IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

DANIEL Fukson
09.06.2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data for analysis was collected from the SpaceX public API and publicly available data on Wikipedia.
  - Data wrangling included extracting of mission outcome to serve as the dependent variable in the Machine Learning models.
  - SQL queries and data visualizations (static plots, interactive maps, and an interactive dashboard) were created to discover insights about the data set and answer questions.
  - Interactive analysis done with the help of Folium maps and Plotly Dash
  - Predictive analysis was pursued using Logistic Regression, SVM (Support Vector Machine), Decision Tree, and KNN (k-Nearest Neighbors) Machine Learning models.
- Summary of all results
  - EDA results
  - Interactive analysis results
  - Predictive analysis results

Introduction

- Project background and context:
  - This project was done for a company that wants to enter the space shuttle launching business and to become a successful bidder against SpaceX company.
  - To do it we need to analyze and predict if the Falcon 9 first stage will land successfully.
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Problem that we want to find answers to:
  - If we can determine what factors are influencing successful landing of the first stage of Falcon 9, we can understand how to build the new spaceship and and its launching infrastructure to gain success in launching the rockets. This information can be used if we want to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Initial data for analysis were collected through the publicly available resources:
    - SpaceX REST API
    - Information from Wikipedia presented as the List of Falcon 9 and Falcon Heavy Launches in HTML format
- Perform data wrangling
  - In preparation to data analysis the Data was properly structured, cleaned, enriched with meaningful content, validated and stored in csv files.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, Support Vector Machine, K-Nearest Neighbors and Decision Trees models have been built and evaluated for the best predicted classifier

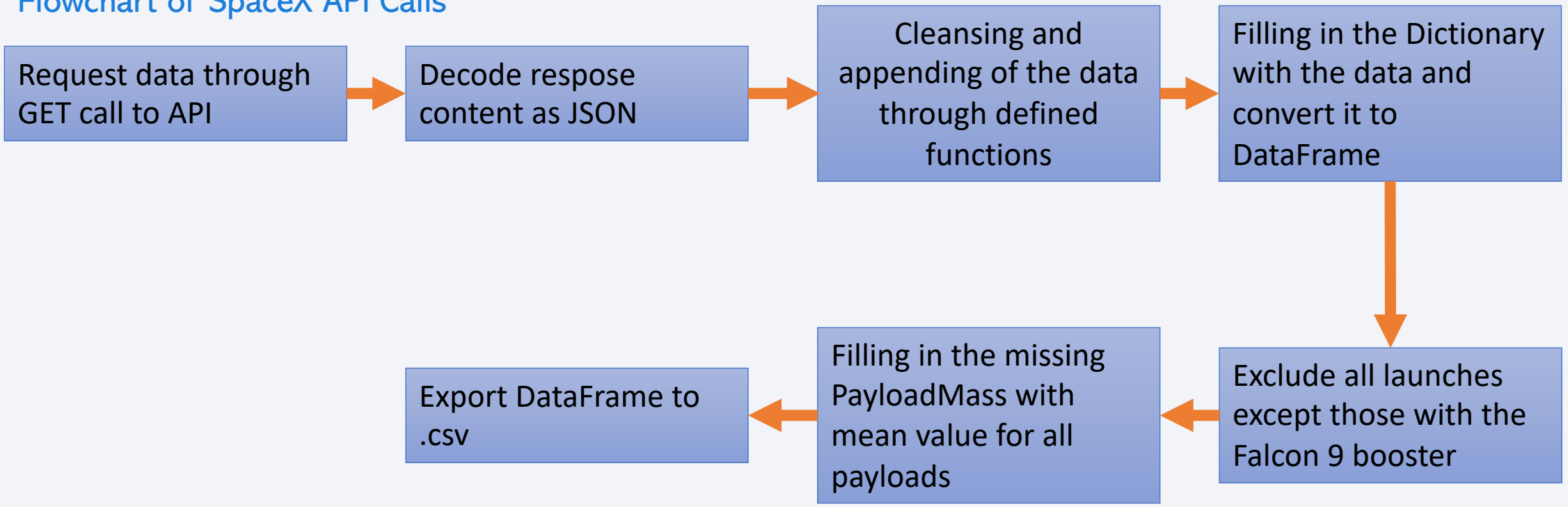
Data Collection

The data sets were collected from the following sources:

- Publicly accessible SpaceX API with historical launch data in JSON format
  - GET request to the SpaceX API to retrieve all available data in JSON format
  - Cleansing of the data, selecting and appending relevant features for future analysis
  - Converting to DataFrrame and filtering on Falcon 9 launches only, filling in missing values
  - Exporting to csv file
- Wikipedia web page with SpaceX launch data in HTML tables from static URL
  - Web scraping Falcon 9 Launch Wiki page from specified URL
  - Use HTLML parser to create a BeautifulSoup object
  - Parsing the launch HTML tables into dictionary and convert it into the DataFrame
  - Exporting to csv file

Data Collection – SpaceX API

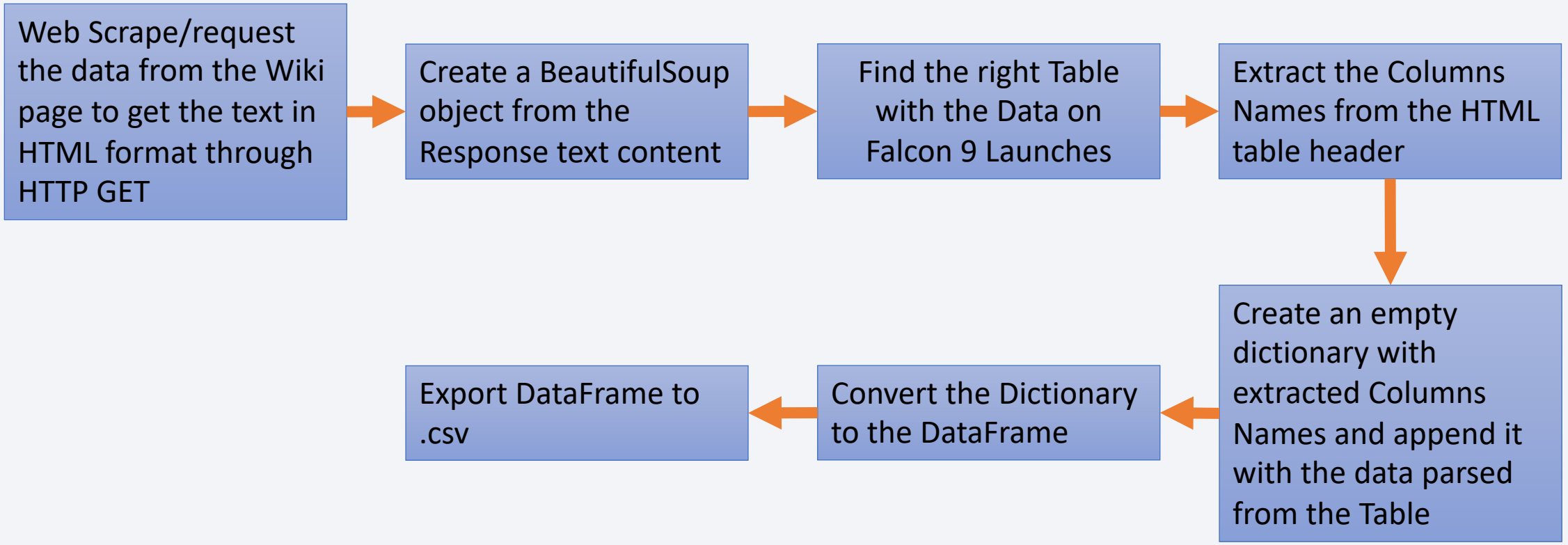
Flowchart of SpaceX API Calls



GitHub URL of the completed SpaceX API calls notebook

Data Collection - Scraping

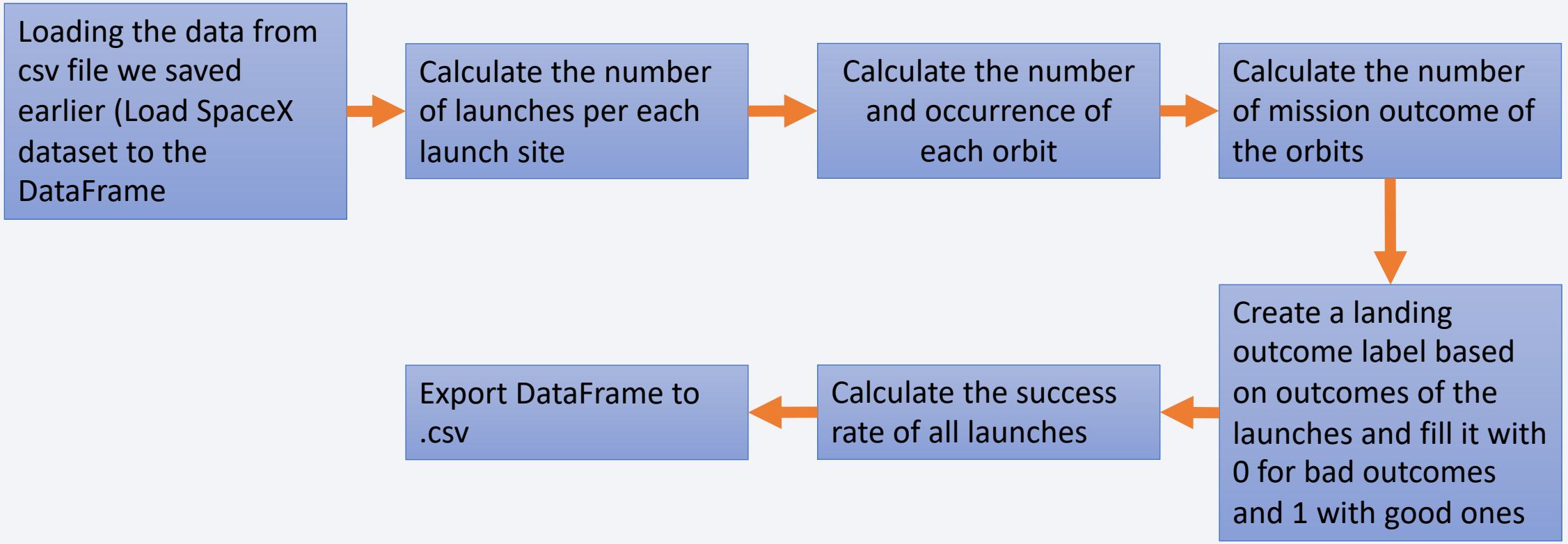
Flowchart of web scraping:



GitHub URL of the completed SpaceX Webscrapping notebook

Data Wrangling

Flowchart of Data Wrangling:



GitHub URL of the completed Data Wrangling notebook

EDA with Data Visualization

The following charts were created to look at Launch Site trends

- Scatterplot to see **Mission Outcome** relationship split by **Launch Site** and **Flight Number**.
- Scatterplot to see **Mission Outcome** relationship split by **Launch Site** and **Payload**.

The following charts were created to look at Orbit Type trends

- Bar chart to see **Mission Outcome** relationship against **Orbit Type**.
- Scatterplot to see **Mission Outcome** relationship split by **Orbit Type** and **Flight Number**.
- Scatterplot to see **Mission Outcome** relationship split by **Orbit Type** and **Payload**.

The following chart was created to look at trends based on time

- Line plot to see **Success Rate** trend by **Year**.

GitHub URL of the completed EDA Data Visualization notebook

EDA with SQL

SQL Queries were performed against the SpaceX Launch data to find out the following information:

- The names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- 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, 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 of the completed EDA with SQL notebook

Build an Interactive Map with Folium

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

- Folium Markers and Circles to mark the Launch Sites locations
- Folium Markers Cluster with Markers to mark the success/failed launches for each Launch Site on the map
- Folium Lines to calculate and show the distance to the Launch Sites proximities
  - Distance from CCAFS LC-40 to the coastline
  - Distance from CCAFS LC-40 to the railroad
  - Distance from CCAFS LC-40 to the closest city

GitHub URL of the completed Interactive Map with Folium Notebook
NBViewer URL of the completed Interactive Map with Folium Notebook

Build a Dashboard with Plotly Dash

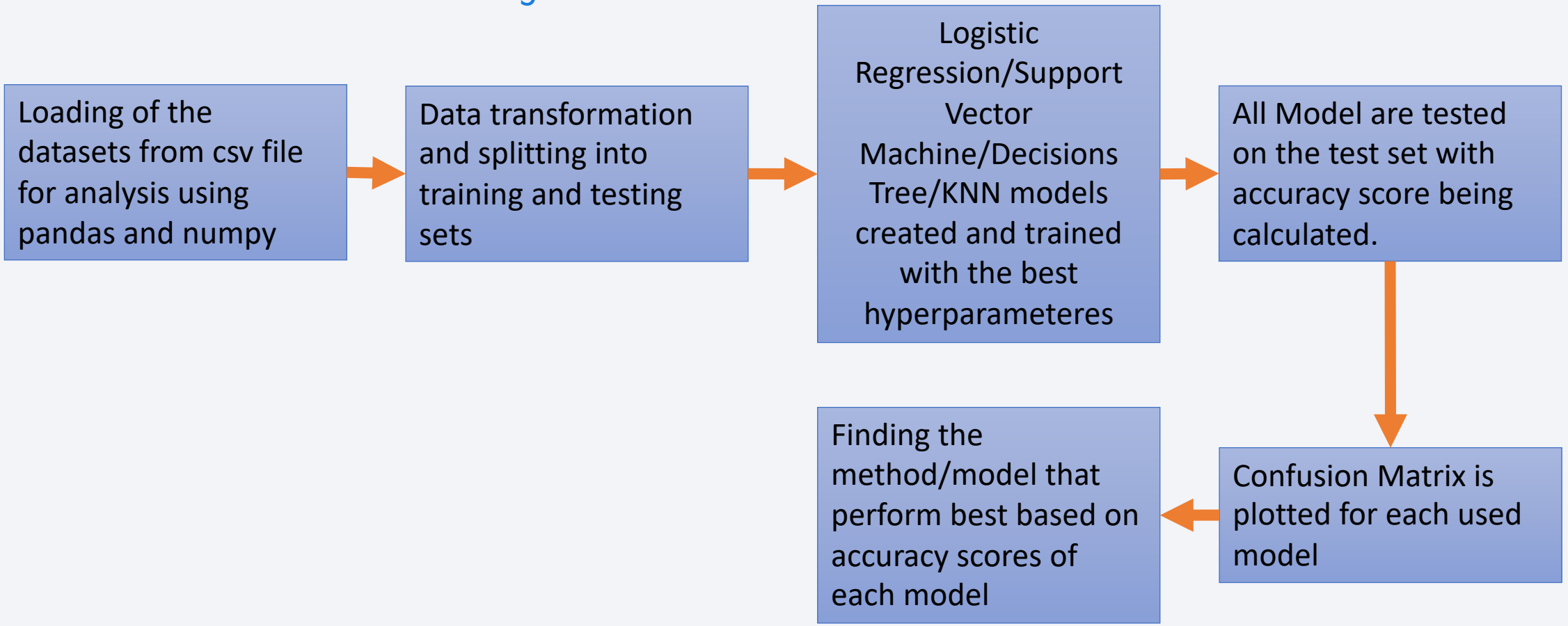
The following plots/graphs interactions have been added to the dashboard:

- The dropdown list for selection of the Launch Site for the Dashboard analysis
- The pie chart that displays the information based on Launch Site selection:
  - When all sites are selected - the distribution of successful Falcon 9 first stage landings between the sites
  - When one particular site is selected - the distribution of successful and failed Falcon 9 first stage landings for that site
- The input slider that used for filtering of the payload mass for the scatterplot
- The scatterplot displays the distribution of Falcon 9 first stage landings split by payload mass, mission outcome and by booster version category.

GitHub URL of the completed Dashboard with Plotly Dash Notebook

Predictive Analysis (Classification)

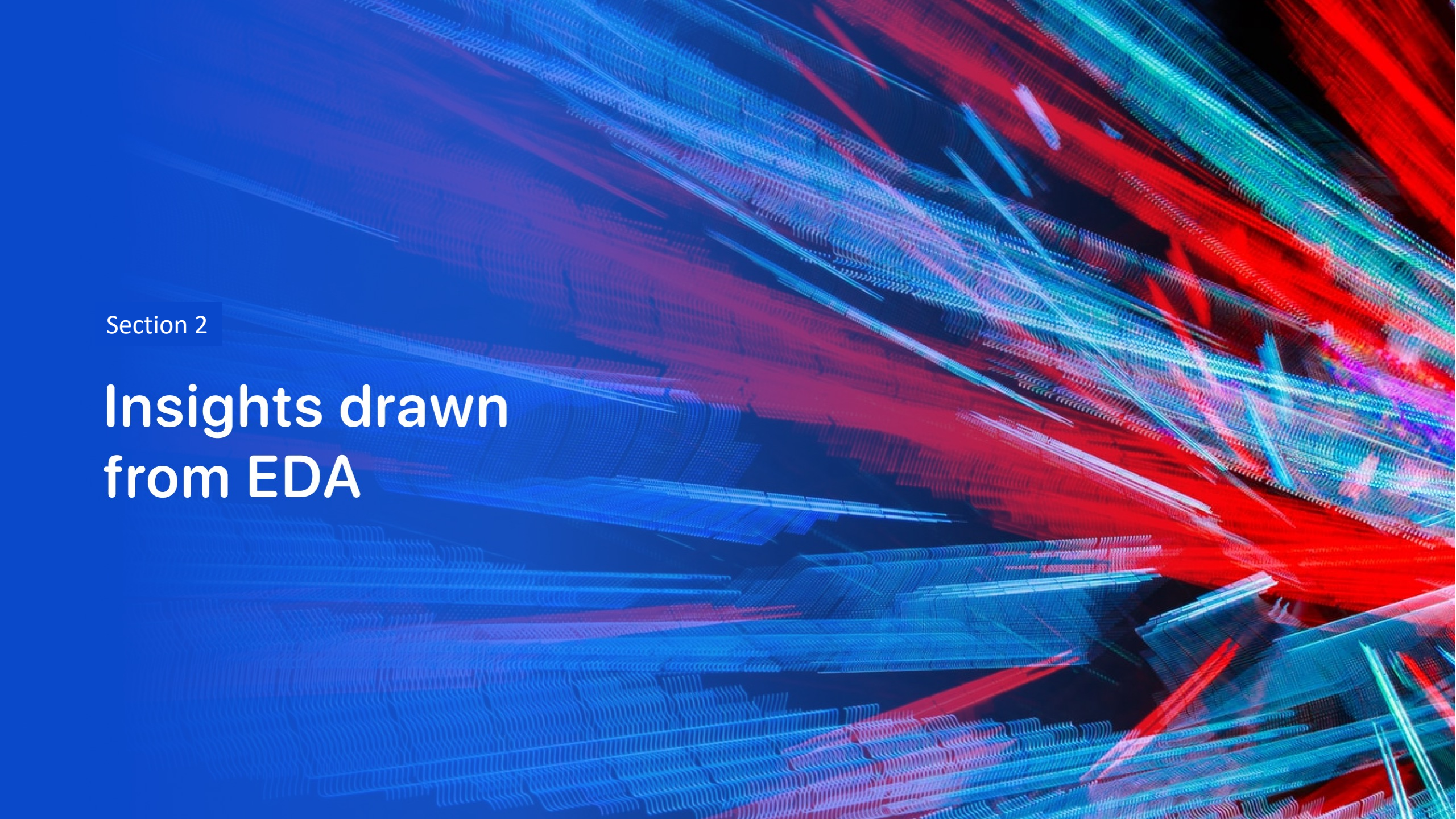
Flowchart of Classification modelling:



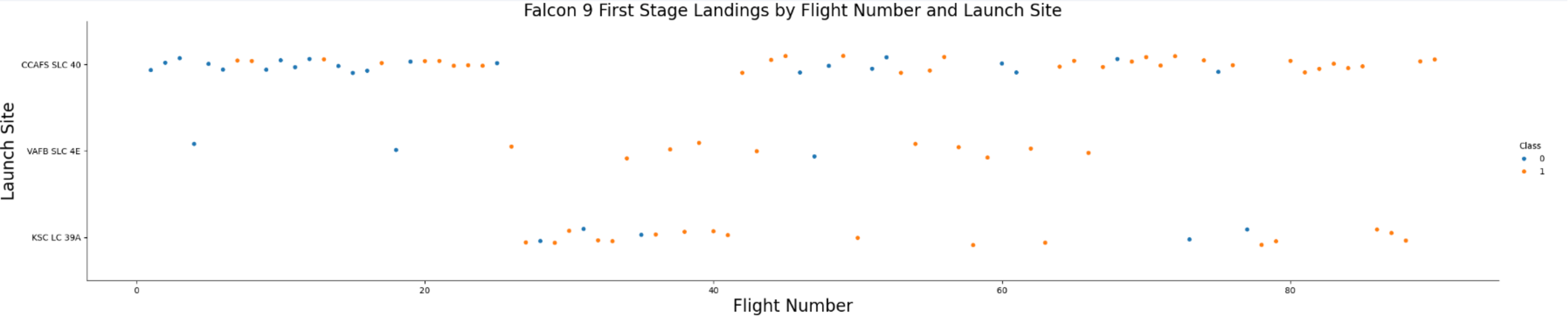
GitHub URL of the completed Predictive Analysis Notebook

Results

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


Section 2
Insights drawn from EDA

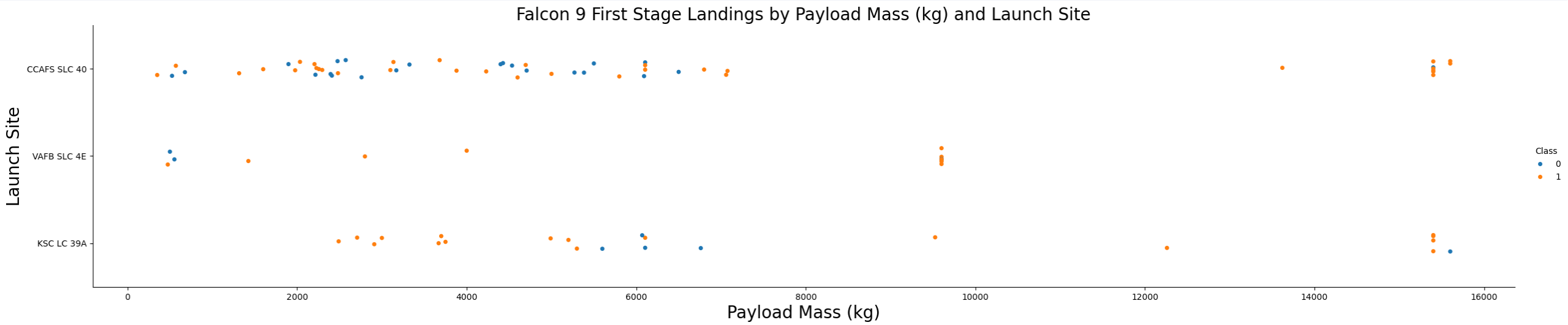
Flight Number vs. Launch Site



Blue colored dots indicates unsuccessful launch; Orange colored dots indicates successful launch

- The number of launches from CCAFS SLC 40 site are significantly higher than form other sites and the number of success landings (Class = 1) are increasing with the increase of the number of launches for this site
- For all site with the number of launches growing the number of successful landing (Class = 1) is also growing
- Successful landing rate for sites are different with the best rate being at KSC LC39A site

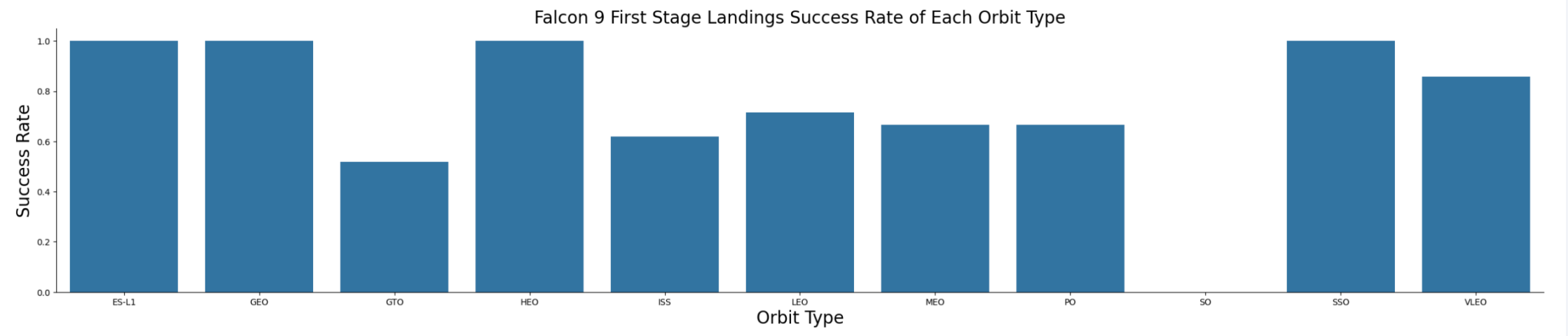
Payload vs. Launch Site



Blue colored dots indicates unsuccessful launch; Orange colored dots indicates successful launch

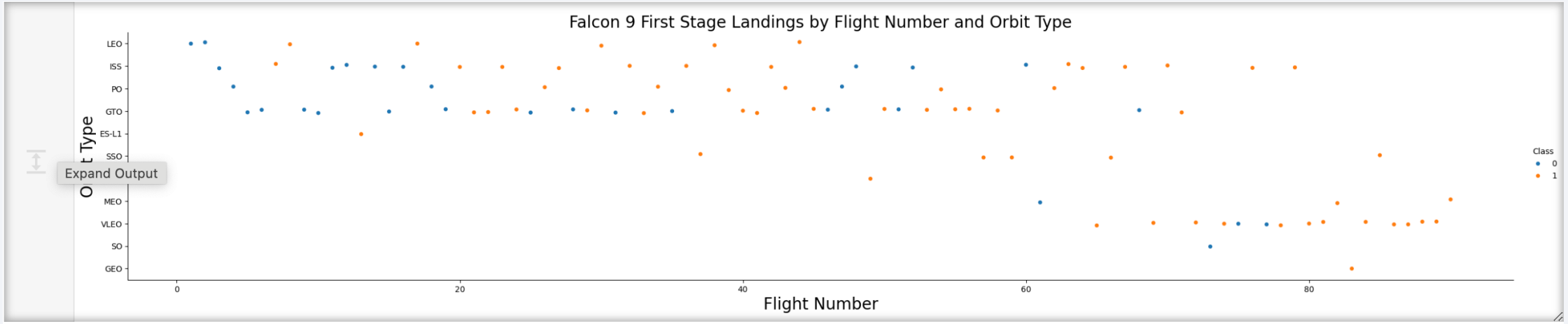
- Most of the payloads for all sites are under 7000 kg but for those launches with the larger payloads success rate is much higher
- In general, there's no much correlation between the Payloads and successes rate of launches

Success Rate vs. Orbit Type



ES-L1, GEO, HEO, SSO have the largest success rates. VLEO success rate is a little bit lower. GTO has the minimums success rate. Other sites are in between on these two groups.

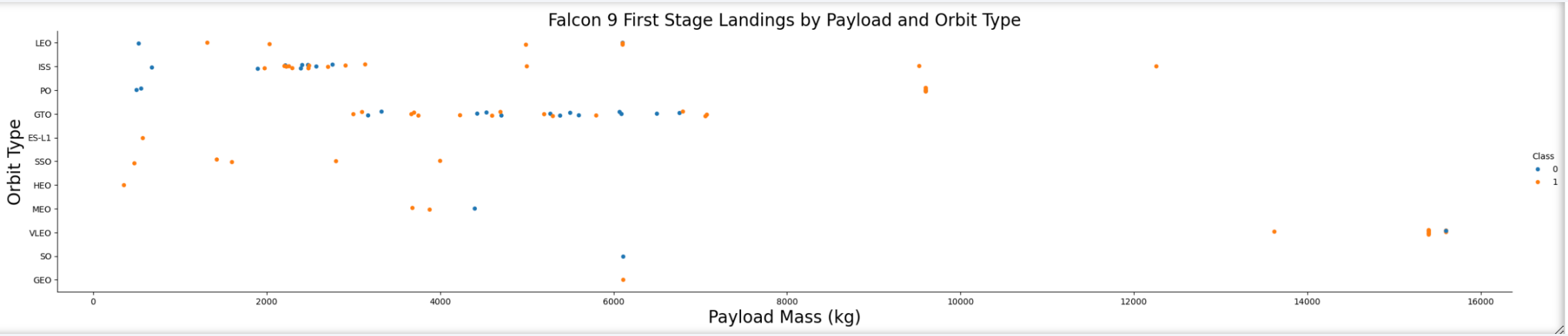
Flight Number vs. Orbit Type



Blue colored dots indicates unsuccessful launch; Orange colored dots indicates successful launch

LEO orbit success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.

Payload vs. Orbit Type

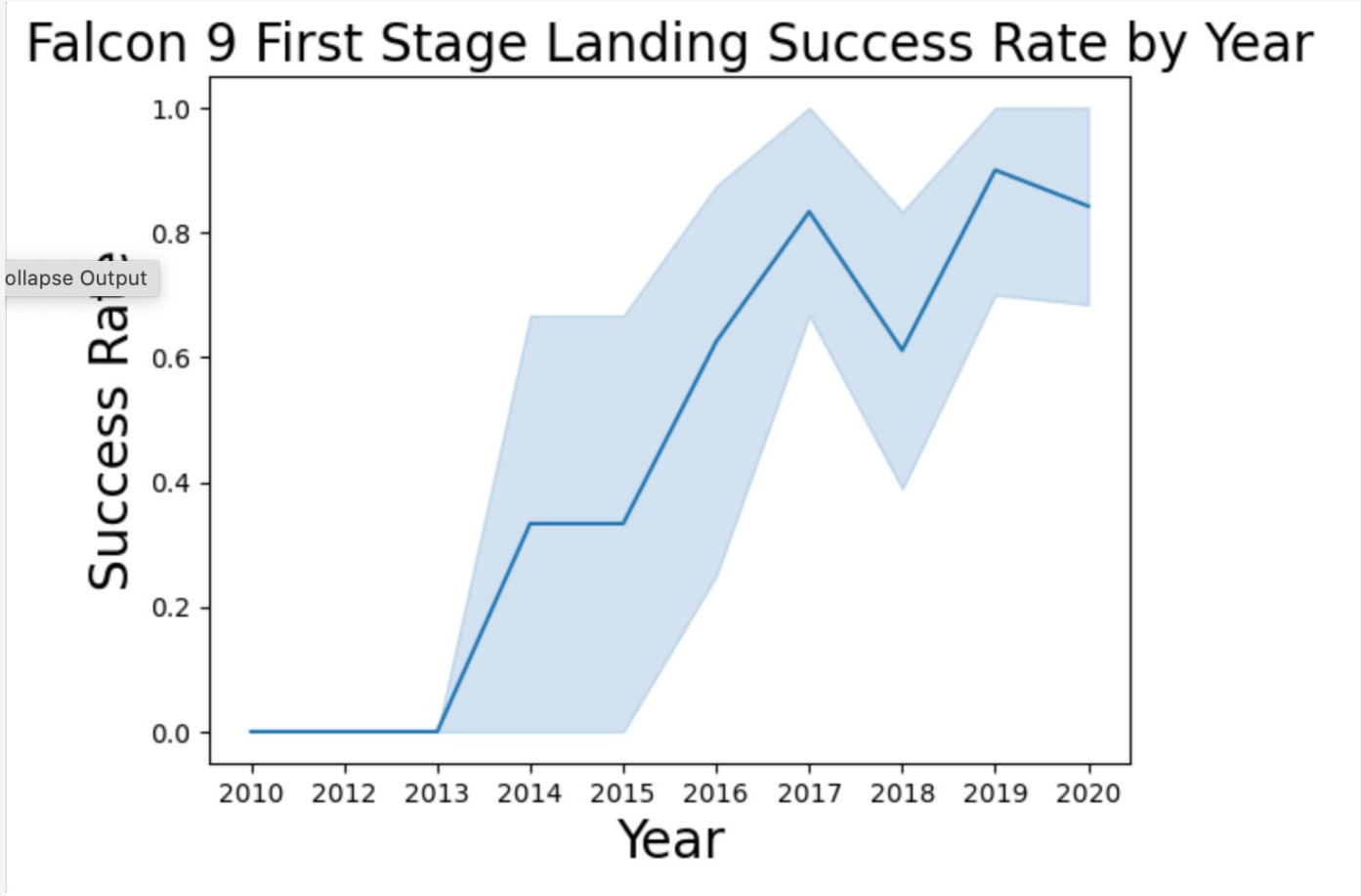


Blue colored dots indicates unsuccessful launch; Orange colored dots indicates successful launch

We can say that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.

Launch Success Yearly Trend

- Success Rate was increasing year over year from 2013 to 2017. 2018 it decreased, but starting from 2019 it stabilized at around 0,8.



All Launch Site Names

SQL Query:

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE;
```

Results:

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

We are getting the results by querying unique values from the field Launch_Site and using SELECT DISTINCT statement

Launch Site Names Begin with 'CCA'

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

SQL Query

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

- Query result:

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

We are getting the results by querying values from the field Launch_Site and using WHERE, LIKE for looking the “CCA” pattern and LIMIT statement to find first 5 records

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

SQL Query

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTABLE \
    WHERE CUSTOMER = 'NASA (CRS)';
```

Query result:

SUM(PAYLOAD_MASS__KG_)
45596

We are getting the results by querying values from the fields Payload_Mass_kg and Customer and using WHERE statement to find all records where Customer is NASA and SUM statement to calculate total of payloads.

Average Payload Mass by F9 v1.1

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

SQL Query

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
    FROM SPACEXTABLE \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

Query result:

AVG(PAYLOAD_MASS__KG_)
2928.4

We are getting the results by querying values from the fields Payload_Mass_kg and Booster_Version and using WHERE statement to find all records where Booster is F9 v1.1 and AVG statement to calculate average of payloads.

First Successful Ground Landing Date

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

SQL Query

```
%sql SELECT MIN(DATE) \
    FROM SPACEXTABLE \
    WHERE Landing_Outcome = 'Success (ground pad)'
```

Query result:

MIN(DATE)
2015-12-22

We are getting the results by querying values from the fields Date and Landing_Outcome and using WHERE statement to find all records where landing was successful and MIN statement to present the first successful landing date.

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 Query:

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

Query result:

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

We are getting the results by querying values from the fields Booster_version, Landing_Outcome and Payload_Mass_kg and using WHERE statement to find all records where landing was successful and BETWEEN statement to present Booster_Version with the mass between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

SQL Query:

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
    FROM SPACEXTABLE \
    GROUP BY MISSION_OUTCOME;
```

Query result:

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

We are getting the results by querying values from the fields Mission_Outcome and using GROUP BY and COUNT statements to find the total number of successful and failed mission outcomes.

Boosters Carried Maximum Payload

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

SQL Query:

```
%sql SELECT BOOSTER_VERSION \
    FROM SPACEXTABLE \
    WHERE PAYLOAD_MASS__KG_ = \
        (SELECT MAX(PAYLOAD_MASS__KG_) \
    FROM SPACEXTABLE);
```

Query result:

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

We are getting the results by querying values from the fields Booster_Version, Payload_Mass_kg and using subquery to select the maximum payload mass of the booster.

2015 Launch Records

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

SQL Query:

```
%sql SELECT substr(Date,6,2) as month, Date, Booster_Version, Launch_Site, Landing_Outcome \
    FROM SPACEXTABLE \
    WHERE Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

Query result:

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

We are getting the results by querying values from the fields Date, Booster_Version, Launch_Site and Landing_Outcome and using WHERE statement to find all records where landing has failed and SUBSTR statement to find the month and the year from the Date field.

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 Query:

```
%sql SELECT Landing_Outcome, count(*) as count_outcomes \
    FROM SPACEXTABLE \
    WHERE DATE between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count_outcomes DESC;
```

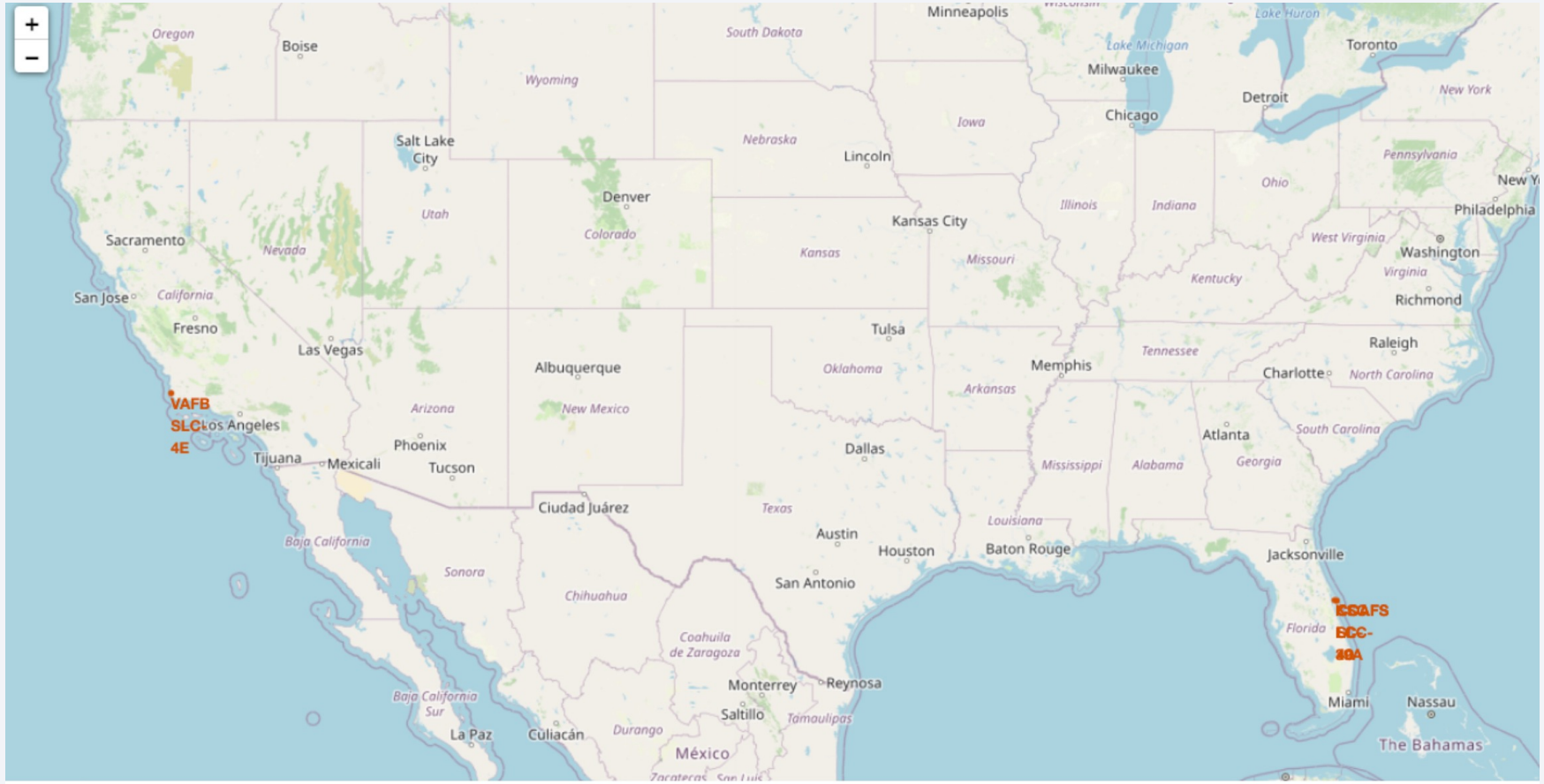
Query result:

Landing_Outcome	count_outcomes
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

We are getting the results by querying values from the fields Landing_Outcome, Date and using COUNT, GROUP BY to calculate numbers of Outcomes grouped by Outcome result and WHERE statement to find all records with the dates from 2010-06-04 till 2017-03-20.


Section 3
Launch Sites
Proximities Analysis

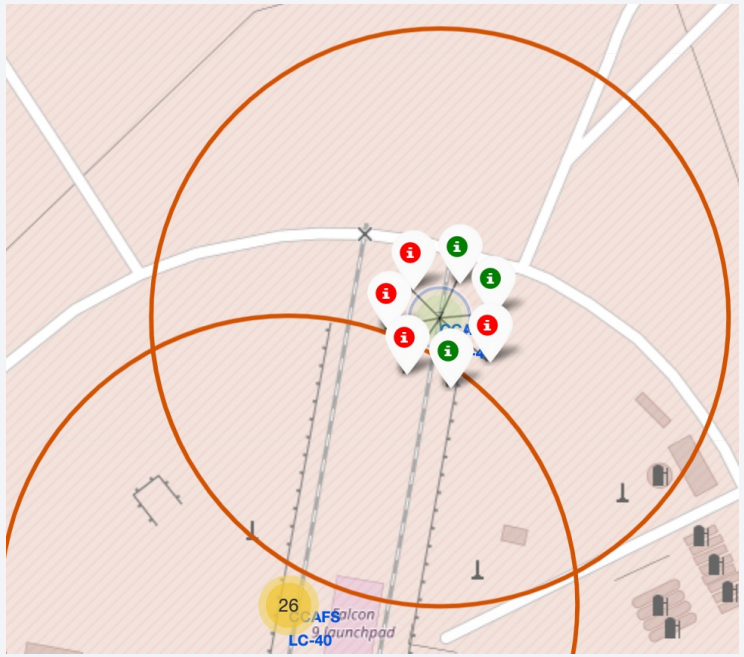
Falcon 9 Launch Site Locations



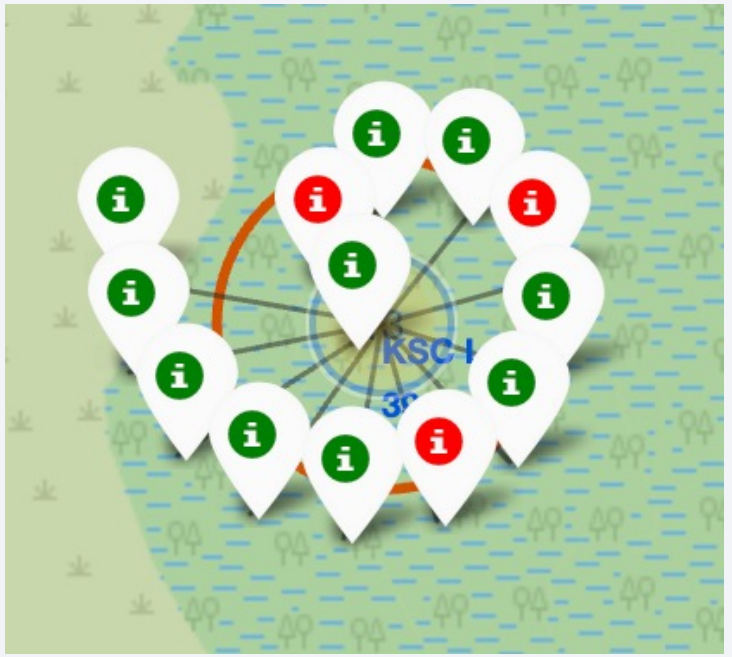
There are four Falcon 9 launch site locations. One location is in Califoria (VAFB SLC-4E) and three in Florida (KSC LC-39A, CCAFS LC-40, CCAFS SLC-40). The locations are at southern part of USA.

Markers of Success/Failed Landings per Launch Site

Site: CCAFS LC-40



Site: CCAFS SLC-40



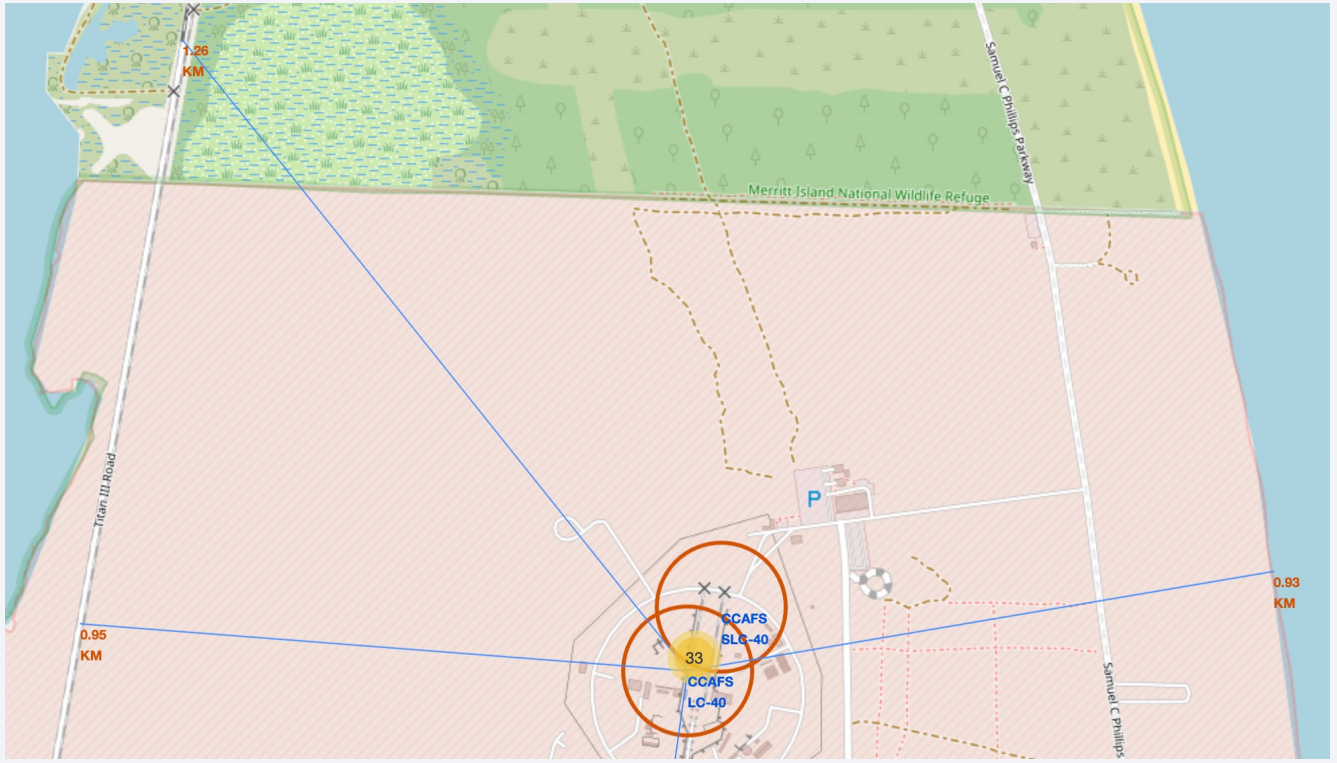
Site: KSC LC-39A

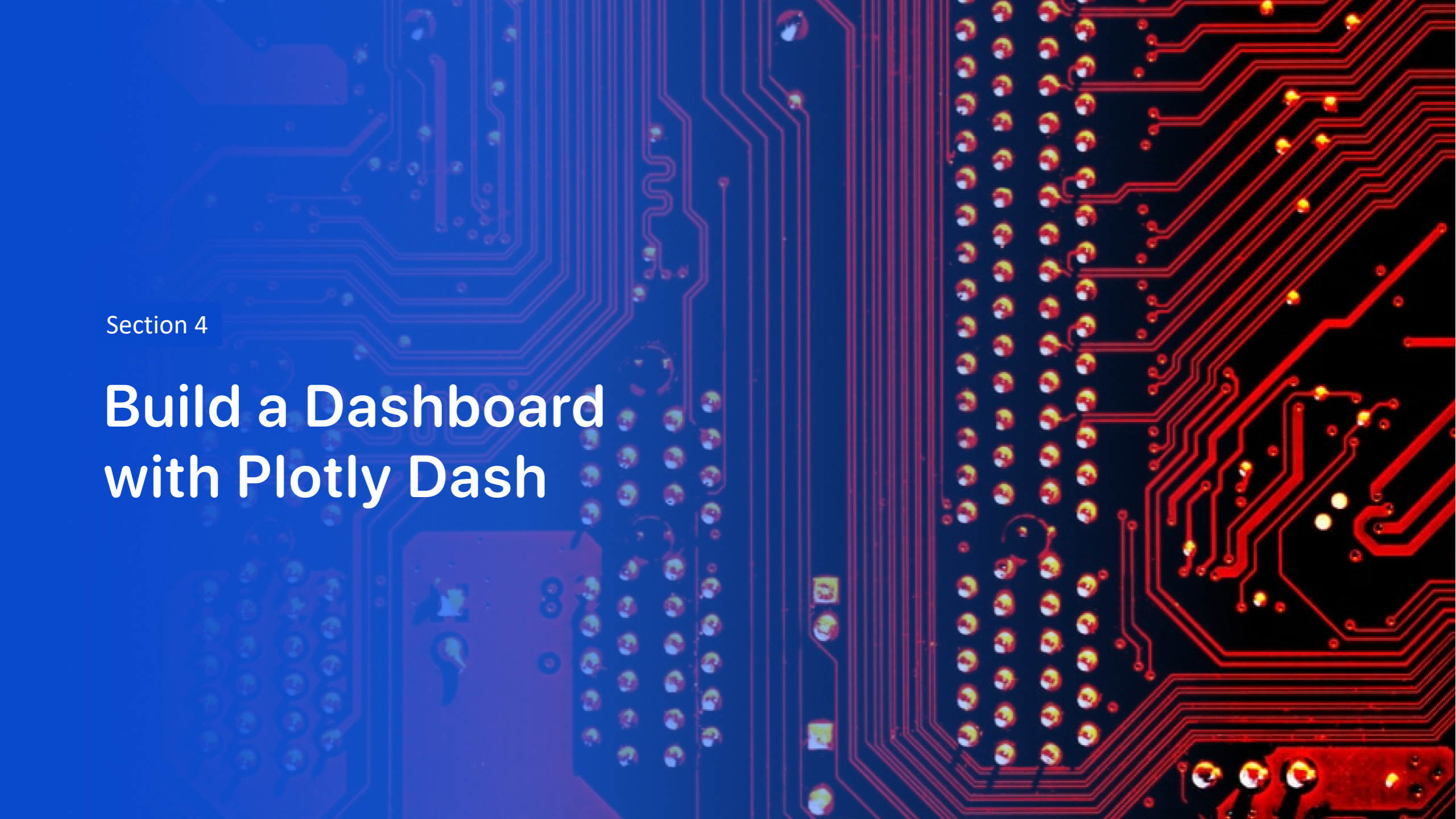
- The markers display the mission outcomes (Success/Failure = Red/Green) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site
- Launch site's success rate for Falcon 9 first stage landings can be realized from the relative number of green success markers to red failure markers.

Distance from Launch Site to Proximities

Selected Launch Site is CCAFS LC-40

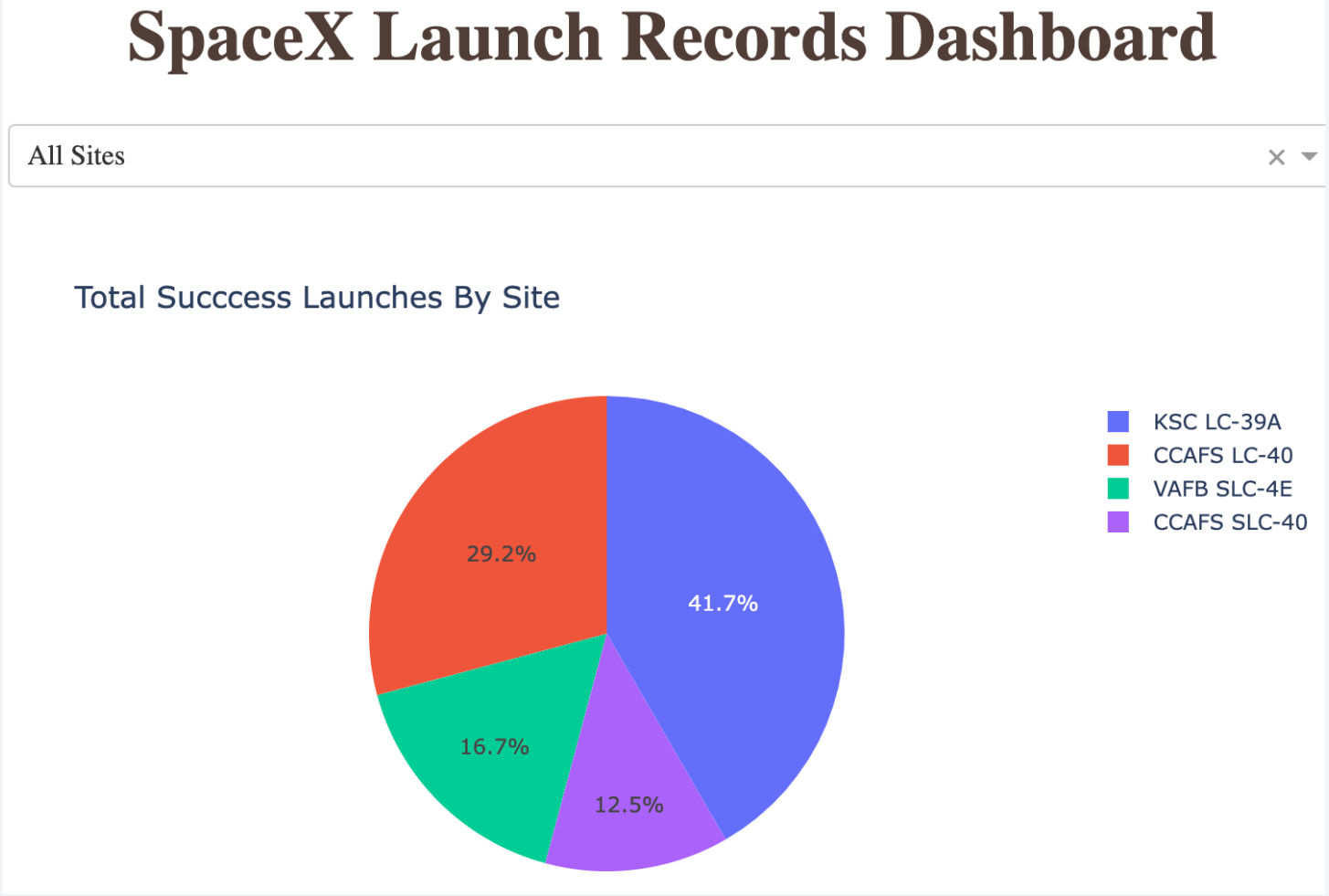
- The coastline is 0.92 km away from CCAFS LC-40
- The railway is 1,26 km away from CCAFS LC-40
- The highway is 0,95 km away from CCAFS LC-40
- The closest city is 19,26 km from CCAFS LC-40
- The CCAFS LC-40 and CCAFS SLC-40 launch sites have coordinates that are close to each other




Section 4
Build a Dashboard with Plotly Dash

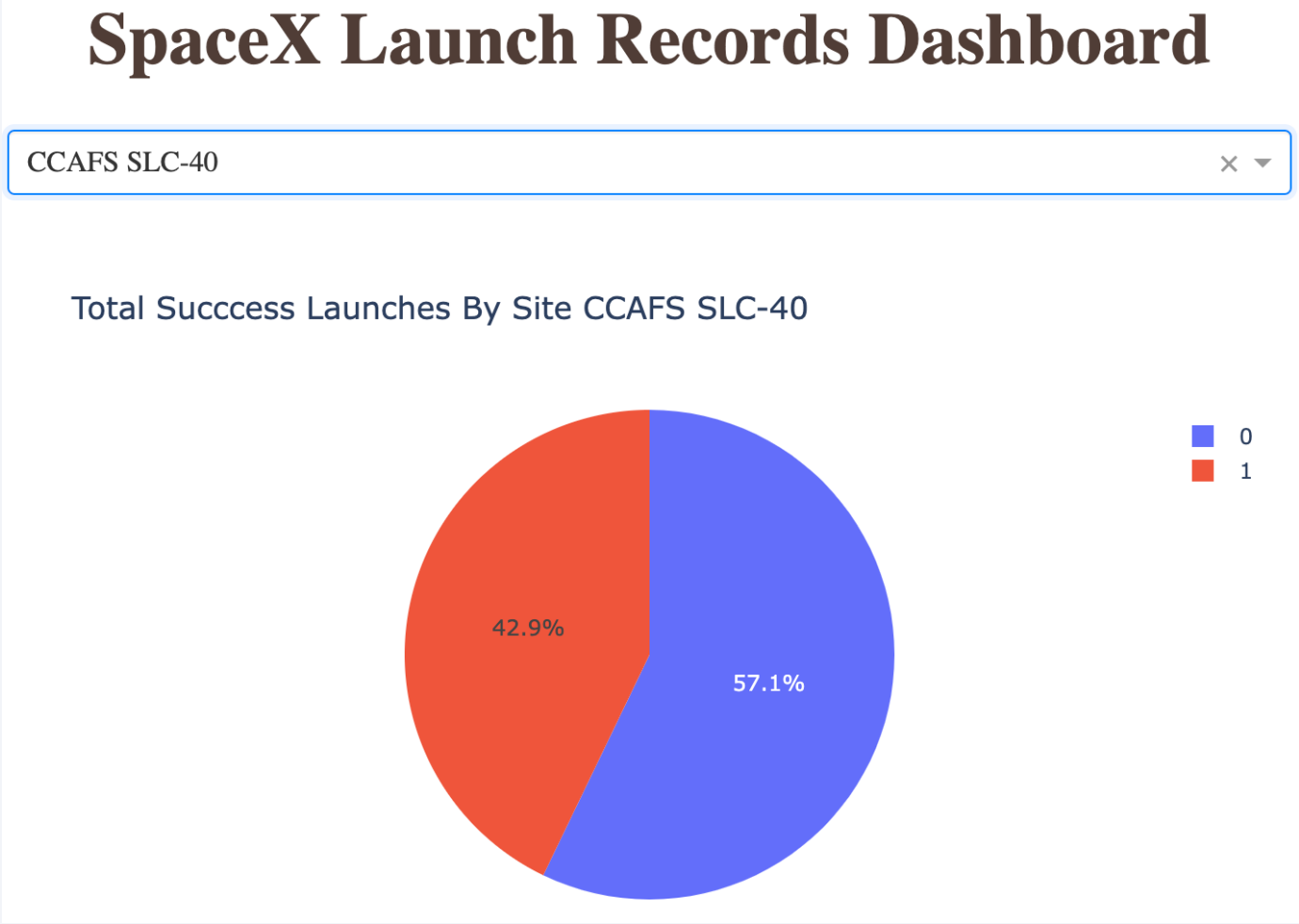
Launch Success Count for All Sites

- The dropdown menu allows the selection of one or all launch sites.
- With all launch sites selected, the pie chart displays the distribution of successful Falcon 9 first stage landing outcomes between the different launch sites.
- The greatest share of successful Falcon 9 first stage landing outcomes (at 41.7% of the total) occurred at KSC LC-39A.

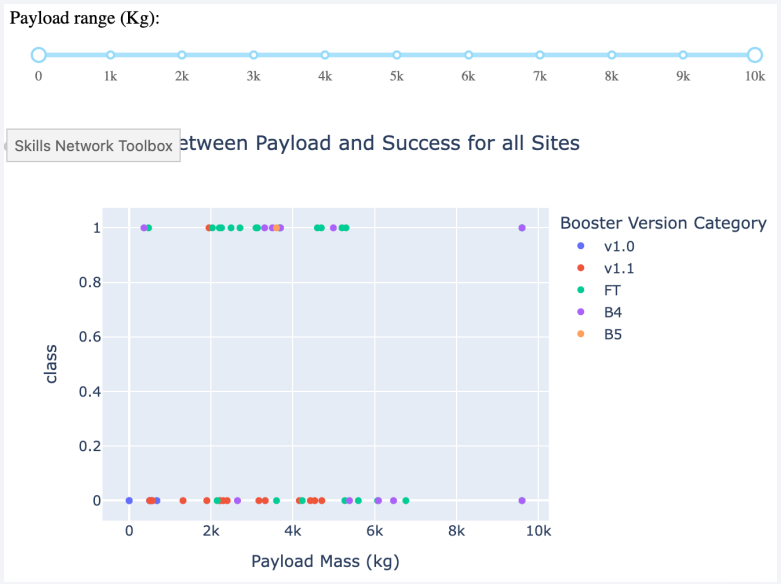


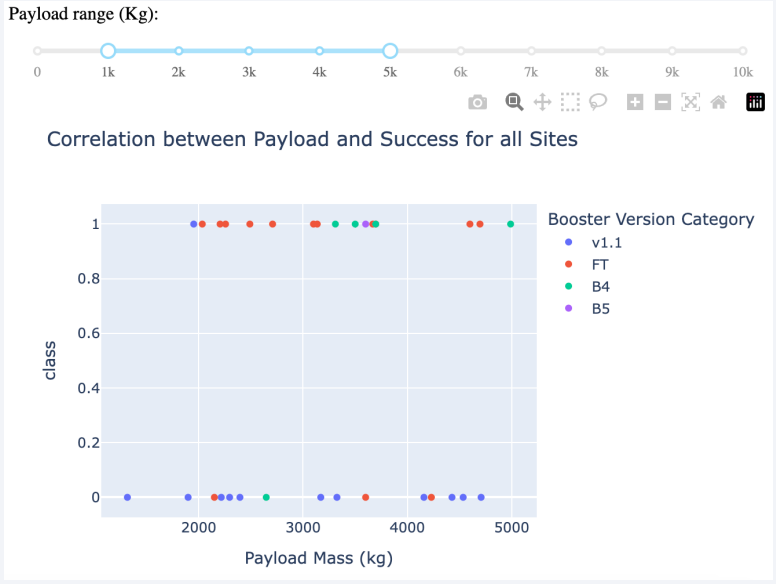
Launch Site with Highest Launch Success Ratio

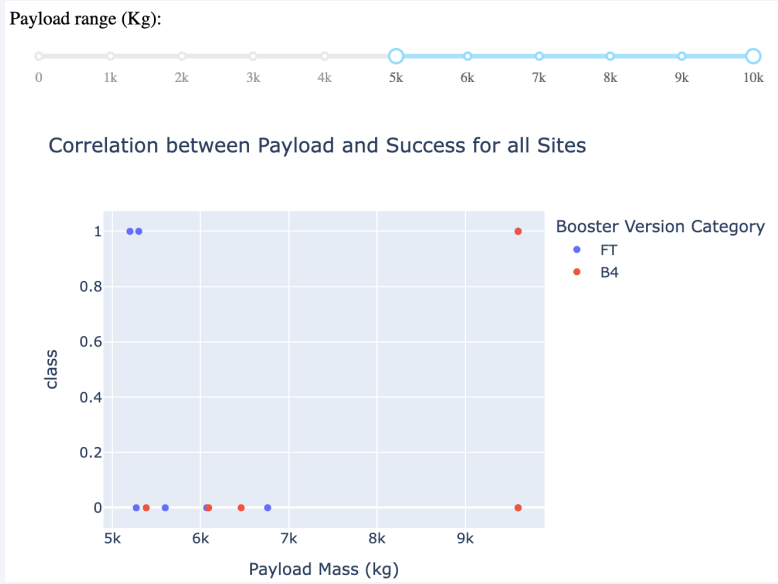
- Falcon 9 failed landings are indicated with the blue and successful with the red color inside the pie chart
- By selecting each landing site at the dropdown menu, we can see ratio of success to failed launches
- CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42.9%)



Payloads vs Launch Outcome







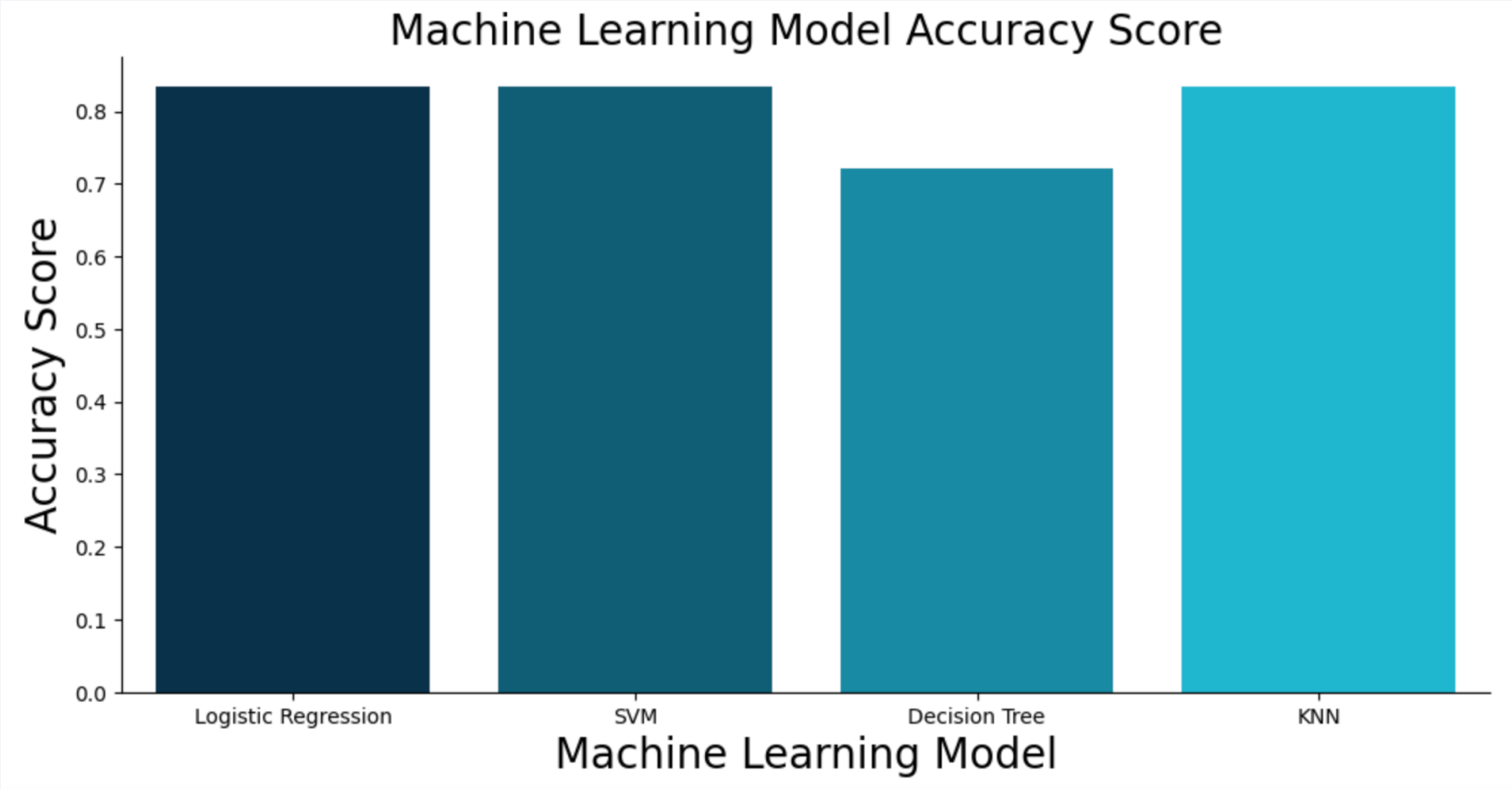
- The payload range from about 2,000 kg to 5,000 kg has the largest success rate
- The payload range from 5,000 kg to 10,000 kg has very low success rate
- FT booster version has the largest success rate with low and high payloads

Section 5

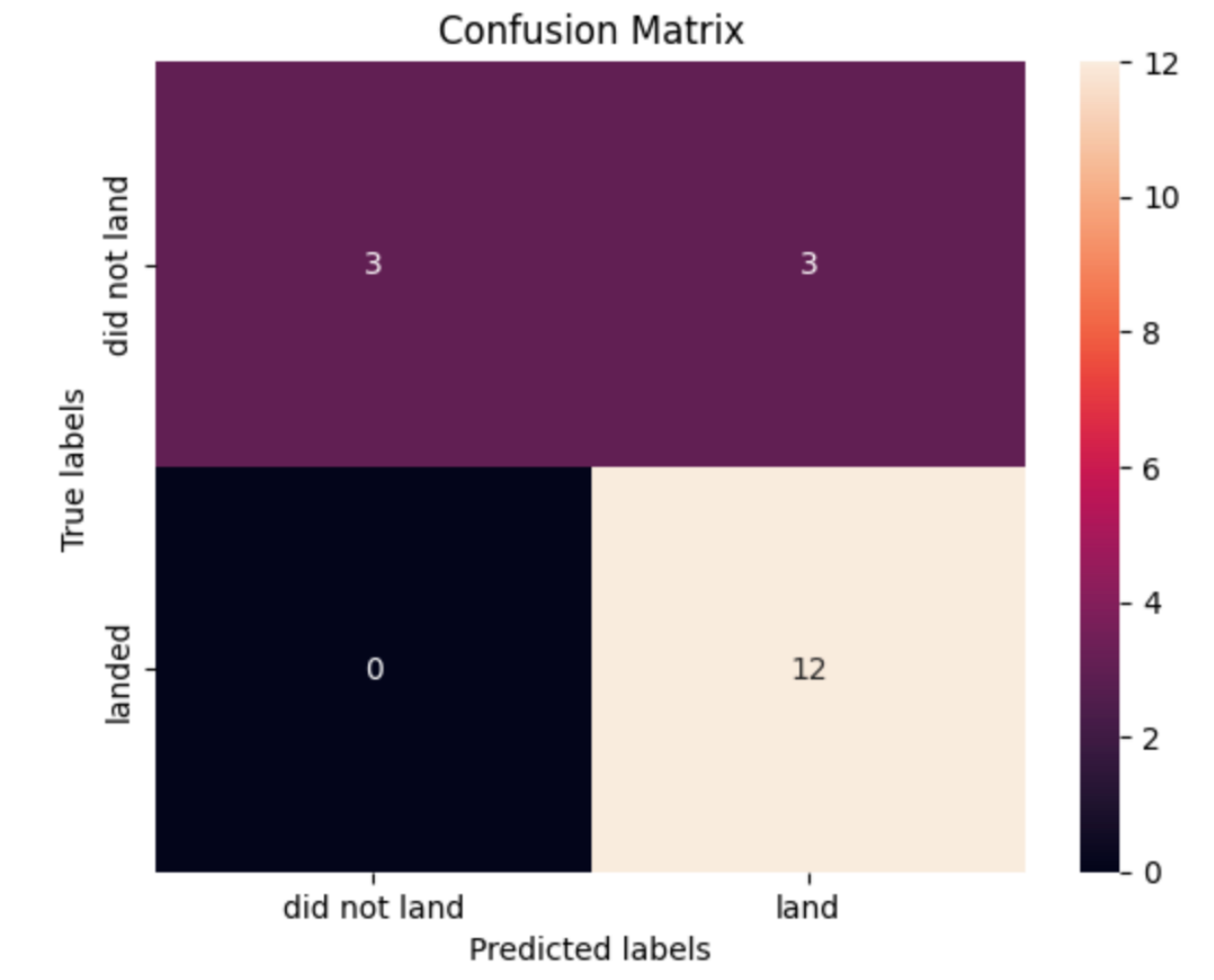
Predictive Analysis (Classification)

Classification Accuracy

As per bar graph built from the Classification Accuracy values Logistic Regression, SVM and KNN models show equally high classification accuracy, while Decision Tree model is lacking the accuracy for the given sets of testing and training data



Confusion Matrix



Confusion Matrix for Logistic Regression, SVM and KNN models shows the following prediction breakdown:

- 12 True Positives
- 3 True Negatives
- 3 False Positives
- No False Negatives

These results are better in comparison with Decision Tree model

Conclusions

In order to identify the factors that are influencing successful landing of the first stage of Falcon 9 we have come to the following conclusions:

- SpaceX does not have a perfect track record of Falcon 9 first stage landing outcomes
- SpaceX's Falcon 9 first stage landing outcomes have been trending towards greater success as more launches are made
- The best launch site is CCAFS SLC-40
- Payloads, Geolocation are one of the main factors in selecting the launch site locations with successful landing output

Appendix

- Initial Data:
  - SpaceX API: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
  - Wikipedia webpage: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Jupyter Notebooks and Dashboard Python File
  - GitHub URL (Data Collection): GitHub URL of the completed SpaceX API calls notebook
  - GitHub URL (Web Scraping): GitHub URL of the completed SpaceX Webscrapping notebook
  - GitHub URL (Data Wrangling): GitHub URL of the completed Data Wrangling notebook
  - GitHub URL (EDA with SQL): GitHub URL of the completed EDA Data Visualization notebook
  - GitHub URL (EDA with Data Visualization): GitHub URL of the completed EDA with SQL notebook
  - GitHub URL (Folium Maps): GitHub URL of the completed Interactive Map with Folium Notebook
  - GitHub URL (Dashboad File): GitHub URL of the completed Dashboard with Plotly Dash Notebook
  - GitHub URL (Machine Learning): GitHub URL of the completed Predictive Analysis Notebook

Appendix

- Data Sets:
  - https://github.com/DanielFukson/DF-spacex-ds-capstone-project/blob/main/1.1_dataset_part_1.csv
  - https://github.com/DanielFukson/DF-spacex-ds-capstone-project/blob/main/2.1_dataset_web_scraped.csv
  - https://github.com/DanielFukson/DF-spacex-ds-capstone-project/blob/main/3.1_dataset_part_2.csv
  - https://github.com/DanielFukson/DF-spacex-ds-capstone-project/blob/main/5.1_dataset_part__3.csv

Thank you!