My Capstone Project in Data Science

Hung Tran

2024-11-29

OUTLINE

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

EXECUTIVE SUMMARY

Summary of methodologies

- Data was collected from the SpaceX public API and public 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

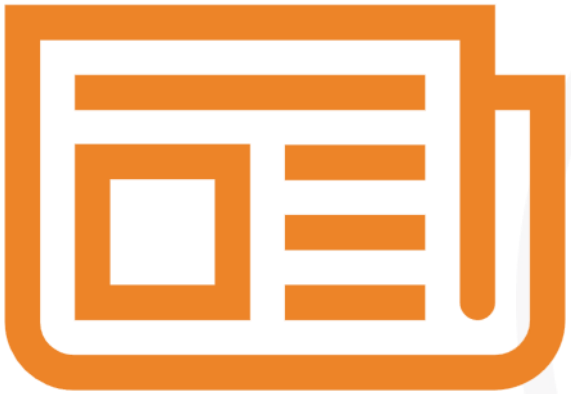
Space Y is a start-up space technology company that would like to compete with SpaceX founded by Billionaire industrialist Allon Musk. My project is to determine the price of each launch by gathering information about Space X and determine the successfulness if SpaceX will reuse the first stage. The outcome of this report is to help Space Y has the insight from the SpaceX's experience. The report will answer following domains:

1.Deliver the overview data that we have on SpaceX Falcon 9 first stage landings.

2.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?

3.Will a future Falcon 9 first stage landing be successful?

METHODOLOGY



Methodology

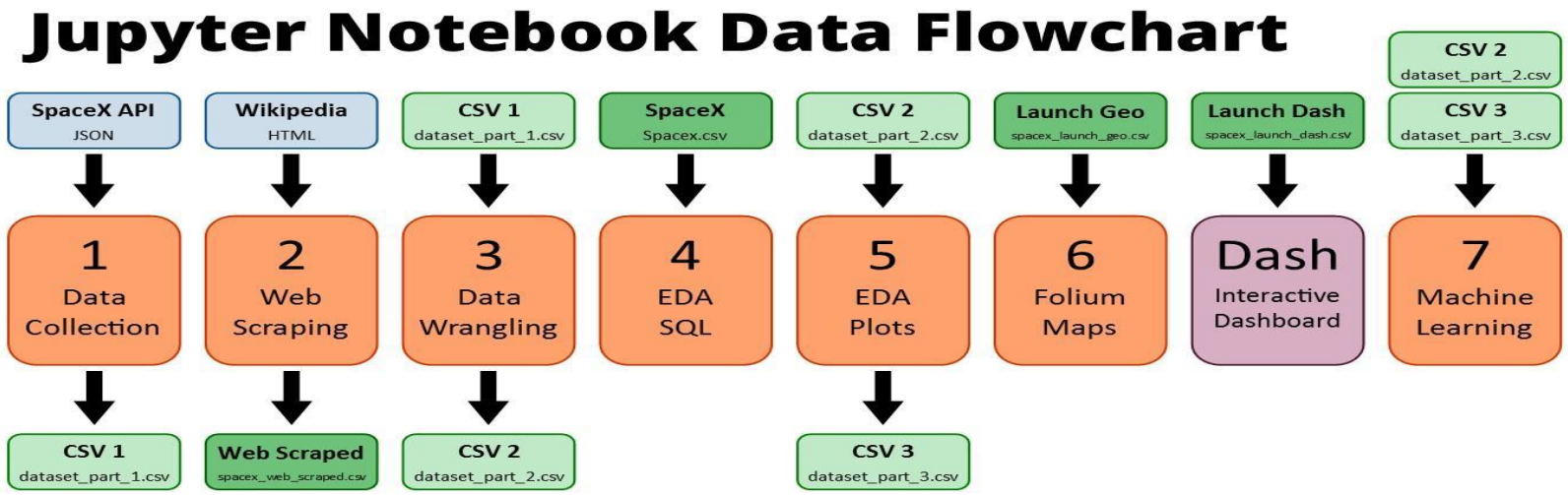
Summary

- 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

The data sets were collected from:

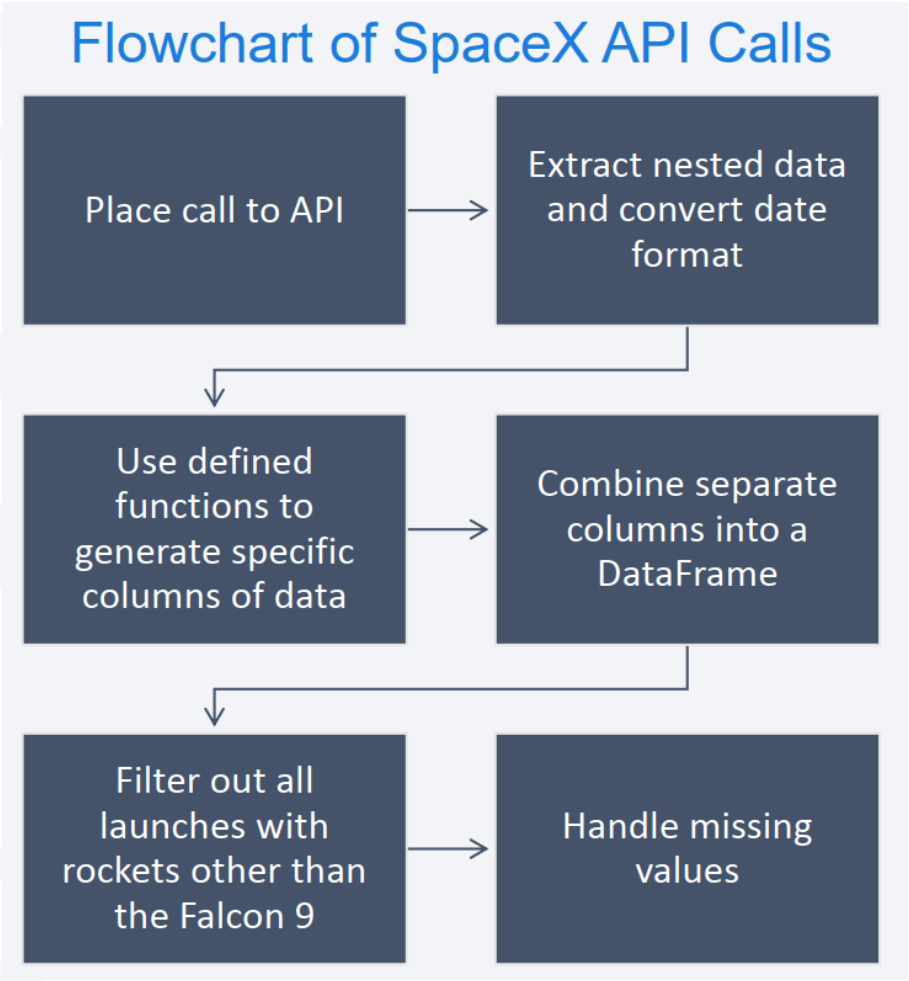
- An IBM copy of SpaceX API with launch data in JSON format.
- Wikipedia page with launch data in HTML tables
- Further data sets were provided. See following flowchart to see how data sources implemented throughout the project.



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
- Data Frame for further analysis.
- GitHub URL (Data Collection):

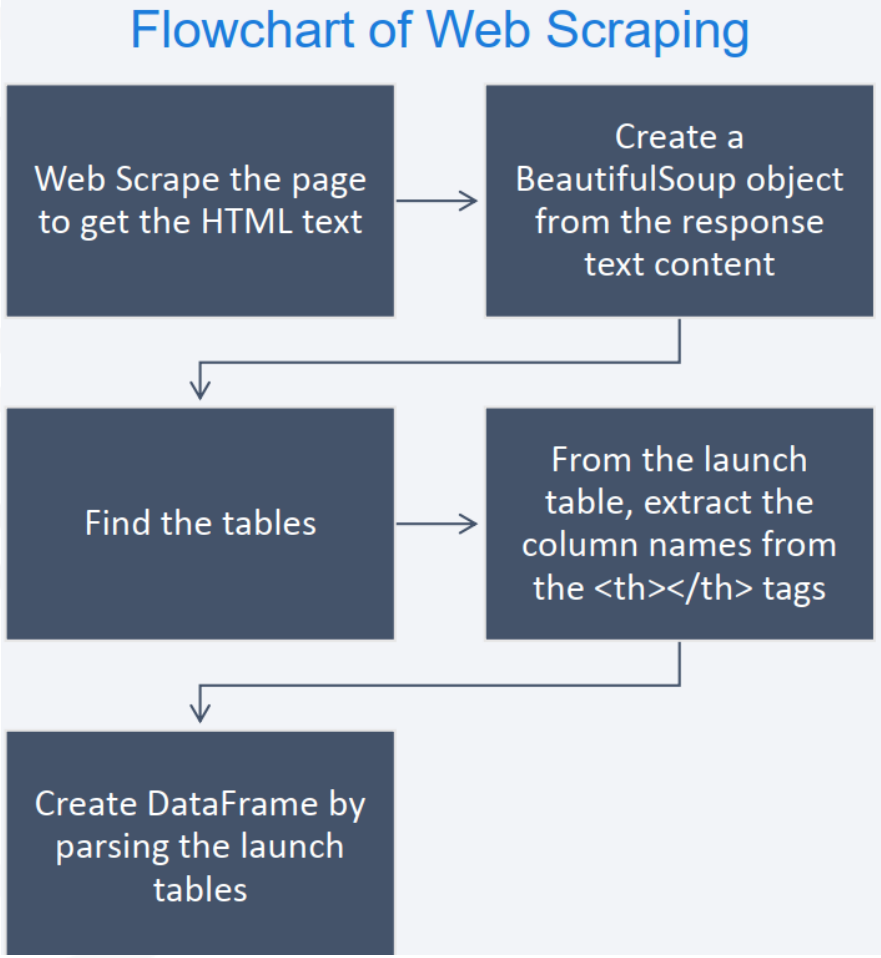
https://github.com/ruahaudau/Module_05_Capstone/blob/main/01.Data-Collection-API.ipynb



Data Collection – Web Scraping

- Wikipedia has a page that has tables of
- These tables can be scraped to extract launch data that can be put into a Pandas DataFrame for further analysis.
- GitHub URL (Web Scraping):

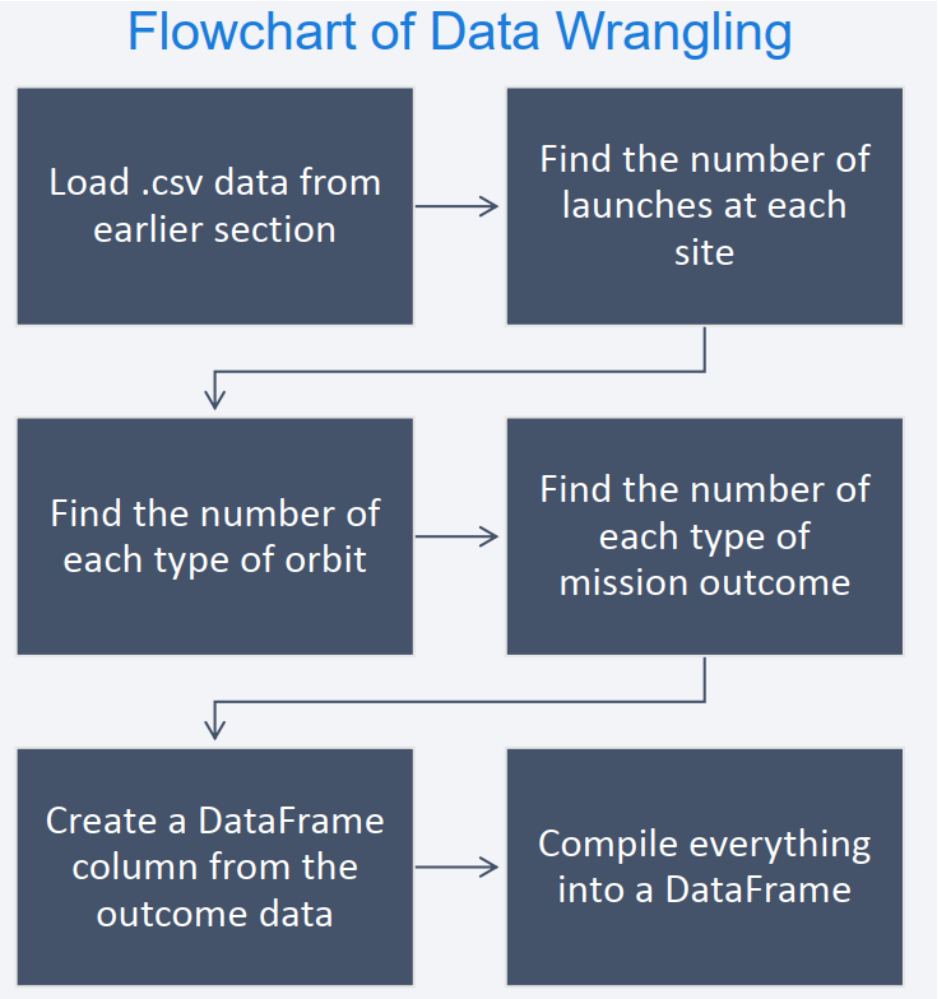
https://github.com/ruahaudau/Module_05_Capstone/blob/main/02.Data-Wrangling.ipynb



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 URL (Data Wrangling):

https://github.com/ruahaudau/Module_05_Capstone/blob/main/02.Data-Wrangling.ipy



EDA 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 (EDA with Data Visualization):

https://github.com/ruahaudau/Module_05_Capstone/blob/main/04.edadataviz.ipynb

EDA with SQL

Queries were written to extract information about:

- Launch sites
- Payload masses
- Dates
- Booster types
- Mission outcomes

GitHub URL (EDA with SQL):

https://github.com/ruahaudau/Module_05_Capstone/blob/main/03.EDA-SQL.ipynb

Interactive Map with Folium

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 (Folium Maps):

https://github.com/ruahaudau/Module_05_Capstone/blob/main/05.Launch_Site_Location.ipynb

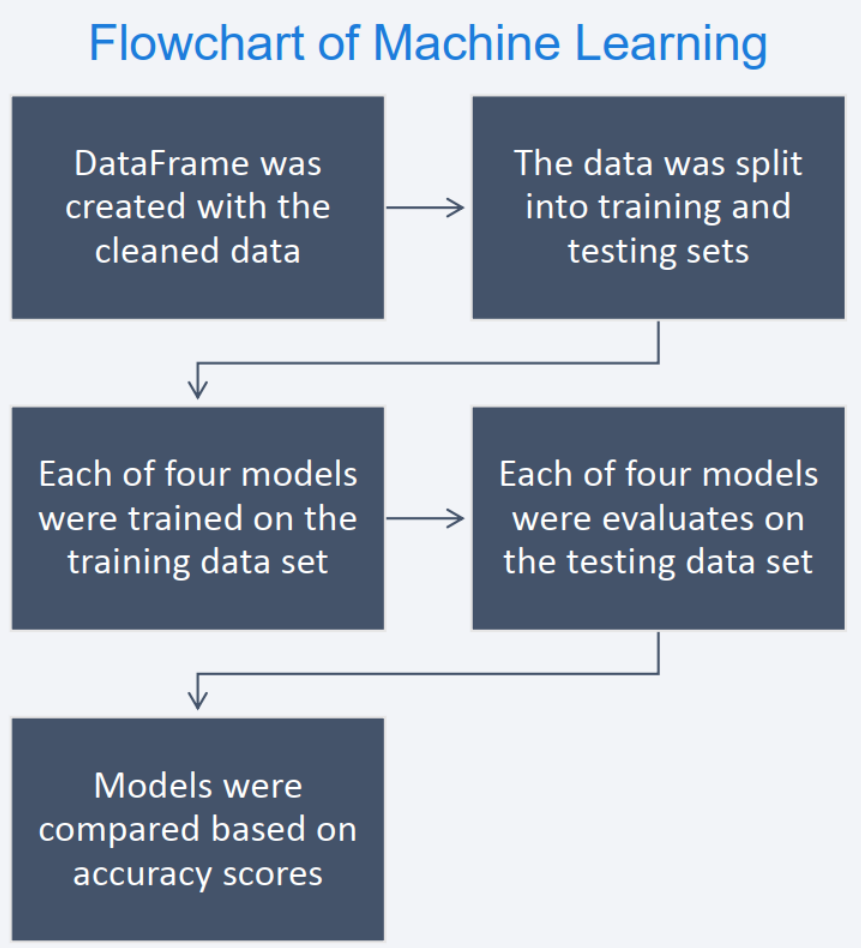
Dashboard with Plotly Dash

- 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 (Dashboard File): https://github.com/ruahaudau/Module_05_Capstone/blob/main/06.Dashboard.py

Predictive Analysis

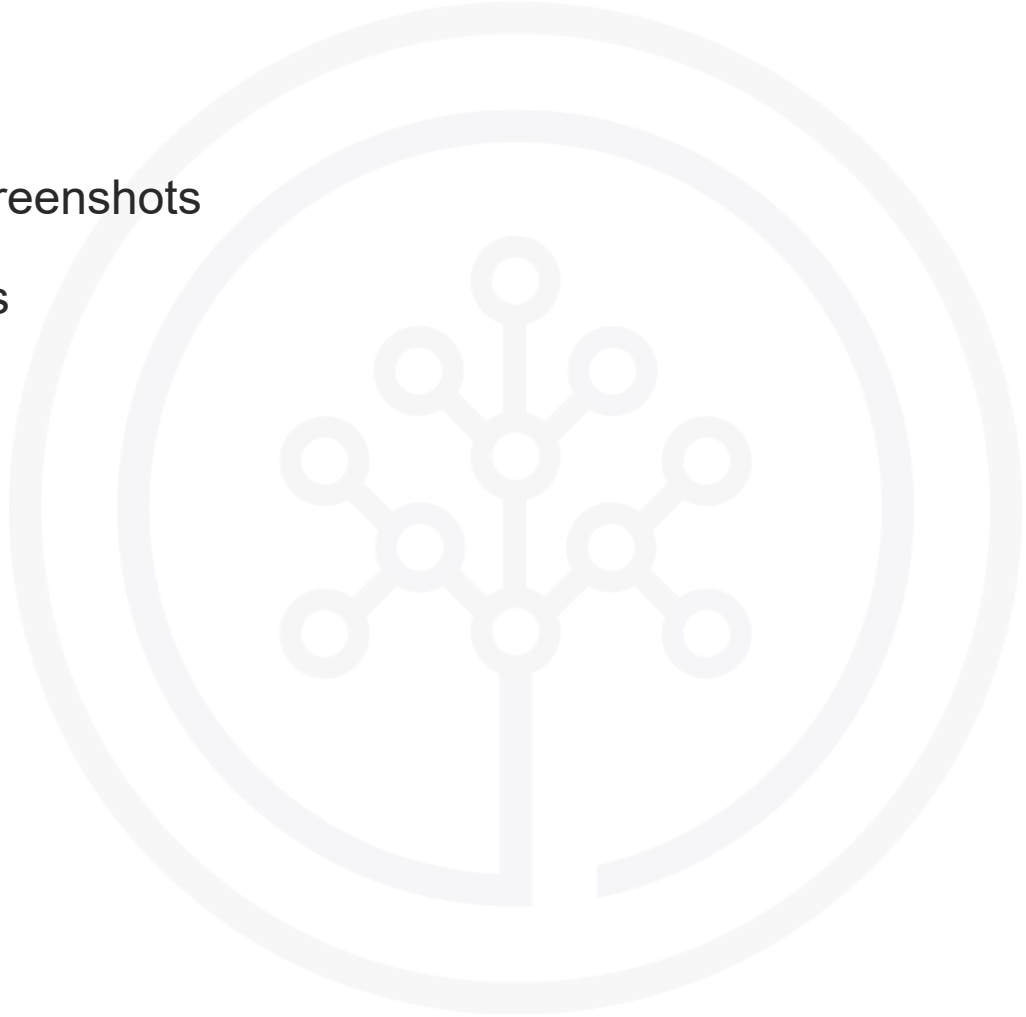
- The dataset was split into training and testing sets.
- Logistic Regression, SVM (Support Vector Machine), Decision Tree, and KNN (k-Nearest Neighbors) machine learning models were trained on the training data set.
- Hyper-parameters were evaluated using GridSearchCV() and the best was selected using '.best_params_'.
- Using the best hyper-parameters, each of the four models were scored on accuracy by using the testing data set.
- GitHub URL (Machine Learning):

https://github.com/ruahaudau/Module_05_Capstone/edit/main/07.SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

- Data analysis results
- Interactive analytics in screenshots
- Predictive analysis results

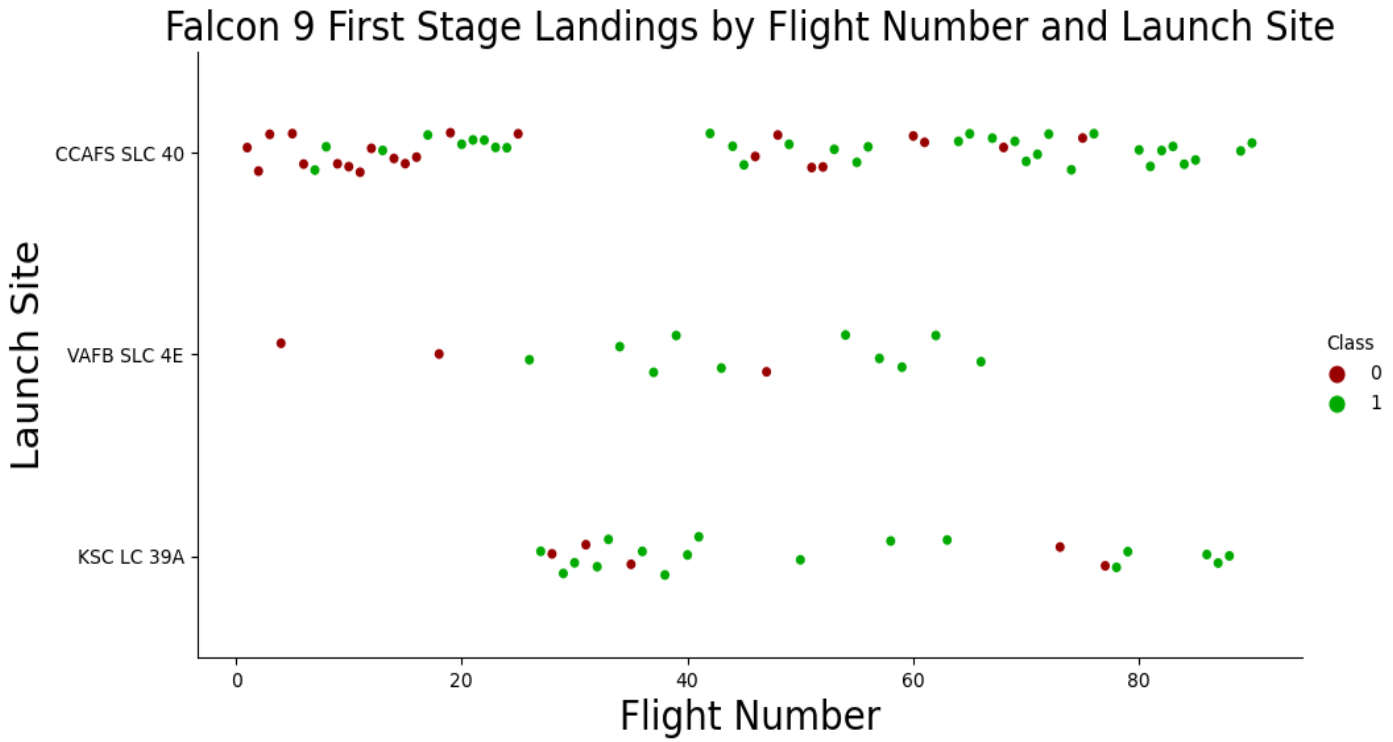


Part 01: Insights from EDA

1.1 Fly 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.

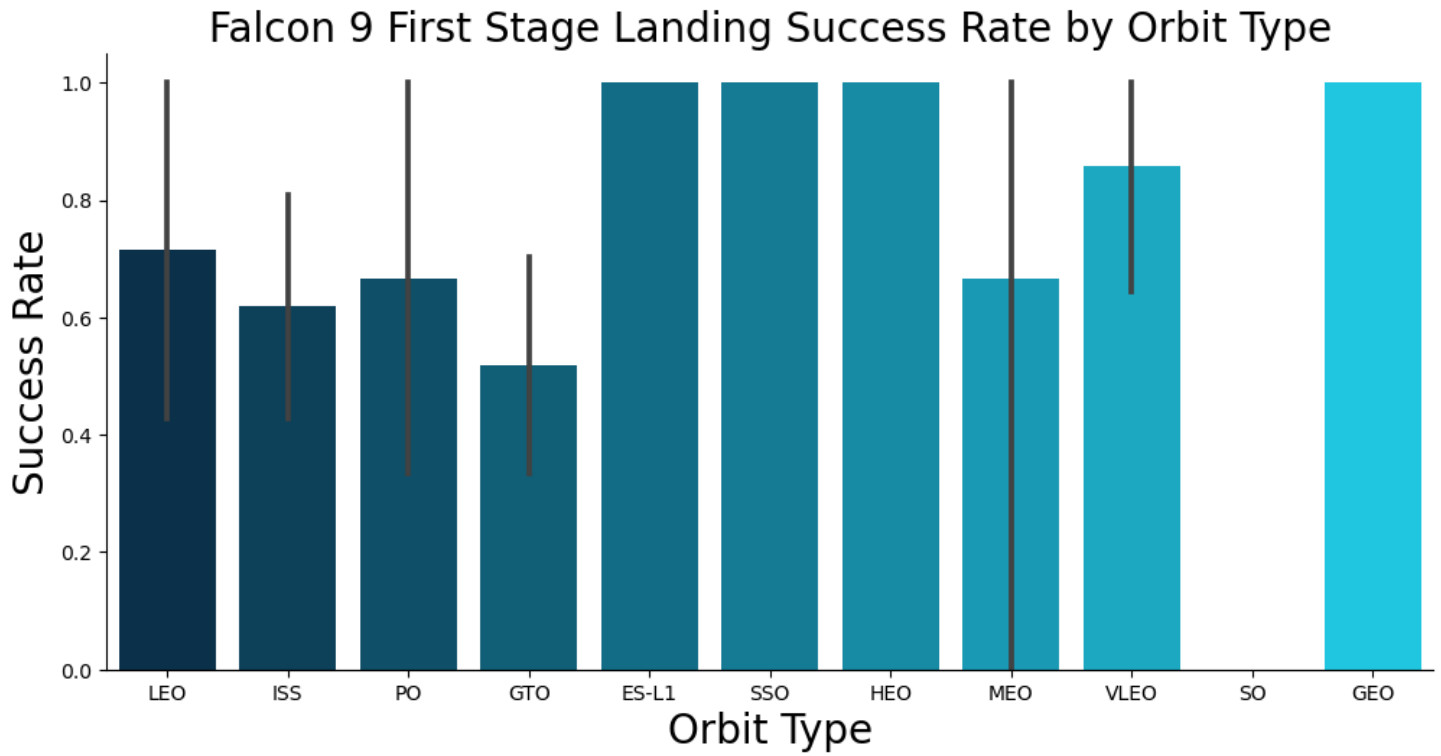


Falcon 9 first stage failed landings are indicated by the '0' Class (● red markers) and successful landings by the '1' Class (● green markers).

Part 01: Insights from EDA

1.2 Success rate vs. Orbit types

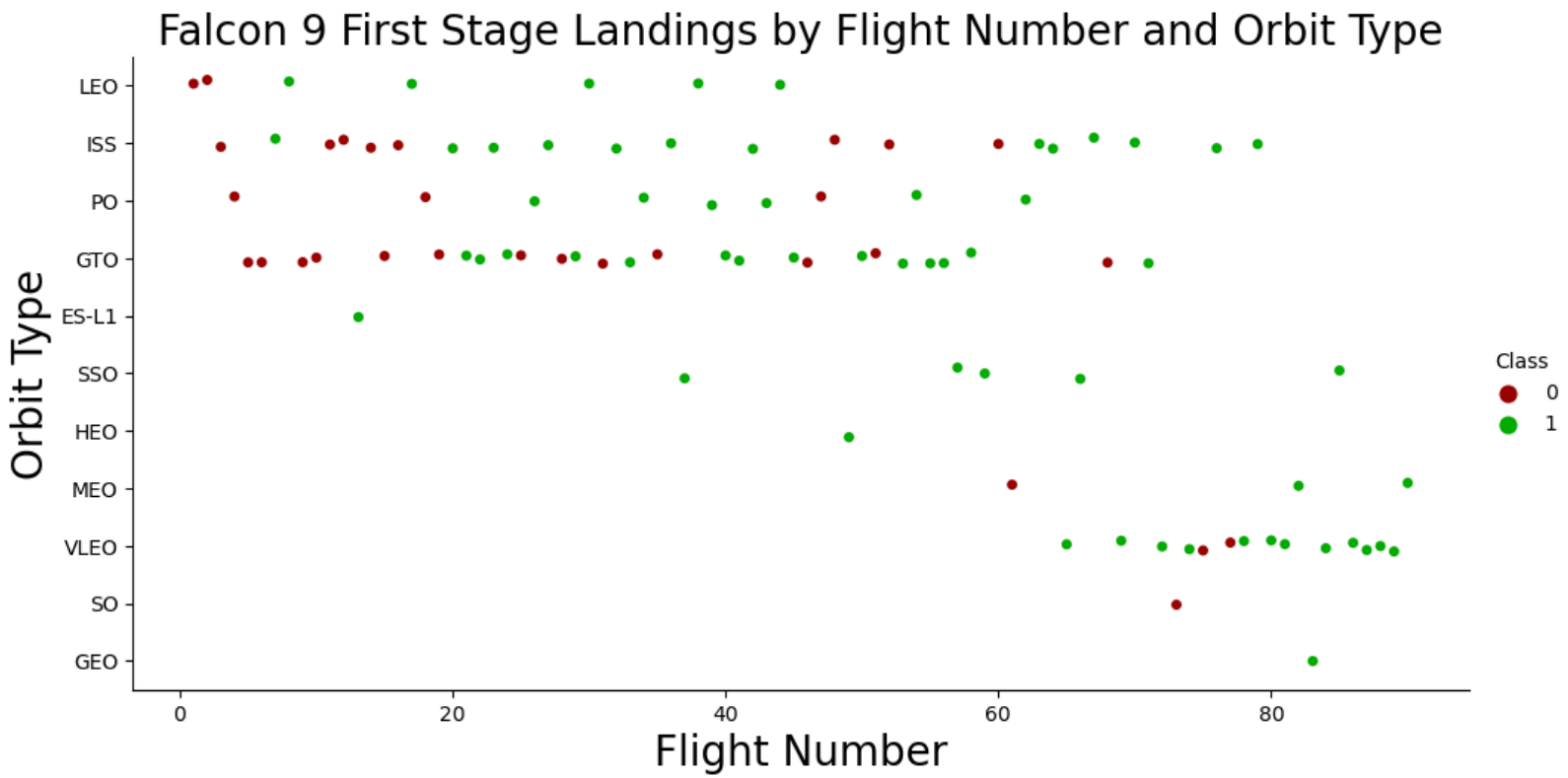
- ES-L1, SSO, HEO and GEO orbits have no failed first stage landings.
- SO orbits have no successful first stage landings.



Part 01: Insights from EDA

1.3 Fly number vs. Orbit types

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

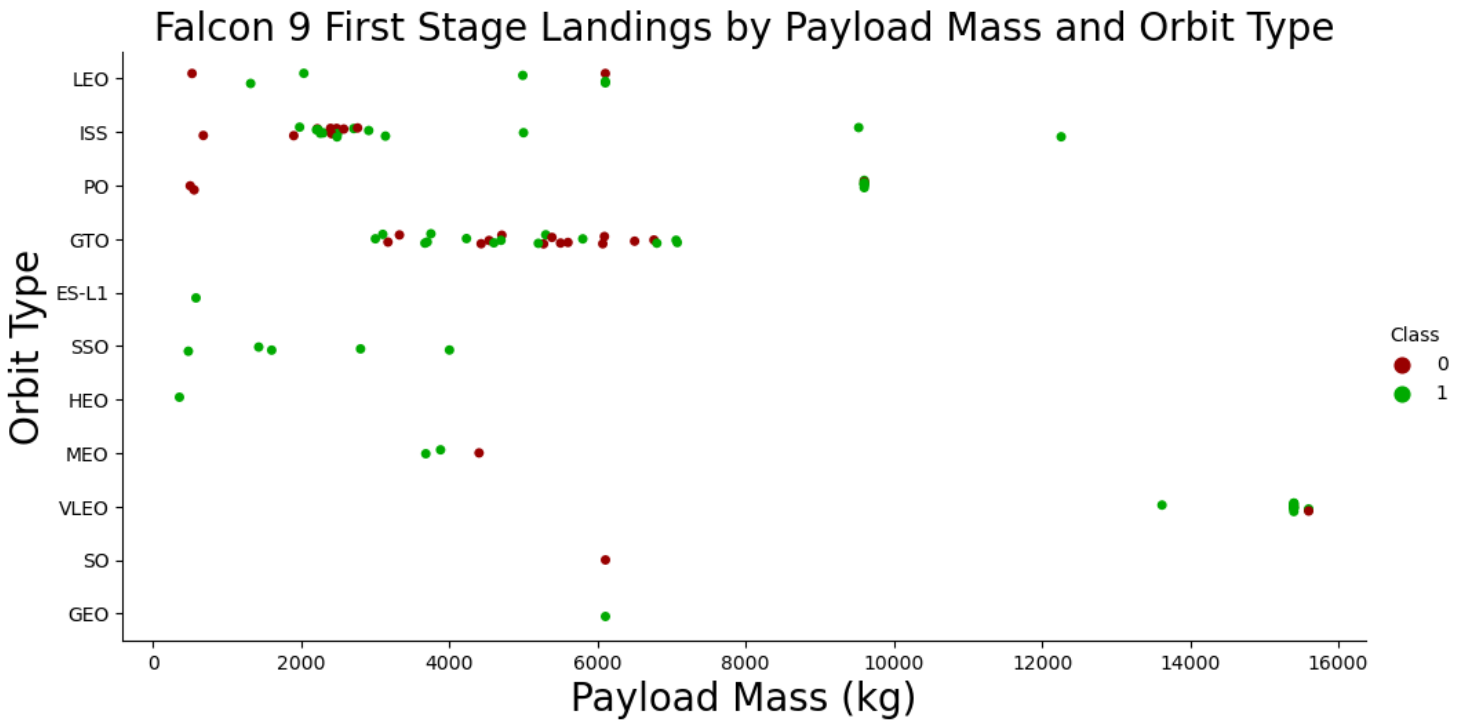


Part 01: Insights from EDA

1.4 Payload vs. Orbit types

- Some orbit types have better success rates than others.

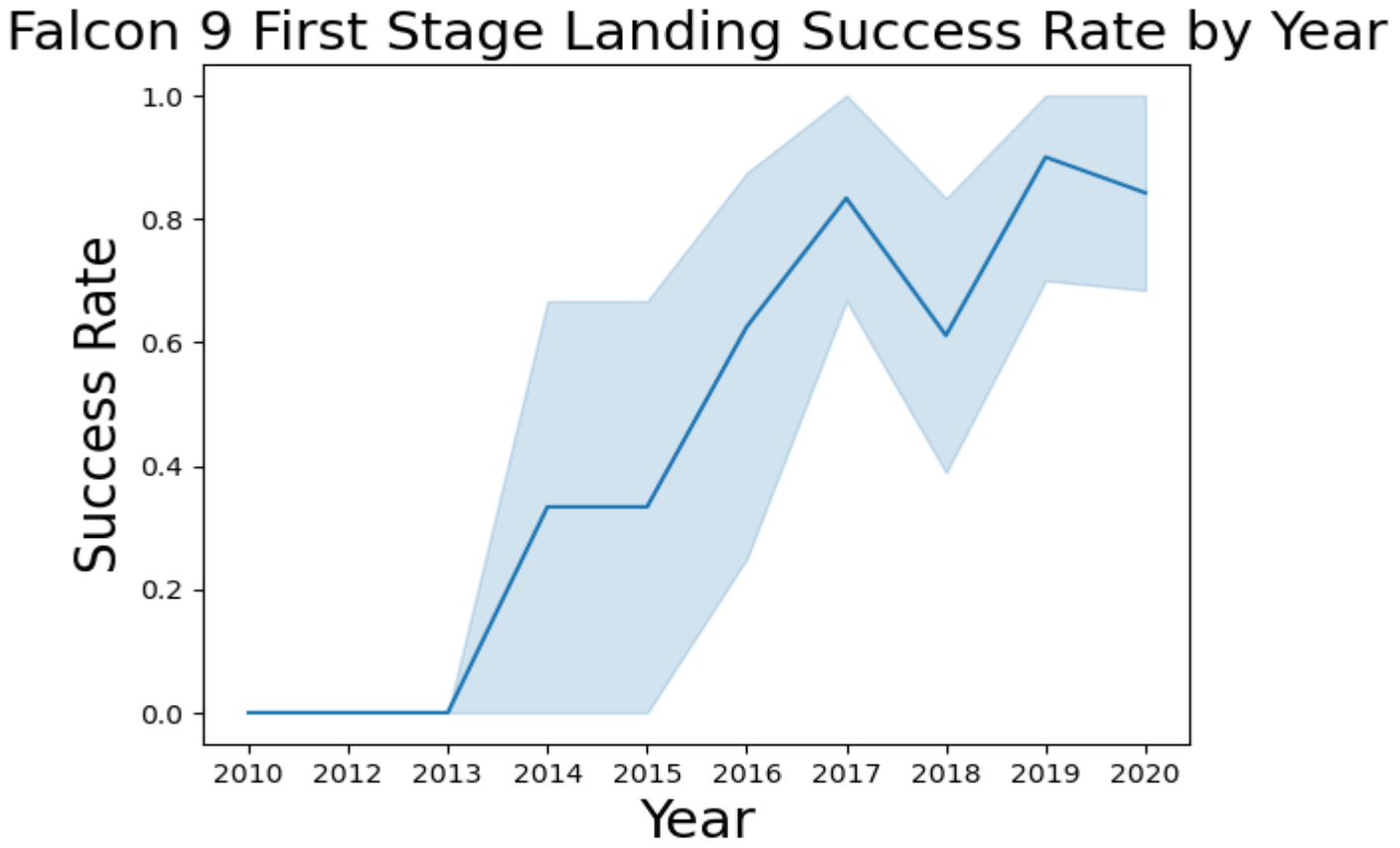
- Success rate appears to have no obvious correlation with payload mass.



Part 01: Insights from EDA

1.5 Landing Success trends

• The success rate has increased significantly over the years.



Part 02: Extracting Data with SQL

2.1 Launch site names

- Question: What are the names of all launch site?
- Query: “*select distinct Launch_Site from SPACEXTBL*”
- Result:

```
[25]: temp7 = "Select distinct Launch_Site from SPACEXTBL"
      query_temp(temp7)

('CCAFS LC-40',)
('VAFB SLC-4E',)
('KSC LC-39A',)
('CCAFS SLC-40',)
```

- Finding: There are 04 launch sites.

Part 02: Extracting Data with SQL

2.2 Launch site name begin with “CCA”

- Question: Find 05 records with launch sites begin with “CCA” ?
- Query: *“select * from SPACEXTBL 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

- Finding: from this query, we could have an overview of data with specific condition, likes the launch site name with “CCA” string.

Part 02: Extracting Data with SQL

2.3 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

- Finding: The total payload mass is 48,213 (Kg)

Part 02: Extracting Data with SQL

2.4 Average Payload Mass

• 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

• Finding: The average payload mass mass carried by booster version F9 v1.1 is 2,928 (Kg).

Part 02: Extracting Data with SQL

2.5 First Successful 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__outcome LIKE ‘Success (ground pad)’ “

• Result:

First Successful Landing Outcome Date
2015-12-22

• Finding: The first successful landing was 22/12/2015.

Part 02: Extracting Data with SQL

2.6 Successful Drone Ship Landing

• 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

• Finding: There are 04 versions of booster which is successful landing on drone ship with payload mass between 4000 and 6000 (Kg).

Part 02: Extracting Data with SQL

2.7 Number of successful and failure missions

• 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

• Finding: There are 61 successful and 40 failed missions.

Part 02: Extracting Data with SQL

2.8 Booster 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

• Finding: The maximum payload carried is 15,600 (kg) and there were 12 missions with vary booster types perform this outcome.

Part 02: Extracting Data with SQL

2.9 Launch Records

• Question: 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

• Finding: There were 02 failed missions in 2015. Both come from CCAFS LC-40 launch site.

Part 02: Extracting Data with SQL

2.10 Rank Landing Outcomes

• Question: 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

• Finding: The landing outcome “No attempt” is the most result.

Part 03: Proximities Analysis

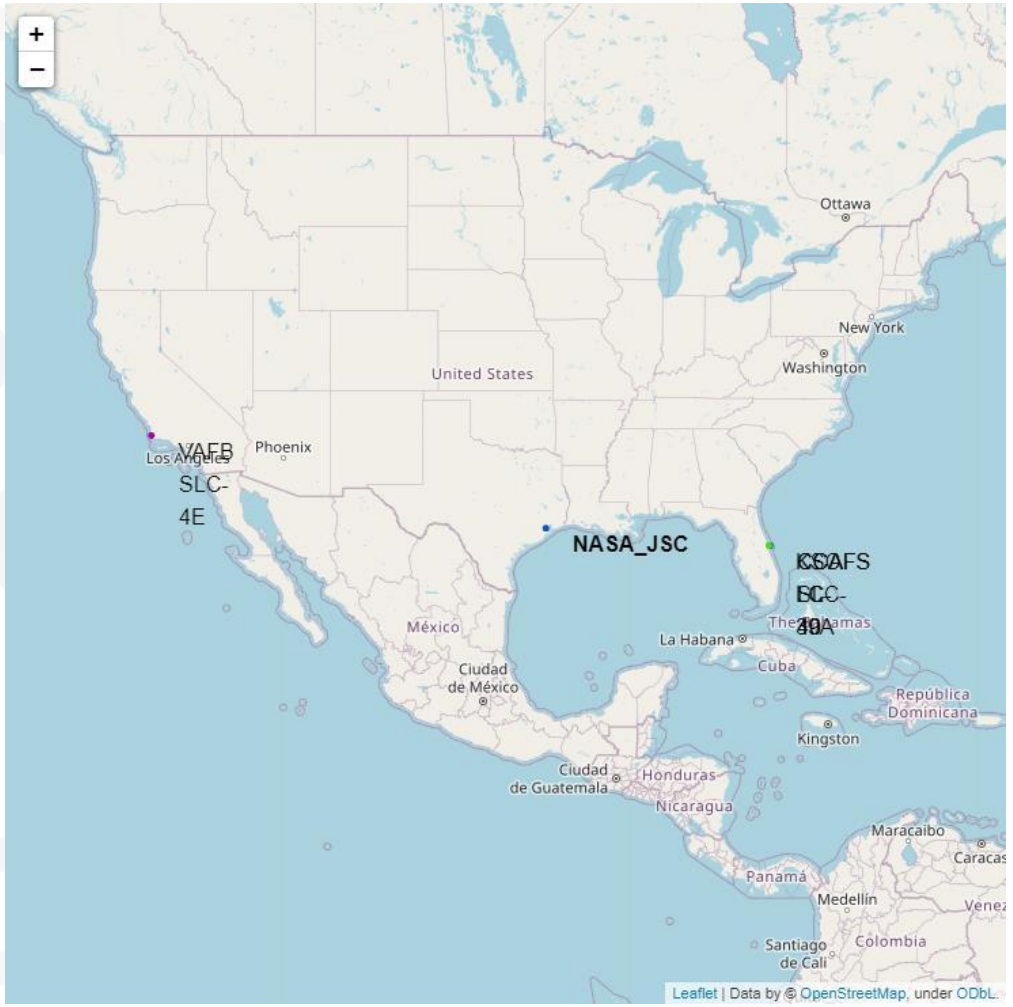
3.1 Launch Site Locations

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

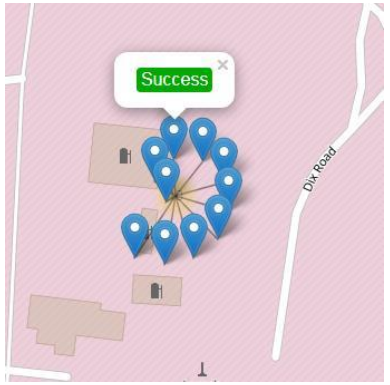


Part 03: Proximities Analysis

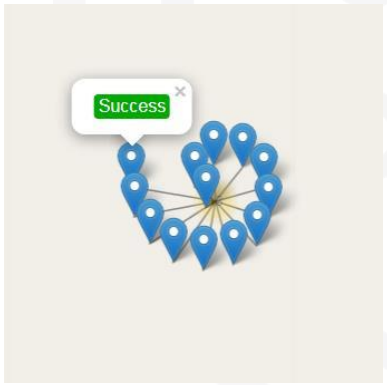
3.2 Map Markers of Successful/ Failure Landing

• 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.

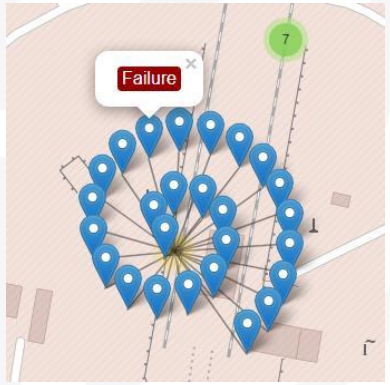
• A sense of a launch site's success rate for Falcon 9 first stage landings can be gleaned from the relative number of green success markers to red failure markers.



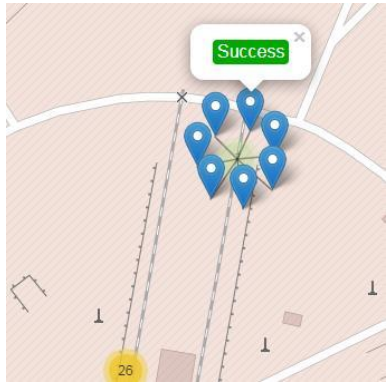
VAFB SLC-4E



KSC LC-39A



CCAFS LC-40



CCAFS SLC-40

Part 03: Proximities Analysis

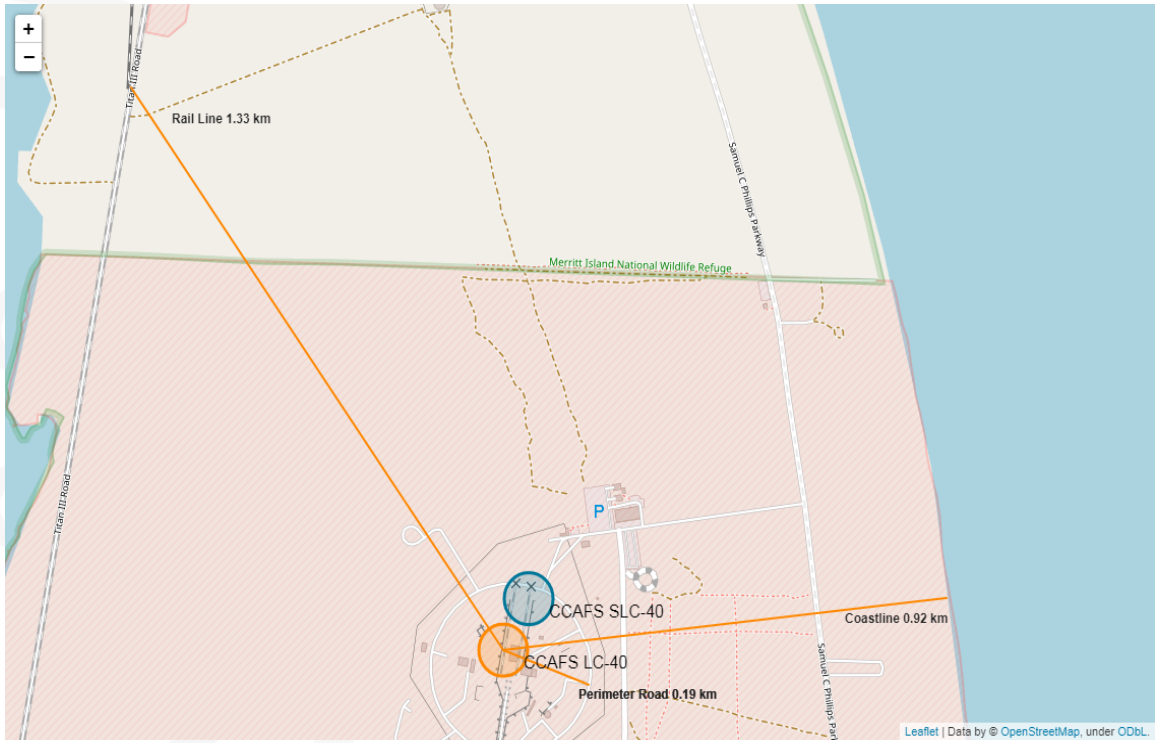
3.3 Launch Site Locations

• 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.



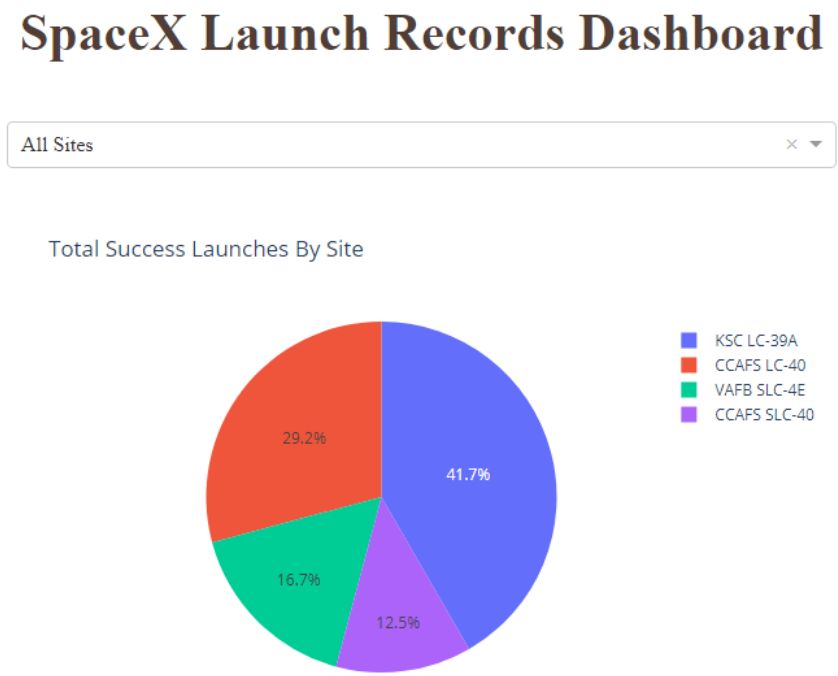
Part 04: Dashboard with Plotly Dash

4.1 Launch Counts

• 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.

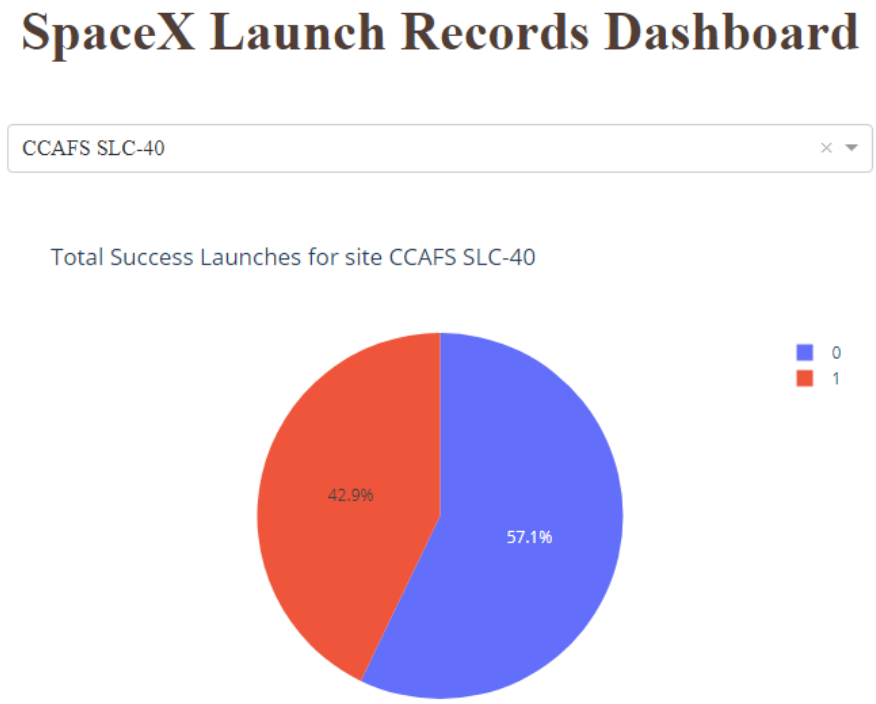


Part 04: Dashboard with Plotly Dash

4.2 Launch Sites with highest successful 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).

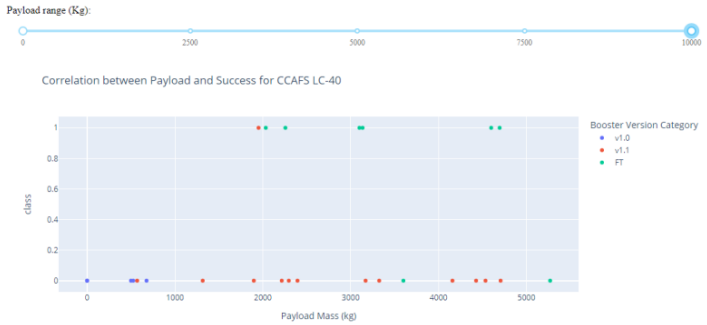
• CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42.9%).



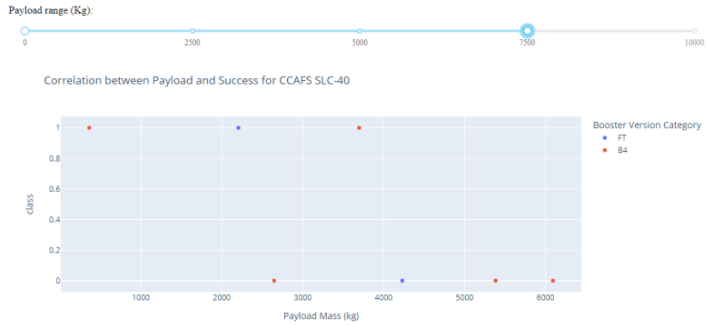
Part 04: Dashboard with Plotly Dash

4.3 Payload vs. Outcomes

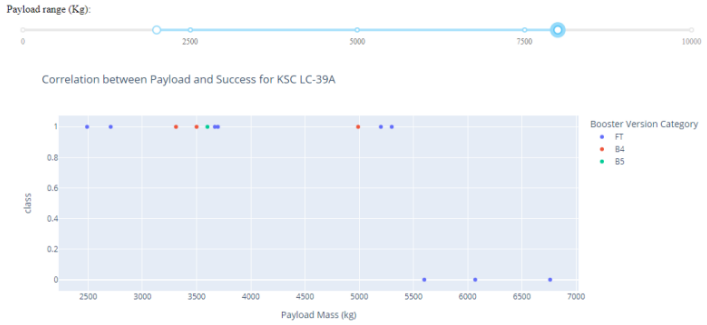
- These screenshots are of the Payload vs. Launch Outcome scatter plots for all sites, with different payload selected in the range slider.
- The payload range from about 2,000 kg to 5,000 kg has the largest success rate.
- The ‘FT’ booster version category has the largest success rate.



CCAFS LC-40



CCAFS SLC-40



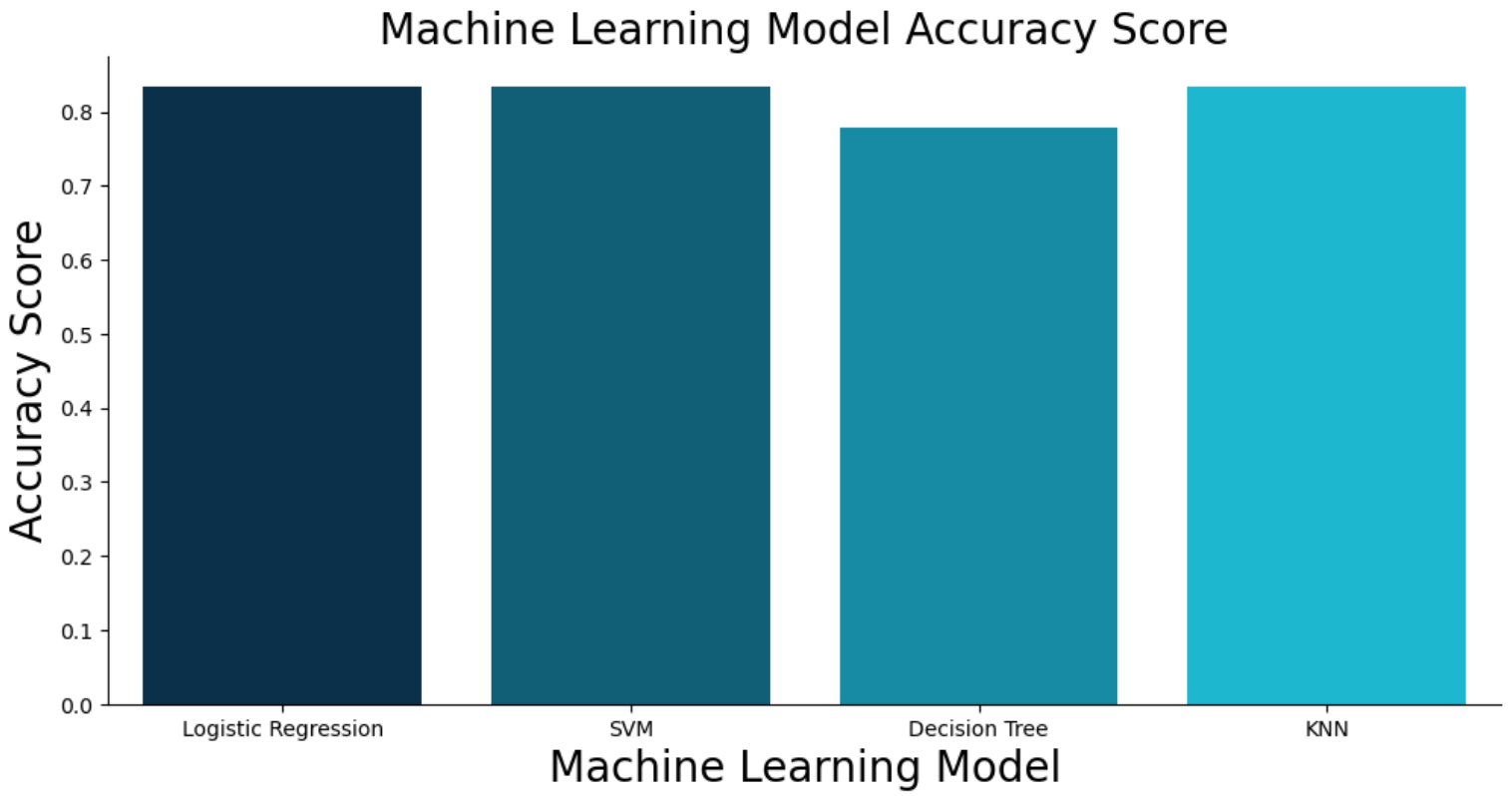
KSC LC-39A

VAFB SLC-4E

Part 05: Predictive Analysis (Classification)

5.1 Classification Accuracy

• All models performed equally well except for the Decision Tree model which performed poorly relative to the other models.



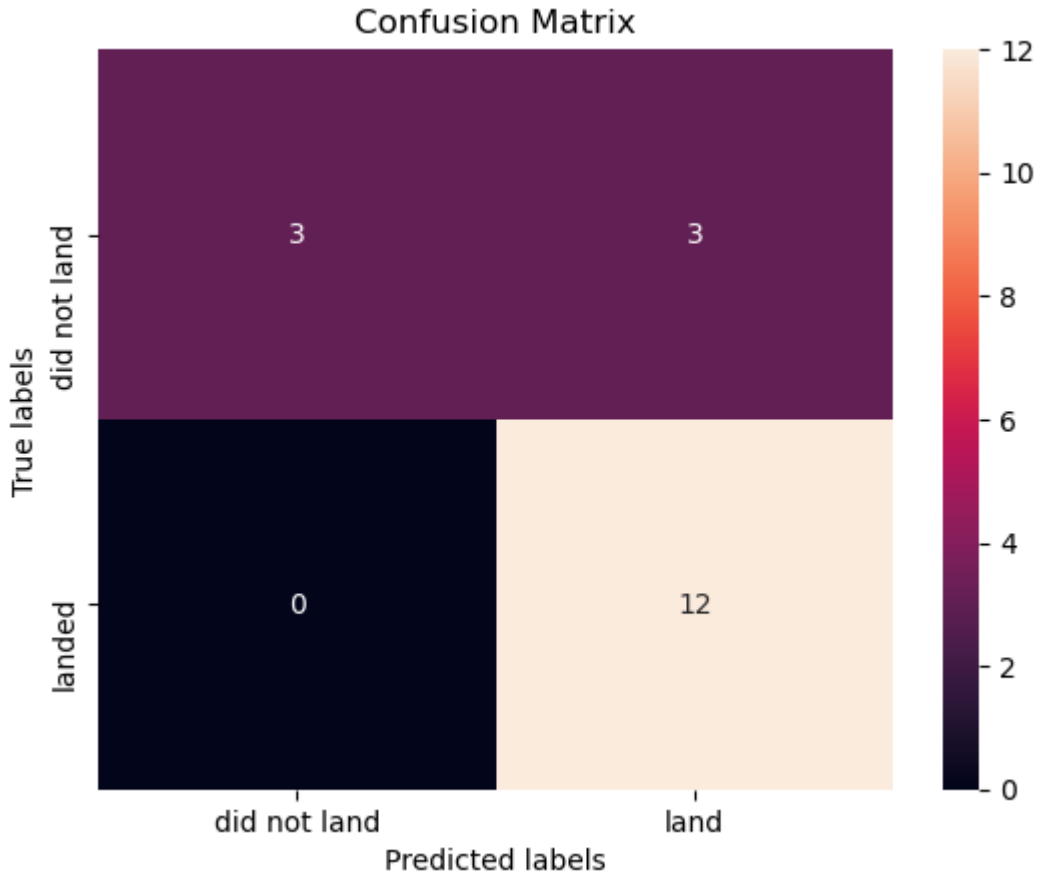
Part 05: Predictive Analysis (Classification)

5.2 Confusion Matrix

- Shown here is the confusion matrix for the Logistic Regression model.
- Confusion matrices can be read as:

True Negative	**False Positive**
False Negative	**True Positive**

- Prediction Breakdown:
  - 12 True Positives and 3 True Negatives
  - 3 False Positives and 0 False Negatives



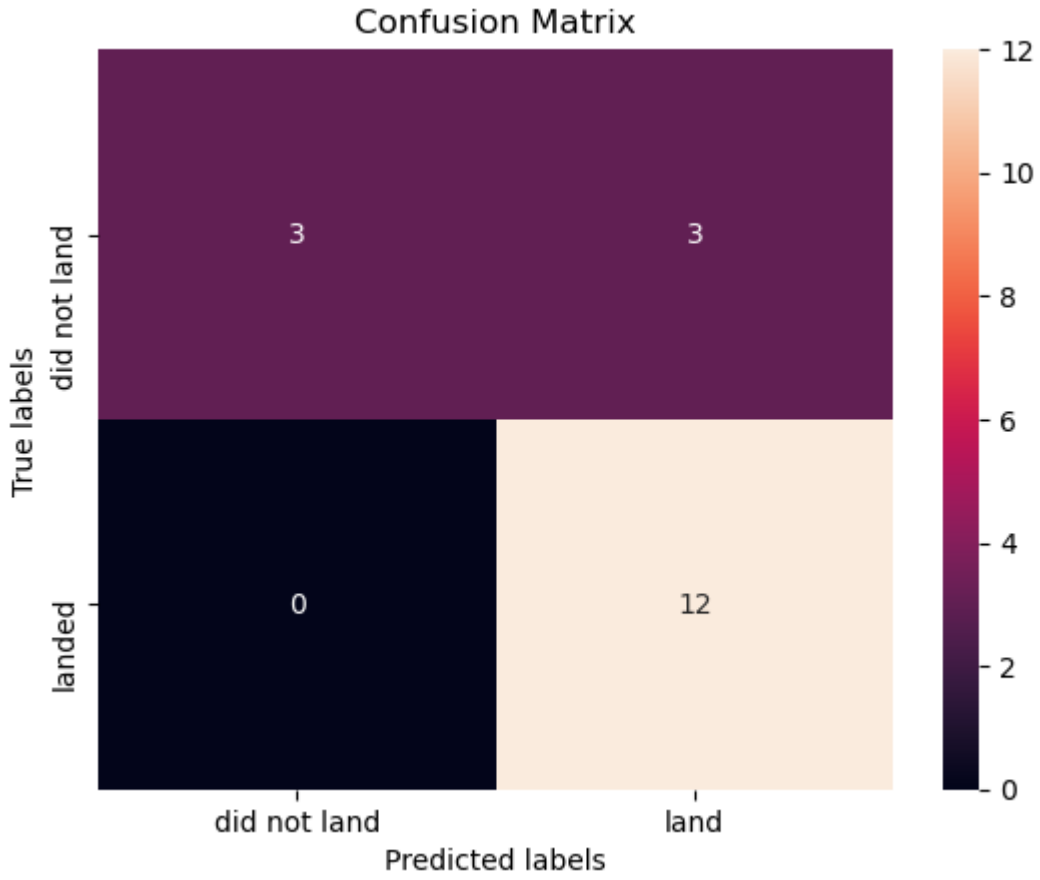
Part 05: Predictive Analysis (Classification)

5.2 Confusion Matrix

- Shown here is the confusion matrix for the Logistic Regression model.
- Confusion matrices can be read as:

True Negative	**False Positive**
False Negative	**True Positive**

- Prediction Breakdown:
  - 12 True Positives and 3 True Negatives
  - 3 False Positives and 0 False Negatives



Conclusion

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

Appendix

1. Initial Data Sets

- **SpaceX API (JSON):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- **Wikipedia (Webpage):** https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- **SpaceX (CSV):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321ENSkillsNetwork/labs/module_2/data/Spacex.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetworkChannel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01
- **Launch Geo (CSV):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
- **Launch Dash (CSV):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv

2. Data Sets (.csv files)

- **GitHub URL (CSV 1):** https://github.com/ruahaudau/Module_05_Capstone/blob/main/dataset_part_1.csv
- **GitHub URL (Web Scraped):** https://github.com/ruahaudau/Module_05_Capstone/blob/main/spacex_web_scraped.csv
- **GitHub URL (CSV 2):** https://github.com/ruahaudau/Module_05_Capstone/blob/main/dataset_part_2.csv
- **GitHub URL (SpaceX):** https://github.com/ruahaudau/Module_05_Capstone/blob/main/Spacex.csv
- **GitHub URL (CSV 3):** https://github.com/ruahaudau/Module_05_Capstone/blob/main/dataset_part_3.csv
- **GitHub URL (Launch Geo):** https://github.com/ruahaudau/Module_05_Capstone/blob/main/spacex_launch_geo.csv
- **GitHub URL (Launch Dash):** https://github.com/ruahaudau/Module_05_Capstone/blob/main/spacex_launch_dash.csv

Appendix

3. Jupyter Notebooks and Dashboard Python File

- **GitHub Data Collection:** https://github.com/ruahaudau/Module_05_Capstone/blob/main/01.Data-Collection-API.ipynb
- **GitHub Web Scraping:** https://github.com/ruahaudau/Module_05_Capstone/blob/main/02__Webscraping.ipynb
- **GitHub Data Wrangling:** https://github.com/ruahaudau/Module_05_Capstone/blob/main/02.Data-Wrangling.ipynb
- **GitHub EDA with SQL:** https://github.com/ruahaudau/Module_05_Capstone/blob/main/04.EDA_SQL.ipynb
- **GitHub EDA with Data Visualization:** https://github.com/ruahaudau/Module_05_Capstone/blob/main/04.edadataviz.ipynb
- **GitHub Folium Maps:** https://github.com/ruahaudau/Module_05_Capstone/blob/main/05.Launch_Site_Location.ipynb
- **GitHub Dashboad File:** https://github.com/ruahaudau/Module_05_Capstone/blob/main/06.Dashboard.py
- **GitHub Machine Learning:** https://github.com/ruahaudau/Module_05_Capstone/blob/main/07.SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb