Winning Space Race with Data Science

Nyv MONDELE MBOLA
04/26/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- 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 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.
- Using historical data, we will predict if the Falcon 9 first stage will land successfully for new launches given some known factors.

Section 1

Methodology

Methodology

- Data collection using requests to SpaceX API and Web Scraping
- Performed data wrangling to clean the data
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Built, tuned, and evaluated classification models

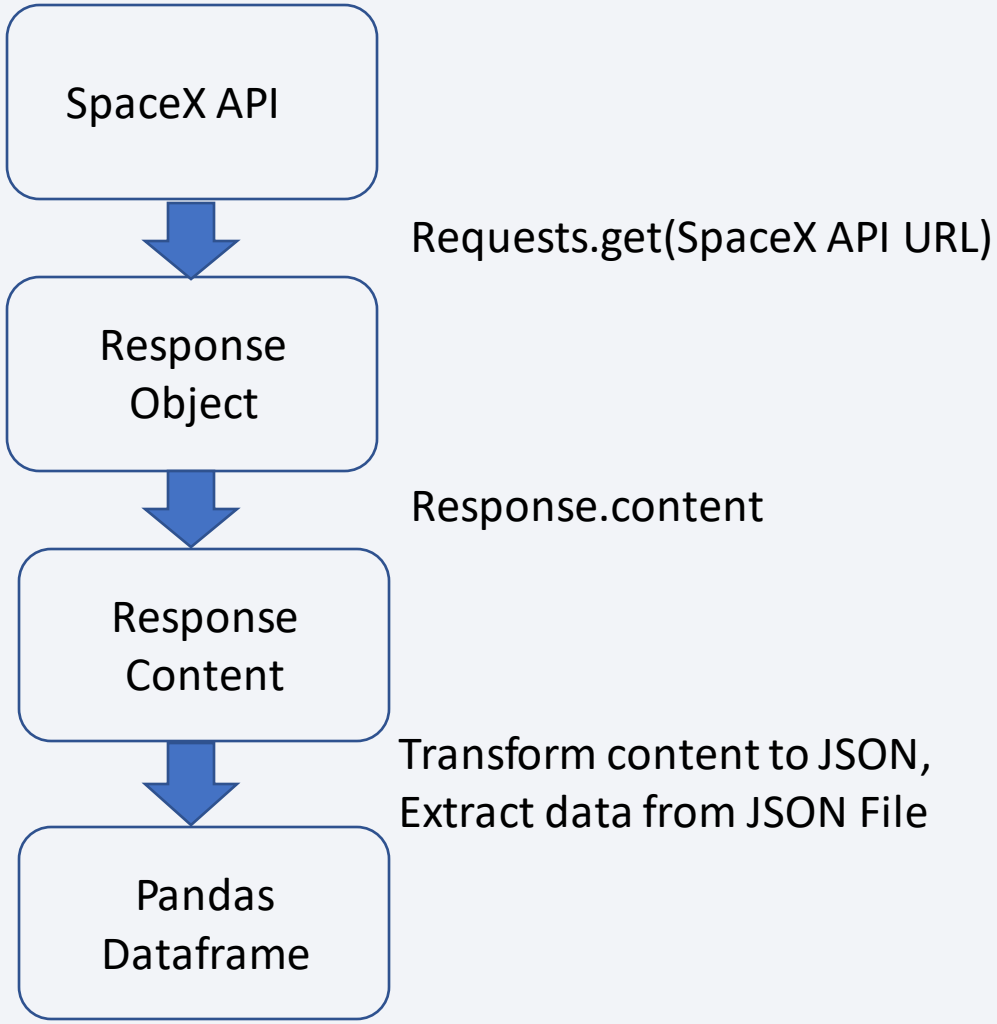
Data Collection

- Data was collected using a combination of Requests to SpaceX API and Web Scraping.

- For each launch, features of Interest included : BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

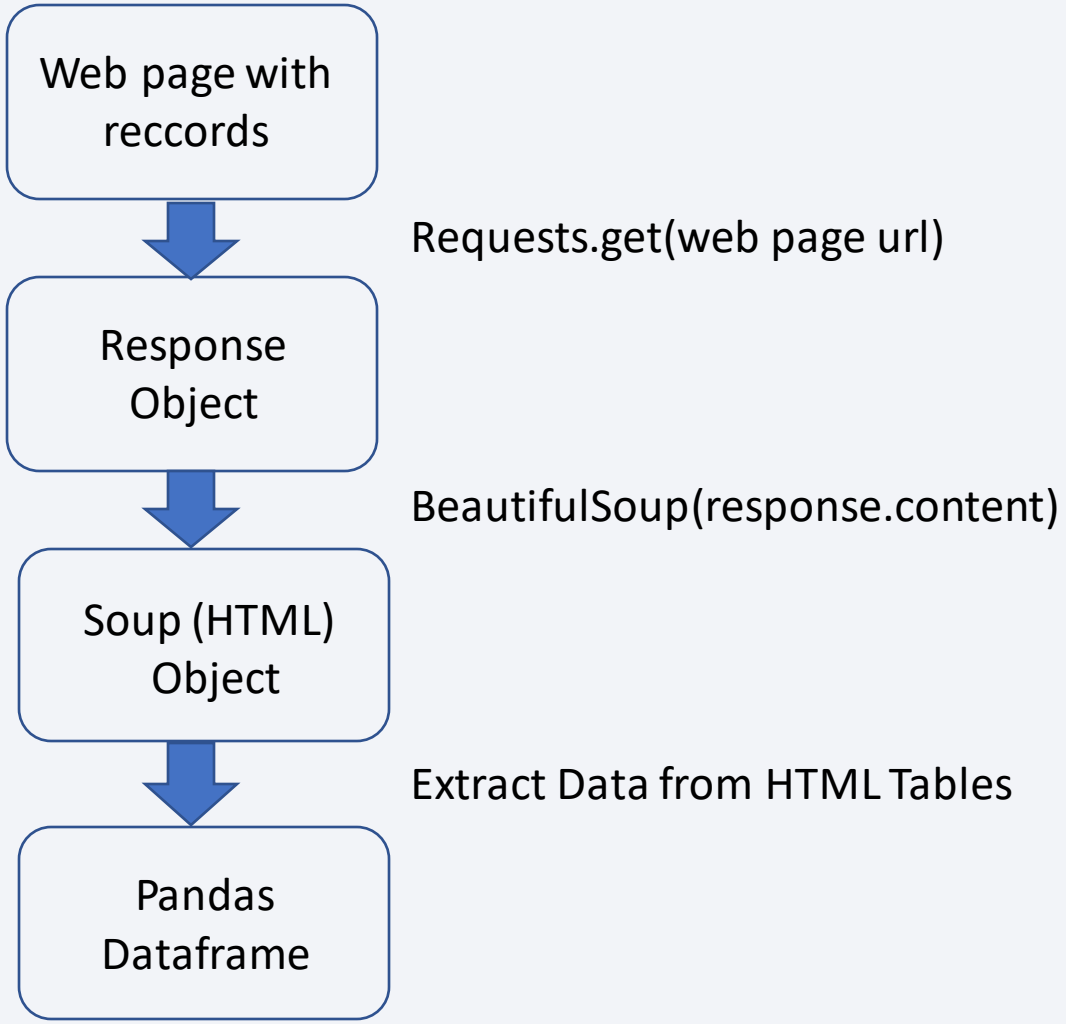
Data Collection – SpaceX API

- The Requests module was used to make HTTP requests to get data from SpaceX API.
-
- The API's Response Content was decoded into JSON and all desired data were extracted then stored in a Pandas DataFrame
- The GitHub URL of the completed SpaceX API calls notebook



Data Collection - Scraping

- The BeautifulSoup from bs4 module was used to scrape a Wikipedia HTML page containing Falcon 9 launch records In form of tables.
- The tables were parsed and converted it into a Pandas data frame
- The **GitHub URL** of the completed web scraping notebook.



Data Wrangling

- We performed Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- NaN in PayloadMass column were replaced with the average value of all PayloadMasses
- We Created a "landing class" column from Outcome column where each successful outcome had a value of 1 and each unsuccessful outcome a value of 0
- The **GitHub URL** of the completed data wrangling related notebooks,.

EDA with Data Visualization

- We used visualization to see the relationship between the features and their impact on the landing outcome.
- For example, scatter plot for Launch Site vs Flight Number, Bar plot for Obits vs Success rate, Line plot for yearly trends in landing outcomes
- The **GitHub URL** of the completed EDA with data visualization notebook.

EDA with SQL

- We performed all SQL queries from the notebook. Tasks included:
  - Display the names of the unique launch sites
  - 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, using a subquery
  - 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
- The GitHub URL of the completed EDA with SQL notebook.

Build an Interactive Map with Folium

- After creating a map to display all the launch sites, we created Circles, Markers, Marker Clusters, Lines and added them to the map
- Circles were used to show launch sites locations, Markers were used to label the points on the map, lines were used to connect two given locations, and clusters were used to group launches made on the same sites or on closer sites and to set their colors based on the landing outcome.
- The **GitHub URL** of the completed interactive map with Folium map.

Build a Dashboard with Plotly Dash

- For figures, Pie chart, and Scatter plot, were used, while Dropdown and Range slider were used to provide interactions with the graphs.
- The pie chart was used to display the success/failure proportion, and using a dropdown menu, the pie chart could be set to display for individual launch site or for all sites combined. In the latter case, site were compared in terms of success rate.
- A scatter plot was used to visualize the relationship between the Payload Mass and the landing outcome, and the impact of the booster version. A range slider was used to select different payload masses.
- The **GitHub URL** of the completed Plotly Dash lab.

Predictive Analysis (Classification)

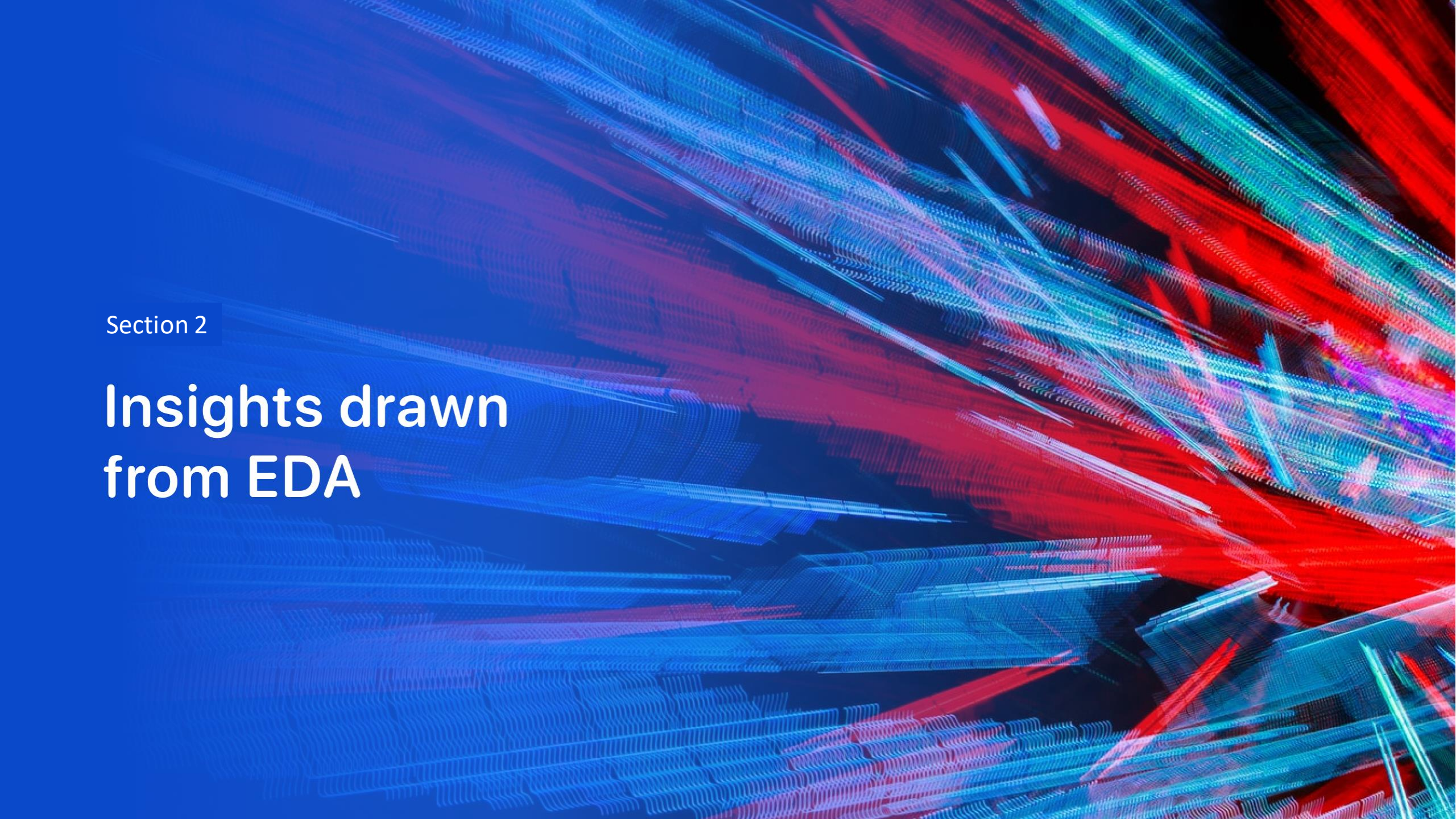
- The features were assigned to the variable X and Standardized, and the target variable was assigned to Y.
- We split the data into training and testing data using the function `train_test_split`. The models, namely KNN, SVM, Decision Tree, and Logistic Regression were trained and hyperparameters were selected using the function `GridSearchCV`.
- The accuracy was calculated on the test data using the method `.score()` and the confusion matrix plotted.
- The best model was selected based on the result from the score.
- The **GitHub URL** of the completed predictive analysis lab.

Results

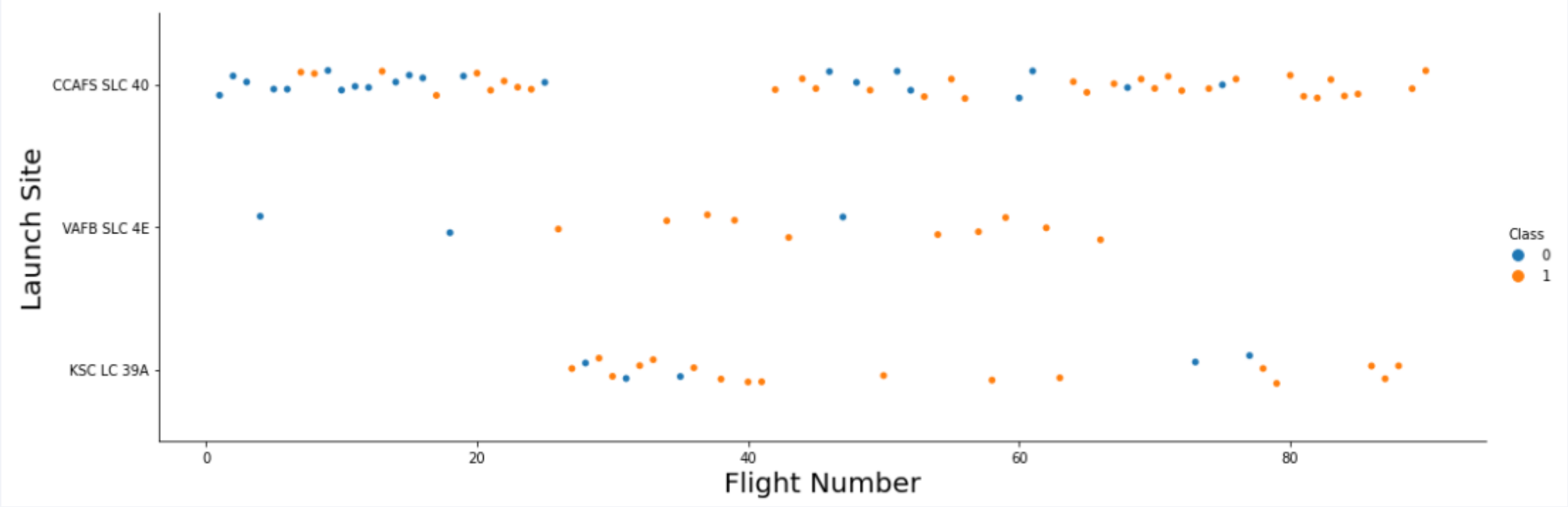
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

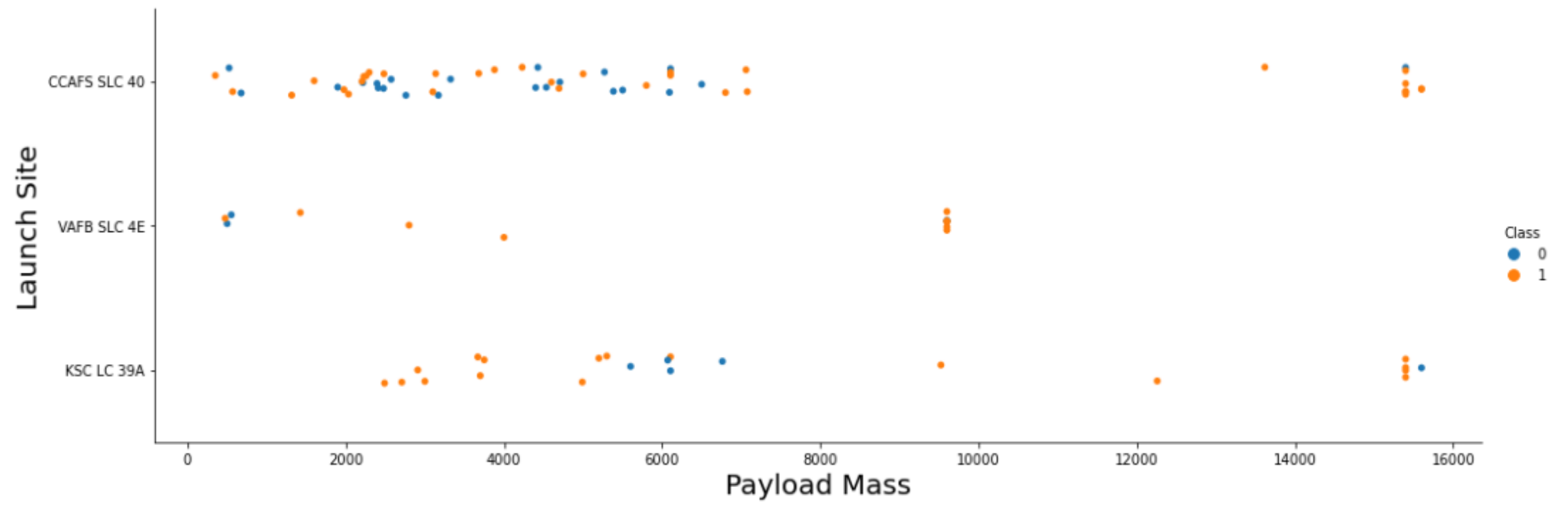


Flight Number vs. Launch Site



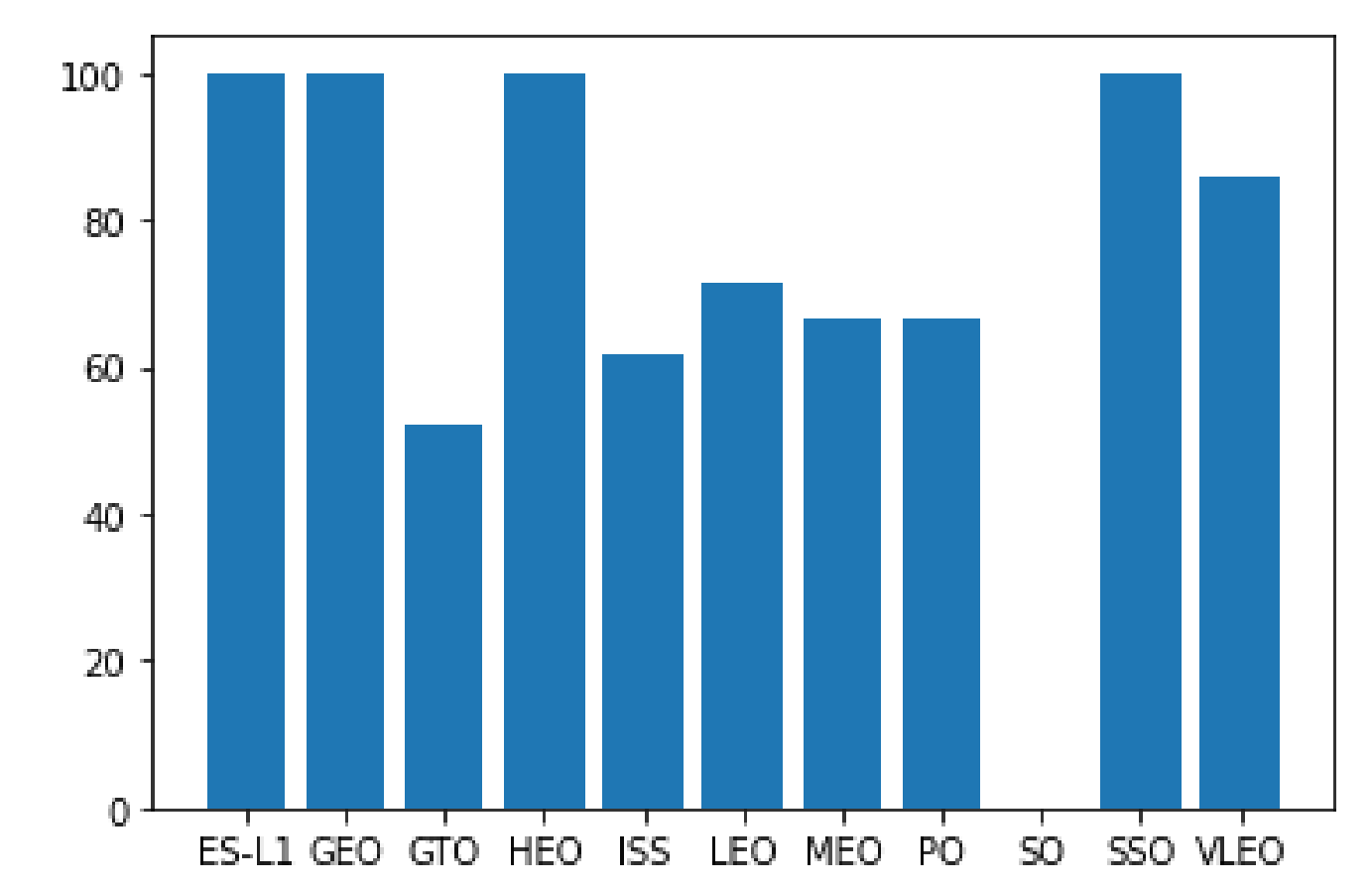
The plot suggests that all Launch Sites have higher success rate with increasing Flight Number

Payload vs. Launch Site



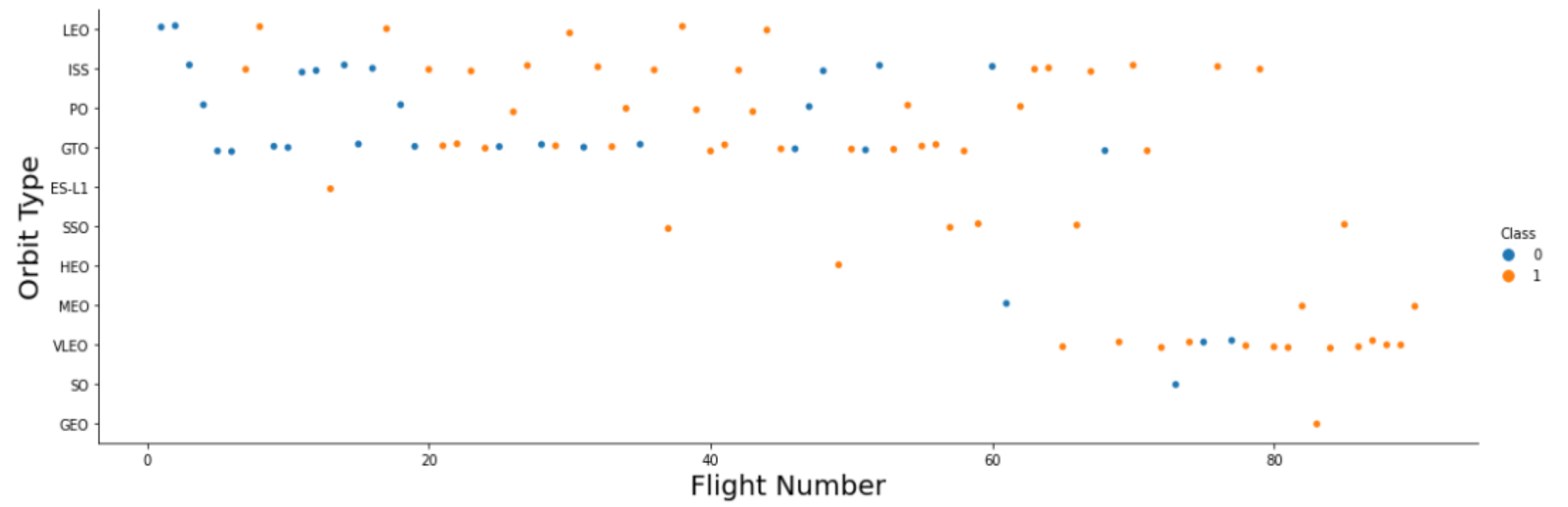
- It appears in **VAFB-SLC 4E** launch site there are no rockets launched for heavy payload mass(greater than 10000) and there seem to be better success rate for higher payload lass in this site
- The payload mass does not appear to have a clear impact on the outcome in the other sites

Success Rate vs. Orbit Type



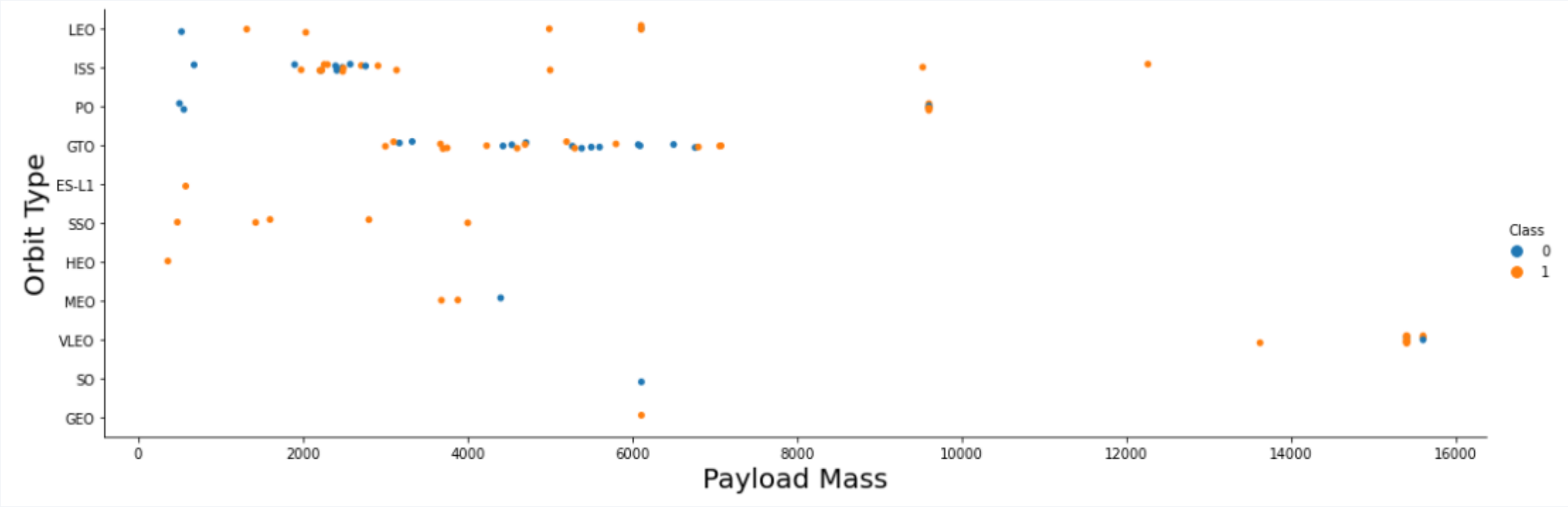
- Orbits with the highest success rate : ES-L1, GEO, HEO, SSO

Flight Number vs. Orbit Type



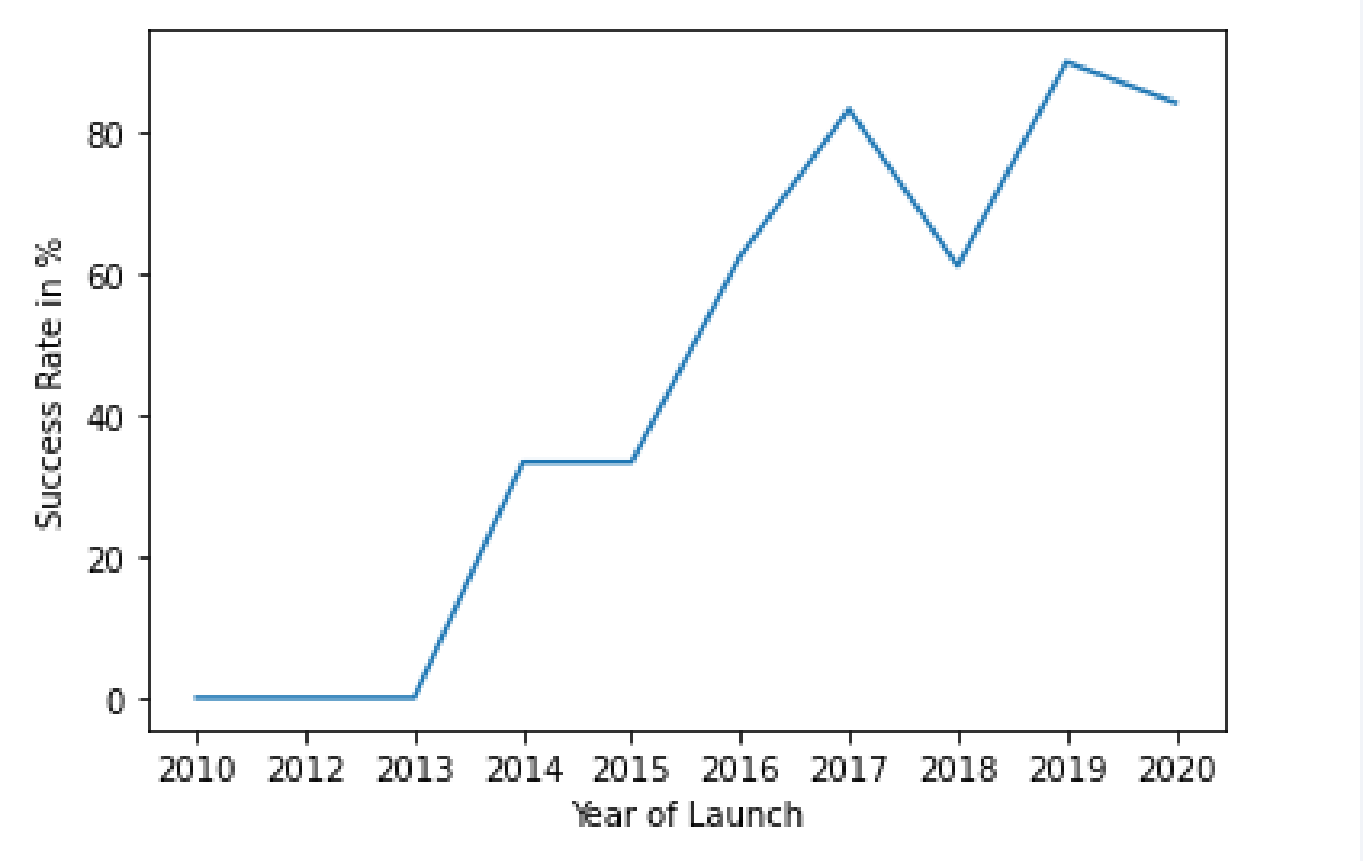
- On one hand, for the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both successful and unsuccessful landing rate are there.

Launch Success Yearly Trend



- The sucess rate since 2013 kept increasing till 2020s

All Launch Site Names

- To find the names of the unique launch sites with SQL:
  - SELECT DISTINCT(Launch_Site) FROM SPACEX
- The query result is shown on the right, and it can be seen that we have four (4) launch sites in total.

Query Result

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

Launch Site Names Begin with 'CCA'

- To find 5 records where launch sites begin with `CCA`:
  - SELECT * FROM SPACEX WHERE Launch_Site LIKE 'CCA%' LIMIT 5
- The query result is shown below, the launch site appears to be CCAFS LC-40

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

- Calculate the total payload carried by boosters from NASA
  - SELECT SUM(payload_mass__kg_) NASA_Total FROM SPACEX WHERE customer='NASA (CRS)'
- The result from the query is shown below, a value of 45 596 Kg.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - SELECT AVG(payload_mass__kg_) Avg_F9_MASS FROM SPACEX WHERE booster_version = 'F9 v1.1'
- The query result is displayed below, the average mass carried by the booster F9 v1.1 is 2 928 Kg

avg_f9_mass
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - SELECT MIN(Date) Success_Date FROM SPACEX WHERE landing__outcome = 'Success (ground pad)'
- The first successful landing outcome occurred in the past, so the MIN() function on the date will return the earliest date where success occurred

success_date
2015-12-22

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
  - SELECT booster_version boosters_names FROM SPACEX WHERE landing__outcome = 'Success (drone ship)' AND (payload_mass__kg_ BETWEEN 4000 AND 6000)

- The result show 4 boosters that fit the given conditions

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
  - SELECT * FROM (SELECT COUNT(*) failure_count FROM SPACEX WHERE mission_outcome LIKE '%Failure%'), (SELECT COUNT(*) success_count FROM SPACEX WHERE mission_outcome LIKE '%Success%')
- We used a main SELECT statement and two subqueries. The main SELECT statement retrieves everything from the two subqueries, where each subquery retrieves a single value and is treated like a table. So this is similar to selecting from two tables.

failure_count	success_count
1	100

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
  - SELECT DISTINCT(booster_version) FROM SPACEX WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEX)
- The WHERE clause was used to restrict results for where the payload is equal to the maximum payload. We get a total of 12 distinct boosters

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - SELECT landing__outcome, booster_version, launch_site FROM SPACEX WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE)=2015
- In 2015, we have only two failures in drone ship

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
  - SELECT landing__outcome, COUNT(landing__outcome) count FROM SPACEX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing__outcome ORDER BY count DESC
- We can see from the result that there are more instances where no attempt was made to successfully land the ship

landing_outcome	COUNT
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

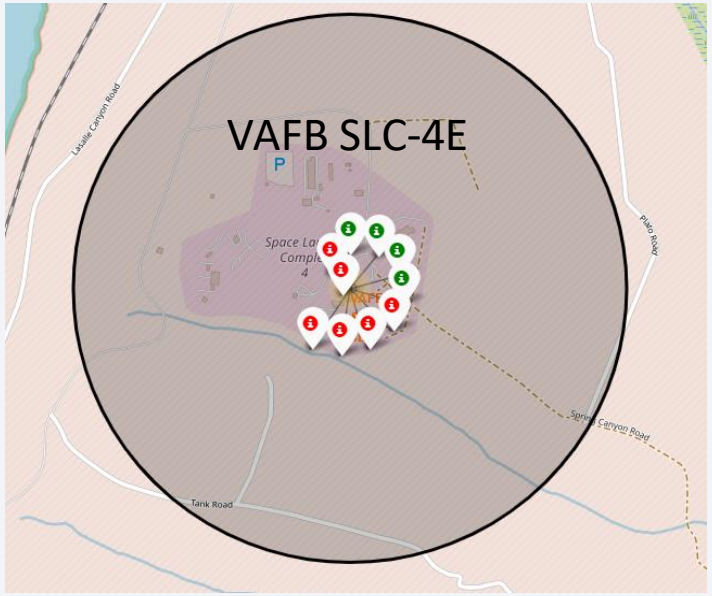

Section 3
Launch Sites
Proximities Analysis

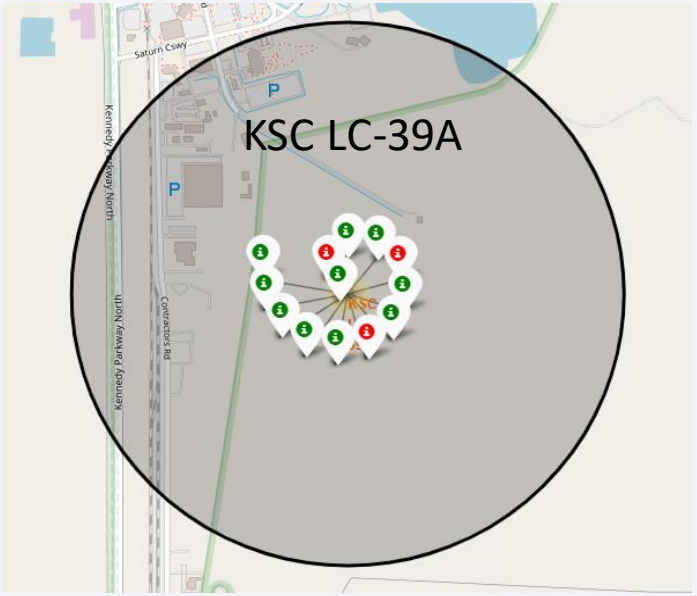
Folium map with marked launch sites

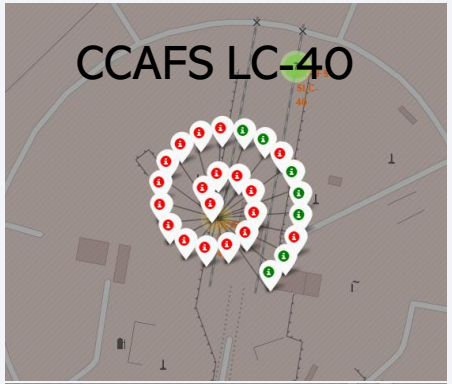
- Except for VAFB SLC 4E which is far to the west, all the remaining 3 sites are closer to one another and on the east
- All launch sites are in the United-States and are very cloe proximity to the coast

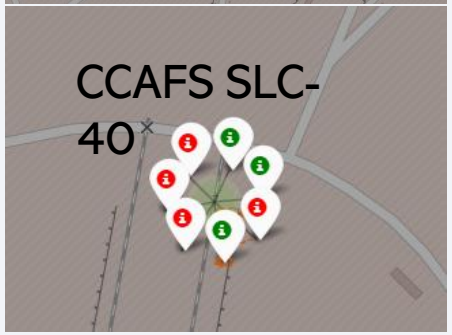
Launch outcomes for each site on the map

- A launch only happens in one of the four launch sites, so many launch records had the same coordinate. Marker clusters were used to simplify a map containing many markers having the same coordinate.





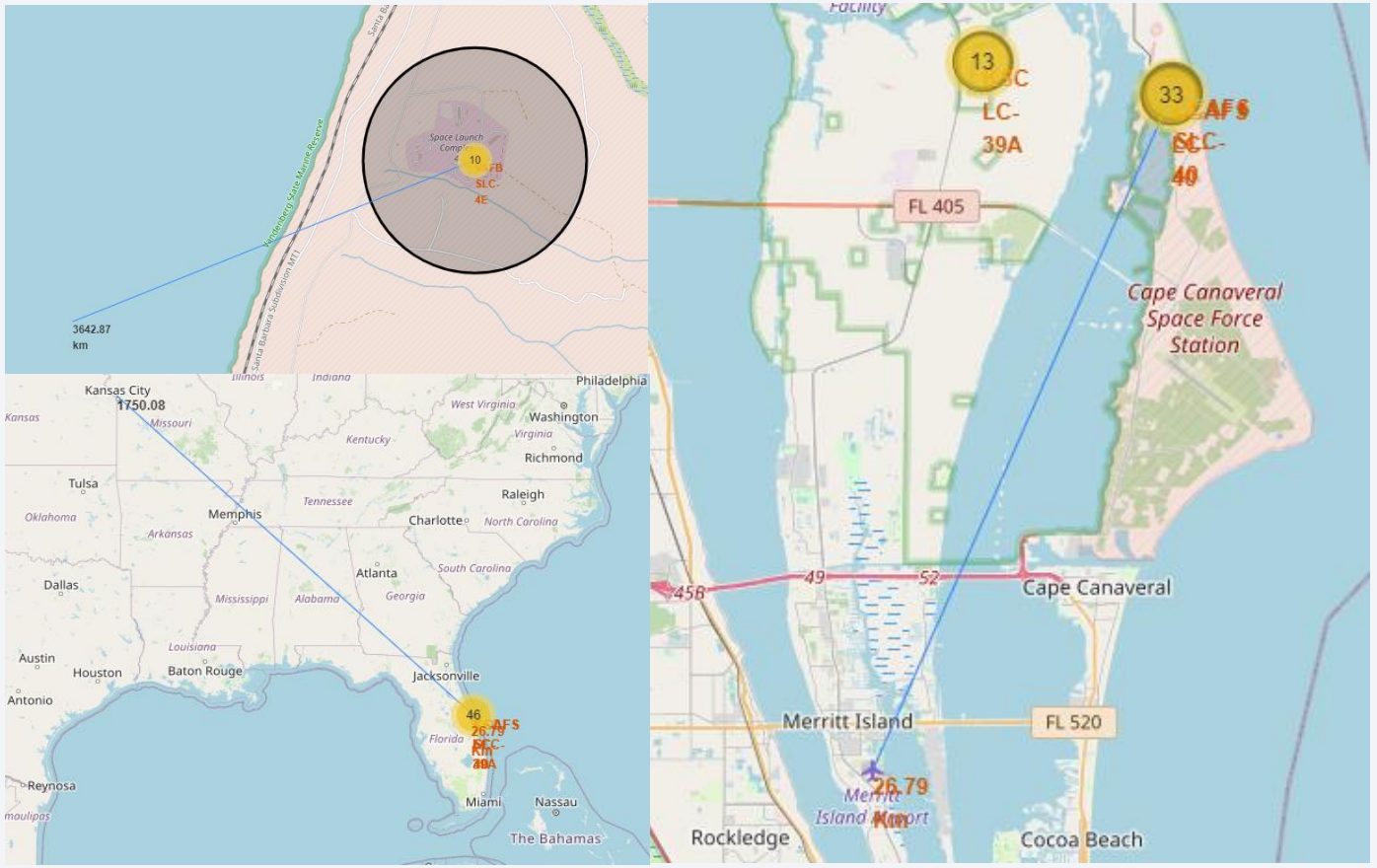


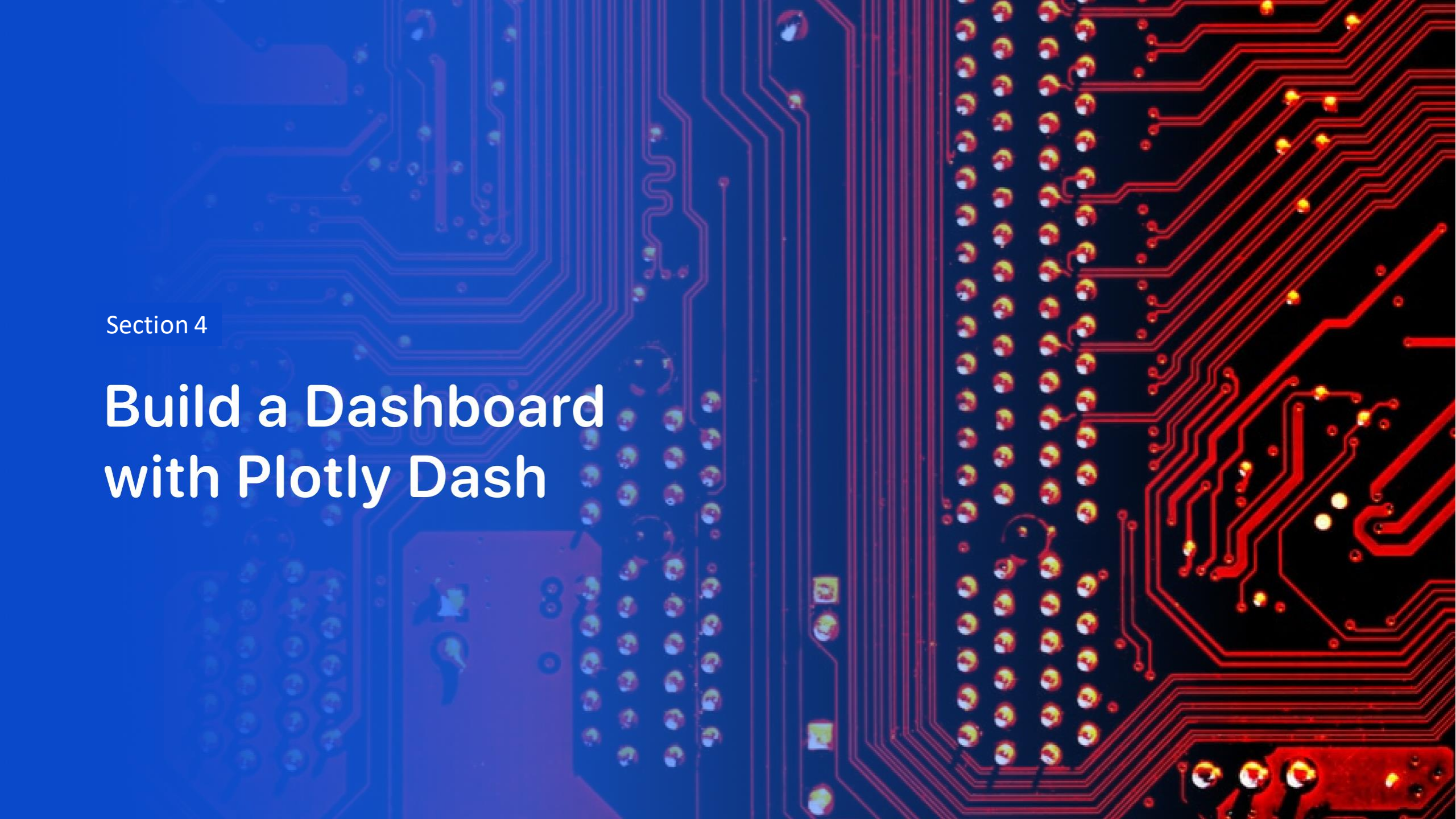


For each site, green marker shows a success and red marker a failure. So we can clearly see KSC LC-39A has the highest success rate

Distance between launch sites and proximities

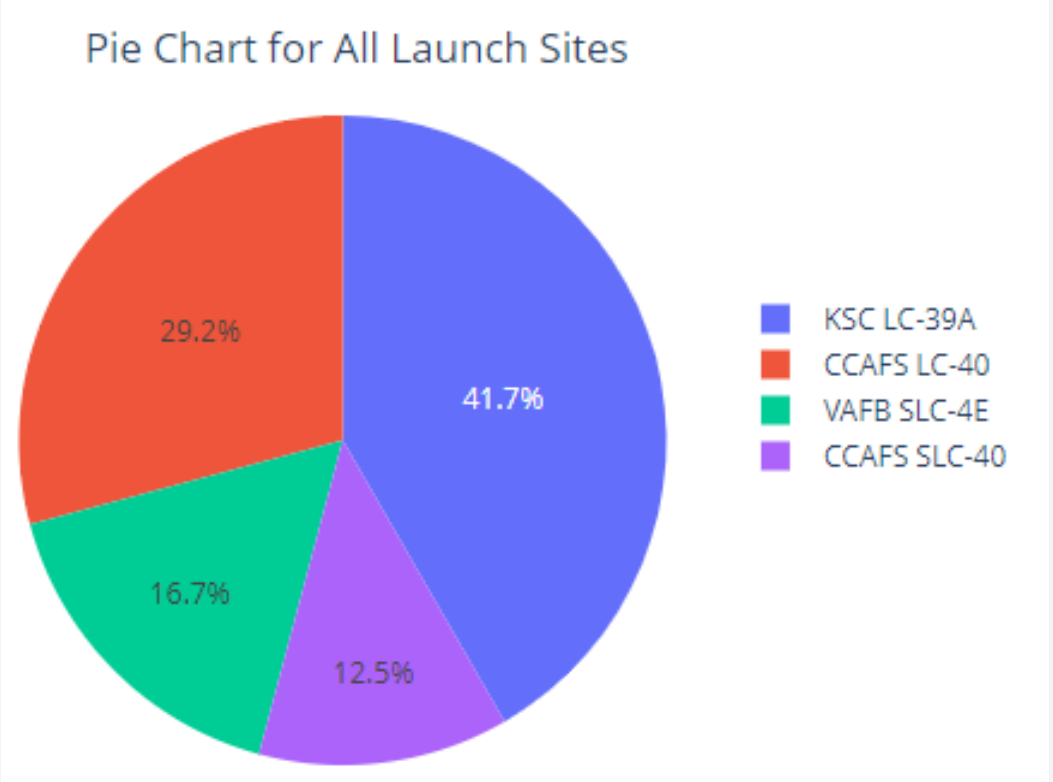
- All sites are in close proximities to coasts, and there is at least one airport in the vicinity.
- Launch sites are far away from major cities




Section 4
Build a Dashboard with Plotly Dash

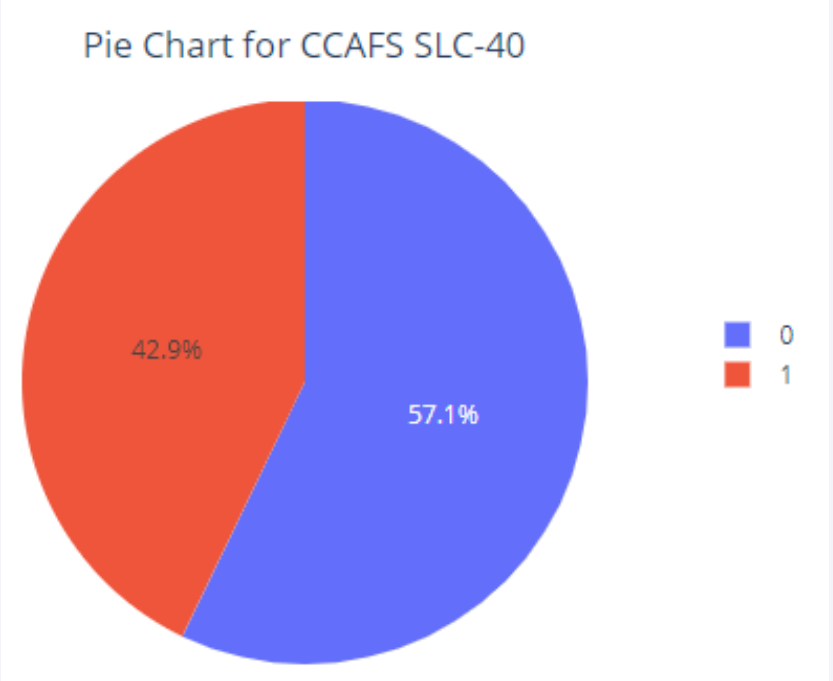
Dashboard, Pie chart for all sites

- The pie chart of launch success count for all sites displays the share of each site in the total success count.
- We see that with 41.7% of all successes, KSC LC-39A is the site with most success, and with 12.5%, CCAFS SLC-40 is the site with the least number of successes.



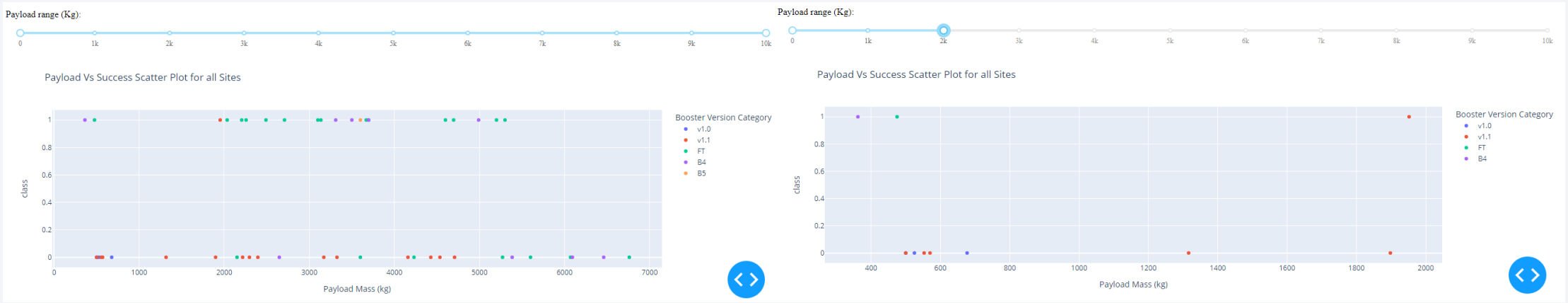
Dashboard, Pie chart for individual site

- The dropdown menu allows the selection of individual launch site and the displayed pie charts shows the proportion of success and failure for the total launch count on the site.
- Here we show the pie chart of the launch site with the highest launch success ratio, and it's CCAFS SLC-40 which has 42.9% success rate



Payload vs Launch Outcome for all sites

- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Payload range does not seem to have great impact on the outcome when all sites are combined
- Booster version FT have the largest success rate while booster version v1.1 has the lowest success rate.

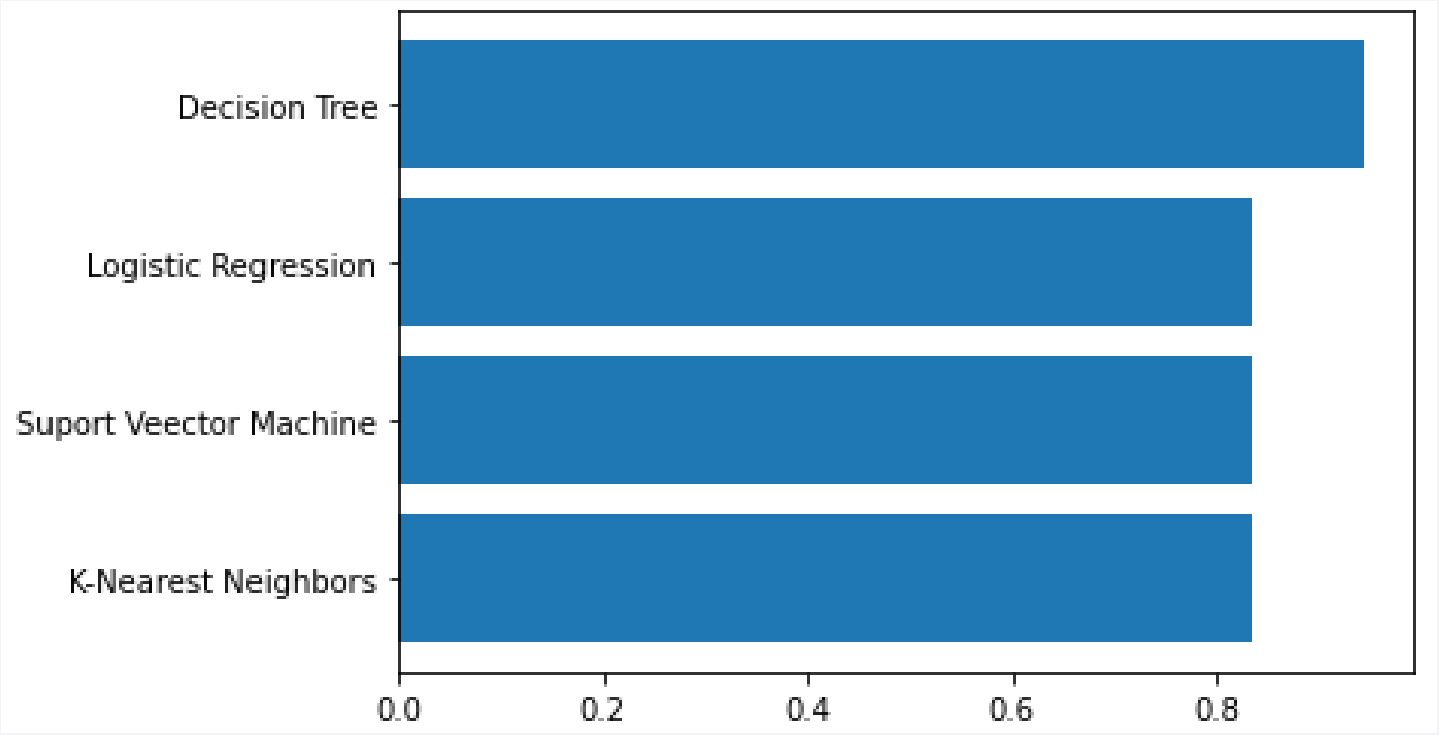


Section 5

Predictive Analysis (Classification)

Classification Accuracy

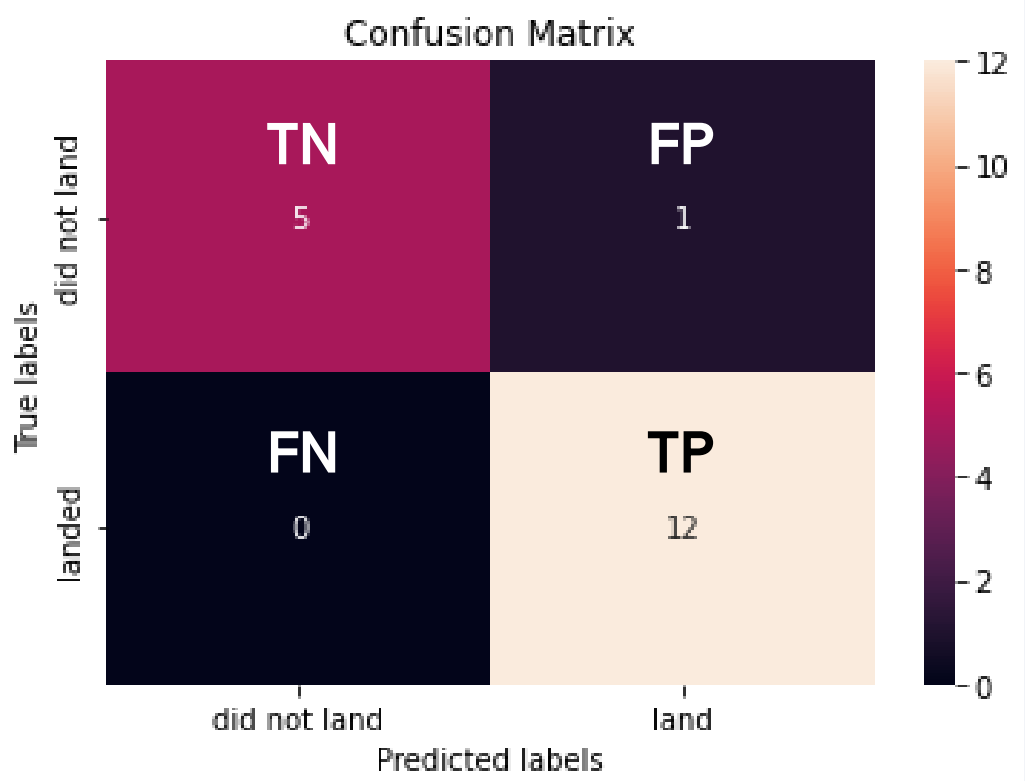
- Visualization of the built model accuracy for all built classification models, in a bar chart



Decision Tree classification has the highest accuracy of about 94% on the test data

Confusion Matrix

- Confusion matrix of the best performing model.



N P

T = True
F = False
N = Negative
P = Positive

We see that this model correctly predicts the outcome for all the test data except a single failure that was wrongly predicted as a success (FP)

Conclusions

- We can accurately predict landing outcome for future space mission using historical data;
- Important factors influencing landing outcome included the year. We can infer that future missions are likely to result in success, but this is probably due to improvement in technology over the years;
- Other factors with clear influence on the landing outcome were the Launch Site, the Booster Version, and the Orbit Type;
- Decision Tree Classifier was the best model to predict the outcome of a mission, and we successfully predicted landing outcomes with over 90% accuracy using this model.

Appendix

- Dash tutorial, found here, were very helpful to understand Dash and build the Dashboard

Thank you!