IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Y.S. Nimesh

Outline

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

Executive Summary

- Methodologies used to analyze data:
  - Data collection through web scraping and SpaceX API
  - Exploratory Data Analysis (EDA) with visualization, SQL, interactive maps and dashboard
  - Machine Learning Modeling
- Summary of all results
  - Data collection was successful via the two resources mentioned
  - EDA was able to identify which variables were most affective for predicting successful launches
  - Machine Learning Modeling determined the best model for predicting the result of a launch.

Introduction

- Project background and context
  - The goal of this project is to predict if the first stage of the Falcon 9 rocket will successfully land. With a successful landing the launch costs SpaceX 62 million dollars, compared to the competitor's 165 million dollar per launch cost. The successful prediction of a landing is helpful to our company because we can then determine what the cost of a launch will be.
- Problems you want to find answers
  - What are the leading factors of a successful landing?
  - How much do these factors contribute to the outcome of a landing?
  - What are the optimal conditions for a successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web Scraping via Wikipedia
- Perform data wrangling
  - Created a landing outcome label from the Outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data into train and test set, subsequently built four classification models and evaluated them based on their accuracy

Data Collection

- Data was collected using SpaceX API and Web Scraping
  - The API collected launch, payload, and rocket information
  - SpaceX API URL: https://api.spacexdata.com/v4/launches/past

Request SpaceX API → API responds with json file → Create dataframe from the json → Clean and export dataframe

- Web Scraping collected launch, landing, and payload information via Wikipedia
  - Wikipedia URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

HTML get request for Wikipedia → Extract data using Beautiful Soup → Create a dataframe → Export dataframe

Data Collection – SpaceX API

- The SpaceX API allowed retrieval of information about the launches
- Steps to extract this data is outlined by the flowchart shown

[Data collection API]

Request API, get response and convert to json file

↓

Transform into a Pandas dataframe, created subset of dataframe keeping important features

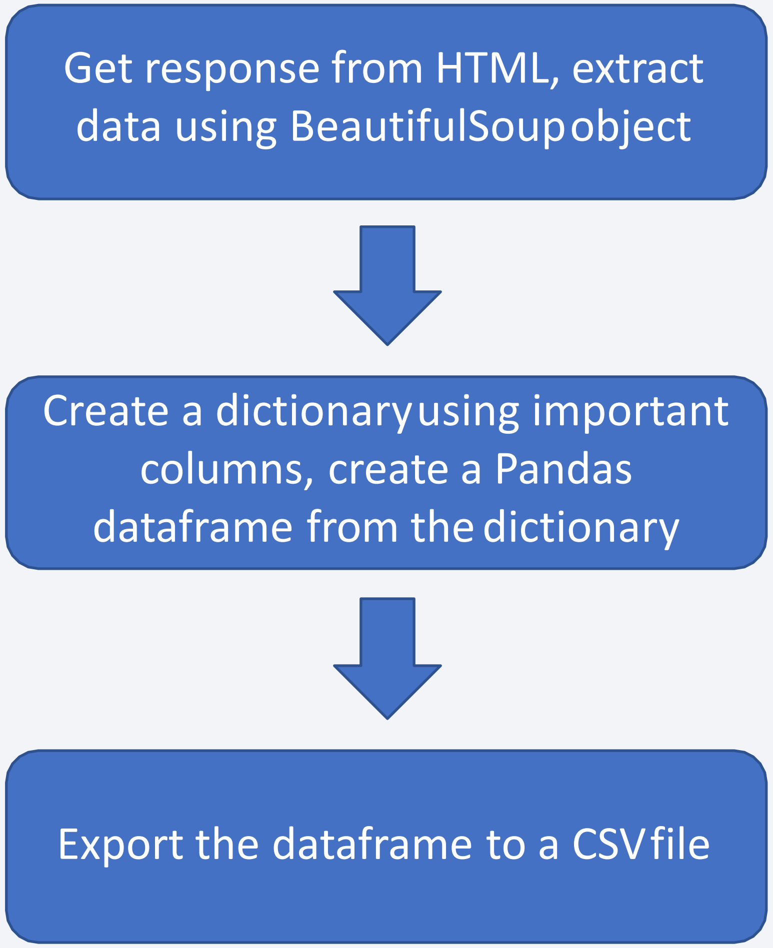
↓

Filter dataframe for Falcon 9 only, deal with missing values, then export dataframe

Data Collection - Scraping

- Web Scraping allowed for retrieval of important launch information from the given Wikipedia page
- The steps to retrieve this data is outlined by the flowchart shown

[Web Scraping Code]



Data Wrangling

- To begin, I performed exploratory data analysis to determine the characteristics of certain columns. Next, I created a landing outcome label from the outcome column where 0 was assigned to a bad outcome and 1 was assigned to a successful outcome. Finally, I exported the data to a CSV for future use.

- [Data Wrangling Code]

EDA with Data Visualization

- Scatter Plots
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Flight Number vs Orbit
  - Payload Mass vs Orbit
  - These plots show the relationship between two given variables

  [EDA with Visualization Code]

- Bar Chart
  - Orbit vs Success Rate
  - This showed which orbits have high success rate
- Line Chart
  - Year vs Success Rate
  - This shows the trend in success rate throughout the years

EDA with SQL

- The following SQL queries were performed to gather and understand the data:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List 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
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names in year 2015
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010 -06-04 and 2017-03-20, in descending order

[EDA with SQL Code]

Build an Interactive Map with Folium

- Objects including markers, circles, lines, and clusters were created using Folium
  - Markers indicate launch sites
  - Circles indicate the area surrounding launch sites
  - Lines show the distance between given coordinates
  - Clusters indicate launches within a given launch site
  - [Folium Code]

Build a Dashboard with Plotly Dash

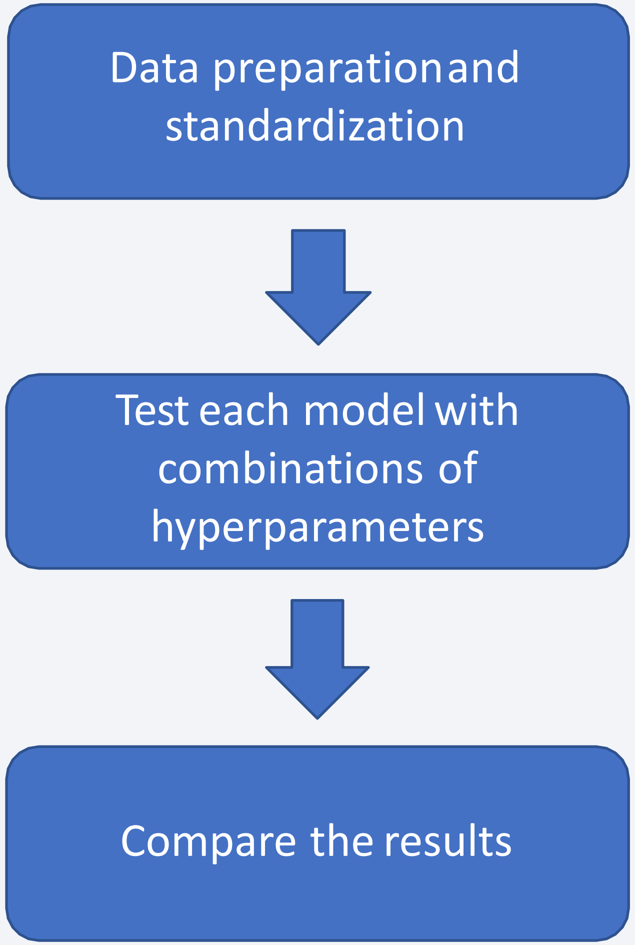
- The following graphs and plots were used to visualize data:
  - Percentage of successful launches by site pie chart
  - Payload range vs booster version graph

- These visualization tools made it possible to analyze relations between payloads and launch sites, which enabled me to identify where is the best launch site according to payloads.

Predictive Analysis (Classification)

- The four models built for comparison were:
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN

[Machine Learning Code]



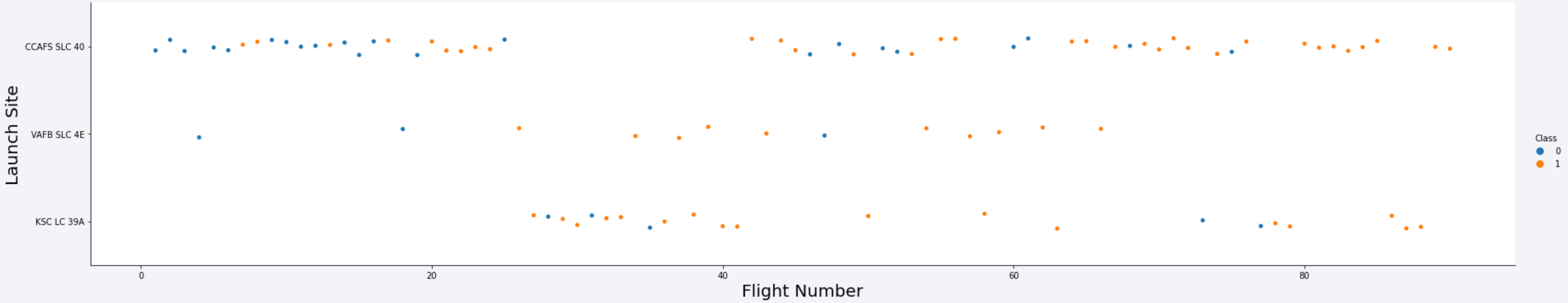
Results

- Exploratory data analysis results:
  - The number of successful landings increased as more missions were launched.
  - The first successful landing was in 2015 which was five years after the first launch.
  - The average payload of the F9 v1.1 booster is 2928kg.
  - SpaceX holds missions at 4 different launch sites.
- Predictive analysis results:
  - The decision tree was the best model for the training data with an accuracy of 88.75%.
  - The other three models had a training accuracy of 83.33%.

Section 2

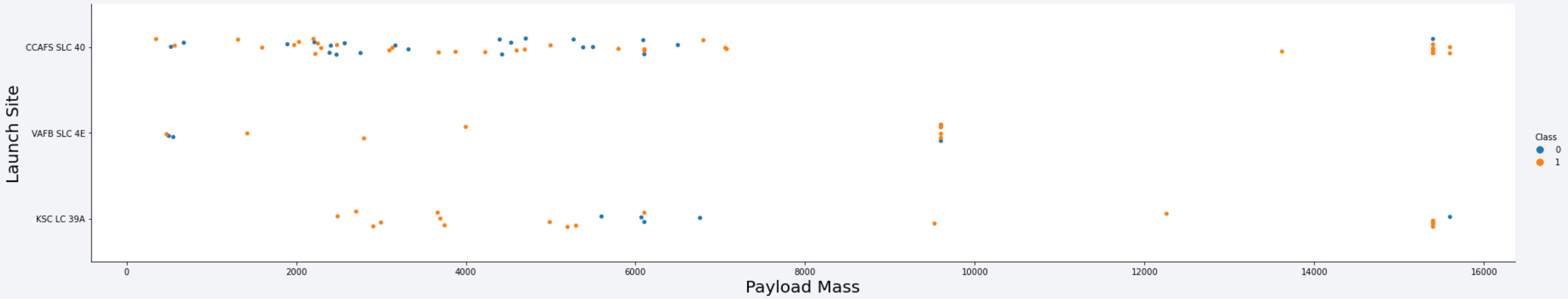
Insights drawn from EDA

Flight Number vs. Launch Site



- With this plot we are able to see that the CCAFS LC-40 launch site is more frequently used.
- The success rate is increasing with the number of launches.

Payload vs. Launch Site

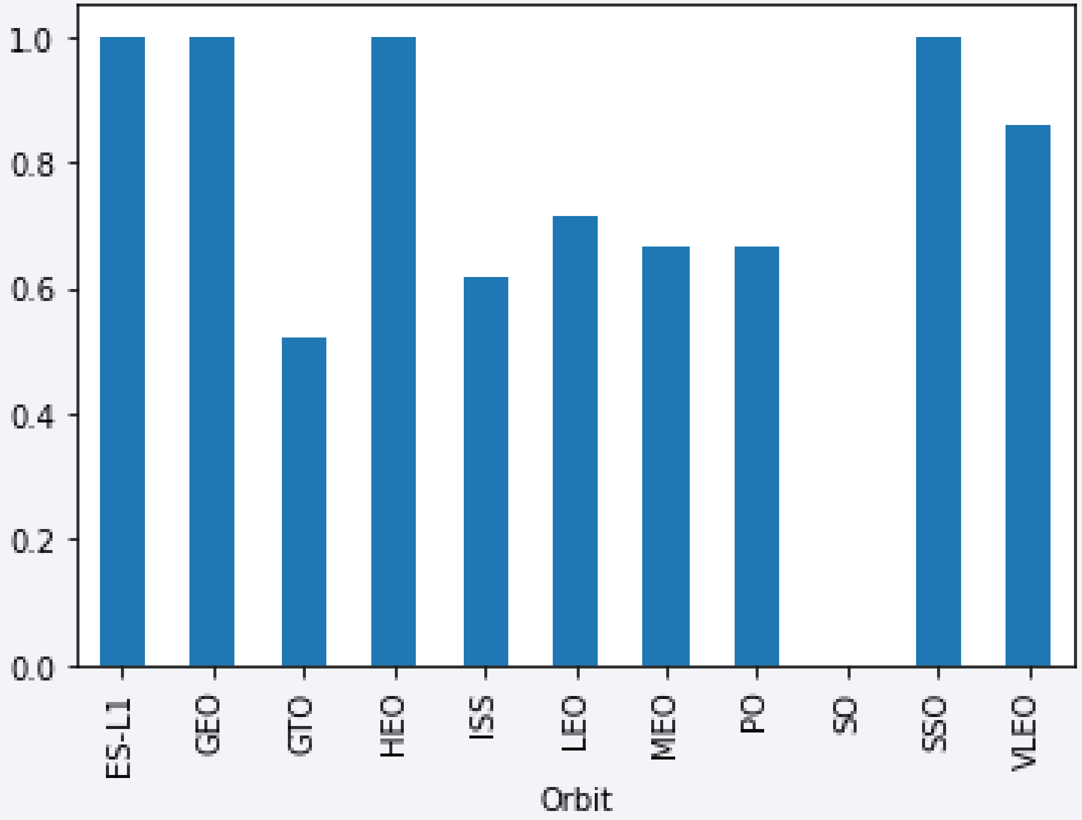


- The VAFB-SLC launch site has no launches for heavy payload mass which would be greater than 10000kg.
- The success rate for the CCAFS LC-40 dramatically increases with payload mass of 13000kg or greater

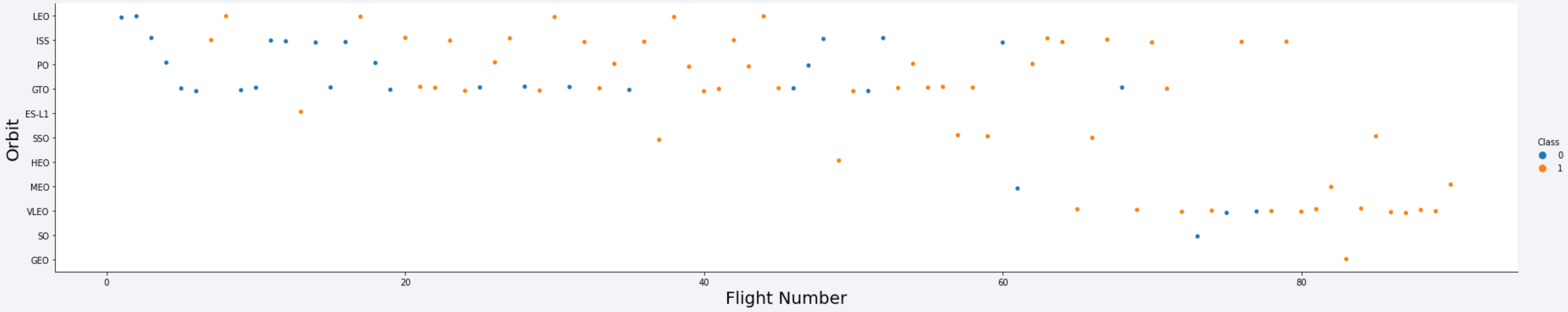
Success Rate vs. Orbit Type

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

- The next highest success rate is the VLEO orbit with the others not mentioned having a lower success rate.

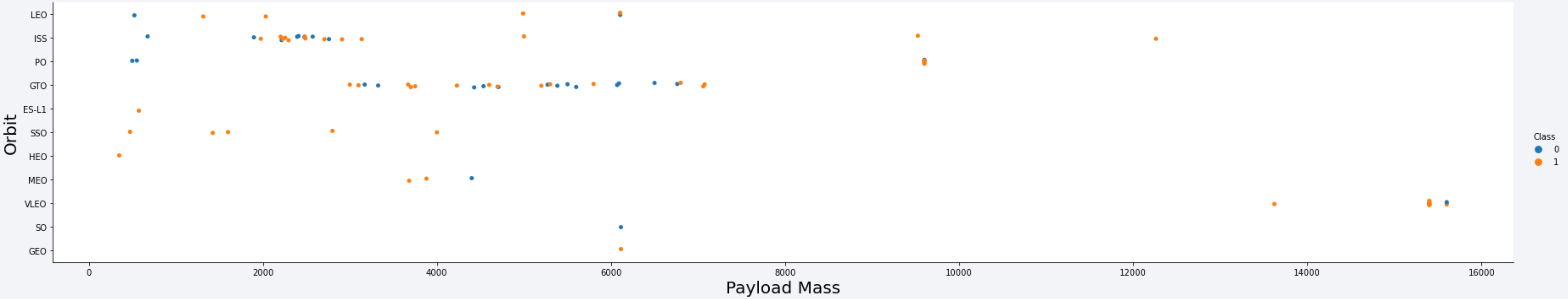


Flight Number vs. Orbit Type



- The success rate is related to the number of flights with the LEO orbit, however there seems to be no relationship between flight number for the GTO orbit.
- The success rate appears to rise with number of flights in the ISS orbit.

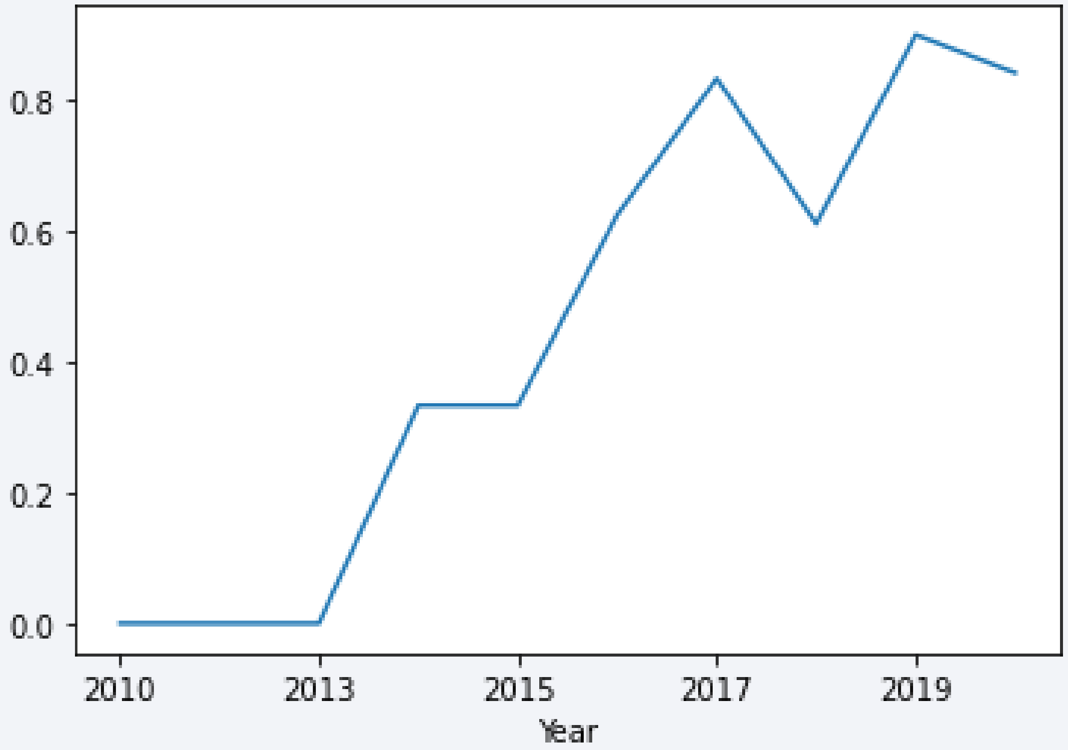
Payload vs. Orbit Type



- There are more successful landings of heavy payloads with the Polar, LEO, and ISS orbits.
- The GTO orbit is not as distinguishable due to successful and failed landings at heavy payload.

Launch Success Yearly Trend

- The success rate has continually increased starting in 2013.
- From 2017-2018 there was a sharp decrease in success rate.



All Launch Site Names

```
%sql select distinct LAUNCH_SITE from SPACEXTBL;
```

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

- This query ignores duplicate entries, which outputs the distinct launch site names.

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5;
```

Date as DD-MM-YYYY	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

- This query uses 'where' followed by 'like' to filter the results to launch sites that start with "CCA".
- The 'limit 5' function outputs the first 5 results.

Total Payload Mass

```
%%sql select sum(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD from SPACEXTBL
where PAYLOAD like '%CRS%';
```

total_payload
111268

- This query gives the sum of payload mass where the payload contains 'CRS', which means it was launched by NASA.

Average Payload Mass by F9 v1.1

```
%%sql select avg(PAYLOAD_MASS__KG_) as AVG_PAYLOAD from SPACEXTBL
where BOOSTER_VERSION = 'F9 v1.1';
```

avg_payload
2928

- This query is calculated by the average payload mass where the booster version is 'F9 v1.1'.

First Successful Ground Landing Date

```
%%sql select min(DATE) as FIRST_GP_SUCCESS from SPACEXTBL
where LANDING__OUTCOME = 'Success (ground pad)';
```

first_gp_success
2015-12-22

- This query shows the first successful landing outcome by ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
select distinct BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ between 4000 and 6000
and LANDING__OUTCOME = 'Success (drone ship)';
```

- This query lists the name of the boosters that have success in drone ship and a payload mass greater than 4000, but less than 6000.

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

Total Number of Successful and Failure Mission Outcomes

```
%sql select MISSION_OUTCOME, COUNT(*) as QUANTITY from SPACEXTBL group by MISSION_OUTCOME;
```

- This query shows the quantity of outcomes grouped by the mission outcome.

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

Boosters Carried Maximum Payload

```
%%sql select distinct BOOSTER_VERSION from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

- This query lists the names of the booster, which have carried the maximum payload mass.

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

2015 Launch Records

```
%%sql select BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL
where LANDING__OUTCOME = 'Failure (drone ship)'
and DATE_PART('YEAR', DATE) = 2015;
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- This query lists the failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015.

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

```
%%sql select LANDING__OUTCOME, COUNT(*) as QUANTITY from SPACEXTBL
where DATE between '2010-06-04' and '2017-03-20'
group by LANDING__OUTCOME order by QUANTITY desc;
```

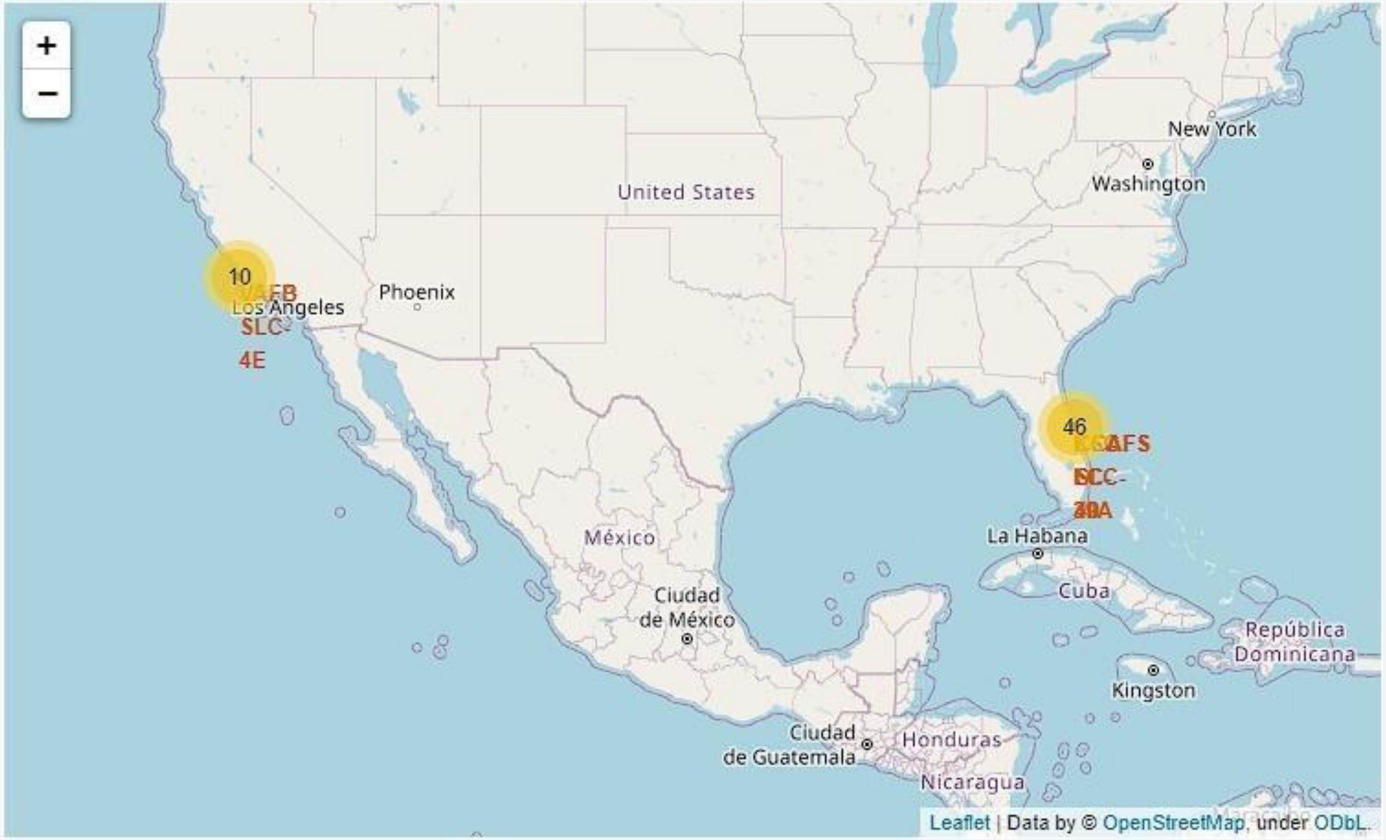
- This query ranks the quantity of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

landing_outcome	quantity
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

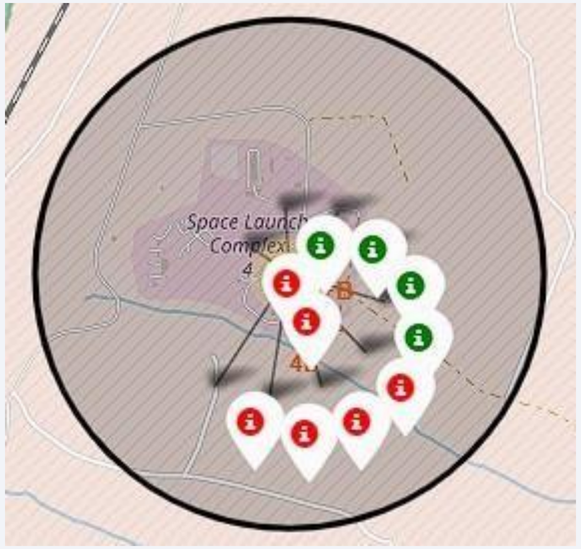
Launch Sites Proximities Analysis

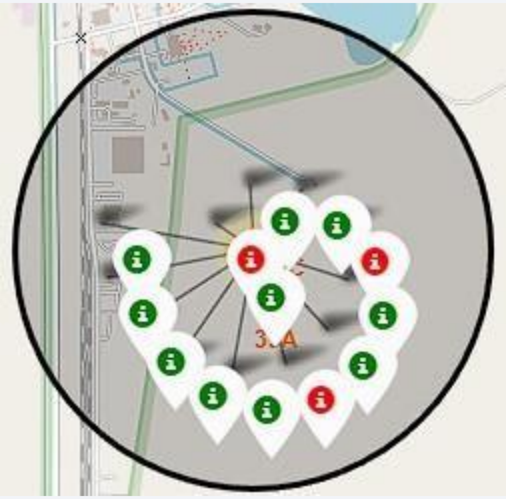
Folium Map: Launch Sites

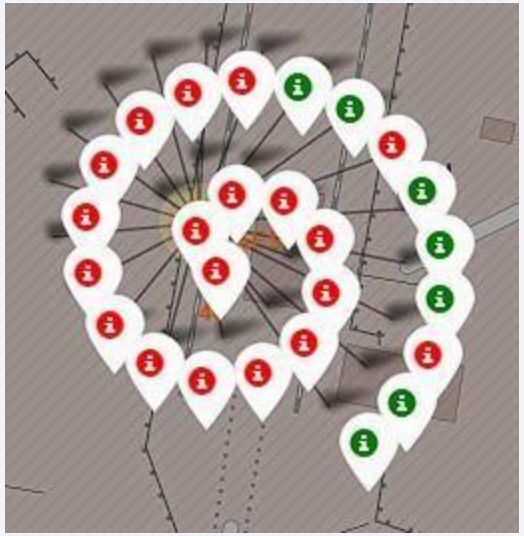


- The map shows us that SpaceX has their launch sites along the coast in the United States

Folium Map: Color Labeled Outcomes

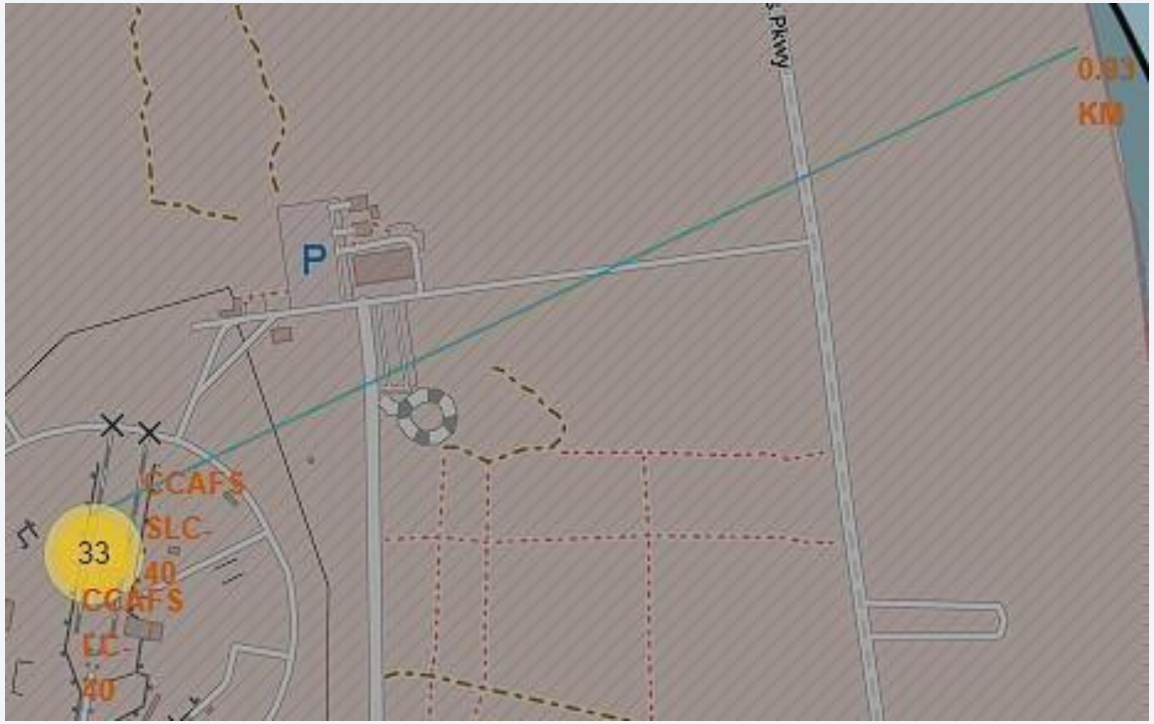






- The green marker illustrates a successful launch, while the red marker illustrates a failed launch.
- The KSC LC-39A launch site, shown in the second picture, has a higher success rate.

Folium Map: CCAFS SLC-40 Distances

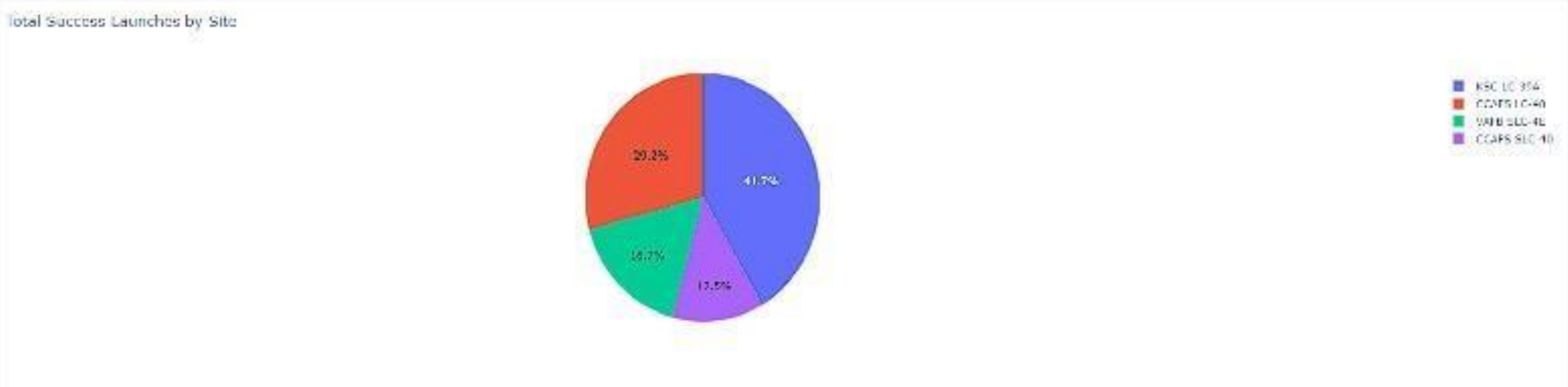


- The CCAFS SLC-40 launch site is very close to the coast. This launch site is also near railroads and highway, however not very close to cities.

Section 4

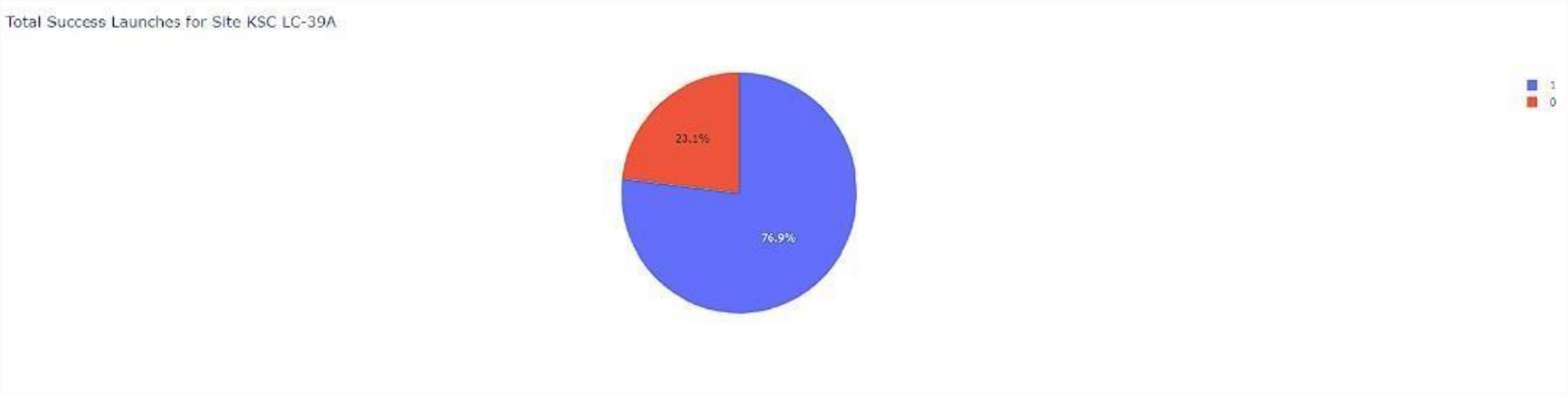
Build a Dashboard with Plotly Dash

Dashboard: Success by Site



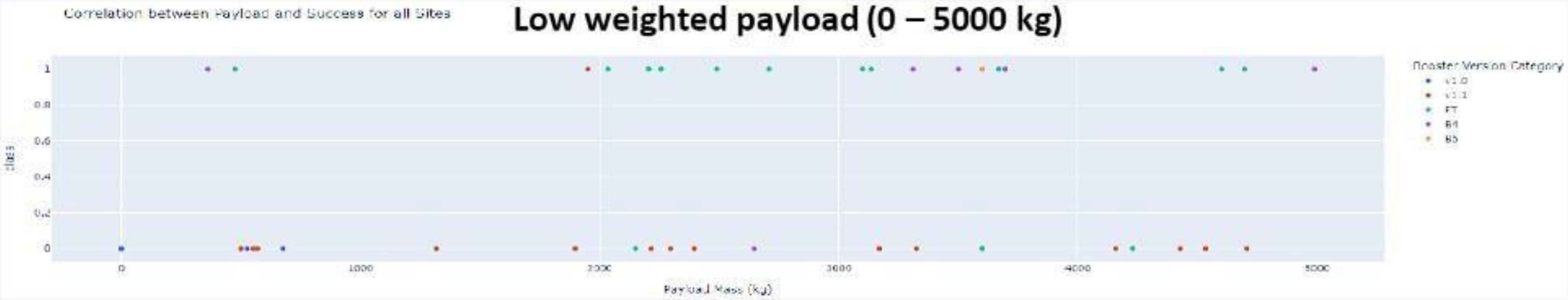
- The site with the most successes is KSC LC-39A

Dashboard: Success Rate of KSC LC-39A



- The KSC LC-39A launch site has the highest success rate of 76.9%

Dashboard: Payload vs Launch Outcome



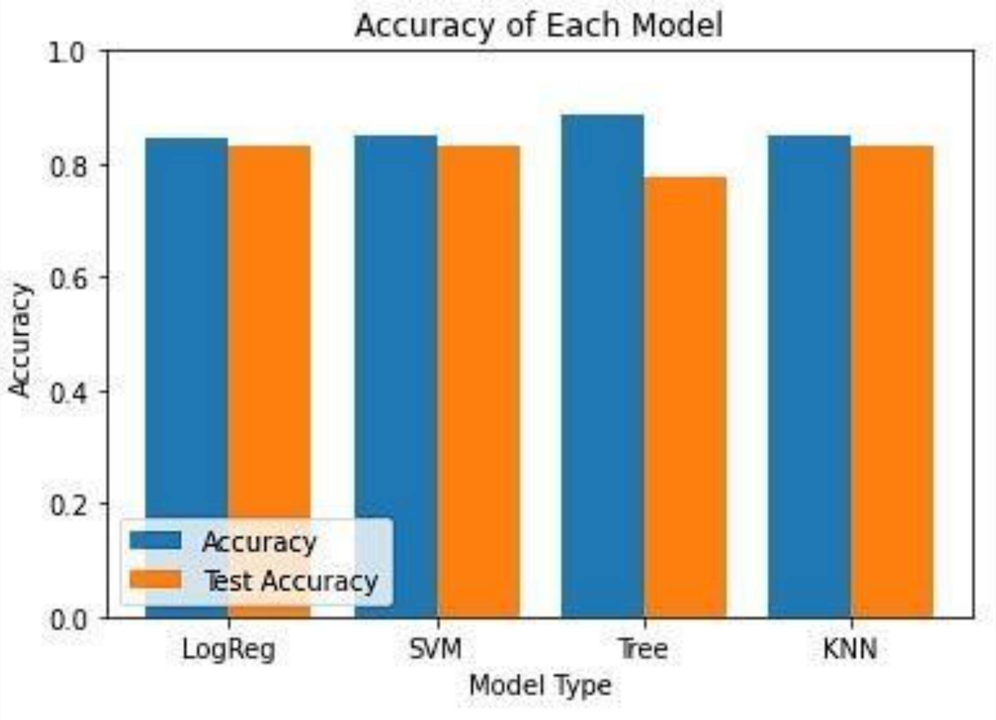
- Payload of 5000kg and less has a higher success rate than payload of 5001kg and greater.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The accuracy of the decision tree model was the highest with 88.75%, but for the testing data it performed the worst of the four with 77.78%.
- The decision tree would be the best model to use and with more test data I believe the accuracy would increase.



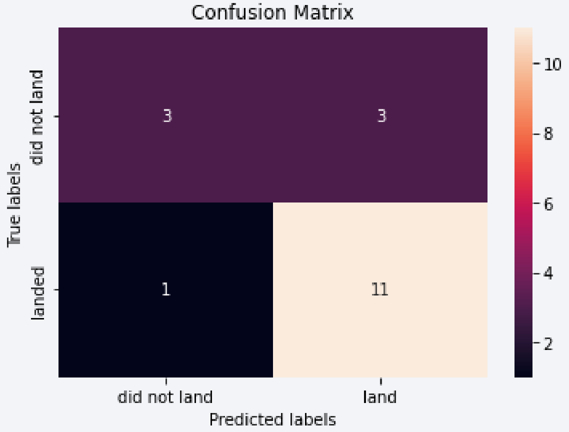
Confusion Matrix

- The decision tree matrix shows that the test accuracy was the worst performer of the four with the others being at 83.33%.
- This is shown in the matrix with the one false negative for the decision tree.

Logistic Regression

SVM

Decision Tree



KNN

Conclusions

- Several factors can lead to the success of a mission such as the launch site, the orbit, and to a great degree the number of previous launches. Undeniably, I can assume that knowledge has been gained between launches which allowed launch failure to turn to launch success.
- The orbits with the best success rates were GEO, HEO, SSO, ES-L1.
- Taking into account the orbits, payload mass can play a role in the success of a mission. For example, some orbits require a light or heavy payload mass, though generally light payloads perform better than heavy payloads.
- I am unable to explain why some launch site were better than others (KSC LC-39A being the best launch site) with the current data at hand. To answer this question, I could gather atmospheric or other relevant data.
- For this dataset, I choose the Decision Tree Model as the best model even though the test accuracy was lower than the other models. I choose the Decision Tree Model because it had a higher train accuracy.

Thank you!