SpaceX With Data Science

Subrat Nanda

11/05/2023

Outline

Executive Summary

❑ Summary of Methodologies :

The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies were used:

- Collect data using SpaceX REST API and web scraping techniques
- Wrangle data to create a success/fail outcome variable
- Explore data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- Analyze the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- Explore launch site success rates and proximity to geographical markers
- Visualize the launch sites with the most successful and successful payload ranges
- Build Models to predict landing outcomes using logistic regression, support vector
- machine (SVM), decision tree, and K-nearest neighbor (KNN)

Cont..

❑ **Results :**

Exploratory Data Analysis:

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

Visualization/Analytics:

- Most launch sites are near the equator, and all are close to the coast

Predictive Analytics:

- All models performed similarly on the test set. The decision tree model slightly outperformed

For more information, find the full project repository:
https://github.com/Subrat-Nanda/IBM-Data-Science-Professional-Certification/Project

Introduction

SpaceX is a revolutionary company that has disrupted the space industry by offering rocket launches specifically Falcon 9 as low as 62 million dollars; while other providers cost upward of 165 million dollars each. Most of these savings is thanks to SpaceX's astounding idea to reuse the first stage of the launch by re-land the rocket to be used on the next mission. Repeating this process will make the price down even further. As a data scientist of a startup rivaling SpaceX, the goal of this project is to create a machine learning pipeline to predict the landing outcome of the first stage in the future. This project is crucial in identifying the right price to bid against SpaceX for a rocket launch.

The problems included:

- Identifying all factors that influence the landing outcome.
- The relationship between each variable and how it is affecting the outcome.
- The best condition needed to increase the probability of a successful landing.

Methodology

Methodology

Executive Summary

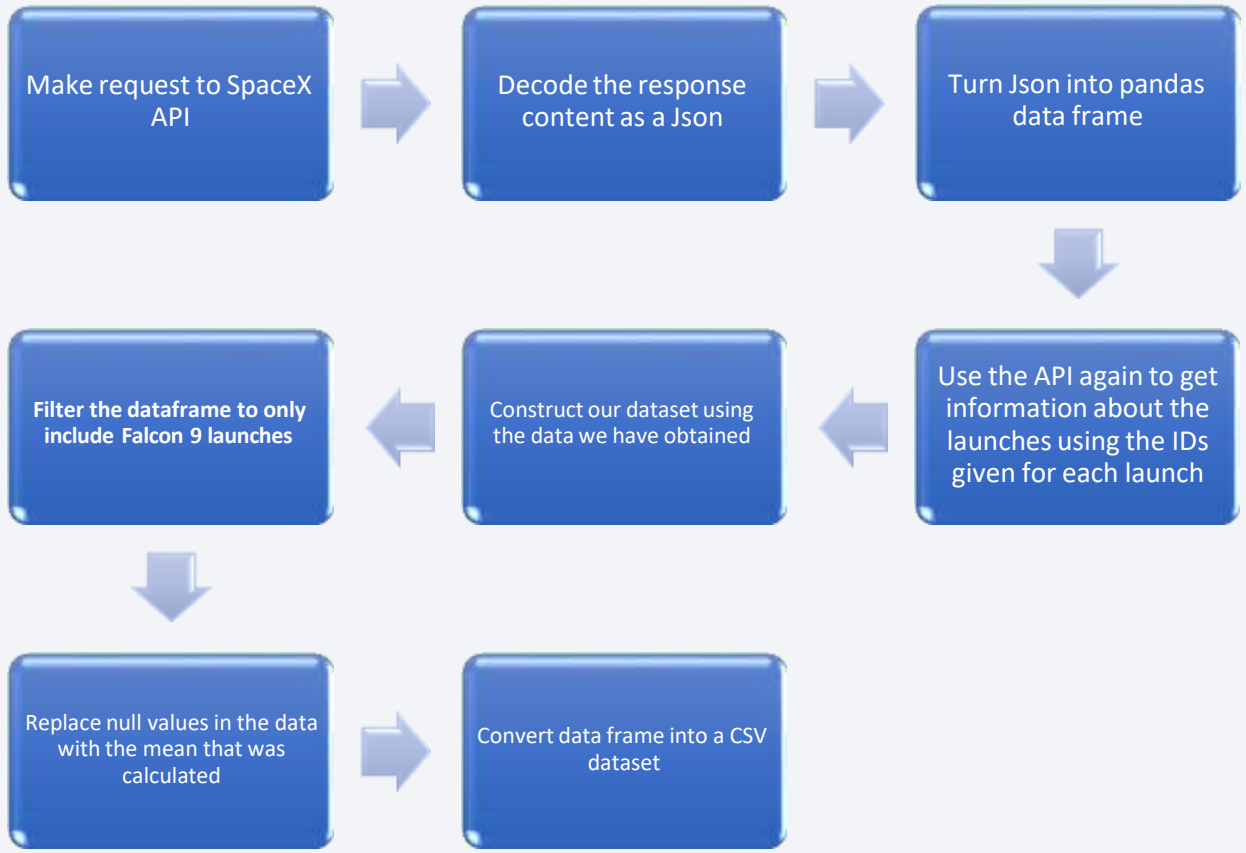
Steps :

- **Collect** data using SpaceX REST API and web scraping techniques.
  - Sources :
  - Space X API (https://api.spacexdata.com/v4/rockets/)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- **Wrangle** data – by filtering the data, handling missing values, and applying one hot encoding – to prepare the data for analysis and modeling
- **Explore** data via EDA with SQL and data visualization techniques
- **Visualize** the data using Folium and ***Plotly Dash***
- **Build Models** to predict landing outcomes using classification models. Tune and evaluate models to find the best model and parameters

Data Collection – SpaceX API

Steps :

- **Request data** from SpaceX API (rocket launch data)
- **Decode response** using .json() and convert to a data frame using .json_normalize()
- **Request information** about the launches from SpaceX API using custom functions
- **Create a dictionary** from the data
- **Create a dataframe** from the dictionary
- **Filter dataframe** to contain only Falcon 9 launches
- **Replace missing values** of Payload Mass with calculated .mean()
- **Export data** to CSV file

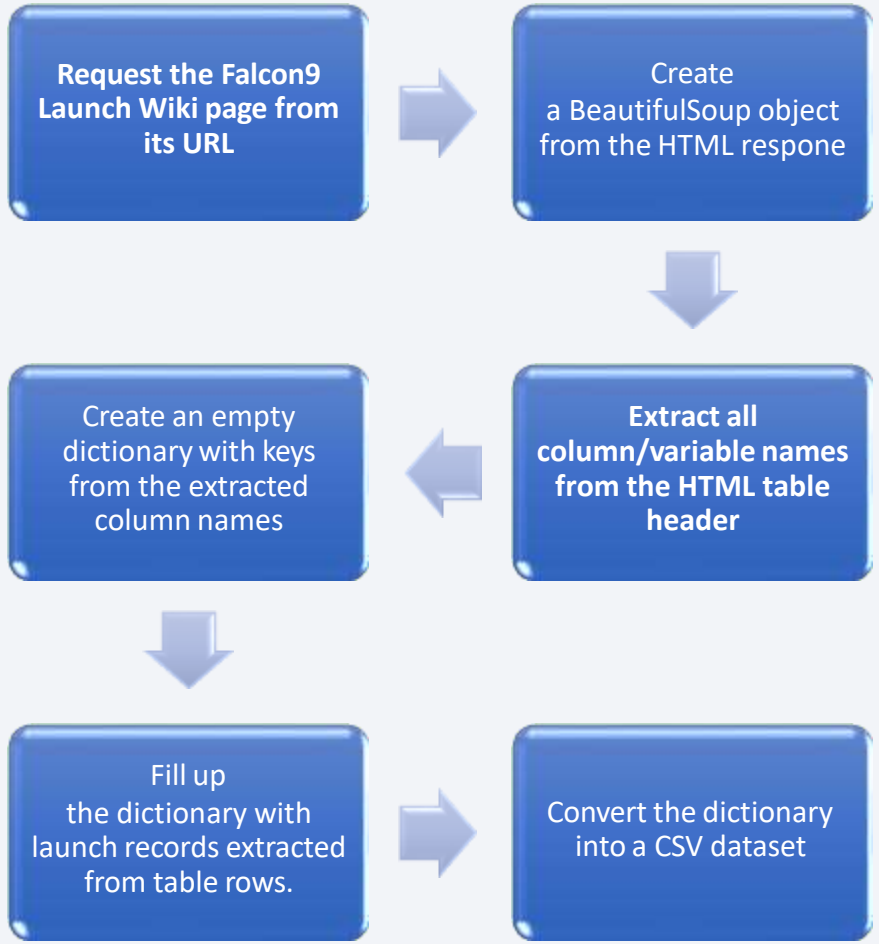


- **URL**: https://github.com/Subrat-Nanda/spacex-data-collection-api.ipynb

Data Collection - Web Scraping

Steps :

- **Request data** (Falcon 9 launch data) from Wikipedia
- **Create BeautifulSoup object** from HTML response
- **Extract column names** from the HTML table header
- **Collect data** from parsing HTML tables
- **Create a dictionary** from the data
- **Create a dataframe** from the dictionary
- **Export data** to CSV file



URL : https://github.com/Subrat-Nanda/Webscraping.ipynb

Data Wrangling

Steps.

- **Perform EDA** and determine data labels
- **Calculate:**
  - # of launches for each site
  - # and occurrence of orbit
  - # and occurrence of mission outcome per orbit type]
- **Create a binary** landing outcome column (dependent variable)
- **Export data** to CSV file

Landing Outcome

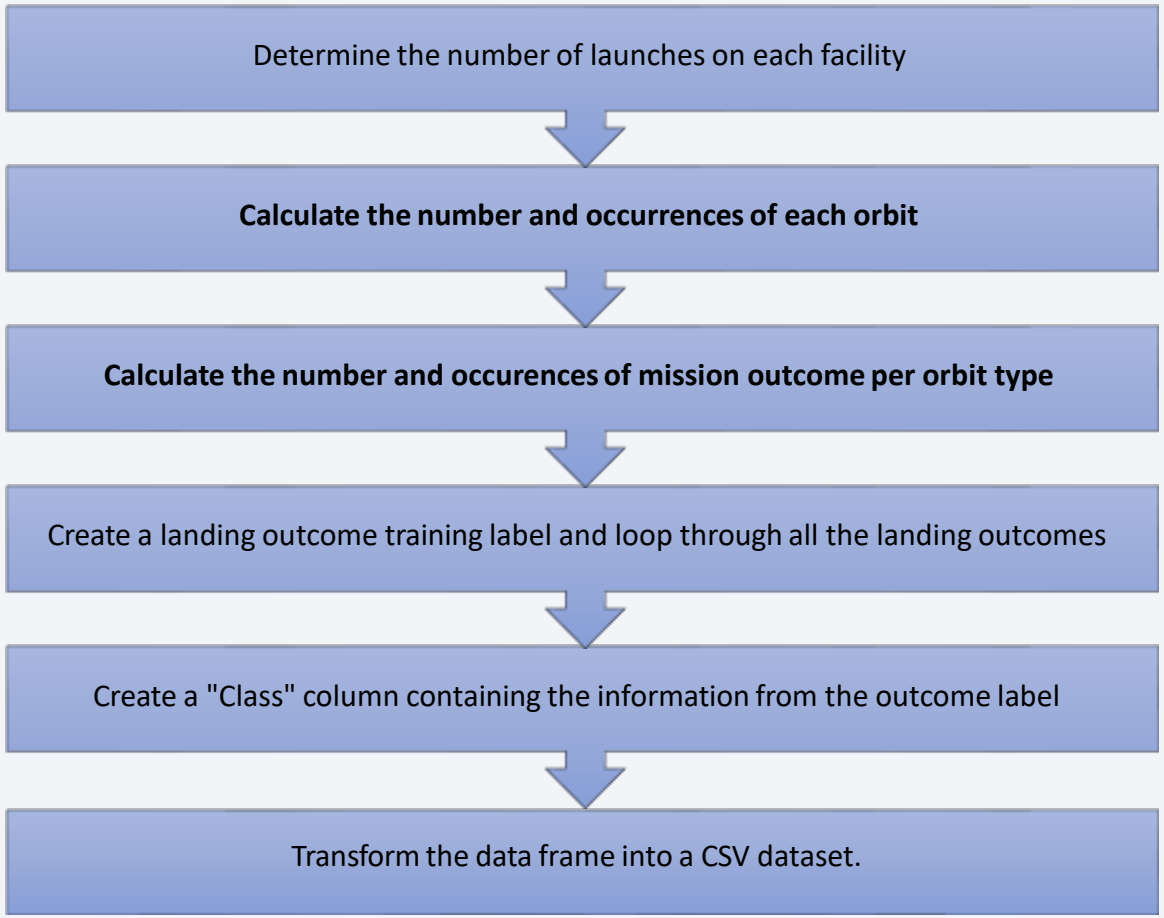
- Landing was not always successful
- **True Ocean:** mission outcome had a successful landing to a specific region of the ocean

Landing Outcome Cont ..

- **False Ocean**: represented an unsuccessful landing to a specific region of the ocean.
- **True RTLS**: meant the mission had a successful landing on a ground pad.
- **False RTLS**: represented an unsuccessful landing on a ground pad.
- **True ASDS**: meant the mission outcome had a successful landing on a drone ship.
- **False ASDS**: represented an unsuccessful landing on the drone ship.
- **Outcomes converted** into -

  1 - successful landing

  0 - unsuccessful landing.

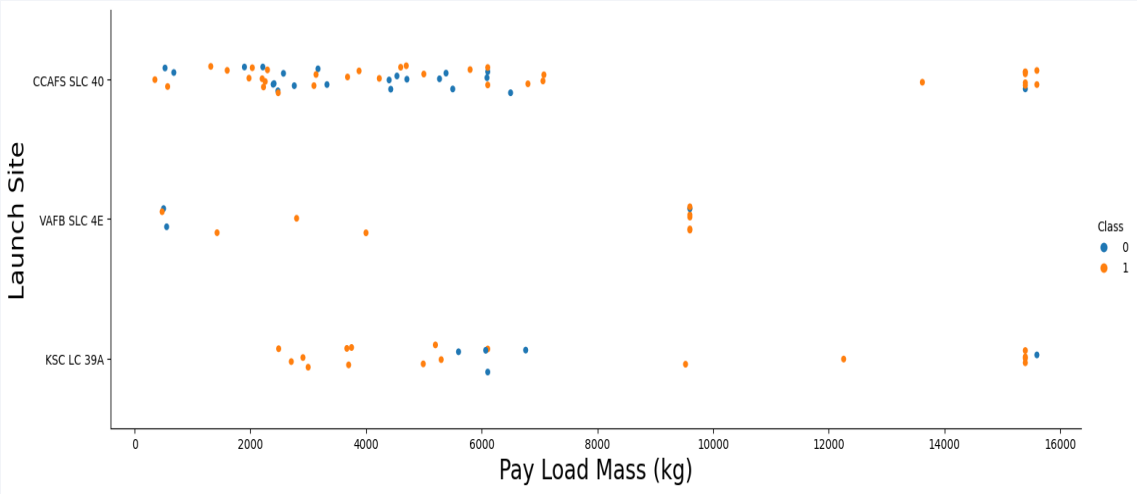


URL: https://github.com/Subrat-Nanda/Data_wrangling.ipynb

EDA with Data Visualization

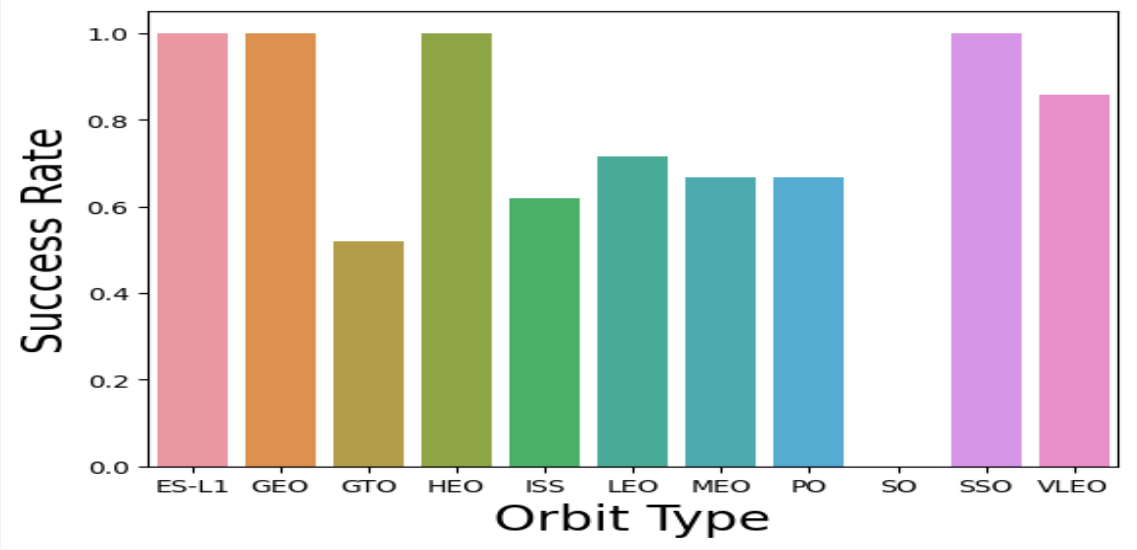
Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type



Analysis

- **View the relationship** by using **scatter plots**. The variables could be useful for machine learning if a relationship exists
- **Show comparisons** among discrete categories with **bar charts**. Bar charts show the relationships among the categories and a measured value



URL :
https://github.com/Subrat-Nanda/EDA_dataviz.ipynb

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begins with the string 'CCA';
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in the ground pad was achieved;
- Names of the boosters which have success in drone ships and have payload mass between 4000 and 6000 kg;
- Total number of successful and failed mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ships, their booster versions, and launch site names for the year 2015; and
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20.
- Source code: **https://github.com/Subrat-Nanda/EDA_SQL.ipynb**

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- Added a **Blue** circle at **NASA Johnson Space Center's** coordinate with a **popup label** showing its name using its latitude and longitude coordinates.
- Added **Red** circles at all launch **site coordinates** with a **popup label** showing its name using its latitude and longitude coordinates

Colored Markers of Launch Outcomes

- Added colored markers of successful (**Green**) and unsuccessful (**Red**) launches at each launch site to show which launch sites have high success rates

Distances Between a Launch Site to Proximities

- Added **colored lines** to **show the distance between** launch site **CCAFS SLC 40** and its proximity to the **nearest coastline, railway, highway, and city.**

Dashboard with Plotly Dash

Launch Sites Dropdown List:
- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Sites):
- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

A slider of Payload Mass Range:
- Added a slider to select the Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
- Added a scatter chart to show the correlation between Payload and Launch Success.

Source Code : SpaceX_Dashboard_with_Plotly_Dash

Predictive Analytics (Classification)

Charts

• **Create** a NumPy array from the Class column

• **Standardize** the data with StandardScaler.Fit & transform the data.

• **Split** the data using train_test_split

• **Create** a GridSearchCV object with cv=10 for parameter optimization

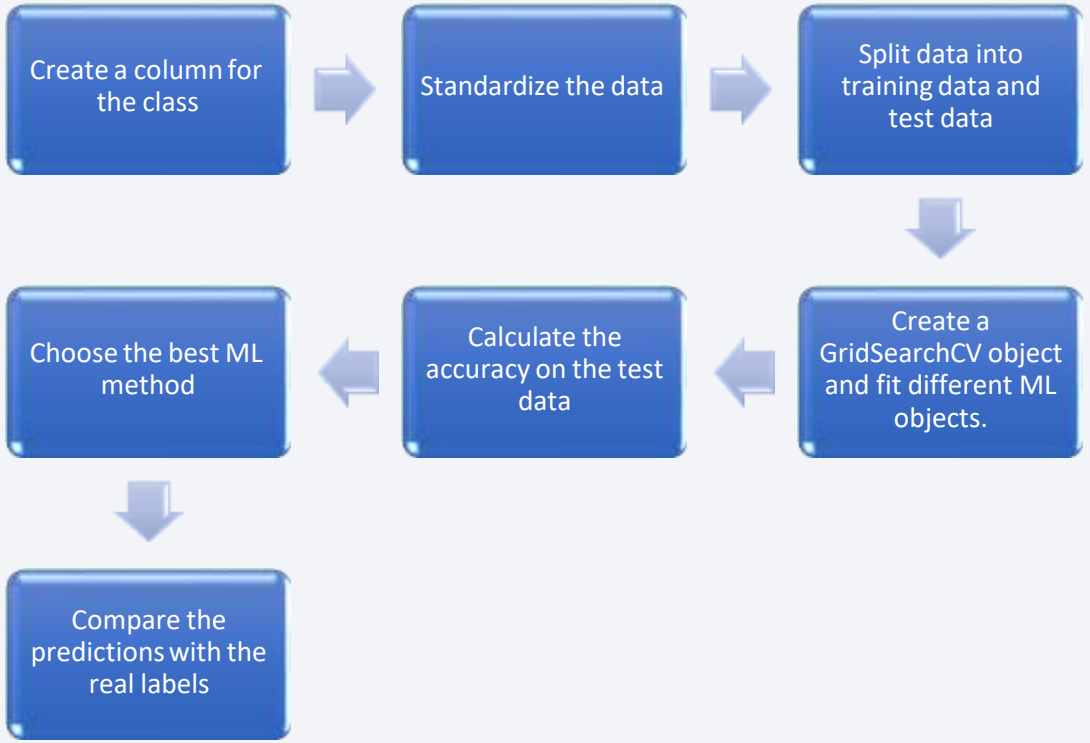
• **Apply** GridSearchCV on different algorithms:

  - logistic regression (LogisticRegression())
  - support vector machine (SVC())
  - decision tree (DecisionTreeClassifier())
  - K-Nearest Neighbor (KNeighborsClassifier())

• **Calculate** accuracy on the test data using .score() for all models

• **Assess** the confusion matrix for all models

• **Identify** the best model using Jaccard_Score, F1_Score & Accuracy



Results Summary

Exploratory Data Analysis :

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate

Visual Analytics :

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics :

- Decision Tree model is the best predictive model for the dataset

Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis

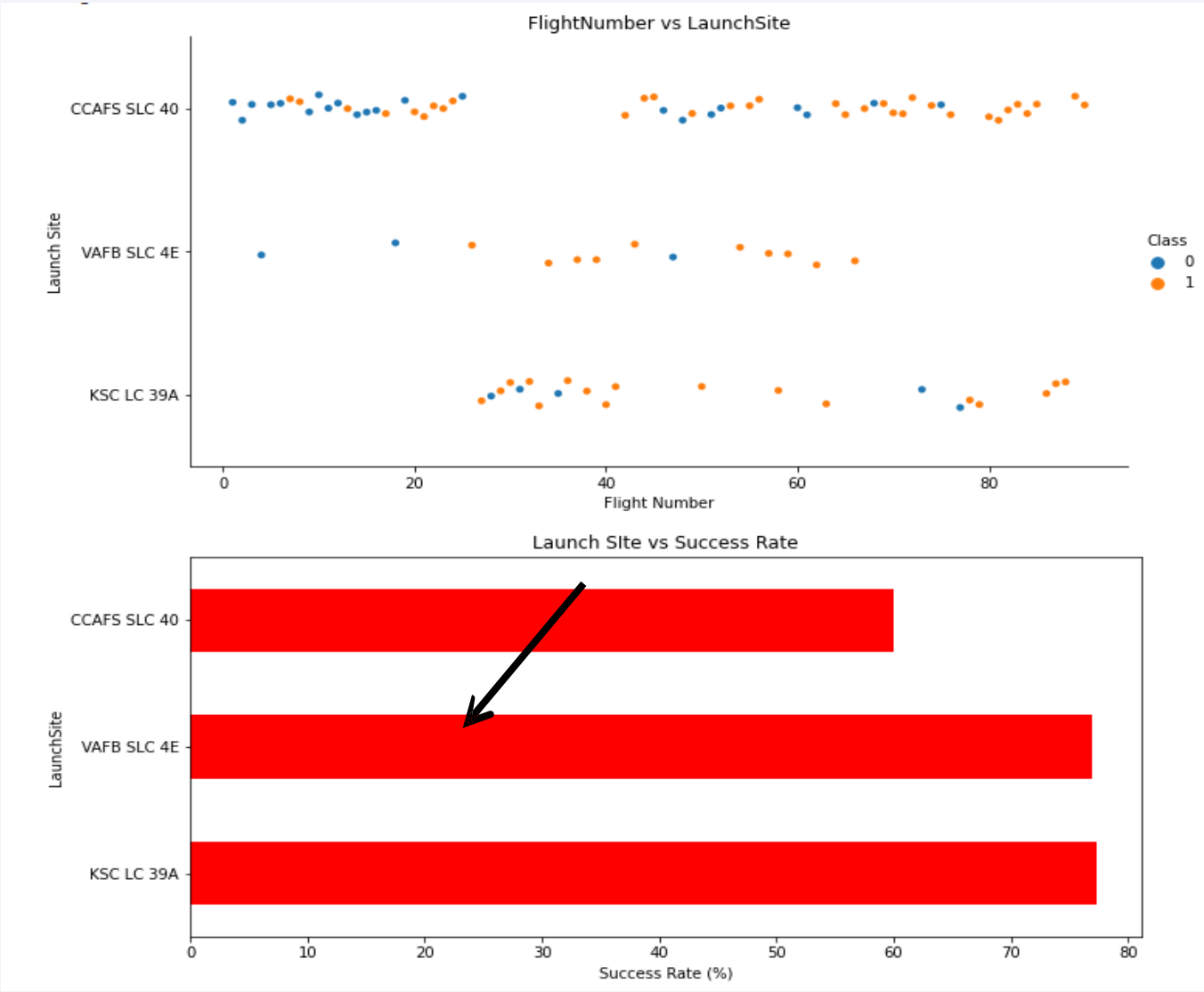
• **Earlier flights** had a **lower success rate** (**blue = fail**).

• **Later flights** had a **higher success rate** (**orange = success**).

• Around half of the launches were from CCAFS SLC 40 launch site.

• VAFB SLC 4E and KSC LC 39A have higher success rates.

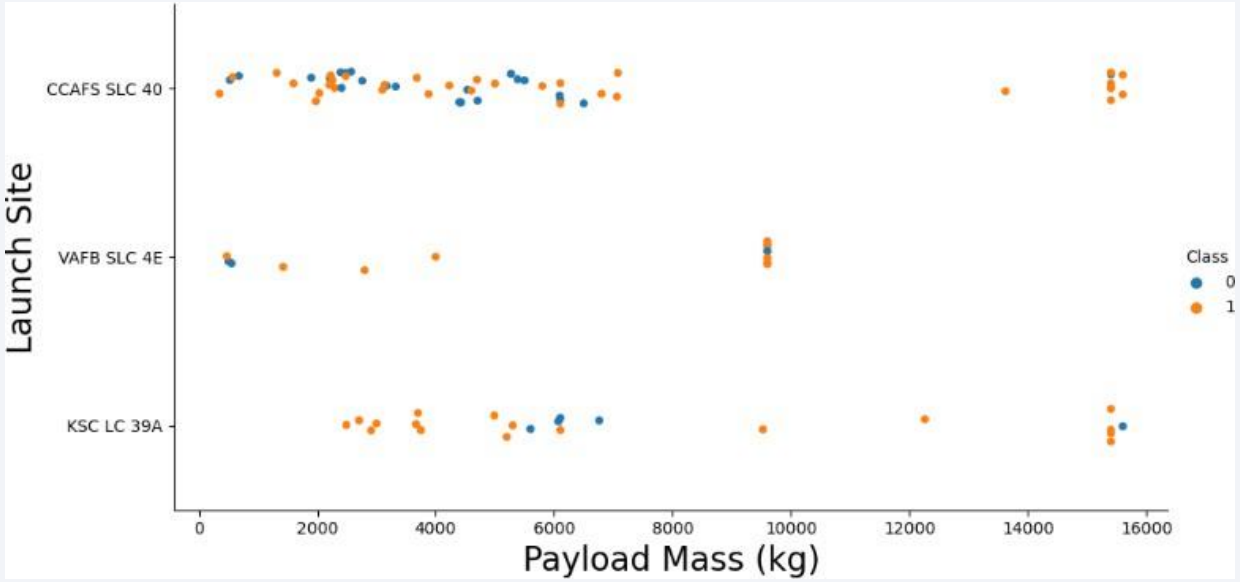
• We can infer that new launches have a higher success rate.



Payload vs. Launch Site

Exploratory Data Analysis

- Typically, the **higher** the **payload mass** (kg), the **higher** the **success rate.**
- Most launches with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



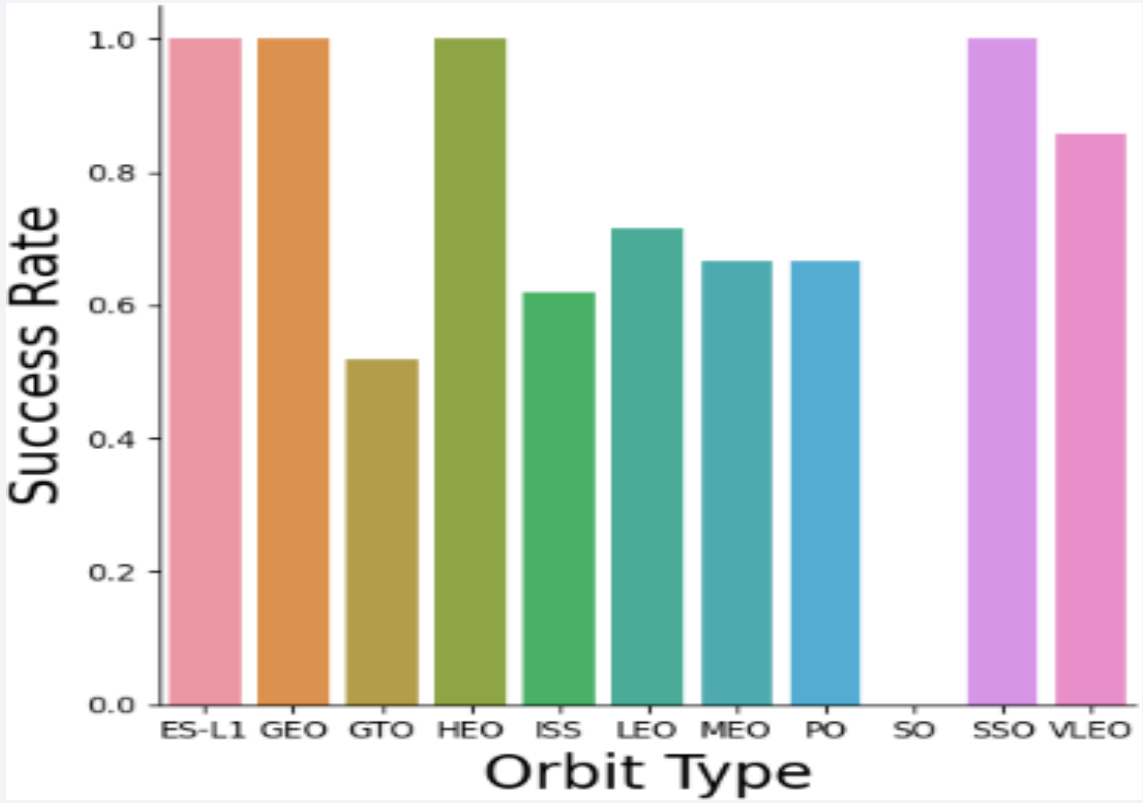
Success Rate vs. Orbit Type

The biggest success rates happen to orbits:

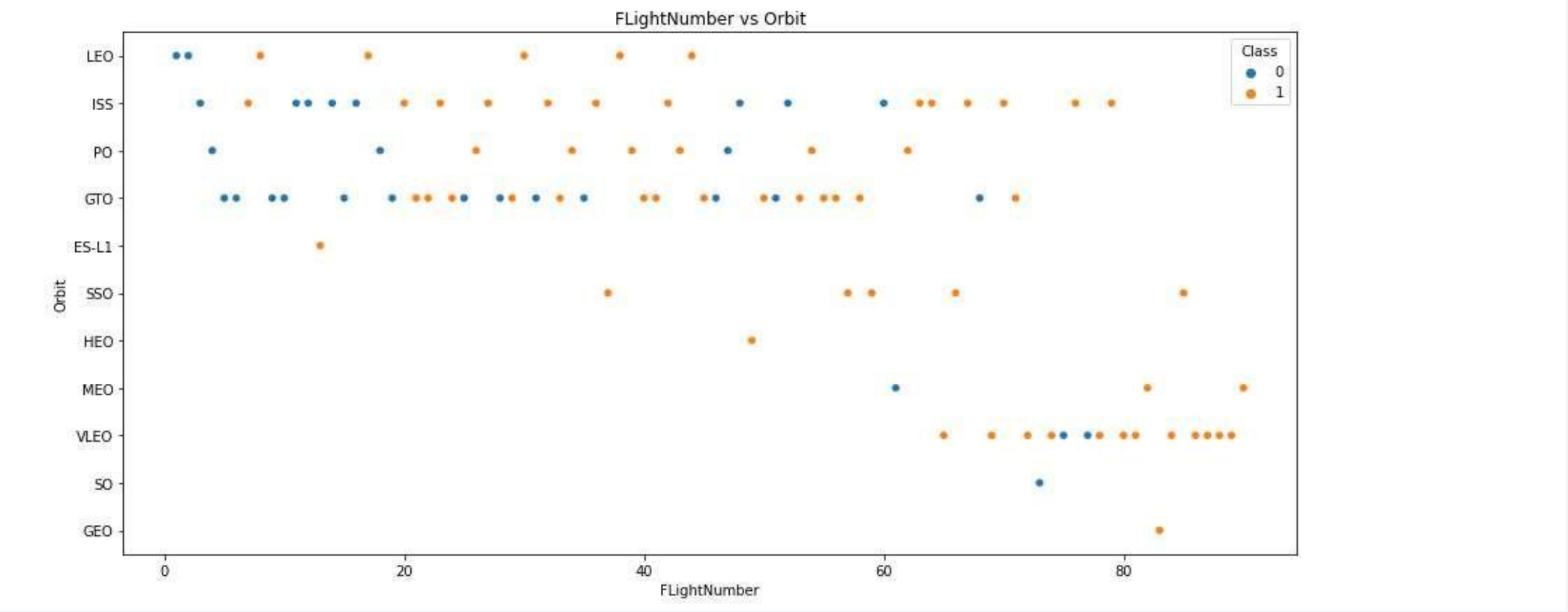
- ES-L1;
- GEO;
- HEO; and
- SSO.

Followed by:

- VLEO (above 80%); and
- LFO (above 70%).

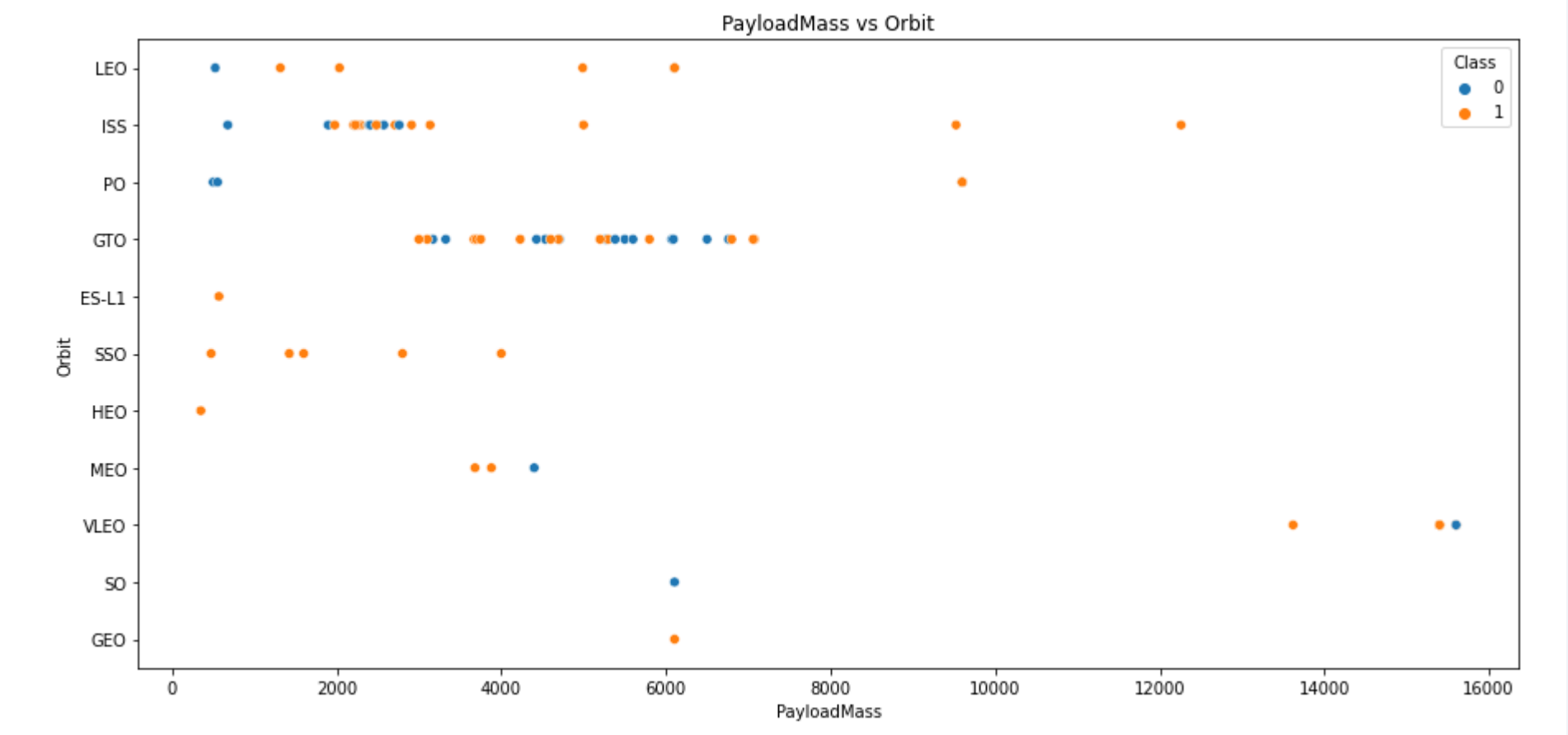


Flight Number vs. Orbit Type



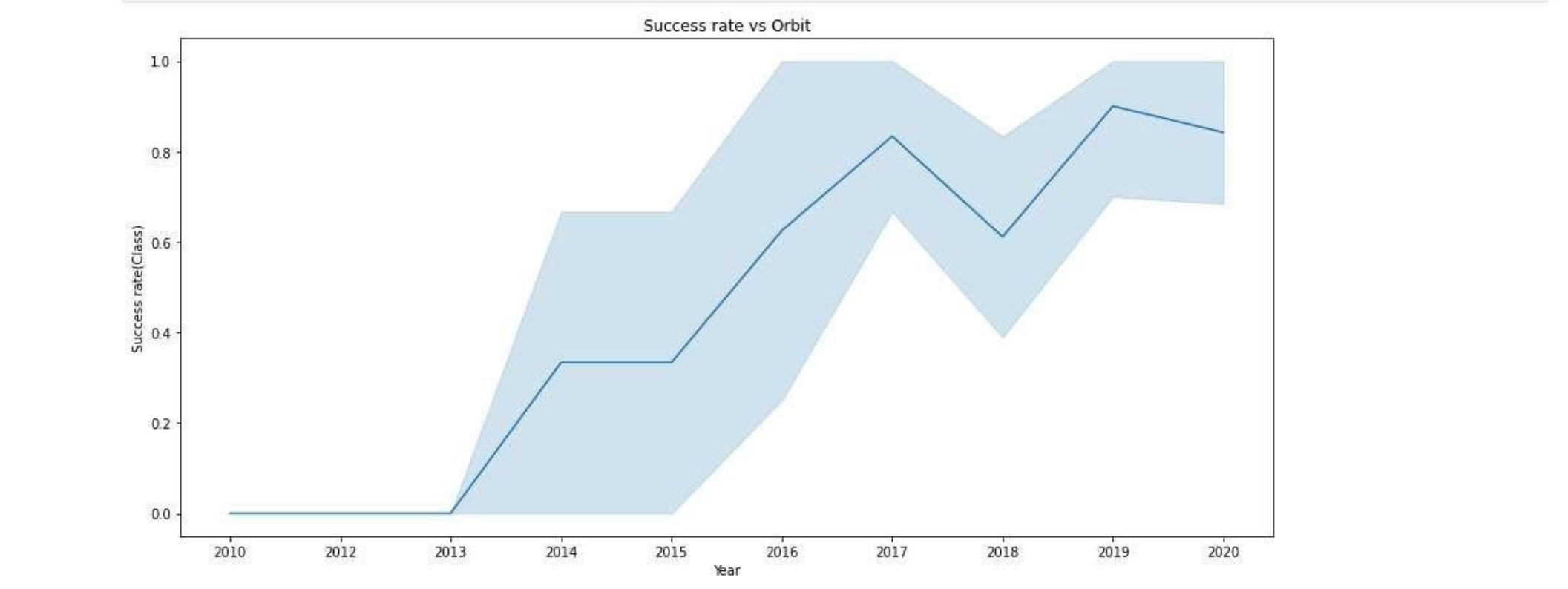
You can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there.

Launch Success Yearly Trend



It is apparent that the success rate has significantly increased from 2013 to 2020.

All Launch Site Names

Given the data, these are the names of the launch sites where different rocket landings where attempted:

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

Launch Site Names Beginning with 'CCA'

Date	Launch_Site	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	CCAFS LC-40	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	CCAFS LC-40	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	CCAFS LC-40	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	CCAFS LC-40	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	CCAFS LC-40	LEO (ISS)	NASA (CRS)	Success	No attempt

These are 5 records where launch sites begin with the letters 'CCA'. As we can see, there are other organizations besides SpaceX that were testing their rockets.

Total Payload Mass

- The information in the table displays the total payload mass carried by boosters launched by NASA.
- It seems that *NASA (CRS)* had a significantly higher total payload mass compared to the rest.

Customer	Total_Payload_Mass
NASA (CRS)	45596
NASA (CCDev)	12530
NASA (CCP)	12500
NASA (CCD)	12055
NASA (CTS)	12050
NASA (CRS), Kacific 1	2617
NASA / NOAA / ESA / EUMETSAT	1192
NASA (LSP) NOAA CNES	553
NASA (COTS)	525
NASA (LSP)	362
NASA (COTS) NRO	0

Average Payload Mass by F9 v1.1

Average_Payload_Mass (kg)	Booster_Version
2928.4	F9 v1.1

- The average payload mass carried by F9 v1.1 was 2928.4 kg.

First Successful Ground Landing Date

Date	Landing_Outcome
22-12-2015	Success (ground pad)

- The first successful ground pad landing took place in December 2015. This was a historic reusable-rocket milestone for both SpaceX and the world.
- Prior to this, no one had ever brought an orbital class booster back intact.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- It appears that there only 4 Boosters with a payload mass between 4000 and 6000.
- It is interesting to see that they all had successful landing outcomes.

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	Outcomes
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- It appears that missions generally tend to be successful with the exception of one failure.

Boosters That Carried the Maximum Payload Mass

- 12 boosters have carried the maximum payload mass of 15600 kg.
- Since the version names are similar, they might be from the same manufactures.

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

2015 Launch Records - Failed Landing Outcomes

Date	Launch_Site	Booster_Version	Landing_Outcome
10-01-2015	CCAFS LC-40	F9 v1.1 B1012	Failure (drone ship)
14-04-2015	CCAFS LC-40	F9 v1.1 B1015	Failure (drone ship)

- It appears that 2 boosters failed to land at the beginning of the year..
- The first successful landing took place later that year in December as we saw earlier.

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

- If we observe the table, it is apparent that the number of successful landings have increased since 2015.
- Before 2013, it seems that there were no attempts to land the boosters.

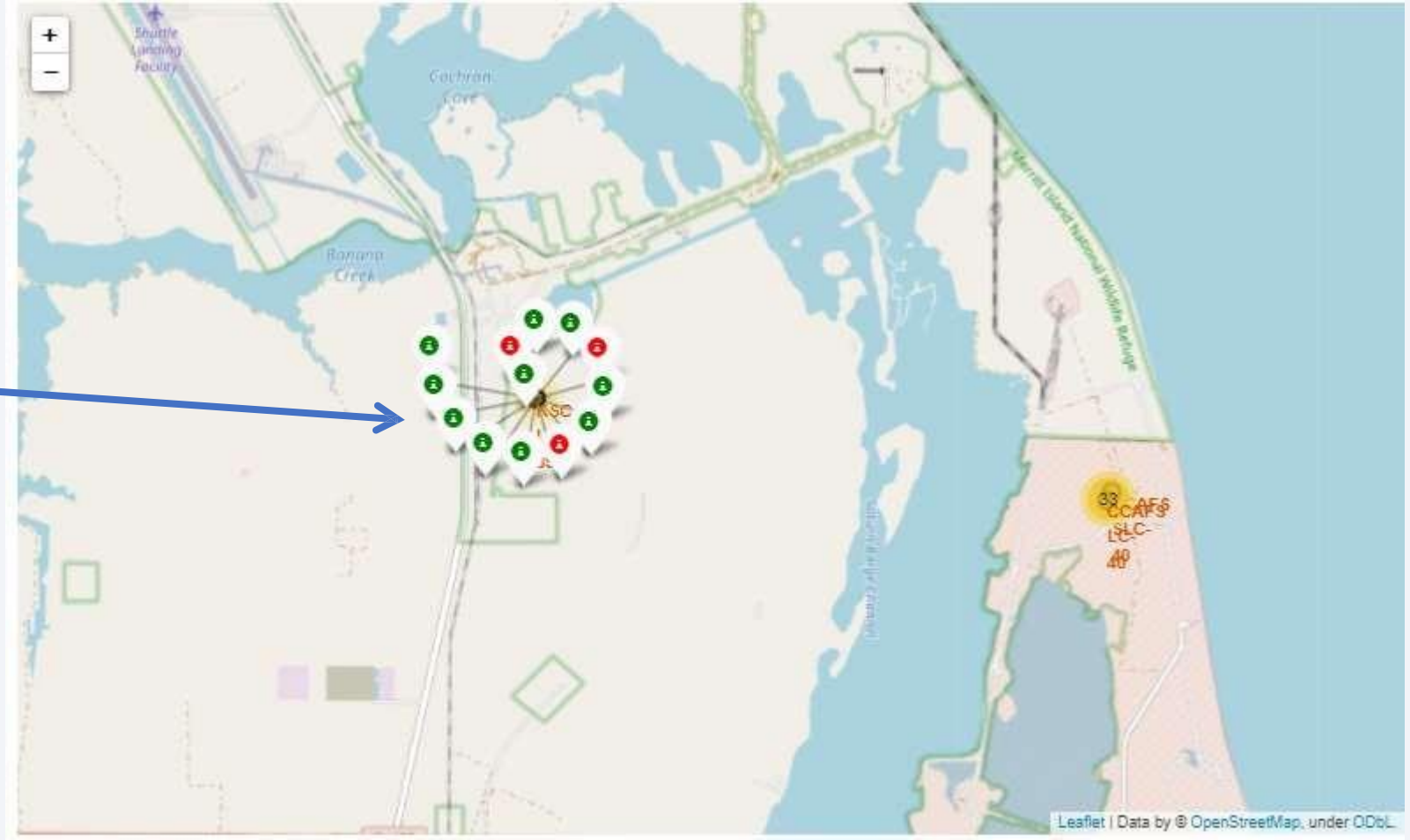
date	Landing_Outcome	Outcomes
2016-04-08	Success (drone ship)	14
2015-12-22	Success (ground pad)	9
2015-06-28	Precluded (drone ship)	1
2015-01-10	Failure (drone ship)	5
2014-04-18	Controlled (ocean)	5
2013-09-29	Uncontrolled (ocean)	2
2012-05-22	No attempt	22

Launch Site Locations

- We can see that all launch sites are in very close proximity to the coast and they are also a couple thousand kilometers away from the equator line.
- It is interesting to see that most launch sites are concentrated near Miami.

Success Rate of Rocket Launches

- The successful launches are represented by a green marker while the red marker represents failed rocket launches.
- It appears that **KSC LC-39A** had the highest success rate of rocket launches compared to other launch sites.



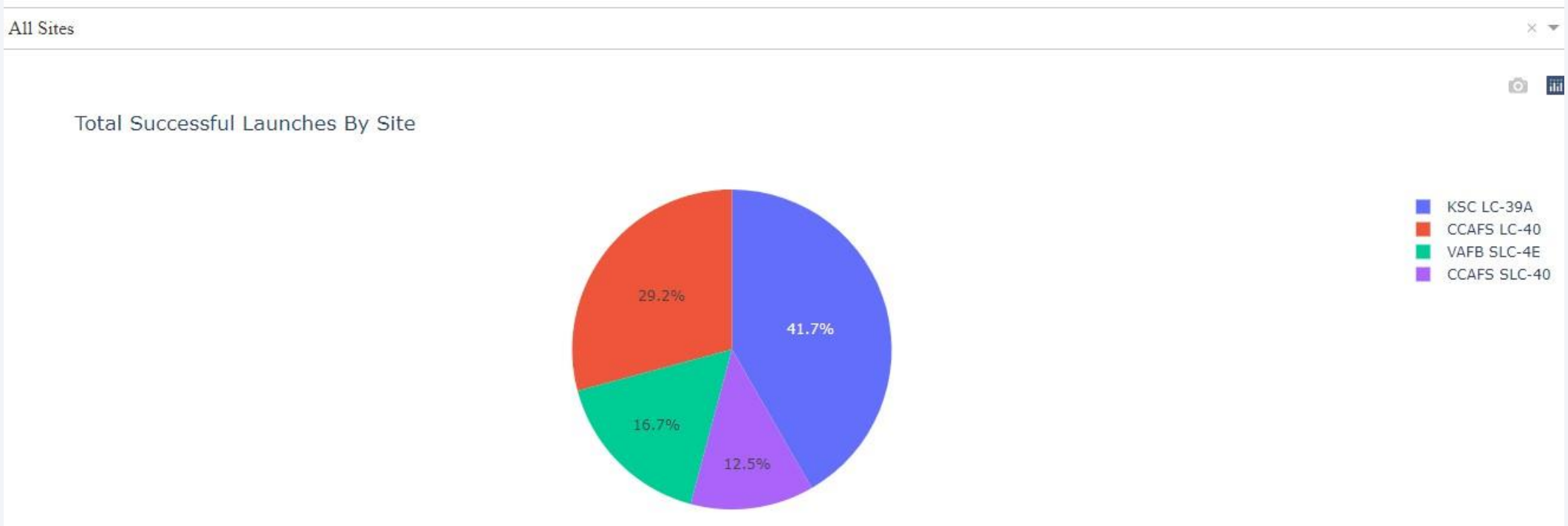
Surrounding Landmarks

- It appears that launch sites are usually set up at least 18 km away from cities. This may be because of the desire to prevent any crashes near populated areas.
- It is also apparent that launch sites are in very close proximity to railways and highways. Perhaps, due to the necessary transportation requirements for rocket parts.
- The sites are close the coast line. This is evident with the many rocket landing tests on water bodies like the ocean.

Map Object	Colour
Nearest Highway	Green
Nearest Railway	Purple
Nearest City	Crimson
Nearest Coastline	Dark Blue

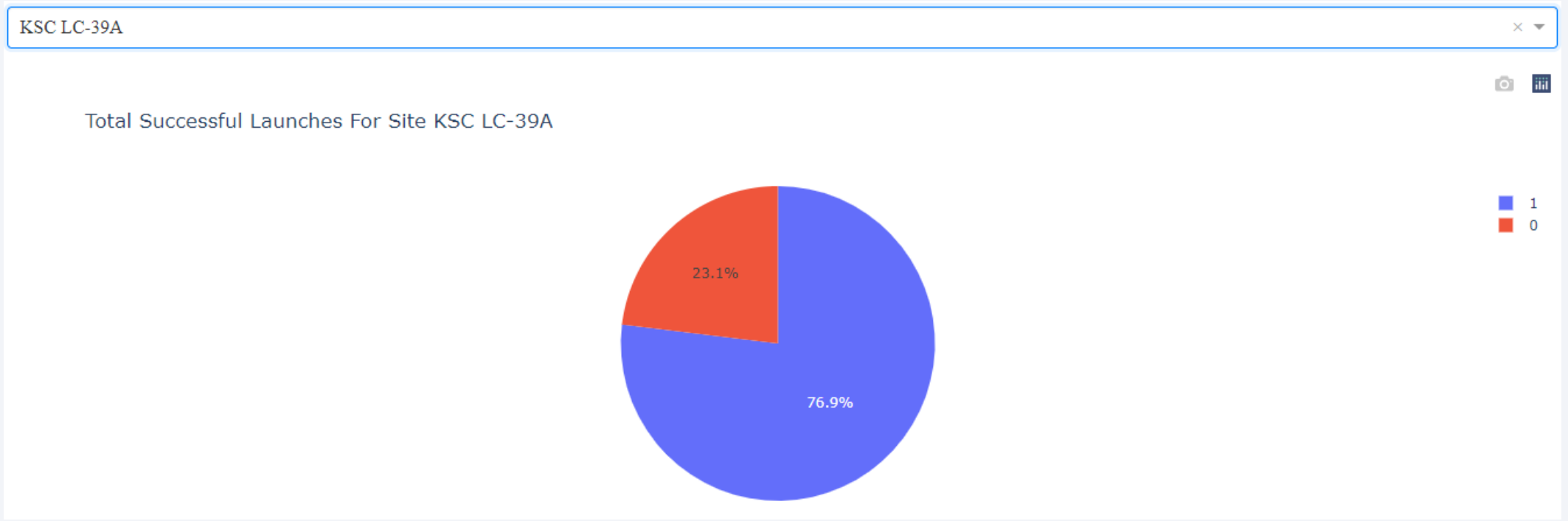
Build a Dashboard with Plotly Dash

Successful Launches by Site



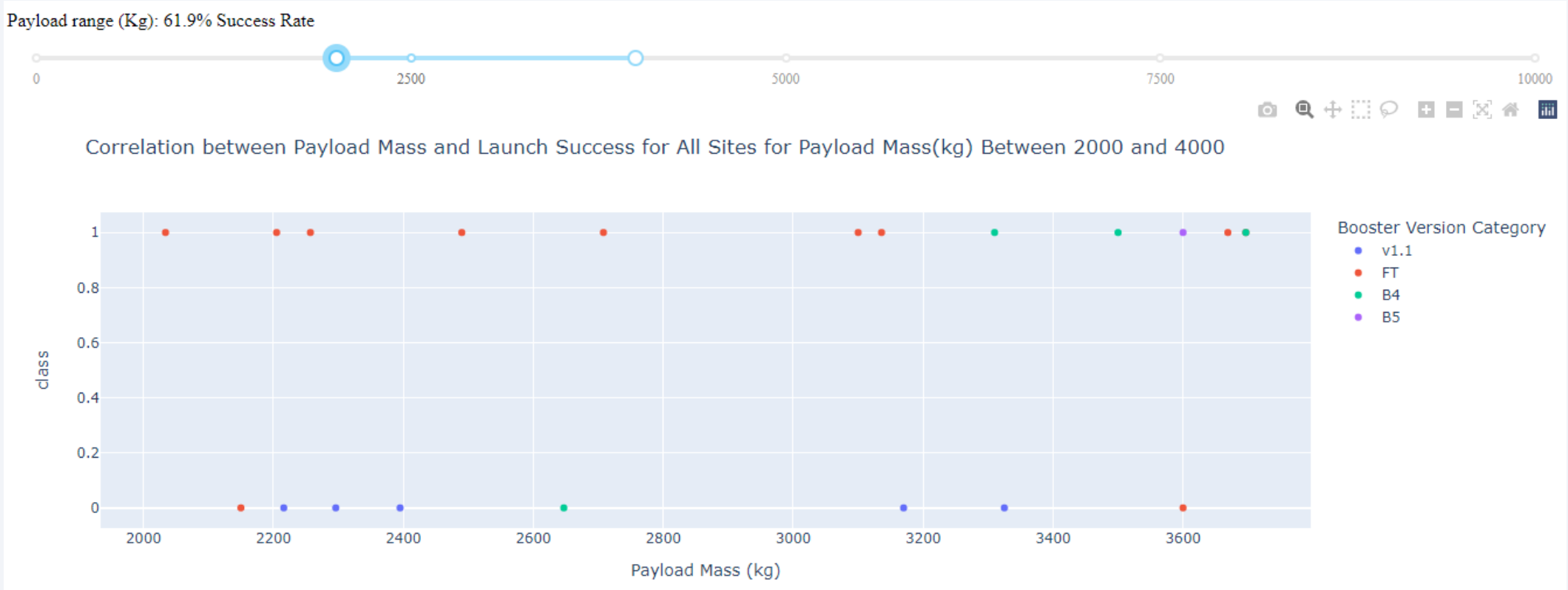
- Site KSC LC-39A has the largest successful launches as well the highest launch success rate.
- More investigation may be needed to determine why KSC LC-39A is the preferred launch site.

Total Successful Launches for Site KSC LC-39A



- As we can see, 76.9% of the total launches at site KSC LC-39A were successful. This is a the highest success rate of all the different launch sites.
- However, this success rate was only around 3% higher than the runner up; site CCAFS LC-40.

Payload Mass vs. Launch Success for All Sites



- It appears that the payload range between 2000 kg and 4000 kg has the highest success rate.
- The launch success rate was also dramatically low between the payload range of 0kg and 2500kg. Perhaps very low masses decrease launch success.
- The booster version **FT**, seems to have a higher success rate than other booster versions

Predictive Analytics (Classification)

Classification Accuracy

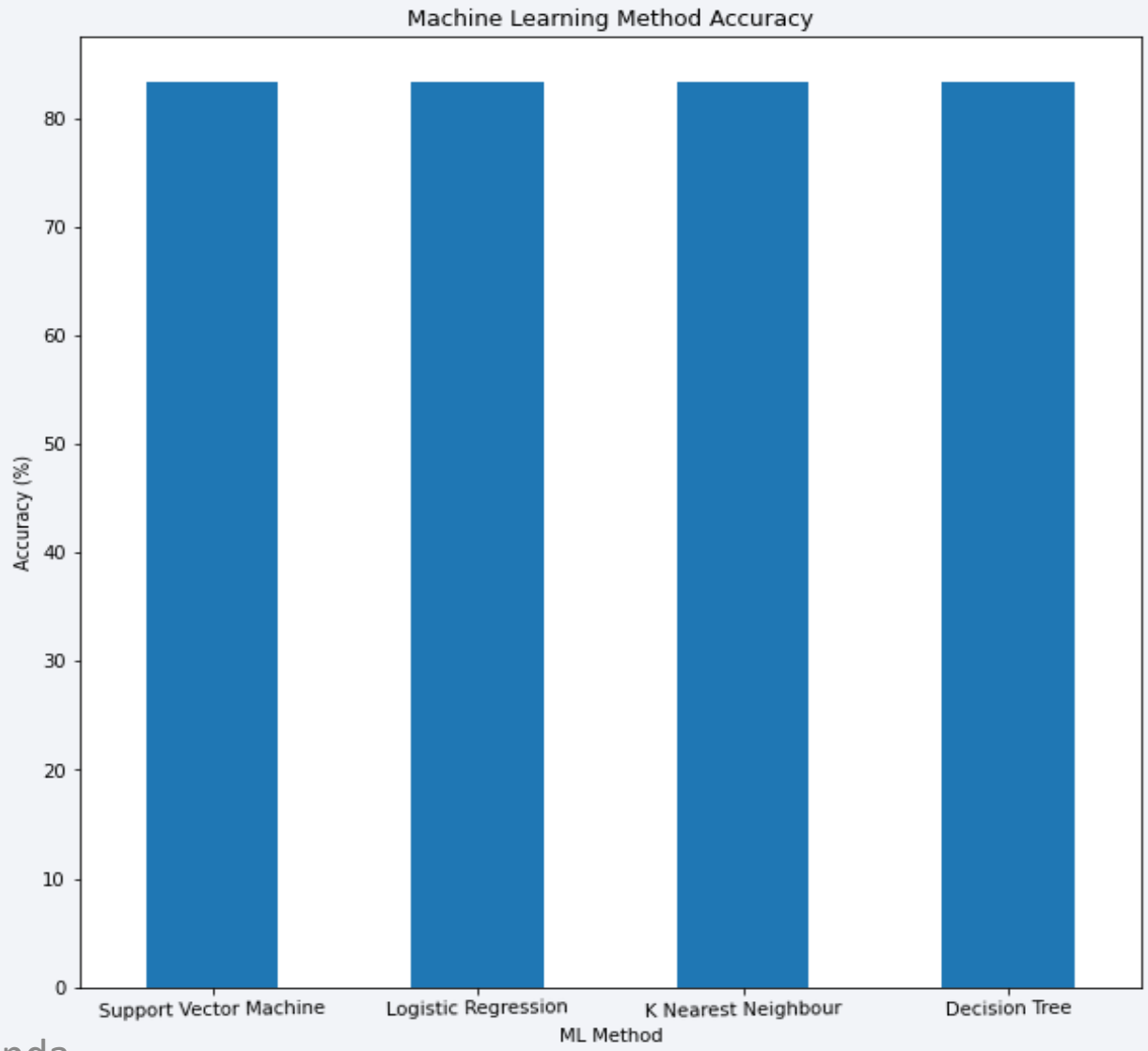
Accuracy :

• All the models performed at about the same level and had the same scores and accuracy. This is likely due to the small dataset. The Decision Tree model slightly outperformed the rest when looking at .best_score_

• .best_score_ is the average of all cv folds for a single combination of the parameters

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

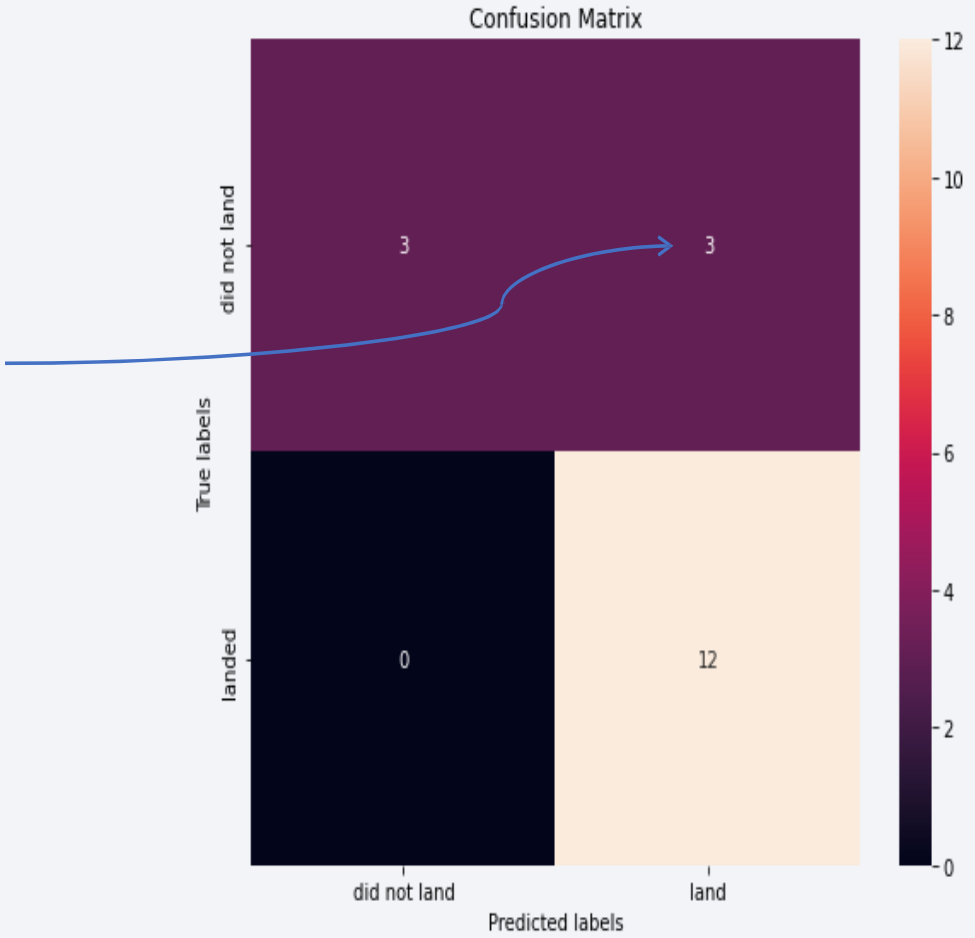
- Since all the methods have an identical accuracy score of 83.33%, we decided to use Logistic Regression for the classification



Confusion Matrix

Performance Summary :

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- Confusion Matrix Outputs:
  - 12 True positive
  - 3 True negative
  - 3 False positive
  - 0 False Negative
- Precision = TP / (TP + FP)
  - 12 / 15 = .80
- Recall = TP / (TP + FN)
  - 12 / 12 = 1
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
  - 2 * (.8 * 1) / (.8 + 1) = .89
- Accuracy = (TP + TN) / (TP + TN + FP + FN) = .833



Conclusions

Research :

- **Model Performance**: The models performed similarly on the test set with the decision tree model slightly outperforms
- **Equator**: Most of the launch sites are near the equator for an additional Natural boost – due to the rotational speed of the earth - which helps save the cost of putting in extra fuel and boosters
- **Coast**: All the launch sites are close to the coast
- **Launch Success**: Increases over time
- **KSC LC-39A**: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits**: ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass**: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Things to Consider:

- **Dataset**: A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set
- **Feature Analysis / PCA**: Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy
- **XGBoost**: Is a powerful model which was not utilized in this study. It would be interesting to see if it outperforms the other classification models

References

- IBM. *Data Science Professional Certificate.* *https://www.coursera.org/professional-certificates/ibm-data-science*

- Source Code: *https://github.com/Subrat-Nanda/IBM-Data-Science-Professional-Certification/DataScienceCapstoneProject*

Appendix

SQLite Data Set

- The table structure belongs to the SQLite data set used for SQL queries.

Tables (1)

Name	Type	Schema
Spacex		CREATE TABLE "Spacex" ("Date" TEXT, "Time(UTC)" TEXT, "Booster_Version" TEXT, "Launch_Site" TEXT, "Payload" TEXT, "PAYLOAD_MASS__KG_" INTEGER, "Orbit" TEXT, "Customer" TEXT, "Mission_Outcome" TEXT, "Landing_Outcome" TEXT)
Date	TEXT	"Date" TEXT
Time(UTC)	TEXT	"Time(UTC)" TEXT
Booster_Version	TEXT	"Booster_Version" TEXT
Launch_Site	TEXT	"Launch_Site" TEXT
Payload	TEXT	"Payload" TEXT
PAYLOAD_MASS__KG_	INTEGER	"PAYLOAD_MASS__KG_" INTEGER
Orbit	TEXT	"Orbit" TEXT
Customer	TEXT	"Customer" TEXT
Mission_Outcome	TEXT	"Mission_Outcome" TEXT
Landing_Outcome	TEXT	"Landing_Outcome" TEXT

Thank you!