IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hafizah Abdul Wahid
18 September 2024

Outline

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

Executive Summary

In this presentation I am investigating what are the key factors that influence the landing outcomes of SpaceX's Falcon 9 launches, and creating a model that will predict the outcomes of future launches.

The dataset is collected from the SpaceX API as well as webscraping. Data analysis is done using visualization, SQL, Folium, and Plotly Dash, the results of which then inform a predictive classification model.

Summary of Results

Key factors identified in the successful landing of a launch are Flight Number, Booster Version, Orbit Type, and Payload Mass. The predictive models identified as the most suitable are Logistic Regression, SVM and KNN.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each. Much of the savings is because SpaceX can reuse the first stage of their launches.

- Therefore if we can determine if the first stage will land, we 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.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX's API and Webscraping
- Perform data wrangling
  - Further classify data into successful and unsuccessful landings
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using Logistic Regression, SVM, Decision Tree and KNN

Data Collection

- Data was collected in two ways:
  - Using the SpaceX API
  - Webscraping

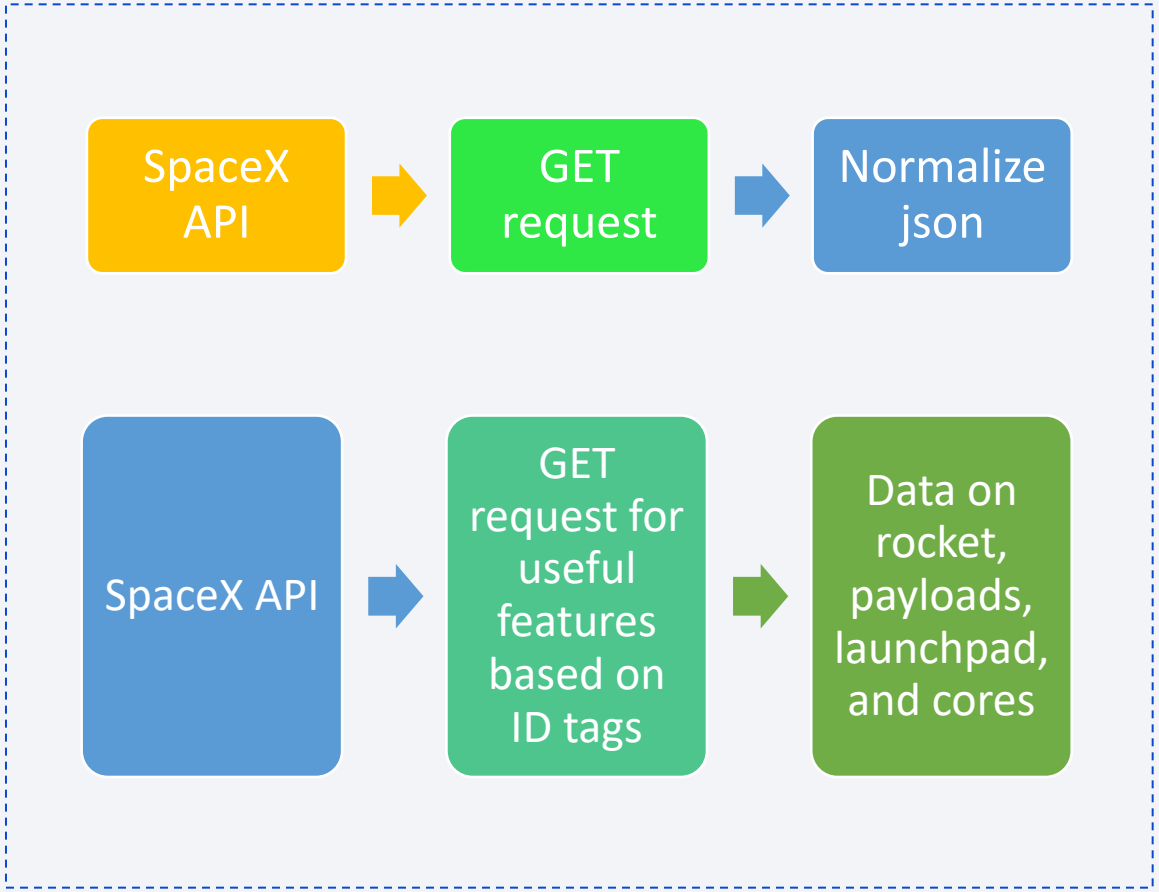
Data Collection – SpaceX API

Data collection via SpaceX API was done in two steps:

- Data harvesting using GET
- Parsing launch data for only useful features to create a new dataframe, which is later filtered for Falcon 9 launches only

Github URL:

https://github.com/hafizah-aw/IBM-SpaceX-Capstone-Project/blob/e713de2ba7bcb59417243ed92af6677baa6cdb87/jupyter-labs-spacex-data-collection-api.ipynb



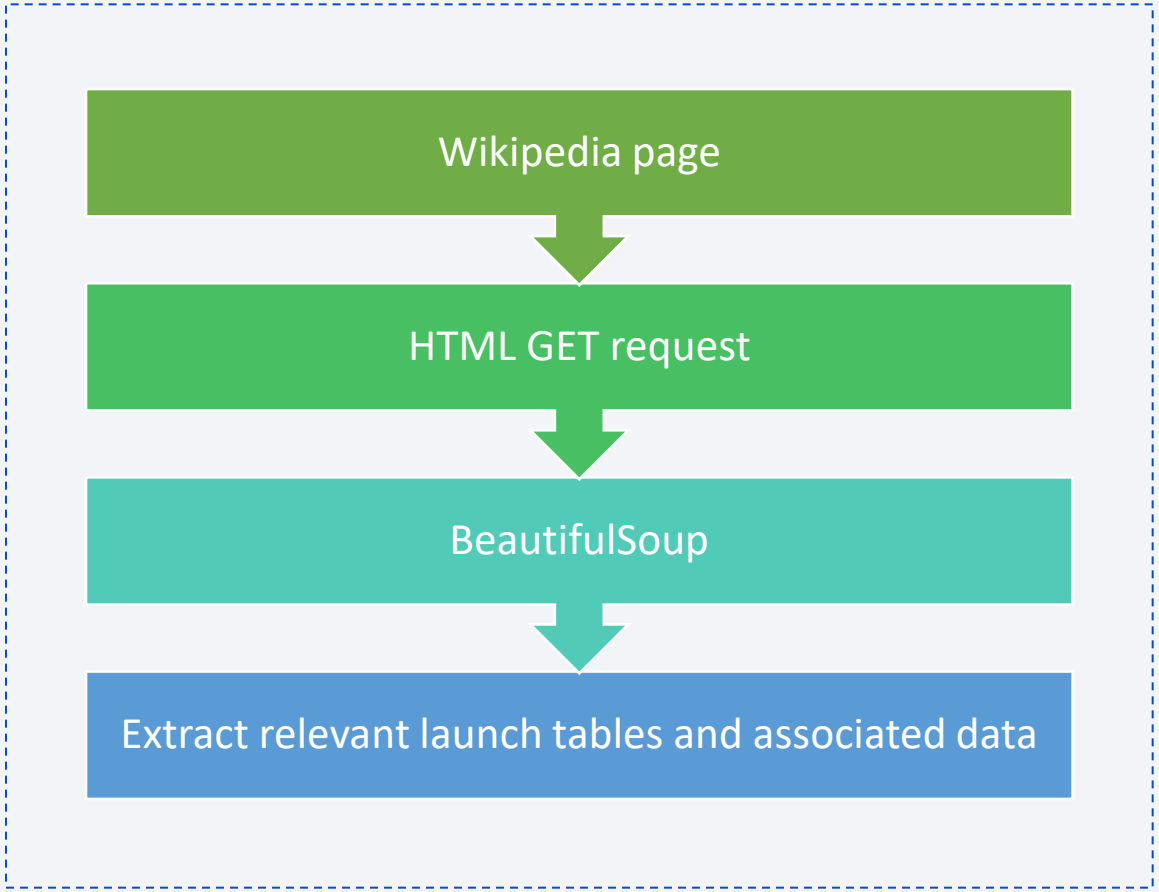
Data Collection - Scraping

Data collection via Webscraping was done in two steps:

- Using BeautifulSoup to extract Falcon 9 launch data from Wikipedia
- Parsing the data and converting it into a dataframe

GitHub URL:

https://github.com/hafizah-aw/IBM-SpaceX-Capstone-Project/blob/b2d9673d65b63dfbd67c7fa4c39ba62b2b21c168/jupyter-labs-webscraping.ipynb



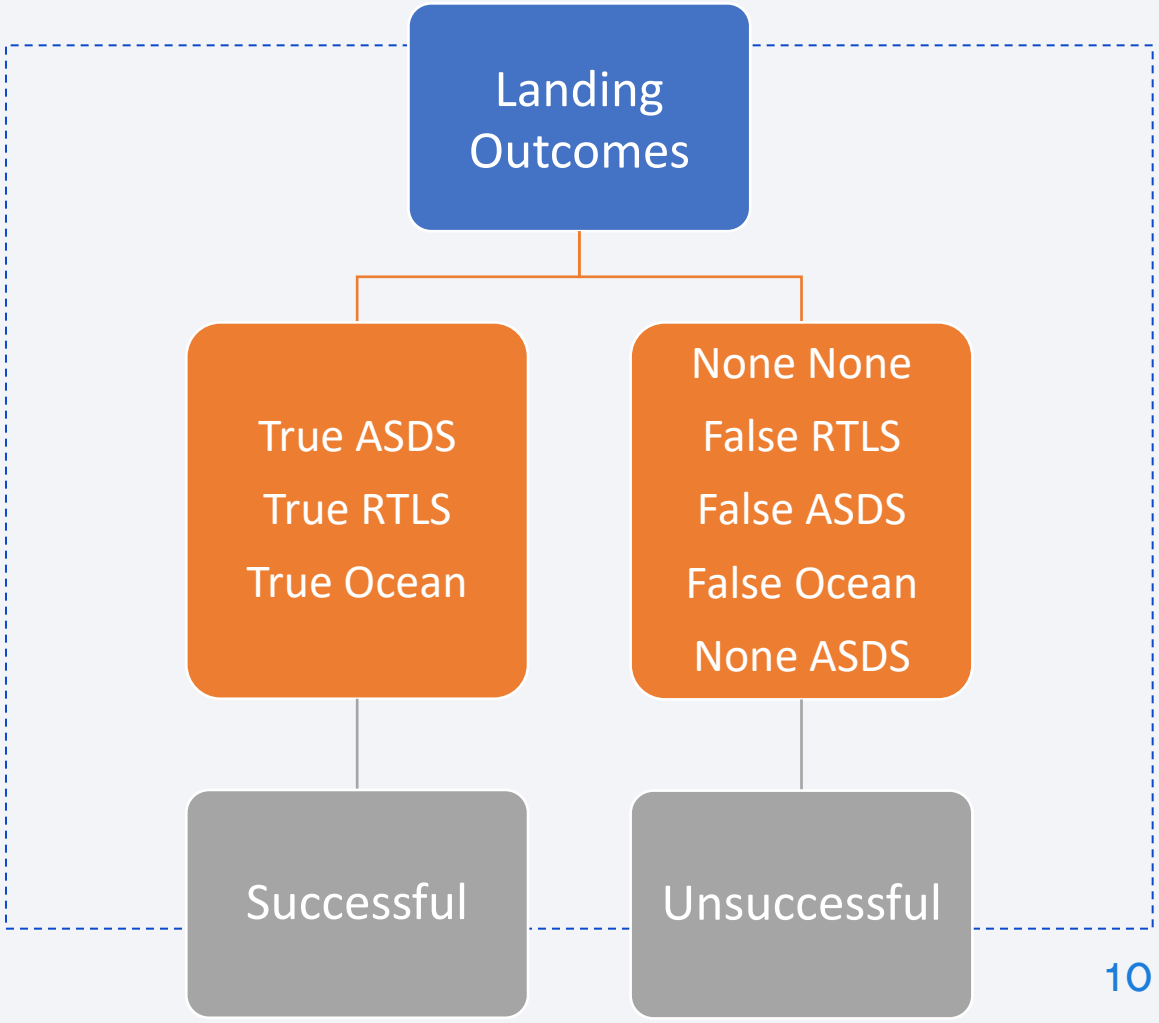
Data Wrangling

The launch data was further processed to identify the types of landing outcomes.

These were further classified as “successful” or “unsuccessful”.

GitHub URL:

https://github.com/hafizah-aw/IBM-SpaceX-Capstone-Project/blob/b2d9673d65b63dfbd67c7fa4c39ba62b2b21c168/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- In exploratory analysis of the data, the following charts were used to find out how different variables affected the success rate of a launch:
  - *Scatter point chart*, to visualize the relationships between selected variables
  - *Bar chart*, to visualize the success rate for a chosen variable
  - *Line chart*, to visualize the any trends in successful landings

- GitHub URL:
  - https://github.com/hafizah-aw/IBM-SpaceX-Capstone-Project/blob/b2d9673d65b63dfbd67c7fa4c39ba62b2b21c168/edadataviz.ipynb

EDA with SQL

- These are the SQL queries I've 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
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

EDA with SQL

- These are the SQL queries I've performed:
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass using a subquery
  - 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
  - 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
- GitHub URL:
  - https://github.com/hafizah-aw/IBM-SpaceX-Capstone-Project/blob/b2d9673d65b63dfbd67c7fa4c39ba62b2b21c168/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- To the interactive map, I added:
  - Circles to represent each of the SpaceX launch sites
  - Marker clusters to show the number of successful and unsuccessful landings at each of these sites
  - Polylines with distance between launch sites and nearby areas of interest such as coastlines, railways, airports, etc.
- On the map, these objects can immediately tell you how often a launch site is used, its landing success rate, and possible reasons for the site location with respect to natural and built features of the surrounding area.

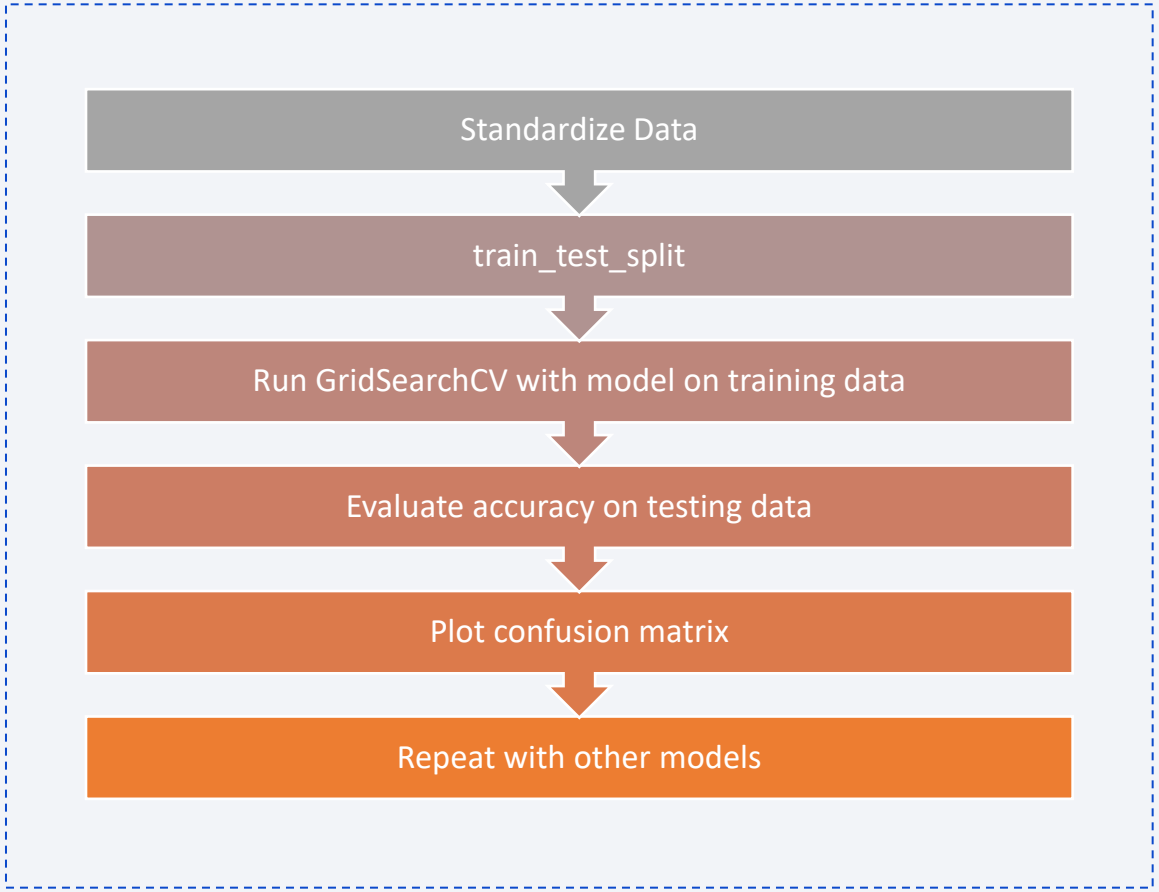
- GitHub URL
  - https://github.com/hafizah-aw/IBM-SpaceX-Capstone-Project/blob/b2d9673d65b63dfbd67c7fa4c39ba62b2b21c168/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- For my dashboard, I have added:
  - A Launch Site Drop-down Input Component
  - A callback function to render success-pie-chart based on selected site dropdown
  - A Range Slider to Select Payload
  - A callback function to render the success-payload-scatter-chart scatter plot
- These plots and interactions can tell you:
  - Which of the sites is most successful and the success rate of each site
  - The highest performing payload range and booster version
- GitHub URL
  - https://github.com/hafizah-aw/IBM-SpaceX-Capstone-Project/blob/b2d9673d65b63dfbd67c7fa4c39ba62b2b21c168/spacex_dash_app.py

Predictive Analysis (Classification)

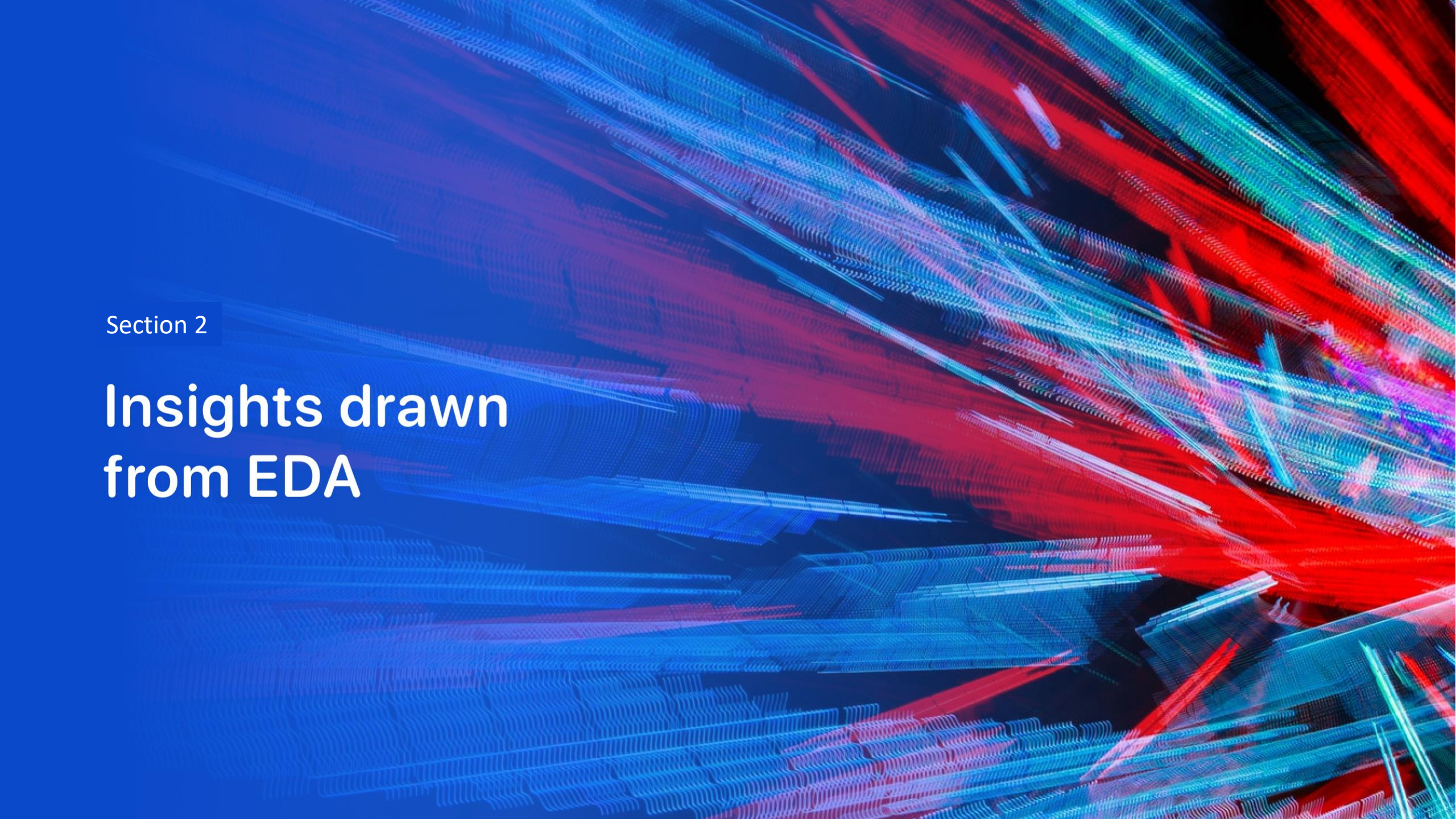
- To find the best performing classification model, the following steps were employed:
  - Standardize data
  - Split data into training and testing data
  - Run GridSearchCV with selected model on training data
  - Evaluate accuracy on test data
  - Plot a confusion matrix
  - Iterate the same process for other models



- GitHub URL
  - https://github.com/hafizah-aw/IBM-SpaceX-Capstone-Project/blob/b2d9673d65b63dfbd67c7fa4c39ba62b2b21c168/SpaceX_Machine%20Learning%20Prediction_Part_5%20(4).ipynb

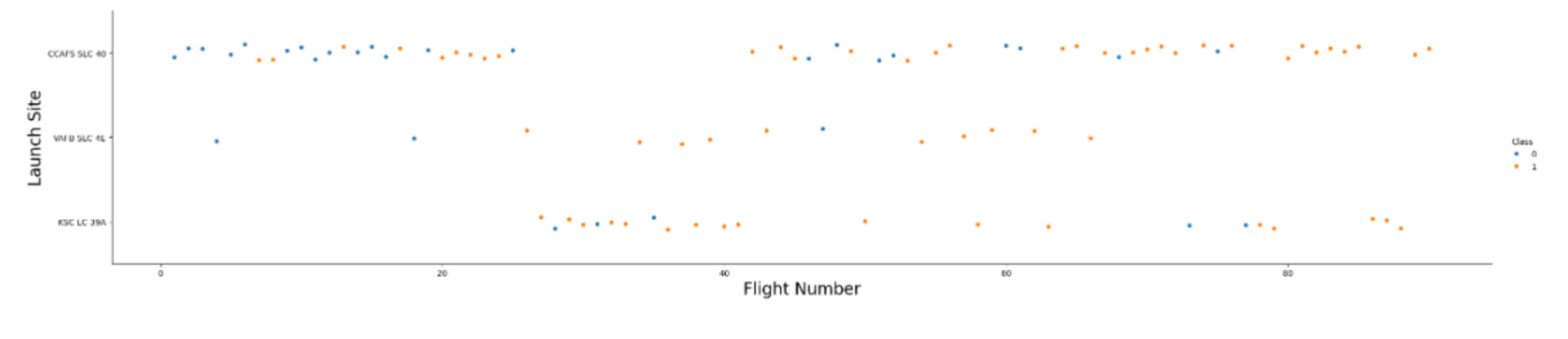
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

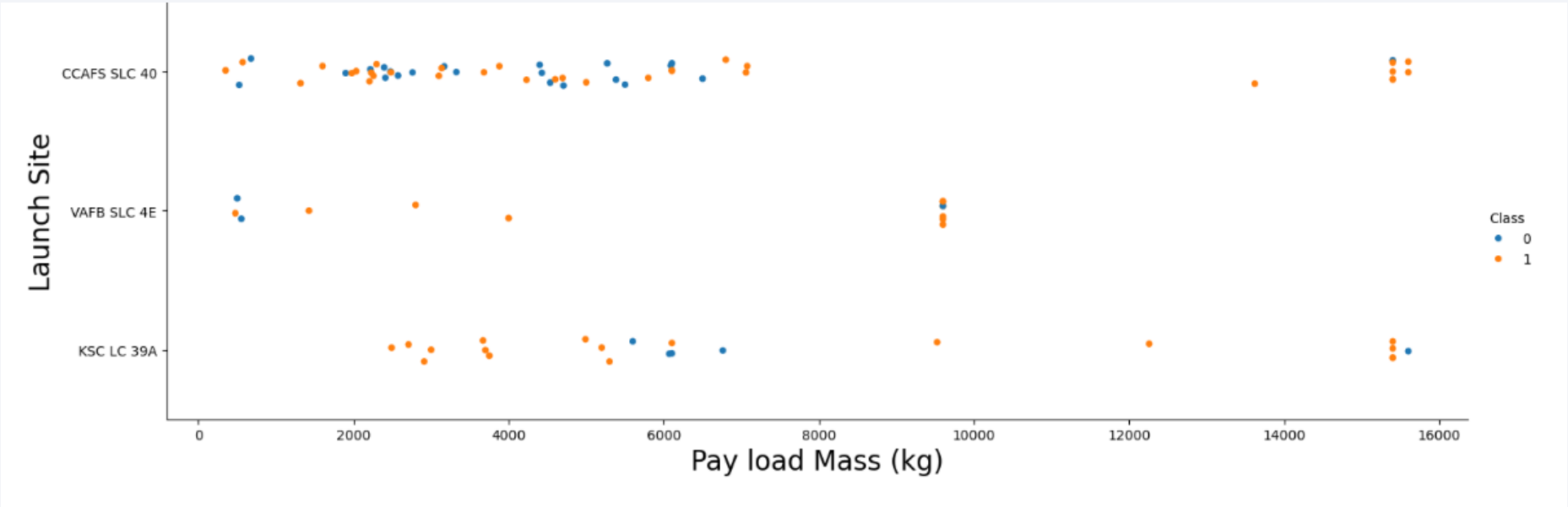
- Scatter plot of Flight Number vs. Launch Site



- Analysis – we can see that CCAFS SLC 40 seems to be a preferred launch site with the highest number of launches, although the launch sites themselves do not appear to be much of a factor in the success rate of launches.

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

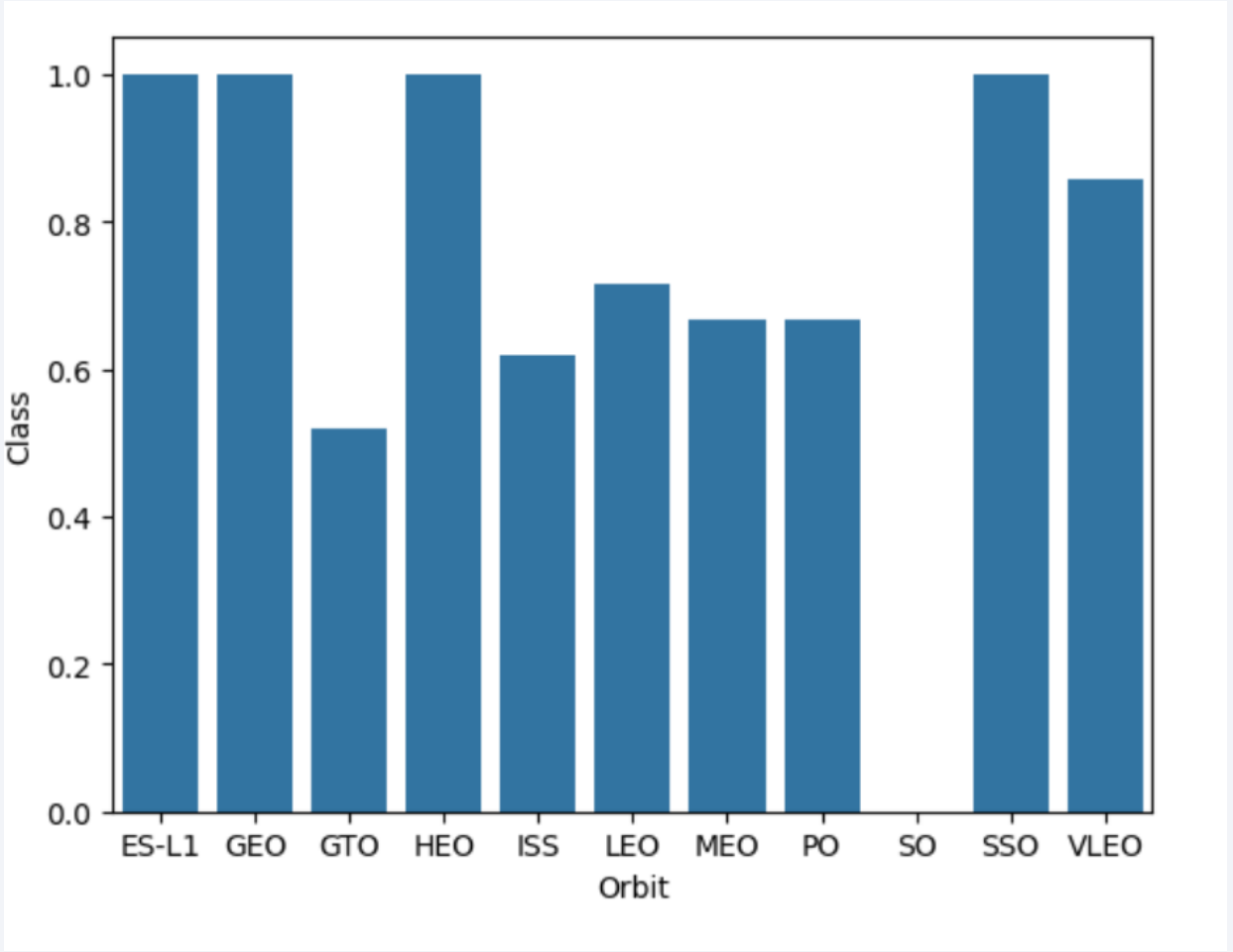


- Analysis – there has been no launches from VAFB SLC 4E with payloads above 10000 kg. Payload or Launch Site do not seem to influence success rate.

Success Rate vs. Orbit Type

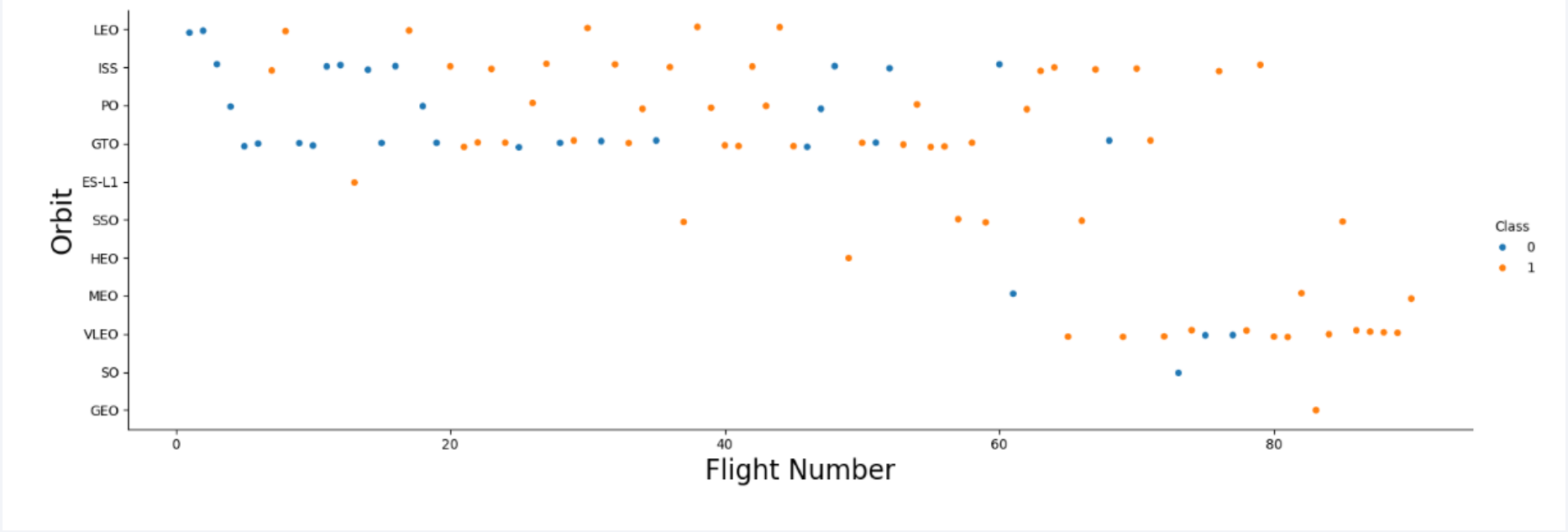
- Bar chart for the success rate of each orbit type

- Analysis – all launches for ES-L1, GEO, HEO and SSO orbits have been successful.



Flight Number vs. Orbit Type

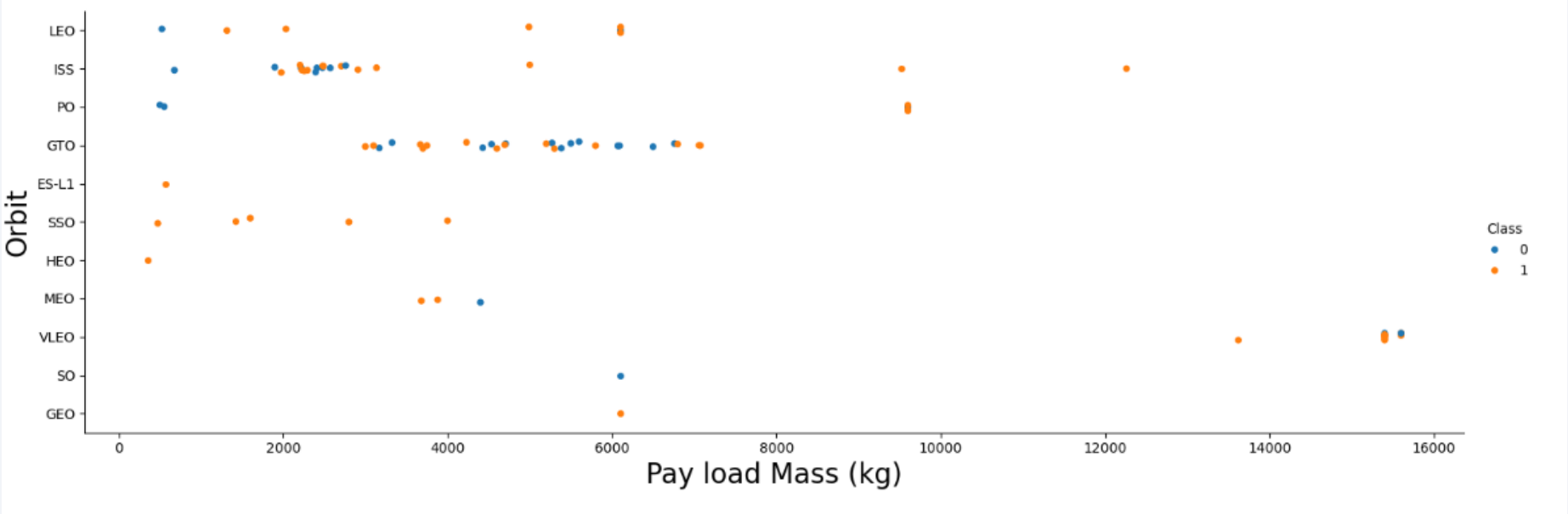
- Scatter point of Flight number vs. Orbit type



- Analysis – it is clear that GEO, HEO, ES-L1 and SSO orbits have extremely high success rates because of low launch numbers. Flight number still appears to be a greater factor in success rate than orbit type.

Payload vs. Orbit Type

- Scatter point of Payload vs. Orbit type

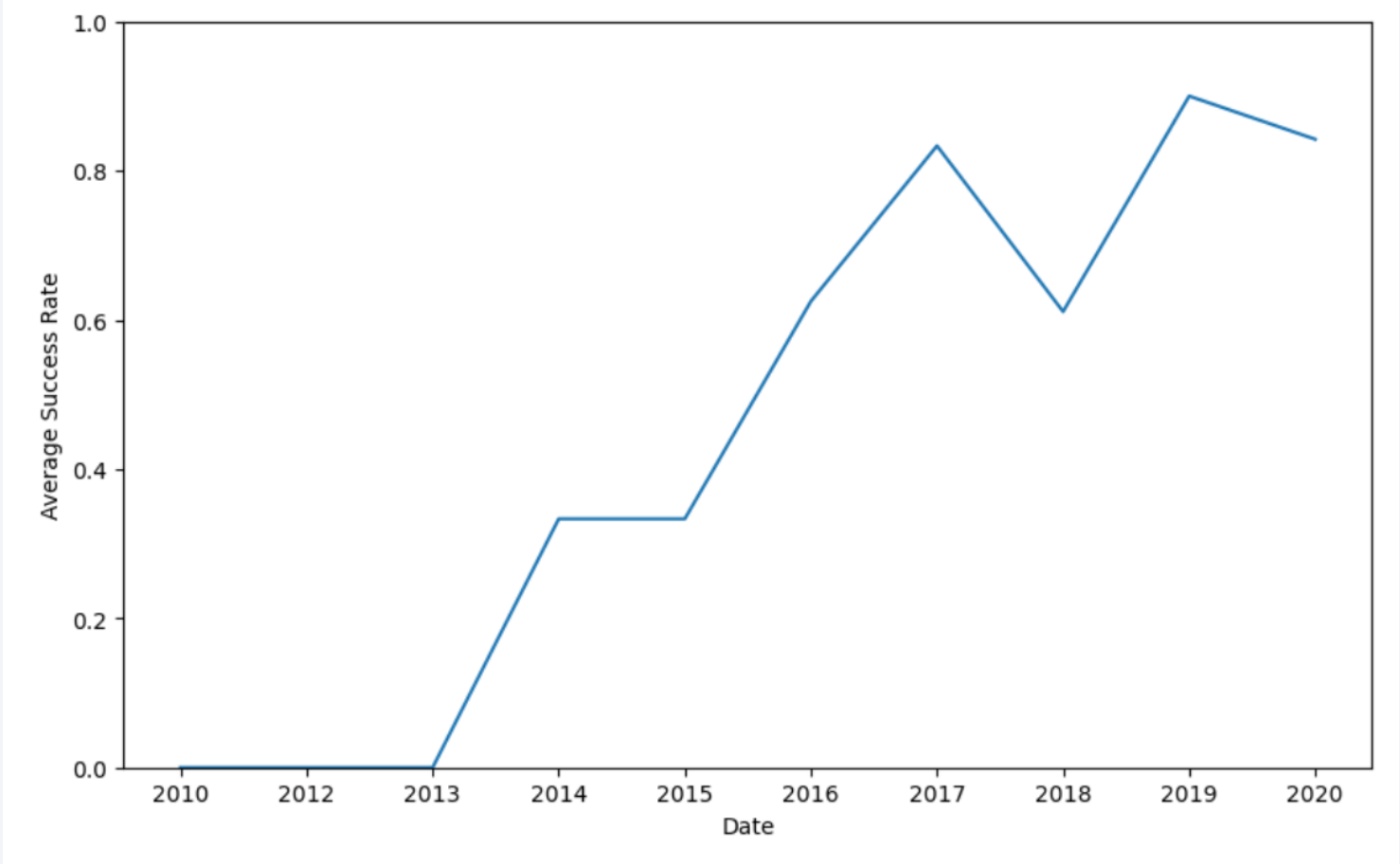


- Analysis - with heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISS. However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- Line chart of yearly average success rate

- Analysis – there is a general trend of increasing success over the years, likely due to better technologies and greater refinement of the Falcon 9 design.



All Launch Site Names

- Find the names of the unique launch sites
- Result - SpaceX only uses 4 launch sites.

Task 1

Display the names of the unique launch sites in the space mission

In [11]:
```
%%sql
SELECT DISTINCT "Launch_Site" FROM SPACEXTBL
```

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

Out[11]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Result – nothing to comment on.

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [13]:
```
%%sql
SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

Out[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- Total payload carried by boosters from NASA
- Result – nothing to comment on.

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [14]:
```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Customer" == 'NASA (CRS)'
```

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

Out[14]:

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1
- Result – the F9 v1.1 carried smaller payloads

Task 4

Display average payload mass carried by booster version F9 v1.1

In [15]:
```
%%sql
SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booster_Version" LIKE 'F9 v1.1%'
```

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

Out[15]:

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

- Result - the first successful landing outcome occurred on 22 Dec 2015

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [16]:
```
%%sql
SELECT DISTINCT "Landing_Outcome" FROM SPACEXTBL
```

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

Out[16]:

Landing_Outcome
Failure (parachute)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
Success (ground pad)
Success (drone ship)
Success
Failure
No attempt

In [17]:
```
%%sql
SELECT MIN("Date") FROM SPACEXTBL WHERE "Landing_Outcome" == 'Success (ground pad)'
```

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

Out[17]:

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Result – nothing to comment on

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [19]:
```
%%sql
SELECT "Booster_Version" FROM SPACEXTBL
    WHERE "Landing_Outcome" == 'Success (drone ship)'
    AND "PAYLOAD_MASS__KG_" BETWEEN 4001 AND 5999
```

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

Out[19]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Result – SpaceX has had a very high mission success rate. Landing outcomes are not considered in the success or failure of a mission.

Task 7

List the total number of successful and failure mission outcomes

In [26]:
```
%%sql
SELECT "Mission_Outcome", COUNT(*) FROM SPACEXTBL GROUP BY "Mission_Outcome"
```

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

Out[26]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Result – nothing to comment on.

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [28]:
```
%%sql
SELECT "Booster_Version" FROM SPACEXTBL
WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

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

Out[28]:

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Result – nothing to comment on.

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [30]:
```
%%sql
SELECT strftime('%m', "Date") AS "Month", "Landing_Outcome", "Booster_Version", "Launch_Site"
FROM SPACEXTBL
WHERE "Landing_Outcome" = 'Failure (drone ship)'
  AND strftime('%Y', "Date") = '2015';
```

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

Out[30]:

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

In [29]:
```
%%sql
SELECT substr("Date", 6,2), "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTBL
    WHERE "Landing_Outcome" == 'Failure (drone ship)'
    AND substr("Date",0,5)='2015'
```

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

Out[29]:

substr("Date", 6,2)	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

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

- 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

- Result – nothing to comment on.

Task 10

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.

In [31]:
```
%%sql
SELECT "Landing_Outcome", COUNT(*) AS "Outcome_Count"
FROM SPACEXTBL
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY "Outcome_Count" DESC;
```

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

Out[31]:

Landing_Outcome	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

Section 3

Launch Sites Proximities Analysis

Map of SpaceX Launch Sites

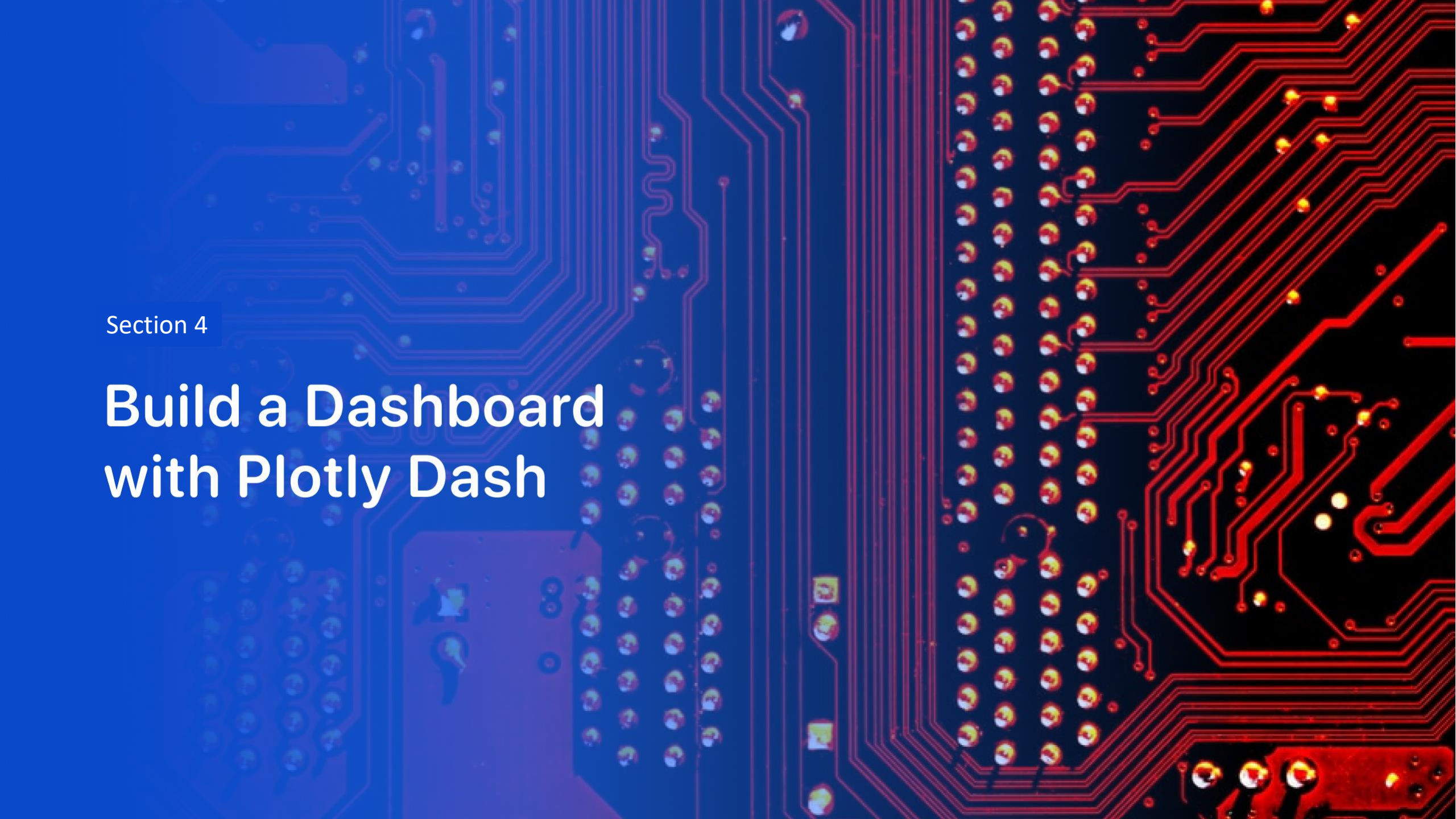
- SpaceX has chosen its 4 launch sites to be on the Western and Eastern coast, with 3 sites clustered close together on the Eastern coast.

Close up of KSC LC-39A Launch Site with Launch Outcomes

- Most missions have launched from launch sites on the Eastern coast, primarily at the CCAFS LC-40 and KSC LC-39A facilities.

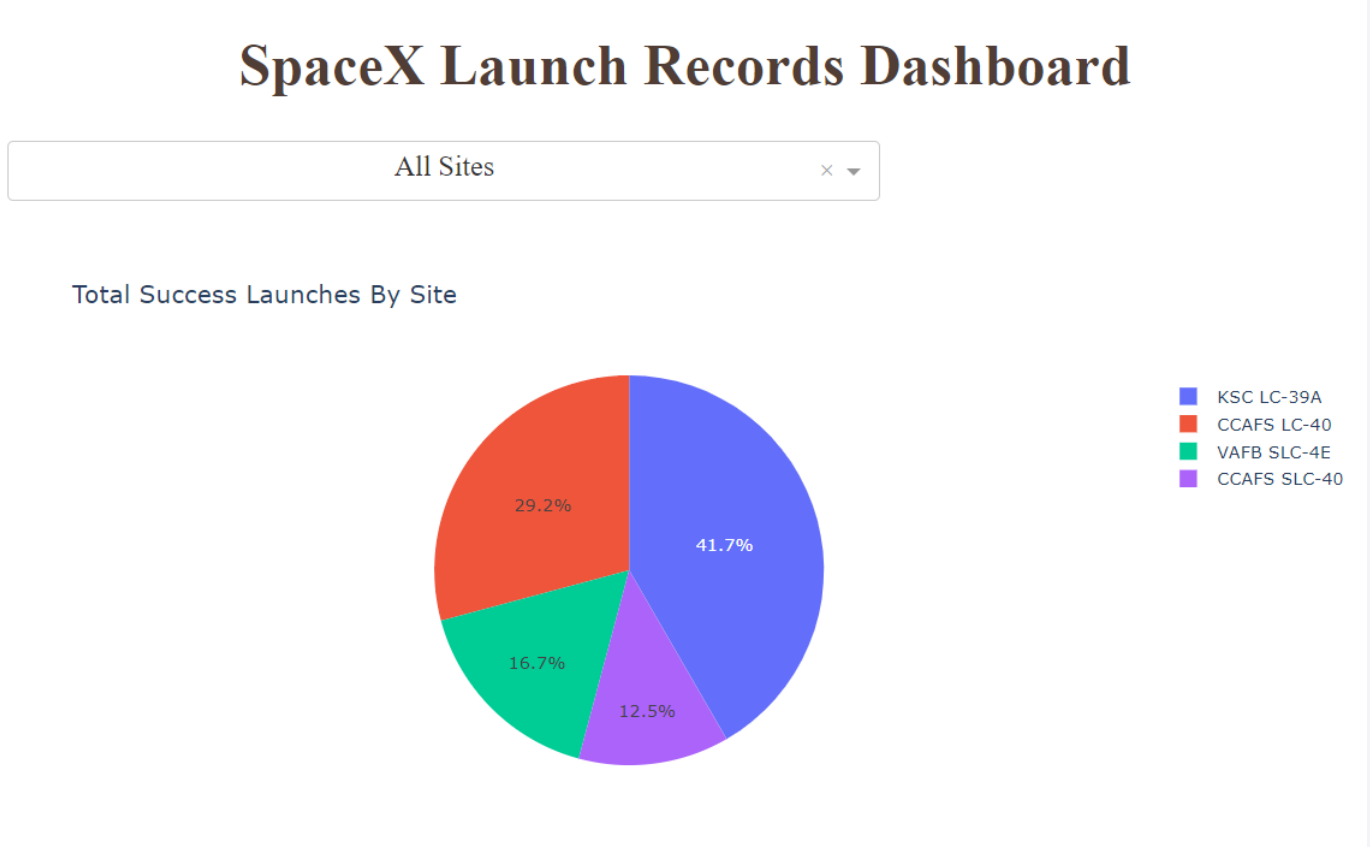
Close up of VAFB SLC-4E facility and surrounding features

- SpaceX's launch sites are all located close to the ocean (1.37km away, in this instance), for ocean landings and also for rockets to safely disintegrate in case missions go awry
- Launch sites are also observed to be close to railways, highways, and airports – presumably for the easy transportation of personnel and rocket components
- They are also situated far away from cities as a safety measure


Section 4
Build a Dashboard with Plotly Dash

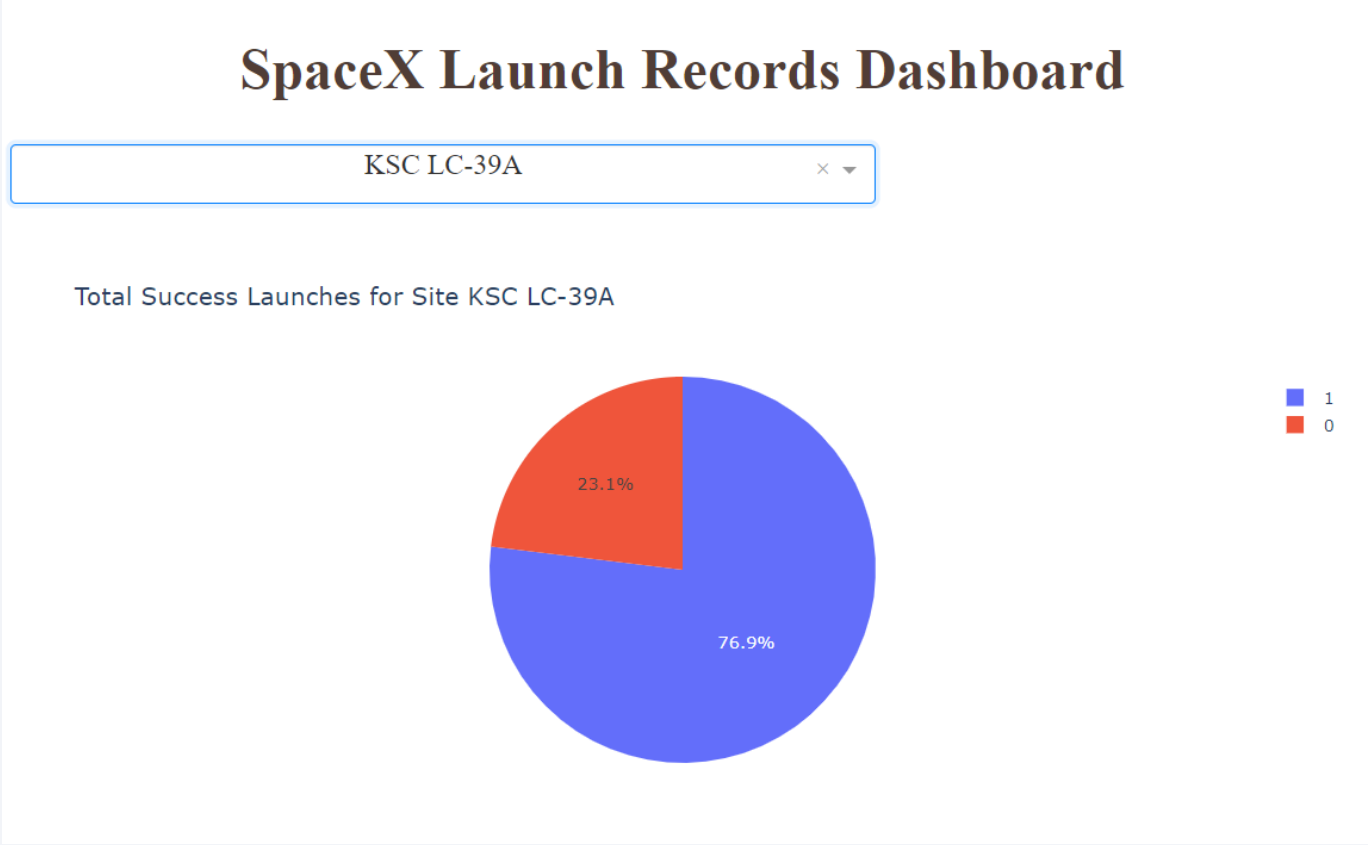
SpaceX Launch Records Dashboard Pt 1

- The pie chart shows the landing success rate of SpaceX's four launch sites
- The site with highest rate of success is KSC LC-39A (41.7%) followed by CCAFS LC-40 (29.2%)
- It is important to note that these sites also have markedly more launches than the others



SpaceX Launch Records Dashboard Pt 2

- The pie chart shows the landing success rate of the launch site with highest success rate, KSC LC-39A
- 76.9% of launches from this site have had successful landing outcomes



SpaceX Launch Records Dashboard Pt 3

- The scatter point chart shows the relationship between Payload Mass, Booster Version and Success rate.
- It is observed that most payloads fall within the selected range (~2000kg - ~5500kg)
- Payload does not seem to influence landing outcomes
- However, ‘FT’ appears to be the best performing booster version across all payloads, and ‘v1.1’, the worst

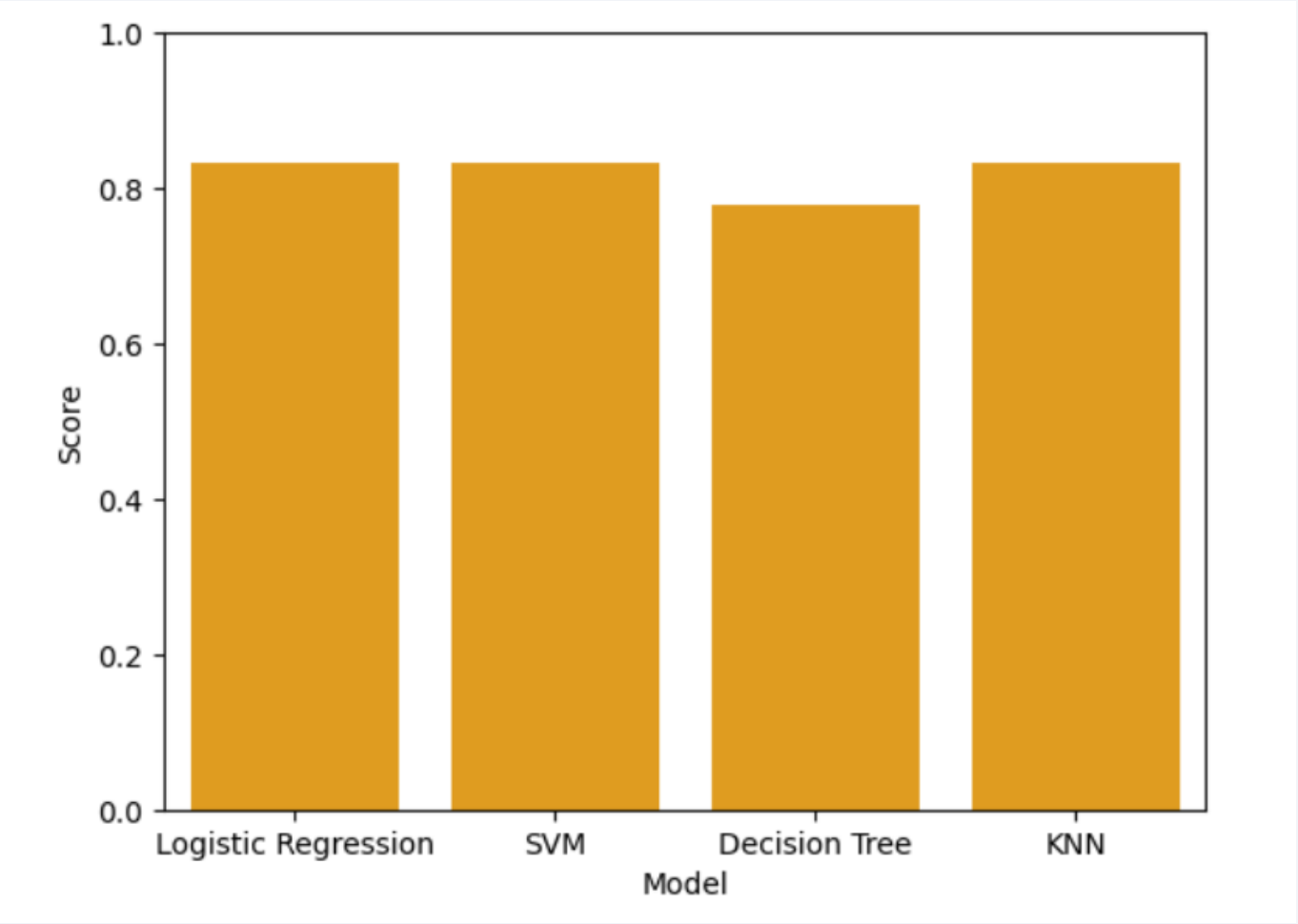
Section 5

Predictive Analysis (Classification)

Classification Accuracy

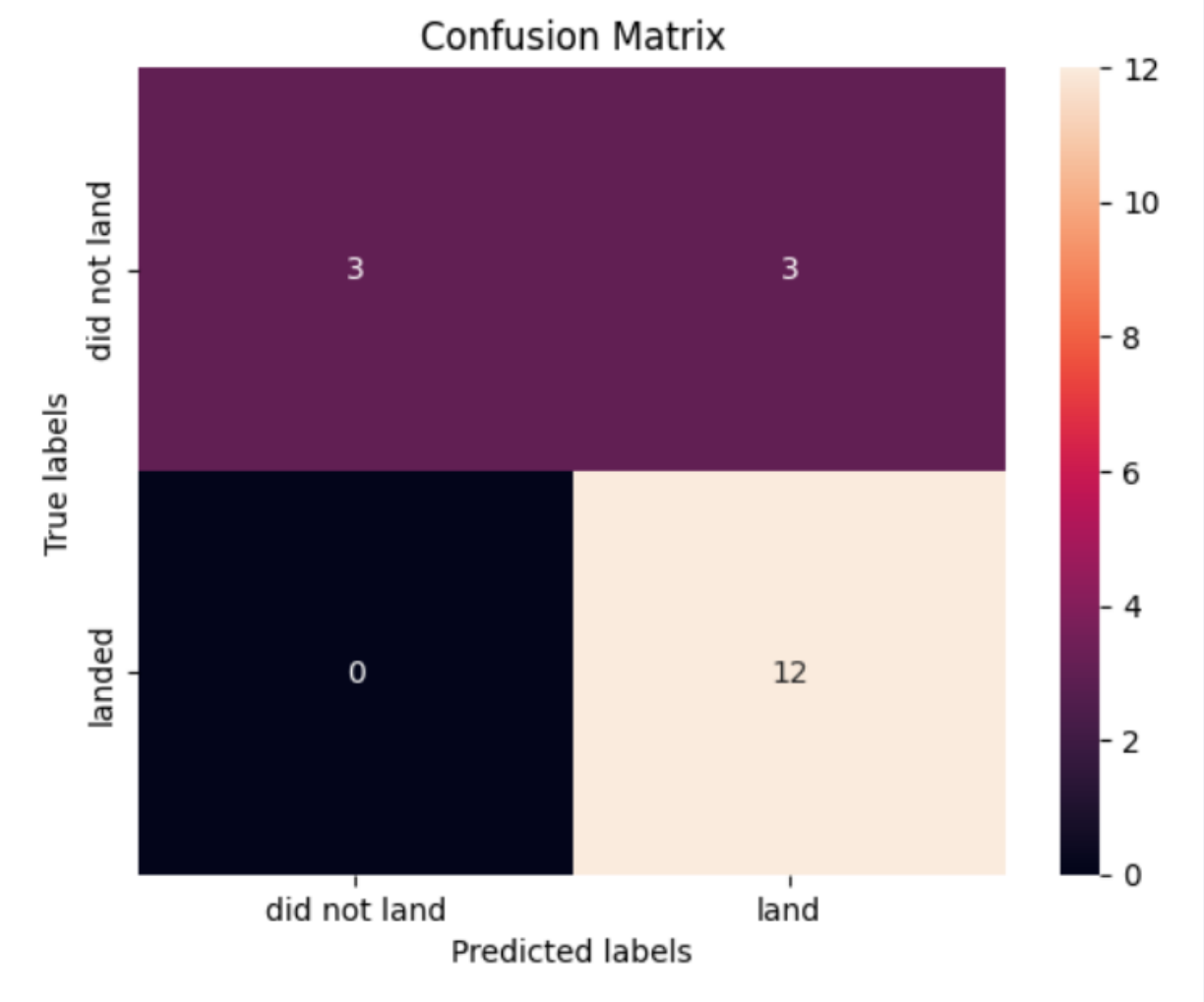
- Bar chart representing accuracy scores for all predictive models tested

- The best performing classification models were Logistic regression, SVM, and KNN with equal accuracy scores of 0.833



Confusion Matrix

- Confusion matrix for Logistic regression, SVM, and KNN models
- Out of the 6 unsuccessful landings, the models correctly predicted 3
- Out of the 12 successful landings, the models correctly predicted 12



Conclusions

- The one factor that influences landing outcomes the most is Flight Number. This makes sense as the more launches SpaceX makes, the better it gets at getting it right
- Orbit type, Payload mass, and Booster version also somewhat influence the success rate of landings – however, the relationships are less straightforward. For instance, only certain orbits have a linear relationship with payload range and success
- Launch sites have the least influence – the data is skewed to favor sites with more launches

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!