IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Leonor Duarte 20/10/2024

https://github.com/leonorduarte/LeonorDuarte-IBM-data-sciense-certificate

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

In this project, a rival company to SpaceX (i.e., SpaceY) uses SpaceX Falcon 9 rocket data to determine the rocket first stage landing successes and the cost of a launch. A summary for the methodologies and results described in this report is outlined below.

Summary of methodologies

- Data was collected from the SpaceX public API and publically available data on Wikipedia. Data wrangling included extracting launch outcome information to serve as the dependent variable in the Machine Learning models.
- SQL queries and data visualizations (static plots, interactive maps, and an interactive dashboard) were created to discover insights about the data set and answer questions.
- Predictive analysis was pursued using Logistic Regression, SVM (Support Vector Machine), Decision Tree, and KNN (k-Nearest Neighbors) Machine Learning models.

Summary of all results

- Launch data include info about flight number, date of launch, payload mass, orbit type, launch site, mission outcome and other variables.
- Logistic Regression, SVM (Support Vector Machine), and KNN (k-Nearest Neighbors) all perform equally well for Machine Learning models on this dataset.

Introduction

- In competition with SpaceX, a rival rocket launch company wants to make predictions about the success/failure of SpaceX Falcon 9 rocket first stage landings.
- What is the nature and extent of the data that we have on SpaceX Falcon 9 first stage landings?
- Which machine learning model would work best (have the highest accuracy) to predict the outcome of a Falcon 9 first stage landing from a future launch?
- Will a future Falcon 9 first stage landing be successful?

Intro - Background

- This capstone project is part of the IBM Data Science Professional Certificate. The goal of the project is to demonstrate proficiency in data science and machine learning techniques using a real-world data and to summarize the results in a report.
- In this project, a rival company to SpaceX (i.e., SpaceY) uses SpaceX Falcon 9 rocket data to determine the rocket first stage landing successes and also uses the rocket data to determine the cost of a launch. Space Y uses the data to bid against SpaceX for a rocket launch. SpaceX advertises Falcon 9 rocket launch cost to be 62 million dollars. Whereas for other companies the cost of a rocket launch is more than 165 million dollars.
- Throughout the project Python Jupyter notebooks are used to perform the data collection and analysis. These Jupyter notebooks and the final *.pdf report are saved in my GitHub repository webpage.
- The major parts of this report include data collection methodology, data wrangling, exploratory data analysis (EDA), interactive data visualization, machine learning (ML) classification model development, and model evaluation. Finally, the accuracy of different ML algorithms are compared in predicting the future landing of the Falcon 9 first stage rocket.

Section 1

Methodology

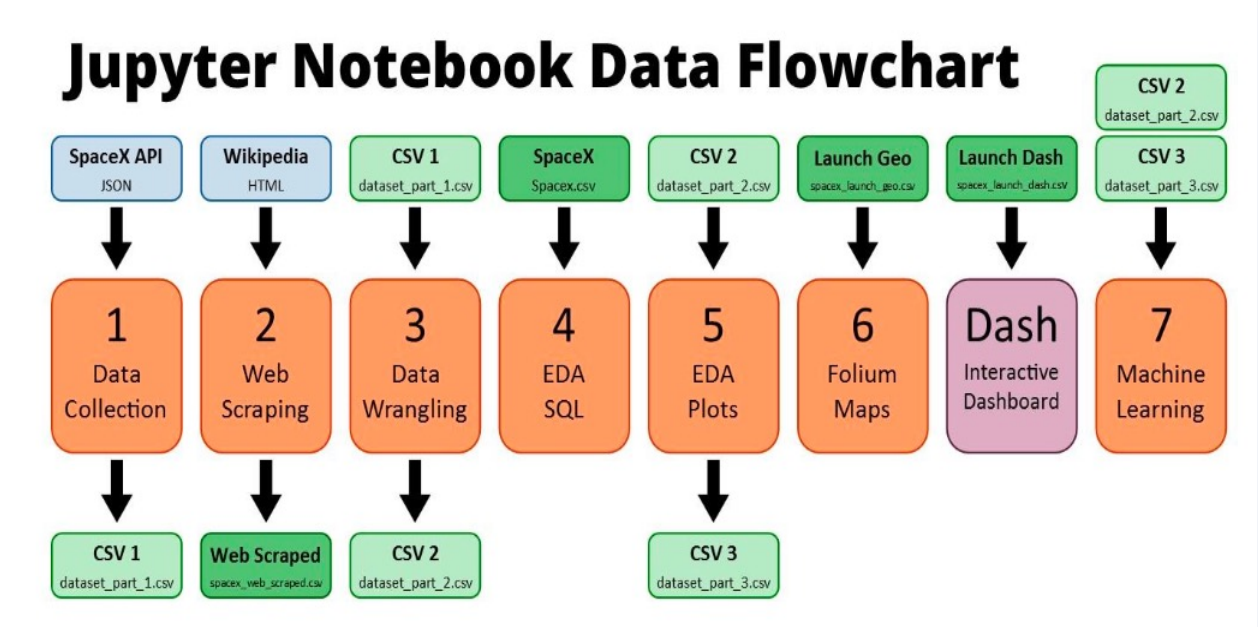
Methodology

Executive Summary

- SpaceX API and Wikipedia launch table data was collected.
- Data was cleaned in preparation for visualizations, queries and machine learning model creation.
- Exploratory data analysis (EDA) was done using visualization and SQL.
- Interactive visual analytics were created using Folium and Plotly Dash.
- Predictive analysis using classification models was done.

Data Collection – SpaceX API

- The data sets were collected from:
- An IBM copy of a call to the publically accessible SpaceX API with launch data in JSON format.
- A permanently linked Wikipedia page with launch data in HTML
- Further data sets were provided. See darker green .csv files in top row of diagram below.
- GitHub URL: https://github.com/leonorduarte/LeonorDuarte-IBM-data-sciense-certificate/blob/main/1Capstone_Leonor%20Duarte_Data_Collection.ipynb

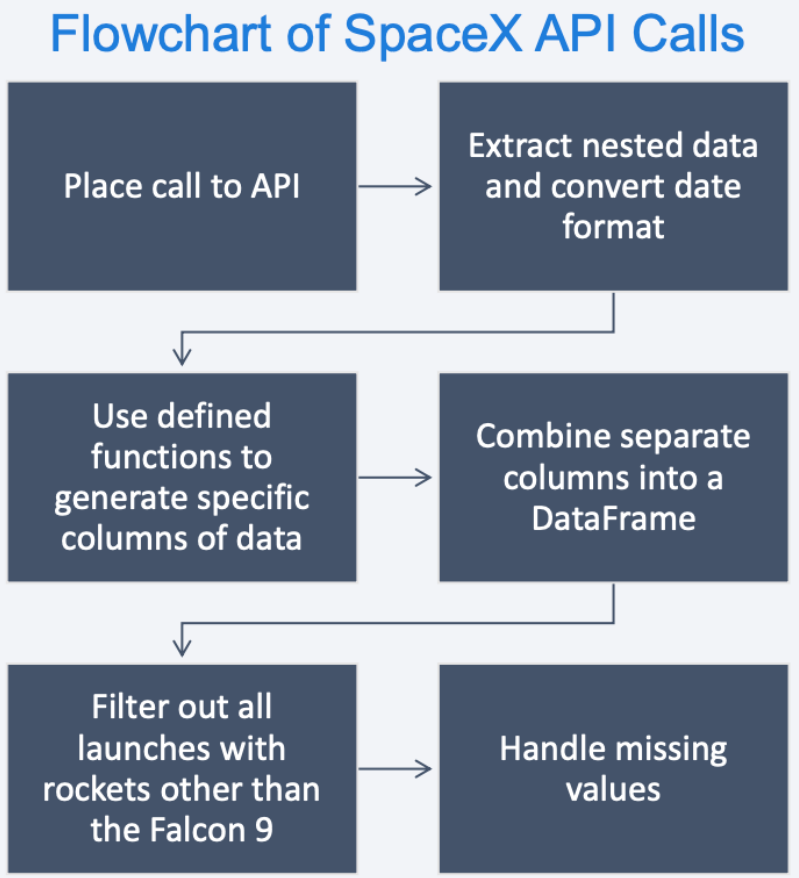


Data Collection – SpaceX API

- The SpaceX API has data available publically.
- Once a GET request has been made to the SpaceX API and the response received, the data can be placed into a Pandas
- DataFrame for further analysis.

GitHub URL (Data Collection):

- https://github.com/leonorduarte/LeonorDuarte-IBM-data-sciense-certificate/blob/main/2_Capstone_LeonorDuarte_Webscraping%20(1).ipynb

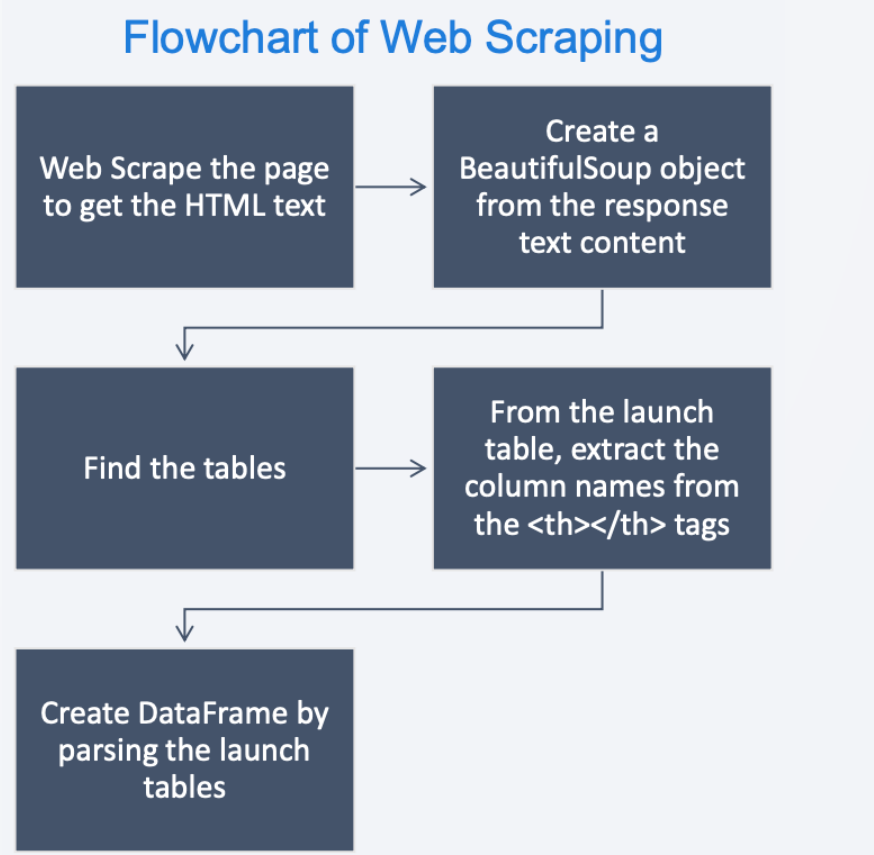


Data Collection – Scraping

- Wikipedia has a page that has tables of data about SpaceX launches.
- These tables can be scraped to extract launch data that can be put into a Pandas
- DataFrame for further analysis.

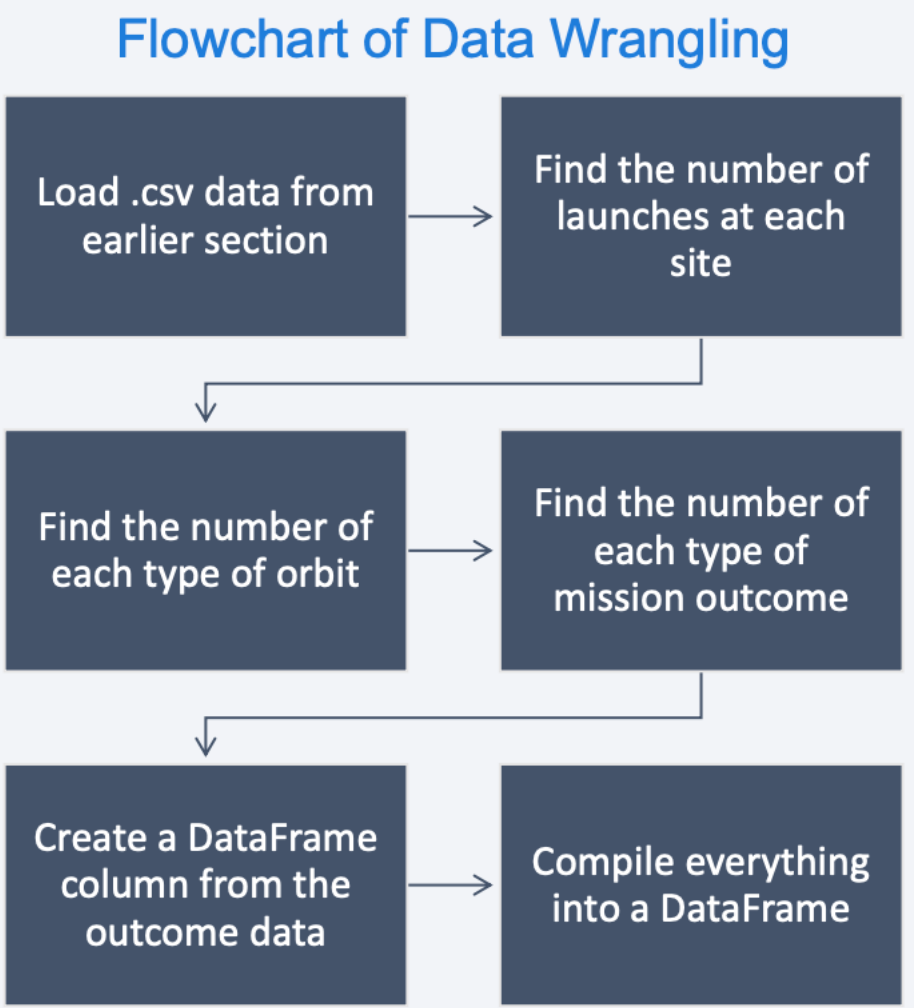
GitHub URL (Data Collection):

- https://github.com/leonorduarte/LeonorDuarte-IBM-data-sciense-certificate/blob/main/2_Capstone_LeonorDuarte_Webscraping%20(1).ipynb



Data Wrangling

- The .csv file from the first section contains the data that needed to be cleaned.
- The launch sites, orbit types and mission outcomes were cleaned up.
- The handful of mission outcome types were converted to a binary classification where 1 means that the Falcon 9 first stage landing was a success and 0 means that it was a failure.
- The new classification was added to the DataFrame for further analysis
- GitHub: https://github.com/leonorduarte/LeonorDuarte-IBM-data-sciense-certificate/blob/main/3-Capstone_LeonorDuarte_Data_Wrangling%20(2).ipynb



EDA with Data Visualization

The following charts were created to look at Launch Site trends.

- Scatterplot to see **mission outcome** relationship split by **Launch Site** and **Flight Number**.
- Scatterplot to see **mission outcome** relationship split by **Launch Site** and **Payload**.

The following charts were created to look at Orbit Type trends

- Bar chart to see **mission outcome** relationship with **Orbit Type**.
- Scatterplot to see **mission outcome** relationship split by **Orbit Type** and **Flight Number**.
- Scatterplot to see **mission outcome** relationship split by **Orbit Type** and **Payload**.

The following chart was created to look at trends based on time

- Line plot to see **mission outcome** trend by **year**.
- GitHub URL: https://github.com/leonorduarte/LeonorDuarte-IBM-data-sciense-certificate/blob/main/4_Capstone_LeonorDuarte_EDA_SQL%20(2).ipynb

EDA with SQL

- Primary goal: With all the launch data, what sort of numerical analysis can we do when we start a deep dive and what useful numbers can we quickly pull without that detailed look?
- Queries just to get used to the data (e.g., unique launch sites, names of boosters in a payload range)
- Useful facts (e.g., number of successful and failed flights, landing outcomes between 2010 and 2017)
- GitHub URL:https://github.com/leonorduarte/LeonorDuarte-IBM-data-sciense-certificate/blob/main/5_Capstone_LeonorDuarte_EDA_Data_Visualization%20(1).ipynb

EDA with SQL

- Primary goal: With all the launch data, what sort of numerical analysis can we do when we start a deep dive and what useful numbers can we quickly pull without that detailed look?
- Queries were written to extract information about:
- Launch sites
- Payload masses
- Dates
- Booster types
- Mission outcomes

GitHub URL: https://github.com/leonorduarte/LeonorDuarte-IBM-data-sciense-certificate/blob/main/4_Capstone_LeonorDuarte_EDA_SQL%20(2).ipynb

Build an Interactive Map with Folium

- Primary goal: “what about the successful missions can be gleaned from the geography of the launch sites?”
- Focused on marking sites of successful and failed launches, where the flight sites are, and their distances from notable landmarks Where should SpaceY keep its launch sites?
- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Markers were added for launch sites and for the NASA Johnson Space Center
- Circles were added for the launch sites.

Lines were added to show the distance to the nearby features:

- Distance from CCAFS LC-40 to the coastline
- Distance from CCAFS LC-40 to the rail line
- Distance from CCAFS LC-40 to the perimeter road

GitHub URL: https://github.com/leonorduarte/LeonorDuarte-IBM-data-sciense-certificate/blob/main/6_LeonorDuarte_Launch_Site_Location%20(1).ipynb

Build a Dashboard with Plotly Dash

- Primary goal: “What feature values map to the most successes?”
- The input dropdown is used to select one or all launch sites for the pie chart and scatterplot.

The pie chart displays one of two things:

- For All Sites – the distribution of successful Falcon 9 first stage landings between the sites
- For One Site – the distribution of successful and failed Falcon 9 first stage landings for that site
- The input slider is used to filter the payload masses for the scatterplot.
- The scatterplot displays the distribution of Falcon 9 first stage landings split by payload mass, mission outcome and by booster version category.

GitHub URL:https://github.com/leonorduarte/LeonorDuarte-IBM-data-sciense-certificate/blob/main/7-App

Predictive Analysis (Classification)

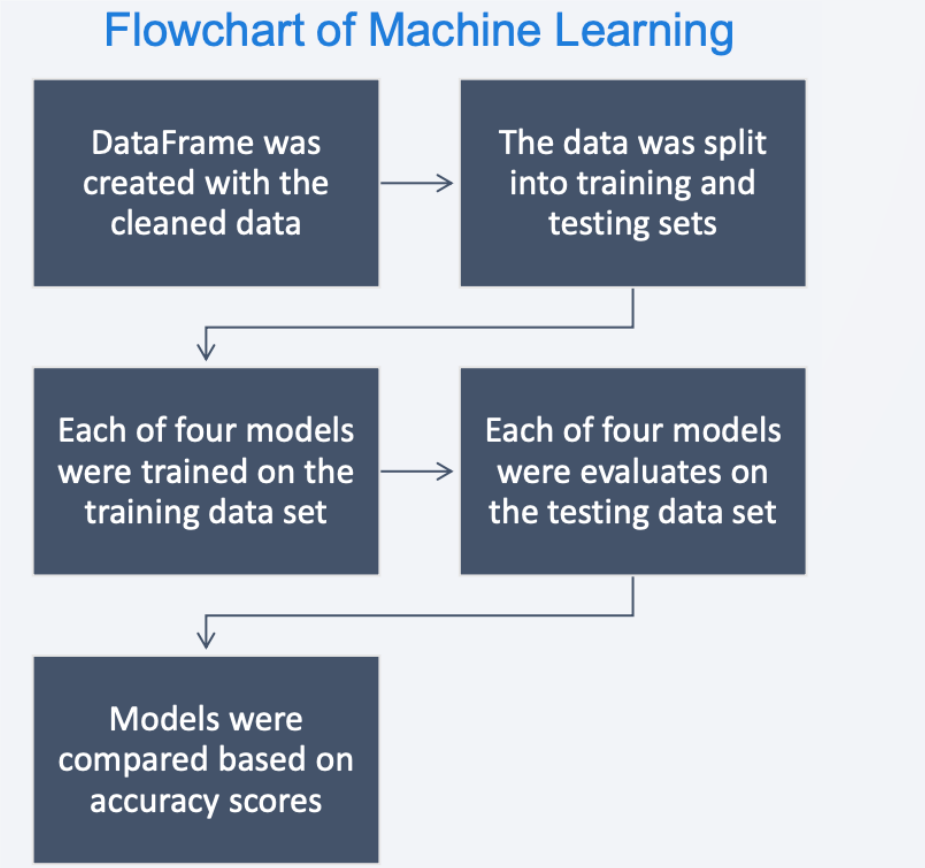
The dataset was split into training and testing sets.

Logistic Regression, SVM (Support Vector Machine), Decision Tree, and KNN (k-Nearest Neighbors) machine learning models were trained on the training data set.

Hyper-parameters were evaluated using GridSearchCV() and the best was selected using '.best_params_'.

Using the best hyper-parameters, each of the four models were scored on accuracy by using the testing data set.

GitHub URL:https://github.com/leonorduarte/LeonorDuarte-IBM-data-sciense-certificate/blob/main/8_Capstone_LeonorDuarte_Machine_Learning_Prediction%20(1)%20(1).ipynb



Results

EDA insights:

- SpaceX got excellent at successful launches with a jump in 2017
- No one feature is particularly predictive

From the interactive analytics:

- The CCAFS launch site, a payload range of 2000 – 3700 kg, and the FT boosters correlated most with a successful launch
- Landing sites stayed near roads and railways but away from cities

Predictive analytics results:

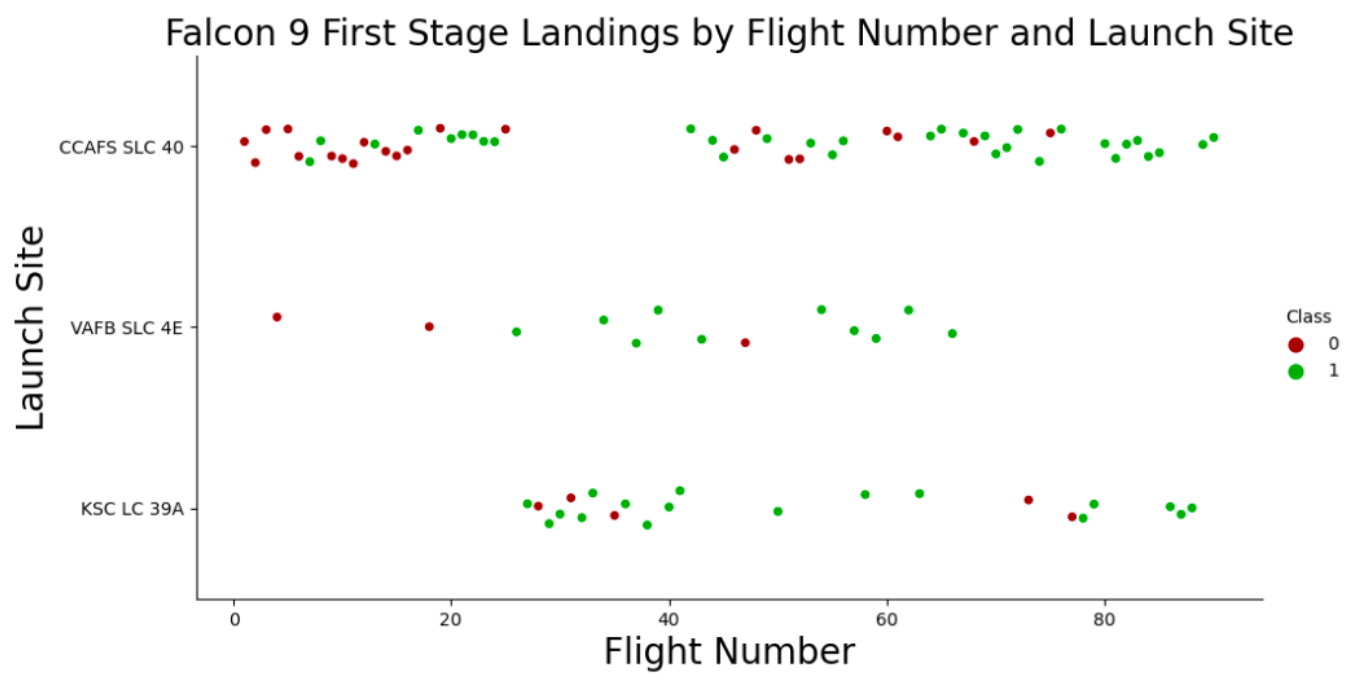
- No one model proved superior as all had a test accuracy of 83.3% though decision trees (DT) had the best training accuracy

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

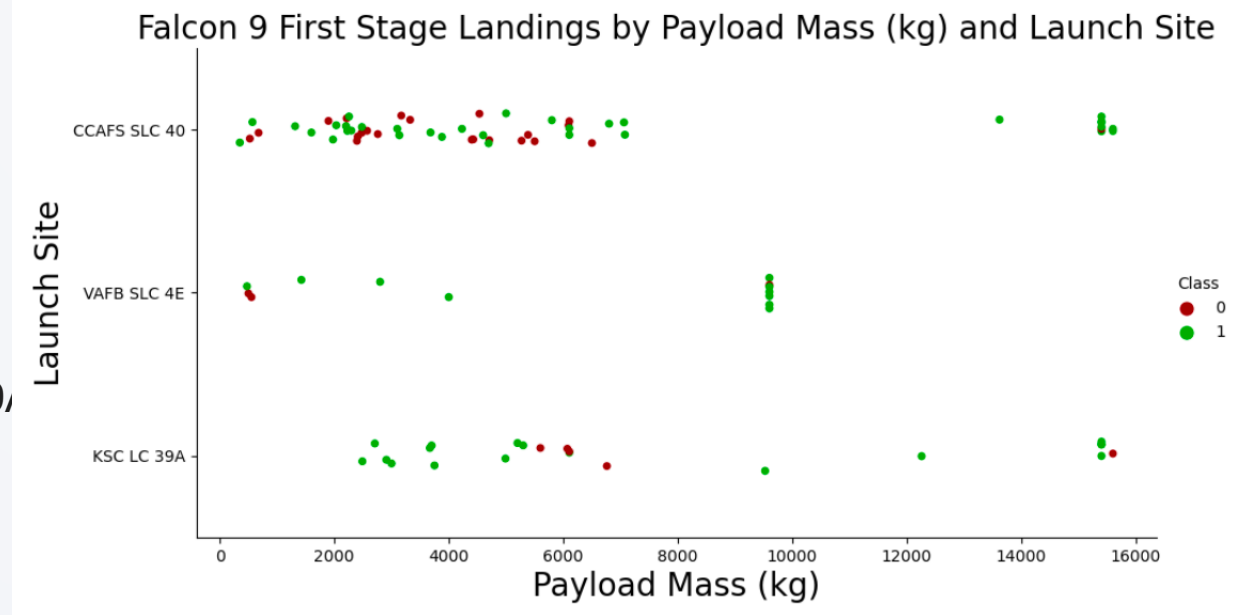
- Success rate varies noticeably with launch site.
- Successful Falcon 9 first stage landings appear to become more prevalent as the flight number increases.



- Falcon 9 first stage failed landings are indicated by the '0' Class (● red markers) and successful landings by the '1' Class (● green markers).

Payload vs. Launch Site

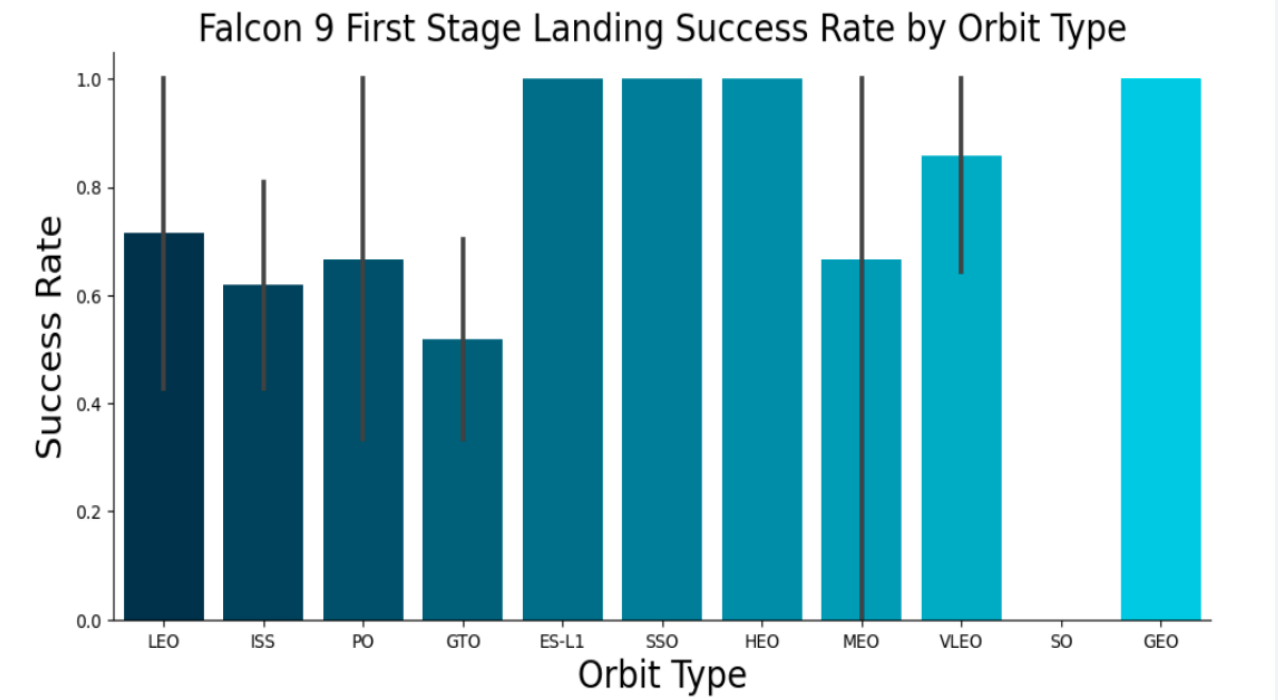
- For the CCAFS SLC 40 launch site, the payload mass and the landing outcome appear to not be strongly correlated.
- The failed landings at the KSC LC 39A launch site are all grouped around a narrow band of payload masses



- Falcon 9 first stage failed landings are indicated by the '0' Class (● red markers) and successful landings by the '1' Class (● green markers).

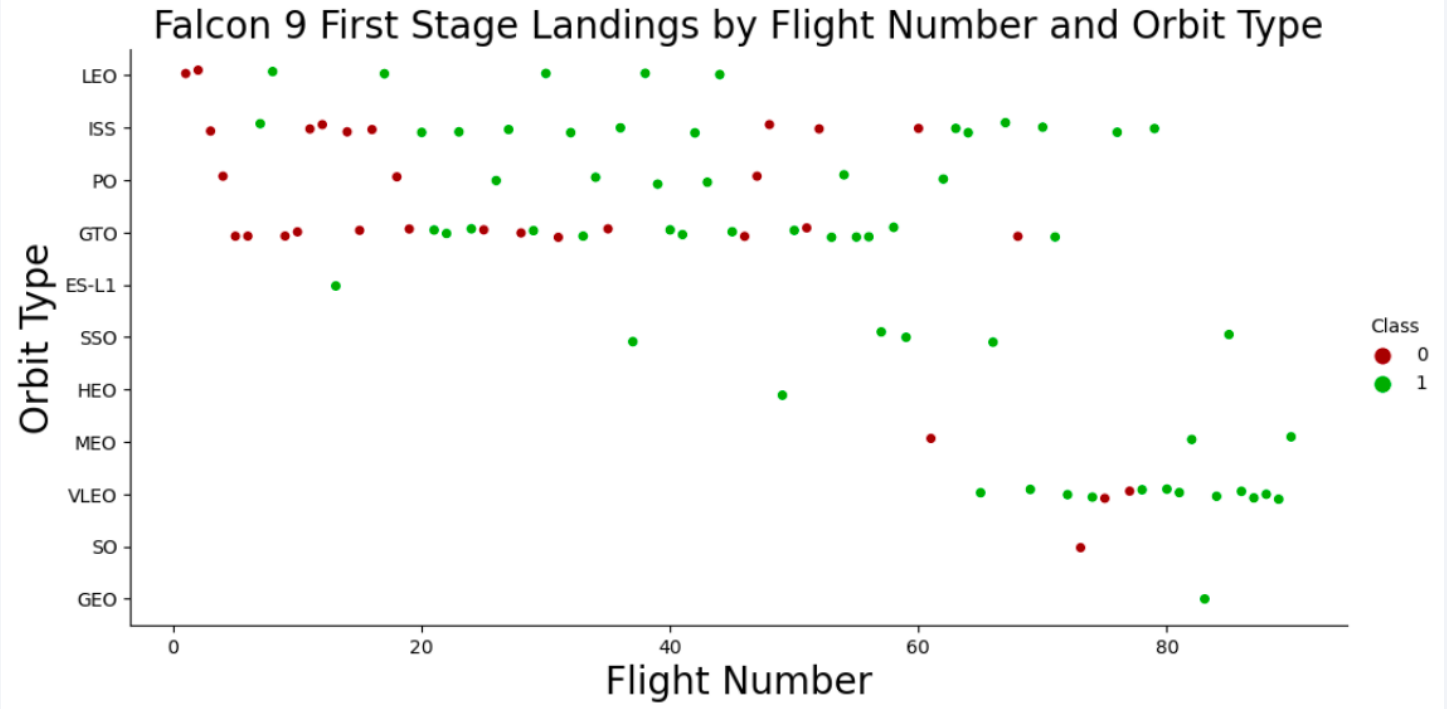
Success Rate vs. Orbit Type

- ES-L1, SSO, HEO and GEO orbits have no failed first stage landings.
- SO orbits have no successful first stage landings.



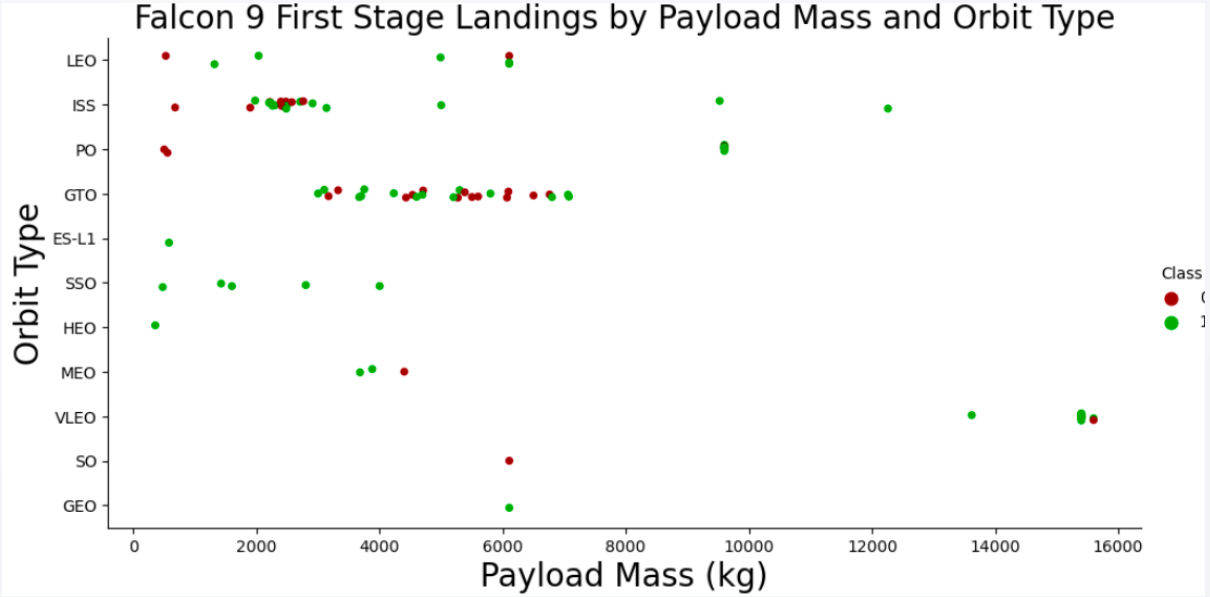
Flight Number vs. Orbit Type

There is a correlation between flight number and success rate with larger flight numbers being associated with higher success rates.



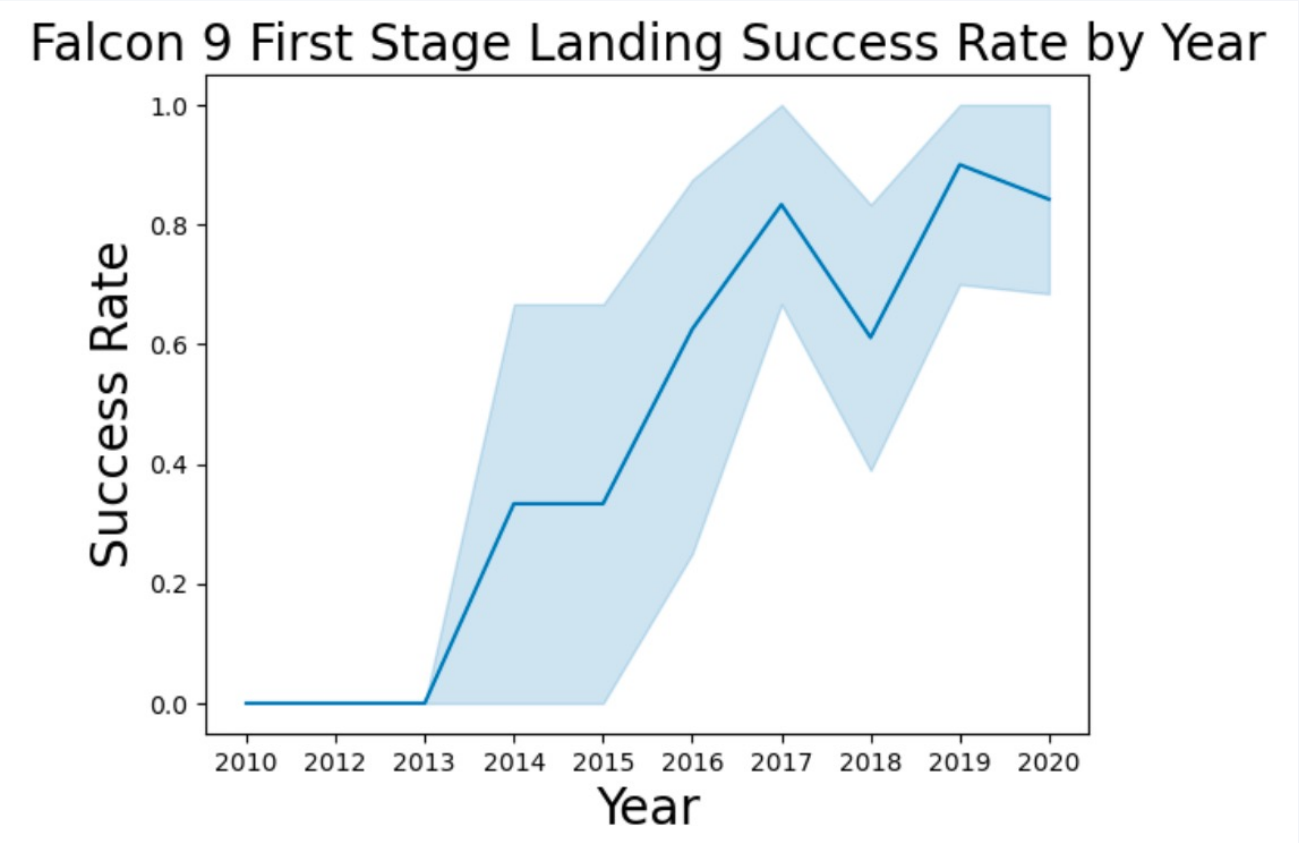
Payload vs. Orbit Type

- Some orbit types have better success rates than others.
- Success rate appears to have no obvious correlation with payload mass.



Launch Success Yearly Trend

- The success rate has increased significantly over the years.



All Launch Site Names

- **Question:** What are the names of the unique launch sites?
- **Query:** SELECT DISTINCT LAUNCH_SITE FROM SPACEXDATASET;
- **Result:**

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

Explanation: There are four unique launch sites.

Launch Site Names That Begin with 'CCA'

- **Task:** Find 5 records with launch sites that begin with `CCA`.
- **Query:** SELECT * FROM SPACEXDATASET WHERE launch_site LIKE 'CCA%' LIMIT 5;
- **Result:**

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation: This is a fairly straightforward sampling mechanism used to gain a sense of the data contained in the database table.

Total Payload Mass

- **Question:** What is the total payload carried by boosters from NASA?
- **Query:** SELECT sum(payload_mass__kg_) AS "Total Payload Mass (kg)" FROM SPACEXDATASET WHERE customer LIKE '%NASA (CRS)%';
- **Result:**

Total Payload Mass (kg)
48213

Explanation: The total payload carried by boosters from NASA is 48,213 kg.

Average Payload Mass by F9 v1.1

- **Question:** What is the average payload mass carried by booster version F9 v1.1?
- **Query:** SELECT sum(payload_mass__kg_) / count(payload_mass__kg_) AS "Average Payload Mass (kg)" FROM SPACEXDATASET WHERE booster_version LIKE 'F9 v1.1';
- **Result:**

Average Payload Mass (kg)
2928

Explanation: The average payload mass carried by booster version F9 v1.1 is 2,928 kg.

First Successful Ground Landing Date

- **Question:** On which date did the first successful landing outcome on ground pad occur?
- **Query:** SELECT min(DATE) AS "First Successful Landing Outcome Date" FROM SPACEXDATASET WHERE landing__outcomeLIKE'Success (ground pad)';
- **Result:**

First Successful Landing Outcome Date
2015-12-22

Explanation: The first successful landing outcome on ground pad occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Question:** What are the names of the boosters which have successfully landed on drone ship and had a payload mass greater than 4000 but less than 6000?
- **Query:** SELECT DISTINCT booster_version FROM SPACEXDATASET WHERE landing__outcome = 'Success (drone ship)' and payload_mass__kg_ BETWEEN 4000 and 6000;
- **Result:**

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

Explanation: The four booster versions that have successfully landed on drone ship with a payload mass greater than 4,000 kg but less than 6,000 kg are listed above

Total Number of Successful and Failure Mission Outcomes

- **Question:** What was the total number of successful and failed mission outcomes?
- **Query:** SELECT (SELECT count(*) FROM SPACEXDATASET WHERE lcase(landing__outcome) LIKE '%success%') AS "Success", count(*) AS "Failure" FROM SPACEXDATASET WHERE lcase(landing__outcome) NOT LIKE '%success%';
- **Result:**

Success	Failure
61	40

Explanation: There were 61 successful and 40 failed mission outcomes.

Boosters Carried Maximum Payload

- **Question:** What were the names of the boosters which have carried the maximum payload mass?
- **Query:** SELECT booster_version, payload_mass__kg_ FROM SPACEXDATASET WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXDATASET);
- **Result:**

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

Explanation: The maximum payload mass carried in this dataset is 15,600 kg. Twelve (12) separate Falcon 9 boosters carried this amount of payload mass.

2015 Launch Records

- **Task:** List the failed landing_outcomes in drone ship, their booster versions, and launch site names for records in year 2015.
- **Query:** SELECT MONTHNAME(DATE) AS “Month”, landing__outcome, booster_version, launch_site FROM SPACEXDATASET WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
- **Result:**

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

Explanation: There were two failed landing outcomes with a drone ship in 2015. Both launched from CCAFS LC-40. One occurred in January and the other in April.

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

- **Task:** Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad) between the date 2010-06-04 and 2017-03-20, in descending order.
- **Query:** SELECT landing__outcome, count(landing__outcome) AS "Count" FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing__outcome ORDER BY count(landing__outcome) DESC;
- **Result:**

landing__outcome	Count
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

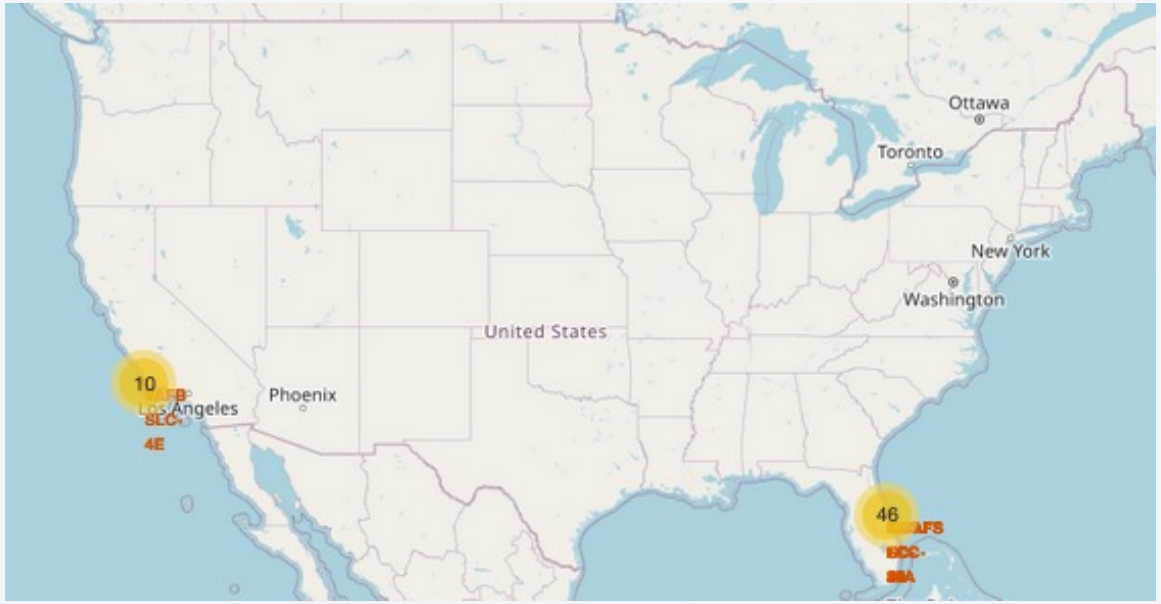
Explanation: The most common landing outcome was ‘not attempted’.

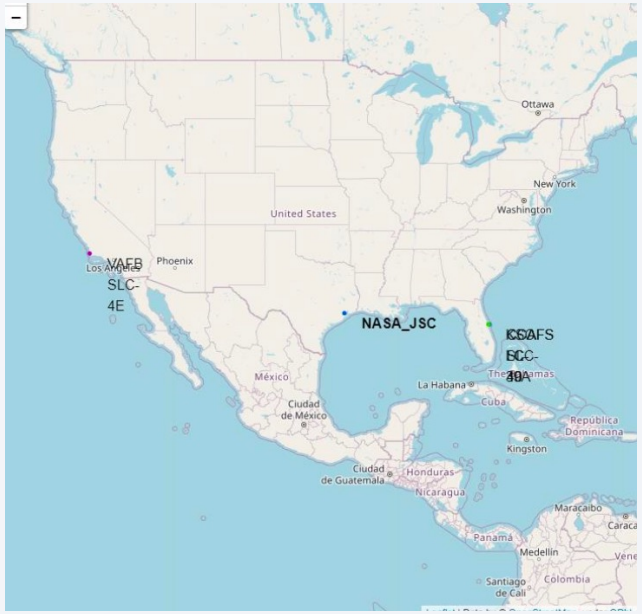
Section 3

Launch Sites Proximities Analysis

Falcon 9 Launch Site Locations

VAFB SLC-4E (California, USA)- Vandenberg Air Force Base Space Launch Complex 4E
KSC LC-39A (Florida, USA)- Kennedy Space Center Launch Complex 39A
CCAFS LC-40 (Florida, USA)- Cape Canaveral Air Force Station Launch Complex 40
CCAFS SLC-40 (Florida, USA)- Cape Canaveral Air Force Station Space Launch Complex 40

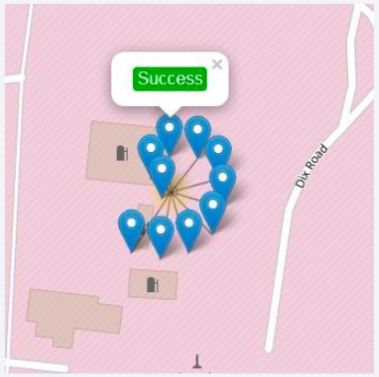




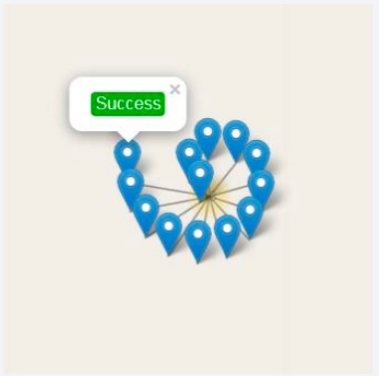
Map Markers of Success/Failed Landings

- The markers display the mission outcomes (Success/Failure) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site.

- A sense of a launch site's success rate for Falcon 9 first stage landings can be gleaned from the relative number of green success markers to red failure markers.

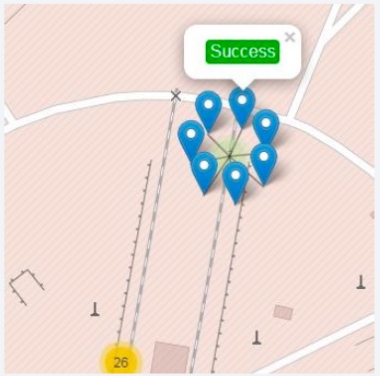


VAFB SLC-4E



KSC LC-39A

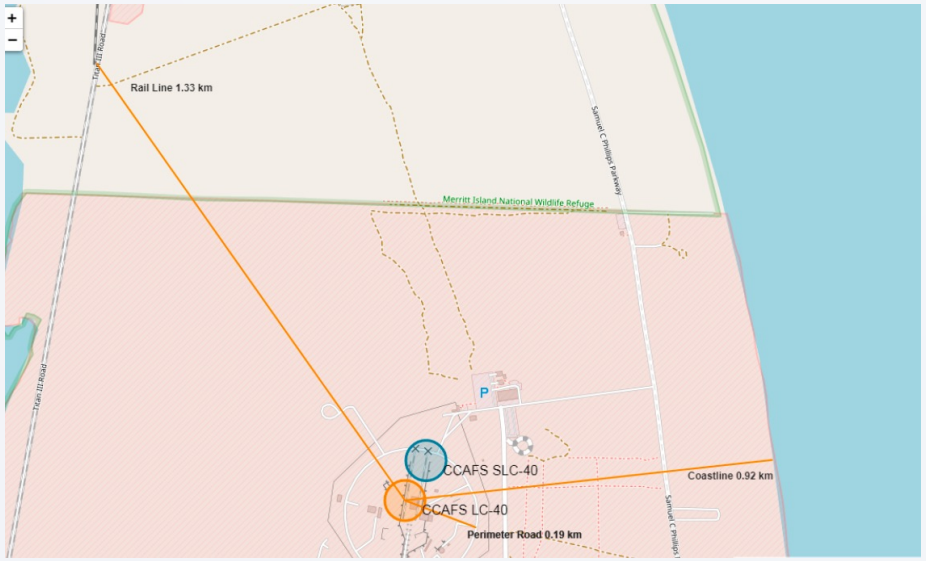
CCAFS LC-40

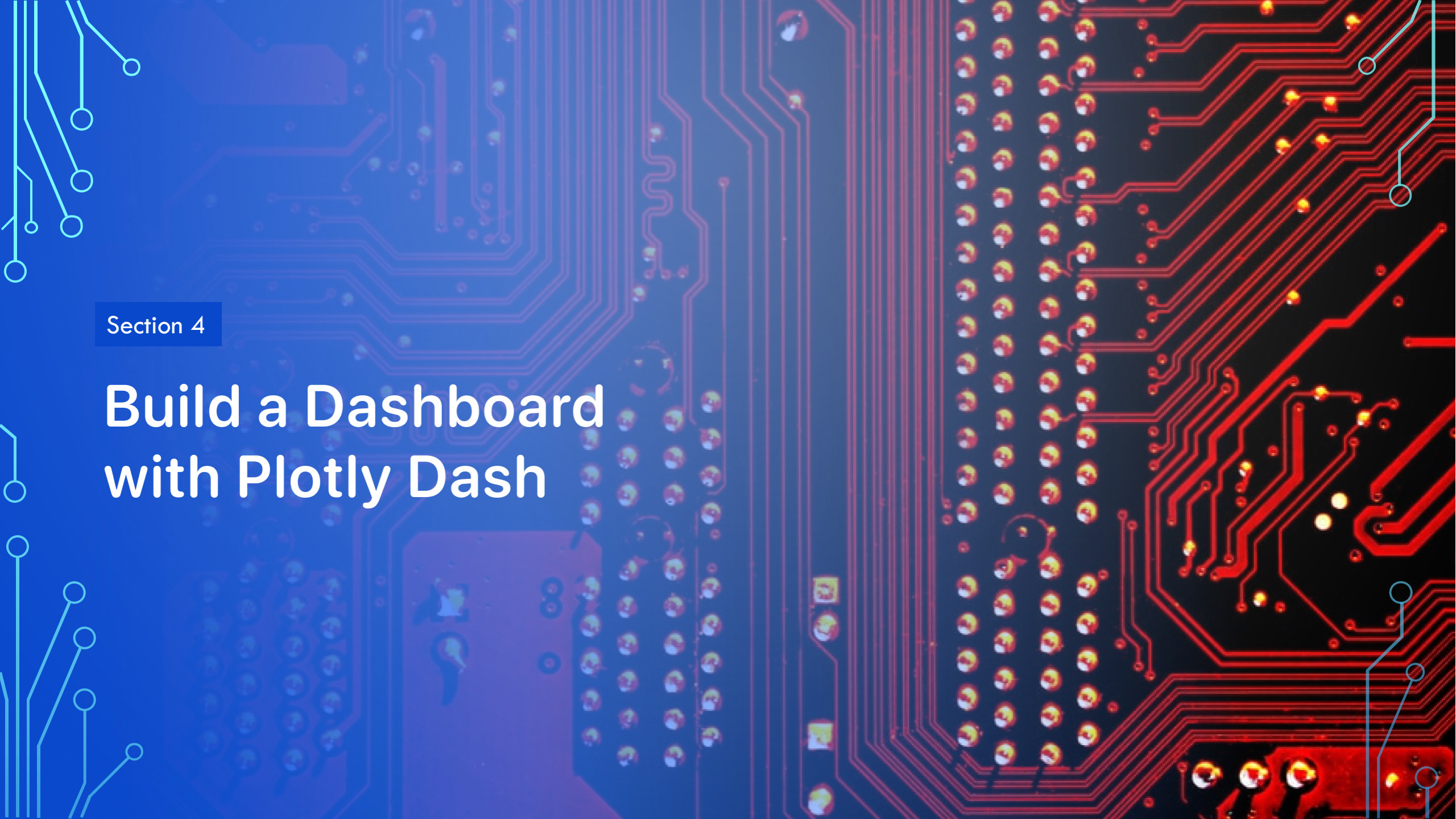


CCAFS SLC-40

Distance from Launch Site to Proximities

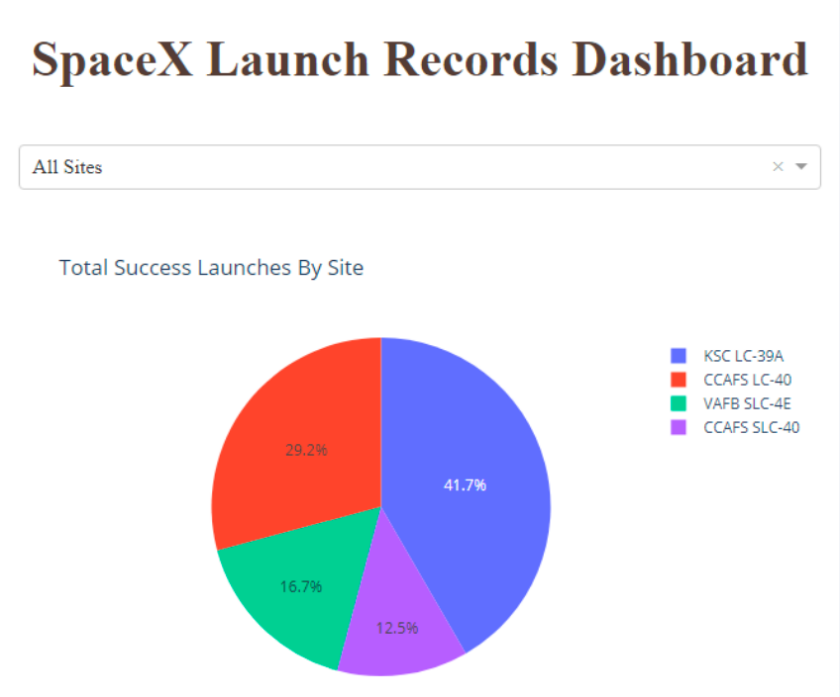
The CCAFS LC-40 and CCAFS SLC-40 launch sites have coordinates that are close to being, but are not exactly, right on top of each other.
The perimeter road around CCAFS LC-40 is 0.19 km away from the launch site coordinates.
The coastline is 0.92 km away from CCAFS LC-40.
The rail line is 1.33 km away from CCAFS LC-40.




Section 4
Build a Dashboard with Plotly Dash

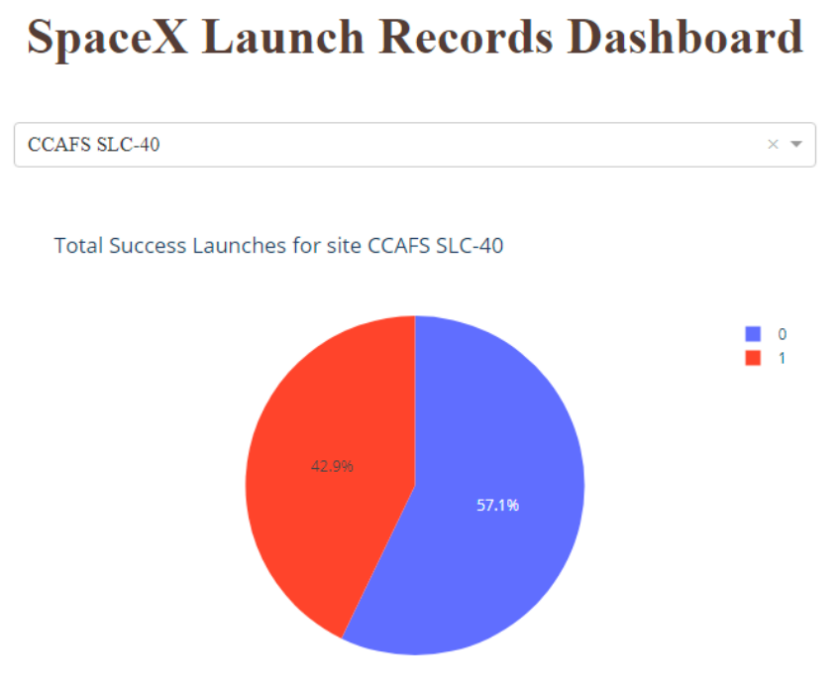
Dashboard for all launch sites

- The dropdown menu allows the selection of one or all launch sites.
- With all launch sites selected, the pie chart displays the distribution of successful Falcon 9 first stage landing outcomes between the different launch sites.
- The greatest share of successful Falcon 9 first stage landing outcomes (at 41.7% of the total) occurred at KSC LC-39A.



Launch Site with Highest Launch Success Ratio

Falcon 9 first stage failed landings are indicated by the '0' Class (■ blue wedge in the pie chart) and successful landings by the '1' Class (■ red wedge in the pie chart). CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42.9%).



Payload vs. Launch Outcome

- These screenshots are of the Payload vs. Launch Outcome scatter plots for all sites, with diferente payload selected in the range slider.
- The payload range from about 2,000 kg to 5,000 kg has the largest success rate.
- The ‘FT’ booster version category has the largest success rate.

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

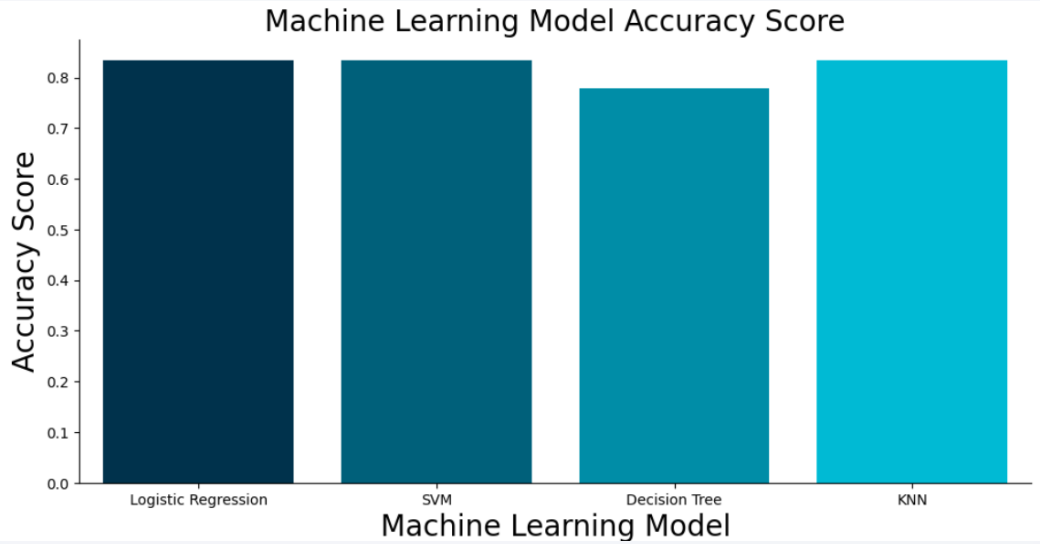
VAFB SLC-4E

Section 5

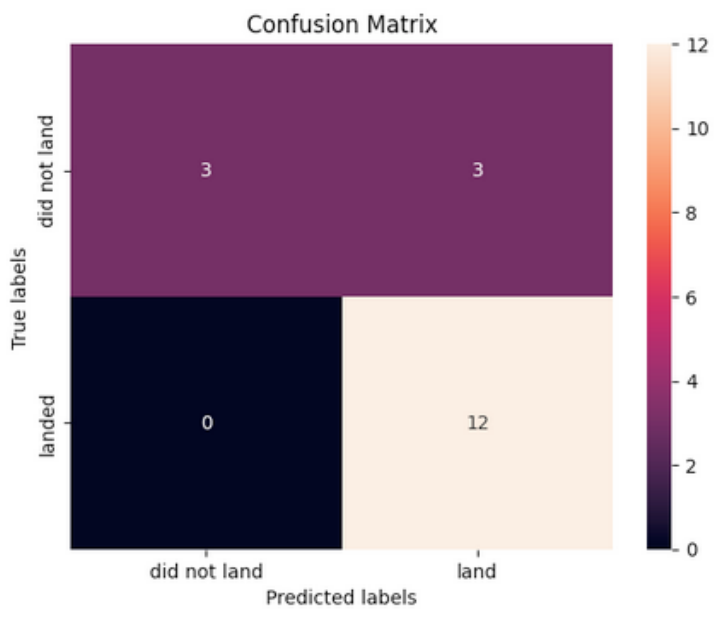
Predictive Analysis (Classification)

Classification Accuracy

- No one model is better than another. All models performed equally well except for the Decision Tree model which performed poorly relative to the other models.



Confusion Matrix



Shown here is the confusion matrix for the Logistic Regression model.

Prediction Breakdown:

- 12 True Positives and 3 True Negatives
- 3 False Positives and 0 False Negatives

Conclusions

- C1: SpaceX does not have a perfect track record of Falcon 9 first stage landing outcomes.
- C2: SpaceX’s Falcon 9 first stage landing outcomes have been trending towards greater success as more launches are made.
- C3: The machine learning models can be used to predict future SpaceX Falcon 9 first stage landing outcomes.

Thank you!