IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Roger Greer

17 June 2022

Outline

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

Executive Summary

- Summary of methodologies utilized
  - Data acquisition through API and web scraping
  - Data wrangling and preprocessing
  - Exploratory Data Analysis with SQL and data visualization
  - Interactive data visualization with Folium
  - Machine Learning model development and predictive analytics
- Summary of all results
  - Findings from Exploratory Data Analysis
  - Interactive analytic graphs and maps
  - Predictive analytics from Machine Learning models

Introduction

- SpaceX has disrupted the space industry by offering a launch vehicle, Falcon 9, at a cost per launch as low as 62 million dollars per launch. This contrasts with other providers with a cost of 165 million dollar per launch. Most of this saving is due to the reusable first stage of the launch vehicle which is callable of landing itself. As a data scientist of a startup rivaling SpaceX, the goal of this project is to create a machine learning pipeline to predict the landing outcome of the first stage which is important in identifying the right price to bid against SpaceX for a rocket launch.
- Problems
  - Data acquisition
  - What factors influence success or failure of first stage landing
  - How do these factors affect the outcome of a landing
  - What will increase the probability of a successful landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Launch data is publicly available through a REST API and web scraping
- Perform data wrangling
  - Data was scaled and encoded in preparation for model evaluation
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four models were developed using optimal parameters and evaluated using a confusion matrices and accuracy assessments

Data Collection

- Datasets were acquired from publicly available sources through REST API and web scraping from spacexdata.com and Wikipedia.
- The REST API used a get request for the launch, sites, and payloads data. The JSON response content was decoded and transformed into a Pandas dataframe using json_normalize(). The data was cleaned, checked for missing values and replaced with mean values.
- The HTML for Falcon 9 launch data was downloaded with Beautiful Soup and parsed into a Pandas dataframe. Further processing removed missing values and special characters.

Data Collection – SpaceX API

- GET request for rocket, launch, and payload data
- Create dataframe from response JSON with json_normalize
- Clean the data and fill missing values with appropriate method

Ref: https://github.com/rgreer2311/IBM-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

Beautiful Soup requests the page HTML and downloads a copy.

Htmp.parser parses the HTML enabling attribute queries for the table data.

Relevant attributes are put into a Pandas dataframe. Further cleaning removes NaN values.

Ref: https://github.com/rgreer2311/IBM-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

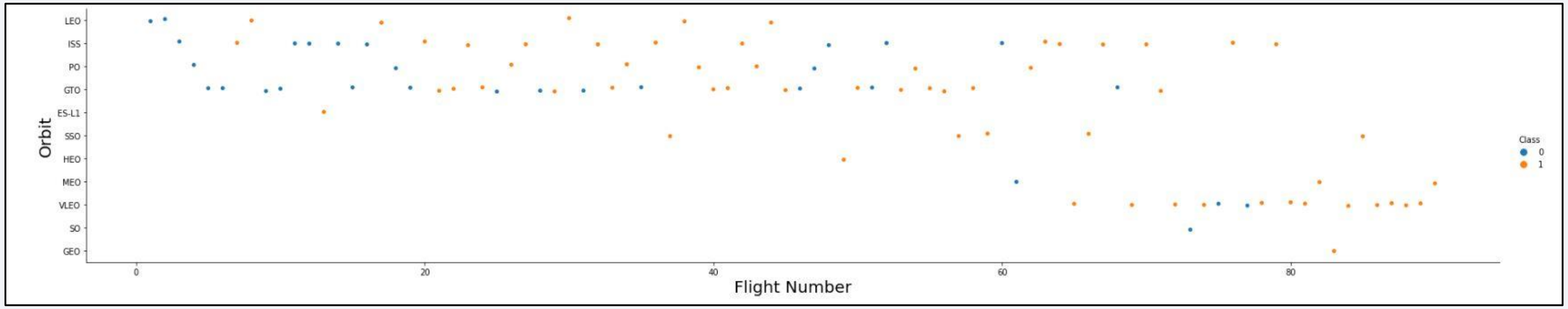
Identify missing and NULL values and types

Generated count summaries for launch, orbit type, and mission outcome

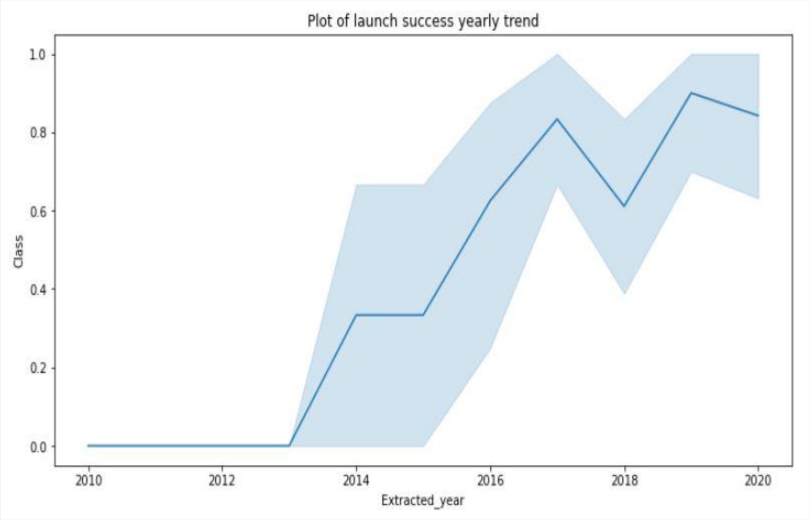
Created class labels for landing outcome for further analysis

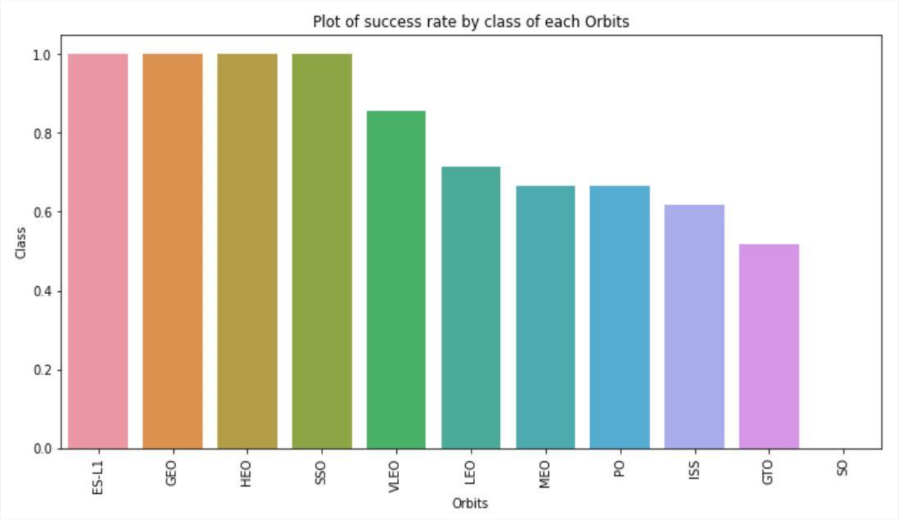
Ref: https://github.com/rgreer2311/IBM-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization



Generated scatter plots to get an idea of the relationships and dependencies between variables to identify patterns.





Line and bar graphs indicate trends and change over time.

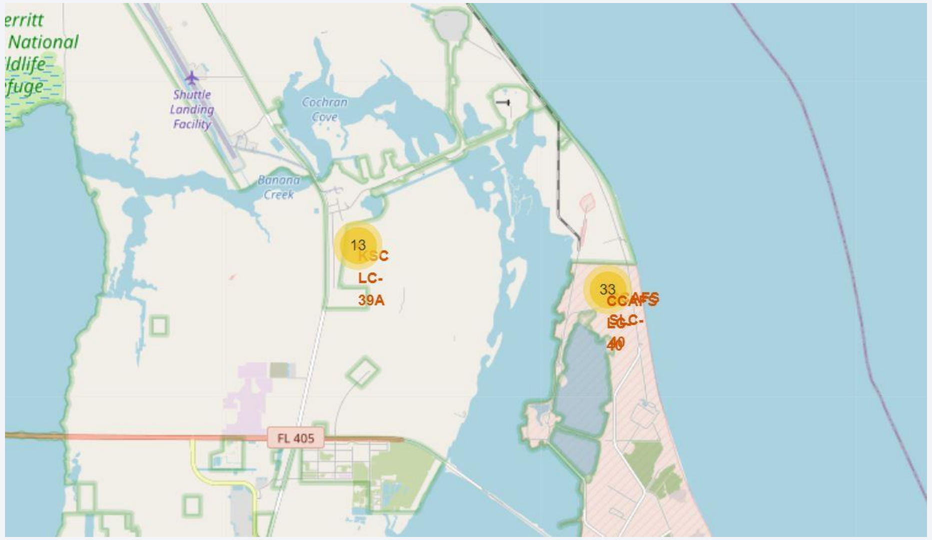
Ref: https://github.com/rgreer2311/IBM-Data-Science-Capstone/blob/main/EDA%20with%20Viz.ipynb

EDA with SQL

- SQL queries enable additional counts and aggregations to get a further understanding of variable relationships:
  - The names of the launch sites
  - The 5 records where launch sites was Cape Canaveral
  - Total payload mass carried by NASA boosters
  - The average payload mass carried by booster F9 v1.1
  - The date of first successful landing outcome on a ground pad was achieved
  - Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 kg
  - Total number of success and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - The failed landing outcomes in drone ship, booster versions, and launch sites names for in year 2015
  - A ranking of the number of landing outcomes between 2010-06-04 and 2017-03-20

Ref: https://github.com/rgreer2311/IBM-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

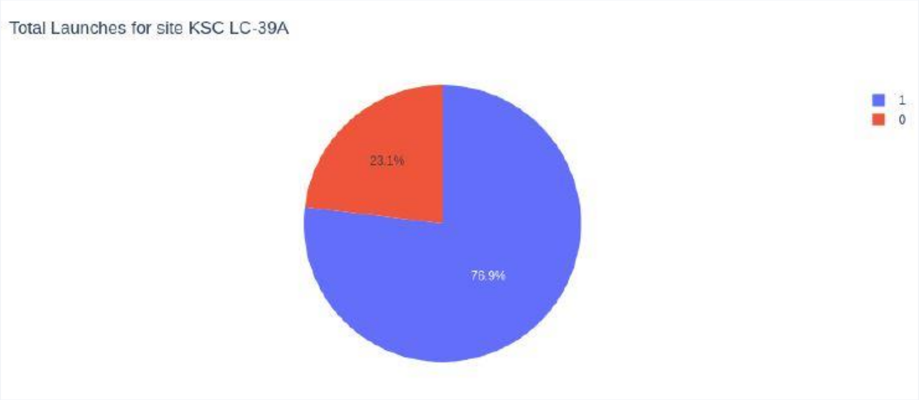
Build an Interactive Map with Folium

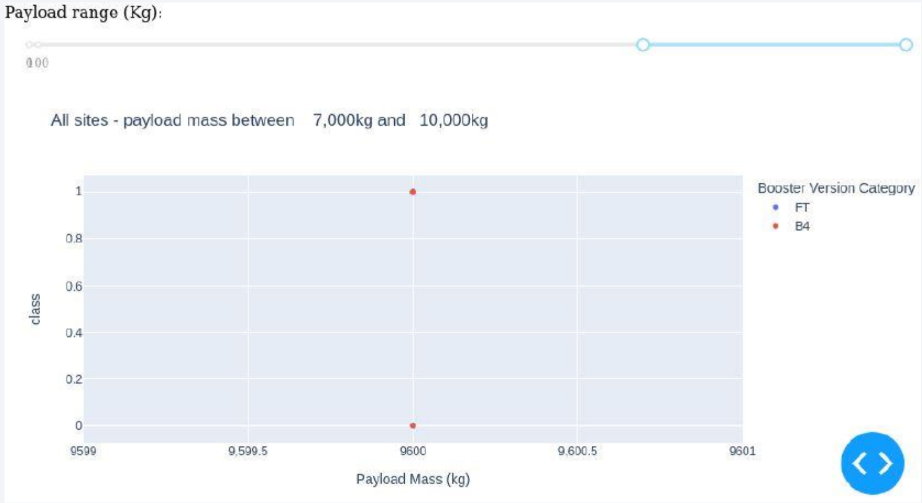


- Interactive maps were created with launch site markers.
- Marker attributes indicate site name, and success/failure in green or red.
- Distance calculations with Haversine’s formula from locations of interest allow spatial modeling of the launch sites, i.e., distance to other sites, oceans, cities, which indicate spatial patterns.

Ref: https://github.com/rgreer2311/IBM-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

Build a Dashboard with Plotly Dash





- An interactive dashboard allows us to share results with a team and stake holders.
- Dashboards allow non-technical users to explore the dataset

Ref: https://github.com/rgreer2311/IBM-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics.ipynb

Predictive Analysis (Classification)

Model Building
- Load data using Numpy and Pandas
- Split the data into test and training sets

Model Evaluation
- Perform accuracy assessment for each model
- Get best hyperparameters

Model Tuning
- Engineer features to tune algorithm

Choose Model
- Select the model with the highest accuracy score

Ref: https://github.com/rgreer2311/IBM-Data-Science-Capstone/blob/main/Machine%20Learning.ipynb

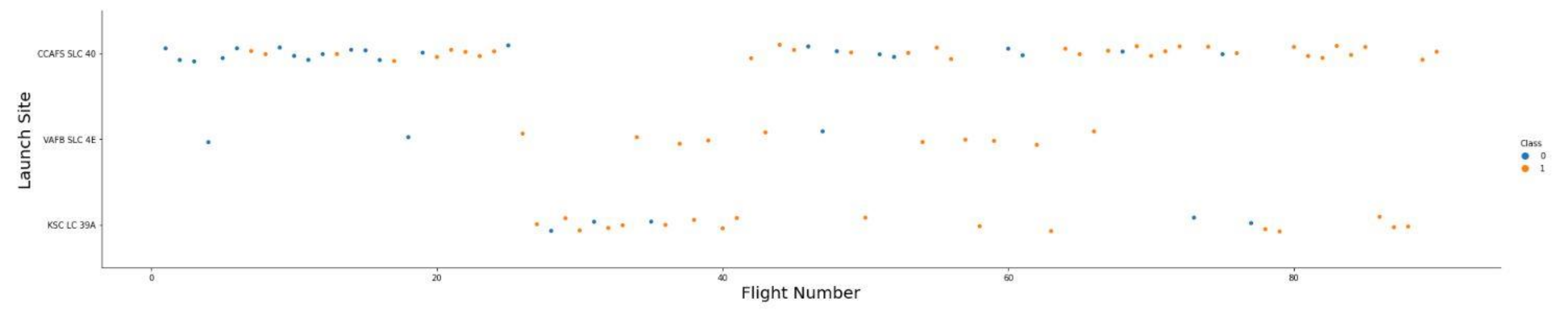
Results

- Exploratory data analysis results
  - Revealed relationships between variables enabling the use of predictive modeling
- Interactive analytics demo in screenshots
  - Indicated spatial patterns in launch sites
- Predictive analysis
  - Indicated that the Decision Tree Classifier was the best having the best accuracy

Section 2

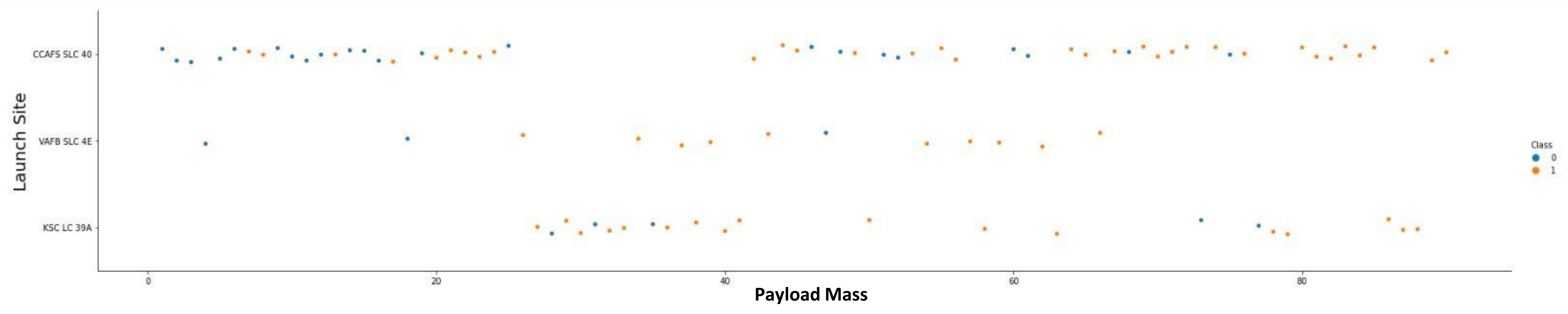
Insights drawn from EDA

Flight Number vs. Launch Site



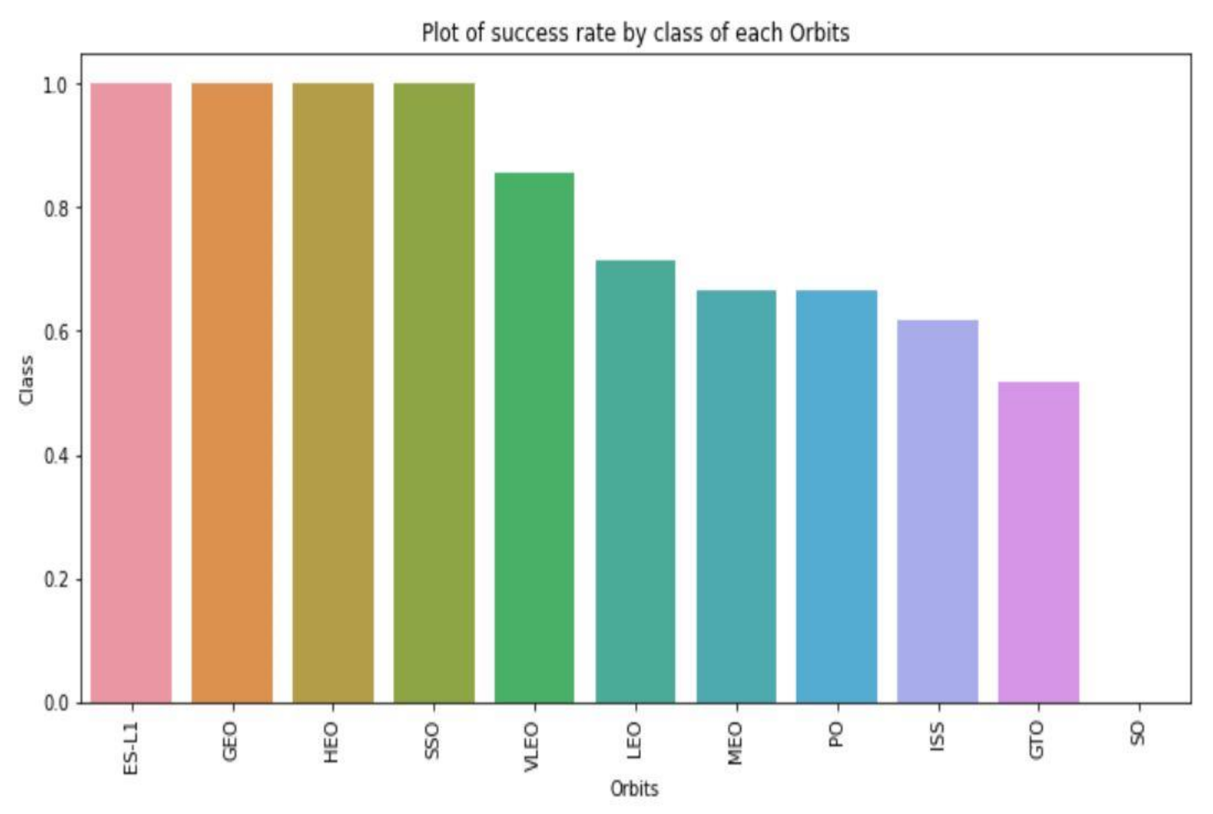
- This scatterplot indicates that the more launches the greater success with Cape Canaveral showing the most variation.

Payload vs. Launch Site



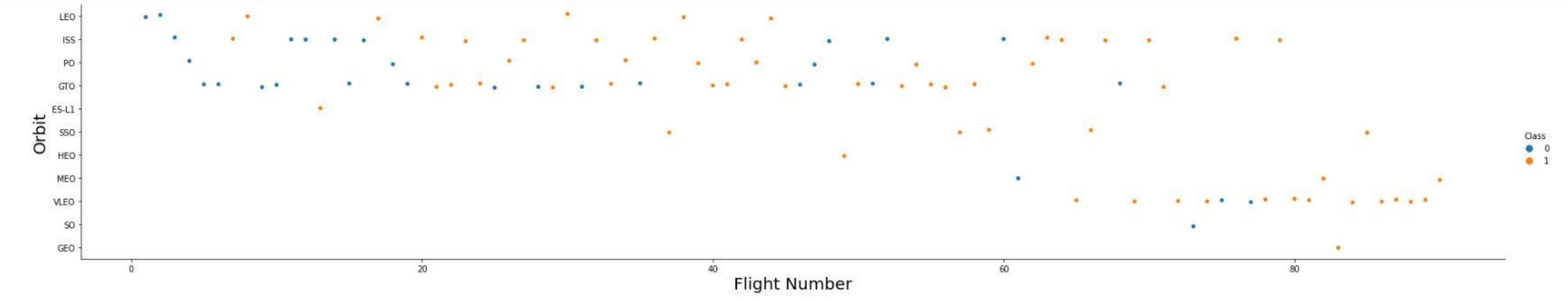
- This scatterplot indicates Cape Canaveral launched the heavier payloads.

Success Rate vs. Orbit Type



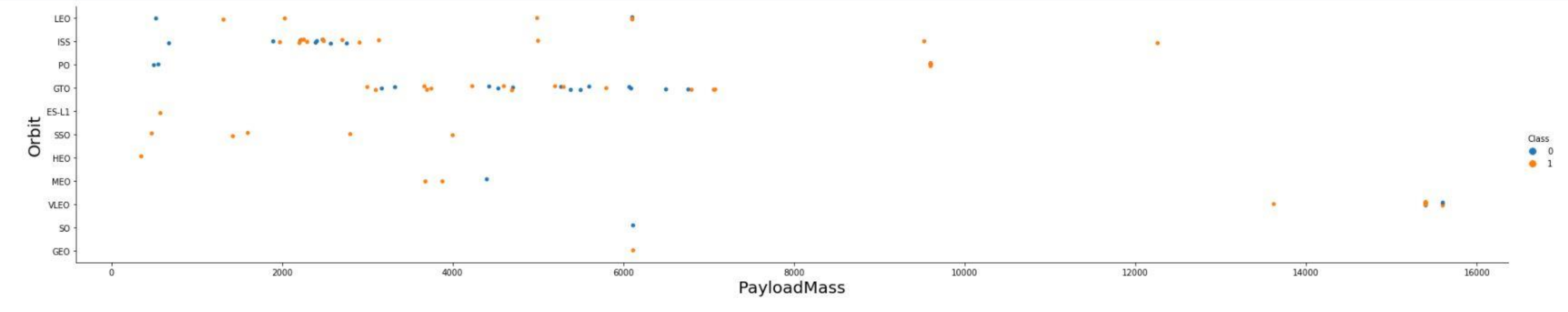
- The bar chart would seem to indicate that certain orbits have more successful outcomes than others.
- However, the number of launches at the launch sites are not proportional to each other, and success rate is not tied to launch site.

Flight Number vs. Orbit Type



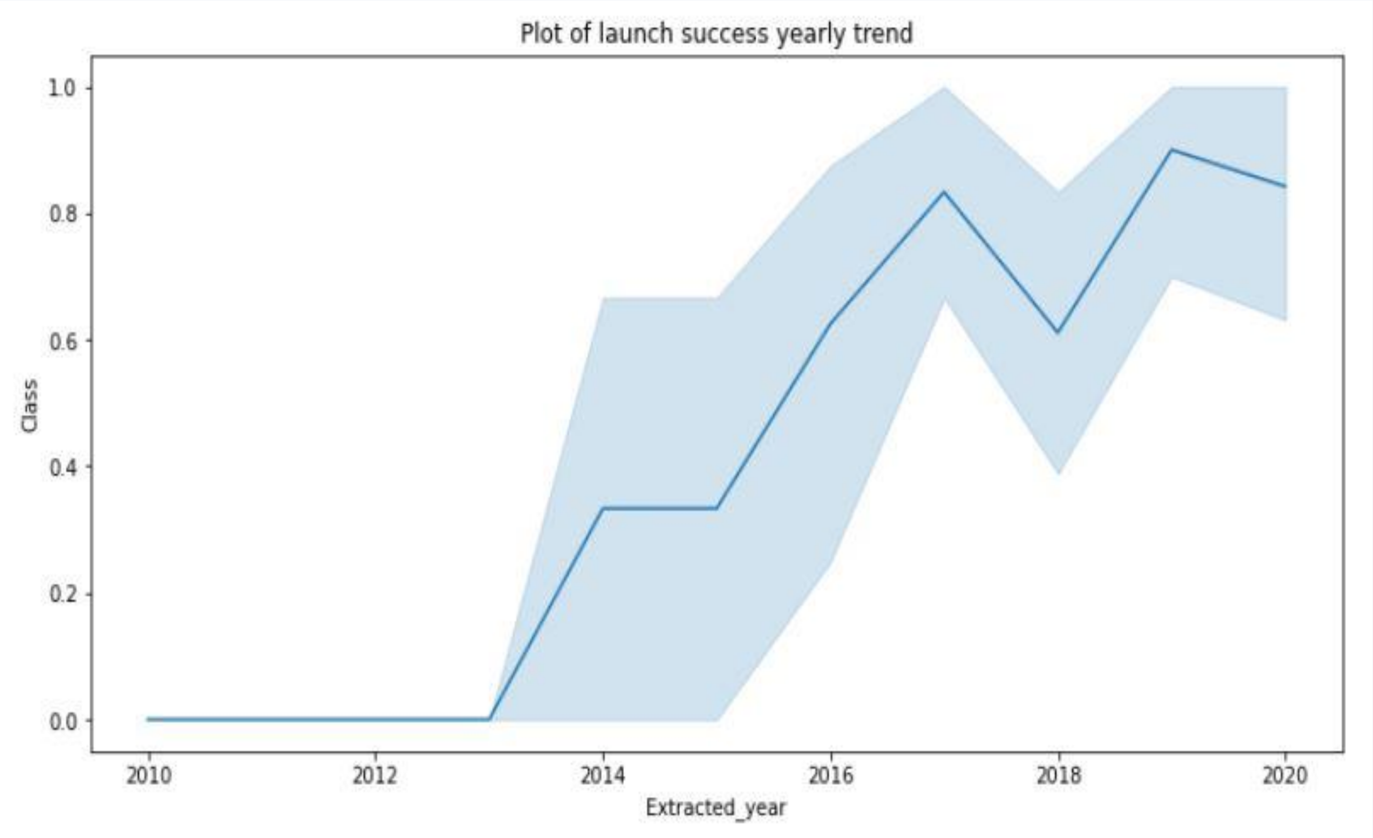
- The scatterplot indicates that the more launches per particular orbit, the greater success rate.
- More data is needed to access orbits with fewer launches.

Payload vs. Orbit Type



- As previously seen, heavier payloads have a higher success rate.
- The scatterplot indicates that it has a positive impact on some orbits, but more data is needed to draw a strong conclusion.

Launch Success Yearly Trend



- There is a definite upward trend in success rate from 2013 to 2020.
- This is likely due to an increase in launches and improved prototypes.

All Launch Site Names

```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

- Using relational database queries, we can identify the distinct launch sites.

Launch Site Names Begin with 'CCA'

```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Launches at Cape Canaveral

Total Payload Mass

```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

	total_payloadmass
0	45596

- The total payload carried by NASA boosters from NASA

Average Payload Mass by F9 v1.1

```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

	avg_payloadmass
0	2928.4

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

First Successful Ground Landing Date

```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

	firstsuccessfull_landing_date
0	2015-12-22

- The first successful landing on a ground pad was on December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

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

Total Number of Successful and Failure Mission Outcomes

```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

```
The total number of successful mission outcome is:
```

	successoutcome
0	100

```
The total number of failed mission outcome is:
```

	failureoutcome
0	1

- Successful and unsuccessful mission outcomes

Boosters Carried Maximum Payload

```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

- The boosters which carried the maximum payload

2015 Launch Records

```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- The failed drone ship landings in 2015

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

```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

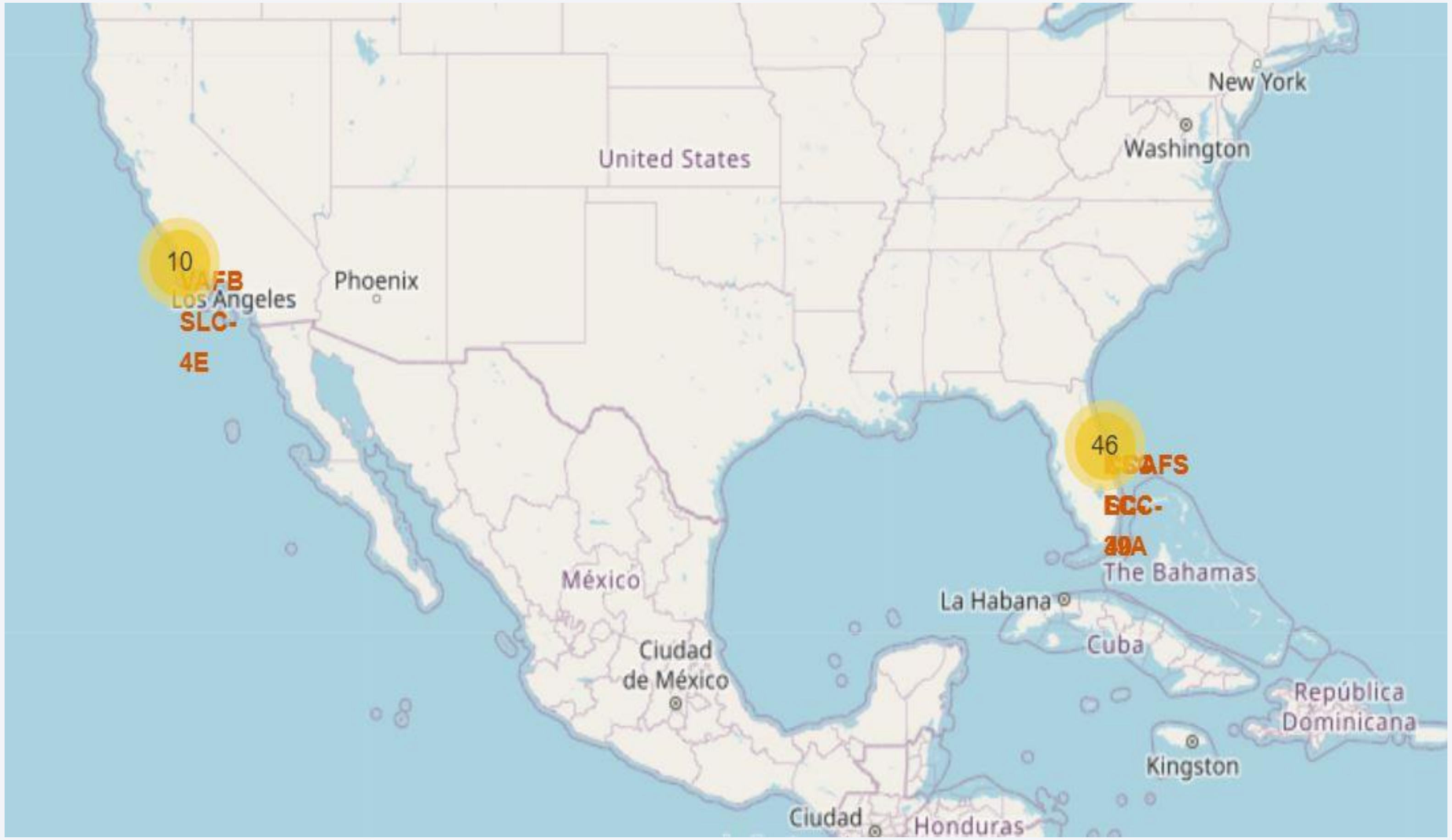
	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

- A ranking of the number of of landing outcomes at landing sites between the 2010-06-04 and 2017-03-20

Section 3

Launch Sites Proximities Analysis

All Launch Sites



- SPACEX launch sites in the contiguous United States

Launch Site by Outcome

- Green and red markers indicate launch success and failure respectively.

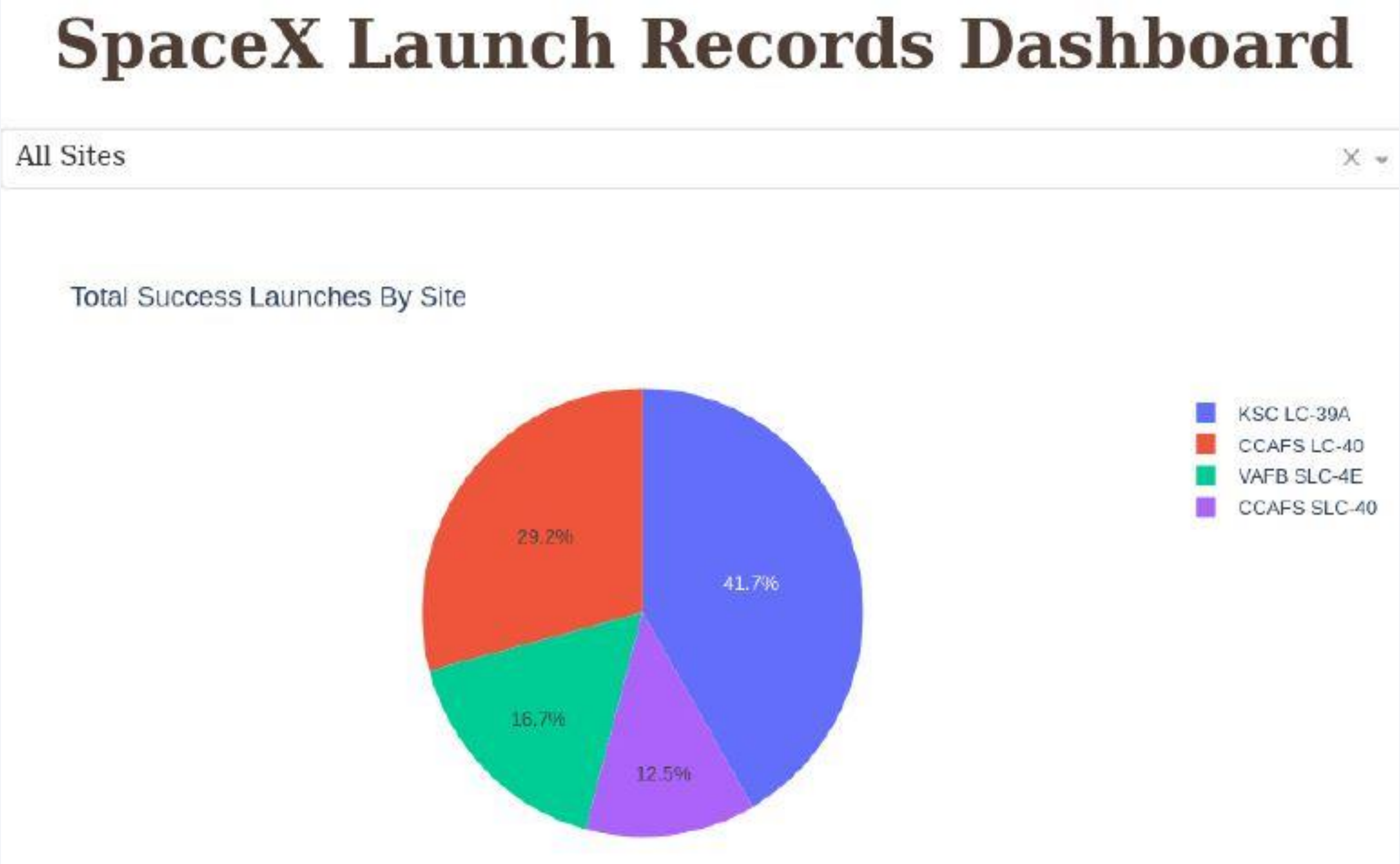
Launch Site Proximity

- Launch sites are located on the coast away from major highways and cities to mitigate risk to people.

Section 4

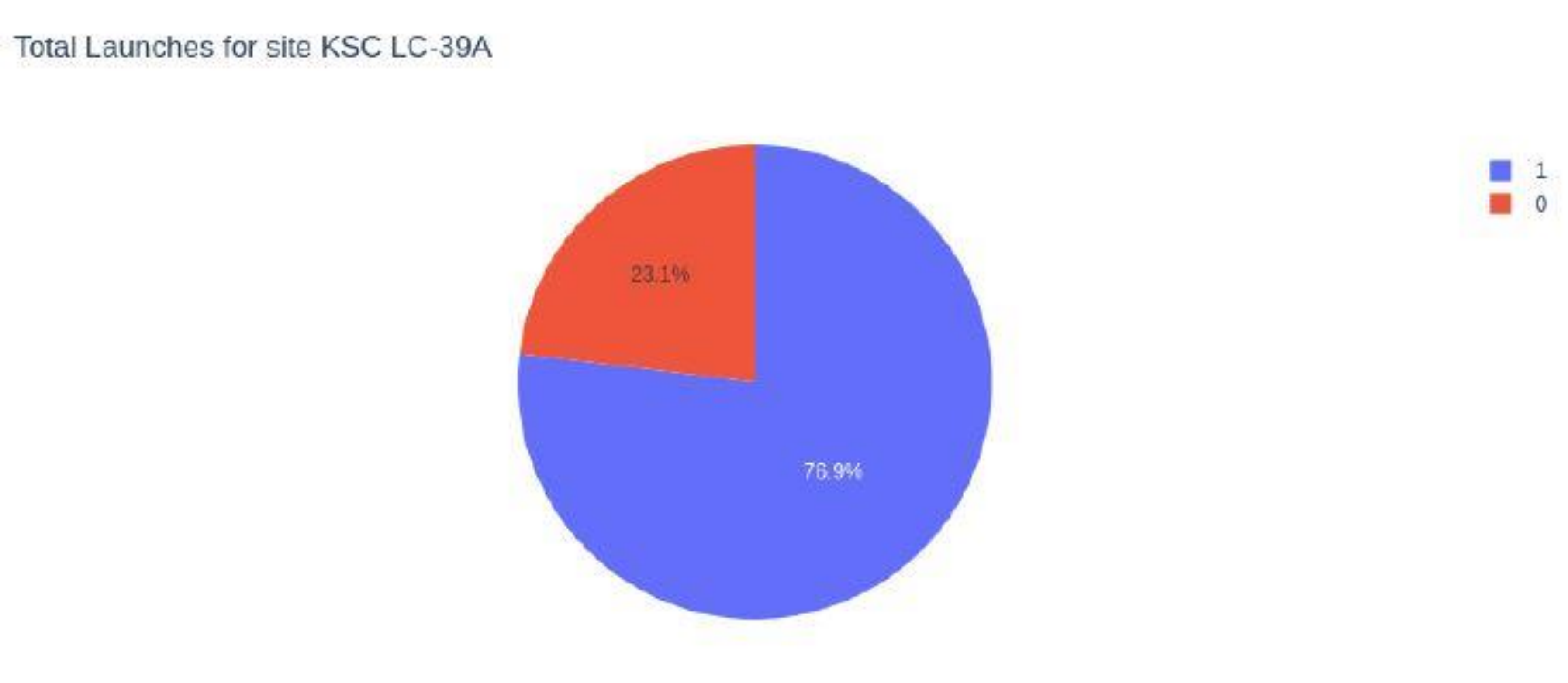
Build a Dashboard with Plotly Dash

Launch Success by Site



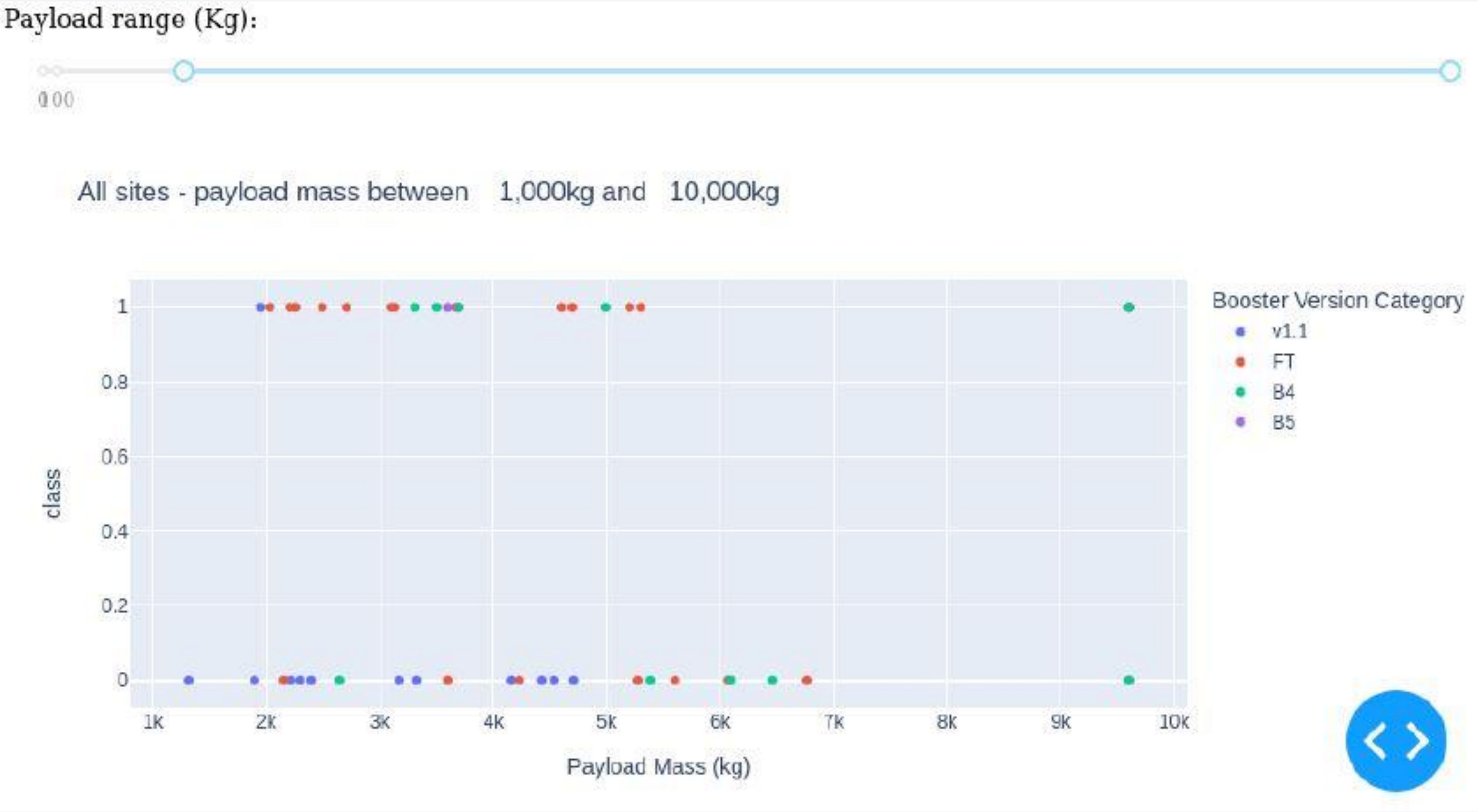
- The pie chart indicates that LSC LC-39A was the most successful

KSC LC-39A Launch Sites



- KSC LC-39A had a 76.9% success rate and a 23.1% failure rate

Payload by Outcome



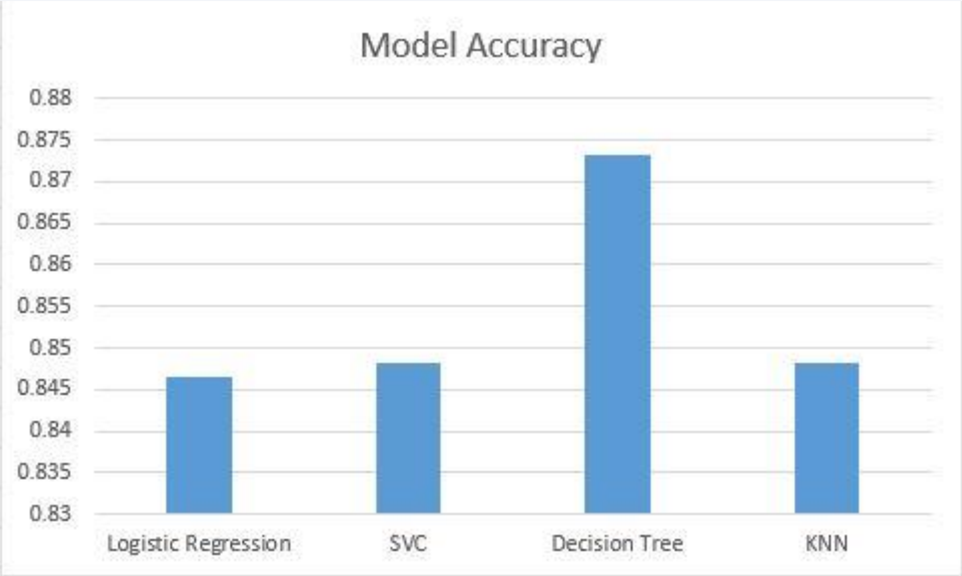
- Lighter payloads had a higher success rate than heavier payloads

Section 5

Predictive Analysis (Classification)

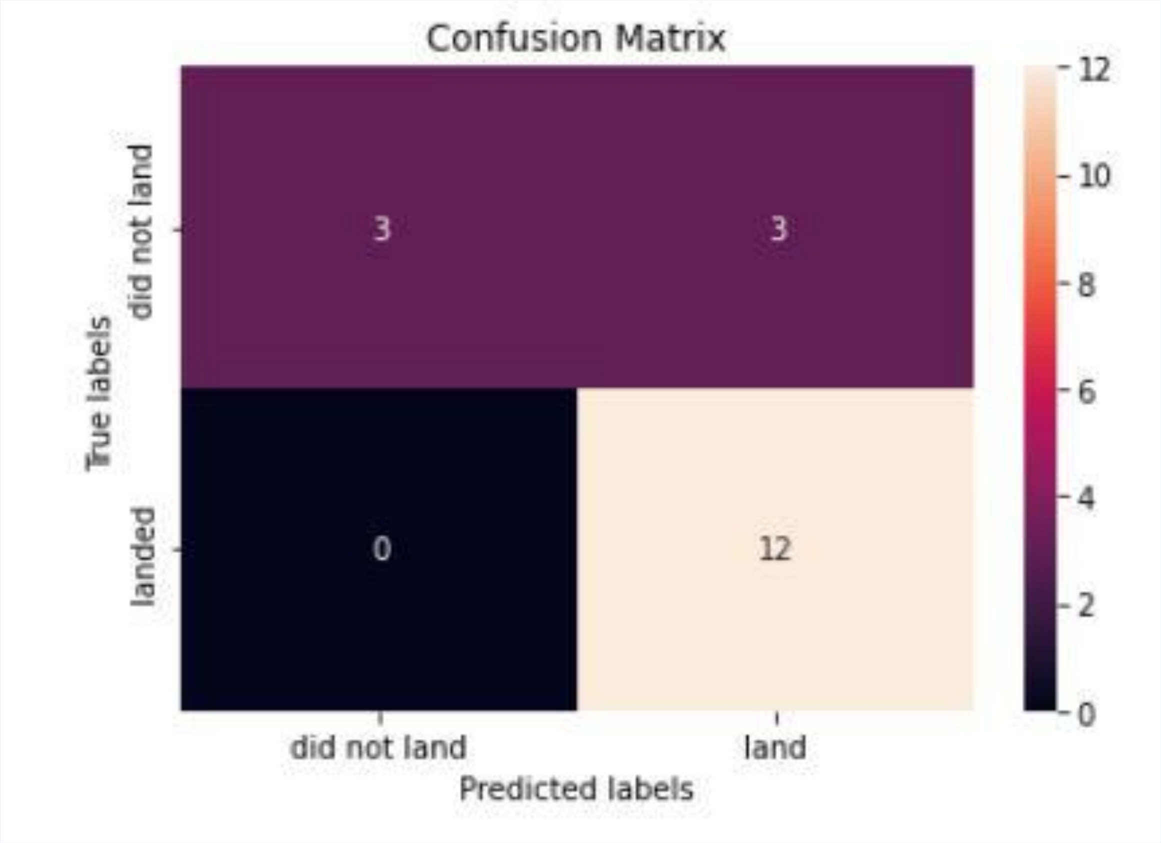
Classification Accuracy

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5,
'splitter': 'random'}
```



- The best performing algorithm was the Decision Tree classifier

Confusion Matrix



- The confusion Matrix indicated the successful a large number of true positives versus true negatives.

Conclusions

- The most successful launch site was KSC LC-39A, Kennedy Space Center.
- Payload mass influences launch success.
- Launch sites on the southern coasts away from population centers are best in terms of damage mitigation.
- Launch successes increased almost exponentially in 2013.
- The Decision Tree classifier can accurately predict successful launches.

Thank you!