IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

AJAY KHANNA
05 October 2023

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- EDA with data visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- EDA results
- Interactive analytics
- Predictive analysis

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.

- **Problems you want to find answers**

The project task is to predicting if the first stage of the SpaceX Falcon 9 rocket will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  SpaceX Rest API

  Web Scrapping from Wikipedia

- Perform data wrangling

  One Hot Encoding data fields for Machine Learning and data cleaning of null values and irrelevant columns.

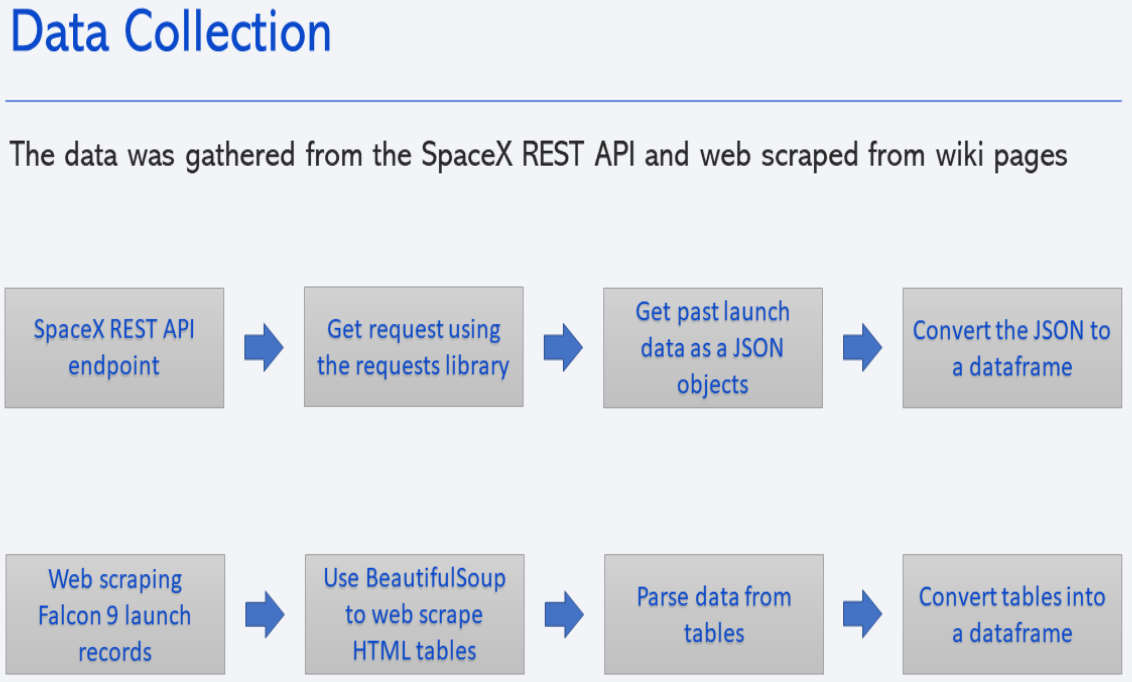
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  LR, KNN, SVM, DT models have been built and evaluated for the best classifier.

Data Collection

The following datasets was collected:

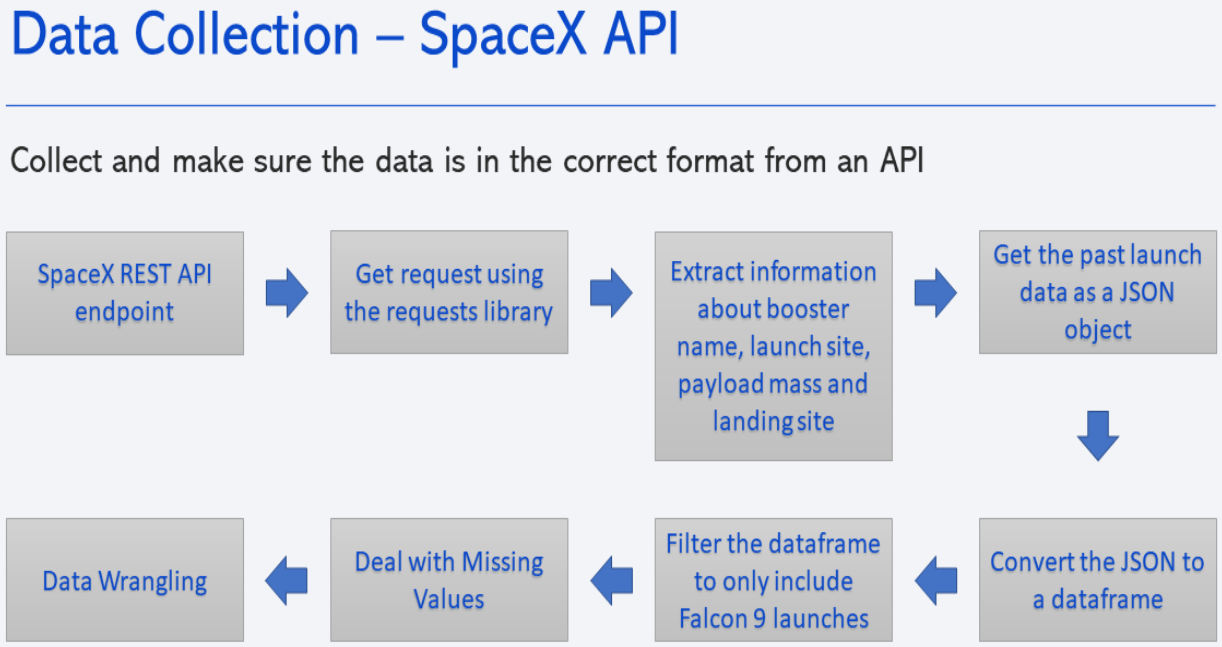
- SpaceX launch data that is gathered from the SpaceX Rest API.
- This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications and landing outcome.
- The SpaceX Rest API endpoints, or URL starts with api.spacexdata.com/v4/.
- Another popular data source for obtaining Falcon 9 Launch data is web scraping wikipedia using BeautifulSoup.



Data Collection – SpaceX API

- Data collection with SpaceX REST calls

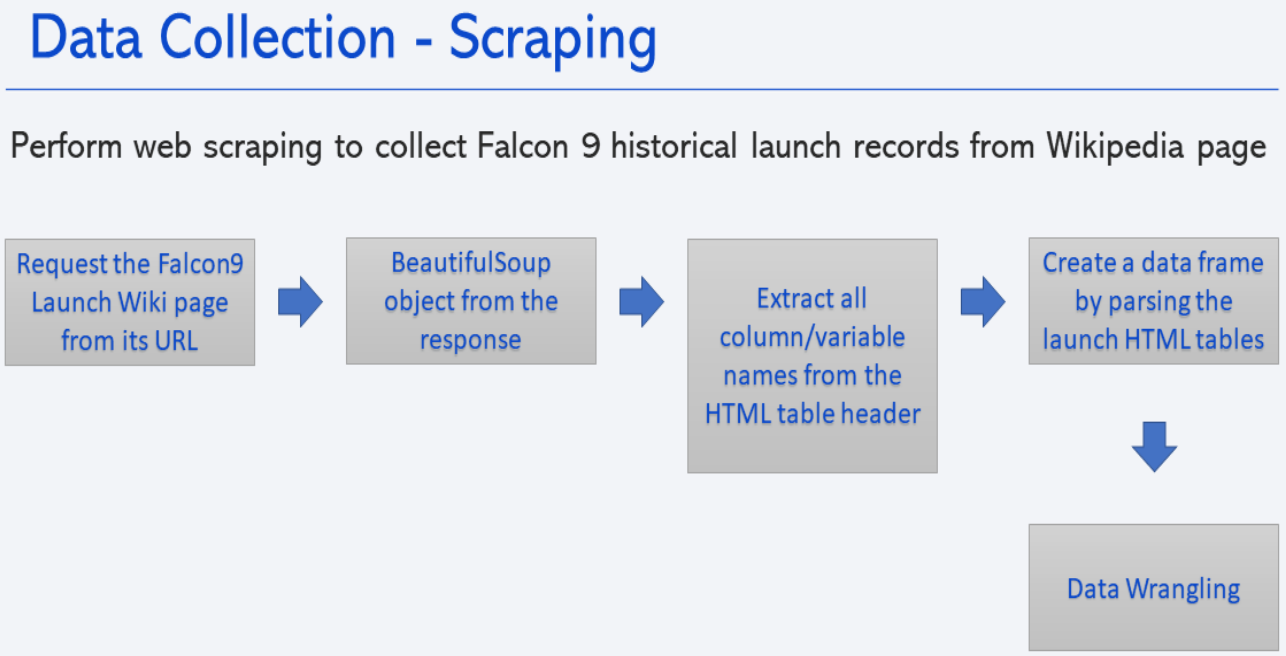
https://github.com/Ajay15Khanna/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

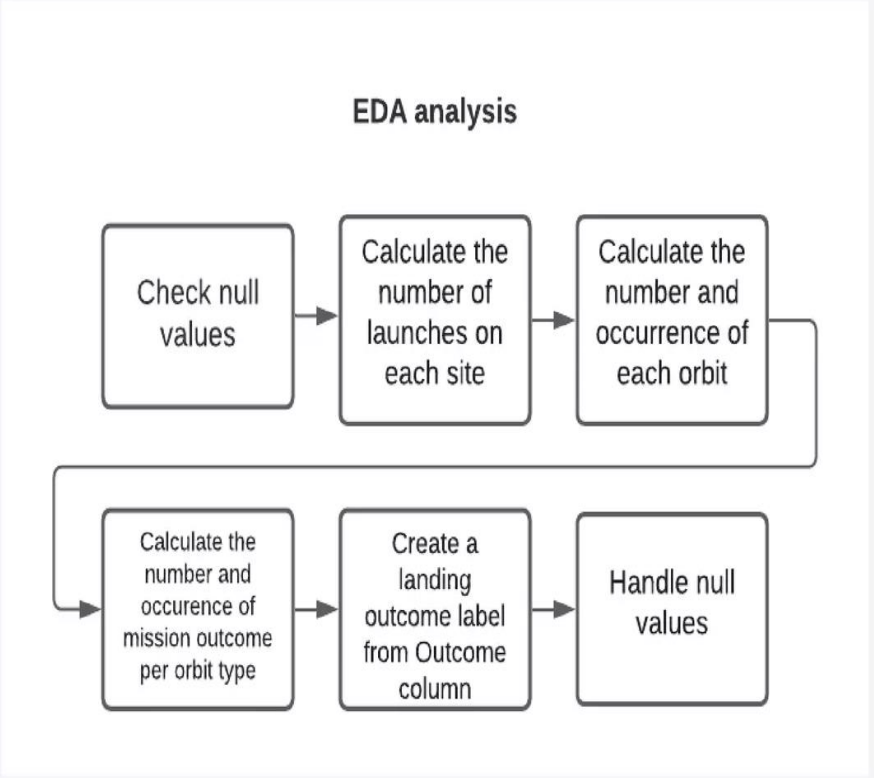
- Web Scrapping from Wikipedia

https://github.com/Ajay15Khanna/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

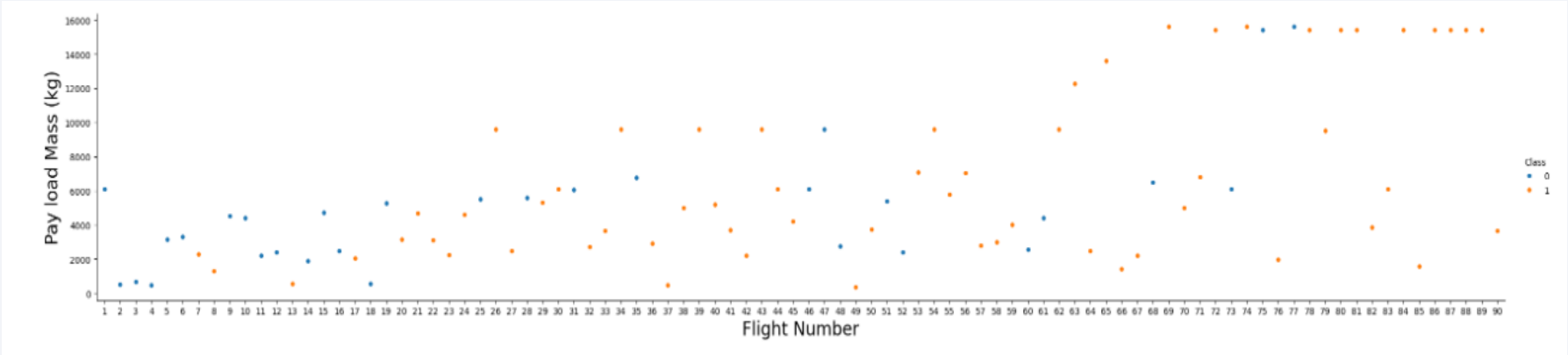


Data Wrangling

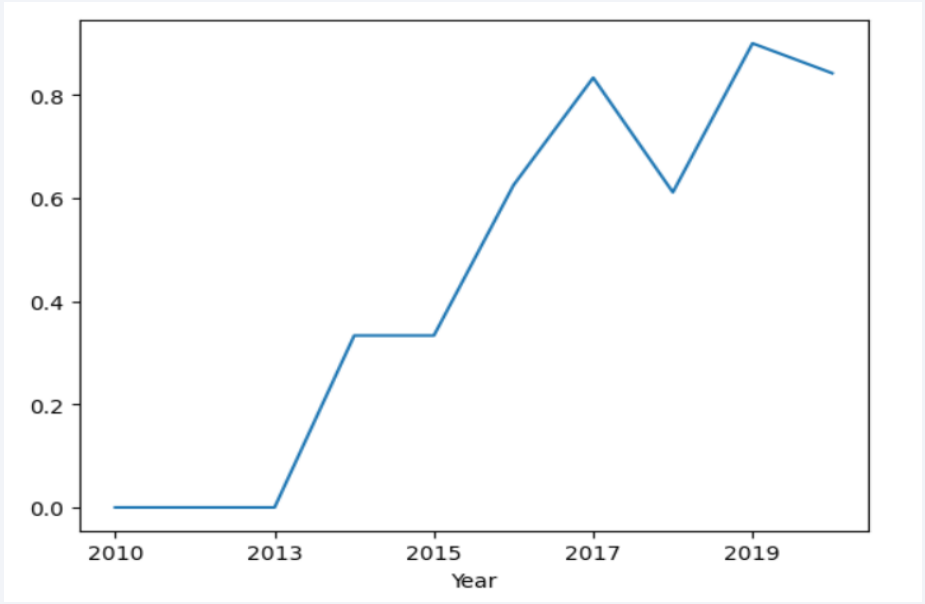
https://github.com/Ajay15Khanna/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

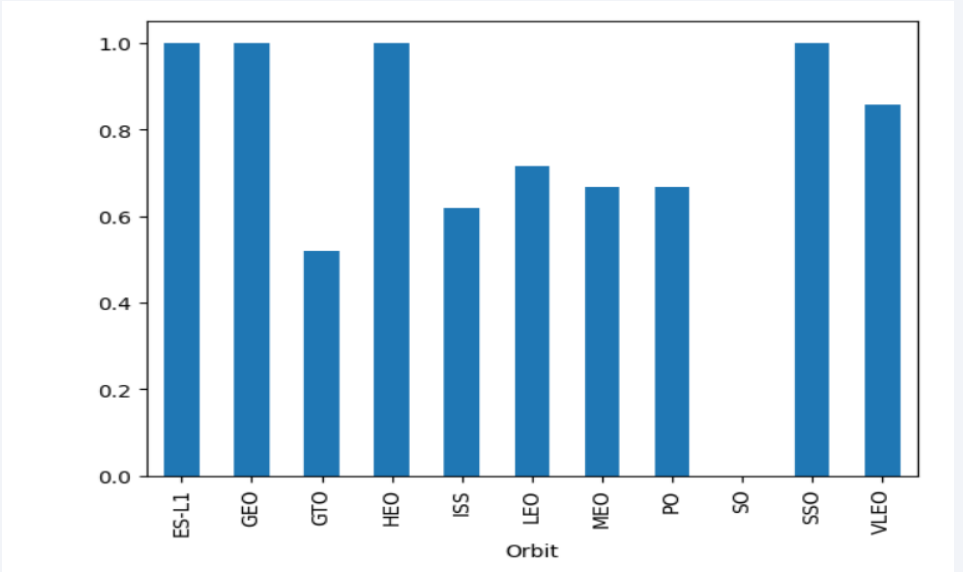


EDA with Data Visualization



https://github.com/Ajay15Khanna/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

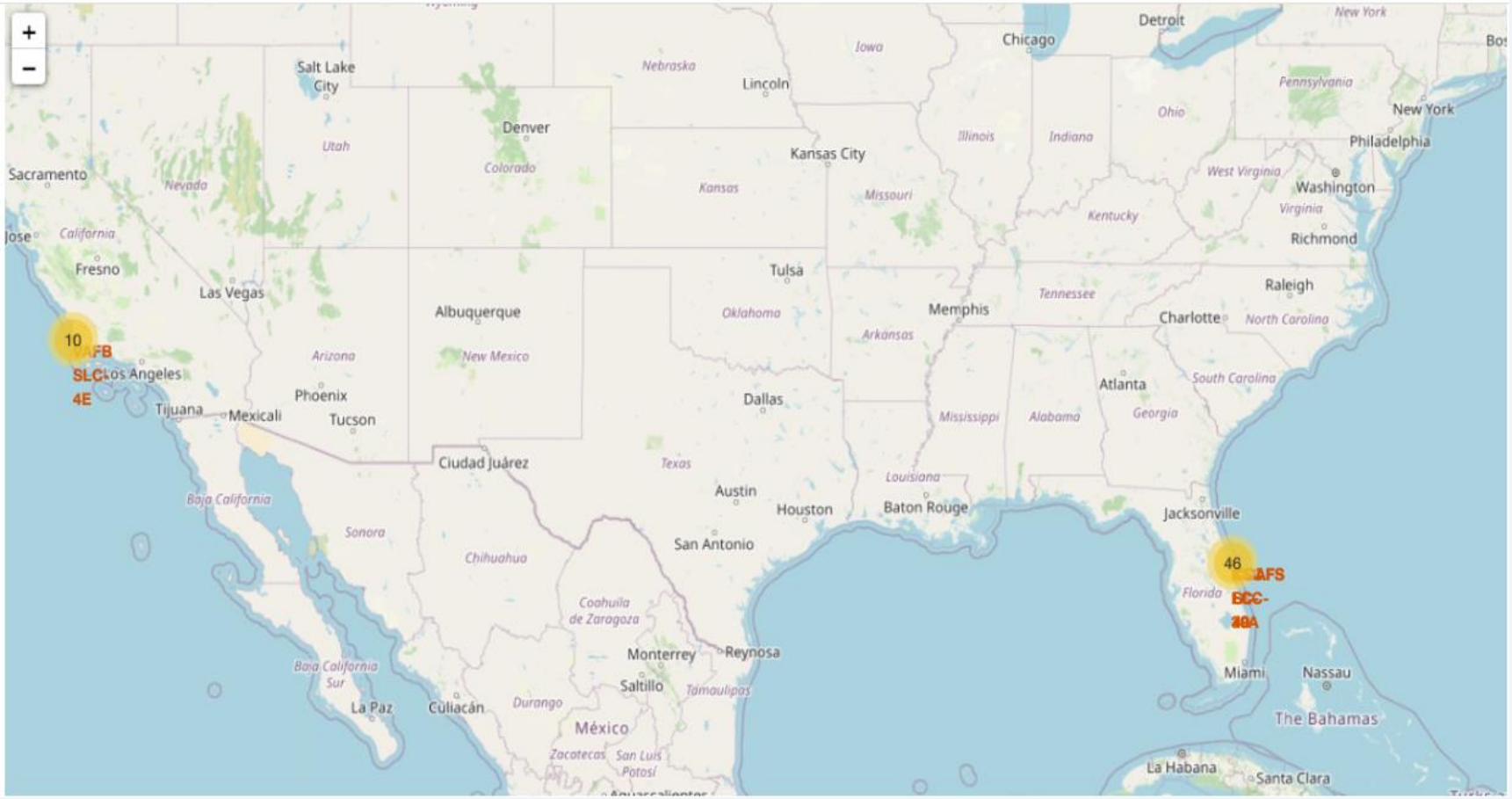




EDA with SQL

- SQL queries performed include: https://github.com/Ajay15Khanna/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_outcomes in ground pad, booster versions, launch_site for the months in year 2017.

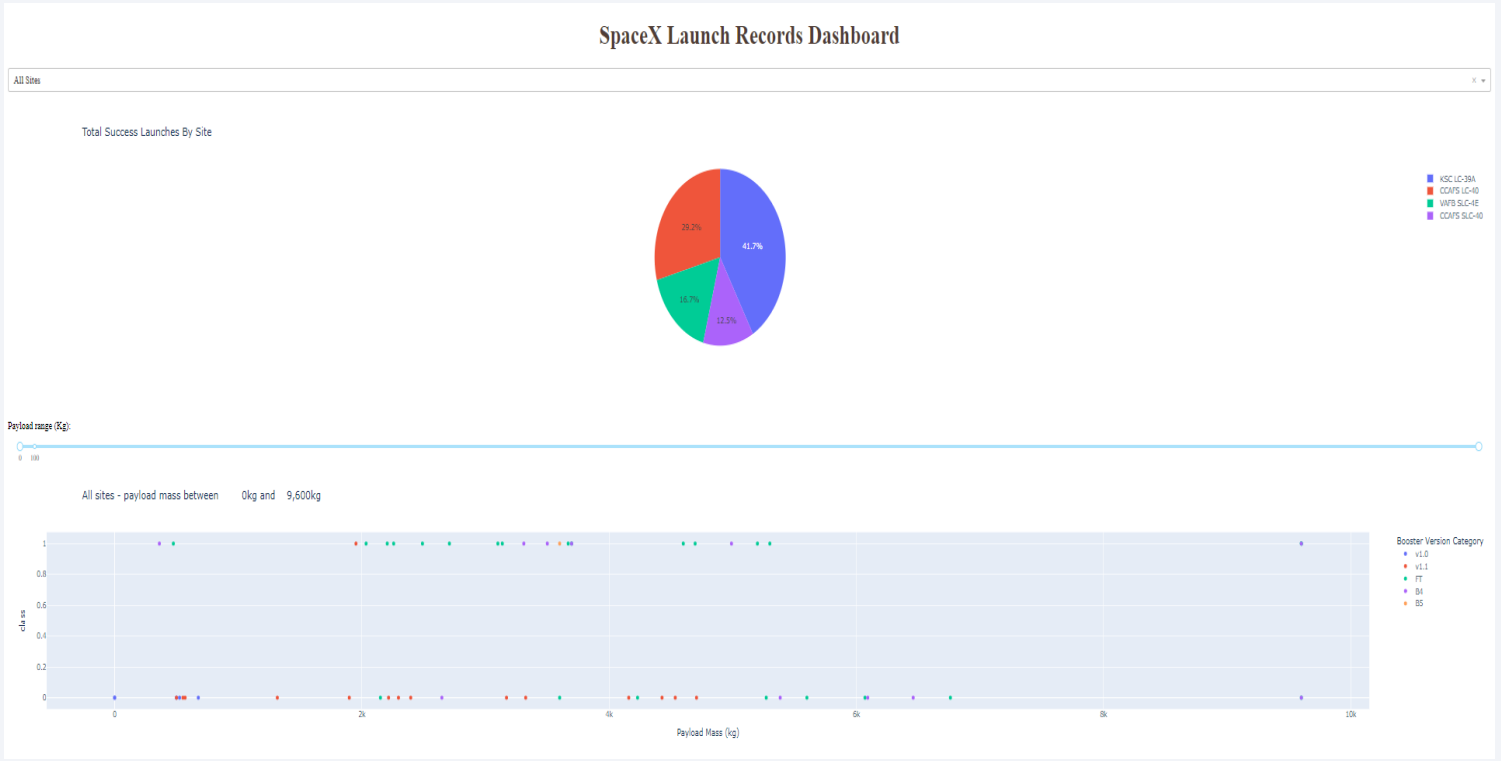
Build an Interactive Map with Folium

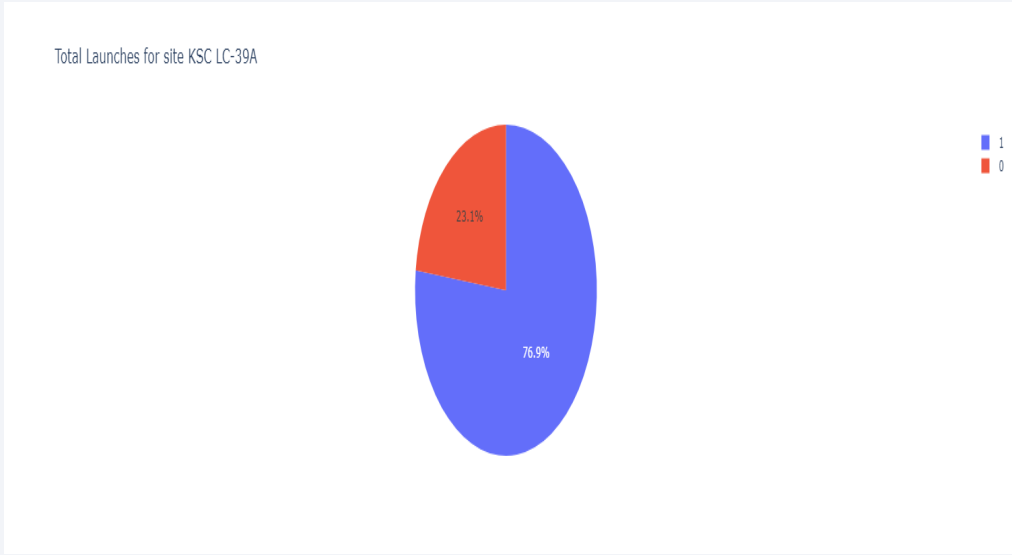


https://github.com/Ajay15Khanna/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Map markers have been added to the map with aim to finding an optimal location for building a launch site.

Build a Dashboard with Plotly Dash





KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

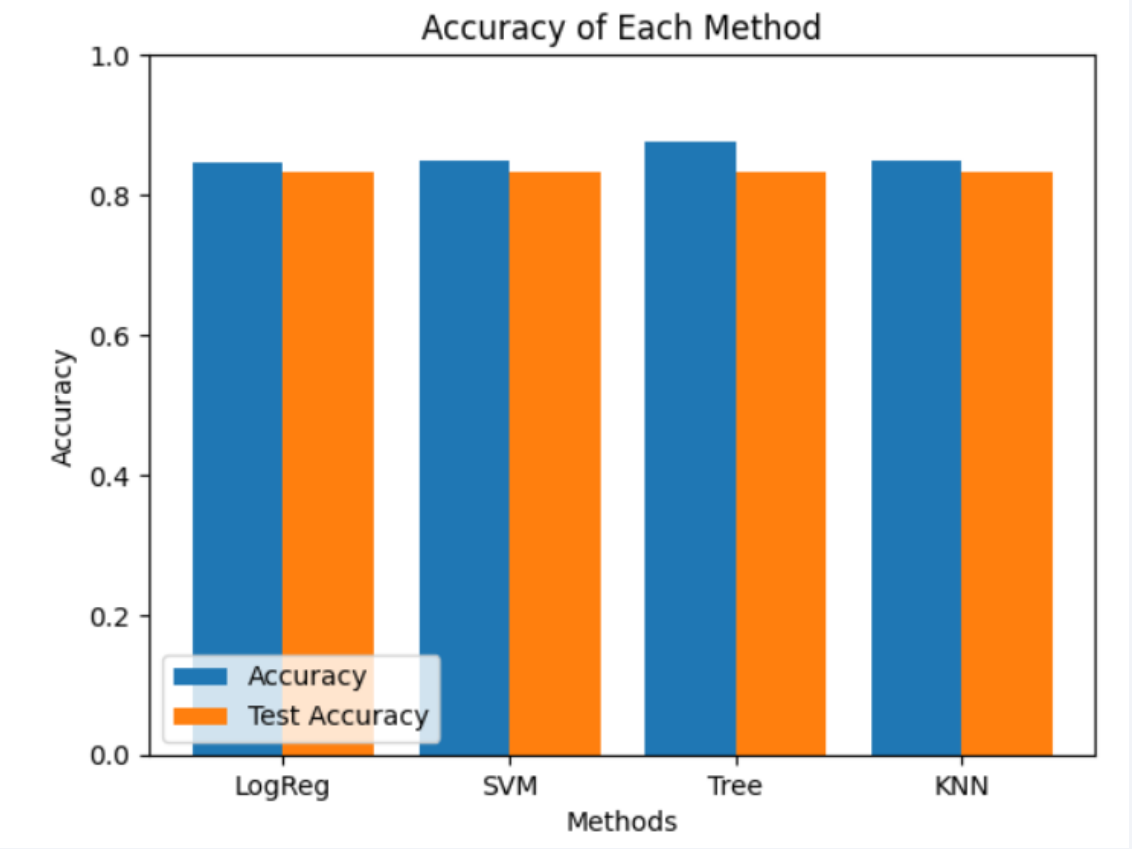
We can see that KSC LC-39A had the most successful launches from all the sites.

https://github.com/Ajay15Khanna/Applied-Data-Science-Capstone/blob/main/Spacex_dash_app.py

Predictive Analysis (Classification)

- The SVM, KNN, and Logistic Regression model achieved the highest accuracy at 83.3%, while the SVM performs the best in terms of Area Under the Curve at 0.958.

https://github.com/Ajay15Khanna/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

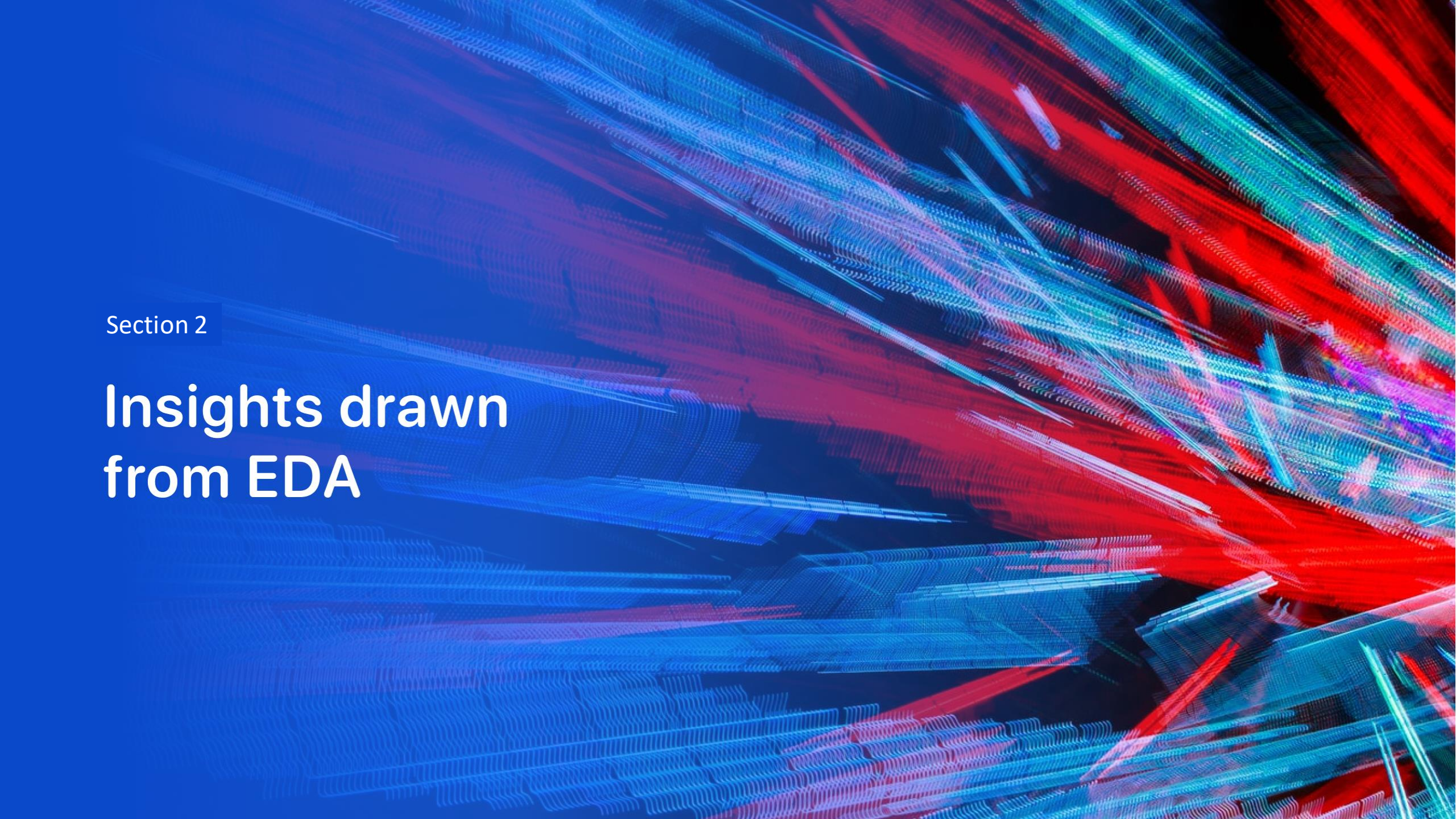


Results

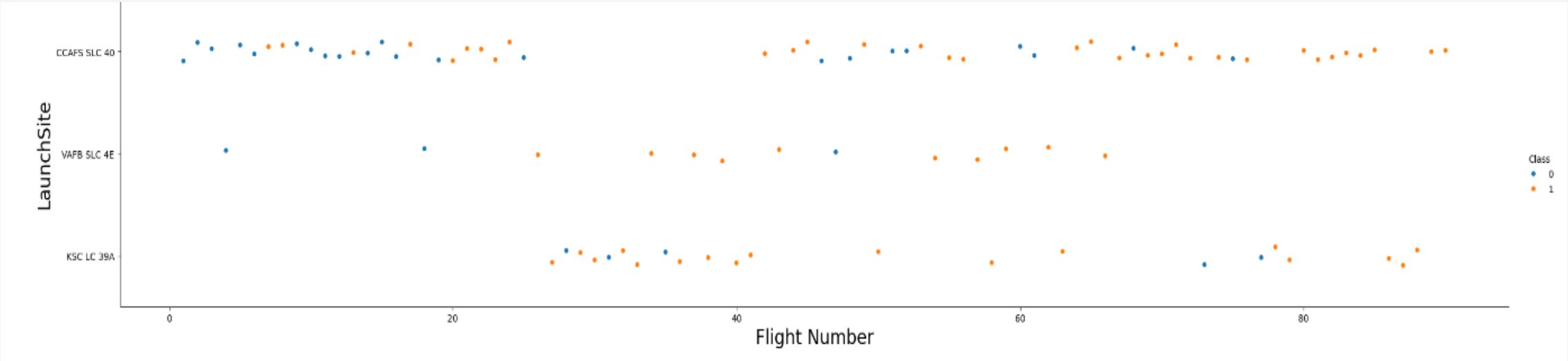
- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low weighted payloads perform better than the heavier payloads.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
- KSC LC-39A had the most successful launches from all the sites.
- Orbit GEO,HEO,SSO,ES L1 has the best Success Rate.

Section 2

Insights drawn from EDA

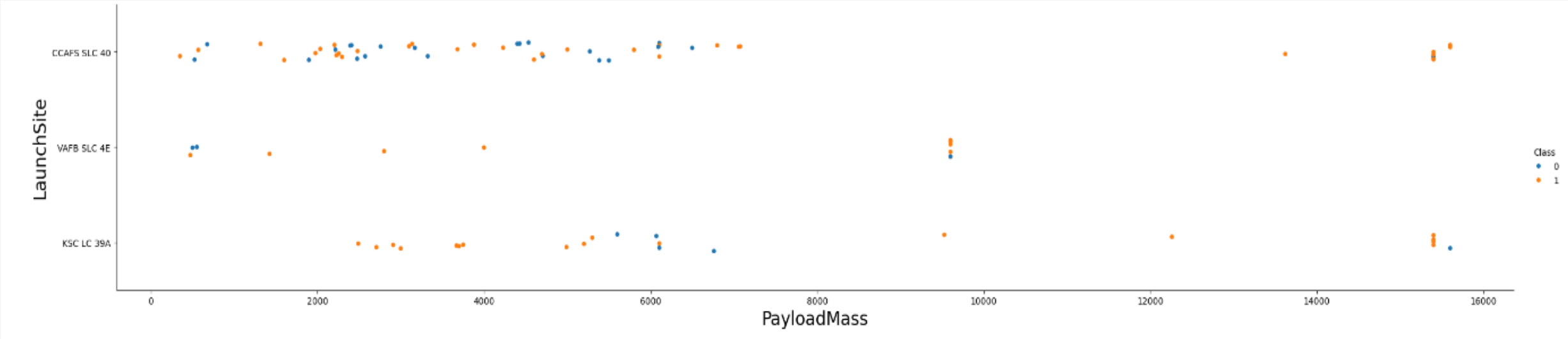


Flight Number vs. Launch Site



- Launches from the site of CCAFS SLC 40 are significantly higher than launches form other sites.

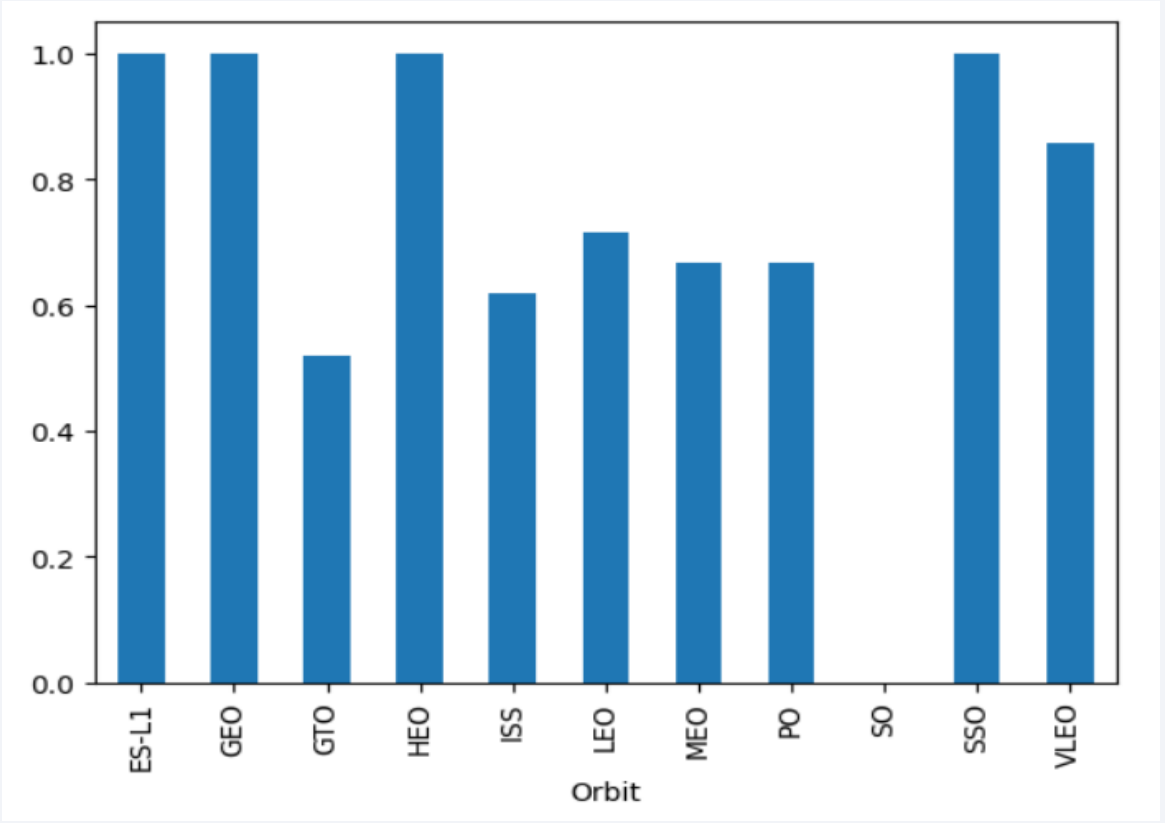
Payload vs. Launch Site



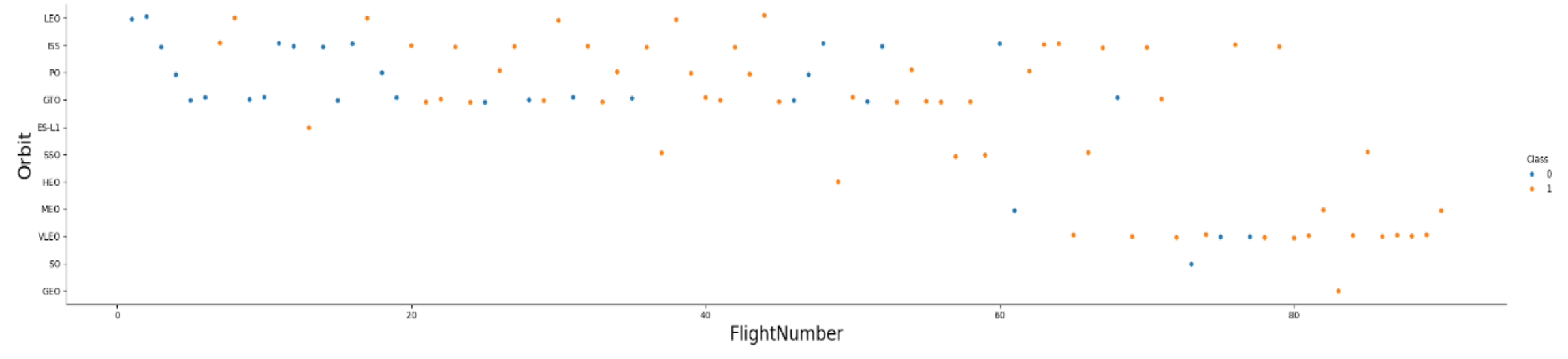
- The majority of Pay Loads with lower Mass have been launched from CCAFS SLC 40.

Success Rate vs. Orbit Type

The orbit types of ES-L1, GEO, HEO, SSO are among the highest success rate.

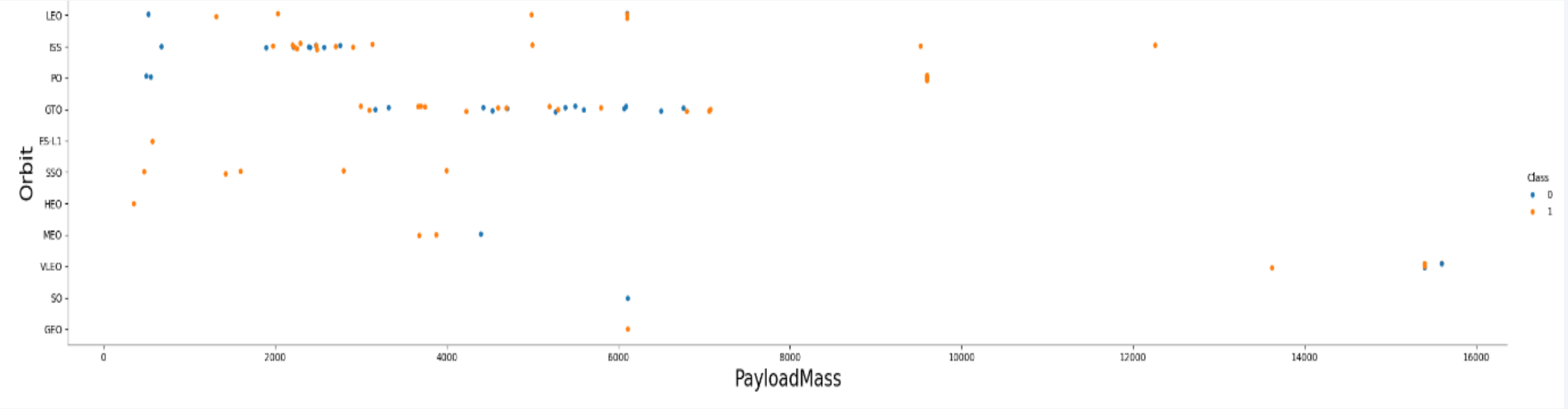


Flight Number vs. Orbit Type



- A trend can be observed of shifting to VLEO launches in recent years.

Payload vs. Orbit Type



There are strong correlation between ISS and Payload at the range around 2000, as well as between GTO and the range of 4000-8000.

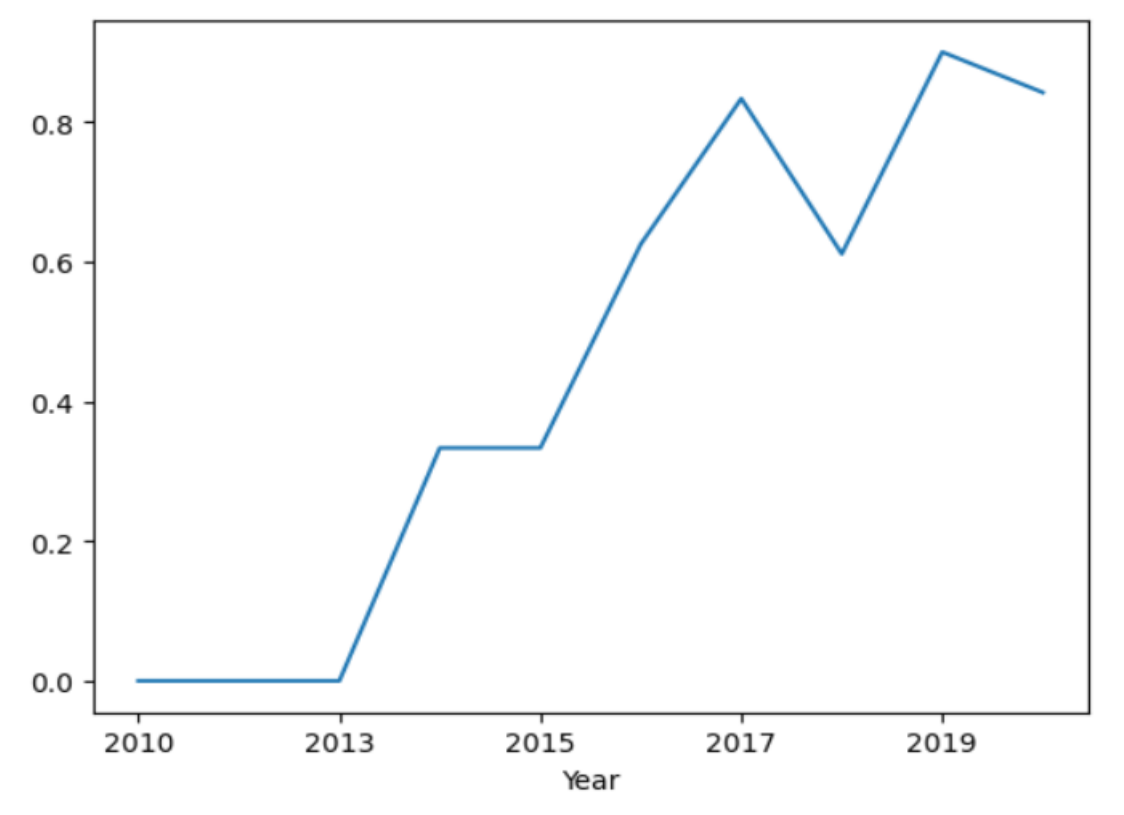
Launch Success Yearly Trend

Launch success rate has increased significantly

Since 2013 and has stablised since 2019,

Potentially due to advance in technology and

Lessons learned.



All Launch Site Names

We can get the unique values by

Using "DISTINCT"

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

Launch Site Names Begin with 'CCA'

We can get only 5 rows by using "LIMIT"

```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' limit 5;
```

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

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)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
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

Total Payload Mass

We can get the sum of all values by using "SUM"

```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD_LIKE_'%CRS%';
```

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

TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1

We can get the average of all values by using "AVG"

```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

We can get the first successful data by using "MIN" , because first date is same with the minimum date

```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

FIRST_SUCCESS_GP
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The payload mass data was taken between 4000 and 6000 only, and the landing outcome was determined to be "success drone ship"

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Success (drone ship)';
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

We can get the number of all the successful and failure mission by using "COUNT"

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

We can get the maximum payload masses by using "MAX"

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VE
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

We need to use substr(Date,4,2) as month to get the months and substr(Date,7,4)='2015' for year

```
%%sql SELECT "Booster_Version","Launch_Site" FROM SPACEXTBL
WHERE "Landing_Outcome"='Failure (drone ship)' AND substr(Date,1,4)='2015';
```

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

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

By using "ORDER" we can order the values in descending order, and with "COUNT" we can count all numbers as we did previously

```
sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY C
```

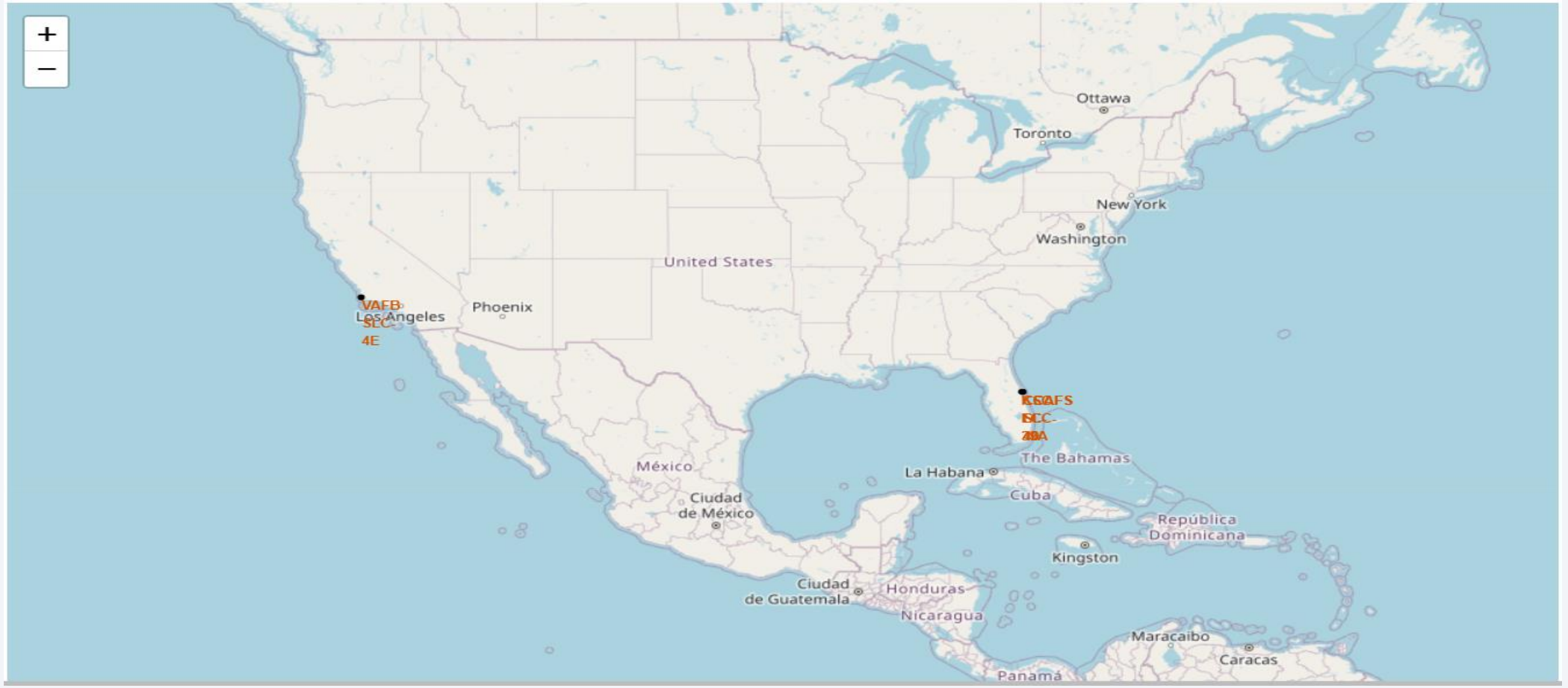
```
 * sqlite:///my_data1.db
Done.
```

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


Section 3
Launch Sites
Proximities Analysis

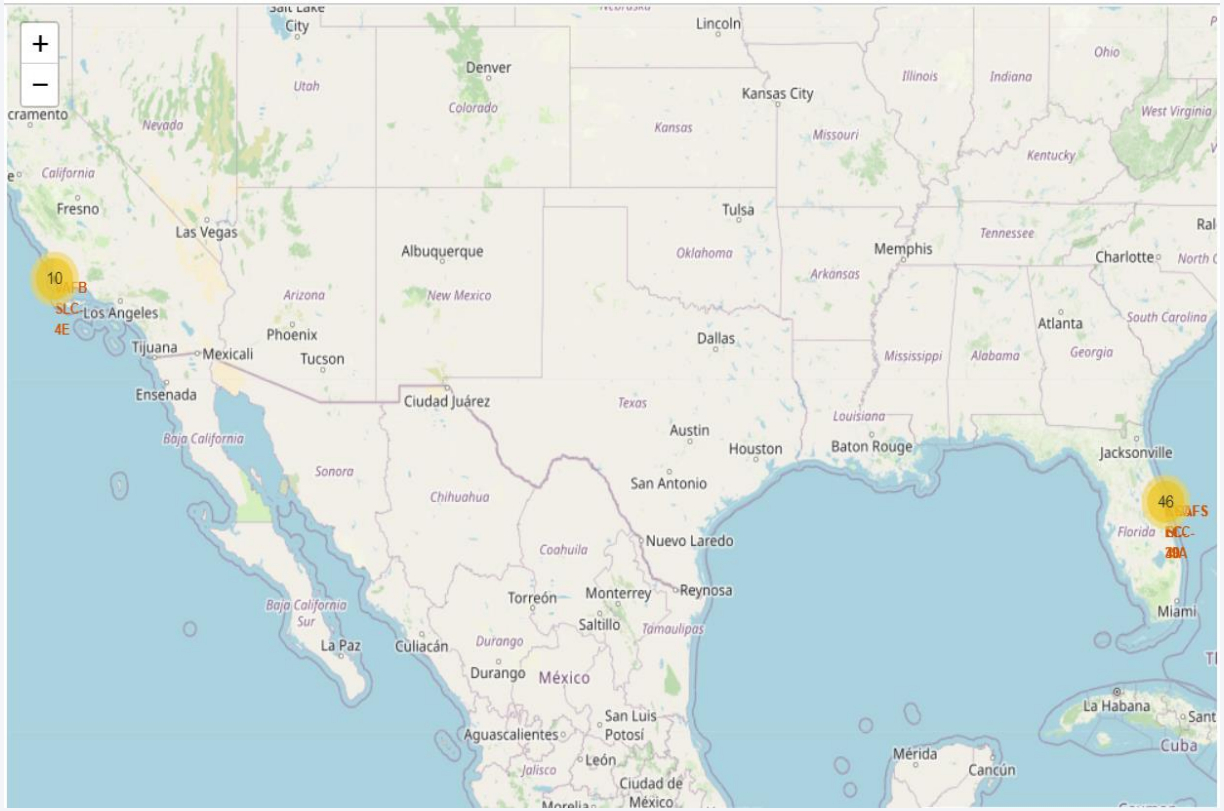
All Launch Sites' Location Markers

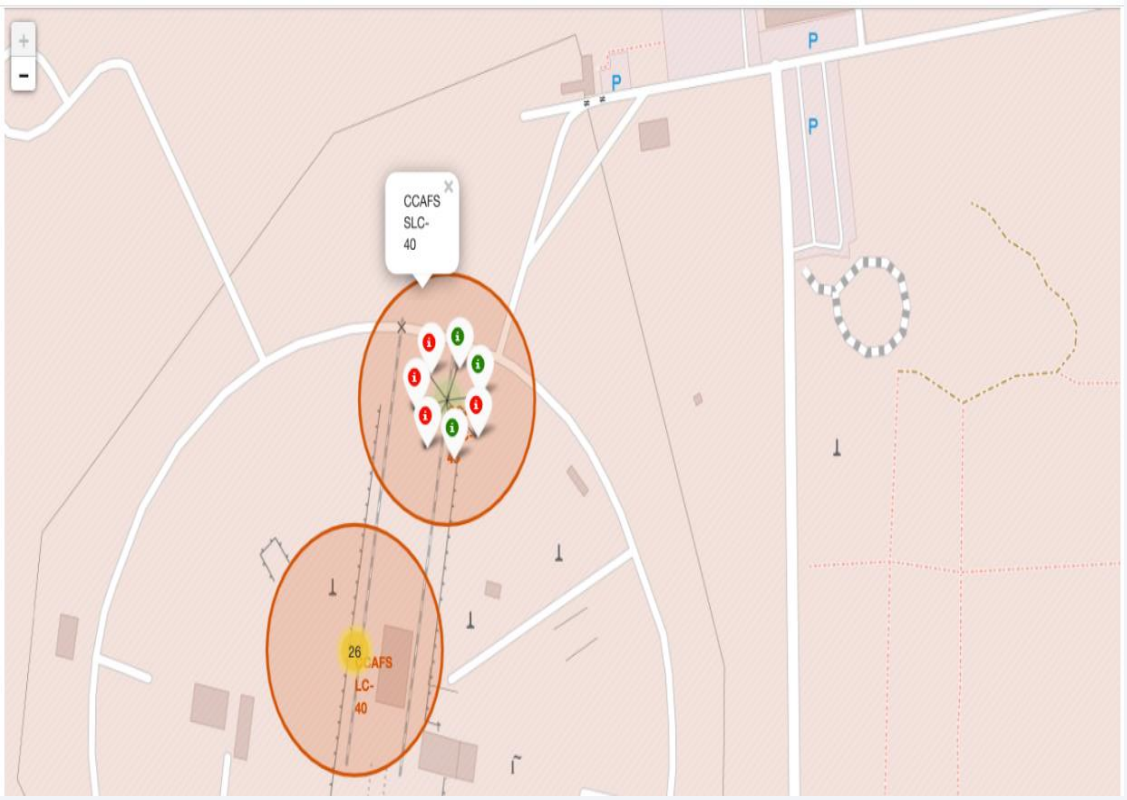
All launch sites are in very close proximity to the coast and into restricted areas.



Success/Failed Launches Marked on the Map

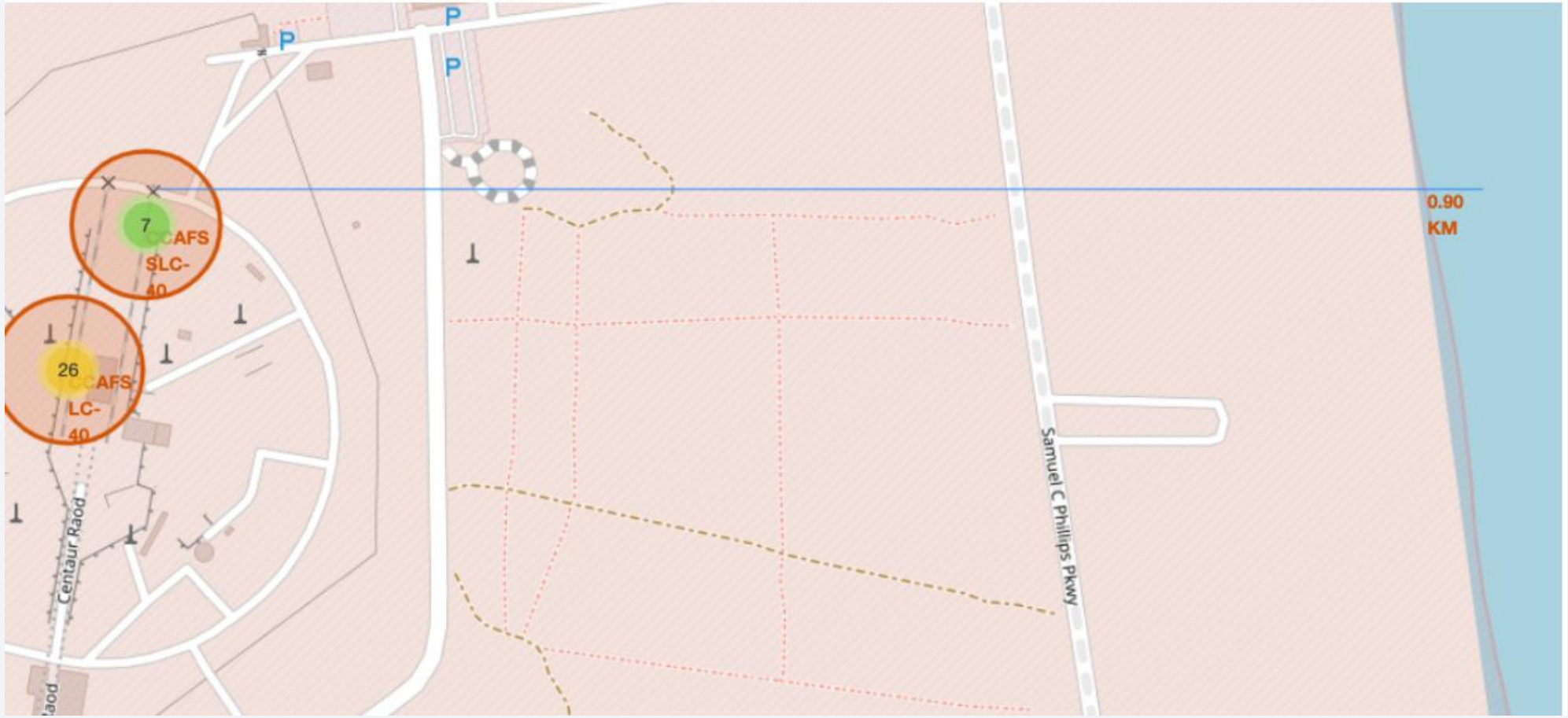
The first map shows clusters for every launch site, the second shows a green marker if a launch was successful, and a red marker if a launch was failed.

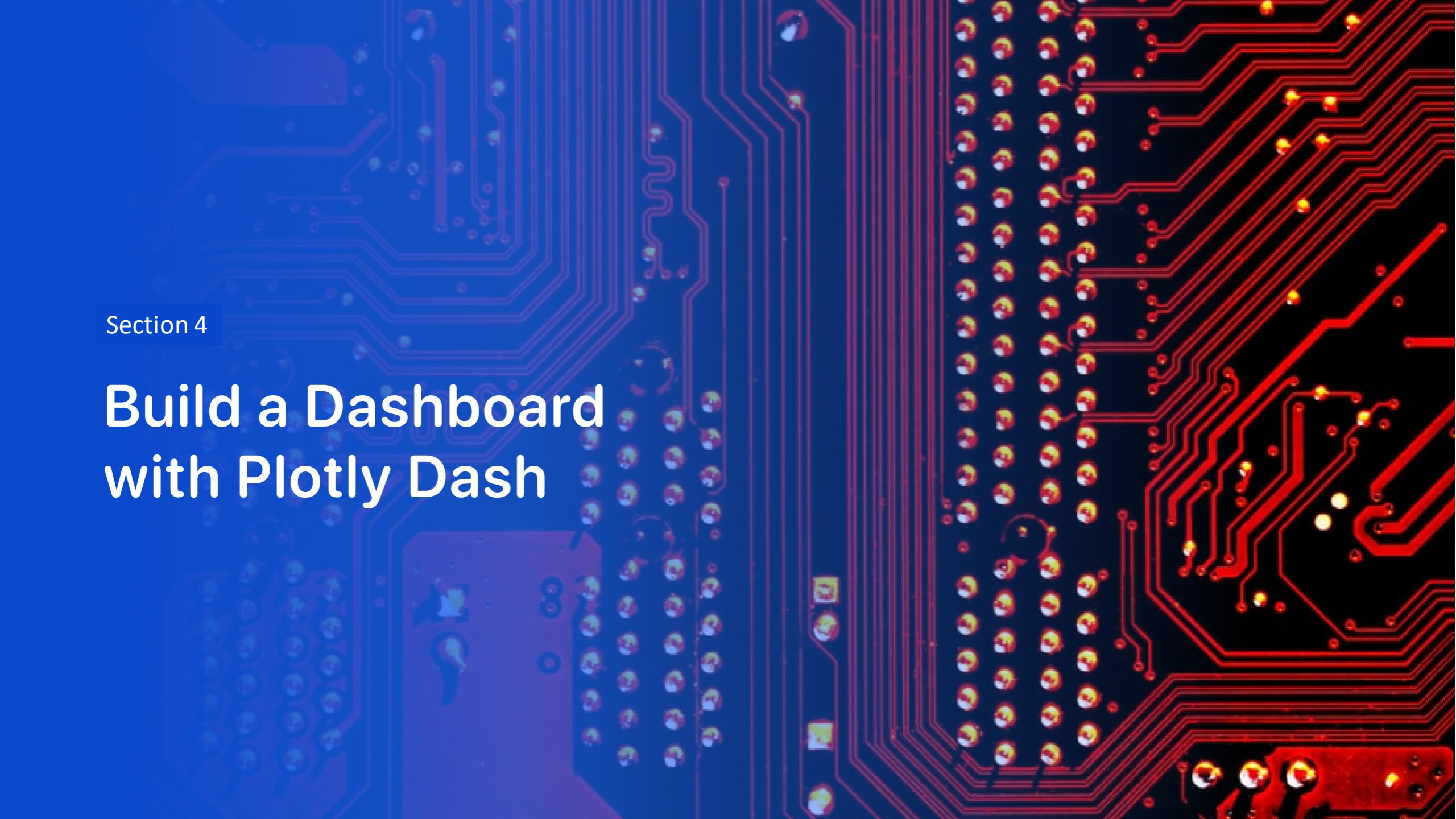




Distance between a launch site to its proximities

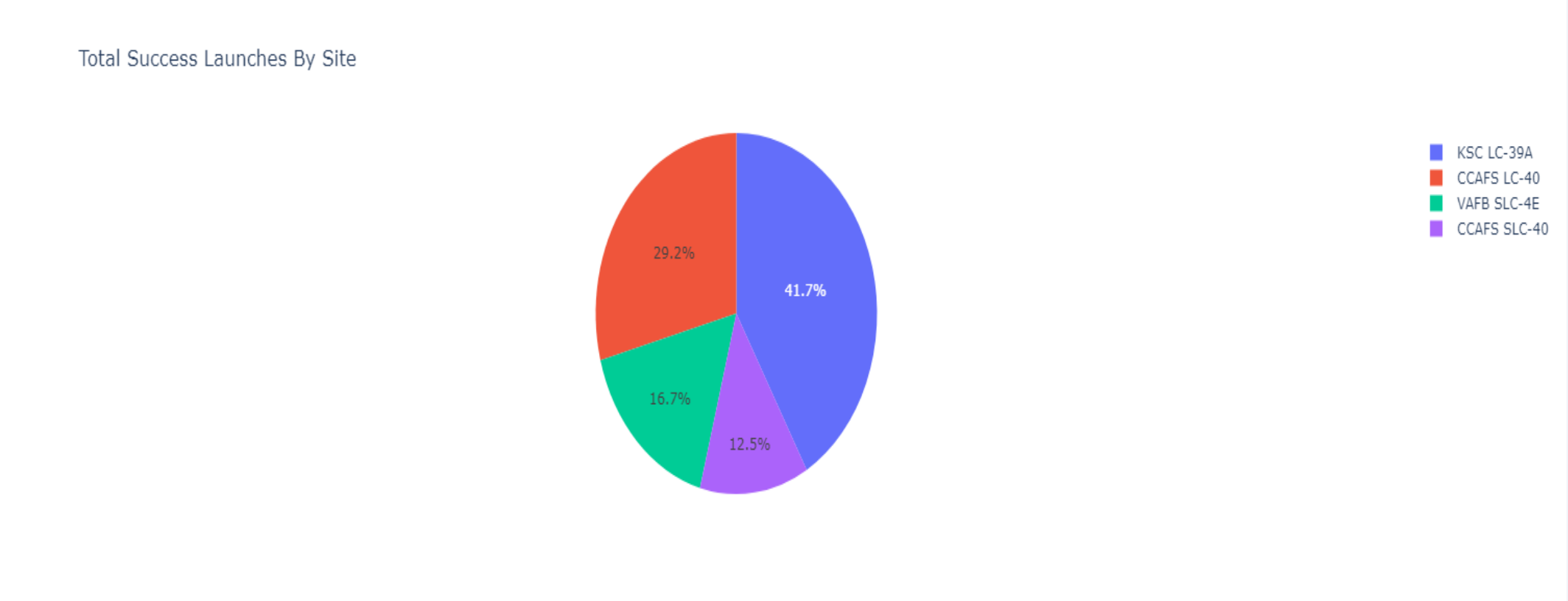
Launch sites are near to railways, roads, highways, and coastline.




Section 4
Build a Dashboard with Plotly Dash

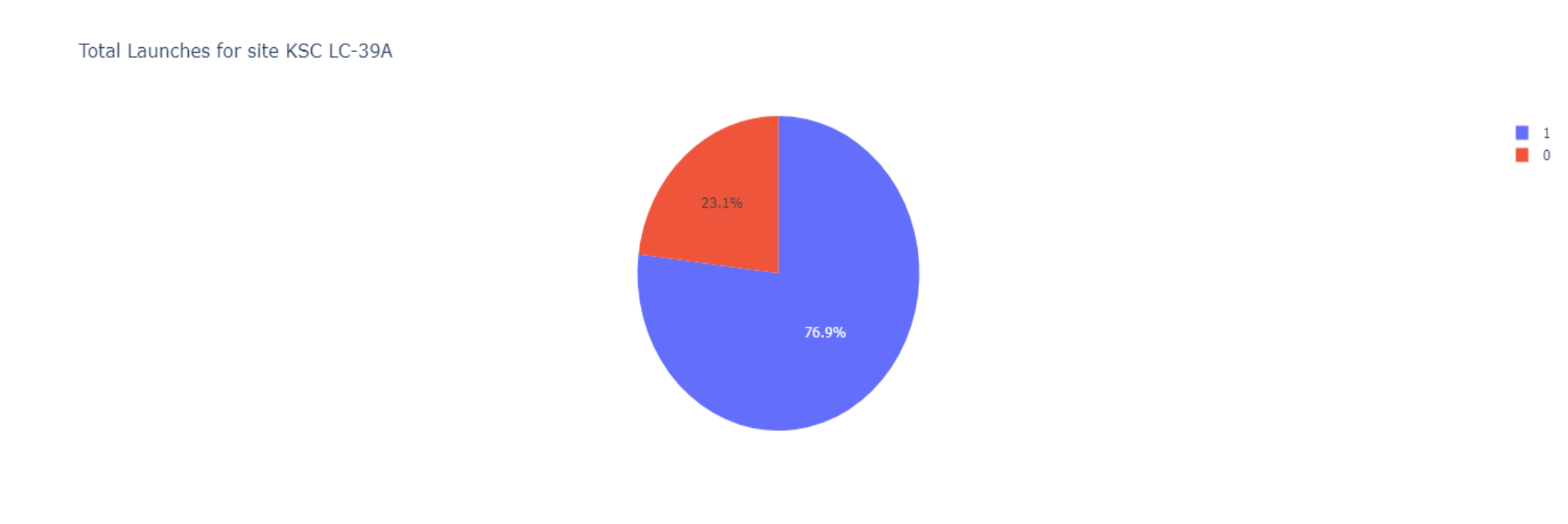
Total success launches by all sites

We can see that KSC LC-39A had the most successful launches from all the sites.



Success Rate by Site

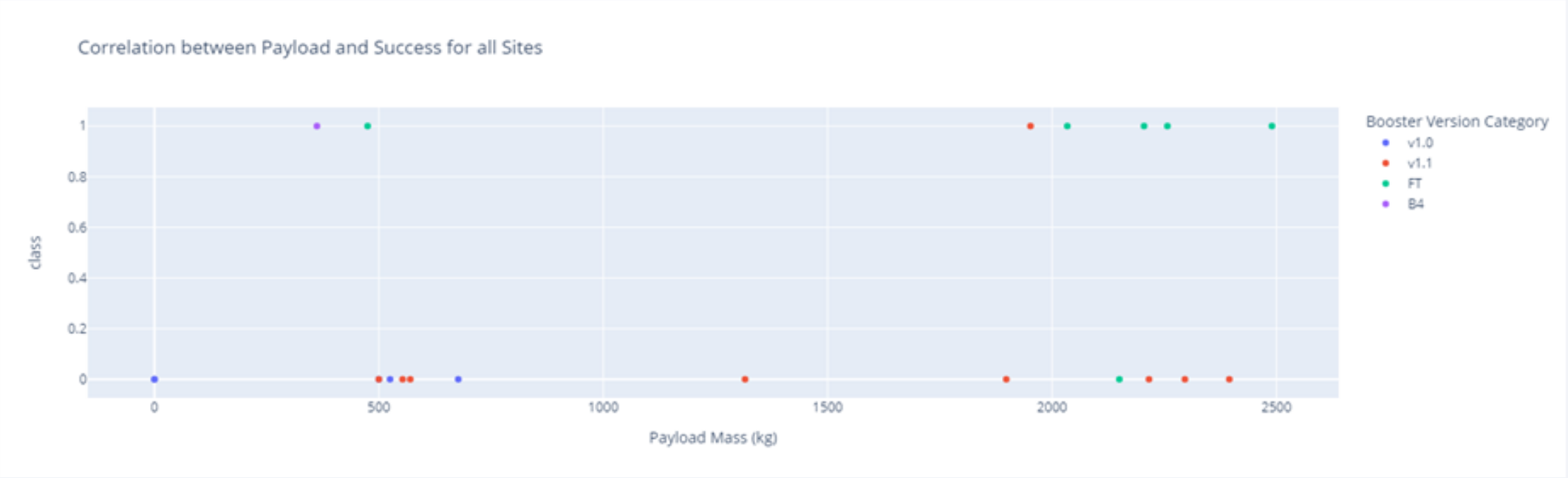
KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate.



Payload vs Launch Outcome

Scatter plot for all sites with 2500(kg), 5000(kg) and 10000(kg) payload ranges.

The 2500-5000(kg) range concentrate the majority of the successfully launches, the 0-2500(kg) Range has most failed launches but all three are similar.

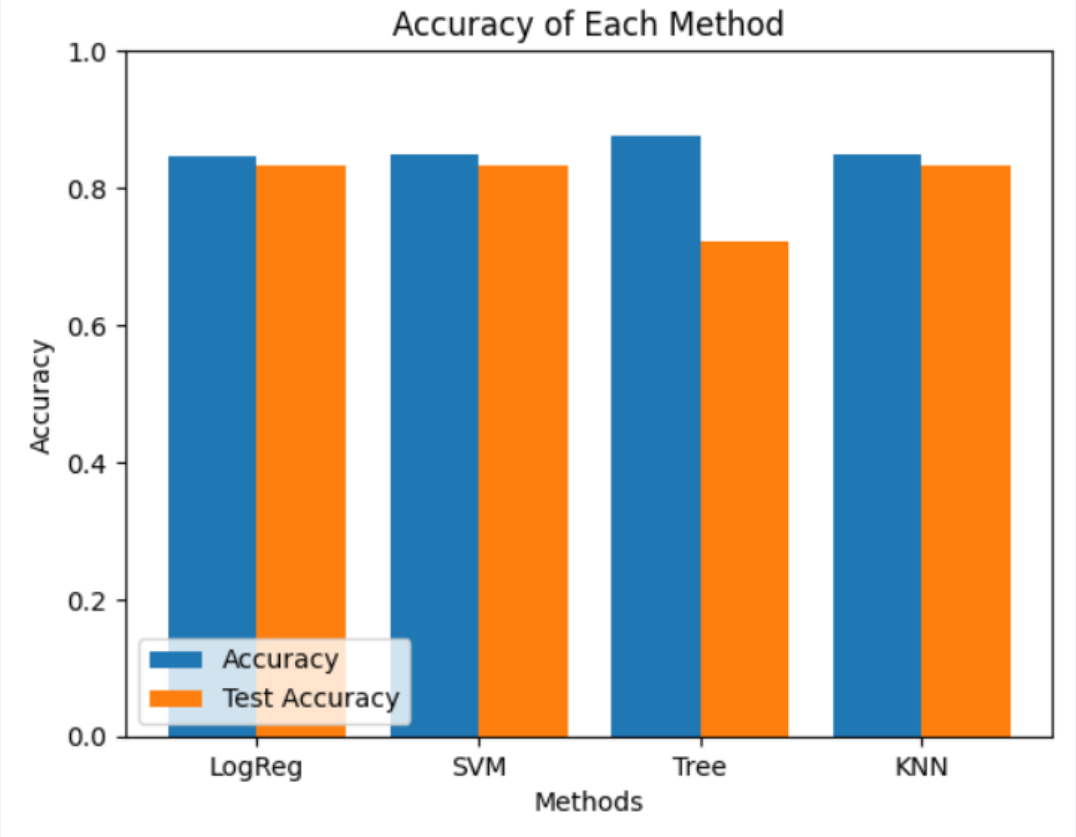


Section 5

Predictive Analysis (Classification)

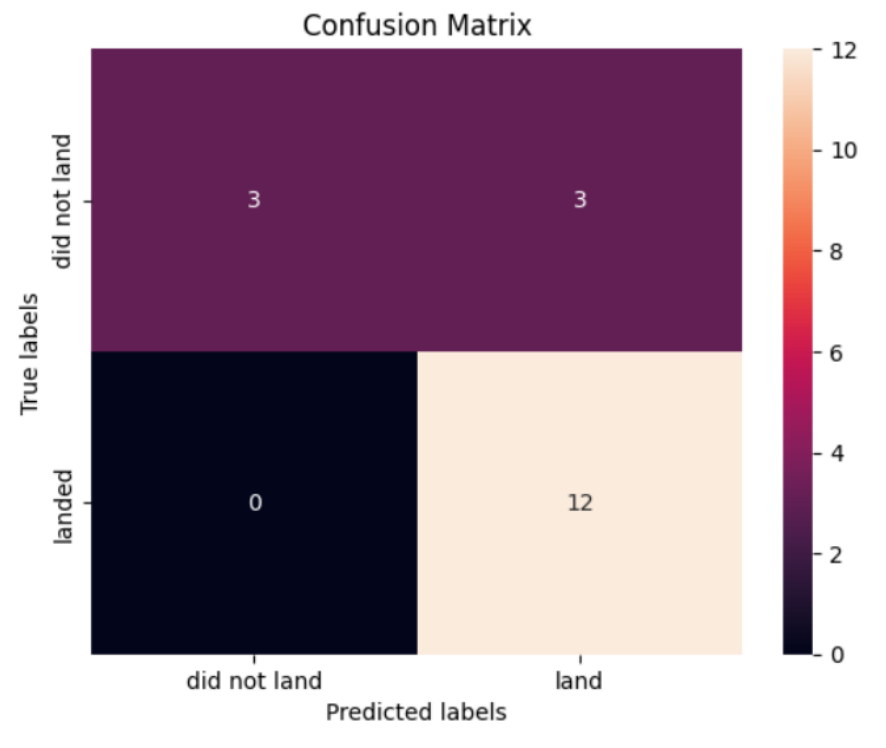
Classification Accuracy

Decision Tree has the highest accuracy with almost 0.89, then comes the remaining models with almost same accuracy of 0.84.



Confusion Matrix

Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low weighted payloads perform better than the heavier payloads.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
- KSC LC-39A had the most successful launches from all the sites.
- Orbit GEO, HEO, SSO, ES L1 has the best success rate.

Appendix

For notebooks, datasets and scripts, follow this GitHub Repository link:

https://github.com/Ajay15Khanna/Applied-Data-Science-Capstone

Thank you!