Winning the Space Race with Data Science

Caoimhe Coveney McKeown

02/12/2024

Outline

1. Executive Summary
2. Introduction
3. Methodology
4. Results
5. Conclusions

Executive Summary

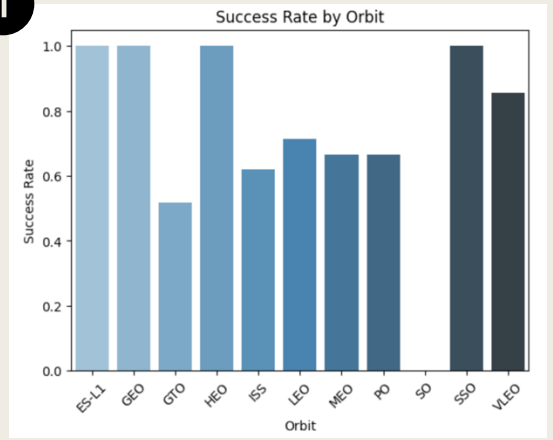
Summary of methodologies:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Interactive Visual Analytics
- Predictive Analysis (Classification)

Summary of all results:

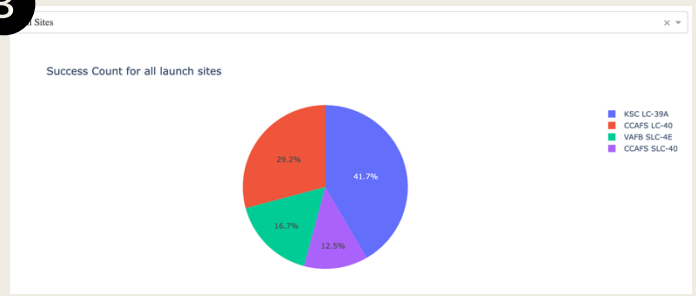
1. Exploratory Data Analysis (EDA) results
2. Geospatial Analytics
3. Interactive Dashboard
4. Predictive Analysis of Classification Models

1

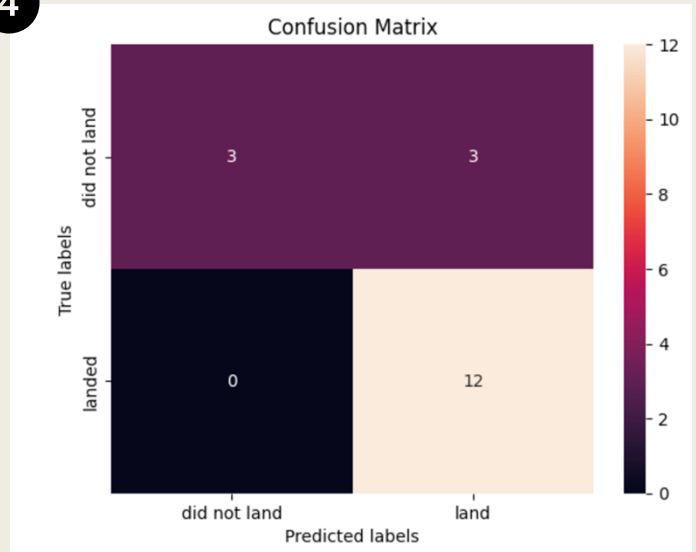


2

3



4



Introduction

Background

SpaceX is one of the most successful companies in the commercial space age, with a goal of making space travel more affordable. Since its launch in 2001, SpaceX has sent spacecrafts to the International Space Station, launched a satellite constellation with the aim of providing internet access and sent manned missions into space. SpaceX has managed to achieve all of this as their rocket launches are relatively inexpensive ($62 million per launch) in comparison with competitors ($165 million per launch). The reduced cost of SpaceX launches is due to their re-use of the first stage of its Falcon 9 rocket – a novel concept. Using data science to determine if the first stage will land, we can determine the price the launch will be. To achieve this, we will use publicly available data as well as machine learning models to predict whether SpaceX can reuse the first stage of their rockets.

Explore

- The impact of payload mass, launch site, number of flights and orbits have on the success rate of the first stage landing
- The rate of landing successes over time
- Find the nest predictive model for successful landing – using binary classification

METHODOLOGY

Section 1

Methodology

1. **Data Collection**
   - *Collect the data using SpaceX REST API and web scraping techniques*
2. **Data Wrangling**
   - *Filter the data and handle missing values*
   - *Prepare the data for analysis and modelling*
3. **Exploratory Data Analysis**
   - *Explore the data with SQL*
   - *Employ data visualization techniques*
4. **Interactive Visual Analytics**
   - *Visualise the data using Folium for geospatial analytics*
   - *Create an interactive Dashboard using Plotly Dash*
5. **Data Modelling and Evaluation**
   - *Split the data into test and train sets using train_test_split function*
   - *Build classification models to predict the landing outcomes*
   - *Plot confusion matrices for the different classification models*
   - *Assess the accuracy of each model to find the best model and parameters*

Data Collection – SpaceX API

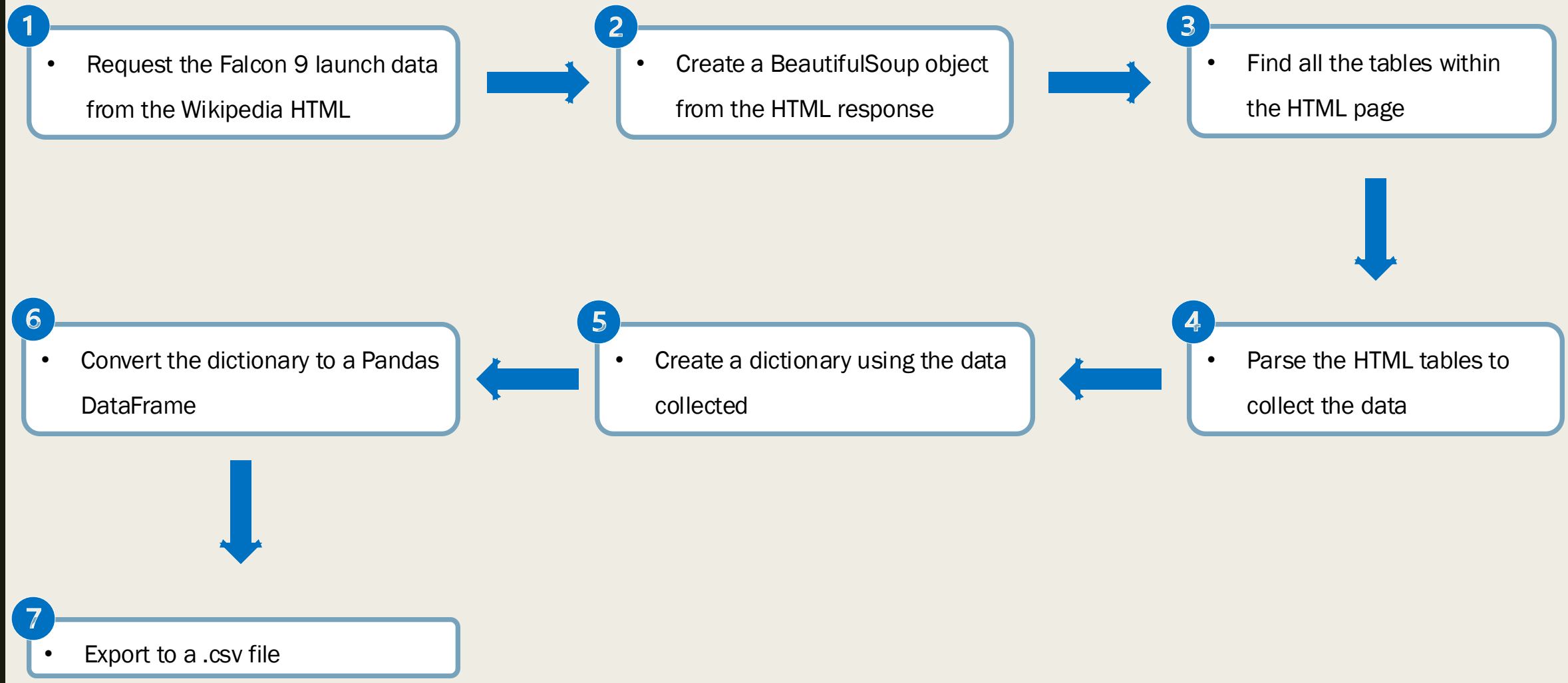
1
- Request data from SpaceX REST API
- Convert response to a .json file and then to a Pandas dataframe

2
- Clean the data using custom logic
- Define lists to store the data in and retrieve data to fill the lists
- Use the lists as values in a dictionary
- Construct the dataset

3
- Create a Pandas dataframe using the constructed dictionary as the dataset

4
- Filter the dataframe so it only contains information on Falcon 9 launches
- Replace the missing values of Payload Mass Kg with the calculated mean of Payload Mass
- Export the data to a .csv file

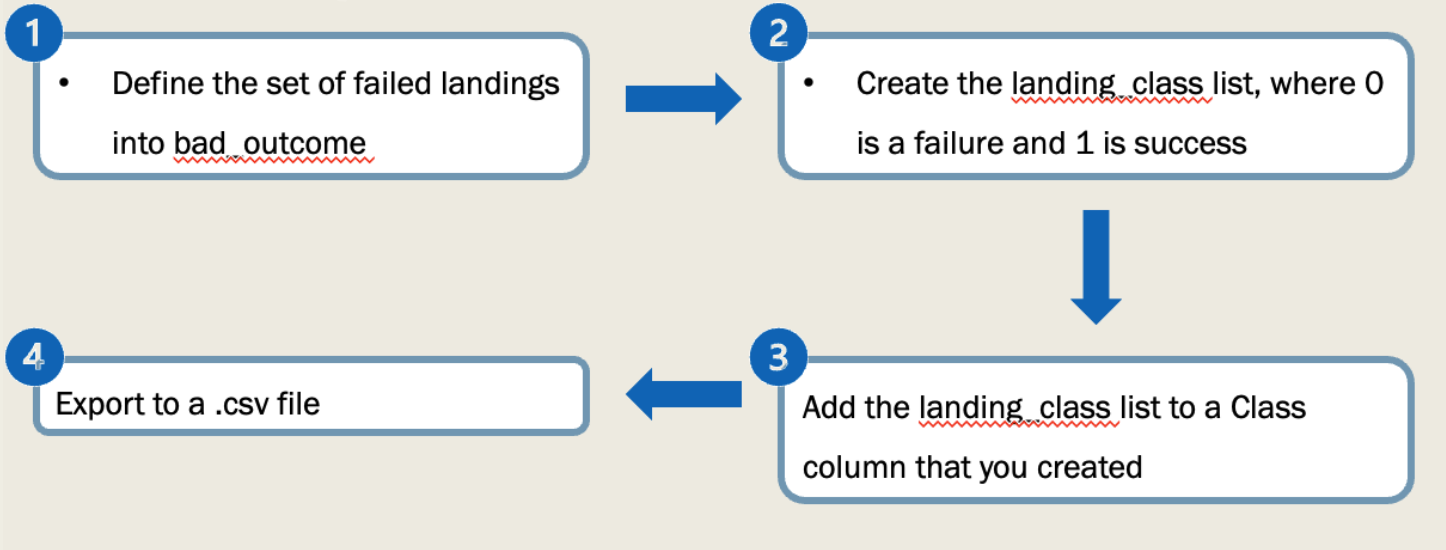
Data Collection – Web Scraping



Data Wrangling

Steps:

- To determine whether a booster will land successfully or not, a binary column was created, where the value of 1 equals a successful landing and 0 equals a failed landing
- This was done by:



Landing Outcomes

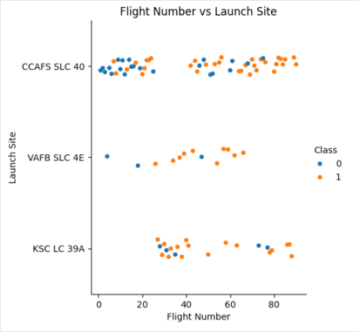
- **True Ocean** – successful mission in specific ocean region
- **False Ocean** – failed mission in specific ocean region
- **True RTLS** – successful mission landed on ground pad
- **False RTLS** – failed mission of landing on ground pad
- **True ASDS** – successful mission landed on drone ship
- **False ASDS** – failed mission of landing on drone ship
- **None ASDS** and **None None** – both represent a failure to land

EDA with Data Visualisation

Scatter Plots

Scatter plots were used to analyse the relationships between the following variables:

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

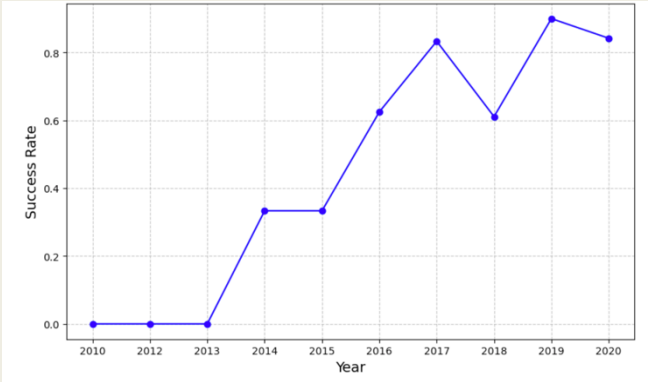


Scatter plots are used to observe relationships between two numeric variables

Line Charts

Line charts were used to visualise the relationship between the success rate and year:

- Launch Success Rate Yearly Trend

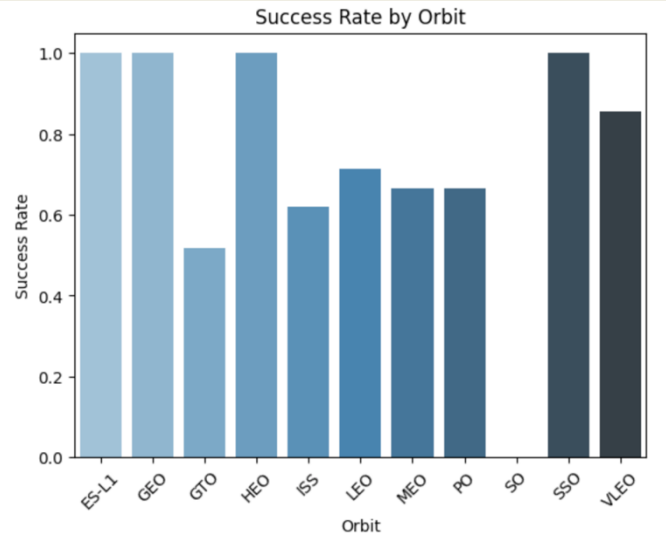


A line chart was used for this graph as it helps visualise the change in a variable over time.

Bar Charts

Bar charts were used in the context of the following graph:

- Success Rate and Orbit Type



Bar charts allow for the comparison between a numerical and categorical value

EDA with SQL

Display

1. Display the names of the unique launch sites
2. Display 5 records where the launch site name begins with ‘CCA’
3. Display the total payload mass carried by boosters from NASA (CRS)
4. Display the average payload mass carried by booster version F9 v1.1

List

1. List the date when the first successful landing on a ground pad was achieved
2. List the names of the booster versions which had a successful land on a drone ship and had a payload mass between 4000 and 6000 Kg
3. List the total number of successful and failed missions
4. List the names of the booster versions which have carried the maximum payload mass
5. List the failed landing outcomes on drone ships, their booster versions and the launch site name in the year 2015
6. Rank the landing outcomes (failure (drone ship) or success (ground pad) between the dates of 2010-06-04 and 2017-03-20 in descending order

Build an Interactive Map with Folium

1. **Mark all the launch sites**
    - *Use Folium to initialise the map*
    - *Add a circle and marker to each site on the map using **folium.Circle** and **folium.Marker** functions respectively*
2. **Mark the successful and failed launches for each site**
    - *Many launches occurred at the same site meaning they have the same co-ordinates, as such we need to cluster them together on the map*
    - *Assign colours to the markers to indicate success or failure before clustering*
        - Failed (class 0) = red
        - Successful (class 1) = green
    - *To cluster the launch outcomes, add a **folium.Marker** to the **MarkerCluster()** object*
    - *Create an icon (green or red) using **icon_color** as the **marker_color***
3. **Distances between the launch site to proximities**
    - *To find the distances from the launch site, use the Lat and Long co-ordinates to calculate these values*
    - *Create a **folium.Marker** object to show these distances once calculated*
    - *Draw a **folium.Polyline** on the map to show the distances between two points (launch site and proximities)*

Build a Dashboard with Plotly Dash

Dropdown List with all the Launch Sites

- Allows for the user of the dashboard to select all launch sites or a specifically selected launch site

Pie Chart Showing Successful Launches

- Allows for the user to see the successful and unsuccessful launches

Slider of Payload Mass Range

- Allows for the user to select the payload mass range they want displayed

Scatterplot Showing Payload Mass (Kg) vs Success Rate by Booster Version

- Allows for the user to see the correlation between Payload Mass (Kg) and the Launch Success

Predictive Analysis (Classification)

Model Development

- Load the dataset
- Standardise and pre-process the data
- Split the data into training and test sets using the train_test_split() function
- Decide on which algorithm is the most appropriate to use
- Create a GridSearchCV object and a dictionary of parameters to be used on each algorithm
- Fit object into the parameters and then use the training set to train the model

Model Evaluation

- Use the output GridSearchCV object to:
  - *Chech the hyperparameters that have been tuned using the best_params function*
  - *Check the accuracy of the model using score and best_score*
- Plot and analyse the confusion matrices of each model

Choose the Best Classification Model

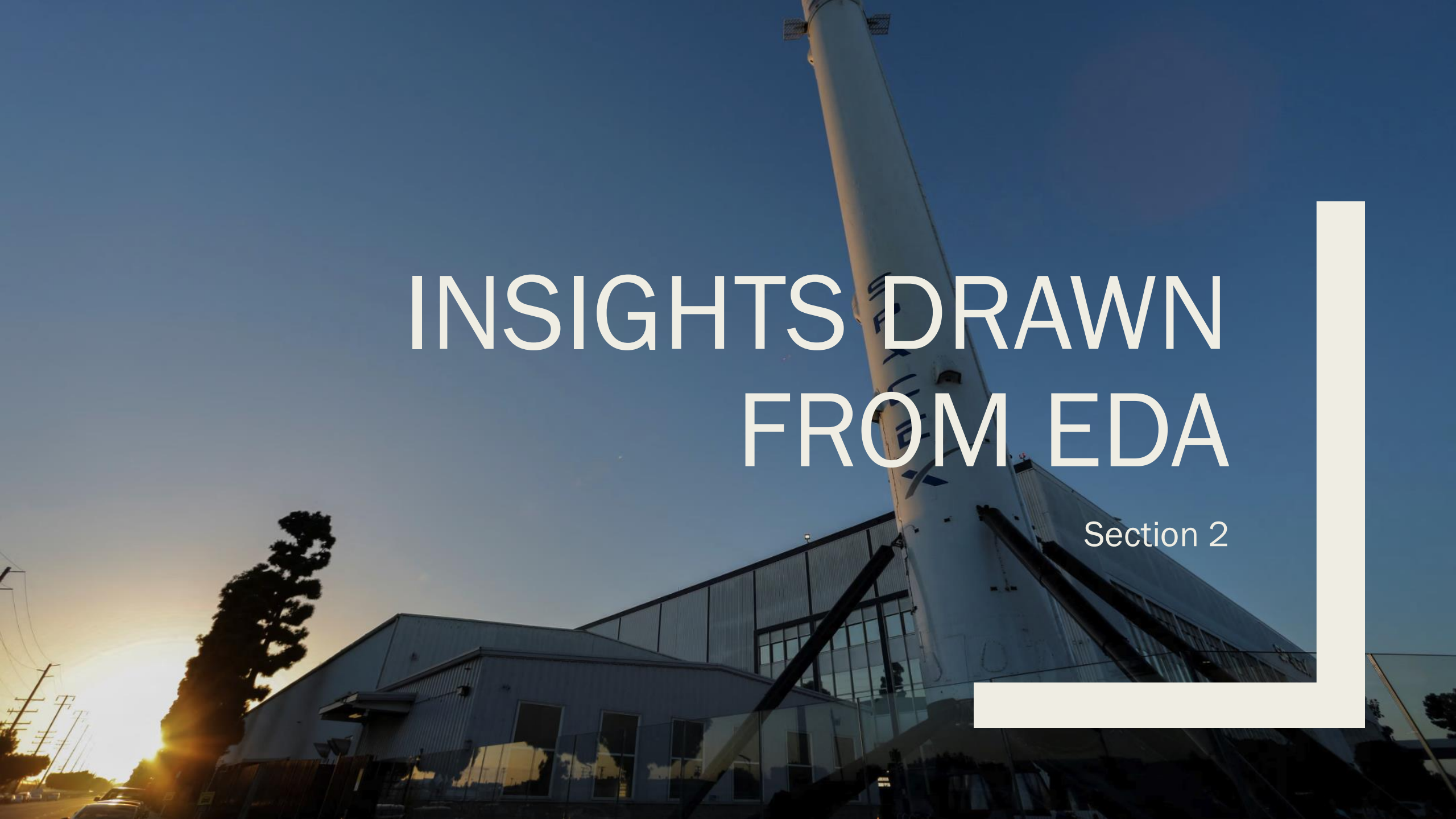
- Review each of the accuracy scores for all the classification models
- The model with the highest accuracy score is the best classification model to use

Results

Exploratory Data Analysis (EDA)

Interactive Analytics

Predictive Analysis (Classification)

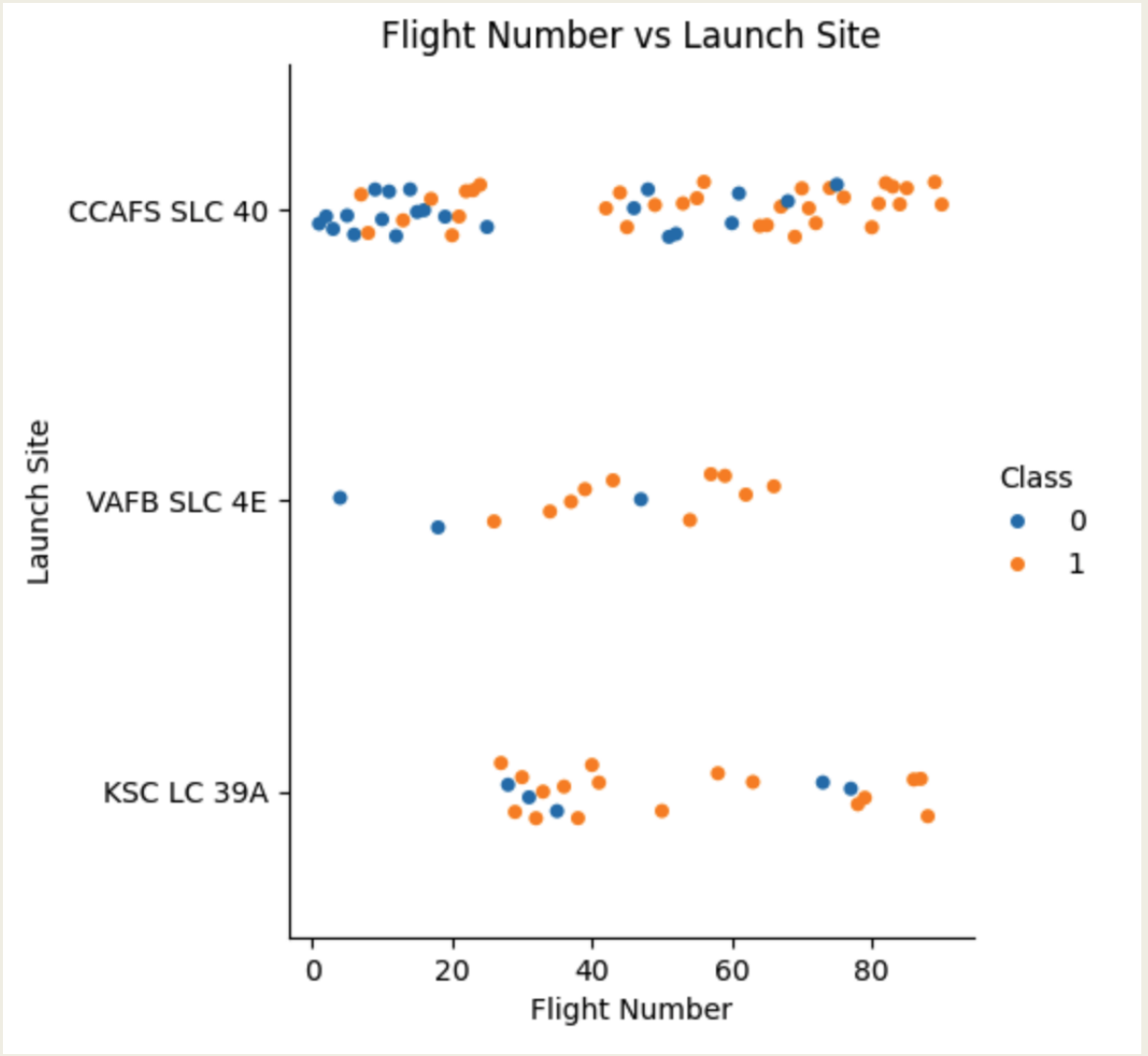


INSIGHTS DRAWN FROM EDA

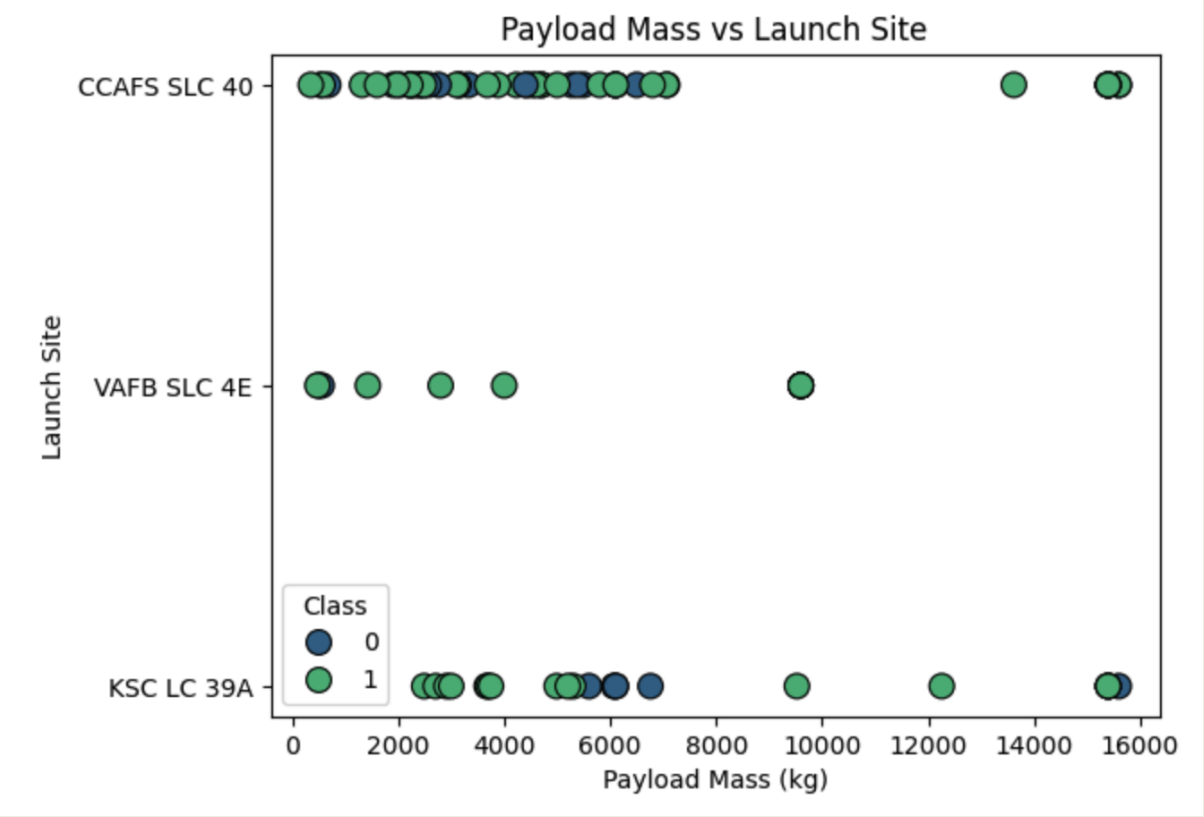
Section 2

Flight Number vs Launch Site

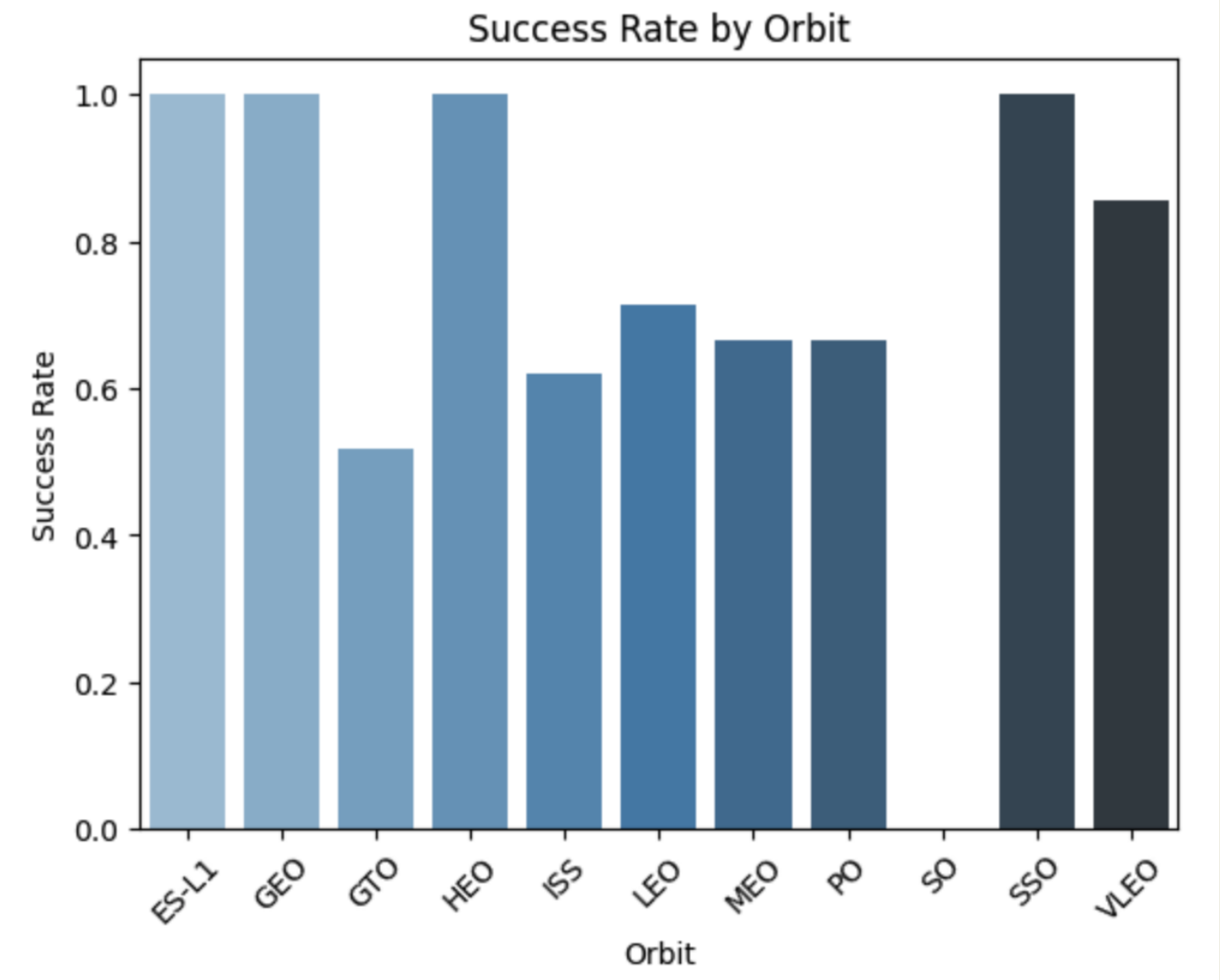
- Lower flights numbers have a lower success rate (blue)
- Higher flight numbers have a higher success rate (orange)
  - *Blue = fail*
  - *Orange = success*
- CCAFS SLC-40 has the most launches of all three launch sites
- Despite this, both other launch sites had higher success rates than CCAFS SLC-40
- Based on this graph we can conclude that more recent launches, and rockets with more launches have a higher rate of success



Payload vs Launch Site



- Overall, higher Payload Mass (Kg) results in higher success rates
- Most launches with a payload mass of over 7,000 Kg were successful
- Site CCAFS SLC-40 contained the most launches of the three sites
- Site VAFB SLC-4E has not launched anything with a mass greater than 10,000 Kg



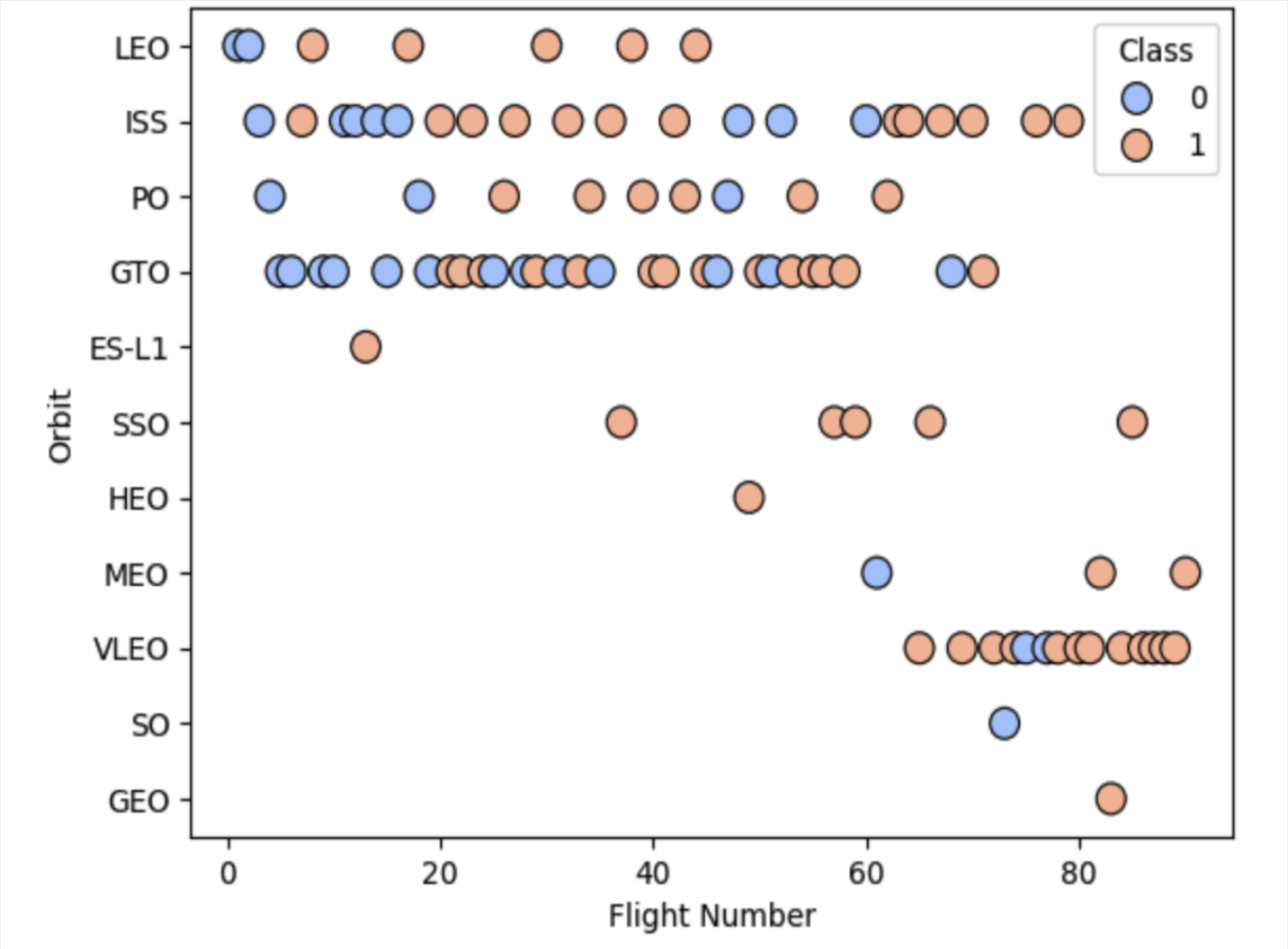
Success Rate vs Orbit Type

Explanation:

- Orbits with a 0% success rate:
  - *SO*
- Orbits with success rate between 50 and 80%:
  - *GTOM ISS, LLEO, MEO and PO*
- Orbits with a 100% success rate
  - *ES-L1, GEO, HEO and SSO*

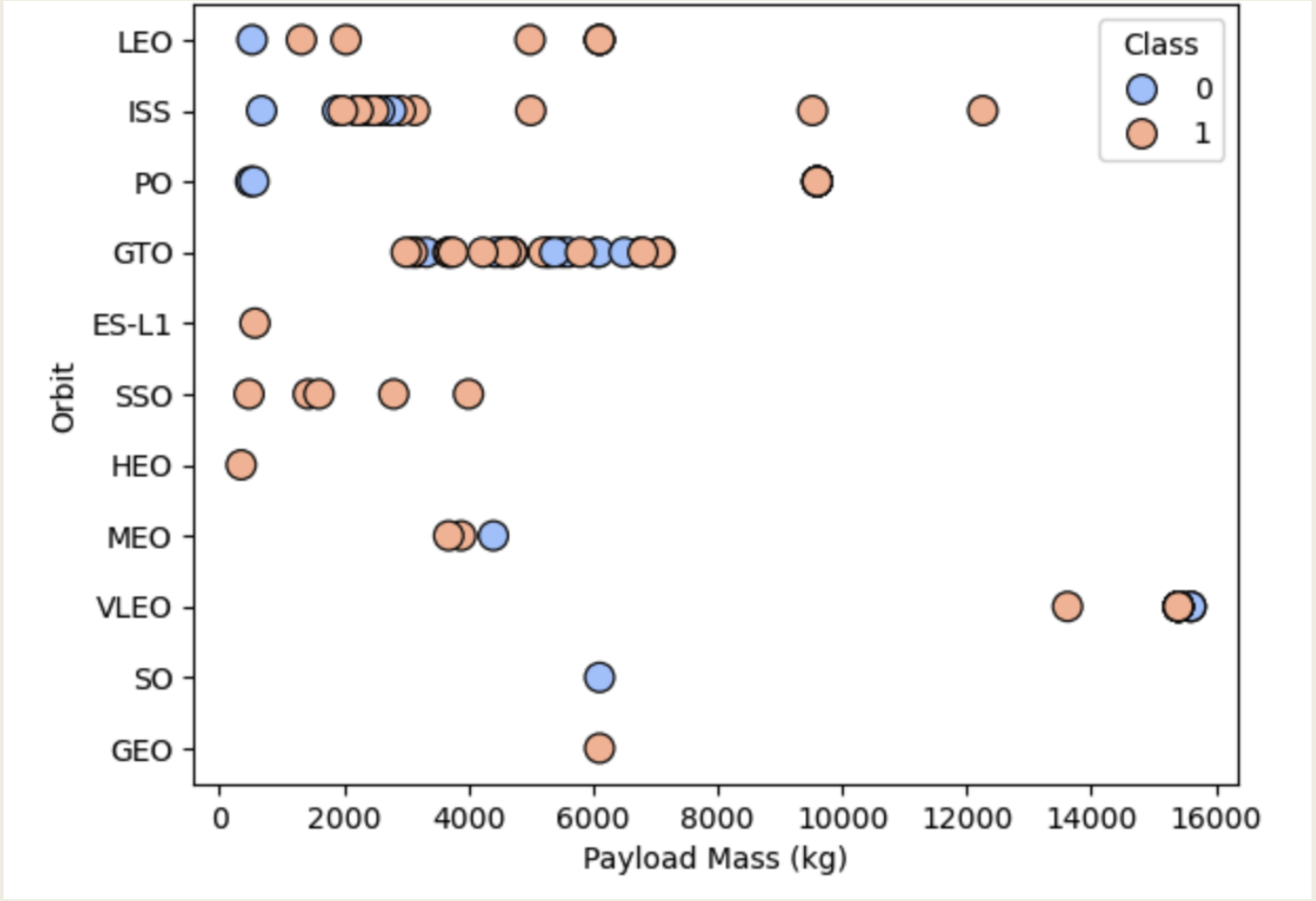
Flight Number vs Orbit Type

- Based on this graph, we can see that the success rate increases with the number of flights for each orbit
  - *The LEO orbit follows this trend closely*
  - *Whereas the GTO orbit does not appear to do so at all*



Payload vs Orbit Type

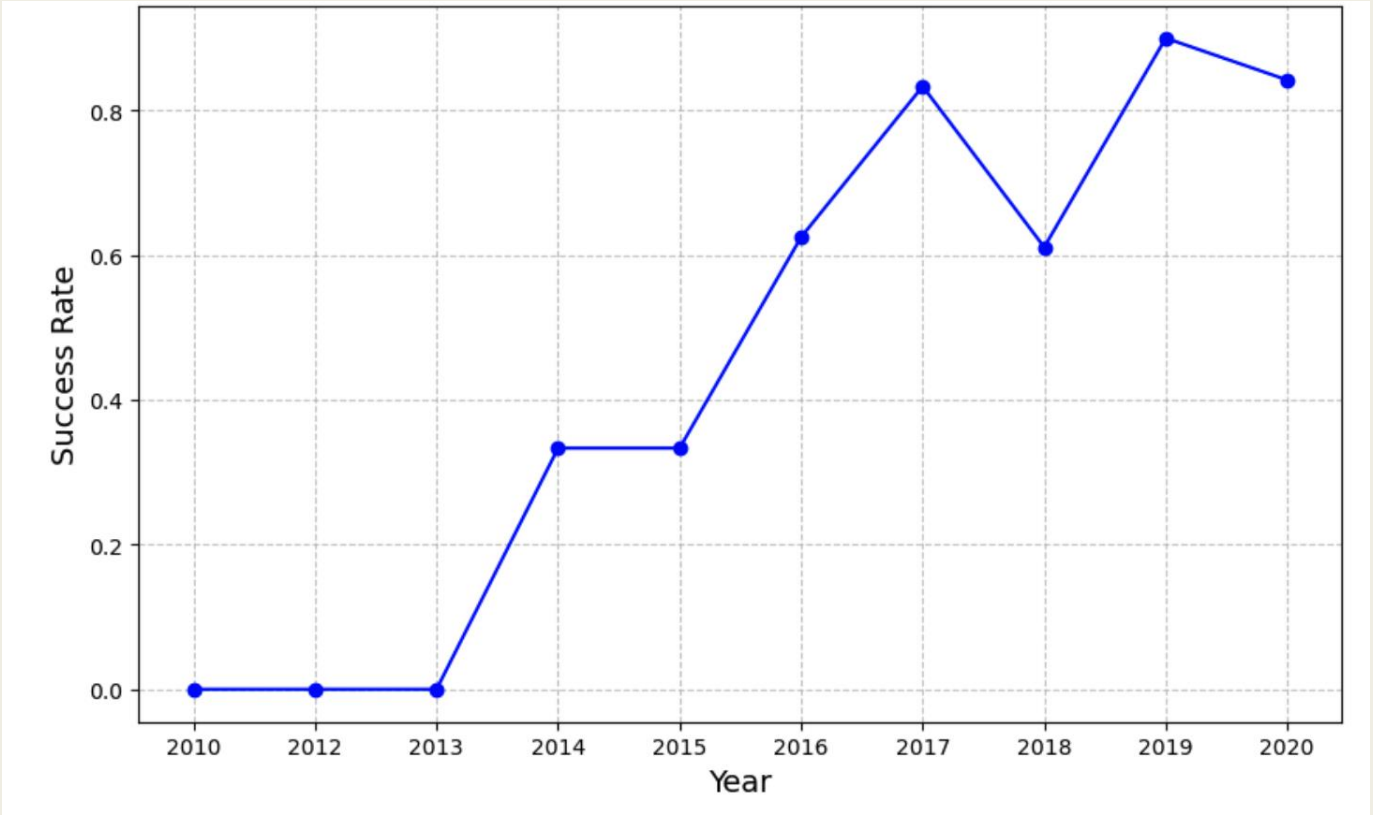
- For orbits LEO, ISS and PO it is clear that heavier payloads increased the success rate
- In contrast, the GTO orbit has mixed results with increased payload mass



Launch Success Yearly Trend

Based on this graph we can conclude:

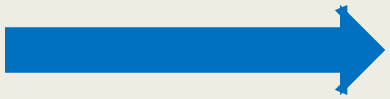
- Success rate improved continuously from
  - *2013 to 2017*
  - *2018 to 2019*
- Success rate decreased from
  - *2017 to 2018*
  - *2019 to 2020*
- Overall, it is clear that the success rate has increased significantly since 2013



All Launch Site Names

Found the names of all launch sites using the following code:

```
%sql select distinct launch_site from SPACEXTABLE
```



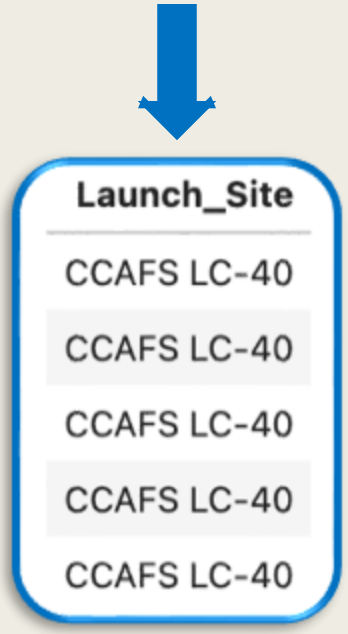
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Using the word UNIQUE to only return unique values from the **Launch_Site** column of the **SpaceXtable** dataset.

Launch Site Names Beginning with ‘CCA’

Using the following code to find 5 records where the launch site begins with ‘CCA’:

```
%sql SELECT LAUNCH_SITE FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```



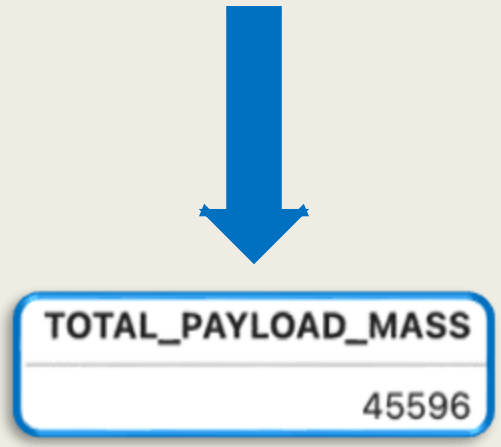
Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- The LIMIT 5 command ensure that only 5 records are returned.
- The LIKE command in combination with ‘CCA%’ returns only values beginning with ‘CCA’.

Total Payload Mass

Calculate the total payload mass carried by boosters from NASA using the following code:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTABLE \
    WHERE CUSTOMER = 'NASA (CRS)';
```



TOTAL_PAYLOAD_MASS
45596

The SUM function calculates the total of the **LAUNCH** column, and the **WHERE CUSTOMER = 'NASA (CRS)'** filters the results to only display the carried boosters from NASA.

Average Payload Mass by F9 v1.1

Calculated the average payload mass carried by booster version F9 v1.1 using the following code:

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEXTABLE WHERE BOOSTER_VERSION = 'F9 v1.1';
```

AVERAGE_PAYLOAD_MASS
2928.4

Using the AVG keyword to calculate the average of the **PAYLOAD_MASS__KG** column, then the WHERE keyword to filter the results to only display the F9 v1.1 booster version

First Successful Ground Landing Date

Find the date of the first successful ground landing using the following code:

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

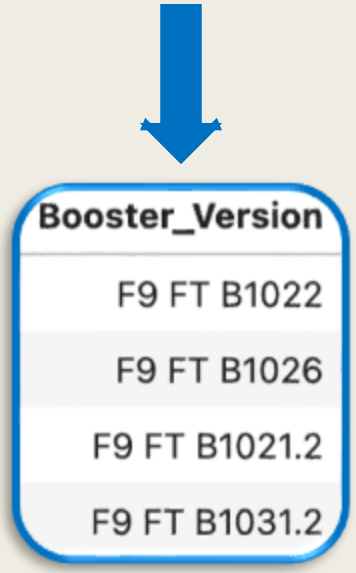
FIRST_SUCCESSFUL_GROUND_LANDING
2015-12-22

- Using the MIN keyword to calculate the minimum date in the DATE column, which indicates the first landing date
- The WHERE keyword is used to filter the results to only show successful ground landing missions.
- The combination of these two functions will return the earliest date of a successful ground landing

Successful Drone Ship Landing with Payload between 4000 and 6000

Listed the names of the boosters which have successfully landed on drone ship and had payload mass between 4000 and 6000 Kg using the following code:

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTABLE 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

- The WHERE function filters the results to only include those that satisfy both listed conditions (successful drone landing and payload between 4000 and 6000 Kg).
- The BETWEEN keyword allows us to ensure that the payload mass for successful drone landings is only between 4000 and 6000 Kg.

Total Number of Successful and Failure Mission Outcomes

Calculated the total number of successful and failed mission outcomes using the following code:

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTABLE GROUP BY MISSION_OUTCOME;
```

Mission_Outcome	TOTAL_NUMBER
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

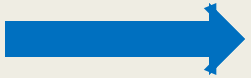
- COUNT was used to calculate the total number of mission outcomes
- GROUPBY keyword was used to group the results by mission outcome (success or failure)

Boosters Carried Maximum Payload

List the names of the boosters which have carried the maximum payload mass.

- The SELECT statement within the brackets searches for the maximum payload, the value of which is used in the WHERE function.
- The DISTINCT function is then used to retrieve only the unique booster version.

```
%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```



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

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

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTABLE \
    WHERE (LANDING_OUTCOME = 'Failure (drone ship)') AND DATE LIKE '2015%';
```

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- WHERE keyword was used to filter the results to show only the failed landing outcomes
- AND keyword shows results from the year 2015 specifically

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

```
%sql SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTABLE \
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
    GROUP BY LANDING_OUTCOME \
    ORDER BY TOTAL_NUMBER DESC;
```

Landing_Outcome	TOTAL_NUMBE
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)	

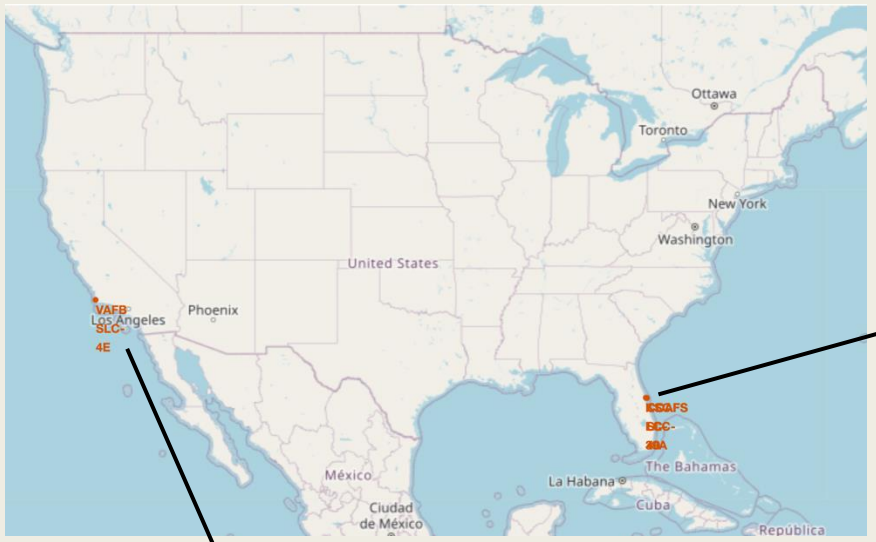
Rank the count of the landing outcomes (such as failure (drone ship) or success (ground pad)) between the dates of 2010-06-04 and 2017-03-20 in descending order

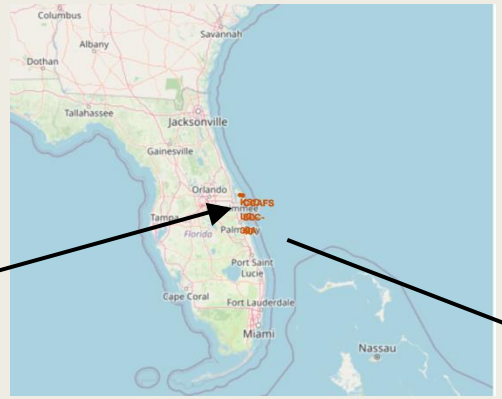
- The WHERE function in combination with the BETWEEN function is used to filter the results to only show dates within the set parameters.
- Then GROUPBY and ORDERBY were used to group and order the results with DESC displaying them in descending order.

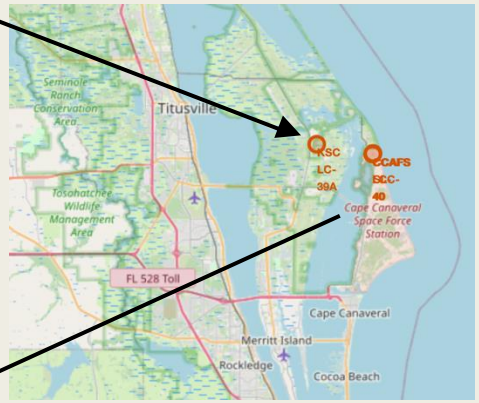
LAUNCH SITES PROXIMITIES ANALYSIS

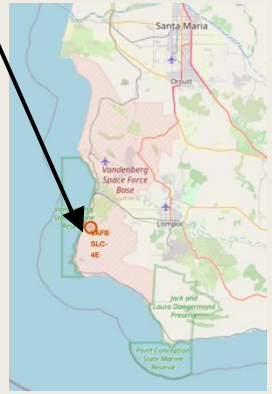
Section 3

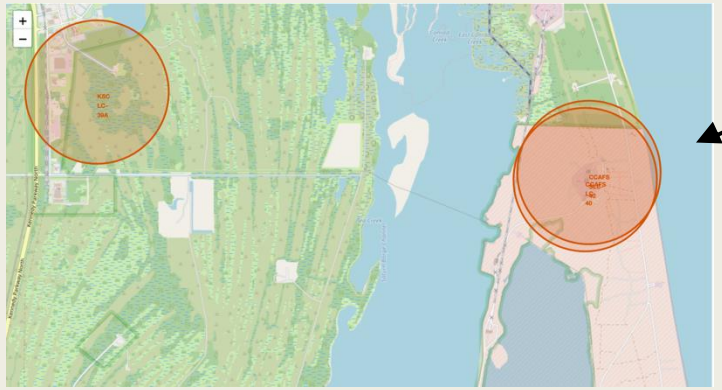
All SpaceX Launch Sites on a Map





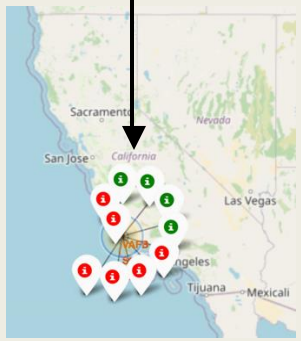




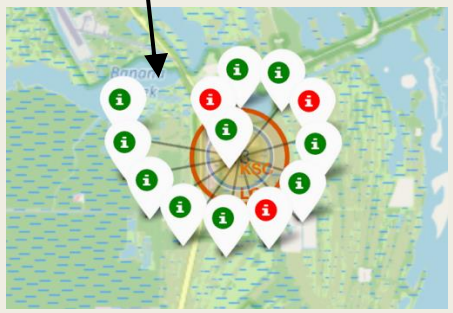


All SpaceX launch sites are located in the United States of America, specifically on the coasts. Launch sites on the East Coast are found in Florida, and the launch site on the West Coast is located in California.

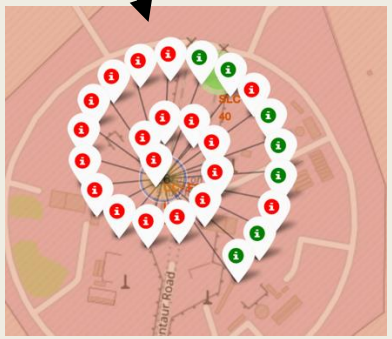
Launch Sites Successes and Failures



VAFB SLC-4E



KSC LC-39A



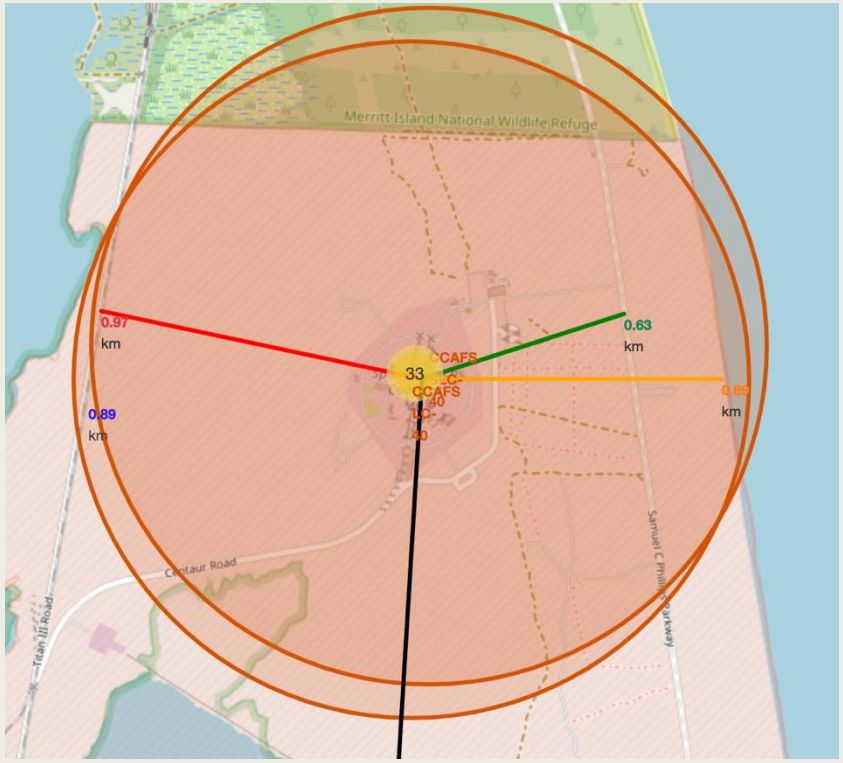
CCAFS LC-40

CCAFS SLC-40

All launches are grouped together into clusters, with green icons representing successful launches and red icons representing failed launches.

Launch Site Proximity to Points of Interest

Using site CCAFS SLC-40 as the example site to understand launch site proximity to points of interest, we can answer the following questions:



Are launch sites in close proximity to railways?

- Yes. The nearest railway is 0.97 Km west of the launch site.

Are launch sites in close proximity to highways?

- Yes. The nearest highway is 0.63 Km east of the launch site.

Are launch sites in close proximity to the coast?

- Yes. The nearest coastline is 0.89 Km east of the launch site

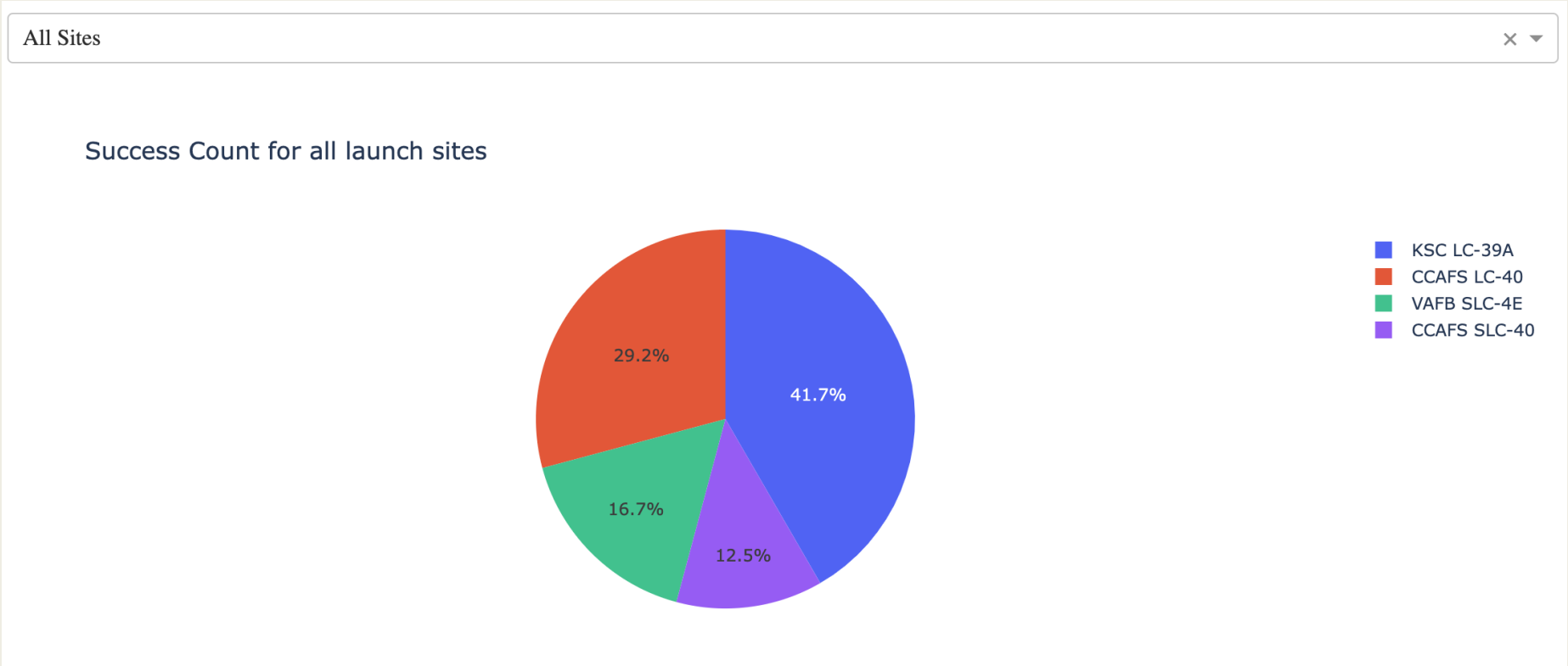
Do launch sites keep certain distance away from cities?

- Yes. The nearest major city is 53.94 Km away

BUILD A DASHBOARD WITH PLOTLY DASH

Section 4

Pie Chart of Successful Launch Counts for All Sites

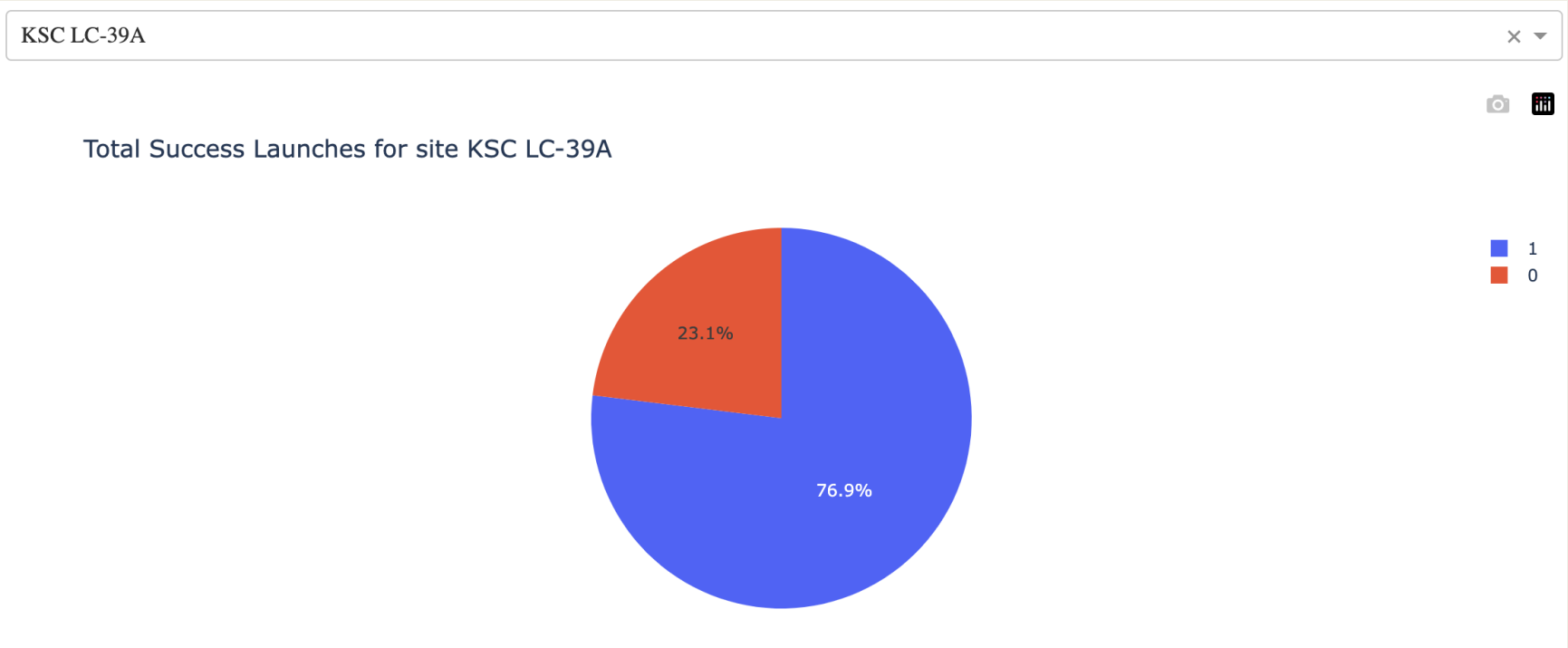


Here we can see that the launch site KSC LC-39A had the most successful launches of all four sites, having 41.7% of launches being successful

Pie Chart of the Launch Site with the Highest Success Ratio

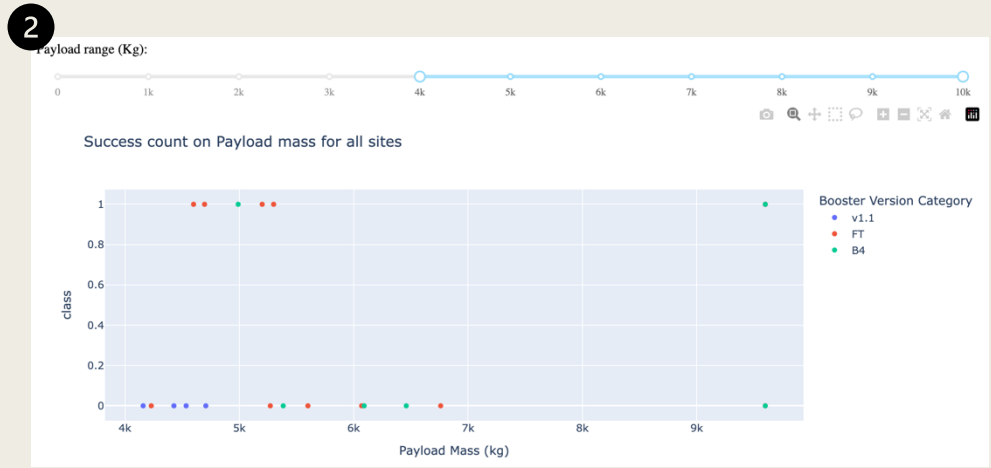
KSC LC-39A had the most successful launches of all four sites, as well as the highest success ratio.

The success ratio for KSC LC-39A was 76.9%



0 = Failure
1 = Success

Launch Outcome vs Payload Mass Scatter Plots for All Sites



- There is a clear gap around the 4,000 Kg mark, and as such I've split the data into lower and higher range Payload Mass.
  - Low Payloads = 0 to 4,000 Kg
  - High Payloads = 4,000 to 10,000 Kg
- Based on this split, we can see that the success rate for high payloads is lower than for the low payloads.
- Additionally, this graph highlights that the v1.1 and B5 booster types have not yet been launched with high payloads.

PREDICTIVE ANALYSIS (CLASSIFICATION)

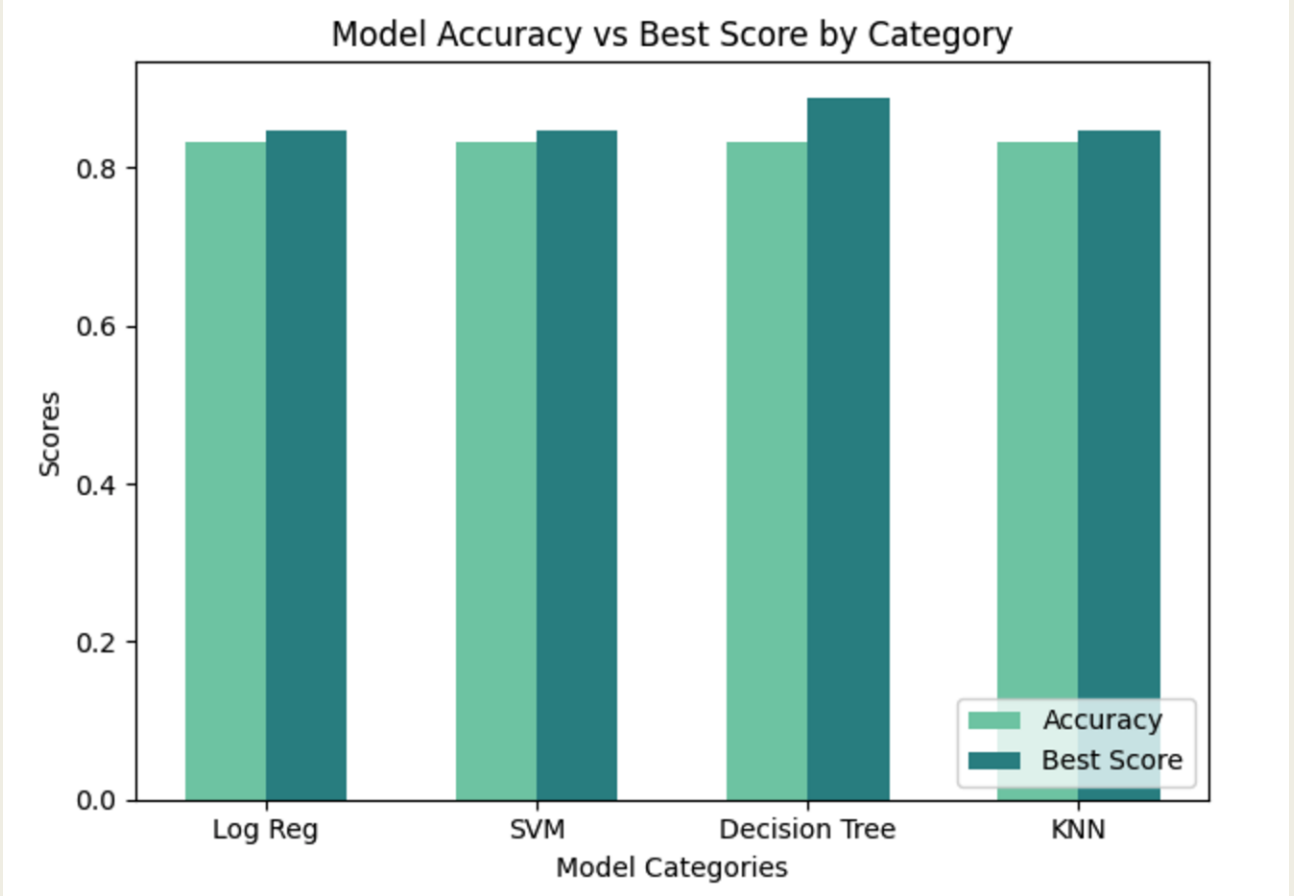
Section 5

Classification Accuracy

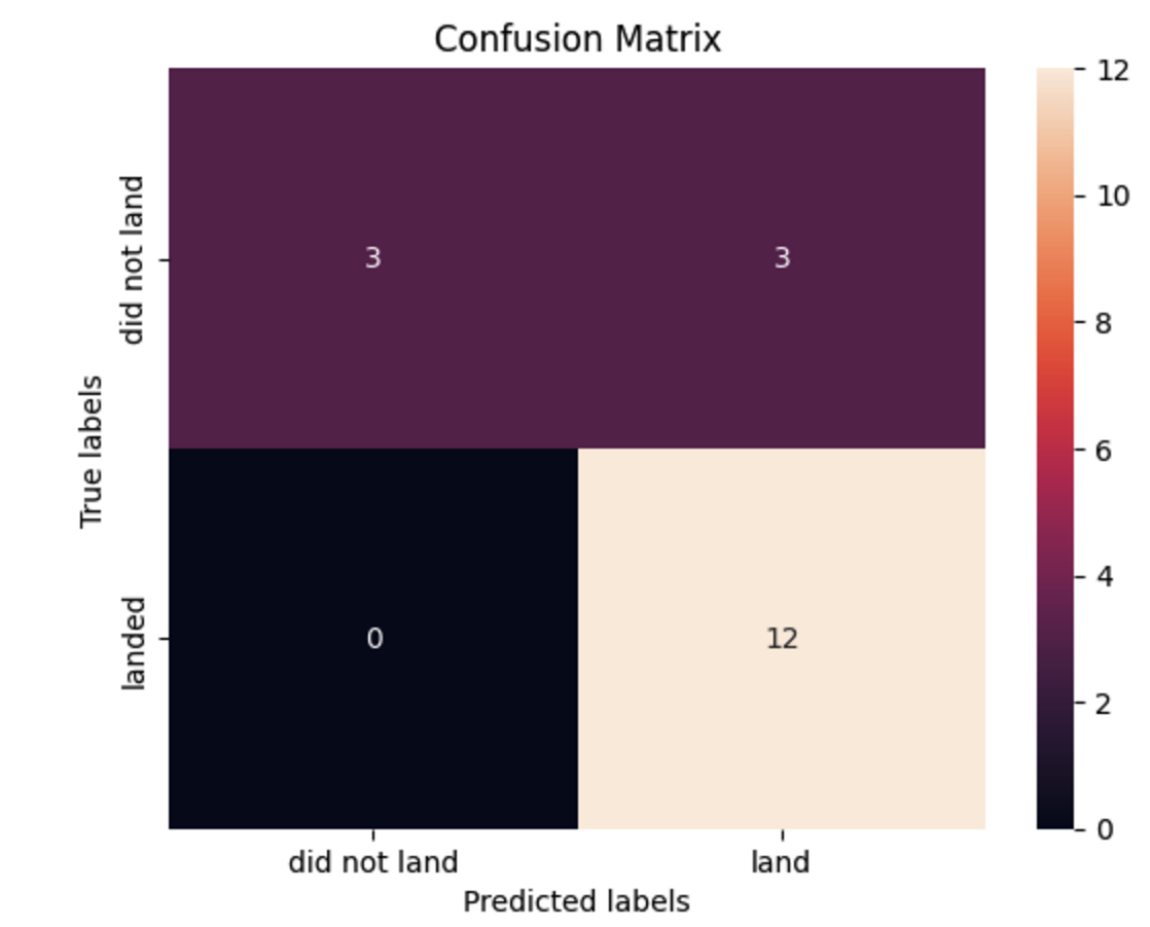
Plotted the model accuracy vs the best score for the model in a bar chart for each of the four model types. The following results can be interpreted from the graph and table output:

- The Decision tree has the highest classification accuracy of all the chosen models
  - Accuracy score of 83.33%
  - Best score is 88.9%

Algorithm	Accuracy	Best Score
Log Reg	0.833333	0.846429
SVM	0.833333	0.848214
Decision Tree	0.833333	0.889286
KNN	0.833333	0.848214



Confusion Matrix



- As shown in the previous slide, the classification model which performed the best was the Decision Tree, which had an accuracy of 88.9%.
- This is explained by the following confusion matrix, which shows only 3 out of 18 results classified incorrectly (3 false positives shown in the top right corner).
- The other 14 results are classified correctly (3 did not land, and 12 landed).

CONCLUSIONS

Conclusions

- The Decision Tree Model is the best algorithm to use for this dataset, with an accuracy of 88.9%.
- The success for higher payloads (>4000 Kg) is lower than for the low payloads.
- Launch site KSC LS-39A has the highest success rate for all launch sites
  - *With a success rate of 76.9%*
  - *Total percentage of launches successful being 41.7%*
- Orbits ES-L1, GEO, HEO and SS0 have a success rate of 100%
  - *This high success rate of ES-L1, GEO and HEO can be explained by the fact that they only have 1 flight into each of these orbits*
  - *Orbit types PO, ISS and LEO have lower success with low payloads and might benefit from higher payloads*
- Increased number of flights results in an increase in success rate at a launch site
  - *Most early flights were unsuccessful, meaning that more flight experience results in higher success rate*
  - *Between 2010 and 2013, all landings were unsuccessful*
  - *After 2016, the landing success rate increased to above 50%*
  - *This means that the success rate of launches has increased over the years*


THANK YOU!