IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohammad Ashna

September 2023

Outline

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

Executive Summary

The objective of this project was to build a machine-learning model capable of predicting the success of first-stage rocket landings for future launches by a new company, ‘**SpaceY.**’ This was accomplished by analyzing historical data from SpaceX rocket launches, which is publicly available via a REST API.

The project involved:

- Collecting and processing data through **web scraping** (Wikipedia) and **SpaceX's REST API.**
- **Cleaning** and refining the data to ensure accuracy and usability for analysis.
- Conducting **exploratory data analysis (EDA)** to uncover patterns and relationships.
- **Visualizing** key insights using Python libraries and interactive dashboards.
- Developing and optimizing **machine-learning** classification models, with a **Decision Tree** achieving the highest accuracy.

The findings indicate that missions involving **Falcon 9 rockets (FT version)** sent to low Earth orbits with payloads between 2,000 and 5,000 kg exhibited the highest success rates. These insights provide a solid foundation for further research and operational planning.

Introduction

The race to explore skies and beyond has been ongoing for more than a century. In the past, aerospace projects were predominantly funded and operated by governments for strategic military objectives. However, the 21st century has witnessed a paradigm shift, with the **private sector** spearheading advancements in space exploration. Companies like SpaceX have revolutionized the industry by significantly reducing the cost of space transportation through innovations such as **reusable rocket boosters.**

This project focuses on developing a **machine-learning** classification model to predict the success of first-stage rocket landings for future launches by '**SpaceY**,' a hypothetical company. By analyzing SpaceX's historical **Falcon 9** landing data, we aim to uncover insights into factors influencing mission success. This study not only highlights the potential of **data science** in addressing real-world challenges but also lays the groundwork for cost-efficient space operations.

Section 1

Methodology

Methodology

- Data collection
  - Scraping data from Wikipedia
  - Obtaining data from SpaceX REST API
- Data wrangling
  - Preprocessing data for the next steps
- Exploratory data analysis (EDA) using visualization and SQL
- Development of Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Building, tuning, and evaluating classification models using Scikit-learn

Data Collection

There are two different ways of collecting needed information:

- **Private** corporate sources, such as the **SpaceX REST API**
- **Public** sources, such as Wikipedia, using **web scraping**

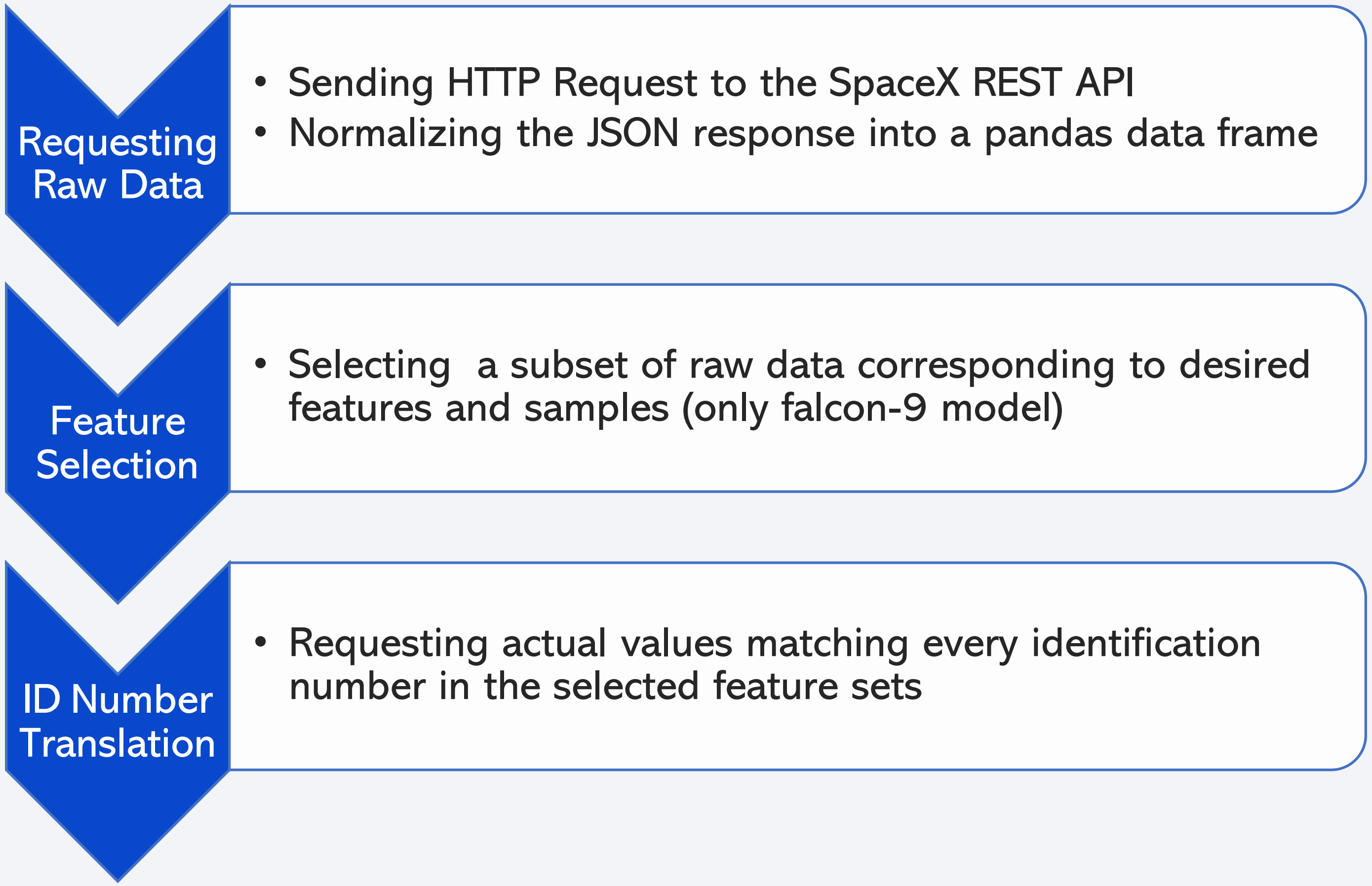
Each source has its own set of challenges and advantages; data extracted using **SpaceX REST API** is more reliable and requires less cleaning, but since the dataset is large, it requires to be filtered down to a specific set of **hand-picked variables** for the sake of **processing efficacy.**

Wikipedia, on the other hand, offers more concise tabular data better suited for the purpose of this study. However, since it is in the **HTML** format, more work is needed to parse the HTML and refine the data to a format appropriate for analysis.

Data Collection – SpaceX API

- The flowchart presents the data collection process using SpaceX REST API

- **For more information, please refer to Jupyter Notebook [GitHub]**



Requesting Raw Data

- Sending HTTP Request to the SpaceX REST API
- Normalizing the JSON response into a pandas data frame

Feature Selection

- Selecting a subset of raw data corresponding to desired features and samples (only falcon-9 model)

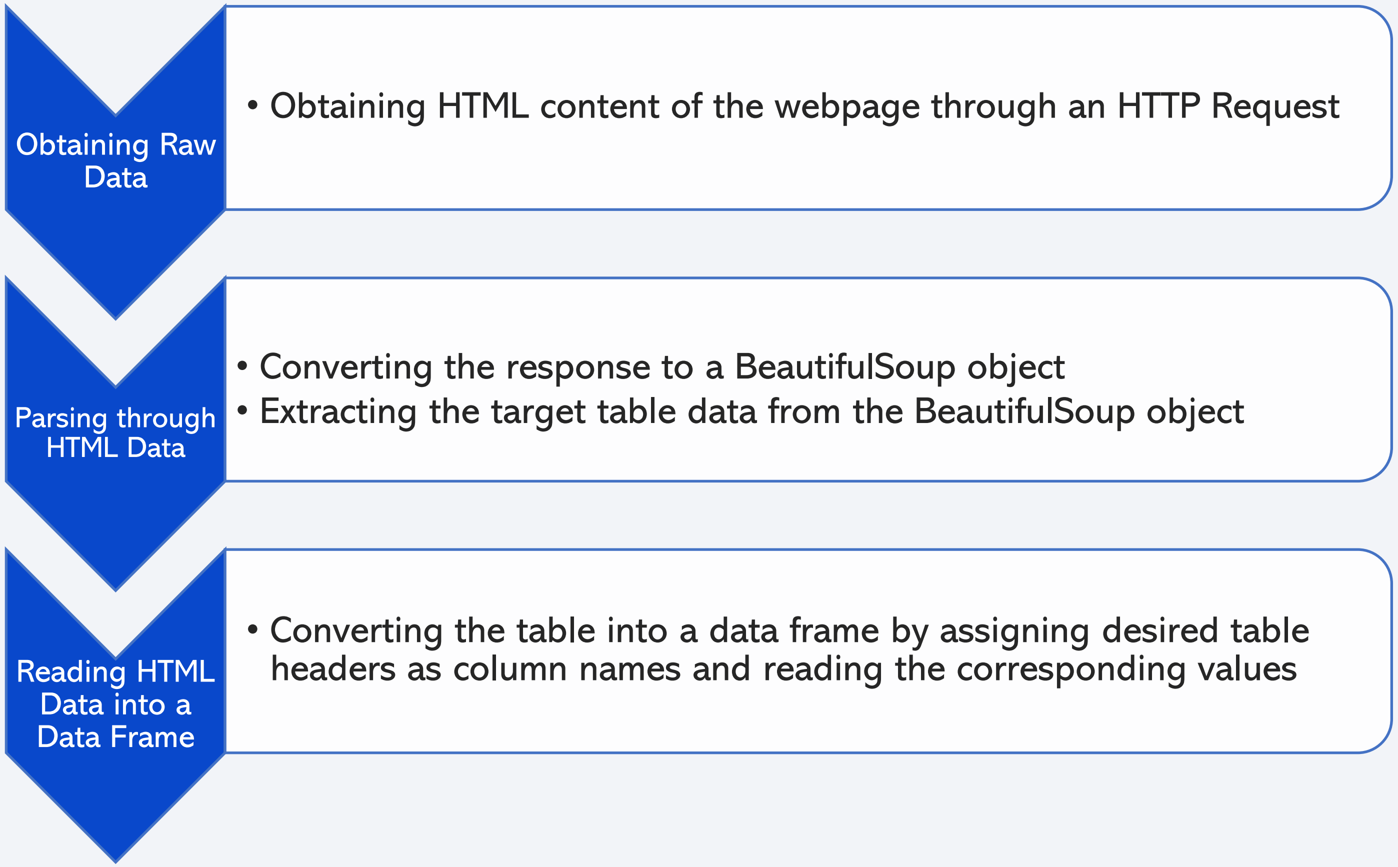
ID Number Translation

- Requesting actual values matching every identification number in the selected feature sets

Data Collection - Scraping

- The flowchart presents the process of collecting data from the source website HTML content using the web scraping method

- **For more information, please refer to Jupyter Notebook on GitHub**



Obtaining Raw Data

- Obtaining HTML content of the webpage through an HTTP Request

Parsing through HTML Data

- Converting the response to a BeautifulSoup object
- Extracting the target table data from the BeautifulSoup object

Reading HTML Data into a Data Frame

- Converting the table into a data frame by assigning desired table headers as column names and reading the corresponding values

Data Wrangling

First, we search for **missing (NaN)** or **vague (e.g., None)** values. In this case, only one set of features, “landing pads,” have missing values. Therefore, we could either omit this feature from the data set altogether or replace the missing values with the mode (the most frequently repeated value), depending on which option would have the **least possible effect** on the outcome.

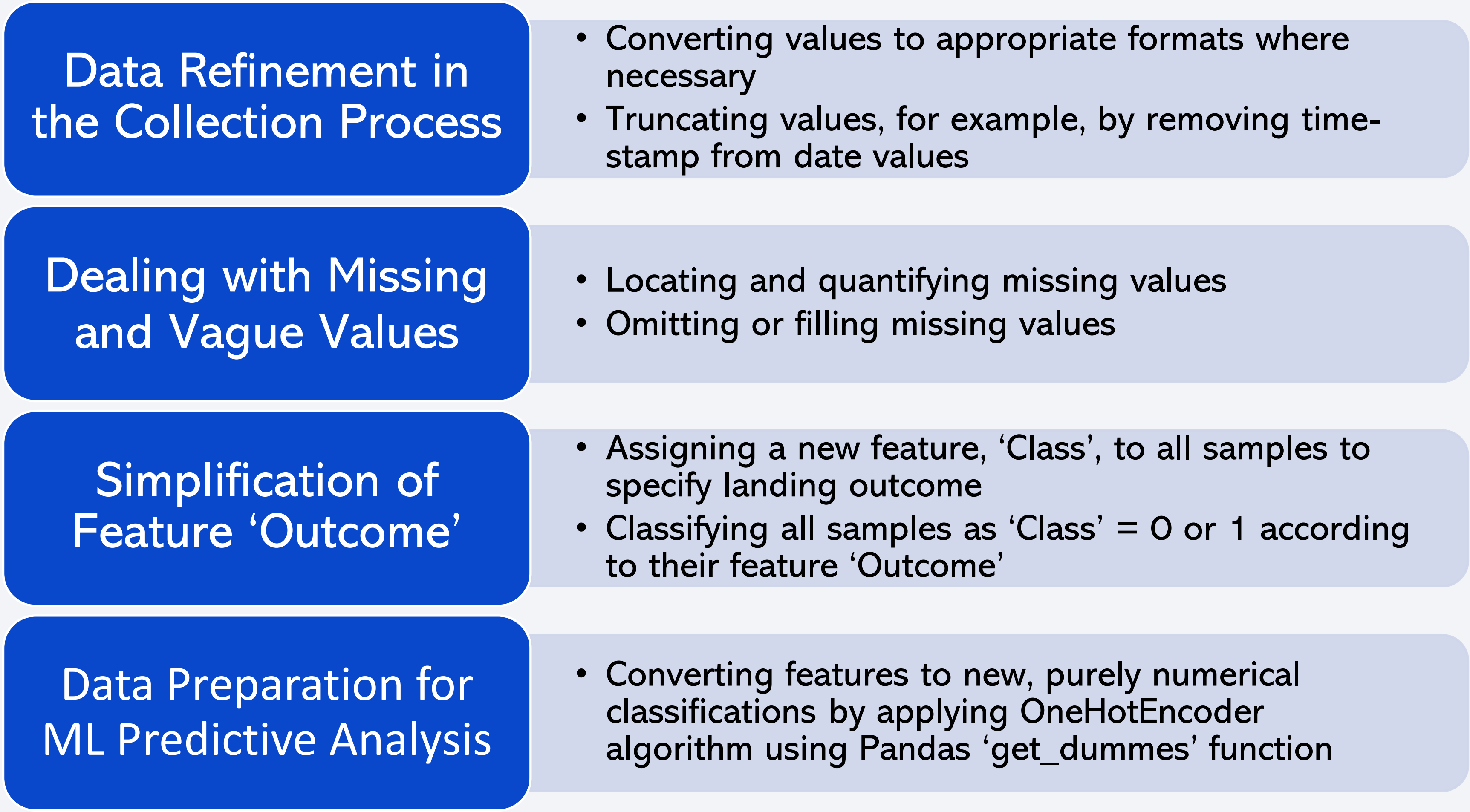
In the next step, we need to **simplify** some features. Since this analysis aims to predict the success of landing, the seven different values of the feature outcome can be **classified into two classes: ‘0’, meaning failure, and ‘1’, meaning success.**

Nonetheless, Some other data refining steps were taken as part of the data collection and EDA with data visualization processes, such as:

- Converting values to **appropriate formats** where necessary
- Truncating values, for example, by **removing time-stamp** from date values
- Data preparation for **machine-learning** analysis using **OneHotEncoder**

For more information, please refer to the Data Wrangling Jupyter Notebook on GitHub

Data Wrangling Flowchart



Data Refinement in the Collection Process

- Converting values to appropriate formats where necessary
- Truncating values, for example, by removing time-stamp from date values

Dealing with Missing and Vague Values

- Locating and quantifying missing values
- Omitting or filling missing values

Simplification of Feature ‘Outcome’

- Assigning a new feature, ‘Class’, to all samples to specify landing outcome
- Classifying all samples as ‘Class’ = 0 or 1 according to their feature ‘Outcome’

Data Preparation for ML Predictive Analysis

- Converting features to new, purely numerical classifications by applying OneHotEncoder algorithm using Pandas ‘get_dummes’ function

EDA with Data Visualization

We plotted **scatterplots** of different features and overlayed each plot with the outcome of the launch (feature ‘Class’) to see the distribution of samples and the relationship between **variables**. Scatterplots of the following feature pairs have been visualized:

- FlightNumber vs. PayloadMass
- FlightNumber vs. Launch Site
- Orbit vs. FlightNumber
- Orbit vs. PayloadMass

A **bar chart** of landing success rate (mean of feature ‘Class’) by **orbit** was also plotted to visualize the relationship between the two features.

Additionally, the **yearly success rate** was also plotted as a line chart to explore the trend of success rate improvement over a decade between 2010 and 2020.

For more information, please refer to EDA with Data Visualization Jupyter Notebook on [GitHub]

EDA with SQL

The following is a list of the SQL queries done as part of this project:

- Selecting distinct launch site names
- Selecting five records where launch sites begin with the string 'CCA’
- Displaying total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by Falcon 9 boosters of version v1.1
- Selecting the date when the first successful landing on the ground pad was achieved
- Listing the names of boosters with payload mass between 4000 and 6000 which had successful landings on a drone ship
- Listing the total number of successful and failed mission outcomes
- Using a subquery to find booster versions that have carried maximum payload mass
- Selecting the records containing booster versions, launch site and month name of failed drone ship landings for each month in the year 2015.
- Sorting the number of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between 2010-06-04 and 2017-03-20, in descending order.

For more information, please refer to EDA with SQL Jupyter Notebook on GitHub

Build an Interactive Map with Folium

The following is a list of items added to the map and their functions:

- **A circle and pop-up** for each launch site to ease finding sites on the map
- **A text marker** indicating the name of each site
- **A marker cluster** displaying an icon for every successful and failed landing in each site using the colors green and red, respectively
- **A text marker** indicating the distance of each site from an adjacent coastline
- **Mouse pointer** position to facilitate finding coordinates of any given point on the map
- **Direct lines** between three launch sites in Florida and a **coastline** in their proximity
- **A direct line** visualizing the distance between the launch site in California and its nearest **railroad**

For more information, please refer to the Interactive Folium Map Jupyter Notebook on GitHub

Build a Dashboard with Plotly Dash

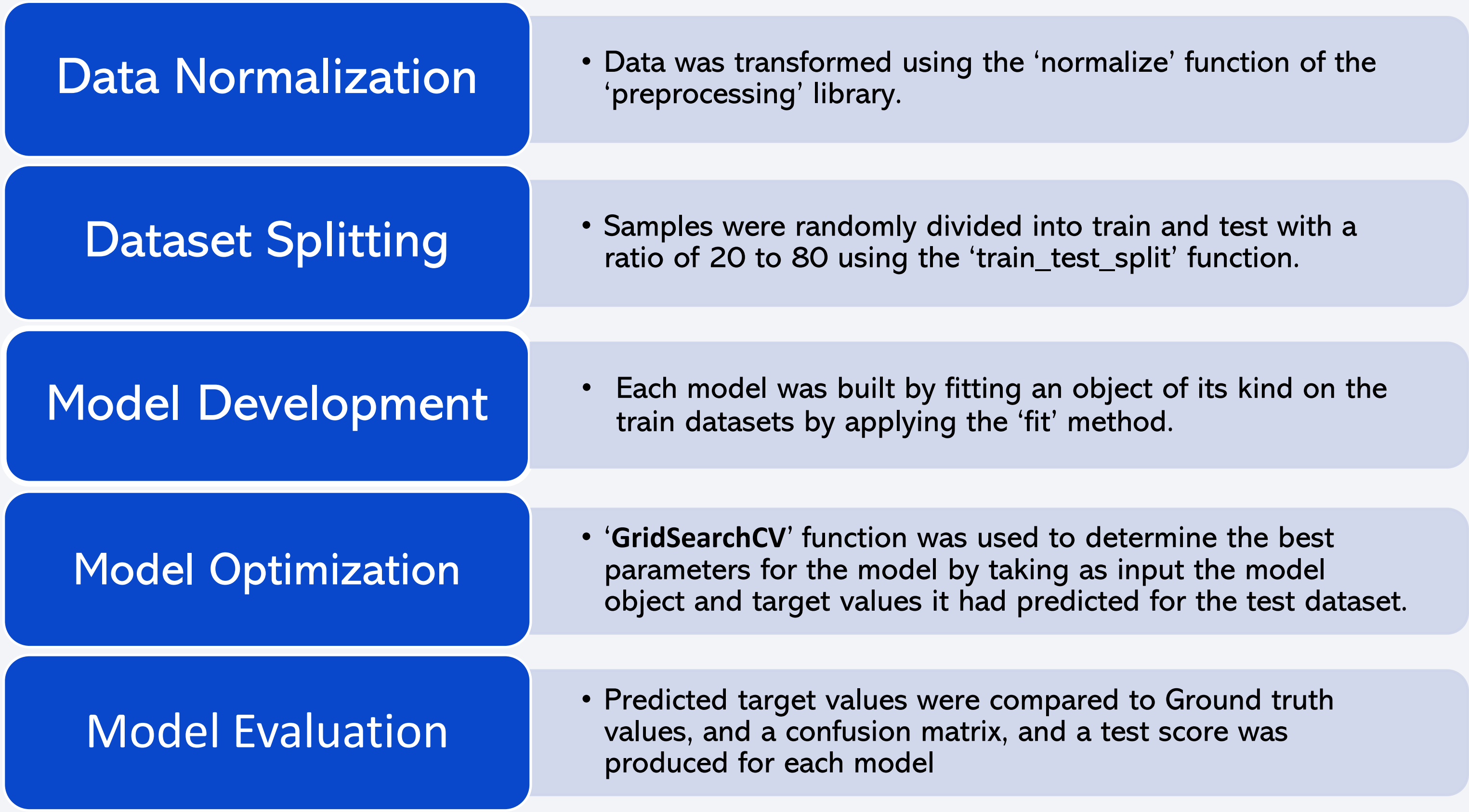
Following is a list of items added to the dashboard application and their functions:

- **A dropdown menu** containing four options for the user to select all or one of the launch sites
- **A range slider** that reads the user input for payload mass in Kg
- **A pie chart** that either compares the proportion of success versus failure for an individual site or illustrates each site’s share of the total number of successful landing missions
- **A scatter plot** presenting the distribution of payload mass and booster version as regards to landing outcome

For more information, please refer to the interactive Plotly Dash Jupyter Notebook on GitHub

Predictive Analysis (Classification)

- Different libraries of the **‘Scikit-Learn’** package were used for each step of the predictive analysis. The flowchart summarizes the process.
- **For more information, please refer to the Interactive Plotly Dash Jupyter Notebook on [GitHub]()**



Data Normalization

- Data was transformed using the ‘normalize’ function of the ‘preprocessing’ library.

Dataset Splitting

- Samples were randomly divided into train and test with a ratio of 20 to 80 using the ‘train_test_split’ function.

Model Development

- Each model was built by fitting an object of its kind on the train datasets by applying the ‘fit’ method.

Model Optimization

- ‘**GridSearchCV**’ function was used to determine the best parameters for the model by taking as input the model object and target values it had predicted for the test dataset.

Model Evaluation

- Predicted target values were compared to Ground truth values, and a confusion matrix, and a test score was produced for each model

Results

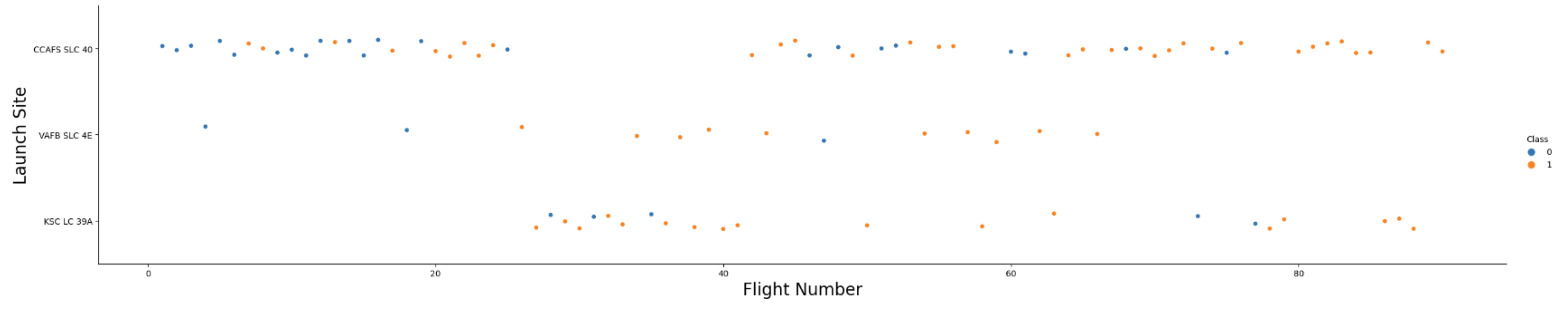
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

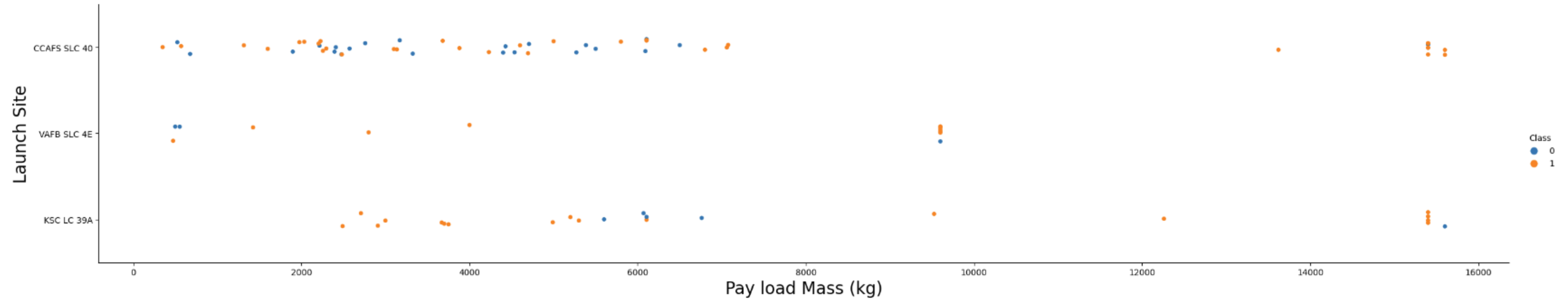
Flight Number vs. Launch Site

- The scatter plot below illustrates the **distribution of 'flight numbers'** across three different launch sites in North America. Most missions have been conducted from the **CCAFS SLC 40.** However, the site **KSC LC 39A** was entirely preferred over the former for the first launches it hosted. **VAFB SLC 4E** operations, on the other hand, were ceased for flight numbers **greater than 50.**



Payload vs. Launch Site

- The plot illustrates the **distribution of 'payload mass (Kg)'** for missions in three U.S. sites. Most rockets launched from the **CCAFS SLC 40** site had a payload lighter than **7000 Killo grams. KSC LC 39A**, the second most used site, had similar results, with the bulk of its cargo being in the thousand **2-7 Kg** range. **VAFB SLC 4E** payloads, however, could be grouped into two groups: the **below-4000** range and **9.5 tonnes.**

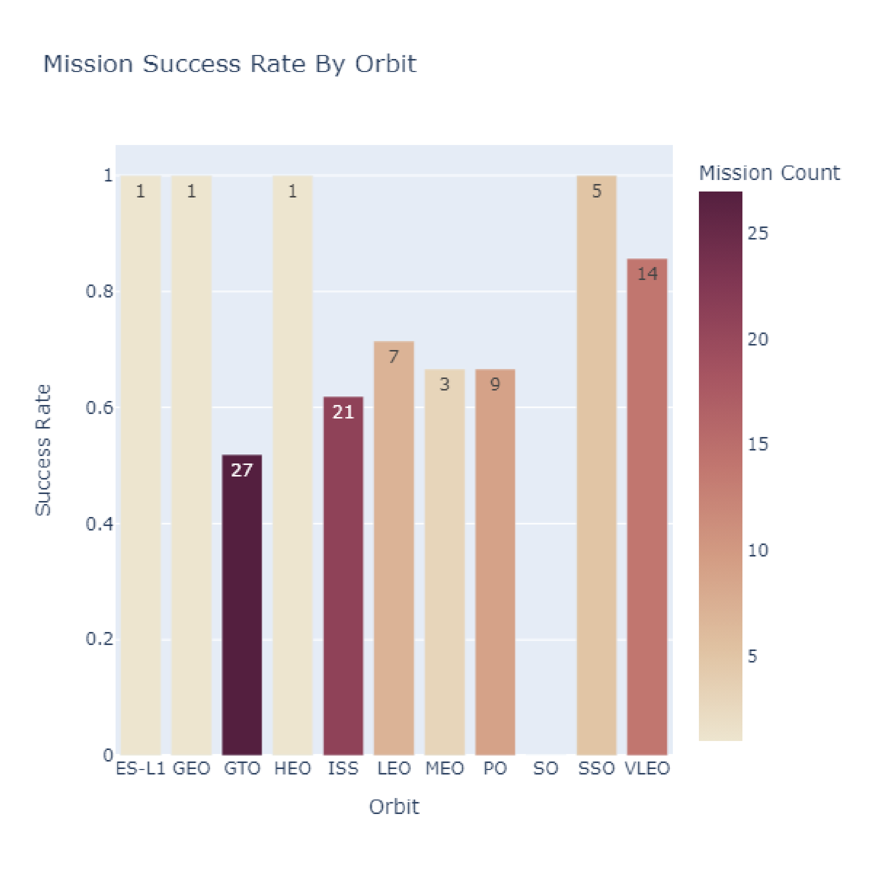


Success Rate vs. Orbit Type

- The bar chart compares the **cumulative success rates** of missions sent to **different categories of earth orbits.** The bars are marked by a **continuous range of colors** according to the number of samples in each category. This figure is additionally shown as a label on the upper end of each bar.

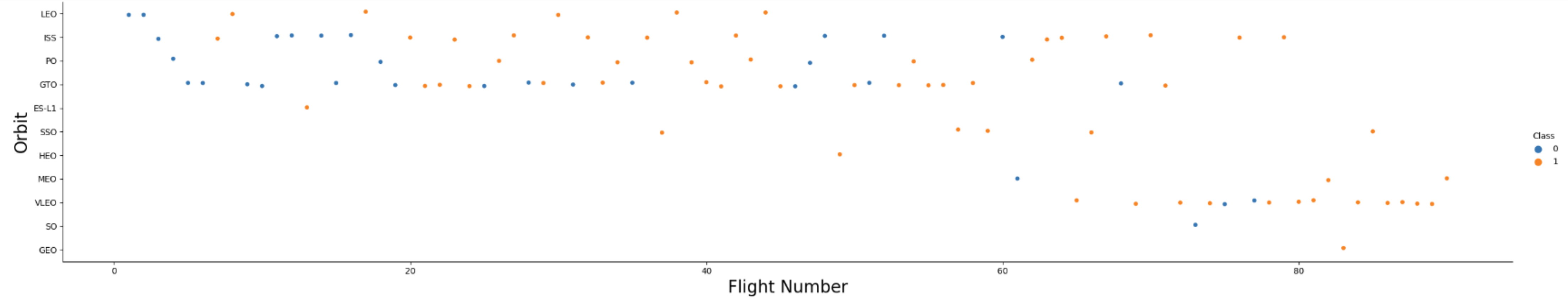
- Despite having **%100 success**, the four most successful orbit types only comprise an **insignificant portion** of all launch records. Therefore, these orbits could be considered **outliers** and subsequently **removed** from the data set

- Overall, missions to orbits of **low and very low** altitudes, including **VLEO, LEO, ISS, and PO**, appear **more successful** than those of higher orbits such as 'GTO.'



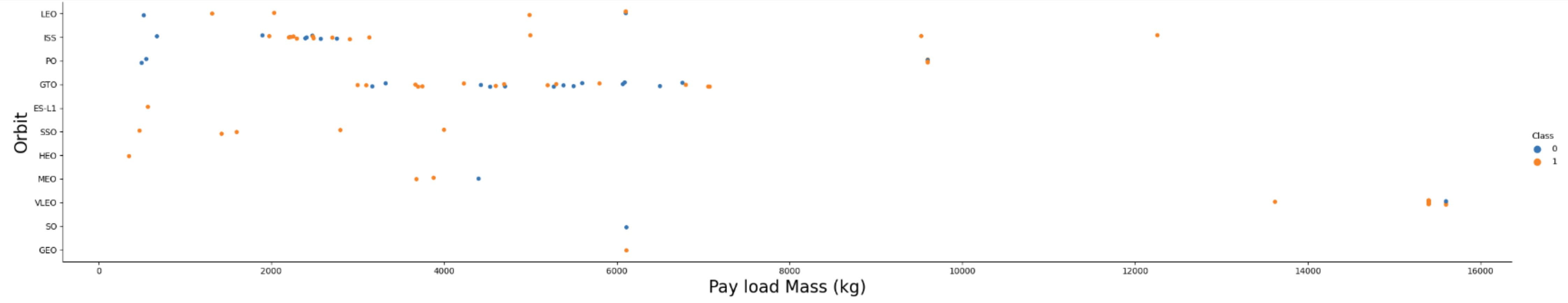
Flight Number vs. Orbit Type

- The plot demonstrates **flight number distribution** of missions to different types of orbits. Most of all flight numbers **below 60** belong to lower earth orbits, among which flight numbers **greater than 20** appear to have performed more successfully. Followed Flight numbers above 60 were mostly associated with **VLEO** and **ISS** and manifest **an overall higher success rate** compared to the rest.



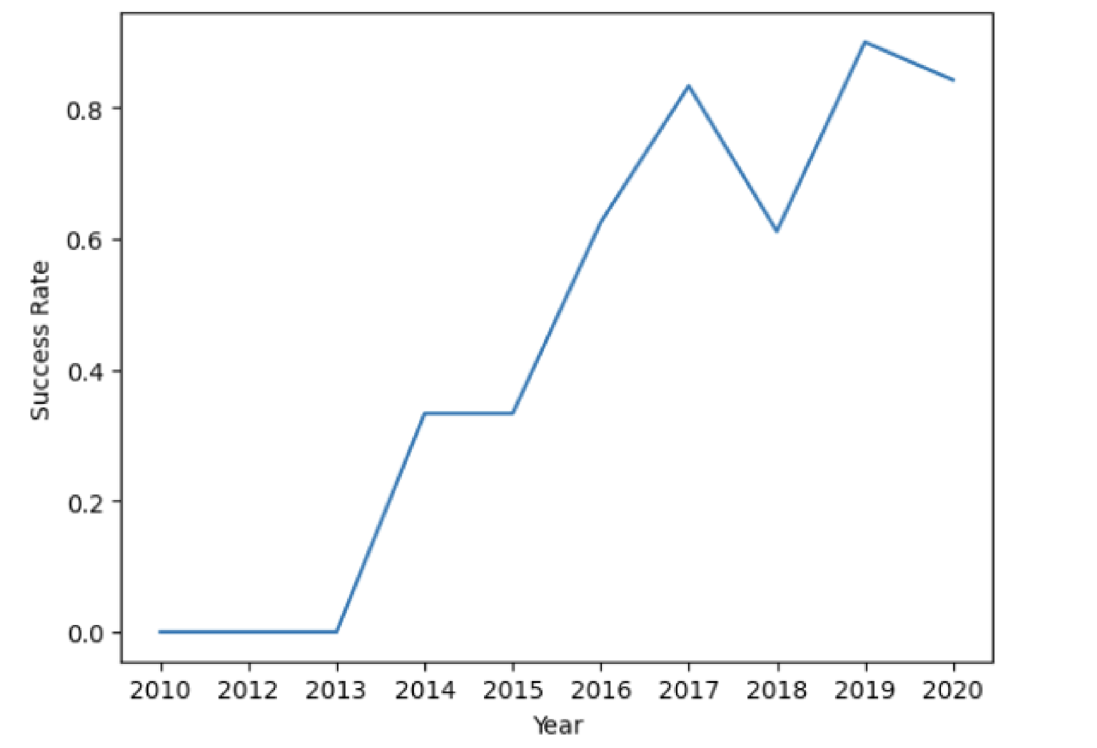
Payload vs. Orbit Type

- The plot demonstrates the **mass distribution** of payloads carried to orbits of distinct categories. All rockets launched to **GTO** had a payload mass in the range of **2000-7000 Kg**, similar to **ISS**, with the bulk of its cargo being **between 2 and 3 tonnes.** Among orbit types with masses above the 9000 level, **VLEO** had the **heaviest** payloads ranging from **13 – 16 thousand Kg.**



Launch Success Yearly Trend

- The line graph shows the trend of **annual mission success rates** over a decade-long period between 2010 and 2020.
- **A general uptrend** is evident for most of the period. However, it was only in **2014** that the **first success rate greater than ‘0’** was recorded at just below 0.4.
- Despite remaining constant for another year, the success rate **surged dramatically** to surpass the %80 level at the end of 2017.
- For the remainder of the period, however, the figure **fluctuated** to finally sit at 0.8 in 2020.



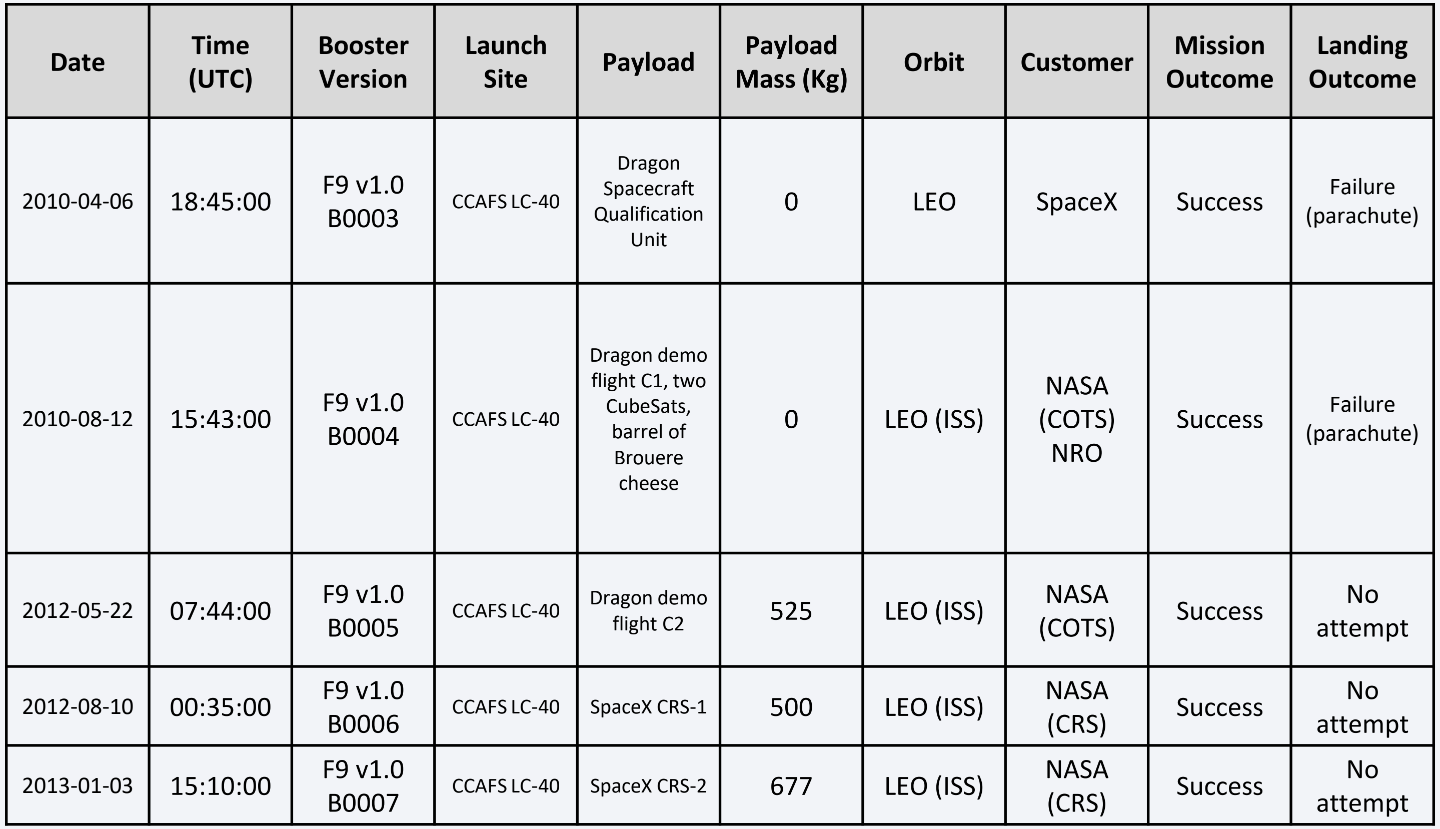
All Launch Site Names

The names of unique launch sites are as follows:

Launch Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

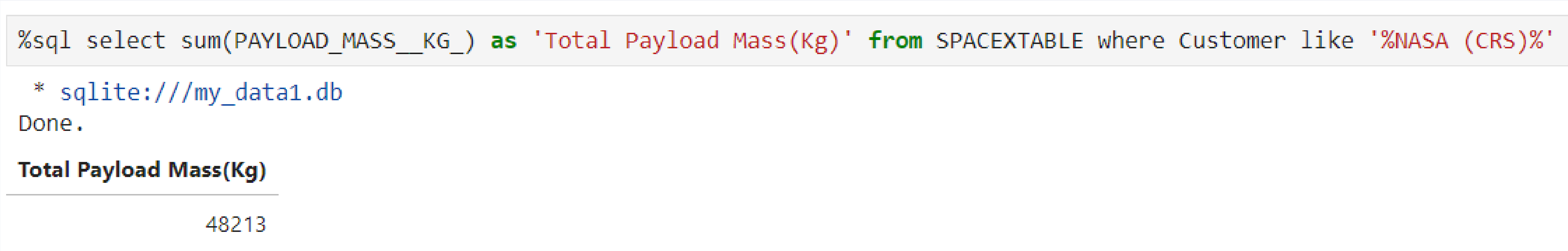
- The table shown here is the result of an inquiry for records where launch sites begin with `CCA`



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

The total payload carried by boosters from NASA is 48213 Kg, which was obtained using the following line of code:



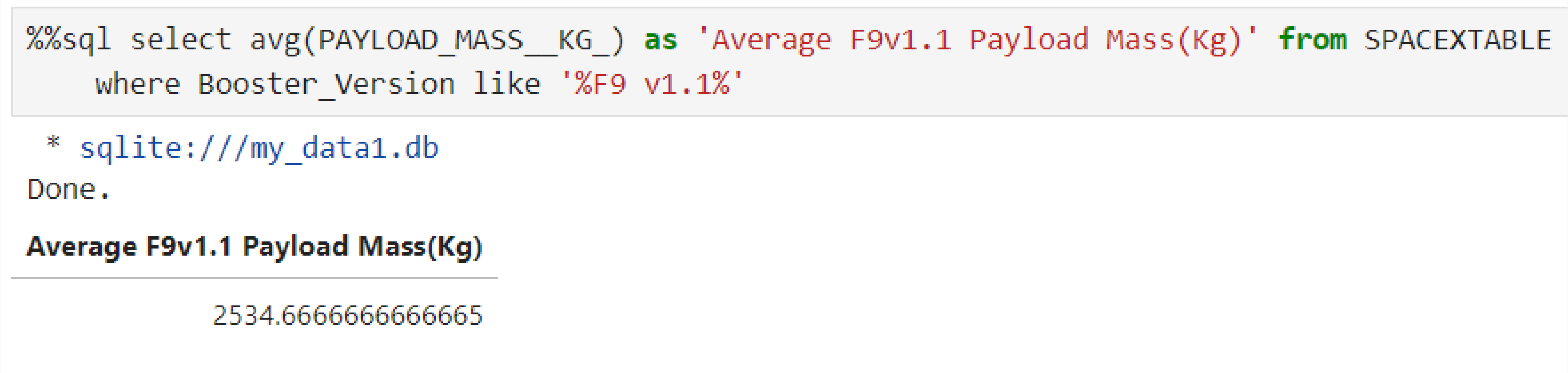
```
%sql select sum(PAYLOAD_MASS__KG_) as 'Total Payload Mass(Kg)' from SPACEXTABLE where Customer like '%NASA (CRS)%'
```

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

Total Payload Mass(Kg)
48213

Average Payload Mass by F9 v1.1

$2534.\overline{6}$ is the average payload mass carried by booster version F9 v1.1, which is calculated through an SQL inquiry as shown below:



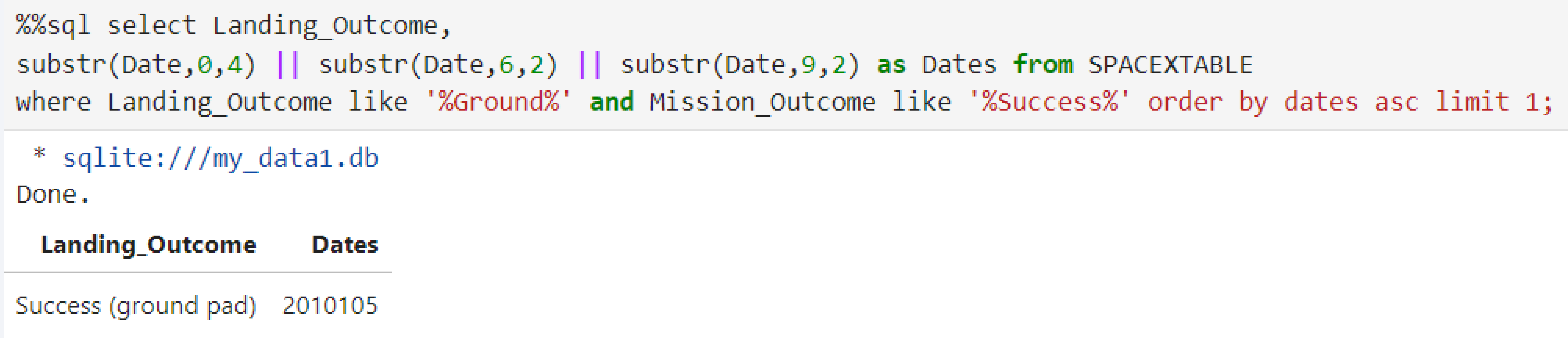
```
%%sql select avg(PAYLOAD_MASS__KG_) as 'Average F9v1.1 Payload Mass(Kg)' from SPACEXTABLE
    where Booster_Version like '%F9 v1.1%'
```

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

Average F9v1.1 Payload Mass(Kg)
2534.6666666666665

First Successful Ground Landing Date

As indicated by the picture below,10 October 2010 was the date when the first successful landing on a ground pad was achieved by a SpaceX Falcon-9 system.



```
%%sql select Landing_Outcome,
substr(Date,0,4) || substr(Date,6,2) || substr(Date,9,2) as Dates from SPACEXTABLE
where Landing_Outcome like '%Ground%' and Mission_Outcome like '%Success%' order by dates asc limit 1;
```

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

Landing_Outcome	Dates
Success (ground pad)	2010105

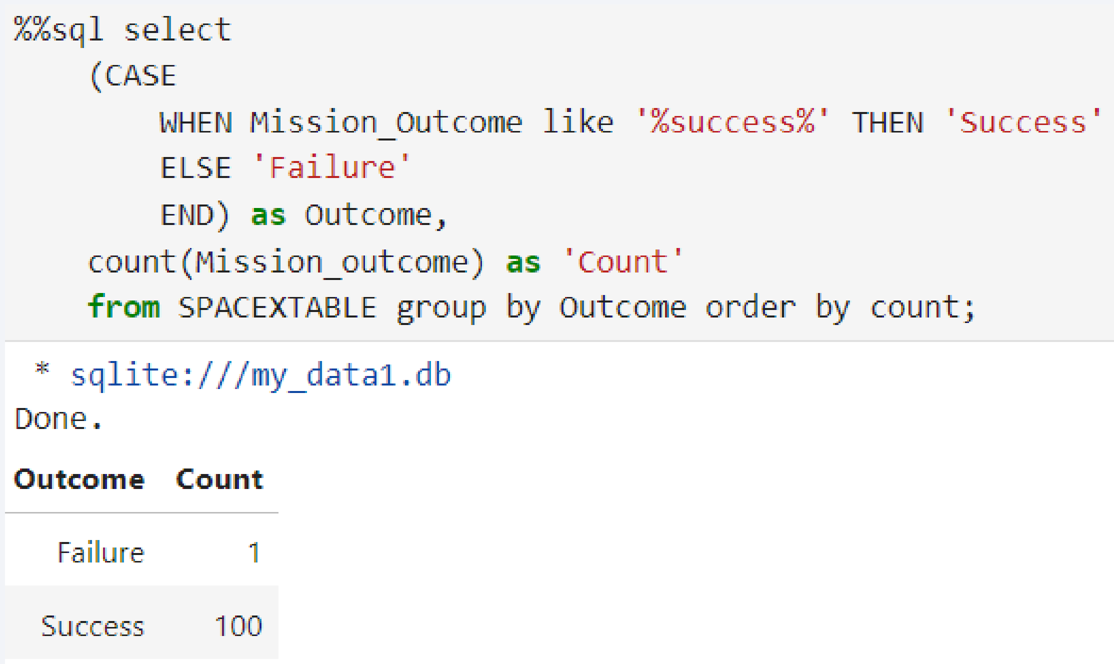
Successful Drone Ship Landing with Payload between 4000 and 6000

The list contains the names of boosters that have successfully landed on drone ships and had payload mass in the 4000–6000 Kg range.

Booster Name	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

As shown by the picture below, the database had 100 records of successful landings compared to one instance of failure.



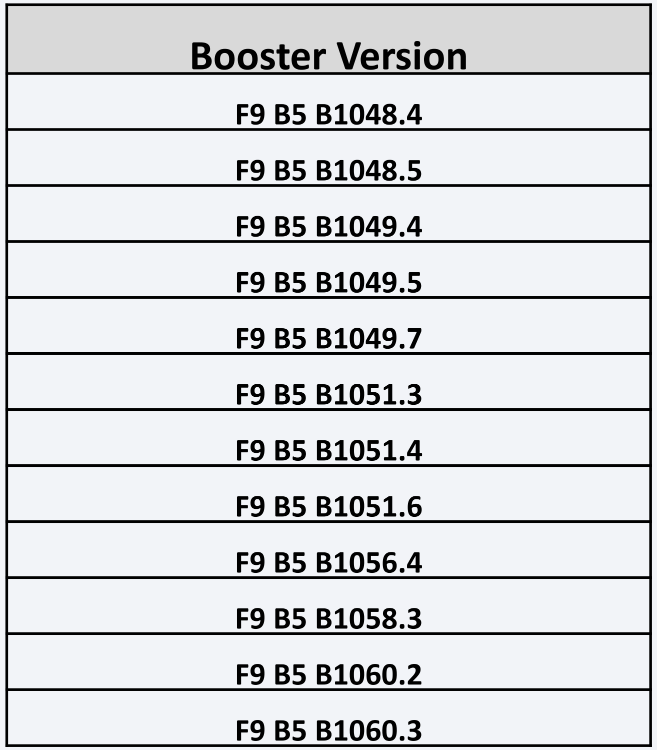
```
%%sql select
    (CASE
        WHEN Mission_Outcome like '%success%' THEN 'Success'
        ELSE 'Failure'
        END) as Outcome,
    count(Mission_outcome) as 'Count'
    from SPACEXTABLE group by Outcome order by count;
```

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

Outcome	Count
Failure	1
Success	100

Boosters Carried Maximum Payload

The table contains a list of boosters that have carried the maximum payload mass of 15600 Kg.



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

2015 Launch Records

The following table contains data on failed drone ship landing attempts in the year 2015:

Year	Month	Landing Outcome	Booster Version	Launch Site	COUNT
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	1
2015	06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40	1
2015	10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	1

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

The table below presents the number of samples for every possible value of landing outcome between 2010-06-04 and 2017-03-20, sorted in descending order:

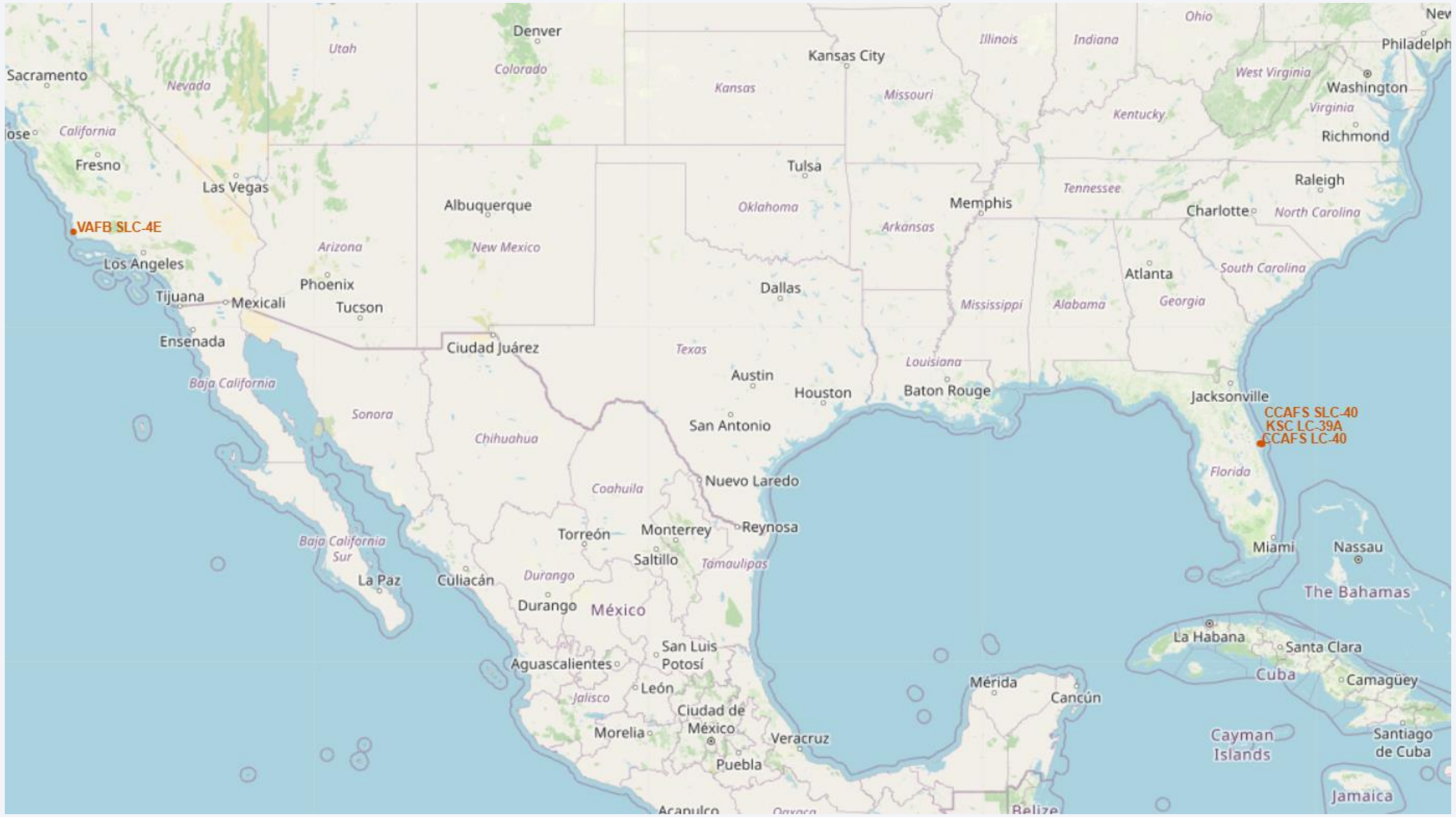
Landing_Outcome	COUNT
No attempt	20
Success	17
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1
Failure	1

Section 3

Launch Sites Proximities Analysis

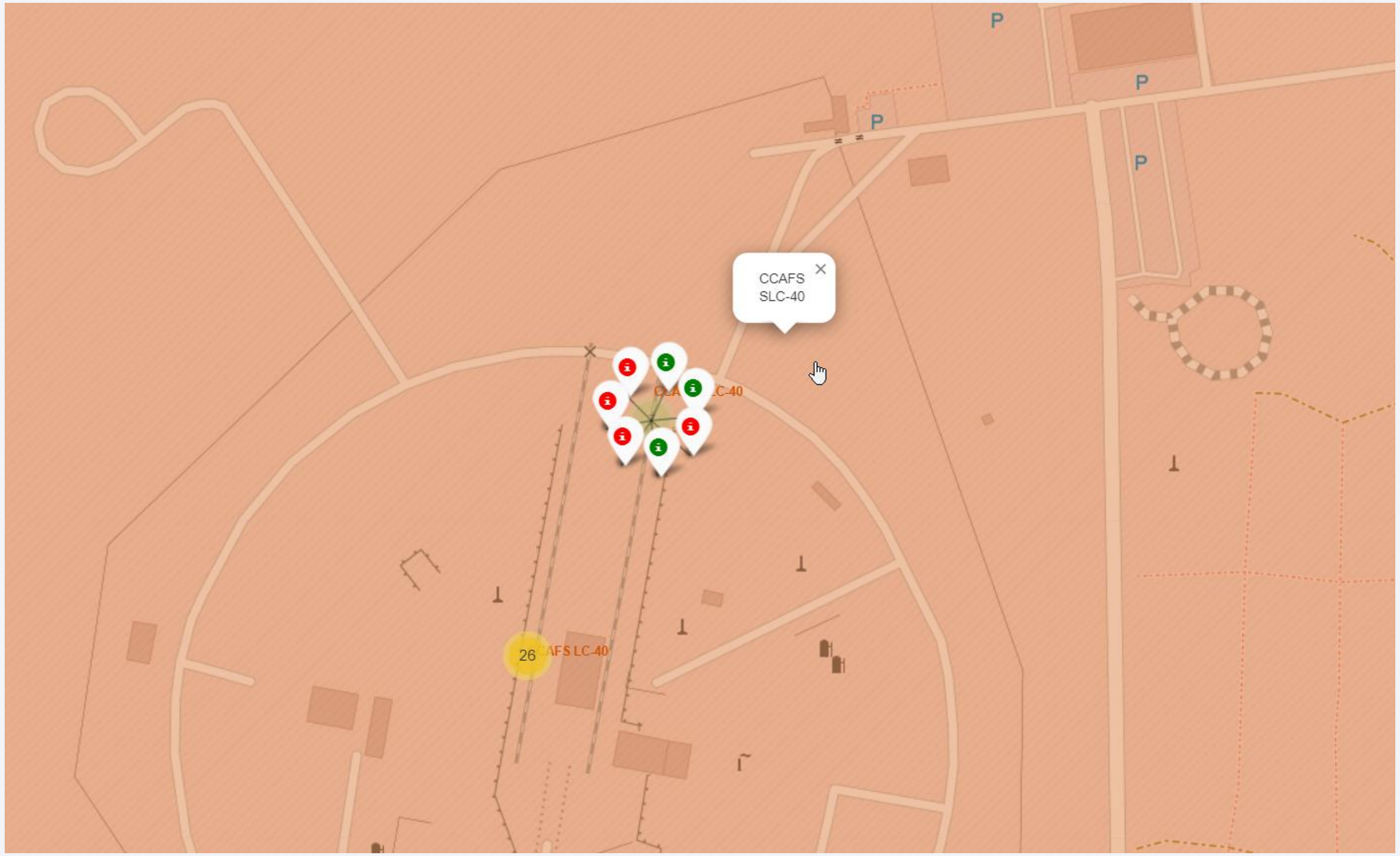
Folium Map: All launch sites location markers on a global map

- The map shows four different launch sites in the USA used by SpaceX for missions involving Falcon-9
- All sites are located near coastlines in furthermost parts of the southern U.S.
- There is only one site in California, whereas there are three other sites near the eastern shores of Florida



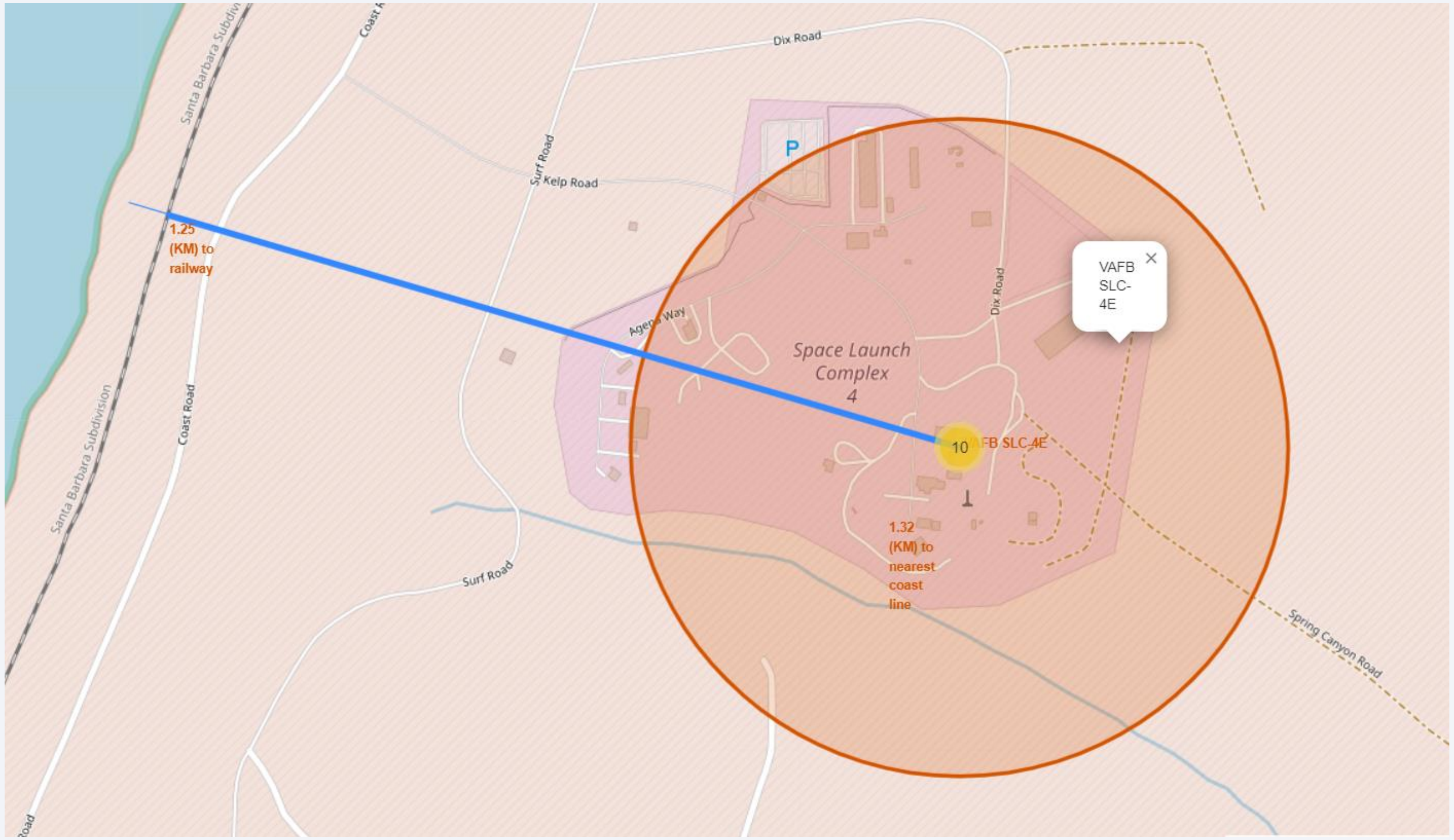
Folium Map: Mission outcomes labelled by color on the map

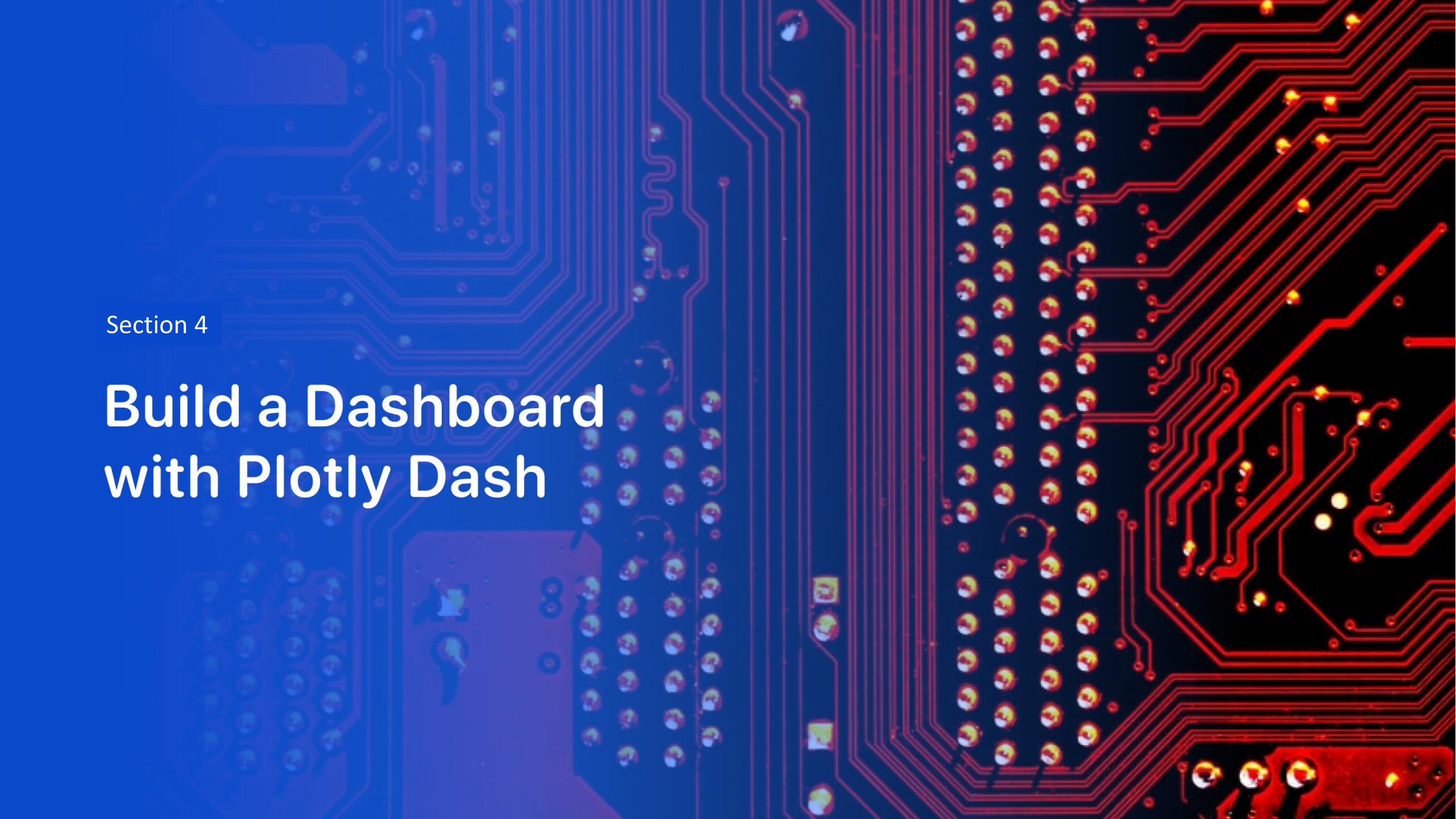
- The map shows color-labeled mission outcomes for every launch site in the U.S.
- Icons marked ‘green’ indicate successful landings, and those colored ‘red’ represent mission failure.



Folium Map: A launch site distance to the nearest railway

- The blue line connects the launch site in California to the nearest railway
- Also, the actual distance between the site and the railway is displayed as a marker on the left side of the picture.



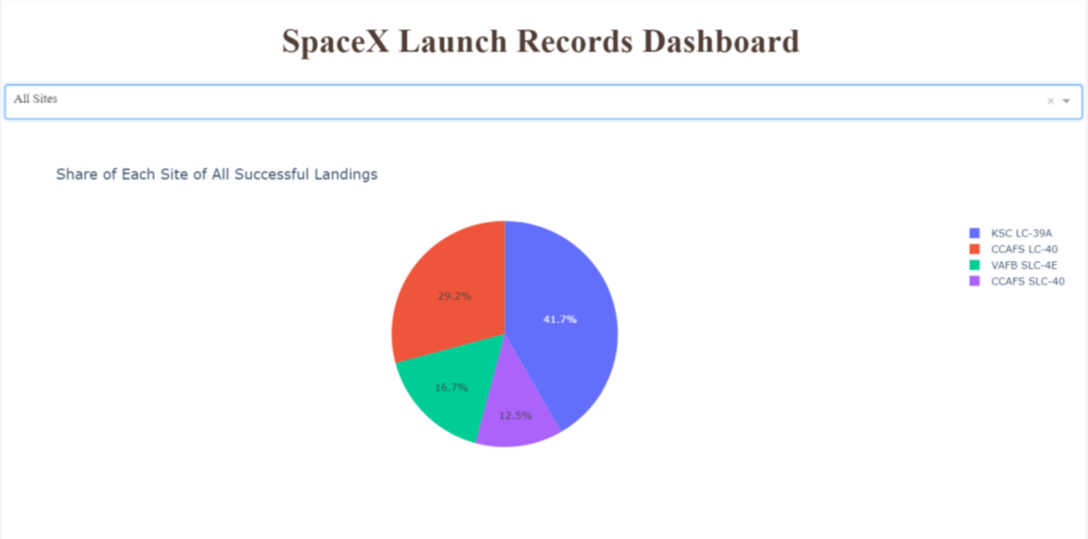


Section 4

Build a Dashboard with Plotly Dash

Sites’ Share of Total Successful Landings

Having the **largest** number of **accomplished** landing missions with approximately 42 percent of all positive outcomes, launch site **KSC LC-39A** is followed by another site located in **Florida**, **CCAF5 LC-40**, at roughly %30 and then **California’s Space Force airbase** with a share of 16.7 percent. Nonetheless, only half a quarter of all successful landings were achieved in the **CCAF5 SLC-40** site, another airbase in the eastern state of **Florida**.

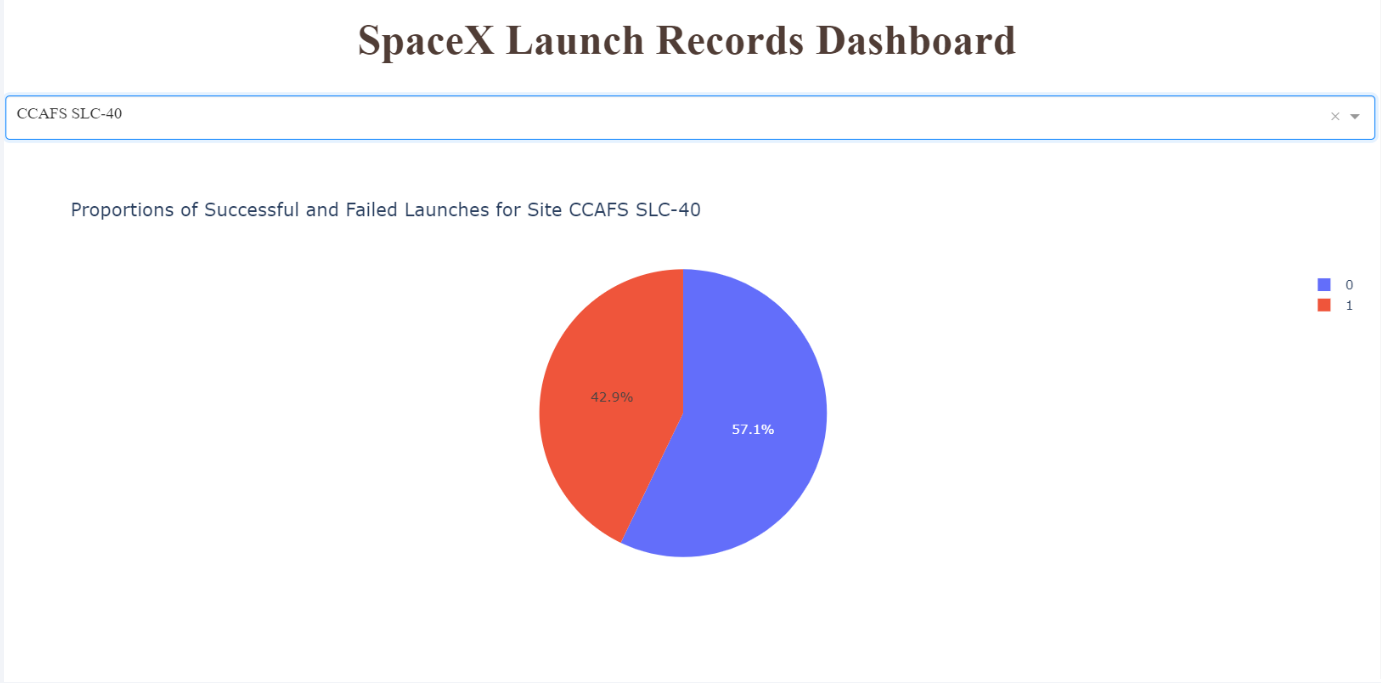


SpaceX Launch Records Dashboard

All Sites

The Site with the Highest Success-to-Failure Ratio

With more than 40 percent success in landing, launch **site CCAFS SLC-40** in **eastern Florida** has the **greatest success-to-failure ratio** of landing missions among all sites in America.



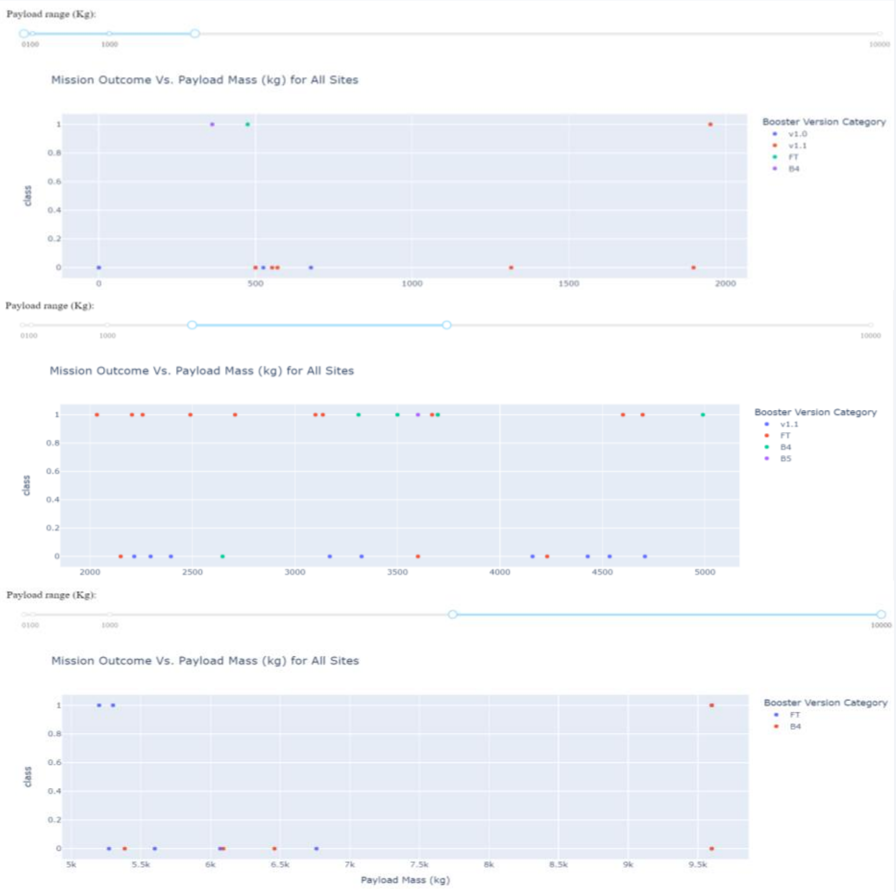
SpaceX Launch Records Dashboard

CCAFS SLC-40

Proportions of Successful and Failed Launches for Site CCAFS SLC-40

Scatterplot Illustrating Payload Vs. Outcome Overlayed by Booster Version

- For payloads in the range **0 – 2000 Kg**, booster **versions 1.0 and 1.1** were predominantly used, and most missions have **failed**, with the bulk of them being accumulated **below the 1K level.**
- A large portion of **successful** missions in the **2-5K** range, on the other hand, were carried out using **FT boosters**. However, the distribution of these accomplished landings seems to be skewed towards the lighter half of the range. Furthermore, **most failures** belong to **version 1.1** and are scattered throughout the range.
- Payloads **over 5000 Kg** were exclusively carried by **FT** and **B4** boosters with similarly **poor success rates.**



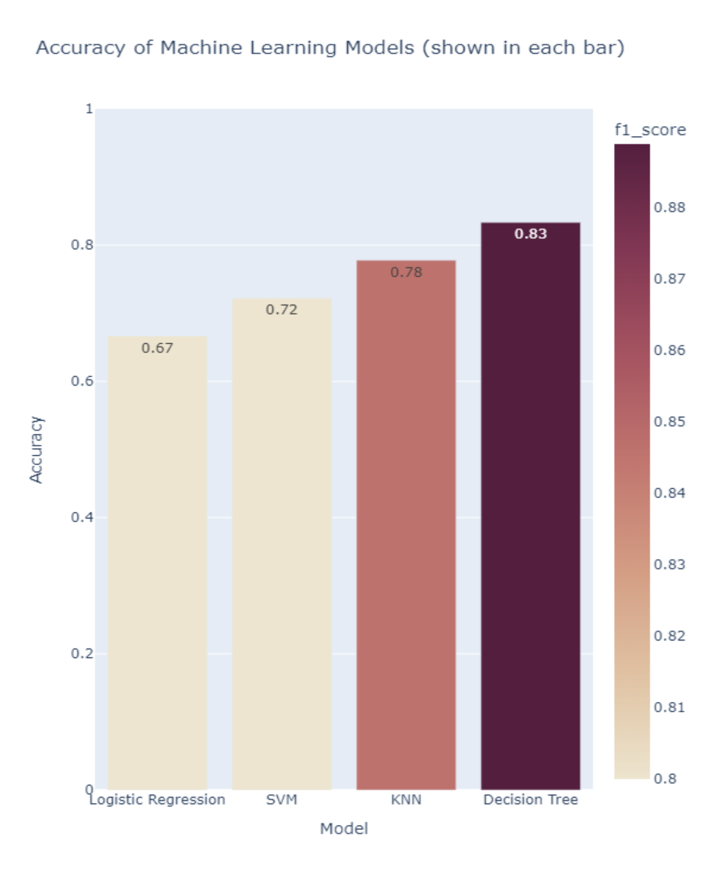
Section 5

Predictive Analysis (Classification)

Classification Accuracy

The bar chart visualizes the accuracy scores of all classification models built. Among these, the **Decision Tree** algorithm achieved the highest accuracy rate (%83). **K-Nearest Neighbor (KNN),** with an accuracy score of 0.78, came in second, followed by **SVM** and **logistic regression** models at 0.72 and 0.67, respectively.

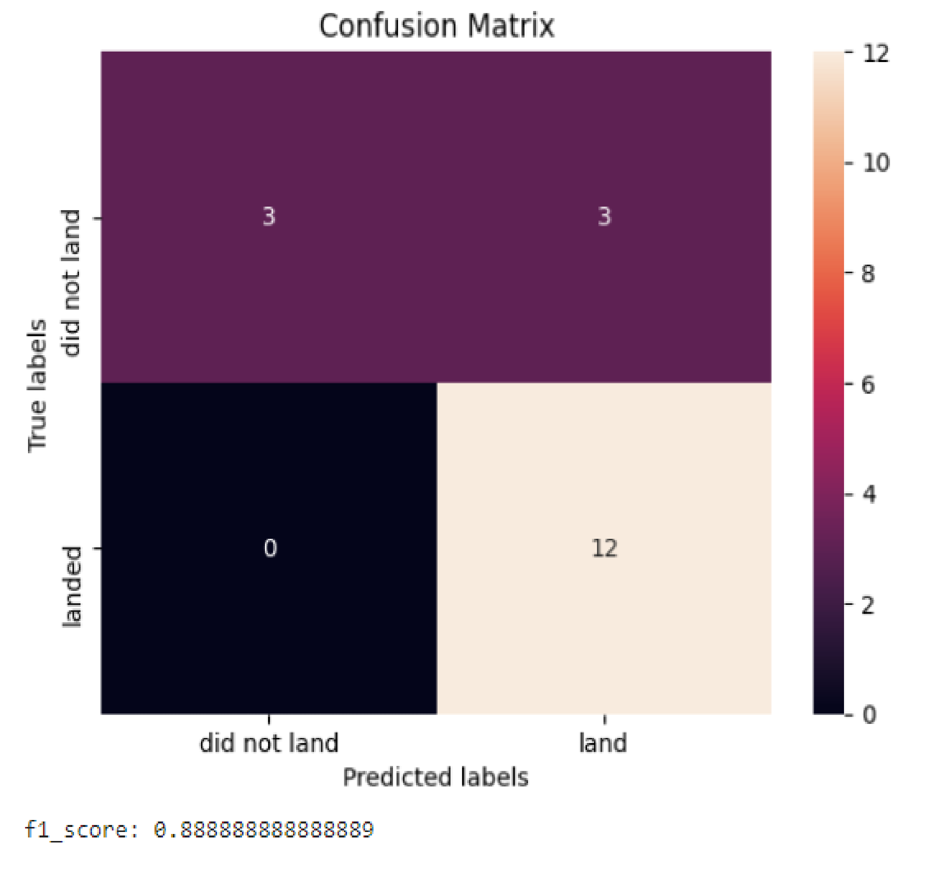
Similarly, the **confusion matrix F1 score**, incorporated into the chart as a continuous color marker, put the **Decision Tree** (F1 score: 0.88) in the first place compared to other models.



Confusion Matrix

As discussed, the **Decision Tree** model performed best regarding its confusion matrix evaluation results. Predicting 12 correct successful landing instances out of 18 total yielded an **F1 score** as high as ~0.89.

However, as this model **failed** to predict any actual values for the category of **failed missions**, further training and optimization are needed to meet the SpaceY criterion.



```
f1_score: 0.888888888888889
```

Conclusions

- **All launch sites** are located in the **southern** parts of the U.S., each in the furthermost points of the East and West. Also, **logistic** considerations, such as proximity to sea or railroad access, seem to have been a decisive factor in choosing site locations.
- Launch site **CCAFS SLC-40 in Florida** had the greatest number of flights while achieving the **highest success rate** among all sites. Interestingly, most of the cargo launched from this site was under 7000 Kg in mass.
- Overall, missions with **payload** masses in the range of **2-5 thousand Kg** had a **higher success rate**, with **version FT** being the **most successful** booster in this mass range.
- **Low earth orbits** could be a good starting point for **further in-depth analysis** as there are comparatively **more data available** on these categories with a **better success rate** than that of higher earth orbits.
- Our **Decision Tree** model performed best in classifying mission outcomes. With **classification** algorithms such as Decision Trees, higher scores may be achieved by factoring in **more features** such as meteorological data or cargo positioning and weight distribution and booster specifications. Also, SpaceY should look into **new strategies** to maximize success, for example, by sending the cargo in lighter pieces that would automatically assemble on site.
- As **many of the crashes are planned** and carried out by SpaceX for research purposes and by considering the insight obtained from the confusion matrix of the best-performing model indicating **poor performance in identifying failures** despite an F1 score of 0.88, it appears that model improvement could be possible by analyzing **planned and unplanned fails** independently.

Appendix

All the files relevant to this project can be accessed via a dedicated GitHub repository at: https://github.com/mohamashna/Winning-the-Space-Race-with-Data-Science

Acknowledgment

I would like to express my heartfelt gratitude to IBM's exceptional educational team and the innovative minds at Coursera for their commitment to providing high-quality learning opportunities worldwide. Their efforts have made it possible for many learners like me to gain transformative skills, opening doors to brighter academic and professional prospects.

I am particularly humbled by the financial aid I received, which allowed me to access the full version of this outstanding series of courses. This opportunity inspired me to put forth my utmost effort to make the most of the program and excel in the field of data science.

As I progress in this constantly evolving field, I feel inspired to give back to the scientific community that has enriched my life with the gift of knowledge. I strive to make meaningful contributions through data-driven research and a lifelong commitment to sharing knowledge.

Thank you!