IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vishal Orsu
01st July 2024

Outline

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

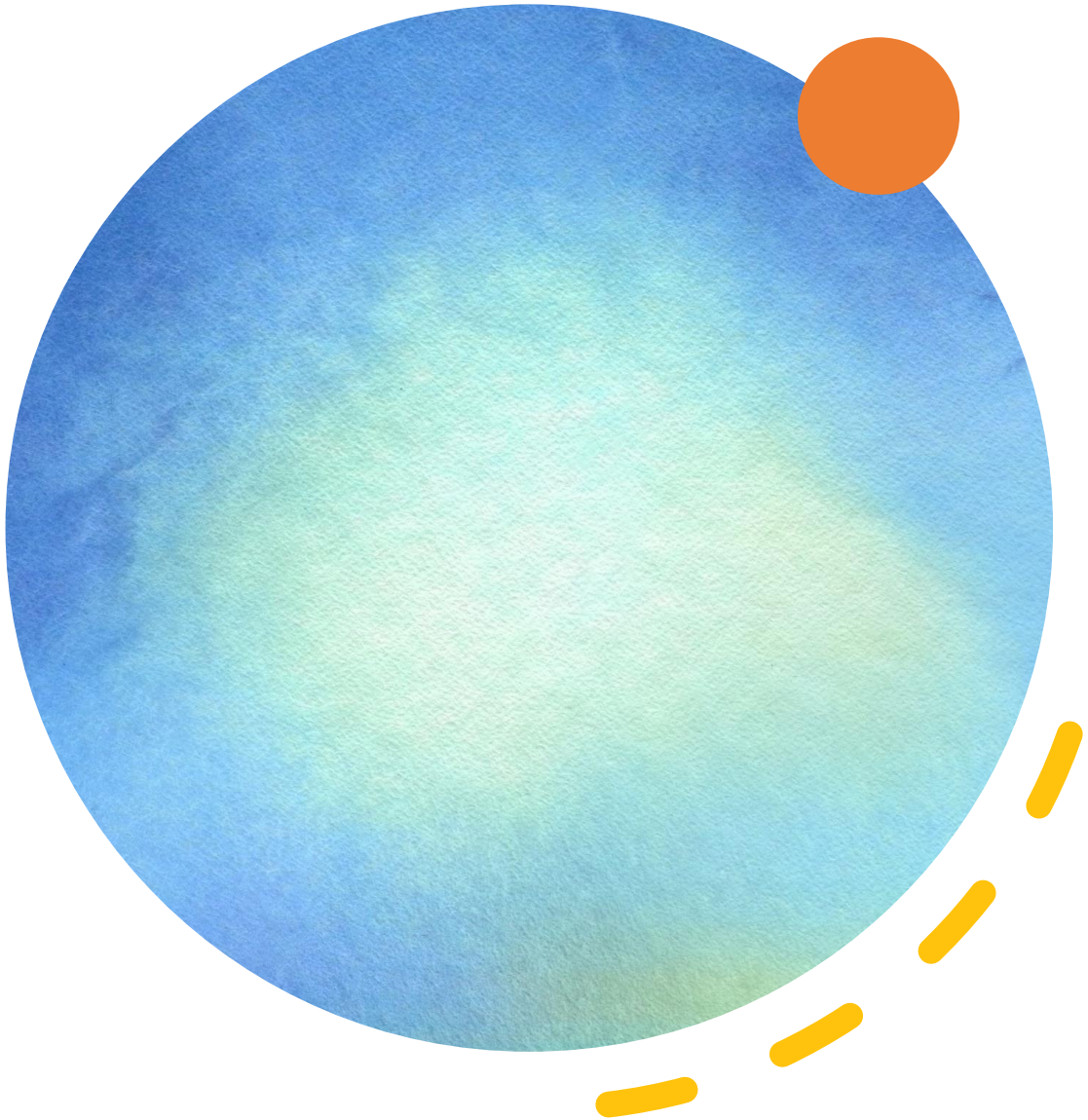
Executive Summary

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

Project background and context

- SpaceX offers Falcon 9 rocket launches at $62 million, whereas other providers charge over $165 million, largely due to SpaceX's ability to reuse the first stage. By predicting the first stage's landing success, we can estimate the launch cost. This project aims to develop a machine learning pipeline to forecast the successful landing of the first stage, aiding other companies in competitively bidding against SpaceX for rocket launches.

Problems you want to find answers

1. What factors contribute to the successful landing of a rocket?

2. What are the key elements that affect the success of rocket landings?

3. What operational requirements are essential for ensuring a successful rocket landing?

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- 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

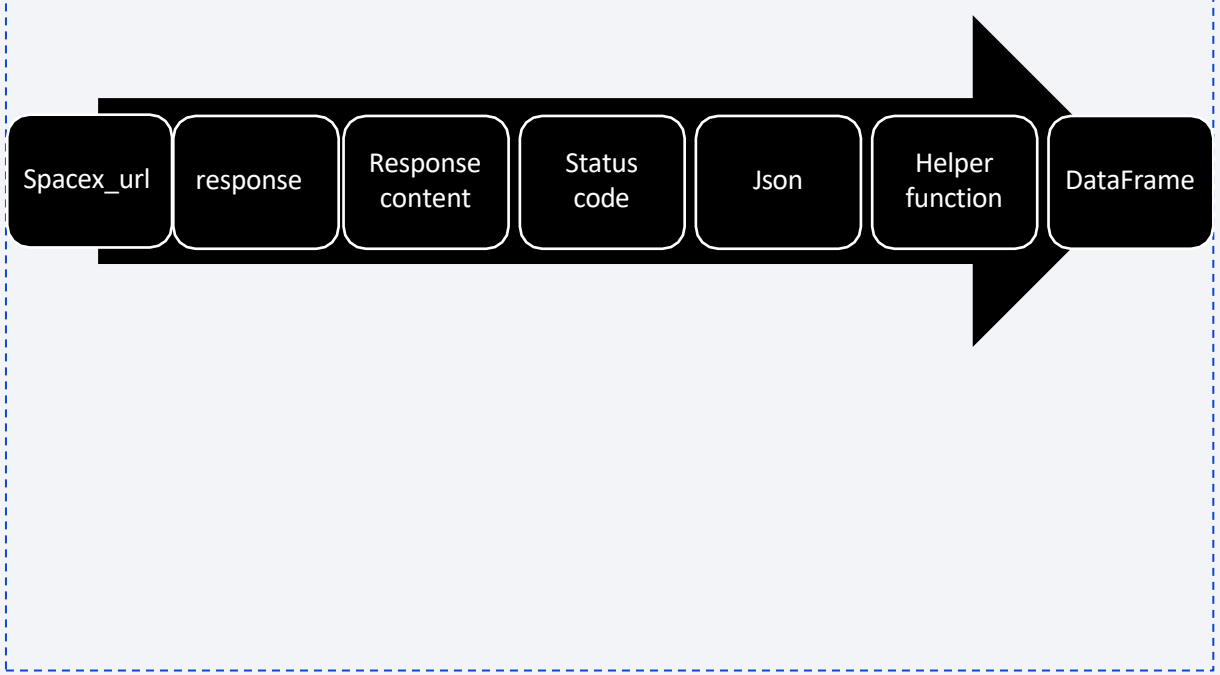
The data was collected using various methods

- Data was collected using GET requests to the SpaceX API.
- The response content was decoded as JSON using the .json() function and converted into a pandas DataFrame using .json_normalize().
- The data was cleaned, missing values were checked, and filled where necessary.
- Web scraping was performed from Wikipedia for Falcon 9 launch records using BeautifulSoup.
- The goal was to extract the launch records as an HTML table, parse the table, and convert it into a pandas DataFrame for future analysis.

Data Collection – SpaceX API

SpaceX URL , response, Json, Data Frame

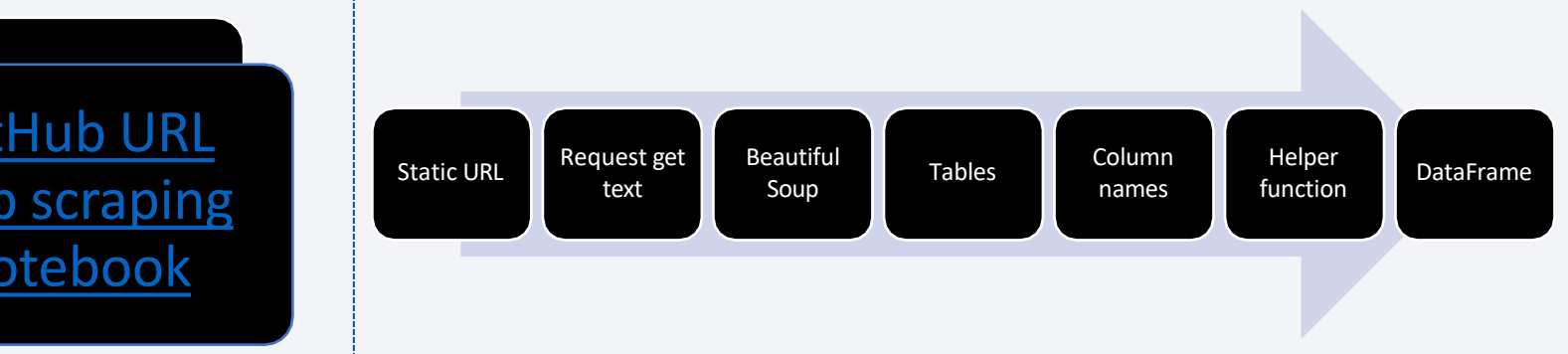
GitHub SpaceX API calls



Data Collection - Scraping

Static URL, Beautiful Soup, Data frame

GitHub URL web scraping notebook



Data Wrangling

- The cleaned dataset was initially imported, and the percentage of missing values in the LaunchingPad column was assessed, as it was the only column still containing missing values indicating instances when no LaunchingPad was used. Subsequently, the datatype of each column was reviewed, revealing four different types: int64, object, float64, and bool. Further analysis included examining the value counts of LaunchSite, with Cape Canaveral Space Launch Complex 40 and VAFB SLC 4E showing the highest counts at 55 each.

- A new feature named "class" was created from the outcomes column, where outcomes labeled as "False" or "None" were categorized as bad, assigned a value of zero (0), while good outcomes received a value of one (1). The success rate of all good outcomes was then calculated, comprising 66.67% of the class feature.

GitHub URL data wrangling related notebooks

EDA with Data Visualization

- Seaborn and Matplotlib are extensively utilized for visualizing static charts, including catplot for displaying the relationship between PayloadMass and LaunchSite, barplot for orbit versus class, and line plot to show the trend of the average success rate of rocket launches over time (2010-2020).
- The analysis involved several key visualizations. A catplot was used to display the relationship between FlightNumber and LaunchSite, highlighting patterns in launch site usage over time. A scatter plot of PayloadMass versus LaunchSite was created to examine payload distribution across different sites. Bar plots and line plots were also used to visualize success rates by orbit type and over time, respectively, providing insights into performance trends and the impact of various factors on launch success.
- GitHub URL EDA with data visualization notebook

EDA with SQL

- Displayed unique launch sites and retrieved records with launch sites starting with 'CCA'.
- Calculated total and average payload masses for specific customer and booster versions.
- Listed the date of the first successful ground pad landing and booster versions with specific landing outcomes and payload ranges.
- Counted successful and failed mission outcomes and identified boosters with maximum payload mass.
- GitHub URL EDA with SQL notebook

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added markers with circles, popup labels, and text labels for all launch sites, including NASA Johnson Space Center, using their latitude and longitude coordinates to display their geographical locations and proximity to the equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- Incorporated colored markers for launch outcomes at each site, using green for success and red for failure, with Marker Cluster to highlight sites with higher success rates.

Distances between a Launch Site to its proximities:

- Visualized distances from the KSC LC-39A launch site to nearby proximities, such as railway, highway, coastline, and the closest city, using colored lines.

GitHub URL: Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

Dashboard Summary

- **Dropdown Menu**: Select launch sites or view all for data filtering.
- **Pie Chart**: Displays total successful launches by site or overall.
- **Payload Range Slider**: Filter data by payload mass range.
- **Scatter Plot**: Correlates payload mass with launch success and booster versions.

Purpose

- **Enhanced Interactivity**: Facilitates site-specific analysis and comparison.
- **Visual Summary**: Quickly assess success rates and trends.
- **Data Exploration**: Analyze payload impacts on launch outcomes.
- **Insight Generation**: Identify patterns in launch success and booster performance.

These features make the dashboard a comprehensive tool for exploring SpaceX launch data efficiently and intuitively.

GitHub URL Plotly Dash lab

Predictive Analysis (Classification)

Model Development Process Summary:

Data Preparation: Loaded and standardized dataset using StandardScaler.

Model Selection: Compared Logistic Regression, SVM, Decision Tree, and KNN.Training and Tuning: Used GridSearchCV for hyperparameter tuning (10-fold CV).

Evaluation: Assessed models on test set for accuracy and F1-score, with confusion matrices for insight.

Best Model: Decision Tree achieved highest accuracy (0.889) and F1-score (0.889).

Recommendation: Decision Tree is recommended for balanced performance; consider ensemble methods for further enhancement.

GitHub URL of predictive analysis lab

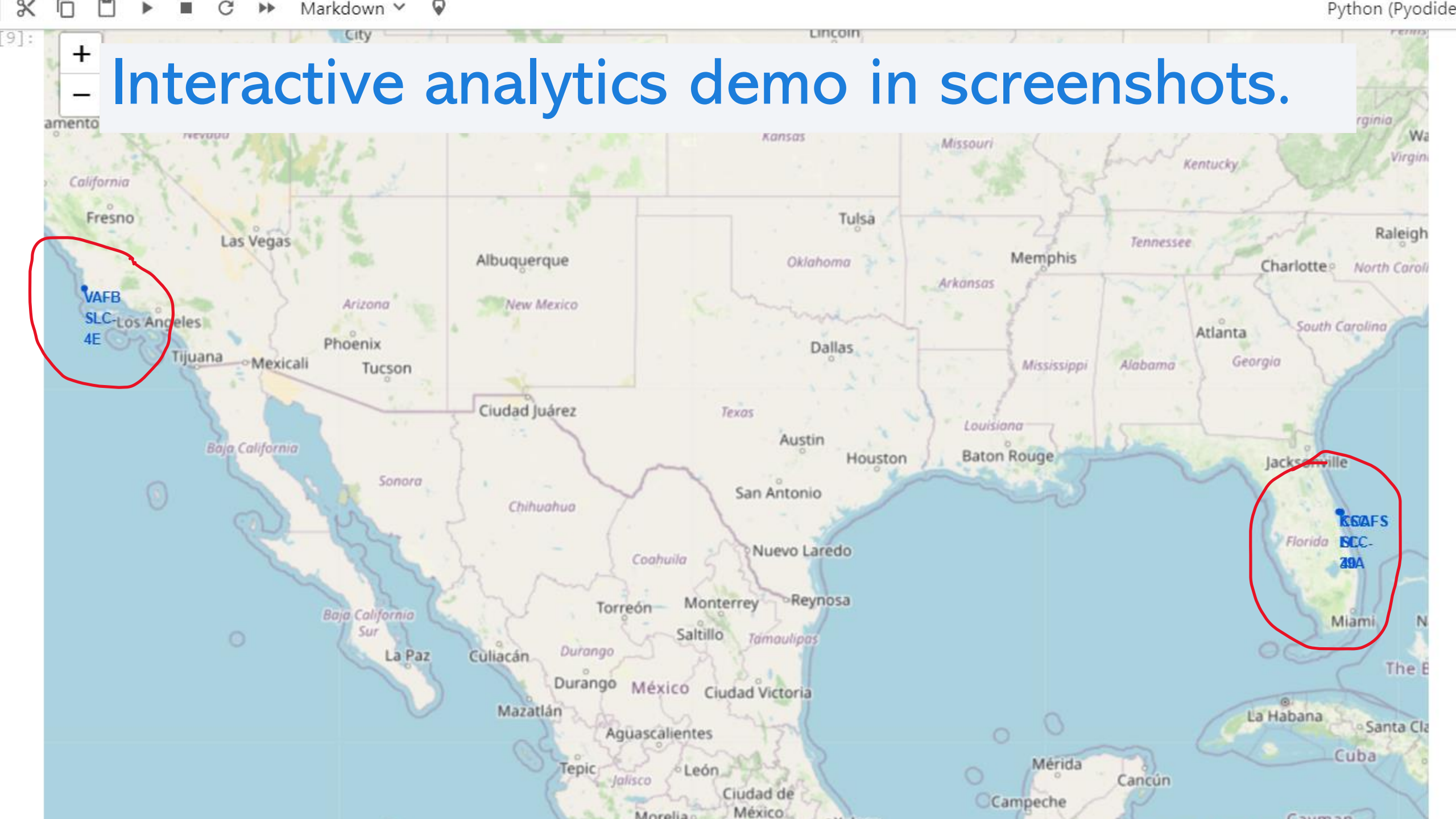
Results

Exploratory Findings and Visual Insights:

During the exploratory analysis, notable findings emerged: NASA's Commercial Resupply Services (CRS) missions handled a total payload mass of 45,596, with the booster version F9 v1.1 averaging 2928.4 in payload mass. The first successful ground pad landing occurred on December 22, 2015. Up to five booster versions successfully landed on drone ships, with payload masses ranging between 4000 and 6000. SpaceX's Falcon 9 achieved a total of 99 successful mission outcomes.

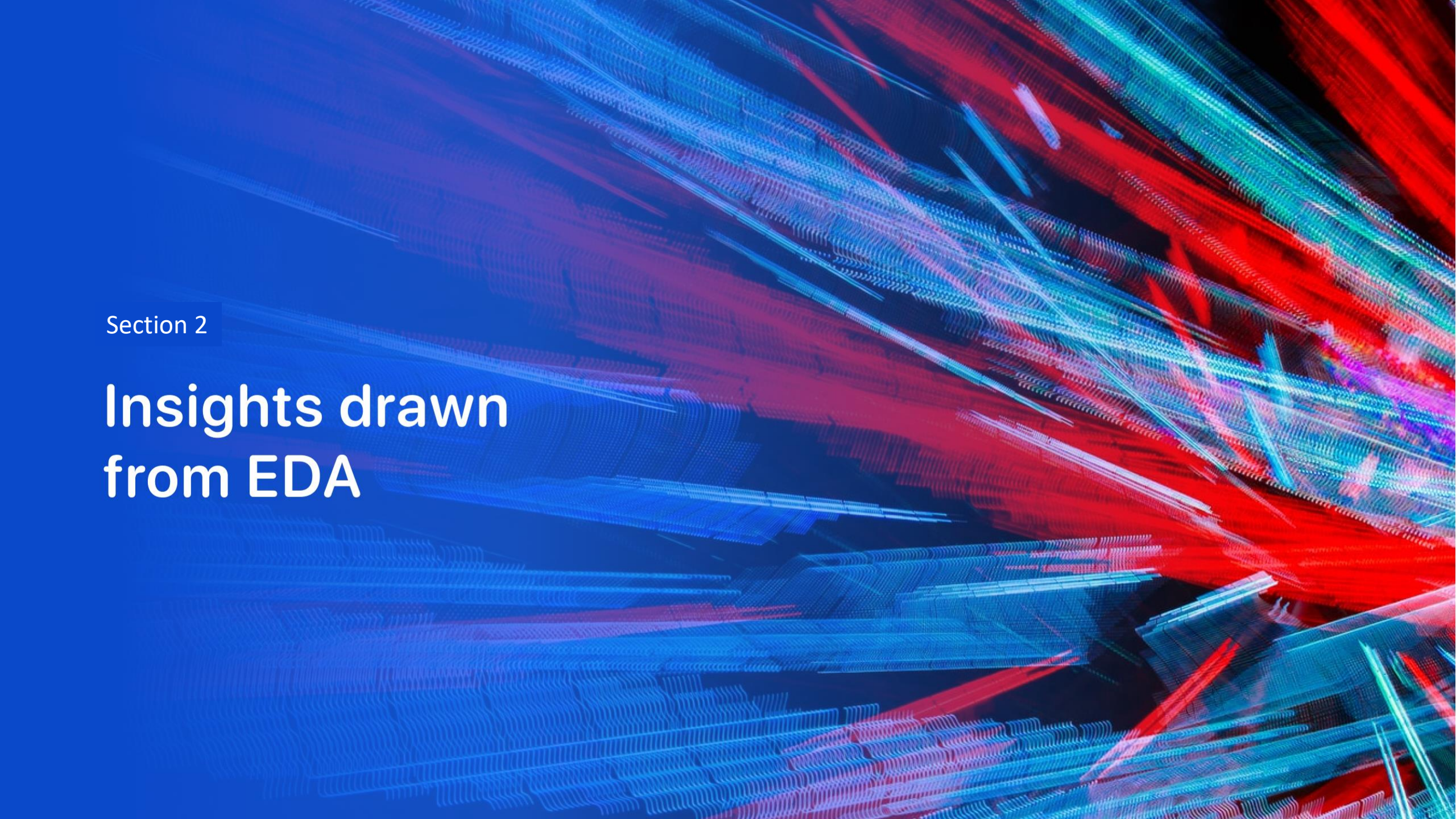
Visualization of the data revealed compelling insights: CCAFS SLC 40 showed an increasing success rate as flight numbers rose. Falcon 9 demonstrated a perfect landing success rate (100%) for orbital types ES-L1, SSO, HEO, and GEO. Over the decade from 2010 to 2020, Falcon 9's first-stage landings displayed significant improvement, reflecting substantial advancements over time.

Interactive analytics demo in screenshots.

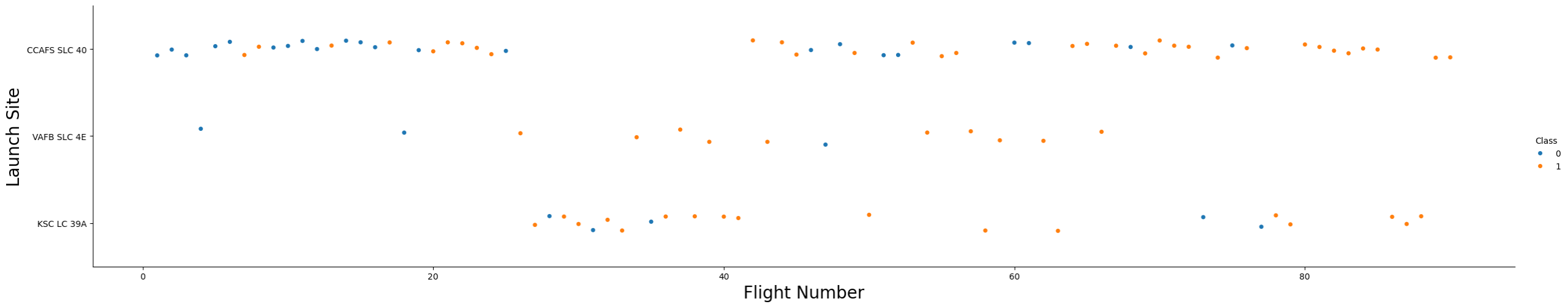


Predictive analysis results.

- Four distinct machine learning algorithms were utilized: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K-Nearest Neighbors. All models exhibited strong performance, achieving accuracies above 83%. Notably, after optimization through GridSearchCV, the Decision Tree Classifier emerged as the top performer with an accuracy of 88.89% and an F1-score of 88.21%. These findings instill confidence in predicting Falcon 9 first-stage landing success with 88.89% accuracy.

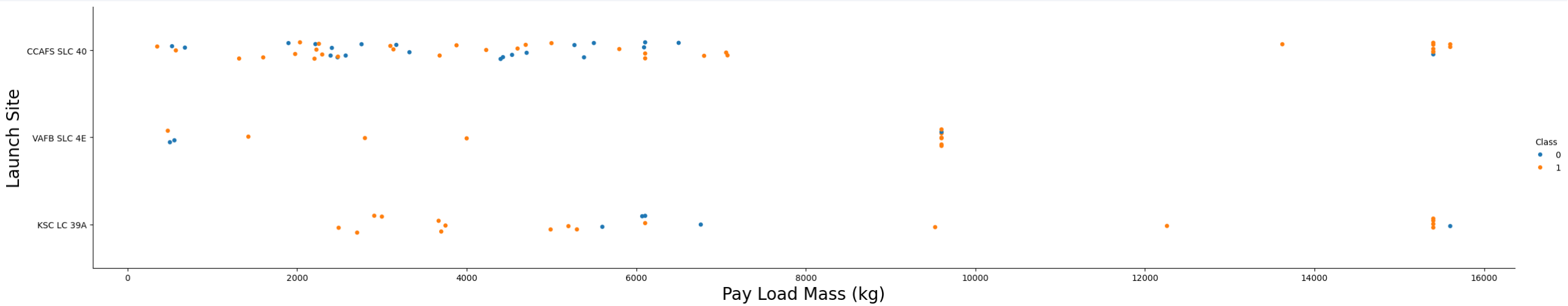

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



In examining the relationship between Flight Number and LaunchSite, it's evident that lower flight numbers (up to 20) show no recorded success or failure metrics for launches from "KSC LC 39A," while two failures are noted for launches from "VAFB SLC 4E." However, as flight numbers exceed 80, there is a notable increase in successful launches from "CCAFS SLC 40," whereas no successes are recorded for launches from "VAFB SLC 4E.

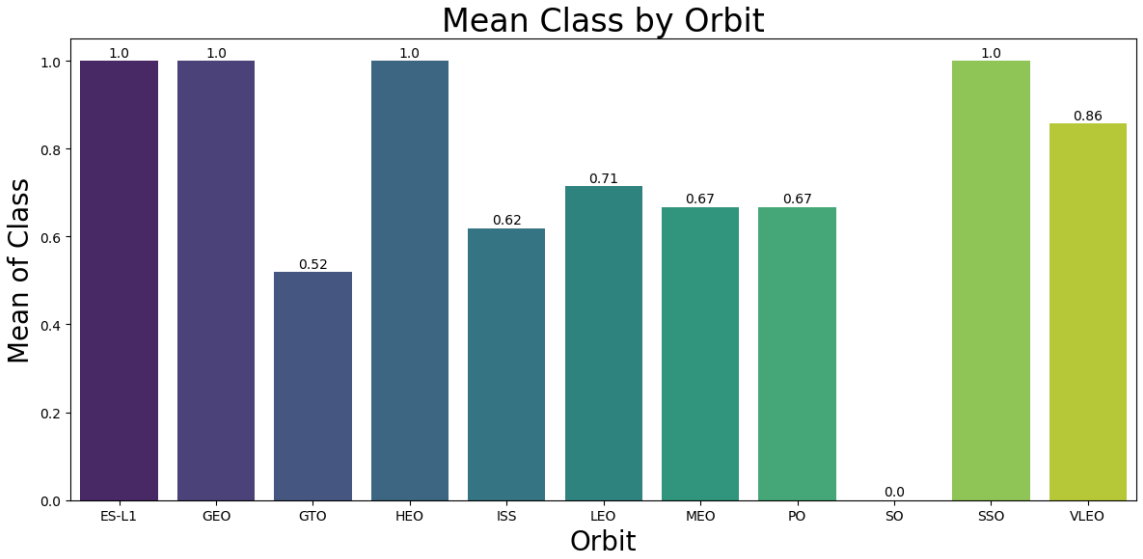
Payload vs. Launch Site



Analyzing the relationship between Flight Number and LaunchSite reveals that flights with lower numbers (up to 20) show no recorded success or failure metrics for launches from "KSC LC 39A," while "VAFB SLC 4E" encountered two failures. However, as flight numbers exceed 80, "CCAFS SLC 40" shows a higher number of successful launches, contrasting with no recorded successes for "VAFB SLC 4E."

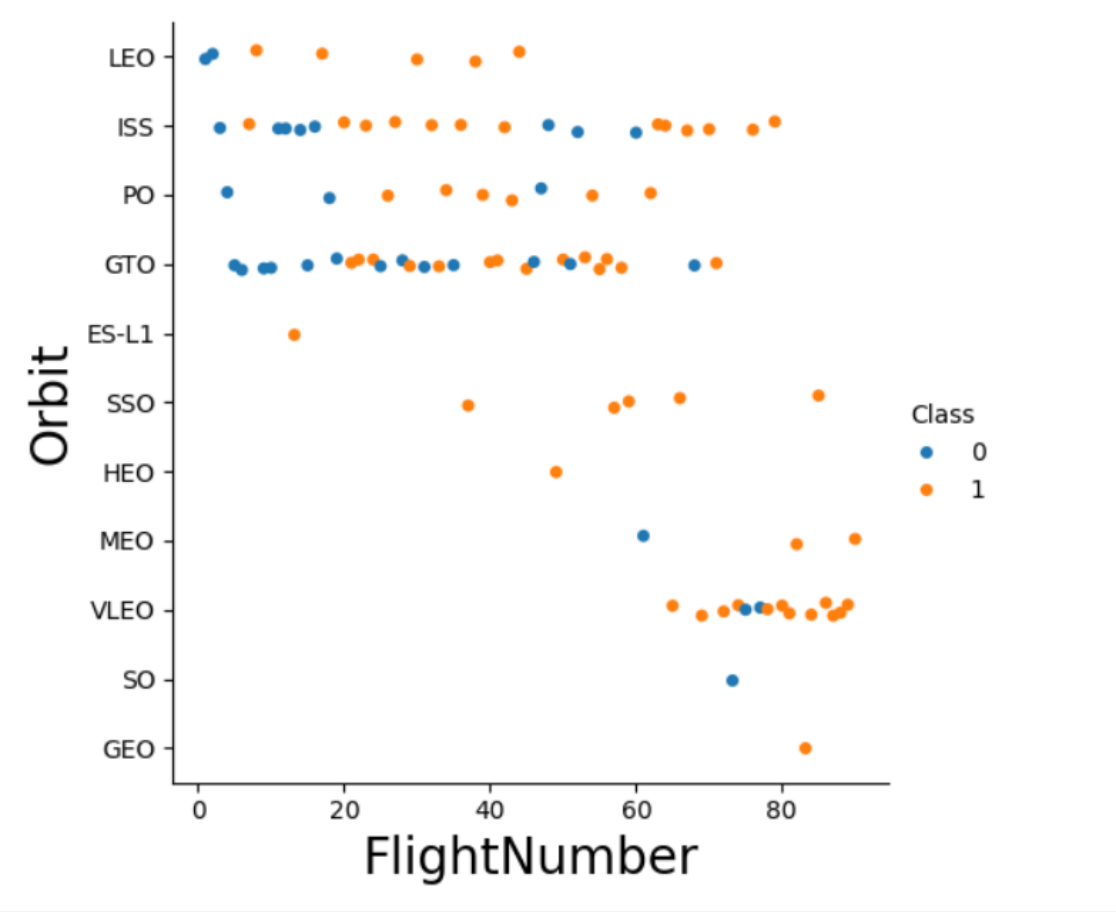
Success Rate vs. Orbit Type

- The highest success rates are seen in ES-L1, GEO, and SSO orbits, indicating these are well-established and reliable orbits for missions.
- The GTO orbit shows a lower success rate, possibly due to the complexities involved in transferring to GEO.
- The PO orbit shows a 0% success rate, indicating that missions targeting this orbit may face significant challenges or that there is limited data.
- VLEO, despite its advantages, shows a slightly lower success rate than the top orbits, potentially due to atmospheric drag and other operational difficulties at lower altitudes.



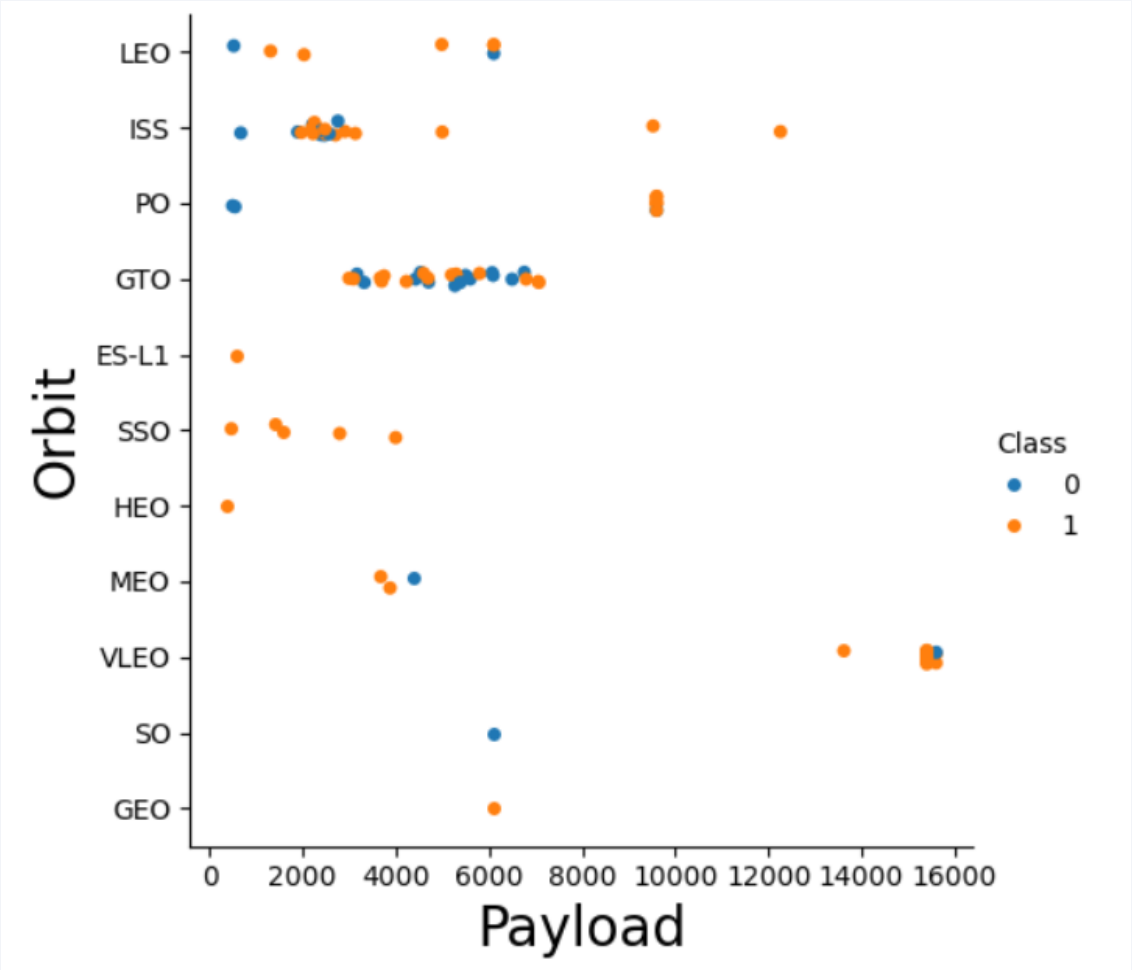
Flight Number vs. Orbit Type

This plot provides a visual representation of the success and failure rates of various orbits across different flight numbers. It helps identify trends and reliability for each orbit type, with LEO, ISS, and GEO showing higher success rates, while PO and GTO have more failures.



Payload vs. Orbit Type

- Payloads in LEO, ISS, and PO orbits are generally lower, indicating that these orbits typically have smaller payloads.
- There is a noticeable clustering of data points in the GTO orbit, suggesting many payloads are sent to GTO.Higher payloads are seen in GEO, suggesting these missions tend to carry heavier payloads.
- Some orbits, like ES-L1, HEO, MEO, VLEO, and SO, have fewer data points, indicating fewer missions or data points available for these orbits.
- The color distribution between blue and orange dots can indicate whether there is any preference or difference in payload distribution between Class 0 and Class 1 across different orbits.

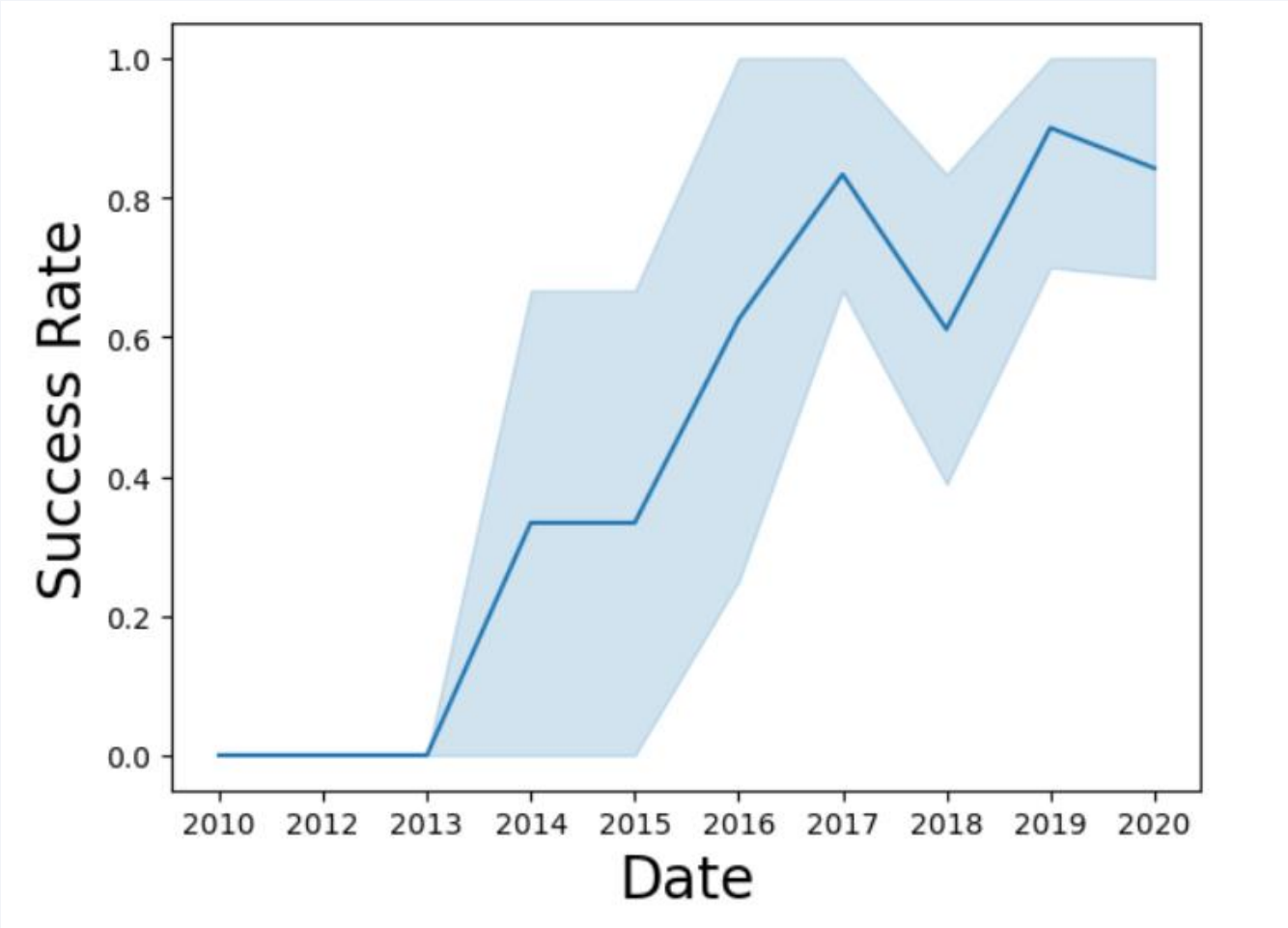


Launch Success Yearly Trend

This plot shows the success rate over time from 2010 to 2020. The key points are:

- The success rate started at 0% and began to increase around 2013.
- There was a significant increase in success rate between 2013 and 2016, reaching about 60%.
- After 2016, the success rate continued to rise, peaking near 90% in 2018.
- The shaded area around the line indicates the confidence interval or variability in the success rate over time.

Overall, the success rate has generally increased over the period from 2010 to 2020.



All Launch Site Names

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

[9]:
```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

[9]:

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

Launch Site Names Begin with 'CCA'

Task 2

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

[10]:
```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

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

[10]:

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

Would you like to receive official Jupyter

Total Payload Mass

- The total payload carried by boosters from NASA

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

```
[11]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

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

```
[12]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

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

[12]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad

Task 5

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

Hint:Use min function

[13]:
```
%sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

[13]:

min(DATE)
2015-12-22

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

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

```
[14]: %sql select Booster_Version from SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 ar
```

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

[14]:

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

Total Number of Successful and Failure Mission Outcomes

- The query counts the total number of mission outcomes that were either successful or failed in flight from a table named SPACEXTBL. The result shows that there are 99 such missions combined (both successful and failed in flight).

List the total number of successful and failure mission outcomes

```
[15]: %sql select count(Mission_Outcome) from SPACEXTBL WHERE Mission_Outcome = 'Success' or Mission_Outcome = 'Failure (in flight)'
```

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

[15]:

count(Mission_Outcome)
99

Boosters Carried Maximum Payload

- List of the names of the booster which have carried the maximum payload mass

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

```
[16]: %sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

[16]:

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 of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
[17]: %sql SELECT SUBSTR(Date,6,2) AS Month, Booster_Ve
```

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

[17]:

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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 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.

[19]:
```
%sql SELECT Landing_Outcome, COUNT(*) AS Numbers FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success%' AND Date BETWEEN '2010-06-04' AND '2017
```

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

[19]:

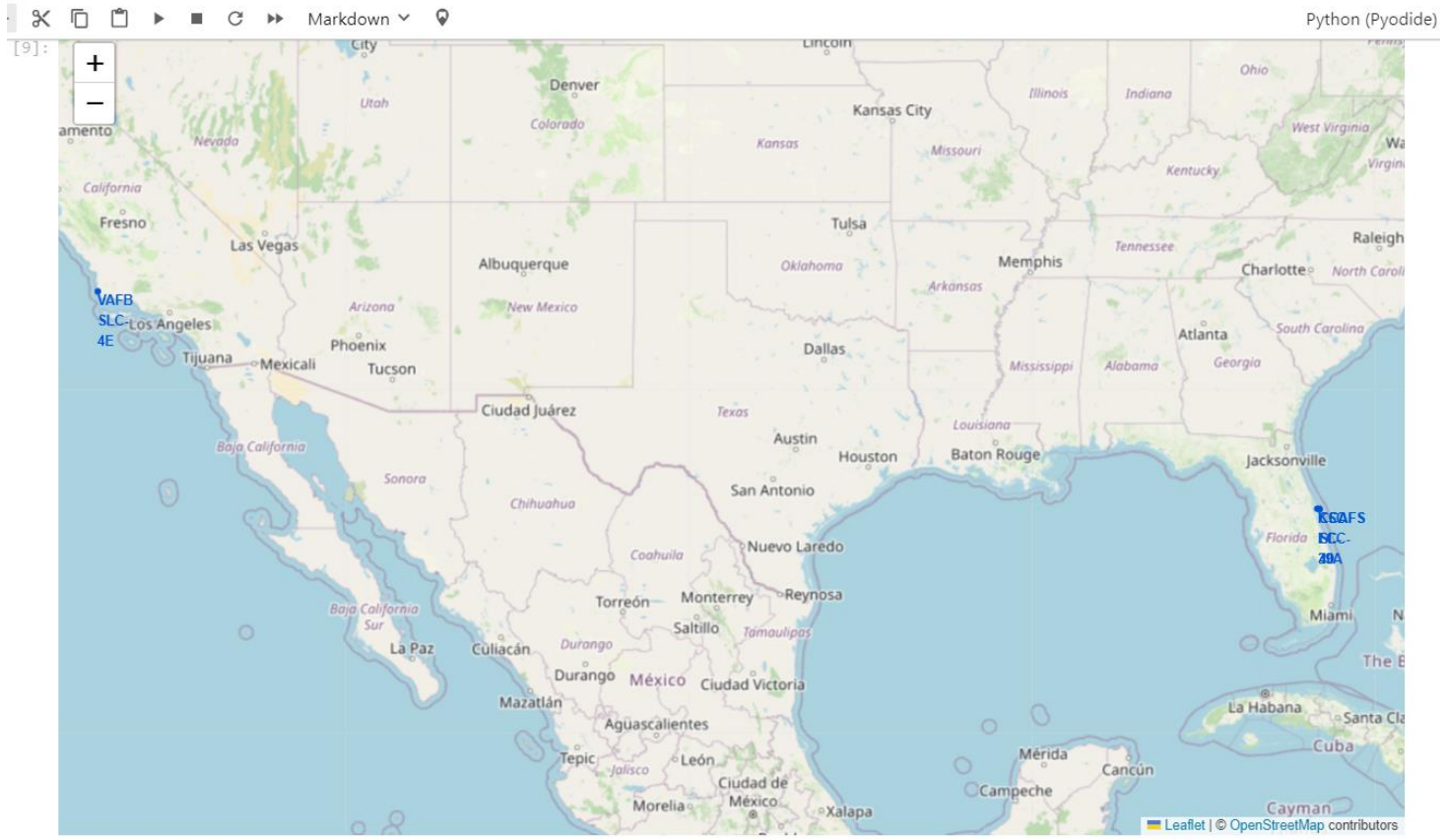
Landing_Outcome	Numbers
Success (drone ship)	5
Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

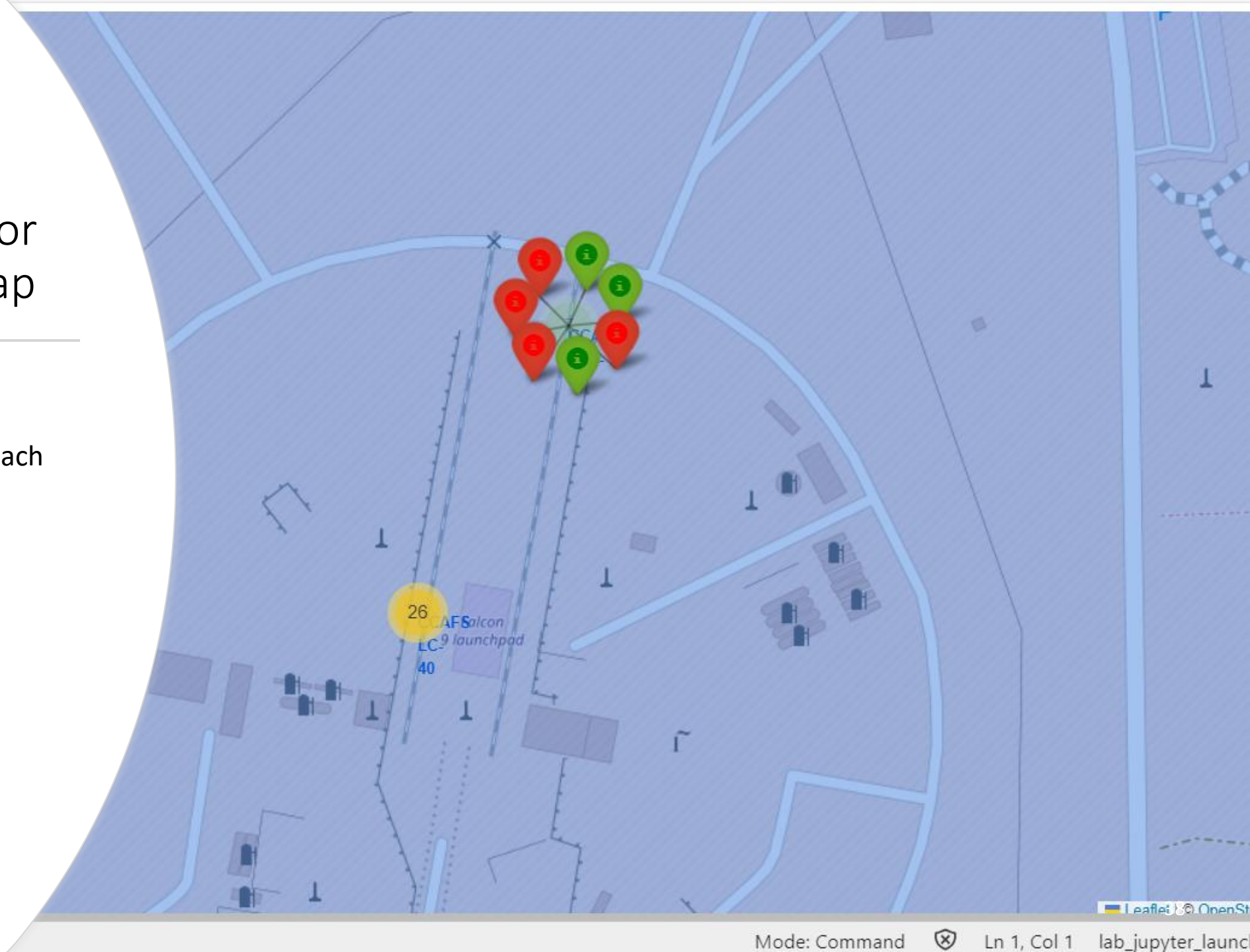
ALL LAUNCH SITES' LOCATION

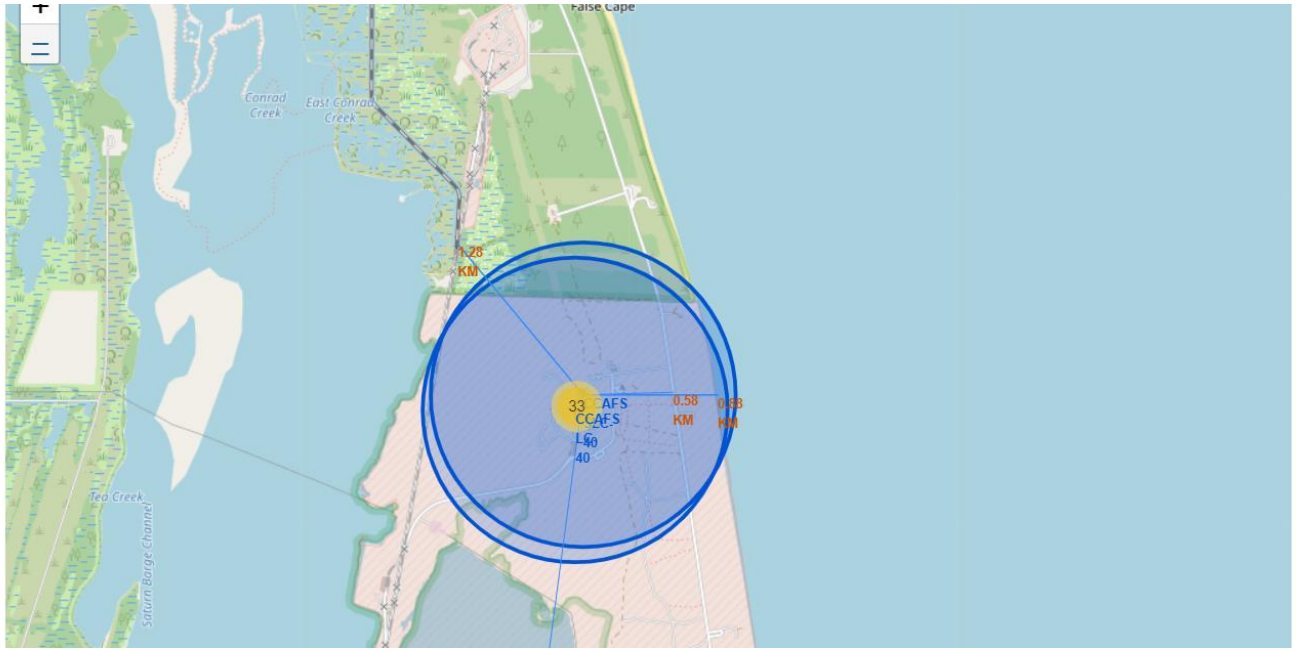
- Launch site locations are located majorly in southeast(California) and southwest(Florida) part of USA. And both are Bay areas.



Launch outcomes for each site on the map

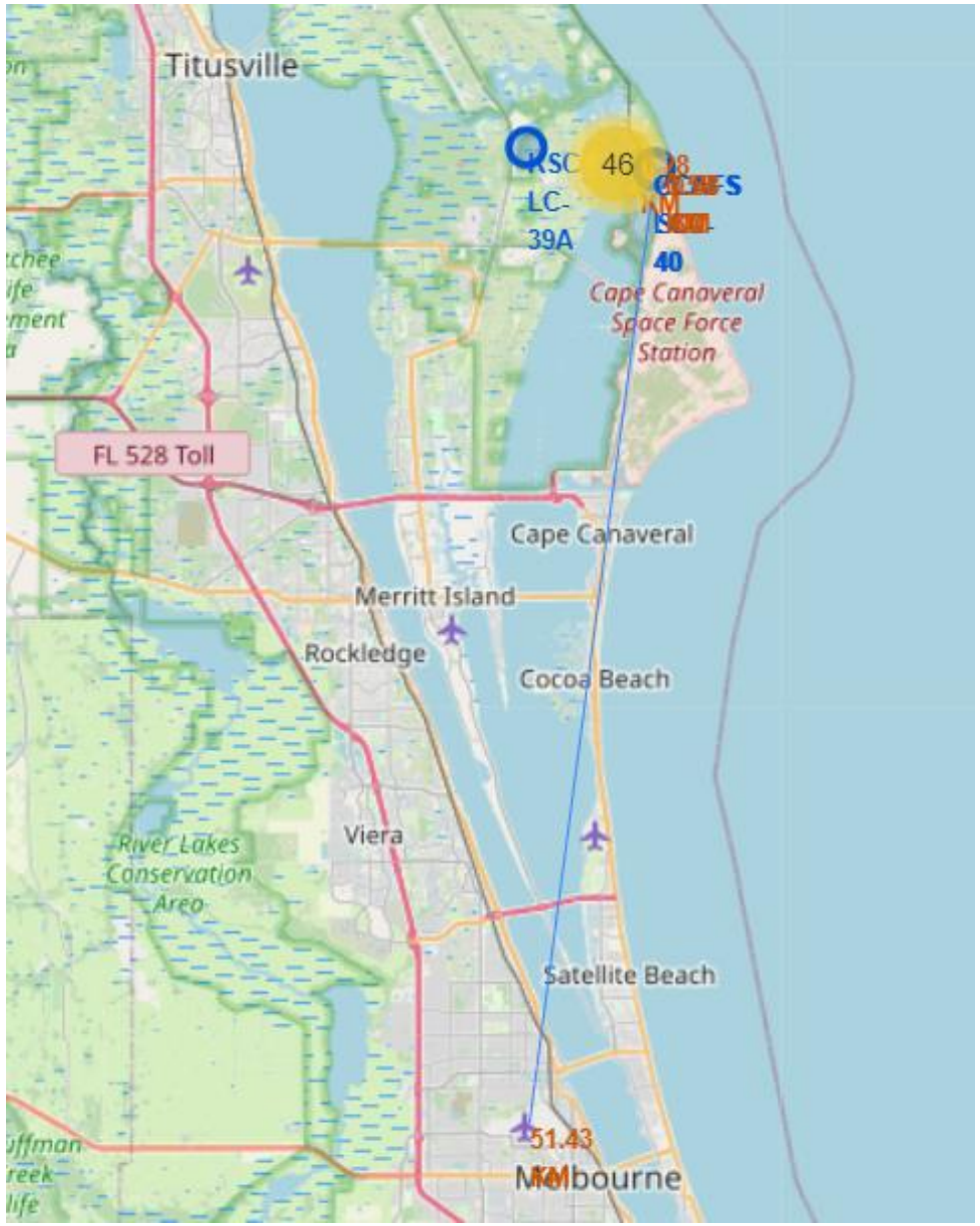
- success/failed launches for each site on the map.

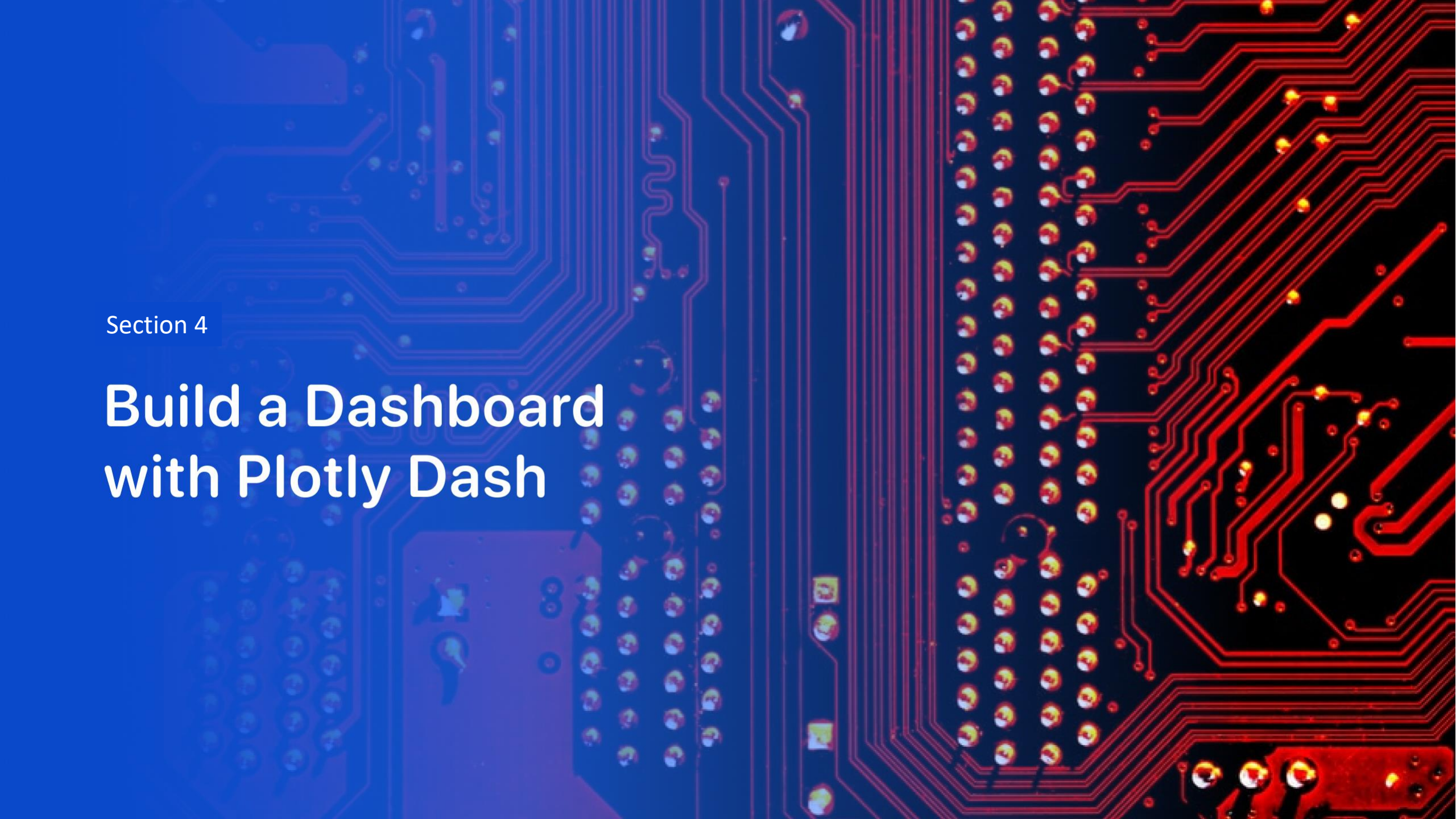




Distance of Launch Sites to Proximities

- proximities such as railway, highway, coastline, with distance calculated and displayed.



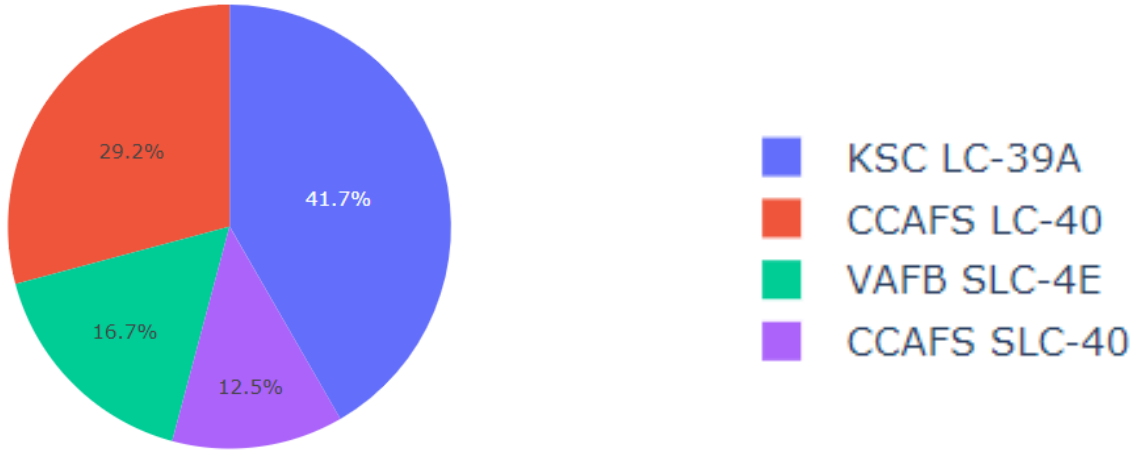

Section 4
Build a Dashboard with Plotly Dash

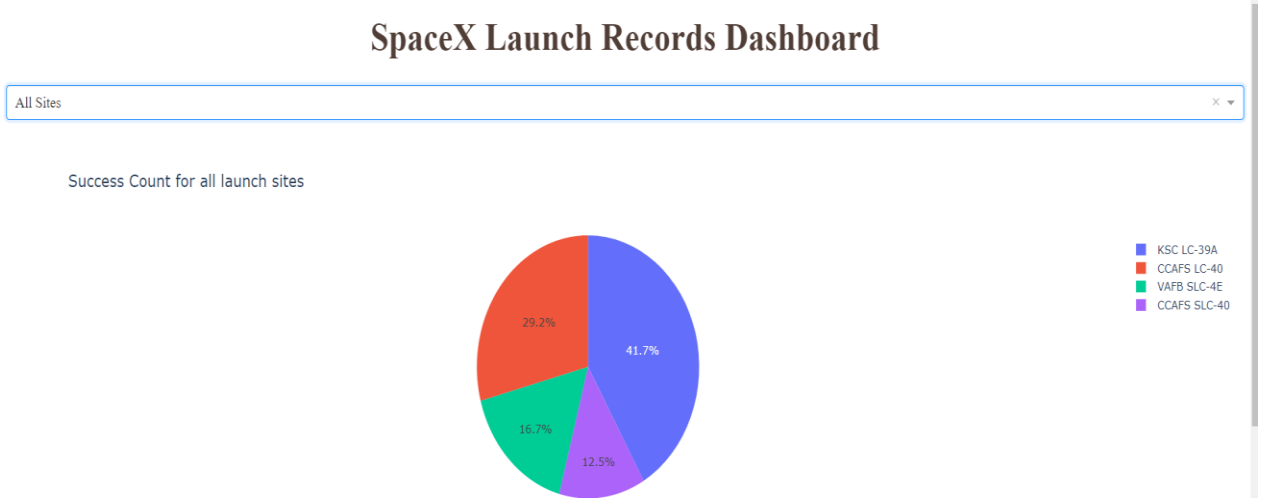
Falcon9 Landing Success Count For All Launch Sites

Success Rates: The size of each slice represents the percentage of successful launches from that particular launch site. For instance, the slice for KSC LC-39A is the biggest, indicating the highest success rate (41.7%) among the four sites for Falcon9 landing.

Focus on Success: The pie chart likely only considers successful launches and doesn't include the total number of launches from each site (which might include failures).

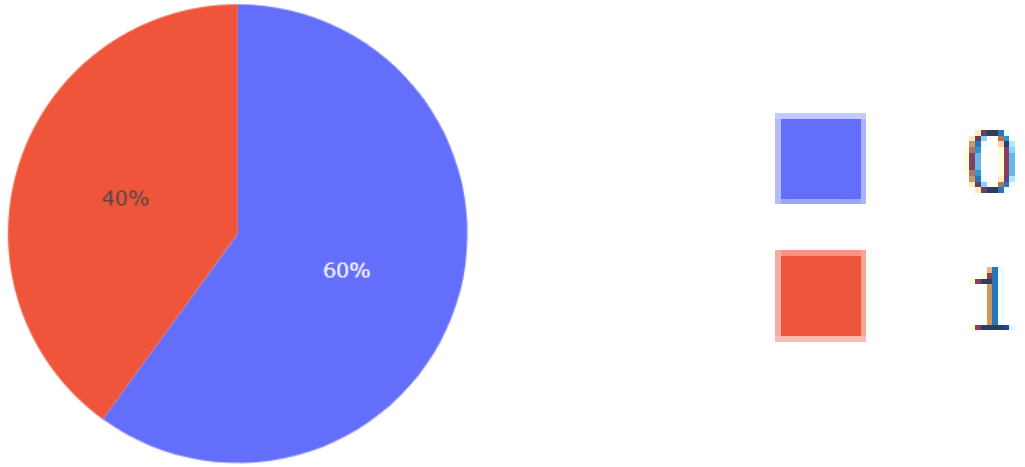
In essence, the pie chart highlights the launch sites with the highest success rates for SpaceX missions.

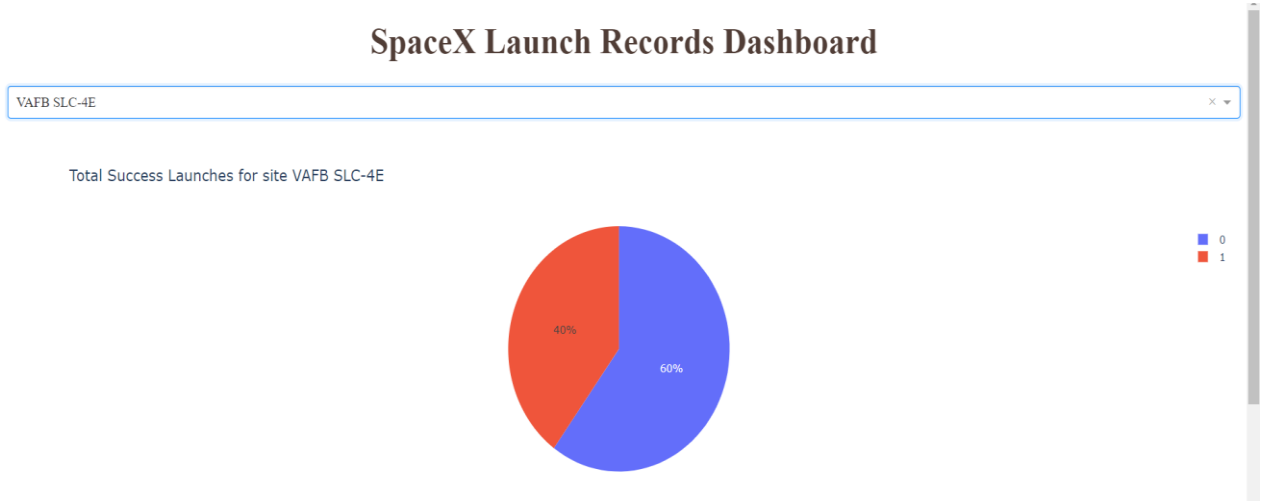




SpaceX Launch Success Rates at VAFB SLC-4E

- The pie chart from a SpaceX Launch Records Dashboard shows a pie chart focused on the total success rate for launches from the VAFB SLC-4E site.
- **Launch Site:** VAFB SLC-4E refers to Vandenberg Air Force Base Space Launch Complex 4E in California.
- **Success Rate:** The pie chart displays a success rate of 60% for SpaceX launches from this site.





Payload vs. Launch Outcome scatter plot

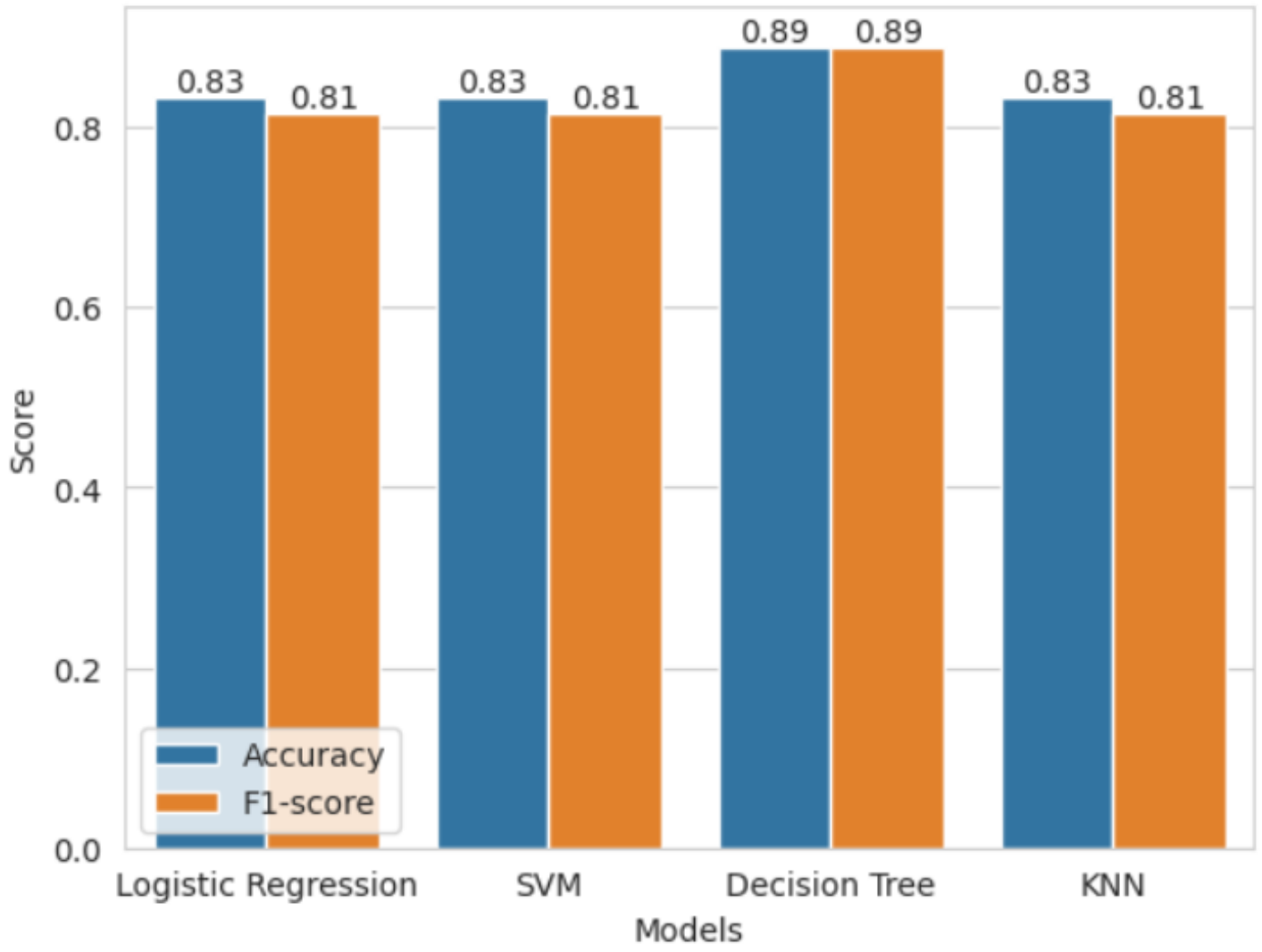
- The scatter plot shows the success rate of SpaceX Falcon 9 launches for various payload masses.
- Each point represents a launch, with the x-axis indicating the payload mass in kilograms and the y-axis indicating the mission outcome (success or failure).
- A clear trend is not visible, suggesting that payload mass may not be a significant factor in launch success for SpaceX Falcon 9 missions within the range of this data.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree Model is the best classifier that has the highest Accuracy of 88.89% when compared to other models.

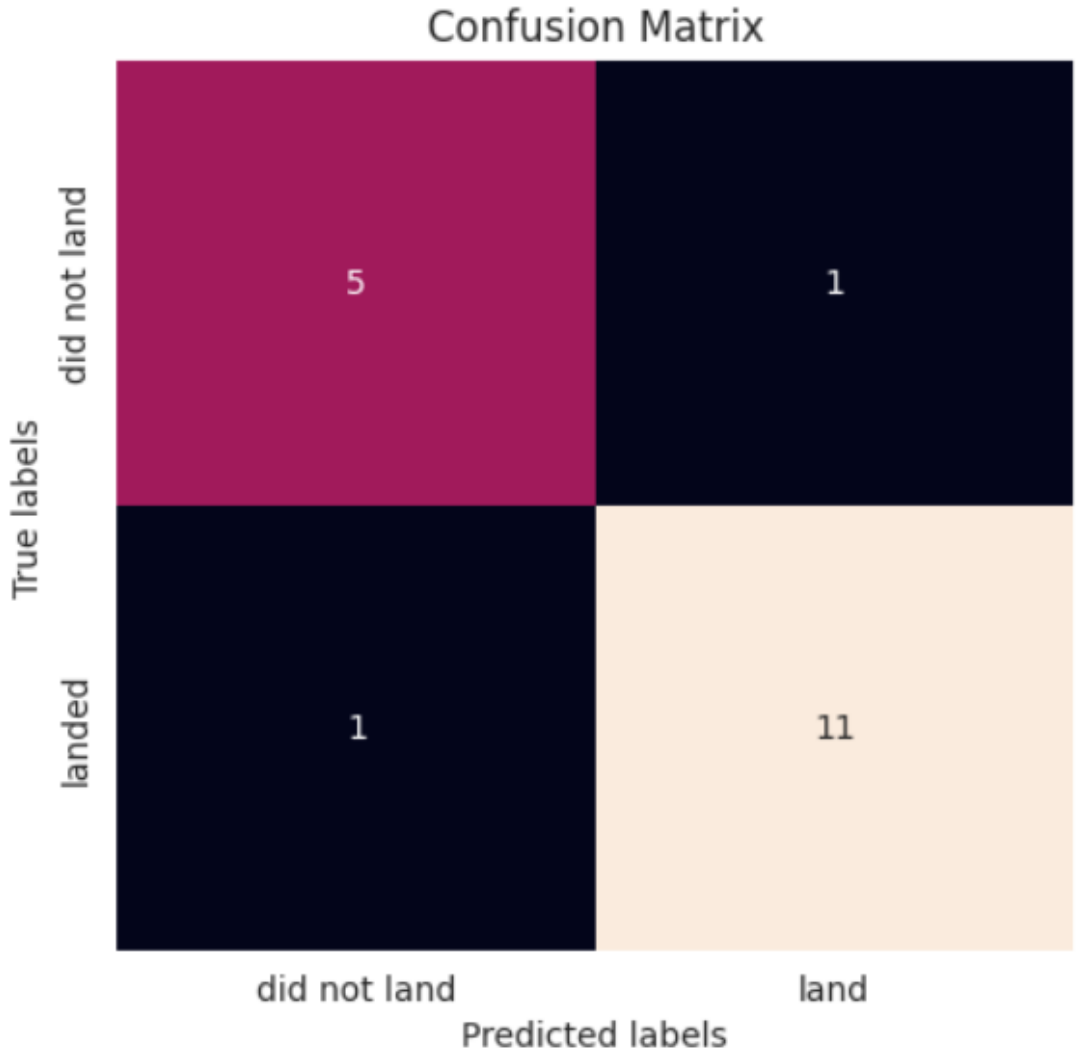


Confusion Matrix

The Decision Tree classifier has the best performance because it shows:

- **High Accuracy**: Correctly predicts 16 out of 18 instances.
- **Low Errors**: Only 1 false positive and 1 false negative.
- **Balanced Performance**: Good precision and recall for both classes.

These factors indicate that the Decision Tree classifier effectively distinguishes between the classes with minimal misclassifications.



Conclusions

- As the flight number increases there's more success rate for Launch Site CCAFS SLC 40, the same relationship was observed for payload mass.
- There's 100% success rate of landing for Orbit type ES-L1, SSO, HEO, and GEO
- There's been stead increase in the success rate of Falcon 9 landing since 2010-2020
- With a very high accuracy of 88.89%, we know that our model can predict the outcome of the landing.

Appendix

- All relevant assets like Python code snippets, SQL queries, charts, and data sets included in this presentation can be found on my **GitHub**.

Thank you!