Winning Space Race with Data Science

Ayush Singh
17th of June 2025

Outline

- Executive Summary
- Introduction
- Methodology
  - EDA with SQL & Visualization
  - Folium Maps
  - Dashboard Plots
- Results
- Conclusion
- Appendix

Executive Summary

- SPACEX is the leader of the most affordable space travel in the USA. The cost of space travel has been significantly reduced because of the possibility to preserve the first stage of rocket launch. By training machine learning models with SPACEX's historical data our results show that using SVM models to predict future outcomes is the most accurate. With predictive outcomes of launches we can understand more about cost.

Introduction

- SPACEX has a goal of making space flights more accessible to all people. Though space travel is expensive.
  - **How can space flight become more accessible?**
  - **What determines the price of a flight?**
- We analyze features of the Falcon 9 launches to understand what factors contribute to landing outcome predictability.
- By training a machine learning model with public information we aim to better understand accurate ways to predict launch success and if the first stage of the rocket can be reused.

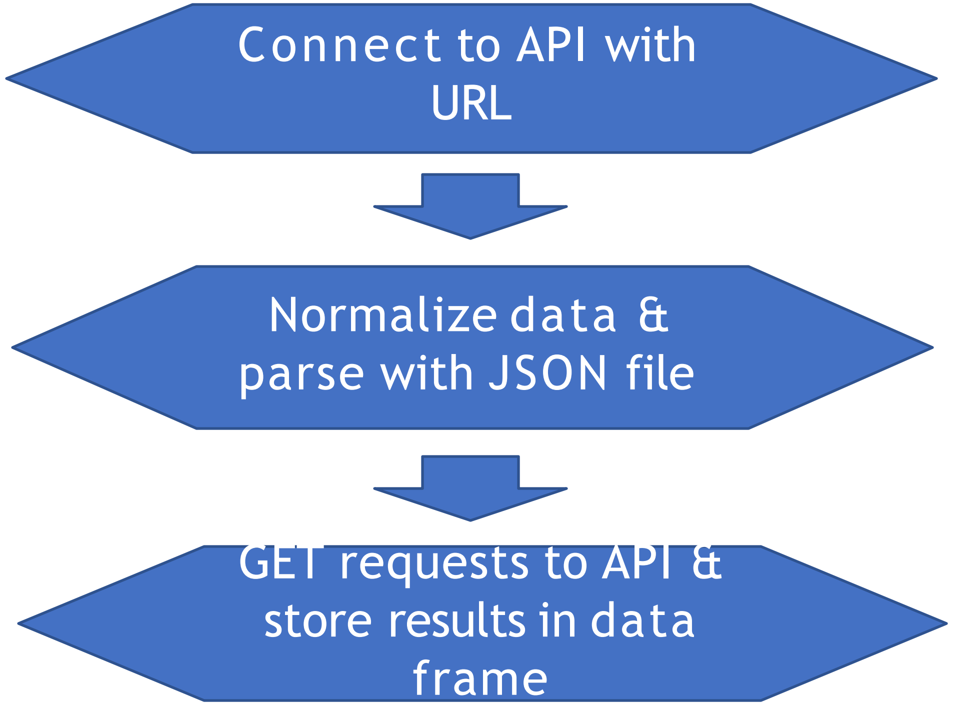
Section 1

Methodology

Methodology

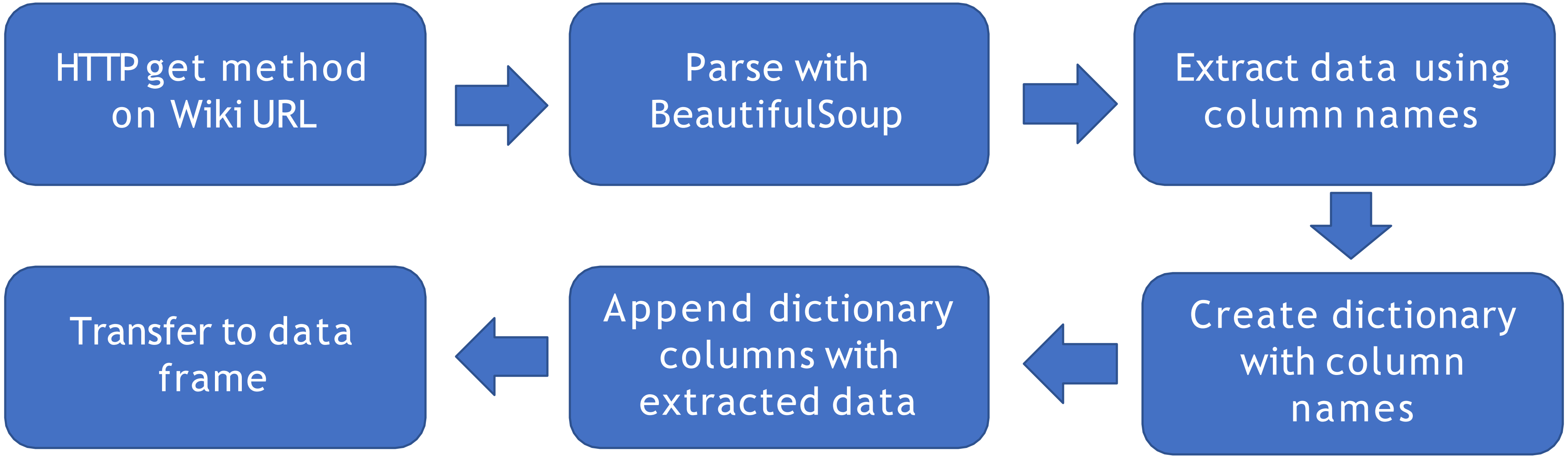
- Predicting the cost of SPACEX flights and first stage landing outcomes
- Data collection methodology:
  - Web scrapping HTML tables from Wikipedia
  - SPACEX API requests into pandas data frame
- Perform data wrangling
  - Removed unimportant data
  - Created training labels and columns
  - Replaced missing values with means
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API



- [Data Collection with API Lab Notebook]

Data Collection - Scraping



[Web Scrapping Lab and Notebook]

Data Wrangling

- Null values in the Payload Mass column were replaced with the mean of the column values
- Value.counts() function shows the need to classify outcome labels to observe outcome variations with other variables
- Created label outcomes (good/bad) from Outcomes column for easier EDA in the future

- [Data Wrangling Lab Notebook]

EDA with Data Visualization

- Seaborn Catplots are specialized scatterplots
  - These plots are best for showing independent data points spread across two dependent variables
- Plotly Barplots
  - For easy to read and clearly deliverable information
- Seaborn Lineplots
  - Useful to show how data evolves over time

- EDA with Visualization Lab and Notebook

EDA with SQL

- Selecting columns From the main table using LIMIT function to specify query
- Using algorithmic functions like sum, avg, min and max
- Selecting successful outcomes using BETWEEN function
- Observing totals of successful and failure outcomes using GROUP BY function
- Sub-queries using two columns
- Using specific DATE functions we observe changes over time
- EDA with SQL Lab and Notebook

Build an Interactive Map with Folium

- Created a Folium map with a center at NASA's JSC space station
- Then added markers for each launch site on the site map
- Then added the launch outcomes with a marker of either red or green as success or fail for each site to see which sites have high success rates
- Analyzed the proximity of each site to nearest coastline, railroad, city and highway
- Folium Maps Lab and Notebook

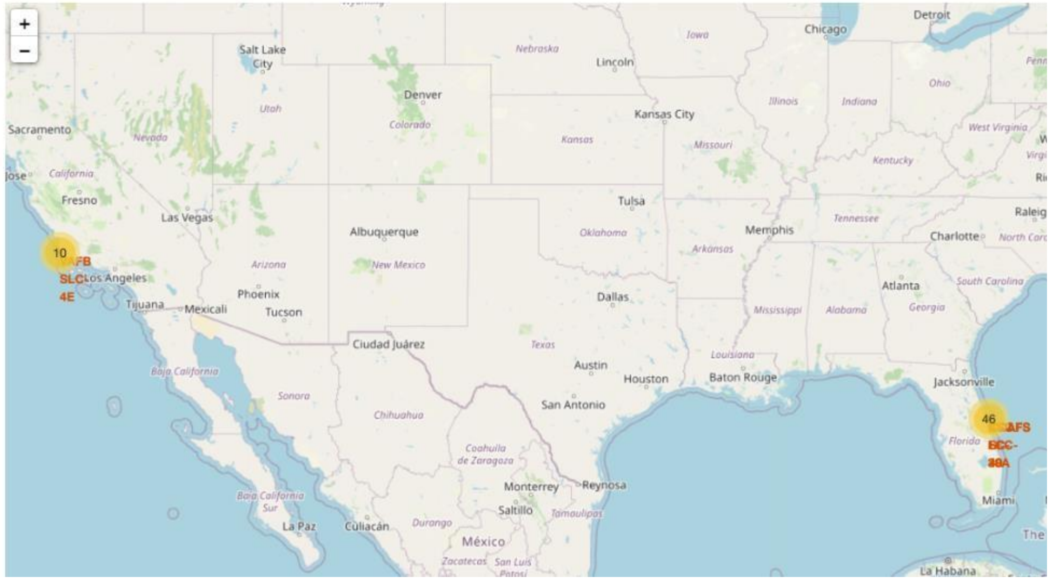
Build a Dashboard with Plotly Dash

- Created a dashboard with a dropdown list to enable Launch Site selection
- Added a pie chart to show the total successful launches count for all sites
- Added a slider to select payload range
- Added a scatter chart to show the correlation between payload and launch success
- Added a callback function for `site-dropdown` as input, `success-pie-chart` as output
- Added a callback function for `site-dropdown` and `payload-slider` as inputs, `success-payload-scatter-chart` as output
- Plotly Dashboard Lab and Notebook

Predictive Analysis (Classification)

- Created an array of the successful and failed launch data
- Split the data into Train and Test data
- Used Logistic Regression, SVM, KNN and Decision Tree Classification methods on the Train/Test data
- Evaluated each model with accuracy and confusion matrix
- Machine Learning Models Lab and Notebook

Results



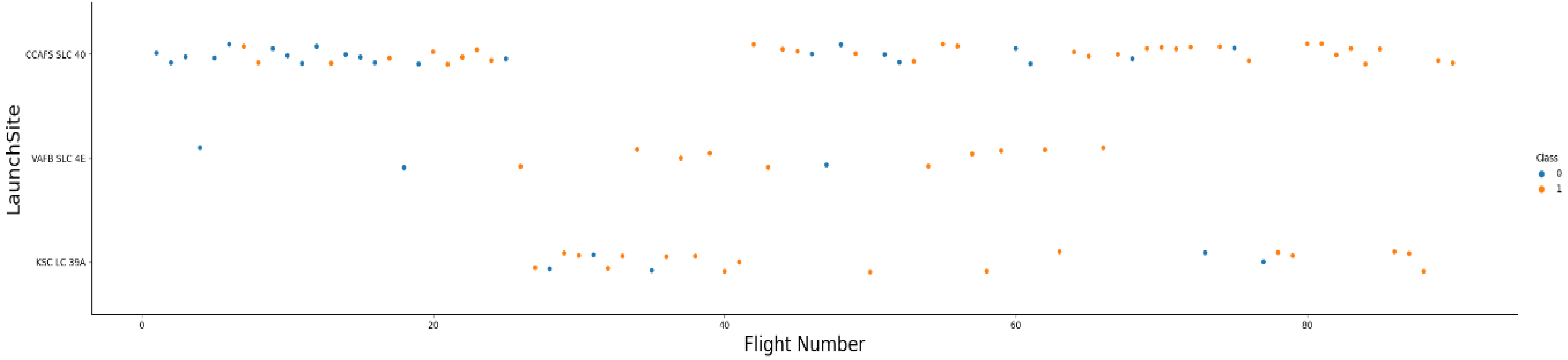
- SVM, KNN and Logistic Regression models are the best for at predictive analysis
- SPACEX has increased the amount of successful launches over the years
- The most successful launches have been from the KSC LC 339A launch site
- Payloads of less weight will perform better overall
- Orbits in LEO perform less successful that other orbits in general
- FT boosters have the highest success rate for Falcon 9

Section 2

Insights drawn from EDA

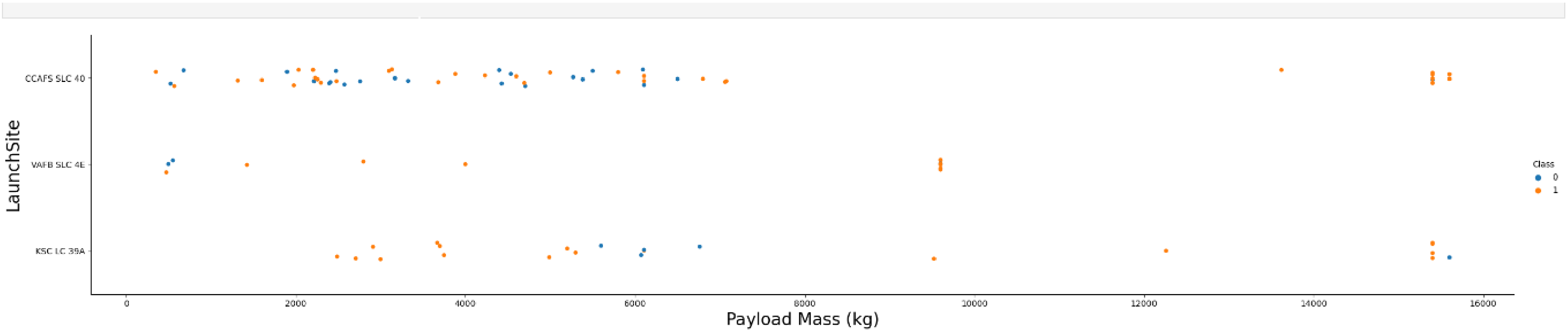
Flight Number vs. Launch Site

- In this catplot we see that the success rate varies among the launch sites when compared to the fight numbers.
- One observation, the CCAFS SLC 40 had more successful flights the higher the flight number is.



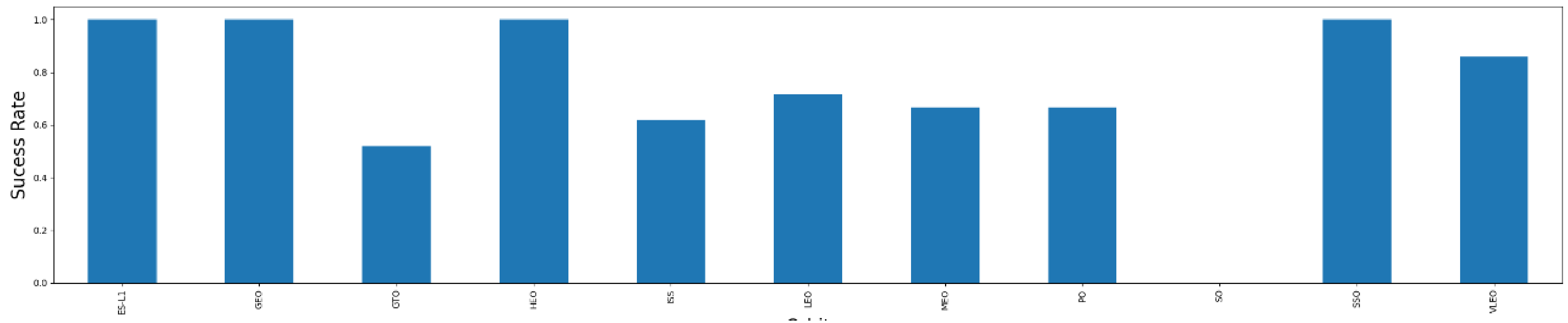
Payload vs. Launch Site

- In this catplot we see most of the results with a Payload under 8000kg with varying results on success.
- We observe that there are less results with a Payload mass above 8000kg though most of those results are successful.



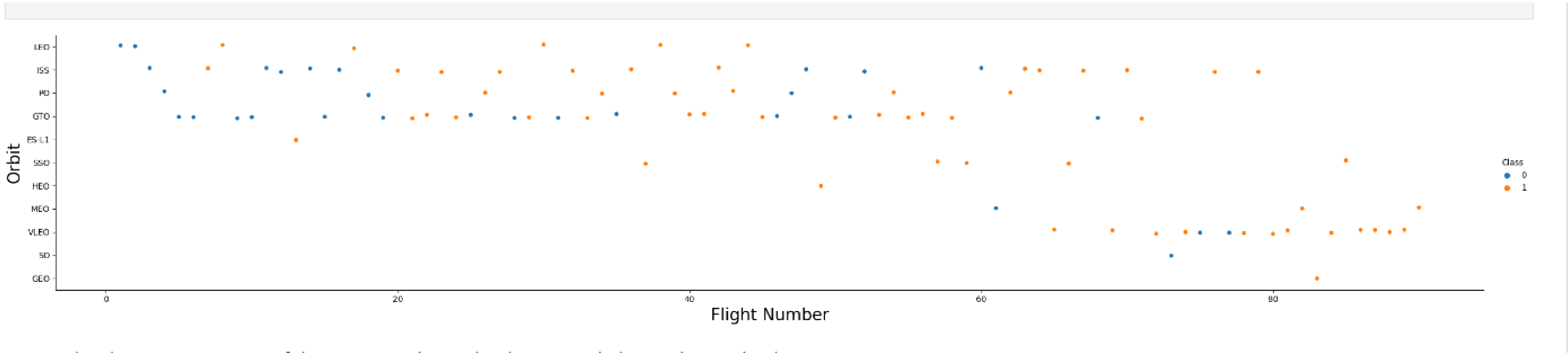
Success Rate vs. Orbit Type

- This bar chart clearly shows which orbit types have more success than others.
- We can observe GTO with the lowest success rate



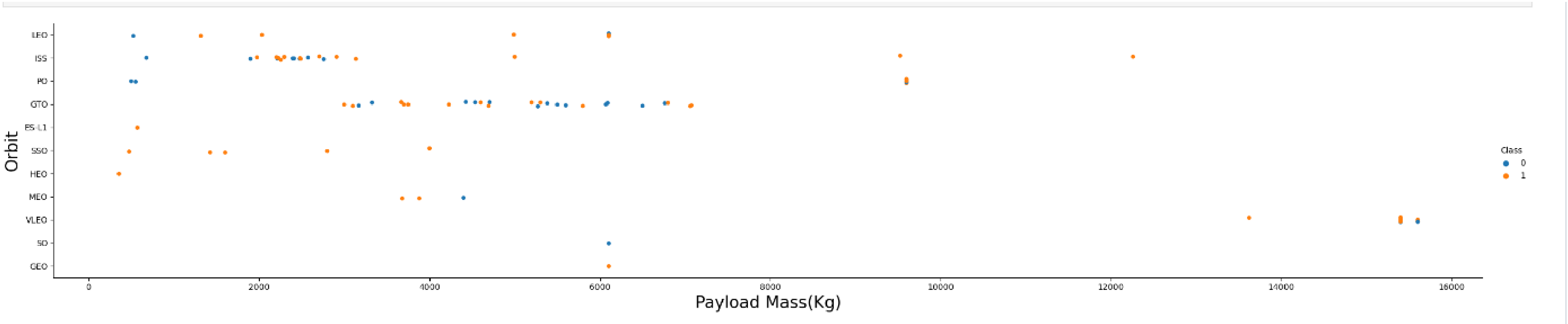
Flight Number vs. Orbit Type

- This catplot show the significant amount of variation in success or failure of the different orbit types and the associated flight number.
- We can observe that we can learn about landing outcome from trying different orbit types



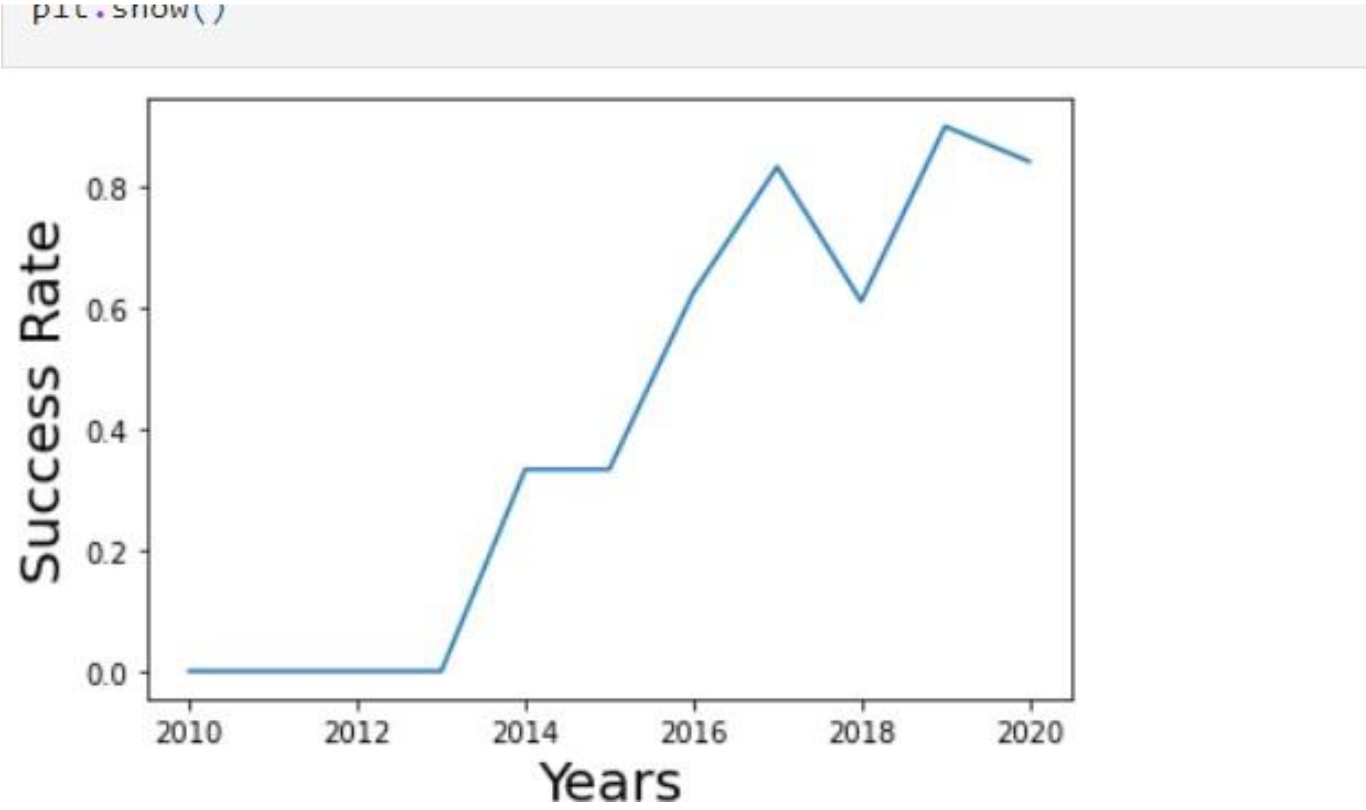
Payload vs. Orbit Type

- This catplot shows the significant variation of the success rate depending on the orbit type and payload mass.
- Similarly to payload mass observed earlier, payload mass above 8000kg has many successes across orbit type



Launch Success Yearly Trend

- This line plot clearly shows the increasing success rate over the years 2010-2020 of SPACEX launches



All Launch Site Names

- Display the names of the unique launch sites in the space mission
- This query show us SPACEX uses 4 different locations for launches

[11]:

```
%sql SELECT Distinct Launch_Site FROM SPACEXTBL
```

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

[11]:

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

Launch Site Names Begin with 'CCA'

- This query shows the full record of launches from Cape Canaveral launch station with a limit of 5 records shown.

```
]: %sql SELECT * from SPACEXTBL where (Launch_Site) LIKE 'CCA%' LIMIT 5;

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

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

Total Payload Mass Launched by NASA (CRS)

- Calculate the total payload mass carried by boosters launched by NASA (CRS)
- This query shows a total of the payload mass in KG carried by NASA (CRS) booster as 45596

Task 3

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

```
%sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL where (Customer) like 'NASA (CRS)';
```

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

payloadmass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- This query shows the average payload mass in Kg of F9 v1.1 boosters as 2928.4

Task 4

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

```
]: %sql select avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL where (Booster_Version) like 'F9 v1.1';
```

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

]:

payloadmass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The first successful landing on ground pad was December 22 2015

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

```
%sql select Booster_Version from SPACEXTBL where "Landing _Outcome" = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4
```

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

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

Total Number of Successful and Failure Mission Outcomes

- This query shows the total number of successful and failure mission outcomes
- There were 98 successful mission outcomes with 4 types of outcomes total

List the total number of successful and failure mission outcomes

```
%sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

missionoutcomes
1
98
1
1

Boosters Carried Maximum Payload

- This query shows the names of the booster which have carried the maximum payload mas

```
%sql select Booster_Version as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_ =(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

boosterversion
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

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

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

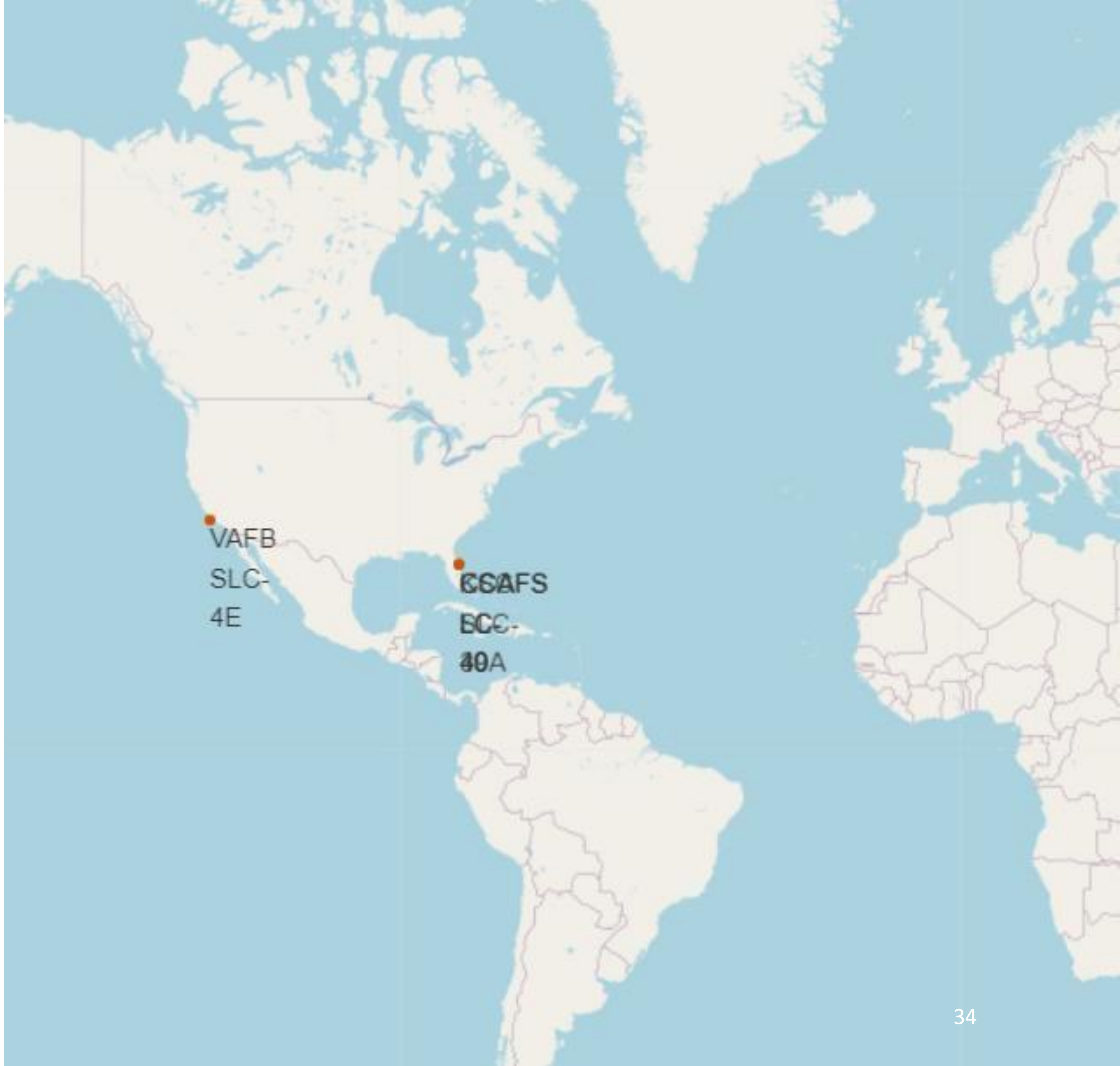
landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

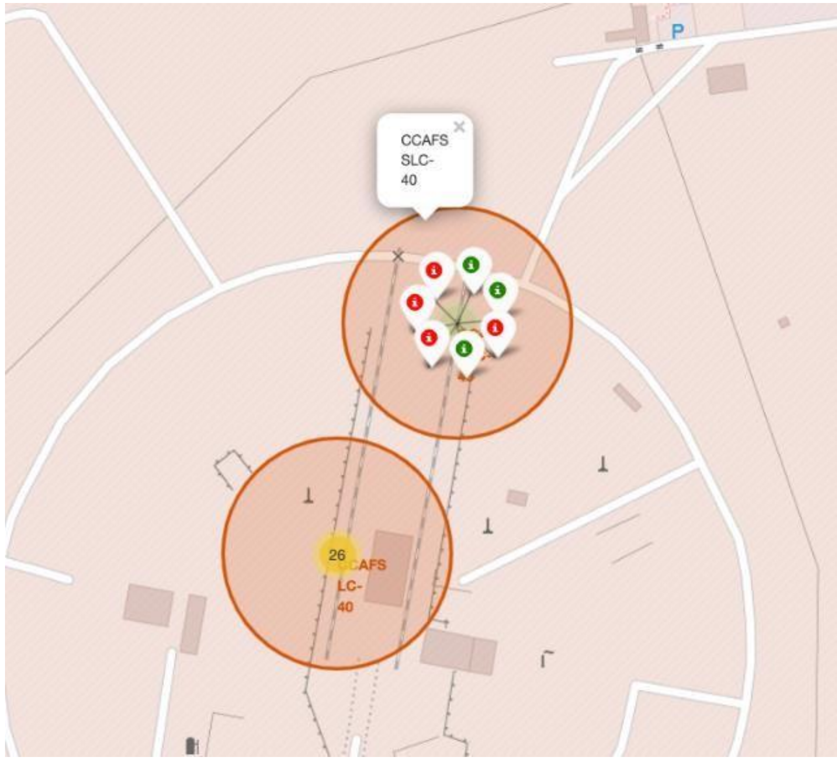
SPACEX Launch Site with Folium Map

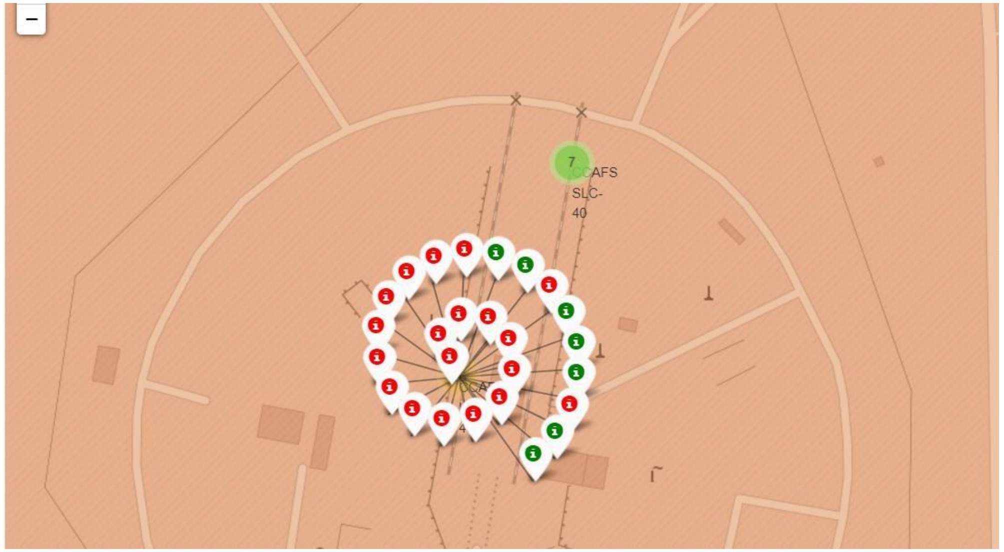
- This Folium Map shows the four launch site locations. We can observe one location on the west coast of the USA and multiple on the East coast of USA



Success/failed launches for each SPACEX site on a Folium Map

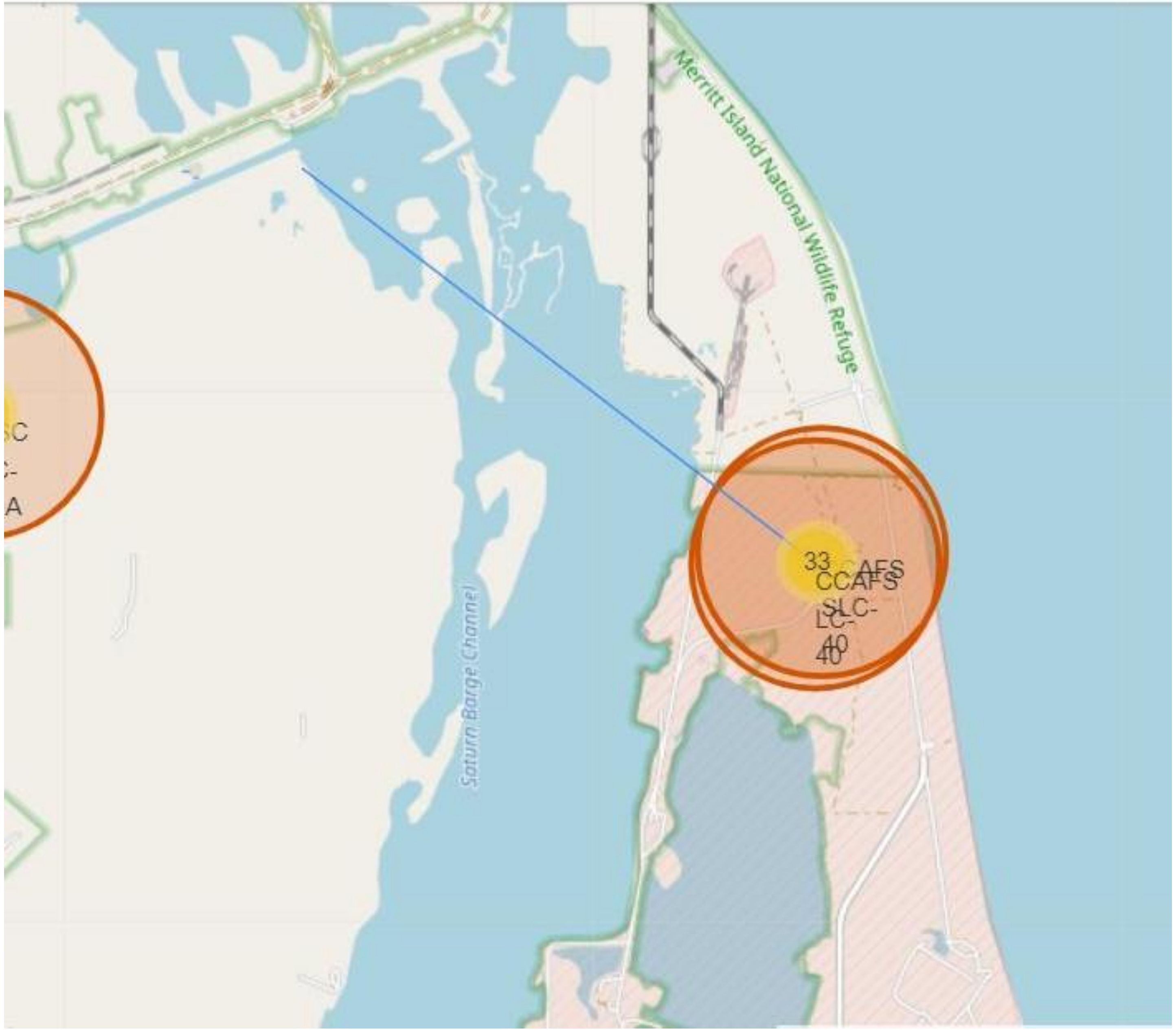
- This map shows us the labeled flights of success or failure marked with a corresponding color from both Cape Canaveral locations





Distances between a launch site and the closest coastline

- This Map shows how we can use the MousePosition function to find close by roads, cities or coastlines from a certain location

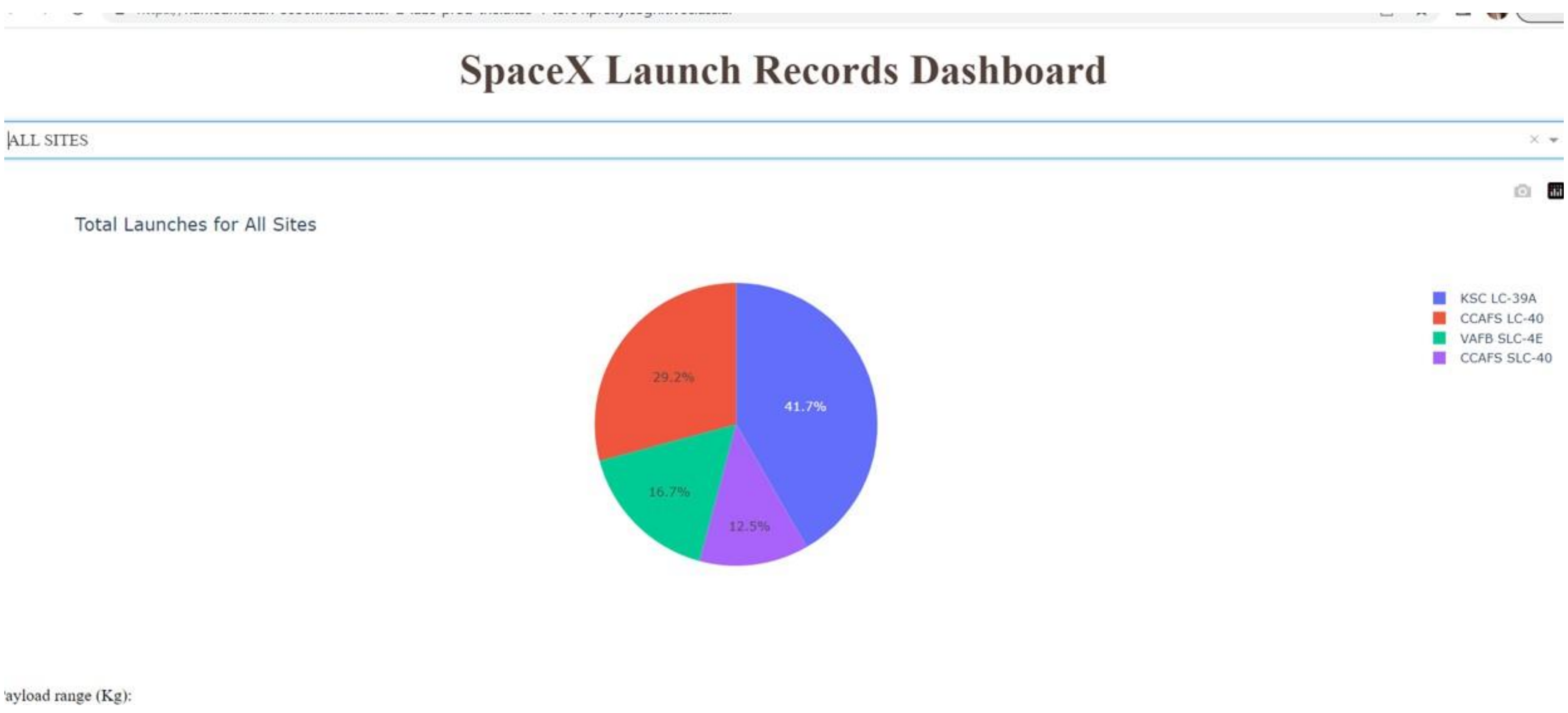


Section 4

Build a Dashboard with Plotly Dash

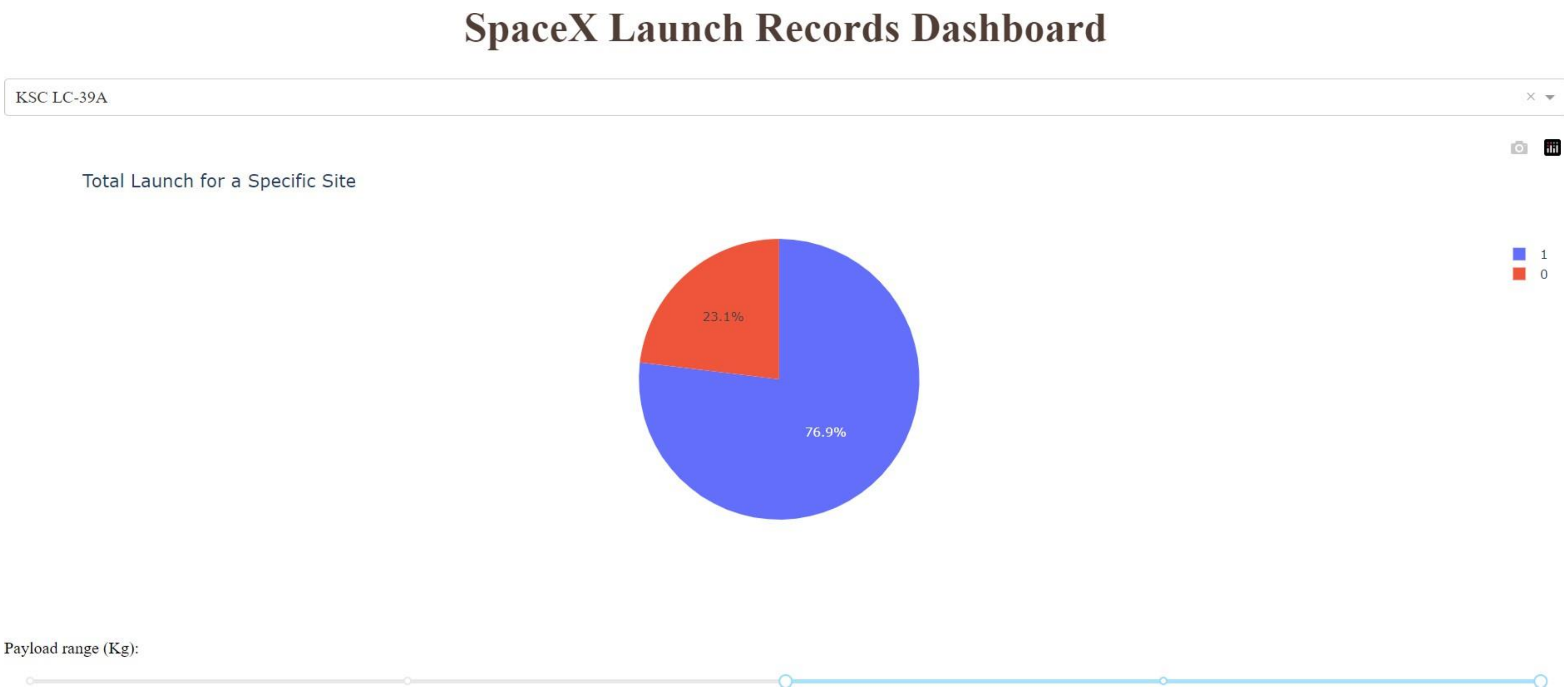
Pie Chart of launch success in all SPACEX locations

- This pie chart shows the percentage of launches from the different SPACEX launch locations



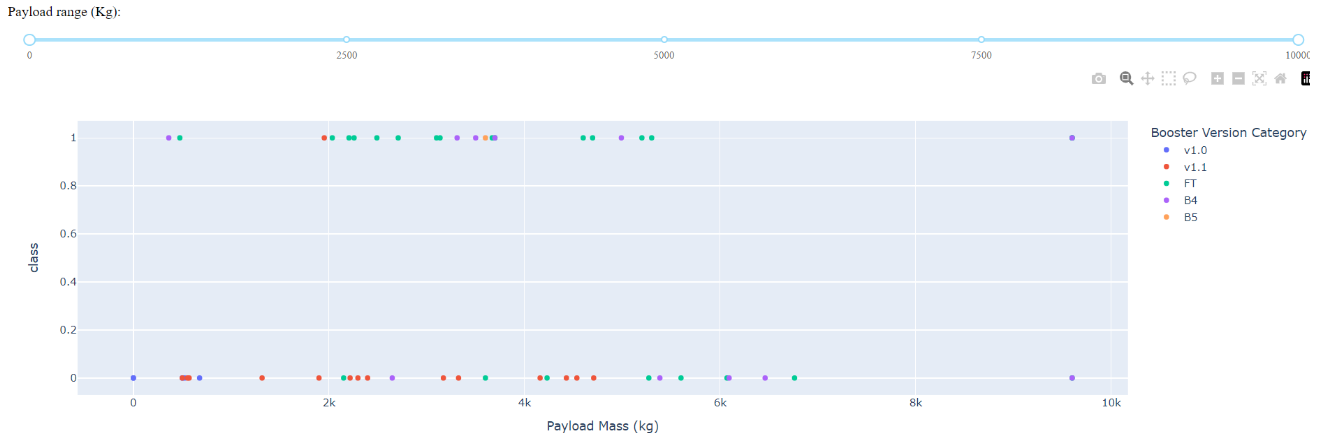
Pie Chart of launch site with highest launch success ratio

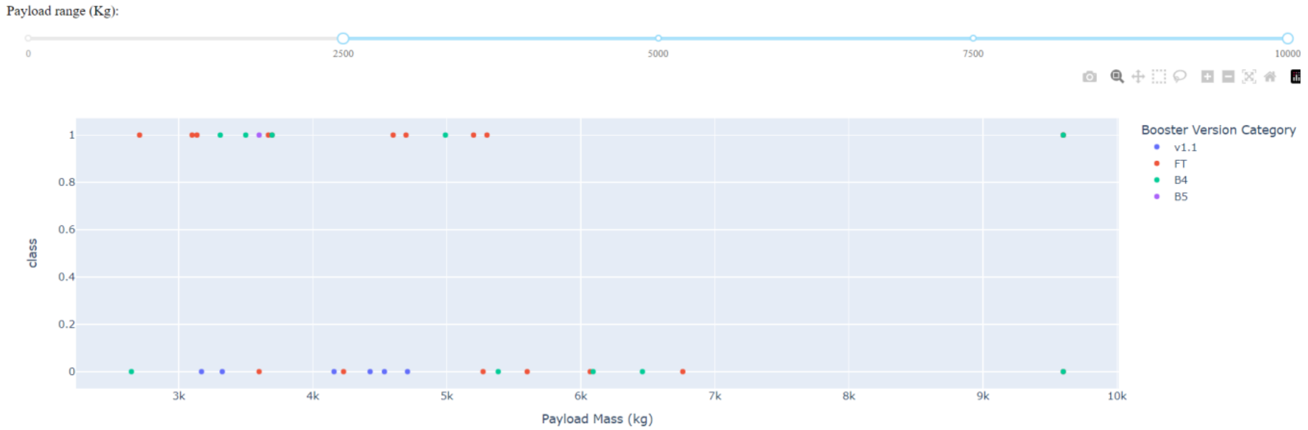
- From the previous pie chart we can see that KSC LC-39A has the highest launch success ratio. This pie chart shows the total success and failures at KSC LC-39A

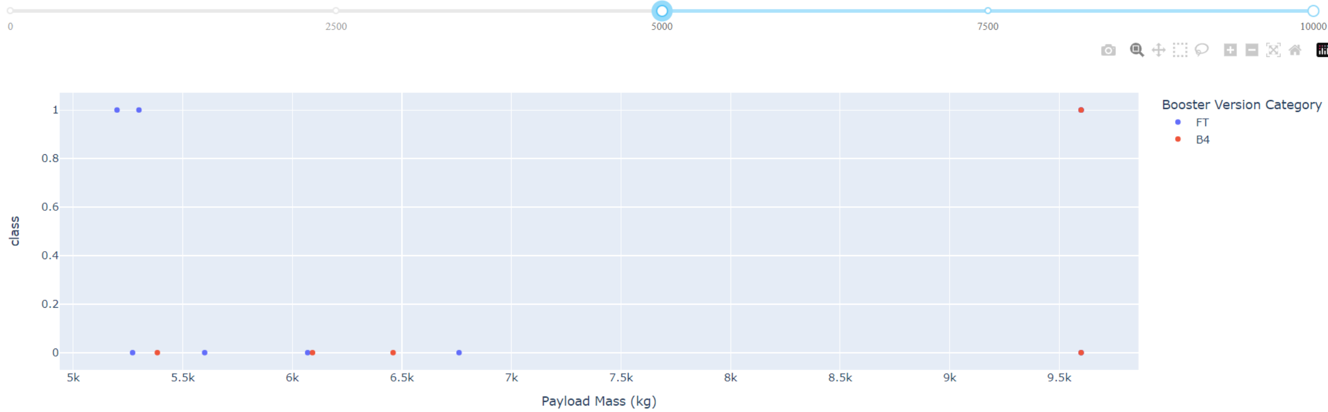


Scatters Plot of Payload vs Launch Outcome at various payloads

- From these scatter plots we can see as the payload increases there are fewer booster versions that have success. The booster version with the most success at higher payloads are FT and B4. Overall the boosters with the most success is FT.





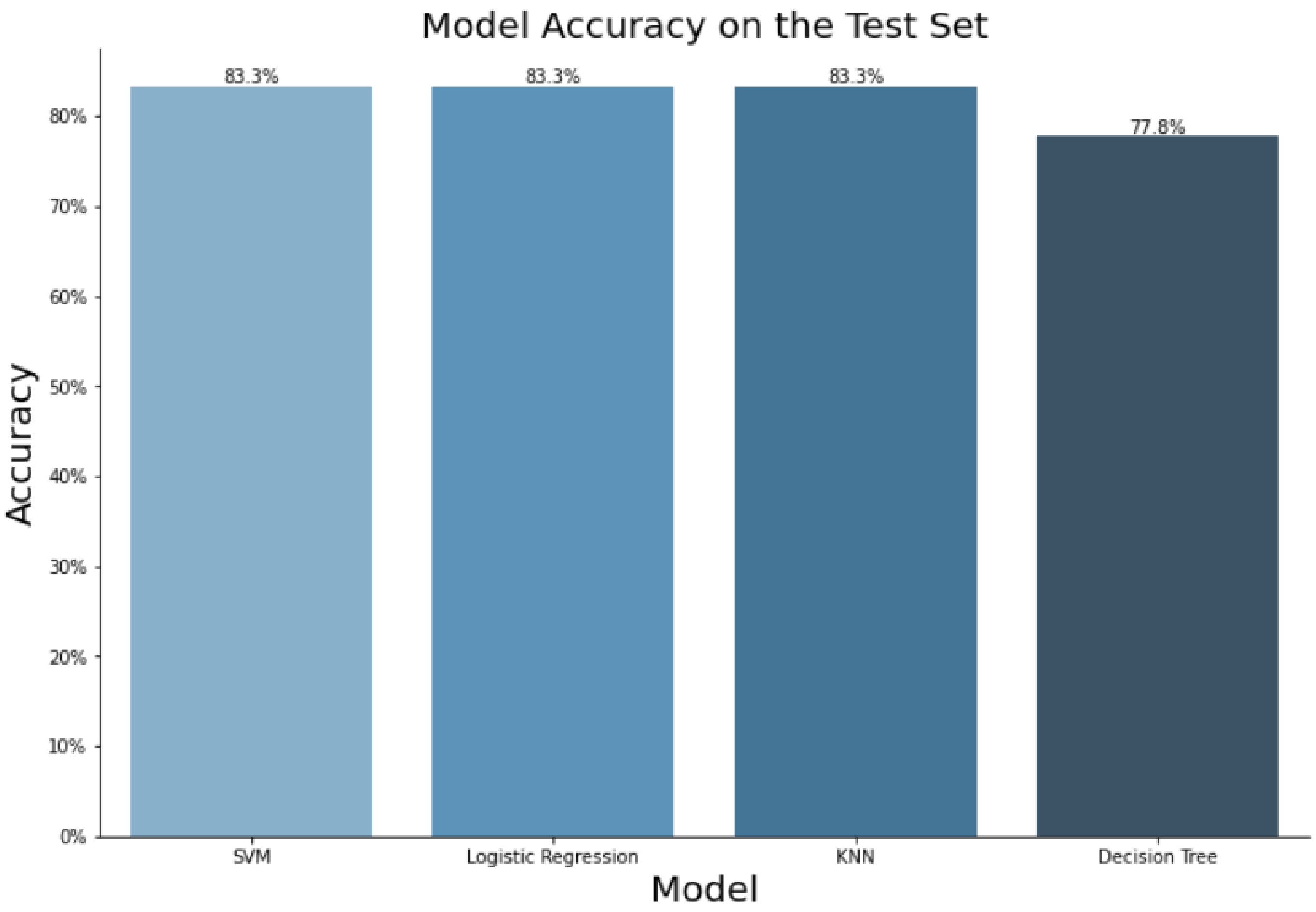


Section 5

Predictive Analysis (Classification)

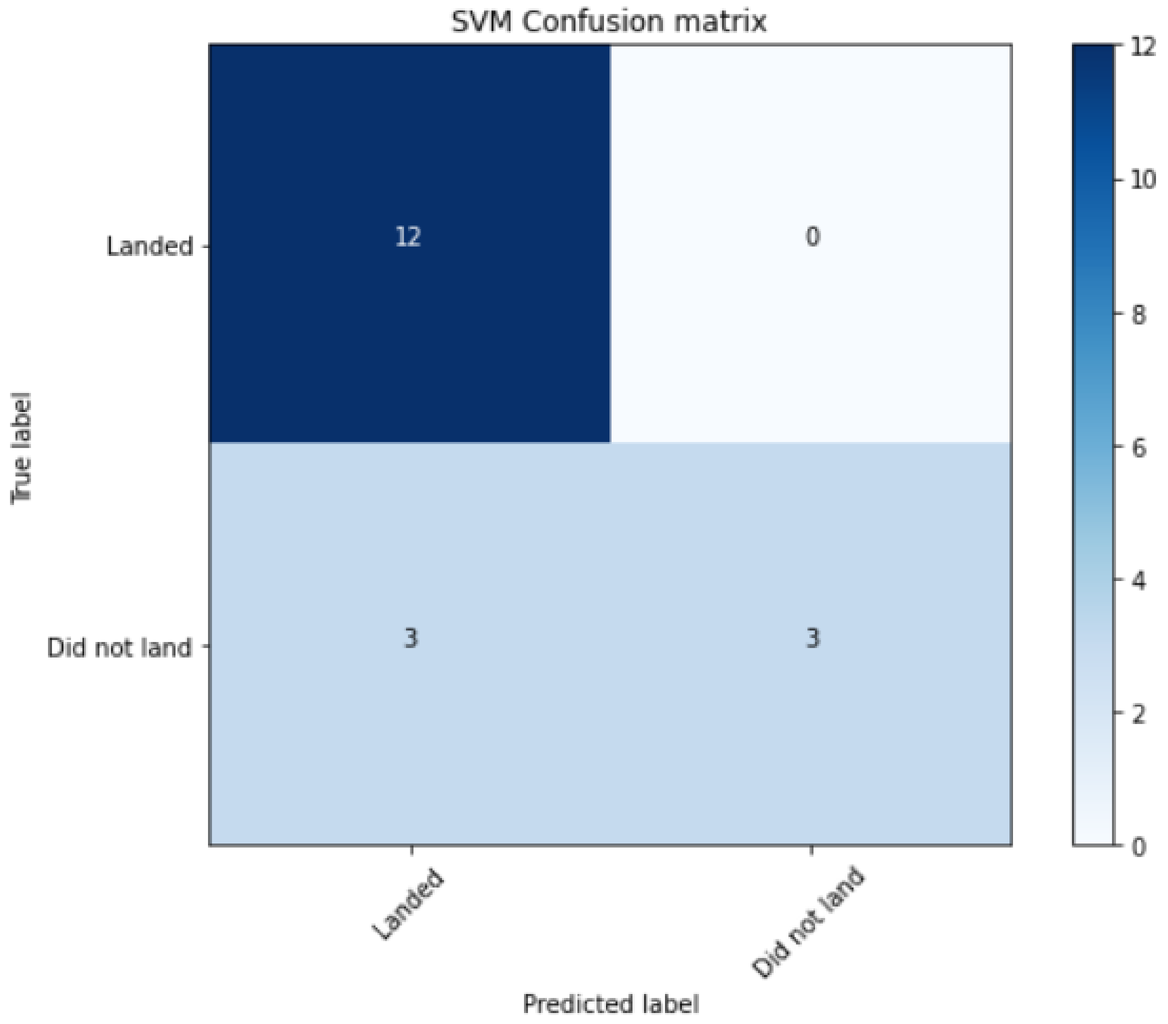
Classification Accuracy

- This bar chart shows the accuracy for each machine learning algorithm we tested.
- SVM, Logistic Regression and KNN classification have the highest classification accuracy.



Confusion Matrix

- This confusion matrix shows the best performing model with zero false positives as SVM.
- Though SVM, KNN, and Logistic have identical accuracy and precision. SVM has the best AUC predictive accuracy making it the best model.



Conclusions

- SVM, KNN and Logistic Regression models are the best for at predictive analysis
- SVM is the best model because it has the greatest AUC for predictability
- SPACEX launches have increased the amount of successful launches over the years
- The most successful launches have been from the KSC LC 339A launch site
- Payloads of less weight will perform better overall
- Orbits in LEO perform less successful that other orbits in general
- FT boosters have the highest success rate for Falcon 9

Appendix

- Github repository for all notebooks used in this presentaions

Thank you!