Winning Space Race with Data Science

Joao Victor Lotfi Barrera
February 21st, 2022

Outline

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

Executive Summary

- The goal of this project is to use SpaceX's data to create a machine learning model which can predict whether the first stage of a rocket launch will successfully land back to Earth.
- The data was collected using SpaceX's API and Web-scraping.
- After cleaning, exploring, and visualizing the data, we were able to observe that succeeding has become more likely each upcoming year due to newer technologies. Also, higher payload masses, choosing specific launch sites, and father orbits all contributed greatly to landing it back.
- Four machine learning models were tested. Logistic Regression, Support Vector Machine, K Nearest Neighbours, and Decision Tree Classifier. They produced similar results, achieving an estimate of 83.33% of accuracy rate.

Introduction

- Context:
  - Commercial Space Age is now here. With the latest technologies, the space industry has become powerful as it has never been. However, costs of launch remain a key barrier for scalability.
  - SpaceX has best pricing ($62 million vs ~$165 million) because its technology allows for the possible recovery of the first stage piece, which can be later reused.
- The problem:
  - How can we use past data from SpaceX to boost a new competitor called SpaceY, which could predict which factors influence the success rate of recovery?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Accessing the JSON result from SpaceX's API, and using beatifulsoup4 library to web-scrape data from Wikipedia's tables.
- Perform data wrangling
  - Reducing the landing column class labels to 0's and 1's, which mean failed and success, respectively.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardizing values and using one-hot encoding to better utilize categorical data, splitting dataset into train/test sets, using 4 model types (SVM, Tree, KNN, and LR), tuning them with GridSearchCV, using r^2 to get accuracy rates, and comparing results in a combined table.

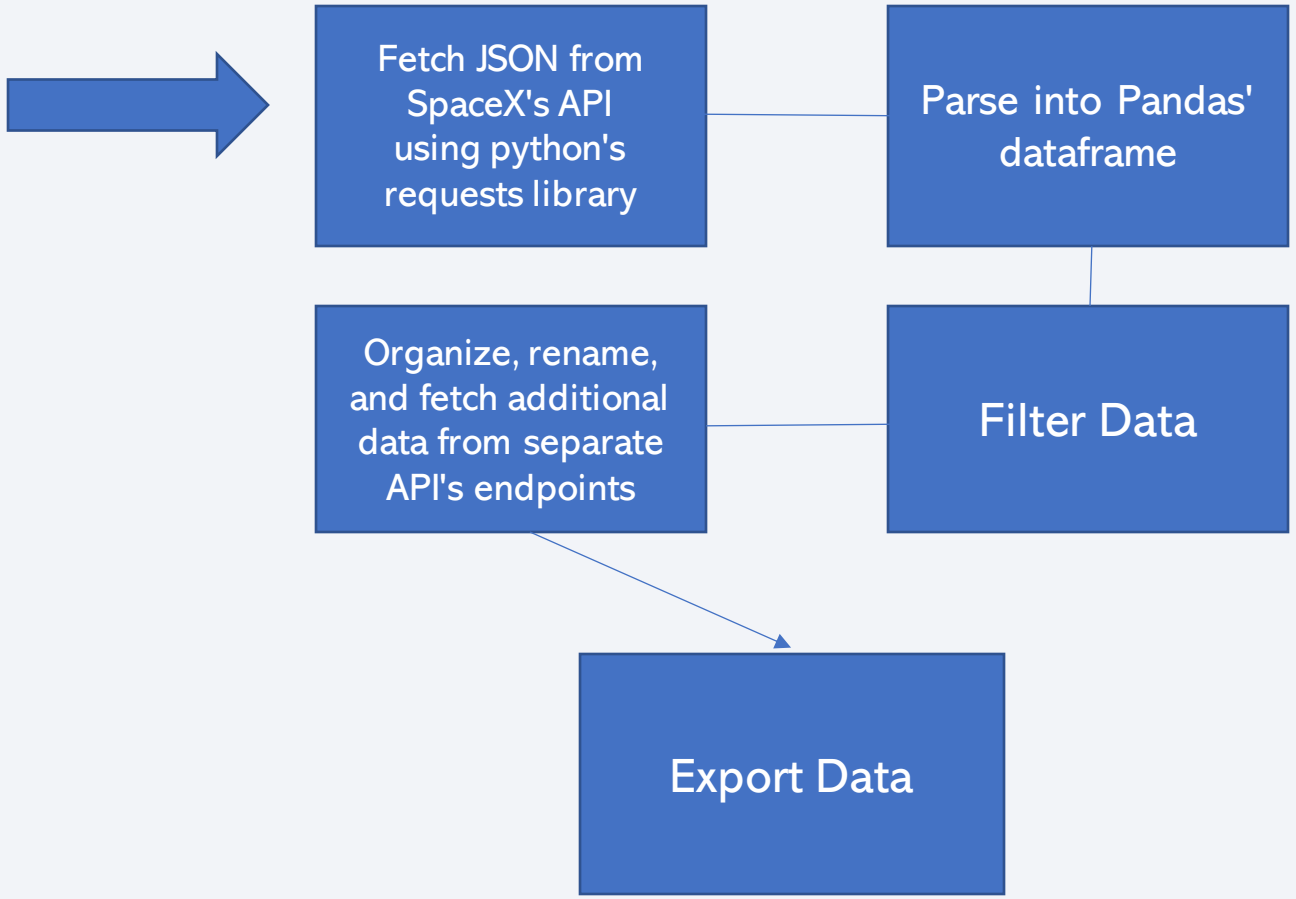
Data Collection

- Data collection originated from 2 paths:
  - SpaceX's API
  - Wikipedia's page for SpaceX's launches.

- The following slides will present flow charts for each of the two processes.

Data Collection – SpaceX API

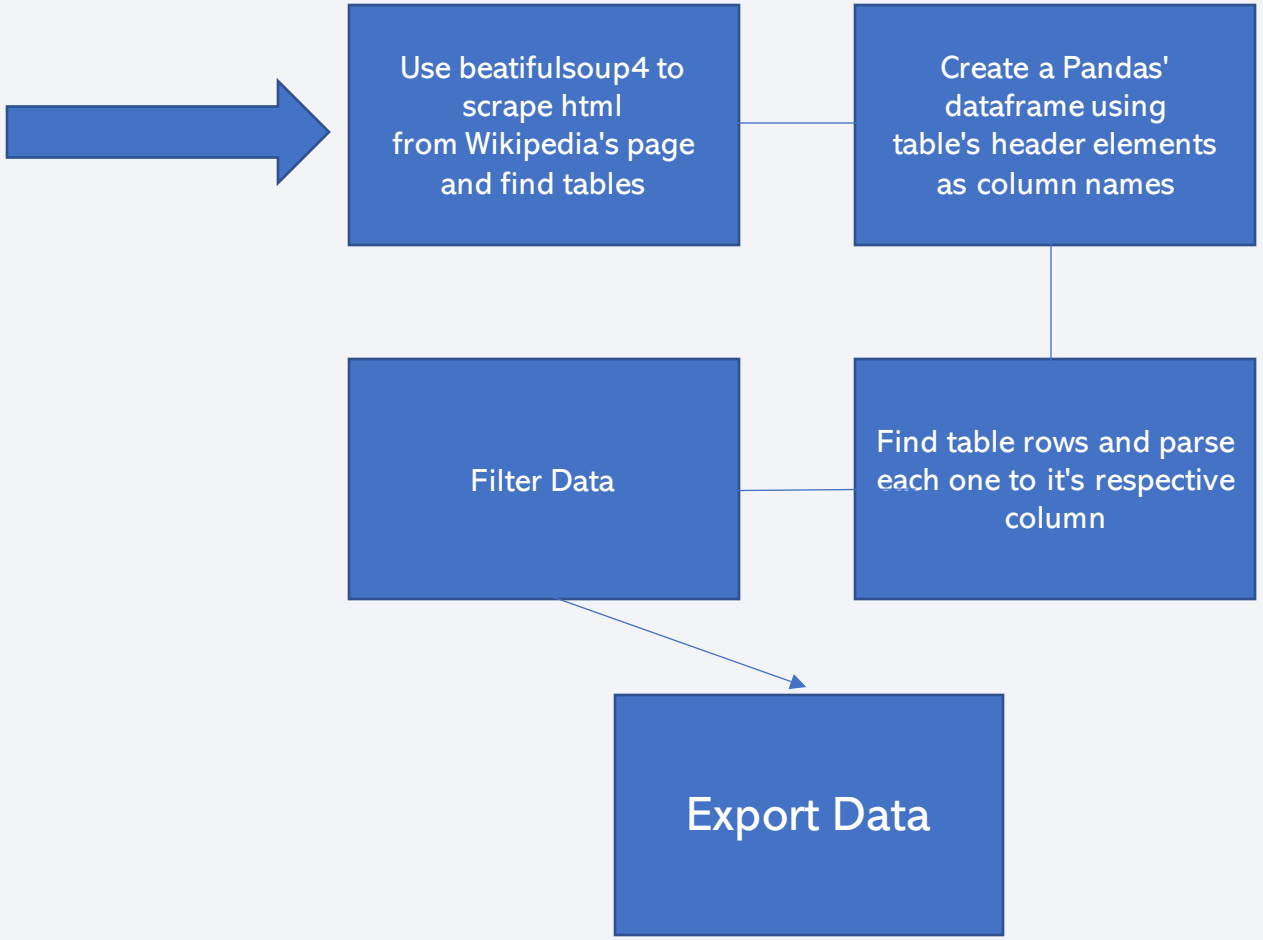
- Here you can visualize how the process of getting data from the API went.



- Here you can find the GitHub URL for the completed notebook containing SpaceX's API work.

Data Collection - Scraping

- Here you can visualize how the process of scrapping data from the html of a page went.

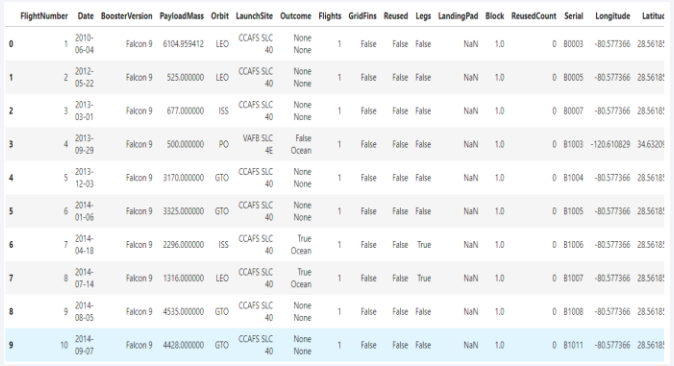


- Here you can find the GitHub URL for the completed notebook containing Wikipedia's table web-scraping work.

Data Wrangling

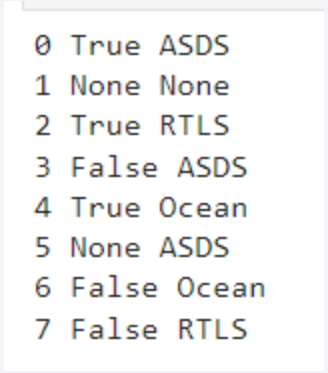
- Used Exploratory Data Analysis to find patterns in the dataset and define labels for the outcome.
- The dataset had a column containing categorical data for all the outcomes. For example, "True ASDS" meant it was a success and it landed on a drone ship. Knowing that, we could reduce all the categorical data to 0's (fail) and 1's (success).

Load Dataset



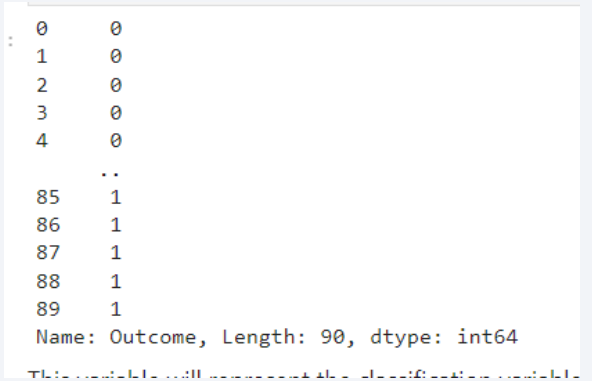
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.5618
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.5618
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.5618
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.6320
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.5618
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1005	-80.577366	28.5618
6	7	2014-04-18	Falcon 9	2296.000000	ISS	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1006	-80.577366	28.5618
7	8	2014-07-14	Falcon 9	1316.000000	LEO	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1007	-80.577366	28.5618
8	9	2014-08-05	Falcon 9	4535.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1008	-80.577366	28.5618
9	10	2014-09-07	Falcon 9	4428.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1011	-80.577366	28.5618

Find unique outcomes



```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 None ASDS
6 False Ocean
7 False RTLS
```

Transform data



```
0     0
1     0
2     0
3     0
4     0
     ..
85    1
86    1
87    1
88    1
89    1
Name: Outcome, Length: 90, dtype: int64
```

- Here you can find the completed notebook for all the data wrangling work.

EDA with Data Visualization

- To better understand the patterns hidden in this stream of data, we explored a few different charts. Here you can see which ones and what for:
- Scatter Plot:
  - Flight Number vs Payload Mass (hue Class)
  - Flight Number vs Launch Site (hue Class)
  - Launch Site vs Payload Mass (hue Class)
  - Flight Number vs Orbit Type (hue Class)
  - Payload Mass vs Orbit Type (hue Class)
- Bar Chart:
  - Orbits vs Avg Success Rate
- Line Chart:
  - Year vs Avg Success Rate

- To view the data visualization process, follow this link.

EDA with SQL

- To solidify our knowledge on the dataset, we performed a few SQL queries to the dataset to filter and order by quite specific data. They are as follows:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster versions which have carried the maximum payload mass
9. List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
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

- To see how we used SQL queries to explore the dataset, visit this link.

Build an Interactive Map with Folium

- To further explore the data, we built an interactive map using Folium, a python library.

- We marked all launch sites, then added a marker cluster to each one of them. These marker clusters contained green markers for successful missions, and red markers for failed missions, for its correspondent launch site. With this, we can better visualize how each launch site did in terms of performance.

- Then, we marked points of interest and calculated the distance between it and a launch site. These points of interest were closest coastline, railway, highway, and city. This way we could observe that launch sites are usually close to coastlines, railways, and highways but are far away from cities.

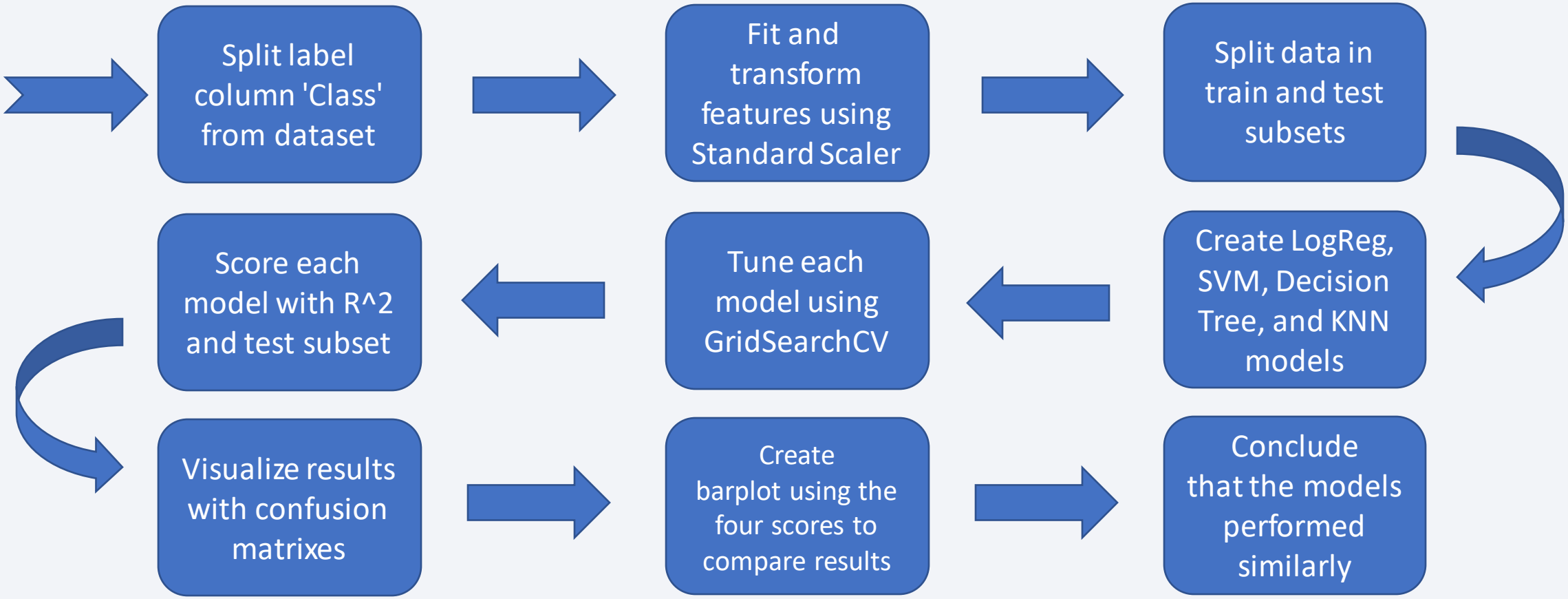
- To see how we created interactive maps with Folium, visit this link.

Build a Dashboard with Plotly Dash

- To consolidate our exploration, we made a dashboard using Ploty and Dash, which are python libraries.
- This dashboard contained 2 graphs and 2 inputs. A change in the input updates the graphs immediately.
- These graphs were a pie chart for successful launches, and a scatter plot for payload mass vs class (hue: booster version category).
- One input was a dropdown menu in which you can select all launch sites or a specific launch site. The other input was a slider for the mass, which one can select the range allowed for mass, in Kilograms.
- This helped answer the following questions:
  - Which site has the largest successful launches? KSC LC-39A with 10.
  - Which site has the highest launch success rate? KSC LC-39A with 76.9% success.
  - Which payload range(s) has the highest launch success rate? 2000 – 5500 kg.
  - Which payload range(s) has the lowest launch success rate? 0-2000 and 5500 – 7000kg.

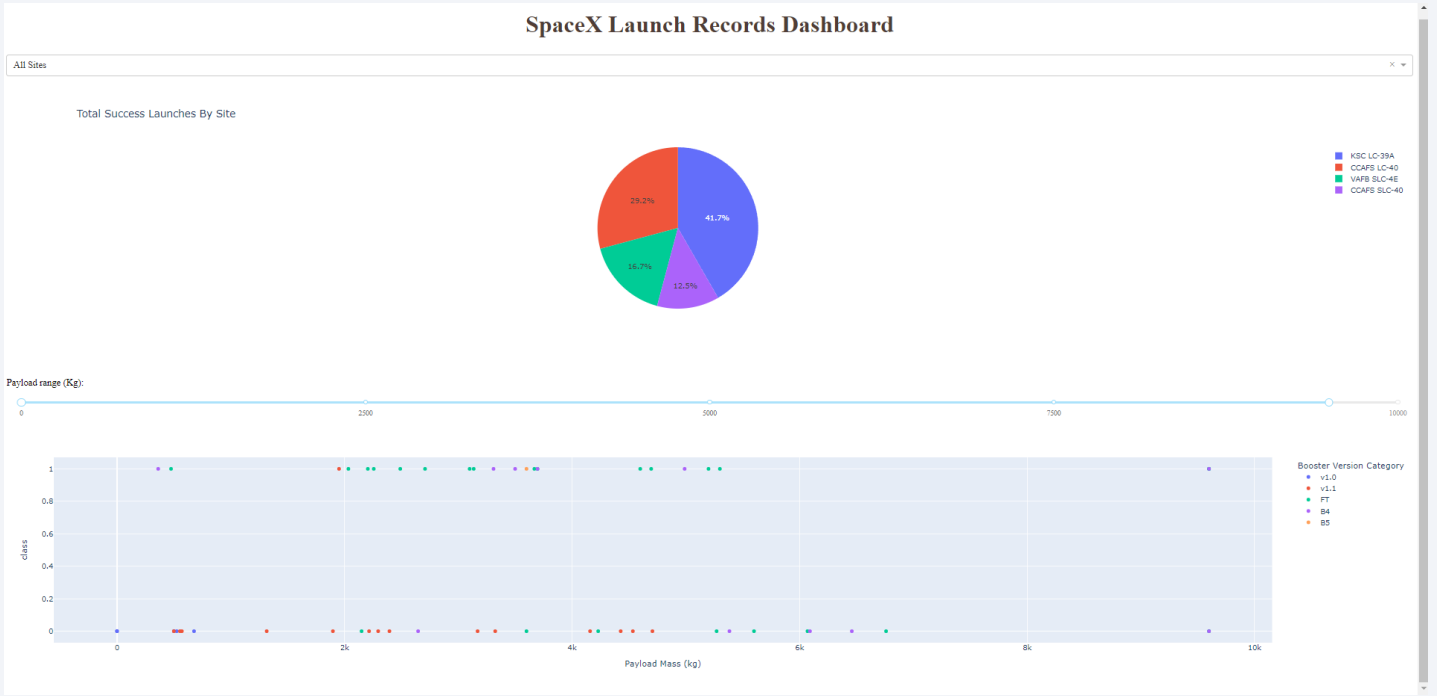
- To see how we built a dashboard using Plotly Dash, visit this link.

Predictive Analysis (Classification)



- To see the predictive analysis and model building process, follow this link.

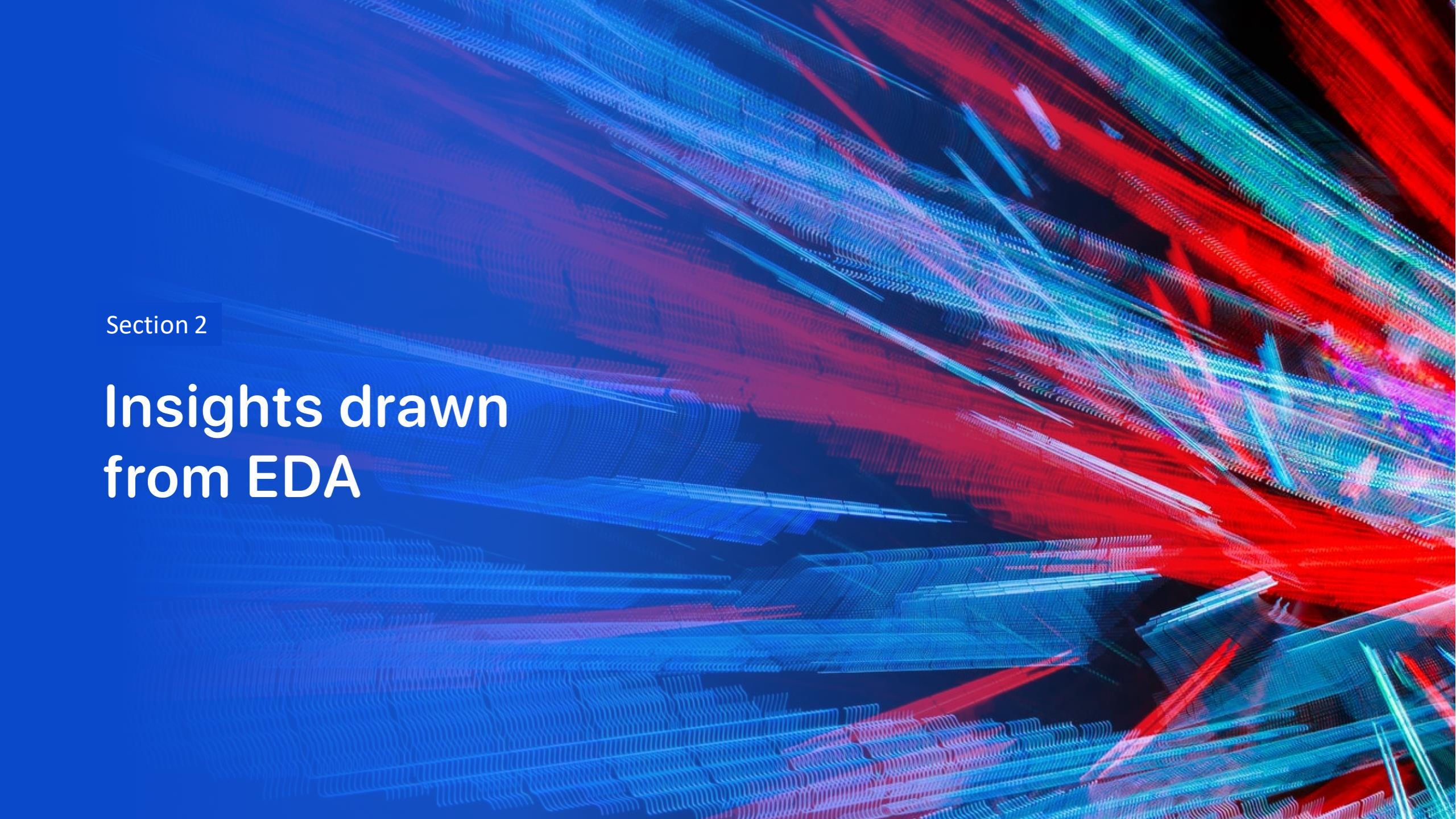
Results



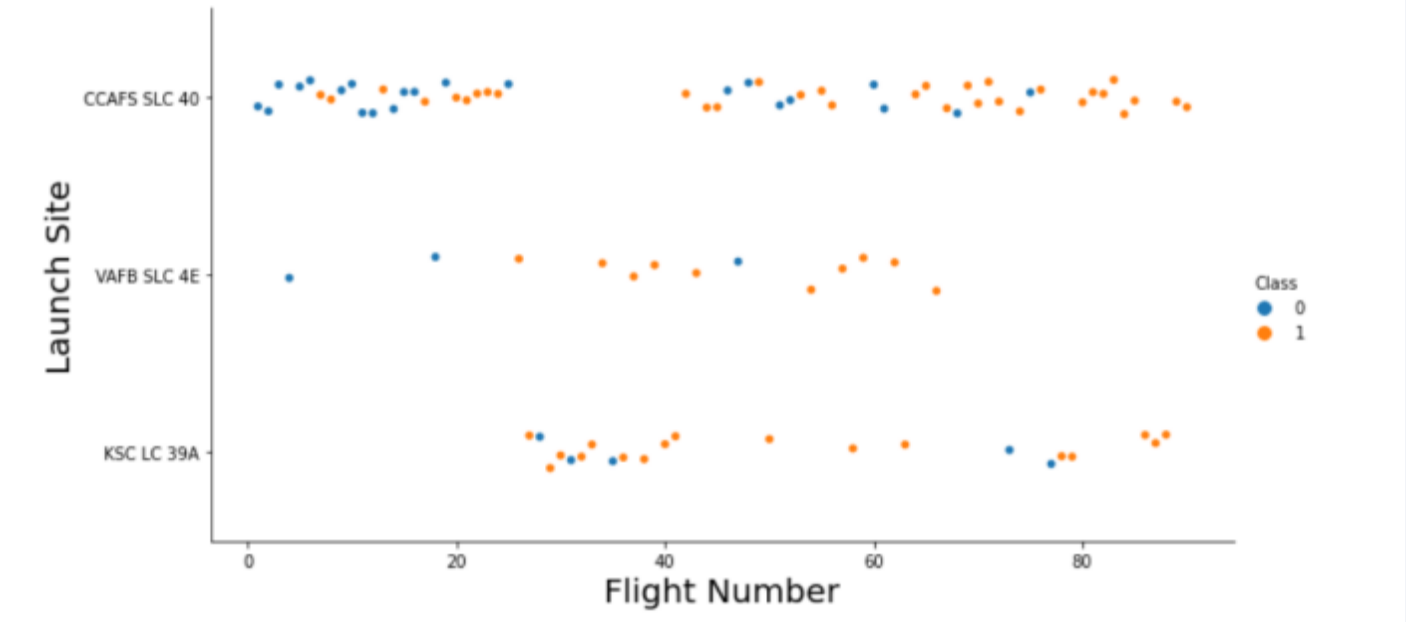
- In the picture above you can see a demo for our Plotly/Dash dashboard.
- The result for the predictive analysis process was that the models performed almost the same, with a success rate of 83.33%
- The following slides will show you a more in-depth analysis to each portion of the exploratory data analysis process.

Section 2

Insights drawn from EDA

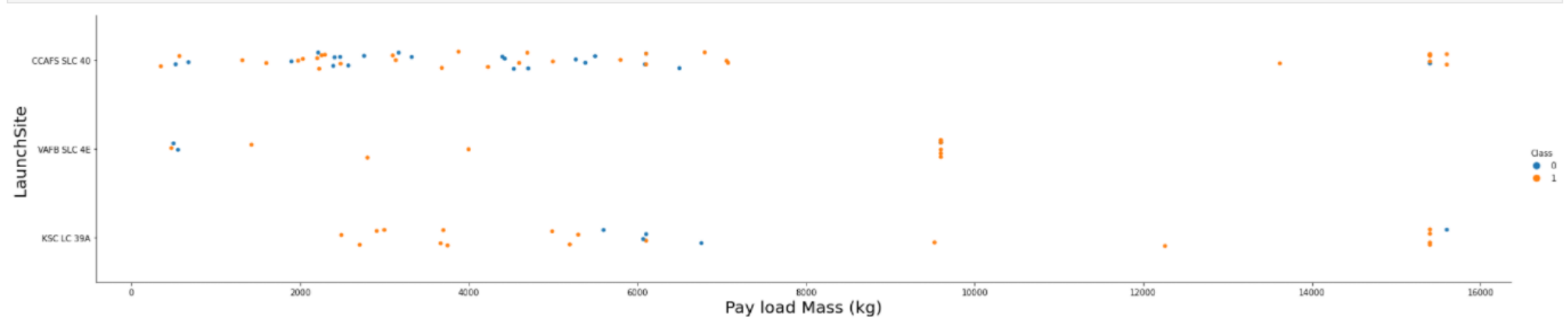


Flight Number vs. Launch Site



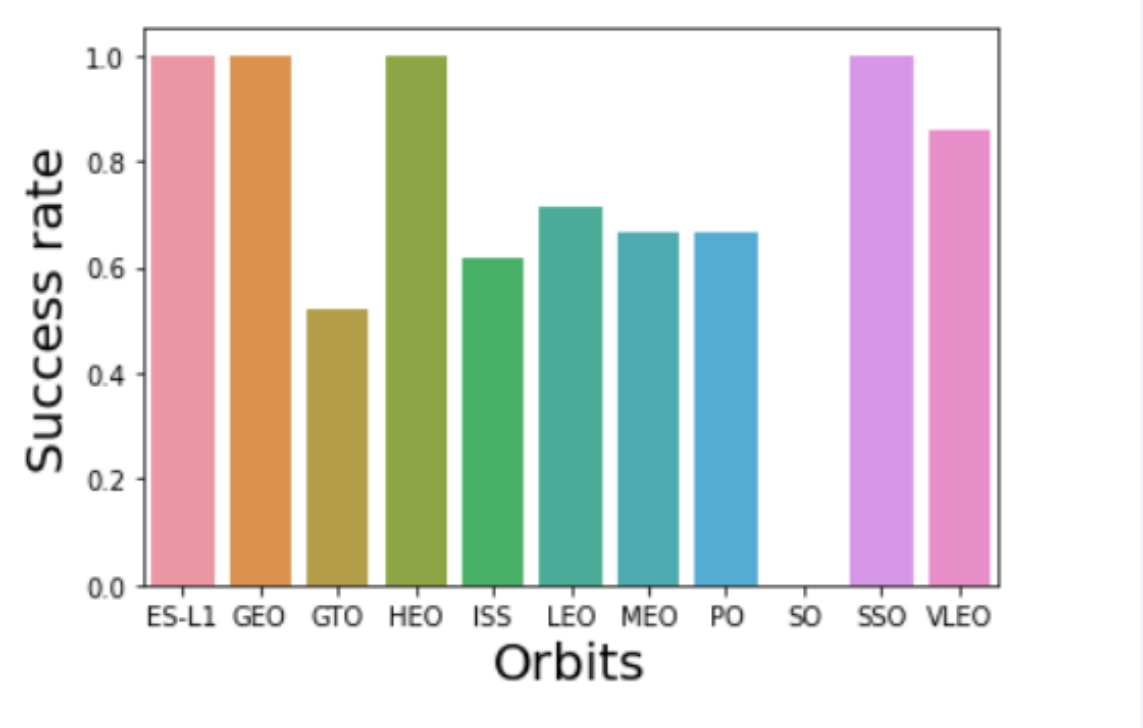
- Success rates (Class 1) increases as the number of flights increase
- For launch site ‘KSC LC 39A’, it takes at least around 25 launches before a first successful launch

Payload vs. Launch Site



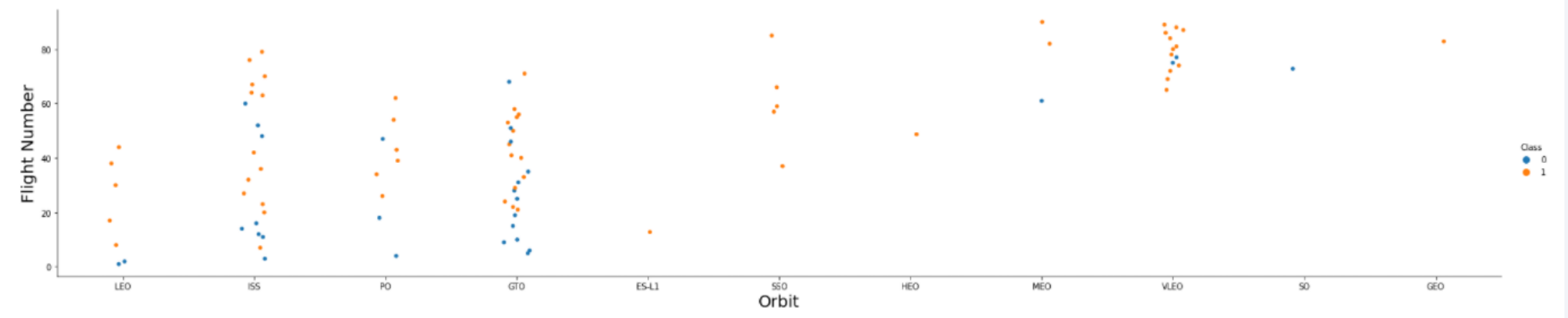
- Most payloads have a mass between 0 and 6000kg.
- Above about 9000 kg we can see that the ratio from successful landings to unsuccessful landings increases in comparison to lower payload masses

Success Rate vs. Orbit Type



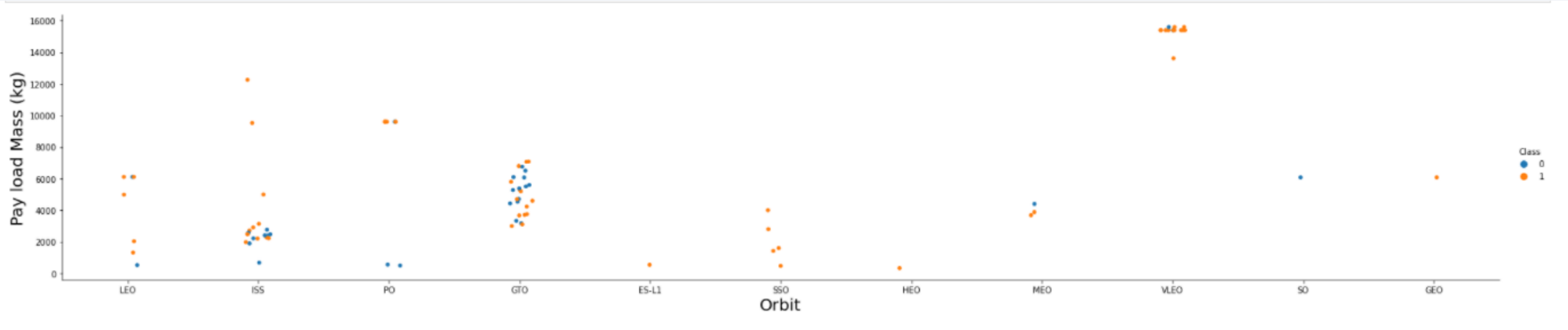
- ES-L1, GEO, HEO, SSO and VLEO have the highest success rates

Flight Number vs. Orbit Type



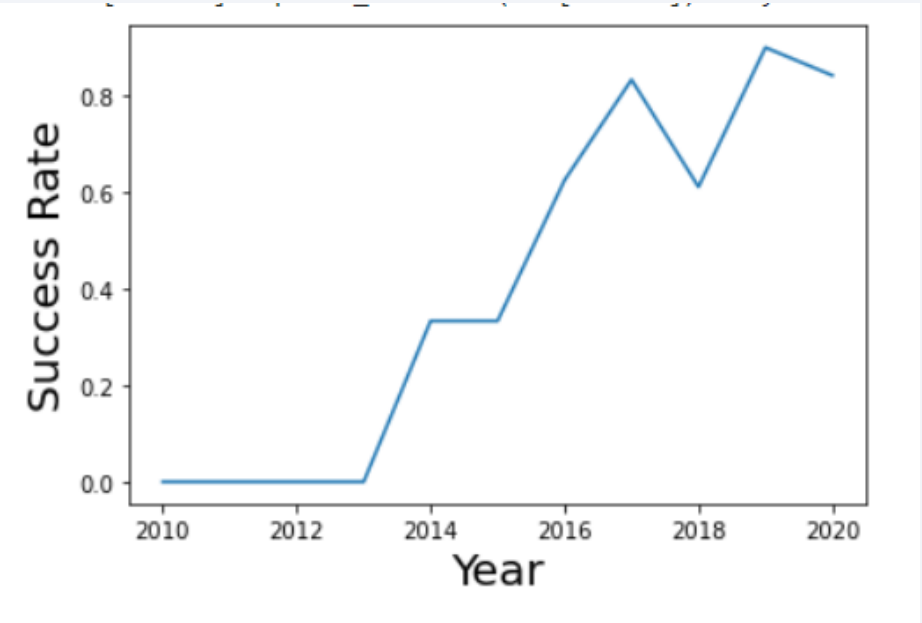
- In the LEO orbit the success appears related to the number of flights
- On the other hand, there seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well because succesful and unsucessful launches happened at the same payload mass range.

Launch Success Yearly Trend



- The figure shows how success rates have gone up significantly each year.
- In 2019 there was a dramatic decrease in success rates (by about 20%).

All Launch Site Names

- In order to display the names of the unique launch sites in the space mission, input the following query:

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL
```

- The result is this:

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

- This query fetches all distinct items from the launch_site column.

Launch Site Names Begin with 'CCA'

- In order to display 5 records where launch sites begin with the string 'CCA', input the following query:

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

- This query selectes everything from all rows where the start of each element in the column launch_site is "CCA", then limits results by the 5 first elements.

- The result is this:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- In order to display the total payload mass carried by boosters launched by NASA (CRS), input the following query:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

- This query sums all values from the payload_mass column where the costumer is NASA.

- The result is this:

Average Payload Mass by F9 v1.1

- In order to display average payload mass carried by booster version F9 v1.1, input the following query:

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
```

- This query averages the entire payload_mass column where the booster version is equal to "F9 v1.1".

- The result is this:

First Successful Ground Landing Date

- In order to display the date when the first successful landing outcome in ground pad was achieved, input the following query:

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%' LIMIT 1
```

- This query selects the smallest date from the date column where the values in the column landing_outcome start with "Success".

- The result is this:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- In order to list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000, input the following query:

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 600
```

- The result is this:

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

- This query selects the booster_version column where the values in the landing_outcome column are equal to "Success (drone ship)" AND the values in the payload_mass column are between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

- In order to list the total number of successful and failure mission outcomes, input the following query:

```
%sql SELECT (mission_outcome), COUNT(*) AS "count" from SPACEXTBL GROUP BY mission_outcome;
```

- The result is this:

mission_outcome	count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- This query selects the mission_outcome column and creates a new column count which contains the amount of rows for each unique value in the mission_outcome column.

Boosters Carried Maximum Payload

- In order to list the names of the booster versions which have carried the maximum payload mass, input the following query:

```
%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

- The result is this:

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- This query finds all the distinct values in the booster version column where the value in the payload mass column is equal to the max value found also in the payload mass column.

2015 Launch Records

- In order to list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015, input the following query:

```
%sql SELECT "Landing _Outcome", BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE "Landing _Outcome" = 'Failure (drone ship)' and YEAR(DATE) = 2015
```

- The result is this:

Landing _Outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- This query selects the items for the landing outcome, booster version, and launch site columns where the value in the landing outcome column is "failure (drop ship)" AND the value for the year in the date column is equal to 2015.

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

- In order to rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order, input the following query:

```
%%sql SELECT "Landing _Outcome", COUNT(*) AS RANK
FROM SPACEXTBL
WHERE DATE > '2010-06-04' and DATE < '2017-03-20'
GROUP BY "Landing _Outcome"
ORDER BY RANK DESC
```

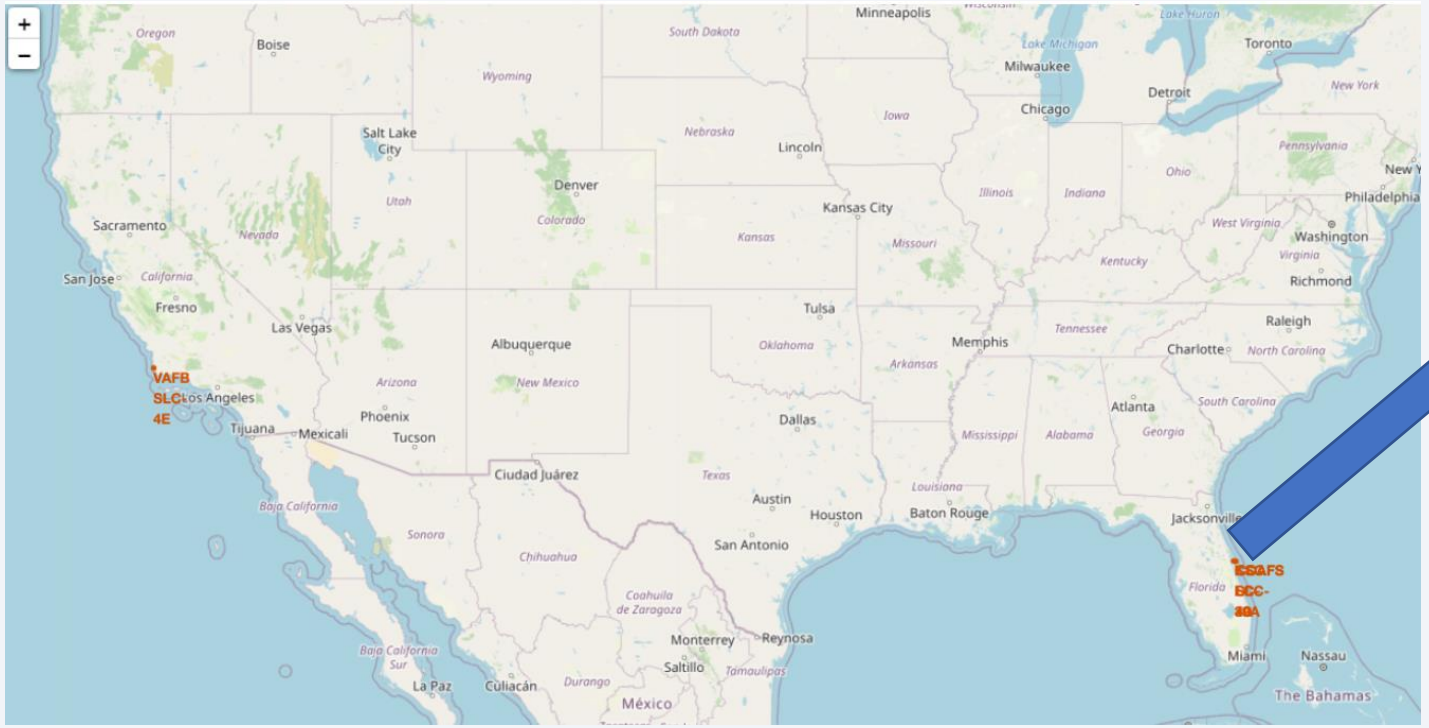
- The result is this:

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

- This query selects the landing outcome column and creates a new column RANK which contains the amount of rows for each unique value in the landing outcome column where the values in the date column are between 2010-06-04 and 2017-03-20, then orders it in descending order.

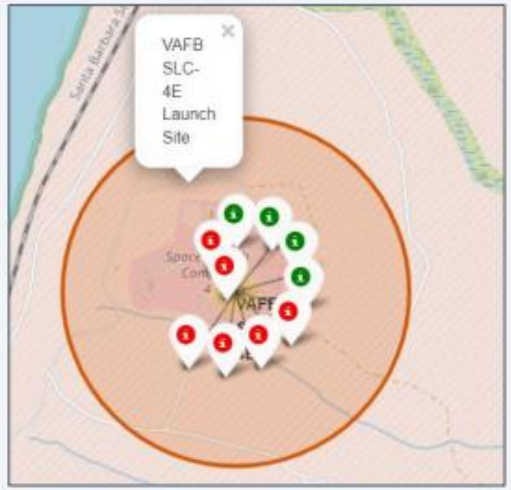

Section 3
Launch Sites
Proximities Analysis

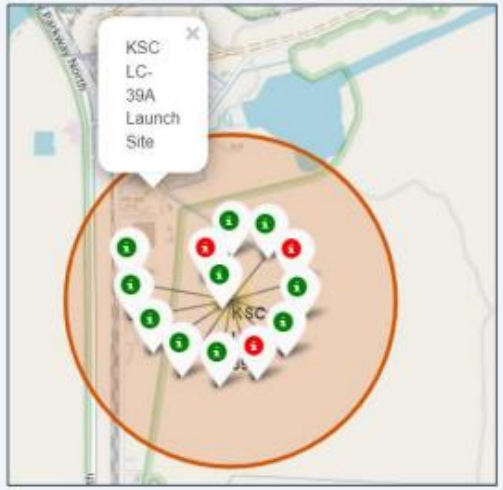
Launch Sites marked on the map

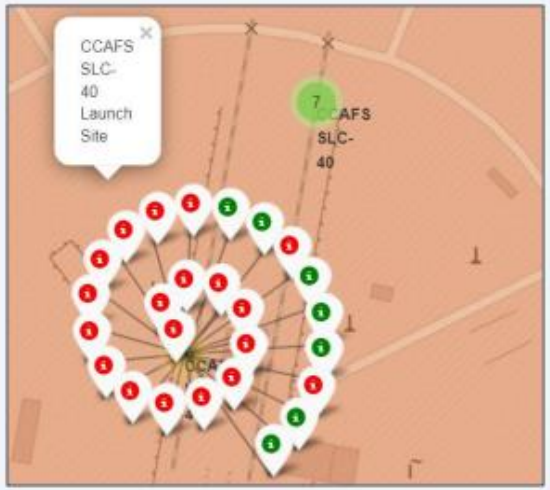


- We can see on the map that both launch sites are in the US. However, only one is located in the west coast, while the other three are located in the east coast.
- We also observed that CCAFS SLC-40 and CCAFS LC-40 are located right next to each other.

Marker Clusters with Success/Failed

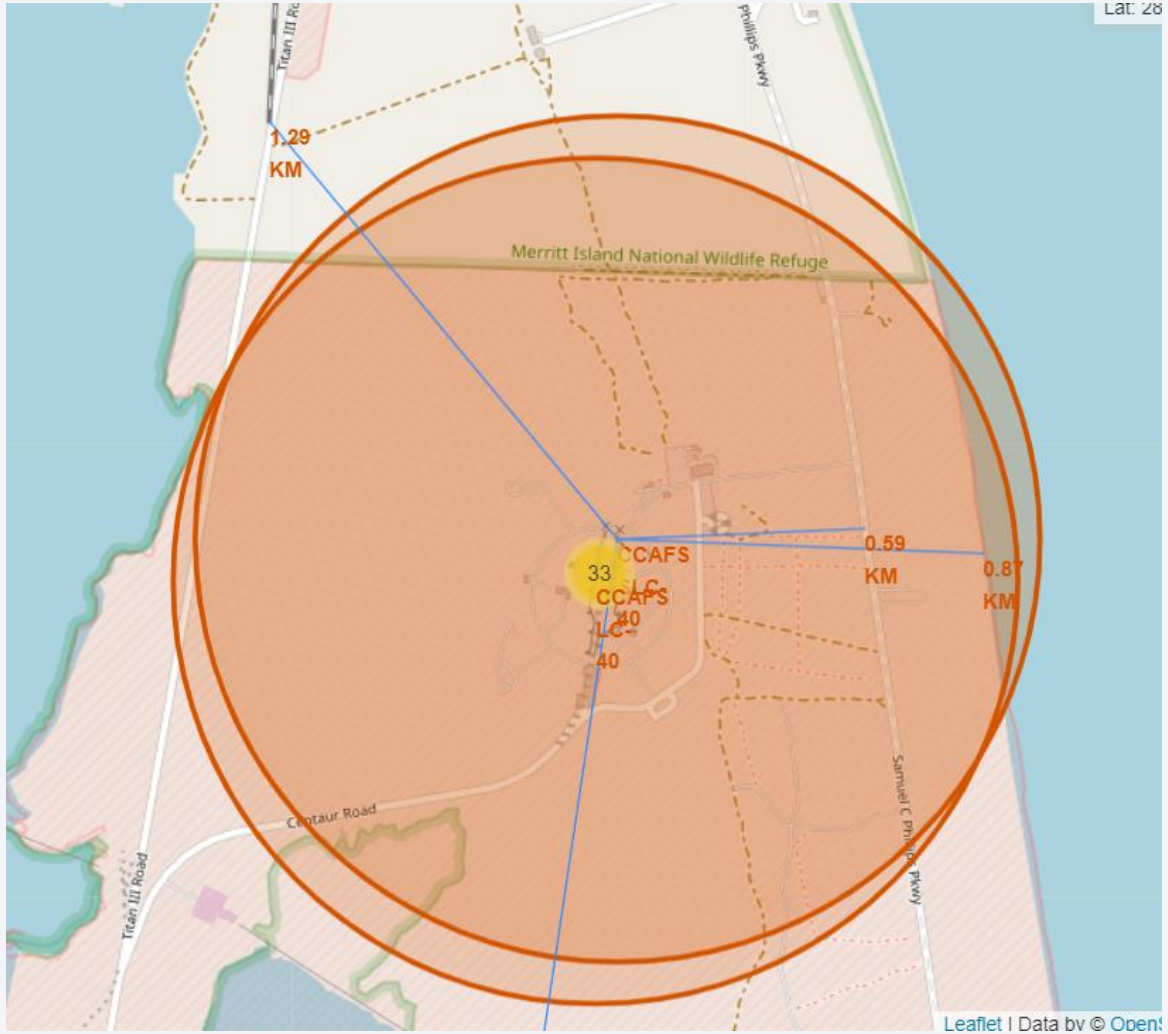




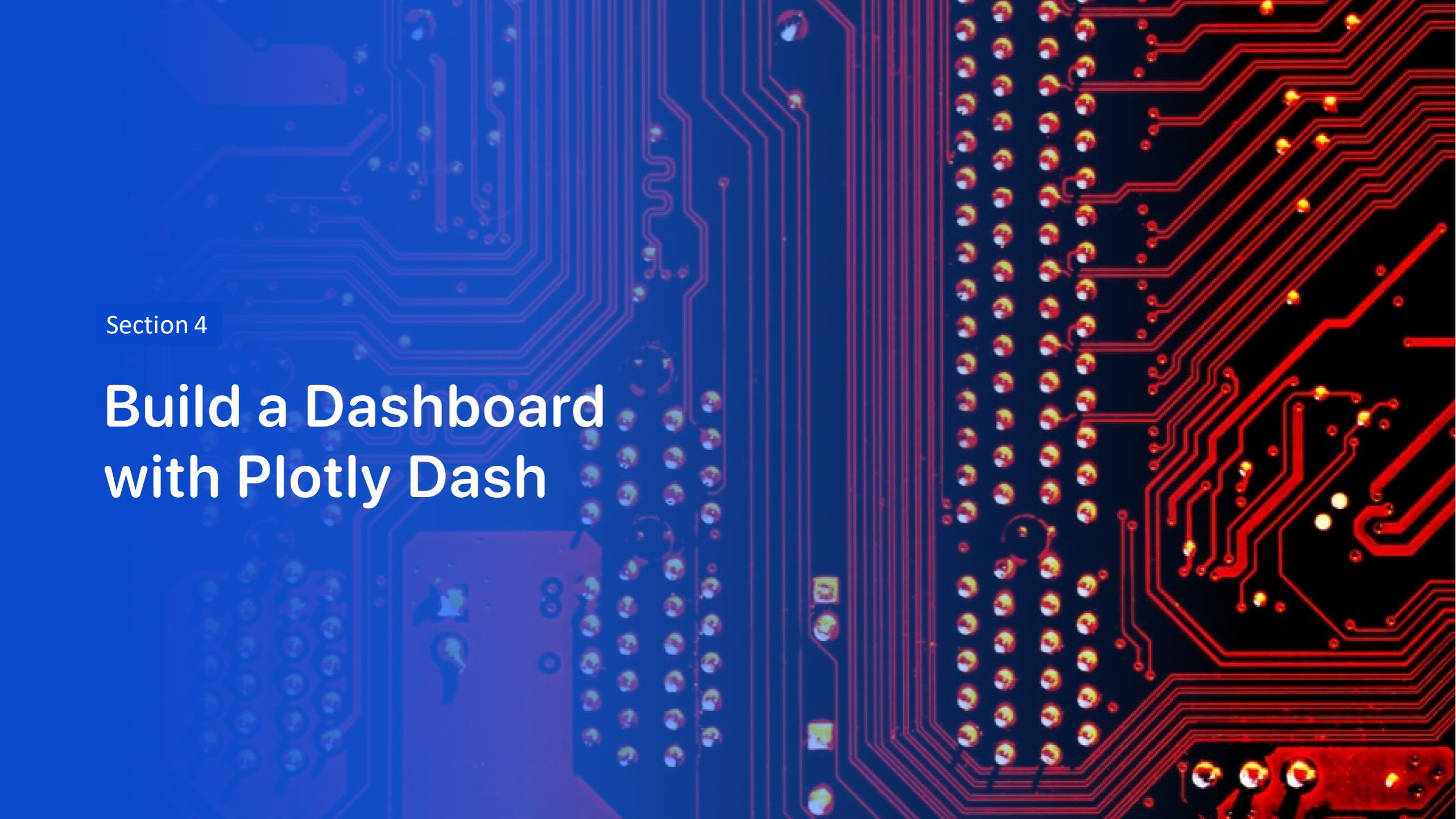


- By creating marker groups with color coding, we can observe that KSC LC-39A had the highest success to fail rate
- On the other hand, CCAFS SLC-40 had the greatest amount of data points but the lowest success to fail rate.

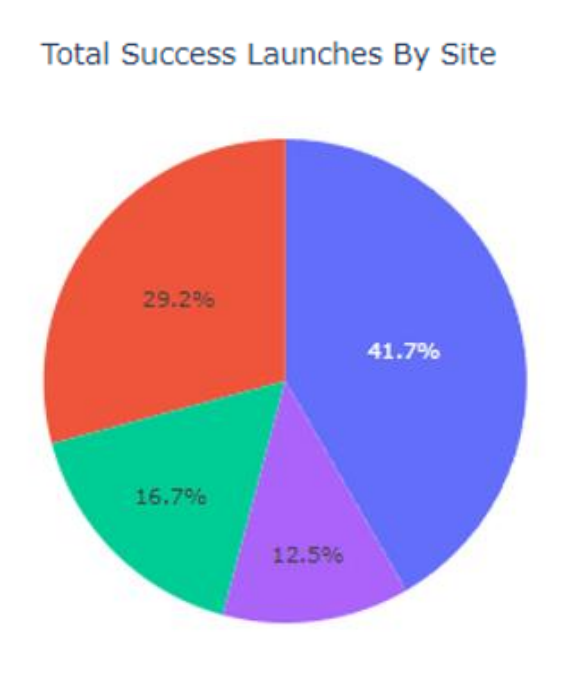
Distance to Points of Interest



- Here we can clearly observe that launch sites are usually close to coast lines, railways, and highways but are quite far from cities.

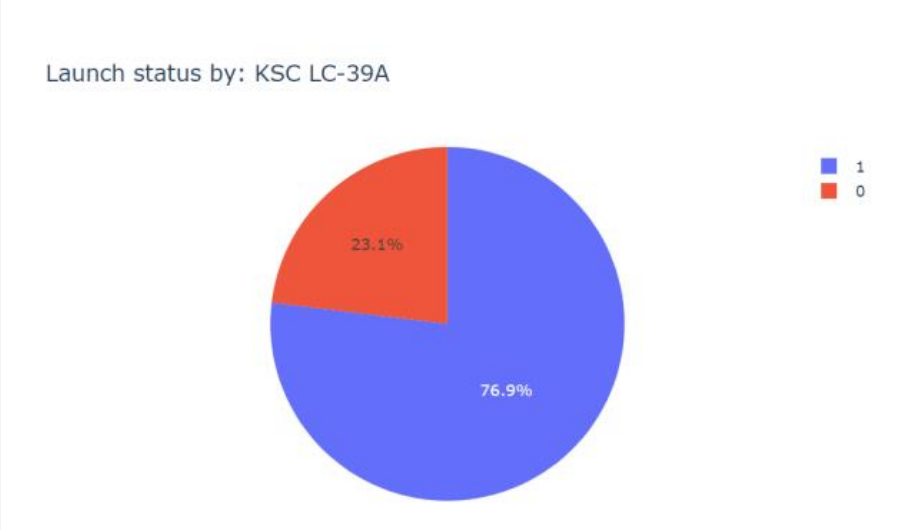

Section 4
Build a Dashboard with Plotly Dash

Launch Success for All Launch Sites



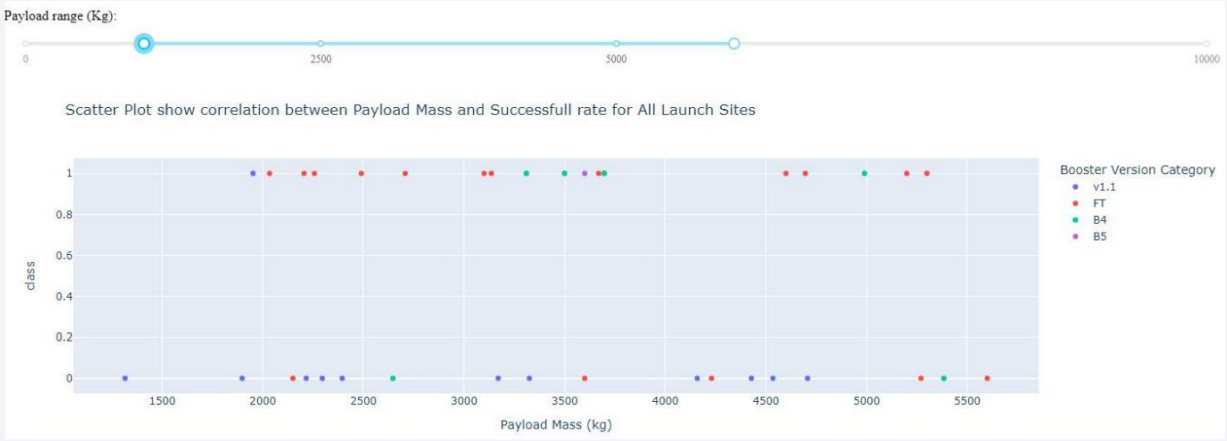
- The data for total successful launches by site is as follows:
  - KSC LC-39A: 41.7%
  - CCAFS LC-40: 29.2%
  - VAFB SLC-4E: 16.7%
  - CCAFS SLC-40: 12.5%
- As we can see, KSC LC-39A had the highest amount of successful launches among all launch sites.

Success Rate Distribution for Highest Success Launch Site



- For this graph, you can see KSC LC-39A is detaily analyzed.
- From all launches in this site, 76.9% succeed and 23.1% failed to land the booster.

Mass vs Class Scatterplot for different Mass Ranges

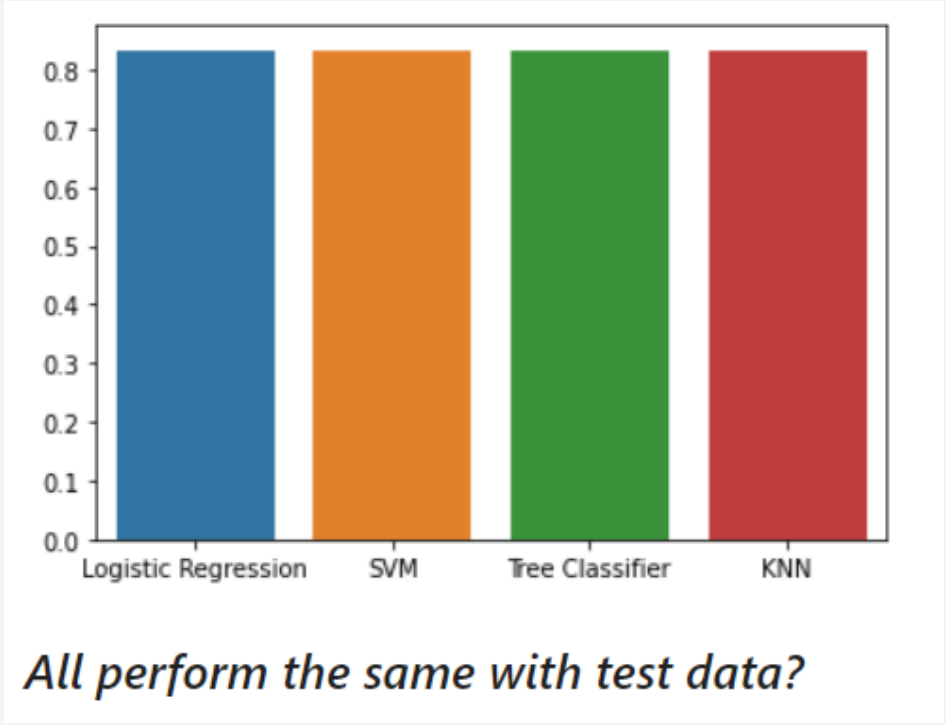


- Here we can see the correlation between payload mass and success/fail.
- Most succesful launches have payload masses from 2000 to 5500 kg.
- Fails are more likely to occur when the payload mass is less than 2000kg, or above 5500kg.

Section 5

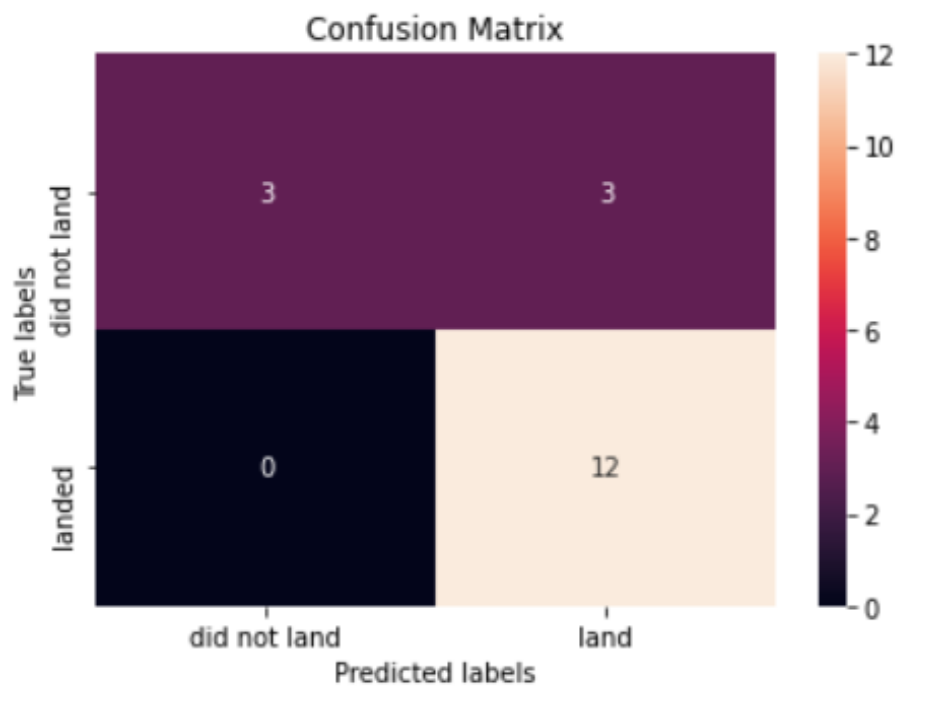
Predictive Analysis (Classification)

Classification Accuracy



- This bar plot demonstrates how each model performed in terms of accuracy.
- As we can see, they all performed equally, with an accuracy rate of 83.33%. Because of that, we can choose any model.

Confusion Matrix



- All models produced equal accuracy rates.
- The picture on the side stands for the confusion matrix created for the tree classifier model (picked randomly).
- Using the matrix we can visualize that for the 18 entries, 15 were predicted correctly and 3 were false positives.

Conclusions

- SpaceY successfully predicts which boosters will land with a small margin of error, and is able to observe which features contribute to this result, making it possible to direct resources to the correct areas. Some of these observations are:
  - As time passes, more and more rockets are launched, technology improves and success rates increase.
  - Launch Sites are strategically chosen far away from cities and close to coast lines.
  - The "KSC LC-39A" launch site performed the best.
  - ES-L1, GEO, HEO, SSO, and VLEO orbits have the highest success rates.
- All tested classification models presented equal results, at 83.33% accuracy rate.
- We should collect more data to improve our model's accuracy and lifespan.

Appendix

- For all notebooks and source code, visit my GitHub repository here.
- Salute to matplotlib, pandas, numpy, plotly, dash, folium, and beautifulsoup4 for being of use in this project.

Thank you!