IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mahsa Alavi
Sep 2023

Executive Summary

- We was working with SpaceX launch data that is gathered from an API, specifically the SpaceX REST API. Another popular data source for obtaining Falcon 9 Launch data is web scraping related Wiki pages. We also converted the data into a data frame and then performed some data wrangling. We performed some Exploratory Data Analysis (EDA) using a database. We was using Folium and Plotly Dash to build an interactive map and dashboard to perform interactive visual analytics. Finally, we used machine learning to determine if the first stage of Falcon 9 will land successfully.
- Launch success has improved over time (2013-2020). KSC LC-39A has the highest success rate among landing sites. ES-L1, GEO, HEO, and SSO orbits achieved the highest success rate. Most launch sites are near the equator, and all are close to the coast. Logistic Regression (LR), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) models are the best in terms of prediction accuracy for this dataset.

Table of Contents

Introduction

- 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.

Section 1

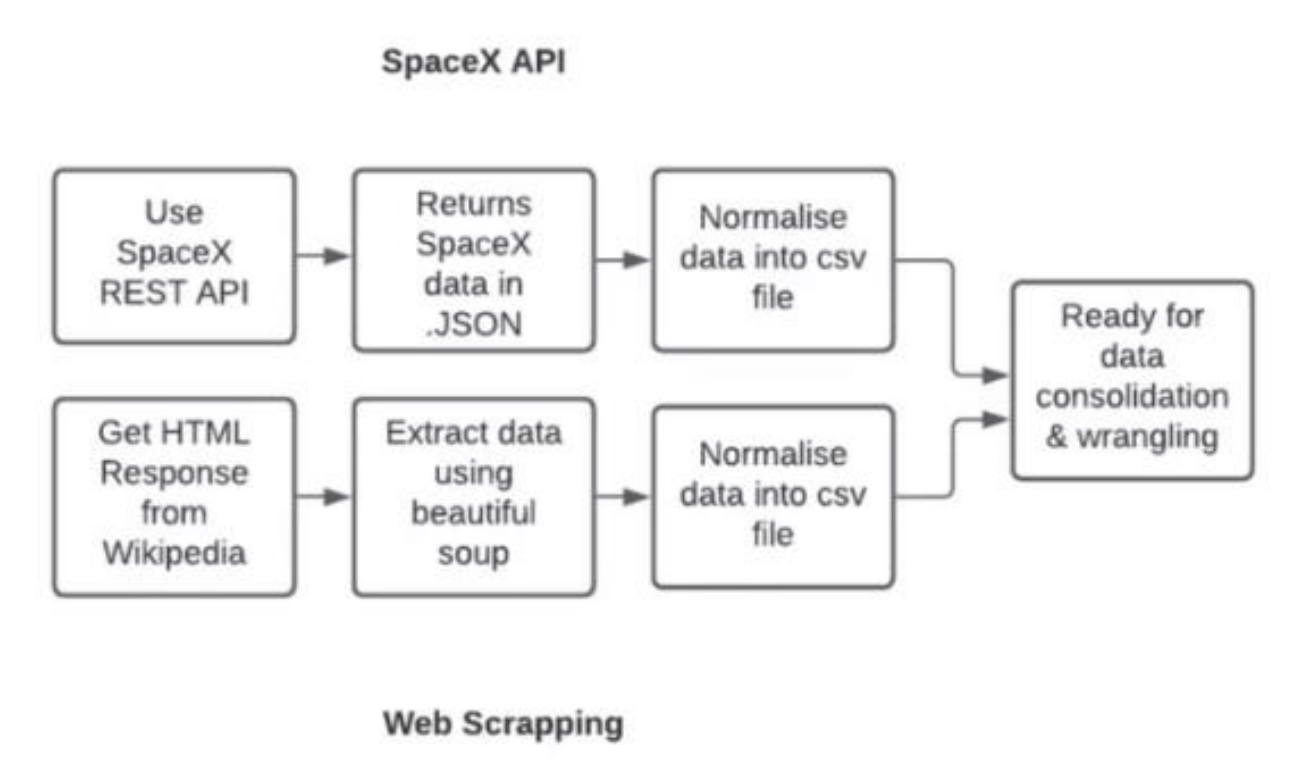
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Using an API, Sampling Data, and Dealing with Nulls
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression/Support Vector Machines/Decision Tree/K-Nearest Neighbors

Data Collection



Data Collection – SpaceX API

- Request rocket launch data from SpaceX API
- Decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()
- Use the API again to get information about the launches using custom functions
- Combine the columns into a dictionary
- Create a Pandas data frame from the dictionary
- Filter the data-frame to only include Falcon 9 launches

https://github.com/mahsa-ak91/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Request the Falcon9 Launch Wiki page from its URL
- Create BeautifulSoup object from the HTML response
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables

https://github.com/mahsa-ak91/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Data Wrangling is the process of cleaning and unifying messy and complex data sets for easy access and Exploratory Data Analysis (EDA).

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome column

https://github.com/mahsa-ak91/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Visualize the relationship between Flight Number and Launch Site
- Visualize the relationship between Payload and Launch Site
- Visualize the relationship between success rate of each orbit type
- Visualize the relationship between Flight Number and Orbit type
- Visualize the relationship between Payload and Orbit type
- Visualize the launch success yearly trend

View relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists.
Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value.

https://github.com/mahsa-ak91/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- 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
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- 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 landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

https://github.com/mahsa-ak91/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Mark all launch sites on a map**
- **Mark the success/failed launches for each site on the map**

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

https://github.com/mahsa-ak91/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Add a dropdown list to enable Launch Site selection

Allow user to select all launch sites or a certain launch site

- Add a pie chart to show the total successful launches count for all sites

Allow user to see successful and unsuccessful launches as a percent of the total

- Add a slider to select payload range

Allow user to select payload mass range

- Add a scatter chart to show the correlation between payload and launch success

Allow user to see the correlation between Payload and Launch Success

Predictive Analysis (Classification)

Four classification models including Logistic Regression, Support Vector Machines, Decision Tree, and K-Nearest Neighbors, were built and compared.

- Create a NumPy array from the column Class in data
- Standardize the data
- Split the data X and Y into training and test data
- Then the models are trained and hyperparameters are selected
- Fit the object to find the best parameters from the dictionary parameters
- Calculate the accuracy on the test data (We can plot the confusion matrix)

https://github.com/mahsa-ak91/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

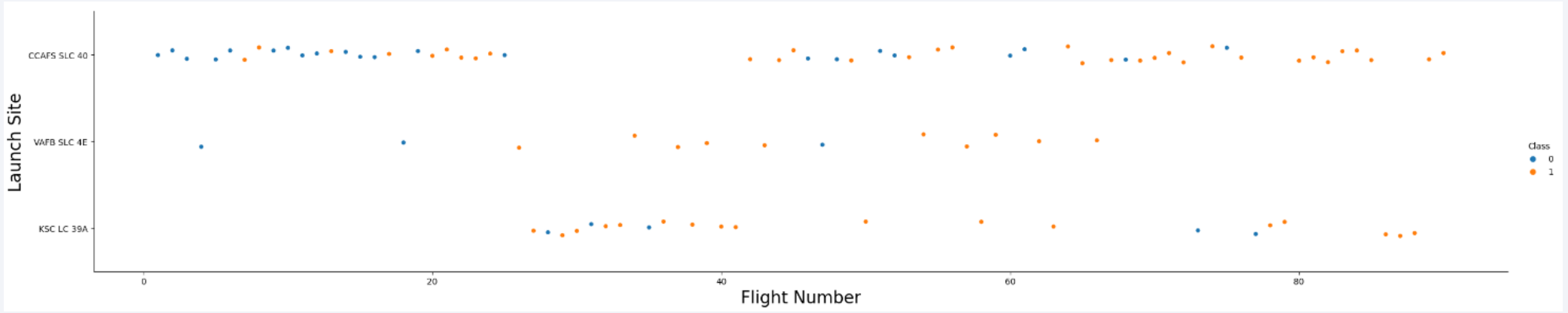
Results

- Exploratory Data Analysis
- Interactive Visual Analytics with Folium
- Interactive Dashboard with Plotly Dash
- Predictive Analysis

Section 2

Insights drawn from EDA

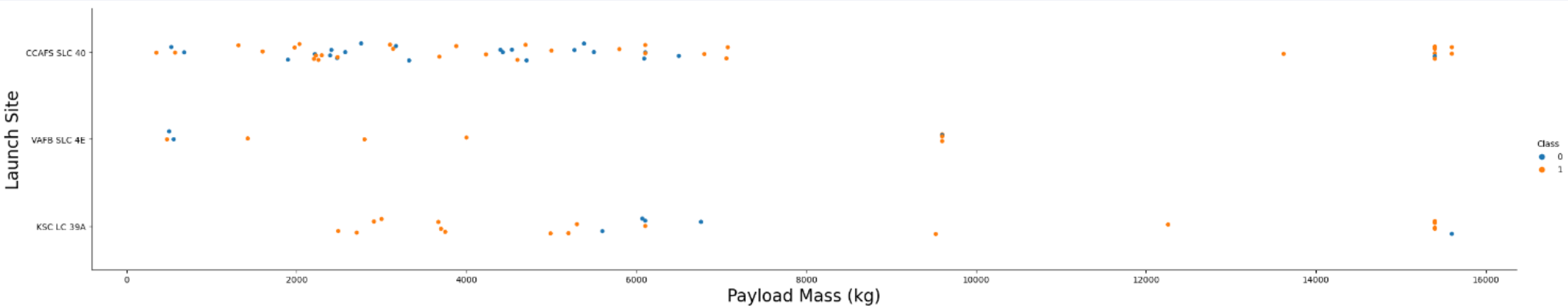
Flight Number vs. Launch Site



Class 0 = Fail
Class 1 = Success

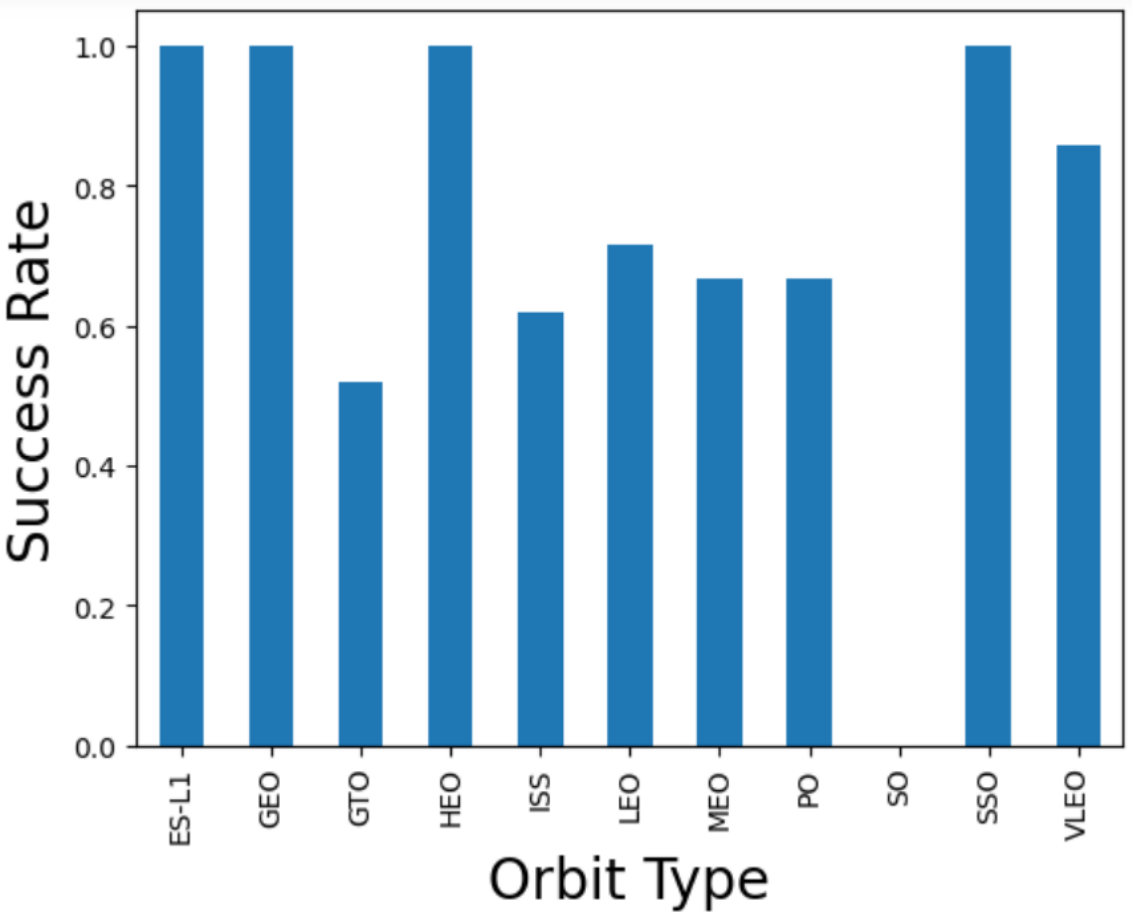
- Launches from the site of CCAFS SLC 40 are significantly higher than launches form other sites.
- There are few launches from VAFB SLC 4E but most have landed successfully.
- Most launches from KSC LC 39A have landed successfully.
- We can infer that new launches have a higher success rate.

Payload vs. Launch Site



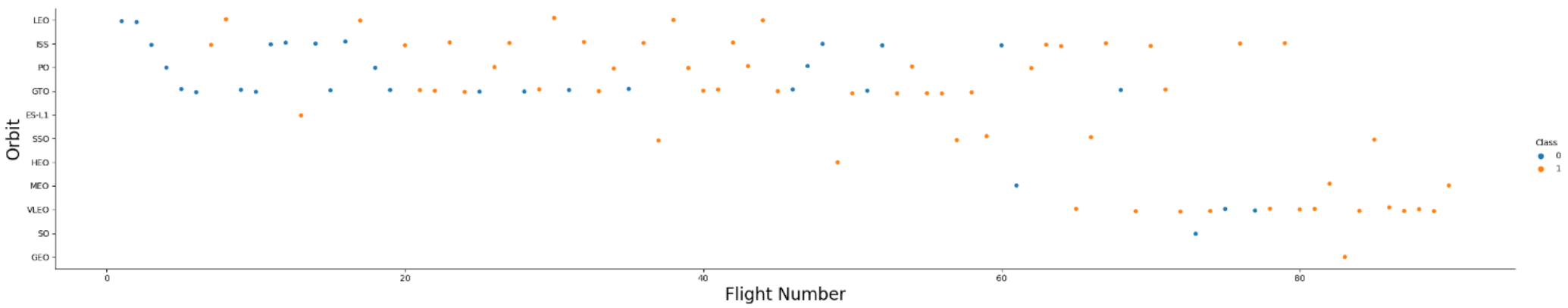
- Most launches with a payload greater than 7,000 kg were successful.
- The majority of payloads with low mass have been launched from CCAFS SLC 40 launch site.
- VAFB SLC 4E has not launched anything greater than 10,000 kg.
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg.

Success Rate vs. Orbit Type



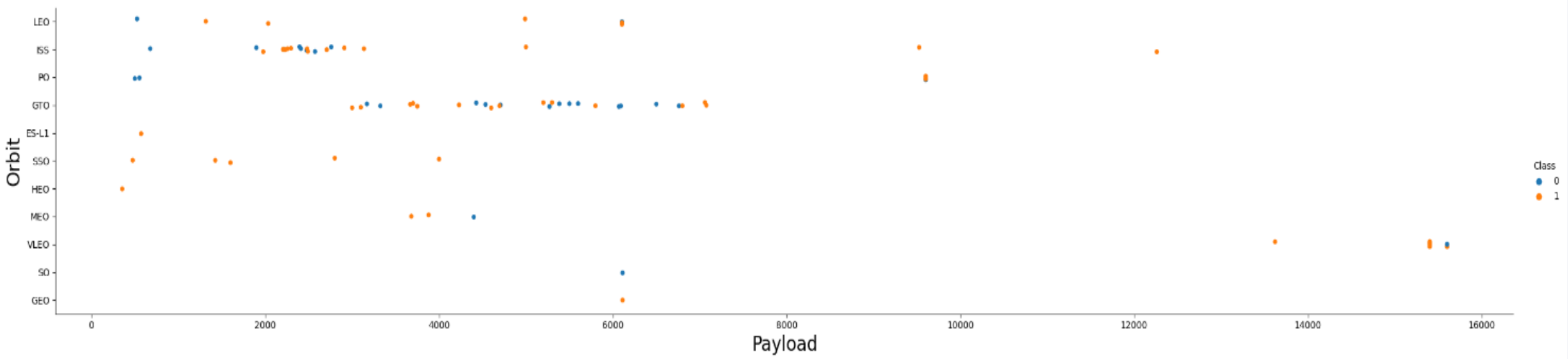
- The orbit types of ES-L1, GEO, HEO, and SSO are the highest success rate.
- Nevertheless, upon closer examination, it becomes evident that several of these orbits are represented by only a single occurrence in the dataset. This scarcity of data points necessitates a more extensive dataset to discern meaningful patterns or trends before definitive conclusions can be drawn.

Flight Number vs. Orbit Type



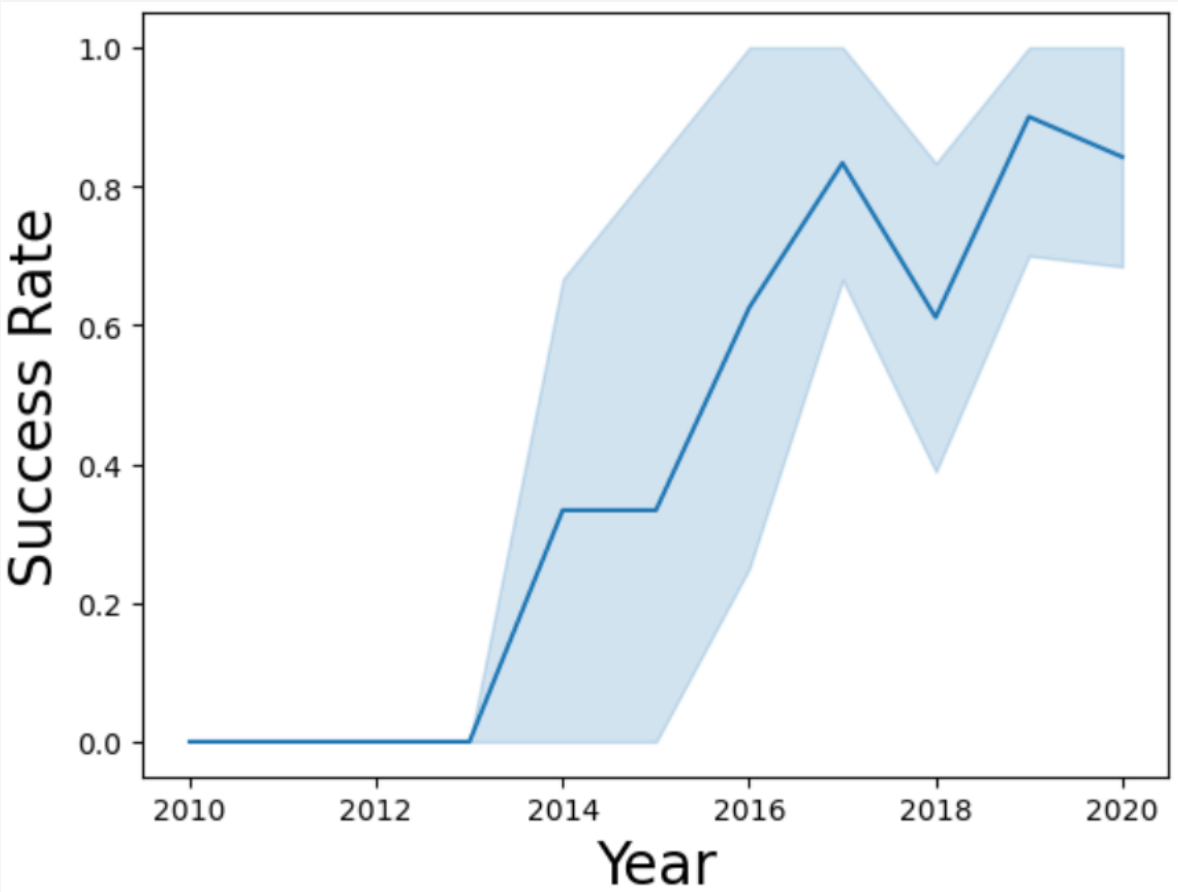
- The success rate typically increases with the number of flights for each orbit, This relationship is highly apparent for the LEO orbit.
- There seems to be no relationship between flight number when in GTO orbit.
- It's important to note that orbits with only one occurrence should be excluded from the aforementioned observation, as these instances require a more substantial dataset to establish any meaningful conclusions.

Payload vs. Orbit Type



- With heavy payloads the successful landing rate are more for PO, LEO and ISS.
- For MEO orbit, heavier payload seems to have a negative impact on the success rate.
- In the case of GTO, there seems no apparent relationship between payload and the rate of success.
- Orbits like SO, GEO, and HEO require more data to identify any potential patterns or trends.

Launch Success Yearly Trend



We can observe that the success rate since 2013 kept increasing till 2020.

All Launch Site Names

The names of the unique launch sites in the space mission

```
]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
 * 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'

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

```
]: %%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

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

]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS)

```
]: %%sql
   SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass
   FROM SPACEXTBL
   WHERE Customer = 'NASA (CRS)';
```

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

]:

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

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

```
]: %%sql
   SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass
   FROM SPACEXTBL
   WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

]:

Average_Payload_Mass
2534.6666666666665

First Successful Ground Landing Date

The date when the first successful landing outcome in ground pad was achieved.

```
]: %%sql
SELECT MIN(Date) AS First_Successful_Landing_In_Ground_Pad
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)';
```

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

]:

First_Successful_Landing_In_Ground_Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
]: %%sql
   SELECT BOOSTER_VERSION
   FROM SPACEXTBL
   WHERE Landing_Outcome = 'Success (drone ship)'
       AND 4000 < PAYLOAD_MASS__KG_ < 6000;

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

]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

```
]: %%sql
SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Total_Number
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

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

]:

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

Boosters Carried Maximum Payload

The names of the Booster_Version which have carried the maximum payload mass.

```
%%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

The records which will display the month names, Failure_Landing_Outcome in drone ship , their Booster_Version, and Launch_Site names for the months in year 2015.

```
]: %%sql SELECT
 CASE
        WHEN Date like '%-01-%' THEN 'January'
        WHEN Date like '%-02-%' THEN 'February'
        WHEN Date like '%-03-%' THEN 'March'
        WHEN Date like '%-04-%' THEN 'April'
        WHEN Date like '%-05-%' THEN 'May'
        WHEN Date like '%-06-%' THEN 'June'
        WHEN Date like '%-07-%' THEN 'July'
        WHEN Date like '%-08-%' THEN 'August'
        WHEN Date like '%-09-%' THEN 'September'
        WHEN Date like '%-10-%' THEN 'October'
        WHEN Date like '%-11-%' THEN 'November'
        WHEN Date like '%-12-%' THEN 'December'
    END AS Month,
    Landing_Outcome AS Failure_Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE Date like'%2015%' and Landing_Outcome LIKE 'Failure (drone ship)';
```

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

]:

Month	Failure_Landing_Outcome	Booster_Version	Launch_Site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
]: %%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Count_of_Landing_Outcomes
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Count_of_Landing_Outcomes DESC
```

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

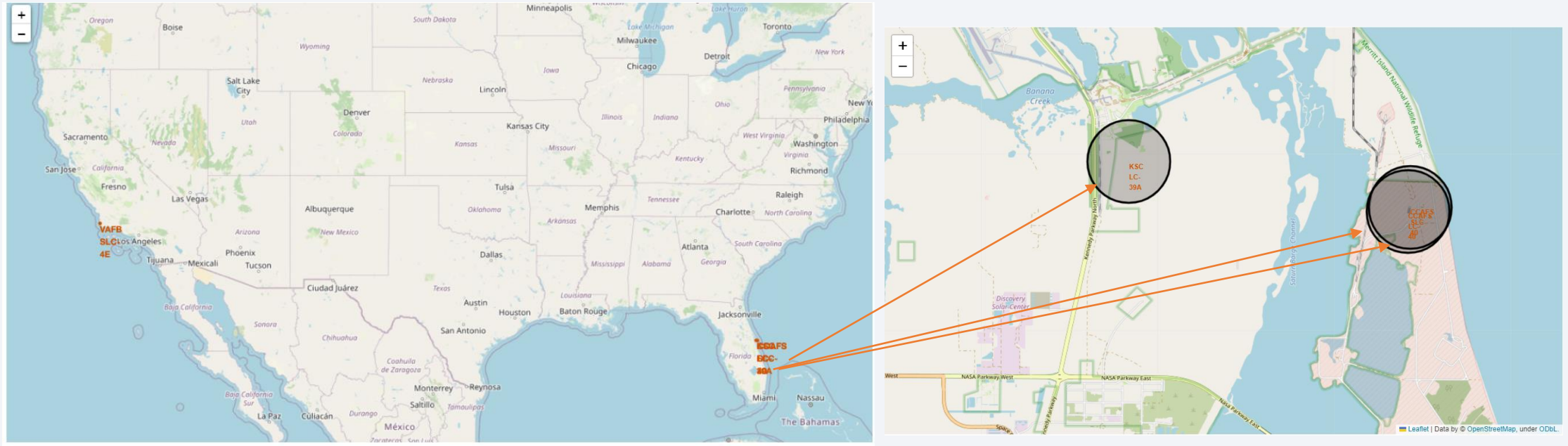
]:

Landing_Outcome	Count_of_Landing_Outcomes
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

As shown, the highest count among the landing outcomes is attributed to drone ship, totaling 11. The launches that have landed on a ground pad have all been successful.

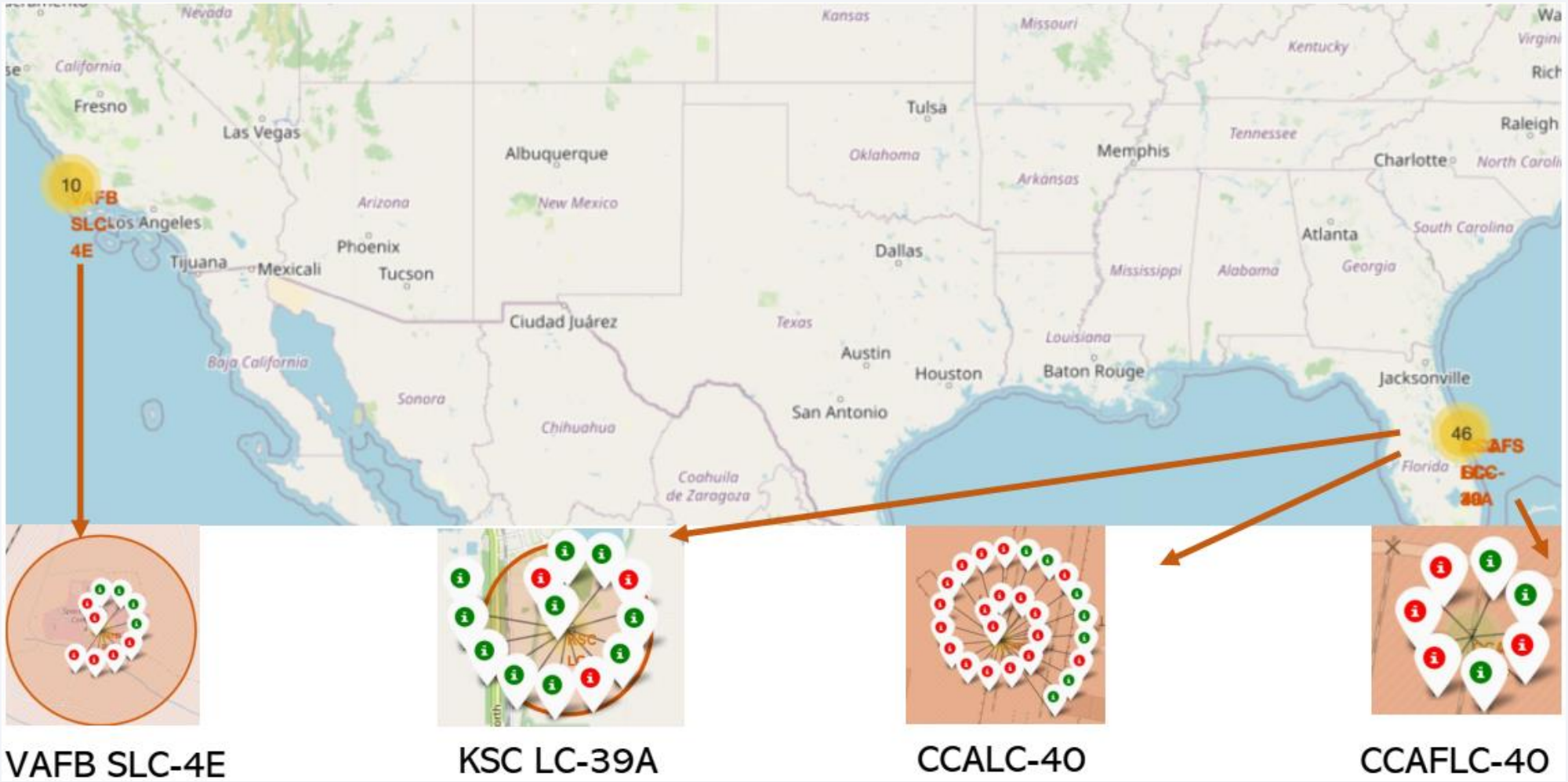

Section 3
Launch Sites
Proximities Analysis

All Launch Sites Marked on a Map

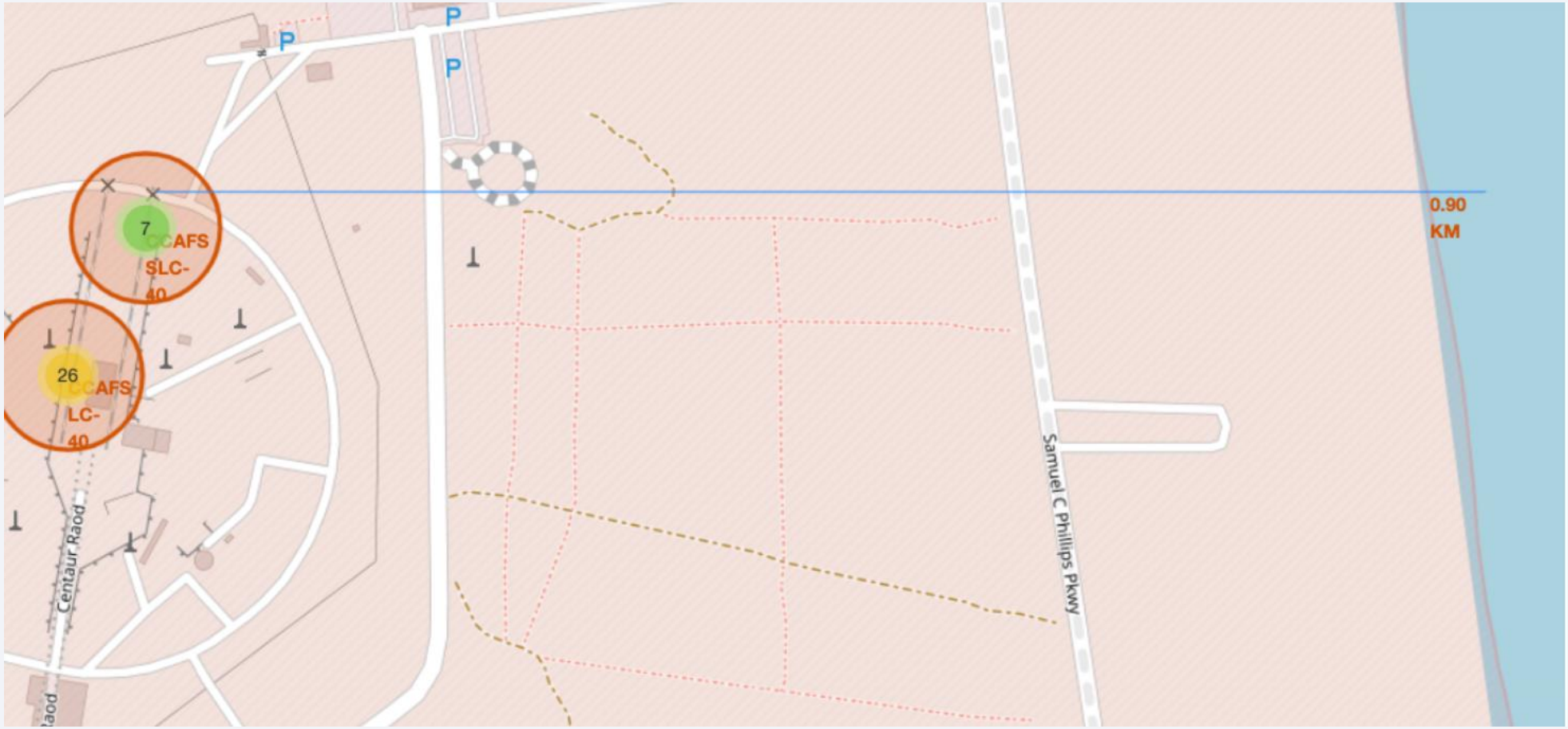


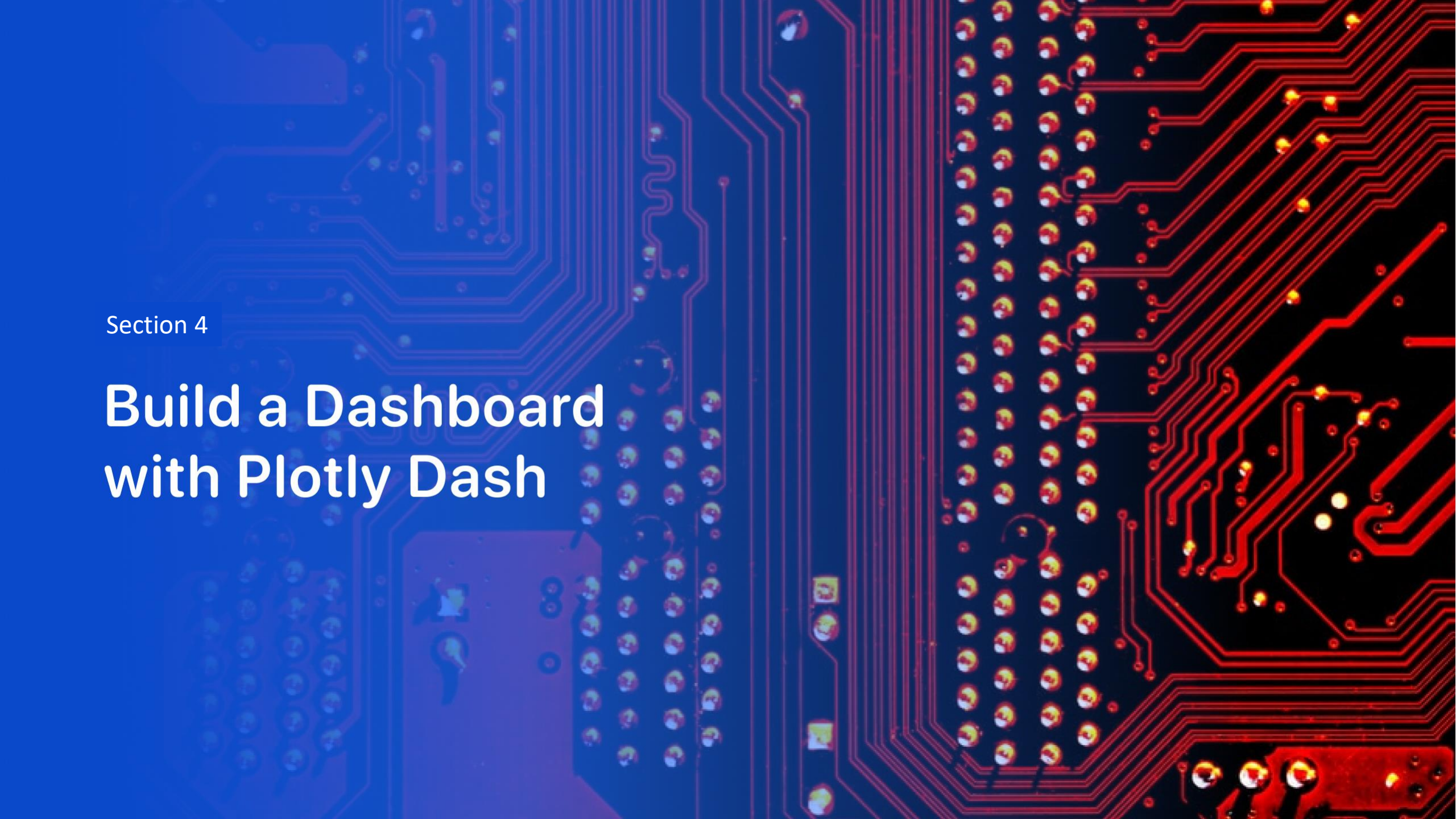
VAFB SLC-4E is situated near the western coastline, while KSC LC-39A, CCAFLC-40, and CCAFSLC-40 are positioned along the eastern coastline.

Success/Failed Launches Marked on the Map

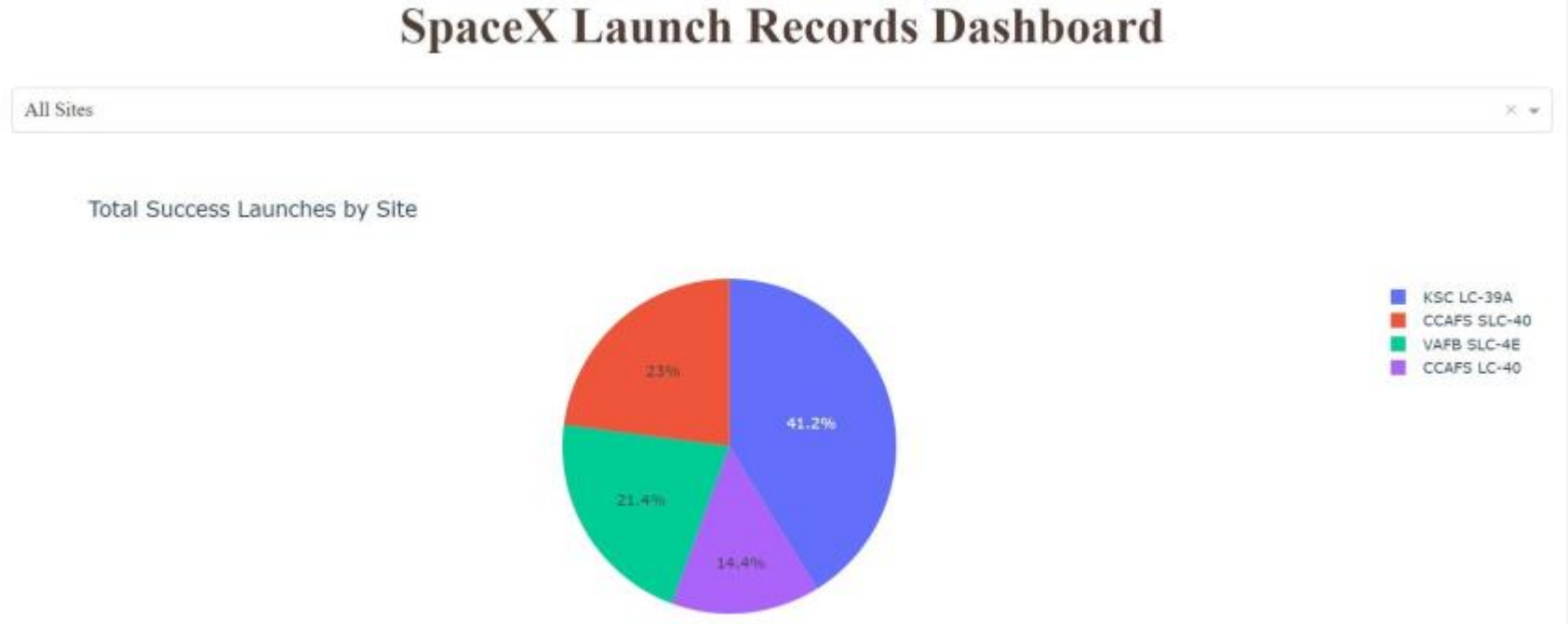


Distances between a Launch Site to its Proximities



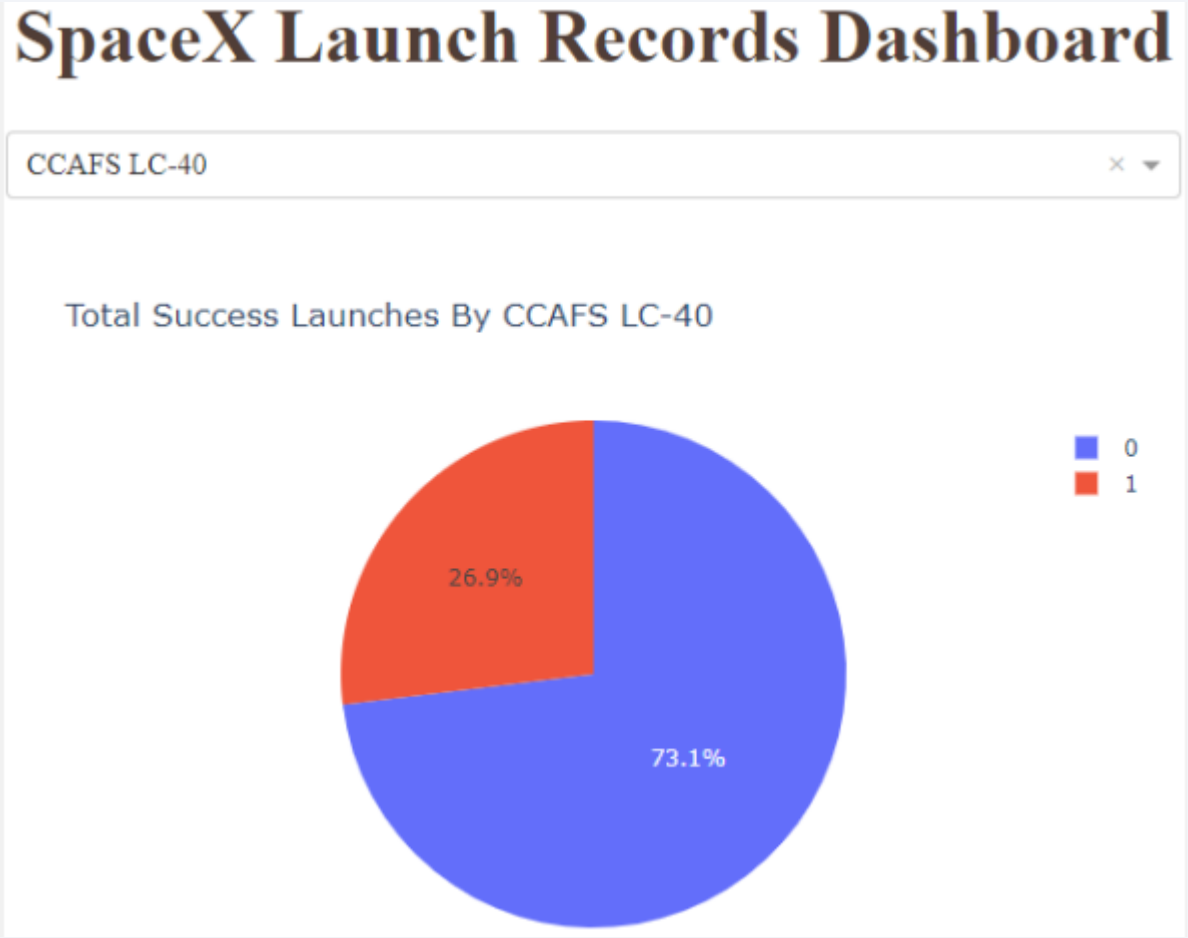

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



KSC LC-39A has the most successful launches amongst launch sites

Launch Site with Highest Launch Success Ratio



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

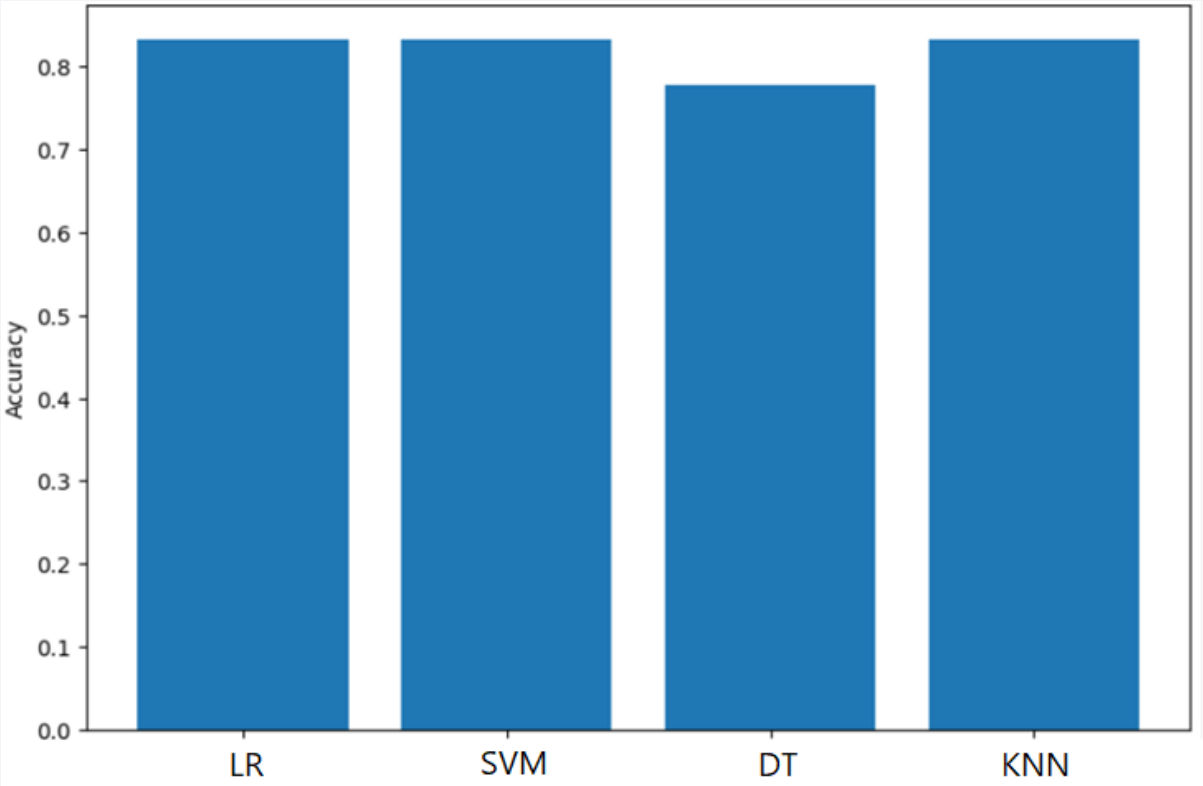
Payload vs. Launch Outcome

Payloads between 2,000 kg and 5,000 kg have the largest success rate

Section 5

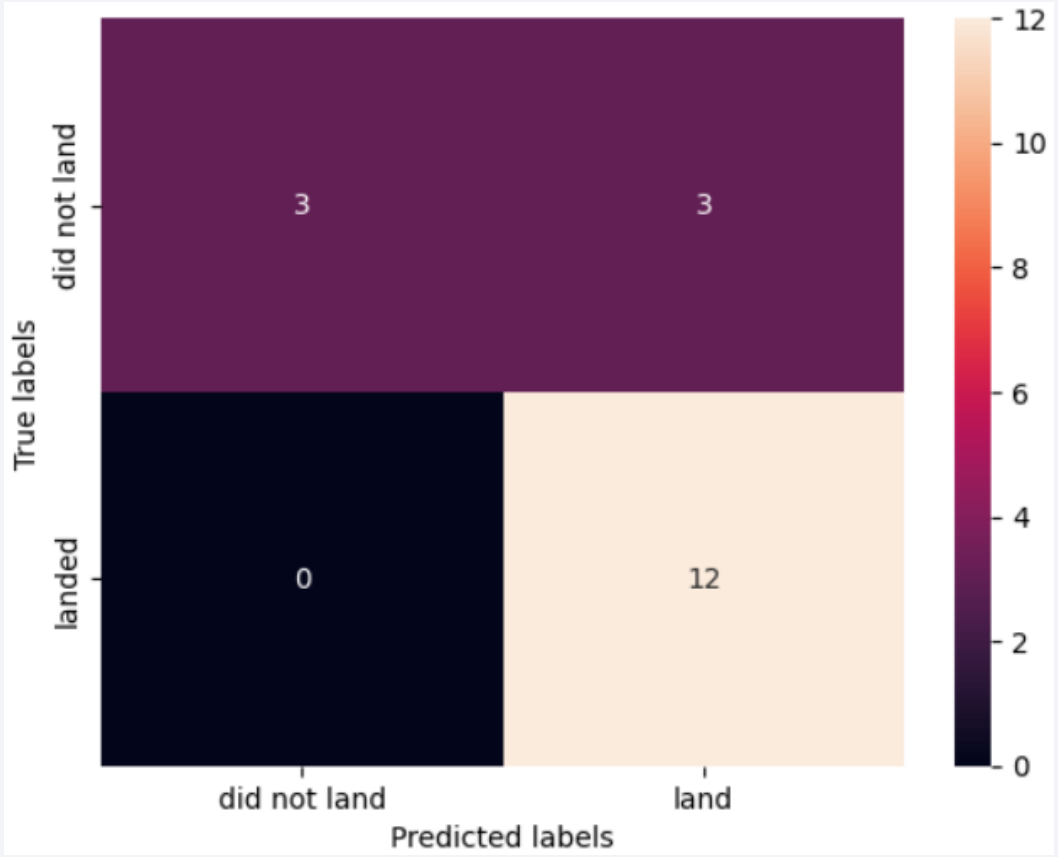
Predictive Analysis (Classification)

Classification Accuracy



After comparing accuracy of above models, they all perform practically the same, except for Decision Tree.

Confusion Matrix



- Confusion matrix of LR, SVM, and KNN
- The fact that there are false positives (Type I error) is not good
- Confusion Matrix Outputs:
  12 True Positive
  3 True Negative
  3 False Positive
  0 False Negative

Conclusions

- The success rates for SpaceX launches are directly proportional time in years they will eventually perfect the launches.
- Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters.
- The Logistic Regression (LR), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) models has the higher accuracy on test data: 0.8334, any of these models can correctly predict the landing of Stage 1 of the Falcon 9 rocket.

Appendix

The Jupyter Notebooks about the analysis are in this GitHub repository:

https://github.com/mahsa-ak91/Applied-Data-Science-Capstone

Thank you!