Outline

- Executive Summary

- Introduction

- Methodology

- Results

- Conclusion

- Appendix

Executive Summary

- The following methodologies were used to pull and analyze the data

  - Web scraping using APIs for Data Collection

  - Data Wrangling, Exploratory Data Analysis, Data Visualization, and Interactive Dashboards

  - Machine Learning

- Summary of all results

  - Since the data was publicly sourced, the Exploratory Data Analysis helped with analyzing launches and payloads

  - Machine Learning helped with predicting which variables are most important for the launches

Introduction

- The objective of the project is to evaluate a new space company, SpaceY to SpaceX to see how well it would compete with SpaceX

- Questions we want answered

  - What is the price of each launch

  - What is most predictive for successful launches

  - Where is the best place for launches

Section 1

Methodology

Methodology

- Data collection methodology:

    - Acquired data via SpaceX API

    - Web scraping from Wikipedia

- Perform data wrangling

    - Filtered data, Inputted Missing Values, One Hot Encoding

- Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models

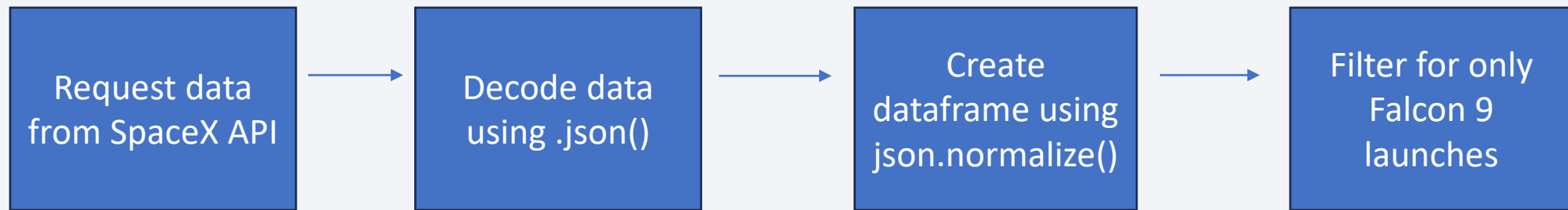   - How to build, tune, evaluate classification models

Data Collection

- The datasets were collected via SpaceX Rest API and data from SpaceX's Wikipedia page

- SpaceX Rest API Columns:

  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- Wikipedia Webscraping Columns:
  - Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch Outcome, Version Booster, Booster landing, Date, Time
  - https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20Data%20Collection.ipynb
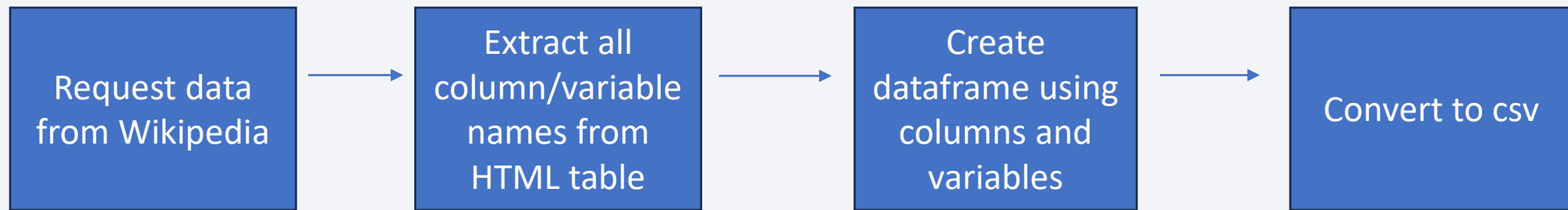
Data Collection – SpaceX API

```
┌─────────────────┐     ┌─────────────────┐     ┌─────────────────┐     ┌─────────────────┐
│  Request data   │ ──▶ │  Decode data    │ ──▶ │     Create      │ ──▶ │ Filter for only │
│ from SpaceX API │     │  using .json()  │     │ dataframe using │     │    Falcon 9     │
│                 │     │                 │     │ json.normalize()│     │    launches     │
└─────────────────┘     └─────────────────┘     └─────────────────┘     └─────────────────┘
```

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20Data%20Collection.ipynb

Data Collection - Scraping

```
┌─────────────┐      ┌─────────────┐      ┌─────────────┐      ┌─────────────┐
│ Request data │─────▶│ Extract all  │─────▶│   Create    │─────▶│Convert to csv│
│from Wikipedia│      │column/variable│     │dataframe using│     │             │
│             │      │ names from   │      │ columns and │      │             │
│             │      │ HTML table   │      │  variables  │      │             │
└─────────────┘      └─────────────┘      └─────────────┘      └─────────────┘
```
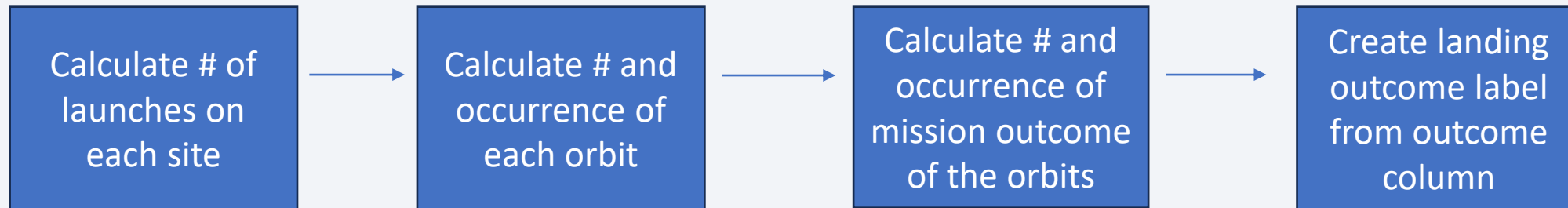
https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20Webscraping.ipynb

Data Wrangling

- Initially did EDA to get value counts for different columns

- Then created a variable for the good and bad outcomes using dictionary keys and mapped to new column named class which classified the good and bad outcomes of landings

- https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20Data%20Wrangling.ipynb

Calculate # of launches on each site	→	Calculate # and occurrence of each orbit	→	Calculate # and occurrence of mission outcome of the orbits	→	Create landing outcome label from outcome column

EDA with Data Visualization

- When visualizing data, scatterplots were used to show relationships in data

- Some charts were:

  - Flight Number and Launch Site

  - Payload and Launch Site

  - Flight Number and Orbit Type

  - Payload and Orbit Type

- https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20EDA%20with%20Visualization.ipynb

EDA with SQL

- SQL Queries:

  - 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 succesful 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

Build an Interactive Map with Folium

- Markers, Circles, and lines were added to help with identifying different points on a map

  - Markers were indicative of launch sites

  - Marker clusters were used to group events together in a specific coordinate, for example, launches

  - Circles were used to highlight different areas around specific coordinates

  - Lines are used to specify the distance between two coordinates

  - https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard I added multiple graphs and charts that were used to visualize both payload and launch success

- I added a Pie Chart to show the total successful launches for all sites, and included a dropdown for specific launch sites

- With the dropdown for specific launch sites, choosing a specific site will show a pie chart showing successful and failed launches from that specific site

- https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20Plotly%20Dash.ipynb

Predictive Analysis (Classification)

Converted target to numpy array → Split into test/train datasets → Tuned hyperparameters for different ML algorithms → Compared ML algorithms with F1 and Jaccard score

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20Machine%20Learning.ipynb
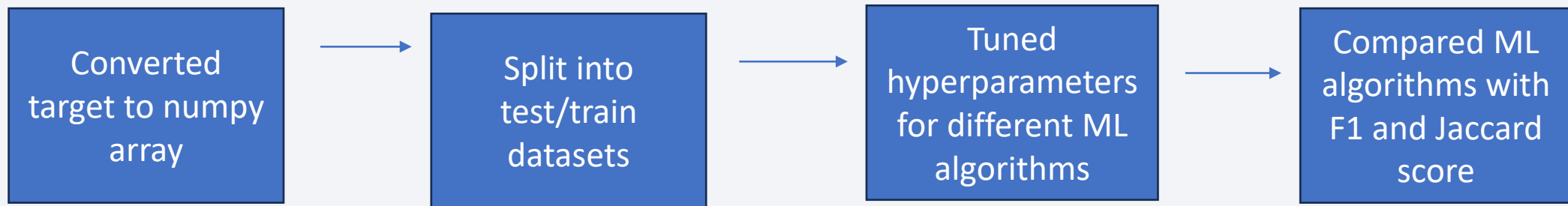
Results

- Exploratory data analysis results

- Interactive analytics demo in screenshots

- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- As the flight numbers go up, more class 1 flights are going to CCAFS SLC 40
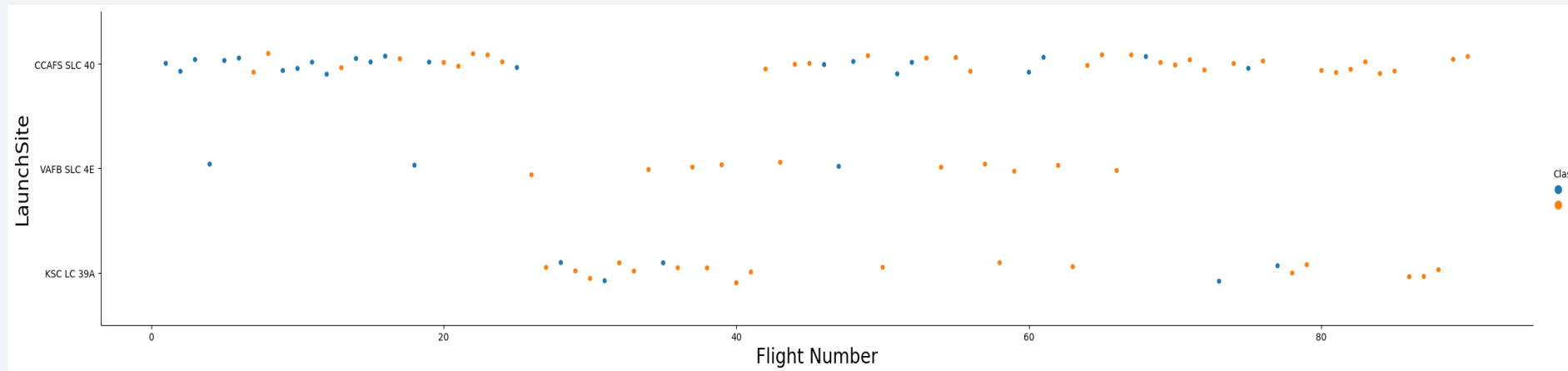
Payload vs. Launch Site

- As the payload mass increases, there are more class 1 flights and less class 0 flights
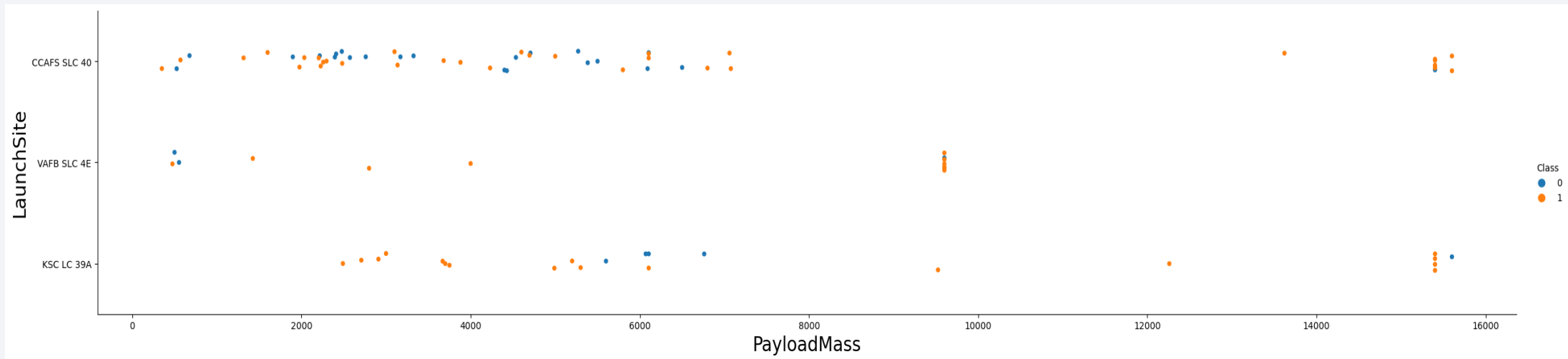
Success Rate vs. Orbit Type

- When looking at the success rate compared to orbit type, there are 4 orbit types that have a 100% success rate

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20EDA%20with%20Visualization.ipynb
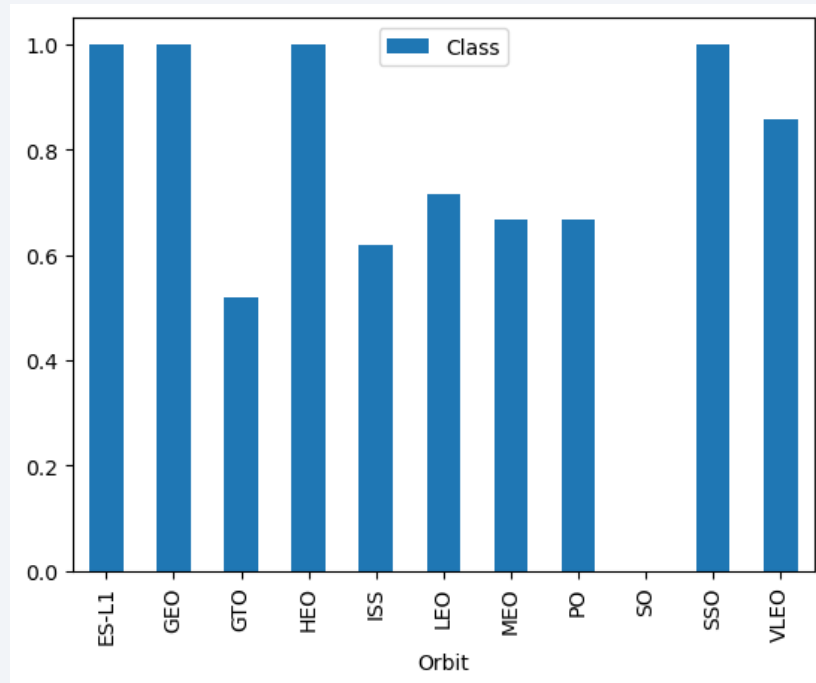
Flight Number vs. Orbit Type

- As the flight number increases the more flights are going to VLEO orbits
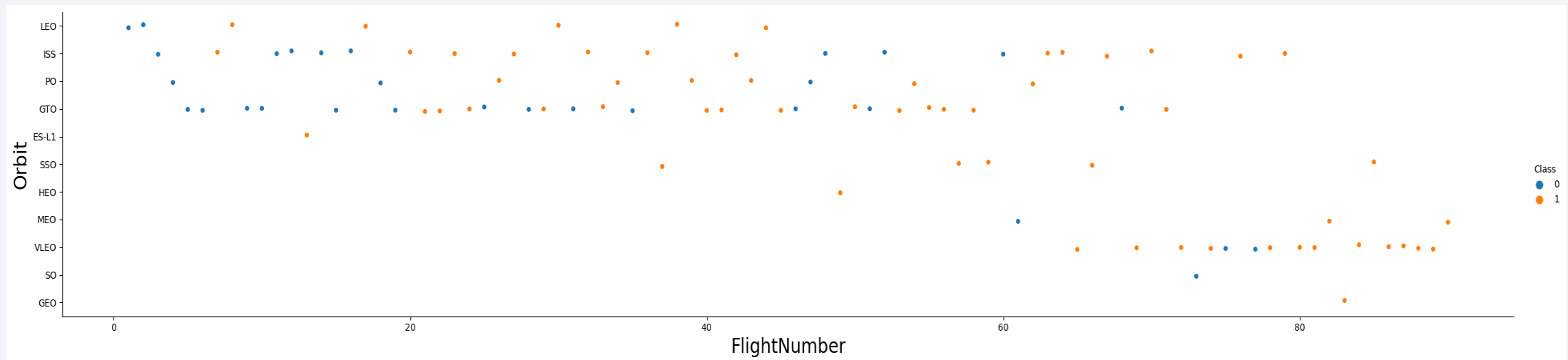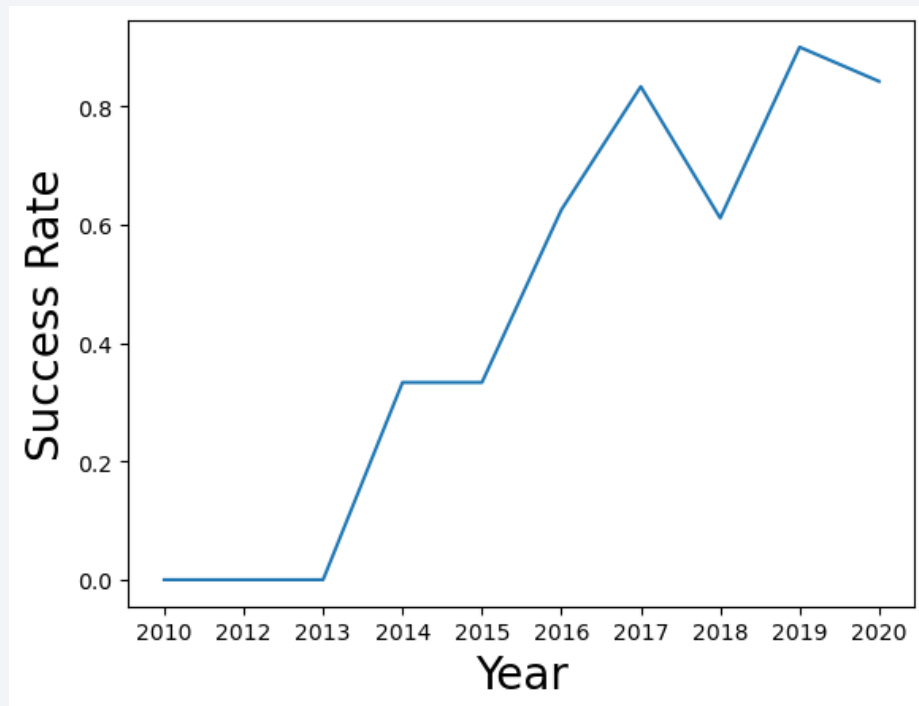
Payload vs. Orbit Type

- When comparing the payload and orbit type, the lesser payloads are using the GTO and ISS orbit types
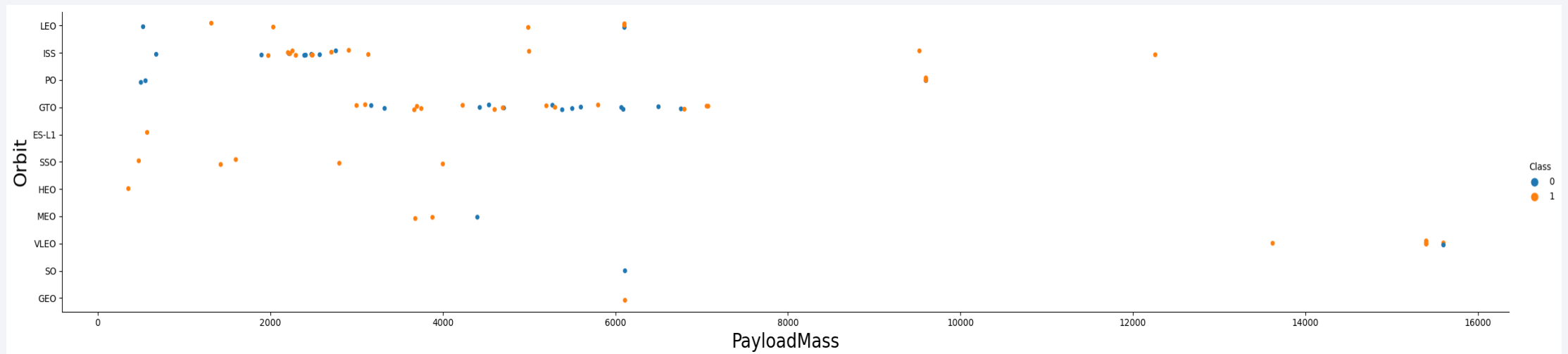
Launch Success Yearly Trend

- When looking at the ling graph, the success rate increases every year

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20EDA%20with%20Visualization.ipynb

All Launch Site Names

- According to the query, there were four unique launch sites, which was found using unique values of "launch_site"

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

Launch Site Names Begin with 'CCA'

- Here are 5 results with the Launch Site starting 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

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20SQL%20EDA.ipynb

Total Payload Mass

- Getting the sum of the payload mass where the payload is like "%CRS%"

total_payload
111268

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20SQL%20EDA.ipynb

Average Payload Mass by F9 v1.1

- Getting the average of the payload but filtering only for booster version F9 v1.1

avg_payload
2928

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20SQL%20EDA.ipynb

First Successful Ground Landing Date

- In order to find the first successful ground landing date, I had to set the minimum date and where the landing outcome was like "success"

first_success_gp

2015-12-22

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20SQL%20EDA.ipynb

Successful Drone Ship Landing with Payload between 4000 and 6000

- When getting the result, I had to do multiple filters where the payload is between 4000 and 6000 and another filter where the landing outcome was like 'Success"

booster_version
F9 B5 B1046.2
F9 B5 B1046.3
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5 B1058.2
F9 B5B1060.1
F9 B5B1062.1

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20SQL%20EDA.ipynb

Total Number of Successful and Failure Mission Outcomes

- When making the query, I had to make a count variable for the mission outcomes, then do a group by for the outcomes

mission_outcome	qty
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20SQL%20EDA.ipynb

Boosters Carried Maximum Payload

- When making the query, I had to set the payload mass equal to a subquery that got the max payload mass

- https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20SQL%20EDA.ipynb

boosterversion
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

- When getting the results, I had to filter the landing outcome to be like "Failure" and the YEAR(DATE) = 2015

booster_version	launch_site	landing_outcome	DATE
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2015-10-01
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	2015-04-14

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20SQL%20EDA.ipynb

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

- When making the query, I made a filter for the date range, as well as making a new variable called "qty" which I used as a groupby for the landing outcomes, then made them in descending order

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

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20SQL%20EDA.ipynb

Section 3

Launch Sites
Proximities Analysis

All Launch Site Locations

- When looking at the launch sites, both are in areas that are warm and have nice weather
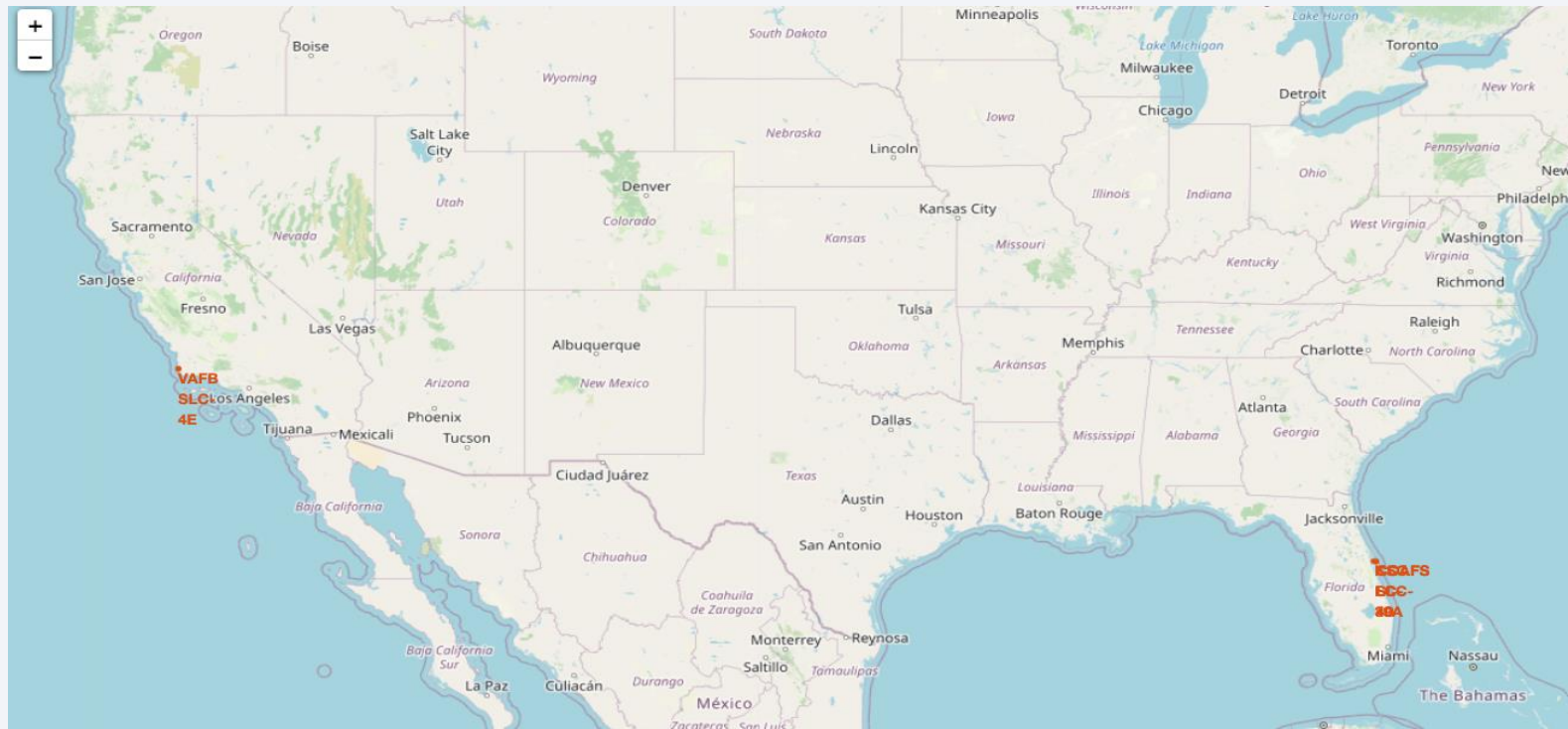
Outcomes By Launch Site

- The launch outcomes for this site show that they are mainly failures and few success rates
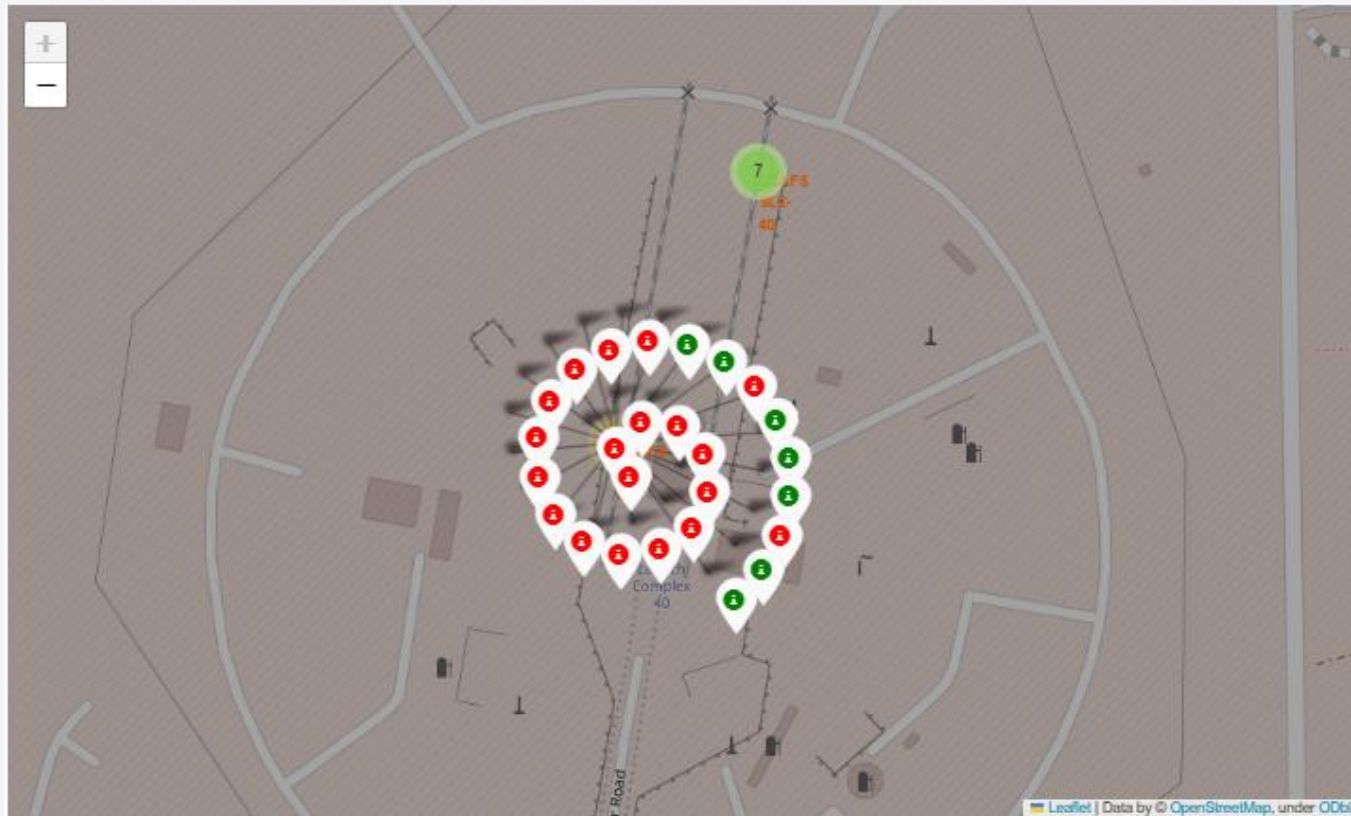
Launch Sites Proximities

- When looking at this specific launch site, it is 5.52 KM away from NASA Parkway East and 0.52 KM from NASA Parkway North
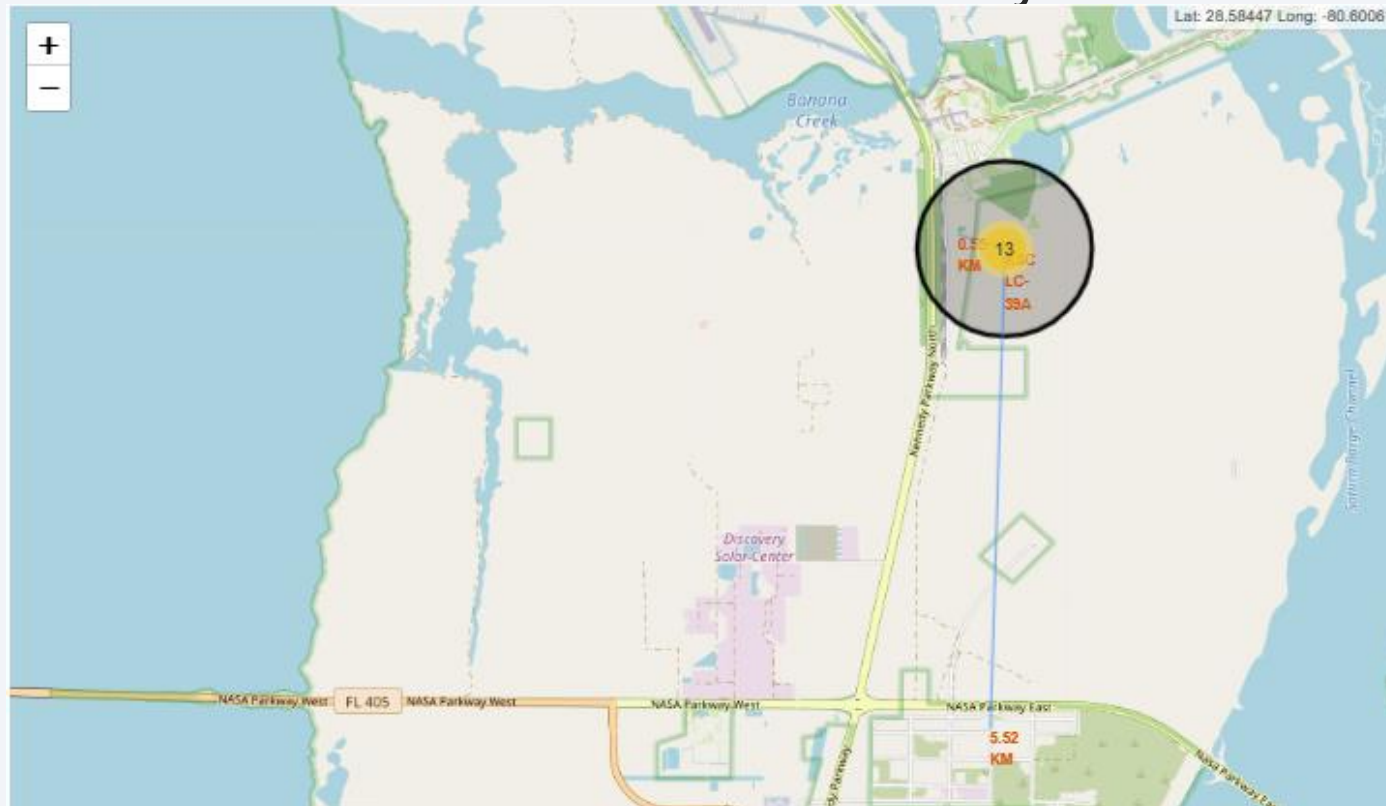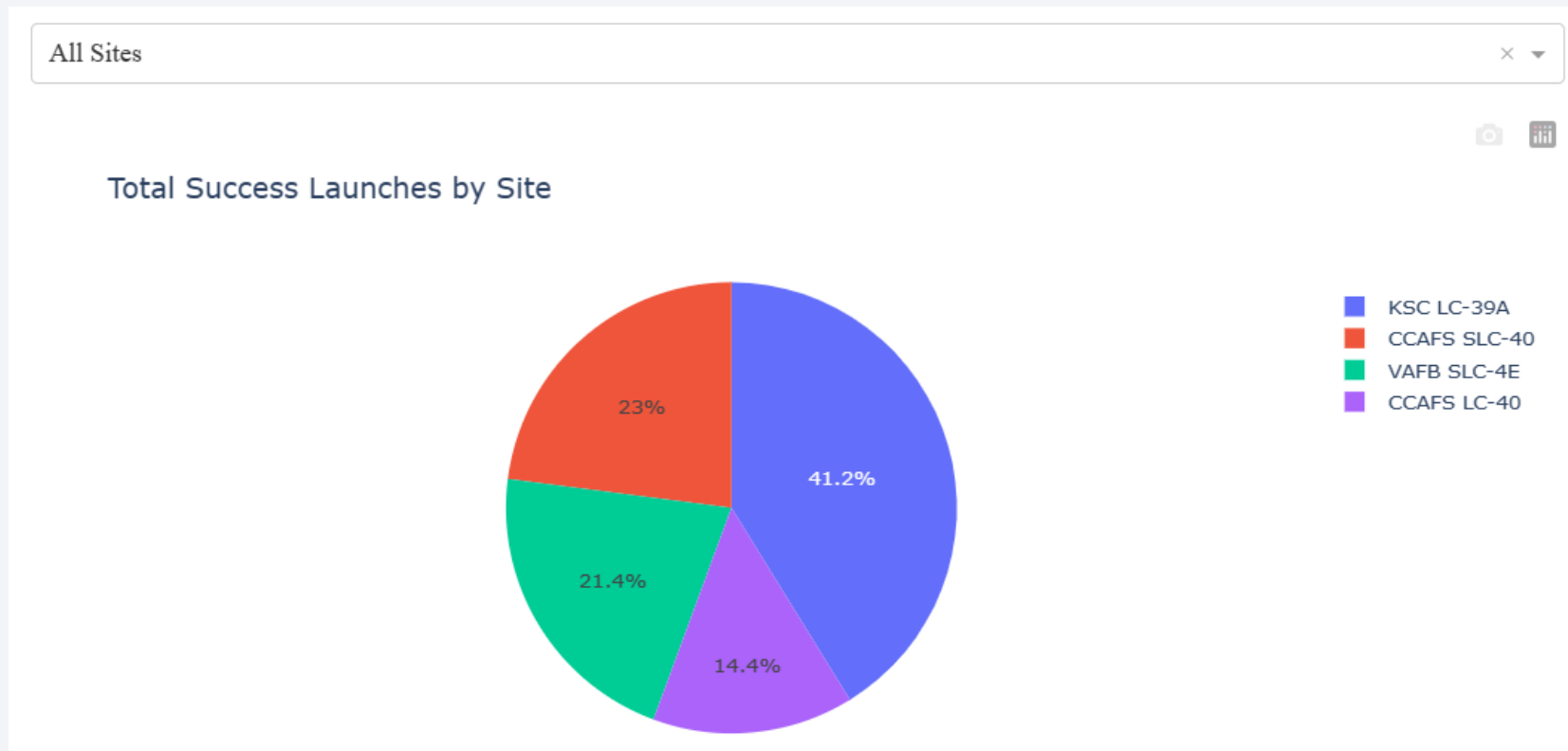
Build a Dashboard
with Plotly Dash

Pie Chart All Sites

- When looking at the pie chart, KSC LC-39A had the highest proportion of successful launches with 41.2%

Pie Chart for KSC LC-39A

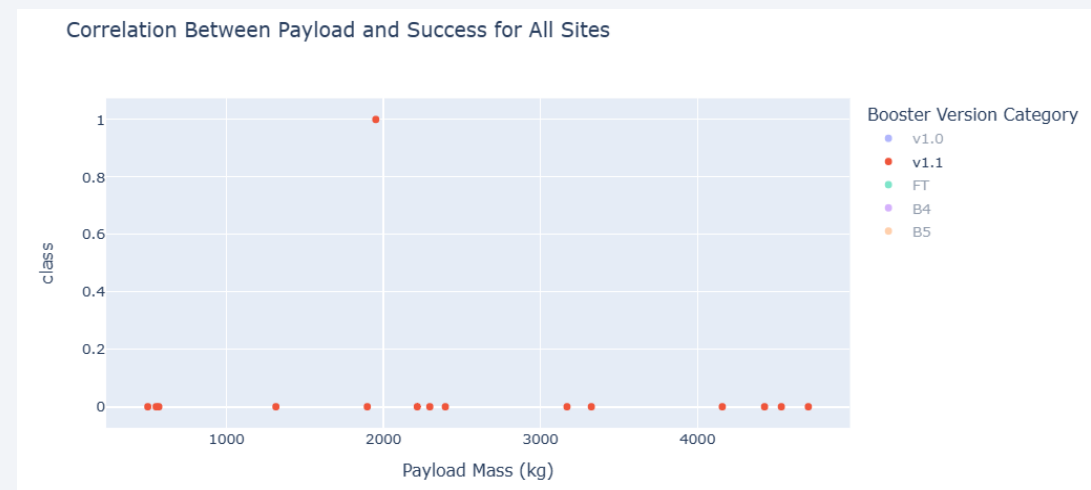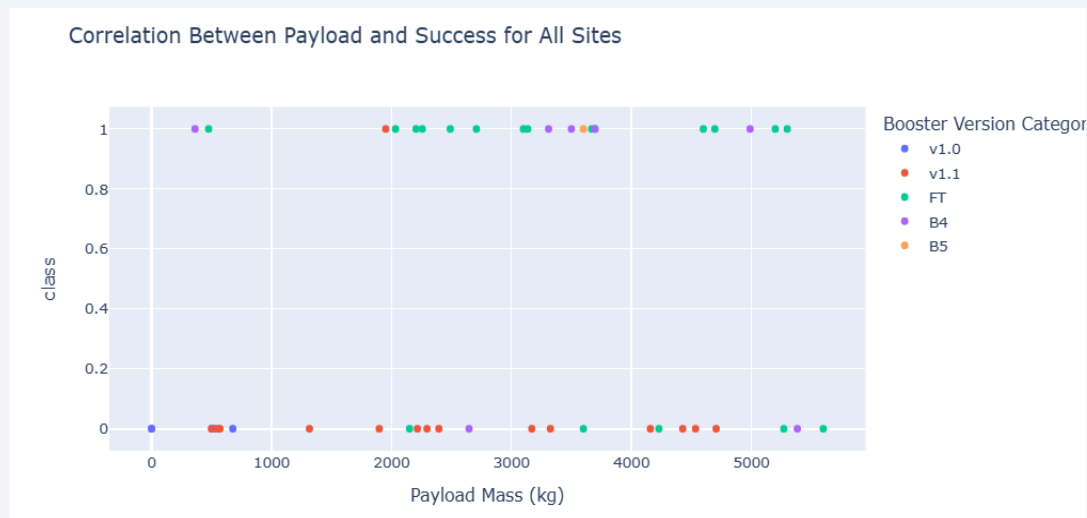- 76.9% of the launches were successful at KSC LC-39C

Total Success Launches for Site KSC LC-39A
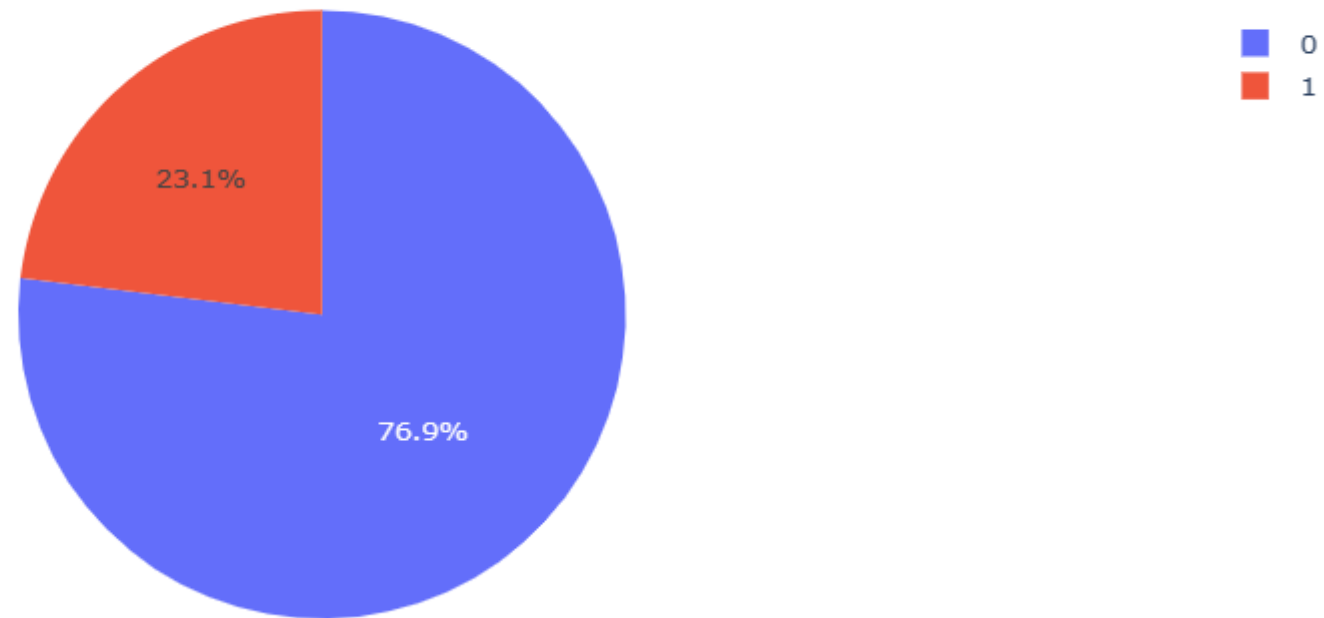
Payload Scatter Chart

- Payloads under 6,000 kg have the highest correlation with successful launches and booster v1.1 has the most successful launches under 6,000 kg

https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20Plotly%20Dash.ipynb

Section 5

Predictive Analysis (Classification)
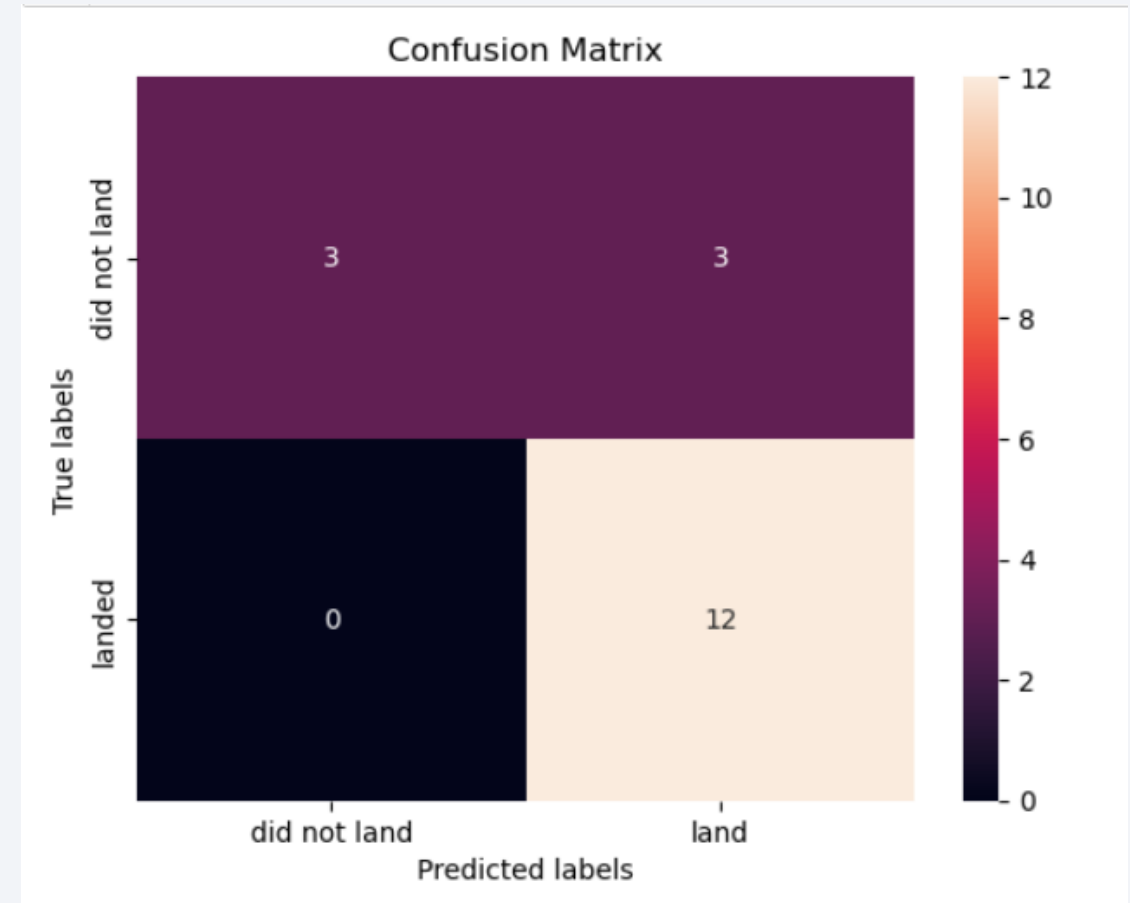
Classification Accuracy

- When looking at the tables, the tree had the highest Jaccard Score, F1 Score, and Accuracy in both tables

- https://github.com/asammarco1999/IBM -Capstone- Project/blob/main/IBM%20Machine%20 Learning.ipynb

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.857143	0.800000
F1_Score	0.888889	0.888889	0.923077	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.840580	0.819444
F1_Score	0.909091	0.916031	0.913386	0.900763
Accuracy	0.866667	0.877778	0.877778	0.855556

Confusion Matrix

- When looking at the confusion matrix, the accuracy of the model did well, with there being 15/18 predicted values that were actual values, going along with the 83.33% accuracy of the models.

- https://github.com/asammarco1999/IBM-Capstone-Project/blob/main/IBM%20Machine%20Learning.ipynb

Conclusions

- Payloads under 6,000 kg have the highest correlation with successful launches and booster v1.1 has the most successful launches under 6,000 kg

- ES-L1, GEO, HEO, and SSO all had 100% success rates

- The tree classifier was the most predictive out of all the ML algorithms

- KSC LC-39A had the most successful launches and a success rate of 76.9%

Appendix

- I would like to thank Coursera and IBM for making the course, I have learned a lot throughout my time doing the course

Thank you!