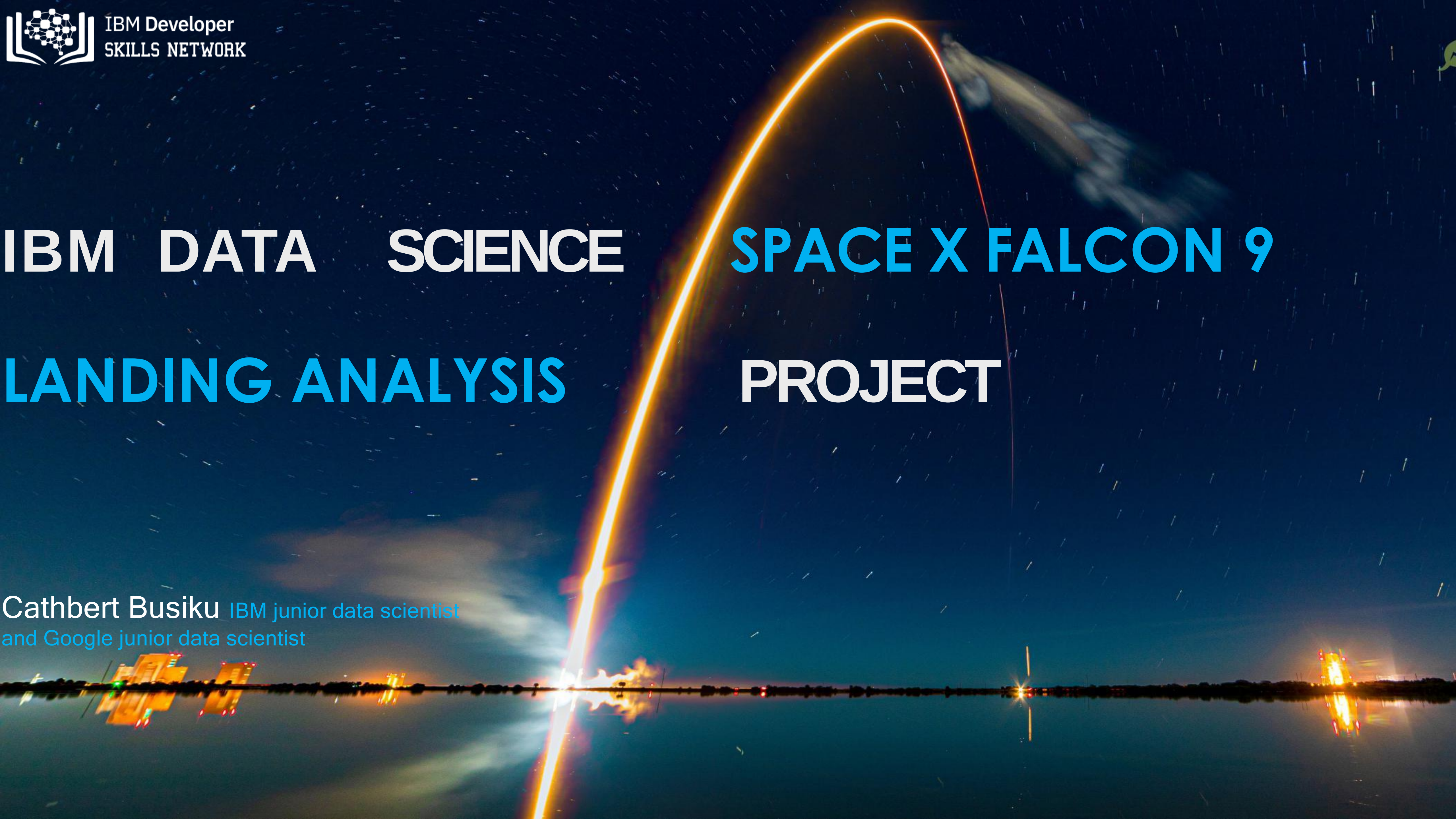
IBM Developer
SKILLS NETWORK



IBM DATA SCIENCE SPACE X FALCON 9 LANDING ANALYSIS PROJECT

Cathbert Busiku IBM junior data scientist and Google junior data scientist

IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

DR. JOSEPH SANTARCANGELO IBM SENIOR DATA SCIENTIST

CATHBERT BUSIKU IBM JUNIOR DATA SCIENTIST

30/04/22

OUTLINE

1. Executive Summary
2. Introduction
3. Methodology
4. Results
5. Conclusions
6. Appendix

EXECUTIVE SUMMARY

- Collected data from public SpaceX API and SpaceX Wikipedia page. Created labels column ‘class’ which classifies successful landings. Explored data using SQL, visualization, folium maps, and dashboards. Gathered relevant columns to be used as features. Changed all categorical variables to binary using one hot encoding. Standardized data and used GridSearchCV to find best parameters for machine learning models. Visualize accuracy score of all models.

Summary of Methodologies:

This project follows these steps:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis
- Interactive Visual Analytics
- Predictive Analysis (Classification)

INTRODUCTION

- SpaceX launches Falcon 9 rockets at a cost of around $62m while other providers launches at a cost of $165m.
- SpaceX launches is considerably cheaper than other providers. This is because SpaceX can successfully land and recover part of rocket (Stage 1) which it can reuse thus reducing the cost.
- Space Y wants to compete with Space X.
- Space Y want to determine the cost of launch by predicting whether the first stage will land.
- This project will ultimately predict if the Space X Falcon 9 first stage will land successfully.

METHODOLOGY SUMMARY

1. Data Collection

- Get data from SpaceX REST API and also Web Scraping from wikipedia
- Then combine the data from the two sources

2. Data Wrangling

- removing NaN values
- Counting the following values using Pandas
  - Number of launches on each site
  - Number and occurrence of each orbit
  - Number and occurrence of mission outcome per orbit type
- Creating a landing outcome label that shows the following:
  - 0 when the booster did not land successfully
  - 1 when the booster did land successfully

3. Exploratory Data Analysis

- Using SQL queries to manipulate and evaluate the SpaceX dataset
- Using Pandas and Matplotlib to visualize relationships between variables, and determine patterns

4. Interactive Visual Analytics

- Geospatial analytics using Folium
- Creating an interactive dashboard using Plotly Dash

5. Data Modelling and Evaluation

- Using Scikit-Learn to:
  - Pre-process (standardize) the data
  - Split the data into training and testing data using `train_test_split`
  - Train different classification models
  - Find the best model
- Plotting confusion matrices for each classification model
- Assessing the accuracy of each classification model

Data Collection Overview

Data was collected from two sources :

1. Data Collection from Space X public API

- data from SpaceX REST API had the following attributes

1. FlightNumber
2. Date
3. BoosterVersion
4. PayloadMass
5. Orbit
6. LaunchSite
7. Outcome
8. Flights
9. GridFins
10. GridFins
11. Reused
12. Legs
13. LandingPad
14. Block
15. ReusedCount
16. Serial
17. Longitude
18. Latitude

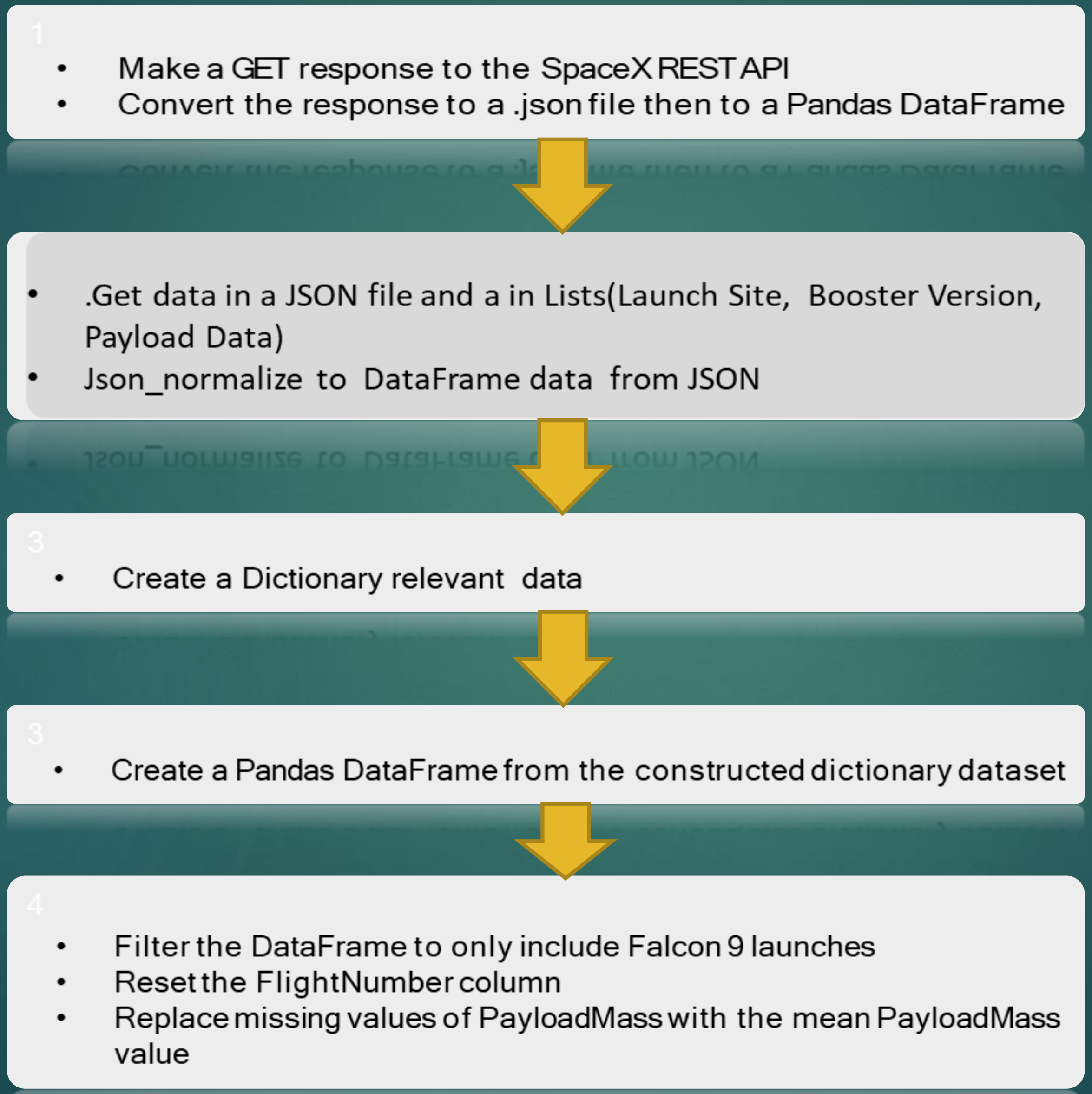
2. Web scraping data from Wikipedia Space X table

- data from web scrapping had the following attributes

1. Flight No.
2. Launch site
3. Payload
4. PayloadMass
5. Orbit
6. Customer
7. Launch outcome
8. Version Booster
9. Booster landing
10. Date
11. Time

- The next slide will show the process of data collection from API and the one after will show the process of data collection from web scraping.

DATA COLLECTION USING SPACE X REST API



1

- Make a GET response to the SpaceX REST API
- Convert the response to a .json file then to a Pandas DataFrame

- .Get data in a JSON file and a in Lists(Launch Site, Booster Version, Payload Data)
- Json_normalize to DataFrame data from JSON

3

- Create a Dictionary relevant data

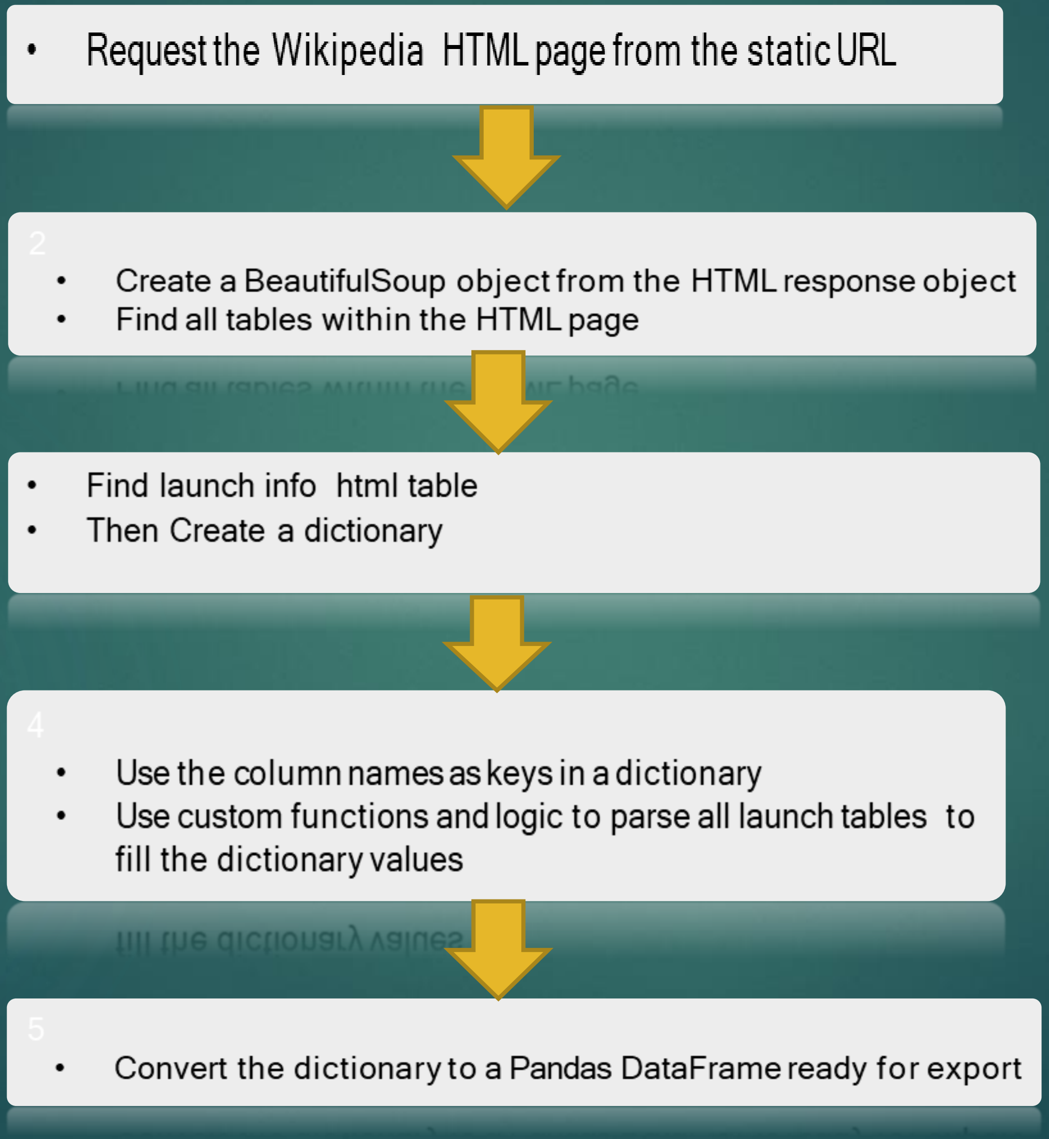
3

- Create a Pandas DataFrame from the constructed dictionary dataset

4

- Filter the DataFrame to only include Falcon 9 launches
- Reset the FlightNumber column
- Replace missing values of PayloadMass with the mean PayloadMass value

DATA COLLECTION USING WEB SCRAPPING



- Request the Wikipedia HTML page from the static URL

2

- Create a BeautifulSoup object from the HTML response object
- Find all tables within the HTML page

- Find launch info html table
- Then Create a dictionary

4

- Use the column names as keys in a dictionary
- Use custom functions and logic to parse all launch tables to fill the dictionary values

5

- Convert the dictionary to a Pandas DataFrame ready for export

DATA WRANGLING

To determine whether a booster will successfully land, it is best to have a binary column, i.e., where the value is 1 or 0, representing the success of the landing.

TYPES OF LANDING OUTCOMES

- The following are landing outcome shown in the `Outcome` column:
  - `True Ocean:` the mission outcome was successfully landed to a specific region of the ocean
  - `False Ocean:` the mission outcome was unsuccessfully landed to a specific region of the ocean.
  - `True RTLS:` the mission outcome was successfully landed to a ground pad
  - `False RTLS:` the mission outcome was unsuccessfully landed to a ground pad.
  - `True ASDS:` the mission outcome was successfully landed to a drone ship
  - `False ASDS:` the mission outcome was unsuccessfully landed to a drone ship.
  - `None ASDS` and `None None` - these represent a failure to land.

1. Defining a set of unsuccessful outcomes, `bad_outcome`
2. Creating a list, `landing_class`, where the element is 0 if the corresponding `row in Outcome` is in the set `bad_outcome`, otherwise, it's 1.
3. Create a `Class` column that contains the values from the list `landing_class`
4. Export the DataFrame as a .csv file.

EXPLORATORY DATA ANALYSIS (EDA) WITH DATA VISUALIZATION

Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

PLOTS USED:

1. SCATTER PLOTS
   - Flight Number and Launch Site
   - Payload and Launch Site
   - Orbit Type and Flight Number
   - Payload and Orbit Type
2. BAR CHART

   A bar chart was produced to visualize the relationship between:
   - Success Rate and Orbit Type
3. LINE CHART
   - Success Rate and Year (i.e. the launch success yearly trend)

EXPLORATORY DATA ANALYSIS (EDA) SQL

The following process were done with the database:

- Loaded data set into IBM DB2 Database.
- Queried using SQL Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

The following are the queries that were performed:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string ‘CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display the average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome on a ground pad was achieved
- List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg
- List the total number of successful and failed mission outcomes

Build an interactive map with Folium

- Folium maps mark Launch Sites as successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.
- This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

Success/failed launches were calculated as follows:

- Before clustering them, assign a marker colour of successful (class = 1) as green, and failed (class = 0) as red.
- To put the launches into clusters, for each launch, add a `folium.Marker` to the `MarkerCluster()` object.
- Create an icon as a text label, assigning the `icon_color` as the `marker_colour` determined previously.

The distances between a launch site to its proximities was calculated as follows:

- To explore the proximities of launch sites, calculations of distances between points can be made using the `Lat` and `Long` values.
- After marking a point using the `Lat` and `Long` values, create a `folium.Marker` object to show the distance.
- To display the distance line between two points, draw a `folium.PolyLine` and add this to the map.

INTERACTIVE DASHBOARD WITH PLOTLY DASH

The following plots were added to a Plotly Dash dashboard to have an interactive visualisation of the data:

1. Pie chart (`px.pie()`) showing the total successful launches per site
   - This makes it clear to see which sites are most successful
   - The chart could also be filtered (using a `dcc.Dropdown()` object) to see the success/failure ratio for an individual site
2. Scatter graph (`px.scatter()`) to show the correlation between outcome (success or not) and payload mass (kg)
   - This could be filtered (using a `RangeSlider()` object) by ranges of payload masses
   - It could also be filtered by booster version

PREDICTIVE ANALYSIS USING CLASSIFICATION

The following steps were taking to find the best performing classification model:

1. MODEL DEVELOPMENT
   - Load dataset
   - Perform necessary data transformations
   - (standardize and pre-process)
   - Split data into training and test data sets, using train_test_split()
   - Decide which type of machine learning algorithms are most appropriate
   - Create a `GridSearchCV` object and a dictionary of parameters
2. MODEL EVALUATION
   - Check the tuned hyper parameters or best_params_
   - Check the accuracy the score and best_score

3. Finding the Best Classification Model
   - Review the accuracy scores for
   - all chosen algorithms
   - The model with the highest accuracy score is determined as the best performing model

RESULTS

Explanatory Data Analysis

Interactive Analysis

Predictive Analysis

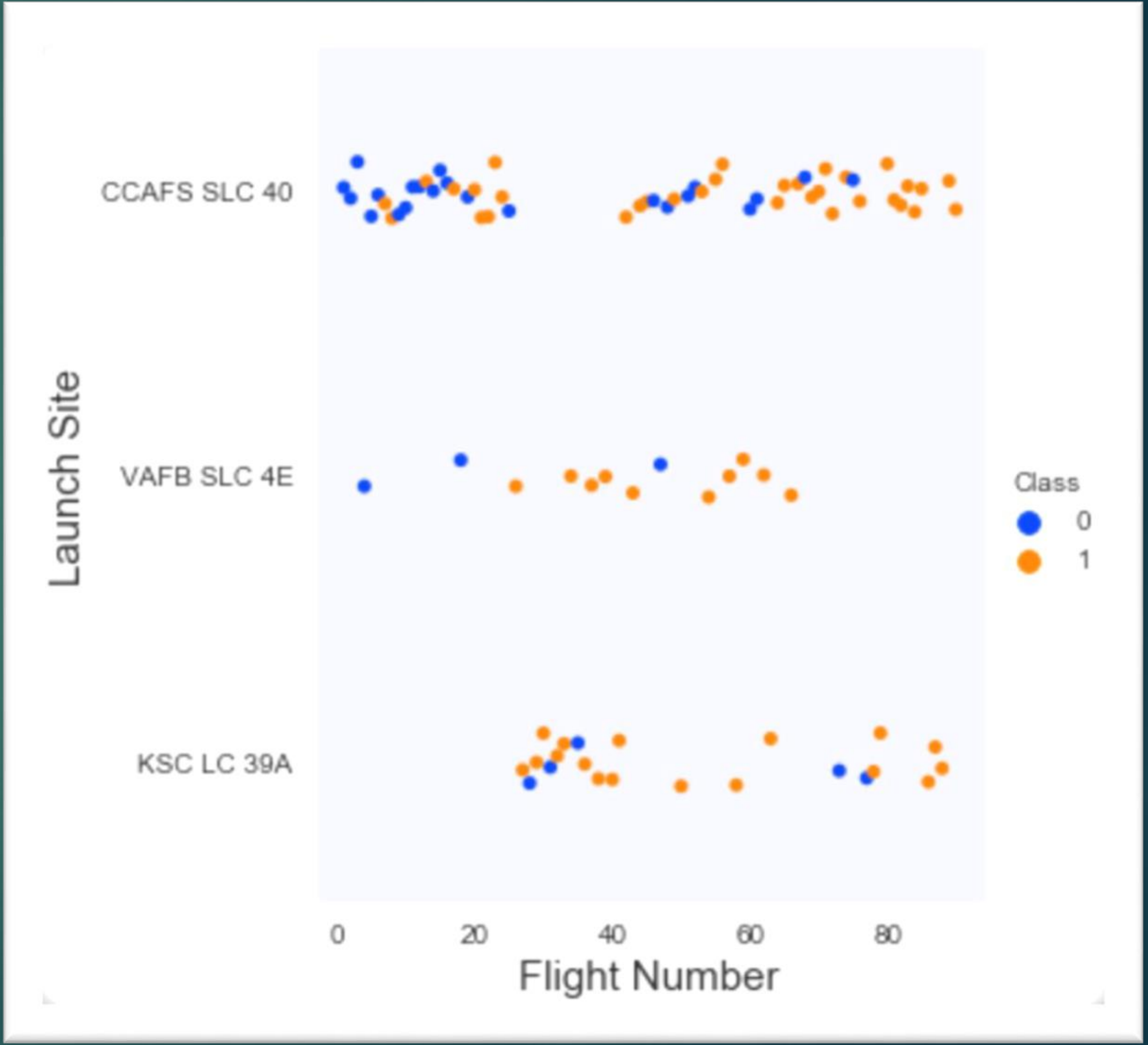
EDA WITH VISUALIZATION

EXPLORATORY DATA ANALYSIS WITH SEABORN PLOTS

LAUNCH SITE VS FLIGHT NUMBER

The scatter plot of Launch Site vs Flight Number shows that:

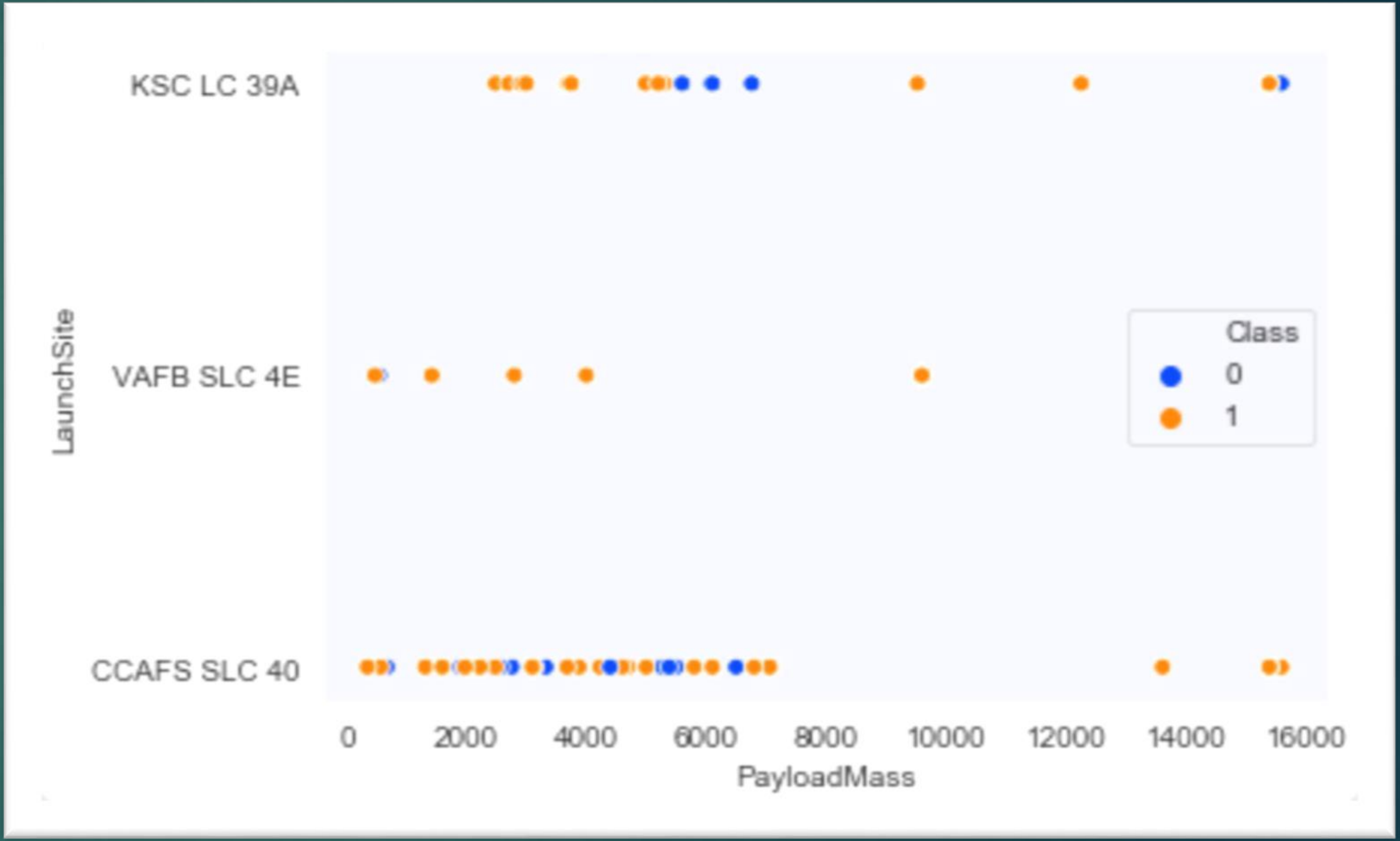
- Orange indicates successful launch; Blue indicates unsuccessful launch.
- Graphic suggests an increase in success rate over time (indicated in Flight Number).
- As the number of flights increases, the rate of success at a launch site increases.
- Most of the early flights (flight numbers < 30) were launched from CCAFS SLC 40, and were generally unsuccessful.
- Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.



LAUNCH SITE VS PAYLOAD MASS

The scatter plot of Launch Site vs Payload Mass shows that:

- Green indicates successful launch; Purple indicates unsuccessful launch.
- Payload mass appears to fall mostly between 0-6000 kg. Different launch sites also seem to use different payload mass.
- Above a payload mass of around 7000 kg, there are very few unsuccessful landings, but there is also far less data for these heavier launches.
- There is no clear correlation between payload mass and success rate for a given launch site.



SUCCESS RATE VS ORBIT TYPE

- ES-L1, GEO, HEO and SSO have 100% success rate (sample sizes in parenthesis)

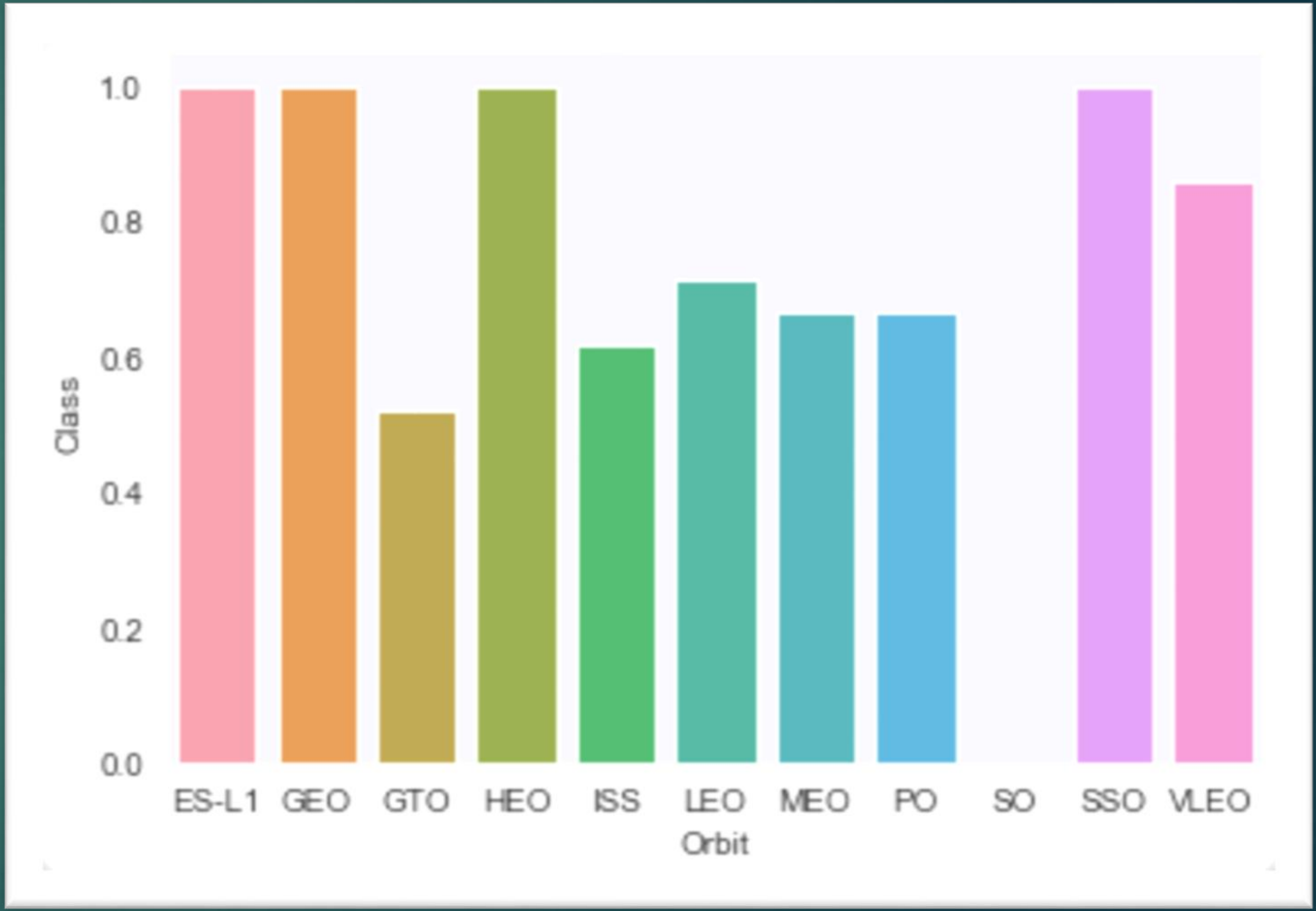
- VLEO has decent success rate and attempts

- SO has 0% success rate

- GTO has the around 50% success rate but largest sample

The orbit with the lowest (0%) success rate is:

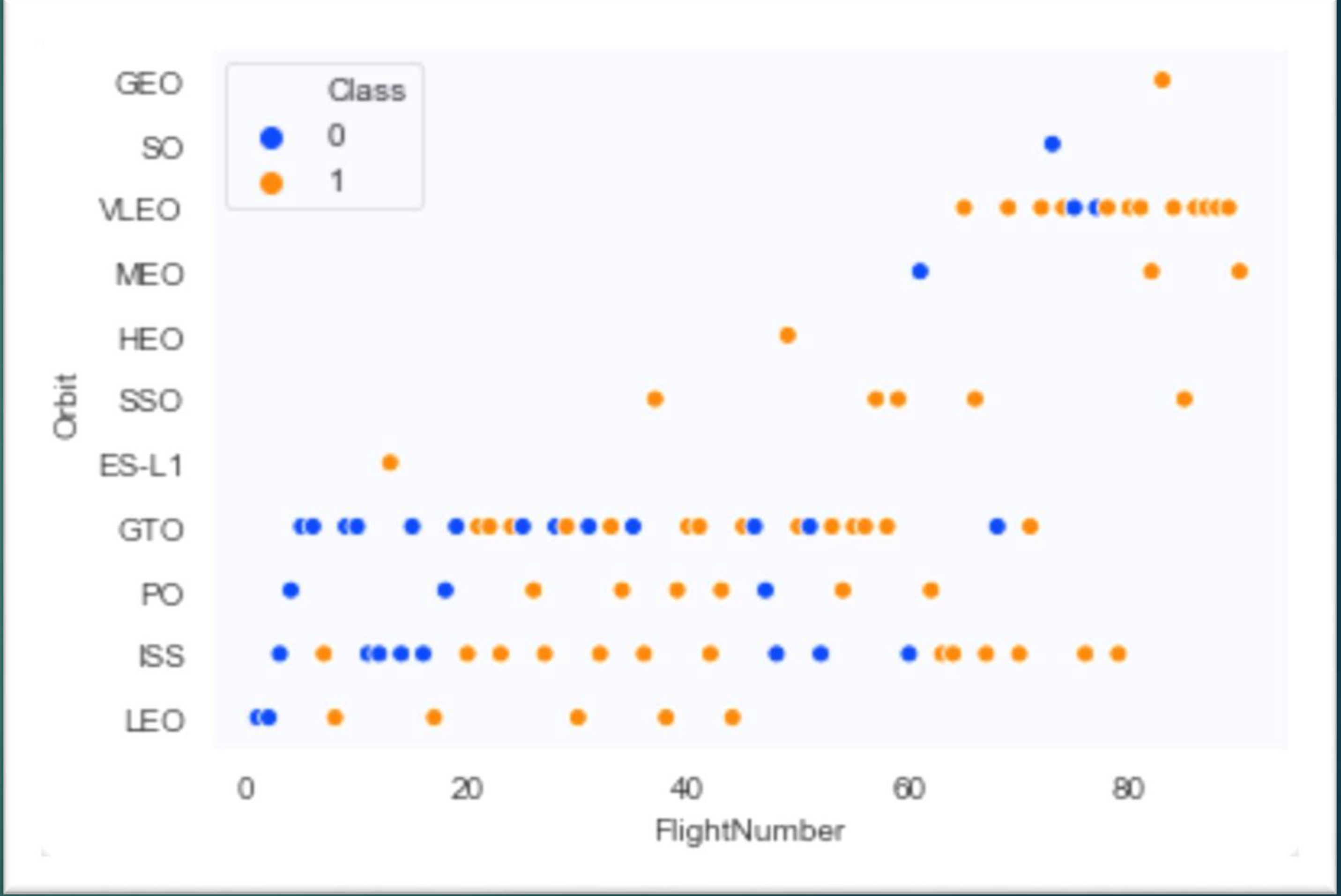
- SO (Heliocentric Orbit)



ORBIT TYPE VS FLIGHT NUMBER

This scatter plot of Orbit Type vs. Flight number shows a few useful things that the previous plots did not, such as:

- The 100% success rate of GEO, HEO, and ES-L1 orbits can be explained by only having 1 flight into the respective orbits.
- The 100% success rate in SSO is more impressive, with 5 successful flights.
- There is little relationship between Flight Number and Success Rate for GTO.
- Generally, as Flight Number increases, the success rate increases. This is most extreme for LEO, where unsuccessful landings only occurred for the low flight numbers (early launches).

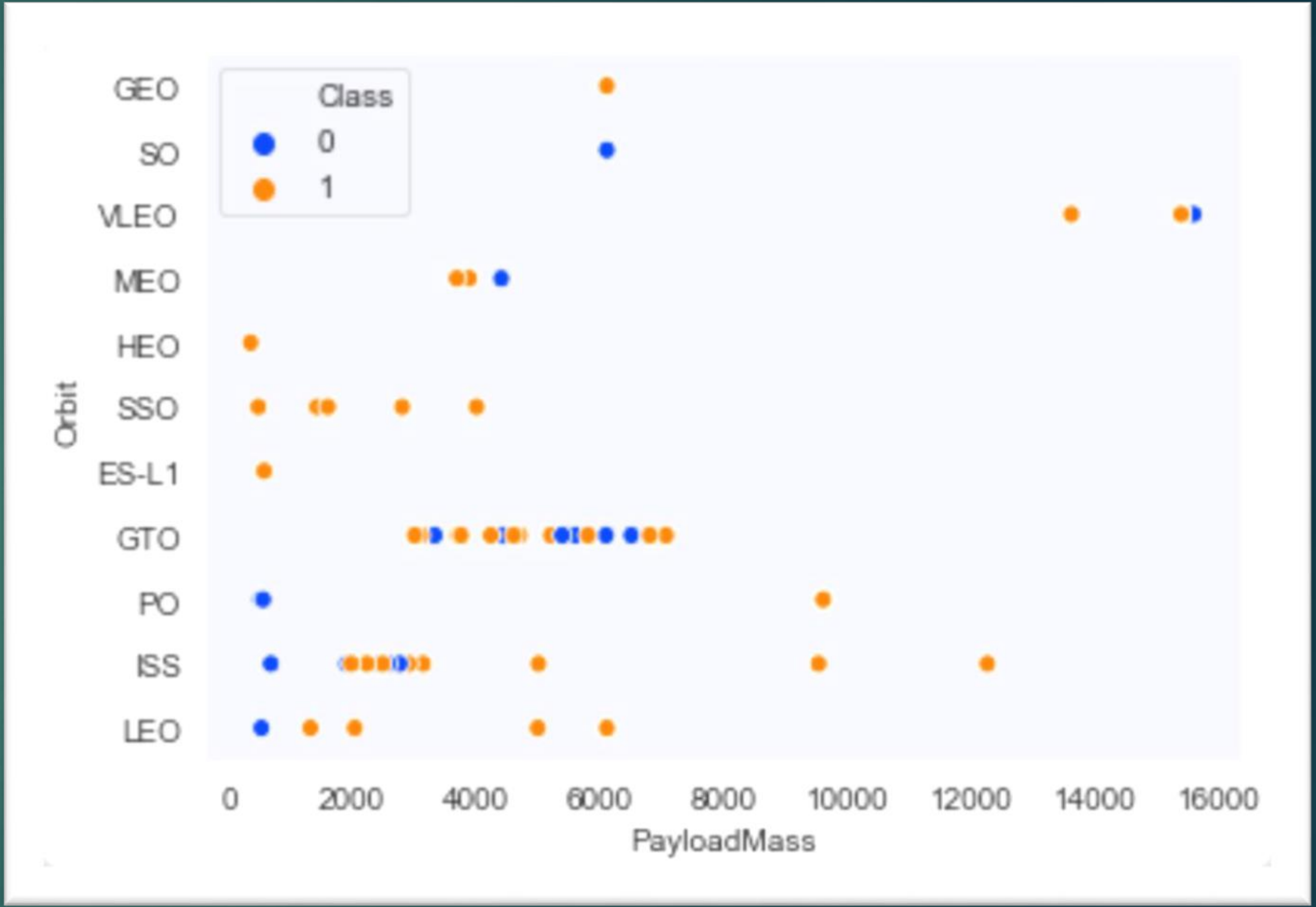


ORBIT TYPE VS PAYLOAD MASS

- Payload mass seems to correlate with orbit

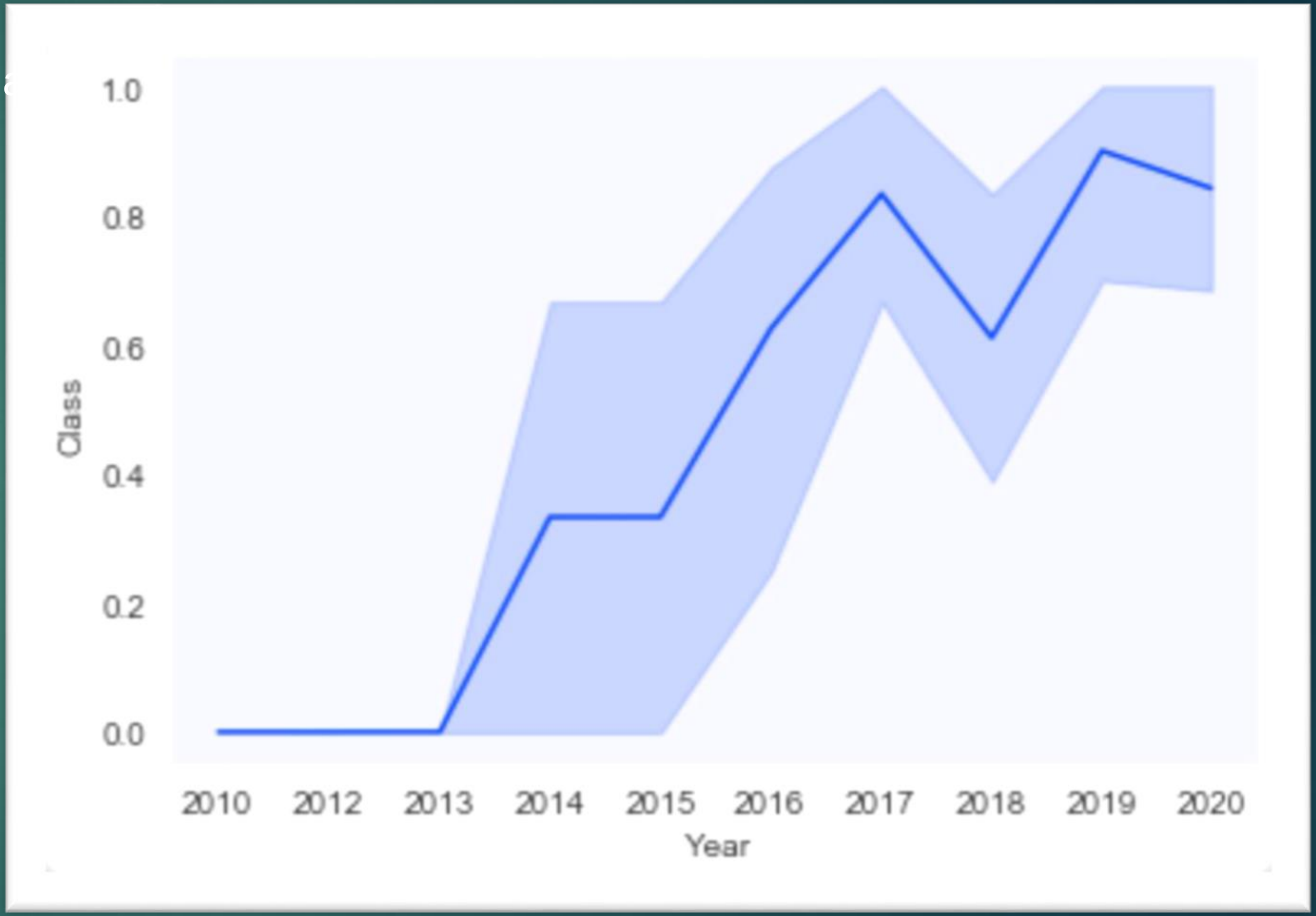
  LEO and SSO seem to have relatively low payload mass

- The other most successful orbit VLEO only has payload mass values in the higher end of the range



LAUNCH SUCCESS YEARLY TREND

- Success generally increases over time since 2013 with a slight dip in 2018
- Success in recent years at around 80%

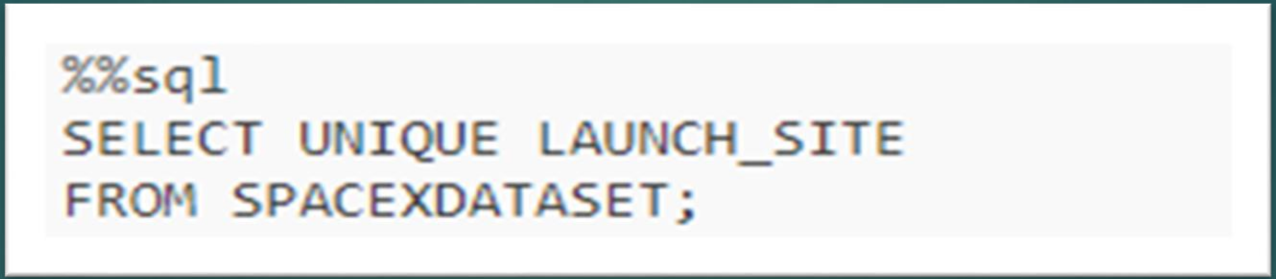


EDA WITH SQL

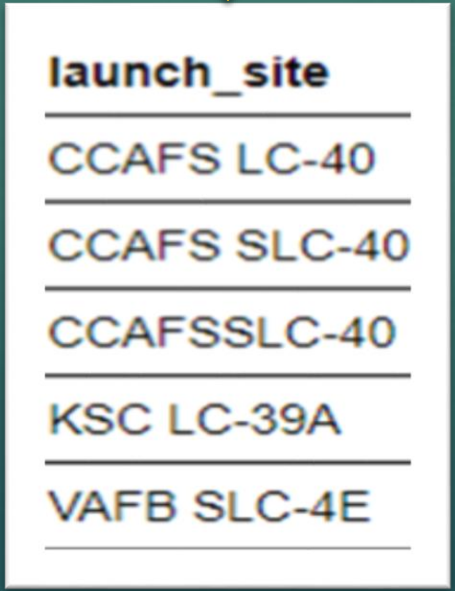
EXPLORATORY DATA ANALYSIS WITH SQL DB2

ALL LAUNCH SITE NAMES

Find the names of the unique launch sites.

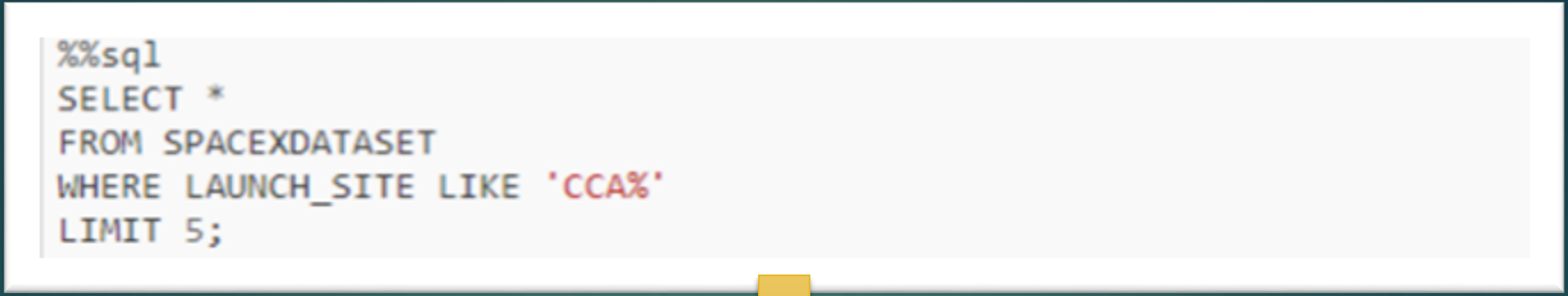


```
%%sql
SELECT UNIQUE LAUNCH_SITE
FROM SPACEXDATASET;
```

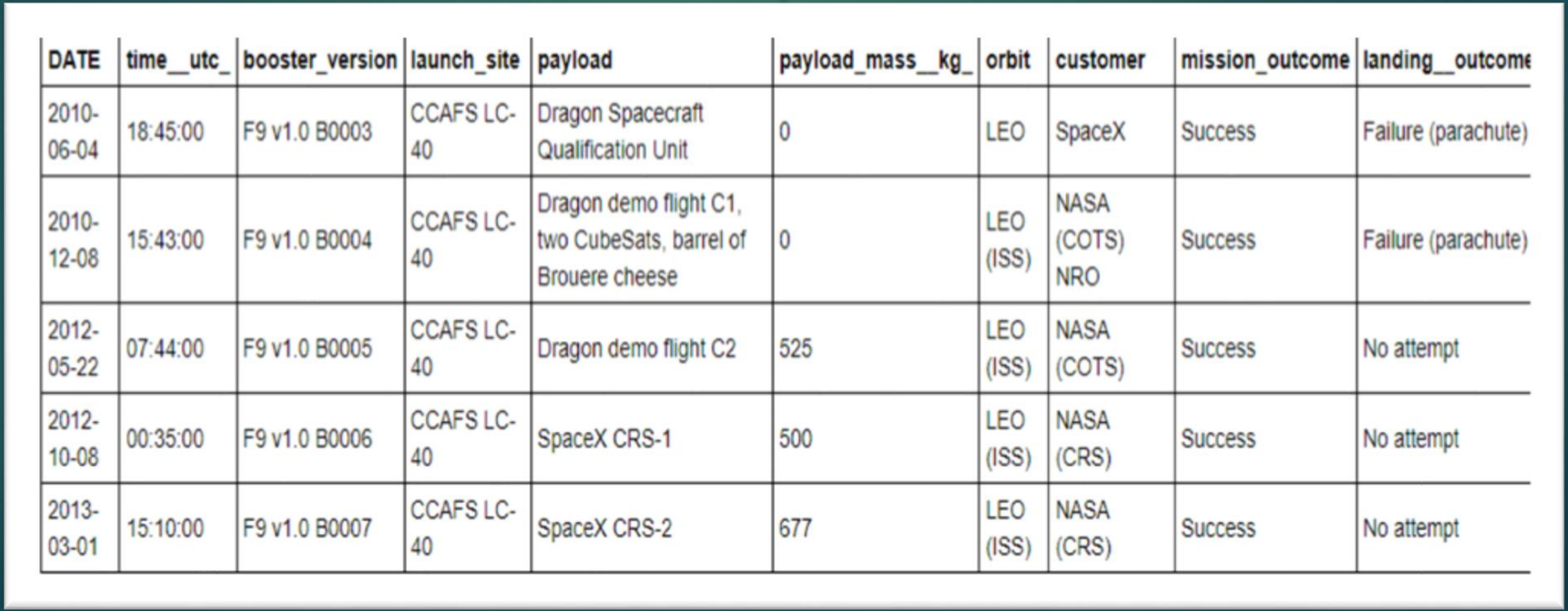


launch_site
CCAFS LC-40
CCAFS SLC-40
CCAFSSLC-40
KSC LC-39A
VAFB SLC-4E

LAUNCH SITE NAMES BEGIN WITH 'CCA'



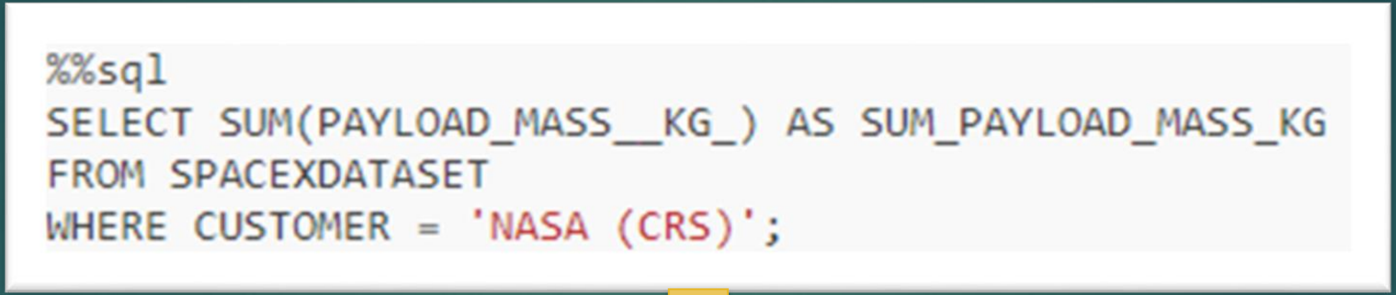
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```



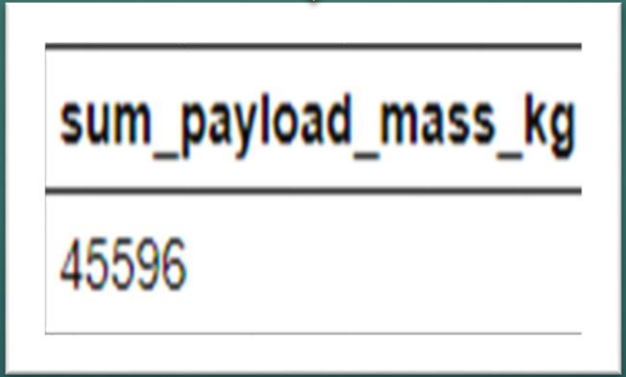
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

TOTAL PAYLOAD MASS

This query sums the total payload mass in kg where NASA was the customer.



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

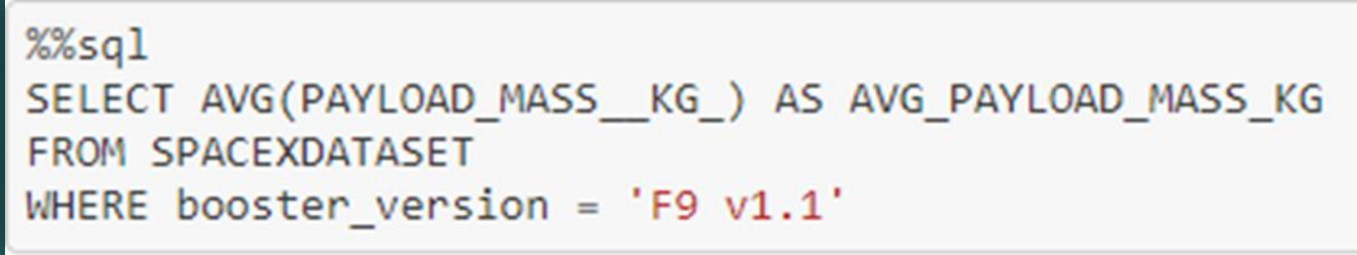


sum_payload_mass_kg
45596

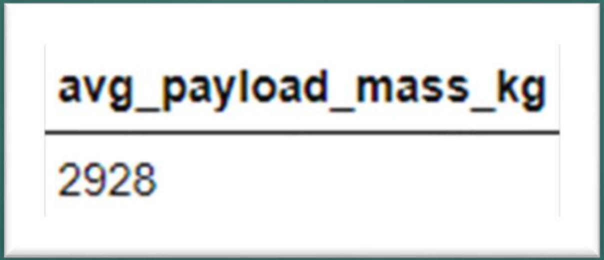
- CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

AVERAGE PAYLOAD MASS BY F9 V1.1

This Query Calculate the average payload mass carried by booster version F9 v1.1



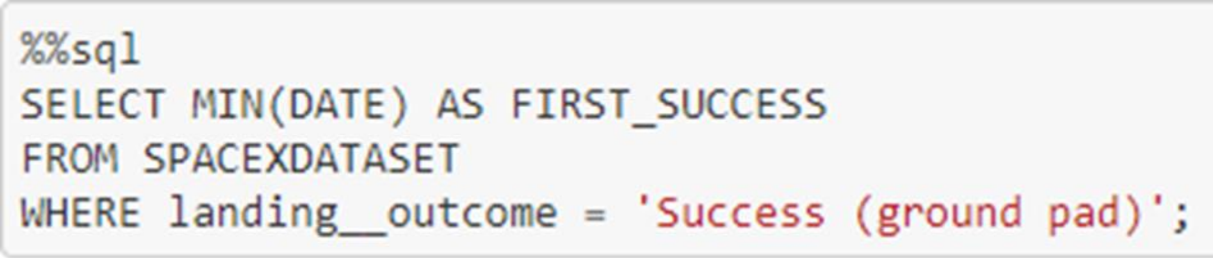
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```



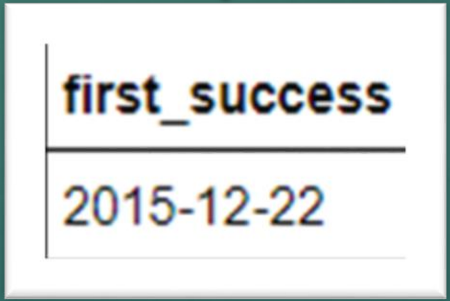
avg_payload_mass_kg
2928

FIRST SUCCESSFUL GROUND LANDING DATE

Find the dates of the first successful landing outcome on ground pad.



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

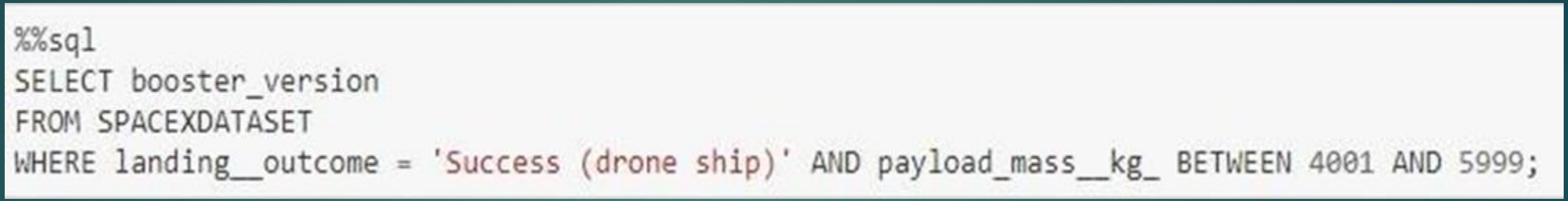


first_success
2015-12-22

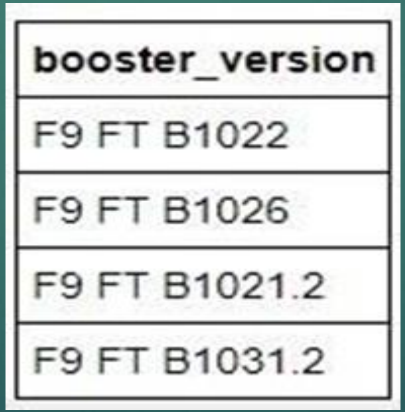
- First ground pad landing wasn’t until the end of 2015.

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively.



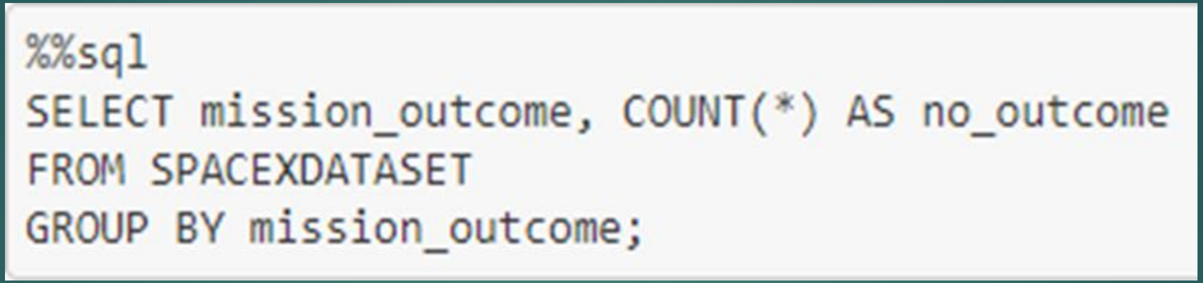
```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```



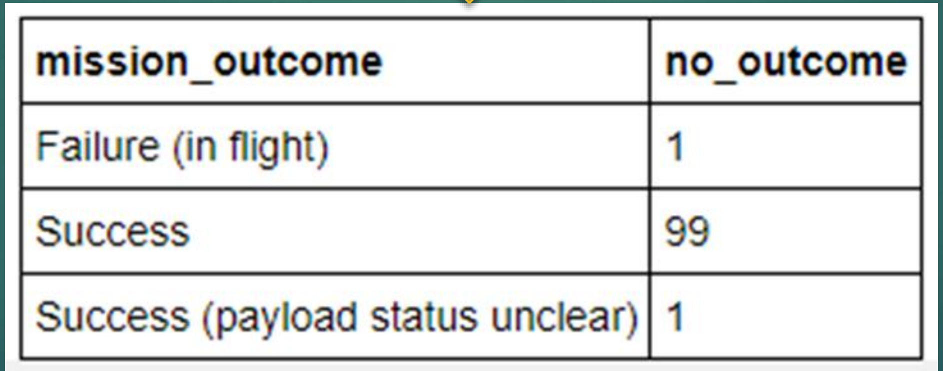
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

This query returns a count of each mission outcome.



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

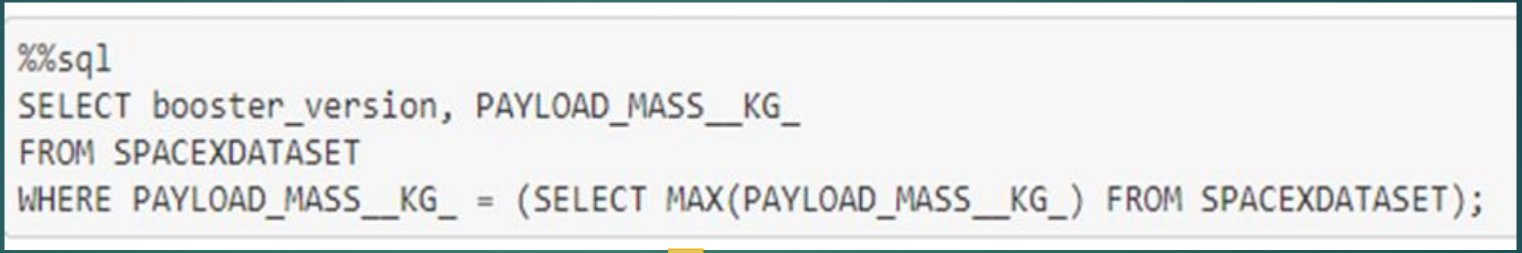


mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

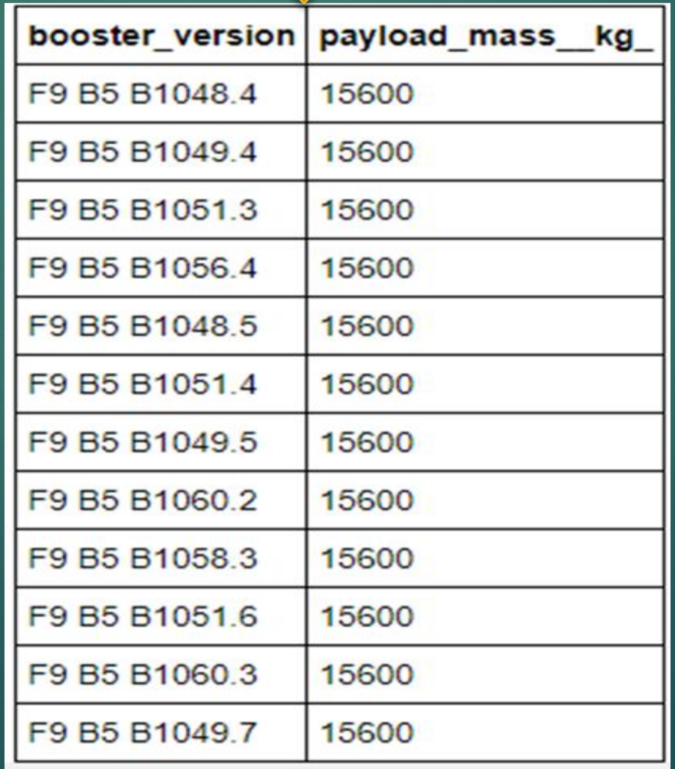
- SpaceX appears to achieve its mission outcome nearly 99% of the time.
- This means that most of the landing failures are intended.

BOOSTERS CARRIED MAXIMUM PAYLOAD

List the names of the booster which have carried the maximum payload mass.



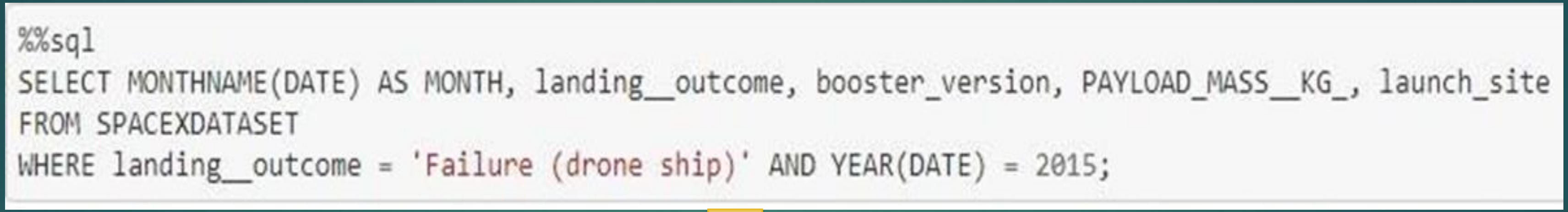
```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```



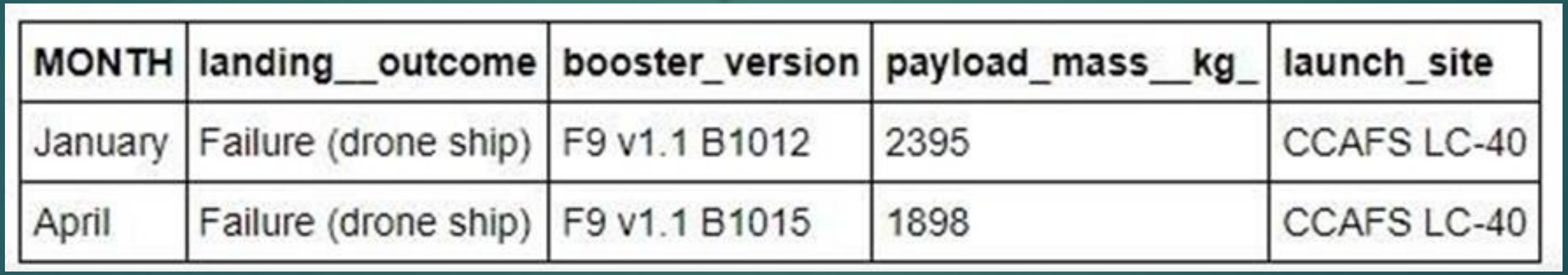
booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 LAUNCH RECORDS

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.



```
%%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;
```

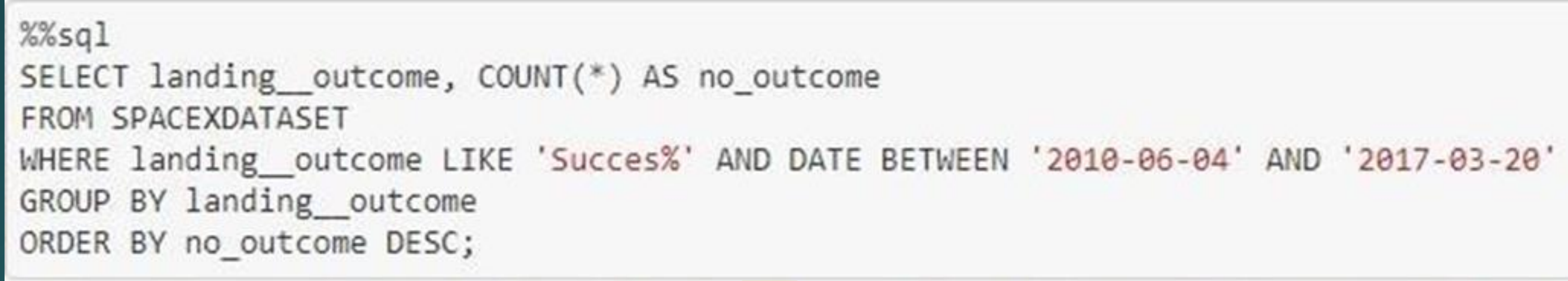


MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

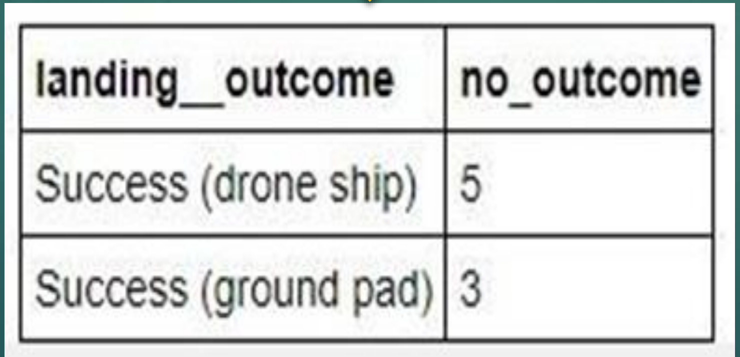
RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.

There are two types of successful landing outcomes: drone ship and ground pad landings.



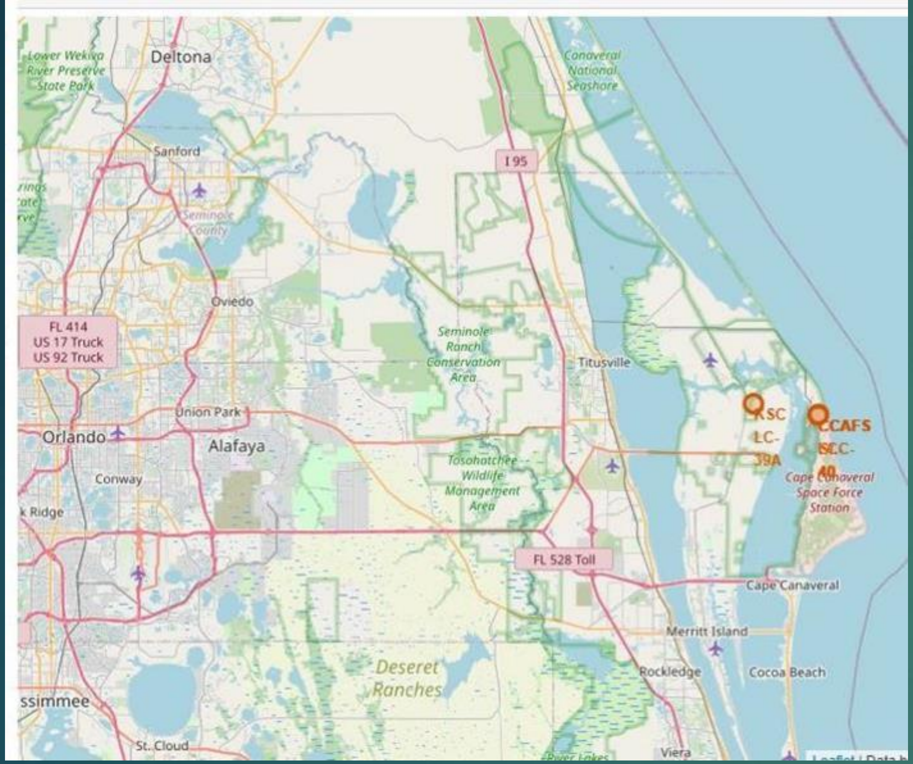
```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

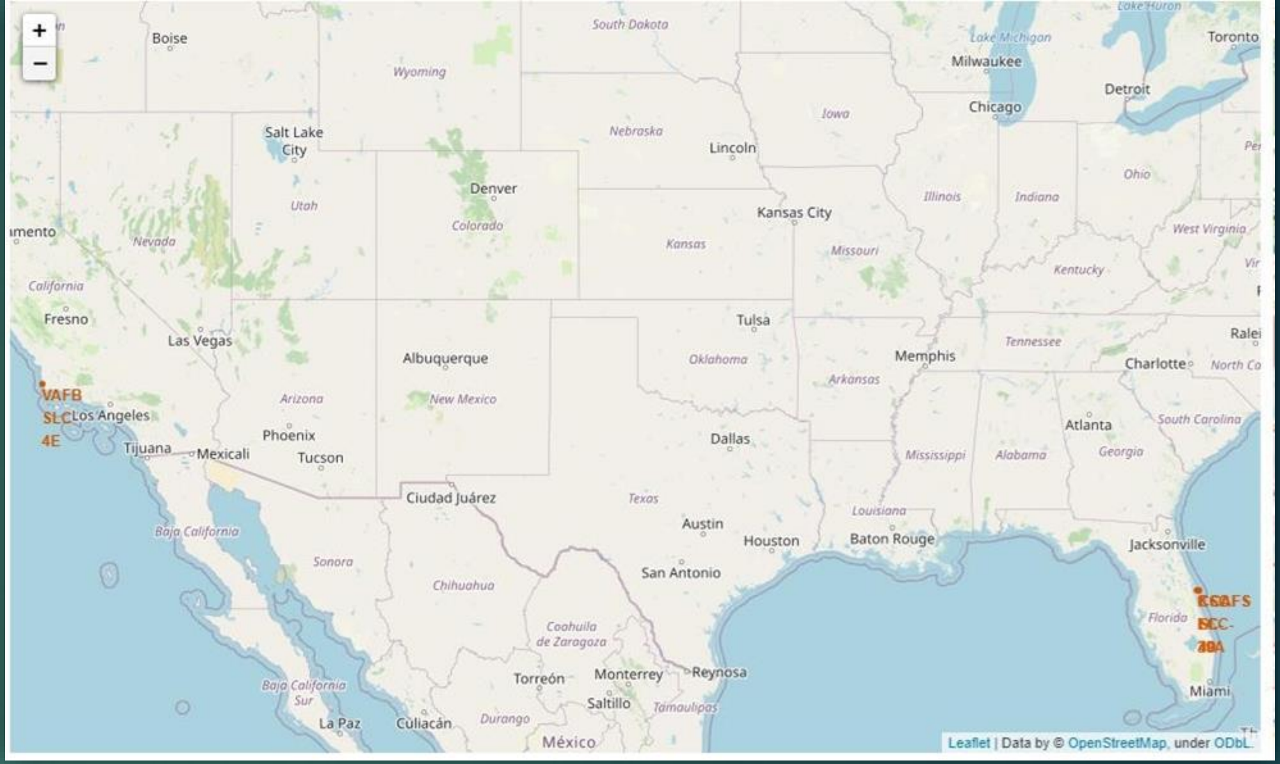


landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

INTERACTIVE MAP ANALYSIS WITH FOLIUM

ALL LAUNCH SITES LOCATIONS ON A MAP

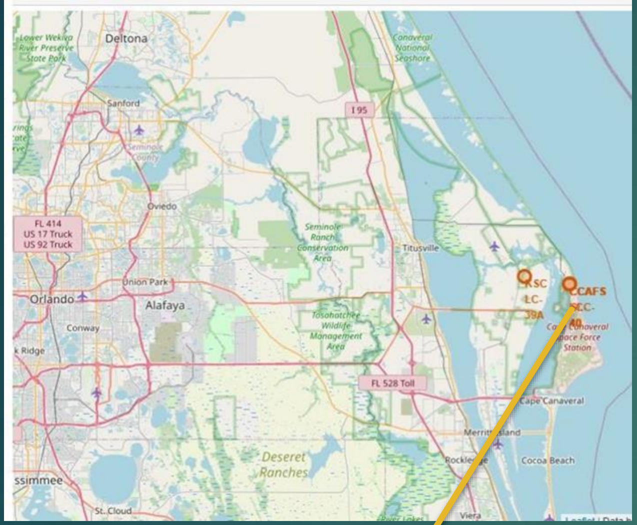


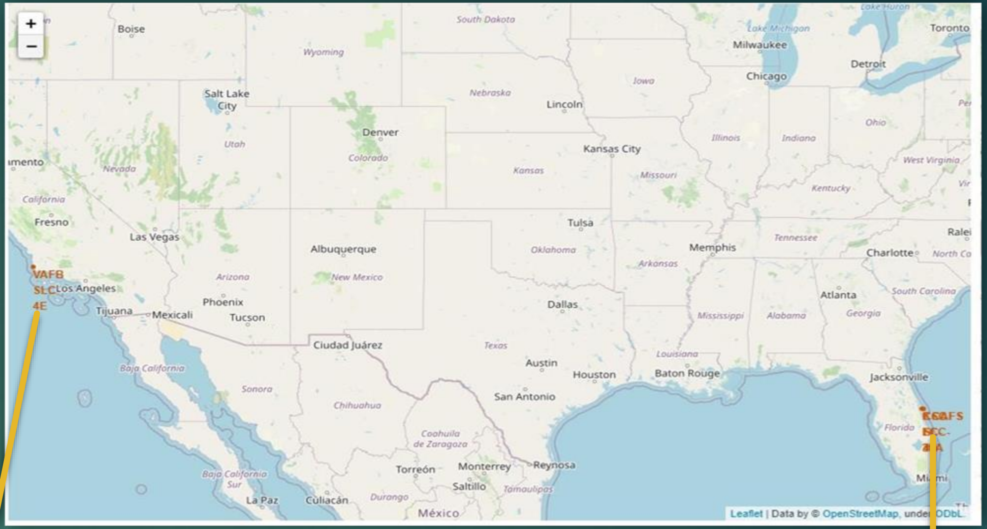


This map shows two Florida launch sites and they are very close to each other. All launch sites are near the ocean.

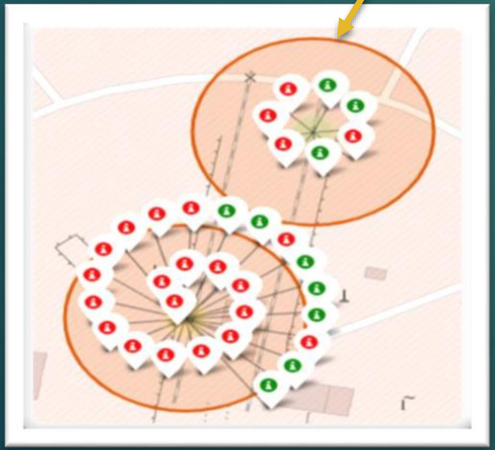
All SpaceX launch sites are on coasts of the United States of America.

SUCCESS AND FAILED LAUNCHES FOR EACH SITE

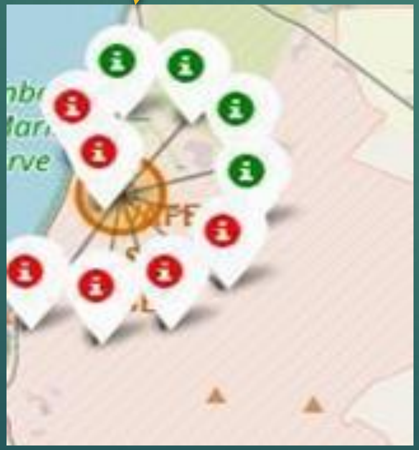




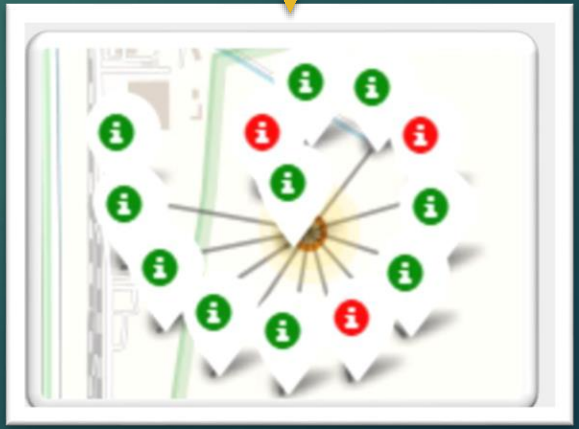
Launches have been grouped into clusters, and annotated with green icons for successful launches, and red icons for failed launches.



CCAFS LC-40



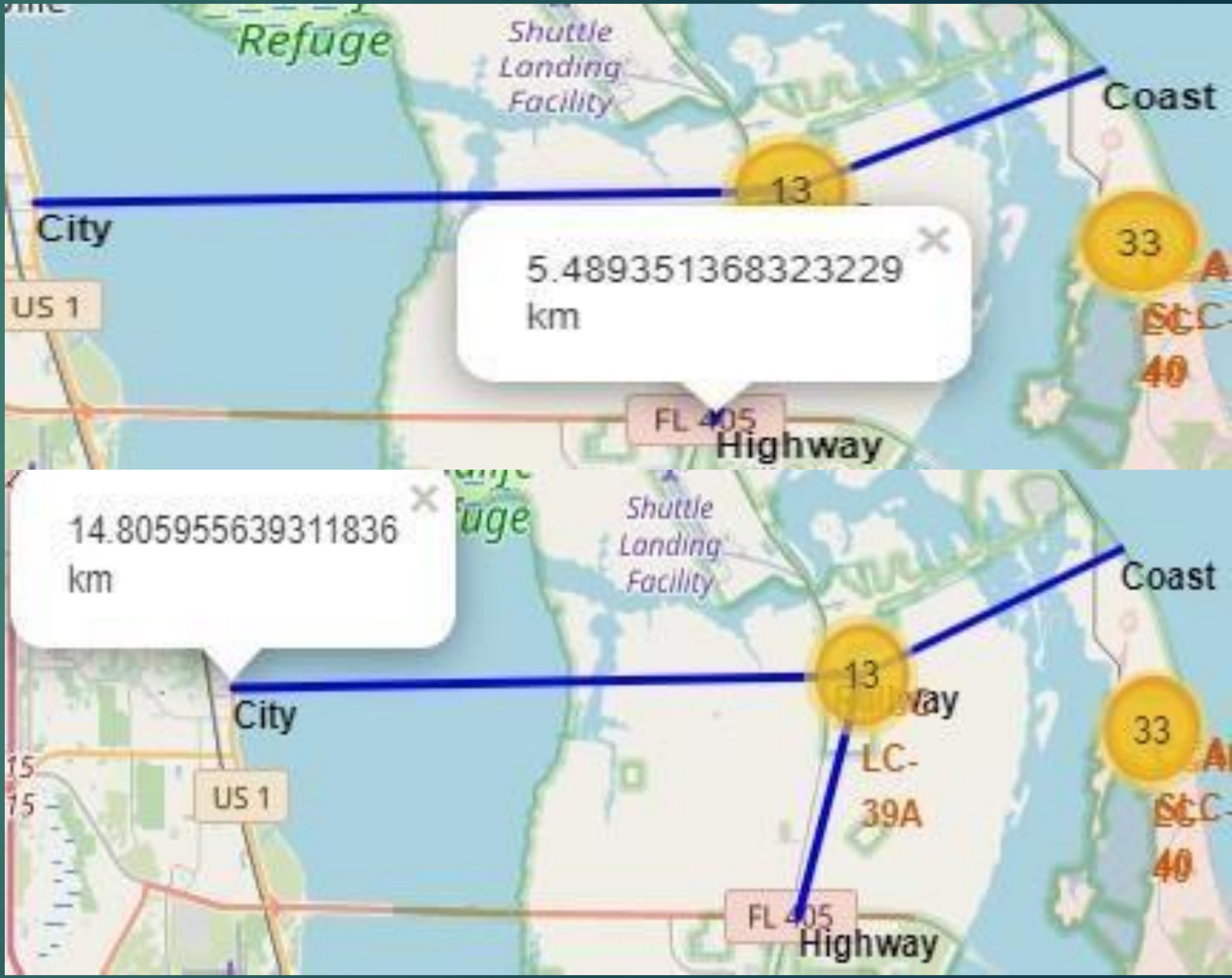
VAFB SLC-4E



KSC LC-39A

PROXIMITY OF LAUNCH SITES TO OTHER POINTS OF INTEREST

- Using launch site KSC LC-39A as an example, it is evident that launch sites are very close to railways for large part and supply transportation.
- Launch sites are close to highways for human and supply transport.
- Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.



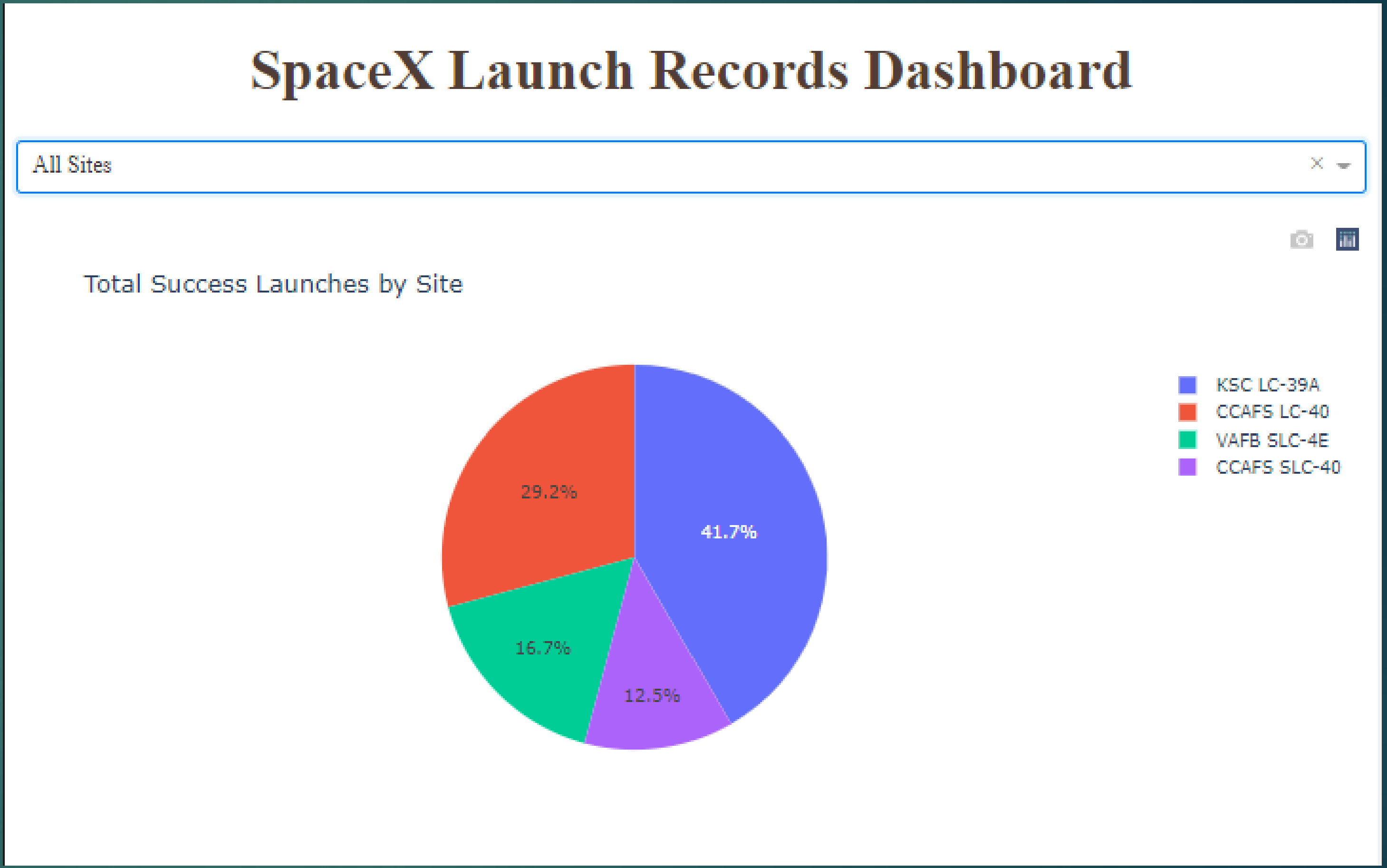
5.489351368323229 km

14.805955639311836 km

INTERACTIVE DASHBOARD USING **PLOTLY DASH**

LAUNCH SUCCESS COUNT FOR ALL SITES

This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

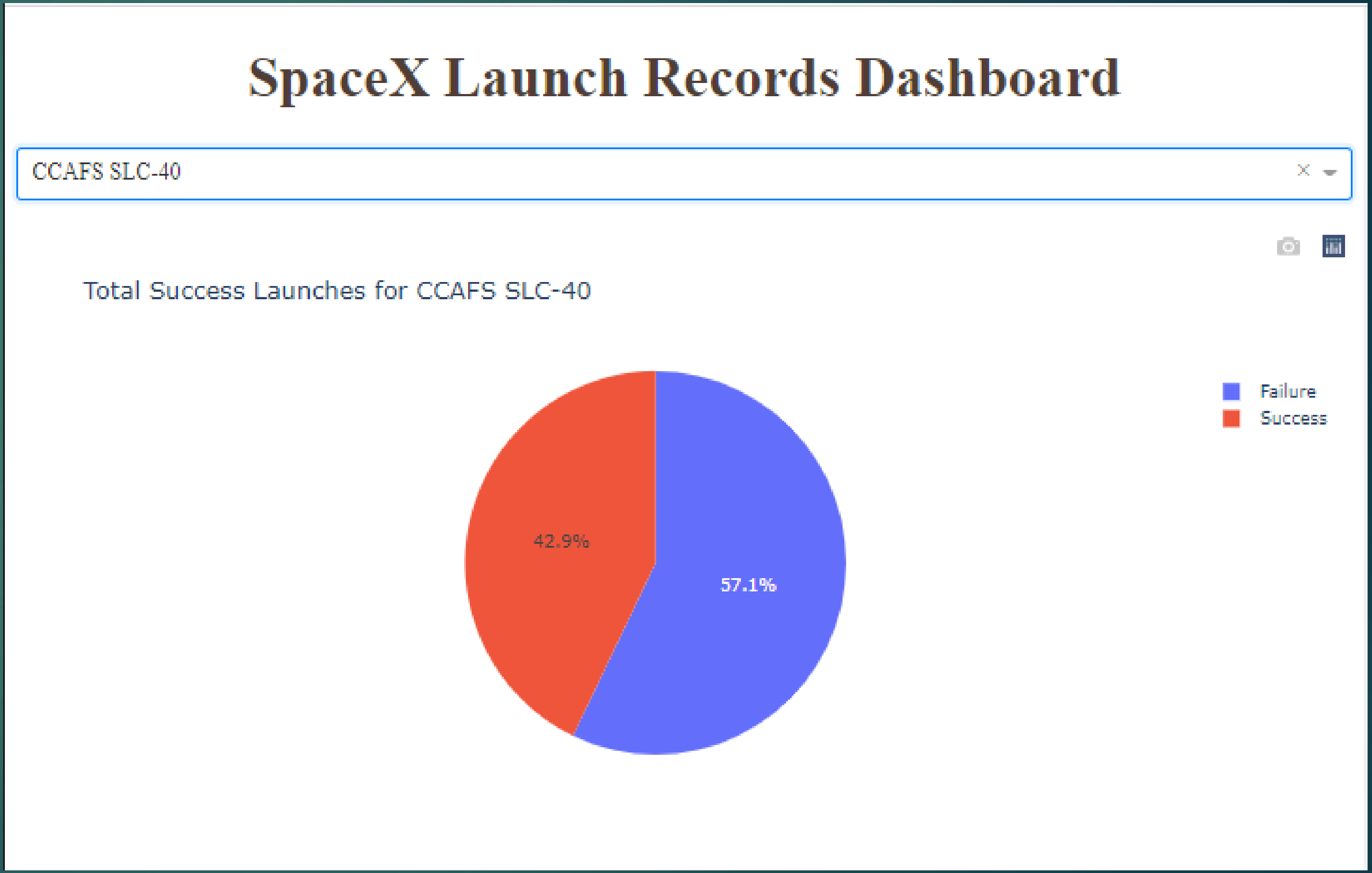


SpaceX Launch Records Dashboard

All Sites

LOWEST SUCCESS LAUNCH SITE

CCAFS SLC-40 has the lowest success rate, with a 57.1% success rate.



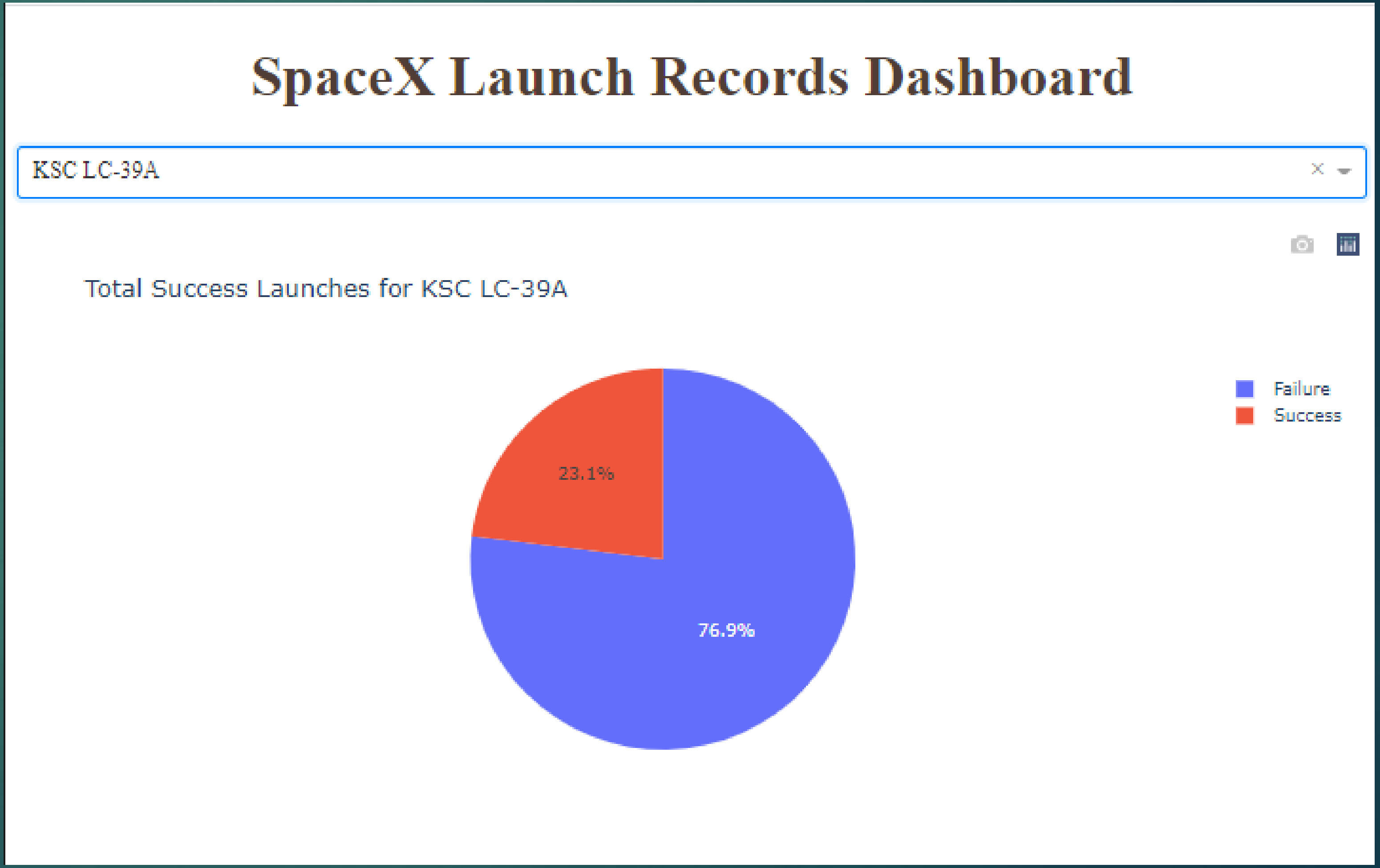
SpaceX Launch Records Dashboard

CCAFS SLC-40

Total Success Launches for CCAFS SLC-40

HIGHEST SUCCESS LAUNCH SITE

KSC LC-39A has the highest success rate, with a 76.9% success rate.



SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for KSC LC-39A

PREDICTIVE ANALYSIS CLASSIFICATION

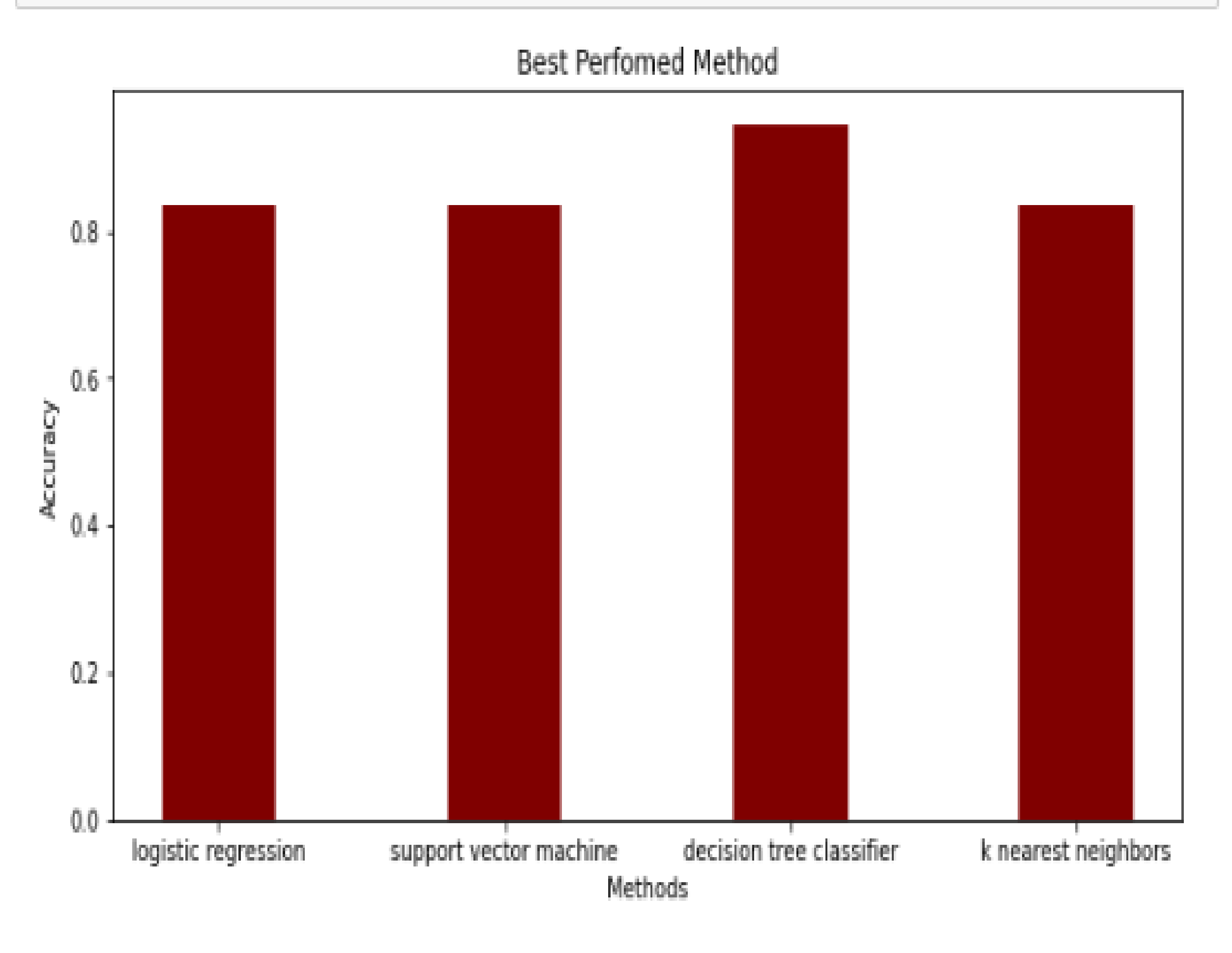
CLASSIFICATION ACCURACY

The Graph shows Accuracy Score and Best Score for each classification algorithm.

The Decision Tree model has the highest classification accuracy:

- The Accuracy Score is 94.44%
- The Best Score is 90.36%

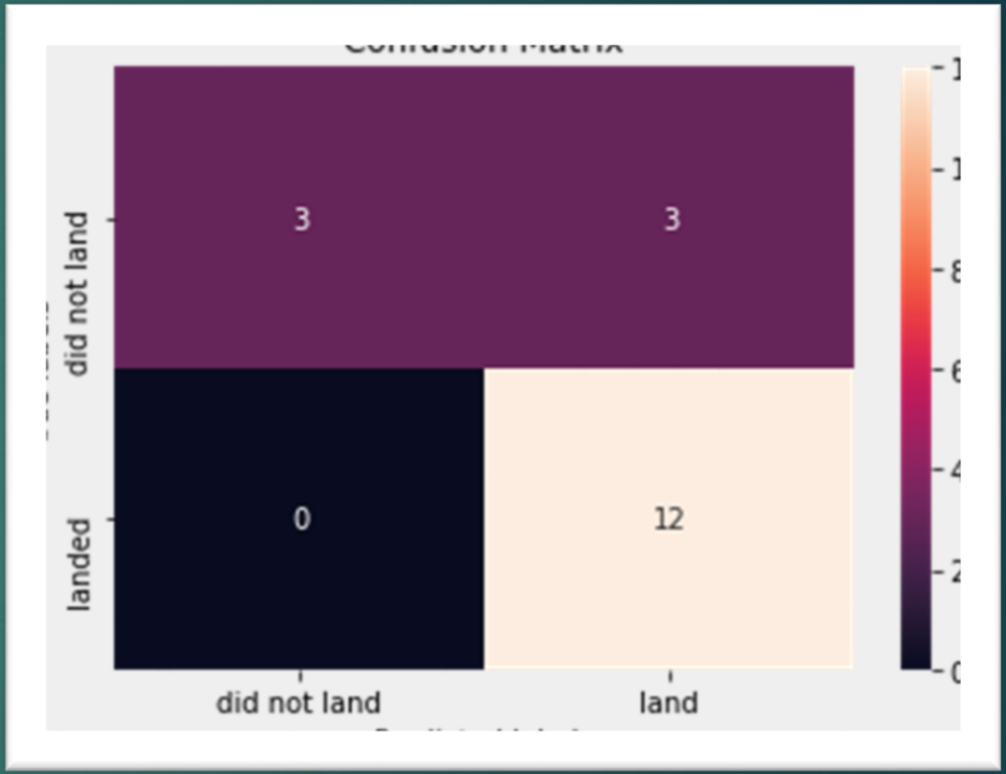
While all other models had virtually the same accuracy on the test set at 83.33% accuracy.



CONFUSION MATRIX

This confusion matrix is the decision tree confusion matrix. It shows that:

- The model predicted 12 successful landings when the true label was successful landing.
- The model predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.



CONCLUSIONS

CONCLUSIONS

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page

PERFORMED EDA WITH VISUALISATION,SQL, FOLIUM AND PREDICTIVE ANALYSIS WHICH PRODUCED THE FOLLOWING RESULTS

- As the number of flights increases, the rate of success at a launch site increases, with most early flights being unsuccessful. I.e. with more experience, the success rate increases.
  - Between 2010 and 2013, all landings were unsuccessful (as the success rate is 0).
  - After 2013, the success rate generally increased, despite small dips in 2018 and 2020.
  - After 2016, there was always a greater than 50% chance of success.
- Orbit types ES-L1, GEO, HEO, and SSO, have the highest (100%) success rate.
  - The 100% success rate of GEO, HEO, and ES-L1 orbits can be explained by only having 1 flight into the respective orbits.
  - The 100% success rate in SSO is more impressive, with 5 successful flights.
  - The orbit types PO, ISS, and LEO, have more success with heavy payloads:
  - VLEO (Very Low Earth Orbit) launches are associated with heavier payloads, which makes intuitive sense.
- The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches, and also the highest rate of successful launches, with a 76.9% success rate.
- The success for massive payloads (over 4000kg) is lower than that for low payloads.
- The best performing classification model is the Decision Tree model, with an accuracy of 94.44%.

APPENDIX

A GitHub repository was created that stores the codes used in this project. It shows the steps from data collection to predictive analysis.

GitHub repository url:

https://github.com/Cathbert-Busiku/Data-science-projects

Instructors:

Instructors: Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo