IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Evgenii Merkushin
01.12.2024

Outline

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

Executive Summary

Summary of methodologies

- SpaceX rocket launch data is collected using SpaceX API and web scraping with Requests and BeutifulSoup Python's packages
- Work with missing values and categorical independent variables (Launch sites, Orbits, Landing outcomes) and create a label with `1` means the booster successfully landed `0` means it was unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Split the data into Train and Test sets, create four classification models: Logistic Regression, SVM, Classification Trees and KNN, tune the best Hyperparameter for each model, evaluate each model based with confusion matrices and find the method performs best using test data

• Summary of all results

The best classification model derived from Jaccard index and 20% Test set to predict success of First Stage Landing is tree. Jaccard index for this model is 92.3%

Introduction

- Project background and context

SpaceX is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010.

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas 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

Therefore, if we can determine 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 SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

SpaceX rocket launch data is collected using SpaceX API and web scraping with Requests and BeutifulSoup Python's packages

- Perform data wrangling

Work with missing values and categorical independent variables (Launch sites, Orbits, Landing outcomes) and create a label with `1` means the booster successfully landed `0` means it was unsuccessful

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Split the data into Train and Test sets, create four classification models: Logistic Regression, SVM, Classification Trees and KNN, tune the best Hyperparameter for each model, evaluate each model based with confusion matrices and find the method performs best using test data

Data Collection

SpaceX API

Request and parse the SpaceX launch data using the GET request

Filter the data frame to only include `Falcon 9` launches

Dealing with Missing Values

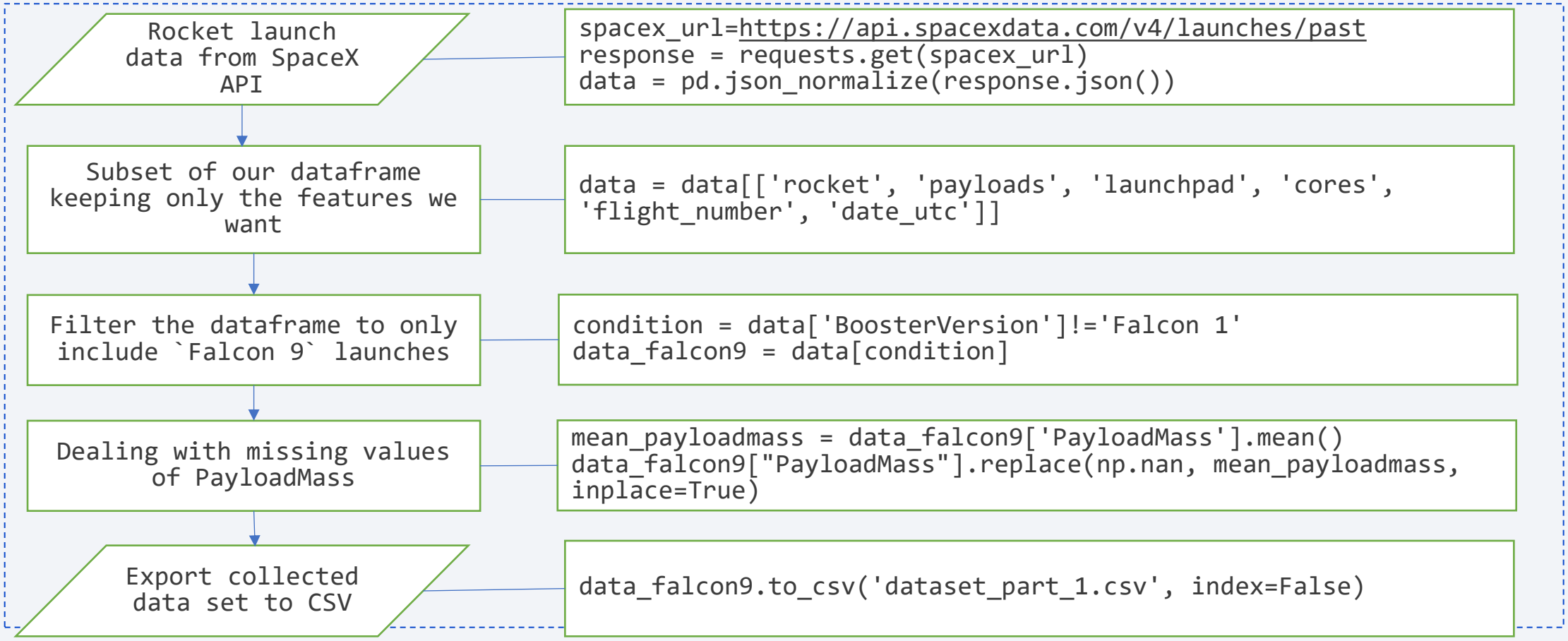
Web Scraping

Request the Falcon9 Launch Wiki page from its URL

Extract all column/variable names from the HTML table header

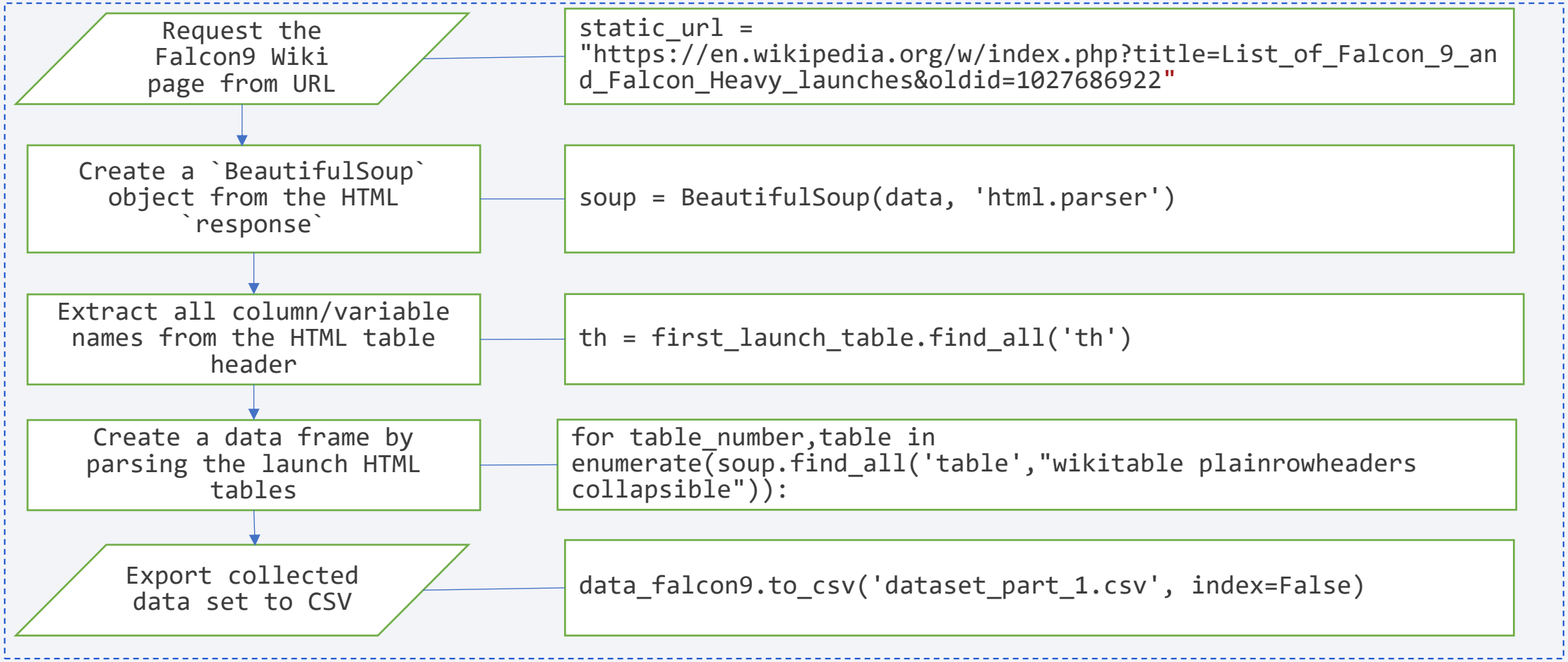
Create a data frame by parsing the launch HTML tables

Data Collection – SpaceX API



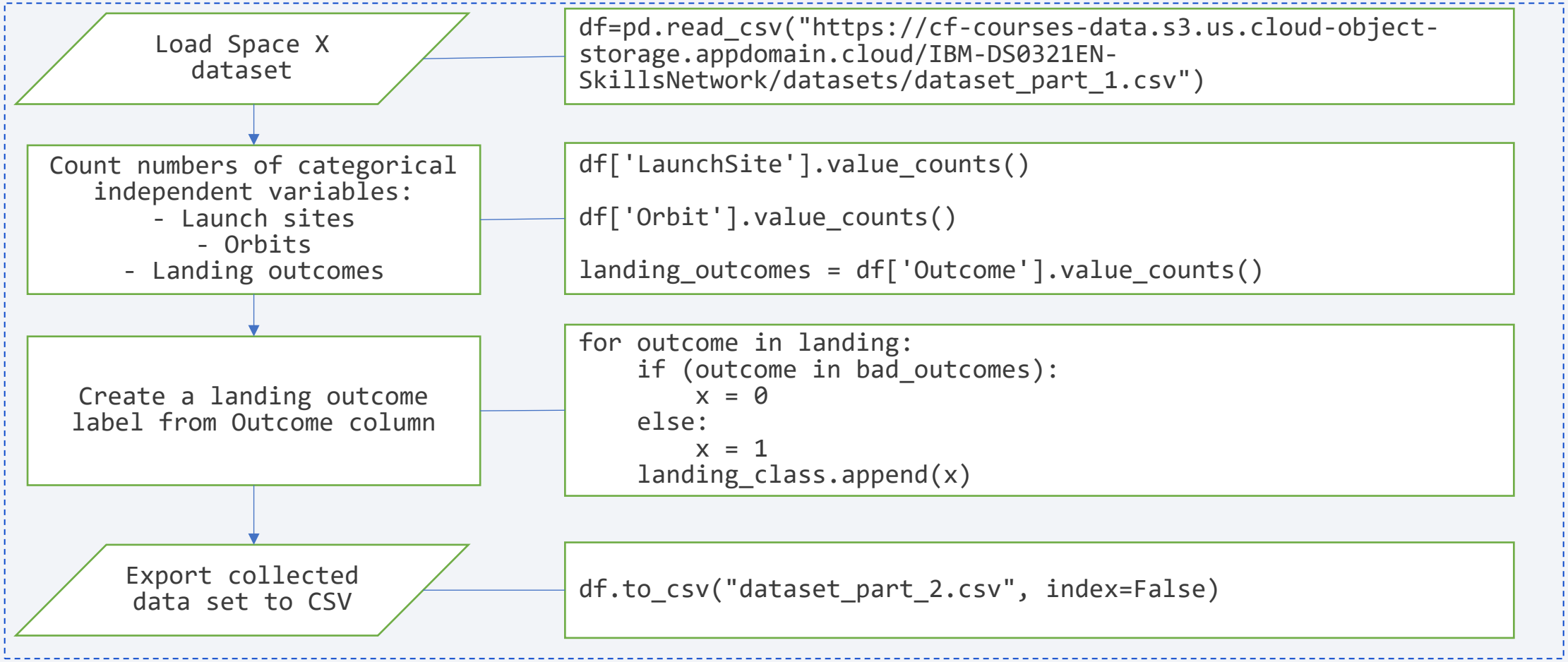
GitHub URL: https://github.com/emerkushin/testrepo/blob/main/collect.ipynb

Data Collection - Scraping



GitHub URL: https://github.com/emerkushin/testrepo/blob/main/webscrap.ipynb

Data Wrangling



GitHub URL: https://github.com/emerkushin/testrepo/blob/main/wrangle.ipynb

EDA with Data Visualization

- Visualize the relationship between Flight Number and Launch Site
- Visualize the relationship between Payload Mass and Launch Site
- Visualize the relationship between success rate of each orbit type
- Visualize the relationship between FlightNumber and Orbit type
- Visualize the relationship between Payload Mass and Orbit type
- Visualize the launch success yearly trend

GitHub URL: https://github.com/emerkushin/testrepo/blob/main/edadataviz.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- 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
- 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

GitHub URL: https://github.com/emerkushin/testrepo/blob/main/eda_sql.ipynb

Build an Interactive Map with Folium

The following map objects were created and added to a folium map:

- `folium.Circle` to add a highlighted circle area with a text label on a specific coordinate
- `folium.Marker` to mark launch sites
- `MarkerCluster` to mark the success/failed launches for each site on the map
- `MousePosition` on the map to get coordinate for a mouse over a point on the map
- `folium.PolyLine` to indicate the distance between a launch site to the selected coastline point

GitHub URL: https://github.com/emerkushin/testrepo/blob/main/folium.ipynb

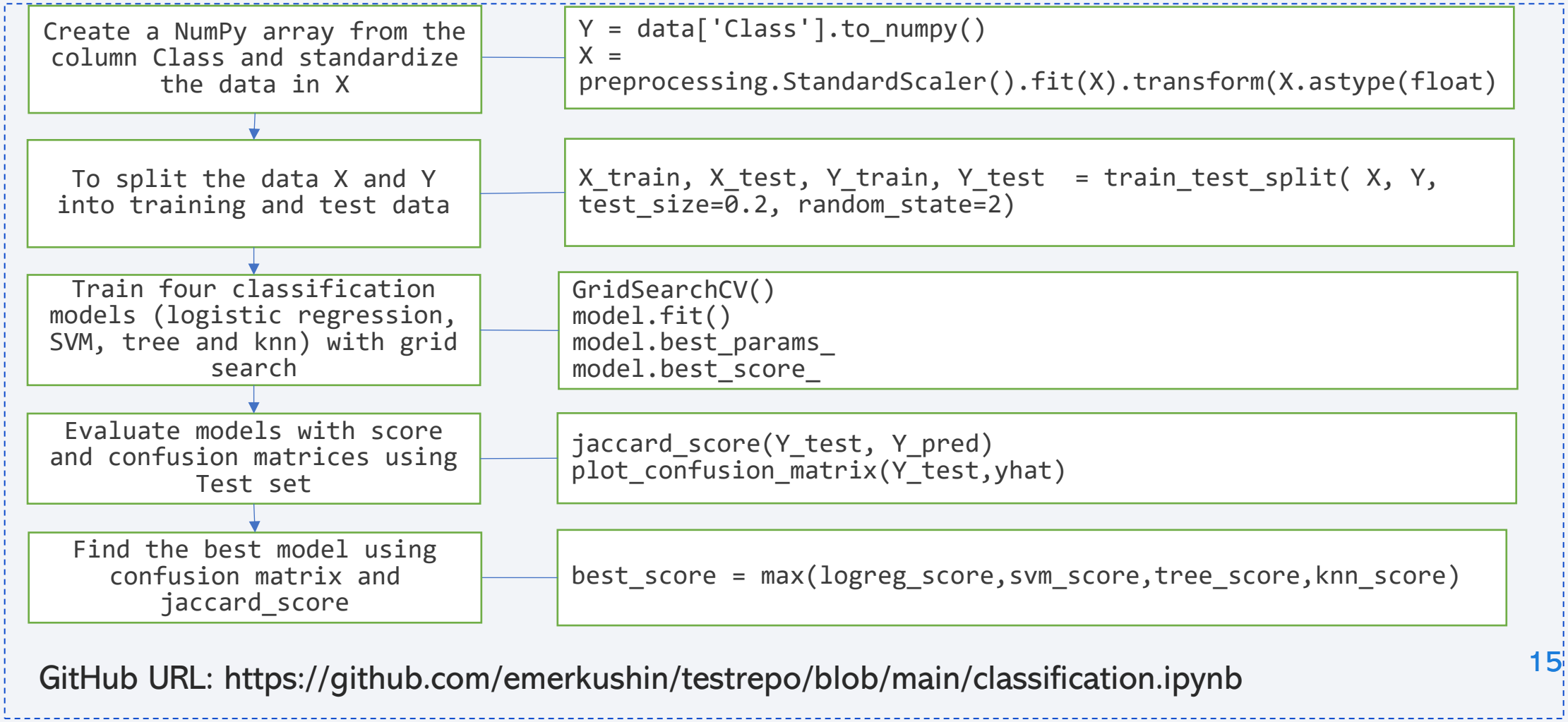
Build a Dashboard with Plotly Dash

The following plots/graphs and interactions were added to a dashboard:

- a pie chart to show the total successful launches count for all sites
- pie charts to display the Success vs. Failed counts for the specific launch site
- a slider to select payload range
- scatter charts to show the correlation between payload and launch success for the specific launch site

GitHub URL: https://github.com/emerkushin/testrepo/blob/main/lab_dash.py

Predictive Analysis (Classification)



GitHub URL: https://github.com/emerkushin/testrepo/blob/main/classification.ipynb

Results

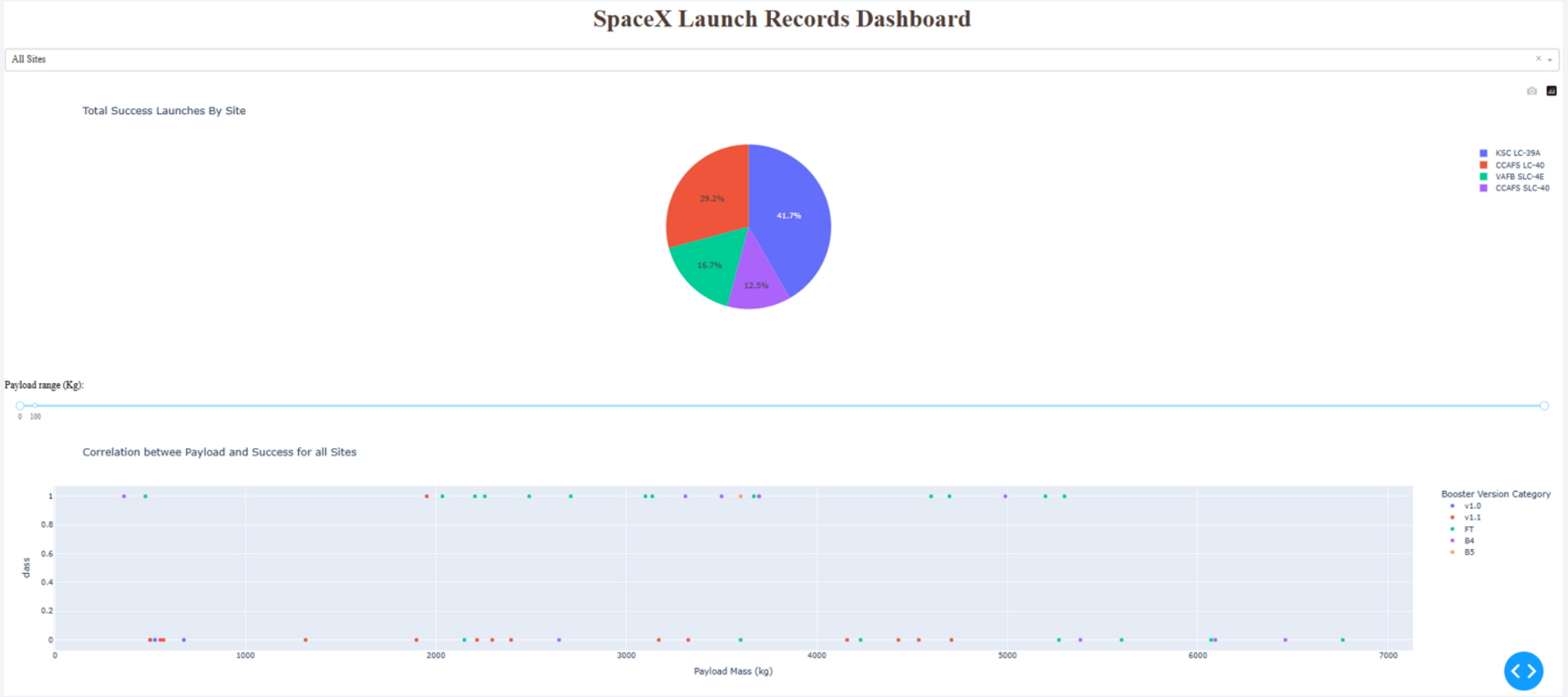
Exploratory data analysis results

We have chosen the following independent variables for our model:
'FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights',
'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount',
'Serial’

We have converted categorical variables Orbits, LaunchSite,
LandingPad, and Serial into dummies variables.

Results

- Interactive analytics demo in screenshots

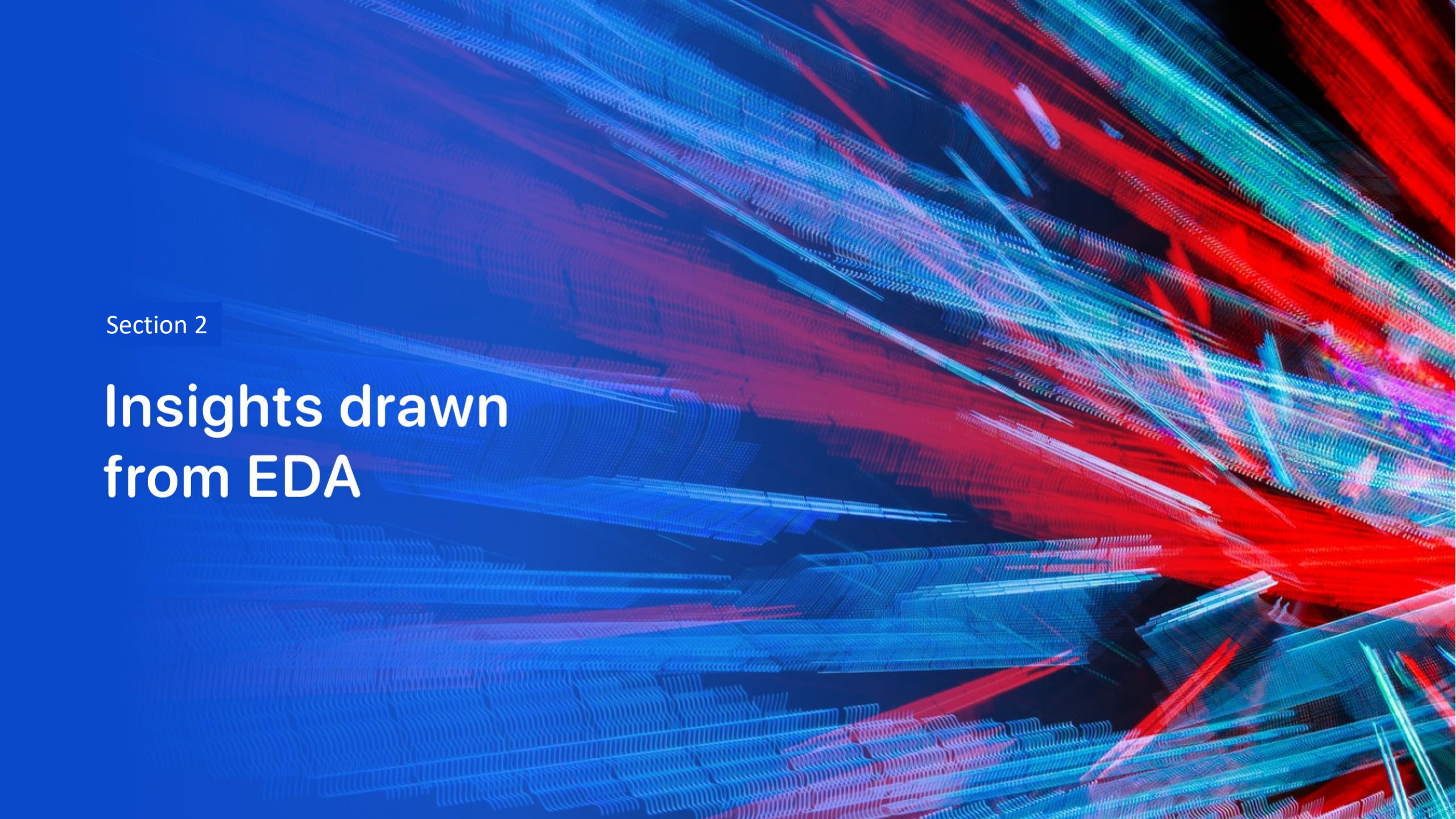


Results

- Predictive analysis results

Four models for classification prediction have been tested: logistic regression SVM, tree and knn.

The best model classification model chosen based on Jaccard index and Test set of 20% of data is tree.

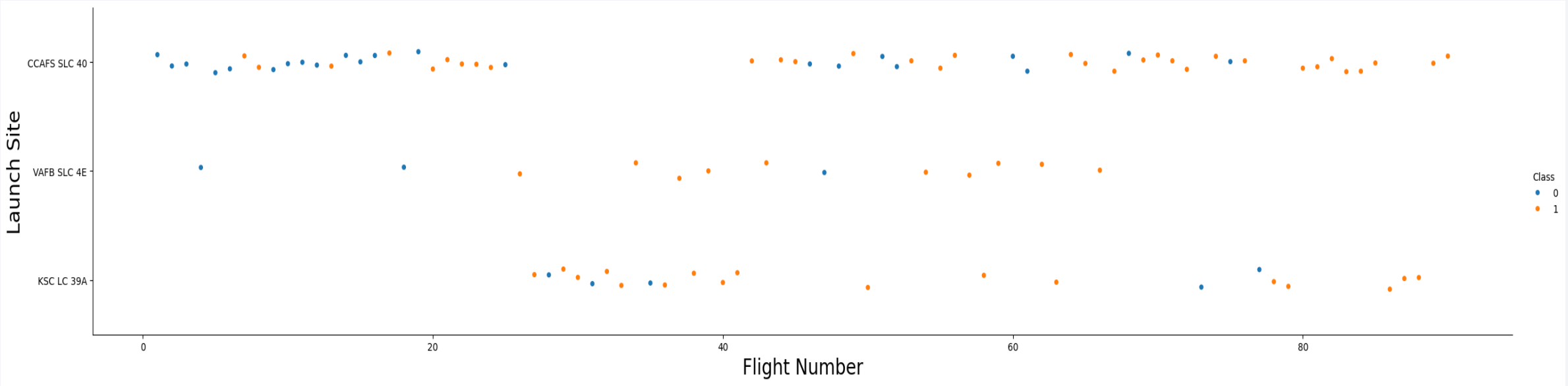

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site

```
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

- Show the screenshot of the scatter plot with explanations



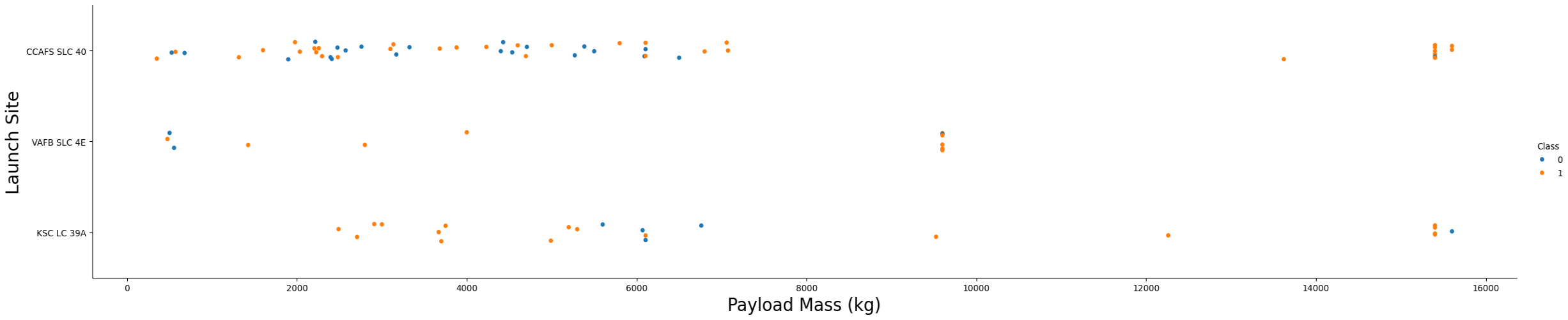
Takeaway: the biggest number of launches was from CCAFS SLC-40 launch site

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

```
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

- Show the screenshot of the scatter plot with explanations



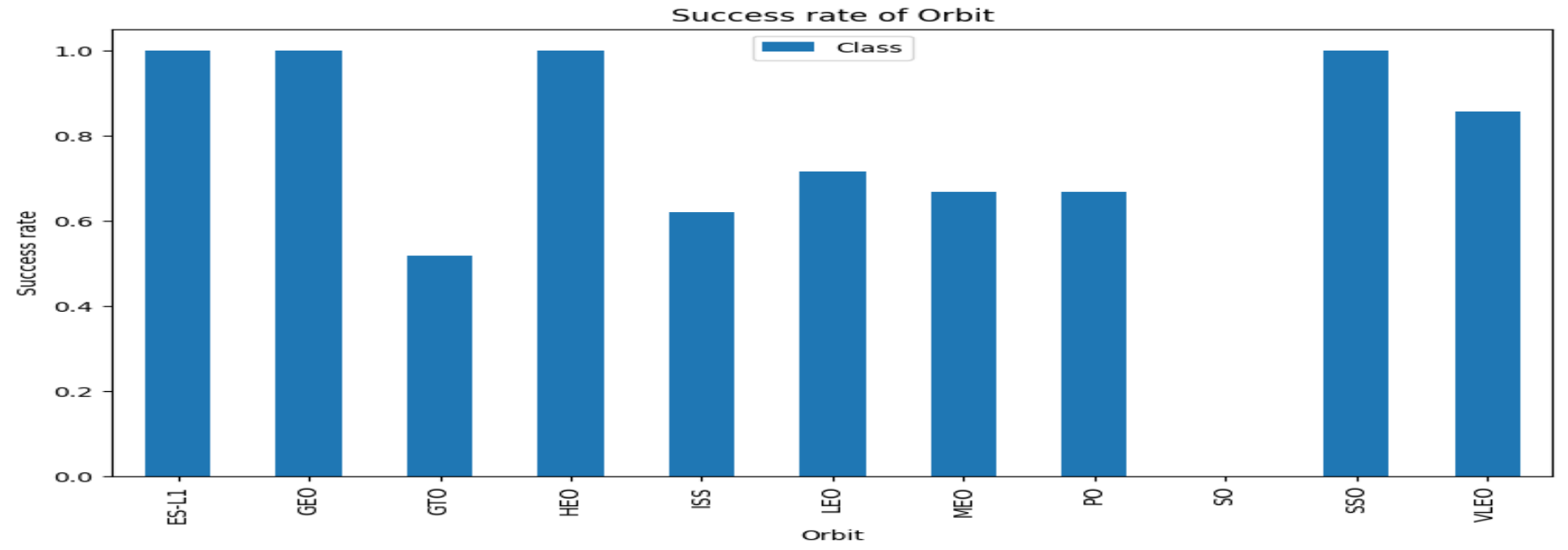
Takeaway: for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

```
df_grouped = df.groupby(['Orbit'], as_index=False).agg({'Class': 'mean'})
df_grouped.set_index('Orbit', inplace=True)
df_grouped.plot(kind='bar', figsize=(10, 6), rot=90)
plt.xlabel('Orbit')
plt.ylabel('Success rate')
plt.title('Success rate of Orbit')
plt.show()
```

- Show the screenshot of the bar chart with explanations



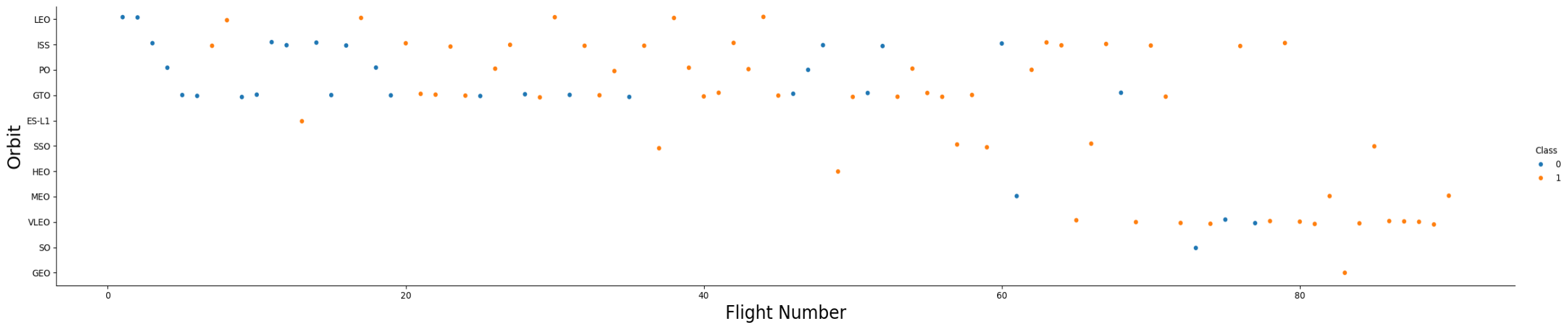
Takeaway: ES-L1, GEO, HEO and SSO orbits have the highest success rates (100%)

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

```
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

- Show the screenshot of the scatter plot with explanations



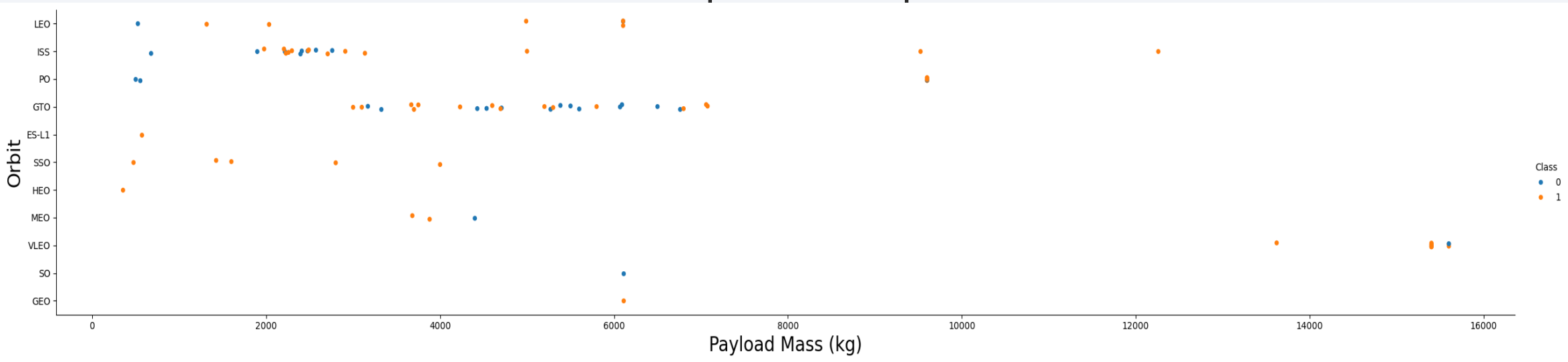
Takeaway: you can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type

- Show a scatter plot of payload vs. orbit type

```
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass (kg)",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

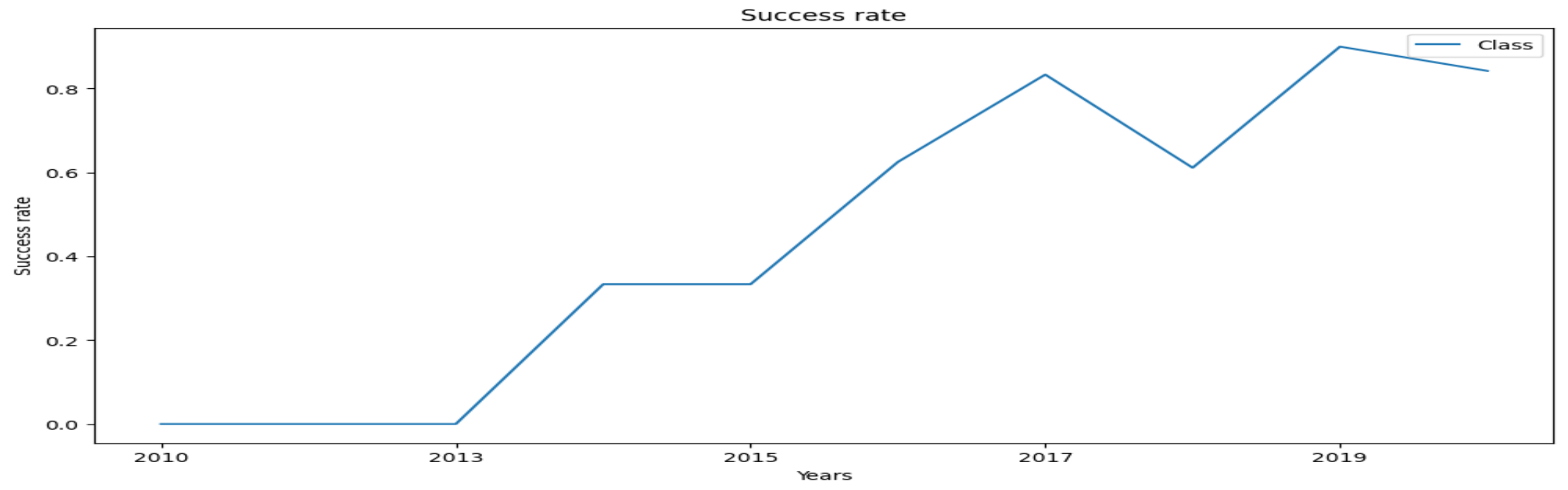
- Show the screenshot of the scatter plot with explanations



Takeaway: with heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



Takeaway: the success rate since 2013 kept increasing till 2020.

All Launch Site Names

- Find the names of the unique launch sites

```
%sql select DISTINCT Launch_Site from SPACEXTABLE;
```

- Present your query result with a short explanation here

```
We have found 4 sites
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
%sql select * from SPACEXTABLE where \
Launch_Site LIKE 'CCA%' LIMIT 5;
```

- Present your query result with a short explanation here

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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where \
Customer = 'NASA (CRS)';
```

- Present your query result with a short explanation here

The result is 45,596 kg

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where \
Booster_Version = 'F9 v1.1';
```

- Present your query result with a short explanation here

The result is 2,928.4 kg

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql select MIN(Date) from SPACEXTABLE where \
Landing_Outcome  = 'Success (ground pad)';
```

- Present your query result with a short explanation here

```
The result is 2015-12-22
```

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select Booster_Version from SPACEXTABLE where \
Landing_Outcome  = 'Success (drone ship)' AND \
PAYLOAD_MASS__KG_ > 4000 AND \
PAYLOAD_MASS__KG_ < 6000;
```

- Present your query result with a short explanation here

The result is 4 booster versions.

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql select COUNT(Mission_Outcome) from SPACEXTABLE where \
Mission_Outcome LIKE 'Success%';
%sql select COUNT(Mission_Outcome) from SPACEXTABLE where \
Mission_Outcome LIKE 'Failure%';
```

- Present your query result with a short explanation here

The result is 100 successful and 1 failure mission outcomes.

COUNT(Mission_Outcome)
100

COUNT(Mission_Outcome)
1

Boosters Carried Maximum Payload

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

```
%sql select Booster_Version from SPACEXTABLE where \
    PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

- Present your query result with a short explanation here

The result is the list of 12 boosters.

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

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

```
%sql select substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTABLE where \
    substr(Date,0,5)='2015' AND \
    Landing_Outcome = 'Failure (drone ship)';
```

- Present your query result with a short explanation here

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

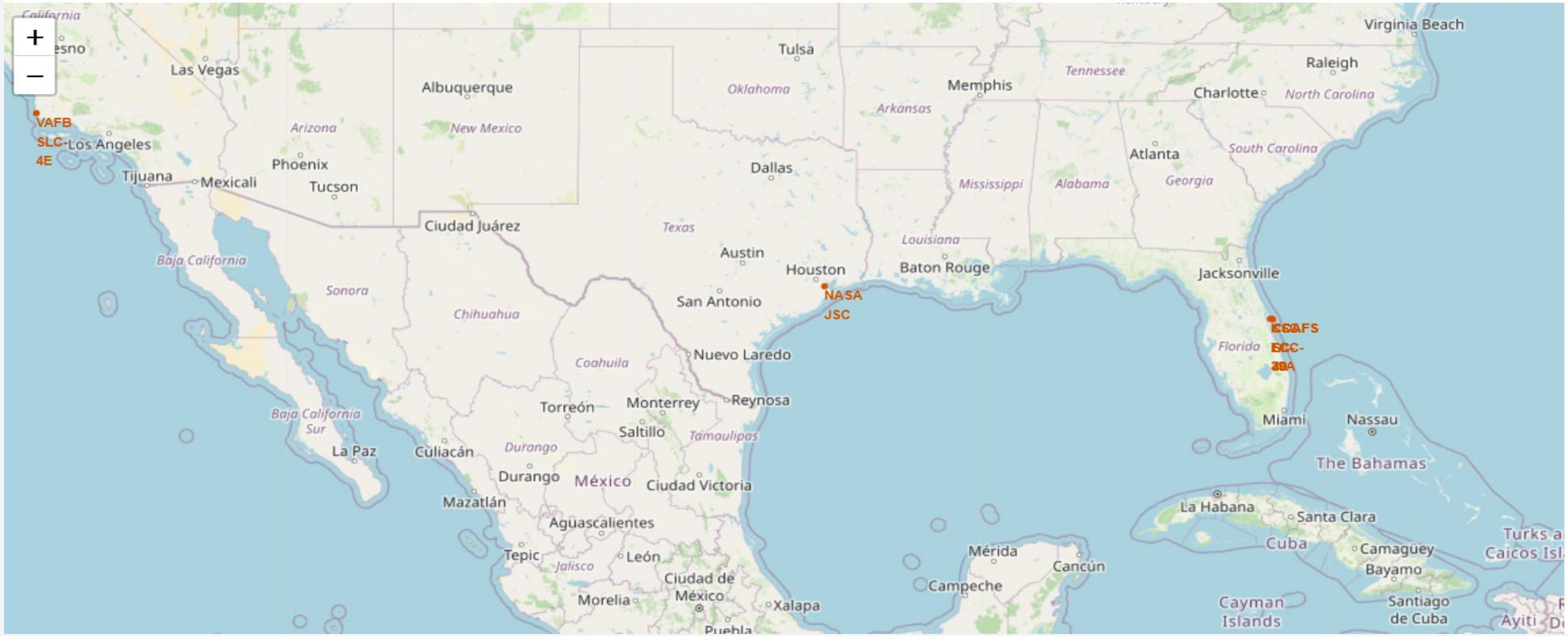
```
%sql select Landing_Outcome, COUNT(Landing_Outcome) \
from SPACEXTABLE where\
Date BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY Landing_Outcome \
ORDER BY COUNT(Landing_Outcome) DESC;
```

Landing_Outcome	COUNT(Landing_Outcome)
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

Section 3

Launch Sites Proximities Analysis

All launch sites' locations



There are two main launch sites for Space X: VAFB SLC-4E in California and 3 close locations in Florida (CCAFC LC-40, KSC LC-39A, CCAFC SLC-40)

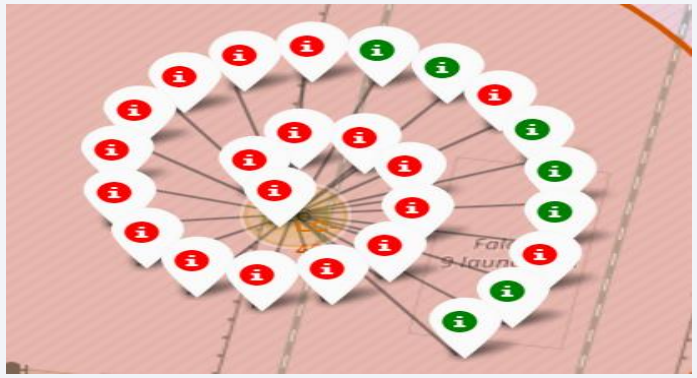
The success/failed launches for each site on the map

The success launch is marked in green, the failed in red.

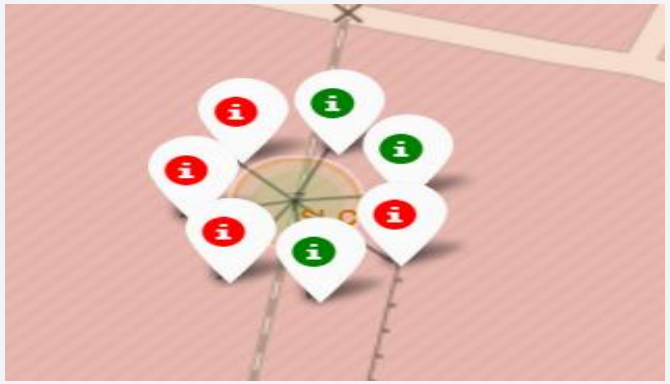
VAFB SLC-4E site in California

KSC LC-39A in Florida

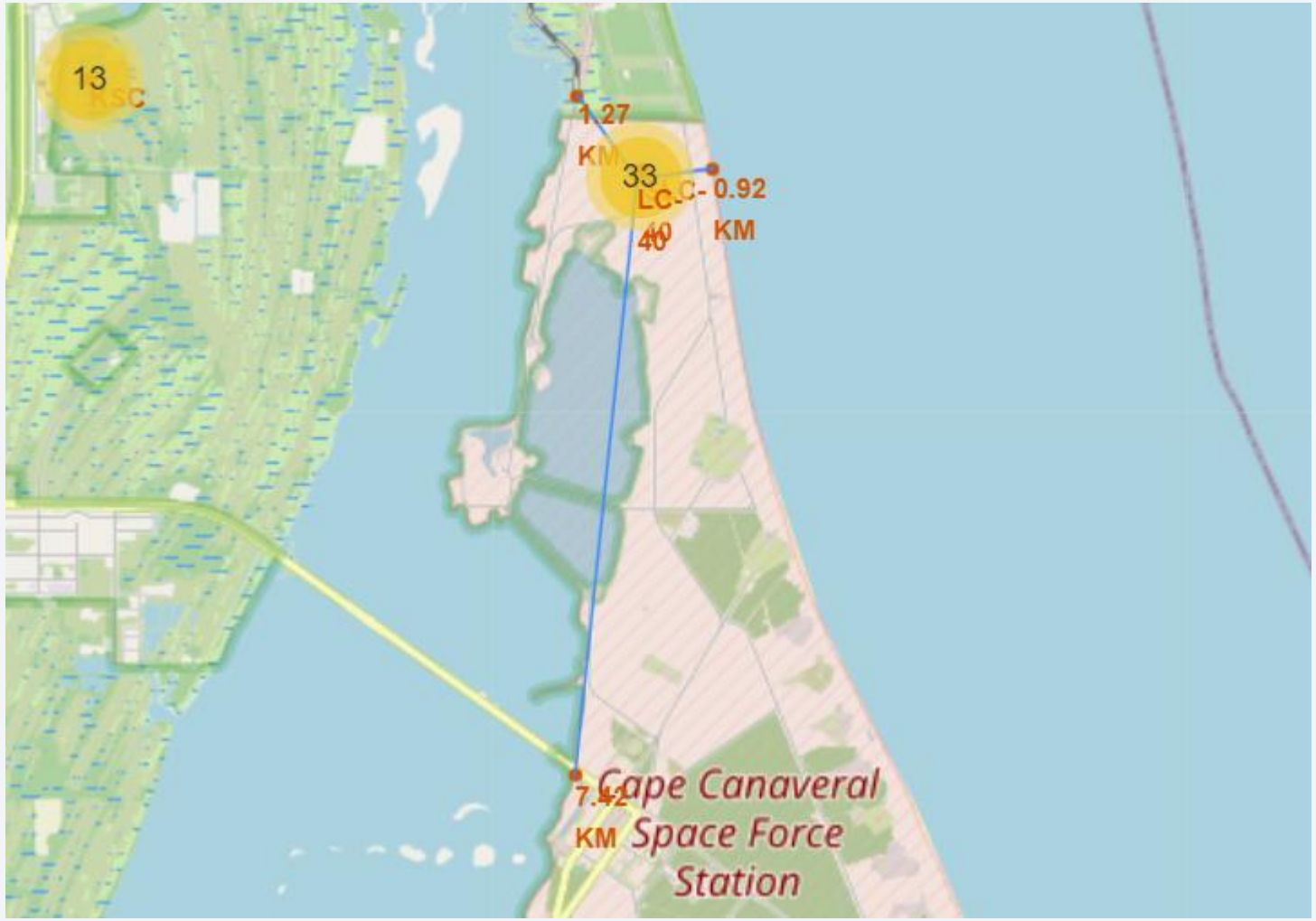
CCAFC LC-40 in Florida



CCAFC SLC-40in Florida

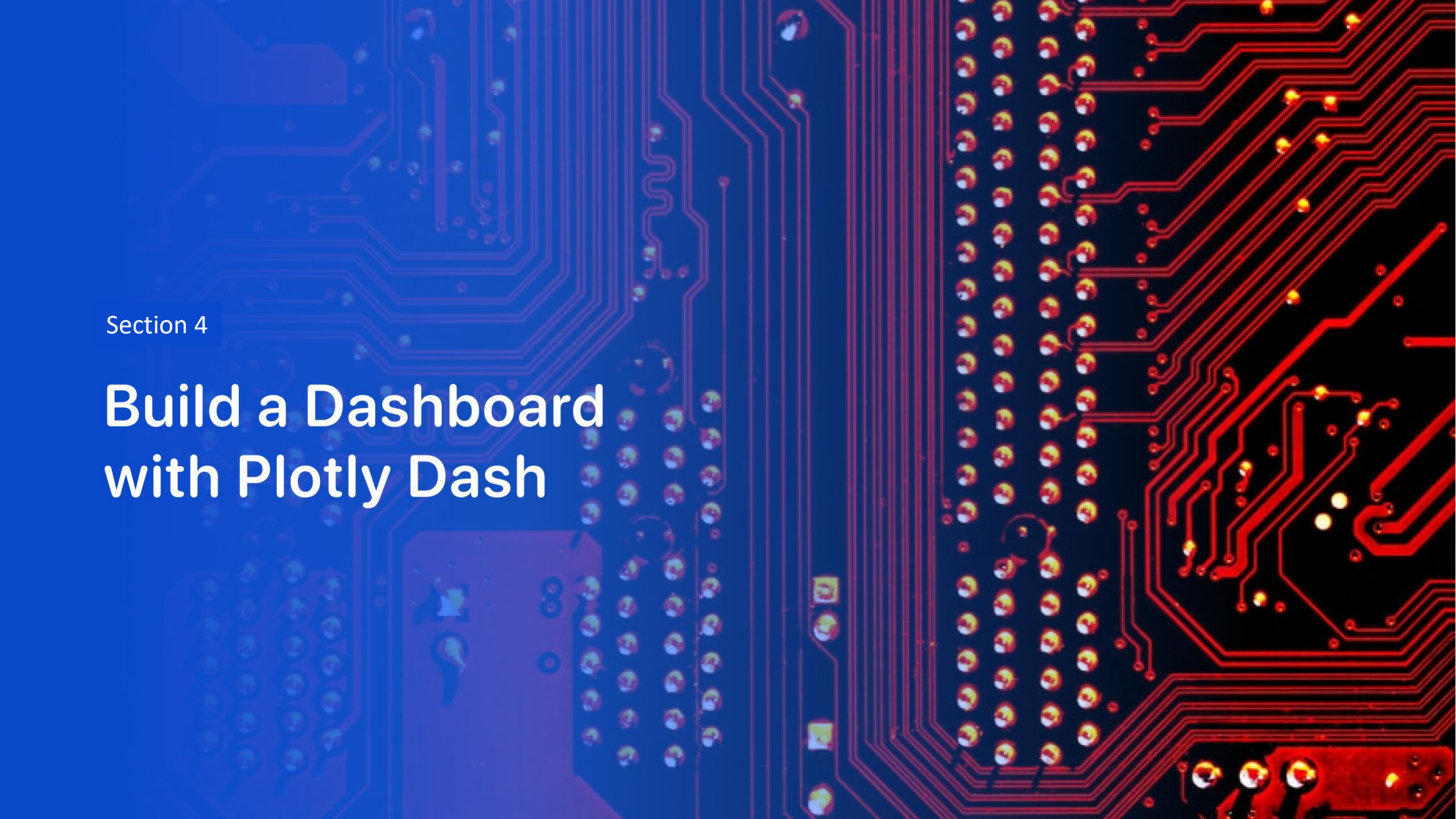


CCAFC LC-40 launch site to its proximities

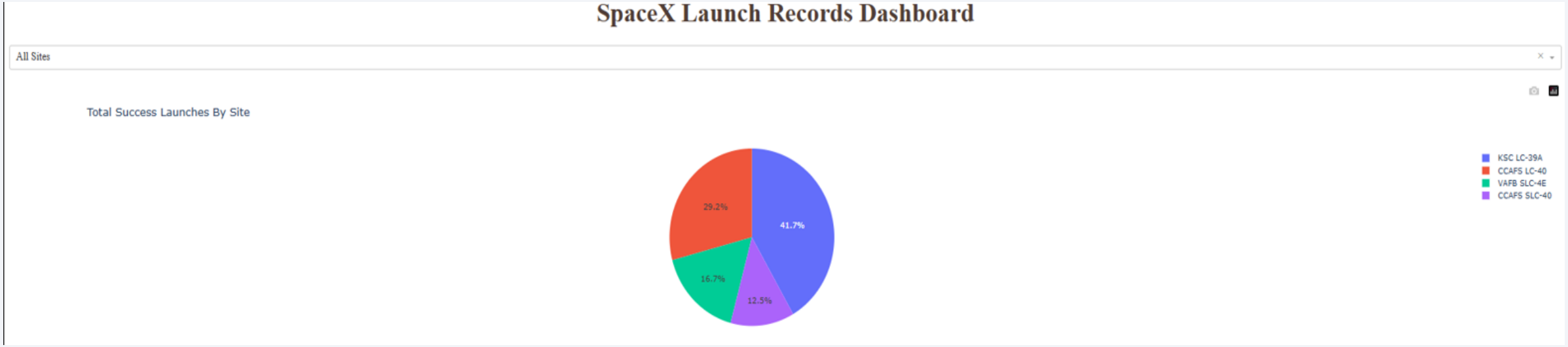


Distances from CCAFC LC-40 launch site:

- To coast: 0.92 km
- To railroad: 1.27 km
- To highway: 7.42 km

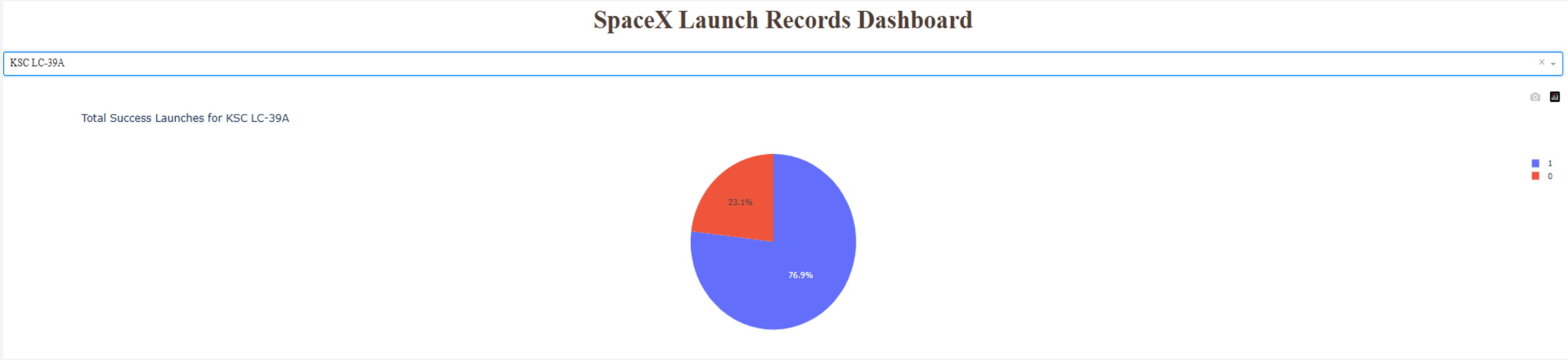

Section 4
Build a Dashboard with Plotly Dash

Total success launches by Site



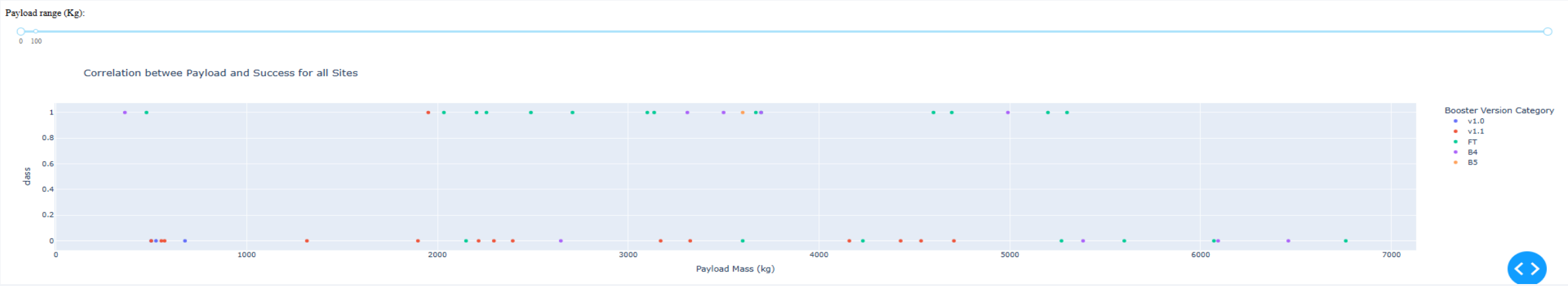
Takeaway: the biggest share of success launches belons to KSC LC-39A launch site

The launch site with highest launch success ratio

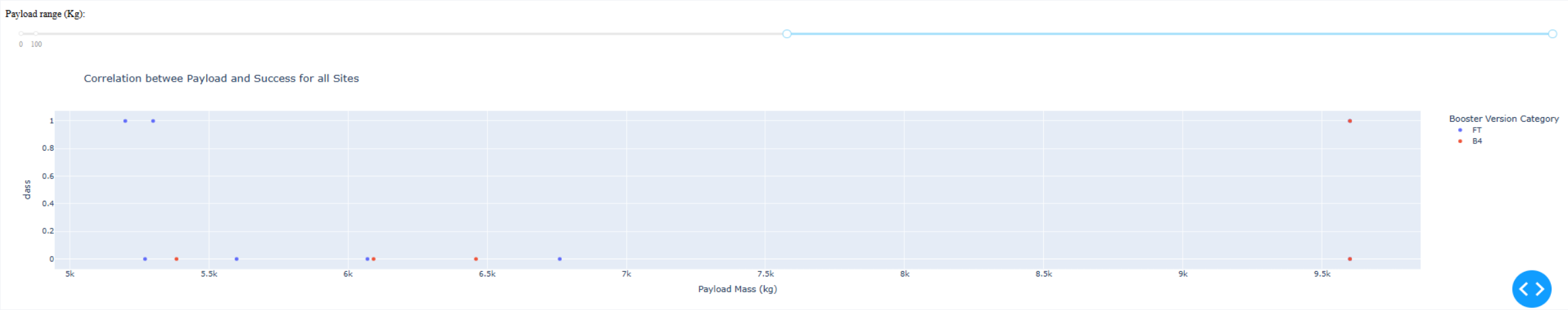


Takeaway: the launch site with highest launch success ratio (76.9%) is KSC LC-39A

Payload vs. Launch Outcome scatter plot for all sites



Takeaway for full payload range: FT booster has the highest number of success launches

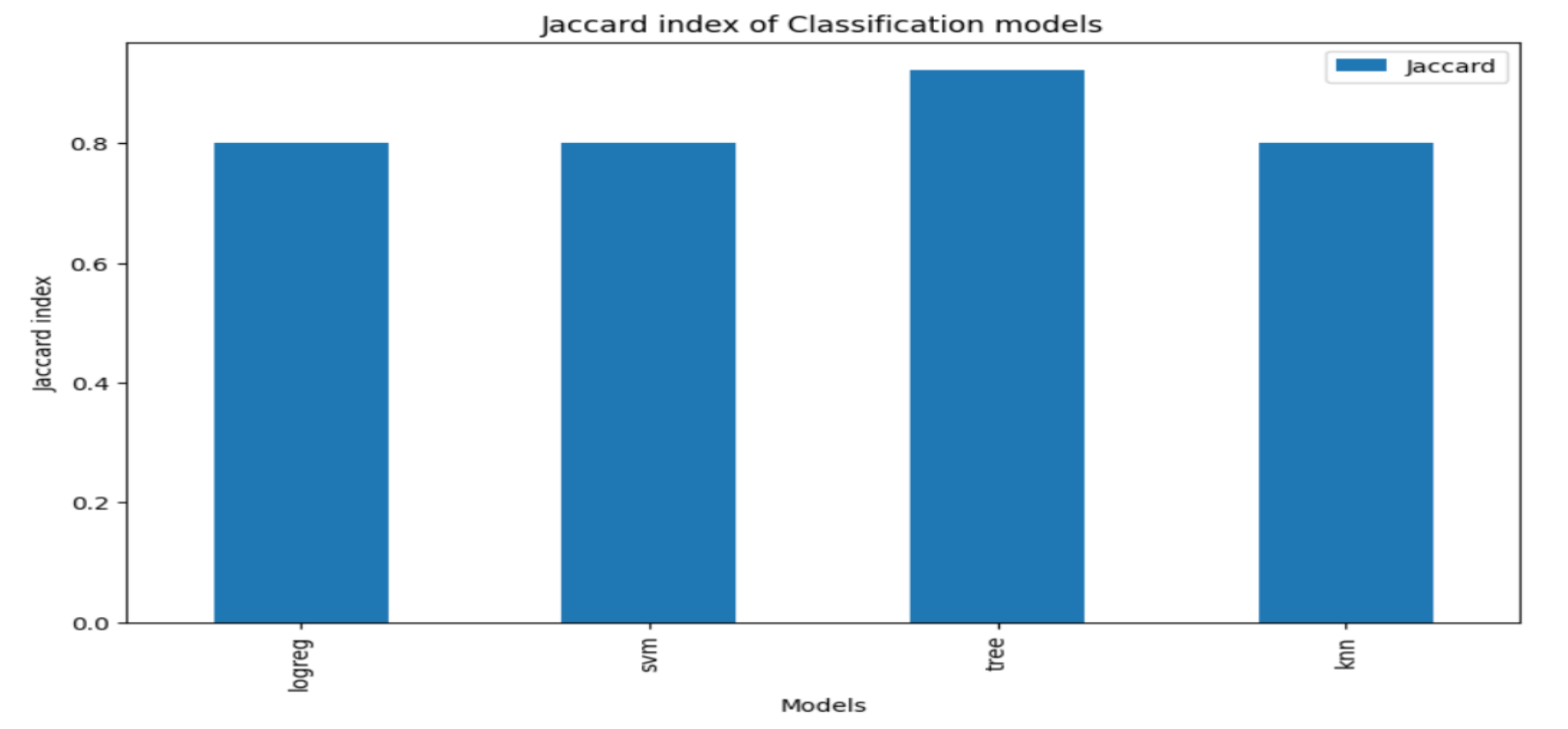


Takeaway for heavier payload range: the number of success launches is crucially diminished

Section 5

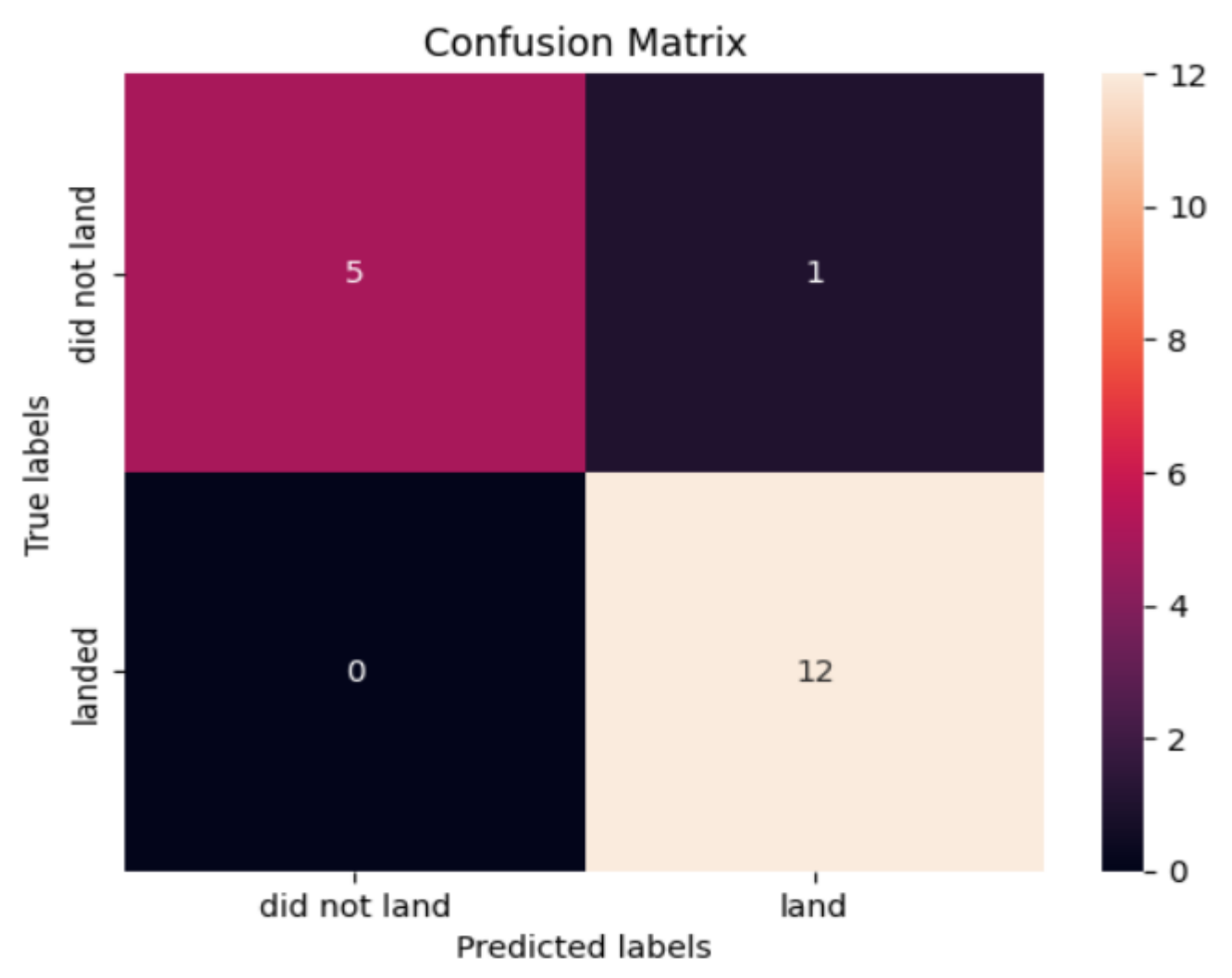
Predictive Analysis (Classification)

Classification Accuracy



Takeaway: the best classification model derived from Jaccard index using Test set to predict success of First Stage Landing is tree.

Confusion Matrix of Tree classification model



Tree classification model has predicted using the Test set:
- All 5 failures of the first stage landing
- 12 out of 13 successful first stage landing

Conclusions

- We have chosen the following independent variables for our model: 'FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial'
- We have converted categorical variables Orbits, LaunchSite, LandingPad, and Serial into dummies variables.
- Four models for classification prediction have been tested using the Test set: logistic regression SVM, tree and knn.
- The best classification model derived from Jaccard index and 20% Test set to predict success of First Stage Landing is tree. Jaccard index for this model is 92.3%
- Tuning parameters of the tree model are the following:

{'criterion': 'entropy', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'best'}

Appendix

https://github.com/emerkushin/testrepo/blob/main/collect.ipynb

https://github.com/emerkushin/testrepo/blob/main/webscrap.ipynb

https://github.com/emerkushin/testrepo/blob/main/wrangle.ipynb

https://github.com/emerkushin/testrepo/blob/main/edadataviz.ipynb

https://github.com/emerkushin/testrepo/blob/main/eda_sql.ipynb

https://github.com/emerkushin/testrepo/blob/main/folium.ipynb

https://github.com/emerkushin/testrepo/blob/main/lab_dash.py

https://github.com/emerkushin/testrepo/blob/main/classification.ipynb

Thank you!