Winning Space Race with Data Science

Tianyuan Zhu
July 23, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection: SpaceX API and web scraping
  - Data wrangling: Deal with missing values, calculate variables for analysis
  - Data visualization: different charts - scatter, bar, line. interactive visual analytics using Folium and Plotly Dash
  - Prediction models investigated: K Nearest Neighbors, Decision Tree, SVM, Logistic Regression
- Summary of all results
  - Location of launch sites are proximity to Equator line, coast and railway.
  - Whether the landing is successful or not are associated with launch site, payload mass, orbit type, boost version, and time.
  - Launch site KSC LC-39A has the highest success rate which is around 77%.
  - The launch success rate increases from 2013 to 2020.
  - The decision tree has the highest classification accuracy which is about 94.4%. Thus, it should be used as the classification method for landing prediction.

Introduction

- Project background and context
  - Companies are making space travel affordable for everyone.
  - SpaceX Falcon 9 rocket: Costs 62 million dollars. Over100 million dollars less than the cost of other providers since SpaceX can reuse the first stage if it lands successfully.
  - Our company (SpaceY) would like to compete with SpaceX.
- Problems you want to find answers
  - Whether SpaceX will reuse the first stage
  - Cost of each launch for SpaceX

Section 1

Methodology

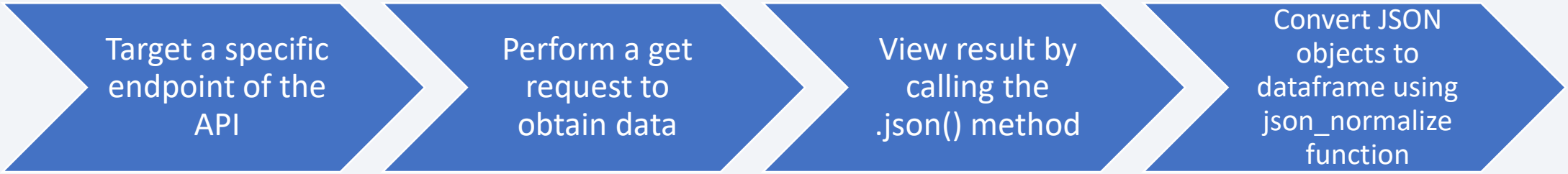
Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API

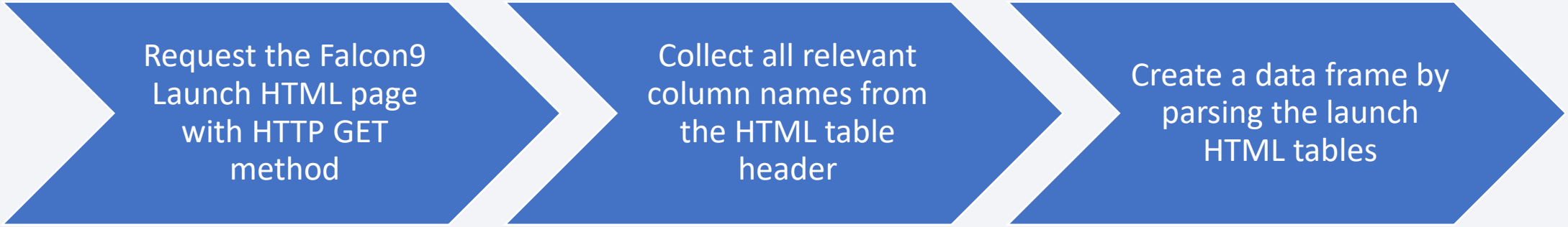
- The SpaceX launch data is collected using an API called SpaceX REST API. The process is shown in the following flowchart:



GitHub URL of the completed SpaceX API calls notebook:
https://github.com/bunny-tz/IBM_DS_Capstone/blob/b96d815f1fae2106d8b6897d6209b9eb403dce63/w1_jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

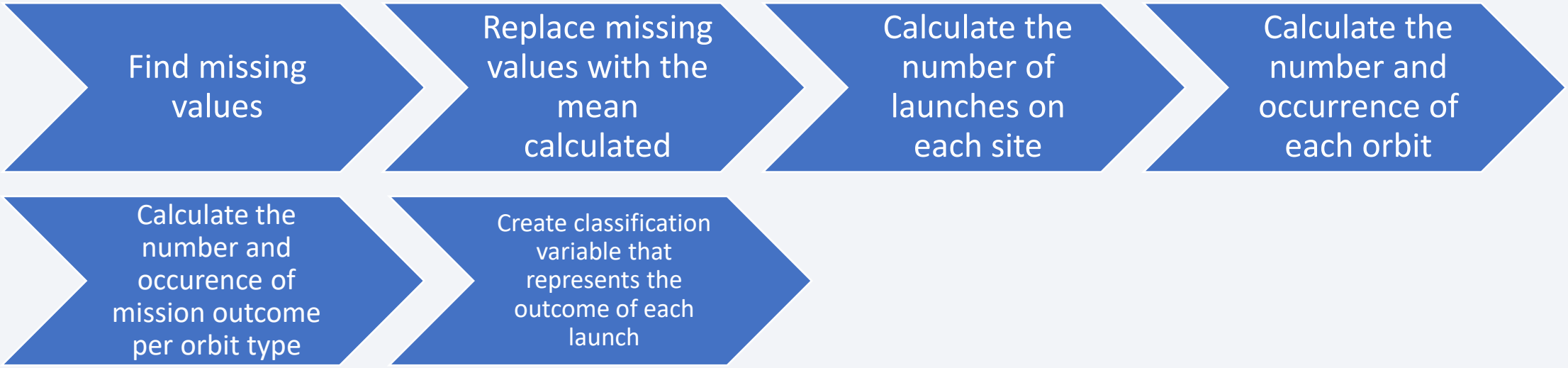
- Web scrap Falcon 9 launch records with BeautifulSoup. The process is shown in the following flowchart:



GitHub URL of the completed web scraping notebook:
https://github.com/bunny-tz/IBM_DS_Capstone/blob/b96d815f1fae2106d8b6897d6209b9eb403dce63/w1_jupyter-labs-webscraping.ipynb

Data Wrangling

- The data wrangling mainly includes two parts: dealing with missing values and calculating variables that are useful for data analysis. The process is shown in the following flowchart:



GitHub URL of the completed data wrangling notebook:
https://github.com/bunny-tz/IBM_DS_Capstone/blob/b96d815f1fae2106d8b6897d6209b9eb403dce63/w1_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter chart:
  - Investigate the relationships: FlightNumber vs. PayloadMass, FlightNumber vs LaunchSite, Payload vs. Launch Site, FlightNumber vs. Orbit type
- Bar chart
  - Investigate the success rate for each orbit type
- Line chart
  - Investigate the yearly trend of success rate
- GitHub URL of the completed EDA with data visualization notebook:
  - https://github.com/bunny-tz/IBM_DS_Capstone/blob/b96d815f1fae2106d8b6897d6209b9eb403dce63/w2_jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Summary of the SQL queries:
  - 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)
  - 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

EDA with SQL Continued

- Summary of the SQL queries:
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - 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 successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- GitHub URL of the completed EDA with SQL notebook:
  - https://github.com/bunny-tz/IBM_DS_Capstone/blob/b96d815f1fae2106d8b6897d6209b9eb403dce63/w2_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

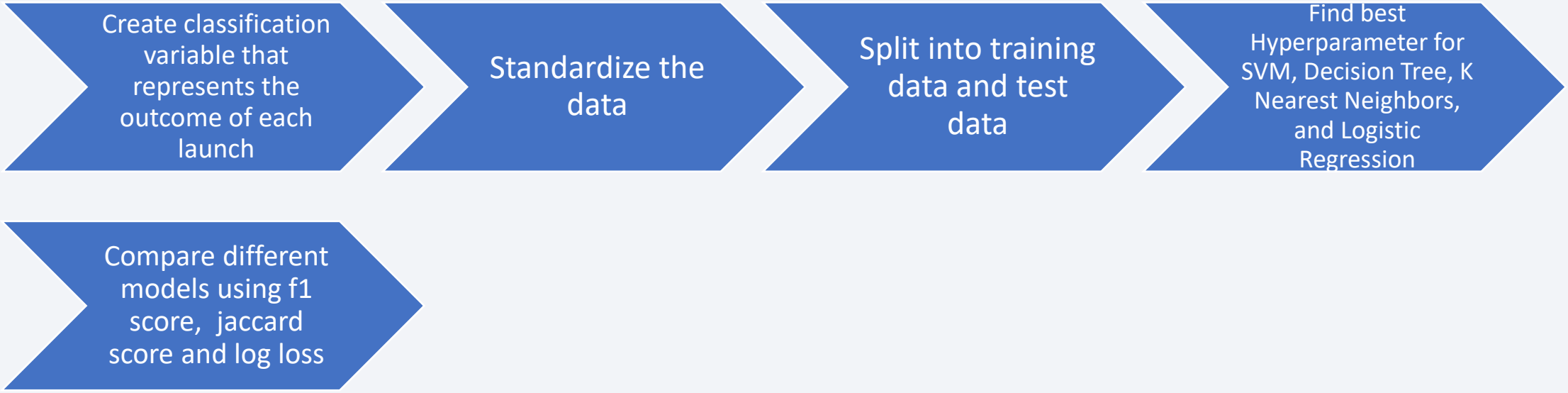
- Circles
  - Mark all launch sites on the map
- Markers
  - Mark the success/failed launches for each site on the map
- Lines
  - Explore and analyze the proximities of launch sites.
- GitHub URL of the completed interactive map with Folium map:
  - https://github.com/bunny-tz/IBM_DS_Capstone/blob/b96d815f1fae2106d8b6897d6209b9eb403dce63/w3_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie chart
  - Show the percentage of success launches for all launch sites or a particular site
- Scatter Chart
  - Display the relationship between Payload Mass, Booster Version and success of the launch
- GitHub URL of the completed Plotly Dash lab notebook:
  - https://github.com/bunny-tz/IBM_DS_Capstone/blob/ffa6143fc990c6dc7ad41feaf47287ed5f505dc3/w3_spacex_dash_app.py

Predictive Analysis (Classification)

- The following flowchart summarizes the process of using machine Learning to do predictive analysis for the first stage landing:



GitHub URL of the completed predictive analysis lab:

https://github.com/bunny-tz/IBM_DS_Capstone/blob/ffa6143fc990c6dc7ad41feaf47287ed5f505dc3/w4_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

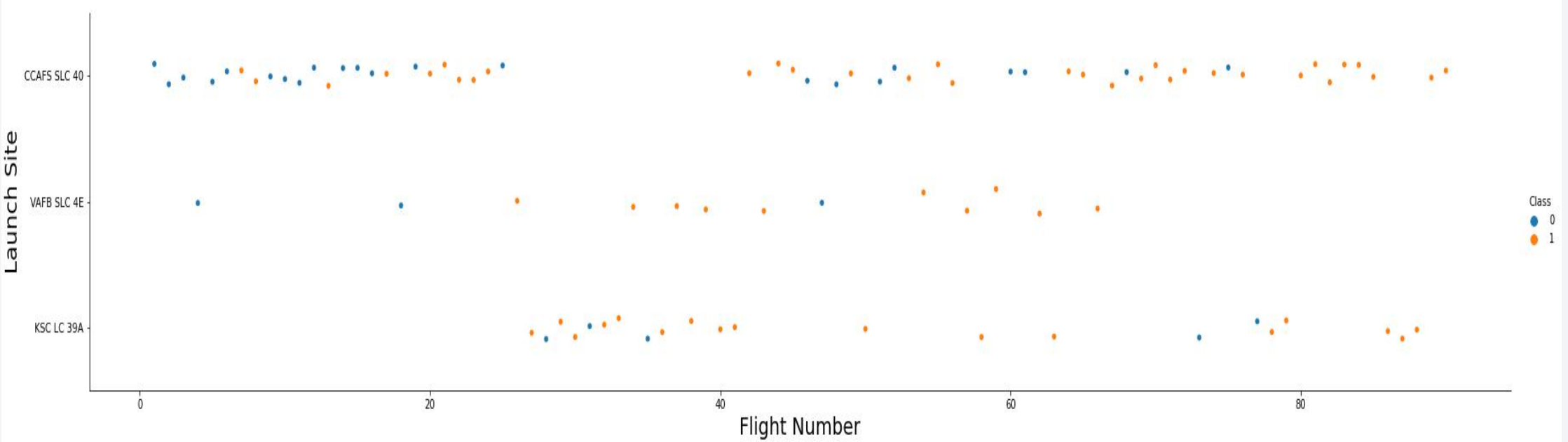
Results

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

Section 2

Insights drawn from EDA

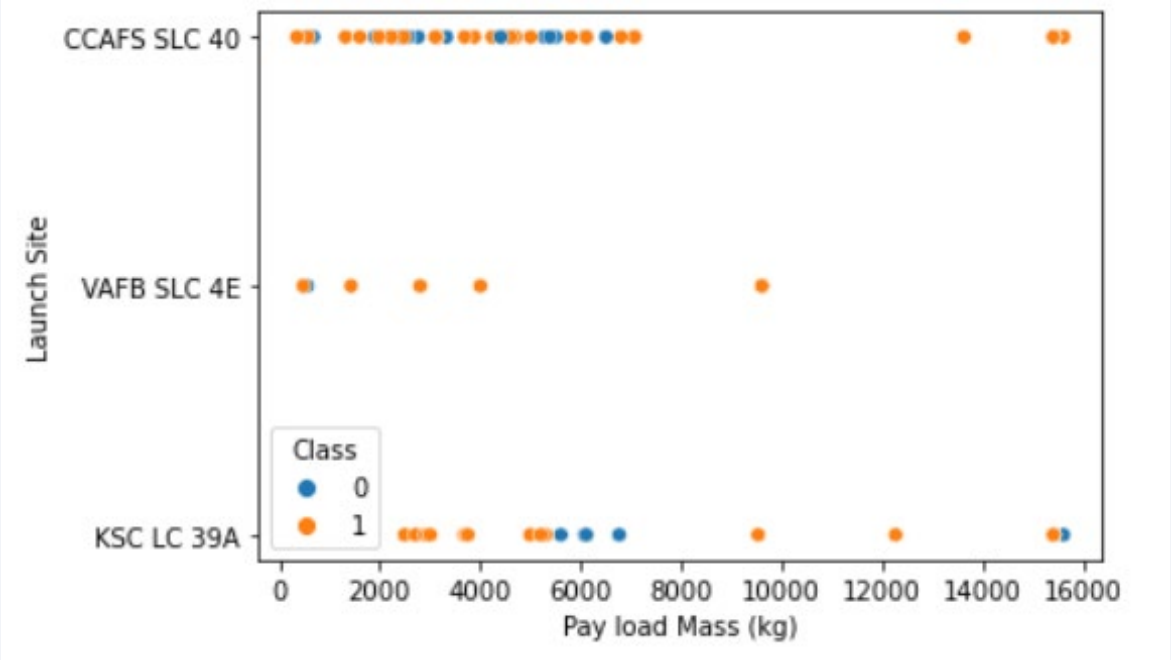
Flight Number vs. Launch Site



- At VAFB SLC 4E and KSC LC-39A , the success rate appears to be associated with number of flights.
- At CCAFS LC-40, there seems no relationship between flight number and success rate.

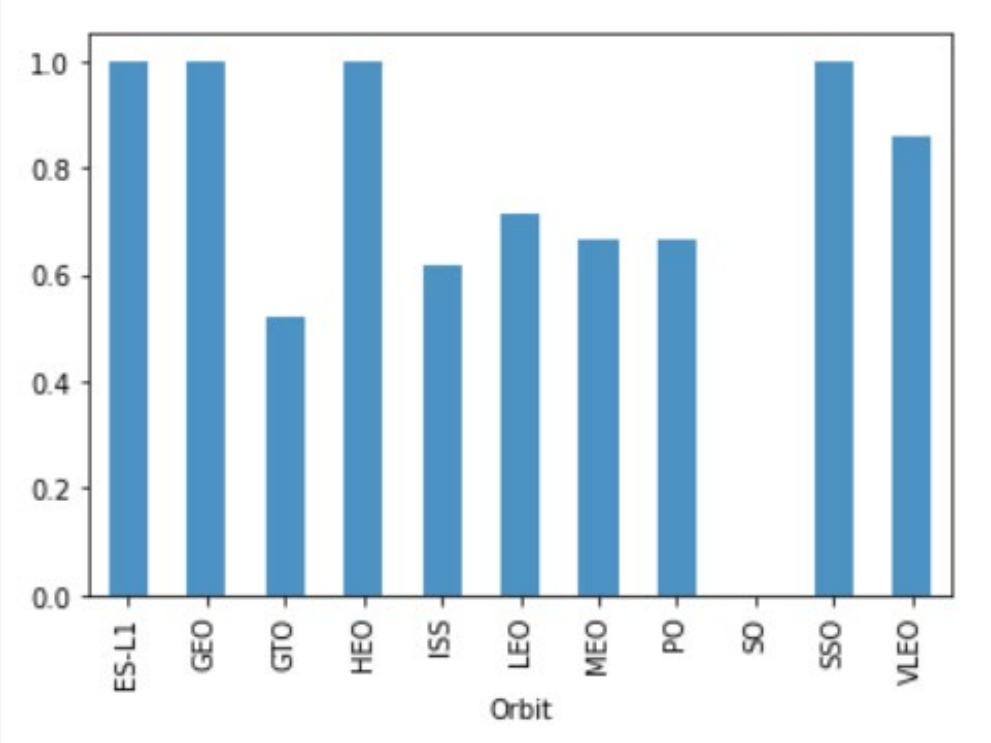
Payload vs. Launch Site

- VAFB SLC 4E has nearly 100% success rate when the pay load mass is relatively low.
- CCAFS LC-40 has low success rate when pay load mass is low but when the pay load mass is above 10,000 kg, the success rate is 100%.
- The overall success rate is high at KSC LC-39A.



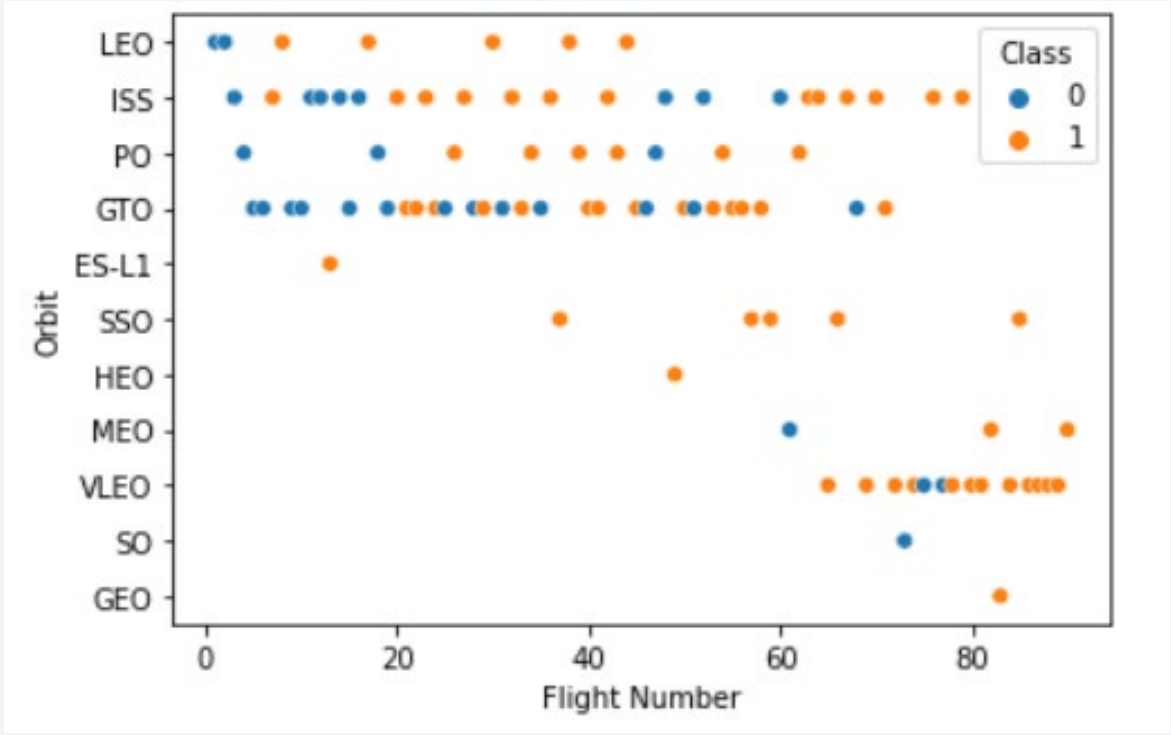
Success Rate vs. Orbit Type

- Orbit ES-L1, GEO, HEO and SSO have the highest success rate of 100%.
- The GTO has the lowest success rate of lower than 60%.
- Other orbits have success rates between 60% and 80%.



Flight Number vs. Orbit Type

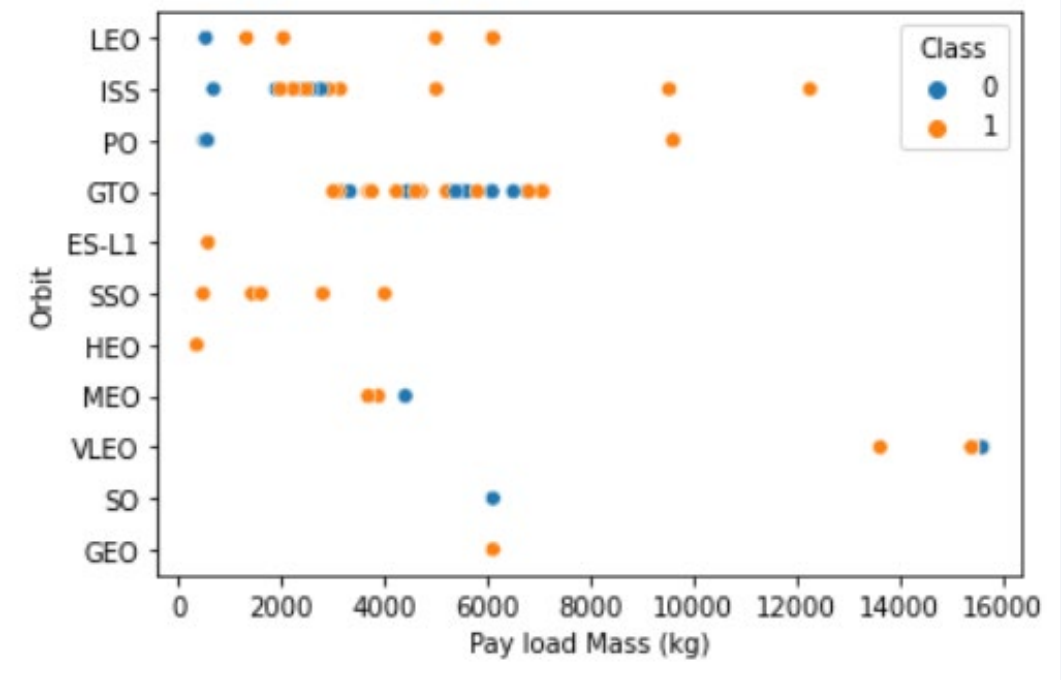
- In the LEO orbit the success rate increase with the number of flights.
- There seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

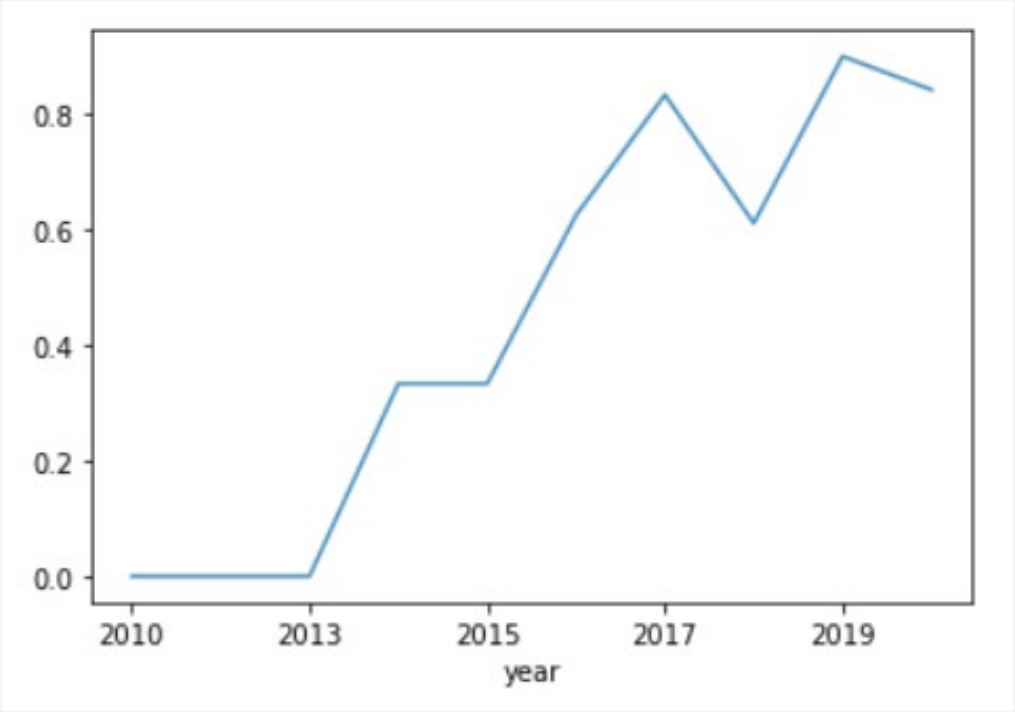
- With heavy payloads the success rate are higher for Polar, LEO and ISS.

- For GTO, we cannot distinguish this well as both successful landing and failure landing are both there.



Launch Success Yearly Trend

- The launch success rate increases from 2013 to 2020.



All Launch Site Names

- Find the names of the unique launch sites
- Four unique launch sites were found as shown in the screenshot.

Display the names of the unique launch sites in the space mission

```
%sql SELECT Distinct LAUNCH_SITE FROM SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- The 5 records are shown as follows:

Display 5 records where launch sites begin with the string 'CCA'

```
%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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- The total payload mass carried by boosters from NASA is 45596 kg.

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

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

TOTAL_PAYLOAD
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg.

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

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The earliest date of first successful landing outcome on ground pad is 01-05-2017.

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

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

MIN(DATE)
01-05-2017

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
- Four booster versions were found as shown below.

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

```
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000
```

```
 * 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

- Calculate the total number of successful and failure mission outcomes
- Total number of successful mission outcomes is 1460 and the total number of failure mission outcomes is 28.

List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, SUM(DATE) AS TOTAL_NUM FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

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

Mission_Outcome	SUM(DATE)
Failure (in flight)	28.0
Success	1429.0
Success	23.0
Success (payload status unclear)	8.0

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- 12 booster versions were found as shown below.

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- There are two failure records, which are in January and April.

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
SELECT SUBSTR(DATE, 4, 2) as MONTH, BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure
```

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

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

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
- Rank of successful landing outcomes is shown below.

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) FROM SPACEXTBL WHERE (DATE BETWEEN '04-06-2010' AND '20-03-2017') AND LANDING_OUTC
```

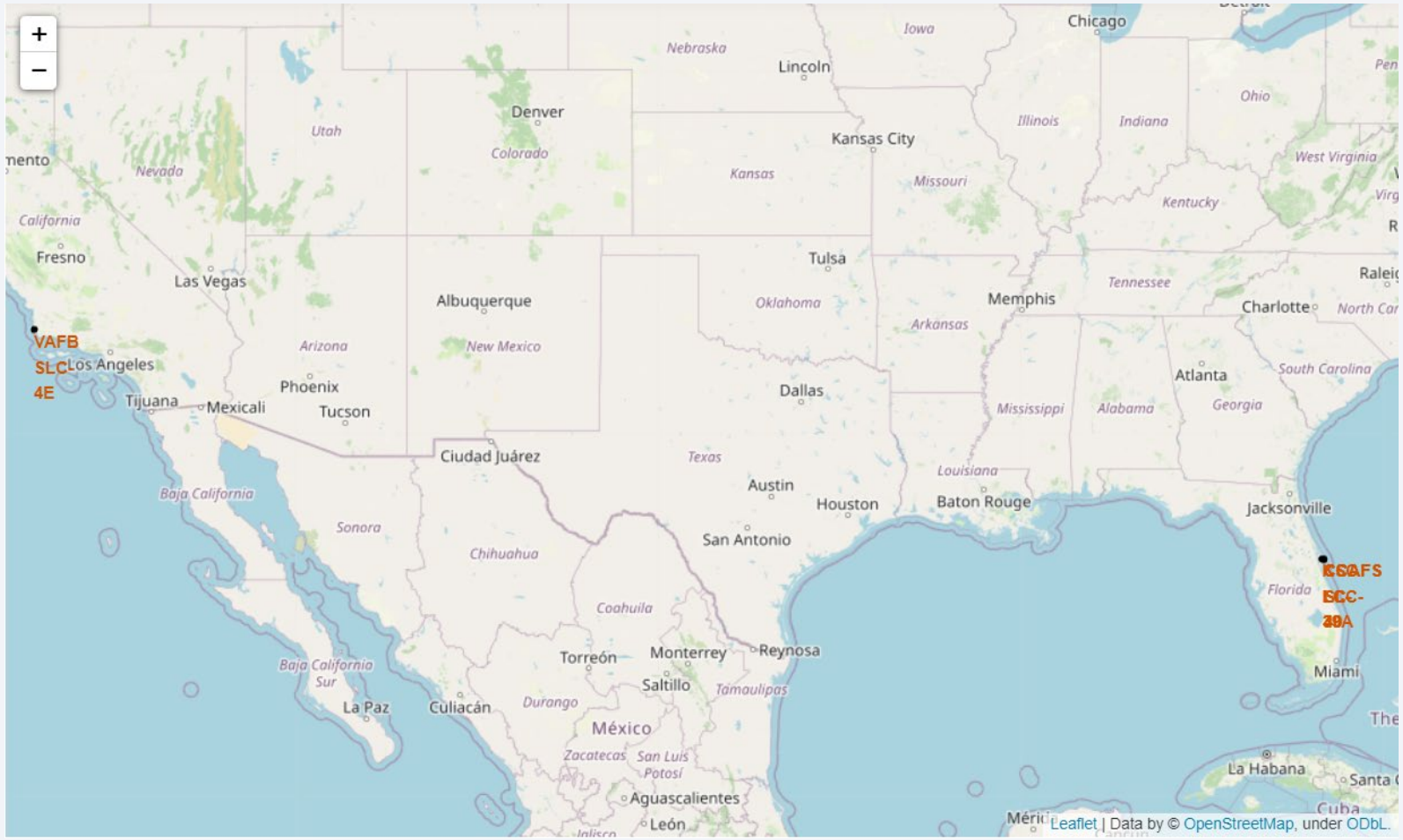
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	COUNT(LANDING_OUTCOME)
Success	20
Success (drone ship)	8
Success (ground pad)	6

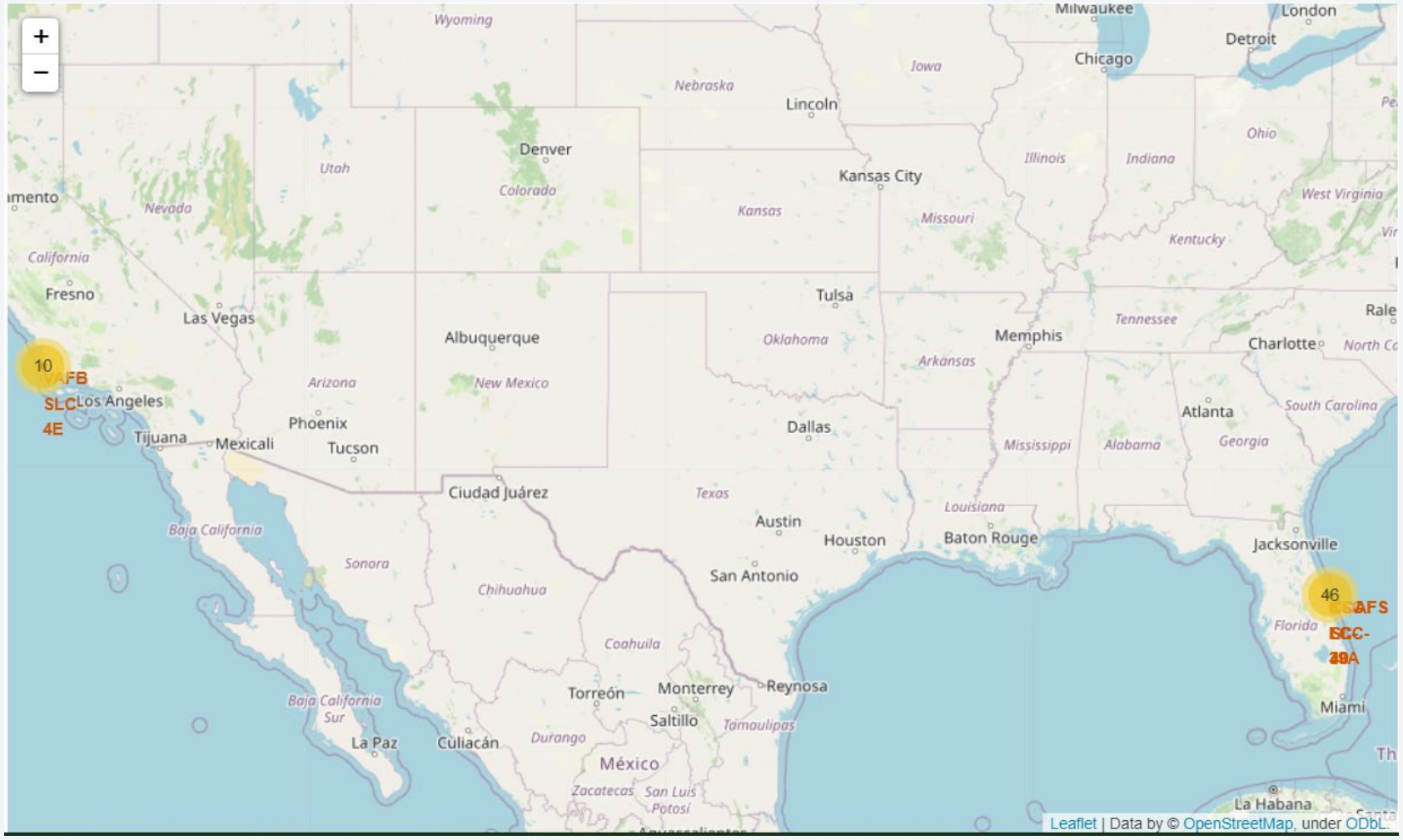

Section 3
Launch Sites
Proximities Analysis

Location of Launch Sites

- All launch sites are in proximity to the Equator line.
- All launch sites are in very close proximity to the coast.

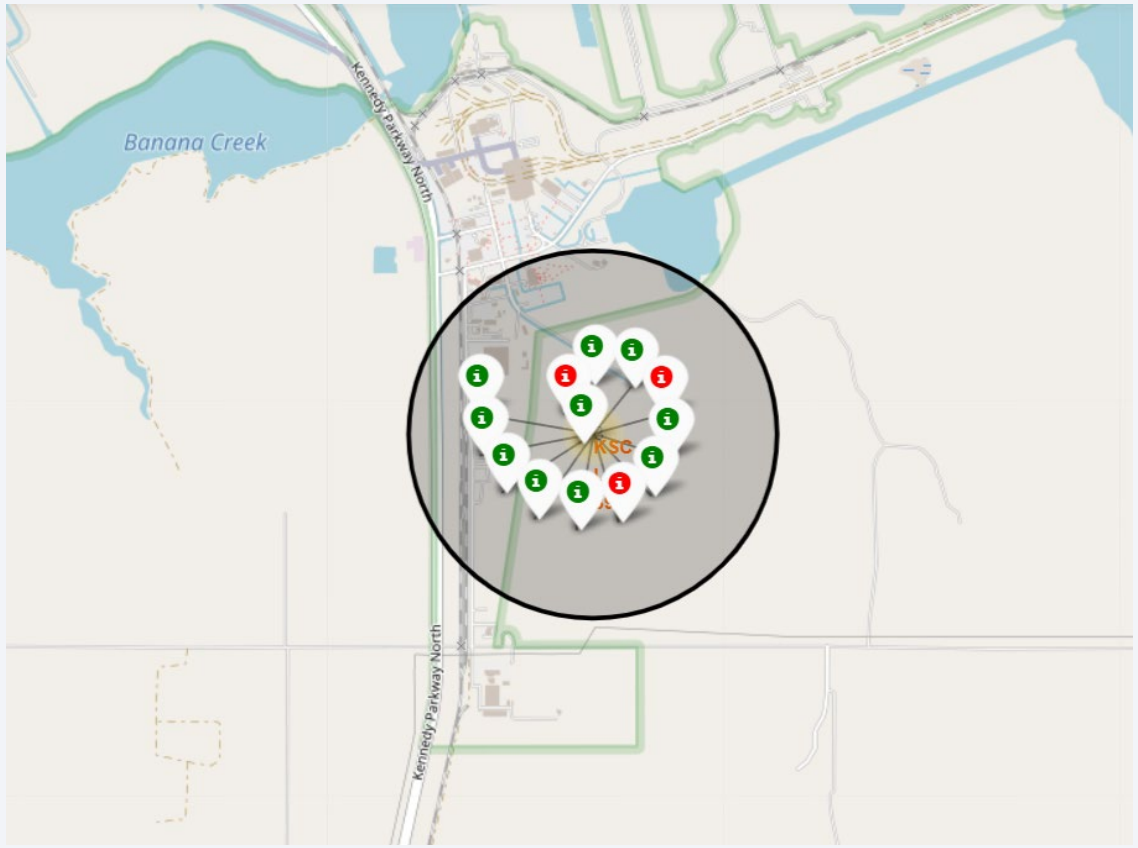


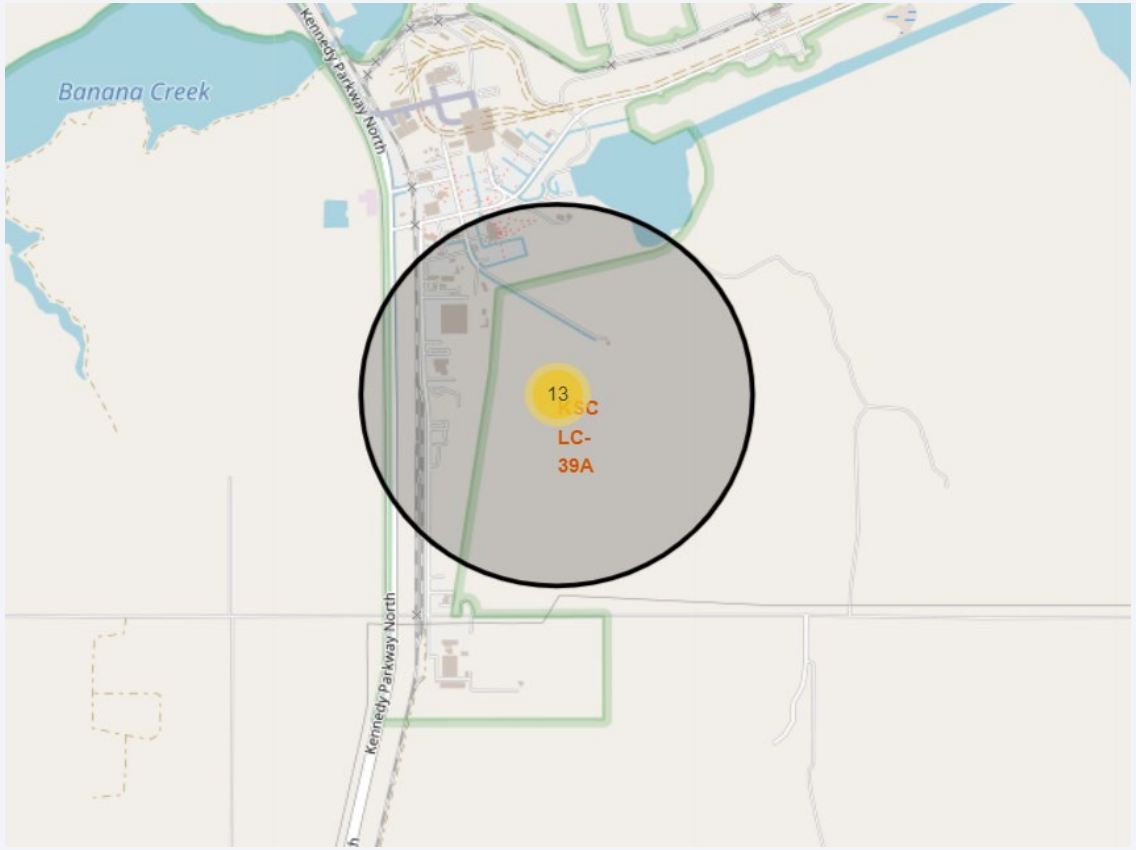
Success/Failed Launches for Each Site



Success/Failed Launches for Each Site

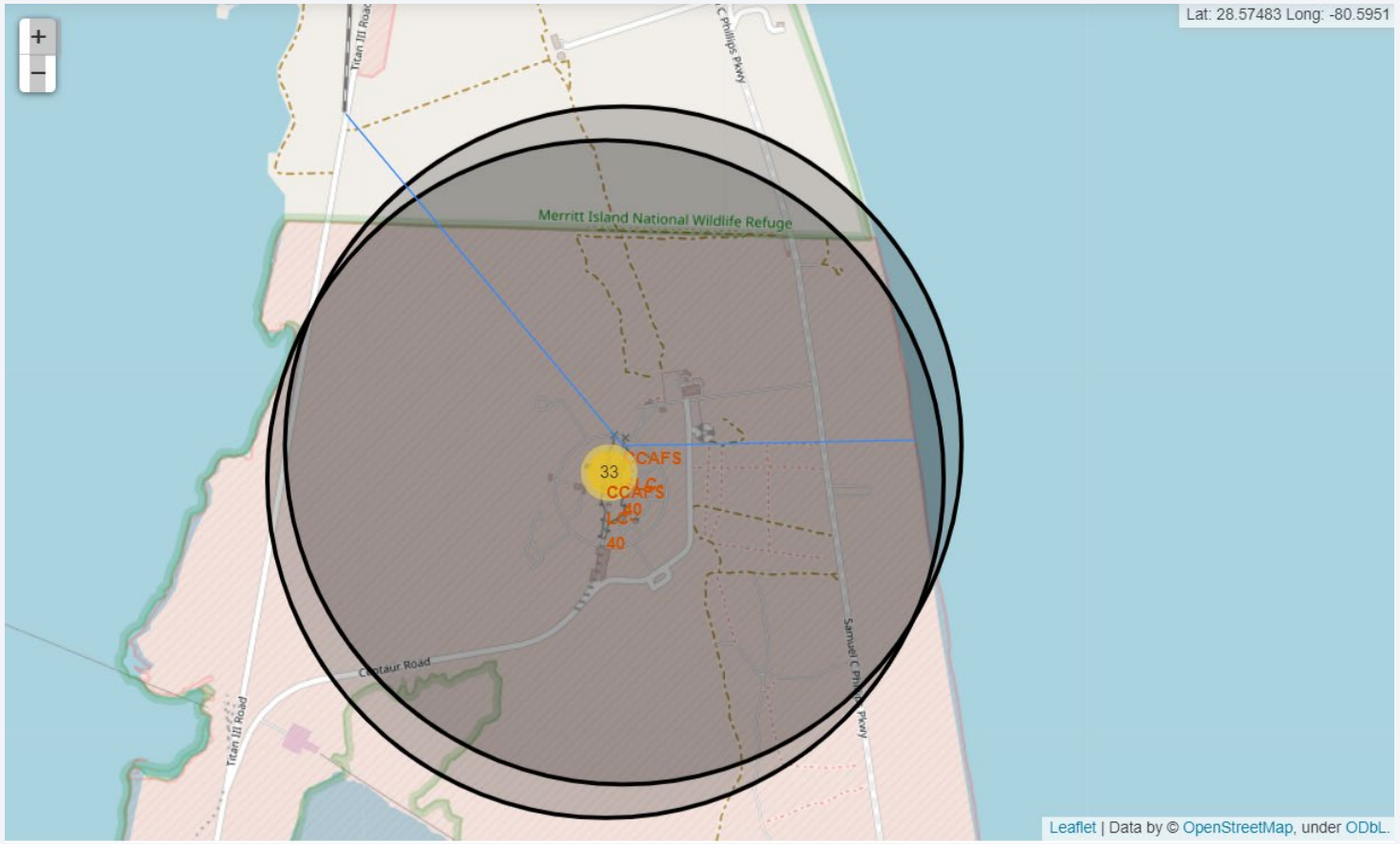
- From the color-labeled markers in marker clusters, it is easy to identify that KSC LC-39A has the highest success rate which is about 77%.

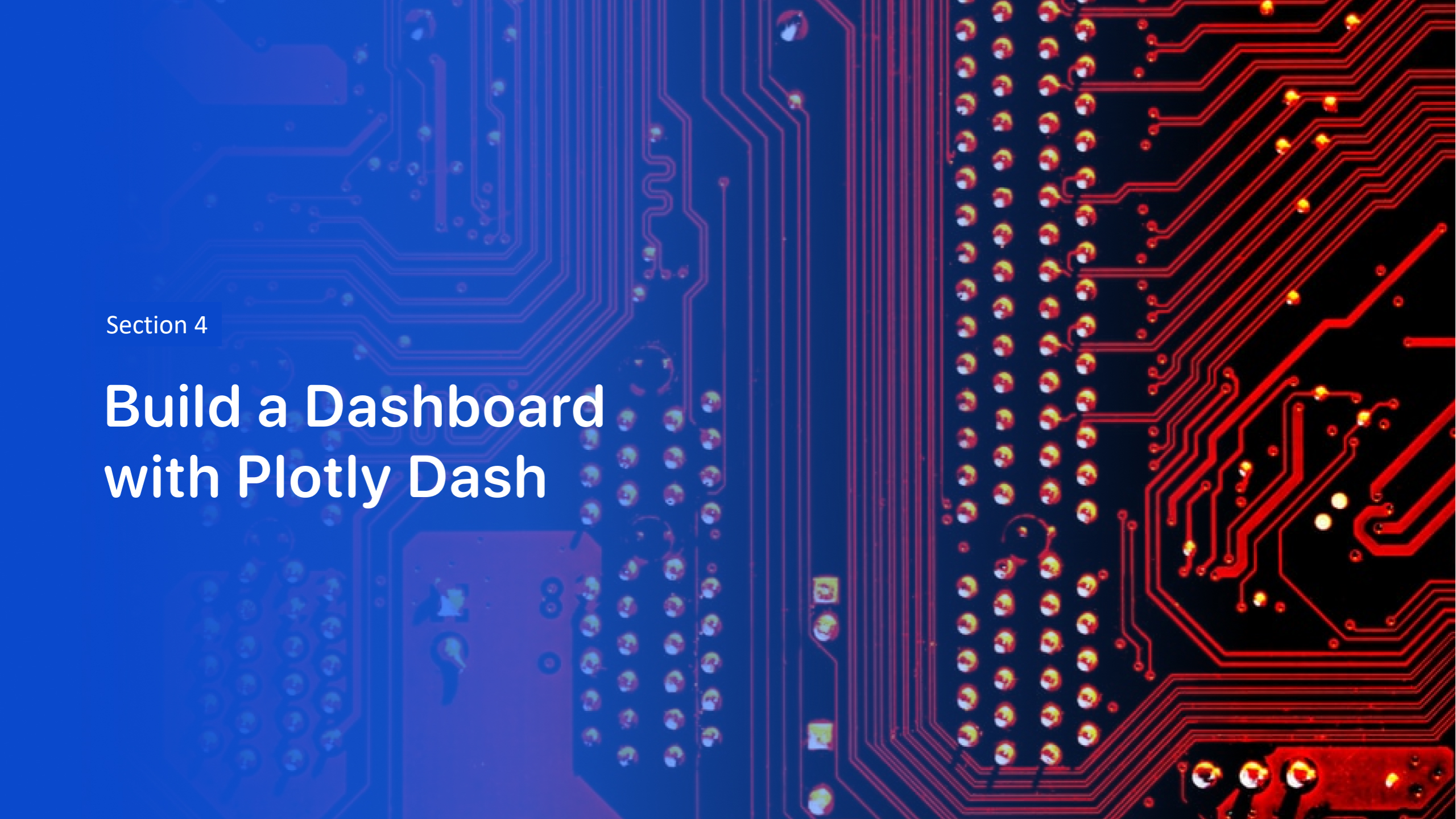




Distances between a Launch Site to its Proximities

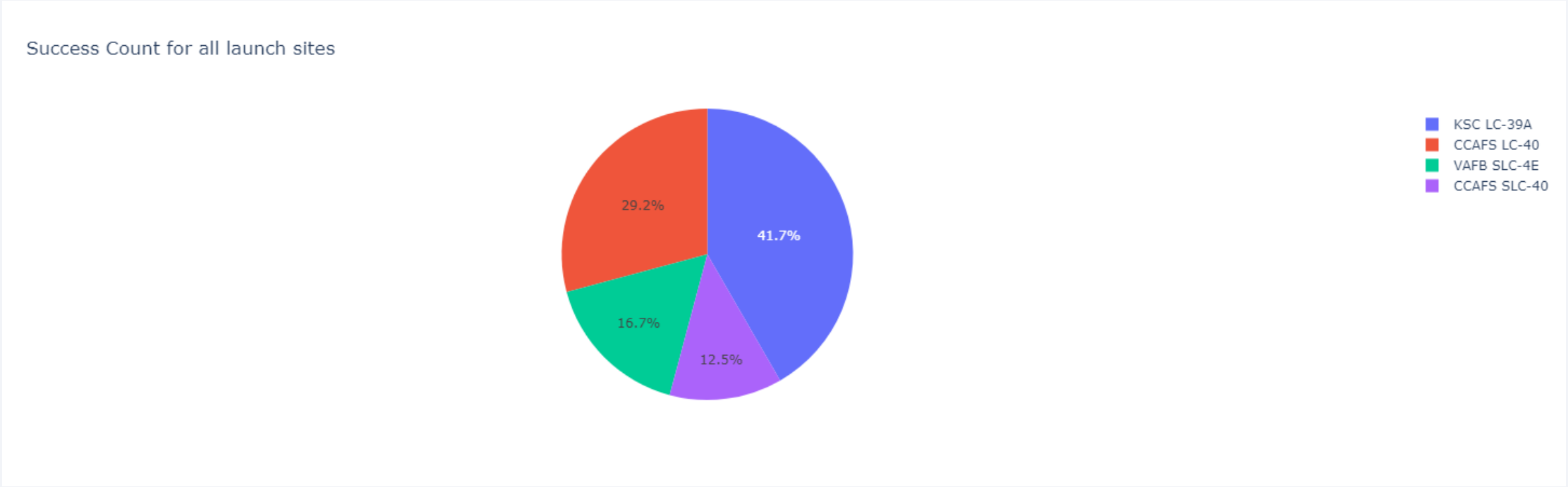
- CCAFS SLC-40 is in close proximity to the coast and railway.




Section 4
Build a Dashboard with Plotly Dash

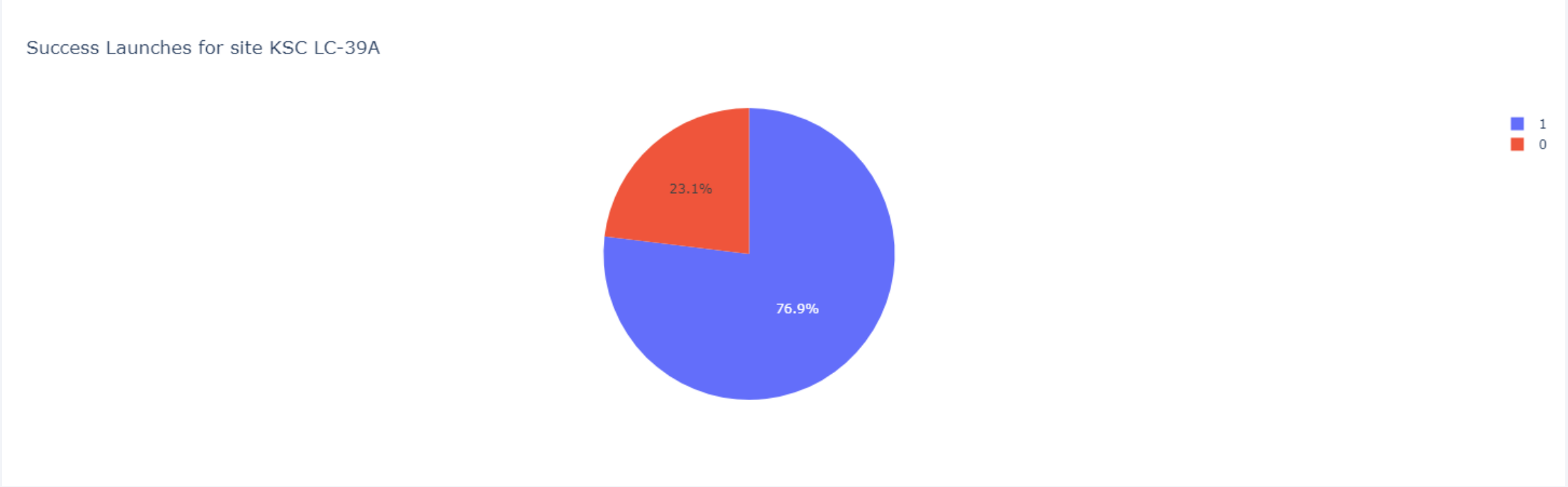
Pie chart of Success Rate for all Launch Sites

- 42% of all the successful launches are at KSC LC-39A.



Pie chart for the Launch Site with Highest Success Rate

- KSC LC-39A has the highest success rate which is around 77%.

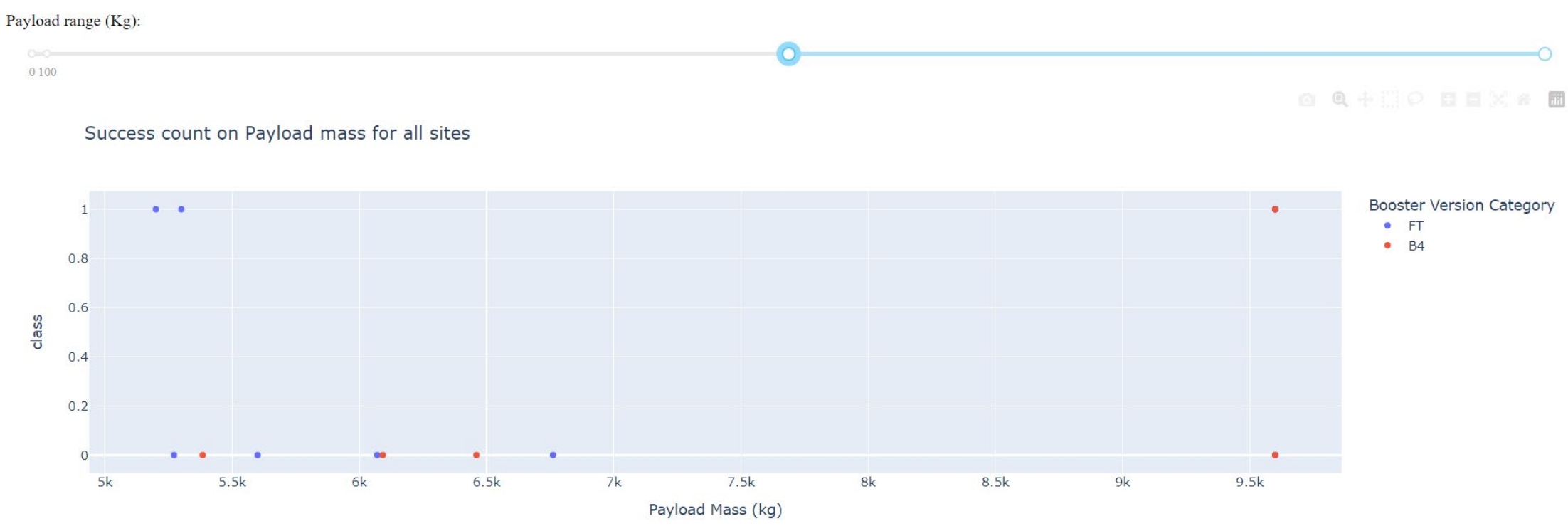


Scatter Plot of Payload vs. Launch Outcome for all Sites

- When Palyload Mass is below 5000 kg, booster version of FT has the highest success rate whereas booster version of v1.1 has the lowest success rate.

Scatter Plot of Payload vs. Launch Outcome for all Sites

- When Palyload Mass is above 5000 kg, only booster version of FT and B4 are applied, and both have relatively low success rate. Success rate of FT is about 33% and success rate of B4 is only 20%.

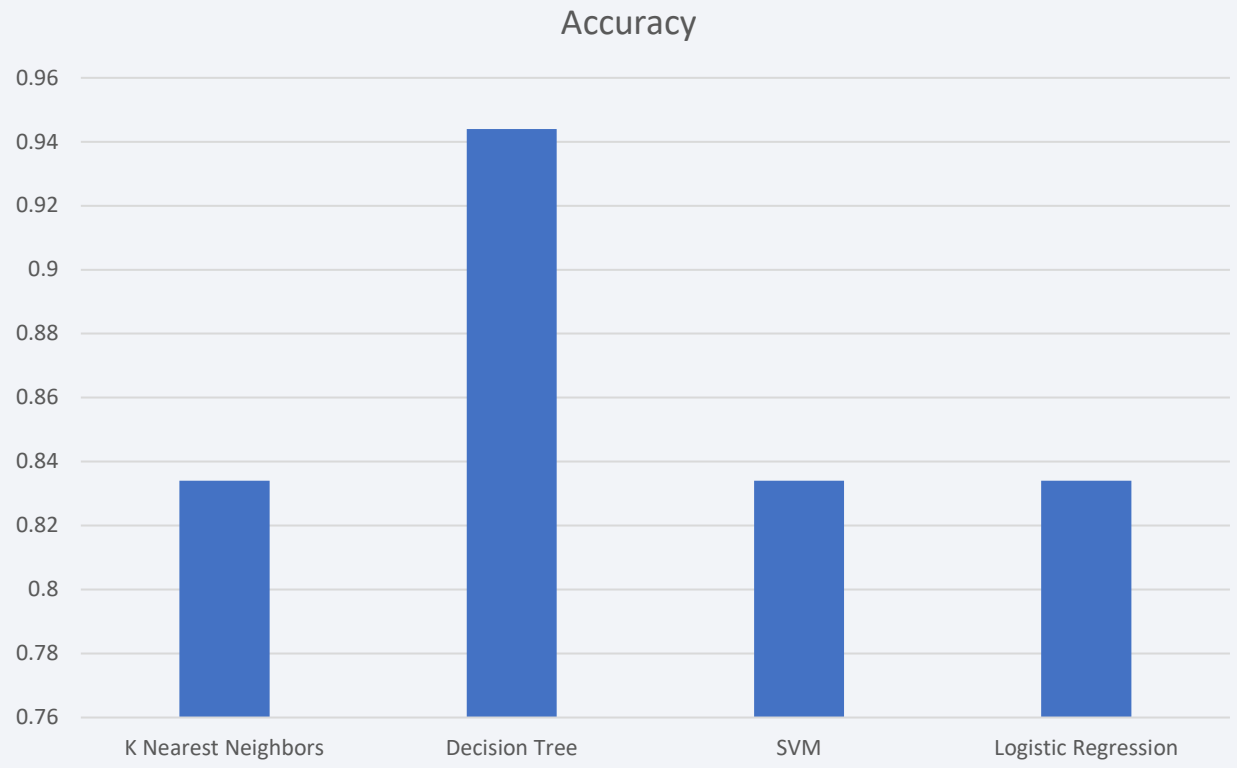


Section 5

Predictive Analysis (Classification)

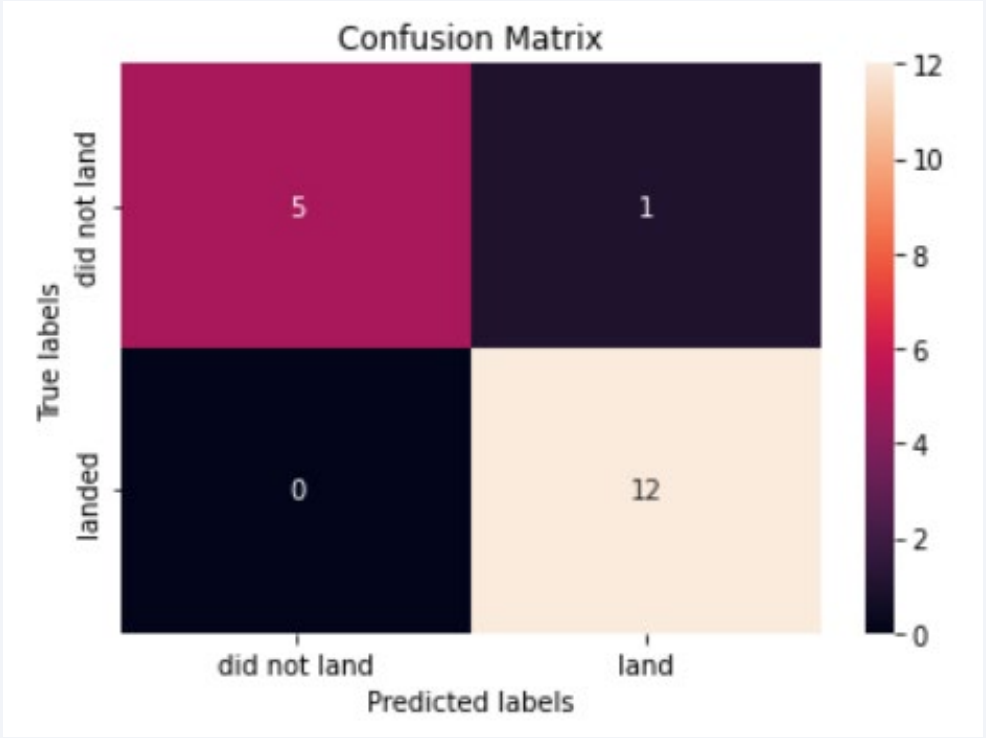
Classification Accuracy

- The decision tree has the highest accuracy which is about 94.4%.
- Best parameters:
  - criterion: entropy,
  - max_depth: 4,
  - max_features: sqrt,
  - min_samples_leaf: 2,
  - min_samples_split: 10,
  - splitter: best



Confusion Matrix

- Confusion matrix for the Decision Tree:
  - For all the 13 successful landings, the decision tree correctly classified 12 times as successful landings.
  - All the 5 unsuccessful landings are correctly classified by the decision tree.



Conclusions

- Location of launch sites are proximity to Equator line, coast and railway.
- Whether the landing is successful or not are associated with launch site, payload mass, orbit type, boost version, and time.
- Launch site KSC LC-39A has the highest success rate which is around 77%.
- VAFB SLC 4E has nearly 100% success rate when the pay load mass is relatively low. CCAFS LC-40 has low success rate when pay load mass is low but when the pay load mass is above 10,000 kg, the success rate is 100%.
- Orbit ES-L1, GEO, HEO and SSO have the highest success rate of 100%.
- When Palyload Mass is below 5000 kg, booster version of FT has the highest success rate whereas booster version of v1.1 has the lowest success rate. When Palyload Mass is above 5000 kg, only booster version of FT and B4 are applied, and both have relatively low success rate.
- The launch success rate increases from 2013 to 2020.
- The decision tree has the highest classification accuracy which is about 94.4%. Thus, it should be used as the classification method for landing prediction.

Appendix

- Table of classification model accuracy comparison using Jaccard Score, F1 Score and LogLoss

Model Type	Jaccard	F1	LogLoss
K Nearest Neighbors	0.80	0.88	NA
Decision Tree	0.92	0.96	NA
SVM	0.80	0.88	NA
Logistic Regression	0.80	0.88	0.48

Thank you!