IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Efran Himel
12-15-2022

Outline

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

Executive Summary

- To determine the success of Class 1 SpaceX rocket launches, I pulled data from public sources (SpaceX, Wikipedia). Then I conducted an initial exploratory analysis of the extracted data to gain a deeper understanding of it. This is done cleaning it, producing initial statistic metrics, and creating dashboards. Finally, the data is fed into various Classification ML models to predict which features had the highest chance of a successful launch.
- After producing several ML models, they all performed equally well producing a score of 83.333 (except for the tree model) so the Logistic Regression model was chosen at the end due to its superior interpretability over the other models. The top 10 features from the model that had the highest impact on launch success were [Serial_B1003,Serial_B1015,Serial_B1016,Serial_B0003,Serial_B1023,Orbit_HEO,Serial_B1038,Serial_B1026,Serial_B1012,LandingPad_5e9e3032383ecb761634e7cb]

Introduction

- In this capstone, I predicted 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. Therefore, if I can determine if the first stage will land, I can determine the cost of a launch. This information can be used if an alternate company (SpaceY) wants to bid against SpaceX for a rocket launch

Section 1

Methodology

Methodology

Executive Summary

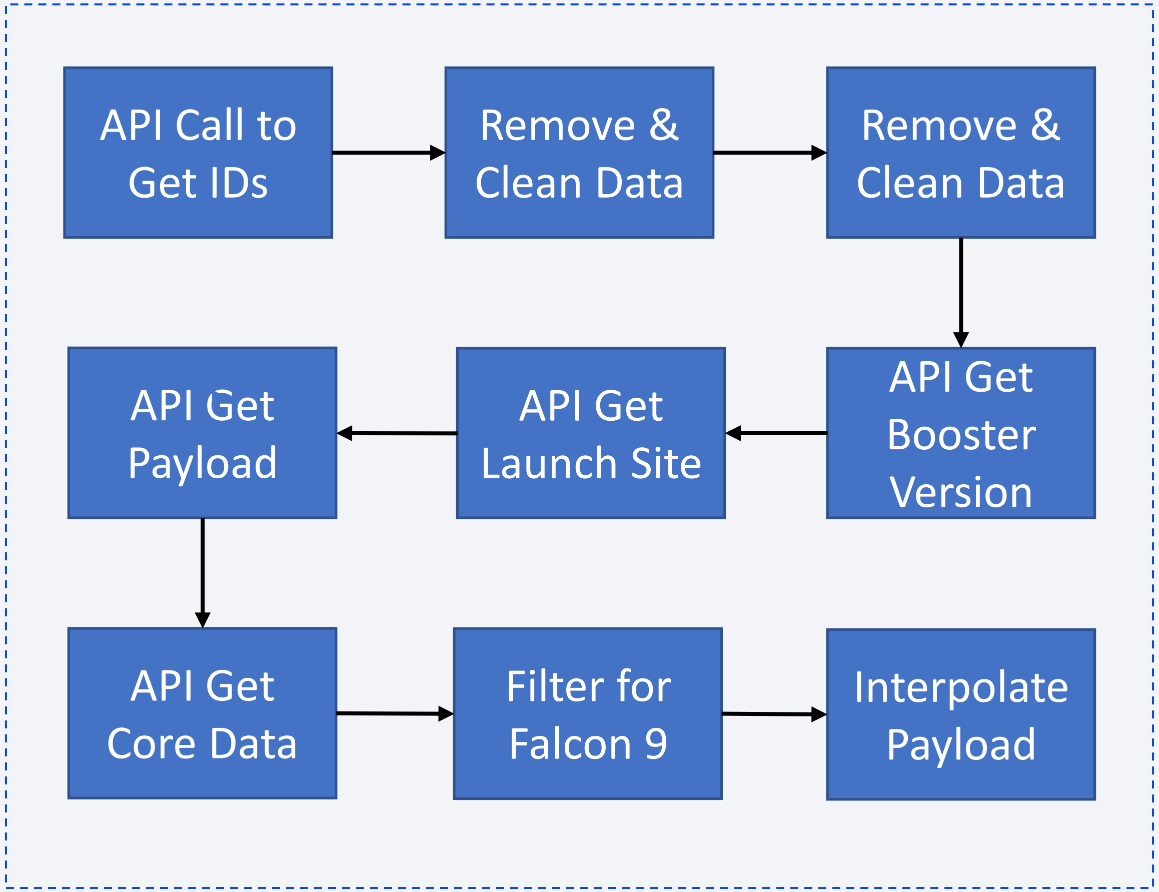
- Data collection methodology:
  - Public data was used for this project. It was collected from the SpaceX API and SpaceX wiki page using request & beatifulsoup4 library
- Perform data wrangling
  - The data was wrangled by filtering irrelevant fields, reformatting strings, and interpolating missing values based on existing data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Sklearn and GridSearch were used to for classification modeling and tuning

Data Collection

- How data sets were collected.
  - The data was collected from the SpaceX API using the python requests library and web scraping the SpaceX Wikipedia page using requests API and beautifulsoup4 libraries
- Data collection process
  - Space Data Collection: Make API call to get IDs of SpaceX flights. Use IDs to make API calls to retrieve ['rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc'] data. With this dataset, filter for only Falcon 9 rockets. Interpolate any missing values for ‘payloads’ by replacing with the mean of ‘payloads’
  - Wikipedia Data Collection: Use beautifulsoup4 to extract html data from SpaceX wiki page to retrieve data from table regarding launch history. After parsing through html, create dataframe of data. Data included the following columns extracted from Wiki ['Flight No.’, 'Launch site’, 'Payload’, 'Payload mass’, 'Orbit’, 'Customer’, 'Launch outcome’, 'Version Booster’, 'Booster landing’, 'Date’, 'Time’,]

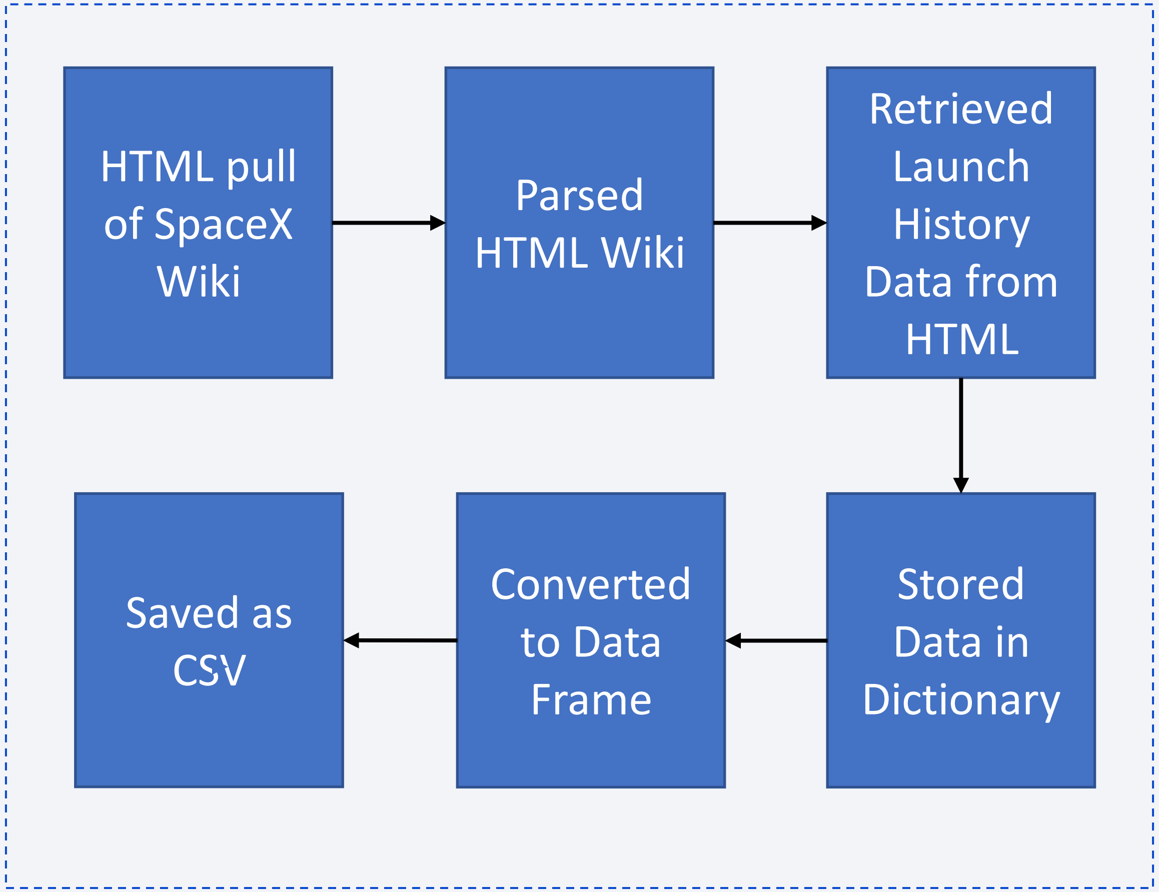
Data Collection – SpaceX API

- Called API to get list of launches & IDS
- Used IDs and made calls to https://api.spacexdata.com/v4/launchpads/ with launch IDs to get [booster version, launch site, payload, core data]
- Filter for Falcon 9
- Payload had missing data so used mean to interpolate missing data
- GitHub Code Link



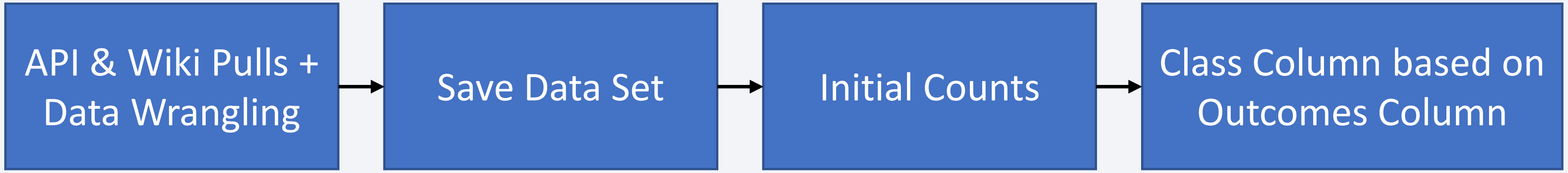
Data Collection - Scraping

- Used beautifulsoup4 & requests library to extract html data from [Wiki]
- Parsed through Wiki html to retrieve data from launch history table
- Converted data into a dictionary. Cleaned some data during conversion process (converting to string, making more readable)
- Converted dictionary into pandas data frame and saved as csv
- [GitHub Code Link]



Data Wrangling

- Initial Data processing was conducting while pulling data from sources (see info and notebook links from slides 8 & 9 for more details)
- After the dataset was extracted from the SpaceX API and Wikipedia, additional data processing was conducted.
  - A few initial calculations were conducted.
    - Conducted value counts for [Launch Site, Orbit, Outcome]
    - Based on landing outcome, created a 'Class' column where 0 = bad outcome and 1 = otherwise
- Flow Chart


API & Wiki Pulls + Data Wrangling → Save Data Set → Initial Counts → Class Column based on Outcomes Column

- GitHub Link

EDA with Data Visualization

- For EDA several scatter plots were created, one bar chart was created, and one line chart was created
  - Scatter Plots: Scatter plots were used to understand the relationship between two variables with each other and used Class to distinguish between the launch failures and success. I compared (Flight Number, Launch Site), (Payload, Launch Site), (Flight Number, Orbit), (Payload, Orbit)
  - Bar Chart: The bar chart was used to understand the success rate of each orbit. ES-L1, GEO, HEO, SSO, VLEO had the highest success rate of 100%
  - Line Chart: The line chart was used to understand how the success rate of launches changed over time by year. From 2010 to 2020 there was a positive updard trend from 0% success rate in 2010 to about 80% in 2020
- GitHub Link

EDA with SQL

- SQL Queries were used to get a general idea various aspects of the data. The Following Queries were conducted.
  - 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.
  - Continue to Next Slide
- GitHub Link

EDA with SQL

- SQL Queries were used to get a general idea various aspects of the data. The Following Queries were conducted.
  - 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.
  - List the failed landing outcomes in drone ship, their booster versions, and launch site names for 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
- GitHub Link

Build an Interactive Map with Folium

- Folium was used to visualize the location of SpaceX launches for their Falcon 9 rockets. Map markers (circles & markers) were placed on a map to visualize the launch sites. Markers were also added for successful & failed launched
- This was done to better visualize the locations of successful launches. This is important data for better understanding the geographic conditions of successful launches.
- GitHub Link

Build a Dashboard with Plotly Dash

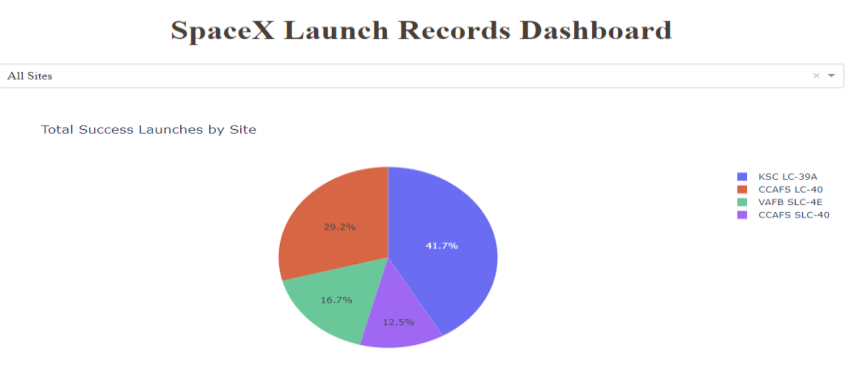
- Another step in the EDA was creating a Plotly dashboard. I created a dashboard that included:
  - A Drop down for all or each individual launch site. This was to filter the data and see if there's a relationship between launch success and site.
  - A Pie chart showing launch success. This helped visualize the results of the down filter.
  - A Slider to indicate launch payload range which helps understand if payload affected launch success.
  - A scatter plot indicating the success of launches based on payload mass and categorized by booster version,
- GitHub Link

Predictive Analysis (Classification)

- After prepping the data and conducting the initial analysis I created several ML classification models. I vectorized the X & Y data, normalized X, and split the data between train & test set. I used gridsearch to find the best parameters for logistic regressions, support vector machine, decision tree, and knn models. I compared their performance with a confusion matrix & best score method in sklearn. All the models except the decision tree got a final score of on the test set 0.8333. Decision tree scored 0.8888. I chose the tree model as the best model because it had the highest predictive score
- GitHub Link

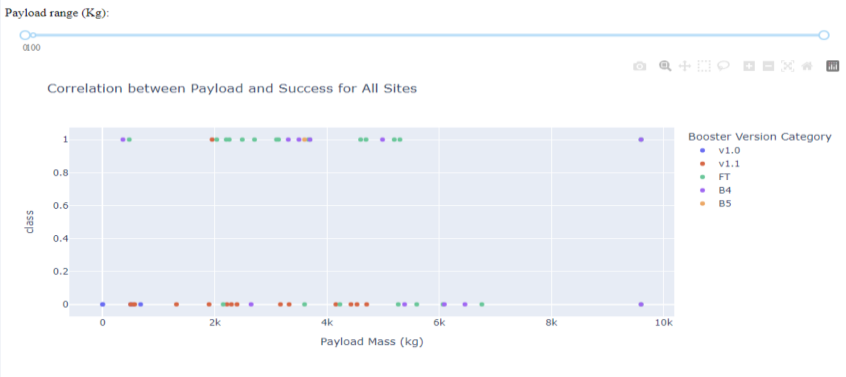
Results

- Exploratory data analysis results & Interactive analytics demo in screenshots



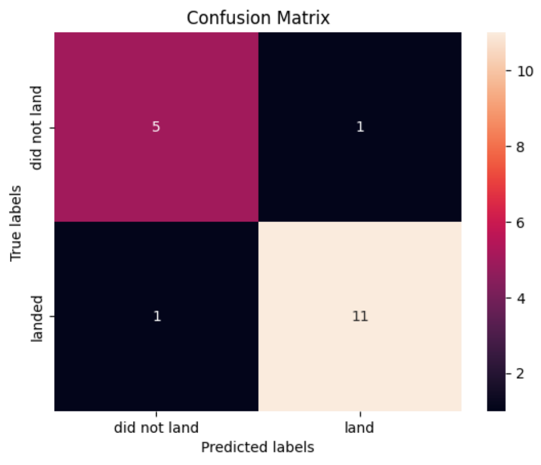
SpaceX Launch Records Dashboard

All Sites



Payload range (Kg):

- Predictive analysis results: Tree model scored highest predictive score on test set with 88.89% accuracy

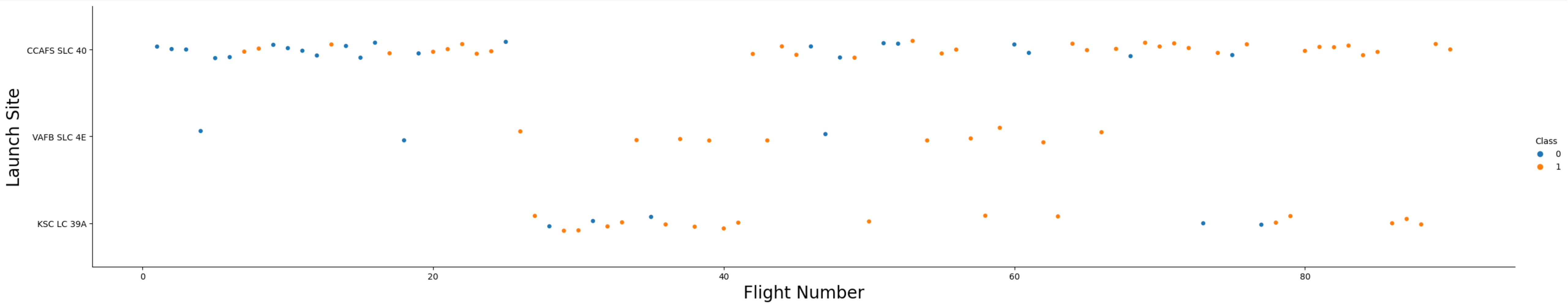


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

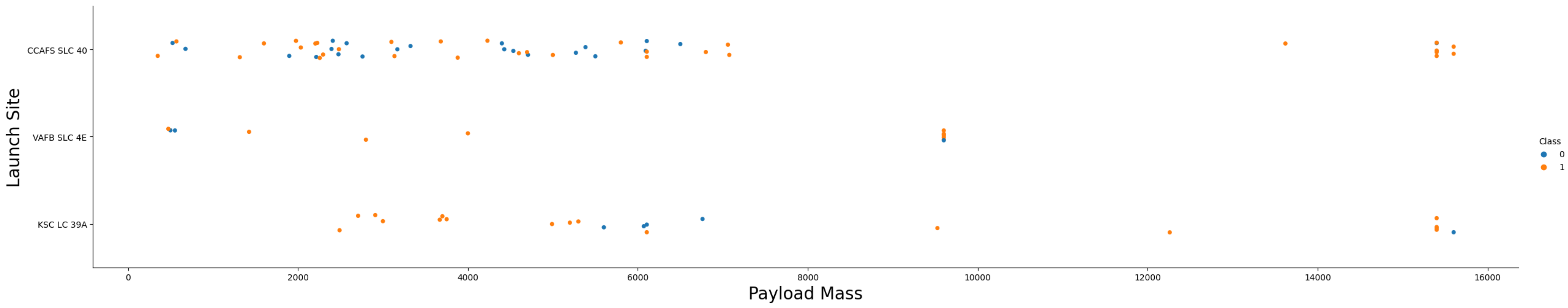
- Show a scatter plot of Flight Number vs. Launch Site



- CCAFS SLC 40 Had the greatest number of launches, and it seems the greatest number of successful flights as well. KSC LC 39A had the fewer launches but proportionally the most successful launches.

Payload vs. Launch Site

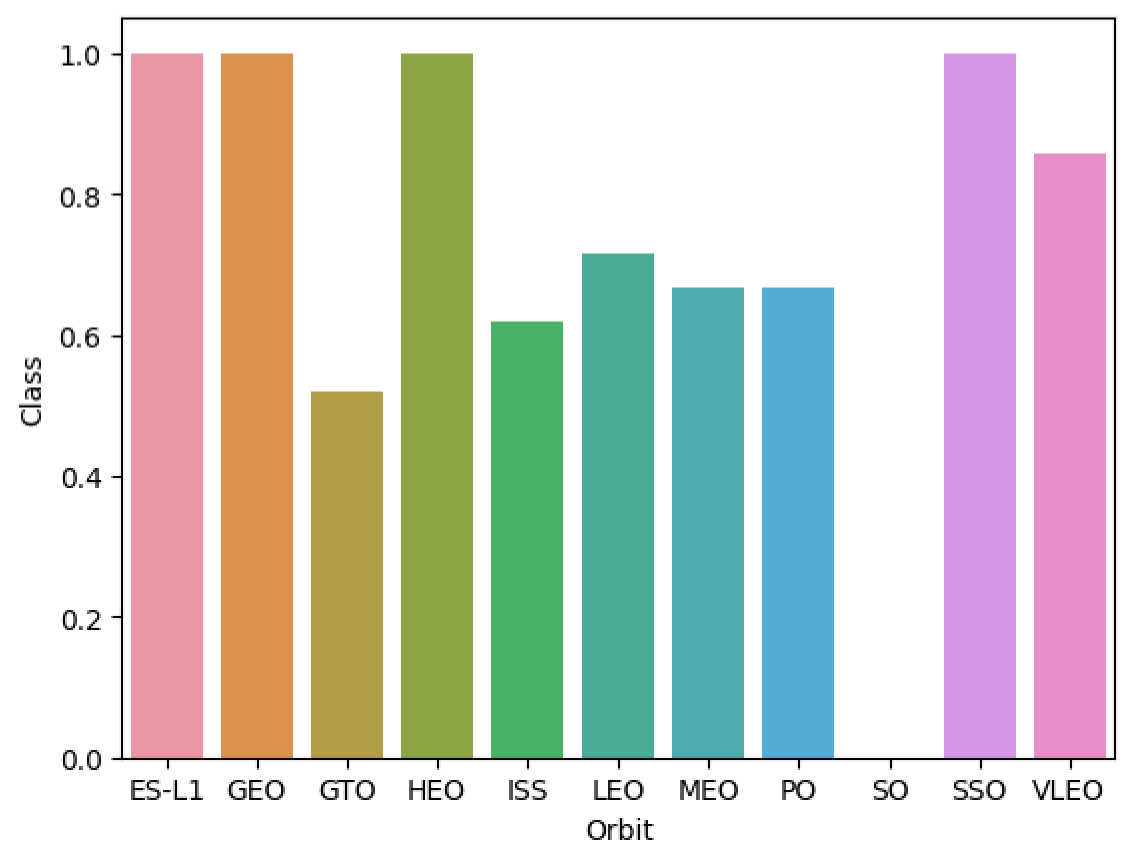
- Show a scatter plot of Payload vs. Launch Site



- Even though most launches are at lower payloads, there’s a proportionally higher number of successful launches at higher payloads.
- CCASFS SLC 40 & KSC LC 39A both seem to be able to handle a wide variety of payloads while maintaining successful launches.

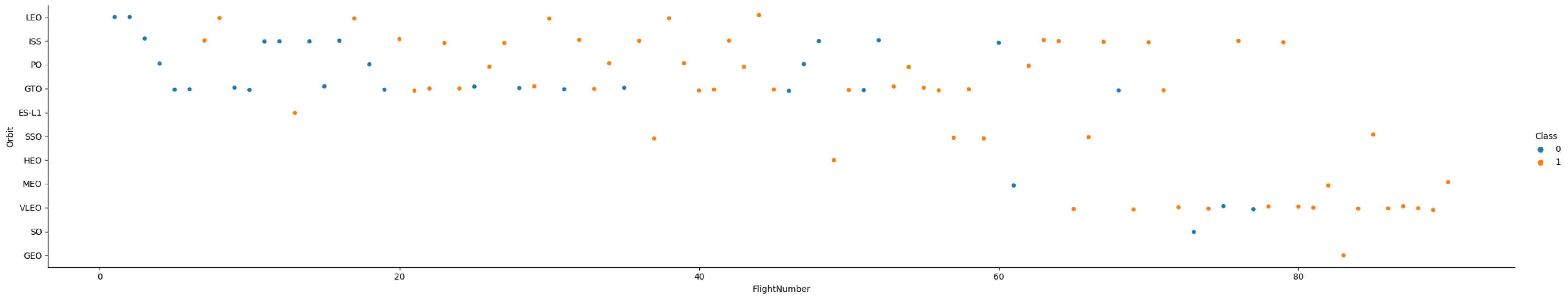
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- Based on the bar chart, orbits [ES-L1, GEO, HEO, SSO] have the highest success rate.
- SO Has the lowest success rate



Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

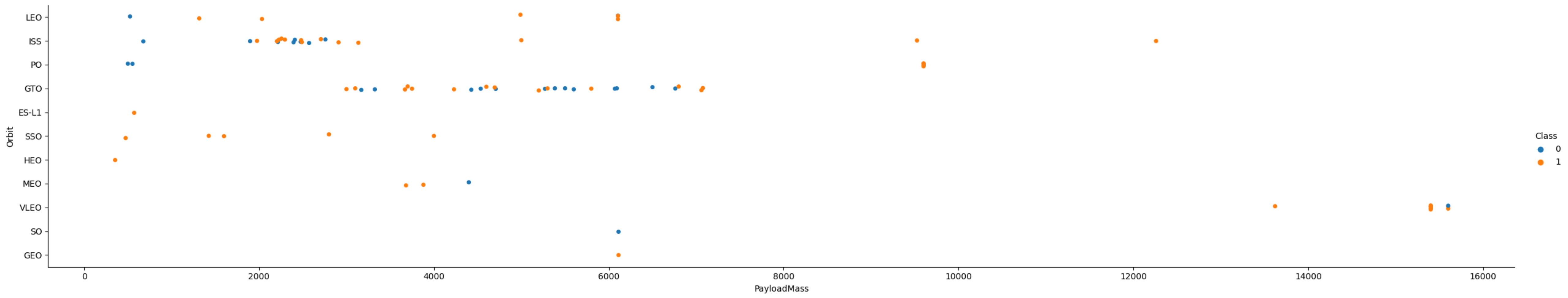


- Even though [ES-L1, GEO, HEO, SSO] had the highest success rate in the bar chart, the scatter plot shows that they had few launches in total, so fewer potential chances of failure.

- VLEO had the second highest success rate and a good number of launches, so it may be a better candidate.

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

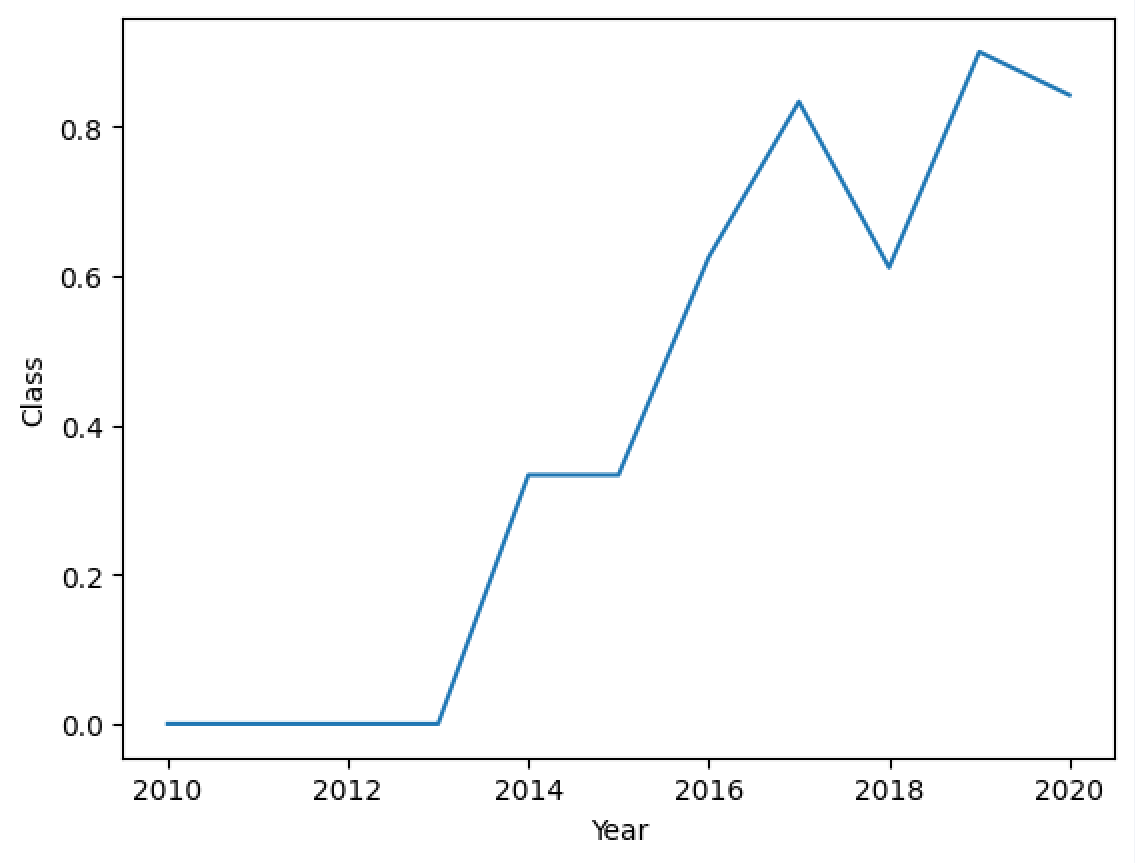


- Most the data is concentrated in the upper left corner of the plot. Seems like most launches are focused on lower orbit lower weight payloads vs high orbit high weight payloads.
- Higher weight payloads seems to be more successful. Potentially do to practice/experience from lower weight payloads?

Launch Success Yearly Trend

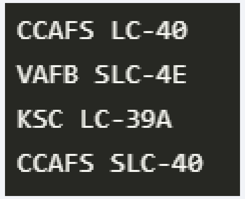
- Show a line chart of yearly average success rate

- Over the years, the success rate of successful launches has gone up from 2010 to 2020



All Launch Site Names

- Find the names of the unique launch sites

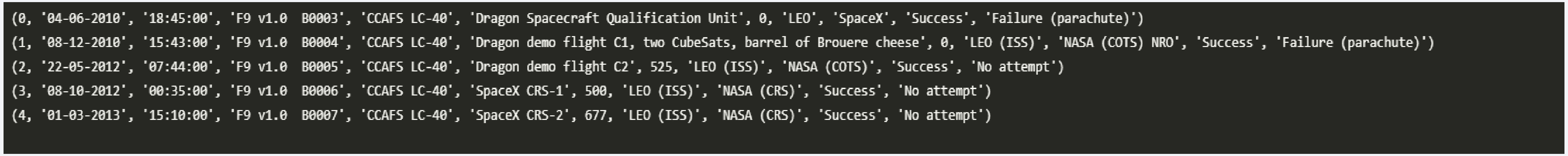


```
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
```

- Present your query result with a short explanation here
  - Query: SELECT DISTINCT Launch_Site from spacex
  - Explanation: selects the distinct values from the spacex table

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`



```
(0, '04-06-2010', '18:45:00', 'F9 v1.0  B0003', 'CCAFS LC-40', 'Dragon Spacecraft Qualification Unit', 0, 'LEO', 'SpaceX', 'Success', 'Failure (parachute)')
(1, '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)')
(2, '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')
(3, '08-10-2012', '00:35:00', 'F9 v1.0  B0006', 'CCAFS LC-40', 'SpaceX CRS-1', 500, 'LEO (ISS)', 'NASA (CRS)', 'Success', 'No attempt')
(4, '01-03-2013', '15:10:00', 'F9 v1.0  B0007', 'CCAFS LC-40', 'SpaceX CRS-2', 677, 'LEO (ISS)', 'NASA (CRS)', 'Success', 'No attempt')
```

- Present your query result with a short explanation here
  - Query: SELECT * FROM spacex WHERE Launch_Site LIKE "CCA%" LIMIT 5
  - Explanation: selects all records from spacex table based on a filter then limits to 5 results

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - 45596
- Present your query result with a short explanation here
  - Query: SELECT SUM(PAYLOAD_MASS__KG_) FROM spacex WHERE Customer = "NASA (CRS)“
  - Explanation: Sums payload column from spacex table after a filter has been applied that only uses NASA as the customer

Average Payload Mass by F9 v1.1

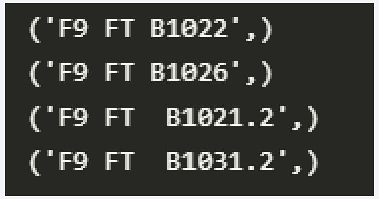
- Calculate the average payload mass carried by booster version F9 v1.1
    - 2928.4
- Present your query result with a short explanation here
    - Query: SELECT AVG(PAYLOAD_MASS__KG_) FROM spacex WHERE Booster_Version = "F9 v1.1“
    - Explanation: Averages payload column after the spacex table filtered the data by booster version

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - 02-03-2019
- Present your query result with a short explanation here
  - Query: SELECT MIN(Date) FROM spacex WHERE "Landing _Outcome" = "Success“
  - Explanation: Filters data for successful landing outcome then uses min to pick the earliest date

Successful Drone Ship Landing with Payload between 4000 and 6000

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

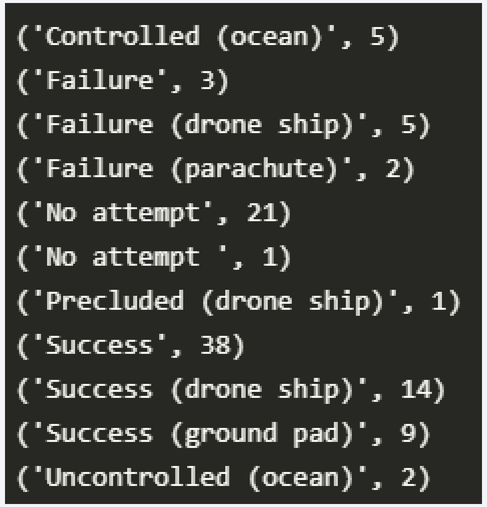


```
('F9 FT B1022',)
('F9 FT B1026',)
('F9 FT  B1021.2',)
('F9 FT  B1031.2',)
```

- Present your query result with a short explanation here
  - Query: SELECT Booster_Version FROM spacex WHERE "Landing _Outcome" = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
  - Explanation: Filtered by landing outcome and payload size then select the booster column

Total Number of Successful and Failure Mission Outcomes

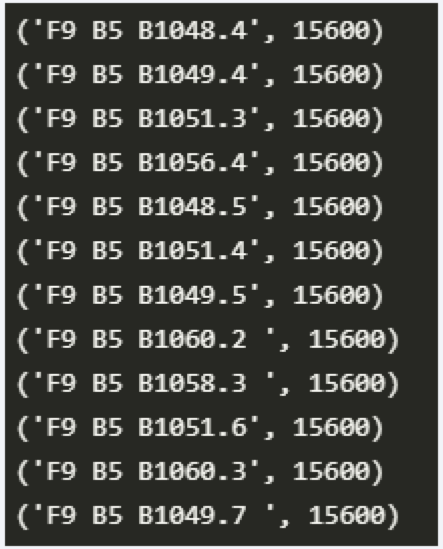
- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here
  - Query: SELECT "Landing _Outcome", COUNT("Landing _Outcome") FROM spacex GROUP BY "Landing _Outcome" ORDER BY "Landing _Outcome"
  - Explanation: Select landing outcome and count of landing outcome from spacex table. Then use group by landing outcome to get the count of success & failures. Order by was used to keep output organized



```
('Controlled (ocean)', 5)
('Failure', 3)
('Failure (drone ship)', 5)
('Failure (parachute)', 2)
('No attempt', 21)
('No attempt ', 1)
('Precluded (drone ship)', 1)
('Success', 38)
('Success (drone ship)', 14)
('Success (ground pad)', 9)
('Uncontrolled (ocean)', 2)
```

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here
  - Query: SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM spacex WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM spacex)
  - Explanation: Used sub query to get max payload then used another query to get all booster versions that had that max payload from the sub query



```
('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)
```

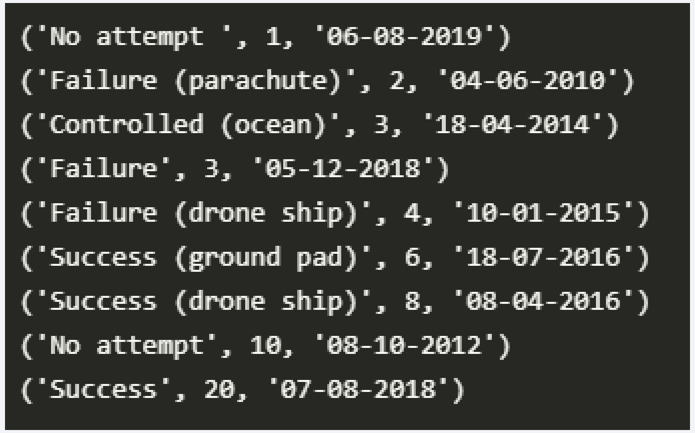
2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - ('Failure (drone ship)', 'F9 v1.1 B1012', 'CCAFS LC-40', '10-01-2015')
  - ('Failure (drone ship)', 'F9 v1.1 B1015', 'CCAFS LC-40', '14-04-2015')
- Present your query result with a short explanation here
  - Query: SELECT "Landing _Outcome", Booster_Version, Launch_Site, Date FROM spacex WHERE "Landing _Outcome" = "Failure (drone ship)" AND Date LIKE "%2015“
  - Explanation: Used where to filter landing outcome and date and then selected all relevant columns for query

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

- Present your query result with a short explanation here
    - Query: SELECT "Landing _Outcome", COUNT("Landing _Outcome") as c, Date FROM spacex WHERE Date >= '04-06-2010' AND Date <= '20-03-2017’ GROUP BY "Landing _Outcome" ORDER BY "c“
    - Explanation: Used Where to get current date range. Selected relevant columns and group by to count outcomes for each category. Used order by to then rank the outcomes



```
('No attempt ', 1, '06-08-2019')
('Failure (parachute)', 2, '04-06-2010')
('Controlled (ocean)', 3, '18-04-2014')
('Failure', 3, '05-12-2018')
('Failure (drone ship)', 4, '10-01-2015')
('Success (ground pad)', 6, '18-07-2016')
('Success (drone ship)', 8, '08-04-2016')
('No attempt', 10, '08-10-2012')
('Success', 20, '07-08-2018')
```

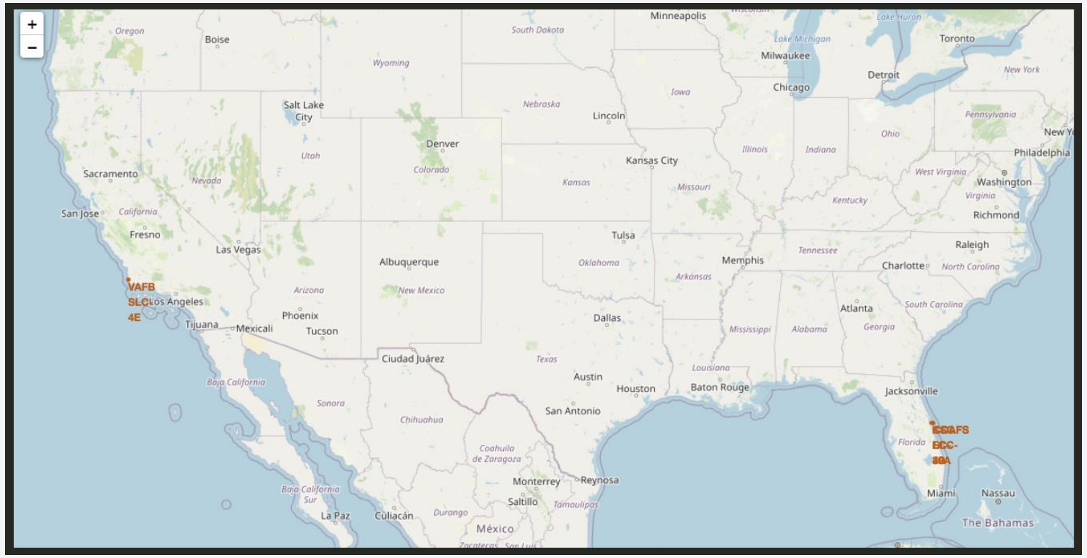
Section 3

Launch Sites Proximities Analysis

All Launch Sites on a Map

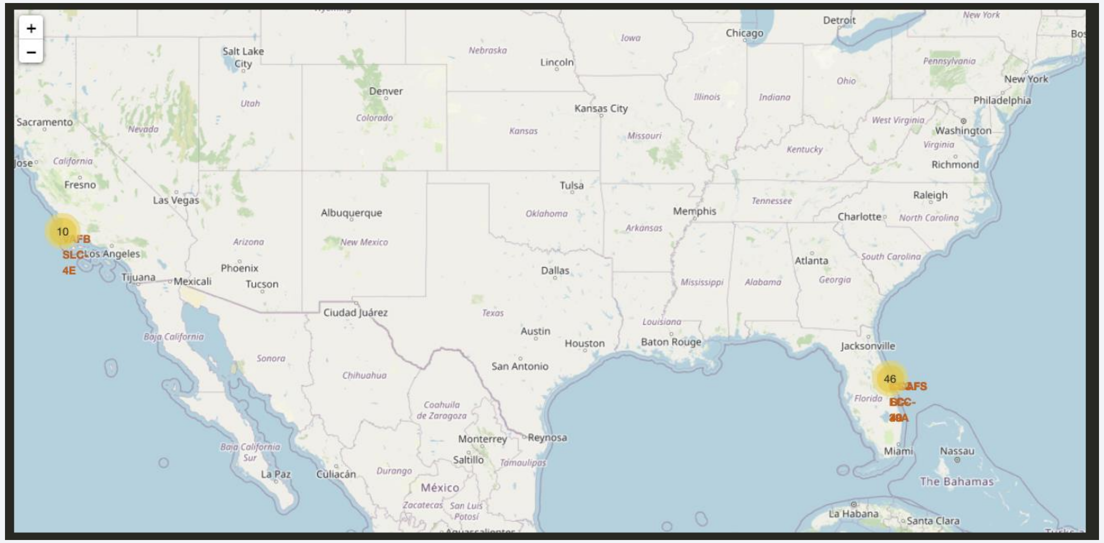
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map

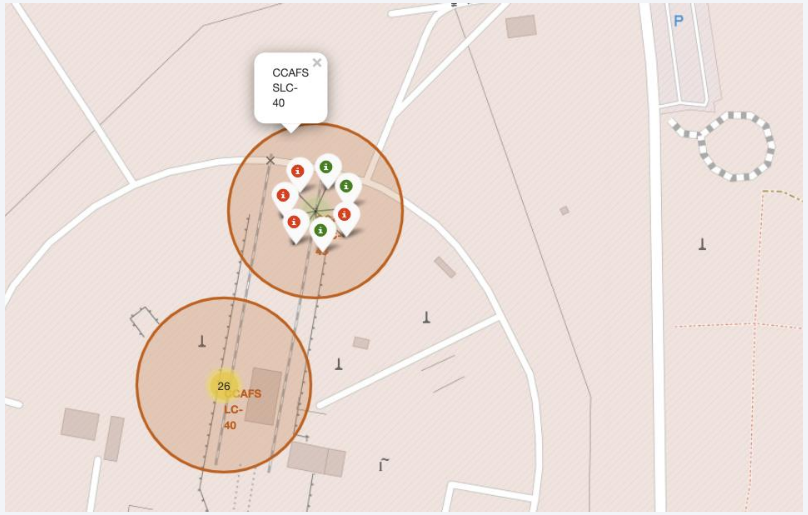
- It seems the important launch sites on concentrated in a few states (CA and FL). This may be because the states have ideal launch conditions for rockets.

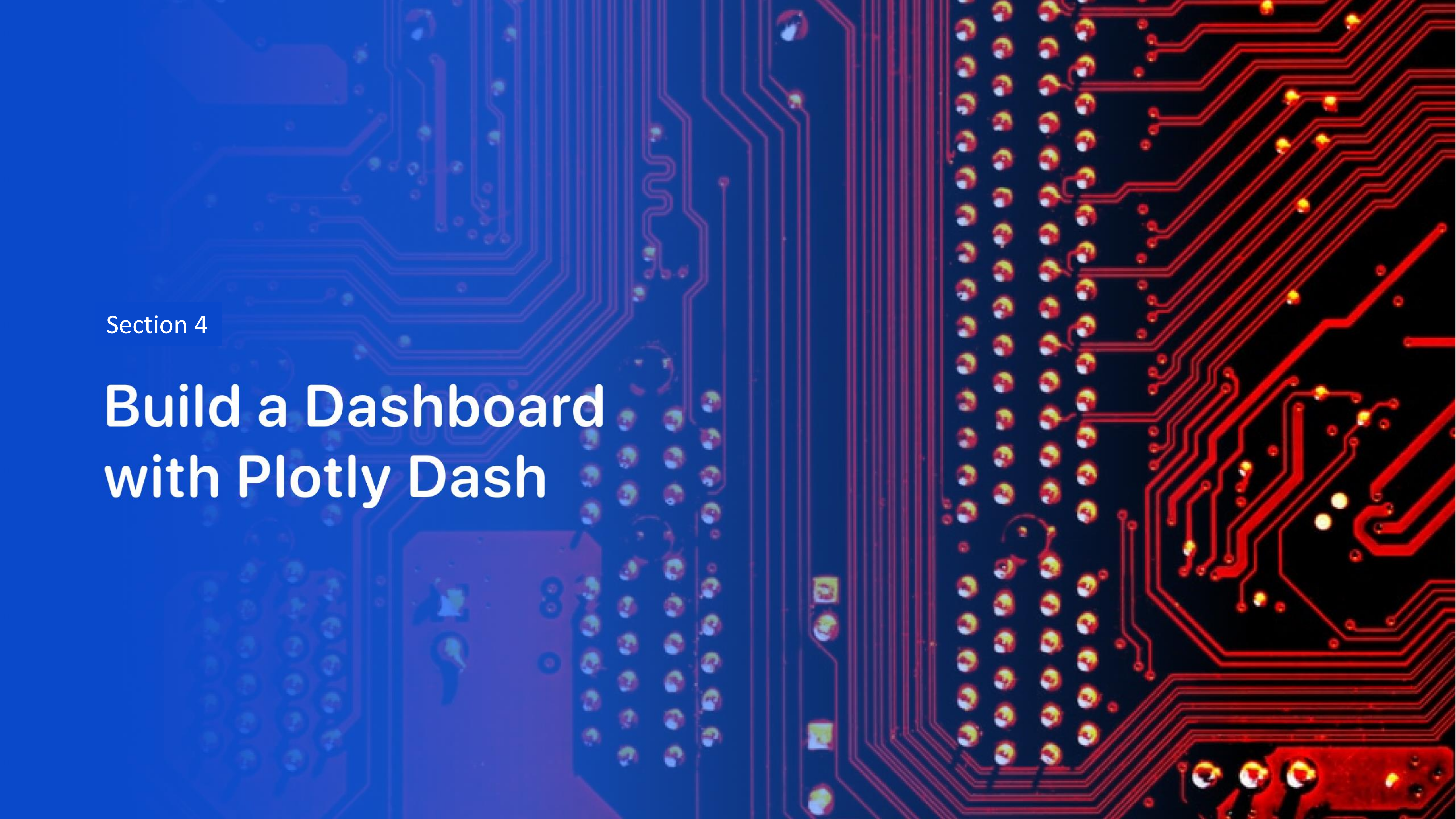


Successful & Failed Launches per Site

- This screenshot indicates where there were more successful launches. It seems Florida had more launches and more successful launches there than California launch locations





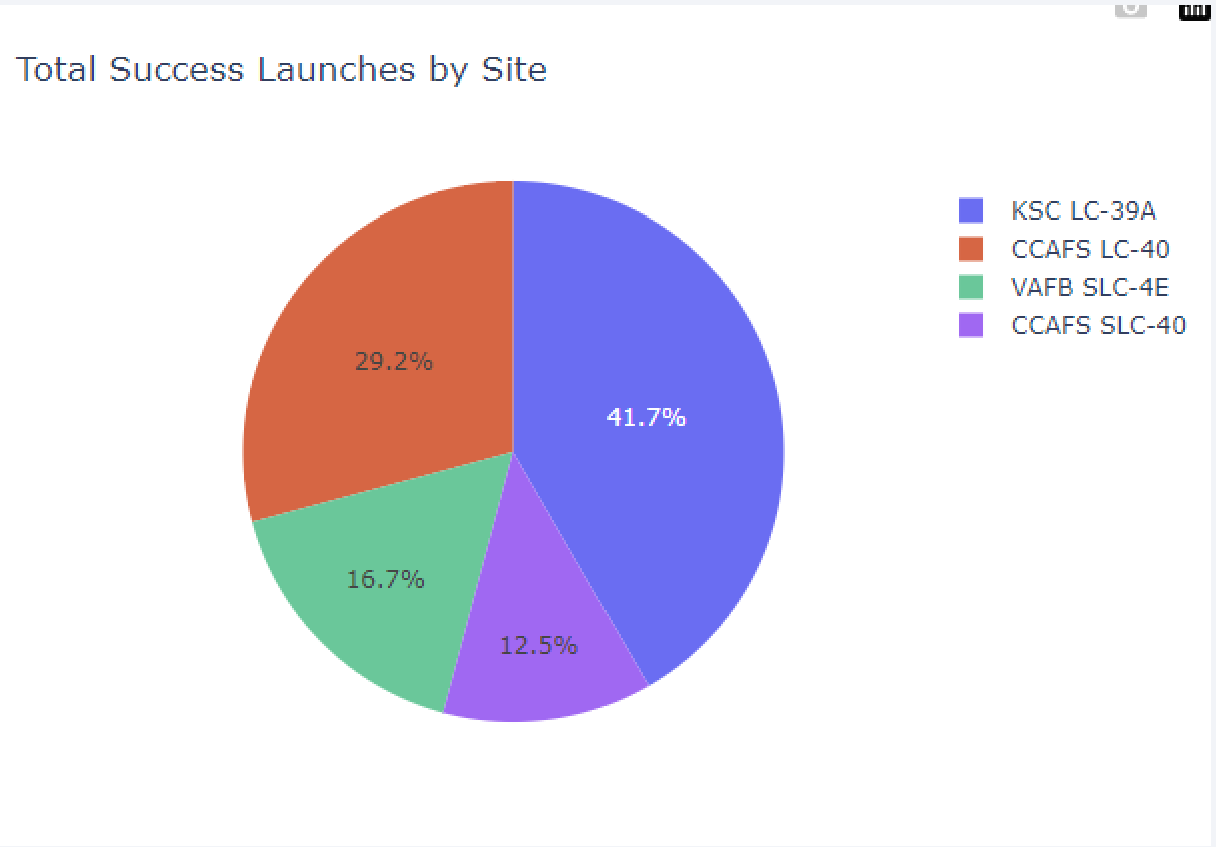


Section 4

Build a Dashboard with Plotly Dash

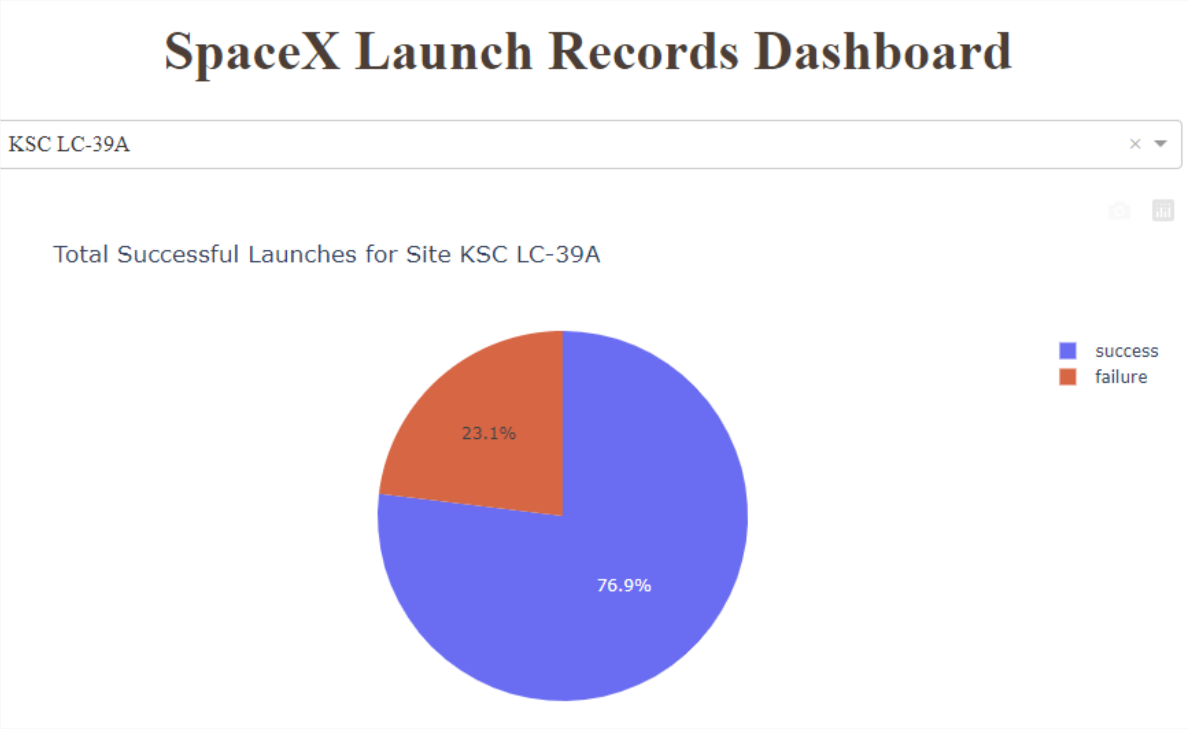
Total Success Launches by Site

- This screenshot shows that KSC LC-39A & CCAFS LC-40 had the highest portion of launches from all the 4 launch sites.
- KSC LC-39A had the greatest number of launches in total
- CCAFS LC-40 had the second most



Launch Site with Highest Success Rate

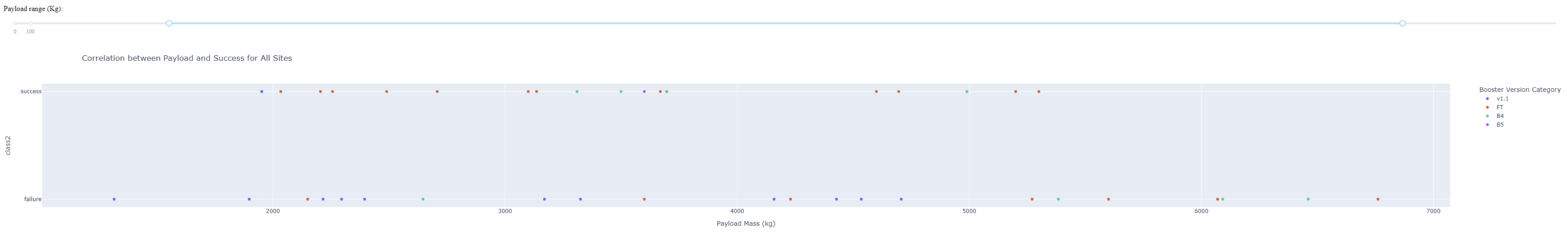
- Out of all the launch sites, KSC LC-39A had the highest success rate out of the 4 launch locations.



SpaceX Launch Records Dashboard

KSC LC-39A

Payload vs. Launch Outcome for All Sites (1000 - 7000)



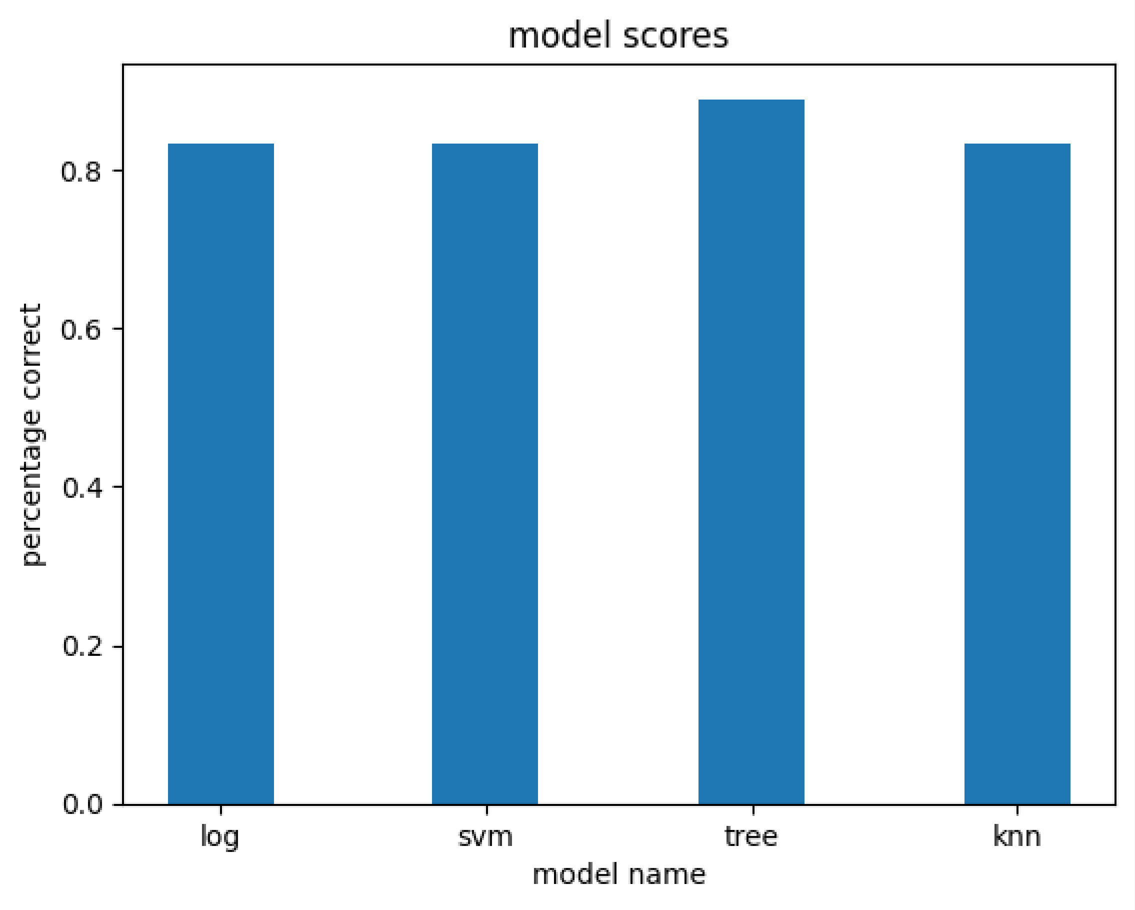
- In the payload range of 1000kg to 7000kg there were more failures than successes. In this range, booster version FT seems to have the highest success rate.

Section 5

Predictive Analysis (Classification)

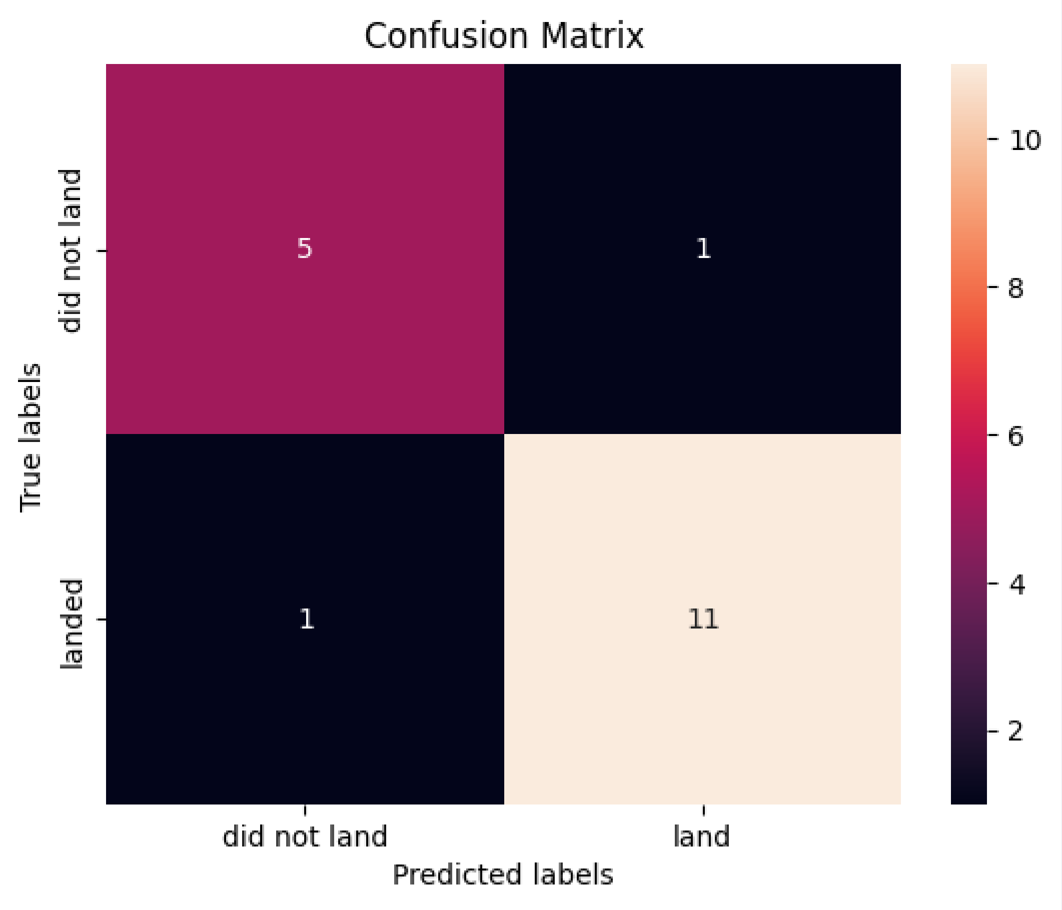
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- Out of all 4 models, the tree model had the highest classification score of 0.88888



Confusion Matrix

- This is the confusion for the tree model which scored the best.
- Its FP and FN scores are the same.
- It guessed mostly correct with TP and TN



Conclusions

- Higher payloads tend to be more successful compared to lower ones.
- Launches have been becoming more successful over time. KSC LC-39A seems to be the best launch site.
- Florida seems to be better location to launch rockets.
- For predicting launch outcomes, a tree-based model would be the best solution.

Thank you!