IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Saidat Ibiribigbe
July 16, 2022

Outline

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

Executive Summary

- The purpose of this project is to explore SpaceX data to predict whether a Falcon 9 launch will land successfully and the costs associated with each landing.
- Exploratory data analysis was performed using various tools in Python. Data visualizations were created to analyze trends across launch sites. Geospatial visualizations were also used to draw conclusions on the locations of various launch sites.
- For predictive modeling, four supervised learning algorithms were explored and the best algorithm was suggested based on its predictive power.

Introduction

- To better understand the different factors that influence Falcon 9 launches and identify potentially important factors that will be useful in predicting launch success or failure, we will be exploring launch data gathered from SpaceX REST API and Wikipedia.
- The information gained from the data exploration will be used to explore various modeling algorithms to determine the best model that presents the most accurate prediction for the proposed business problem.
- Drawing insight from modeling results, we will discuss potential process improvements and alternate methods are proposed at the end of the presentation.

Section 1

Methodology

Methodology

Executive Summary

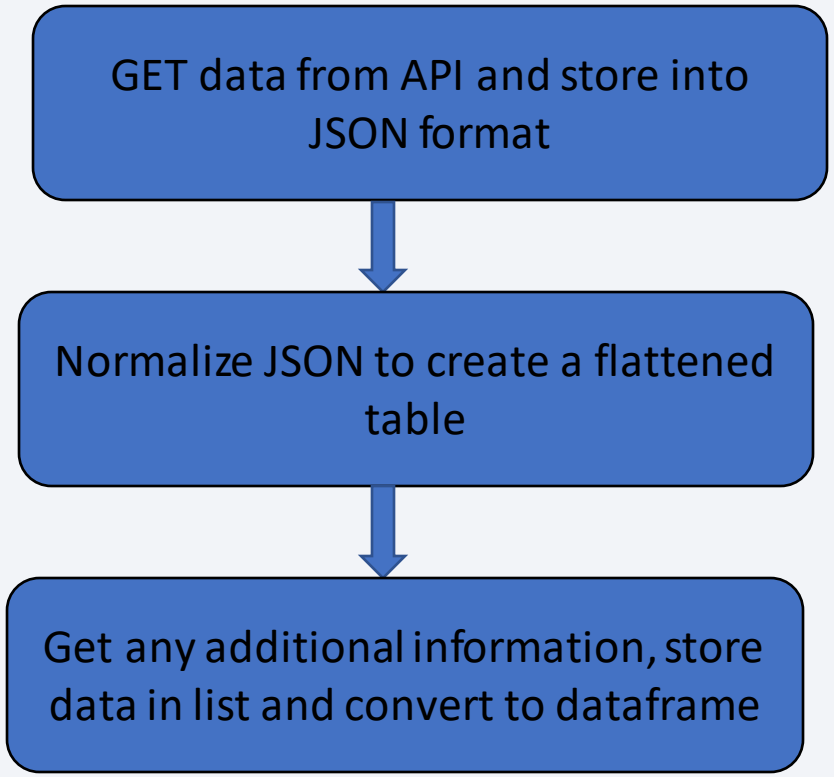
- Data collection methodology:
  - Data was collected from SpaceX REST API. Additional data was also collected using web scraping
- Perform data wrangling
  - Launch data was further filtered to only include Falcon 9 launches
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was preprocessed, then split into training and test sets. Four supervised learning algorithms were explored and ranked based on their predictive accuracy

Data Collection

- Datasets were collected from SpaceX Rest APIs. api.spacexdata.com/v4/launches/past
- Additional data that referenced Falcon 9 or Falcon 9 heavy launches were also scraped from Wikipedia. https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

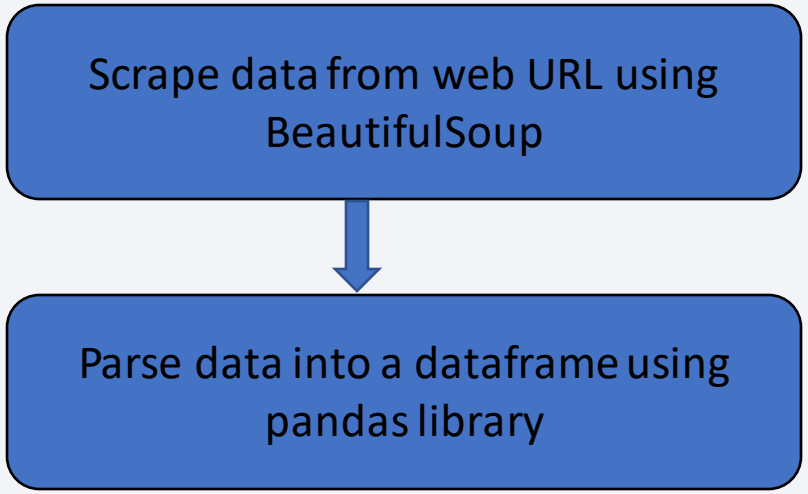
Data Collection – SpaceX REST API

- Data was retrieved using get requests from the REST API at the following URL (api.spacexdata.com/v4/launches/past). Data was then further filtered and preprocessed to only include Falcon 9 launches.
- Notebook URL: (https://github.com/Sayyddah/SpaceX-Project/blob/master/SpaceX%20-%20Data%20Collection%20API.ipynb)



Data Collection - Scraping

- Data was also scraped from from wikepedia URL (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches) using the BeautifulSoup library and its contents were parsed into a pandas data frame.
- Notebook URL: https://github.com/Sayyddah/SpaceX-Project/blob/master/SpaceX%20-%20Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- We filtered data from the REST API to only retain Falcon 9 launch data.
- The replace function was used to replace the missing values in the payload column with the mean of the column.
- The dataset was analyzed for missing values. Number of launches per site was also determined.
- Notebook URL: https://github.com/Sayyddah/SpaceX-Project/blob/master/SpaceX%20-%20Data%20Wrangling.ipynb

EDA with Data Visualization

- Relationships between flight number and payload mass, flight number and launch site, as well the success rate by orbit type were explored. Various visualizations were produced including scatterplots, bar graphs and line charts which we discuss in the Insights drawn section of this presentation.
- Notebook URL: https://github.com/Sayyddah/SpaceX_Project/blob/master/SpaceX%20-%20EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Several queries were ran on the SpaceX data to include:
  - List unique launch sites
  - Display total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Rank the count of landing outcomes between the date June 4, 2010 and March 20, 2017 in descending order
- Notebook URL: https://github.com/Sayyddah/SpaceX-Project/blob/master/SpaceX%20EDA%20-%20SQL.ipynb

Build an Interactive Map with Folium

- Several map objects such as markers, circles, lines, etc. were created and added to a map using folium
- These objects were added to identify launch site locations and assess their proximities to significant map locations like cities, coastlines, highways and railway.
- Notebook URL: https://github.com/Sayyddah/SpaceX-Project/blob/master/SpaceX%20-%20Folium_%20Interactive%20Visual%20Analytics.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard was created in Dash that included pie charts, drop downs, sliders and scatterplots
- These visualizations were used to analyze the changes in success rate across different launch sites, payload mass and booster versions.
- Notebook URL: https://github.com/Sayyddah/SpaceX-Project/blob/master/SpaceX%20-%20Dash%20App.py

Predictive Analysis (Classification)

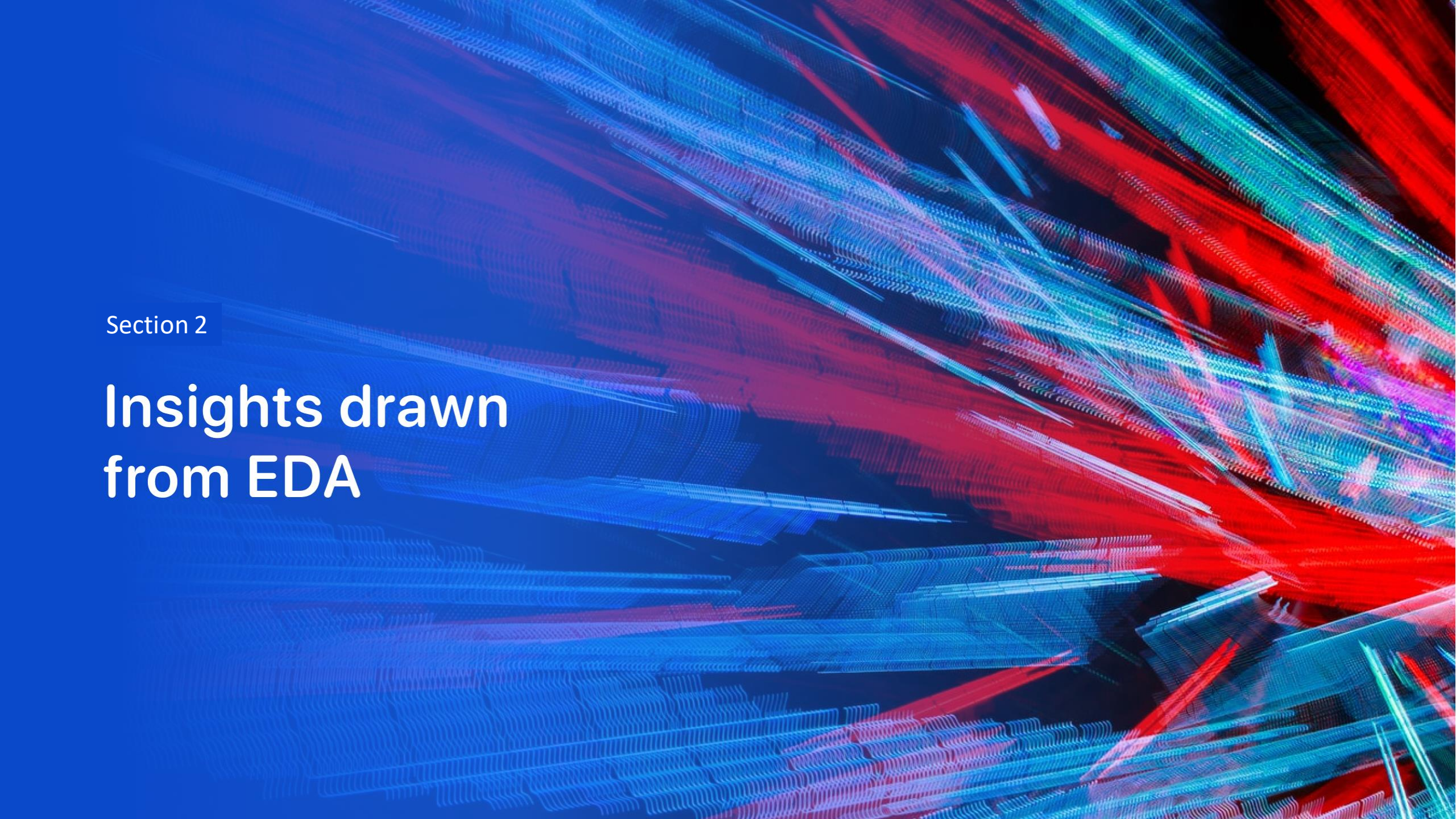
- The input variable data was scaled and preprocessed.
- The data was split into training and test datasets in using the 80/20 rule.
- Four supervised learning algorithms were explored: Logistic Regression, Support Vector Machine, Decision Trees, and K-Nearest Neighbors.
- GridSearchCV was used to select the best parameters for each algorithm. The algorithm was ran on the training model and the accuracy score was determined on the test model to evaluate each model's predictive performance.
- Notebook URL: https://github.com/Sayyddah/SpaceX-Project/blob/master/SpaceX%20-%20First%20Stage%20Landing%20Prediction.ipynb

Results

- From the outcome of the data analysis, we can say that the most important factors that influence the success of a launch include but are not limited to: launch site, payload mass and launch year.
- From our predictive analysis, for future launch predictions, a decision tree model should be used. This is because its high test accuracy score of 89% provides us with the best chance of accurately predicting a launch outcome.

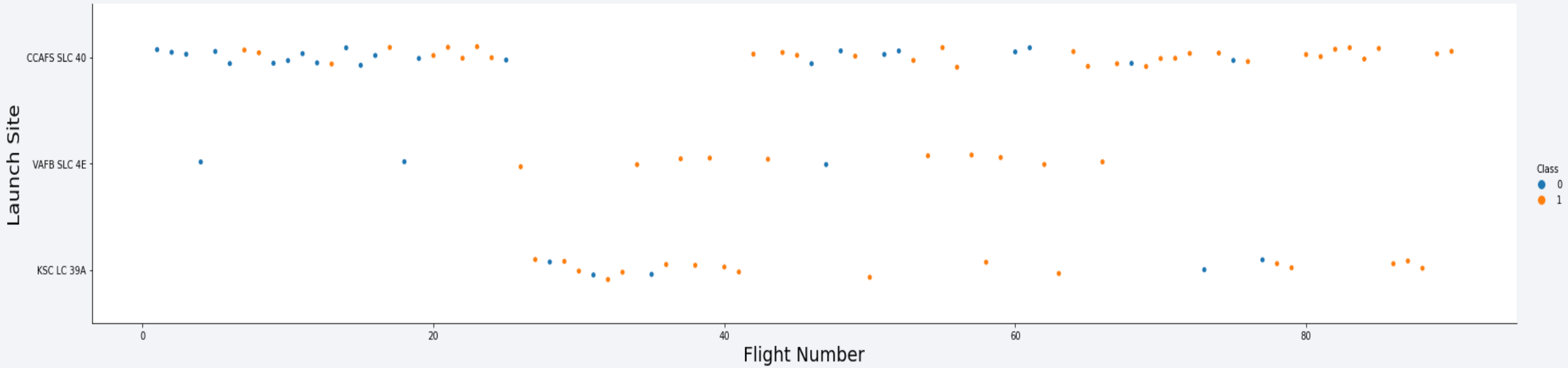
Section 2

Insights drawn from EDA



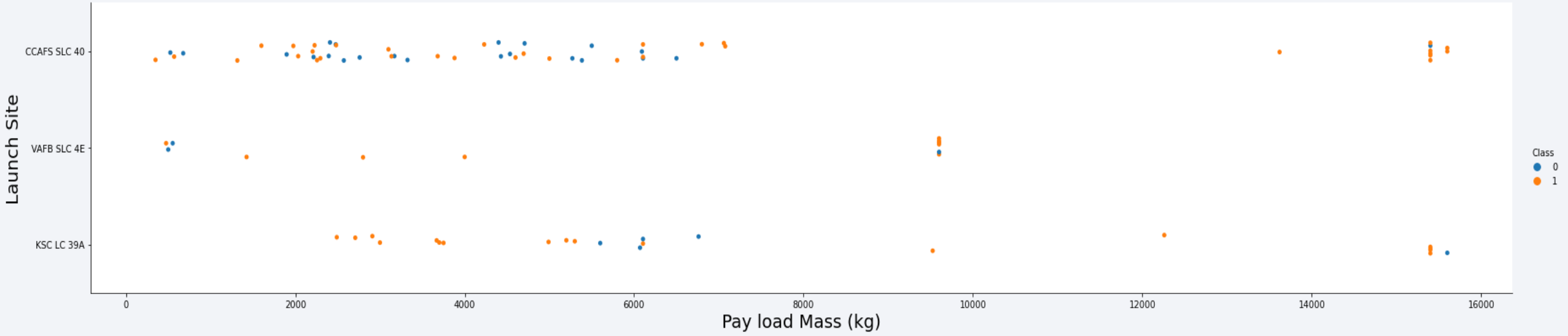
Flight Number vs. Launch Site

From the plot below, we observe that KSC LC 39A has higher flight numbers which means that, launches did not start at this site until a later time. CCAFS SLC 40 has the most flights recorded.



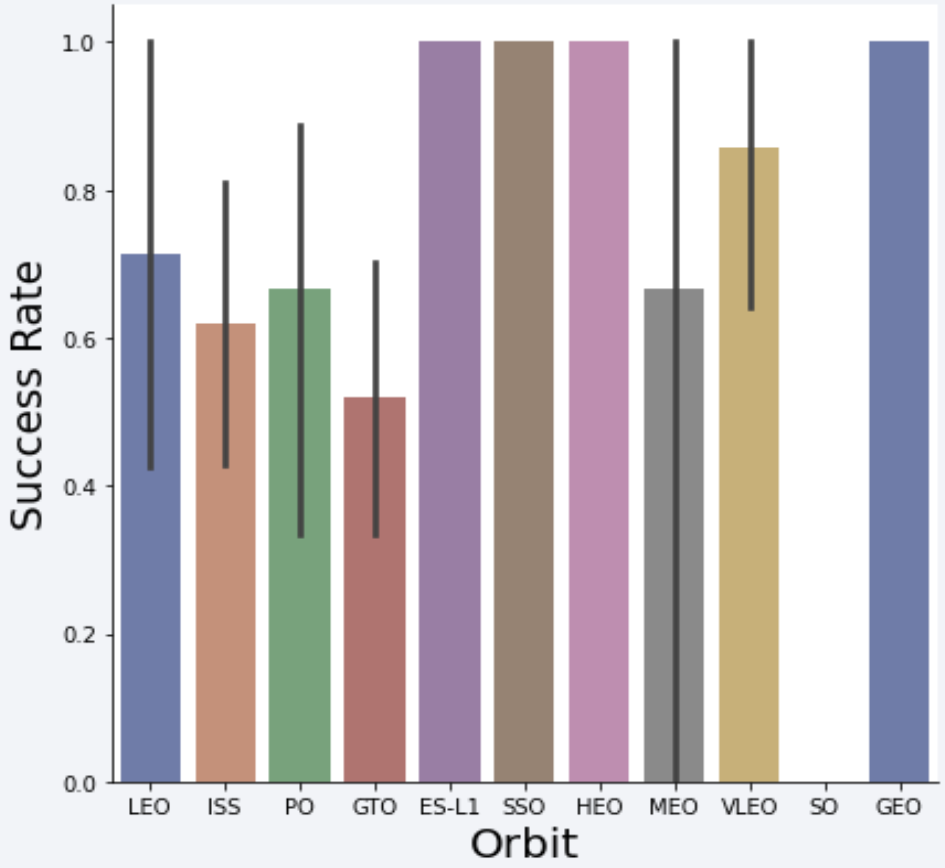
Payload vs. Launch Site

From the plot below, we observe that the VAFB-SLC launch site does not record a launch with a payload mass above 10,000 kg. Additionally, for all launch sites there are fewer launches recorded with a payload mass higher than 10,000 kg.



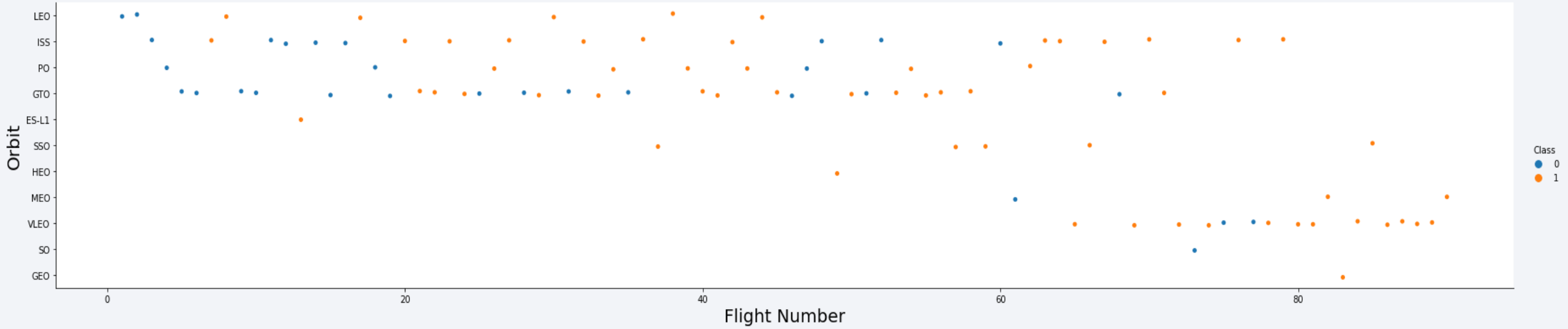
Success Rate vs. Orbit Type

- As seen in the bar graph, four orbits – GEO, HEO, SSO, ES-L1, all have a 100% success rate. This could be due to the fact that there are fewer launches in these orbits or launches in the orbit occurred at a later time, increasing the chances of success.



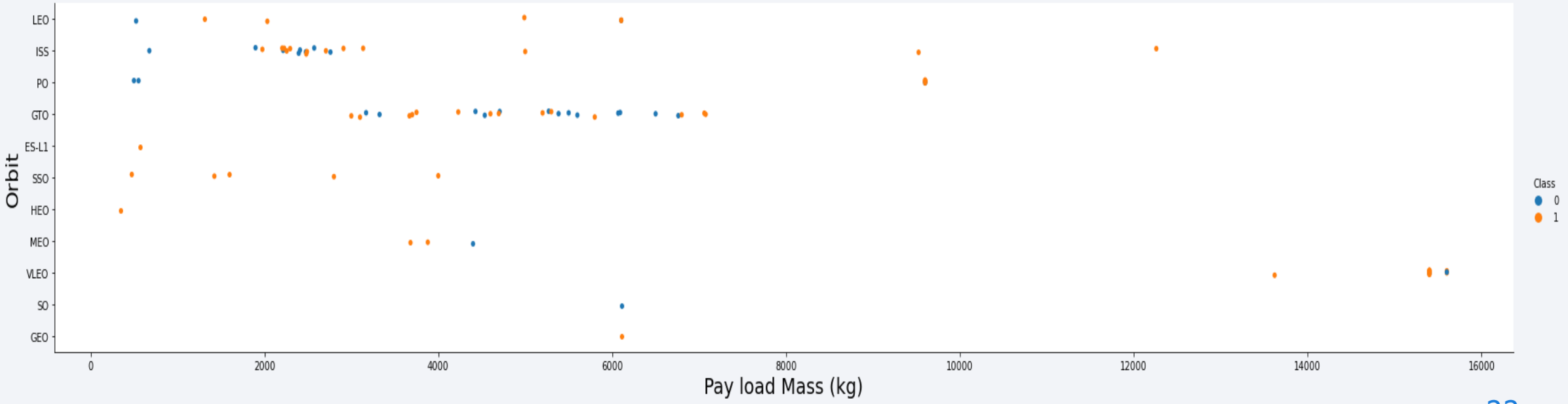
Flight Number vs. Orbit Type

From the plot below, we see that have GEO, SO, VLEO, MEO, HEO, SSO launch sites have higher flight numbers which means that they were launched at a later time. These flights also seem to be more likely to be successful. This could be due to be the fact that lessons learned in earlier failed launches are applied in later launches to improve the probability of success.



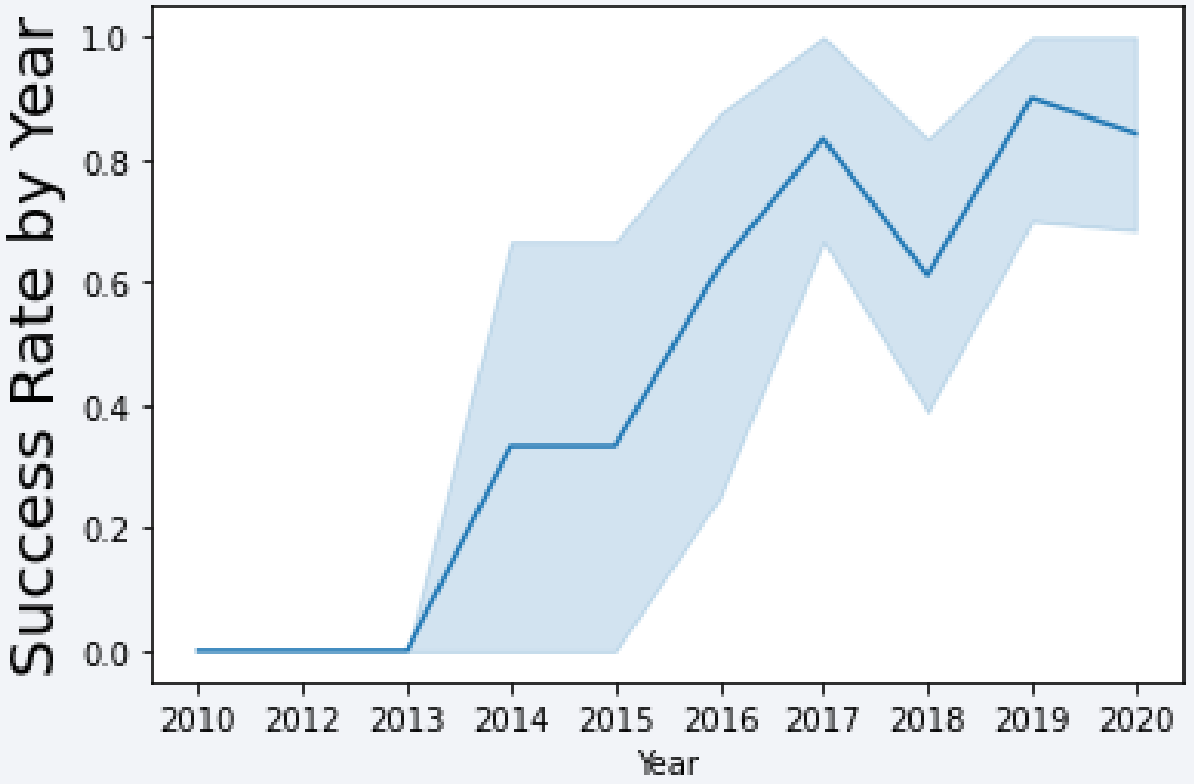
Payload vs. Orbit Type

From the plot below, we observe that most orbit type recorded lower payload masses. Additionally, higher payload masses do not seem to have any impact to success rates with regards to orbit type.



Launch Success Yearly Trend

The line plot of
the success rate by year shows
a steady increase
in launch success rate between
2013 with the exception of a
20% decrease in success rate
in from 2018 to 2019.



All Launch Site Names

- Query to find unique names of launch sites:
  - `%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX`
- Results:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
  - Query: `%sql SELECT * FROM SPACEX WHERE UPPER(LAUNCH_SITE) LIKE 'CCA%' LIMIT 5`
- Results:

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

- Calculate the total payload carried by boosters from NASA
  - Query: `%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)'`
- The total payload mass by NASA boosters is 45,596 kilograms.

1
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - Query: `%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1'`
- The average payload mass carried by booster version F9 v1.1 is 2,928 kilograms.

1
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - Query: `%sql SELECT MIN(DATE) FROM SPACEX WHERE LANDING__OUTCOME = 'Success'`
- The date of the first successful landing outcome in ground pad was July 22nd, 2018.

1
2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - Query: `%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE LANDING__OUTCOME = 'Success' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000`
- There are 8 booster versions that successfully landed on drone ship with payload mass between 4000 and 6000 kilograms. They are all in the B5 category.

booster_version
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
  - Query: `%sql SELECT MISSION_OUTCOME, COUNT(*) AS Outcome_Count FROM SPACEX GROUP BY MISSION_OUTCOME`
- From the output, we see that all missions were recorded as successful with the exception of one attempt.

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
  - Query: `%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX)`
- A sub-query that first calculates the maximum payload mass in the table was used to generate this output. 12 launches were reported.

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Query: `%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE (LANDING__OUTCOME LIKE '%Failure%' AND YEAR(DATE) = 2015)`
- There were two drone ship landing outcome failures recorded in 2015. Both occurred at the CCAFS LC-40 launch site.

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

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

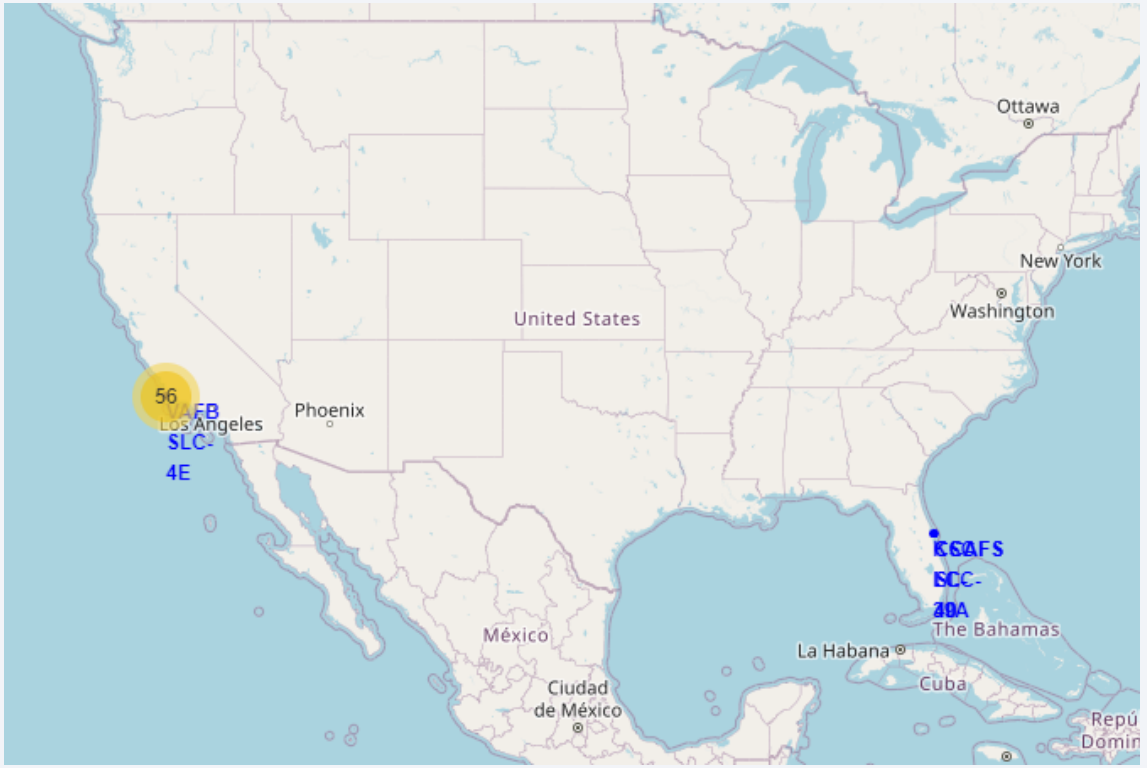
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
  - Query: `%sql SELECT LANDING__OUTCOME, COUNT(*) AS NumofOccurence FROM SPACEX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY NumofOccurence DESC`
- The query resulted in seven groups, with "No attempt" landing outcome having the highest number of occurrence at 10.

landing__outcome	numofoccurence
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

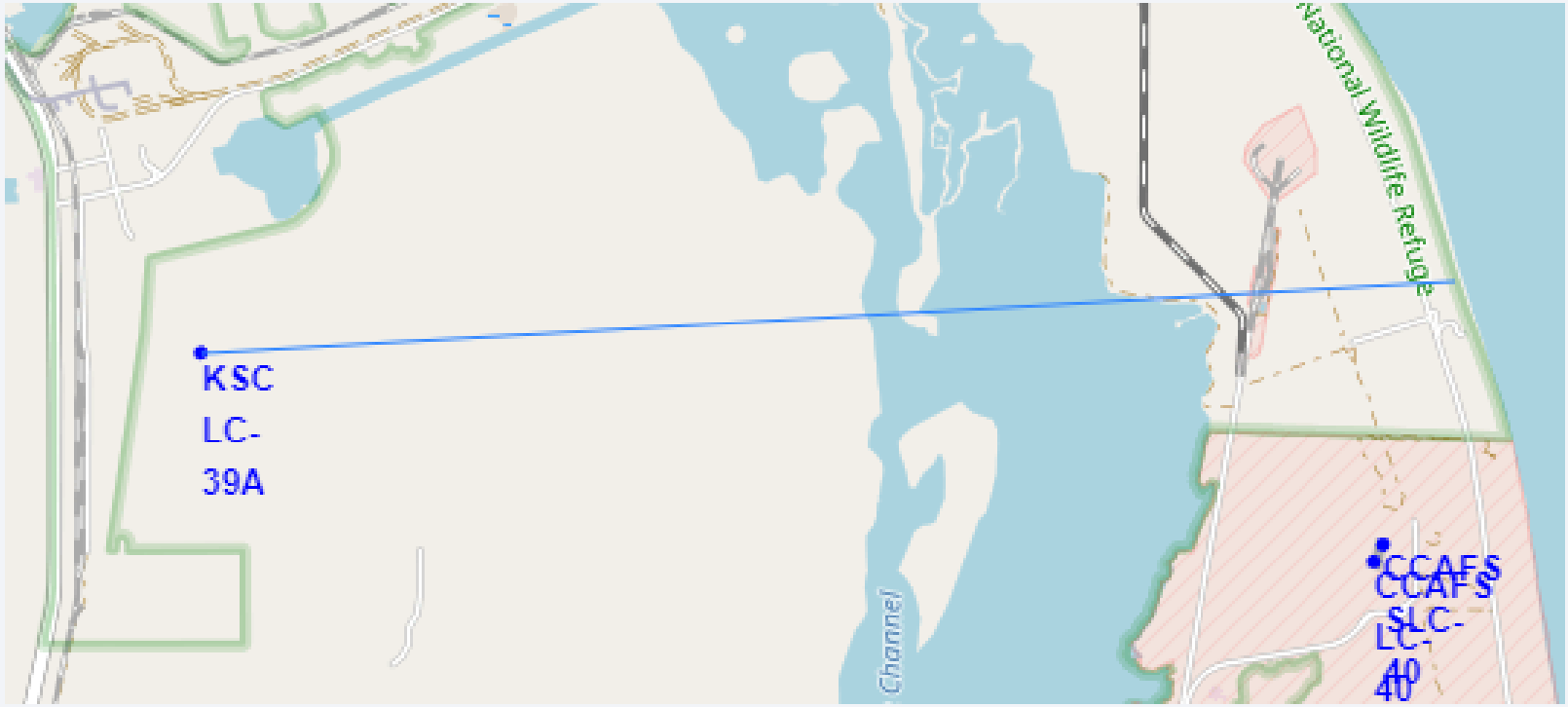
Launch sites with markers

- Below is a map showing the location of the launch sites with markers and color included



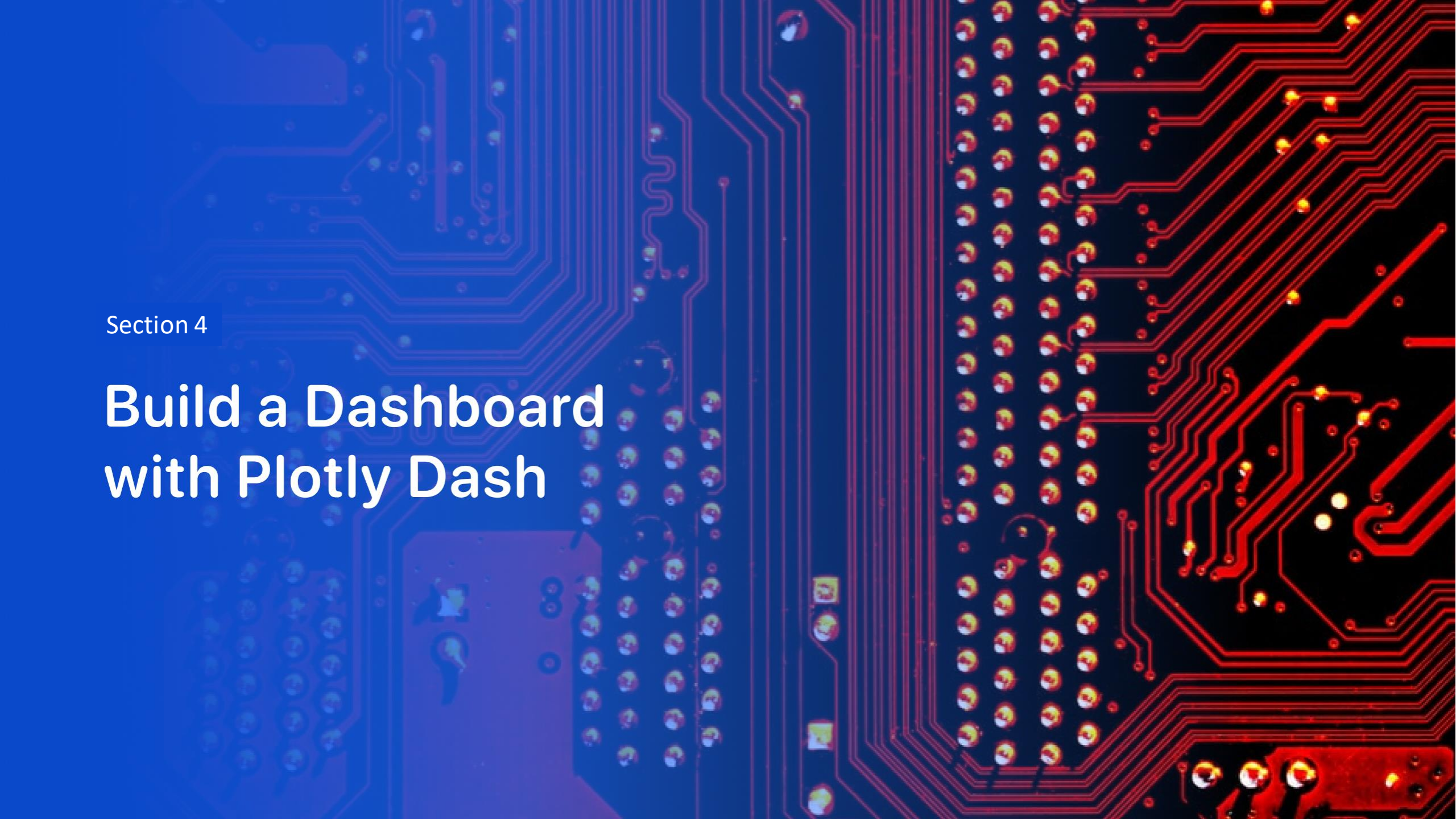
Distance from KSC to closest coastline

- The map below shows the distance from KSC LC-39A launch site to the closest coast line. The distance from the launch site to the coast line is about 7.3 kilometers. We also observe that the CCAFS launch sites are closer to the coastline.



<Folium Map Screenshot 3>

- The map below shows the distance from KSC LC-39A launch site to the to various geographic locations. We observe that the furthest distance is between the launch site and the Titusville city line with a distance of about 16.2 kilometers.

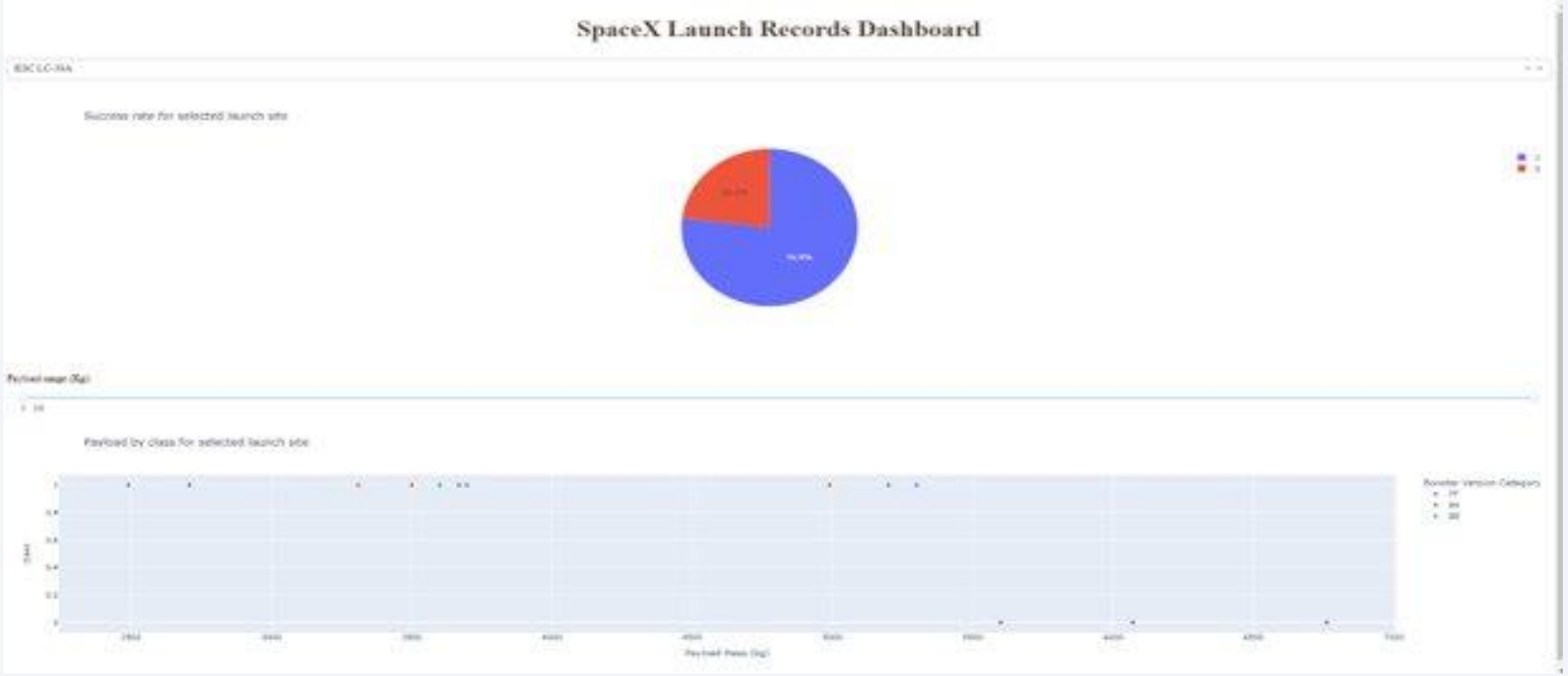

Section 4
Build a Dashboard with Plotly Dash

Success count for all launch sites

- The pie chart below shows the launch success count for all sites. We observe that the KSC LC-39A has the highest success count, implying that launches at the site are the most likely to be successful.

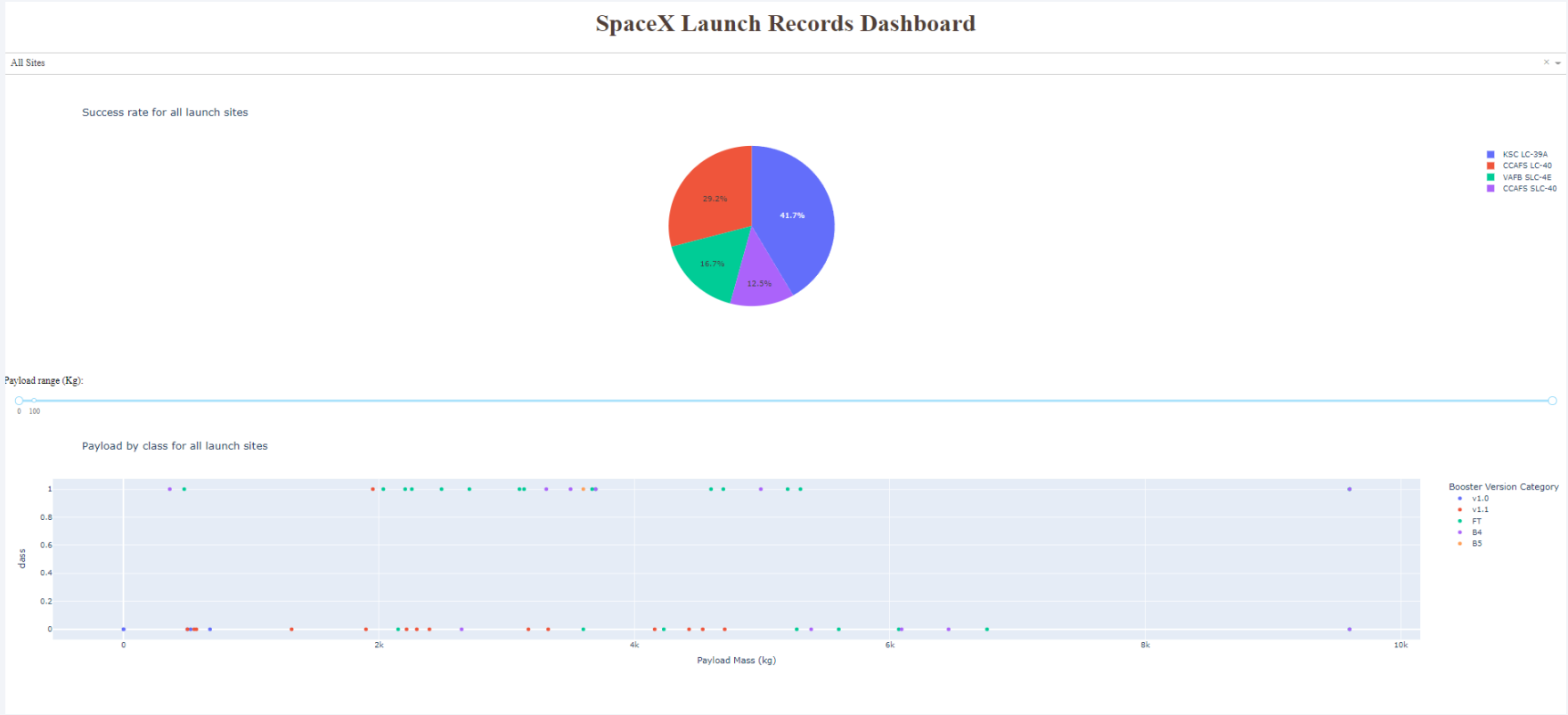
KSC LC-39A Success rate

- Upon further exploration, we observe that launch site also KSC LC-39A has the highest success rate (77%) with success more likely in payloads mass between 2500 and 5500 kg.



Payload vs. Launch Outcome

Analyzing the launch data for all sites, we observe that booster version FT has higher success rate across all payload ranges. We also observe that launches with payload mass less than about 5500 kilograms are more likely to be successful.

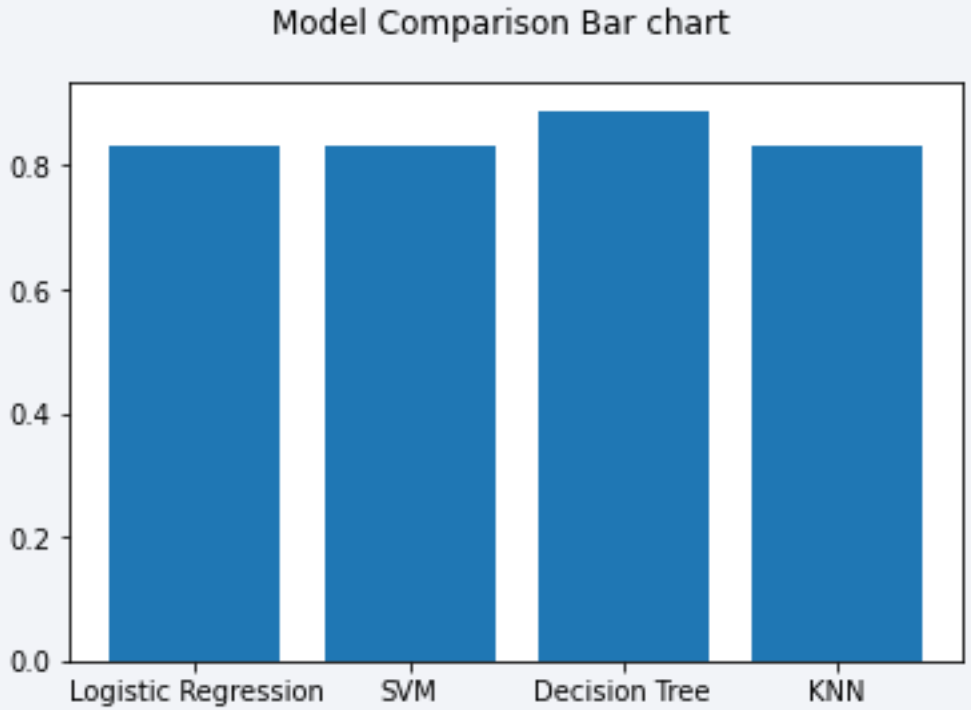


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- As seen on the bar chart on the right, we observe that the decision tree model produces the highest accuracy making it the best model to use. It is worthy of note that the accuracy score on all models are fairly high.



Confusion Matrix

The decision tree has the highest test accuracy with a value of 89%. Looking at it's confusion matrix, we see that it is very good at predicting landings. 2 launches that did not land were however correctly misclassified as landed.

Conclusions

- From the data analysis, we can say that the most important factors that influence the success of a launch include but are not limited to: launch site, payload mass and launch year.
- With this information, for future launch predictions, a decision tree model should be used. This is because its high test accuracy score 89% provides us with the best chance of accurately predicting launch outcome.
- Since we only explored supervised learning models for this study, an improvement in prediction accuracy can be achieved by exploring more robust techniques like random forests or gradient boosting. We can also explore unsupervised learning algorithms to help find more hidden and actionable insights and relationship in our data.

Appendix

Below is a code snippet used to create the bar chart displaying the accuracy of the classification models.

```
models = []
names = []
outcome = []
parameters = []
y_pred = []

models.append(('Logistic Regression', lr, parameters_lr))
models.append(('SVM', svm, parameters_svm))
models.append(('Decision Tree', tree, parameters_tree))
models.append(('KNN', knn, parameters_knn))

for model_name, model, parameter in models:
    cv = GridSearchCV(model, cv = 10, param_grid = parameter)
    fit = cv.fit(X_train, Y_train)
    yhat = cv.predict(X_test)
    result = cv.score(X_test, Y_test)
    outcome.append(result)
    names.append(model_name)
    y_pred.append(yhat)

fig = plt.figure()
fig.suptitle('Model Comparison Bar chart')
ax = fig.add_subplot(111)
plt.bar(names, outcome)
ax.set_xticklabels(names)
plt.show()
plot_confusion_matrix(Y_test,y_pred[2])
outcome
```

Thank you!