IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jean Nelson
12/25/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection through API connection
  - Data wrangling
  - Exploratory data analysis with visualization and SQL queries
  - Data analysis using interactive dashboards and maps
  - Machine learning predictions

- Summary of all results
  - Given various factors such as launch site, orbit, and payloads, we can predict with over 80% accuracy whether the first stage of Falcon 9 will land successfully.

Introduction

- Project background and context

  SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. The goal of this project is to predict if the Falcon 9 first stage will land successfully.

- Problems you want to find answers

  - Are there relationships between the different factors and landing outcomes?
  - What factors and their interactions contribute to successful landings?
  - Can we predict an outcome given a launch condition?

Section 1

Methodology

Methodology

Executive Summary

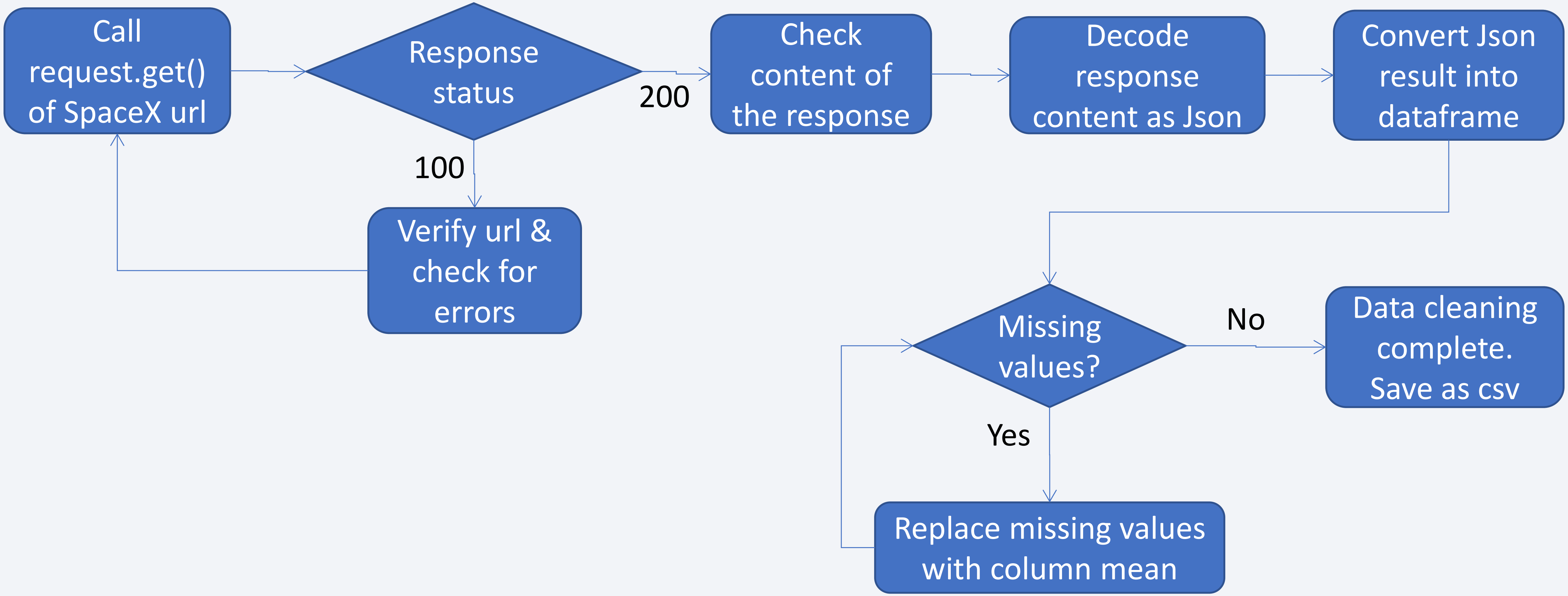
- Data collection methodology:
    - Collected launch data by making requests to the SpaceX API
- Perform data wrangling
    - Decoded the response content as a Json and turn it into a Pandas dataframe
    - Replaced numerical missing values with the mean of the column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Split data into training and test
    - Built several classification models by varying the hyperparameters using Gridsearch
    - Compared the accuracy of each model

Data Collection

Collect launch data through the SpaceX API

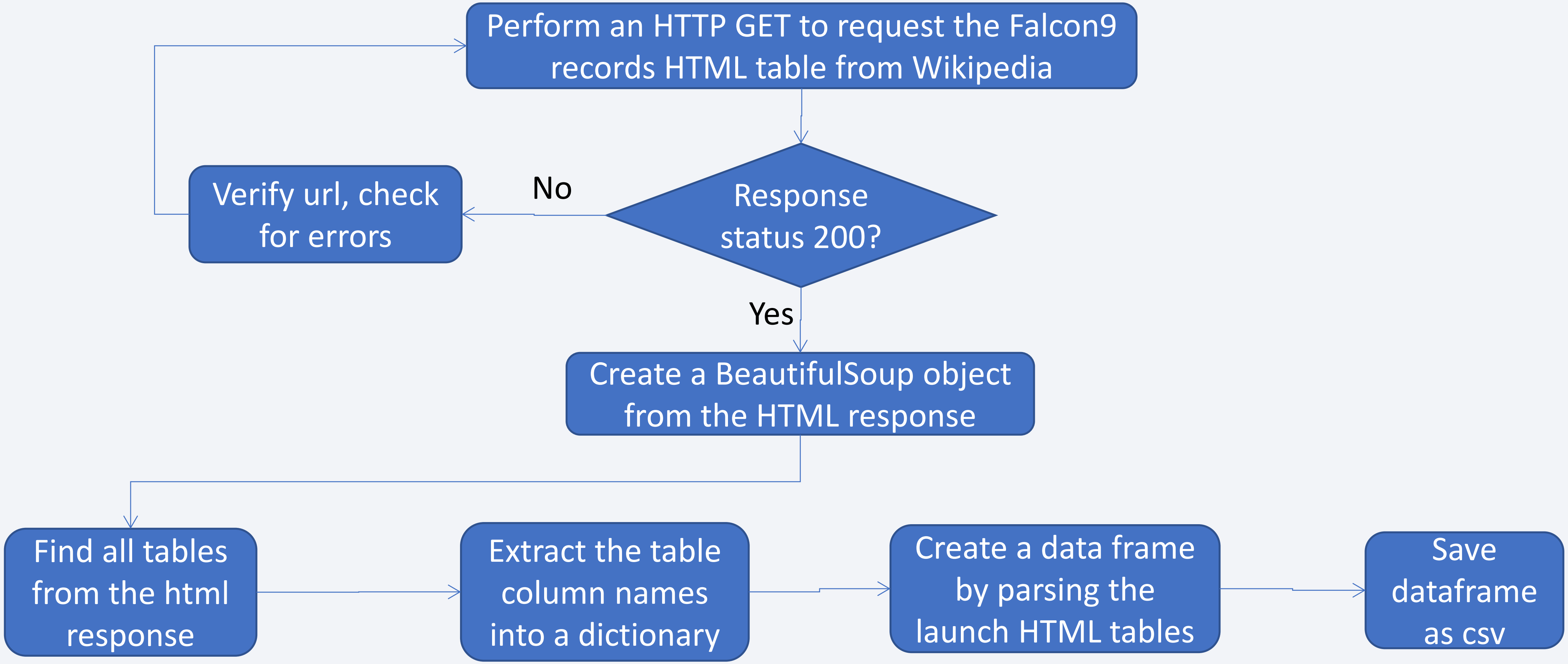
- Make a get request to the SpaceX API
- Decode the response content as a Json and turn it into a Pandas dataframe
- Clean the dataframe:
  - Filter Falcon 9 booster version
  - Replace missing values for payload mass with the mean of payload masses
  - Create landing outcome label as binary (good or bad)

Data Collection – SpaceX API



Github Link: https://github.com/eddyson8/IBM_DS_Capstone/blob/main/Lab1_spacex-data-collection-api.ipynb

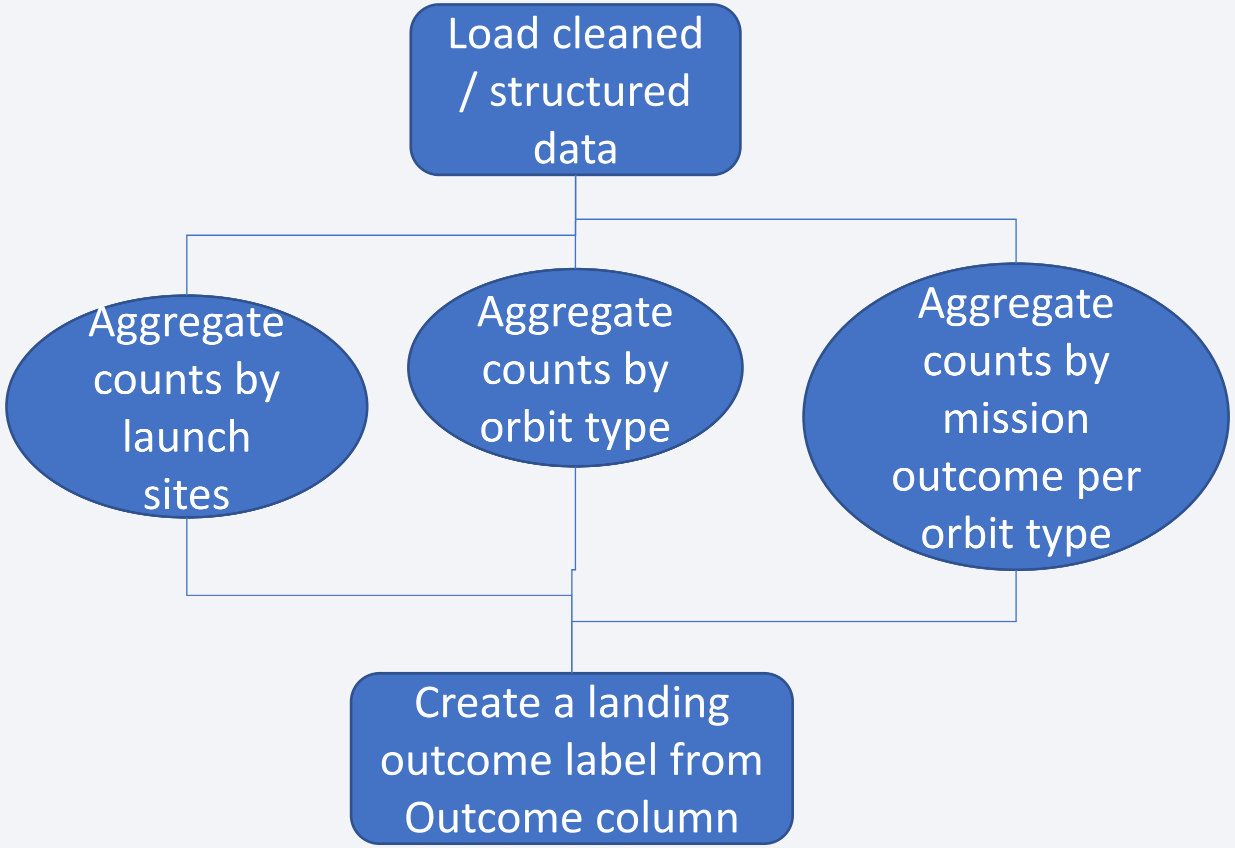
Data Collection - Scraping

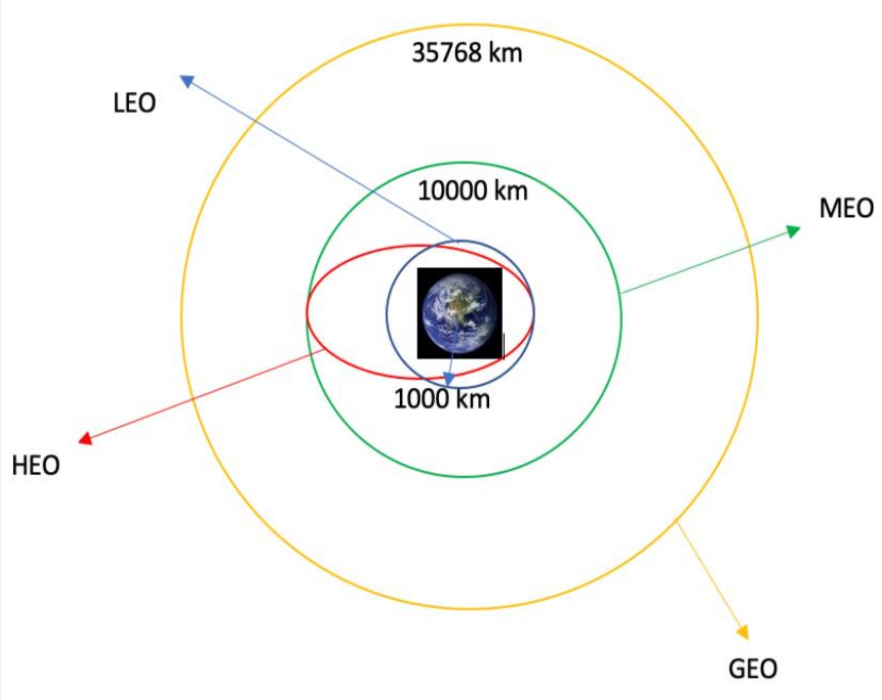


https://github.com/eddyson8/IBM_DS_Capstone/blob/main/Lab1b-webscraping.ipynb

Data Wrangling

- Performed some Exploratory Data Analysis to understand the data and prepare for further analysis
- Determined training labels for supervised models





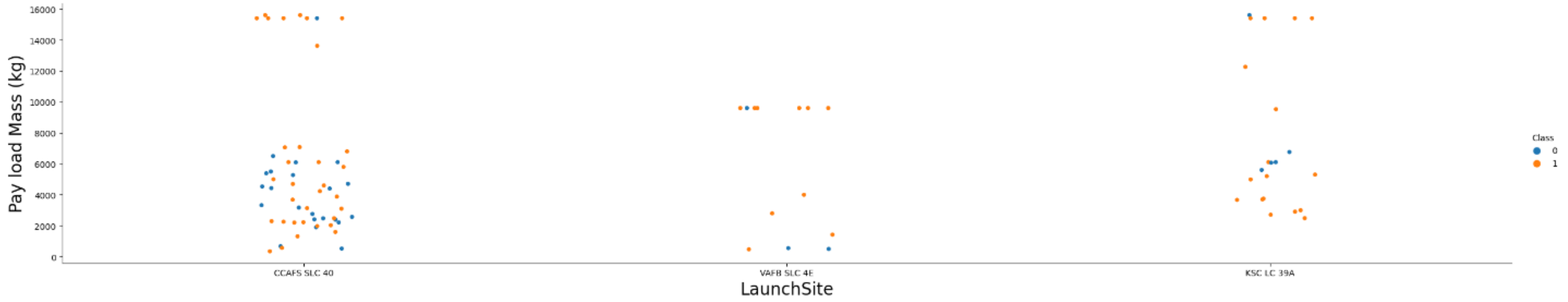
https://github.com/eddyson8/IBM_DS_Capstone/blob/main/Lab2_spacex-Data%20wrangling.ipynb

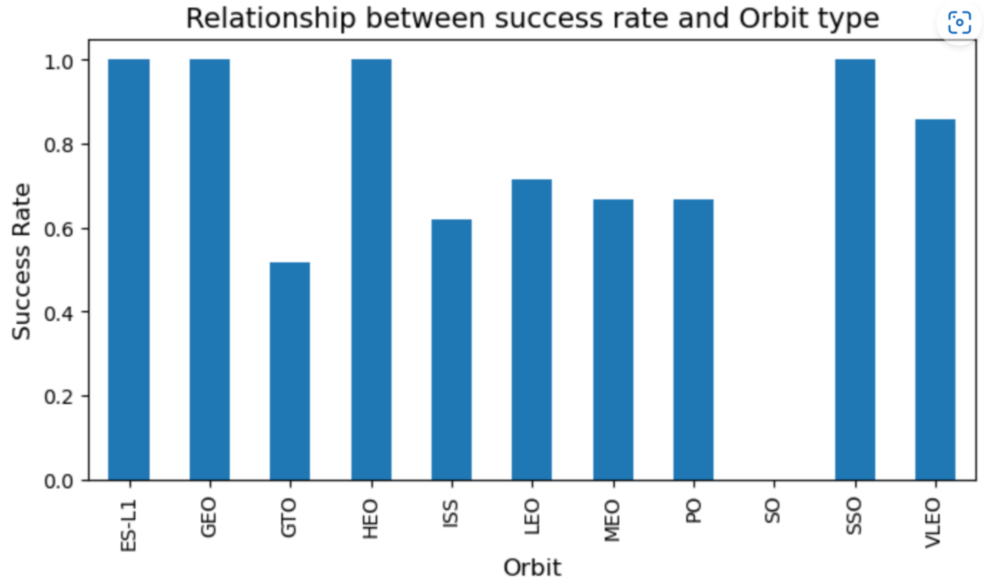
EDA with Data Visualization

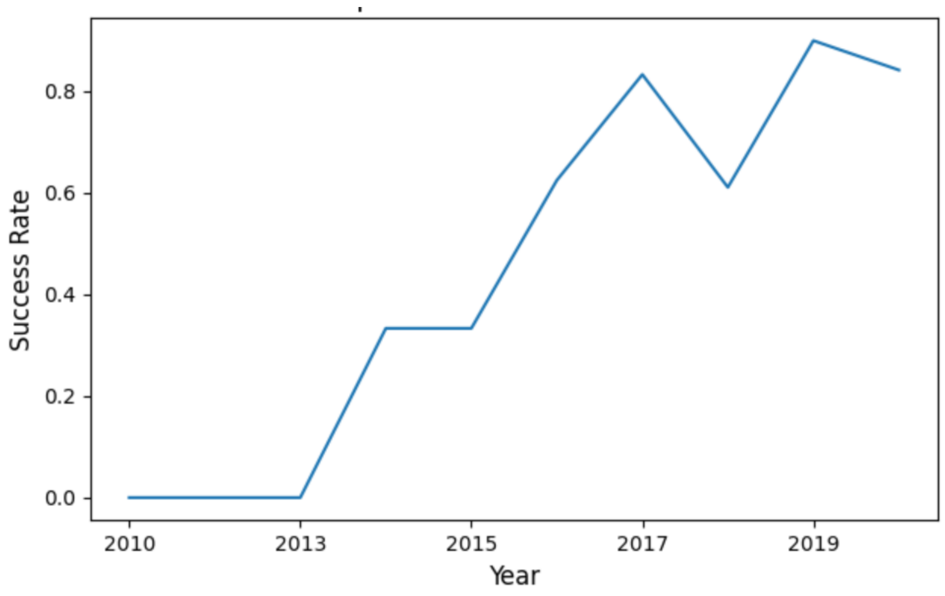
Performed some Exploratory Data Analysis with visualizations to find some useful patterns in the data

- Visualized the relationship between Flight Number and Payload by landing outcomes
  - The higher the payload mass, the more likely a successful landing outcome
- Visualized the relationship between Flight Number and Launch Site by landing outcomes
  - VAFB SLC 4E has higher success rate, but with fewer launches historically
  - As number of flights increases at a launch site, the success rate tends to increase as well, possibly due to learning curve
- Visualized the relationship between Payload and Launch Site by landing outcomes
  - Launch sites CCAFS LC-40 and KSC LC-39A have higher success rates with higher payloads
- Visualized the relationship between success rate and orbit type
  - The following orbit types have success rates greater than 95 percent: ES-L1, GEO, HEO, SSO
- Visualized the relationship between FlightNumber and Orbit type by landing outcomes
- Visualized the relationship between Payload and Orbit type by landing outcomes
- Visualized the launch success yearly trend: success rate tends to increase over time from 2013

EDA with Data Visualization (2)







These plots highlight the relationship between:

1) payload, launch sites, and success rate,

2) orbit and success rate, and

3) the trend of success rate over time.

https://github.com/eddyson8/IBM_DS_Capstone/blob/main/Lab4_eda-dataviz.ipynb

EDA with SQL

Summary of SQL queries performed to get insights from the data:

- Load the SQL extension and establish a connection with the database
- Displayed the names of the unique launch sites and number of launches at each
- Displayed the total payload mass carried by boosters launched by one particular customer, NASA (CRS)
- Displayed the average payload mass carried by one particular booster version, F9 v1.1
- Displayed the date when a particular landing outcome first succeeded, ground pad.
- Displayed the booster versions for a particular landing outcome with a given payload range (successful drone ship landing with payload between 4000 and 6000 kg)
- Displayed the names of the booster versions which have carried the maximum payload mass, using a subquery
- Display month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015, using substring

https://github.com/eddyson8/IBM_DS_Capstone/blob/main/Lab3_eda-sql.ipynb

Build an Interactive Map with Folium

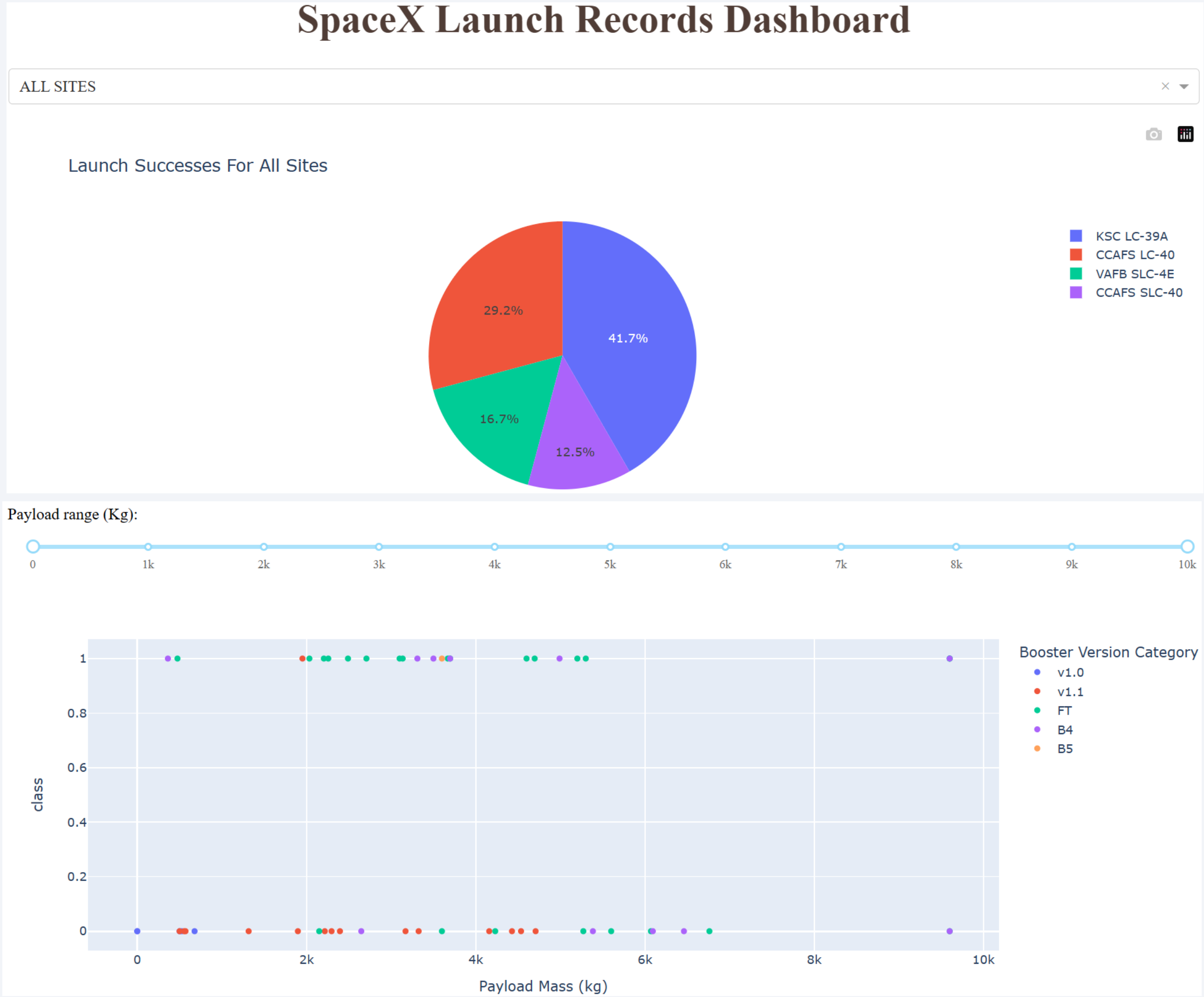
- Plotted the latitude and longitude coordinates of each launch site and marked them with a circle of 1000 meters radius to be visible on a map. They are all in close proximity to the coast
- Added the launch outcomes for each site, color-coded, to see which ones have high success rates. Added a marker cluster to display the number of launches in a cluster when zoomed out.
- Added a MousePosition to get the lat/long coordinates when moused over a point on the map. The moused over coordinates can be used to calculate distances from a point to a launch site.

IBM Watson Link: https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/84d141e2-af91-4f3a-9e9c-e8b2afc9857b/view?access_token=ded32e688366842f0cf481a72e03d51a3bad1984fc0adf4aef20d9514fdb0332

Github Link: https://github.com/eddyson8/IBM_DS_Capstone/blob/main/Lab5_Data_Visualization_Folium_map.ipynb

Build a Dashboard with Plotly Dash

Built a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time



- Created a pie chart to display the percentage of all successful launches from all the site. Most of the successful launches are from launch site KSC LC-39A.
    - Users also have the option to view pie charts showing the success rate from each site by selecting from a dropdown menu. Thee majority of the launches from KSC LC-39A, 77%, were successful.
- Created a scatter plot of payload mass vs class, color coded by booster version, and a slider of payload mass, giving the user the option to vary the payload range to view.
    - Users can view a scatter plot of all launch site or a selected site from the dropdown.

https://github.com/eddyson8/IBM_DS_Capstone/blob/main/Lab6_Dashboard.ipynb

Predictive Analysis (Classification)

Flow chart to summarize how the classification models were built, evaluated, improved, and optimized

Prepare the data for input into machine learning models

Load csv into dataframe → Convert categorical variables into dummy variables → Create features array and response array → Standardize the features array → Split the features and response into training and test sets

Build machine learning models, tune hyperparameters, and calculate accuracies

Create model object → Create parameters dictionary with values to be used for parameter optimization → Create GridSearchCV object → Fit the model object and parameters into the GridSeachCV object → Fit the training set into the model → Calculate the model accuracy on the validation data with the best parameters found → Calculate the model accuracy on the test data

Models Tested: logistic regression, SVM, Decision Tree, KNN

https://github.com/eddyson8/IBM_DS_Capstone/blob/main/Lab7_ML.ipynb

Results

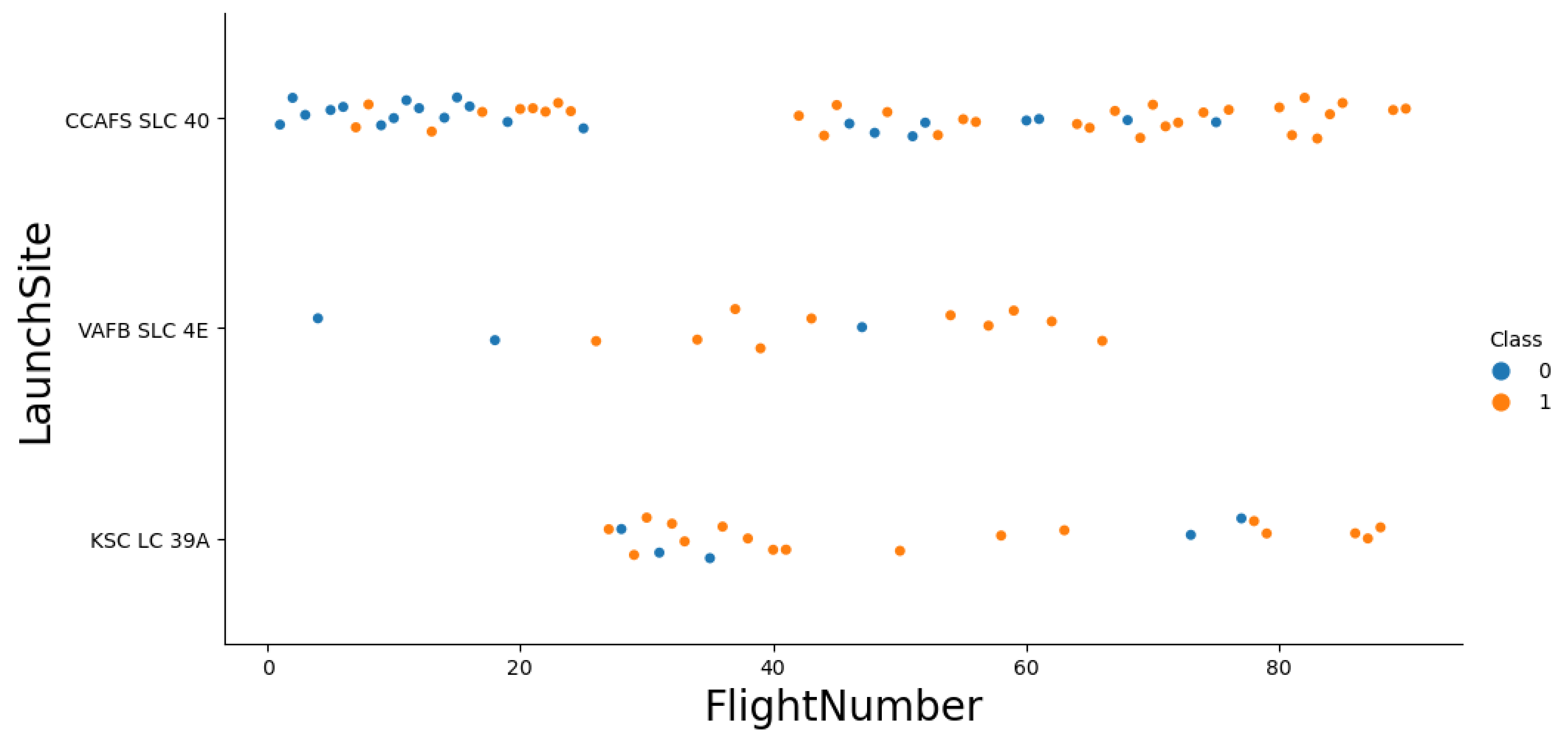
- Exploratory data analysis results
  - Higher payload mass corresponds with more likely a successful landing outcomes
  - Orbit sites ES-L1, GEO, HEO, SSO almost always result in successful outcomes
  - Success rate tends to increase over time
- Interactive analytics demo in screenshots
  - Launch site KSC LC-39A showed higher success rate
- Predictive analysis results
  - The highest accuracy score for the validation sets was found with the decision tree model. However the accuracy score was the same, 0.83, in the test sets from all the models.

Section 2

Insights drawn from EDA

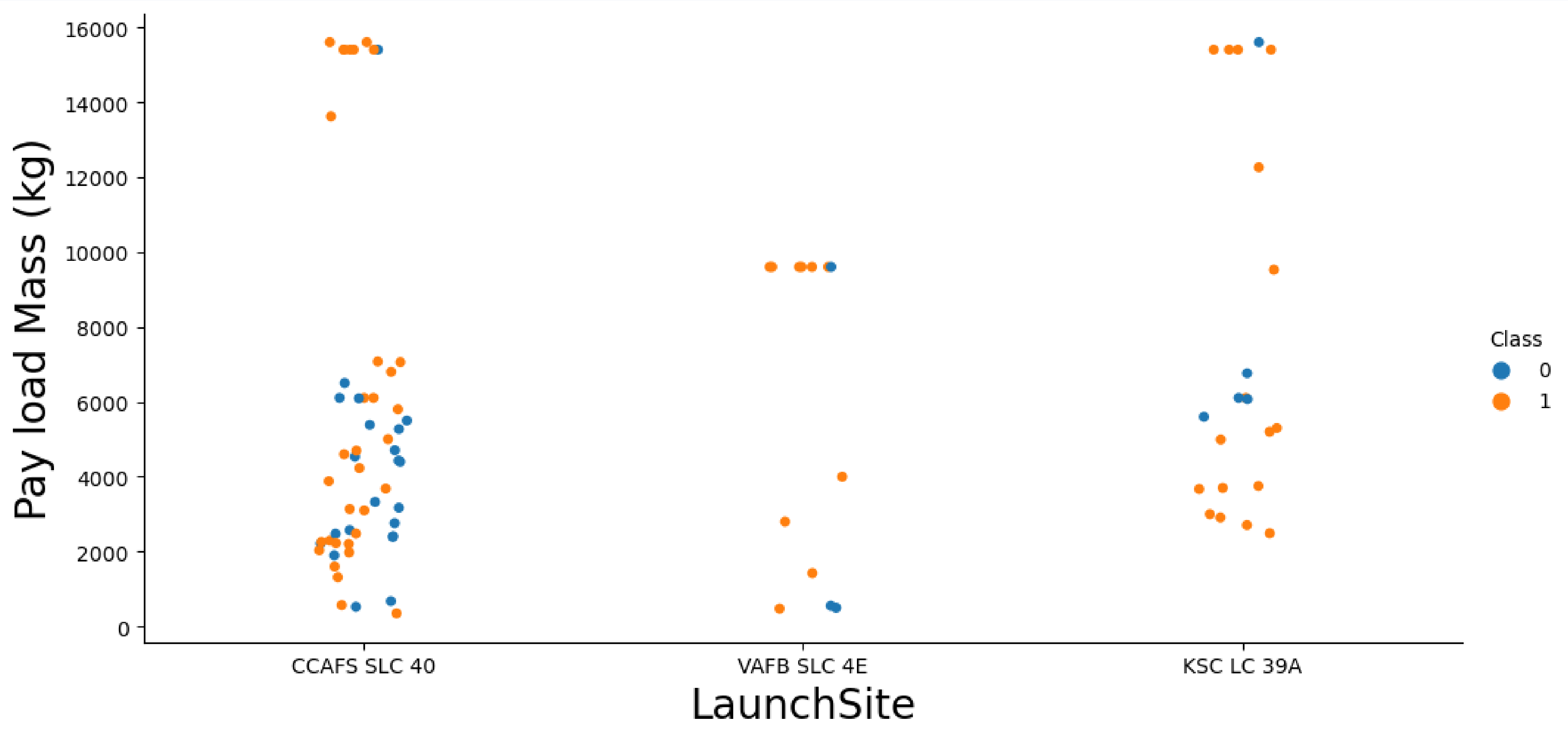
Flight Number vs. Launch Site

More flights at a launch site corresponds with higher success rate



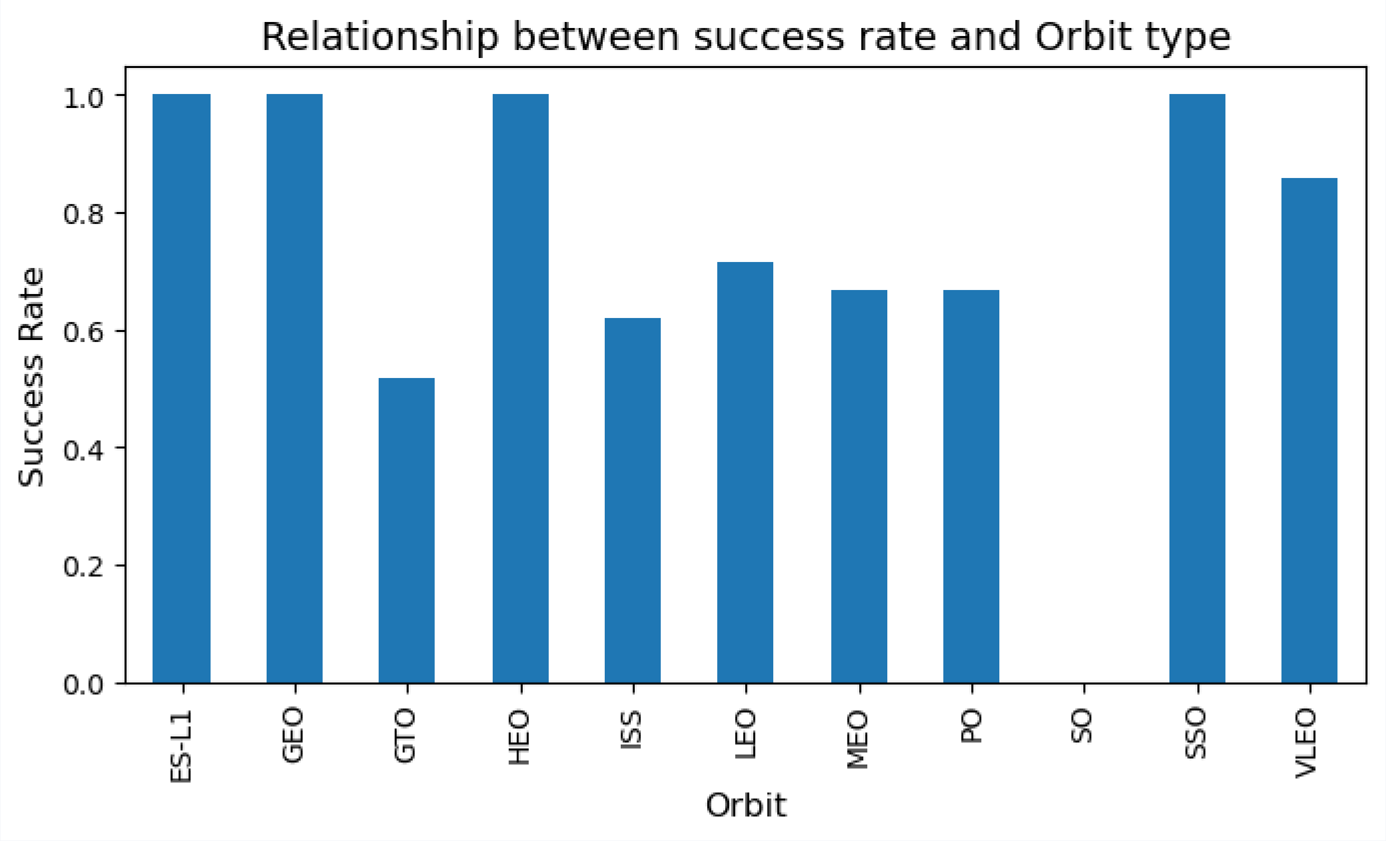
Payload vs. Launch Site

- KSC LC_39A and VAFB SLC_4E have higher success rates than CCAFS SLC_40.
- Success rate higher with larger payload mass.



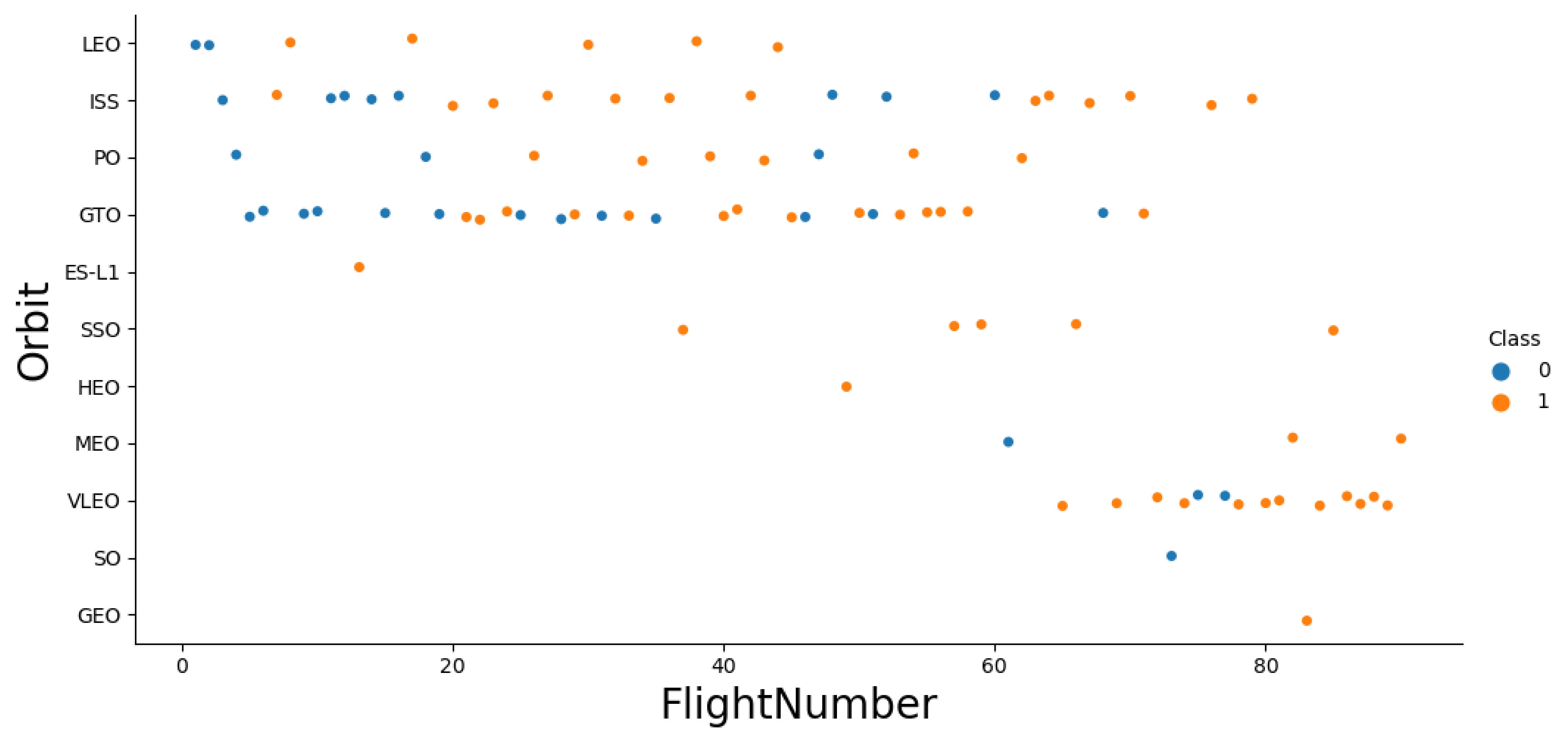
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, and SSO have very high success rate, close to 100% (although number of flights might be relatively less).
- Orbit type GTO have the lowest success rate, close to 50%.



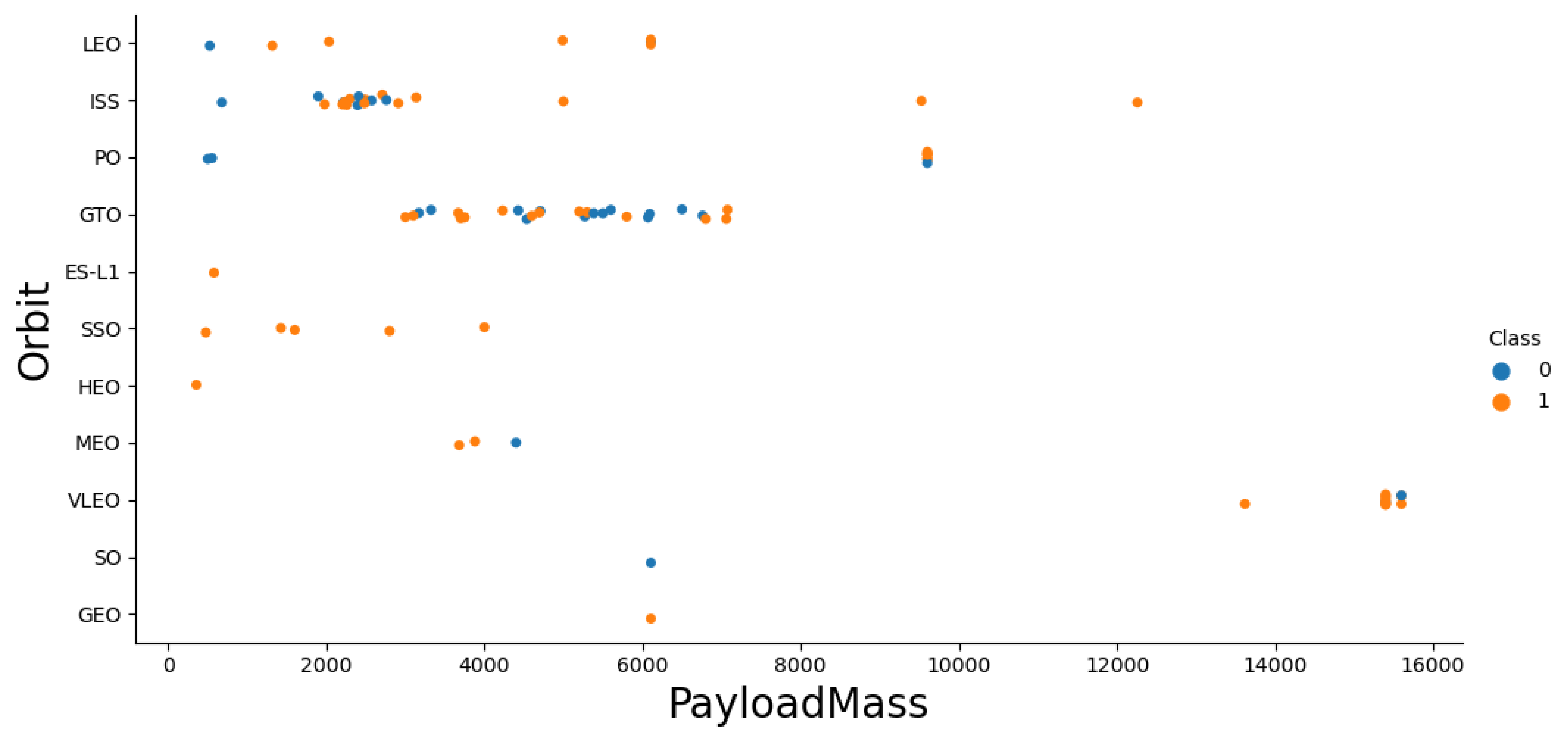
Flight Number vs. Orbit Type

- VLEO shows high successful when number of flights at an orbit is taken into account



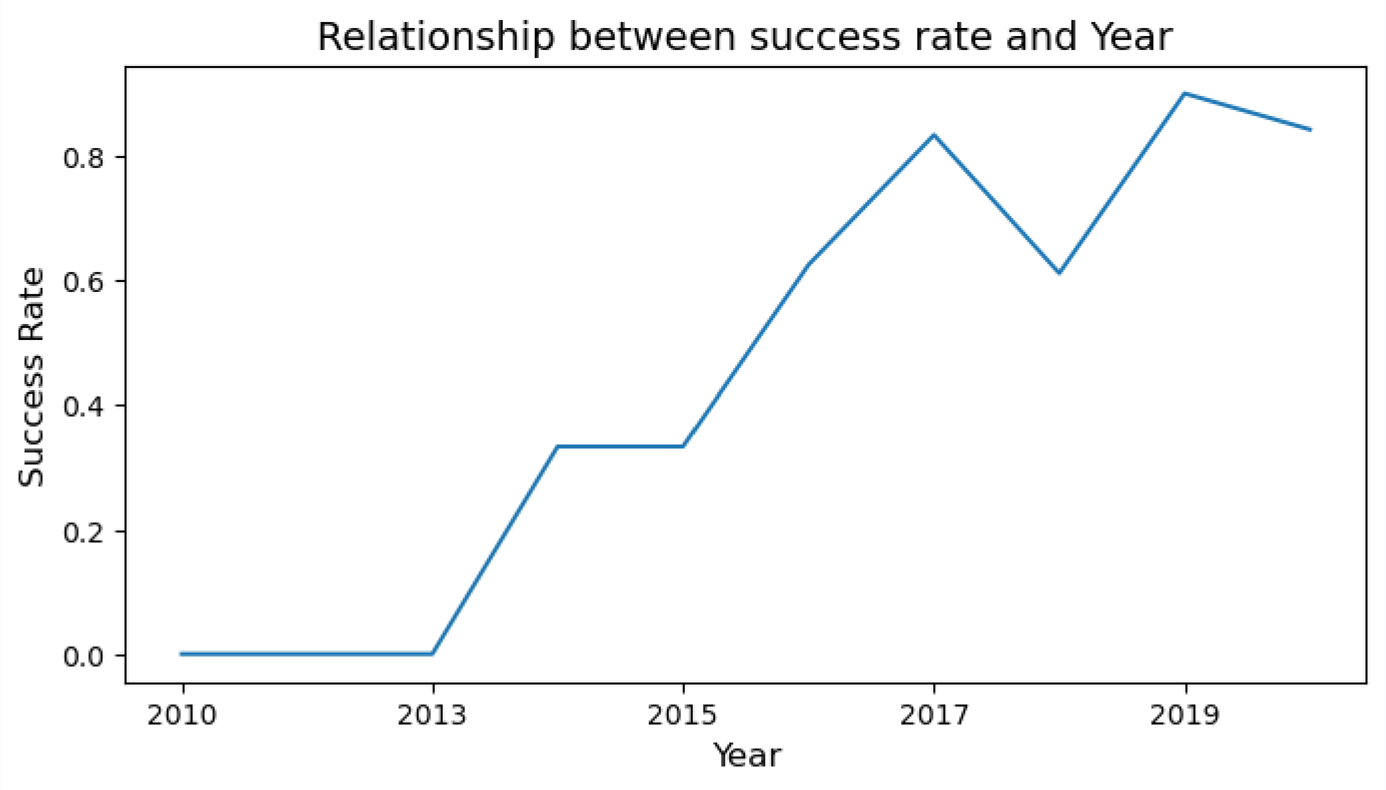
Payload vs. Orbit Type

- Successes appear to be more consistent at higher payload for orbit ISS.



Launch Success Yearly Trend

- Success rate has been in an upward trend since 2013. There was a notable drop in success rate in 2018.



All Launch Site Names

Table displaying the names of the unique launch sites. There were four launch sites used.

```
%sql SELECT LAUNCH_SITE, COUNT(*) AS NUM_LAUNCHES FROM SPACEX GROUP BY LAUNCH_SITE
```

launch_site	num_launches
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

Launch Site Names Begin with 'CCA'

Result showing 5 records where launch sites begin with `CCA`

```
%sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

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

Query and result for calculating the total payload carried by boosters from customer NASA. Total payload from NASA was 45,596.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "NASA_CRS_PAYLOAD_MASS_SUM" \
FROM SPACEX \
WHERE CUSTOMER = 'NASA (CRS)'
```

nasa_crs_payload_mass_sum
45596

Average Payload Mass by F9 v1.1

Query and result to calculate the average payload mass carried by booster version F9 v1.1. The average payload was 2,928.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "F9-v1.1_AVG_PAYLOAD_MASS" \
FROM SPACEX \
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

F9-v1.1_AVG_PAYLOAD_MASS
2928

First Successful Ground Landing Date

Finding the dates of the first successful landing outcome on ground pad. The date was 22 December 2015.

```
%sql SELECT MIN(DATE) AS "First successful gound pad landing date" \
FROM SPACEX \
WHERE Landing__Outcome = 'Success (ground pad)'
```

First successful gound pad landing date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql SELECT BOOSTER_VERSION \
FROM SPACEX \
WHERE Landing__Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Total Number of Successful and Failure Mission Outcomes

Result displaying the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS "Mission Success/Failure" \
FROM SPACEX \
GROUP BY MISSION_OUTCOME
```

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

Boosters Carried Maximum Payload

Table displays the names of the booster which have carried the maximum payload mass using a subquery.

```
%sql SELECT Booster_Version \
FROM SPACEX \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX)
```

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

Table displaying the month, year, booster versions, and launch sites of failed landing_outcomes in drone ship in year 2015

```
%sql SELECT MONTHNAME(DATE) AS MONTH, YEAR(DATE) AS FAIL_YEAR, LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE \
FROM SPACEX \
WHERE LANDING__OUTCOME LIKE 'Failure (d%' AND SUBSTR(DATE,1,4)='2015'
```

MONTH	fail_year	landing__outcome	booster_version	launch_site
January	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Table listing, in descending order, the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20

```
%sql SELECT LANDING__OUTCOME, COUNT(*) AS NUM_SUCCESSES FROM SPACEX \
WHERE LANDING__OUTCOME LIKE 'Success%' AND  DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY LANDING__OUTCOME
```

landing__outcome	num_successes
Success (drone ship)	5
Success (ground pad)	3

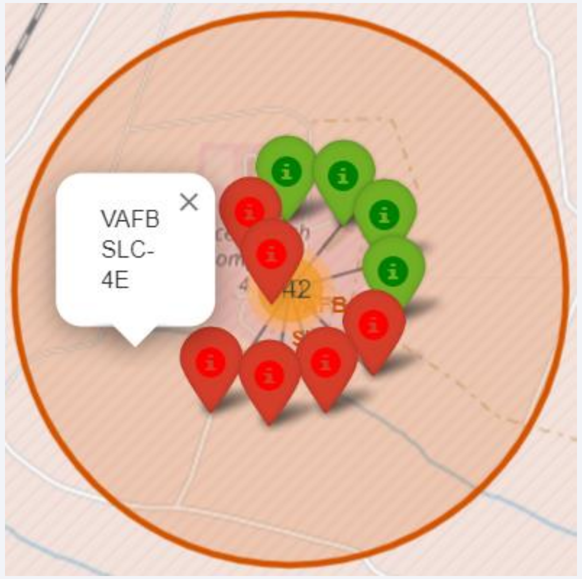
Section 3

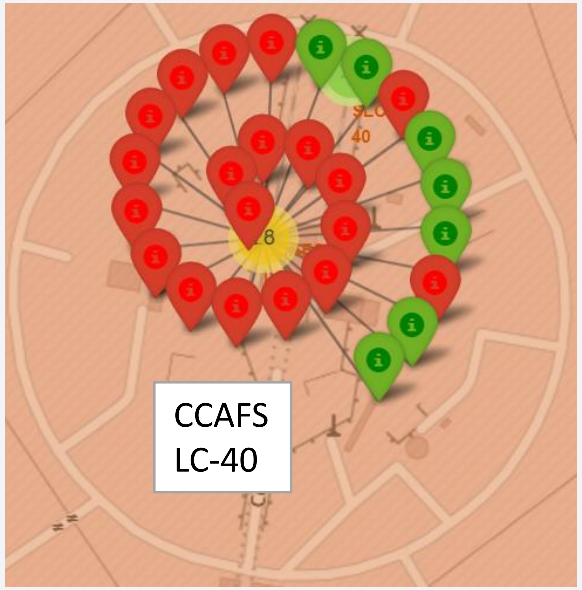
Launch Sites Proximities Analysis

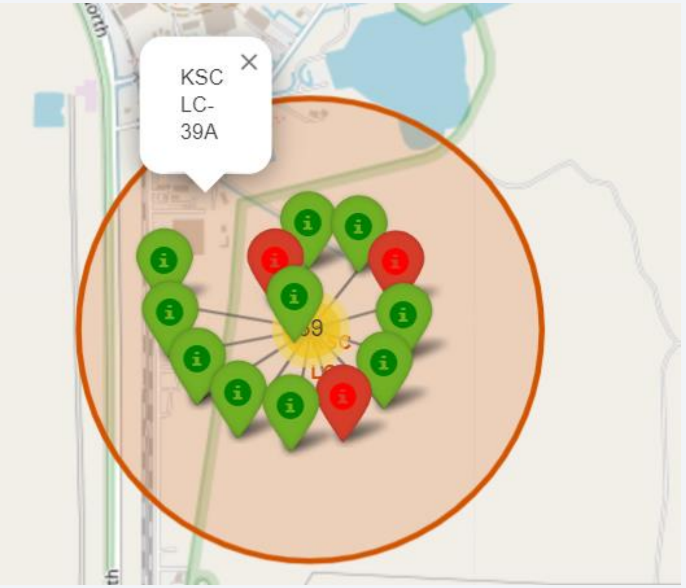
SpaceX Launch Sites

The locations marked in red on the map are SpaceX launch sites on the east coast of the US in Florida and the west coast in California

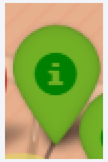
Success/Failed Launches for each Site

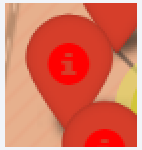




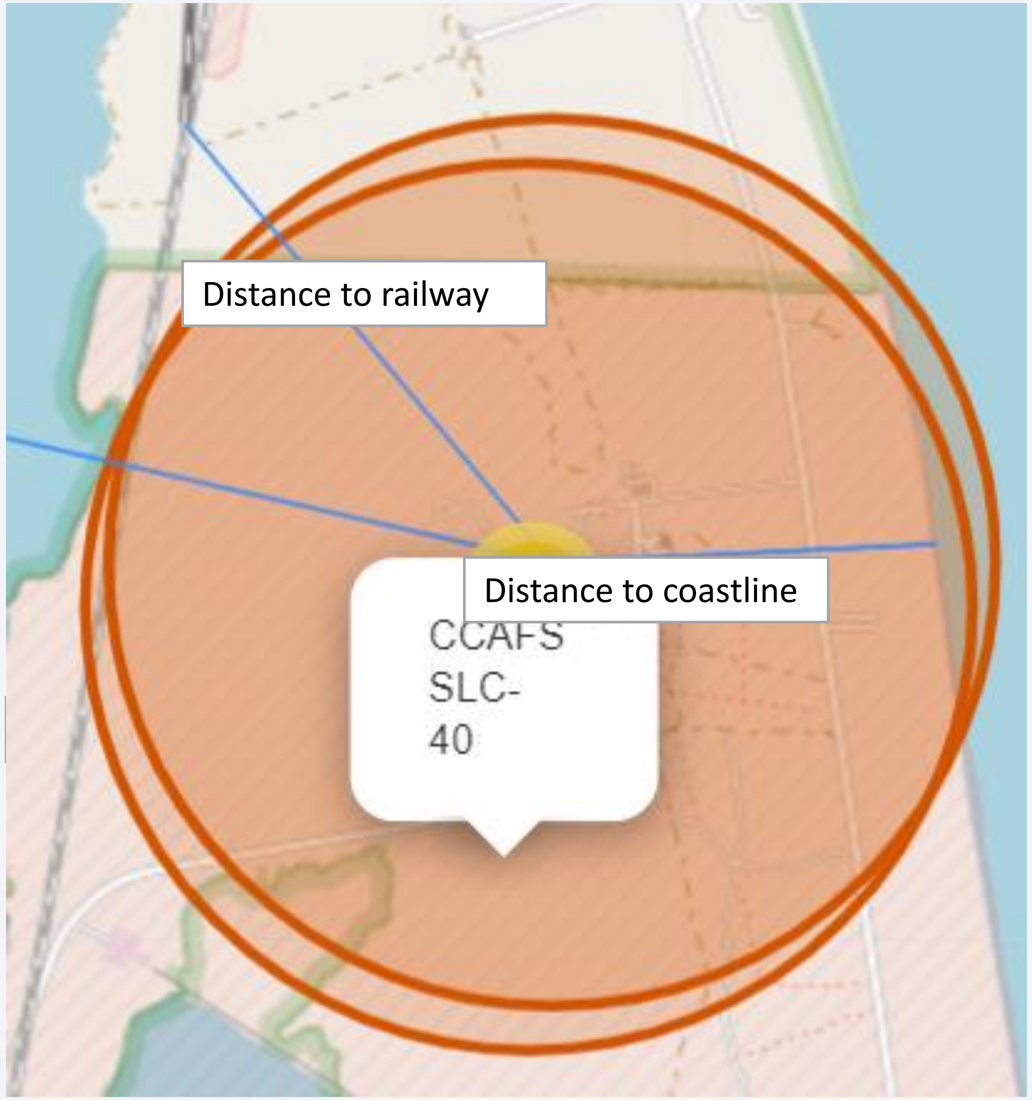


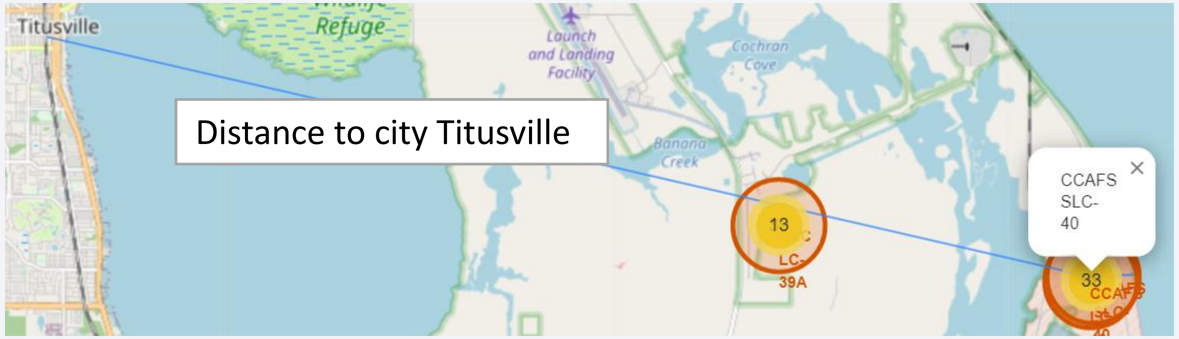
The figures represent map location of the four launch sites as labeled. Green indicates successful launches and red, the failures.

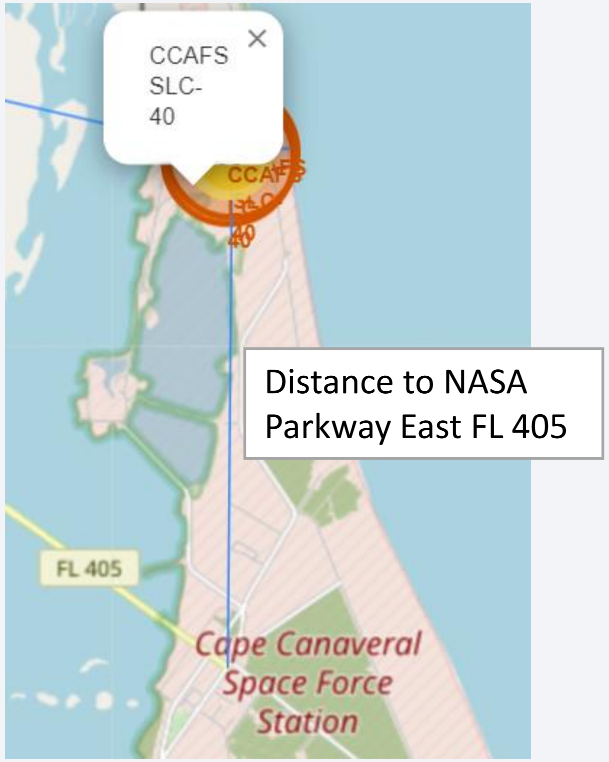

Successful Launch


Failed Launch

Launch Site Proximities to landmarks







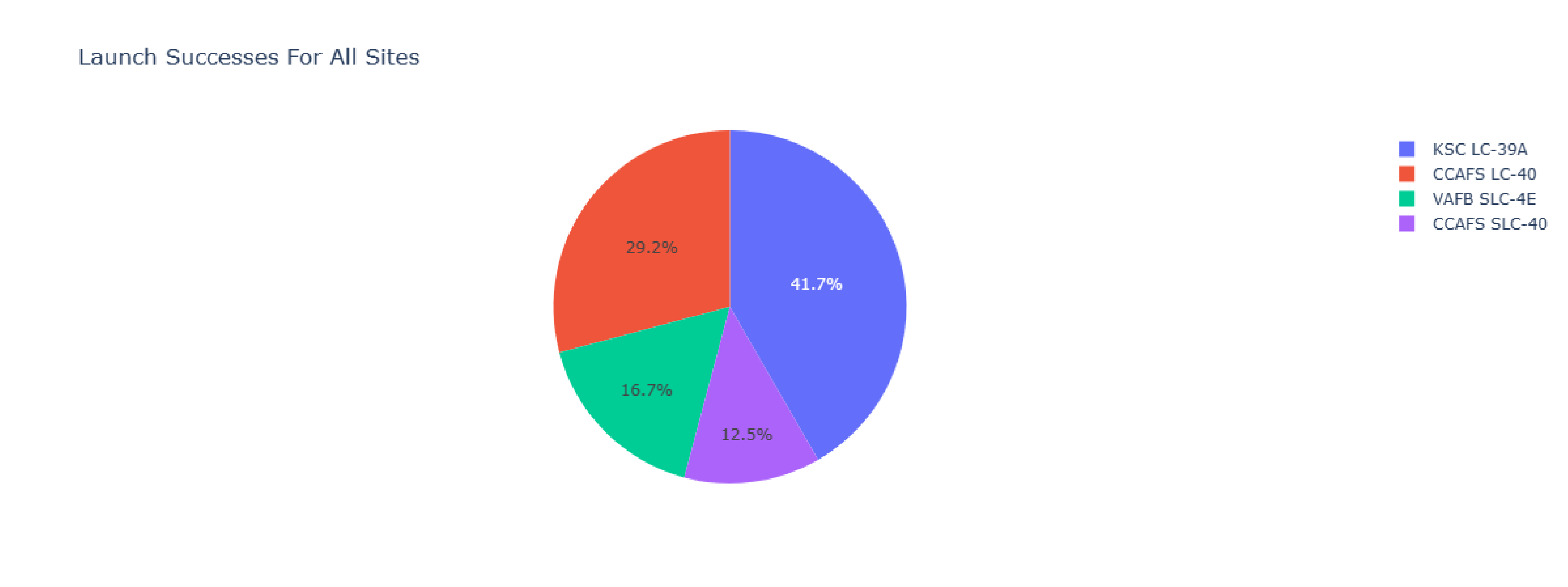
The figures above show distances between launch site CCAFS SLC-40 and nearby landmarks such as railway, highway, and coastline.

Section 4

Build a Dashboard with Plotly Dash

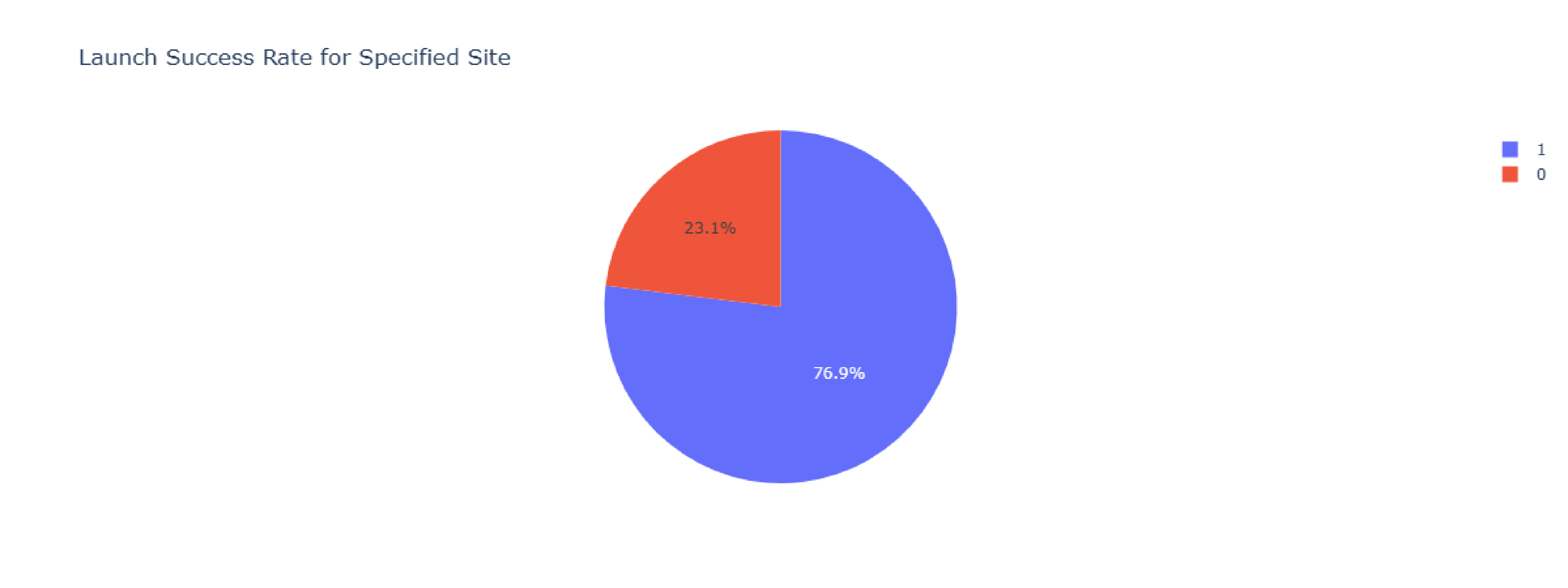
Launch Successes by Sites

- The pie chart represents launch success count for all sites represented as a percentage of all success.
- We can see that most of the successful launches occurred at site KSC LC-39A



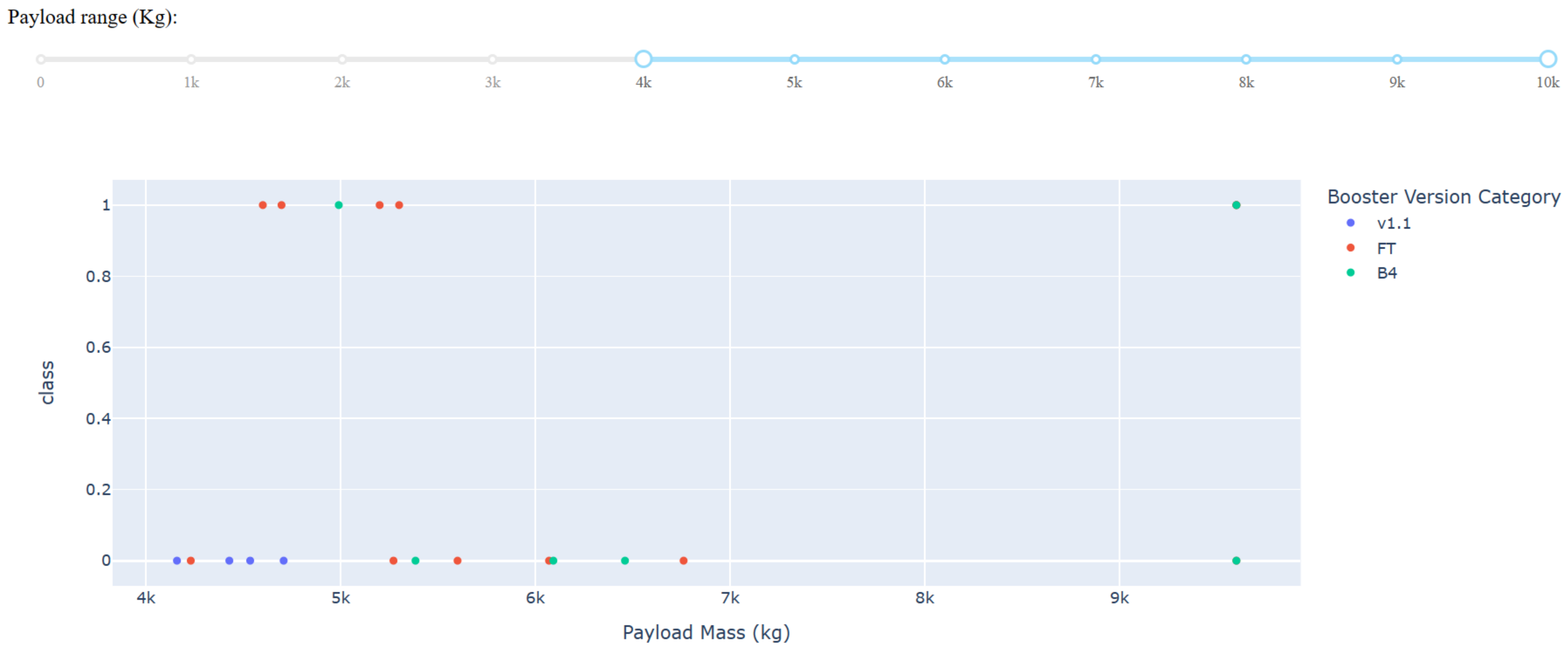
Site off Highest Success Ratio

The piechart the success rate for the launch site with highest launch success ratio, KSC LC-39A.



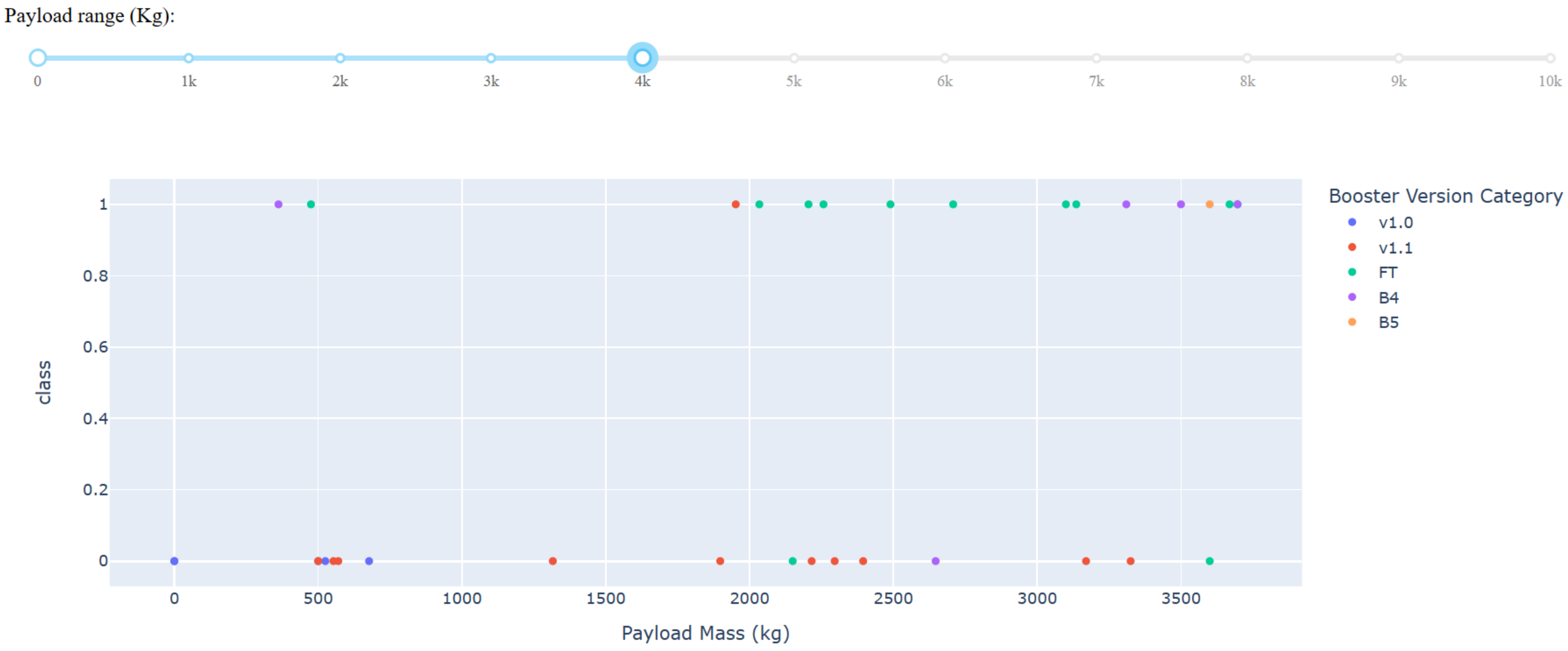
High Payload Mass Success Ratio

- Scatter plot of Payload vs. Launch Outcome for high payload mass (between 4000 and 10000 kg), with successes and failures distinguished by booster versions.
- High payload mass has high success ratio, particularly for booster version v1.1 and B4.



Low Payload Mass Success Ratio

- Scatter plot of Payload vs. Launch Outcome for low payload mass (between 0 and 4000kg), with successes and failures distinguished by booster versions.
- Low payload mass has low success ratio, particularly for booster version FT and B4 between 2000 and 4000 kg.



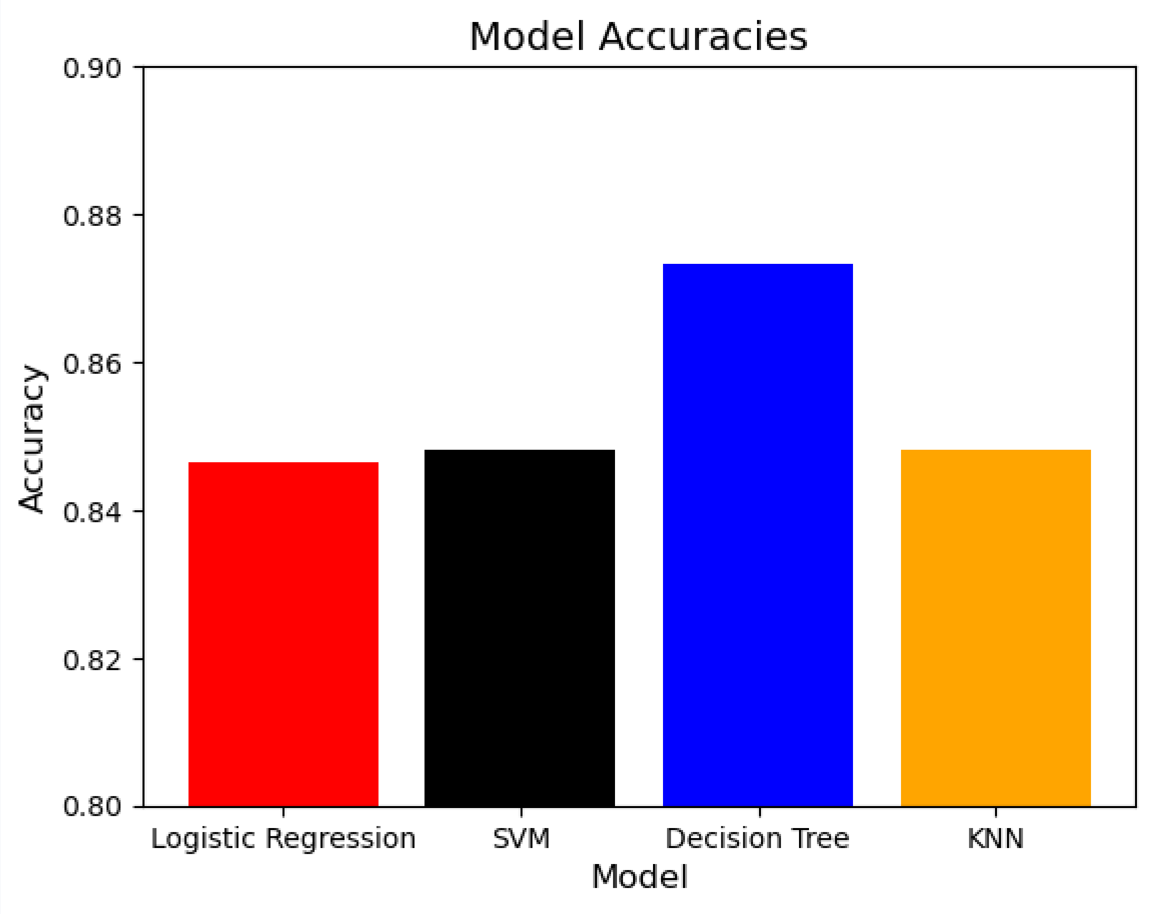
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The bar chart displays the accuracy for all the classification models built.

- Decision Tree has the highest accuracy from the validation set.

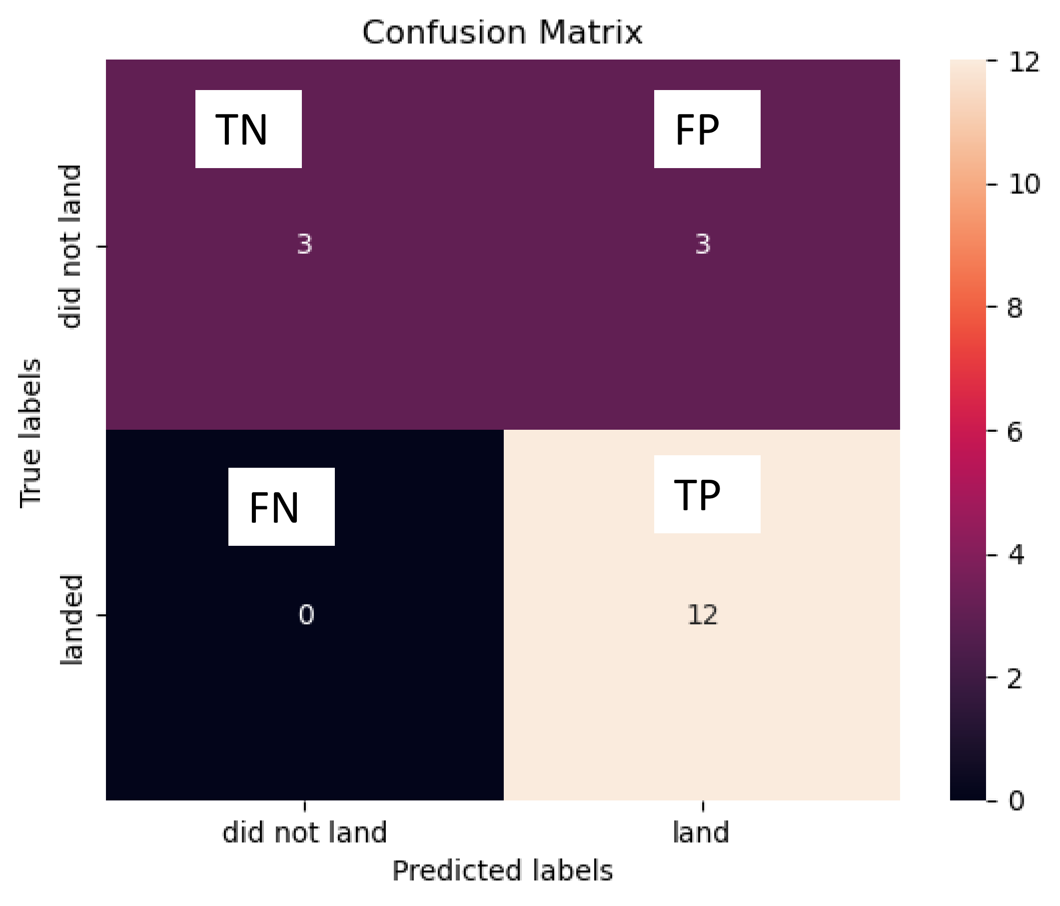


Confusion Matrix

- Figure showing the confusion matrix for the decision tree classification model.
- 15 launces were predicted to land successfully while 12 actually did.

True positive = 12

False positive = 3



Conclusions

- We can conclude that launch success rate vary by payload mass, launch site, and orbit.
  - Success rate higher with higher payload mass
  - Orbit sites ES-L1, GEO, HEO, SSO have consistently more successful outcomes
  - A launch is more likely to succeed at launch site KSC LC-39A
  - Success rate tends to increase over time
- A decision tree classification model can be used to accurately predict whether or not a launch will succeed or not

Appendix

Links to python notebooks for project:

https://github.com/eddyson8/IBM_DS_Capstone/tree/main

Thank you!