IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Karolina Cukier

30.04.2024

Outline

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

Executive Summary

- **Summary of methodologies**

The study aimed to uncover the key elements contributing to the successful landing of SpaceX's Falcon 9 rockets. To achieve this goal, the research employed various approaches and techniques, such as gathering data through the SpaceX REST API and web scraping, processing and organizing the data to create a variable indicating success or failure, investigating the data through visualization techniques, focusing on various factors, analyzing the data using SQL to compute statistics, exploring the success rates of launch sites and their proximity to geographical landmarks, presenting visualizations highlighting launch sites with the highest success rates and successful payload ranges, constructing predictive models, including logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN), to forecast landing outcomes.

- **Summary of all results**

The analysis of the data show that rocket launch has improved over time, the highest success rate have ESL1, GEO, HEO, and SSO orbits and the Kennedy Space Center Launch Complex 39A is the landing site with the best success rate. The decicion tree model had the best accuraccy. All the others prediction models have shown the same accuraccy (around 83%).

Introduction

- **Project background**

SpaceX, renowned in the space industry, is dedicated to democratizing space travel by making it more affordable. Its cost-effectiveness stems from its innovative approach of reusing the first stage of its Falcon 9 rocket, which reduces launch expenses to approximately $62 million per launch. In contrast, competitors unable to reuse the first stage face significantly higher costs (exceeding up to 2.6 times of SpaceX cost). By predicting the successful landing of the first stage, we can ascertain the launch price.

- **Outcome**

Our aim is to predict the success of the initial phase of rocket retrieval, ultimately optimizing how resources are allocated. Through this predictive ability, we strive to improve mission success rates and generate significant cost savings.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web scraping
- Data wrangling
  - Sorting important data for the project, addressing any missing values and preparing data for EDA
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Developing predictive models to anticipate landing results using classification techniques. Assessing models to identify the most effective model and optimal parameters.

Data Collection – SpaceX API

Data retriving from the **REST API** using function **GET**

Decoding content as a Json using **.json()** turning into a dataframe using **.json_normalize()**

Using the API to get information about the launches

Creating dictonary

Creating dataframe using **pd.DataFrame** form the dictonary

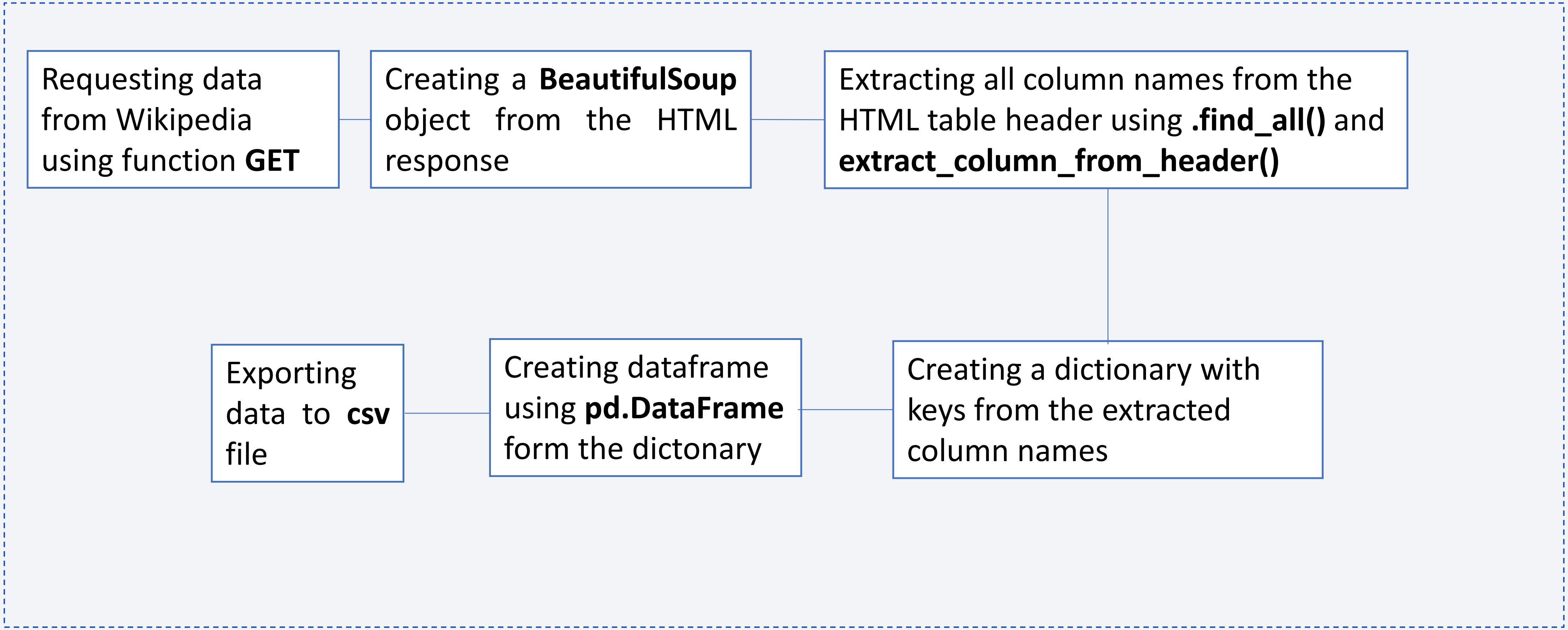
Filter the dataframe to only include Falcon 9 launches

Dealing with missing values using **.mean()** and **.replace()**

Exporting data to **csv** file

GitHub Notebook

Data Collection - Scraping



GitHub Notebook

Data Wrangling

Importing dataset, calculating the number of launches on each site using **value_counts()**

Calculating the number and occurrence of each orbit using **value_counts()**

Calculate the number and occurence of mission outcome of **the orbits using value_counts()**

Create a landing outcome label from Outcome column

Exporting data to **csv** file

GitHub Notebook

EDA with Data Visualization

- **Charts**

Flight Number vs. Payload Mass

Flight Number vs. Launch Site

Payload Mass vs. Launch Site

Orbit vs. Class

Flight Number vs. Class

Flight Number vs. Orbit

Year vs Success Rate

Scatter plots are useful to determinate if relationship between various factors can be used for machine learning.

Differences between distinct categories are shown with bar charts. These visualizations depict the connections between the categories and a quantified value.

GitHub Notebook

EDA with SQL

- **SQL queries**

- the names of the unique launch sites in the space mission

- 5 records where launch sites begin with the string 'CCA'

- the total payload mass carried by boosters launched by NASA (CRS)

- average payload mass carried by booster version F9 v1.1

- the date when the first succesful landing outcome in ground pad was acheived

- the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

- the total number of successful and failure mission outcomes

- the names of the booster_versions which have carried the maximum payload mass. Use a subquery

- the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

- Rank of 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 Notebook]()

Interactive Map with Folium

- **Colored Circles**

- blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name based on its coordinate values
- red circles at all launch sites coordinates with a popup label showing its name using its name based on its coordinate values

- **Markers**

- green markers for successful launches
- red markers for unsuccessful launches to display high success rates

- **Lines**

- lines displaying distance distance to a closest city, railway, highway, etc.from Space Launch Complex 40

GitHub Notebook

Dashboard with Plotly Dash

- **Dropdown Lisy for Launch Sites**

Enable users to choose either all launch sites or a specific one.

- **Payload Mass Range Slider**

Enable users to select the range of payload masses.

- **Pie Chart Displaying Successful Launches**

Show the proportion of successful and unsuccessful launches relative to the total.

- **Scatter Plot Illustrating Payload Mass vs. Success Rate by Booster Version**

Visualize the relationship between payload mass and launch success, categorized by booster version.

GitHub Notebook

Predictive Analysis (Classification)

- Create NumPy array from the Class
- Standardize, fit and transform the data
- Split the data in the test and training data
- Create a logistic regression object and apply it on following algorithms:

- logistic regression
- support vector machine
- decision tree
- K Nearest Neighbor

- Calculate accuracy on the test data for all models
- Plot the confusion matrix for all models
- Assess the best model comparing accuracy using the method **score**

GitHub Notebook

Results

- **Exploratory data analysis results**

There has been a noticeable enhancement in launch success over the years. Among the landing sites, KSC LC-39A stands out with the highest rate of success. Orbits such as ES-L1, GEO, HEO, and SSO have consistently achieved a 100% success rate.

- **Visual analytic results**

The majority of launch sites are situated in proximity to the equator and are adjacent to coastlines. Launch facilities are strategically positioned to ensure they are sufficiently distant from populated areas, highways, and railways, minimizing potential damage in the event of a failed launch, while still maintaining accessibility for personnel and logistical support.

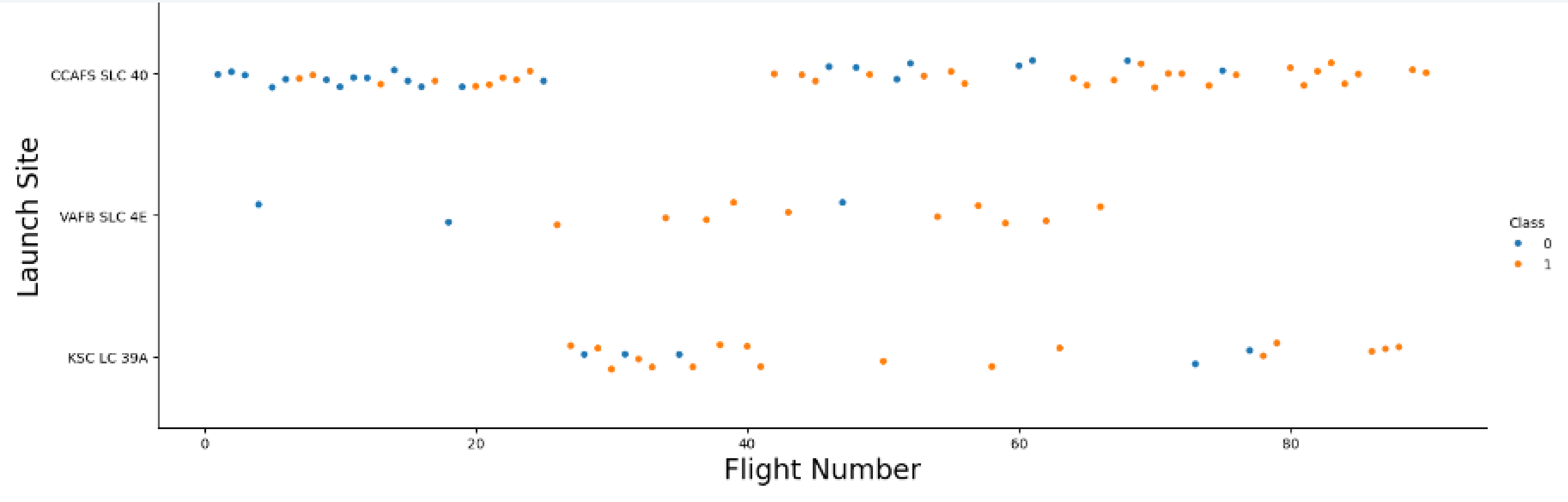
- **Predictive analysis results**

The best predictive model is a Decision Tree.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



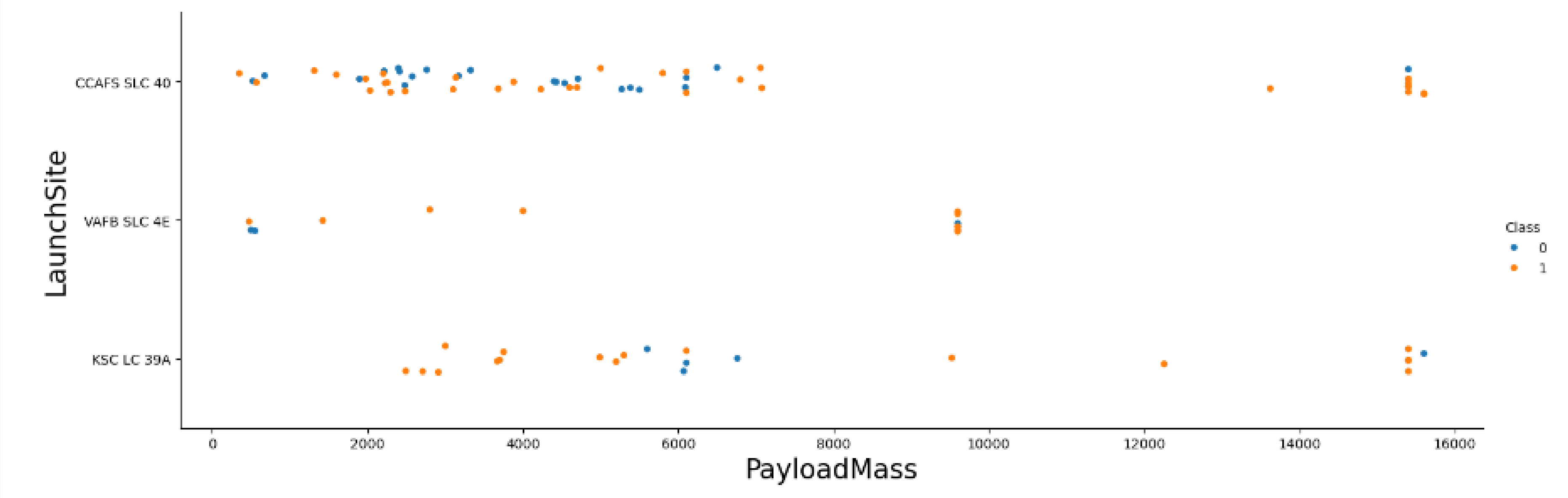
0 – fail
1 - success

VAFB SLC 4E and KSC LC 39A exhibit elevated success rates compared to other sites.

Approximately half of the launches originated from the CCAFS SLC 40 launch site.

It can be deduced that recent launches tend to achieve a higher rate of success

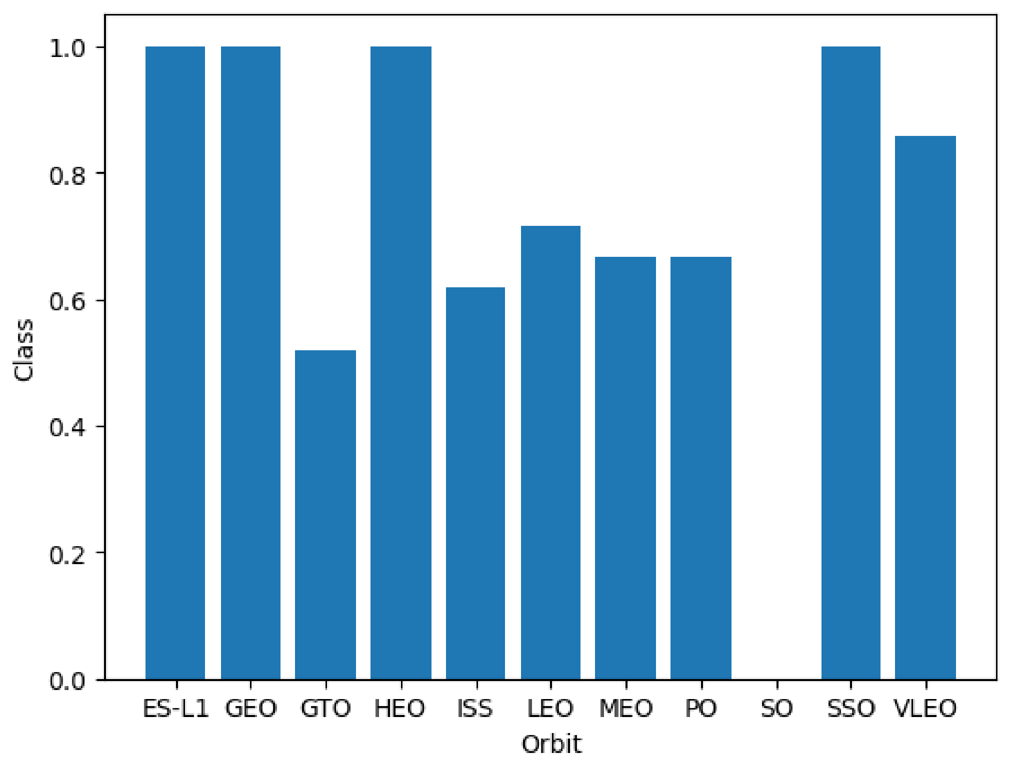
Payload vs. Launch Site



0 – fail
1 - success

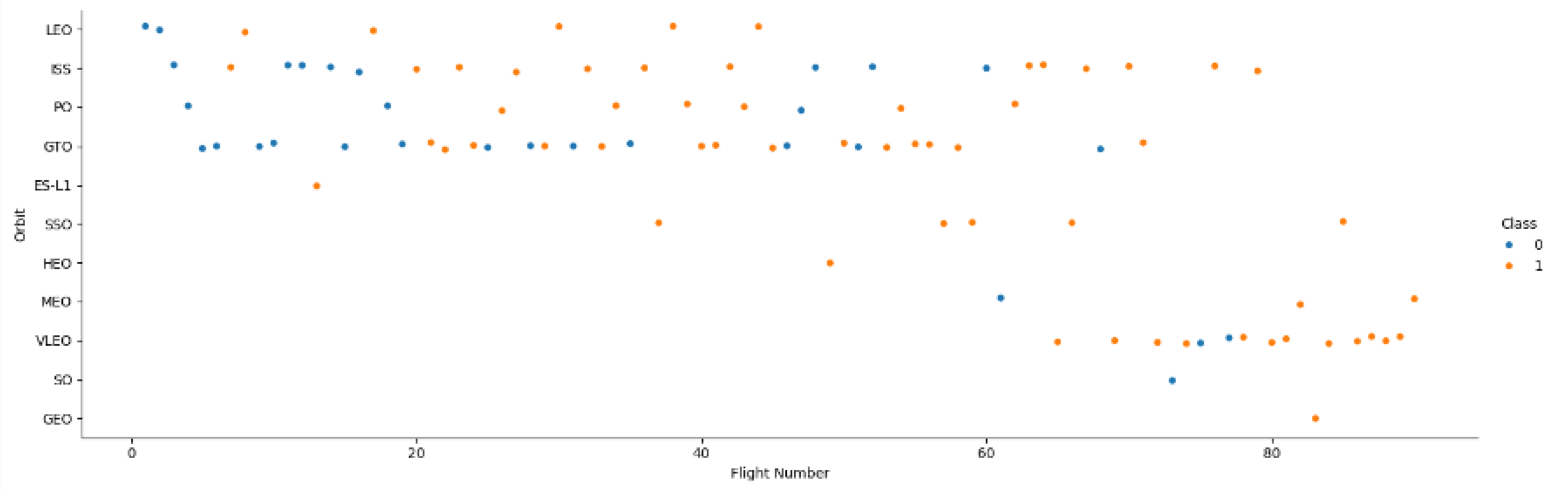
In general, a correlation exists between higher payload masses (measured in kilograms) and increased success rates. The majority of launches with payloads exceeding 7,000 kg resulted in successful outcomes. For launches under 5,500 kg, KSC LC 39A boasts a flawless success rate of 100%.

Success Rate vs. Orbit Type



ES-L1, GEO, HEO, SSO has the highest success rate.
GTO, ISS and SO has the lowest success rate.

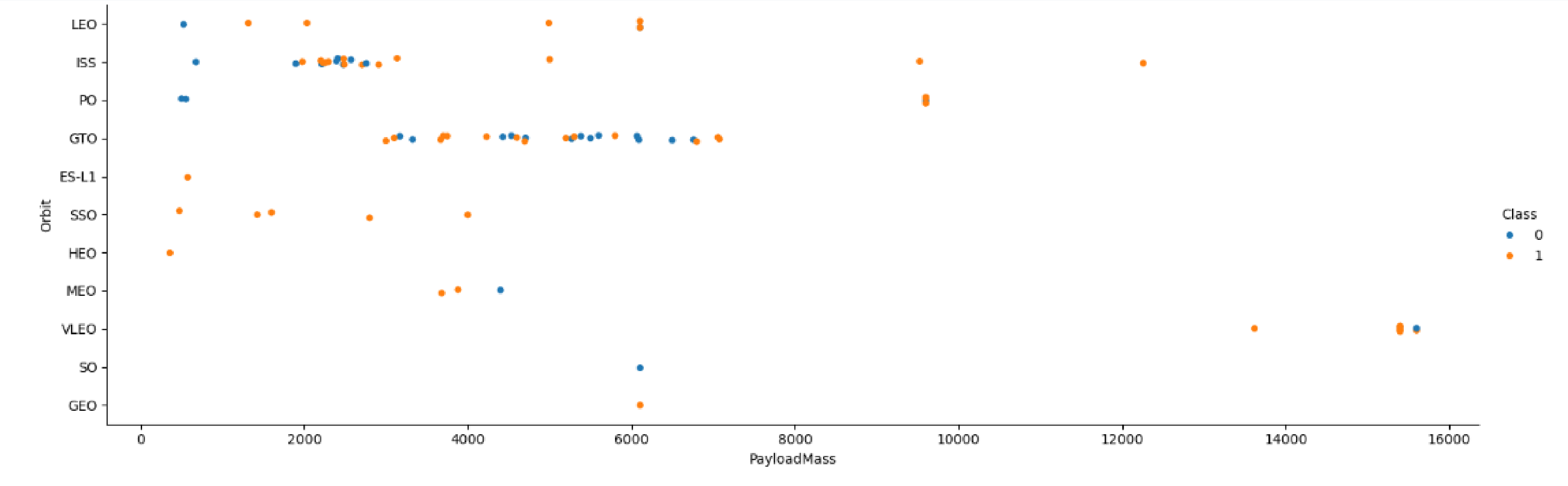
Flight Number vs. Orbit Type



0 – fail
1 - success

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.

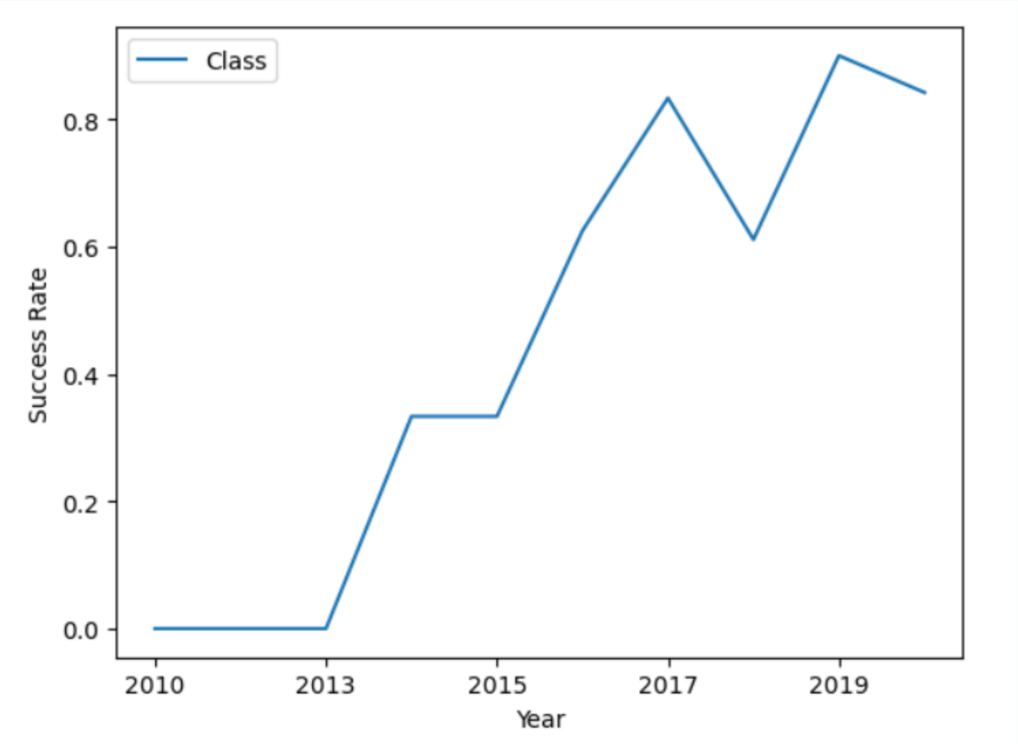
Payload vs. Orbit Type



0 – fail
1 - success

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 here.

Launch Success Yearly Trend



The sucess rate since 2013 kept increasing till 2020.

Launch Information

- Unique launch sites

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

- Find 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Payload Mass Information

- Total payload carried by boosters from NASA

sum(PAYLOAD_MASS__KG_)
45596

- Average payload mass carried by booster version F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

Landing and Mission Information

- Date of the first successful landing outcome on ground pad

min(Date)
2015-12-22

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Mission and Boosters Information

- Total number of successful and failure mission outcomes

Success	Failure
61	10

- Names of the booster which have carried the maximum payload mass

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 in drone ship, their booster versions, and launch site names for in year 2015

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

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

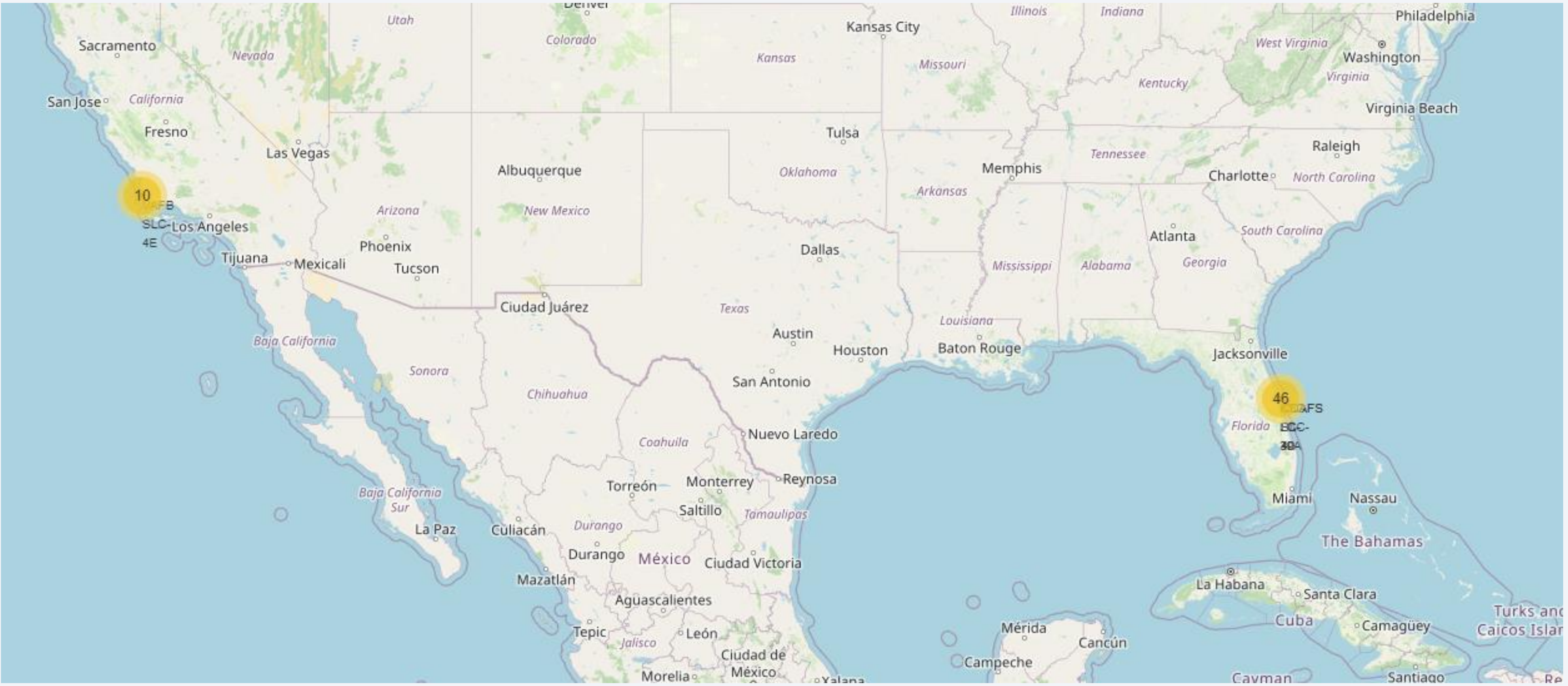
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

ranking	outcome_count	Landing_Outcome
1	10	No attempt
2	5	Success (drone ship)
2	5	Failure (drone ship)
4	3	Success (ground pad)
4	3	Controlled (ocean)
6	2	Uncontrolled (ocean)
6	2	Failure (parachute)
8	1	Precluded (drone ship)

Section 3

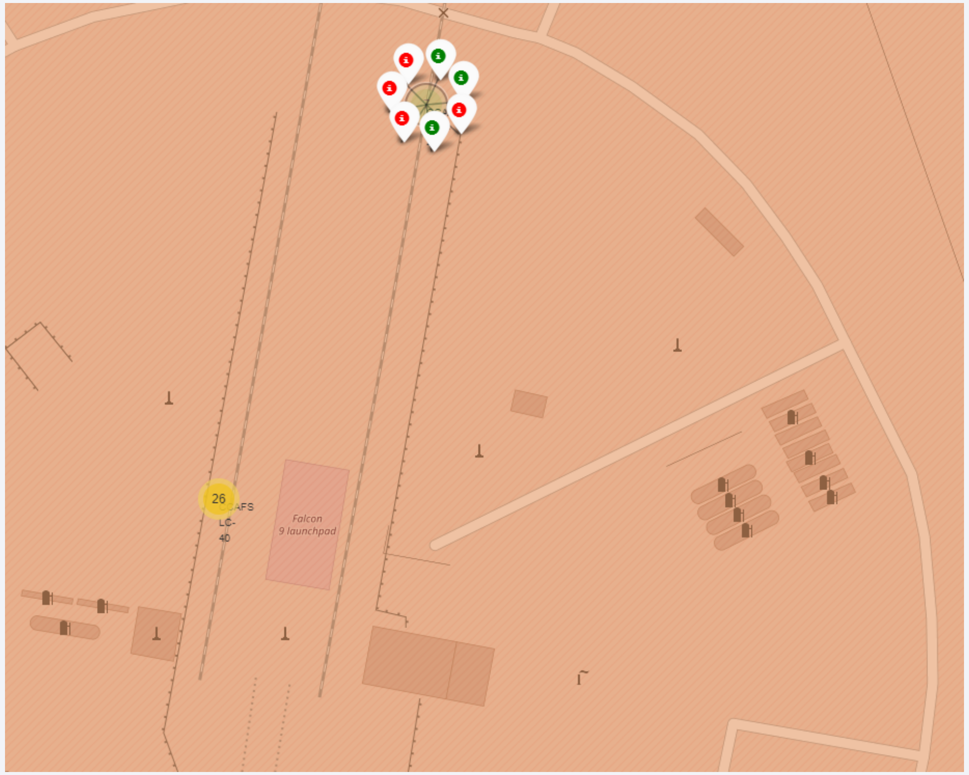
Launch Sites Proximities Analysis

Launch Sites



Launch sites located nearer to the equator benefit from advantages when launching into equatorial orbits. The proximity to the equator allows rockets to leverage the Earth's rotational speed, resulting in additional momentum for prograde orbits. This natural boost from Earth's rotation minimizes the need for extra fuel and boosters, thereby reducing costs associated with launching payloads into space.

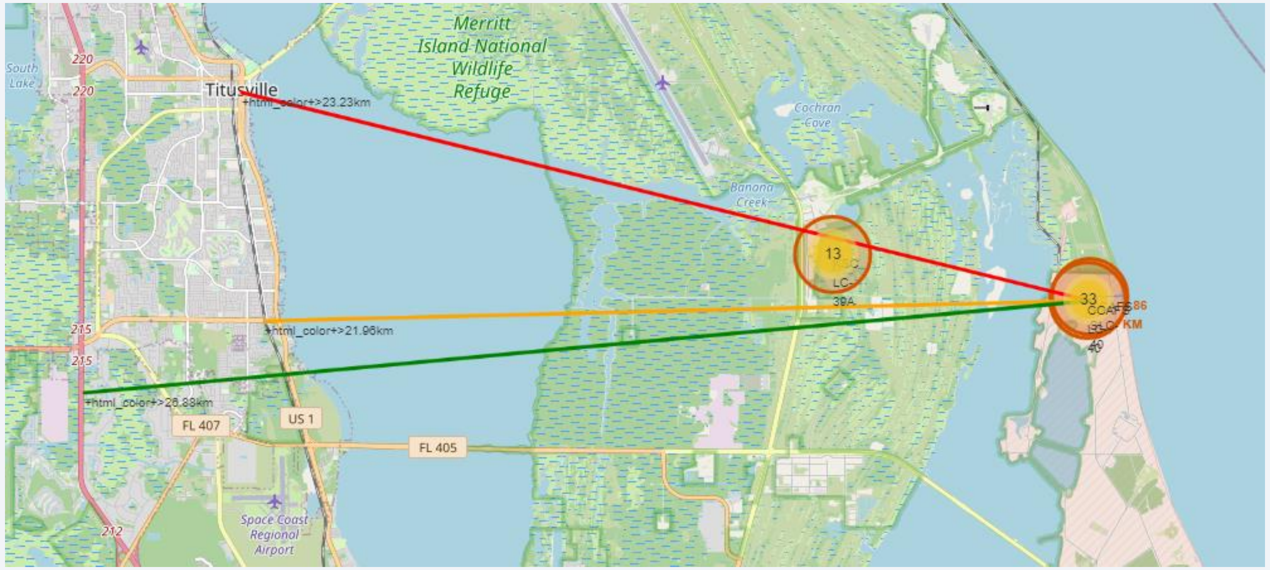
Launch Outcomes



Green markers - successful launches
Red markers - unsuccessful launches

Launch site CCAFS SLC 40 has 42.85% success rate.

Distances from Launch Site CCAFS SLC 40



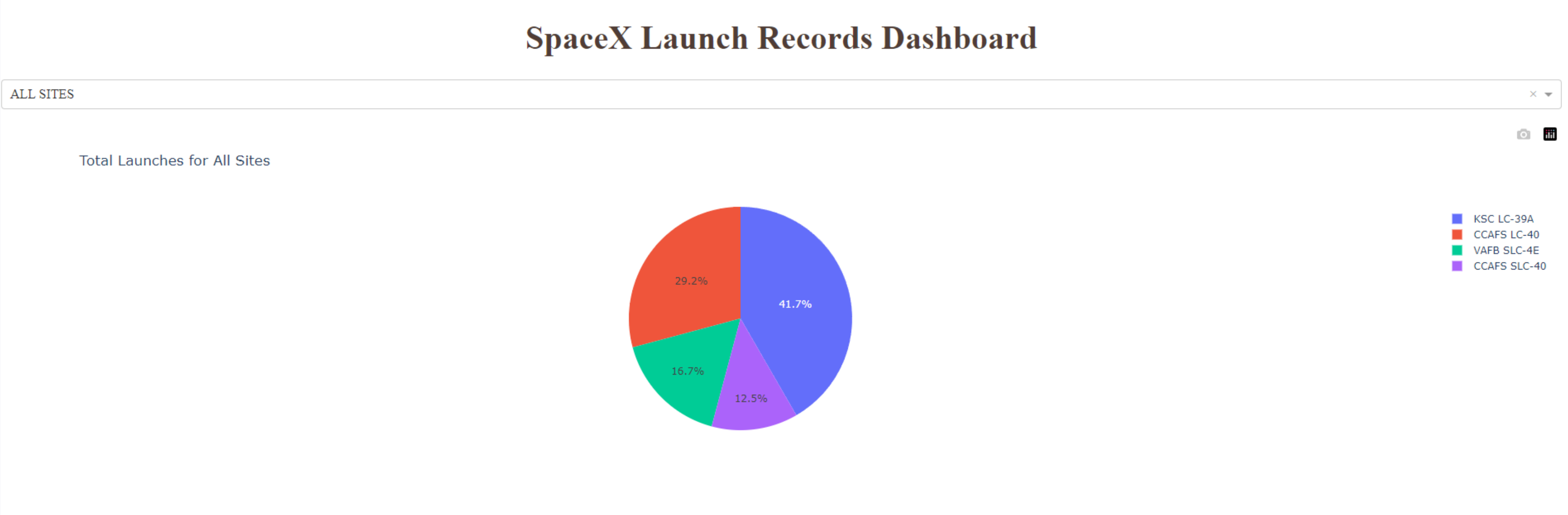
City Distance: 23.23 km
Railway Distance: 21.96 km
Highway Distance: 26.88 km
Coastline Distance: 0.86 km

The proximity of launch sites to surrounding areas holds significant importance for safety considerations. It is imperative for launch sites to maintain a safe distance from inhabited properties to mitigate the risk of damage in the event of a launch failure. Simultaneously, they must remain sufficiently close to ensure timely access for logistical support, including material transportation and medical assistance, to effectively facilitate launch operations.

Section 4

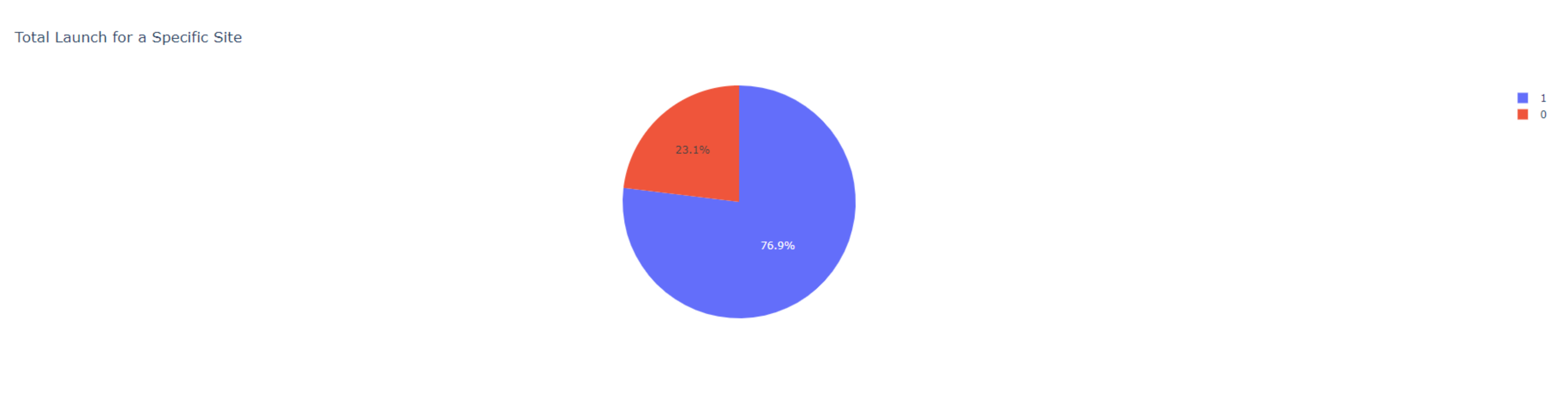
Build a Dashboard with Plotly Dash

Total Launches for all Sites



KSC LC 39A has the most successful launches amongst launch sites

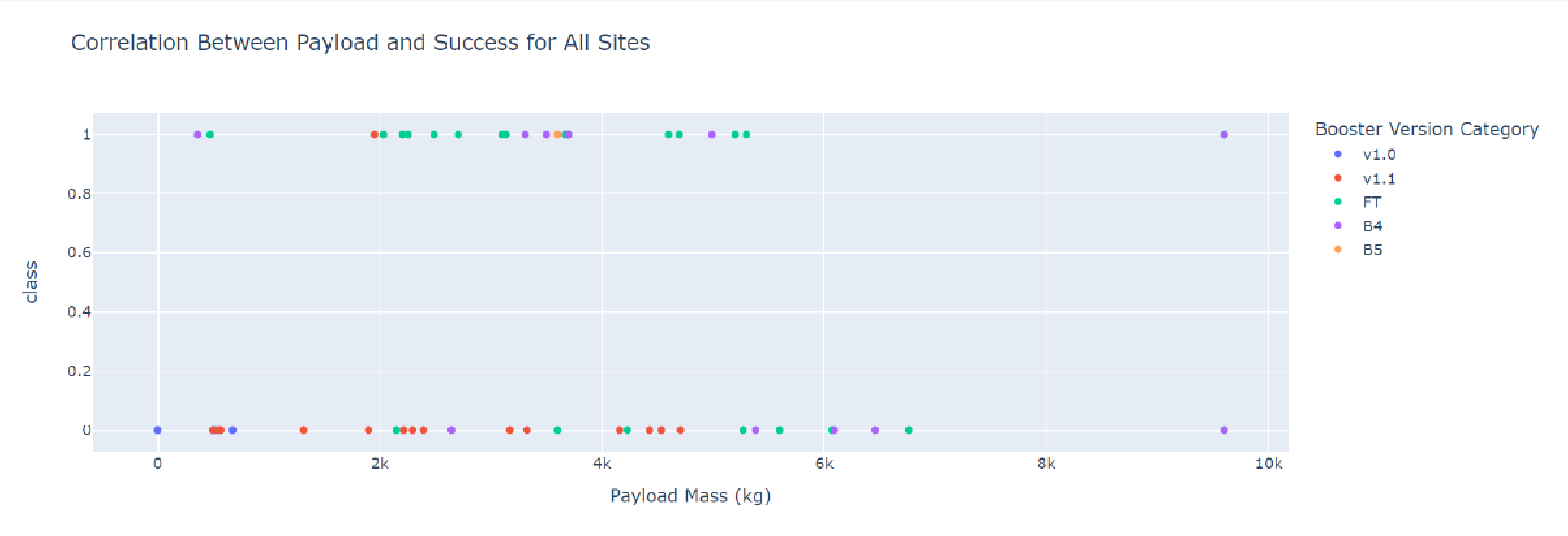
Launch Site with highest Launch Success Ratio



0 – fail
1 - success

KSC LC 39A has the highest success rate with 76.9% success

Success Rate by Payload Mass

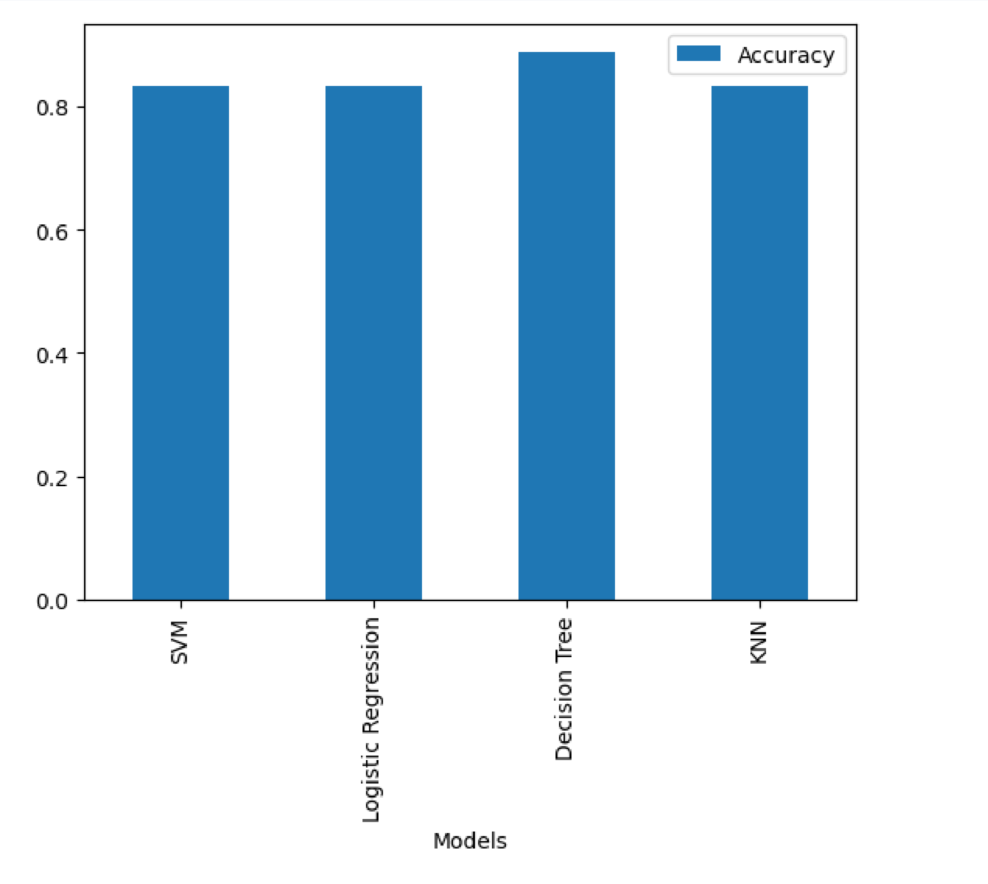


2,000 kg and 5,000 kg Payload Mass have the highest success rate

Section 5

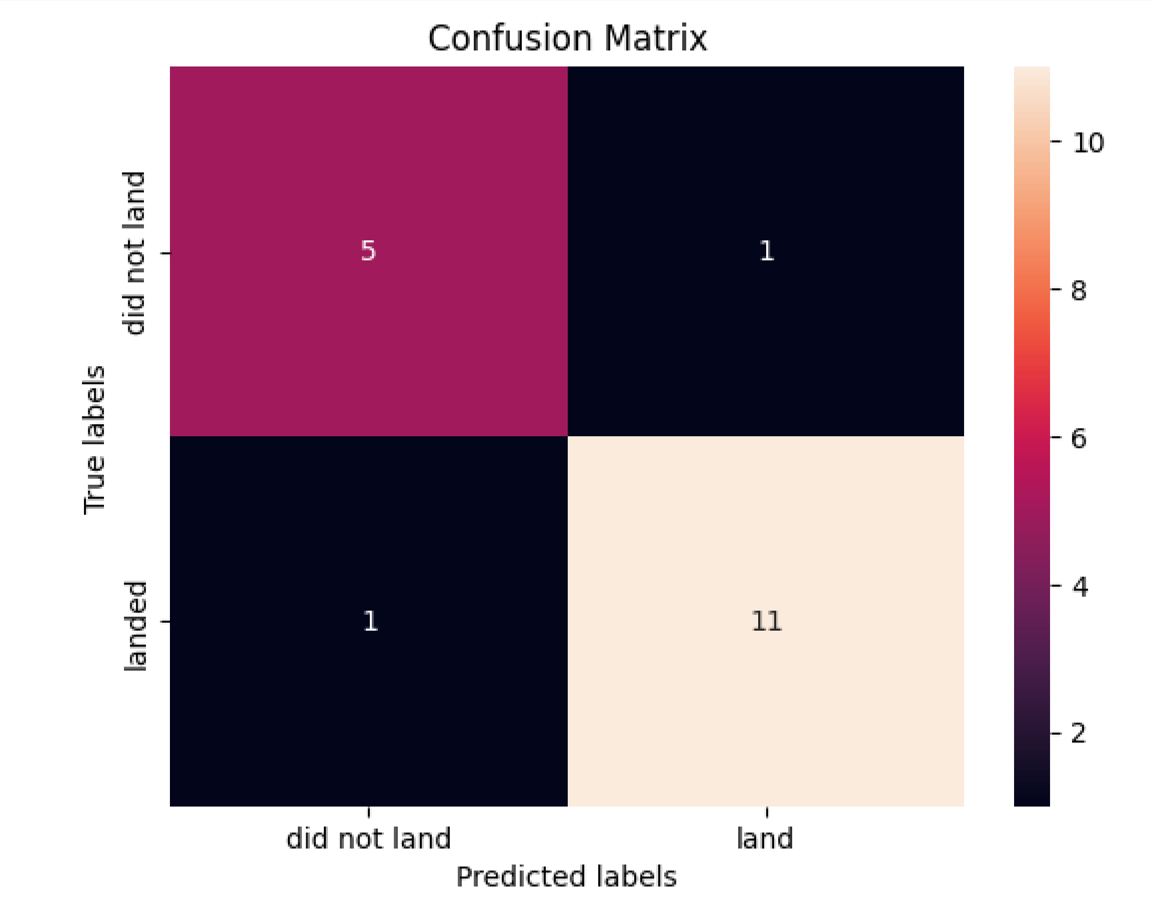
Predictive Analysis (Classification)

Classification Accuracy



The decicion tree model had the best accuraccy. All the others prediction models have shown the same accuraccy (around 83%).

Confusion Matrix



A confusion matrix for decision tree summarizes the performance of an algorithm:

11 True positive
5 True negative
1 False positive
1 False Negative

Accuracy : 0.8910714285714285

Conclusions

- The performance of the models on the test set was generally comparable, with the decision tree model showing a slight advantage.
- The majority of launch sites are situated near the equator, capitalizing on the Earth's rotational speed to reduce the need for additional fuel and boosters, thus saving costs.
- There is a discernible upward trend in launch success rates over time.
- Across all launch sites, there is a positive correlation between higher payload masses (measured in kilograms) and increased success rates.
- KSC LC 39A launch site boasts the highest success rate among all sites. Notably, it achieves a flawless success rate for launches weighing less than 5,500 kg.
- Orbits such as ES L1, GEO, HEO, and SSO have consistently achieved a perfect success rate.

Thank you!