IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<KOTOUNNOU Prudence>
<21 juin 2023>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis using SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

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. 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. The goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers
  - Will the Falcon 9 first stage land successfully.
  - What is the price of each launch ?
  - Will the SpaceX reuse the first stage

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - DatafromSpaceXwasobtainedfrom2sources:
    - SpaceXAPI(https://api.spacexdata.com/v4/rockets/)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Perform summarizing to show some characteristics, create a landing outcome label based on outcome data
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Using machine learning to determine if the first stage of Falcon 9 will land successfully. Split data into training data and test data to find the best Hyperparameter for SVM, Classification Trees, and Logistic Regression.

Data Collection

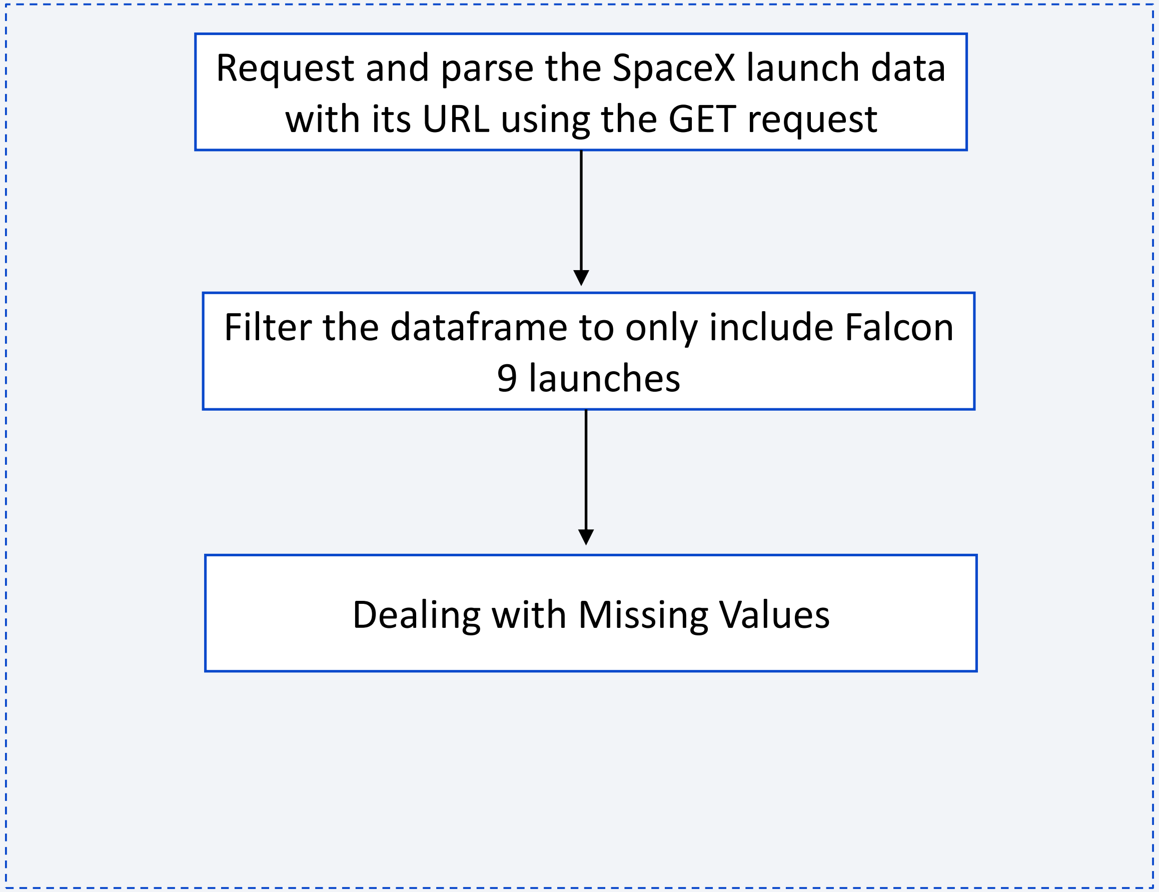
- Describe how data sets were collected.
- Datasets were collected from SpaceX API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- SpaceX has a public API from where data can be obtained and then used. This API was used to get the data and then wrangle it according to the flowchart beside

- GitHub URL of the completed SpaceX API calls notebook

https://github.com/Prudence-K/data_sc_fin/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Request and parse the SpaceX launch data with its URL using the GET request

Filter the dataframe to only include Falcon 9 launches

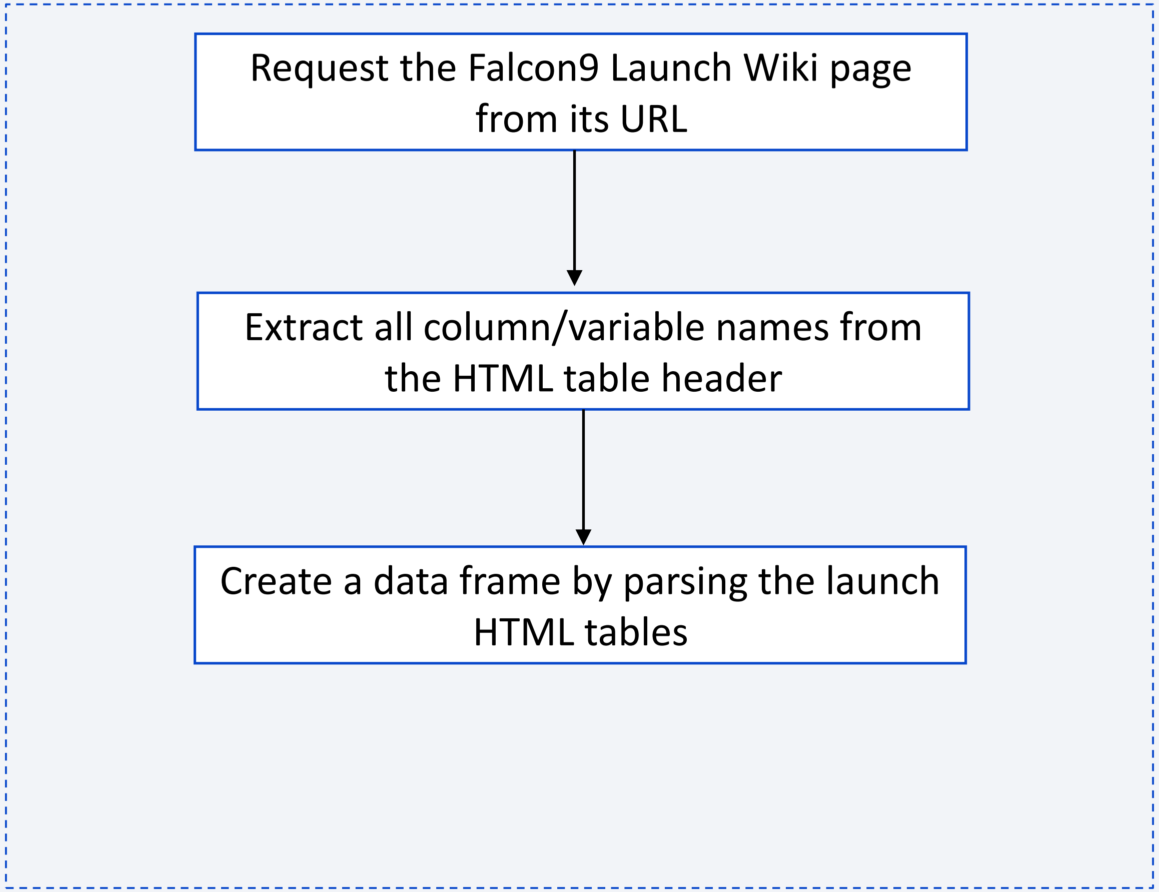
Dealing with Missing Values

Data Collection - Scraping

- Using the wiki page URL, we request the falcon9 and the extract column names from the HTLM table header. The final step is creating a data frame by parsing the launch HTLM tables

- GitHub URL of the completed web scraping notebook

https://github.com/Prudence-K/data_sc_fin/blob/main/jupyter-labs-webscraping(1).ipynb



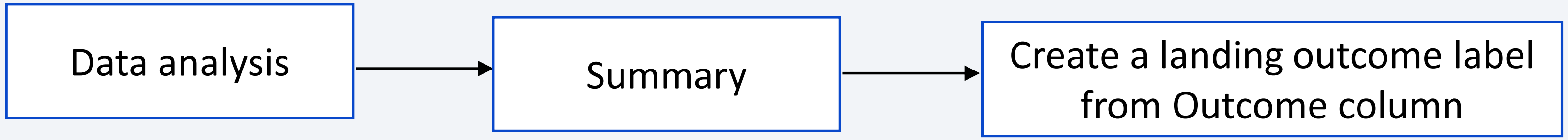
Request the Falcon9 Launch Wiki page from its URL

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- Doing some Exploratory Data Analysis (EDA) on the dataset.
- Get summary launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type.
- Create the landing outcome label.



Data analysis

Summary

Create a landing outcome label from Outcome column

- GitHub URL

https://github.com/Prudence-K/data_sc_fin/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite(1).ipynb

EDA with Data Visualization

- To explore data, scatterplots and bar plots were used to visualize the relationship between pair of features: Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit
- GitHub URL

https://github.com/Prudence-K/data_sc_fin/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- These SQL queries were performed
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begins with the string 'CCA’;
  - Total pay load mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing out comes in droneship, their booster versions, and launch site names for in year 2015; and
  - Rank of the count of landing outcomes (such as Failure (droneship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
- GitHub URL

https://github.com/Prudence-K/data_sc_fin/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
- Why those objects are added?
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site ;
  - Lines are used to indicate distances between two coordinates.

- GitHub URL

https://github.com/Prudence-K/data_sc_fin/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash where pie charts and scatter graph are plotted.
  - Pie charts show the total launches by a certain sites
  - Scatter graph show the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- GitHub URL

  https://github.com/Prudence-K/data_sc_fin/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Load data using numpy and pandas, transformed the data, split our data into training and testing.
- Build different machine learning models and tune different hyperparameters using GridSearchCV.
- Find the best performing classification model.
- GitHub URL

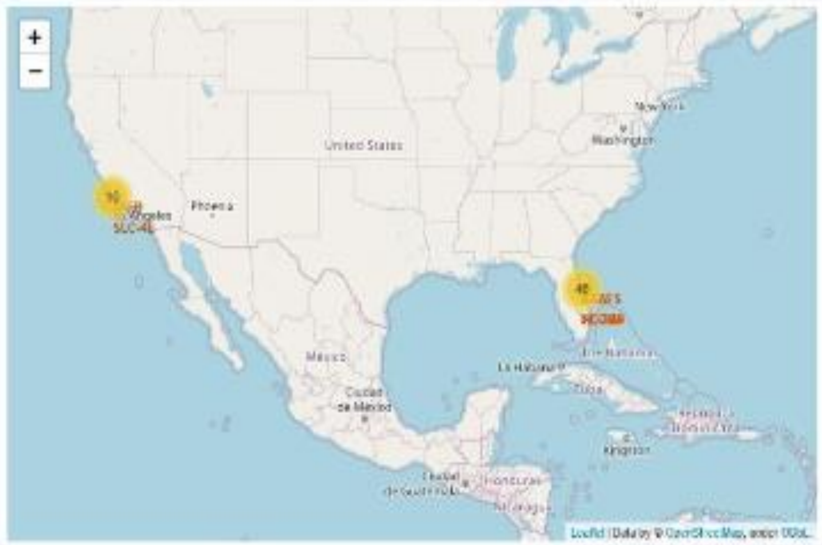
  https://github.com/Prudence-K/data_sc_fin/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

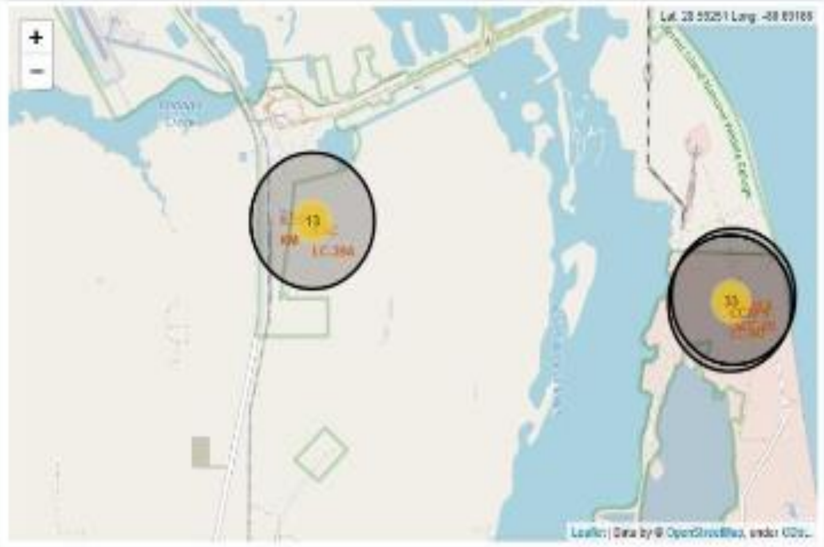
Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

Results

- Interactive analytics demo in screenshots




+
−
Lat: 28.55251 Long: -80.60188
13
LC-39A
33
Leaflet | Data by © OpenStreetMap, under ODbL.

Results

- Predictive analysis results

  Accuracy for Logistics Regression method: 0.834

  Accuracy for Support Vector Machine method: 0.834

  Accuracy for Decision tree method: 0.888

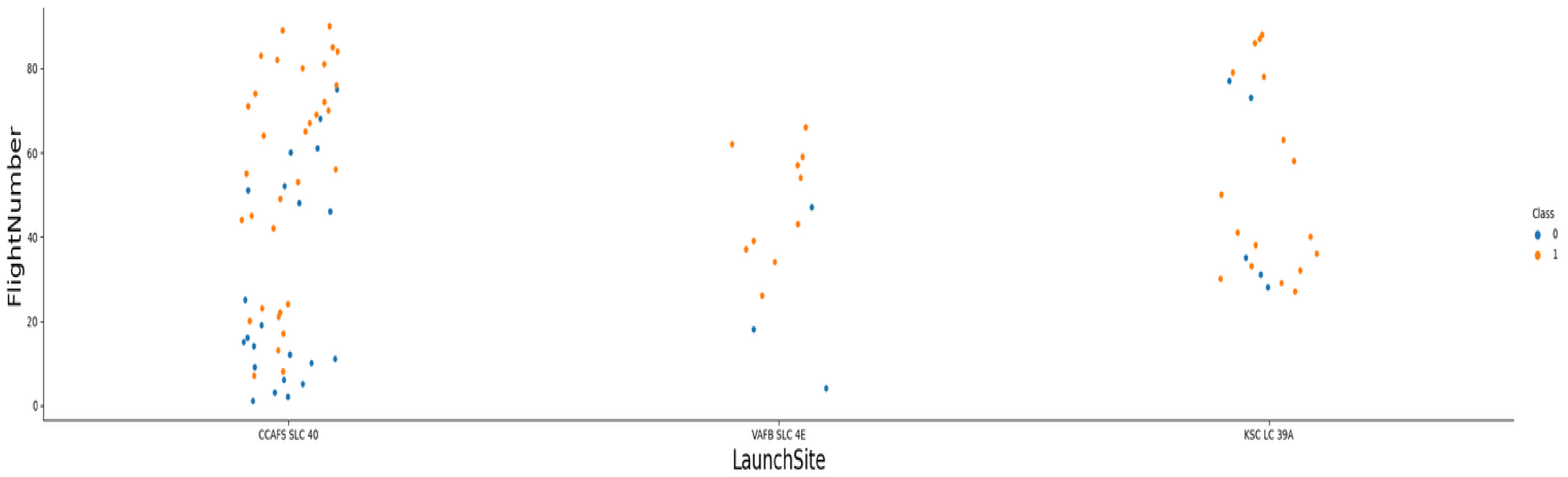
  Accuracy for K nearsdt neighbors method: 0.834

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

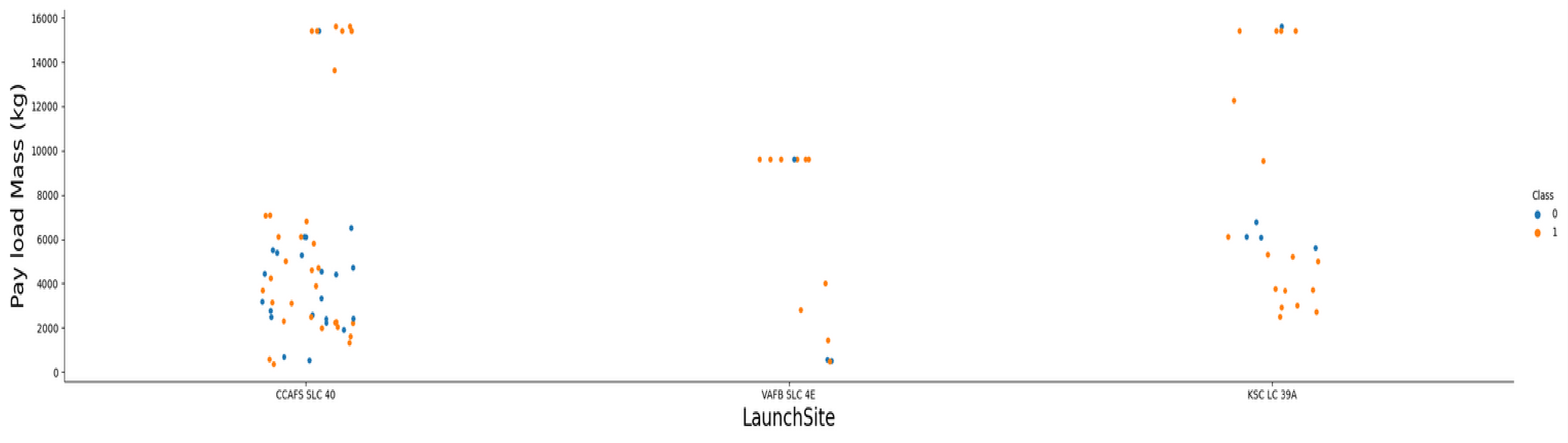
- Scatter plot of Flight Number vs. Launch Site



- the larger the flight amount at a launch site, the greater the success rate at a launch site.

Payload vs. Launch Site

- Payload vs. Launch Site

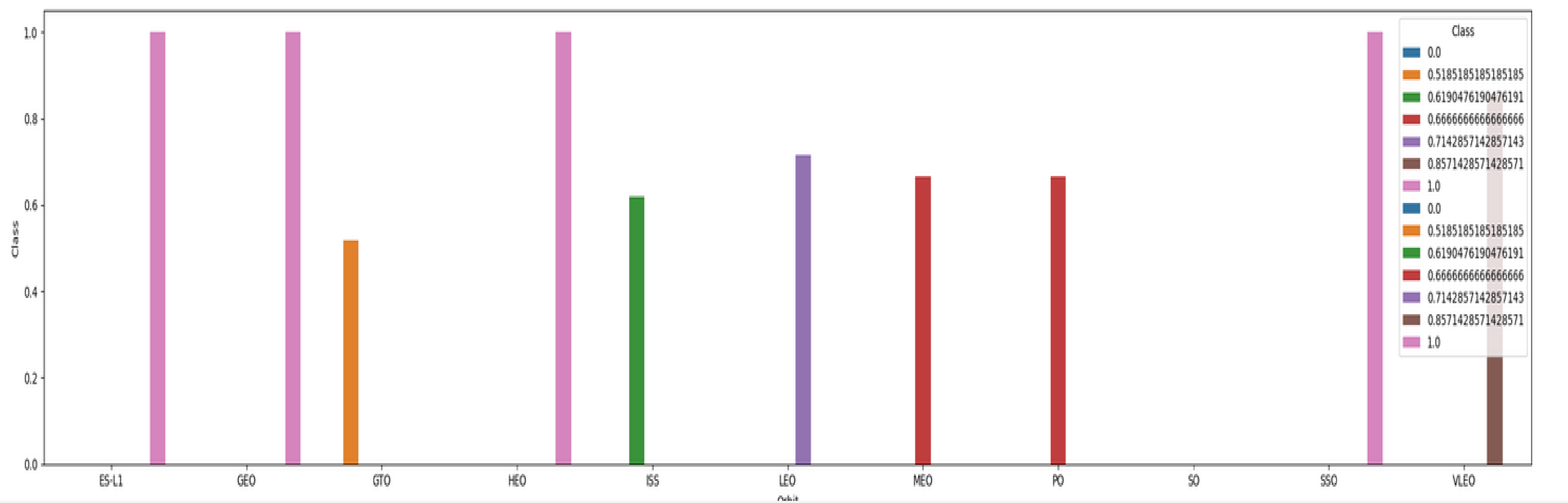


- Show the screenshot of the scatter plot with explanations

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

Success Rate vs. Orbit Type

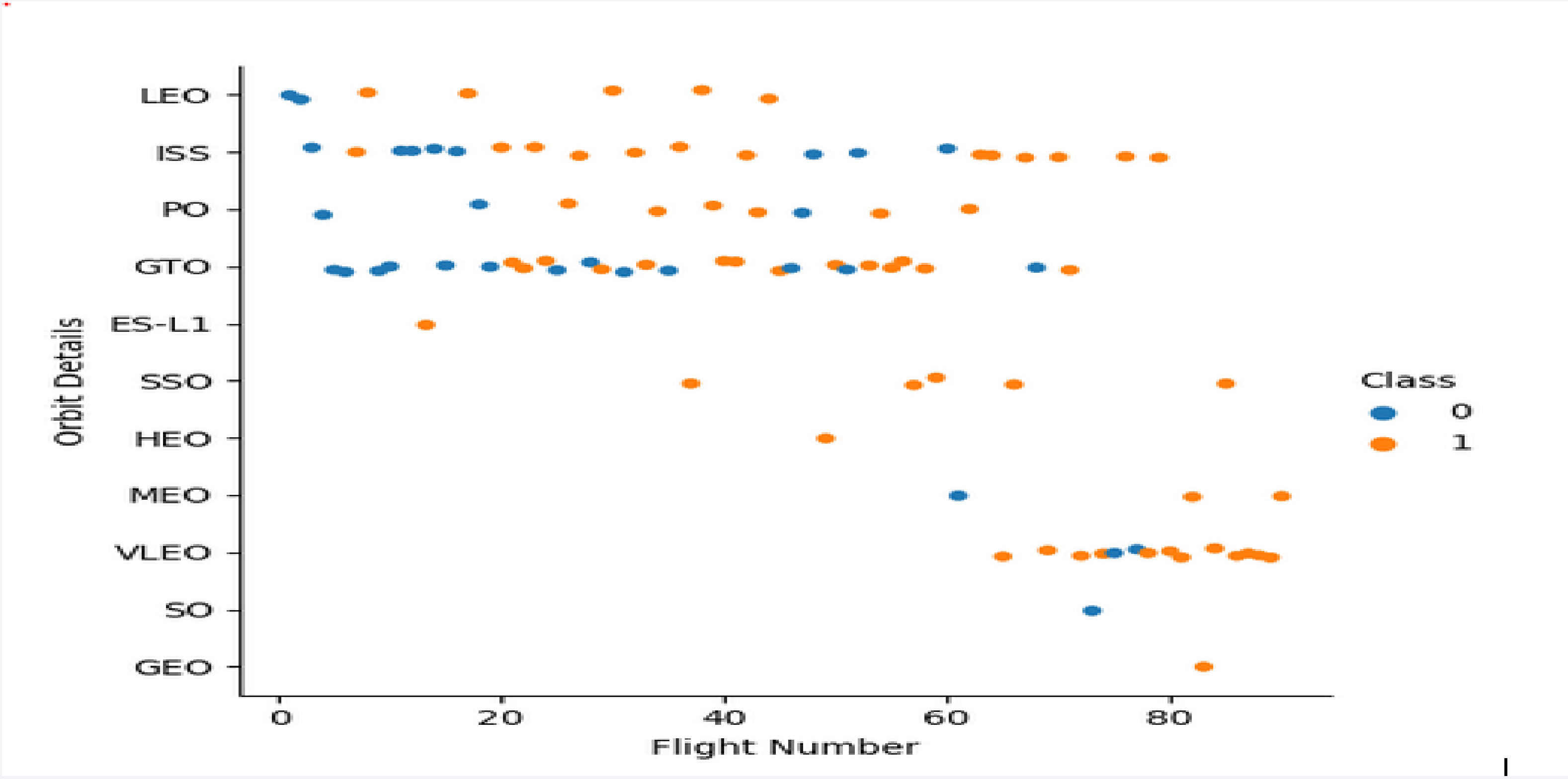
- Success rate per orbit type



- ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

Flight Number vs. Orbit Type

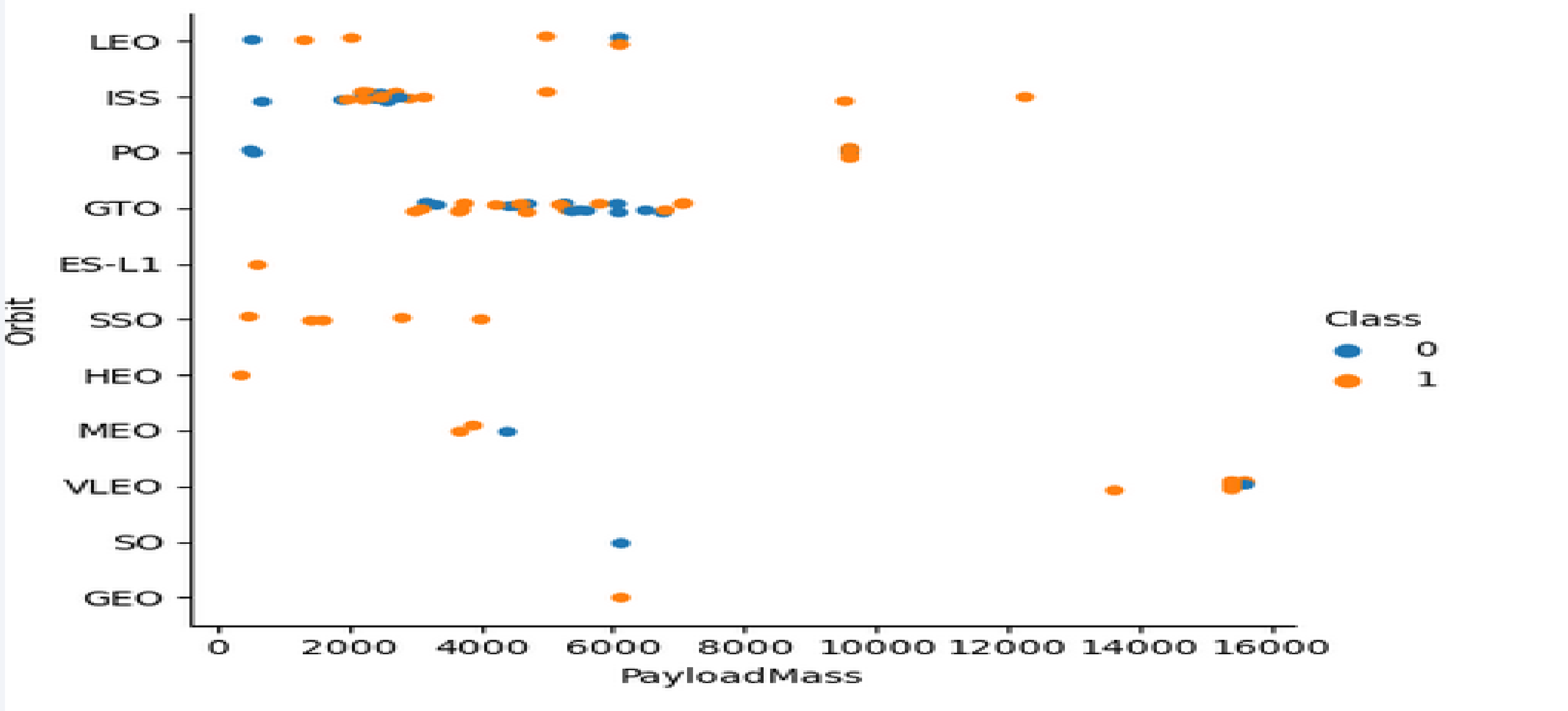
- Show a scatter point of Flight number vs. Orbit type



- the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

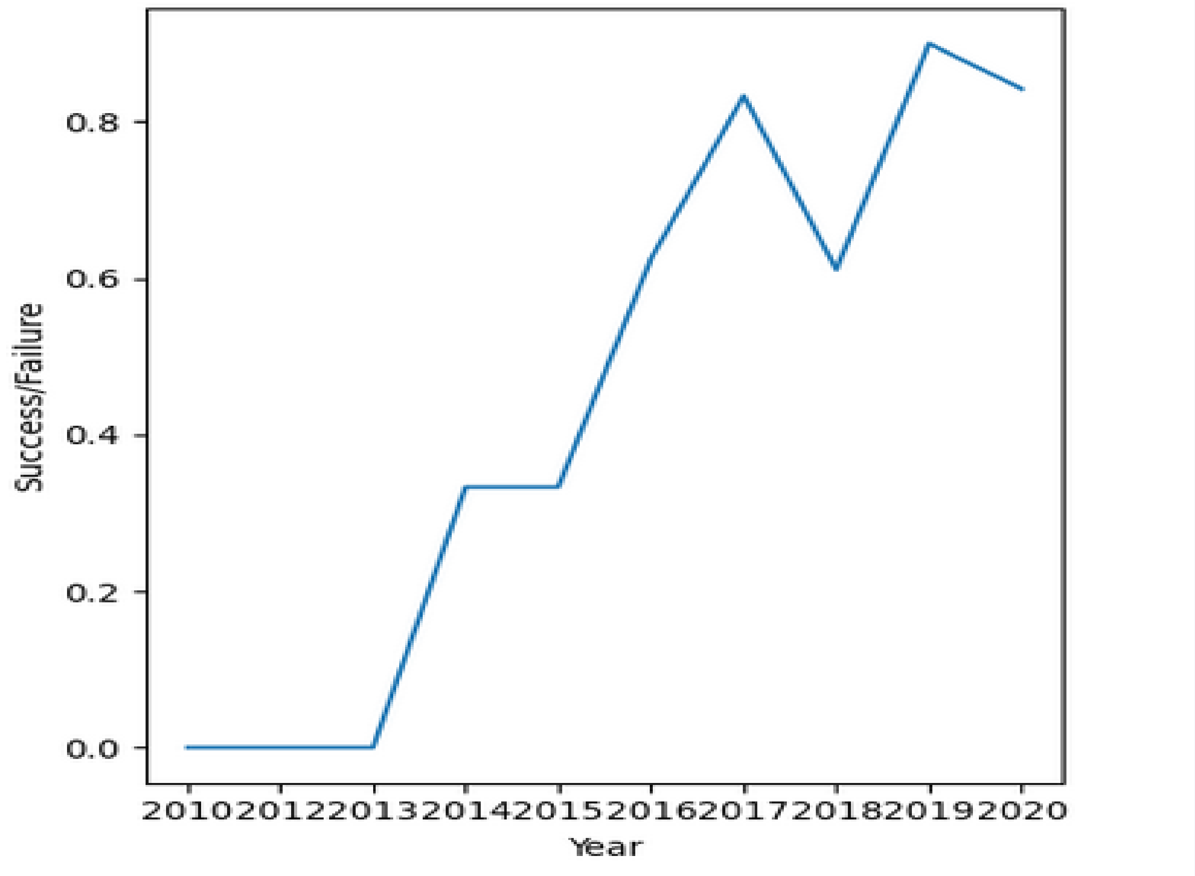
- Show a scatter point of payload vs. orbit type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



- The success rate since 2013 kept increasing till 2020

All Launch Site Names

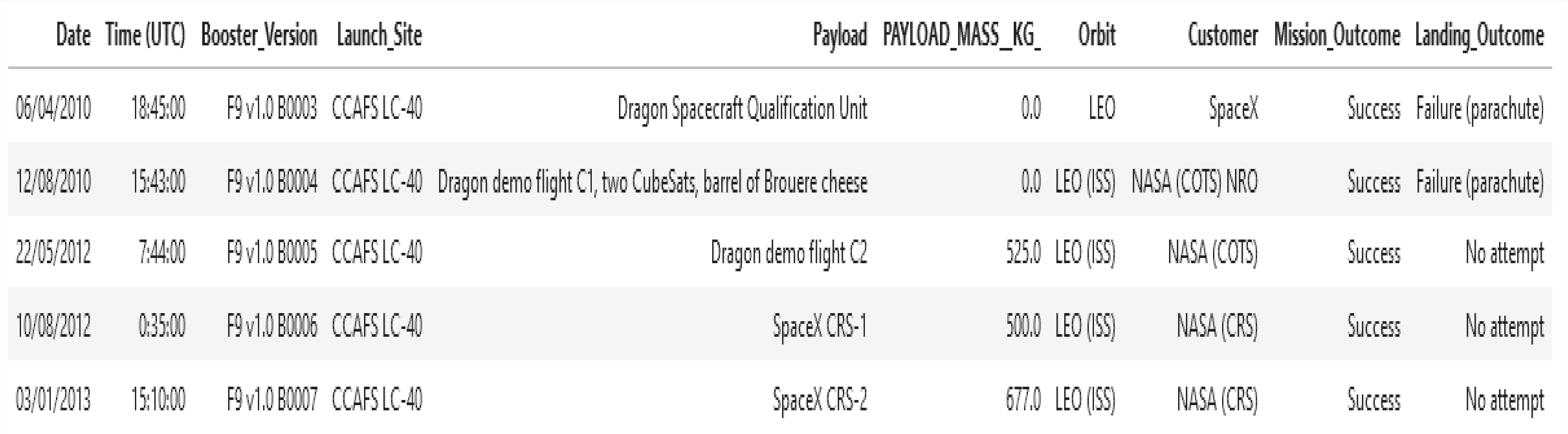
- Names of the unique launch sites

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

- There a four launch sites as shown above

Launch Site Names Begin with 'CCA'

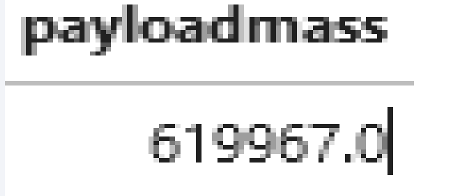
- 5 records where launch sites begin with `CCA`



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

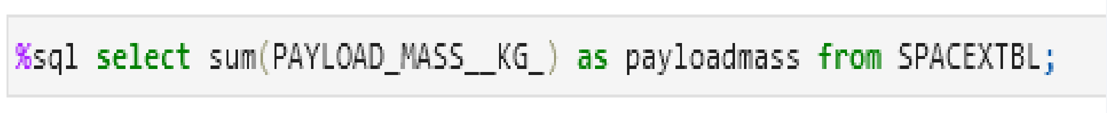
Total Payload Mass

- Total payload carried by boosters from NASA



payloadmass
619967.0

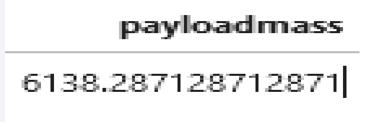
- The total payload carried by boosters from NASA is 619967. It is obtained by running the following query :



```
%sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL;
```

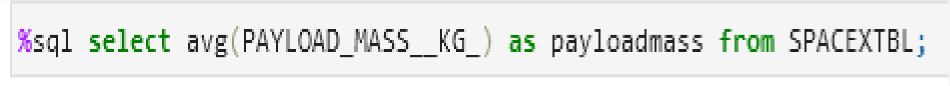
Average Payload Mass by F9 v1.1

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



payloadmass
6138.287128712871

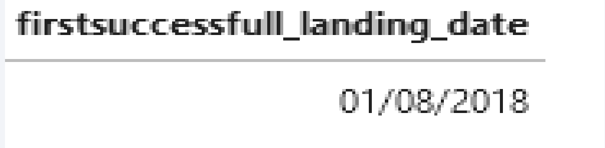
- The average payload mass carried by booster version F9 v1.1 is 6138,3. It is obtained by running the following query :



```
%sql select avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL;
```

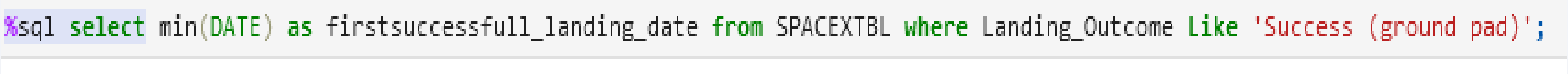
First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad



firstsuccessfull_landing_date
01/08/2018

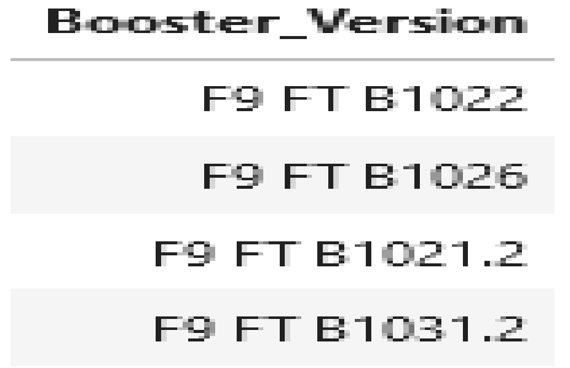
- The first successful landing outcome on ground pad was on 01/08/2018. The result is obtained by executing the following query :



```
%sql select min(DATE) as firstsuccessfull_landing_date from SPACEXTBL where Landing_Outcome Like 'Success (ground pad)';
```

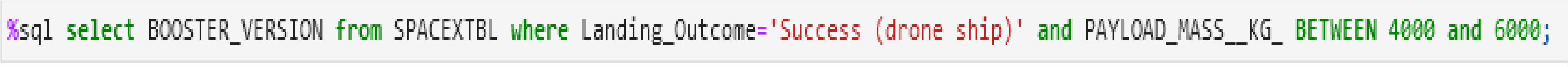
Successful Drone Ship Landing with Payload between 4000 and 6000

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



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

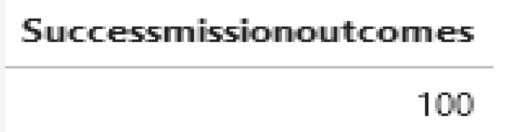
- 4 boosters have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000. The query is



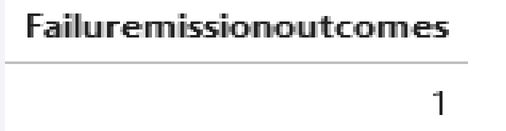
```
%sql select BOOSTER_VERSION from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes



Successmissionoutcomes
100

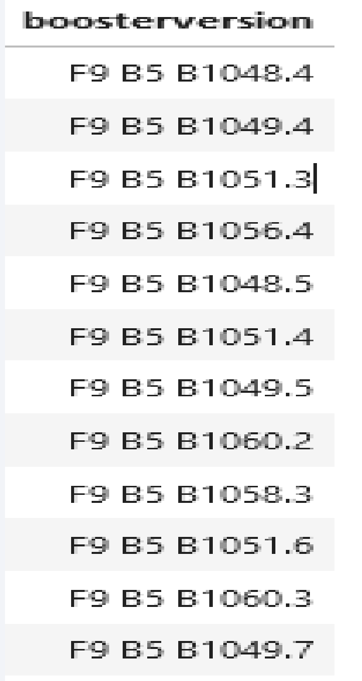


Failuremissionoutcomes
1

- We used wildcard like '%' to filter for **WHERE** MissionOutcome was a success or a failure.

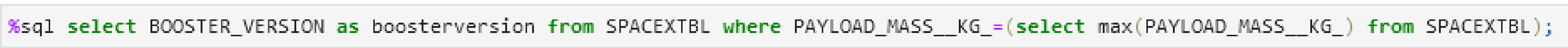
Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass



boosterversion
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

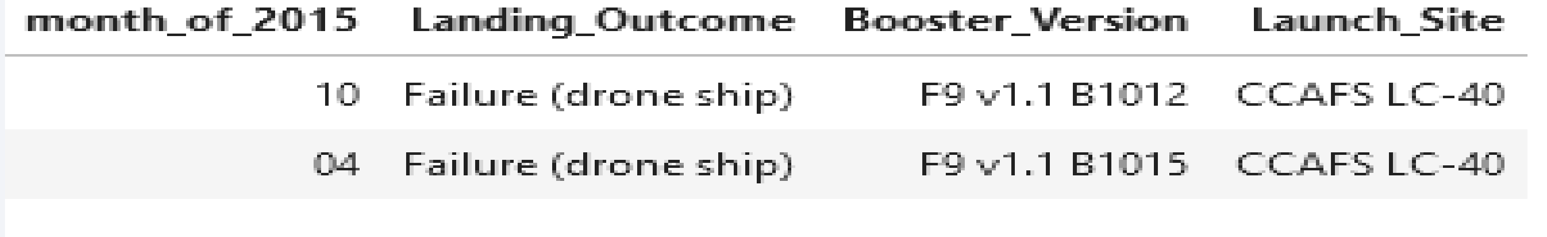
- 12 booster have carried the maximum payload mass. The result is obtained by executing the following query:



```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

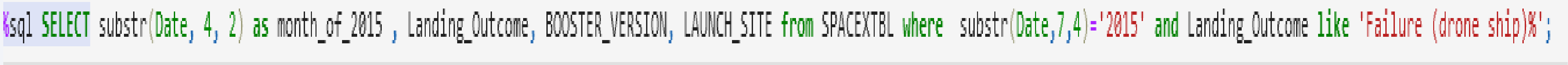
2015 Launch Records

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



month_of_2015	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

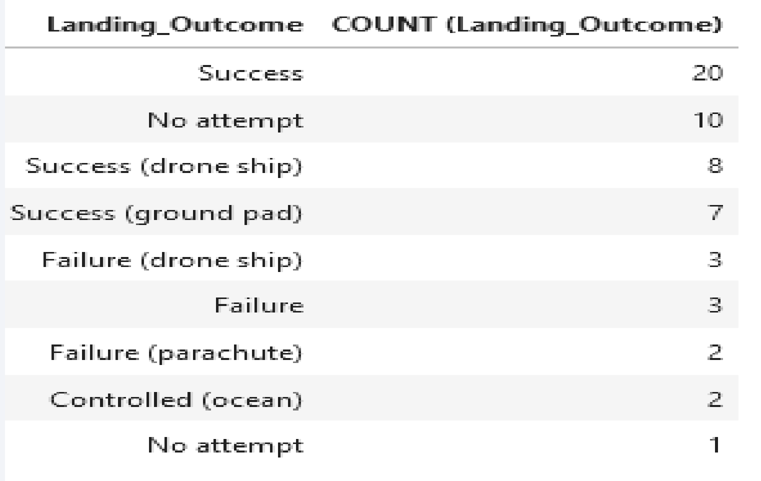
- There are 2 drones ship which fail in 2015. The query is



```
%sql SELECT substr(Date, 4, 2) as month_of_2015 , Landing_Outcome, BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL where  substr(Date,7,4)='2015' and Landing_Outcome like 'Failure (drone ship)%';
```

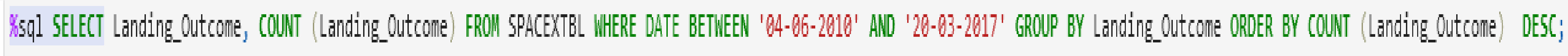
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order



Landing_Outcome	COUNT (Landing_Outcome)
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

- We selected Landing outcomes and the COUNT of landing outcomes from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2010-03-20. then, we applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

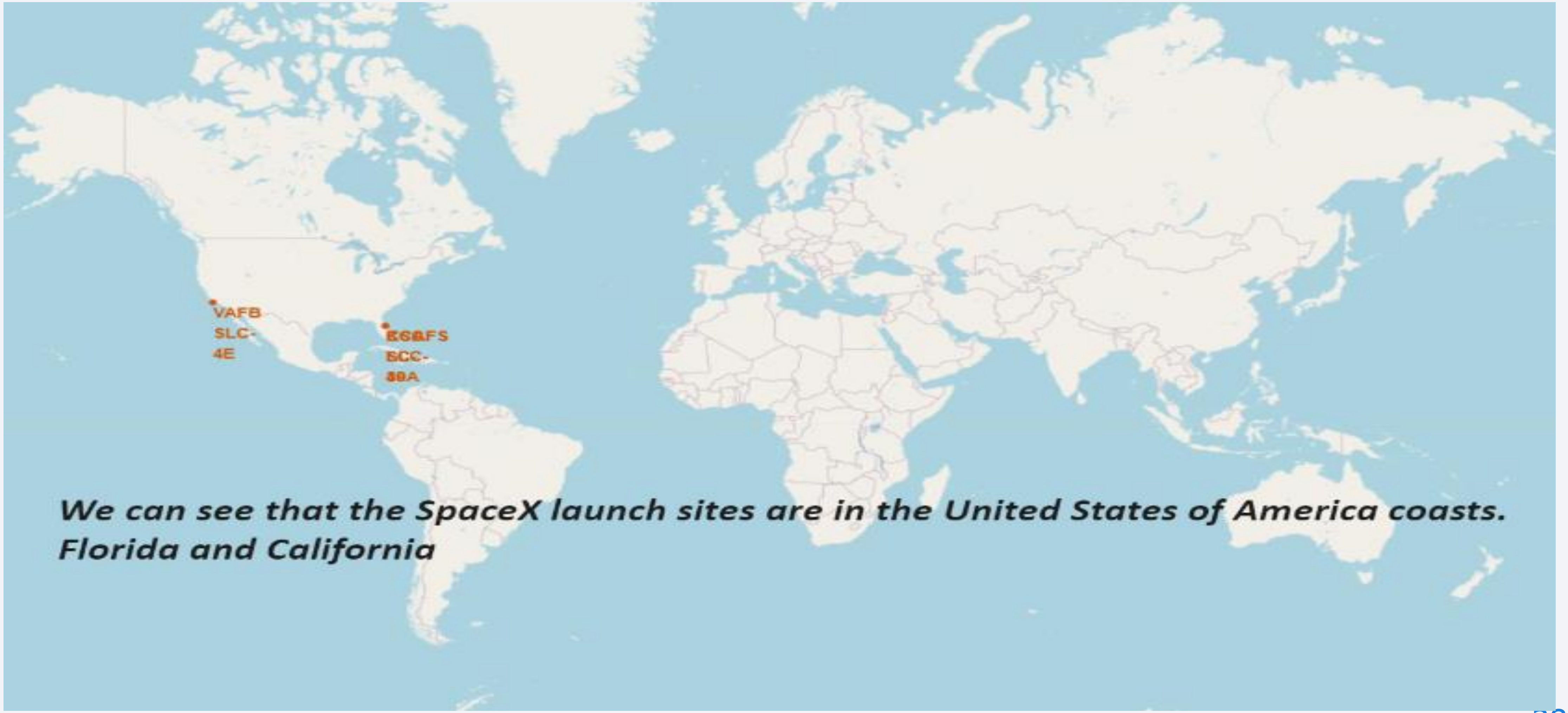


```
%sql SELECT Landing_Outcome, COUNT (Landing_Outcome) FROM SPACEXTBL WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY Landing_Outcome ORDER BY COUNT (Landing_Outcome)  DESC;
```

Section 3

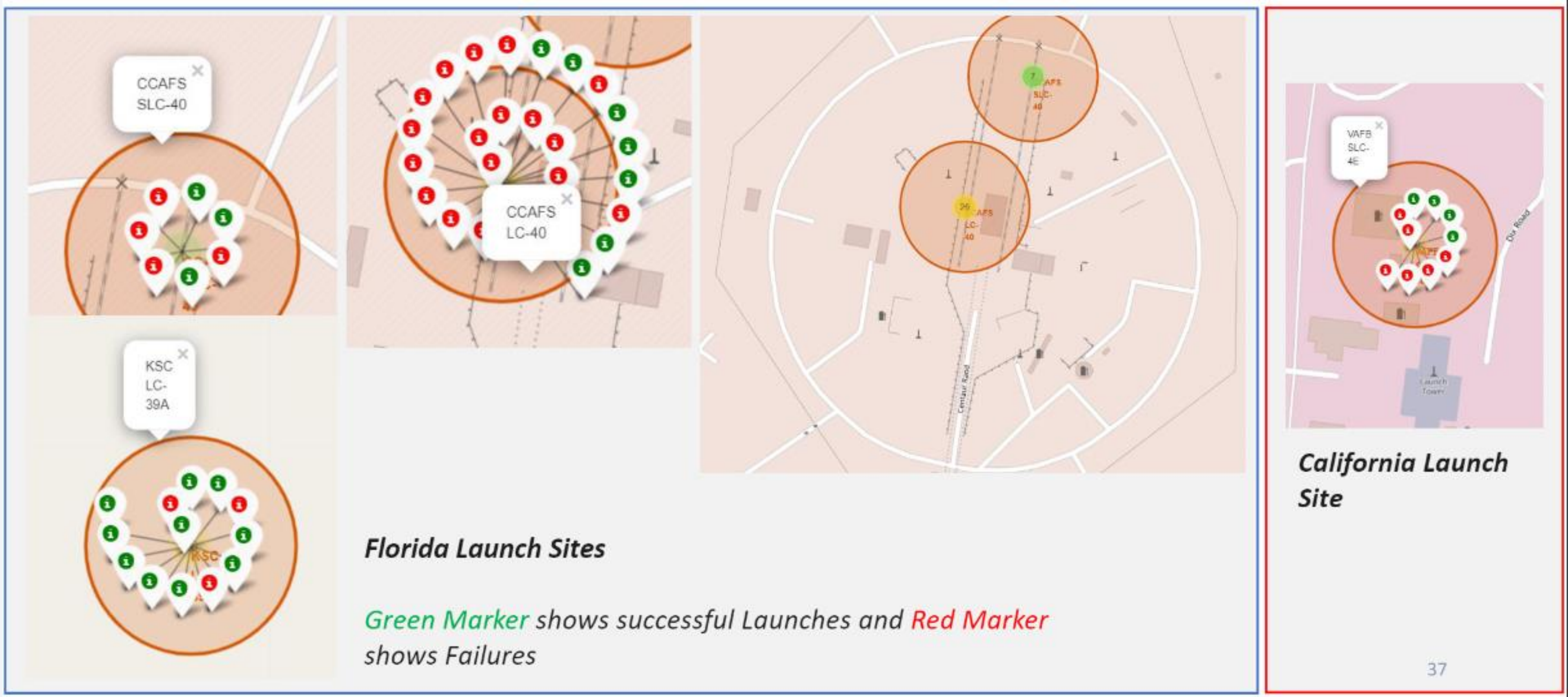
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>



We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

<Folium Map Screenshot 2>

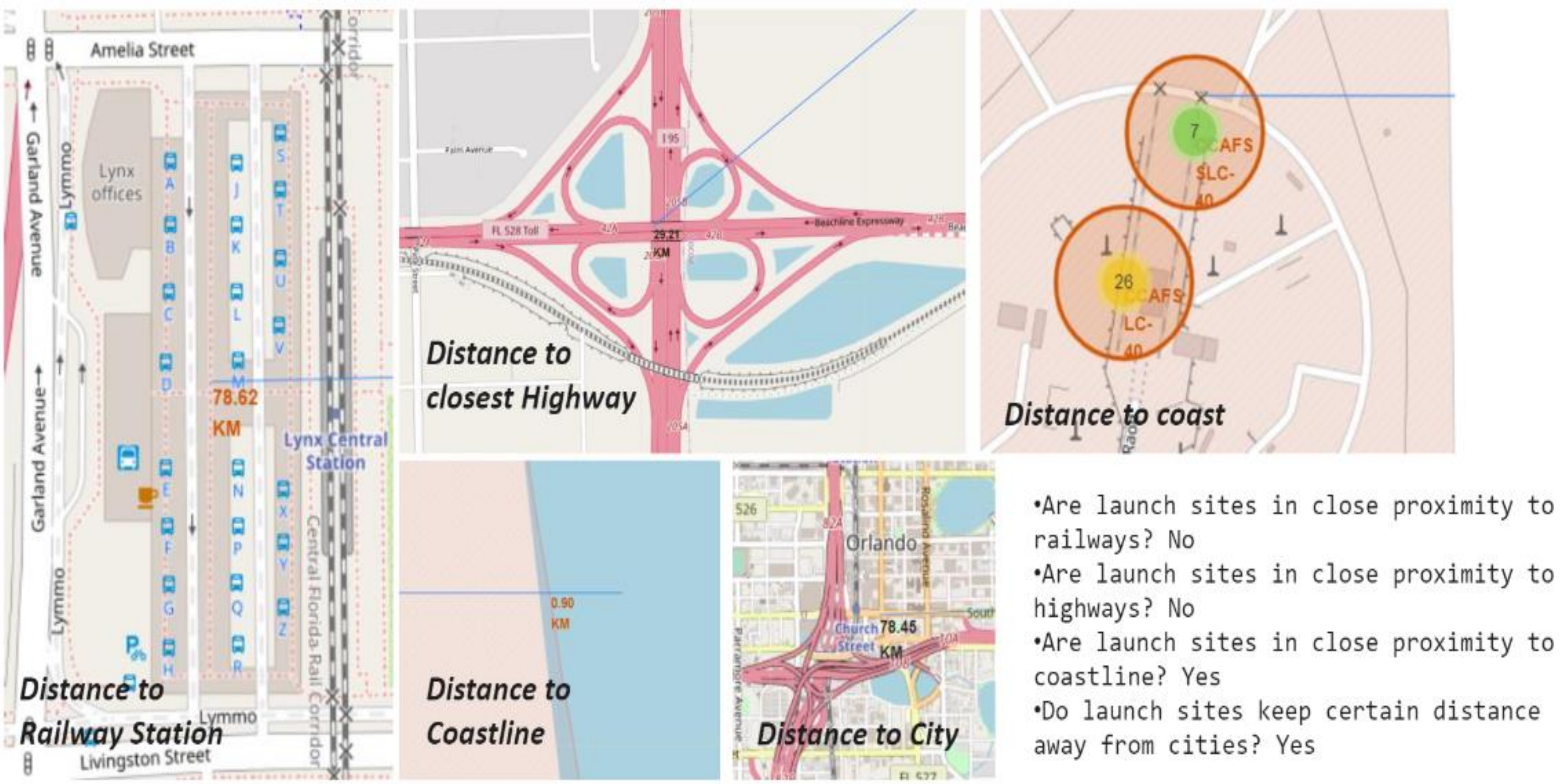


Florida Launch Sites

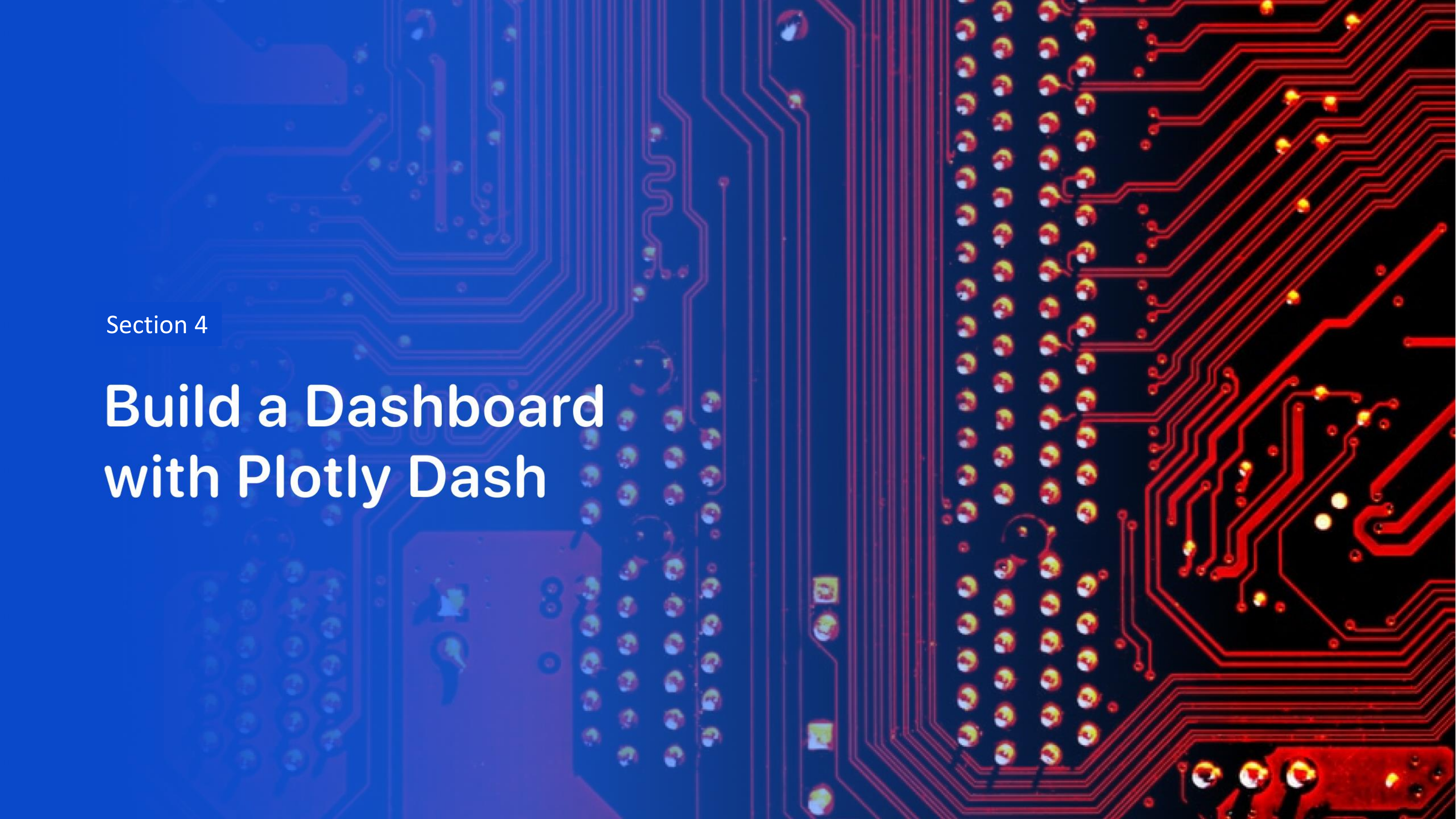
Green Marker shows successful Launches and Red Marker shows Failures

California Launch Site

<Folium Map Screenshot 3>



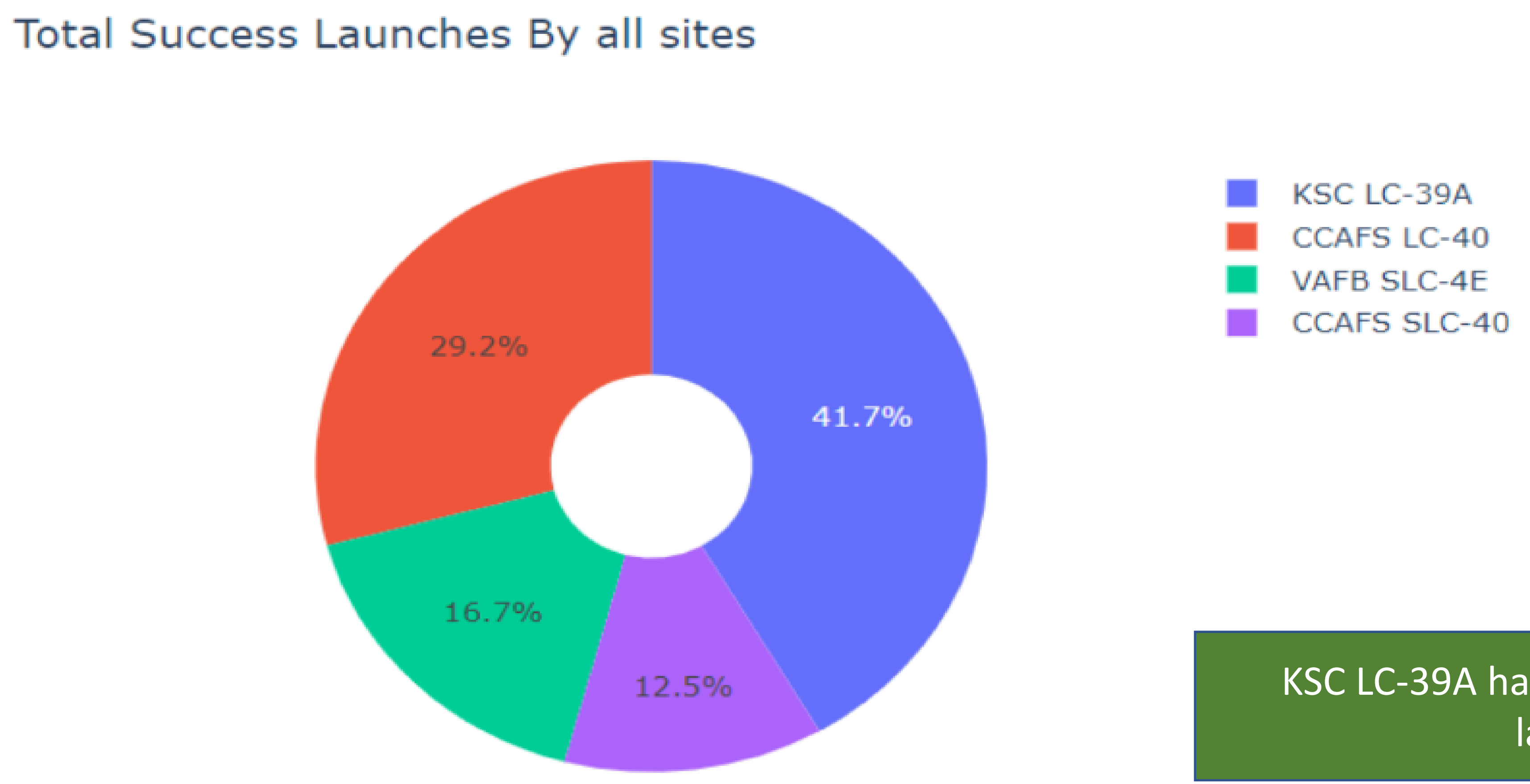
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes



Section 4

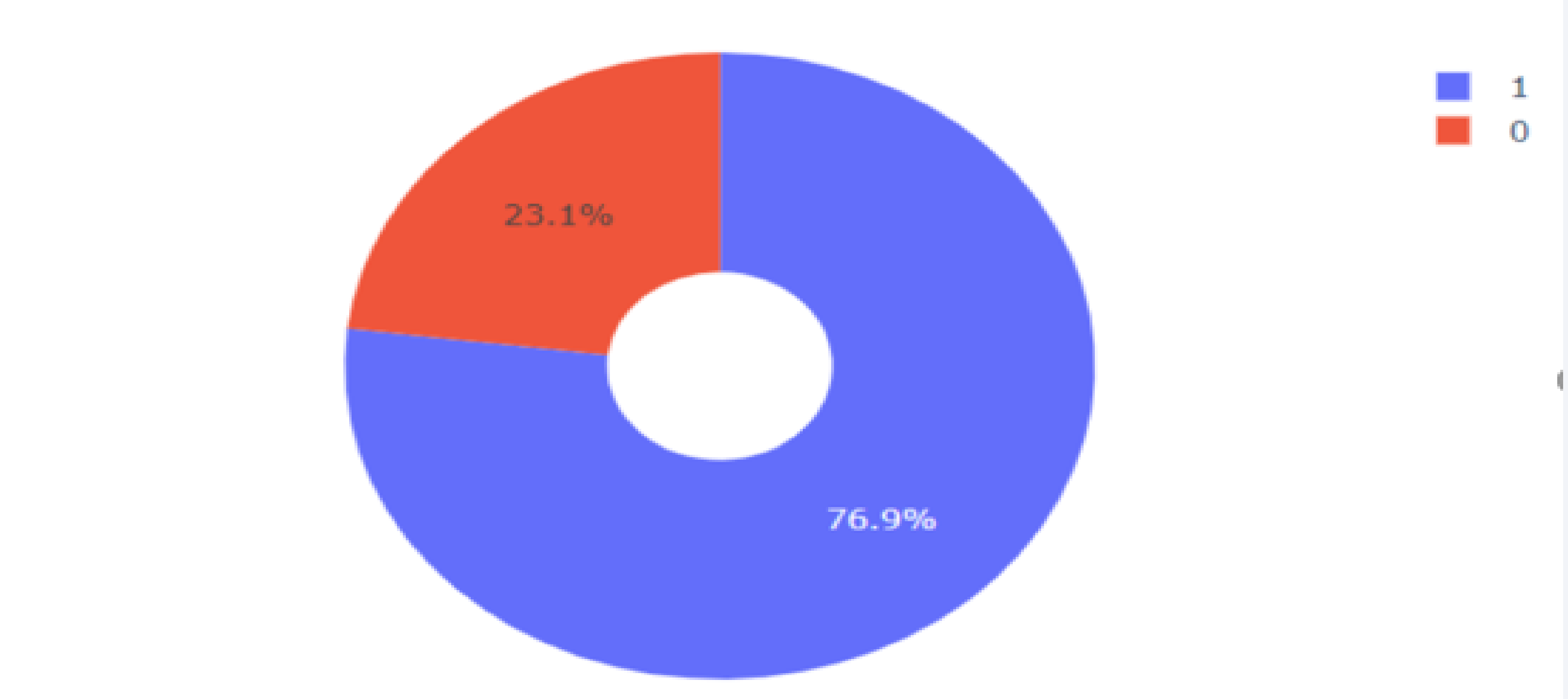
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

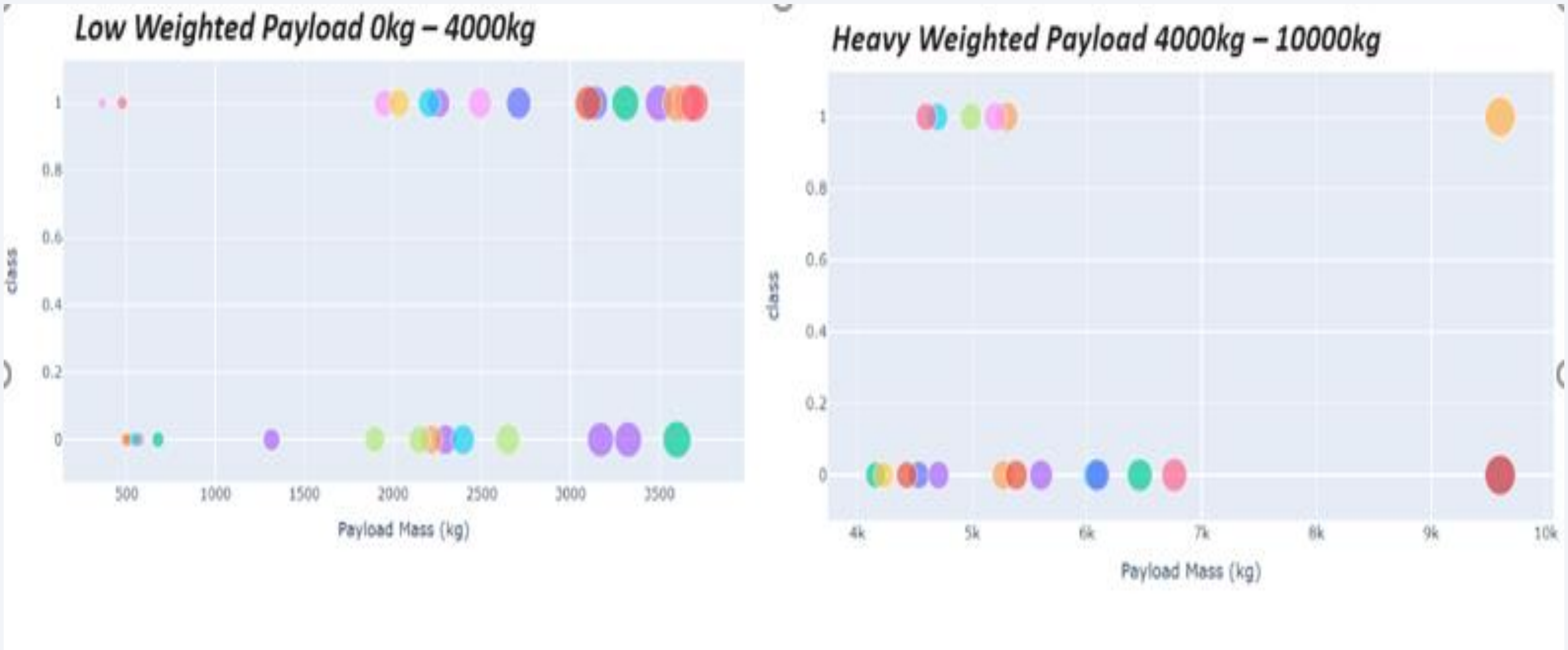


KSC LC-39A had the most successful launches

<Dashboard Screenshot 2>


1
0
23.1%
76.9%

<Dashboard Screenshot 3>



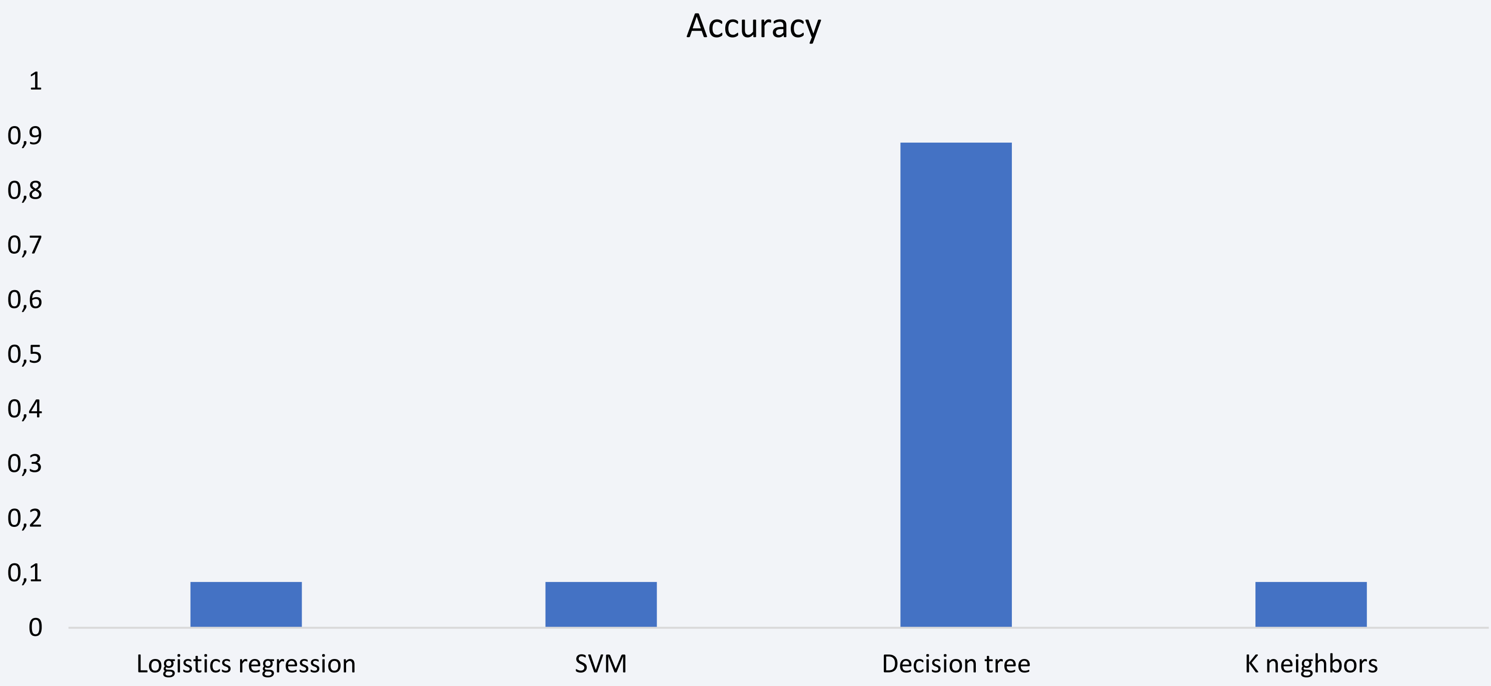
Low Weighted Payload 0kg – 4000kg

Heavy Weighted Payload 4000kg – 10000kg

Section 5

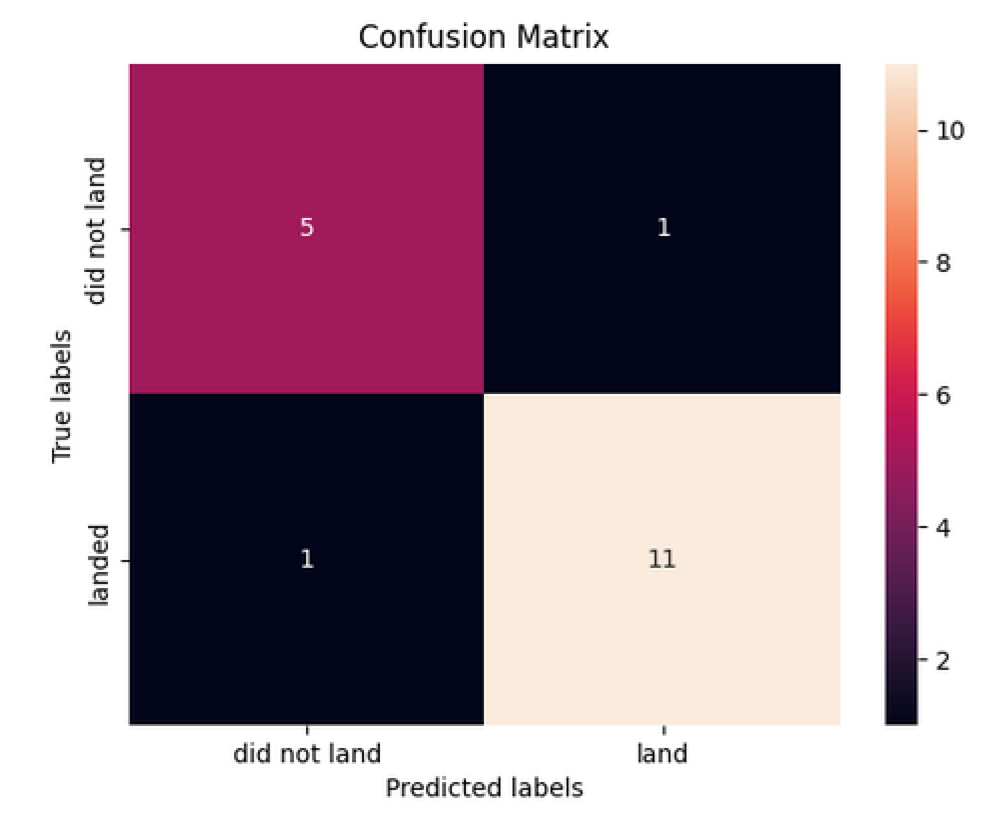
Predictive Analysis (Classification)

Classification Accuracy



The best model is decision tree

Confusion Matrix



Conclusions

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

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!