Winning Space Race with Data Science

Areeb Shafqat

7th November 2021

Outline

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

Executive Summary

This report consists of details regarding the capstone data science project carried out which aims to predict whether a first stage Falcon 9 rocket launch is successful or not. After requesting rocket launch data from SpaceX API in a json file and web scraping from Wikipedia for historical launch records, helper functions, the pandas library, Beautiful Soup and others are implemented to obtain flight details such as orbit type, success outcome, launch site and more in a tabular pandas dataframe format. Subsequently, using exploratory data analysis with SQL, data wrangling, location analysis through folium, and other data visualization techniques, a suitable dataset with understanding of relationship between variables and the significance of each was generated. Lastly, machine learning models like logistic regression were trained to predict success or failure for a given Falcon 9 launch.

Some results such as the “KSC LC-39 A” launch site having the highest success rate in the first stage of the launch were discovered and others will be mentioned in the later stages of this report. With the machine learning models, all models used after Hyperparameter tuning (KNNs, SVMs, Decision Trees and Logistic Regression) generated the same results on the test dataset but the one that had the highest accuracy on the training set was the Decision Trees model and hence, overall, it seems to be the winner.

Introduction

- Space X can reuse the first stage of a Falcon 9 rocket launch easily, saving them millions of dollars
- To compete with them, a machine learning model can be built to successfully predict the outcome of the first stage in a Falcon 9 rocket launch
- If a landing is predicted successfully, other providers can also come near the $62 million mark like SpaceX when advertising their rocket launches.
- In order to complete this task, problems like classifying the data into successful and unsuccessful or finding which features affect the success outcome the most need to be tackled.
- Certain relationships between variables may be of interest for example correlations between launch site, payloads, orbit types, gridfins of a rocket and success outcome would help a company craft a successful first stage launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using get requests on the SpaceX Rest API and Web Scraping from Wikipedia
- Perform data wrangling
  - Converting JSON to a dataframe, removing null values, one hot encoding, and classifying outcomes into either 0 or 1
- 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 ML classification models

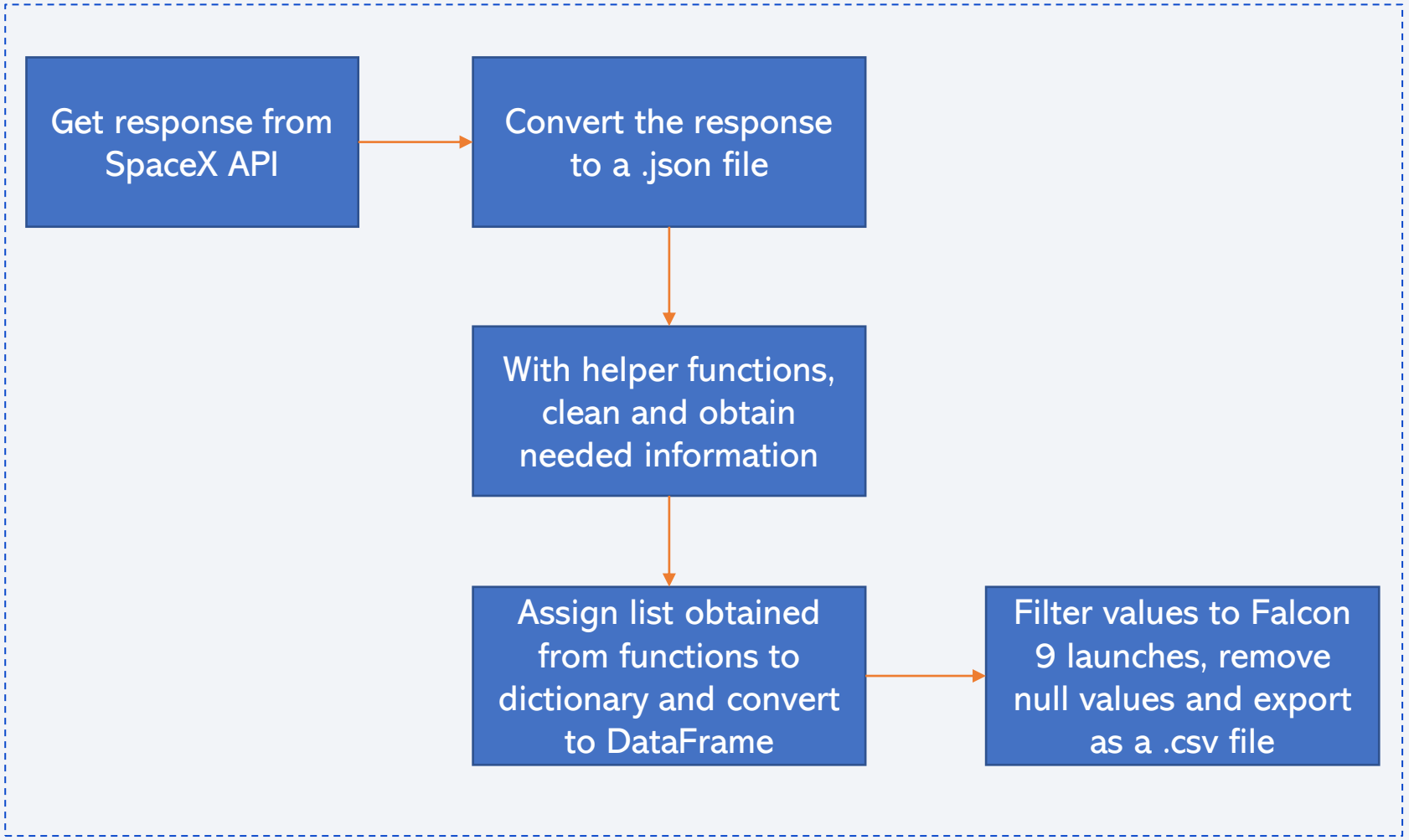
Data Collection

- SpaceX Rest API
  - ✓ Implementing get request and helper functions to extract information from API using identification numbers to obtain booster or launch site names, payload specifications, landing outcomes and more. The static response API used is as follows:

    https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json

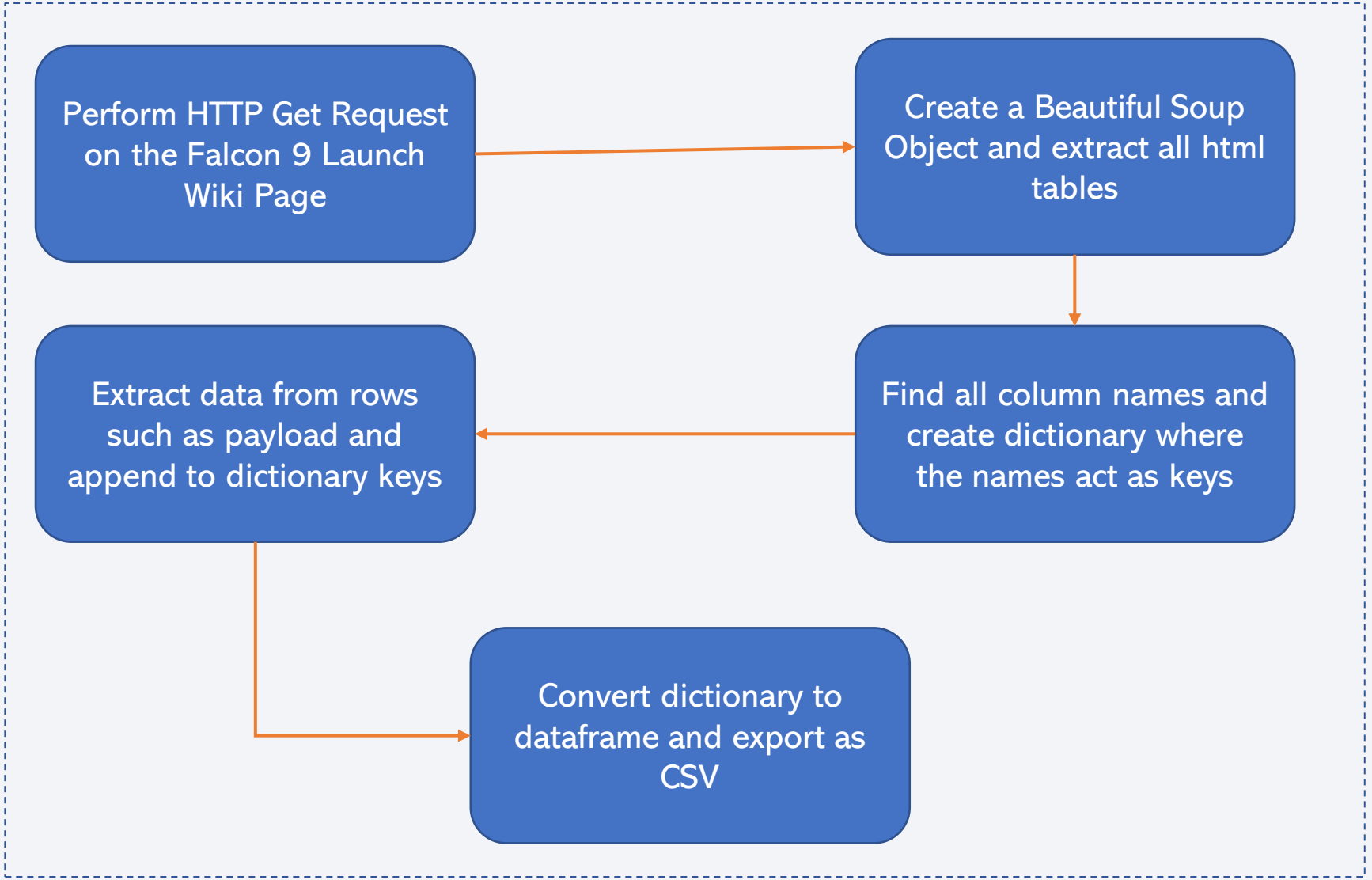
  - ✓ Convert JSON file into a dataframe and processing the data
- Web Scraping
  - ✓ Another common source is scraping the Wikipedia page with Beautiful Soup and HTML parsing

Data Collection – SpaceX API



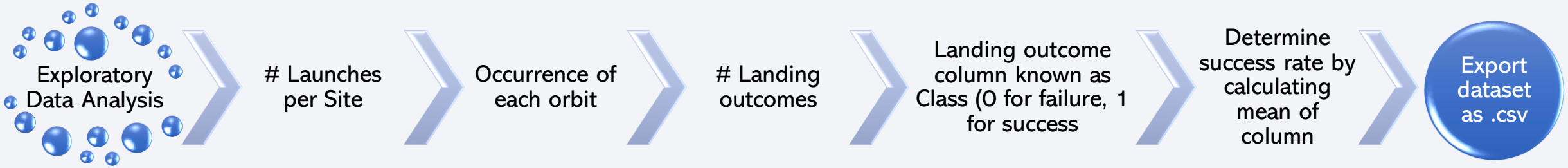
Notebook link to GitHub

Data Collection - Scraping



[Notebook link to GitHub]

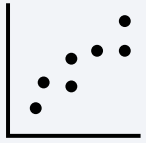
Data Wrangling



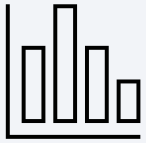
Notebook link to GitHub

EDA with Data Visualization

Types of charts plotted:

i. Scatter Graph 

These plots were implemented to explore correlations between variables such as between Launch Site and Pay Load Mass (kg) where it was shown site VAFB-SLC does not launch heavy payload rockets.

ii. Bar Graph 

A bar figure can simplify comparing various categories in data sets for example, visualizing success rate of different orbit types.

iii. Line Plot 

Line graphs can help in identifying trends and making future predictions of data variables and in our case, a line plot was drawn to show the overall increasing success rate from 2013 till 2020.

[Notebook link to GitHub]

EDA with SQL

After loading the dataset in a Db2 database, SQL queries as summarized below were executed to better comprehend the data:

- Displaying unique launch sites in the space mission
- Displaying initial 5 records in the data where the launch sites begin with string 'CCA'
- Calculating total booster payload launched by NASA (CRS)
- Showing the average payload mass carried by booster F9 v1.1
- Finding the date of the first successful ground pad landing outcome
- Listing booster names with payload between 4000 – 6000 kg and landing outcome a success in drone ships
- Displaying total number of successful and failed rocket launch outcomes
- Listing booster names that carried highest payload mass
- Displaying launch site names and booster versions that produced fail drone ship landing outcomes in 2015
- Listing count of each landing outcome (e.g., Failure (drone ship)) in descending order between dates 4/6/2010 and 20/3/2017

Notebook link to GitHub

Build an Interactive Map with Folium

To visualize and discover certain factors that can lead to selecting an optimal location for a launch, Folium is used to design interactive maps:

- Firstly, the latitude and longitude coordinates at each launch site were taken and a circle marker was added around each site with labels.
- A mouse position was added to obtain coordinates of each point on the map
- A column called marker color was added which assigns green to a record in the dataframe if the outcome is a success or otherwise red. Marker clusters were used to visualize colored outcomes on each launch site.
- Lastly, using Haversine's formula, the distances from a launch site to various landmarks or proximities were calculated to identify trends and patterns of what is around a launch site. Lines are drawn additionally to measure the distances to the landmarks.

Examples of landmark trends:

Are Launch sites near railways: Yes Are Launch sites near highways: Yes Are Launch sites away from cities: Yes

Notebook link to GitHub

Build a Dashboard with Plotly Dash

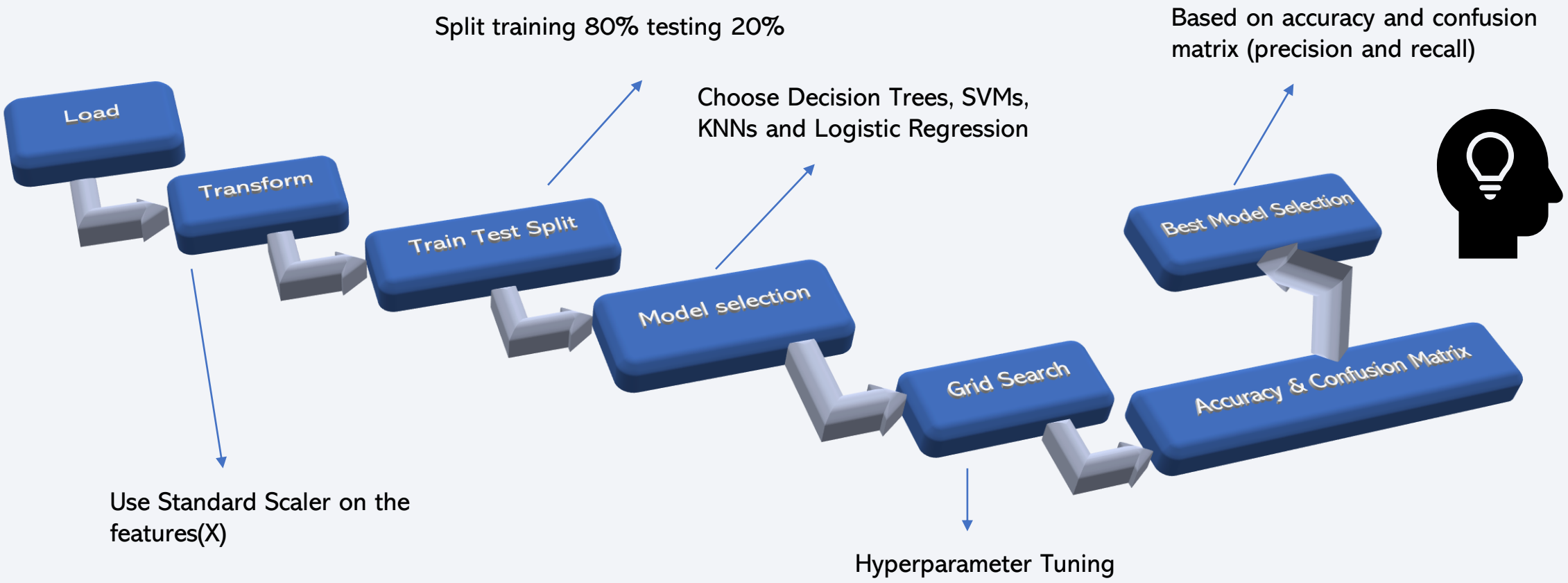
The interactive dashboard consists of a pie chart and scatter plot:

- ❑ The pie chart displays proportion of launches per launch site
- ❑ The chart can be made to display proportion of successful and failed launches of each launch site
- ❑ Additionally, the chart can be altered based on payload mass variations

- ✓ The scatter plot explores the relationship between payload mass and launch success for varying booster versions
- ✓ Insights such as which booster versions and payload mass have the highest success rate can be found.

Notebook link to GitHub

Predictive Analysis (Classification)



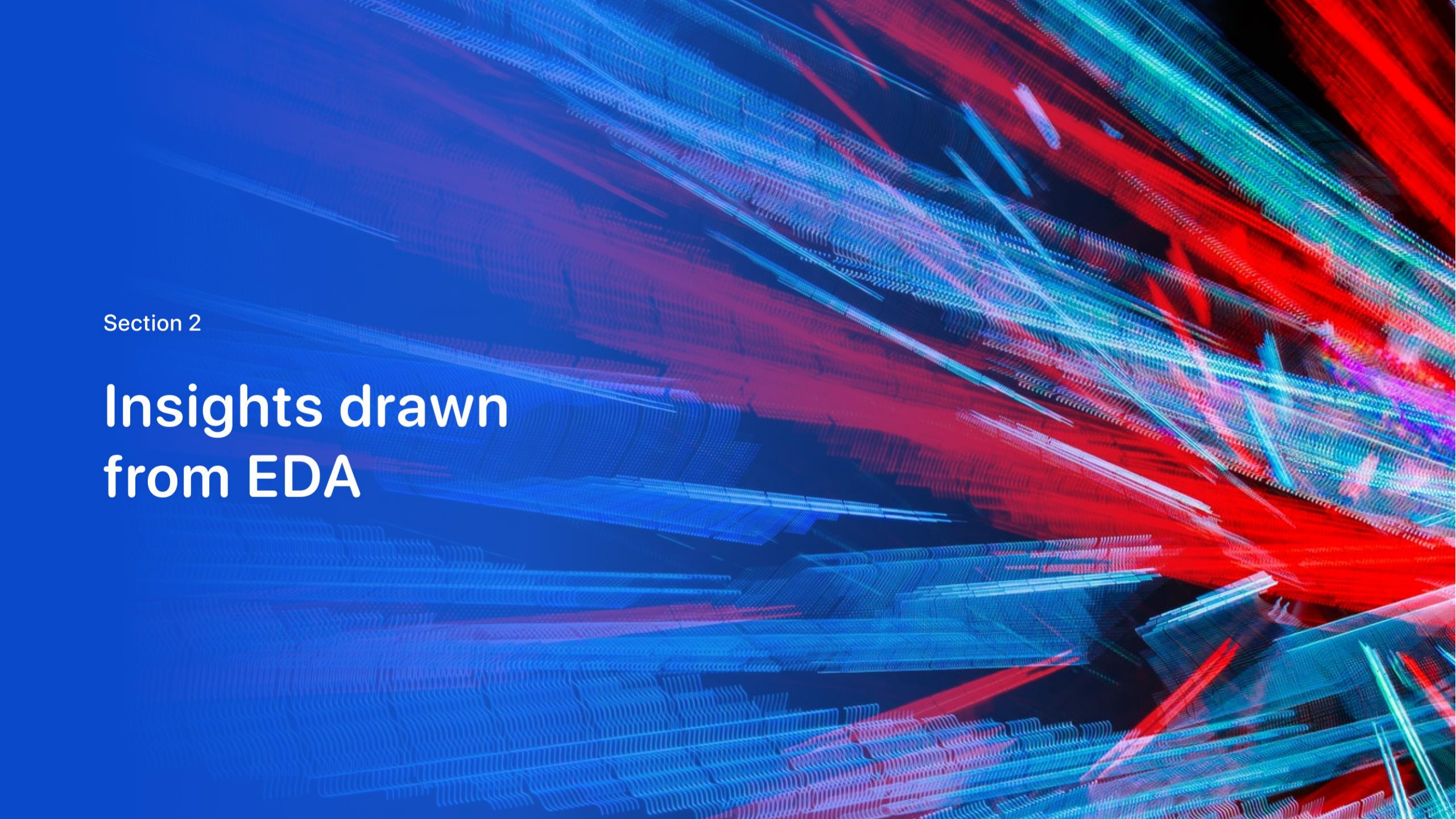
[Notebook link to GitHub]

Results

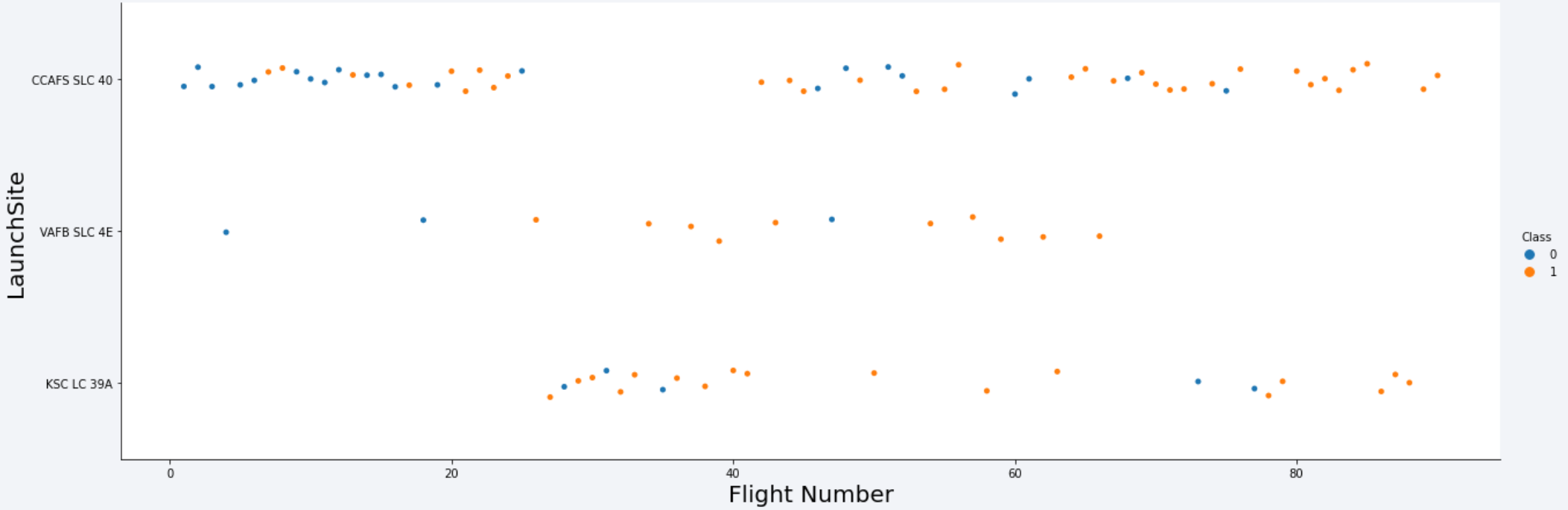
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

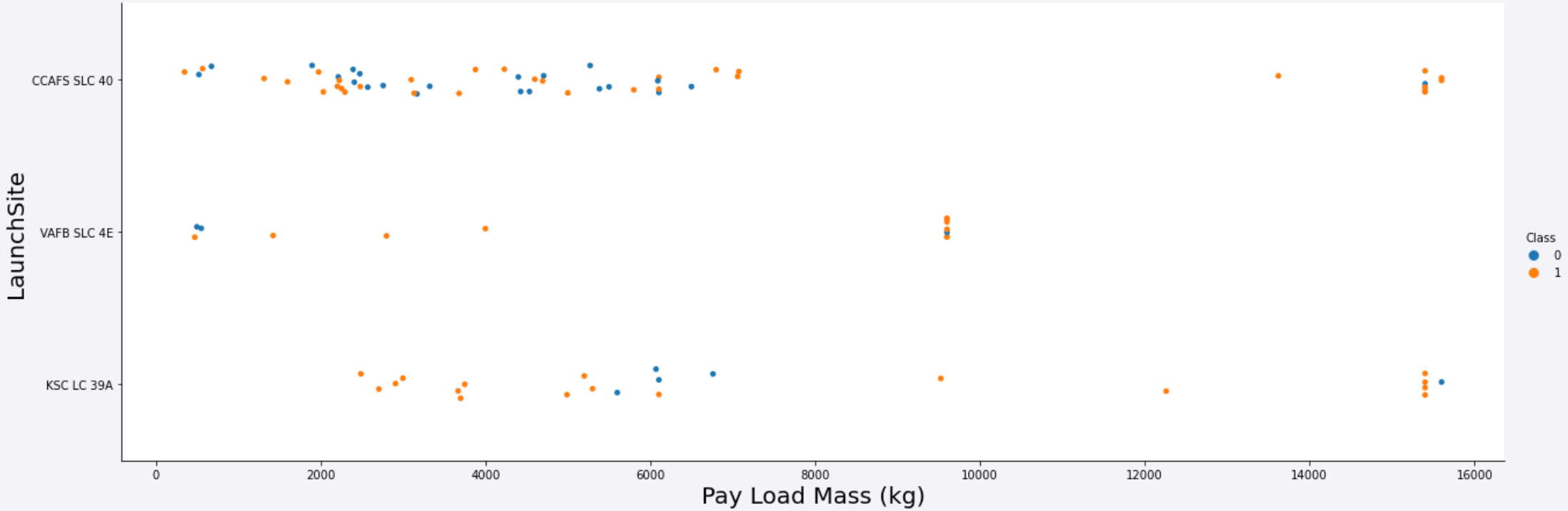


Flight Number vs. Launch Site



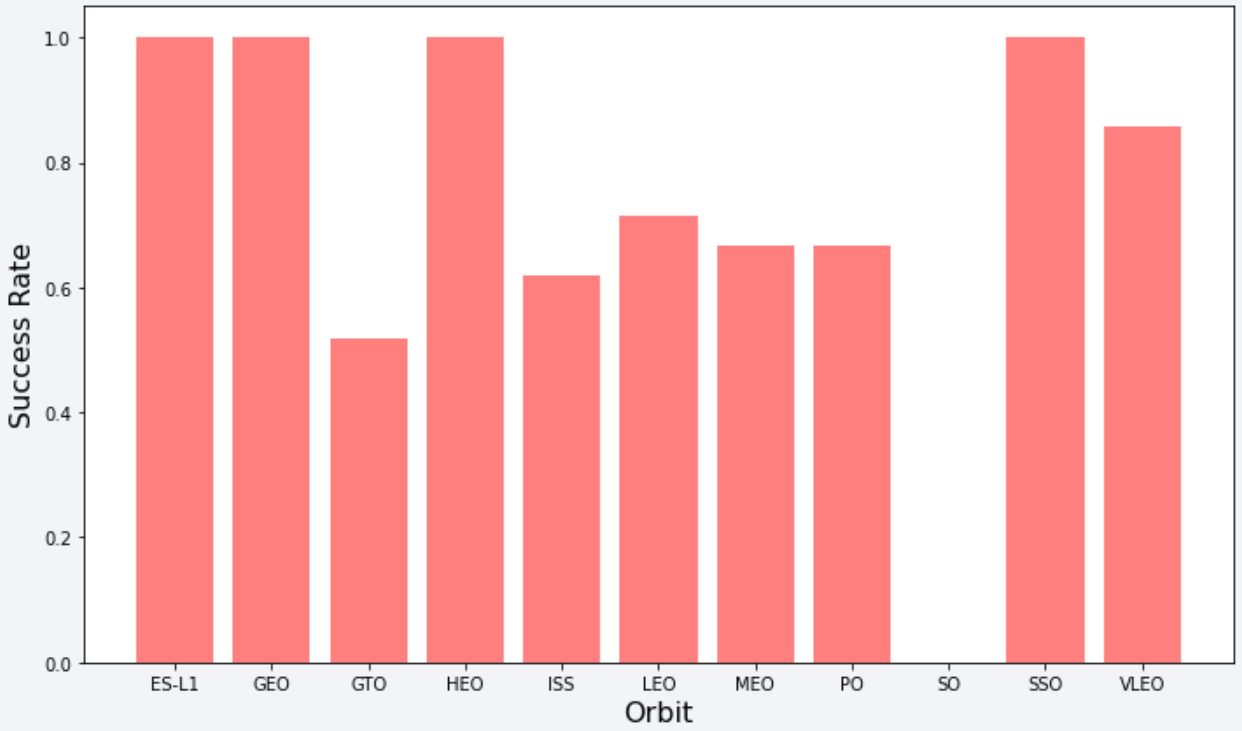
This shows that with greater number of a flights at a given launch site, the higher the success rate of a Falcon 9 rocket launch

Payload vs. Launch Site



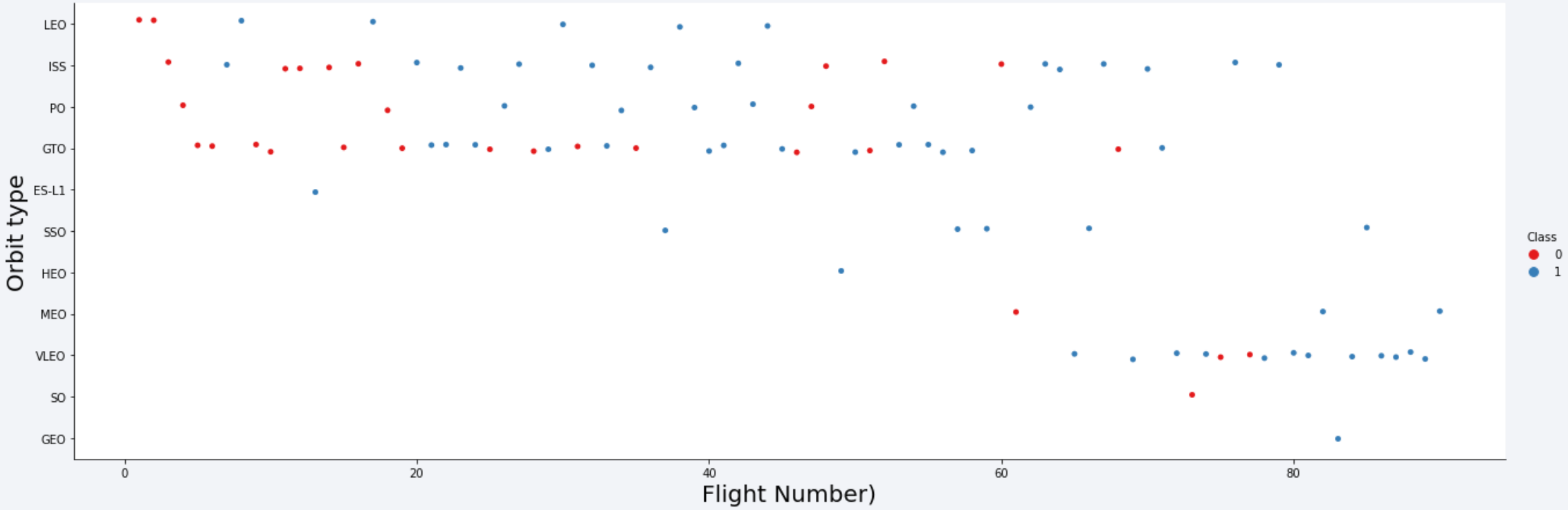
This figure shows launch sites such as CCAFS SLC 40, a higher success rate is followed by higher pay loads. However, this pattern is not clear enough as shown in the graph to be able to deduce if the success rate of a launch site is dependant on Pay Loads.

Success Rate vs. Orbit Type



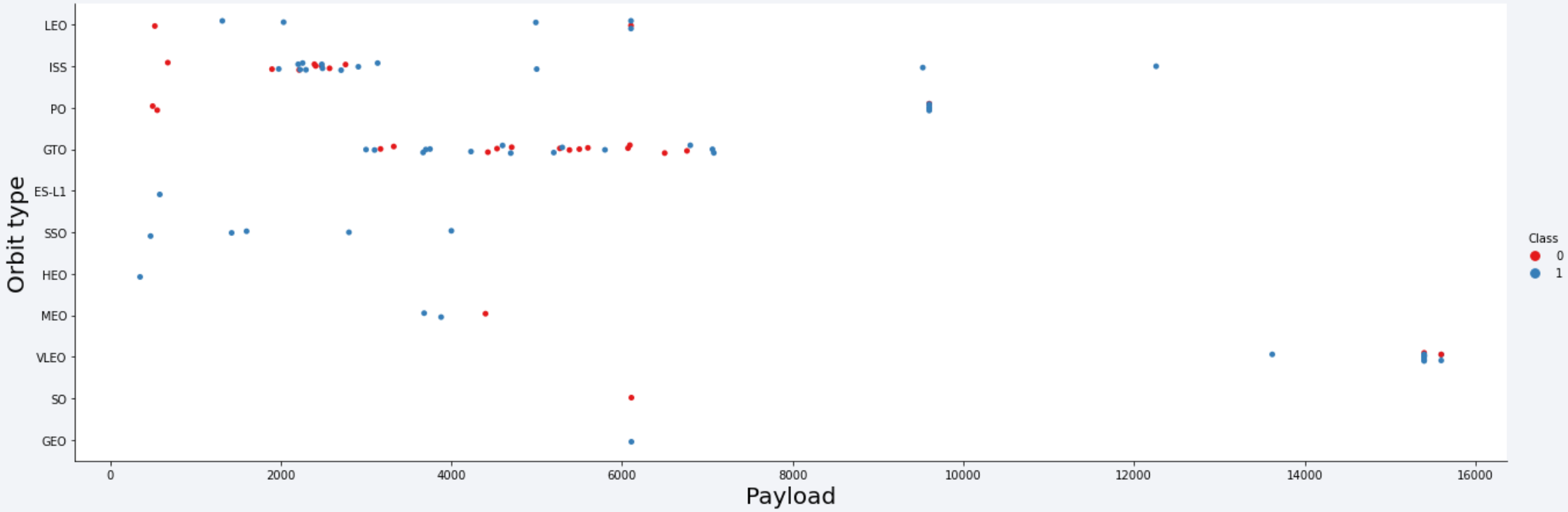
SSO, ES-L1, GEO, HEO and lastly, VLEO (although not to the extent of the previous ones mentioned) have the highest launch success rates. These all have varying orbit radii and hence, a correlation between orbit type and success rate unlikely.

Flight Number vs. Orbit Type



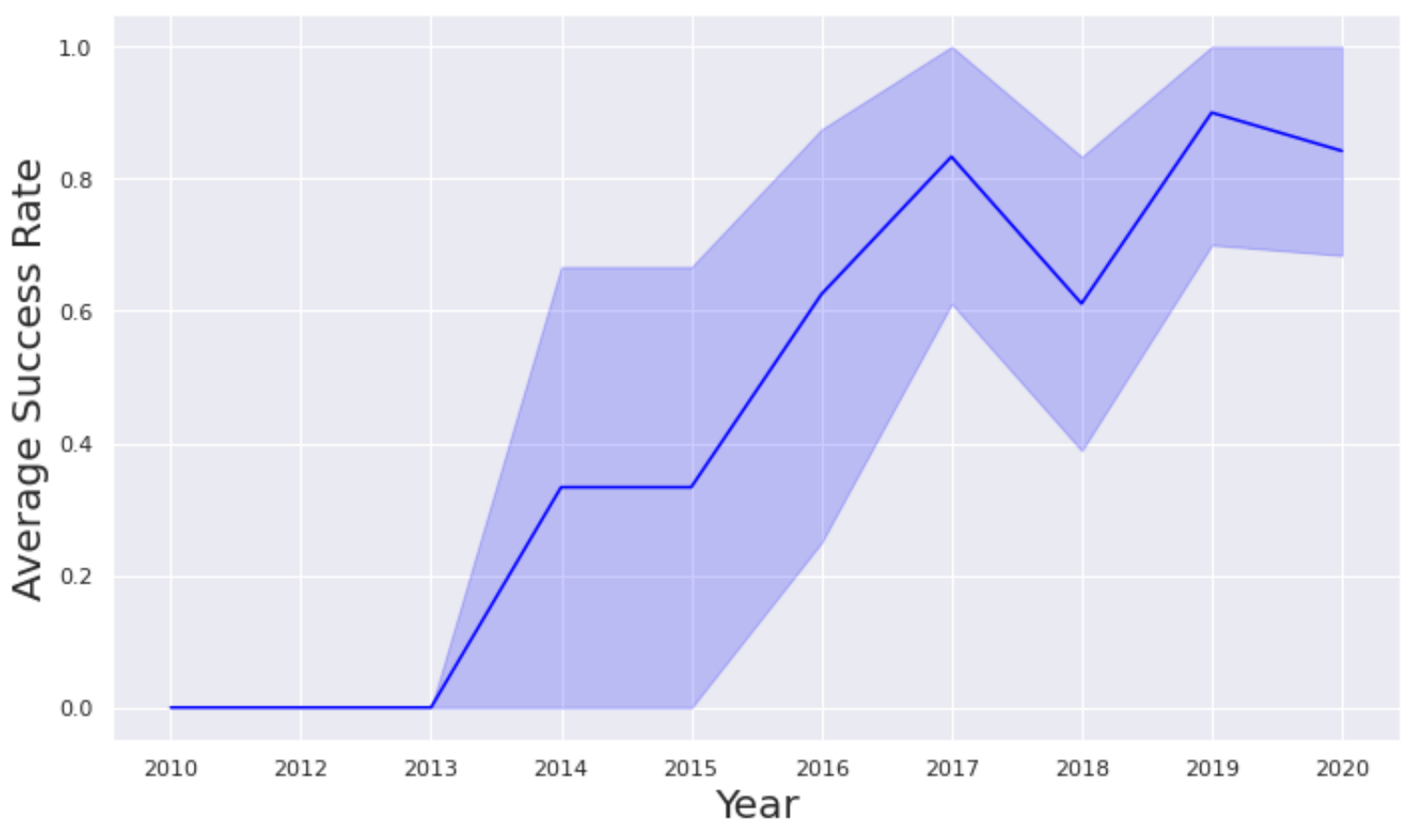
Noticing the LEO orbit, there is a clear trend of success rate increasing with the number of flights however, taking GTO on the other hand, we see no such pattern

Payload vs. Orbit Type

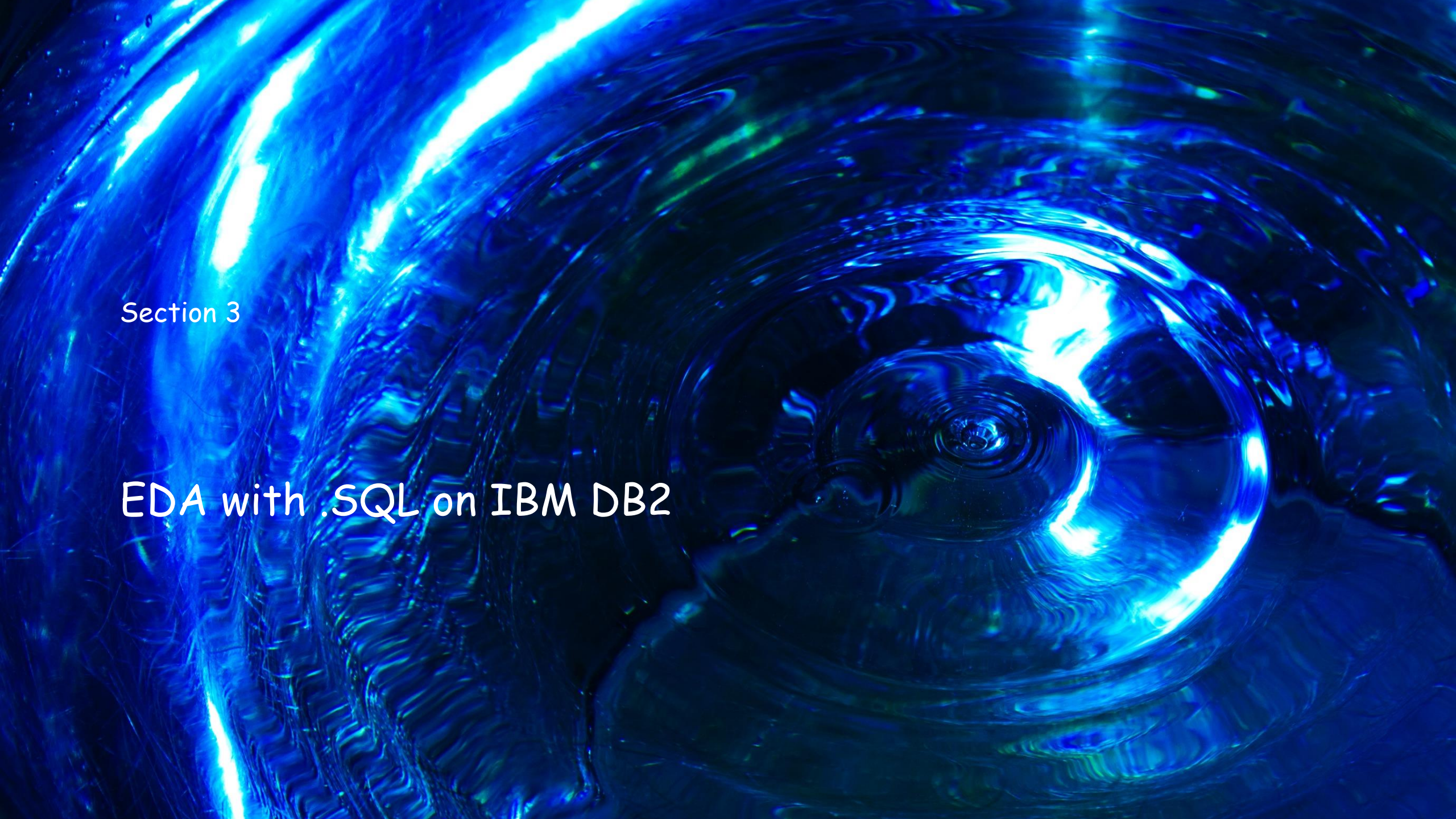


Heavy Payloads have a positive influence on ISS, LEO and Polar Orbits and a negative influence on GTO orbit types.

Launch Success Yearly Trend



The success rate has on average, kept increasing from 2013 till 2020 where we can predict that it will continue to increase.



Section 3

EDA with .SQL on IBM DB2

All Launch Site Names

SQL Query: SELECT DISTINCT (Launch_Site) from SPACEXTBL;

Unique Launch Site Names
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

The word ‘DISTINCT’ will show only unique site names in the Launch_Site column

Launch Site Names Begin with 'CCA'

SQL Query: SELECT * from SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

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

The word '%' wild card implies that the launch site name will start with 'CCA', and the 'LIMIT' clause displays only the top 5 records

Total Payload Mass

SQL Query: SELECT SUM (payload_mass_kg_) as “Total NASA Payload” FROM SPACEXTBL WHERE CUSTOMER LIKE ‘NASA%’;

Total NASA Payload
99980

- ✓ The ‘SUM’ function aggregates all the values in the Payload Mass column
- ✓ The ‘WHERE’ and ‘LIKE’ clauses filter the table to only include payloads of customers starting with ‘NASA’.

Average Payload Mass by F9 v1.1

SQL Query: SELECT AVG (payload_mass_kg_) as “Total F9 v1.1 Payload” FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE ‘F9 v1.1%’;

Total F9 v1.1 Payload
2534

- ✓ The ‘AVG’ function calculates the mean of all the values in the Payload Mass column
- ✓ The ‘WHERE’ and ‘LIKE’ clauses filter the table to only include payloads of booster versions starting with ‘F9 v1.1’ (F9 version 1.1 boosters).

First Successful Ground Landing Date

SQL Query: SELECT MIN (DATE) as “First Successful Landing” FROM SPACEXTBL WHERE LANDING_OUTCOME LIKE ‘Success%’;

First Successful Landing
2015-12-22

- ✓ The ‘MIN’ function calculates the minimum date in DATE column
- ✓ The ‘WHERE’ and ‘LIKE’ clauses filter the table to only include successful landing outcomes.

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query: SELECT Booster_Version FROM SPACEXTBL WHERE LANDING_OUTCOME LIKE 'Success (drone ship)%' AND 4000 < PAYLOAD_MASS_KG_ < 6000;

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

- ✓ The 'SELECT' function gets only the booster versions
- ✓ The 'WHERE' and 'LIKE' clauses filter the table to only include successful landing outcomes on drone ships.
- ✓ The 'AND' clause is used to specify an additional condition of payload being between 4000 kg and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

SQL Query: SELECT Mission_Outcome, COUNT(Mission_Outcome) AS "# Success and Failure Outcomes" FROM SPACEXTBL GROUP BY Mission_Outcome;

mission_outcome	# success and failure outcomes
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- ✓ The 'SELECT' and 'Count' functions gets only the type and count of mission outcomes
- ✓ The 'GROUP BY' clause categorizes the outcomes into distinct mission outcome.
- ✓ From the result, total number of success outcomes is 100 and failed is only 1.

Boosters Carried Maximum Payload

SQL Query: SELECT booster_version FROM SPACEXTBL WHERE PAYLOAD_MASS_KG_=

(SELECT MAX(PAYLOAD_MASS_KG_) FROM SPACEXTBL);

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

- ✓ The ‘WHERE’ statement finds booster versions of highest payloads.
- ✓ A subquery is used after the ‘WHERE’ clause to find the maximum payload value.

2015 Launch Records

SQL Query: SELECT Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE Landing_Outcome

LIKE 'Fail%' AND YEAR(DATE) = 2015;

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

- ✓ The 'SELECT' function outputs only the landing outcomes, booster versions and launch sites.
- ✓ The 'WHERE' clause finds the failed landing outcomes that occurred in 2015.
- ✓ The 'YEAR' function selects only the year from the DATE column.

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

SQL Query: SELECT Landing_Outcome, Count(*) AS Count_Outcomes FROM SPACEXTBL

WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome

ORDER BY Count_Outcomes DESC;

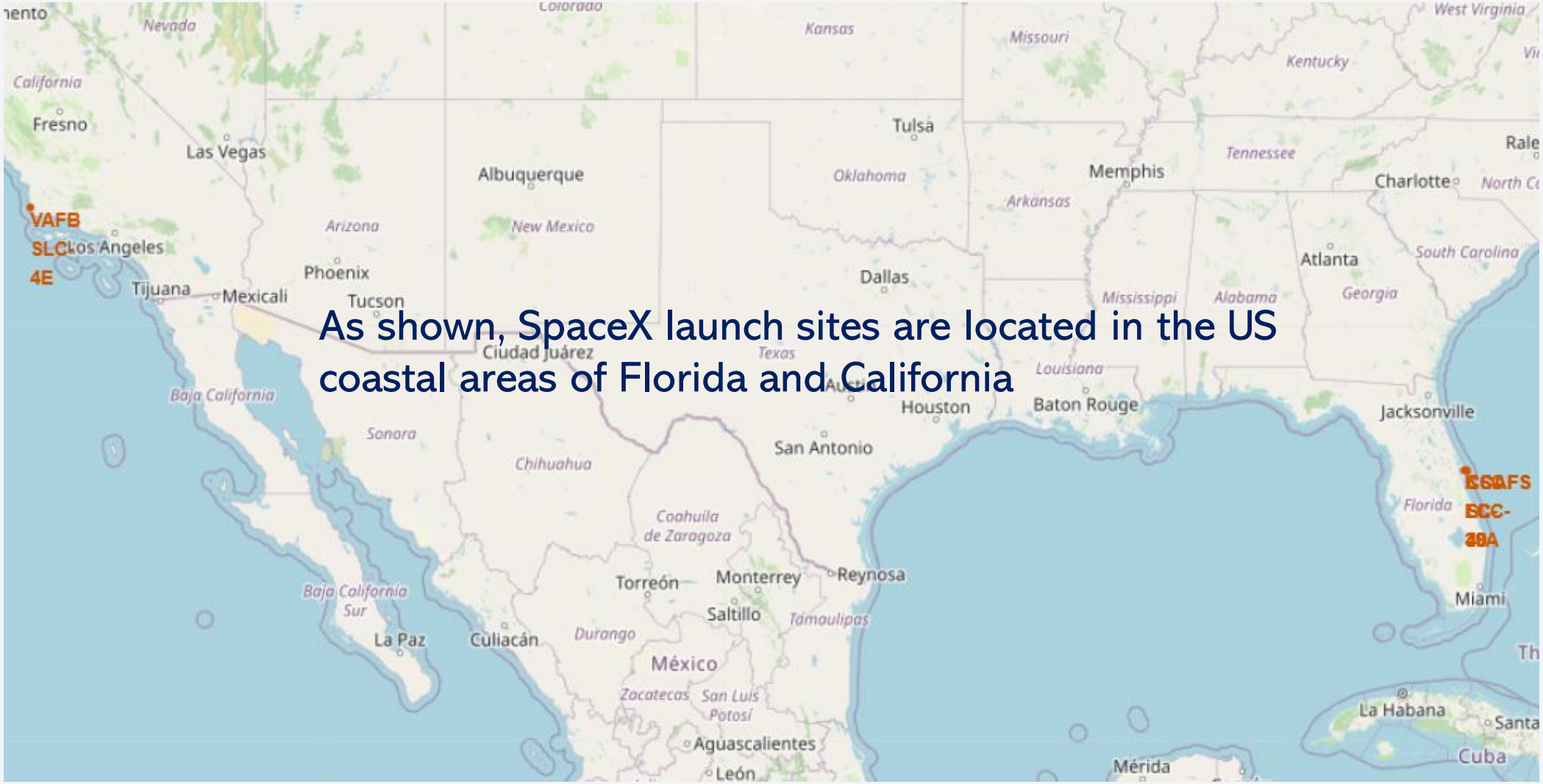
landing_outcome	Count_Outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- ✓ The 'SELECT' function outputs only the landing outcomes and the occurrence of each outcome
- ✓ The 'WHERE' clause finds dates between 4th June 2010 and 20th March 2017
- ✓ The 'GROUP BY', 'ORDER BY' and 'DESC' statements ensures the landing outcomes are used as index and the table is displayed in descending order according to the count of each landing outcome

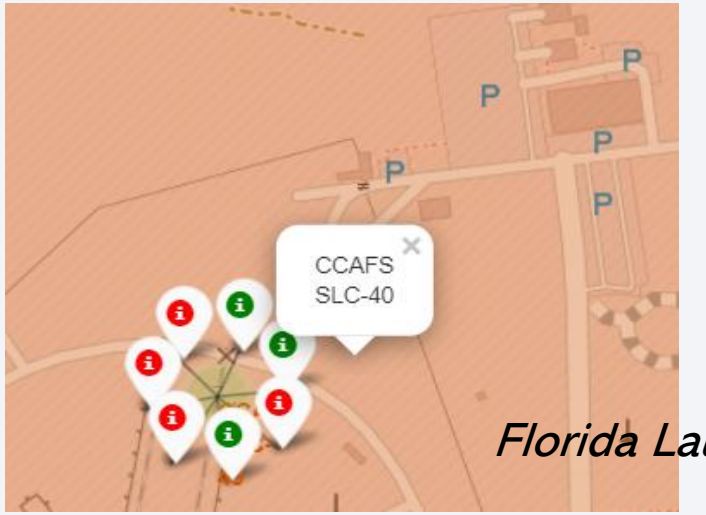
Section 4

Launch Sites Proximities Analysis

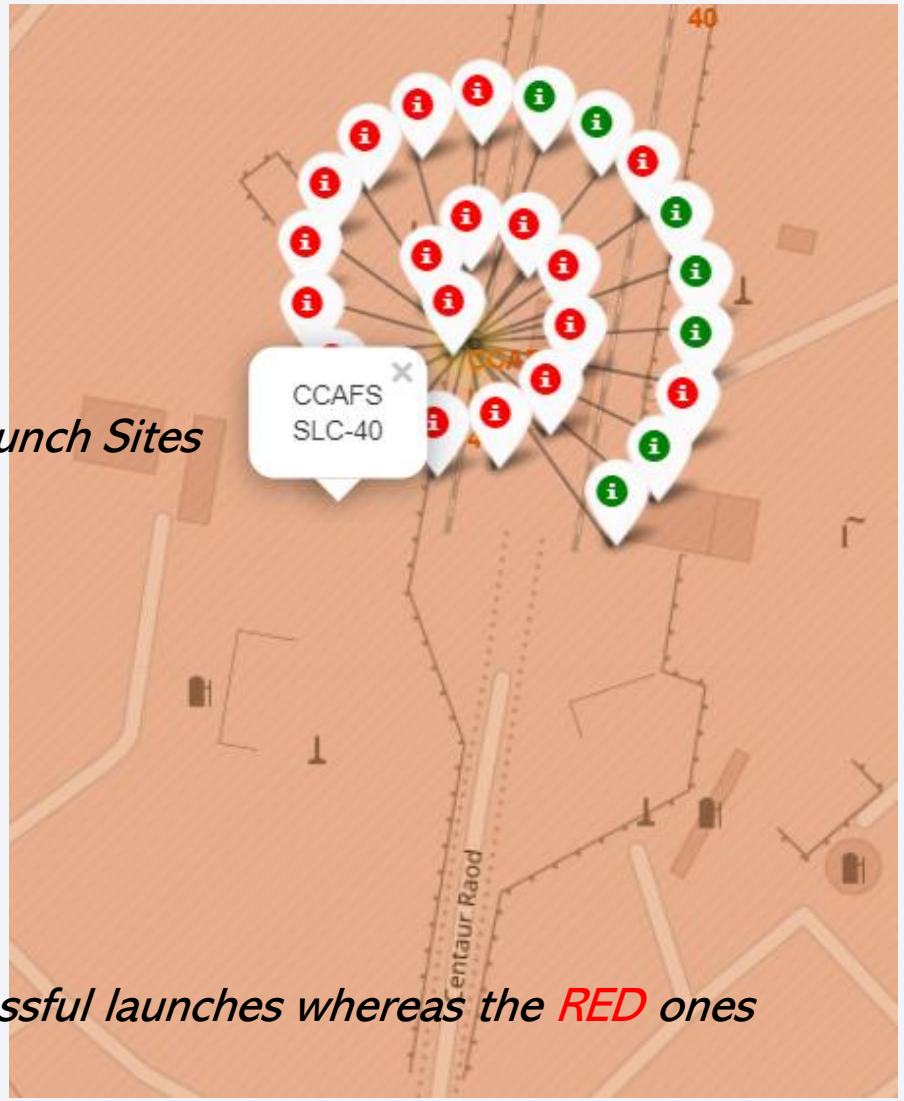
Global Map Markers of all Launch Sites



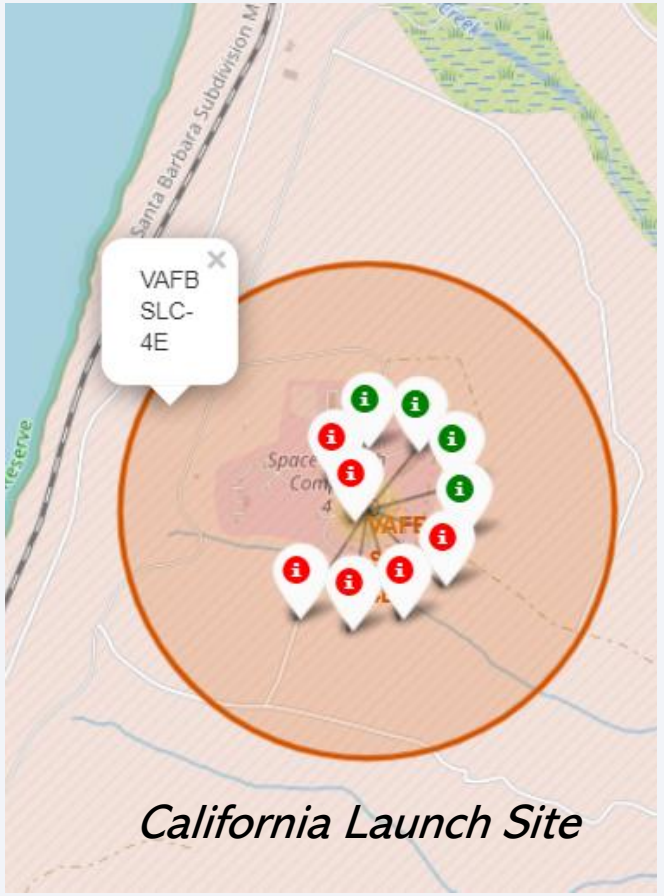
Color Labelled Launch Outcomes



Florida Launch Sites

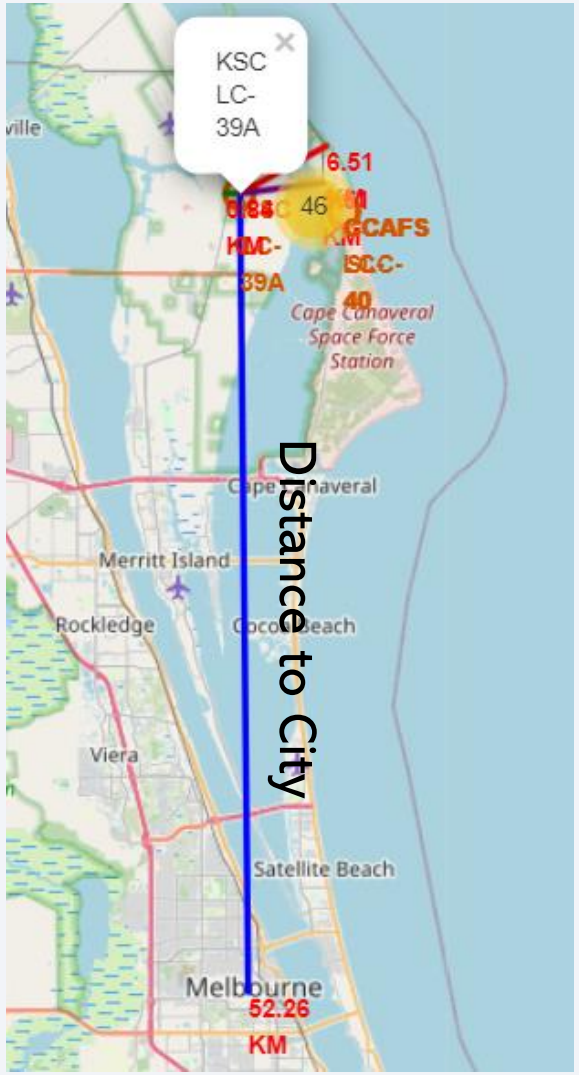


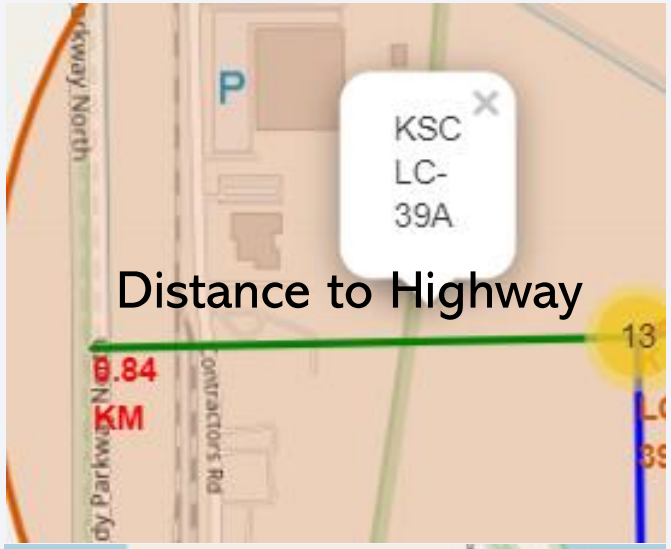
The GREEN markers signify successful launches whereas the RED ones show failures

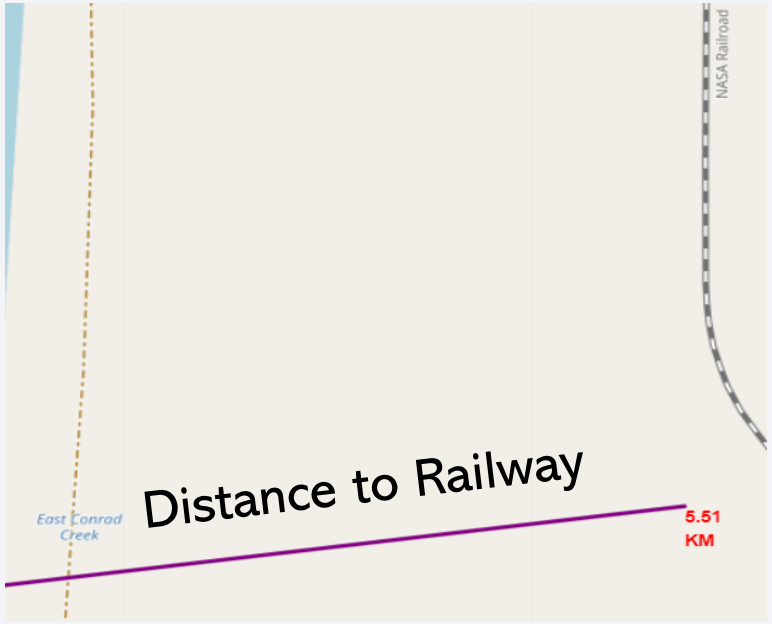


California Launch Site

Launch Site Distances to its Proximities



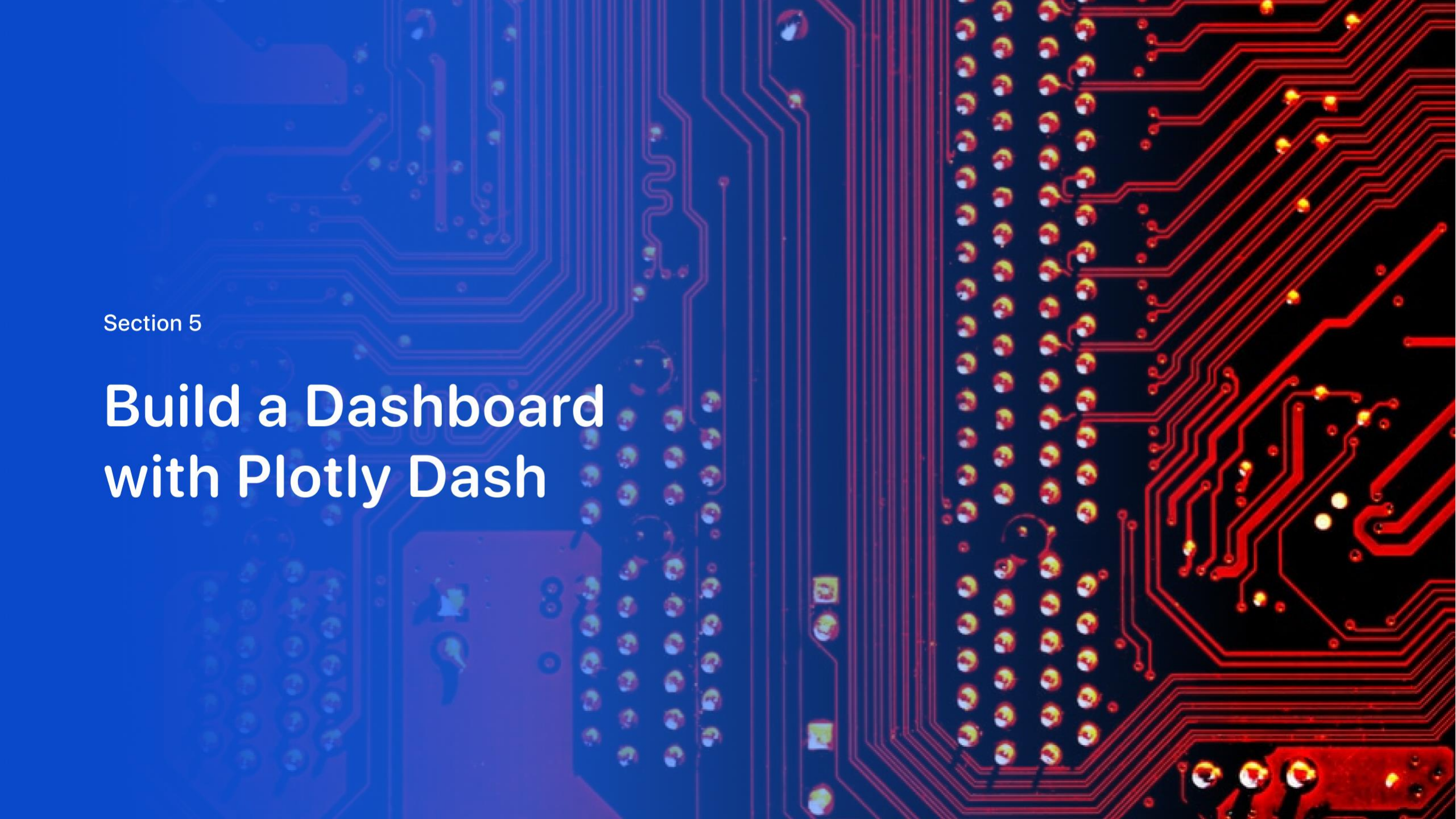




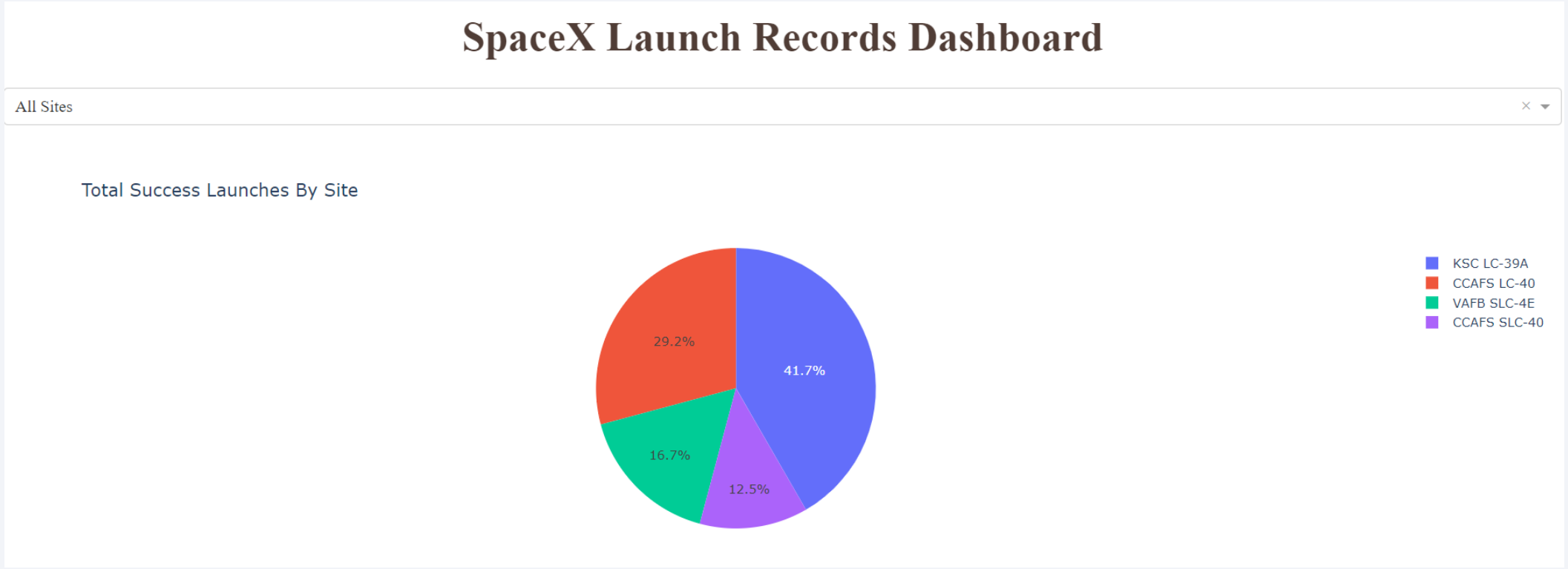
Are launch sites in close proximity to railways? Yes
Are launch sites in close proximity to highways? Yes
Are launch sites in close proximity to coastline? Yes
Do launch sites keep certain distance away from cities? Yes

Section 5

Build a Dashboard with Plotly Dash

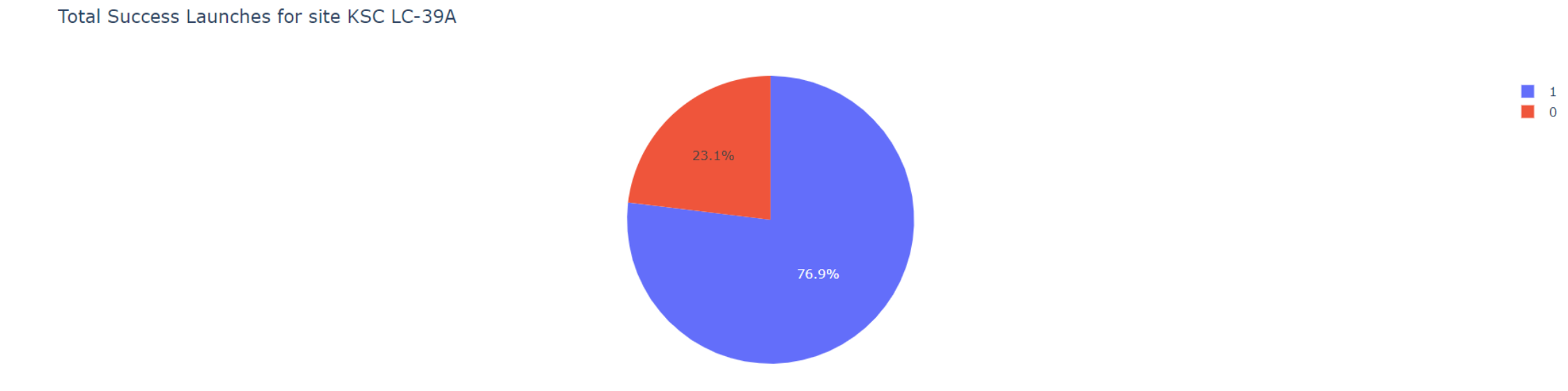


Success Proportions of each Launch Site



We see that the KSC LC-39A (Florida) has the highest number of successful launches compared to other sites

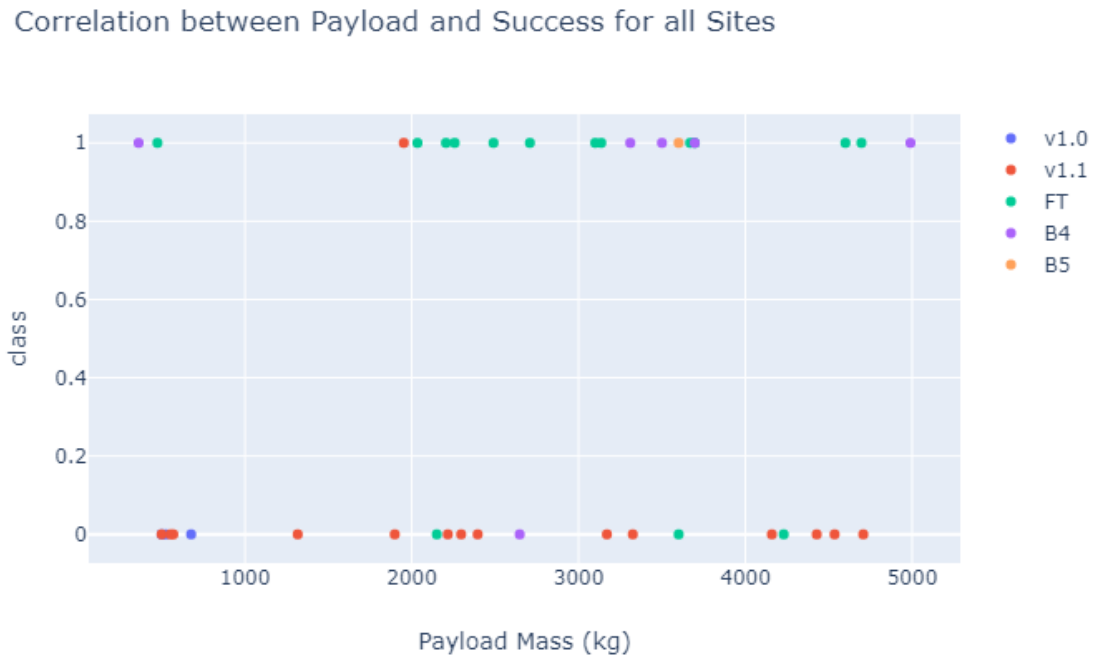
Launch Site with Highest Success Ratio



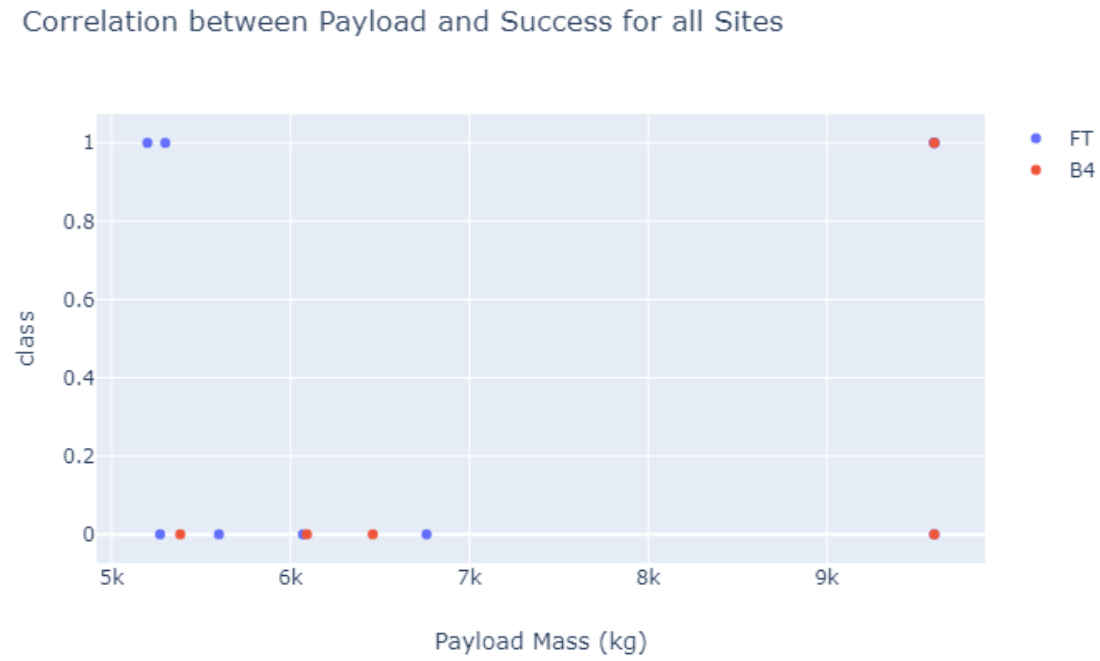
We see that the KSC LC-39A (Florida) achieved a success rate of ~77% with only a failure rate of ~23% and therefore, should be selected for successful launches in the future

Payload vs. Launch Outcome for All Sites

Low Payloads between 0 – 5000 kg



High Payloads between 5000 – 10,000 kg



From the plots shown above, it is evident that lower payloads result in higher success rates and the FT booster version has the largest success rate amongst all other boosters.

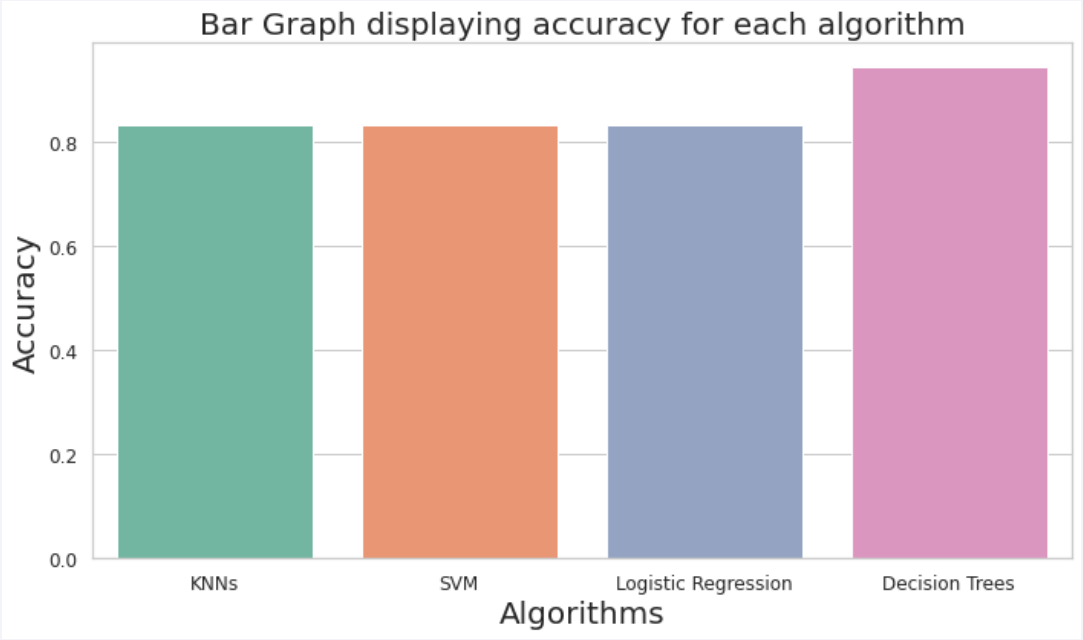
Section 6

Predictive Analysis (Classification)

Classification Accuracy

After performing Grid Search on each algorithm shown in the Bar Chart, the Decision Tree algorithm is clearly the winner when ran on the validation data set with an accuracy of 94%!

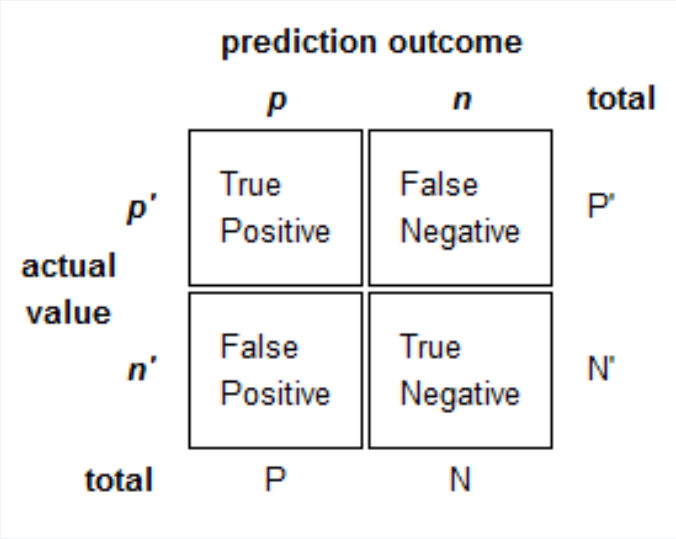
Algorithms	Accuracy
KNNs	0.833333
SVM	0.833333
Logistic Regression	0.833333
Decision Trees	0.944444



Best Hyperparameters for Decision Tree Algorithm:
tuned hyperparameters : (best parameters) {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 4, 'splitter': 'best'}

Confusion Matrix for Tree Algorithm

Examining the matrix, we see that the decision tree algorithm is 100% accurate when predicting launches that landed (No False Negatives) whilst at the same time, almost nearly perfect in the False Positives aspect.



Concluding Remarks

- Lower payloads result in more successful launches
- There is positive correlation with the number of years passed and the launch success rate
- This can also be observed in the scatter plot between launch sites and flight number
- Launch site KSC LC-39A has the highest success ratio and the most successful launches
- Orbit types GEO, HEO, SSO and ES-L1 had the best success rate
- The Decision Tree Classifier is shown the most effective Machine Learning Algorithm for this SpaceX in the category of SVM, Logistic Regression and KNNs

Thank you!