IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eric Yang
Jan 4th, 2024

Outline

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

Executive Summary

- This project will explore data gathered from SpaceX's public records and find useful insights through various methods.
- The key components of this project involve:
  - Data Collection and Wrangling
  - Exploratory Data Analysis with SQL and Visualization
  - Interactive Visualization Analytics with Folium
  - Machine Learning Predictions
- The key finds from the data exploration looked at the connection between certain features and their mission success rate.
- The machine learning experimentation found that from the various models tested, the decision tree model yielded the best accuracy in predicting whether the Falcon 9 first stage would land successfully.

Introduction

The commercial space age is upon us and companies are racing to reduce their costs for rocket launches. SpaceX is one of the more accomplished companies who advertises their cost of launch for the Falcon 9 rocket to be at $62 million while other companies have costs upwards of $165 million. Most of these reduced costs come from the fact that SpaceX can reuse their first stage.

For our company SpaceY to compete with SpaceX, we will be gathering data from their launches to help us determine the costs of launches, whether the first stage will land and train machine learning models to predict if the first stage will be reused.

Section 1

Methodology

Methodology

Executive Summary

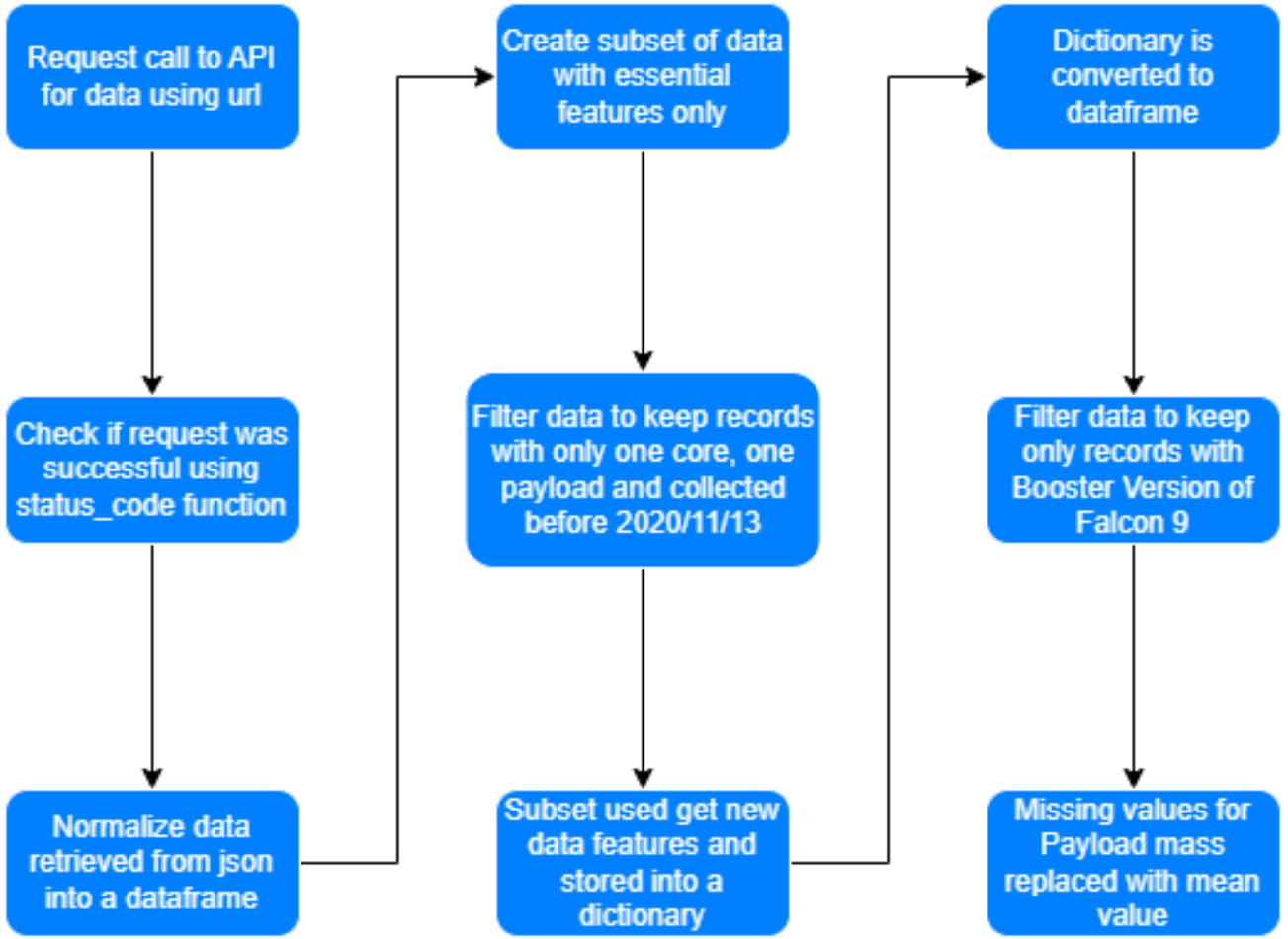
- Data collection methodology:
  - Through the SpaceX API
  - Through Web Scraping with the BeautifulSoup library
- Perform data wrangling
  - New class was created based on the outcome feature using binary values to represent the success or failure of a mission.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The dataset is first standardized and then split into a training and testing set. For each model, the best hyperparameters were found through tuning and evaluated based on their accuracy to identify the best performing model

Data Collection

- Datasets were collected through two methods: SpaceX API and Web Scraping.
- From the SpaceX API, data was extracted as a json file that contained various key data features.
- While the data gathered through web scraping was gathered through the SpaceX Wikipedia page using the BeautifulSoup library.
- Additional data wrangling methods were applied for both methods to clean and filter the data.

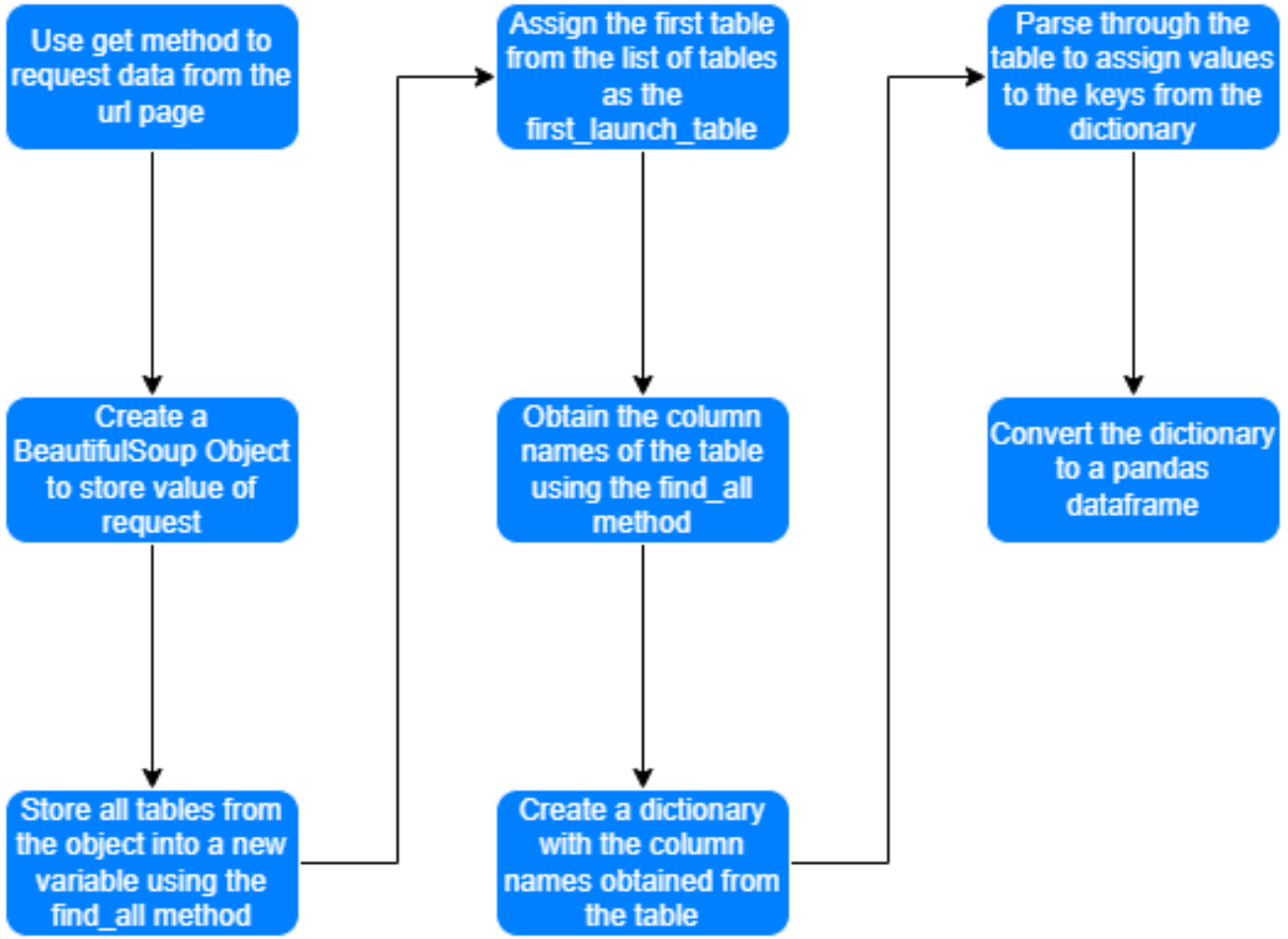
Data Collection – SpaceX API

- https://github.com/EricYang/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb



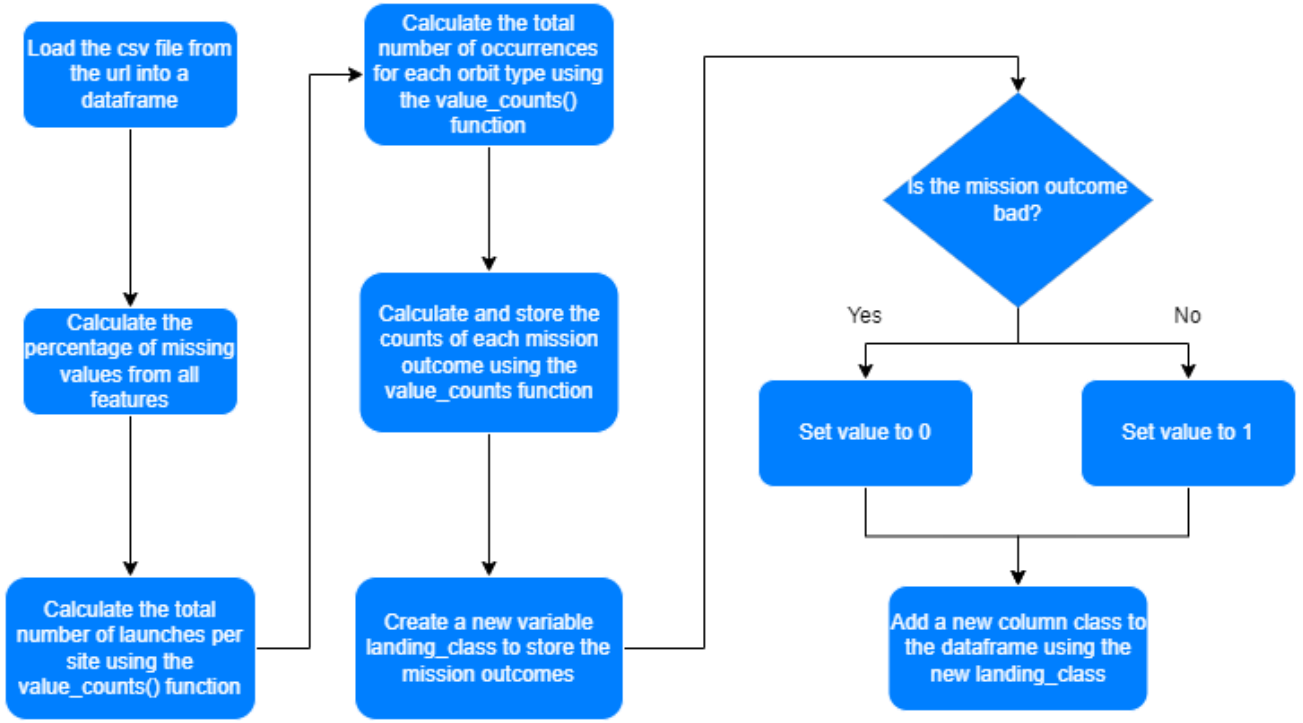
Data Collection - Scraping

- https://github.com/EricYang/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collecton%20with%20Web%20Scraping.ipynb



Data Wrangling

- https://github.com/EricYang/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization

- Scatter plot
  - Several scatter plots were used to try to identify any correlation between 2 variables, such as the Launch site and payload mass
- Bar chart
  - The bar chart helped showcase and compare values within a category
- Line chart
  - The line chart helped showcase the trend over time of a category
- https://github.com/EricYangg/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

- Create: created table entry from the imported dataset
- Distinct: pulled unique values from a feature
- Where: used to set condition on the desired data
- Sum, Avg, Min: mathematical operations performed on a feature
- Between: condition used in where clause to set two numerical boundaries for feature
- Group By: group data by feature
- Subquery: using the result of a query inside of another query to perform more complex queries
- https://github.com/EricYangg/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

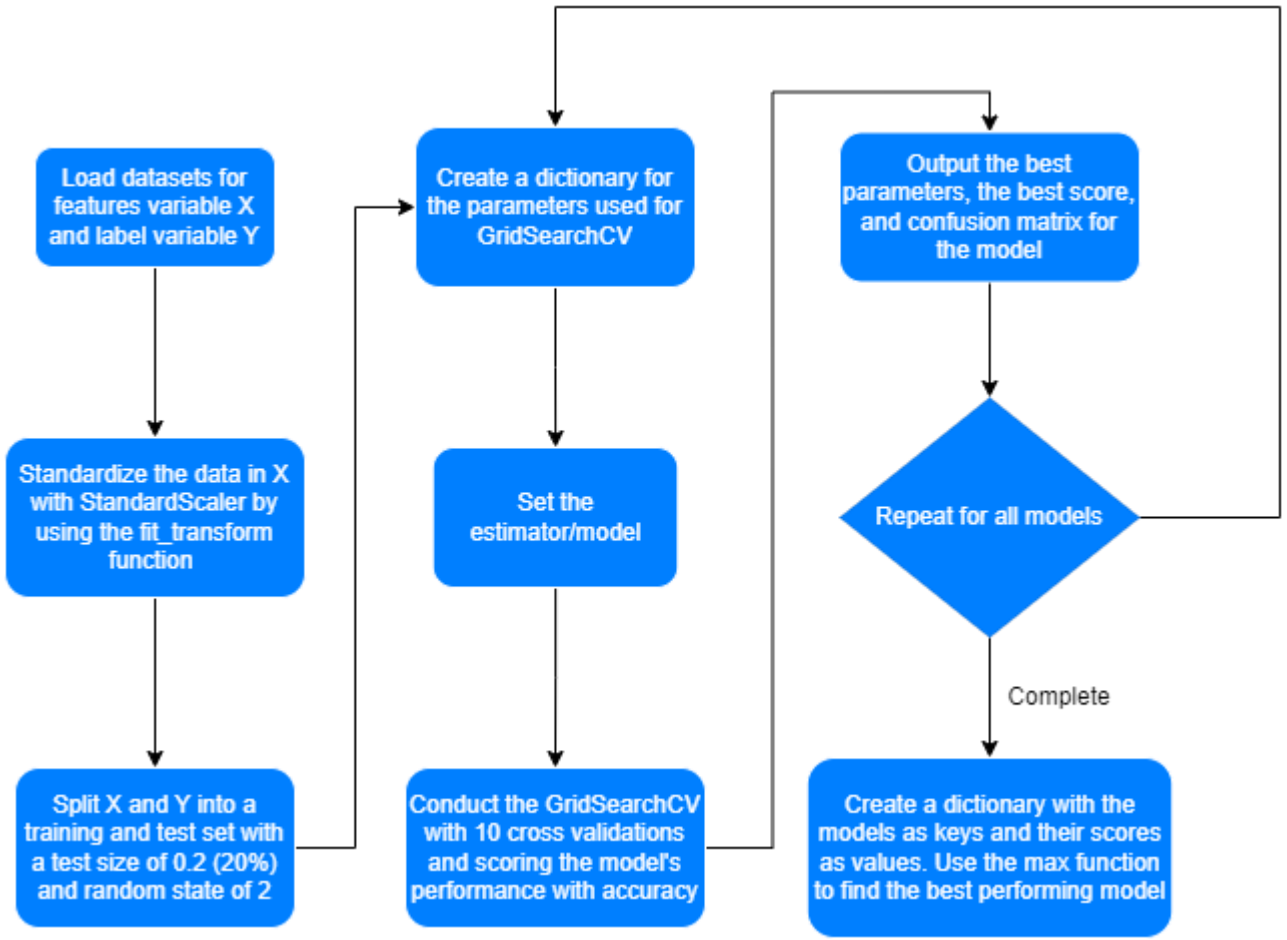
- Markers and Circles
  - Used to highlight key coordinates on the map such as launch sites as well as provide a label to give a brief description about the marker
- MarkerCluster
  - Used to group markers that are in a close range of each other allowing for the visual to be easier to understand
- PolyLine
  - Used to draw a line between two coordinates to help visualize distance between two points on the map
- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- https://github.com/EricYangg/IBM-Applied-Data-Science-Capstone/blob/main/Interative%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Pie Chart
  - This chart was used to success rate of each launch site
- Scatter Plot
  - This plot was used to showcase the success rate in relation to the payload mass
- Site Selection
  - This dropdown menu allowed the user to select a specific launch site to focus on or view all sites for both plots above
- Payload Mass Range
  - This scale was used to set upper and lower limits for the scatter plot's x axis, payload mass. This allowed the user to filter specific payload ranges to get a better understanding of the data
- Summarize what plots/graphs and interactions you have added to a dashboard
- https://github.com/EricYangg/IBM-Applied-Data-Science-Capstone/blob/main/Interactive%20Dashboard%20with%20Plotly%20Dash.py

Predictive Analysis (Classification)

- Data is first standardized and then split into 80/20 training and test split
- For each model
    - Create a dictionary of parameters for grid search
    - Set the estimator/model for the search
    - Conduct GridSearchCV with a cross validation of 10 and scoring the performance on accuracy
    - Output the best parameters, best score and confusion matrix
- Find best model by creating a dictionary with the model names and scores. Use the max function to find the model with the highest score.
- https://github.com/EricYangg/IBM-Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Predictions.ipynb



Results

- Exploratory Data Analysis Results:
  - 4 unique launch sites – See Table 1
  - Between 2010-06-04 and 2017-03-20, the most frequent landing outcome was No Attempt, and the least was Precluded (Drone Ship) – See Table 2
- Interactive Analytics Results
  - Launch sites are located close to the coastlines and in the southern areas of the country – Image 1
  - Further breakdown in Section 3
- Predictive analysis results
  - Decision Tree was best performing model

Table 1

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

Table 2

Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

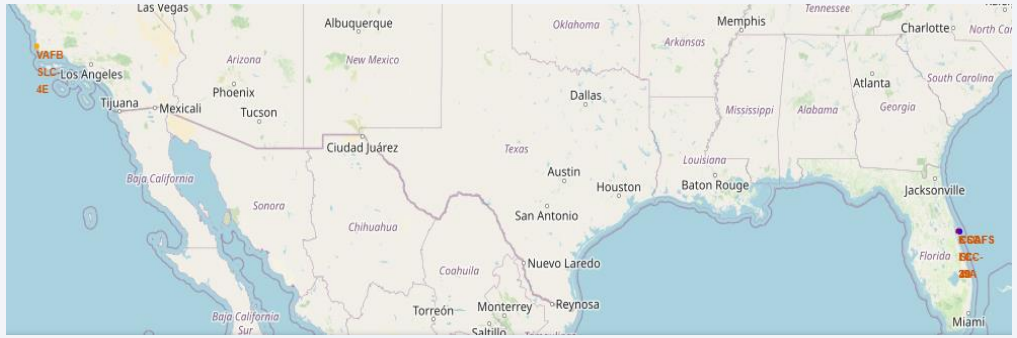
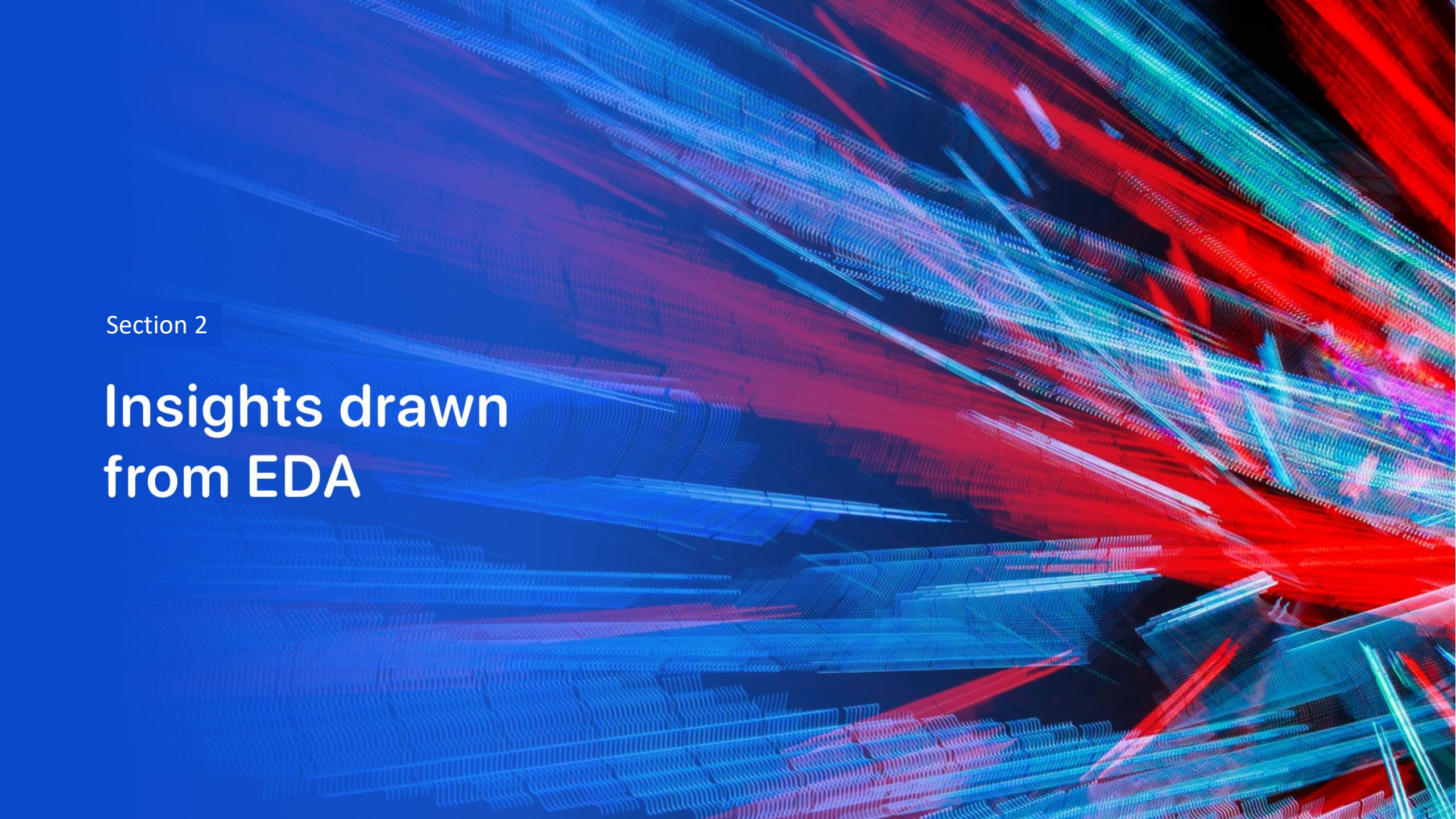
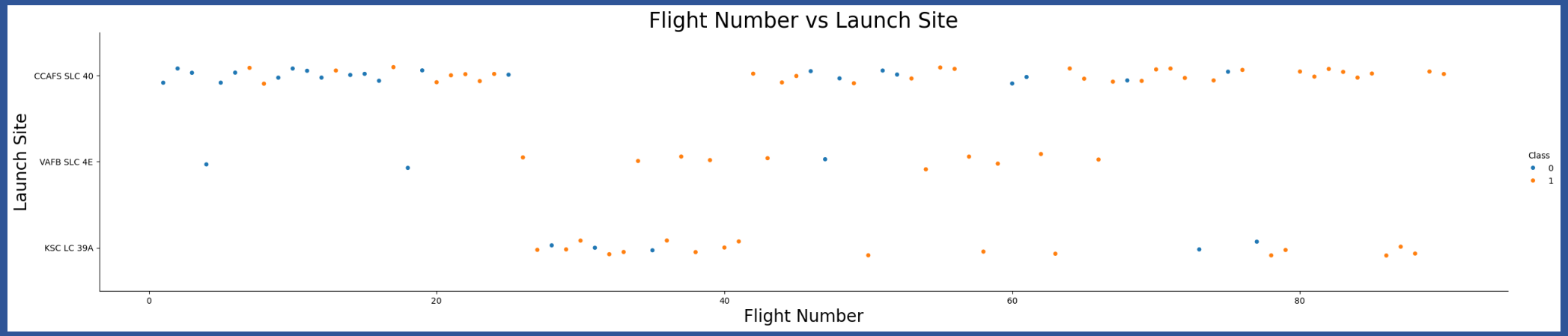


Image 1

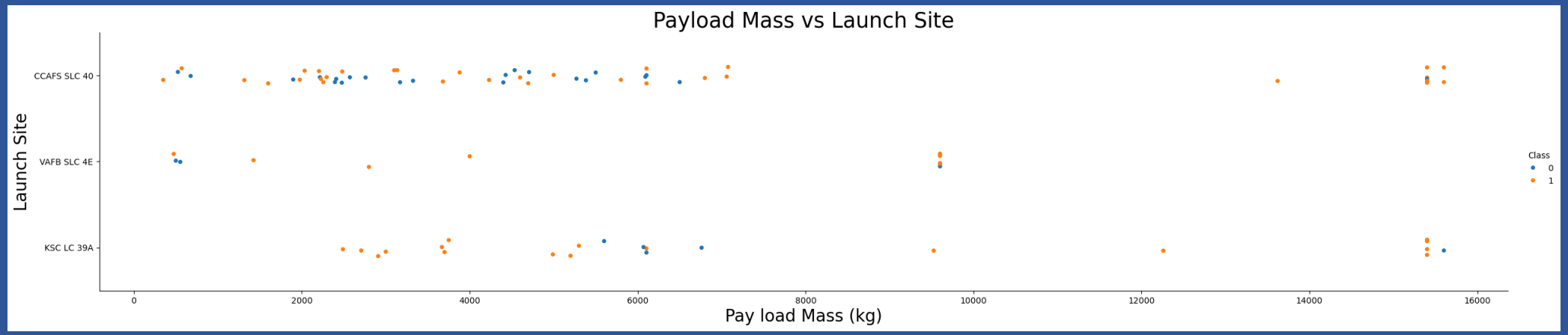

Section 2
Insights drawn from EDA



Flight Number vs. Launch Site

- For the CCAFS LC-40 launch site, we can see that flight numbers span across the whole range except between 25-40. It also has the most entries out of all the sites.
- For the VAFB SLC 4E launch site, we can see it has the least entries amongst the sites and has a maximum flight number of around 70.
- For the KSC LC-39A launch site, its lowest flight number is around 25 and reaches upwards towards 90 for its maximum flight number.

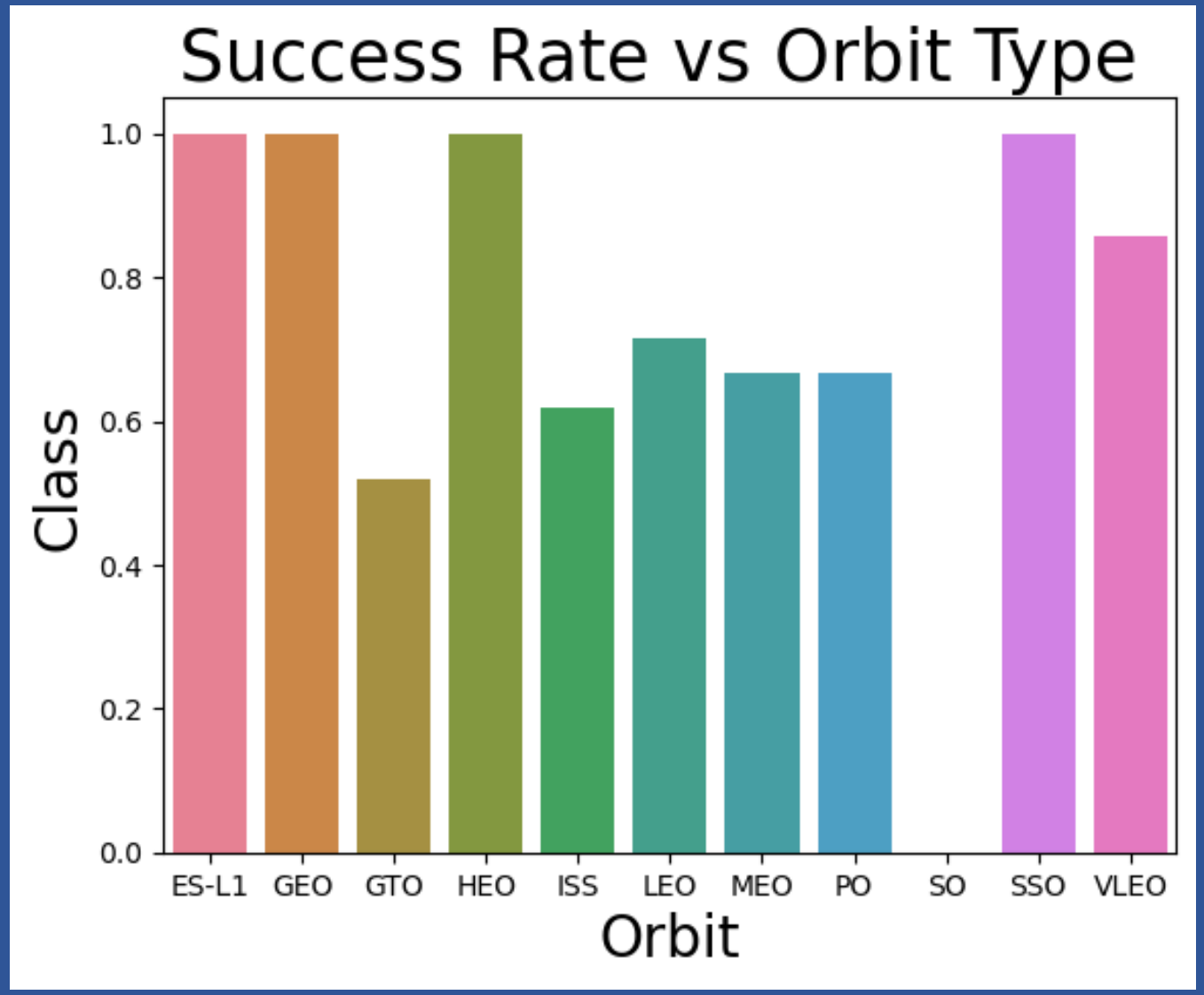
Payload vs. Launch Site

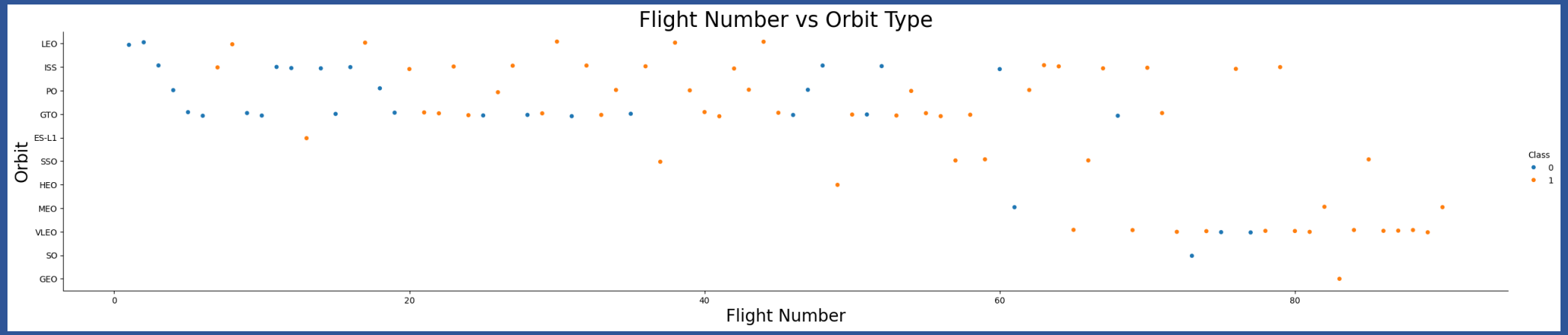


- For the CCAFS LC-40 launch site, it mostly dealt with payloads in the 0-8000kg range but did have some missions with payloads in the 12000-16000kg range.
- For the VAFB SLC 4E launch site, it had a few payloads that ranged from 0-4000kg but also had multiple missions with a payload of about 9500kg.
- For the KSC LC-39A launch site, its payloads were mostly in the 2000-8000kg range but had a few missions that had payloads ranging from 9000-16000kg.

Success Rate vs. Orbit Type

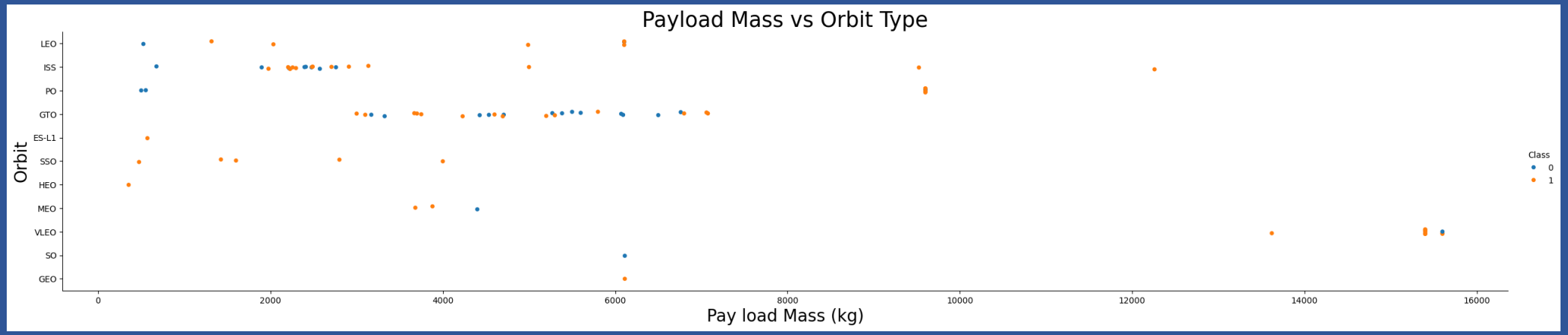
- The SO orbit type has a success rate of 0
- All orbit types except SO have a mission success rate of 50% or higher
- ES-L1, GEO, HEO, and SSO all have a 100% success rate on their missions.





Flight Number vs. Orbit Type

- The SO, HEO, ES-L1, and GEO orbit types have the least number of missions, all having only one.
- Lower flight numbers consisted of mostly LEO, ISS, PO, and GTO orbit types.
- Higher flight numbers consisted mostly of LEO and VLEO orbit types with other orbit types having a few missions.

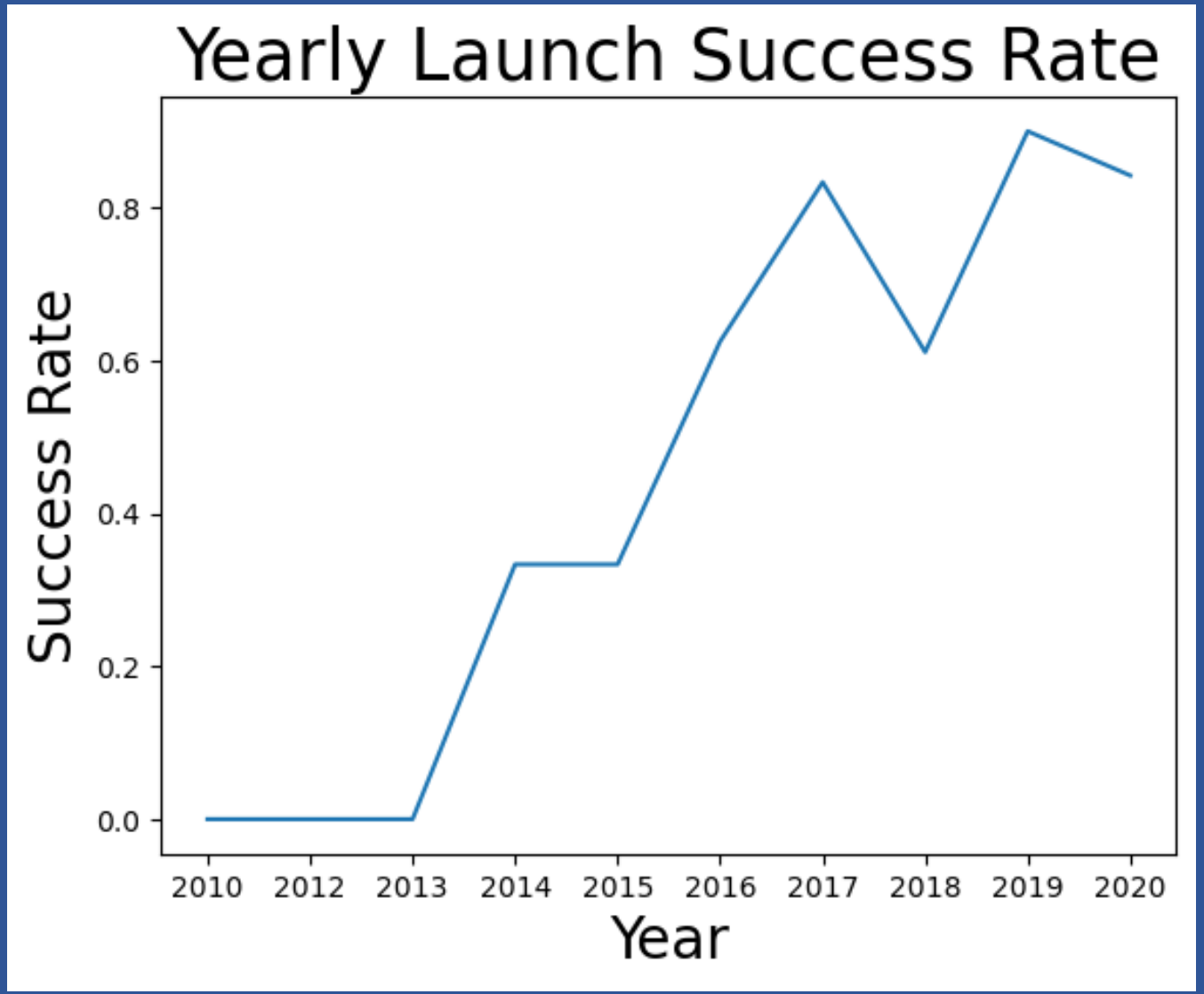


Payload vs. Orbit Type

- Most missions carried payloads between 0-8000kg with the remaining portion carrying payloads of 9000-16000kg.

Launch Success Yearly Trend

- The success rate generally increased every year
- The year with the highest success rate is 2019
- No successful missions were completed between 2010-2013



All Launch Site Names

Query

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

Result

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

- The distinct function is used to find all unique entries for a feature
- There are a total of 4 unique launch sites in the dataset

Launch Site Names Begin with 'CCA'

Query

```
%%sql
SELECT
    *
FROM
    SPACEXTABLE
WHERE
    Launch_Site like 'CCA%'
LIMIT
    5
```

Result

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The where clause is used to filter the launch site to start with CCA and limit is used to view only to first 5 entries
- The first five logged entries with launch sites starting with CCA all had successful outcomes and were all from the same launch site

Total Payload Mass

Query

```
%%sql
SELECT
    Customer,
    SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass
FROM
    SPACEXTABLE
WHERE
    Customer = 'NASA (CRS)'
```

Result

Customer	Total_Payload_Mass
NASA (CRS)	45596

- The SUM function is used to find the total value of a feature
- The total payload mass carried across all missions for the customer, NASA (CRS) was 45596 kg

Average Payload Mass by F9 v1.1

Query

```
%%sql
SELECT
    Booster_Version,
    AVG(PAYLOAD_MASS__KG_) as Average_Payload_Mass
FROM
    SPACEXTABLE
WHERE
    Booster_Version Like 'F9 v1.1'
```

Result

Booster_Version	Average_Payload_Mass
F9 v1.1	2928.4

- The AVG function is used to find the average value of a feature
- The average payload mass carried across all missions with the booster F9 v1.1 was 2928.4 kg

First Successful Ground Landing Date

Query

```
%%sql
SELECT
    Landing_Outcome,
    Min(Date)
FROM
    SPACEXTABLE
WHERE
    Landing_Outcome = 'Success (ground pad)'
```

Result

Landing_Outcome	Min(Date)
Success (ground pad)	2015-12-22

- The MIN function is used to find the smallest value in a feature
- The first successful landing outcome on ground pad was on December 22nd, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

Query

```
%%sql
SELECT
    Booster_Version,
    Landing_Outcome,
    PAYLOAD_MASS__KG_
FROM
    SPACEXTABLE
WHERE
    Landing_Outcome like 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ Between 4000 And 6000
```

Result

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

- The WHERE clause is used to filter the landing outcome and also set a range for the payload mass with the between function
- There are a total of four boosters which have landed on drone ship with a payload between 4000kg and 6000kg

Total Number of Successful and Failure Mission Outcomes

Query

```
%%sql
SELECT
    Mission_Outcome,
    Count(Mission_Outcome) as Total_Mission_Outcomes
From
    SPACEXTABLE
Group By
    Mission_Outcome
```

Result

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

- The COUNT function is used to get the number of entries for each feature and the group by clause is used to group values in a feature that are the same
- There is only 1 failure and 100 successful mission outcomes in total

Boosters Carried Maximum Payload

Query

```
%%sql
SELECT
    Booster_Version,
    PAYLOAD_MASS__KG_
FROM
    SPACEXTABLE
WHERE
    PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

Result

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- A subquery is used to find the maximum payload and then used as the condition in the WHERE clause
- These are all boosters that have carried the maximum payload of 15600 kg

2015 Launch Records

Query

```
%%sql
SELECT
    substr(Date, 6, 2) As Month,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM
    SPACEXTABLE
WHERE
    substr(Date, 0, 5) = '2015'
    AND Landing_Outcome ='Failure (drone ship)'
```

Result

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

- These are the records of failed landings in drone ship in the year 2015

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

Query

```
%%sql
SELECT
    Landing_Outcome,
    Count(Landing_Outcome) as Count
FROM
    SPACEXTABLE
WHERE
    Date Between '2010-06-04' AND '2017-03-20'
GROUP BY
    Landing_Outcome
Order By
    Count(Landing_Outcome) Desc
```

Result

Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

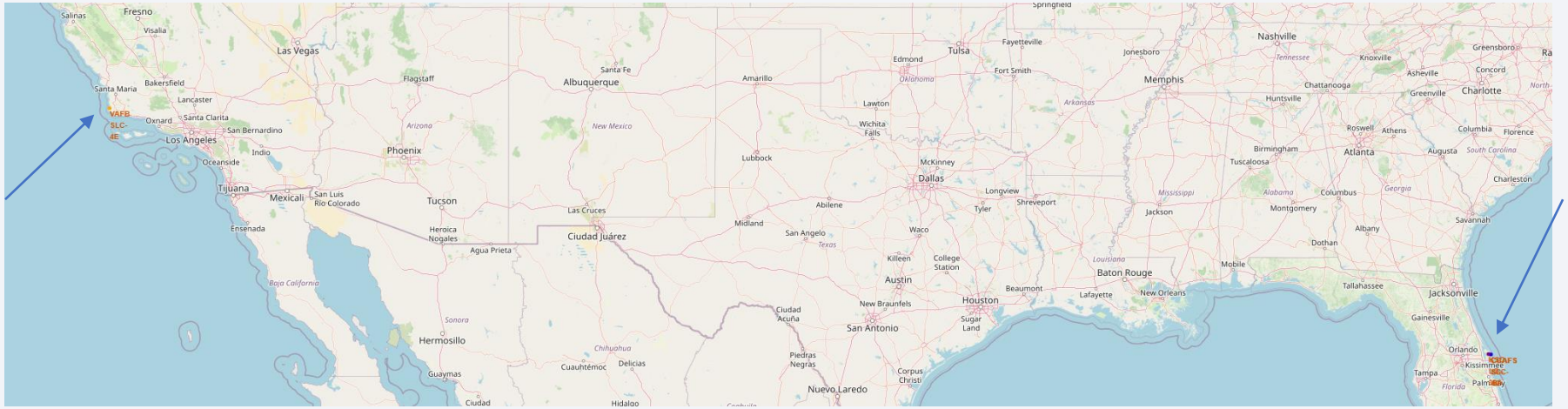
- This table shows the total number of records for each landing outcome between 2010-06-04 to 2017-03-20 in descending order with No attempt having the most records

Section 3

Launch Sites Proximities Analysis

SpaceX Launch Site Locations

- This map has icons indicating the locations of launch sites with labels for the name of the site
- There are a total of 4 Launch sites
- All Sites on the coasts in The United States near the equator

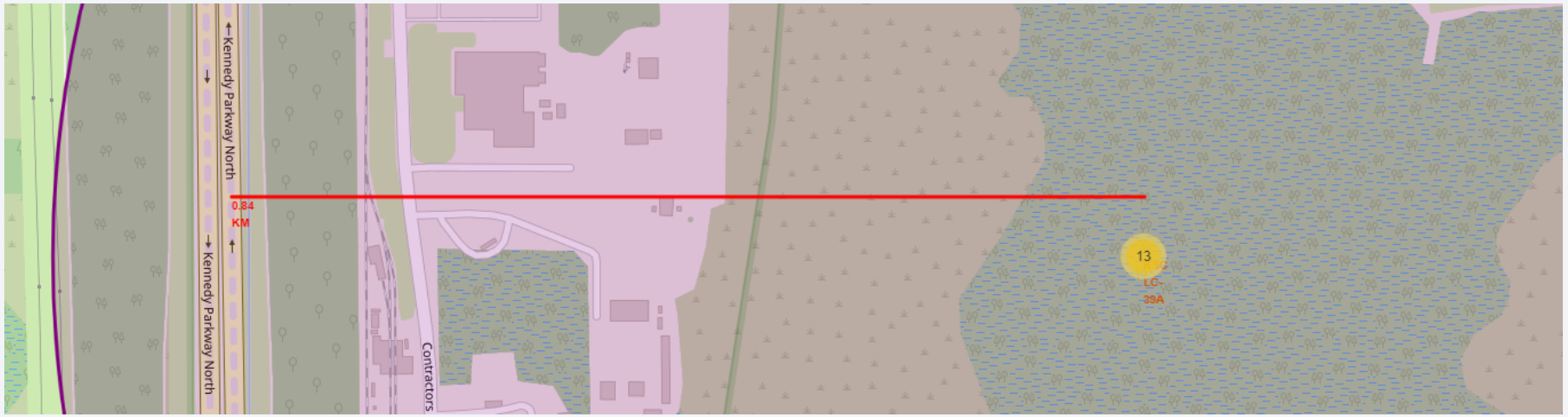


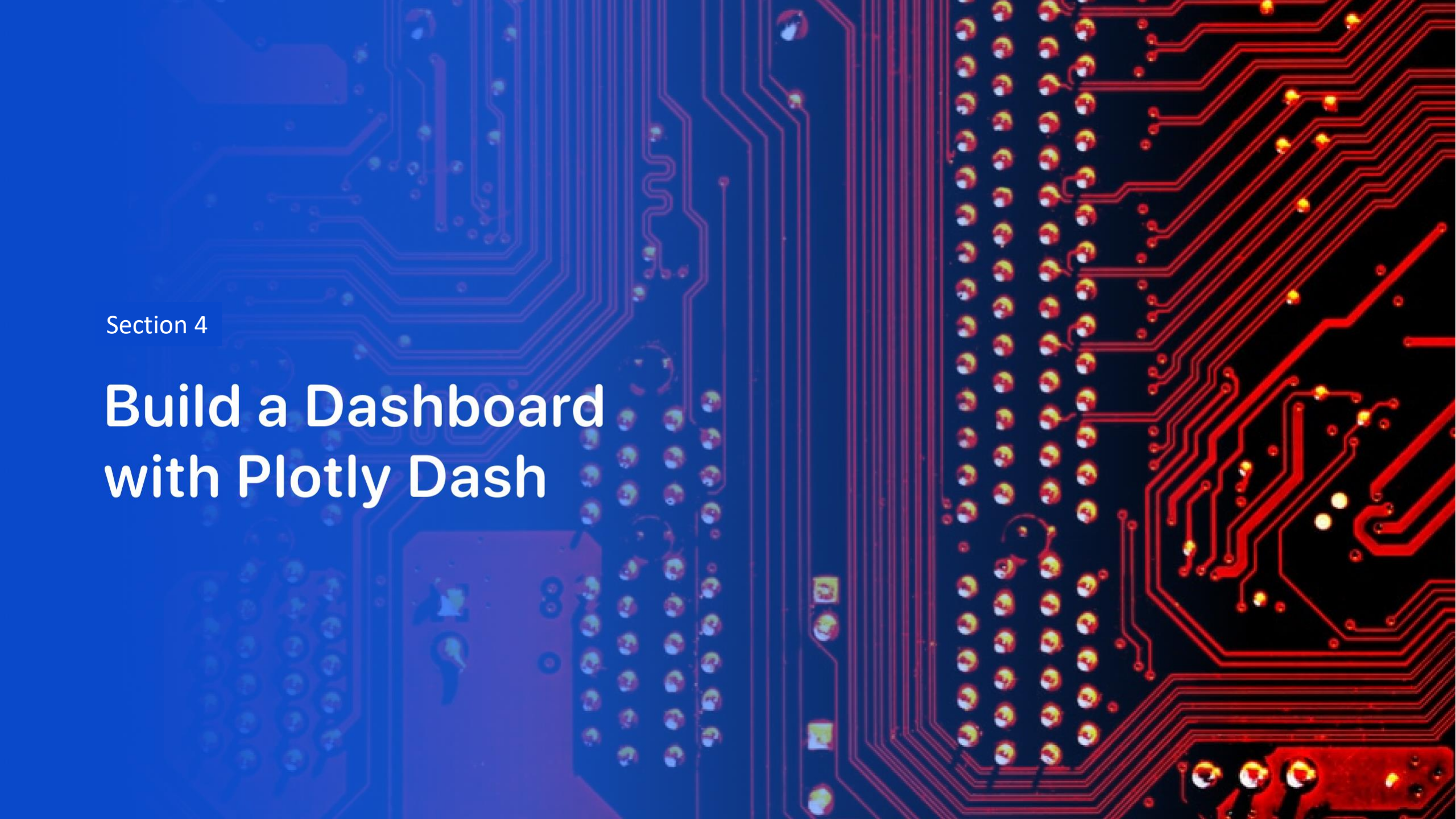
Launch Site Mission Outcome Clusters

- When unexpanded, the total number of missions for a launch site is shown
- When expanded the outcomes of missions for each launch site are shown in colour labeled. Red indicating a mission had failed and green indicating a successful mission.

Launch Site to Highway Distance Calculation

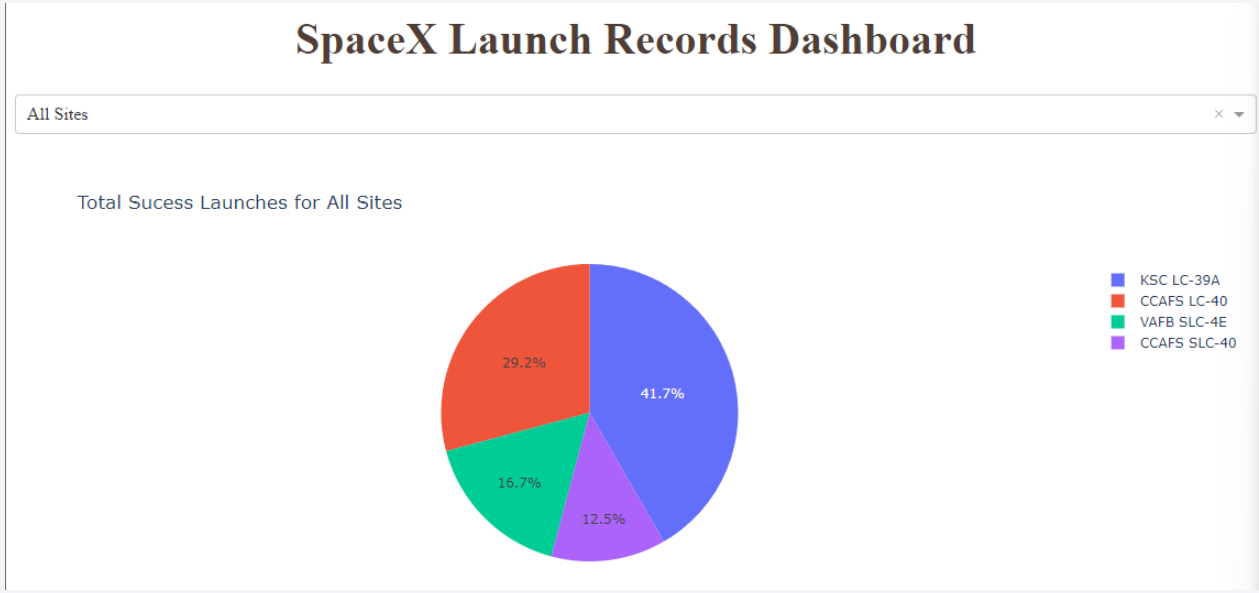
- The red line drawn in the screenshot is created using the PolyLine function
- The distance value label is calculated by obtaining the coordinate values which can be found using the MousePosition function which indicates the longitude and latitude values of the point on the map.
- In this screenshot, the distance from the launch site and the highway is 0.84 km.




Section 4
Build a Dashboard with Plotly Dash

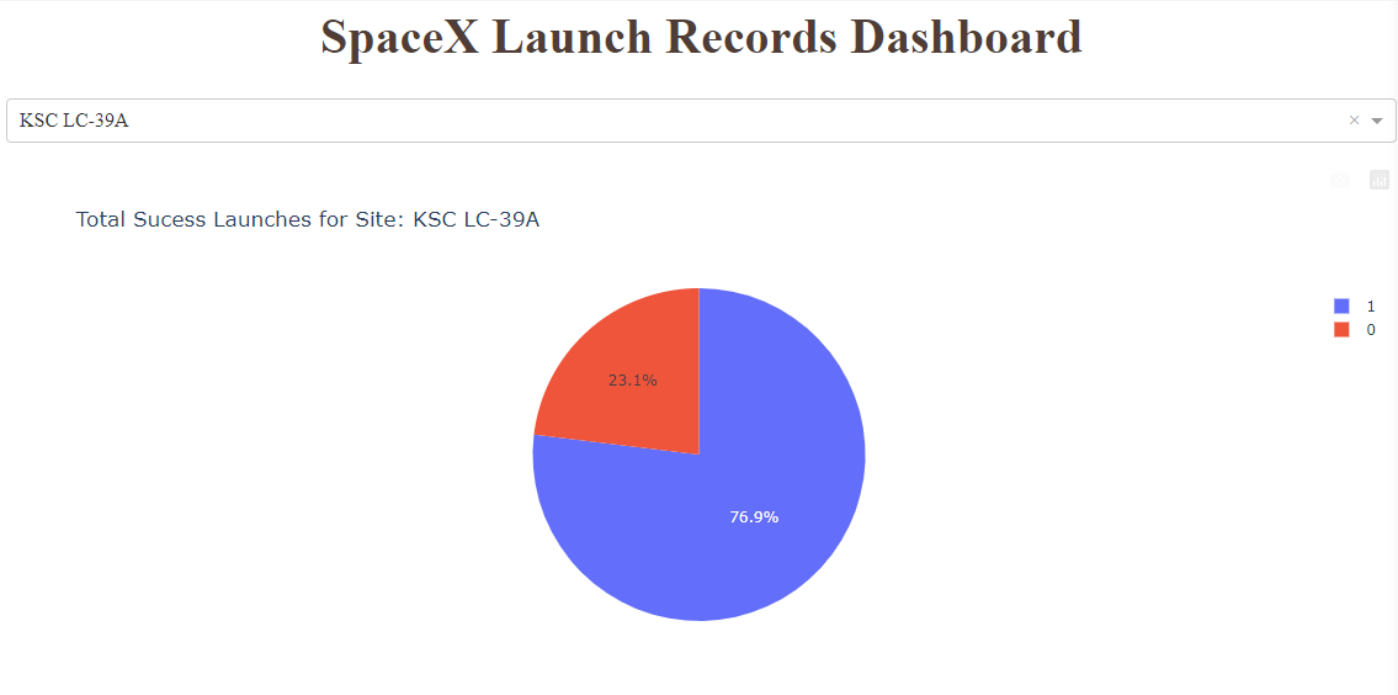
Pie Chart for SpaceX Launch Site Outcomes

- This pie chart by default will show the total successful launches for all sites
- The dropdown at the top allows the viewer to select a specific site
- The legend on the right is also dynamic and when the dropdown is on a specific site, the legend will indicate the mission outcome as the pie chart will show the total outcomes for the site



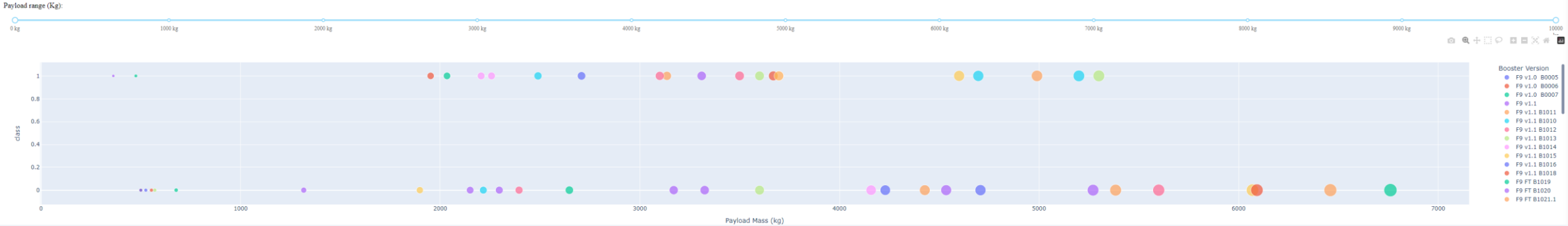
Launch Site with the Highest Success Rate

- The site with the highest success rate on missions is KSC LC-39A with a mission success rate of 76.9%



Payload Mass vs Outcome Scatter Plot

- This is a dynamic scatter plot where the payload mass range can be adjusted by the viewer using the slider at the top of the image to see a specified payload range.
- The payload range with the highest success rate is between 3000-4000kg.
- There have been no successful missions with payloads above the mass of 6000kg.

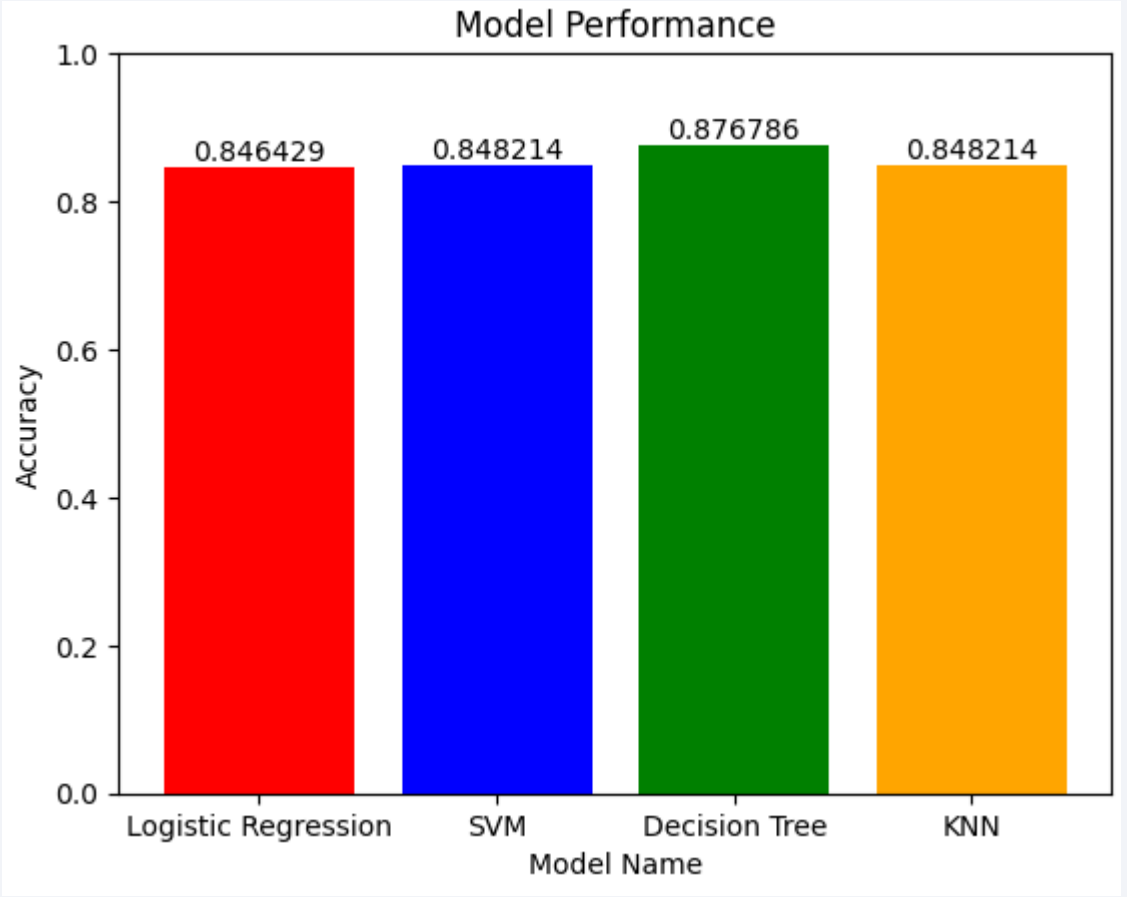


Section 5

Predictive Analysis (Classification)

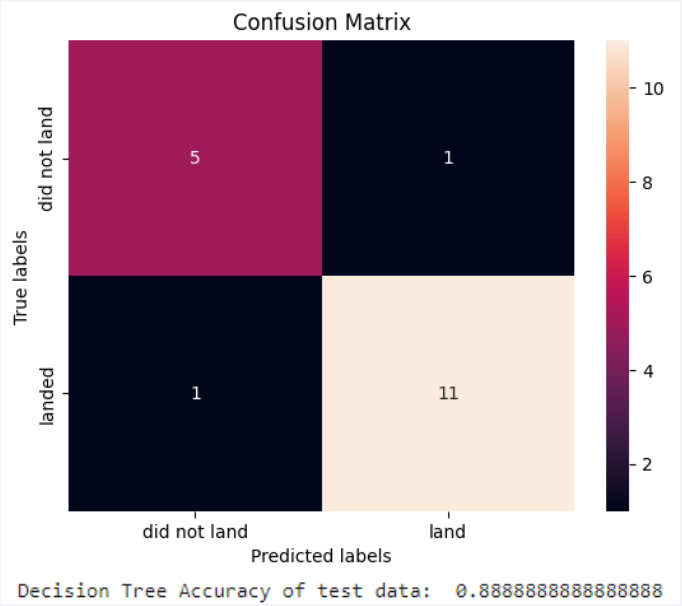
Classification Accuracy

- The Decision Tree classifier has the highest accuracy score amongst all models
- All other models performed the same as each other



Confusion Matrix

- The Decision Tree performed best on the test data with an accuracy of 0.88
- Out of 18 predictions, the model predicted correctly 16/18 labels correctly



Conclusions

- The general trend in mission success rate increased each year between 2013 to 2020, with 2019 having the highest success rate
- The orbits: ES-L1, GEO, HEO, and SSO all have a 100% success rate for their missions.
- All launch sites are located on the coasts of The United States near the equator
- The launch site with the greatest mission success rate is KSC LC-39A
- The most optimal machine learning algorithm to predict landing outcomes is the decision tree classifier

Appendix

GitHub Repo: https://github.com/EricYangg/IBM-Applied-Data-Science-Capstone/tree/main

Thank you!