IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zain Makhdum
24th December 2024

Outline

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

Executive Summary

Summary of methodologies

- Collection of data by use of API
- Collection of data utilizing Web scrapping approach
- Wrangling of data gathered/obtained
- Analytical procedures with Data Visualization
- Exploratory Data Analysis with SQL
- Enhancing visualization analytics by making use of Folium and Ploty Dash
- Applying Machine Learning models for predictive analysis for e.g. SVM, Logistic Regression, KNN and Decision Tree

Summary of all results

- There has been a marked improvement in the launch success rate over time.
- Launch Site KSC LC-39A shows the highest success rate in comparison to the rest of the launch sites.
- Orbits pertaining to ES-L1, SSO, HEO, and GEO were highlighted as having the highest success rates.

Introduction

Project background

SpaceX, a highly successful commercial space firm, that has made space travel affordable its main goal. The business advertises its Falcon 9 rocket launches on its website, costing 62 million dollars in comparison to the 165 million going market price.

It attributes the saving of 103 million to the fact that the first stage of landing can be reused.

Hence the cost of launch will be heavily dependent on the result of the first stage. Our procedures will be centered around the data pertaining to this scenario.

Questions for which answers are sought

Does the rate of successful landings increase or decrease over the years?

How do the various features/variables in the collected dataset correlate with the success of landing?

Which area proximities have the highest launch success rate?

What is the best predictive model to be used for this scenario?

Section 1

Methodology

Methodology

Executive Summary

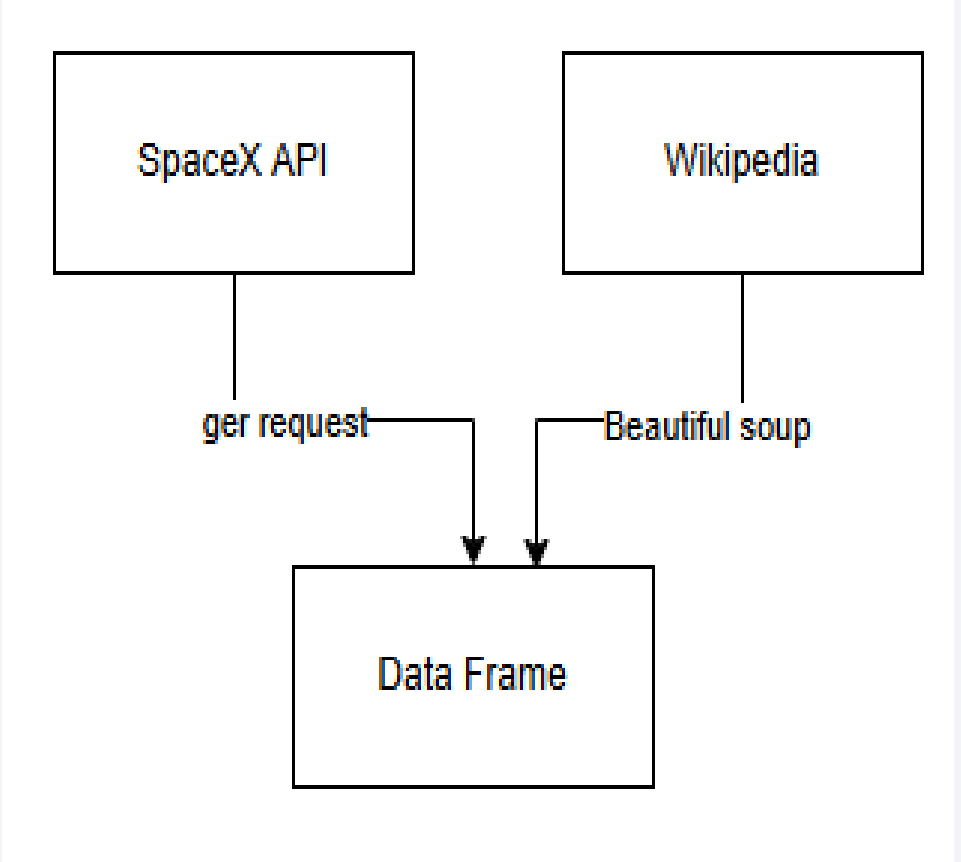
- Data collection methodology:
  - Collected by SpaceX API calls and Webscrapping Wikipedia SpaceX info
- Perform data wrangling
  - Identifying/addressing missing values and working with categorical columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Splitting data into training and testing
  - Working with 4 machine learning algorithms and their respective evaluations for the highest accuracy

Data Collection

Describe how data sets were collected

Data was collected in two phases for this project for the purpose of complete/exhaustive information on the topic. Firstly, we requested rocket launch data from SpaceX API. The result was converted into a pandas data frame.

In the second phase, we webscrapped Falcon 9 launch records from Wikipedia (also converted into a data frame).

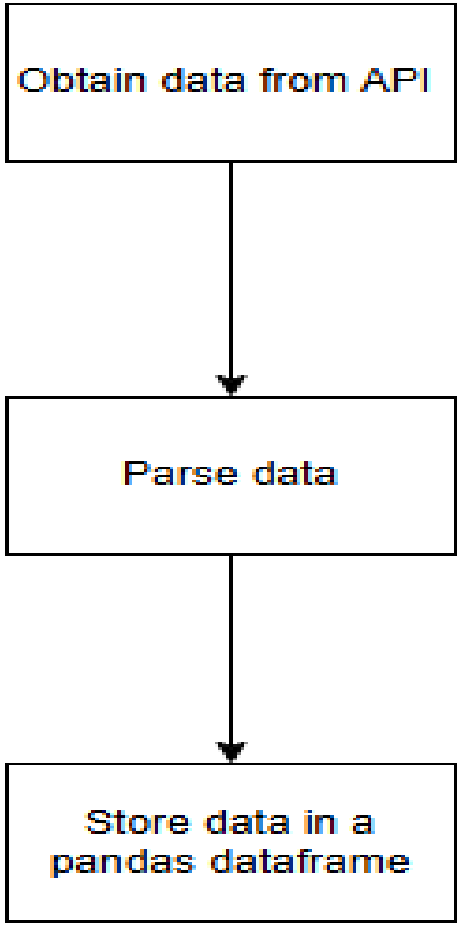


Data Collection – SpaceX API

We initiated collecting data from the SpaceX API by importing the necessary libraries and writing request. We then created a URL GET request to obtain the data in JSON format.

This data was subsequently converted into a data frame by extracting relevant columns/information such as geospatial details, rocket type, orbit etc

GITHUB link

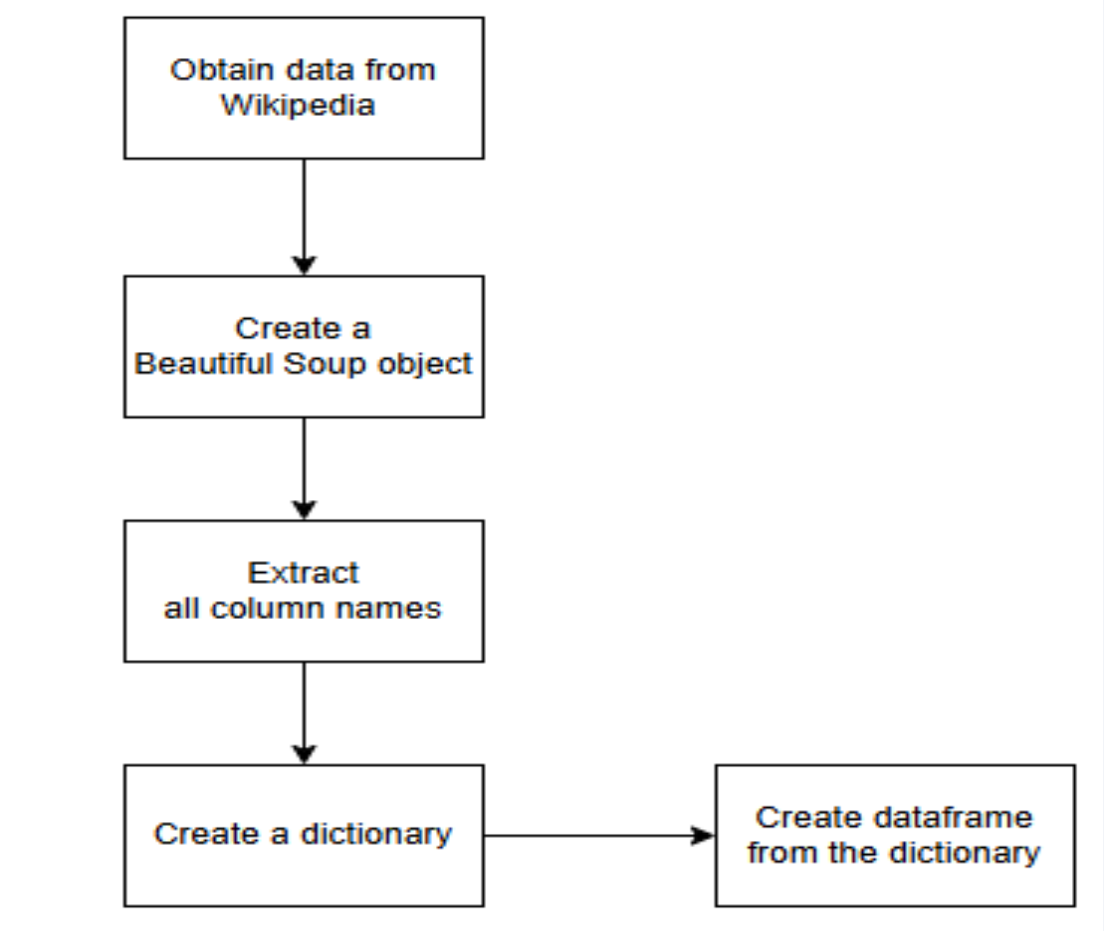


Data Collection - Scraping

We requested data from Wikipedia pertaining to Falcon 9 launches. Beautiful Soup library was used to extract tables and columns from the HTML response.

Finally, a data frame was created using the extracted data.

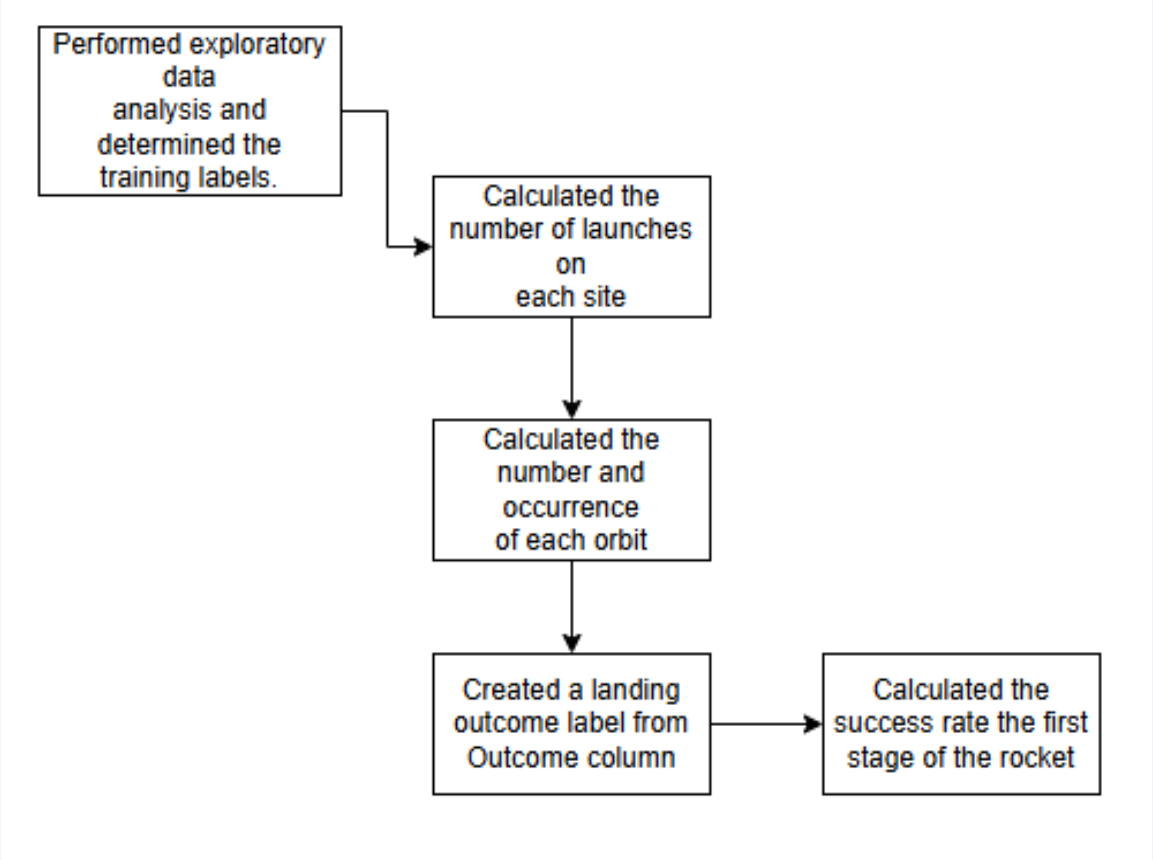
GITHUB link



Data Wrangling

This stage heavily encompasses data preprocessing and Exploratory Data Analysis which includes:

- Cleaning the data
- Choosing the features from the dataset to work on
- Making calculations for exploring and analyzing data
- Converting landing outcomes into categorical data containing ‘1’ and ‘0’



GITHUB link

EDA with Data Visualization

Charts plotted

- Flight Number vs Payload Mass (kg)
- Flight Number vs Launch Site
- Payload Mass (kg) vs Launch Site
- Flight Number vs Orbit Type
- Payload Mass (kg) vs Orbit Type
- Orbit Type vs Success Rate
- Success Rate Yearly Trend

GITHUB link

- Success Rate Yearly Trend

Charts significance

- Scatter plots illustrate how variables relate to one another. They could be incorporated into a machine learning model (if a relationship exists) for predictive analysis
- Line charts display data patterns throughout time
- Comparisons between distinct categories are displayed in bar charts. The objective is to demonstrate the connection between a measured value and the particular categories

EDA with SQL

The following SQL queries were written:-

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship / booster versions / launch site names for the months in year 2015
- Ranking 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

GITHUB link Yearly Trend

Build an Interactive Map with Folium

Launch site markers

- Created blue circle at NASA Johnson Space Center's coordinate with popup text revealing its name
- Created red circles at all launch sites coordinates with popup text

Launch outcomes markers

- Created colored markers for successful/failed launches at each launch site to determine success rates

Colored lines to mark distance

- Created colored lines to mark distance from site CCAFS SLC 40 to the nearest coastline, railway, highway, and city

GITHUB link

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

Enables selection of all launch sites/ specific launch site

Slider of Payload Mass Range

Enables selection of payload mass range

Pie Chart Showing Successful Launches

Enables visualization of successful/failed launches as a percent of the total launches

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

Enables visualization of correlation between Payload Mass and Launch Success

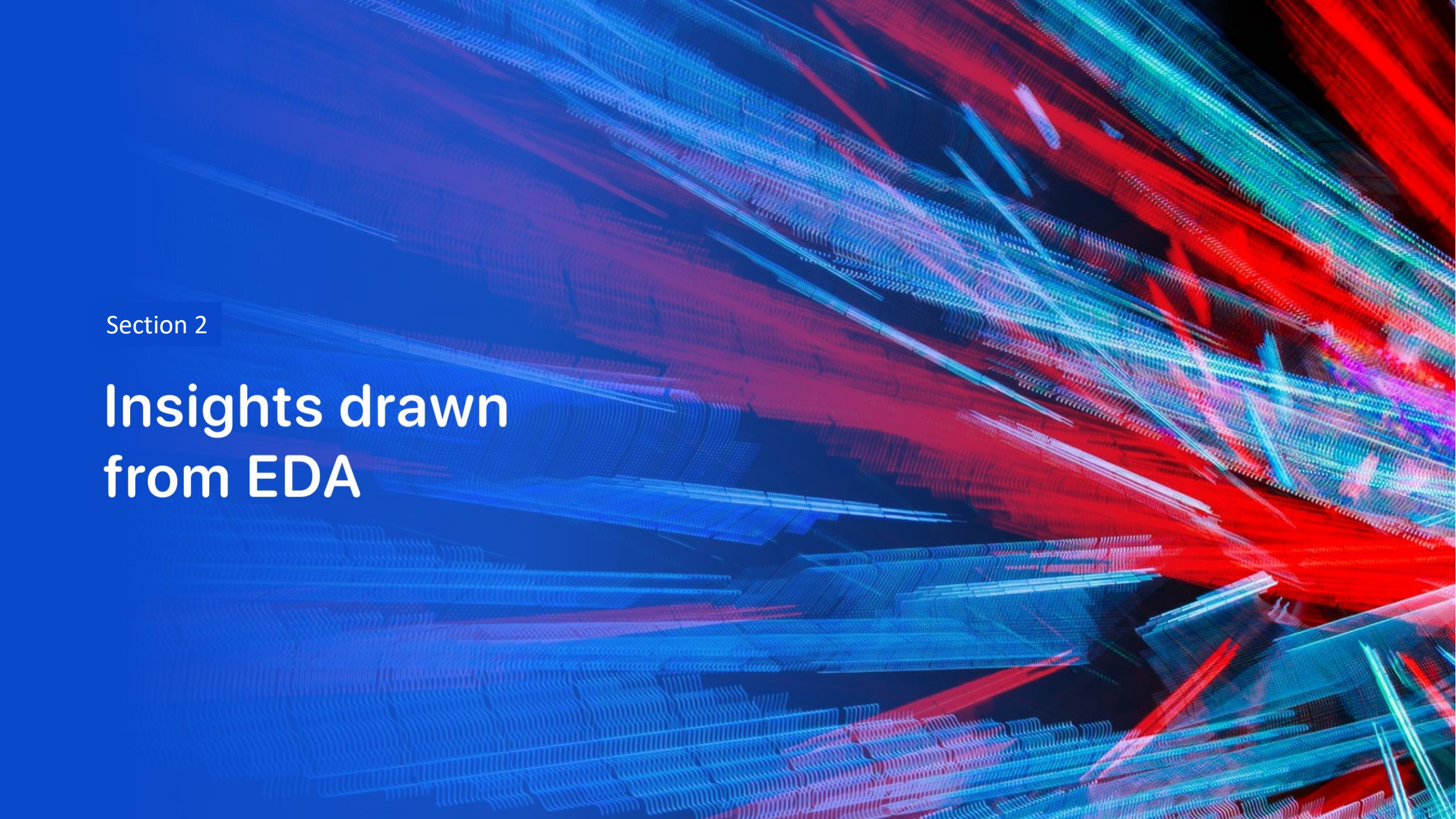
Predictive Analysis (Classification)

- Creating NumPy array from the Class column
- Standardizing the data with StandardScaler function
- Splitting the data using train_test_split (20:80)
- Tuning different hyper parameters via GridSearch on various models: logistic regression, support vector machine, decision tree and K Nearest Neighbor
- Calculating accuracy on the test data with .score() for the above algorithms
- Assessing/Identifying the most optimized model using Confusion Matrix, Jaccard score and F1 score metrics

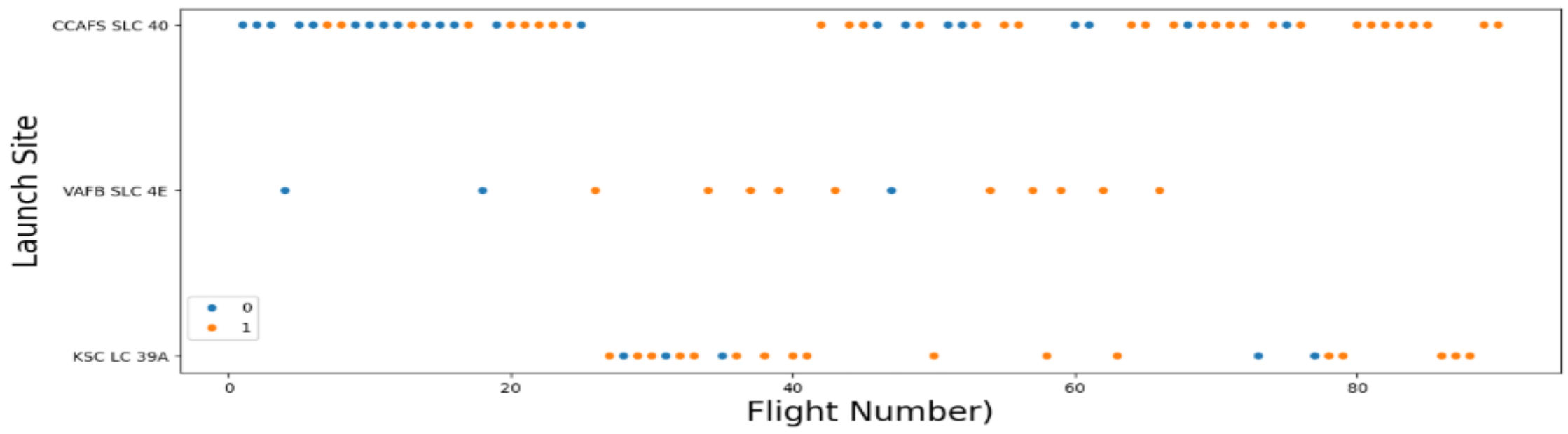
GITHUB link

Results

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

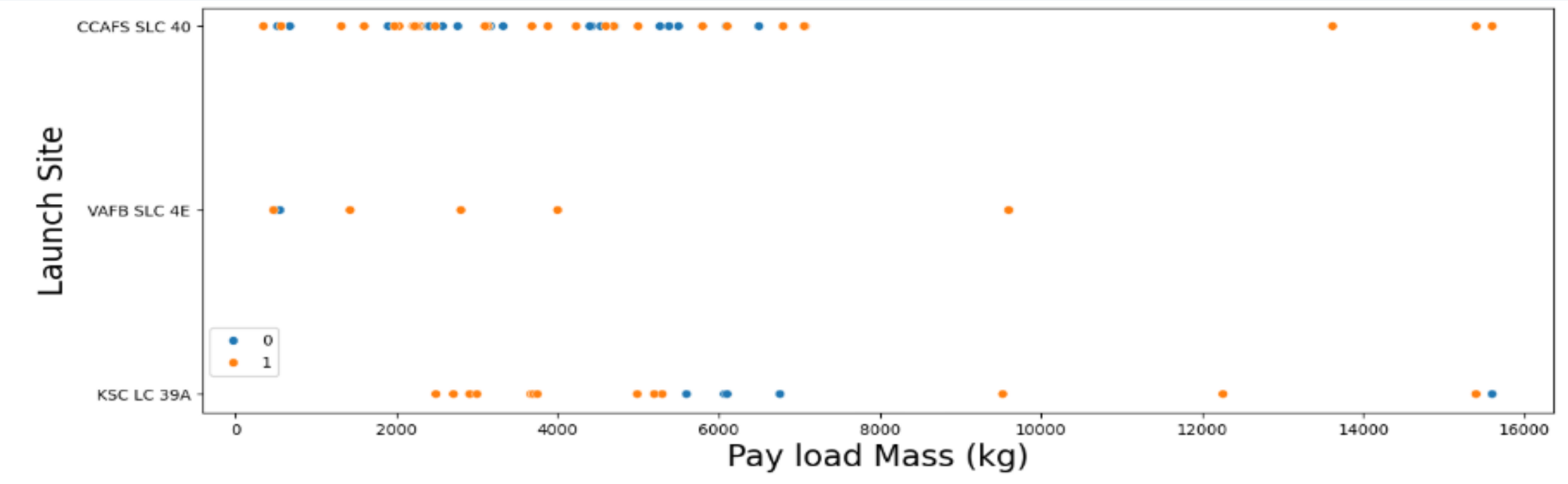

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- We can deduce that as the years have gone by, the launches had a higher possibility of success (namely VAFB SLC 4E and KSC LC 39A)
- CCAFS SLC 40 launch site has been identified as the most frequently used site for launches

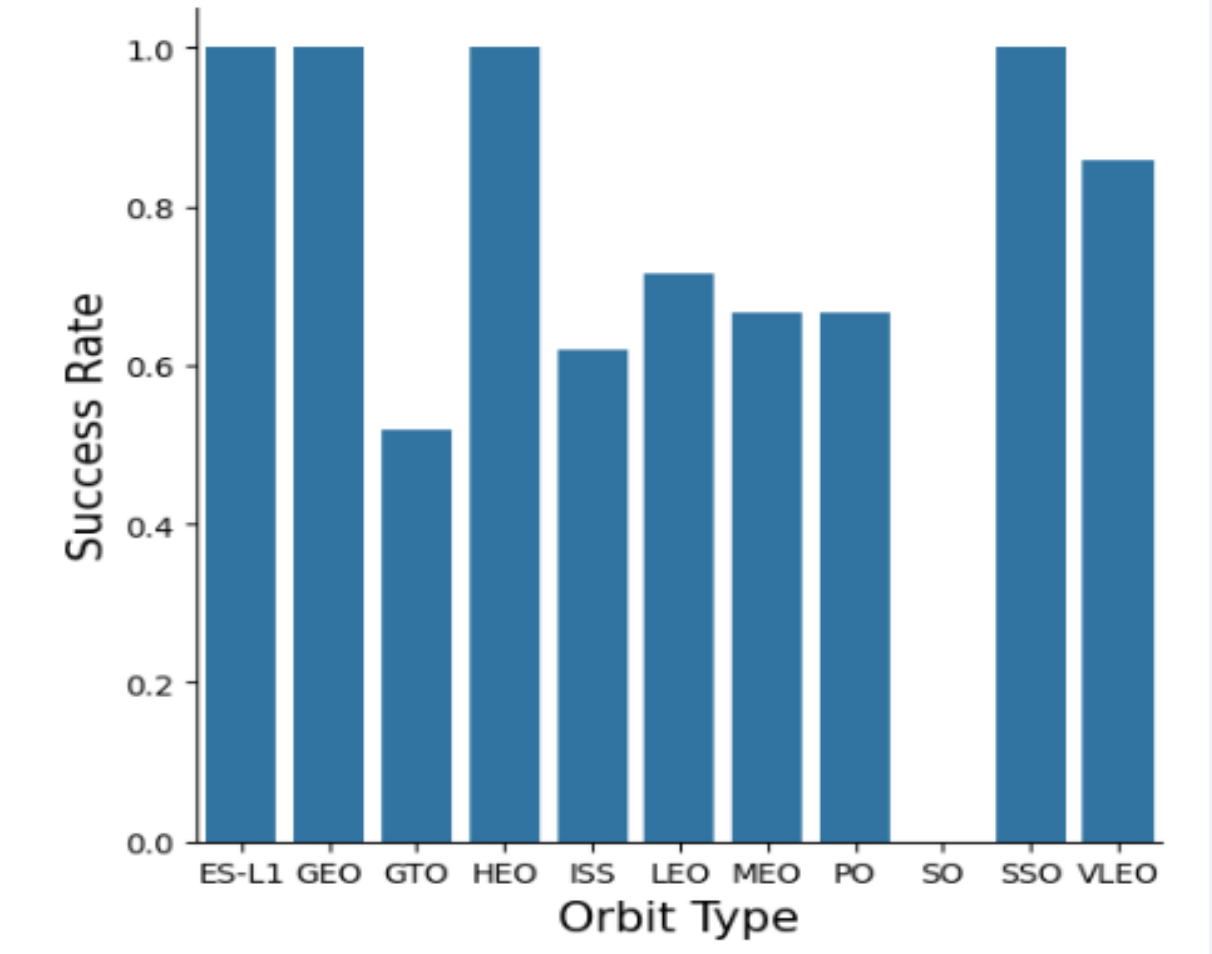
Payload vs. Launch Site



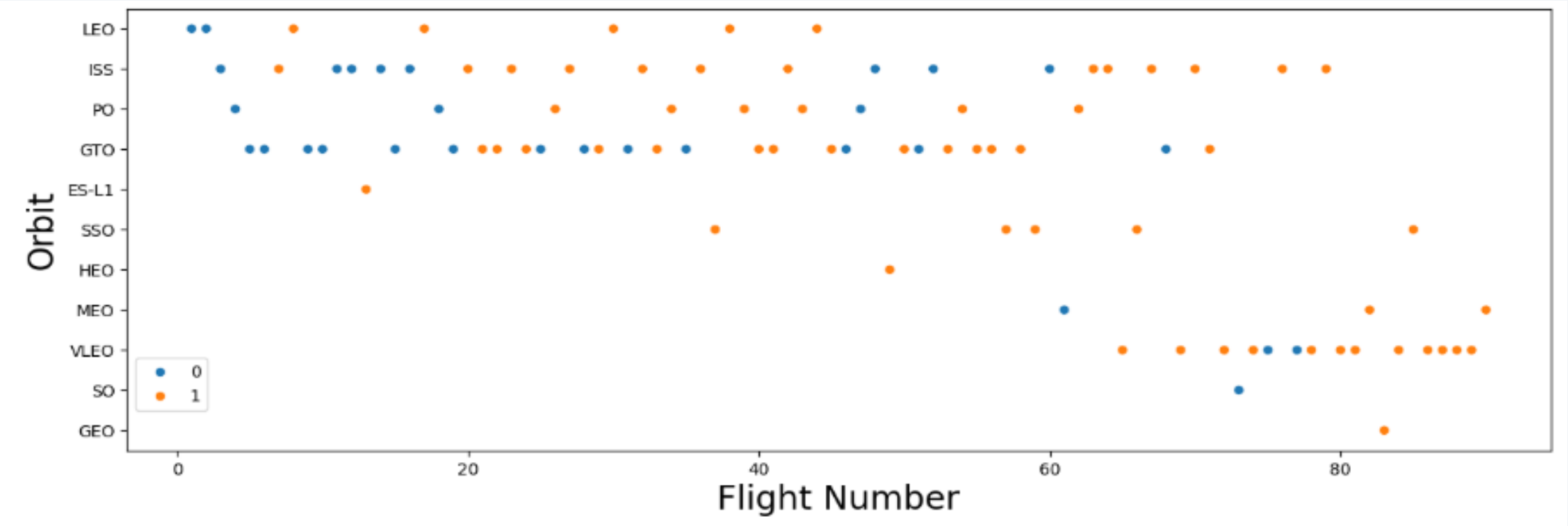
- We can deduce that overall the Pay load Mass and the successful launches have a proportional relationship i.e. the greater the payload weight, the higher the success
- However, KSC LC 39A has a whooping 100% success rate for launches less than 5,000 kg of payload mass

Success Rate vs. Orbit Type

- We can deduce that best orbits with 100% success rate are ES-L1, GEO, HEO and SSO
- The worst performing orbit has been SO
- The rest have achieved moderate success

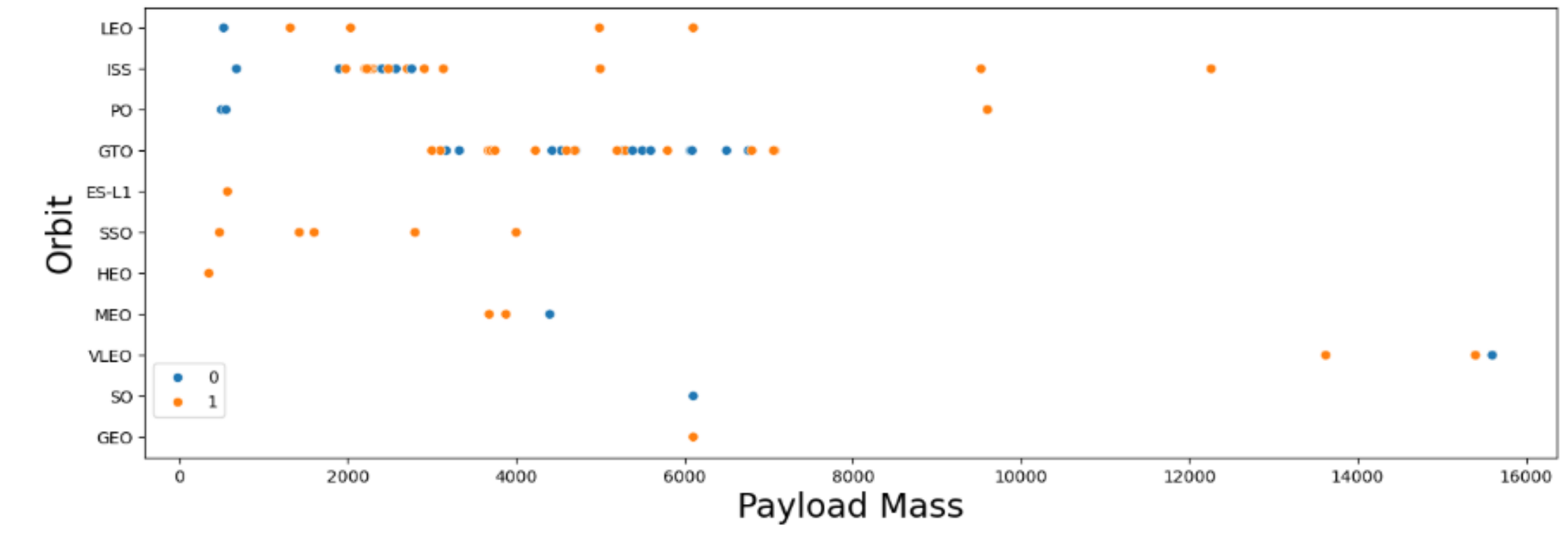


Flight Number vs. Orbit Type



- We can deduce that for Leo orbit success rate is directly proportional to number of flights
- However, for GTO, no such relationship exists

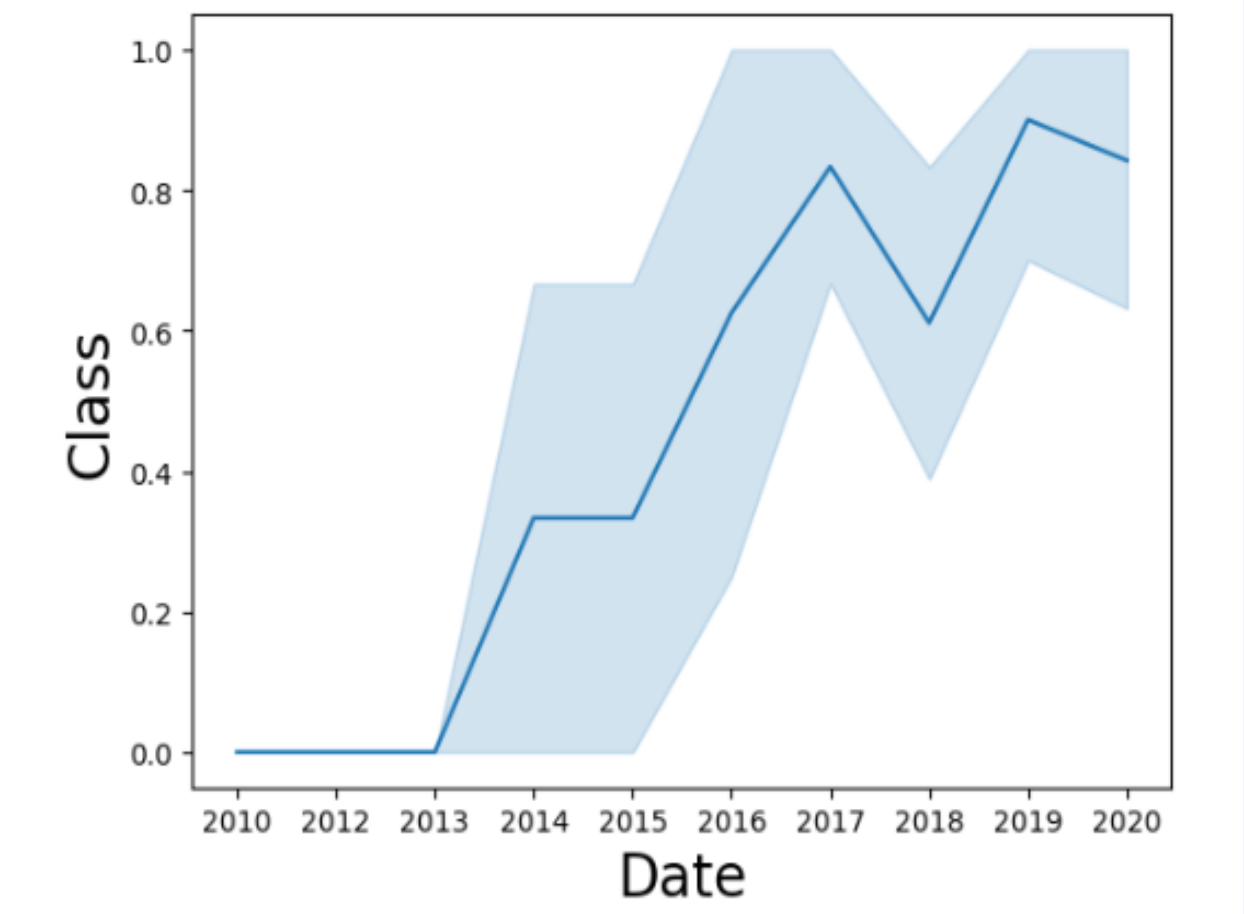
Payload vs. Orbit Type



- We can deduce that higher payload mass has a positive influence on PO,LEO and ISS orbits
- On the other hand, GTO orbit has mixed success with heavier payloads

Launch Success Yearly Trend

- The success rate soared from 2013 to 2017 and 2018 to 2019
- However , it has decreased from 2017 to 2018 and from 2019 to 2020
- Overall, on the average the success rate has increased during the defined time frame



All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql select distinct launch_site from SPACEXTABLE ;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Displaying the names of the unique launch sites

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Displaying 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

TOTAL_PAYLOAD
45596

Displaying the total payload mass carried by boosters launched by

NASA (CRS)

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE Booster_Version == "F9 v1.1"
```

```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

Displaying average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(Date) as LaunchDate FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

LaunchDate
2015-12-22

Listing the date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%%sql SELECT Booster_Version, PAYLOAD_MASS__KG_
    FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(mission_outcome) as counts from SPACEXTABLE GROUP BY mission_outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

Listing the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
SELECT Booster_Version from SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ IN (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

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

Listing the names of the booster versions which have carried the maximum payload mass

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql SELECT substr(Date, 6,2) as Month, Booster_Version, Launch_Site, Landing_Outcome
    FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date,0,5)='2015'
```

```
 * sqlite:///my_data1.db
Done.
```

Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Listing the failed landing outcomes in month, their booster versions and launch site names for the year 2015

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.

```
%%sql

SELECT Landing_Outcome, COUNT(*) as Count
    FROM SPACEXTABLE
        WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY Landing_Outcome
        ORDER BY Count DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

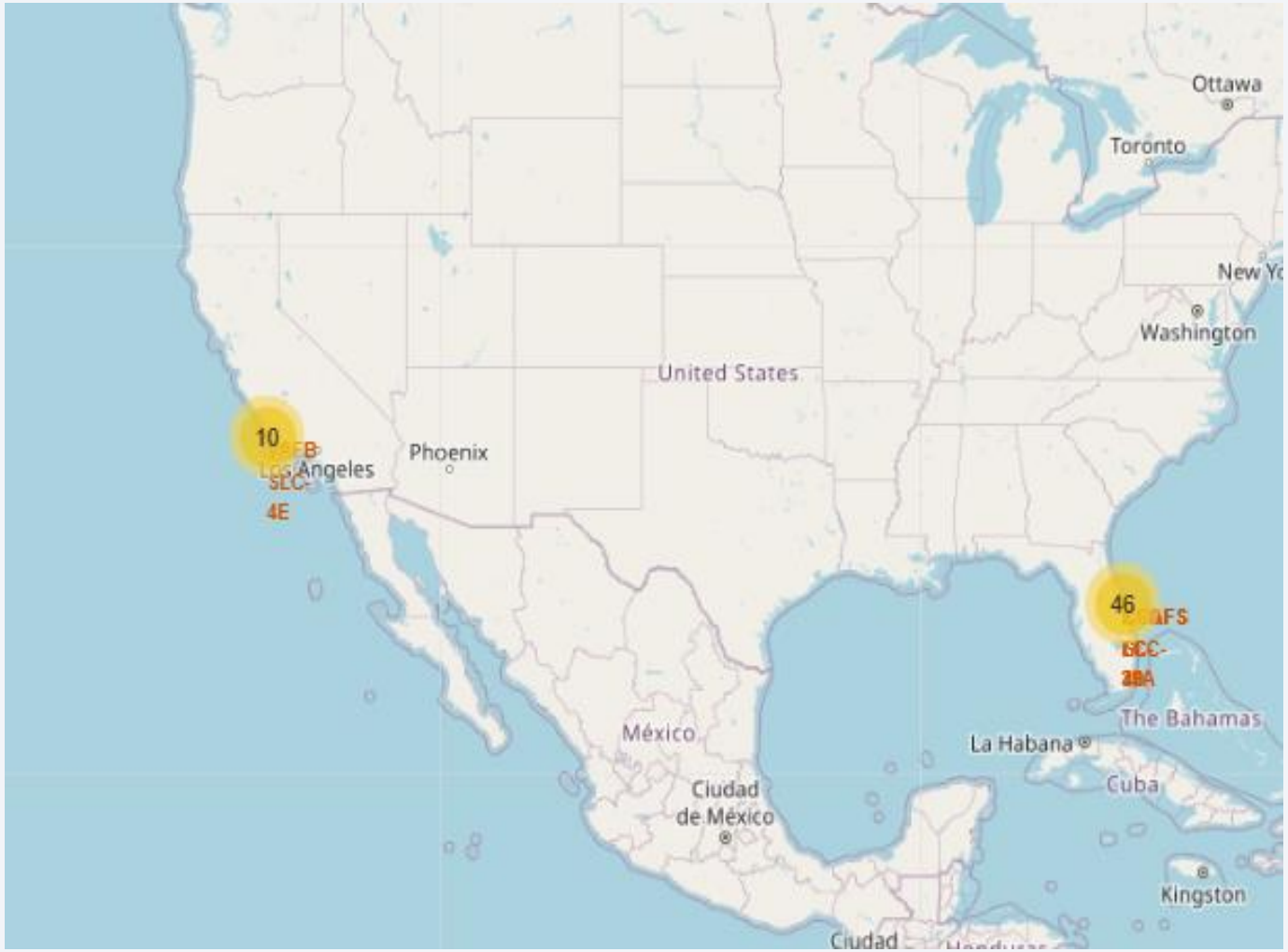
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Ranking 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.

Section 3

Launch Sites Proximities Analysis

Launch sites on the global map

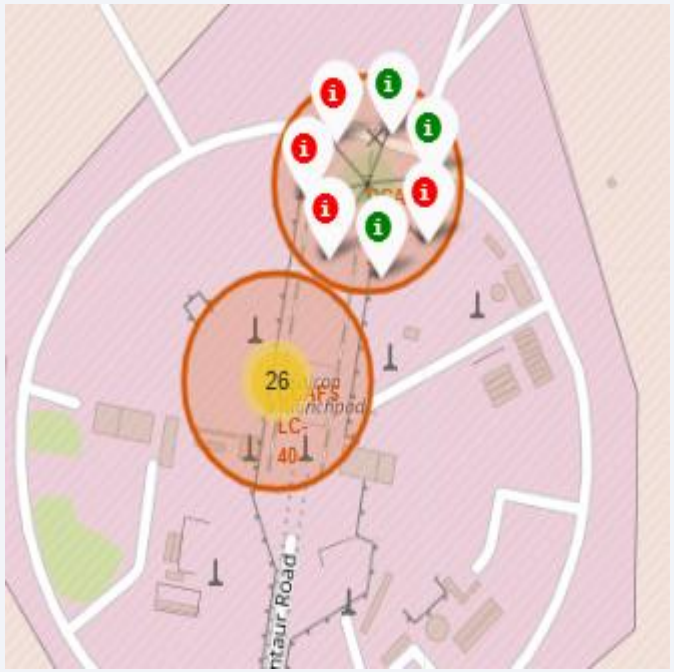


As seen, the launch point has intentionally been kept closer to the equator and the sea.

Rockets launched from such sites, get an additional boost due to the rotational speed of earth

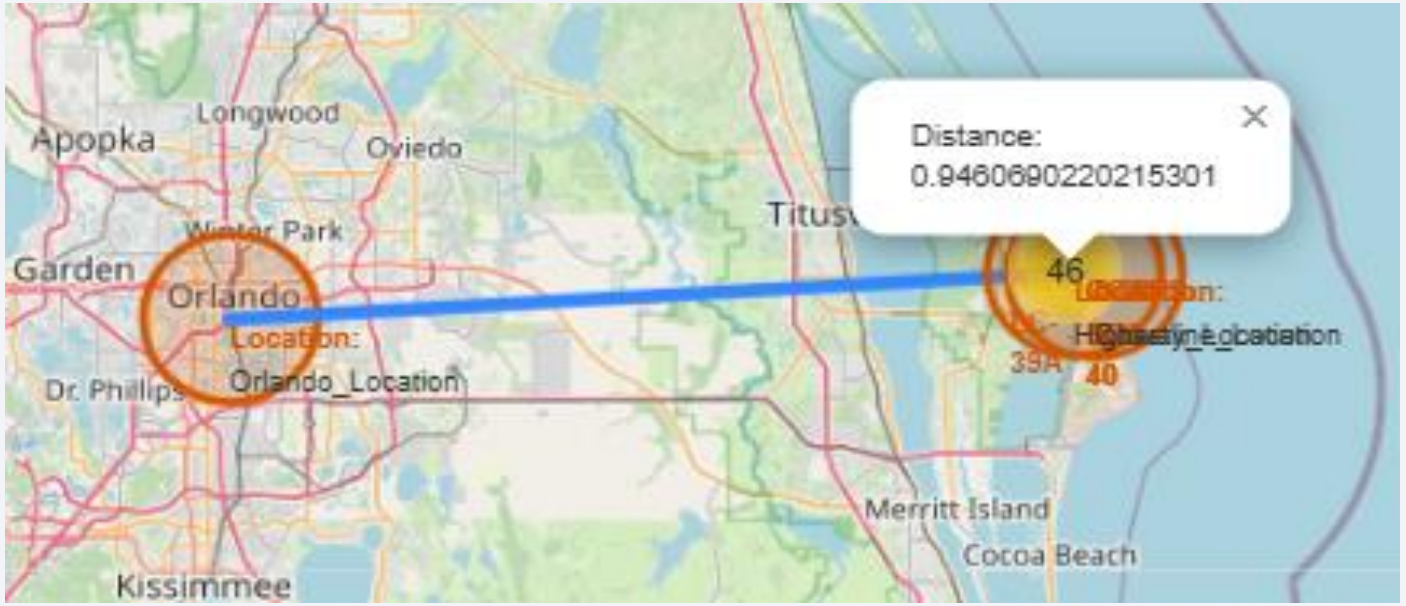
Also, the risk of damage caused by falling debris from faulty launches (rockets exploding in mid air), is minimized due to the launch being in coastal areas

Launch outcomes marked on map

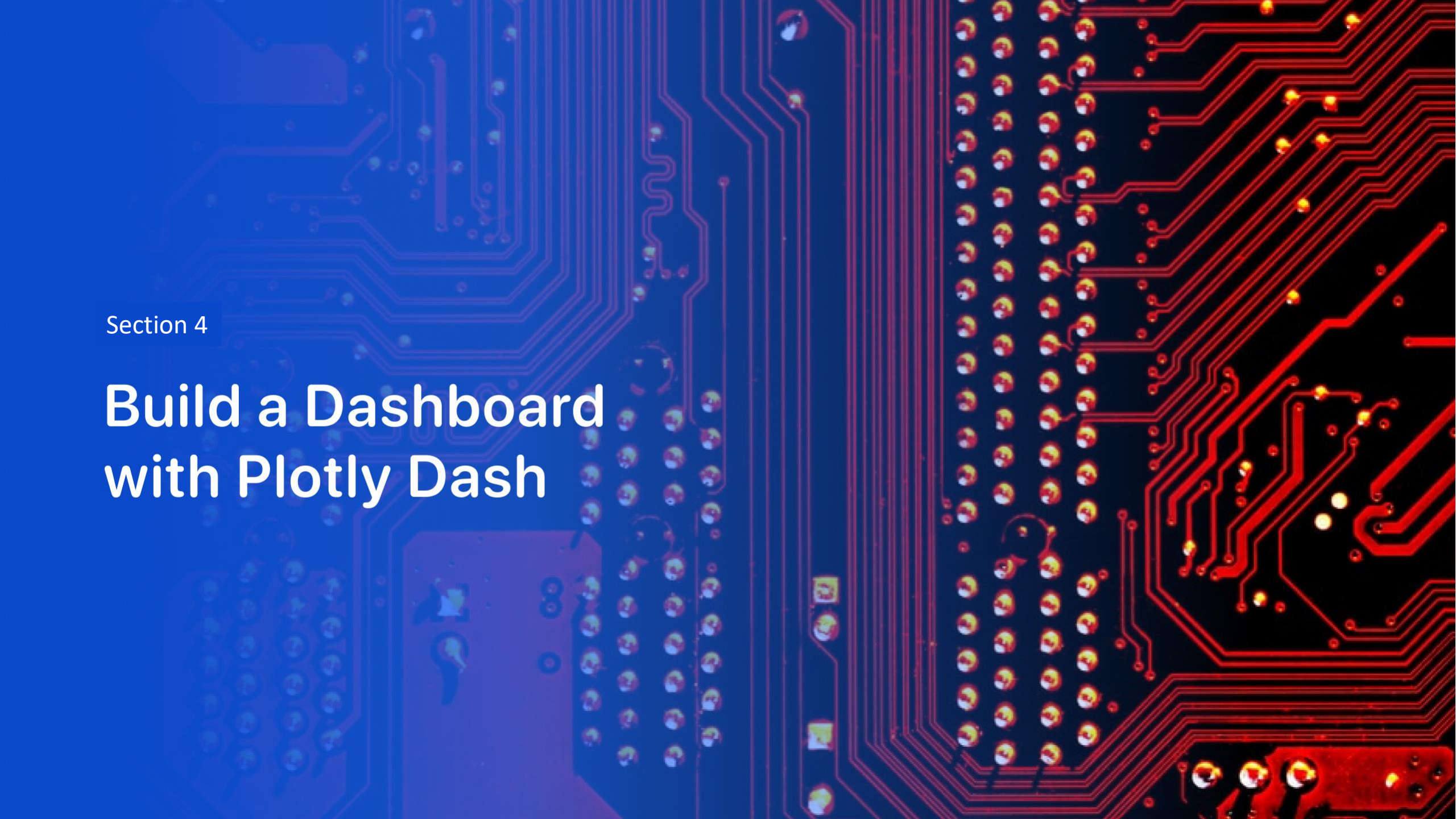


The red markers denote the failed launches and the green ones show the successful ones. Launch Site KSC LC-39A is leading in the number of successful launches.

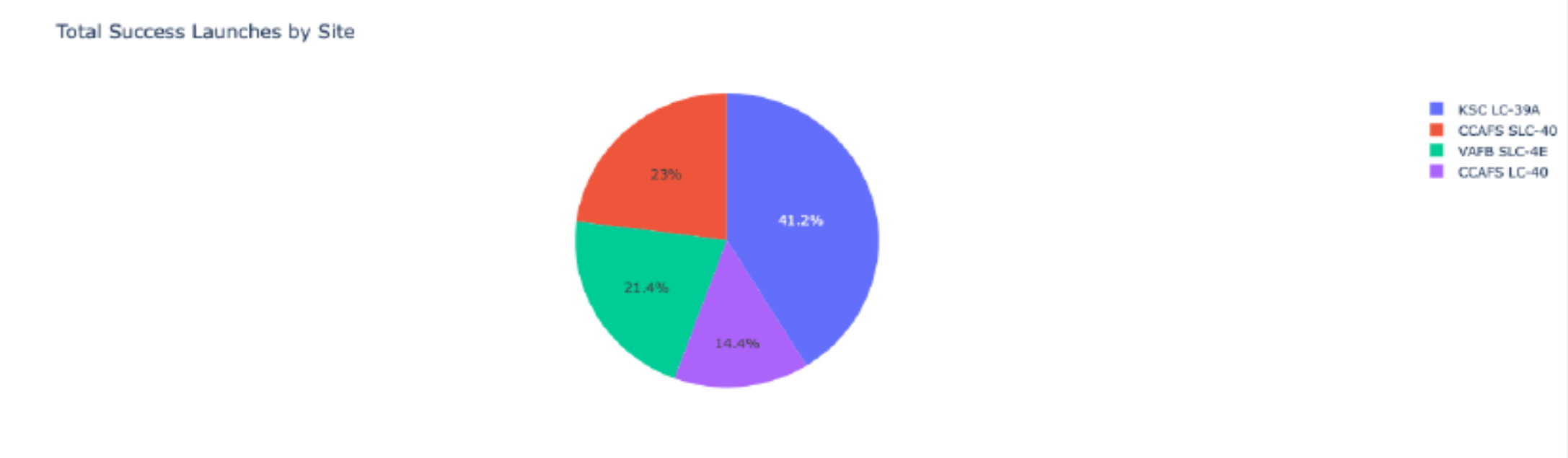
Distance proximities from Launch sites



The distance from cities (Orlando), railways and roads need to be taken into account to understand what would be at risk in the scenario of a failed launch or a mid air destruction of a rocket.


Section 4
Build a Dashboard with Plotly Dash

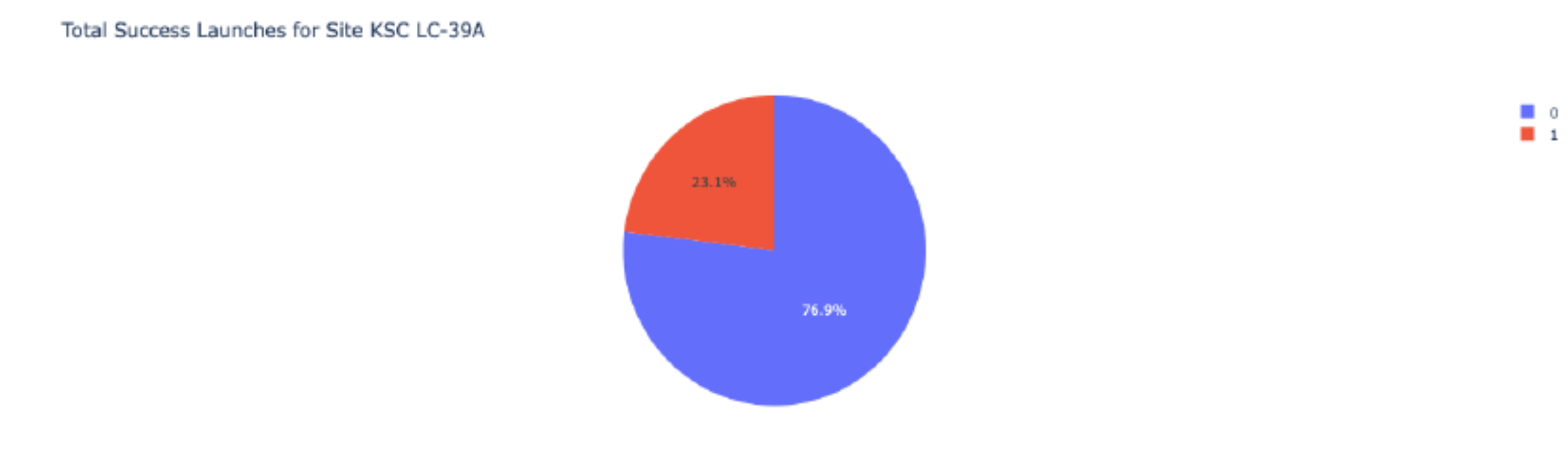
Pie shart of total success launches



KSC LC-39A has exhibited the highest proportion of successful landings

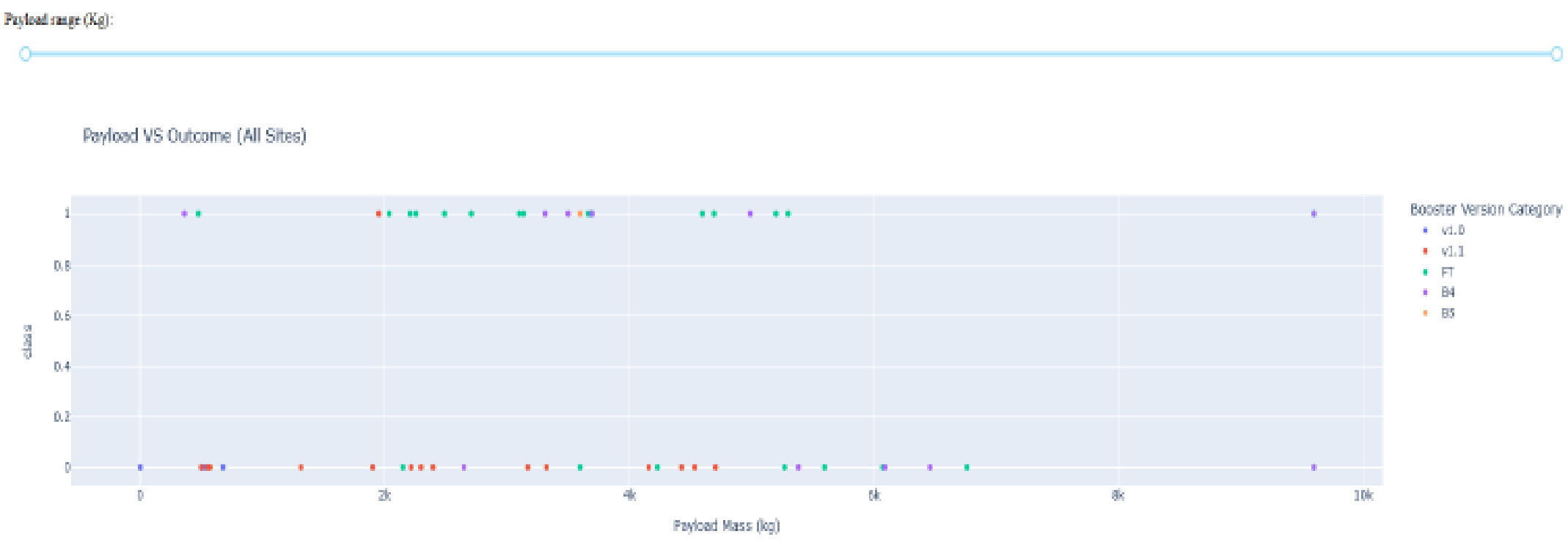
On the other hand, VAFB SLC-4E and CCAFS SLC-40 showed the lowest.

Pie chart of launch site with highest launch success ratio



KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload Mass vs. Launch Outcome for all sites



The chart shows that payloads between 2000 and 5000 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

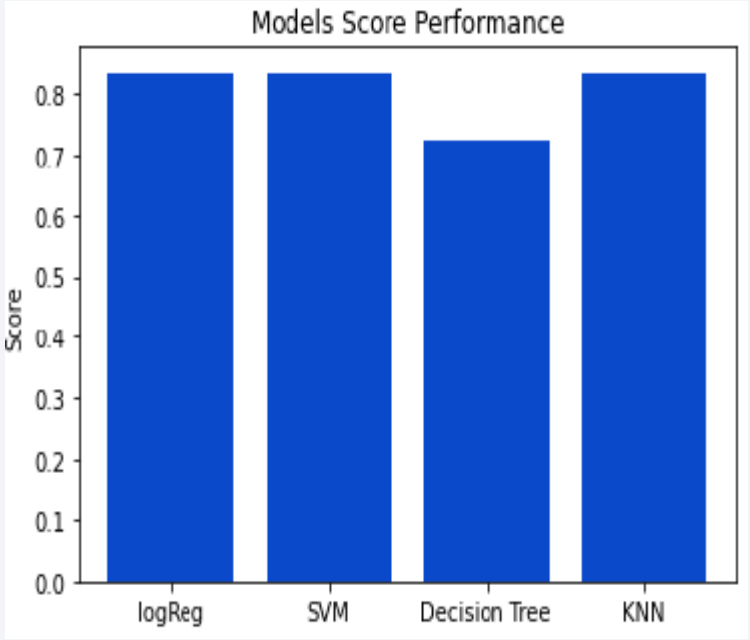
Accuracy pertaining to 18 sample set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.733333	0.800000
F1_Score	0.888889	0.888889	0.846154	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Accuracy pertaining to entire data set

]:

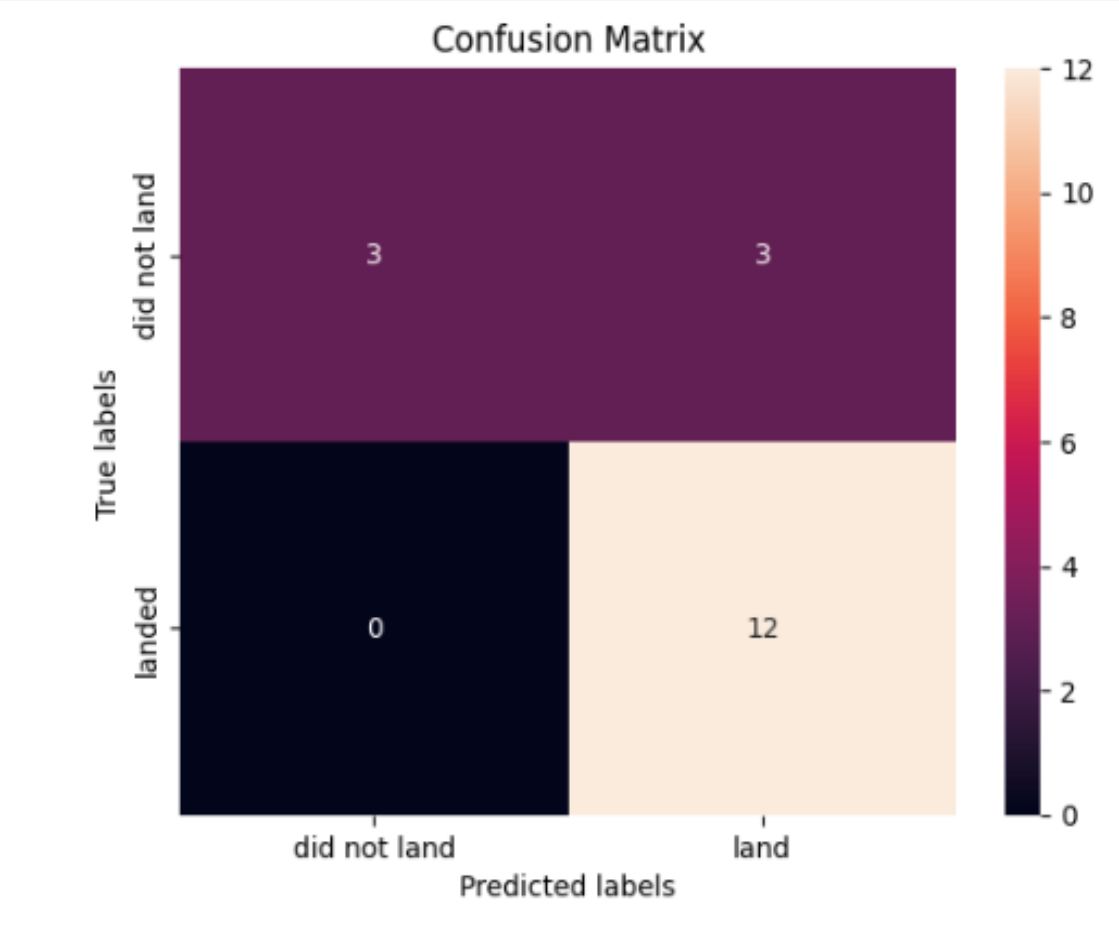
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.800000	0.819444
F1_Score	0.909091	0.916031	0.888889	0.900763
Accuracy	0.866667	0.877778	0.844444	0.855556



As seen from the comparison score tables/bar chart, we are unable to determine the best model for this project due to having obtained very similar scores pertaining to Logistic Regression, SVM and KNN.

However, SVM emerges as the best model for this project when the whole data set is taken into account.

Confusion Matrix



The confusion matrix is an effective method to check whether the models can differentiate between the various classes.

All the modes considered for this report had the same confusion matrix.

Scrutinizing the confusion matrix, we can highlight the 3 false positives (predicted as landed but not actually landed). This is concerning and needs to be further investigated.

Conclusions

- The optimal algorithm for this dataset is SVM
- Majority of launch sites are located close to the Equator and coastal areas
- The success rate of launches rises with time
- The highest success rate among all launch sites is KSC LC-39A
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

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!