IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kristofers Ejugbo,
June 13, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Two data sources were utilized in this analysis: data was collected through the Spacex API and web scraping.
  - Data wrangling and Exploratory Data Analysis (EDA) were performed using Python and SQL.
  - Data visualization was carried out using multiple Python libraries, including Matplotlib, Folium, and Plotly Dash.
  - Machine learning techniques were applied for predictive analysis.
- Summary of all results
  - Data was successfully gathered using both the Spacex API and web scraping methods.
  - EDA results provided insights into the most appropriate features for machine learning.
  - Testing multiple machine learning algorithms resulted in identical accuracy results for all methods except Decision Tree, indicating the robustness of the chosen features and methods.

Introduction

- Project background and context
  - SpaceX lists Falcon 9 rocket launches on its website at a price of $62 million, while other providers charge upwards of $165 million per launch.
  - The significant cost savings are largely due to SpaceX's ability to reuse the first stage of their rockets.
  - Therefore, predicting whether the first stage will successfully land is key to estimating the cost of a launch.
  - This information is valuable for alternative companies looking to compete with SpaceX on rocket launch bids
- Problems you want to find answers
  - What are the features that determine a successful rocket launch?
  - How can a successful rocket launch be predicted?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web scraping from Wikipedia.com
- Perform data wrangling
  - Categorical values were converted using One-Hot Encoding method
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Training and Testing Data sets were processed through multiple ML algorithms.

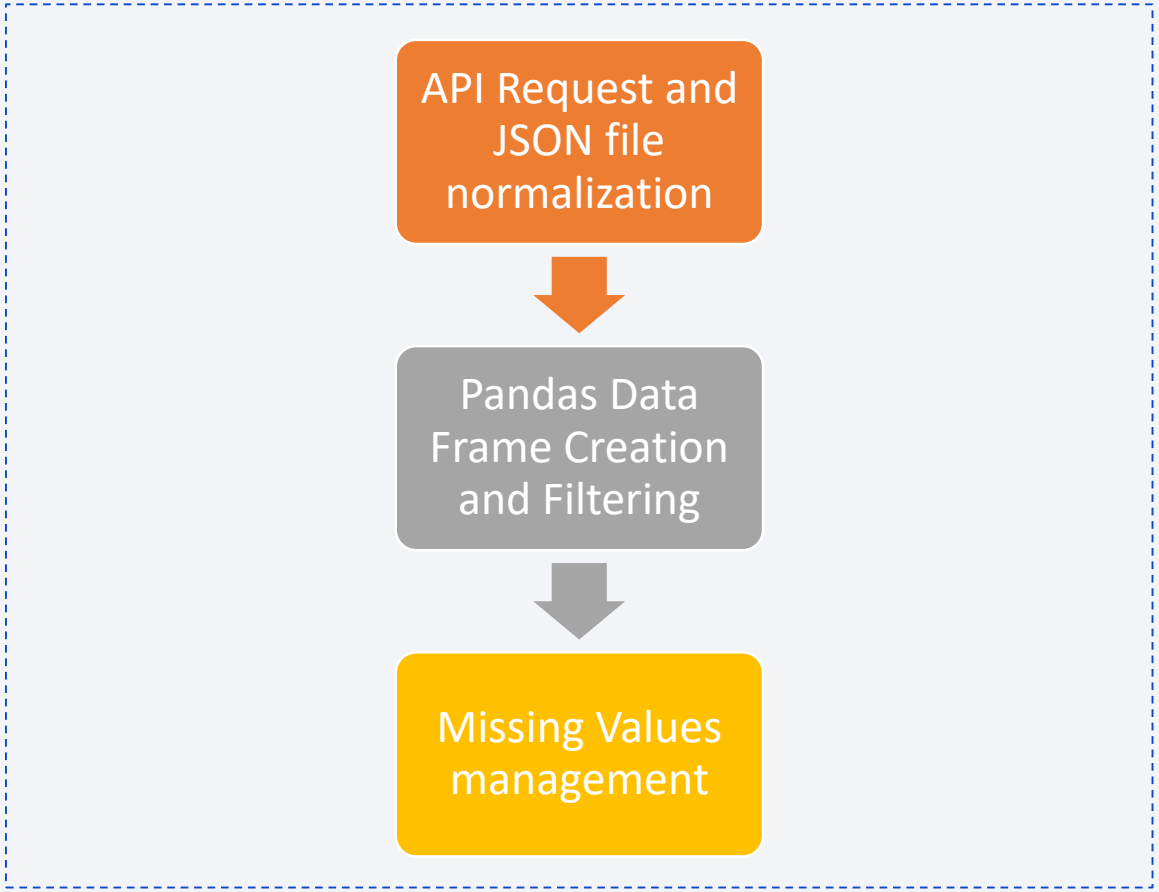
Data Collection

- SpaceX API:
  - Response Object creation for an API request and status confirmation
  - JSON file normalization and conversion into Pandas DataFrame
  - Data Enrichment with additional API calls
  - Data Cleaning
- Wikipedia Web Scarping
  - Web Scraping Data using BeautifulSoup
  - Data Extraction from HTML and conversion into Pandas DataFrame

Data Collection – SpaceX API

- Get Request was used to acquire the data set
- Dataset was converted into Pandas Data Frame
- Data Set filtered to include only Falcon 9 launches
- Missing Values replaced with mean values

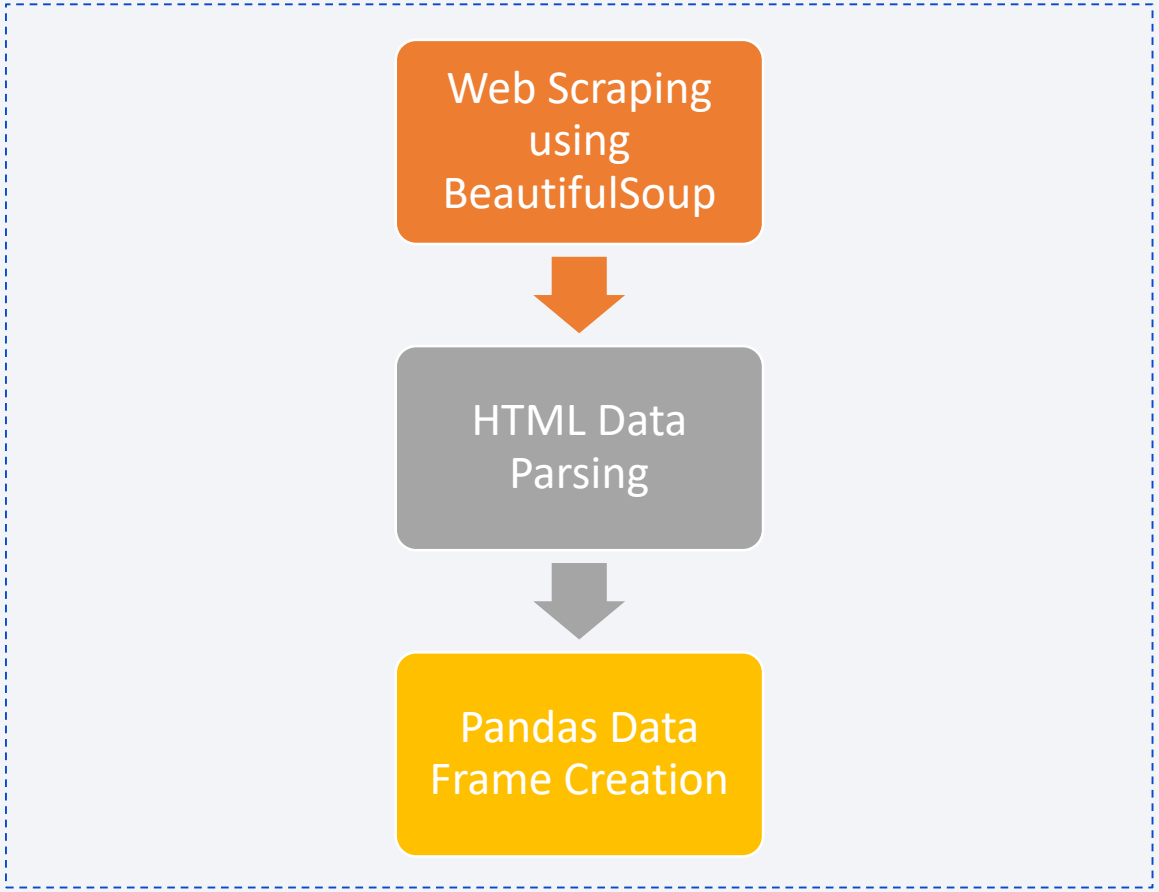
- https://github.com/chrisejugbo/Capstone_Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web Scraping Data using BeautifulSoup
- Data Extraction from HTML and conversion into Pandas DataFrame

- https://github.com/chrisejugbo/Capstone_Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data Wrangling process included Exploratory Data Analysis and One Hot coding for Landing Outcomes

- https://github.com/chrisejugbo/Capstone_Project/blob/main/3_spacex_Data_wrangling.ipynb

EDA with Data Visualization

- Graphs used in Data Visualization:
  - Scatter plot visualizing the relationship between Flight Number and Payload Mass,
  - Scatter plot visualizing the relationship between Flight Number and Launch Site,
  - Scatter plot visualizing the relationship between Payload and Launch Site,
  - Bar chart visualizing the relationship between success rate of each orbit type,
  - Scatter plot visualizing the relationship between Flight Number and Orbit type,
  - Scatter plot visualizing the relationship between Payload and Orbit type,
  - Bar chart visualizing the launch success yearly trend
- https://github.com/chrisejugbo/Capstone_Project/blob/main/5_jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries performed
  - Displaying the names of the unique launch sites,
  - Displaying 5 records where launch sites begin with the string 'CCA‘,
  - Displaying 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 when the first successful landing outcome in ground pad was achieved,
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000,
  - Listing the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass. Use a subquery
  - Listing the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015,
  - Ranking 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.
- https://github.com/chrisejugbo/Capstone_Project/blob/main/4_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium Library was used to visualize SpaceX, for example:
  - Markers were used:
    - To mark all Launch Sites,
    - To mark Launch Outcomes,
    - To mark distances in km between launch sites and its proximities.
  - Circles were used:
    - To Circle a radius of 1000 around the Launch Sites,
    - To Circle a radius of 1000 around the NASA Johnson Space Center.
  - Lines were used:
    - To visualized distances between Launch Sites and its proximities

- https://github.com/chrisejugbo/Capstone_Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Graphs Created:
  - Pie Chart that depicts Total Success Launches by Site
  - Scatter Chart that depicts Correlation between Payload and Success Rate. Color coded by Booster Version Category
- Interactions created:
  - Drop Down List to select the Launch Site
  - Slider to select Payload Range
- https://github.com/chrisejugbo/Capstone_Project/blob/main/spacex_dash_app.py

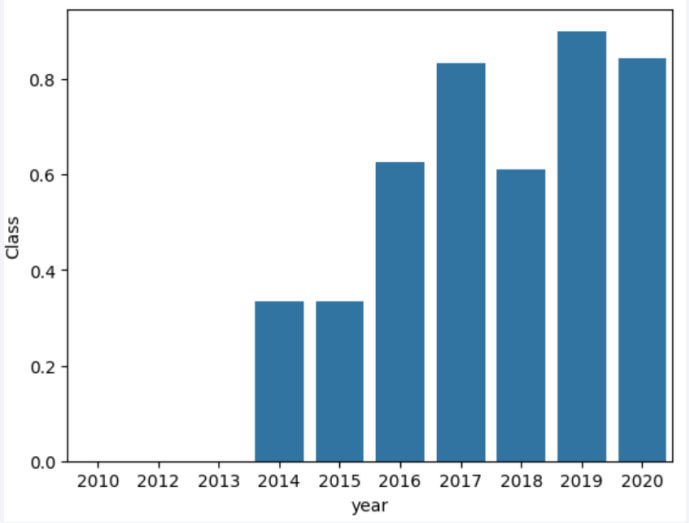
Predictive Analysis (Classification)

- ML Modes used to predict Landing Outcome:
  - Logistics Regression
  - Support Vector Machine
  - Decision Tree
  - K Nearest Neighbours

- https://github.com/chrisejugbo/Capstone_Project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

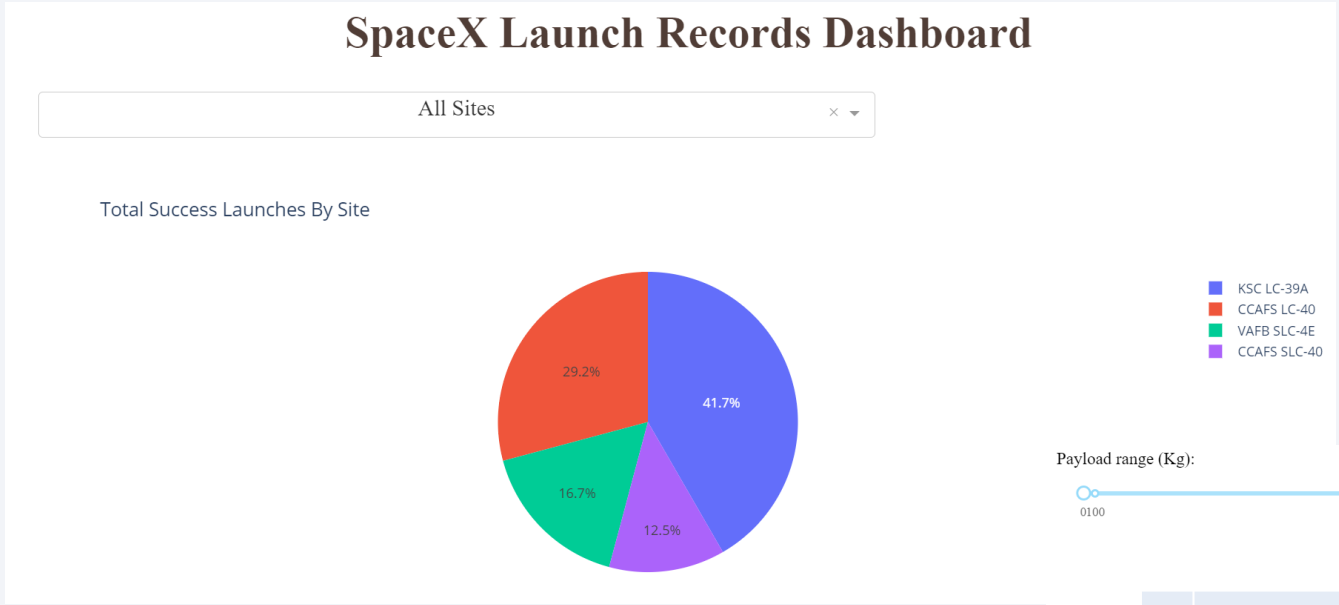
Results

- Exploratory data analysis results
  - SpaceX uses 4 launches sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40,
  - Average Payload of Falcon 9 , Booster v1.1 B1003 is 2534 KG
  - First Successful Launch happened on December 22nd, 2015,
  - Four Orbits have a 100% success rate
  - Success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing:



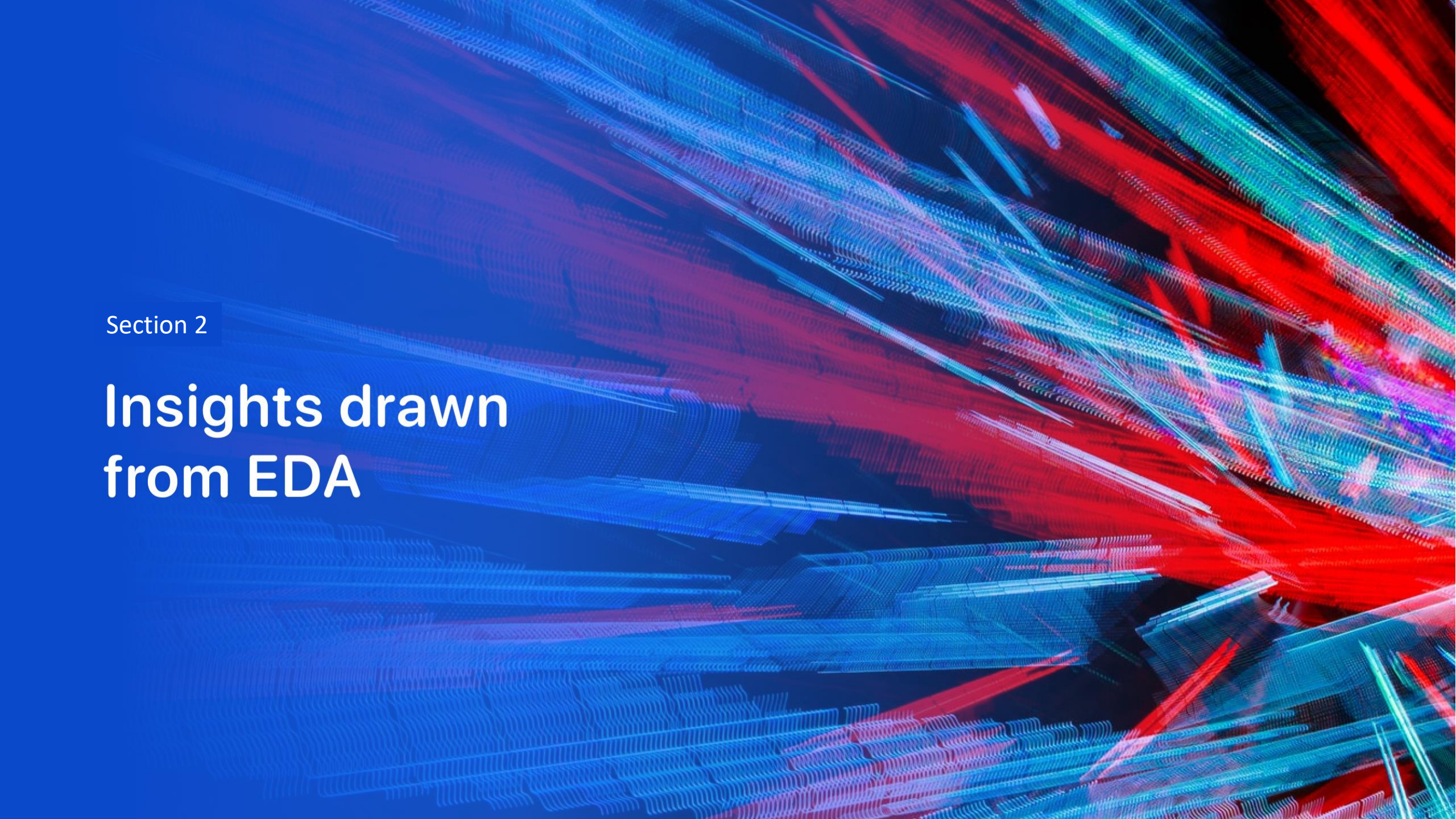
Results

- Interactive analytics demo in screenshots

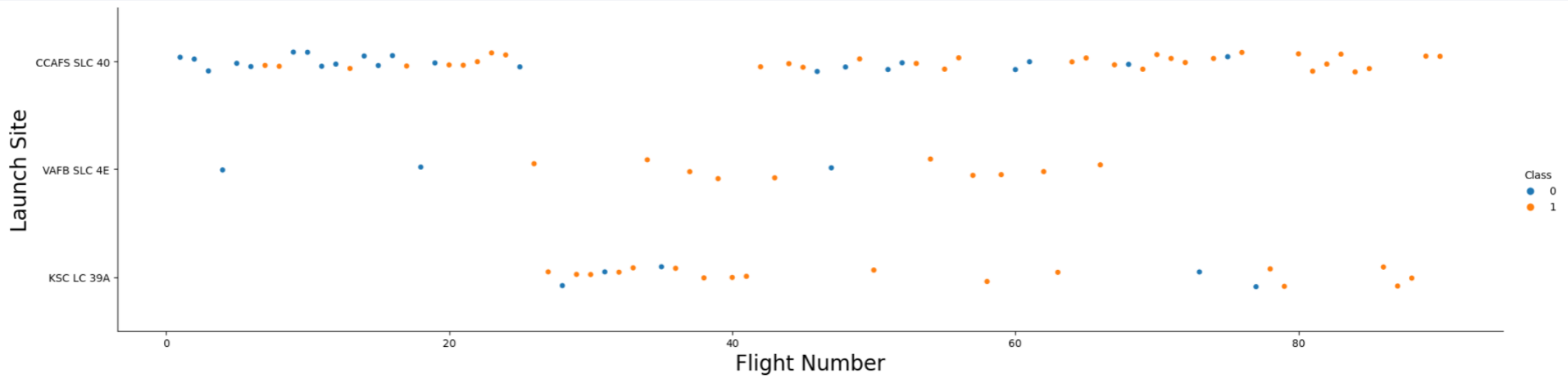


Results

- Predictive analysis results
  - Logistics Regression Accuracy: 0.83
  - SVM Acuracy: 0.83
  - Decision Tree Accuracy: 0.78
  - KNN Accuracy: 0.83
- All models provide except Decision Tree give the same accuracy

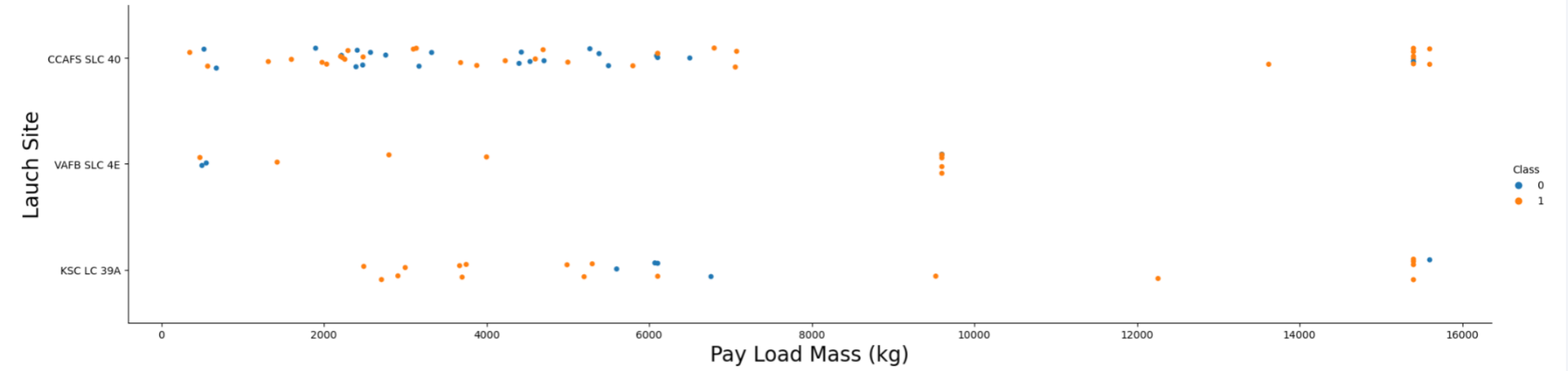

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Successful launches become more frequent as Flight Number increases
- There are significantly fewer launches with VAFB SLC 4E
- KSC LC 39A has a significantly higher success rate

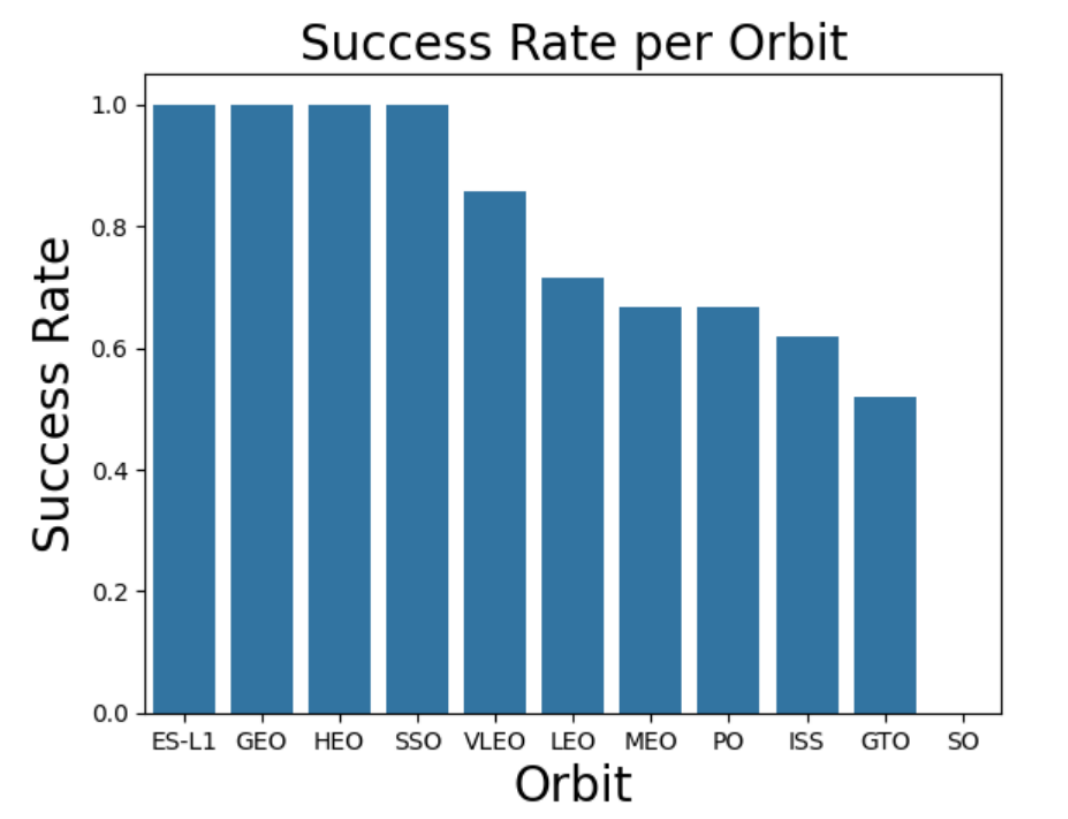
Payload vs. Launch Site



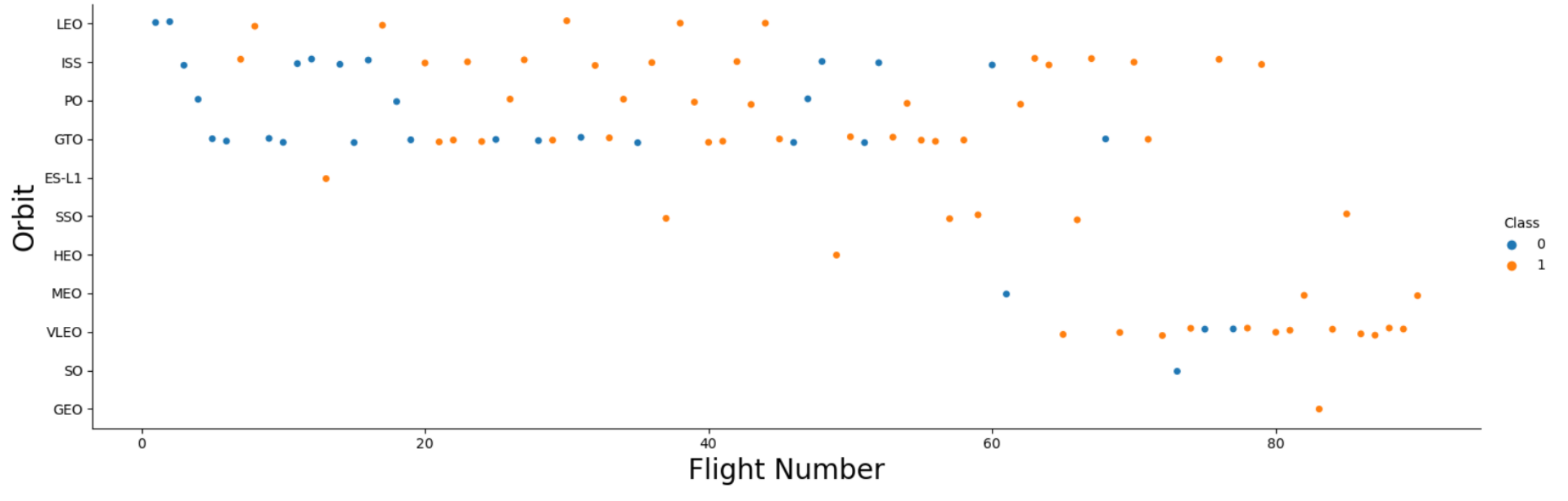
- Almost all flights with a payload mass over 6000 kg are successful
- Higher payload masses are more frequently launched from KSC LC 39A and CCAFS SLC 40, which indicates that these sites prefer higher payload flights

Success Rate vs. Orbit Type

- Four Orbits shows 100% success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- VLEO, LEO, MEO, PO & ISS show medium success rate
- Low Success rates:
  - GTO – below 50%
  - SO – 0% success rate

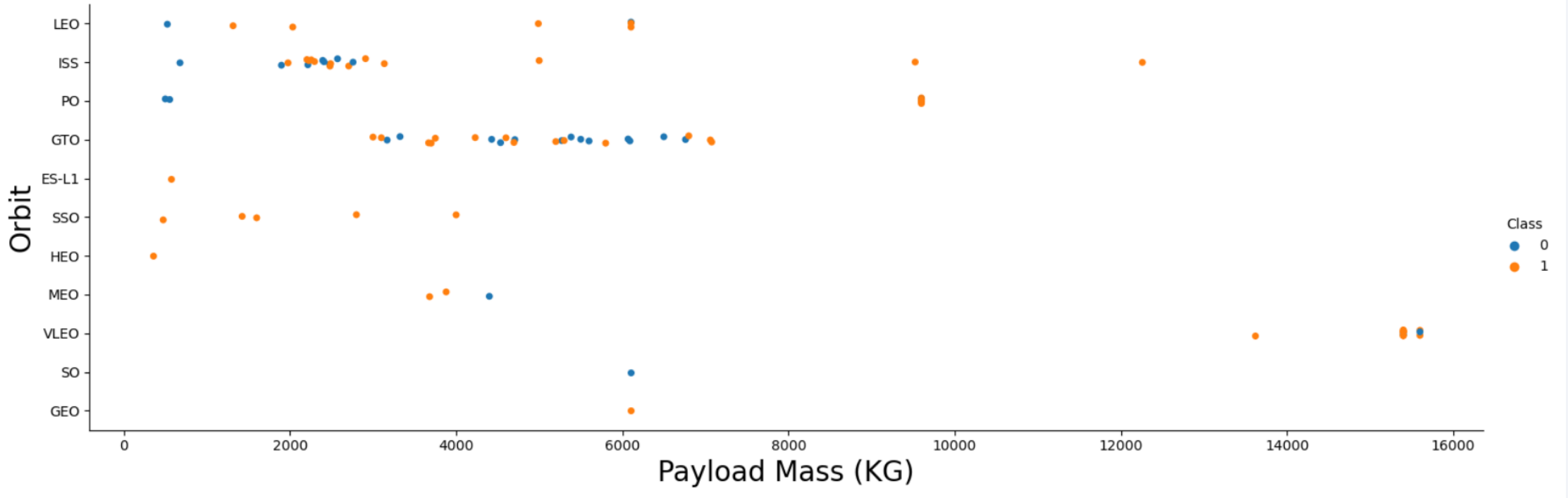


Flight Number vs. Orbit Type



- There’s a visible increase in successful flights as the flight number increases across all Orbit Types.
- Orbit types ES-L1, GEO, HEO, SSO have the highest success rate, however, they also have the least amount of flights.
- Therefore, Orbit Type alone cannot be a single identifier of success

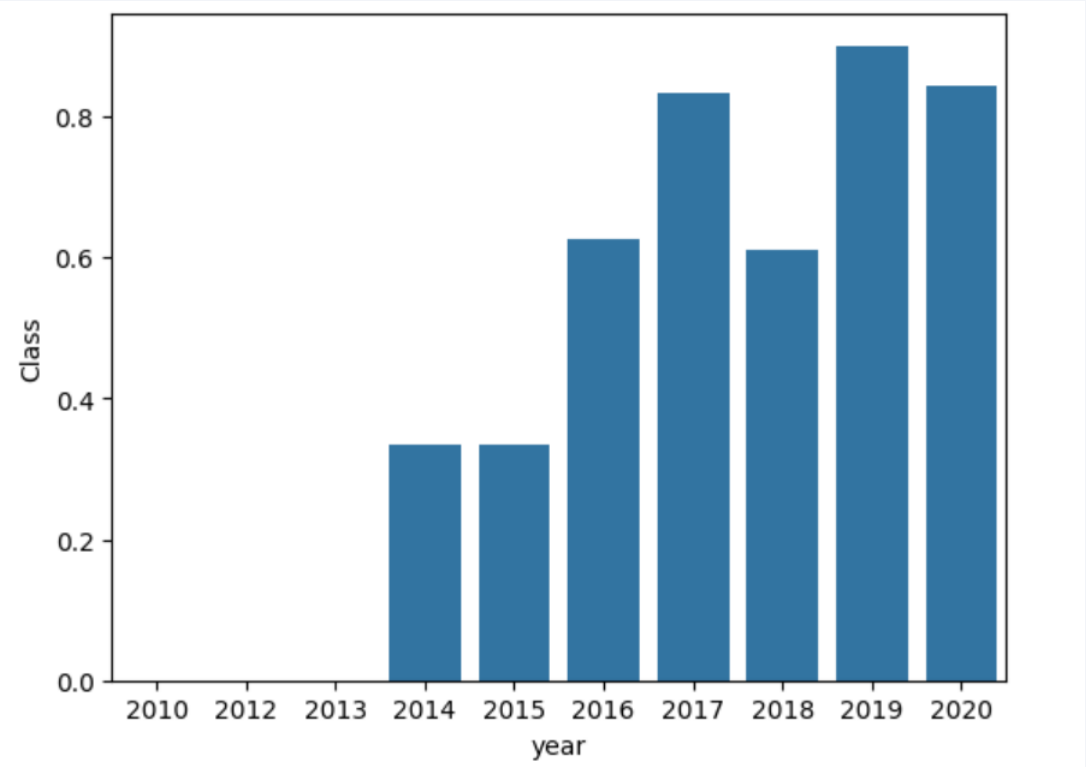
Payload vs. Orbit Type



- Higher payloads typically show greater success rate
- Payload doesn’t show an impact on success rate in GTO orbit

Launch Success Yearly Trend

- Success rate kept increasing from 2015 to 2017
- In 2018 the success rate took a significant plunge
- 2019 has the highest success rate



All Launch Site Names

- Launch Site Names:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- Select Distinct Query was used to gather unique Launch Site names

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

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

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

Launch Site Names Begin with 'CCA'

- WHERE and LIKE statements where used to find records where Launch Site names start with “CCA”
- LIMIT statement was used to only show 5 records

In [26]:
```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE Launch_Site LIKE "CCA%"
LIMIT 5
```

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

Out[26]:

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

Total Payload Mass

- WHERE clause was used to select NASA (CRS) as the customer
- Total Payload was calculated by doing a sum of PAYLOAD_MASS__KG_

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

```
%%sql
SELECT Customer, sum(PAYLOAD_MASS__KG_) AS Total_Payload
FROM SPACEXTABLE
WHERE Customer = "NASA (CRS)"
GROUP BY Customer
```

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

Customer	Total_Payload
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Average Payload mass was calculated by avg() function
- F9 v1.1 was filtered in the WHERE clause

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

```
%%sql
SELECT Booster_Version, avg(PAYLOAD_MASS__KG_) AS Average_Payload_Mass
FROM SPACEXTABLE
WHERE Booster_Version LIKE "F9 v1.1%"
```

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

Booster_Version	Average_Payload_Mass
F9 v1.1 B1003	2534.6666666666665

First Successful Ground Landing Date

- The earliest date was selected using min(Date)
- The Successful landing was filtered using WHERE clause

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

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- WHERE clause was used to filter:
  - Landing Outcome = "Success (drone ship)“
  - Payload Mass BETWEEN 4000 AND 6000

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

Total Number of Successful and Failure Mission Outcomes

- Mission outcomes Success vs Failure counts were acquired using the Group by statement and Select count(*)

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

Boosters Carried Maximum Payload

- WHERE clause and a Subquery was used to gather the Booster Versions who have had a Payload Mass that’s equal to 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

2015 Launch Records

- Year was acquired using the substr function
- Where clause used to filter year and Landing outcome

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

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

- WHERE and BETWEEN clause was used to filter dates between:
  - 2010-06-04 and 2017-03-20
- Order By clause together with DESC was used to rank them in a descending order.

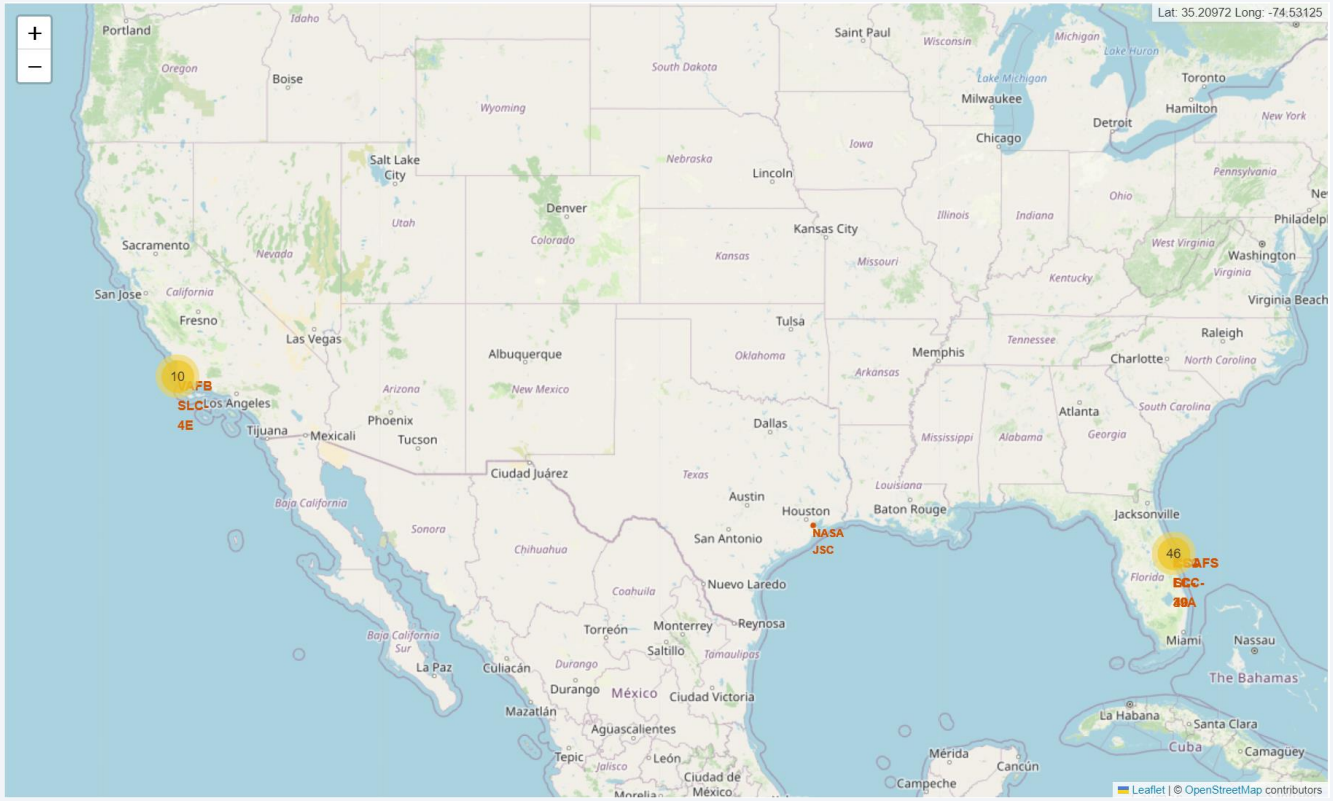
Landing_Outcome	count(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

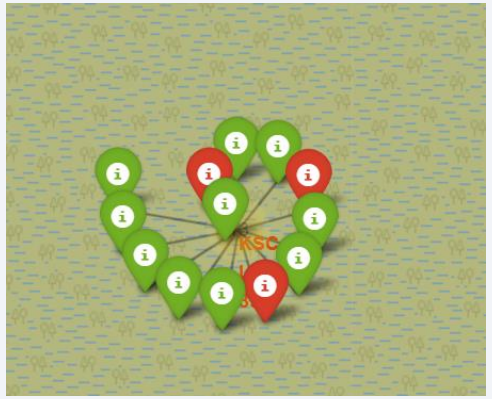
- Launch sites are located very close to the coasts
- There’s only 1 site on west coast
- There are 3 sites on the east coast



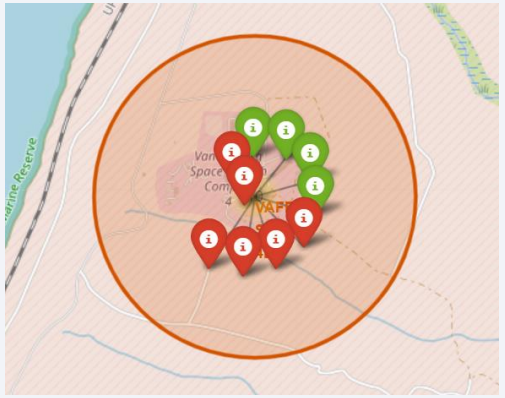
Launch Success per Site

- CCAFS SLC-40 – has the least amount of launches
- KSC LC-39A – has the best success rate
- CCAFS LC-40 – has the most amount of launches

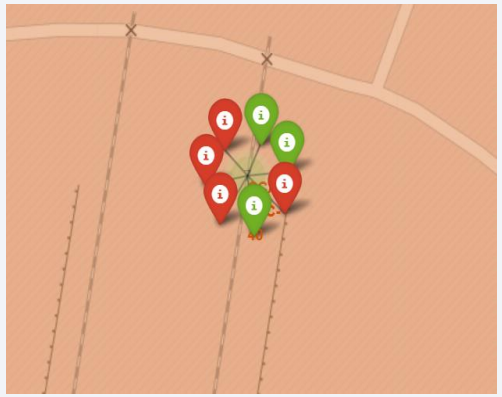
KSC LC-39A



VAFB SLC-4E



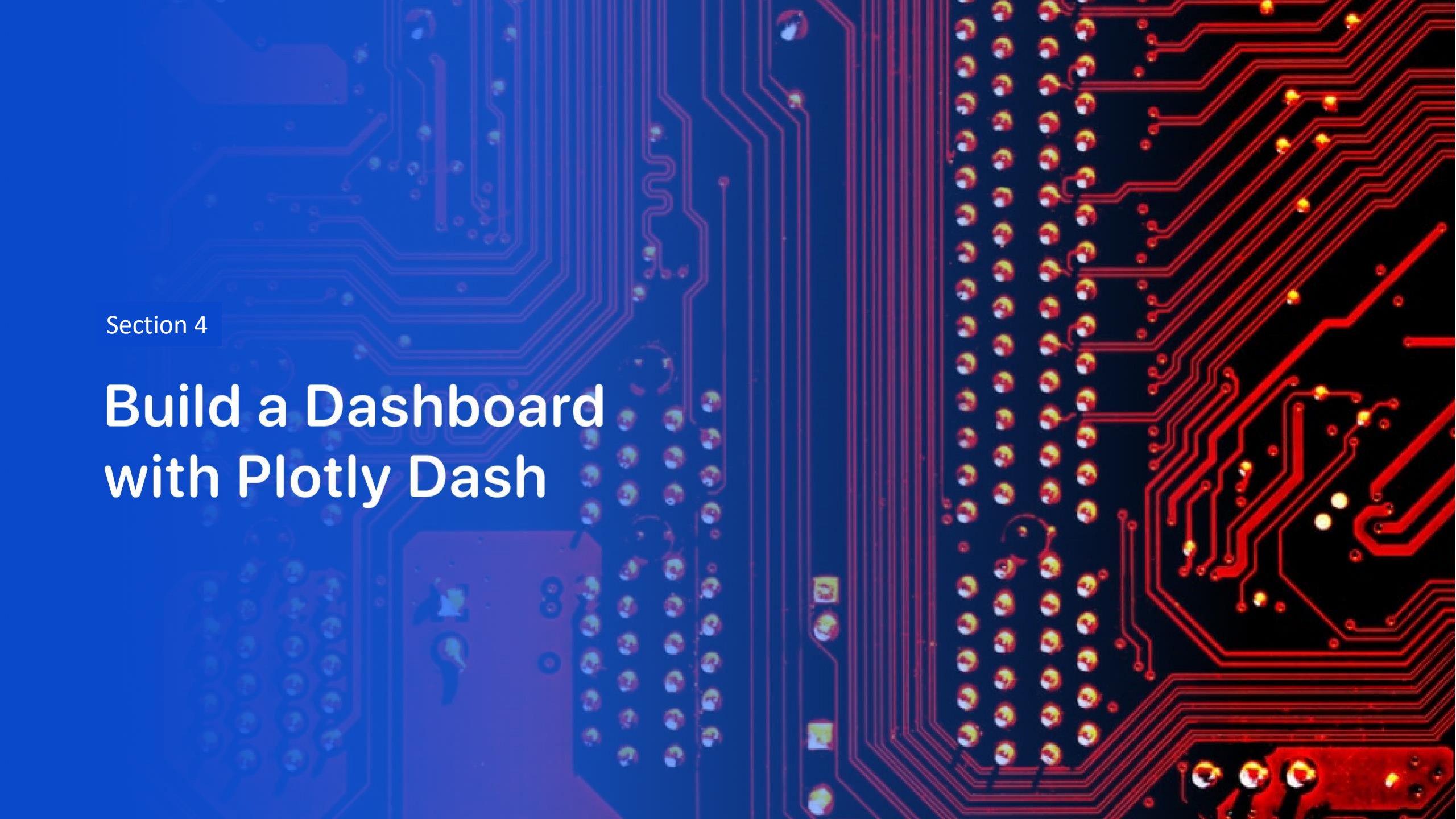
CCAFS SLC-40



CCAFS LC-40

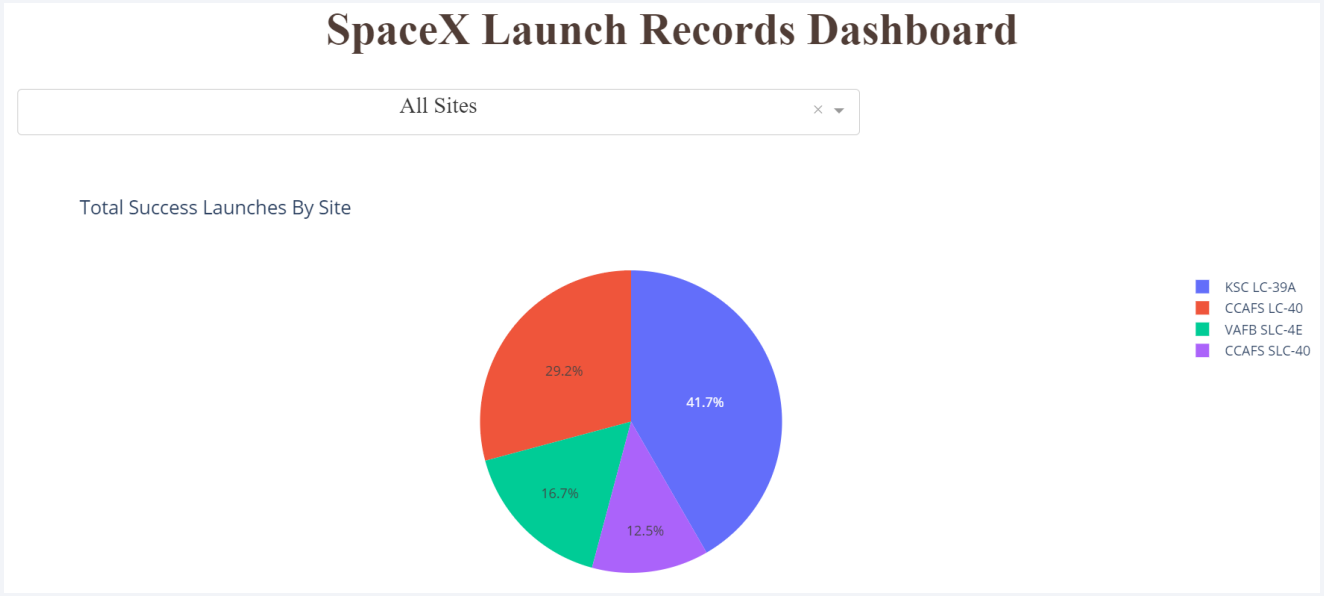
Launch Site proximities

- Launch Site is relatively close to Railways, Coastlines and Highways
- Launch Site has a much greater distance to the nearest city


Section 4
Build a Dashboard with Plotly Dash

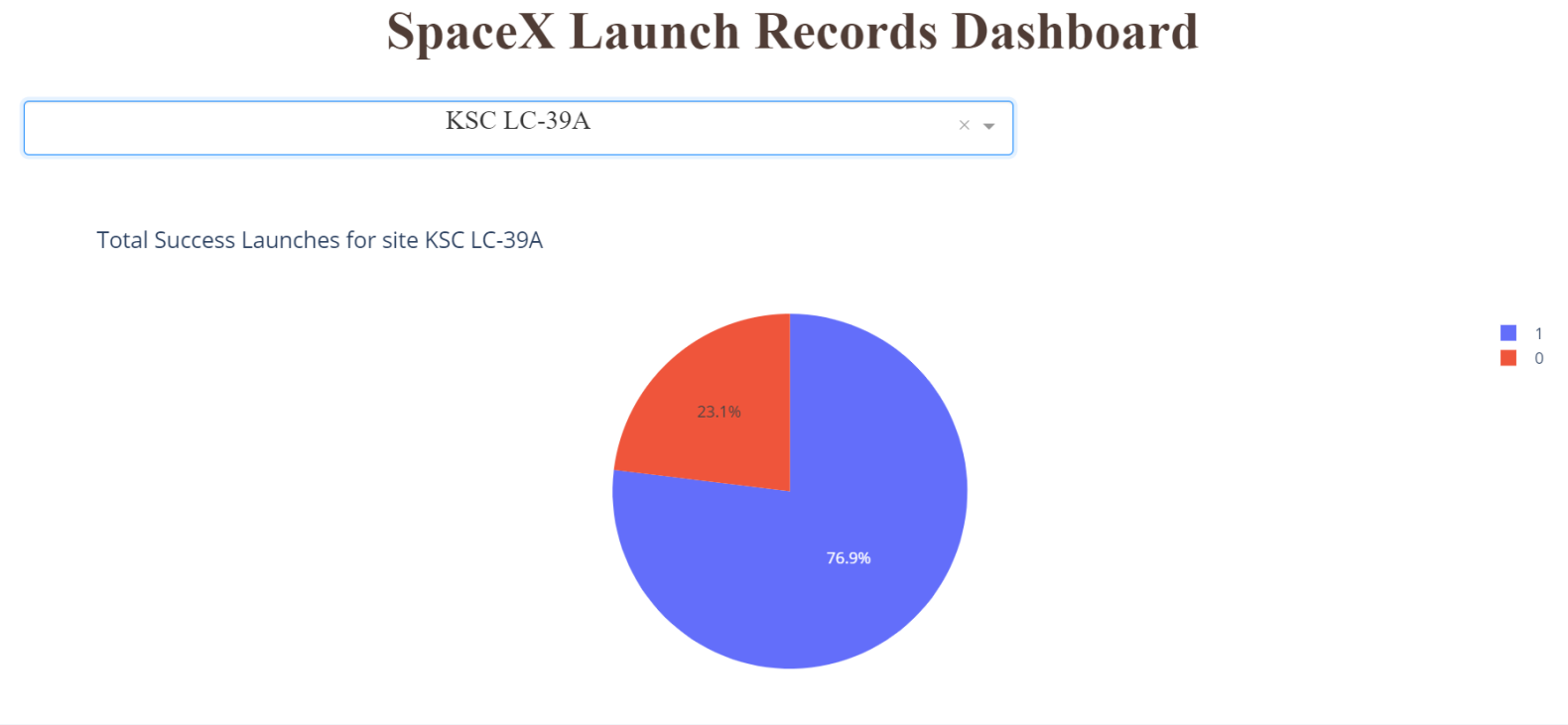
Success Rate per Launch Site

- Success Rate is the highest for KSC LC-39A and CCAFS LC_40 site
- CCAFS SLC-40 has the lowest success rate, even though it is located very close to CCAFS LC_40



Successful vs unsuccessful launches for KSC LC-39A

- Successful launches: 76.9%
- Unsuccessful launches: 23.1%



Success Rate per Payload Mass

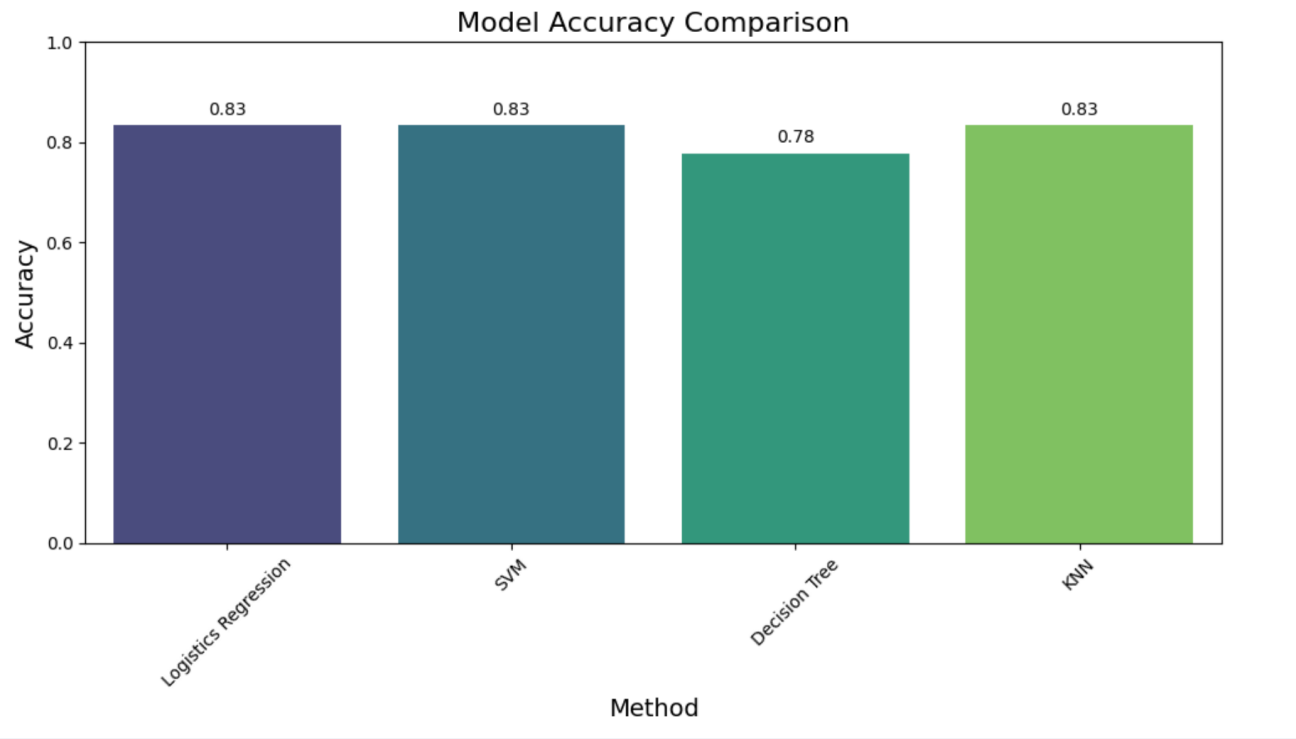
- There are only 2 booster version categories for launches with apayload mass above 6000 kg
- B4 is the only Booster that has a successful launches above 9000 kg payload

Section 5

Predictive Analysis (Classification)

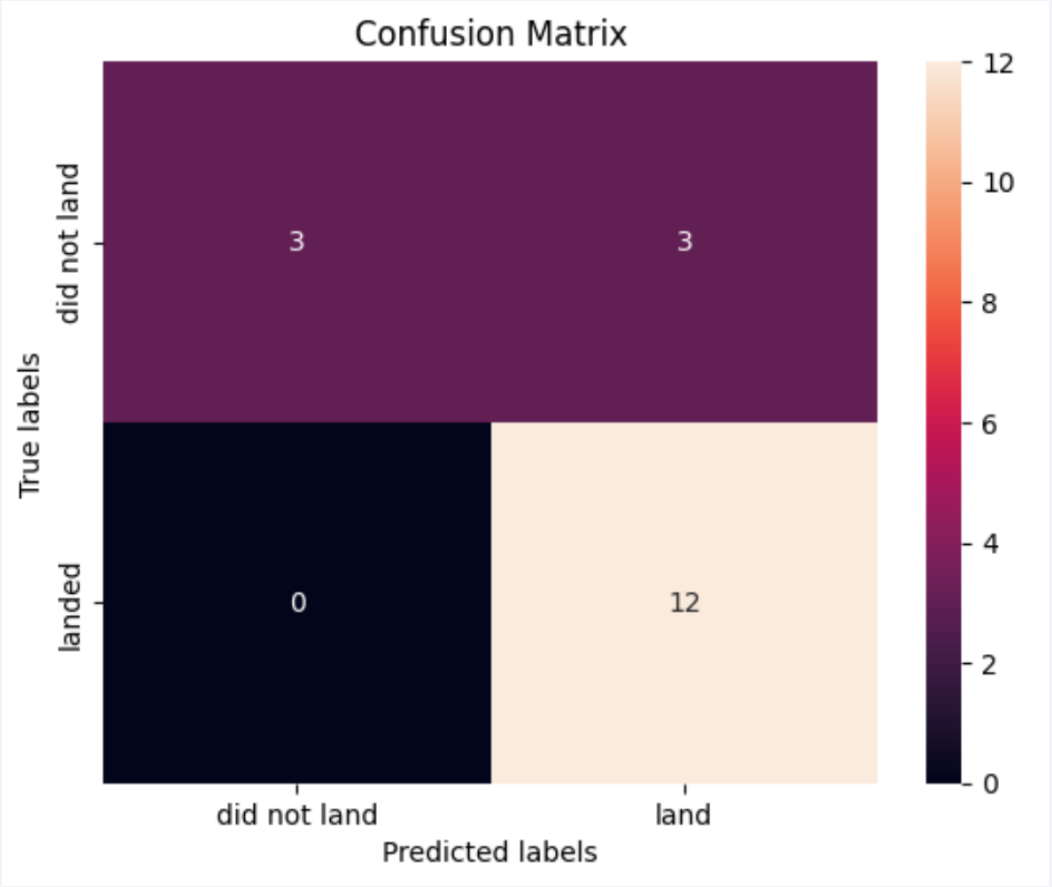
Classification Accuracy

- After testing four models, Decision Tree shows the lowest accuracy at 0.77
- Rest of the models show the same accuracy score at 0.83 thus being the highest



Confusion Matrix

- The Confusion Matrix shows high accuracy:
  - True Negatives: did not land and predicted as did not land: 3
  - False Positives: did not land but predicted as landed: 3
  - False Negatives: landed but predicted as did not land: 0
  - True Positives: landed and predicted as landed: 12
- There are 15 True Statements and 3 False Statements



Conclusions

- Launches with Payload Mass above 8000 have a significantly higher success rate
- All classification models except for Decision Tree provided high accuracy at 0.83
- Logistics Regression, SVM and KNN can all be used for predictions.

Appendix

- Read Me page: https://github.com/chrisejugbo/Capstone_Project/blob/main/ReadMe

- Data Set links
  - https://github.com/chrisejugbo/Capstone_Project/blob/main/dataset_part_1.csv
  - https://github.com/chrisejugbo/Capstone_Project/blob/main/dataset_part_2.csv
  - https://github.com/chrisejugbo/Capstone_Project/blob/main/dataset_part_3.csv
  - https://github.com/chrisejugbo/Capstone_Project/blob/main/spacex_web_scraped.csv

Thank you!