IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Sai Krishna Vadlamudi>
<May 2023>

Outline

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

Executive Summary

- Summary of methodologies and results
  - **Data collection:** Collected Falcon 9 historical launch records through SpaceX-API (get request to the SpaceX API), and Webscraping SpaceX records from Wikipedia.
  - **Data Wrangling:**
  - **Exploratory Data Analysis:** Using Python and SQL queries calculated and displayed
    - a) the number of launches on each site.
    - b) the number and occurrence of each orbit.
    - c) the number and occurrence of mission outcome per orbit type.
    - d) the names of the unique launch sites in the space mission.
    - e) the average and total payload mass carried by boosters.
    - f) the total number of successful and failure mission outcomes.
  - **Data Visualization:** Performed Launch Sites Locations Analysis with Folium and developed Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.
  - **Machine Learning Prediction:** Logistic Regression, SVM, Decision Tree and KNN models.

Introduction

- Project background and context
  - Space X 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 Space X can reuse the first stage.
- Problems you want to find answers
  - Determining 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 space X for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collected Falcon 9 historical launch records through SpaceX-API (get request to the SpaceX API), and
  - Webscraping SpaceX records from Wikipedia.
- Perform data wrangling
  - Replaced missing values by mean values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The training data is divided into validation data, a second set used for training data; then the models are trained and hyperparameters are selected using the function GridSearchCV.

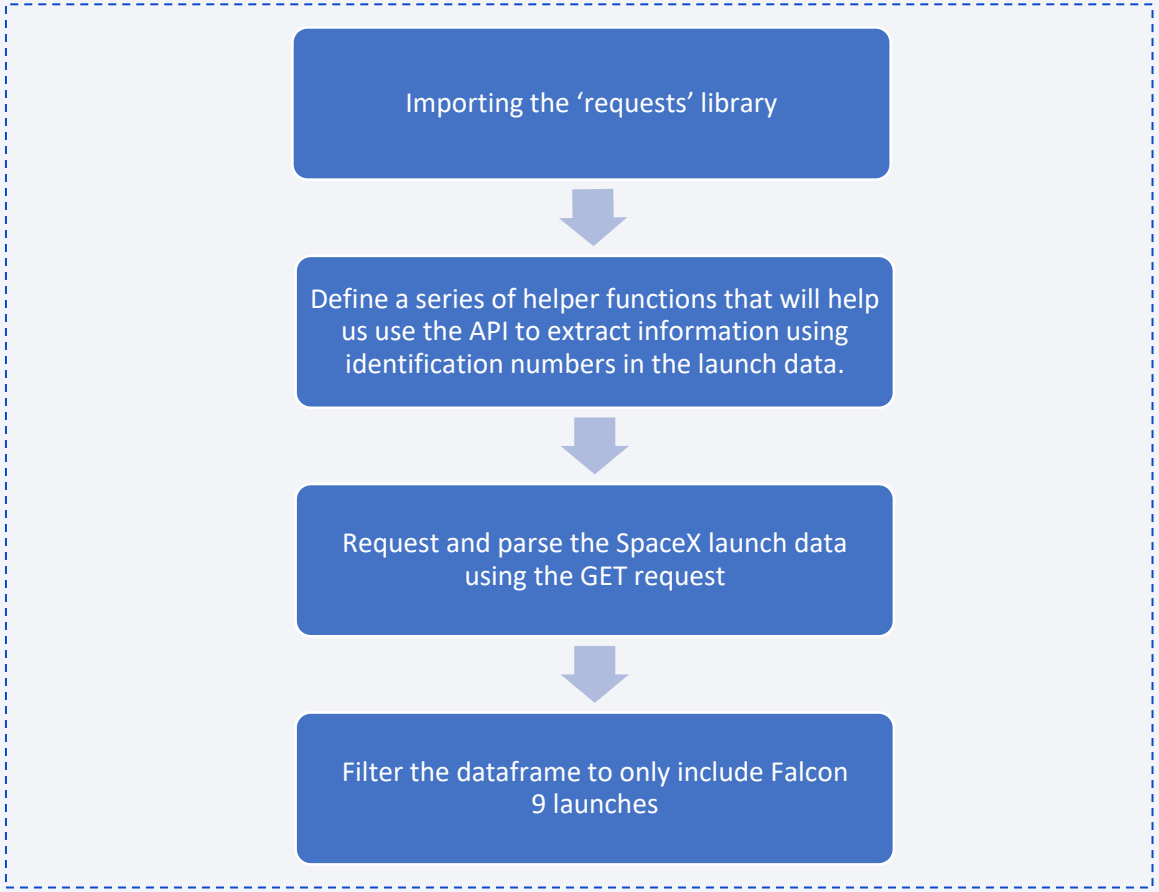
Data Collection

- SpaceX REST API
  - Data is extracted by making a get request to the SpaceX API
  - Defined a series of helper functions that will help us use the API to extract information using identification numbers in the launch data

- Webscraping of SpaceX Wikipedia Page
  - Data is extracted from the launch records that are stored in a HTML table of Wikipedia page titled ‘List of Falcon 9 and Falcon Heavy launches’
    https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

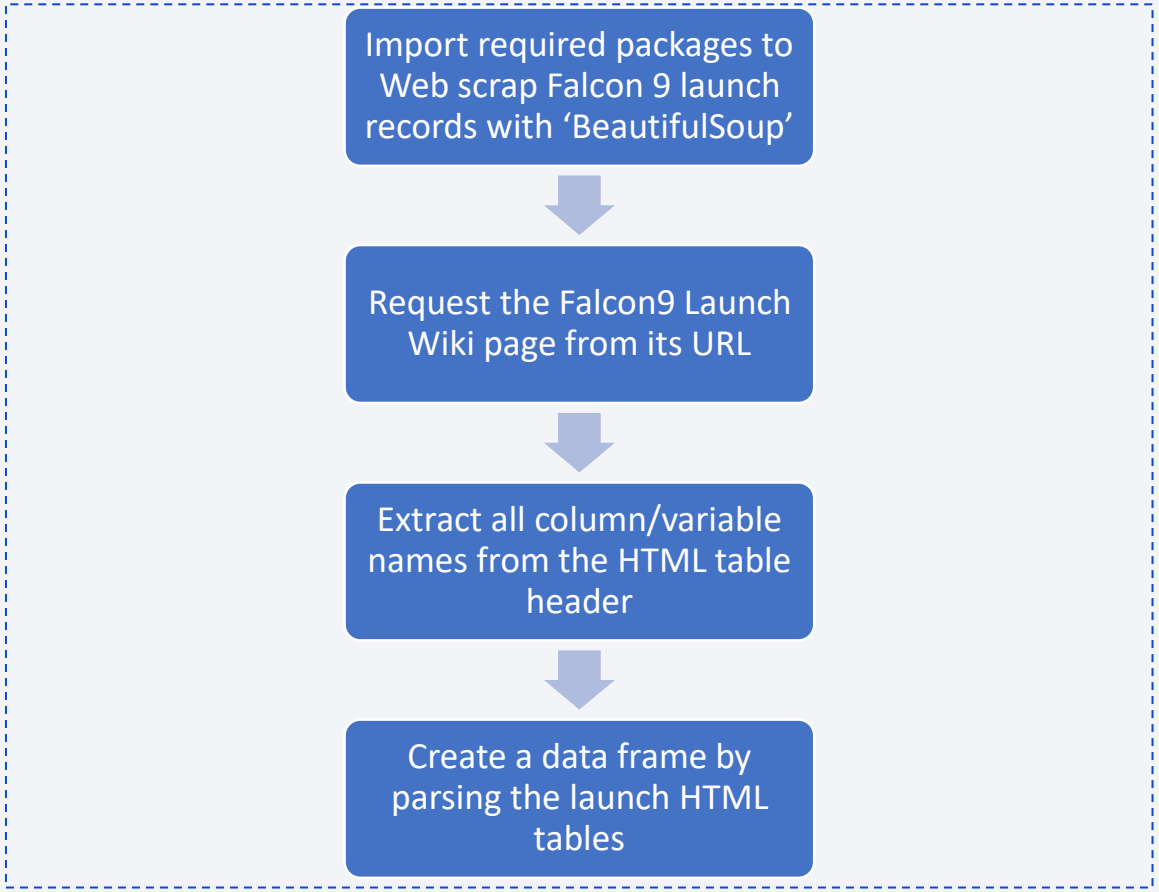
Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook https://github.com/saivadlamudi18/Data-Science/blob/main/Lab%201%20Data%20Collection.ipynb



Data Collection - Scraping

- GitHub URL of the completed web scraping notebook: https://github.com/saivadlamudi18/Data-Science/blob/main/Lab%201a%20Web%20scraping%20Records%20from%20Wikipedia.ipynb

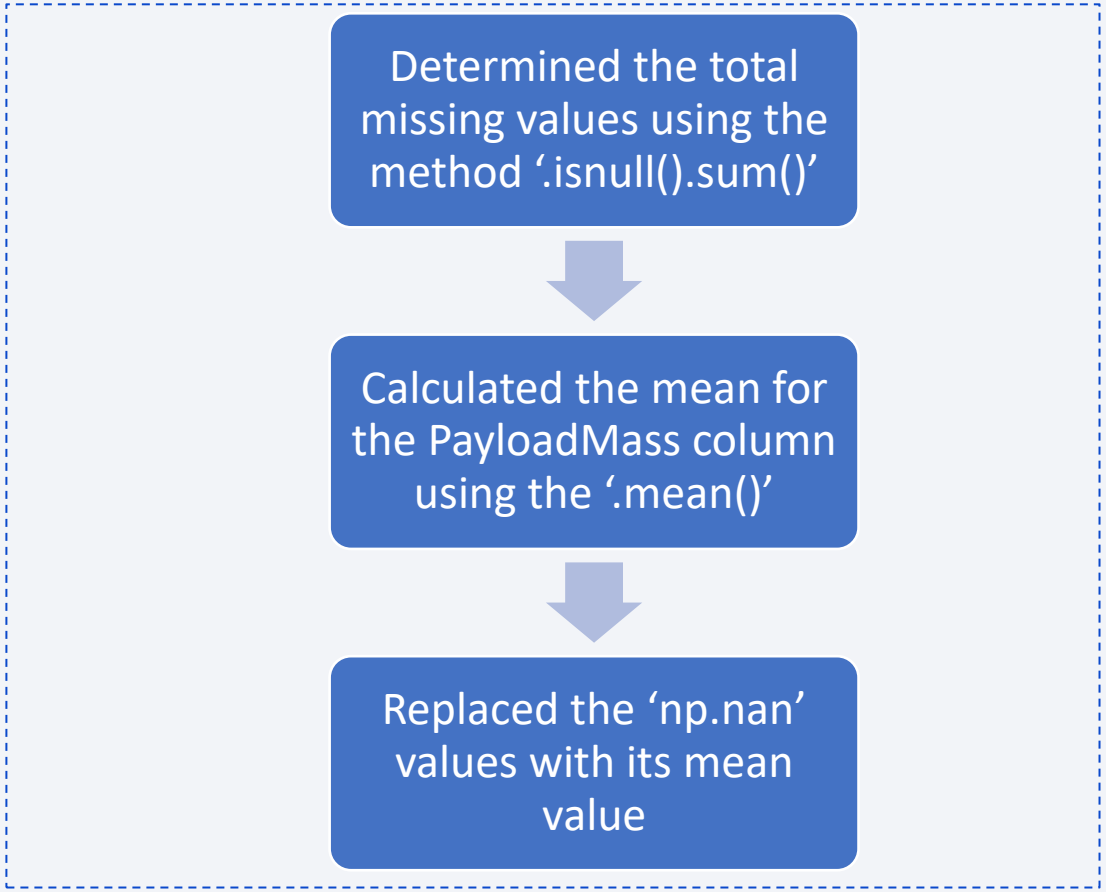


Data Wrangling

- The total missing values in the column were determined and they were replaced by the mean value.
- GitHub URLs of completed data wrangling related notebooks

https://github.com/saivadlamudi18/Data-Science/blob/main/Lab%201%20Data%20Collection.ipynb

https://github.com/saivadlamudi18/Data-Science/blob/main/Lab%202%20Data%20wrangling.ipynb



EDA with Data Visualization

- Charts plotted
  - **Scatter plot: Flight Number vs Payload mass** to see how the FlightNumber (indicating the continuous launch attempts) and Payload mass would affect the launch outcome.
  - **Scatter plot: Flight Number vs Launch Site** to study how the patterns found would affect the launch outcome.
  - **Scatter plot: Payload vs Launch Site** to study which launchsites are used for heavy payloads and vice versa.
  - **Bar plot** to see the relationship between **success rate of each orbit type**(to find which orbits have high success rate)
  - **Scatter plot: Flight Number vs Orbit type,** for each orbit, to see if there is any relationship between Flight Number and Orbit type
  - **Scatter plot: Payload vs Orbit type** to reveal the relationship between Payload and Orbit type.
  - **Line chart:** to get the launch success yearly trend.
- GitHub URL of completed EDA with data visualization notebook, https://github.com/saivadlamudi18/Data-Science/blob/main/Lab%204%20EDA%20with%20Visualization.ipynb

EDA with SQL

- Displayed the names of the unique launch sites in the space mission.
- Displayed 5 records where launch sites begin with the string 'CCA'.
- Displayed the total payload mass carried by boosters launched by NASA (CRS).
- Displayed average payload mass carried by booster version F9 v1.1.
- Listed the date when the first succesful landing outcome in ground pad was acheived.
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listed the total number of successful and failure mission outcomes.
- Listed the names of the booster versions which have carried the maximum payload mass.
- Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- GitHub URL of your completed EDA with SQL notebook, https://github.com/saivadlamudi18/Data-Science/blob/main/Lab%203%20EDA%20with%20SQL.ipynb

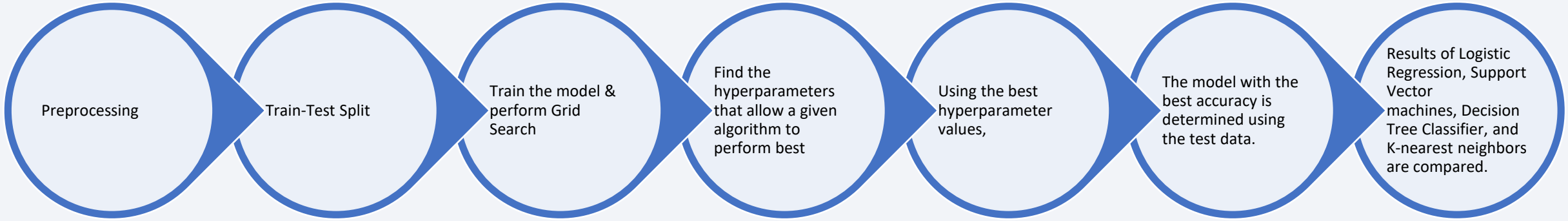
Build an Interactive Map with Folium

- Marked all launch sites on a map to identify their geographical locations.
- Marked the success/failed launches for each site on the map.
- Calculated the distances between a launch site to its proximities to analyze the proximities of launch sites.
- GitHub URL of completed interactive map with Folium map, https://github.com/saivadlamudi18/Data-Science/blob/main/Lab%205%20Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Added a Launch Site Drop-down Input Component
- Added a callback function to render success-pie-chart based on selected site dropdown
- Added a Range Slider to Select Payload
- Added a callback function to render the success-payload-scatter-chart
- GitHub URL of your completed Plotly Dash lab, https://github.com/saivadlamudi18/Data-Science/blob/main/Lab%205a%20Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash

Predictive Analysis (Classification)

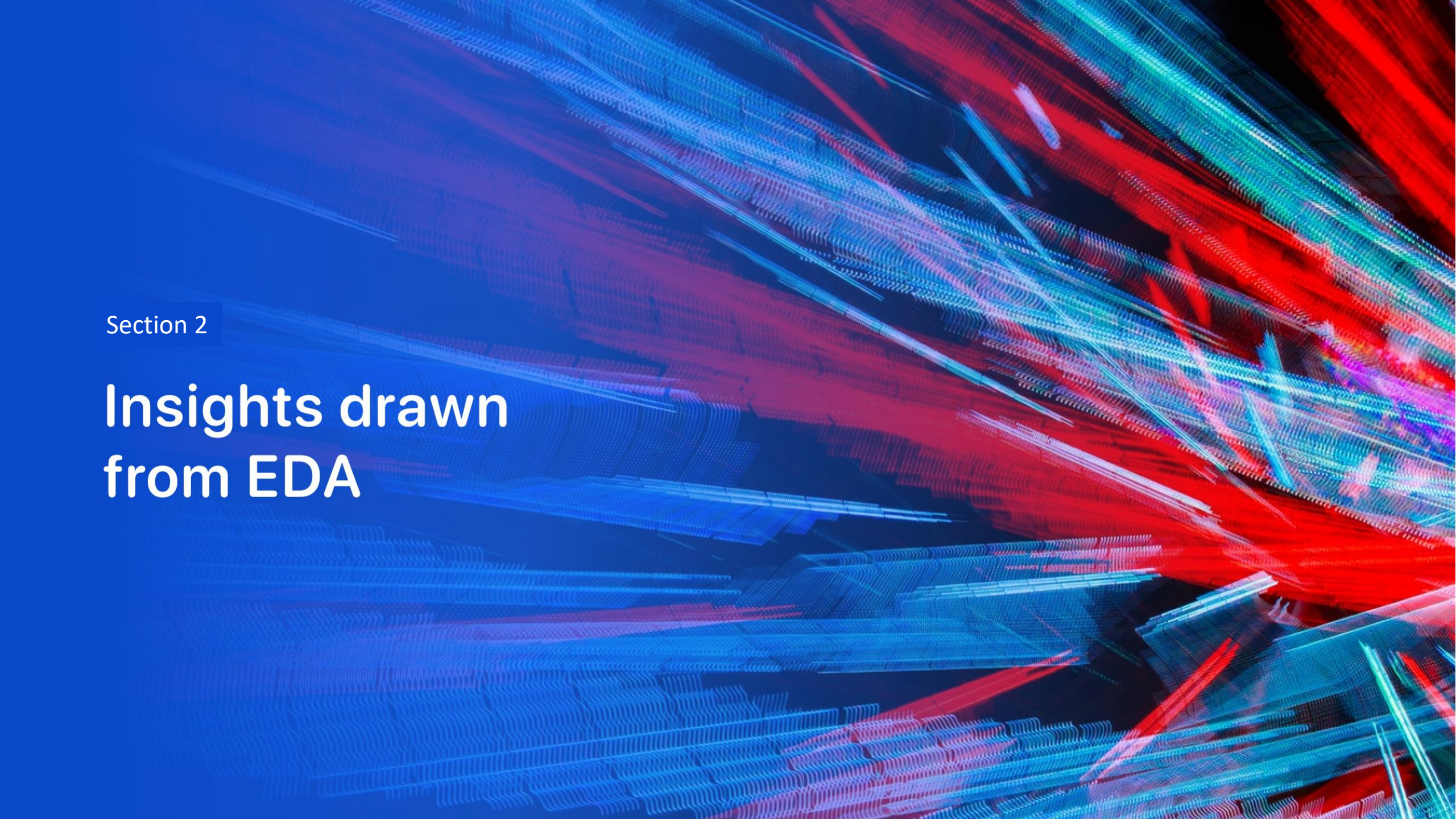


- A machine learning pipeline is built to predict if the first stage of the Falcon 9 lands successfully.
- The data is split into training and testing data. We will train the model and perform Grid Search, allowing us to find the hyperparameters that allow a given algorithm to perform best.
- Four classification models are tested. Using the best hyperparameter values, the model with the best accuracy using the test data is determined.

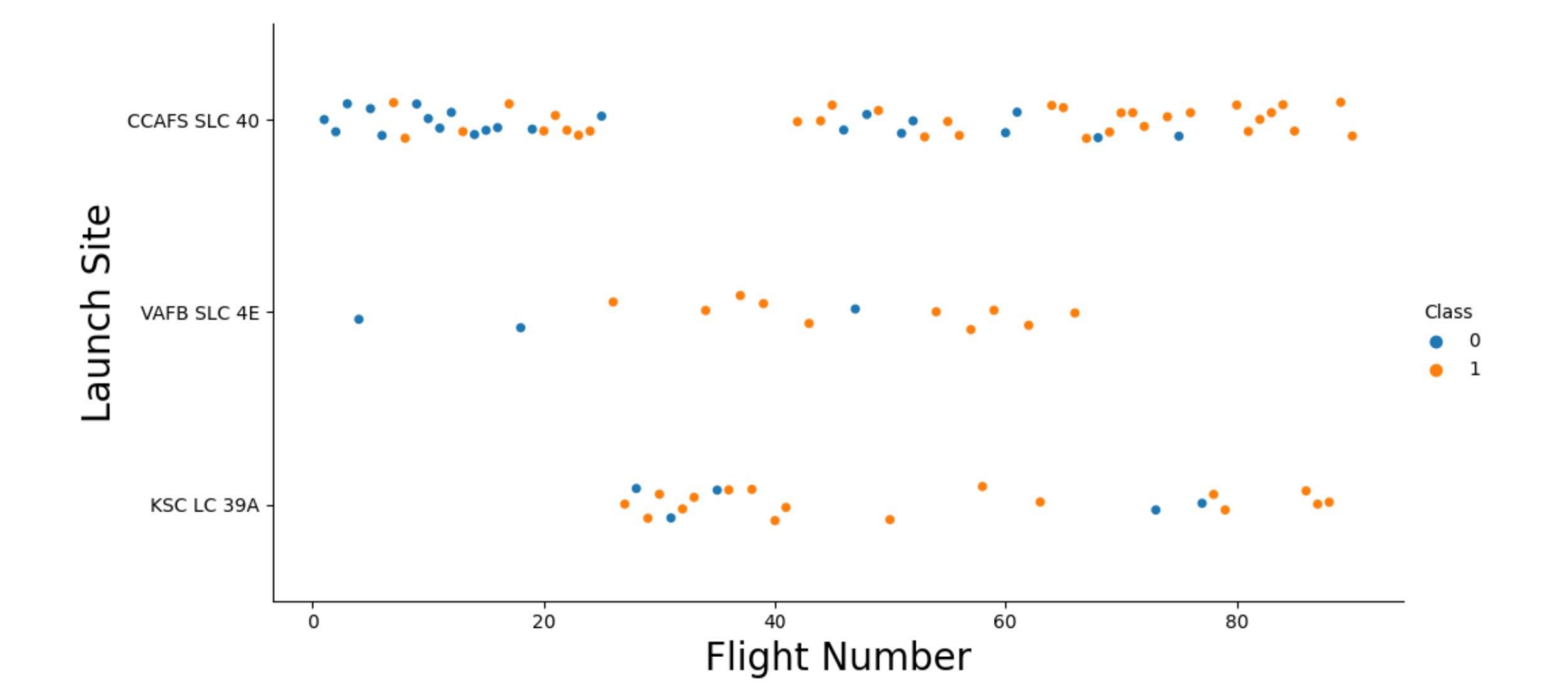
https://github.com/saivadlamudi18/Data-Science/blob/main/Lab%206%20Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results
  - Success rate increases with increase in flight number across all three launch sites.
  - All launch sites in proximity to the Equator line and coastline.
- Interactive analytics demo in screenshots
  - KSC LC-39A can be considered as best launch site because above 75% of launches were successful.
- Predictive analysis results
  - Prediction with models Logistic Regression, SVM and KNN achieved accuracy over 84% on the test data which was better than the accuracy achieved by using decision trees model.

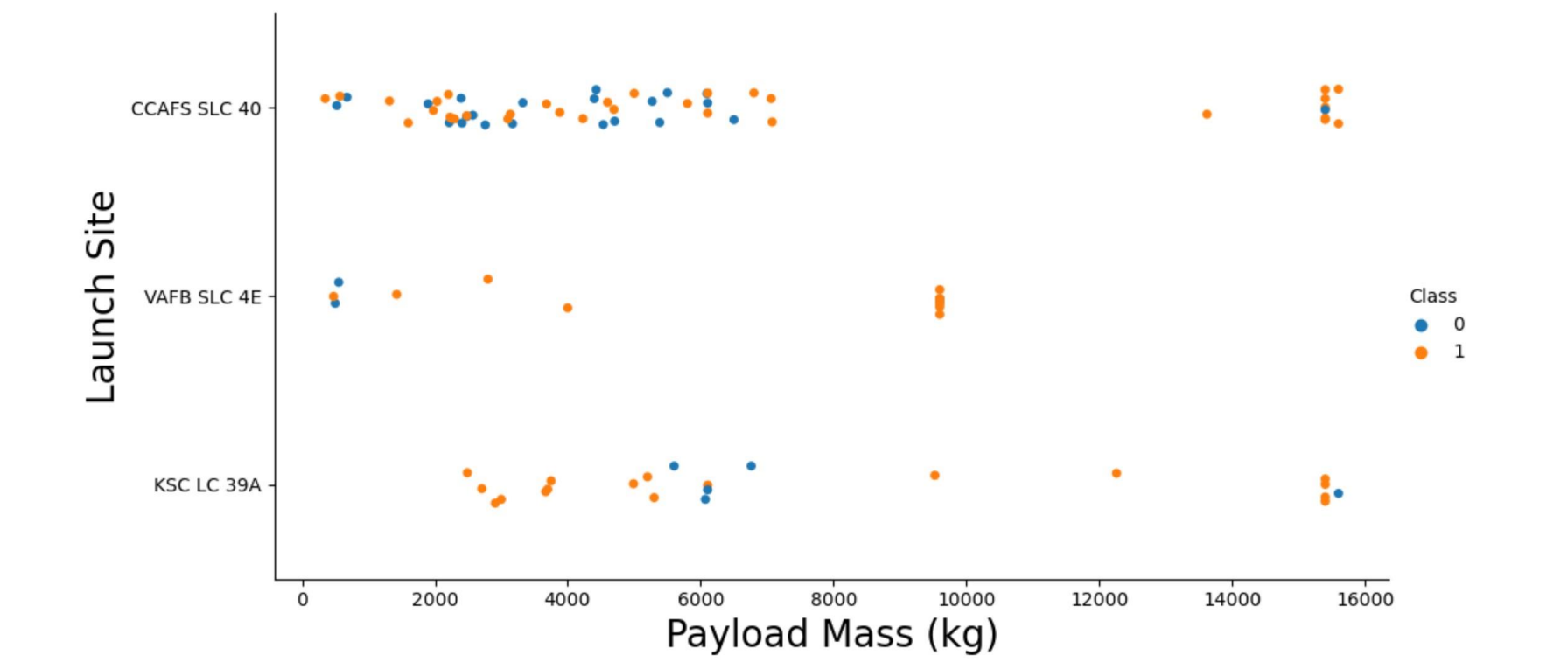

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



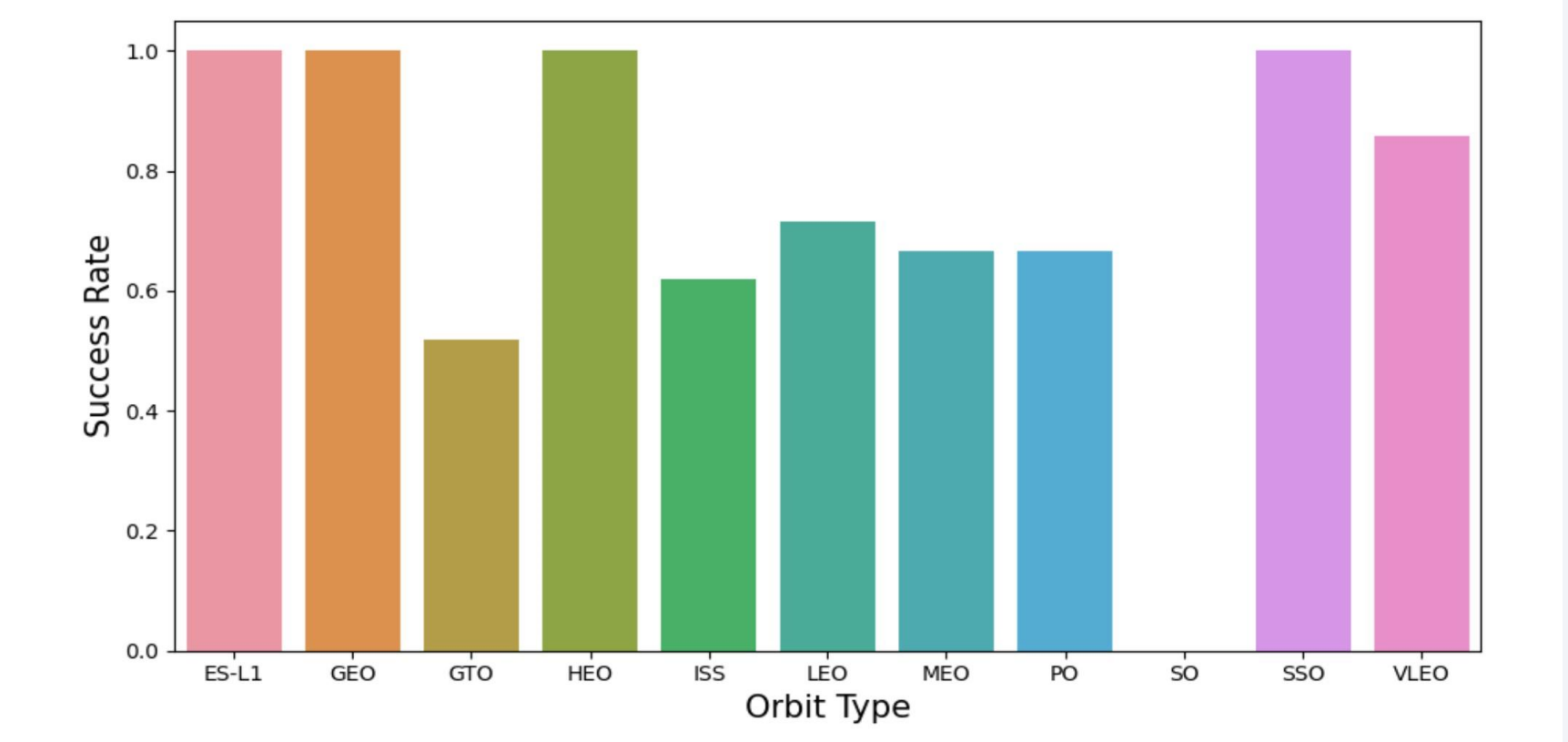
- Across all three sites it is noticed that success rate increases with increase in flight number

Payload vs. Launch Site



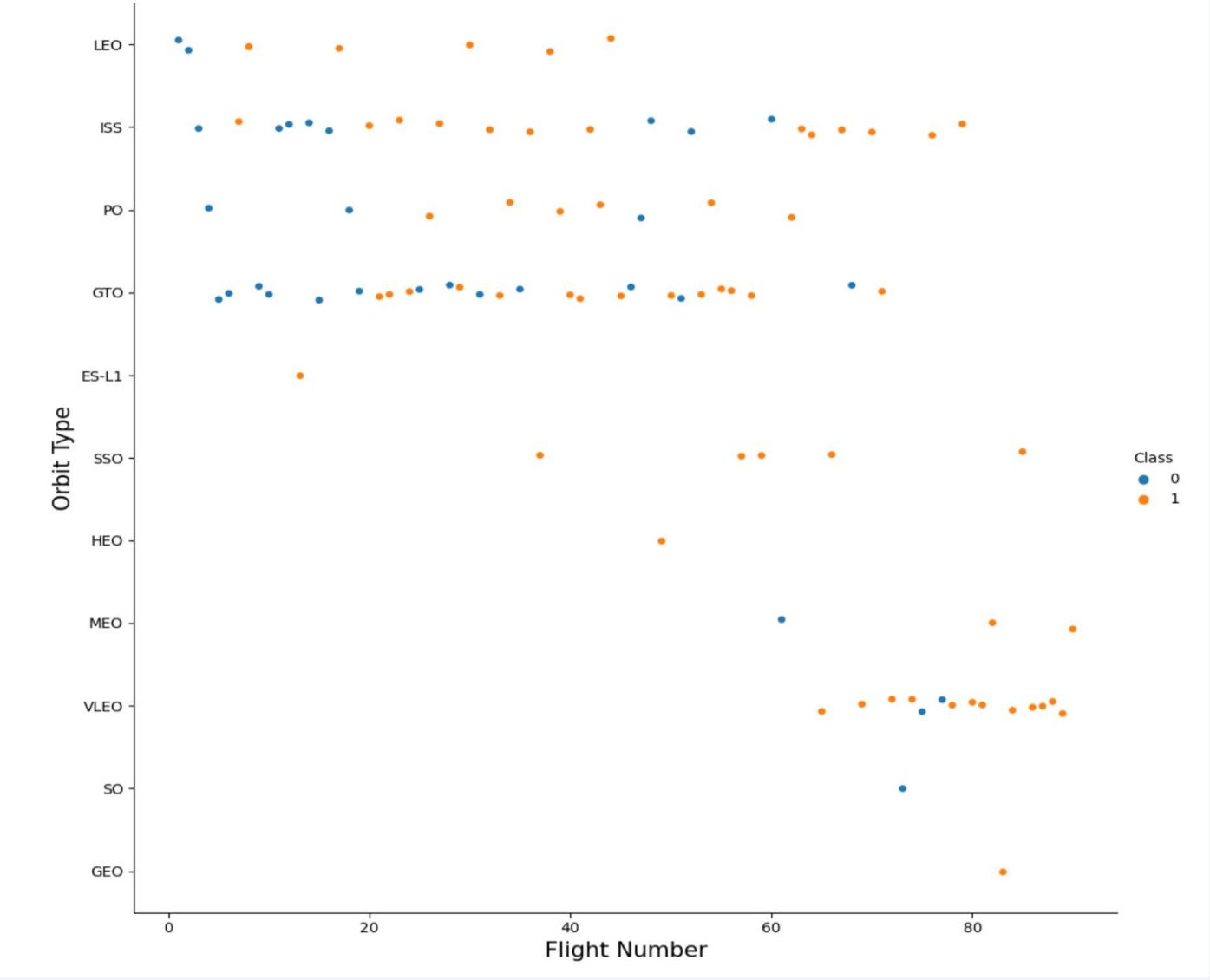
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



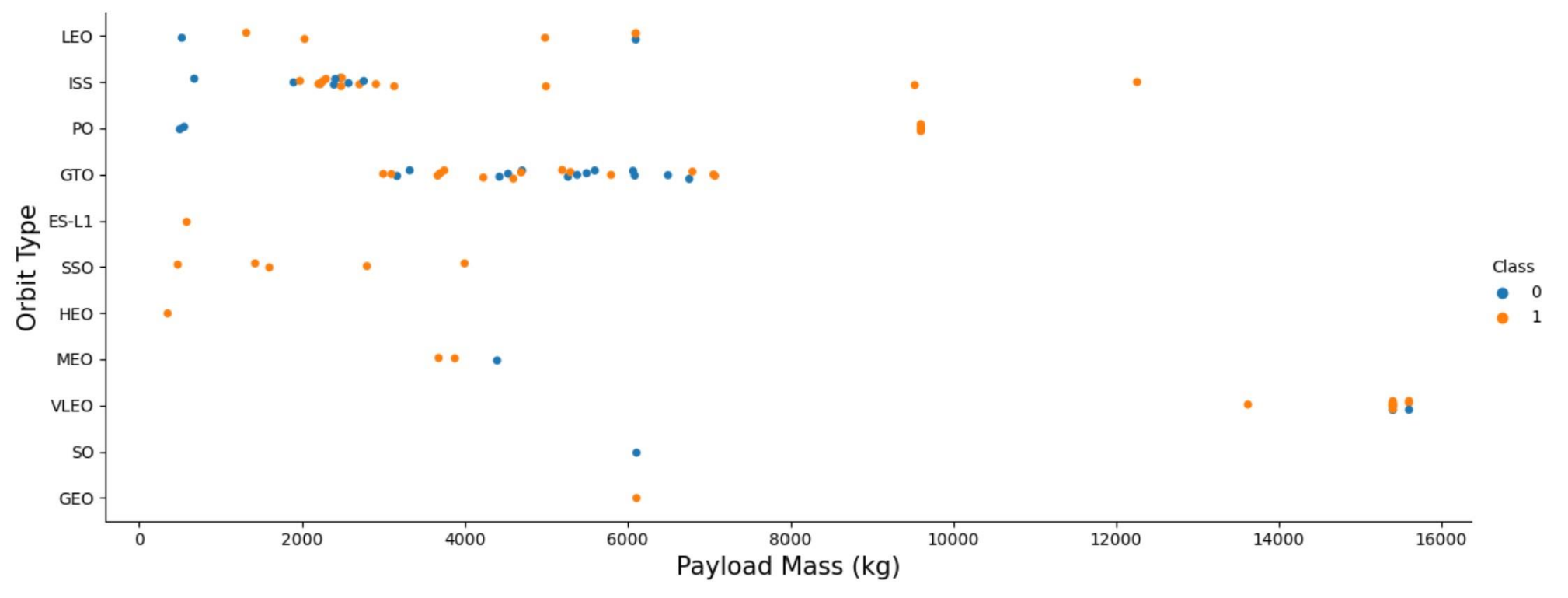
- Orbits ES-L1, GEO, HEO and SSO have the highest(100%) success rate.

Flight Number vs. Orbit Type



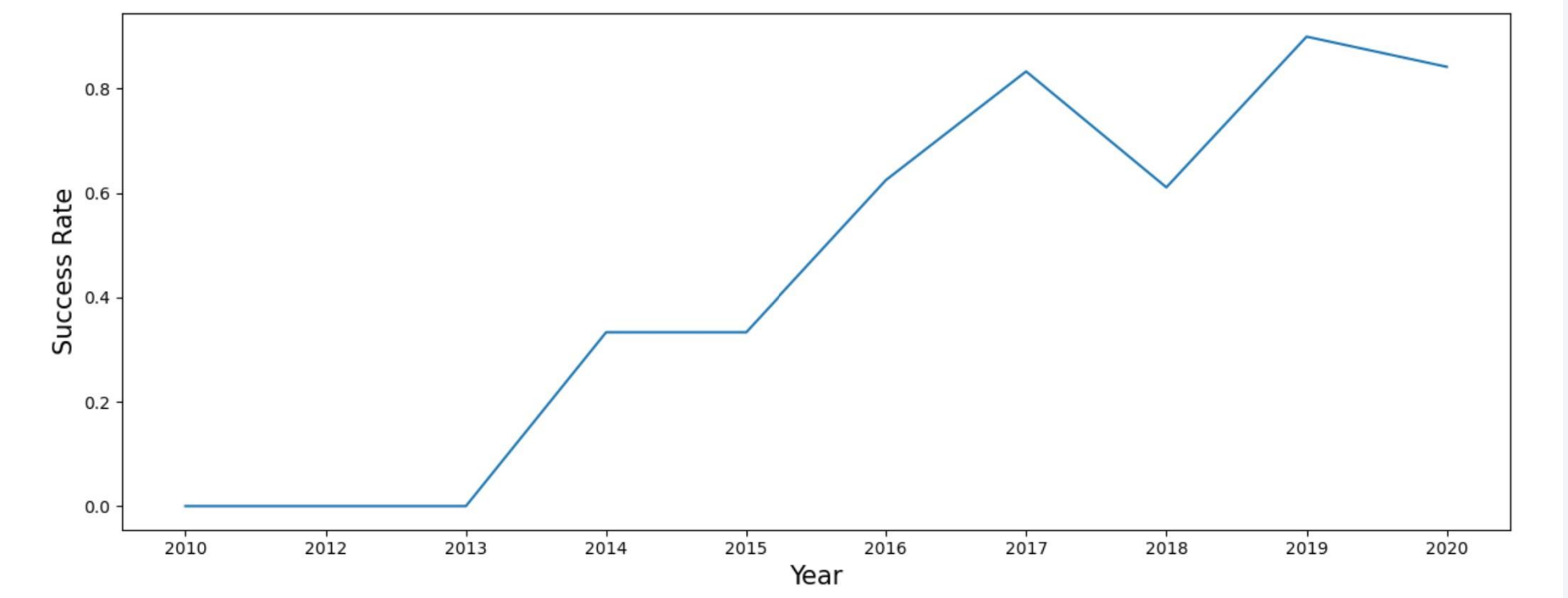
- In the LEO orbit, the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



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

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2017

All Launch Site Names

- Displayed the names of the unique launch sites in the space mission
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

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

Launch Site Names Begin with 'CCA'

- **Displayed 5 records where launch sites begin with the string 'CCA'.**

Out[8]:

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

- Calculated the total payload carried by boosters from NASA

Out[9]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculated the average payload mass carried by booster version F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

- Calculated the total number of successful and failure mission outcomes

```
In [13]: %sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight
 * sqlite:///my_data1.db
Done.
```

Out[13]:

count(MISSION_OUTCOME)
99

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

```
In [14]: %sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

Out[14]:

Booster_Version
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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

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

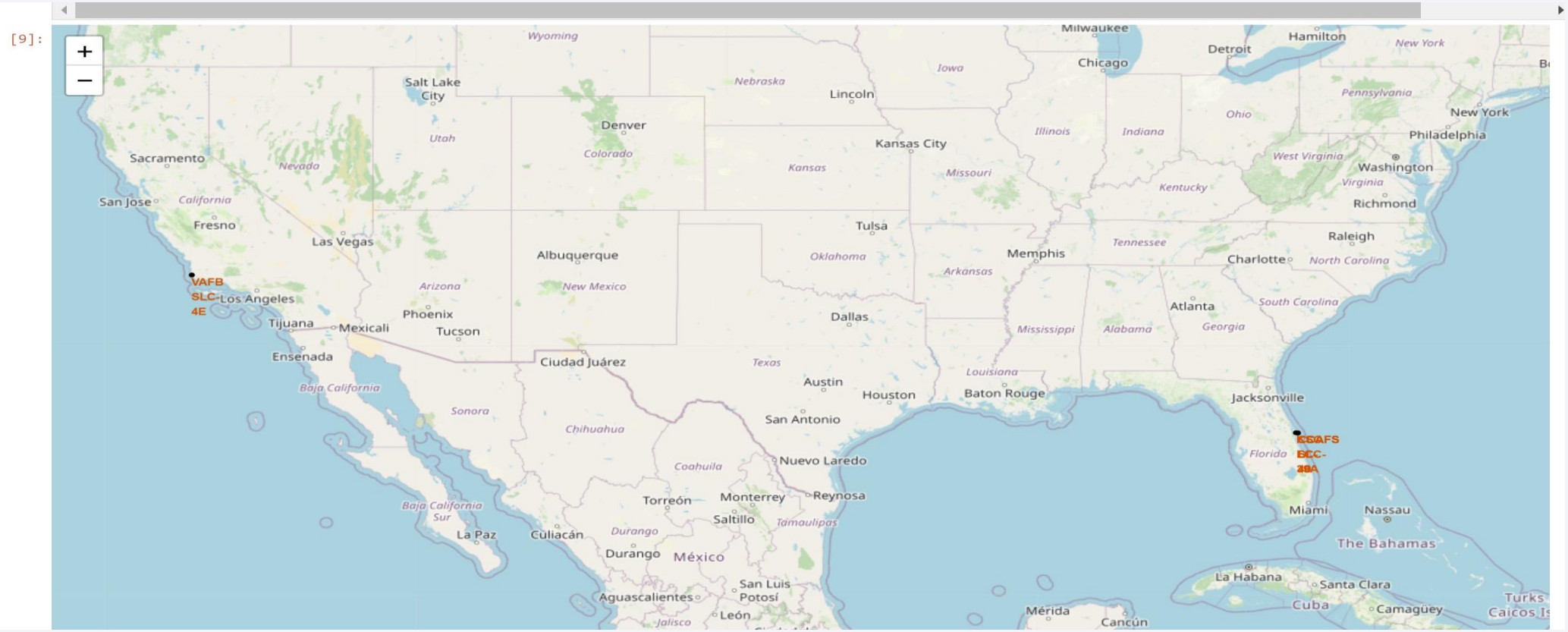
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

landing_outcome	qty
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

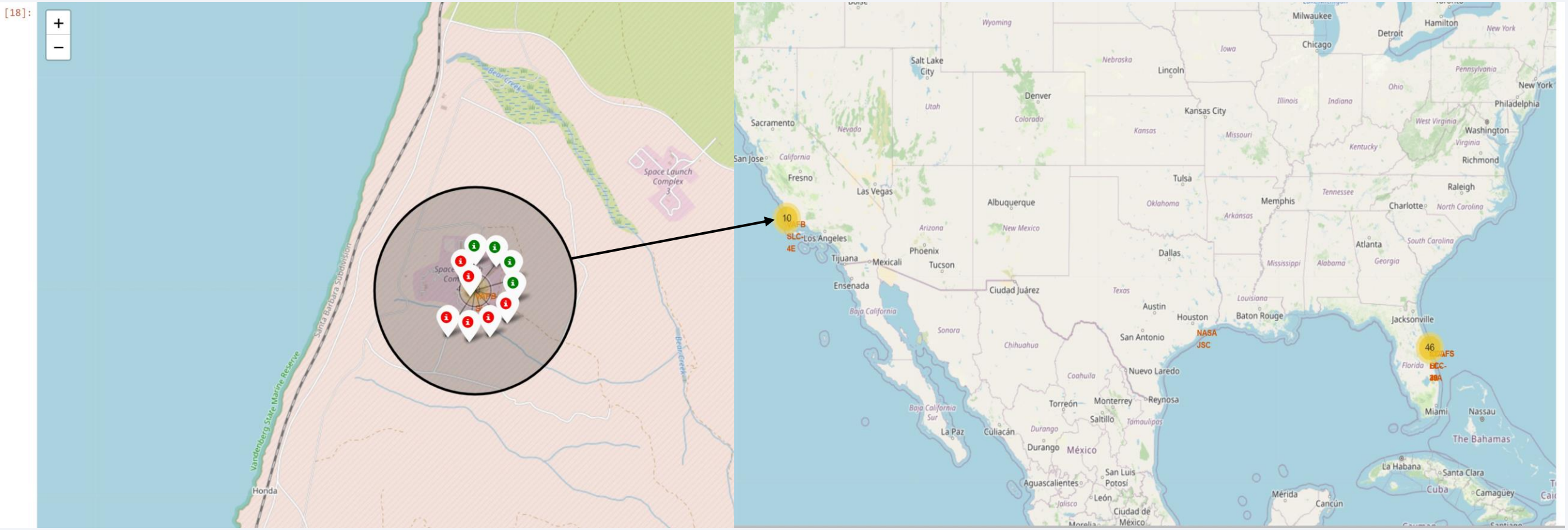
Launch Sites Proximities Analysis

Launch sites on a map



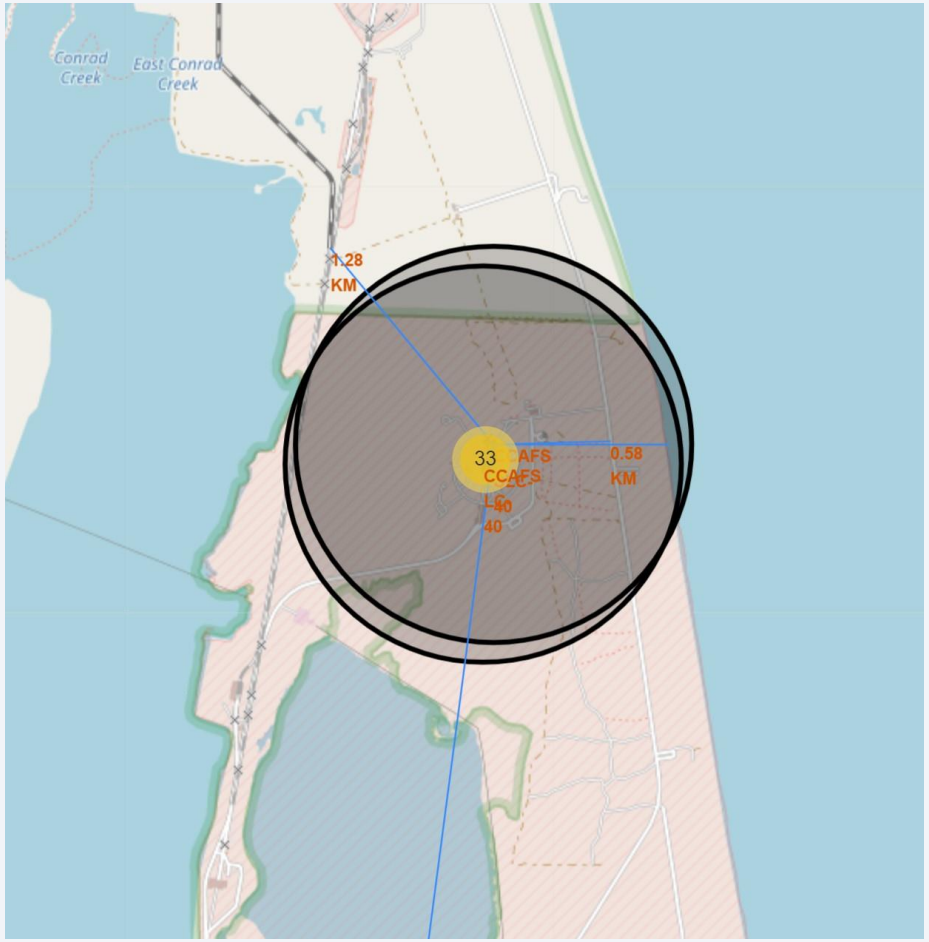
- All launch sites in proximity to the Equator line which means less fuel to get into space from the equator due to the Earth's rotation.
- All launch sites in very close proximity to the coastline for safety reasons

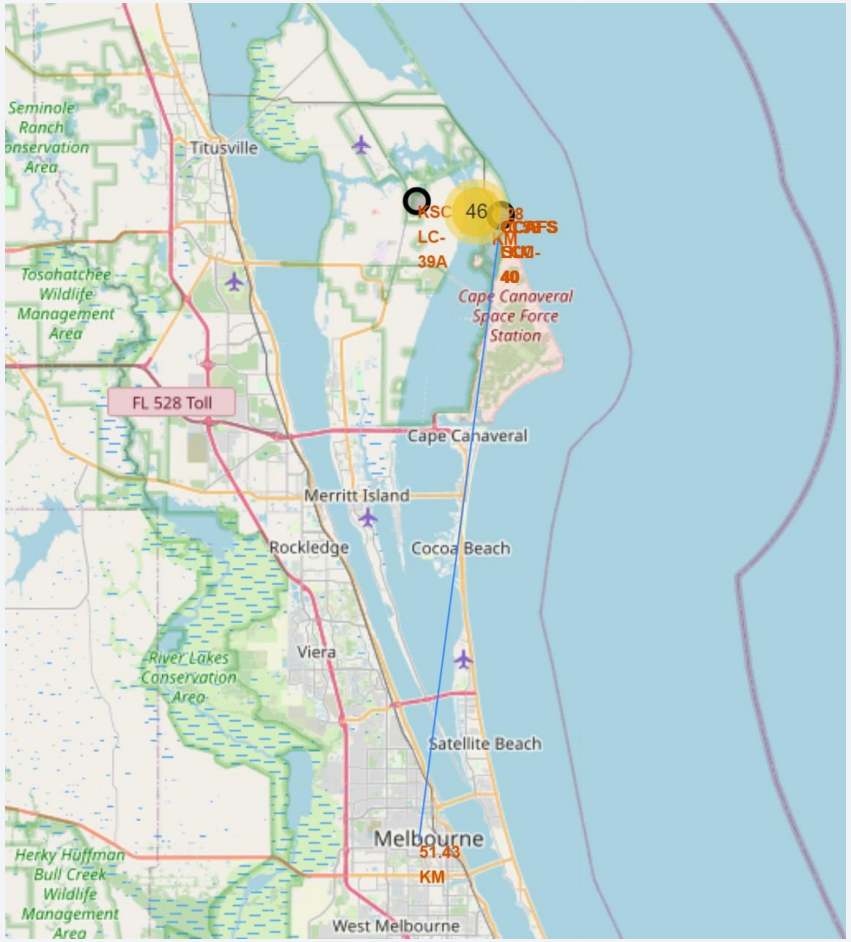
Success/failed launches for each site on the map



- From the color-labeled markers in marker clusters, it is easy to identify which launch sites have relatively high success rates. The green markers indicate success and red ones indicate failed launches.

Distances between a launch site to its proximities

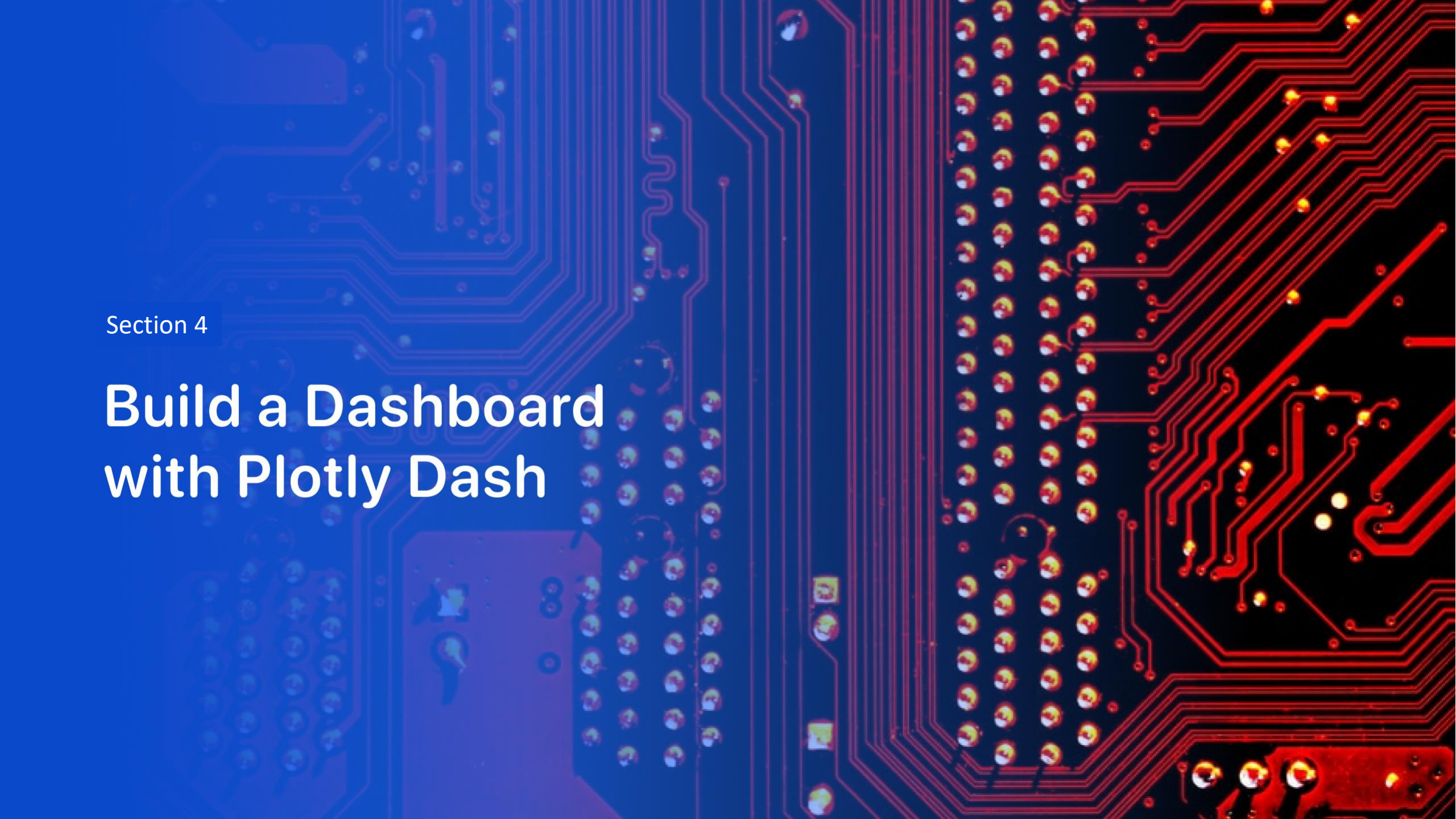




- Figures representing the distance between Launch sites and nearest city, Railway, Highway and seacoast.

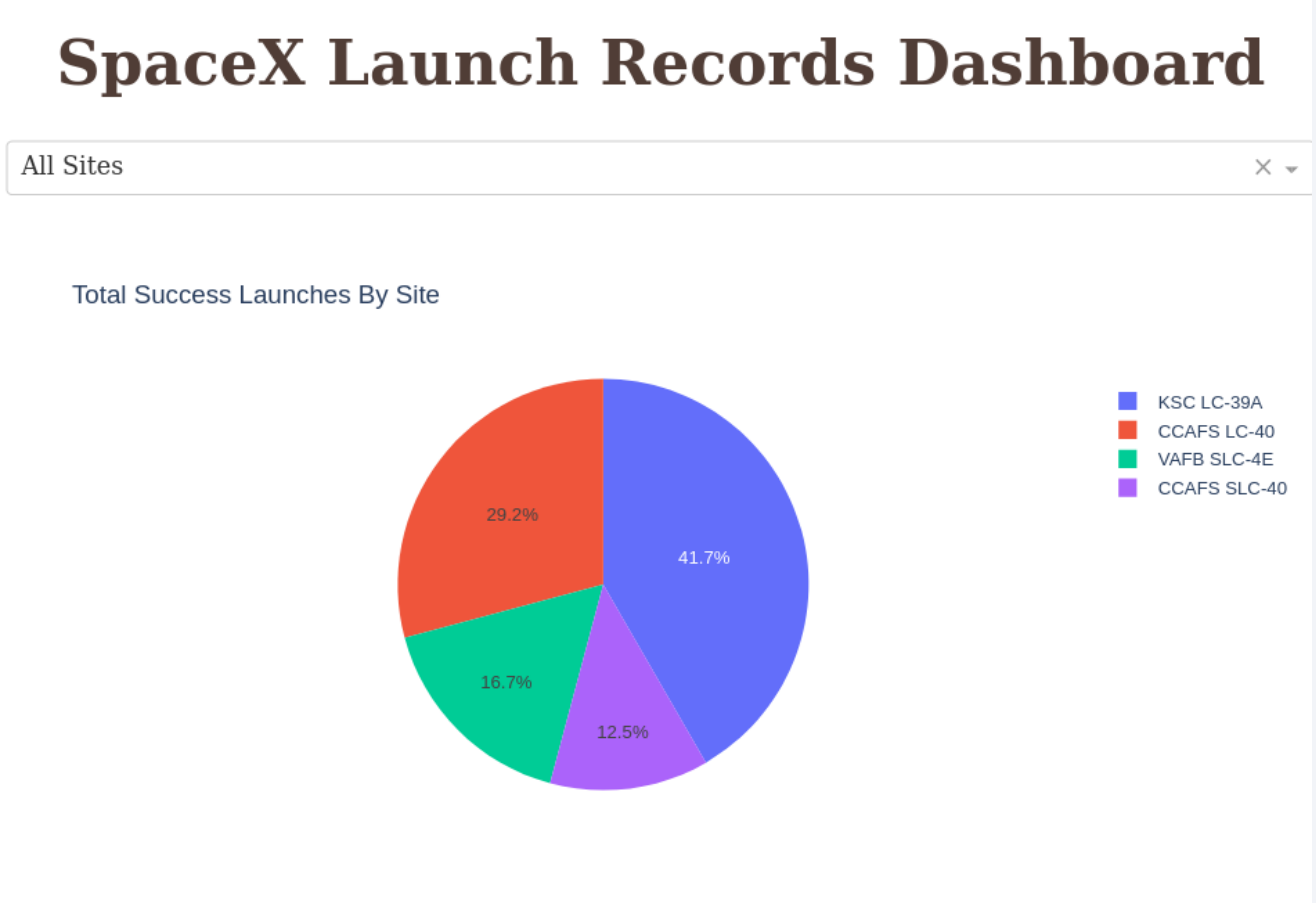
Distances between a launch site to its proximities

- Launch sites are in close proximity to railways and highways, so that it facilitates transporting heavy cargo and workers to the launch site.
- Launch sites keep certain distance away from cities so that it reduces risk from noise and air pollution, and serious hazzards to population living in urban areas.
- Close proximity to coastline also adds weight to the safety factors mentioned above, this benefits the crew if they want attempt water landing and prevents risk from falling debris.

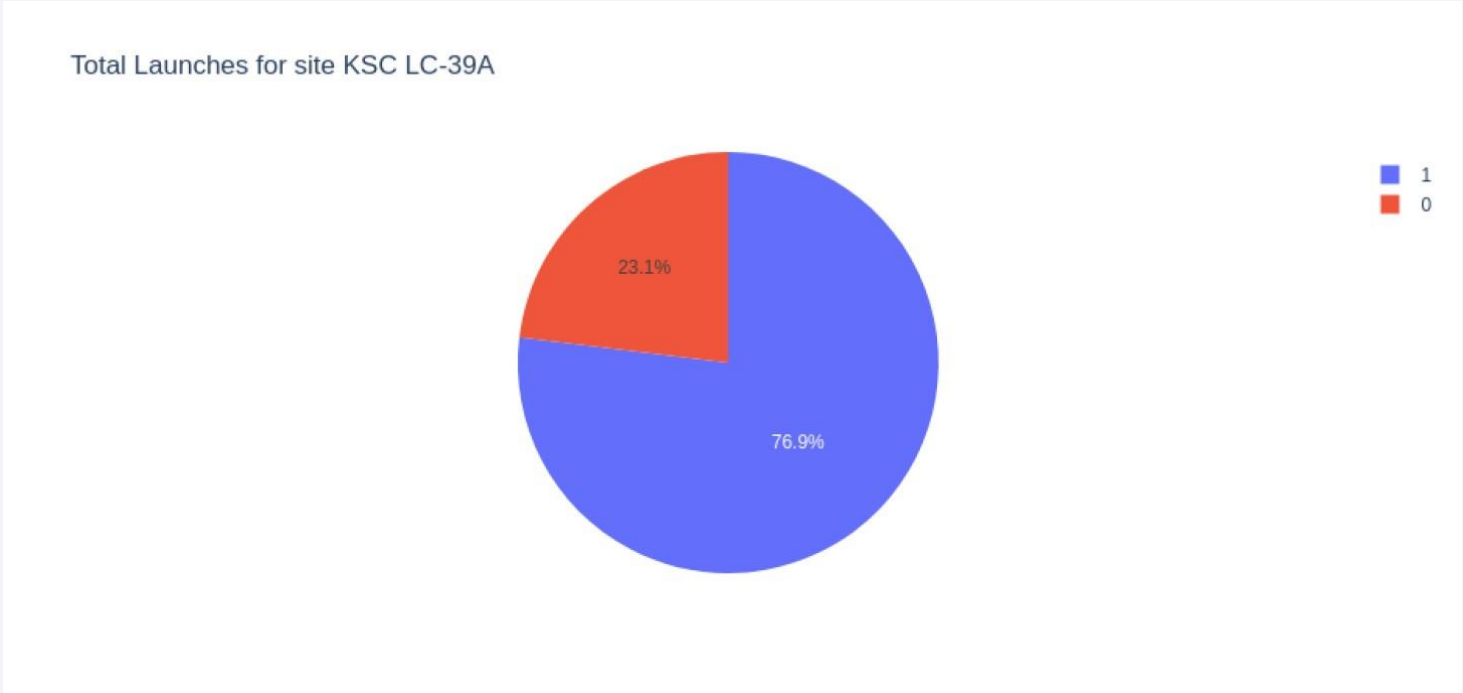

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites

- KSC LC-39A stands top while CCAFS SLC-40 stands last in the total successful launches.

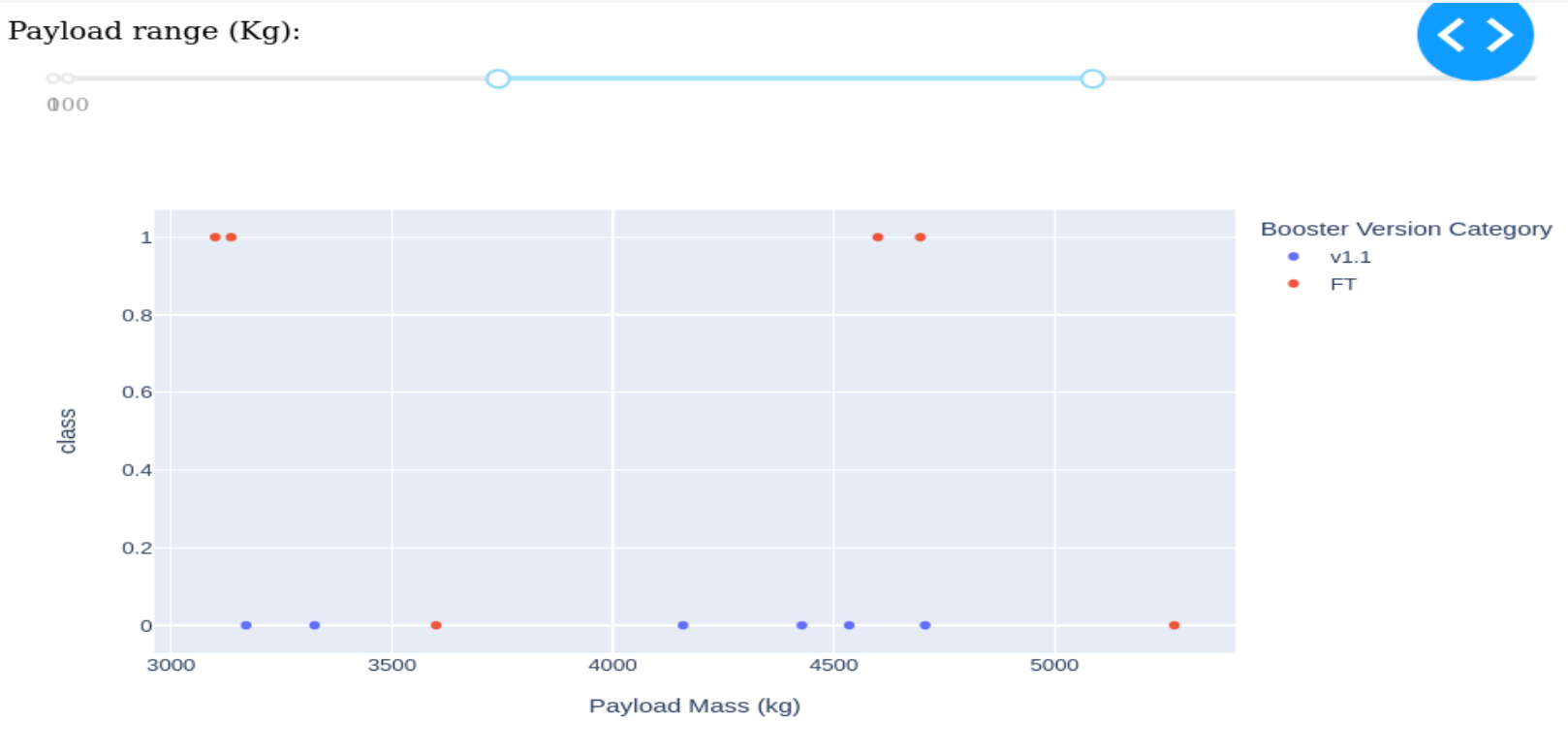


Launch site with highest launch success ratio



- More than of 75% of launches in KSC LC-39A were successful

Payload vs. Launch Outcome for all sites



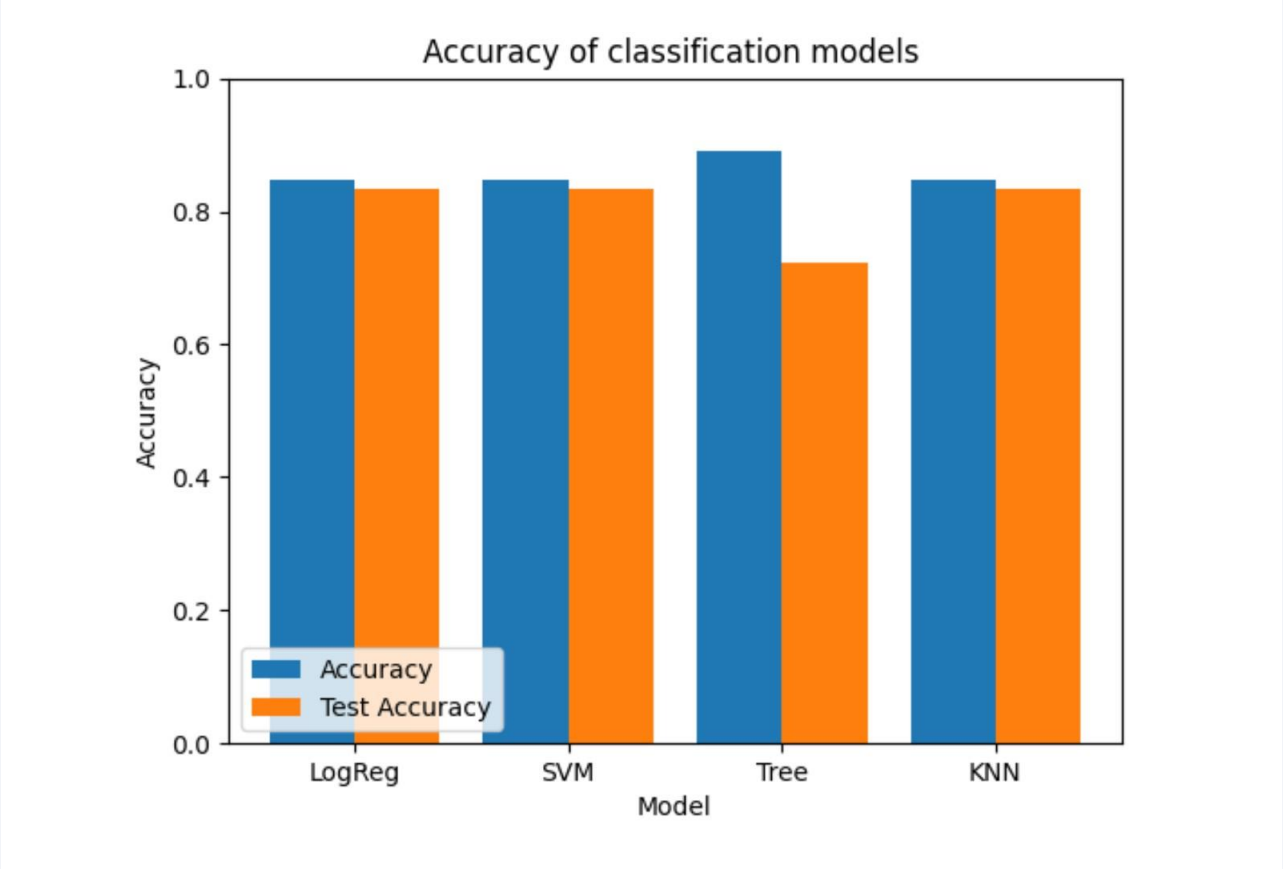
- For the above selected payload range booster version ‘FT’ had the largest success rate.

Section 5

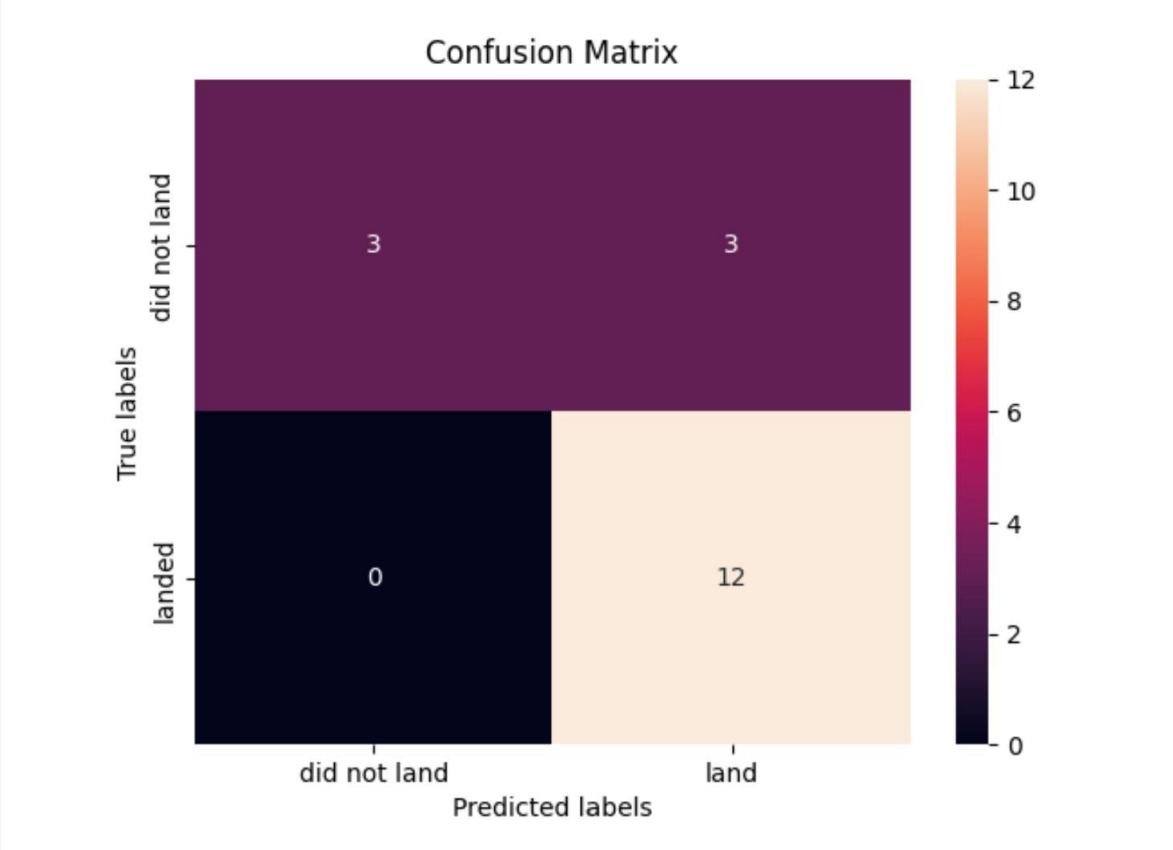
Predictive Analysis (Classification)

Classification Accuracy

- In my study, except the decision tree model, the remaining three models achieved accuracy over 84% on the test data.



Confusion Matrix



Conclusions

- Success rate increases with increase in flight number across all three launch sites.
- KSC LC-39A can be considered as best launch site because above 75% of launches were successful.
- All launch sites in proximity to the Equator line and coastline.
- Prediction with models Logistic Regression, SVM and KNN achieved accuracy over 84% on the test data which was better than the accuracy achieved by using decision trees model.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!