IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohsen Doust
10/23/2023

Outline

Executive Summary

- This project is focused on predicting the success of Falcon 9 first stage landings, crucial for reducing SpaceX launch costs. Methodologies:
- **Problem Definition**: Determine Falcon 9 first stage landing probability for cost estimation.
- **Data Handling**: Python and Pandas for dataset collection, cleaning, and preprocessing.
- **EDA**: Explored dataset for insights into landing success factors.
- **Machine Learning**: Logistic Regression, SVM, Decision Tree, and KNN models for predictions.
- **Hyperparameter Tuning**: GridSearchCV optimized model performance.
- **Model Evaluation**: Accuracy and confusion matrices assessed model reliability.
- **Conclusion**:
- Accurate Support Vector Machine predictions offer insights, optimizing SpaceX launch costs.

Introduction

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. In this project, we will predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by using SpaceX API and by web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled `List of Falcon 9 and Falcon Heavy launches`
- Perform data wrangling
  - performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Used several classification models and compared the result to Find the method performs best

Data Collection

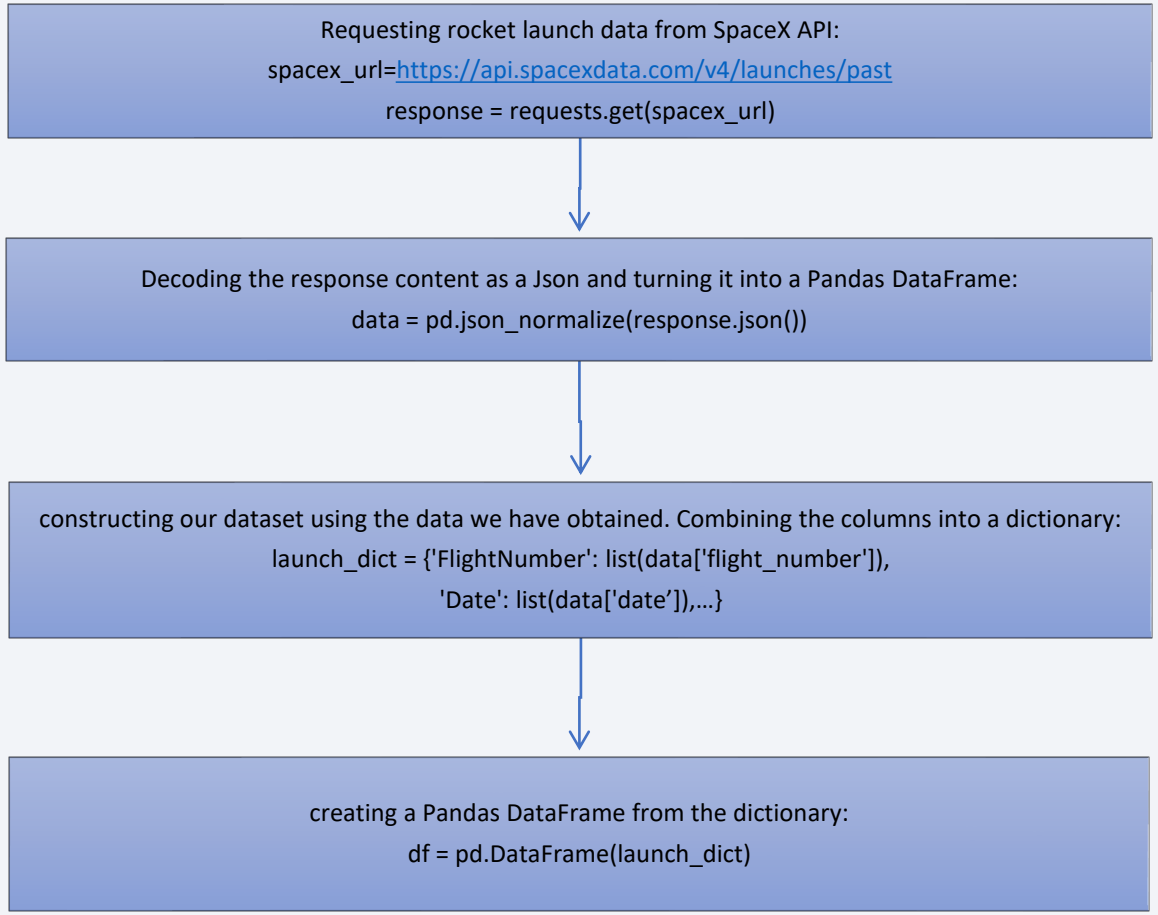
- Data was collected by making a get request to the SpaceX API. Some data wrangling (like dealing with missing data) and formatting was applied.

- Some data was also collected by web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled `List of Falcon 9 and Falcon Heavy launches`: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

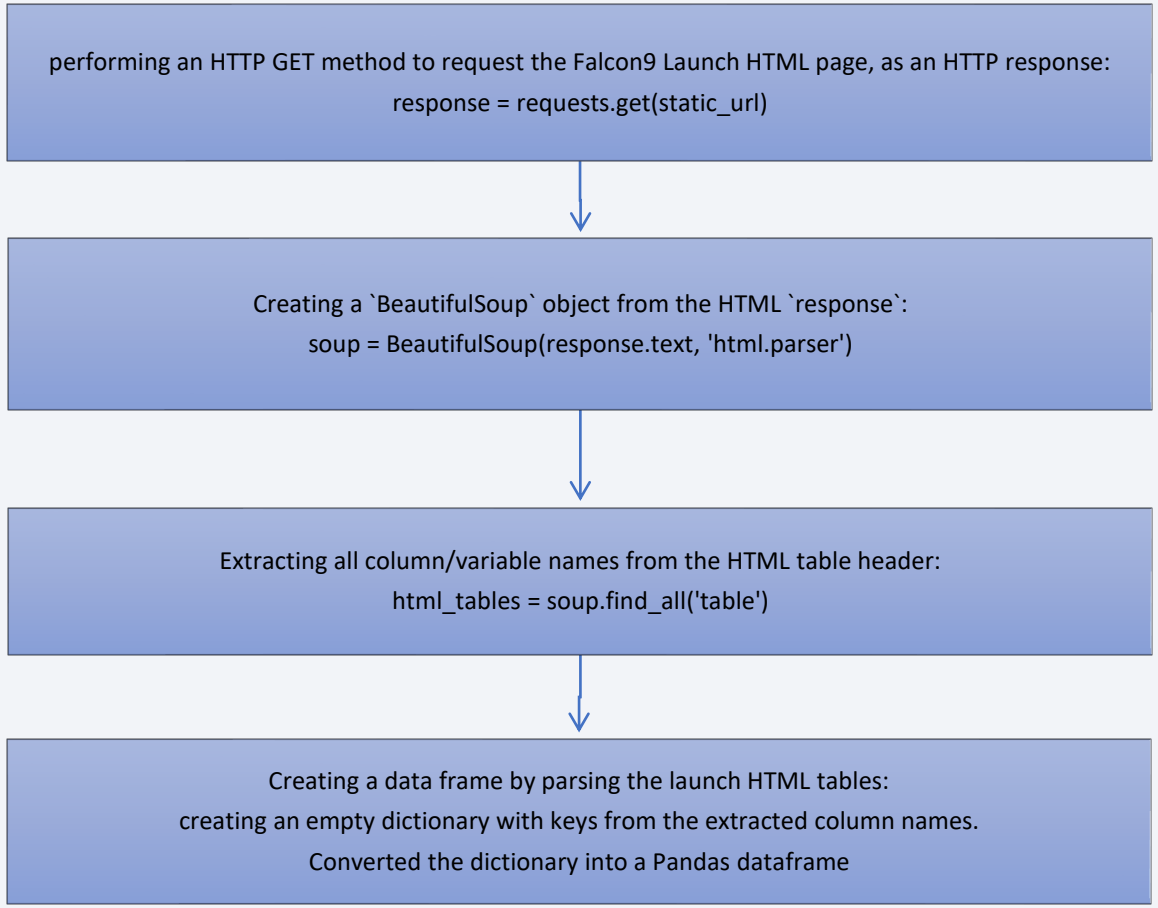
- We made a request to get rocket launch data from SpaceX API and turned it to Pandas DataFrame.

- GitHub: Collecting the data



Data Collection - Scraping

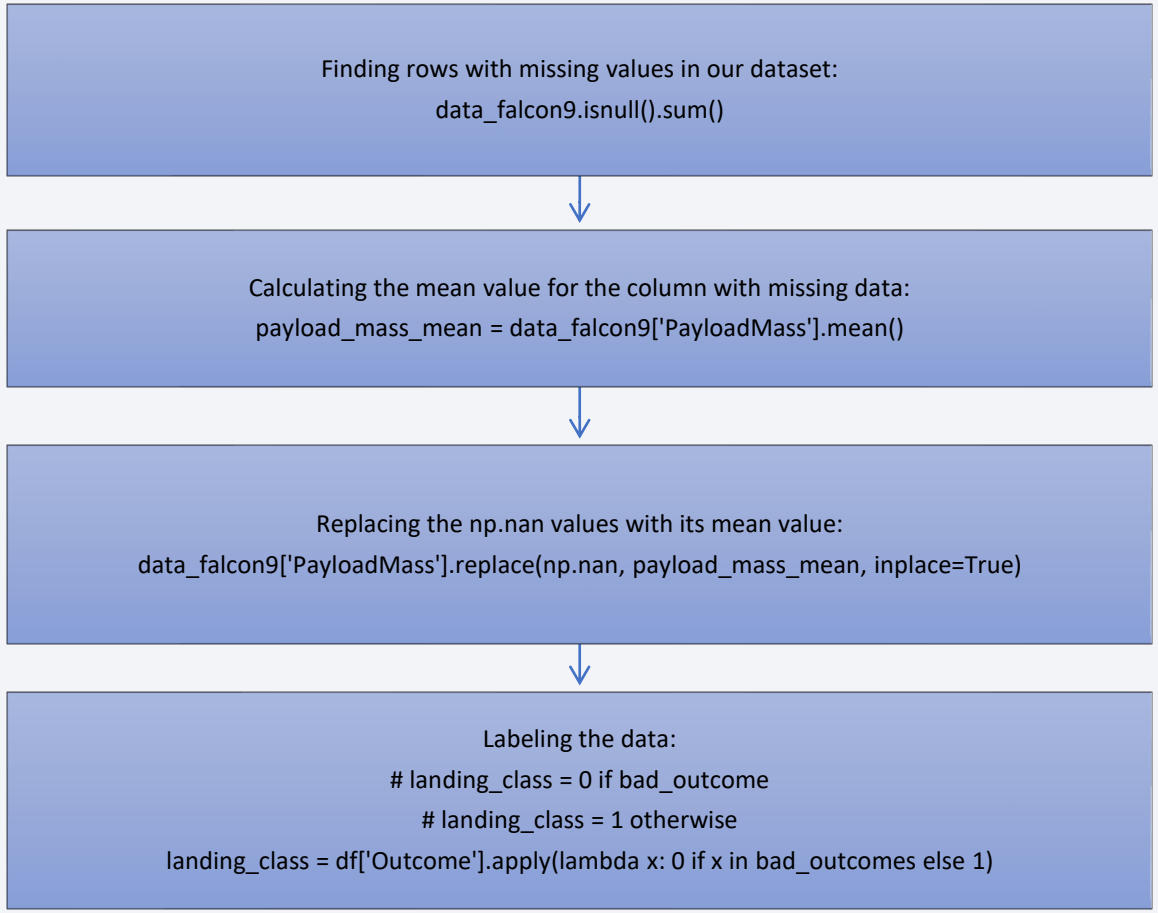
- We performed web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled `List of Falcon 9 and Falcon Heavy launches`
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- GitHub: Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia



Data Wrangling

- We looked for missing values in our dataset and replaced the null values with the mean value of the column
- We also labeled all different kind of landings (landing to the ocean, to a ground pad, or on a drone ship) as 1 for successful landing and 0 for unsuccessful landing. We needed this for training supervised models

- GitHub Data wrangling



EDA with Data Visualization

- Using data visualization, we tried to find the relationship between different variables.
- We used cat plots and scatter plots to visualize the relationships between numerical categorical variables like FlightNumber and PayloadMass, or between Flight Number and Launch Site or between Payload and Launch Site
- We used bar charts to visually check if there is any relationship between success rate and orbit type.
- We also visualized the launch success yearly trend. We used lineplot to show the trend.

- GitHub: EDA with data visualization

EDA with SQL

- To understand the SpaceX Dataset, we loaded the dataset into a table in a Db2 database. Then executed some SQL queries to:
  - Display the names of the unique launch sites in the space mission.
  - 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 total number of successful and failure mission outcomes
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad))

- GitHub: EDA with SQL

Build an Interactive Map with Folium

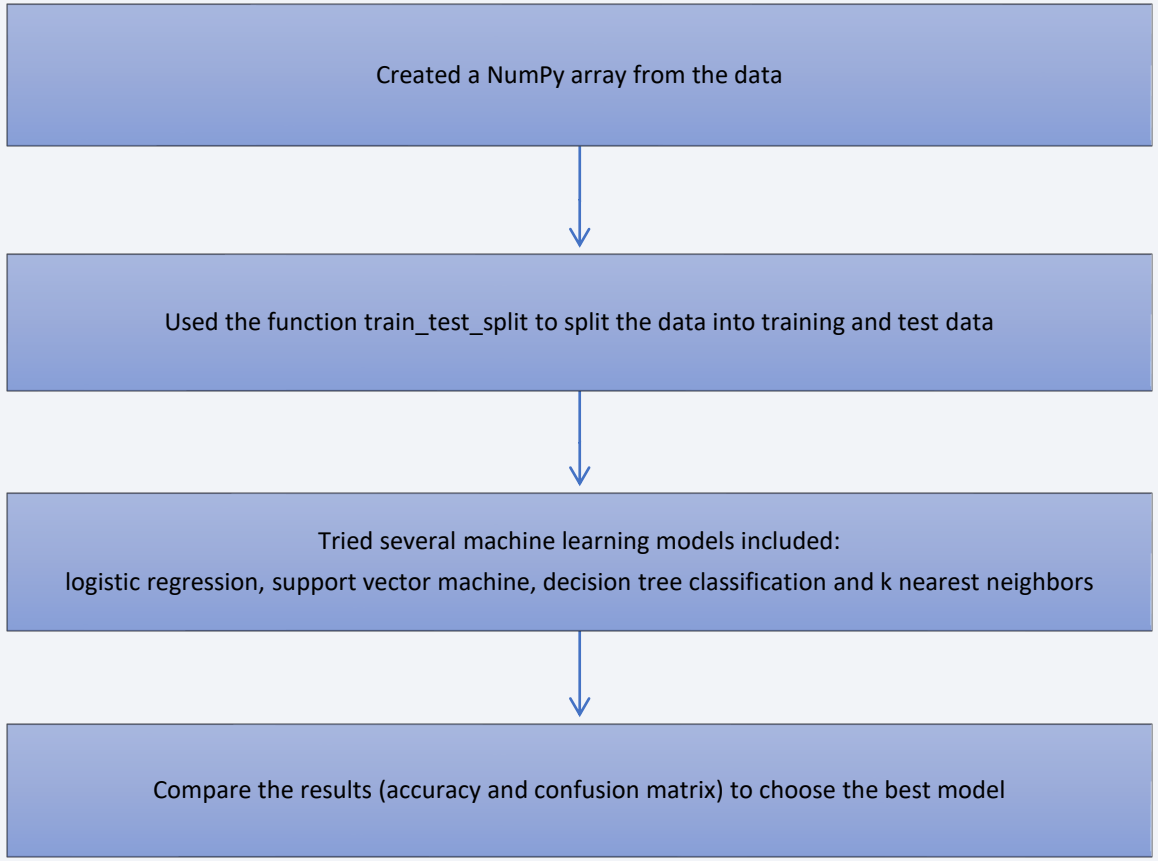
- To find out if there are any relationships between launch success rate and location and proximities of a launch site, we created an interactive map with Folium.
- We marked all launch sites and the success/failed launches for each site on the map. Then we calculated the distances between a launch site to its proximities. After you plot distance lines to the proximities, you can answer the following questions easily:
  - Are launch sites in close proximity to railways?
  - Are launch sites in close proximity to highways?
  - Are launch sites in close proximity to coastline?
  - Do launch sites keep a certain distance away from cities?
- GitHub: Launch Sites Locations Analysis with Folium

Build a Dashboard with Plotly Dash

- We made an interactive dashboard and added dropdown lists and sliders to interact with the pie charts and a scatter point chart.
- These charts will help us to answer some questions like:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?
- We will present some screenshots from the dashboard in this presentation, and we provided the link to the source files here:
- GitHub: Dashboard Application with Plotly Dash

Predictive Analysis (Classification)

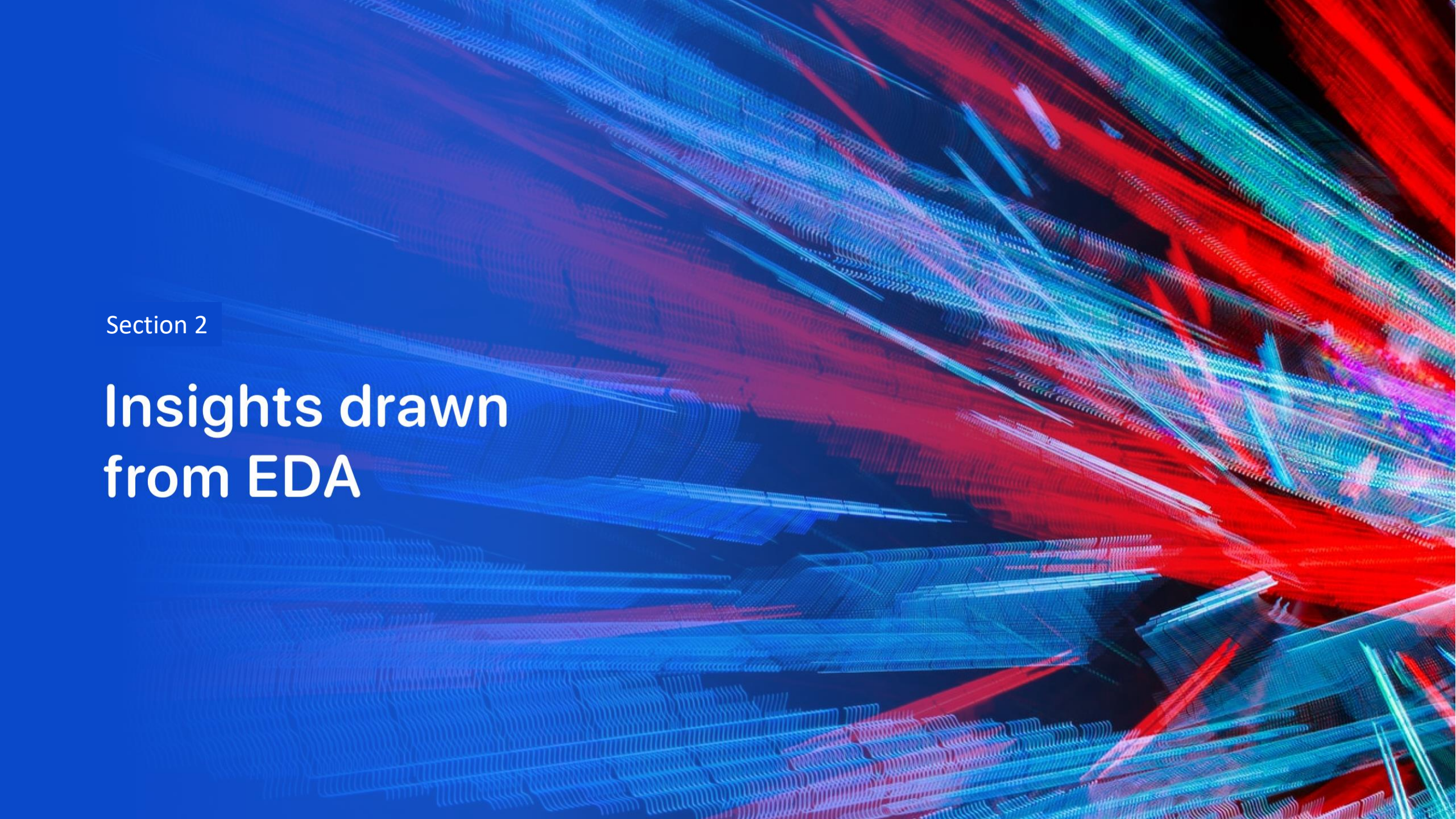
- Most of the cost savings that Space X advertises on Falcon 9 rocket launches, are because Space X can reuse the first stage. We created a machine learning pipeline to predict if the first stage will land, given the data from the preceding stages. The flowchart is this page will show the steps we took to make this happen.



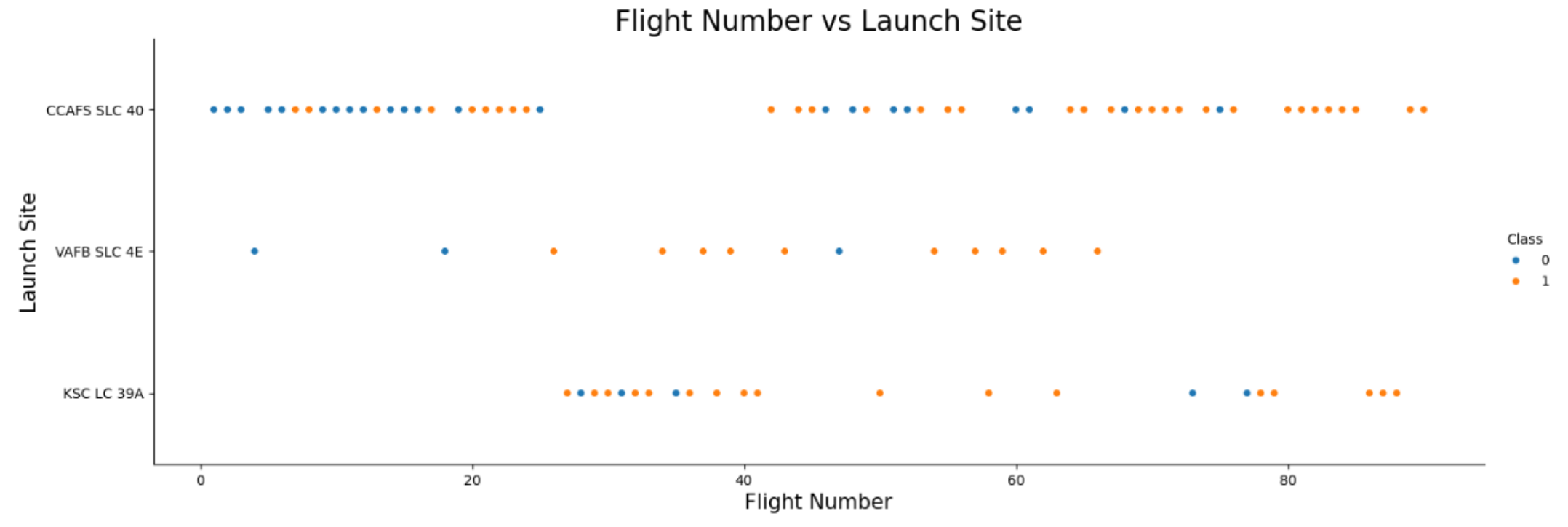
- GitHub: Machine learning predictive analysis

Results

- Exploring the data, we were able to get some insight like the names of the unique launch sites in the space mission, or the average payload mass carried by booster version F9 v1.1, or total number of successful and failure mission outcomes. We also made some visualizations from the data to find the relationship between different variables. We used cat plots and scatter plots to visualize the relationships between numerical categorical variables like FlightNumber and PayloadMass, or between Flight Number and Launch Site or between Payload and Launch Site
- We also created a machine learning pipeline to predict if the first stage will land, given the data from the preceding stages. We tried several machine learning models included: logistic regression, support vector machine, decision tree classification and k nearest neighbors. Compare the results (accuracy and confusion matrix) to choose the best model.

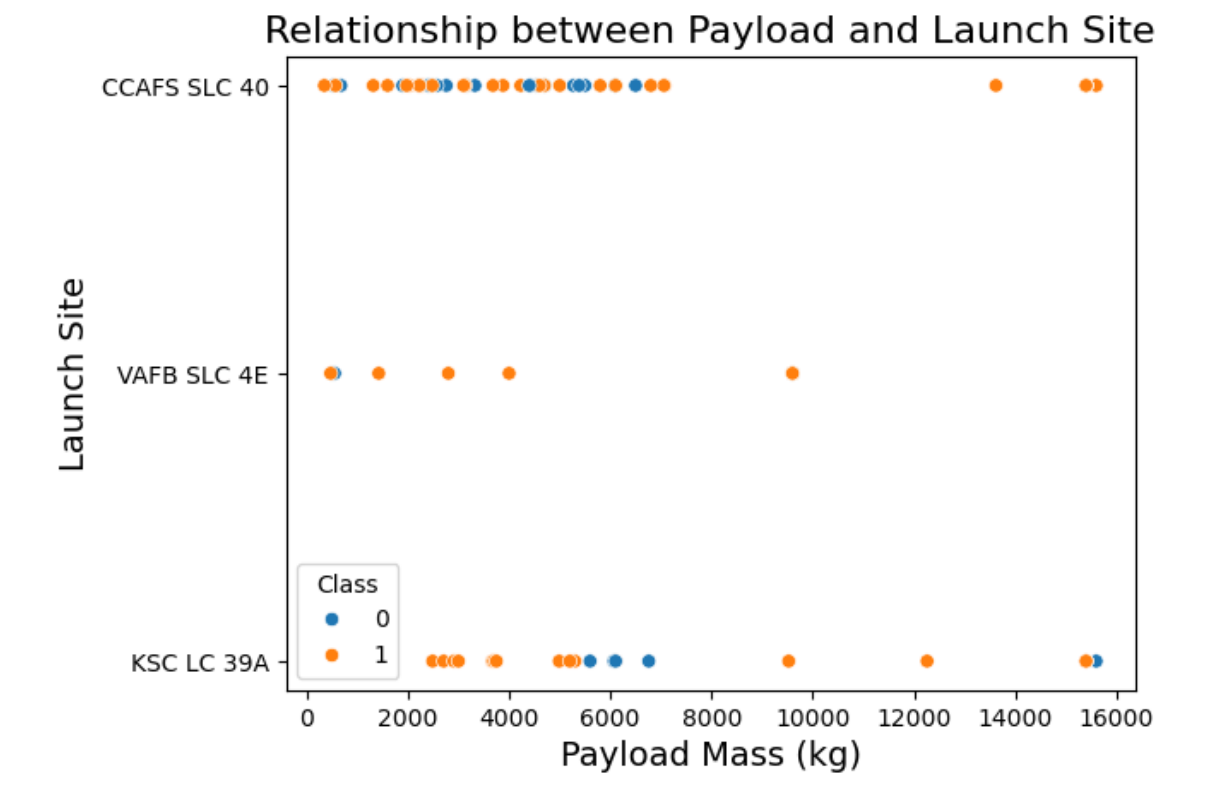

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



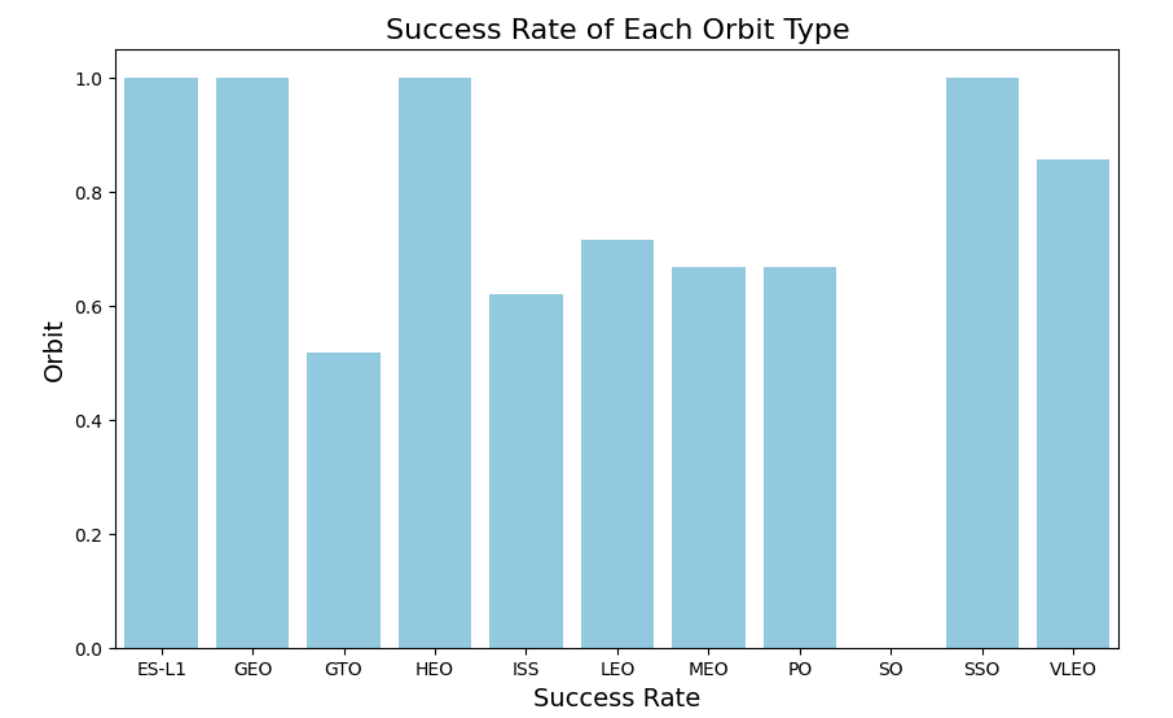
As the Flight Number increases, there is a general trend of more launches occurring over time. The distribution of launches varies across different launch sites, with some sites having a higher success rate than others. The patterns suggest that, over time, there has been an improvement in the success rate of launches at these sites.

Payload vs. Launch Site



VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000)

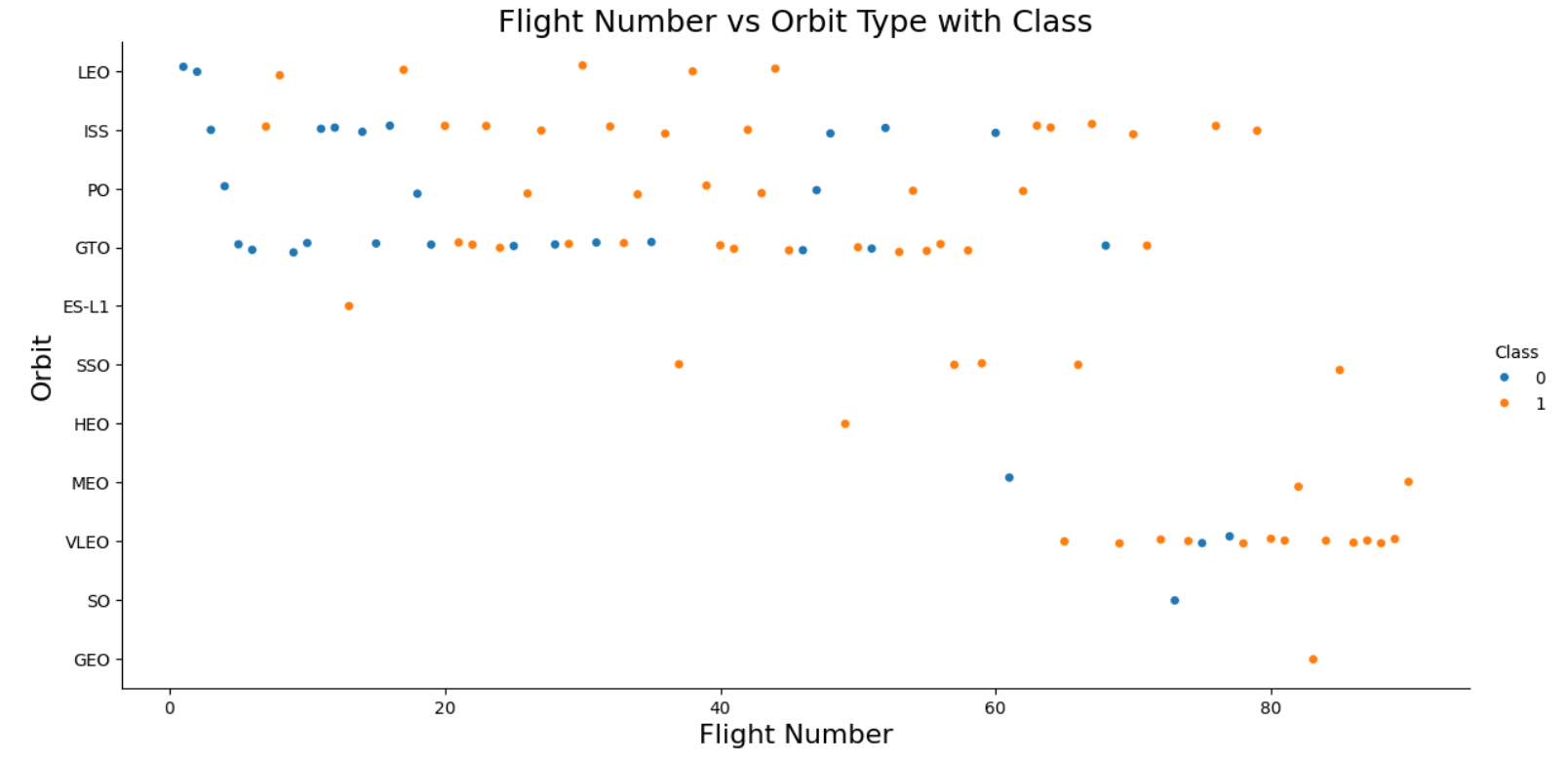
Success Rate vs. Orbit Type



Observations:

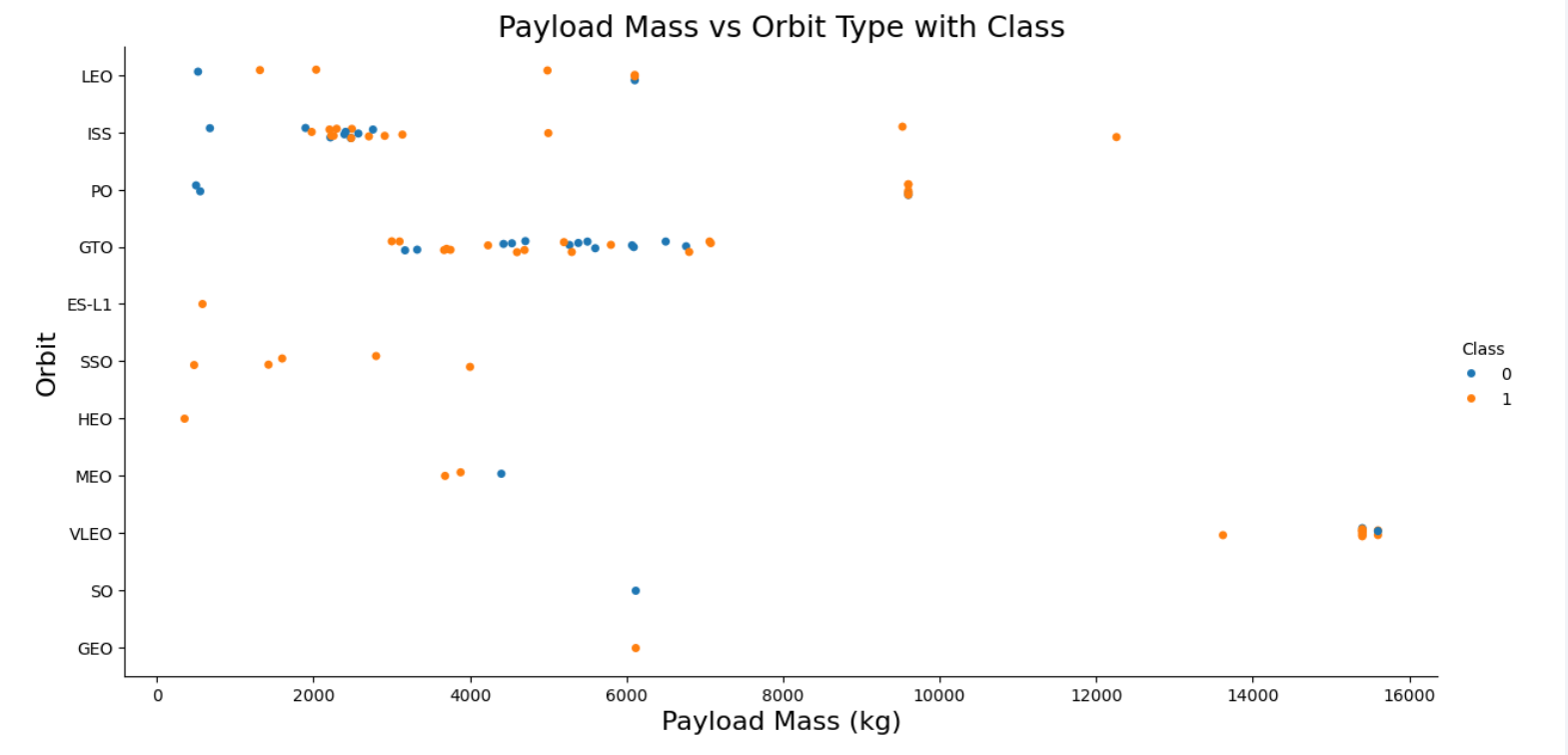
- Orbits related to Earth observation, communication, and space station missions tend to have higher success rates.
- Geostationary orbits (GEO) also show a high success rate.
- Orbits associated with higher-energy trajectories, such as GTO (Geostationary Transfer Orbit), have a lower success rate compared to others.

Flight Number vs. Orbit Type



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

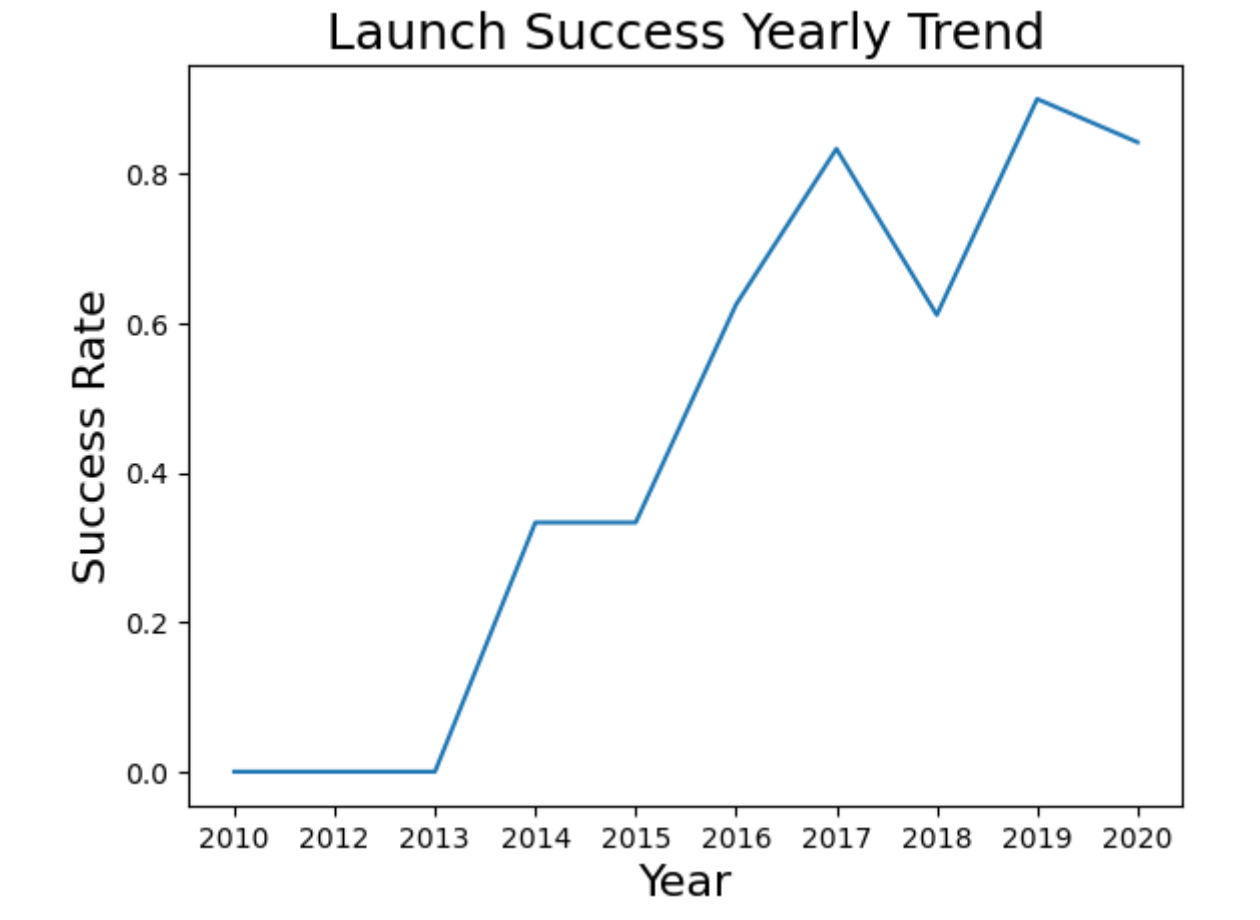
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2020



All Launch Site Names

- We used this query to find the names of the unique launch sites
- We can see the use of select **Distinct** to get this result

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

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

Launch Site Names Begin with 'CCA'

- Using this query, we found 5 records where launch sites begin with `CCA`
- We can see the use of “where … like …” and “limit 5” to get this result

```
%%sql SELECT * FROM SPACEXTABLE
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- We calculated the total payload carried by boosters from NASA
- We can see the use of “where … like …” to get this result

```
%%sql
SELECT Customer, SUM("PAYLOAD_MASS__KG_") as "Total Payload Mass"
FROM SPACEXTABLE
WHERE "Customer" LIKE 'NASA (CRS)';
```

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

Customer	Total Payload Mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1
- We can see the use of AVG to get this result

```
%%sql
SELECT Booster_Version, AVG("PAYLOAD_MASS__KG_") as "Average Payload Mass"
FROM SPACEXTABLE
WHERE "Booster_Version" = 'F9 v1.1';
```

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

Booster_Version	Average Payload Mass
F9 v1.1	2928.4

First Successful Ground Landing Date

- We found the date of the first successful landing outcome on ground pad
- We can see the use of min function to get this result

```
%%sql
SELECT MIN("Date") as "First Successful Landing Date"
FROM SPACEXTABLE
WHERE "Landing_Outcome" LIKE 'Success%'
```

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

First Successful Landing Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- We can see the use of Between to apply the condition

```
%%sql
SELECT "Booster_Version", "PAYLOAD_MASS__KG_"
FROM SPACEXTABLE
WHERE "Landing_Outcome" LIKE 'Success (drone ship)%'
      AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000;
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- We calculated the total number of successful and failure mission outcomes
- To get this result, we used “Count” and “Group by”

```
%%sql
SELECT "Mission_Outcome", COUNT(*) as "Count"
FROM SPACEXTABLE
GROUP BY trim("Mission_Outcome");
```

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

Mission_Outcome	Count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- We listed the names of the boosters which have carried the maximum payload mass
- We use max function to get this result

```
%%sql
SELECT DISTINCT "Booster_Version", "PAYLOAD_MASS__KG_"
FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_")
    FROM SPACEXTABLE
);
```

```
 * 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

- We listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- We used multiple conditions for this query

```
%%sql
SELECT
    SUBSTR("Date", 6, 2) as "Month",
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE SUBSTR("Date", 0, 5) = '2015'
      AND "Landing_Outcome" LIKE 'Failure (drone ship)%';
```

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

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

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

- We ranked 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

- We used multiple conditions and “Group by” for this query

```
%%sql
SELECT
    "Landing_Outcome",
    COUNT(*) as "Count"
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY "Count" DESC;
```

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

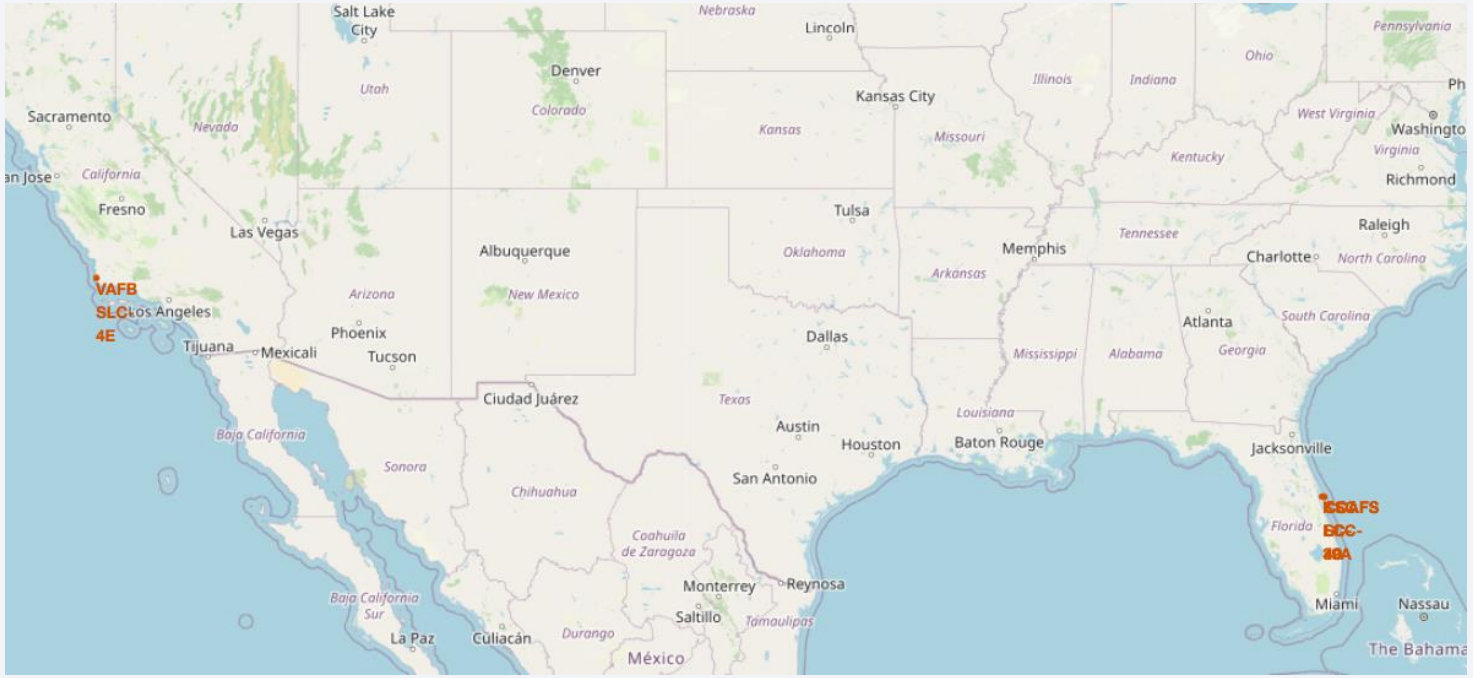
Landing_Outcome	Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

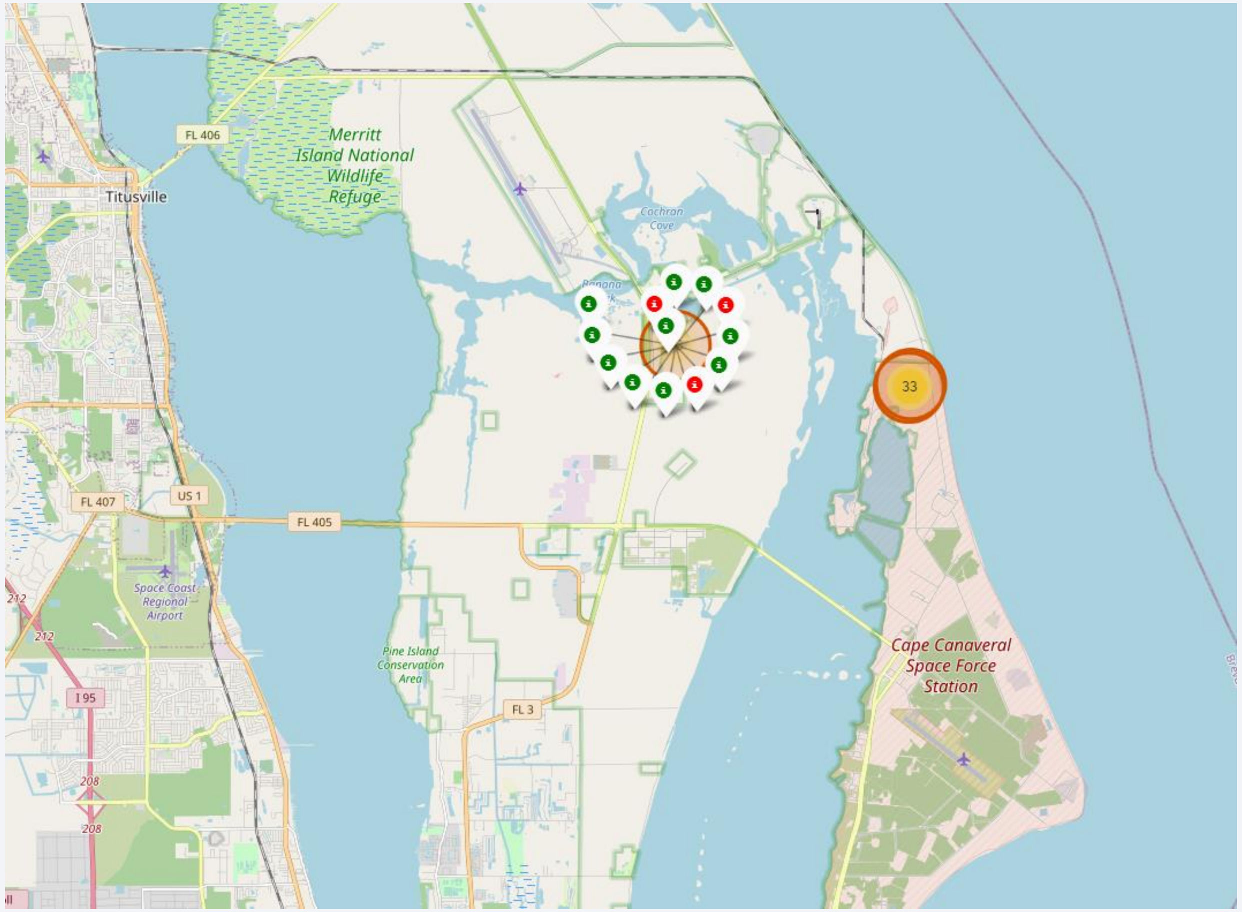
Launch sites' locations on the map

- We created a map with Folium and marked all launch sites' locations on the map



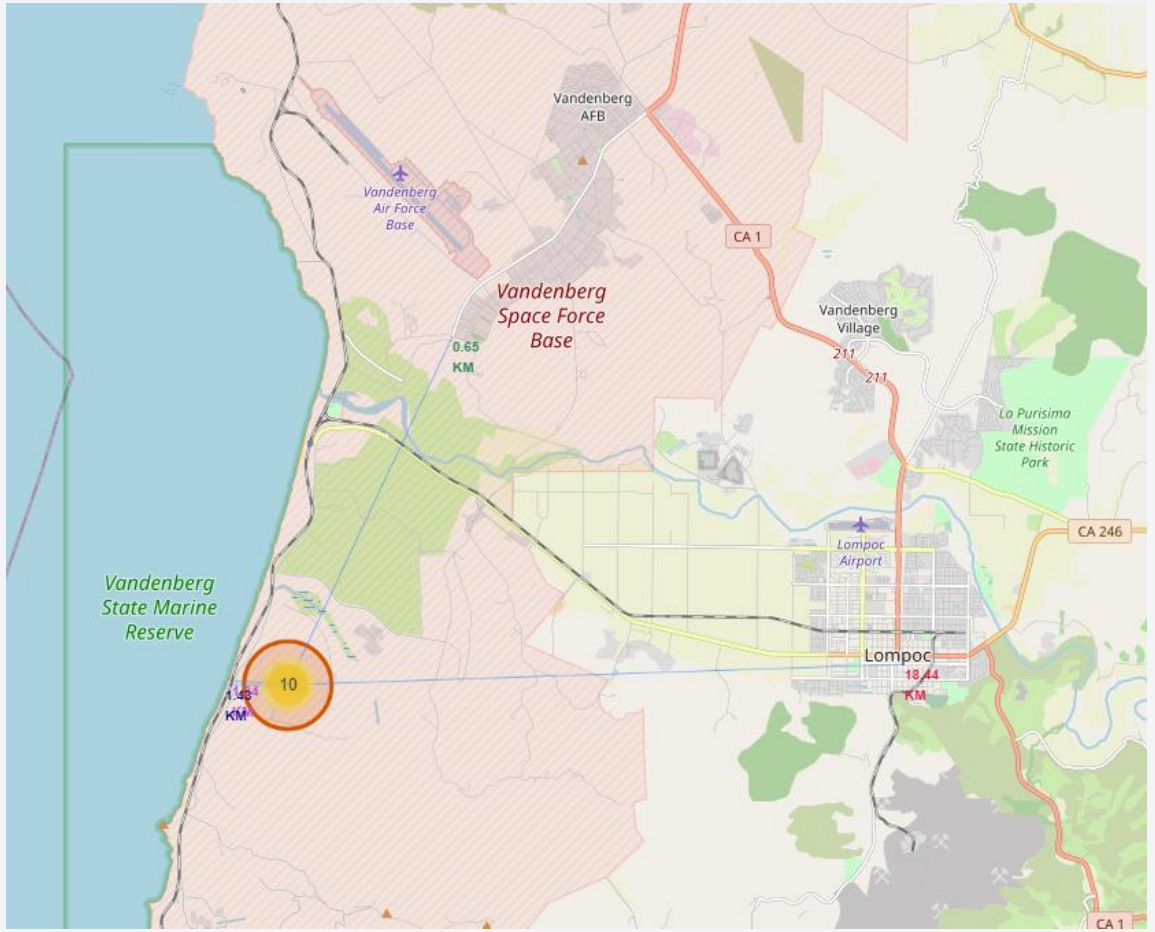
Color-Labeled Launch Outcomes

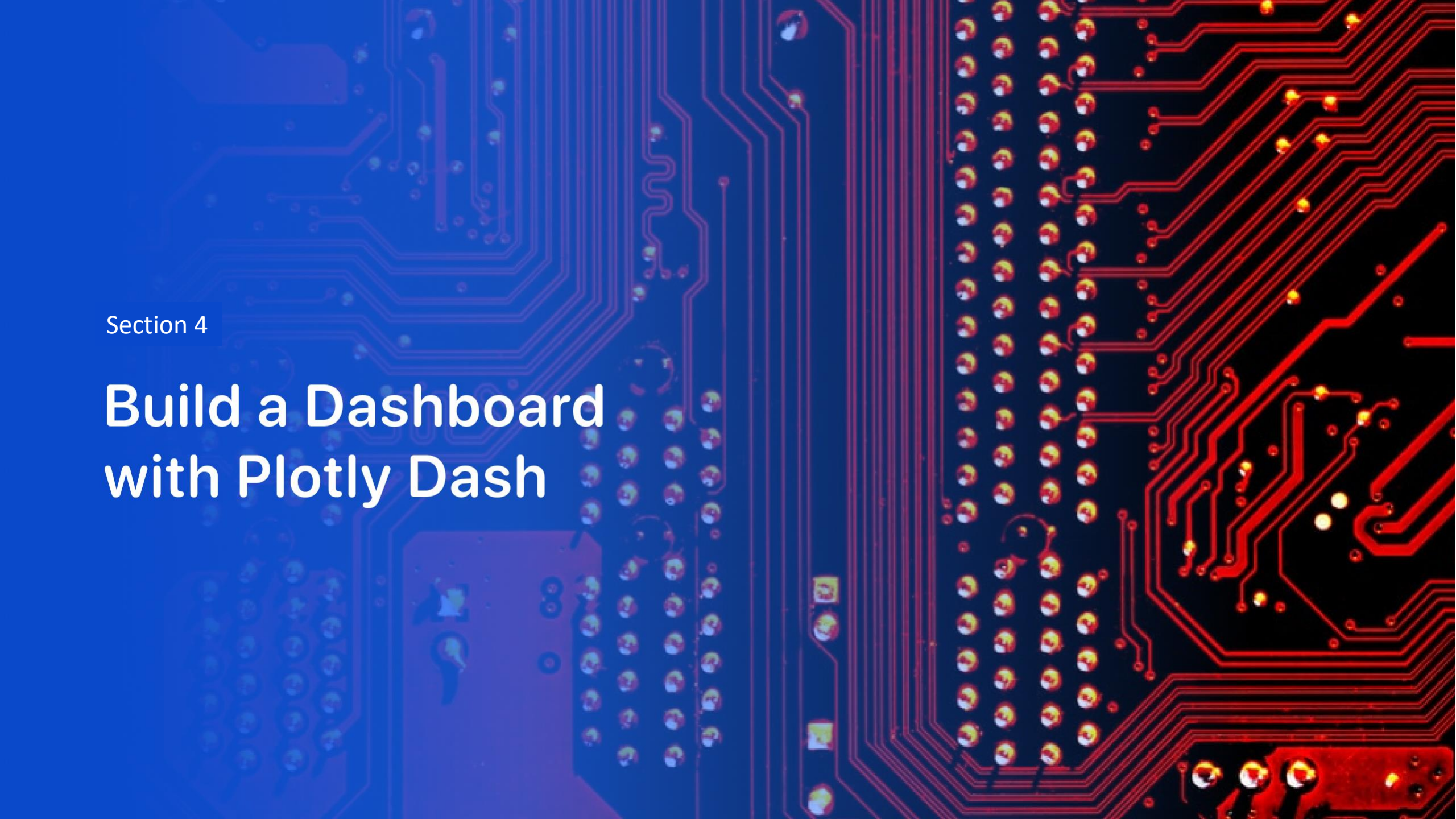
- We updated our map to color-label launch outcomes on the map



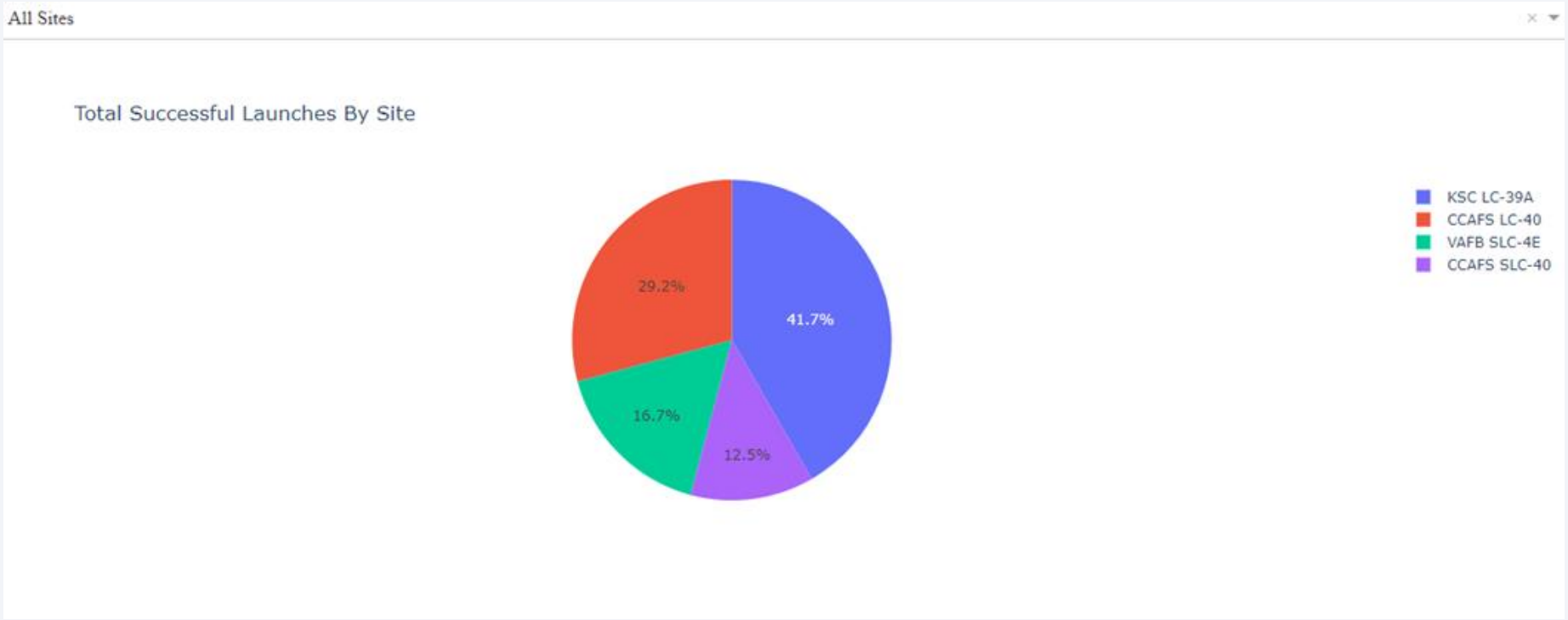
Site distance to its proximities

- We updated the map to draw a PolyLine between a launch site to the selected coastline point



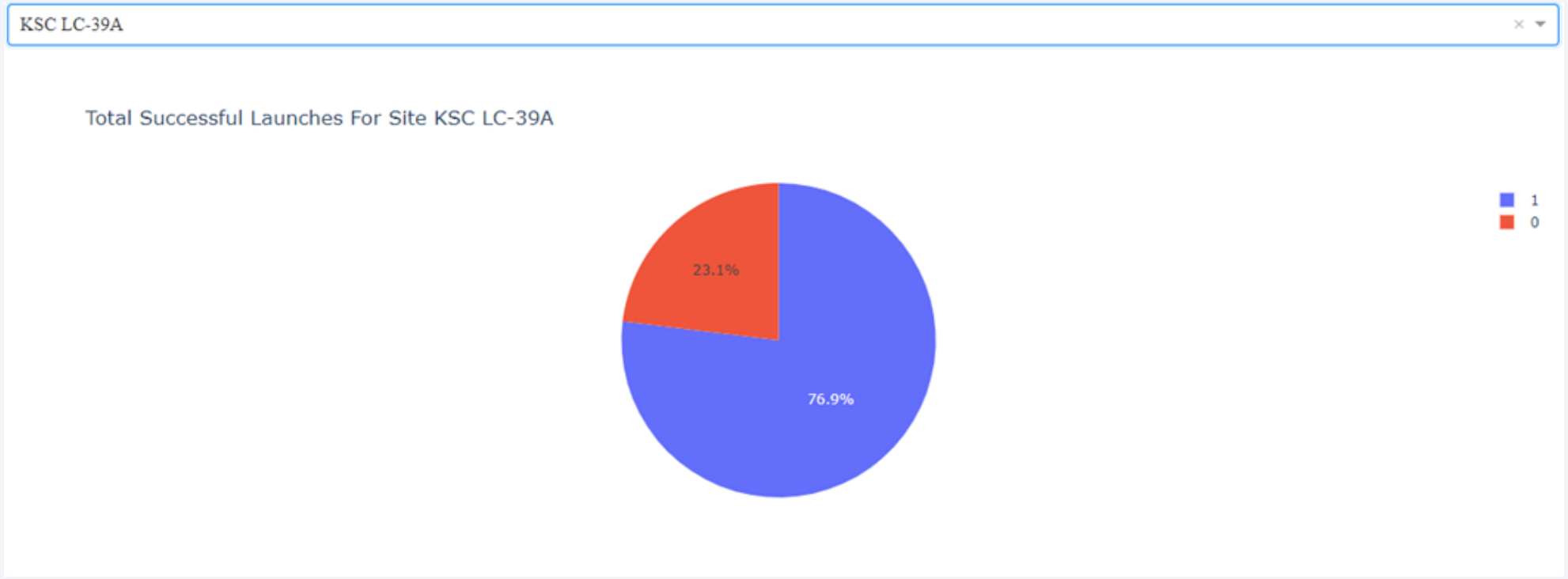

Section 4
Build a Dashboard with Plotly Dash

Successful Lunches by Site



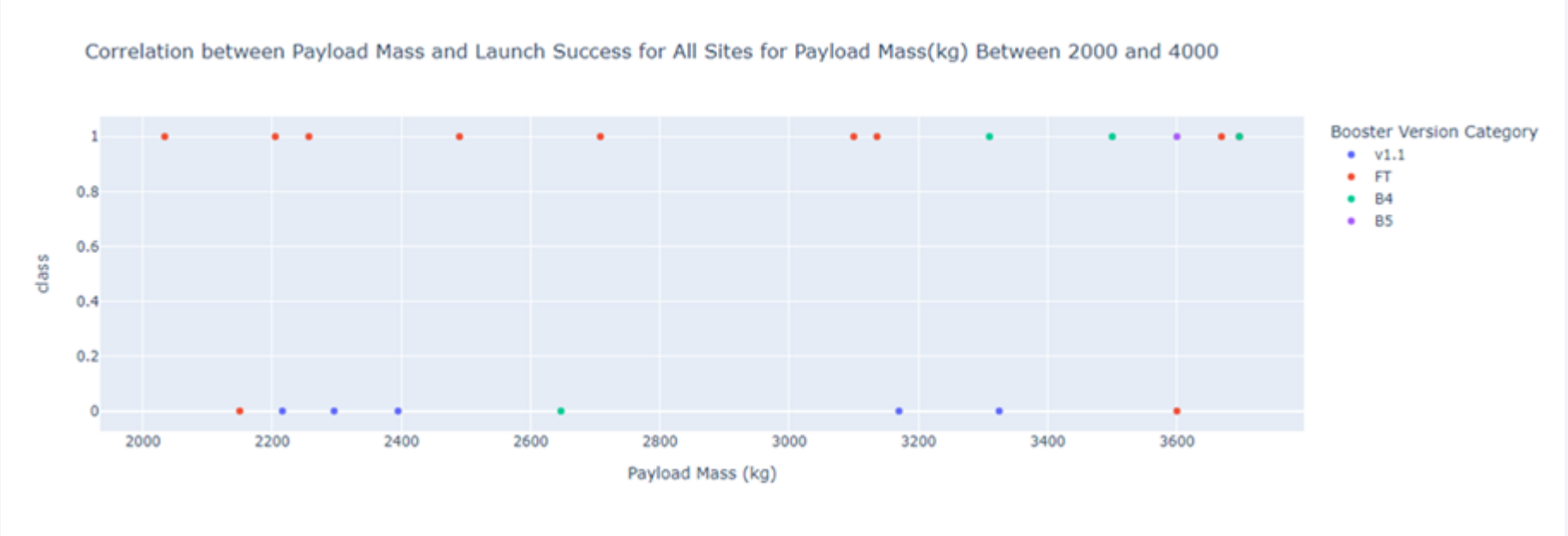
- As we see, KSC LC-39A has the most successful lunches

Success Ratio for the Site with the Highest Lunch



- Success ratio for KSC LC-39A (The site with most successful)lunches

Payload vs. Launch Outcome

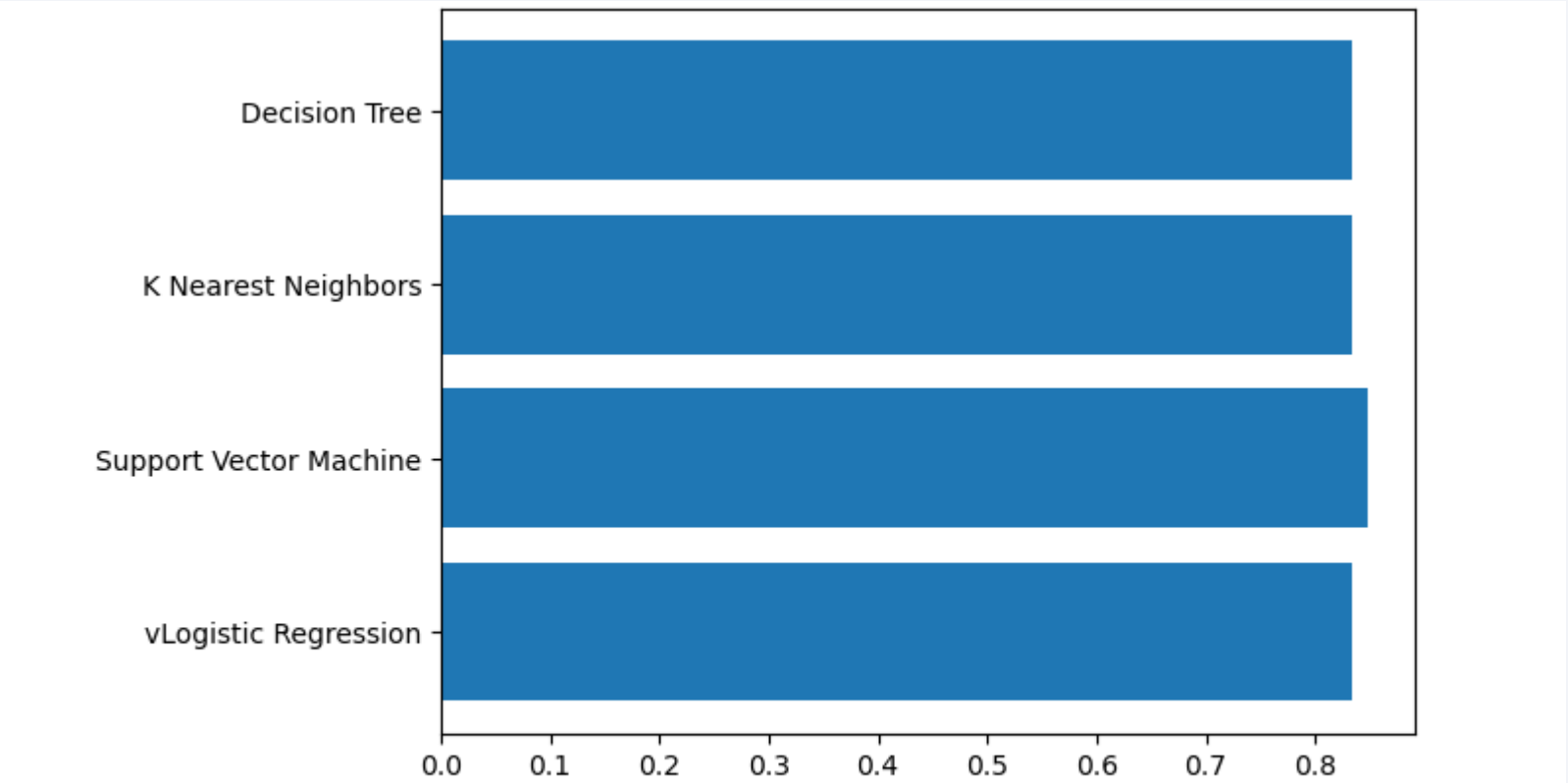


- Correlation between payload Mass and Launch Success

Section 5

Predictive Analysis (Classification)

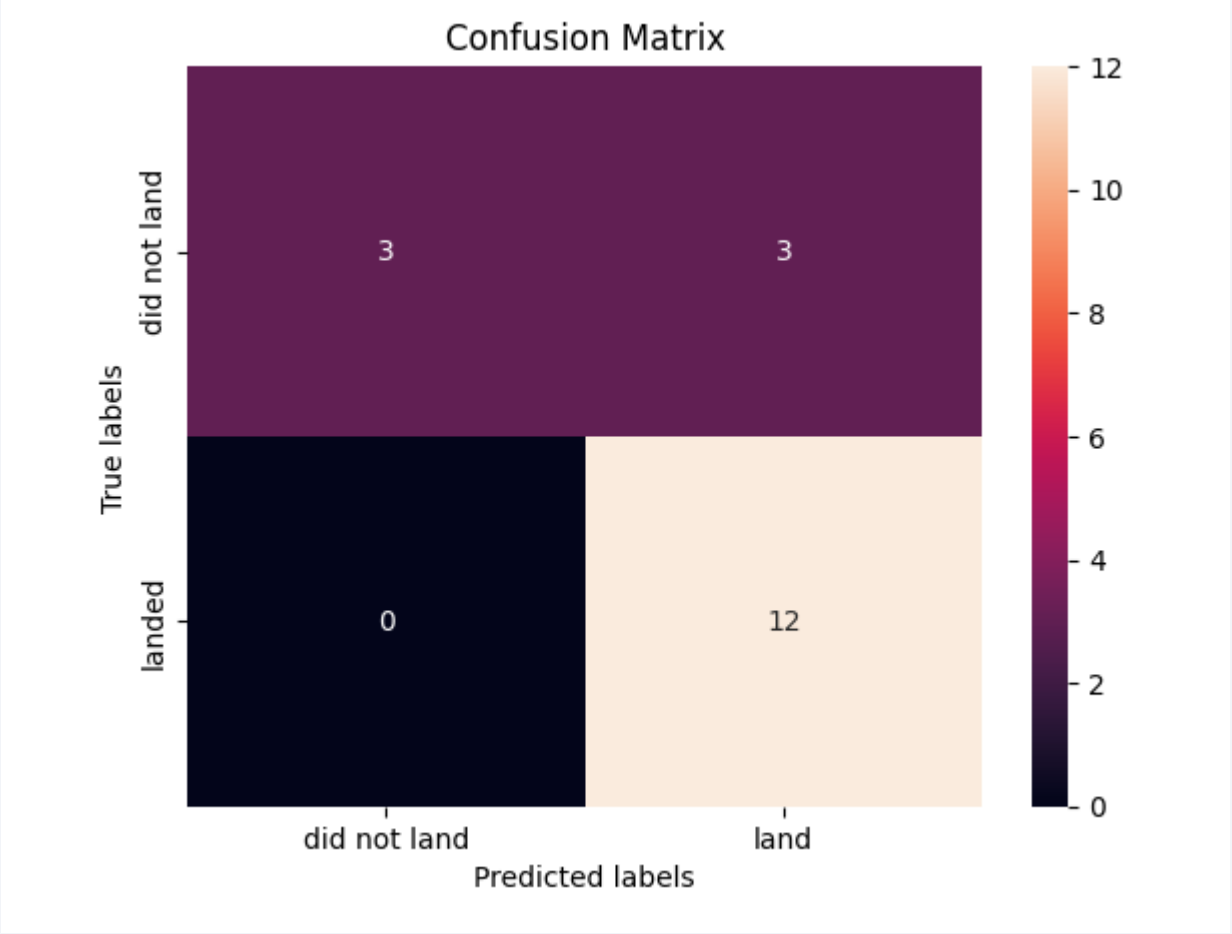
Classification Accuracy



Comparing the accuracy results, we can see support vector machine performs the best.

Confusion Matrix

- The confusion matrix of the Support Vector Machine model.
- As we see, there are three incorrect prediction.



Conclusions

- Model Excellence: Logistic Regression excelled with an 82% accuracy, surpassing SVM, Decision Tree, and KNN. Its selection establishes a reliable basis for decision-making.
- Financial Impact: Cost estimation models revealed substantial cost reduction with successful Falcon 9 landings, emphasizing the practical significance in aerospace decision-making.
- Decision-Making Foundation: Logistic Regression's high accuracy positions it as an optimal predictor, providing a sturdy foundation for strategic planning.
- Holistic Approach: Integrated methodologies offer a comprehensive solution to real-world challenges, with versatile applications in the aerospace industry.
- Industry Relevance: Beyond Falcon 9, the project's methodologies hold broader applications, contributing to efficient resource allocation across the sector.
- Continuous Enhancement: Logistic Regression's success lays the groundwork for ongoing improvements, exploring advanced features and techniques.
- Educational Significance: Serving as an educational resource, the project exemplifies the application of data science in solving real-world challenges.
- Scalability Potential: Developed methodologies offer scalability for similar aerospace challenges, serving as adaptable templates for decision support systems.

Appendix

- Everything related to this project, including this presentation, the Python notebooks, and the data sets are available here:
- https://github.com/msdoust/IBM-DataScience_SpaceX_Landing

Thank you!