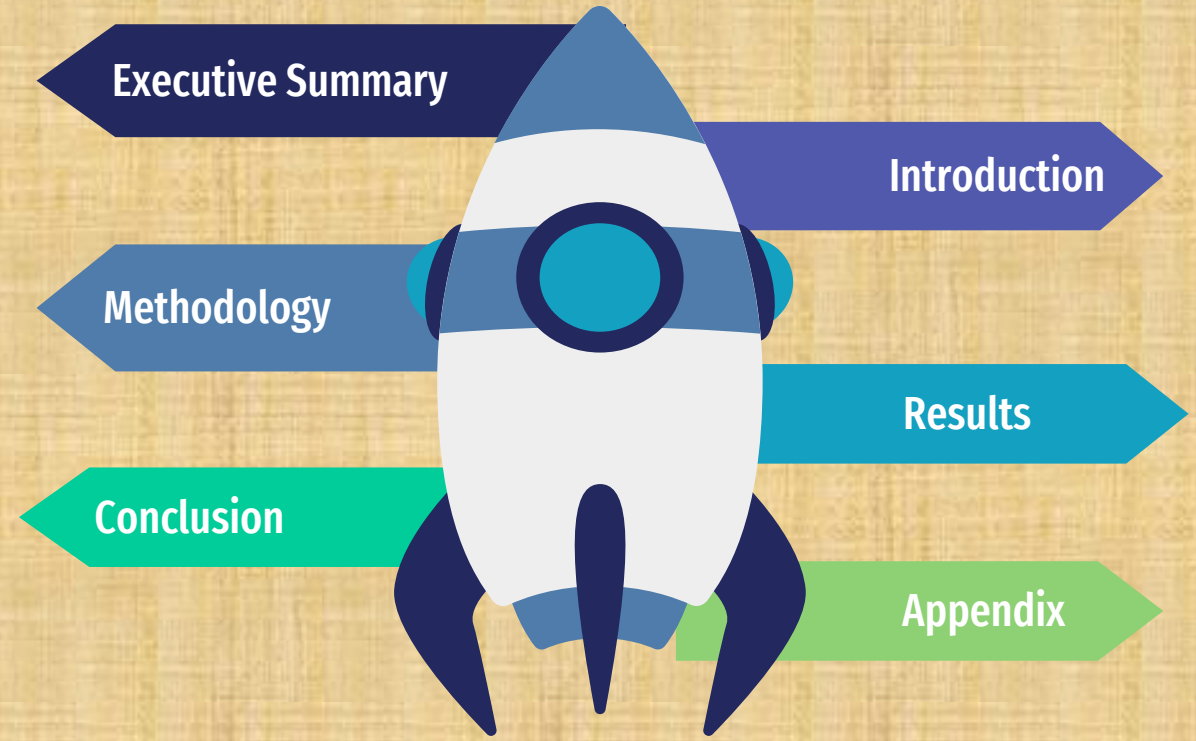
IBM Data Science Capstone Project: Landing Cost Prediction of the Falcon 9

BY: HERBERT BETT

DATE: 17/02/2023

TABLE OF CONTENTS



INTRODUCTION

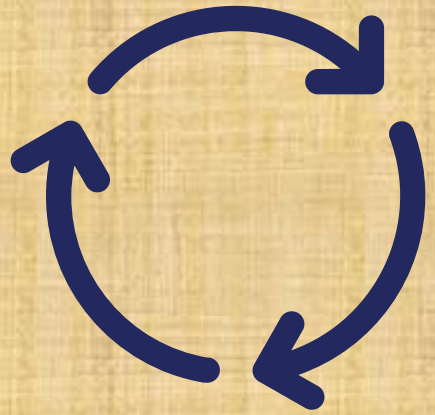
- Normally, rockets are destroyed during their first space flight. Making rockets practically reusable has drawn interest and support from all over the world in an effort to dramatically lower the cost of space exploration.

- Commercial aerospace firms like SpaceX have made major advancements in developing the technologies related to this phenomenon. Due to how reasonably priced rocket launches wind up being, this has proven to be convenient. This means that the cost can be calculated if it can be determined whether the first stage will land.

PROJECT OBJECTIVES

- Build an interactive dashboard to extract actionable insights from the collected data
- Extract data using the SpaceX REST API and populate it into a Pandas DataFrame.
- Train a classification model to determine whether the first stage of the launch will land instead of using rocket science

METHODOLOGY



Methodology

Data Collection

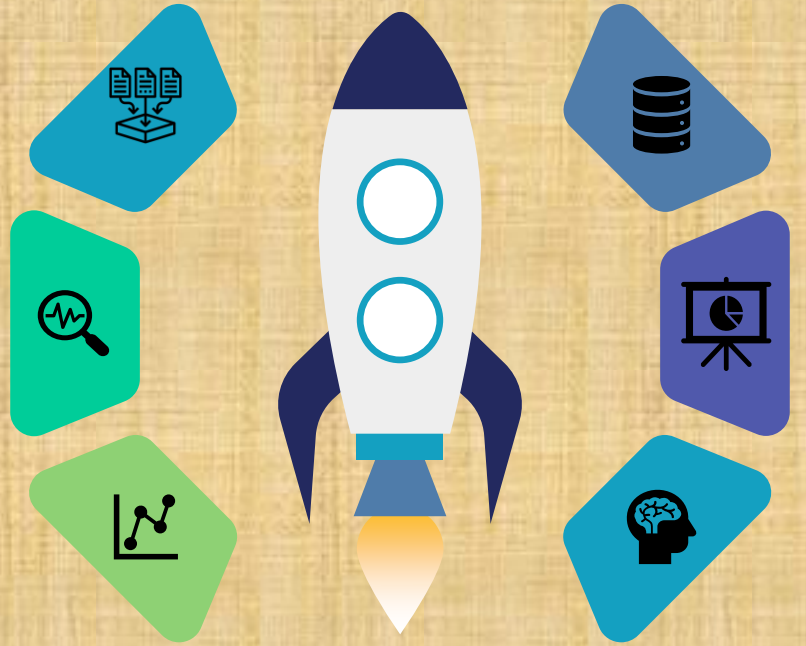
Involved collecting data from the SpaceX API and Web Scraping.

Data Wrangling

Involving cleaning and Transforming the obtained data.

EDA using Visuals

Data was analyzed using the Seaborn and Matplotlib libraries.



EDA Using SQL

SQL is used to perform further data analysis.

Interactive Visual Analytics

The Folium and Plotly Dash libraries are used to extract further insights.

Predictive Analysis

Training of classification model to predict whether first stage will land.

Data Collection

- The rocket launch data was collected from the SpaceX REST API using the requests library in Python
- Web Scraping was also used to collect some of the data using the bs4 library in Python.
- The obtained data was then populated into a Pandas dataframe.
- Subsequently, the data was cleaned with the exception of the **Landing Pad** column since the null values represented a situation when the identified pad was not utilized during the launch.

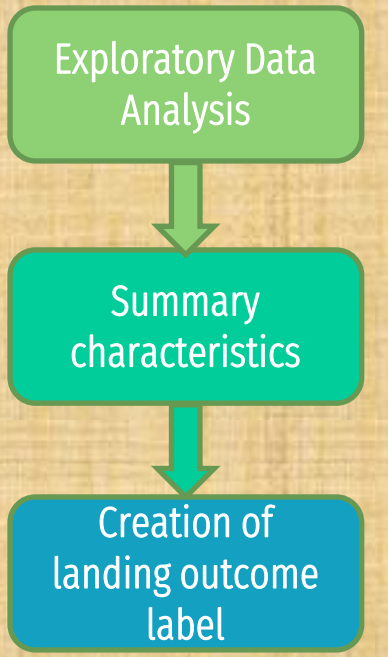
Dataset Description

Key Features	Description
Booster Version	Version of booster i.e. F9 B5 B1049.4
Launch Site	Name of site where rocket launch took place.
Payload Mass	Mass in kilograms of payload.
Orbit Type	The path to be taken in space.
Cores	Identification of core used
Outcome	Result after launch i.e. Success or Failure.

Data Wrangling

The Data Wrangling process consisted of 3 steps :

- Exploratory Data Analysis
- Calculation of key summary characteristics such as the no. of launches per launch site , the number and occurrence of each orbit type and mission outcome per orbit type .
- Finally, an outcome label was created using one-hot encoding.



Exploratory Data Analysis using Visualization

- In this section, the **seaborn** and **Matplotlib** libraries were used to generate graphs and plots that were used to extract further insights from the dataset.

- Scatter plots were used mainly to identify relationships between the predictor variables and the response variable or to observe whether there was any correlation between the predictor variables.

- Some of the visualizations were also used to capture any patterns in the data.

Exploratory Data Analysis using SQL

- SQL (Structured Query Language) was used to query the data that was contained in a SQLite database.

- The following are some of the queries that were performed on the data:
  - 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.

Interactive Visual Analytics using Folium

- Interactive visual analytics add a different dimension when it comes to extracting actionable insights from data.

- In this case, the Folium library was used to create interactive maps that were used to clearly show the locations of the various launch sites relative to key infrastructure such as roads or the proximity of the same sites to various landmarks such as coastlines.

- This was done using markers, clusters, polylines , icons among other features.

Interactive Visual Analytics using Plotly Dash

- The Dash library is used to create a dashboard app that consists of various features such as a dropdown for the various launch sites and a range slider that was used to obtain the successful and unsuccessful launches in various payload ranges by booster version.

- The dashboard app consisted of two sections that would display a pie chart and a scatter plot respectively based on the selections from the dropdown and range slider.

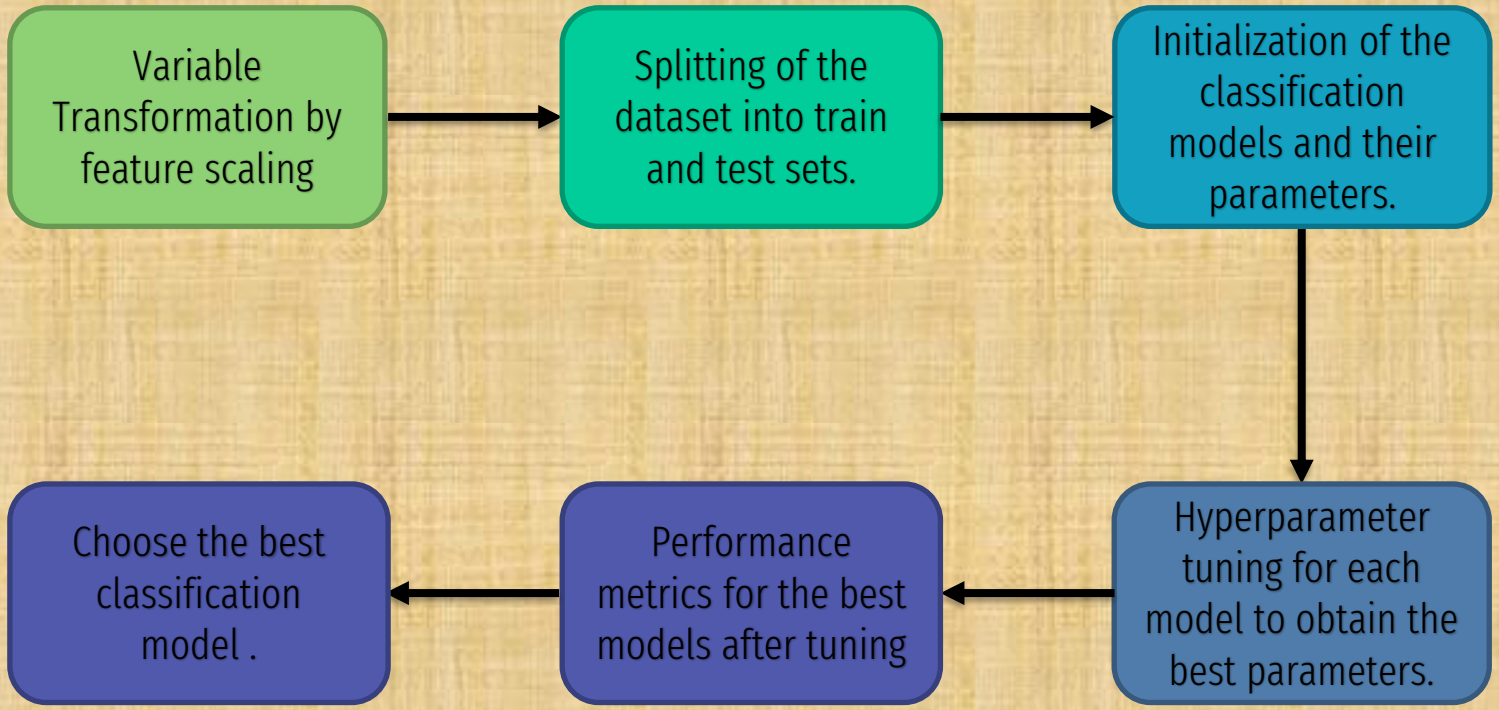
Predictive Analysis

- This is the final and key step in the process and is used to determine whether the first stage of the launch will land.

- Since we want to predict the outcome of a particular activity, classification models were used to predict the launch outcome.

- The following were the classification models used:
  - ❑ Logistic Regression
  - ❑ Support Vector Machines
  - ❑ K Nearest Neighbors
  - ❑ Decision Trees

Predictive Analysis Process



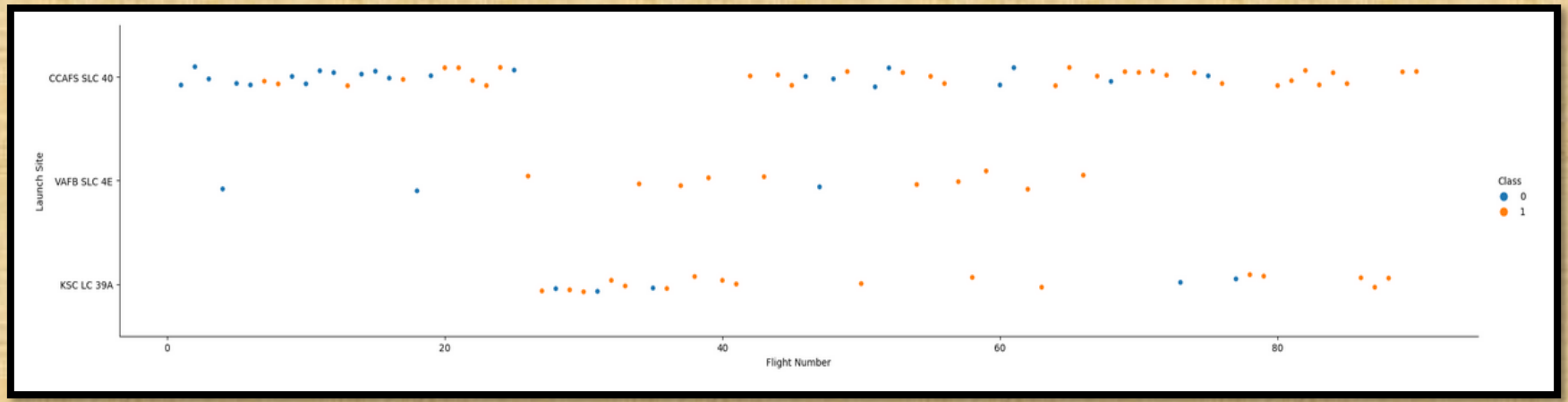
EDA RESULTS

Exploratory Data Analysis Results

- There were four distinct launch sites used in Falcon 9 launches.
- The average payload mass of rockets using the F9 V1.1 booster version was 2928.4 kg.
- The first successful ground pad landing took place nearly 8 years since the first launch.
- Almost 99% of the missions were successful.
- The number of successful landing outcomes increased from 2013 onwards.
- Most Falcon 9 booster versions were successful at landing in drone ships having payload masses above the average.
- Most of the launch sites were located at a close proximity to the coastlines and far away from critical infrastructure

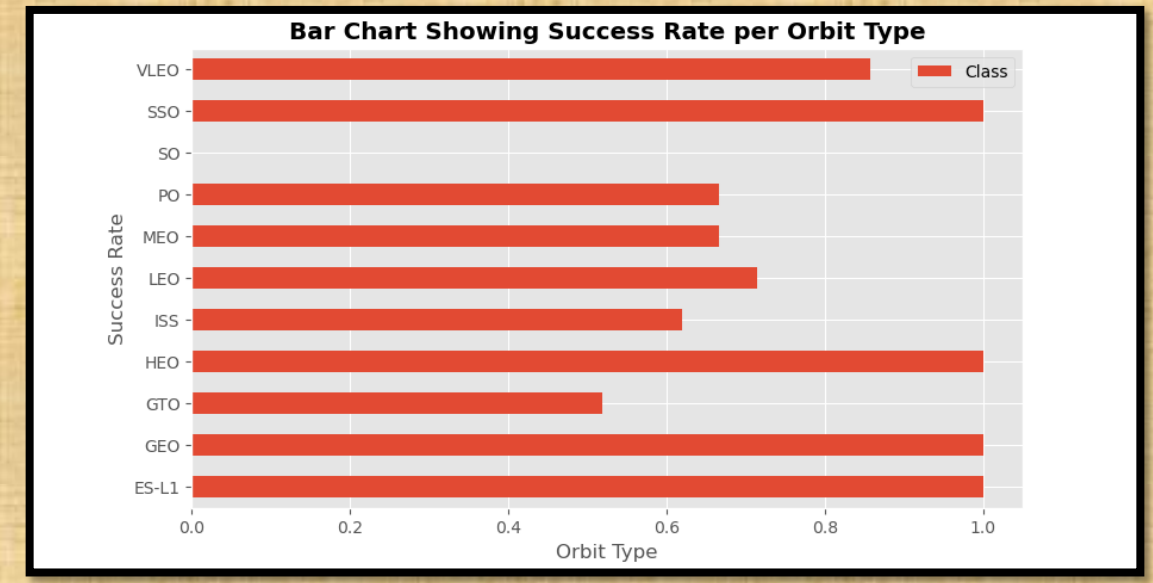
EDA USING VISUALIZATION RESULTS

Launch Site vs Flight Number



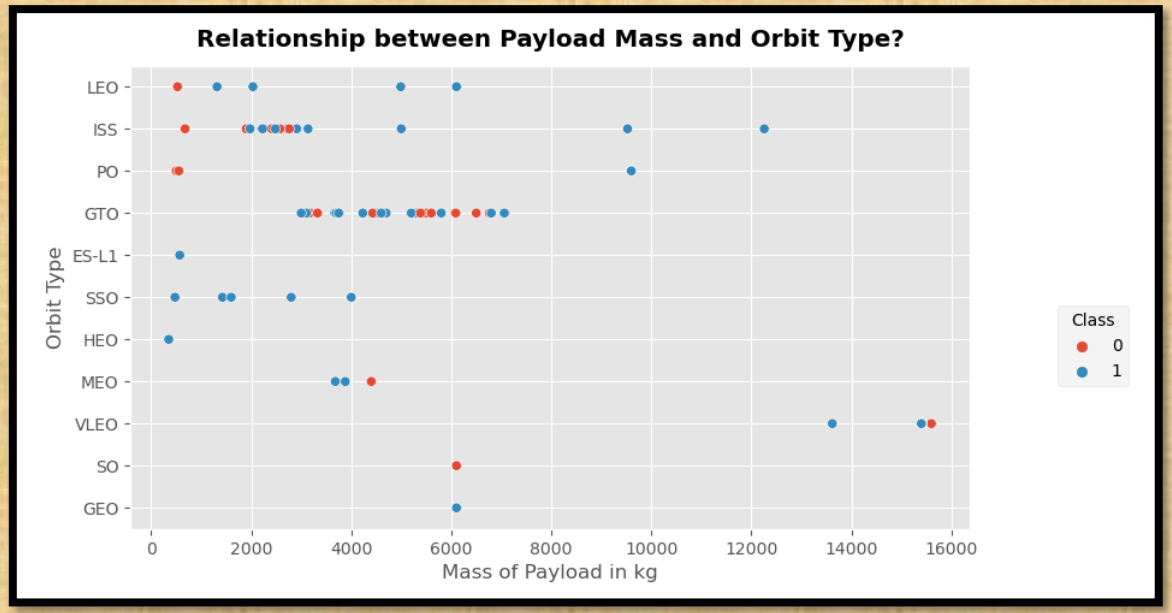
- Most launches take place at the **CCAFS SLC 40** most of which are successful while the **VAFB SLC 4E** has the least amount of launches most of which are successful.

Success Rate per Orbit type



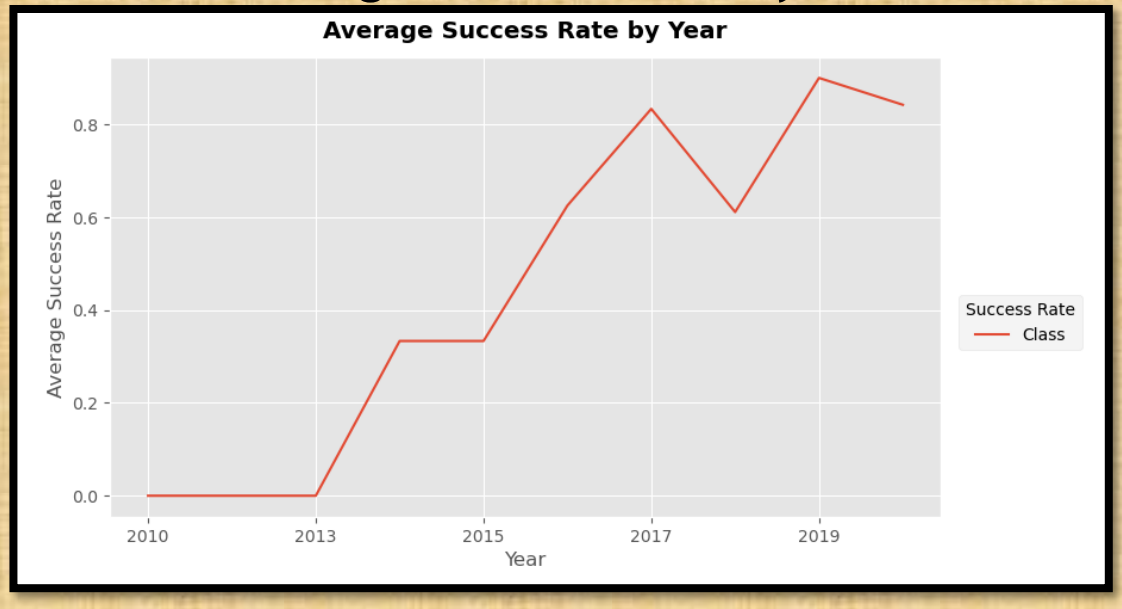
- **SSO, HEO, GEO and ES-L1** have the highest success rate when they are used during the launches.
- The **SO** orbit type has a **zero success rate** when used for Falcon 9 launches.

Orbit Type vs Payload Mass



- The payload mass range for rockets launched in the GTO orbit is **2000 – 8000 kg.**
- The **VLEO** orbit type is used when rockets contain **heavy payloads.**
- The **ES-L1** and **HEO** orbit types are the **least utilized** orbit types.

Average Success Rate by Year



- The average success rate of the rocket launches increases drastically **after 2013** as shown by the sharp increase in slope of the graph.
- The stagnation between **2010 – 2013** can be attributed to the various **launch failures** when SpaceX started out experiments on reusable rockets.

EDA USING SQL RESULTS

Launch Sites Used

The launch sites that were mainly used include:

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

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

EDA USING SQL RESULTS

Payload Mass carried by NASA(CRS) launches

- The total payload mass carried by NASA (CRS) is **45596 kg .**

Customer	Total mass
NASA (CRS)	45596

EDA USING SQL RESULTS

Average Payload Mass carried by F9 V1.1 boosters

- The average payload mass carried by **F9 v1.1** boosters is **2928.4 kg**

Booster Version	Average mass
F9 v1.1	2928.4

EDA USING SQL RESULTS

First Successful landing on a Ground Pad

- The first successful ground pad landing occurred on **01-05-2017**

MIN(“Date”)
01-05-2017

EDA USING SQL RESULTS

Boosters with a successful landing outcome on a drone ship and a Payload mass between 4000 – 6000 kg

- The booster versions with a **successful** landing outcome on a **drone ship** and a payload mass between **4000 and 6000 kg** are as shown below.

Booster Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

EDA USING SQL RESULTS

Proportion of successes and failures in landing outcome

- The proportion of **successful** and **unsuccessful** landing outcomes is as shown below:

Mission Outcome	Total outcomes
Failure (in flight)	1
Success	99
Success (payload status unclear).	1

EDA USING SQL RESULTS

Booster Versions that carry the maximum Payload mass

- The booster versions that carry the maximum payload mass recorded are as shown below:

Booster Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4

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

EDA USING SQL RESULTS

Booster Versions that failed to land on a drone ship in 2015

- The launch sites where there were failed landings on drone ships and their booster versions are shown below:

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

EDA USING SQL RESULTS

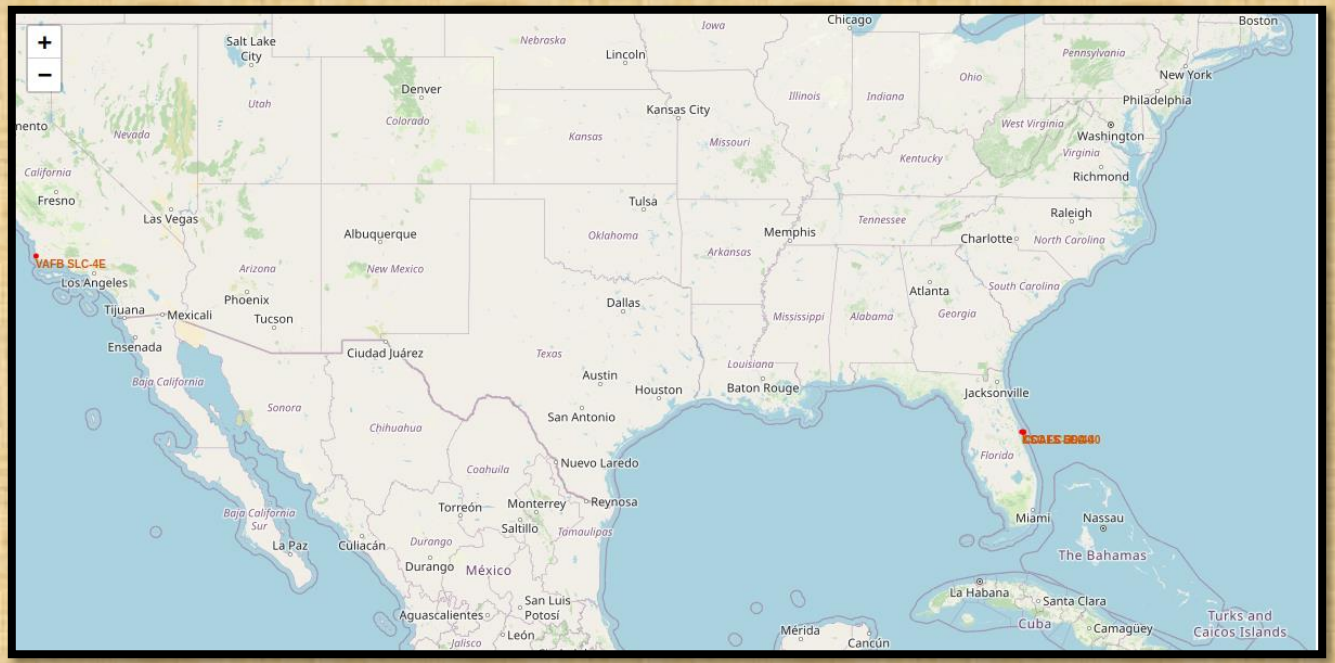
Total possible outcomes ranked by count between 04-06-2010 and 20-03-2017

- Ranking the all the possible outcomes by count between **04-06-2010** and **20-03-2017** yields:

Landing Outcome	Total count
Success	20
No attempt	11
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2

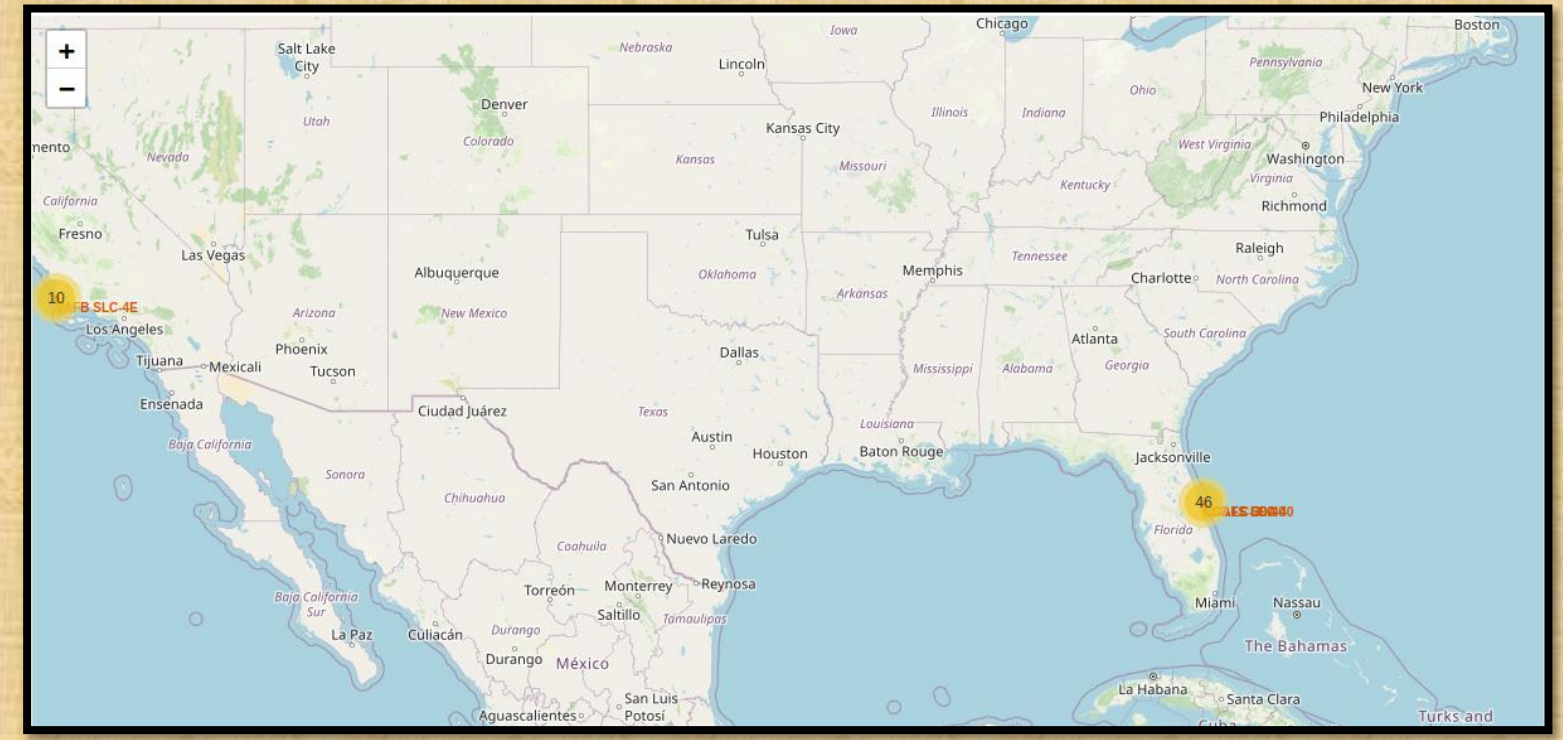
INTERACTIVE VISUAL ANALYTICS USING FOLIUM

General location of launch sites



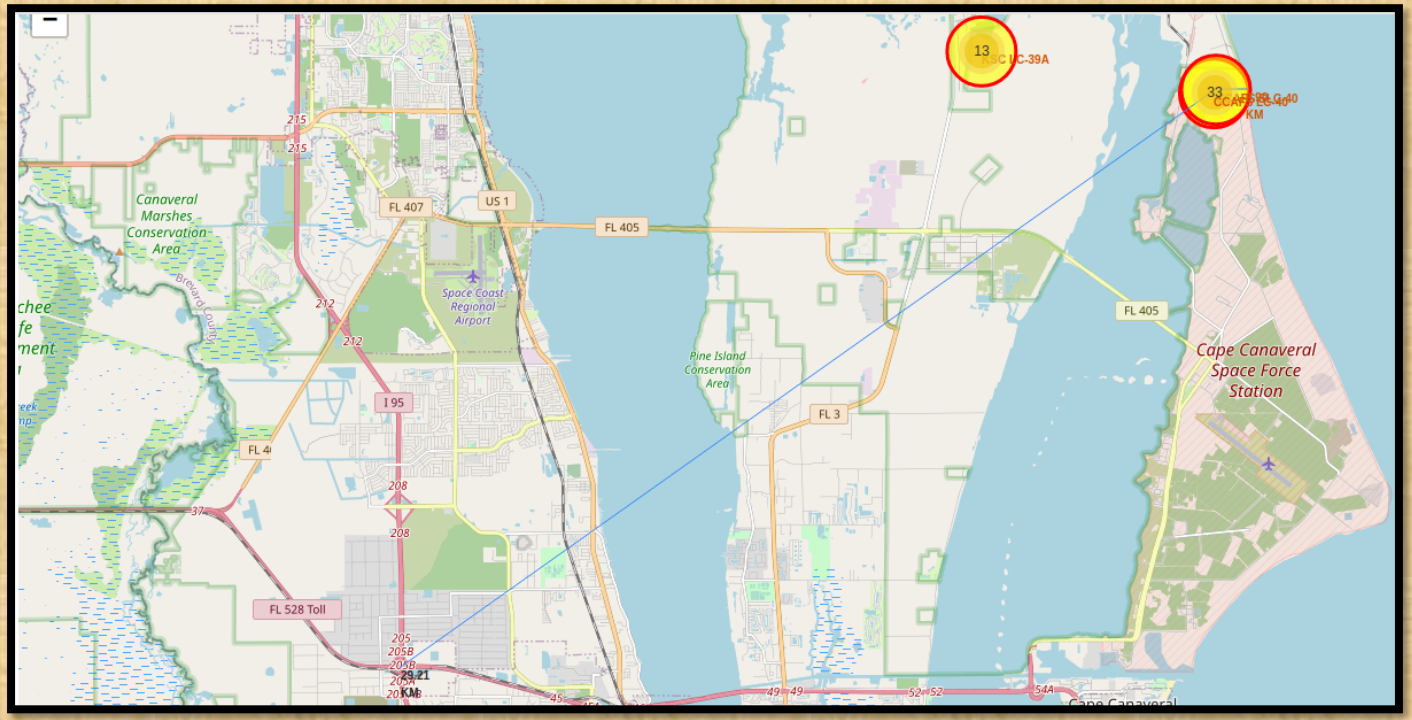
- All the launch sites are located at or near coastlines as shown above to ensure safety of people and avoidance of critical infrastructure damage.

- Upon further analysis, it is observed that most of the launch sites are located at or near the **coastline of Florida** (46) while the rest are found on the **coast of California** (10) as shown below.



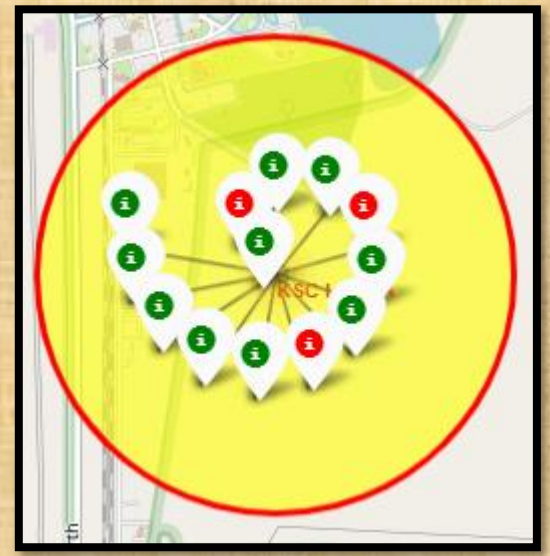
Proximity analysis on the Launch Sites.

- Proximity analysis was carried out by computing the distance between the launch sites and critical infrastructure such as roads , railways and buildings using the **haversine distance** and drawing a line between the launch sites and the various structures.
- Upon performing this computation and visualizing it on the map, it can be concluded that the launch sites are located as far as possible from populated areas and key infrastructure as said earlier.
- An example of this can be seen on the next slide.



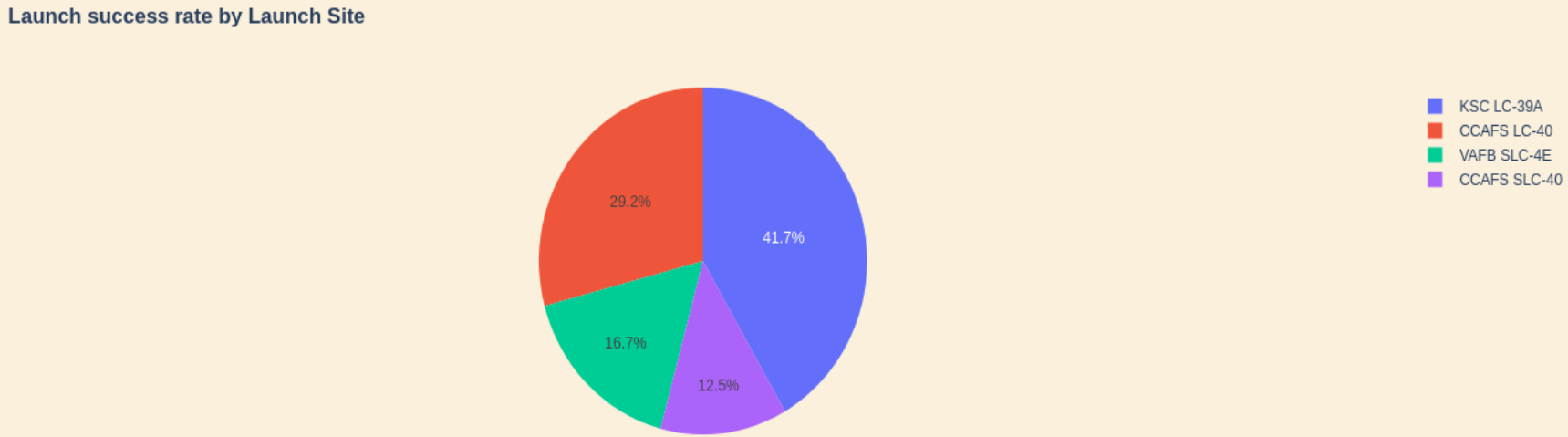
- From the above graphic, it can be seen that the launch sites are located as far as possible from populated areas.

- Markers and icons with special color coding were used to represent the landing outcomes in the various launch sites with green and red used to show successful and unsuccessful landing outcomes respectively as shown below.



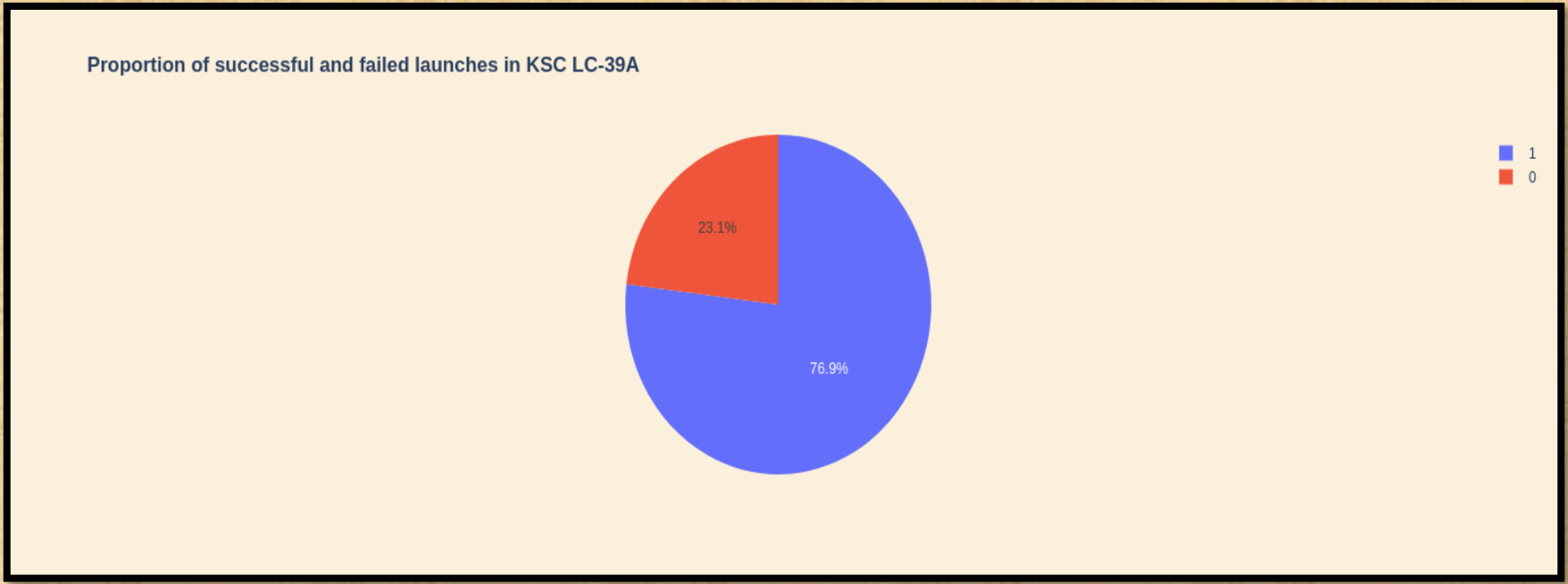
INTERACTIVE VISUAL ANALYTICS USING PLOTLY DASH

Launch Success Rate by Launch Site

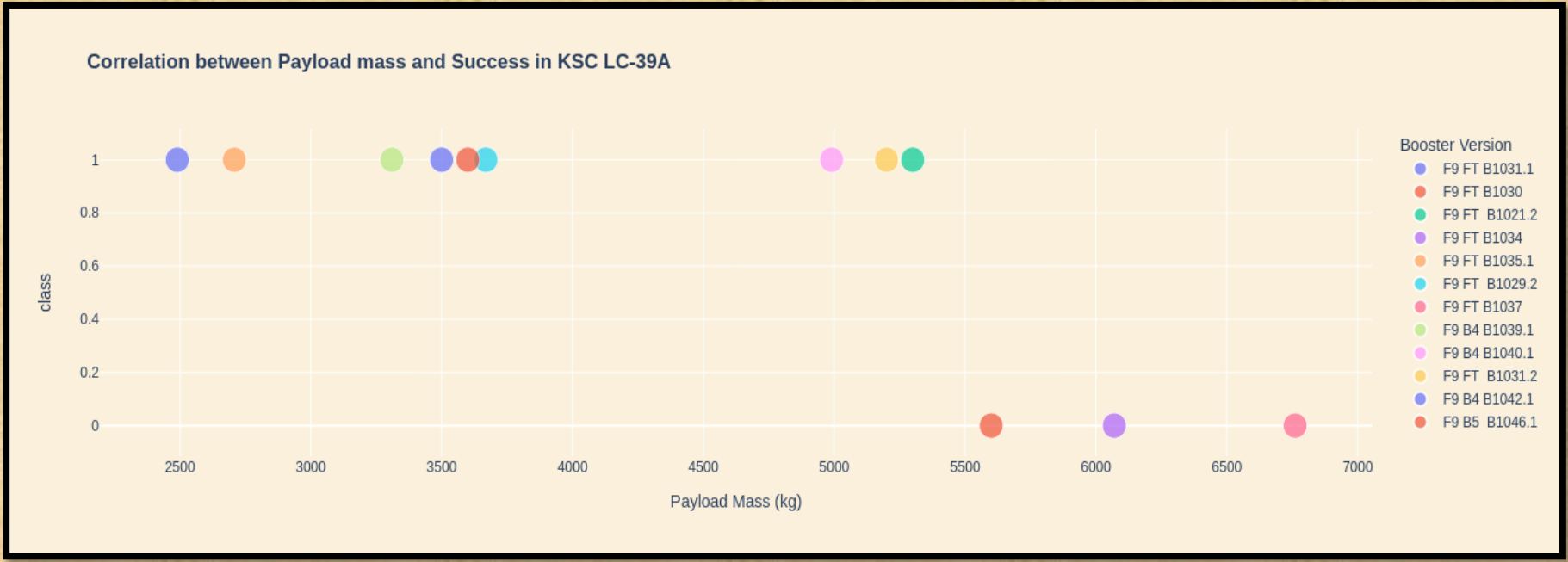


- The **KSC LC-39A** launch site accounts for **41.7%** of the successful outcomes achieved in all sites

- The same launch site has a success rate of **76.9%** in its launches as shown in the pie chart below:

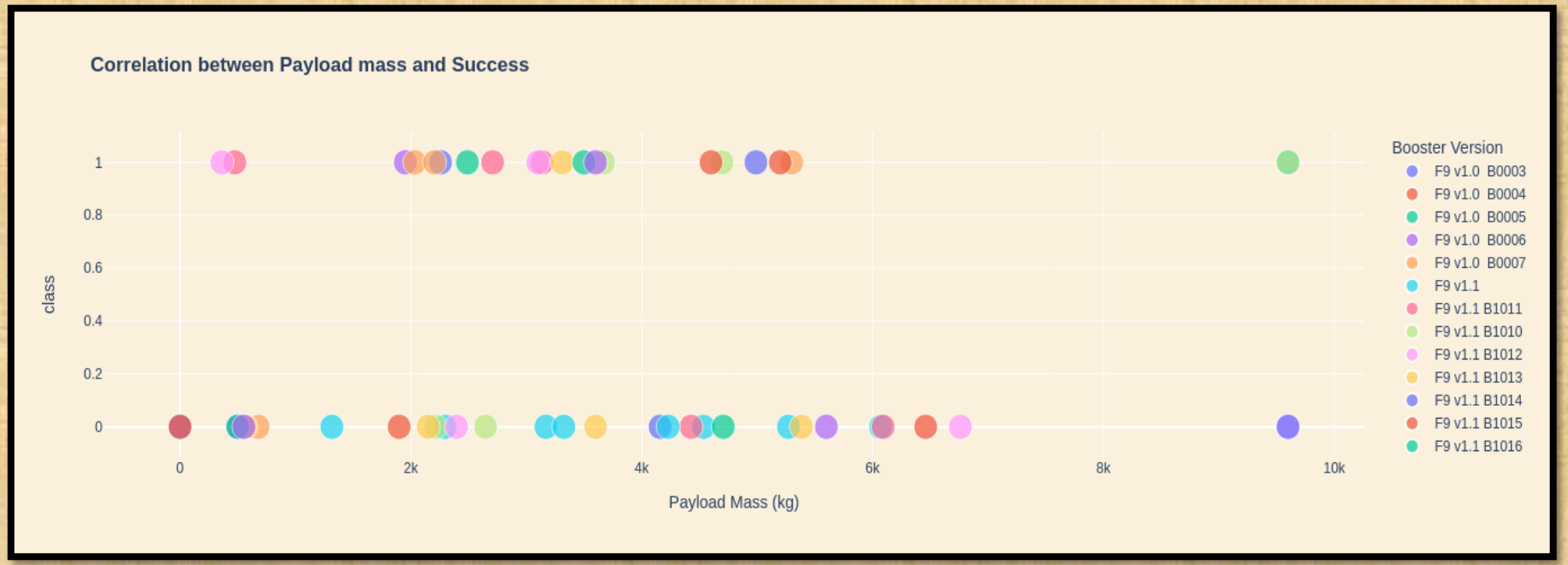


Scatter plot of Outcome against Payload mass



- From the scatter plot for launch site **KSC LC-39A**, it can be concluded that rockets launched from this location succeed when the payload mass is between **0 – 5500 kg**

- From the scatter plot for all sites, most of the launching outcomes were unsuccessful as in the figure below.

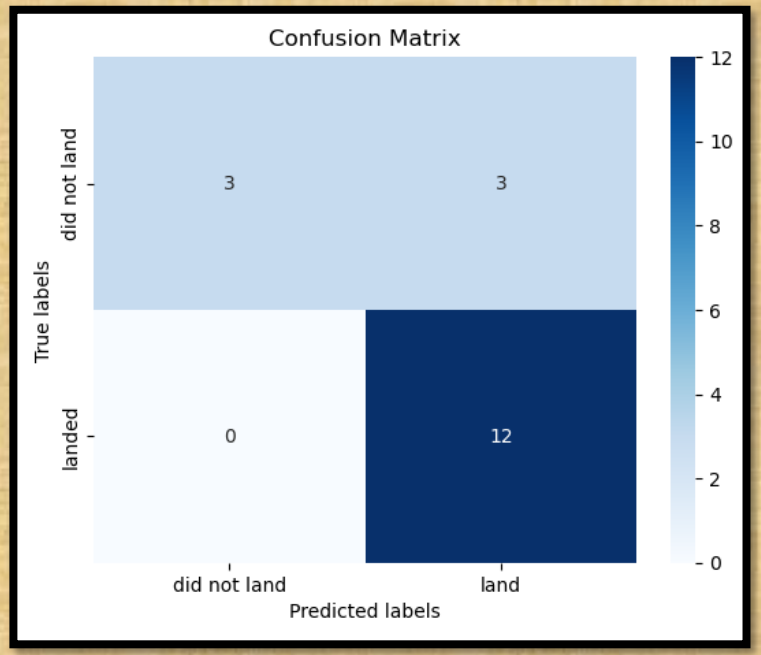


PREDICTIVE ANALYSIS

- After exhaustive analysis of the dataset, predictive analysis was carried out to **find out whether the first stage will land.**
- The predictive analysis process was as outlined in the predictive analysis slide of the methodology section.
- The said process was applied using **four** classification algorithms namely:
  - ❑ Logistic Regression
  - ❑ Support Vector Machines
  - ❑ K-nearest neighbors
  - ❑ Decision Trees

Logistic Regression

- Model features and parameters:
  - Model: LogisticRegression()
  - Solver: lbfgs
  - Penalty: L2 (Ridge)
  - Regularization strength, C : [0.01, 0.1, 1]

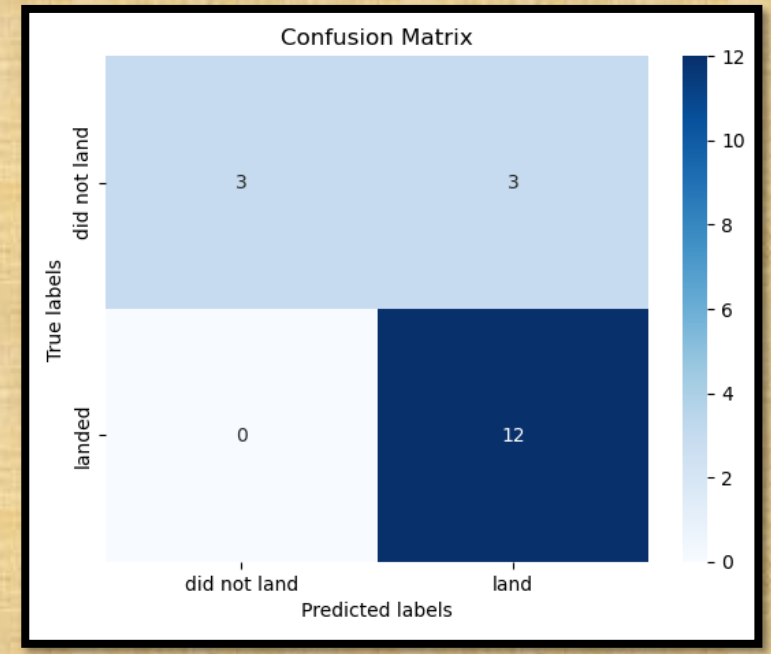


Validation score	Test score
0.8464	0.8333

Support Vector Machines

- Model features and parameters:
  - Model: SVC()
  - Kernels: [linear, poly, sigmoid, rbf]
  - Gamma: np.logspace(-3, 3, 5)
  - Regularization strength, C: np.logspace(-3, 3, 5)

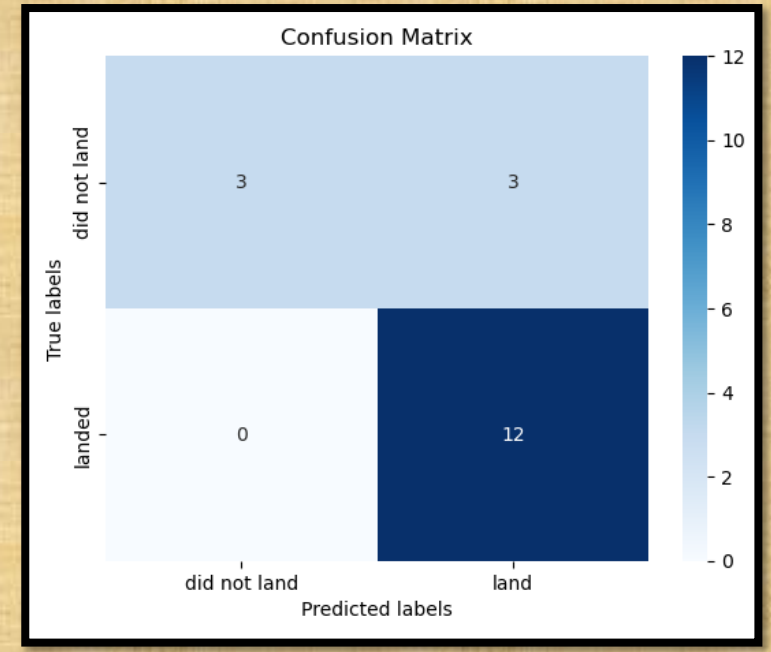
Validation score	Test score
0.8482	0.8333



Decision Tree

- Model features and parameters:

Feature / Parameter	Value
Model	DecisionTreeClassifier
Max features	[‘auto’, ‘sqrt’]
Splitter	[‘best’, ‘random’]
Criterion	[‘entropy’, ‘gini’]
Min samples leaf	(1, 2, 4)
Min samples split	(2, 5, 10)
Max_depth	[2*n for n in range(1,10)]

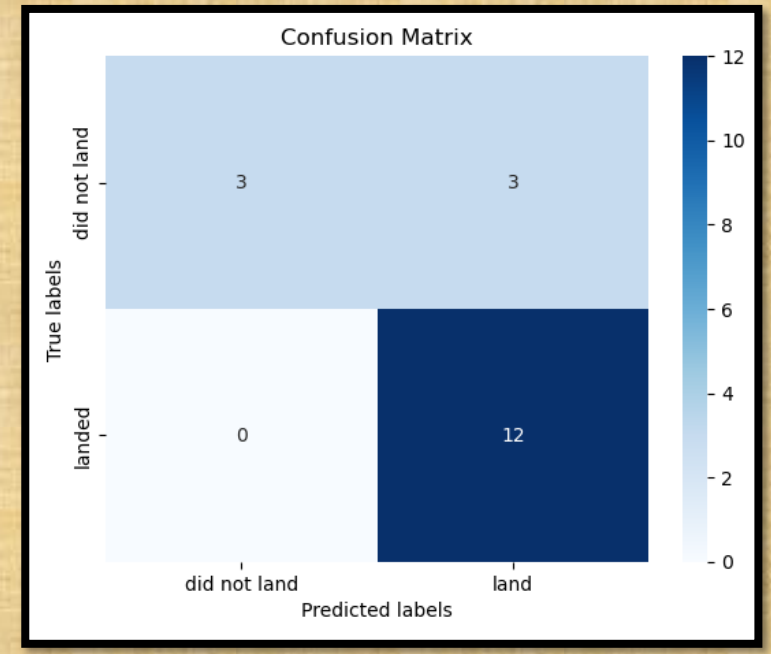


Validation score	Test score
0.8893	0.9444

K-Nearest Neighbors

- Model features and parameters:
  - Model: KNeighborsClassifier()
  - n_neighbors: range(1,11)
  - Algorithm: [‘ball_tree’, ‘kd_tree’, ‘auto’, ‘brute’]
  - p: [1, 2]

Validation score	Test score
0.8482	0.8333



- Based on the performance metrics as shown on the table, it can be concluded that:
  - The Logistic Regression, Support Vector Machine and K-Nearest Neighbors models perform worse on the test set since all their test accuracy scores are lower than the validation score.
  - The Decision Tree classification model performs exceptionally well on the test since its test score is significantly larger than its validation score and thus one is inclined to choose it.
- The final test scores are shown below:

Estimator	Test score
Decision Tree Classifier	0.9444
Logistic Regression	0.8333
Support Vector Machines	0.8333
K-Nearest Neighbors	0.8333

Conclusions

- Based on the visual analysis carried out, the most successful launch site is the **KSC LC-39A.**

- From the maps generated using Folium, almost all launch sites are located at or near coastlines to prevent damage of critical infrastructure such as roads, railways and buildings caused by debris during launch.
- The **Decision Tree Classifier model** was identified as the best model to predict whether the first stage would land.
- The **Florida coastline** accounts for nearly all the launch sites associated with the Falcon 9.

THANK YOU