IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zakaria Hassan Mohamed

30/11/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection via API & Web Scraping
  - Exploratory Data Analysis (EDA) with Data Visualization
  - EDA with SQL
  - Interactive Visual Analytics with Folium
  - Dashboards with Plotly Dash
  - ML Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive maps and dashboard
  - Predictive results

Introduction

- Project background and context
  - The objective of this project is to predict whether the first stage of the Falcon 9 will land successfully. The ability to land the first stage, saving them upwards to 100 million dollars, is what sets SpaceX apart from their competition. Therefore, we can calculate the cost of a launch by determining if the first stage will land, we can determine the cost of a launch.
- Problems we want to solve
  - What are the main characteristics of successful or failed landings ?
  - What are the effects of each relationship of the rocket variables on the success or failure of a landing ?
  - What are the conditions which will allow SpaceX to achieve the best landing success rate ?

Section 1

Methodology

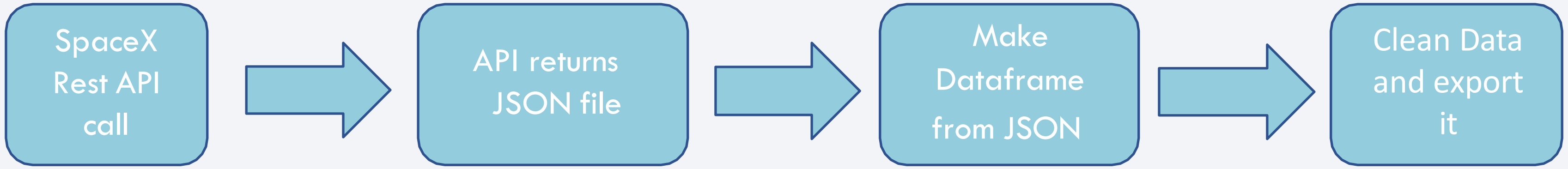
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - One Hot Encoding for classification models
  - Dropping unnecessary 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
  - Knn, Trees, Logistic Regression, SVM

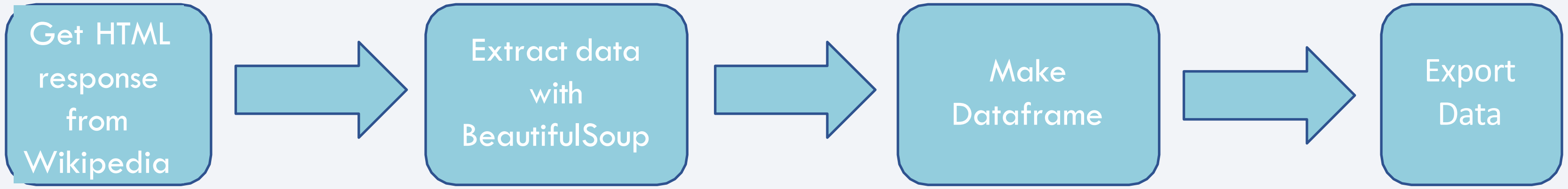
Data Collection

- Datasets are collected from Rest SpaceX API and via web scraping Wikipedia
  - The information obtained by the API are rockets, launches and payload information.
    - The Space X REST API URL is api.spacexdata.com/v4/



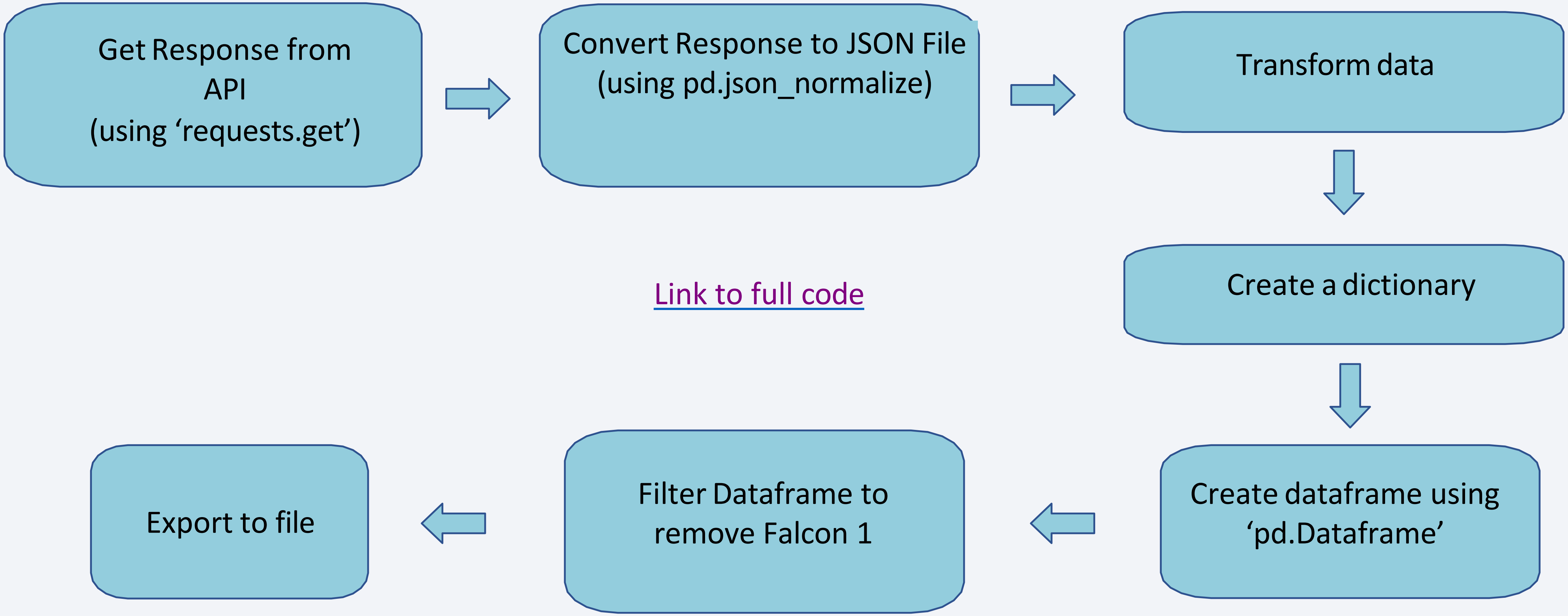
SpaceX Rest API call → API returns JSON file → Make Dataframe from JSON → Clean Data and export it

  - The information obtained by the webscrapping of Wikipedia are launches, landing, payload information.
    - URL is https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922



Get HTML response from Wikipedia → Extract data with BeautifulSoup → Make Dataframe → Export Data

Data Collection - SpaceX API



Get Response from API (using 'requests.get')

Convert Response to JSON File (using pd.json_normalize)

Transform data

Create a dictionary

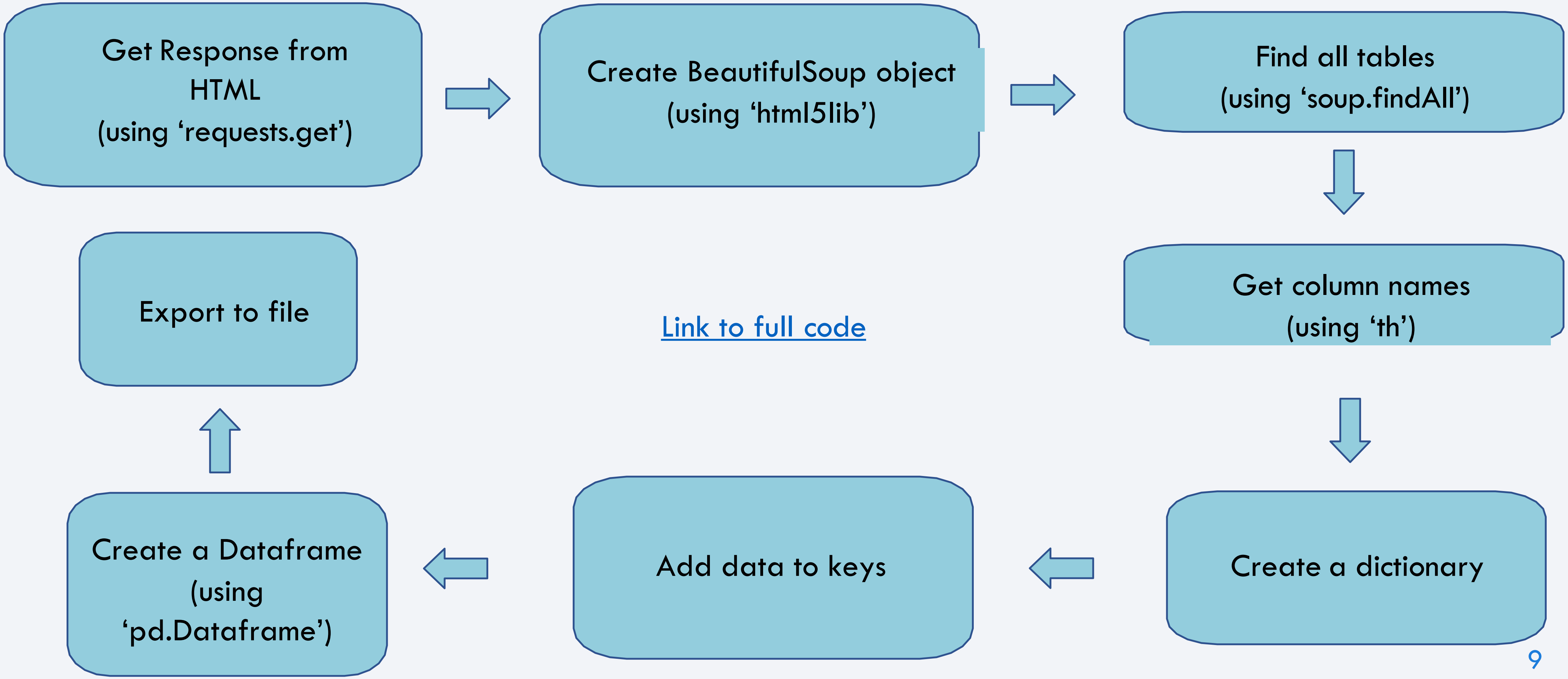
Create dataframe using 'pd.Dataframe'

Filter Dataframe to remove Falcon 1

Export to file

Link to full code

Data Collection - Web Scraping



Get Response from HTML (using ‘requests.get’)

Create BeautifulSoup object (using ‘html5lib’)

Find all tables (using ‘soup.findAll’)

Get column names (using ‘th’)

Create a dictionary

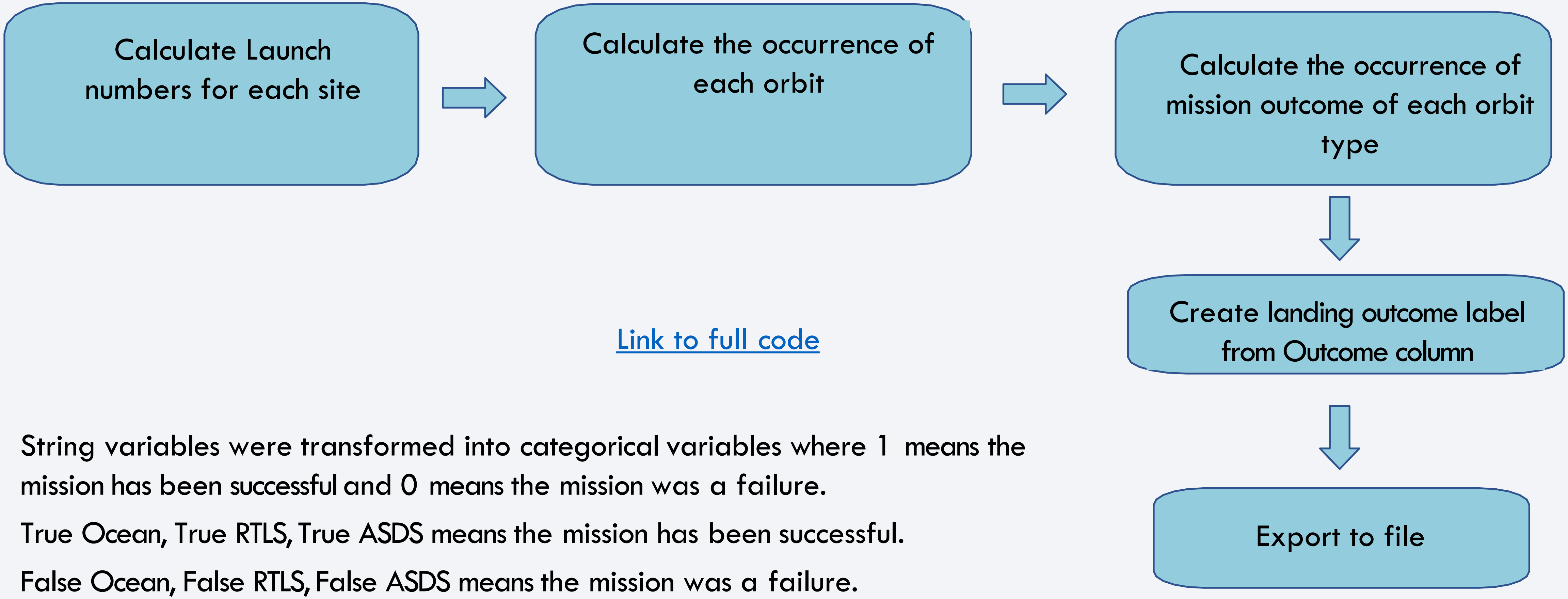
Add data to keys

Create a Dataframe (using ‘pd.Dataframe’)

Export to file

[Link to full code]()

Data Wrangling



Calculate Launch numbers for each site
Calculate the occurrence of each orbit
Calculate the occurrence of mission outcome of each orbit type
Create landing outcome label from Outcome column
Export to file

[Link to full code]()

- String variables were transformed into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure.
- True Ocean, True RTLS, True ASDS means the mission has been successful.
- False Ocean, False RTLS, False ASDS means the mission was a failure.

EDA with Data Visualization

- Scatter Graphs
  - Scatter graphs were used to show the correlation between different variables and helps us to discover relationships between the different variables.

- Line Graph
  - *Line graphs were used to show the trends of the success rate through the years.*
  - *Line graphs can help visualise patterns and make predictions for unseen data.*

- Bar Chart
  - *Bar charts were used to show the relationship between the success rate and the different orbit types.*

.

Link to full code

EDA with SQL

- These are the following SQL queries that were performed on the dataset:
  - Displaying the names of the unique lauunch 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 successful landing outcome in ground pad was achieved.
  - 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.
  - List the records which will display the month names, faiilure landing_ouutcomes in drone ship, booster versions, launch_site for the months in year 2015.
  - Rank the count of successful landiing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Link to full code

Build an Interactive Map with Folium

- Folium map object is a map centered on NASA Johnson Space Center at Houson, Texas
  - Red circle at NASA Johnson Space Center's coordinate with label showing its name *(folium.Circle, folium.map.Marker).*
  - Red circles at each launch site coordinates with label showing launch site name *(folium.Circle, folium.map.Marker, folium.features.DivIcon).*
  - The grouping of points in a cluster to display multiple and different information for the same coordinates *(folium.plugins.MarkerCluster).*
  - Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing. *(folium.map.Marker, folium.Icon).*
  - Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them. *(folium.map.Marker, folium.PolyLine, folium.features.DivIcon)*
- These objects were created in order to understand better the problem and the data. We can show easily all launch sites, their surroundings and the number of successful and unsuccessful landings.

Link to full code

Build a Dashboard with Plotly Dash

- The dashboard is of Spaxex Launch Records and it has a dropdown, a pie chart, scatter plots and a rangeslider
  - Dropdown allows a user to choose the launch site or all launch sites.
  - Pie chart shows the total success and the total failure for the launch site chosen with the dropdown component.
  - Scatter chart shows the relationship between two variables, particularly Success vs Payload Mass.
  - Rangeslider allows a user to select a payload mass for a specified range.

Link to full code

Predictive Analysis (Classification)

- Data preparation
  - Load dataset
  - Normalize data
  - Split data into training and test sets.
- Model preparation
  - Selection of machine learning algorithms
  - Set parameters for each algorithm to GridSearchCV
  - Training GridSearchModel models with training dataset
- Model evaluation
  - Get best hyperparameters for each type of model
  - Compute accuracy for each model with test dataset
  - Plot Confusion Matrix
- Model comparison
  - Comparison of models according to their accuracy
  - The model with the best accuracy will be chosen (see Notebook for result)

Link to full code

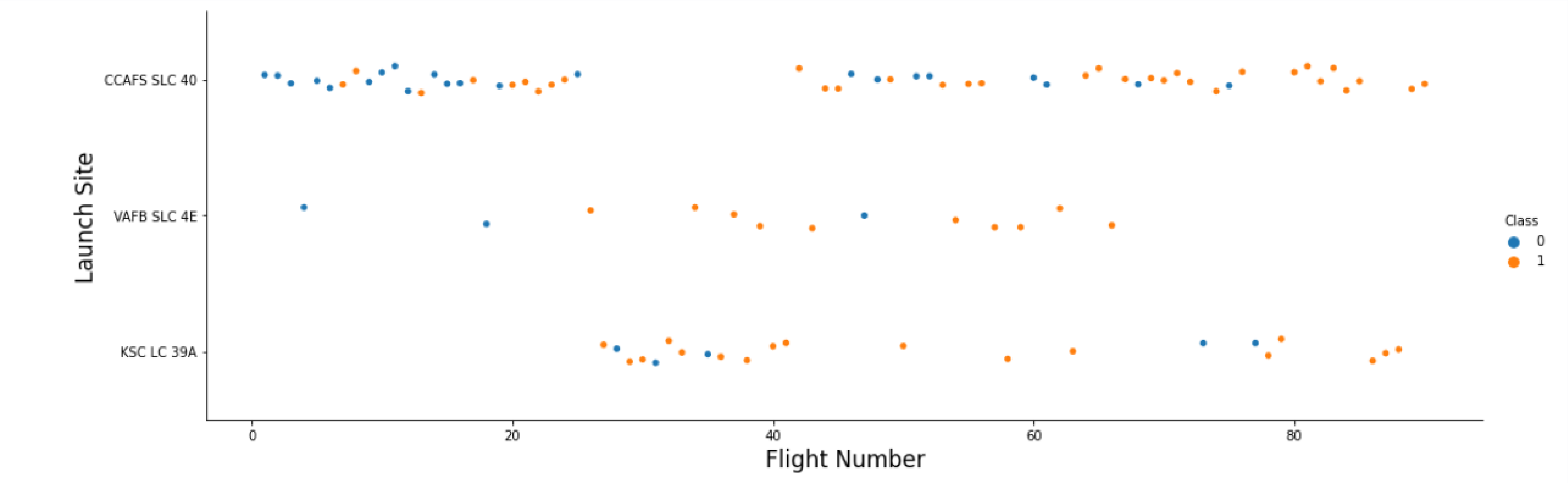
Results

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

Section 2

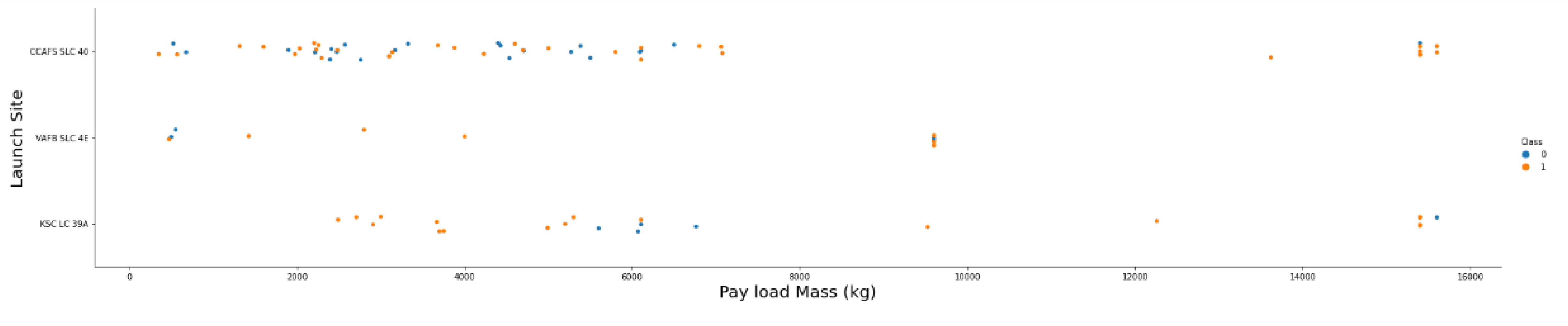
Insights drawn from EDA

Flight Number vs. Launch Site



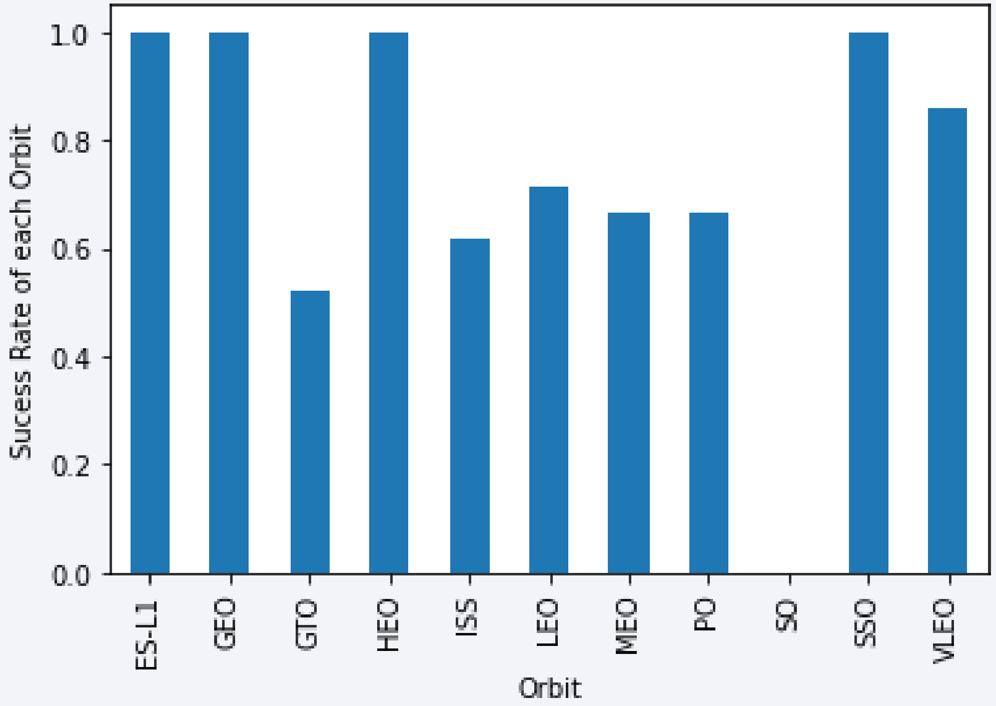
- The 0 (blue) represents a failed launch and the 1 (orange) represents a successful launch.
- From here we can see that as the flight number increases, the success rate increase for all launch sites.

Payload vs. Launch Site



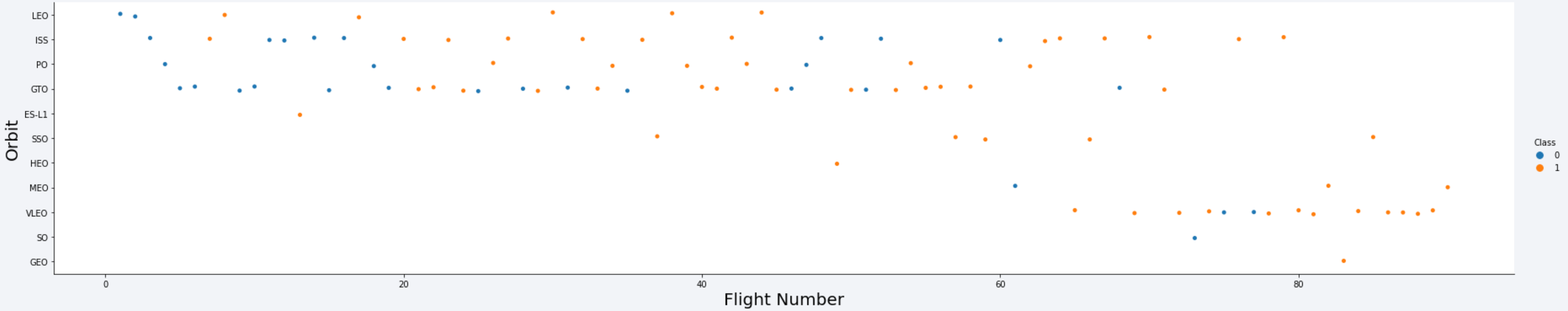
- The 0 (blue) represents a failed launch and the 1 (orange) represents a successful launch.
- From here it seems a heavier payload slightly correlate to a successful landing depending on the launch site.
- CCAFS SLC 40 performs best at a payload mass above 13000 kg than under 6000 kg, whereas KSC LC 39A performs best at under 4000kg.

Success Rate vs. Orbit Type



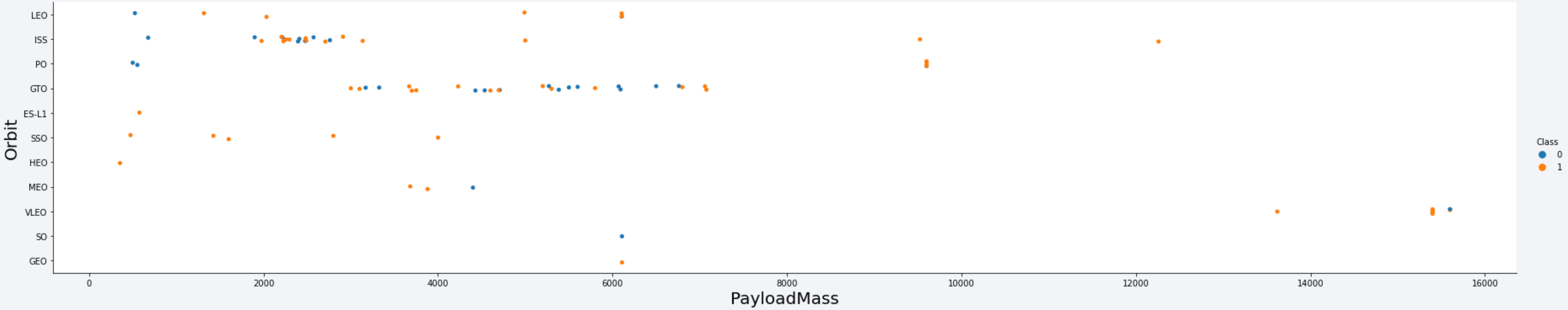
From this bar chart, we can see that ES-L1, GEO, HEO and SSO have the best success rate.

Flight Number vs. Orbit Type



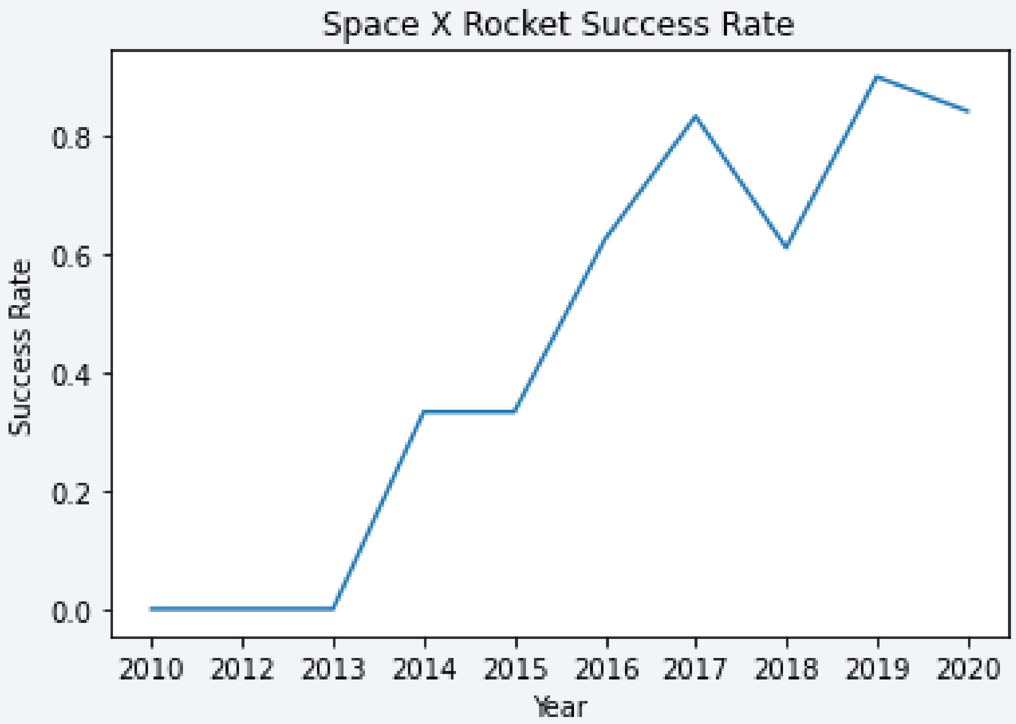
- The 0 (blue) represents a failed launch and the 1 (orange) represents a successful launch.
- From this scatter plot, we can notice that there is a general positive correlation between higher flight numbers and successful outcomes for most orbit types.

Payload vs. Orbit Type



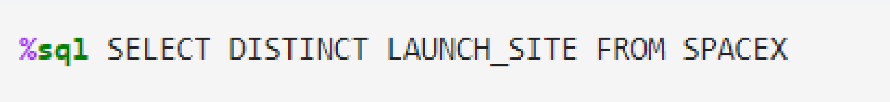
- The 0 (blue) represents a failed launch and the 1 (orange) represents a successful launch.
- From this, we can see that the payload mass can impact the success rate of some orbit types. However, more data is needed of higher payload masses to derive conclusions.

Launch Success Yearly Trend

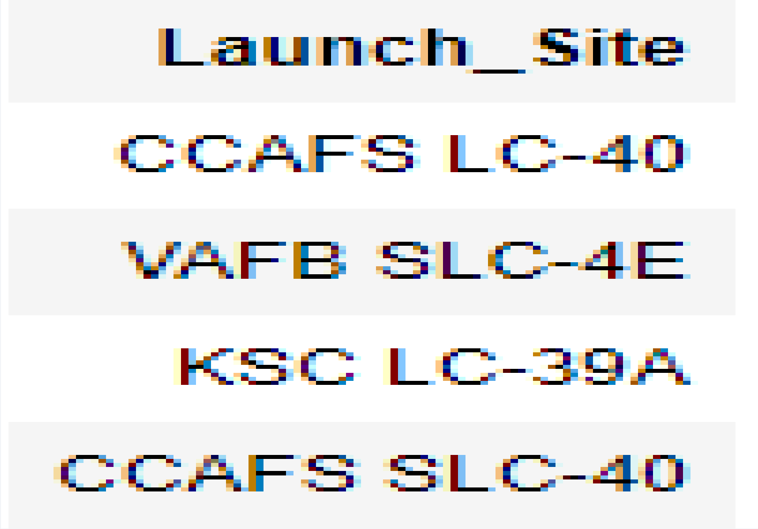


We can see that the SpaceX success rate has generally increased since 2013 with the biggest leap being from 2015 to 2017.

All Launch Site Names



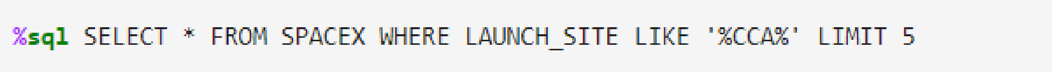
```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX
```



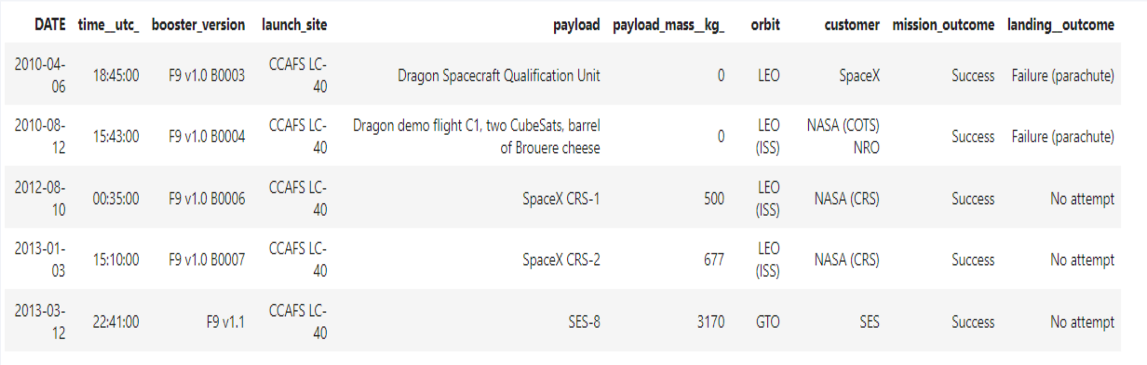
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

The use of DISTINCT in the query allows to remove duplicate LAUNCH_SITE.
The output shows that there were 4 launch sites.

Launch Site Names Begin with 'CCA'



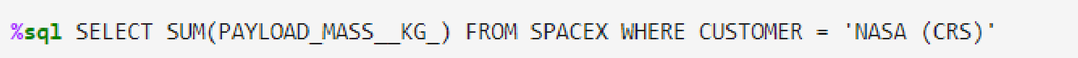
```
%sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE '%CCA%' LIMIT 5
```



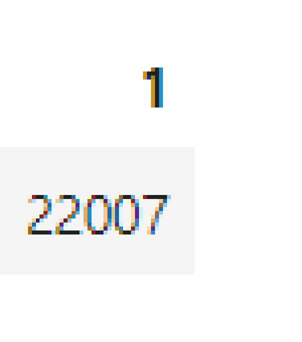
DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

The WHERE clause followed by LIKE clause filters launch sites that contain the substring CCA. LIMIT 5 shows 5 records from filtering.

Total Payload Mass



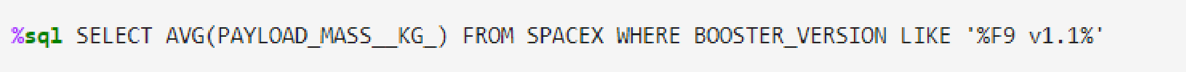
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)'
```



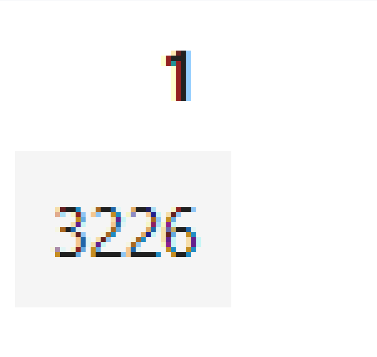
```
    1
22007
```

The result of the query shows that the total of payload mass for NASA (CRS) is 22007 kg.

Average Payload Mass by F9 v1.1



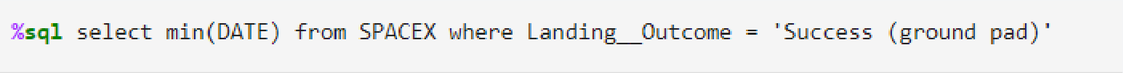
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEX WHERE BOOSTER_VERSION LIKE '%F9 v1.1%'
```



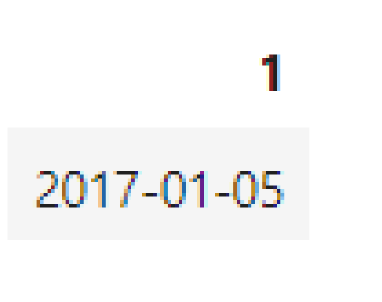
1
3226

The result of the query shows that the average payload masses where the booster version contains the substring F9 v1.1 was 3226 kg.

First Successful Ground Landing Date



```
%sql select min(DATE) from SPACEX where Landing__Outcome = 'Success (ground pad)'
```

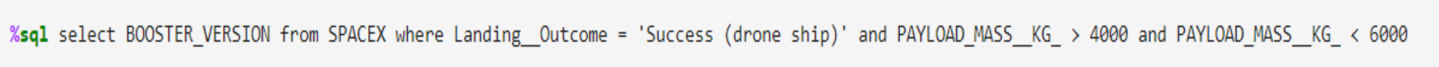


1
2017-01-05

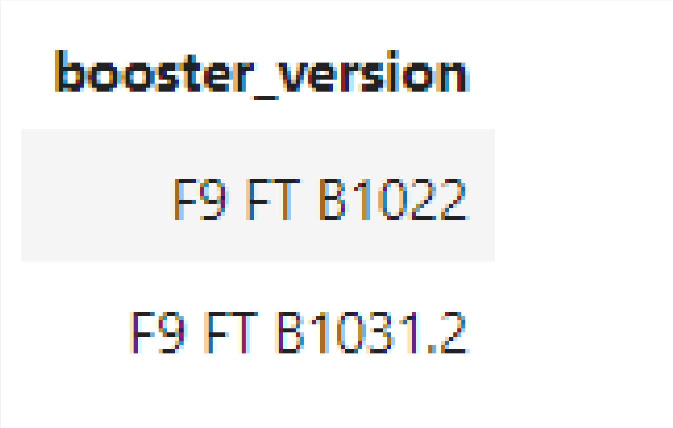
The result of the query shows that the first successful ground landing was in the 5th January 2017.

With the MIN function, we select the record with the oldest date.

Successful Drone Ship Landing with Payload between 4000 and 6000



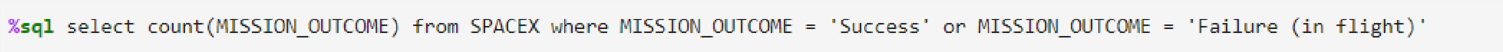
```
%sql select BOOSTER_VERSION from SPACEX where Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```



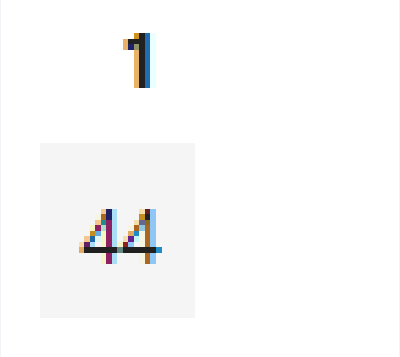
booster_version
F9 FT B1022
F9 FT B1031.2

The result of the query shows that only 2 booster versions with a payload mass between 4000kg and 6000 kg had a successful landing.

Total Number of Successful and Failure Mission Outcomes



```
%sql select count(MISSION_OUTCOME) from SPACEX where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
```



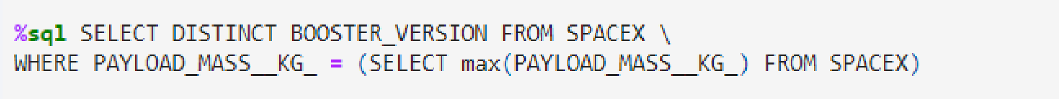
1
44

The result of the query shows that the total number of successful and failed missions were 44.

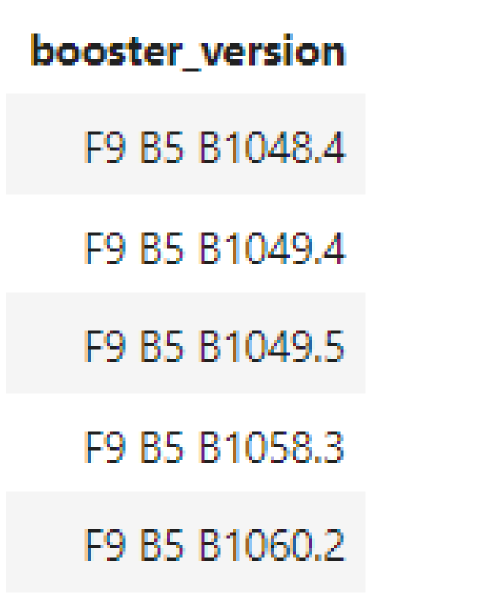
The WHERE clause followed by LIKE clause filters mission outcome.

The OR combines the two outcomes.

Boosters Carried Maximum Payload



```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEX \
WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEX)
```

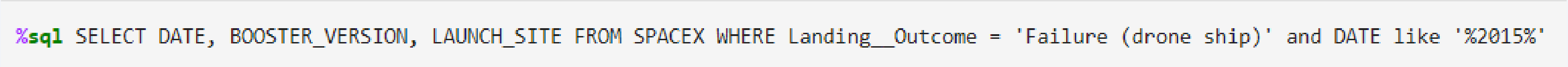


booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1058.3
F9 B5 B1060.2

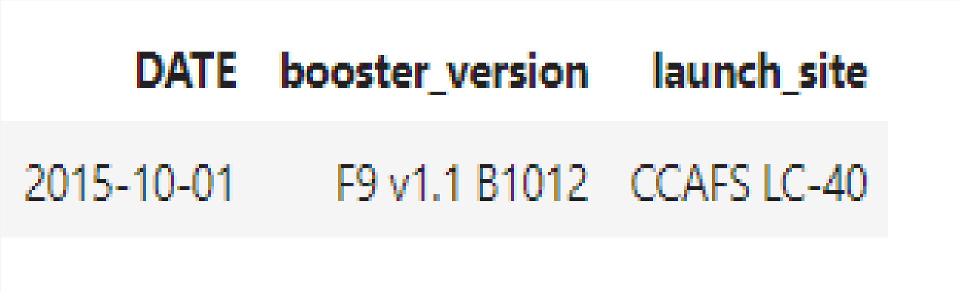
The result of the query shows the booster versions with the heaviest payload.

The max function was used to return the booster versions with the heaviest payload mass.

2015 Launch Records



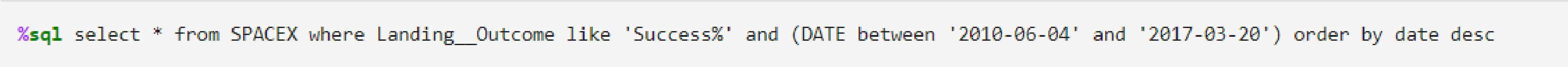
```
%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE Landing__Outcome = 'Failure (drone ship)' and DATE like '%2015%'
```



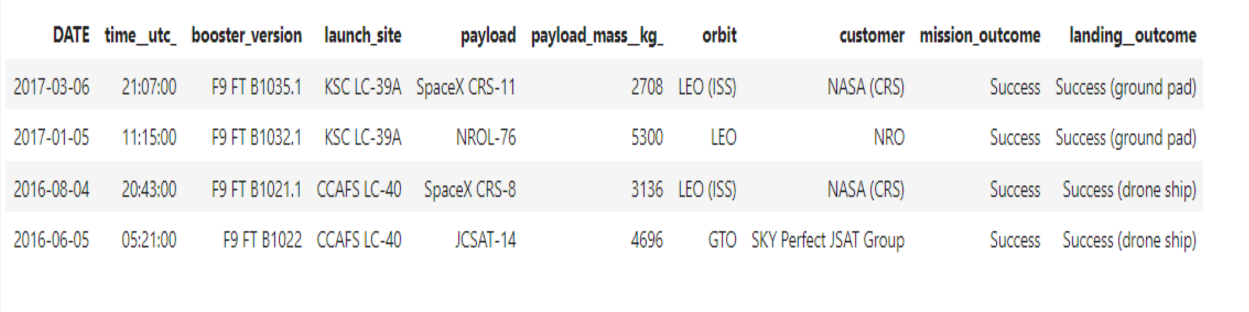
DATE	booster_version	launch_site
2015-10-01	F9 v1.1 B1012	CCAFS LC-40

The result of the query returns the date, booster version, and launch site of an unsuccessful launch in 2015.
The LIKE '%2015%' will find the date with the number '2015'.

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



```
%sql select * from SPACEX where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc
```



DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2017-03-06	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-01-05	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2016-08-04	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2016-06-05	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)

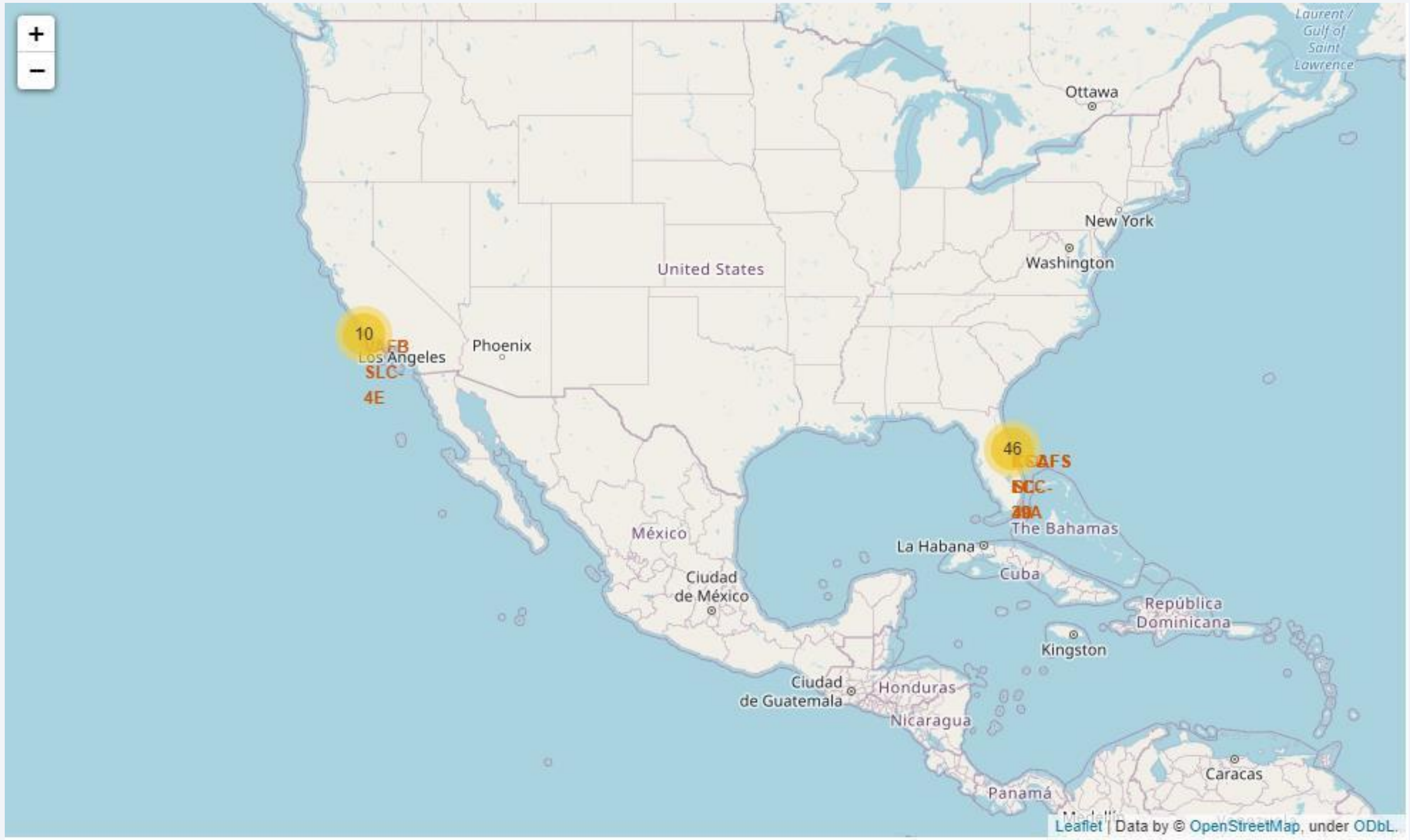
The result of the query ranks the successful landing outcomes by order of date from 2010-06-04 to 2017=03=20

The ORDER BY date desc shows results in decreasing order.

Section 4

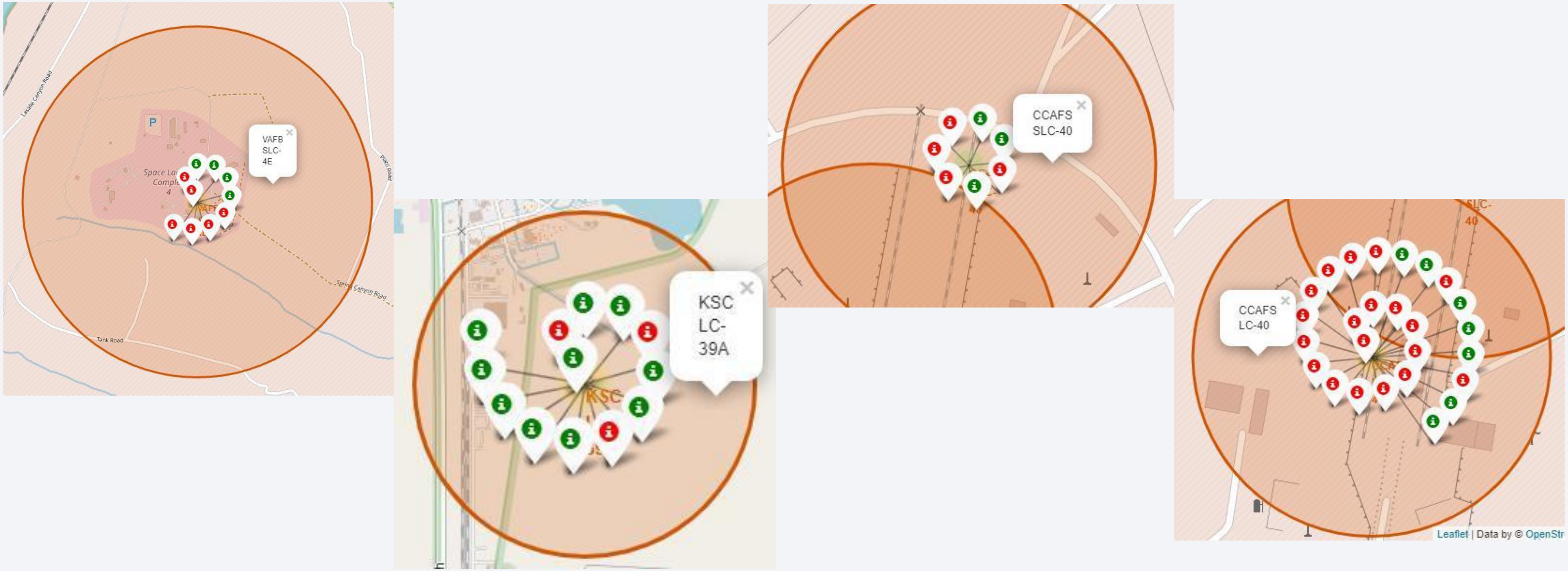
Launch Sites Proximities Analysis

Folium map - Ground stations



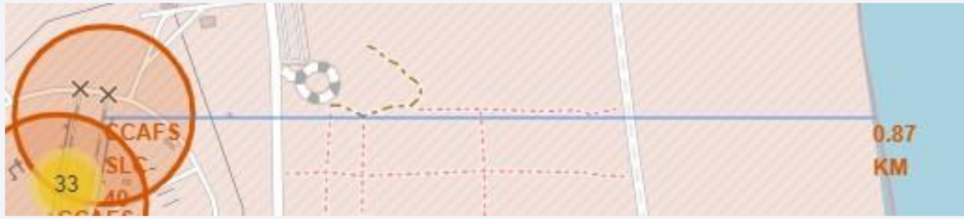
Space X launch sites are shown to be located on the East and West coasts of the United States of America.

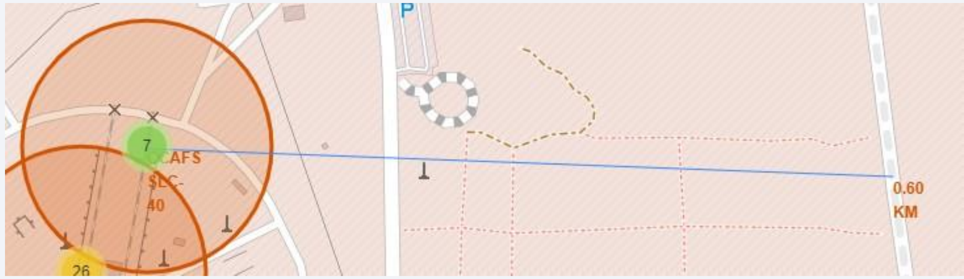
Folium map - Color Labeled Markers

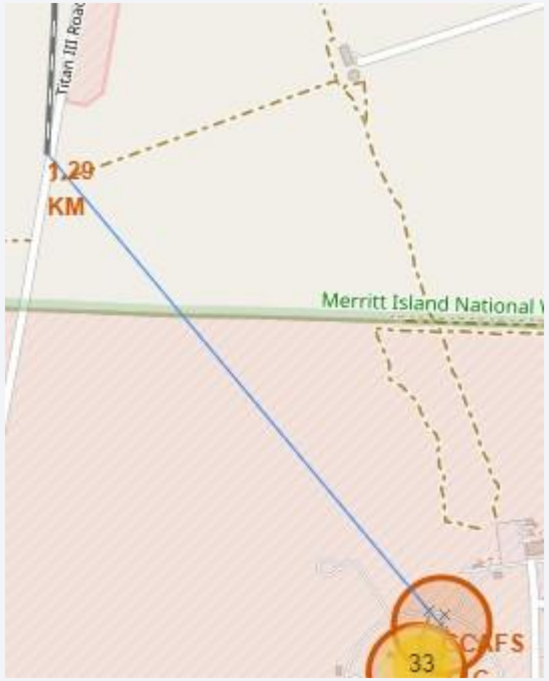


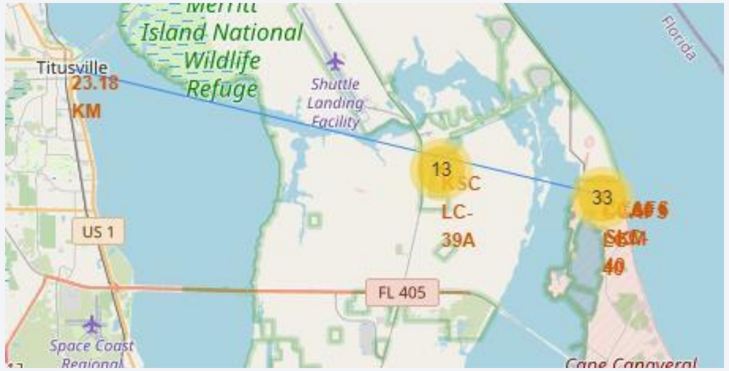
- The green marker represents successful launches.
- The red marker represents unsuccessful launches.
- The KSC LC-39A launch site has a higher launch success rate.

Folium Map - Distances between CCAFS SLC-40 and its proximities









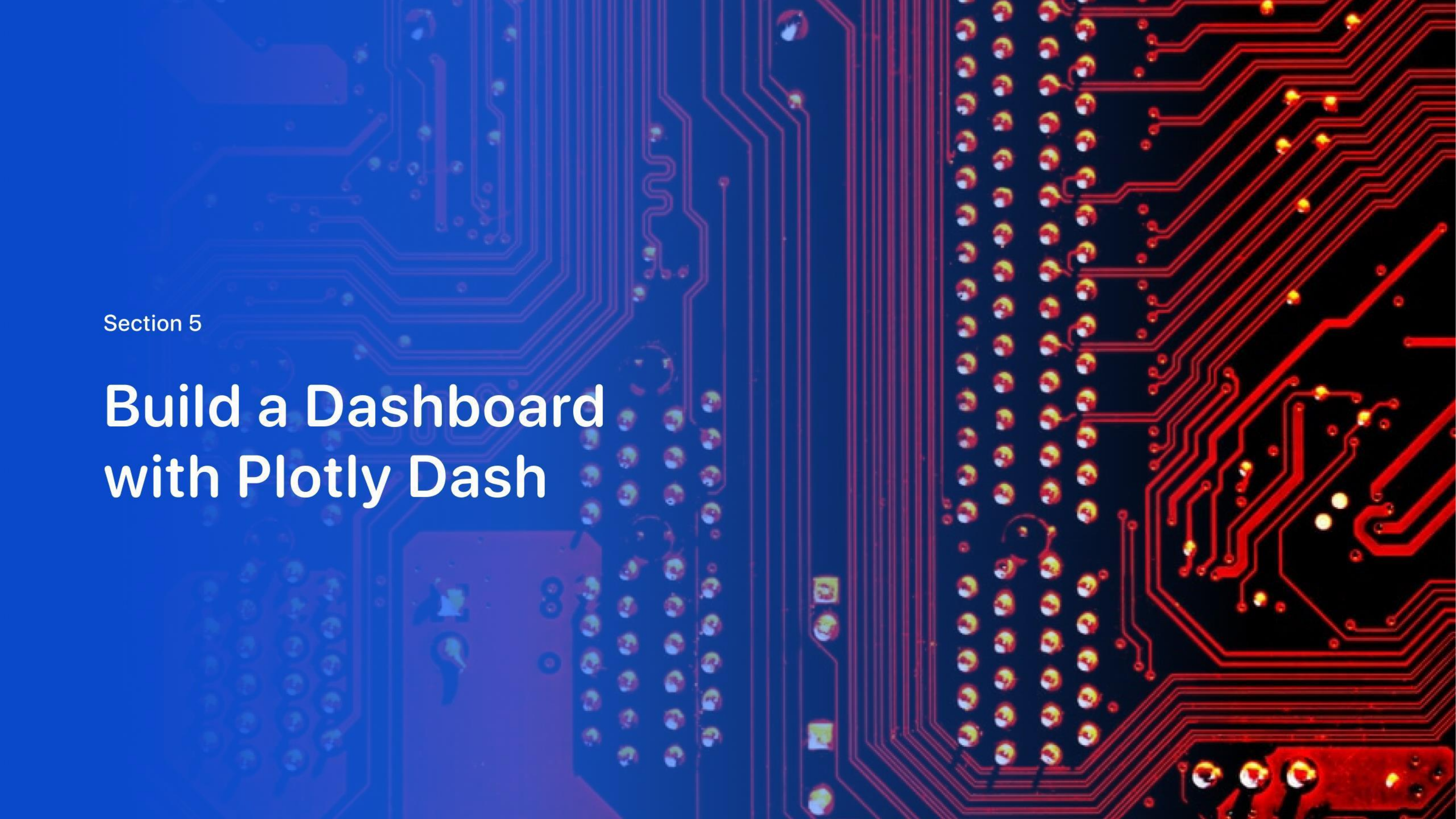
Proximities CCAFS SLC-40 is close to:

- Railway
- Coastline
- Highway

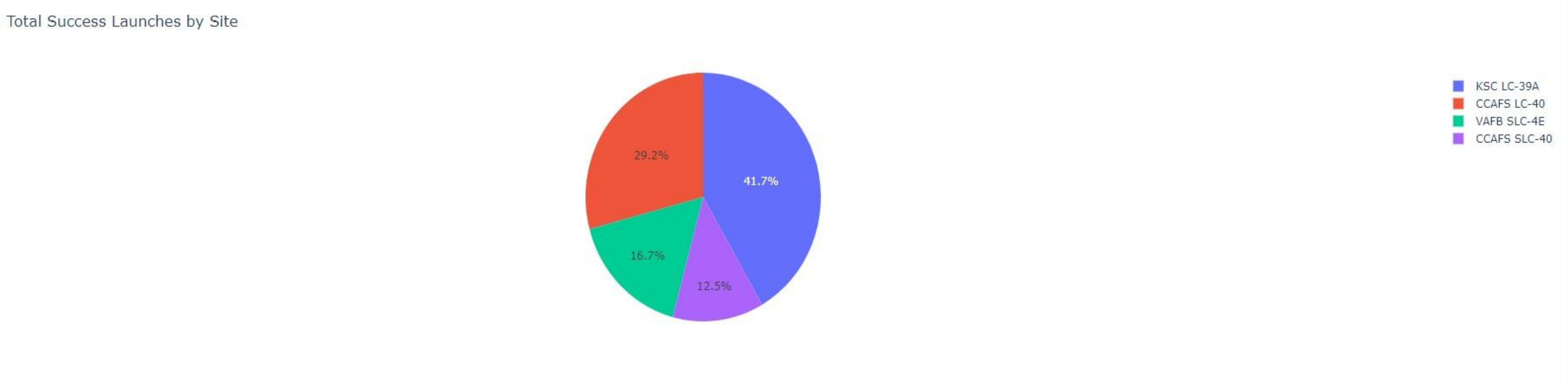
These findings show us that the CCAFS SLC-40 launch site is very accessible which is important for logistics and resources.

Section 5

Build a Dashboard with Plotly Dash

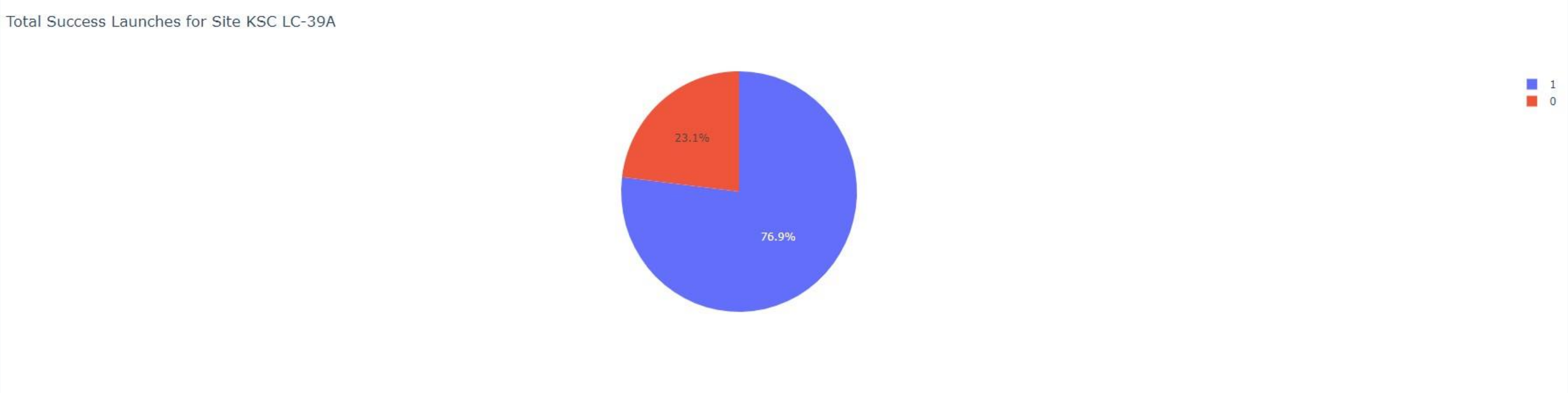


Dashboard - Total success by Site



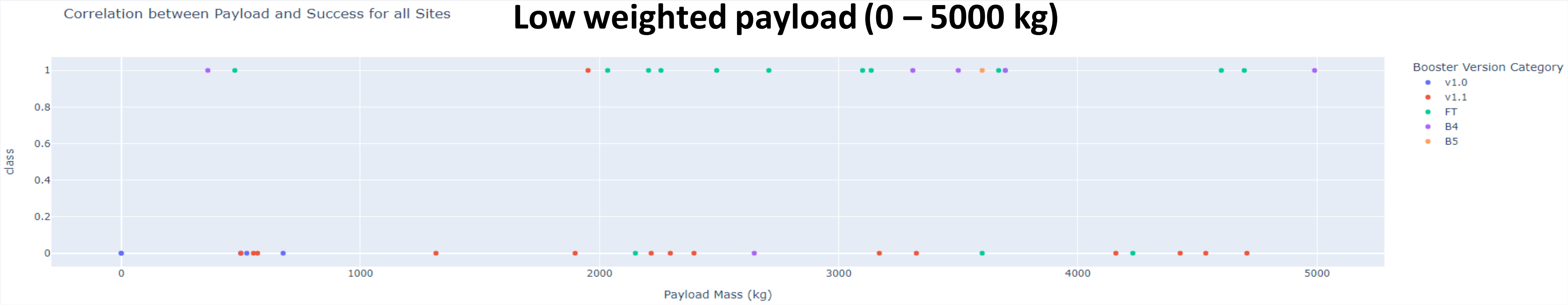
We see that KSC LC-39A has the best success rate of launches.

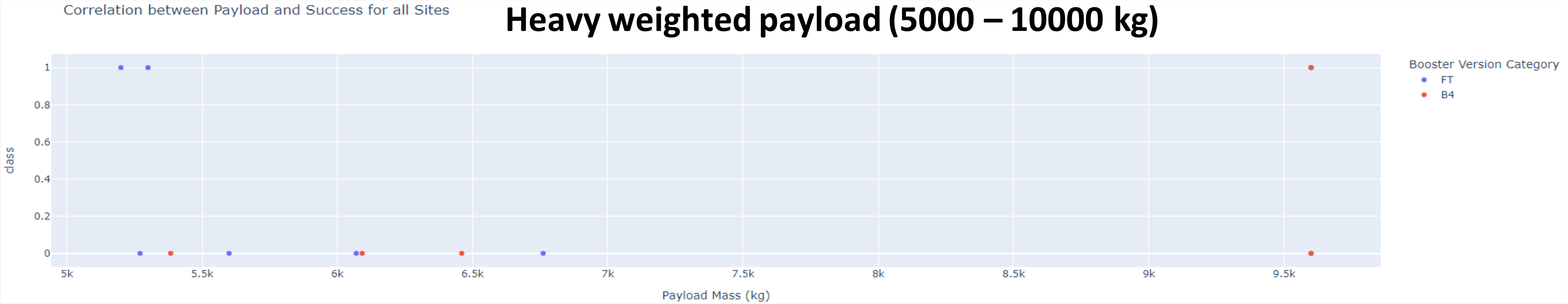
Dashboard - Total success launches for Site KSC LC-39A



- The 1 (blue) represents successful launches and the 0 (orange)represents failed launches.
- KSC LC-39A has achieved a 76.9% success rate.

Dashboard - Payload mass vs Outcome for all sites with different payload mass selected



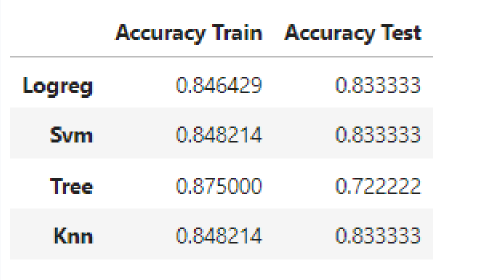


Low weighted payloads have a better success rate than the heavy weighted payloads.

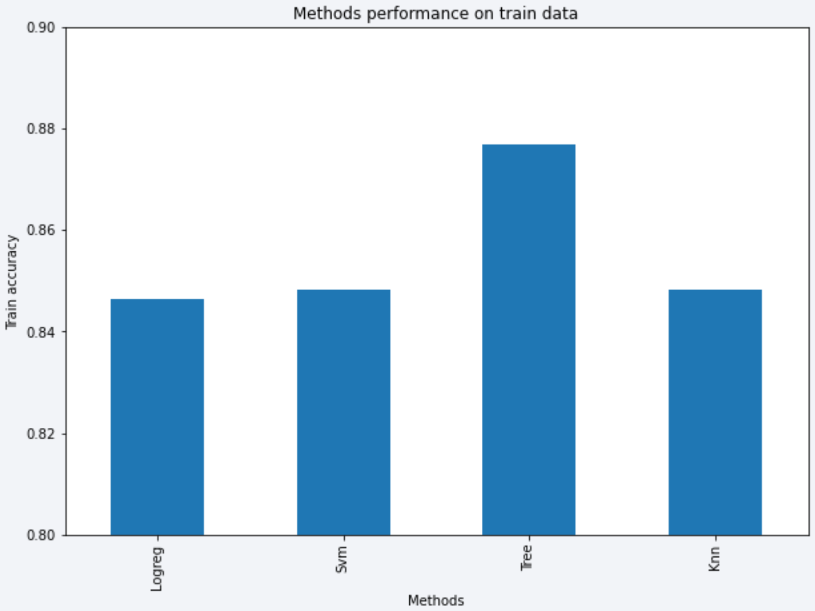
Section 6

Predictive Analysis (Classification)

Classification Accuracy



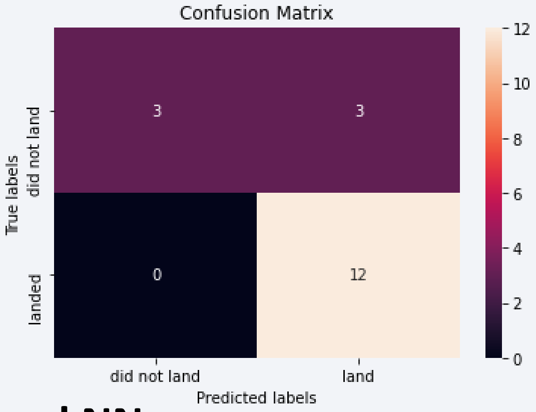
	Accuracy Train	Accuracy Test
Logreg	0.846429	0.833333
Svm	0.848214	0.833333
Tree	0.875000	0.722222
Knn	0.848214	0.833333



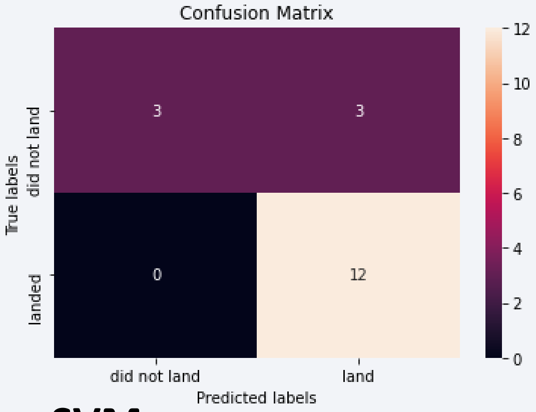
- All methods performed generally well.
- The decision tree had the highest accuracy train, but with an accuracy test that is lower than all other tests.
- KNN and the SVM have the better balance of the two scores.
- We would need more test data to determine the most accurate test with more certainty.

Confusion Matrix

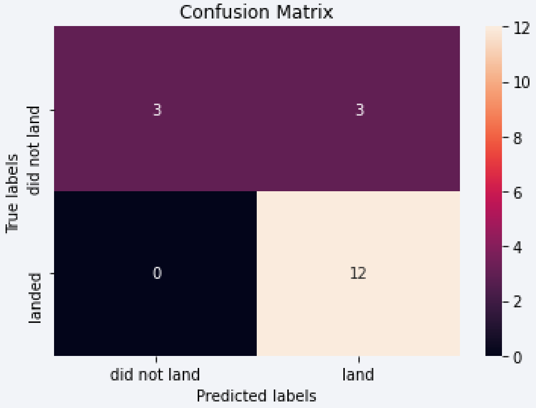
Logistic regression



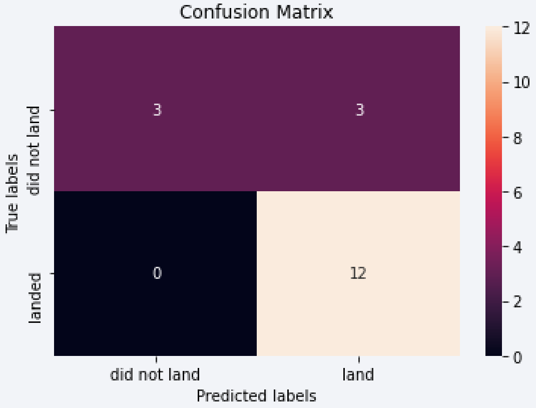
Decision Tree



kNN



SVM



The Confusion matrices are all equal.

Conclusions

- We have learnt that there are key factors in determining the success of a launch which includes the launch site, the orbit as well as the number of previous launches. This information gain can better prepare SpaceX to maximise their likelihood of having successful outcomes.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- Generally, low weighted payloads perform better than heavy weighted payloads.
- KSC LC-39A is the best launch site, however more data is needed to understand the other factors at play
- The SVM classification model is our best model as it had a better balance between the accuracy train and the accuracy test than the Decision Tree and a higher accuracy train the logistic regression model. We prefer the SVM over the KNN model although they performed the same on accuracy tests because it is easier to compute and interpret.

Thank you!