IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pablo Gomes de Miranda
2023-07-11

Outline

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

Executive Summary

- Methodologies:
  - Data collected with Web Scraping and SpaceX API;
  - Exploratory Data Analysis and Data Wrangling with SQL and Python;
  - Interactive Visual Analytics;
  - Predictive Analysis with Machine Learning.

- Results:
  - Data collected from public sources;
  - Insights acquired from Data Analysis;
  - Prediction with Machine Learning.

Introduction

- What is this **about**?
  - **SpaceY**, a tech company focused on space exploration, spacecraft manufacturer, launcher and satellite communications **a technology company focused on space** exploration **wants to acquire information** that will **enable the improvement of its operations**.
- What questions we need to **answer**?
  - Successful landing of the first stage of the rockets;
  - Best place to launch its rockets.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data obtained and cleaned from **two sources**: 1 - SpaceX API, 2 – Web Scraping Falcon 9 launch wiki.
- Perform data wrangling:
  - Missing values from '**PayloadMass**' feature replaced by its mean values.
  - Data Analysis and **training label** determined.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:
  - The data was sampled into **training** and **test sets** and **different models** were trained and its respective metrics compared.

Data Collection

1) SpaceX API:

1) Data requested and parsed from SpaceX API from the following URL: https://api.spacexdata.com/v4/launches/past

2) Response decoded as a Json and turned into a Pandas data frame.

3) Data frame filtered to include Falcon 9 launches only.

2) Web Scraping:

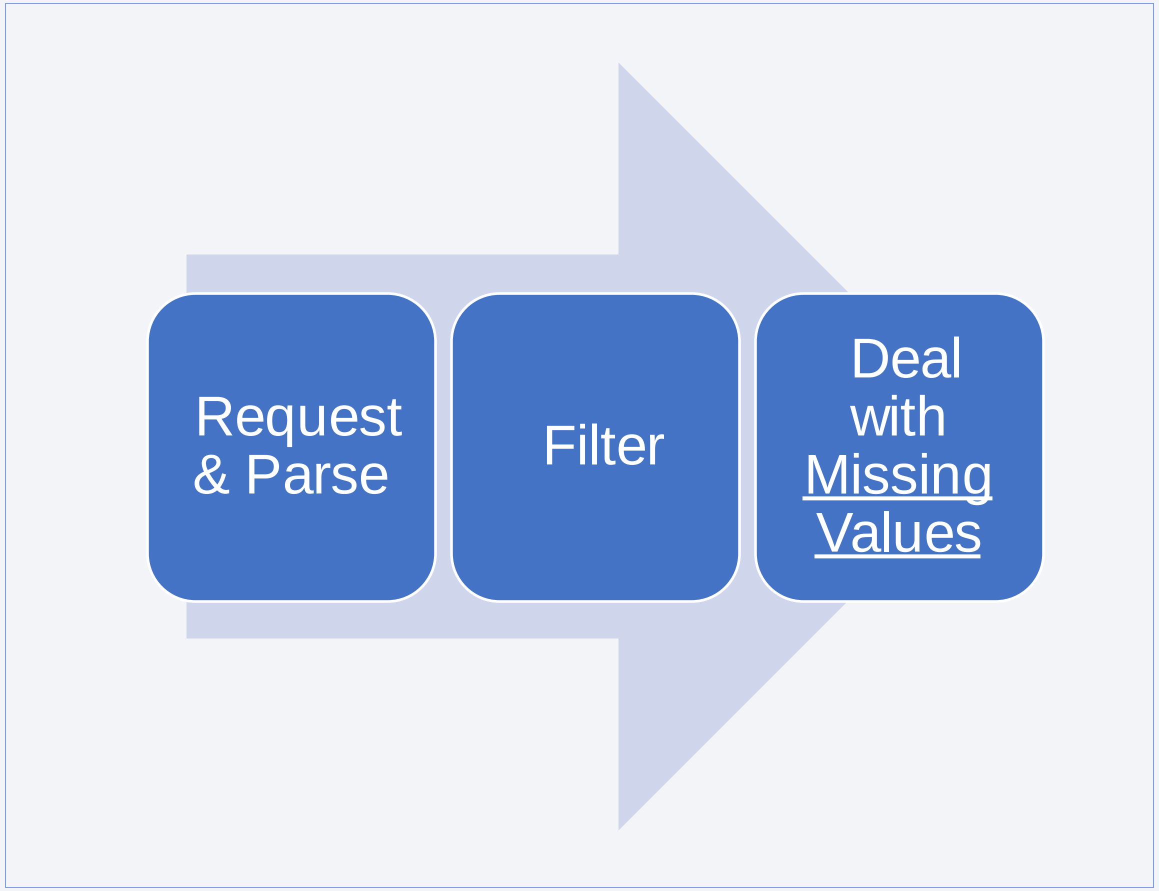
1) Data requested from Falcon 9 Launch Wiki page from the following URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

2) Column/variable names extracted from the HTML table header.

3) Data frame created by parsing the launch HTML tables.

Data Collection – SpaceX API

- Request & Parse
- Filter
- Deal with missing values



Notebook for reference:

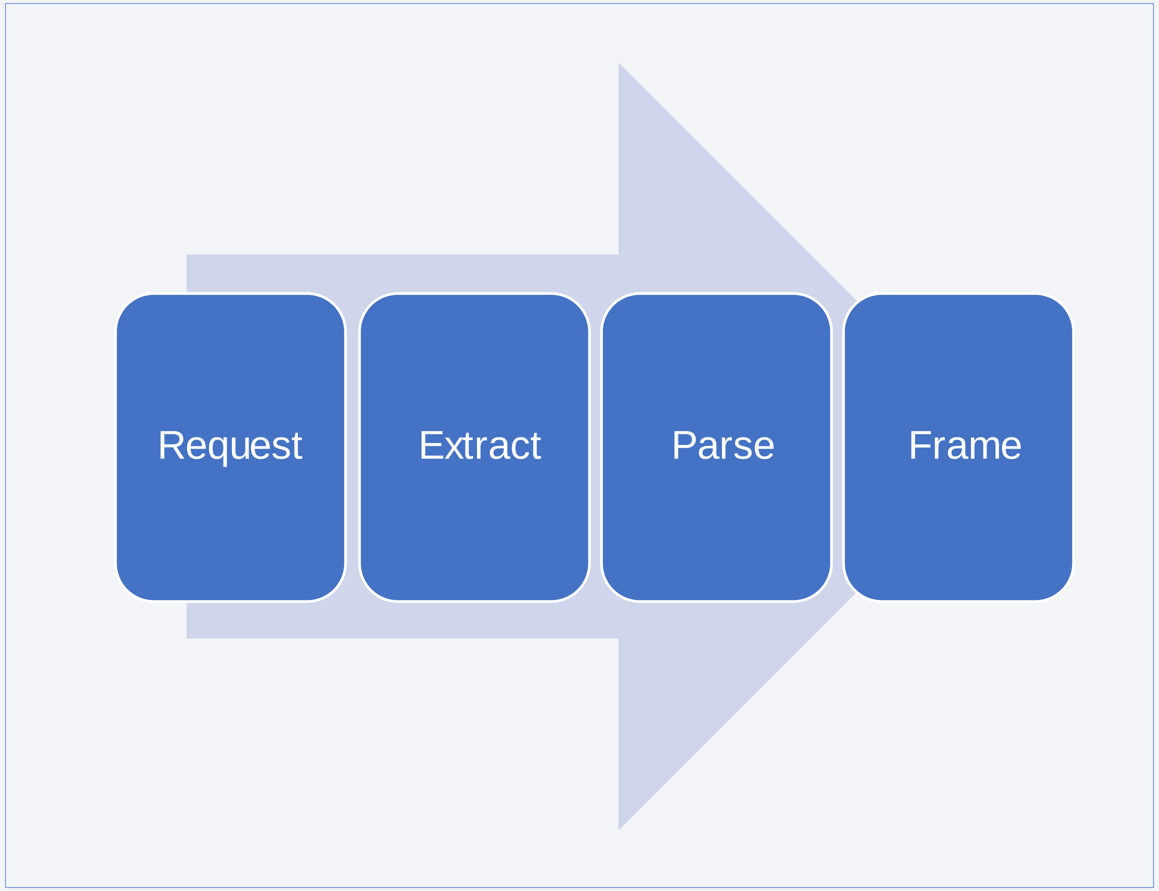
https://github.com/pgdemiranda/ibm-capstone_DS/blob/main/1%20-%20jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Request
- Extract
- Parse
- Frame
  - Data frame created by parsing the HTML tables.

Notebook for reference:

https://github.com/pgdemiranda/ibm-capstone_DS/blob/main/2%20-%20jupyter-labs-webscraping.ipynb

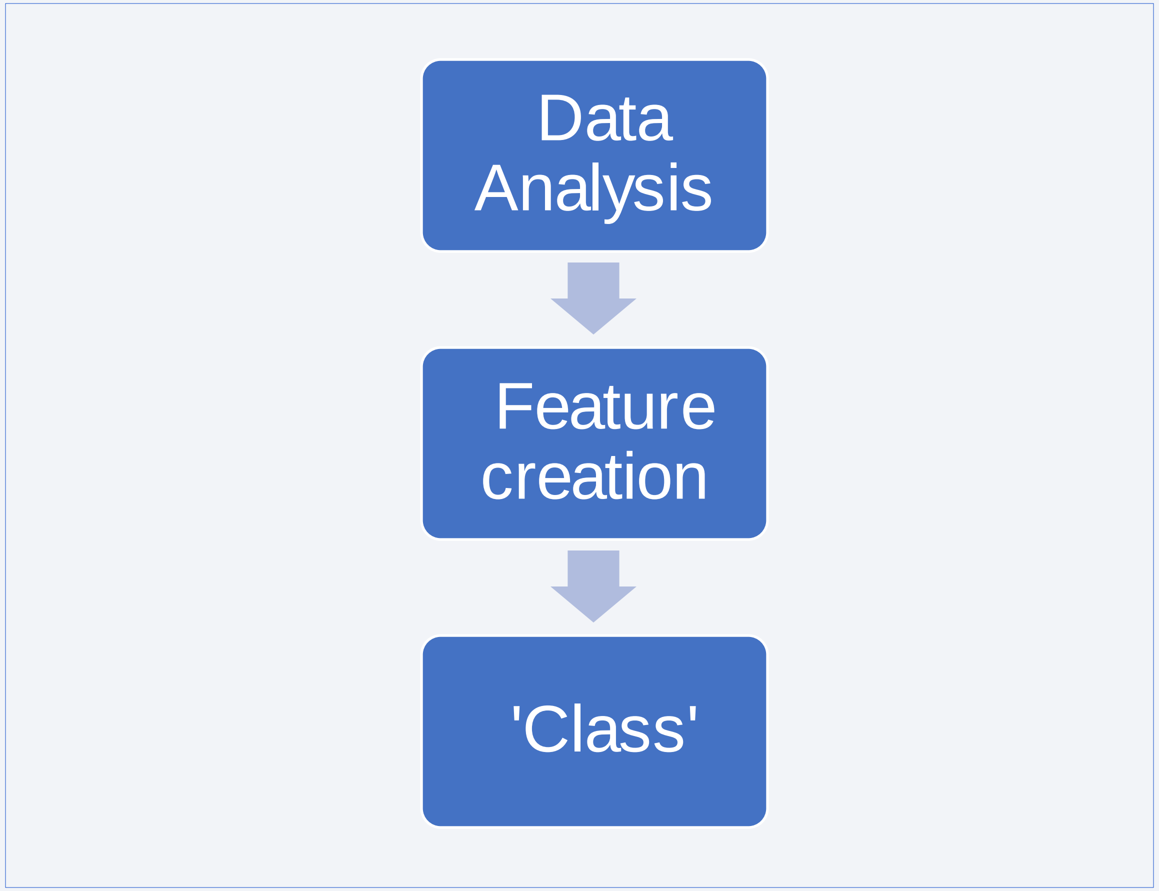


Data Wrangling

- Data Analysis: calculate the number of:
  - Launches on each site;
  - Occurrence of each orbit;
  - Mission outcome of the orbits.
- Determine Training Labels (Feature Creation)
  - Create a new landing outcome feature from Outcome column.
  - Selecting 'Class' label for future prediction.

Notebook for reference:

https://github.com/pgdemiranda/ibm-capstone_DS/blob/main/3%20-%20labs-jupyter-spacex-Data%20wrangling.ipynb



Data Analysis

Feature creation

'Class'

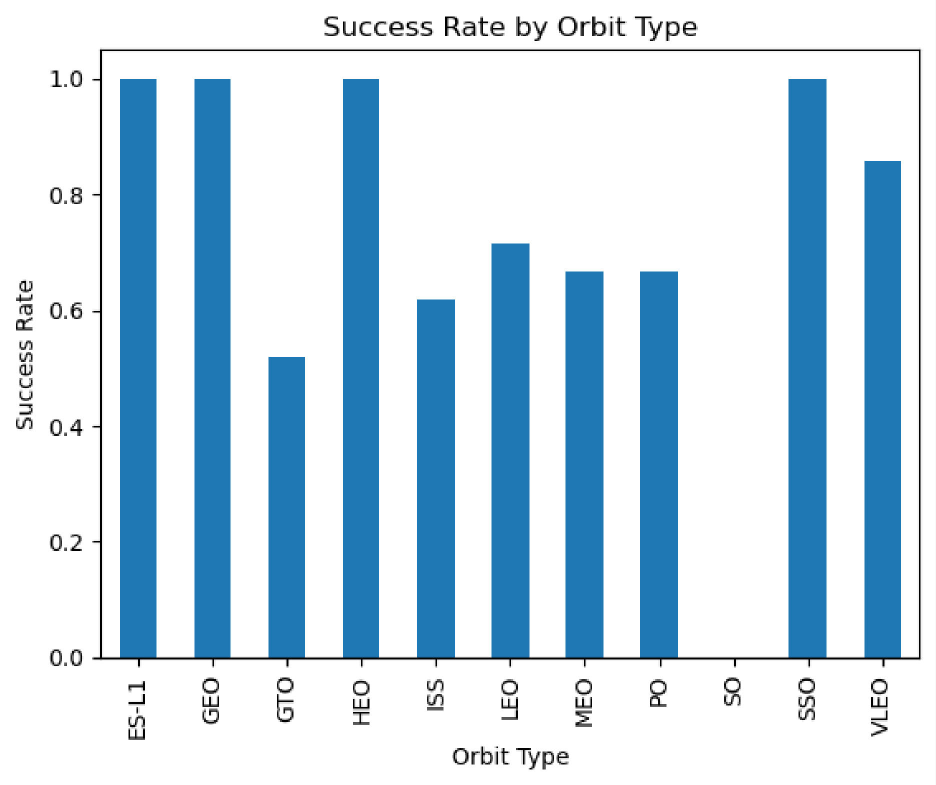
EDA with Data Visualization

Graphs plotted to gather insights from our Data:

- **Scatterplots** to study the relationship between Flight Number and Pay Load Mass, Launch Site and Orbit Type.
- Additional **Scatterplot** to examine the relationship between **Payload and Launch Site**;
- A **Bar Chart** to display the elements of **Orbit Type** category;
- A **Line Plot** to study the **Launch Success** Yearly Trend.

Notebook for reference:

https://github.com/pgdemiranda/ibm-capstone_DS/blob/main/5%20-%20jupyter-labs-eda-dataviz.ipynb



Example of the Bar Chart

EDA with SQL

In order to **acquire more insight**, a series of **SQL queries** were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name 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;
- Date when the first successful landing outcome in ground pad was achieved.
- Names of the boosters which have success in drone ship and have a payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions and launch site names from the months of 2015;
- Rank of the count of landing outcomes between 2010-06-04 and 2017-03-20.

Notebook for reference:

https://github.com/pgdemiranda/ibm-capstone_DS/blob/main/4%20-%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

This is a list of objects, and its **purpose**, displayed on a Map with Folium:

- **Markers** to indicate the launch sites;
- **Circles** to highlight areas around coordinates of interest, like NASA Johnson Space Center;
- **Lines** to display the distance between these coordinates and shorelines, highways, railways and cities;
- **Marker Clusters** to group a series of elements in each coordinate, making it easier to observe the existence of these elements before zooming the map.

Notebook for reference:

https://github.com/pgdemiranda/ibm-capstone_DS/blob/main/6%20-%20lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Plots, graphs and interactions on dashboard:
  - **Pie Chart** and **Dropdown**: Percentage of Launches by Site;
  - **Scatterplot**, **Dropdown** and **Slider**: Payload Range;

- The goal is to understand where is the **best site to launch the rockets according to the payload**.

Dashboard for reference:

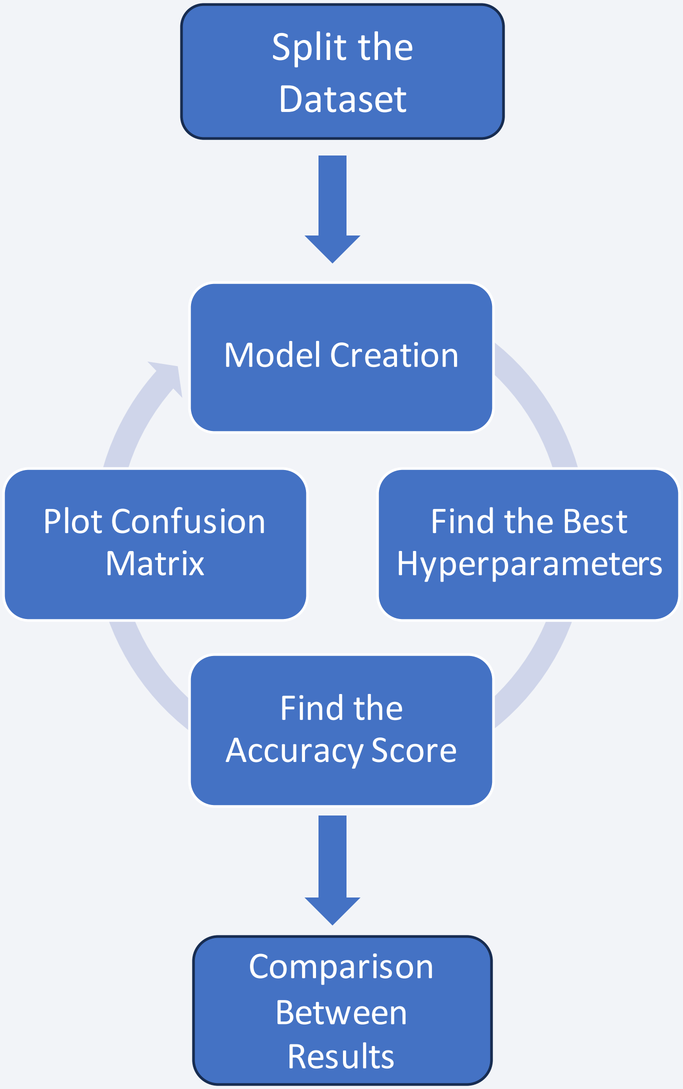
https://github.com/pgdemiranda/ibm-capstone_DS/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Four Machine Learning Classification models were trained, and its results compared:

- Logistic Regression;
- Support Vector Machine;
- Decision Tree;
- K-Nearest Neighbors.

The dataset was splitted **once** between sets of Train and Test and the results were also compared at the **end** to select the **best classification model.**



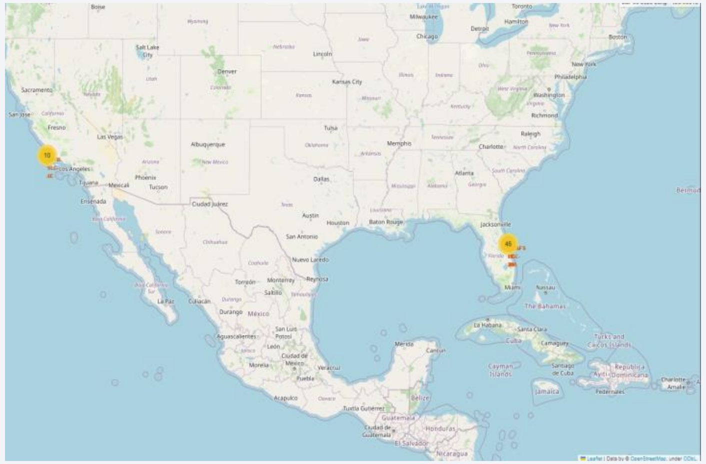
Notebook for reference:

https://github.com/pgdemiranda/ibm-capstone_DS/blob/main/7%20-%20SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

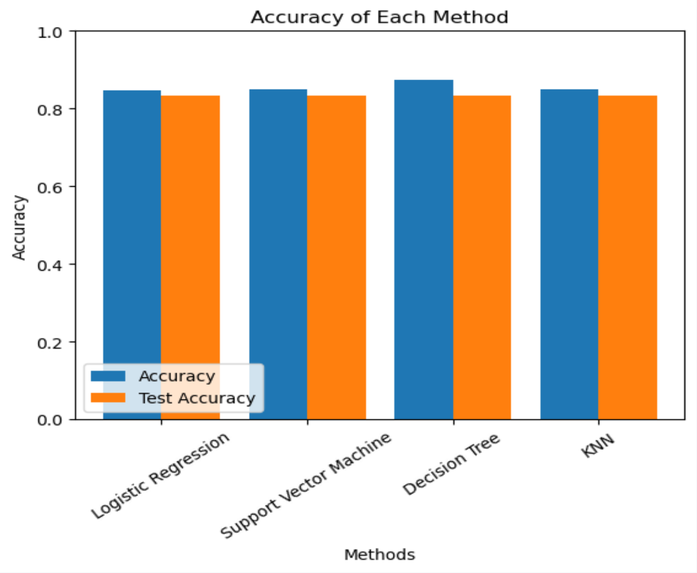
Exploratory Data Analysis:

- SpaceX employed 4 different launch sites since its first launch done by SpaceX and NASA; the first successful landing outcome happened in 2015, the same year when 2 boosters failed at landing in drone ships;
- Most of the Falcon9 boosters were successful at landing in drone ships; the average payload of F9 v1.1 booster is 2,928Kg;
- As the years passed, the successful landing outcomes rates also got better.



Interactive Analytics:

- East Coast concentrate most of SpaceX launches. All launch sites are close to the shorelines, connected to highways, while also maintaining a good distance from the cities nearby.



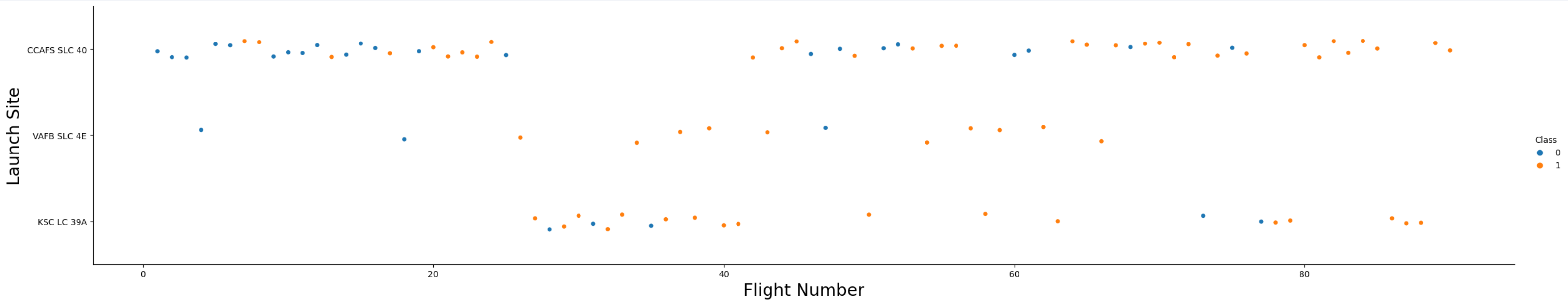
Predictive Analysis:

- Decision Tree Classifier returned the best results of Accuracy and Test Accuracy without committing overfit.

Section 2

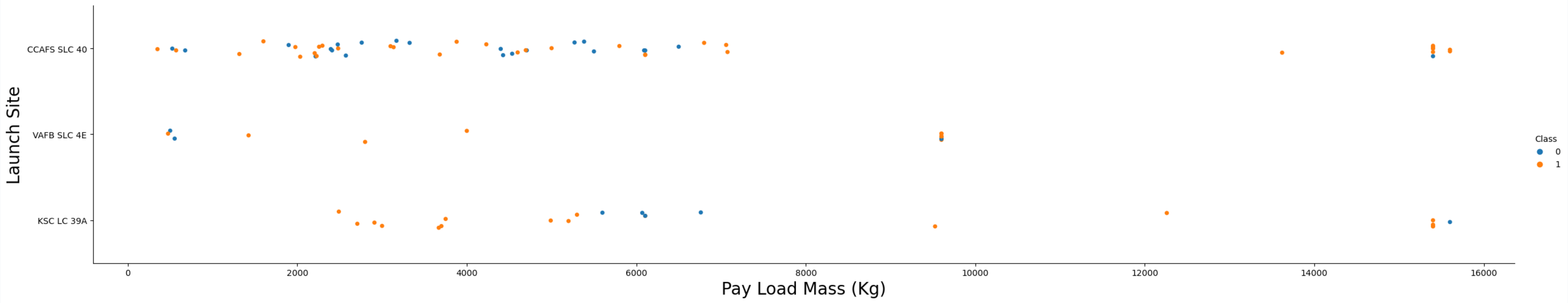
Insights drawn from EDA

Flight Number vs. Launch Site



- Most of the launches were conducted at CCAFS SLC 40, followed by a period when most of the launches were transferred to KSC LC 39A.
- Failure rate at CCAFS SLC 40 was high at the beginning, which could justify the operation transfer to other launch sites.
- Few launches took place at VAFB SLC 4E, but they did with a high success rate.

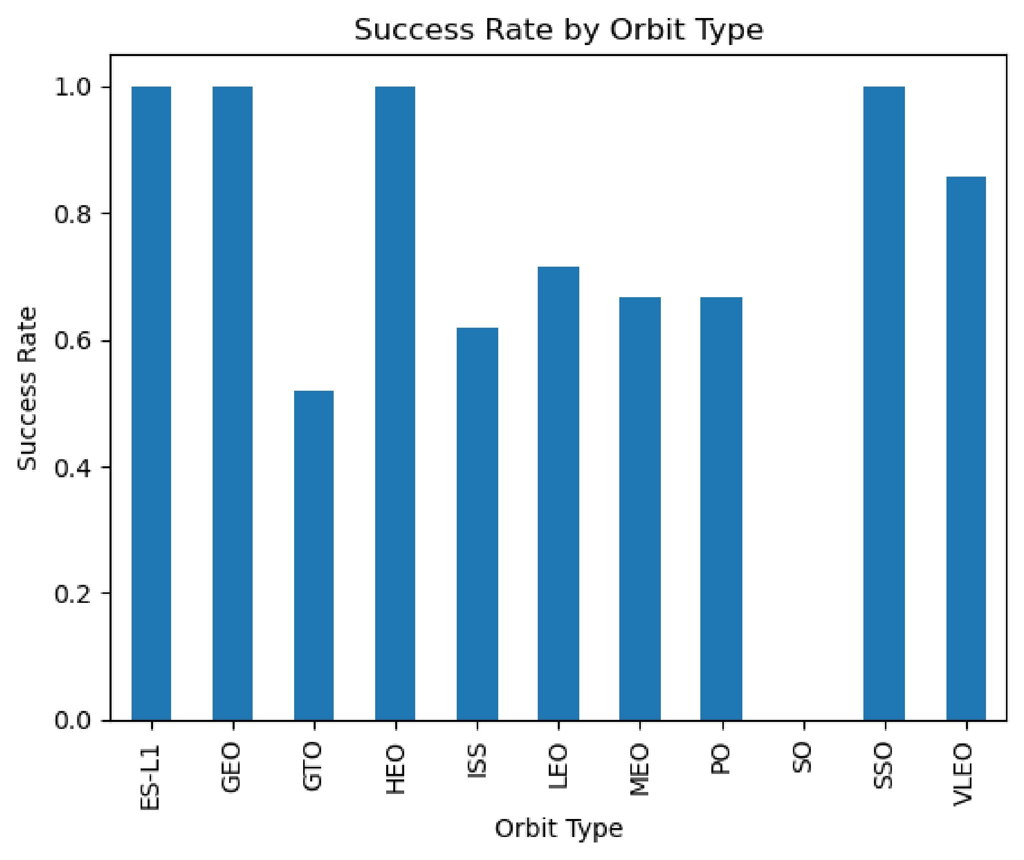
Payload vs. Launch Site



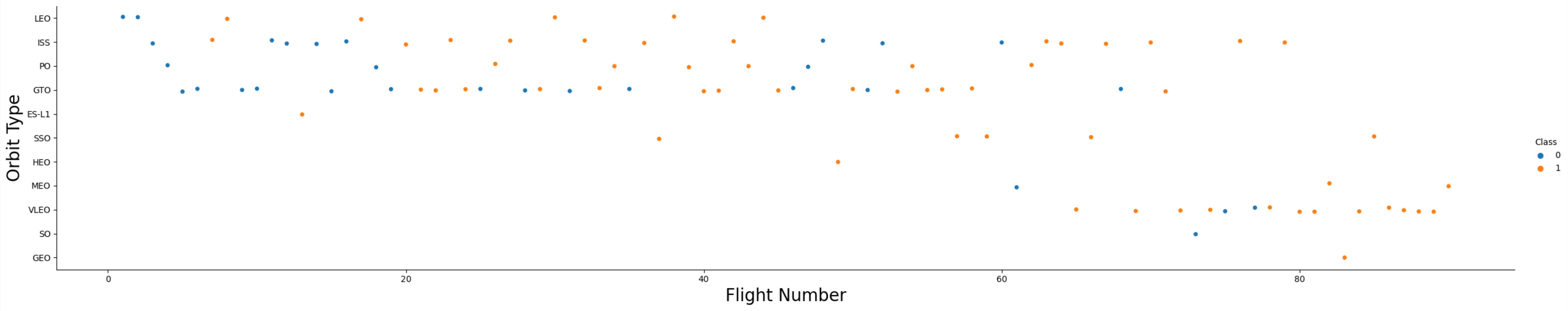
- At VAFB SLC 4E, there wasn't a rocket launched with a payload mass higher than 10000 Kg.
- CCAFS SLC 40 and KSC LC 39A both launched rockets with the higher payload, but only CCAFS SLC 40 was successful while not having the best launch score.

Success Rate vs. Orbit Type

- Best Orbit Type (close to 100%):
  - ES-L1;
  - GEO;
  - HEO;
  - SSO.
- Promising Orbit Type:
  - VLEO (> 80%);
  - LEO (> 70%).

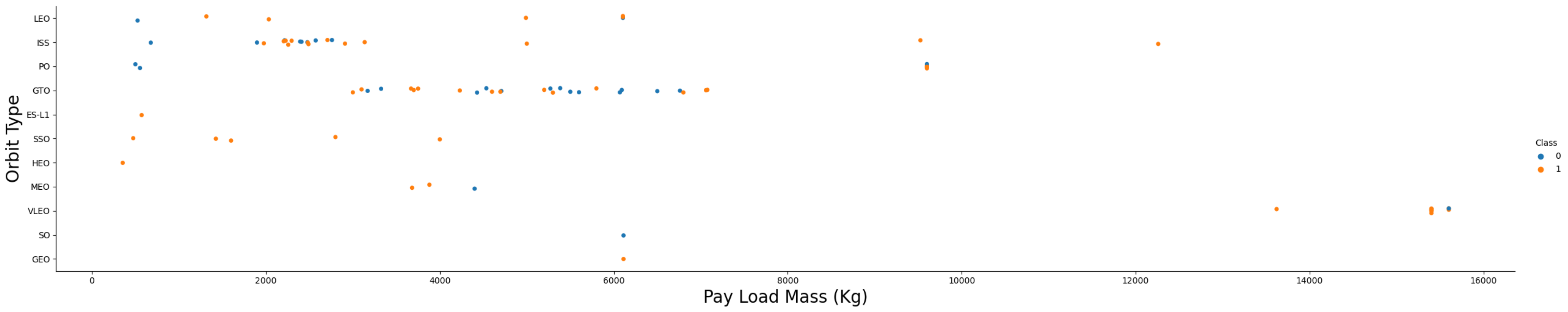


Flight Number vs. Orbit Type



- Experience brings success: as more mission happens, all orbits show improvement.
- Missions on VLEO only appears later but are very promising.

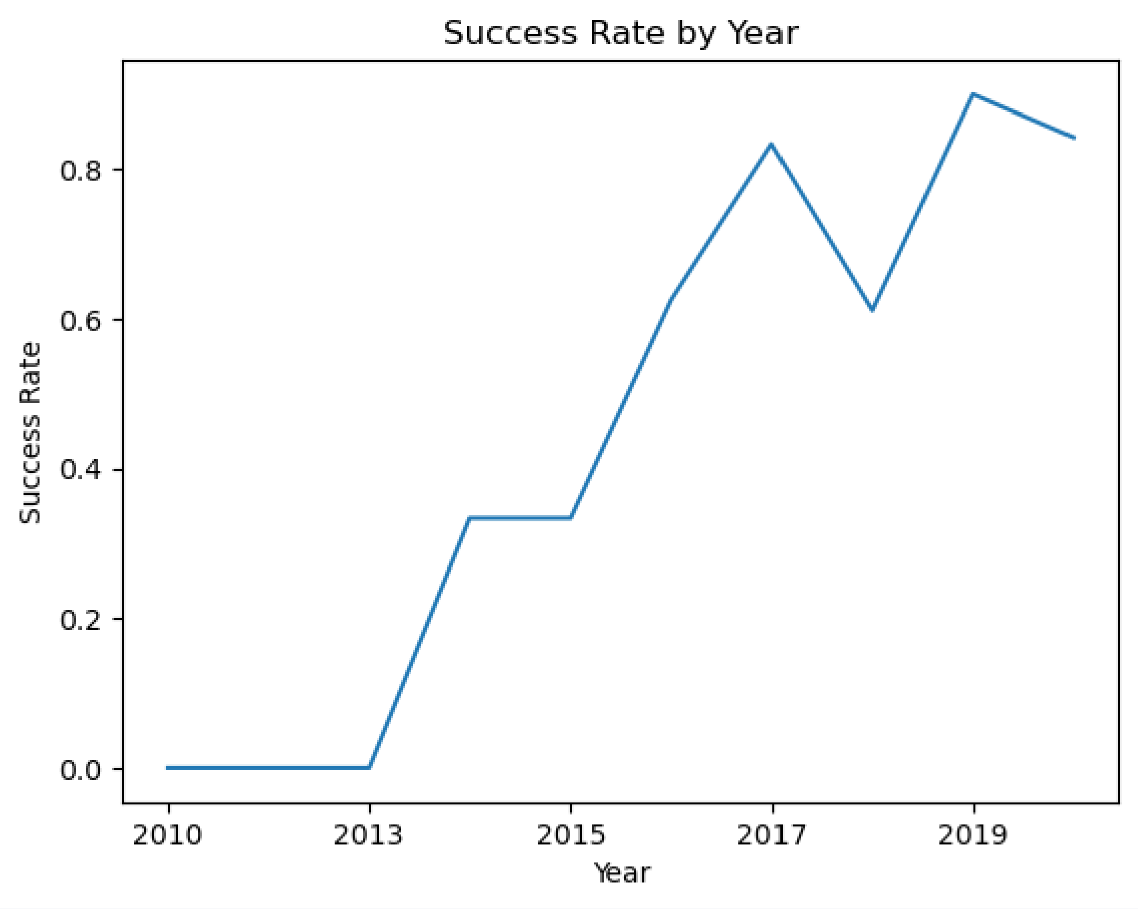
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO it isn't possible to distinguish this well, as both positive and negative landing are there and overlapping each other.

Launch Success Yearly Trend

The success rate increased since 2013 but decreased between 2017 and 2019.

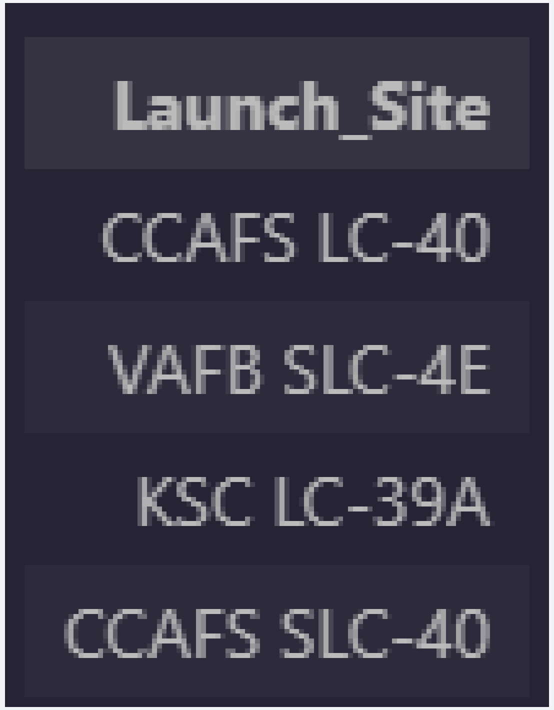


All Launch Site Names

- SQL Query where distinct launch sites were select from the database:

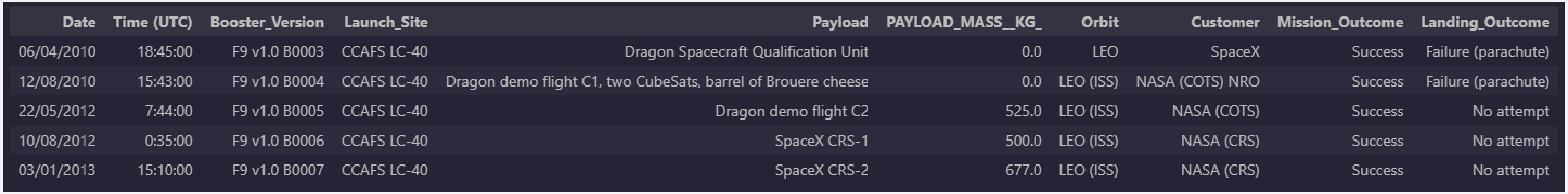
```
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

- There are four launch sites:
    - CCAFS LC-40
    - VAFB SLC-4E
    - KSC LC-39A
    - CCAFS SLC-40



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

Launch Site Names Begin with 'CCA'



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

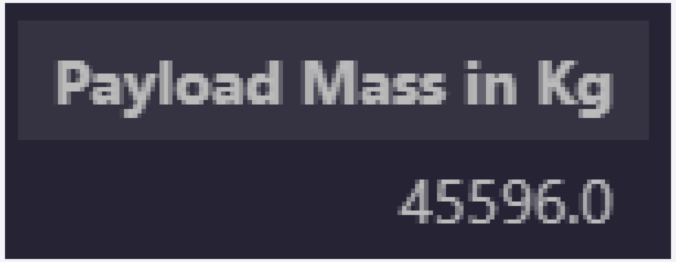
SQL Query where all data from from the database was selected where the name of the launch site begin with 'CCA':

```
SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Total Payload Mass

SQL Query selecting the sum of payload mass from the database but for the missions from NASA (CRS):

```
SELECT SUM(PAYLOAD_MASS__KG_) AS 'Payload Mass in Kg' FROM
SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

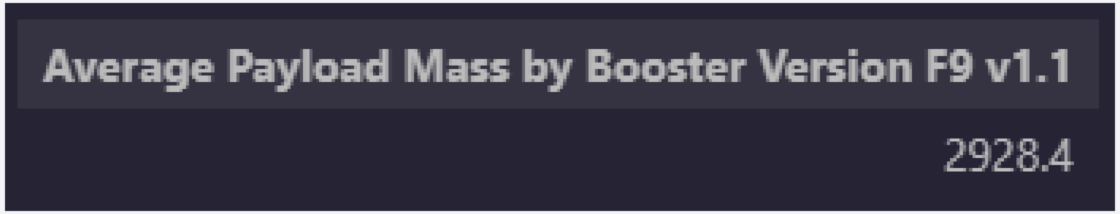


Payload Mass in Kg
45596.0

Average Payload Mass by F9 v1.1

SQL Query selecting the average payload mass in Kg for F9 v1.1 from the database:

```
SELECT AVG(PAYLOAD_MASS__KG_) AS 'Average Payload Mass by Booster
   Version F9 v1.1' FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9
   v1.1'
```

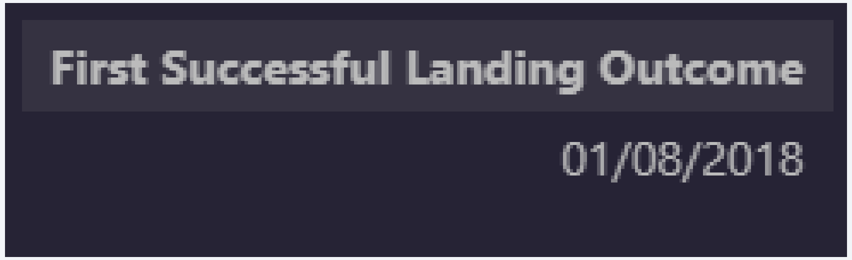


Average Payload Mass by Booster Version F9 v1.1
2928.4

First Successful Ground Landing Date

SQL Query selecting the date (dd-mm-YY) of the first successful landing outcome from the database:

```
SELECT MIN(DATE) as 'First Successful Landing Outcome' FROM
   SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

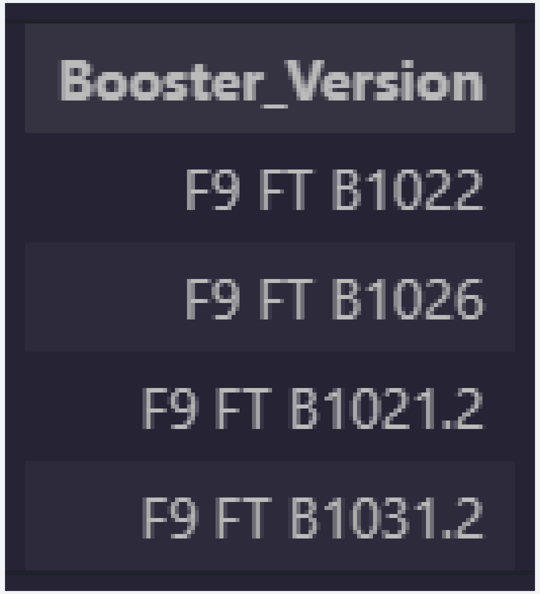


First Successful Landing Outcome
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 from the database:

```
SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_
   BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Success (drone
   ship)';
```

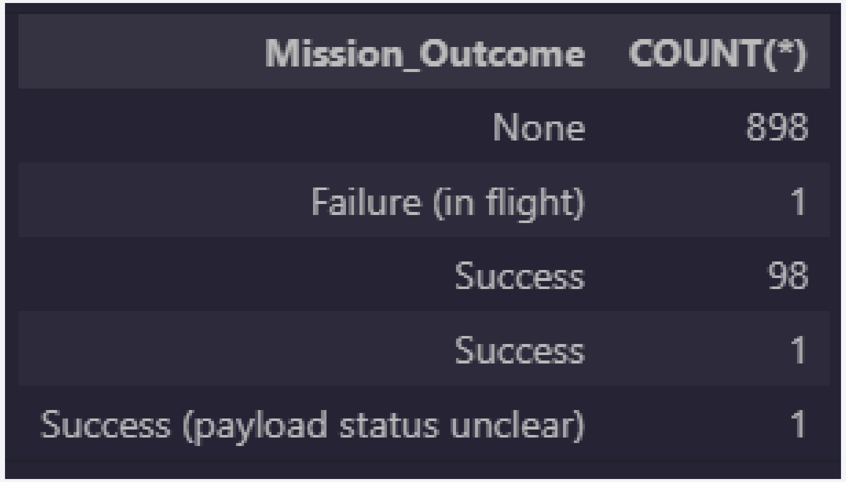


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

SQL Query calculating the total number of successful and failure mission outcomes from our database ordering by mission outcome:

```
SELECT MISSION_OUTCOME, COUNT(*) FROM SPACEXTBL GROUP BY
   MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

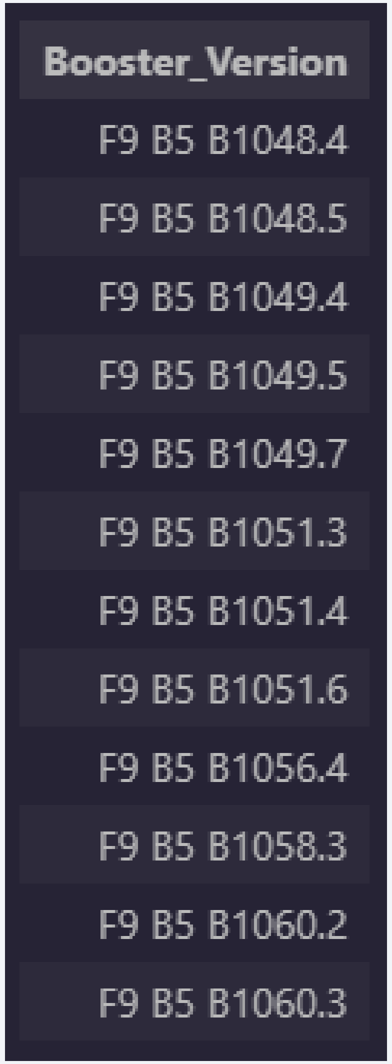


Mission_Outcome	COUNT(*)
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

SQL Query listing the names of the booster which have carried the maximum payload mass from our databased and ordered by boosters:

```
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE
   PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM
   SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

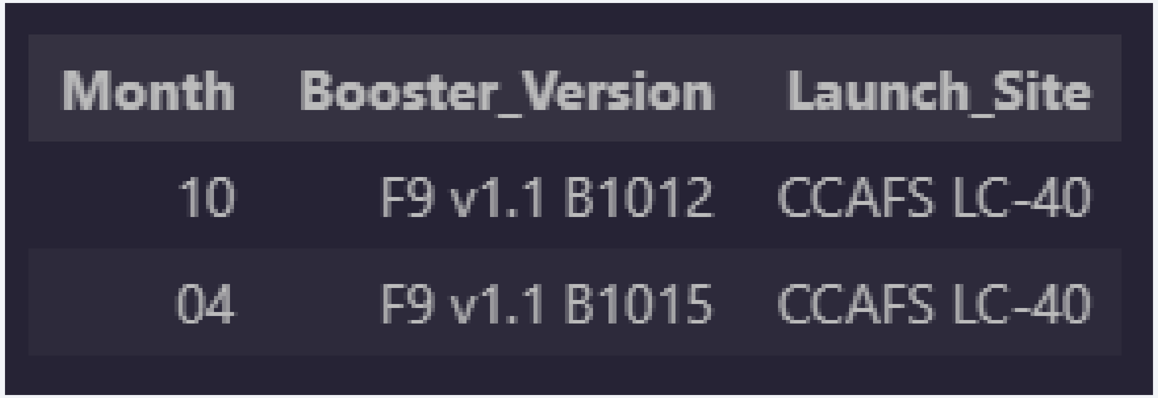


Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

SQL Query listing the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 (and the months) from our database:

```
SELECT substr(DATE, 4, 2) AS 'Month', BOOSTER_VERSION,
LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure
(drone ship)' AND substr(DATE, 7, 4) = '2015';
```

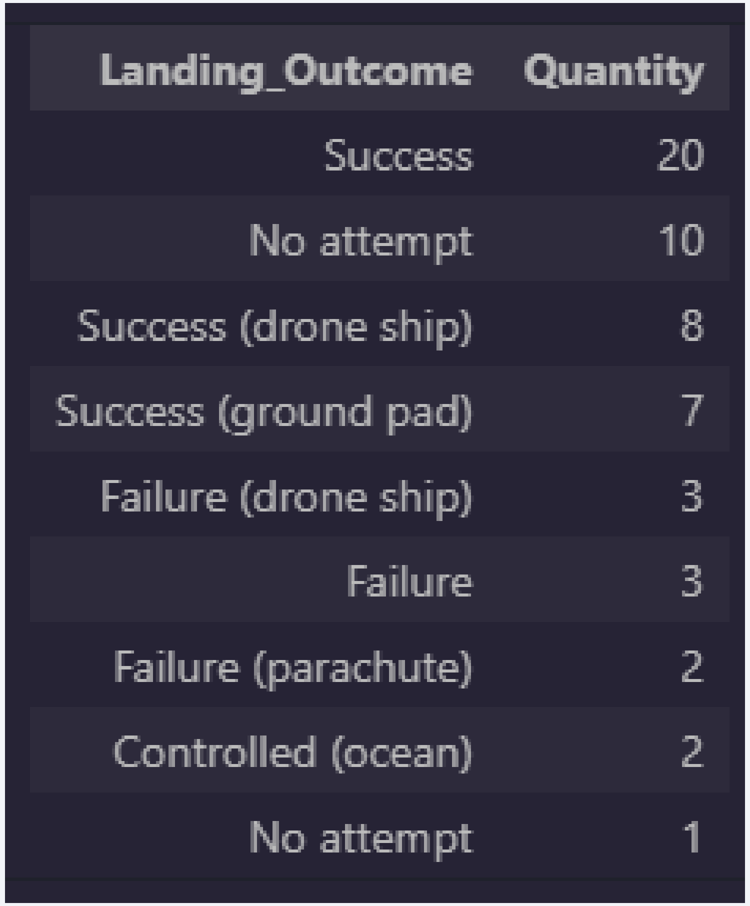


Month	Booster_Version	Launch_Site
10	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

SQL Query ranking 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, from our database:

```
SELECT LANDING_OUTCOME, COUNT(*) AS Quantity FROM SPACEXTBL WHERE
   DATE BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY
   LANDING_OUTCOME ORDER BY Quantity DESC;
```

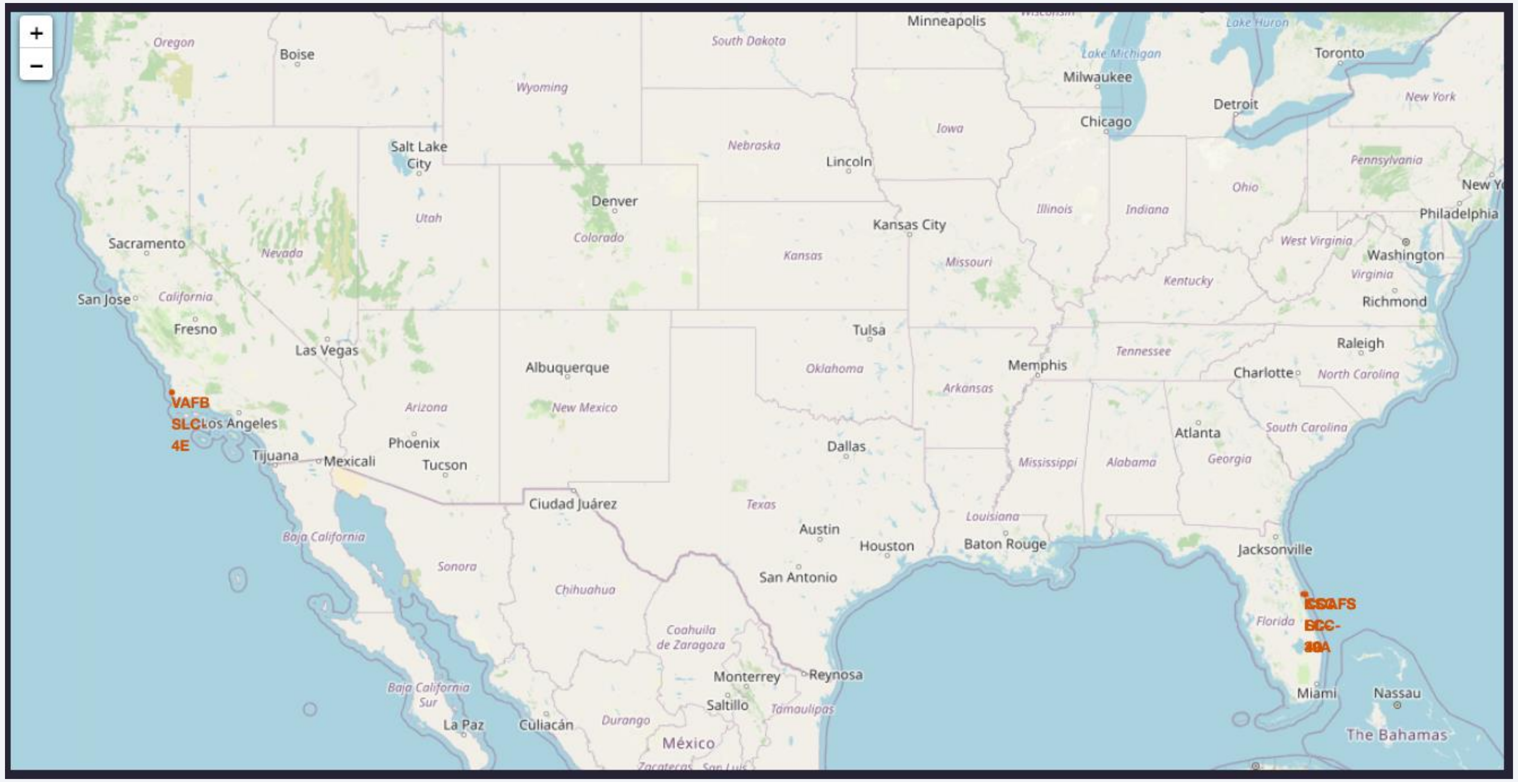


Landing_Outcome	Quantity
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

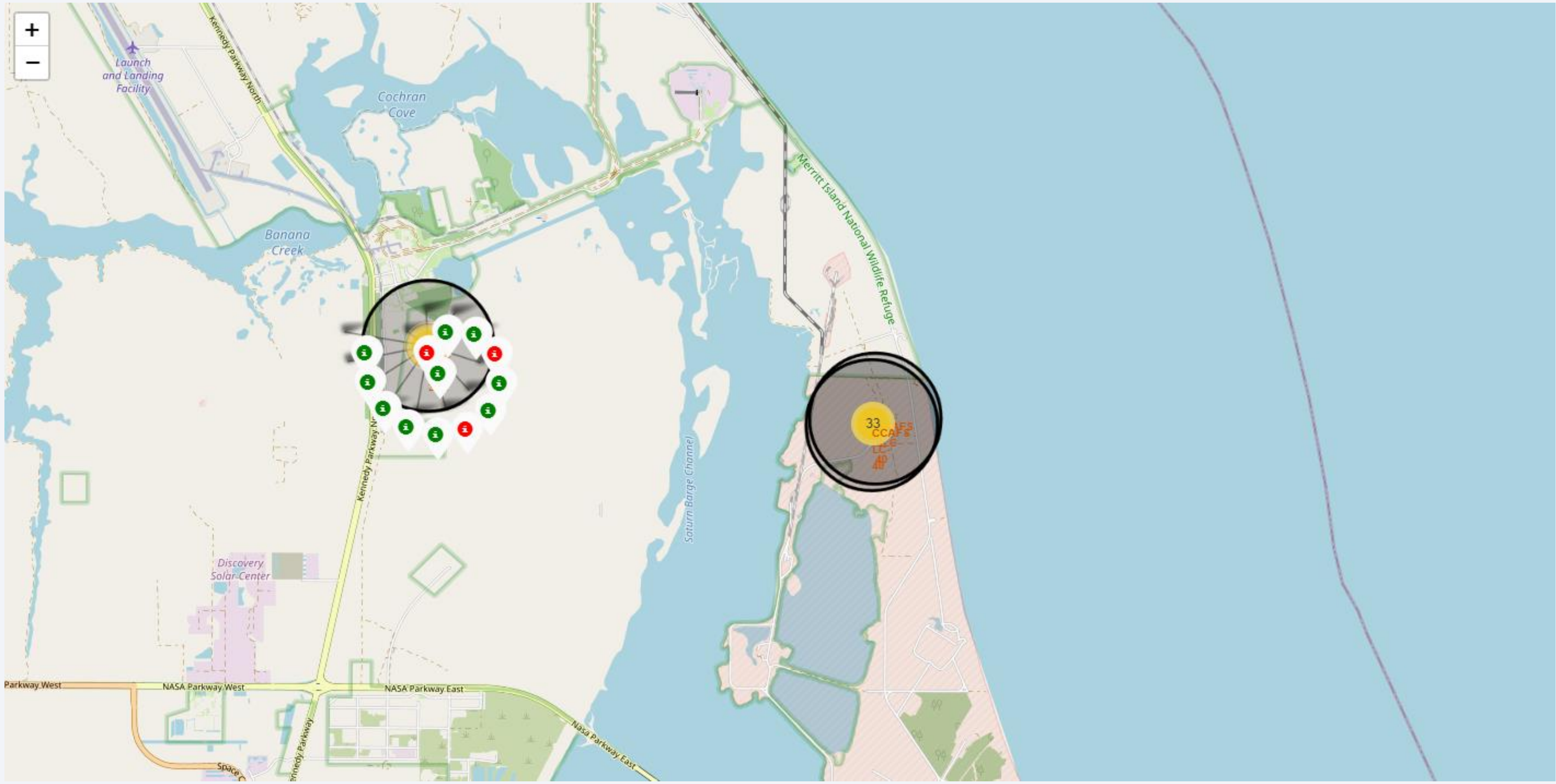
Launch Sites Proximities Analysis

All Launch Sites



- Launch sites can be found in both West and East coasts, but not close to the equator lines, because the US is a country positioned on the global north.
- These sites are close to the coastlines, highways, railways, but not that close to big cities.

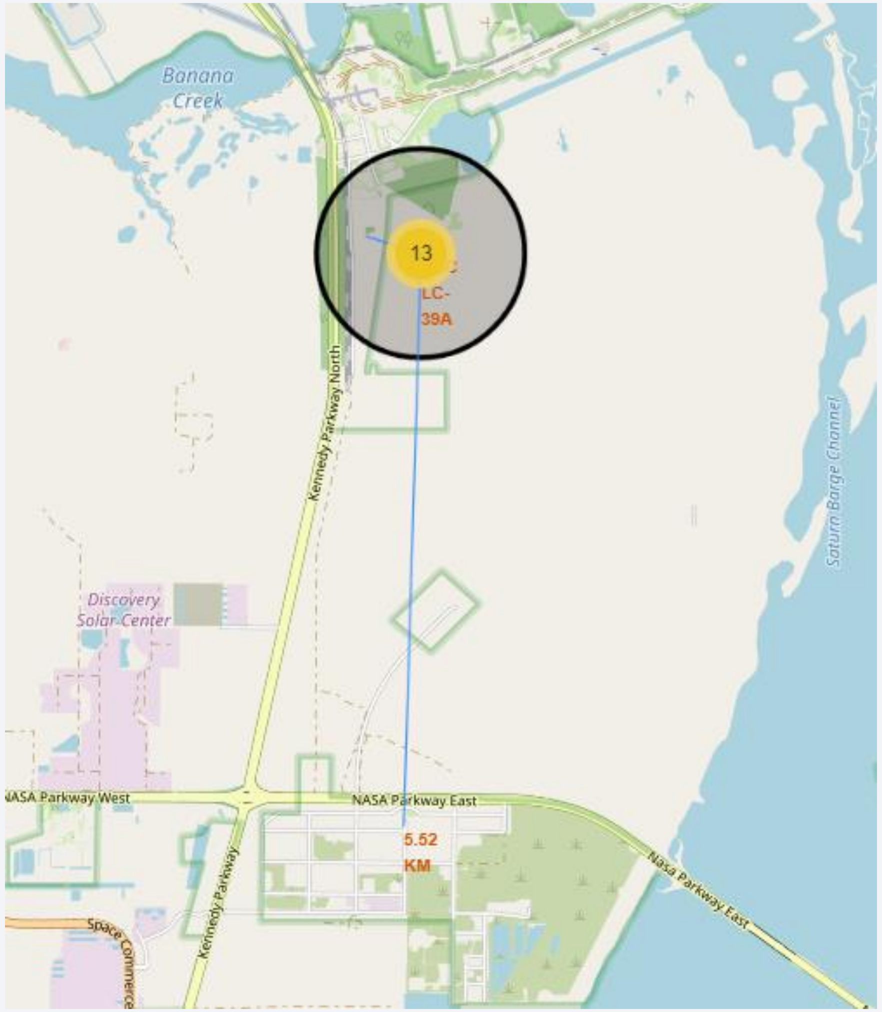
Launch Outcomes on the Map

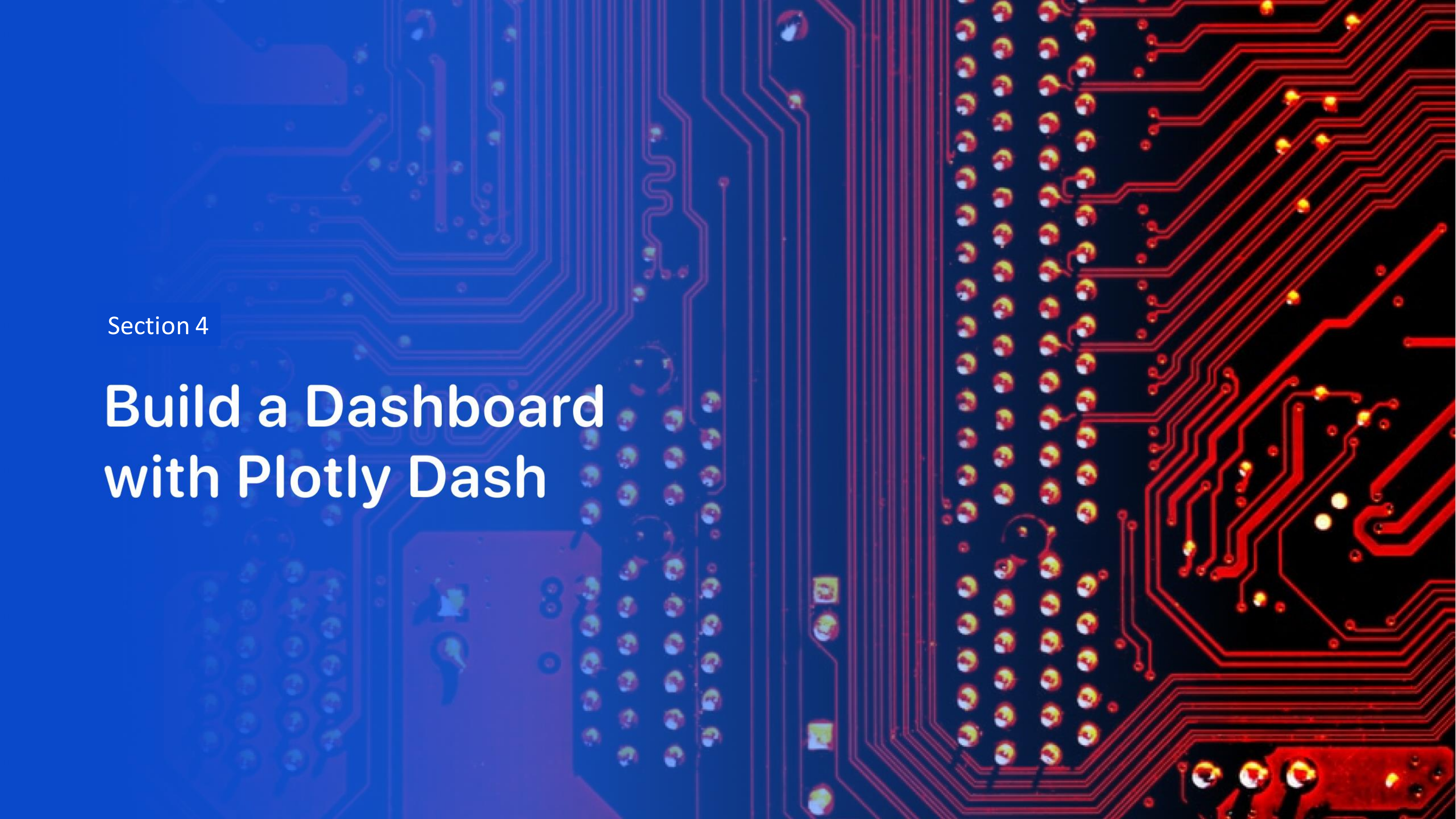


- Green Markers are for successful outcomes, while Red Markers are for failure outcomes.

Logistic and Safety

- Example from KSC LC-39A
- To illustrate how good and safe was the launch site, it is possible to examine that the next highway is 5.52Km.
- NASA Parkway cut the Highway FL 405.

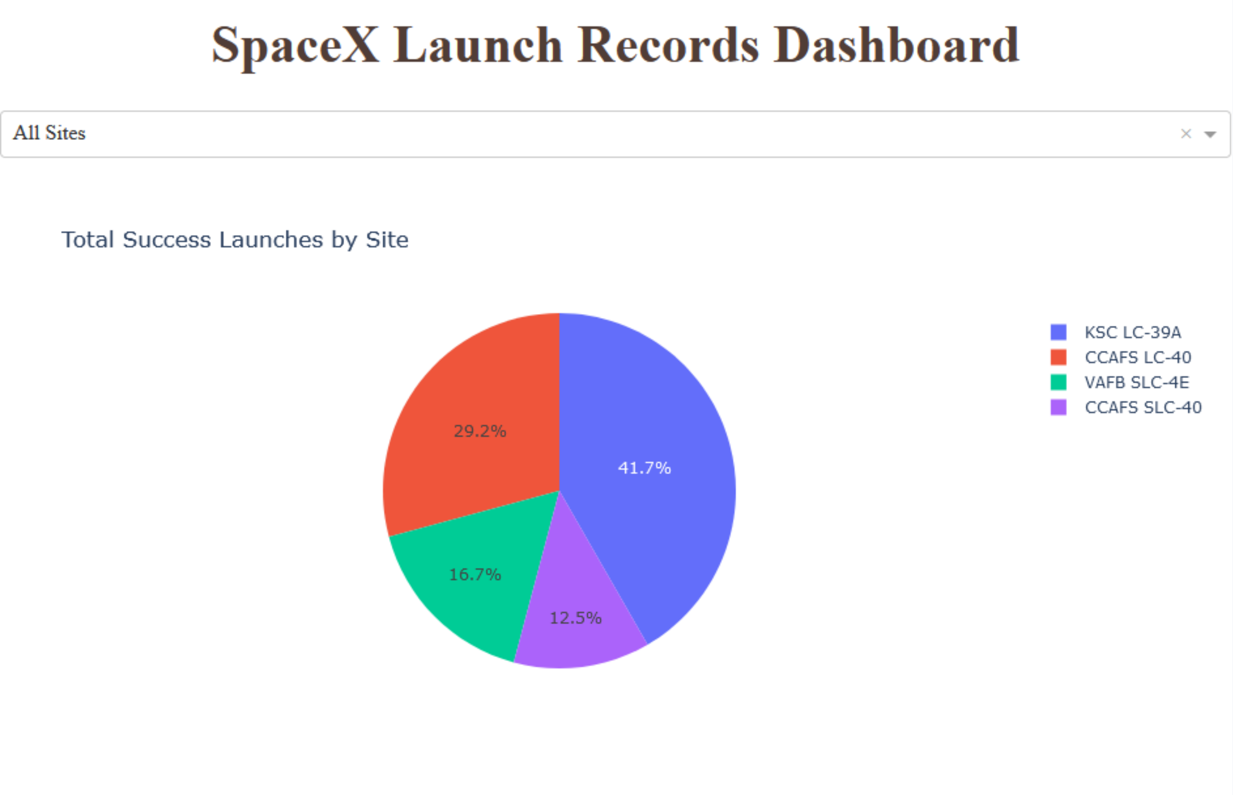




Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site

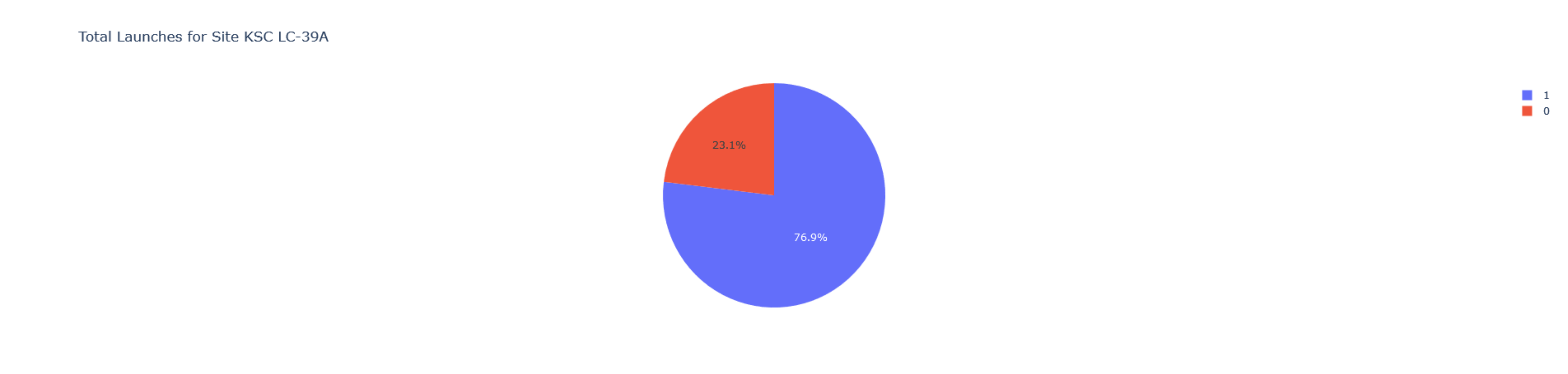


SpaceX Launch Records Dashboard

All Sites

- Some launch sites have a better success rate than others.
- Better opportunities come from KSC LC-39A.

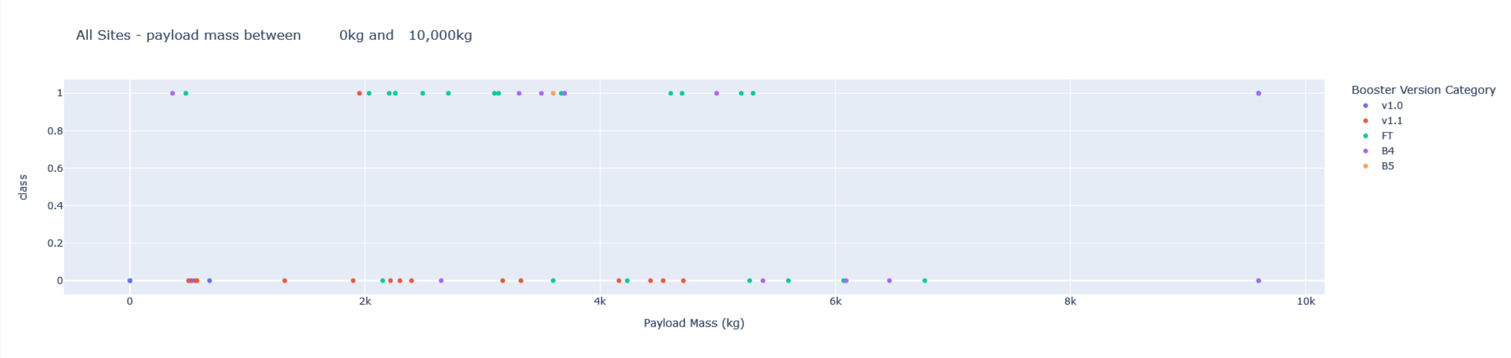
Launch Success Ratio for KSC LC-39A



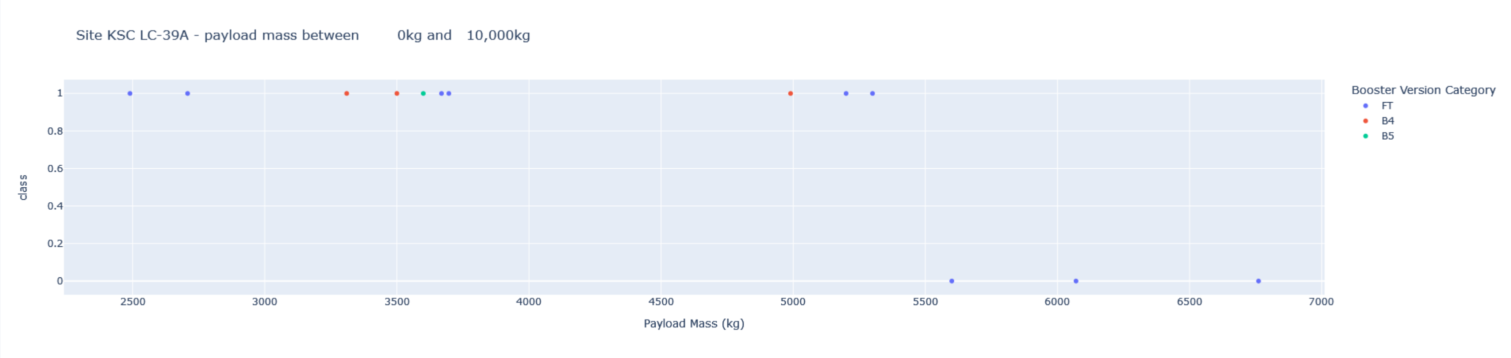
- 1 = Success
- 0 = Failure
- At KSC LC-39A site, 76.9% of all launches are successful!

Payload vs Launch Outcome

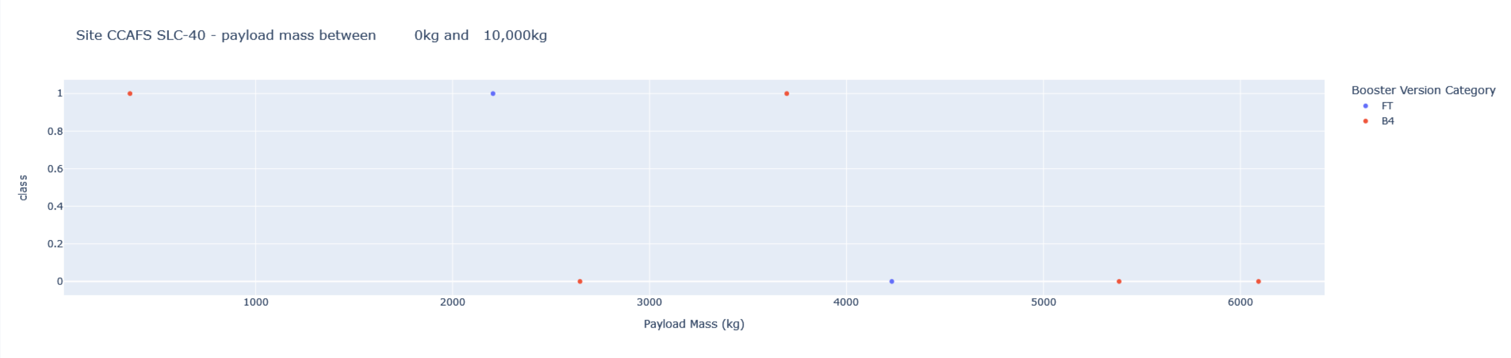
- All launches outcomes for max payload range;



- On KSC LC-39A payloads from more than 5000Kg are all FT boosters;



- The launch site with less success, CCAFS SLC-40, had so few launches, that we can't infer much from its visualization.



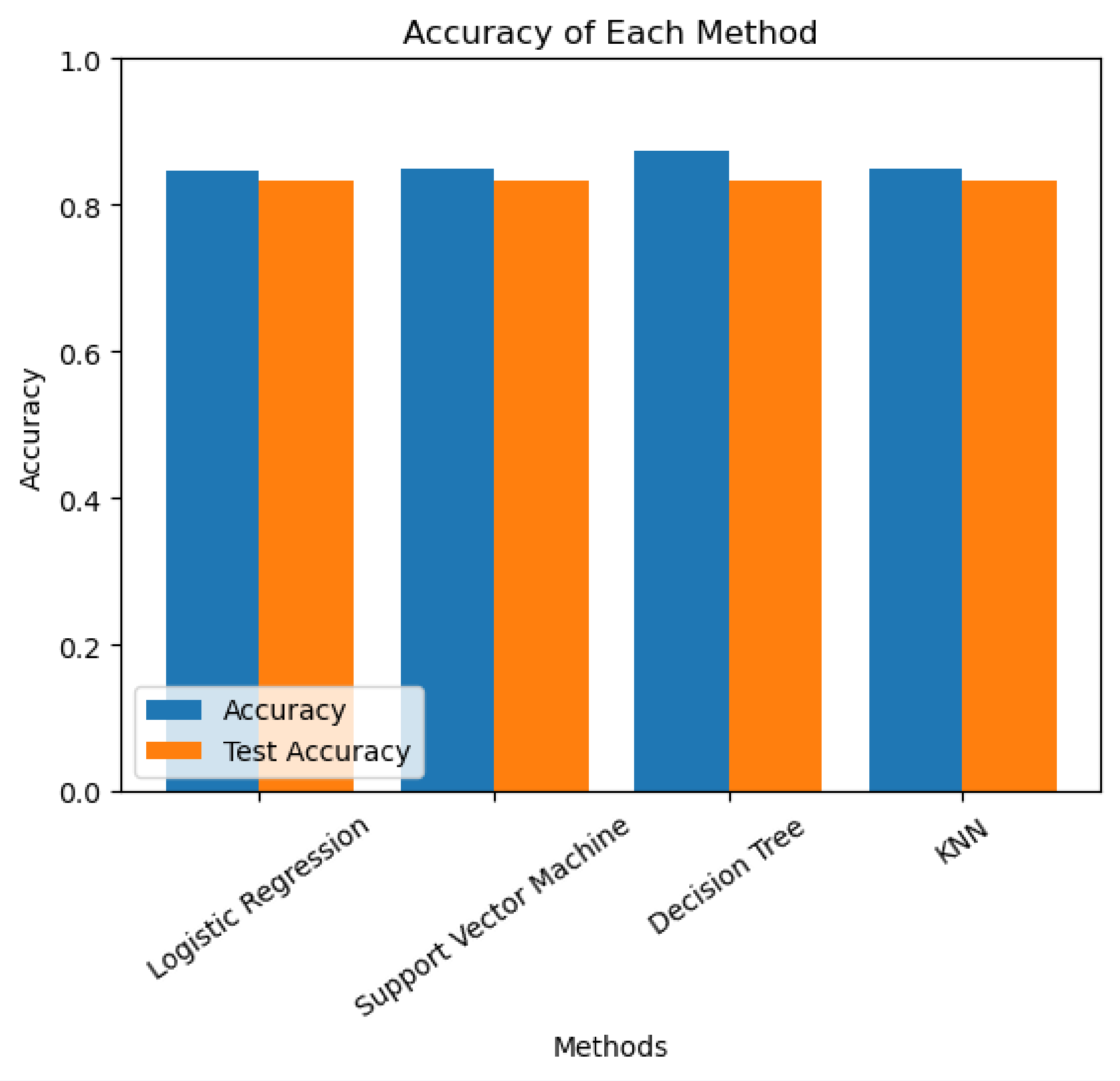
Section 5

Predictive Analysis (Classification)

Classification Accuracy

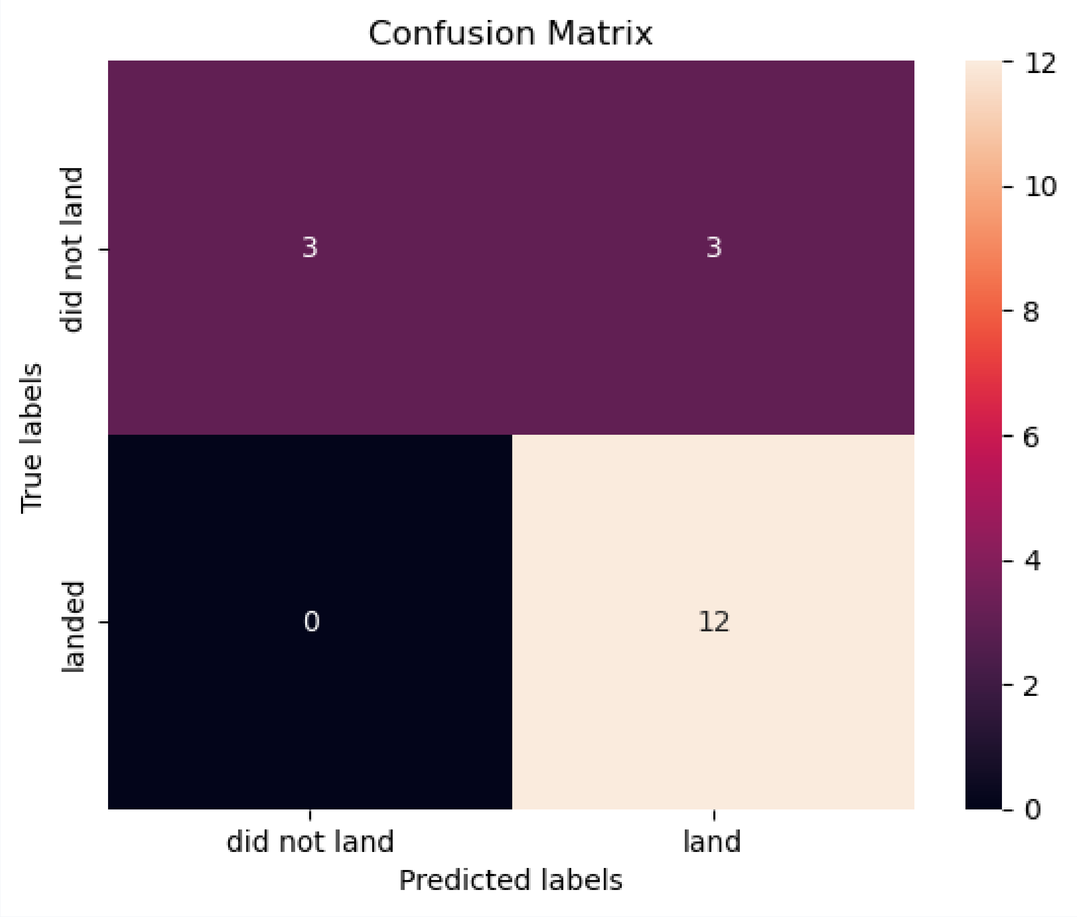
- 4 ML Classification models were trained, and it is possible to examine their Accuracy and **Test Accuracy Score.**

- **Decision Trees** displayed the higher Accuracy Score, but the Test Accuracy Score of the models are the same.



Confusion Matrix

- This is the Confusion Matrix of the Decision Tree Classifier:
    - 3 True Positives
    - 3 False Positives
    - 0 False Negatives
    - 12 True Negatives
- The model has correctly classified 3 positive instances (TP), incorrectly classified 3 negative instances as positive (FP), and correctly classified 12 negative instances (TN).
- **There are no false negative instances (FN), which means that all actual positive instances were correctly classified as positive.**



Conclusions

- Different datasets were analyzed, from different sources and conclusions were drawn from it;
- KSC LC-39A is the best launch site;
- Successful landing outcomes improve with time and experience, this appears to be not only trial and error, but also launching rockets from different sites.
- Although our Decision Tree Classifiers could be used to predict successful landings, it can be refined with more data and better hyperparameters.

Appendix

- It appears that there is a small **Data Leakage** when using 'GridSearchCV' to find the best parameters while also applying Cross-Validation.
  - This should be dealt with a Holdout set when splitting the datasets between train and test. It is possible to split these in three datasets: Train, Test and **Validation**.

- While the 'mean()' function was used to deal with missing values, it is possible to train a KNN MI model to also deal with this problem.

Thank you!