IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Dlayne Blair>
<Feb 19 2025>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

SpaceX is an American company recognized as a leader in space exploration and rocket science advancements. It has consistently provided cost-effective rocket launches at $62 million, compared to other providers charging $165 million. I work as a data scientist at the emerging SpaceY company, which aims to compete with SpaceX. Through my research, I will evaluate the feasibility of surpassing SpaceX in the industry.

Introduction

Since 1957, nations worldwide have been competing to expand beyond Earth, whether through satellite launches or space exploration. These missions, however, come at a significant cost, with the average launch of a space rocket requiring approximately $165 million. SpaceX has revolutionized the space industry by drastically reducing launch costs to just $60 million, thanks to its cutting-edge technology and innovative approach to rocket reusability.

One of SpaceX's most groundbreaking achievements is the safe landing and reusability of the first stage of its rockets, a key factor in its success in the modern space race. This advancement not only makes space travel more affordable but also paves the way for more frequent and sustainable missions.

In this research, we aim to explore the crucial attributes and variables that influence successful rocket landings. By analyzing factors such as launch location, payload mass, rocket orbit, and mission conditions, we will apply data science methodologies—leveraging IBM's tools—to gain insights into the mechanics of SpaceX's success. Through data analysis and visualization, we will uncover the key drivers behind successful rocket recoveries and how they impact the future of space exploration.

Section 1

Methodology

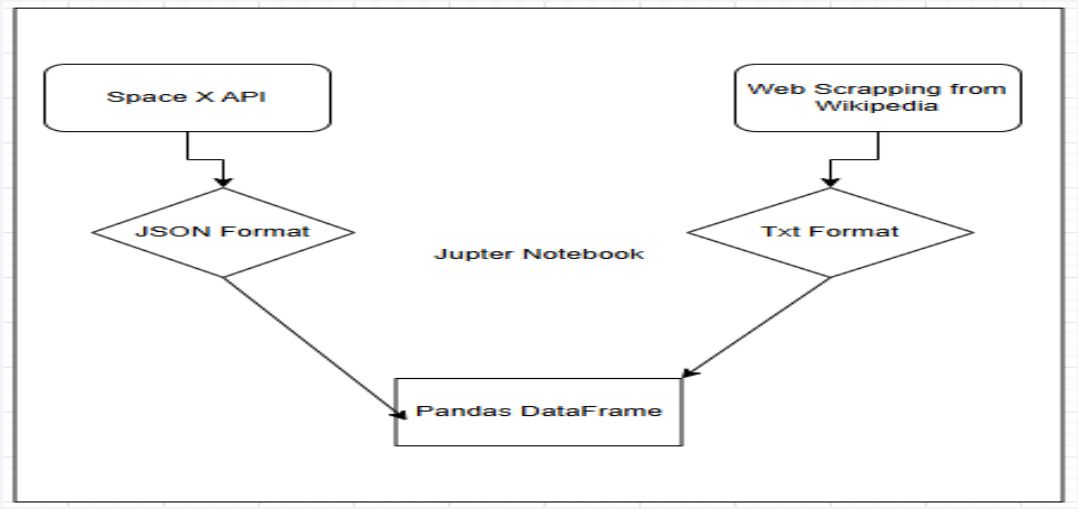
Methodology

Executive Summary

- Data collection methodology:
  - Data will be collected using SpaceX rest API and the use of web scrapping of data from various Wikipedia webpages.
- Perform data wrangling
  - We will use a series of python libraries such as Pandas and Numpy to pre-process data. The removal of unnecessary columns and normalization and standardization of data are a few techniques we will use.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We will begin by splitting the data into training and test sets, identify the optimal algorithm and parameters through Grid Search, and then deploy the best-performing model using the selected algorithm with optimized parameters.

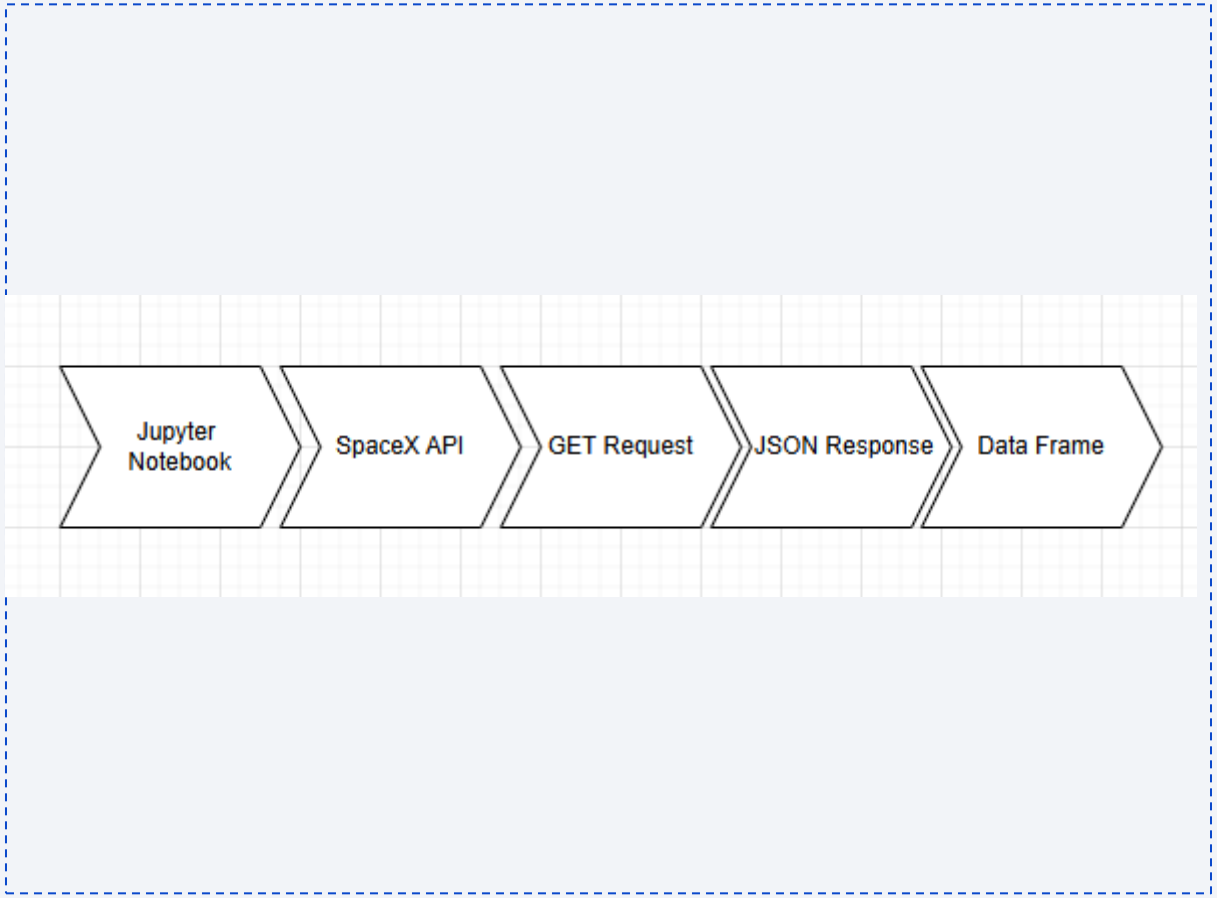
Data Collection

- Describe how data sets were collected.
  - We used two main sources for the collection of information.
    - SpaceX API: An open-source REST API used to access data on launches, rockets, capsules, Starlink satellites, launchpads, and landing pads.
    - Wikipedia: A free online encyclopedia, collaboratively created and edited by volunteers worldwide, and hosted by the Wikimedia Foundation.
- The process of how data was collected



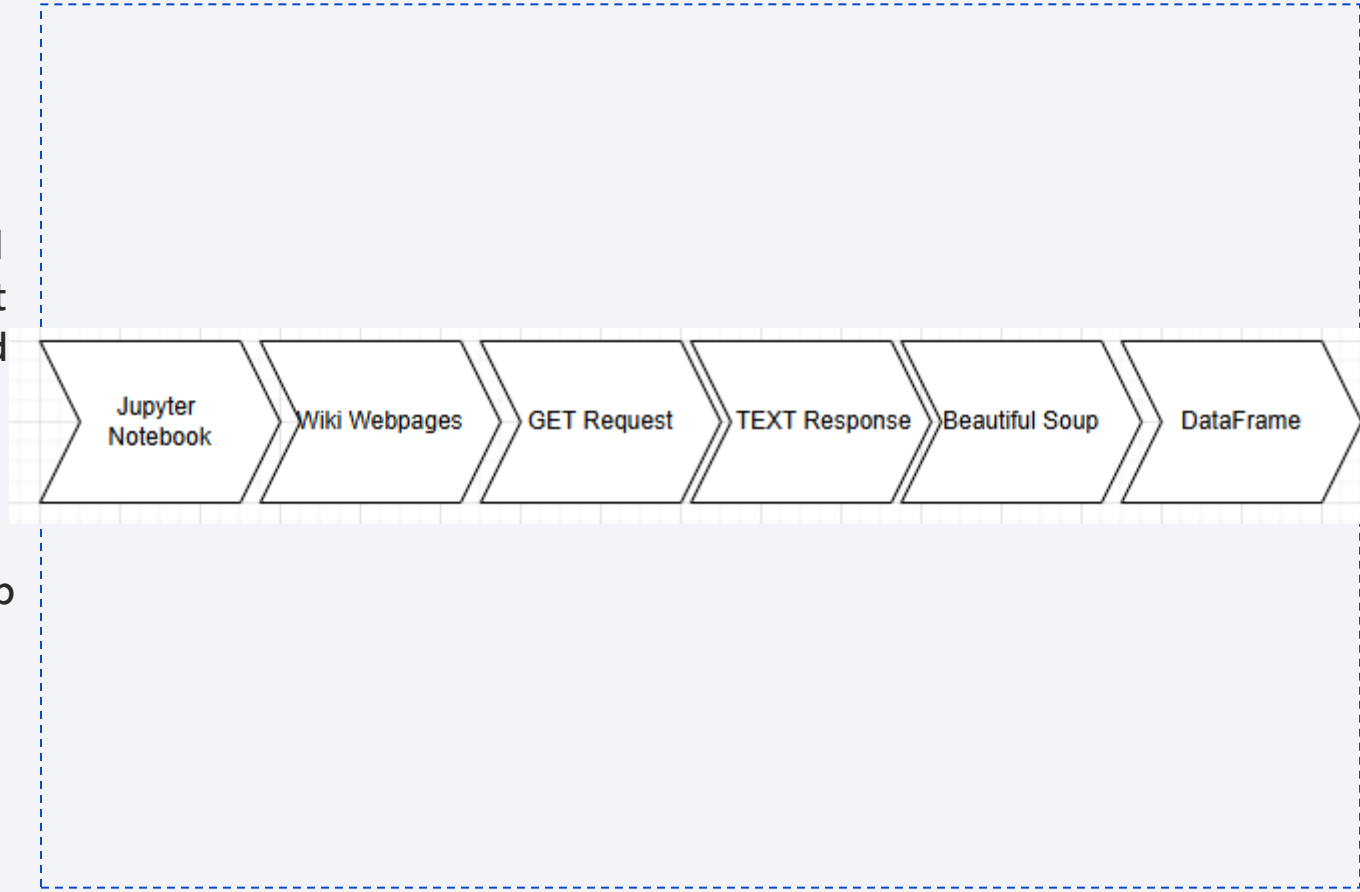
Data Collection – SpaceX API

- We began data collection from the SpaceX API by importing essential libraries such as Pandas, NumPy, and Requests. Next, we established a URL GET request, retrieving the data in JSON format. The JSON response was then converted into a DataFrame by extracting key information, including geospatial data, rocket type, orbit, flight number, and more
- Please click here to view the completed SpaceX API calls notebook



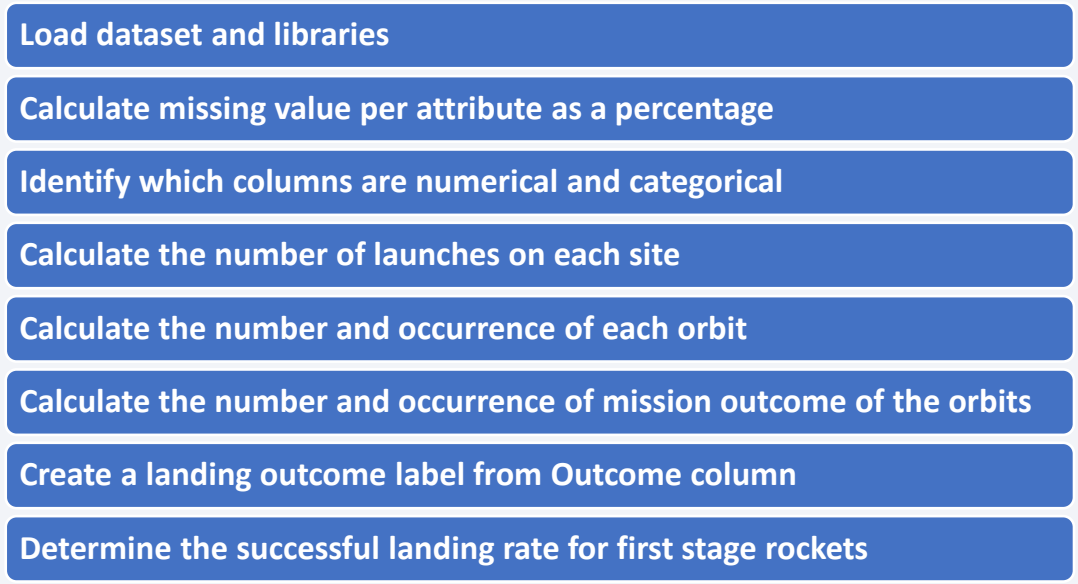
Data Collection - Scraping

- As before, we began by importing the necessary Python libraries, BeautifulSoup and Requests, to perform our task. This time, we used a Wikipedia webpage titled ‘SpaceX Falcon 9 First Stage Landing Prediction’ as our data source. We initiated an HTTP GET request, receiving the response in text format. Using BeautifulSoup, we efficiently extracted the relevant tables and columns from the response, which were then converted into a Pandas DataFrame for further analysis
- Please click here to view the completed SpaceX web scraping notebook



Data Wrangling

- At this stage, we began by importing Pandas and NumPy, followed by loading the collected data from the previous stage. Our goal was to conduct exploratory data analysis (EDA) to clean the dataset and identify the most relevant features for training a machine learning model.
- Please click here to view the completed data wrangling notebook



EDA with Data Visualization

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

- At this stage we completed the Exploratory Data Analysis process by finding the correlation between various dependents and independents with the aid of various visualization tools. The main libraires used in this section were the seaborn and matplotlib. We also worked on converting types of columns to suit our data
- Please click here to view the completed EDA with data visualization notebook.

EDA with SQL

- Here we queried the dataset using SQL commands to complete the tasks seen to the right.
- Please click here to view the completed EDA with SQL notebook.

Task Performed
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 succesful landing outcome in ground pad was acheived
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. Use a subquery
List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

Build an Interactive Map with Folium

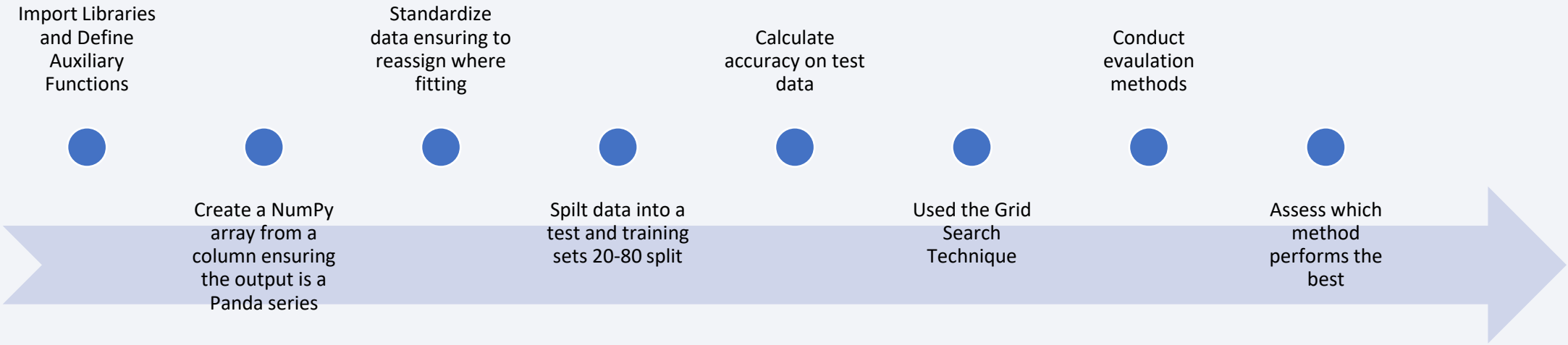
- Geospatial data was used to place markers, circles and lines on a map to identify various launch sites.
  - **TASK 1:** Mark all launch sites on a map – we identified 4 locations along with their latitude and longitude
  - **TASK 2:** Mark the success/failed launches for each site on the map - placed launches into 2 groups
  - **TASK 3:** Calculate the distances between a launch site to its proximities – after calculating the distances we attempted to draw polylines to show distance.
- Please click here to view the completed interactive map with Folium map.

Build a Dashboard with Plotly Dash

- Add a Launch Site Drop-down option for each Launch Site for ease in selecting 1 or all sites
- Attempted to create pie charts that shows the count for successful launches per site.
- Add a Range Slider to Select Payload
- Attempted to create scatterplots to render the successful launches and and their payload to discern correlation.
- Please click here to view the completed Plotly Dash lab.

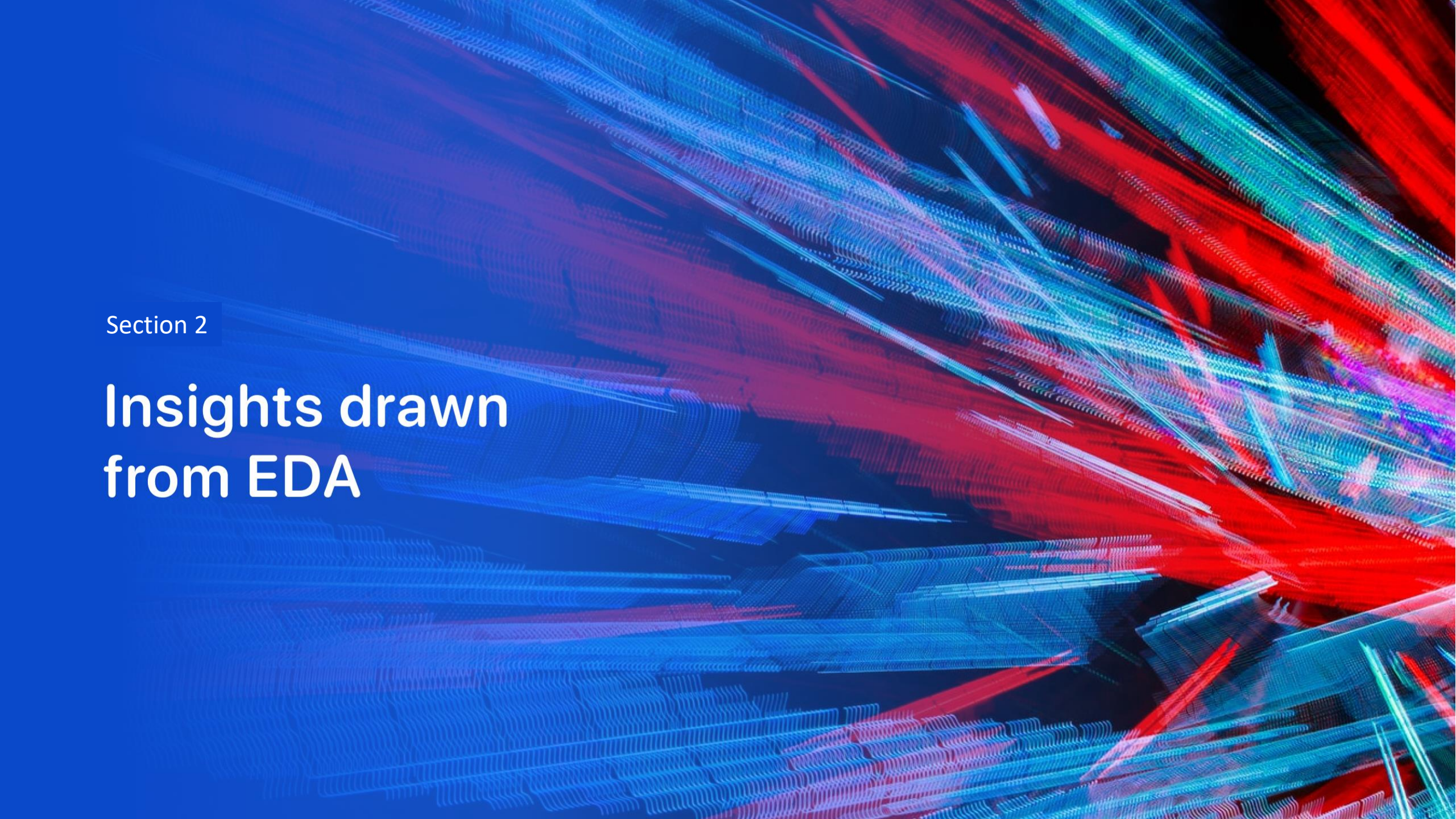
Predictive Analysis (Classification)

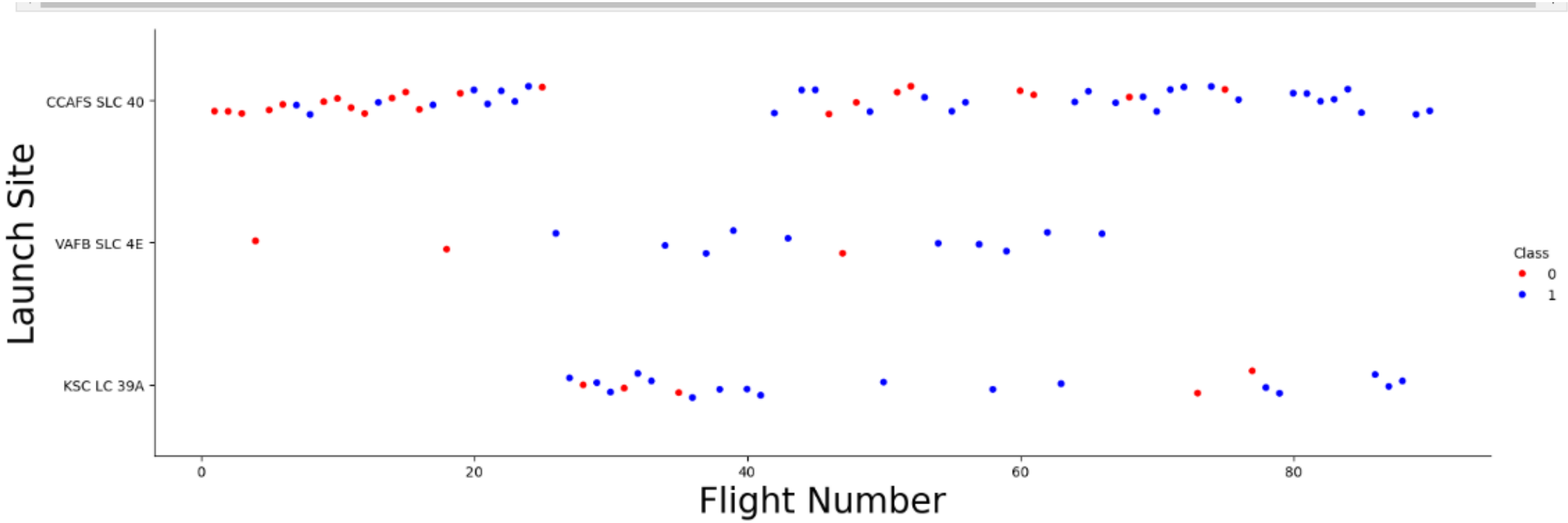
- To build the best-performing classification model, we started by importing the necessary libraries and loading the dataset. The data was standardized to eliminate bias and then split into 80% training and 20% testing sets. We initialized four classification algorithms—Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN)—and used the Grid Search technique to fine-tune their hyperparameters. To evaluate model performance, we employed key metrics, including the Confusion Matrix for visualization, the F1 Score for balanced accuracy, and the Jaccard Score for similarity measurement. After comparing the performance of all models, the best-performing algorithm was selected for deployment, ensuring optimal classification accuracy.
- Please click here to view the completed predictive analysis lab.



Results

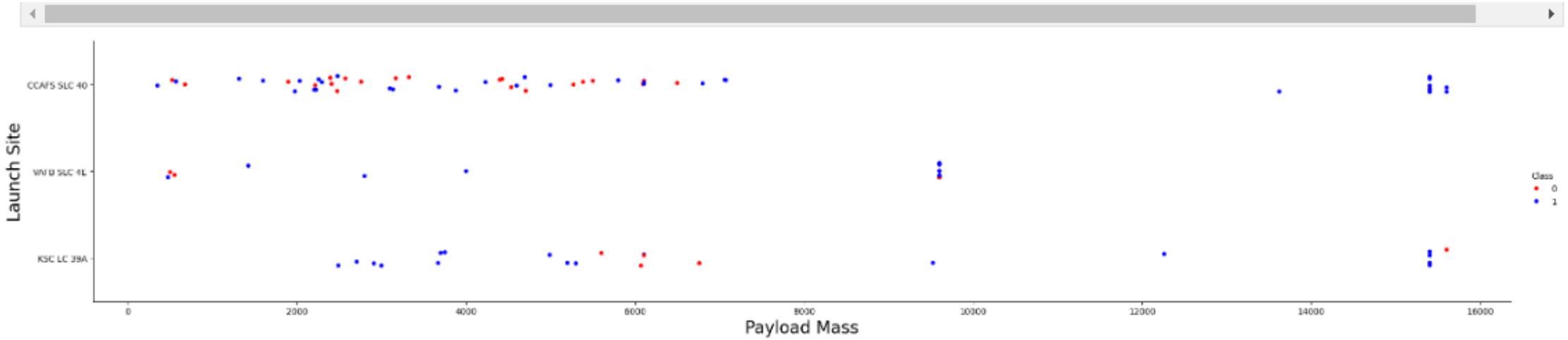
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA



Flight Number vs. Launch Site

- From the plot we understood that CCAFS SLC 40 is the most used site for launching rockets, The least can be seen to be VAFB SLC 4E with only 13 launches however it has only 3 failures making it one of the most successful launch rate.

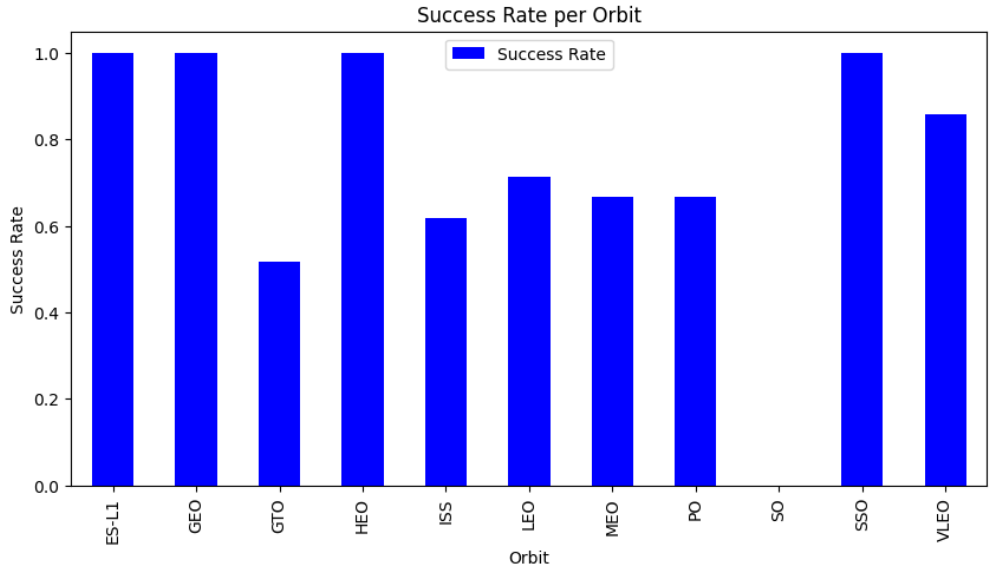


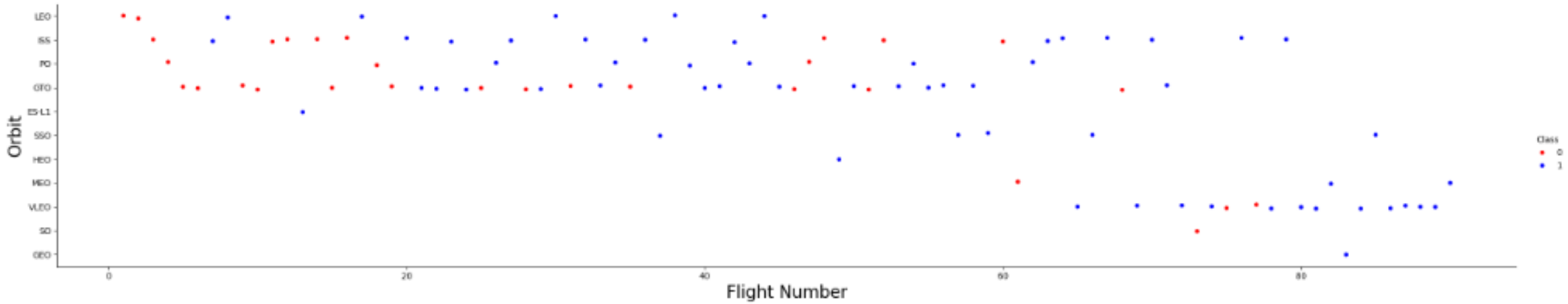
Payload vs. Launch Site

- The plot suggest that there is no correlation between the payload size and the launch sites as they both contain similar amounts of failures and successes.

Success Rate vs. Orbit Type

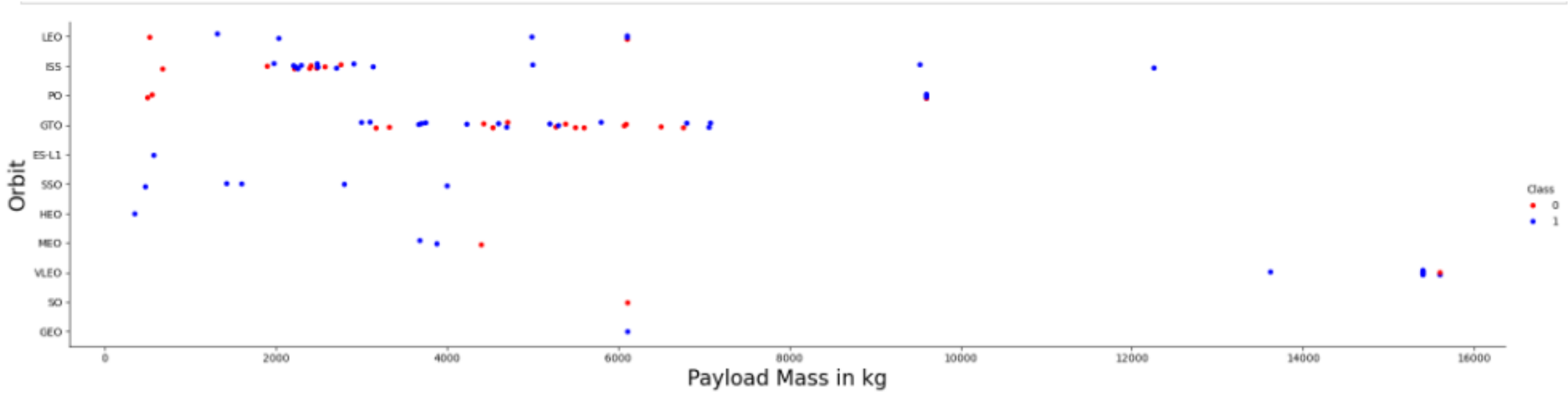
- The bar chart allows us at a glance to determine the best orbits for successful returns of the first stage rockets. ES-11, GEO, HEO, and SSO have the joint highest whilst GTO the lowest.





Flight Number vs. Orbit Type

- There is a gradual increase in the number of success as there is an increase in flight number, for LEP, ISS and VLEO orbits but GTO doesn’t show this correlation.



Payload vs. Orbit Type

- The red dots show us that lower payload causes unsuccessful landing we see this in LEO and ISS as well as PO orbits. GTO shows the payload has no effect on success.

Launch Success Yearly Trend

There is a positive trend seen in the success rate as the time goes by.

All Launch Site Names

- We used SQL queries to identify distinct launch_sites from SPACEXTABL table. This produce 4 results.

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

Launch Site Names Begin with 'CCA'

- To identify launch sites beginning with ‘CCA’ we queried the table SPACEXTBL for launch_site that started with CCA. To do this we used a wildcard format ‘CCA%’ ensuring that the first 3 letters CCA furthermore, be present at the beginning of the name. Furthermore, the use of LIMIT(5) was used to produce only 5 results.

Date	Time (UTC)	Booster_Version	Launch_Site	
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Sp Qua
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	der C
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	der
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	

Total Payload Mass

- To get he total payload mass we queried the sum of table payload_mass_kg from the spacextabl and ensured that only customers equal to ‘NASA (CRS)’ was produced.

sum(payload_mass__kg_)
45596

2928.4

Average Payload Mass by F9 v1.1

- We selected the avg function for the column payload_mass_kg from SPACEXTBL where the booster_version had to equal ‘F9 v1.1’.

Task 5

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

Hint:Use min function

In [20]:
```
%sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

Out[20]:

min(DATE)
2015-12-22

First Successful Ground Landing Date

From my query the first successful ground landing date was the 22nd December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- To identify the successful drone ship landings, we ensured that the column that contained the names of the boosters (booter_version) was selected from the table SPACEXTBL. We then queried the tables to produce results where landing_outcome was equal to ‘ Success (drone ship)’ and where payload_mass_kg was between 4000 to 6000.

Booster_Version
F9 FT B1029.1
F9 FT B1036.1
F9 B4 B1041.1

Task 7

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(mission_outcome) as counts from SPACEXTBL GROUP BY mission_outcome
```

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

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

- I attributed the difference in Success to be a trimming in the dataset issue, however 99 successful missions occurred in total.

Total Number of Successful and Failure Mission Outcomes

Boosters Carried Maximum Payload

- 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

To gather the max payload carried by booster the below was used:

```
select distinct booster_version from SPACEXTBL where payload_mass__kg_ in
(select max(payload_mass__kg_) from SPACEXTBL);
```

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

2015 Launch Records

In 2015 2 failed landings were recorded , coincidentally both occurred at the same launch site. Booster 12 and 15 both recorded failures.

select Landing_Outcome, booster_version, launch_site from SPACEXTBL where (Landing_Outcome = 'Failure (drone ship)' and date like '2015%')

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.
- select Landing_Outcome, count(*) as Landing_Outcome_count from SPACEXTBL where DATE between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count(Landing_Outcome) desc

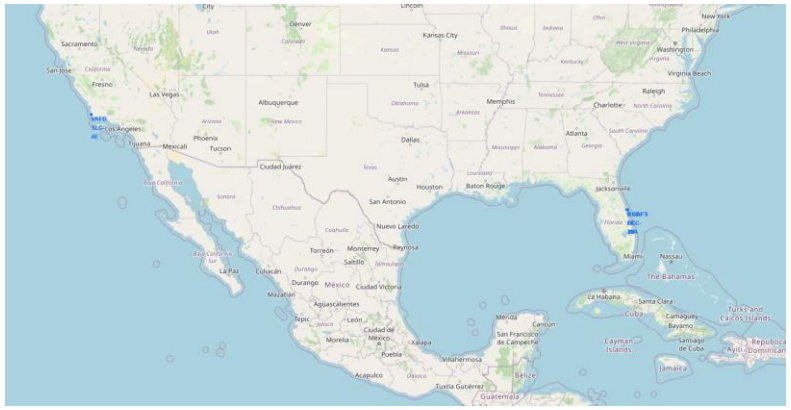
Landing_Outcome	Landing_Outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

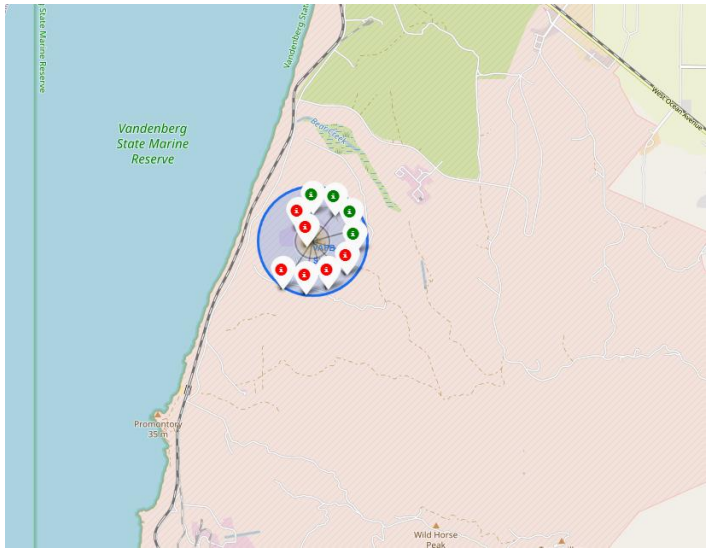
Launch Sites Proximities Analysis

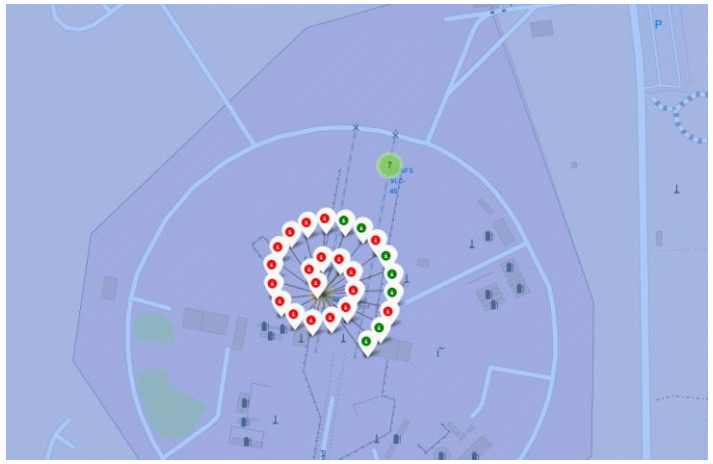
Launch sites Folium Map

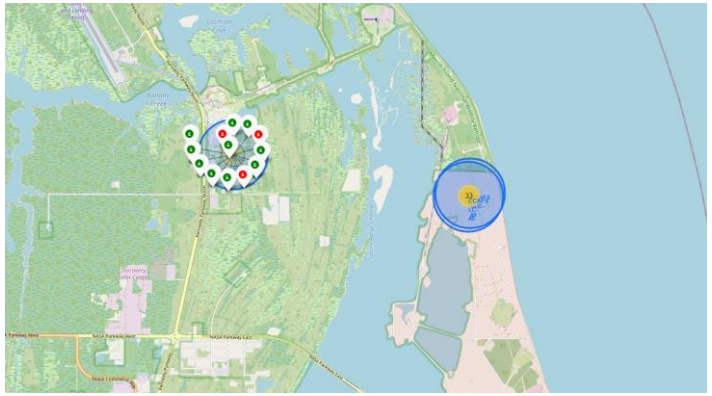
Based on my investigation, SpaceX launch sites are strategically located at lower latitudes to take advantage of Earth's rotational velocity. This positioning helps reduce fuel requirements, as rockets launched closer to the Equator gain additional speed from Earth's rotation. Additionally, launch sites are typically situated near coastal locations and over open water to minimize risks to populated areas in case of launch failures. Coastal proximity also enhances booster recovery operations, as first-stage boosters can land on drone ships or safely return to land after launch.



Launch Site	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610745







Success Rate per Location

- We can see that at the various locations the markers present indicate success rate. With Green being successful and Red a failed return.

Location	Lat	Long
Orlando_Location	28.52300	-81.38260
Coastline_Location	28.56146	-80.56746
Highway_Location	28.56270	-80.58703

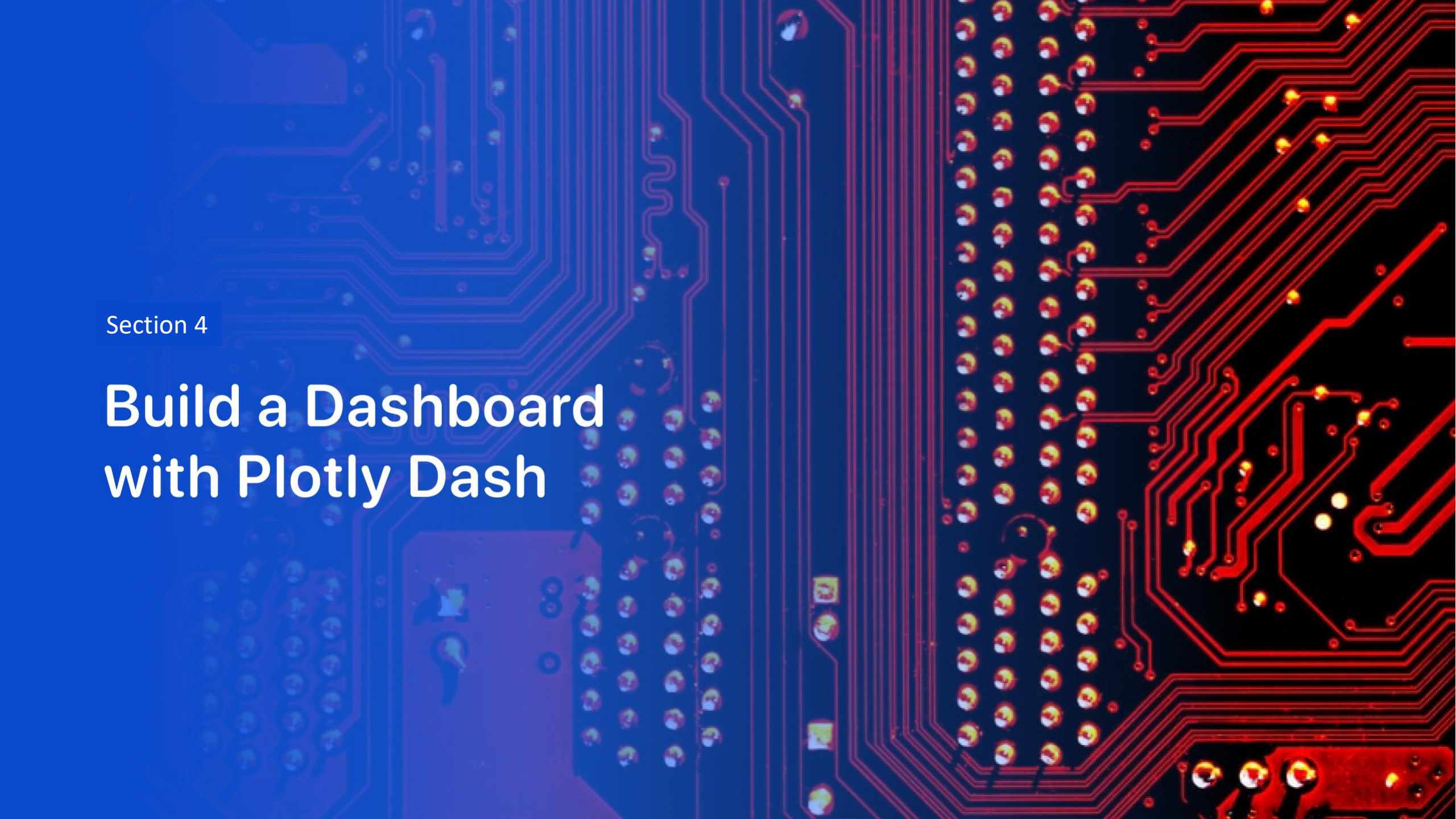
Proximities closest to launch site CCAFS LC-40

The location of CCAFS LC-40 is Lat: 28.5623 and Long: -80.5774as such the calculated distances would be

Orlando_Location: ~79.6 km

Coastline_Location: ~0.9 km

Highway_Location: ~0.9 km


Section 4
Build a Dashboard with Plotly Dash

Launch Success per Site

KSC LC 39A Dashboard: Total Launch Success

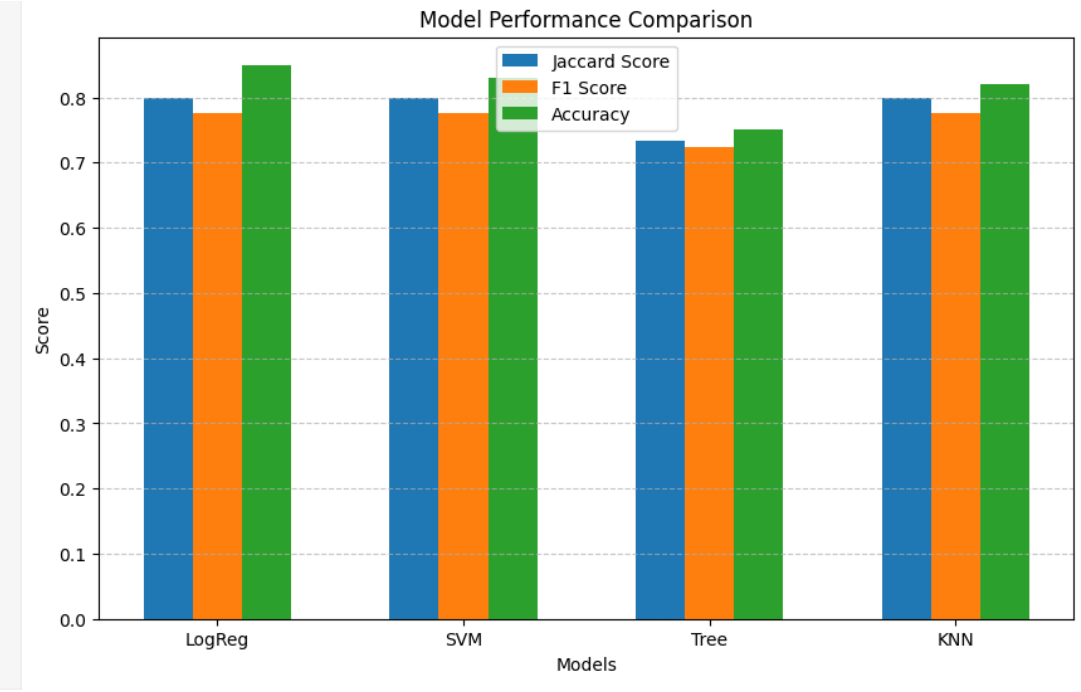
Payload Correlation per Site

Section 5

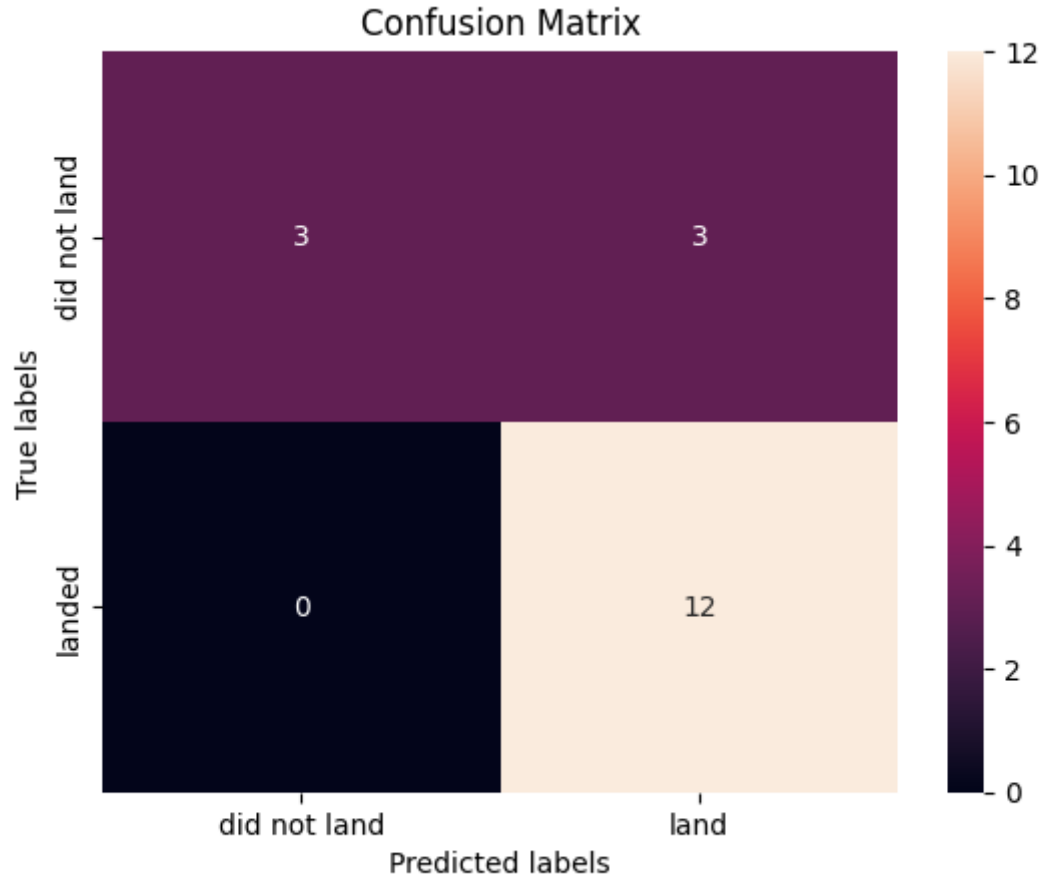
Predictive Analysis (Classification)

Classification Accuracy

- From the graph Decision Tree performed the worst out of all models used



Confusion Matrix



- The confusion matrix of the Logistic Regression and the KNN provided similar results

Conclusions

A successful first-stage return significantly reduces rocket launch costs, making reusability a key factor in cost efficiency.

Various factors—including booster type, target orbit, and launch site characteristics—play a significant role in determining the probability of a successful first-stage landing.

Falcon 9 launch facilities are positioned near major highways, railways, and coastal areas to streamline transportation and operational logistics. Additionally, sites with a higher launch frequency tend to exhibit greater success rates.

SpaceX's launch reliability has shown a steady upward trend since 2013, with the KSC LC-39A facility recording the highest number of successful missions

ES-L1, GEO, HEO, and SSO, orbits have been associated with higher launch success rates, highlighting the importance of orbital positioning in mission planning.

Thank you!