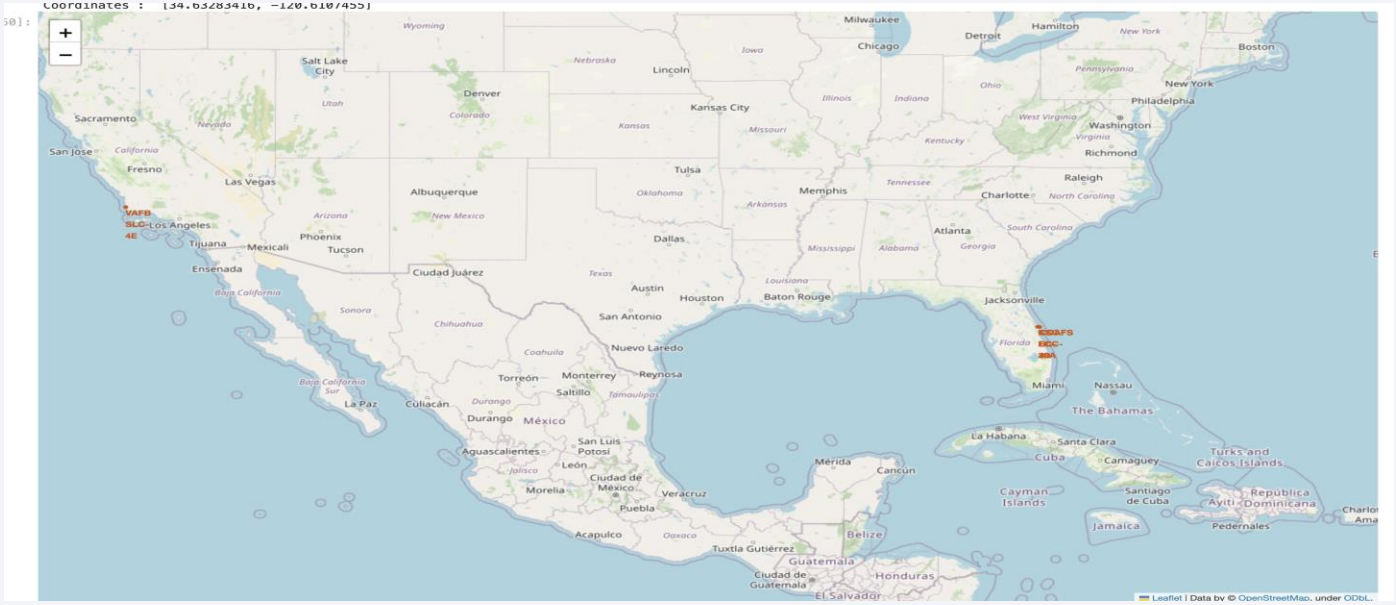
Rank of the count of successful landing outcomes between 2010-06-04 and 2017-03-20 in descending order:

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

The table above displays the rank of the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20.

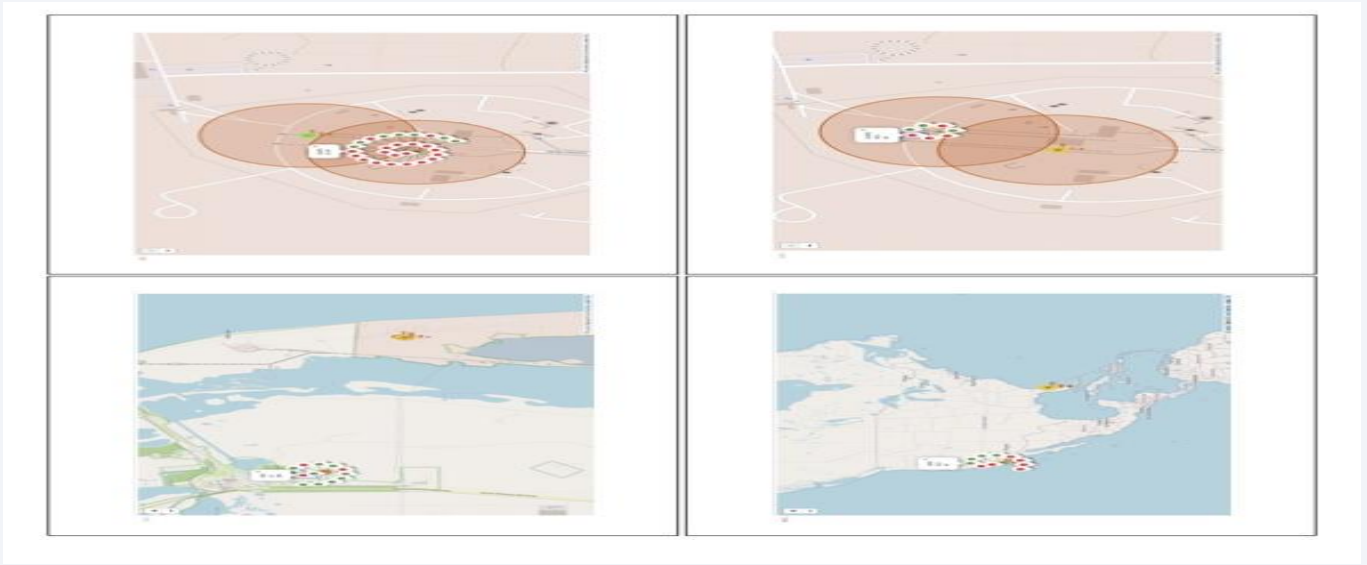

Section 3
Launch Sites
Proximities Analysis

Launch Sites



From the map above, we can see that all launch sites are located near the coast line.

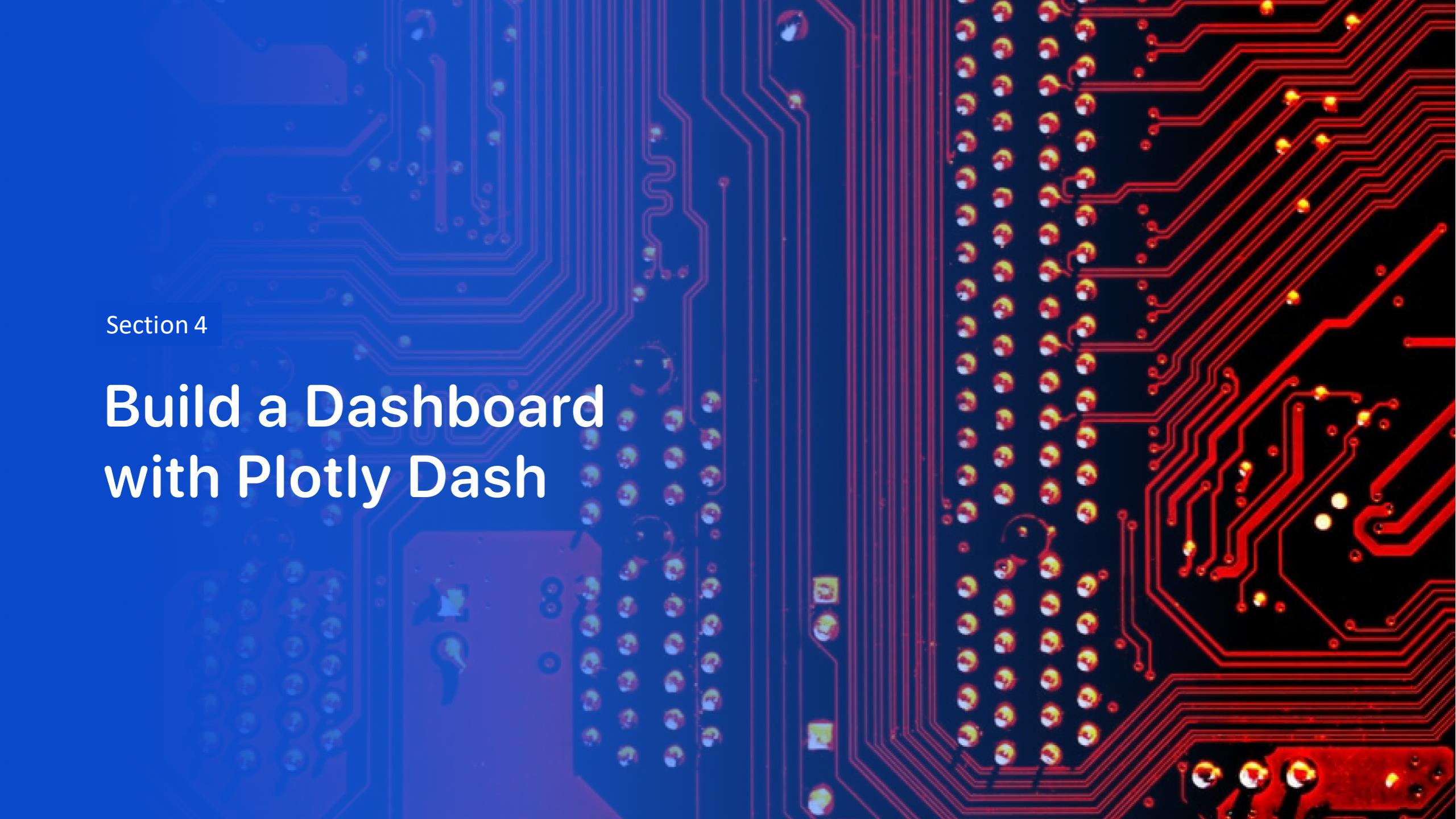
Mission Outcomes By Launch Sites



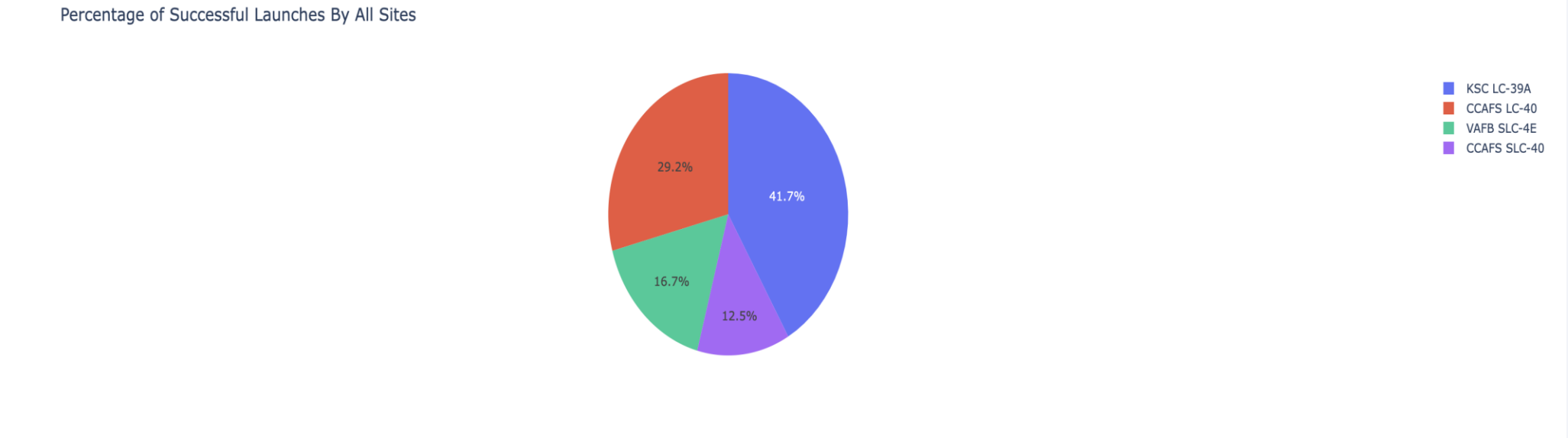
From the figure above, green marker represents successful launches while red marker represents unsuccessful outcomes. It can be seen that KSC LC-39A has the highest success rate of the launch sites.

Proximity of CCAFS SLC-40 To The Coast Line

From the map above, it can be seen that CCAFS SLC-40 is in close proximity to the coastline.

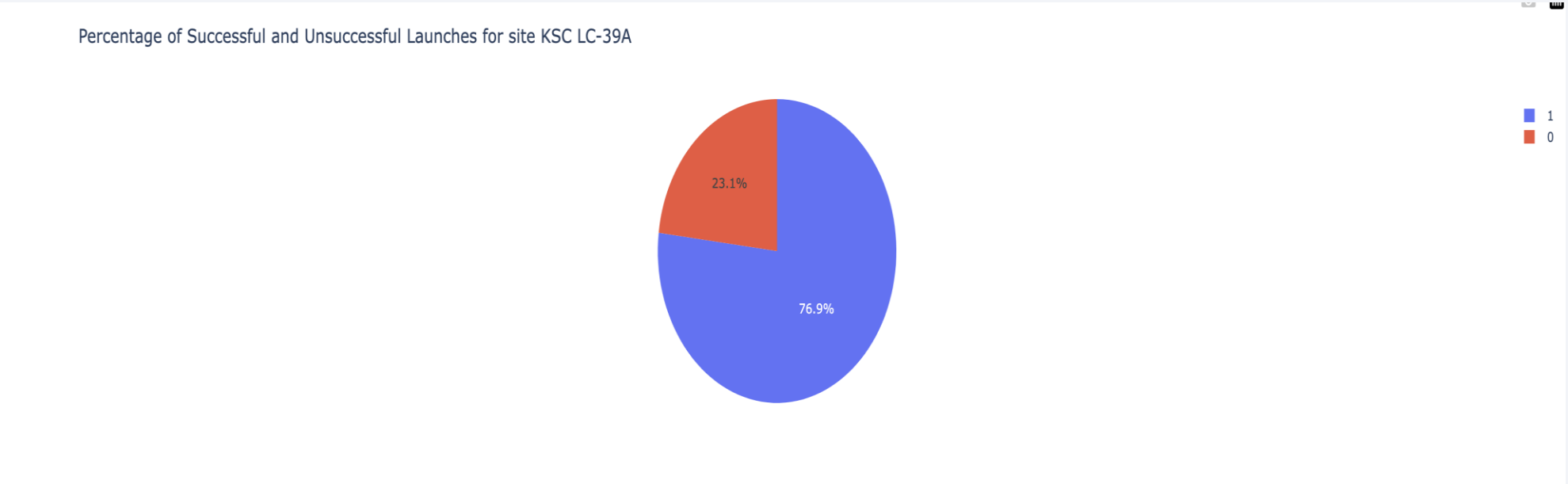

Section 4
Build a Dashboard with Plotly Dash

Percentage of Successful Launches By All Sites



From the figure above, it can be seen that KSC LC-39A has the highest success rate of all the launch sites.

Launch Success ratio for Site KSC LC-39A



1 indicates successful launches while 0 indicates unsuccessful launches. It can be seen that 76.9% of the launches are successful while 26.1% are unsuccessful.

Payload Vs Launch Outcome With Different Payload

The graph above shows the scatter plot of Launch Outcome and Payload range between 0 and 3000 kg

Payload Vs Launch Outcome With Different Payload

The graph above shows the scatter plot of Launch Outcome and Payload range between 0 and 5000 kg

Payload Vs Launch Outcome With Different Payload

The graph above shows the scatter plot of Launch Outcome and Payload range between 0 and 7000 kg

Payload Vs Launch Outcome With Different Payload

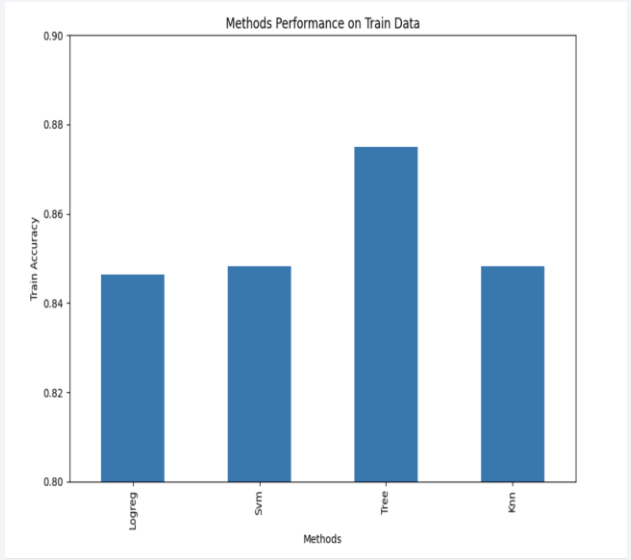
The graph above shows the scatter plot of Launch Outcome and Payload range between 0 and 10000 kg.
It can be concluded that both successful and unsuccessful launches increase with heavier payload for all sites.

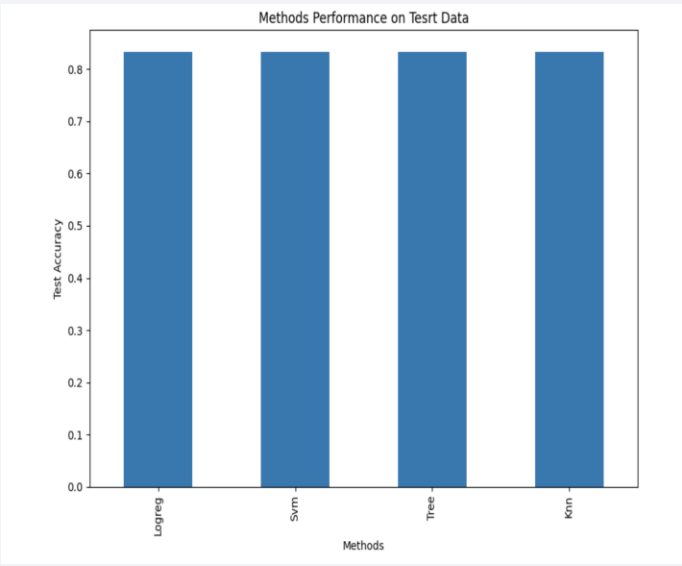
Section 5

Predictive Analysis (Classification)

Classification Accuracy

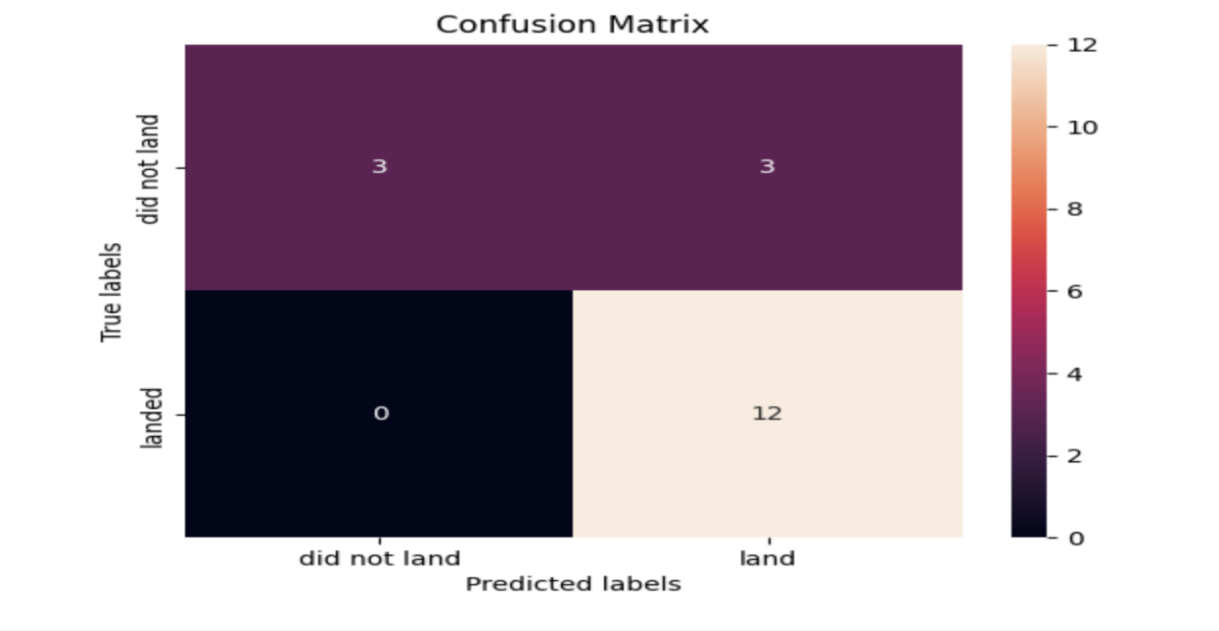
	Accuracy Train	Accuracy Test
Logreg	0.846429	0.833333
Svm	0.848214	0.833333
Tree	0.875000	0.833333
Knn	0.848214	0.833333





The test accuracies on all the four models are virtually the same but the train accuracy of Decision tree model (more than 0.87) is higher than the rest of them. Therefore Decision tree is the preferred model.

Confusion Matrix



Out of 6 rockets that did not land, the classifier correctly predicted 3 of them. On the other hand, out of 12 rockets that actually landed, the classifier correctly predicted all of them. We have equal true negative and false negative while the true positive is large (100%).

Conclusions

- CCAFS SLC-40 has the highest number of launches and most successful recent launches
- Orbit ES-L1, GEO, HEO and SSO have the highest success rate of all the orbit types.
- KSC LC-39A has the highest success rate of all the launch sites.
- Decision Tree Classifier is the best model to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully.

Thank you!