IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matthew R Vortex
05/31/2025

Outline

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

Executive Summary

- Methodologies used
  - Data Collection: SpaceX REST API, webscraping Wikipedia
  - Exploratory Data Analysis: data wrangling (pandas), and data visualization (matplotlib, seaborn) including interactive dashboards (plotly, dash) and geospatial plotting (folium)
  - Machine Learning: architectures including Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbors were evaluated with multiple hyperparameters through GridCVSearch.
- Summary of all results
  - Data Collection: Successfully read data from multiple sources into dataframes
  - Exploratory Data Analysis: Identified relationships and correlations among variables
  - Machine Learning: Generated an 88% accurate model using Support Vector Machine

Introduction

- The consumer space age is upon us. In order for Space Y to be competitive in our field, we must understand our costs versus those of our competitors.
- Space X advertises a Falcon 9 Launch Cost of $62MM vs up to $165MM for competitors.
- Understanding their ability to reuse their first stage is critical to our strategy.
- This project seeks to develop a Machine Learning Model suitable for predicting successful first stage recovery of Space X launches.
- This information is a key pillar of our strategy to bring Space Y to the front of space exploration through fiscal management.

Section 1

Methodology

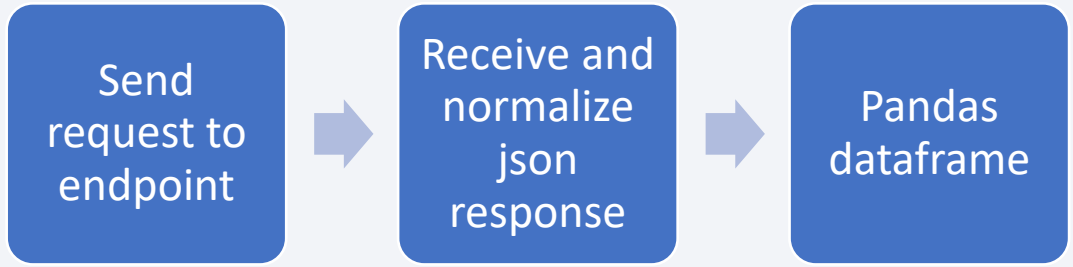
Methodology

Executive Summary

- Data collection methodology:
  - Collect data through SpaceX API calls and webscraping Wikpedia.
- Perform data wrangling
  - Encode landing results and landing method into a binary success variable
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Utilize Logistic Regression, Support Vector Machine (SVM), Decision Trees, and K-Nearest Neighbor (KNN) algorithms in conjunction with GridSearchCV to find optimal model and hyperparameters.

Data Collection

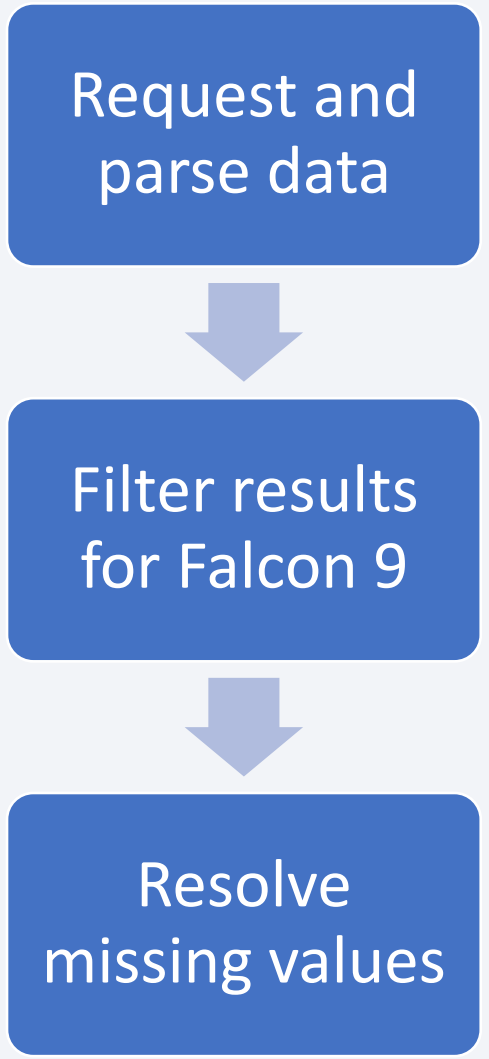
- Space X REST API calls to collect past launch data. Target data includes reuse and recovery data.



- Webscraping Wikipedia tables (with requests, BeautifulSoup) into dataframes. Targets include Payload Mass, Orbit, Booster Version and Outcome.

Data Collection – SpaceX API

- Data was retrieved with API call. Request was normalized, parsed, and filtered for Falcon 9 data. Then, missing payloads were replaced with the mean.
- https://github.com/Mdawg27265/DataCapstone/blob/c6c8151552212f1425b0529ad6ff1508cdd9dd11/jupyter-labs-spacex-data-collection-api%20(1).ipynb

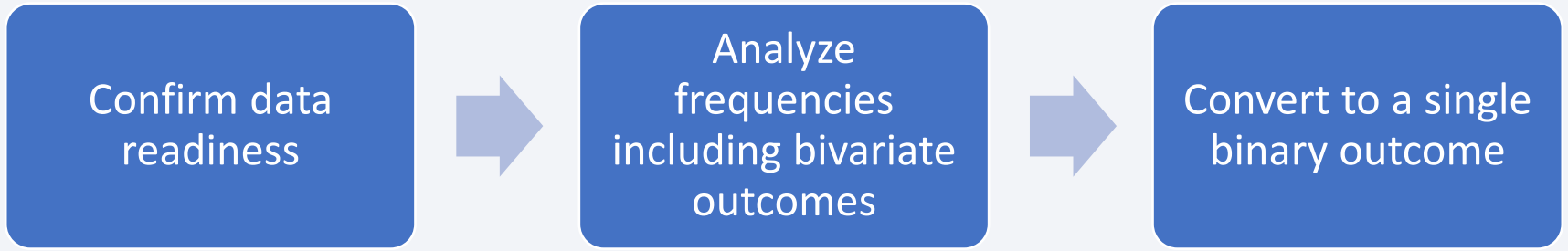


Data Collection - Scraping

- Data was retrieved by sending GET request to Wikipedia page with tables. Target table was parsed into dataframe.
- https://github.com/Mdawg27265/DataCapstone/blob/c6c815155221 2f1425b0529ad6ff1508cdd9dd11/jupyter-labs-webscraping%20(1).ipynb

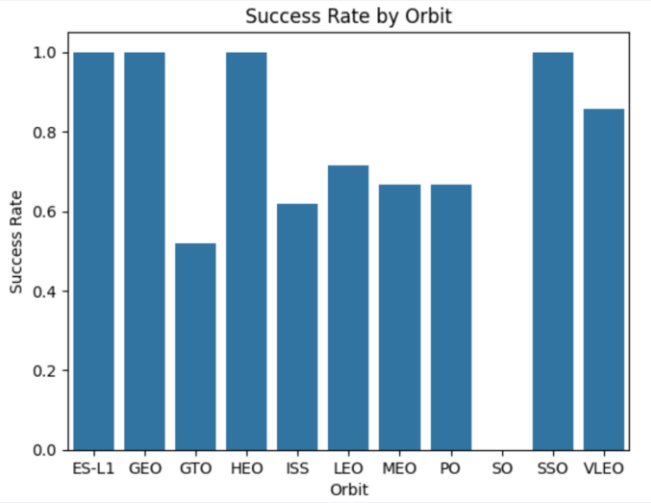
Data Wrangling

- Confirmed missing data and data types, calculated number of launches by site, calculated the frequency of orbit types, calculate the frequency of different outcomes by landing type and transform into a binary outcome variable.
- https://github.com/Mdawg27265/DataCapstone/blob/149c2dd92d401421becb499bb82a72df053c7cff/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb



EDA with Data Visualization

- Visual Analysis
  - Catplots and barplots were used to explore relationships between variables.
  - Variables considered include Outcome with respect to Flight Number, Payload Mass, Launch site, Orbit
- Feature Engineering by one hot encoding dummies for categorical variables.
- https://github.com/Mdawg27265/DataCapstone/blob/48e75bb1b02beb82b6203cc782bf9e09b05401d8/edadataviz%20(1).ipynb



EDA with SQL

- SQL Queries were developed to gain the following insight:
  - names of the unique launch sites
  - records where launch sites begin with the string 'CCA'
  - total payload mass carried by boosters launched by NASA
  - average payload mass carried by booster version F9 v1.1
  - date of the first successful landing outcome in ground pad
  - names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - total number of successful and failure mission outcomes
- https://github.com/Mdawg27265/DataCapstone/blob/48e75bb1b02beb82b6203cc782bf9e09b05401d8/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

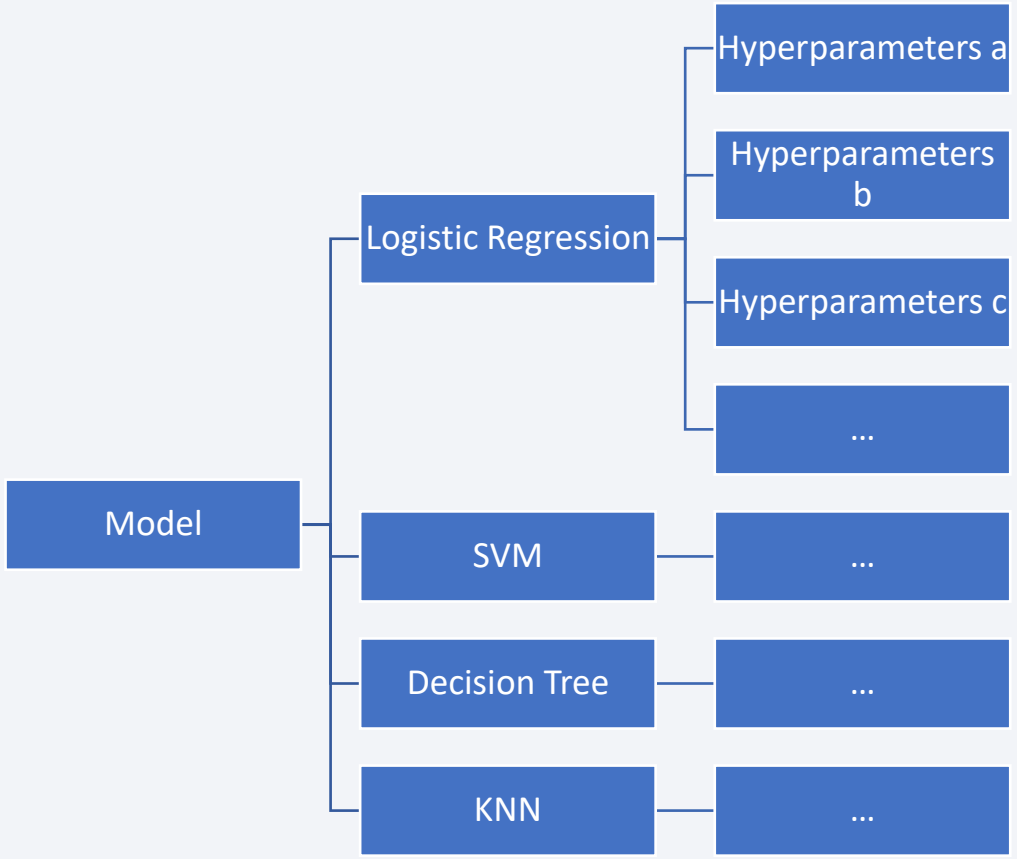
- A map was developed with Folium. The map was centered on Johnson Space Center in Texas.
- In Folium, objects are added to convey information
  - Circles: added to the different launch sites to visualize where launches occur from.
  - Circle popups display additional info, the site name
  - Markers: added to the different launch sites to visualize launch outcomes. Contains label.
  - Marker Clusters: associate multiple Markers with a specific location.
  - Lines: display distances to points of interest
- https://github.com/Mdawg27265/DataCapstone/blob/ccbe151a85b5730e1a43115a813e59c17a5adc88/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- A dashboard was developed for a clean interactive view of the data.
  - Percentage of successful launches by site
  - Payload Mass
- This dashboard helps to review the outcome with respect to the relationship between launch site and payload mass.
- https://github.com/Mdawg27265/DataCapstone/blob/190b265e200659db003787830aa174a3ce8acb0a/spacex-dash-app.py

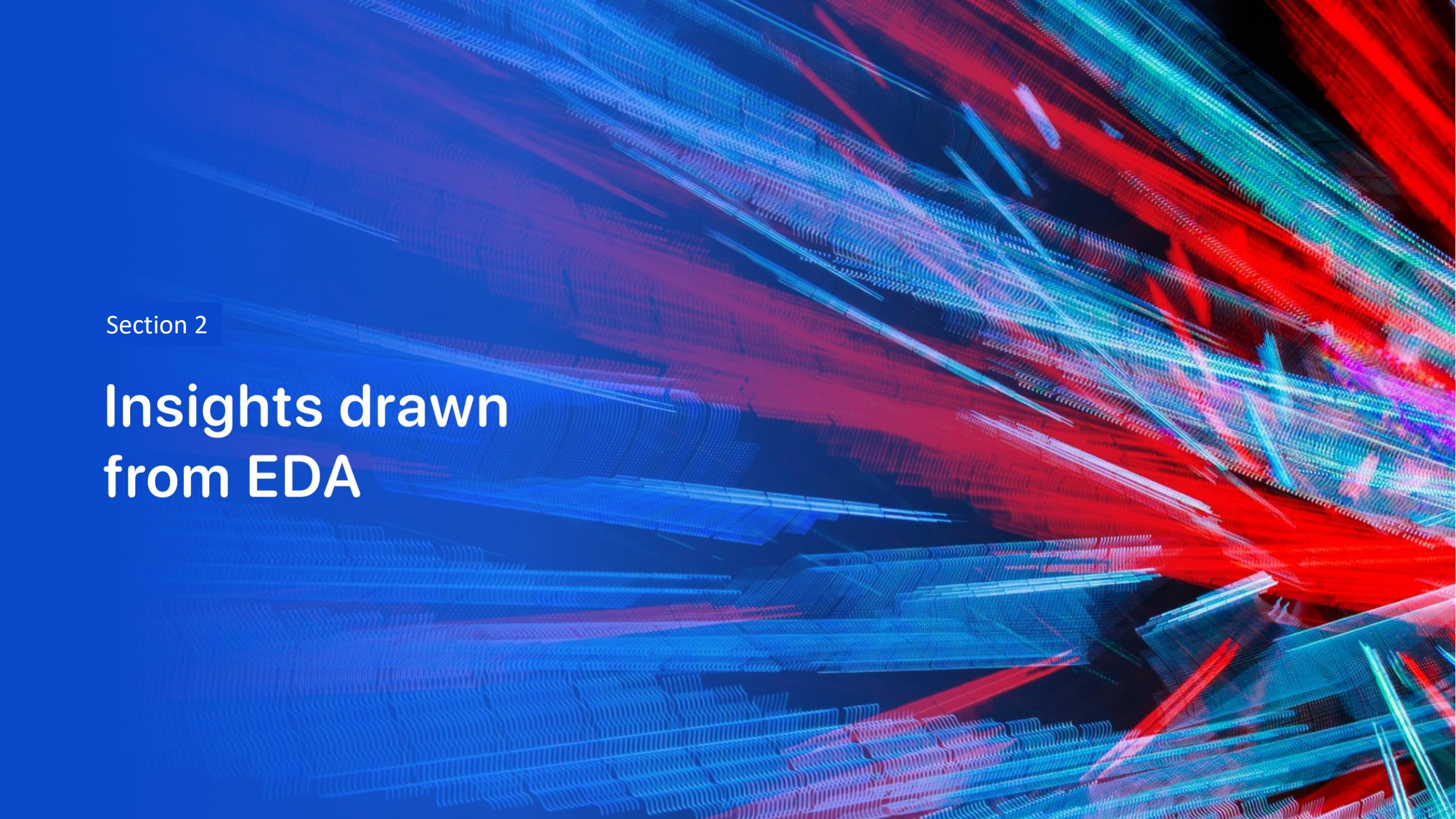
Predictive Analysis (Classification)

- Machine Learning Models used:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbors (KNN)
- GridSearchCV performed multiple iterations of each architecture to identify optimal hyperparameters.
- https://github.com/Mdawg27265/DataCapstone/blob/6d08379c439fe8e8a7eebf2ed8913329ee6035ca/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

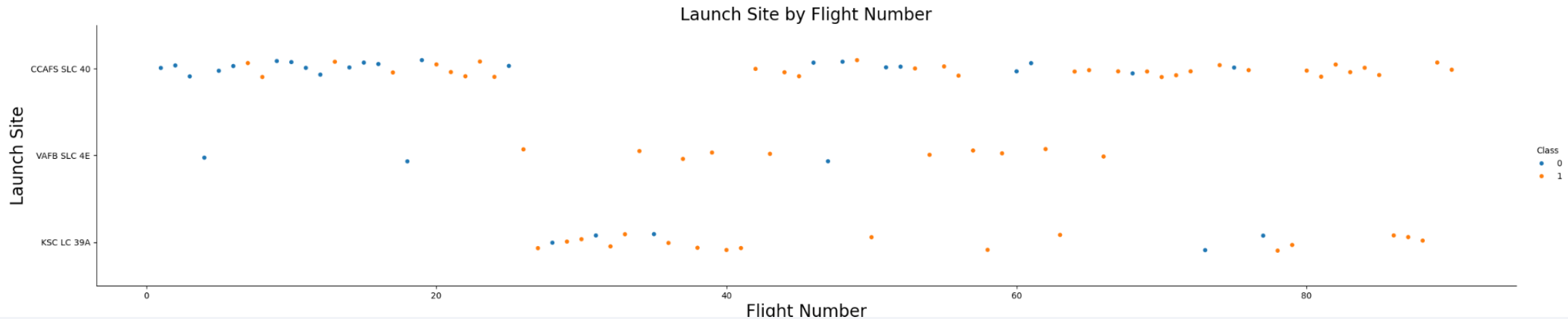


Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
  - Mapping of launch sites and performance
  - Interactive Dashboard
- Predictive analysis results
  - Comparison of model performance
  - Confusion Matrix for best performing model

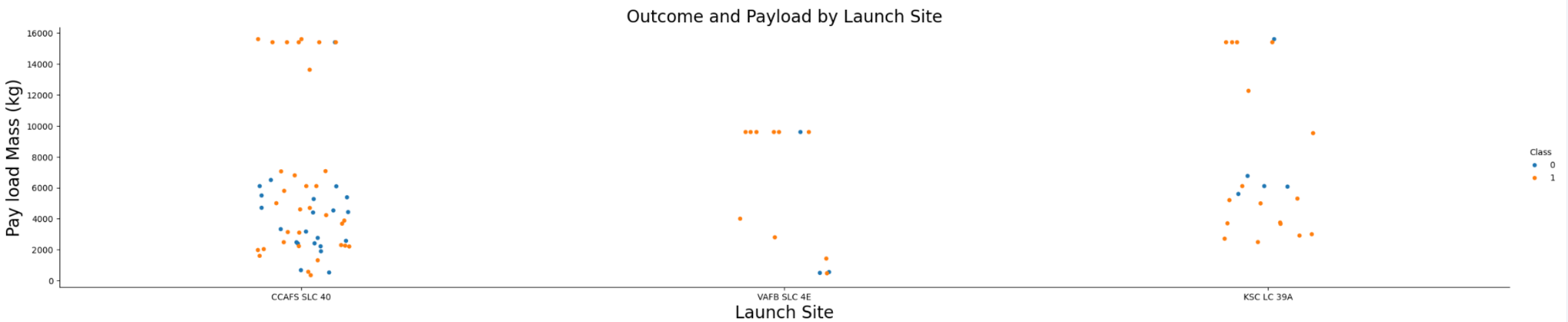

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Later launches have a higher rate of success

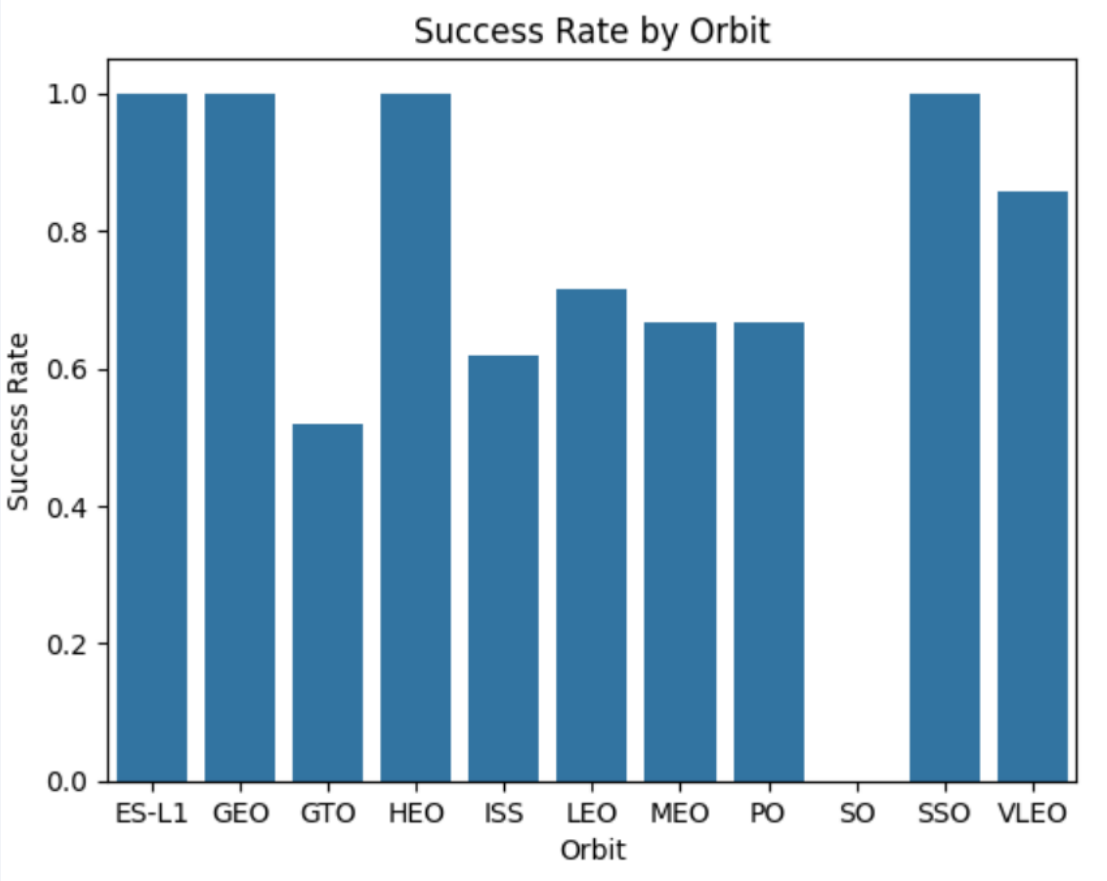
Payload vs. Launch Site



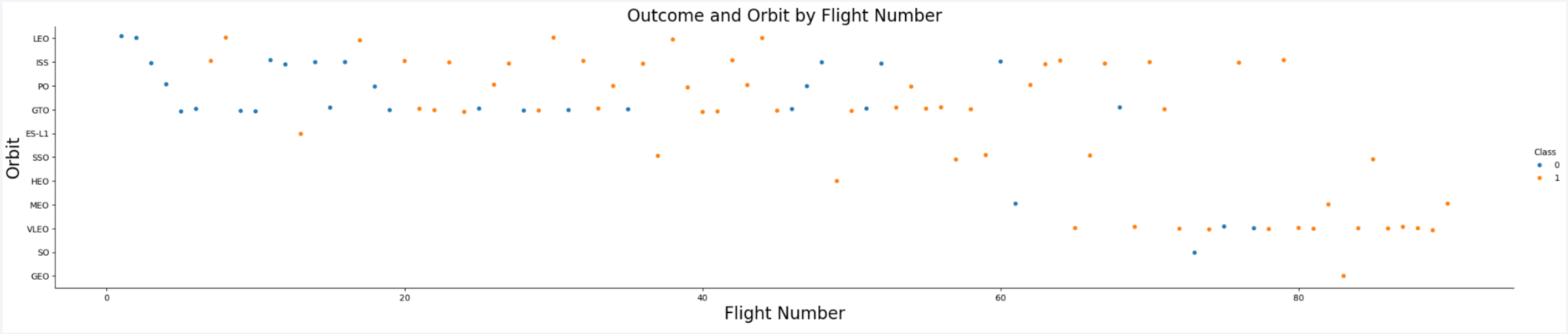
- CCAFS SLC 40 has universal success with higher payload masses.
- VAFB SLC-4E has not launched payloads > 10K kg
- KSC LC39A has high success at lower and higher masses

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO are associated with 100% success
- GTO has lowest success rate at approx. 50%.

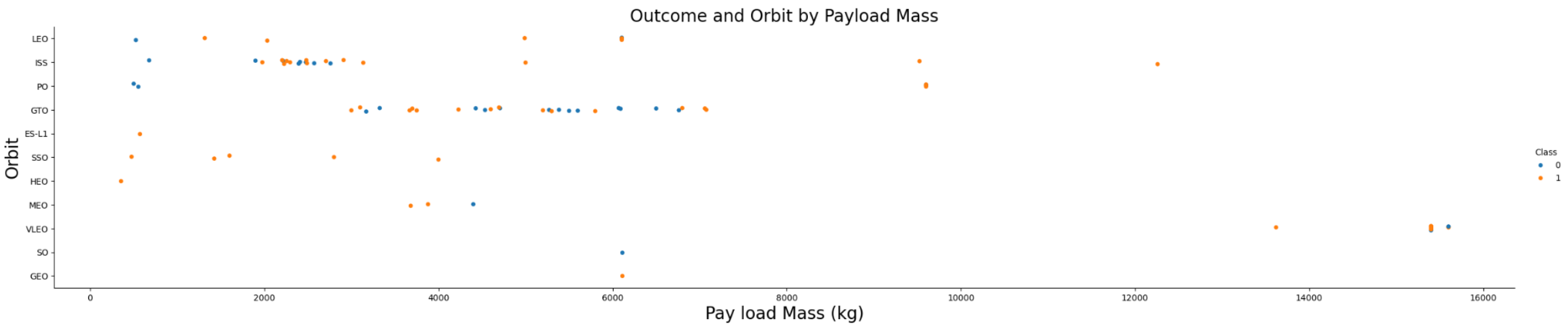


Flight Number vs. Orbit Type



- Most flights go to GTO.
- Early flights did not go to VLEO, but later flights do frequently.

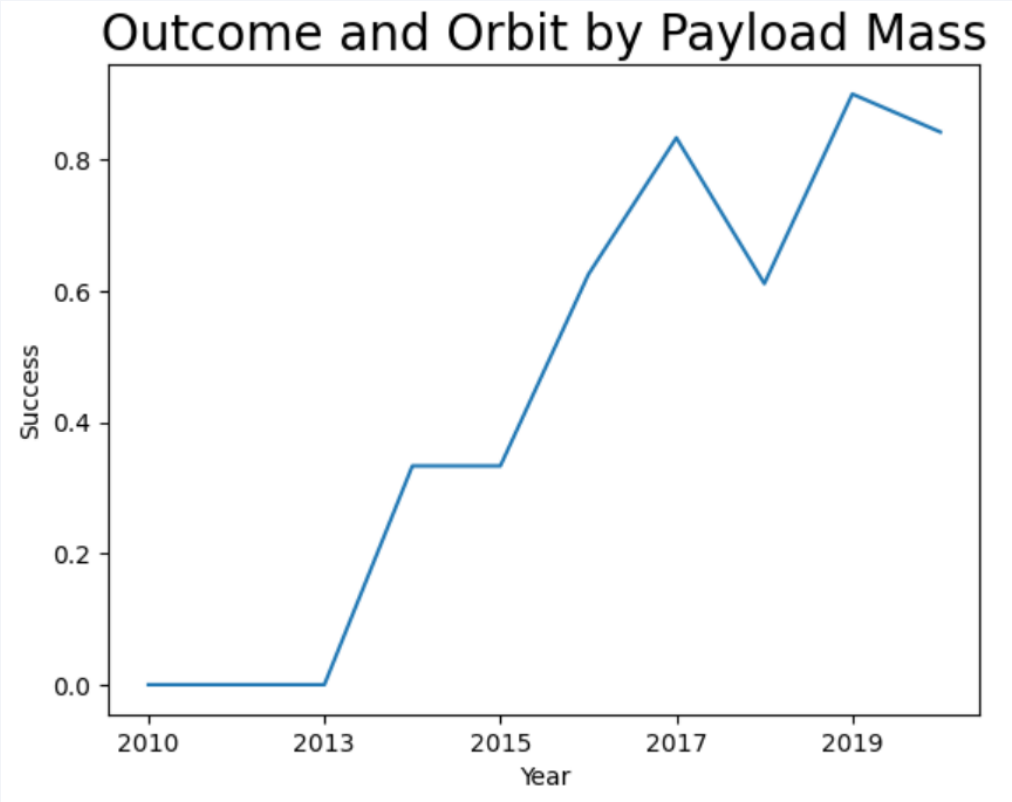
Payload vs. Orbit Type



- Launches heading for the ISS Orbit are typically 2000-3000kg.
- GTO Orbit launches tend toward 4000-6000kg.

Launch Success Yearly Trend

- Success rate has generally increased since 2013.
- 2020 showed a slight decrease in success rate, but still second best performance.



All Launch Site Names

- Find the names of the unique launch sites

In [12]:
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[12]:

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`

In [15]:
```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[15]:

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

In [23]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass FROM SPACEXTBL WHERE Customer LIKE 'NASA (CRS)%';
```

```
 * sqlite:///my_data1.db
Done.
```

Out[23]:

total_payload_mass
48213

Average Payload Mass by F9 v1.1

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

```
[13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS average_payload_mass FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%';
 * sqlite:///my_data1.db
Done.
```

[13]:

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

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

```
[15]: %sql SELECT MIN(Date) AS first_ground_pad_landing FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

[15]:

first_ground_pad_landing
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

```
[16]: %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
 * sqlite:///my_data1.db
Done.
```

[16]:

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

2015 Launch Records

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

- Present your query result with a short explanation here

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

- Present your query result with a short explanation here

Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites

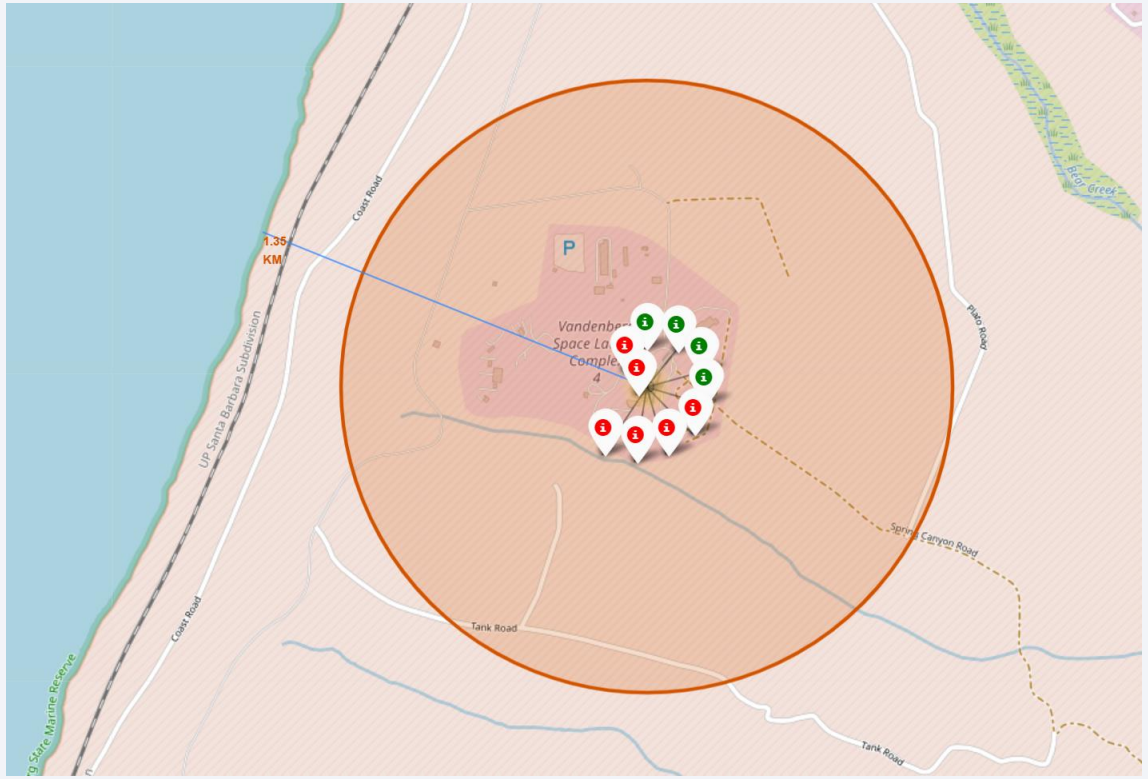
- There is a launch site in Southern California
- There are three launch sites on the East Coast of Florida.

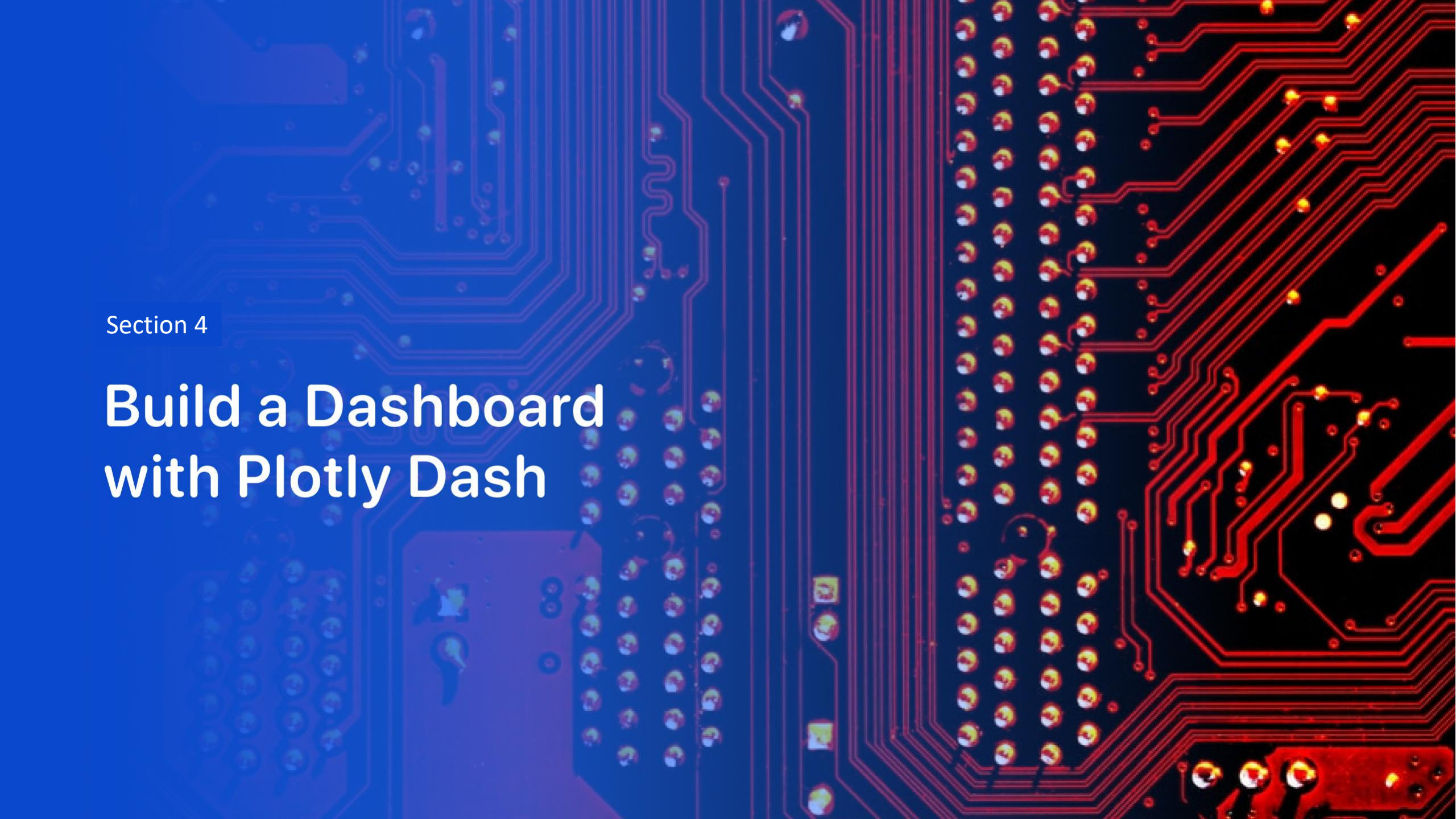
<Folium Map Screenshot 2>

- There are more launches in Florida than in California

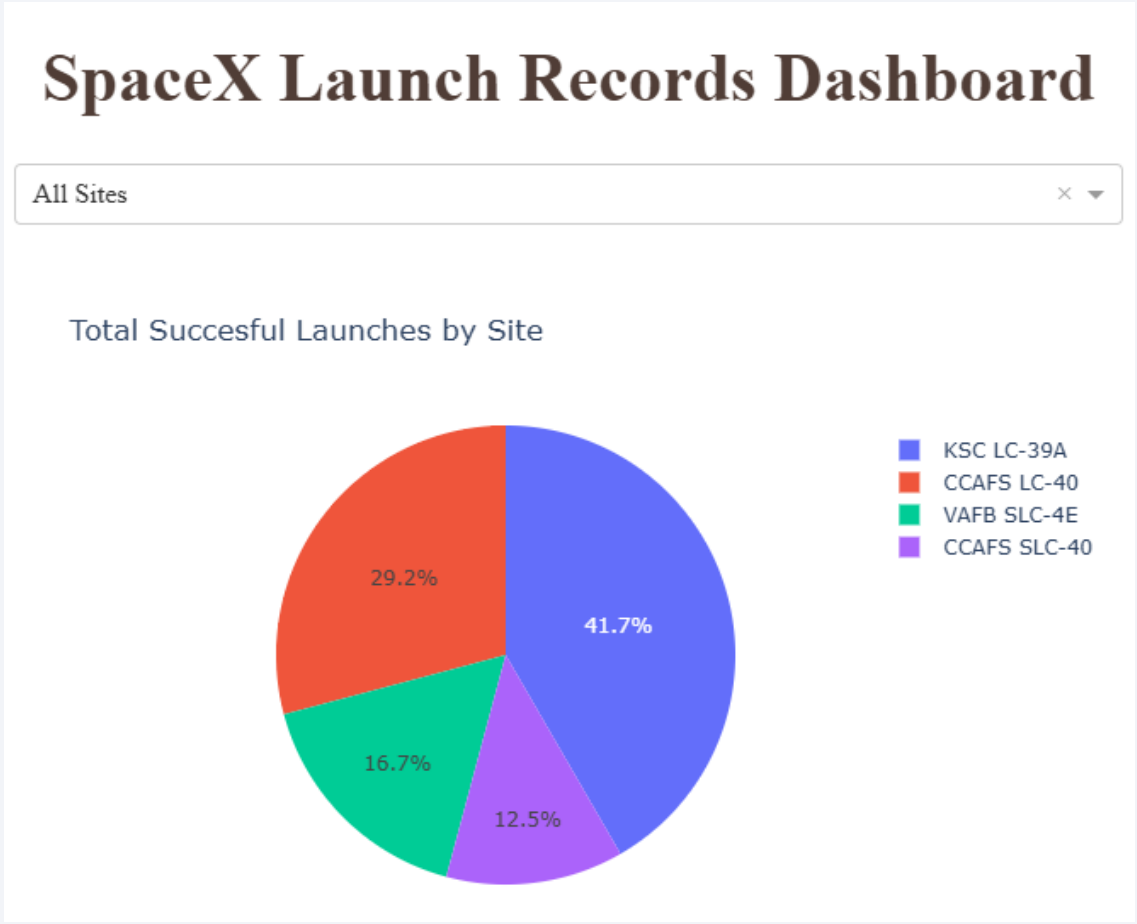
VAFB Results and proximity to coastline

- Folium map showing Vandenburg Space Complex (VAFB SLC-4E)
- Proximal to railroads and coastline, but not directly adjacent to any populated towns.
- 1.35km to coast



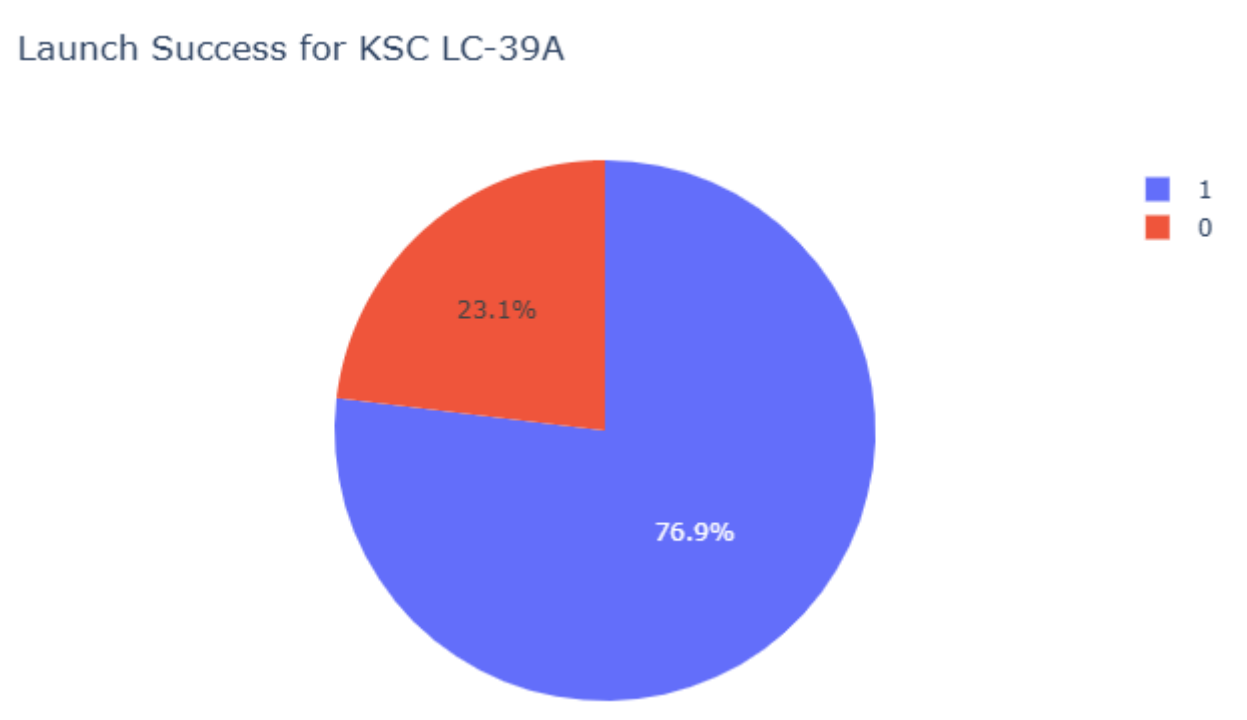

Section 4
Build a Dashboard with Plotly Dash

Launch Success relationship to Launch Site



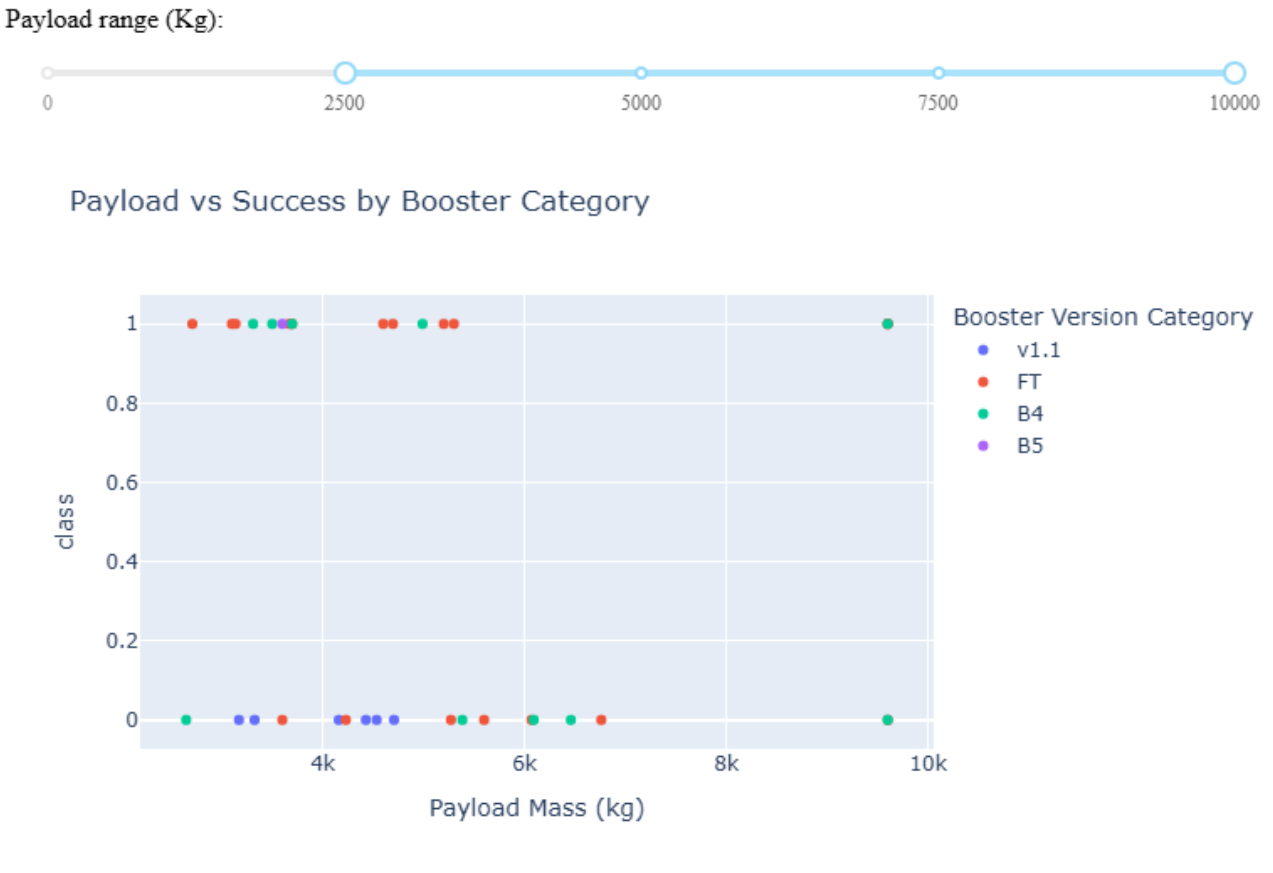
- The majority of successful recoveries have been launched from KSC LC-39A.
- CCAFS SLC-40 launch site is associated with the least amount of successful recoveries.

Results for Best Performing Site (KSC LC-39A)



- 22 launches from KSC LC-39A in Cape Canaveral, FL
- 76.9% success rate is highest among all Launch Sites

Payload vs Success



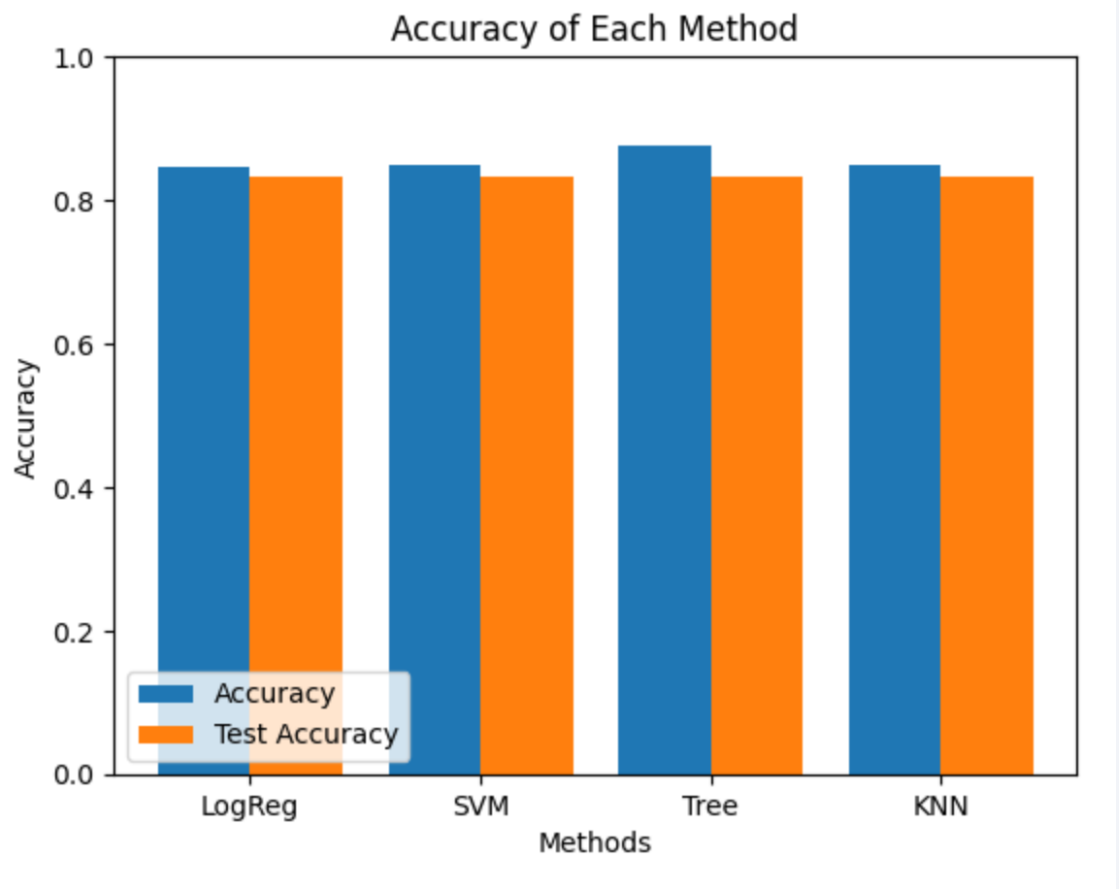
- The payload mass feature is examined for the launch sites and booster categories.
- In the 2k-6k range, the FT and B4 boosters are associated more with successful outcomes.
- Heavier payloads seem less likely to succeed, although there is limited data at heavier payloads.

Section 5

Predictive Analysis (Classification)

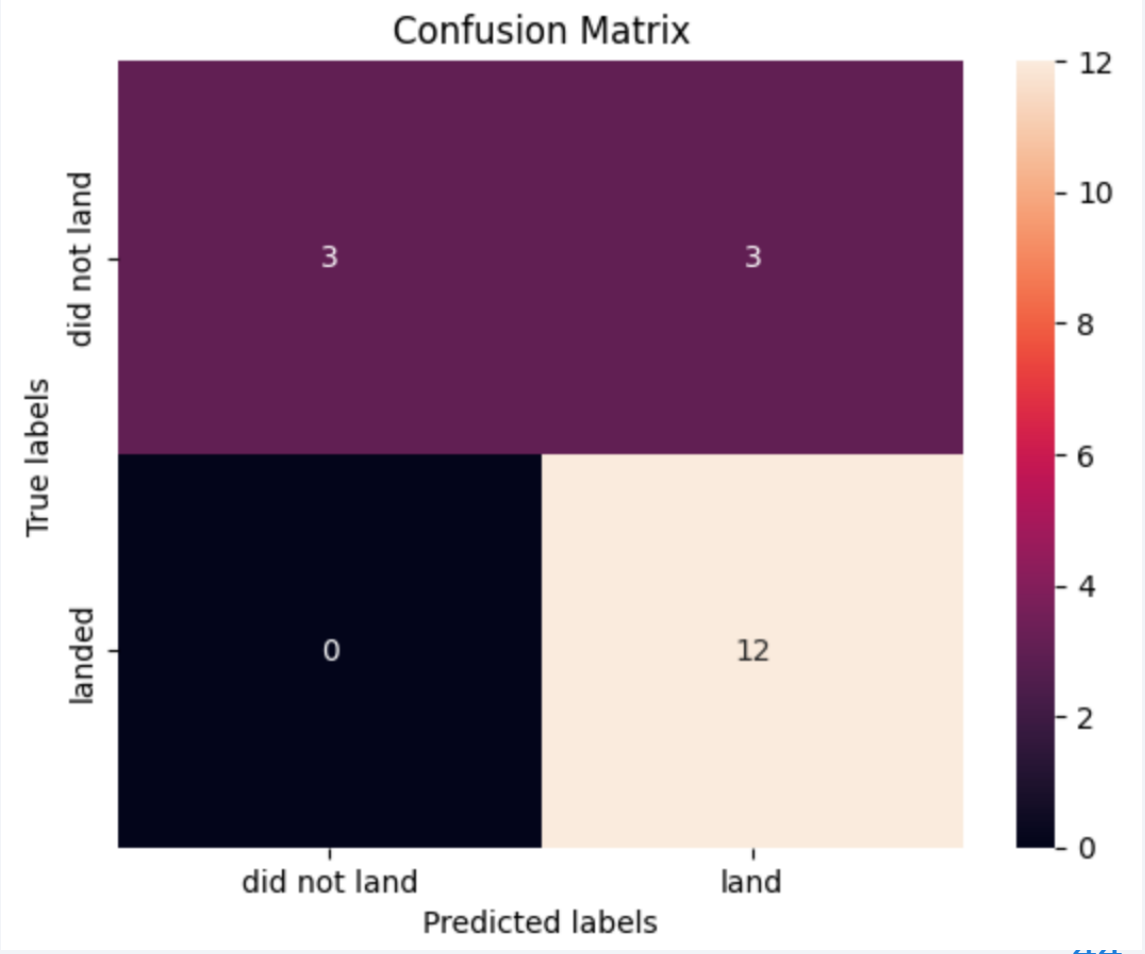
Classification Accuracy

- Decision Tree had the highest accuracy among all model architectures at 87.5%
- Accuracy on testing data was slightly lower at 83.3% or 1 incorrect classification out of 12.



Decision Tree Confusion Matrix

- The model exhibited difficulty with False Alarms (i.e., predicting a landing when craft does not).
- In the case in which the craft landed successfully, the model was able to predict 100%



Conclusions

- Using publicly available information, we can collect data on SpaceX launch performance
- Taking into account features like Orbit, Launch Site, Booster Reusage, we can build and evaluate ML models to predict the chances of successful recovery of Stage 1
- This ML model will allow us to anticipate competitor success and strategize more robustly.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!