IBM Developer
SKILLS NETWORK

SPACEX

Winning the Space Race with Data Science

Oscar G
12/29/2022

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- Data Visualization
- Data Analysis with SQL
- Building an Interactive Map with Folium
- Building a dashboard with Plotly Dash
- Predictive Analysis (classification)

Summary of all results

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

Introduction

Background

SpaceX is an American aerospace manufacturer and space transport services company founded in 2002 by Elon Musk. The company has developed a reusable launch system for rockets and spacecraft, with the ultimate goal of reducing space transportation costs and enabling the colonization of Mars. For this project, I've created a data analysis model designed to predict and determine the financial viability of Space X's first stage rockets and their ability to land based on primary datasets.

Problem

- What variables help us predict the success of first stage rocket landings?
- What does the rate of success look like over time?
- What can predictive modeling (classification) tell us about the future of these landings?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Space X REST API
    - Webscraping from Wikipedia
- Perform data wrangling
    - Data filtering
    - Replacing missing values
    - Preparing the data for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Tuned classification models

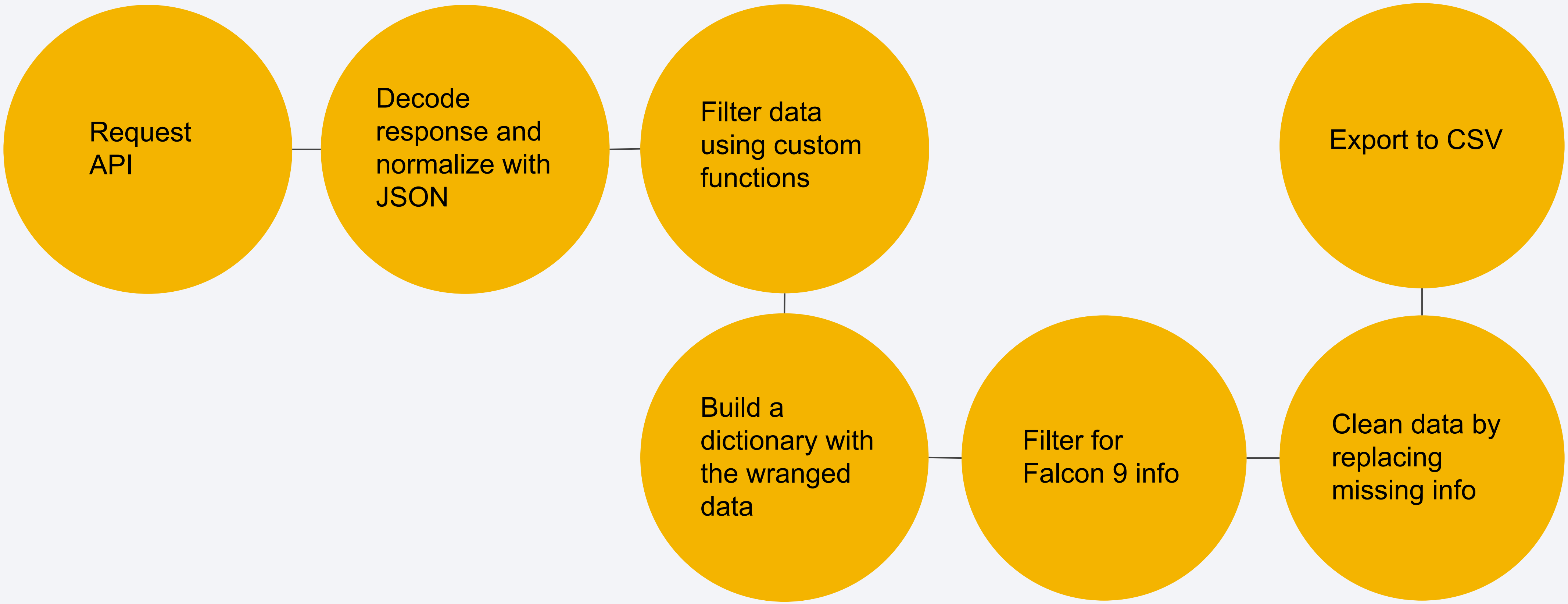
Data Collection

For this project I used two data collection methods: Rest API and Webscraping

APIs allow developers to access and retrieve data from a specific website or service using a set of predefined instructions, often in the form of a request-response model. Web scraping, on the other hand, involves extracting data from a website by parsing through the HTML or XML code of a webpage. Web scraping can be used to extract specific data elements from a webpage or to download the entire webpage for further processing.

To get information about Space X's rocket launches I converted REST API into a pandas dataframe and cleaned the data. As for the webscraping, I used BeautifulSoup to pull information from Wikipedia, also converting it into a datafram for analysis.

Data Collection – SpaceX API



Github URL:
https://github.com/oscarhgriffin/IBM-Data-Science-Professional-Certificate/blob/master/IBM%20Data%20Science%20Capstone/Final_Assignment1.1%20Collecting%20Data.ipynb

Data Collection - Scraping

1. Request Wikipedia HTML info
2. Create a BeautifulSoup Object
3. Extract column names from the table
4. Parse the data and create a dictionary
5. Create a dataframe
6. Export data to CSV

Github URL:
https://github.com/oscarhgriffin/IBM-Data-Science-Professional-Certificate/blob/master/IBM%20Data%20Science%20Capstone/Final_assignment1.2%20Web%20Scraping.ipynb

Data Wrangling

Data wrangling is the process of cleaning, organizing, and transforming raw data into a usable format for analysis. This often involves correcting errors or inconsistencies in the data, filling in missing values, and converting the data into a consistent format.

For this project I developed a training label converting the information into binary outcome: ‘1’ for successful landing and ‘0’ for unsuccessful landings.

Github url: https://github.com/oscarhgriffin/IBM-Data-Science-Professional-Certificate/blob/master/IBM%20Data%20Science%20Capstone/Final_Assignment1.3%20Data%20Wrangling.ipynb

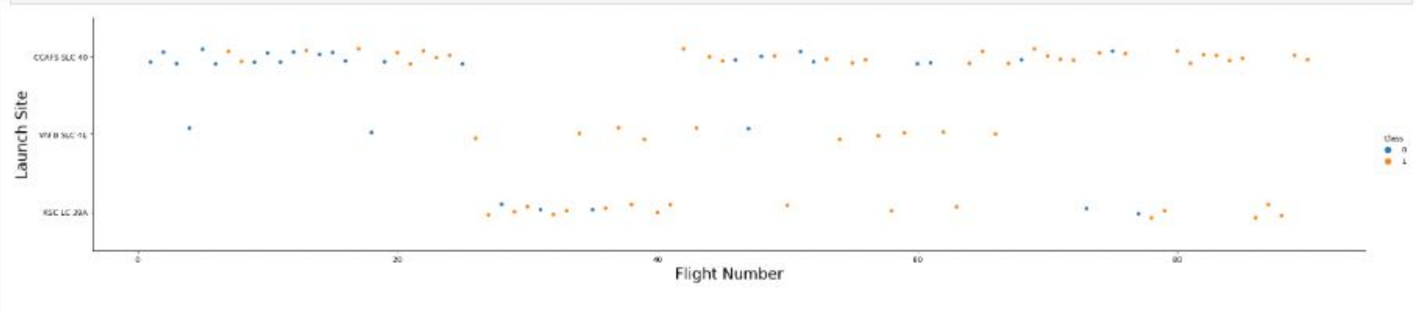
Wrangling Process

1. Determine initial training labels
2. Calculate launches
3. Calculate mission outcomes
4. Create outcome label from the output column
5. Export to CSV

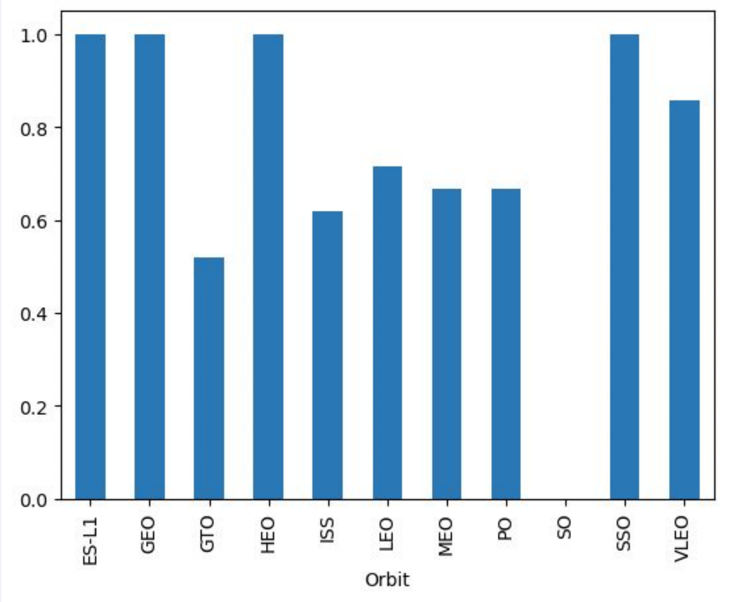
EDA with Data Visualization

Charts plotted:

- Payload Mass vs. Flight Number
- Flight Number vs Launch Site
- Payload Mass vs. Launch Site
- Success Rate vs Orbit Type
- Payload vs Orbit Type
- Flight number vs Orbit Type



Scatterplots for showing dependency between variables and bar charts for comparing discrete categories.



Github url:
https://github.com/oscarhgriffin/IBM-Data-Science-Professional-Certificate/blob/master/IBM%20Data%20Science%20Capstone/Final_Assignment2.1%20Analysis%20with%20SQL.ipynb

EDA with SQL

SQL Queries:

- Display unique launch site names
- Display first 5 rows where the column contains “CCA”
- Display total payload mass carried by boosters launched by NASA
- Display average payload mass carried by booster version f9
- List the date when the first success landing was achieved
- List the names of the boosters which have success in drone ship and have a payload mass between 4000 and 6000
- List total number of successful and failed landings
- List number of booster versions carrying max payload mass
- List failed landing outcomes for drone ship, booster versions, and launch sites for 2015
- Rank the landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

Github URL:
https://github.com/oscarhgriffin/IBM-Data-Science-Professional-Certificate/blob/master/IBM%20Data%20Science%20Capstone/Final_Assignment2.2%20Pandas%20and%20Matplotlib.ipynb

Build an Interactive Map with Folium

- Created a map object, folium.circle, to highlight a specific coordinate: NASA Johnson Space Center.
- Used folium.circle, and folium.marker to indicate all launch sites using their corresponding coordinates to show their location on the map.
- Used colored markers to indicate successful and failed launches using Marker Cluster.
- Determined distances between launch sites, adding colored lines to show proximity.

Github URL:
https://github.com/oscarhgriffin/IBM-Data-Science-Professional-Certificate/blob/master/IBM%20Data%20Science%20Capstone/Final_Assignment3%20Visual%20Analytics%20and%20Dashboards.ipynb

Build a Dashboard with Plotly Dash

Launch Sites Dropdown

- Created a dropdown list for uses to select Launch Site

Pie Chart of Successful Launches

- Created a pie chart to show percentage of successful launches

Slider for Payload Range

- Added a slider to enter payload range

Scatter chart: payload mass vs Success Rate

- Added a scatter chart to show the relationship between the Payload and Launch success variables.

Github URL:
https://github.com/oscarhgriffin/IBM-Data-Science-Professional-Certificate/tree/master/IBM%20Data%20Science%20Capstone/Interactive%20Dashboard%20using%20Plotly

Predictive Analysis (Classification)

Build

- Load the dataset into NumPy/Pandas.
- Transform the data and split for training and test datasets.
- Determine ML type to use.
- Set parameters to GridSearchCV and fit it to the dataset

Evaluate

- Check the accuracy of each model
- Tune hyperparameters
- Plot confusion matrices

Refine

- Calculate the accuracy of the test data using method.score()
- Review the confusion matrices for all models

Finalize

- Determine the best method by measuring the accuracy score and reviewing metrics

Github URL: https://github.com/oscarhgriffin/IBM-Data-Science-Professional-Certificate/blob/master/IBM%20Data%20Science%20Capstone/Final_Assignment4%20Predictive%20Analysis.ipynb

Results

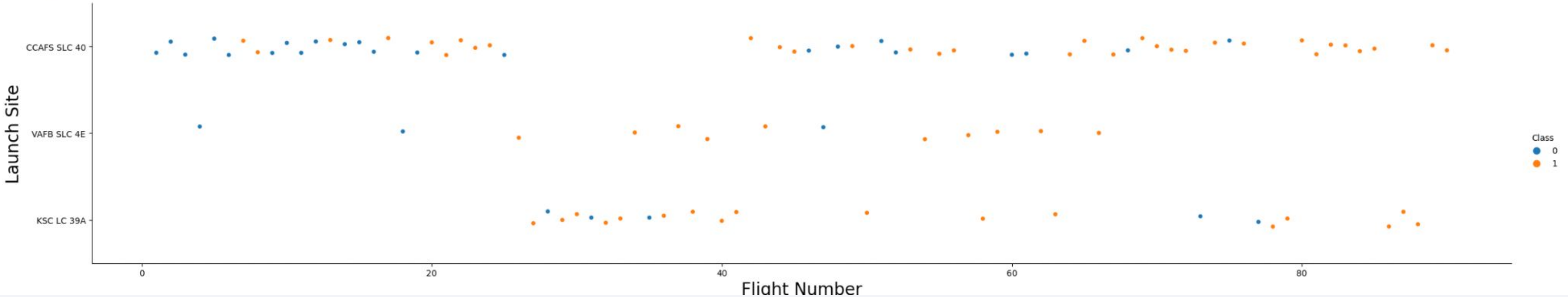
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

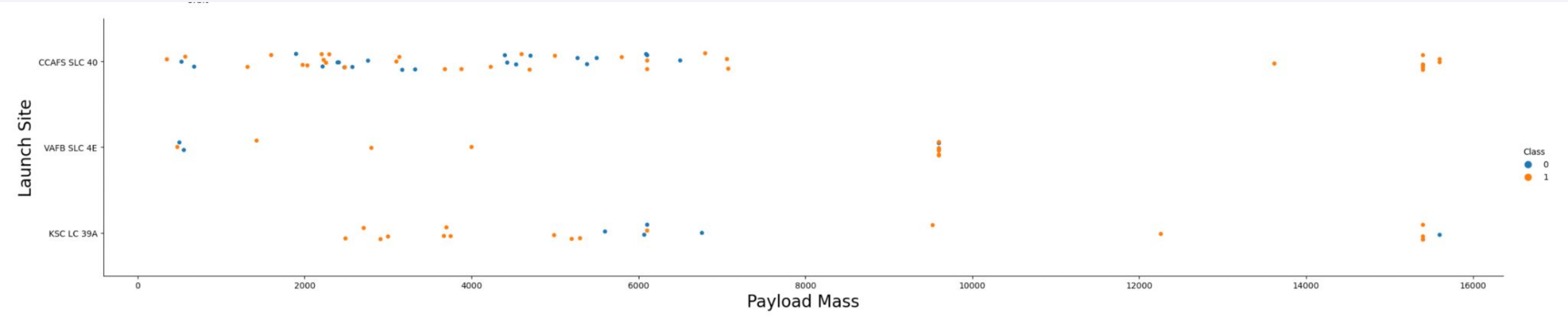
```python
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



Insights:

- Flight success increased in later flights.
- CCAFS SLC 40 launch site bore a large majority of the tests.

Payload vs. Launch Site



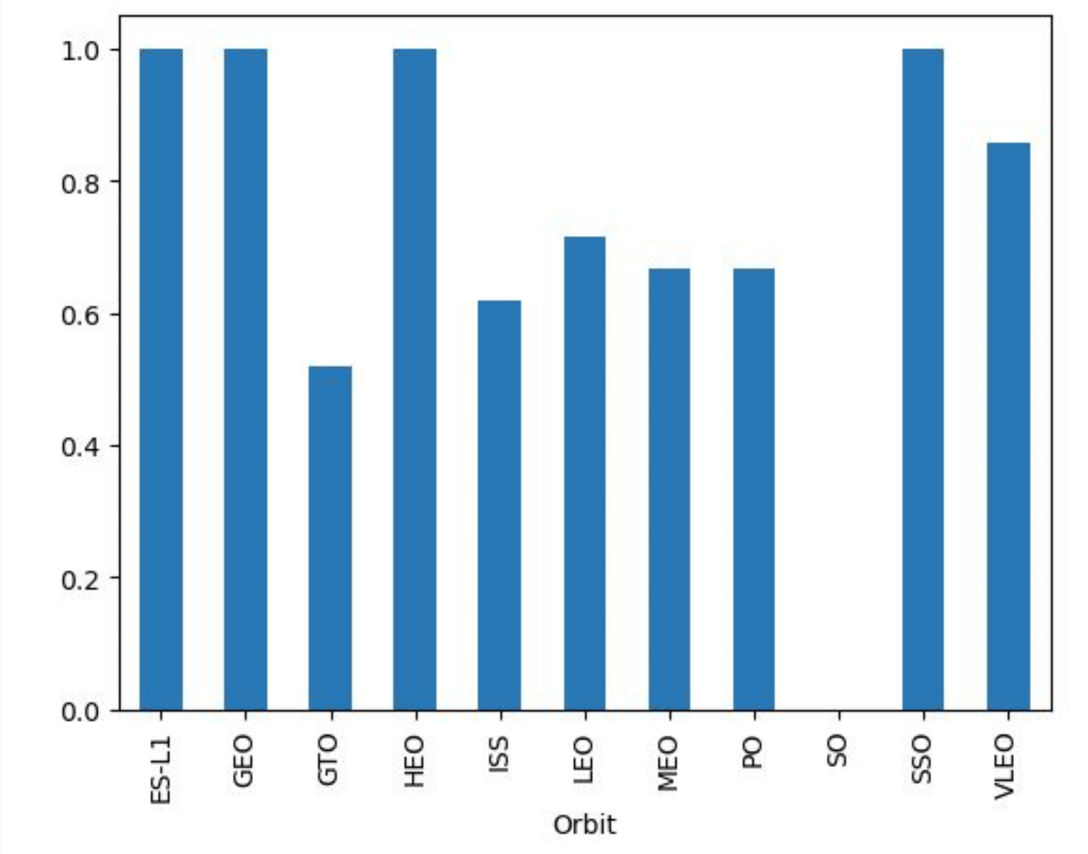
Insights:

- While larger payload mass appear to have higher success rate, there aren't enough data points to draw a strong correlation.
- Launches were between 0 and about 7,000 kg.
- Again here the CCAFS SLC 40 launch site has the most launch data.

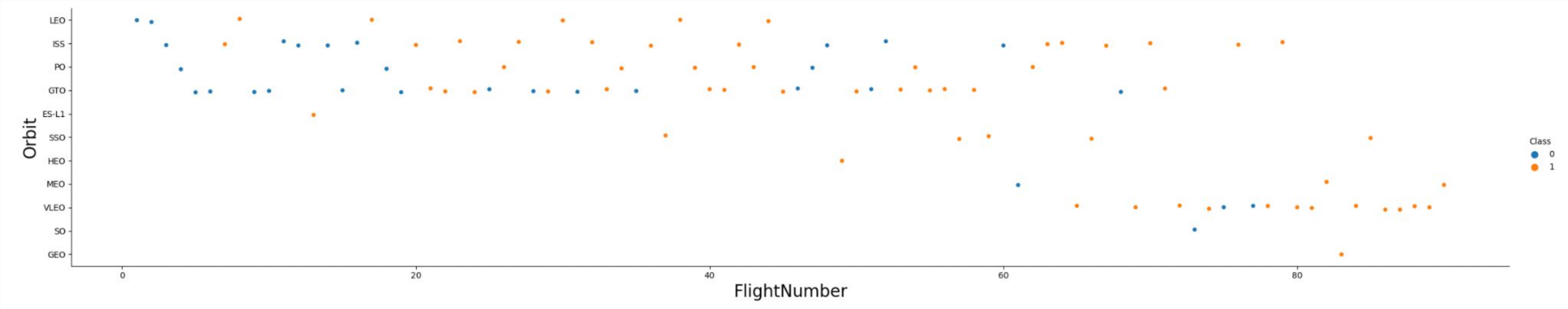
Success Rate vs. Orbit Type

Insights

- Orbits with the highest success rate
  - ESL1, GEO, HEO, SSO
- Orbit with the lowest success rate
  - SO
- Orbits with variable success
  - GTO, ISS, LEO, MEO, PO, VLEO



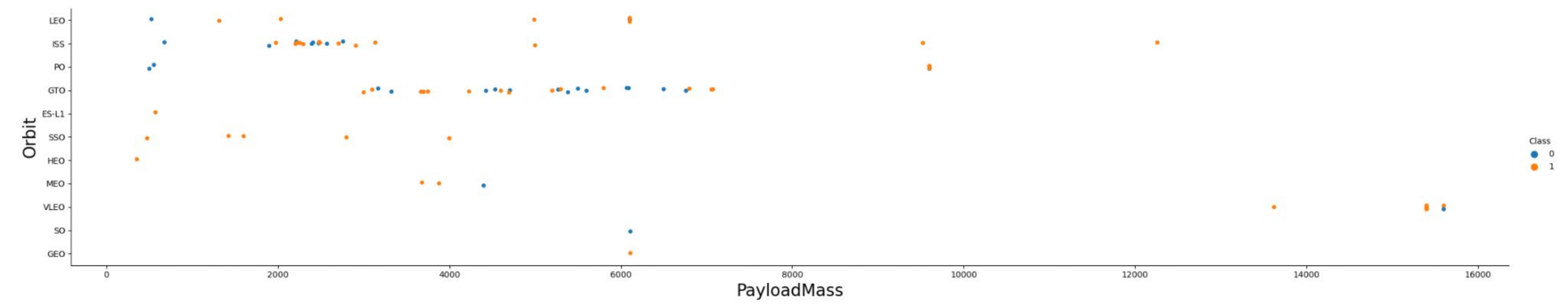
Flight Number vs. Orbit Type



Insights:

- Space X experienced varying results between LEO, ISS, PO and GTO orbits in early flights.
- VLEO shows a more consistent success rate in most recent flights.

Payload vs. Orbit Type



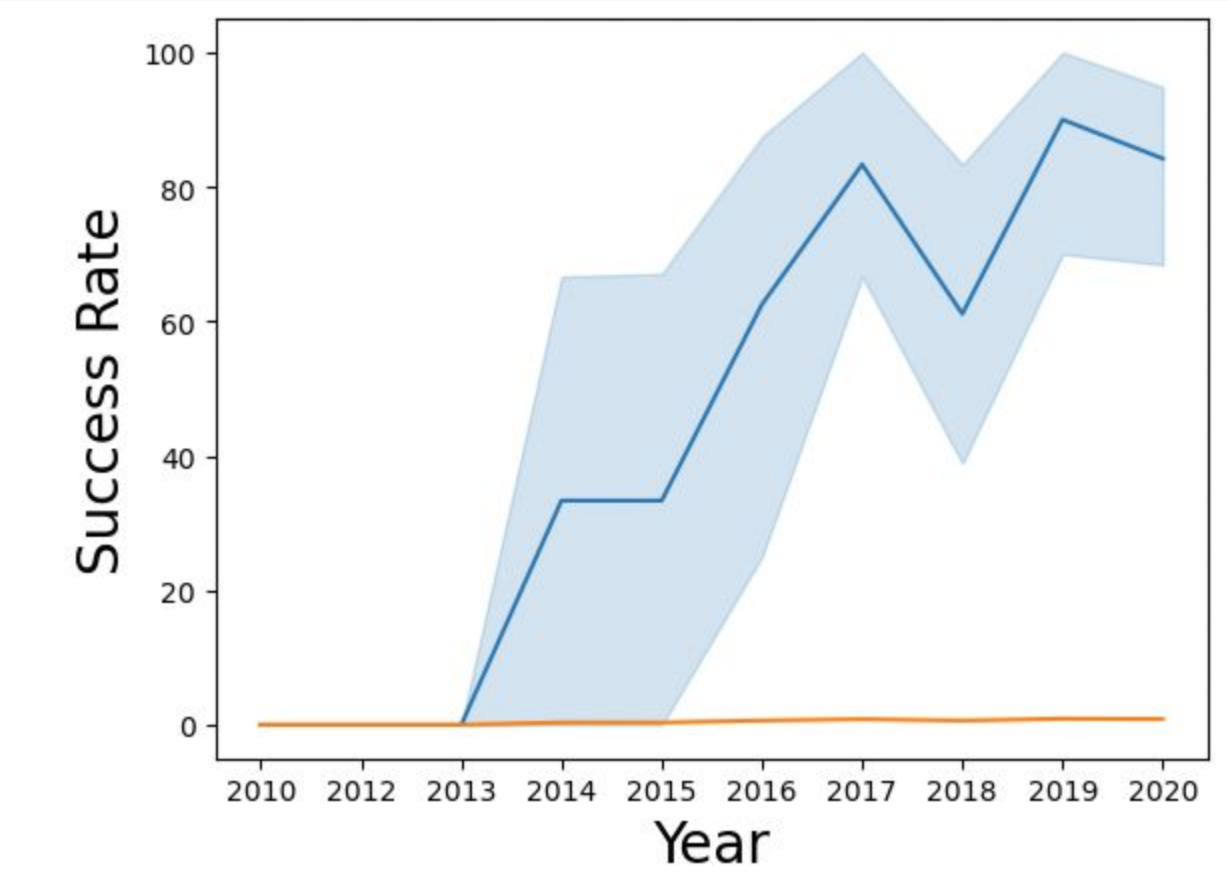
Insights:

- Payload mass shows a strong correlation to orbit.
- Payload mass below 6000 for SSO shows a 100% success rate.
- Generally, payload masses in the low to mid range with SSO orbits would have the highest success rate.

Launch Success Yearly Trend

Insights

- Success rates have trended upwards since 2013.



All Launch Site Names

This process shows the syntax for querying site names from the Space X database.

```
%sql select DISTINCT LAUNCH_SITE from SPACEXDATA
```

```
 * ibm_db_sa://jwj46821:***@9938aec0-8105-433e-8bf9-0fbb
gj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/BLUDB
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXDATA where launch_site like 'CCA%' limit 5
```

```
* ibm_db_sa://jwj46821:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/BLUDB
Done.
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	0:35:00	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

This syntax displays 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as sum from SPACEXDATA where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://jwj46821:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.datab
ases.appdomain.cloud:32459/BLUDB
Done.
```

SUM
45596

Summing the total payload mass in KG for rockets launched by NASA

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as Average from SPACEXDATA where booster_version like 'F9 v1.1%'
 * ibm_db_sa://jwj46821:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdom
Done.
```

average
2534

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

First Successful Ground Landing Date

Finding the dates of the first successful landing outcome on ground pad

```
%sql select min(date) as Date from SPACEXDATA where mission_outcome like 'Success'
```

```
 * ibm_db_sa://jwj46821:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde0
Done.
```

DATE
2010-04-06

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXDATA where (mission_outcome like 'Success') AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome like 'Success (drone ship)')
```

```
 * ibm_db_sa://jwj46821:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/BLUDB
Done.
```

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXDATA GROUP by mission_outcome ORDER BY mission_outcome
 * ibm_db_sa://jwj46821:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/BLUDB
Done.
```

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

Calculating the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

Listing the names of the booster which have carried the maximum payload mass

```
maxm = %sql select max(payload_mass__kg_) from SPACEXDATA
maxv = maxm[0][0]
%sql select booster_version from SPACEXDATA where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXDATA)
```

```
 * ibm_db_sa://jwj46821:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/BLUDB
Done.
 * ibm_db_sa://jwj46821:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/BLUDB
Done.
```

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

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

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site from SPACEXDATA where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

```
 * ibm_db_sa://jwj46821:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/BLUDB
Done.
```

MONTH	landing__outcome	booster_version	launch_site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Ranking 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

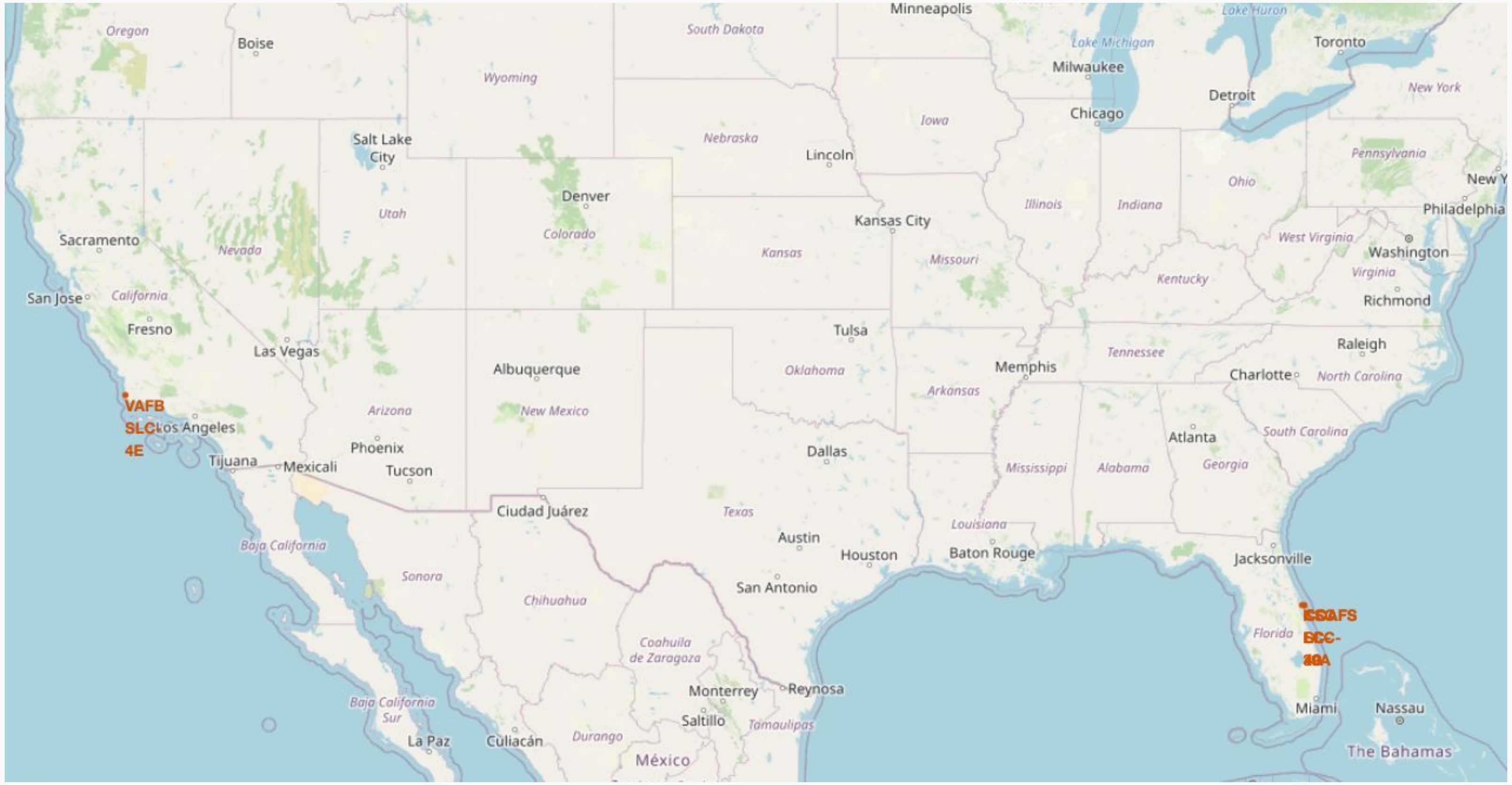
```
%sql select landing__outcome, count(*) as count from SPACEXDATA where Date >= '2010-06-04' AND Date <= '2017-03-20' GROUP by landing__outcome ORDER BY count Desc
```

```
 * ibm_db_sa://jwj46821:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/BLUDB
Done.
```

landing__outcome	COUNT
No attempt	11
Failure (drone ship)	5
Success (drone ship)	5
Success (ground pad)	5
Controlled (ocean)	4
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis



Launch sites are in very close proximity to coastlines. Florida especially is much close to the equator than other launch sites.

Marking Launch Sites on a Map

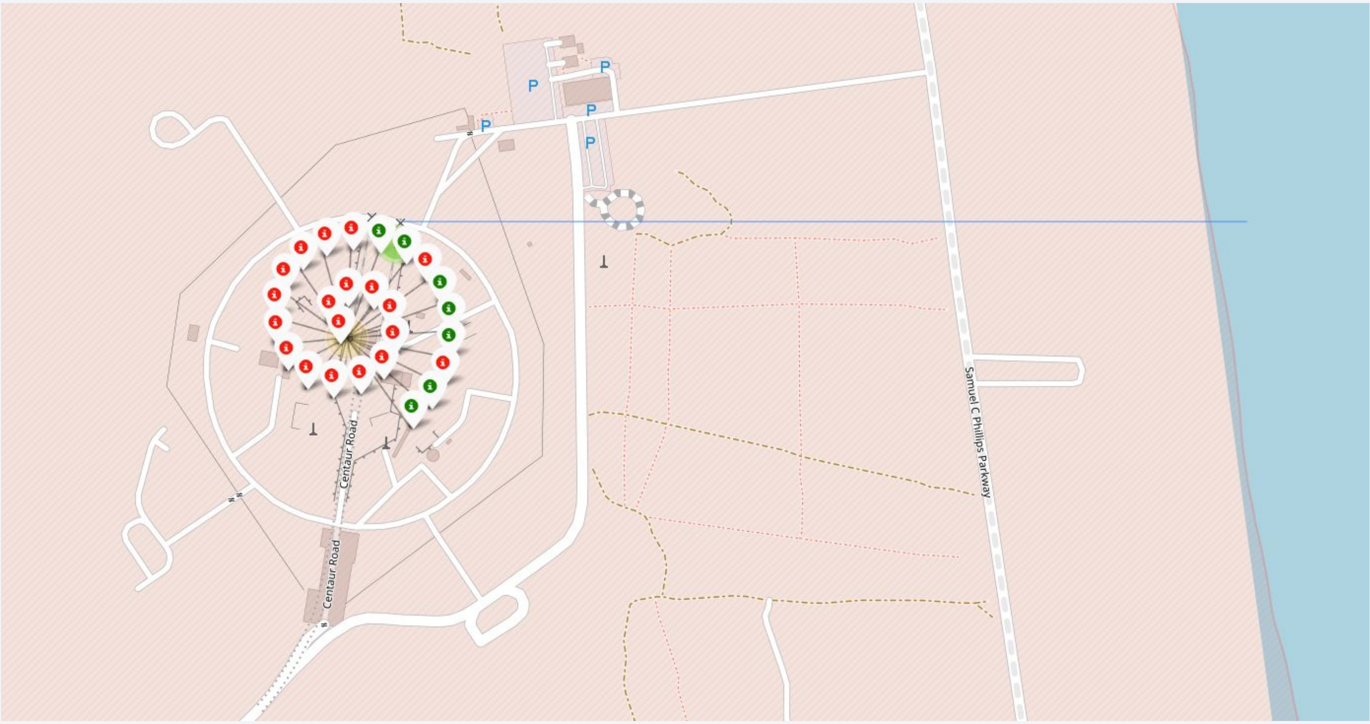
Color-Labeled Launch Markers

Green markers shown here on the right indicate successful launches, whereas red markers show unsuccessful launches.

Marking distances between selected points

Exploring our interactive map to show selected launch sites and their proximities to railways, highways, coastlines, with distance calculated and displayed.

In the example here, the blue line indicates distance from the launch site to the coastline.

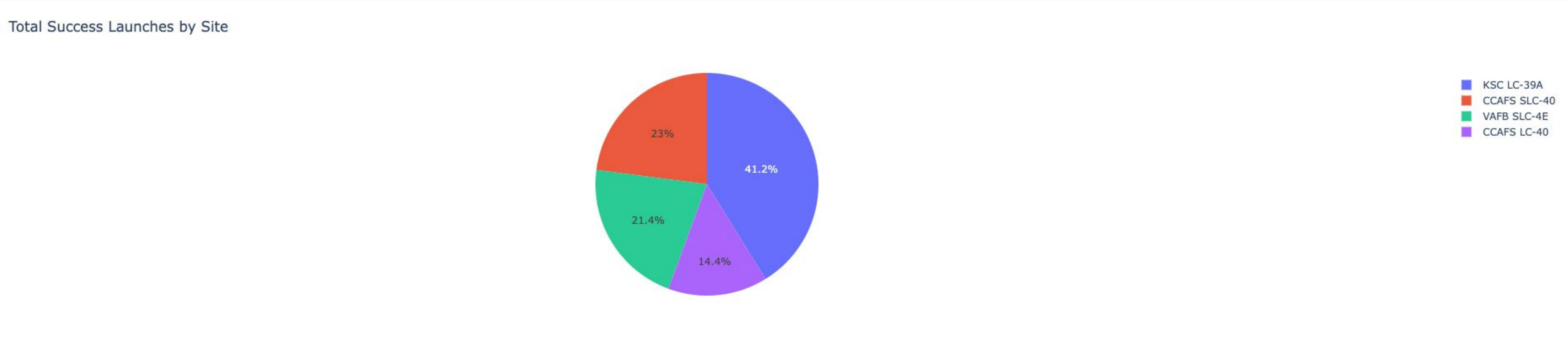


Section 4

Building a Dashboard with Plotly Dash

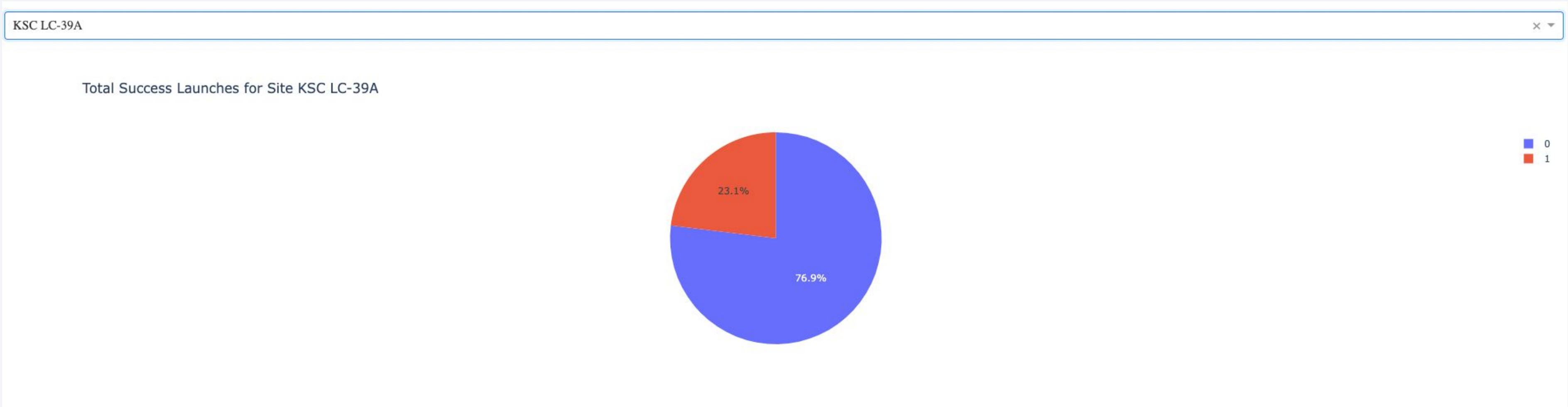
Successful Launches by Launch Site

This diagram shows the screenshot of launch success counts for all sites, in a pie chart format



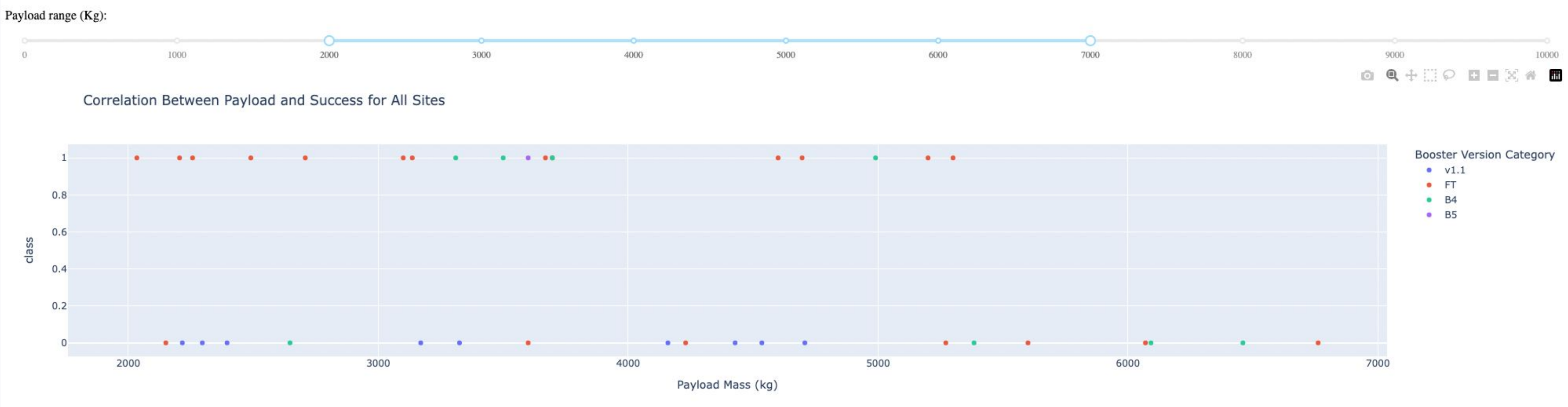
Highest Launch Success Ratio

Toggling launch sites in the dropdown menu we can see that KSC LC-39A has the highest success ratio among all launch sites.



Payload Mass vs Launch Outcome (all sites)

Scatterplots on the right show success rates by Payload Mass, and Booster Version Category. The highest success rates appear to be between 2k and 6k for payload mass.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

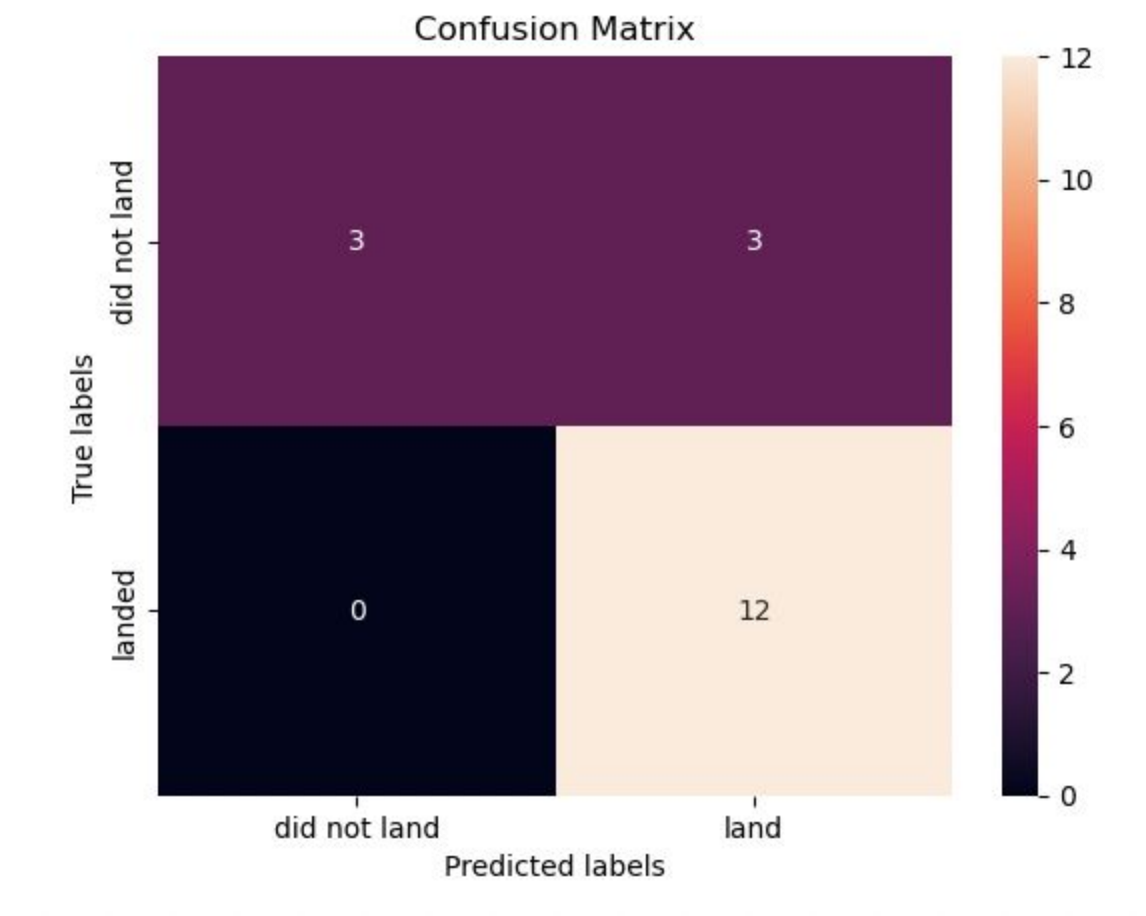
Using Jaccard_Score, and F1_Score, we can calculate the classification accuracy of LogReg, SVM, Tree, and K Nearest Neighbors models.

In this case, all models showed the same accuracy rate of 83%.

Because all accuracy rates are the same, we're unable to determine the best model to use.

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

Confusion Matrix



This confusion model shows the how all models performed in the accuracy tests.

Our models predicted 3 unsuccessful landings and 3 successful landings when true label was unsuccessful.

Overall our models predict successful landings.

Conclusions

- We were able to successfully predict first stage successful landings for future Space X launches
- We were able to create a machine learning model with 83% accuracy across multiple algorithms.
- Were able to identify strong correlations between Payload Mass, Orbits, Booster Versions and Landing Outcomes in order to predict future launch success rates.
- We were able to create a useful, interactive dashboard for visualizing collected data from the Space X dataset.

Thank You!