<SpaceX Falcon 9 Project>

<Hima Tummalapalli>

<05/15/2024>

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- We want to be able to predict if the first stage of a launch will land successfully, so that we can use that information to decide how to bid against our competitor, SpaceX, for a rocket launch.
- **Methodology:**
    - Use the data on the SpaceX website, collecting it through the SpaceX API or through webscraping
    - Explore the data for relations, patterns, and trends using SQL, Pandas, and Matplotlib
    - Visualize the data in an interactive way using Folium and Plotly Dash
    - Train some ML models using Logistic Regression, Support Vector Machines, Decision Trees, and K-Nearest Neighbors using our data to make predictions on rocket launches
- **Findings:**
    - Most successful launch site: KSC LC-39A; least successful site: CCAFS LC-40; most used site: CCAFS SLC-40
    - Most successful orbit with more than one launch: SSO
    - Between June 2010 and June 2017: most successful landing method: ground pad; most used landing method: drone ship; many no attempts
    - Most successful booster version: FT; least successful booster version: v1.1
    - KSC LC-39A's launches contained: Number of flights between 20 and 80, Payload mass above 2500 kg
    - Best ML model algorithms to use: Support Vector Machines, K-Nearest Neighbors
- **Further Actions:**
    - Analyze SpaceX data more thoroughly for clearer insights
    - Hone the SVM or KNN model to be more accurate

INTRODUCTION

- We are working for a company called SpaceY that is a competitor of SpaceX
- SpaceY wants to figure out whether to bid against SpaceX for a rocket launch
- We will analyze the data on SpaceX's Falcon9 to help SpaceY make that decision
    - We will try to predict using SpaceX's data whether the first stage of a rocket will land because that is what determines the cost of a launch
    - The first stage of the rocket propels the rocket into orbit and returns and lands the rocket back on Earth, carrying the second stage (and payload) as it launches
    - We refer to the rockets as boosters

DATA COLLECTION METHODOLOGY

Using the SpaceX API

- Requested information on past launches using the get method from the SpaceX API
- Requested further information, such as booster version, launch site, etc., using the previously retrieved information, and created a DataFrame
- Removed Falcon 1 information from the DataFrame and replaced PayloadMass missing values with mean

<u>Using Webscraping</u>

- Webscraped information from an HTML table on Wikipedia containing launch records using BeautifulSoup by
    - Extracting the column names by iterating through the table headers and created a DataFrame using them
    - Iterating through the rows of the table and added each of the values to the corresponding column in the DataFrame

DATA WRANGLING METHODOLOGY

Did some Data Analysis such as

- Finding out the data types of each column
- Calculating the percentage of missing values in each attribute using isnull() and dividing sum by length of DataFrame
- Using value_counts()
  - Counting the number of launches on each site, number and occurrence of each orbit, and number of and occurrence of mission outcome

Created a landing outcome label in the DataFrame where 0 indicates a bad outcome and 1 indicates a good outcome for each record

EXPLORATORY DATA ANALYSIS METHODOLOGY

Using SQL Queries:

- Finding the unique launch sites using the DISTINCT keyword
- Displaying 5 records with launch sites beginning with 'CCA' using the LIKE keyword
- Displaying the total payload mass carried by boosters launched by the customer NASA (CRS) using the SUM() function
- Displaying the average payload mass carried by booster version F9 v1.1 using the AVG() function
- Finding Date of first successful landing outcome in ground pad using ORDER BY and LIMIT
- Listing Booster Version that have success in drone ship with payload mass greater than 4000 but less than 6000 using AND and BETWEEN
- Counting the total number of successful and failed mission outcomes using COUNT() and OR
- Listing Booster Version that carried the maximum payload mass using a subquery and MAX()
- Listing month, landing outcome, booster version, and launch site for launches in 2015 and landing outcome of failure in drone ship using substr()
- Ranking count of landing outcomes between June 4, 2010 and March 20, 2017 in descending order using BETWEEN, GROUP BY, ORDER BY, and DESC

EDA WITH SQL RESULTS

```
%sql select DISTINCT("Launch_Site") from SPACEXTABLE
```

```
* sqlite:///my_data1.db
Done.
```

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

CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40 are the 4 Launch Sites in the SPACEXTABLE

EDA WITH SQL RESULTS

```
%sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5
```

```
* sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

The 5 results displayed are from Launch Site CCAFS LC-40 (since it begins with CCA). We can see that all 5 launches were successes. We can also see the information regarding each record's payload, orbit, customer, payload mass, etc. Note: some of the feature values have been excluded from the picture due to limited space. Interestingly, 2 of the boosters' payload mass was 0 kg.

EDA WITH SQL RESULTS

```
%sql select SUM("PAYLOAD_MASS__KG_") from SPACEXTABLE where CUSTOMER == 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

SUM("PAYLOAD_MASS__KG_")
45596

The total payload mass carried by boosters launched by NASA (CRS) is 45596 kg.

EDA WITH SQL RESULTS

```
%sql select AVG("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" == 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

AVG("PAYLOAD_MASS__KG_")
2928.4

The average payload mass carried by boosters of version F9 v1.1 is 2928.4 kg.

EDA WITH SQL RESULTS

```
%sql select Date from SPACEXTABLE where "Landing_Outcome" == 'Success (ground pad)' order by Date limit 1
```

```
 * sqlite:///my_data1.db
Done.
```

Date
2015-12-22

The Date of the first successful landing outcome in ground pad is December 22, 2015.

EDA WITH SQL RESULTS

```
%sql select "Booster_Version" from SPACEXTABLE where "Landing_Outcome" = 'Success (drone ship)' and "PAYLOAD_MASS
```

```
* sqlite:///my_data1.db
Done.
```

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

F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2 are the booster versions that were successful in drone ship carrying a payload mass greater than 4000 but less than 6000.

EDA WITH SQL RESULTS

```
%sql select COUNT("Mission_Outcome") from SPACEXTABLE where "Mission_Outcome" like '%Success%' or '%Failure%'
```

```
 * sqlite:///my_data1.db
Done.
```

COUNT("Mission_Outcome")
100

There were 100 total mission outcomes that were either a success or failure.

EDA WITH SQL RESULTS

```
%sql select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select MAX("PAYLOAD_MASS__KG_") from
```

```
* sqlite:///my_data1.db
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

This list of 12 items contains the booster versions that have carried the highest payload mass.

EDA WITH SQL RESULTS

```
%sql select substr(Date, 6, 2), substr(Date, 0, 5), "Landing_Outcome", "Booster_Version", "Launch_Site" from SPAC
```

```
 * sqlite:///my_data1.db
Done.
```

substr(Date, 6, 2)	substr(Date, 0, 5)	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

These two records refer to launches in 2015 with a landing outcome of "failure in drone ship." You can see the month, year, landing outcome, booster version, and launch site for these launches. Both launches were from CCAFS LC-40.

EDA WITH SQL RESULTS

```
%sql select "Landing_Outcome", COUNT("Landing_Outcome") from SPACEXTABLE where Date between '2010-06-04' and '201
```

```
* sqlite:///my_data1.db
Done.
```

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

This list ranks the counts of all landing outcomes between June 4, 2010 and March 20, 2017 in descending order . "No attempt" has the highest count, which means there was no attempt to recover the first stage in 10 of the launches. "Precluded (drone ship)" has the lowest count, which means 1 launch was prevented from happening in the first place. More launches seemed to happen in the drone ship than in any other place, success or failure.

EXPLORATORY DATA ANALYSIS METHODOLOGY

Using Pandas and Matplotlib (with Visualization)

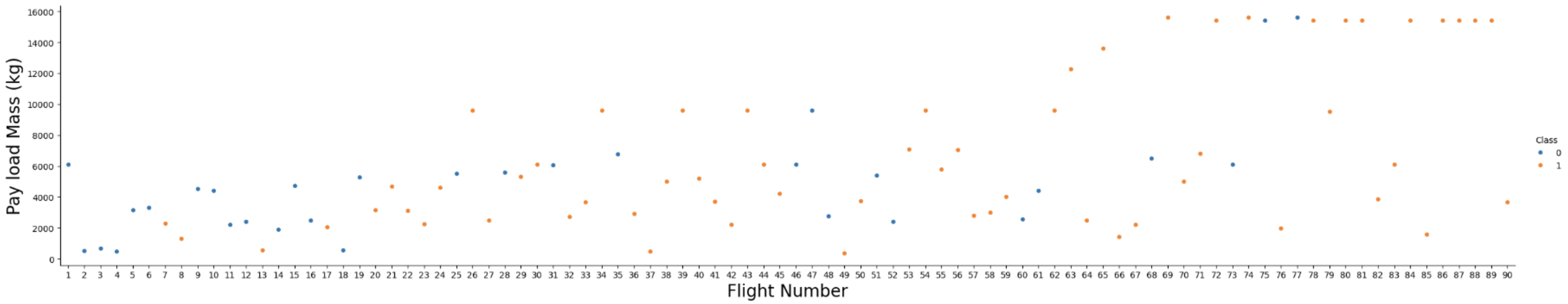
❿ Plotted various charts and graphs using Seaborn to understand relationships in the data:

- Scatterplot of Flight number vs Payload mass (catplot, hue="Class")
- Scatterplot of Flight number vs Launch Site (catplot, hue="Class")
- Scatterplot of Payload mass vs Launch Site (catplot, hue="Class")
- Bar chart of Orbit vs Class (barplot)
- Scatterplot of Flight number vs Orbit (catplot, hue="Class")
- Scatterplot of Payload mass vs Orbit (catplot, hue="Class")
- Line Chart for the Launch Success Yearly Trend

❿ Feature Engineering:

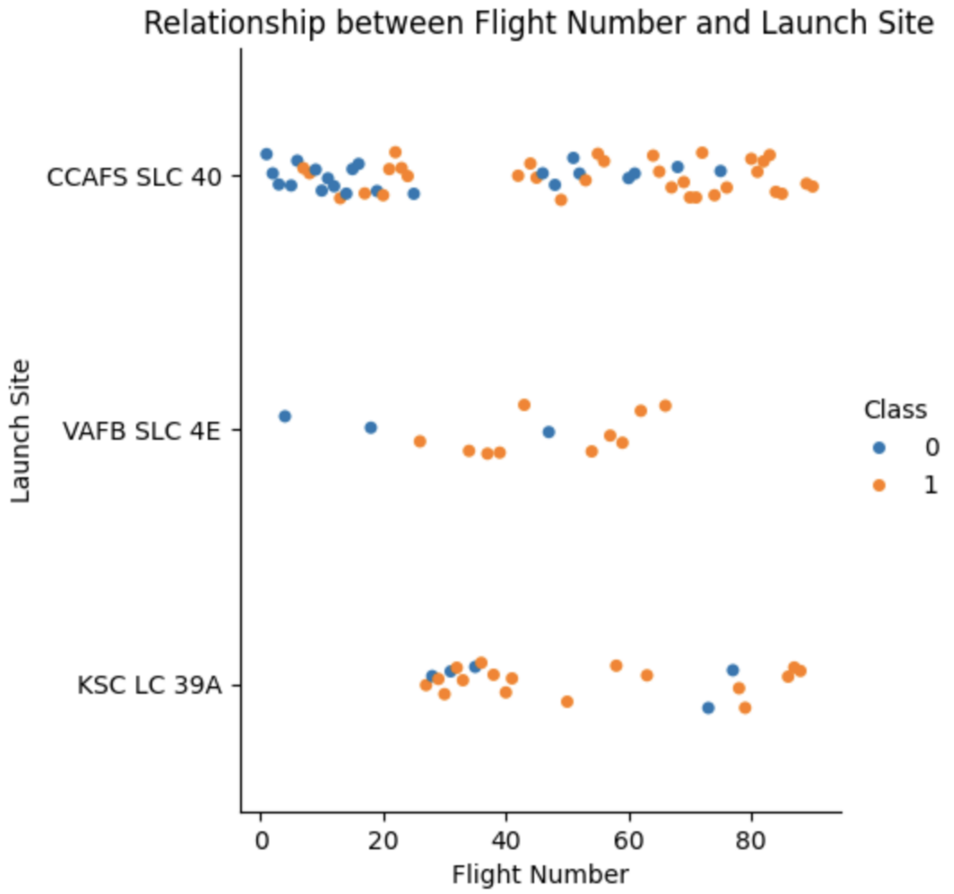
- Filtered DataFrame to only have the features used for predicting success
- One hot encoding categorical values using get_dummies()
- Turning all values in DataFrame to floats using astype()

EDA WITH VISUALIZATION RESULTS



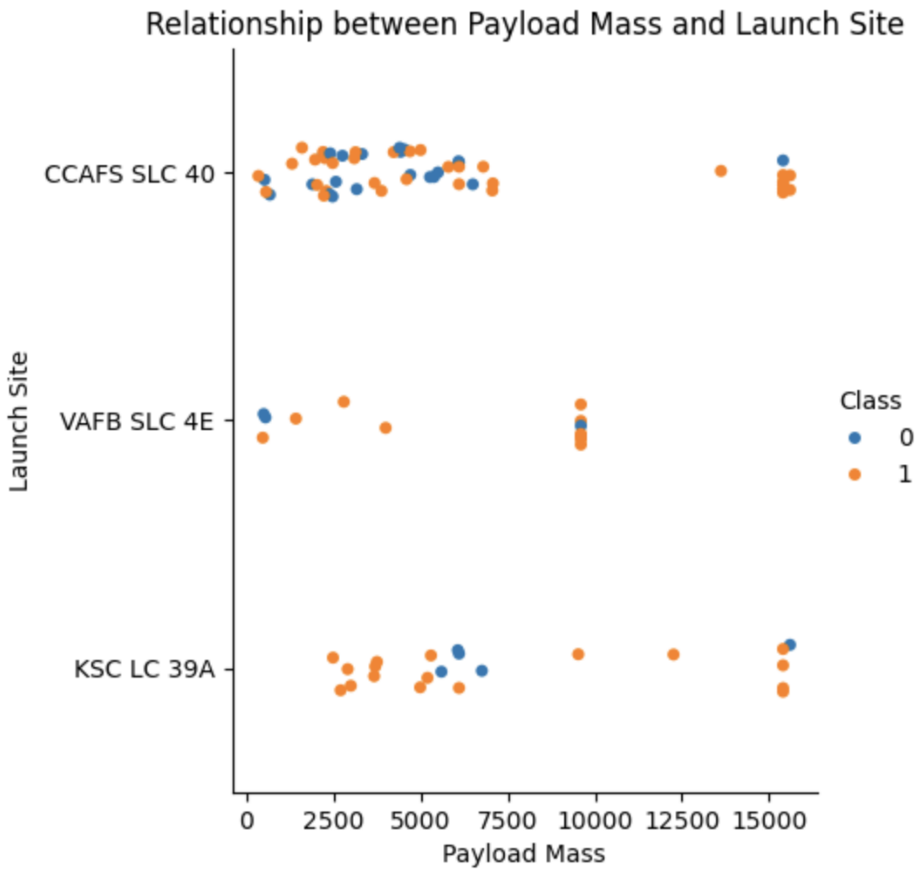
This graph refers to flight number vs payload mass with a hue of class. It seems to be that the higher the number of flights, the more varied the payload masses carried are (payload masses range between low and high). It also seems like there is a cluster of failure outcomes at lower numbers of flights and lower payload masses, and there are more successful outcomes than failures at the higher numbers of flights and higher payload masses.

EDA WITH VISUALIZATION RESULTS



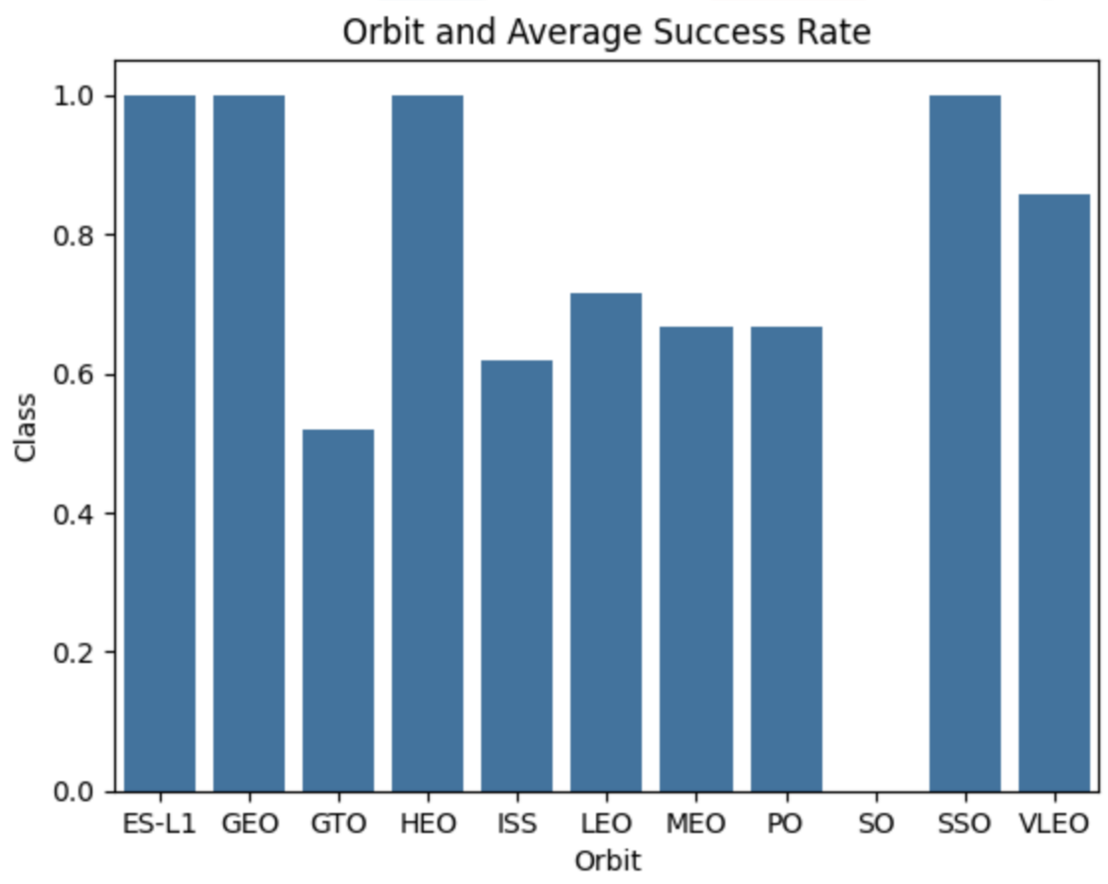
This graph refers to number of flights vs launch site with a hue of class. CCAFS SLC-40 seems to have the highest number of launches, with launches ranging the whole span of number of flights besides ~22-38. VAFB SLC 4E seems to have most launches with a number of flights in the range 20 – 60. KSC LC 39A has all of its flights above 20, with the biggest cluster between 20 and 40. CCAFS SLC 40 tends to have more failures at lower numbers of flights and successes at higher numbers of flights. VAFB SLC 4E generally has a lot more successes, with most of the failures at lower numbers of flights. KSC LC 39A also has a lot more successes, with some failures at smaller numbers of flights and some at larger ones.

EDA WITH VISUALIZATION RESULTS



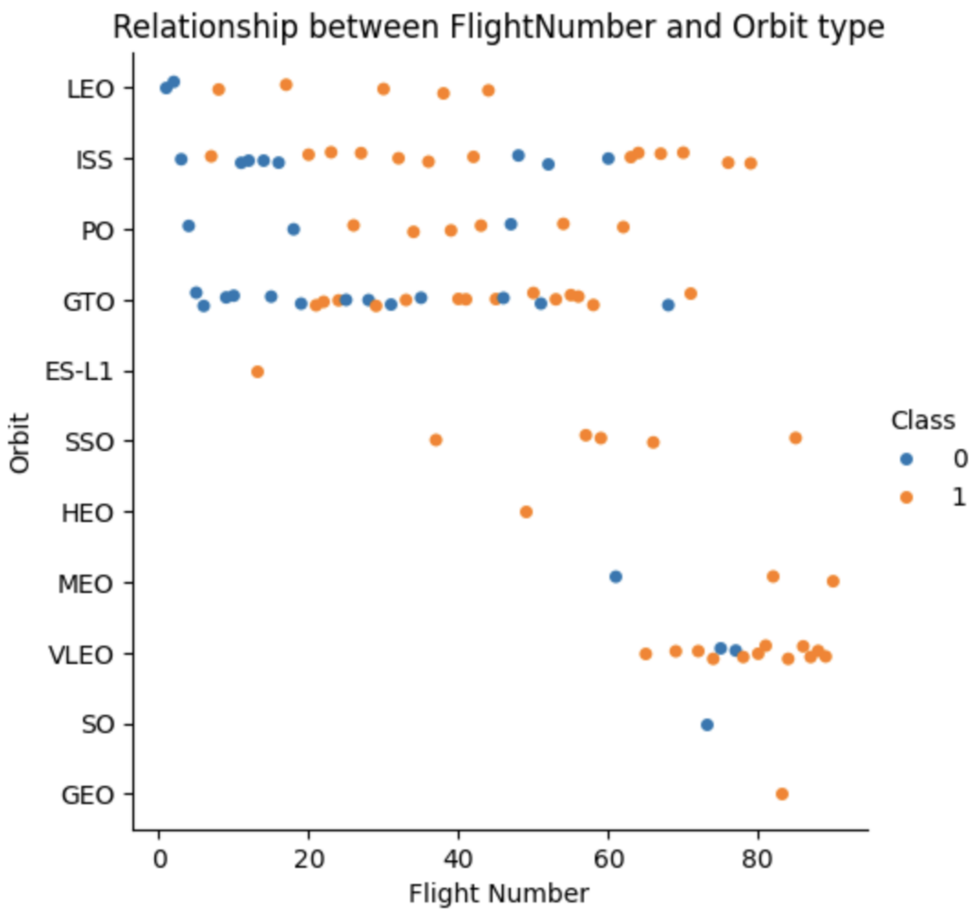
This graph refers to payload mass vs launch site with a hue of class. CCAFS SLC40 seems to have the highest number of launches, with payload masses mostly clustered between 0 and 7500 kg, and a few around 15000 kg. VAFB SLC 4E seems to have some payload masses between 0 and 4000 kg, and the rest at about 9000 kg. KSC LC 39A has all of its payload masses above 2500 kg, with the biggest cluster between 2500 and 7500 kg. CCAFS SLC 40 tends to have roughly an equal number of successes and failures between 0 and 7500 kg, but more successes than failures at 15000 kg. VAFB SLC 4E generally has a lot more successes, with 1 in each of its ranges. KSC LC 39A also has a lot more successes, with most failures between 5000 and 7500 kg .

EDA WITH VISUALIZATION RESULTS



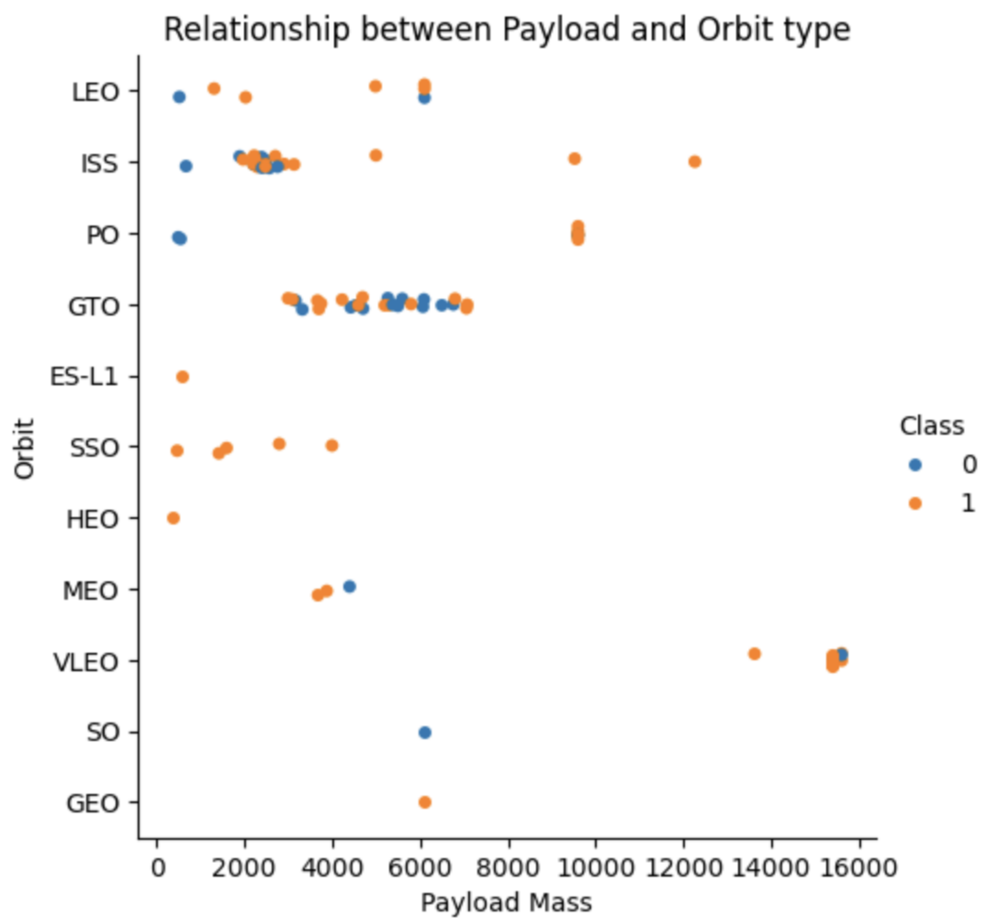
This graph refers to the average success rate of each orbit. Launches to ES-L1, GEO, HEO, and SSO had the highest average success rate (1.0). Launches to GTO had the lowest average success rate, 0.5, which means that half of the times, the first stage was successful at this orbit.

EDA WITH VISUALIZATION RESULTS



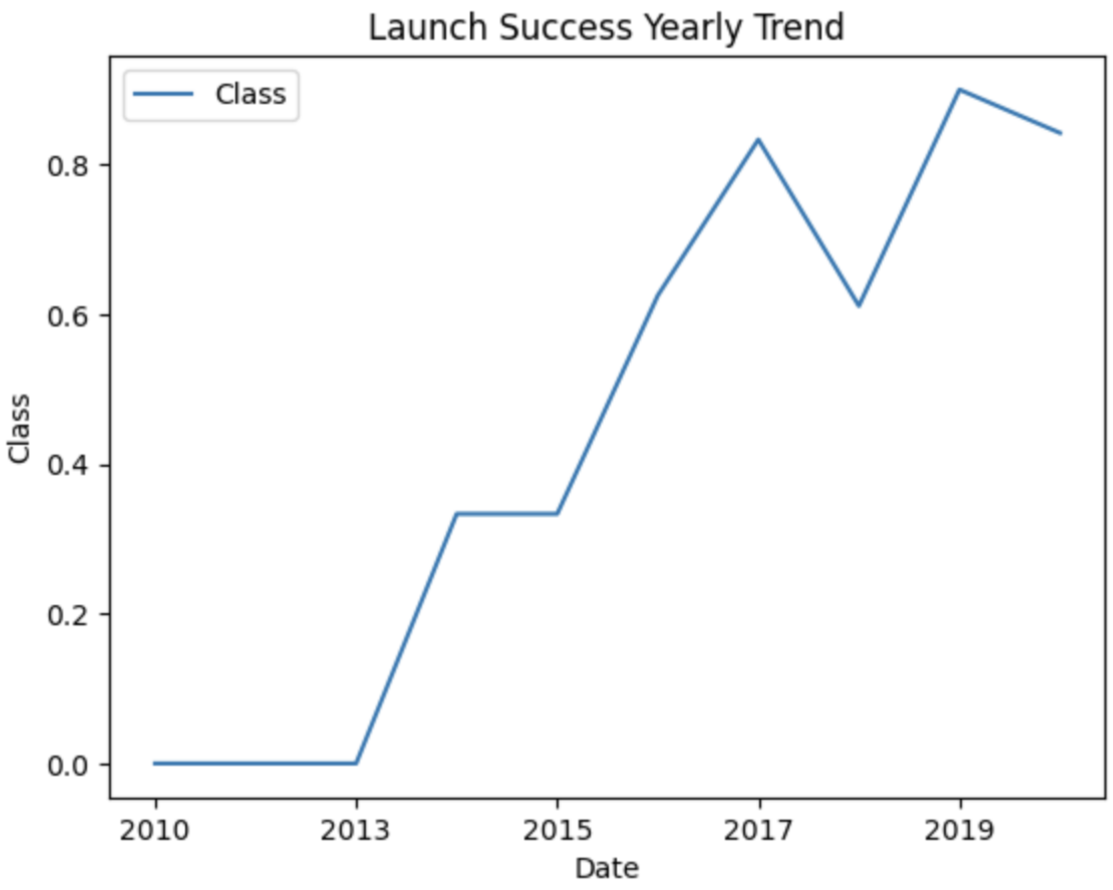
This graph refers to number of flights vs orbit type with a hue of class. At orbits LEO, ISS, PO, and GTO, the number of flights were pretty spread out, but a lot of them were closer to lower numbers. SSO, MEO, and VLEO had flight numbers on the higher end. ES-L1, SSO, SO, and GEO only had one flight. ES-L1's was pretty low, HEO's was in the middle, and SO and GEO had high numbers of flights. For orbits like LEO, PO (somewhat), GTO, and MEO, the overall trend is that higher numbers of flights tended to have more success. SSO is the only orbit that had all numbers of flights be pretty successful.

EDA WITH VISUALIZATION RESULTS



This graph refers to payload vs orbit type with a hue of class. Most of the orbits had most of its first stages containing payload masses under 10000 kg. First stages at VLEO had payload masses around 13000 kg and above, while ISS had one around 12000 kg. ES-L1, HEO, SO, and GEO only have 1 payload mass. For Orbits like PO, LEO, and ISS, higher payload masses were more successful. However, GTO doesn't seem to have a trend. SSO, ES-L1, HEO, and GEO have only really had good success rates. MEO and VLEO have had its failures at the highest payload masses they've had.

EDA WITH VISUALIZATION RESULTS



This graph refers to the yearly trend of launch success from the year 2010 to 2019. There seems to be an overall positive trend, meaning that over time, the success rate of first stages has increased. The trend was stagnant between 2010 and 2013, and for a period between 2013 and 2015. There was also a dip from about 2017 to 2018, with a dip starting from 2019. It is possible that the last dip could have been a result of COVID-19. Additionally, the lack of improvement between 2010 and 2013 from an extremely low success rate could have been because the Falcon 9 was first launched in 2010, and the first version may have needed many improvements.

INTERACTIVE VISUAL ANALYTICS METHODOLOGY

Using Folium:

- Creating a Map containing:
  - A circle indicating the NASA Johnson Space Center
  - A circle and marker for each Launch Site
  - Creating marker clusters for each launch outcome for each site (red = bad, green = good)
  - Created markers with distance for the closest coastline, railway, highway, and city to one of the launch sites

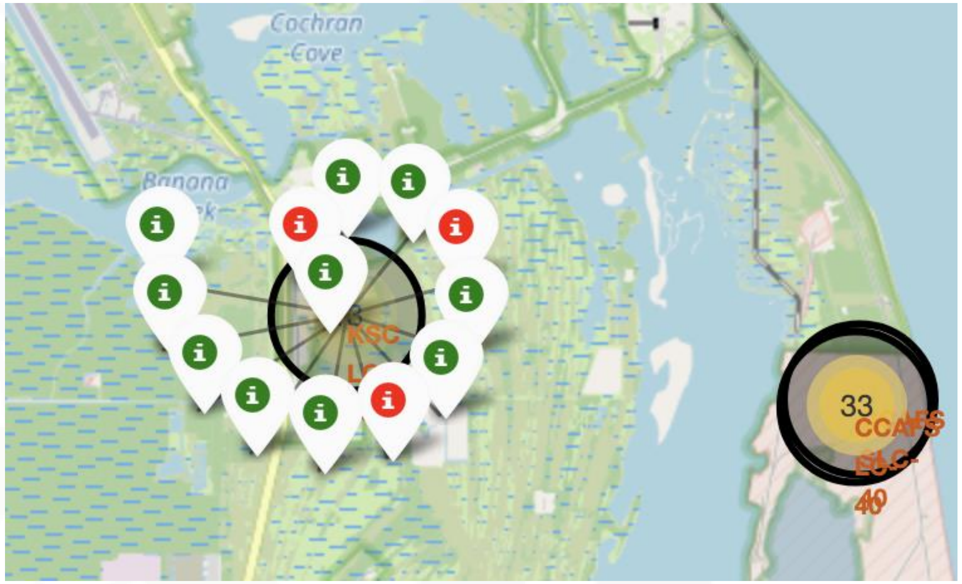
Using Plotly Dash

- Created a Dashboard using html.Div containing:
  - A dropdown for launch site selection (dcc.DropDown)
  - Pie chart showing success counts for all sites, or the success vs failed counts if a specific site is selected
  - Slider to select payload range (dcc.RangeSlider)
  - Scatter Chart showing Payload vs Launch Success
- Created Charts using Callback functions

INTERACTIVE VISUAL ANALYTICS RESULTS

These circles and markers indicate the four launch sites of the Falcon 9 rockets. VAFB SLC-4E is located on the coastline of California, while the other 3, CCAFS LC-40, KSC LC-39A, and CCAFS SLC-40, are located on the coastline of Florida. CCAFS LC-40 and CCAFS SLC-40 are pretty much located in the same area. That's because CCAFS LC-40 was simply renamed to CCAFS SLC-40. They are located at some of the closest points in the United States to the equator.

INTERACTIVE VISUAL ANALYTICS RESULTS

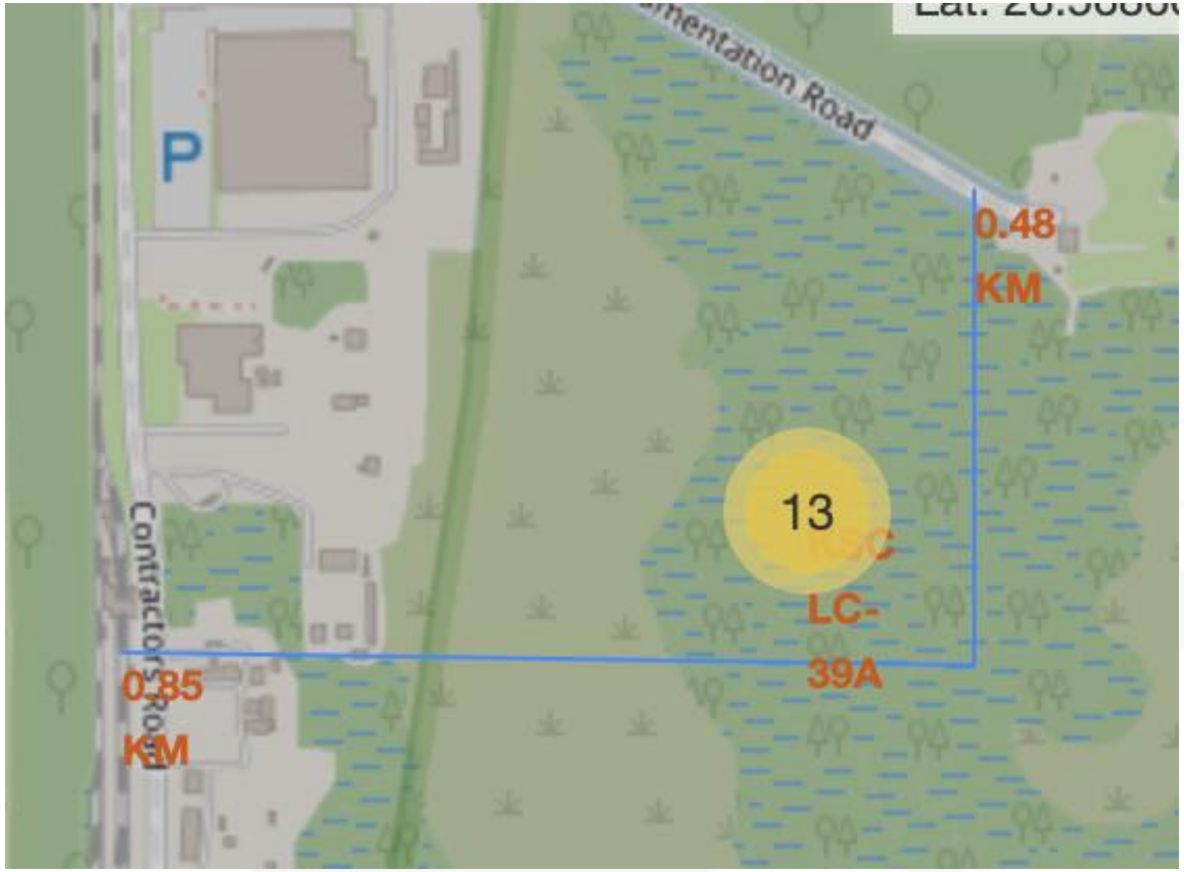


These images indicate the addition of marker clusters for the launch outcomes at each site. From my findings, VAFB SLC-4E has a 40% success rate (4/10), KSC LC-39A has a 77% success rate (10/13), CCAFS LC-40 has a 27% success rate (7/26), and CCAFS SLC-40 has a 43% success rate (3/7).

INTERACTIVE VISUAL ANALYTICS RESULTS

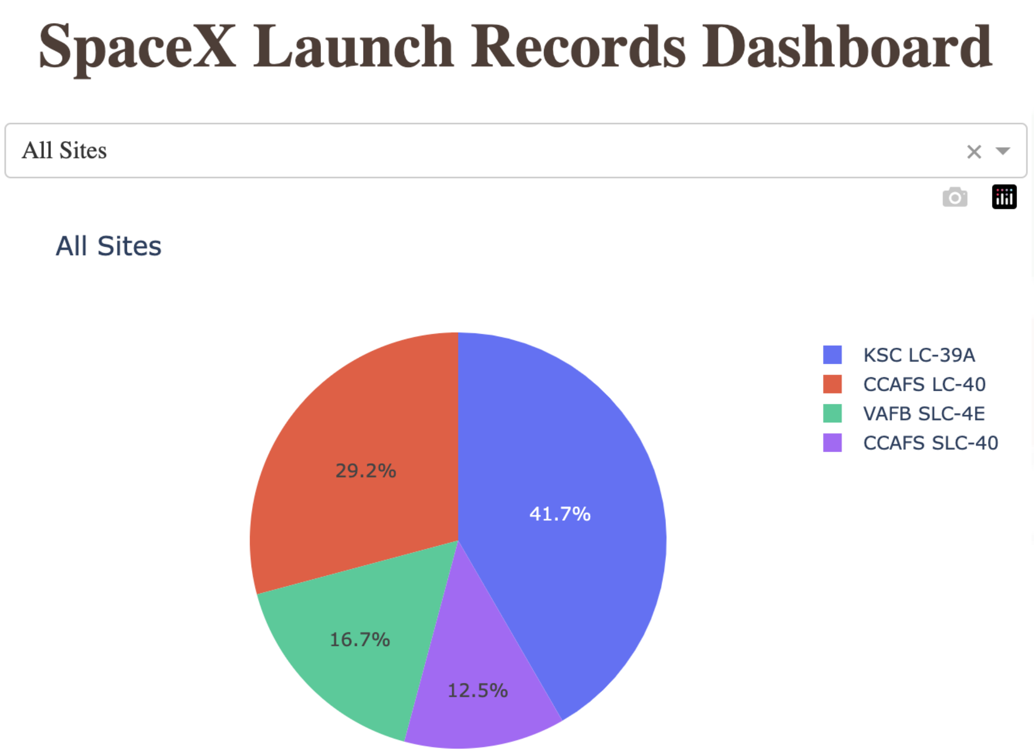
This image indicates the distance of the KSC LC-39A launch site from the nearest coastline point and the nearest city, Titusville. It is approximately 3.10 km from the nearest coastline point and 17.27 km from the nearest city, making it a lot closer to the coastline than to cities.

INTERACTIVE VISUAL ANALYTICS RESULTS



This image indicates the distance of the KSC LC-39A launch site from the nearest highway and the railroad. It is approximately 0.48 km from the nearest highway and 0.85 km from the nearest railroad. Therefore, the launch site is closer to highways and railroads than the coastline and cities.

INTERACTIVE VISUAL ANALYTICS RESULTS

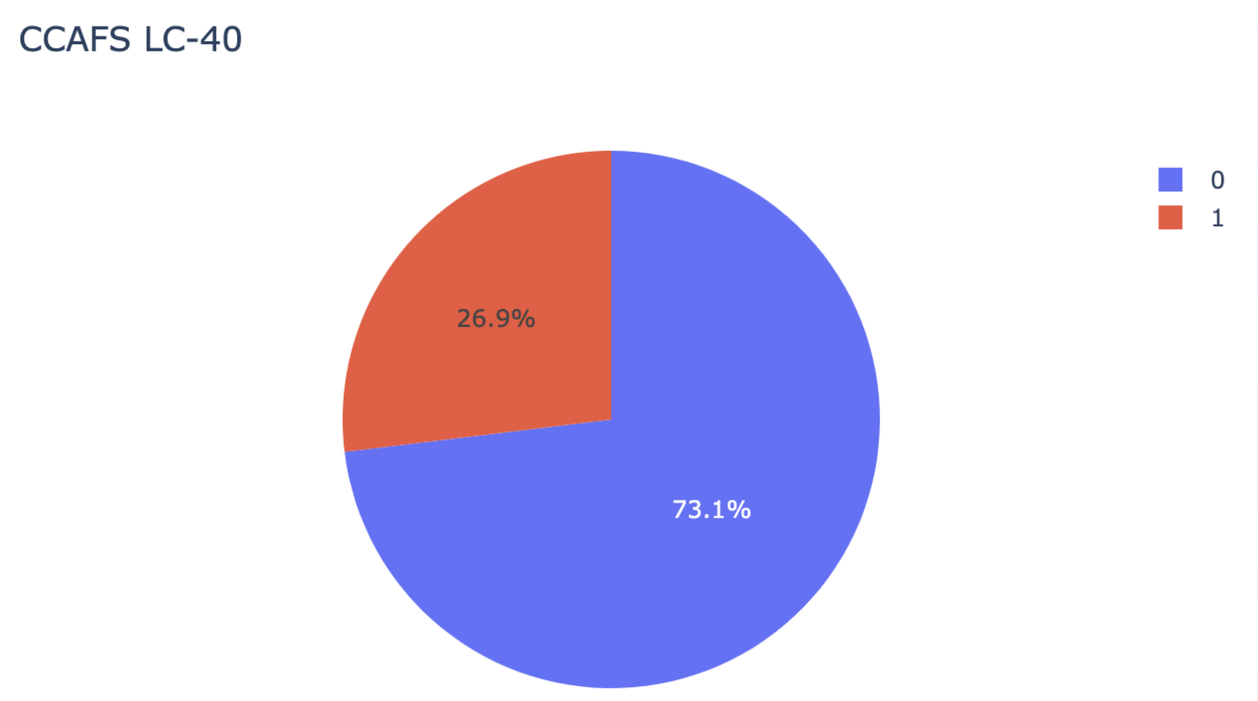


This image indicates the percentage of the success count coming from each of the launch sites. We can see that KSC LC-39A contains the highest percentage of the success count (has the highest number of successes), while CCAFS SLC-40 contains the lowest percentage (has the lowest number of successes).

INTERACTIVE VISUAL ANALYTICS RESULTS

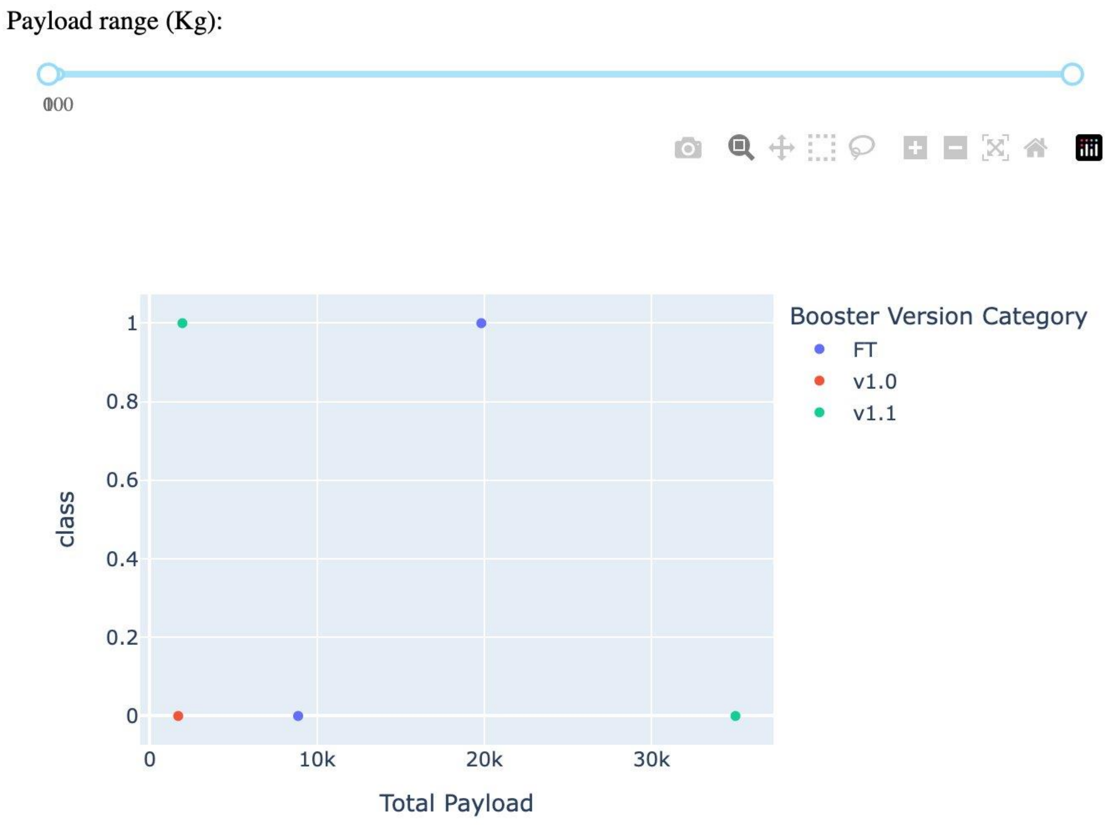
This image indicates the relation between payload mass and class, with different colors for different booster versions. It seems like there are more rockets that have had failures than successes. There also does not seem to be much of a relation between payload mass and class when the range of payload mass is set to the entire spectrum. It does seem like v1.1 had had many more failures than successes, while FT has had more successes than failures by a little bit.

INTERACTIVE VISUAL ANALYTICS RESULTS



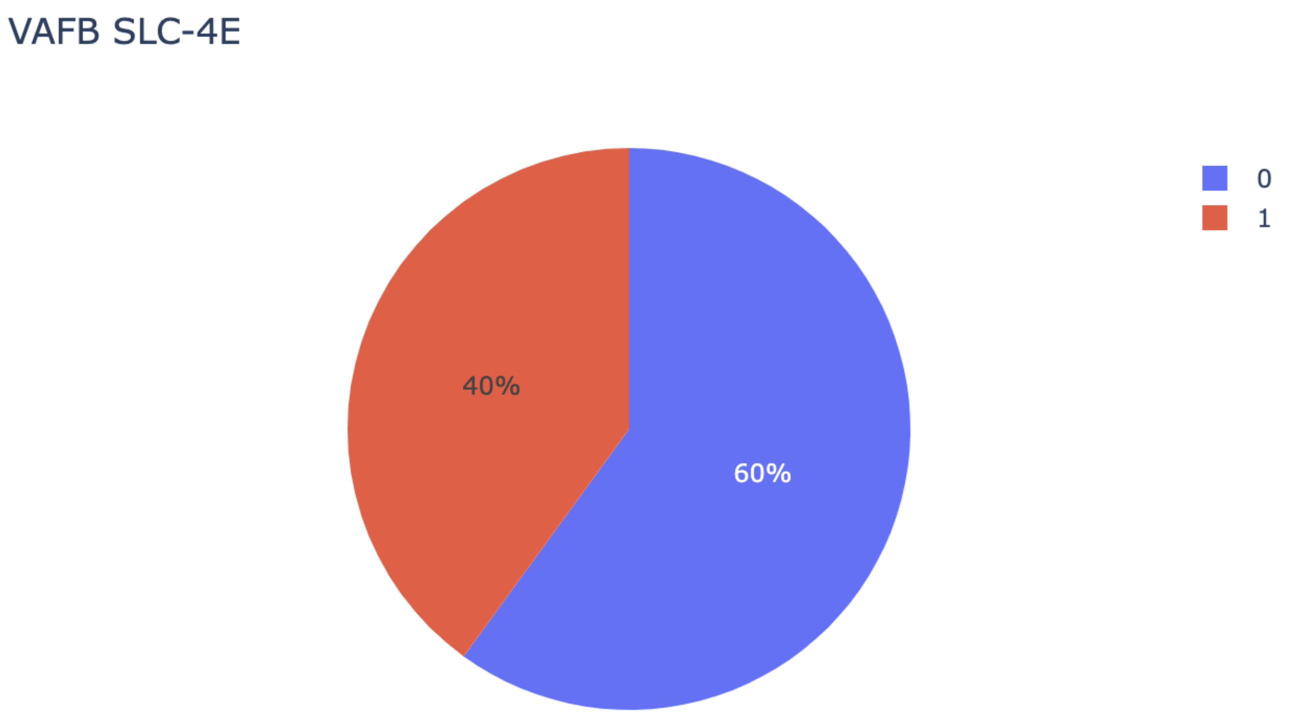
This image indicates the percentage of successes and percentage of failures at the CCAFS LC-40 launch site. There seems to have been a little under 3 times as many failures as successes at this site.

INTERACTIVE VISUAL ANALYTICS RESULTS



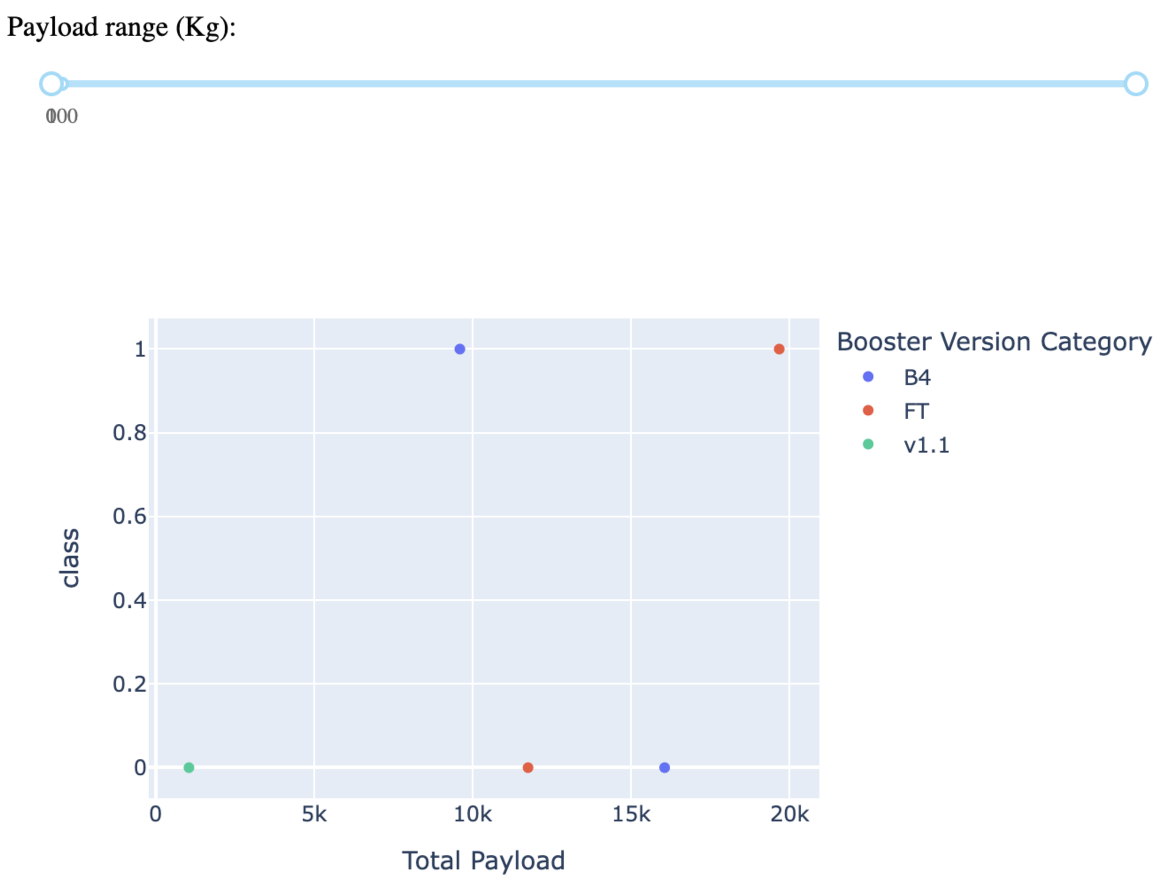
This image indicates the relation between payload mass and class at the CCAFS LC-40 launch site, with different colors for the different booster versions that have launched there. There are 3 failures and 2 successes, where v1.0 has not had a success. There also does not seem to be much of a relation between payload mass and class when the range of payload mass is set to the entire spectrum.

INTERACTIVE VISUAL ANALYTICS RESULTS



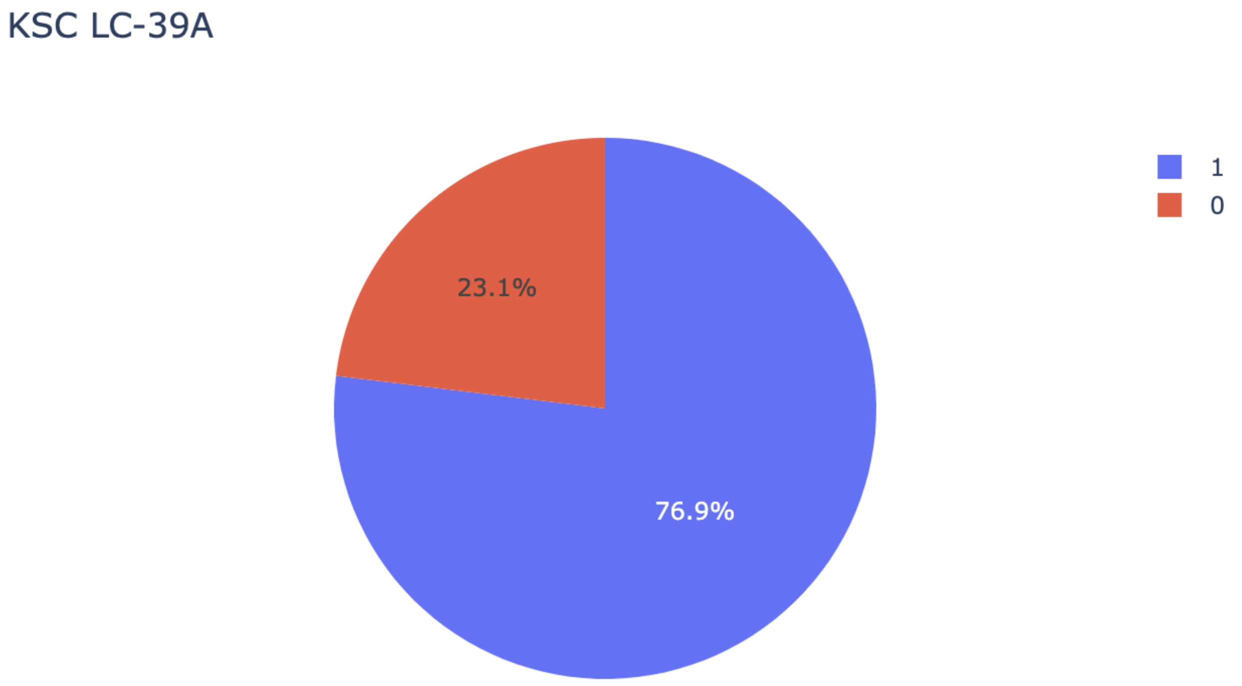
This image indicates the percentage of successes and percentage of failures at the VAFB SLC-4E launch site. There seems to be a 2:3 proportion between successes and failures at this site.

INTERACTIVE VISUAL ANALYTICS RESULTS



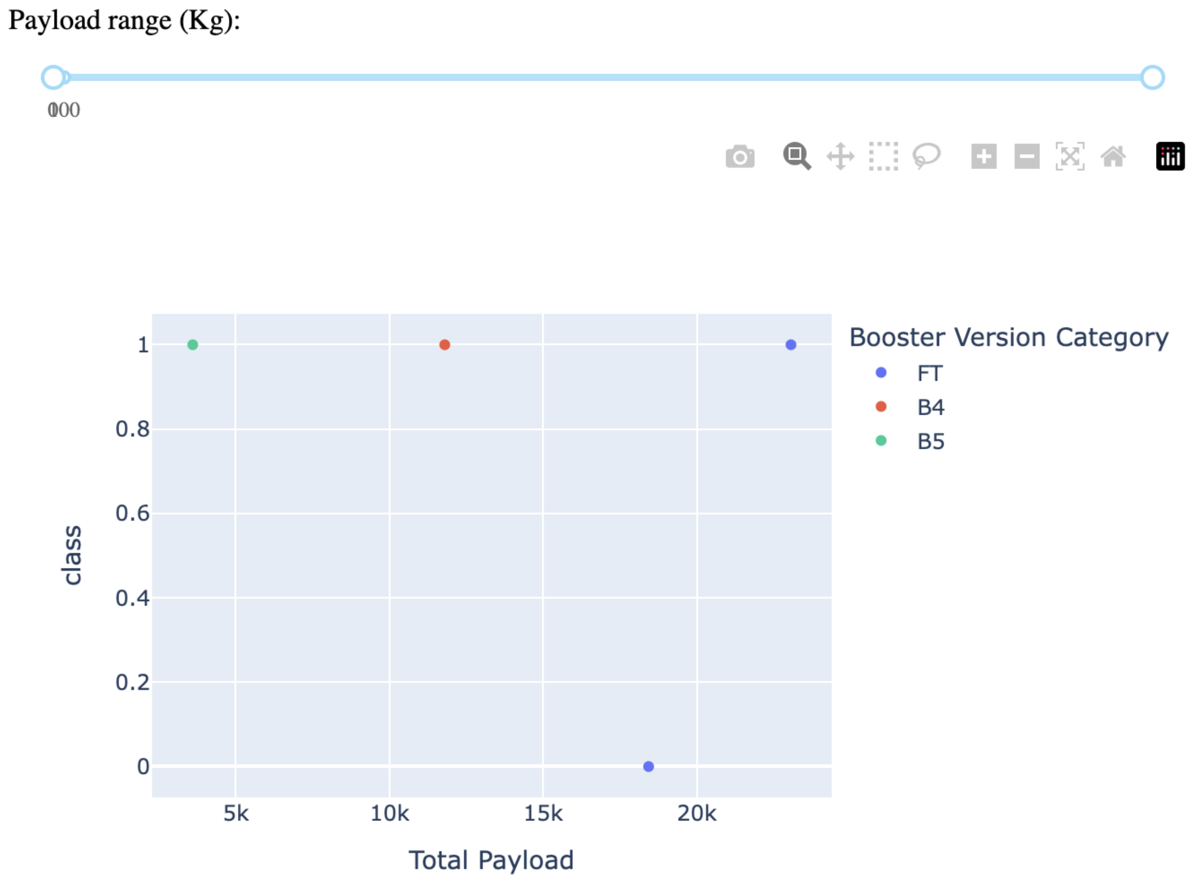
This image indicates the relation between payload mass and class at the VAFB SLC-4E launch site, with different colors for the different booster versions that have launched there. There are 3 failures and 2 successes, where v1.1 has not had a success. There also does not seem to be much of a relation between payload mass and class when the range of payload mass is set to the entire spectrum, but it does seem like overall, the successes have a higher payload range than the failures.

INTERACTIVE VISUAL ANALYTICS RESULTS



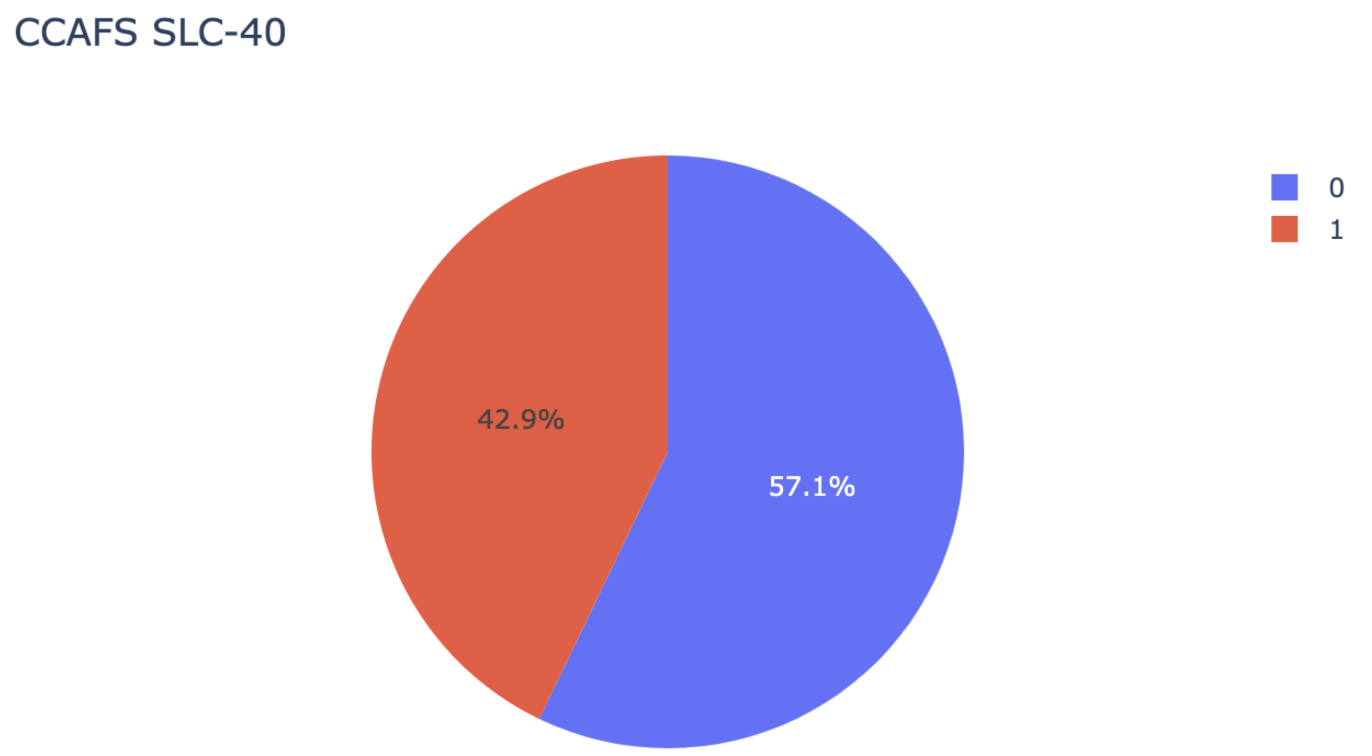
This image indicates the percentage of successes and percentage of failures at the KSC LC-39A launch site. There seems to have been a little over 3 times as many successes as failures at this site.

INTERACTIVE VISUAL ANALYTICS RESULTS



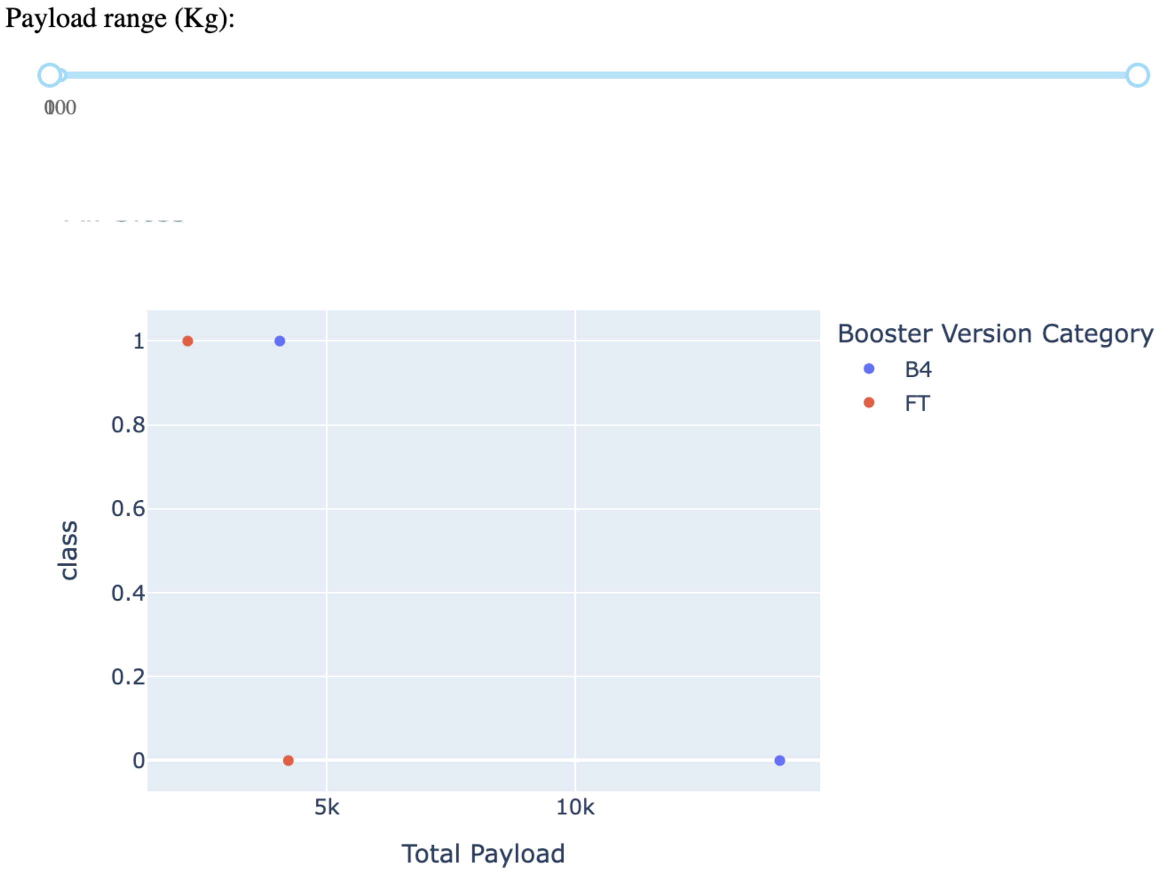
This image indicates the relation between payload mass and class at the KSC LC-39A launch site, with different colors for the different booster versions that have launched there. There is 1 failure and 3 successes, and the failure comes from FT. All booster versions have had a success though. There also does not seem to be much of a relation between payload mass and class when the range of payload mass is set to the entire spectrum.

INTERACTIVE VISUAL ANALYTICS RESULTS



This image indicates the percentage of successes and percentage of failures at the CCAFS SLC-40 launch site. There seems to have be a bit more failures than successes at this site.

INTERACTIVE VISUAL ANALYTICS RESULTS



This image indicates the relation between payload mass and class at the CCAFS SLC-40 launch site, with different colors for the different booster versions that have launched there. There are 2 failures and 2 successes, where both booster versions have had a success and a failure. It does seem like the higher payload masses have been failures, and the lower ones have been successes.

PREDICTIVE ANALYSIS METHODOLOGY

Used scikit-learn to find the best ML model to use for prediction by:

- Splitting off the class label column from the rest of the DataFrame
- Standardized the data in the feature columns
- Split the X and Y data into training and testing sets
- Tuned hyperparameters, checked the accuracy, and plotted the confusion matrix for ML algorithms such as Logistic Regression, Support Vector Machine, Decision Tree, and K_Nearest Neighbors
- Compared the accuracies of all the models

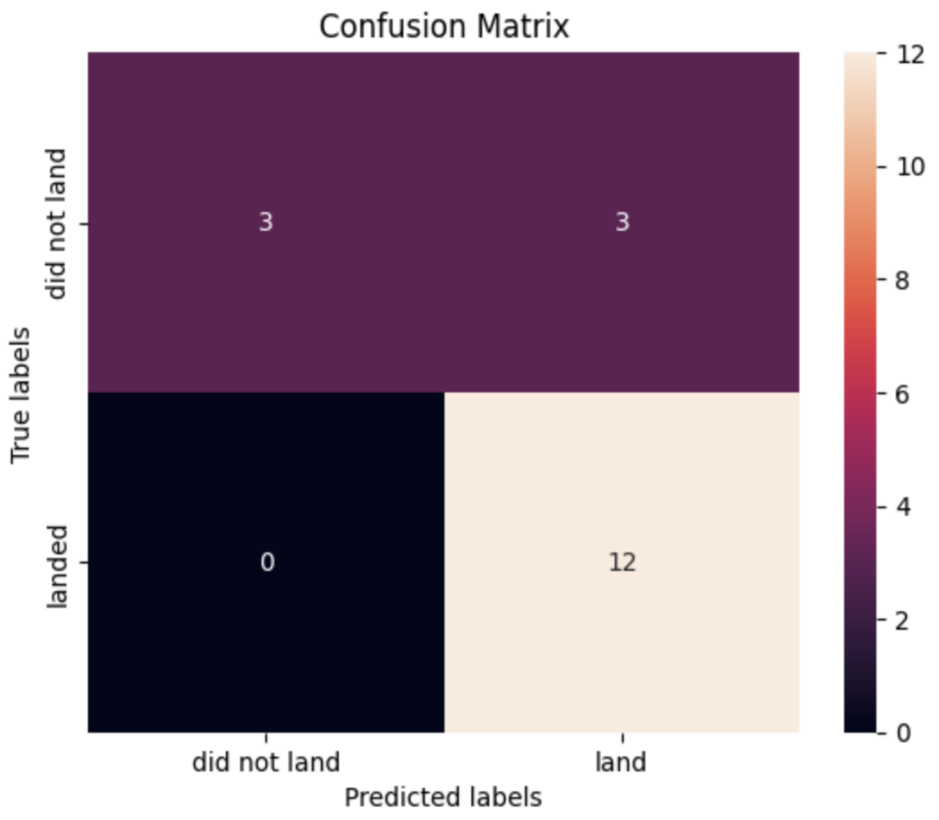
PREDICTIVE ANALYSIS RESULTS

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

Test set accuracy:

```
0.8333333333333334
```

These values refer to the results of GridSearch with Logistic Regression. The C hyperparameter controls the regularization of the model, and its best performing value is 0.01. Penalty controls the size of the slopes/coefficients, and the best performing norm is l2. Solver is the algorithm to use for optimizing the result, and the best performing algorithm is lbfgs. The best accuracy score we got for all iterations of the model is 84.6%.



The confusion matrix shows that, out of the 18 test records, 3 of them were correctly predicted to not land, 12 were correctly predicted to land, 0 were incorrectly predicted to not land, and 3 of them were incorrectly predicted to land, meaning there were false positives.

PREDICTIVE ANALYSIS RESULTS

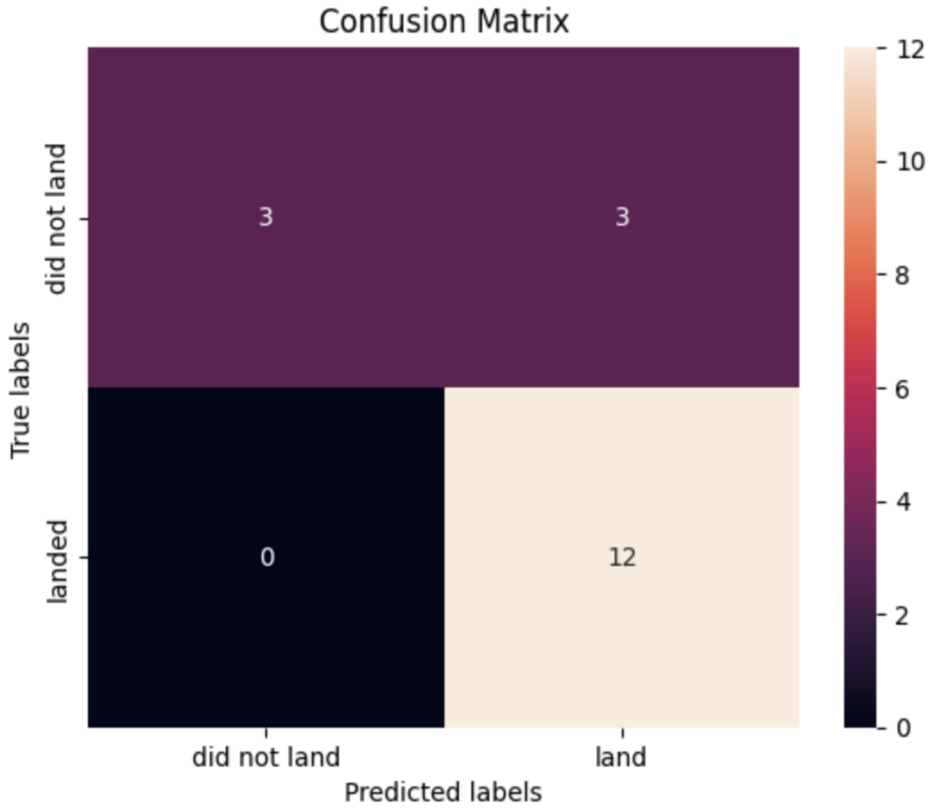
```
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856
```

Test set accuracy:

```
0.8333333333333334
```

These values refer to the results of GridSearch with Support Vector Machines. The C hyperparameter controls the regularization of the model, and its best performing value is 1.0. Gamma affects the complexity of the decision boundary, and its best performing value is around 0.032. The kernel function is used to find the hyperplane separating the data points of different classes, and the best performing function is sigmoid. The best accuracy score we got for all iterations of the model is 84.8%.

The confusion matrix shows that, out of the 18 test records, 3 of them were correctly predicted to not land, 12 were correctly predicted to land, 0 were incorrectly predicted to not land, and 3 of them were incorrectly predicted to land, meaning there were false positives.



PREDICTIVE ANALYSIS RESULTS

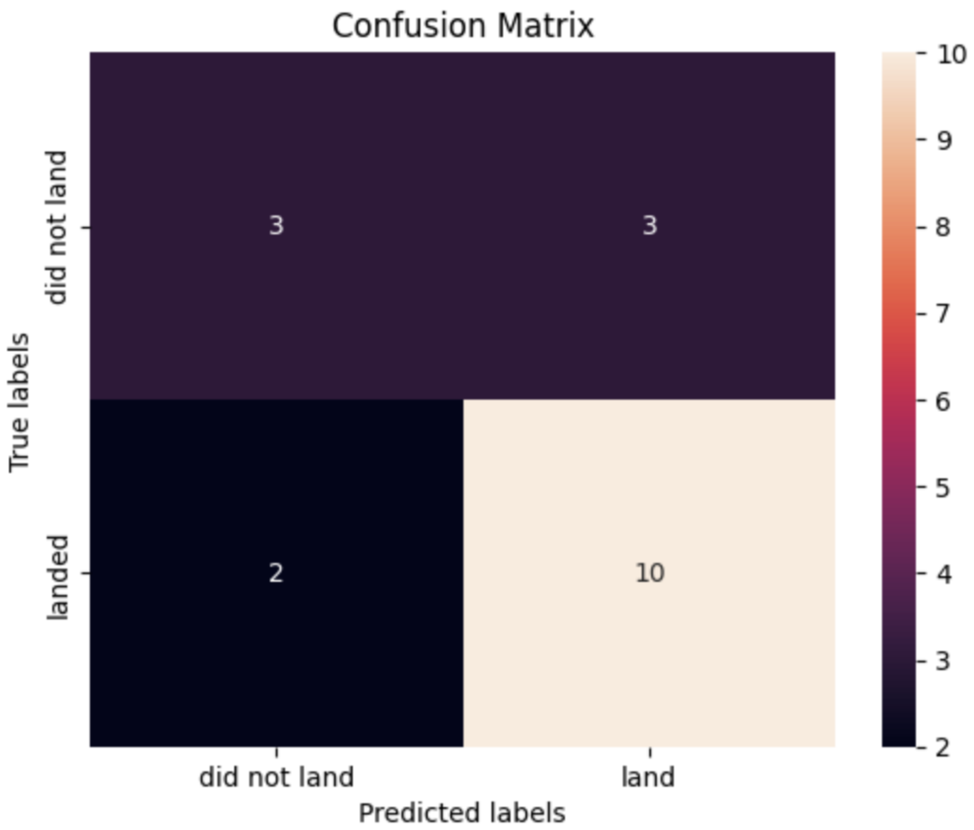
```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 12, 'min_samples_leaf': 4, 'min_sampl
es_split': 10, 'splitter': 'best'}
accuracy : 0.8714285714285713
```

These values refer to the results of GridSearch with Decision Tree. The criterion hyperparameter refers to the impurity metric to use when deciding which feature to split by, and the best performing metric is gini. Max depth refers to the maximum depth of the tree, and the best performing value is 12. Min samples leaf refers to the minimum amount of samples that must be at a leaf node (cannot split more), and the best performing value is 4. Min samples split refers to the minimum amount of samples that are needed to split using a feature, and the best performing value is 10. Splitter refers to the split strategy, and the best performing strategy is best split. The best accuracy score we got for all iterations of the model is 87.1%.

The confusion matrix shows that, out of the 18 test records, 3 of them were correctly predicted to not land, 10 were correctly predicted to land, 2 were incorrectly predicted to not land, and 3 of them were incorrectly predicted to land, meaning there were false positives and false negatives.

Test set accuracy:

```
0.7222222222222222
```



PREDICTIVE ANALYSIS RESULTS

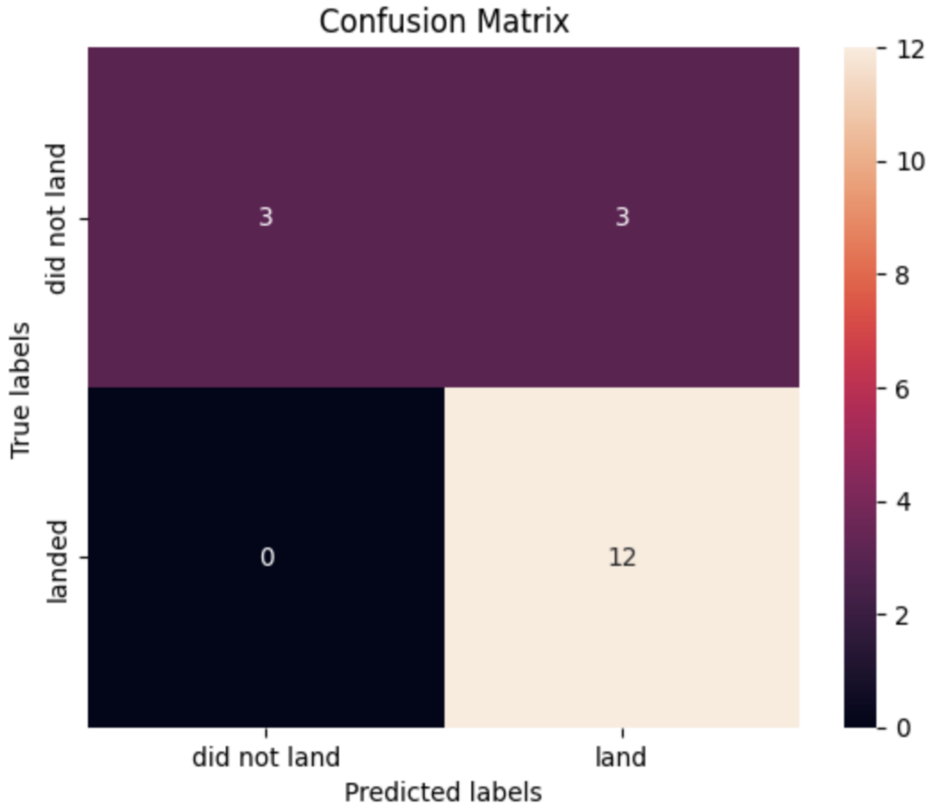
```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858
```

Test set accuracy:

```
0.8333333333333334
```

These values refer to the results of GridSearch with K-Nearest Neighbors. The algorithm hyperparameter refers to the method to find the nearest neighbors, and its best performing method is auto. N Neighbors refers to how many neighbors to use, and its best performing value is 10. P refers to the distance metric, and its best performing metric is manhattan distance (p=1). The best accuracy score we got for all iterations of the model is 84.8%.

The confusion matrix shows that, out of the 18 test records, 3 of them were correctly predicted to not land, 12 were correctly predicted to land, 0 were incorrectly predicted to not land, and 3 of them were incorrectly predicted to land, meaning there were false positives.



PREDICTIVE ANALYSIS RESULTS

```
Logistic Regression Accuracy:  0.8333333333333334
SVM Regression Accuracy:  0.8333333333333334
Decision Tree            Accuracy:  0.7222222222222222
KNN Regression Accuracy:  0.8333333333333334
```

These values refer to the accuracy, calculated using the score() method. All models except for decision tree had an accuracy score of 0.8333.

DISCUSSION

- Between June 2010 and March 2017, no failures with landings on ground pad, but of those attempted, the most number of launches used drone ship
- According to the flight number vs payload mass graph, there is a higher chance of success with a higher number of flights
- Orbits ES-L1, GEO, HEO, and SSO have the highest success rate, but SSO is the only orbit with more than one launch and all were successes. All of the launches had lower payload masses
- The CCAFS SLC-40 launch site seems to have the highest number of launches, but KSC LC-39A has the highest success count and the highest success rate (76.9%) out of all launch sites,
- Launches at KSC LC-39A all had numbers of flights ranging from 20 to 80 and payload masses above 2500 kg
- Overall, as time passes, the launch success rate increases

DISCUSSION

- All launch sites are near coastlines and as close to the equator as possible within the US. They also tend to be pretty far away from cities.
- Booster version v1.1 has a huge skew towards failures, while FT tends to be more successful of a booster version. KSC LC-39A and CCAFS SLC-40 are the only sites that have launched FT but not v1.1, and they have higher success rates than the other two sites.
- The CCAFS LC-40 launch site has the lowest success rate of 26.9%
- Based on overall comparison and individual best accuracy scores of models, the models that predict the best would be K-Nearest Neighbors or Support Vector Machines

CONCLUSION

- We want to be able to predict if the first stage of a launch will land successfully
- In order to do so, we can create a Machine Learning Model using KNN or SVM with the data from the SpaceX website
- From the findings, we can conclude that the launches that are most likely to be successful will have these qualities:
    - Launched from KSC LC-39A
    - Number of flights between 20 and 80
    - Payload mass above 2500 kg
    - Booster version FT
    - Launched to Orbit SSO
    - Possibly doing landings on ground pad
- We can use our insights and the ML model to predict the first stage success of various kinds of launches, and use it to decide how to bid against our competitor, SpaceX
- Before we do so, it would be good to do a more thorough analysis of the data and work more closely on creating a model with a higher accuracy than the ones we have currently

APPENDIX

- Learn more about SpaceX rockets and stages here:
    - TechTarget: What is SpaceX?
    - Wikipedia: Booster (rocketry)
    - Wikipedia: Falcon9 First Stage Landing Tests

APPENDIX

- Learn more about ML Algorithms and Hyperparameters at
  - Scikit-learn.org
  - Gemini AI
  - Logistic Regression and Regularization