IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anna Guseva
February 2025

Outline

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

Executive Summary

1. Data for the analysis was collected from a public API and a Wikipedia page on Falcon 9 launches.
2. Data was filtered, missing values were handled, normalization was performed to make data fit for machine learning purposes.
3. Comprehensive analysis including SQL queries, data visualization, creation of interactive maps and dashboards, as well as training machine learning models, was performed.
4. The following conclusions were made:

- The success rate of SpaceX launches was increasing since 2013 till 2017, was stable in 2014, and started increasing again after 2015.
- There is a direct correlation between the success rate and the number of launches.
- Launch sites are located in close proximity to the coast, and far from cities and towns.
- Machine learning models can be used to predict launch success. Decision tree model has the highest accuracy at the moment.

Introduction

- The commercial space age is already here, companies are making space travel affordable for everyone. SpaceX — the most successful commercial space company — advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, whereas other providers cost upward of 165 million dollars each. Much of these savings is due to the fact that SpaceX can reuse the first stage of the launch.
- Thus, the main goal of this project is to determine the price of each launch. We will accomplish it by training a machine learning model and using public information to predict if SpaceX will reuse the first stage.
- This information is invaluable for SpaceY which aims at bidding against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
  - Data was collected from a public API, as well as from a Wikipedia page.
- Data wrangling
  - Data was filtered, missing values were handled, and data was normalized in preparation for machine learning
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium maps and Plotly Dash
- Predictive analysis using classification models
  - Logistic regression, SMV, KNN, and Decision Tree methods were used, and it was established that Decision Tree has the highest accuracy at the moment

Data Collection

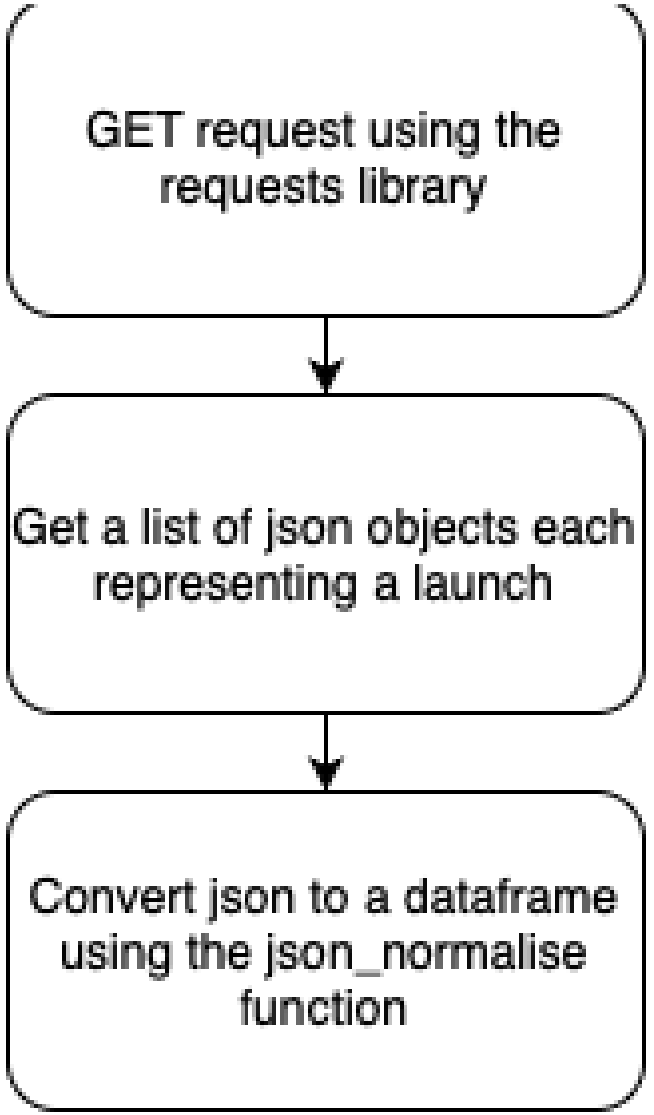
The following data collection methods were used:

1) collection from SpaceX API (https://api.spacexdata.com)

2) webscraping of the Wikipedia page on Falcon 9 launches (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

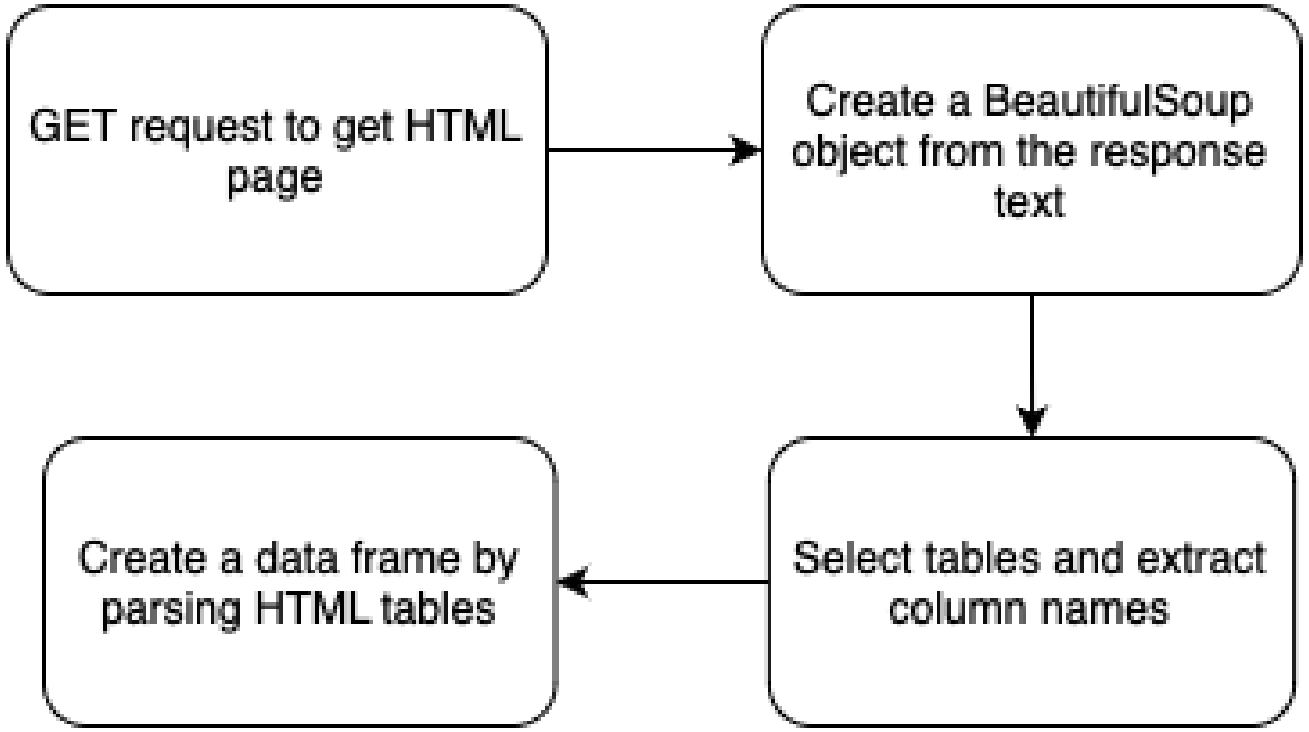
Data Collection – SpaceX API

- We were working with the endpoint api.spacexdata.com/v4/launches/past to get past launch data.
- See the flowchart for process stages
- API Data Collection notebook is available at: https://github.com/tengrin-me/DS-Capstone-Project/blob/main/1_Spacex_Data_Collection-API.ipynb



Data Collection - Scraping

- Data was collected from the Wikipedia page on Falcon 9 launches: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- See the flowchart for process stages
- API Data Collection notebook is available at: https://github.com/tengrin-me/DS-Capstone-Project/blob/main/2_Webscraping.ipynb



Data Wrangling

The following data wrangling methods were implemented:

- API data was filtered to include only Falcon 9 launches;
- Missing payload values were replaced with a mean value;
- 'Class' column specifying whether the launch was a success or a failure was added;
- Dummy variables for categorical columns were created;
- All numeric columns were cast to 'float64';
- Data was standardized for machine learning

The data wrangling Jupiter notebook is available at:

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/3_Data_wrangling.ipynb

EDA with Data Visualization

The following charts were plotted to visualize the relationship between different parameters:

1. a scatter plot showing the impact of Flight Number and Payload Mass on the launch outcome;
2. a scatter plot showing the impact of Flight Number and Launch Site on the launch outcome;
3. a scatter plot showing the impact of Payload Mass and Launch Site on the launch outcome;
4. a bar chart showing the impact of Orbit type on the launch outcome;
5. a scatter plot showing the impact of Flight Number and Orbit on the launch outcome;
6. a scatter plot showing the impact of Payload Mass and Orbit on the launch outcome;
7. a line plot showing the yearly trend;

The Jupiter notebook with the corresponding plots is available at:

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/5_EDA%20with%20Visualization.ipynb

EDA with SQL

The following SQL queries were performed:

- the names of the unique launch sites in the space mission;
- 5 records where launch sites begin with the string 'CCA';
- the total payload mass carried by boosters launched by NASA (CRS);
- the average payload mass carried by booster version F9 v1.1;
- the date when the first successful landing outcome in ground pad was achieved;
- the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000;
- the total number of successful and failure mission outcomes;
- the names of the booster_versions which have carried the maximum payload mass;
- the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015;
- the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

The Jypiter notebook is available at:

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/4_EDA_with_SQL.ipynb

Build an Interactive Map with Folium

An interactive Folium map was created which included the following objects:

- markers to indicate the NASA Johnson Space Center and launch sites;
- circles to indicate the launch sites;
- lines to show the distance between the launch site and the coastline, as well as between the launch site and the nearest town

The Jypiter notebook with the interactive maps is available at:

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/6_Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

An interactive Plotly Dashboard was created which included the following plots and interactions:

- dropdown input component allowing to select one or all of the launch sites;
- a callback function to render success-pie-chart based on selected site dropdown;
- a Range Slider to Select Payload;
- a callback function to render the success-payload-scatter-chart scatter plot

The Python script is available at:

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/7_spacex_dash_app.py

Predictive Analysis (Classification)

The preprocessed and standardized data was split into the training and test sets, then the following machine learning models were created and trained:

1. Logistic Regression
2. Support Vector Machine
3. Decision Tree
4. K-Nearest Neighbours

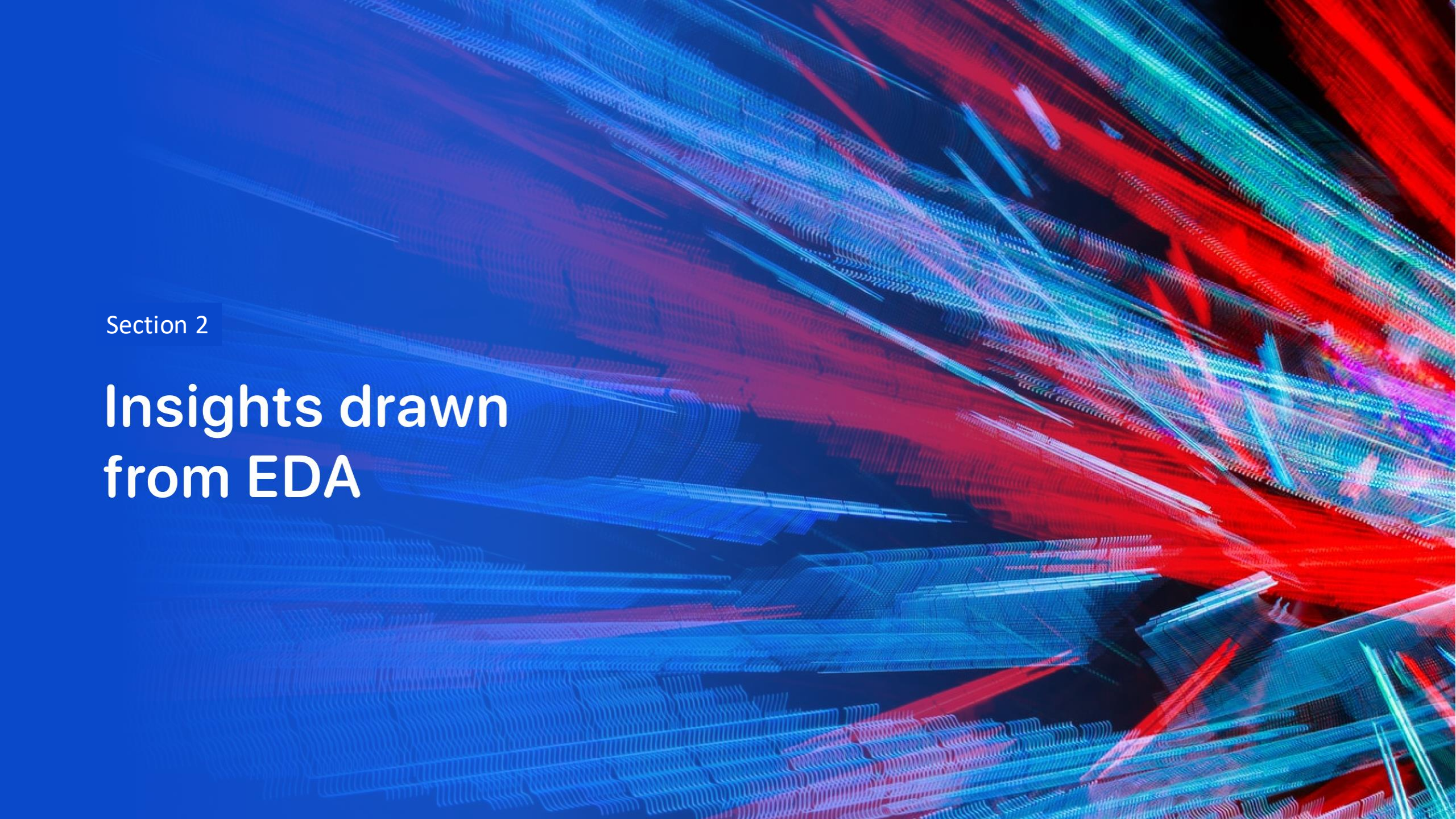
The 4 models were then evaluated on the test set, and the model with the highest accuracy was identified. The Decision Tree model has the highest accuracy at the moment.

The Jypiter notebook with predictive analysis is available at:

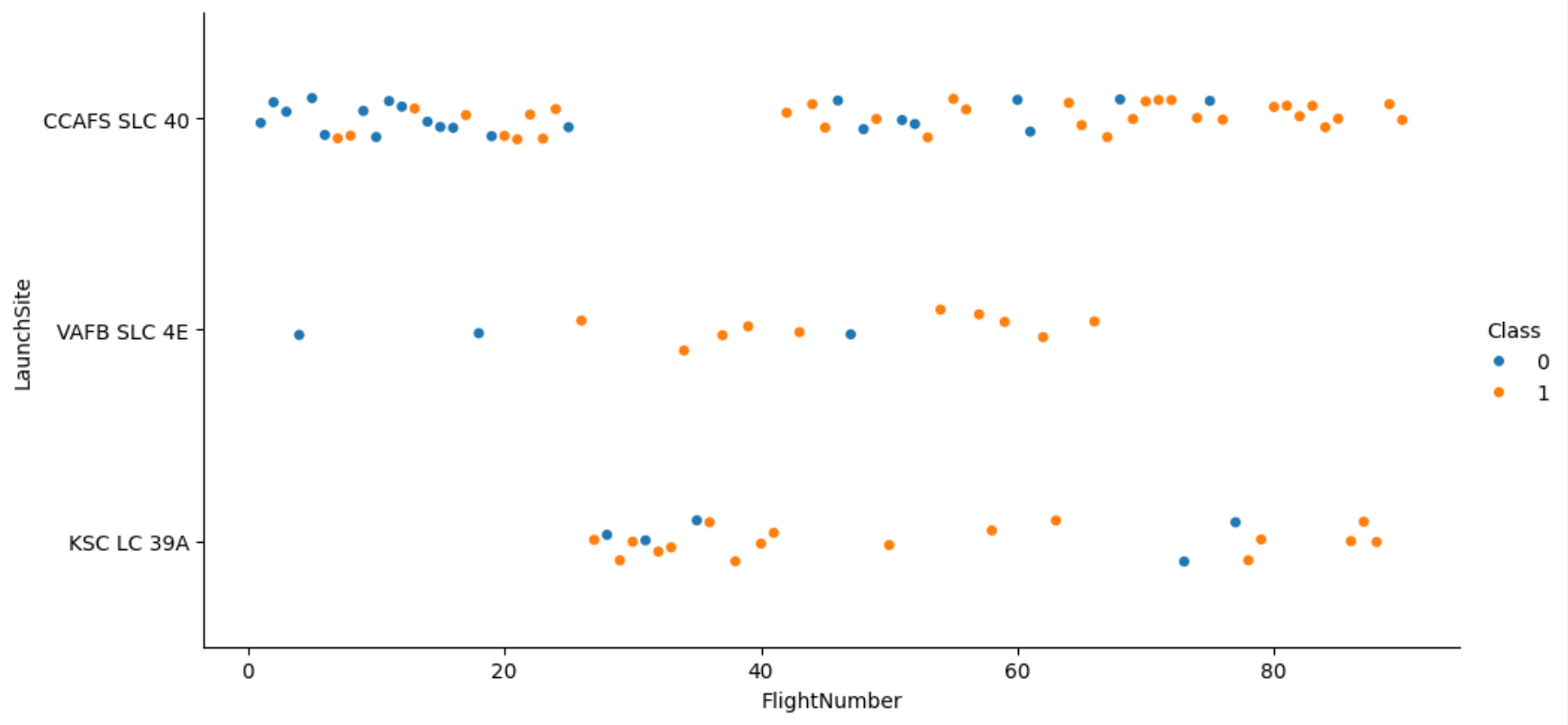
https://github.com/tengrin-me/DS-Capstone-Project/blob/main/8_Machine%20Learning%20Prediction.ipynb

Results

1. Insights from exploratory data analysis
    1.1 Plots showing the relationship between various parameters
    1.2 SQL queries
2. Interactive analytics insights
    2.1 Interactive Folium maps
    2.2 Plotly Dashboard
3. Predictive analysis results (classification)

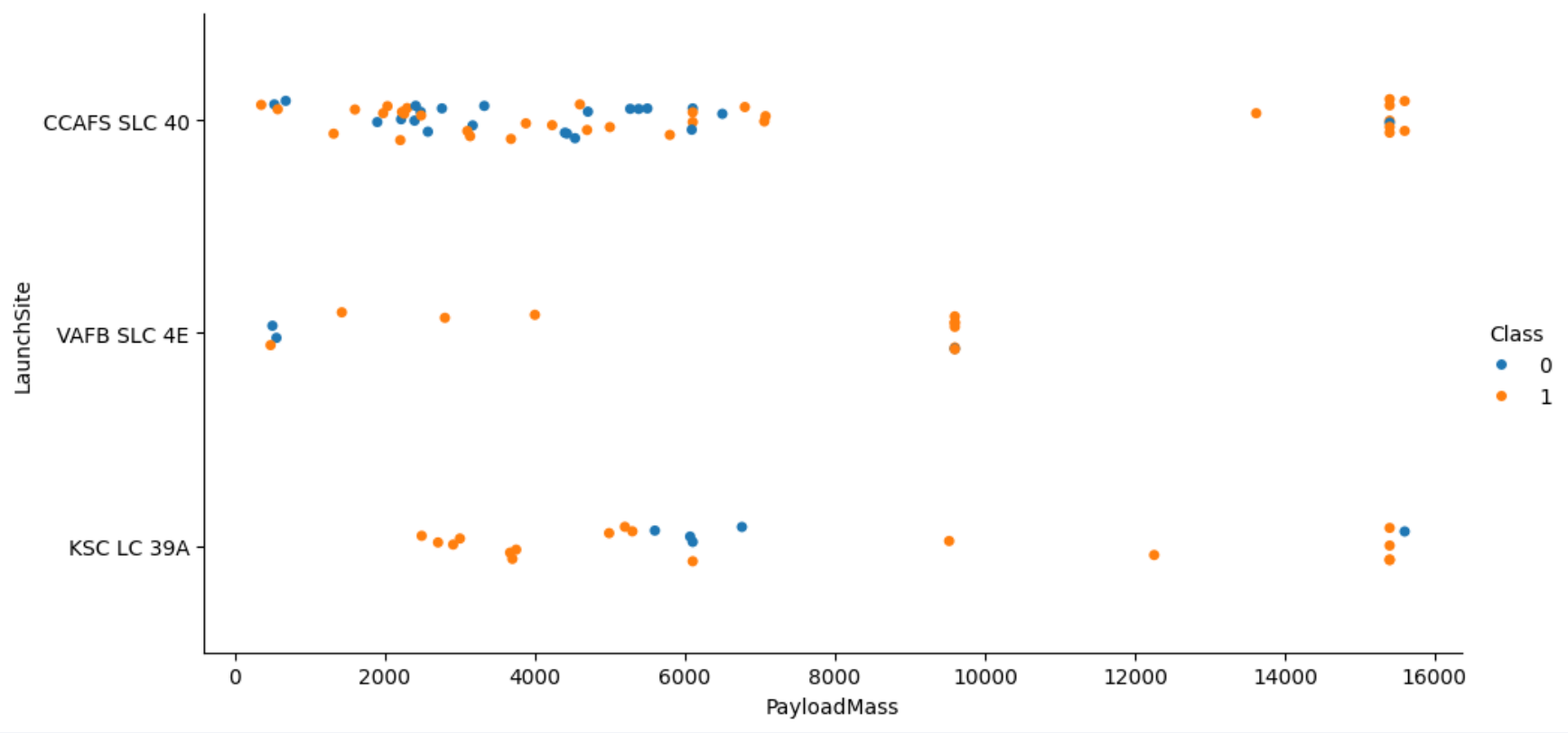

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Success rate varies a lot depending on the launch site
- For VAFB SLC 4E and KSC LC 39A sites the success rate increases along with the number of flights

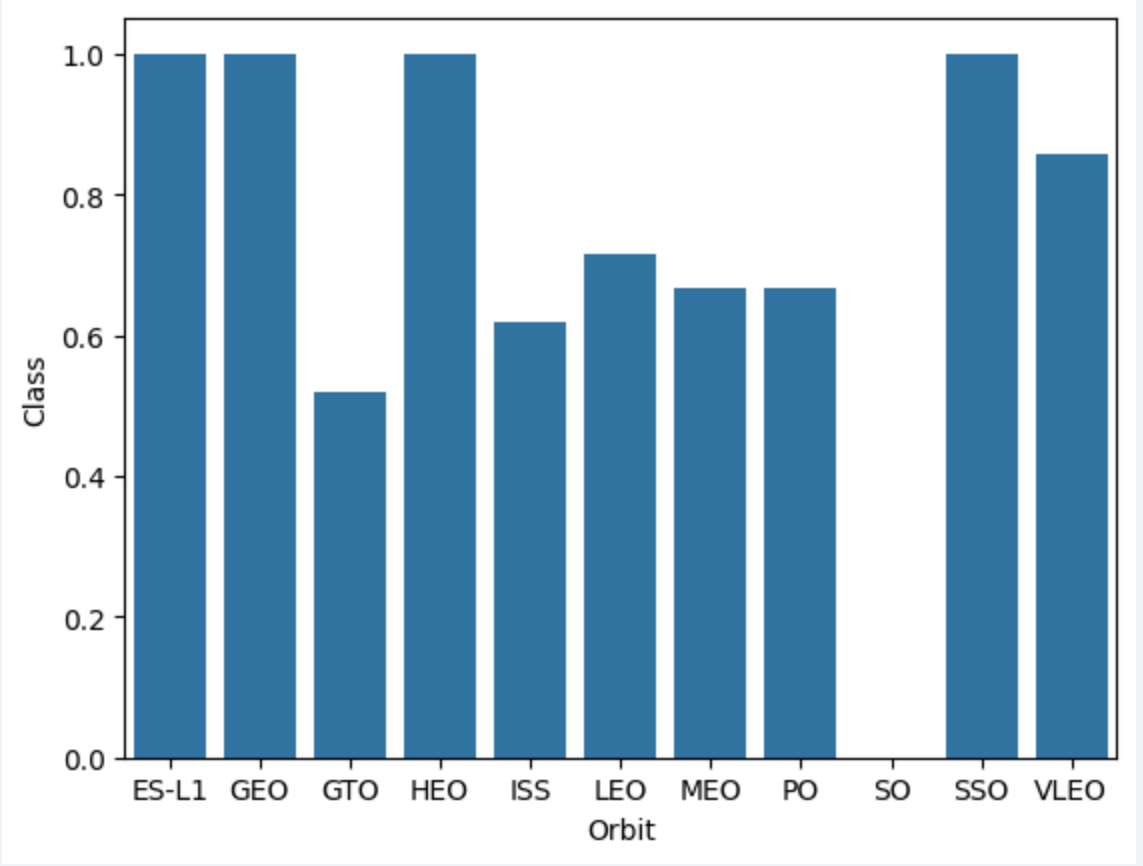
Payload vs. Launch Site



- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)
- There is no strong correlation between the launch site, the payload mass and the success rate

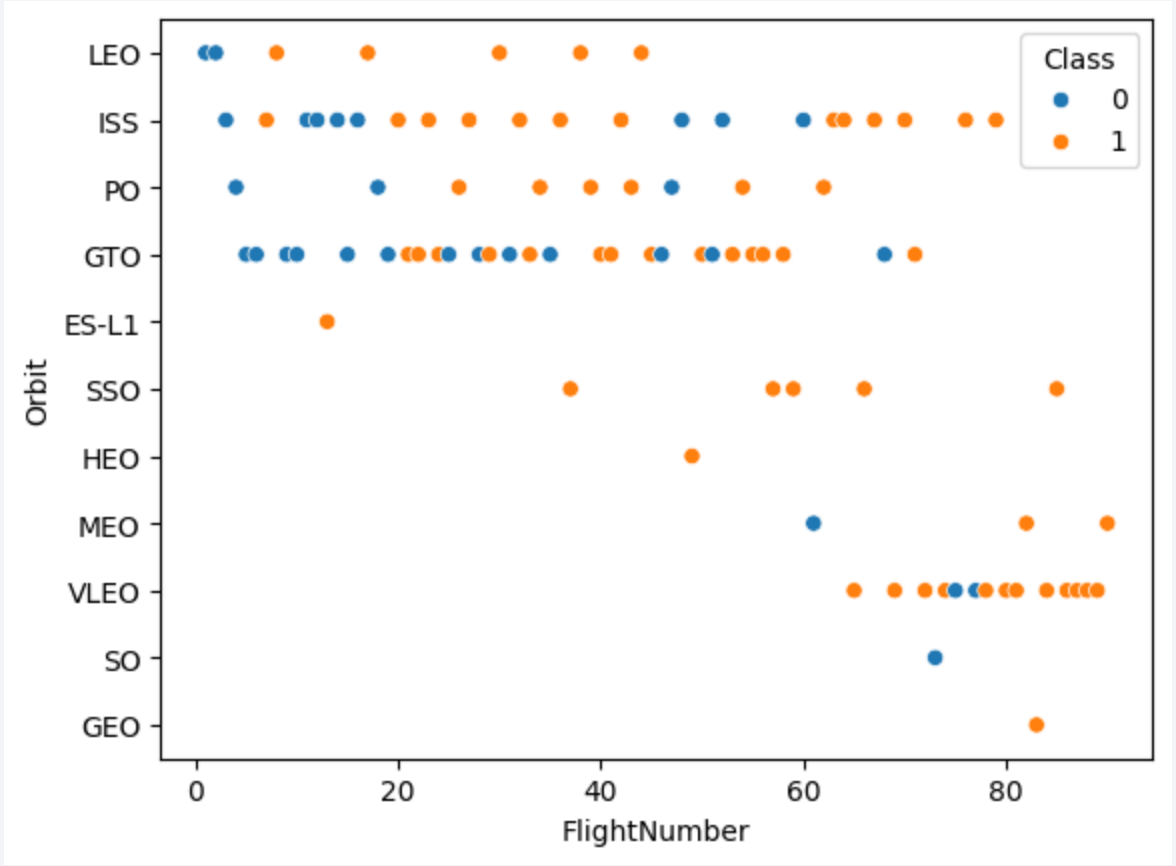
Success Rate vs. Orbit Type

- All launches for ES-L1, GEO, HEO, and SSO orbits were successful
- There were no successful launches for the SO orbit



Flight Number vs. Orbit Type

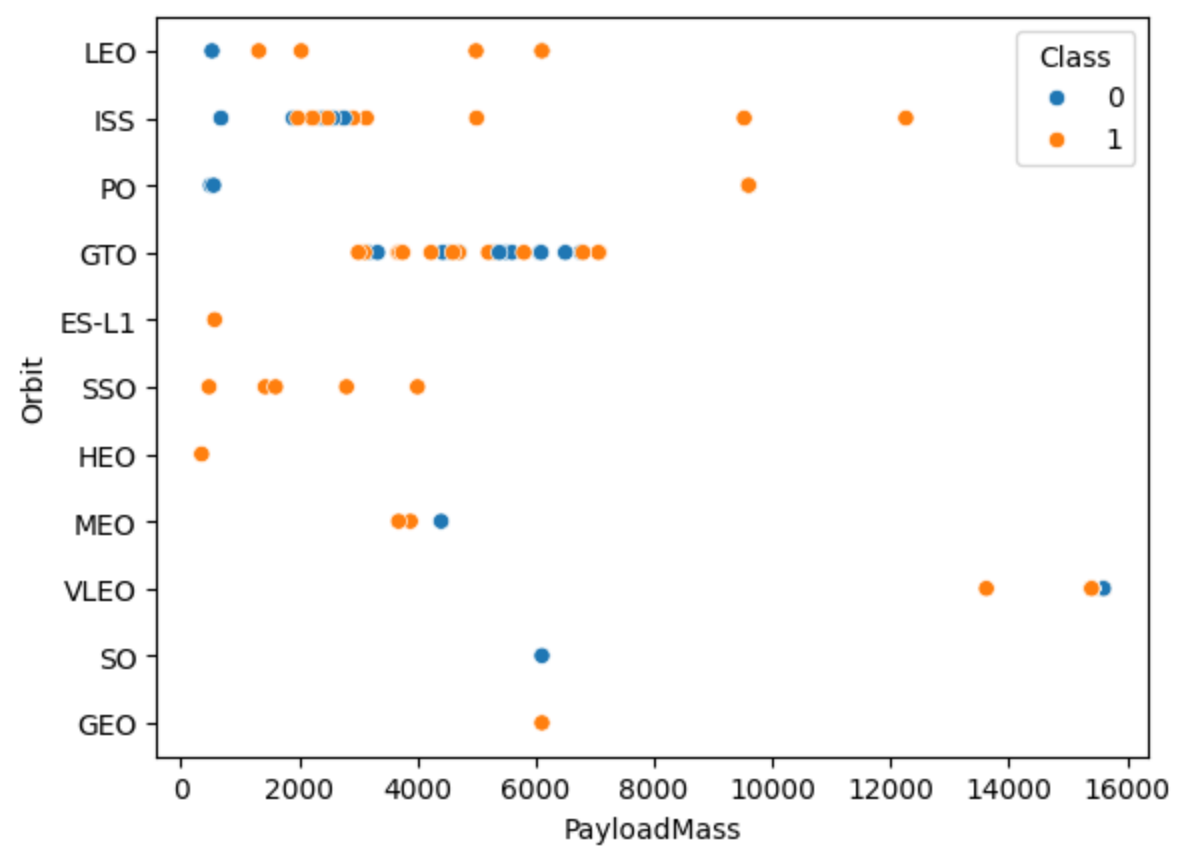
In the LEO orbit the Success appears to be related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit



Payload vs. Orbit Type

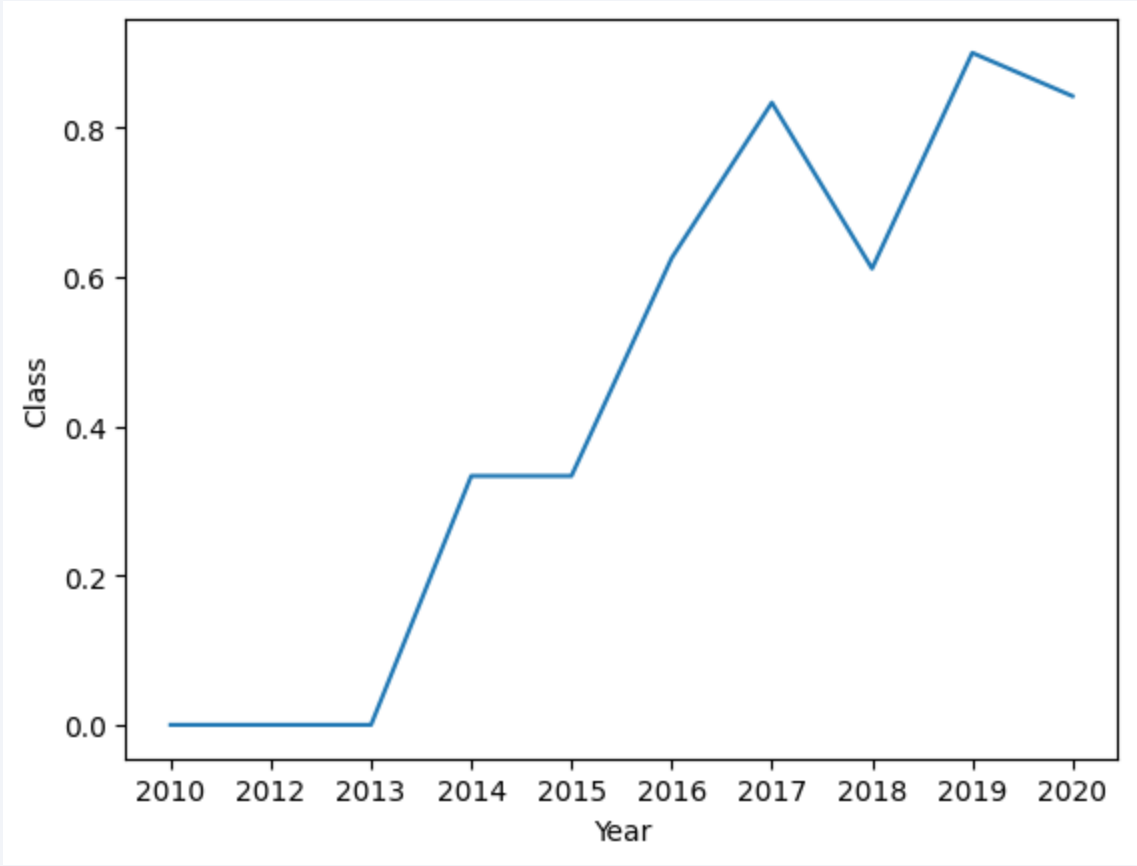
- With heavy payloads the successful landing or positive landing rate are higher for Polar, LEO and ISS orbits

- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful missions) are present



Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing



All Launch Site Names

The list of all launch site names was obtained using the following SQL query:

```
%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'

The 5 records where launch sites begin with `CCA` were obtained using the following SQL query:

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

```
%sql SELECT * from SPACEXTBL where (Launch_Site) LIKE 'CCA%' LIMIT 5;
```

[20] Python

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

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

Total Payload Mass

The total payload carried by boosters from NASA was calculated using the following SQL query:

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

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 was calculated using the following SQL query:

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

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad were identified using the following SQL query:

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

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were listed using the following SQL query:

```
%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

The total number of successful and failure mission outcomes was calculated using the following SQL query:

```
%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 which have carried the maximum payload mass were listed using the following SQL query:

```
%sql SELECT DISTINCT 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 failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 were listed using the following SQL query:

[43]
```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
    FROM SPACEXTBL where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 were ranked in descending order using the following SQL query:

```
%sql SELECT [Landing_Outcome], count(*) as Count FROM SPACEXTBL WHERE DATE \
    between '2010-06-04' and '2017-03-20' group by [Landing_Outcome] order by Count DESC;
```

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

Landing_Outcome	Count
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

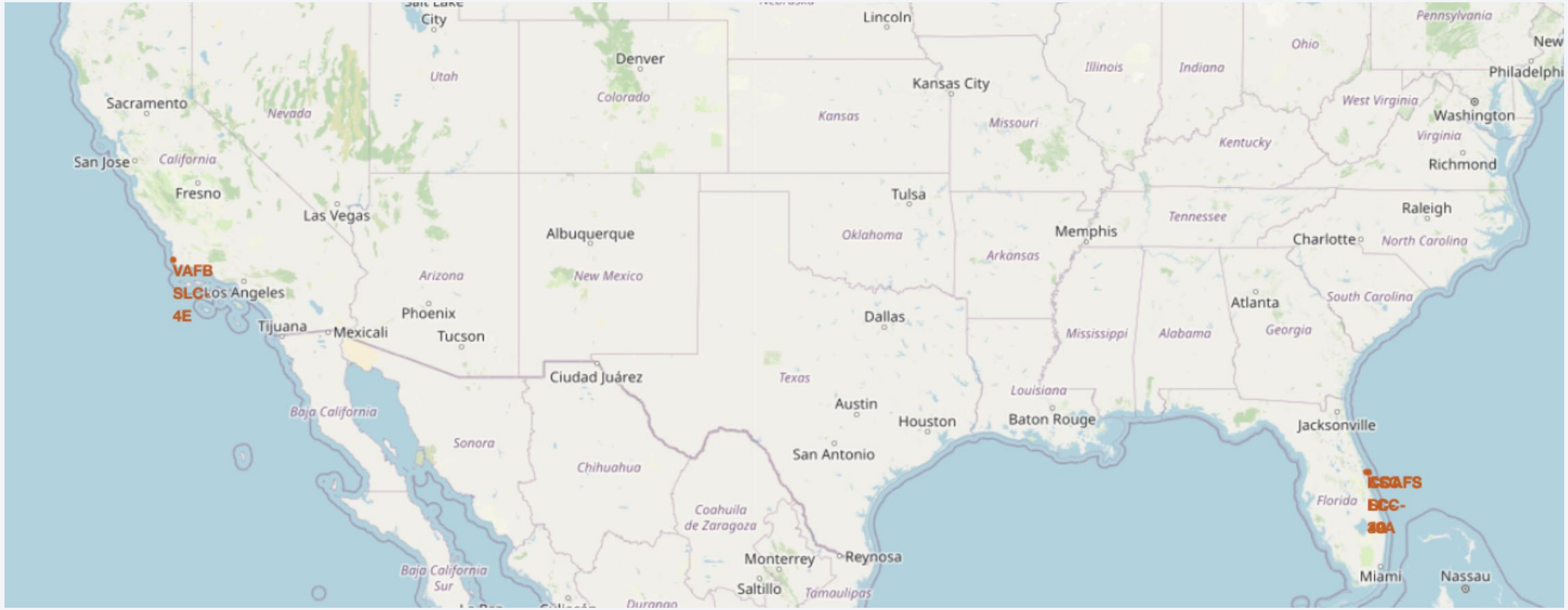
Section 3

Launch Sites Proximities Analysis

Locations of All Launch Sites

The map shows locations of all launch sites

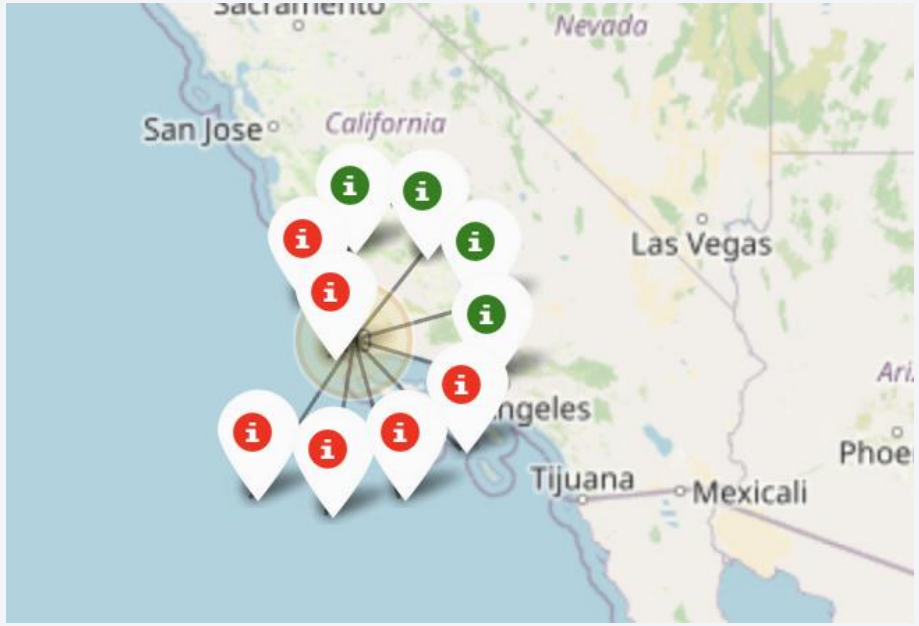
All sites are located in close proximity to the coast.

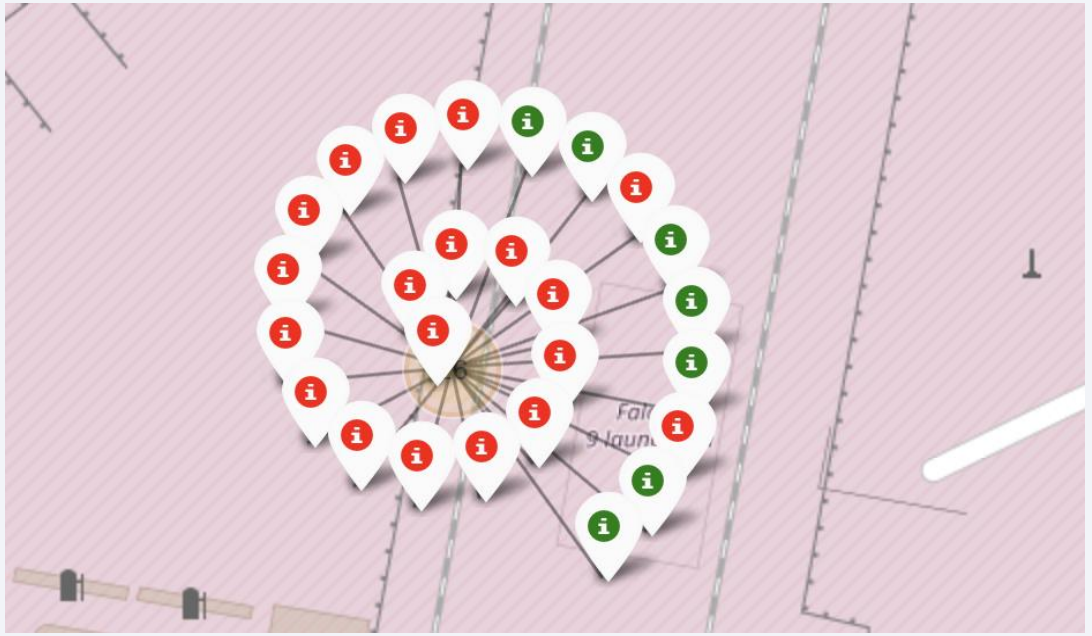


Success/Failed Launches for Each Site

The following screenshots show examples of the color-labeled launch outcomes.

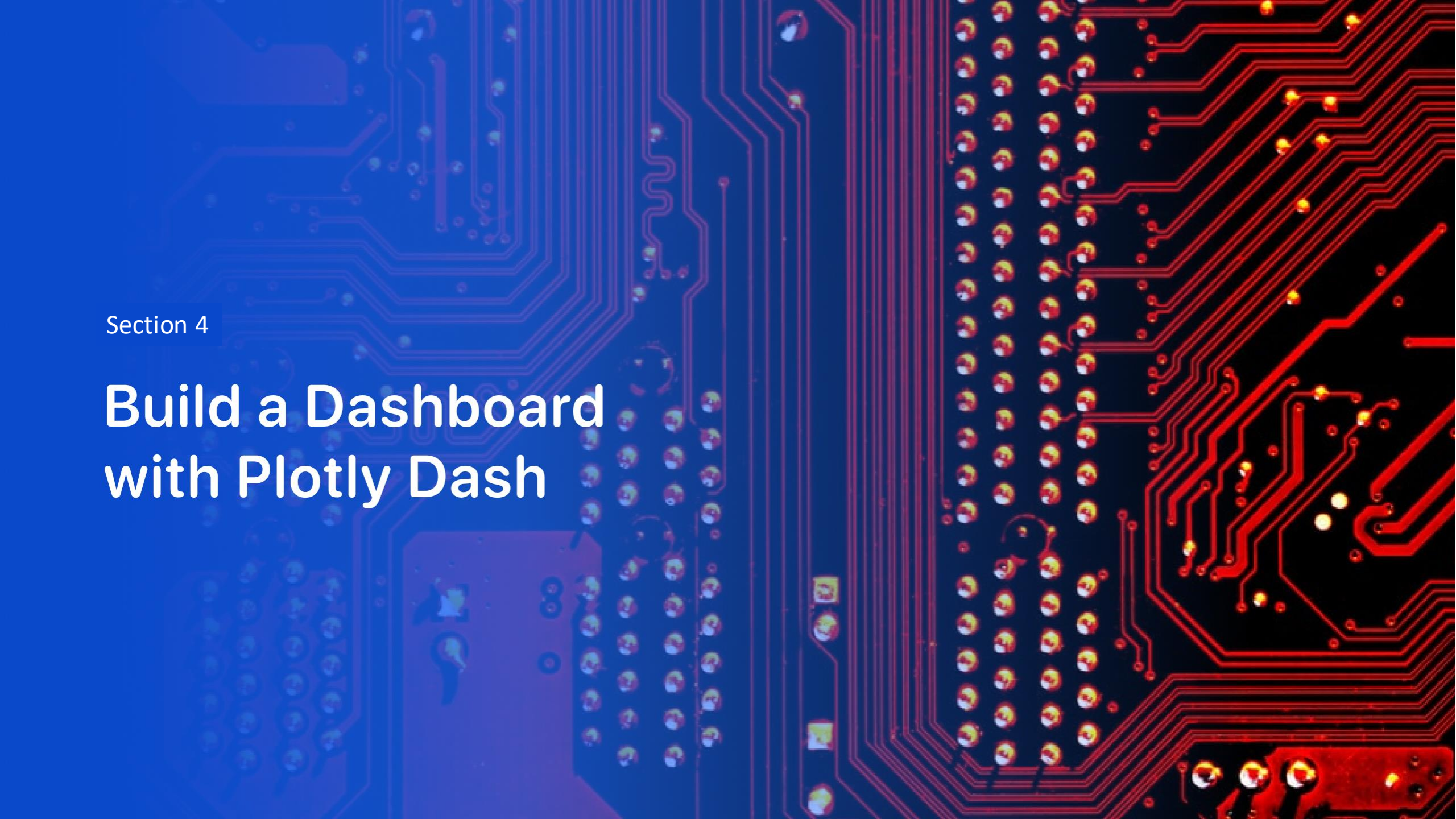
Such visualization helps to identify which launch sites have relatively high success rates.





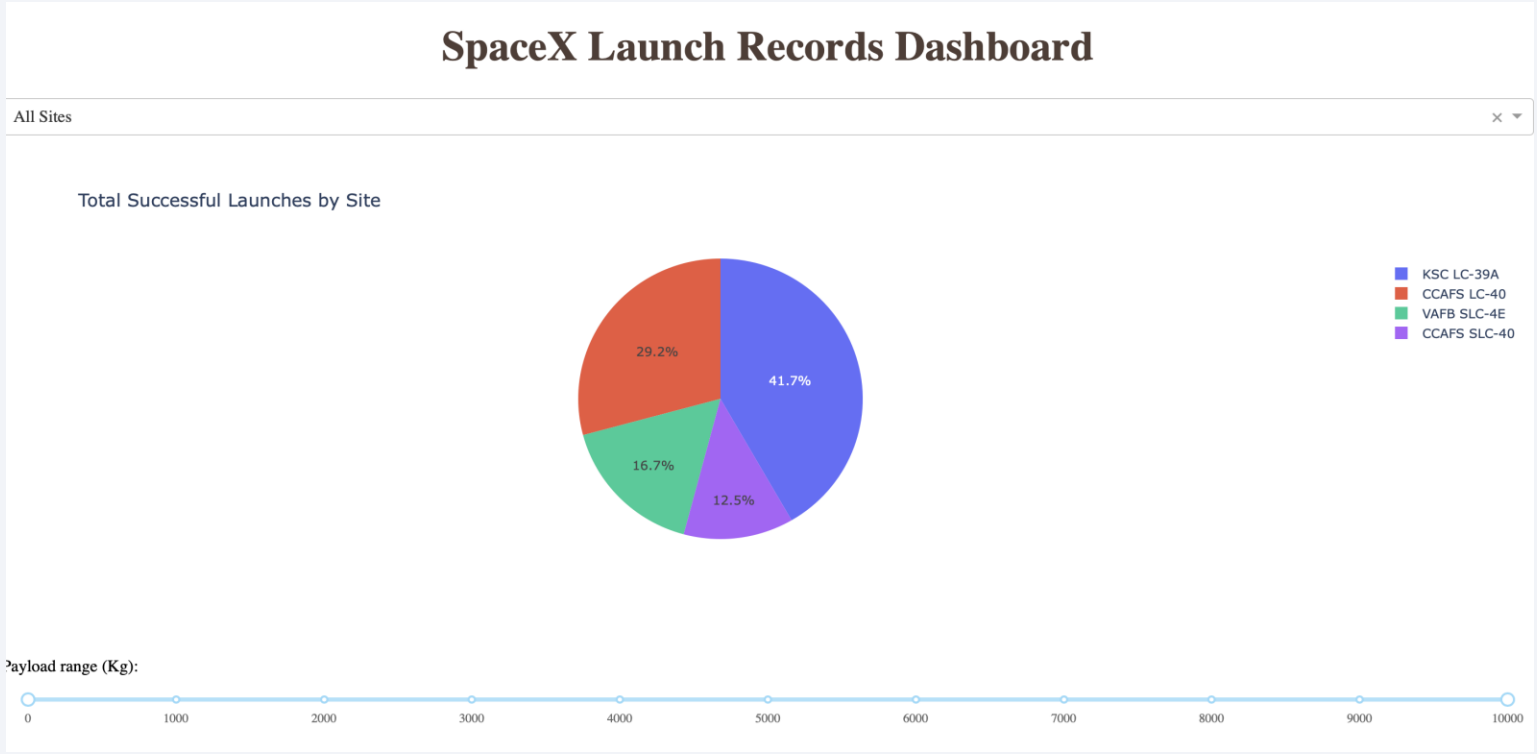
Launch Site Proximity to Coastline and Town

The following screenshots show the distance between one of the launch sites and the coastline, as well as the nearest town


Section 4
Build a Dashboard with Plotly Dash

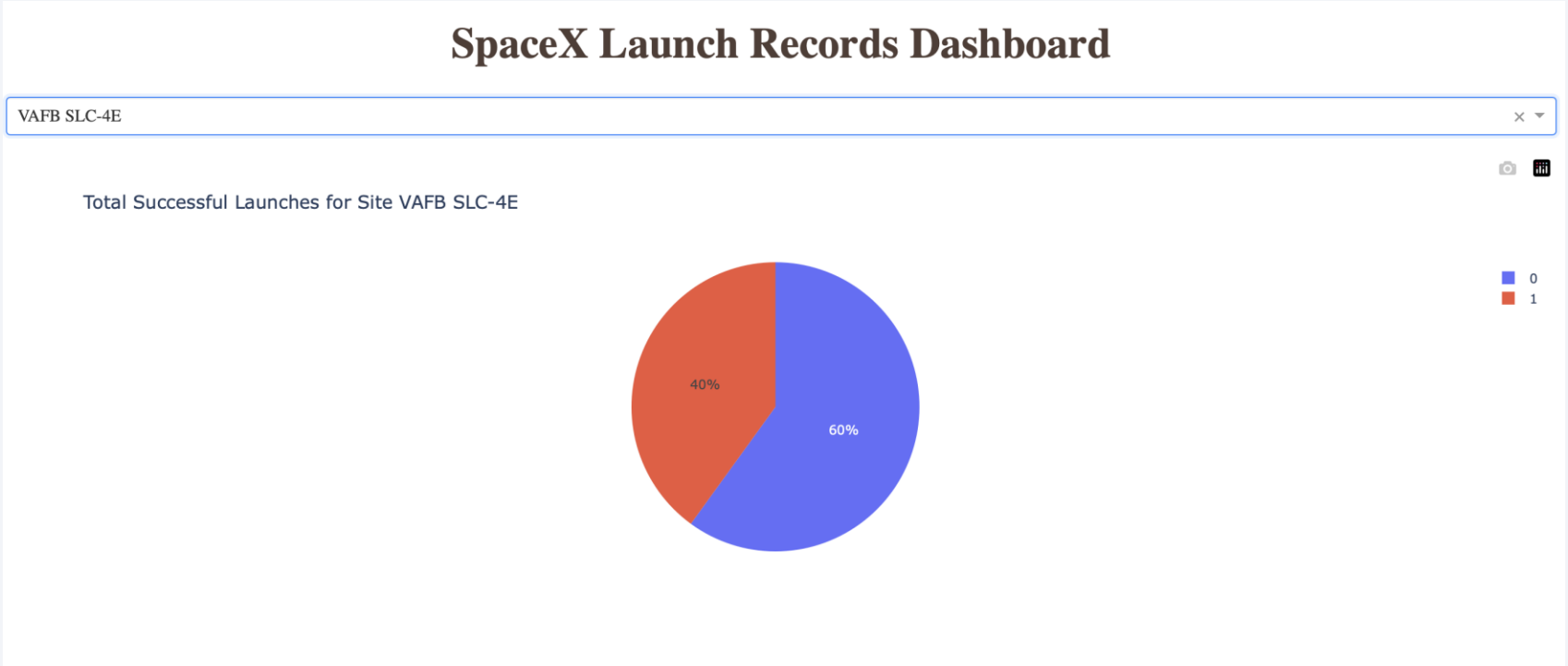
Total Successful Launches by Site

The pie chart shows the count of successful launches for all sites



Total Successful Launches for One Site

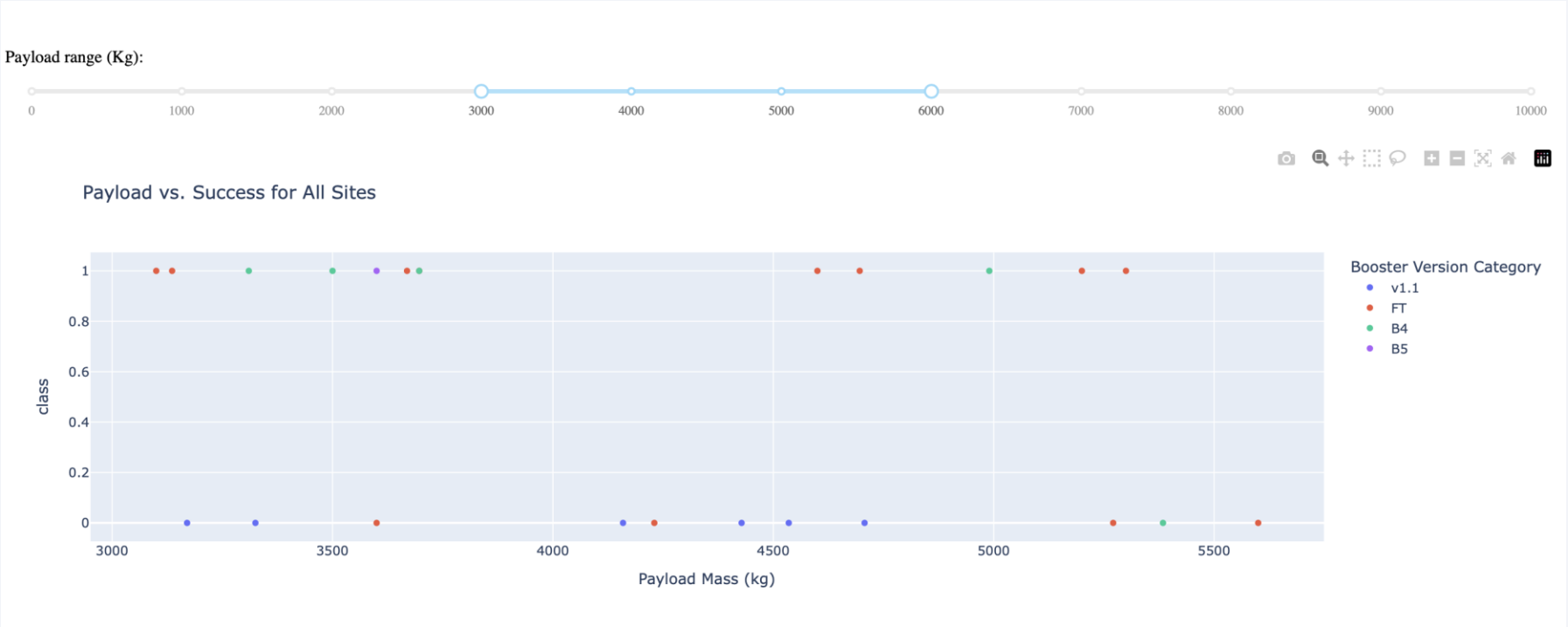
The dropdown menu allows to select a particular launch site. A pie chart showing the statistics of launch failure/success is displayed then



Payload vs. Launch Outcome

A scatter plot showing the relationship between the payload mass and the launch outcome is shown below.

A range slider to select payload has been implemented on the dashboard.

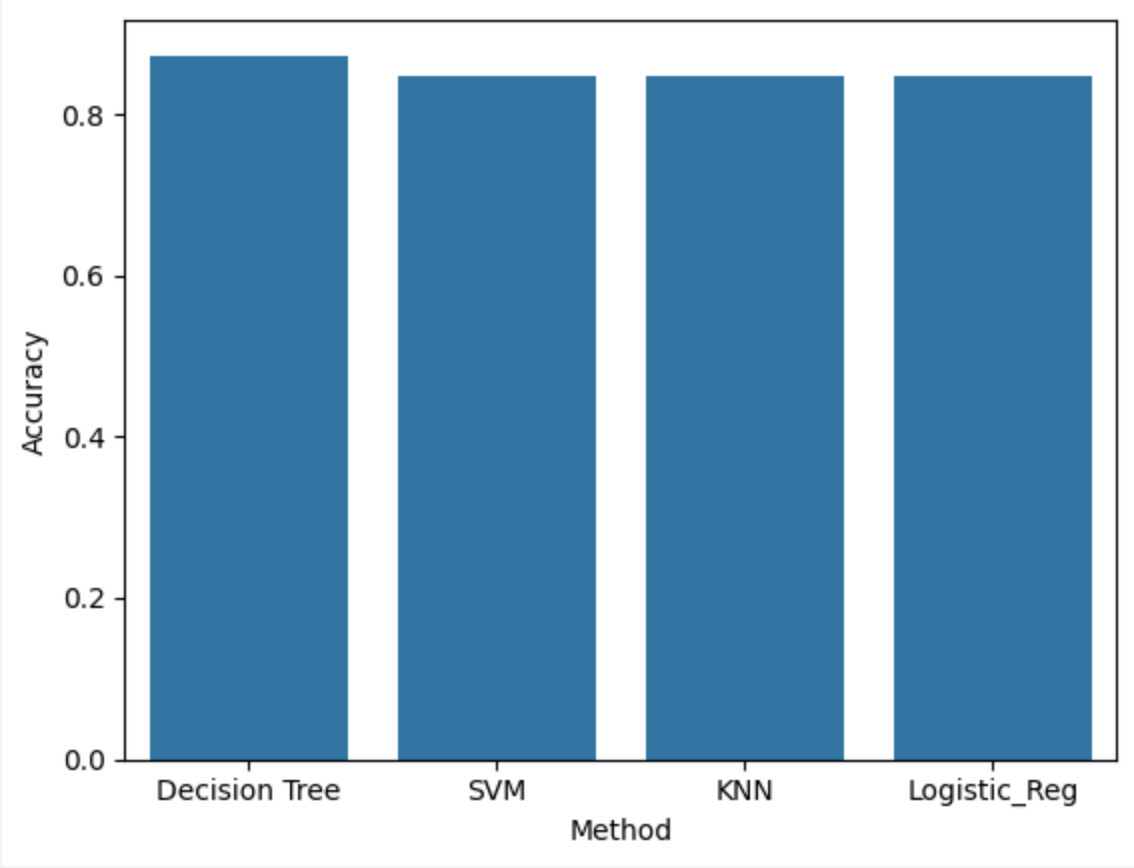


Section 5

Predictive Analysis (Classification)

Classification Accuracy

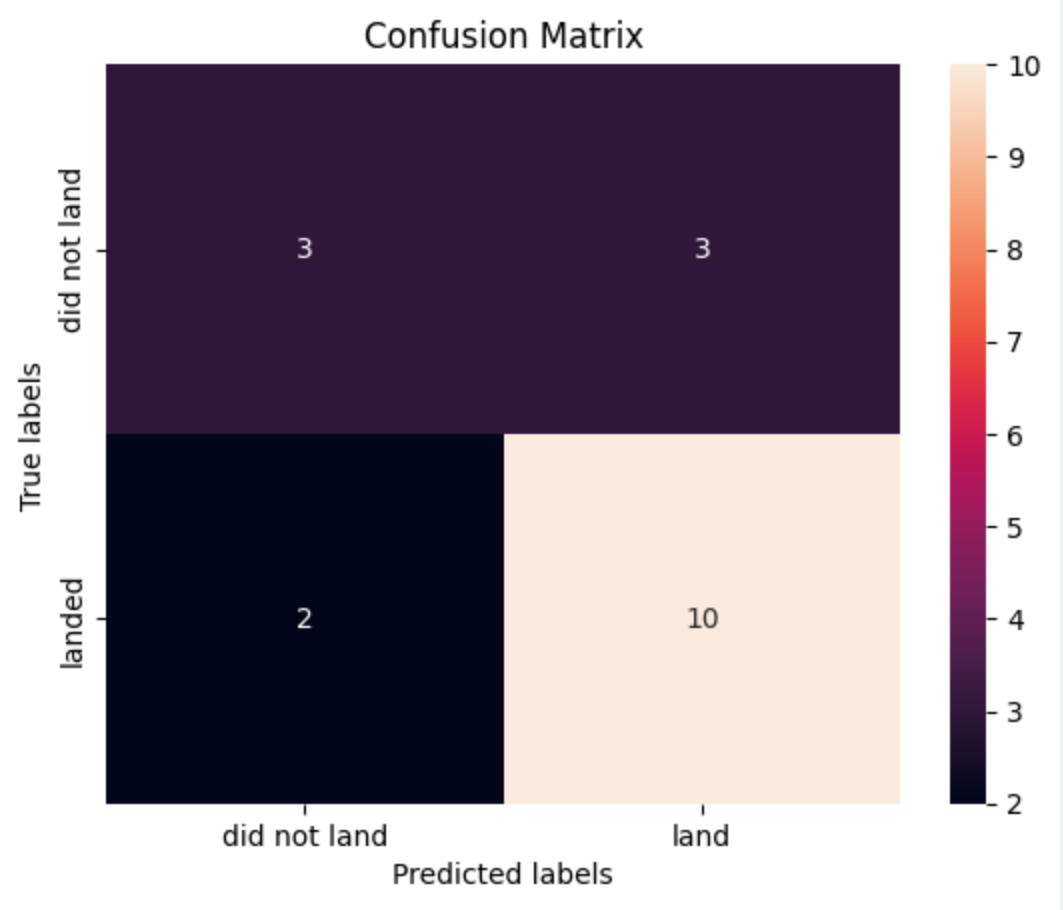
Out of the 4 models, the Decision Tree had the highest accuracy equal to 0.87



Confusion Matrix

The figure shows the Confusion Matrix for the Decision Tree model.

Examining the confusion matrix, we can see that the method can distinguish between the different classes. The major problem is false positives.



Conclusions

1. The success rate of SpaceX launches was increasing since 2013 till 2017, was stable in 2014, and started increasing again after 2015.

2. There is a direct correlation between the success rate and the number of launches.

3. Launch sites are located in close proximity to the coast, and far from cities and towns.

4. Machine learning models can be used to predict launch success. Decision tree model has the highest accuracy at the moment.

Appendix

Sources:

SpaceX API

https://api.spacexdata.com

Wikipedia page on Falcon 9 launches

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Cleaned data sets:

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/dataset_part_1.csv

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/dataset_part_2.csv

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/dataset_part_3.csv

Appendix

Jupiter Notebooks:

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/1_Spacex_Data_Collection-API.ipynb

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/2_Webscraping.ipynb

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/3_Data_wrangling.ipynb

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/4_EDA_with_SQL.ipynb

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/5_EDA%20with%20Visualization.ipynb

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/6_Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/7_spacex_dash_app.py

https://github.com/tengrin-me/DS-Capstone-Project/blob/main/8_Machine%20Learning%20Prediction.ipynb

Thank you!