IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mahak Rajvanshi

16 May 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  While other companies charge upwards of $165 million for each launch, Space X promotes Falcon 9 missions on its website at a cost of 62 million dollars. A large part of the savings comes from Space X's ability to reuse the first stage. The cost of a launch may be determined if we know if the first stage will land. If another business wishes to compete with Space X for a rocket launch, this information can be used. Predicting whether the first stage will land successfully is the project's purpose.

- Problems you want to find answers

  - In the event of a successful landing, what are the determining factors?
  - A rocket's ability to land safely depends on the interplay of several factors..
  - A successful landing program requires what operational parameters to be in place..

Section 1

Methodology

Methodology

Executive Summary

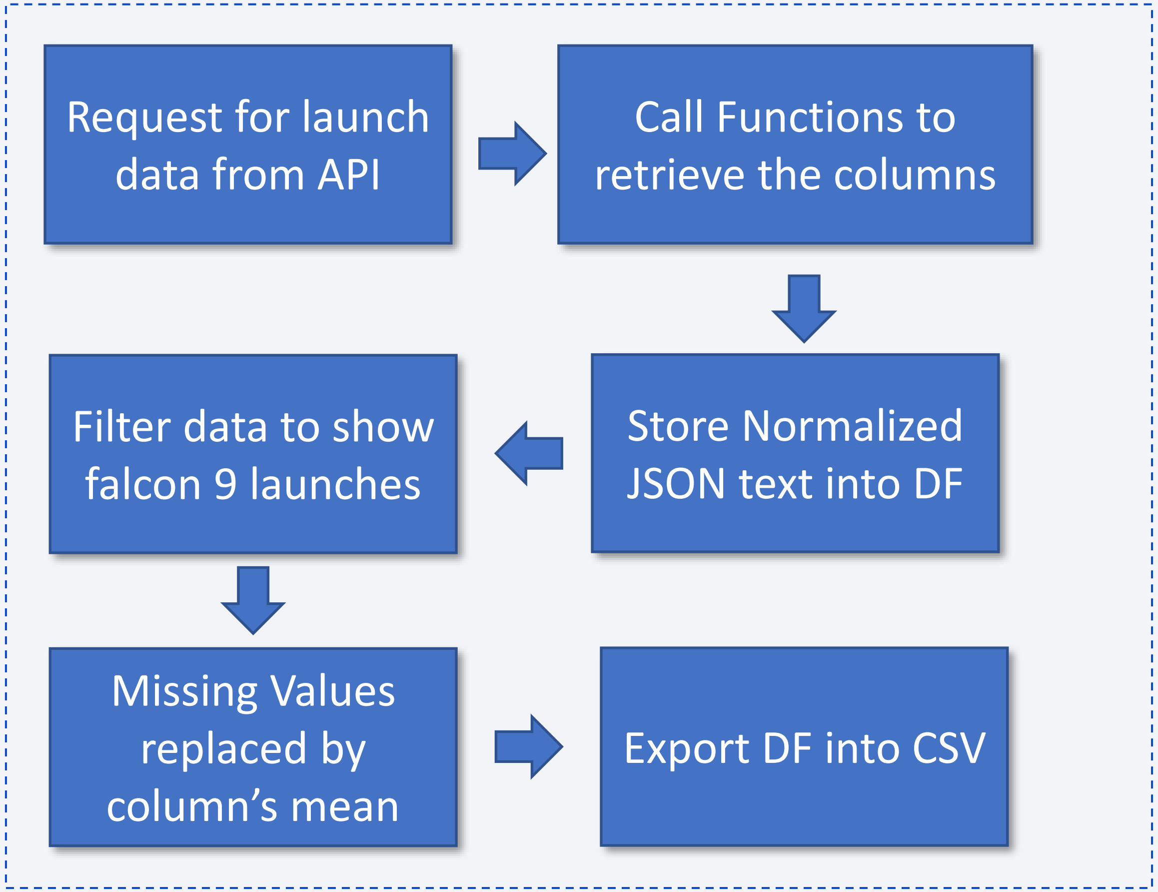
- Data collection methodology:
  - SpaceX API and Wikipedia site scraping were used to get data.
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Visualized several columns to show the relationships.
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models
  - Evaluated the most accurate and predictive models

Data Collection

- The data was collected using various methods as below:
  - The SpaceX API's get request method was used to gather the data.
  - We then used the.json() function call to decode the response content into a JSON object, then the.json normalize method to transform it into a pandas dataframe ().
  - After that, we went through the data and made sure there were no missing entries..
  - Additionally, we used BeautifulSoup to scrape Wikipedia for Falcon 9 launch data..
  - Extract the launch records as HTML tables, parse the tables and transform them to pandas dataframe for further research..

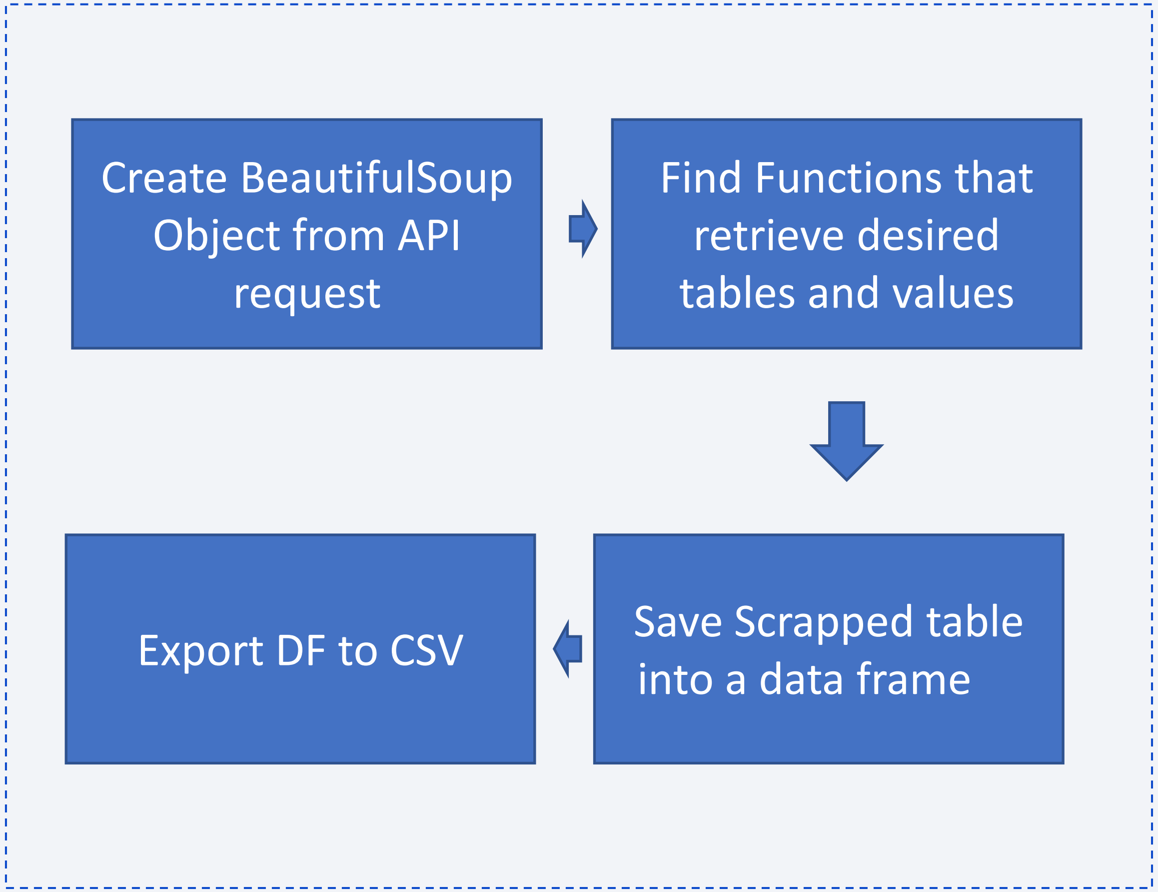
Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.
- Github Link: https://github.com/mrajvanshi/IBM_Capstone_SpaceY_Data_Science/blob/main/SpaceY%20Data%20Collection%20API.ipynb



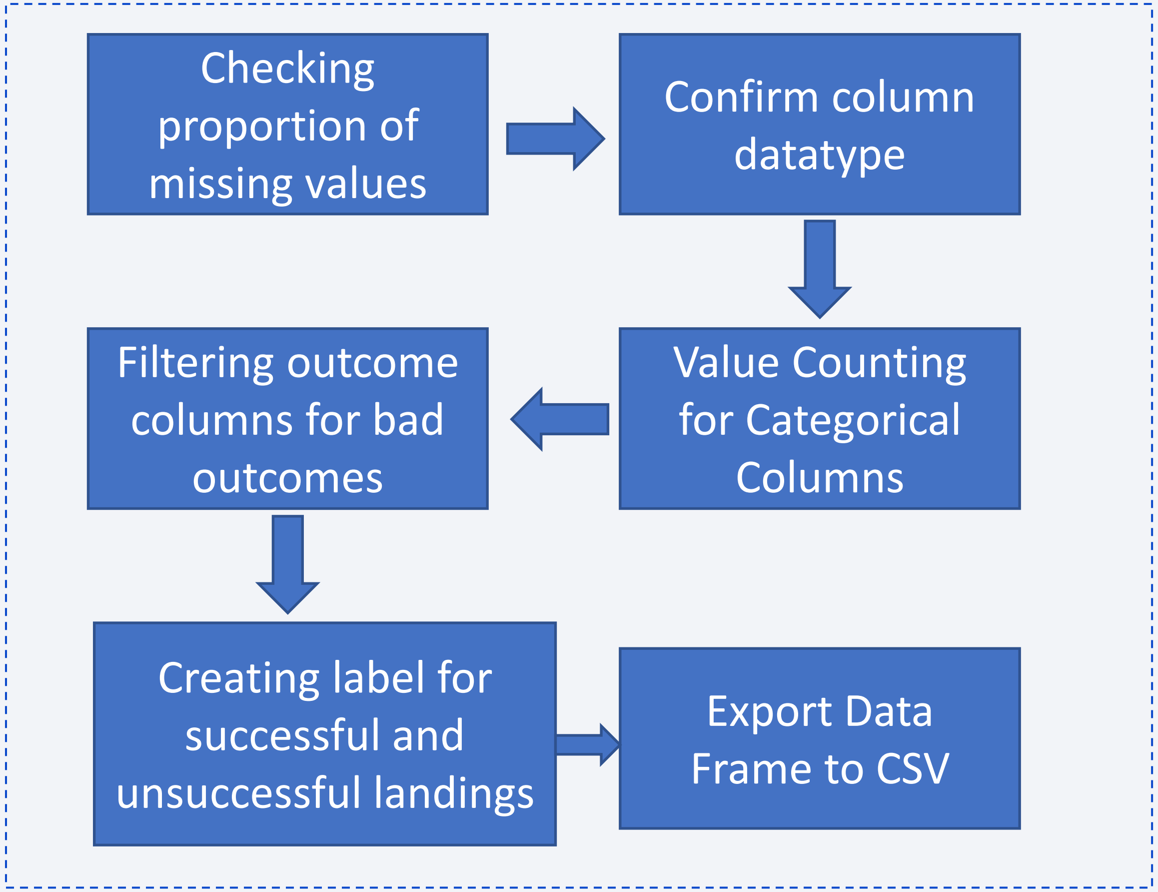
Data Collection - Scraping

- With BeautifulSoup, we used web scraping to scrape Falcon 9 launch data.
- We analyzed the data and created a pandas dataframe from it..
- Github Link: https://github.com/mrajvanshi/IBM_Capstone_SpaceY_Data_Science/blob/main/SpaceY%20Data%20Collection%20With%20Web%20Scraping.ipynb



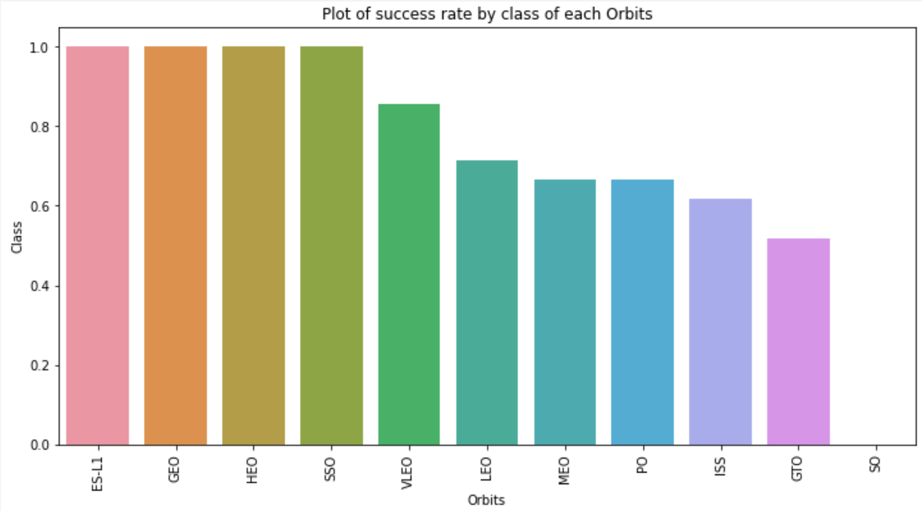
Data Wrangling

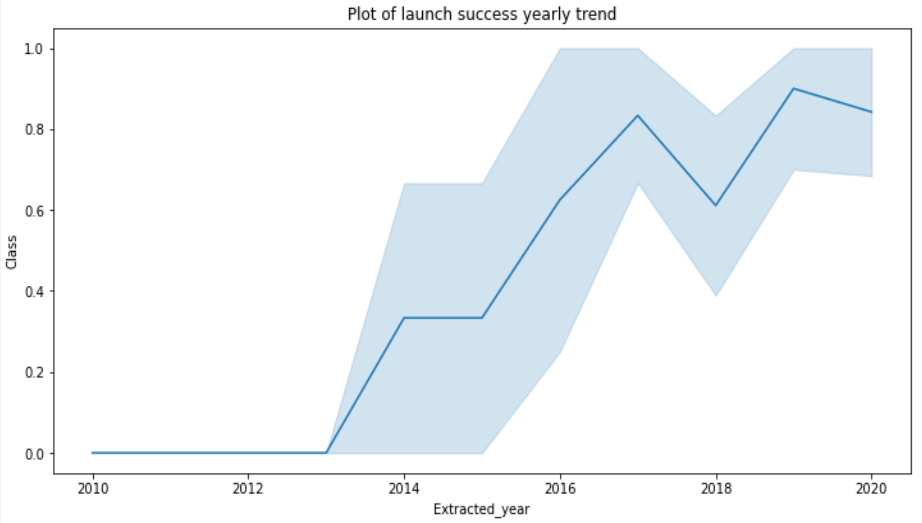
- The training labels were generated by exploratory data analysis.
- The number of launches at each location, as well as the number and frequency of each orbit, were analyzed.
- The landing result label was generated from the outcome column and the data was sent to a csv file.
- Github Link: https://github.com/mrajvanshi/IBM_Capstone_SpaceY_Data_Science/blob/main/SpaceY%20Data%20Wrangling.ipynb



EDA with Data Visualization

- We analyzed the data by displaying the association between flight number and launch site, payload and launch site, success rate of each orbit type, flight number and orbit type, and the launch success yearly trend.
- We plotted scatter plots, bar plots and line plots.
- Github Link: https://github.com/mrajvanshi/IBM_Capstone_SpaceY_Data_Science/blob/main/SpaceY%20EDA%20with%20Data%20Visualization.ipynb
- Dash App Github Link: https://github.com/mrajvanshi/IBM_Capstone_SpaceY_Data_Science/blob/main/SpaceY%20Dash%20App.py



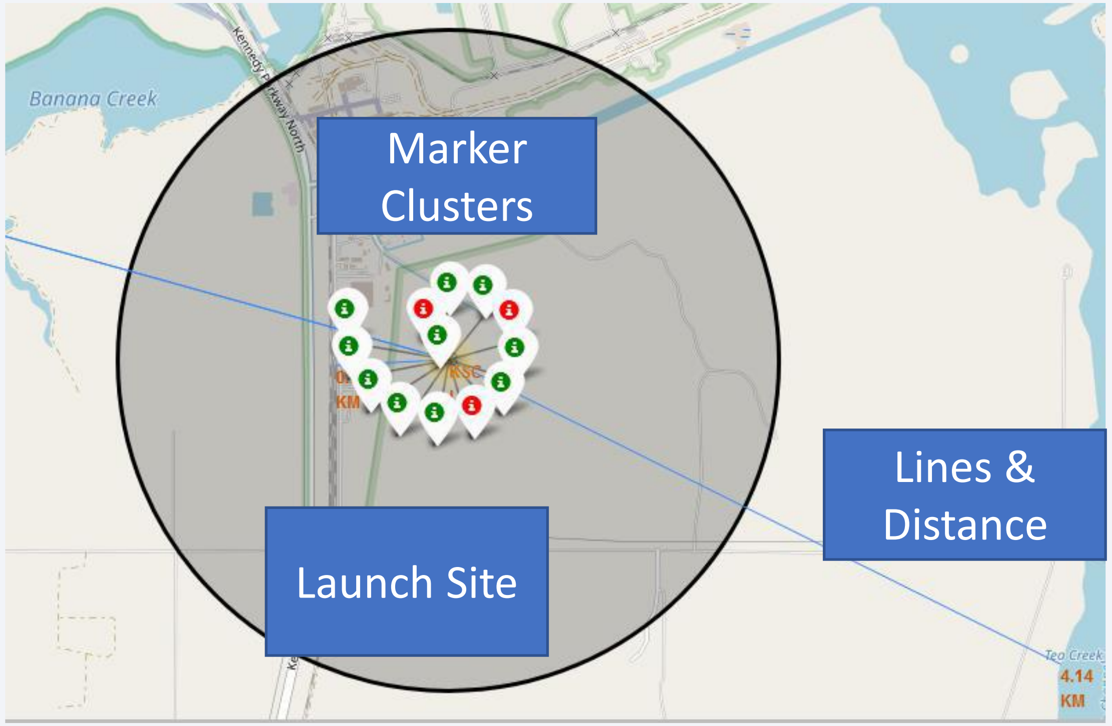


EDA with SQL

- We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:
  - Display the names of unique launch sites in the space mission.
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display the average payload mass carried by booster version F9 v1.1
  - List the total number of successful and failure mission outcomes
  - List the failed landing outcomes in drone ship, their booster version and launch site names.
  - Display 5 records where launch sites begin with the string 'CCA' and many others
- Github link to the notebook is: https://github.com/mrajvanshi/IBM_Capstone_SpaceY_Data_Science/blob/main/SpaceY%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- All launch sites were identified and map items such as markers, circles, and lines were added to the folium map to indicate whether a launch was successful or unsuccessful.
- Class 0 and 1 were allocated to the two possible outcomes of a feature launch: failure or success. For example, a score of 0 indicates failure, whereas a score of 1 indicates success.
- Color-labeled marker clusters helped us identify which launch sites have an above-average success rate.
- We estimated the distances between a launch site and its immediate surroundings. For example, we answered a few questions:
  - Are launch locations located near railways, highways, and the seashore, or are they isolated?
  - Is there a set distance between launch locations and cities?
- Github Link: https://github.com/mrajvanshi/IBM_Capstone_SpaceY_Data_Science/blob/main/SpaceY%20Launch%20Site%20Location.ipynb



Build a Dashboard with Plotly Dash

- Plotly dash was used to create an interactive dashboard.
- Pie charts depicting the overall number of launches by a specific site were created.
- In order to better understand the link between output and payload mass (Kg), we created a scatter plot.
- The highest success rate is achieved by KSC LC 39A.
- Github Link: https://github.com/mrajvanshi/IBM_Capstone_SpaceY_Data_Science/blob/main/SpaceY%20Dash%20App.py

Predictive Analysis (Classification)

- We used numpy and pandas to load the data, transform it, and divide it into training and testing sets.
- We used GridSearchCV to build and tune a variety of machine learning models and parameters.
- By employing feature engineering and algorithmic tuning, we enhanced our model's accuracy.
- The best classification model was discovered.
- Github Link: https://github.com/mrajvanshi/IBM_Capstone_SpaceY_Data_Science/blob/main/SpaceY%20Machine%20Learning%20Prediction.ipynb

Results

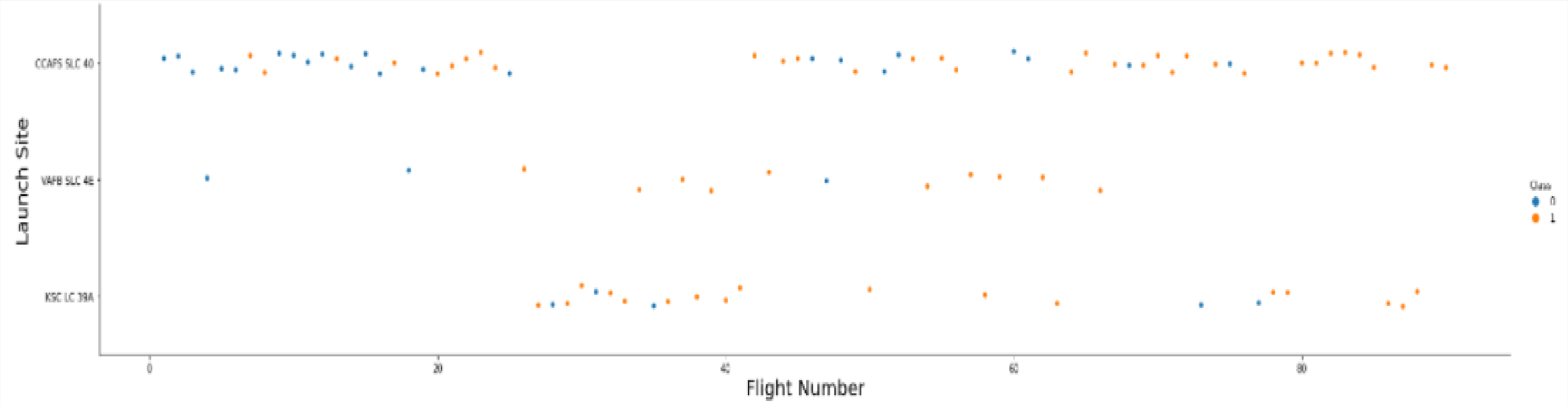
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

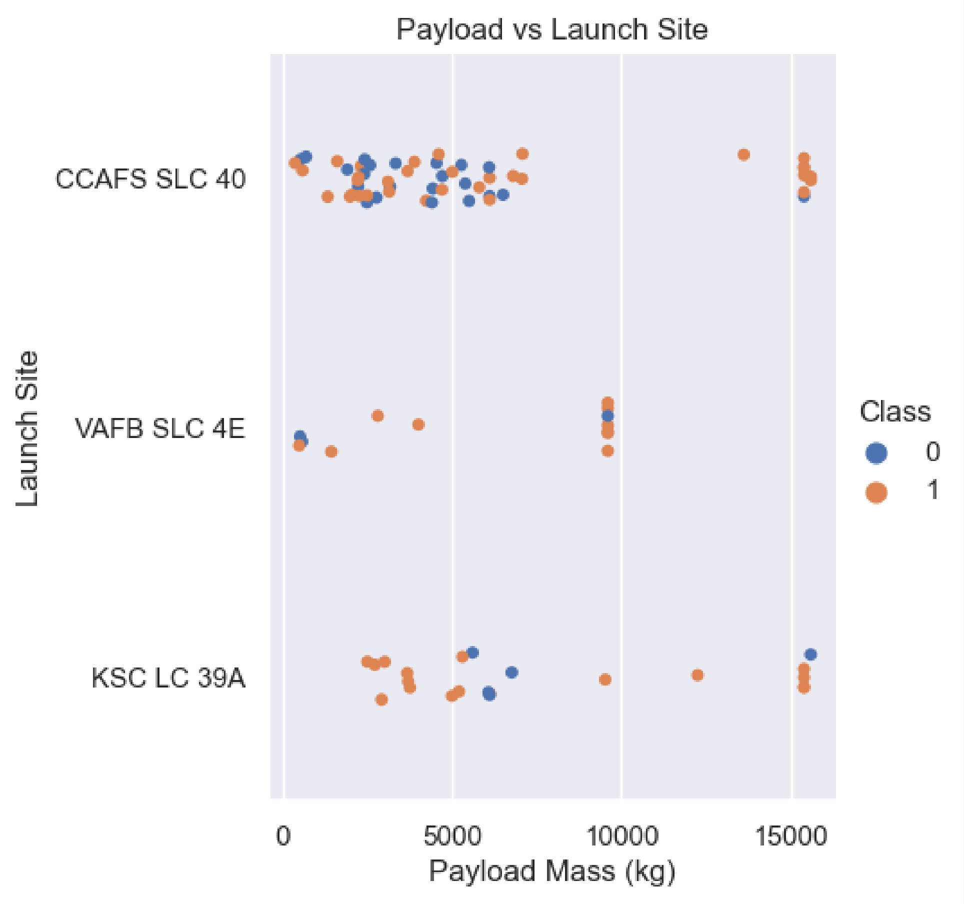
Flight Number vs. Launch Site

- Using a plot, we discovered that the more flights a launch location had, the more likely it was to succeed.



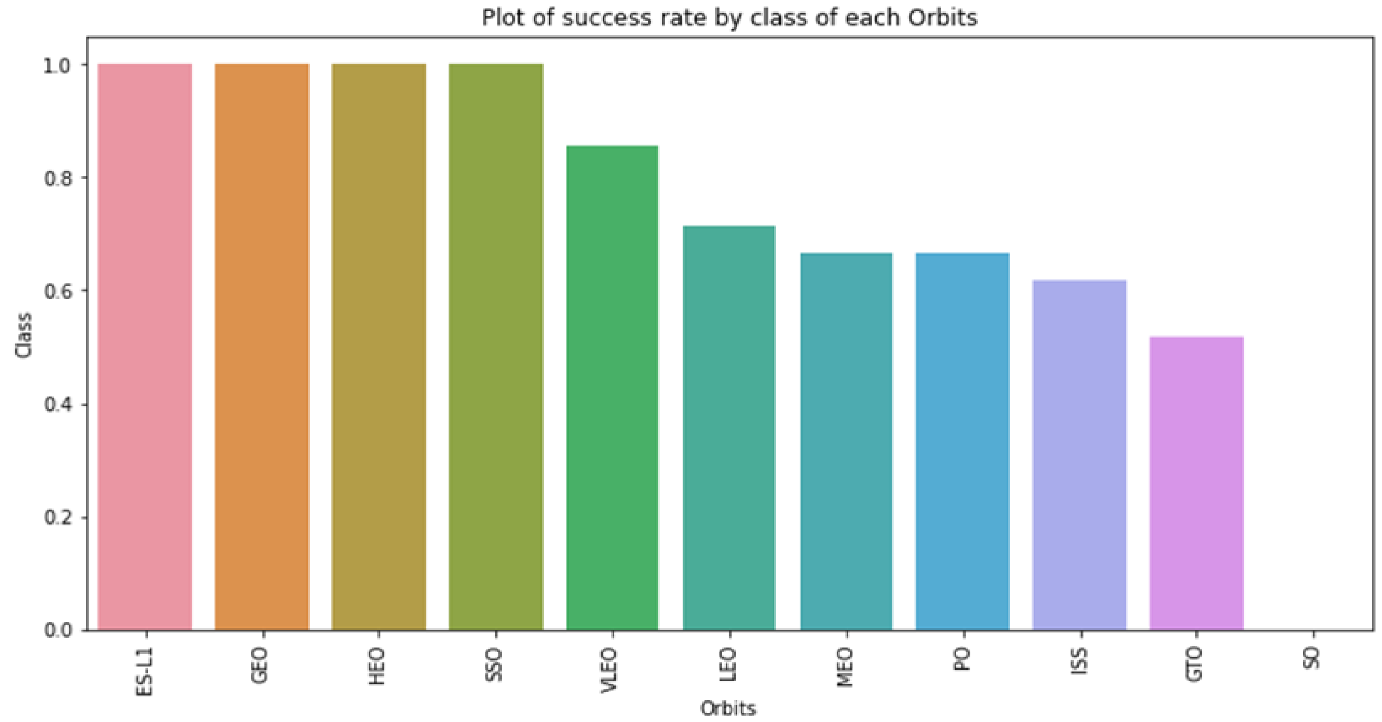
Payload vs. Launch Site

- Having a larger payload mass for a launch site increases a rocket's chances of success.



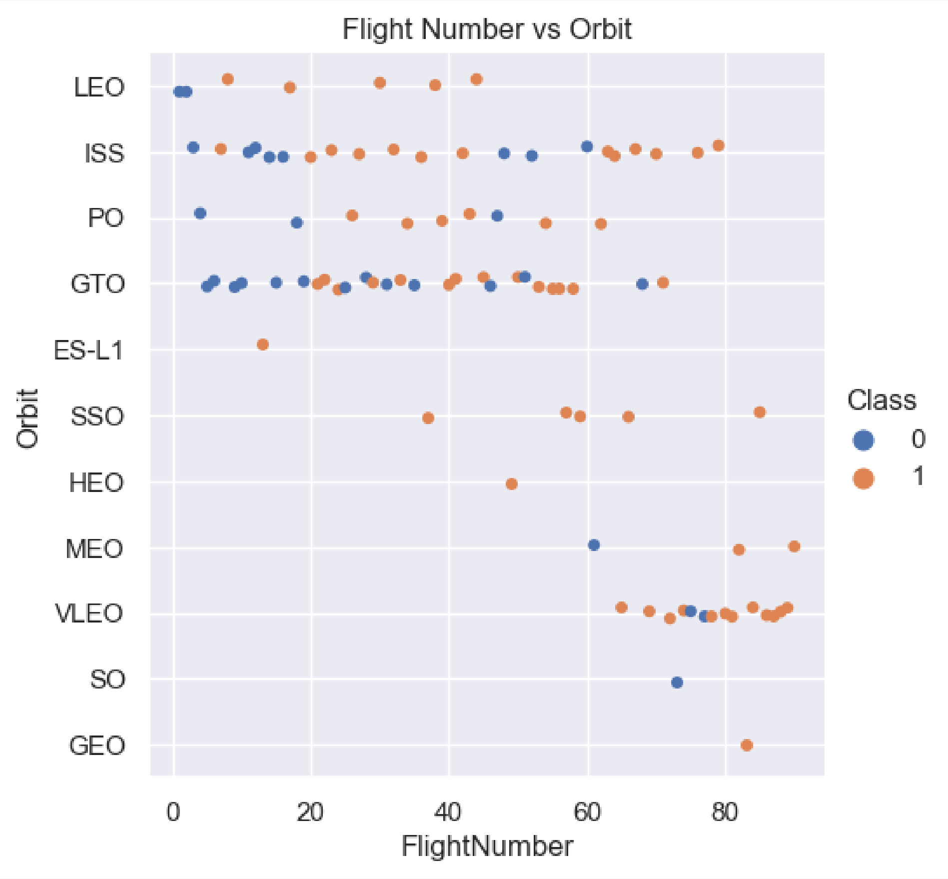
Success Rate vs. Orbit Type

- ES L1, GEO, HEO, and SSO are the only orbits with a success rate of 100 percent; all others, except for SO, have success rates that are only somewhat good.



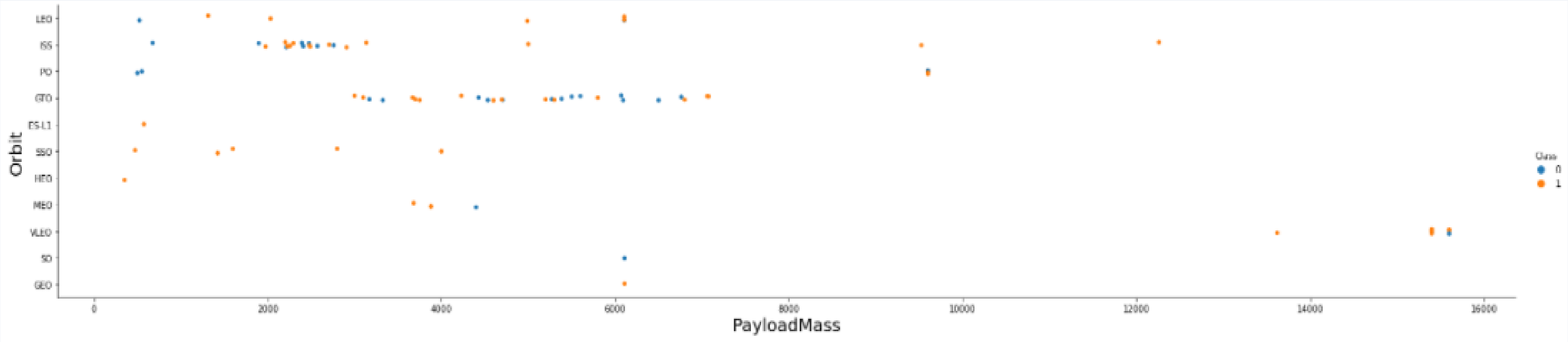
Flight Number vs. Orbit Type

- The higher the flight number, the greater the success rate for each orbit.



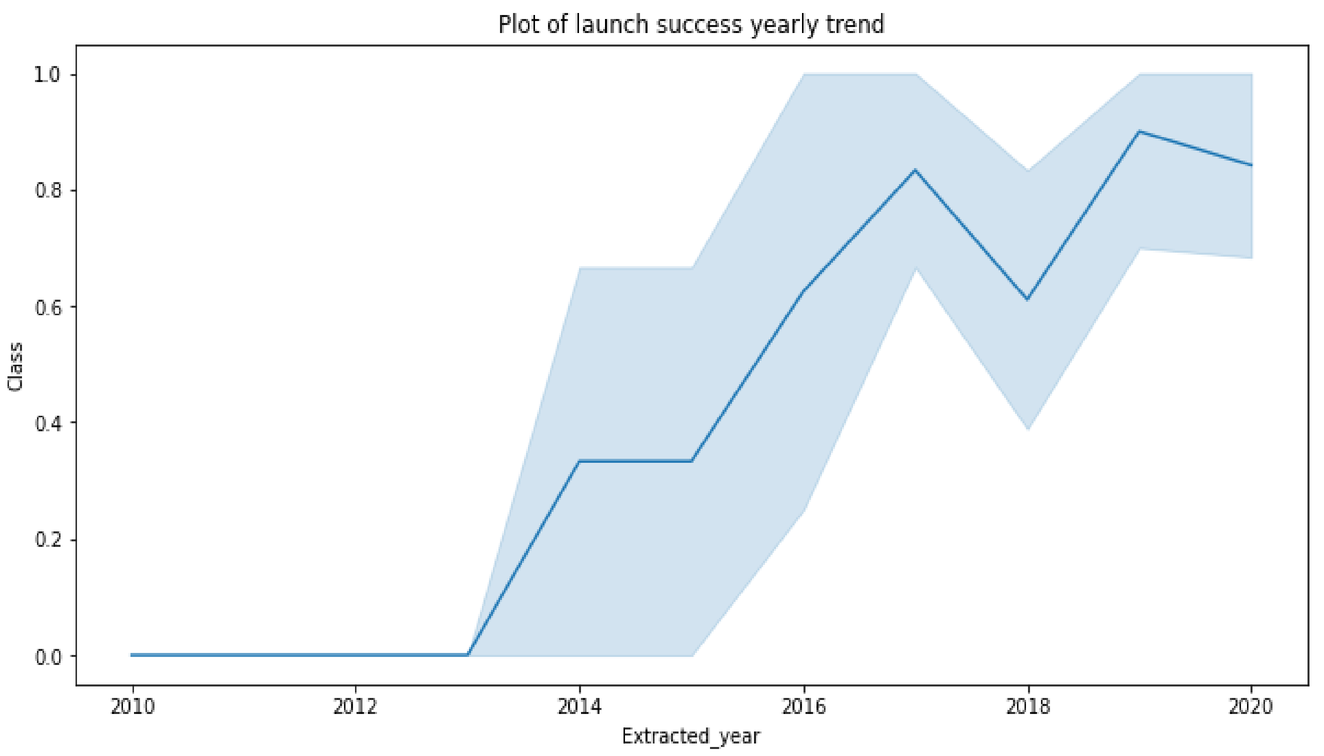
Payload vs. Orbit Type

- Heavy payloads are more likely to land successfully in PO, LEO, and ISS orbits.



Launch Success Yearly Trend

- We can see from the graph that the success rate has been steadily rising since 2013, except year 2018.



All Launch Site Names

- This query returns a list of Launch Sites with unique identifiers.
- SpaceX makes use of four different launch pads. 3 are on the east coast while the rest are on the west.

[5]:
```
query = "SELECT DISTINCT(Launch_Site) FROM Spacex"
pd.read_sql_query(query, conn)
```

[5]:

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

Launch Site Names Begin with 'CCA'

- In order to display the five entries with launch sites that begin with "CCA," we performed this query.

[6]:
```
query = """
    SELECT *
    FROM Spacex
    WHERE Launch_Site like 'CCA%'
    LIMIT 5"""
pd.read_sql_query(query, conn)
```

[6]:

	Date	Time(UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	08-12-2010	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	22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Using the following query, we were able to determine the overall weight of NASA's boosters: 45596.

[7]:

```
query = """
    SELECT SUM(CAST(PAYLOAD_MASS__KG_ as INT))
        AS NASA_PAYLOAD_SUM_KG
    FROM Spacex
    WHERE Customer = 'NASA (CRS)'
"""
pd.read_sql_query(query, conn)
```

[7]:

	NASA_PAYLOAD_SUM_KG
0	45596

Average Payload Mass by F9 v1.1

- The average payload that usually gets carried by F9 v1.1 is 2534.67 kg

[8]:

```
query = """
    SELECT AVG(CAST(PAYLOAD_MASS__KG_ as INT))
        AS F9v1_1_PAYLOAD_AVG_KG
    FROM Spacex
    WHERE Booster_Version LIKE 'F9 v1.1%'
"""
pd.read_sql_query(query, conn)
```

[8]:

	F9v1_1_PAYLOAD_AVG_KG
0	2534.666667

First Successful Ground Landing Date

- We discovered that the first successful landing on the launch pad occurred on December 22nd, 2015, according to records.

In [14]:
```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

Out[14]:

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of boosters that have successfully landed on the drone pad with payloads between 4000 and 6000 kg are included here.

[36]:

```
query = """
    SELECT DISTINCT(Booster_Version)
    FROM Spacex
    WHERE (SELECT CAST(PAYLOAD_MASS__KG_ as INT) AS PAYLOAD_MASS_KG
        WHERE PAYLOAD_MASS_KG > 4000 AND PAYLOAD_MASS_KG < 6000)
        AND Landing_Outcome like '%drone%'
"""
pd.read_sql_query(query, conn)
```

[36]:

	Booster_Version
0	F9 FT B1020
1	F9 FT B1022
2	F9 FT B1026
3	F9 FT B1021.2
4	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Most of the time, each mission succeeds and only one fails.

[11]:
```python
query = """
    SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Count
    FROM Spacex
    GROUP BY Mission_Outcome
"""
pd.read_sql_query(query, conn)
```

[11]:

	Mission_Outcome	Count
0	Failure (in flight)	1
1	Success	99
2	Success (payload status unclear)	1

Boosters Carried Maximum Payload

- There are a total of 12 boosters that can carry a maximum weight of 15600 kg.

```
query = """
    SELECT DISTINCT Booster_Version, CAST(PAYLOAD_MASS__KG_ AS INT)
        AS Payload_KG
    FROM Spacex
    WHERE Payload_KG IN (
        SELECT MAX(CAST(PAYLOAD_MASS__KG_ as INT)) FROM Spacex
    )
"""
pd.read_sql_query(query, conn)
```

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

2015 Launch Records

- Only two flights in 2015 were unable to touch down on a drone ship.

[10]:

```
query = """
    SELECT Landing_Outcome,
        Booster_Version,
        Launch_Site,
        Date
    FROM Spacex
    WHERE
        Date LIKE '%2015%' AND
        Landing_Outcome = 'Failure (drone ship)'
"""
pd.read_sql_query(query, conn)
```

[10]:

	Landing_Outcome	Booster_Version	Launch_Site	Date
0	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	10-01-2015
1	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	14-04-2015

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

- Ranked 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.

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad))

In [19]:

```
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)
```

Out[19]:

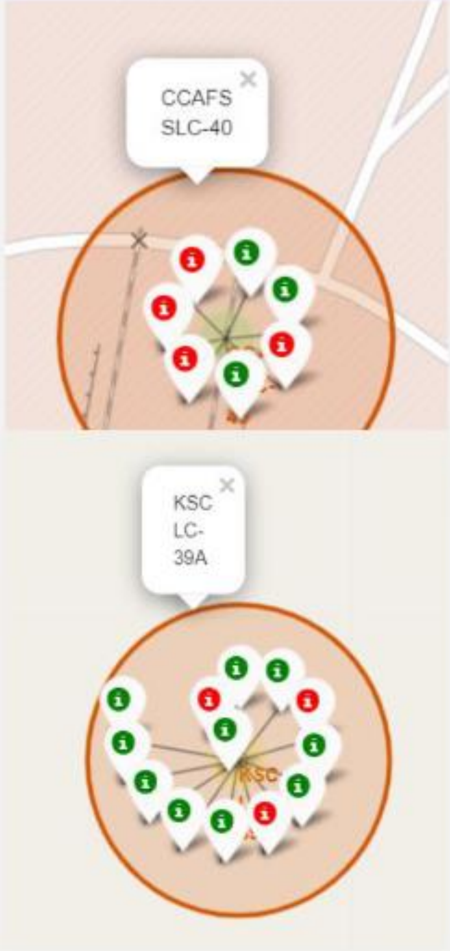
	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

Locations of Launch Sites

Markers with colored labels indicating the locations of the launch sites



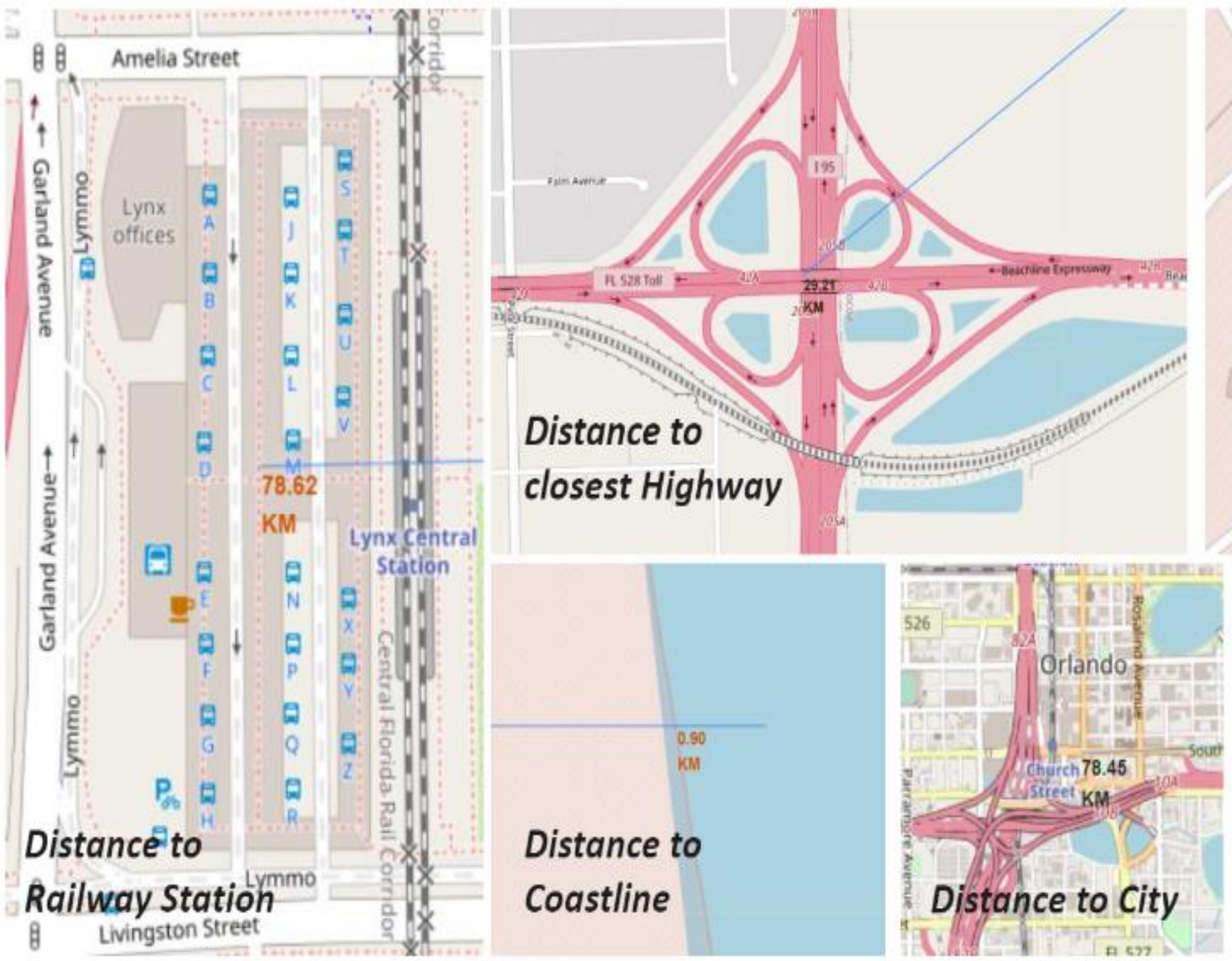
Florida Launch Sites

Green Marker shows successful Launches and Red Marker shows Failures

California Launch Site

37

Launch Sites Proximity

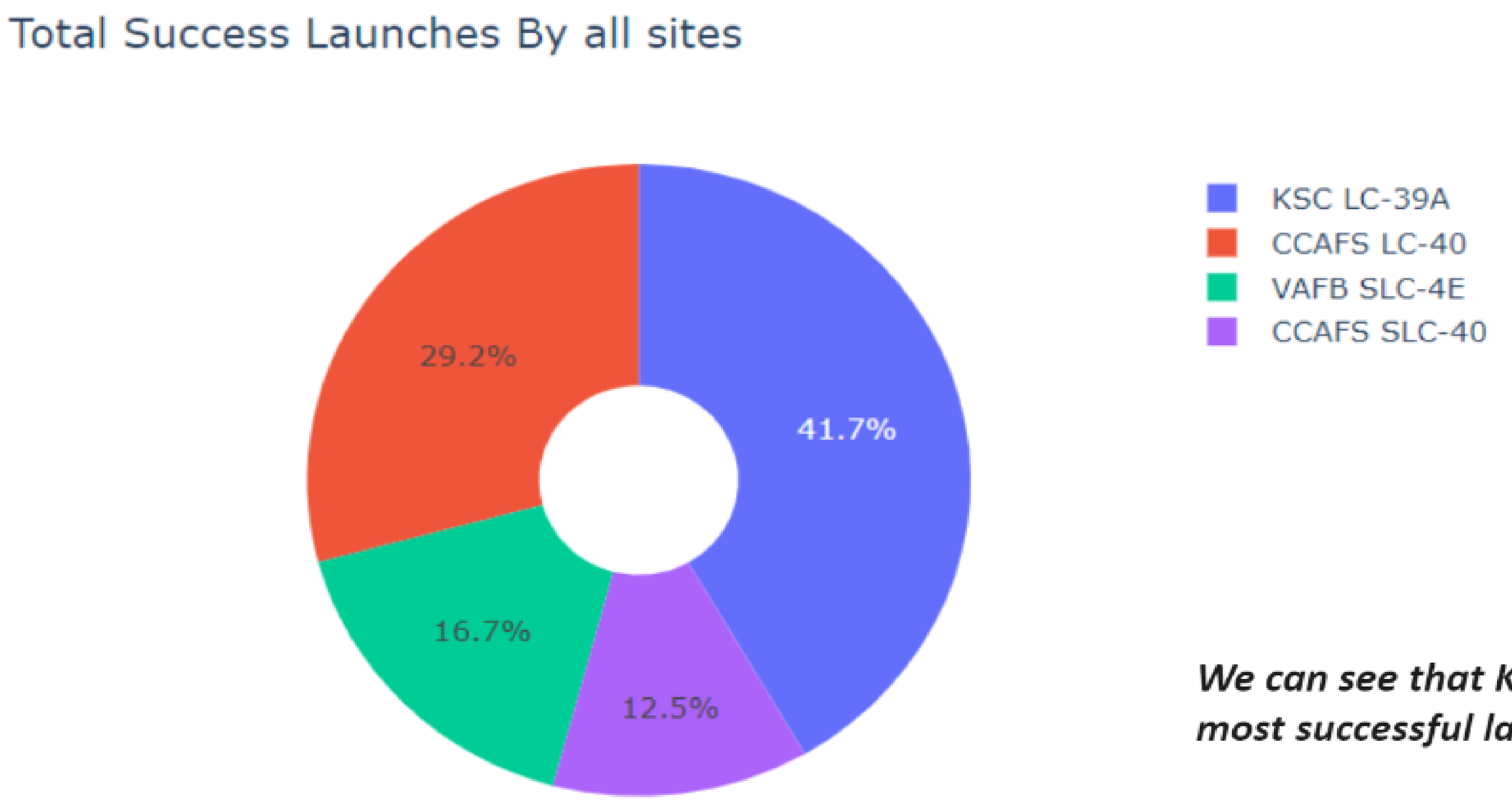


- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Section 4

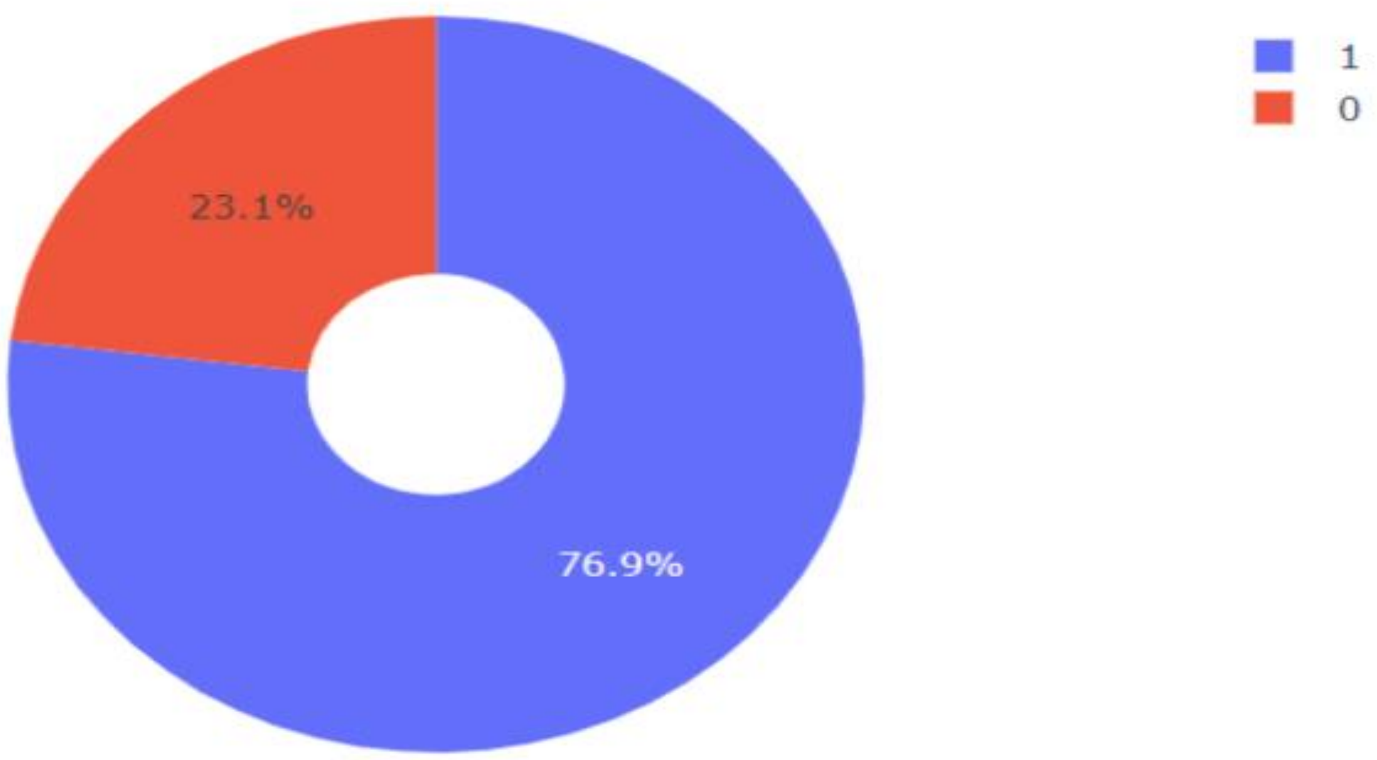
Build a Dashboard with Plotly Dash

All Launch Sites Success Rates



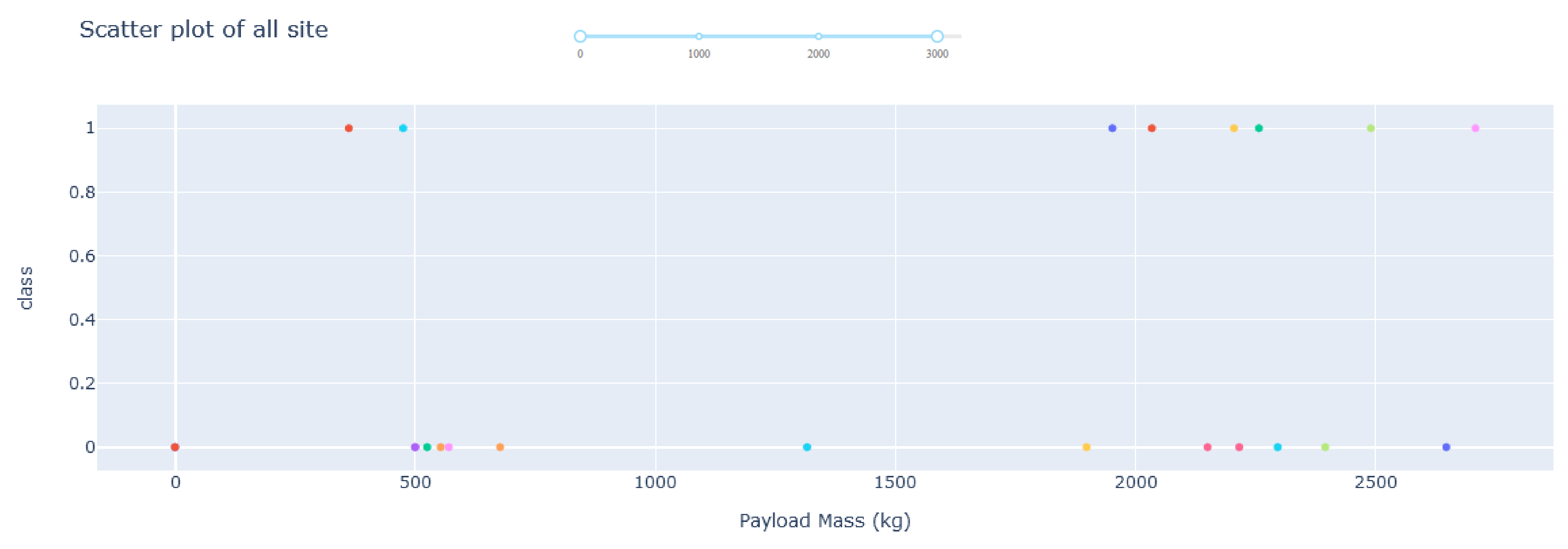
We can see that KSC LC-39A had the most successful launches from all the sites

Highest Success Rate Launch Site



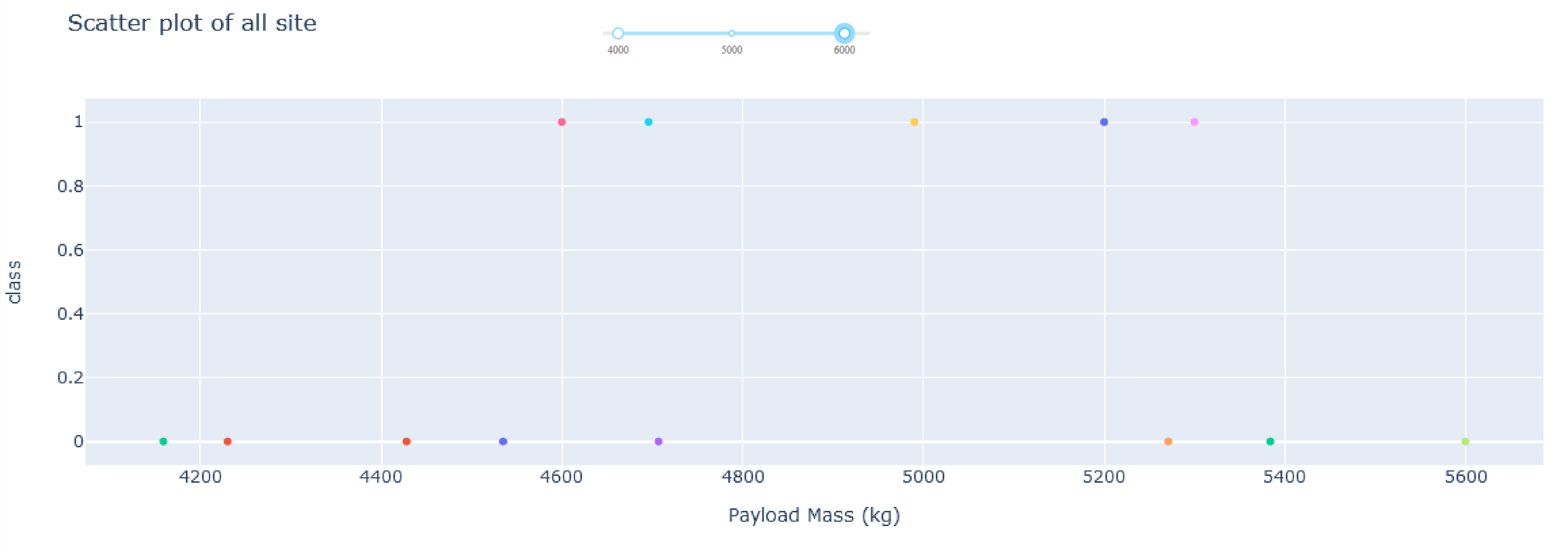
KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Success Rate at Different Payload Range



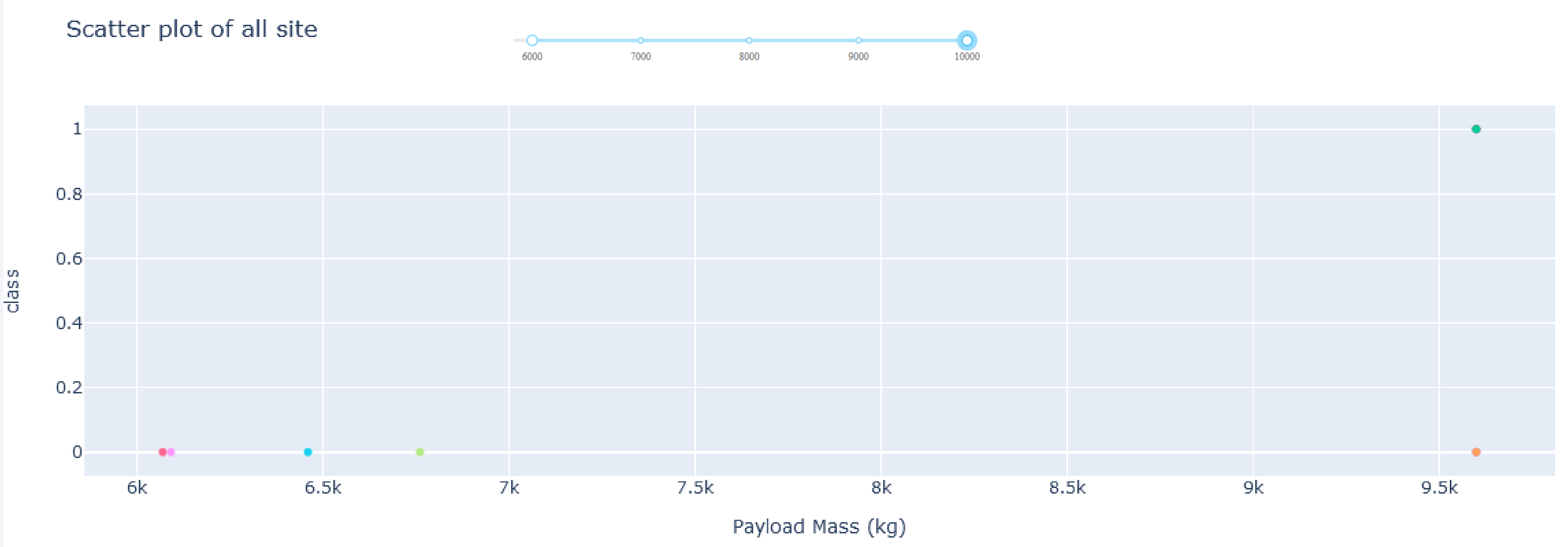
From 0 –3000 kg there are 6 successes

Success Rate at Different Payload Range



From 4000 –6000 kg there are 5 successes

Success Rate at Different Payload Range



From 6000 –10000 kg there is 1 success

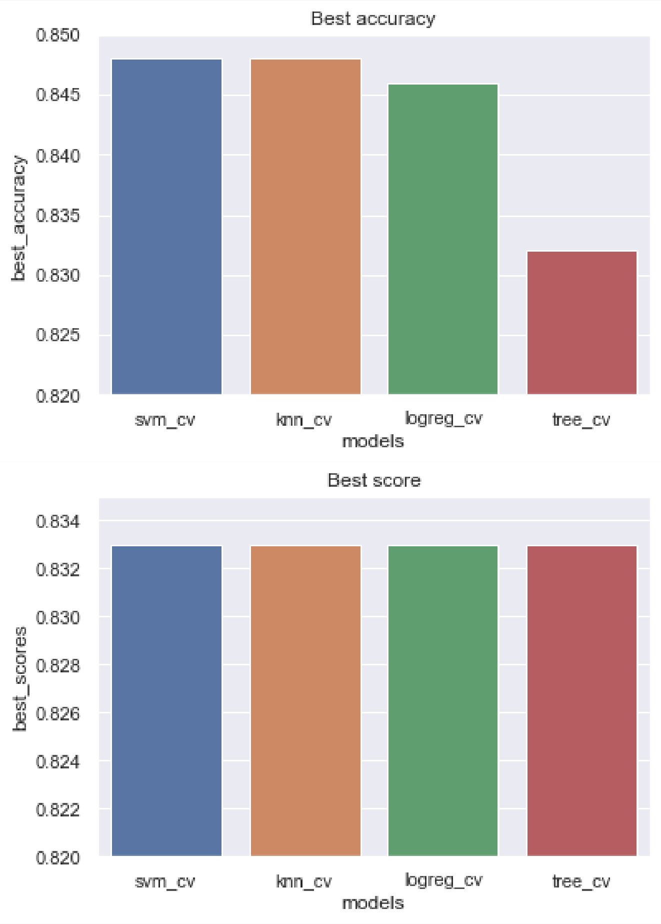
Section 5

Predictive Analysis (Classification)

Classification Accuracy

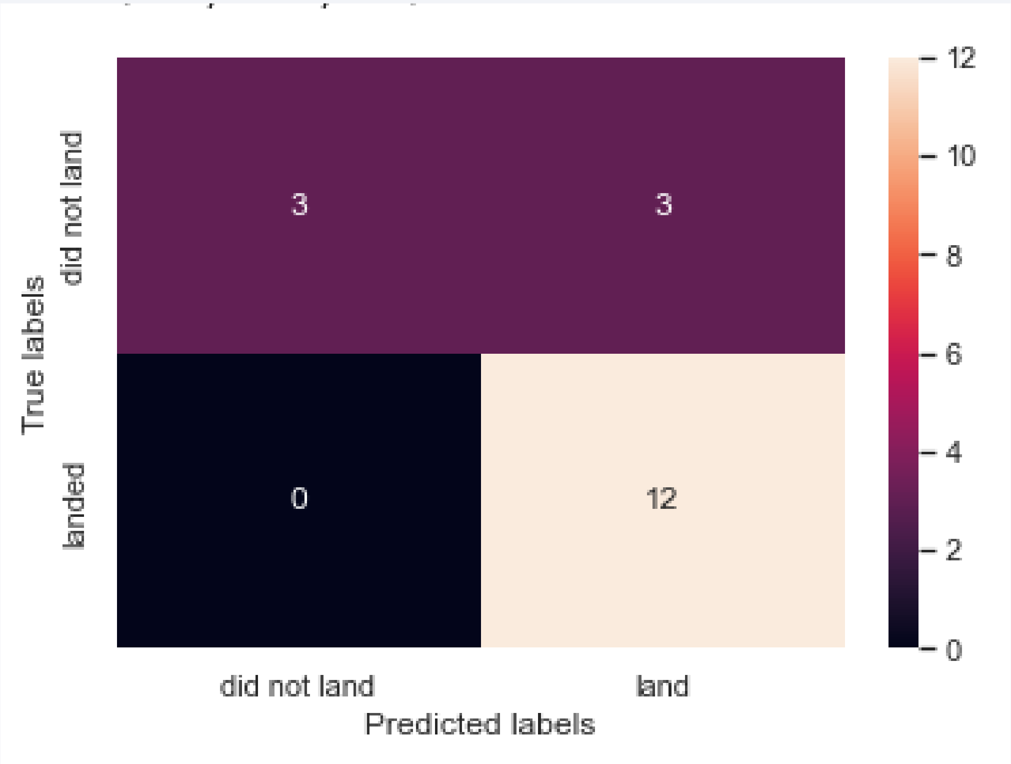
- Models with the best accuracy & scores are SVM and KNN with scores of 0.848 & 0.833

	models	best_accuracy	best_scores
0	svm_cv	0.848	0.833
1	knn_cv	0.848	0.833
2	logreg_cv	0.846	0.833
3	tree_cv	0.832	0.833



Confusion Matrix

- The confusion matrix for KNN and SVM is the same. There were three incorrect predictions for "did not land" and three correct predictions for the same class in the KNN and SVM plots. While they accurately predicted 12 of the "landed" class predictions with zero errors.



Conclusions

- The higher the number of flights from a launch site, the more successful the launch site is likely to be.
- From 2013 through 2020, the launch success rate for SpaceX is expected to rise.
- The most successful orbits were ES-L1, GEO, HEO, SSO, and VLEO.
- The most successful launches were at KSC LC-39A.
- The highest performing models are KNN and SVM.

Thank you!