IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ravipati Gowtham Aditya
15-05-2025

Outline

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

Executive Summary

- Today, SpaceX is known as one of the top companies in the world for launching space rockets. This is because of the progress they've made in rocket technology, which has helped make space missions cheaper and more practical. While most rocket companies charge around 165 million dollars for a launch, SpaceX offers the same service for only 62 million dollars. This big cost saving has led organizations like NASA to work with SpaceX.
- In this report, I'm acting as a data scientist for a new rocket company called SpaceY, which wants to compete with SpaceX. SpaceX was started by billionaire Allon Musk. My job is to use data science—not rocket science—to find out if we can compete with SpaceX. I'm doing this by collecting information about SpaceX, analyzing the data, building machine learning models, and creating dashboards for my team.

Introduction

- Since 1957, the countries around the world are competing to expand beyond Earth, whether through satellites launches or space exploration, these missions require huge amounts of money where a launch of one space rocket costs in average 165 million dollars, a company like Space X changes the equation by reducing this amount of money massively to only 60 million dollars due to its unique and advanced technologies in returning the first stage of rocket structure.
- As mentioned above one of the key factor of SpaceX's success in the space race is the first stage of rockets return through a safe landing, from this point we will discover together what are the main attributes or variables that control these successful landings, through asking questions about the nature of the launch process , the payload mass of rocket, launch location, rocket orbit, and more by analyzing and visualizing these information through the data science methodology provided by IBM

Section 1

Methodology

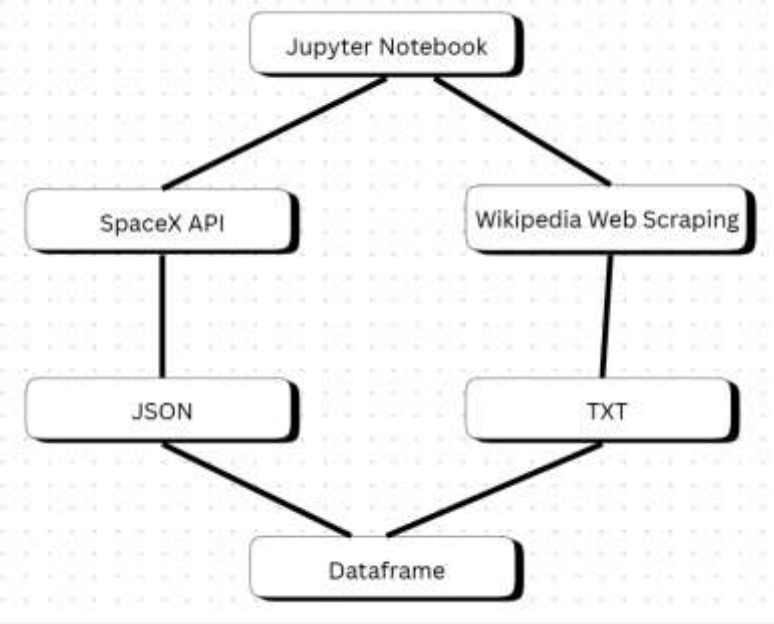
Methodology

Executive Summary

- Data collection methodology:
  - The data was collected using SpaceX rest API in addition of using data web scraping on Wikipedia webpages.
- Perform data wrangling
  - The data was preprocessed using Panadas and NumPy, some of main technique are used: OneHot encoding, unnecessary columns removal, data normalization and standardization.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Using libraries such as seaborn and matplotlib for visualization and SQL for data quires.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Using the following libraries Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Starting with splitting the data into train and test sets
  - Identifying the best algorithm and parameters through hyperparameters tuning using Grid Search
  - Adopting the best algorithm and parameters for the purpose of model deployment.

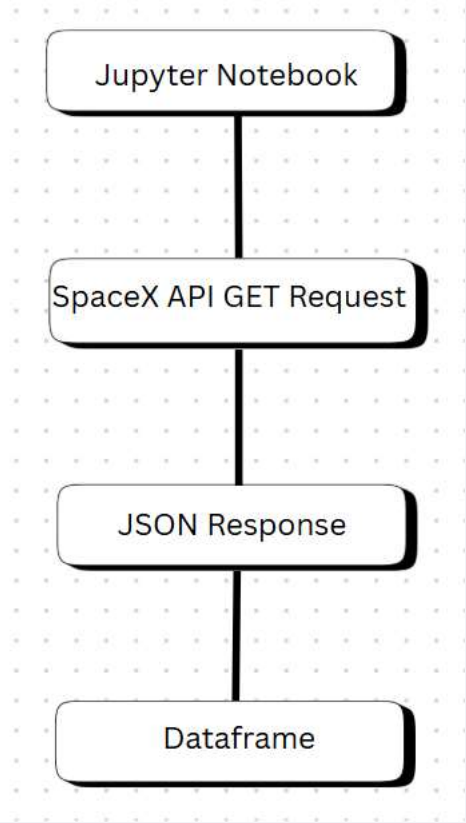
Data Collection

- We have collected the data from two main sources:
- SpaceX API: Open Source REST API for launch, rocket, core, capsule, starlink, launchpad, and landing pad data.
- Wikipedia: is a free online encyclopedia, created and edited by volunteers around the world and hosted by the Wikimedia Foundation The Process of data collection:



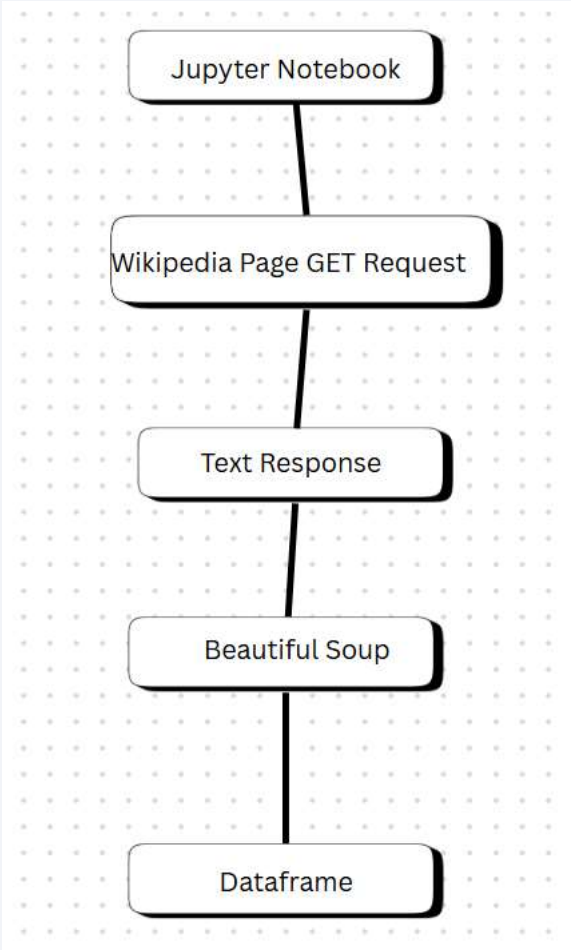
Data Collection – SpaceX API

- We started the data collection from SpcaeX API by importing the required libraries such as pandas, NumPy and Request, then we established a URL GET request, this request is raised as JSON file to be finally converted to a data frame through choosing the required information like the geospatial info, rocket type, orbit, flight number and more.
- Data Collection



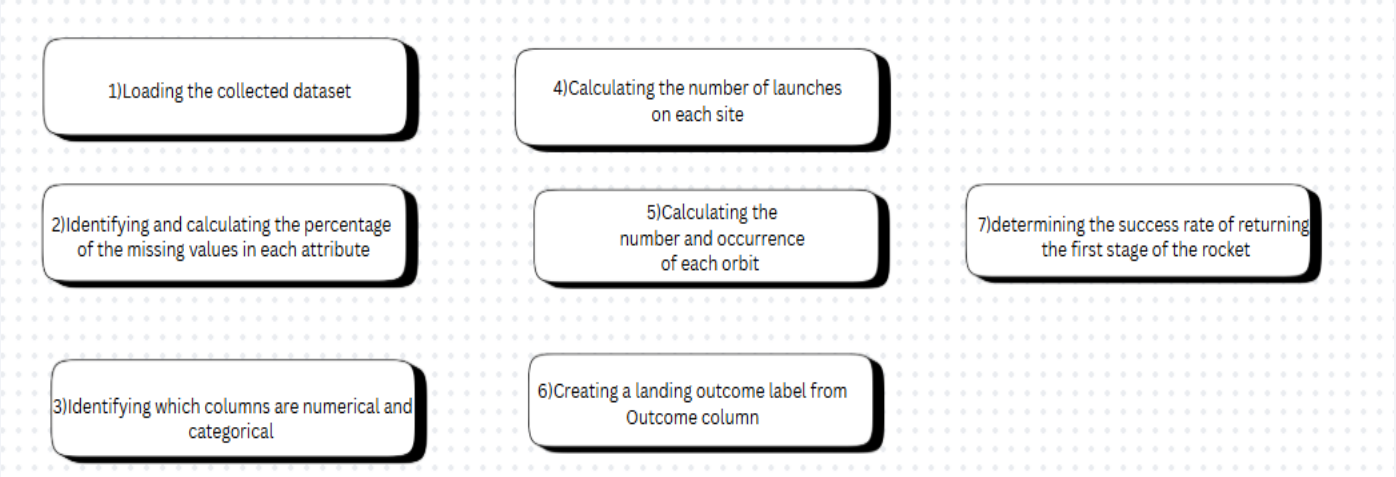
Data Collection - Scraping

- As we have done before we start by importing the required Python libraries beautiful soup and request to perform our task, and this time, we have used a webpage on Wikipedia called “Space X Falcon 9 First Stage Landing Prediction” as a data source, then we initialized an HTTP Get Request and the response was as a text format, then we used the beautiful soup library to extract the tables and columns effectively from the text response to be converted later to a pandas' data frame
- Web Scraping



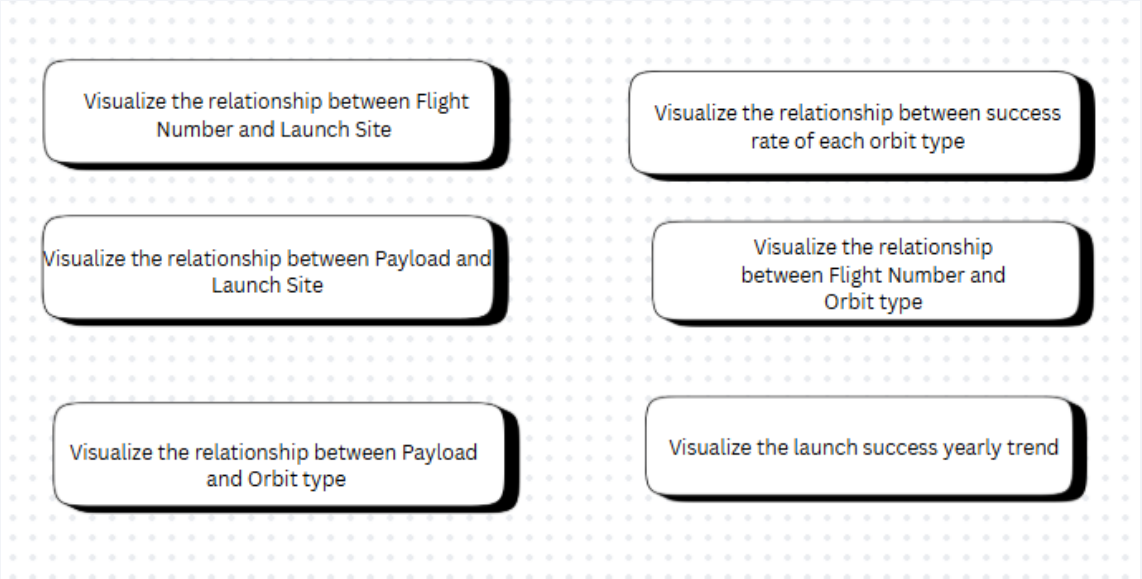
Data Wrangling

- In this stage we started by importing pandas and NumPy, loading our collected data in the previous stage to perform our exploratory data analysis which aimed to clean the data and choose the valid features for training a machine learning model.
- Data Wrangling



EDA with Data Visualization

- In this stage we completed our EDA process through finding the correlation between the features and the target using different visualization tools via seaborn and matplotlib furthermore we have performed feature engineering by converting categorical features into dummy values.
- EDA visualization



EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - 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).
  - 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 using a subquery.
  - List the failed landing outcomes in drone ship, their booster versions, and launch site names for the 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
- EDA with SQL

Build an Interactive Map with Folium

- In this stage we used folium library to represent our work as geospatial data by drawing markers circles and lines on an interactive map
- We started our interactive map by drawing 4 circles on 4 different sites belongs to Falcon 9 rockets lunches
- We put markers to on the same sites to represent the successful/failed first stage of rockets return using marker objects
- Finally, we calculated the distances between the launch site (CCAFS LC40) to its proximities 1-the closest city, 2-coastline, and 3-highway. Then we drew polylines to represent these distances using PolyLine object
- Folium Map

Build a Dashboard with Plotly Dash

- Added a dropdown list to enable Launch Site selection including the following options:
  - All Sites,
  - CCAFS LC-40,
  - CCAFS SLC-40
  - VAFB SLC-4E
  - KSC LC-39A
- we added a pie chart to show the total successful launches count for all sites
- Added a slider to select payload which ranges from 0 -10000
- Added a scatter chart to show the correlation between payload and launch success
- Plotly Dash

Predictive Analysis (Classification)

- Importing the required libraries.
- Loading the cleaned data.
- Standardizing the data to prevent the bias.
- Splitting the data into 20% for testing data and 80% training data.
- Initializing 4 different classification algorithms:
  - Logistic Regression (LR)
  - Support Vector Machine (SVM)
  - Decision Tree (DT)
  - K nearest neighbors (KNN)
- Using Grid Search technique to find the best parameters
- Using Evaluation techniques including, Confusion matrix , F1 score, for finding the best algorithm
- [Machine Learning]

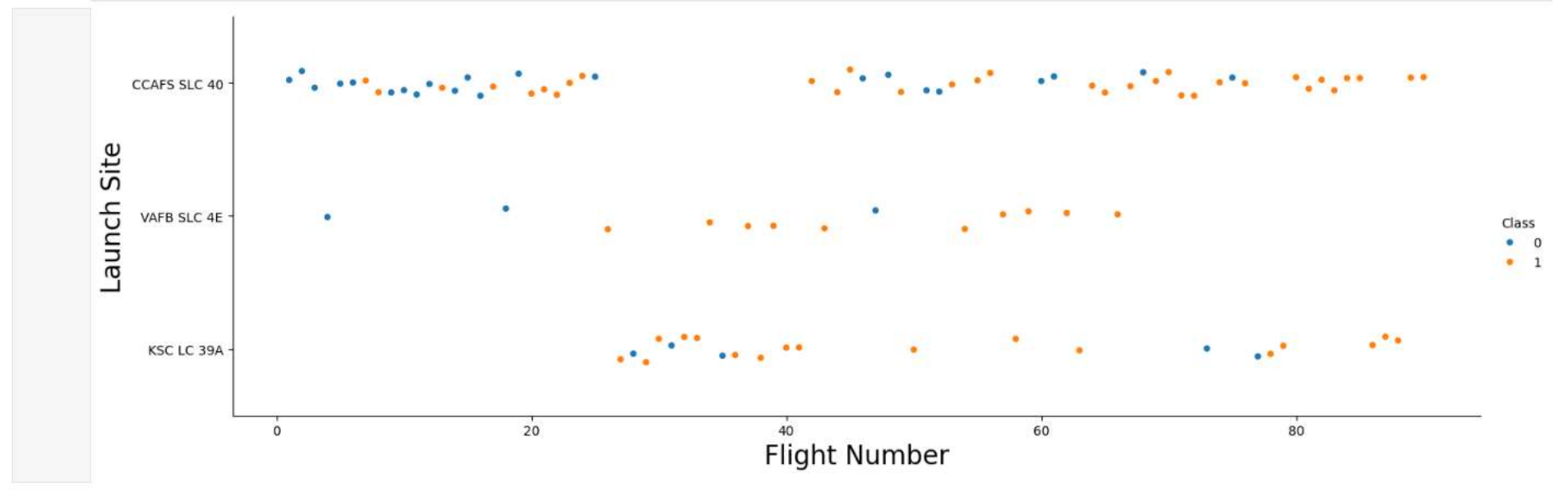
Results

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

Section 2

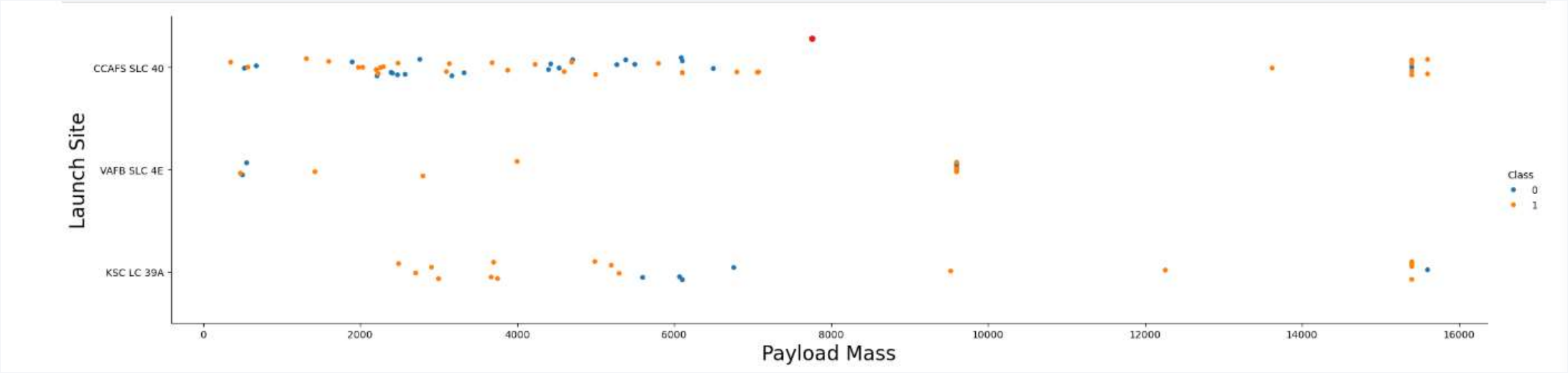
Insights drawn from EDA

Flight Number vs. Launch Site



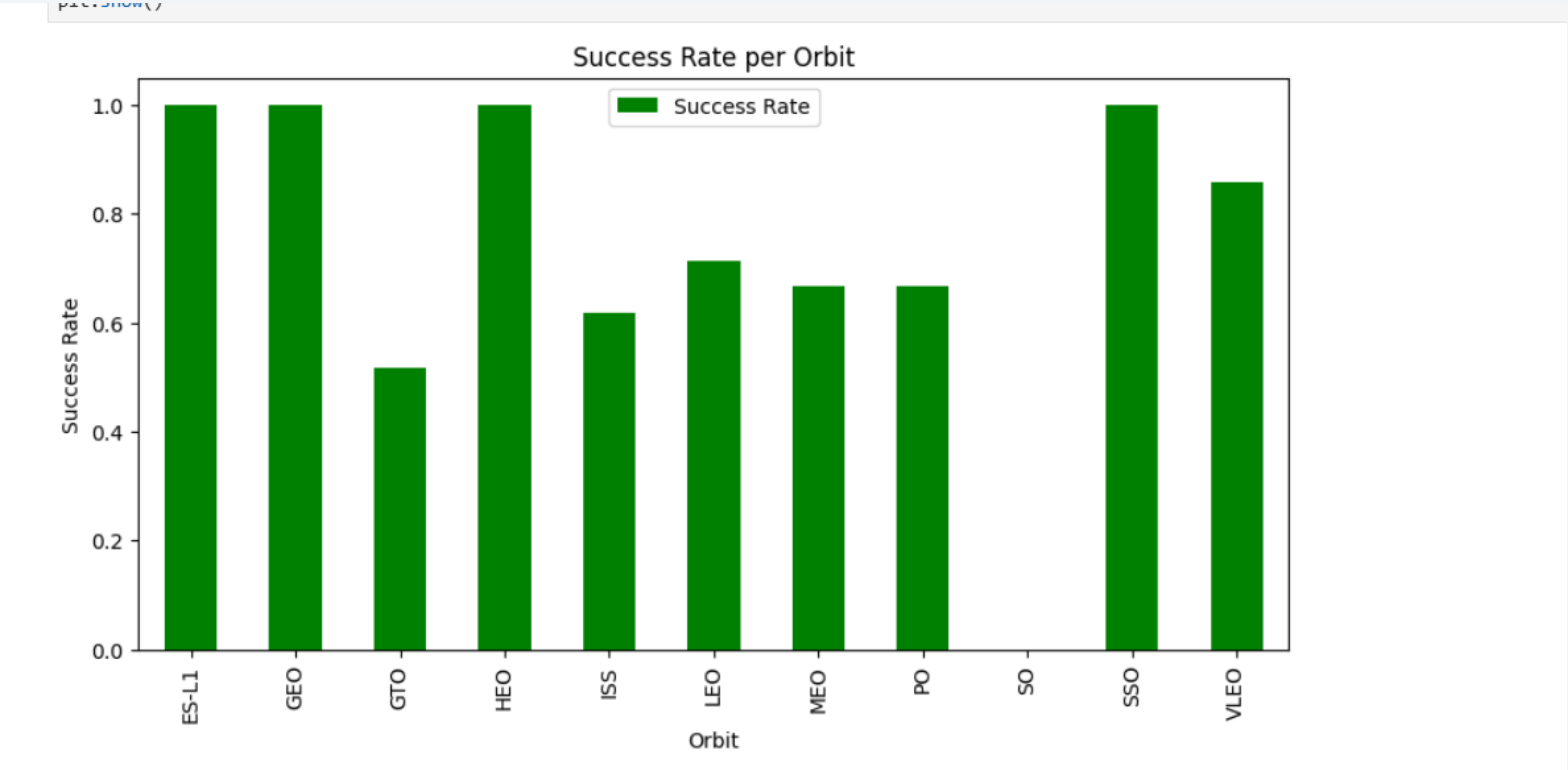
- CCAFS SLC 40 : is the most usable site for launching SpaceX's rockets and it has 55 trials, 33 of them are successful and 22 of them are failed # 60% success rate
- VAFB SLC 4E : is the least usable site for launching SpaceX's rockets and it has 13 trials, 10 of them are successful and 03 of them are failed # 77% success rate
- VAFB SLC 4E : is a moderate site in terms of launching SpaceX's rockets and it has 22 trials, 17 of them are successful and 05 of them are failed # 77% success rate

Payload vs. Launch Site



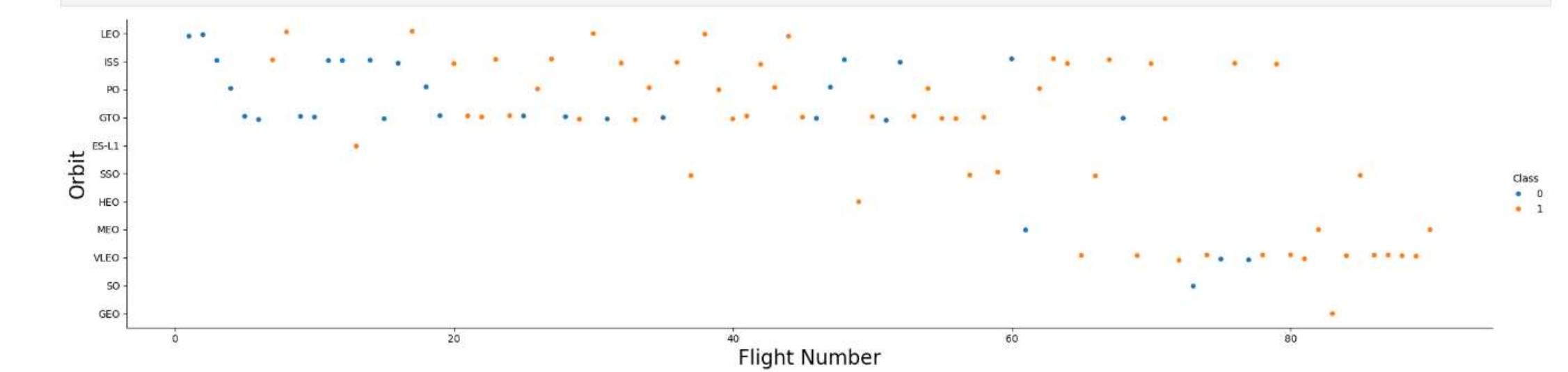
- There is no strong relationship between the payload mass and the success of first stage return since there are approximately equivalent numbers of failed and successful trials.

Success Rate vs. Orbit Type



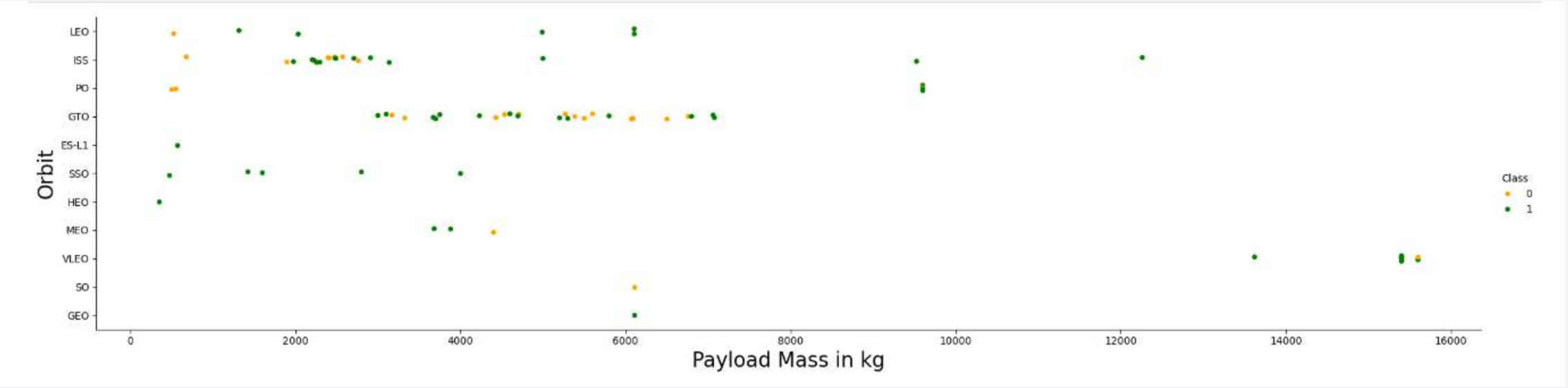
- The best orbits in terms of successful first stage returns are ['ES-L1', 'GEO', HEO, SSO]
- The worst orbit is 'GTO’

Flight Number vs. Orbit Type



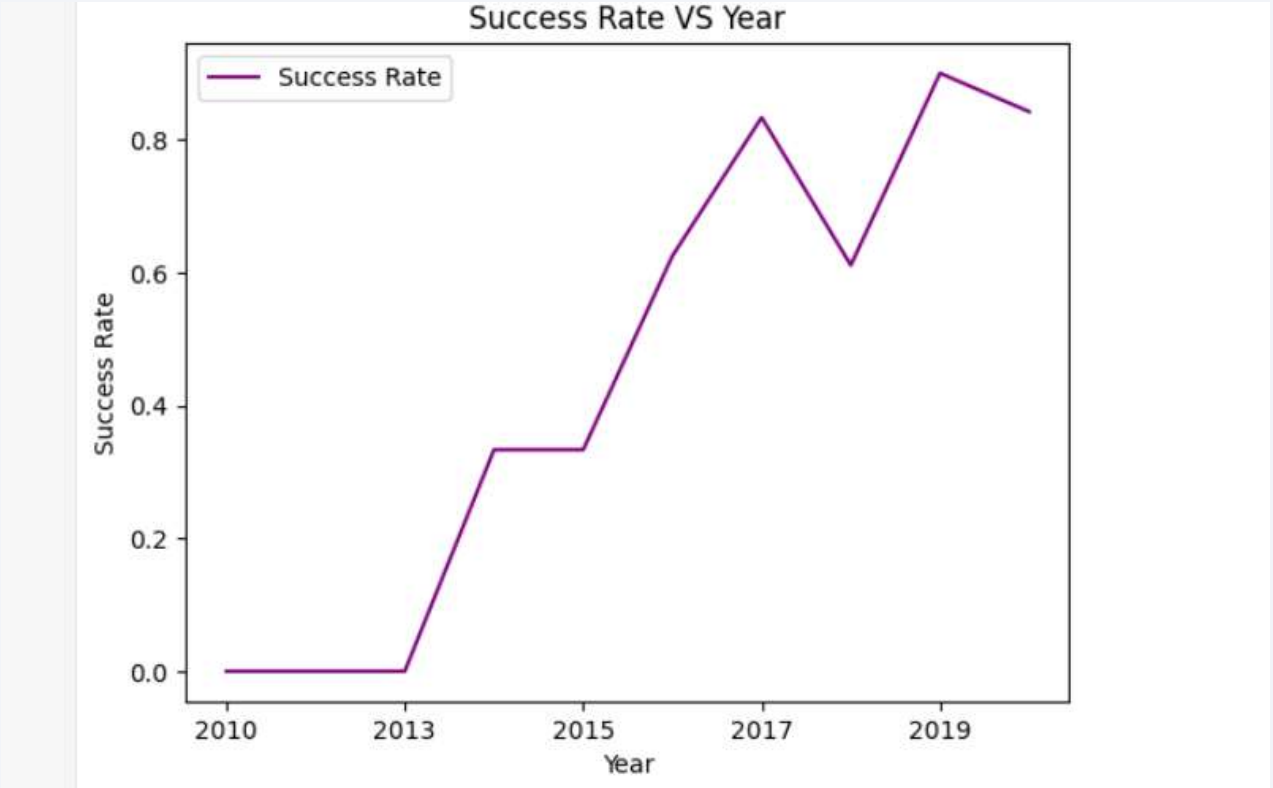
- In the LEO orbit the success is related to the number of flights whereas there is no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- 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



- you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

```
[11]: %sql select distinct launch_site from SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

[11]:

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

- The query is:
  - %sql select distinct launch_site from SPACEXTBL
- We get all the unique launch sites

Launch Site Names Begin with 'CCA'

```
[12]: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
 * sqlite:///my_data1.db
Done.
```

[12]:

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

- The query is
  - %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
- We use like to fetch from the database where all the launch sites start with CCA and limit to get only 5 values

Total Payload Mass

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

```
[13]: %sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';
```

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

[13]:

sum(payload_mass__kg_)
45596

- The query is
  - %sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)'
- We fetch the total weight using sum() function and we filter using the condition where the customer is NASA

Average Payload Mass by F9 v1.1

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

```
[14]: %sql select avg(payload_mass__kg_) as avg_mass_F9 from SPACEXTBL where booster_version = 'F9 v1.1'
```

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

[14]:

avg_mass_F9
2928.4

- The query is
  - %sql select avg(payload_mass__kg_) as avg_mass_F9 from SPACEXTBL where booster_version = 'F9 v1.1’
- We fetch the average weight using average() function and we filter using the condition where the booster version is F9 v1.1

First Successful Ground Landing Date

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

Hint:Use min function

```
[19]: %sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

[19]:

min(DATE)
2015-12-22

- The query is
  - %sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)’
- The date of the first successful landing was **22-12-2015**

Successful Drone Ship Landing with Payload between 4000 and 6000

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

[21]:
```
%sql select booster_version from SPACEXTBL\
where (Landing_Outcome = 'Success (drone ship)' and (payload_mass__kg_ > 4000 and  payload_mass__kg_ > 6000));
```

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

[21]:

Booster_Version
F9 FT B1029.1
F9 FT B1036.1
F9 B4 B1041.1

- The query is
  - %sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)
- The result was
  - F9 FT B1029.1
  - F9 FT B1036.1
  - F9 B4 B1041.1

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
[24]: %sql select mission_outcome, count(mission_outcome) as counts from SPACEXTBL GROUP BY mission_outcome
```

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

[24]:

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

- We have only 1 failure while 99 successful outcomes for the mission

Boosters Carried Maximum Payload

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

```
[25]: %sql select distinct booster_version from SPACEXTBL\
      where payload_mass__kg_ in (select max(payload_mass__kg_) from SPACEXTBL);
```

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

[25]:

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

- The booster versions that carry the maximum payload starts with F9 B5 and ranges from B1048 up to B1060

2015 Launch Records

```
[28]: %sql select Landing_Outcome, booster_version, launch_site from SPACEXTBL\
      where (Landing_Outcome = 'Failure (drone ship)' and date like '2015%')
```

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

[28]:

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

- We have two failed landing in 2015 on a drone ship which both in the same site, CCAFS LC-40 and with same booster version F9 v1.1

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

[29]:
```
%sql select Landing_Outcome, count(*) as counts_of_landing_outcomes from SPACEXTBL\
where DATE between '2010-06-04' and  '2017-03-20' group by Landing_Outcome\
order by count(Landing_Outcome) desc
```

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

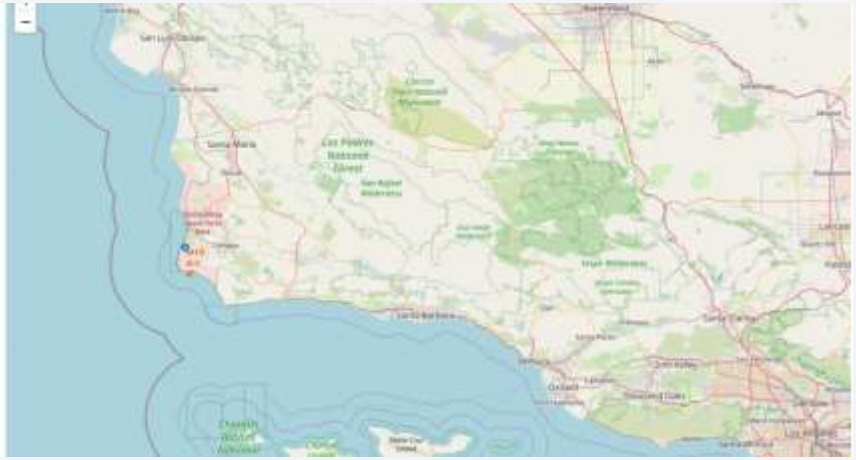
[29]:

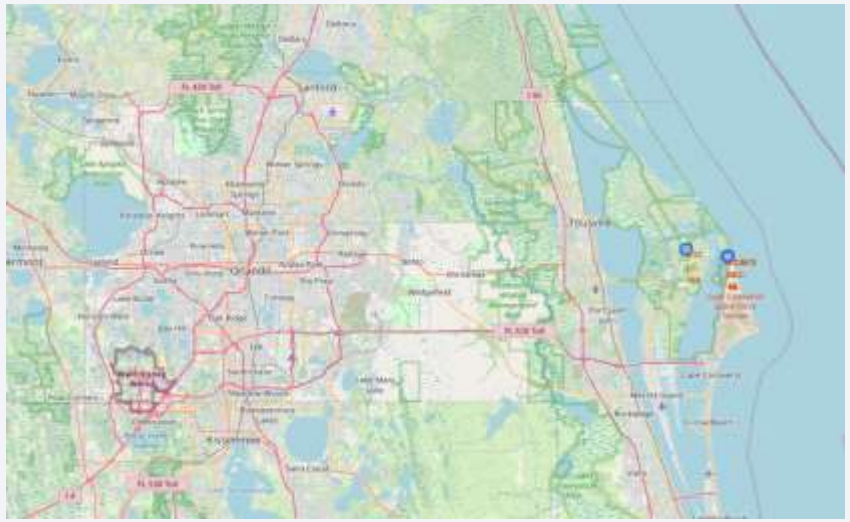
Landing_Outcome	counts_of_landing_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

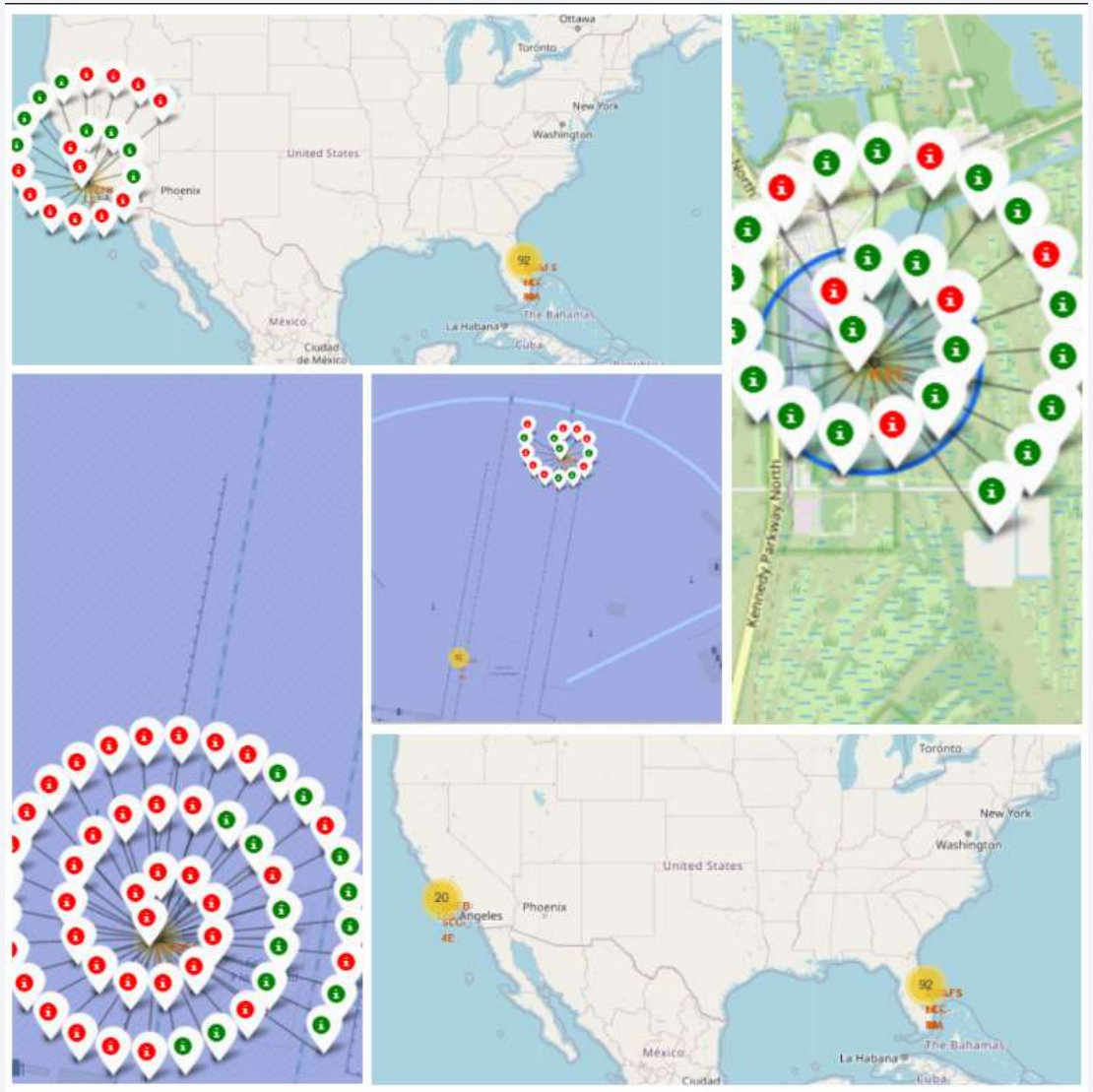
<Folium Map Screenshot 1>





- All site locations are near the coast and Equator line,
- The launch sites are distributed in two states California and Florida and they are in close proximity to the coast

<Folium Map Screenshot 2>



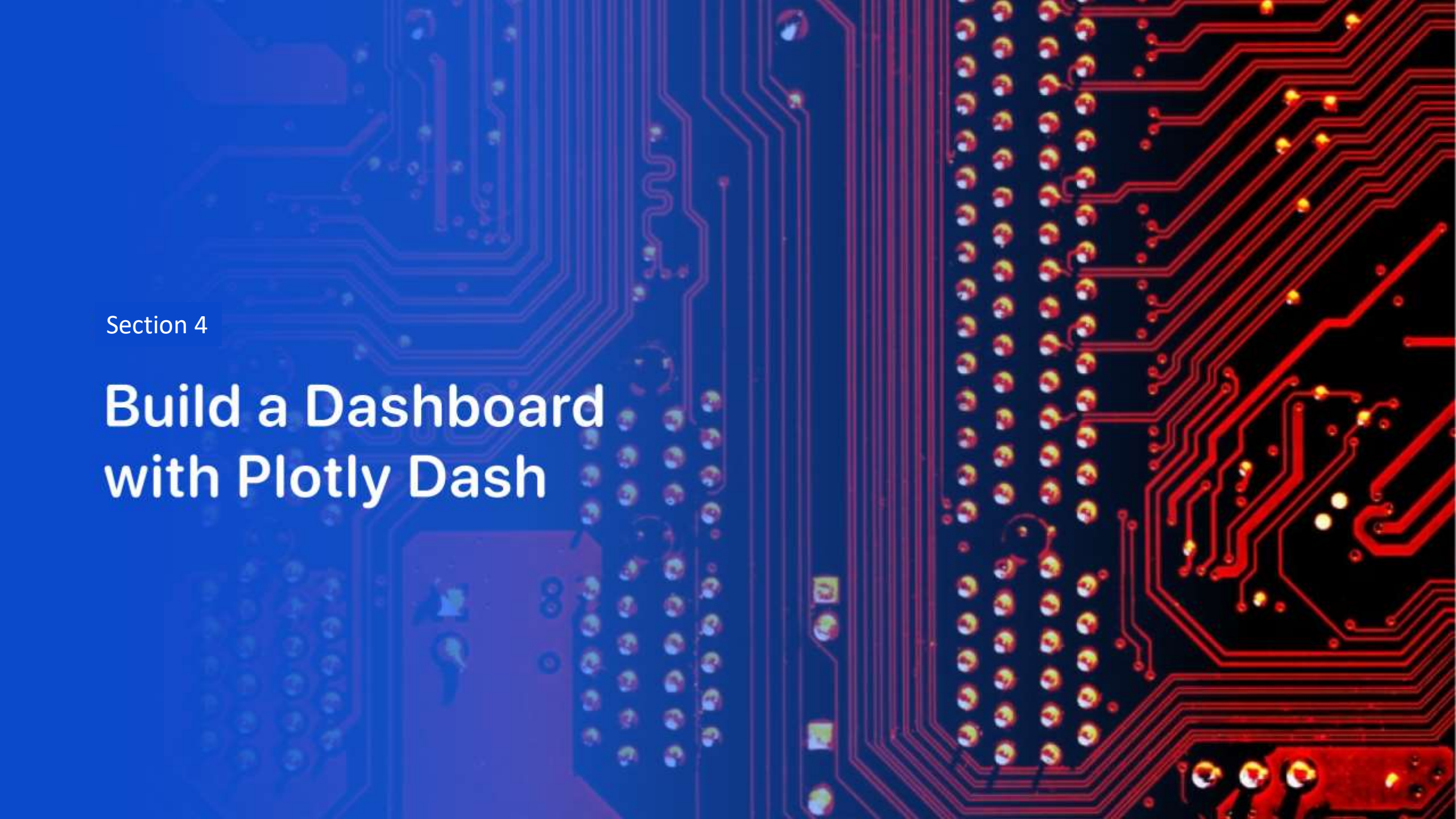
- We have marked green color markers which represents launch sites with high success rates and red color markers for sites with high failed rate

<Folium Map Screenshot 3>

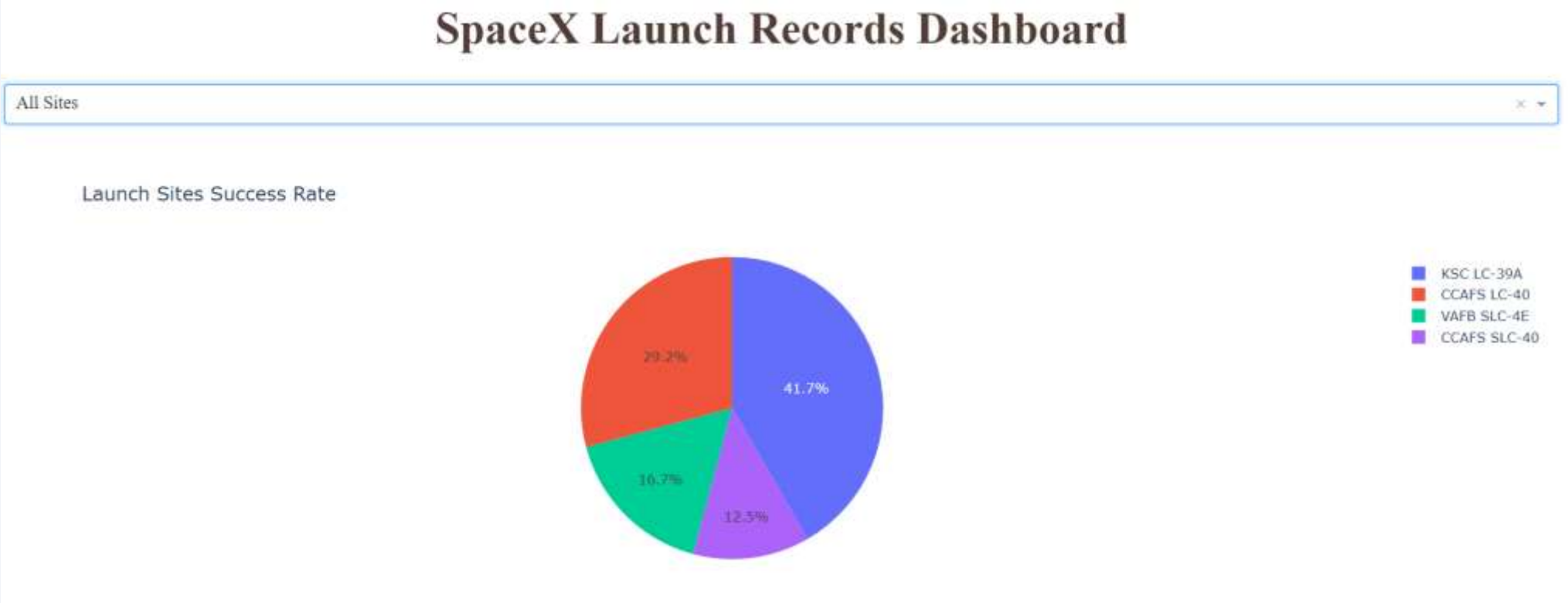
- Are launch sites in close proximity to railways? Yes
- Are launch sites in close proximity to highways? Yes
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Section 4

Build a Dashboard with Plotly Dash

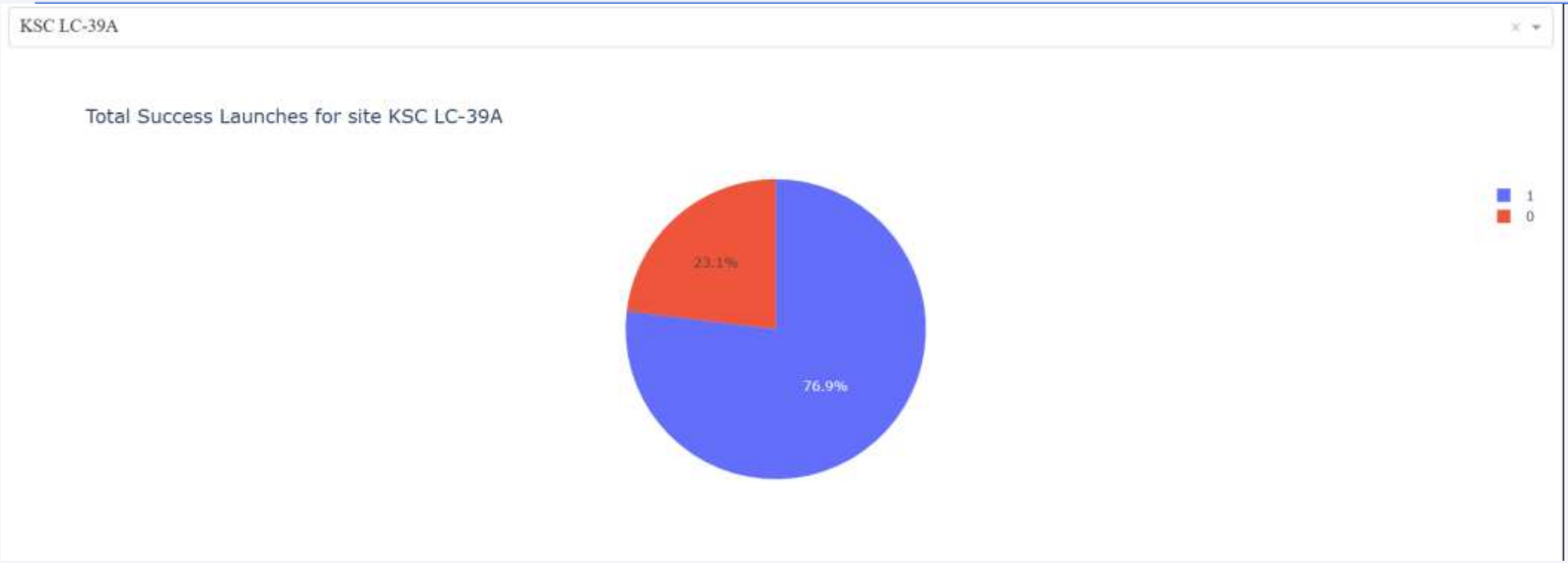


<Dashboard Screenshot 1>



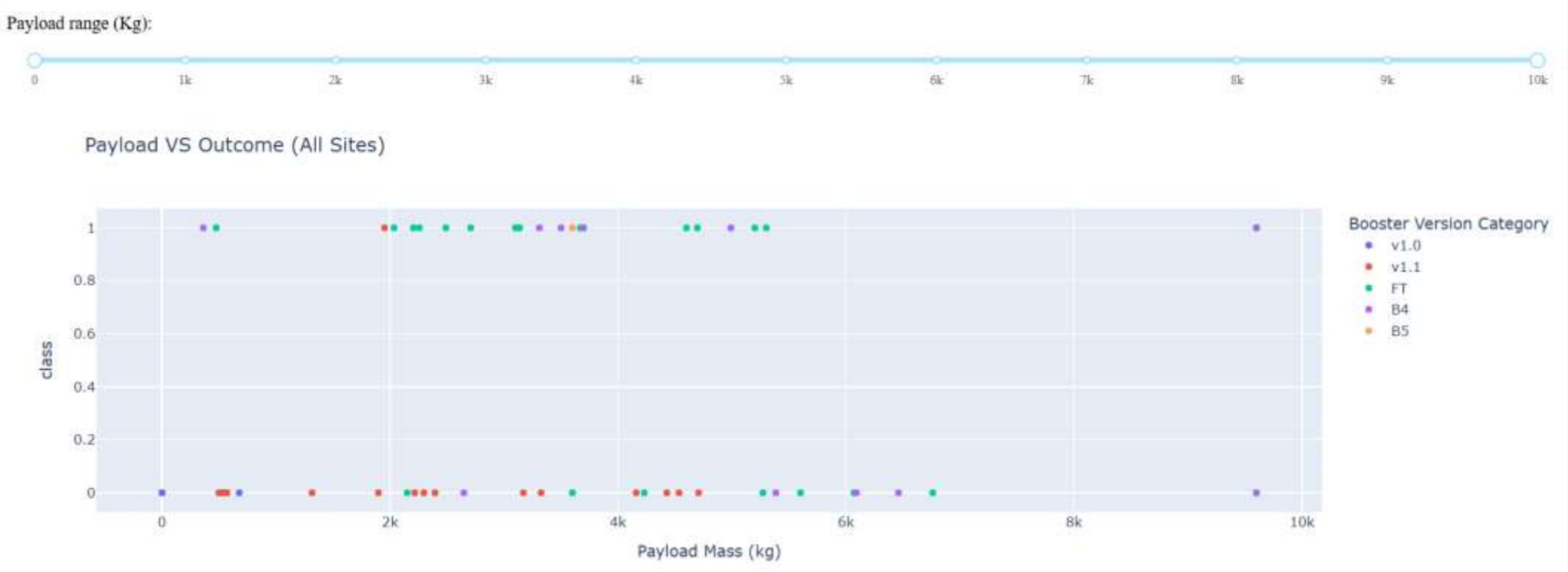
- The best is KBC LC-39A with 41.7% success rate
- The one with the least success rate is CCAFS SLC-40 with 12.5% success rate

<Dashboard Screenshot 2>



- The launch site with highest ratio is KSC LC-39A
- Successful missions-76.9%
- Failure missions-23.1

<Dashboard Screenshot 3>

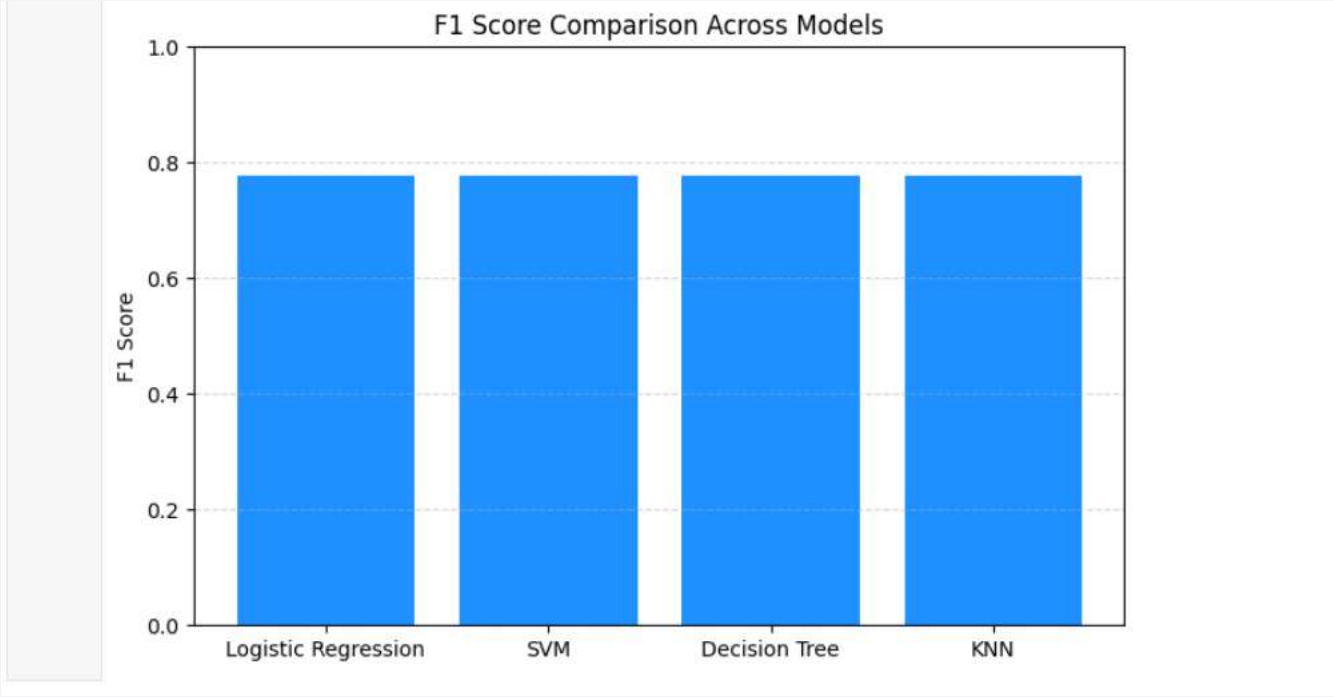


- We can infer that if the payload mass is less then 4000 kg then the outcome is more likely to be successful.

Section 5

Predictive Analysis (Classification)

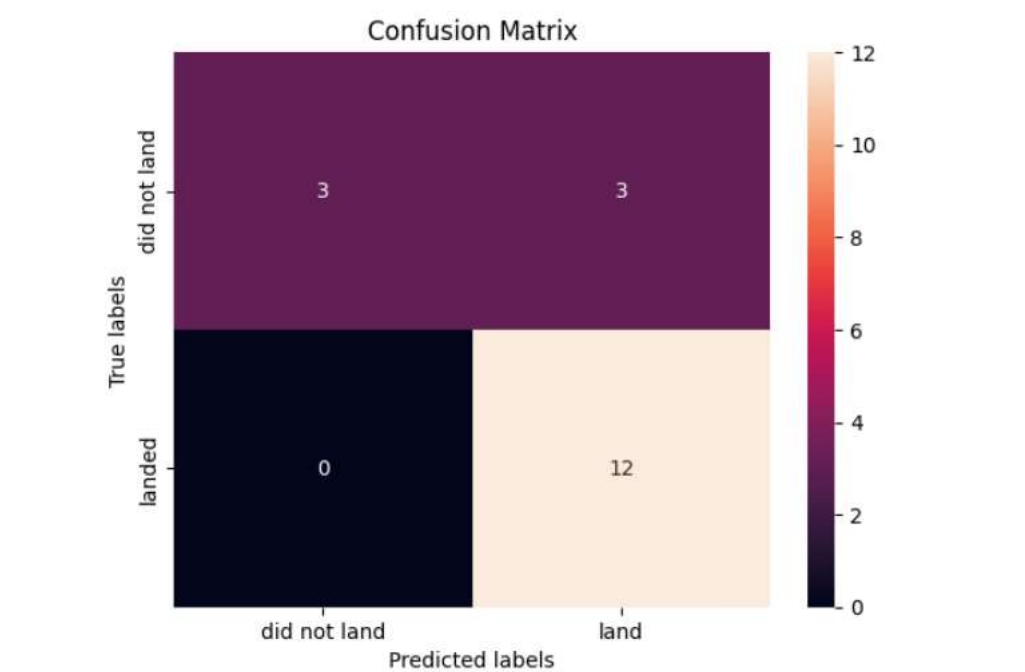
Classification Accuracy




```
Logistic Regression:
Jaccard Score of = 0.8
F1 Score = 0.7777777777777777
-----------------------------------
SVM:
Jaccard Score of = 0.8
F1 Score = 0.7777777777777777
-----------------------------------
Decision Tree:
Jaccard Score of = 0.8
F1 Score = 0.7777777777777777
-----------------------------------
KNN:
Jaccard Score of = 0.8
F1 Score = 0.7777777777777777
-----------------------------------
```

- After analyzing the 4 models it is being concluded that all the models have the same jaccard score of .8 and the same F1 score of .77 hence all models are perfectly suitable.

Confusion Matrix



- All the models have shown the same confustion matrix.
- True Positive = 12
- False Positive = 0
- True Negative = 3
- False Negative = 3

Conclusions

- A successful first stage return, leads to huge savings in terms of rockets lunches cost
- A wide range of attributes affects the possibility of a successful first stage return. In our model there were 83 attributes were taken into consideration.
- Launch sites were all close to a highway, railway, and coastline proximities, which aimed in transportation cost-reduction
- Orbit and booster version affect success rate.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!