IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eveliina Hampus
26.5.2024

Outline

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

Executive Summary

- Summary of methodologies

Data collection methodology included using Python's libraries for API requests, Numpy, Pandas and Datetime. Performing GET request and have the data return as JSON. From there to filter and collect information that was later used for our predictions. These would include rocket booster name, payload mass and orbit, launchpad and its location, outcome, type of landing, number of flights and few more. Data wrangling would include filtering data, replace NaN values with mean method.

Some data was collected via webscraping the Wikipedia page of Falcon 9 and Falcon Heavy launches. In this information was collected with Python as well, performing GET requests, creating beautiful soup object, parse HTM, extract names from table headers, create empty dictionaries from keys of extracted column names, parse them, fill a dictionary, and convert into a data frame.

External CSV file was used, transferred to database and performed data collection with Python as well as SQL queries with the help of SQL Alchemy.

The data was then visualized with Python's visualization libraries such as Matplotlib and Seaborn, and with interactive methods such as Plotly and Dash.

In the last part, machine learning techniques were used to create models that best predict these changing variables. Method of Decision Trees, K-Nearest Neighbors and SVM were used.

Executive Summary

- Summary results

Exploratory Data Analysis (EDA) with data visualization introduces parameters that are placed as variables towards each other. In the first visualization the payload mass was plotted against flight number. The scatterplot indicated that good portion of first 20 attempts of launches were failed, with all payloads.

After that the successful launch was more likely to result. Following scatterplot was with launch site against flight number. The first launches were mostly from certain CCAFS SLC-40 launch site, henceforth there would be difficult to estimate this variable, as against flight number which seems to indicate stronger relationship for failures.

Payload mass against launch site indicated that payload mass less than 10 000 were more likely to be launched, and payload mass greater than that was completely omitted from testing of launch site VAFB-SLC 4E. The orbit type against payload mass would indicate that orbit types such a s Polar, Leo and ISS were more likely to be successfully landed with heavy payloads. Success rate over time would increase steadily after first three years of failures.

Executive Summary

Exploratory Data Analysis (EDA) with SQL give specific answer of queries. In the launch site, only four were used. These were CCAFS LC-40, VAFB LSC-4E, KSC LC-39A and CCAFS SLC-40. Then the total payload mass was calculated, the boosters carried mass total of 45 596kg. Average booster mass carried by version F9 v.1.1 was 2928,4kg and first successful landing outcome of ground was in December 22nd in 2015.

The boosters that were successful in carrying mass between were 4000 to 6000 kilograms were F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2. The total number of successful mission outcomes 100 against 1 failed one. Note that successful mission does not indicate necessarily successful landing or launching, but more like how outcome was to of planned one.

Lastly it was examined which boosters could carry maximum payload, and then study carefully records from year 2015.

Executive Summary

Exploratory Data Analysis (EDA) with Interactive Dashboard summarize basic statistics. Total successful launches were dominated by launch site KSC LC-39A with 41,7% share, followed by CCAFS LC-40 with 29,2% share, VAFB SLC-4E with 16,7% share, and last performing launch site was CCAFS SLC-40 with 12,5% share.

In CCAFS LC-40 had 73,1% failed and 26,9% successful launches. CCAFS SLC-40 has 57,1% failed and 42,9% successful launches. The best performing launch site KSC LC-39A has 76,9% successful and 23,1% failed launches. VAFB SLLC-4E performed 60% failed and 40% failed launches.

Lastly the machine learning models were built, evaluated, and improved. The decision tree model has best accuracy of 94%, whereas SVM and KNN both performed with 83% accuracy. The best performing model for predicting landing outcome was decision tree.

Introduction

- Project background was to explore SpaceX data of rocket launches. From the data I was to clean, process, and predict best parameters which will affect success rate of the launches.

- Question here was which parameters would clearly affect on success of the launch. In many processes the parameters examined were related the Flight Number, Launch Site, Payload Mass or Orbit type targeted.

- Lastly it was to predict with different machine learning models how the landing would end up, and find best performing model.

Section 1

Methodology

Methodology

Executive Summary

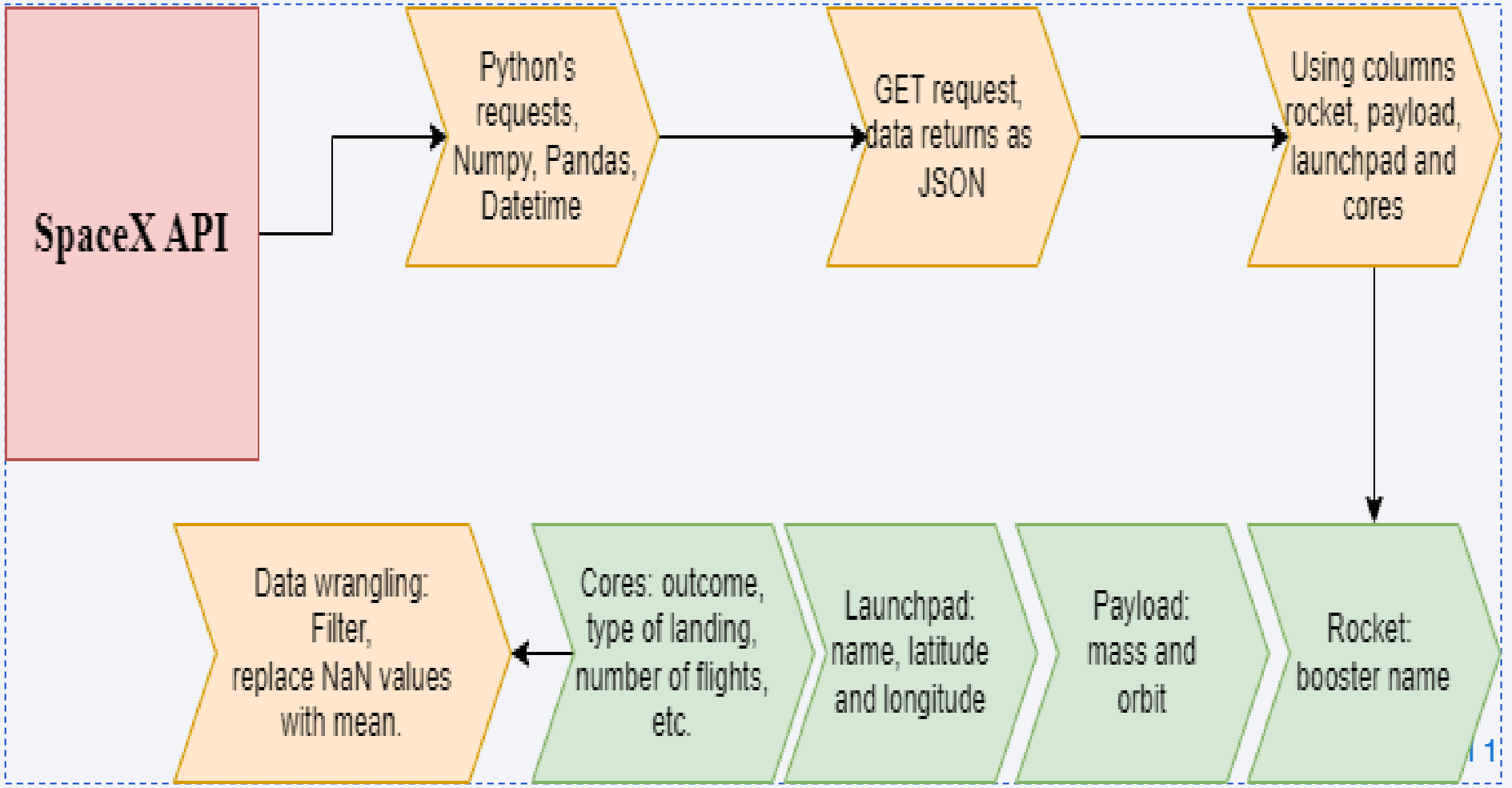
- Data collection methodology:
  - Webscraping and API requests
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Datasets were collected by two different methods: using webscraping by Python's beautiful soup library and requesting information from spacexdata API.
- The site used as the object of webscraping was Wikipedia page of Lists Falcon 9 and Falcon Heavy Launches.
- The HTTP requests were targeted to SpaceX API .

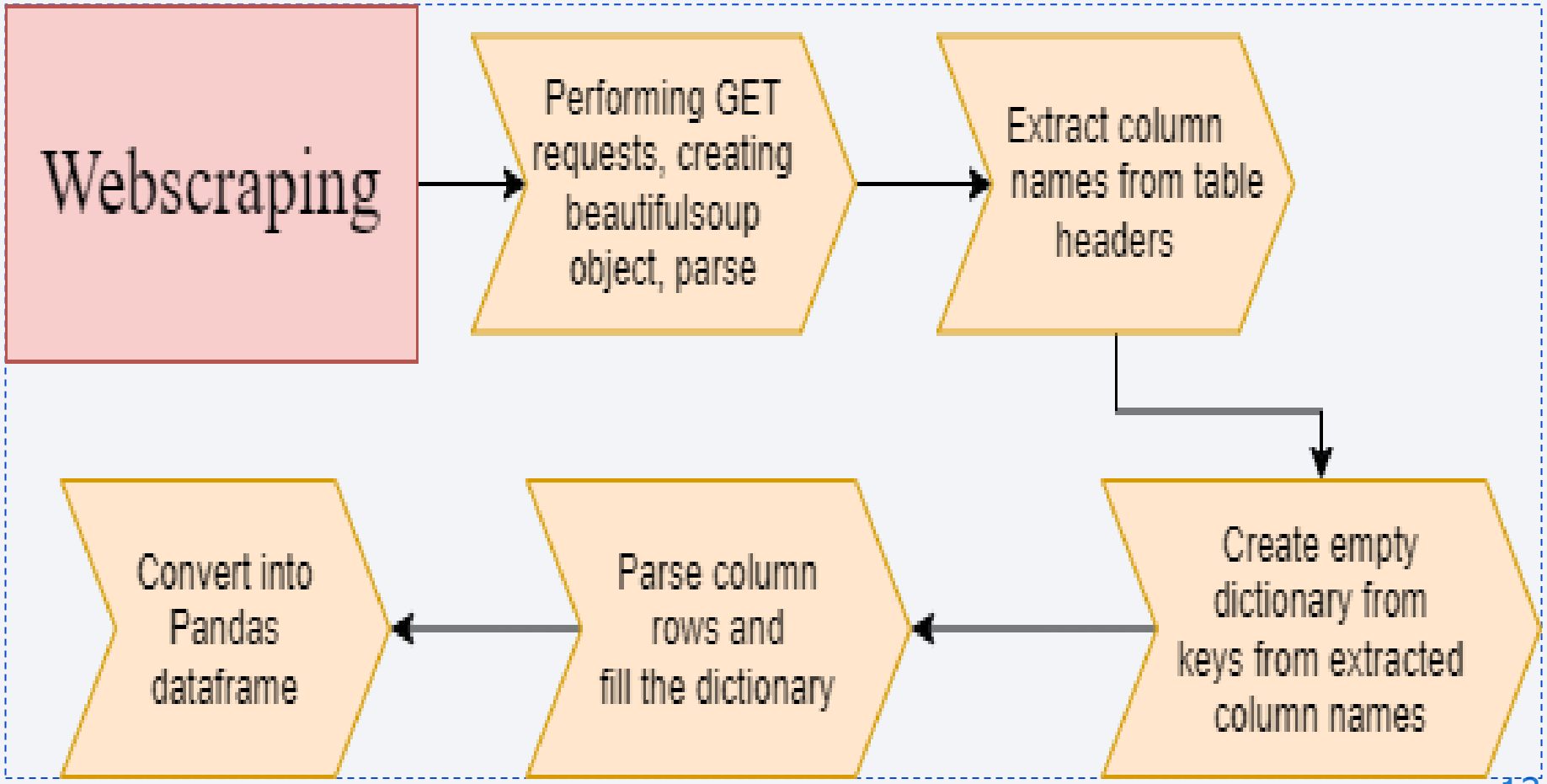
Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook.



Data Collection - Scraping

- GitHub URL of the completed web scraping notebook.



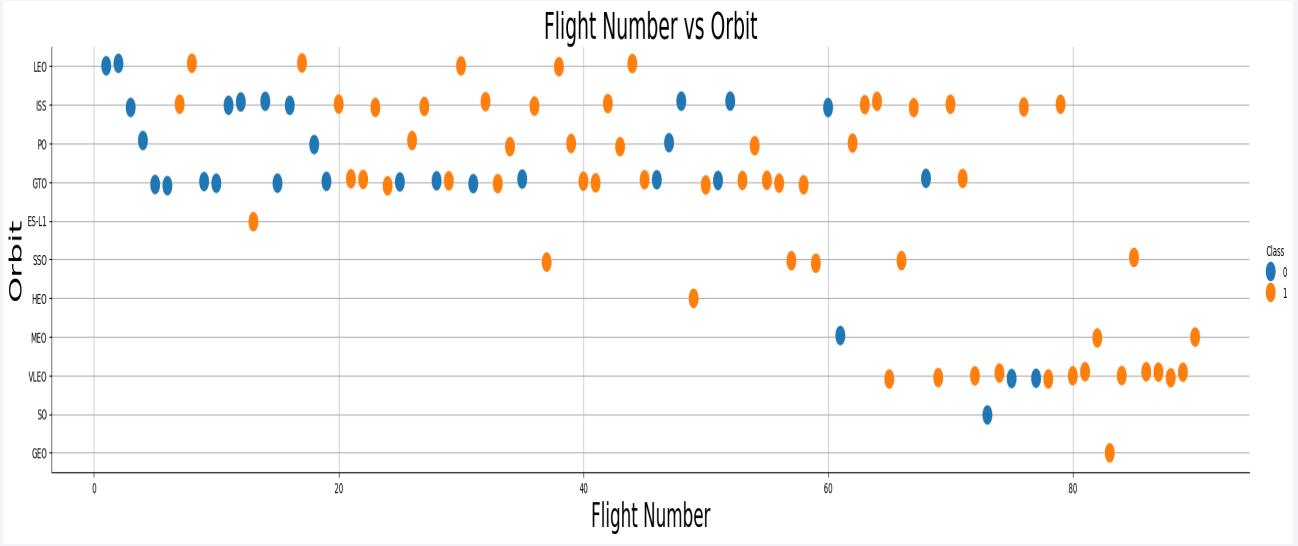
Data Wrangling

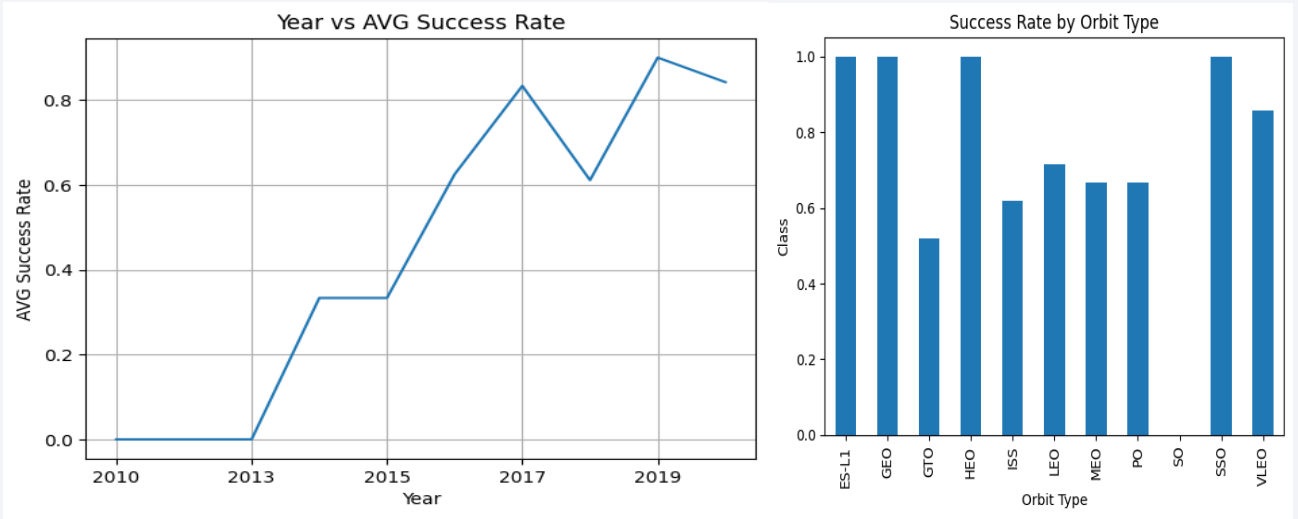
- GitHub URL of completed data wrangling related notebooks.

- Calculating number and occurrence of each site, each orbit, mission outcome
- Create landing outcome label from outcome column

EDA with Data Visualization

- GitHub URL of completed EDA with data visualization notebook.
- For visualization purposes, scatter plots, line charts and bar charts were used due both being best present different variables over each other.





EDA with SQL

SQL QUERIES

- SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
- SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- SELECT SUM(Payload_mass__kg_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
- SELECT AVG(Payload_mass__kg_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
- SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
- SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND Payload_mass__kg_ > 4000 AND Payload_mass__kg_ < 6000;

EDA with SQL

- SELECT Mission_Outcome, COUNT(*) AS Count FROM SPACEXTABLE GROUP BY Mission_Outcome;
- SELECT Booster_Version FROM SPACEXTABLE WHERE Payload_mass__kg_ = (SELECT MAX(Payload_mass__kg_) FROM SPACEXTABLE);
- SELECT SUBSTR(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE SUBSTR(Date, 0, 5) = '2015' AND Landing_Outcome = 'Failure (drone ship)';
- SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;
- GitHub URL of completed EDA with SQL notebook.

Build an Interactive Map with Folium

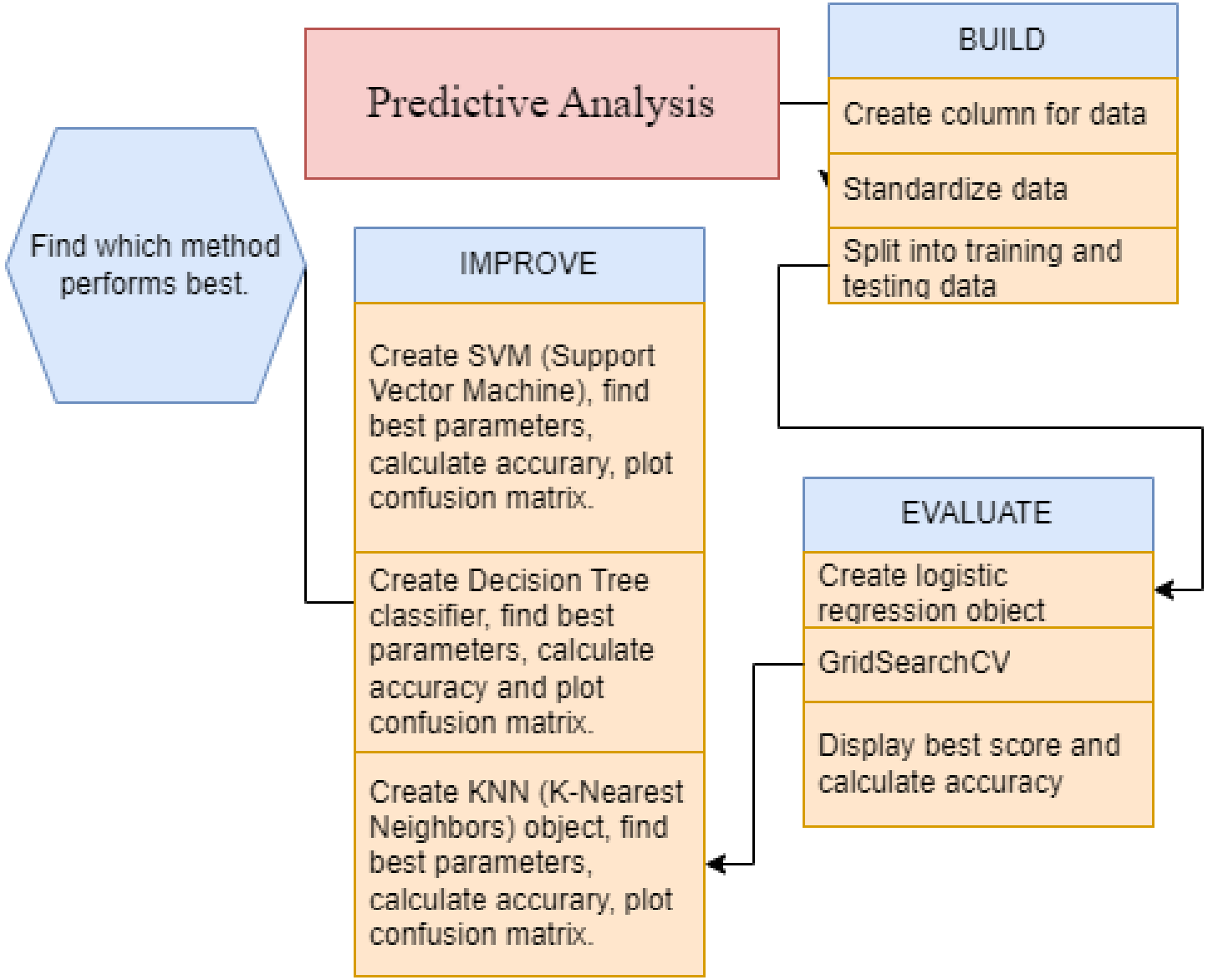
- Added circles and markers to indicate location of each launch site.
- Marked successful launches with green marker and failed with red marker.
- Used polyline to indicate proximity to sea and another to proximity of nearest airport (Lompoc Airport) from launch site VAFB SLC-4E.
- GitHub URL of completed interactive map with Folium map.

Build a Dashboard with Plotly Dash

- Pie charts and scatterplots were used, as well as payload slider for interaction of payload variables.
- Pie charts were to indicate success/failure ratio of each launch site, as well as total successes.
- Scatterplot were to indicate Payload vs Launch Outcome relationships.
- GitHub URL of completed Plotly Dash lab.

Predictive Analysis (Classification)

[GitHub URL of completed predictive analysis lab.]

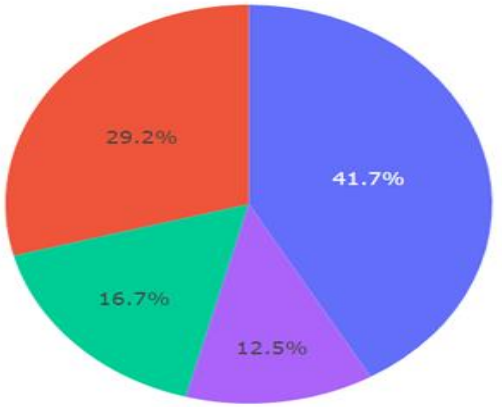


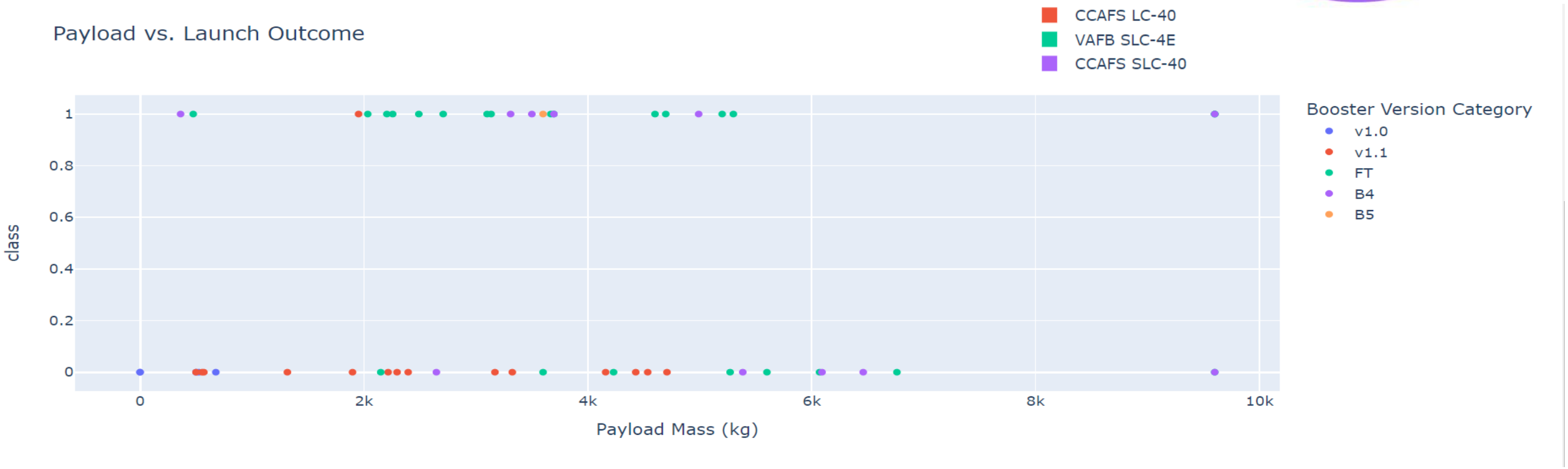
Results

- Exploratory data analysis results
- As Flight numbers increase, the likeliness of launch will increase.
- Payload mass has some significancy of likeliness of which orbit will be successfully targeted at.
- Some launch sites were used more, hence more likeliness to succeed in launches.

Results

Interactive analytics demo in screenshots

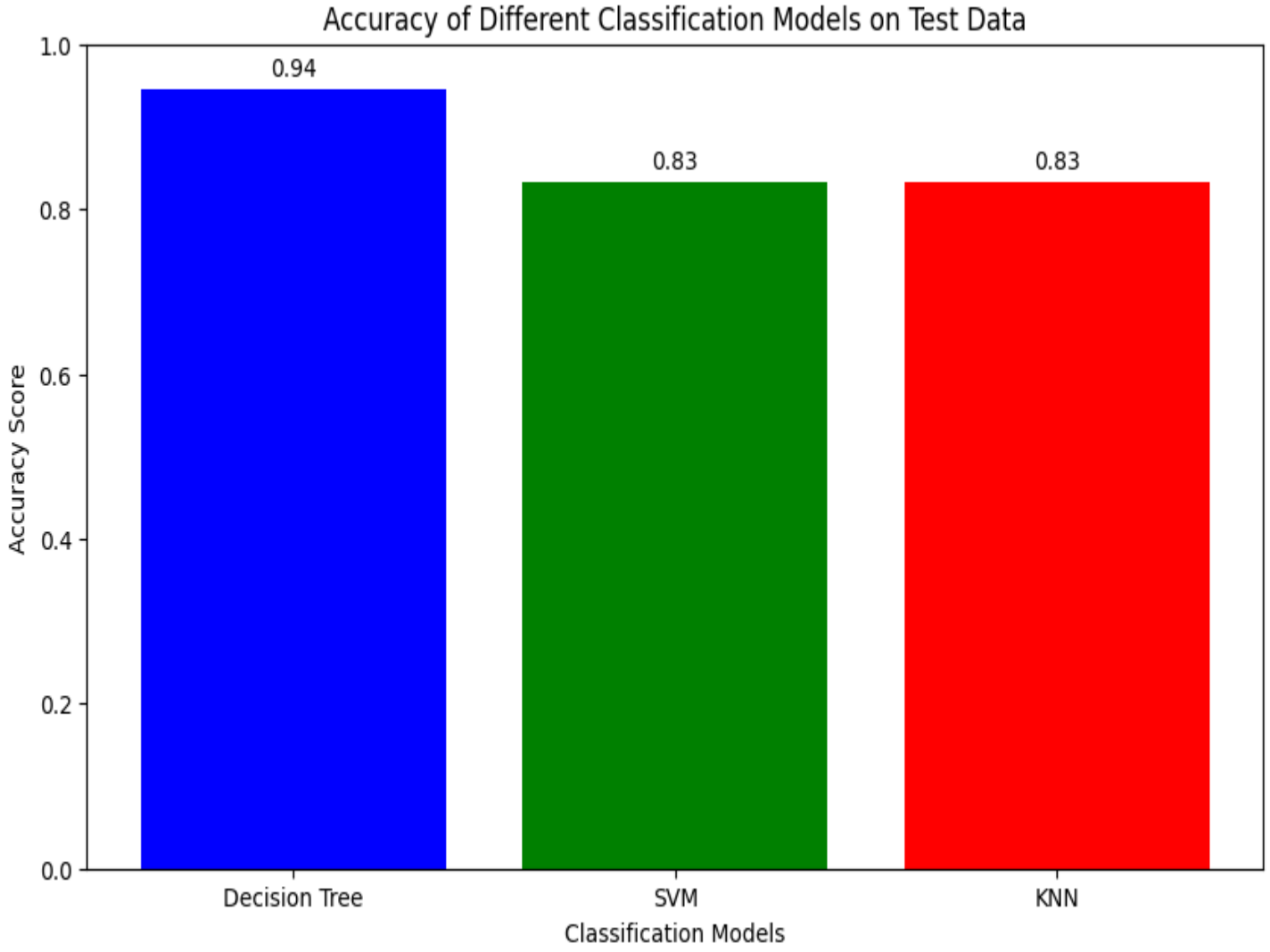


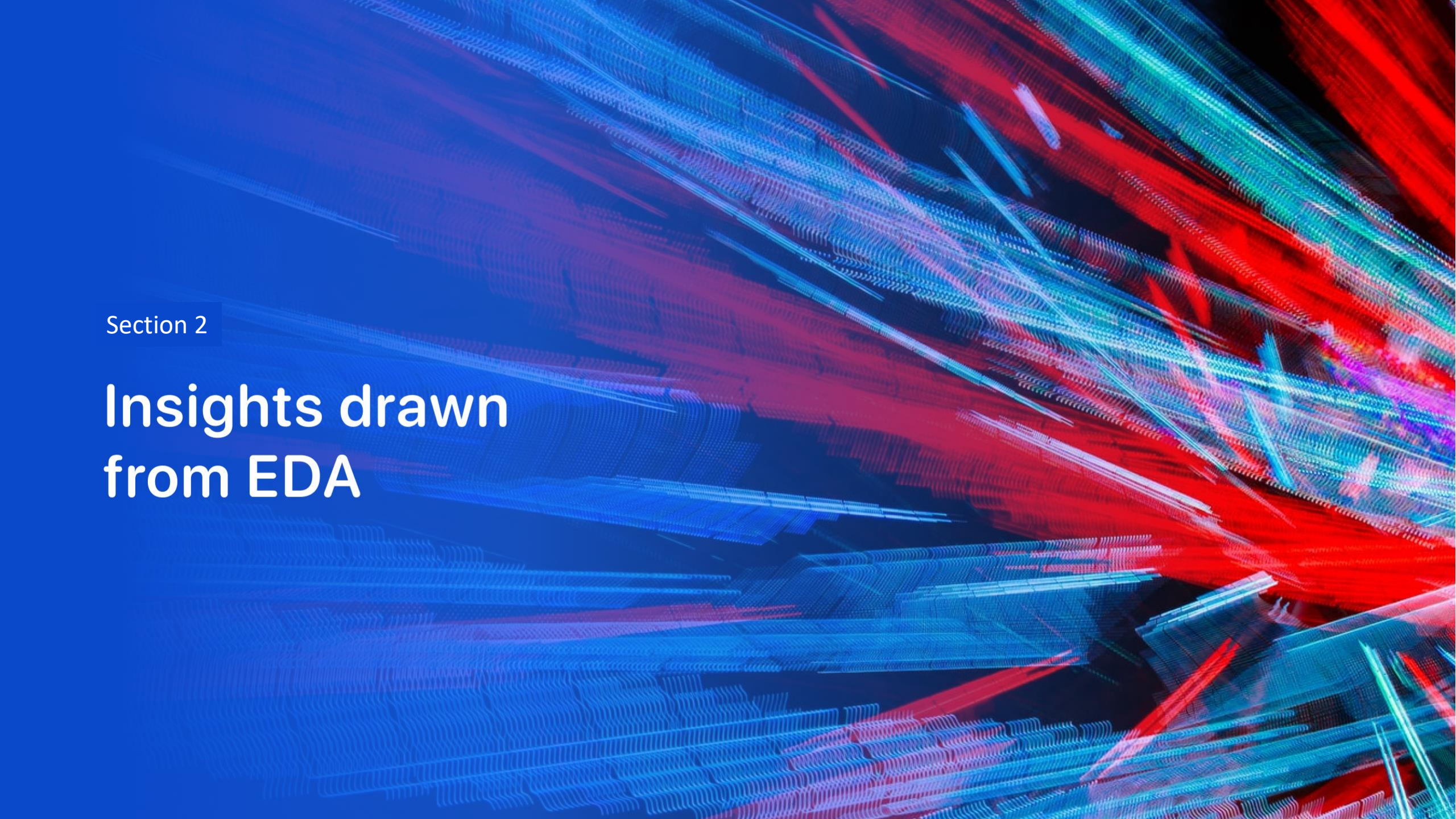


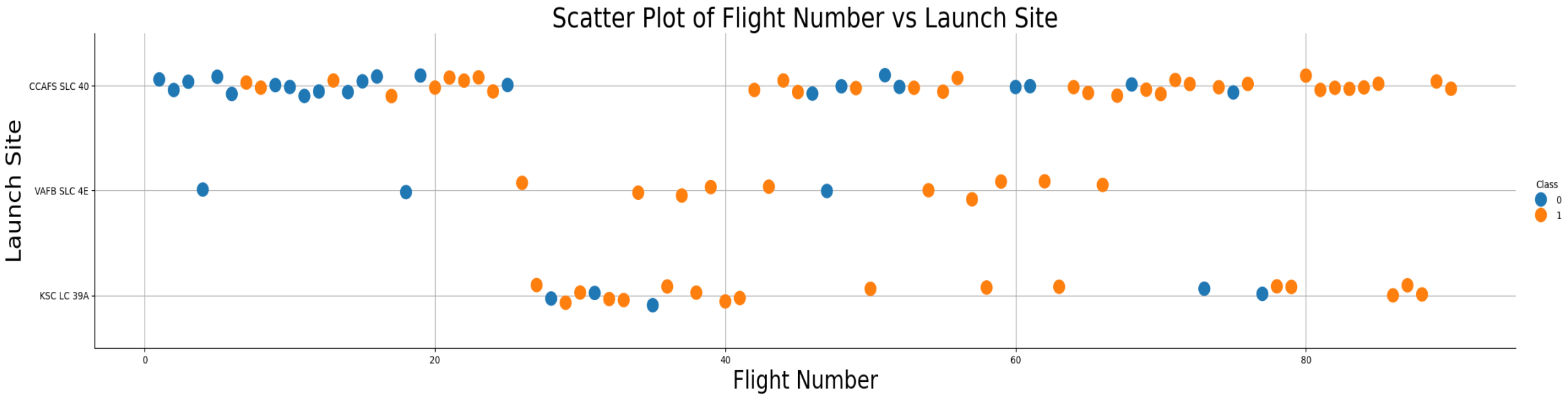
Results

Predictive analysis results

- Accuracy of Decision Tree on test data: 0.9444444444444444
- Accuracy of SVM on test data: 0.8333333333333334
- Accuracy of KNN on test data: 0.8333333333333334
- Decision Tree performs the best.

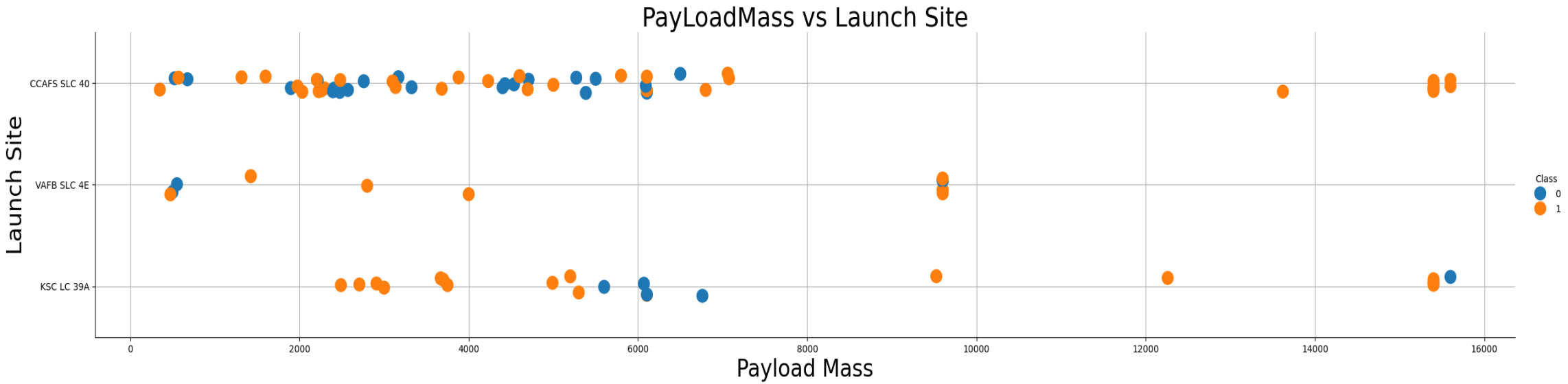



Section 2
Insights drawn from EDA



Flight Number vs. Launch Site

- Scatterplot reveals as Flight Number increases, the likeliness of success which is indicated with class value of 1 and orange color.
- Most failures were from CCAFS SLC-40 launch site.

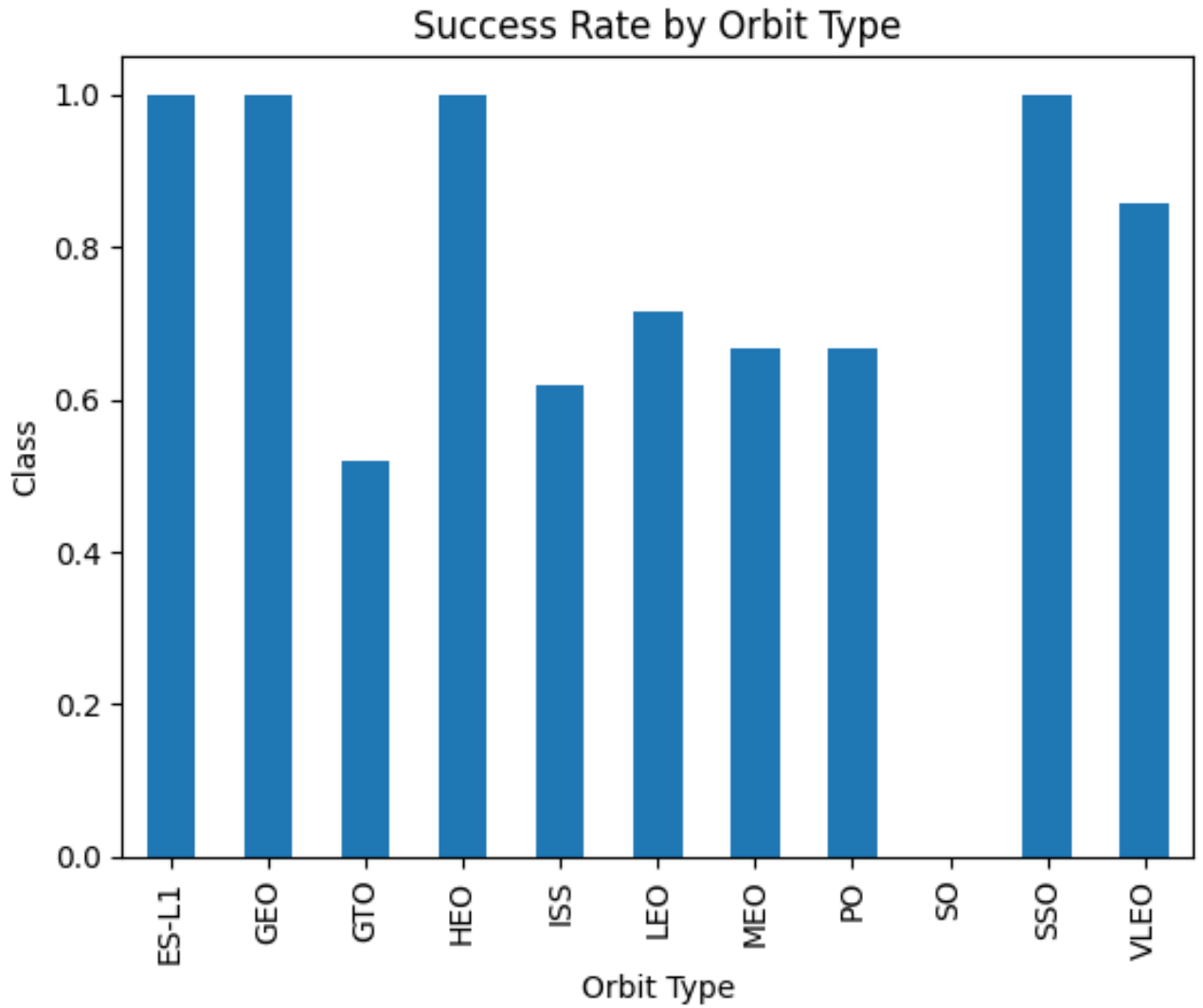


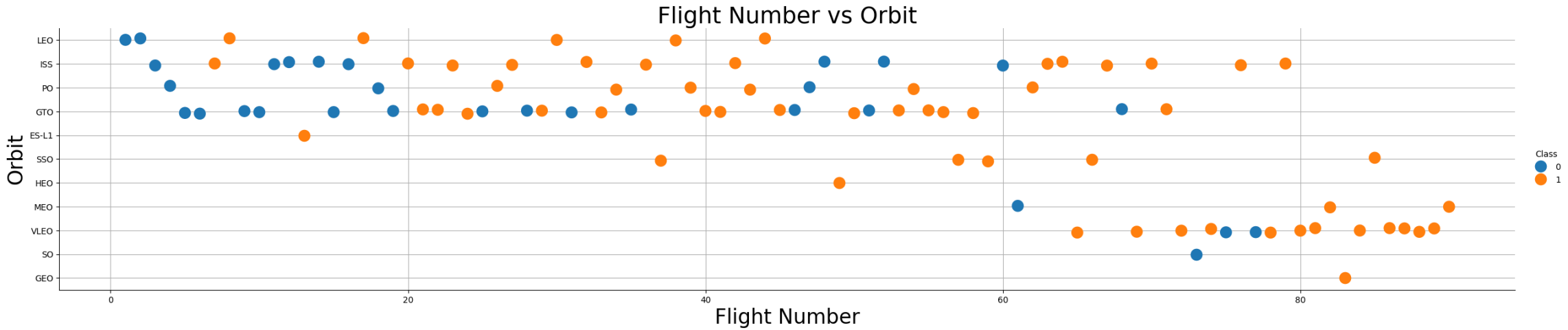
Payload vs. Launch Site

- CCAFS SLC 40 launch site doesn't seem suggest any relationship between launch site with payload mass.
- In launch site KSCLC 39A smaller payload masses were clearly more successful in launches.

Success Rate vs. Orbit Type

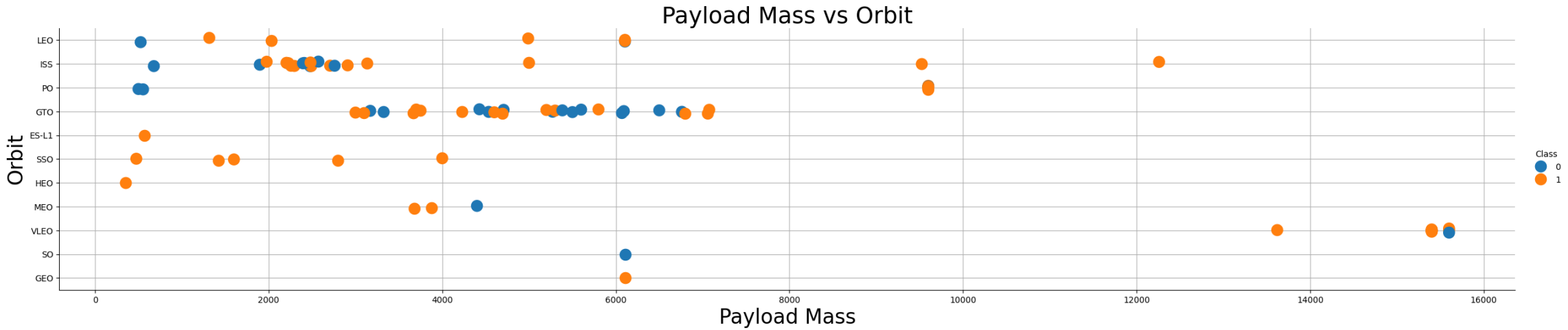
- Success rates were highest with orbits ES-L1, GEO, HEO and SSO.
- Success rate was lowest in SO, where all were failed.
- Orbits GTO, ISS, LEO, MEO and PO were moderate in success.
- VLEO had slightly higher than average success rate compared to previously mentioned moderate successful orbits.





Flight Number vs. Orbit Type

- With most orbits, the more launches were performed, the more likely it was to be successful.
- First 20 launches seemed mostly end up with failure.

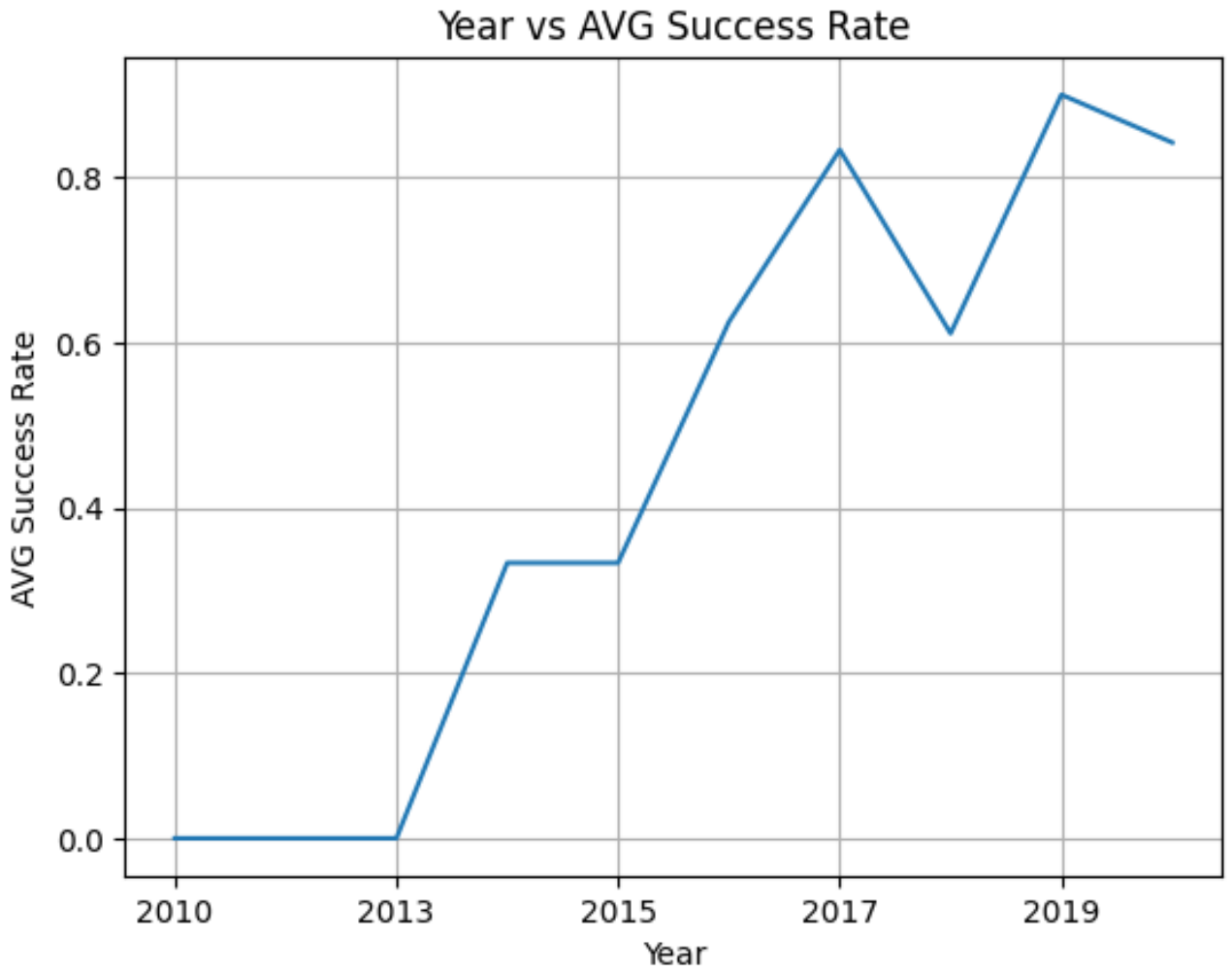


Payload vs. Orbit Type

- Only orbits ISS, PO and VLEO were targeted with payload mass greater than 8000.
- Number of attempts were greatest for orbits ISS and GTO specifically.
- Most attempts were with payload mass less than 8000.

Launch Success Yearly Trend

- Launch success was increasing steadily over the year, only 2018 was seemed to be more unsuccessful year.
- Records increased again in 2019 but faced a small slump in success right on next year.
- Between 2010 to 2013 all launches were unsuccessful.



All Launch Site Names

- There are four launch sites.
- Names of all the launches sites are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

- Query returns 5 records of launch site CCAFS LC-40.
- Records were dated for 2010-06-04, 2010-12-08, 2012-05-22, 2012-10-08 and 2013-03-01.
- *('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', '7:44:00', 'F9 v1.0 B0005', 'CCAFS LC-40', 'Dragon demo flight C2', 525, 'LEO (ISS)', 'NASA (COTS)', 'Success', 'No attempt')*
- *('2012-10-08', '0:35:00', '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

- Total payload mass carried by boosters launched by NASA (CRS): 45596 kg

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1: 2928.4 kg

First Successful Ground Landing Date

- Date of the first successful landing outcome in ground pad: 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- F9 FT B1022, F9 FT B1026, F9 FT B1021.2 F9 FT B1031.2 are the boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

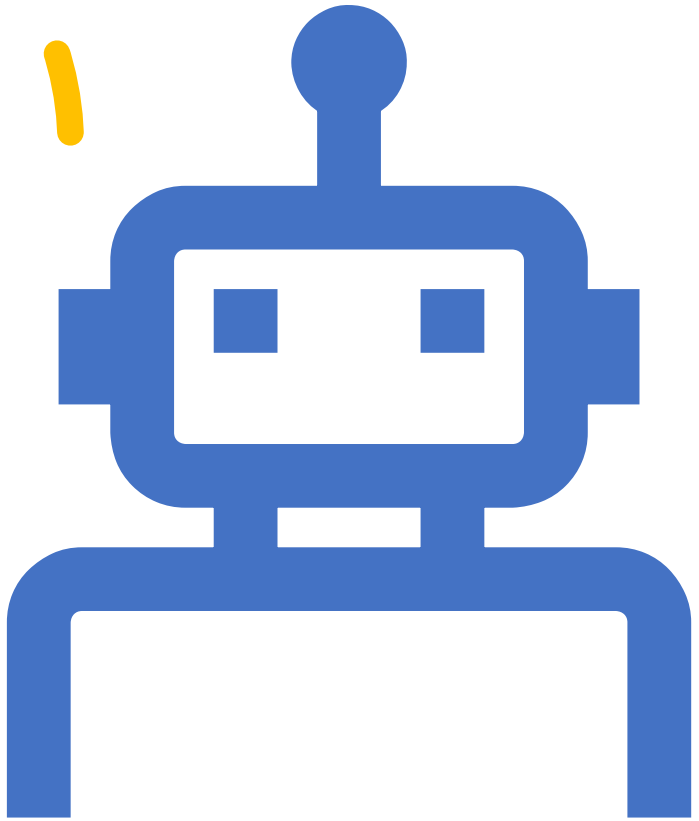
Total Number of Successful and Failure Mission Outcomes

- Failure (in flight) : 1
- Success : 98
- Success : 1
- Success (payload status unclear) : 1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass:

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 and 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

Month: 01 Landing Outcome: Failure (drone ship) Booster Version: F9 v1.1 B1012 Launch Site: CCAFS LC-40

Month: 04 Landing Outcome: Failure (drone ship) Booster Version: F9 v1.1 B1015 Launch Site: 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

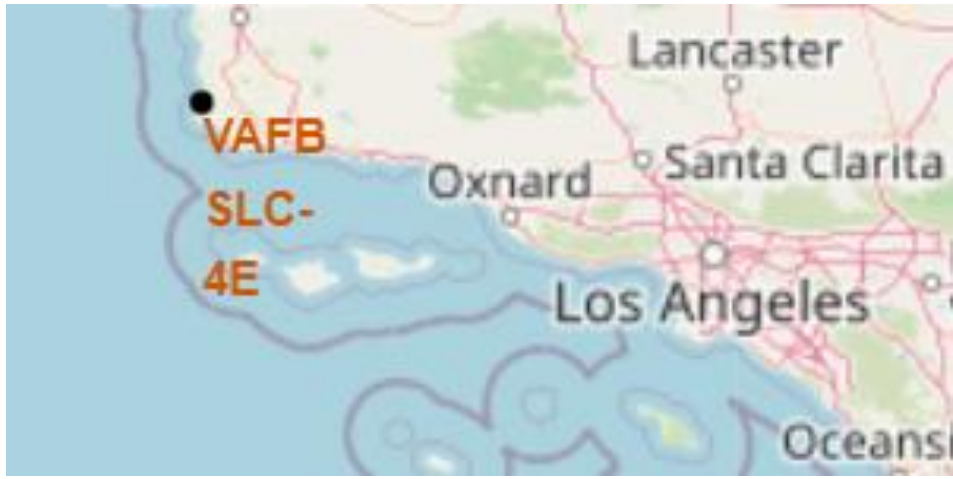
Landing Outcome: No attempt Count: 10 Landing Outcome: Success (drone ship) Count: 5 Landing Outcome: Failure (drone ship) Count: 5 Landing Outcome: Success (ground pad) Count: 3 Landing Outcome: Controlled (ocean) Count: 3 Landing Outcome: Uncontrolled (ocean) Count: 2 Landing Outcome: Failure (parachute)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

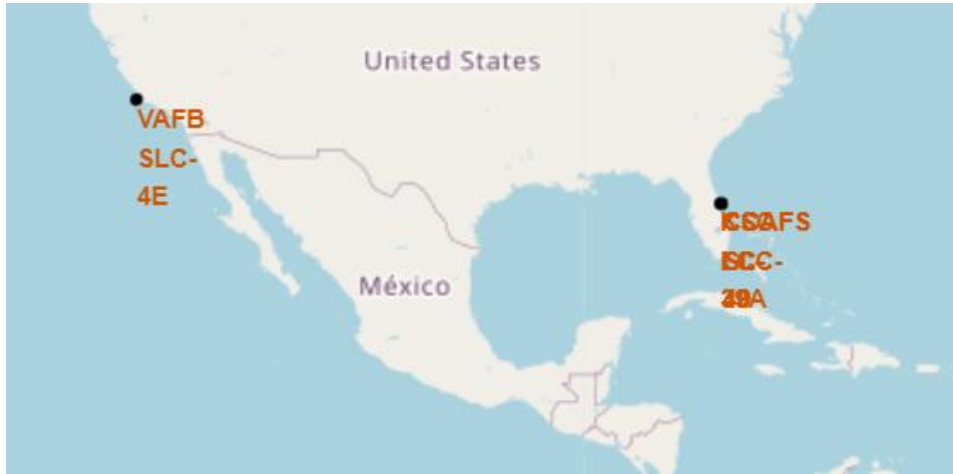
Section 3

Launch Sites Proximities Analysis

Location of Launch Sites

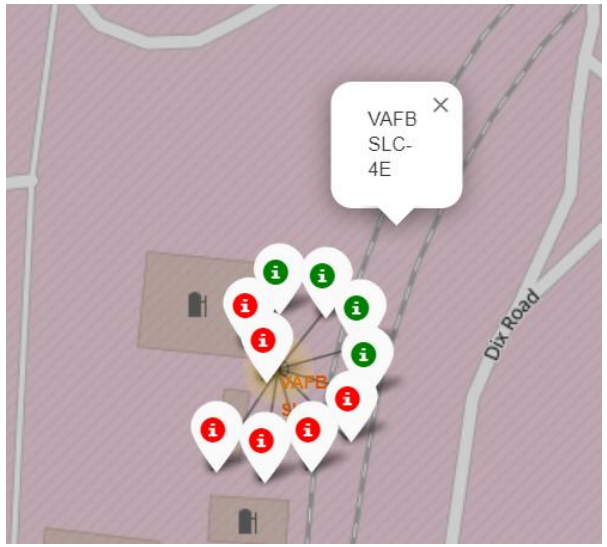
- West coast of USA has only one launch site.
- East coast has three launch sites, where sites CCAFS SLC-40 and CCAFS LC-40 almost overlap fully.

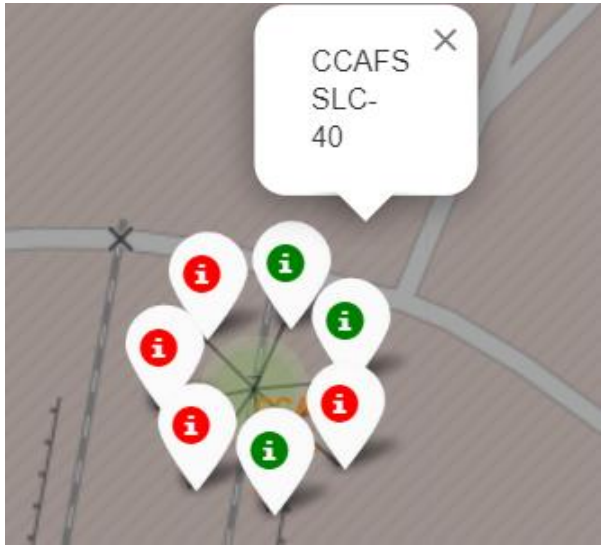


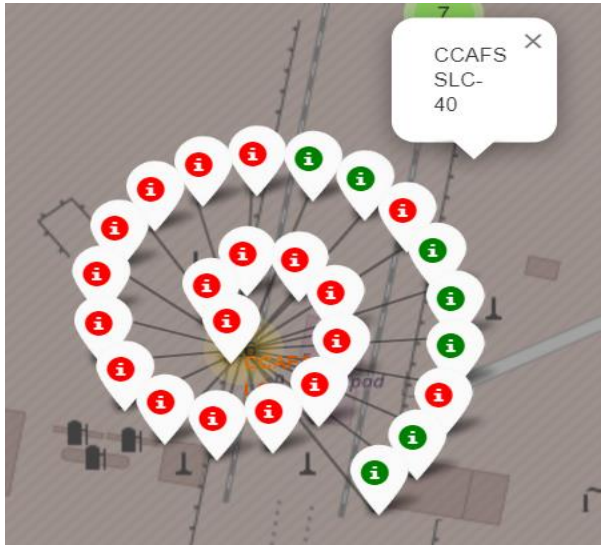


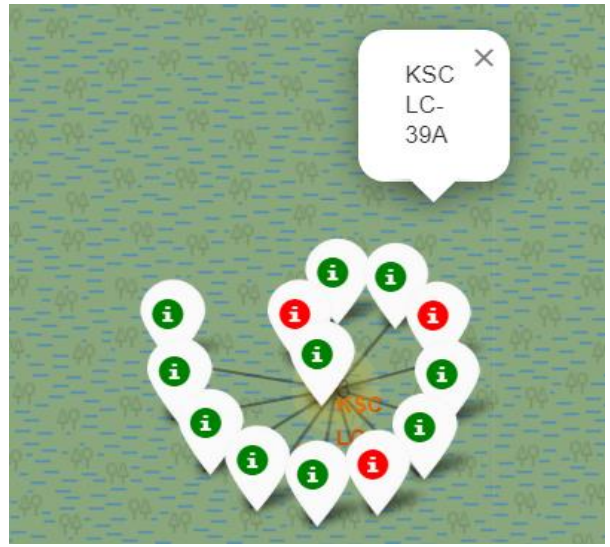
Success and Failure of Each Launch Site

- Red indicate a failed launch, and green marker indicate successful launch.



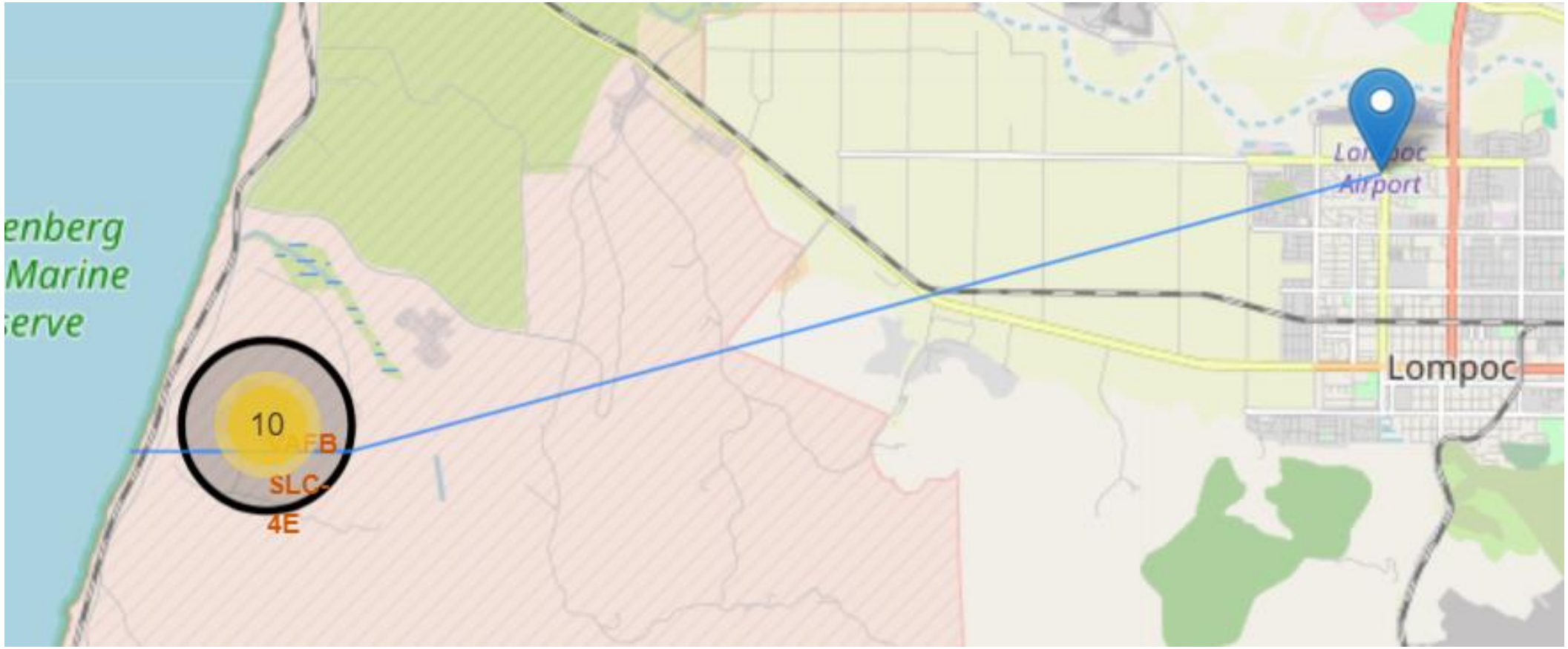


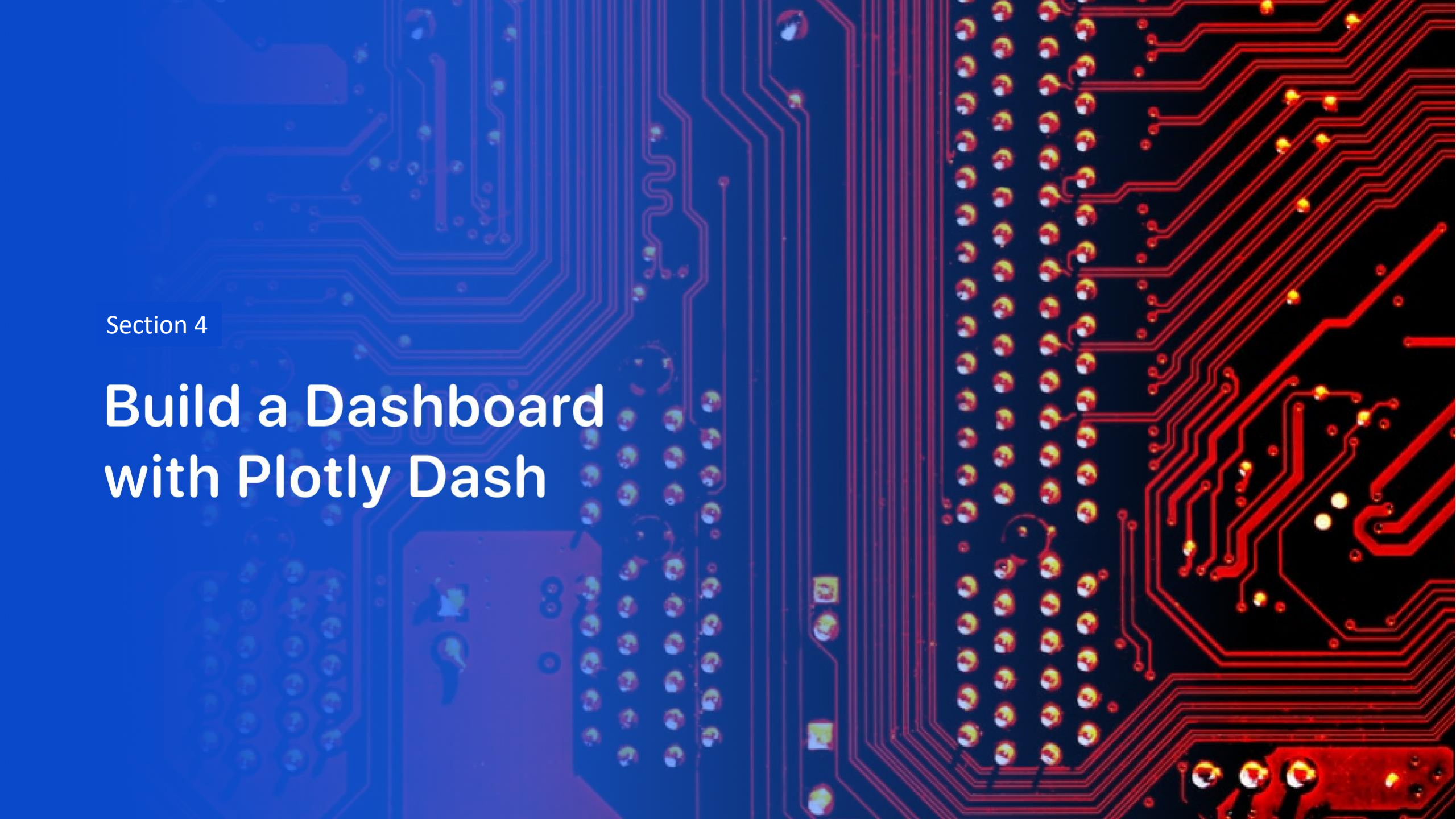




Launch Site VAFB SLC-4E

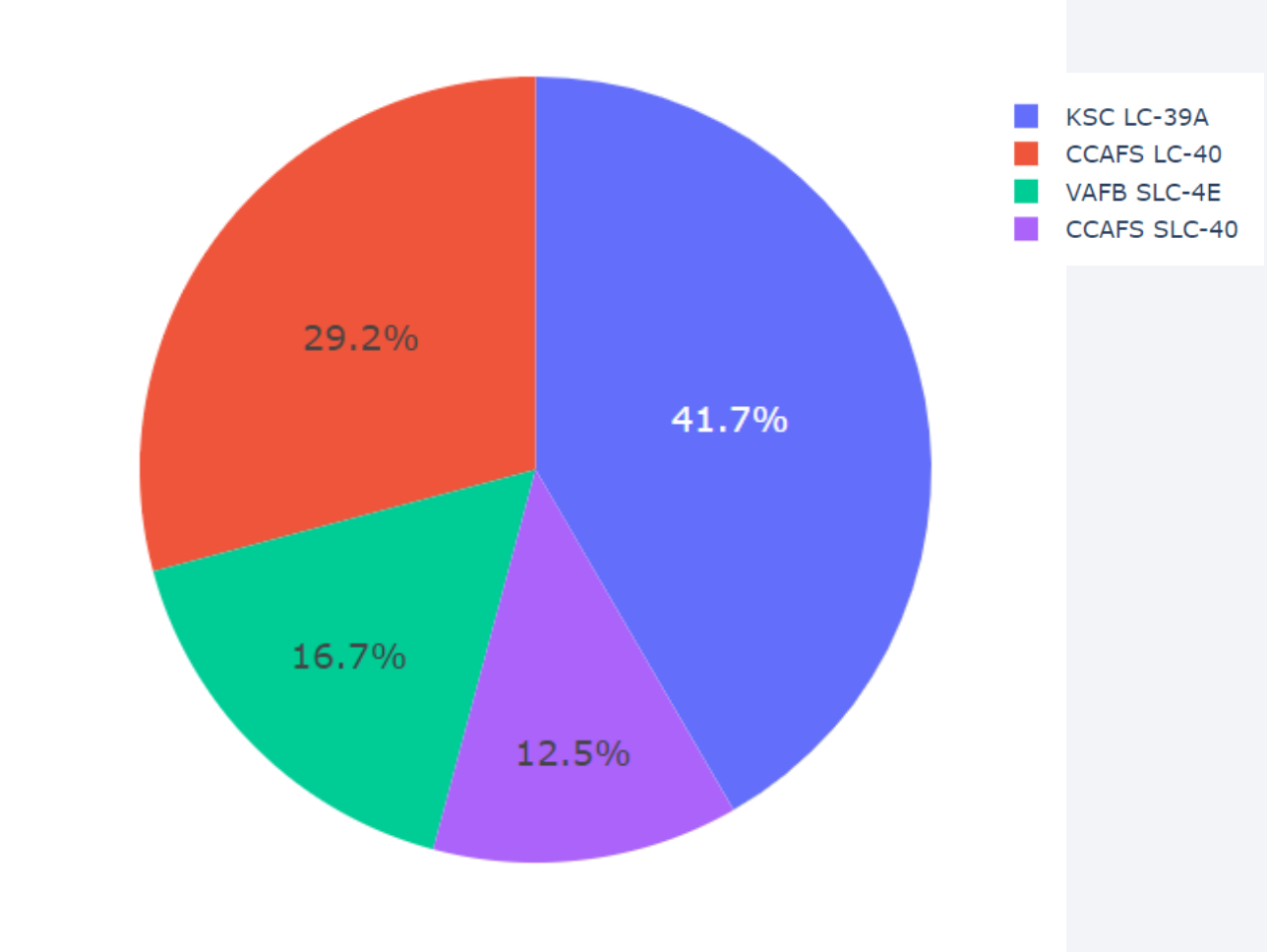
Proximities of nearest coastline and airport.




Section 4
Build a Dashboard with Plotly Dash

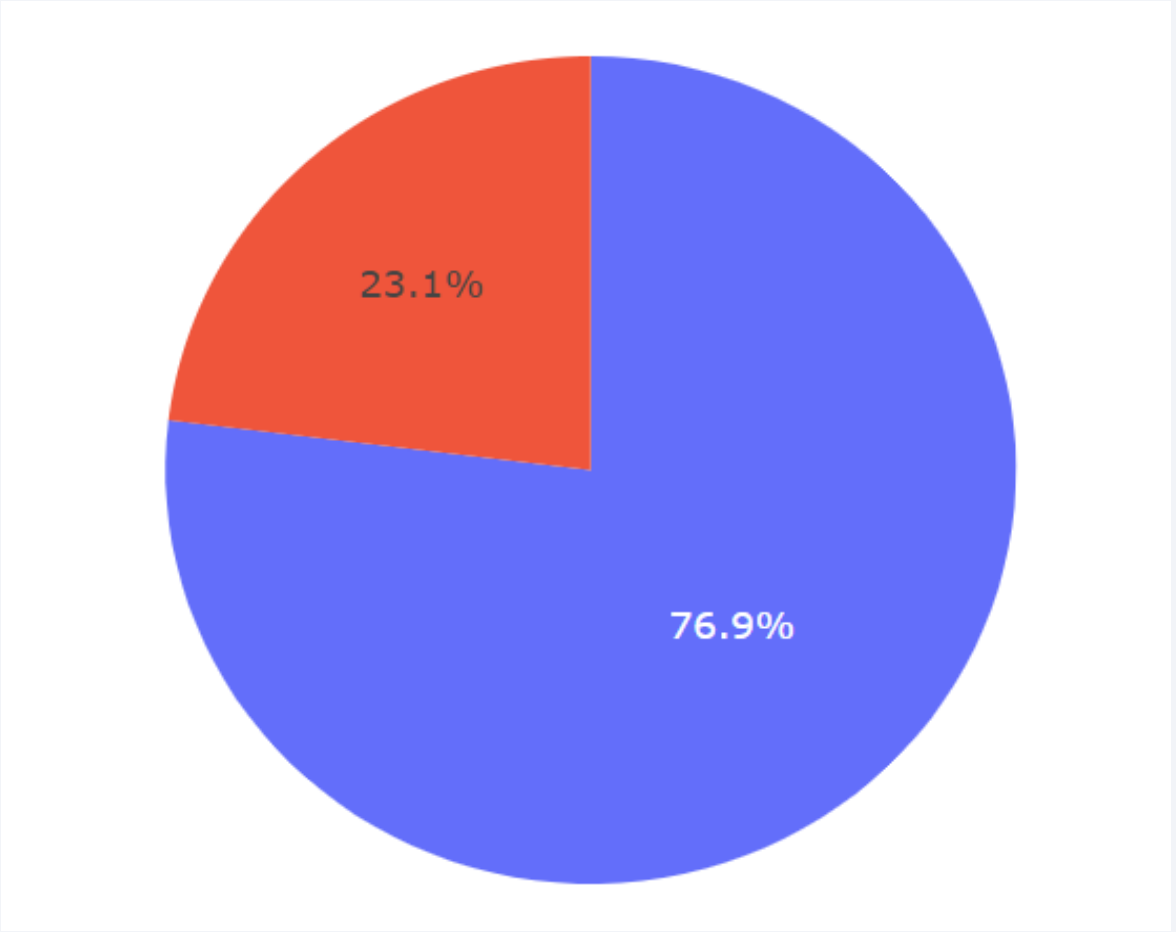
Total of Successful Launches

- Total percentage of successful launches comparing to each other.
- Launch site KSC LC-39A performed best.
- Launch site CCAFS SLC-40 performed worst.



Highest Success Ratio of Launches

- Highest success ration for launches was at launch site KSC LC-39A .
- 76,9% of launches were successful.
- 23,1% of launches failed.



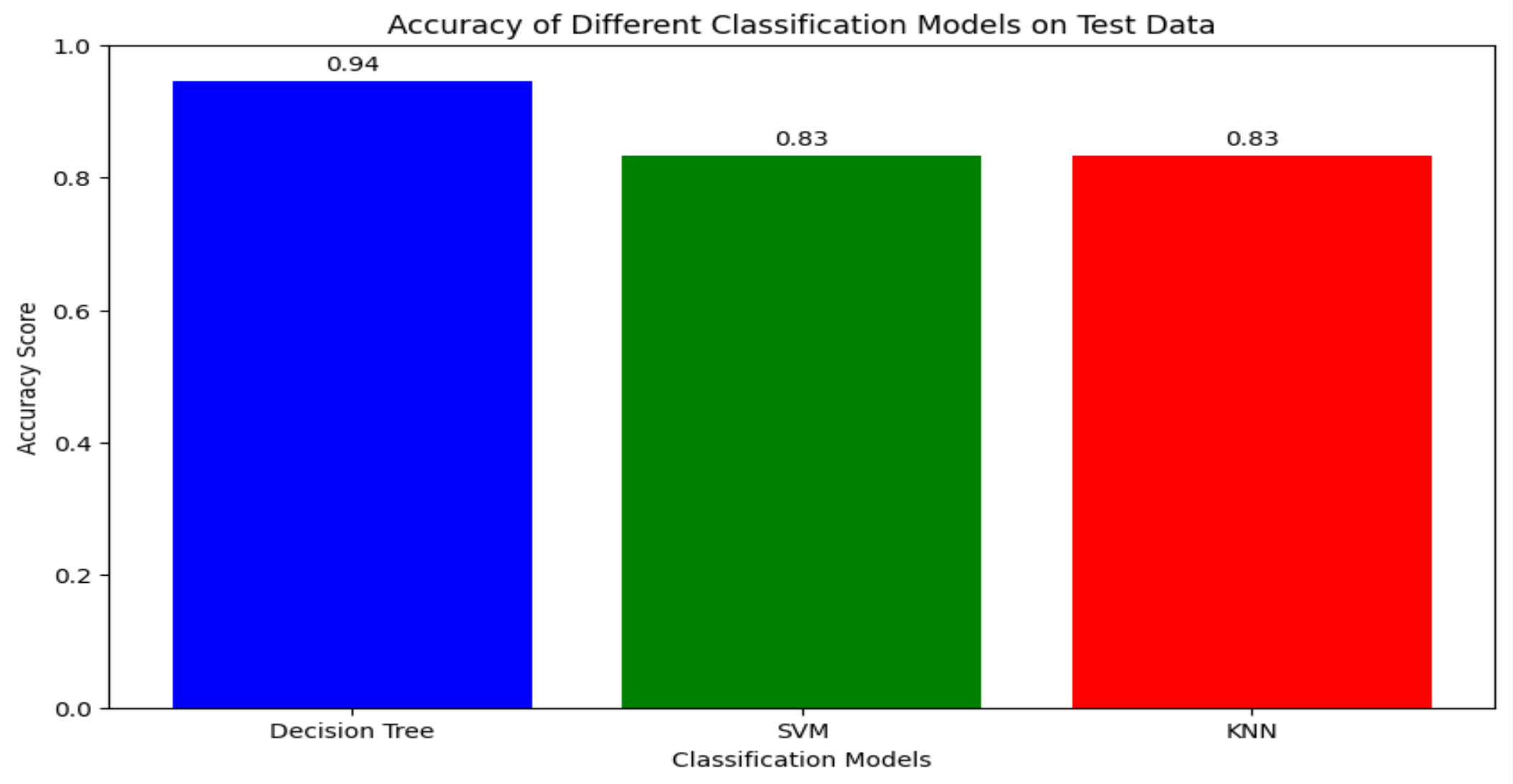
Payload vs. Launch Outcome

- Booster version FT has best success outcome.

Section 5

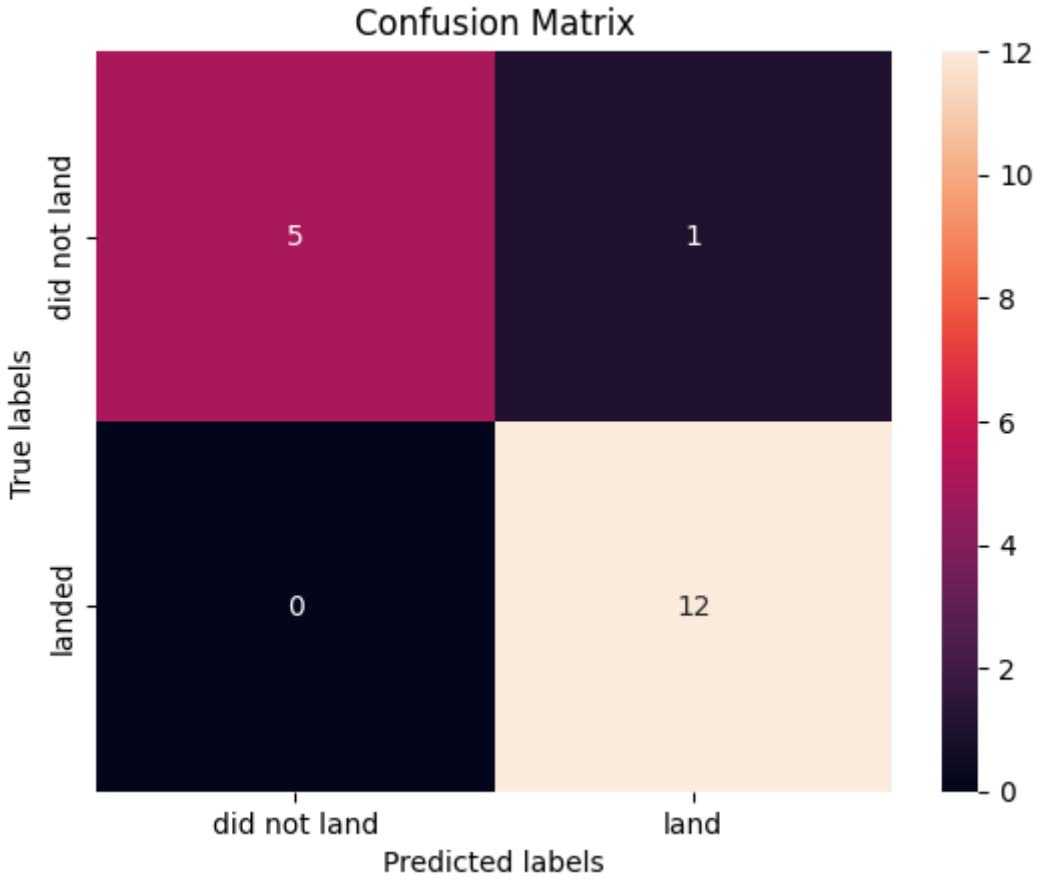
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

- Decision tree performed best.
- Confusion matrix shows that 5 attempts didn’t land and 12 that landed. These both are true values.
- Parts marked with black background color are false positive and negative values, where model failed predict correctly.



Conclusions

- Decision tree performed best, whilst K-Nearest Neighbours and SVM performed equally well.
- Decision tree model failed with a one false positive value and zero false negative value that actually landed.
- Accuracy of decision tree was 94%.

Appendix

- Github repository for all IPNB files.

Thank you!