IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Isaac Emmanuel Thomas
11/14/2023

Outline

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

Executive Summary

- The goal of this project was to analyze SpaceX Falcon 9 data collected through various sources and use several Machine Learning models to help predict if the SpaceX Falcon 9 first stage will land successfully.
- The following methods were the main steps in this project:
  - Collecting data through API and Web scraping.
  - Transforming the data through Data Wrangling.
  - Conducting Exploratory Data Analysis (EDA) with SQL as well as Data Visuals.
  - Building an interactive map with folium to analyze launch site proximity.
  - Building a dashboard to analyze launch records interactively with Plotly Dash.
  - Building and testing various models to predict if the first stage of Falcon 9 will land
  - successfully.

Executive Summary

- This report will share results in various formats such as:
  - Data Analysis results
  - Data visuals
  - Interactive dashboards
  - Predictive model analysis results
- It is also concluded that decision tree may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully.

Introduction

With the recent successes in private space travel, space industry is becoming more and more mainstream and accessible to general population. Cost of launch continues to remain a key barrier for new competitors to enter the space race. In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. Most unsuccessful landings are planned. Sometimes, SpaceX will perform a controlled landing in the ocean. The main question that we are trying to answer is, for a given set of features which include its payload mass, orbit type, launch site, etc., what are the impact of different parameters/variables on the landing outcomes and also what are Correlations between launch sites and success rates and finally will the first stage of the rocket land successfully?

Section 1

Methodology

Methodology

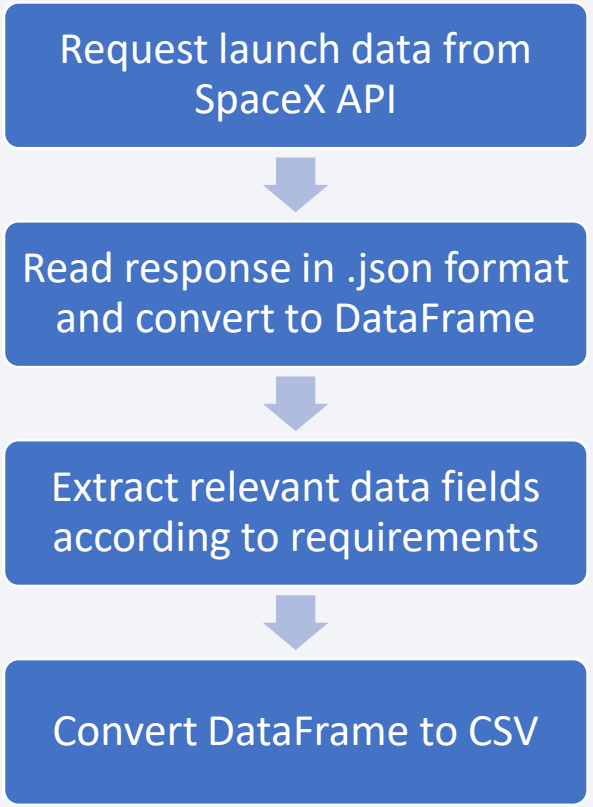
- Data collection methodology:
  - SpaceX API
  - Web scrap Falcon 9 and Falcon Heavy launch records from Wikipedia
- Perform data wrangling
  - Data was labelled for training the supervised models by converting mission outcomes in to classes (0-unsuccessful, 1-successful)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Loaded the data; standardized and transformed data; train/test split data; Find best classification algorithm (Logistic regression, SVM, decision tree, & KNN) based on evaluation scores.

Data Collection

Data collection is the process of gathering data from available sources. This data can be structured or unstructured. For this capstone project, data was collected from SpaceX API and Web scrapping Wiki pages for relevant launch data that can be used.

Data Collection – SpaceX API

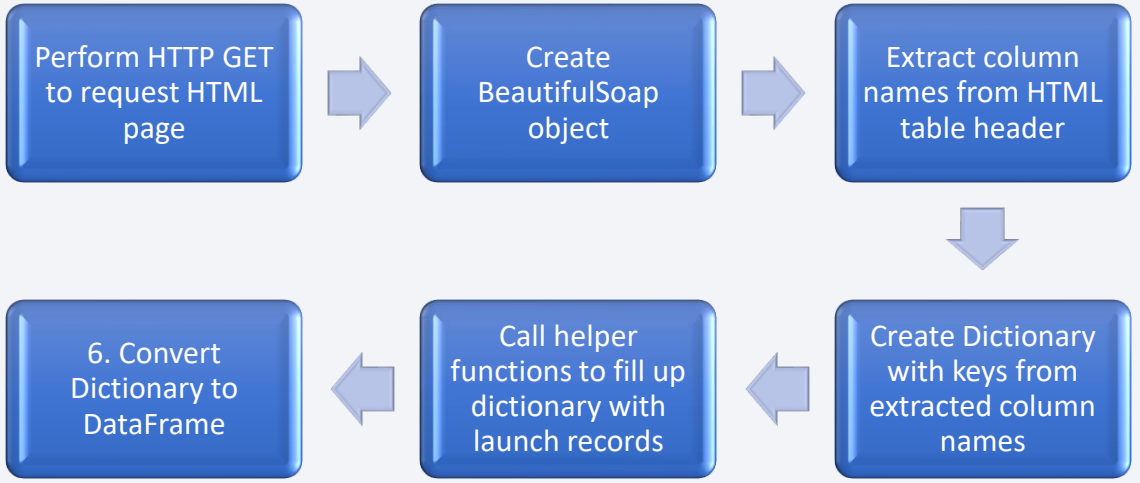
- The API used is https://api.spacexdata.com/v4/rockets/
- The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.
- Every missing value in the data is replaced by the mean of the column that the missing value belongs to.
- There are 90 rows or instances and 17 columns or features.
- The flow chart shows the data collection process.



GitHub

Data Collection - Scraping

- The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The website contains only the data about Falcon 9 launches.
- There are 121 rows or instances and 11 columns or features.
- The picture below shows the first few rows of the data:



GitHub

Data Wrangling

- The data is processed so that there are no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.
- The following landing scenarios were considered while labelling:
  - True Ocean means the mission outcome was successfully landed to a specific region of the ocean
  - False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean
  - RTLS means the mission outcome was successfully landed to a ground pad
  - False RTLS means the mission outcome was unsuccessfully landed to a ground pad
  - True ASDS means the mission outcome was successfully landed on a drone ship
  - False ASDS means the mission outcome was unsuccessfully landed on a drone ship
- In the end, there are 90 rows and 83 columns.

GitHub

EDA with Data Visualization

As part of the Exploratory Data Analysis (EDA), following charts were plotted to understand the data and gain further insights into the dataset:

1. Scatter plot: Shows relationship or correlation between two variables making patterns easy to observe The following were visualized by Scatter plot:
   - Relationship between Flight Number and Launch Site
   - Relationship between Payload and Launch Site
   - Relationship between Flight Number and Orbit Type
   - Relationship between Payload and Orbit Type
2. Bar Chart: Commonly used to compare the values of a variable at a given point in time. Length of each bar is proportional to the value of the items that it represents. The following was visualized by Bar chart:
   - Relationship between success rate of each orbit type
3. Line Chart: Commonly used to track changes over a period of time. It helps depict trends over time. The following was visualized by Line chart :
   - Average launch success yearly trend

GitHub

EDA with SQL

To understand the data better, 10 SQL queries were done on the data. The queries were as follows and the results can be seen in the GitHub link:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

GitHub

Build an Interactive Map with Folium

Folium interactive map helps analyze geospatial data to perform more interactive visual analytics and better understand factors such location and proximity of launch sites that impact launch success rate.

The Folium library was used to:

- Mark all launch sites on a map
- Mark the succeeded launches and failed launches for each site on the map
- Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

Building the Interactive Map with Folium helped answered following questions:

Are launch sites in close proximity to railways? Are launch sites in close proximity to highways? Are launch sites in close proximity to coastline? Do launch sites keep certain distance away from cities? These questions were answered with the help of the interactive map.

GitHub

Build a Dashboard with Plotly Dash

Plotly Dash web application was built to perform interactive visual analytics on SpaceX launch data in real-time. The application has a Launch Site Drop-down, Pie Chart, Payload range slide, and a Scatter chart.

1. A Launch Site Drop-down Input component was added to the dashboard to provide an ability to filter Dashboard visual by all launch sites or a particular launch site
2. A Pie Chart was added to the Dashboard to show total success launches when 'All Sites' is selected and show success and failed counts when a particular site is selected
3. A Payload range slider was added to the Dashboard to easily select different payload ranges to identify visual patterns
4. 4. A Scatter chart was added to observe how payload may be correlated with mission outcomes for selected site(s). The color-label Booster version on each scatter point provided missions outcomes with different boosters.

GitHub

Predictive Analysis (Classification)

The machine learning prediction phase included the following steps:

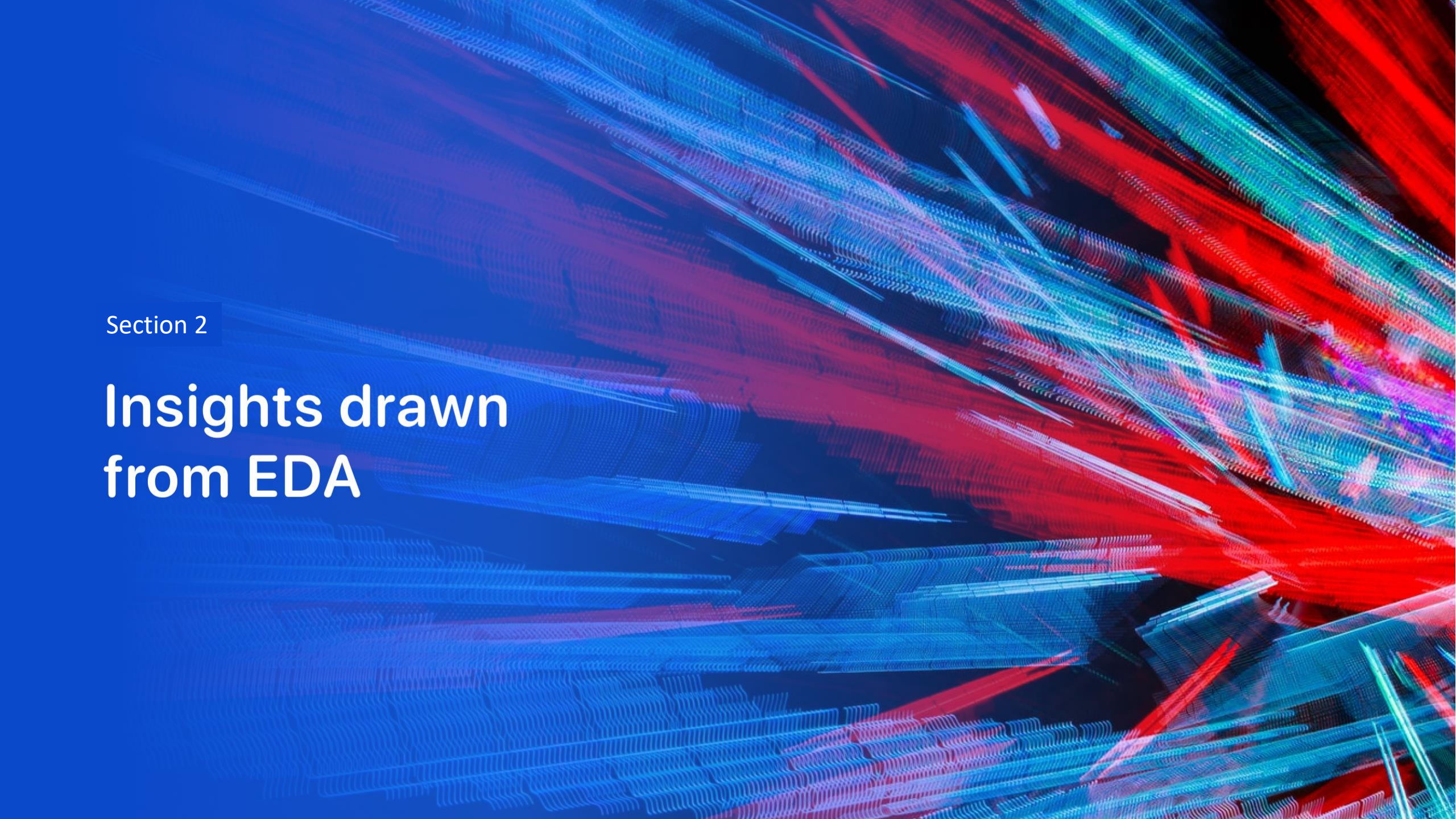
- Standardizing the data
- Splitting the data into training and test data
- Creating machine learning models, which include:
    - Logistic regression
    - Support vector machine (SVM)
    - Decision tree
    - K nearest neighbors (KNN)
- Fit the models on the training set
- Find the best combination of hyperparameters for each model
- Evaluate the models based on their accuracy scores and confusion matrix

GitHub

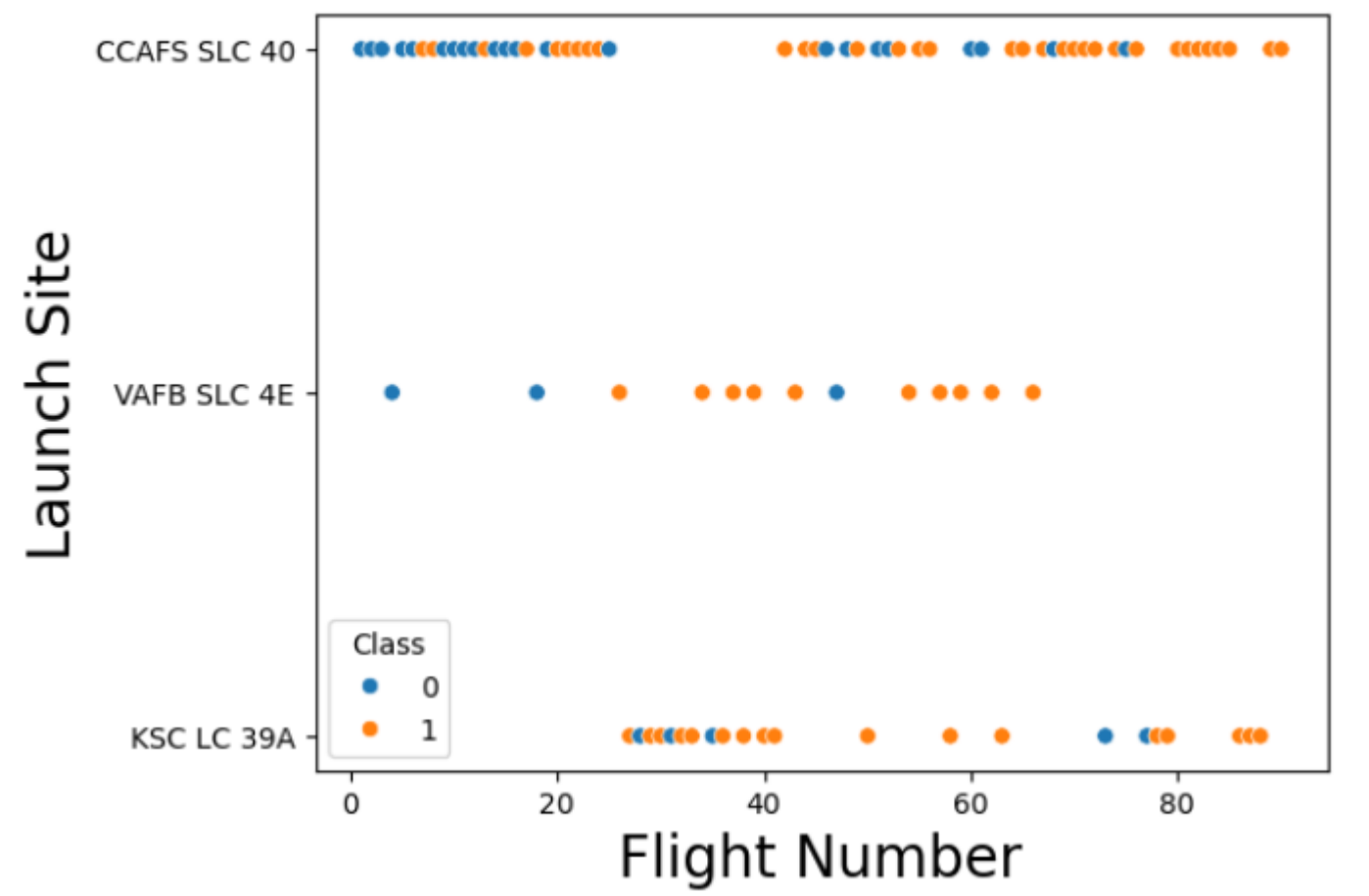
Results

The results are split into 5 sections:

- EDA with Visualization
- EDA with SQL
- Folium Map
- Plotly Dash
- Predictive Analysis

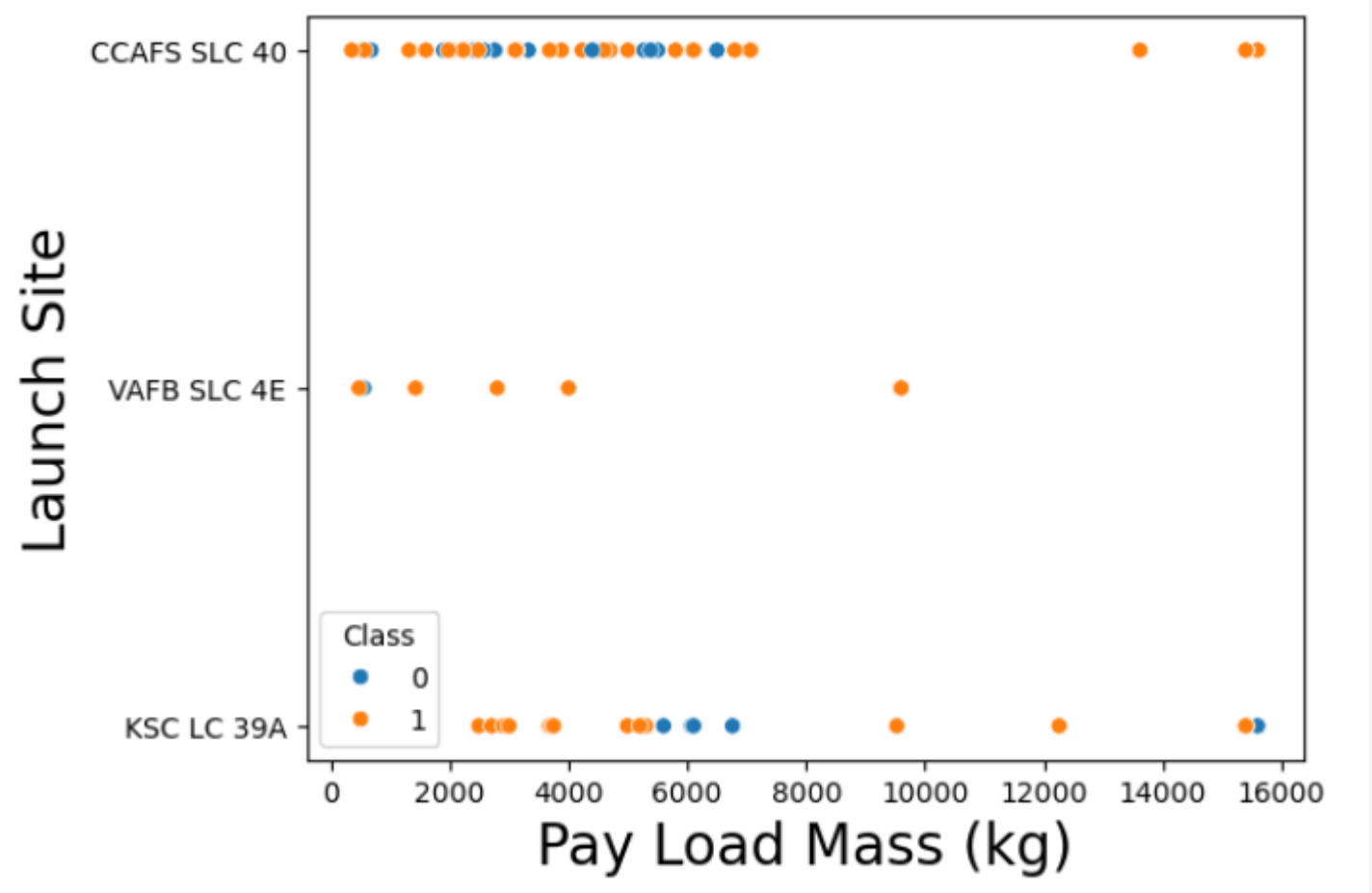

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



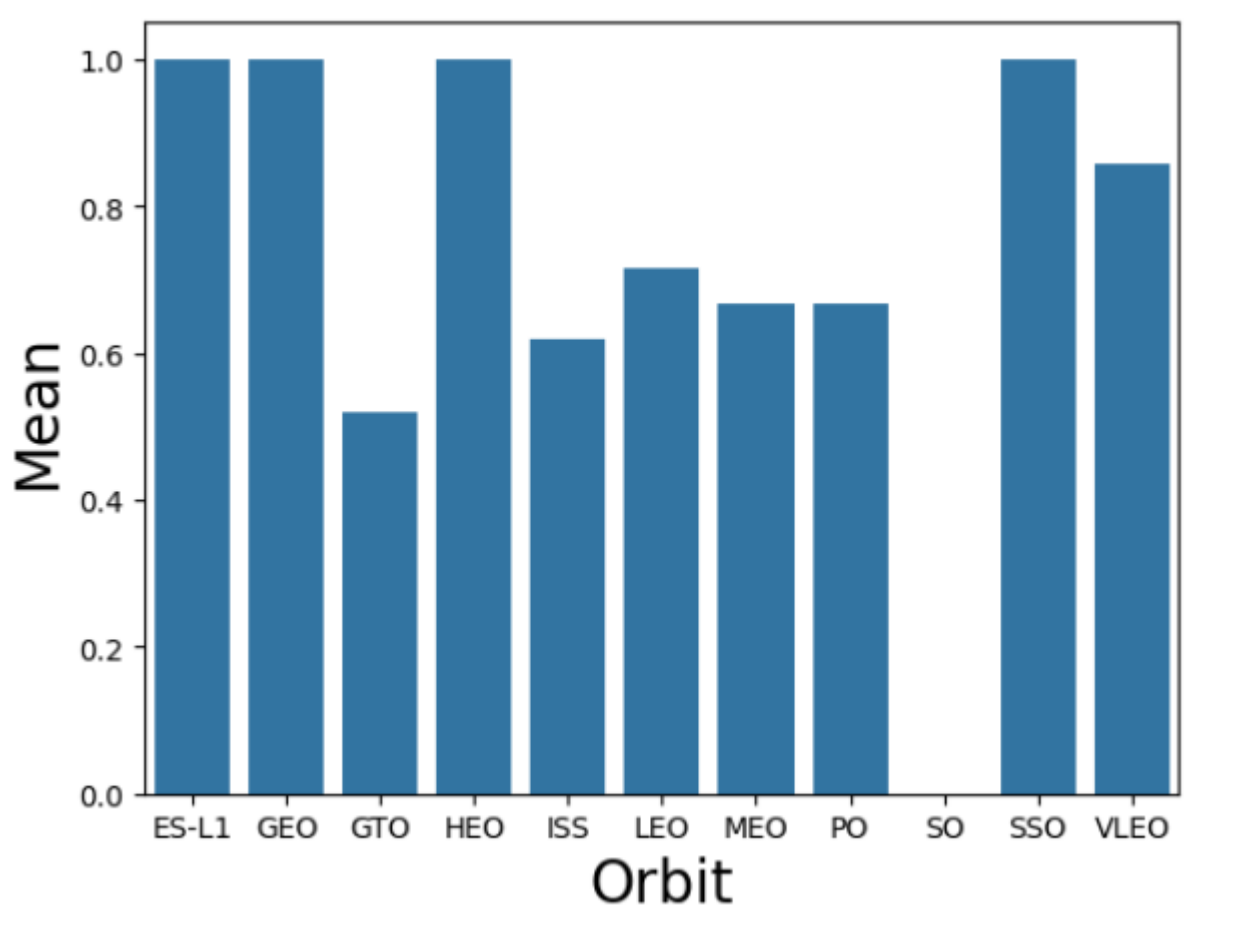
- For launch site 'KSC LC 39A', it takes around 25 launches before a first successful launch
- Success rates increases as the number of flights increase

Payload vs. Launch Site



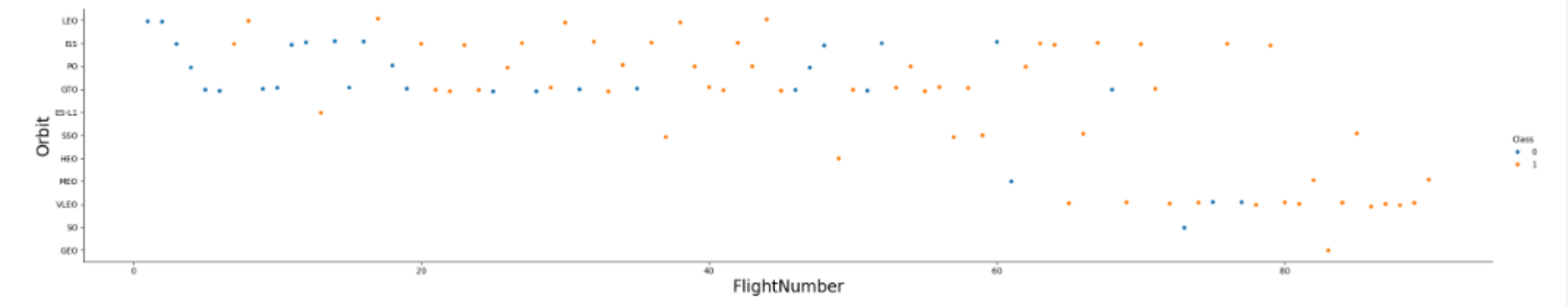
- Percentage of successful launches for launch site 'VAFB SLC 4E' as the payload mass increases. However, there are no rockets launched for payload greater than 10,000 kg

Success Rate vs. Orbit Type



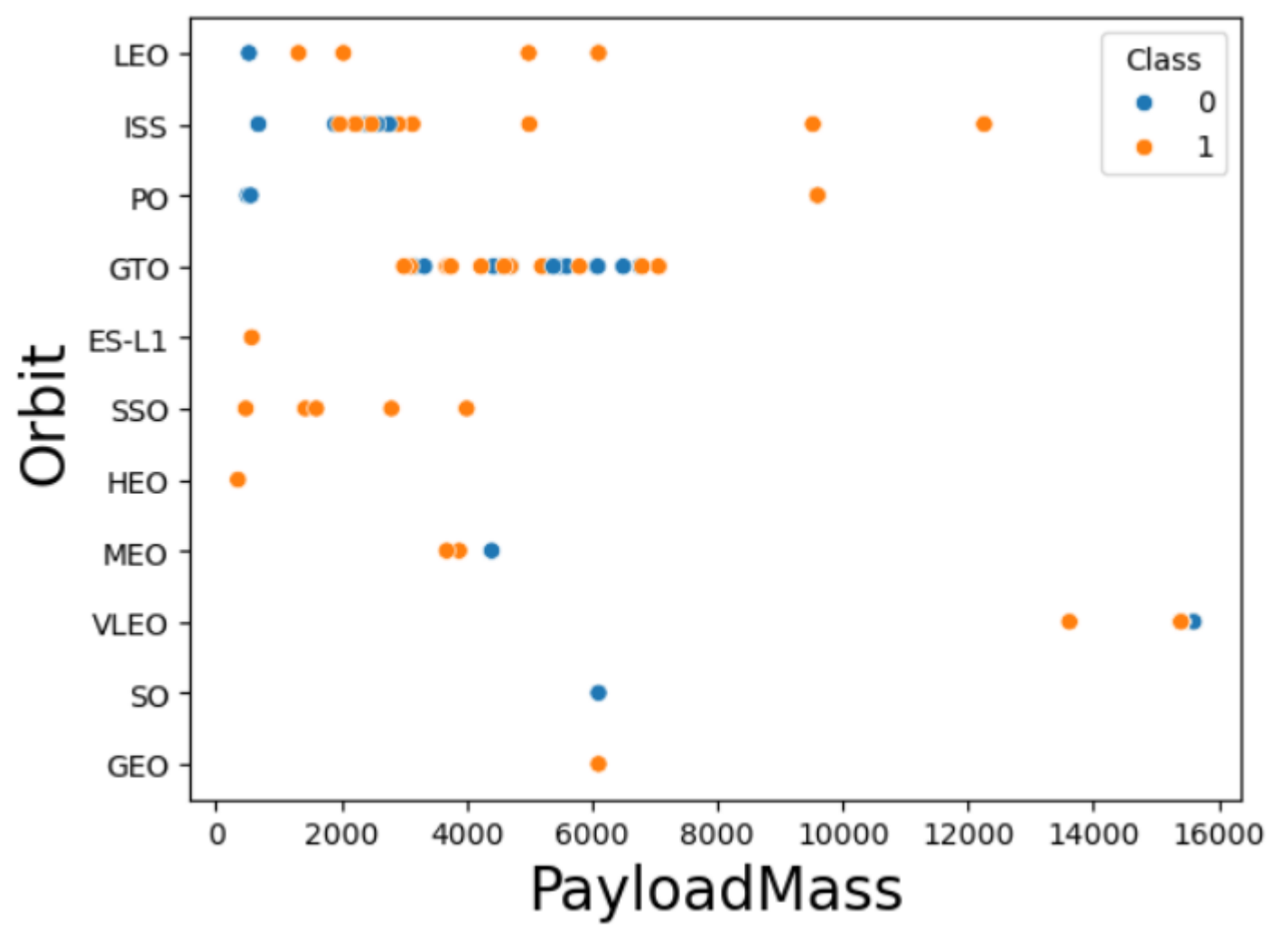
- Orbits ES-LI, GEO, HEO, and SSO have the highest success rates

Flight Number vs. Orbit Type



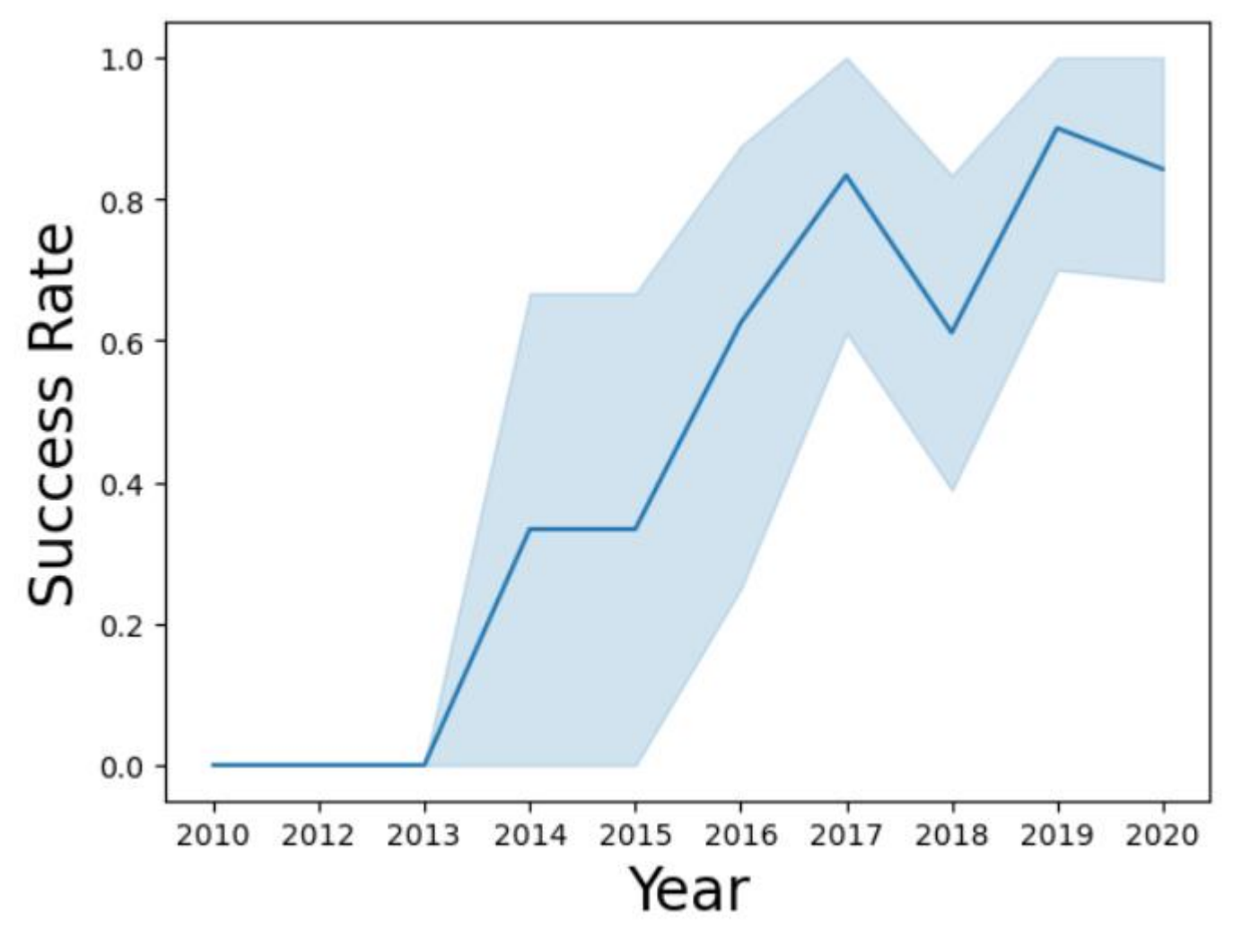
- For most orbits (LEO, ISS, PO, SSO, MEO, VLEO) successful landing rates appear to increase with flight numbers
- For orbit VLEO, first successful landing doesn't occur until 60+ number of flights

Payload vs. Orbit Type



- The successful landing rates appear to increase with pay load for orbits LEO, ISS, PO, and SSO

Launch Success Yearly Trend



- Success rate remained the same between 2010 and 2013 as well as between 2014 and 2015
- The success rate decreased between 2017 and 2018 and between 2019 and 2020

All Launch Site Names

Query: %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;

Result:

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

Explanation: ‘distinct’ returns only unique values and there are 4 unique launch sites.

Launch Site Names Begin with 'CCA'

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
6/4/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
12/8/2010	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
10/8/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
3/1/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation: Using ‘Like’ with ‘CCA%’ returns records where the ‘Launch_Site’ name starts with “CCA”. Limit 5, limits the number of returned records to 5.

Total Payload Mass

Query: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload mass by NASA (CRS)"
FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';

Result:

Total Payload mass by NASA (CRS)
45596

Explanation: Sum() calculates the sum and returns the total.

Average Payload Mass by F9 v1.1

Query: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average payload mass by Booster Version F9 v1.1" FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';

Result:

Average payload mass by Booster Version F9 v1.1
2928.4

Explanation: Avg() returns the average.

First Successful Ground Landing Date

Query: %sql SELECT MIN(DATE) AS "Date of first successful landing outcome in ground pad"
FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';

Result:

Date of first successful landing outcome in ground pad
1/8/2018

Explanation: We use Min() here to find the first date.

Successful Drone Ship Landing with Payload between 4000 and 6000

Query: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

Result:

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

Explanation: The above query returns booster versions that were successful between 4000 AND 6000.

Total Number of Successful and Failure Mission Outcomes

Query: %sql SELECT number_of_success_outcomes AS "Number of successful outcomes" , number_of_failure_outcomes AS "Number of failure outcomes" FROM (SELECT COUNT(*) AS number_of_success_outcomes FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%') success_table, (SELECT COUNT(*) number_of_failure_outcomes FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Failure%') failure_table

Result:

Number of successful outcomes	Number of failure outcomes
100	1

Explanation: Count() is used to count the data in a particular column and Group BY is used to arrange identical data in a column. Subqueries are used to get the count.

Boosters Carried Maximum Payload

Query: %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

Result:

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

Explanation: The main query returns booster versions and respective payload mass where payload mass is maximum with value of 15600. The sub query returns the maximum payload mass by using Max() on the pay load mass column.

2015 Launch Records

Query: %sql SELECT substr(Date,-7,2) AS "Month" , LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL where LANDING_OUTCOME = 'Failure (drone ship)' AND (substr(Date,7,4) = '2015' OR substr(Date,6,4) = '2015');

Result:

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Explanation: The output contains the month, landing outcome, booster version, and the launch site where landing outcome is failed in drone ship and the year is 2015.

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

Query: %sql SELECT landing_outcome, count(Landing_Outcome) from SPACEXTBL where DATE BETWEEN '04/06/2010' and '20/03/2017'group by Landing_Outcome order by count(Landing_outcome) desc;

Result:

Landing_Outcome	Count
Success	15
No attempt	6
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	1

Explanation: The output of the query is a ranked list of landing outcome counts per the specified date range.

Section 3

Launch Sites Proximities Analysis

SpaceX Falcon9 – All Launch Sites on Map

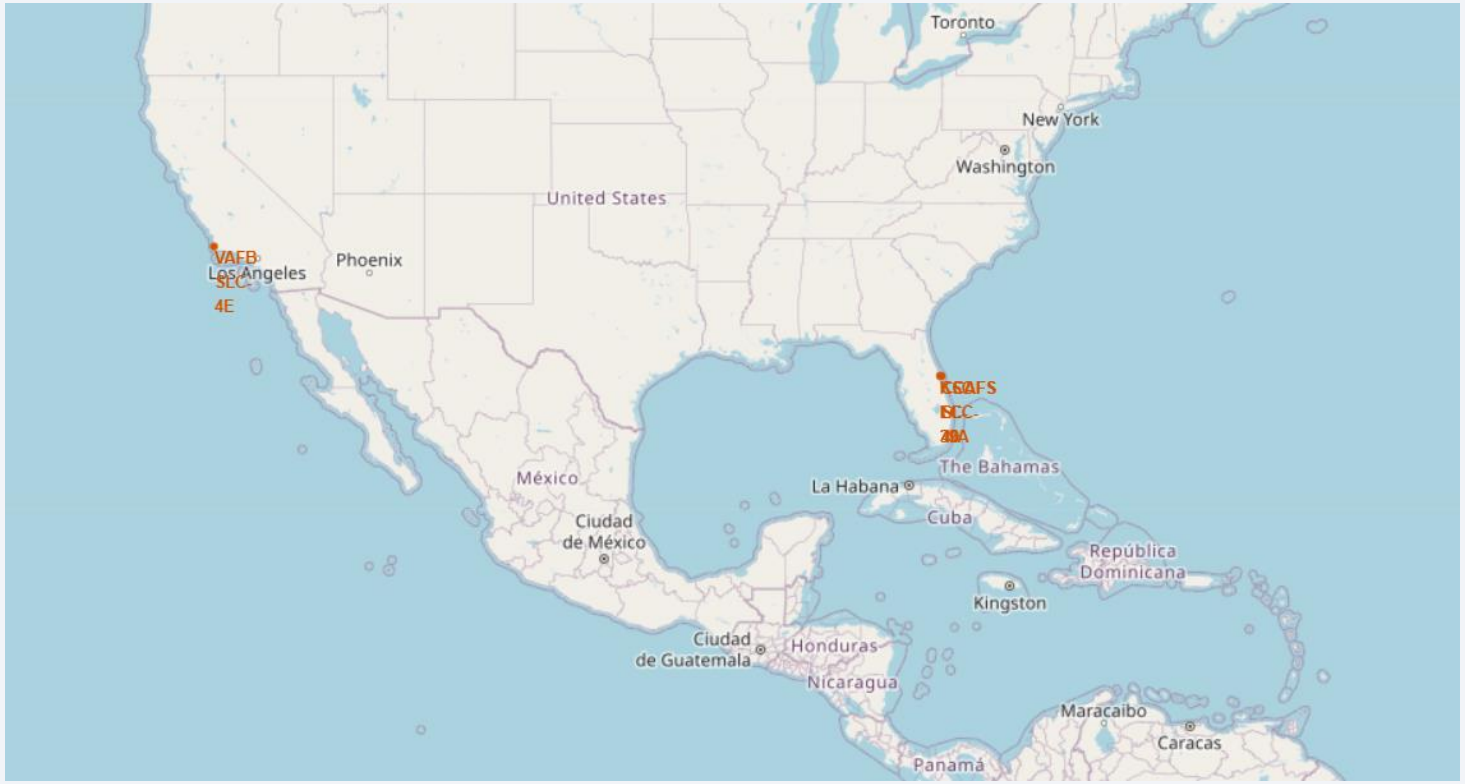
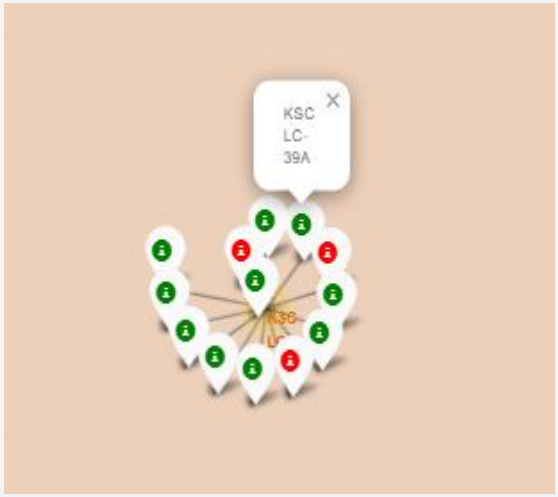
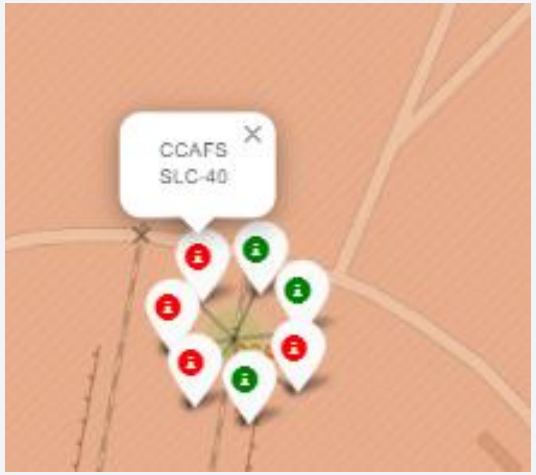


Figure on the left displays the Global map with Falcon 9 launch sites that are located in the United States (in California and Florida). Each launch site contains a circle, label, and a popup to highlight the location and the name of the launch site. It is can also be seen that all launch sites are near the coast. The launch sites on display are:

- VAFB SLC-4E (CA)
- CCAFS LC-40 (FL)
- KSC LC-39A (FL)
- CCAFS SLC-40 (FL)

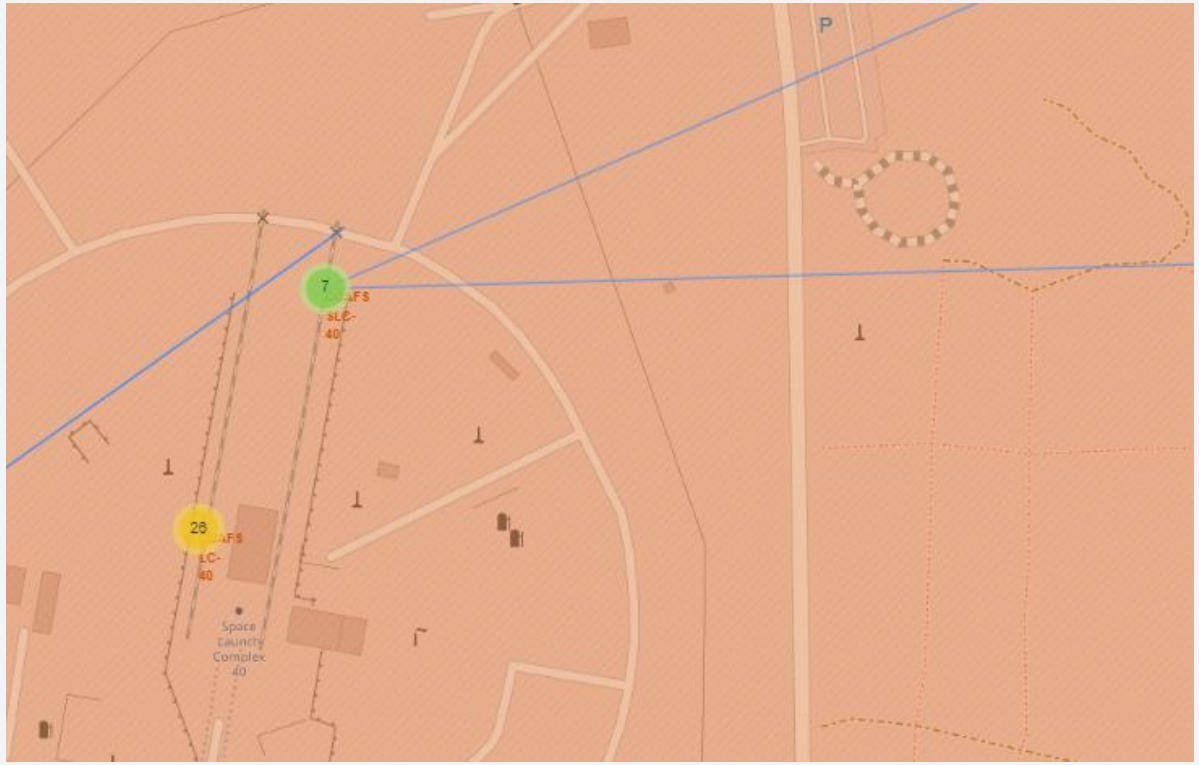
SpaceX Falcon9 – Successful/Failed launches for all Launch Sites



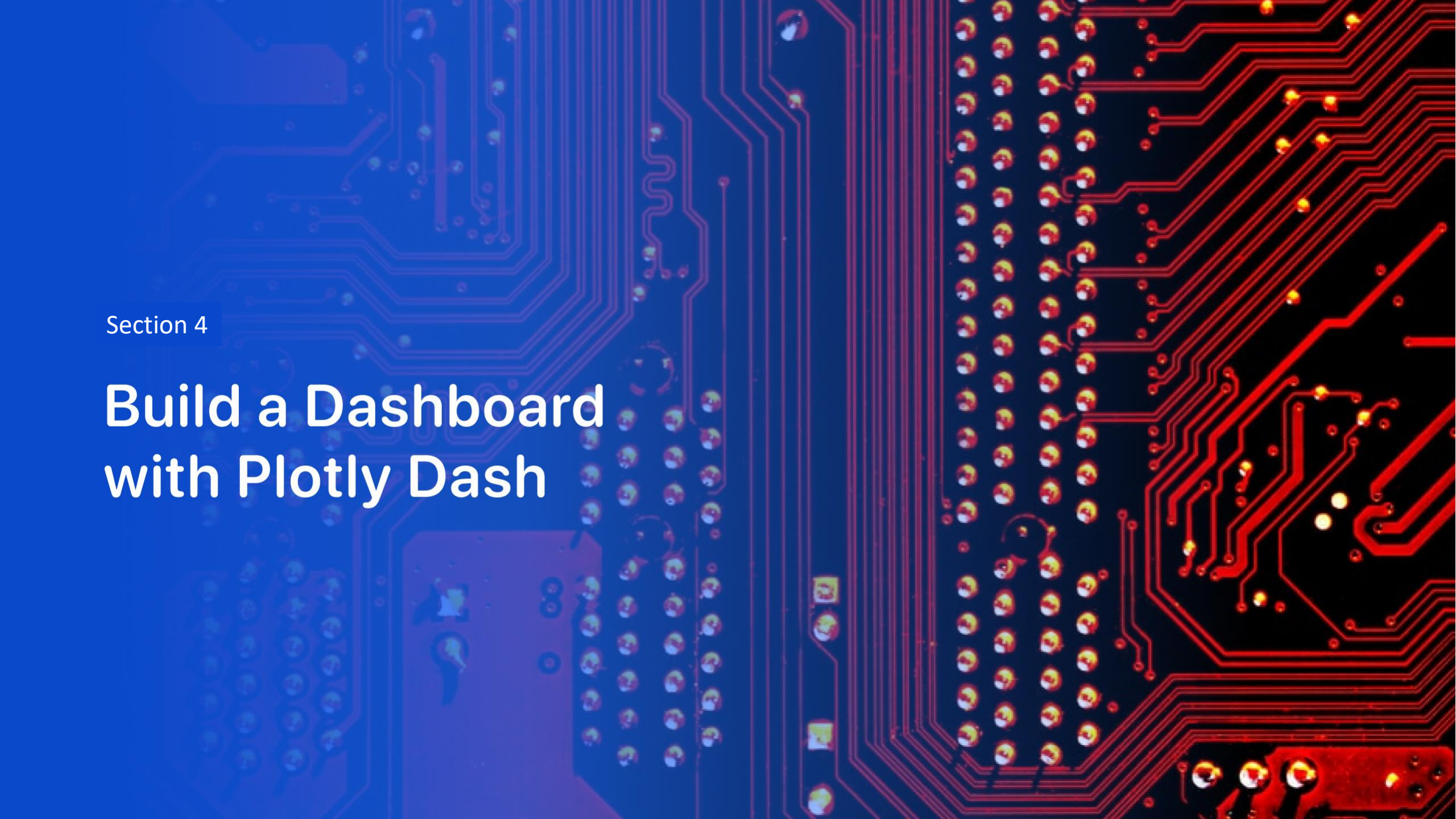


If we zoom in on one of the launch site, we can see green and red markers. Each green marker represents a successful launch while each red marker represents a failed launch. It can be noticed that the KSC LC-39A Launch Site has the greatest number of successful launches.

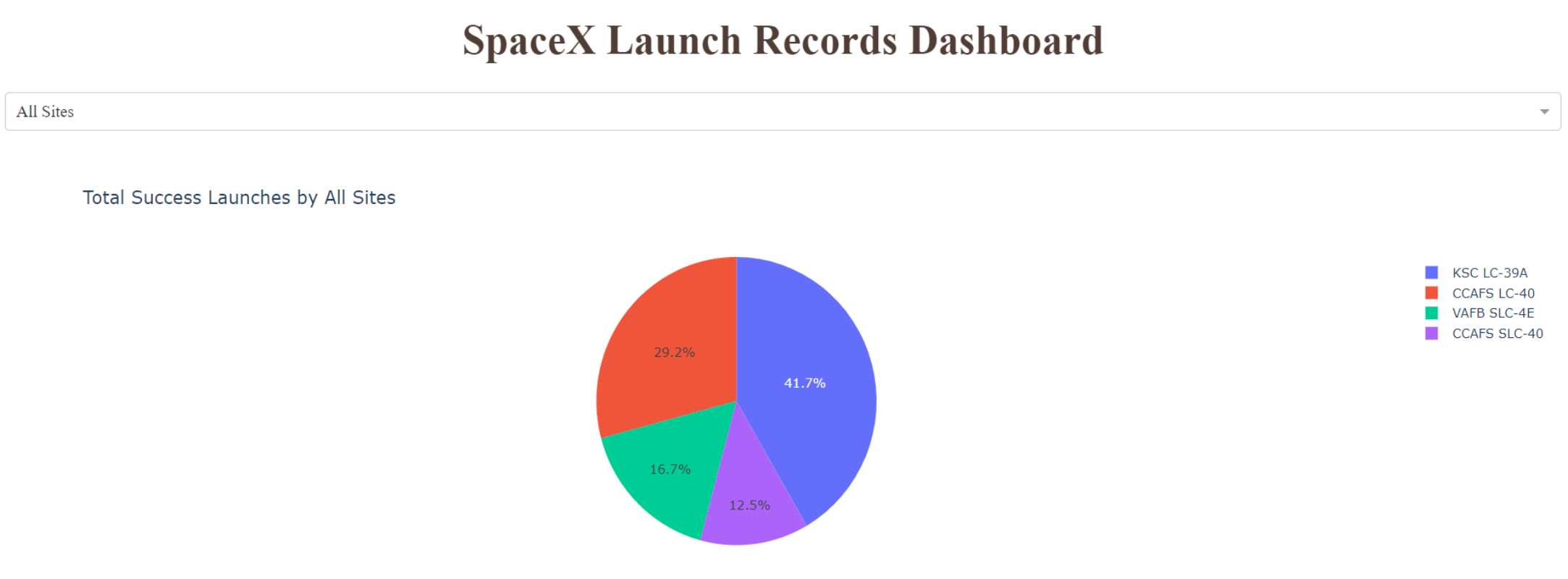
SpaceX Falcon9 – Launch Sites to closest proximities



The figure provides a zoom in view into other proximities such as coastline, railway, and highway with respective distances and lines from the Launch Site.

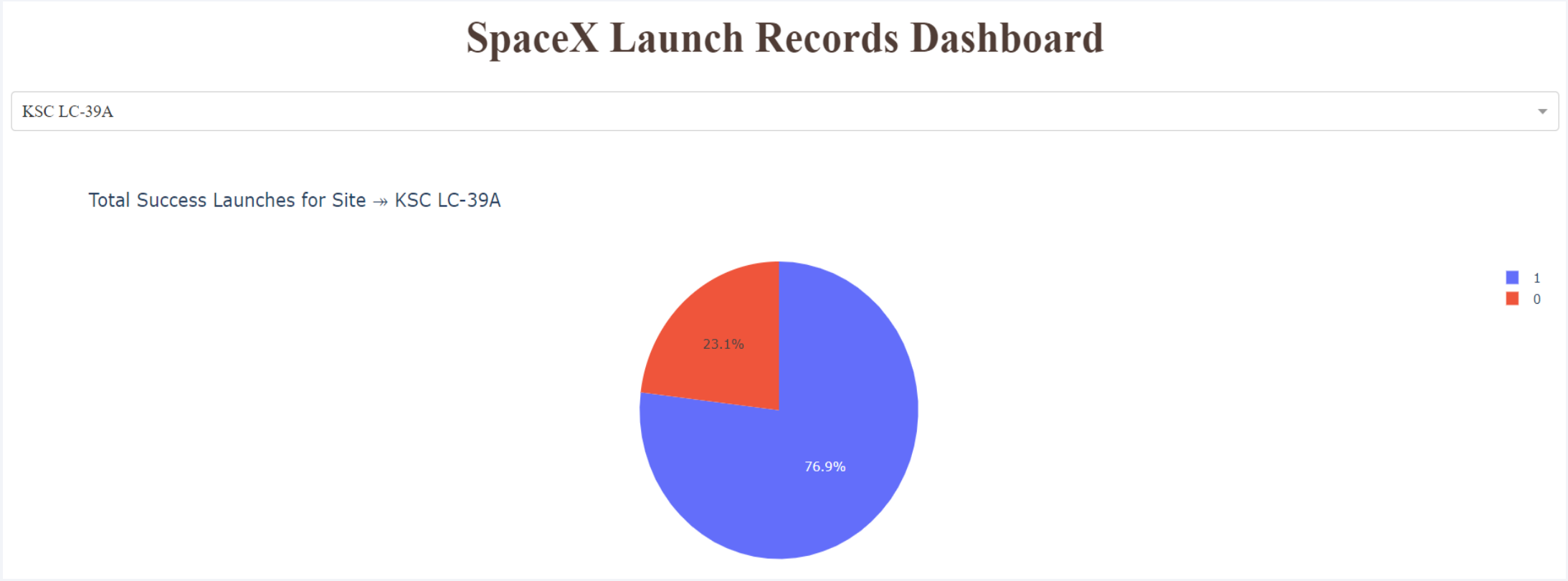

Section 4
Build a Dashboard with Plotly Dash

Successful Launches for all Launch Sites



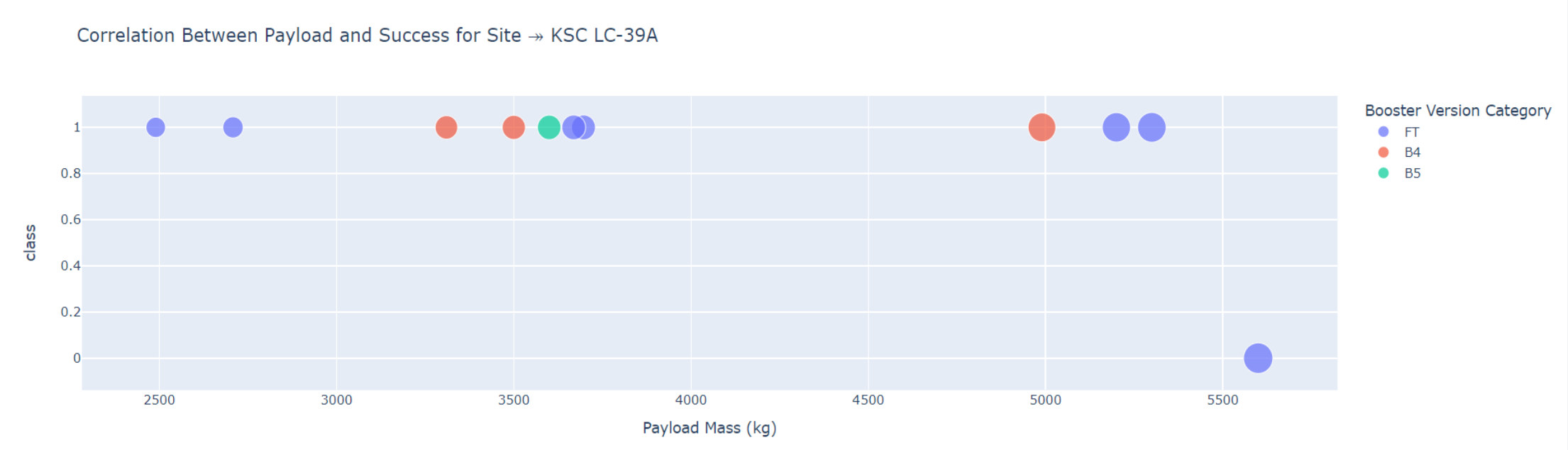
Launch Site ‘CCAFS SLC40’ has the lowest launch success rate and Launch Site ‘KSC LC-39A’ has the highest launch success rate.

Highest Launch Success Ratio



KSC LC-39A Launch Site has the highest launch success rate. Launch success rate is 76.9%.

Payload vs Launch Outcome Plot

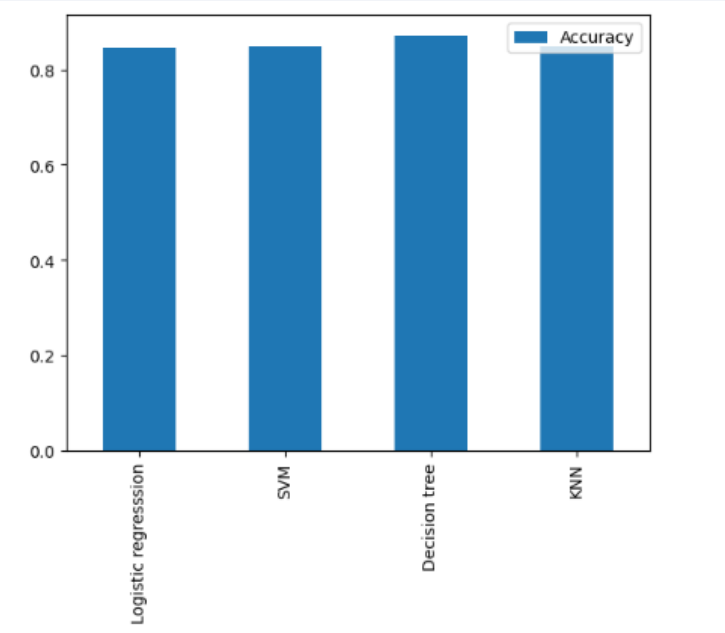


Most successful launches are in the payload range from 2000 to about 5500 and Booster version category ‘FT’ has the most successful launches.

Section 5

Predictive Analysis (Classification)

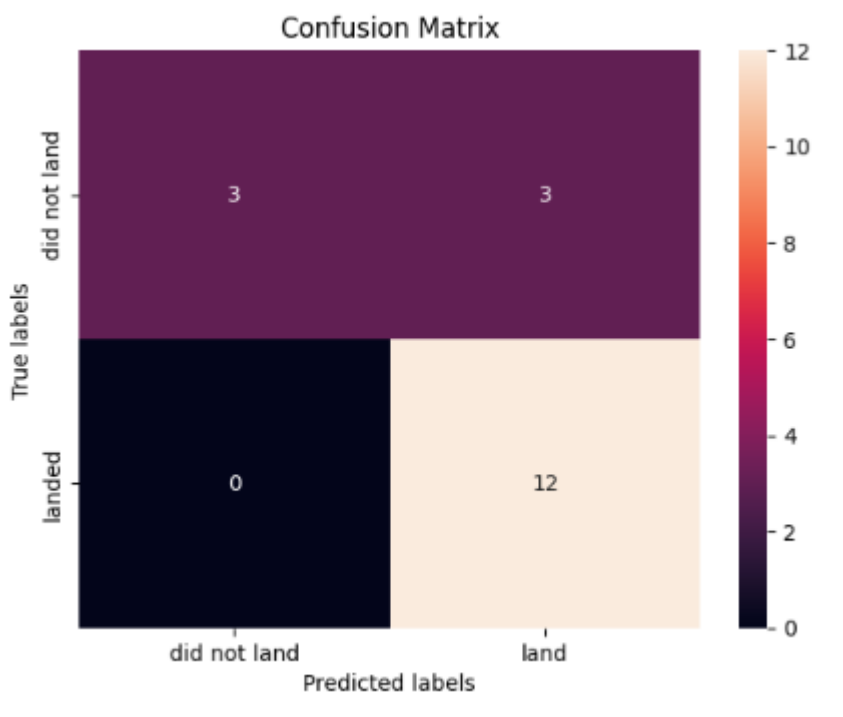
Classification Accuracy



	Accuracy
Logistic regresssion	0.846429
SVM	0.848214
Decision tree	0.875000
KNN	0.848214

Putting the results of all 4 models side by side, we can see that they all share the same accuracy score and confusion matrix when tested on the test set. Therefore, their GridSearchCV best scores are used to rank them instead. Based on the GridSearchCV best scores, we can find that Decision Tree algorithm has the highest best score with a value of 0.875.

Confusion Matrix



The confusion matrix is same for all the models (Logistic Regression, Support Vector Machines, Decision Tree, K Nearest Neighbors). The classifier made 18 predictions: 12 times it was True positive, 3 times it was True negative and 3 times it was False positive. Accuracy Score on the test data is the same for all the classification algorithms based on the data set with a value of 0.83

Conclusions

- As part of this capstone project, we try to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch.
- It was also found that each feature of a Falcon 9 launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way
- Various information was gathered such as Launch success rate increased by about 80% from 2013 to 2020, Launch Site 'KSC LC-39A' has the highest launch success rate and Launch Site 'CCAFS SLC40' has the lowest launch success rate, Launch sites are located strategically away from the cities and closer to coastline, railroads, and highways and many more information was found.
- The predictive model produced by decision tree algorithm performed the best among the 4 machine learning algorithms employed.

Appendix

The link to every python notebook can be found below.

LINK

Thank you!