IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gilberto Sosa
January 11th 2025

Outline

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

Executive Summary

Methodologies used

- Data collection
- Data collection with Webscraping
- Data Wrangling
- Exploratory Analysis
- Data Analysis
- Exploratory Data Analysis
- Interactive Visual Analytics
- Predictive Analysis

Summary of all results

- Since 2014, success rates have increased and continue to increase while unsuccessful launch rates have decreased.
- Orbit types ES-L1, GEO, HEO, SSO and SO need more launches to determine their real success rate.
- Excluding the orbit types that need more launches, orbit VLEO has the highest and more accurate success rate out of all the orbit types.
- Site KSC LC-39A has the highest successful launch ratio out of all launch sites.
- The SVM model is the most accurate model for the task.

Introduction

Project background and context

The project's purpose was to acquire data from SpaceX and analyze it. All of this was to gain insight on the data that was obtained and reach to conclusions.

Problems you want to find answers

- Determine the cost of a rocket launch
- Determine the success rate of rocket landings
- Predict if a rocket will land successfully
- Finding an optimal location to build a launch site

Section 1

Methodology

Methodology

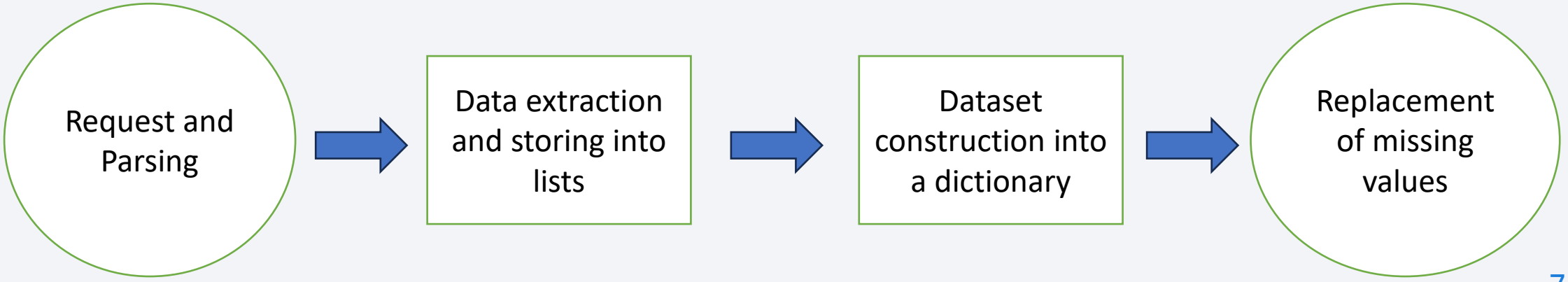
Executive Summary

- Data collection methodology:
  - A GET request was made to the SpaceX API and Webscraping was used to obtain Falcon 9 launch records from Wikipedia.
- Perform data wrangling
  - The data obtained was cleaned form missing values and certain information was extracted from the raw data for example, number of launches per site, number and occurrence of each orbit and a landing outcome label was added to the 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
  - The data was standardized, then split into training and test data after that logistic regression was used and the accuracy was calculated. Different methods were applied, and confusion matrixes were used for a better visualization of results. Finally, the best method was found.

Data Collection

- How data sets were collected.

The first set of data was obtained by using a GET request to the SpaceX API. Functions were used to get specific data like the mass of the payload and the outcome of landings. The data was extracted based on keywords like ‘Rocket’ and ‘Payloads’. The columns were combined into a dictionary and missing data was replaced.



Data Collection – SpaceX API

https://github.com/Gil-ib/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

1. Make GET requests to SpaceX API
2. Json response was parsed
3. Specific data was extracted using keywords (e.g. ‘rocket’, ‘payloads’, launchpad’)
4. Data obtained was stored in lists
5. Data was merged into a dictionary
6. Missing values were replaced

Data Collection - Scraping

https://github.com/Gil-ib/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

1. Helper functions were defined

2. HTTP GET method was used to request the Falcon9 Launch HTML page

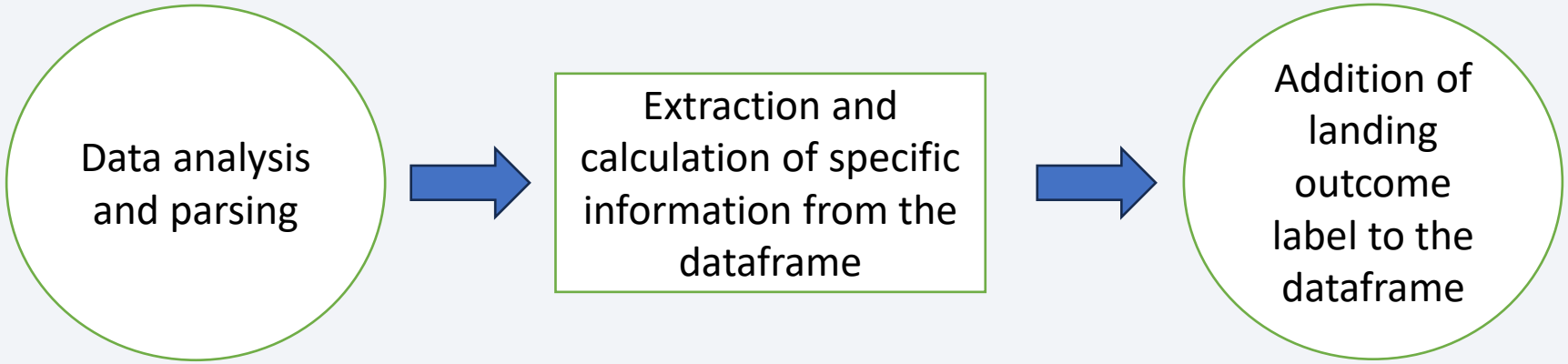
3. Column names and variables were extracted from HTML table header

4. Data frame was created by parsing HTML tables

Data Wrangling

How data was processed

The data was first analyzed and was parsed. The percentage of missing values in each column was calculated and the data type (numeral, categorical) of each column was identified. Afterwards, calculations were made to obtain specific information (e.g. number of launches on each site, number and occurrence of each orbit, number and occurence of mission outcome of the orbits). Finally, a landing outcome label was added to the dataframe.



https://github.com/Gil-ib/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- What charts were plotted and why you used those charts

First, categorical charts were plotted to compare different variables and to observe if there were any correlations between them.

A bar chart was plotted to see if there was any relation between success rate and orbit type.

A line chart was plotted to observe the yearly trend of launch success.

https://github.com/Gil-ib/Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

SQL Queries used

- %sql DROP TABLE IF EXISTS SPACEXTABLE;
- %sql create table SPACEXTABLE as select * from SPACEXTBL where Date is not null
- %sql select distinct launch_site from SPACEX;
- %sql select * from SPACEX where launch_site like 'CCA%' limit 5;
- %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEX where customer = 'NASA (CRS)';
- %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEX where booster_version like '%F9 v1.1%';
- %sql select min(date) as first_successful_landing from SPACEX where Landing_Outcome = 'Success (ground pad)';
- %sql select booster_version from SPACEX where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
- %sql select mission_outcome, count(*) as total_number from SPACEX group by mission_outcome;
- %sql select booster_version from SPACEX where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEX);
- %%sql select landing_outcome, count(*) as count_outcomes from SPACEX where date between '2010-06-04' and '2017-03-20' group by landing_outcome order by count_outcomes desc;

https://github.com/Gil-ib/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

1. Markers were added to mark launch sites from the dataframe on the map.
2. Circle markers were added to represent if a launch was successful (green) or failed (red)
3. Lines were added to visualize the proximity of launch sites
4. Lines were also drawn to see the proximity of launch sites to railways, highways, cities and more.

https://github.com/Gil-ib/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard

1. Interaction added to choose a launch site.
2. Pie chart added to visualize total successful launches depending on the site picked.
3. Added a range slider to select the payload mass, this changes the rendering in the scatter plot.
4. Added callback function to render a scatterplot that compares the class to the payload mass.

https://github.com/Gil-ib/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model

1. The data was first split into training and testing data.
2. A logistic regression object, a support vector machine object, a tree classifier object and a K nearest neighbor object were created.
3. For each object, the method ‘score’ was used to calculate the accuracy.
4. Confusion matrixes were plotted for each object.
5. Tables were created to compare the accuracy obtained from each object and determine which method was that performed the best.

https://github.com/Gil-ib/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results

Unique Launch Sites

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

Launch sites that begin with the string 'CCA'

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	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

Results

Total Payload Mass carried by Boosters launched by NASA (CRS)

total_payload_mass
45596

Date when the first succesful landing outcome in ground pad was achieved

first_successful_landing
2015-12-22

Total number of successful and failure mission outcomes

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Average payload mass carried by booster version F9 v1.1

average_payload_mass
2534.6666666666665

Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

[85]:

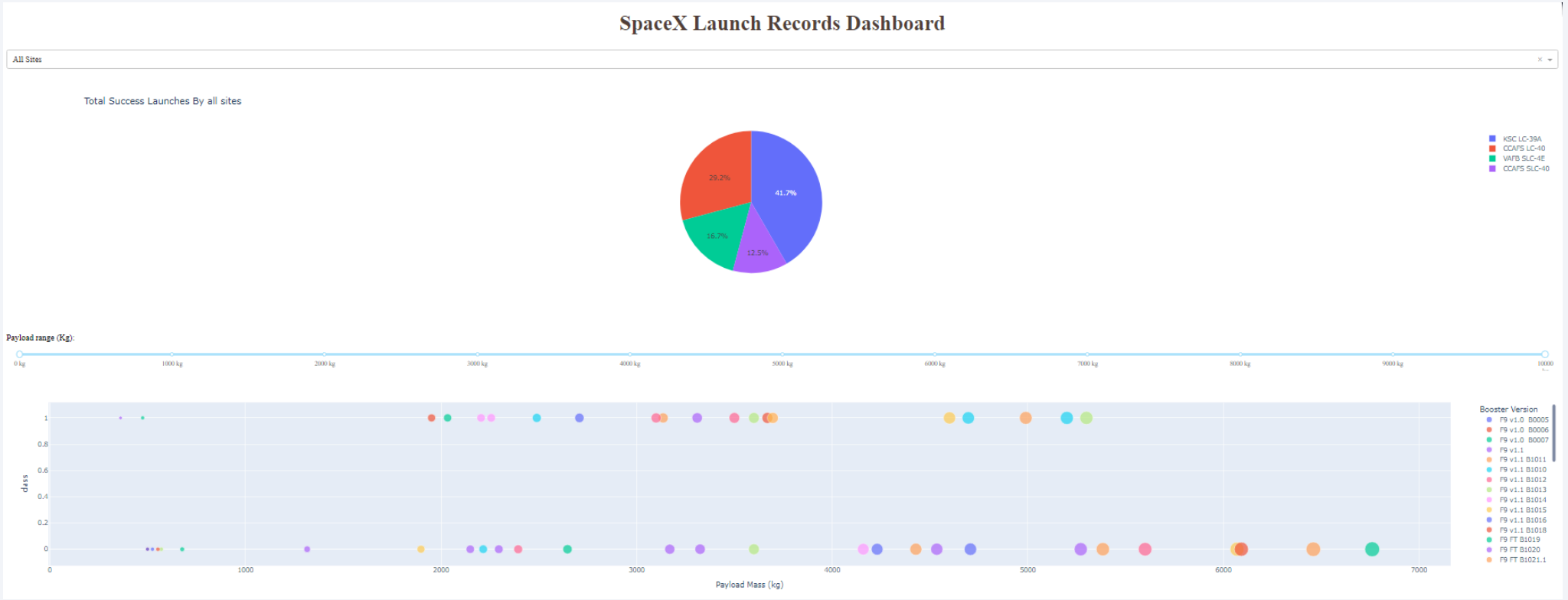
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Names of the booster_versions which have carried the maximum payload mass

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

Results

- Interactive analytics demo in screenshots



Results

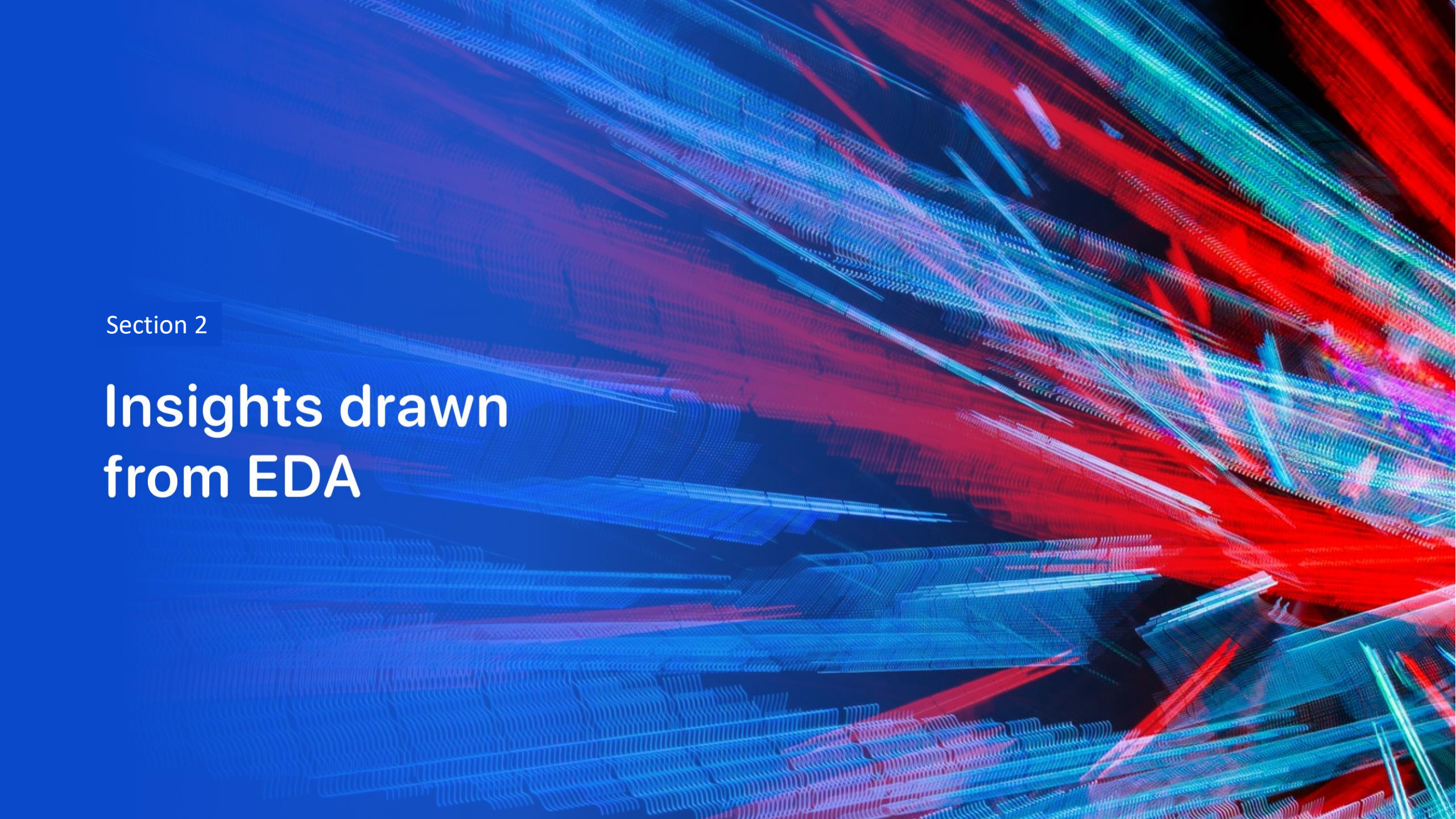
- Predictive analysis results

Results from test Sets

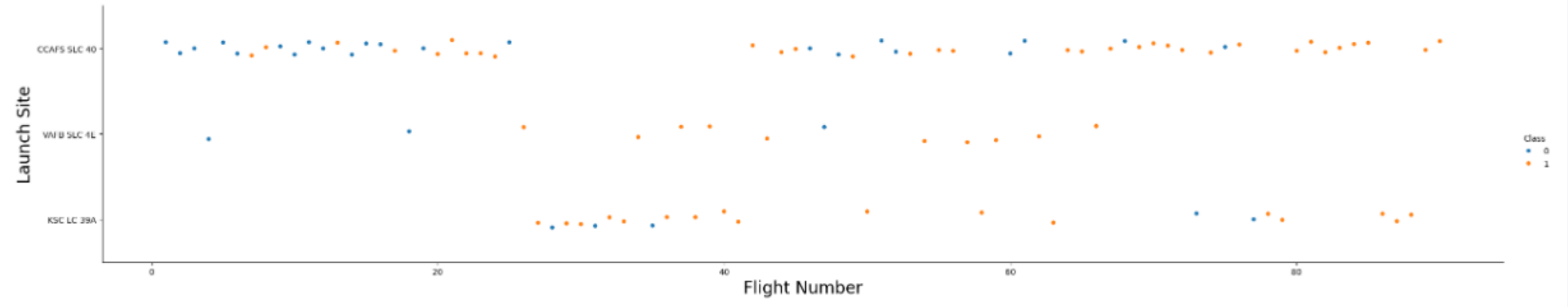
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Results from whole Dataset

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.867647	0.819444
F1_Score	0.909091	0.916031	0.929134	0.900763
Accuracy	0.866667	0.877778	0.900000	0.855556


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



Class 0: First Stage landed unsuccessfuly

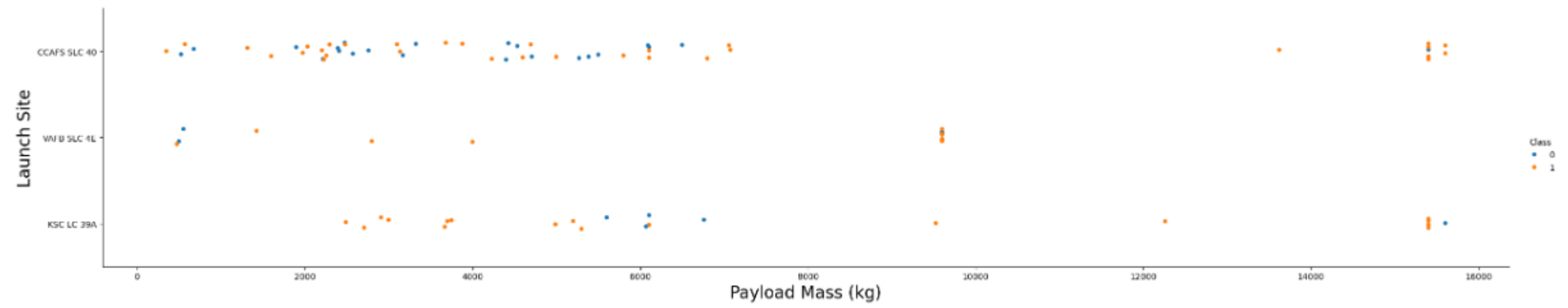
Class 1:First Stage landed successfully

Insight

The graph shows us the different results of various launches depending on the launch sight. On site CCAFS SLC 40 the most launches were made, and most were successful. On site VAFB SLC 4E the least number of launches were made but the majority were also successful. On site KSC LC 39A the second most launches were made, and the majority were successful.

This graph shows us that on any launch site the majority of the launches made were successful.

Payload vs. Launch Site



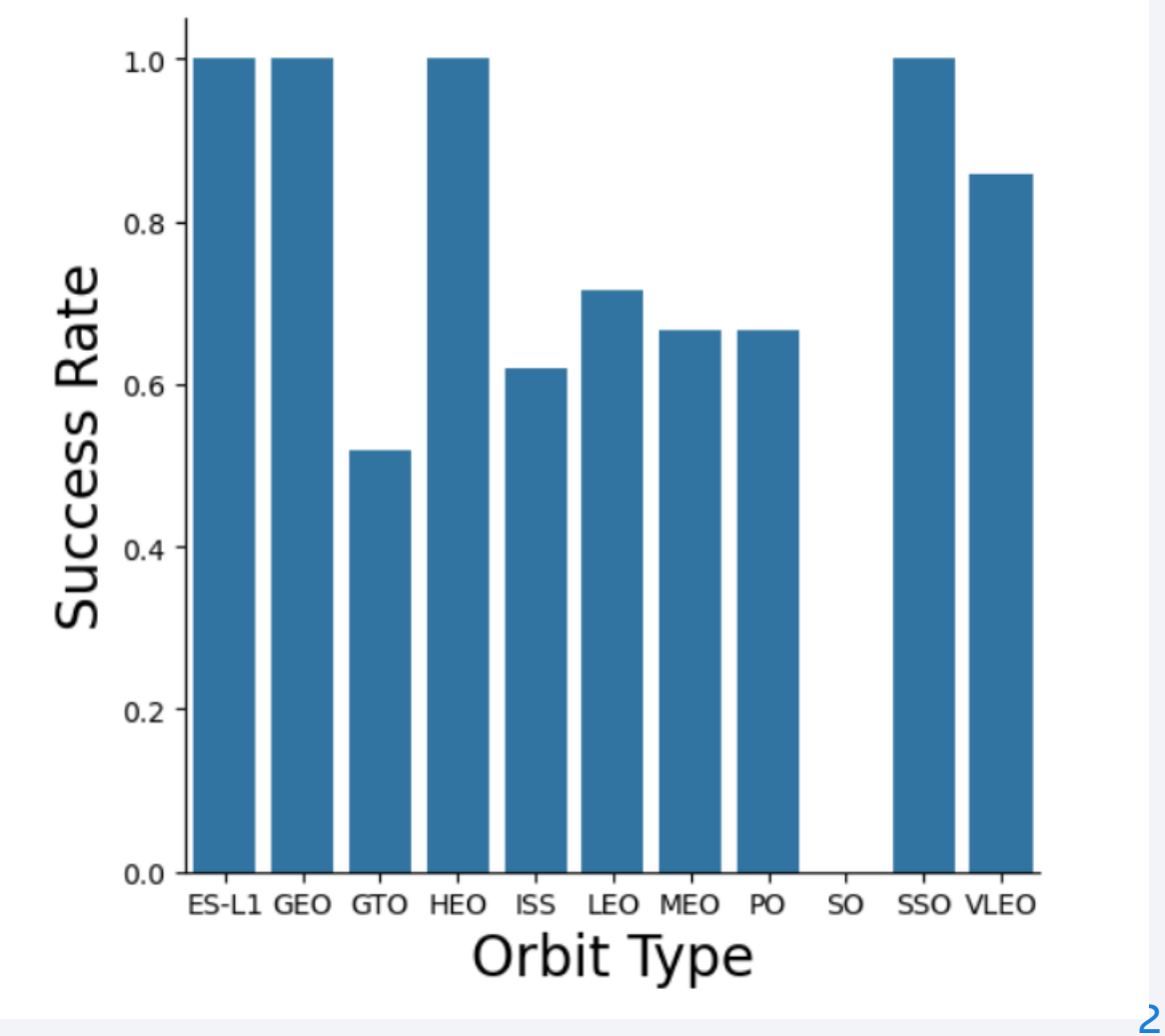
Insight

Most of the data points are found between 0 and 8,000 kg of payload. Of any site, most of the launches within this range were successful. The launches that are found in the higher ranges of payload weight are also successful. This graph doesn't indicate as much, only that more launches were done on sight CCAFS SLC 40 and KSC LC 39A and that the launches with a higher payload weight that have been done on these two sites almost all of them have been successful.

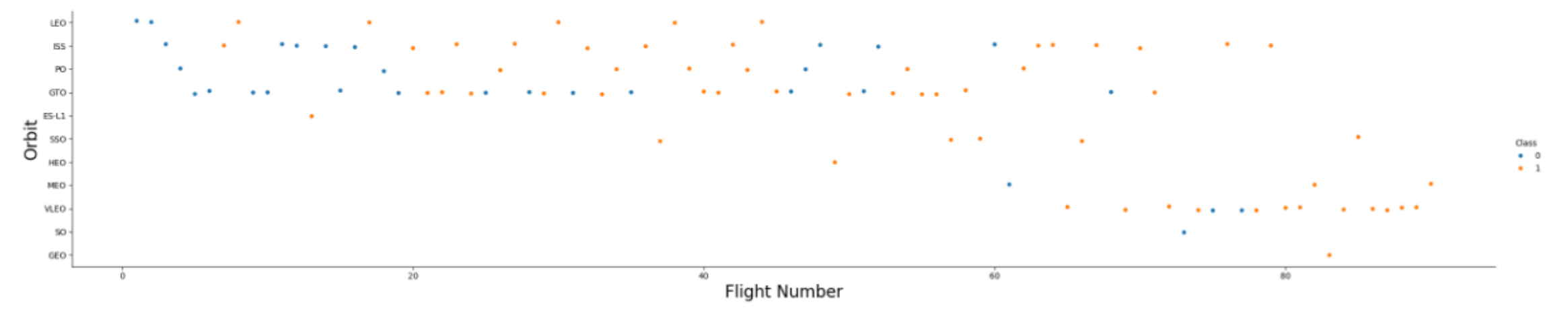
Success Rate vs. Orbit Type

Insight

The graph shown indicates that there are 4 orbit types that have the most success rate of them all, ES-L1, GEO, HEO and SSO. The orbit type VLEO also has a high success rate, almost as high as the previously mentioned. The rest of the orbit types have an intermediate success rate, except for one. The SO orbit type has a success rate of 0%.



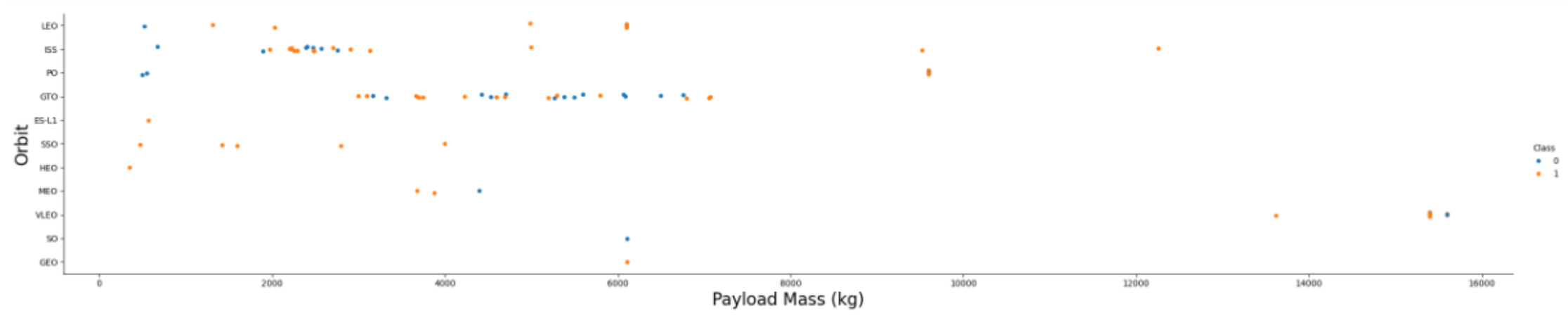
Flight Number vs. Orbit Type



Insight

This scatter plot clarifies some things about the previous bar chart. flight, the orbit types that show a 100% success rate in the bar chart have had a very small number of flights done in that orbit type. For example, orbit types GEO and HEO only have 1 flight. The reason orbit SO has a 0% success rate was also because it only has 1 flight, and it was unsuccessful. These flights need more tests to see their true success rate. The other orbit types have more flights done and it can be said that the orbit type VLEO has the most success rate of them all, this is while excluding the orbit type with very little flights done.

Payload vs. Orbit Type



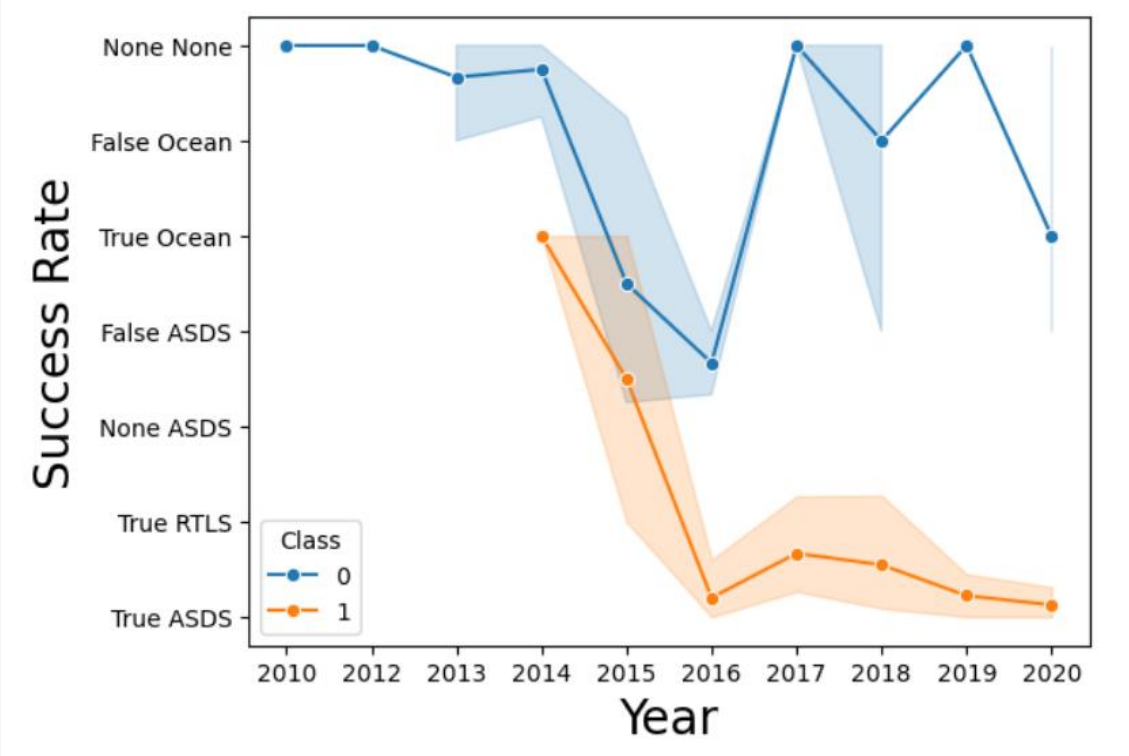
Insight

A trend cannot be identified for most orbit types except for two, ISS and GTO. For ISS it seems there seems to be about a 50% success rate with payloads between 2000 and 4000 kg. For GTO the higher the payload weight the lower the success rate becomes.

Launch Success Yearly Trend

Insight

The graph shows that since the year 2014 the success rate has been increasing and the rate of completely unsuccessful launches has been lowering.



All Launch Site Names

Query

%sql select distinct launch_site from SPACEX;

This query obtains all the distinct values of the launch_site column from the SPACEX table.

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

Launch Site Names Begin with 'CCA'

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	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

Query

%sql select * from SPACEX where launch_site like 'CCA%' limit 5;

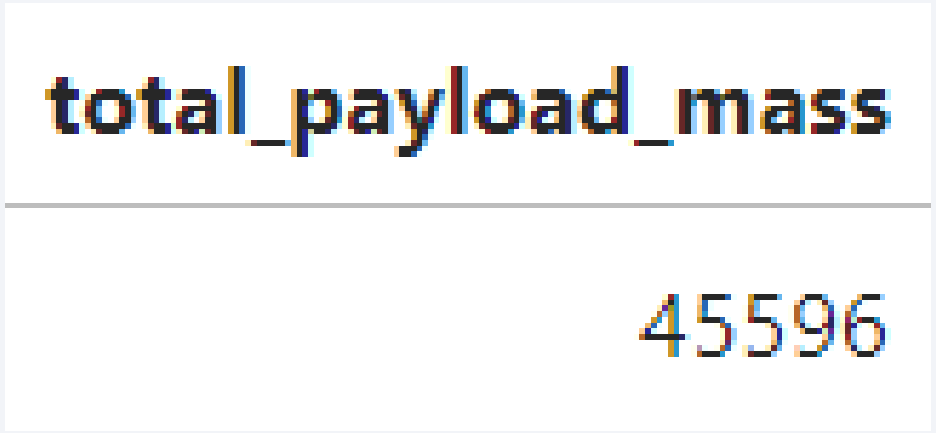
This query obtains 5 rows from the SPACEX table where the launch_site column text start with the string "CCA". What ‘CCA%’ does is that it filters values that have CCA at the beginning. The text that comes after the “CCA” string doesn’t matter, as long as it has “CCA” at the beginning it will show it.

Total Payload Mass

Query

%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEX where customer = 'NASA (CRS)';

This query calculates the total payload mass for launches done by "NASA (CRS)".
‘sum(payload_mass__kg_)’ adds all the payload mass values and the result is inserted into the variable ‘total_payload_mass’.



total_payload_mass
45596

Average Payload Mass by F9 v1.1

Query

%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEX where booster_version like '%F9 v1.1%';

This query calculates the average payload mass for launches that used the booster version "F9 v1.1". “avg(payload_mass__kg_)” obtains the average of the payload masses and the value is stored into the ‘average_payload_mass’ variable.

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

Query

%sql select min(date) as first_successful_landing from SPACEX where Landing_Outcome = 'Success (ground pad)';

first_successful_landing
2015-12-22

This query identifies the earliest date when a successful landing occurred on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

Query

%sql select booster_version from SPACEX where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;

This query retrieves the booster versions that successfully landed on a drone ship and carried a payload within 4000 and 6000 kg.

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

Total Number of Successful and Failure Mission Outcomes

Query

%sql select mission_outcome, count(*) as total_number from SPACEX group by mission_outcome;

This SQL query counts the number of missions for each unique mission outcome in the SPACEX table, it groups the results by mission_outcome.

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Query

%sql select booster_version from SPACEX where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEX);

This query retrieves the booster versions which have carried the maximum payload mass.

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Query

```
%%sqlSELECT CASE WHEN substr(date, 6, 2) = '01' THEN 'January' WHEN substr(date, 6, 2) = '02' THEN 'February' WHEN substr(date, 6, 2) = '03' THEN 'March' WHEN substr(date, 6, 2) = '04' THEN 'April' WHEN substr(date, 6, 2) = '05' THEN 'May' WHEN substr(date, 6, 2) = '06' THEN 'June' WHEN substr(date, 6, 2) = '07' THEN 'July' WHEN substr(date, 6, 2) = '08' THEN 'August' WHEN substr(date, 6, 2) = '09' THEN 'September' WHEN substr(date, 6, 2) = '10' THEN 'October' WHEN substr(date, 6, 2) = '11' THEN 'November' WHEN substr(date, 6, 2) = '12' THEN 'December' END AS month, date, booster_version, launch_site, landing_outcome FROM SPACEXWHERE landing_outcome = 'Failure (drone ship)' AND substr(date, 0, 5) = '2015';
```

Query

This query retrieves all of the records that have failure outcomes in drone ship in the year 2015. It also shows the month names, the booster versions and launch site.

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

Query

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEX
where date between '2010-06-04' and '2017-03-20'      group by
landing_outcome      order by count_outcomes desc;
```

This SQL query counts the number of missions for each landing_outcome within the date range from June 4th 2010, to March 20th 2017, in the SPACEX table. It groups the results by landing outcome and sorts them in descending order of the count.

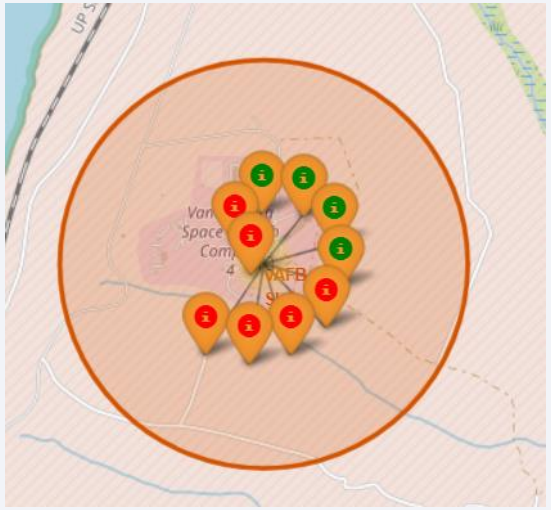
Section 3

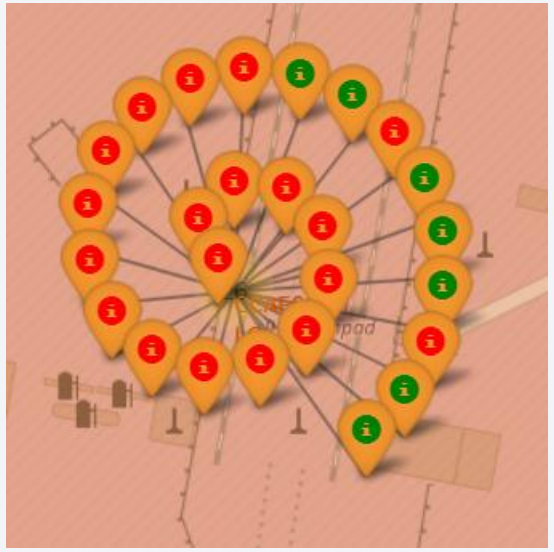
Launch Sites Proximities Analysis

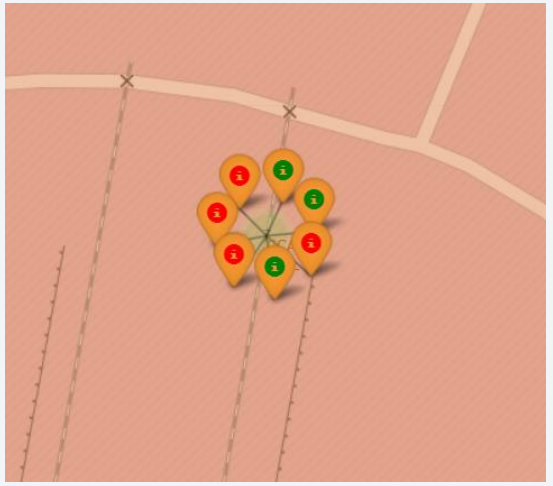
Folium Map, all Launch Sites

This map shows all of the launch sites in the global map. It's important to know where they are located.

Folium Map: Launch Outcomes

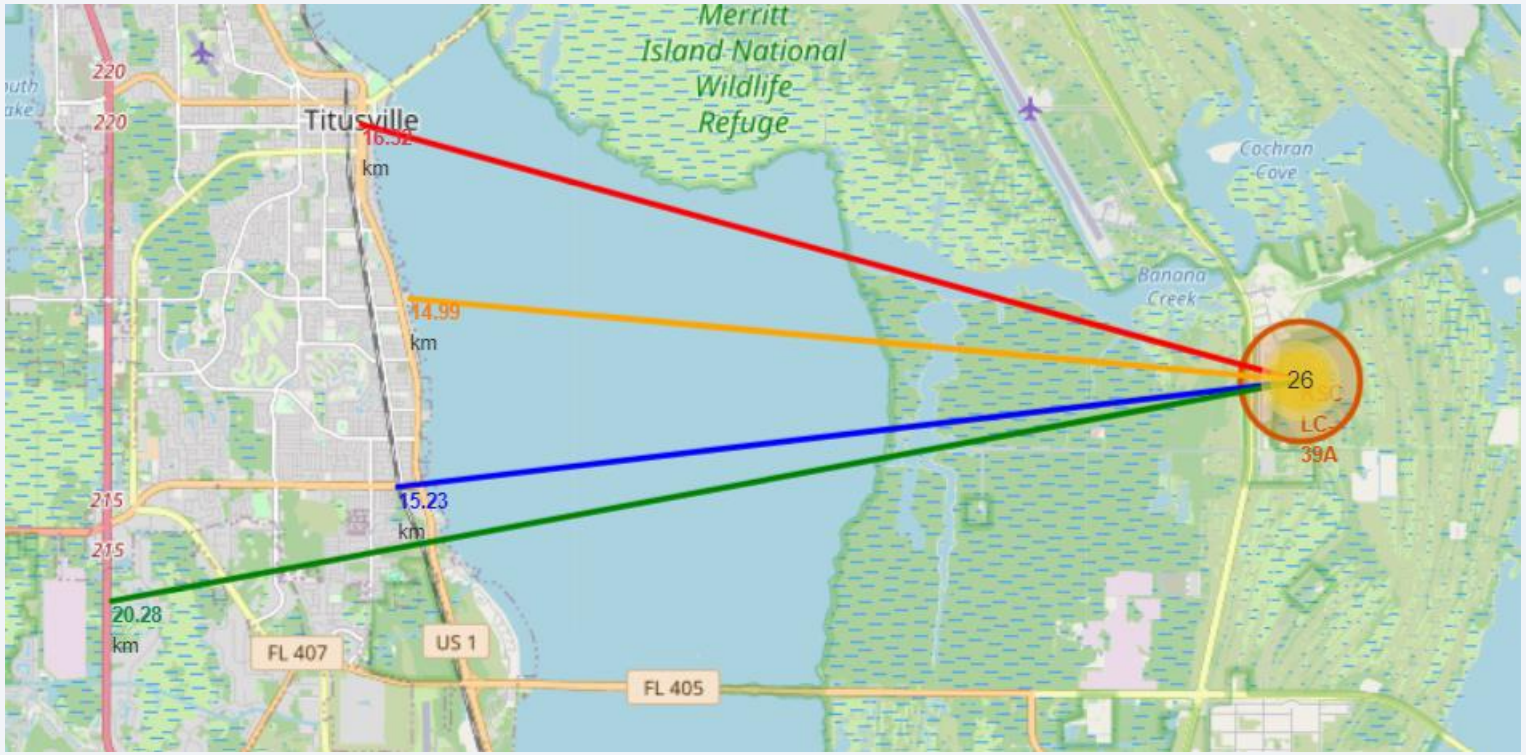




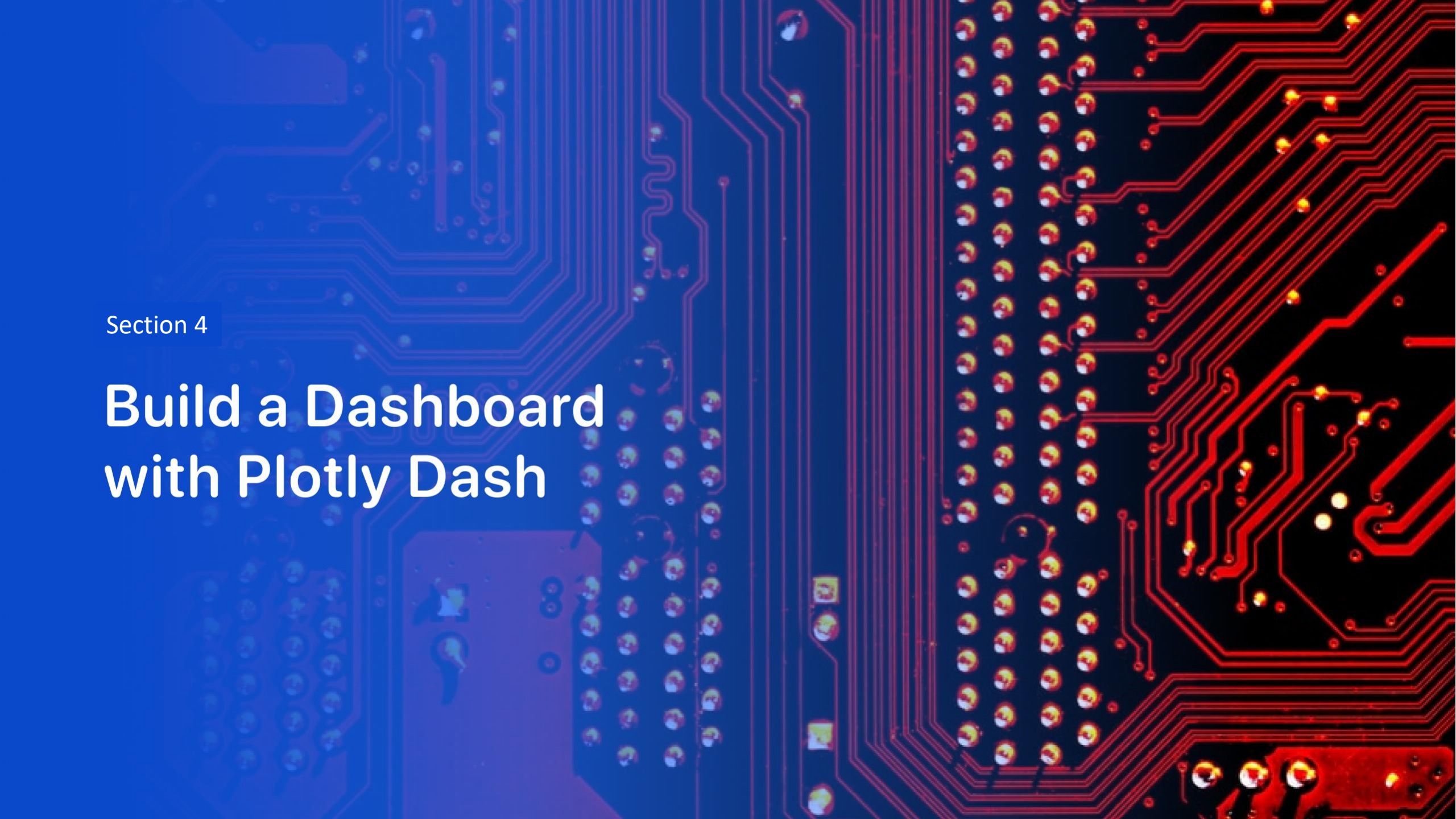


The markers on the map show the launch amounts per site and their respective outcomes. This indicates the success rate of launches per launch site.

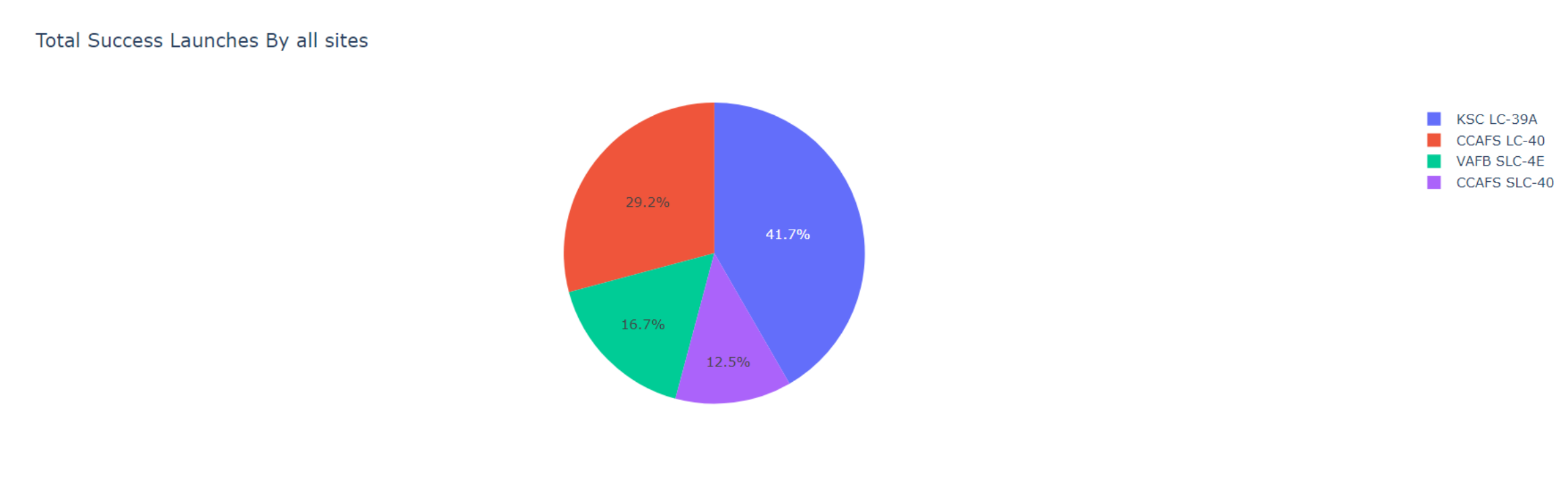
Folium Map: Proximities from Launch Sites



This map shows the proximity of launch site KSC LC-39A to a City (Red), Coastline (Yellow), Railway (Blue), Highway (Green).

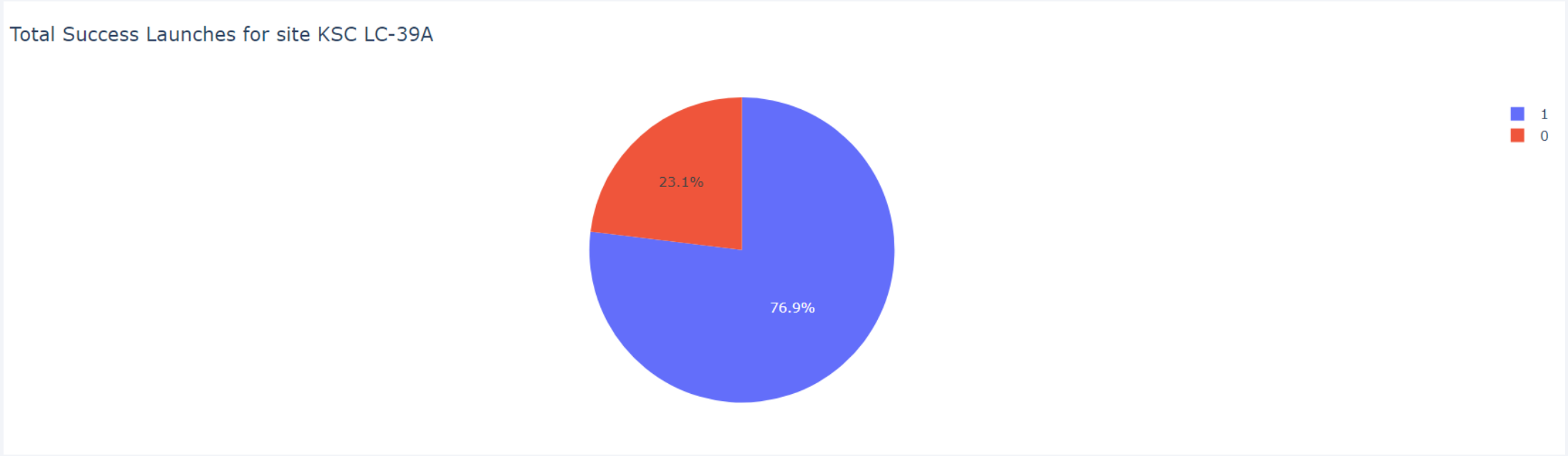

Section 4
Build a Dashboard with Plotly Dash

Dashboard: Launch Successes in All Launch Sites



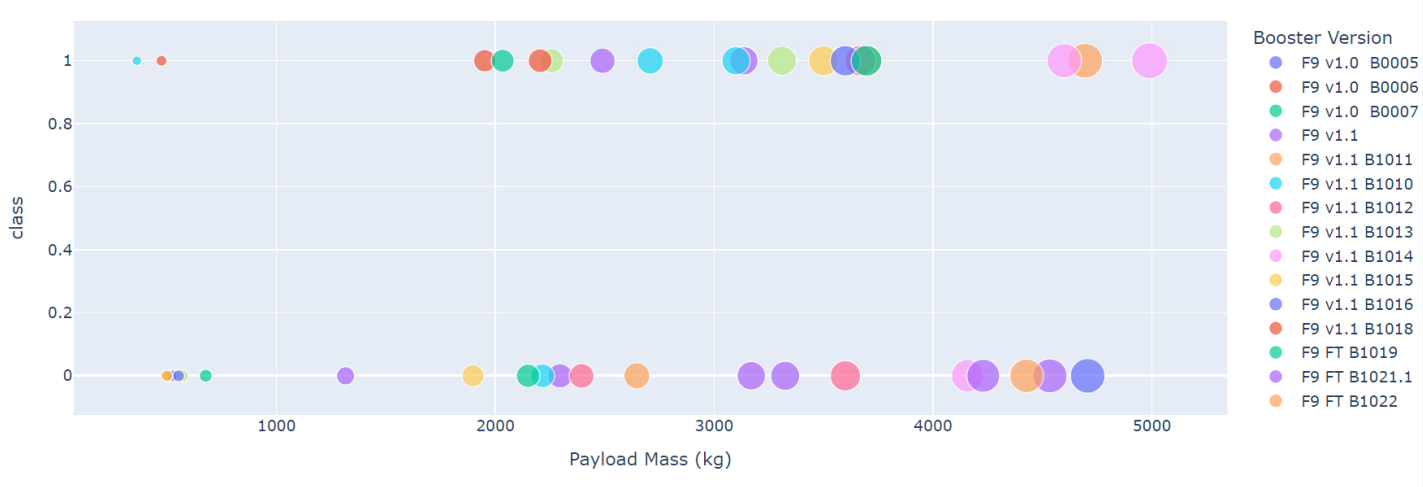
This pie chart shows what percentage of successful launches corresponds to each site out of number the successful launches. This pie chart shows that site KSC LC-39A and CCAFS LC 40 have been the sites that have had the biggest number of successful launches. This could indicate many things like favorable conditions for launches like weather and terrain.

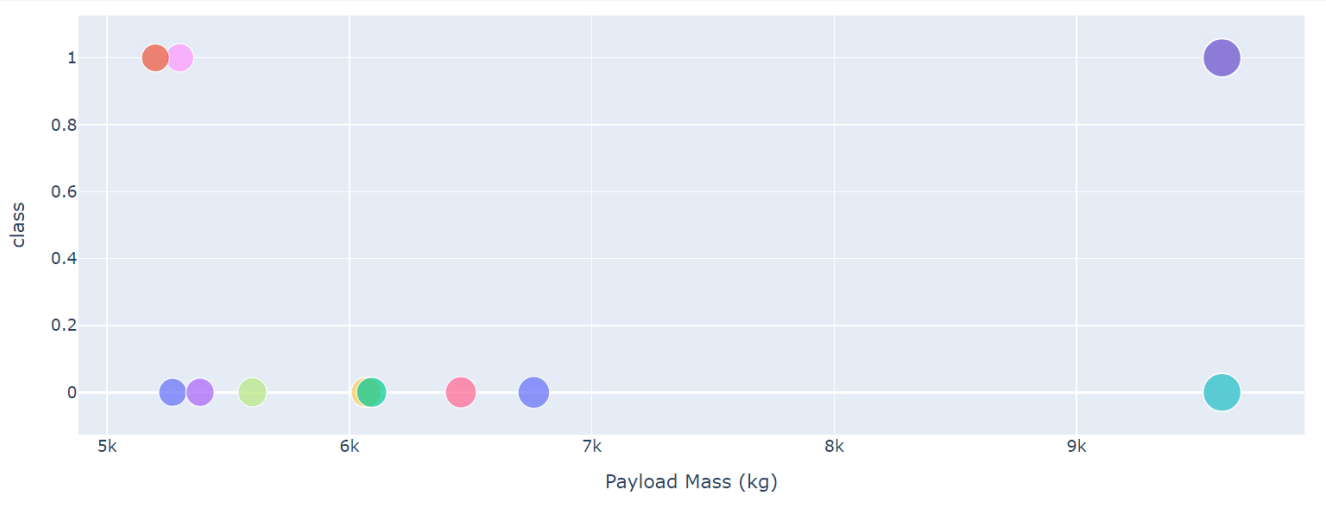
Dashboar: Highest launch success ratio



This pie chart shows the site with the highest launch success ratio. This means that out of all the launch sites, site KSC LC-39A has the highest percentage of successful launches done.

Dashboard: Payload vs. Launch Outcome





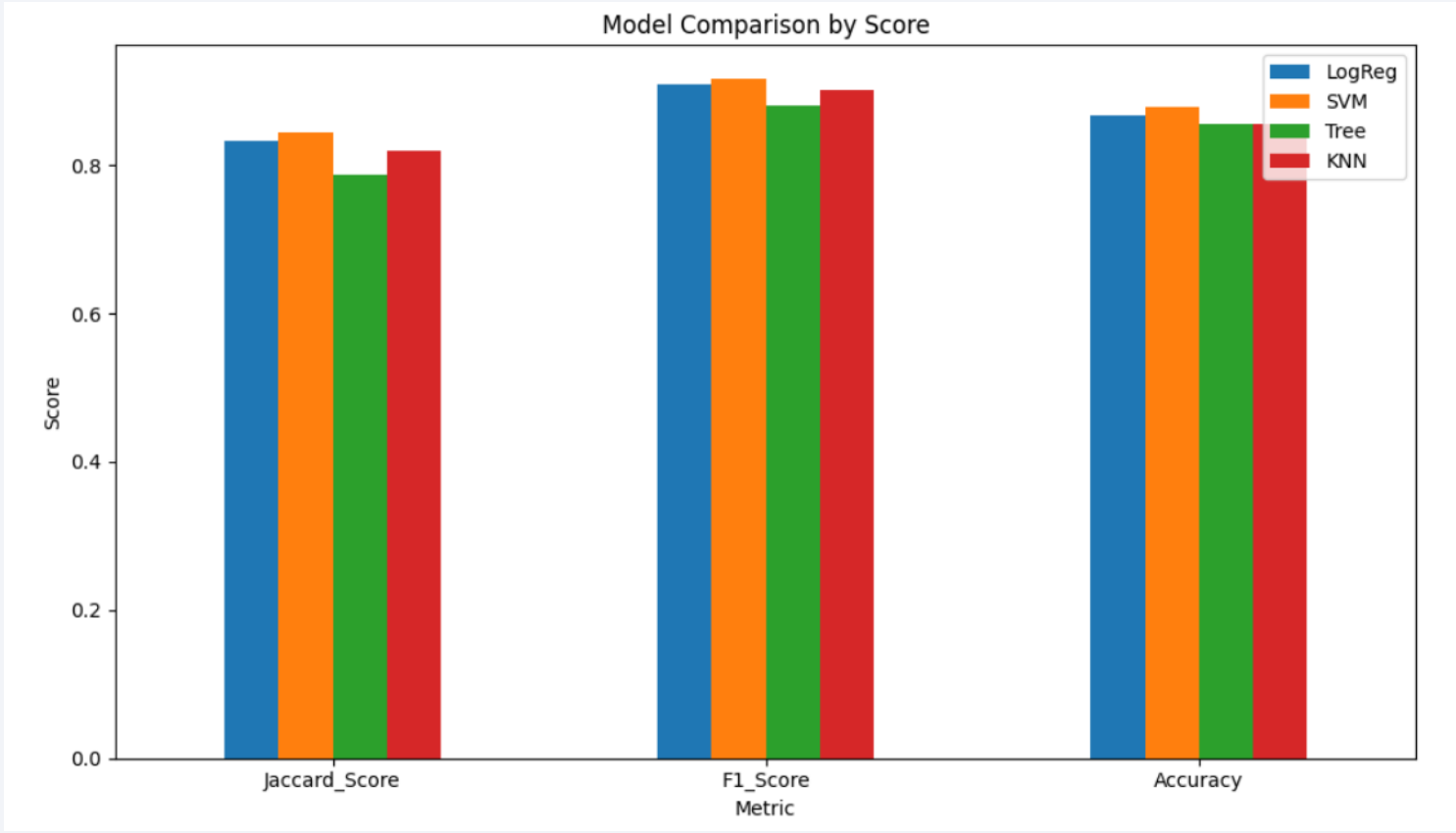
These plots show the ranges between payloads of 0 to 4000kg and 4000 to10,000kg. It indicates that launches done within the 0 to 4000kg range have a higher success rate than those done in ranges of 0 to 10,000.

Section 5

Predictive Analysis (Classification)

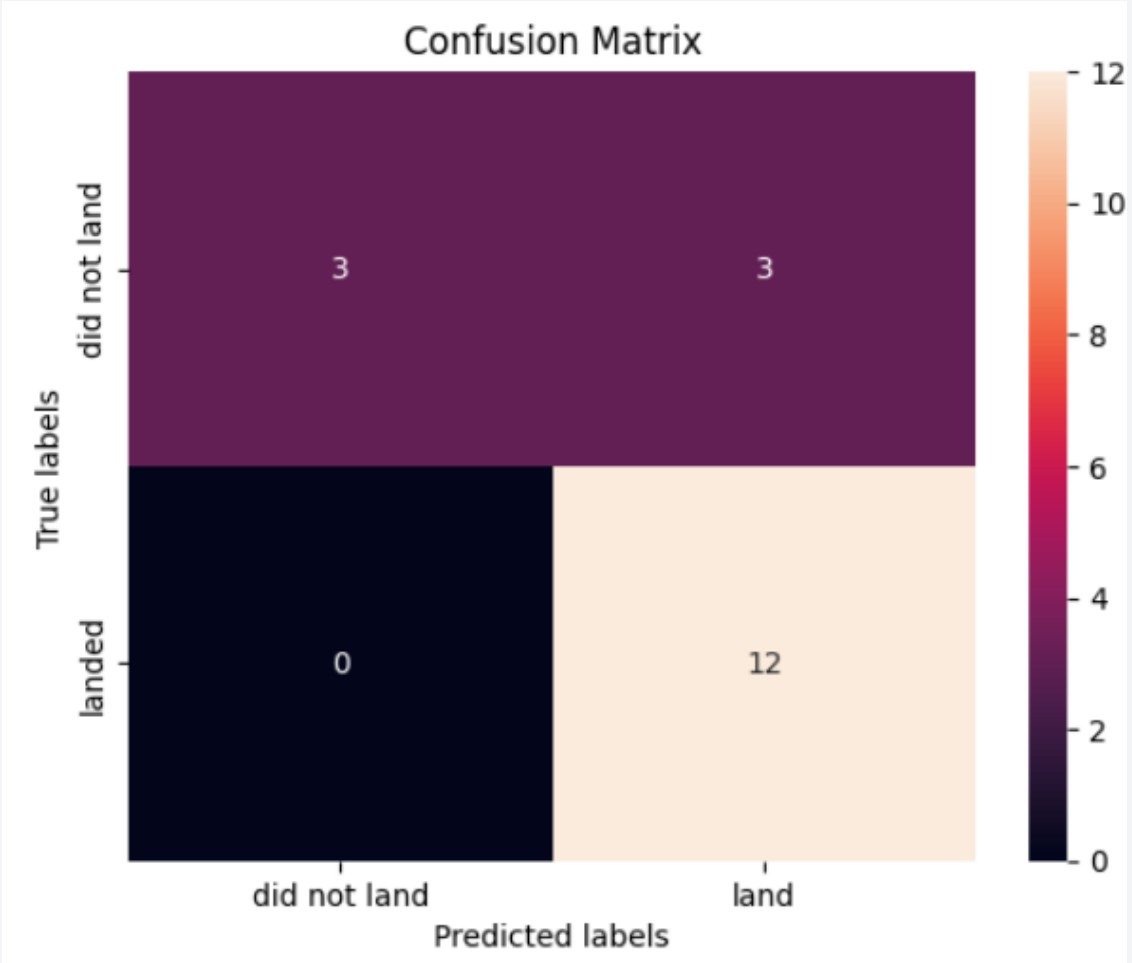
Classification Accuracy

By observing the bar chart and comparing values, the model that has the most classification accuracy is the SVM model, followed by LogReg, then KNN and finally Tree.



Confusion Matrix: SVM

The confusion indicates a very high recall, meaning that it can identify all the true positives (landed-landed). It has good precision and accuracy. It is good that it has no false negatives, but the model can be made better to reduce the number of false positives (predicted landing but didn't land).



Conclusions

- Since 2014, success rates have increased and continue to increase while unsuccessful launch rates have decreased.
- Orbit types ES-L1, GEO, HEO, SSO and SO need more launches to determine their real success rate.
- Excluding the orbit types that need more launches, orbit VLEO has the highest and more accurate success rate out of all the orbit types.
- Site KSC LC-39A has the highest successful launch ratio out of all launch sites.
- The SVM model is the most accurate model for the task.

Appendix

Github link of Lab Notebooks and .py Dash app.

https://github.com/Gil-ib/Applied-Data-Science-Capstone/blob/main/README.md

Thank you!