Winning Space Race with Data Science

Archie Goli
9/3/2023

Outline

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

Introduction

- Commercial spaceflight has become a reality over the past few decades.
- Perhaps the most successful company in the running is **SpaceX**.
- This is because their rocket launches (in the context of Falcon 9) are ***relatively inexpensive*** ($62 million vs. upwards of $165 million).
- The reason SpaceX launches are inexpensive is because they can **reuse** the **first stage** of their rocket for **every launch**.
- In this project, we seek to determine whether the first stage of the rocket can land to estimate the relative cost of a given launch.

Executive Summary

- In this project, we:
  - Extracted data through an open-source SpaceX REST API and used web scraping to gather additional historical data
  - Conducted exploratory data analysis through SQL, Matplotlib, and Pandas
  - Created visualizations using Folium and generated an interactive analytical dashboard of results with Python's Plotly and Dash frameworks.
  - Trained three different machine learning models (Decision Tree Classifier, K-Nearest Neighbors, and SVM) and analyzed which would be the most accurate at predicting whether a given launch would land the first stage successfully.
- Overall, it was determined that approximately **65-70%** of all SpaceX launches typically land their first stages successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - 2 methods: SpaceX REST API + Web scraping through Python
- Perform data wrangling
  - Examined success rate by orbit type, launch site location, and payload mass
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built and evaluated 3 different machine learning models on key performance indicators, including accuracy and confusion

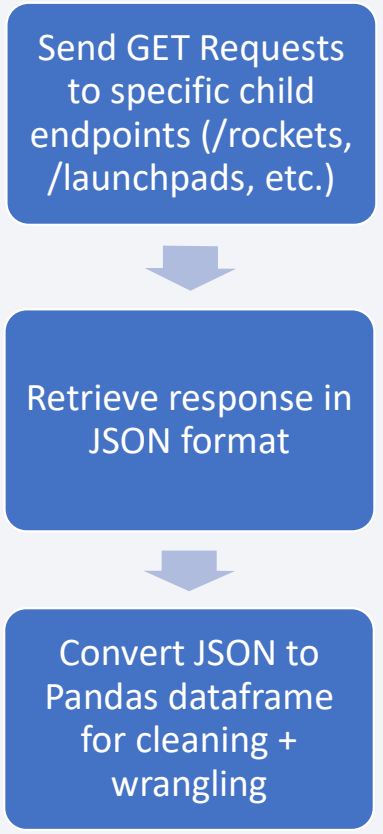
Data Collection

- Data was collected through two sources:
  - SpaceX REST API → open-source API with information on all Falcon 9 launches
    - Used specific endpoints to retrieve data such as landing status, payload mass, type of landing (drone ship, ocean), etc.
  - Web scraping on Wikipedia
    - The Wikipedia site on SpaceX Falcon 9 launches includes more relevant information on the launches conducted to this day.

Data Collection – SpaceX API

- The open-source SpaceX API was accessible through a parent endpoint: https://api.spacexdata.com/v4/

- The notebook with API calls and results is stored on GitHub and can be accessed through this link.



Data Collection - Scraping

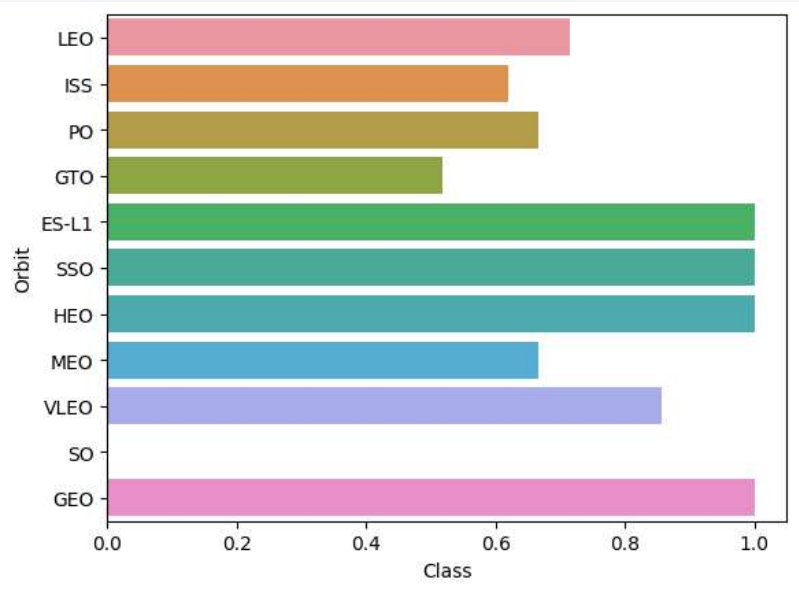
- Webscraping was conducted with the BeautifulSoup library in Python.
- Launch records were retrieved from a Wikipedia Page titled “List of Falcon 9 and Falcon Heavy Launches” (accessible here).
  - Data such as the version of the booster, the status of the landing, the launch site, and payload mass, as well as the dates and times of the launch and landing, were recorded.
- The notebook with webscraping results is stored on GitHub and can be accessed through this link.

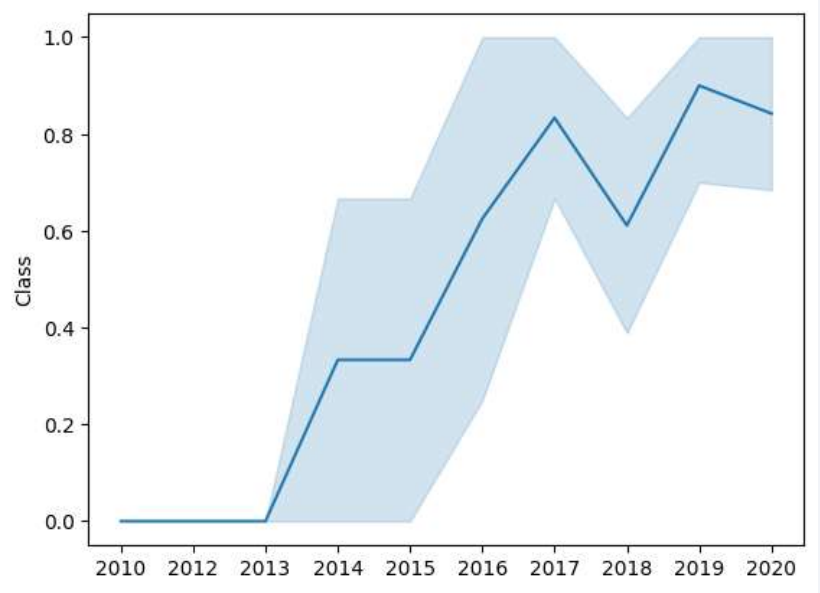
Data Wrangling

- Calculated the number of launches at each launch site, separated by the types of orbits (LEO, VLEO, GTO, SSO, etc.)
  - **GTO (Geosynchronous Orbit)** was the most common type of orbit for rocket launches, followed by **ISS (International Space Station)** and **VLEO (Very Low Earth Orbit)**
- Approximately **67%** of launches from 2010 to 2020 had successful first-stage landings, and **62%** of successful landings were on drone ships
- The notebook with data wrangling results is stored on GitHub and can be accessed through this link.

EDA with Data Visualization

- Exploratory Data Analysis was used to visualize what factors were most important in determining the success of a first-stage landing.
  - We generated visualizations that demonstrated the success rate of each landing by type of orbit, as well as the historical success rate of landings from 2013 to 2020 (seen below).
- The notebook with data exploration results is stored on GitHub and can be accessed through this link.





EDA with SQL

- Additional exploratory analysis was conducted through SQL for more powerful and meaningful results.
- Through this analysis, we determined that the overall mission success is not necessarily correlated to the success of the landing or even launch.
  - Within the dataset (which contained 101 launches from 2010 – 2020), **100 missions** were successful, while only **1** was a mission failure.
- The average payload mass for each launch was found to be **2,534.67** kilograms, and the first successful landing outcome was achieved in 2015.
- The notebook with SQL data exploration results is stored on GitHub and can be accessed through this link.

Build an Interactive Map with Folium

- Folium was used to generate geographical representations of launch sites within the dataset.
- Each launch site was denoted with an **orange dot**, and the clustering of launch sites was usually in 3 locations: Florida, Texas, and California.
  - In addition, launch sites were close to the coast and away from densely populated areas.
- The notebook with Folium visualization results is stored on GitHub and can be accessed through this link.

Build a Dashboard with Plotly Dash

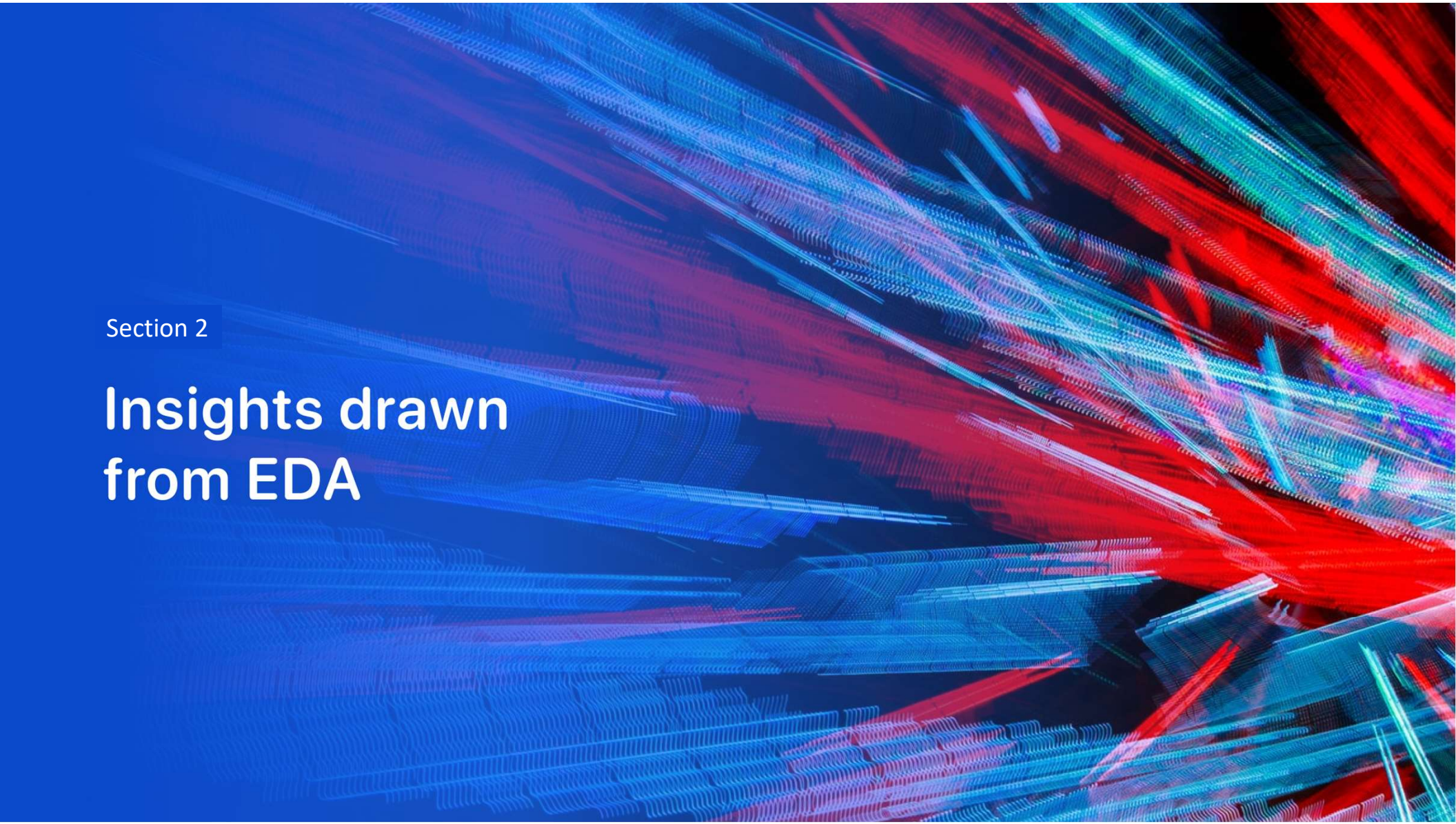
- Through exploratory data analysis (EDA), the most prevalent stratifications of landing statuses were in relation to the launch site and the mass of the payload on the given rocket.
- The interactive Plotly Dash dashboard allowed us to **filter** based on these constraints and more easily demonstrate the success rates across launch sites and payload mass ranges.
- The dashboard code is publicly available on GitHub and can be accessed through this link.

Predictive Analysis (Classification)

- With predictive analysis, we started with the most *common* machine learning models used for classification tasks:
  - K-Nearest Neighbors (KNN)
  - Support Vector Machine (SVM)
  - Logistic Regression
  - Decision Tree Classification
- To evaluate the accuracy of the model after training, we utilized a separate **validation/test set** (with labels) and determined the best accuracy across epochs.
- The notebook with the generated machine-learning models and results is available on GitHub and can be accessed through this link.

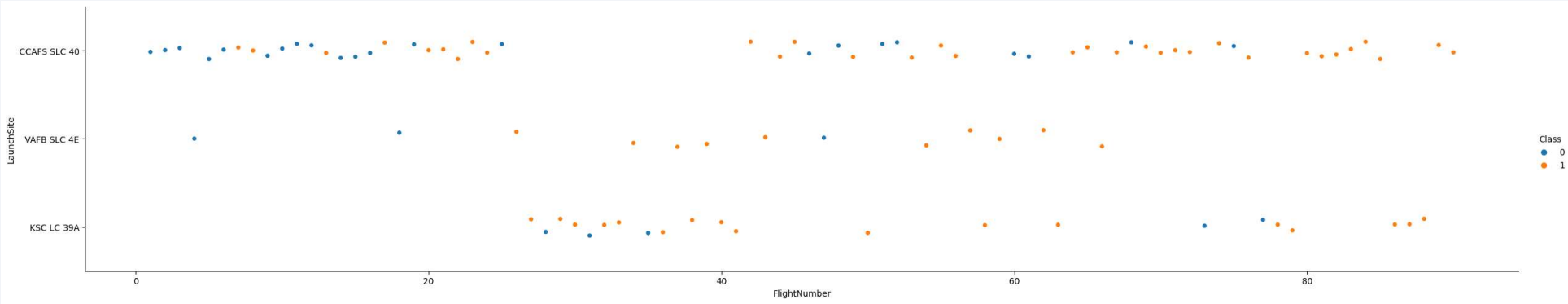
Results

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



Section 2

Insights drawn from EDA

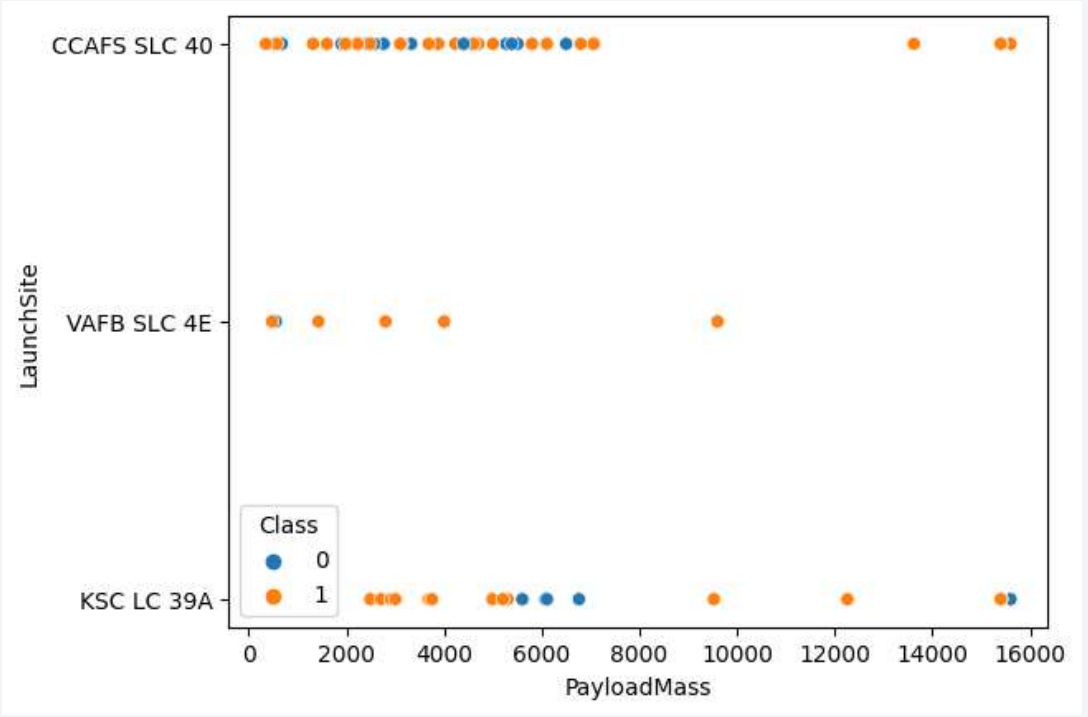


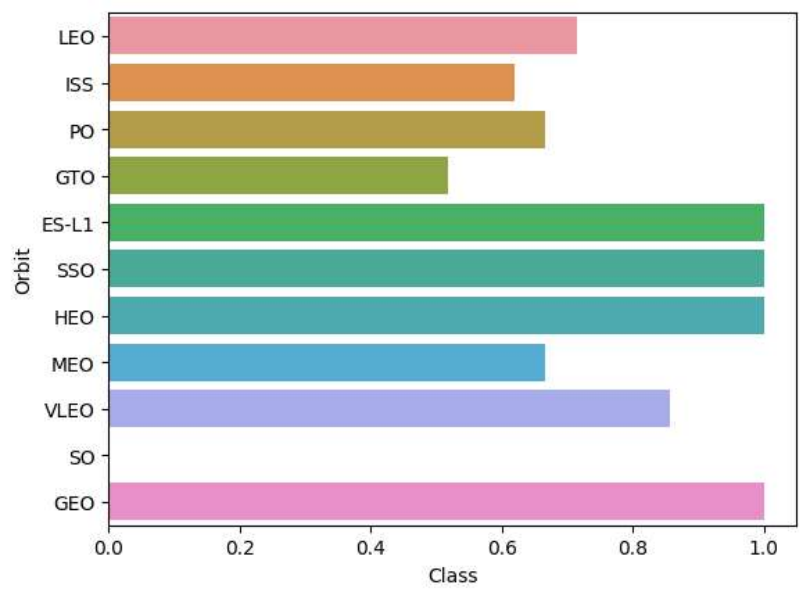
Flight Number vs. Launch Site

- Within the scatterplot, **0** represents failed first-stage landings, and **1** represents successful first-stage landings
- It's clear that the higher the flight number (i.e. the later the launch occurred in time), the more likely the first-stage landing would be successful (which corroborates the improvements across time we saw in our other exploratory data analysis visualizations!)

Payload vs. Launch Site

- The greatest number of launches have occurred at the CCAFS SLC-40 launch site, and at that location, rockets with payloads between 12000 and 16000 kilograms.



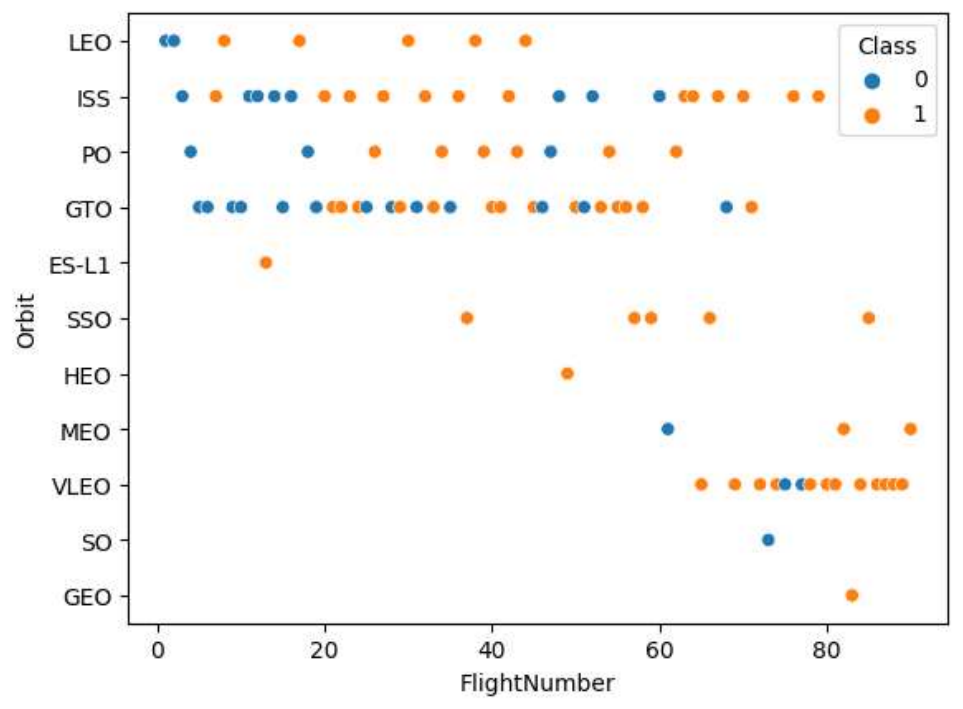


Success Rate vs. Orbit Type

- The **most successful** orbit types include GEO, HEO, SSO, and ES-L1.
- The **least successful** orbit type is GTO.

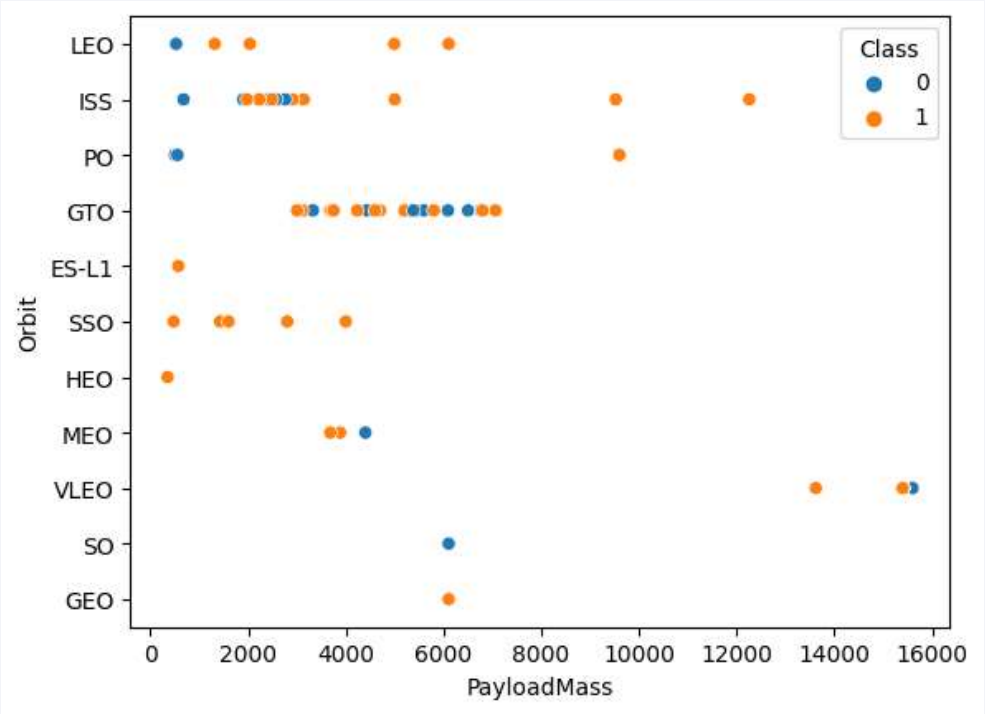
Flight Number vs. Orbit Type

- With this visualization, it's clearer to see why certain orbits have been more or less successful overall.
- SpaceX has seemed to *transition* towards the majority of their later launches and flights being VELO orbits.



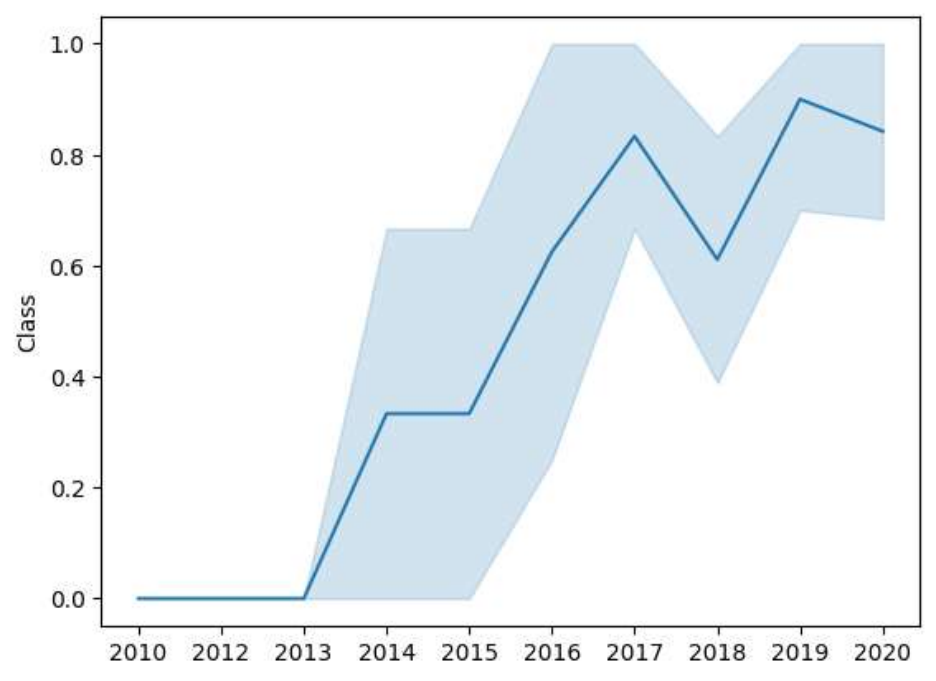
Payload vs. Orbit Type

- However, there's no clear relationship between the payload mass and the type of orbit.



Launch Success Yearly Trend

- Fortunately, the proportion of launch successes has increased consistently from 2013 onwards.
- The y-axis represents a decimal proportion of successful launches (if multiplied by 100, you would retrieve a percentage).



All Launch Site Names

- The distinct launch sites include:
  - CCAFS LC-40
  - VAFG SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- This was most evident through the following SQL query:

```
SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

Launch Site Names Begin with 'CCA'

- Here are 5 launch records whose site names began with “CCA”:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The **total payload mass** carried by all NASA boosters amounted to **99,980 kilograms** across launches from 2010 – 2020.
- In this case, NASA was a customer of SpaceX that utilized certain launches for their own purposes.

Average Payload Mass by F9 v1.1

- The **average payload mass** carried by booster version F9 v1 was **2,534.67 kilograms**.

First Successful Ground Landing Date

- The **first successful landing outcome** on ground pad occurred on **December 22nd, 2015.**

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which have **successfully** landed on a drone ship and had payload mass greater than 4000 but less than 6000 kilograms include:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The **total number** of **successful** missions has been: 100 outcomes.
- The **total number** of **failed** missions has been: 1 outcome.

Boosters Carried Maximum Payload

- The booster versions that have carried the **maximum** payload mass have been:

  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

Here are the **failed** drone ship landing outcomes in the year 2015.

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

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

- In descending order, here are the ranked landing outcomes between June 4th, 2010, and March 20th, 2017:

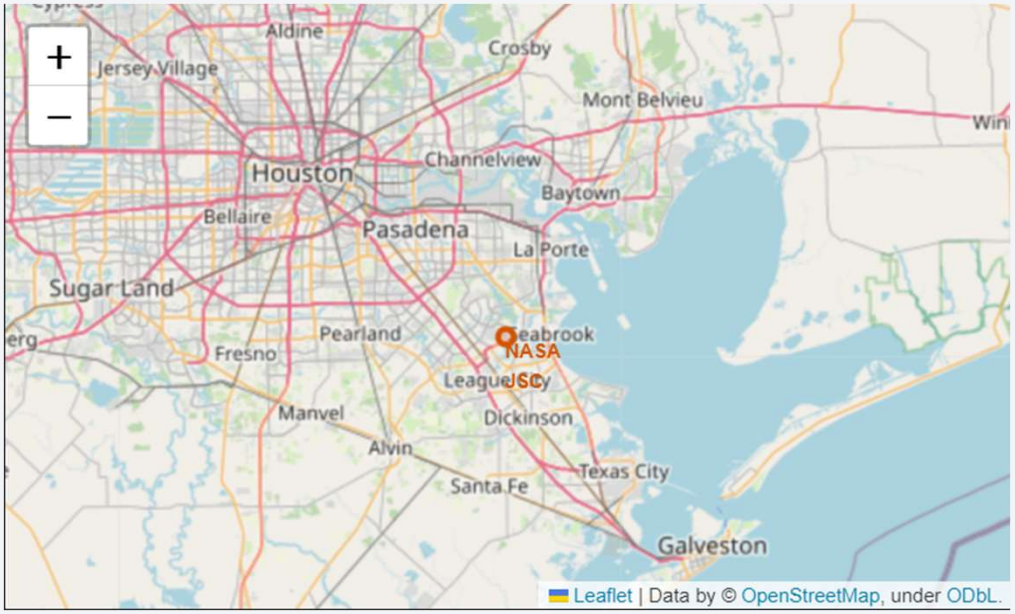
Landing_Outcome	Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

Texas Launch Site(s)

- One Launch Site is the **NASA Johnson Space Center** in Houston, Texas:

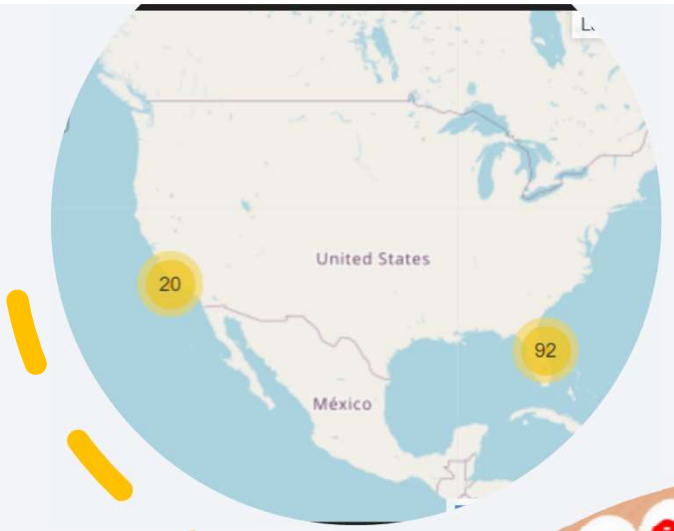


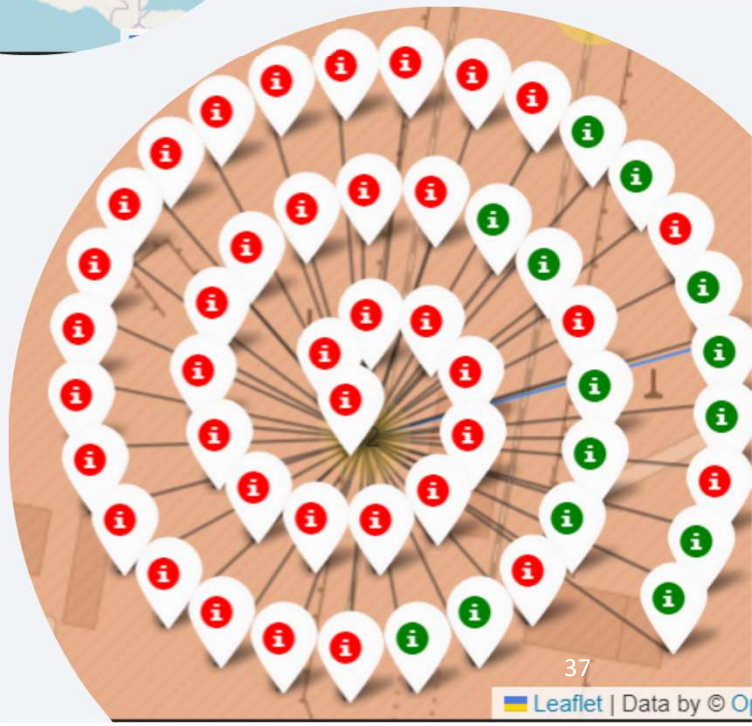
Additional Launch Sites

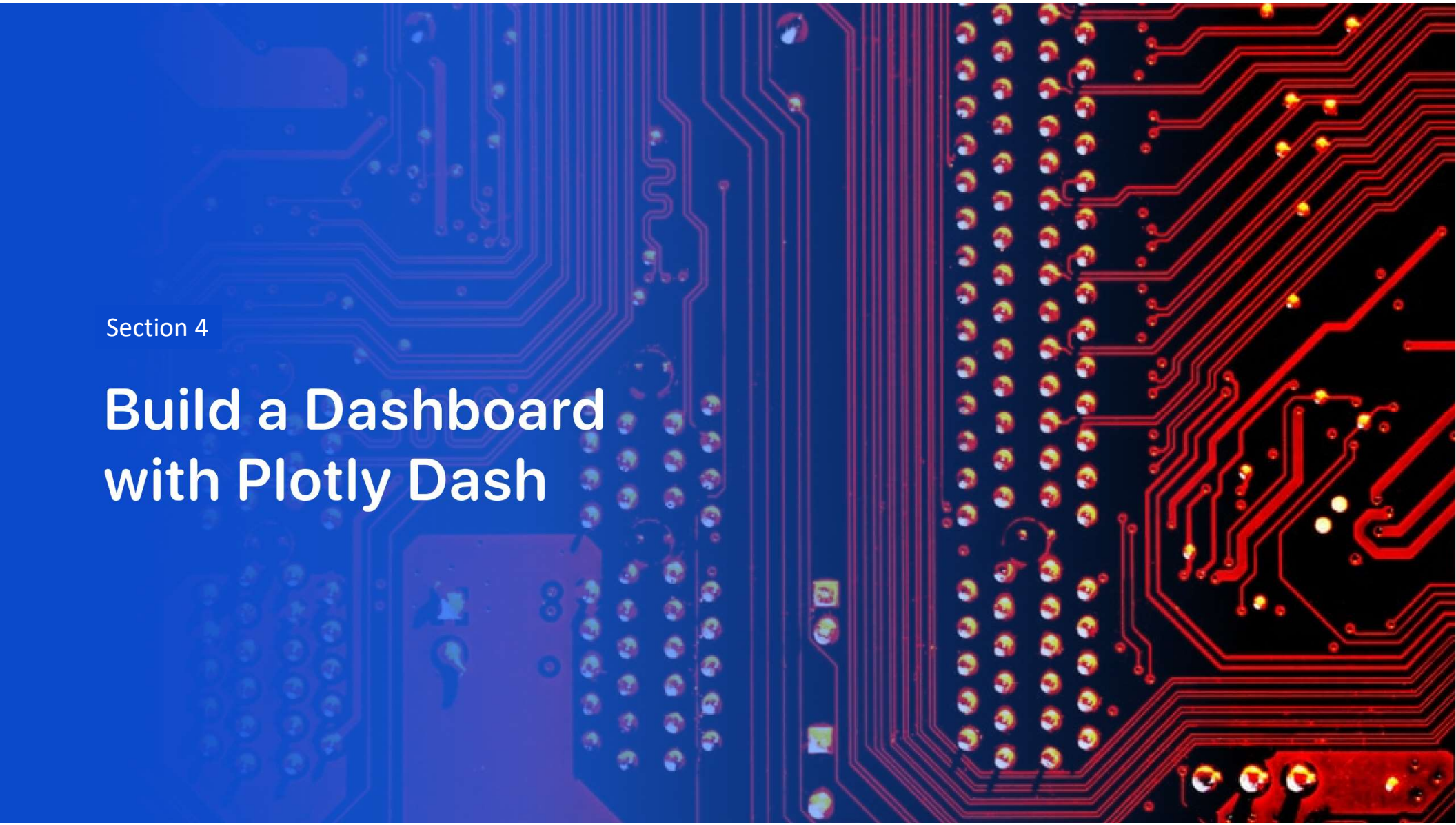
- Additional Launch Sites have been located near the costs of California and Florida:

Launch Successes and Failures

- With Folium, we can visualize the total number and relative *status* of launches across our dataset.
- Most launches have taken place across the coast of Florida.
- In addition, these launches have been further split into several sub-cluster locations. We have visualized the successes and failures of certain launches at the most common launch site in Florida.

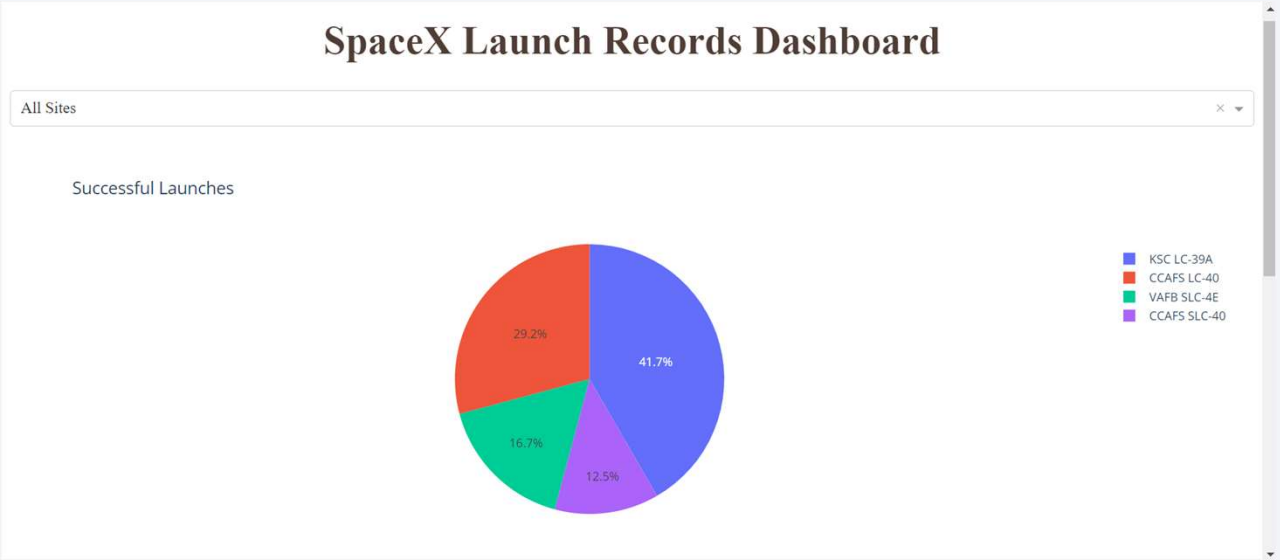






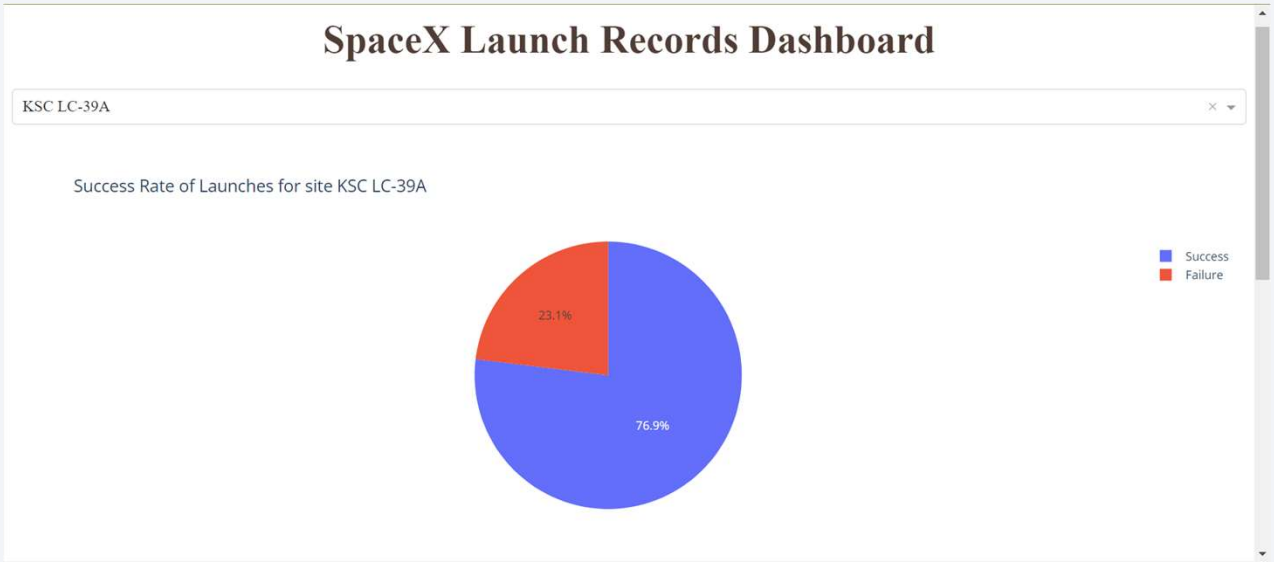
Section 4

Build a Dashboard with Plotly Dash



- A significant portion of successful launches have taken place at the **KSC LC-39A** Launch Site.

Successful Launches, Total

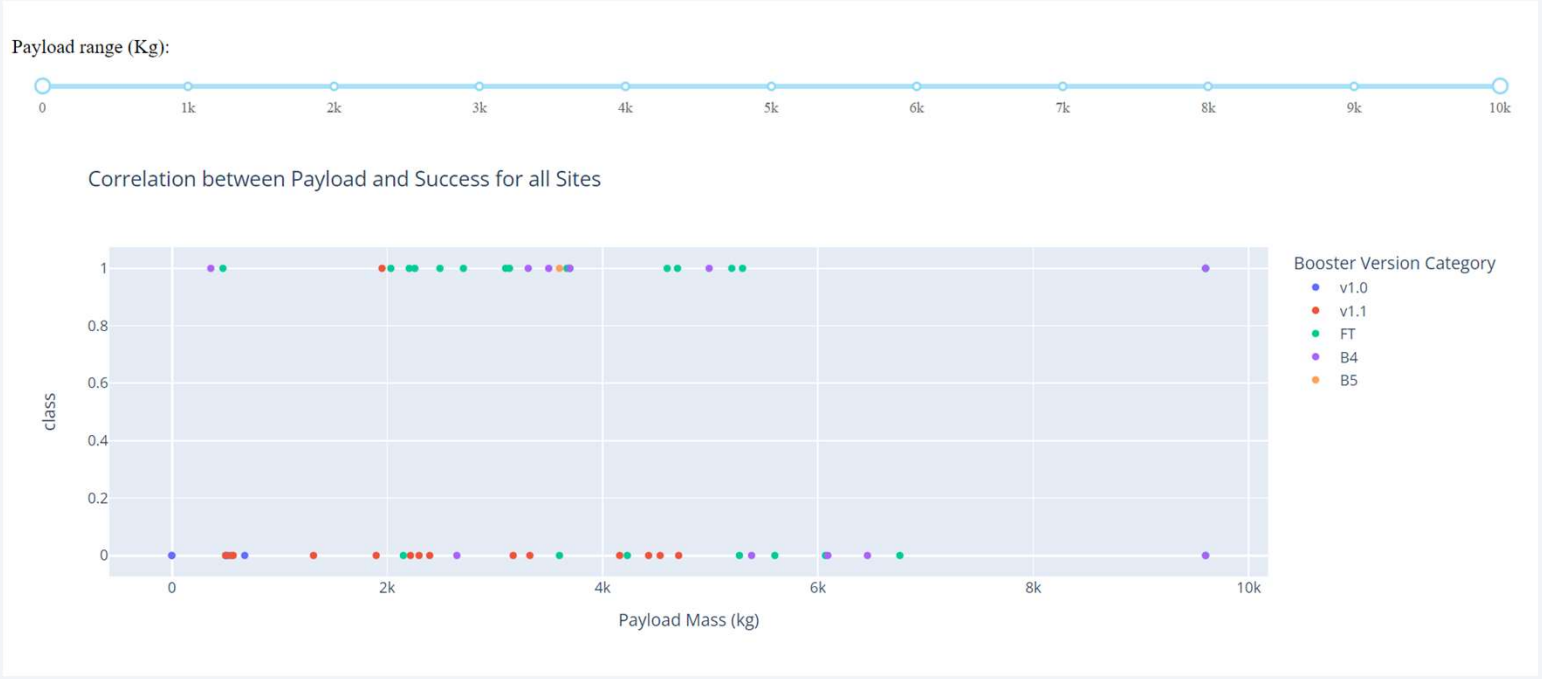


- The **KSC LC-39A** Launch Site also has the greatest *proportion* of successful launches, with over 70% of launches being positive.

Highest Success Rates

Payload vs. Launch Outcome

We have also visualized the relationship between payload mass and launch outcomes. Here is a screenshot of the scatterplot with the entire range of possible payloads for all launch sites within the dataset.

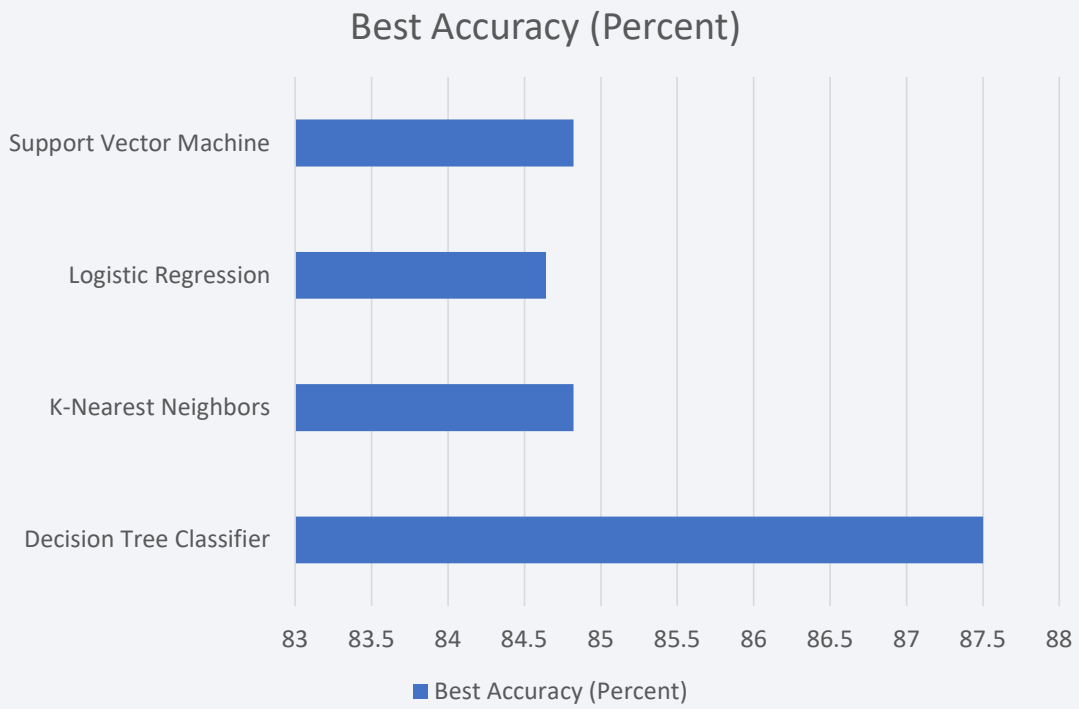


Section 5

Predictive Analysis (Classification)

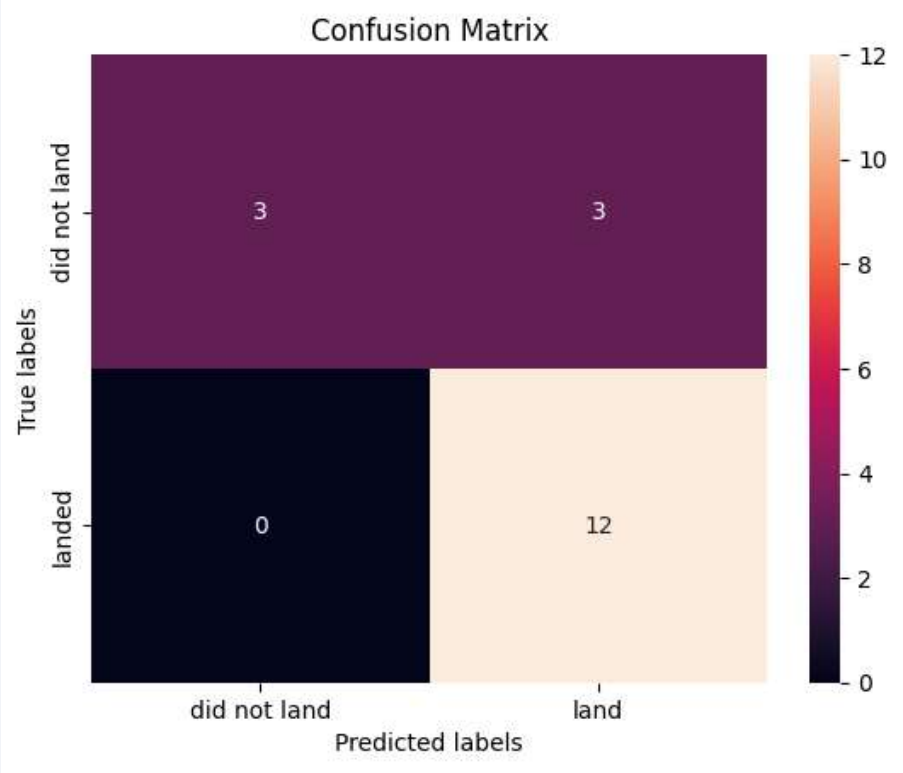
Classification Accuracy

- The bar chart demonstrates the classification accuracy for the four machine learning models utilized for prediction.
- It is clear that the **Decision Tree Classifier** performed best on our dataset.



Confusion Matrix

- While the Decision Tree Classifier performed best, its weakness was **false positives**: predicting that a certain landing was successful when in fact it wasn't.



Conclusions + Appendix

- Determining the importance of factors such as the size of the payload, the location of the launch site, the **date** the launch occurred, etc. can allow us to make better predictions about whether the first stage of a SpaceX rocket launch and landing will be successful or not.
- All the demonstrated notebooks are stored on a master public GitHub repository, which can be accessed through this link.

Thank you!