IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ted Letherer

12/21/2023

Outline

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

Executive Summary

This data science project aims to predict the successful landing of SpaceX's Falcon 9 first stage using data scraped from the web. The motivation behind this project is the significant cost savings associated with the reusability of the Falcon 9's first stage.

SpaceX advertises Falcon 9 rocket launches on its website at a cost of $62 million, while other providers' costs exceed $165 million each. A substantial portion of these savings is due to SpaceX's ability to reuse the first stage of the Falcon 9.

By predicting whether the first stage will land successfully, we can estimate the cost of a launch. This information could be invaluable to alternative companies wishing to compete with SpaceX for rocket launch contracts.

The project utilized Exploratory Data Analysis (EDA) and Machine Learning methods to analyze the data and build predictive models. After evaluating various models, a **Decision Tree model** was found to be the most effective, achieving an impressive **accuracy of 0.90** in predicting the success or failure of the first stage landing.

The results of this project could potentially disrupt the market dynamics in the space industry by providing competitors with valuable insights into SpaceX's cost structure.

The project underscores the power of data science in driving business strategy and competitiveness in the high-stakes world of space exploration. The success of the Decision Tree model in this project highlights the potential of machine learning in making accurate predictions that can inform strategic decisions.

Introduction

The advent of the commercial space age has revolutionized the realm of space travel, making it more affordable and accessible than ever before. Among the trailblazers in this field, SpaceX has distinguished itself through its innovative approach and cost-effective solutions, particularly with its Falcon 9 rocket launches.

A significant factor contributing to SpaceX's cost-effectiveness is the reusability of the Falcon 9's first stage. By successfully landing the first stage, SpaceX can reuse it for subsequent launches, resulting in substantial cost savings. This unique capability sets SpaceX apart from other providers whose costs often exceed $165 million per launch.

In this project, we leverage data science techniques to predict the successful landing of the Falcon 9's first stage. This prediction is not merely an academic exercise; it has practical implications for estimating the cost of a launch. Such information is invaluable for new entrants in the space industry, like Space Y, that aim to compete with SpaceX.

To make these predictions, we delve into various factors that could influence the success of a launch. We also analyze the trend of mission success over time to understand if more recent missions have been more successful than earlier ones. Furthermore, we aim to identify the ideal payload range for each orbit type, which can provide insights into optimizing mission planning and resource allocation.

The key questions we aim to answer through this project are:

1. What are the factors that most influence success?
2. Have more recent missions been more successful than earlier ones?
3. Is there an ideal payload minimum and maximum for each orbit type?

By answering these questions, we hope to provide strategic insights that can drive decision-making in the competitive space industry. This project underscores the potential of data science in shaping the future of space exploration and commercial competitiveness. It highlights how data-driven insights can inform strategic decisions, optimize resources, and potentially disrupt market dynamics in the high-stakes world of space exploration.

Section 1

Methodology

Methodology

Executive Summary

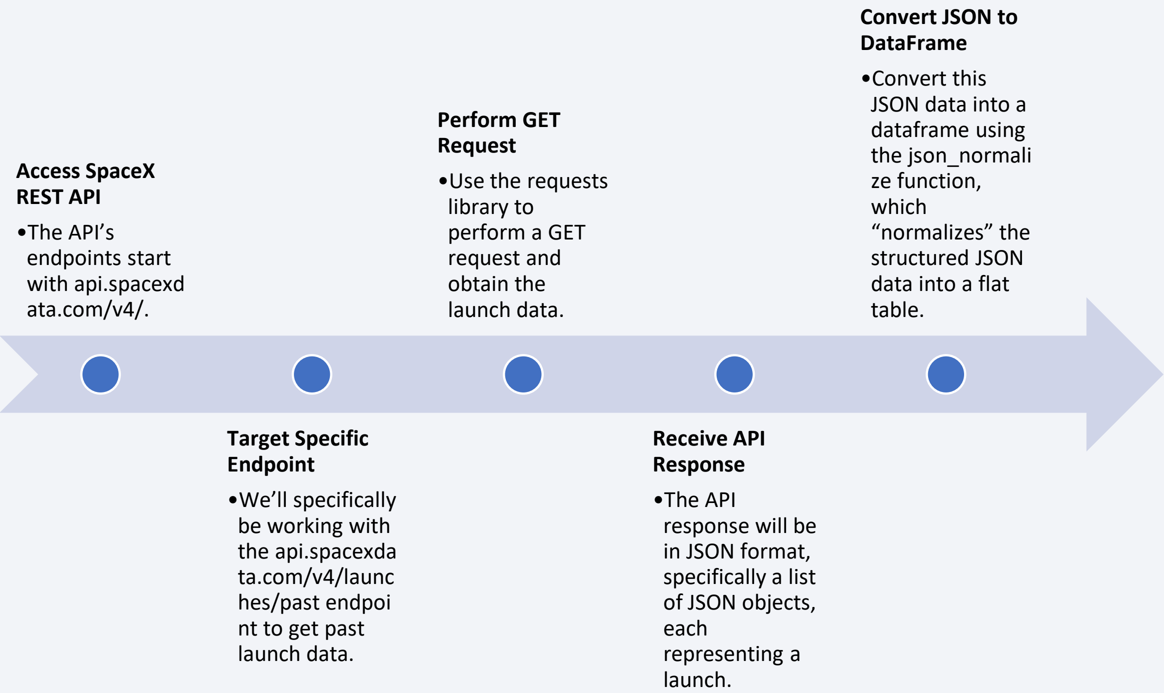
- Data collection methodology:
  - API
  - Web Scraping
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

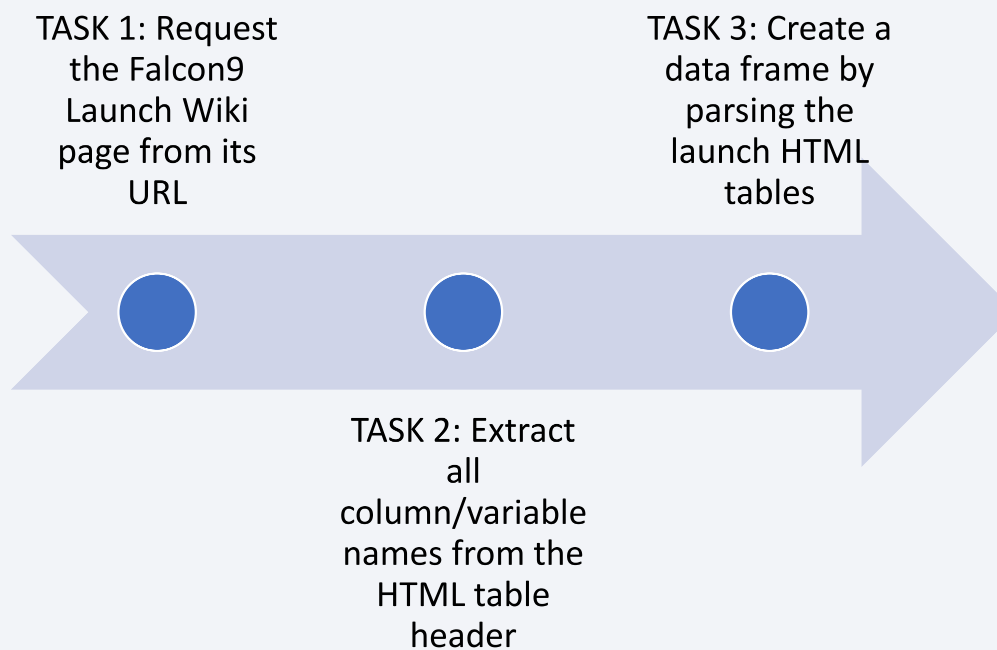
Data Collection – SpaceX API

- The API's endpoints start with api.spacexdata.com/v4/, and we'll specifically be working with the api.spacexdata.com/v4/launches/past endpoint to get past launch data. We'll perform a GET request using the requests library to obtain this data.
- The API response will be in JSON format, specifically a list of JSON objects, each representing a launch. We'll then convert this JSON data into a dataframe using the json_normalize function, which "normalizes" the structured JSON data into a flat table. This process transforms our JSON data into a more manageable table format for further analysis.
- https://github.com/indyted314/urban-tomato-soup/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



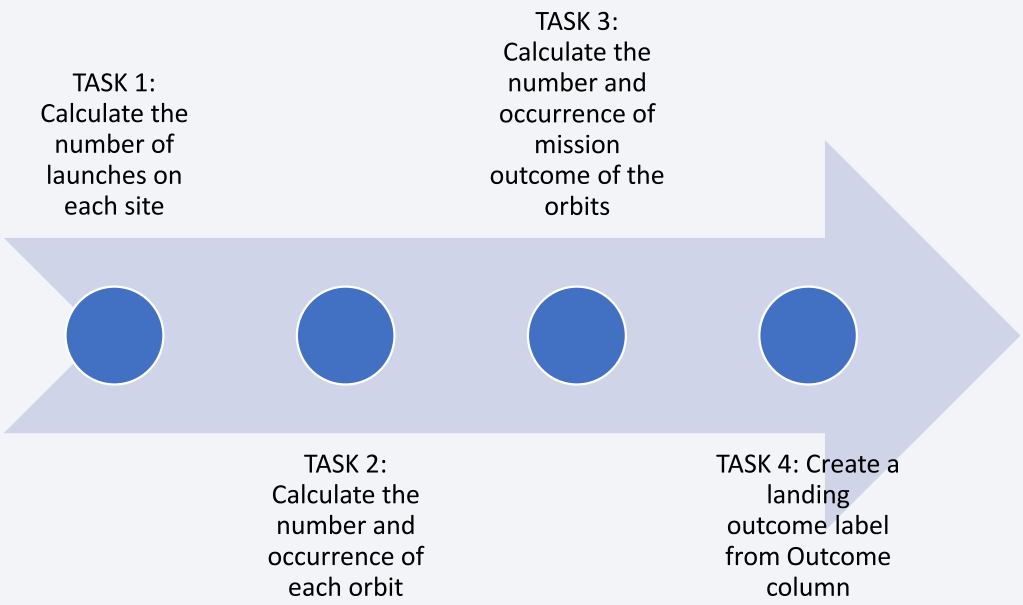
Data Collection - Scraping

- Another popular data source for obtaining Falcon 9 Launch data is web scraping related Wiki pages. We used the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.
- https://github.com/indyted314/urban-tomato-soup/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- The data from the tables needs to be parsed and converted into a Pandas dataframe for further analysis.
- The goal is to transform this raw data into a clean dataset that can be used for data wrangling, sampling, and dealing with null values. Some columns, like ‘rocket’, contain ID numbers instead of actual data, requiring the use of the API again to gather specific data for each ID. This data, related to the booster, launchpad, payload, and core, is stored in lists for dataset creation. The launch data includes Falcon 1 booster data, which needs to be filtered out to focus on Falcon 9. Null values in the data, specifically in the ‘PayloadMass’ column, are replaced with the mean of the PayloadMass data. Null values in the ‘LandingPad’ column, representing when a landing pad is not used, will be dealt with later using one-hot encoding.
- https://github.com/indyted314/urban-tomato-soup/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



---Internal Use---

EDA with Data Visualization

- A number of different charts were plotted to explore the data including:
  - FlightNumber vs. PayloadMass
  - FlightNumber vs LaunchSite
  - Payload Vs. Launch Site
  - Success Rate of each Orbit Type
  - Flight Number and Orbit Type.
  - Orbit Type vs Payload
  - Success Rate since 2013 to 2020
- https://github.com/indyted314/urban-tomato-soup/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- The following SQL queries you performed:
    - Display the names of the unique launch sites in the space mission
    - Display 5 records where launch sites begin with the string 'CCA'
    - Display the total payload mass carried by boosters launched by NASA (CRS)
    - Display average payload mass carried by booster version F9 v1.1
    - List the date when the first successful landing outcome in ground pad was achieved.
    - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    - List the total number of successful and failure mission outcomes
    - List the names of the booster versions which have carried the maximum payload mass.
    - List 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 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.

- https://github.com/indyted314/urban-tomato-soup/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb *(Note: I did complete the notebook but it would not save. I do show my work in later slides)*

Build an Interactive Map with Folium

- We created map objects including markers, circles, lines and added these objects to a folium map
- TASK 1: Mark all launch sites on a map
- TASK 2: Mark the success/failed launches for each site on the map
- TASK 3: Calculate the distances between a launch site to its proximities

- https://github.com/indyted314/urban-tomato-soup/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- We created a dashboard in Plotly dashboard that did the following:
  - dropdown list to enable Launch Site selection
  - pie chart to show the total successful launches count for all sites
  - slider to select payload range
- https://github.com/indyted314/urban-tomato-soup/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Perform exploratory Data Analysis and determine Training Labels
  - create a column for the class
  - Standardize the data
  - Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
  - Find the method performs best using test data
- https://github.com/indyted314/urban-tomato-soup/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

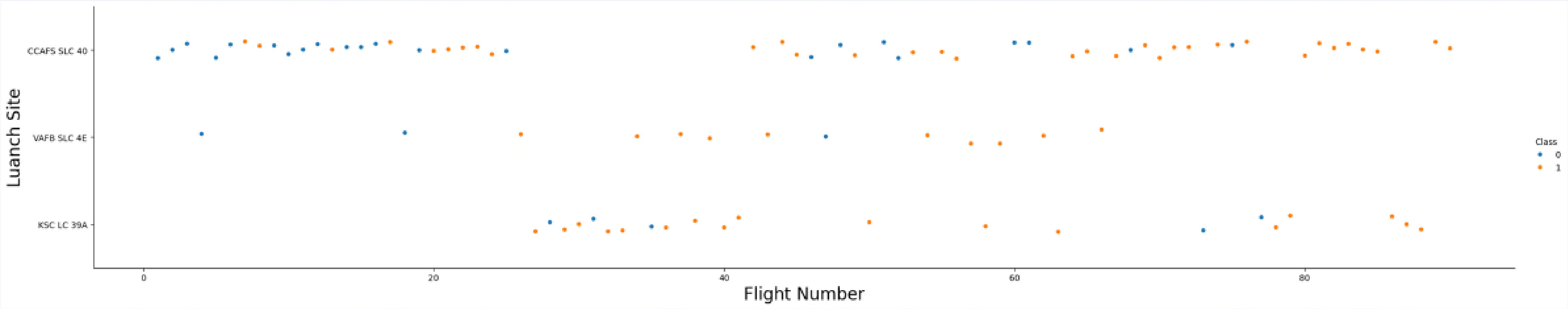
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

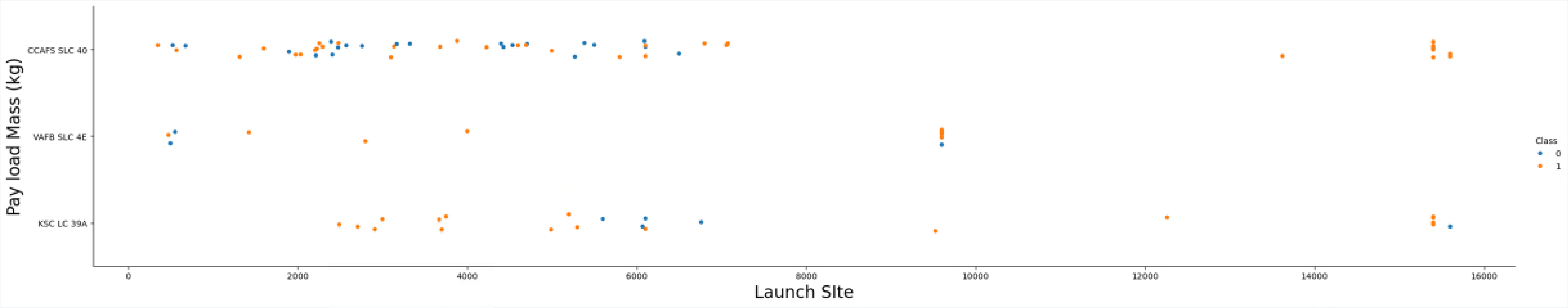
Insights drawn from EDA

Flight Number vs. Launch Site



- Not too much was learned here. Successful missions ('1' in the key) are in orange and unsuccessful missions ('0' in the key) are in blue. It does appear that there are more successful missions with the higher flight numbers. Most missions were at CCAPS SLC 40 site.

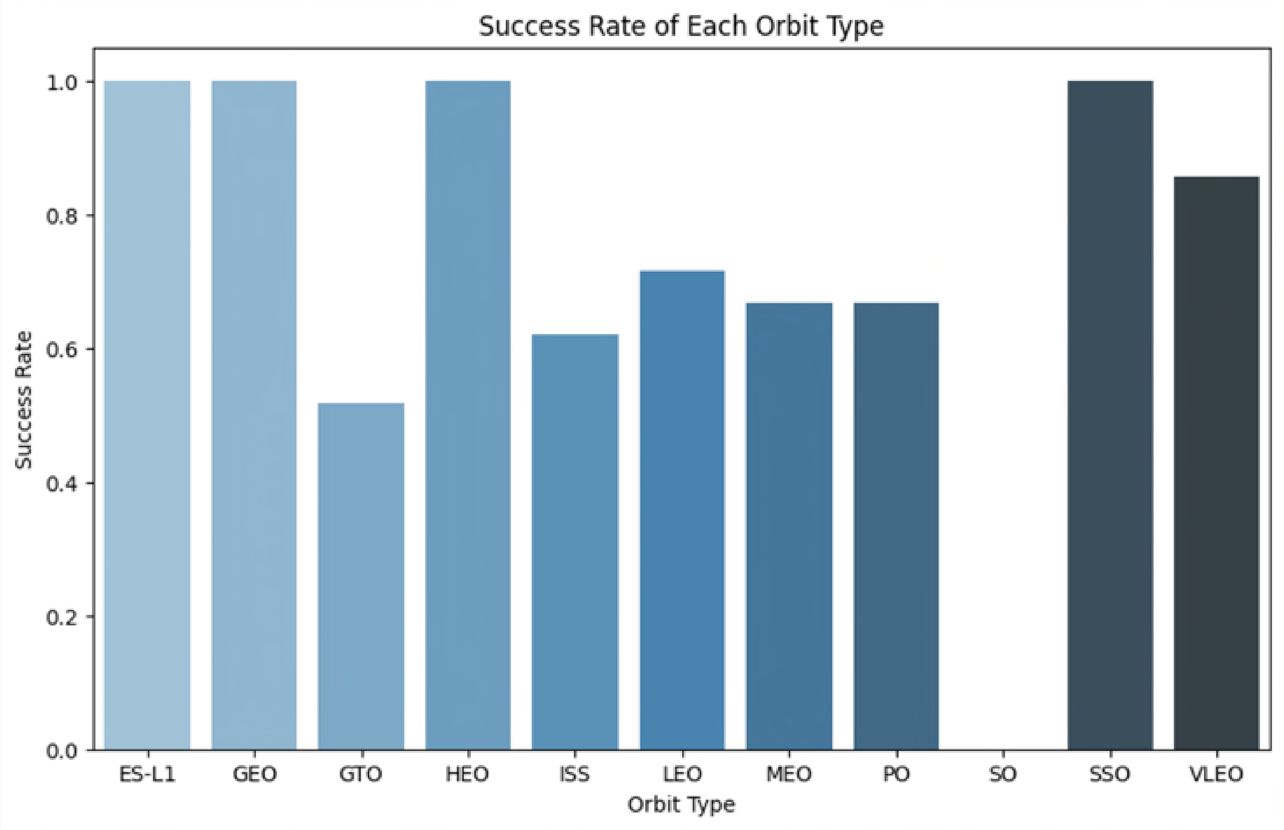
Payload vs. Launch Site



- Again, successful missions ('1' in the key) are in orange and unsuccessful missions ('0' in the key) are in blue. No other patterns emerged.

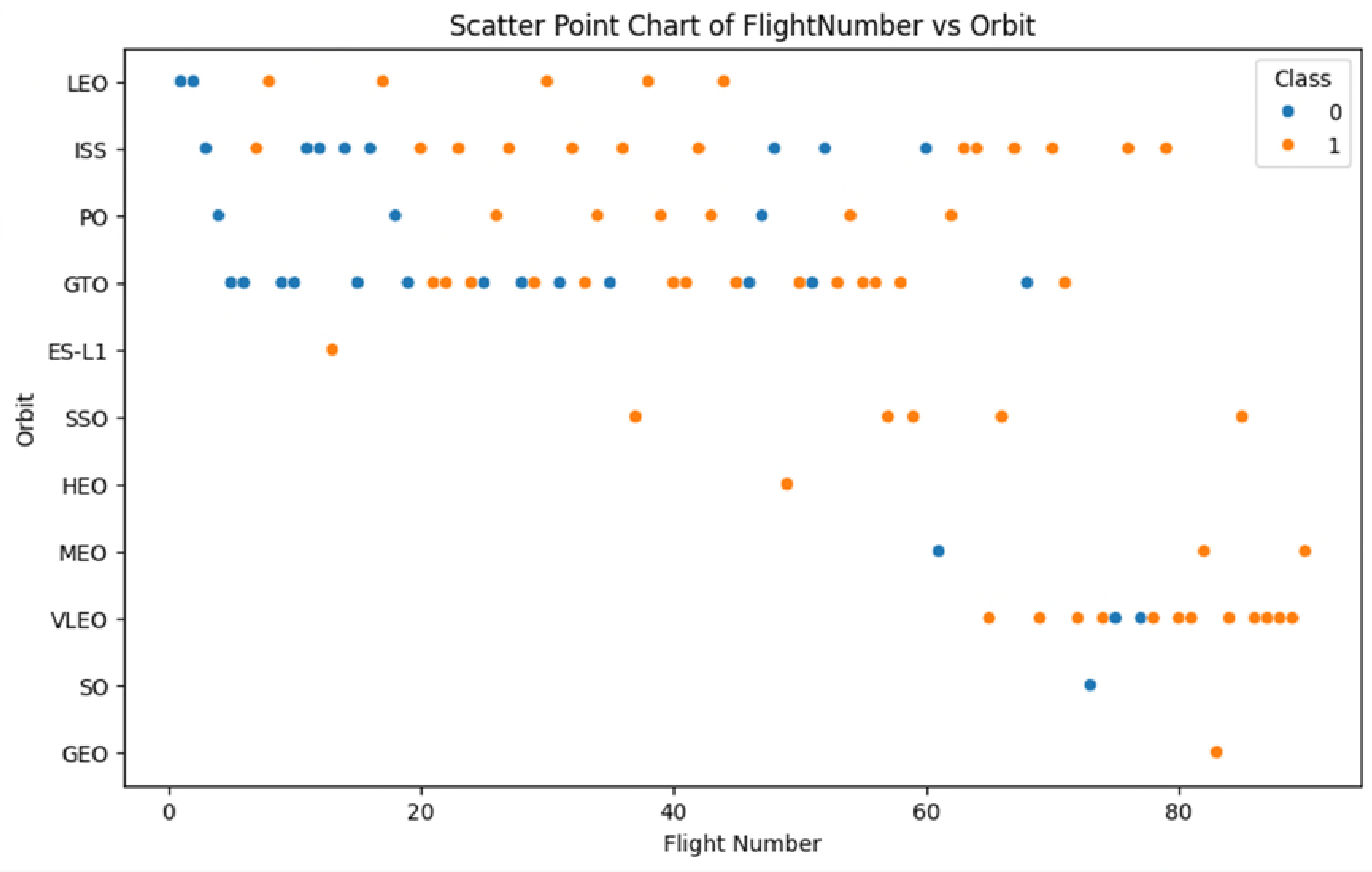
Success Rate vs. Orbit Type

- Here we see the most successful orbit types are
  - ES-L1
  - GEO
  - HEO
  - SSO



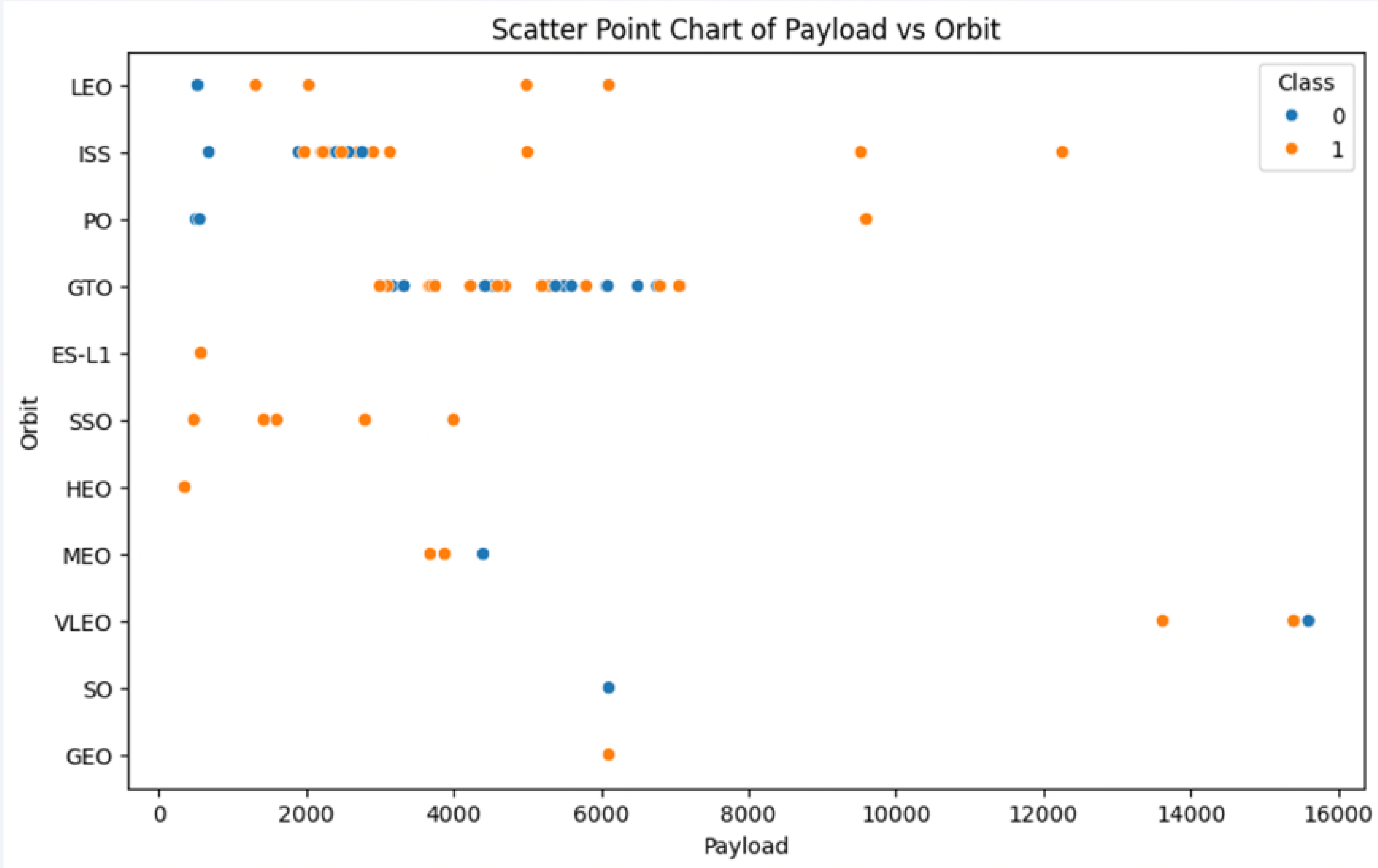
Flight Number vs. Orbit Type

- Here we get an idea of the number of flights for each orbit, their success, and the if they occurred early in the program or later on.



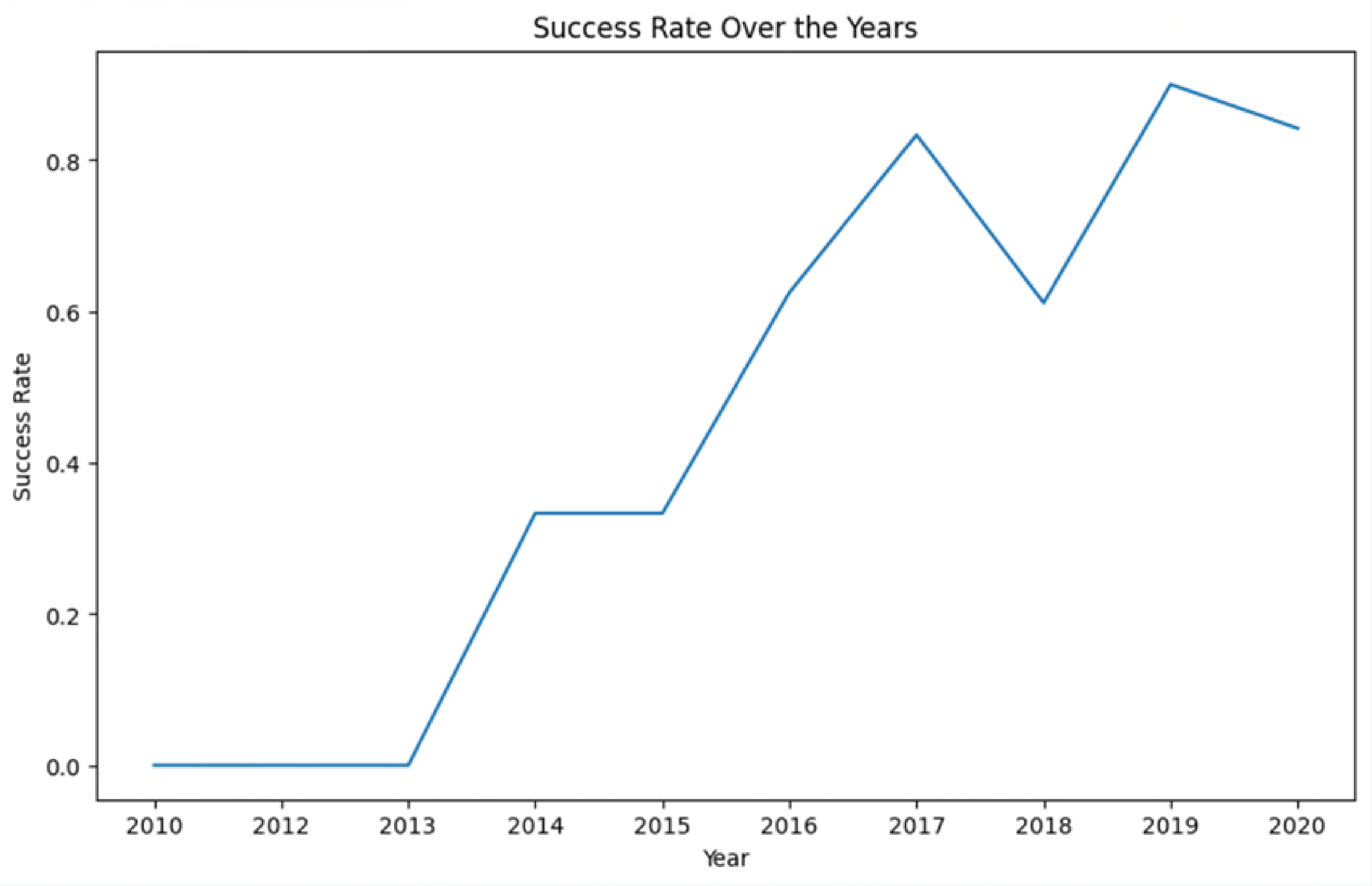
Payload vs. Orbit Type

- Here we see that the ideal payload is different for different orbit types



Launch Success Yearly Trend

- It is clear that we success rate has steadily been increasing from 20120 to 2020.



All Launch Site Names

- Name of the four unique launch sites

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

Launch Site Names Begin with 'CCA'

- 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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Total_Payload_Mass
48213

Average Payload Mass by F9 v1.1

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

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

First_Successful_Landing_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the 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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Success_Count
100

Failure_Count
1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

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

- 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

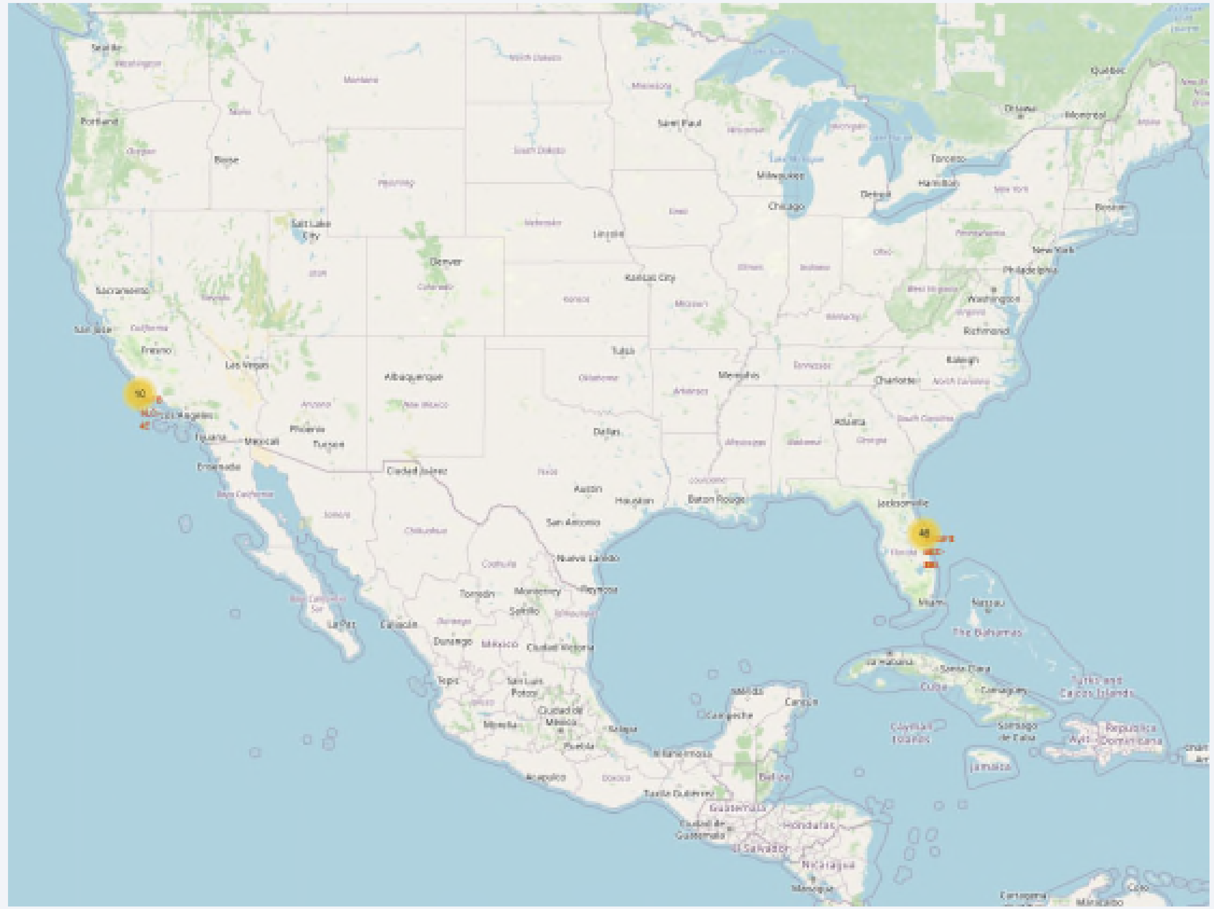
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

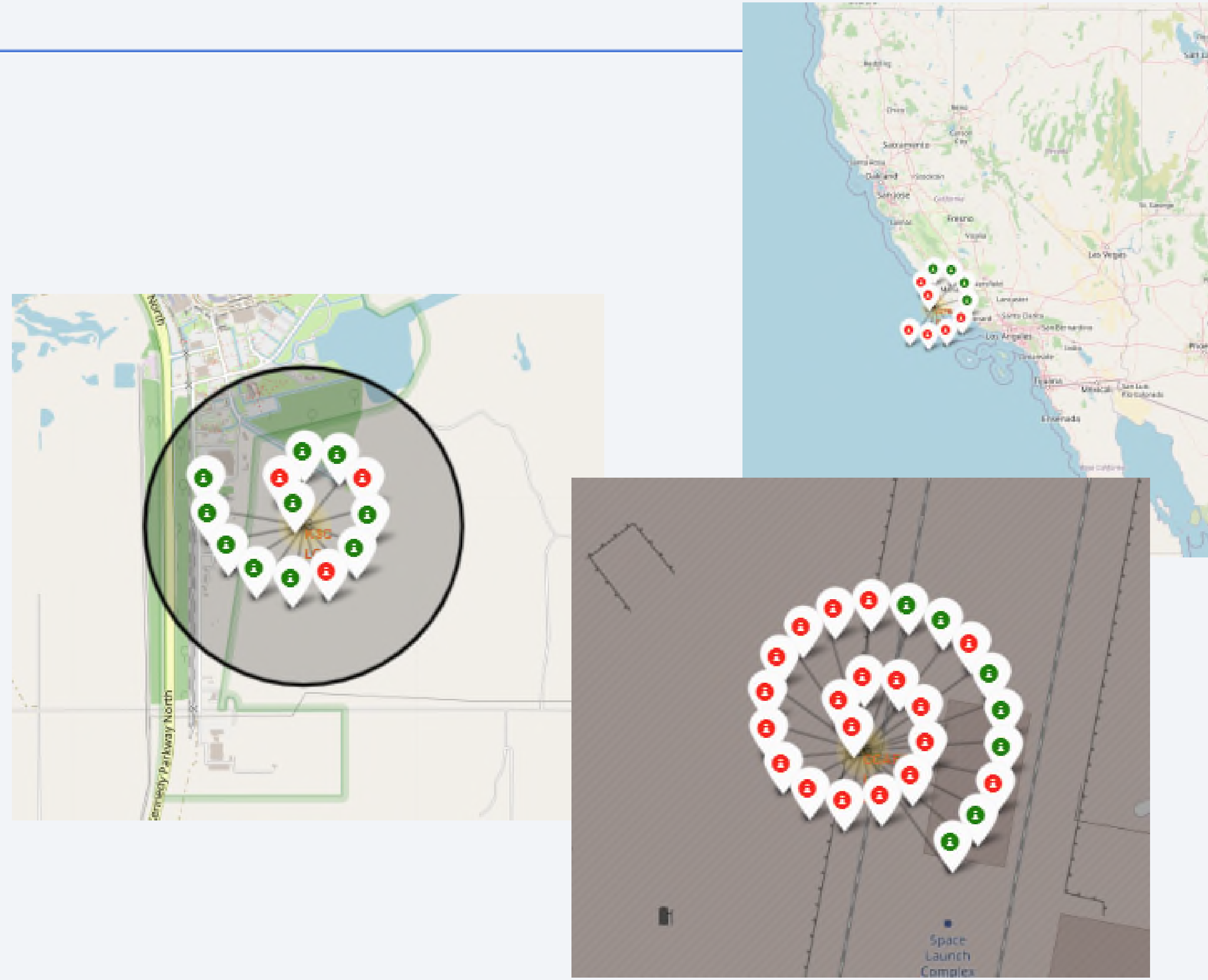
Mark all launch sites on a map

- We see here that the launch sites are near two areas, one on the Atlantic coast of Florida, the other on the Pacific coast in southern California



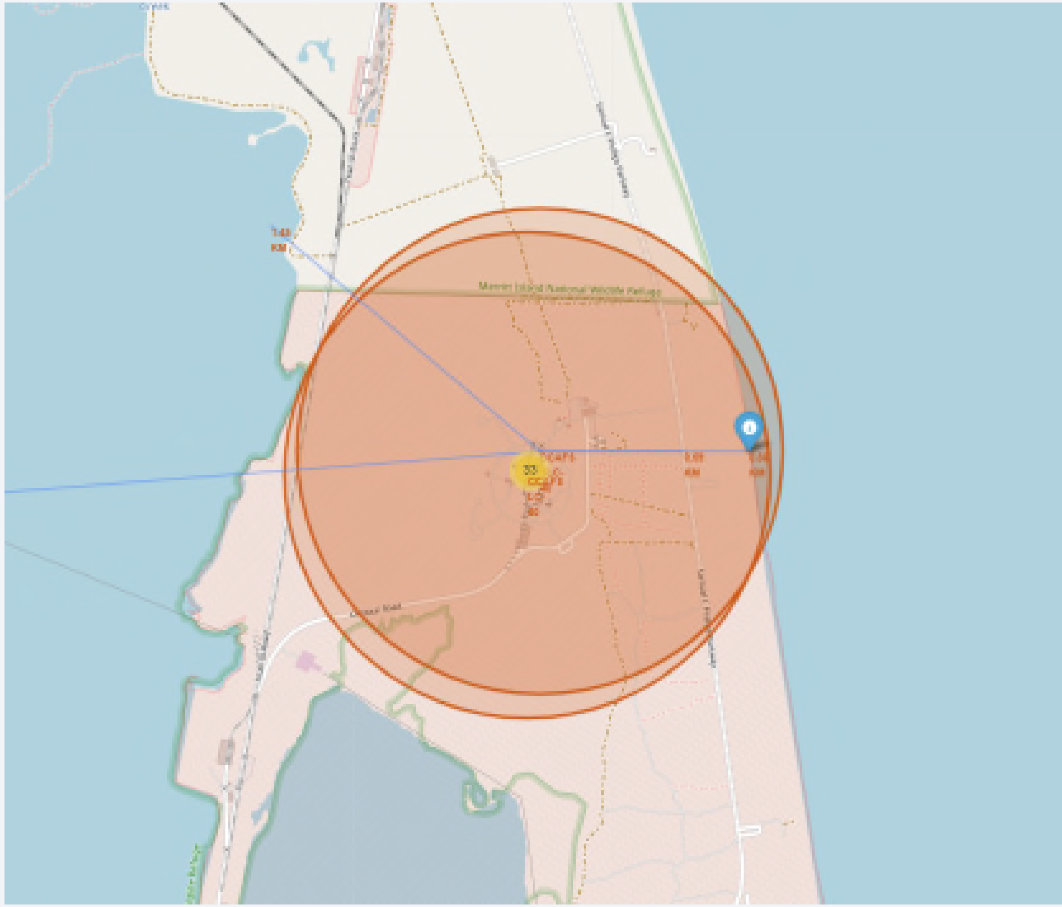
Mark the success/failed launches for each site on the map

- Each launch was marked with either red or green marker. Green markers are successful launches, red or unsuccessful



Calculate the distances between a launch site to its proximities

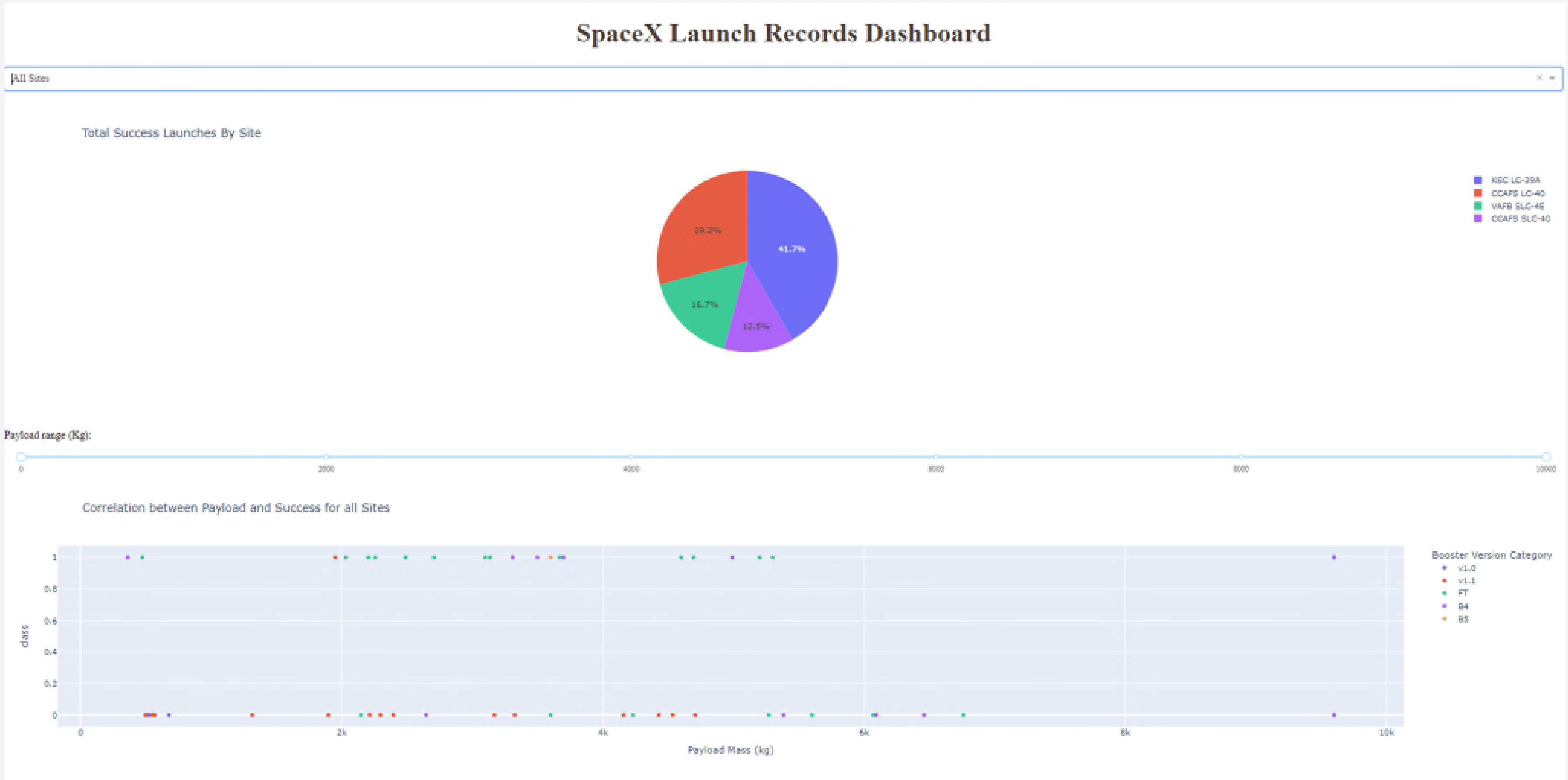
- Distances between the launch site and the nearest coast, highway, railroad, and city were calculated. Each distance was marked with a line and the distance in KM was labelled.



Section 4

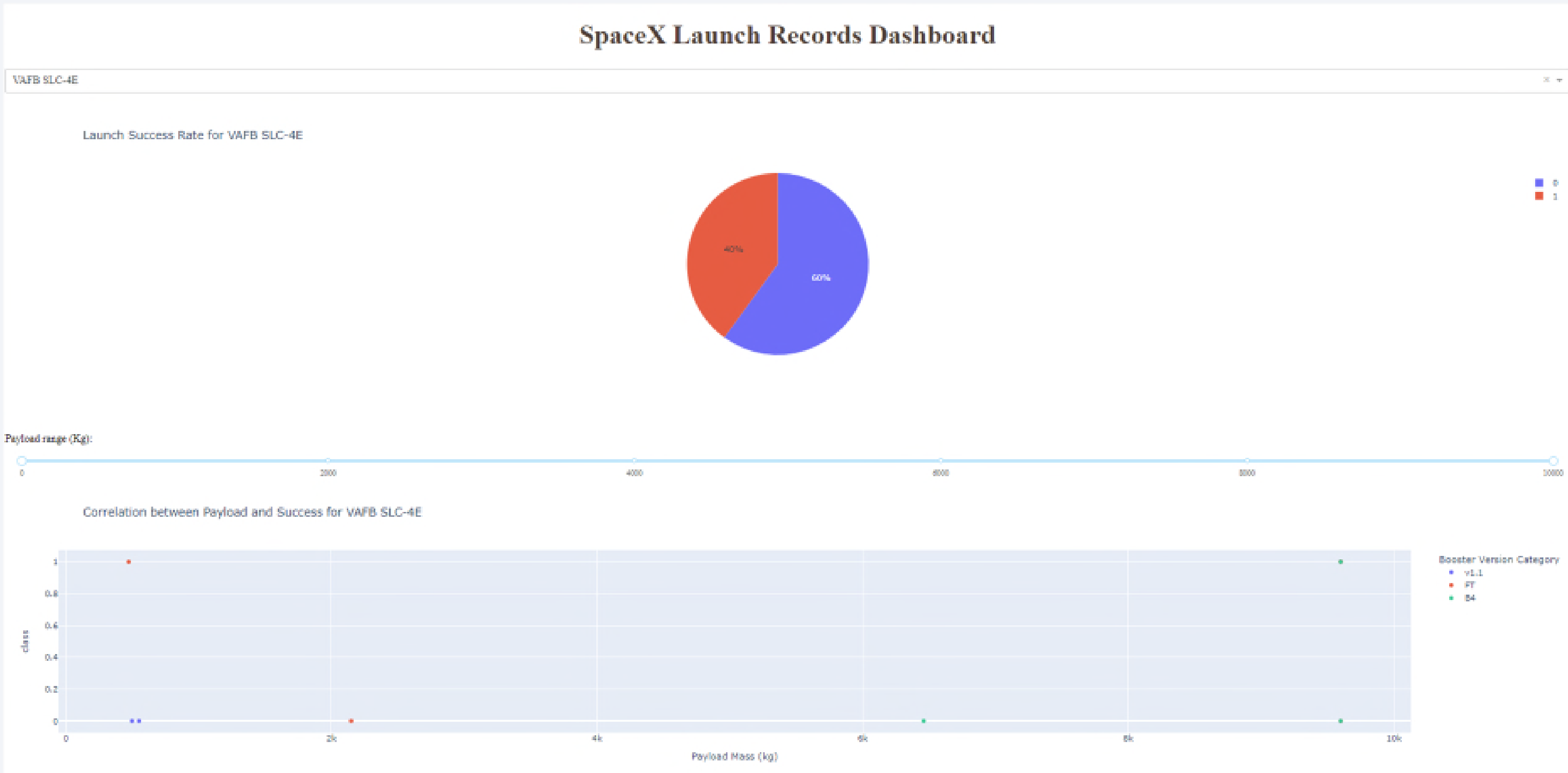
Build a Dashboard with Plotly Dash

Add a dropdown list to enable Launch Site selection



- This dashboard shows launches by site, payload, and success rate

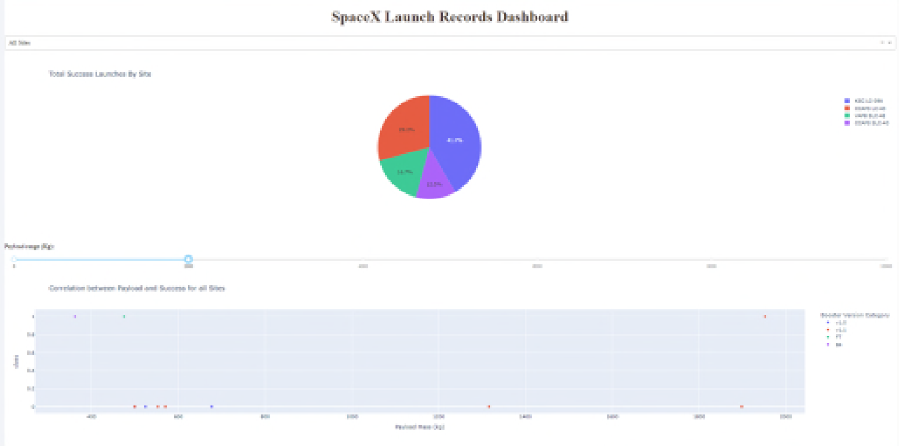
Launch site with highest launch success ratio

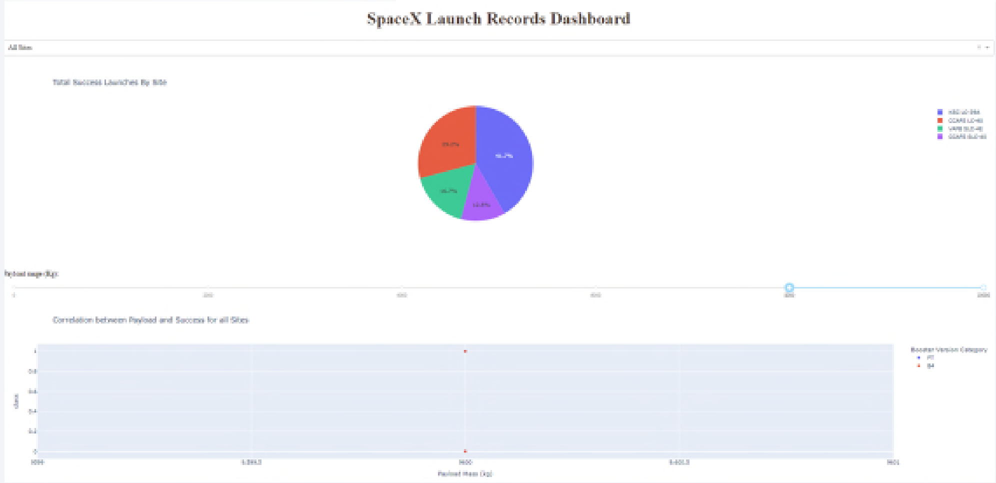


- Highest success rate was site VAFB SLC-4E which had a success rate of 40%

Adjusting Payload Range

- Here we tried a couple different payload ranges to observe the effect on success





Section 5

Predictive Analysis (Classification)

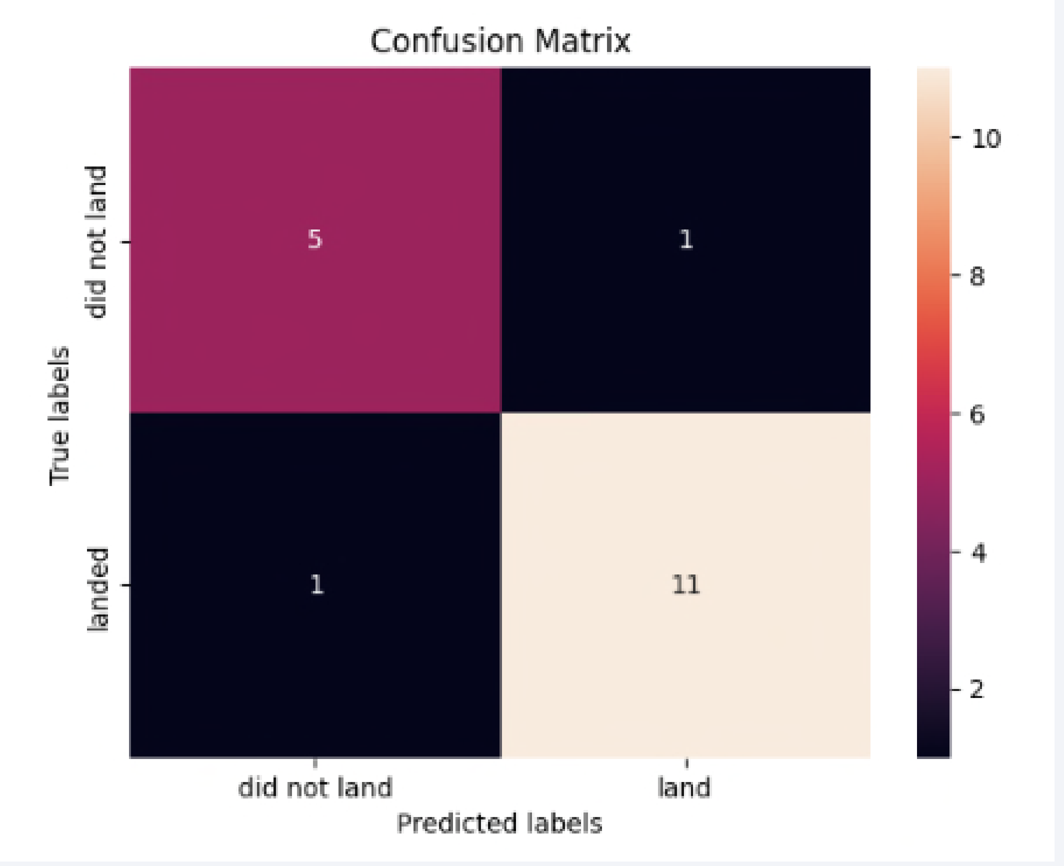
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

- Here we see that the Classification Tree has the highest accuracy but that all models did perform similarly

Confusion Matrix for Classification Tree model

- The classification tree model had only 1 false negative and 1 false positive



Conclusions

- Steady improvements in mission success have been happened over the last decade
- There are different ideal payload ranges for each orbit type
- We are able to use a Classification Tree model to accurately predict the success outcome at 90%

Appendix

- All notebooks and associated data and research are located on GitHub: https://github.com/indyted314/urban-tomato-soup

Thank you!