Winning Space Race with Data Science

Neil Truskolaski

https://github.com/Jtrusko/Applied_Data_Science_Capstone

04/28/2025

Outline

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

Executive Summary

- This analysis leveraged data from the SpaceX API and Wikipedia to build machine learning models predicting rocket landing success. Key steps included data labeling (‘class’ column), exploratory data analysis using SQL and various visualization techniques, feature engineering (one-hot encoding, standardization), and model training with GridSearchCV optimization.
- Four models were created: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors, all achieving approximately 83.33% accuracy. While the models demonstrate a reasonable predictive capability, they exhibit a bias towards predicting successful landings.
- This suggests that acquiring a more comprehensive and representative dataset is crucial for developing more robust and dependable landing prediction models, which could have significant implications for optimizing launch operations and cost-efficiency.

Introduction

Project Background:

- The commercial space industry is increasingly competitive, with cost a major factor.
- SpaceX's cost advantage (\$62M vs. \$165M for competitors) is largely due to its ability to recover and reuse the first stage of its Falcon rockets.

Business Problem:

- Space Y seeks to enter this market and needs to develop a similar capability.
- Additionally, Space Y recognizes that reusability is essential to compete with SpaceX's pricing.
- This project will deliver a machine learning solution that predicts the likelihood of successful Stage 1 rocket recovery, enabling Space Y to make informed decisions about recovery strategies and potentially reduce launch costs.

Section 1

Methodology

Methodology

Executive Summary

This project aimed to predict successful SpaceX Stage 1 rocket landings through the following process:

- **Data Collection:** Data was gathered from the SpaceX public API and the SpaceX Wikipedia page.
- **Data Wrangling:** Initial processing involved classifying landings as either successful or unsuccessful.
- **Exploratory Data Analysis (EDA):** The data was then explored using visualization techniques and SQL queries.
- **Interactive Visual Analytics:** Folium and Plotly Dash were used to create interactive visualizations.
- **Predictive Analysis:** Finally, classification models were built and tuned using GridSearchCV.

Four machine learning models were evaluated: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K-Nearest Neighbors. All models achieved similar accuracy (approximately 83.33%), with a tendency to over-predict successful landings. The project suggests that additional data may improve model accuracy and reliability.

Data Collection

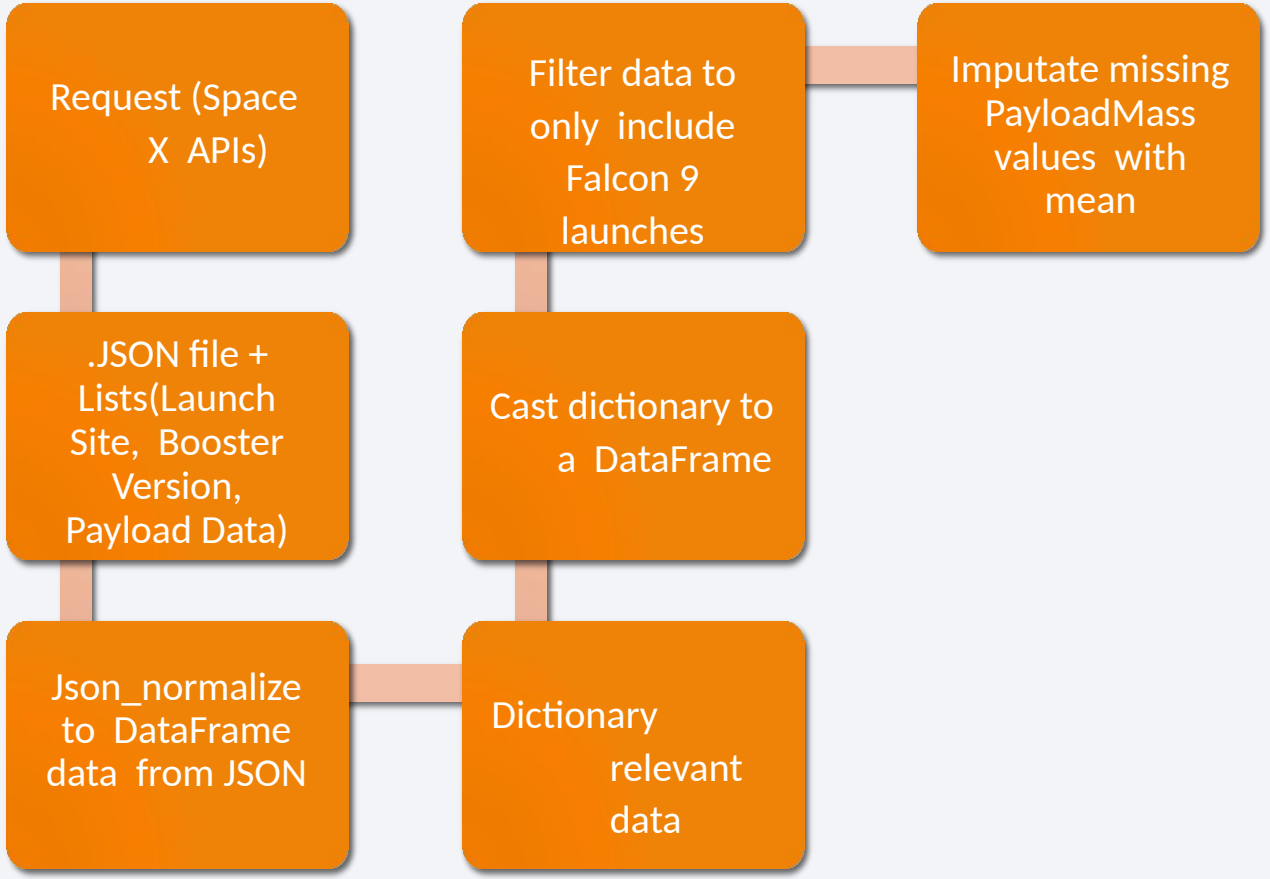
To compile a robust dataset, data was collected from the following sources:

- **SpaceX Public API:** This API provided a structured source of real-time and historical launch information. The following attributes were extracted:
  - *FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude*
  -
- **SpaceX Wikipedia:** Web scraping was utilized to capture additional launch details from a tabular format on the SpaceX Wikipedia page. The extracted data includes:
  - *Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time*

The integration of these two sources provided a rich dataset for analysis.

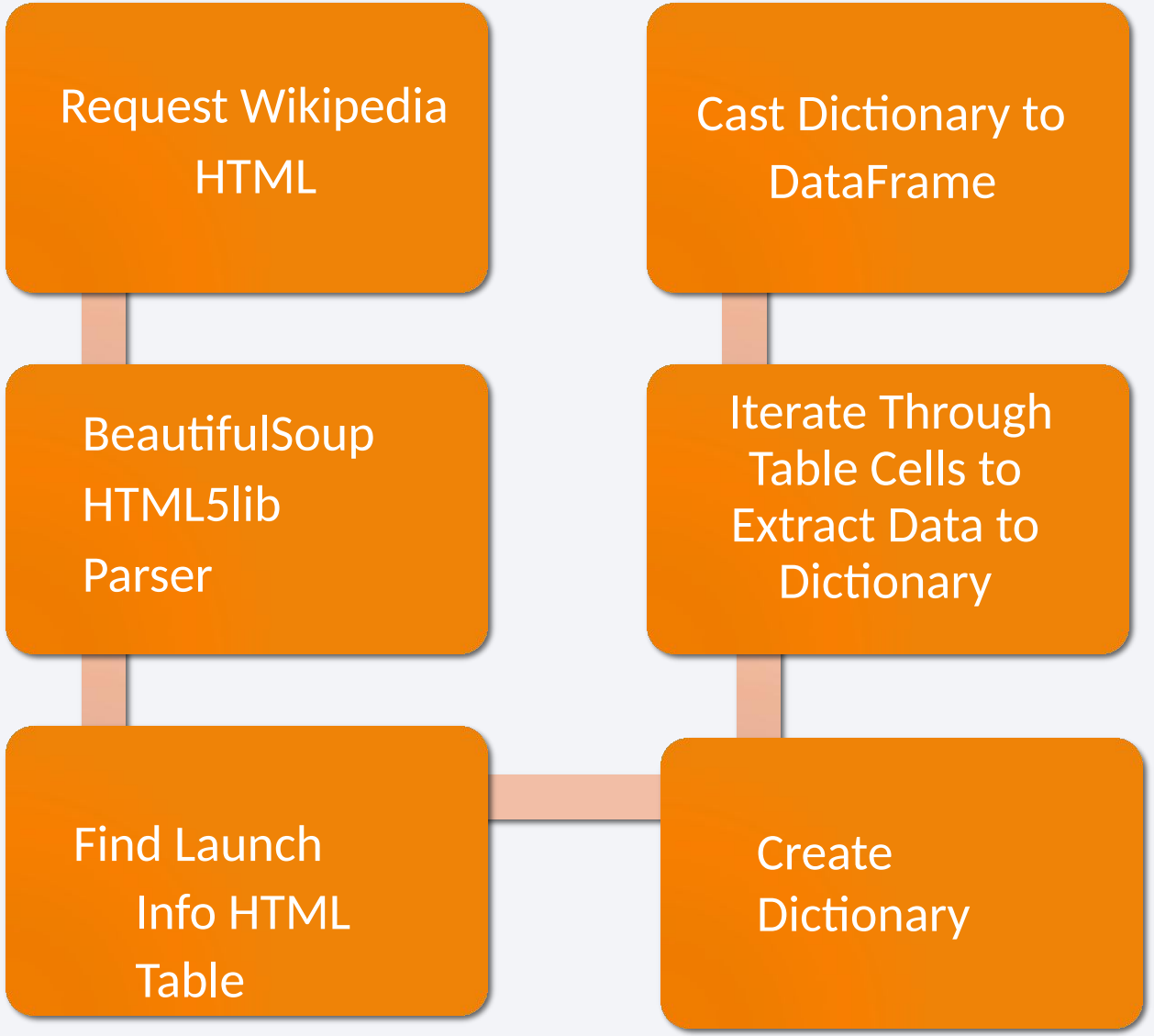
Data Collection – SpaceX API

- Here you can see the flow chart for the Data Collection API.
- Below is the link to the file.
- https://github.com/Jtrusko/Applied_Data_Science_Capstone/blob/main/Module%201/DataCollectionApi.ipynb



Data Collection - Scraping

- Seen here is the Web Scraping flow chart.
- Below is a link to the file.
- https://github.com/Jtrusko/Applied_Data_Science_Capstone/blob/main/Module%201/DataCollectionWebscraping.ipynb



Data Wrangling

The data wrangling process focused on generating a clear target variable for analysis:

- **Target Variable Creation:** A new column, Class, was engineered to represent the success or failure of rocket landings.
- **Outcome Column Analysis:** The original Outcome column included combined information on mission success and landing location.
- **Binary Label Encoding:**
  - Outcomes indicating successful landings were mapped to a value of 1:
    - *True ASDS*
    - *True RTLS*
    - *True Ocean*
  - Outcomes representing unsuccessful landings were mapped to a value of 0:
    - *None None*
    - *False ASDS*
    - *None ASDS*
    - *False Ocean*
    - *False RTLS*

https://github.com/Jtrusko/Applied_Data_Science_Capstone/blob/main/Module%201/Data%20wrangling.ipynb

EDA with Data Visualization

- **Flight Number vs. Payload Mass (Scatter Plot)**
  This chart was used to explore the relationship between mission experience (as measured by flight number) and payload weight. It helps determine whether success rates improve over time or with heavier/lighter payloads.

- **Flight Number vs. Launch Site (Scatter Plot)**
  By plotting flight numbers against different launch sites, we aimed to identify patterns in launch frequency and site usage across missions. This can reveal operational preferences or capacity trends.

- **Payload Mass vs. Launch Site (Scatter Plot)**
  This was used to compare how payload capacity differs across launch sites. It allows us to see which launch sites typically handle larger payloads and how that relates to mission success.

- **Success Rate by Orbit Type (Bar Chart)**
  A bar chart was plotted to compare success rates across various orbit types (e.g., LEO, GTO, etc.). This visualization helps determine which orbits have higher reliability and may influence strategic planning for future missions.

https://github.com/Jtrusko/Applied_Data_Science_Capstone/blob/main/Module%202/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- **Initial Data Exploration**
  SELECT * FROM SPACEXTBL LIMIT 5
  a. Previewed the dataset structure and sample values.
- **Launch Frequency Analysis**
  SELECT COUNT(*), GROUP BY Launch_Site
  a. Counted total launches and compared activity across sites.
- **Success Rate Evaluation**
  WHERE Class = 1, SUM(Class)/COUNT(*)
  a. Filtered for successful launches and calculated success percentages per site and booster version.
- **Payload Impact Study**
  Queried Payload_Mass__kg_ vs. Class
  a. Assessed whether payload mass influenced launch outcomes.
- **Booster Version Performance**
  GROUP BY Booster_Version
  a. Compared launch success across different Falcon 9 booster versions.

https://github.com/Jtrusko/Applied_Data_Science_Capstone/blob/main/Module%202/EDA_SQL.ipynb

Build an Interactive Map with Folium

To visualize SpaceX launch site data, I used **Folium** to create an interactive map with the following objects:

- **Circles**: Placed at each launch site's geographic coordinates to highlight the location area. Circles help visually represent the proximity and clustering of sites.
- **Markers with Labels**: Added text-based markers using `DivIcon` to label each site directly on the map for clear identification.
- **Popups**: Each circle includes a popup showing the full launch site name for better user interaction and site recognition.
- **Lines (in later tasks)**: Used to draw paths from each launch site to nearby coordinates, illustrating spatial relationships and aiding in distance analysis.

These visual elements were chosen to provide an intuitive understanding of spatial patterns in launch success and geography.

https://github.com/Jtrusko/Applied_Data_Science_Capstone/blob/main/Module%203/Folium%20Interactive%20Visual%20Map.ipynb

Build a Dashboard with Plotly Dash

This dashboard provides a dynamic interface for exploring SpaceX launch data, featuring the following interactive elements:

- **Launch Site Selection Dropdown:**
  - Purpose: Filters data and visualizations by launch site, enabling site-specific analysis.
  - Interaction: Selection updates the pie chart and scatter plot.
- **Pie Chart (Total Success Launches):**
  - Purpose: Visualizes launch success proportions, showing overall or site-specific success rates.
  - Interaction: Updates based on the launch site dropdown selection.
- **Payload Mass Range Slider:**
  - Purpose: Filters data by payload mass range to analyze its influence on launch success.
  - Interaction: Adjusting the slider filters the scatter plot data.
- **Scatter Plot (Payload vs. Launch Success):**
  - Purpose: Illustrates the relationship between payload mass and launch success, color-coded by booster version.
  - Interaction: Updates based on both launch site selection and payload mass range.

https://github.com/Jtrusko/Applied_Data_Science_Capstone/blob/main/Module%203/spacex_dash_app.py

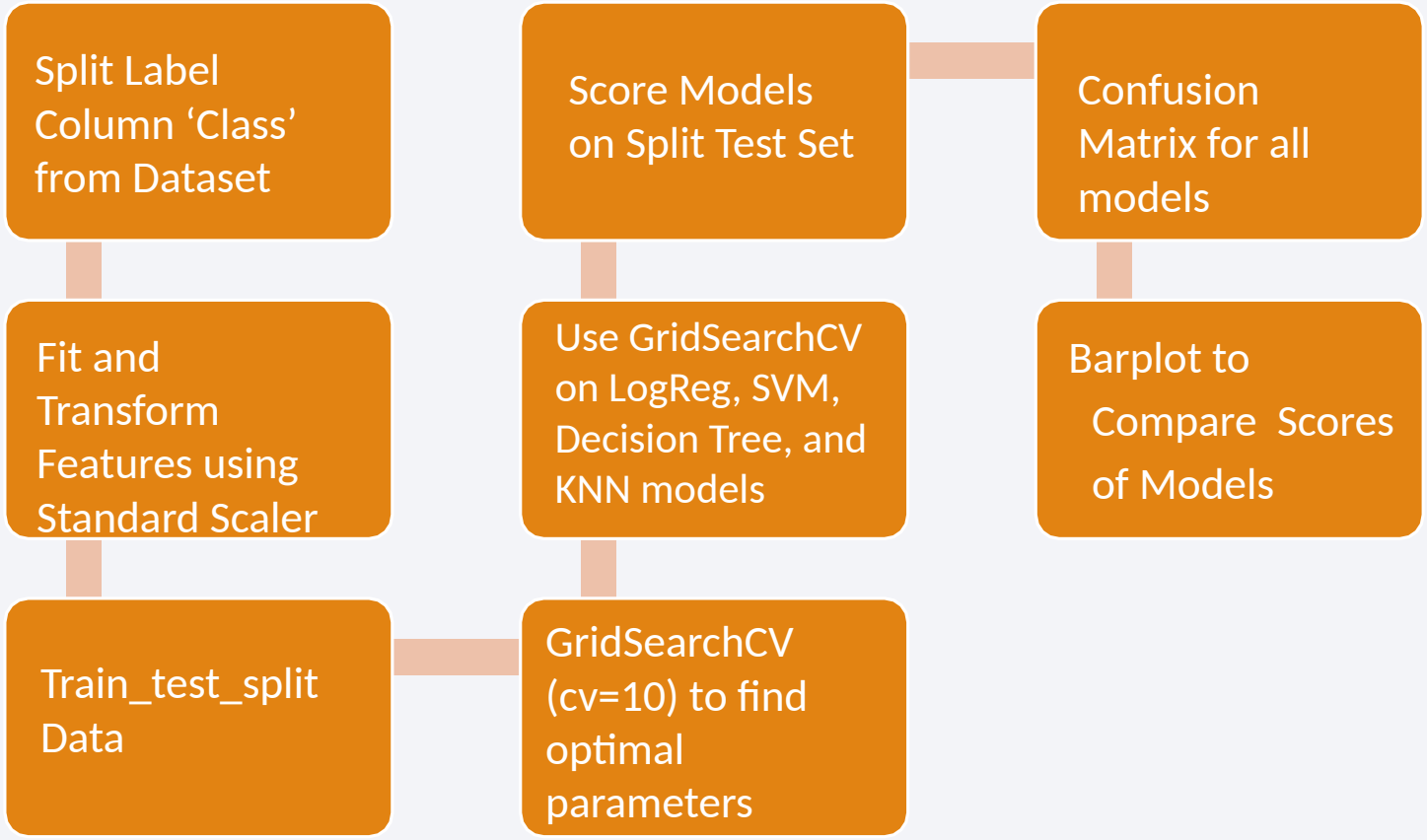
Predictive Analysis (Classification)

This project employed the following process to develop and evaluate classification models for predicting SpaceX launch outcomes:

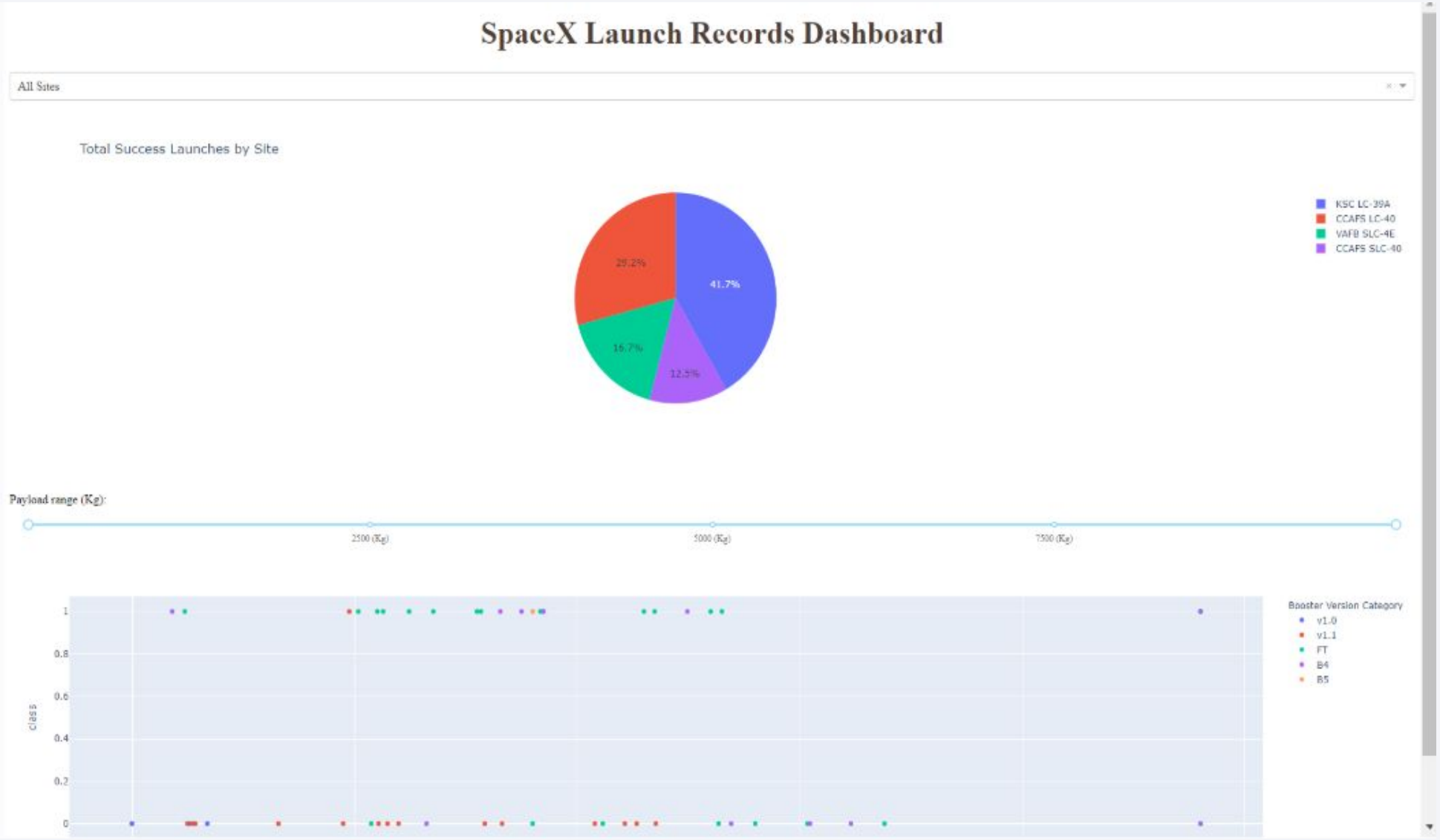
- **Data Preparation:**
  - Features were selected and engineered from the processed dataset.
  - Categorical variables were converted to binary using one-hot encoding.
  - Numerical features were standardized to ensure consistent scaling.
  - The data was split into training and testing sets to evaluate model performance.
- **Model Selection and Training:**
  - Four classification models were chosen: Logistic Regression, Support Vector Machine (SVM), Decision Tree Classifier, and K-Nearest Neighbors (KNN).
- **Hyperparameter Tuning:**
  - GridSearchCV was used to optimize the hyperparameters for each model. This involved defining a parameter grid and performing cross-validation to identify the best-performing parameter combinations.
- **Model Evaluation:**
  - The trained models were evaluated on the test set.
  - Accuracy score was used as the primary metric to assess model performance.
  - Confusion matrices were generated to analyze the types of errors made by each model.
- **Model Comparison:**
  - The accuracy scores of all four models were compared to determine the best-performing model.

Predictive Analysis (FlowChart)

https://github.com/Jtrusko/Applied_Data_Science_Capstone/blob/main/Module%204/Machine%20Learning%20Prediction.ipynb



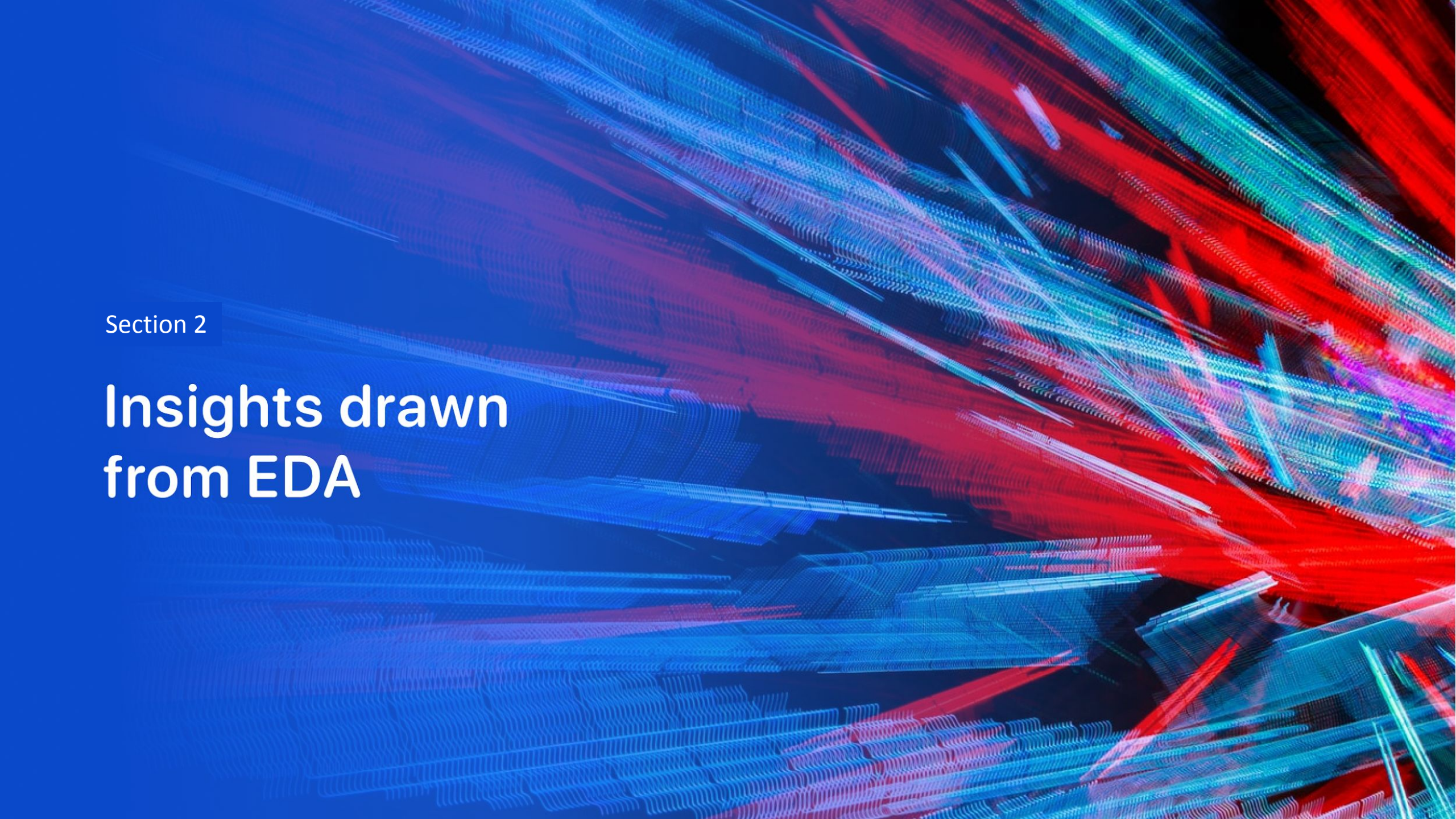
Results



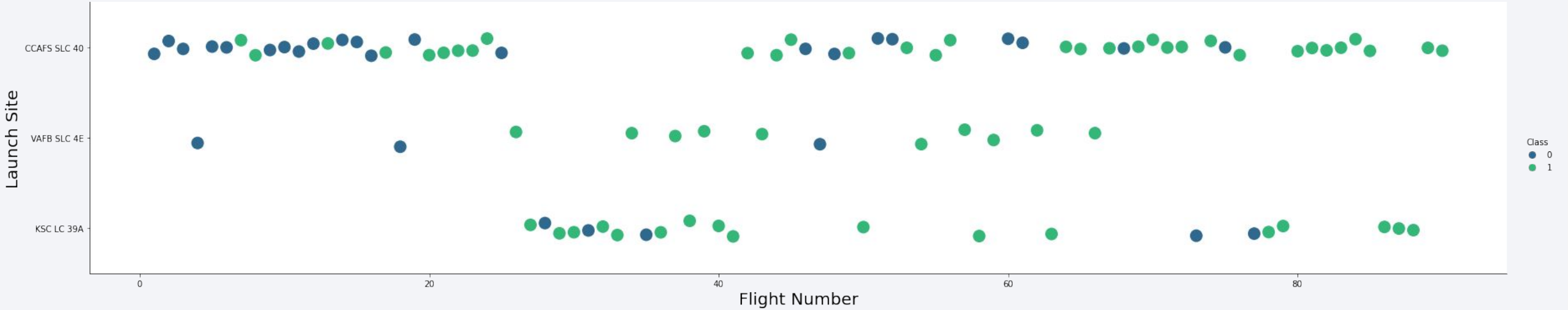
- This slide previews the Plotly dashboard. Subsequent slides will cover EDA visualizations, SQL analysis, the Folium map, and model results (approximately 83% accuracy)

Section 2

Insights drawn from EDA

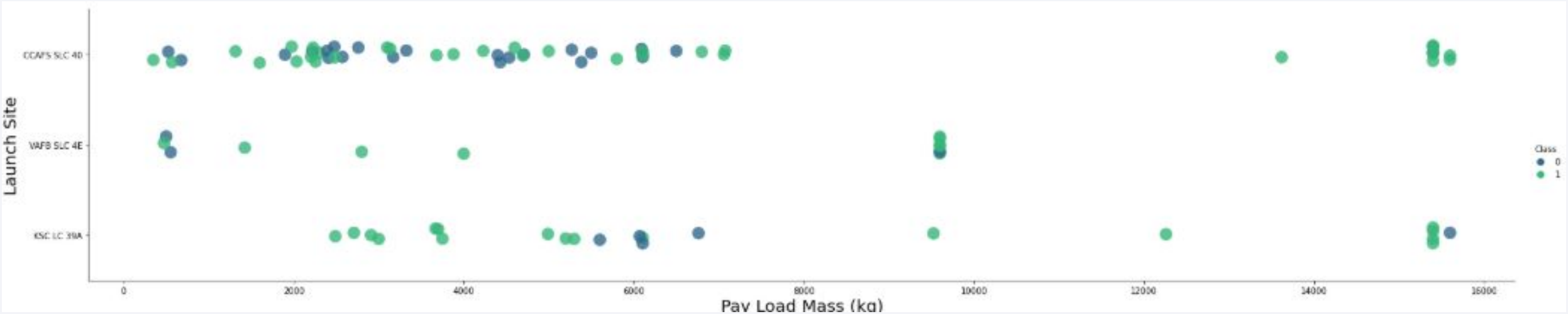


Flight Number vs. Launch Site



- The scatter plot illustrates the launch success rate (indicated by color: **Green** for success, **Blue** for failure **this is true for all future slides**) across different flight numbers and launch sites.
- A visual trend suggests an increasing proportion of successful launches as the flight number progresses, with a noticeable shift towards predominantly successful outcomes after approximately flight 20.
- Cape Canaveral Space Force Station (CCAFS) shows the highest density of launch records compared to the other sites (KSC LC-39A and VAFB SLC-4E), indicating it was the most frequently used launch site during this period.

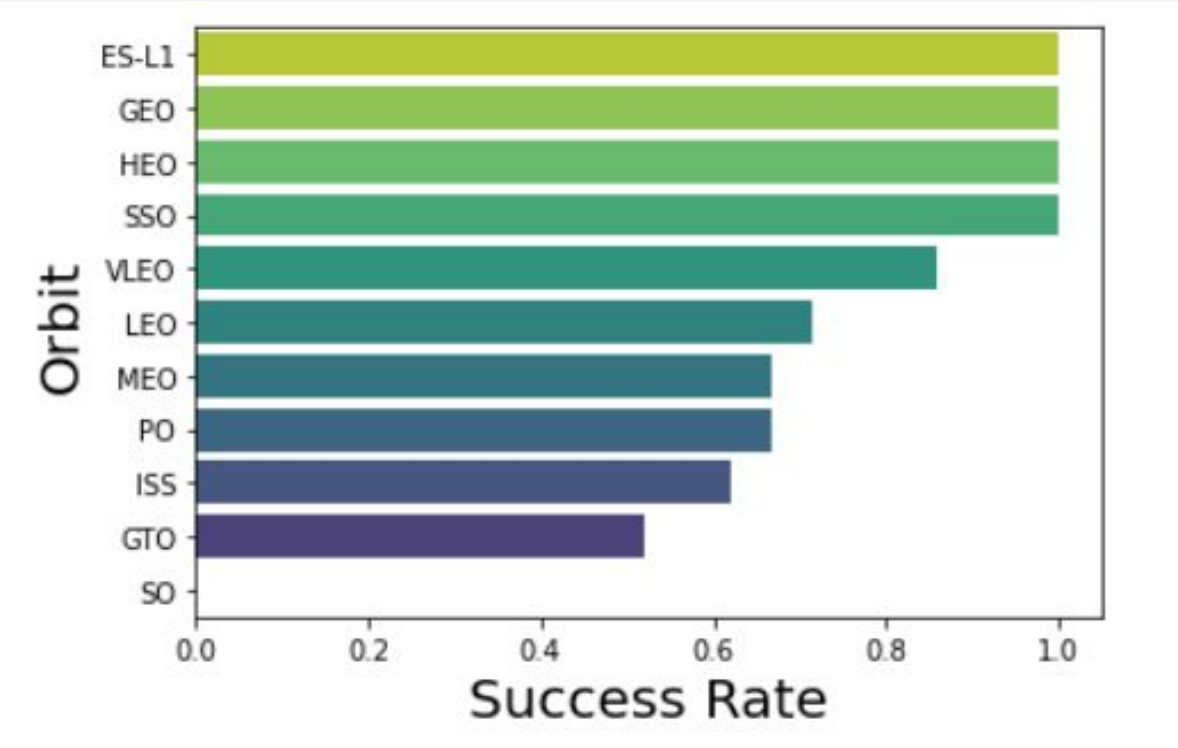
Payload vs. Launch Site



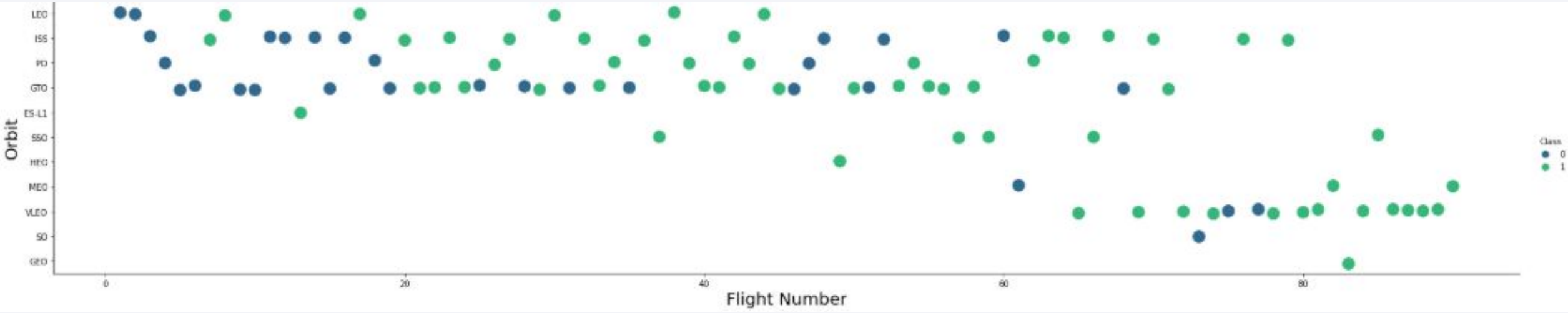
- The majority of payload masses for these launches fall within the 0 to 6000 kg range, as visually indicated by the density of data points in that area.
- Furthermore, the scatter plot suggests a tendency for different launch sites to handle distinct ranges of payload mass.

Success Rate vs. Orbit Type

- This bar chart visualizes the success rate of SpaceX launches across various orbital trajectories.
- The success rate is scaled from 0% to 100%.
    - Notably, ES-L1 (sample size: 1), GEO (sample size: 1), and HEO (sample size: 1) exhibit a 100% success rate, as does SSO (sample size: 5).
    - VLEO (sample size: 14) demonstrates a decent success rate with a reasonable number of launch attempts.
    - In contrast, SO (sample size: 1) has a 0% success rate.
    - GTO (sample size: 27), which has the largest sample size among the orbits, shows an approximate 50% success rate.
- This data highlights the varying degrees of success associated with different orbital destinations and the relative frequency of launches to those orbits.

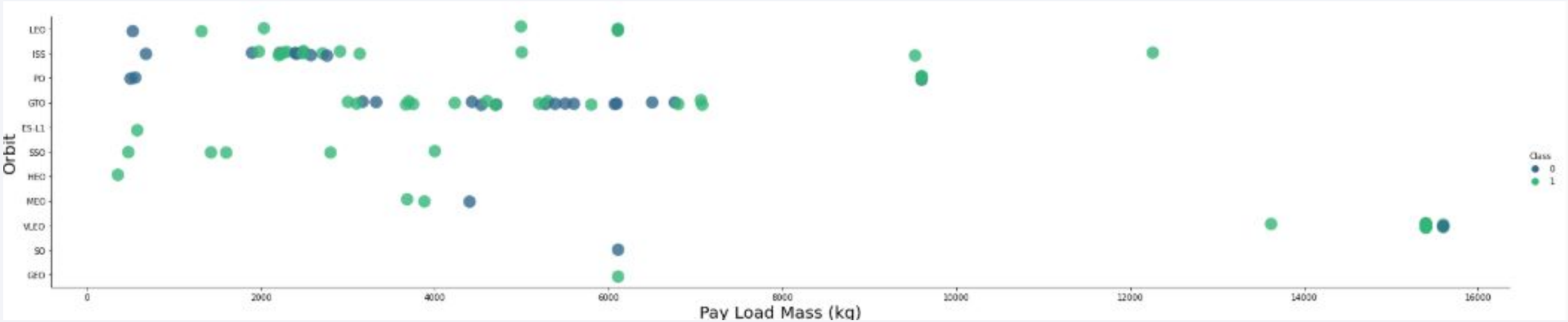


Flight Number vs. Orbit Type



- This scatter plot reveals a potential evolution in SpaceX's preferred launch orbits over time (indicated by Flight Number).
- Early launches were predominantly to Low Earth Orbit (LEO), which experienced moderate success rates (indicated by the distribution of colors).
- There's a noticeable return to Very Low Earth Orbit (VLEO) in more recent launches.
- Overall, the visual data suggests that SpaceX achieves **higher success rates in lower Earth orbits (LEO, VLEO) and Sun-Synchronous Orbit (SSO)**, as indicated by the prevalence of successful launch outcomes (likely green dots) in those orbit types across the flight history.

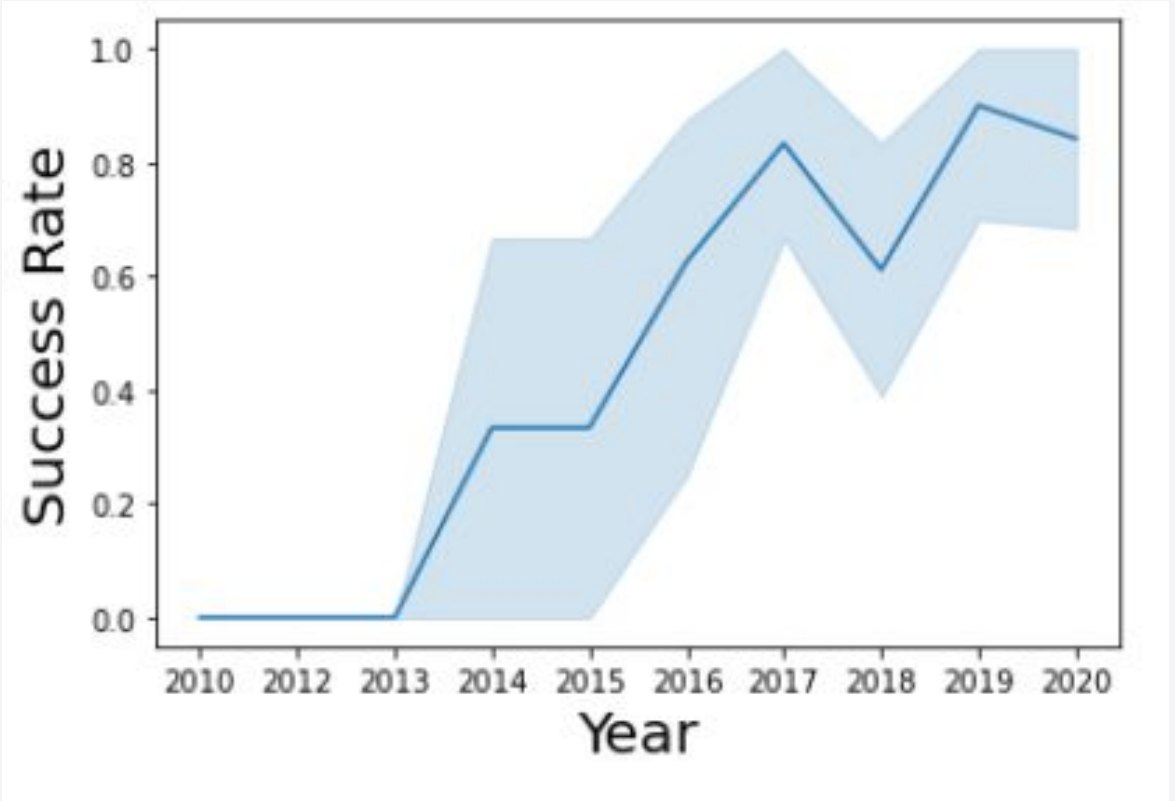
Payload vs. Orbit Type



- This scatter plot suggests a correlation between payload mass and the achieved orbit.
- Low Earth Orbit (LEO) and Sun-Synchronous Orbit (SSO) launches appear to be associated with relatively lower payload mass values.
- In contrast, Very Low Earth Orbit (VLEO), another orbit with a notable success rate (based on previous observations), seems to be utilized for launches carrying payloads in the higher end of the observed mass range.
- This indicates that the mission requirements and the target orbit significantly influence the amount of payload that can be carried.

Launch Success Yearly Trend

- This line plot illustrates the trend of SpaceX launch success rate over time, along with a shaded area representing the 95% confidence interval.
- Overall, there's a general upward trend in success rate observed since 2013.
    - A notable dip in success rate occurs around 2018 before recovering.
    - In recent years (around 2019-2020), the success rate appears to have stabilized at approximately 80%, with the confidence interval suggesting a relatively consistent performance within that timeframe.



All Launch Site Names

```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;

         * ibm_db_sa://ftb12020:***@0c77d6f2
        Done.
```

Out[4]:

launch_site
CCAFS LC-40
CCAFS SLC-40
CCAFSSLC-40
KSC LC-39A
VAFB SLC-4E

The SQL query shown was executed to retrieve a list of all the unique launch site names present in the SPACEXDATASET table. The UNIQUE *(also known as DISTINCT)* keyword ensures that each launch site is listed only once, eliminating any duplicates.

Explanation:

The query identified the following unique launch sites used by SpaceX:

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

Launch Site Names Begin with 'CCA'

```
In [5]: %%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

The SQL query above was executed to retrieve all columns for the first 5 records in the `SPACEXDATASET` table where the `Launch_Site` name begins with the characters 'CCA', (the `LIKE 'CCA%'` clause filters the results to include only those launch sites that start with CCA), and `LIMIT 5` restricts the output to the first five matching records.

The query returned the following first five launch records from sites starting with 'CCA':

- **Date:** 2010-06-04, **Launch Site:** CCAFS LC-40
- **Date:** 2010-12-08, **Launch Site:** CCAFS LC-40
- **Date:** 2012-05-22, **Launch Site:** CCAFS LC-40
- **Date:** 2012-10-08, **Launch Site:** CCAFS LC-40
- **Date:** 2013-03-01, **Launch Site:** CCAFS LC-40

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

- The SQL query above calculates the sum of the PAYLOAD_MASS__KG_ for all launch records in the SPACEXDATASET where the CUSTOMER is 'NASA (CRS)'. The result is aliased as SUM_PAYLOAD_MASS_KG.

- The query determined that the total payload mass carried by SpaceX boosters for NASA (Commercial Resupply Services) missions in this dataset is **45,596 kilograms**. This figure represents the **cumulative weight** of all cargo delivered to space under the NASA (CRS) program based on the available launch records.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-8
Done.
```

avg_payload_mass_kg
2928

- The SQL query above calculates the average of the PAYLOAD_MASS__KG_ for all launch records in the SPACEXDATASET where the booster_version is 'F9 v1.1'. The result is aliased as AVG_PAYLOAD_MASS_KG.

- The query determined that the average payload mass carried by the SpaceX booster version F9 v1.1 in this dataset is **2928 kilograms**. This figure represents the mean weight of the payloads launched using this specific booster version across all recorded missions.

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

- The SQL query above finds the earliest (minimum) date from the DATE column in the SPACEXDATASET table for records where the landing__outcome is 'Success (ground pad)'. The result is aliased as FIRST_SUCCESS.
- The query determined that the date of the first successful landing outcome on a ground pad in this dataset is **2015-12-22**. This indicates the point in time when SpaceX first achieved a successful recovery of a stage on a designated ground landing site, according to the recorded data.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

- The SQL query above retrieves the `booster_version` for all launch records in the `SPACEXDATASET` where the `landing__outcome` was 'Success (drone ship)' and the `payload_mass__kg_` was strictly greater than 4000 and strictly less than 6000 (using the `BETWEEN` operator with the specified boundaries).
- The query identified the following booster versions that successfully landed on a drone ship while carrying a payload mass between 4001 and 5999 kilograms:
  - F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

This list provides specific booster identifiers that achieved this particular combination of successful landing method and payload weight range.

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

The SQL query above counts the number of occurrences for each unique value in the `mission_outcome` column of the `SPACEXDATASET` table. The results are grouped by `mission_outcome`, and the count for each outcome is aliased as `no_outcome`.

The query determined the following counts for each mission outcome:

- **Failure (in flight):** 1
- **Success:** 99
- **Success (payload status unclear):** 1

This shows that the vast majority of missions in this dataset were recorded as a 'Success', with a small number of 'Failure (in flight)' and 'Success (payload status unclear)' outcomes

Boosters Carried Maximum Payload

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
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

arried the maximum payload

anation here

2015 Launch Records

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

The SQL query above retrieves the `booster_version` and `PAYLOAD_MASS__KG_` for all records in the `SPACEXDATASET` where the `PAYLOAD_MASS__KG_` is equal to the maximum payload mass found in the entire dataset (determined by the subquery `SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET`).

The query identified the following booster versions that carried the maximum payload mass of **15600 kilograms**:

- 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

This list indicates multiple booster versions were used to launch the heaviest payloads recorded in this dataset.

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

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

The SQL query above counts the occurrences of different landing outcomes that start with(%) 'Success' within the specified date range (between June 4th, 2010, and March 20th, 2017). The results are grouped by `landing__outcome`, the counts are aliased as `no_outcome`, and the final output is ordered in descending order based on the count.

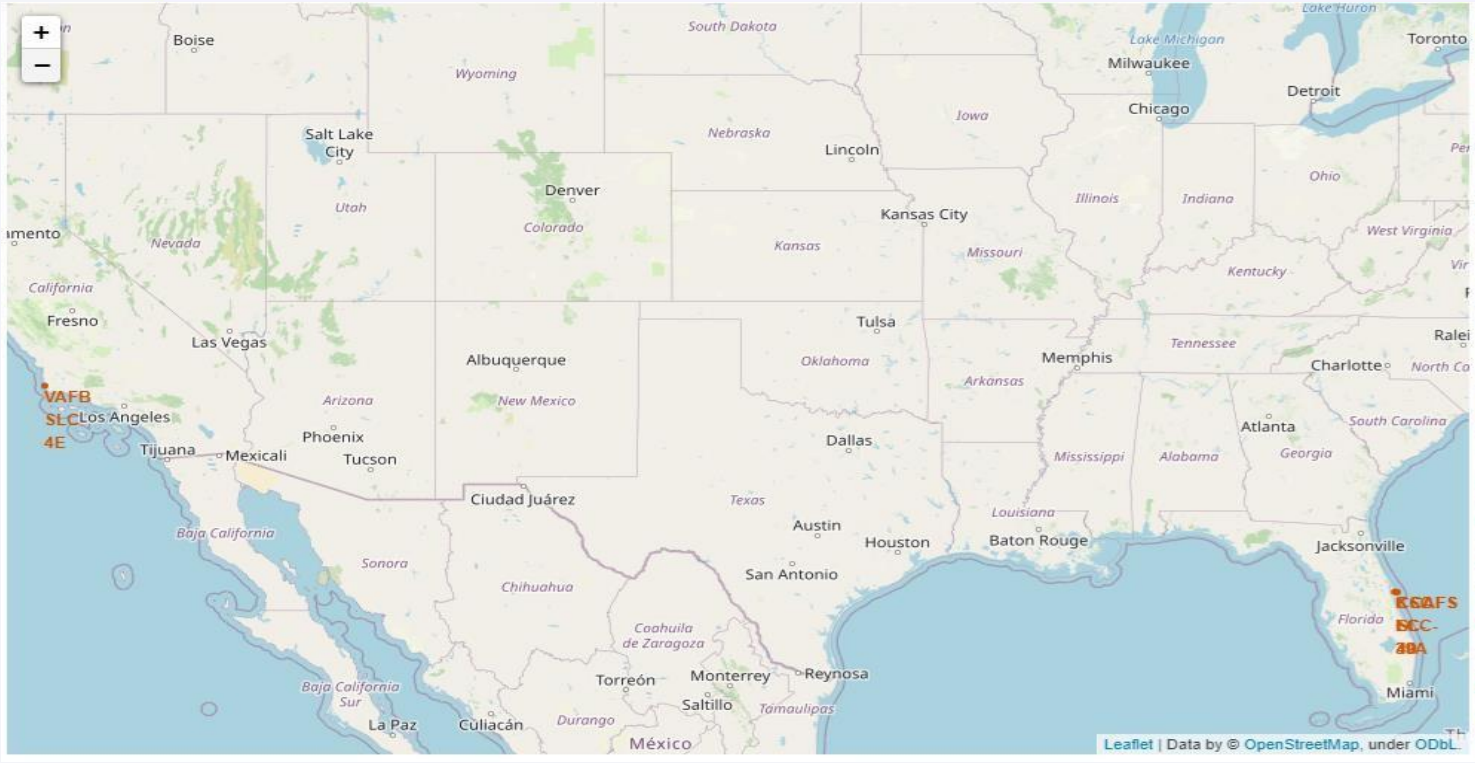
Within the specified timeframe, the ranking of successful landing outcomes is as follows:

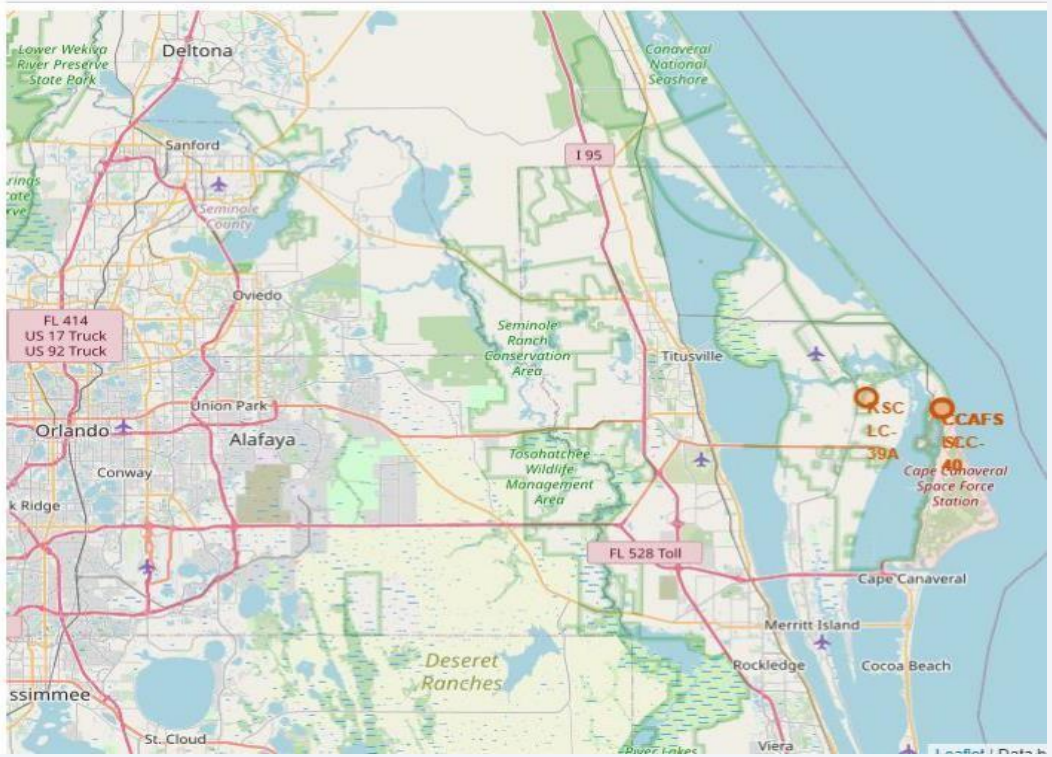
1. **Success (drone ship):** 5 Occurrences
2. **Success (ground pad):** 3 Occurrences
3. **Success (drone + ground):** 8 Occurrences

This ranking indicates that successful landings on a drone ship were more frequent than successful landings on a ground pad during the period between June 2010 and March 2017, according to the data.


Section 3
Launch Sites
Proximities Analysis

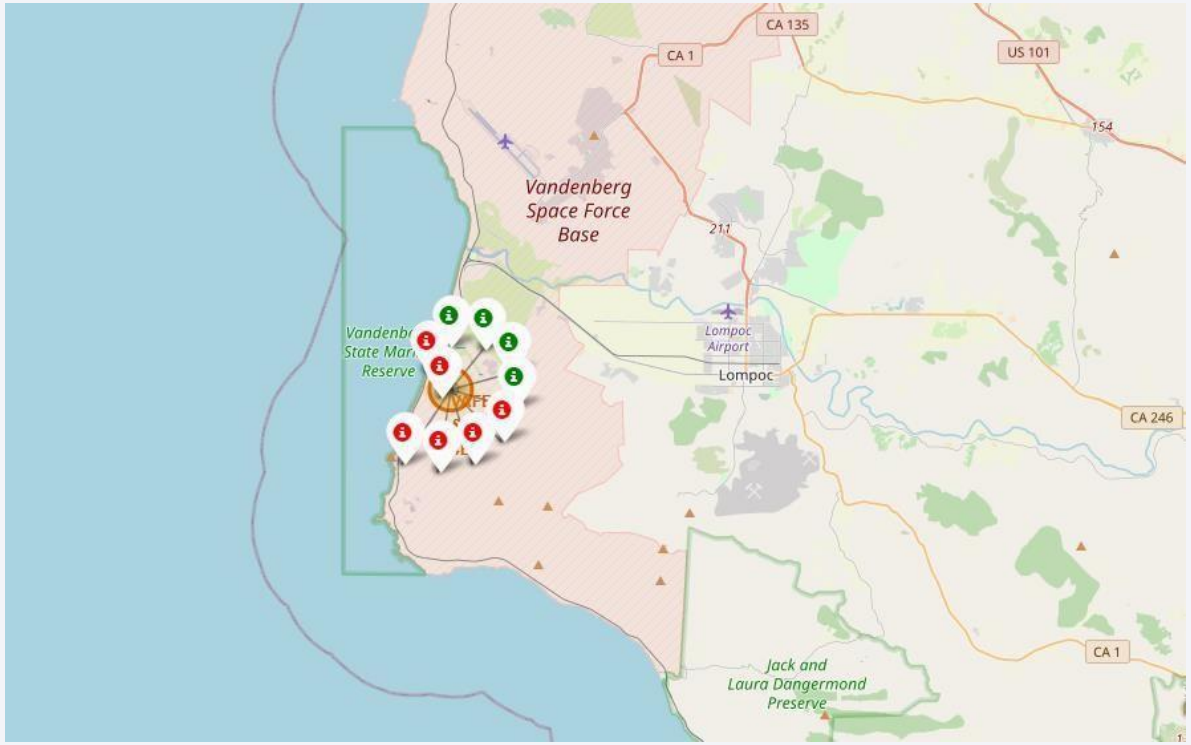
SpaceX Launch Site Locations - Global Overview





- The left map displays the global locations of all SpaceX launch sites relative to the US. The right inset map provides a closer view of the two closely situated launch sites in Florida. Notably, all identified launch sites are strategically positioned near the ocean.

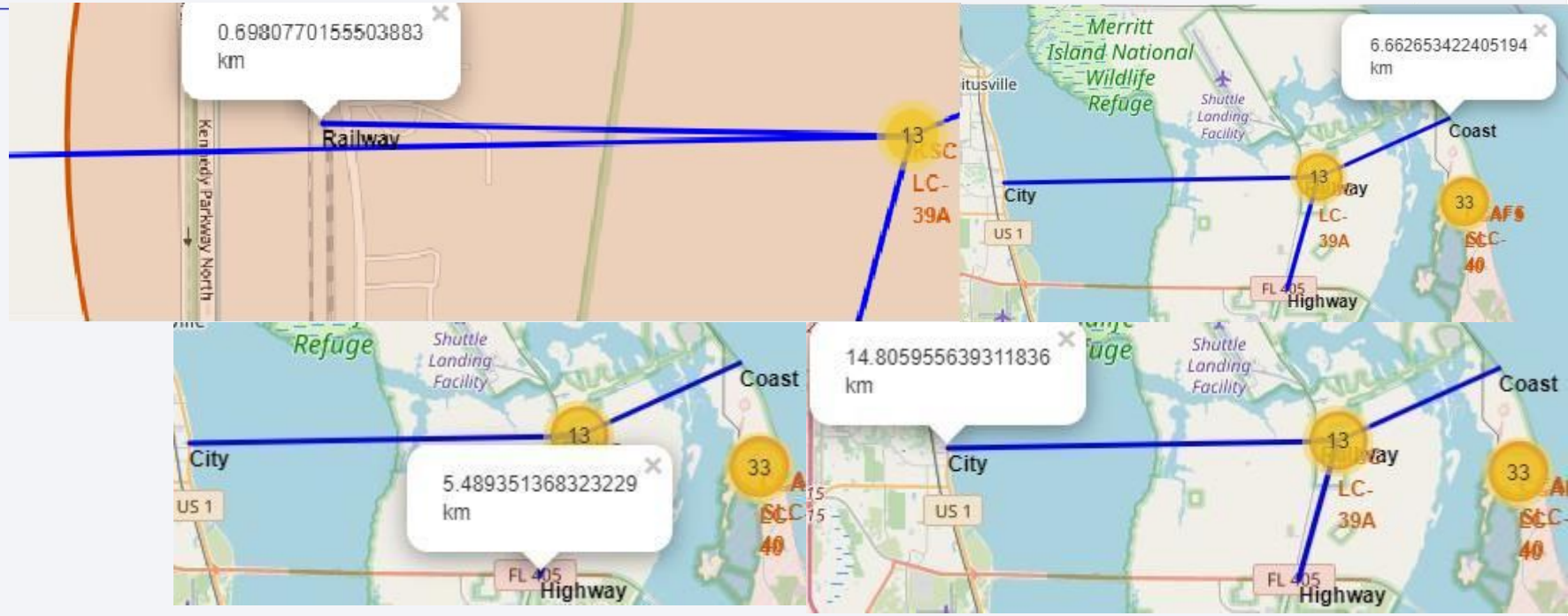
Launch Outcomes at VAFB SLC-4E



- The interactive Folium map shows launch outcomes at VAFB SLC-4E via color-coded icons (green = success, red = failure). Clicking the cluster displays 4 successful and 6 failed landings.

Proximity Analysis for Florida Launch Sites

The combined view focuses on the geographical proximity of the Florida launch sites to key infrastructure and natural features.

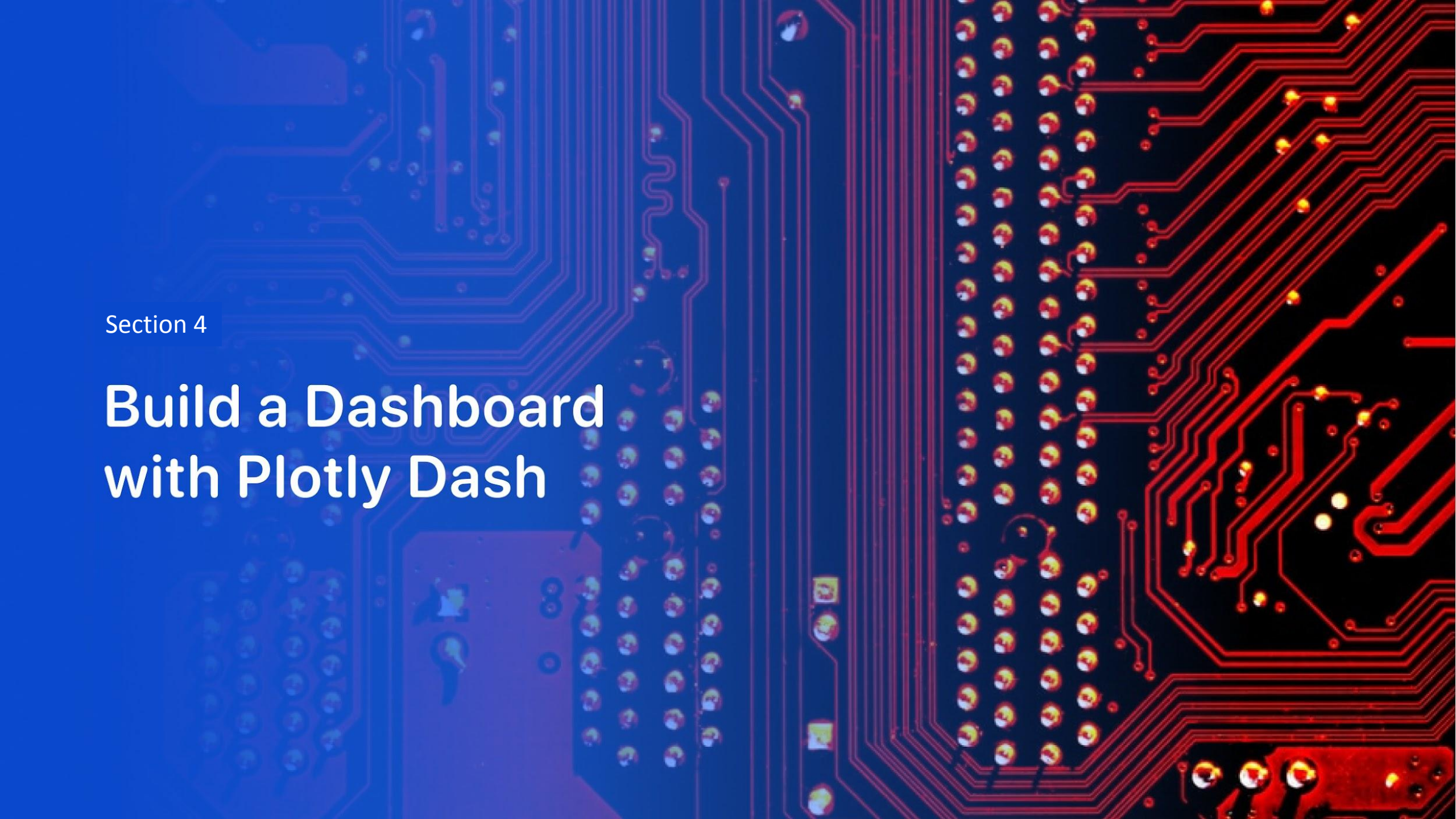


- **Top Left:** Shows the distance from one of the launch sites to a **railway line**. The calculated distance is displayed as approximately **0.698 km**. This indicates a relatively close proximity, which could be relevant for transporting large rocket components.

- **Top Right:** Shows the distance from one of the launch sites to the **coastline**. The calculated distance is approximately **5.66 km**. This proximity to the coast is a common characteristic of launch sites, providing safety for launches over water and access to various orbital inclinations.

- **Bottom Left:** Shows the distance from one of the launch sites to a **highway** (labeled "Highway"). The calculated distance is approximately **5.49 km**. Proximity to major transportation arteries is important for logistics and personnel access.

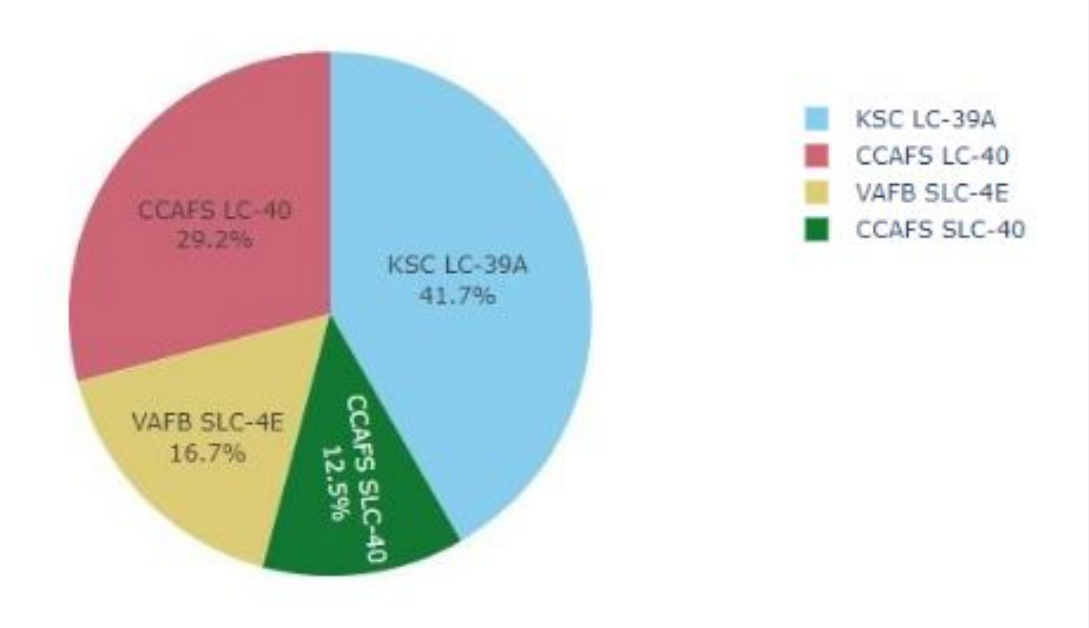
- **Bottom Right:** A distance of approximately **14.8 km** separates the launch site from a **city,** highlighting the importance for employee staffing and housing considerations at a nearby location.


Section 4
Build a Dashboard with Plotly Dash

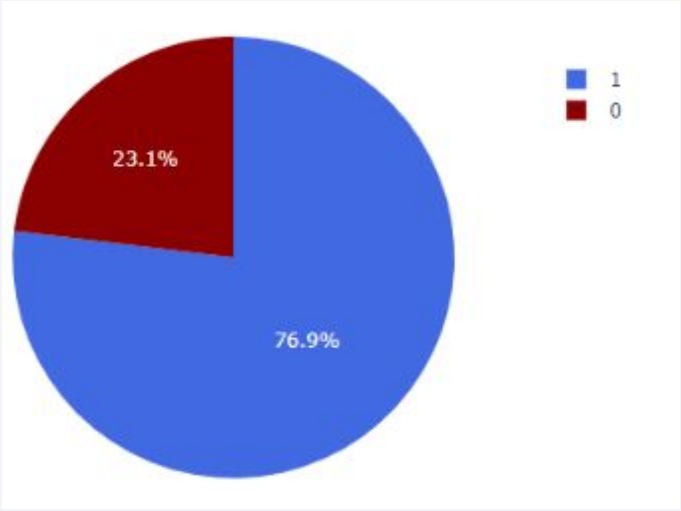
Distribution of Successful Launches by Site

This pie chart visualizes the proportion of successful SpaceX launches attributed to each launch site.

- **KSC LC-39A (Light Blue - 41.7%):** Kennedy Space Center Launch Complex 39A is responsible for 41.7% of successful launches, the largest share from a single site.
- **CCAFS (Red & Green - 41.7% Total):**
  - **CCAFS LC-40 (Red - 29.2%)**
  - **CCAFS SLC-40 (Green - 12.5%)**
- **VAFB SLC-4E (Yellow - 16.7%)**



Launch Outcomes at KSC LC-39A (Highest Success)



KSC LC-39A
Success Rate = **Blue**

This pie chart visualizes the distribution of launch outcomes specifically for Kennedy Space Center Launch Complex 39A (KSC LC-39A). The chart clearly shows the proportion of successful and failed landings at this particular launch site.

- **Successful Landings (Green):** 10
- **Failed Landings (Red):** 3

Key Finding:

KSC LC-39A demonstrates a high success rate, with 10 successful landings compared to 3 failed landings, within the data represented in the pie chart. This indicates a strong track record for this specific launch complex.

Payload Mass vs. Success vs. Booster Version Category

This scatter plot displays Payload Mass (kg) vs. Launch Outcome (Class: 1 = Success, 0 = Failure), with booster version categories color-coded. The dashboard includes a Payload range selector (0-10000 kg, though max payload is 15600 kg). Point size indicates the number of launches.

- In the 0-6000 kg range shown, there are two failed landings with payloads of 0 kg

Section 5

Predictive Analysis (Classification)

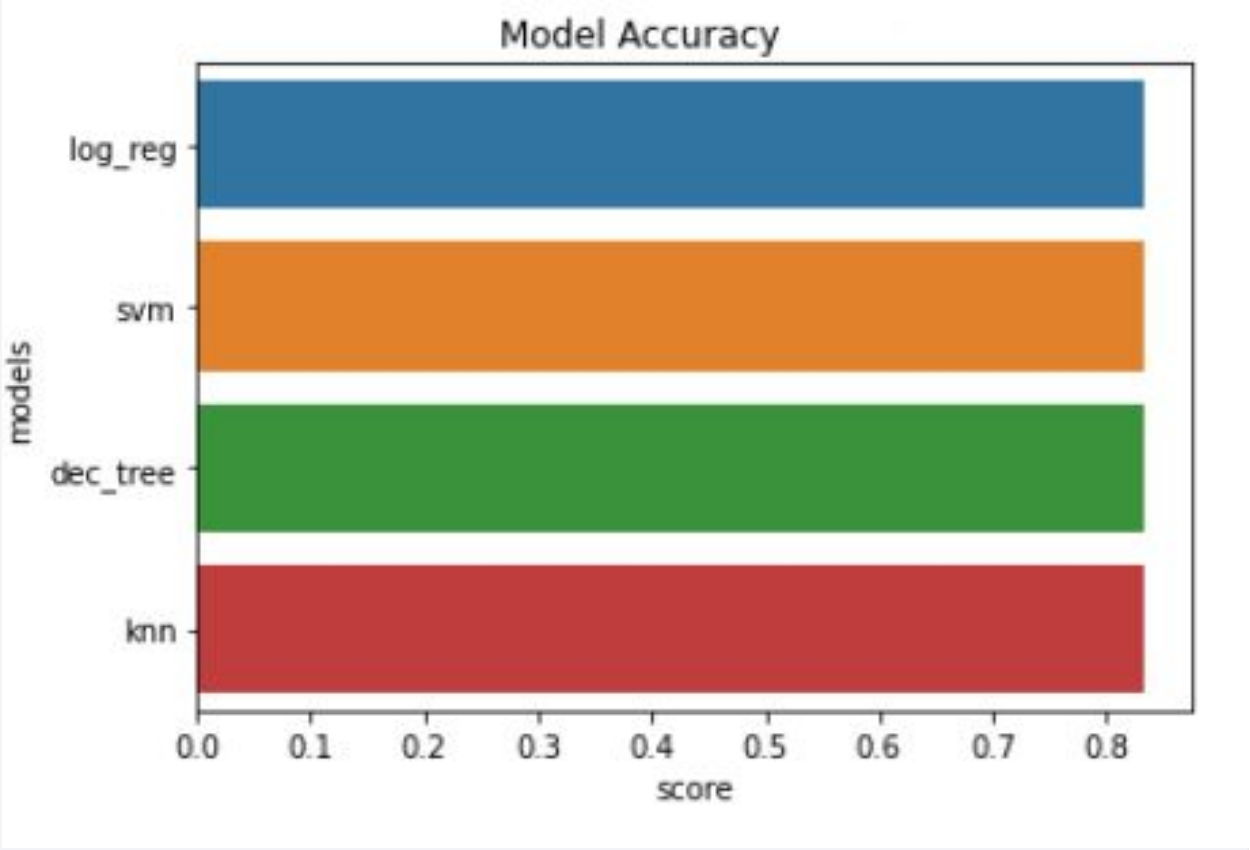
Classification Accuracy

This bar chart compares the accuracy of four classification models: Logistic Regression, SVM, Decision Tree, and KNN.

- All models achieved virtually the same accuracy: 83.33% on the test set.

Key Points:

- The test set sample size is small (n=18), which can increase accuracy variance.
- Decision Tree Classifier accuracy can vary significantly across repeated runs due to the small test set.
- More data is needed to reliably determine the best-performing model.



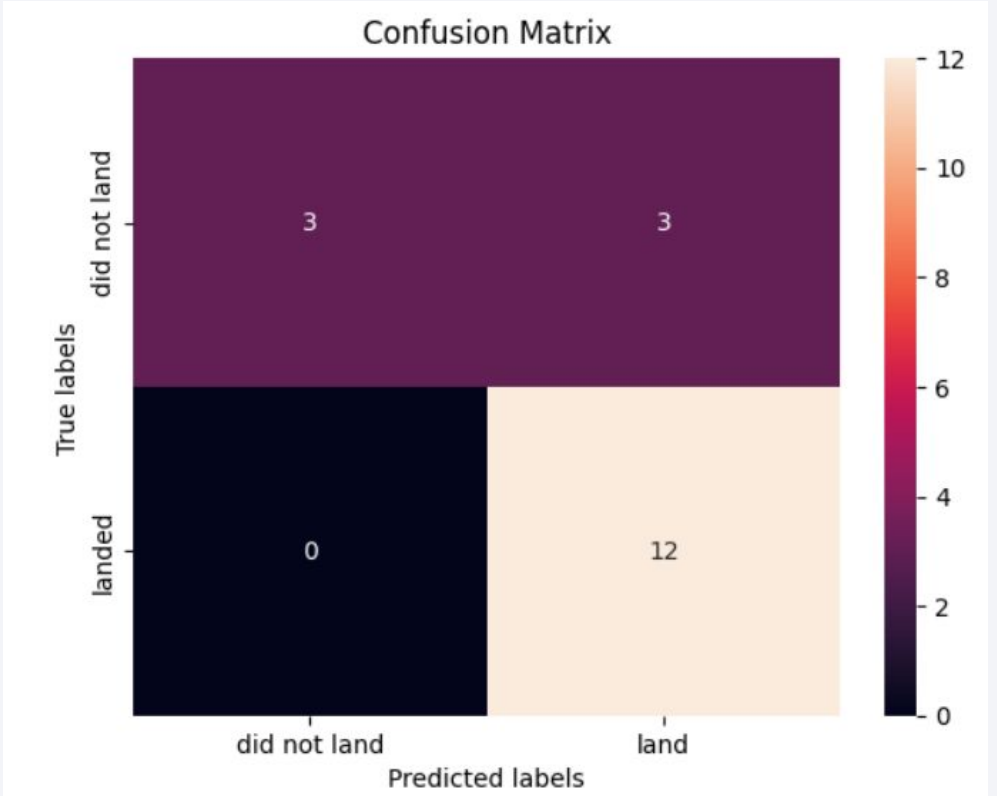
Confusion Matrix

This confusion matrix represents the performance of all four classification models (Logistic Regression, SVM, Decision Tree, and KNN), as they all exhibited identical results on the test set.

- **True Positives:** The models correctly predicted **12** successful landings.
- **True Negatives:** The models correctly predicted **3** unsuccessful landings.
- **False Positives:** The models incorrectly predicted **3** successful landings when the actual outcome was unsuccessful.

Key Findings:

- The models accurately predicted both successful and unsuccessful landings in most cases.
- The models exhibit a tendency to **over-predict successful landings**, as indicated by the false positives. This suggests a potential bias towards classifying outcomes as successful.



Conclusions (Part 1)

Key Conclusion:

- We successfully developed a machine learning model capable of predicting Falcon 9 first stage landing outcomes with approximately 83% accuracy
- This provides a foundation for informed decision-making in launch operations.

Business Impact for SpaceY:

- **Cost Reduction:** By leveraging the model to predict landing success, SpaceY can potentially reduce the financial risk associated with failed landing attempts. ***Each successful landing saves an estimated $100 million in recovery/replacement costs.***
- **Competitive Advantage:** The model empowers SpaceY to optimize launch strategies, increasing their competitiveness against SpaceX by improving mission reliability and cost-efficiency.
- **Operational Efficiency:** Data-driven predictions allow for better planning of recovery operations and resource allocation, streamlining launch procedures.

Predictions and Recommendations:

- **Informed Launch Decisions:** SpaceY can use the model's predictions to assess the viability of a launch based on the probability of a successful landing. This enables a more proactive approach to risk management.
- **Scenario Planning:** The model facilitates scenario planning, allowing SpaceY to evaluate different launch parameters and their potential impact on landing success.

Conclusions (Part 2)

Next Steps:

- **Enhanced Data Acquisition:**
  - Prioritize the collection of more extensive and diverse data to improve model robustness and generalizability.
  - Explore incorporating additional relevant features, such as environmental conditions, hardware specifications, and historical maintenance records.
- **Model Refinement:**
  - Conduct further model evaluation and selection to identify the optimal algorithm for this prediction task.
  - Implement techniques to address class imbalance (if present) and reduce the model's tendency to over-predict successful landings.
  - Fine-tune model hyperparameters to optimize performance and reduce variance.
- **Real-time Integration:**
  - Investigate the feasibility of integrating the model into real-time launch control systems to provide dynamic decision support during launch operations.
  - Develop a system for continuous model monitoring and retraining to ensure sustained accuracy and adapt to evolving launch conditions.

Appendix

- **GitHub Repository:**
  https://github.com/Jtrusko/Applied_Data_Science_Capstone/tree/main

- **Book References:**

- Grus, Joel. *Data Science from Scratch: First Principles with Python*. O'Reilly Media, 2019. ISBN: 978-1-492-04114-9.
- Nield, Thomas. *Essential Math for Data Science: Take Control of Your Data with Fundamental Linear Algebra, Probability, and Statistics*. No Starch Press, 2022. ISBN: 978-1-098-10293-7.

Thank you!