IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Amer Chowdhury
12/12/22

Outline

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

Executive Summary

Summary of methodologies

- Data Requirements - To provide information related to successful launches etc.
- Data Collection - Web scraping and utilising SpaceX API
- Data Wrangling - Processing data by removing missing values and organising landing outcomes in a way that can be used to train the model ie Good and Bad outcomes as 1 and 0.
- EDA - With Data Visualisation - Building interactive maps with Folium and Dashboards with Dash
- EDA - With Querying using SQL
- Model Building - Predictive Classification Model
- Evaluation - Train Test Split Result Analysis

Summary of all results

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

Introduction

- Project background and context
  - The aim of this project was to predict if the Falcon 9 first stage will land successfully. 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.

- Problems you want to find answers
  - 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

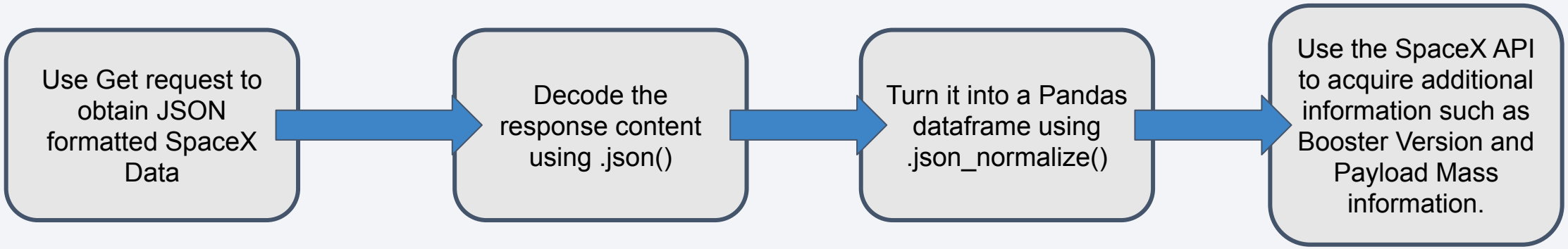
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected via web scraping and via the SpaceX API to gather more relevant information
- Perform data wrangling
  - Data Wrangling was carried out by identifying missing values followed by replacing values which were NaN - “Not a Number” with the respective mean value instead.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create a column for the class which we want to classify.
  - Standardize the data to improve consistency
  - Split the data into training data and test data
  - Find the best Hyperparameters for various machine learning techniques such as Support Vector Machines, Classification Trees and Logistic Regression.
  - Use the R^2 score and accuracy to evaluate the classification models.

Data Collection

- Datasets were collected from launch records posted on Wikipedia. These results were web scraped using Beautiful Soup. Next using Get requests we turned JSON formatted rocket information into a tabular structure by normalizing it into a Pandas dataframe before using the SpaceX API to retrieve additional information using the identification numbers in the Launch Data



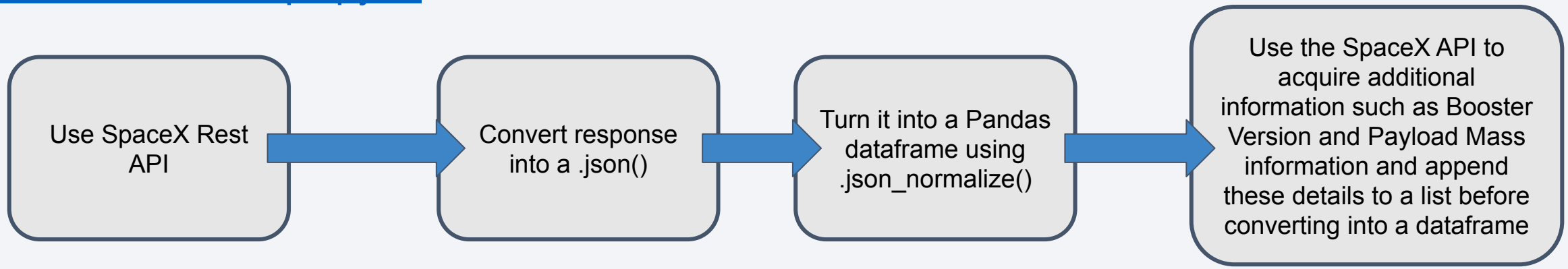
Data Collection – SpaceX API

- Data collection with SpaceX REST calls using key phrases and flowcharts

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Completed SpaceX API calls notebook :

https://github.com/Amertastic/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



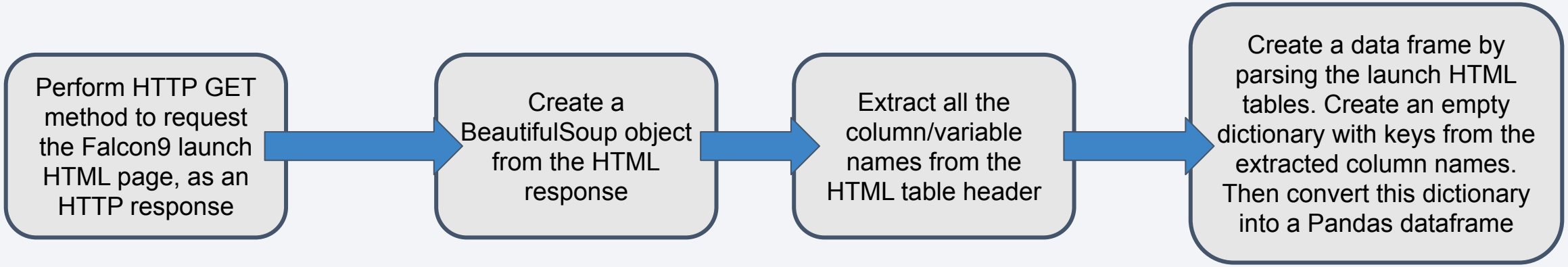
Data Collection - Scraping

- Web scraping process using key phrases and flowcharts

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	<generator object Tag._all_strings at 0x7fcc3b...	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	<generator object Tag._all_strings at 0x7fcc3b...	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	<generator object Tag._all_strings at 0x7fcc3d...	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	<generator object Tag._all_strings at 0x7fcc3b...	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	<generator object Tag._all_strings at 0x7fcc3b...	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

GitHub URL of the completed web scraping notebook:

https://github.com/Amertastic/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

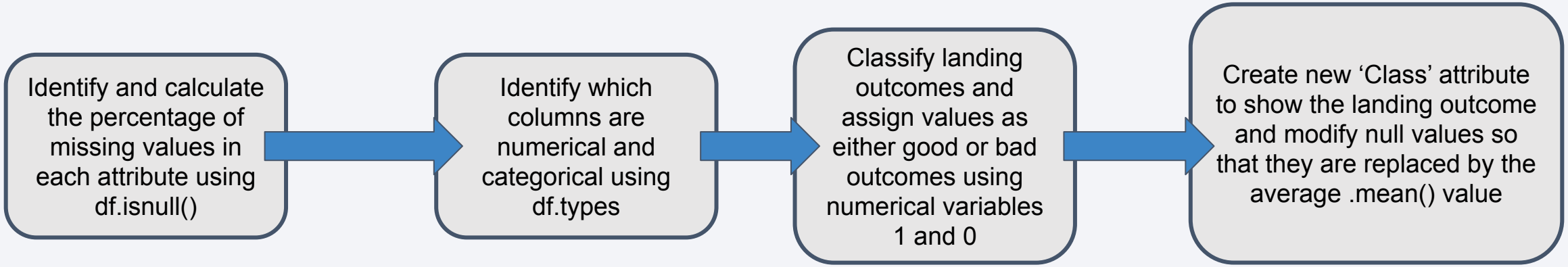


Data Wrangling

Data Wrangling Summary :

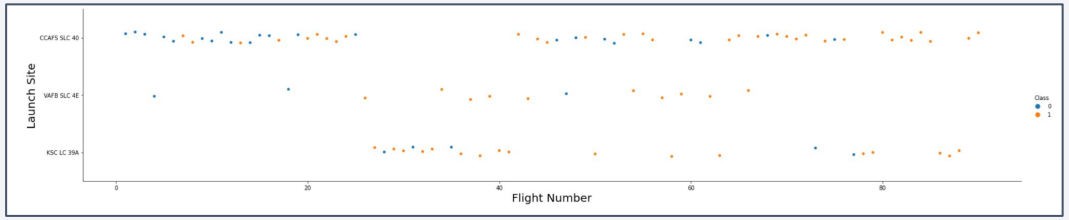
- Remove missing values and replace with mean values
- Classify landing outcomes as Good or Bad with 1s and 0s respectively

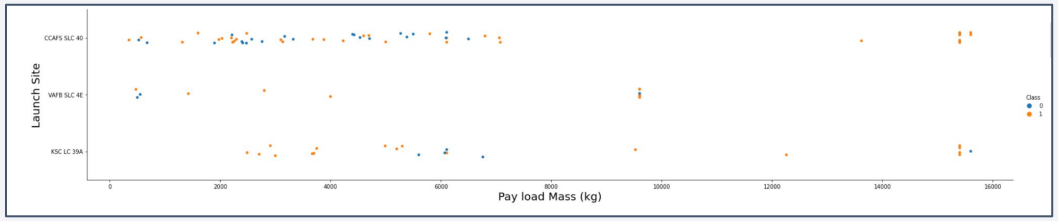
GitHub URL of the completed notebook :
https://github.com/Amertastic/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

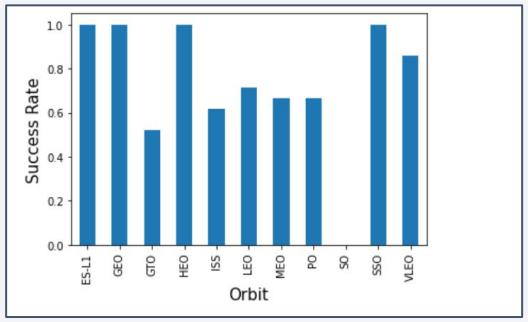


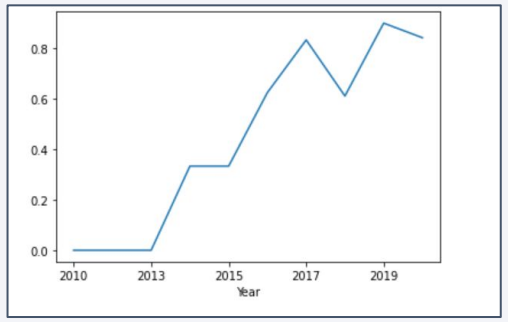
EDA with Data Visualization

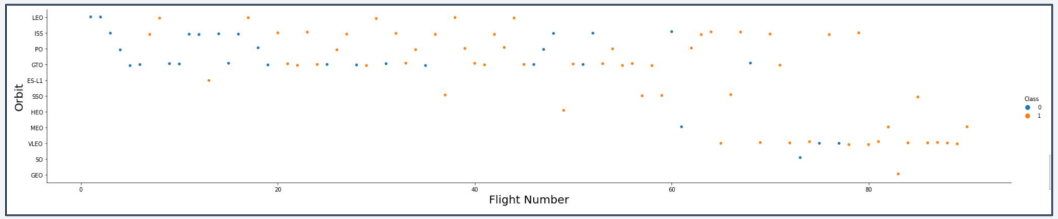
The following charts were used to help identify trends in the data which could be used for training the classification model.

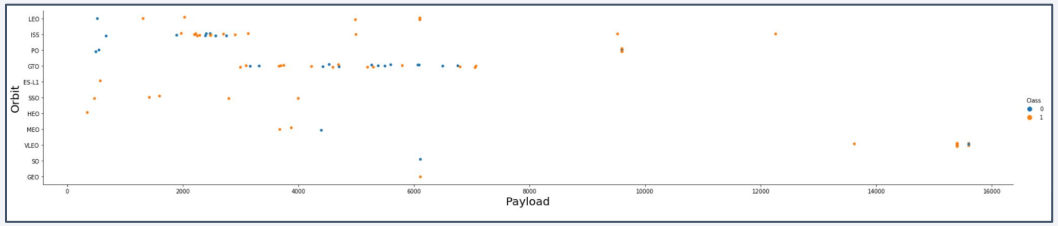












GitHub URL :
https://github.com/Amertastic/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

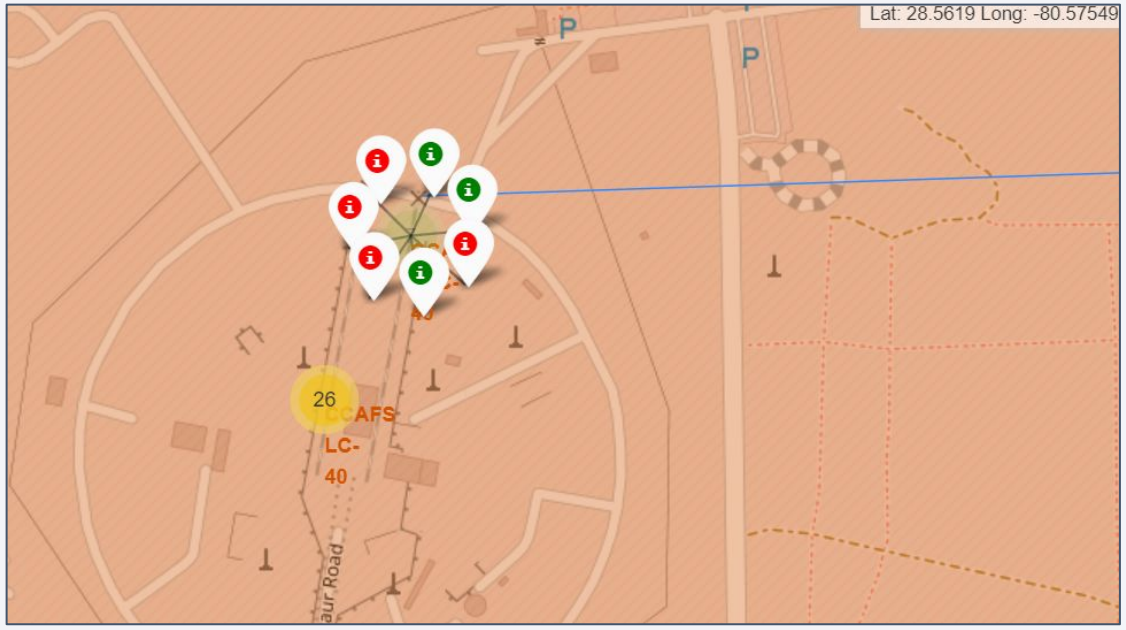
Summary of the SQL queries you performed:

- 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 a 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.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub URL :
https://github.com/Amertastic/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

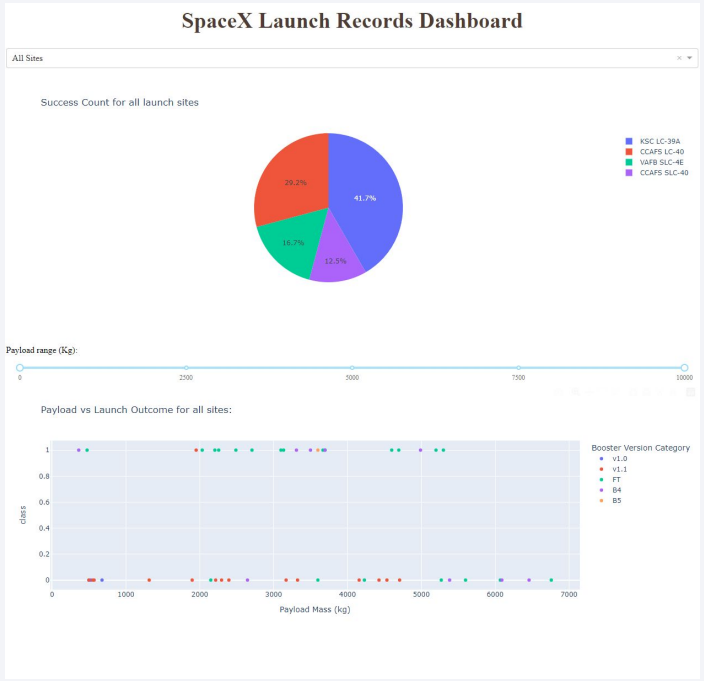
Map objects such as markers, circles, lines and clusters were created and added to a Folium map.

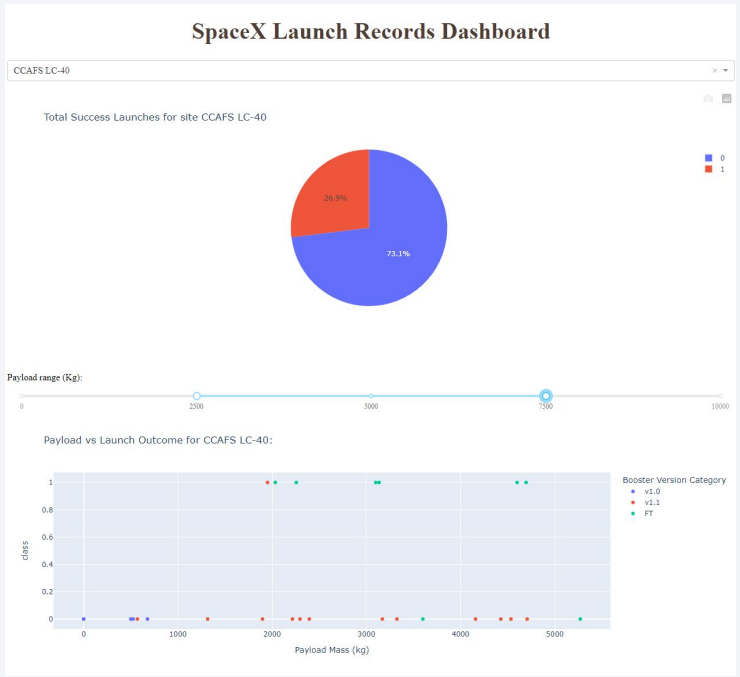


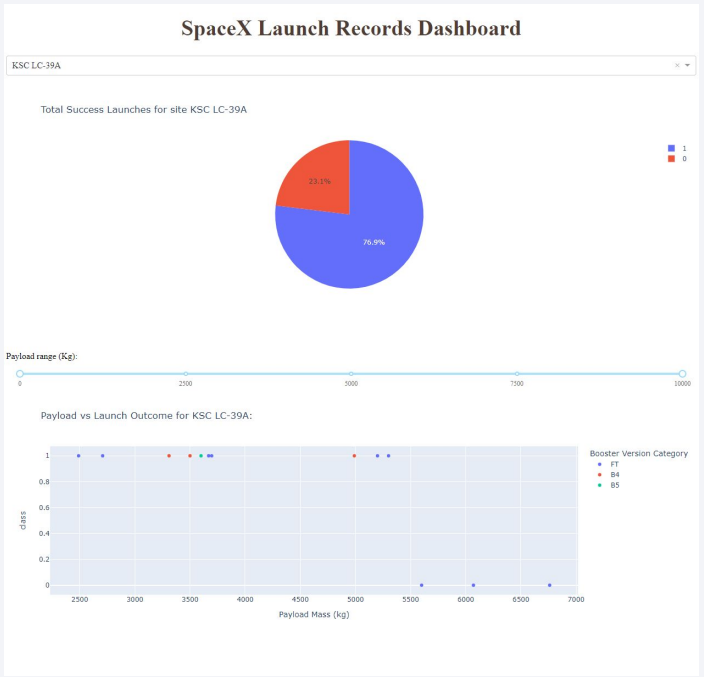
This was carried out in order to help perform more interactive visual analytics in order to find some geographical patterns about launch sites.

GitHub URL : https://github.com/Amertastic/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash







A Plotly Dash application was created to perform visual analytics on SpaceX launch data in real-time.

GitHub URL :
https://github.com/Amertastic/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Summary of Machine Learning Model creation steps:

- Standardize the data. This is an important step to do before training for most machine learning models.
- Split the Data into training and test sets
- Use GridSearchCV to find the best hyperparameters for each of the Machine Learning models tested.
- Create models for Linear Regression, Support Vector Machine, Decision Tree Classifier and K Nearest Neighbors.
- Identify the method which performs best by using model evaluation criteria such as R^2 error, .score() and .best_score_ to determine accuracy

The best hyperparameters were determined using GridSearchCV cross validation

GitHub URL : https://github.com/Amertastic/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

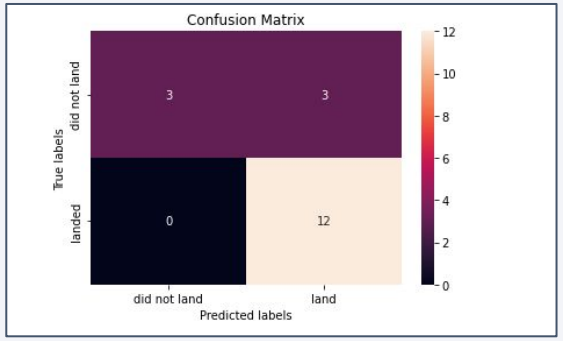
Results

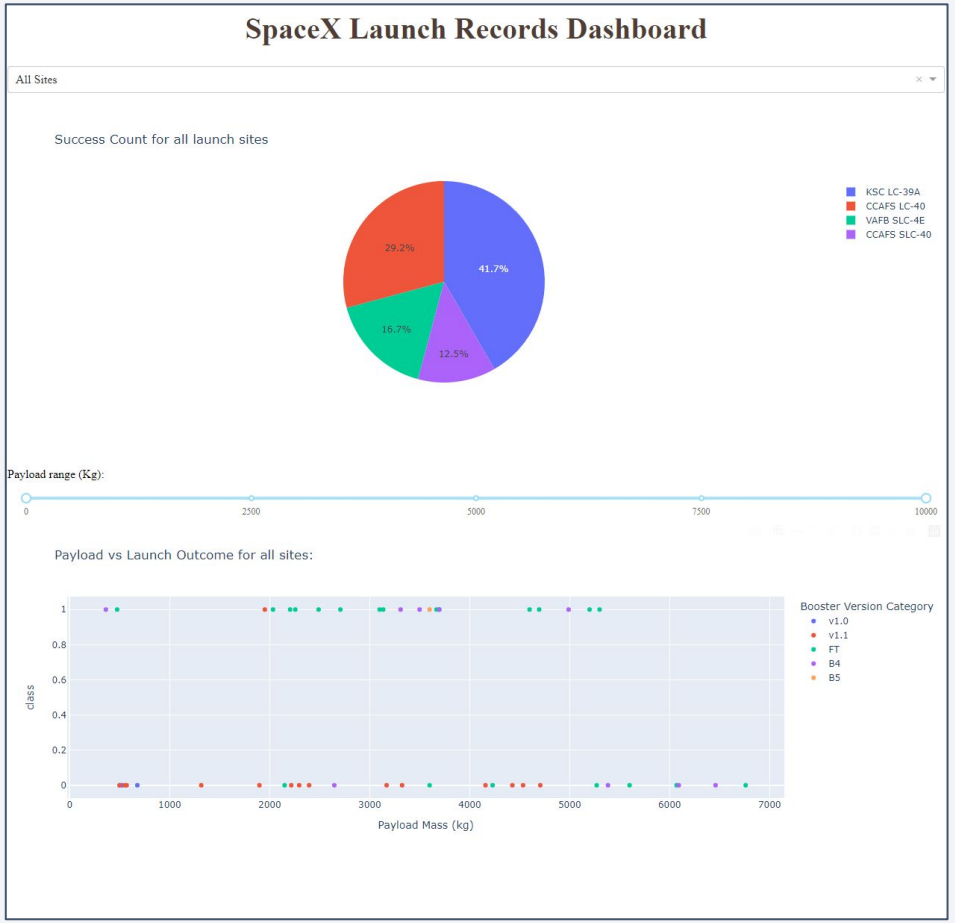
Each of the models Logistic Regression, SVM, Decision Tree and KNN had the same R^2 error or .score() value of 0.83333..

The Decision Tree had a higher accuracy .best_score_ value of 0.84214 …

KSC LC-39A had the highest number of successful launches out of all the launch sites.

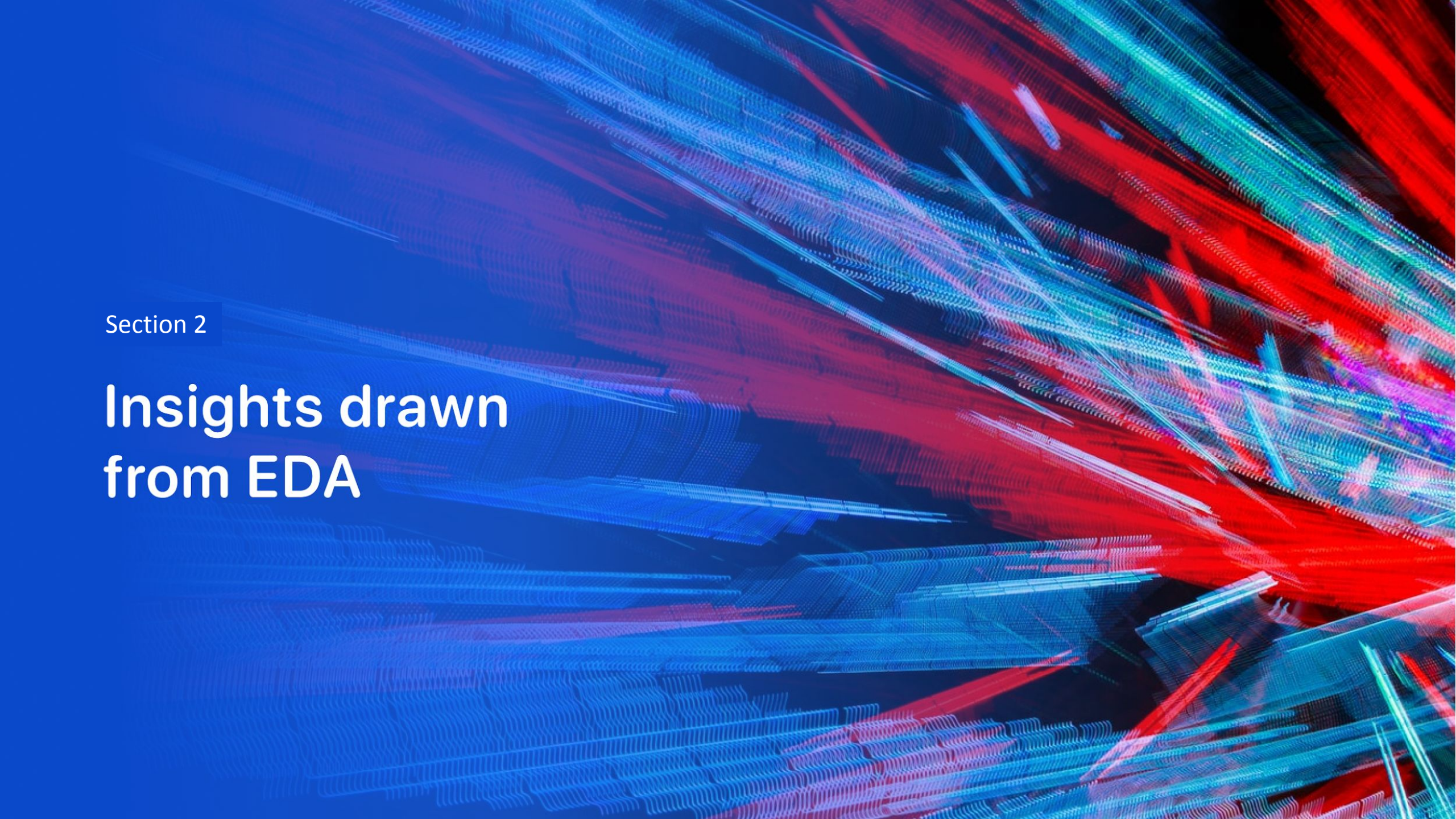
The major problem with the model was false positives when using all methods.





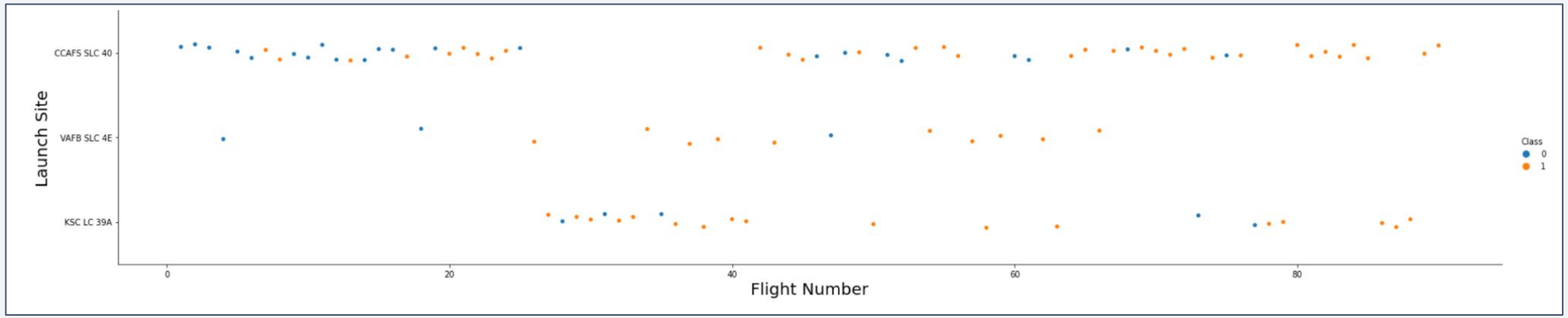
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

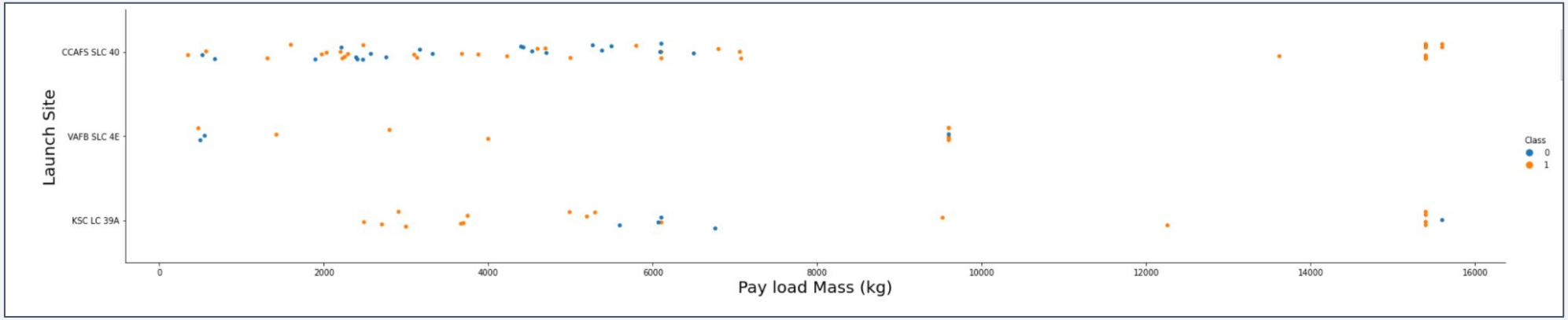
Plot of Flight Number vs. Launch Site



Explanation : There are more launches from the CCAFS SLC 40 site than the other two sites. The number of successful landings have increased since first using the Launch Site

Payload vs. Launch Site

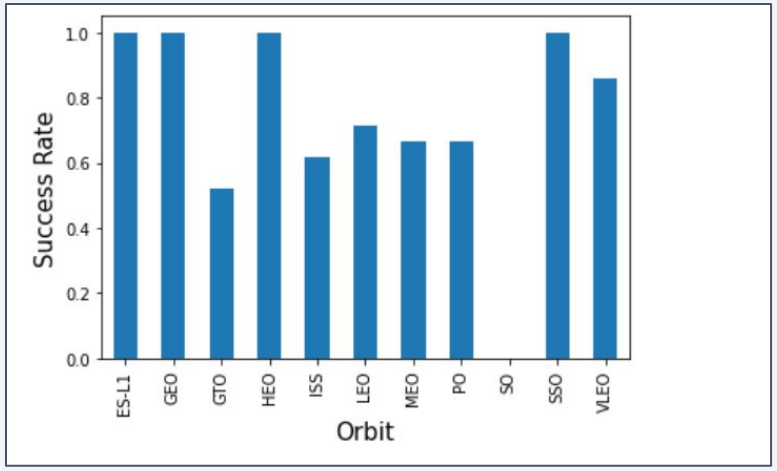
Scatter plot of Payload vs. Launch Site



Explanation : There appears to be a higher success rate for landings when the Pay Load Mass is above 8000 kg. Most of the Launches under 8000 kg took place at the CCAFS SLC 40 Launch Site.

Success Rate vs. Orbit Type

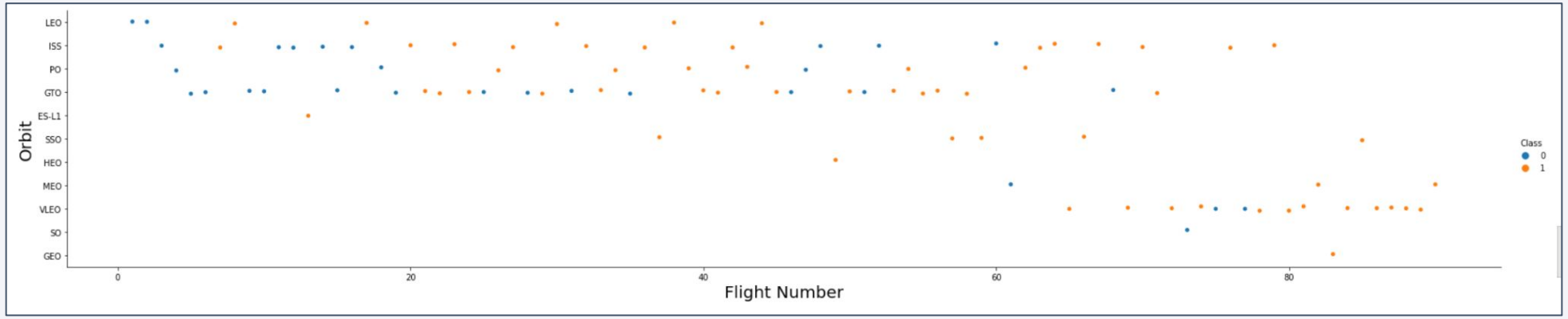
Bar chart for the success rate of each orbit type



Explanation : The orbit types ES-L1, GEO, HEO and SSO have the highest success rate.

Flight Number vs. Orbit Type

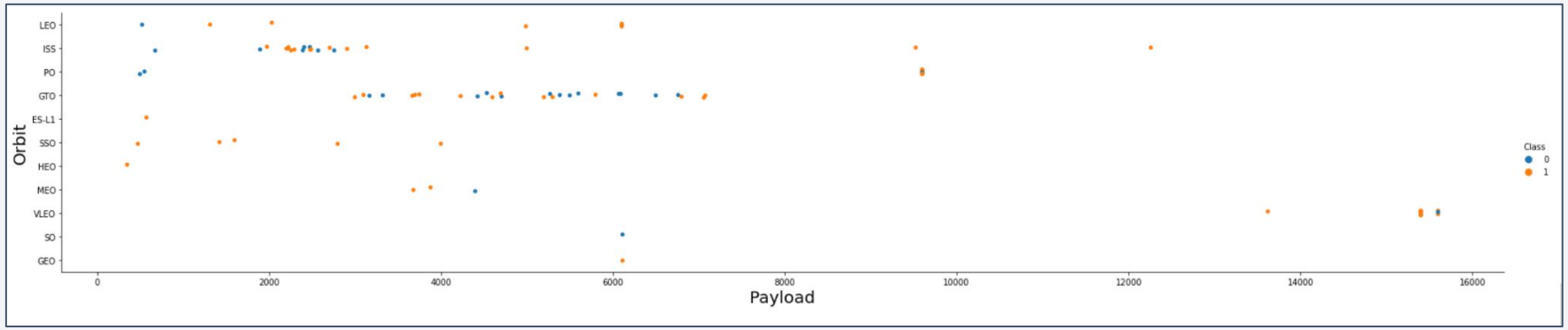
Scatter point of Flight number vs. Orbit type



Explanation : It would appear that a large number of the recent launches have been Very Low Earth orbit - (VLEO)

Payload vs. Orbit Type

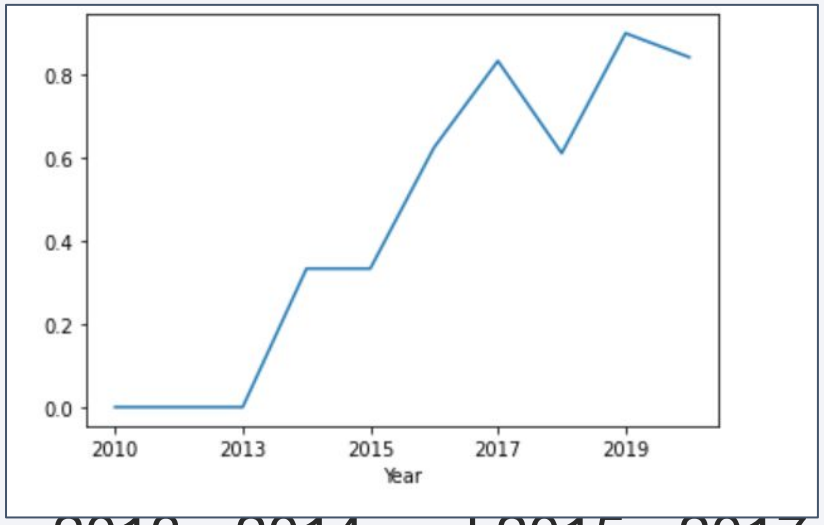
Scatter point of payload vs. orbit type



Explanation : Payload masses between 2000 kg and 3000 kg tend to be orbit type ISS. Payload masses between 3500 kg and 7000 kg tend to be orbit type GTO.

Launch Success Yearly Trend

Line chart of yearly average success rate



Explanation : Between 2013 - 2014 and 2015 - 2017 average launch success rates increased significantly. After 2018 the success rate seems to have plateaued

All Launch Site Names

The names of the unique launch sites

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

Query result with a short explanation here

```
# Unique Launch Sites
%sql SELECT DISTINCT("Launch_Site") FROM SPACEXTBL
```

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	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Query result with a short explanation here

```
%sql SELECT * FROM SPACEXTBL WHERE ("Launch_Site") LIKE 'CCA%' LIMIT(5)
```

Total Payload Mass

The total payload carried by boosters from NASA

sum("PAYLOAD_MASS__KG_")
48213

Query result with a short explanation here

```
%sql SELECT sum("PAYLOAD_MASS__KG_")  FROM SPACEXTBL WHERE "CUSTOMER" LIKE "%CRS%"
```

Average Payload Mass by F9 v1.1

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

avg("PAYLOAD_MASS__KG_")
2928.4

Query result with a short explanation here

```
%sql SELECT avg("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booster_Version" LIKE 'F9 v1.1'
```

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad

MIN(DATE)	Landing _Outcome
01-05-2017	Success (ground pad)

Query result with a short explanation here

```
%sql SELECT MIN(DATE), "Landing _Outcome" FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%ground%'
```

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Query result with a short explanation here

```
%sql SELECT "Booster_Version", "Landing _Outcome", "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE ("Landing _Outcome" LIKE '%Success%drone%') AND ("PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000)
```

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

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

Query result with a short explanation here

```
%sql SELECT "Mission_Outcome", COUNT(*) FROM SPACEXTBL GROUP BY "Mission_Outcome"
```

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass

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

Query result with a short explanation here

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

2015 Launch Records

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

Month	Year	Landing _Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Query result with a short explanation here

```
%sql SELECT substr(Date,4,2) AS "Month", substr(Date,7,4) AS "Year", "Landing _Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTBL WHERE "Landing _Outcome" LIKE ("Failure%drone%") AND "Date" LIKE "%2015%"
```

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

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

Date	Landing _Outcome	Landing_Outcome_Count
08-01-2018	Success (ground pad)	6
11-05-2018	Success (drone ship)	8
06-12-2020	Success	20

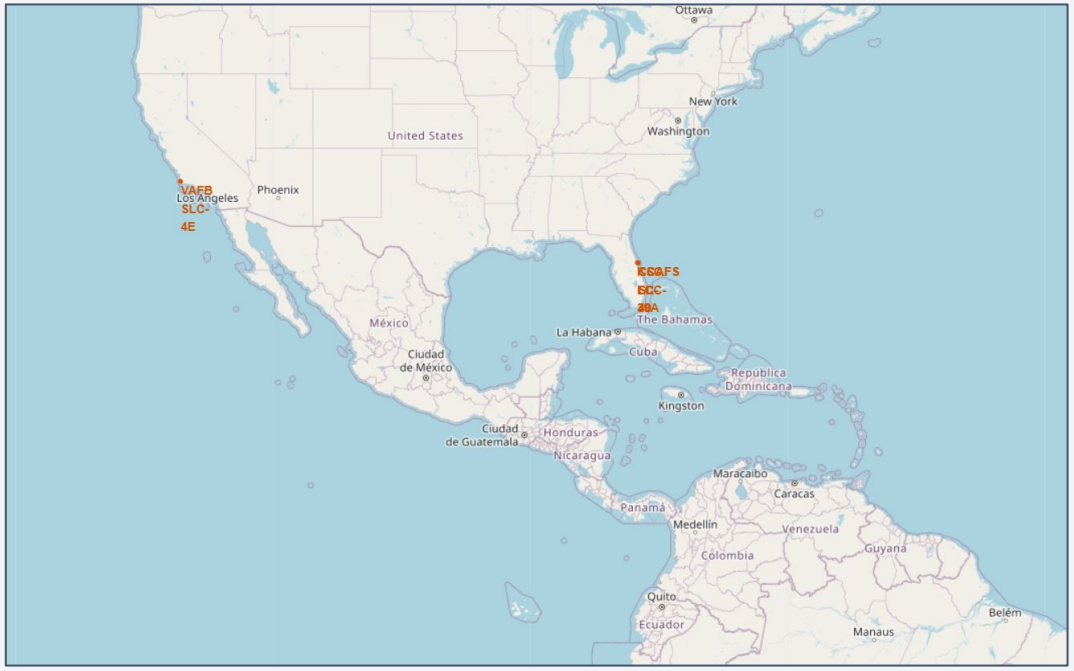
Query result with a short explanation here

```
%sql SELECT "Date", "Landing _Outcome", COUNT("Landing _Outcome") as "Landing_Outcome_Count" FROM SPACEXTBL \
WHERE ("Landing _Outcome" LIKE '%Success%') AND (DATE between '04-06-2010' and '20-03-2017') \
GROUP BY "Landing _Outcome" ORDER BY ("Landing _Outcome") DESC
```


Section 3
Launch Sites
Proximities Analysis

Folium - All Launch Sites On A Map

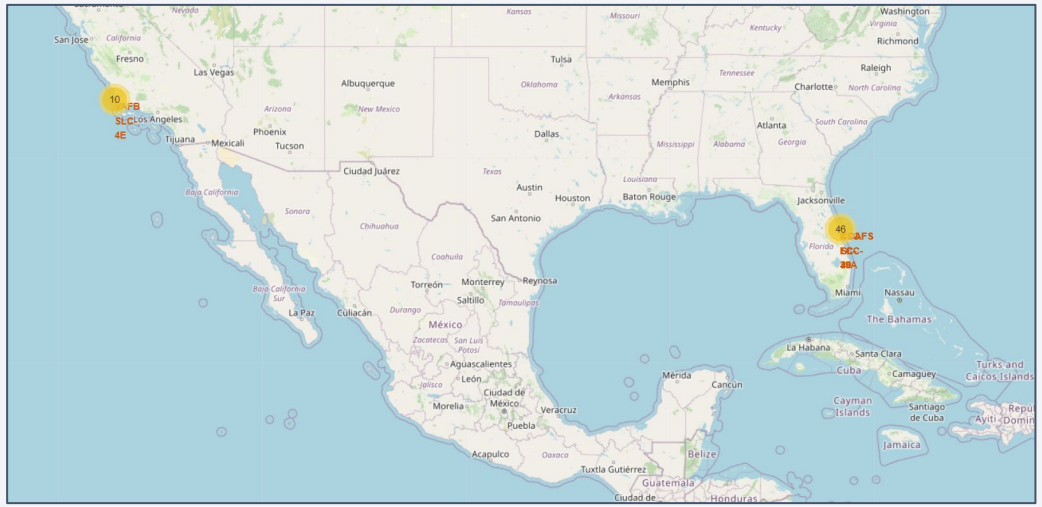
Screenshot to include all launch sites' location markers on a global map

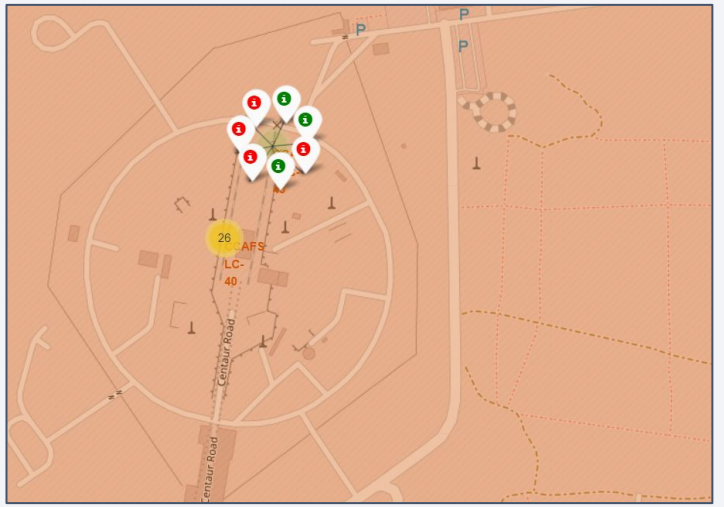


Some launch sites are in very close proximity to the coast and others in proximity to the equator line.

Folium - Success / Failed Launches For Each Site On The Map

Screenshot to show the color-labeled launch outcomes on the map

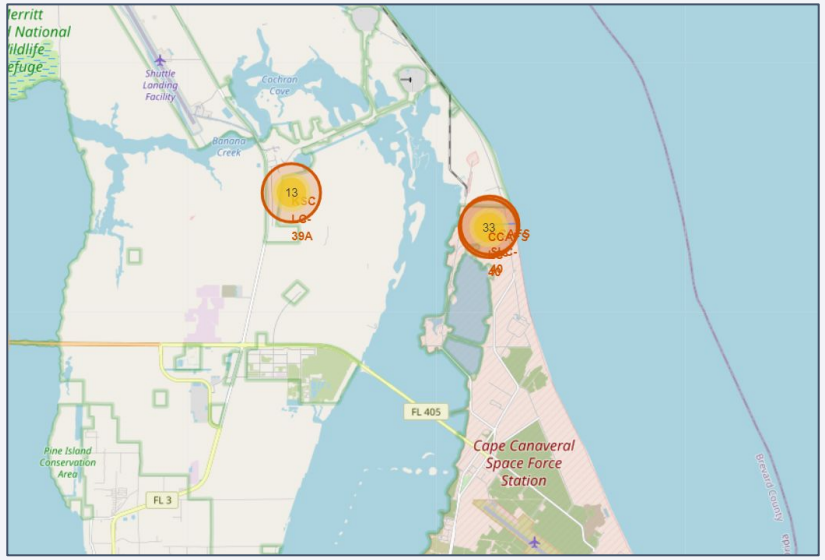




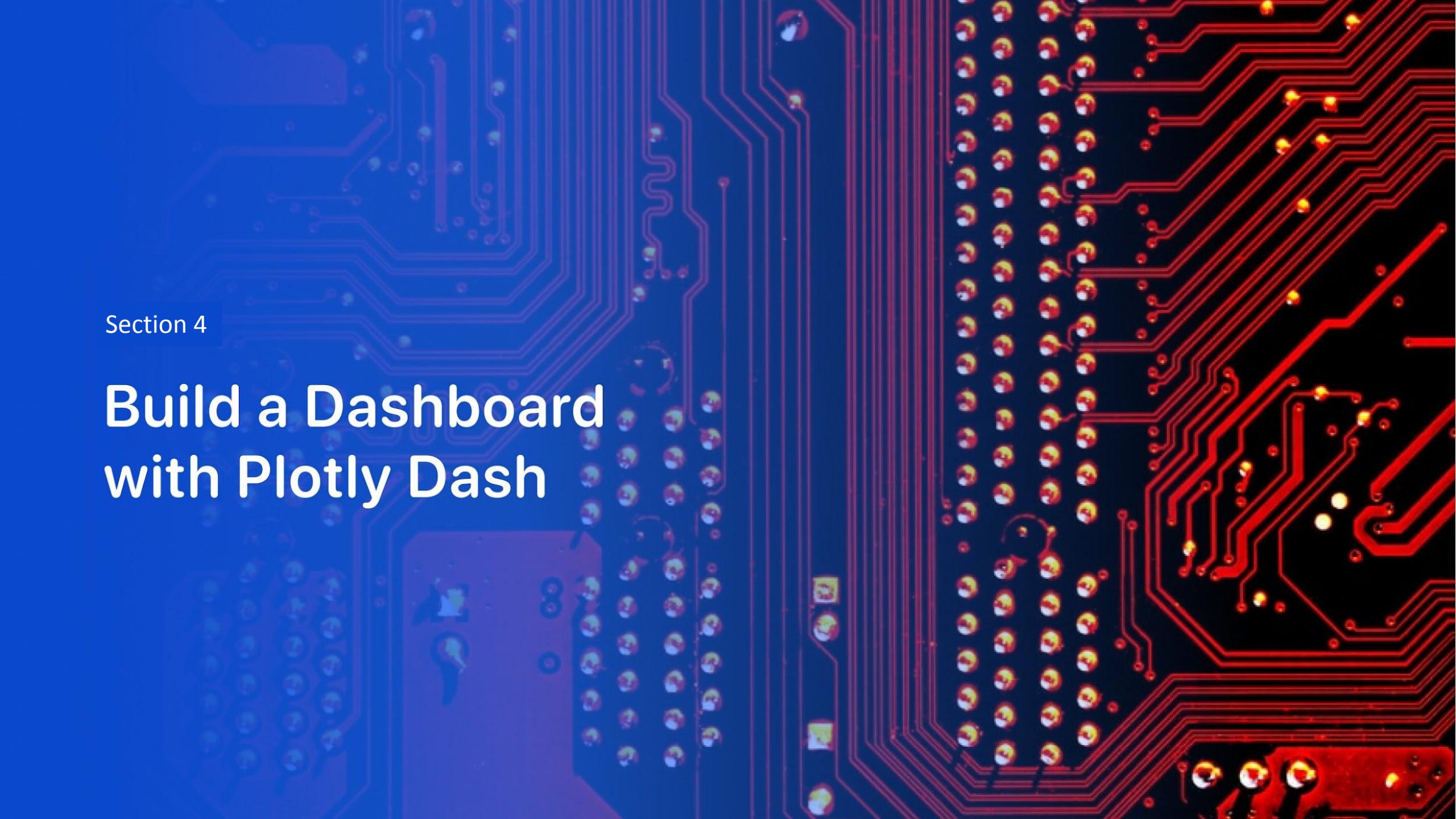
Using the color-labeled markers in the clusters, we can easily identify which launch sites have relatively high success rates.

Folium - Distances between a launch site to its proximities

Screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

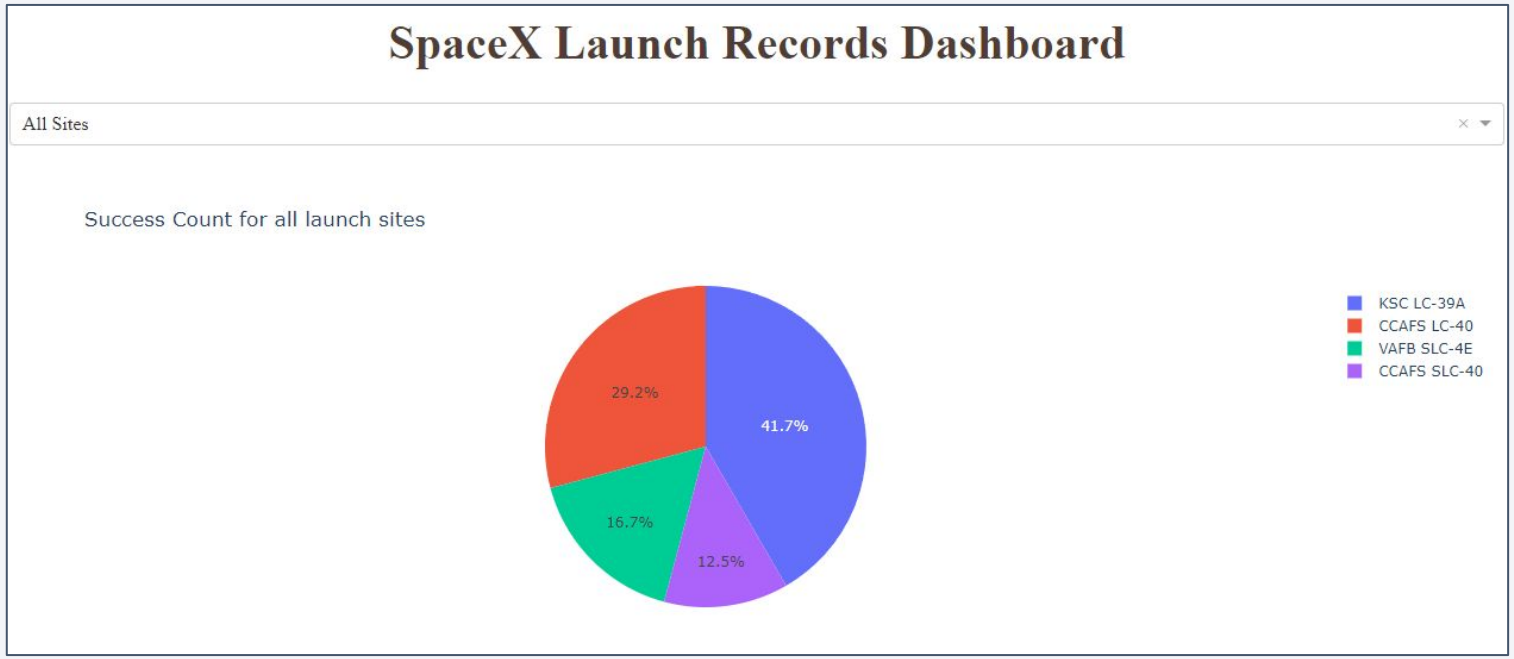


Some launch sites are in close proximity to railways while others are close to highways. Most launch sites are kept a certain distance away from cities for safety reasons.


Section 4
Build a Dashboard with Plotly Dash

Dashboard - Launch success count for all sites

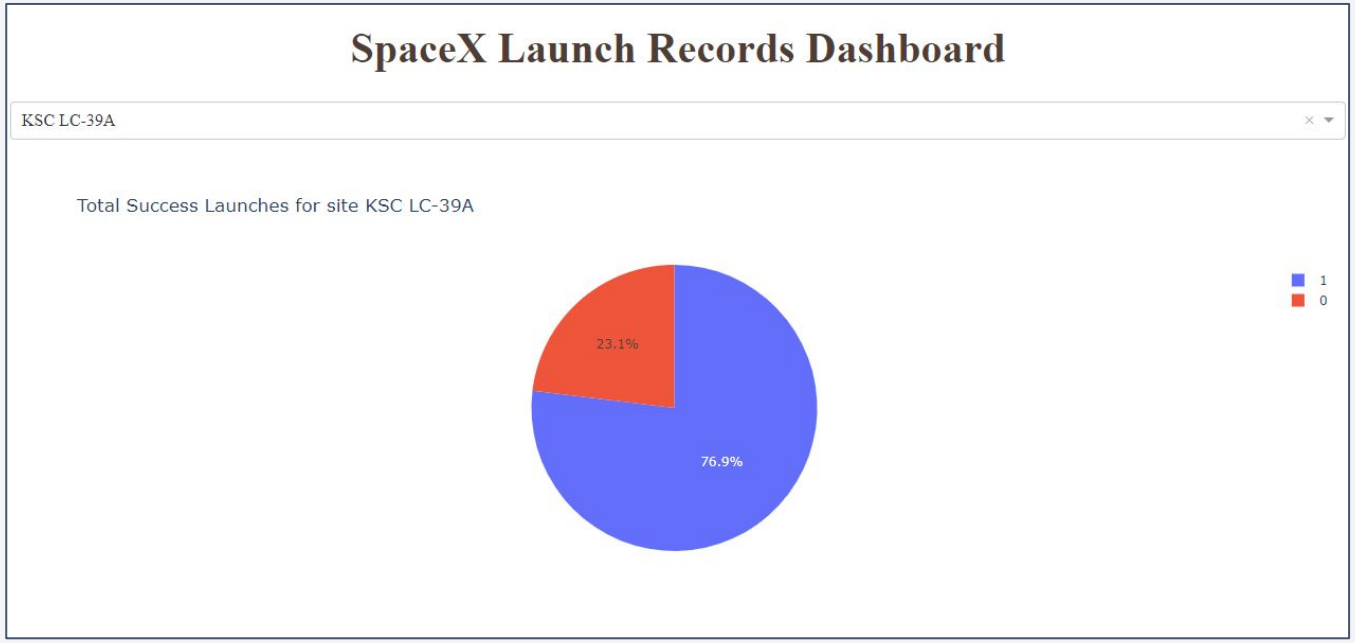
Screenshot of launch success count for all sites, in a piechart



KSC LC-39A had the highest number of successful launches out of all the launch sites.

Dashboard - Launch site with the highest launch success ratio

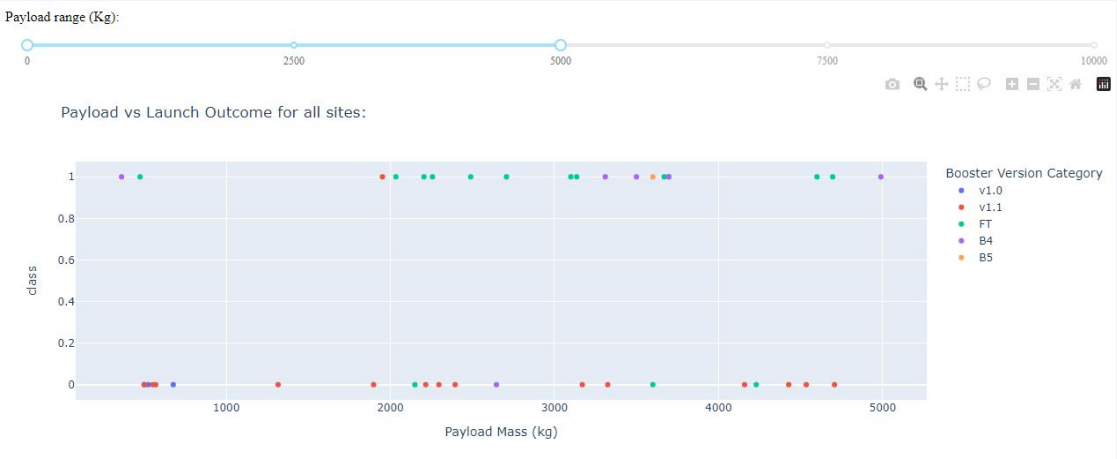
The screenshot of the piechart for the launch site with highest launch success ratio



76.9% Success Rate and 23.1% Failure Rate

Dashboard - Payload vs. Launch Outcome Scatter Plot For All Sites

Screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



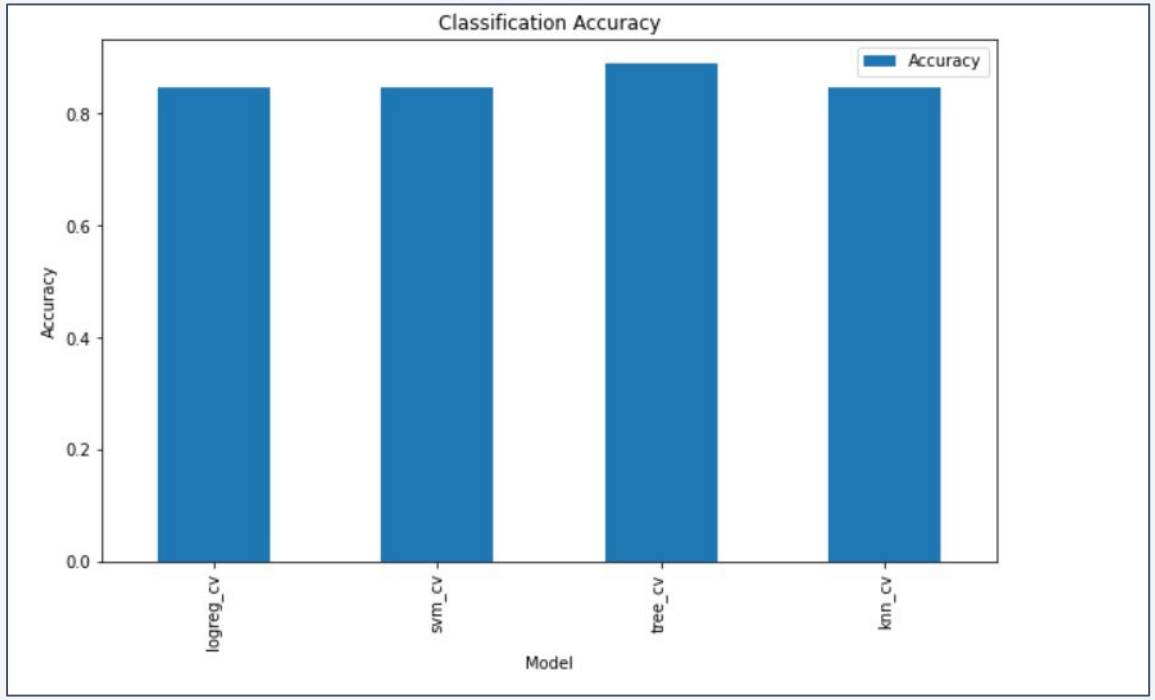
It appears that when the Payload weight is greater than 5000 it is less likely for the Launch Outcome to be positive / successful.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

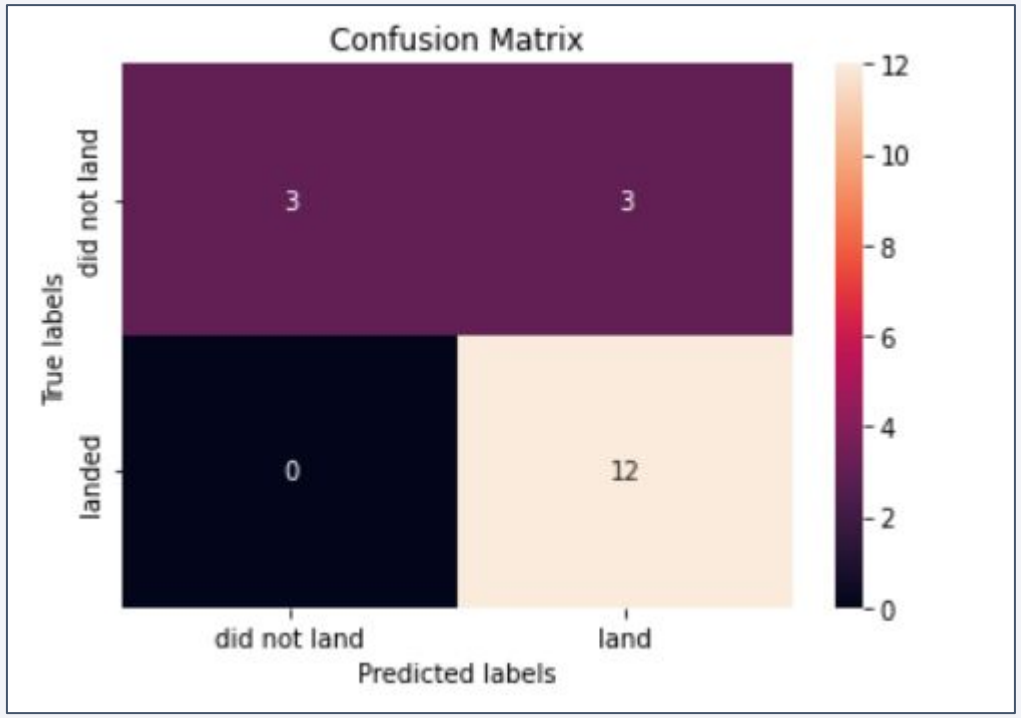
- Visualize the built model accuracy for all built classification models, in a bar chart



The highest classification accuracy belongs to the tree_cv model, which is the Decision Tree Classification model

Confusion Matrix

The confusion matrix of the best performing model with an explanation



We see that the major problem is false positives.

Conclusions

All the models had issues with false positives. This could potentially be fixed by analyzing more data or attempting to use other machine learning methods.

The best model with the highest accuracy out of the models analysed and compared was the Decision Tree Classification model.

It appears like the more launches SpaceX carry out the higher their success rate which suggests they are learning from their mistakes and improving their rockets each time.

KSC LC-39A had the highest number of successful launches out of all the launch sites.

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

  Not Applicable

Thank you!