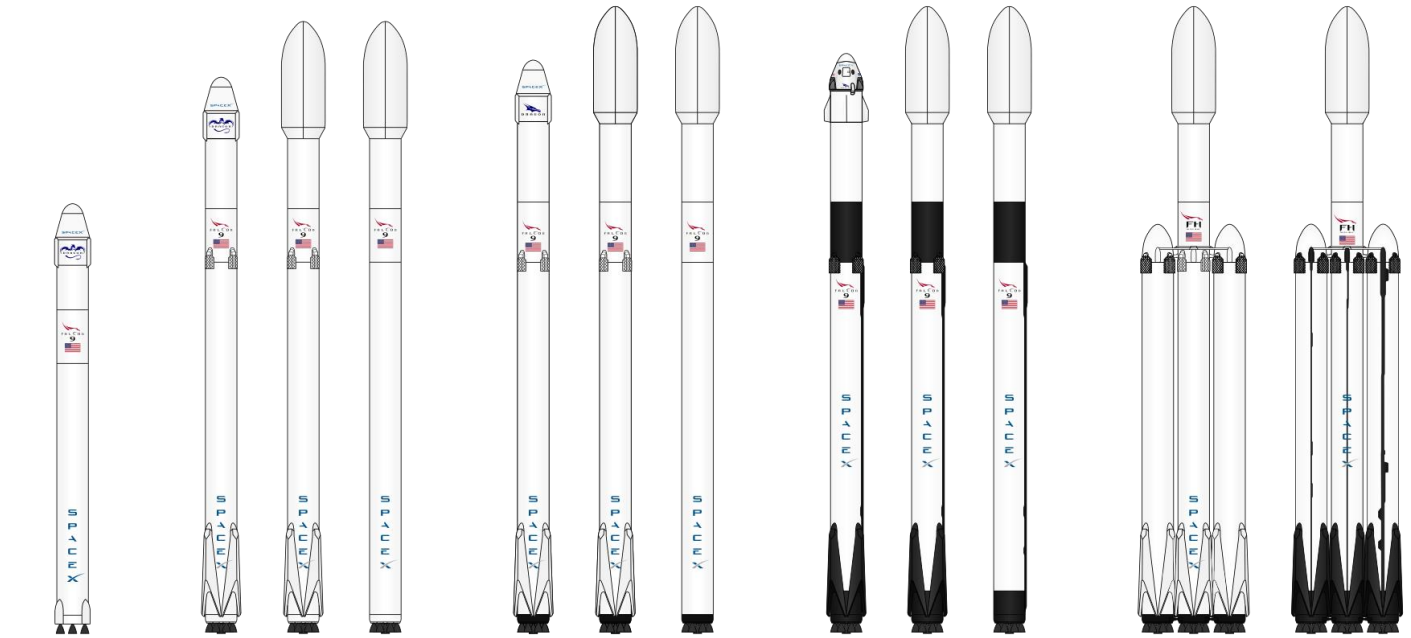

IBM Developer
SKILLS NETWORK
coursera
Winning Space Race
with Data Science
Natthapong Sueviriyapan
22 October 2022

Outline

IBM Data Science Capstone Project: Space X Falcon 9 Landing Analysis and Prediction

1. **Executive Summary**
2. **Introduction:** Overview and background
3. **Methodology**
4. **Results**
   - Exploratory data analysis results
   - Interactive visual analytics
   - Predictive analysis using classification models
5. **Conclusion and innovative insights**
6. **Acknowledgements**
7. **Appendix**



SPACEX

1. Executive Summary

1.) Executive Summary (1/2)

❑ Project Scenario Overview

- **SpaceY (by Allon Mask)** is a new commercial rocket launch provider who **wants to bid against SpaceX.**
  - SpaceX (by Elon Musk) promoted Falcon 9 rocket launches for a cost of 62 million dollars when the first stage of their rockets can be reused.
  - Also, SpaceX public statements indicate a 1st stage Falcon 9 booster to cost upwards of $15 million to build without including R&D cost recoupment or profit margin.

❑ Scope and Objective

- To **analyze** what **key factors** can **influence successful launches/landings**
- To **find** the **best way** to **estimate** the **total cost for launches by predicting successful landings of the first stage of rockets**
- To **evaluate** the **viability of the new company Space Y to compete with Space X.**

SPACE Y

Allon Mask

Photo based on Photo by **Britta Pedersen-Pool/Getty Images**

1.) Executive Summary (2/2)

❑ The methodologies used in the capstone project

- **Data Collection** (SPACEX API and Web Scraping)
- **Data Wrangling**
- **Exploratory Data Analysis** (EDA) using Visualization + SQL
- **Interactive Visual Analytics**: Folium maps + Dashboard using Plotly
- **Predictive Analysis** using Supervised Machine Learning
    - Logistic regression, Support Vector Machine (SVM), Decision Tree classifier, and K nearest neighbors (KNN)

❑ Summary of all results

- Given mission parameters (such as payload mass, flight numbers, launch location, and orbit types), **the models produced in this report were able to predict the first-stage rocket booster landing successfully**
    - **EDA** allowed to **identify which features at the best to predict success of launching** from training supervised models
    - **Machine learning** predictions using different techniques showed the best **model** to **predict which characteristics are important to drive this opportunity** by the best way, using all collected data.
        - **All** the **models** produced similar results with **accuracy rate of 83.33%**. Yet, a better model determination and accuracy can be achieved with more data
- As a result, **SpaceY will be able to make more informed bids against SpaceX by using 1st stage landing predictions as a proxy for the cost of a launch.**

2. Introduction: Overview and Background

2.) Introduction (1/2)

Project background and context

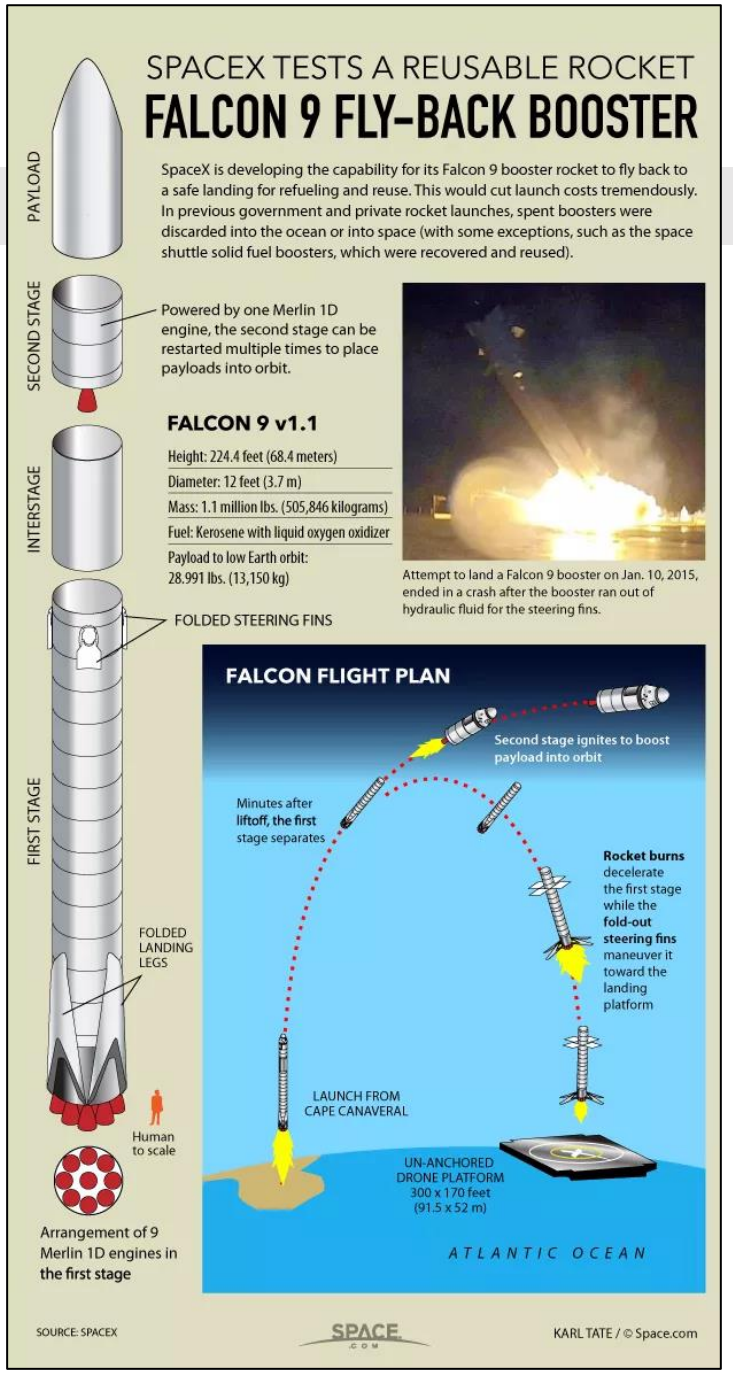
- The **main goal** of this study was **to forecast where the Falcon 9's first stage will land.**
  - According to SpaceX, the cost of launching a Falcon 9 rocket was $62 million.
  - The expense of other providers is greater than $165 million.
  - It is possible for SpaceX to reuse the first stage, which results in cost savings.
  - The launch costs are determined by whether or not the first stage successfully lands.

- The **company** that is **in competition with SpaceX for a rocket launch** (i.e., **Space Y by Allon Mask**) **can utilize this information to their advantage**.
  - Using publicly available data and machine learning, several machine learning models were built for Space Y to compete against SpaceX using data from a public SpaceX API and web scraping SpaceX's Wikipedia page.
  - The ability to predict when Stage 1 will land will save tens of millions of dollars.
  - This will allow SpaceY to make better informed bids against SpaceX because they will know when to expect SpaceX to include the cost of a sacrificed first stage rocket.

2.) Introduction (2/2)

The prior Falcon 9 rocket launch data was employed to forecast the likelihood of the booster landing back on the pad, which is influenced/correlated with several factors/conditions.

- **What** are the **primary criteria for** a **successful** or **unsuccessful landing**?
- **How** do **variables** like payload mass, launch site, number of flights, and orbits **affect first-stage landing success**?
- What **other conditions** will allow SpaceX **to attain the highest landing success rate**?
- Is the **rate of successful landings increasing over time?**
- What is the **optimal algorithm for binary classification** that can develop the model **to predict landing outcomes**?



Source: https://www.space.com/28167-spacex-risky-reusable-rocket-landing-infographic.html

3. Methodology

3.) Methodology (1/2)

Data collection → Data wrangling → Exploratory data analysis → Interactive visual analytics → Predictive analysis

3.1 Collect data

- **Combining Space X data from 2 sources/ways**: **A)** SPACEX API and **B)** Web Scrapping
- **Gathering information** on targeted variables

3.2 Perform data wrangling

- Filtering the data/Dealing with missing values/Dropping irrelevant columns/Using One Hot Encoding to prepare the data to a binary classification (Transforming data for machine learning)
  - **Classifying a landing outcome label** as successful and unsuccessful otherwise

3.3 Perform exploratory data analysis (EDA)

- **A) Using Pandas and Matplotlib** for the relationship between variables and to obtain some preliminary insights about how each important variable would affect the success rate (**Scatter and bar graphs to show patterns between data**)
- **B) Using SQL queries** to manipulate and evaluate the SPACE X dataset

3.) Methodology (2/2)

Data collection → Data wrangling → Exploratory data analysis → **Interactive visual analytics → Predictive analysis**

3.4 Perform interactive visual analytics

- A) Performing **launch sites locations analysis with Folium**
- B) Creating an **interactive dashboard using Plotly Dash**

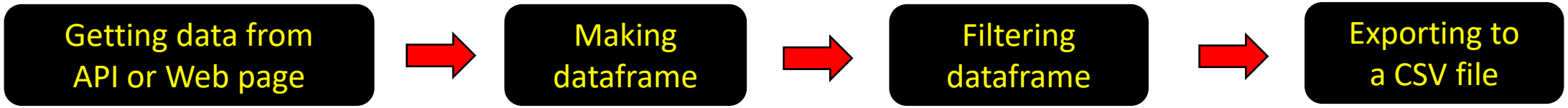
3.5 Perform predictive analysis using classification models

- **Building**, **tuning**, and **evaluating** of **classification models using Scikit-Learn** to:
  - Pre-process (normalize) the data
  - Split the data into training and testing data using train_test_split
  - Train four different classification models
  - Find hyperparameters using GridSearchCV
- **Assessing** the **accuracy of each classification model** to ensure the best results
- **Plotting confusion matrices** for each classification model

3.1) Data Collection

SPACE API & Web scrapping

Genialized data collection steps



Data from Space X was combined from 2 sources in order to get complete information about the launches for a more detailed analysis.

- **A) Making a get request to the open-source SpaceX API (https://api.spacexdata.com/v4/rockets/)** to gain historical launch data
  - Extractingthe data in the form of JSON and transforming it to a dataframe using built-in python pandas method.
  - Space X API data Columns such as Flight No., Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Custormer, Landing Pad, Reused Count, Serial, Longitude, Latitude

- **B) Web Scrapping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)** to gain historical launch data
  - Web scraping SpaceX launches from a table in Space X's Wikipedia entry and converting it into a dataframe
  - Wikipedia Web-scraped data Columns such as Flight No., Launch site, Payload Mass, Orbit, Launch outcome, Booster Version, Booster landing, Date

3.1A) Data Collection – SpaceX API

Goal: To acquire historical launch data from open-source API for SpaceX in the proper, useful format

Task 1. Requesting and parsing the SpaceX launch data

- Imported associated libraries and defined auxiliary functions to help manage data
- Perfomed the GET request method to obtain rocket launch data from SpaceX API
- Decoded the response to JSON file + Turned it into a Pandas dataframe

Task 2. Making and Filtering dataframe (only include Falcon 9 launches)

- Called the defined custom functions to get/filter data about the launches
- Created a dictionary with the collected dataset and transformed it to a dataframe
- Removed the unwanted Falcon launches keeping only the Falcon 9 launches.

Task 3. Pre-wrangling and exporting data

- Replaced missing payload mass values from classified missions with mean
- Exported the ready dataframe to a CSV file

[Github URL: Data_Collection_SpaceX_API]

3.1B) Data Collection – Web Scraping

Goal: To further acquire historical Falcon 9 launch data (HTML table) from Wikipedia in the proper, useful format

Task 1. Requesting the Falcon9 Launch Wiki page from its URL

- Performed an HTTP GET method to request the Falcon9 Launch HTML page
- Created a BeautifulSoup object from the HTML response

Task 2. Processing web scraped HTML table and Collecting data

- Searched for all tables within the HTML page
- Extracted all relevant column/variable names from the HTML table header

Task 3. Creating and exporting the dataframe

- Created a dictionary to be be converted into a Pandas dataframe
- Created a dataframe by parsing the launch HTML tables with custom functions
- Exported the ready dataframe to a CSV file

GitHub URL: Data_Collection_Web_Scrapping_Wiki

3.2) Data Wrangling

Goal: To mainly perform basic EDA on cleaned data and determine training labels (Boolean values)

Task 1. Calculating the number of launches on each site

- Imported associated libraries and defined auxiliary functions to help manage data
- Analyzed the collected SpaceX dataset (e.g., missing values or type of columns)
- Determined the number of launches on each site

Task 2. Calculating the number and occurrence of each orbit

- Determined the number and occurrence of each orbit

Task 3. Calculating the number and occurence of mission outcome per orbit type

- Determined the number of landing mission outcomes per orbit type

Task 4. Creating a landing outcome label and Exporting the data

- Assigned binary classification variable that represents the two groups of outcome of each launch (1= Successful landing; 0= failed landing)
- Exported the udpated dataframe to a CSV file

[Github URL: Data_Wrangling]

3.3A) EDA with Data Visualization

Goal: To Perform detailed EDA and Feature Engineering using Pandas and Matplotlib

Summary of the chart types of EDA-based visualization performed on multiple variables to get some preliminary insights about how each important variable would affect the success rate (Features could be used for training the machine learning model and prediction)

Scatter charts were created to depict the c orrelations between

- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Flight Number vs. Orbit Type
- Payload vs. Orbit Type

Why: This sort of chart is excellent for observing relationships, or correlations, between two numerical variables. Display the dependency of each selected pair of attributes to predict which aspects will result in the highest probability of success in landing outcomes.

A bar chart was created to depict the association between

- Success Rate vs. Orbit Type

Why: It is helpful to compare a numerical value to a category variable. Depending on the amount of data, either horizontal or vertical bar charts can be employed. Also, it is easy to discover which orbits have the best chances of success.

A line chart was created to represent the links between:

- Success Rate and Year (launch success yearly trend 2010-2020)

Why: Line chart, which has numerical values on both axes, is commonly used to depict the change of a variable over time. Furthermore, it may benefit in future prediction.

[Github URL: EDA_with_Visualization]

3.3B) EDA with SQL

Goal: To further analyze and gain useful insights from the data set

Summary of the SQL queries performed

1. Displayed the **names** of the **unique launch sites** in the space mission
2. Displayed **5 records** where **launch sites begin with** the string '**CCA**'
3. Displayed the **total payload mass** carried by **boosters launched by NASA** (CRS)
4. Displayed the **average payload mass** carried **by booster version F9 v1.1**
5. Listed the **date** when the first **successful landing outcome on a ground pad** was achieved
6. Listed the **names of the boosters** which had **success on a drone ship** and a **payload mass between 4000 and 6000 kg**
7. Listed the **total number of successful** and **failed mission outcomes**
8. Listed the **names of the booster versions** which have carried the **maximum payload mass**
9. Listed the records displaying display the **month names**, **failure landing outcomes in drone ship ,booster versions, launch site** for the months in year **2015**.
10. **Ranked** the count of **landing outcomes between the date 04-06-2010 and 20-03-2017** in descending order

Github URL: EDA_with_SQL

3.4A) Build an Interactive Map with Folium

Goal: To analyze the existing launch site locations and geospatial proximities that may influence the launch success rate

- The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.
- Finding an optimal location for building a launch site certainly involves many factors. To visualize the launch data on an interactive map, the following steps were conducted.

In this project, a Folium map object was created with an initial center location to be NASA Johnson Space Center in Houston, Texas.

TASK 1: Marking all launch sites on a map

- Added marker with colored circles (showing highlighted areas), popup /text Labels (e.g., showing names) of
  - NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - All Launch Sites using their latitude and longitude coordinates to display their geographical locations and proximities (such as Railway, Highway, Coast, and City) to Equator lines and coastlines.

TASK 2: Marking the successful/failed launches for each site on the map

- Added colored Markers of all launch records: success (Green/class =1) and failed (Red class =0) launches using Marker Cluster (a good way to simplify a map containing many markers having the same coordinate) to identify which launch sites have relatively high success rates from the color-labeled markers.

TASK 3: Calculating the distances between a launch site to its proximities

- Added colored Lines to show distances between two points on the map based on their coordinates (Latitude and Longitude), such as the distance between the key Launch Site and its proximities

Github URL: Launch_Site_Location_with_Folium

3.4B) Build a Dashboard with Plotly Dash

Goal: To quickly explore and manipulate data in an interactive and real-time manner

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

- **Pie chart** was used to visualize the launch site success rate by showing the total successful launches per site
  - The chart (colored coded by launch site) could also be filtered using drop-down menu to see the success/failure ratio for an individual site of interest.

- **Scatter chart** was used to show the 2-D correlation between landing mission outcome (success or not) and payload mass (kg).
  - The chart could also be filtered by a fixed range (0-10000 kg) of payload masses or booster version (color-coded).

[Github URL: Spacex_Dash_App_with_Plotly]

3.5) Predictive Analysis (Classification)

Goal: To develop, evaluate, and find the best performing classification models for the optimal prediction

Task 1. Preparing the dataset

- Imported associated libraries and defined auxiliary functions to help manage data
- Loaded the featured engineered dataset, standardized, and transformed the data
- Split the data into training data (80%) and test data (20%)

Task 2. Building the machine-learning model

- Created a GridSearchCV object and a dictionary of parameters
- Fitted the training data to the four selected algorithms (1 logistic regression, 2 support vector machine, 3 decision tree and 4 kKnearest neighbors)

Task 3. Evaluating the four distinct machine-learning models

- Checked the tuned hyperparameters using the output GridSearchCV object
- Checked the accuracy (on the test data) using the output GridSearchCV object
- Plotted and examined the Confusion Matrices

Task 4. Comparing and finding the best classification model

- Reviewed the accuracy results for four classification models to find the best one

[Github URL: SpaceX_Machine Learning Prediction]

4. Results

4.1 Exploratory data analysis results

4.1.1 With visualization

4.1.2 With SQL

4.2 Interactive analytics demo in screenshots

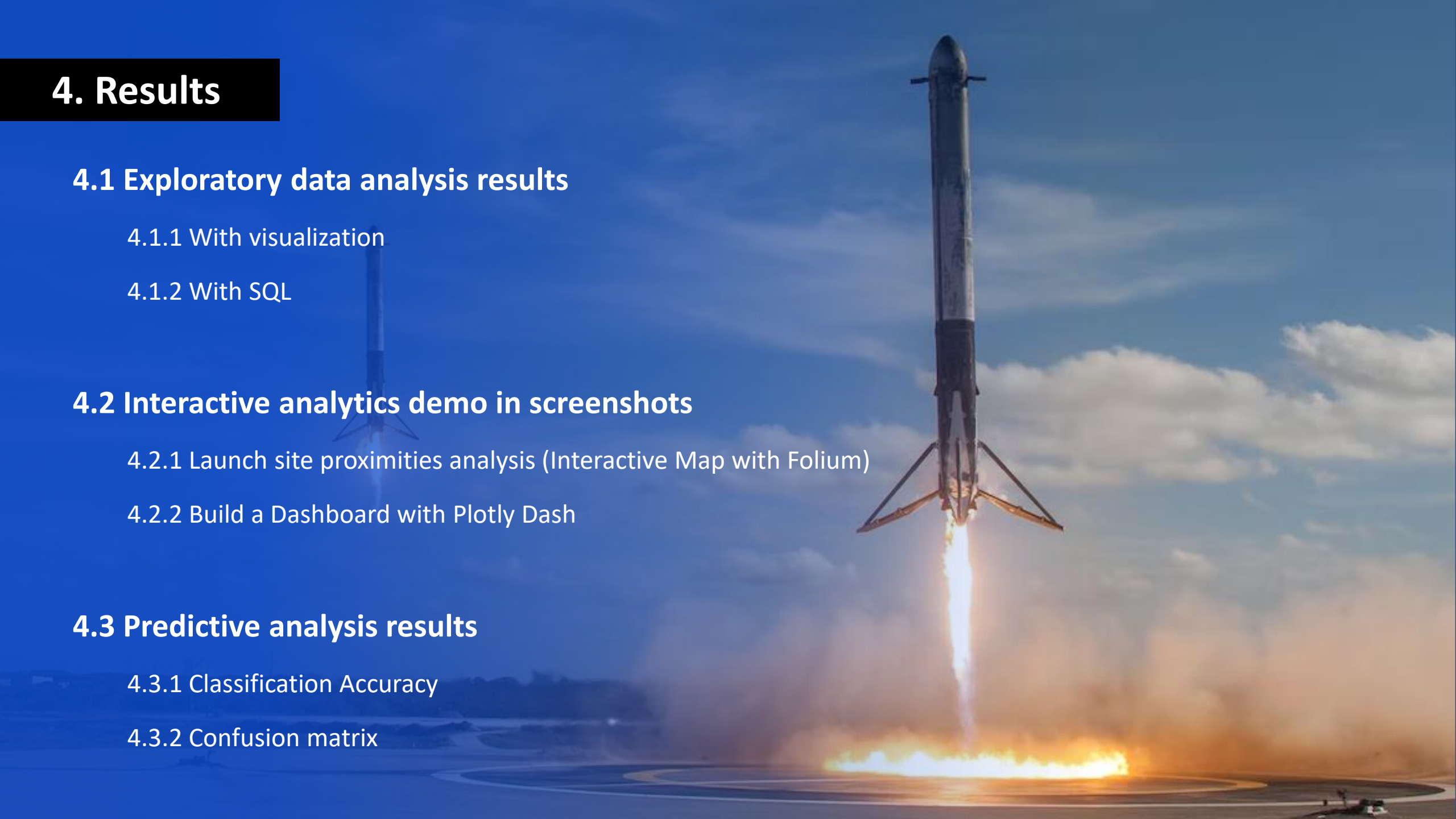
4.2.1 Launch site proximities analysis (Interactive Map with Folium)

4.2.2 Build a Dashboard with Plotly Dash

4.3 Predictive analysis results

4.3.1 Classification Accuracy

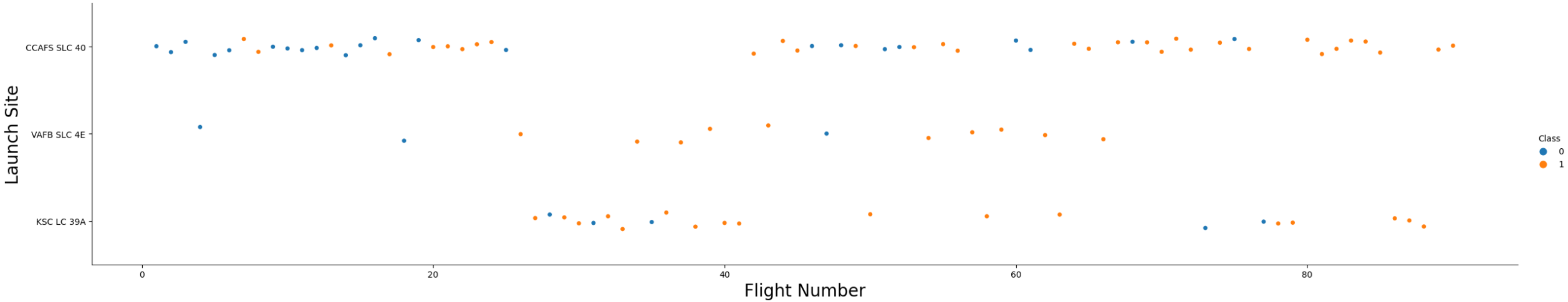
4.3.2 Confusion matrix



4.1.1 EDA - WITH VISUALIZATION

4.1.1.A) Flight Number vs. Launch Site
4.1.1.B) Payload vs. Launch Site
4.1.1.C) Success rate vs. Orbit type
4.1.1.D) Flight Number vs. Orbit type
4.1.1.E) Payload vs. Orbit type
4.1.1.F) Launch Success Yearly Trend

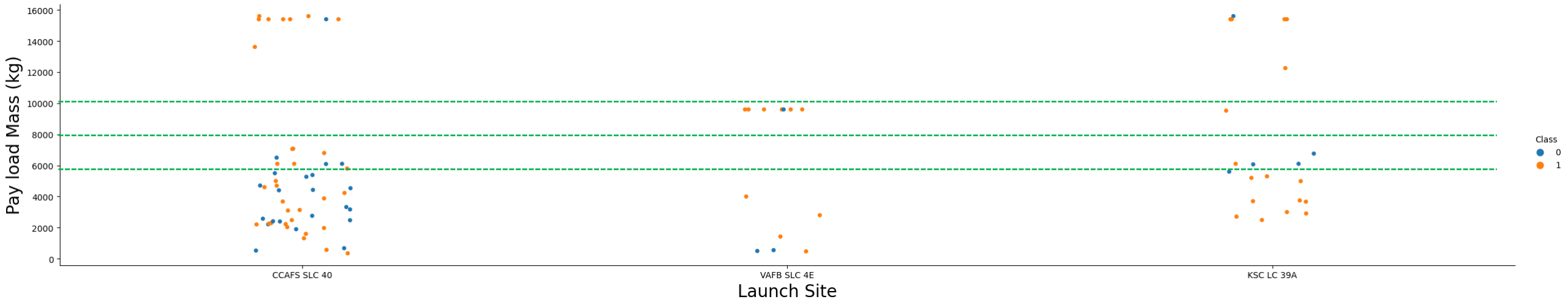
4.1.1A) Flight Number vs. Launch Site



Explanation:

- The first flights were generally unsuccessful, but **after about 25 flight**s, there are **far more successful landings** (Class = 1), with the **CCAFS SLC 40 launch site accounting for roughly half of all launches**.
- There are **no early flights from KSC LC 39A**; nonetheless, launches from this location are more successful.
- The **success rate for each site** appears to be **growing as the number of flights increases.**

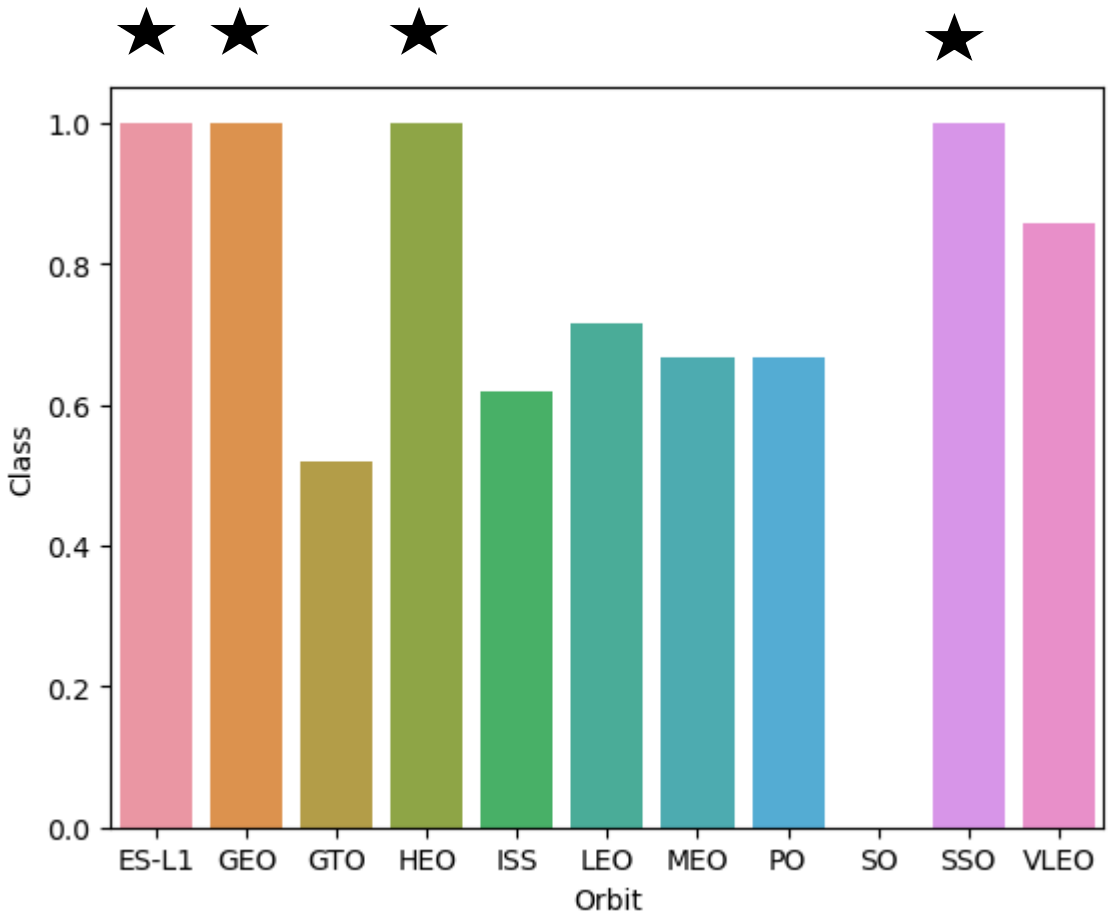
4.1.1B) Payload vs. Launch Site



Explanation:

- **Different launch sites appear to use varying payload masses,** with the majority of launches from CCAFS SLC 40 utilizing comparatively lighter payloads (with some outliers).
- The majority of **launches weighing more than 8000 kg** were **successful.**
- Notably, KSC LC-39A has a **100% success rate** for **payloads weighing less than 6000 kg.**
- There are **no rockets** launched **for heavy-payload mass (greater than 10000) at** the **VAFB-SLC** launch site.
- Overall, there is **no evident relationship between payload mass and launch site success rate**.

4.1.1C) Success Rate vs. Orbit Type

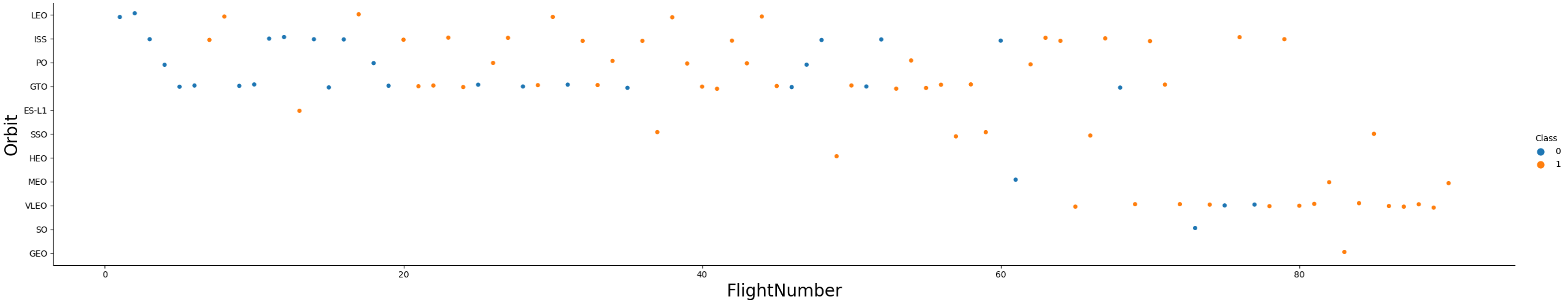


Explanation:

- The following orbits have the best success rate (class 1.0 = **100% marked with a black star**):
  - ES-L1 (Earth-Sun First Lagrangian Point)
  - GEO (Geostationary Orbit)
  - HEO (High Earth Orbit)
  - SSO (Sun-synchronous Orbit)
- **In contrast, SO (Heliocentric Orbit) is the o**rbit with the lowest (0%) success rate

More information and short description about orbit type can be found in Github URL: Data Wrangling

4.1.1D) Flight Number vs. Orbit Type

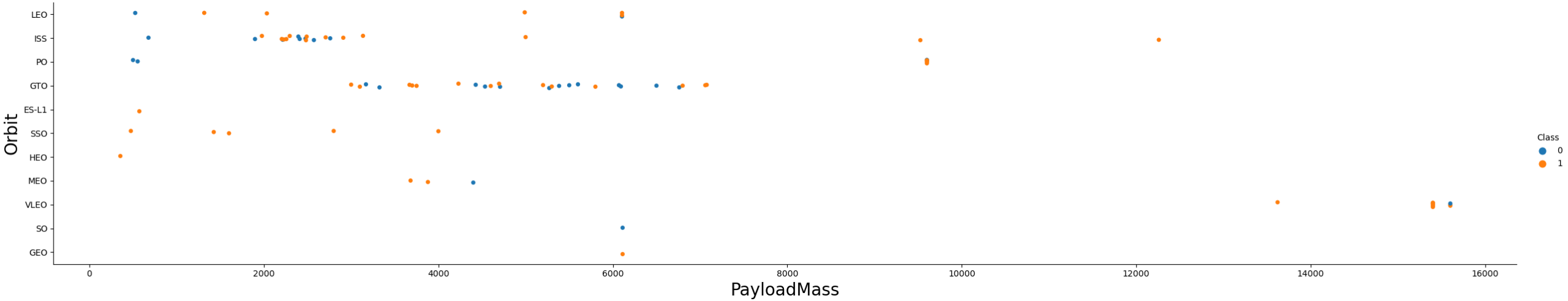


Explanation:

- **In general, as the number of flights grows, so does the success rate** (Class = 1), but there is little correlation between flight number and success rate for GTO.
- Noticeably, the **100% success record of GEO, HEO, ES-L1, and SSO** orbits may be explained by having only 1-5 missions in each orbit.
- SpaceX **began with LEO orbits** and had **moderate success**.
- In recent launches, **LEO has been replaced by VLEO**. SpaceX appears to fare better in lower orbits or those that are Sun-synchronous. **Because of the recent rise in frequency**, VLEO orbit appears to be a new economic prospect.

More information and short description about orbit type can be found in Github URL: Data Wrangling

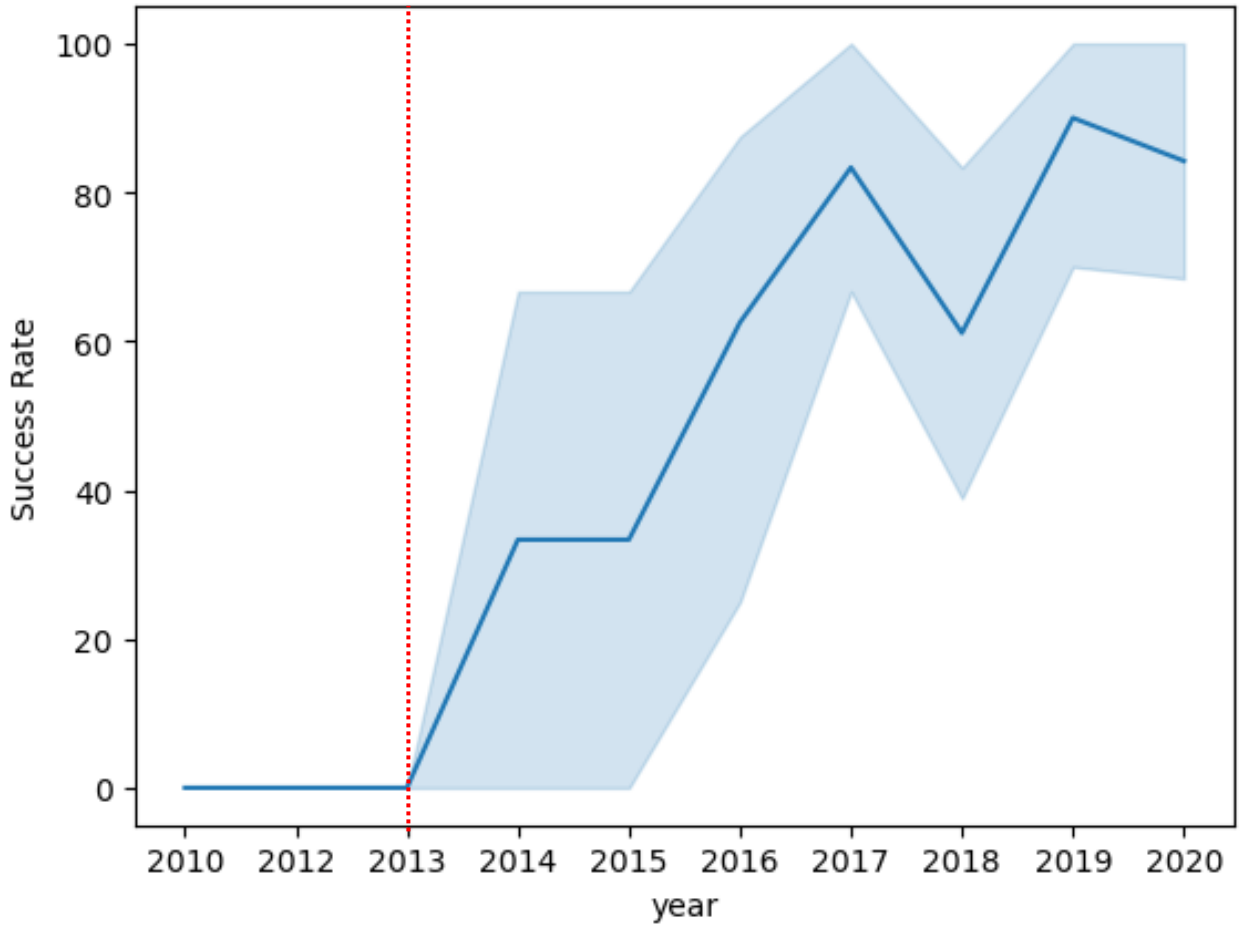
4.1.1E) Payload vs. Orbit Type



Explanation:

- **With heavier payloads**, the **successful landing rate** (Class =1) is **higher for PO,LEO, and ISS.**
- The other **most successful VLEO and SSO** orbits only have **payload mass values in** the **higher and lower end of the range, respectively**.
- For **GTO**, the **correlation between payload mass** and **success rate** is **ambiguous**.
- There are **a few launches to the orbits SO, GEO, and HEO.**

4.1.1F) Launch Success Yearly Trend



Explanations:

- Since 2013, success has typically increased, with a tiny fall in 2018 and 2020.
- There was always a **better than 50% likelihood of success after 2016**, while success has been **about 80% in recent years**

4.1.2 EDA - WITH SQL

4.1.2.A) All Launch Site Names
4.1.2.B) Launch Site Names Beginning with `CCA`
4.1.2.C) Total Payload Mass
4.1.2.D) Average Payload Mass by F9 v1.1
4.1.2.E) First Successful Ground Landing Date
4.1.2.F) Successful Drone Ship Landing with Payload Between 4000 and 6000
4.1.2 G) Total Number of Successful and Failure Mission Outcomes
4.1.2 H) Boosters that Carried Maximum Payload
4.1.2 I) 2015 Launch Records
4.1.2 J) Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

4.1.2A) All Launch Site Names

SQL Query

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL;
```

Result

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

Explanation:

- The use of **DISTINCT** in the query enabled for the removal of duplicate **LAUNCH SITE** entries, yielding only unique values from the **SPACEXTBL** table's LAUNCH SITE column.

4.1.2 B) Launch Site Names Begin with 'CCA'

SQL Query

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

Result

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation:

- The **LIMIT 5** keyword retrieved **just 5 records**, whereas the **LIKE** keyword with the wild card 'CCA%' retrieved string values **beginning with 'CCA'.**
- Thus, **Five records of Cape Canaveral launches** are shown above.

4.1.2C) Total Payload Mass

SQL Query

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

Result

sum
45596

Explanation:

- The **SUM** keyword was used to compute the **overall payload mass in kilograms**, **where NASA (CRS) is the customer.**
  - CRS stands for Commercial Resupply Services, indicating that these payloads were delivered to the International Space Station (ISS).

4.1.2D) Average Payload Mass by F9 v1.1

SQL Query

```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1'
```

Result

Average
2928.4

Explanation:

- The **AVG** keyword computed the **average of the PAYLOAD MASS KG_ column**, while the **WHERE** keyword restricted the results to **only the F9 v1.1 booster version.**

4.1.2E) First Successful Ground Landing Date

SQL Query

```
%sql select date as date from SPACEXTBL where "landing _outcome" like 'Success (ground pad)'
```

Result

date
22-12-2015
18-07-2016
19-02-2017
01-05-2017
03-06-2017
14-08-2017
07-09-2017
15-12-2017
08-01-2018

Explanation:

- The **WHERE** keyword restricted the results to only **successful ground pad landings.**
- This query provided the dates of all successful ground pad landings.
  - So, **the first successful ground pad landing is 22 December 2015.**

4.1.2F) Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query

```
%sql select booster_version from SPACEXTBL where  (payload_mass__kg_ BETWEEN 4000 AND 6000) AND ("landing _outcome" like 'Success (drone ship)')
```

Result

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

Explanation:

- The **WHERE** keyword was used to **limit the results to those that meet both conditions** (as the **AND** keyword was also used). The **BETWEEN** keyword allows you to choose **between 4000 x 6000 values and landing on a drone ship.**
- This query returned the **four booster types with successful drone ship landings** and payload masses ranging **from 4000 to 6000.**

4.1.2G) Total Number of Successful and Failure Mission Outcomes

SQL Query

```
%sql select mission_outcome, count(*) as Count from SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

Result

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

Explanation:

- The **COUNT** keyword was used to calculate the **total number of mission outcomes**
- The **GROUPBY** keyword was used to **categorize these results**.
- This query provided the total number of mission outcomes.
  - **SpaceX appears to complete its missions nearly 99% of the time**.
  - Notably, one launch had an unknown payload status, while one failed in flight.

4.1.2H) Boosters Carried Maximum Payload

SQL Query

```
%sql select booster_version, payload_mass__kg_ FROM SPACEXTBL WHERE payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
```

Result

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

Explanation:

- The **SELECT** statement within the brackets **determined the maximum payload**, which is then used in the **WHERE** condition **with another subquery.**
- This query **returned all the F9 B5 B10xx.x booster variations**

4.1.2I) 2015 Launch Records

SQL Query

```
%sql select substr(Date, 4, 2) as Month, "landing _outcome", booster_version, launch_site  from SPACEXTBL where (date like '%2015')  AND ("landing _outcome"  like 'Failure (drone ship)')
```

Result

Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Explanation:

- The **WHERE** keyword was used to narrow down the results **to only failed landing outcomes AND** only for **2015.**
- Substr function processes a date to determine whether it is a month or a year.
  - **Substr(DATE, 4, 2) displays the month.**
- This query gave the month, booster version, launch site where the landing was failed in 2015.

4.1.2J) Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

SQL Query

```
%sql select date, count("landing _outcome") as Count from SPACEXTBL  where Date >= '04-06-2010' AND Date <= '20-03-2017' AND "landing _outcome" LIKE '%Success%' Group by "landing _outcome" \
    Order by count("landing _outcome") Desc
```

Result

Date	Count
07-08-2018	20
08-04-2016	8
18-07-2016	6

Explanation:

- The **WHERE** keyword was combined with the **BETWEEN** keyword to limit the results to dates that fall within the range supplied.
- The results were then sorted and ordered using the keywords **GROUP BY and ORDER BY, with DESC specifying the descending order.**
- This query provided **a list of successful landings that occurred between 2010-06-04 and 2017-03-20.**

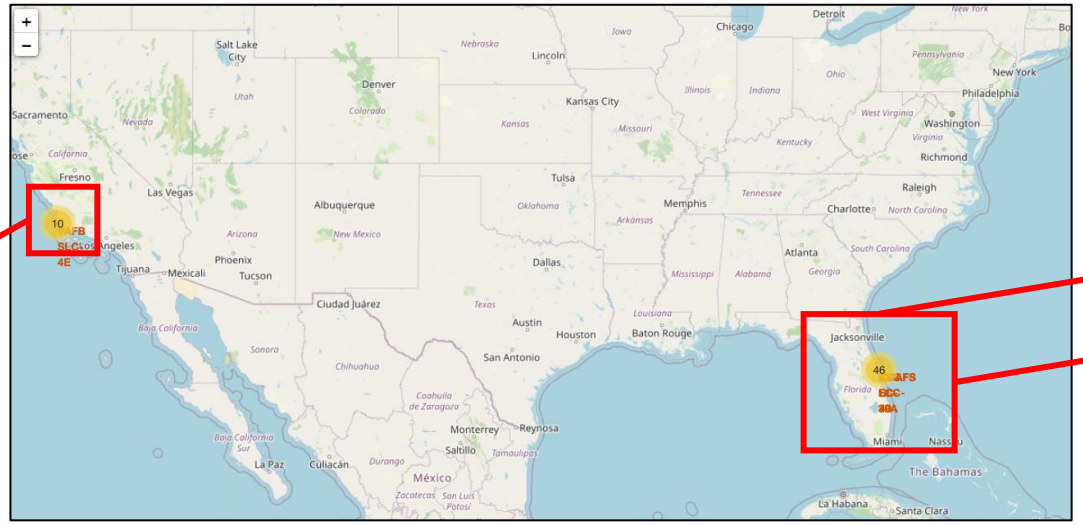
4.2.1 Launch sites proximities analysis

4.2.1A) All launch sites' location
4.2.1B) The color-labeled launch outcomes (Successful vs. Failed launches)
4.2.1C) The key proximities of the selected launch site

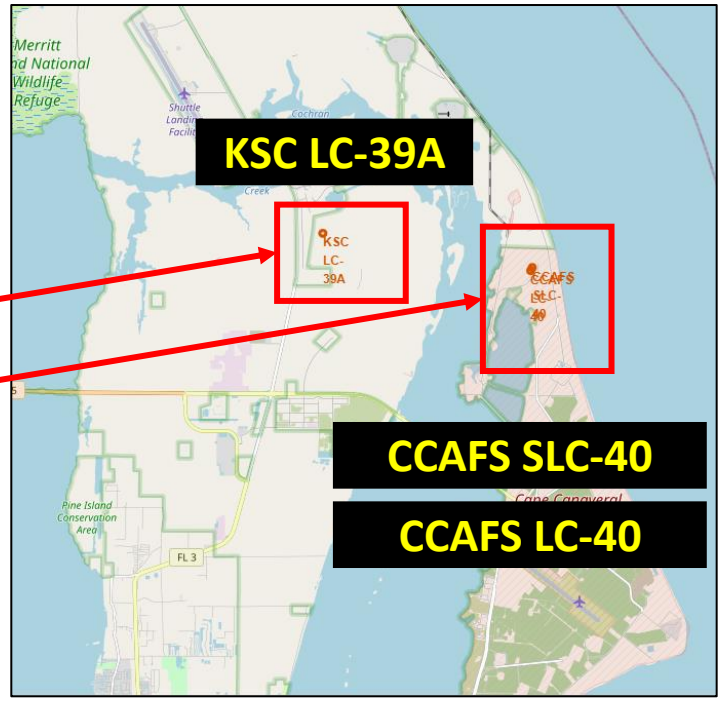
4.2.1A) All launch sites on the global map

California: 10 launch sites

The US on the global map



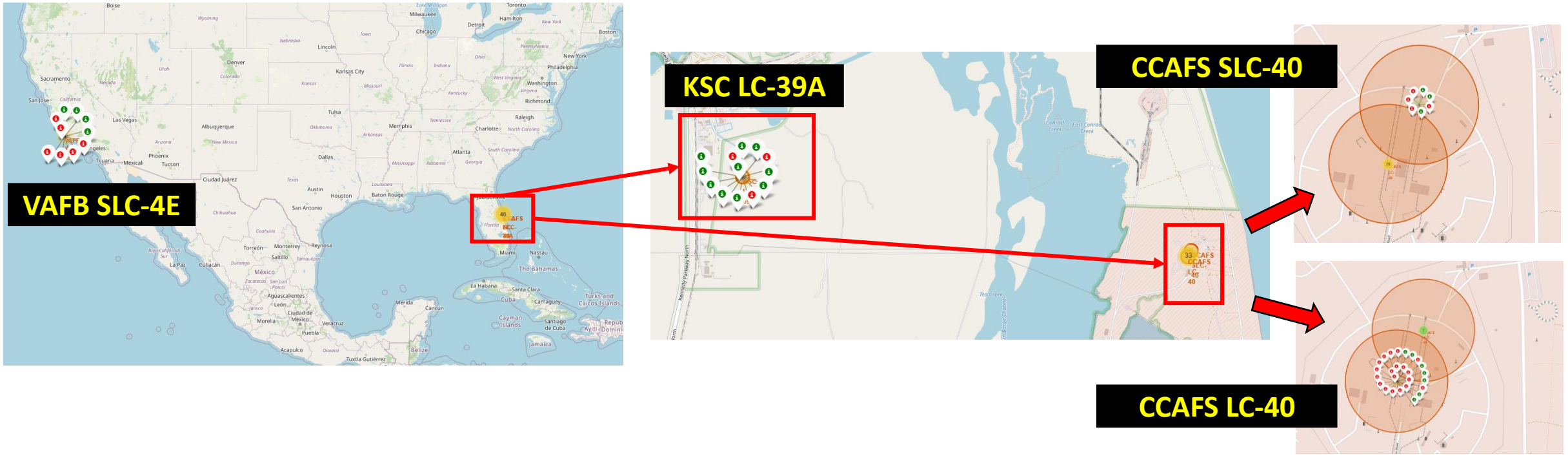
Florida: 46 launch sites



Explanation:

- **All SpaceX launch sites** are on the **coasts of the United States**, most notably in **Florida** and **California**.
- Using interactive analytics, it was possible to discover that launch sites used to be insecure regions with an excellent logistic infrastructure surrounding them.
- **To decrease the chance of debris falling or exploding near humans, all launch locations are very close to the coast.**

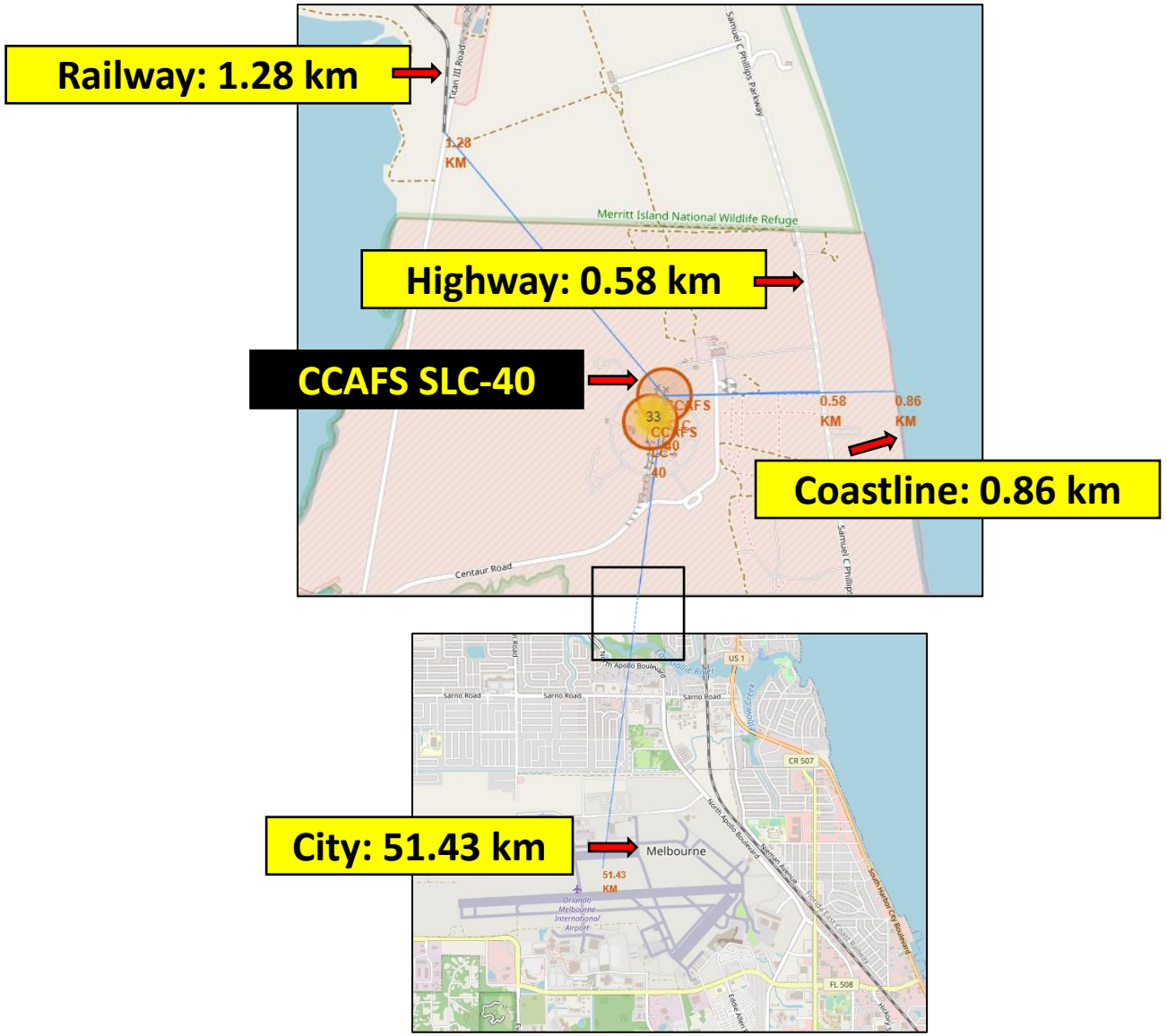
4.2.1B) The color-labeled launch outcomes on the map



Explanation:

- **Clusters on the Folium map** can be clicked on from the color-labeled markers to display each **successful landing (green icon)** and **unsuccessful landing (red icon)** .
- Visualizing the rocket landing results for each launch site reveals which launch locations, especially **KSC LC-39A**, have relatively **high success rates**.

4.2.1C) Key proximities of the selected launch site (CCAFS SLC-40)



Explanation:

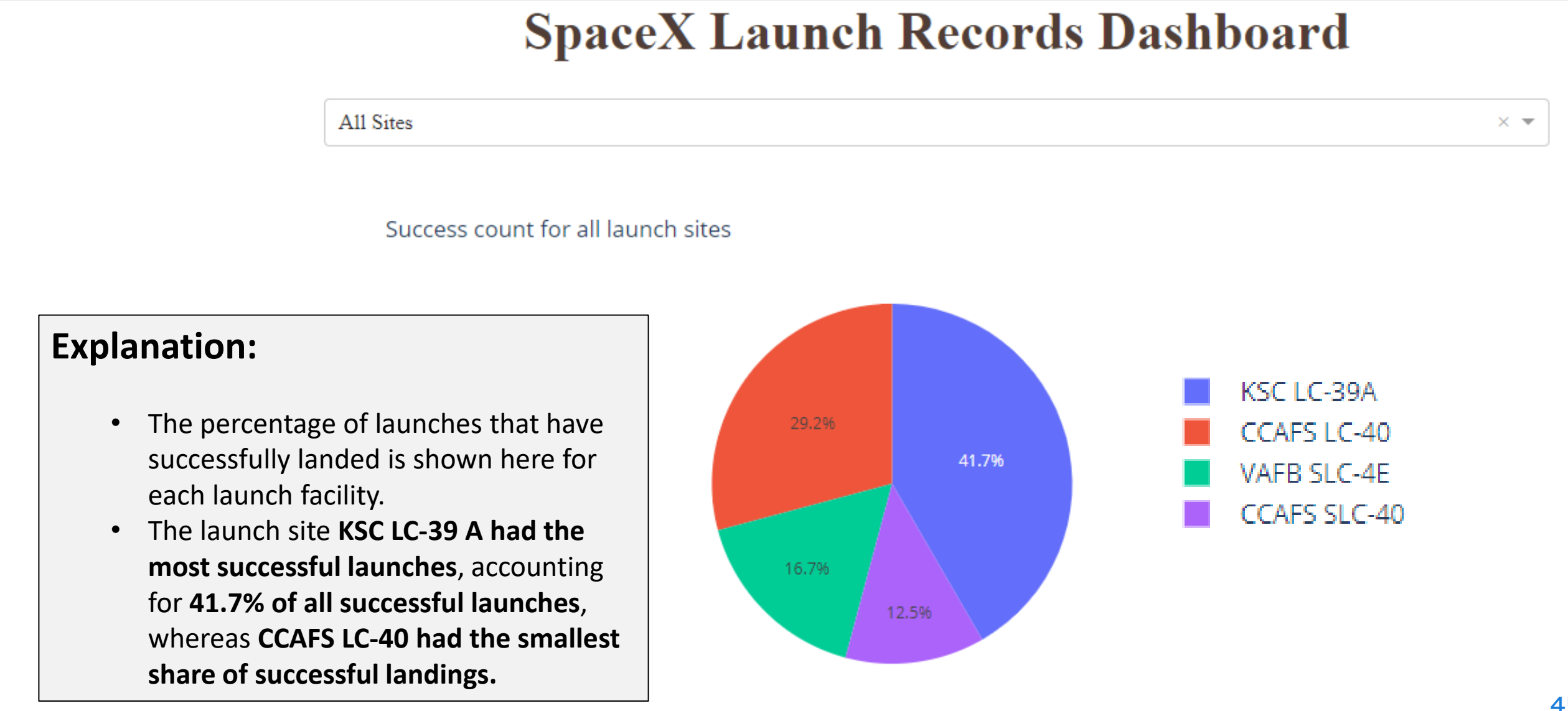
By visualizing the railway, highway, coastline, and city proximities for each launch site (**Using CCAFS SLC-40 as an example**), we can understand how close they are and the potential need for being close to such proximities :

- Transportation of heavy cargo
  - **1.28 km from the railway**
- Transportation of personnel and equipment
  - **0.58 km from the highway**
- The safety reason to abort the launch and attempt water landing, thereby reducing the risk of falling debris to densely populated areas
  - **0.86 km from the coastline**
  - **51.43 km from the city (Melbourne)**

4.2.2 Build a dashboard with Plotly Dash

4.2.2A) Launch success count for all sites
4.2.2B) The launch site with the highest launch success ratio
4.2.2C) Payload vs. Launch Outcome vs. Booster for all sites

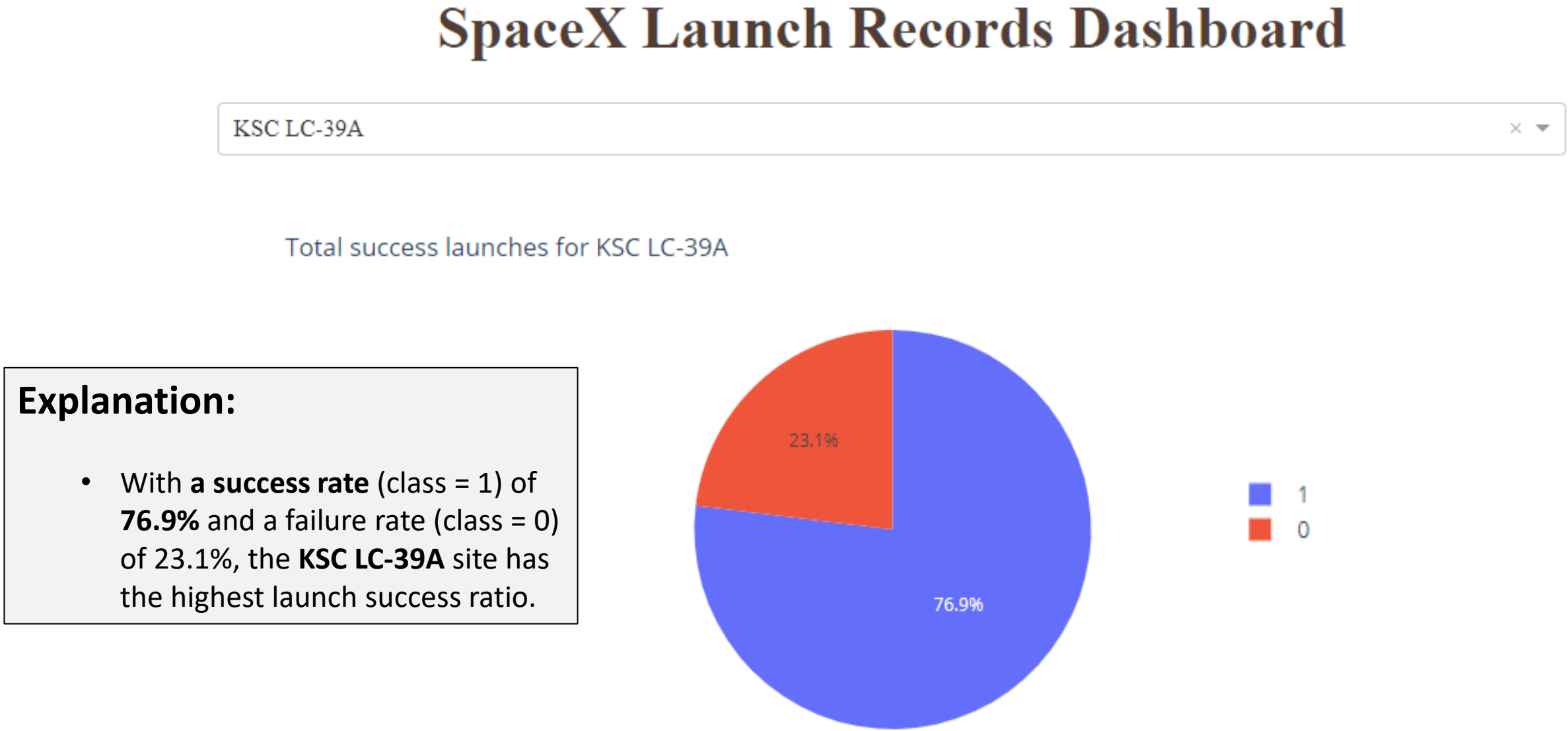
4.2.2A) Launch success count for all sites



Explanation:

- The percentage of launches that have successfully landed is shown here for each launch facility.
- The launch site **KSC LC-39 A had the most successful launches**, accounting for **41.7% of all successful launches**, whereas **CCAFS LC-40 had the smallest share of successful landings.**

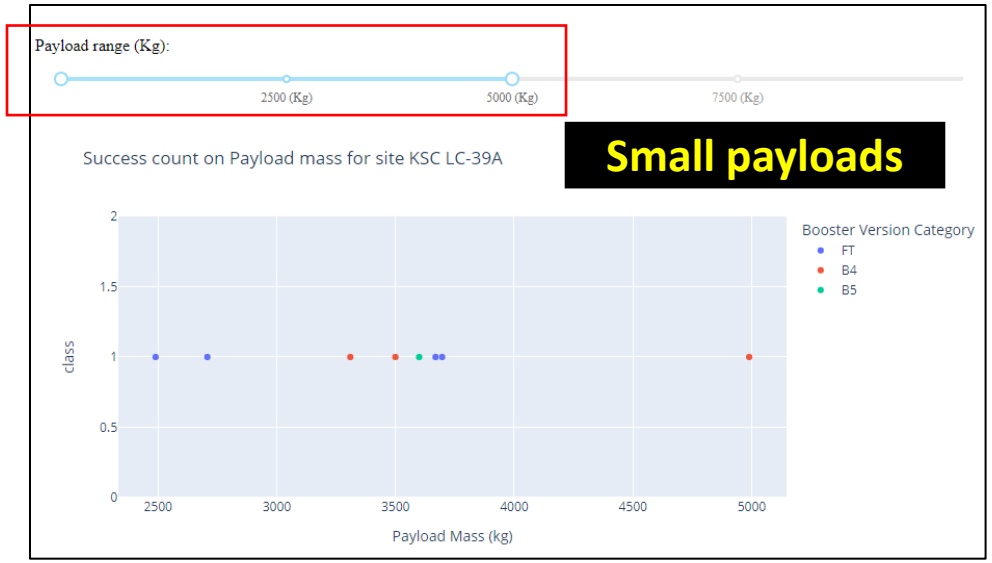
4.2.2B) The launch site with the highest launch success ratio



Explanation:

- With **a success rate** (class = 1) of **76.9%** and a failure rate (class = 0) of 23.1%, the **KSC LC-39A** site has the highest launch success ratio.

4.2.2C) Payload vs. Launch Outcome vs. Booster for all sites



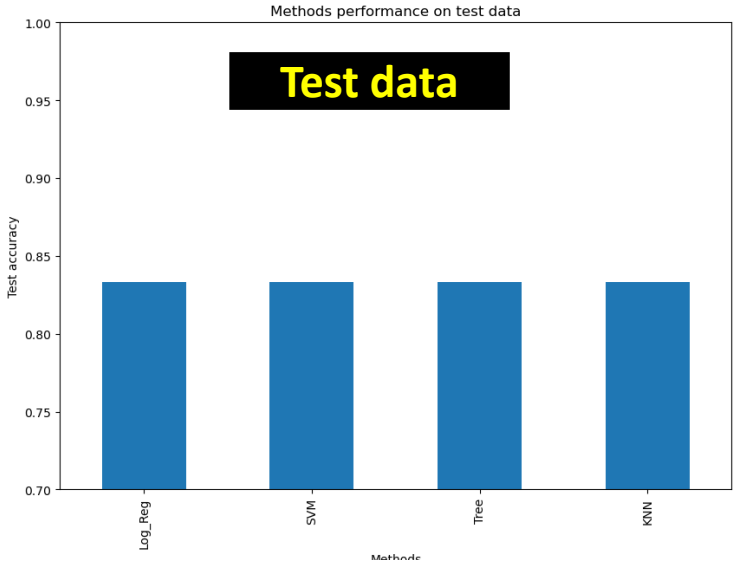
Explanation:

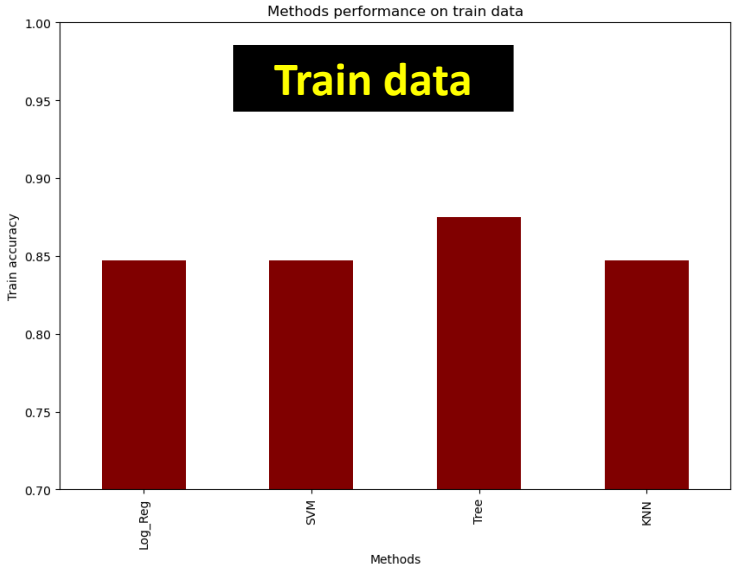
- The Payload range selection is available on the Plotly dashboard. However, instead of the maximum Payload of 15600, this was set from 0 to 10000. It is reasonable to divide the data into two ranges:
  - **0 – 5000 kg (for small payloads)**
  - **5000 – 10000 kg (massive payloads)**
- Class marks a successful landing with a 1 and a failure with a 0.
- These data show that **the probability of success decreases with increasing payload size**.
- **Large payloads** have **not been launched on** several booster types (**v1.0, v.11,** and **B5**).

4.3 Predictive Analysis (Classification)

4.3.1 Classification Accuracy
4.3.2 Confusion Matrix

4.3.1) Classification Accuracy





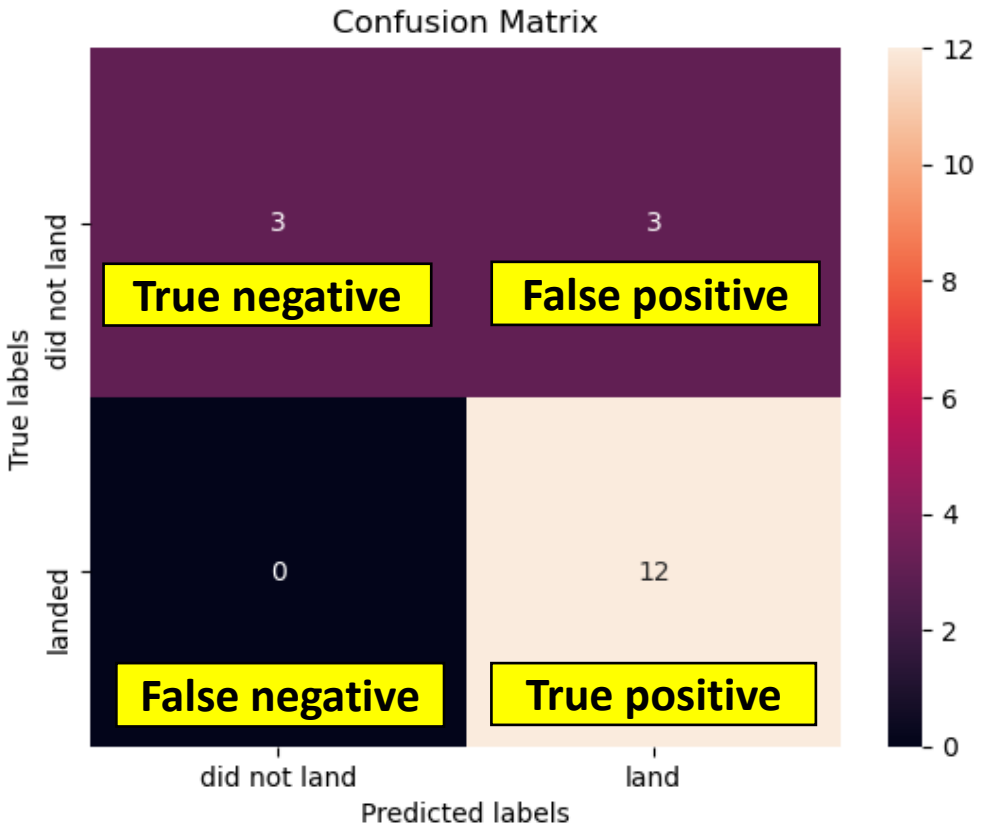
Explanation:

- Plotting the Accuracy Score and Best Score for each classification algorithm yields the following result:
  - The **decision tree** model has the **highest accuracy** on the **train set (87.50%)** compared to the **others**, which have the same accuracy at **84.73%**.
  - However, **all models had the same accuracy** on the **test set**, **83.33%**
- It should be emphasized that the test data size is tiny, with only 18 samples, which can cause considerable variations in accuracy findings.
  - More data will most likely be required to select the optimum model.

- ❑ Log_reg = Logistic regression
- ❑ Tree = Decision tree
- ❑ SVM = Support vector machine
- ❑ KNN = K-nearest neighbor

4.3.2) Confusion Matrix

The confusion matrices of the four models (Logistic regression/Decision tree/SVM/KNN) are identical.



Explanation:

- When the **true label** was **unsuccessful landings**, the **models predicted three unsuccessful landings** (**True negative** or TN).
- When the **true label** was **unsuccessful landings**, the **models predicted three successful landings** (**False positive** or FP).
- When the **true label** was **successful landings**, the **models predicted 0 unsuccessful landings** (**False negative** or FN).
- When the **true label** was **successful landings**, the **models predicted 12 successful landings** (**True positive** or TP).

5. Conclusion and innovative insights

5.) Conclusion and innovative insights (1/2)

Mission success can be explained/predicted by the following factors/models. It can also be assumed that expertise gained between launches, allowing for a successful launch.

1. Flight number

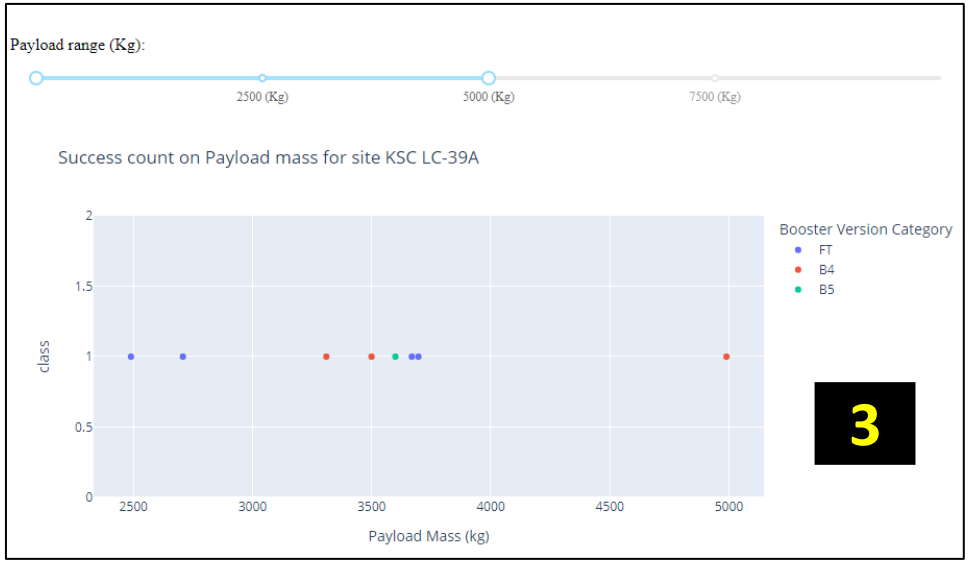
- As the **number of flights increases**, a **launch site's success rate generally rises**, with most early flights failing. Experience likely enhances success rate.

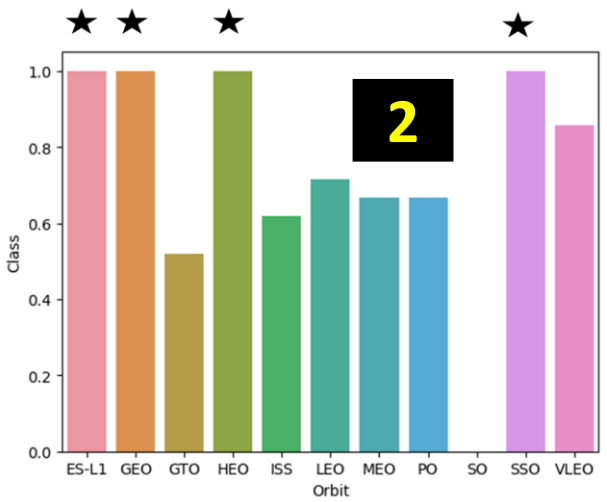
2. Orbit types

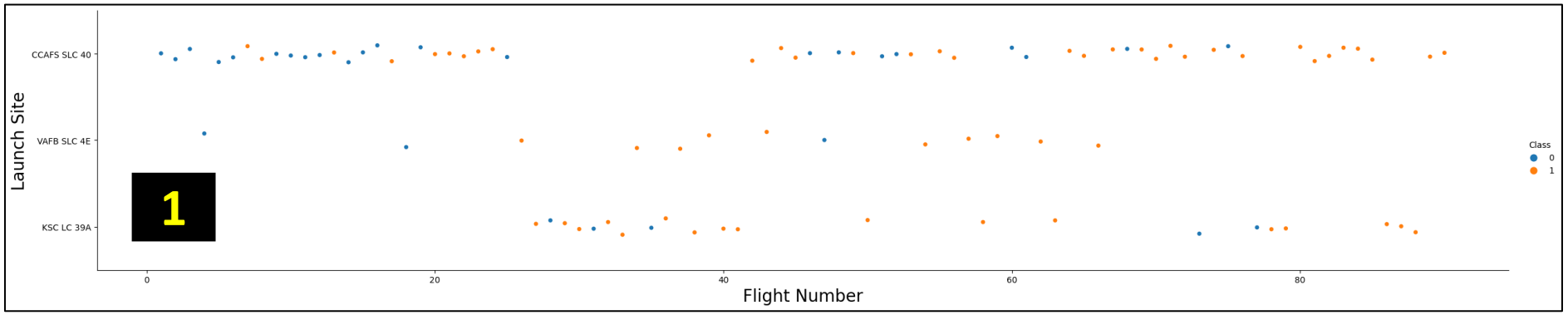
- ES-L1, GEO, HEO, and SSO have the **highest success rate** (SSO's 100% success rate with 5 flights is more noteworthy).

3. Payload mass

- **Depending on orbits, payload mass can affect mission success.** Some orbits demand hefty or small payloads. **Low-weight payloads likely perform better than hefty ones**. Massive payloads (above 5000kg) have lower success rates. Increasing payload seems to reduce success.

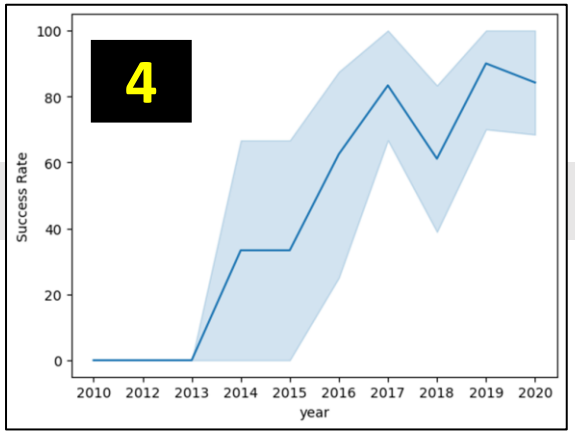




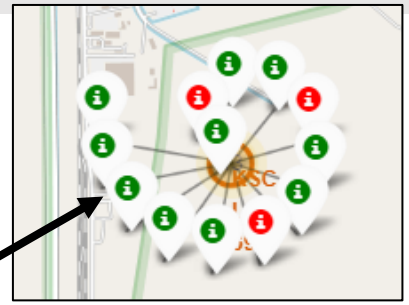


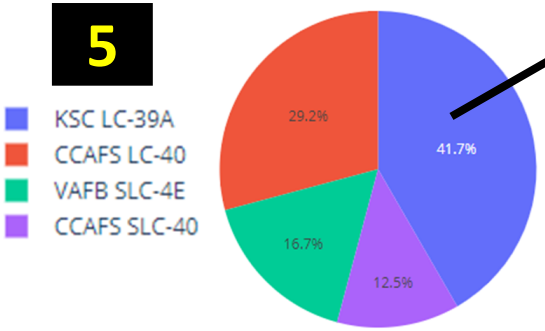
5.) Conclusion and innovative insights (2/2)

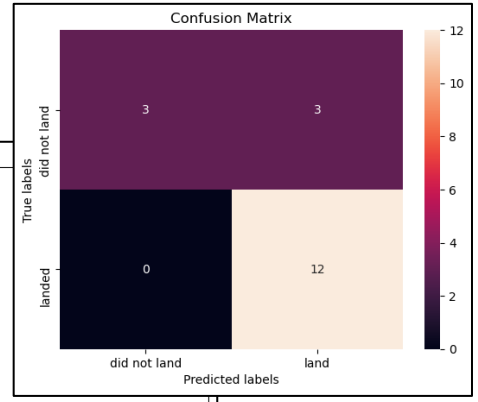
4. Yearly trend
   - **After 2016, success had a 50% likelihood**. The launch success rate seems to rise with the growth of techniques, rockets, and experience.
5. Launch site
   - **Most launch places** are **near the Equator and** the **coast.**
   - **KSC LC-39 A** had **41.7% of all successful launches** and **76.9% success rate**.
     - It is unclear from the data why some launch sites are superior (KSC LC-39A is the best launch site). To solve this challenge, we may collect atmospheric or other geospatial data on landing success.
6. Predictive models
   - **Logistic regression, Decision tree, SVM, and KNN** had the **same 83.33% accuracy on the test dataset** and confusion matrix result.
   - Thus, Allon Mask of SpaceY may use these models to forecast with great accuracy if a launch will have a successful Stage 1 landing before launch.
   - **Other ways to improve the model**:
     - 1 **Develop the best model and hyperparameters** and **re-fit utilizing the other train/test dataset ratio** instead of 80% train and 20% test data to improve model accuracy.
     - 2 **Add new launch data** to the dataset and model.

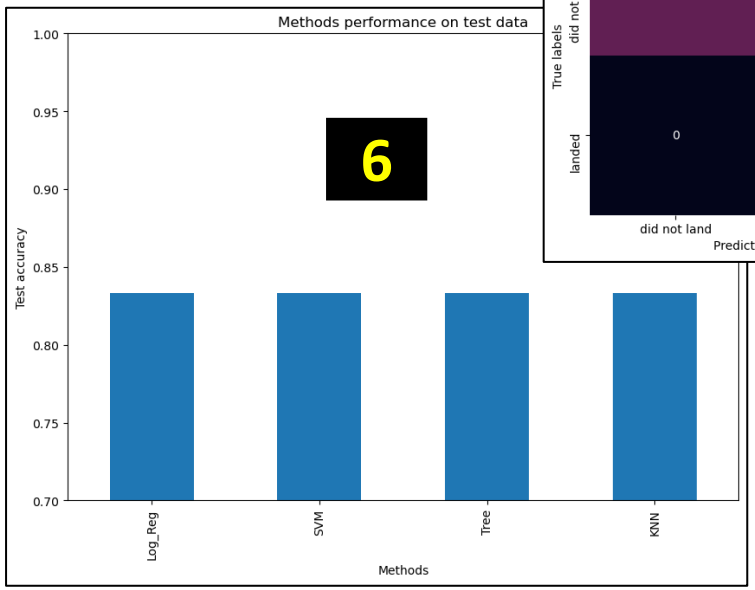


Green = Successful; Red = Unsuccessful









6. Acknowledgements

6.) Acknowledgments

Special thanks to the course organizers, the instructor team, and all anonymous peer reviewers

coursera

IBM Data Science Professional Certificate

Kickstart your career in data science & ML. Build data science skills, learn Python & SQL, analyze & visualize data, build machine learning models. No degree or prior experience required.

★★★★½ 4.6 57,815 ratings

IBM Professional Certificate — Data Science — Powered By coursera

SKILLS NETWORK — IBM Developer

Instructors

Rav Ahuja
Global Program Director
IBM Skills Network
1,295,328 Learners
33 Courses

Alex Aklson
Ph.D., Data Scientist
718,728 Learners
22 Courses

Aije Egwaikhide
Senior Data Scientist
IBM
371,538 Learners
6 Courses

Svetlana Levitan
Senior Developer Advocate with IBM Center for Open Data and AI Technologies
316,829 Learners
1 Course

Romeo Kienzler
Chief Data Scientist, Course Lead
IBM Watson IoT
444,395 Learners
9 Courses

Polong Lin
Data Scientist
213,073 Learners
6 Courses

Joseph Santarcangelo
Ph.D., Data Scientist at IBM
IBM Developer Skills Network
807,818 Learners
24 Courses

Azim Hirjani
Cognitive Data Scientist
89,424 Learners
1 Course

Hima Vasudevan
Data Scientist
IBM
307,857 Learners
4 Courses

Saishruthi Swaminathan
Data Scientist and Developer Advocate
IBM CODAIT
183,166 Learners
2 Courses

SAEED AGHABOZORGI
Ph.D., Sr. Data Scientist
294,779 Learners
4 Courses

Yan Luo
Ph.D., Data Scientist and Developer
IBM
155,852 Learners
8 Courses

7. Appendix

7.) Appendix

- **Data sources regarding SpaceX historical launches used in the project can be found here:**
  - SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - Web Scrapping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

- **The relevant Python codes, SQL queries, Notebook outputs, and data sets for this project can be found here:**
  - https://github.com/Feem-NS/IBM-Data-Science-Professional-Certification/tree/main/IBM-Module-10%20Applied-Data-Science-Capstone%20(Project)

- **More information about the IBM Data Science Professional Certificate offered by Coursera and IBM can be found here:**
  - Overview of 10 courses:
    - https://www.coursera.org/professional-certificates/ibm-data-science#courses
  - The applied data science capstone (10th course):
    - https://www.coursera.org/learn/applied-data-science-capstone?specialization=ibm-data-science

Thank you!