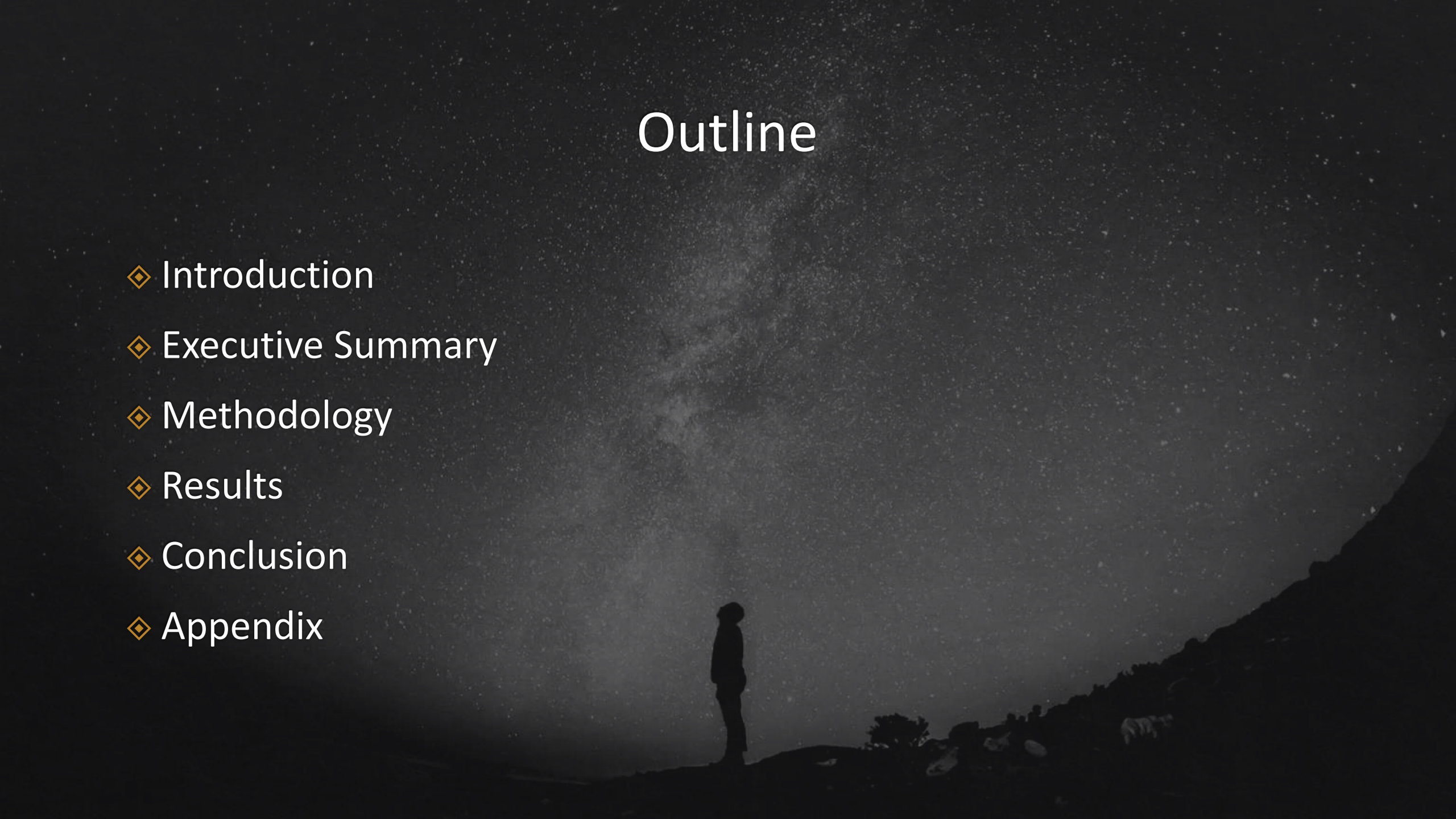
WINNING THE SPACE RACE WITH DATA SCIENCE

AAKIF ALTAF

Outline

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



Introduction

The aim of this report is to predict if the first stage of SpaceX Falcon 9 rocket launches will land successfully.

- SpaceX Falcon 9 rocket launches cost 62 million dollars
- Other providers cost upwards of 165 million dollars
- Much of the savings is because SpaceX can reuse the first stage of Falcon9 rocket launches

If we can determine if the first stage will land, we can estimate the launch cost. This information could be beneficial to companies competing against SpaceX.

Executive Summary of Methodology

Data Collection

Data Wrangling

Exploratory Data Analysis

Geospacial Data Visualization

Dashboard

Predictive Analysis

- Data extraction through an API
- Webscraping with Beautiful Soup html parser
- Transforming and loading data with Pandas
- Sorting and filtering data with SQL
- Data visualization using Plotly and Seaborn
- Interactive maps of launch sites using Folium
- SpaceX launch dashboard built in Plotly Dash
- Supervised Machine Learning with sklearn

Data Collection

- Use API requests on SpaceX REST API
- Webscrape SpaceX Wikipedia entry
- Combine into dataframe to get complete dataset

SpaceX Data Collection

1

Request/Parse Data

- Request data from API
- Decode content as .JSON format
- Convert to data frame

2

Filter Dataframe

- Isolate Falcon 9 launches within the data frame

3

Remove Missing Values

- Use isnull() to replace missing values with calculated .mean()
- Export to CSV

Source: SpaceX Data API

Falcon9 Launch Data Collection

1

Pull data from URL

- Use get() method to extract data from Wikipedia
- Create Beautiful Soup object from html

2

Extract Column Names

- Extract all of the column names from the html table header (‘th’ element)

3

Create a Dataframe

- Parse the html tables into a Pandas DataFrame
- Export to CSV

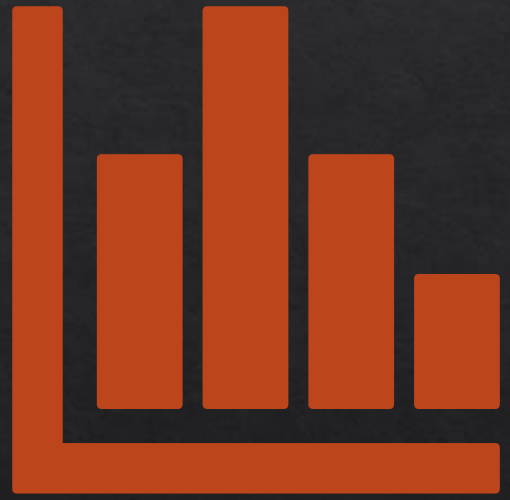
Source: SpaceX Wikipedia

Data Wrangling

- Clean dataset using Pandas and Numpy
- Convert complicated outcomes into Boolean results
- 1 = booster successfully landed
- 0 = booster failed to land

Data Wrangling

Valuecounts()	Calculate number of launches on each site
Valuecounts()	Calculate number and occurrence of each orbit type
Valuecounts()	Calculate number and occurrence of mission outcomes per orbit type
Boolean Label 0 = failed 1 = success	Create landing outcome label: 0=failure and 1=success
pd.df.to_csv()	Export data to CSV



Exploratory Data Analysis

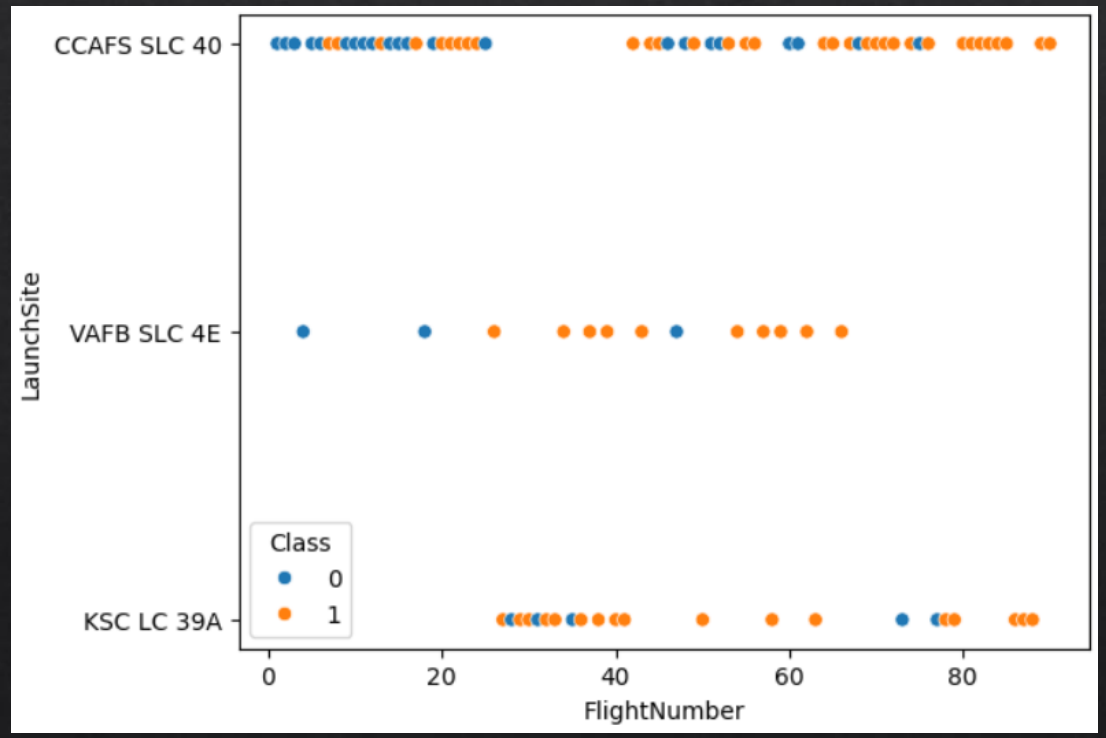
- Query the dataset using SQL to prepare for EDA data visualization
- Plot charts using Plotly and Seaborn to gain a better understanding of the dataset

Exploratory Data Analysis in SQL

Use SQL queries to perform initial EDA:

- Select names of unique launch sites
- Filter records where launch sites begin with ‘CCA’
- Calculate total payload mass carried by boosters launches by NASA
- Determine avg payload mass carried by booster version F9
- List date of first successful landing
- List names of boosters successfully carrying payloads 4000 – 6000kg
- List total number of successful and failed outcomes
- List names of boosters which have carried max payload mass

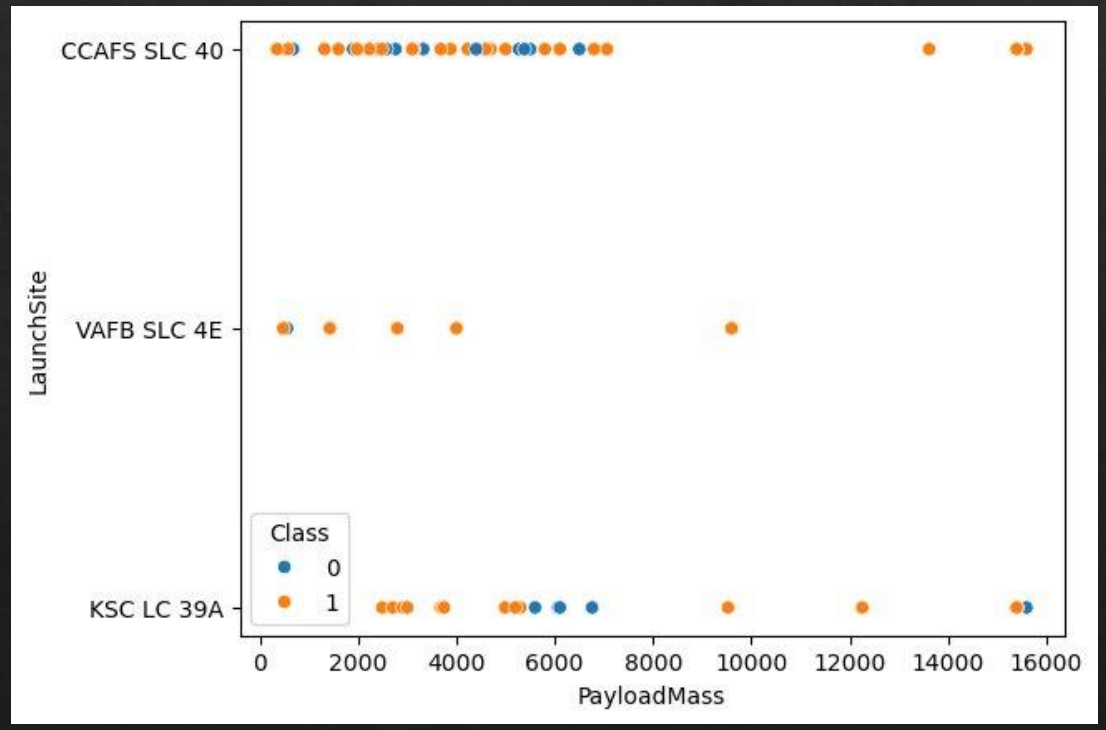
Exploratory Data Analysis in Python



Success Rates by Flight Number and Launch Site

- The chance of success (orange) increases as the number of launches increase
- Launch site CCAFS SLC 40 has seen the most rocket launches
- KSC LC 39A has the highest number of successful launches

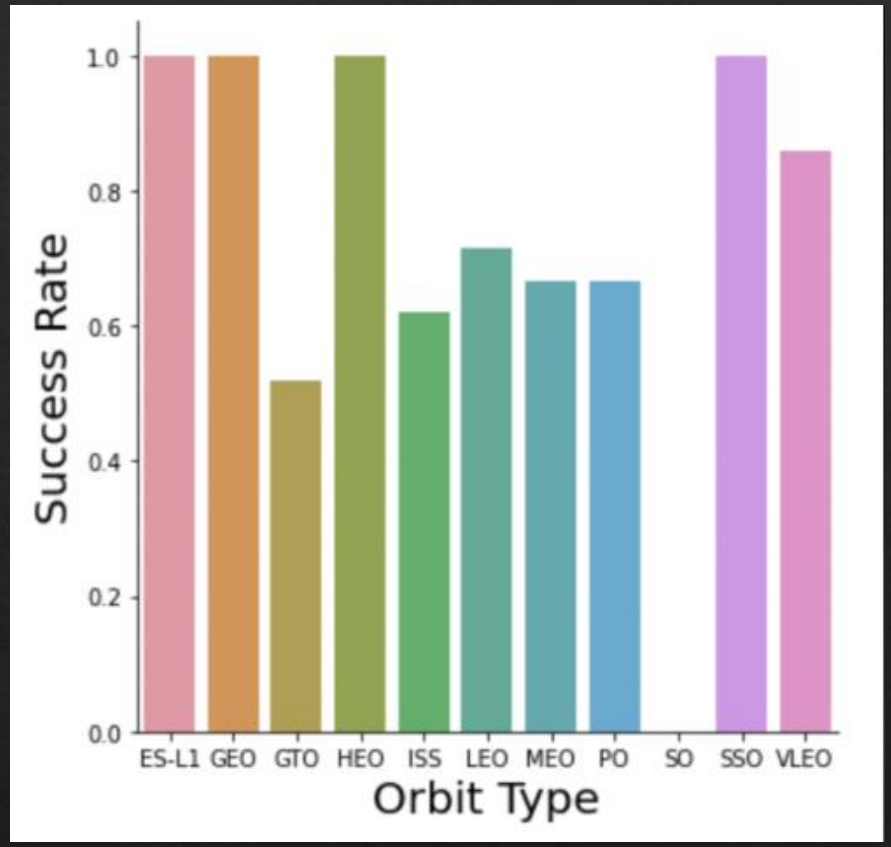
Exploratory Data Analysis in Python



Success Rates by Payload Mass and Launch Site

- Higher payload mass generally results in a higher success rate
- Most launches over 7000kg were successful
- KSC LC 39A has 100% success rate for mass < 5500kg as well

Exploratory Data Analysis in Python



Success Rates by Orbit Type

100% success:

- ES-L1, GEO, HEO, SSO

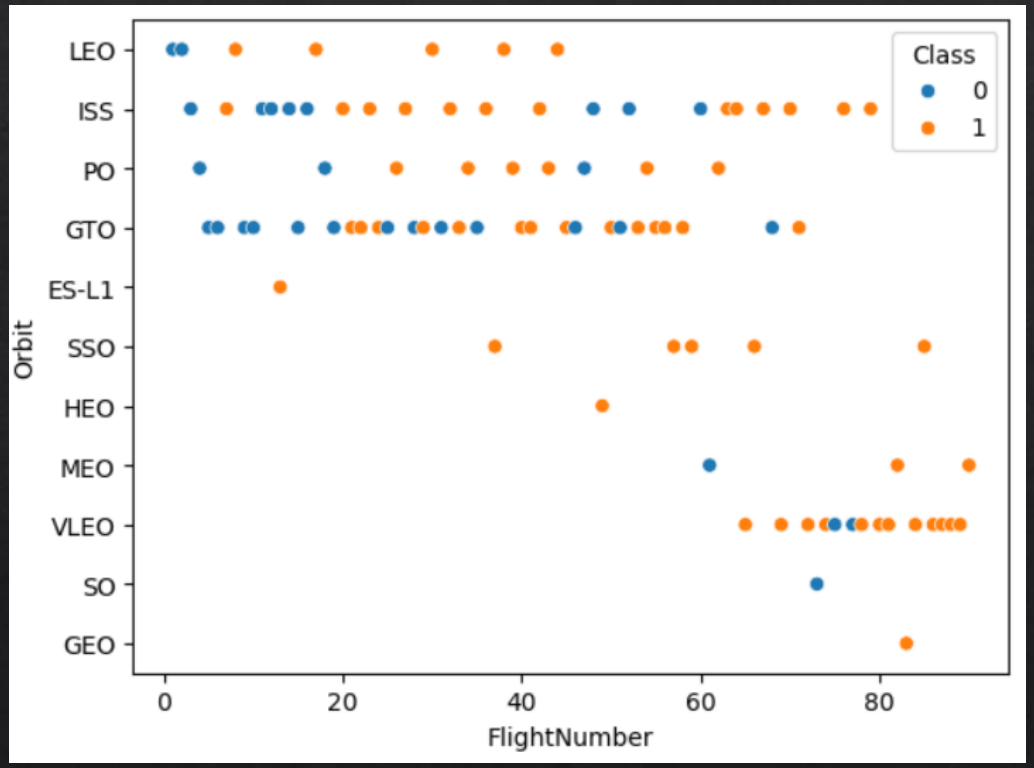
50-85% success:

- GTO, ISS, LEO, MEO, PO

0% success:

- SO

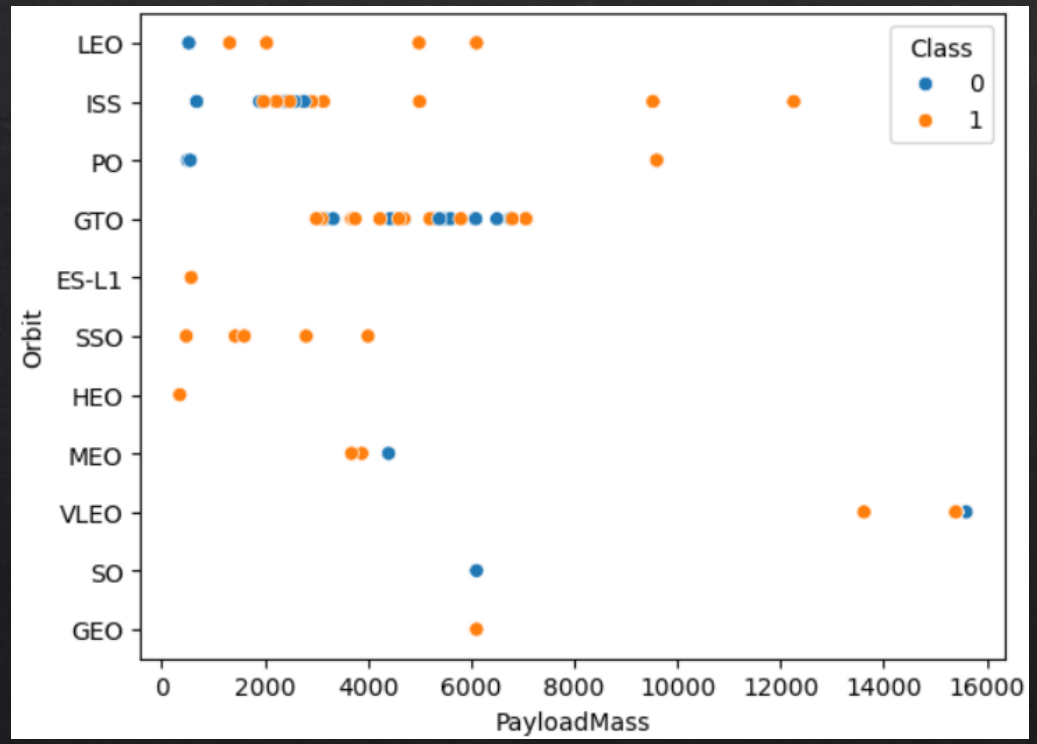
Exploratory Data Analysis in Python



Success Rates by Flight Number and Orbit Type

- The LEO orbit's success is related to the number of flights
- There is no relationship between flight number in GTO orbit
- ES-L1, GEO, and HEO orbit types succeeded on their only launch
- SSO is 100% successful across 5 launches

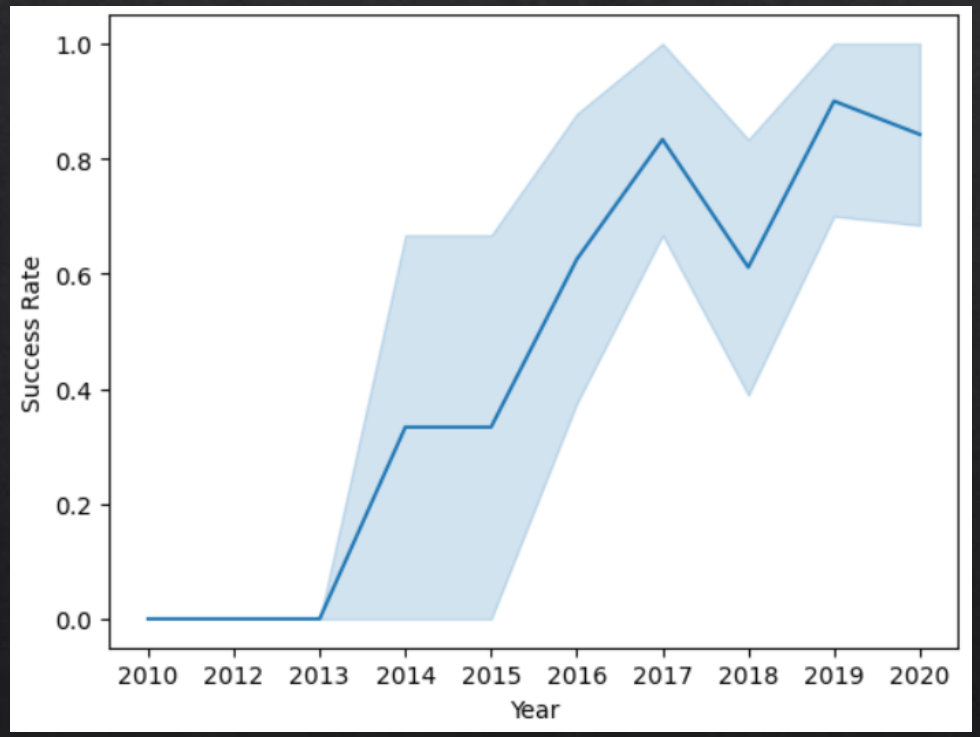
Exploratory Data Analysis in Python



Success Rates by Payload Mass and Orbit Type

- If the payload mass is greater than 8000kg the Polar, LEO, and ISS orbit types appear to be more successful
- GTO seems to have no relationship between payload mass and success rate

Exploratory Data Analysis in Python



Success Rates 2013-2020

- The success rate of SpaceX Falcon 9 rocket launches has clearly increased over time

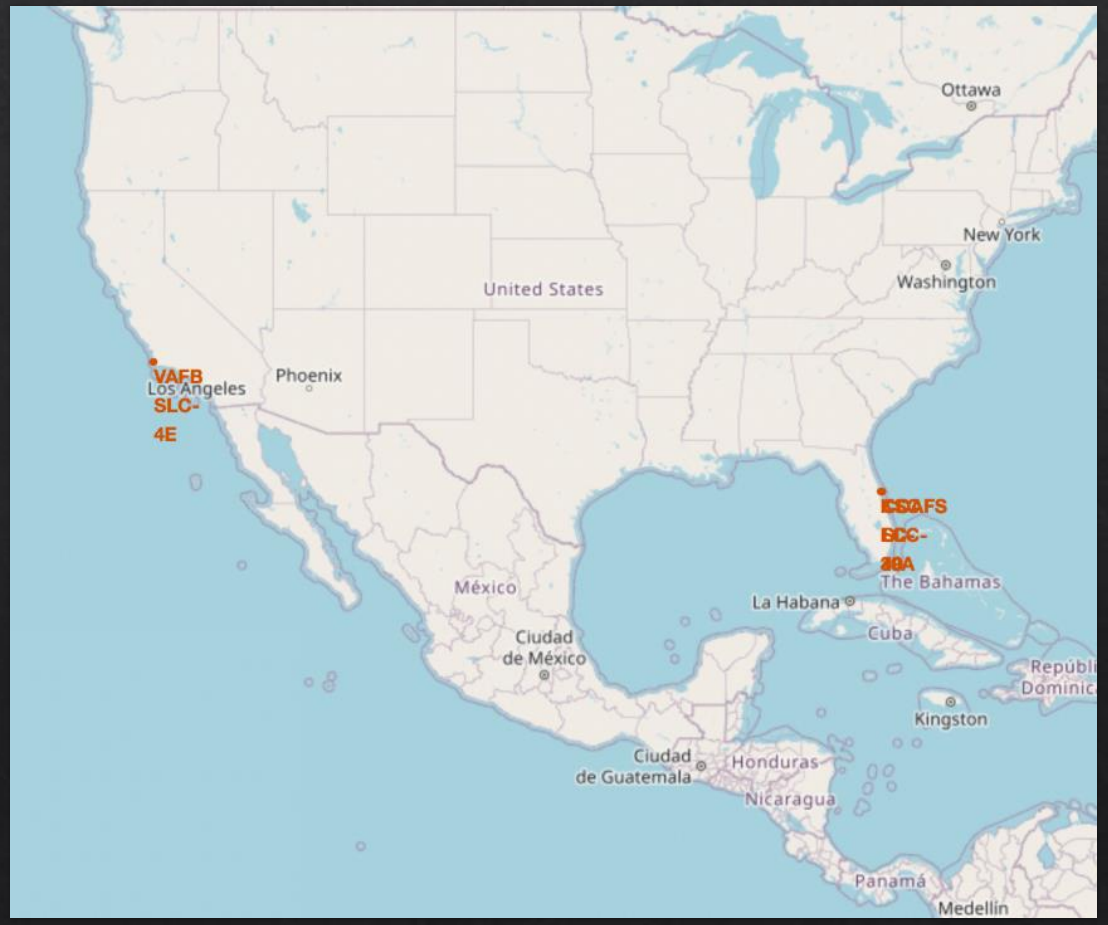
Geospacial Data Visualization

- Create an interactive map using folium to visualize the success rate of launches and proximity to geographical locations of interest

Geospacial Data Visualization

This map displays the location of SpaceX launch sites

- Note that all launch sites are in proximity to the equator line
  - The land at the equator is moving 1670 km per hour, so launching from the equator makes the spacecraft move faster and stay in orbit once it is launched
- Note that all launch sites are very close to the coast
  - Rockets launched over the ocean reduce the risk of debris dropping or exploding near people

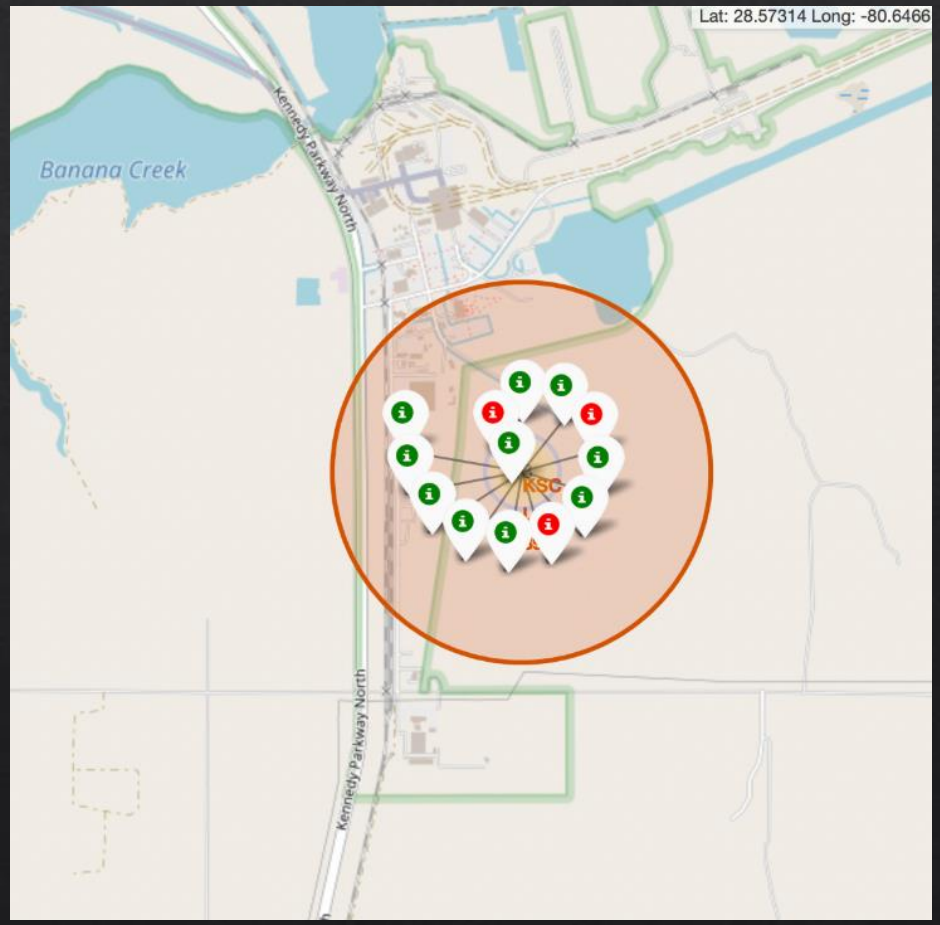


Geospacial Data Visualization

This map displays the success rates of launches at the KSC LC-39A launch site

- Green = Successful Launch
- Red = Failed Launch

We can clearly verify from this graphic that the KSC LC-39A launch site has a very high success rate

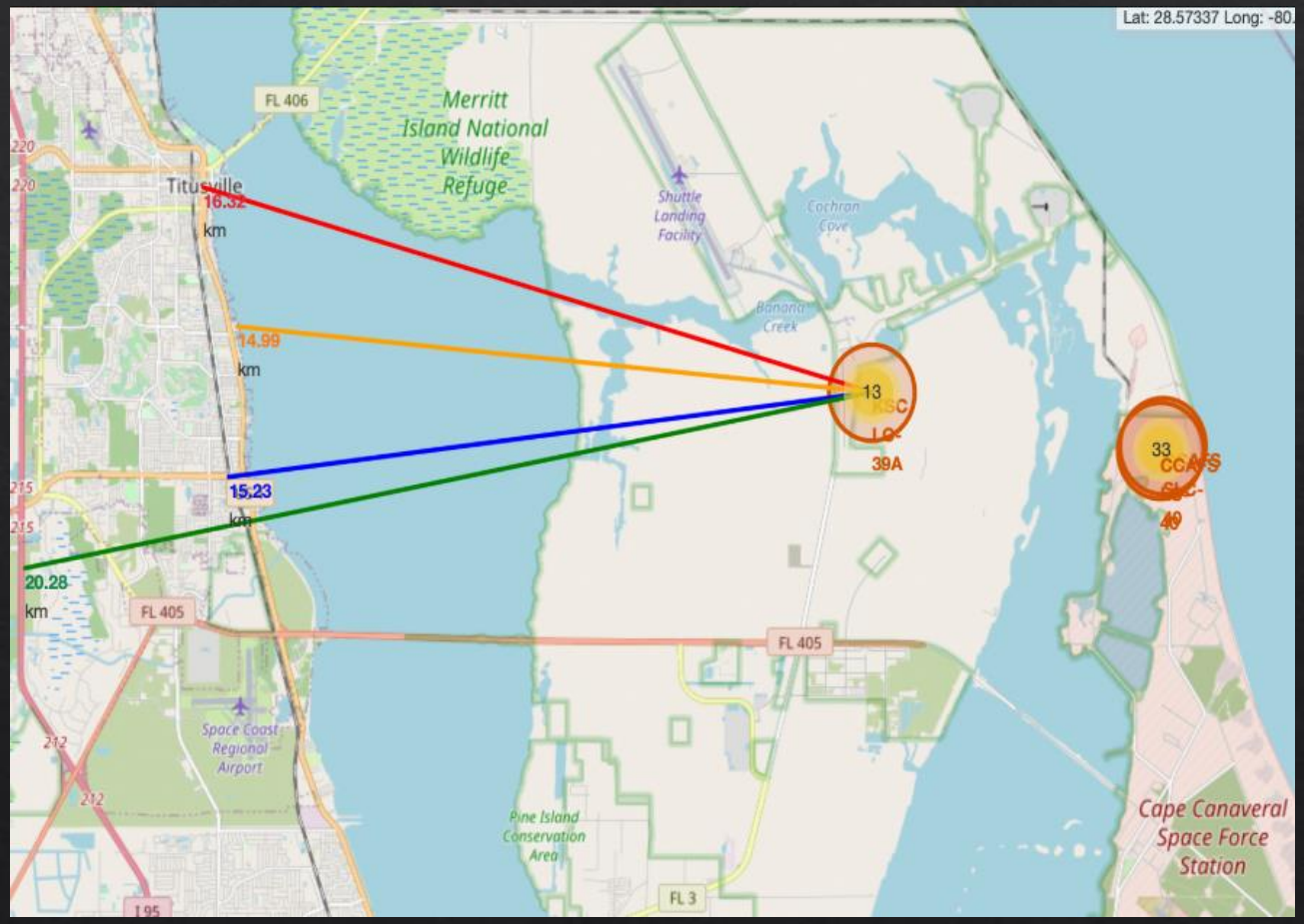


Geospacial Data Visualization

This map displays the distance from KSC LC-39A launch site to some important geographic locations:

- Coastline – 14.99km
- Railway – 15.23km
- Titusville – 16.32km
- Highway – 20.28km

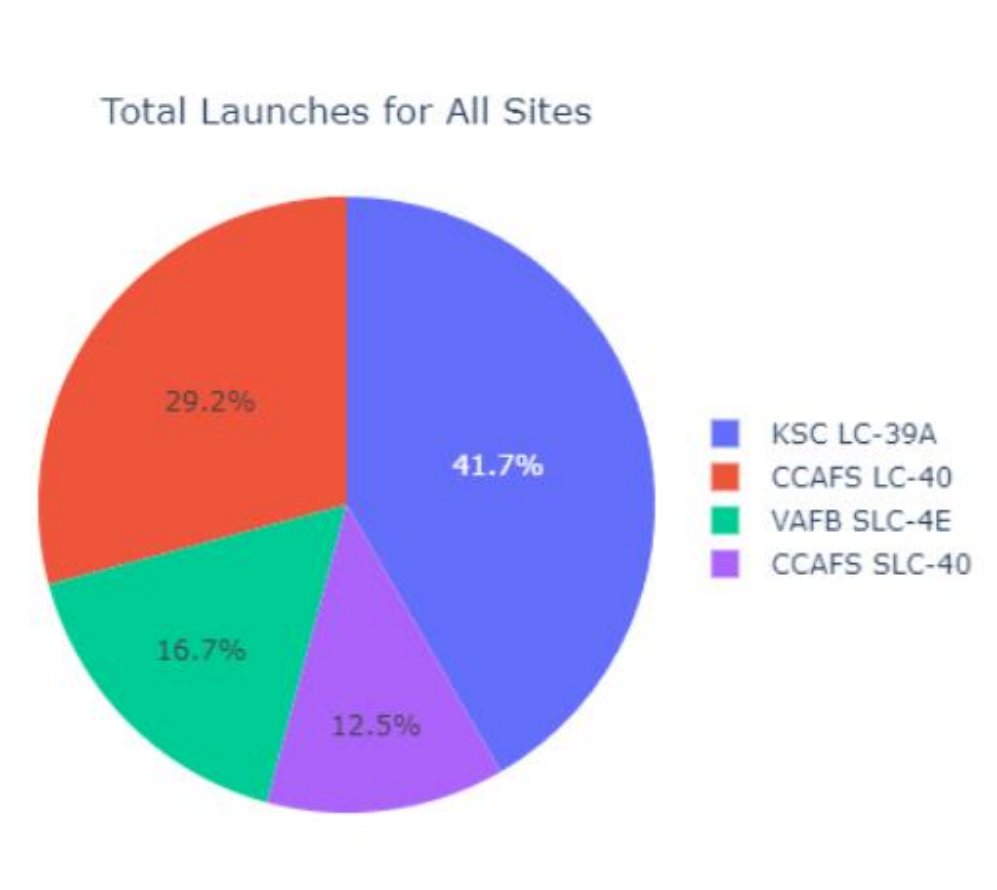
Failed rockets can cover distances such as 15-20km in a matter of seconds, which could be dangerous in populated areas.



Launch Analysis Dashboard

- Create an interactive dashboard using Plotly Dash to summarize findings from EDA process

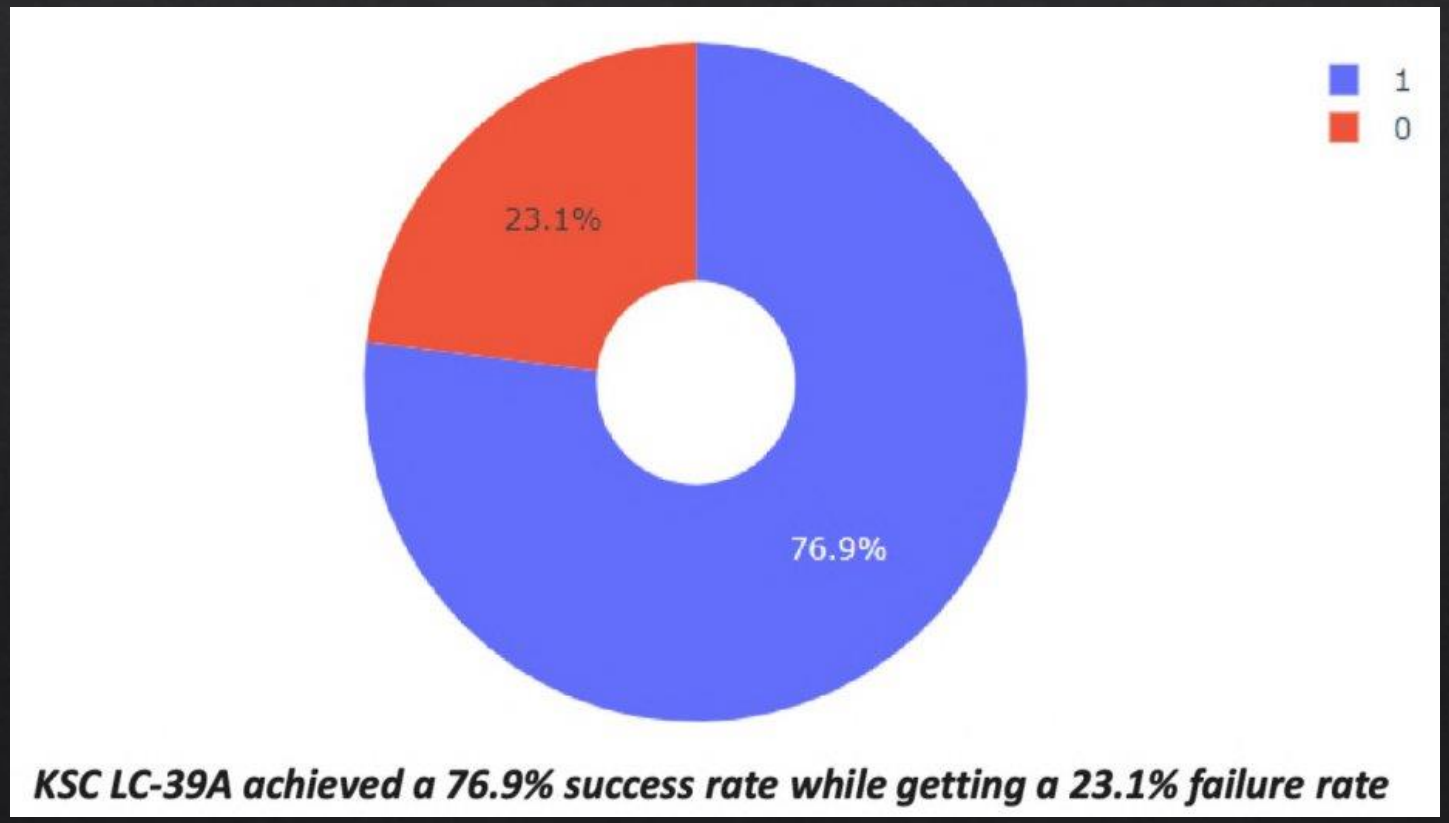
Launch Analysis Dashboard



This screenshot from the Launch Analysis Dashboard displays the total number of launches organized by launch site.

- We can see that the KSC LC-39A launch site accounts for 41.7% of all launches
- We could assume that the high success rate at KSC LC-39A might be correlated with the high number of launches

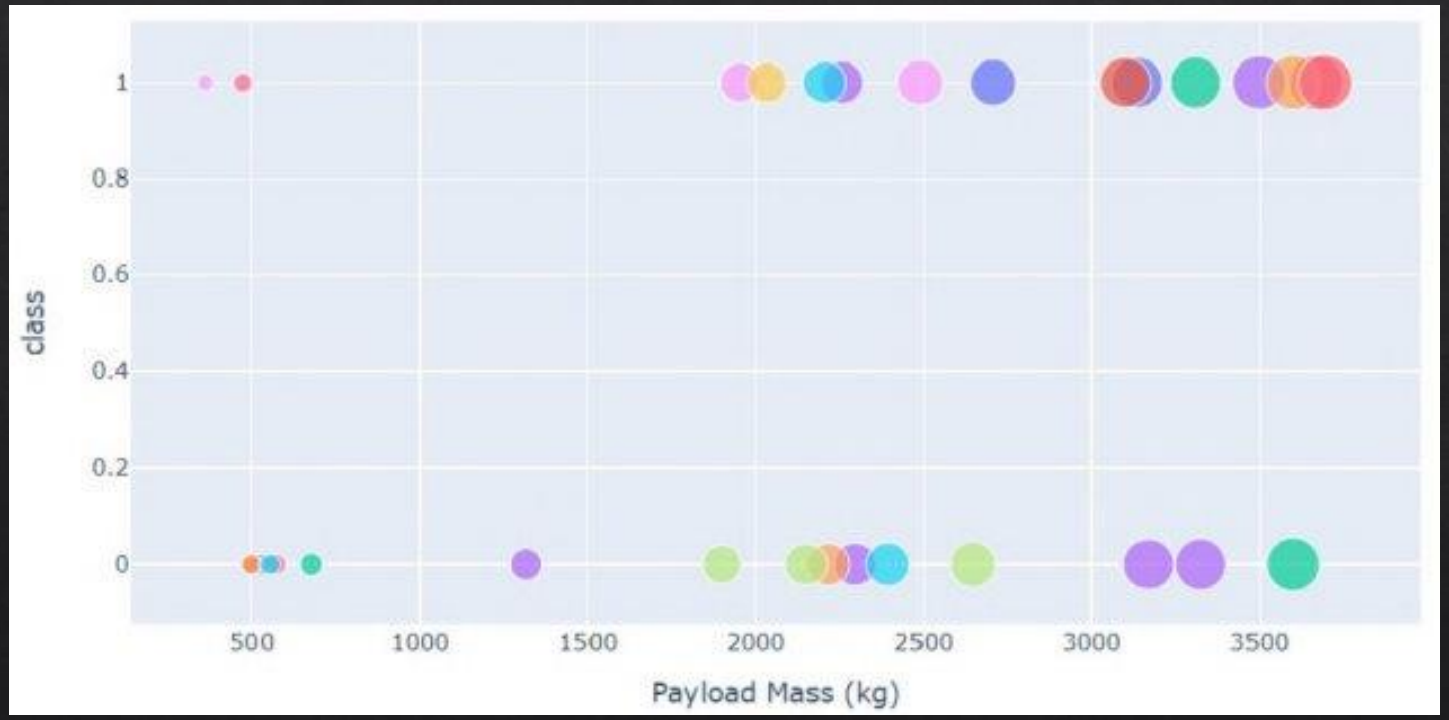
Launch Analysis Dashboard



This screenshot from the dashboard visualizes the success rate of launches from KSC LC-39A.

- We can see that this site had a 76.9% success rate, the highest among all launch sites

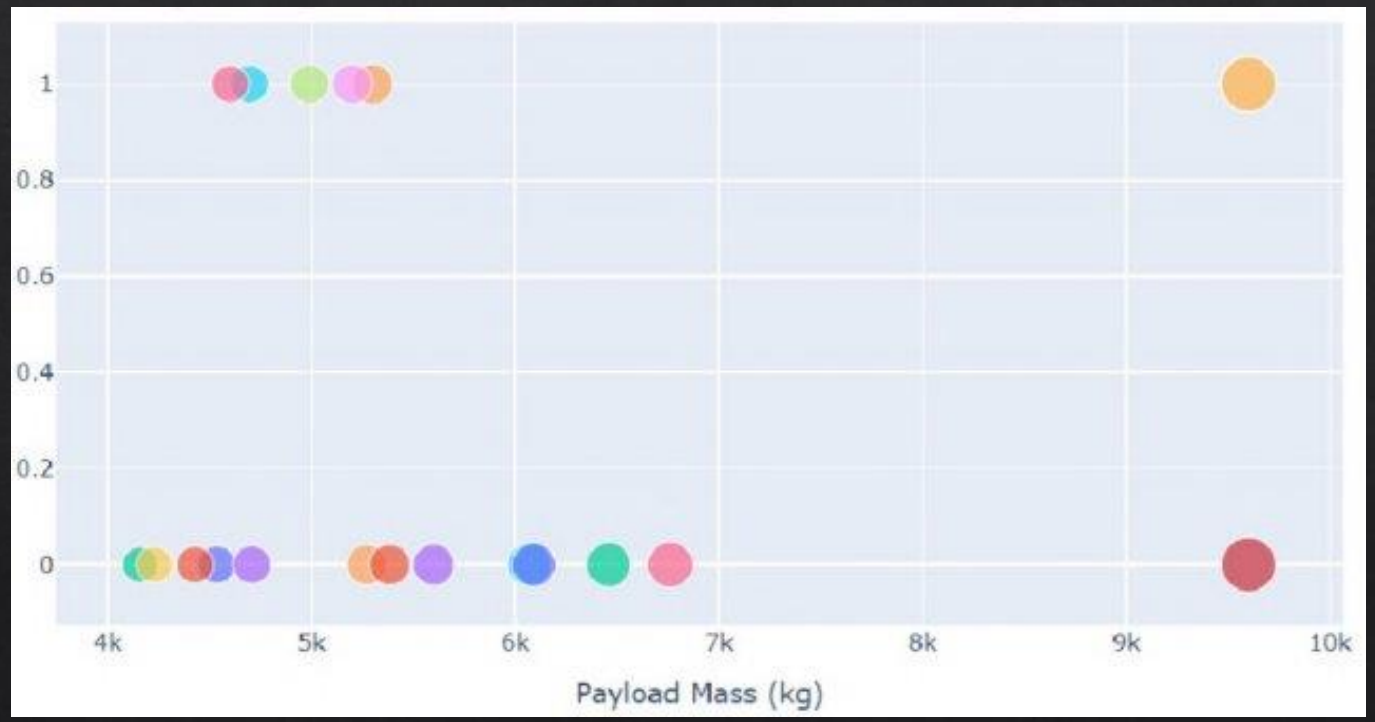
Launch Analysis Dashboard



This chart visualizes success rate by orbit type and payload mass.

- 1=Success and 0=Failure
- When payload mass is too low, the launches are prone to failure.

Launch Analysis Dashboard



This chart also visualizes success rate by orbit type and payload mass.

- 1=Success and 0=Failure
- When payload mass is too high, the launches are also prone to failure.

Launch Analysis Dashboard

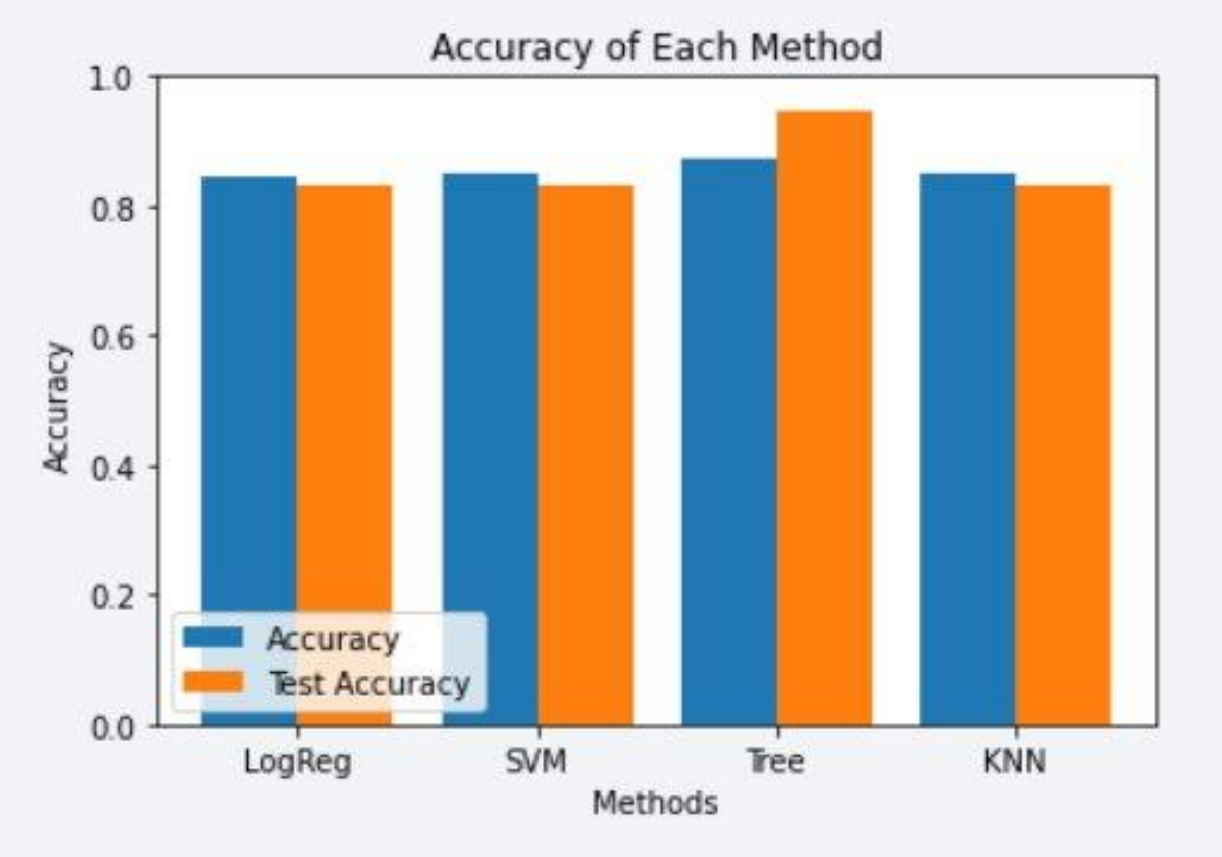
We have confirmed the following from the launch analysis dashboard:

- KSC LC-39A has the highest number of launches (41.7% of total)
- KSC LC-39A is the most successful launch site (76.9% success rate)
- If payload mass is too low or too high, the launches are more prone to failure
- A payload mass between 2000kg and 5000kg appears to offer the highest change of success

Predictive Analysis

- Separate the dataset into training and testing sets
- Train classification machine learning models and calculate accuracy on the test data to determine most effective model

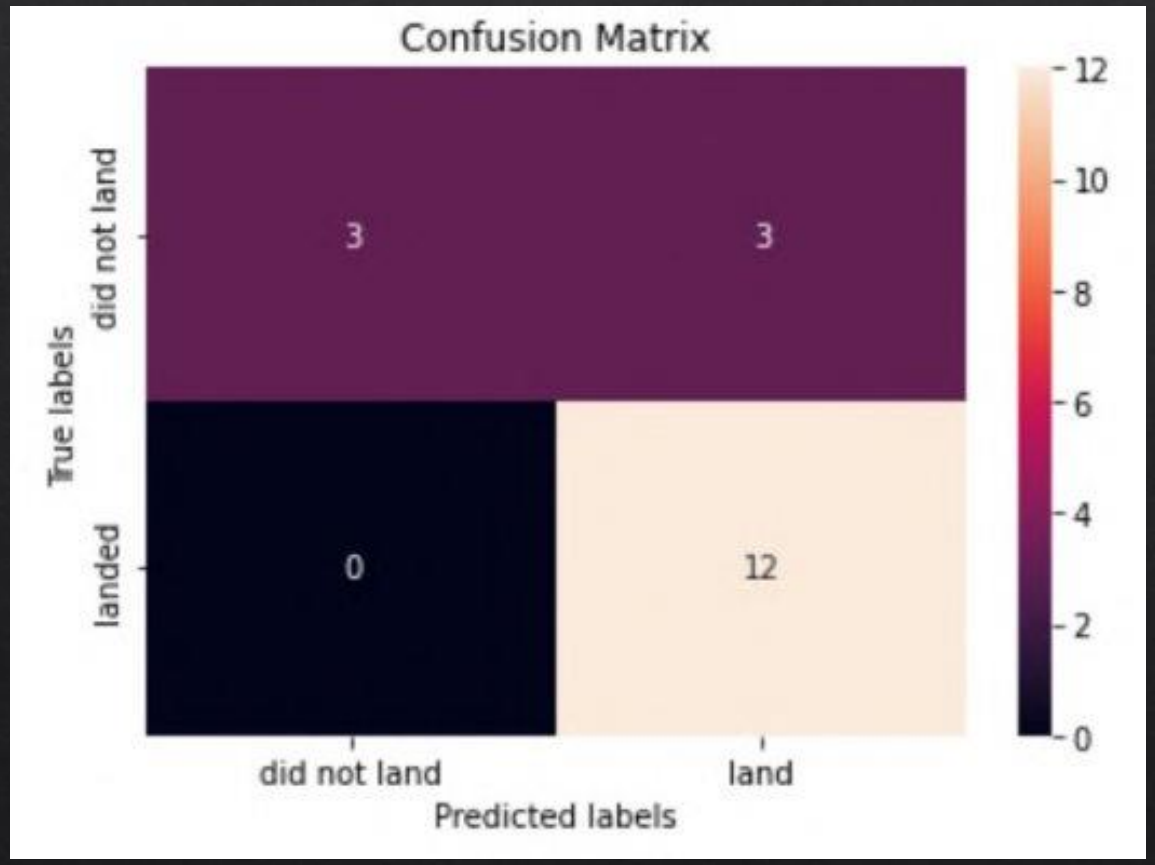
Predictive Analysis



We tested four different supervised machine learning classification algorithms on the dataset.

The bar chart to the left clearly shows that a Decision Tree Model resulted in the highest accuracy in predicting the success or failure of future SpaceX Falcon 9 rocket launches.

Predictive Analysis



The confusion matrix displayed on the left indicates that a decision tree model was 83.33% accurate in predicting whether rockets **will** land successfully compared to the testing data set.

The model is 100% accurate when predicting a rocket **will not** land successfully.

Conclusions

- Chances of success have increased over time
- KSC LC 39A is the most successful launch site with 76.9% success rate
- Medium payloads (2000-5000kg) succeed most often compared to small or large payloads
- The most accurate model for predicting the success of launches is a Decision Tree Model
- The Decision Tree model is very accurate (83.33%) at predicting whether a rocket will successfully land

Appendix

- https://www.coursera.org/professional-certificates/ibm-data-science
- https://github.com/IAaqib78/Labs
- https://api.spacexdata.com/v4/launches/past
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

IBM Data Science Capstone Github Repository


Thank you.