IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Carlo De Rossi
February 2025

Outline

Executive Summary

Summary of methodologies

- First Lab: Collect data from the SpaceX API, Ensure data is in the correct format, Export data to a CSV file.
- Second Lab:Perform web scraping to collect Falcon 9 historical launch records from Wikipedia.Export the collected data.
- Third Lab: Conduct Exploratory Data Analysis (EDA) to find patterns in the data. Determine the label for training supervised models. Label data: `1` for successful booster landing, `0` for unsuccessful.
- Fourth Lab: Perform EDA using SQL and the SQLite database.
- Fifth Lab: Conduct EDA and visualization. Invistigate which features are more revelant for missions' success. Perform Feature Engineering to prepare data for prediction.
- Sixth and Seventh Labs: Perform visual analytics.Create an interactive dashboard to show launch locations and their success.
- Eighth and Last Lab: Use collected data to predict if the stage will land based on observations and features.

- Summary of all results
  Decision Tree shows the best accuracy for prediction the mission's success
- This model also shows a higher False Negative rate (1 vs 0) than the other models
- Most successful Launch site is Kennedy Space Center Launch Complex 39A (KSC LC-39A)
- ES-L1 (Earth-Sun Lagrange Point 1), HEO (Highly Elliptical Orbit), and GEO (Geostationary Orbit) have proven to be very successful target orbits for satellite deployments and other space missions.

Introduction

Project Background and Context:

SpaceX has revolutionized the space launch industry with its Falcon 9 rocket, offering launches at $62 million compared to competitors' prices of $165 million or more. This remarkable cost advantage stems primarily from SpaceX's ability to reuse the rocket's first stage, which returns to Earth and lands after launch. This innovation has fundamentally changed the economics of space access.

Problem Statement:

The central challenge is to develop a predictive model that can determine whether a Falcon 9 rocket's first stage will successfully land after launch. This prediction capability is valuable because the success or failure of first stage recovery directly impacts the launch cost. By accurately predicting landing success, we can help estimate the true cost of SpaceX launches.

Key Questions to Address:

We aim to develop a reliable Prediction Model (based on available features) that can predict the successful landing of the first stage. This analysis would be particularly valuable for competing space launch companies trying to understand SpaceX's cost structure and formulate competitive bids against them in the launch services market.

Business Impact:

The insights from this analysis could help other companies make informed decisions when competing with SpaceX for launch contracts. Understanding the probability of successful first stage recovery, and thus the actual cost structure of SpaceX launches, would enable more strategic pricing and bidding strategies in the commercial space launch market.

Section 1

Methodology

Methodology

Executive Summary

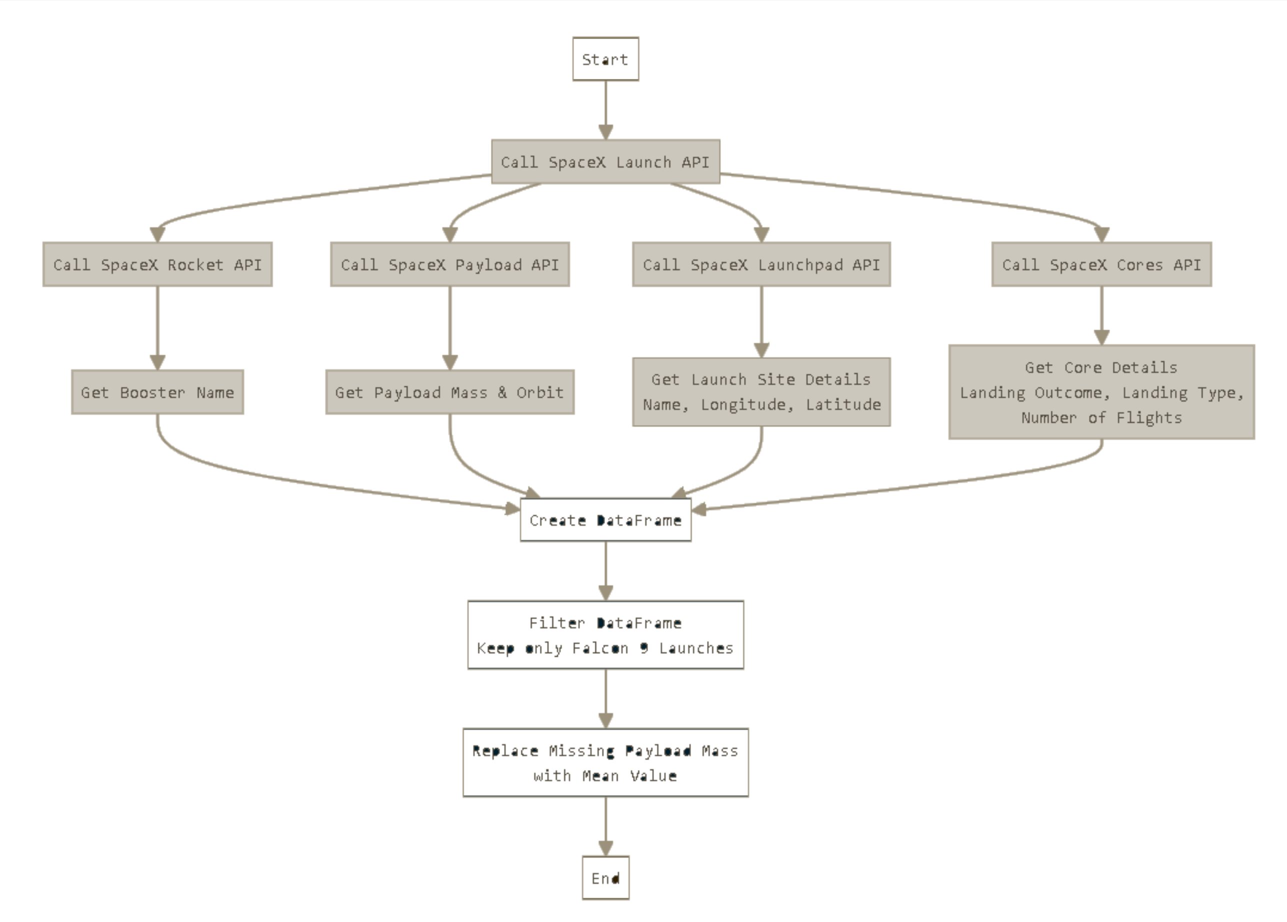
- Data collection methodology:
  - For collecting data both SpaceX REST APIs and Wikipedia WebScraping were used
- Perform data wrangling
  - Data clean up and convert launch outcomes into Training Labels (`1` means the booster successfully landed `0` means it was unsuccessful).
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Prepare Train and Test Datasets to train and validate different classification algorithms
  - Evaluate and compare the algorithms' prediction accuracy

Data Collection

- Collect the launch related data from 2 different sources:
    - The different SpaceX APIs
    - Web scraping the Wikipedia page
- After the initial collection is done
    - Create a data frame
    - Keep only the data related to the 'Falcon 9' launches
    - Replace missing payload mass with the mean values
    - Export the data into a CSV file

Data Collection – SpaceX API

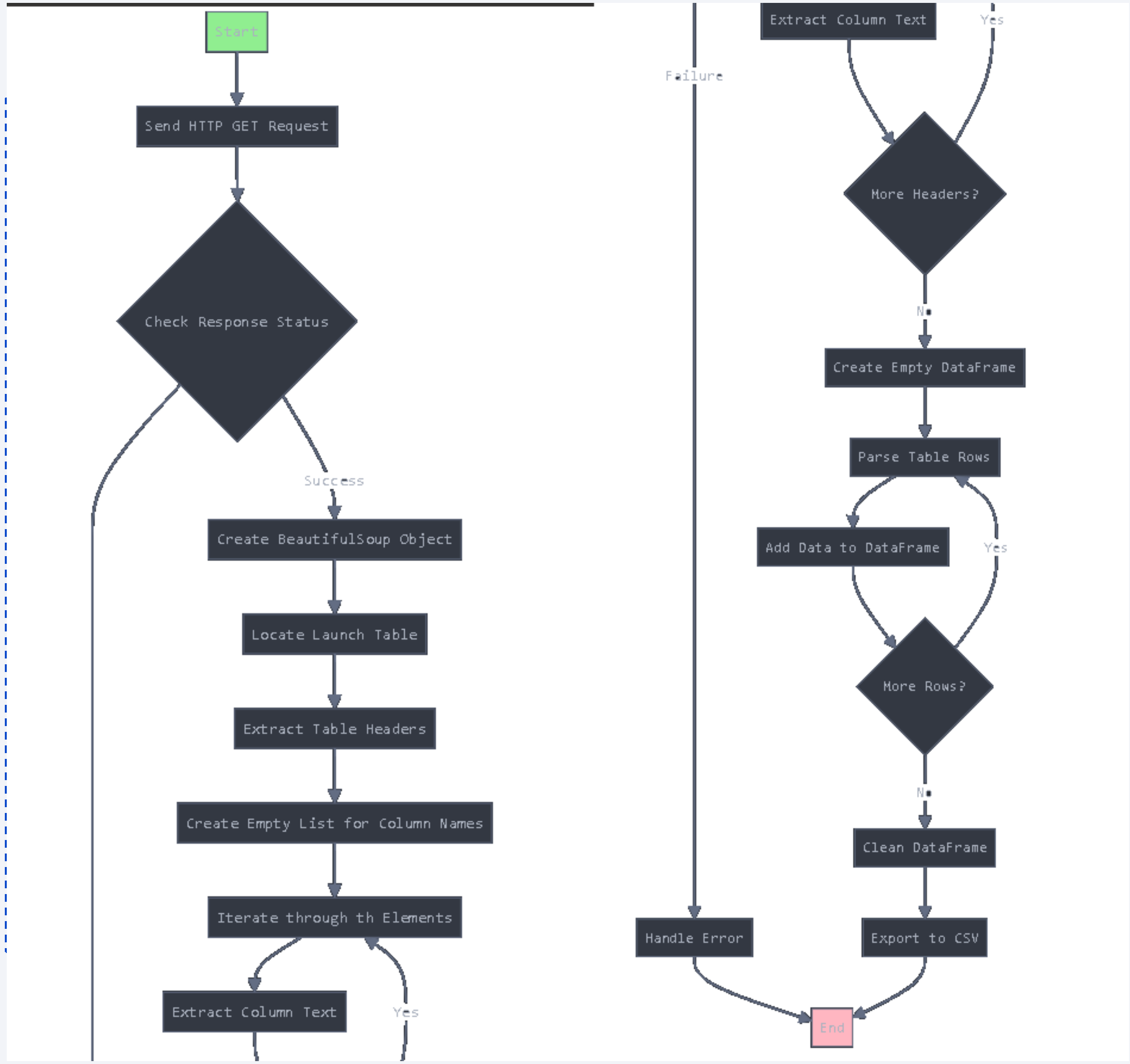
- Use the different SpaceX APIs to collect relevant data for each launch and construct a DataFrame
- Filter only data relevant to "Falcon 9" related launches
- Clean up missing values, replacing them with the mean value
- Export to CSV file (End process)
- See also: [CDR GitHub - jupyter-labs-spacex-data-collection-api.ipynb]()



Data Collection - Scraping

- perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response
- Create a BeautifulSoup object from the HTML response
- collect all relevant column names from the HTML table header
- iterate through the <th> elements and extract column name one by one
- Create a data frame by parsing the launch HTML tables
- Export the data as CSV file

- See also: [CDR GitHub - jupyter-labs-webscraping.ipynb]()



Data Wrangling

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome column

- See also: CDR GitHub - labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Analyzed the different impact of success of a launch for
  Orbits
  Booster versions
  Payload Mass
  Number of Flight
  Launch Sites
- And show a clear trend of improvement during the years, with minor setbacks

- See also: [CDR GitHub - edadataviz.ipynb]

EDA with SQL

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

- See also: CDR GitHub - jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Visualize on a Folium map the location of the launch centers, and the number of launches successful (green) or failed (red) and their combination (resulting in a yellow circle)
- Measured and displayed the distance of launch centers from coastal lines, cities and highways, showing that best location for the launch centers are near coastal lines and away from cities, highways, etc.
- See also: CDR - GitHub - jupyther-labs-launch_site_locations.ipynb

Build a Dashboard with Plotly Dash

- Display Highest number of success launch counts (KSC LC-39A) and lowest rate of success launches (CCAFS SLC-40) per launch center as a pie chart
- For the most successful launch center display the percentage of success
- Display graph showing that B4 and FT Booster versions are the only ones successfully launching higher payload masses
- The visualization where created for extracting this useful information from the data
- See also: CDR GitHub - SpaceX Dash App

Predictive Analysis (Classification)

Model Building

- Model Selection: Chose several classification algorithms including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).

- Feature Scaling: Applied `StandardScaler` to standardize the numeric features for better model performance.

- Model Training: Trained each model on the training dataset using the `fit` method.

Model Evaluation

- Accuracy Measurement: Measured the accuracy of each model on the test dataset to compare their performance.

- Confusion Matrix Analysis: Evaluated the models using confusion matrices to understand the True Positives, True Negatives, False Positives, and False Negatives.

Identifying the Best Model

- Accuracy Comparison: Compared the accuracy scores of different models.

- Decision Tree Excellence: Identified the Decision Tree model as the best-performing model with the highest accuracy and balanced error rates.

- False Positives vs. False Negatives: Prefer False Positives over False Negatives to avoid missing out on recovering rockets, which is crucial for cost efficiency.

Visualization and Interpretation

- Accuracy Visualization: Created a bar plot to visualize the accuracy of different models with rotated x-axis labels for better readability.

- Model Interpretation: Explained why the Decision Tree model is preferred, considering its high accuracy, balanced performance, and ability to capture complex patterns.

See also: [CDR GitHub - SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb]

Results

- Exploratory data analysis results
  The APIs are more effective and less '*cumbersome*' for retrieving data
  The WebScraping of Wikipedia, while interesting for NLP projects, does not seems reliable or efficient as a method
  DataSet is mostly clear with minimal clean up and feature engineering needed

- Interactive analytics demo in screenshots
  The visualization shows clearly the most successful version and type of boosters, as well as of the launch center. They also give an interesting insight about most successful orbits and the role that the featires play in the success or not of the missions

- Predictive analysis results
  Clearly the Decision Tree is the prediction method that shows the best accuracy, whereas the less accurate model do not deliver any False Negative

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

1. **CCAFS SLC 40**:

- There's a mix of successful (Class 1) and unsuccessful (Class 0) launches across a wide range of flight numbers. This indicates variability in launch outcomes at this site.

- As the flight numbers increase, there appears to be a tendency towards more successful launches, suggesting improvements over time.

2. **VAFB SLC 4E**:

- There are fewer data points compared to other sites, but both successful and unsuccessful launches are present.

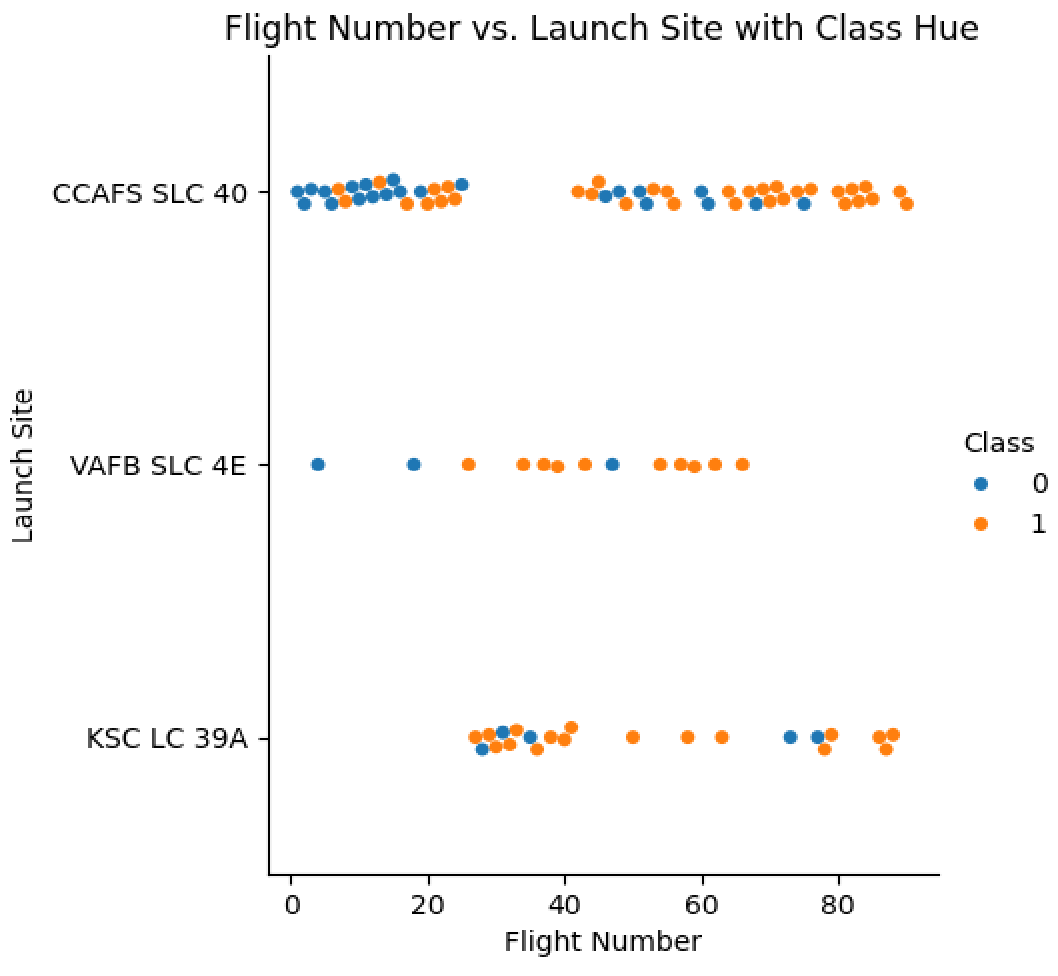
- The limited number of flights makes it harder to discern clear trends, but there is still a noticeable mix of outcomes.

3. **KSC LC 39A**:

- Similar to CCAFS SLC 40, there are clusters of both successful and unsuccessful launches across different flight numbers.

- Higher flight numbers generally show more successful launches, indicating progress in reliability and success rates over time.

Overall, the scatter plot illustrates a mix of outcomes across all launch sites, with a general trend of **increasing success rates as flight numbers grow**. This likely reflects technological advancements and experience gained over time.



Payload vs. Launch Site

1. **CCAFS SLC 40** (Cape Canaveral Space Force Station Space Launch Complex 40):

   - Payload masses are mostly clustered between 0 and 7500 kg, with a few outliers around 10000 kg and 15000 kg.

   - Both successful (Class 1) and unsuccessful (Class 0) payloads are present in this range.

2. **VAFB SLC 4E** (Vandenberg Space Force Base Space Launch Complex 4 East):

   - Payload masses are primarily between 0 and 5000 kg, with a significant number of payloads around 10000 kg.

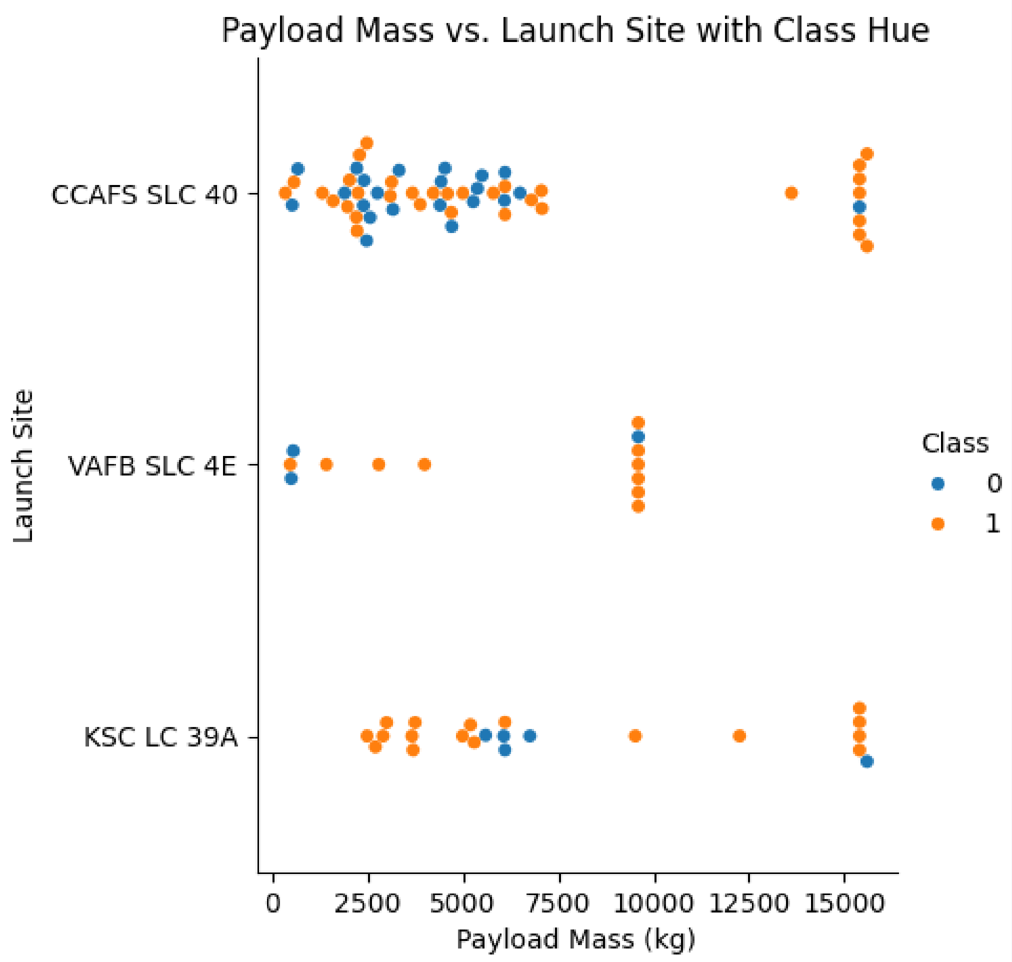
   - All payloads in the 10000 kg range are classified as successful (Class 1).

3. **KSC LC 39A** (Kennedy Space Center Launch Complex 39A):

   - Payload masses are distributed between 0 and 15000 kg, with noticeable clusters around 5000 kg and 15000 kg.

   - Both successful (Class 1) and unsuccessful (Class 0) payloads are present in these clusters.

In summary, the scatter plot highlights the distribution of payload masses across different launch sites and shows the success (Class 1) or failure (Class 0) of the launches. This information is useful for analyzing the performance and reliability of launches from different sites based on payload mass.



Success Rate vs. Orbit Type

Based on the Bar chart, the most successful orbits, with a success rate of 1.0, are:

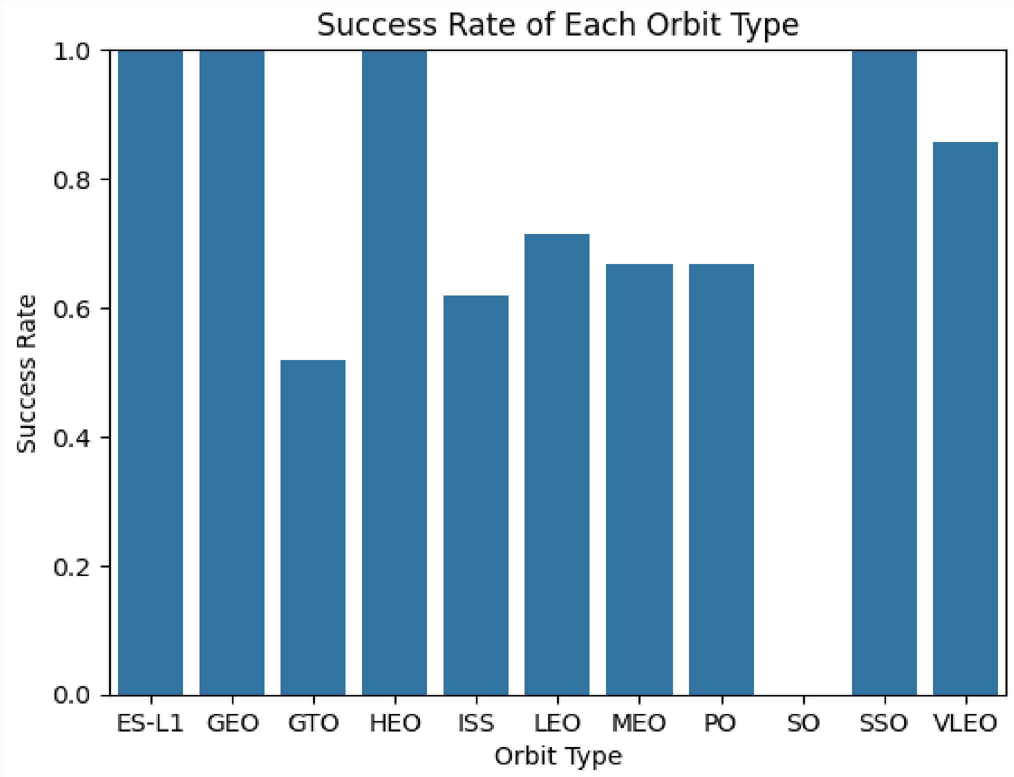
ES-L1 (Lagrange Point): Located between the Sun and Earth, this orbit has shown consistent success.

GEO (Geosynchronous Orbit): This high Earth orbit allows satellites to match Earth's rotation.

SSO (Sun-synchronous Orbit): A nearly polar orbit around a planet, also known for its consistent success rate.

HEO (Highly Elliptical Orbits): An elliptical orbit, with high eccentricity

These orbits have demonstrated a perfect success rate, making them the most reliable according to the data presented.



Flight Number vs. Orbit Type

1. LEO (Low Earth Orbit): The LEO category shows a mix of both Class 0 and Class 1 flights, with flight numbers distributed across the entire range, indicating variability in success rates.
2. ISS (International Space Station): Similar to LEO, ISS has both successful and unsuccessful launches spread over different flight numbers.
3. PO (Polar Orbit): There are fewer data points for PO, with both classes represented, showing some level of variability.
4. GTO (Geostationary Transfer Orbit): GTO has a dense cluster of data points with both classes, suggesting a significant number of launches with mixed outcomes.
5. ES-L1 (Earth-Sun Lagrange Point 1): This category has only one data point, which is classified as a successful launch (Class 1).
6. SSO (Sun-Synchronous Orbit): Most data points in SSO are classified as successful (Class 1).
7. HEO (Highly Elliptical Orbit): HEO only has Class 1 data points, indicating all recorded launches in this category were successful.
8. MEO (Medium Earth Orbit): There is a single data point in MEO, which is an unsuccessful launch (Class 0).
9. VLEO (Very Low Earth Orbit): VLEO has a dense cluster of data points with both classes, showing mixed outcomes.
10. SO (Sub-Orbital): SO has a few data points, with both classes represented.
11. GEO (Geostationary Orbit): GEO has a single data point, which is a successful launch (Class 1).

In summary, we can observe that certain orbits like **ES-L1**, **HEO**, and **GEO** have shown consistent success. On the other hand, orbits like **LEO**, **ISS**, and **GTO** have mixed outcomes, suggesting variability in launch success rates.

Payload vs. Orbit Type

1. LEO (Low Earth Orbit): This orbit type shows a wide range of payload masses, from small to very large payloads, with a mix of Class 0 and Class 1. Successful (Class 1) and unsuccessful (Class 0) launches are both represented across different payload masses.

2. ISS (International Space Station): Payloads destined for the ISS also exhibit a variety of masses, mostly successful (Class 1). There is a higher concentration of successful launches in the mid to higher payload mass range.

3. GTO (Geostationary Transfer Orbit): This orbit type has a diverse range of payload masses, with a significant number of successful launches (Class 1). There are some unsuccessful launches (Class 0), but the majority appear to be successful.

4. SSO (Sun-Synchronous Orbit): SSO has mainly small to medium payload masses, with most launches being successful (Class 1). There are a few unsuccessful launches (Class 0) as well.

5. HEO (Highly Elliptical Orbit): All payloads for HEO are represented as successful (Class 1), indicating consistent success for this orbit type.

6. VLEO (Very Low Earth Orbit): This category shows a range of payload masses with both successful (Class 1) and unsuccessful (Class 0) launches.

7. PO (Polar Orbit): PO has a few payloads with both successful (Class 1) and unsuccessful (Class 0) launches, indicating variability in success.

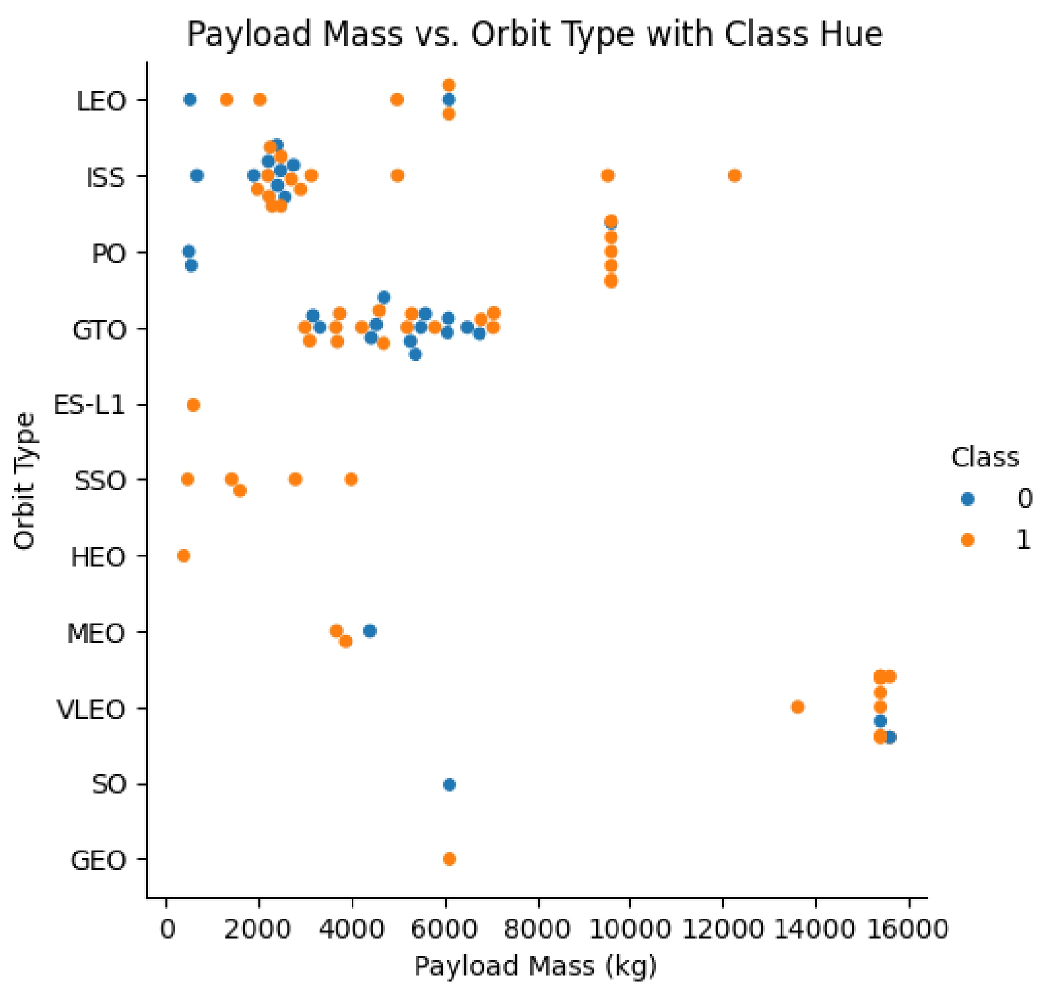
8. ES-L1 (Earth-Sun Lagrange Point 1): There is only one data point for ES-L1, which is successful (Class 1).

9. MEO (Medium Earth Orbit): MEO has one data point that is an unsuccessful launch (Class 0).

10. SO (Sub-Orbital): There are a few data points for SO, with both successful and unsuccessful launches.

11. GEO (Geostationary Orbit): GEO has one data point, which is a successful launch (Class 1).

In summary, different orbit types show a variety of payload masses and success rates. Orbits like **HEO** and **GEO** have consistent success in the given data, while orbits like **LEO, ISS,** and **GTO** show a mix of outcomes.



Launch Success Yearly Trend

1. 2010 to 2013: The success rate remained at 0.0, indicating no successful launches during these years.

2. 2014: There was a significant increase in the success rate to approximately 0.4, marking the beginning of successful launches.

3. 2014 to 2016: A steady increase in success rate, reaching around 0.7 by 2016, suggesting continuous improvements in launch technology or procedures.

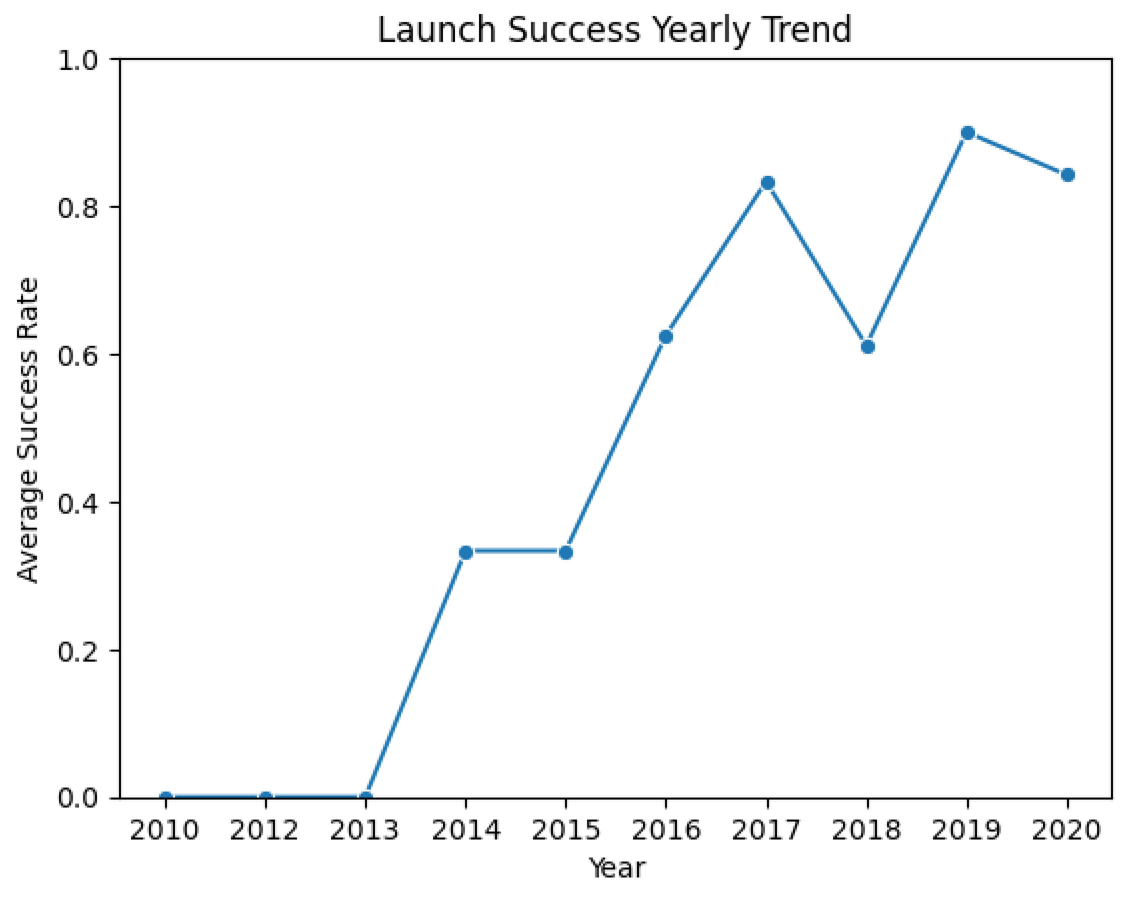
4. 2017: The success rate peaked at around 0.9, indicating a very high success rate for launches.

5. 2018: A drop in success rate to about 0.6, which could be due to specific challenges or anomalies encountered that year.

6. 2019: The success rate increased again to around 0.85, showing recovery and continued improvements.

7. 2020: A slight decrease to around 0.8, indicating some variability but maintaining a high success rate overall.

This graph highlights the progress (and small fluctuations) in launch success rates over a decade, showcasing **significant improvements** with occasional setbacks.



All Launch Site Names

The unique launch sites are:

1. CCAFS SLC-40 (Cape Canaveral Air Force Station Space Launch Complex 40). Location: Cape Canaveral, Florida, USA.

2. VAFB SLC-4E (Vandenberg Air Force Base Space Launch Complex 4 East). Location: Vandenberg Space Force Base, California, USA.

3. KSC LC-39A (Kennedy Space Center Launch Complex 39A). Location: Merritt Island, Florida, USA.

4. CCAFS SLC-40 (Cape Canaveral Space Force Station Space Launch Complex 40). Location: Cape Canaveral, Florida, USA.

```
%%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'

None of the 5 attemps were successful.

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

[15]

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Selecting only the Nasa Customer, we calculate the sum of the payloads' mass in KG.

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

[17]

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE
WHERE Customer LIKE 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Using as condition the Booster Version to be 'F9 v1.1' we select the average of the Payload mass in KG

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

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

[18]

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Using as a condition the Landing Outcome to be successful on ground pad, we select the minimal value for the corresponding dates

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

Hint:Use min function

[19]

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

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Using as a conditions

- Landing Outcomes successful on drone ship
- Payload mass greater than 4000 but less than 6000

We select the corresponding Booster Version

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
  AND PAYLOAD_MASS__KG_ > 4000
  AND PAYLOAD_MASS__KG_ < 6000;
```

[29]

```
 * 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

Grouping the results by mission outcomes, we count the total number of successful and failure

List the total number of successful and failure mission outcomes

[22]

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

```

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

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

Boosters Carried Maximum Payload

Using a sub-query for selecting the maximum payload mass,

We select the corresponding names of the booster

[23]

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

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

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

2015 Launch Records

Using as conditions

Landing Outcomes equals to Failure in drone ship

And year 2015

We select the corresponding month, booster versions, and launch site names

[24]

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

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

Month	Landing_Outcome	Booster_Version	Launch_Site
01	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

Using as condition the date between 2010-06-04 and 2017-03-20, we select the count of the landings outcomes , grouping them and sorting them in descending order

```
%%sql

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

```

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

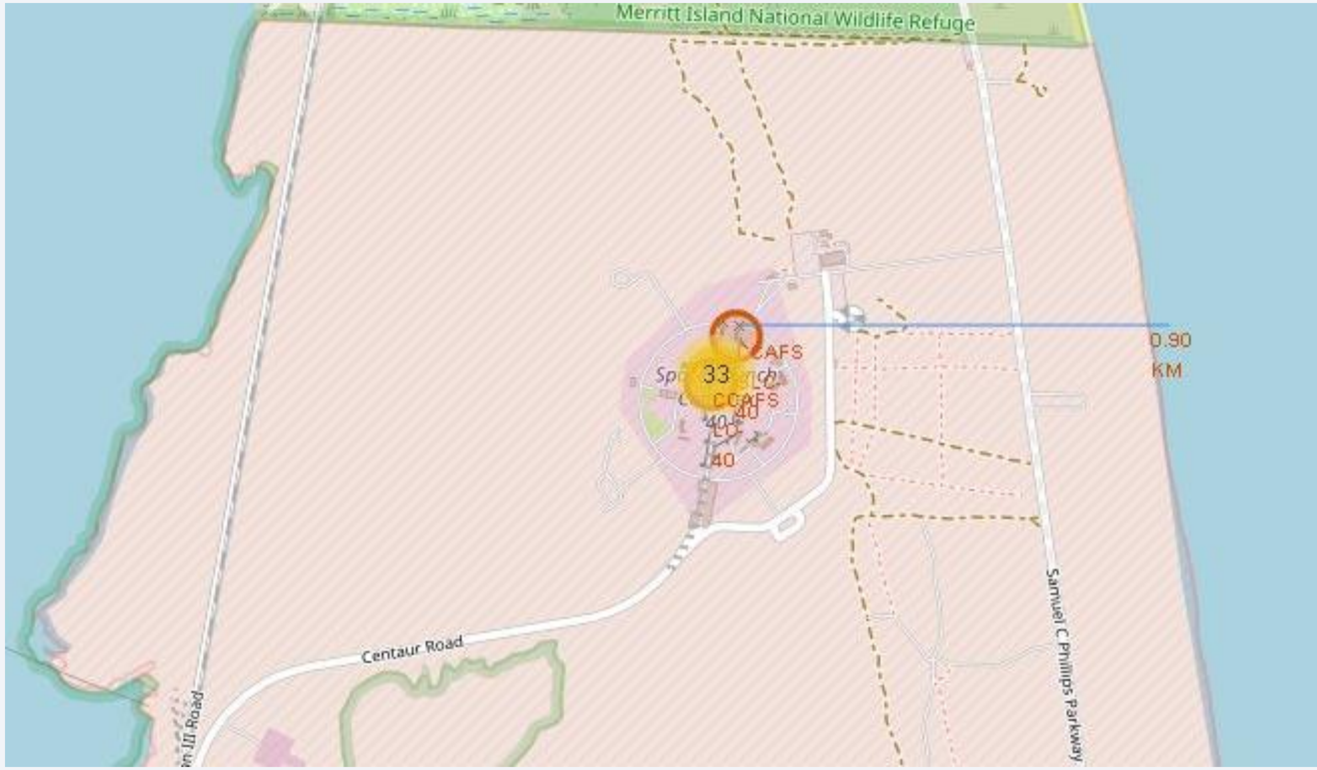
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

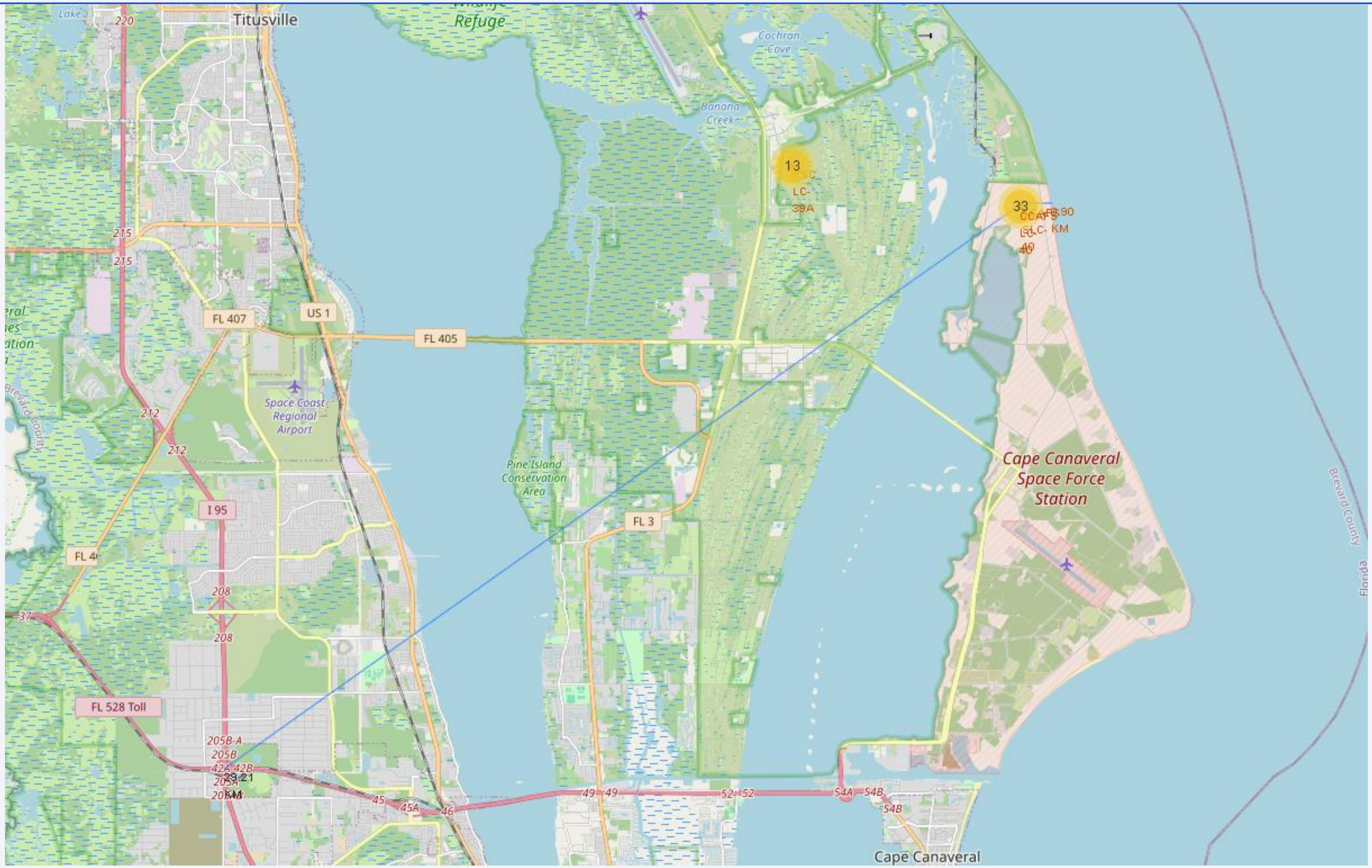
Launch Sites Proximities Analysis

Cape Canaveral Distance from CoastLine and Vanderberg

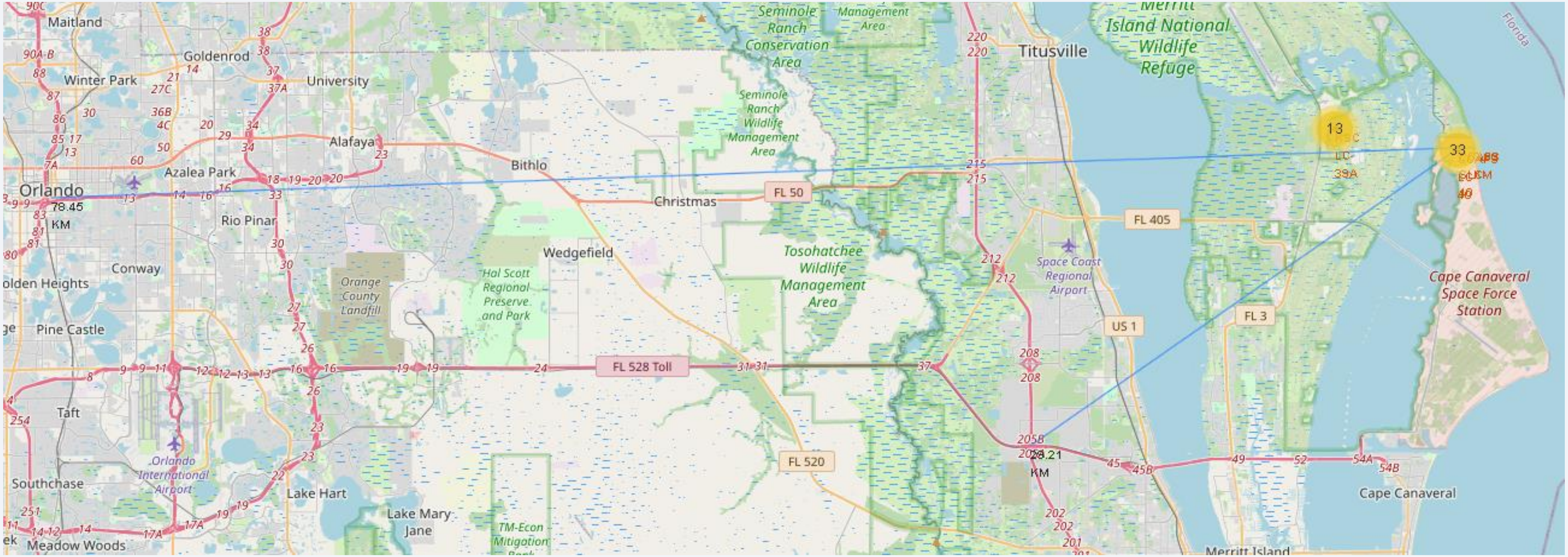
- Cape Canaveral distance from Coast Line



Cape Canaveral Distance from Highway



Cape Canaveral distance from Orlando

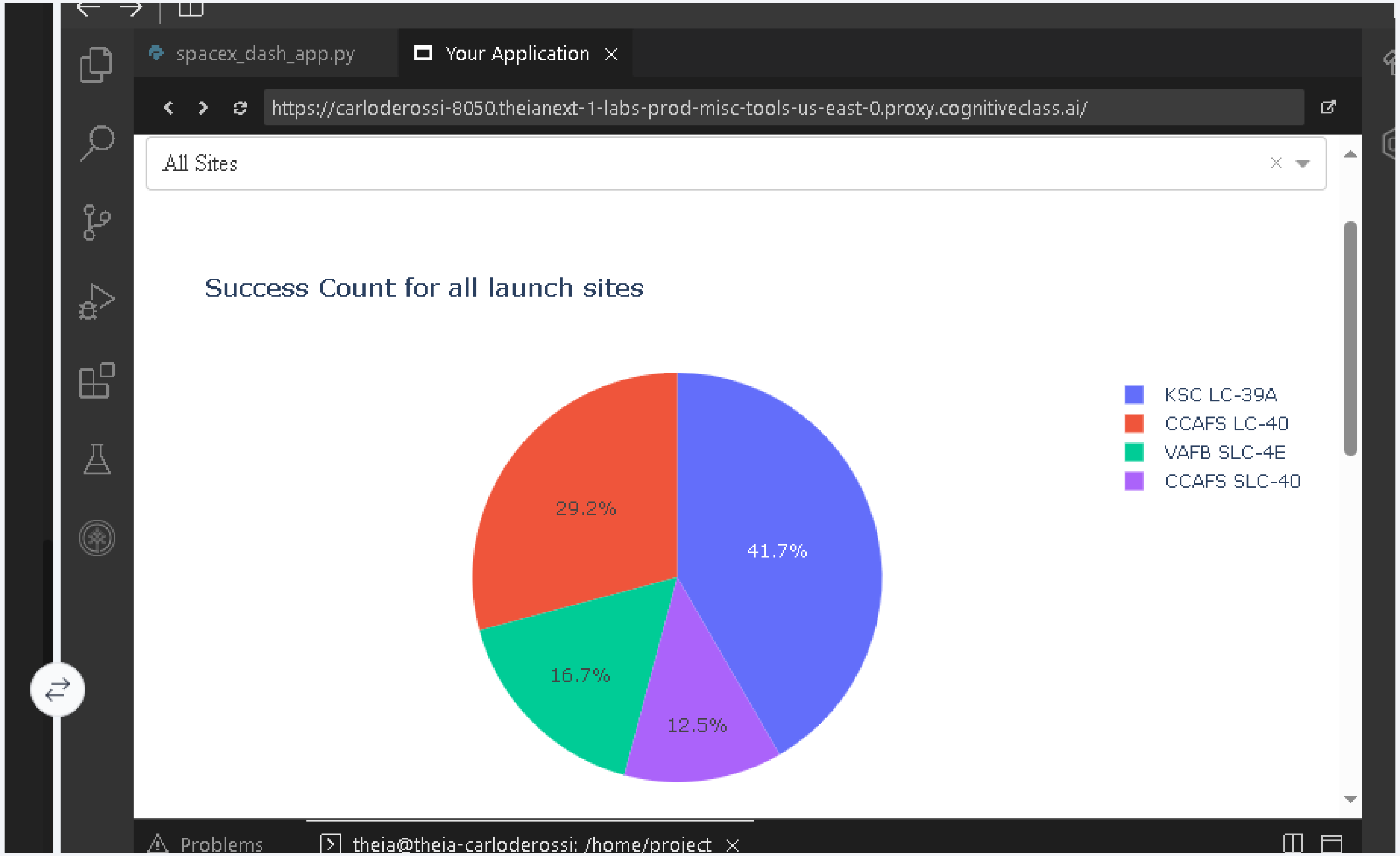


Section 4

Build a Dashboard with Plotly Dash

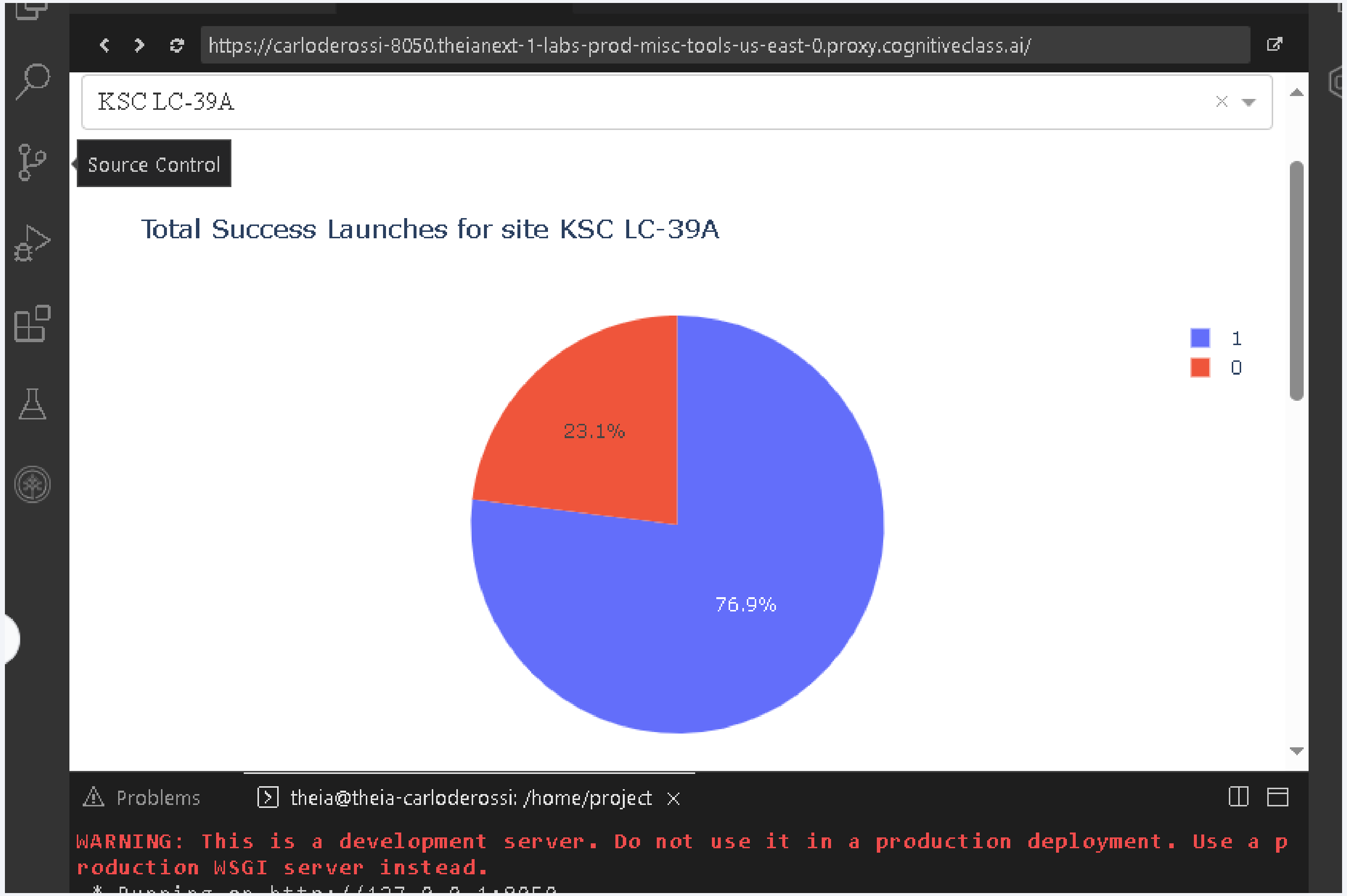
Success Count for all launch sites

- Highest number of success launch counts are in the KSC LC-39A launch center
- Lowest rate of success launches are in CCAFS SLC-40 center



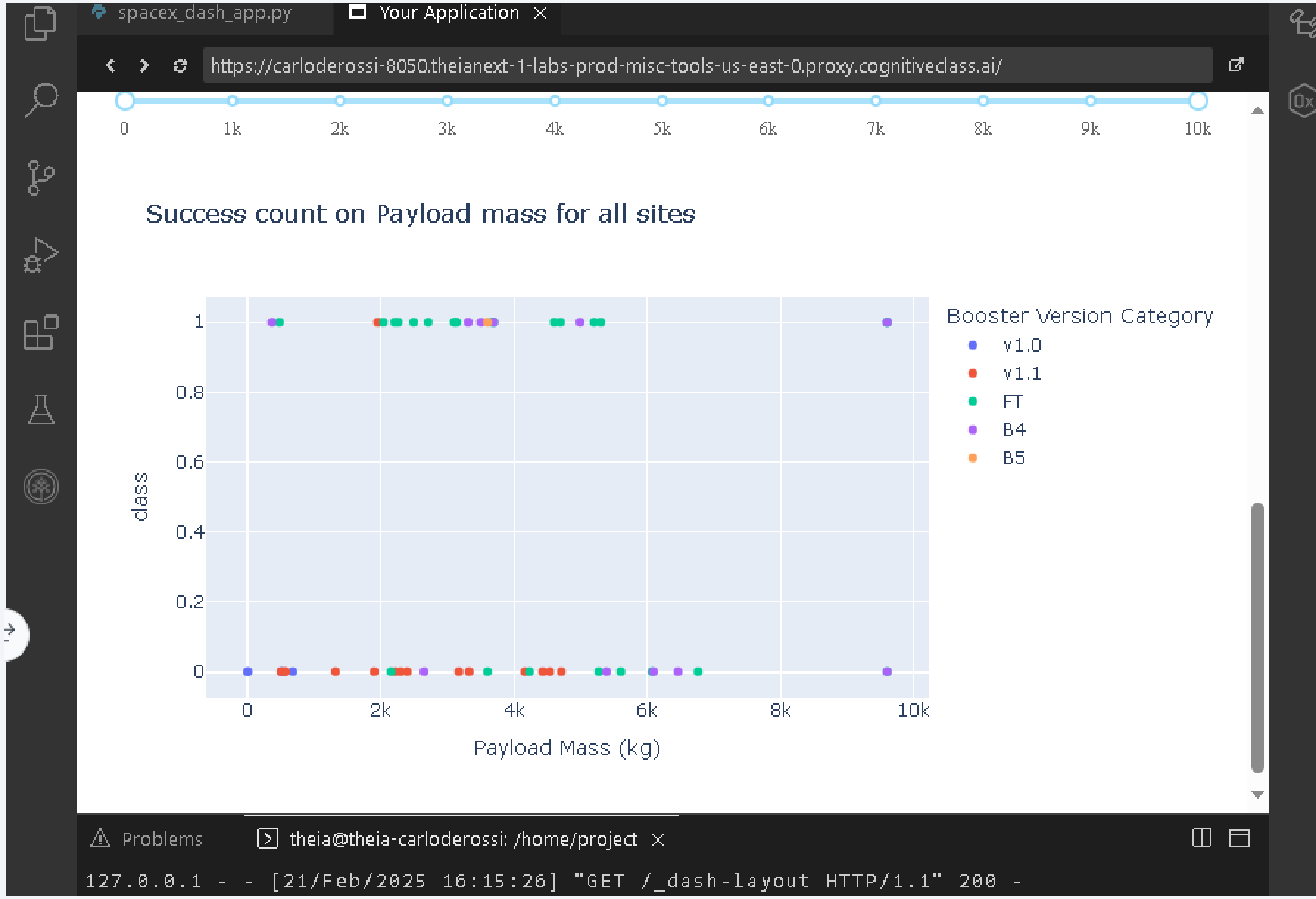
Total Success launches for site KSC LC-39A

- KSC LC-39A has a 76.9% success rate versus a 23.1% failure
- With this, it is the most succesful launch site



Success versus payload mass for Booster Category

- B4 and FT Booster versions are the only ones successfully launching higher payload masses



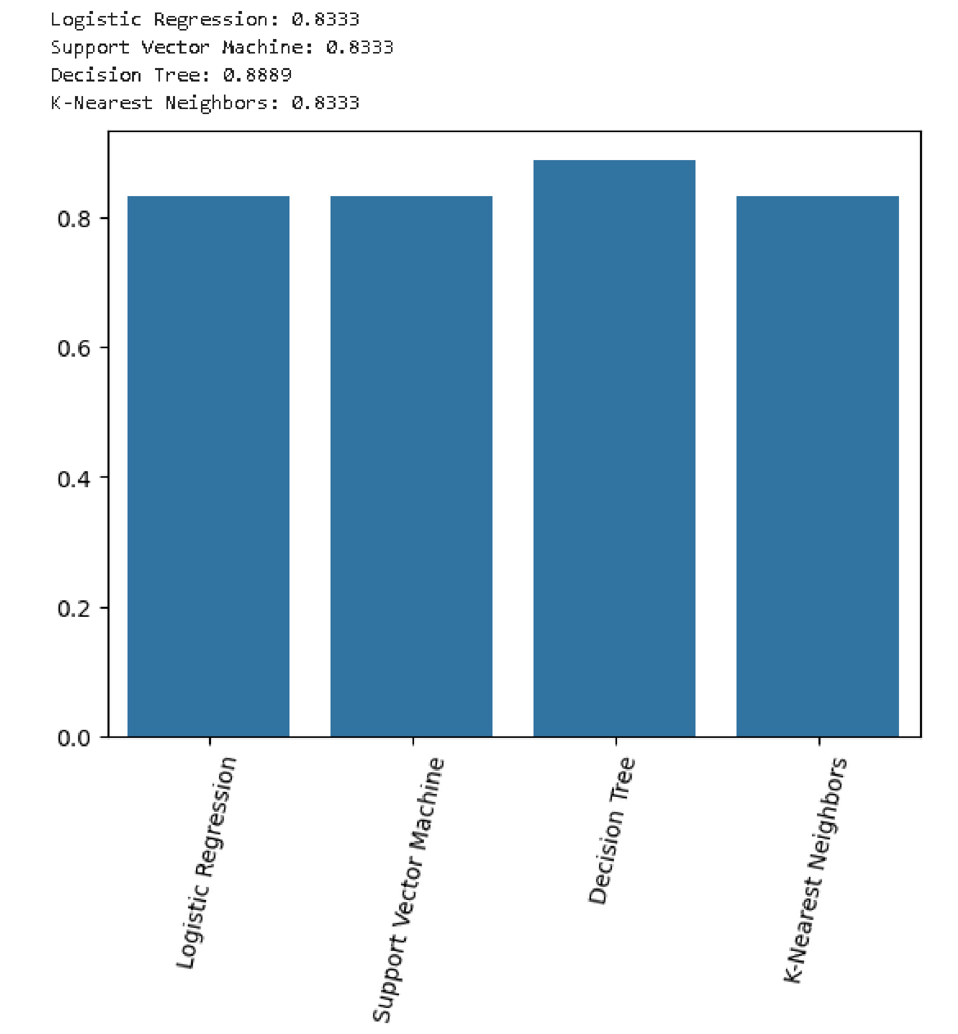
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Decision Tree: The Decision Tree model has the highest accuracy at **0.8889**. This means that this model correctly predicts the success of SpaceX launches almost 90% of the time.

Decision Trees are intuitive and easy to interpret, and their high accuracy suggests that this model might be capturing the underlying patterns in the data effectively.

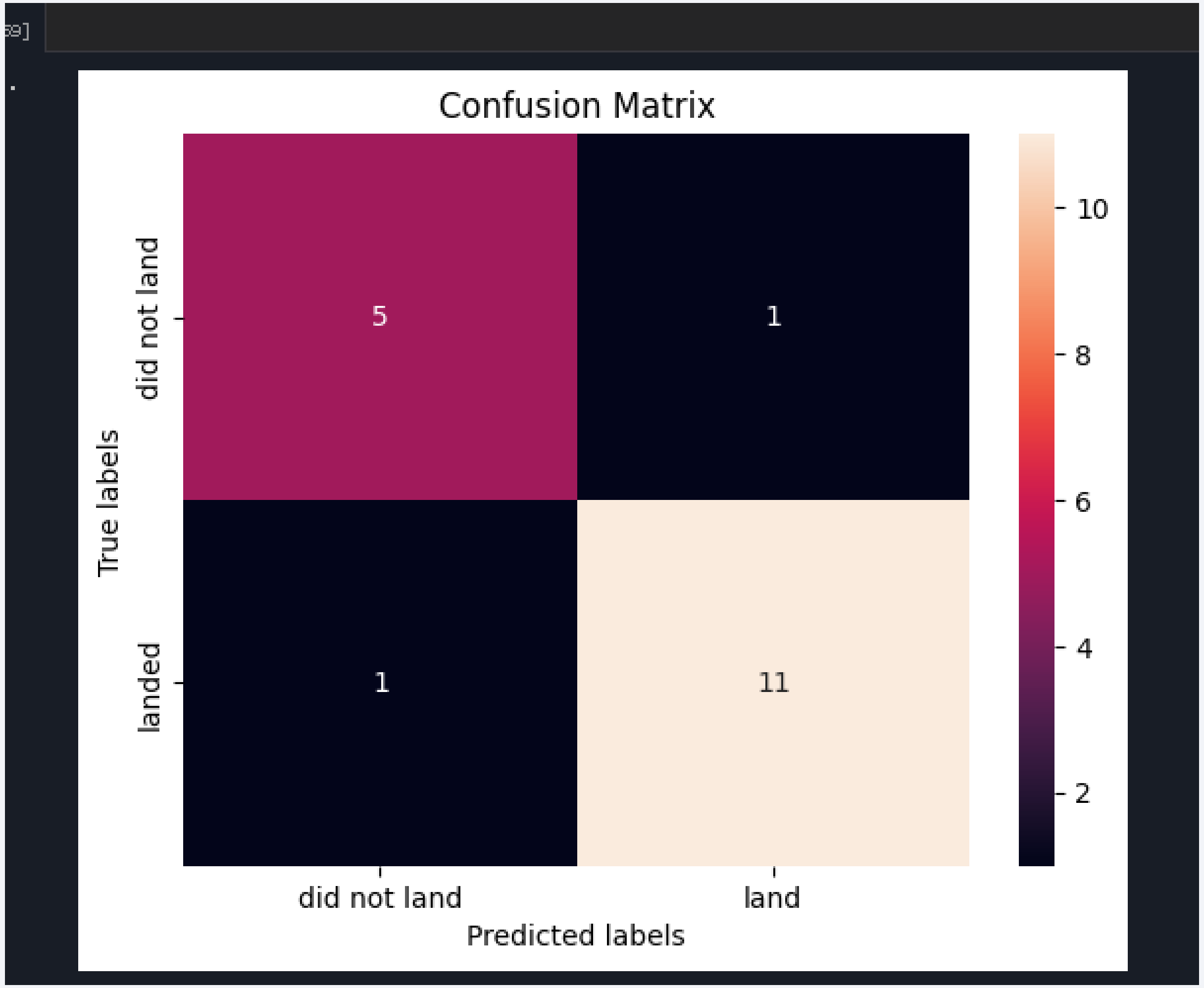


Confusion Matrix

The confusion matrix of the Decision Tree show its better accuracy.

On the other hand, this model has more False Negatives (predicting failure when it actually succeeds) than other models.

Financially, this is a much safer error than False Positives (predicting success when it fails).



Conclusions

- Decision Tree shows the best accuracy for prediction the mission's success
- This model also shows a higher False Negative rate (1 vs 0) than the other models
- Most successful Launch site is Kennedy Space Center Launch Complex 39A (KSC LC-39A)
- ES-L1 (Earth-Sun Lagrange Point 1), HEO (Highly Elliptical Orbit), and GEO (Geostationary Orbit) have proven to be very successful target orbits for satellite deployments and other space missions.

Appendix

- Please see all relevant source codes here:
- https://github.com/carloderossi/DataScienceCapstoneProj/tree/main

Thank you!