Winning Space Race with Data Science

Alireza Aminidad

April 10th 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

The space industry has seen a surge of interest in reusable rocket technology, with Space X pioneering the way by offering Falcon 9 rocket launches at a significantly lower cost than other providers. This cost savings is made possible by Space X's ability to reuse the first stage of the rocket, which reduces the overall cost of production and launch. By reusing the first stage, Space X can minimize the resources required to build a new rocket for each launch, resulting in a more sustainable and cost-effective approach to space travel.

Predicting the successful landing of the first stage is crucial in determining the overall cost of the launch, making it a valuable tool for companies competing for rocket launch bids. In this project, we aim to develop a machine learning process that can accurately forecast the first stage's successful landing. By leveraging machine learning algorithms and analyzing a range of data sources we hope to create a robust and reliable model that can help optimize the landing process and reduce the risk of mission failure.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Wikipedia – Web Scrapping
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

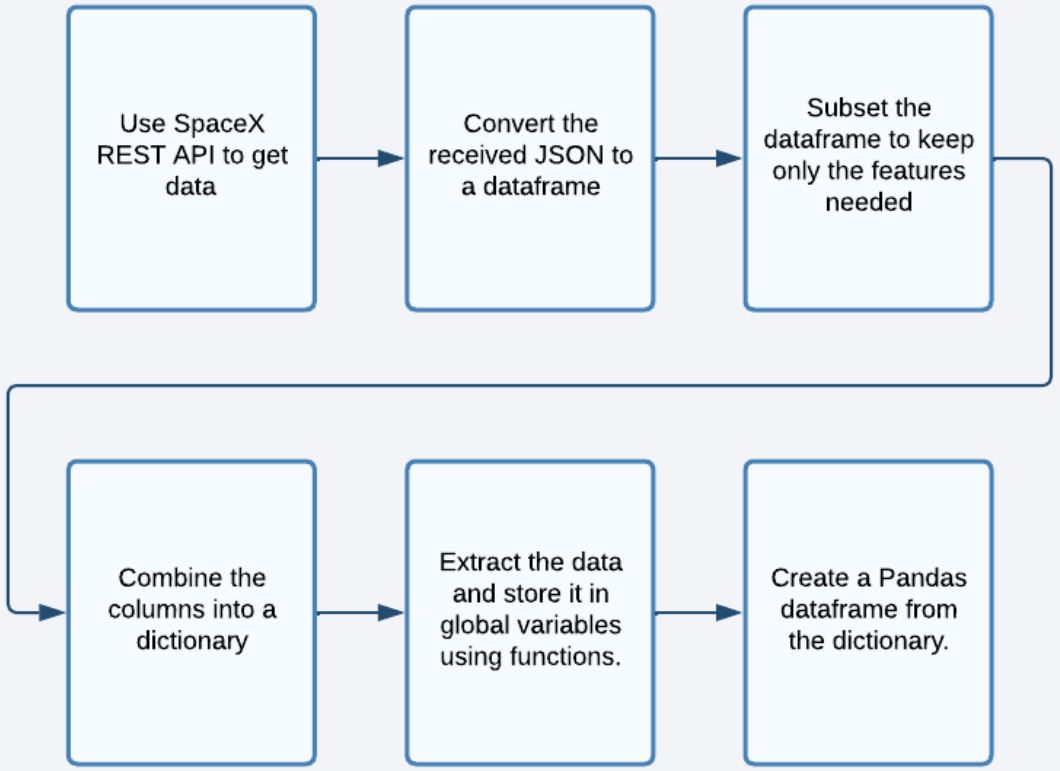
Various techniques were utilized to gather the data. Specifically, the data collection was performed through a GET request to the SpaceX API.

The response content was decoded as JSON using the ".json() function", and the resulting JSON data was transformed into pandas dataframe using ".json_normalize()".

The data was subsequently cleaned, and missing values were addressed as needed. Additionally, web scraping was performed on Wikipedia using BeautifulSoup to extract Falcon 9 launch records in the form of an HTML table, which was then parsed and converted into a pandas dataframe for further analysis.

Data Collection – SpaceX API

The data was retrieved from the SpaceX API using a GET request, after which data cleaning, basic wrangling, and formatting were performed to ensure accuracy, organization, and readability.



GitHub Link:

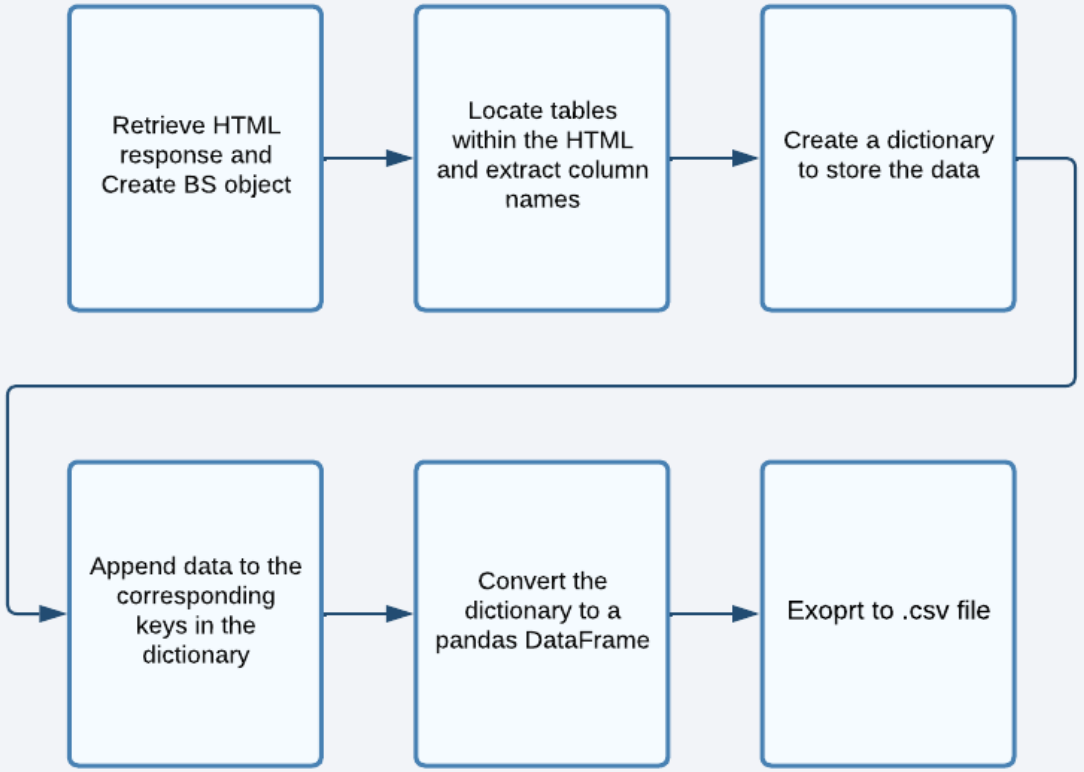
Step 1: Data Collection

Data Collection – Web Scraping

The data was retrieved from the SpaceX API using a GET request, after which data cleaning, basic wrangling, and formatting were performed to ensure accuracy, organization, and readability.

GitHub Link:

Step 1.1: Web-Scrapping



Data Wrangling

- The dataset includes mission outcomes, such as “True Ocean” indicating a successful landing in a specific oceanic region, and “False Ocean” indicating an unsuccessful landing in a specific oceanic region.
- Similarly, “True RTLS” represents a successful ground pad landing, while “False RTLS” represents an unsuccessful ground pad landing. “True ASDS” indicates a successful drone ship landing, while “False ASDS” indicates an unsuccessful drone ship landing.
- These outcomes are converted into training labels, where a value of 1 denotes a successful booster landing, and a value of 0 denotes an unsuccessful landing.

GitHub link:

Step 2: Data Wrangling

Perform exploratory data analysis (EDA) using visualization and SQL

- EDA is an approach to analyzing data that involves summarizing its main characteristics and identifying patterns, trends, and relationships among variables. Visualization is an essential tool for EDA as it allows for the representation of data in a visual format, making it easier to identify patterns and trends.
- SQL, on the other hand, is a programming language used for managing and analyzing relational databases.
- By combining visualization and SQL, one can perform a comprehensive EDA that involves querying and analyzing data from databases and representing it visually to gain insights into the underlying patterns and relationships.

GitHub link:
Step 3: Data Visualization
Step 3.1: EDA with SQL

Building an Interactive Map with Folium

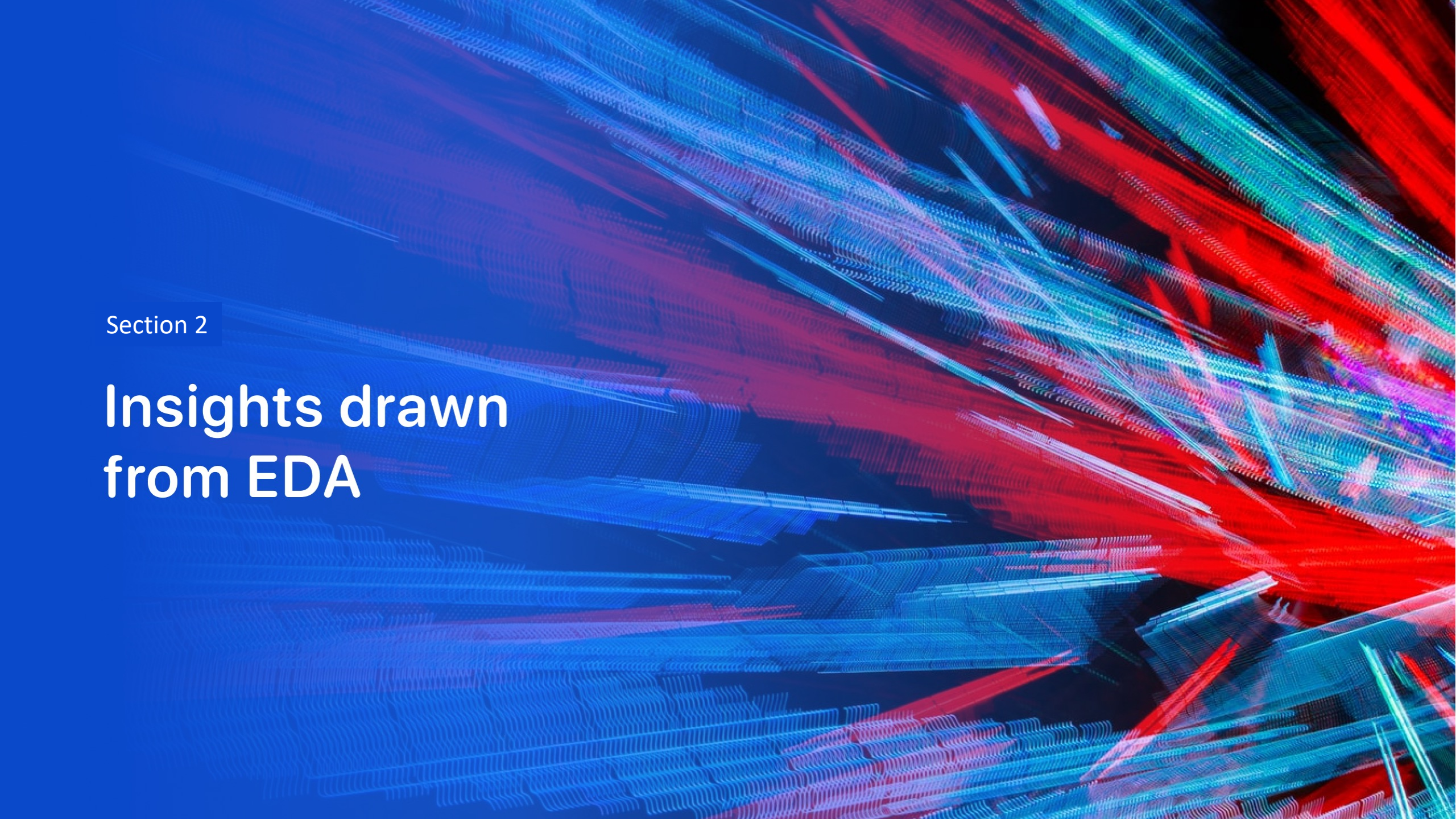
- The launch data was visualized as an interactive map by adding circle markers with labels around each launch site using latitude and longitude coordinates.
- The dataframe launch_outcomes was assigned to classes 0 and 1 with green and red markers respectively using MarkerCluster(). The Haversine formula was used to calculate the distance between the launch site and various landmarks to determine patterns around the launch site.
- The map showed that launch sites are not in close proximity to railways or highways but are near the coastline. Additionally, launch sites tend to keep a certain distance away from cities.

GitHub link:
Step 4: Launch Sites Locations Analysis with Folium

Perform predictive analysis using classification models

- Predictive analysis involves using statistical and machine learning techniques to analyze historical data and make predictions about future outcomes.
- Classification models are a type of machine learning algorithm used to classify data into different categories or classes.
- Building a classification model involves selecting a suitable algorithm, preparing the data, and training the model.
- Tuning the model involves optimizing its hyperparameters to improve its performance.
- Evaluating the model involves testing its accuracy, precision, recall, and other metrics to determine its effectiveness in predicting outcomes.
- By following these steps, one can develop accurate and reliable classification models that can be used to make predictions about future events.

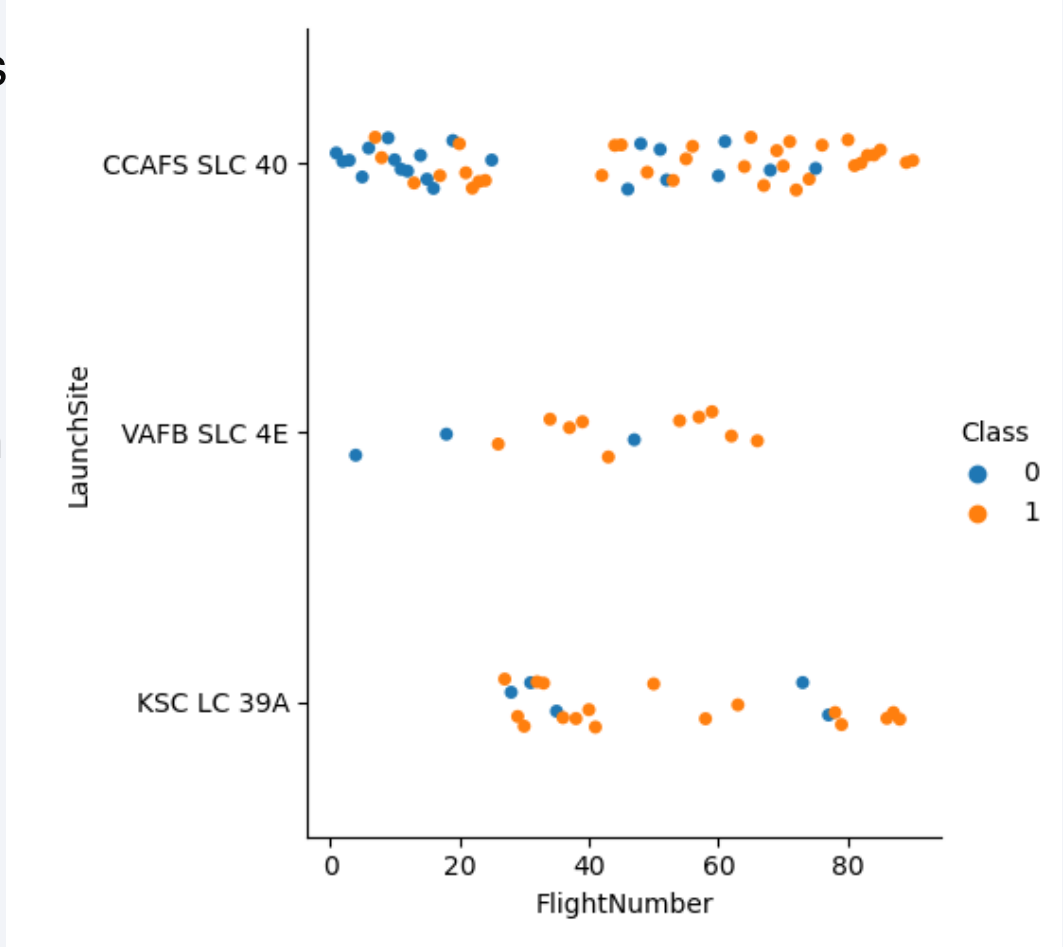
GitHub link:
Step 5: Predictive Analysis


Section 2
Insights drawn from EDA

EDA with Data Visualization

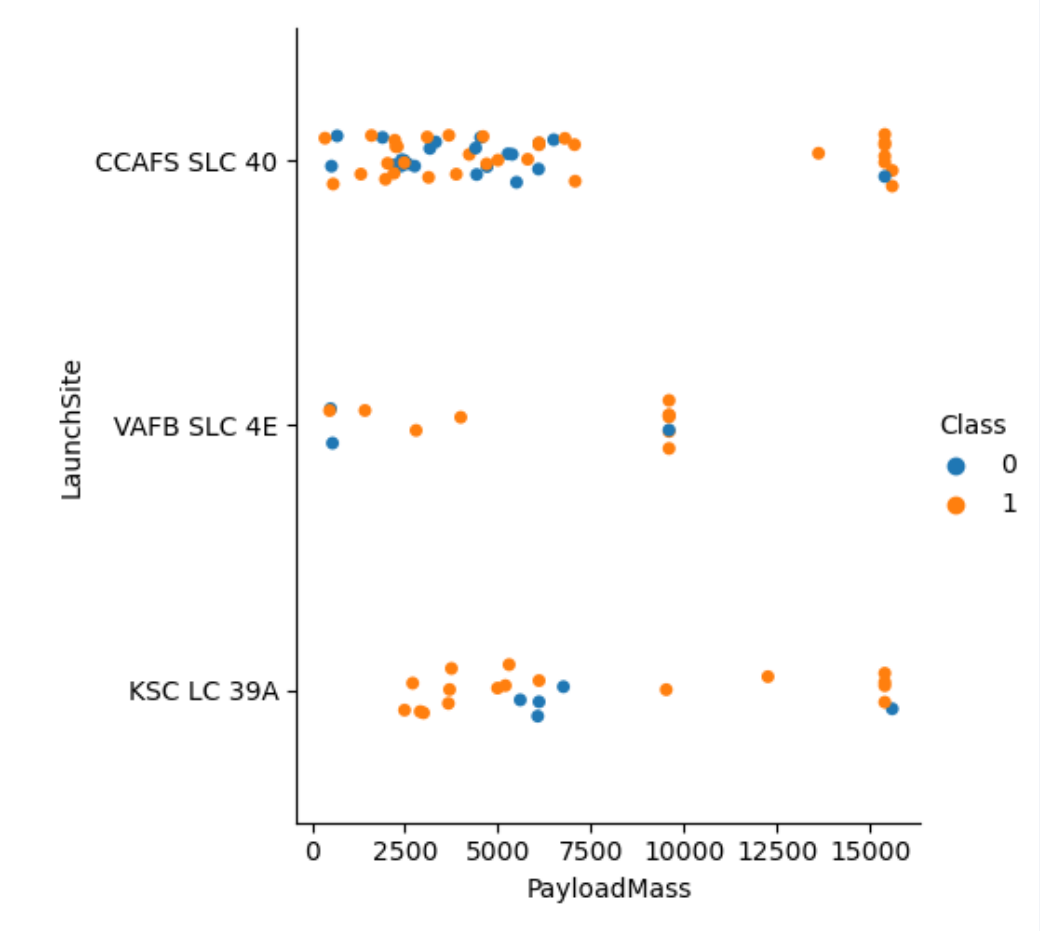
Graph 1: Flight Number vs Launch Site by Class

There seems to be a positive correlation between the number of flights conducted at a launch site and the success rate of launches at that site. In other words, an increase in the number of flights is associated with a higher likelihood of successful launches from the same site.



EDA with Data Visualization

Graph 2: Payload Mass vs Launch Site by Class



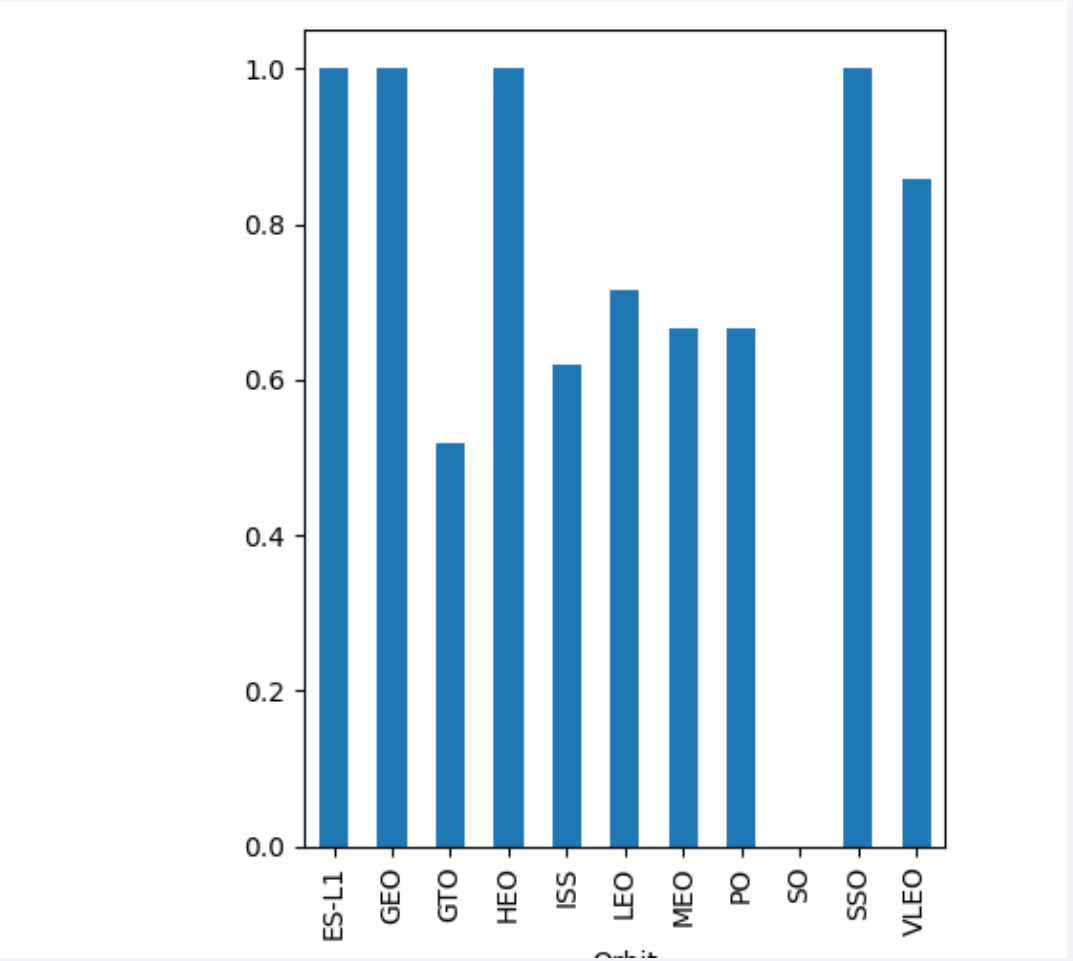
It appears that the success rate of a rocket launched from Launch Site CCAFS SLC 40 is positively correlated with the payload mass.

However, it is not possible to make a conclusive decision about the dependence of the launch site on the payload mass for a successful launch based solely on the provided visualization, as no clear pattern can be discerned.

EDA with Data Visualization

Graph 3: Success rate by Orbit Type

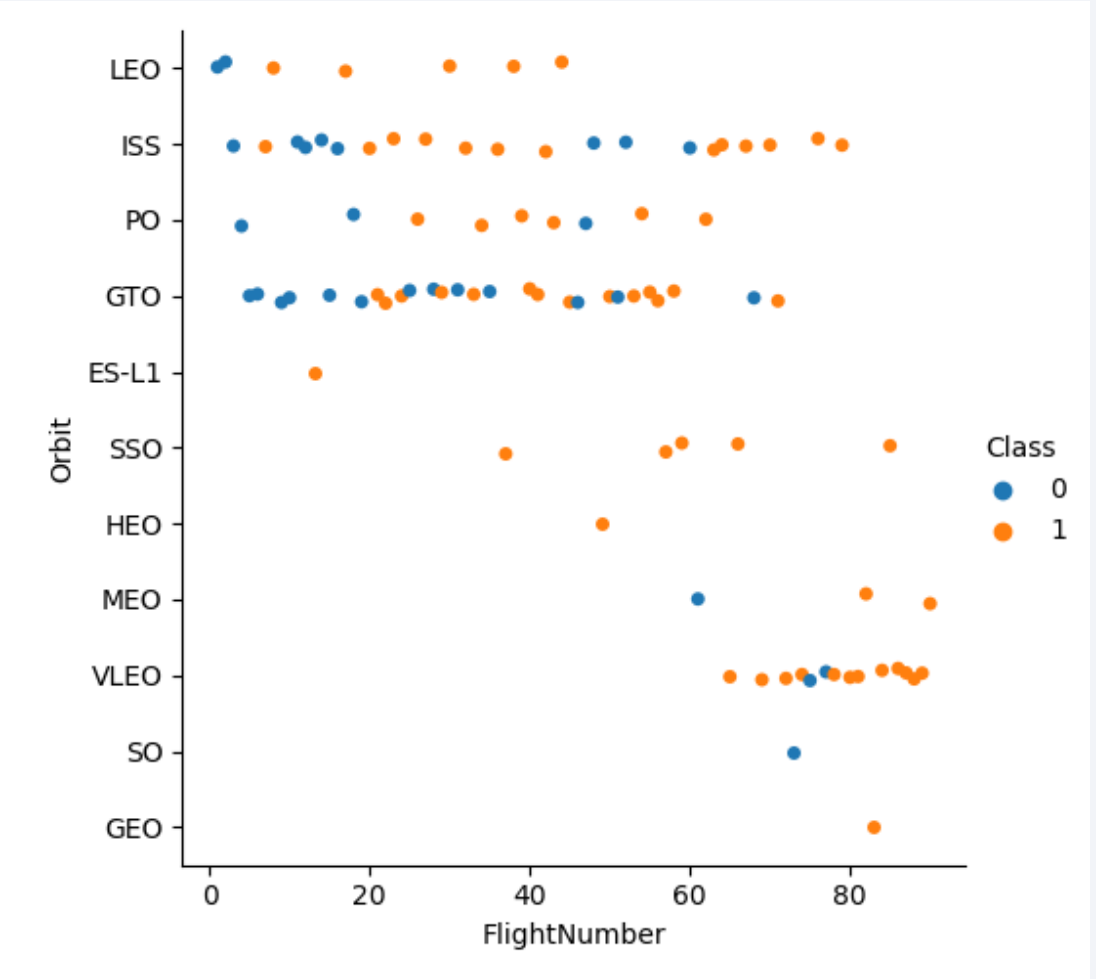
Orbits GEO,HEO,SSO,ES-L1 have the best Success Rates



EDA with Data Visualization

Graph 4: Flight number by Orbit Type

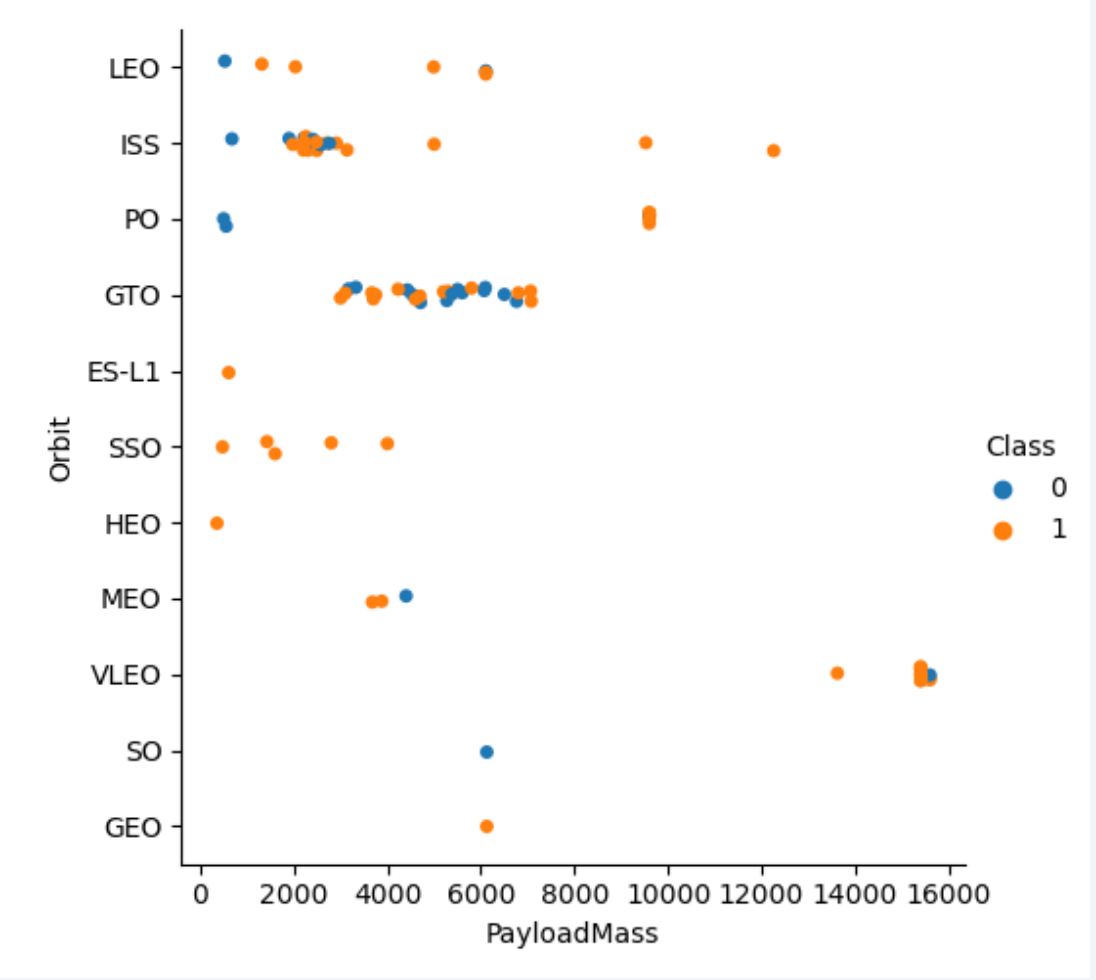
It can be observed that in the LEO orbit, the success rate of launches appears to be positively correlated with the number of flights conducted. However, in the case of the GTO orbit, no significant relationship between the number of flights and the success rate is apparent.



EDA with Data Visualization

Graph 5: Payload Mass by Orbit Type

It should be noted that heavy payloads tend to have a negative impact on launches in the GTO orbit, while they have a positive influence on launches in the GTO and Polar LEO (ISS) orbits.

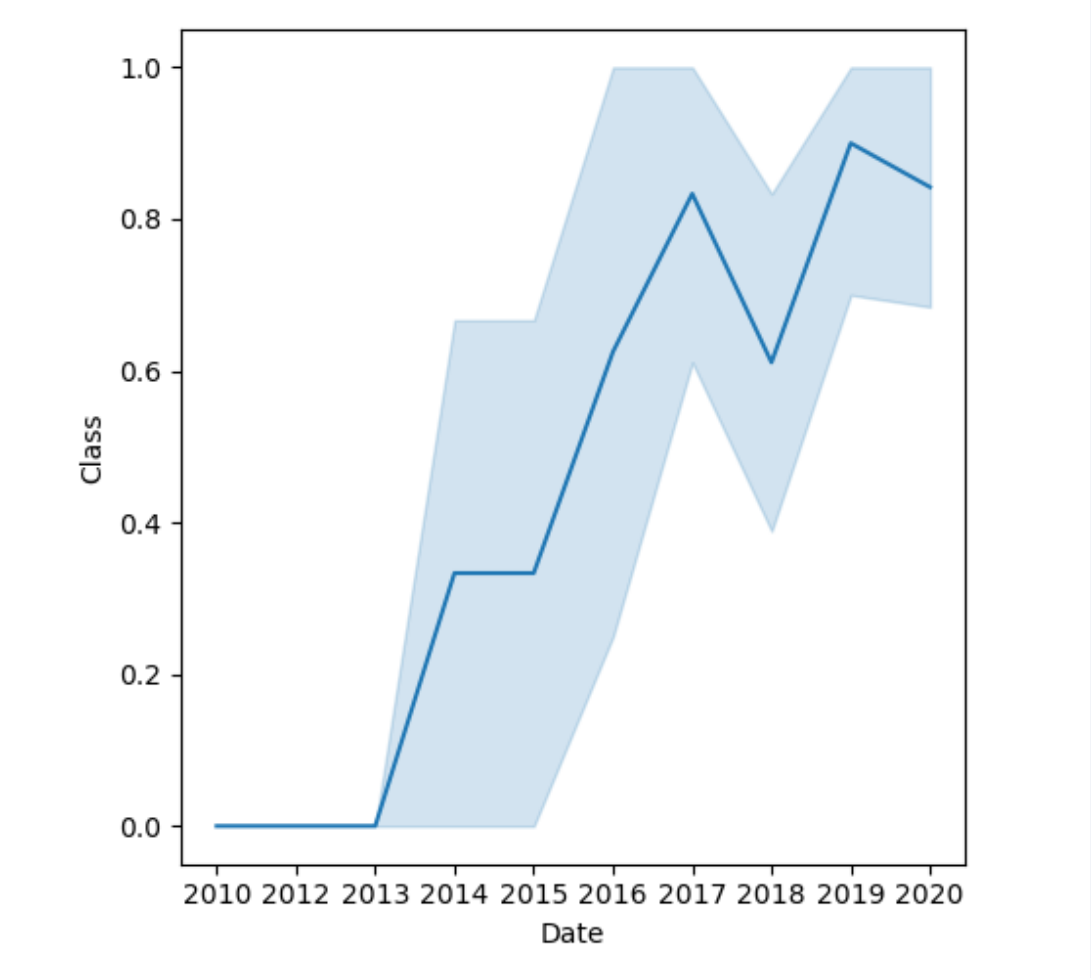


EDA with Data Visualization

Graph 6: Launch Success Rate Trend by Year

The graph “Launch Success Rate Trend by Year” shows the trend of launch success rate over the years. It plots the average success rate for each year from the data provided.

The x-axis represents the year, and the y-axis represents the average success rate. The graph shows that the launch success rate has been increasing since 2013, with some fluctuations in between.



EDA with SQL

- Unique launch sites in the space mission are displayed.
- Five records where launch sites begin with ‘KSC’ are displayed.
- The total payload mass carried by NASA (CRS) boosters is displayed.
- The average payload mass carried by booster version F9 v1.1 is displayed.
- The date of the successful drone ship landing outcome is listed.
- The names of boosters with successful ground pad landing and payload mass between 4000 and 6000 are listed.
- The total number of successful and failed mission outcomes is listed.
- The names of booster versions with maximum payload mass are listed.
- The records displaying month names, successful ground pad landing outcomes, booster versions, and launch sites for the year 2017 are listed.
- The count of successful landing outcomes between 2010-06-04 and 2017-03-20 is ranked in descending order.

Predictive Analysis (Classification)

- The model building process began by loading the dataset into NumPy and Pandas and transforming the data. The data was then split into training and test sets, and the number of test samples was checked. The type of machine learning algorithms to use was decided, and parameters and algorithms were set to GridSearchCV. The datasets were then fitted into the GridSearchCV objects and trained.
- In the model evaluation phase, the accuracy of each model was checked, and tuned hyperparameters were obtained for each type of algorithm. A confusion matrix was plotted to visualize the results.
- To improve the model's performance, feature engineering and algorithm tuning were performed. Feature engineering involves creating new features from existing ones or selecting relevant features to improve the model's performance. Algorithm tuning involves adjusting the model's hyperparameters to improve its accuracy.
- The best performing classification model was chosen based on its accuracy score. A dictionary of algorithms with their corresponding scores was provided at the bottom of the notebook for reference.

All Launch Site Names

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

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

[7] Python

```
 * 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'

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

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

Add Code Cell (⌘Enter)

[24] Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_O
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
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	
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

Total Payload Mass

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS 'total payload mass carried by boosters launched by NASA (CRS)
```

[9] Python

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

total payload mass carried by boosters launched by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

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

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

[10] Python

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%';
```

[47] Python

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

MIN(DATE)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

List 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 PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000 AND "Landin
```

[55] Python

```
 * 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

List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, count(*) FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

[66] Python

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

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

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

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

[68] Python

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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.

```
%sql SELECT substr(DATE,4,2) as Month_Number, "Landing _Outcome" as Landing_Outcome, BOOSTER_VERSION, LAUNCH_SITE \
FROM SPACEXTBL WHERE DATE LIKE '%2015';
```

Python

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

Month_Number	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

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

```
%sql SELECT substr(DATE,7,4) as Year_ , "Landing _Outcome", count(*) as Rank \
FROM SPACEXTBL \
WHERE "Landing _Outcome" LIKE 'Success%'\
GROUP BY substr(DATE,7,4) ORDER BY Rank DESC;
```

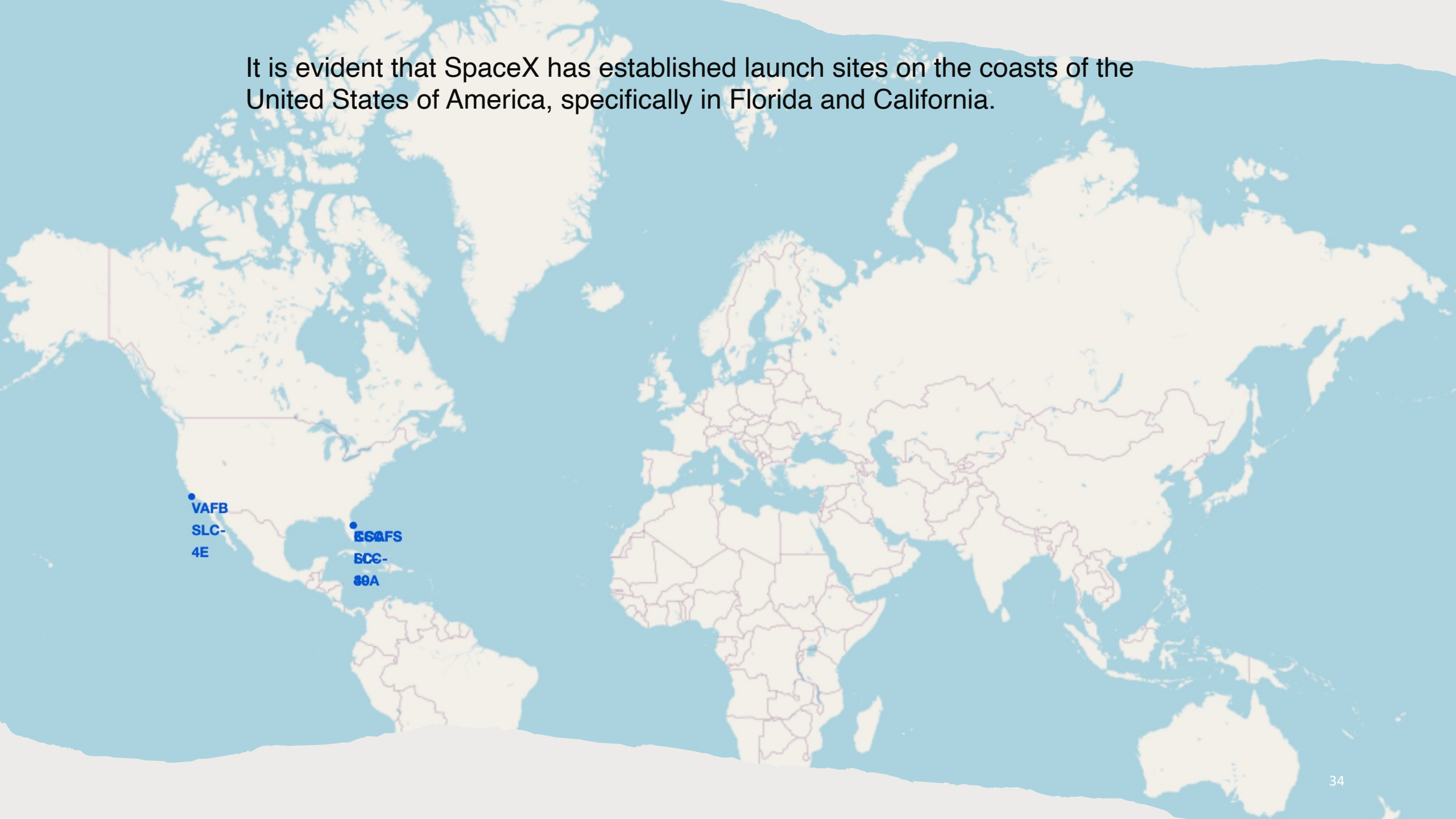
[135] Python

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

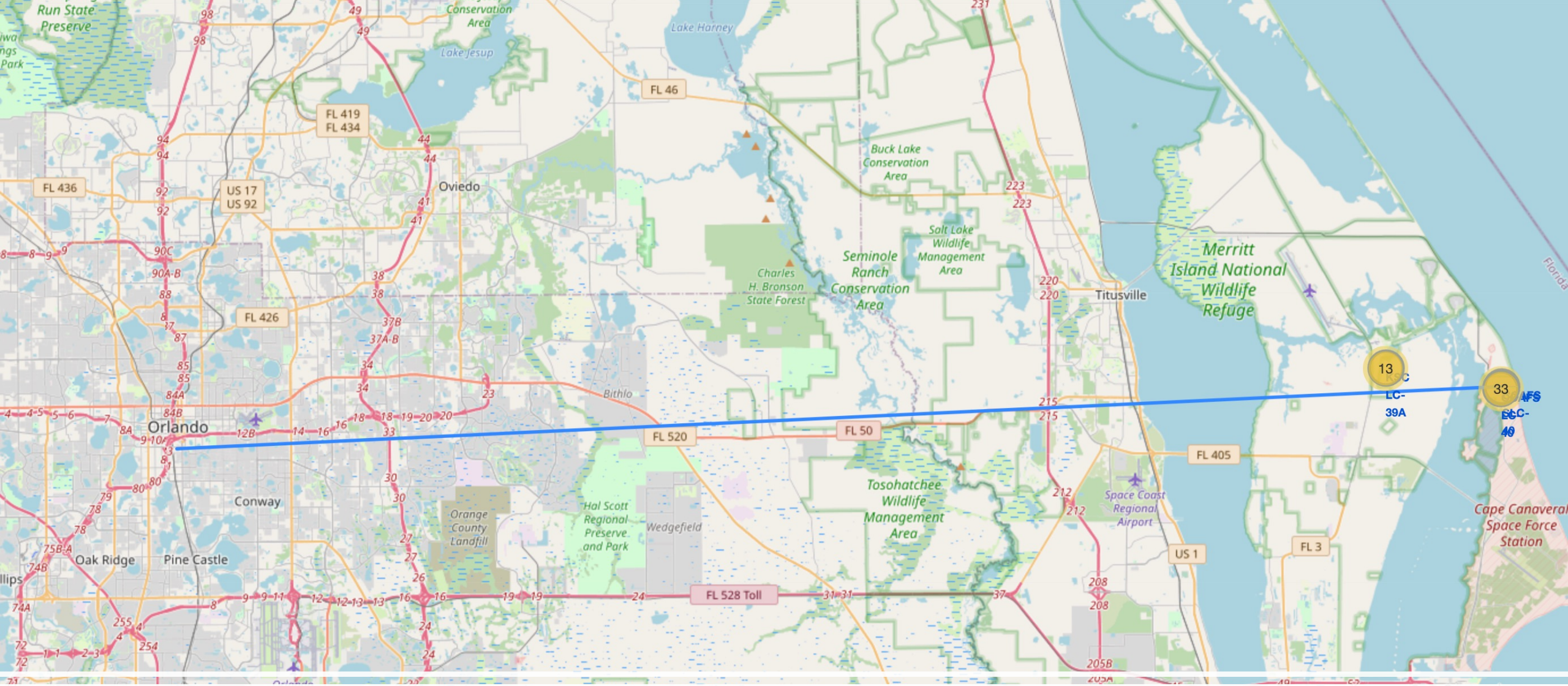
Year_	Landing _Outcome	Rank
2020	Success	21
2017	Success (drone ship)	14
2019	Success	10
2018	Success (ground pad)	10
2016	Success (drone ship)	5
2015	Success (ground pad)	1


Section 3
Launch Sites
Proximities Analysis

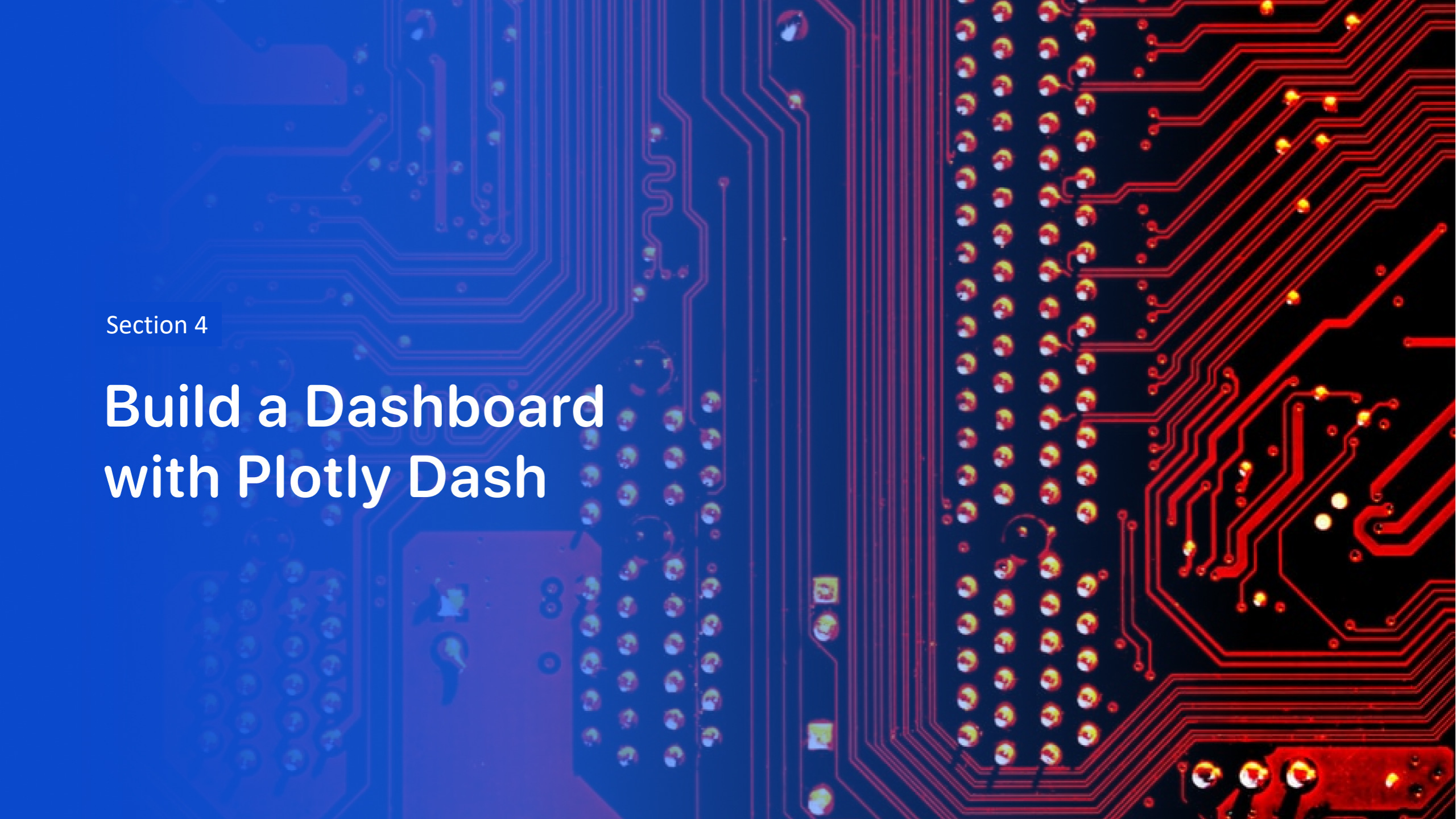
It is evident that SpaceX has established launch sites on the coasts of the United States of America, specifically in Florida and California.



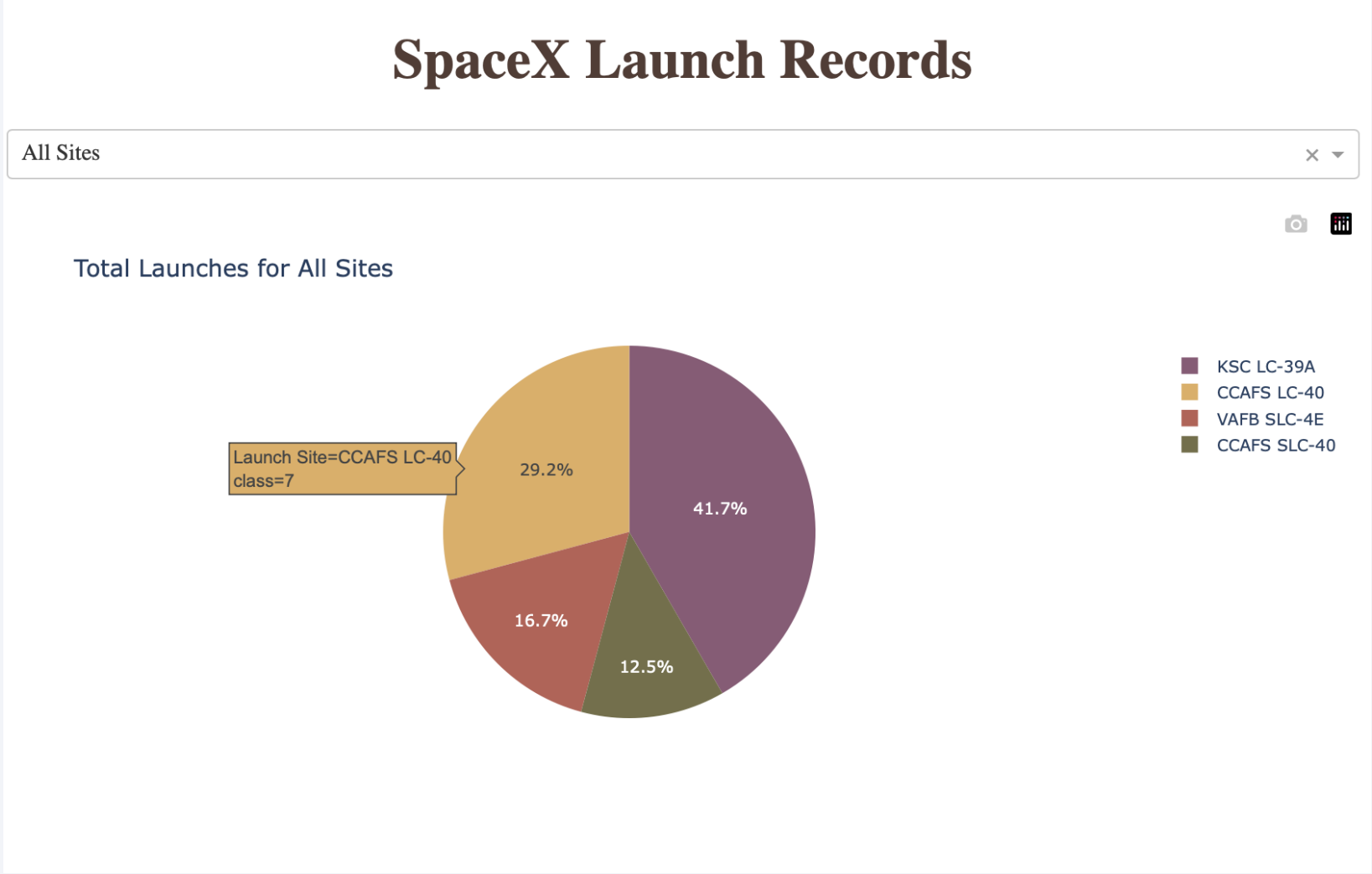
Successful launches are indicated by green markers, while red markers represent failures.



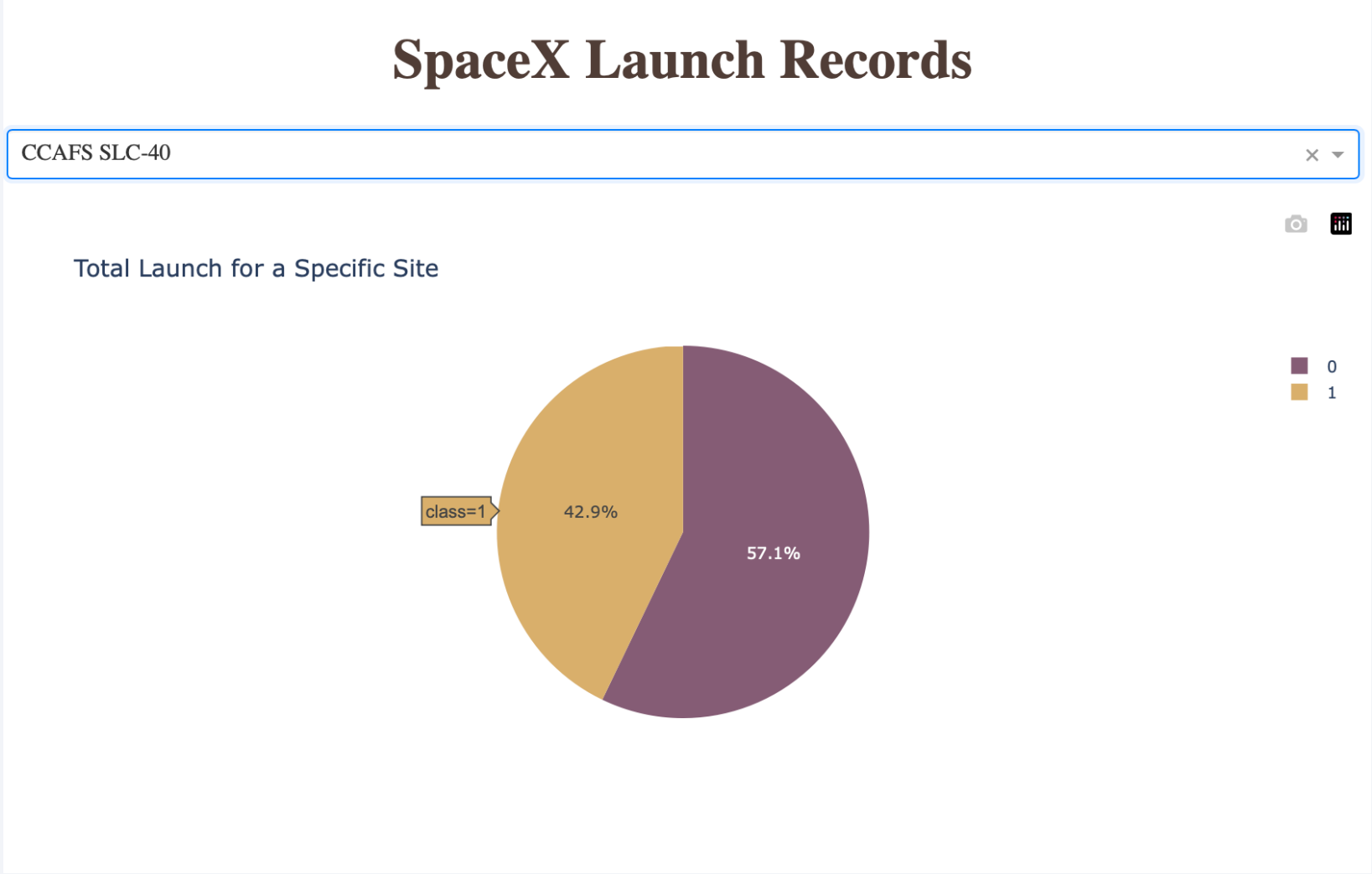
Distance to Closest City


Section 4
Build a Dashboard with Plotly Dash

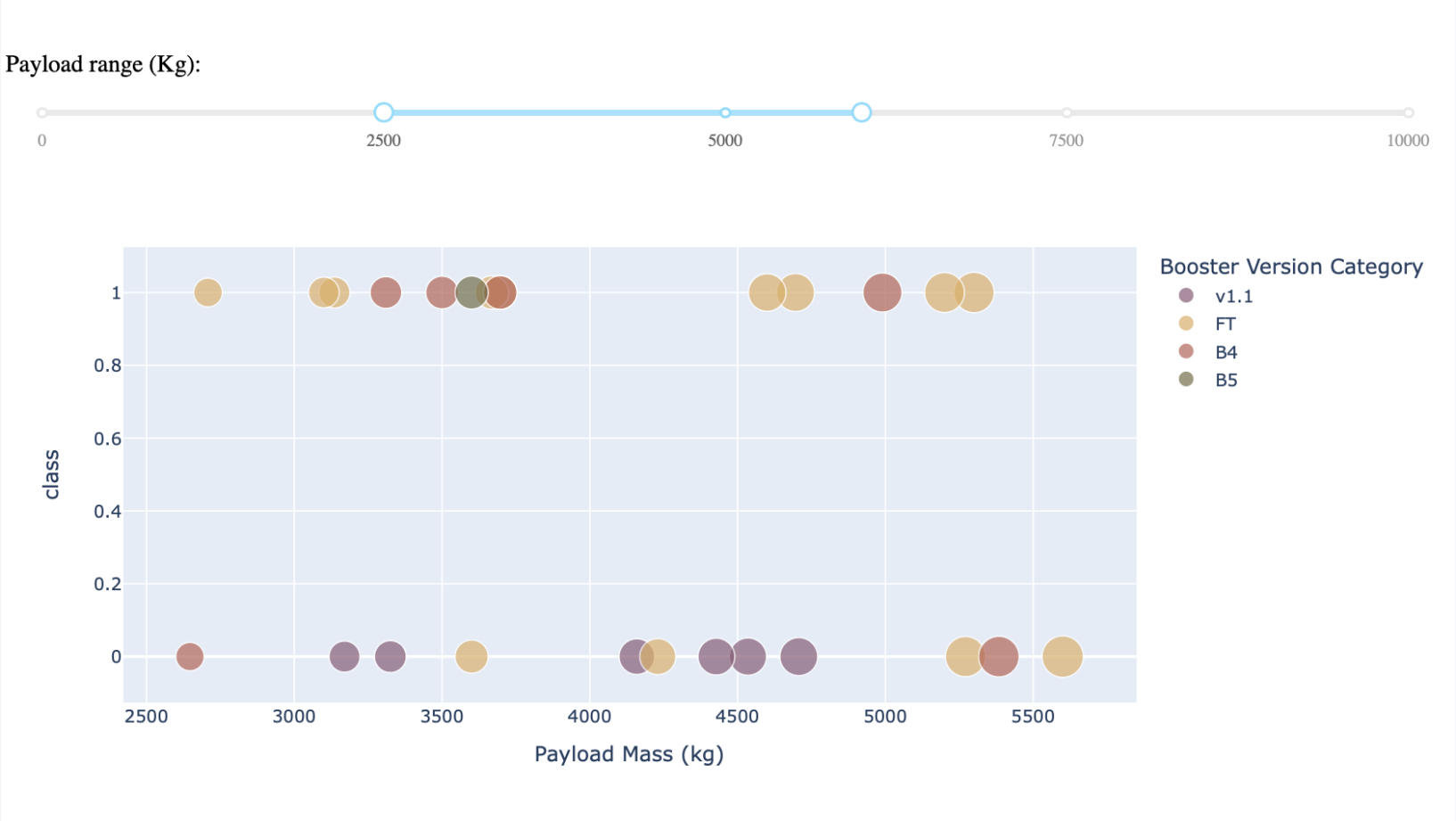
Launch success count for all sites



The launch site with highest launch success ratio



Payload vs. Launch Outcome scatter



Section 5

Predictive Analysis (Classification)

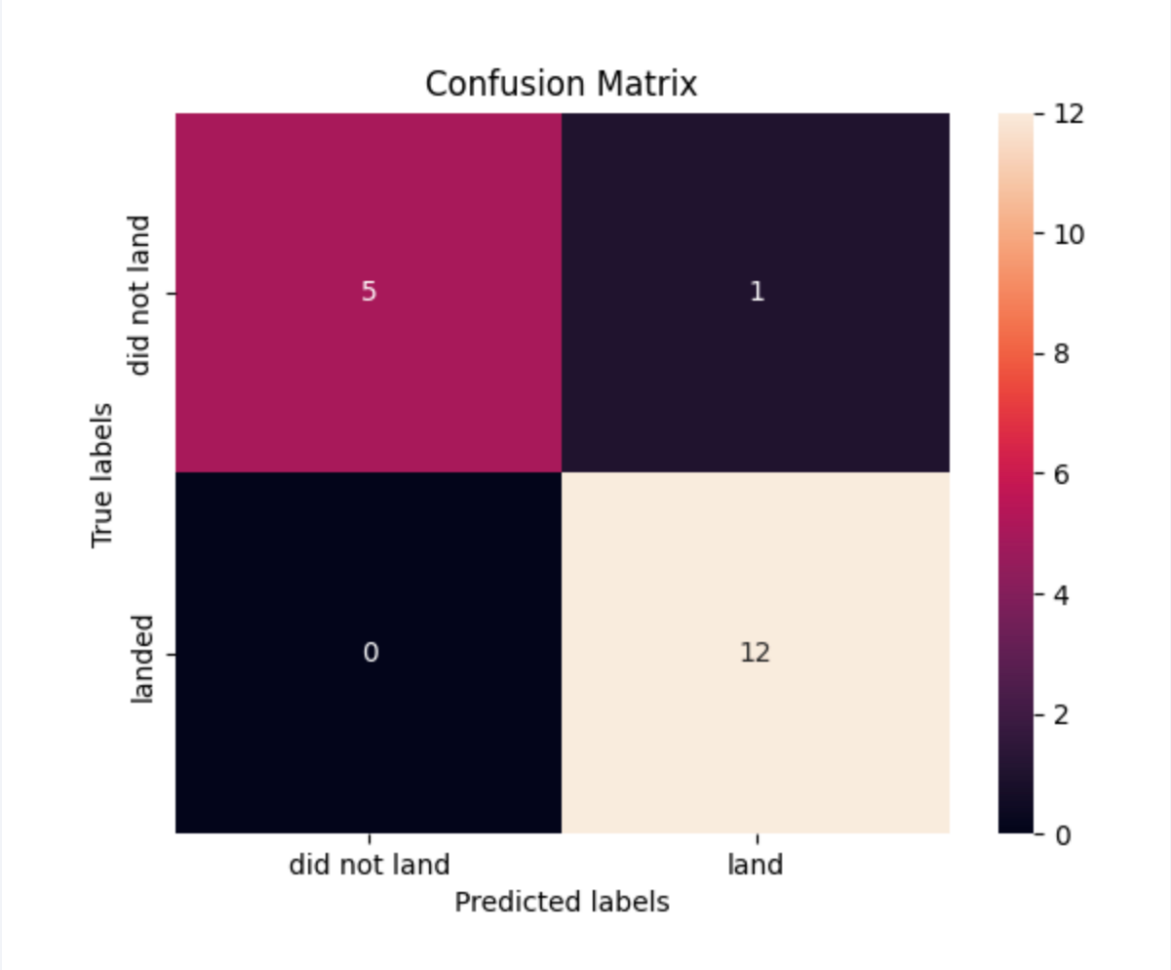
Classification Accuracy

It seems that after we identified the optimal hyperparameters for the all of the models and validated it with with the test data, we obtained an accuracy of 94.445% with SVM.

	Model	Score
0	Logistic Regression	0.944
1	SVM	0.944
2	Decision Tree	0.888
3	KNN	0.888

Confusion Matrix

- The given confusion matrix shows a binary classification problem with 5 positive and 13 negative instances. The model achieved an accuracy of 94.44%. It has a high precision of 80% but a lower recall of 80% for the positive class.



Conclusions

- The SVM algorithm is the optimal choice for this dataset.
- Orbits GEO, HEO, SSO, and ES-L1 exhibit the highest success rates, while KSC LC-39A has the most successful launches among all launch sites.
- Notably, lighter payloads demonstrate superior performance compared to heavier ones.
- Additionally, the success rate of SpaceX launches demonstrates a direct correlation with the time invested in perfecting the launches.

Thank you!