IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christopher Cumplido
10/17/2023

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

- According to SpaceX, the Falcon 9 rocket costs 62 million dollars while other providers cost up to 165 million dollars. SpaceX reduces most of its cost from the reuse of its rockets

Problems you want to find answers for

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

Section 1

Methodology

Methodology

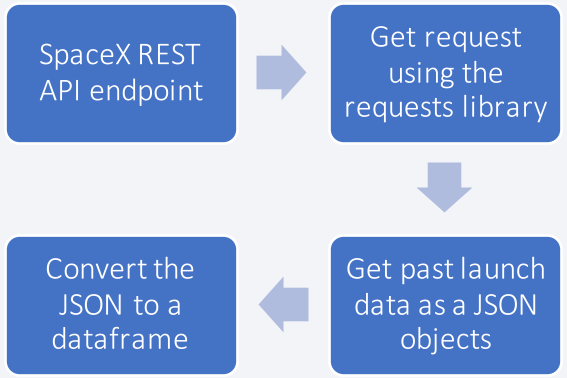
Executive Summary

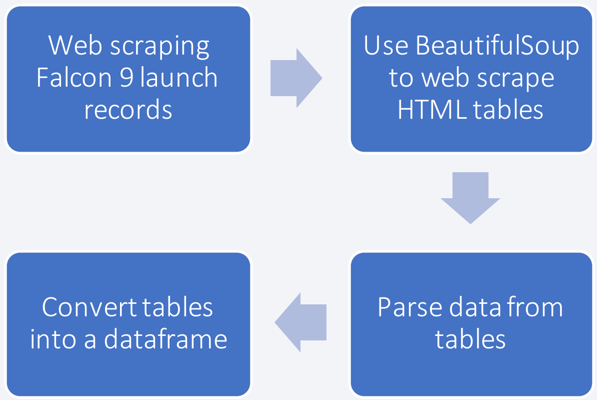
- Data collection methodology:
  - SpaceX Rest API
  - Web scraping from Wikipedia
- Perform data wrangling
  - One Hot Encoding data fields for machine learning and data cleaning of null values and irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive 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 were 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/
- Falcon 9 launch data can also be obtained by web scraping wikipedia using BeautifulSoup

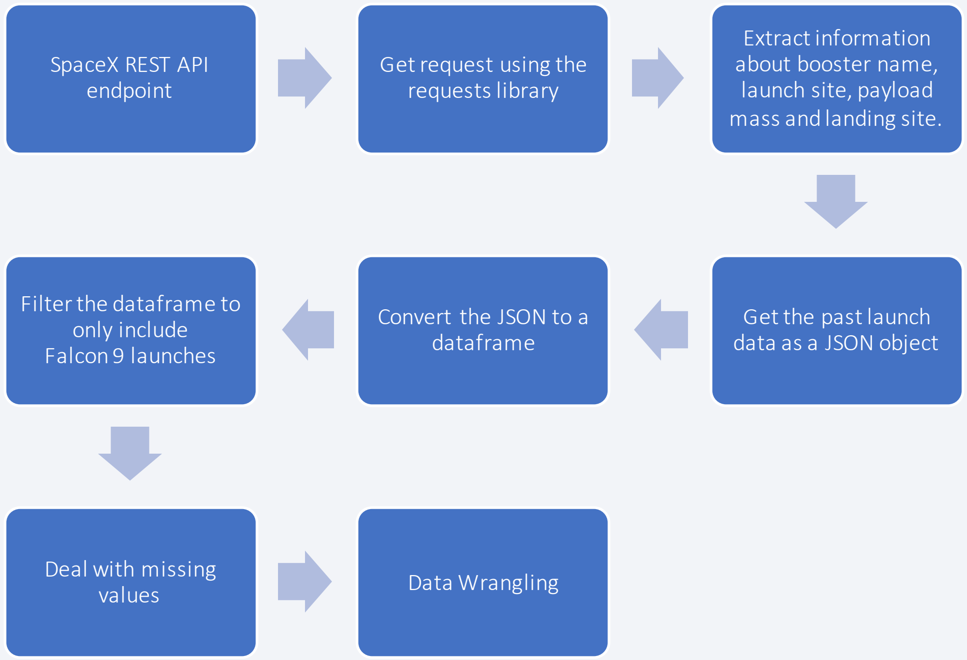




Data Collection – SpaceX API

- Data collection with SpaceX REST calls

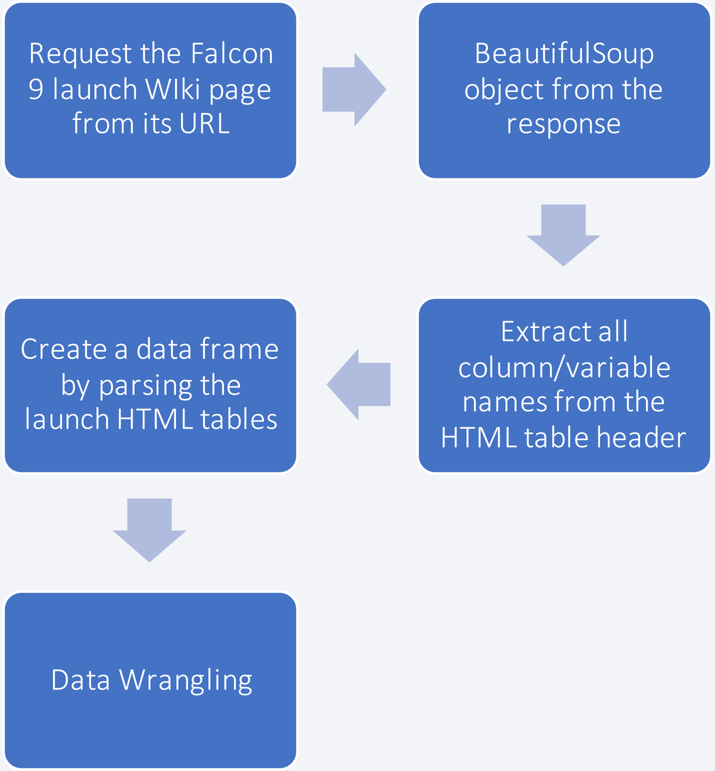
https://github.com/cwrite0/Final-Assignment/blob/954c67cc5e8143c2e8e0971c8dea7fe6549c7de4/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web scraping from WIkipedia

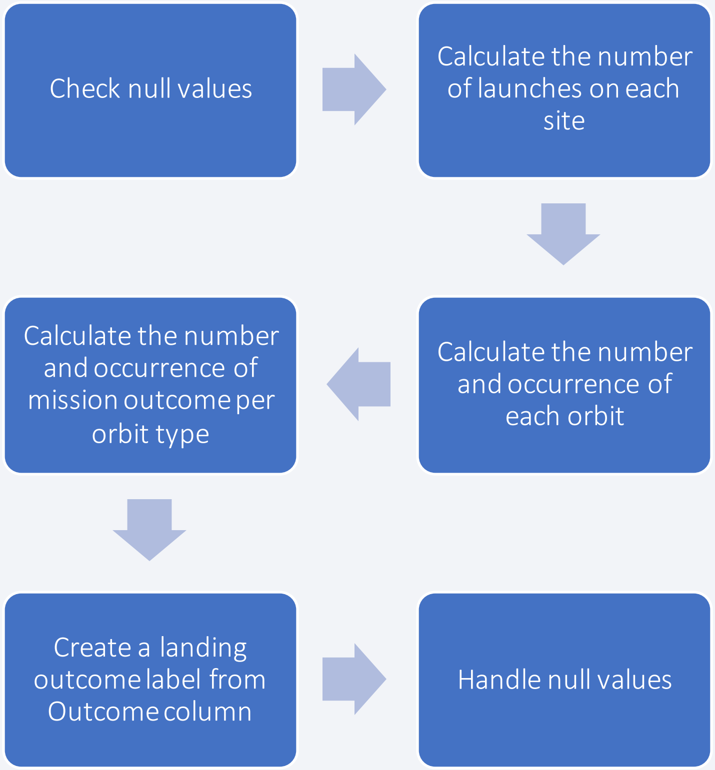
https://github.com/cwrite0/Final
-
Assignment/blob/954c67cc5e81
43c2e8e0971c8dea7fe6549c7de
4/jupyter-labs-
webscraping.ipynb



Data Wrangling

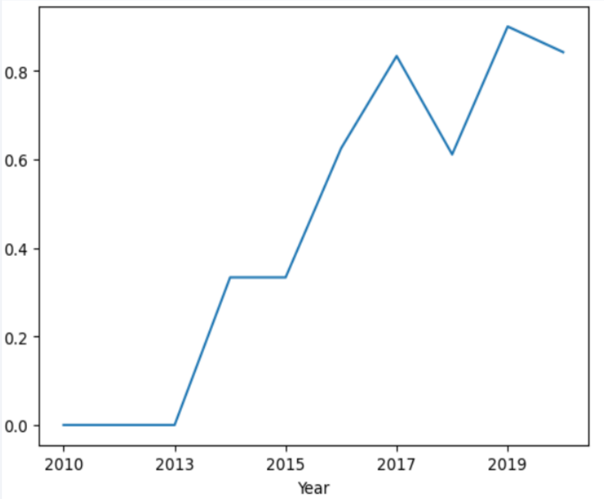
- https://github.com/cwrite0/Final-Assignment/blob/954c67cc5e8143c2e8e0971c8dea7fe6549c7de4/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

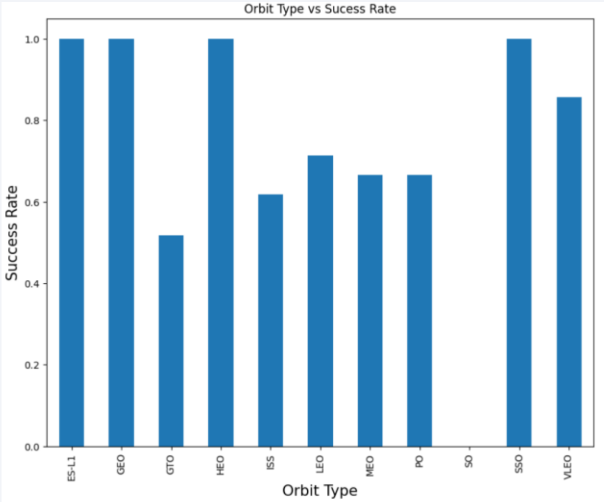
EDA Analysis

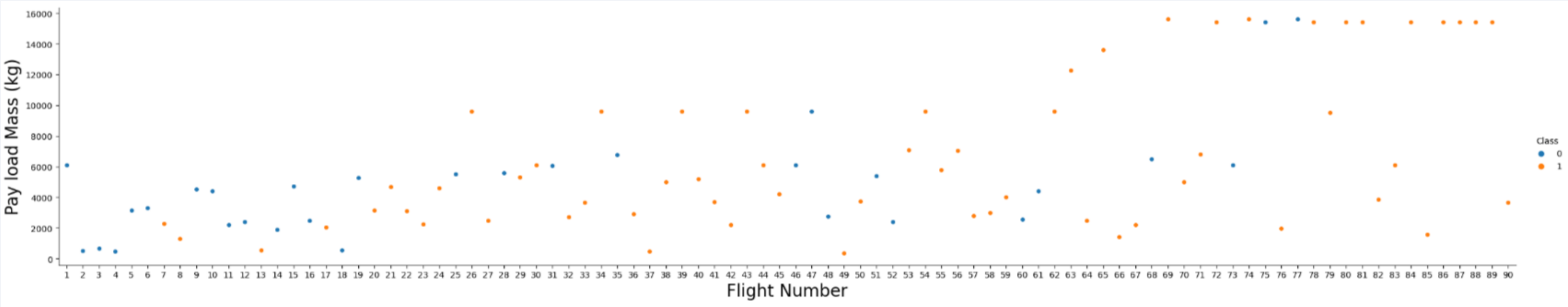


EDA with Data Visualization

- https://github.com/cwrite0/Final-Assignment/blob/954c67cc5e8143c2e8e0971c8dea7fe6549c7de4/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb







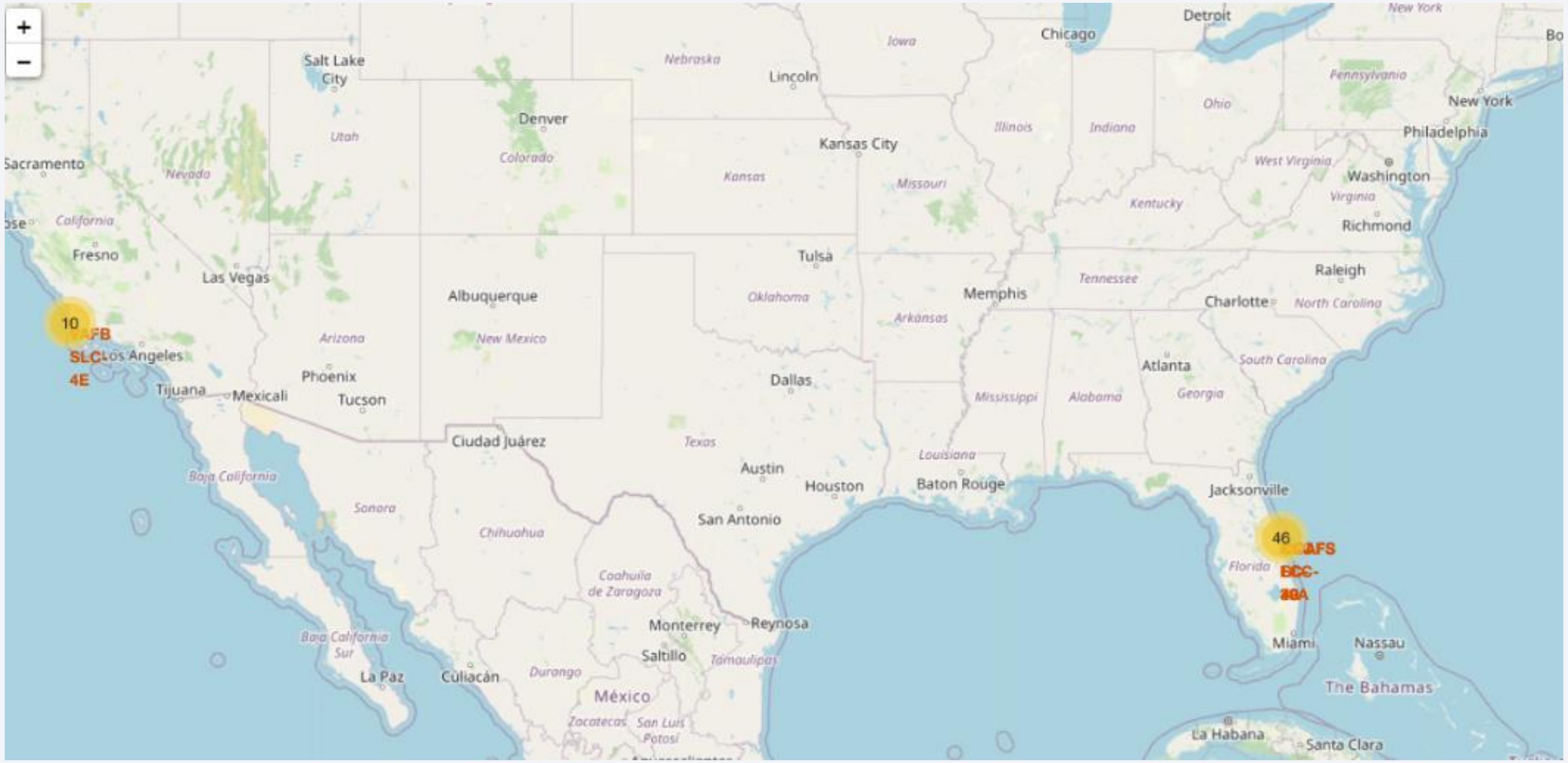
EDA with SQL

SQL queries performed:

- Names of unique launch sites in the space mission
- 5 records of launch sites beginning with the string 'KSC'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date where successful landing outcome in drone ship was achieved
- Names of boosters that had success in ground pad and have a payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions that carried the maximum payload mass
- Records of month names, successful landing outcomes on ground pad, booster versions, and launch site for the months of 2017

https://github.com/cwrite0/Final-Assignment/blob/954c67cc5e8143c2e8e0971c8dea7fe6549c7de4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

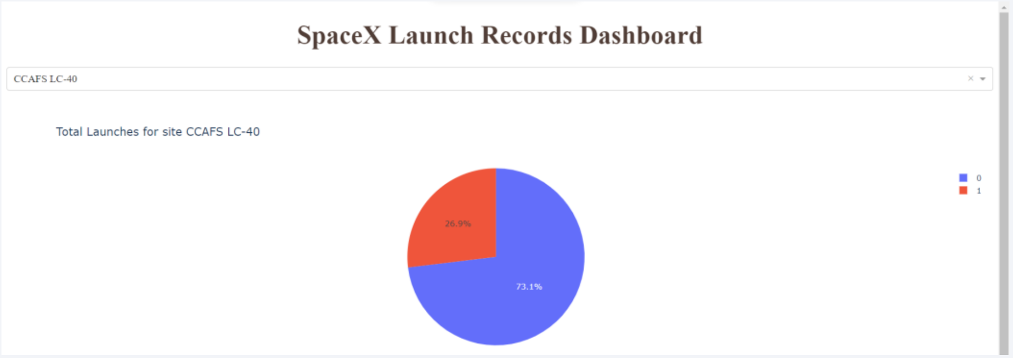
Build an Interactive Map with Folium

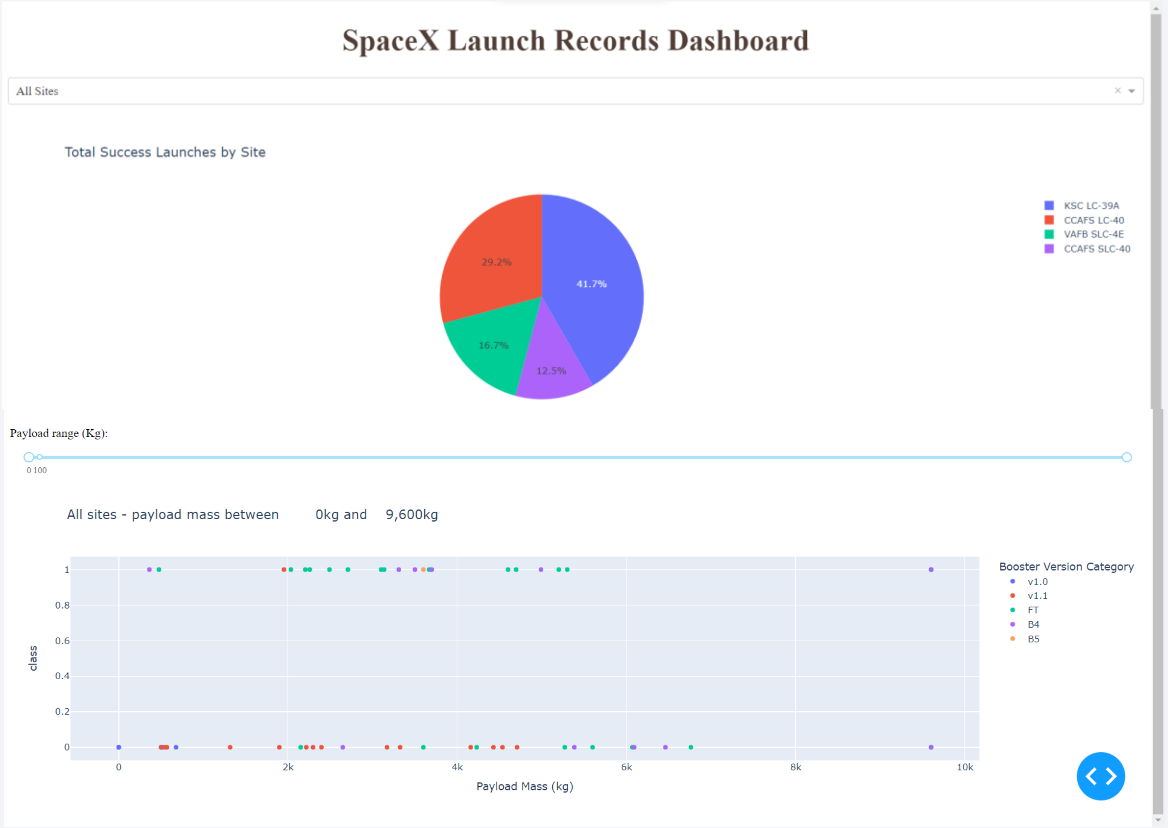


https://github.com/cwrite0/Final-Assignment/blob/954c67cc5e8143c2e8e0971c8dea7fe6549c7de4/lab_jupyter_launch_site_location.jupyterlite.ipynb

Map markers added to the interactive map to find the optimal location for building a launch site

Build a Dashboard with Plotly Dash





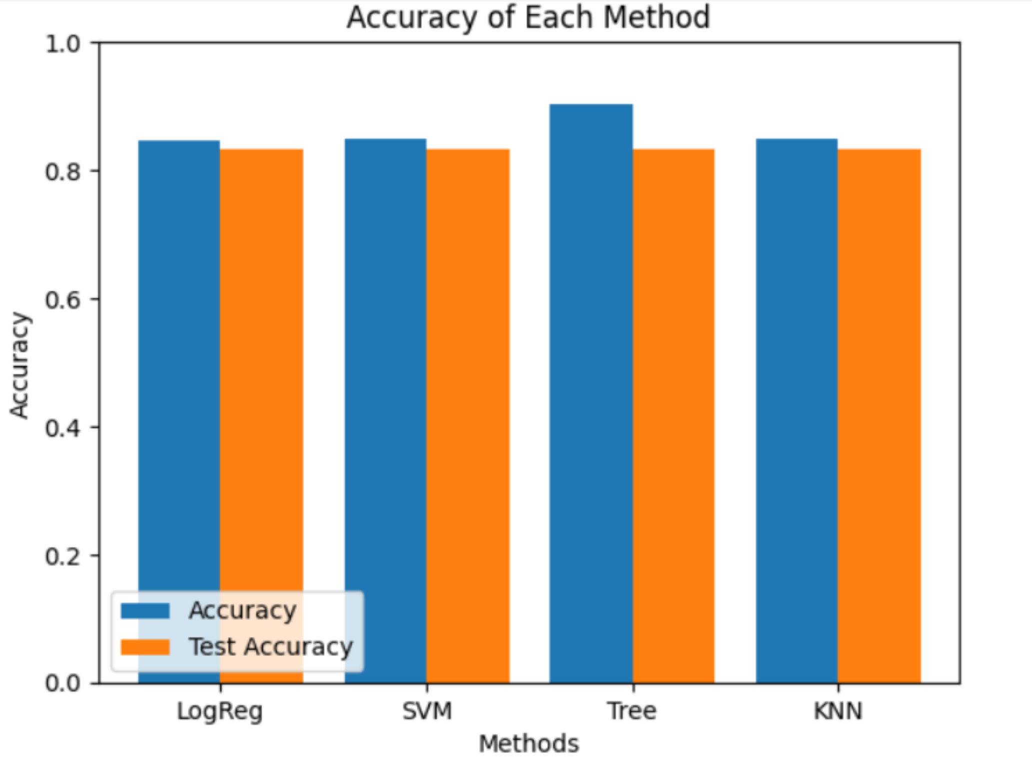
KSC LC-39A had the most successful launches from all sites

CCAFS LC-40 had a success rate of 73.1% and a failure rate of 26.9%

https://github.com/cwrite0/Final-Assignment/blob/954c67cc5e8143c2e8e0971c8dea7fe6549c7de4/spacex_dash_app.py

Predictive Analysis (Classification)

- The SVM, KNN, Decision Tree and Logistic Regression model achieved test accuracy at 83.33%, while the Decision Tree method performs the best at 90.36%



https://github.com/cwrite0/Final-Assignment/blob/954c67cc5e8143c2e8e0971c8dea7fe6549c7de4/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

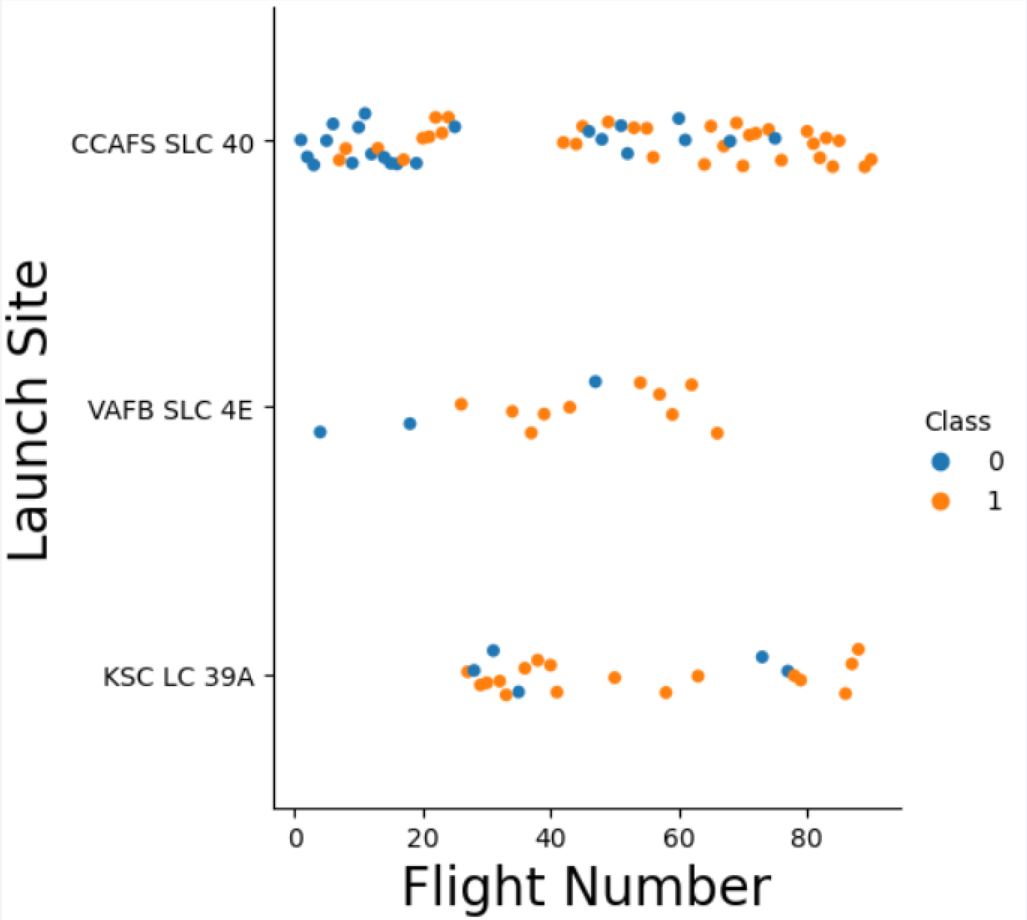
- The SVM, KNN, Decision Tree and Logistic Regression model achieved the best test accuracy
- Low weight payload has a better success rate than heavier payloads
- SpaceX launch success rate is closely related to time in years. More time, more successful launches
- KSC LC-39A had the most successful launches from all sites
- Orbit GEO, HEO, SSO, ES L1 had the highest success rate

Section 2

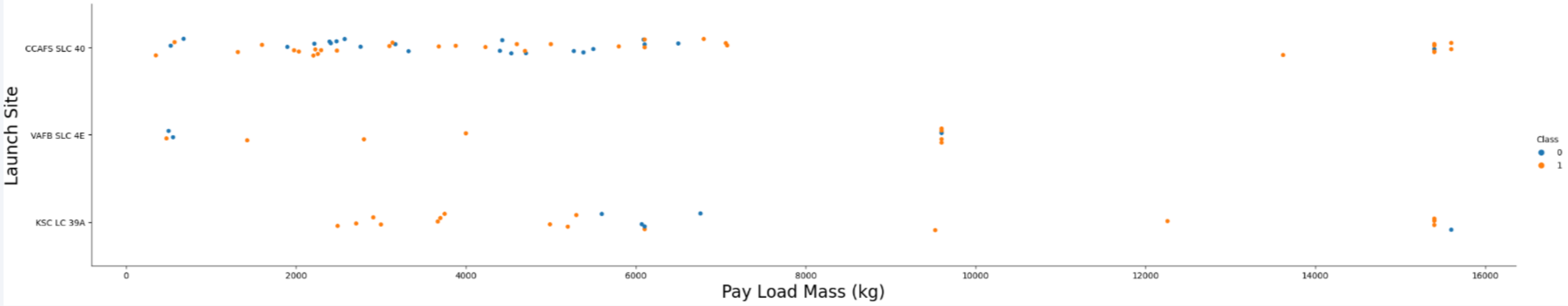
Insights drawn from EDA

Flight Number vs. Launch Site

- CCAFS SLC 40 has more launches than other sites



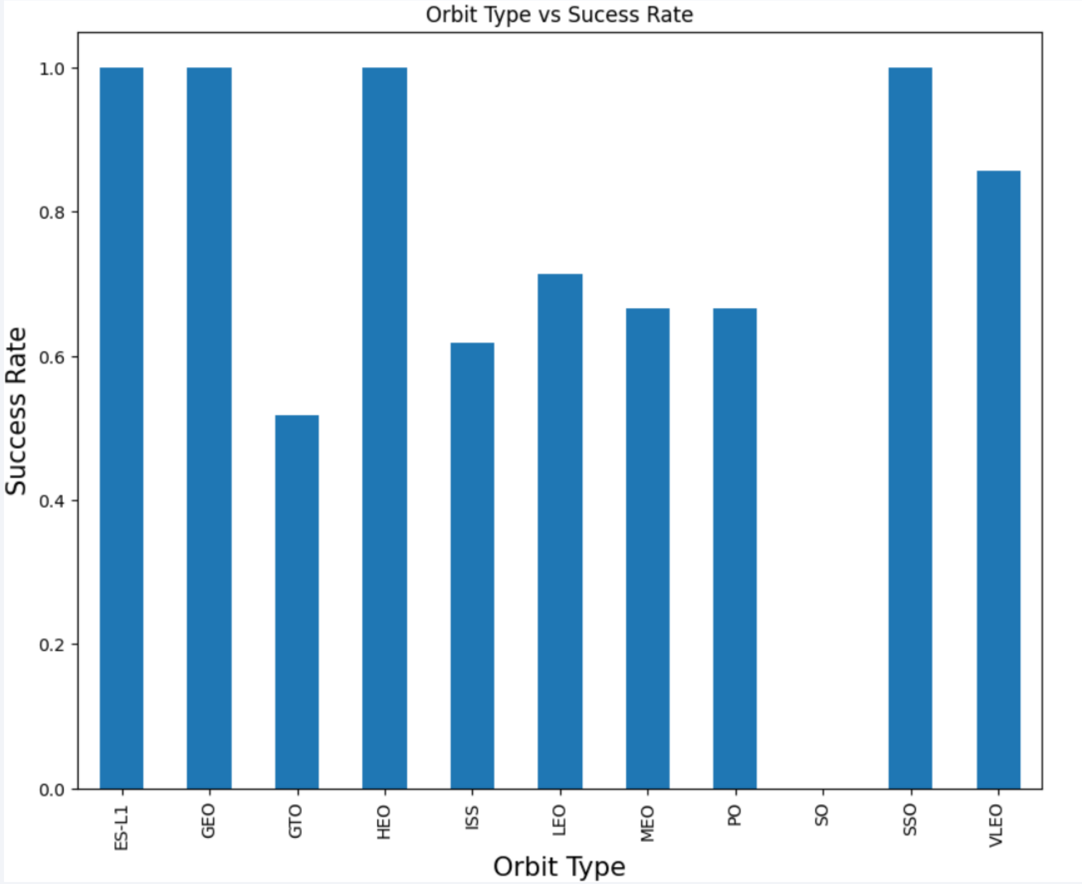
Payload vs. Launch Site



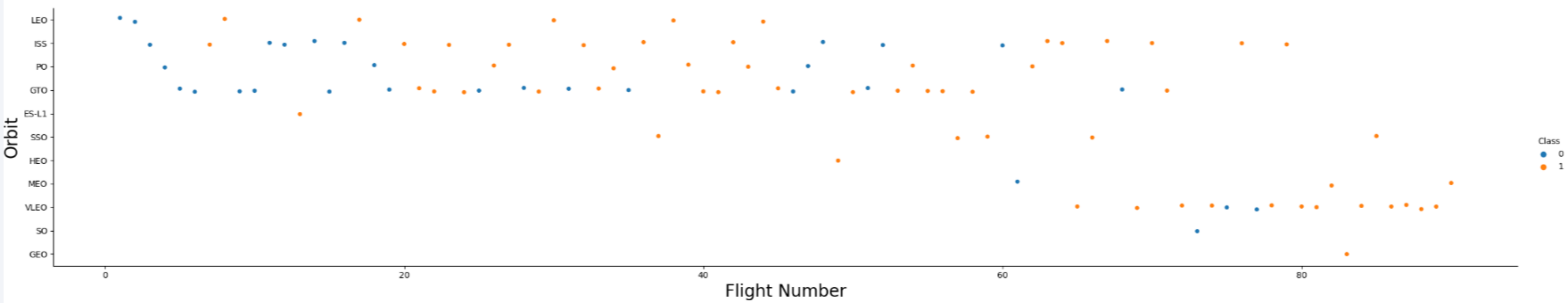
- Most low pay load mass launches occurred in CCAFS SLC 40

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO has the highest success rate

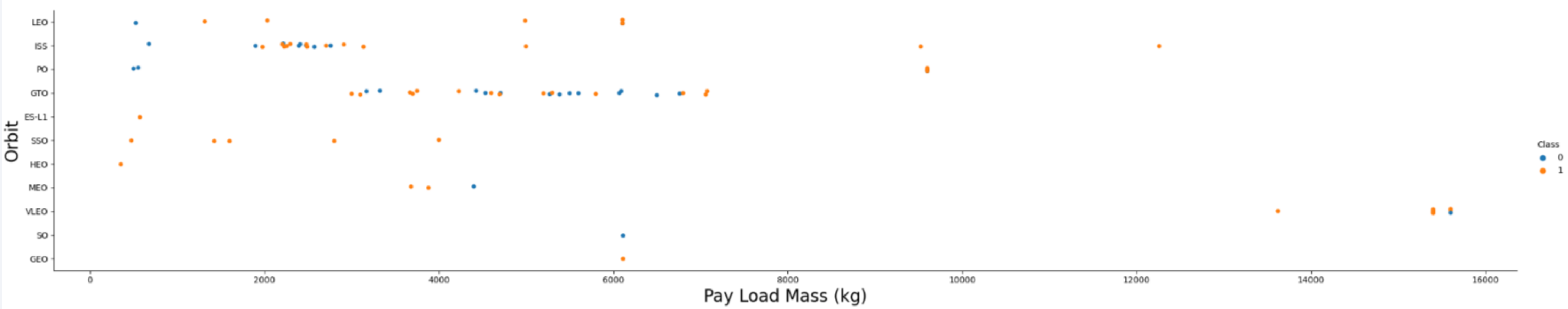


Flight Number vs. Orbit Type



- In the LEO orbit, 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

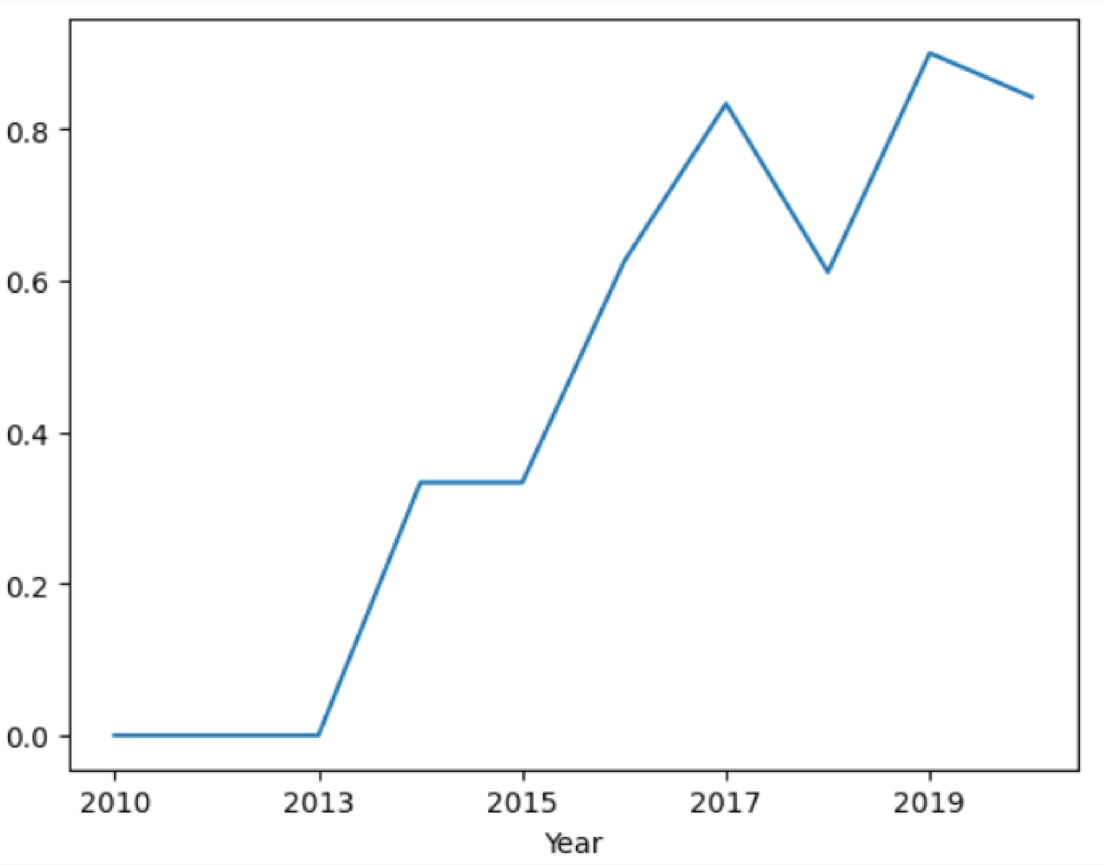
Payload vs. Orbit Type



- With a heavier pay load, we see a higher landing success rate with VLEO
- ISS and GTO we see a positive and negative landing rate between 2000-4000kg and 4000-8000kg, respectively

Launch Success Yearly Trend

- Launch success rate has consistently increased since 2013, implying a higher success rate over time



All Launch Site Names

- We can display the names of the unique launch sites in the space mission using DISTINCT with SQL

In [25]:
```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

Out[25]:

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

Launch Site Names Begin with 'CCA'

- By using LIMIT, we can limit the amount of launch site names that begin with 'CCA'

In [26]:

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

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

Out[26]:

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

- To obtain the sum of all pay load mass we can use SUM

```
In [27]: sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';

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

Out[27]:

TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1

- Using AVG we can obtain the average payload mass

```
In [18]: sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';

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

Out[18]:

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- By using MIN we can obtain the date of the first successful ground landing

```
In [19]: sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';

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

Out[19]:

FIRST_SUCCESS_GP
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Drone ship landing was determined a success using payload data of 4000-6000kg

```
In [20]: sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Succ

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

Out[20]:

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

Total Number of Successful and Failure Mission Outcomes

- Using COUNT allows us to determine the number of successful and failed missions

In [21]:
```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

Out[21]:

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

Boosters Carried Maximum Payload

- Using MAX, we can determine which boosters had the maximum payload mass

```
In [22]:  sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

 * sqlite:///my_data1.db
Done.
Out[22]:
```

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 can use substr(Date,6,2) as month to determine the months
- And use substr(Date,0,5)='2015' for year

```
In [23]: %%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.
```

Out[23]:

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

- Using ORDER, we can organize the values in descending order
- Using COUNT, we can count all numbers

```
In [24]: sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDIN(
 * sqlite:///my_data1.db
Done.
Out[24]:
```

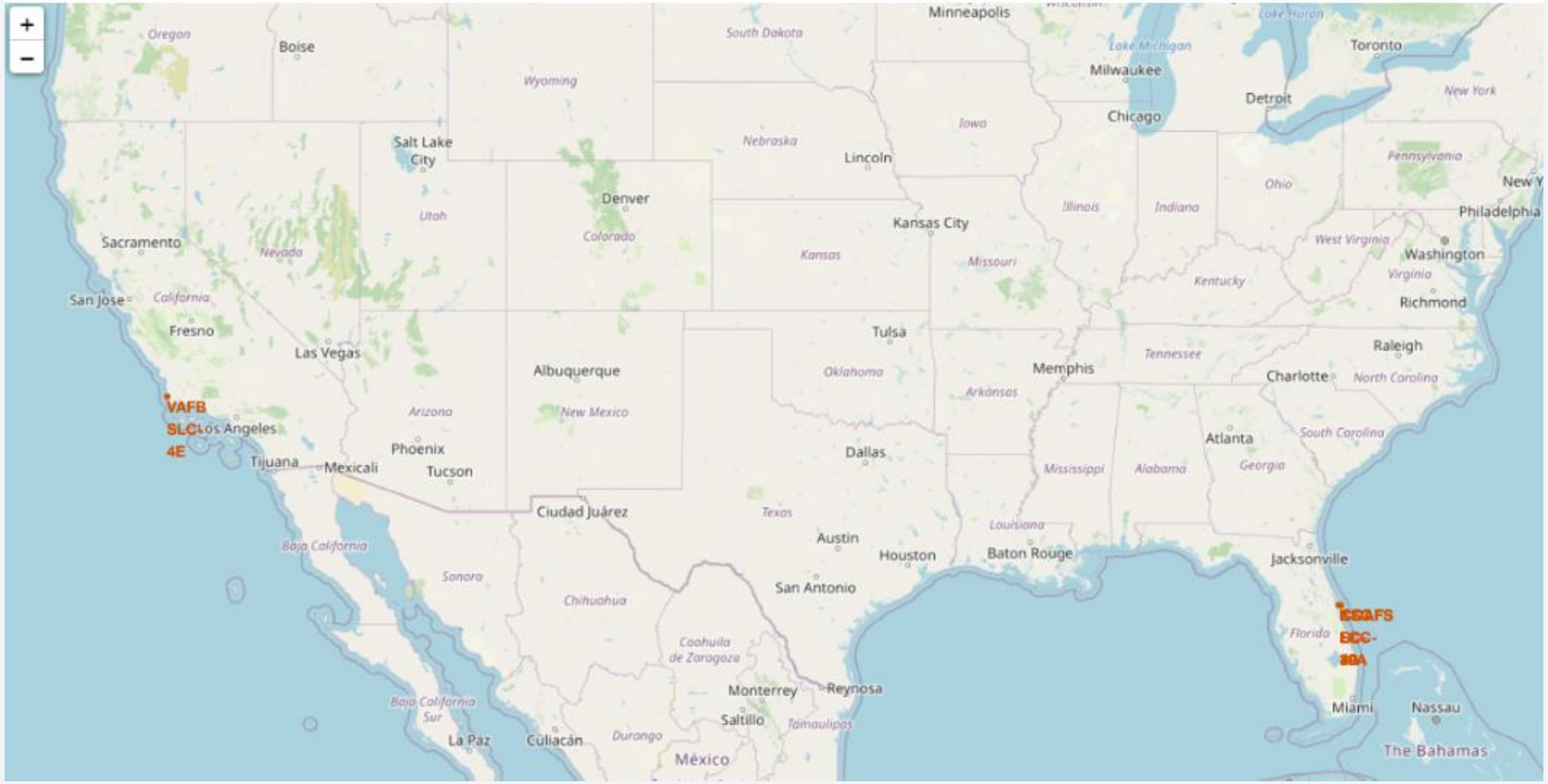
Landing_Outcome	QTY
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

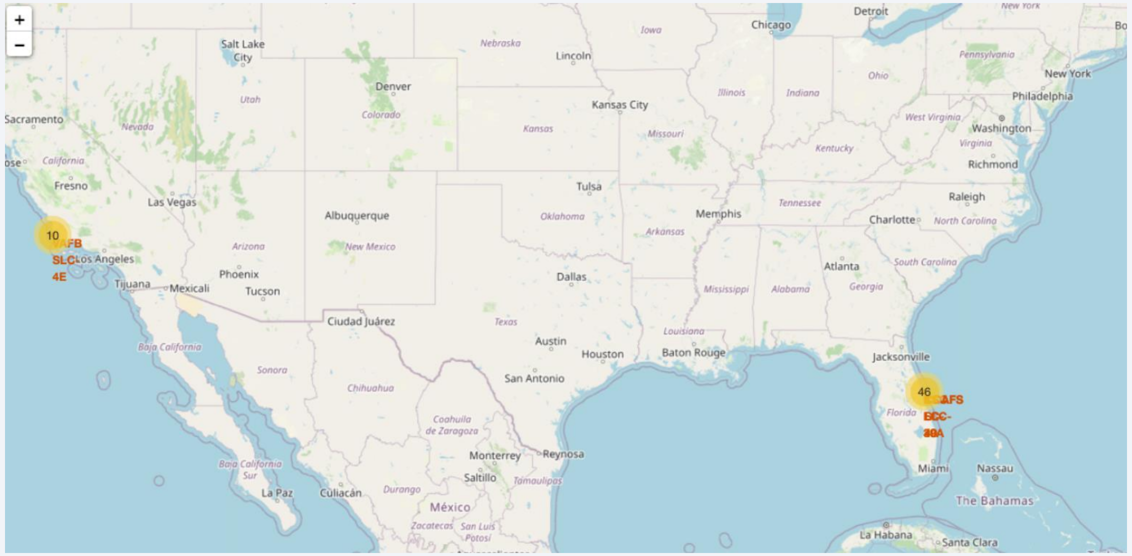
Launch Site Location Markers

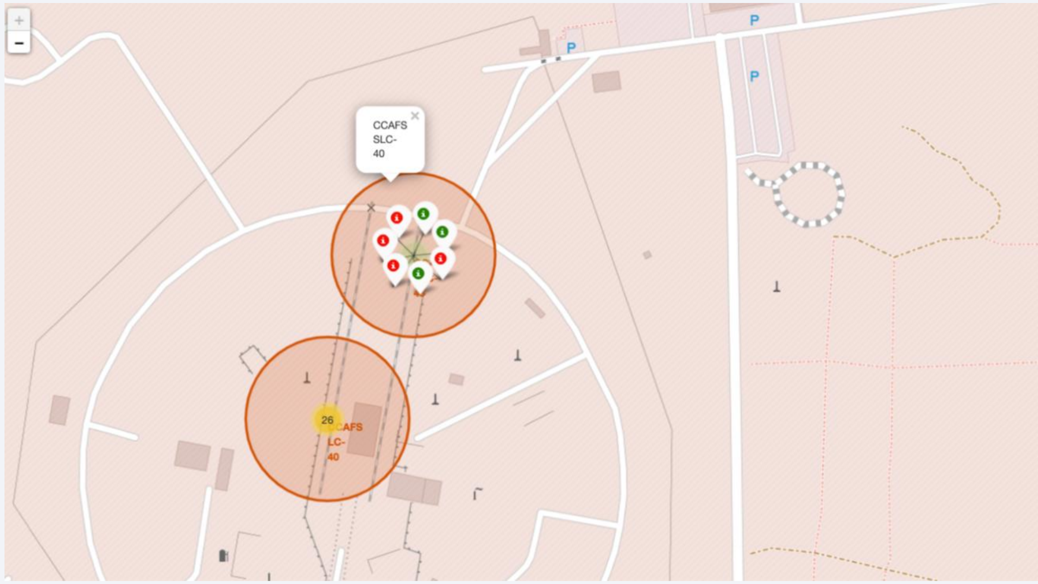
- All launch sites are in close proximity to the coast



Success/Failed Launch Site Locations

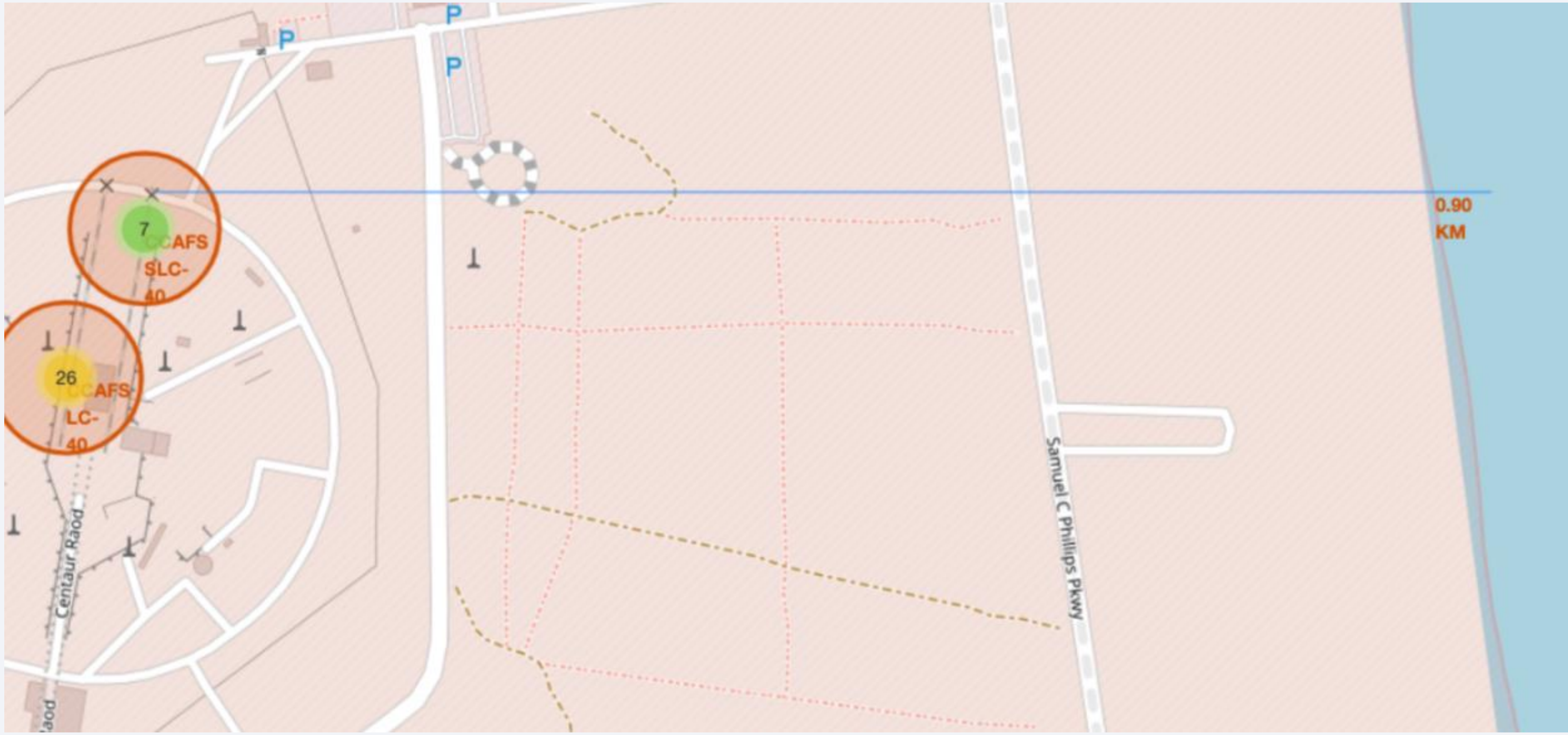
- Clusters are shown for every launch site
- Green markers represent a successful launch while a red marker represents a failed launch





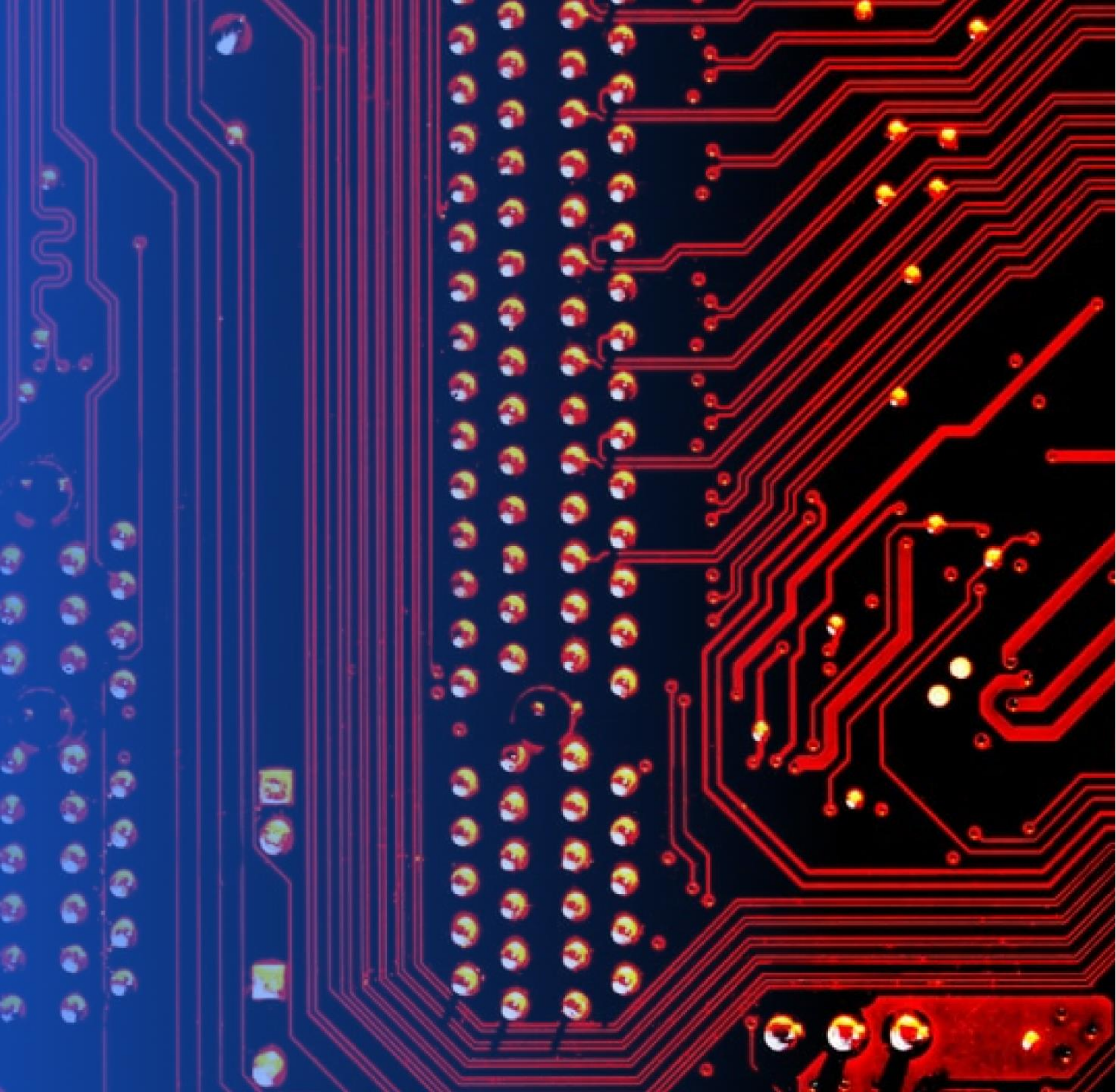
Distance of Launch Site to its Proximities

- Launch sites are close to railways, roads, highways, and the coastline



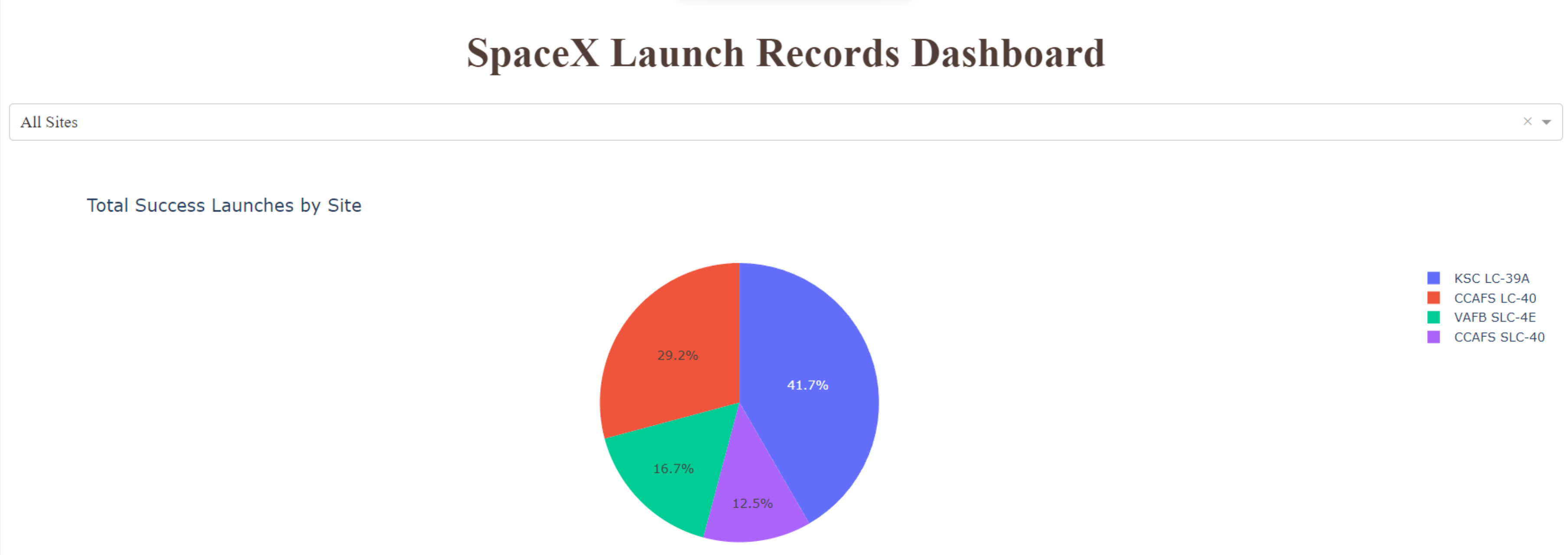
Section 4

Build a Dashboard with Plotly Dash



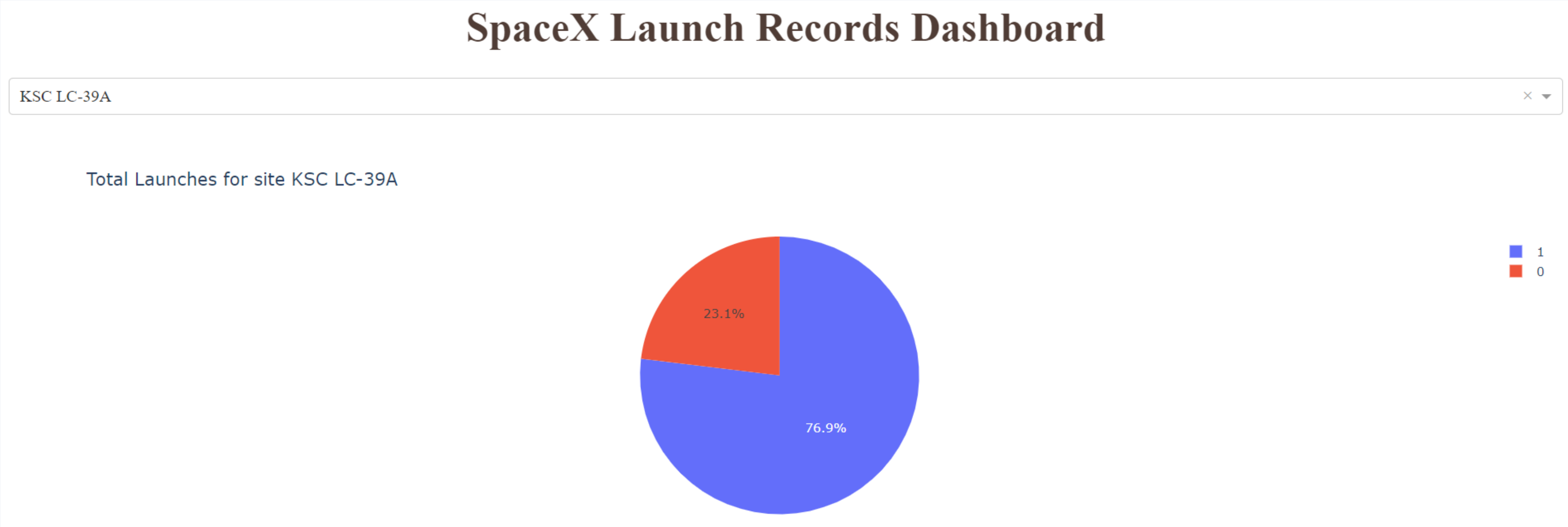
Total Success Launches by Site

- KSC LC-39A is shown to have the most successful launches



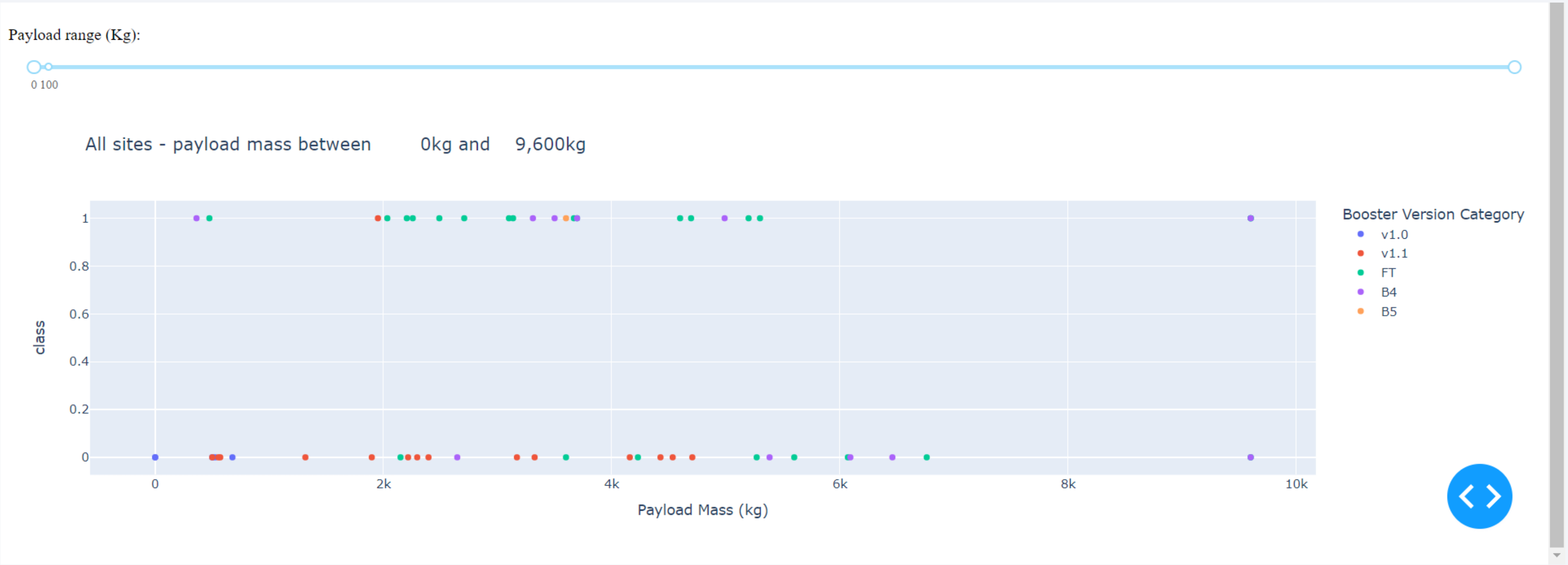
KSC LC-39A Success Rate

- KSC LC-39A had the highest success rate of 76.9% and a failure rate of 23.1%



Payload vs Launch Outcome

- The 2000-4000kg range has the majority of successful launches
- The 0 to 4500kg range has the majority of failed launches

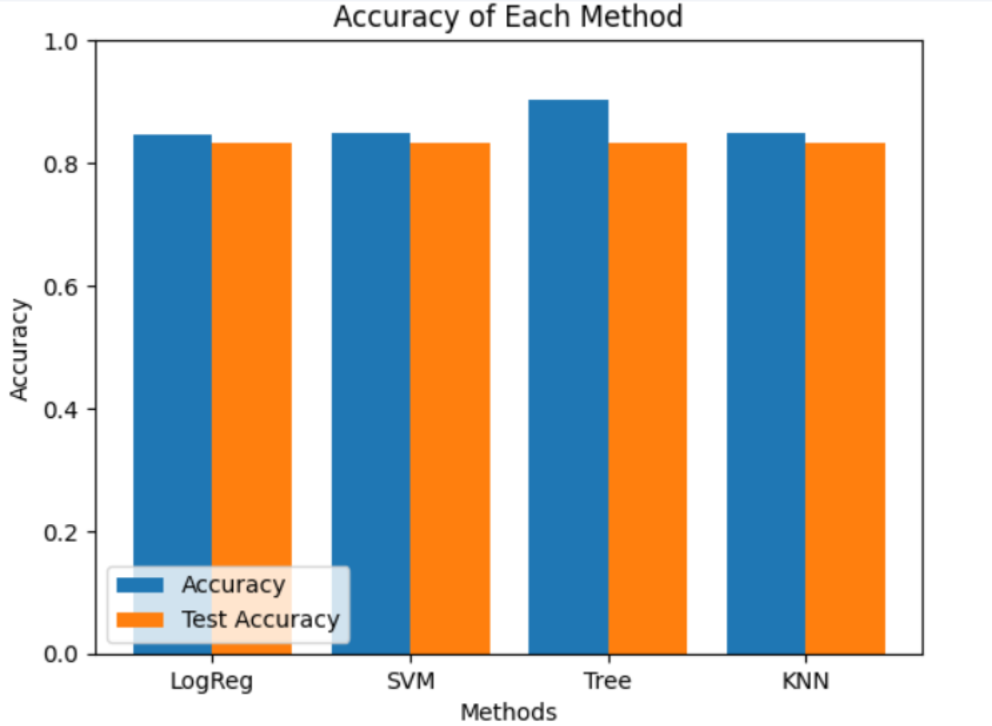


Section 5

Predictive Analysis (Classification)

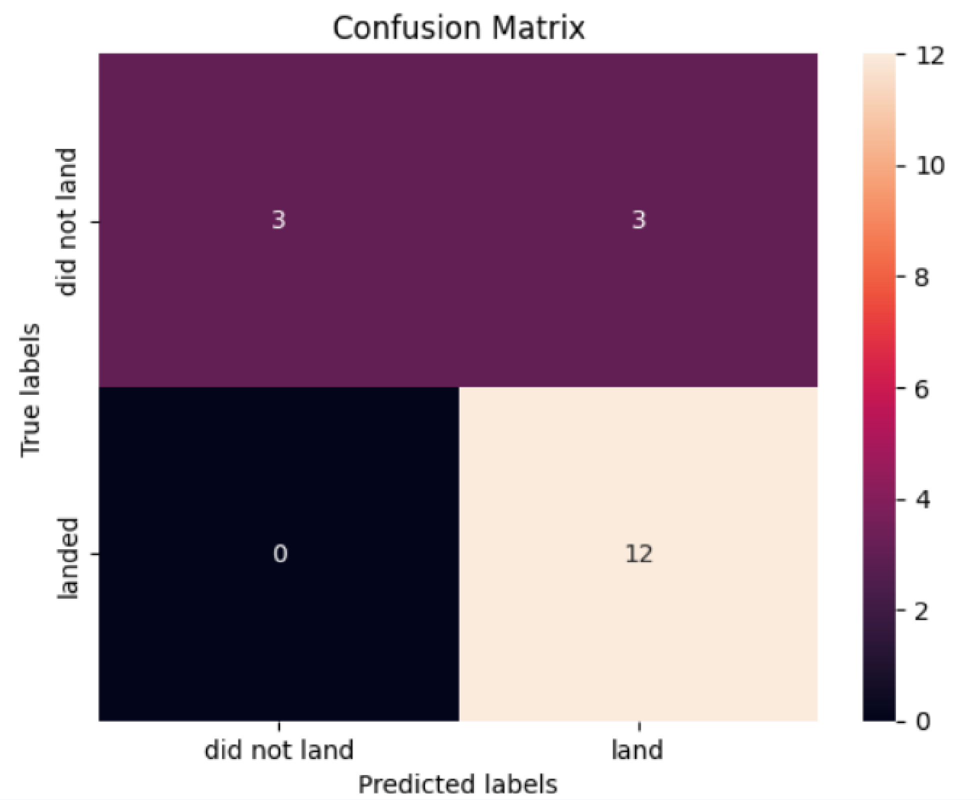
Classification Accuracy

- The Decision Tree method has the highest accuracy rate of 0.9 followed by the rest of the methods with an accuracy rate of 0.8



Confusion Matrix

- The confusion matrix for the decision tree method tells us that it is able to tell apart the different classes
- False positives are an issue, however



Conclusions

- The Decision Tree method has the best prediction accuracy for our dataset
- Payloads of a lower weight have a higher success rate than heavier
- The rate of successful SpaceX launches consistently improves over time
- Orbit GEO, HEO, SSO, and ES L1 had the highest success rate

Appendix

- Please follow this repository link for notebooks, datasets, and scripts:

https://github.com/cwrite0/Final-Assignment.git

Thank you!