IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

BRYAN IGBOEGWU
16th September 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- **Summary of Methodologies**

- The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies where used:

- **Collect** data using SpaceX REST API and web scraping techniques
- **Wrangle** data to create success/fail outcome variable
- **Explore** data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- **Analyze** the data with SQL, calculating the following statistics: total payload, payload
- range for successful launches, and total # of successful and failed outcomes
- **Explore** launch site success rates and proximity to geographical markers
- **Visualize** the launch sites with the most success and successful payload ranges
- **Build Models** to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

Introduction

- **Background**

- SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

- **Explore**

- How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- Rate of successful landings over time
- Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Executive Summary

- **Steps**

I. **Collect** data using SpaceX REST API and web scraping techniques

II. **Wrangle** data – by filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling

III. **Explore** data via EDA with SQL and data visualization techniques

IV. **Visualize** the data using Folium and Plotly Dash

V. **Build Models** to predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters

Data Collection

- **Steps**

  I. **Request data** from SpaceX API (rocket launch data)

  II. **Decode response** using .json() and convert to a dataframe using .json_normalize()

  III. **Request information** about the launches from SpaceX API using custom functions

  IV. **Create dictionary** from the data

  V. **Create dataframe** from the dictionary

  VI. **Filter dataframe** to contain only Falcon 9 launches

  VII. **Replace missing values** of Payload Mass with calculated .mean()

  VIII. **Export data** to csv file

Data Collection - Scraping

- **Steps**

- **Request data** (Falcon 9 launch data) from Wikipedia
- **Create BeautifulSoup object** from HTML response
- **Extract column names** from HTML table header
- **Collect data** from parsing HTML tables
- **Create dictionary** from the data
- **Create dataframe** from the dictionary
- **Export data** to csv file

https://github.com/igboegwu/IBM-Data-science-Project/blob/dddff153b26f3602a40d86711ea3f7a8e6e65106/02_SpaceX_Web_Scraping.ipynb

Data Wrangling

- **Perform EDA** and determine data labels
- **Calculate:**
  - # of launches for each site
  - # and occurrence of orbit
  - # and occurrence of mission
  - outcome per orbit type]
- **Create binary** landing outcome
- column (dependent variable)
- **Export data** to csv file

Landing Outcome

- Landing was not always successful
- **True Ocean:** mission outcome had a successful landing to a specific region of the ocean

Landing Outcome Cont

- **False Ocean:** represented an unsuccessful landing to a specific region of ocean
- **True RTLS:** meant the mission had a successful landing on a ground pad
- **False RTLS:** represented an unsuccessful landing on a ground pad
- **True ASDS:** meant the mission outcome had a successful landing on a drone ship
- **False ASDS:** represented an unsuccessful landing on drone ship
- **Outcomes converted** into 1 for a successful landing and 0 for an unsuccessful landing

EDA with Data Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Analysis

- **View relationship** by using **scatter plots**. The variables could be useful for machine learning if a relationship exists
- **Show comparisons** among discrete categories with **bar charts**. Bar charts show the relationships among the categories and a measured value.

https://github.com/igboegwu/IBM-Data-science-Project/blob/dddff153b26f3602a40d86711ea3f7a8e6e65106/05_SpaceX_EDA_Data_Visualization.ipynb

EDA with SQL

- **Queries**
- **Display:**
- Names of unique launch sites
- 5 records where launch site begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.
- **List:**
- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have
- payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

https://github.com/igboegwu/IBM-Data-science-Project/blob/dddff153b26f3602a40d86711ea3f7a8e6e65106/04_SpaceX_EDA_SQL.ipynb

Build an Interactive Map with Folium

- **Markers Indicating Launch Sites**

- Added **blue** **circle** at **NASA Johnson Space Center's coordinate** with a
- **popup label** showing its name using its latitude and longitude coordinates

- Added **red** **circles** at **all launch sites coordinates** with a **popup label**
- showing its name using its name using its latitude and longitude coordinates

Colored Markers of Launch Outcomes

- Added **colored markers** of **successful** (green) and **unsuccessful** (red) **launches** at each launch site to show which launch sites have high success rates

Distances Between a Launch Site to Proximities

- Added **colored lines** to **show distance between** launch site **CCAFS SLC-40 and** its proximity to the **nearest coastline, railway, highway, and city**

Build a Dashboard with Plotly Dash

https://github.com/igboegwu/IBM-Data-science-Project/blob/dddff153b26f3602a40d86711ea3f7a8e6e65106/07_SpaceX_Interactive_Visual_Analytics_Plotly.py

- **Dropdown List with Launch Sites**
- Allow user to select all launch sites or a certain launch site
- **Pie Chart Showing Successful Launches**
- Allow user to see successful and unsuccessful launches as a percent of the total
- **Slider of Payload Mass Range**
- Allow user to select payload mass range
- **Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version**
- Allow user to see the correlation between Payload and Launch Success

Predictive Analysis (Classification)

- **Charts**

- **Create** NumPy array from the Class column
- **Standardize** the data with StandardScaler. Fit and transform the data.
- **Split** the data using train_test_split
- **Create** a GridSearchCV object with cv=10 for parameter optimization
- **Apply** GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
- **Calculate** accuracy on the test data using .score() for all models
- **Assess** the confusion matrix for all models
- **Identify** the best model using Jaccard_Score, F1_Score and Accuracy

Results

- Exploratory Data Analysis
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- Visual Analytics
- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities
- Predictive Analytics
- Decision Tree model is the best predictive model for the dataset

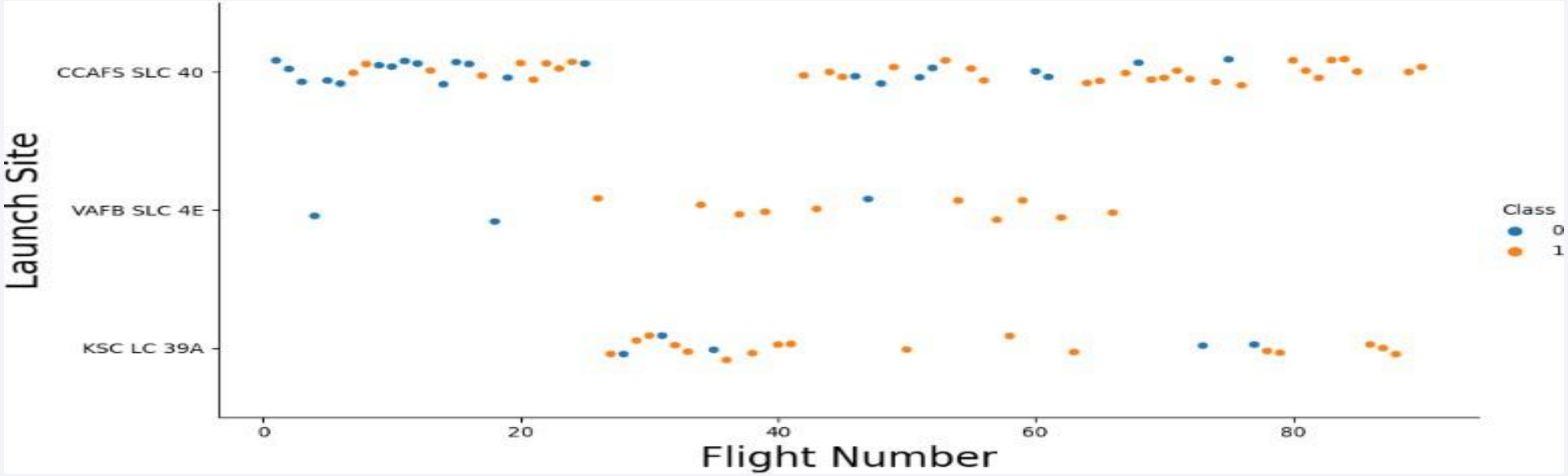
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- **Exploratory Data Analysis**

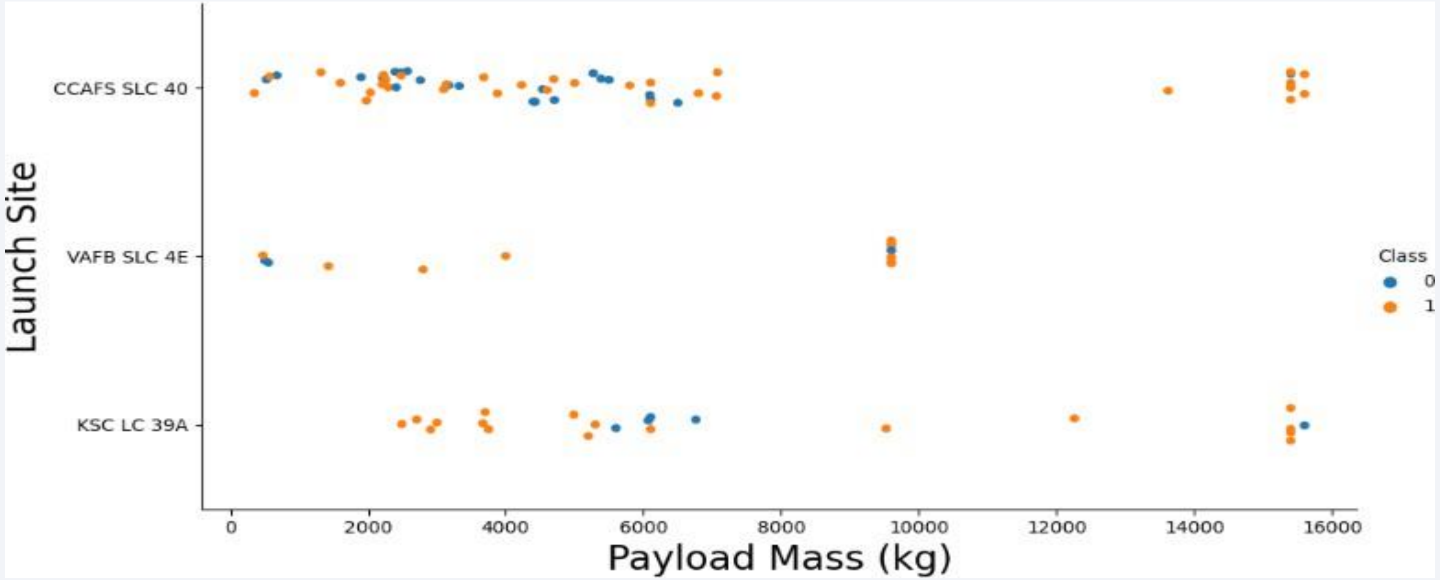
- **Earlier flights** had a **lower success rate** (**blue = fail**)
- **Later flights** had a **higher success rate** (**orange = success**)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



Payload vs. Launch Site

- **Exploratory Data Analysis**

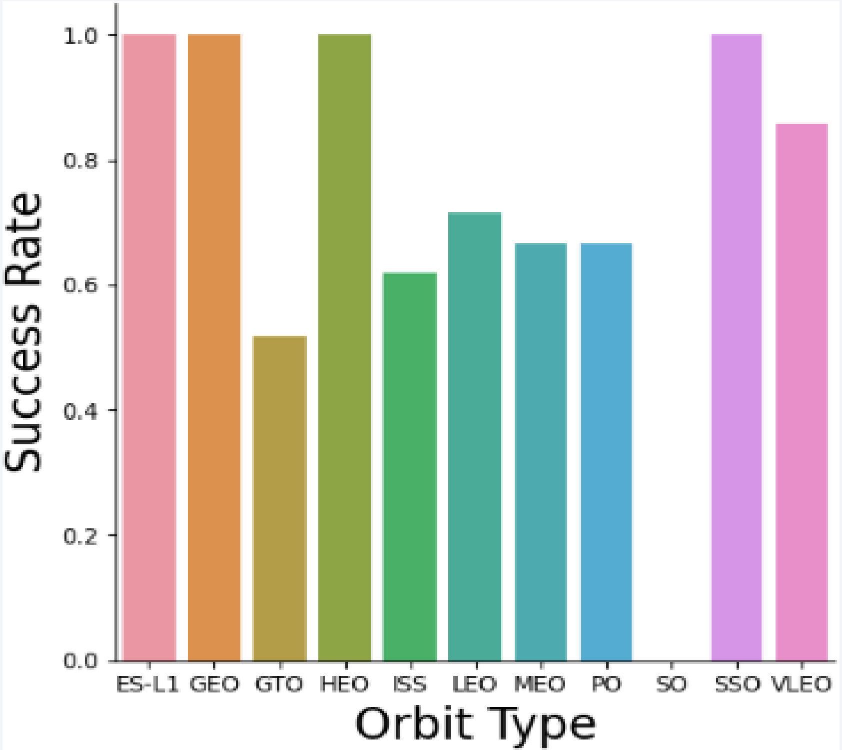
- Typically, the **higher** the **payload mass** (kg), the **higher** the **success rate**
- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Success Rate vs. Orbit Type

- **Exploratory Data Analysis**

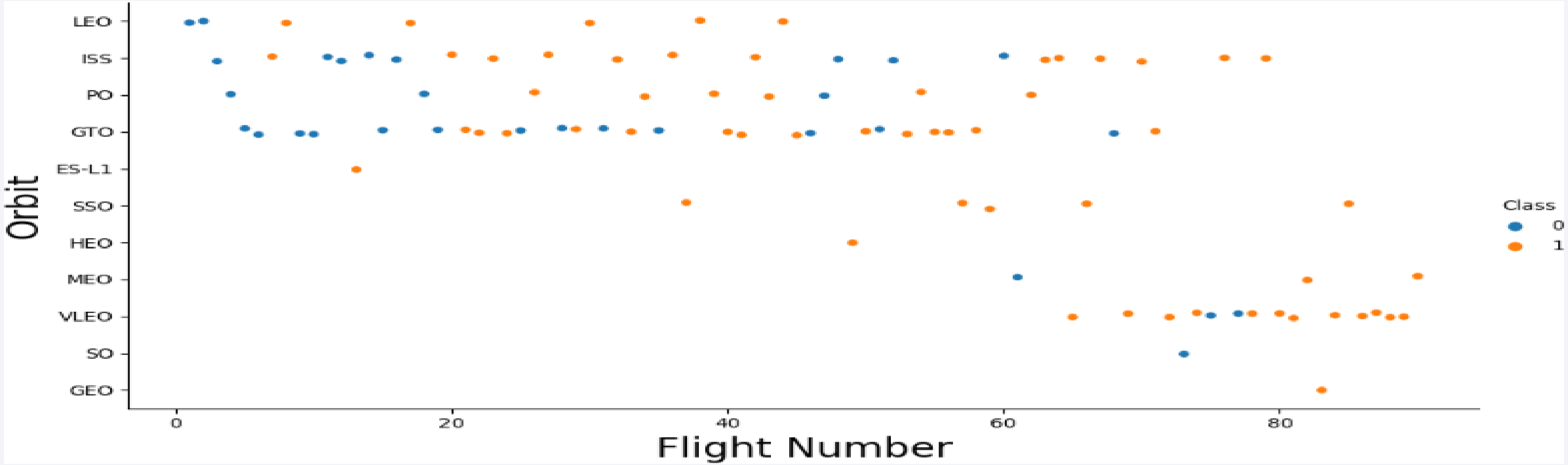
- **100% Success Rate**: ES-L1, GEO, HEO and SSO
- **50%-80% Success Rate**: GTO, ISS, LEO, MEO, PO
- **0% Success Rate**: SO



Flight Number vs. Orbit Type

- **Exploratory Data Analysis**

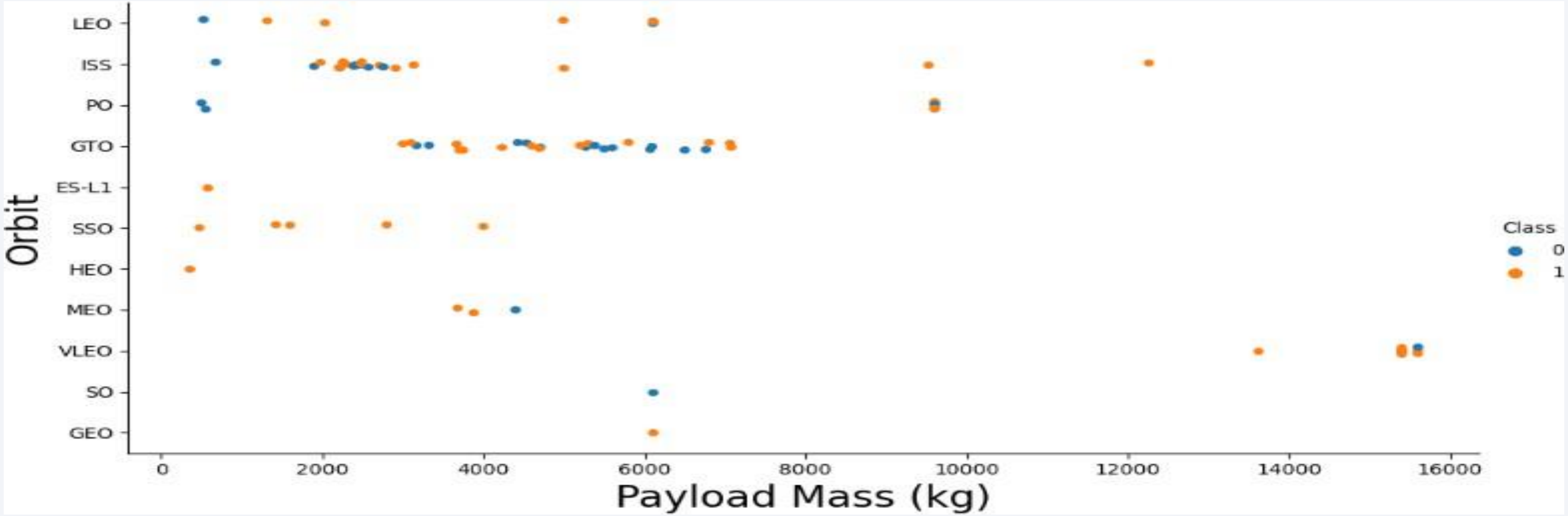
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



Payload vs. Orbit Type

- **Exploratory Data Analysis**

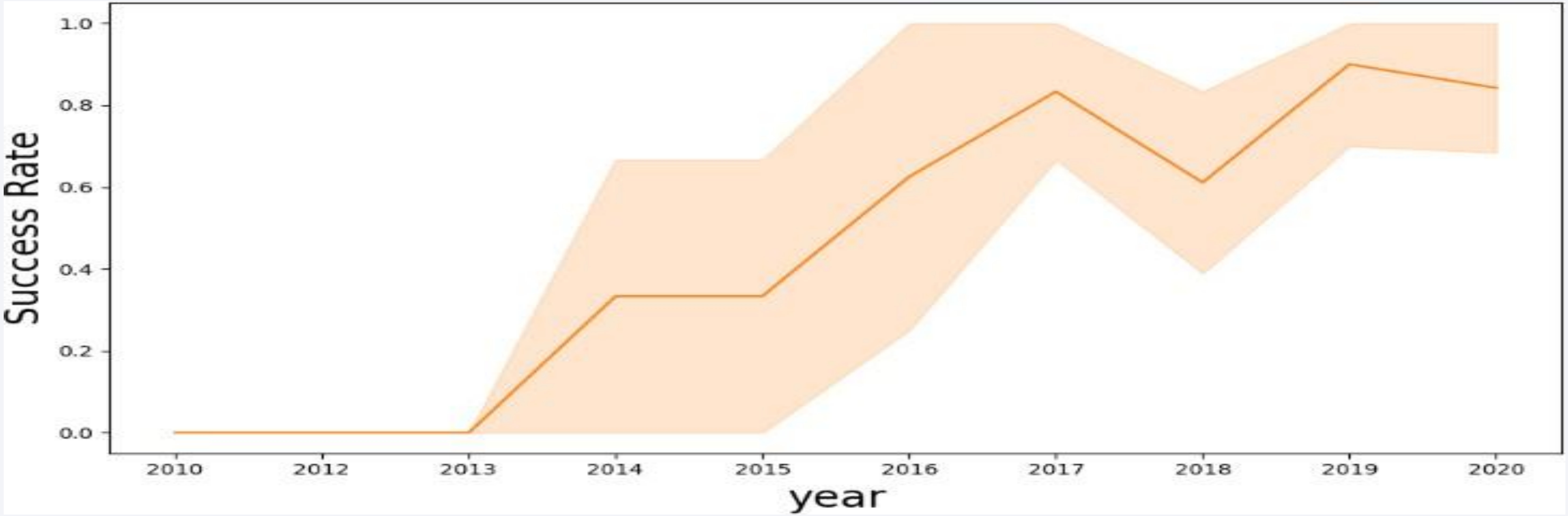
- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

Exploratory Data Analysis

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



All Launch Site Names

- **Launch Site Names** **Landing Outcome Cont.**
- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
%sql SELECT * \
    FROM SPACEXTBL \
    WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
   sqlite:///my_data1.db
Done.
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- **45,596 kg** (total) carried by boosters launched by NASA(CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4l
   sqlite:///my_data1.db
Done.
```

1
45596

Average Payload Mass by F9 v1.1

- **2,928 kg** (average) carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
     FROM SPACEXTBL \
     WHERE BOOSTER_VERSION = 'F9 v1.1';

 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.

    1
2928
```

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

- **Ranked Descending**
- Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

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

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

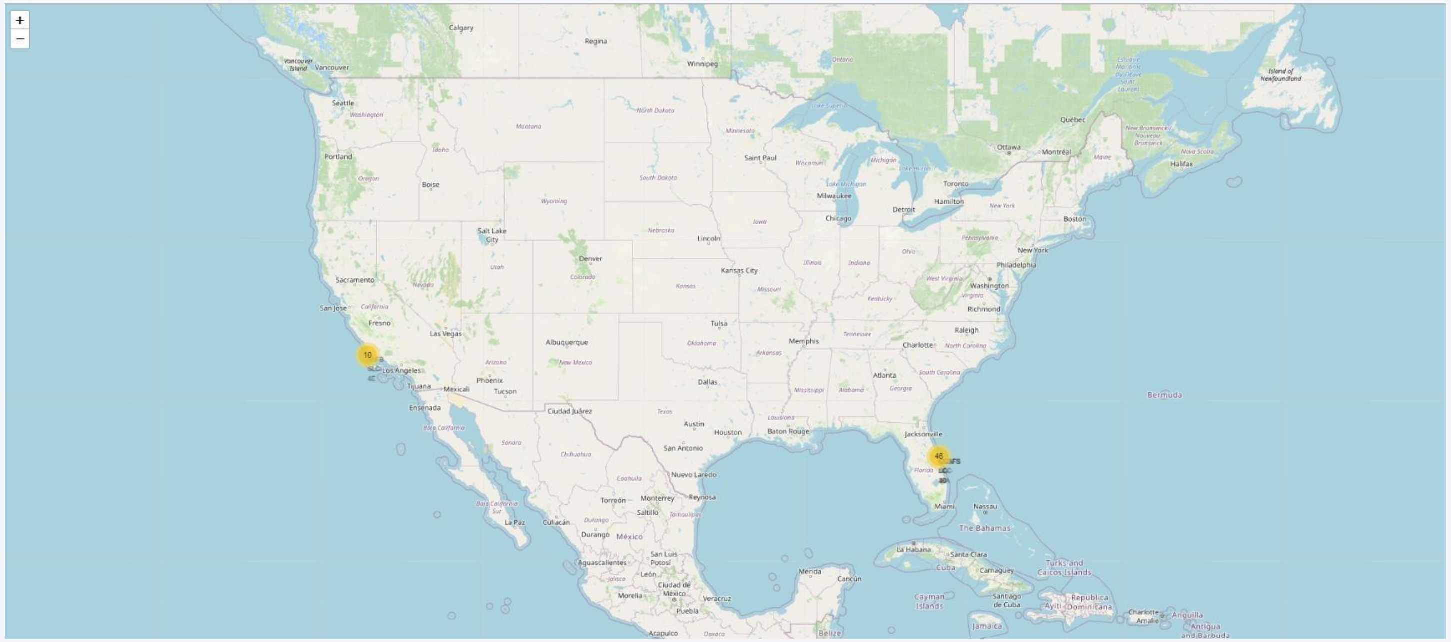
Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- **With Markers**

- **Near Equator**: the closer the launch site to the equator, the **easier** it is **to launch** to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an **additional natural boost** - due to the rotational speed of earth - that **helps save the cost** of putting in extra fuel and boosters.



<Folium Map Screenshot 2>

- **At Each Launch Site**
- **Outcomes**:
- **Green** markers for successful launches
- **Red** markers for unsuccessful launches
- Launch site **CCAFS SLC-40** has a **3/7 success rate** (**42.9%**)

<Folium Map Screenshot 3>

- **CCAFS SLC-40**
- **.86 km** from nearest coastline
- **21.96 km** from nearest railway
- **23.23 km** from nearest city
- **26.88 km** from nearest highway

Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

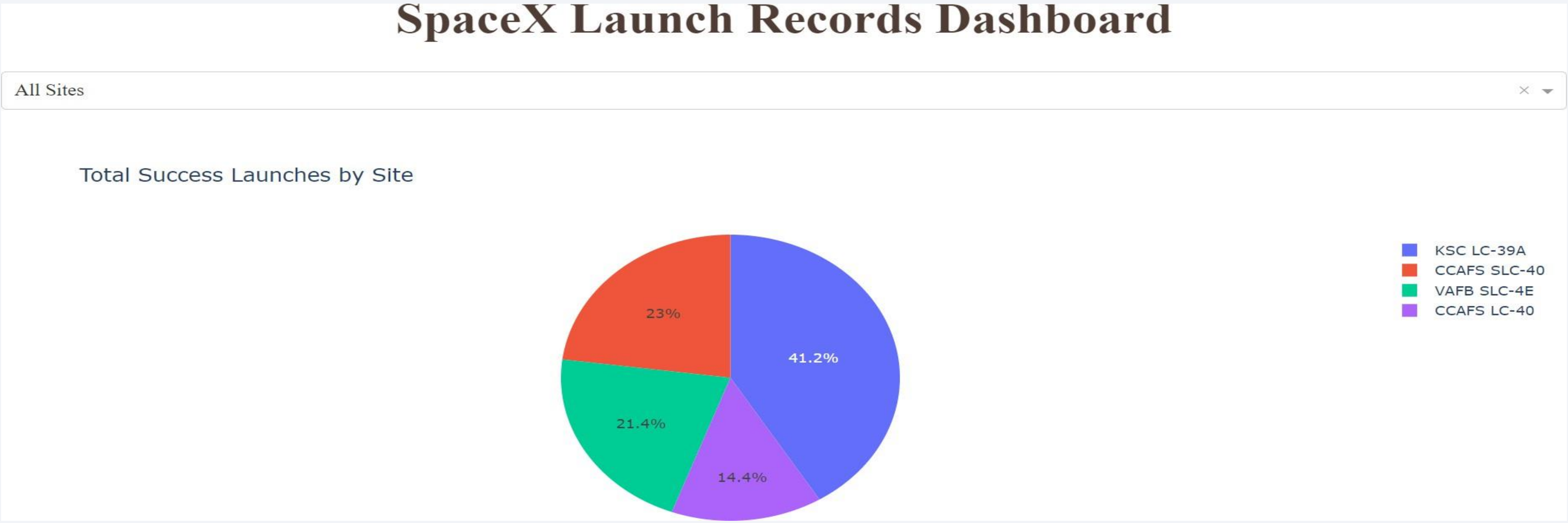
- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- **Success as Percent of Total**
- **KSC LC-39A** has the **most successful launches** amongst launch sites (**41.2%**)



Confusion Matrix

- **Performance Summary**
- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- Confusion Matrix Outputs:
  - 12 True positive
  - 3 True negative
  - **3 False positive**
  - 0 False Negative
- **Precision** = TP / (TP + FP)
  - 12 / 15 = .80
- **Recall** = TP / (TP + FN)
  - 12 / 12 = 1
- **F1 Score** = 2 * (Precision * Recall) / (Precision + Recall)
  - 2 * (.8 * 1) / (.8 + 1) = .89
- **Accuracy** = (TP + TN) / (TP + TN + FP + FN) = .833

Conclusions

- **Research**
- **Model Performance**: The models performed similarly on the test set with the decision tree model slightly outperforming
- **Equator**: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast**: All the launch sites are close to the coast
- **Launch Success**: Increases over time
- **KSC LC-39A**: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits**: ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass**: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Conclusions

- Things to Consider
- Dataset: A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set
- Feature Analysis / PCA: Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy
- XGBoost: Is a powerful model which was not utilized in this study. It would
- be interesting to see if it outperforms the other classification models

Thank you!