IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jonathan Dedinata
May 22, 2023

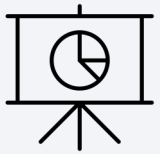
Outline

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

Executive Summary

Methodologies

- Data were collected and data wrangling was performed to make the data usable.
- Exploratory Data Analysis was performed to gain insight on what data features were usable.
- We used said insights to make an interactive plot to see how the features relate to each other.
- We used the data we gathered to create a predictive classification model.



Result

- We decided on a Decision Tree classification model to predict whether the first stage of the Falcon 9 will land successfully based on multiple factors
- We saw that as the year goes by, Falcon 9 first stage landing success rates increases. Additionally, specific rocket booster versions with payload mass below 6000 kg have a higher success rate.
- We also saw that specific launch sites have a higher success ratio in comparison to other launch sites.
- We can predict with an 83.33% accuracy using the data we gathered and explored.

Introduction

Problem

- SpaceX advertises the Falcon 9 launches on its website with a cost of 62 million dollars whereas other providers cost upwards of 165 million dollars. Much of these savings can be attributed to SpaceX reusing its first stage materials when locking their rockets.
- With this in mind, we would like to see if we can determine whether the first stage will land and this information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Target

- Obtain and parse through rocket launch data to gain insight on SpaceX rocket launches.
- Build a reliable classification model that can be used to predict whether the Falcon 9 first stage will land successfully.

Section 1

Methodology

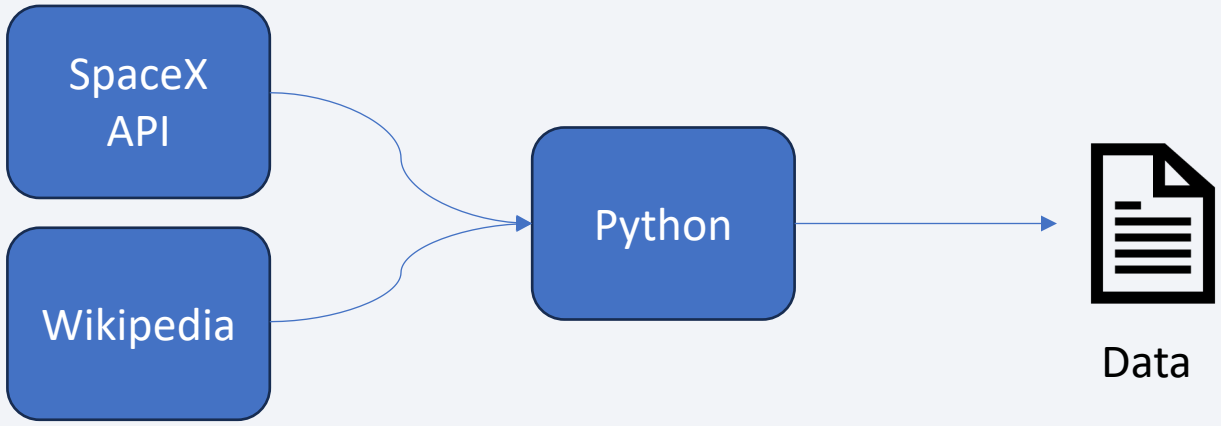
Methodology

Executive Summary

- Data collection methodology:
  - Requested data through SpaceX API and web scraping Wikipedia
- Perform data wrangling:
  - Processed data using the Pandas library and encoded landing outcomes into ordinal values
- Perform exploratory data analysis (EDA) using visualization and SQL:
  - Explored data on IBM db2 using SQL to see landing outcomes on select booster versions
- Perform interactive visual analytics using Folium and Plotly Dash:
  - Visualized analytics based on launch site and what is around it using Folium
  - Visualized the success and failure rates of each launch site and what payloads each booster version carried using Plotly
- Perform predictive analysis using classification models:
  - Performed grid search on Logistic Regression, SVM, Decision Trees and KNN to find the best performing model using Scikit-Learn

Data Collection

- The data was collected through Python by accessing SpaceX API[1] and web scraping Wikipedia[2]
- The raw files were converted using Python into usable and readable data for further processing, removing data with missing values in the process
- The data is further filtered through to only include Falcon 9 launches and rows with missing payload mass values are replaced with the mean value of the column

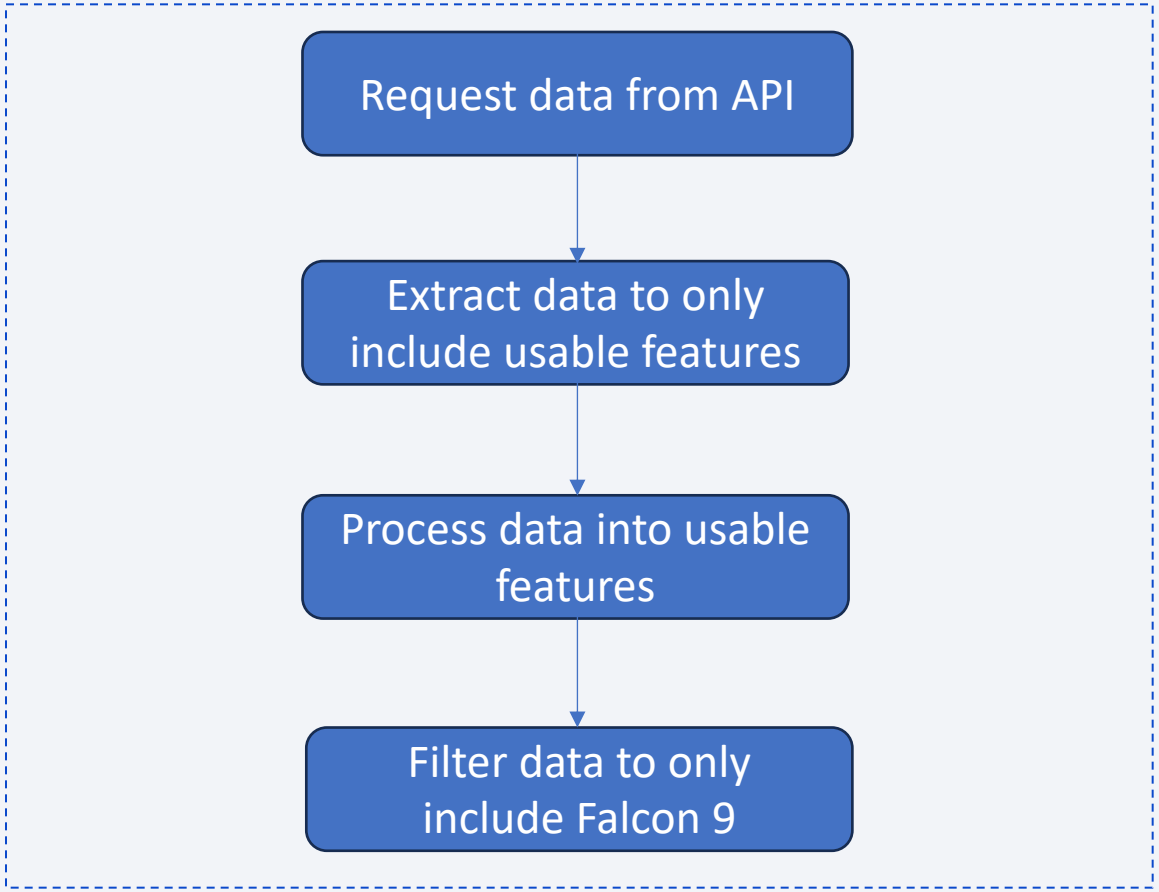


- [1] URL: https://api.spacexdata.com/v4/launches/past
- [2] URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

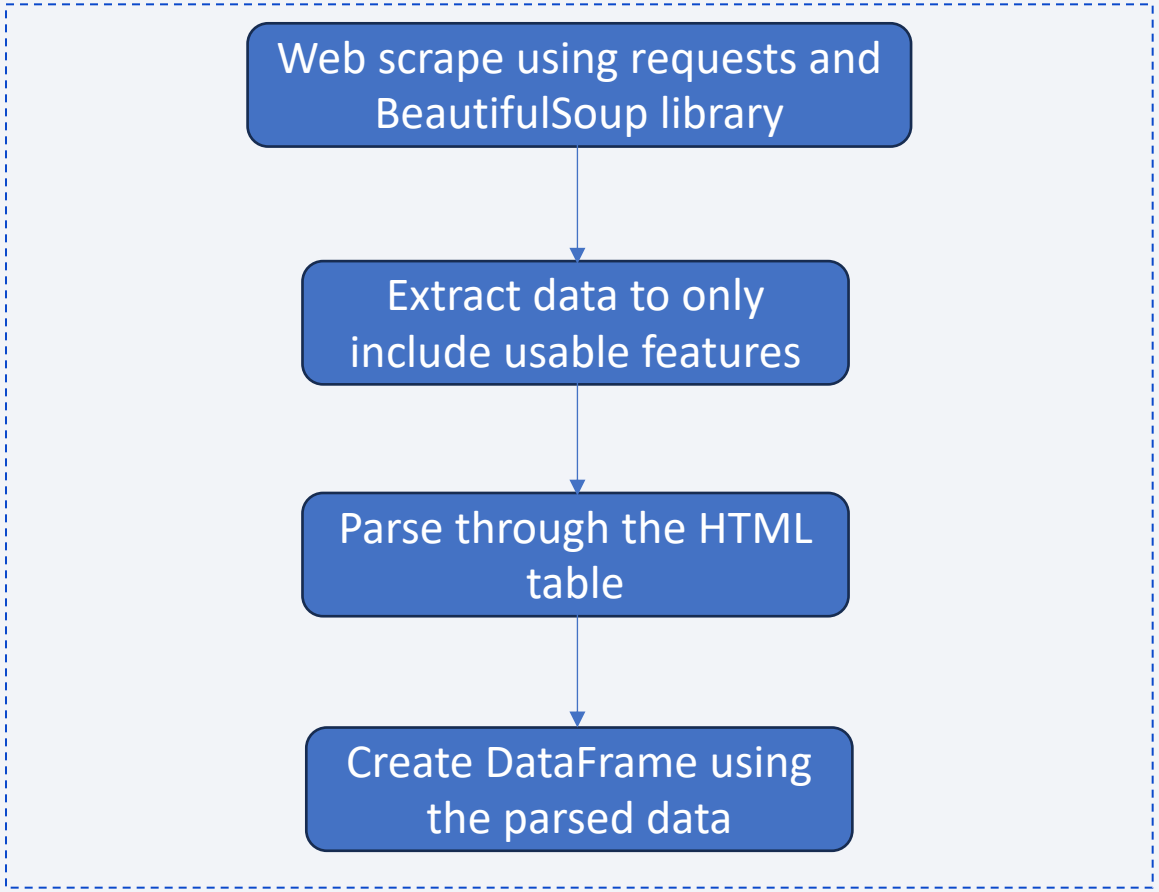
- The data collected through the SpaceX API is requested using the requests library.
- We extract the features that we are interested in. As such, we only take the rocket, payload, lauchpad, cores, flight number, and flight date.
- We further process rocket, payload, launchpad, and cores into features we can use by processing them with functions.
- After processing, we filter the data to only include Falcon 9 launches and removed unusable data, i.e. data with missing values.

- https://github.com/Jonathan-Dedinata/Data-Science-Capstone/blob/main/Data%20Collection/jupyter-labs-spacex-data-collection-api.ipynb



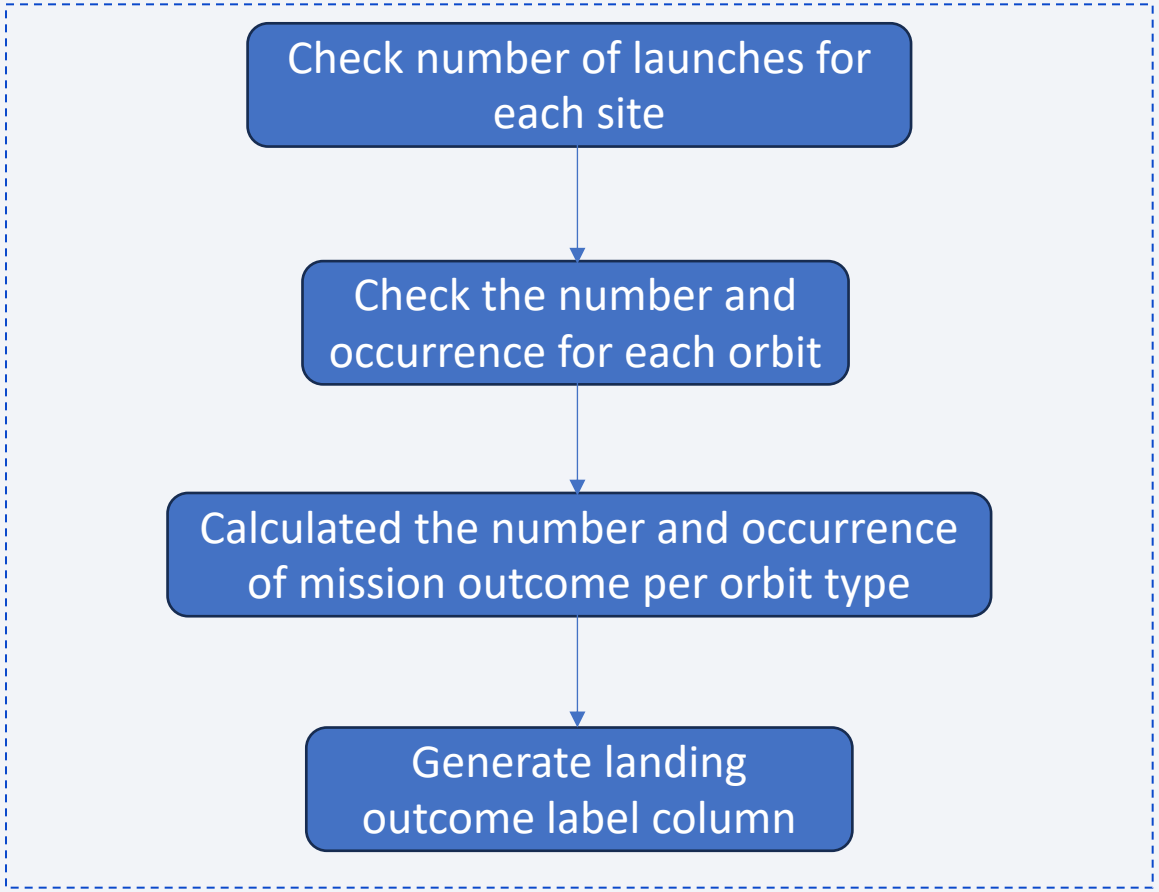
Data Collection - Scraping

- The data collected through web scraping Wikipedia is requested using the requests and BeautifulSoup library.
- We extract the features that we are interested in. As such, we only take the Flight number, Date and Time, Launch Site, Payload, Payload mass, Orbit, Customer, and Launch Outcome.
- We further process the aforementioned features into usable format by parsing through the table.
- After processing, we used the data to create a DataFrame.
- https://github.com/Jonathan-Dedinata/Data-Science-Capstone/blob/main/Data%20Collection/jupyter-labs-webscraping.ipynb



Data Wrangling

- The data was processed using the Pandas and NumPy library
- We checked the number of launches for each site, and the number and occurrence for each orbit
- We then calculated the number and occurrence of mission outcome per orbit type
- Using the Outcome column, we generated a landing outcome label for each flight data
- https://github.com/Jonathan-Dedinata/Data-Science-Capstone/blob/main/Data%20Wrangling/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization

- The following charts were plotted:
  - Scatter plot to find the relationship between:
    - Flight Number and Launch Site
    - Payload Mass and Launch Site
    - Flight Number and Orbit Type
    - Payload Madd and Orbit Type
  - Bar plot to find the success rate on each orbit
  - Line plot to observe the average success rate yearly trend
- https://github.com/Jonathan-Dedinata/Data-Science-Capstone/blob/main/Exploratory%20Data%20Analysis/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- SQL queries were performed to find the following:
  - Names of unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the 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 achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

- https://github.com/Jonathan-Dedinata/Data-Science-Capstone/blob/main/Exploratory%20Data%20Analysis/Complete%20EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

- The following markers were added into the Folium map:
  - Launch Site coordinates and their corresponding site names
  - A marker cluster which consists of markers that indicates how many launches were done in a site and how many were a success/failure, with success being a green marker and failure being a red marker
  - Coordinates to the nearest coast area, highway, railway, and city from a launch site and their corresponding distance from the launch site
- The launch site coordinates were added for context and the marker cluster was added to give more visual information for each launch site. The coordinates to coastal areas, highway, railway, and city were to gain insight as to what exactly makes a launch site successful

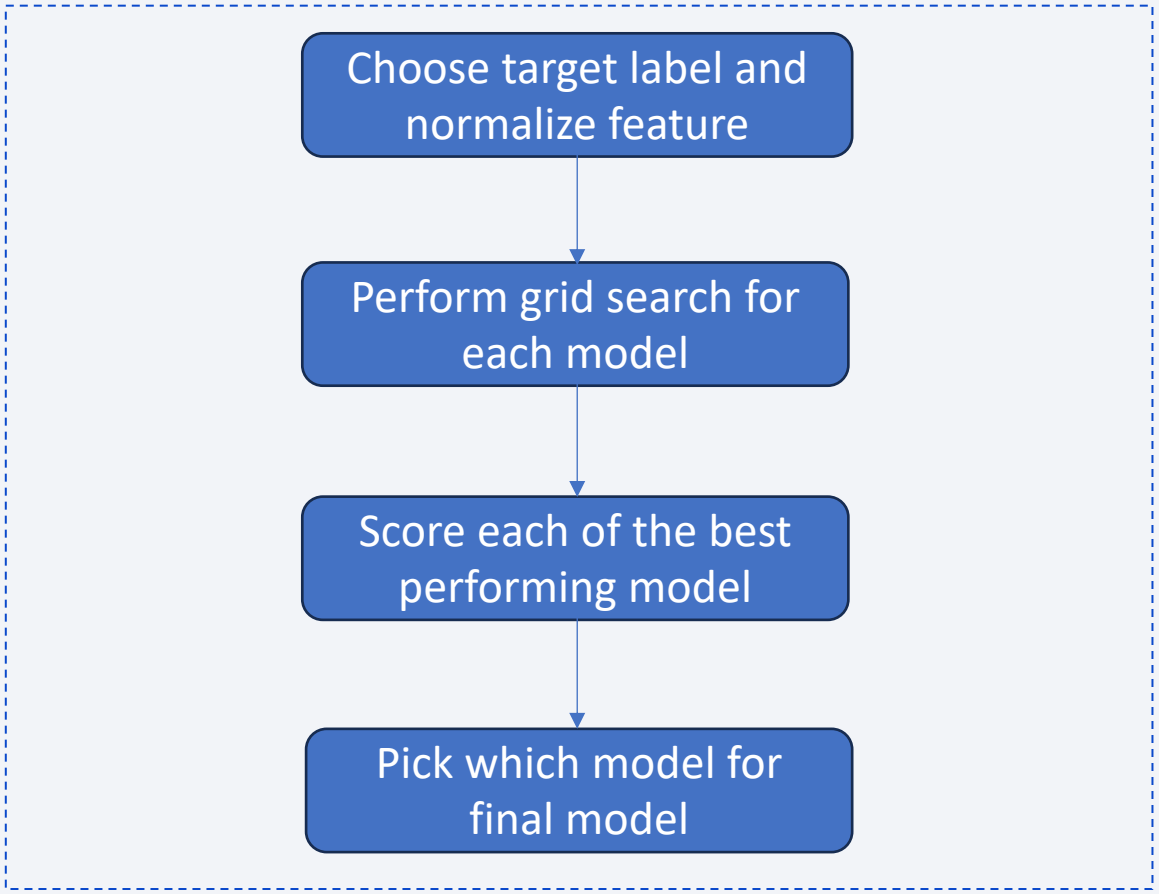
- https://github.com/Jonathan-Dedinata/Data-Science-Capstone/blob/main/Data%20Visualization/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- The following plots and graphs were added into the dashboard:
  - A pie chart showing successful launches by launch site
  - Pie charts showing the success vs failure rates for each launch site
  - A scatter plot of the booster version vs their payload mass
- The pie charts can be accessed through a dropdown menu and the scatter plot can be edited using a slider located above it
- The pie charts were added to observe which launch site showed the highest success rate
- The scatter plot was added to observe which payload ranges and booster version showed the highest success rate
- https://github.com/Jonathan-Dedinata/Data-Science-Capstone/blob/main/Data%20Visualization/spacex_dash_app.py

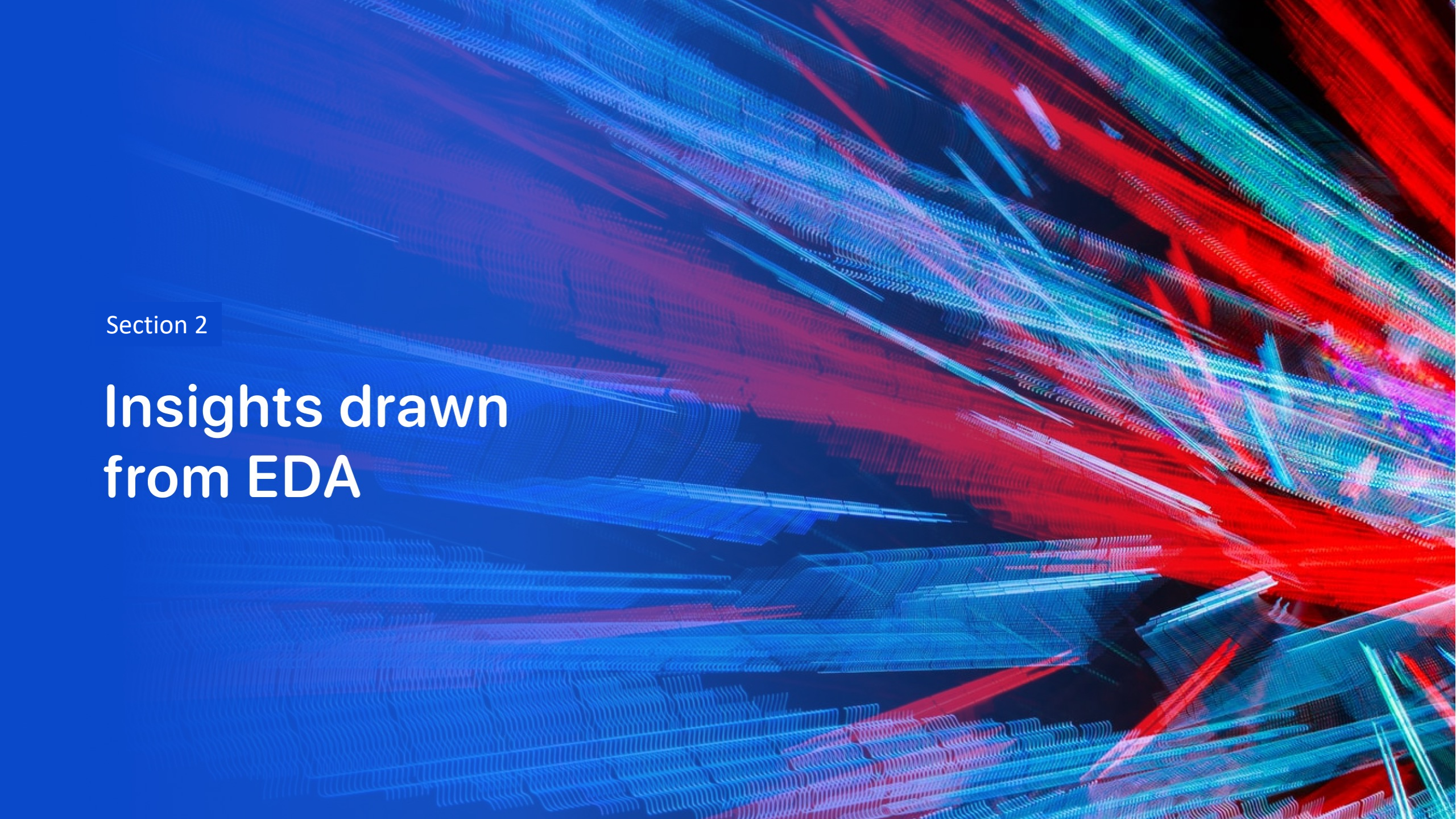
Predictive Analysis (Classification)

- During the predictive analysis section, we first choose the target label, which in our case is the landing outcome. We then normalized the rest of the data features and split the data into training and testing data.
- We then perform grid search with training data on 4 different classification models and testing each model's parameters. The models we used are as follows:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Trees
  - K-Nearest Neighbor (KNN)
- Each model is then tested for accuracy using the test data and we pick which model is best for our predictive analysis
- https://github.com/Jonathan-Dedinata/Data-Science-Capstone/blob/main/Prediction%20Analysis/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

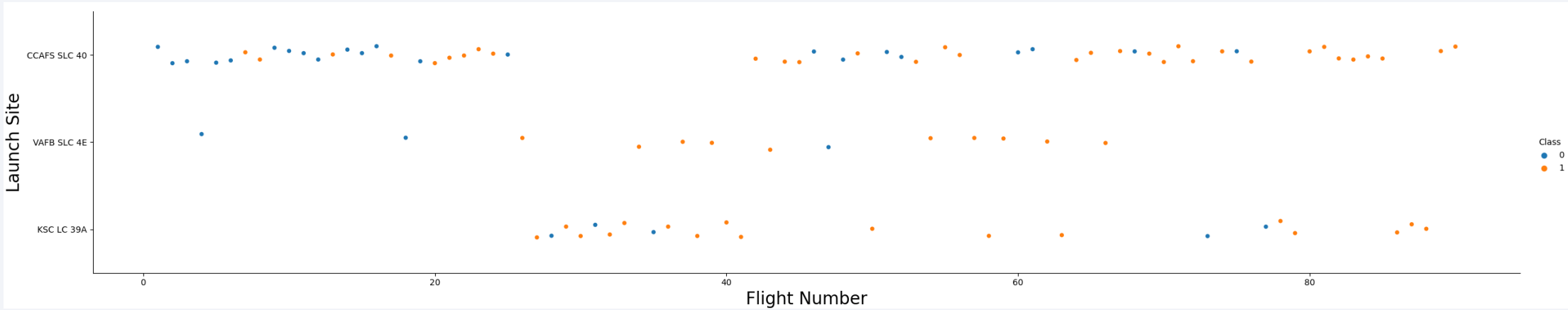


Results

- Exploratory data analysis results:
  - The data shows that as flight number increases, the more likely it is for the first stage to land successfully. Additionally, as the payload mass increases, the less likely it is for the first stage to return. We can also see that LEO orbits are more successful as the flight number increases. For Polar, LEO and ISS, the more the payload mass is, the more successful they are. We also see an increasing success rate trend as the year goes by.
- Interactive analytics:
  - Shown with Plotly, it is shown which site had the highest success to failure ratio and which site overall had the highest success rate. It also showed which booster version carrying how much payload had the highest success rate.
- Predictive analysis results:
  - Through grid searching, we can see which classification model performed the best on predicting what features are needed to make a launch successful. We can see that all the models performed identically with each other. Further testing may be needed but so far, all models performed with 83.33% accuracy.

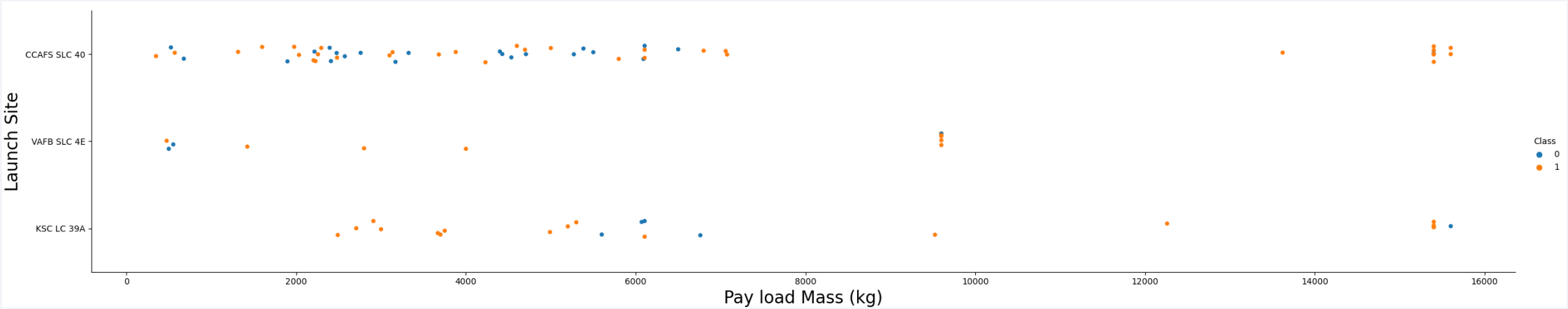

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



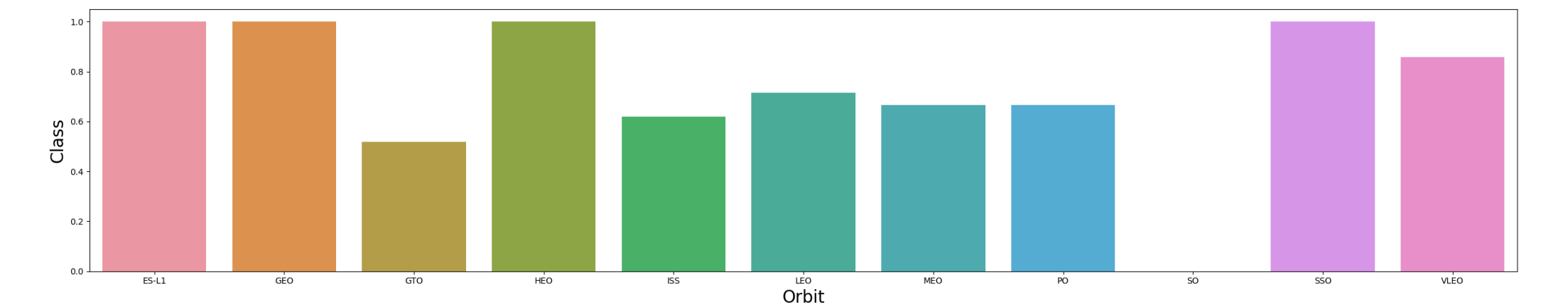
- Shown above is the scatter plot of Flight Number vs. Launch Site. We can see that as the flight number goes up, the outcome leans toward successful regardless of the launch site. As such, we can infer that flight number has a relationship with the success of launches.

Payload vs. Launch Site



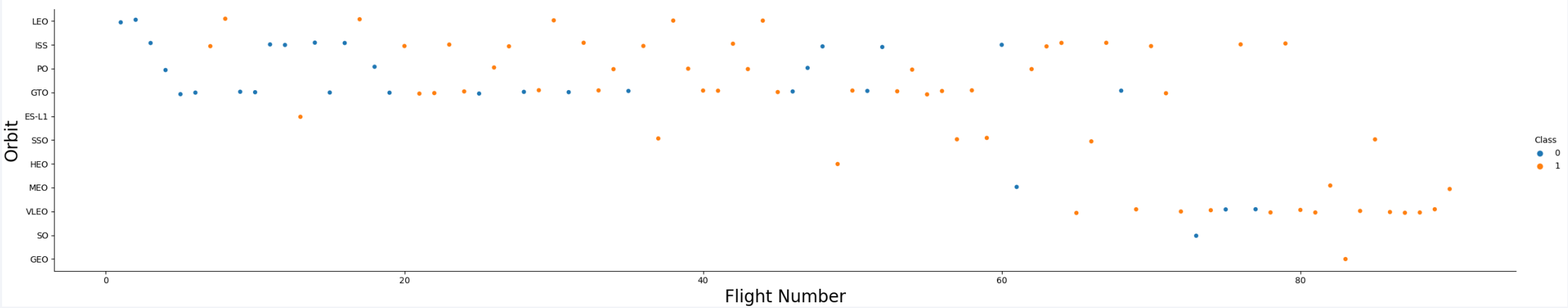
- Shown above is the scatter plot of Payload vs. Launch Site. We can see that the VAFB SLC 4E site does not launch any rockets with a payload higher than 10000 kg. Additionally, we can see that payloads above 8000 kg seem to perform very well in comparison to rockets launched lower than or equal to 6000 kg.

Success Rate vs. Orbit Type



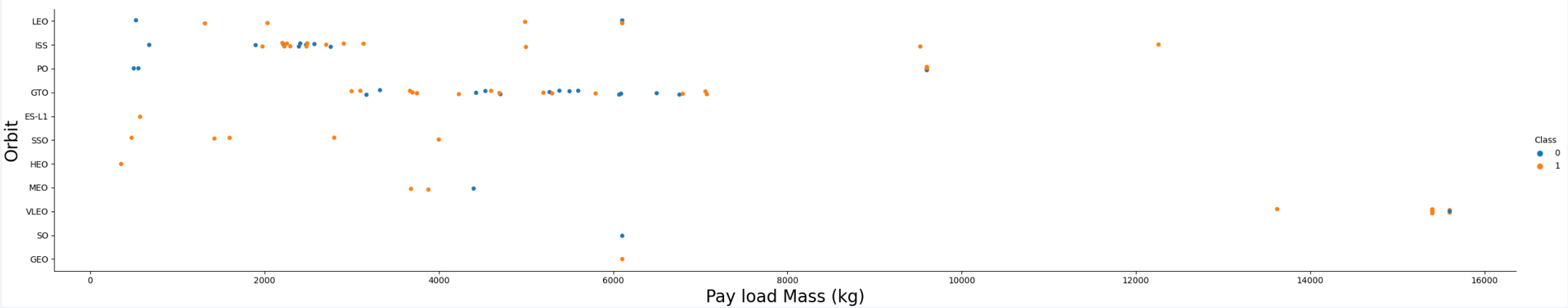
- Shown above is the bar chart for the success rate of each orbit type. We can see that the success rate is 0 when the orbit type is SO. We can also see that there are multiple orbit types that shows a 100% success rate for rocket launches. Given the randomness of the data, we are not sure that we can infer whether success rate is related to orbit type.

Flight Number vs. Orbit Type



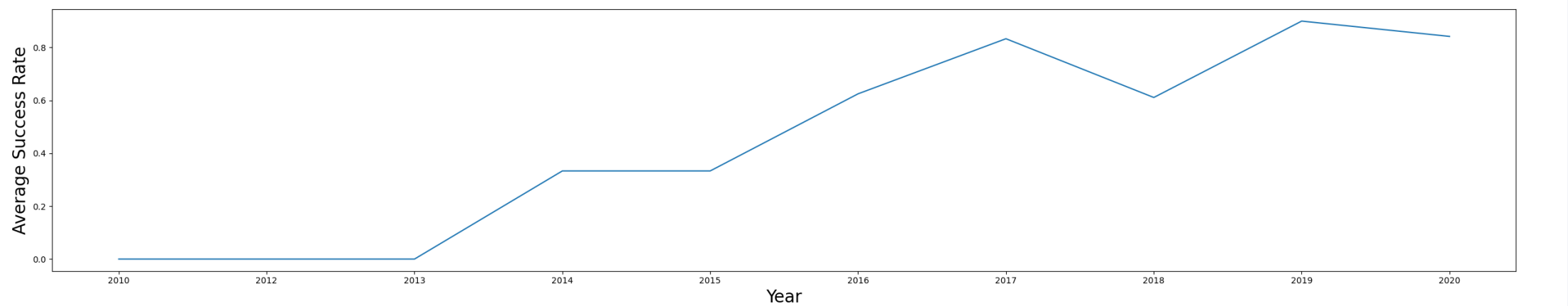
- Shown above is the scatter point of Flight number vs. Orbit type. For the LEO, ISS, and PO orbit, the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- Shown above is the scatter point of payload vs. orbit type. PO, LEO, and ISS seems to show more successful landing rates with heavier payload mass. However, for the GTO orbit type, we cannot infer the same since both success and fail landing rates are present and is not clearly divisible.

Launch Success Yearly Trend



- Show above is the line chart of yearly average success rate. We can see the trend that as the year goes by, the success rate goes up.

All Launch Site Names

- There are 4 launch sites contained in the data. The following are the names of the unique launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- The result was obtained using the following query:
  - %sql SELECT DISTINCT(launch_site) FROM SPACEX

Launch Site Names Begin with 'CCA'

- The following is the result of the query to find 5 records where the name of the launch site begins with ‘CCA’:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- The result was obtained by using the following query :
  - %sql SELECT * FROM SPACEX WHERE launch_site LIKE 'CCA%' LIMIT 5

Total Payload Mass

- Total payload carried by boosters from NASA: 48213 kg
- The result was obtained by using the following query :
  - %sql SELECT SUM(payload_mass_kg_) FROM SPACEX WHERE customer like '%NASA (CRS)%'

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1: 2928 kg
- The result was obtained by using the following query:
  - %sql SELECT AVG(payload_mass_kg_) FROM SPACEX WHERE booster_version = 'F9 v1.1'

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad: 2015-12-22
- The result was obtained by using the following query :
  - %sql SELECT MIN(DATE) FROM SPACEX WHERE landing_outcome like '%ground pad%'

Successful Drone Ship Landing with Payload between 4000 and 6000

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

booster_version	payload_mass_kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- The result was obtained by using the following query:
  - %sql SELECT booster_version, payload_mass_kg_ FROM SPACEX WHERE landing_outcome = 'Success (drone ship)' and payload_mass_kg_ < 6000 and payload_mass_kg_ > 4000

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- The result was obtained by using the following query:
  - %sql SELECT mission_outcome, COUNT(*) FROM SPACEX GROUP BY mission_outcome

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

booster_version	payload_mass_kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- The result was obtained by using the following query:
  - %sql SELECT booster_version, payload_mass_kg_ FROM SPACEX WHERE payload_mass_kg_ = (SELECT MAX(payload_mass_kg_) FROM SPACEX)

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

DATE	landing_outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The result was obtained by using the following query:
  - %sql SELECT DATE, landing_outcome, booster_version, launch_site FROM SPACEX WHERE landing_outcome = 'Failure (drone ship)' and year(DATE) = 2015

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

landing_outcome	2
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

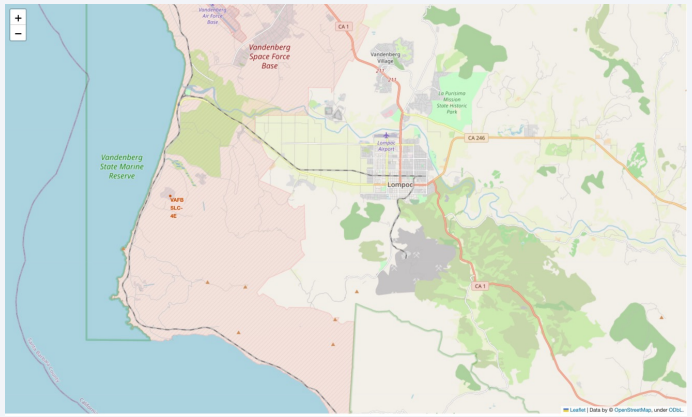
- The result was obtained by using the following query:
  - %sql SELECT landing_outcome, COUNT(*) FROM SPACEX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing_outcome ORDER BY COUNT(*) DESC


Section 3
Launch Sites
Proximities Analysis

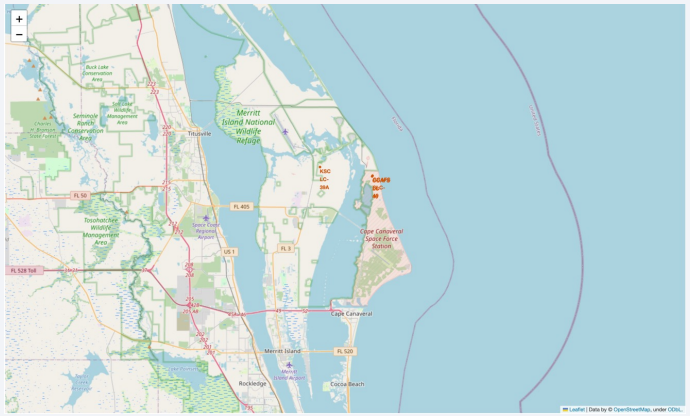
Location of each launch site on the map

- The following are the locations of each launch sites marked on a map:

All



VAFB SLC 4E

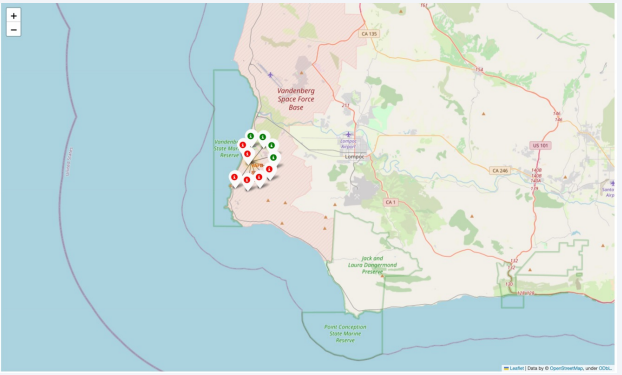


CCAFS S/LC and KSC LC

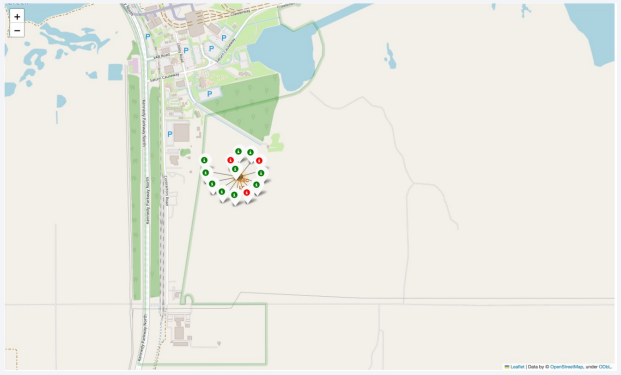
- We can see that all launch sites are located on the extremities of the country, i.e. near coastlines. We can also see how the CCAFS and KSC launch sites are located right next to each other.

Launch outcomes of each launch site

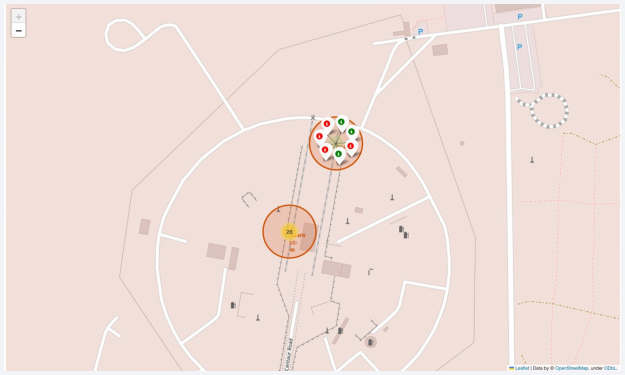
- The following are the launch outcomes for each launch site marked on the map:



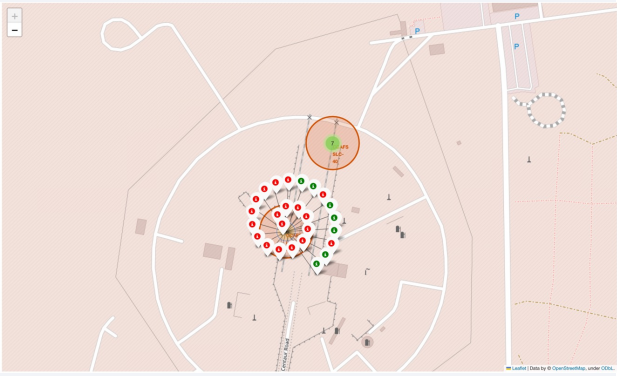
VAFB SLC 4E



KSC LC 39-A



CCAFS SLC-40

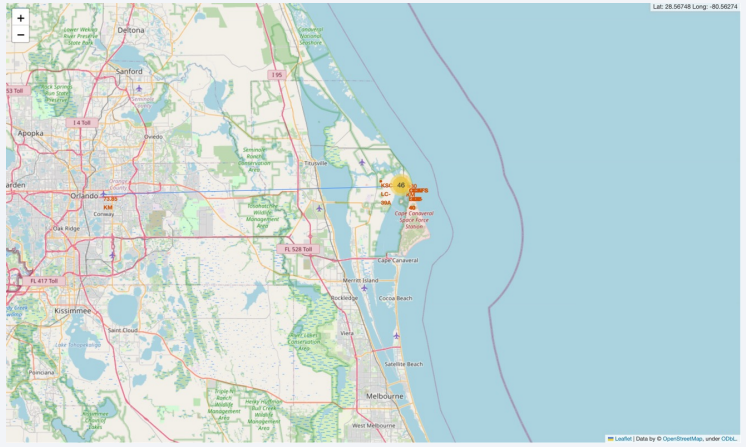


CCAFS LC-40

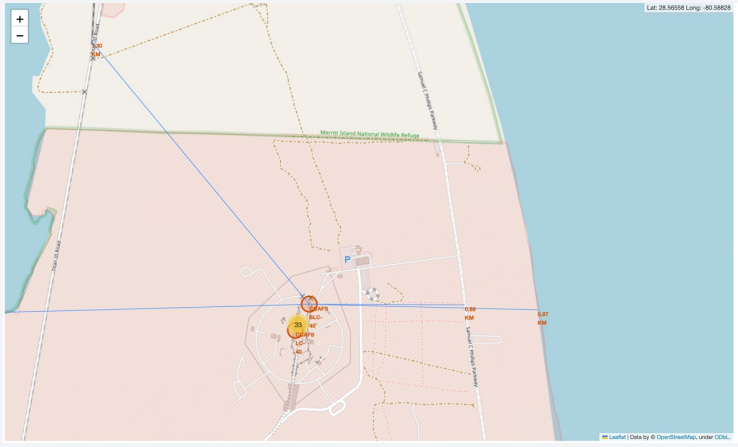
- We can see that the KSC LC 39-A launch site has more successful launch outcomes compared to the other 3 launch sites.

Distance from CCAFS SLC-40 to POI

- The following shows the distance from the launch site CCAFS SLC-40 to near point of interests such as city, railway, highway, and coastline, with distance calculated and displayed

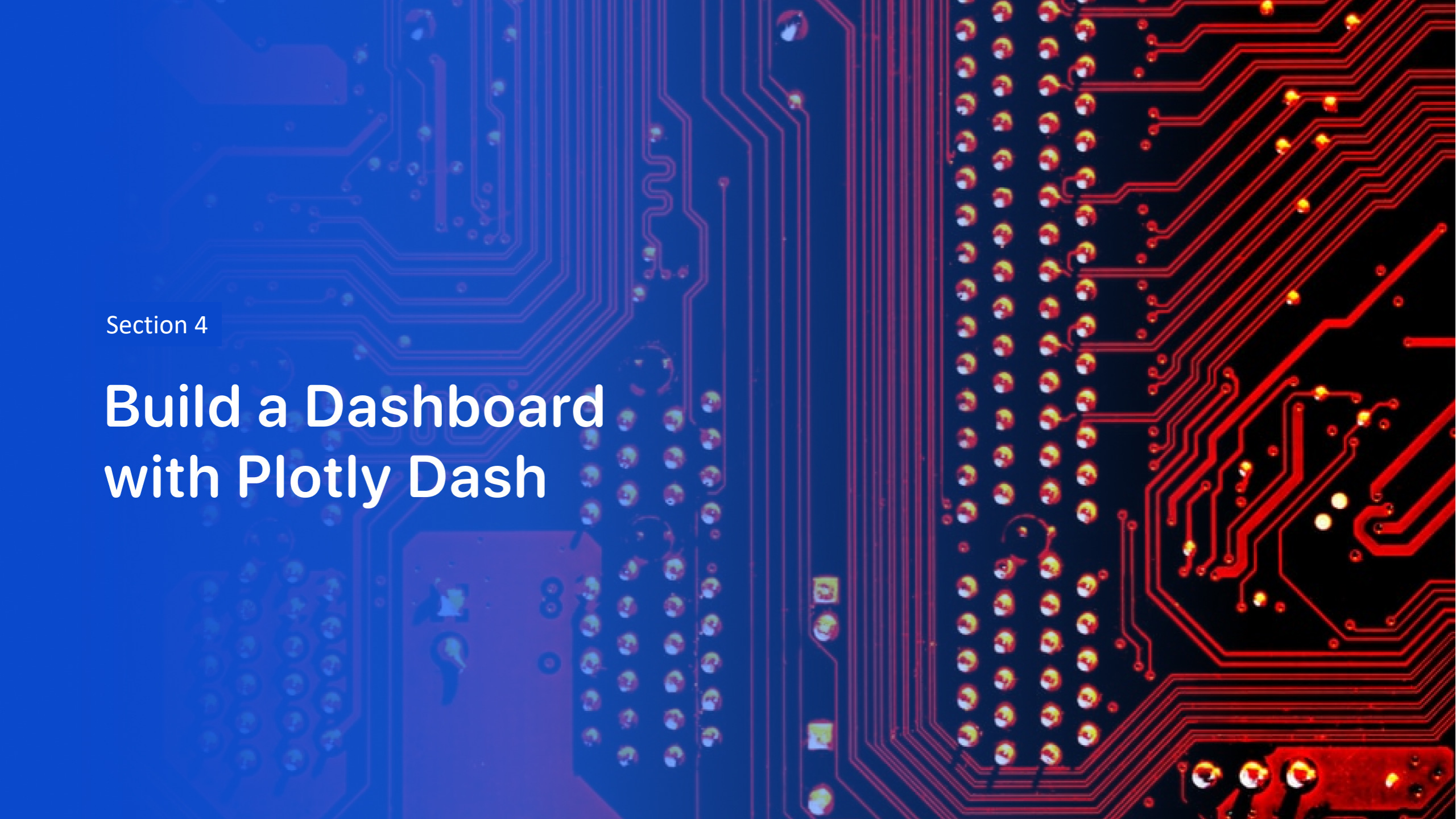


Distance to city



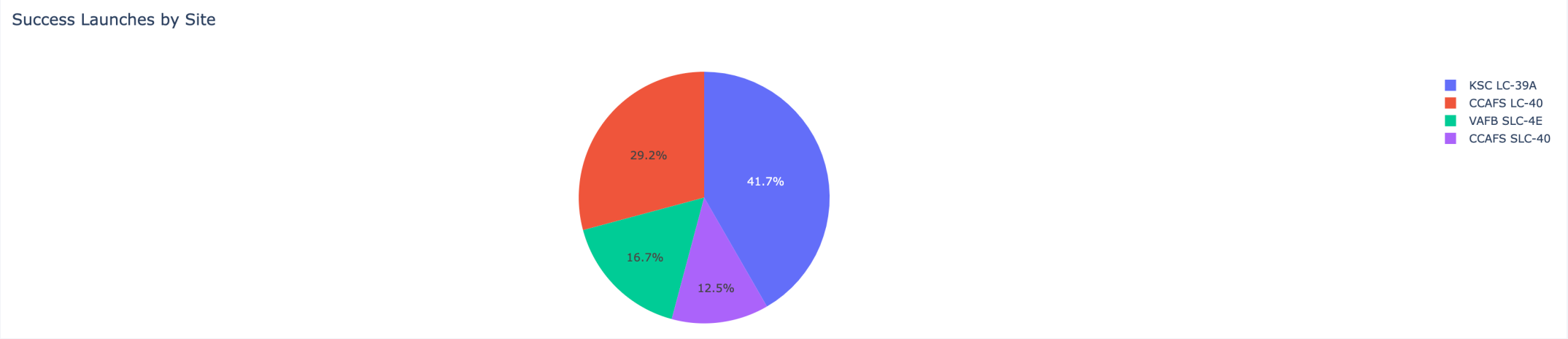
Distance to railway, highway, and coastline

- We can see from above that the launch site CCAFS SLC-40 is very near to a railway, a highway and the coastline. However, it is far from the city comparatively.


Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site

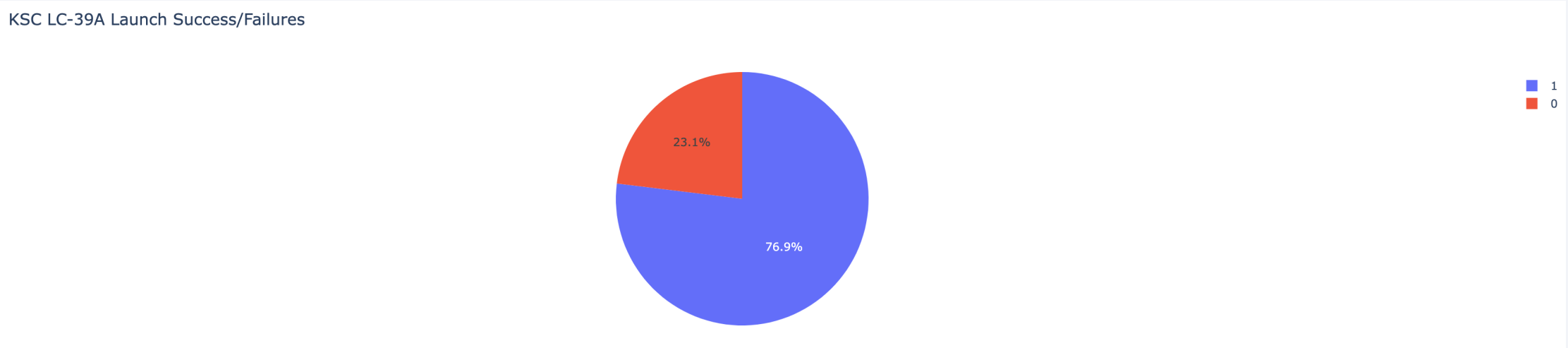
- Shown below is a pie chart of launch success count for all sites



- We can see from the above pie chart that the KSC LC-39A launch site has the most successful launches out of all the launch sites.

KSC LC-39A Success/Failure Ratio

- Shown below is a pie chart showing the KSC LC-39A launch success/failure ratio, which is the highest among the launch sites.



- In the case of this pie chart, 1 indicates a successful launch and 0 shows otherwise.

<Dashboard Screenshot 3>

- Shown below is a scatter plot Payload vs. Launch Outcome for all sites, with different payload selected in the range slider

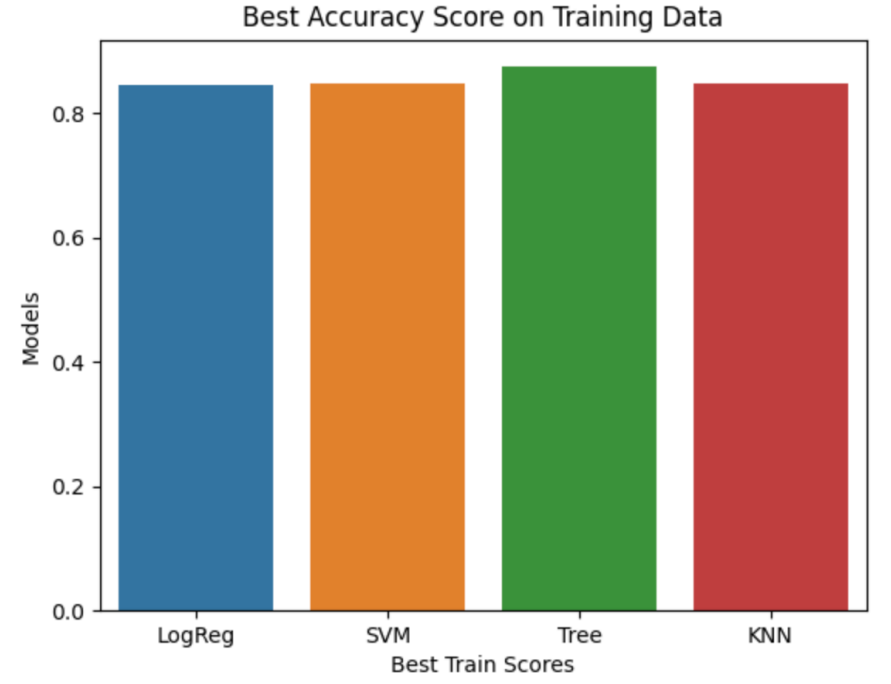
- From the plot above, we can see that the FT booster version shows the highest success rate. Additionally, nearly all of the successful launches had a payload mass lower than 6000 kg.

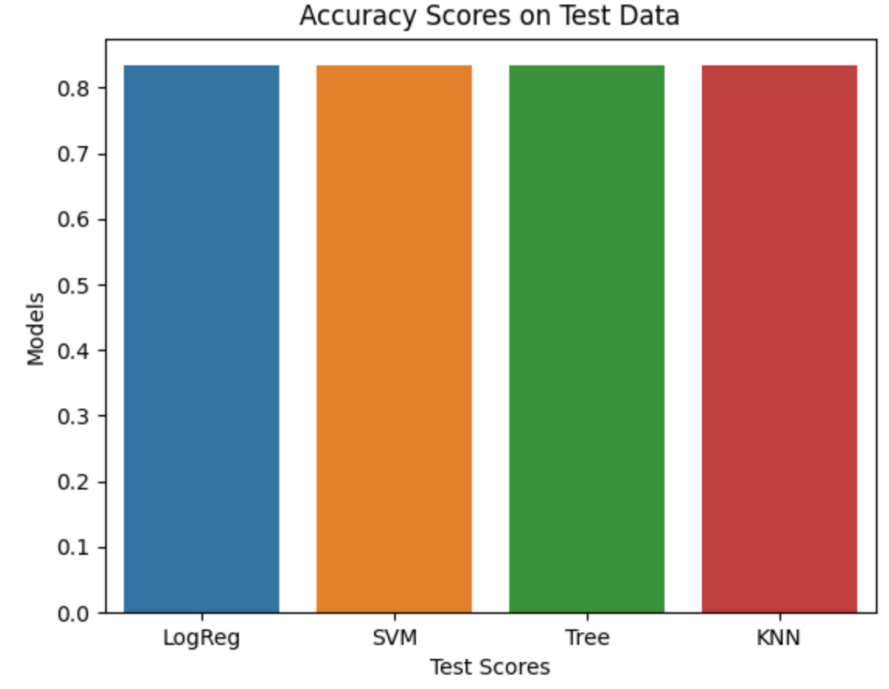
Section 5

Predictive Analysis (Classification)

Classification Accuracy

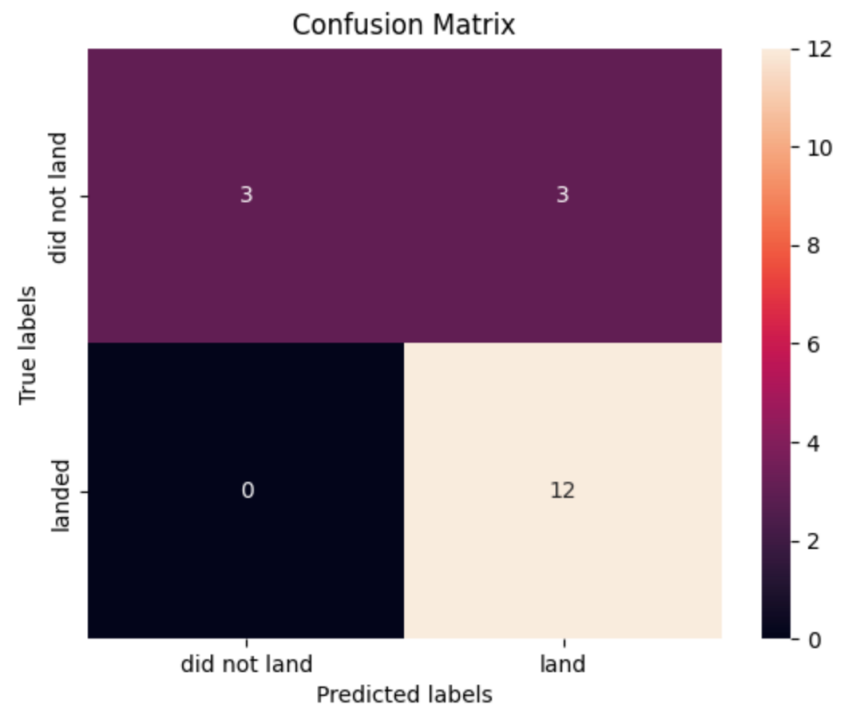
- Shown below are 2 bar charts with the accuracies for each classification model, one evaluated on training data and the other on test data.
- We can see that all models perform similarly in testing data, though the Decision Tree model has a slight edge in training data.





Confusion Matrix

- Shown below is the confusion matrix of the classification models built.
- We can see that the model has an issue with False Positives, with an accuracy rate of 50% for predicting negative values.



Conclusions

- Using the data we gathered, we can infer the following:
  - Flight number is a good indicator on whether a landing outcome is successful or not. Additionally, as the year goes by, the success rate trends upwards.
  - The KSC LC-39A launch site has a higher success rate in comparison to other launch sites.
  - Booster version FT has a higher success rate compared to other booster versions. Additionally, payloads with less than 6000 kg has a higher chance to get the first stage to land successfully.
  - Using these data, we built a predictive classification model and decided on the Decision Tree since it has a higher training accuracy, though its test accuracy is on par with the rest of the models.
  - The Decision Tree was able to predict with an 83.33% testing accuracy.

Thank you!