Outline

- Executive Summary

- Introduction

- Methodology

- Results

- Conclusion

- Appendix

Executive Summary

Summary of Methodologies:

This project follows these steps:
Data Collection
Data Wrangling
Exploratory Data Analysis
Interactive Visual Analytics
Predictive Analysis (Classification)

Summary of Results:

This project produced the following outputs and visualizations:
1. Exploratory Data Analysis (EDA) results
2. Geospatial analytics
3. Interactive dashboard
4. Predictive analysis of classification models

Introduction

- SpaceX launches Falcon 9 rockets at a cost of around $62m. This is considerably cheaper than other providers (which usually cost upwards of $165m), and much of the savings are because SpaceX can land, and then re-use the first stage of the rocket.

- If we can make predictions on whether the first stage will land, we can determine the cost of a launch, and use this information to assess whether or not an alternate company should bid and SpaceX for a rocket launch.

- This project will ultimately predict if the Space X Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

1. Data Collection
   - Making GET requests to the SpaceX REST API
   - Web Scraping

2. Data Wrangling
   - Using the `.fillna()` method to remove NaN values
   - Using the `.value_counts()` method to determine the following:
     - Number of launches on each site
     - Number and occurrence of each orbit
     - Number and occurrence of mission outcome per orbit type
   - Creating a landing outcome label that shows the following:
     - 0 when the booster did not land successfully
     - 1 when the booster did land successfully

3. Exploratory Data Analysis
   - Using SQL queries to manipulate and evaluate the SpaceX dataset
   - Using Pandas and Matplotlib to visualize relationships between variables, and determine patterns

4. Interactive Visual Analytics
   - Geospatial analytics using Folium
   - Creating an interactive dashboard using Plotly Dash

5. Data Modelling and Evaluation
   - Using Scikit-Learn to:
     - Pre-process (standardize) the data
     - Split the data into training and testing data using `train_test_split`
     - Train different classification models
     - Find hyperparameters using `GridSearchCV`
   - Plotting confusion matrices for each classification model
   - Assessing the accuracy of each classification model

Data Collection

- Using the SpaceX API to retrieve data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

1
- Make a GET response to the SpaceX REST API
- Convert the response to a .json file then to a Pandas DataFrame

2
- Use custom logic to clean the data (see Appendix)
- Define lists for data to be stored in
- Call custom functions (see Appendix) to retrieve data and fill the lists
- Use these lists as values in a dictionary and construct the dataset

3
- Create a Pandas DataFrame from the constructed dictionary dataset

4
- Filter the DataFrame to only include Falcon 9 launches
- Reset the FlightNumber column
- Replace missing values of PayloadMass with the mean PayloadMass value

Data Collection - Scraping

- Web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches.

1
- Request the HTML page from the static URL
- Assign the response to an object

2
- Create a BeautifulSoup object from the HTML response object
- Find all tables within the HTML page

3
- Collect all column header names from the tables found within the HTML page

4
- Use the column names as keys in a dictionary
- Use custom functions and logic to parse all launch tables (see Appendix) to fill the dictionary values

5
- Convert the dictionary to a Pandas DataFrame ready for export

Data Wrangling

- Context:
- The SpaceX dataset contains several Space X launch facilities, and each location is in the LaunchSite column.
- Each launch aims to a dedicated orbit, and some of the common orbit types are shown in the figure below. The orbit type is in the Orbit column.
- Initial Data Exploration:
- Using the .value_counts() method to determine the following:
- Number of launches on each site
- Number and occurrence of each orbit
- Number and occurrence of landing outcome per orbit type

EDA with Data Visualization

SCATTER CHARTS

Scatter charts were produced to visualize the relationships between:

- Flight Number and Launch Site
- Payload and Launch Site
- Orbit Type and Flight Number
- Payload and Orbit Type

BAR CHARTS

A bar chart was produced to visualize the relationship between:

- Success Rate and Orbit Type

LINE CHARTS

Line charts were produced to visualize the relationships between:

- Success Rate and Year (i.e. the launch success yearly trend)

EDA with SQL

To gather some information about the dataset, some SQL queries were performed.

- The SQL queries performed on the data set were used to:

1. Display the names of the unique launch sites in the space mission

2. Display 5 records where launch sites begin with the string 'CCA'

3. Display the total payload mass carried by boosters launched by NASA (CRS)

4. Display the average payload mass carried by booster version F9 v1.1

5. List the date when the first successful landing outcome on a ground pad was achieved

6. List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg

7. List the total number of successful and failed mission outcomes

8. List the names of the booster versions which have carried the maximum payload mass

9. List the failed landing outcomes on drone ships, their booster versions, and launch site names for 2015

10. 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

Build an Interactive Map with Folium

The following steps were taken to visualize the launch data on an interactive map:

1. Mark all launch sites on a map
   - Initialise the map using a Folium `Map` object
   - Add a `folium.Circle` and `folium.Marker` for each launch site on the launch map

2. Mark the success/failed launches for each site on a map
   - As many launches have the same coordinates, it makes sense to cluster them together.
   - Before clustering them, assign a marker colour of successful (class = 1) as green, and failed (class = 0) as red.
   - To put the launches into clusters, for each launch, add a `folium.Marker` to the `MarkerCluster()` object.
   - Create an icon as a text label, assigning the `icon_color` as the `marker_colour` determined previously.

3. Calculate the distances between a launch site to its proximities
   - To explore the proximities of launch sites, calculations of distances between points can be made using the `Lat` and `Long` values.
   - After marking a point using the `Lat` and `Long` values, create a `folium.Marker` object to show the distance.
   - To display the distance line between two points, draw a `folium.PolyLine` and add this to the map.

Build a Dashboard with Plotly Dash

1. Pie chart (`px.pie()`) showing the total successful launches per site
   - This makes it clear to see which sites are most successful
   - The chart could also be filtered (using a `dcc.Dropdown()` object) to see the success/failure ratio for an individual site

2. Scatter graph (`px.scatter()`) to show the correlation between outcome (success or not) and payload mass (kg)
   - This could be filtered (using a `RangeSlider()` object) by ranges of payload masses
   - It could also be filtered by booster version

13

Predictive Analysis (Classification)

The following steps were taking to develop, evaluate, and find the best performing classification model:

Model Development

- To prepare the dataset for model development:
  - Load dataset
  - Perform necessary data transformations (standardise and pre-process)
  - Split data into training and test data sets, using `train_test_split()`
  - Decide which type of machine learning algorithms are most appropriate
- For each chosen algorithm:
  - Create a `GridSearchCV` object and a dictionary of parameters
  - Fit the object to the parameters
  - Use the training data set to train the model

Model Evaluation

- For each chosen algorithm:
  - Using the output GridSearchCV object:
    Check the tuned hyperparameters (`best_params_`)
    Check the accuracy (`score` and `best_score_`)
  - Plot and examine the Confusion Matrix

Finding the Best Model

- Review the accuracy scores for all chosen algorithms
- The model with the highest accuracy score is determined as the best performing model

14

Results

- Exploratory data analysis results

- Interactive analytics demo in screenshots

- Predictive analysis results

Section 2

# Insights drawn from EDA

Flight Number vs. Launch Site

The scatter plot of Launch Site vs. Flight Number shows that:
- As the number of flights increases, the rate of success at a launch site increases.
- Most of the early flights (flight numbers < 30) were launched from CCAFS SLC 40, and were generally unsuccessful.
- The flights from VAFB SLC 4E also show this trend, that earlier flights were less successful.
- No early flights were launched from KSC LC 39A, so the launches from this site are more successful.
- Above a flight number of around 30, there are significantly more successful landings (Class = 1).
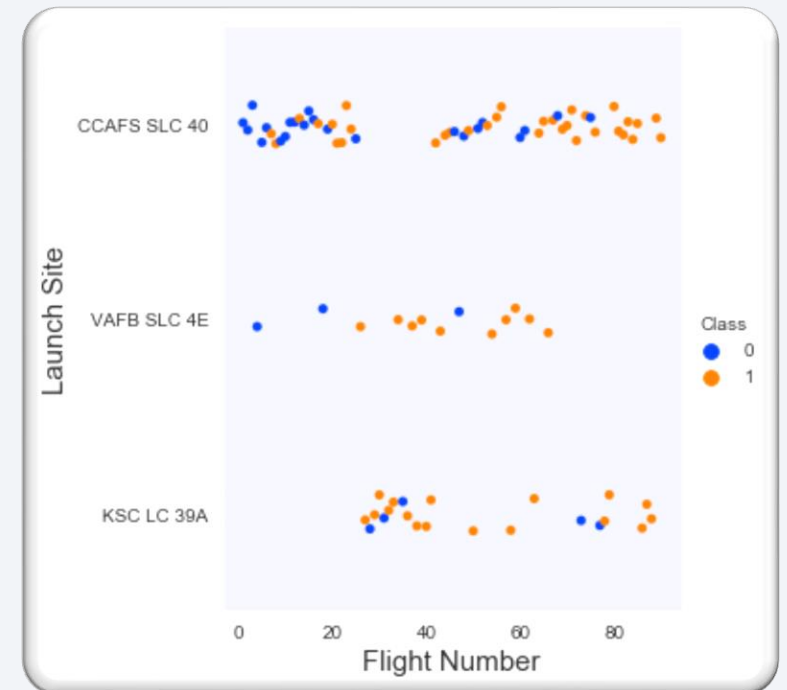
Payload vs. Launch Site

- The scatter plot of Launch Site vs. Payload Mass shows that:

- Above a payload mass of around 7000 kg, there are very few unsuccessful landings, but there is also far less data for these heavier launches.

- There is no clear correlation between payload mass and success rate for a given launch site.

- All sites launched a variety of payload masses, with most of the launches from CCAFS SLC 40 being comparatively lighter payloads (with some outliers).
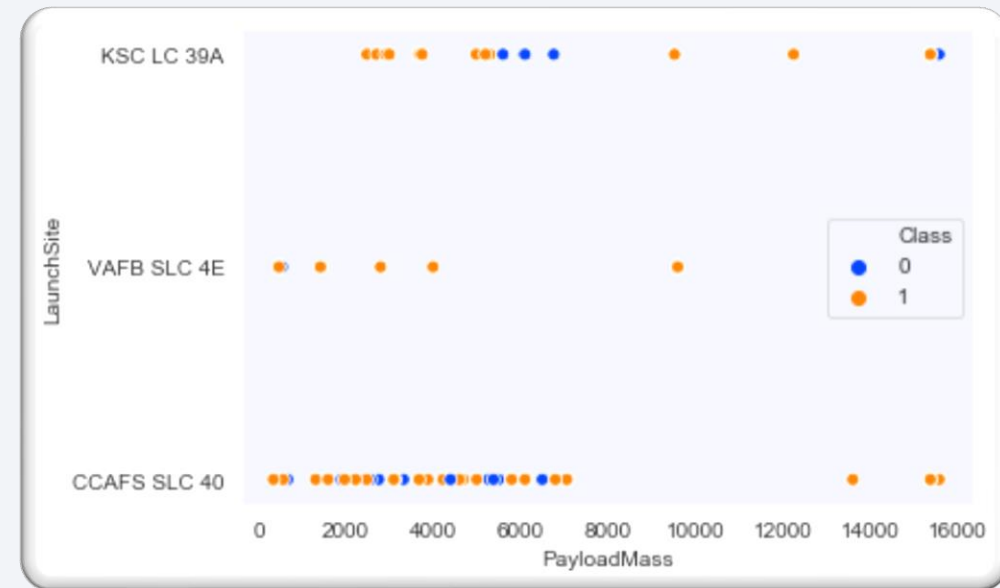
Success Rate vs. Orbit Type
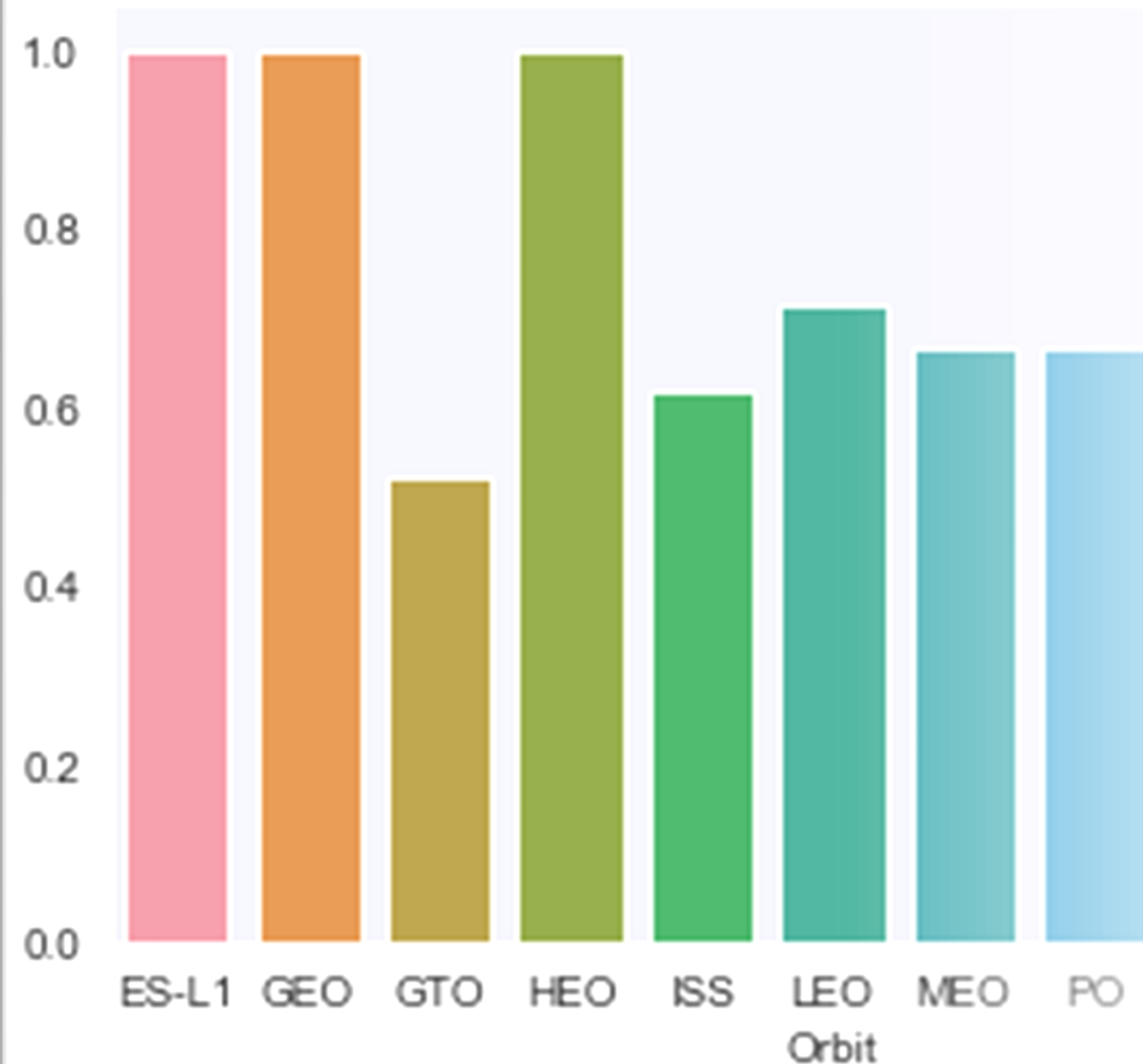
- The bar chart of Success Rate vs. Orbit Type shows that the following orbits have the highest (100%) success rate:
 - ES-L1 (Earth-Sun First Lagrangian Point)
  - GEO (Geostationary Orbit)
  - HEO (High Earth Orbit)
  - SSO (Sun-synchronous Orbit)

- The orbit with the lowest (0%) success rate is:
 - SO (Heliocentric Orbit)

Flight Number vs. Orbit Type

- This scatter plot of Orbit Type vs. Flight number shows a few useful things that the previous plots did not, such as:

- The 100% success rate of GEO, HEO, and ES-L1 orbits can be explained by only having 1 flight into the respective orbits.

- The 100% success rate in SSO is more impressive, with 5 successful flights.

- There is little relationship between Flight Number and Success Rate for GTO.

- Generally, as Flight Number increases, the success rate increases. This is most extreme for LEO, where unsuccessful landings only occurred for the low flight numbers (early launches).
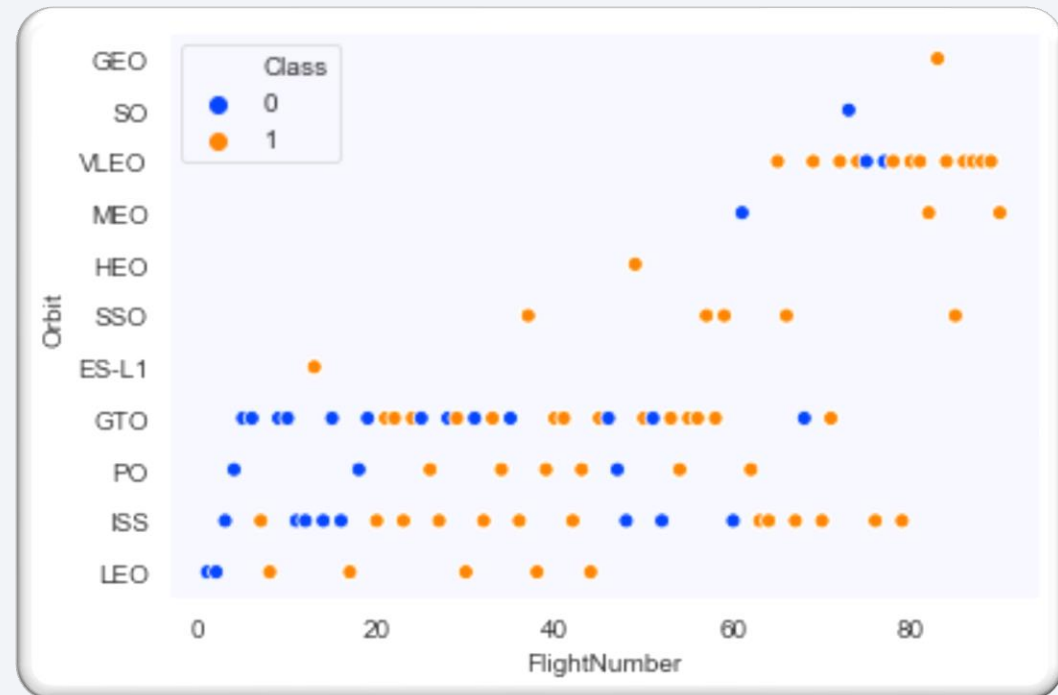
Payload vs. Orbit Type

- This scatter plot of Orbit Type vs. Payload Mass shows that:

- The following orbit types have more success with heavy payloads:
  - PO (although the number of data points is small)
  - ISS
  - LEO

- For GTO, the relationship between payload mass and success rate is unclear.

- VLEO (Very Low Earth Orbit) launches are associated with heavier payloads, which makes intuitive sense.
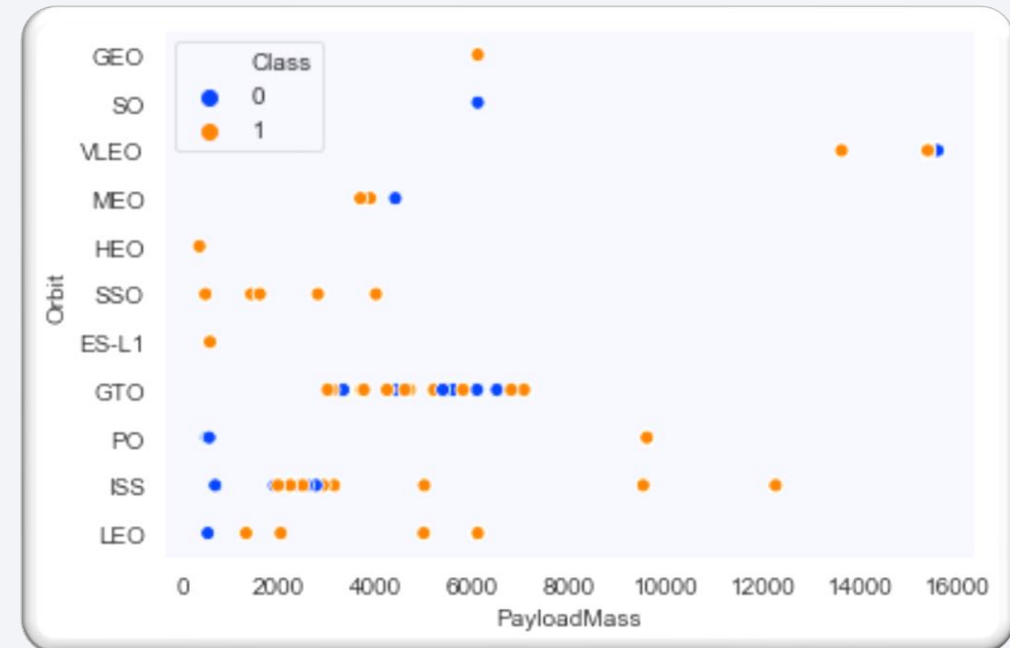
Launch Success Yearly Trend

- The line chart of yearly average success rate shows that:

- Between 2010 and 2013, all landings were unsuccessful (as the success rate is 0).

- After 2013, the success rate generally increased, despite small dips in 2018 and 2020.

- After 2016, there was always a greater than 50% chance of success.

All Launch Site Names

- Find the names of the unique launch sites

- Present your query result with a short explanation here

- %sql SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXTBL;

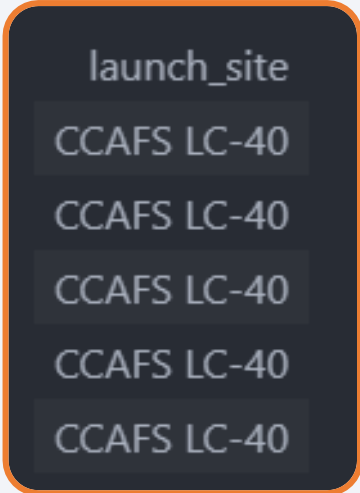| launch_site |
| --- |
| CCAFS LC-40 |
| CCAFS SLC-40 |
| KSC LC-39A |
| VAFB SLC-4E |

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

- Present your query result with a short explanation here

%sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5'

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

- Present your query result with a short explanation here

%sql SELECT SUM(PAYLOAD_MASS_KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL \

WHERE CUSTOMER = 'NASA (CRS)':   = 45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

- Present your query result with a short explanation here

%sql SELECT AVG(PAYLOAD_MASS_KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEXTBL \

WHERE BOOSTER_VERSION = 'F9 v1.1'; = 2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

- Present your query result with a short explanation here

%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTBL \

WHERE LANDING_OUTCOME = 'Success (ground pad)';

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

- Present your query result with a short explanation here

%sql SELECT BOOSTE_VERSION FROM SPACEXTBL \

      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

- Calculate the total number of successful and failure mission outcomes

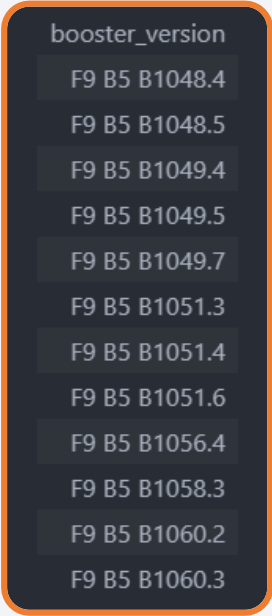- Present your query result with a short explanation here

%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME

mission_outcome	total_number
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

- Present your query result with a short explanation here

%sql SELECT DISTINCT (BOOSTER_VERSION) FROM SPACEXTBL \

WHERE PAYLOAD_MASS_KG_ = (SELECT MAX(PAYLOAD_MASS_KG_)

FROM SPACEXTBL)

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

2015 Launch Records

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

- Present your query result with a short explanation here

%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL

WHERE (LANDING_OUTCOME = 'FAILURE (drone ship)') AND (EXTRACT(YEAR FROM DATE) = '2015')

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
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

landing_outcome	total_number
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

- Present your query result with a short explanation here

%sql SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL \

WHERE DATE BETWEEN '2010-06-04' AND '2-17-03-20' \
GROUP BY LANDING_OUTCOME \
ORDER BY TOTAL_NUMBER DESC;

Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- Replace <Folium map screenshot 1> title with an appropriate title

- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map

- Explain the important elements and findings on the screenshot

\<Folium Map Screenshot 2\>

- Replace \<Folium map screenshot 2\> title with an appropriate title

- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map

- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title

- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

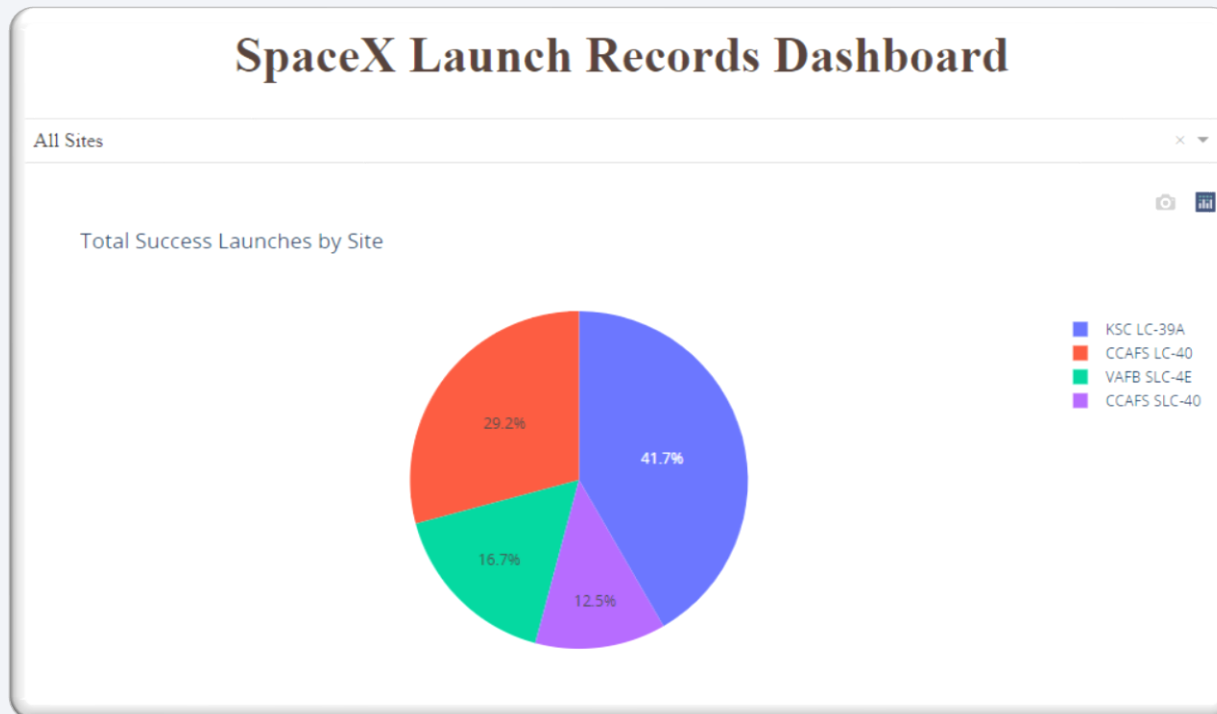- Explain the important elements and findings on the screenshot

Section 4

Build a Dashboard
with Plotly Dash

<Dashboard Screenshot 1>
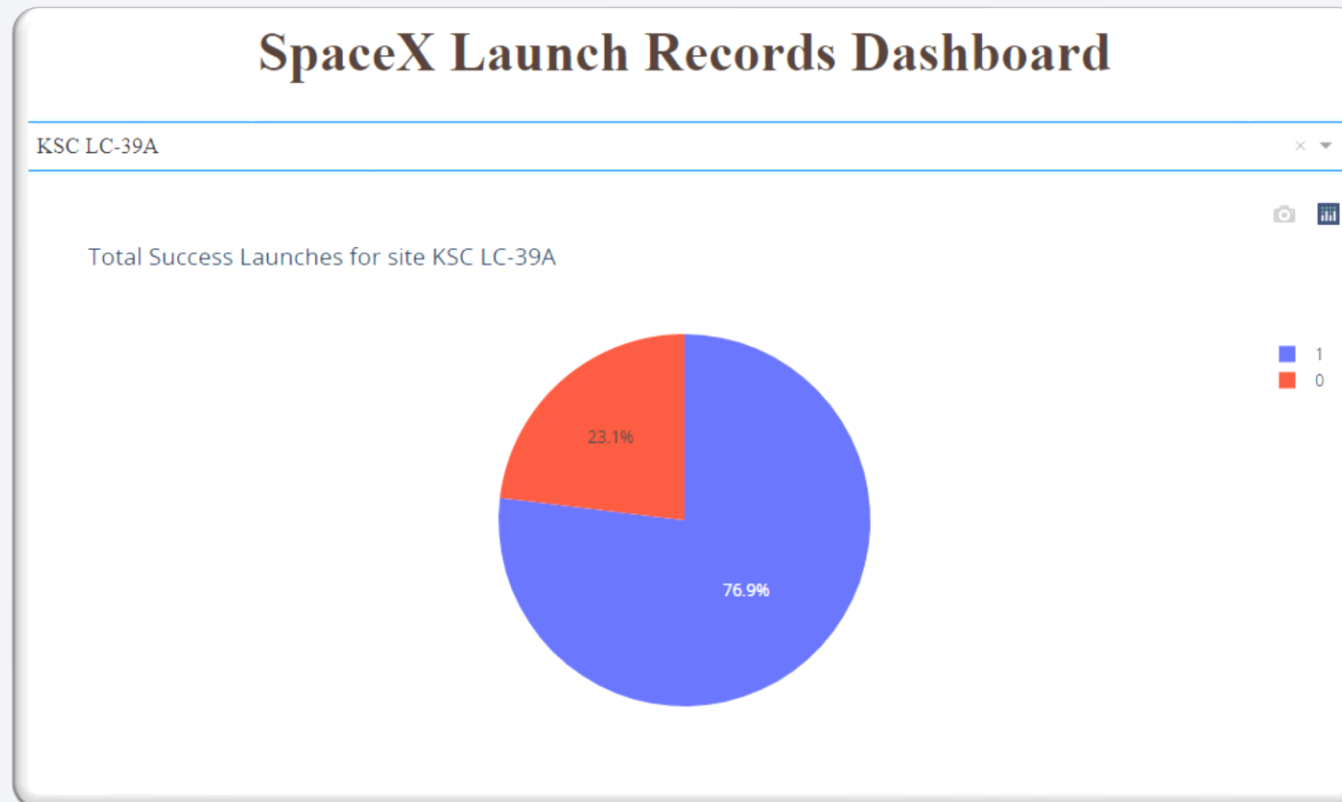
- Replace <Dashboard screenshot 1> title with an appropriate title

- Show the screenshot of launch success count for all sites, in a piechart

- Explain the important elements and findings on the screenshot

- The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches.

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title

- Show the screenshot of the piechart for the launch site with highest launch success ratio

- Explain the important elements and findings on the screenshot

The launch site KSC LC-39 A also had the highest rate of successful launches, with a 76.9% success rate.

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title

- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

# Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

- Find which model has the highest classification accuracy

 Plotting the Accuracy Score and Best Score for each classification algorithm produces the following result:
  - The Decision Tree model has the highest classification accuracy
    - The Accuracy Score is 94.44%
    - The Best Score is 90.36%

Algorithm	Accuracy Score	Best Score
Logistic Regression	0.833333	0.846429
Support Vector Machine	0.833333	0.848214
Decision Tree	0.944444	0.903571
K Nearest Neighbours	0.888889	0.876786
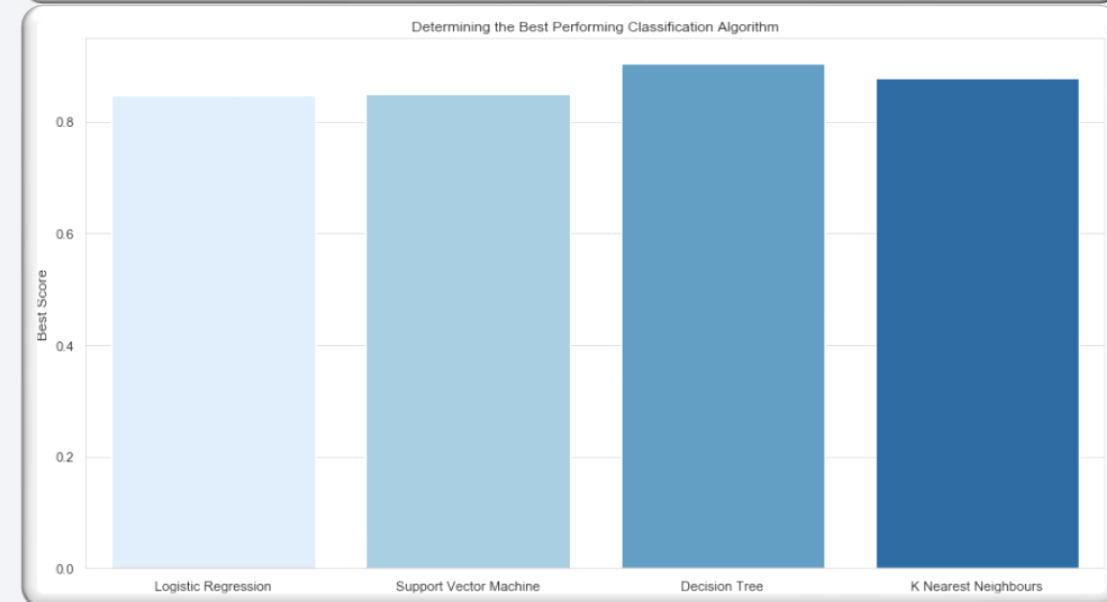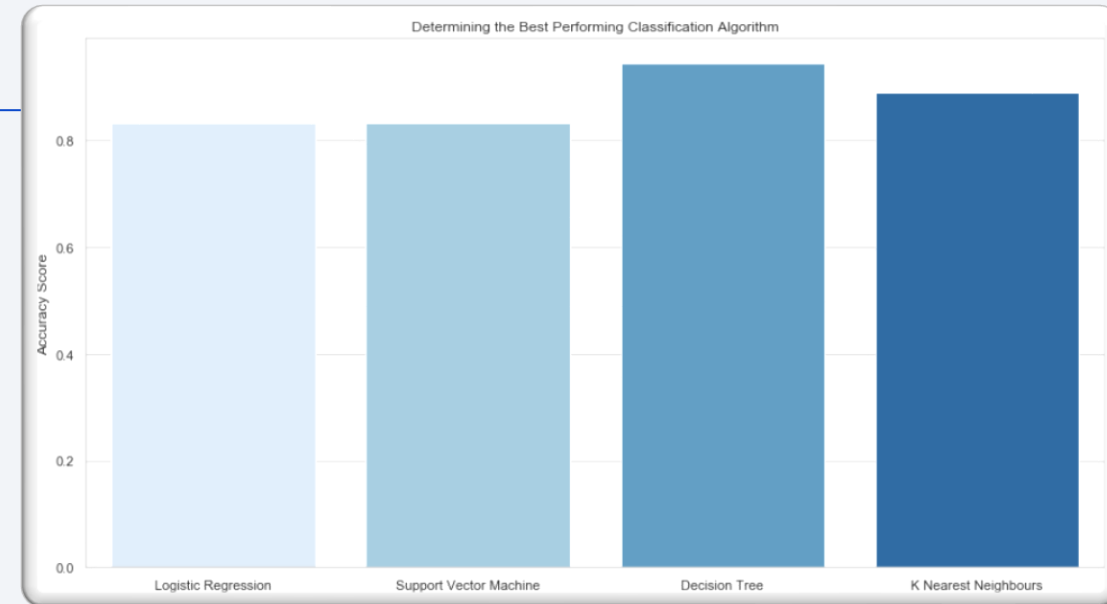
Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation

- As shown previously, best performing classification model is the Decision Tree model, with an accuracy of 94.44%.

- This is explained by the confusion matrix, which shows only 1 out of 18 total results classified incorrectly (a false positive, shown in the top-right corner).

- The other 17 results are correctly classified (5 did not land, 12 did land).
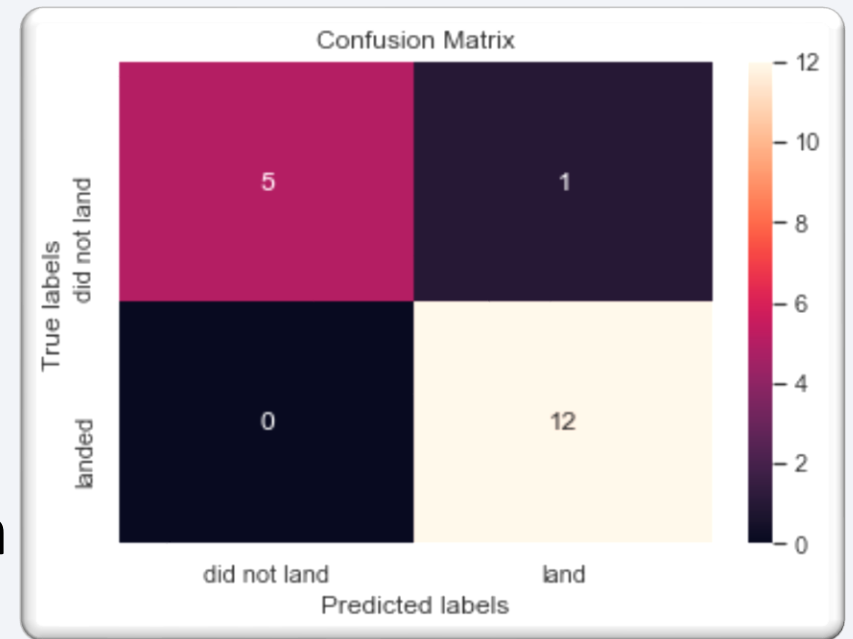
Conclusions

- As the number of flights increases, the rate of success at a launch site increases, with most early flights being unsuccessful. I.e. with more experience, the success rate increases.
  - Between 2010 and 2013, all landings were unsuccessful (as the success rate is 0).
  - After 2013, the success rate generally increased, despite small dips in 2018 and 2020.
  - After 2016, there was always a greater than 50% chance of success.

- Orbit types ES-L1, GEO, HEO, and SSO, have the highest (100%) success rate.
    The 100% success rate of GEO, HEO, and ES-L1 orbits can be explained by only having 1 flight into the respective orbits.
    The 100% success rate in SSO is more impressive, with 5 successful flights.
    The orbit types PO, ISS, and LEO, have more success with heavy payloads:
    VLEO (Very Low Earth Orbit) launches are associated with heavier payloads, which makes intuitive sense.

- The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches, and also the highest rate of successful launches, with a 76.9% success rate.

- The success for massive payloads (over 4000kg) is lower than that for low payloads.

- The best performing classification model is the Decision Tree model, with an accuracy of 94.44%.

Appendix

- N/A

Thank you!