IBM Developer
SKILLS NETWORK

Winning space race with data science

Shahira Najia Jamil
29 April 2024

Outline

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

Executive Summary

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

Machine learning models on the data set were performed equally. Launch data included information 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).

Introduction

Having undertaken the role of a Data Scientist, I have been tasked to predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of US$62 million. Other provider's cost is around US$ 165 million each, much of the which due to the first stage reuse by SpaceX. The first stage is determined through successful rocket landings and the factors that contributed to the success or failure. This presentation aims at understanding the launch performances of Falcon 9 that are also applied to other rockets, where Data Science have been useful in improving the future space exploration endeavours.

Problems

Firstly, we develop the problem statements to ascertain the objective of this project and they are as such:

- Conditions will the rocket be able to land safely
- Effects of each relationship of rocket variables on outcome
- Criteria that will aid SpaceX to achieve best results

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  Data is compiled through SpaceX API and web scrapping from Wikipedia. Data was cleaned to prepare for visualisation, queries and machine learning model creation

- Perform data wrangling
  One hot encoding data fields for machine learning and dropping irrelevant columns (Transforming data for Machine Learning).

Methodology (cont'd)

- Perform exploratory data analysis (EDA) using visualization and SQL
  Visualisations such as scatter and bar plots that show data behaviour

- Perform interactive visual analytics using Folium and Plotly Dash
  Such as Folium and Plotly Dash visualization

- Perform predictive analysis using classification models
  Develop and assess classification models

Data collection

Getting data from API or webpage → Develop data frame → Filter data frame according to requirement → Export to flat file (csv)

Data collection entails information gathering process from an established system. Collected data is used to Explanation relevant questions and evaluated the outcomes. Data collection process for this project includes using get request to the SpaceX API. Data is decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize(). Data undergoes cleaning and checked for missing values and fill in missing values where necessary. Then, web scraping is performed from Wikipedia for Falcon 9 launch records with BeautifulSoup. The objective is to extract the launch records as HTML table, parse the table and convert it to a pandas data frame for future analysis.

Data Collection – SpaceX API

SpaceX API Calls Flowchart

1. Place call to API
2. Extract nested data and convert the data format
3. Utilise defined functions to develop specific data columns
4. Combine separate columns into a DataFram
5. Filter out all launches with rockets that are not the Falcon 9

The SpaceX API has data available publicly. 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/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%201(A)%20Data%20collection%20API%20checkpoint.ipynb

Data Collection – Scraping

Web Scraping Flowchart

Web scrap the page to obtain HTML text

↓

Create a BeautifulSoup object from response text content

↓

Find tables

↓

Extract column names from <th></th> tags (from launch table)

↓

Create DataFrame by parsing launch tables

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

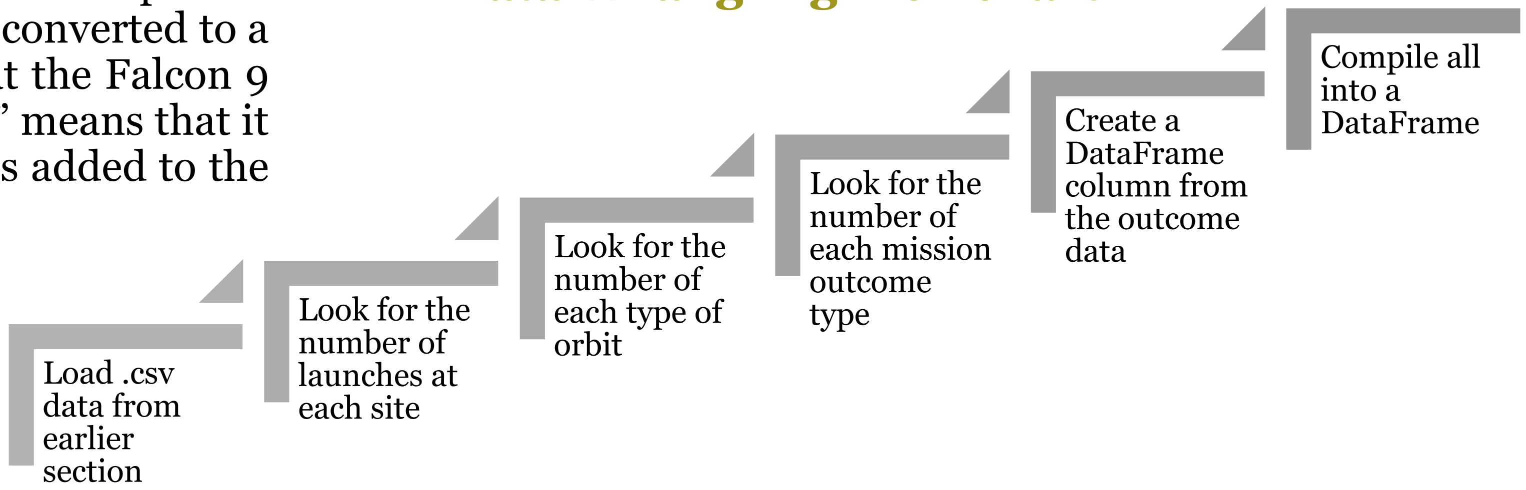
GitHub URL (Web Scraping):

https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%201(B)%20Data%20Collection%20with%20Web%20Scraping%20lab.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.

Data Wrangling Flowchart



GitHub URL (Data Wrangling):

https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%201(C)%20Data%20Wrangling.ipynb

EDA with Data Visualisation

- **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 Visualisation): https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%202(B)%20EDA%20with%20Visualisation%20lab.pdf

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/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%202(A)%20EDA%20with%20SQL%20(2).ipynb

Build an Interactive Map with Folium

Markers, circles and lines were added to the Folium map for the reasons below:

- **Markers** for launch sites and for the NASA Johnson Space Center

- **Circles** for the launch sites.

- **Lines** 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/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%203A%20Interactive%20Visual%20Analytics%20with%20Folium%20lab.pdf

Build a 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:
  (1) For All Sites – the distribution of successful Falcon 9 first stage landings between the sites
  (2) 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/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%203(B)Plotly%20Dash.py

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 was scored on accuracy by using the testing data set.

Machine Learning Flowchart

1. DataFrame was developed with cleansed data
2. The data was split into training and testing sets
3. Each of the four models was trained via training dataset
4. Each of the four models was evaluated via testing dataset
5. Models were compared according to accuracy results

GitHub URL (Machine Learning):

https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%204%20Machine%20Learning%20Prediction%20lab.pdf

Results

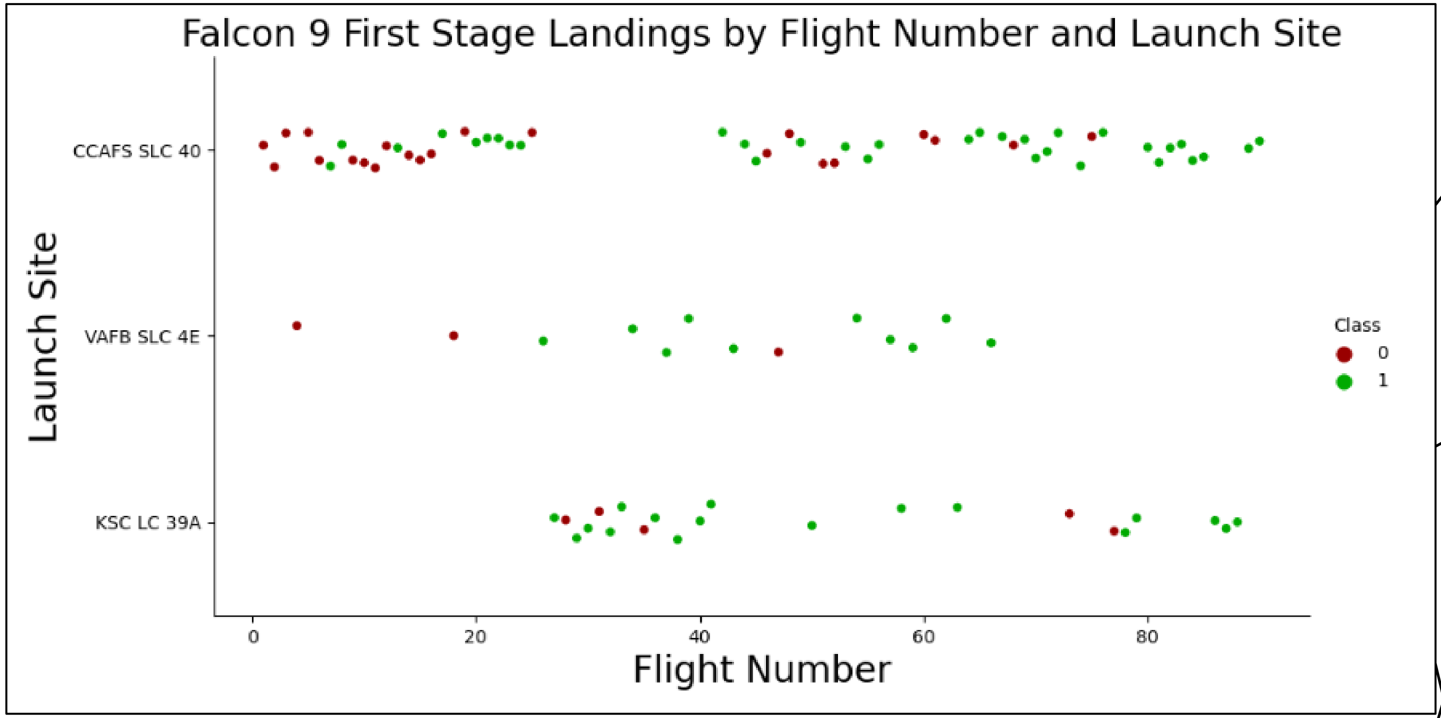
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Data 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.

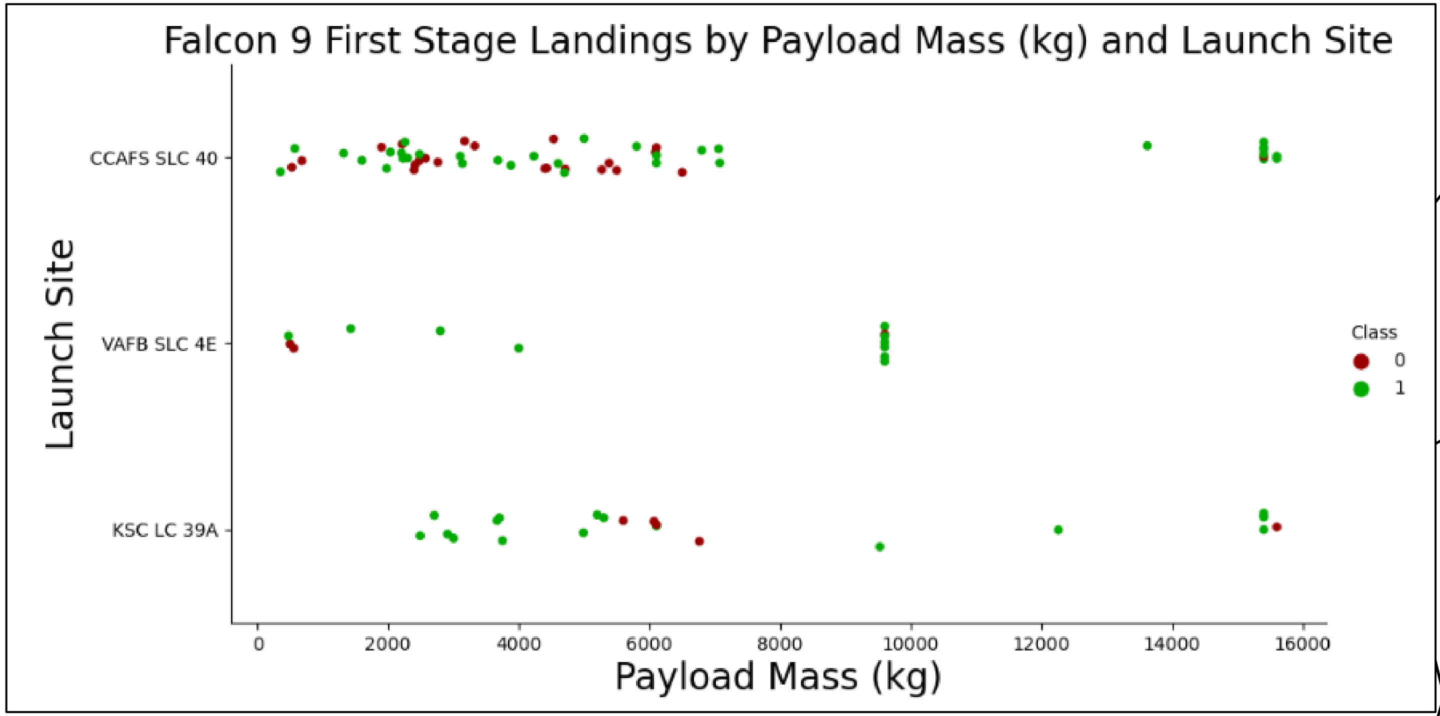


Falcon 9 first stage failed landings are indicated by the '0' class (red markers) and successful landings by the '1' class (green markers)

Payload vs. Launch Site

- For the CCAFS SLC 40 launch site, the payload mass and the landing outcome appear to not be strongly correlated.

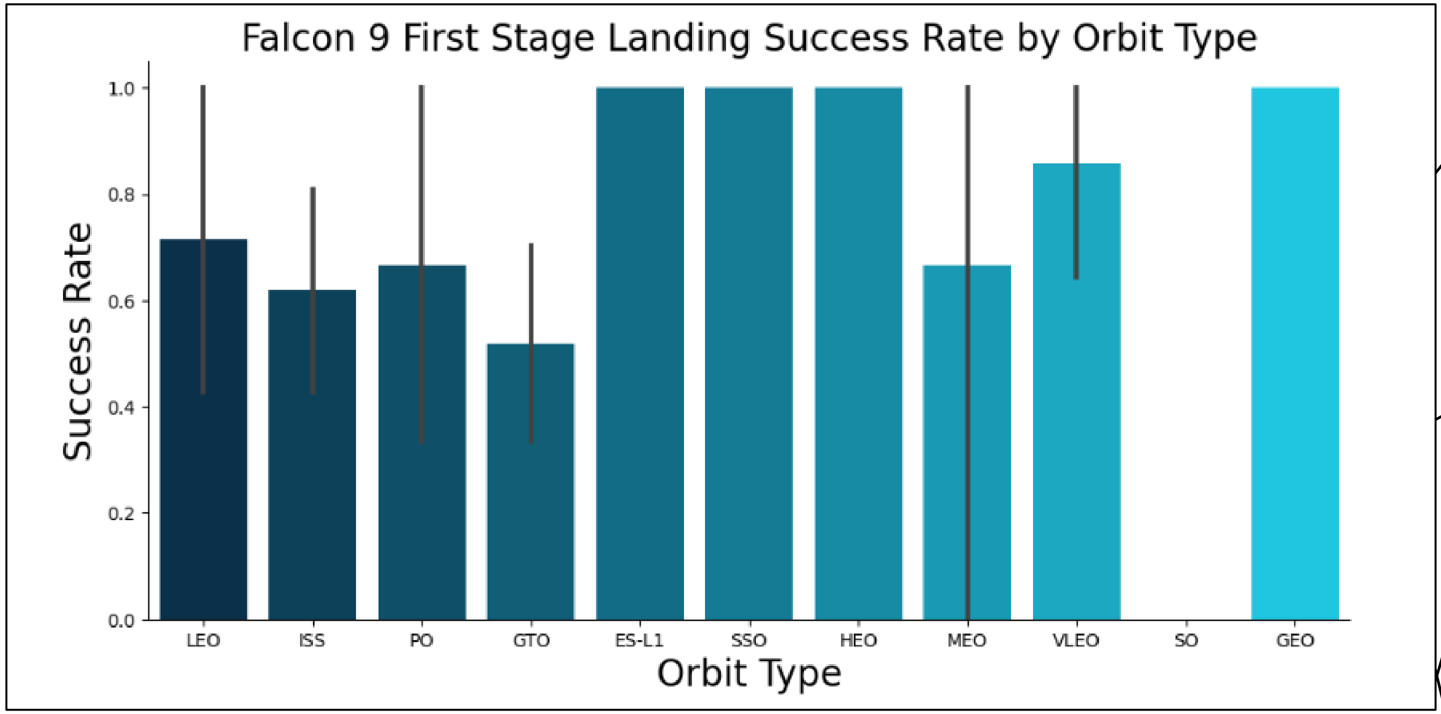
- The failed landings at the KSC LC 39A launch site are all grouped around a narrow band of payload masses.



Falcon 9 first stage failed landings are indicated by the ‘0’ class (red markers) and successful landings by the ‘1’ class (green markers)

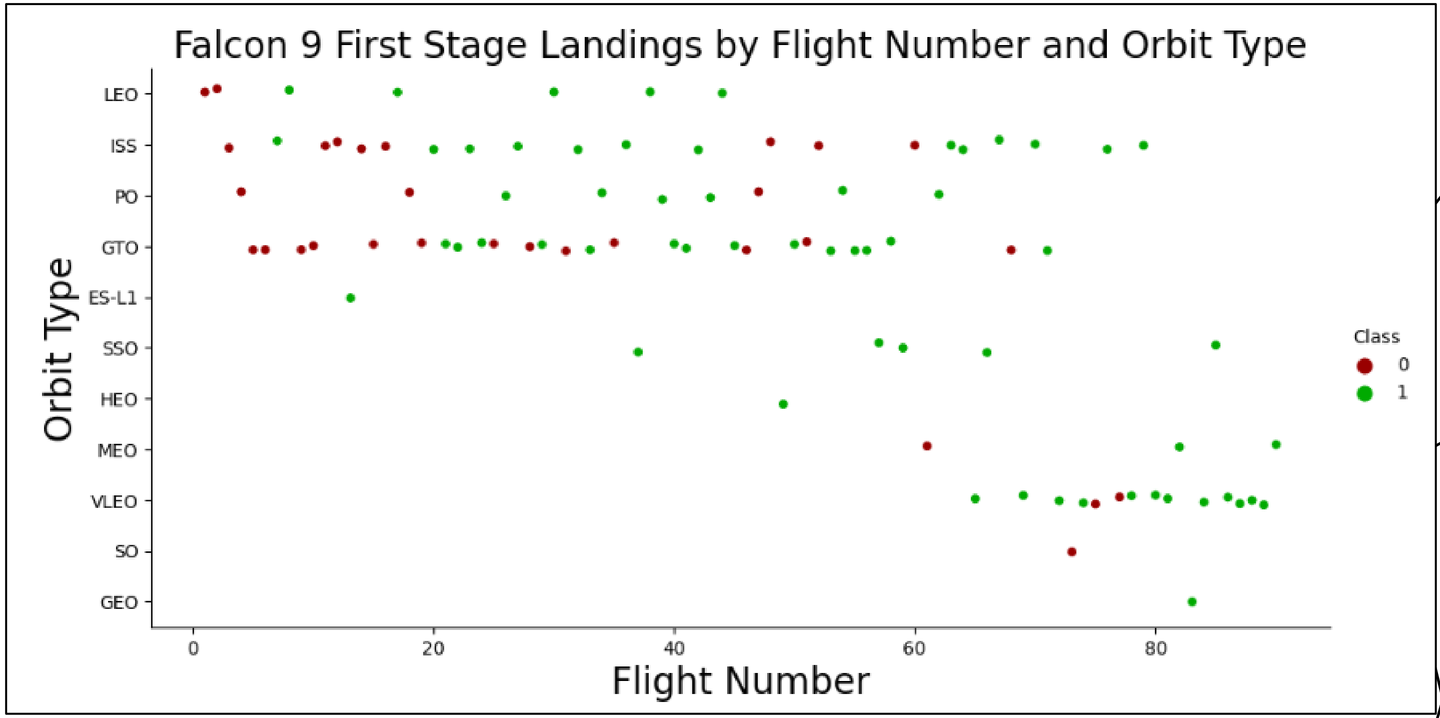
Success Rate vs. Orbit Type

- 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

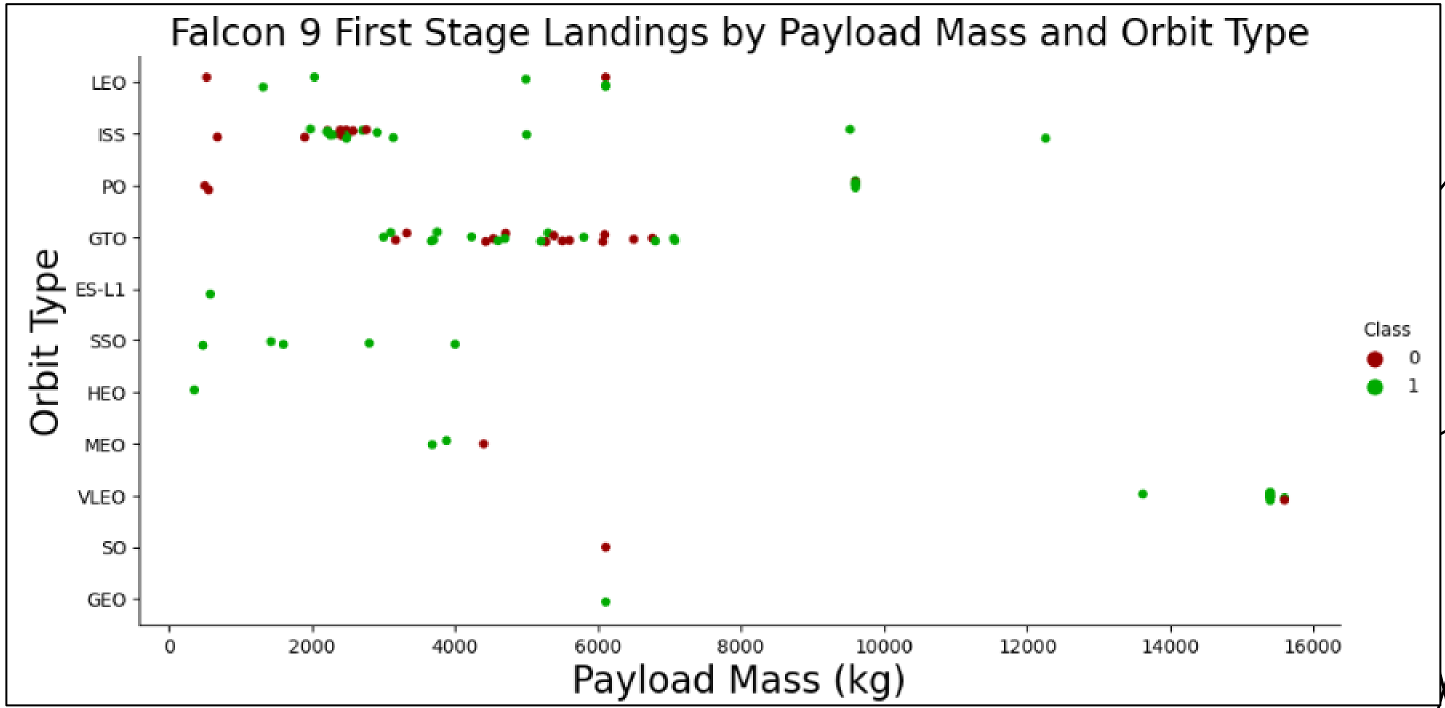
- There is a correlation between flight number and success rate with larger flight numbers being associated with higher success rates



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. Orbit Type

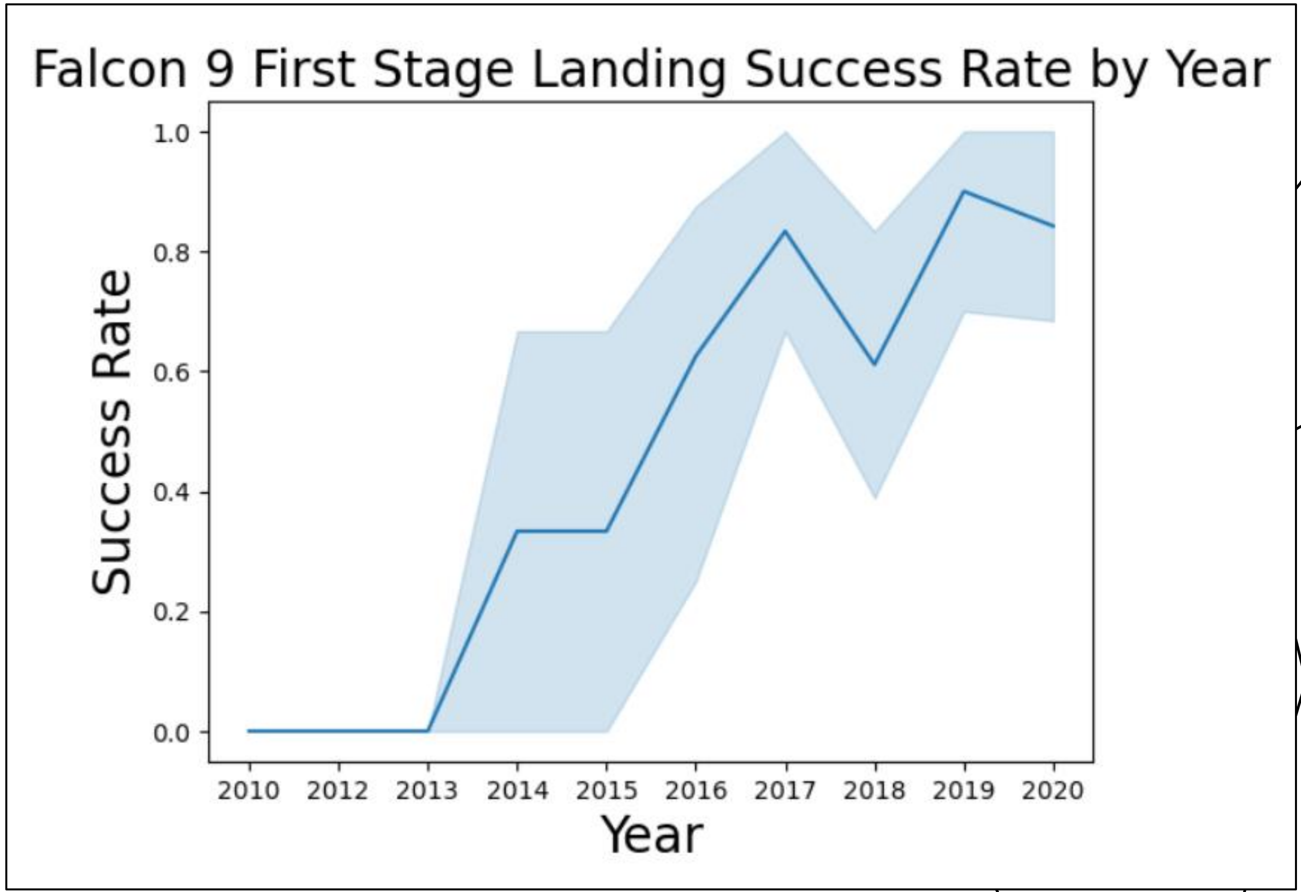
- Some orbit types have better success rates than others.
- Success rate appears to have no obvious correlation with payload mass.



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

Launch Success Yearly Trend

- The trend shows the increasing success rate that appears to be significant 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 Begins 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 sampling mechanism is used to gather pattern of the data from the database table. Launch Site Names That Begin with 'CCA' SELECT * FROM SPACEXDATASET WHERE launch_site LIKE 'CCA%' LIMIT 5;

Total Payload Mass

- Question: What is the total payload mass carried by booster 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__outcome LIKE ‘Success (ground pad)';
- Result:

First Successful Landing Outcome Date
2015-12-22

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

Successful Drone 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

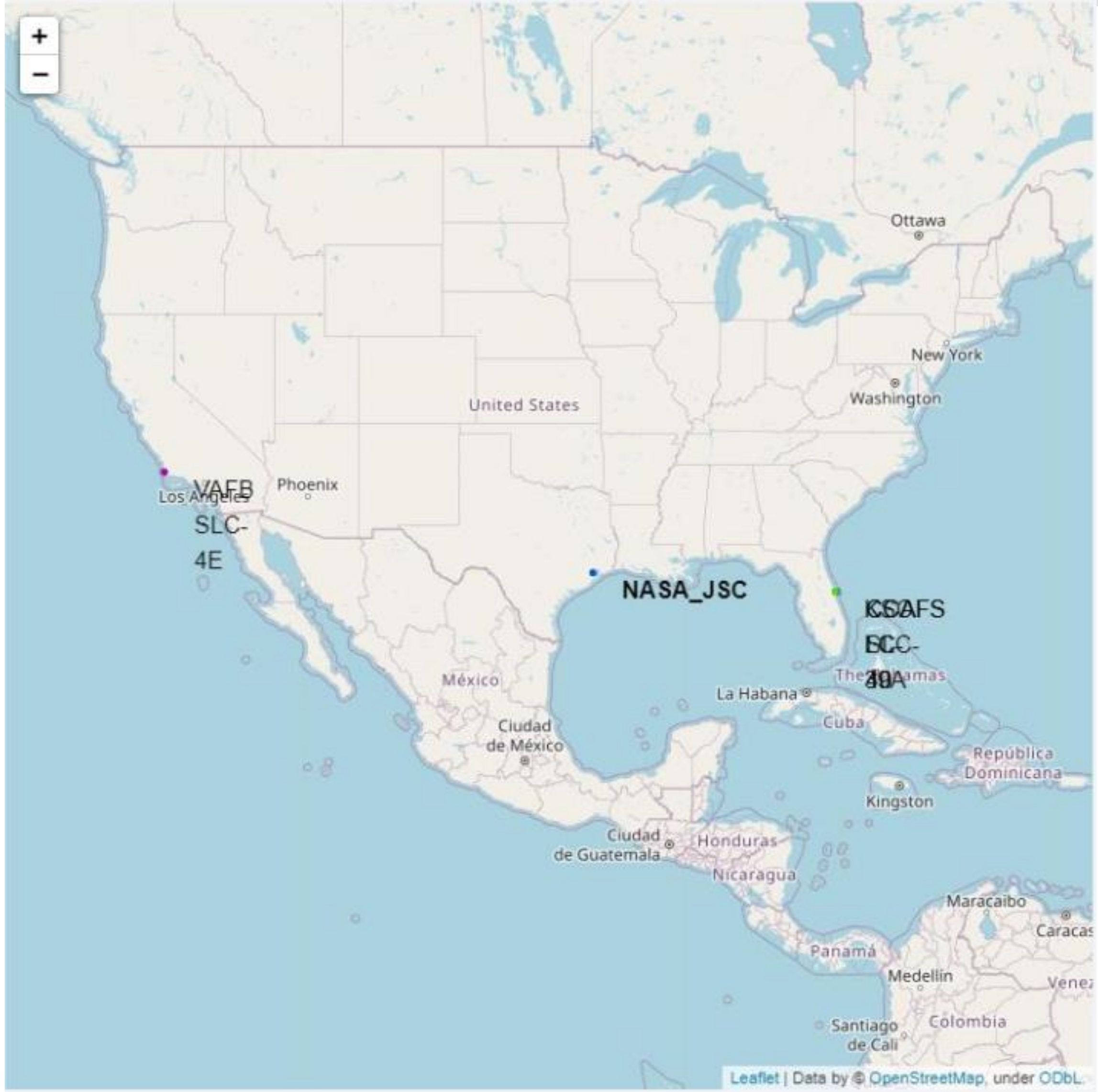
- Explanation: The above list shows the outcomes in descending order. The most common landing outcome was 'not attempted'.

Section 3

Launch sites proximity analysis

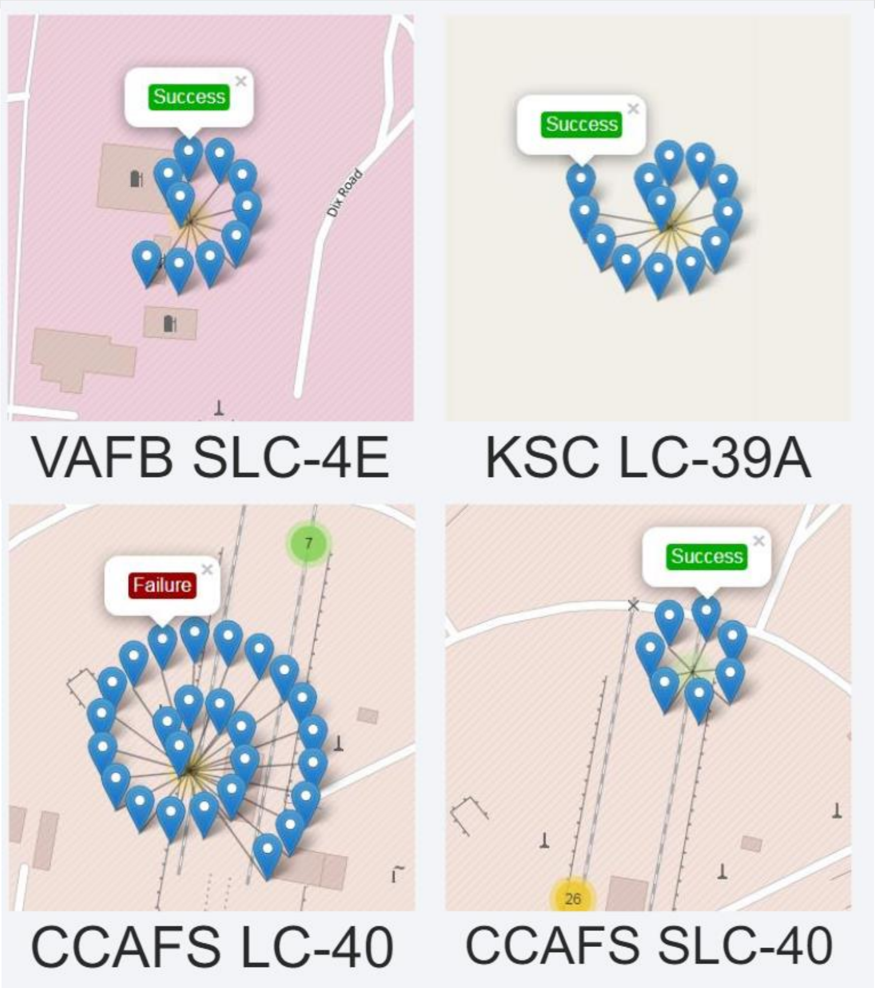
Falcon 9 Launch Site Locations

- **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



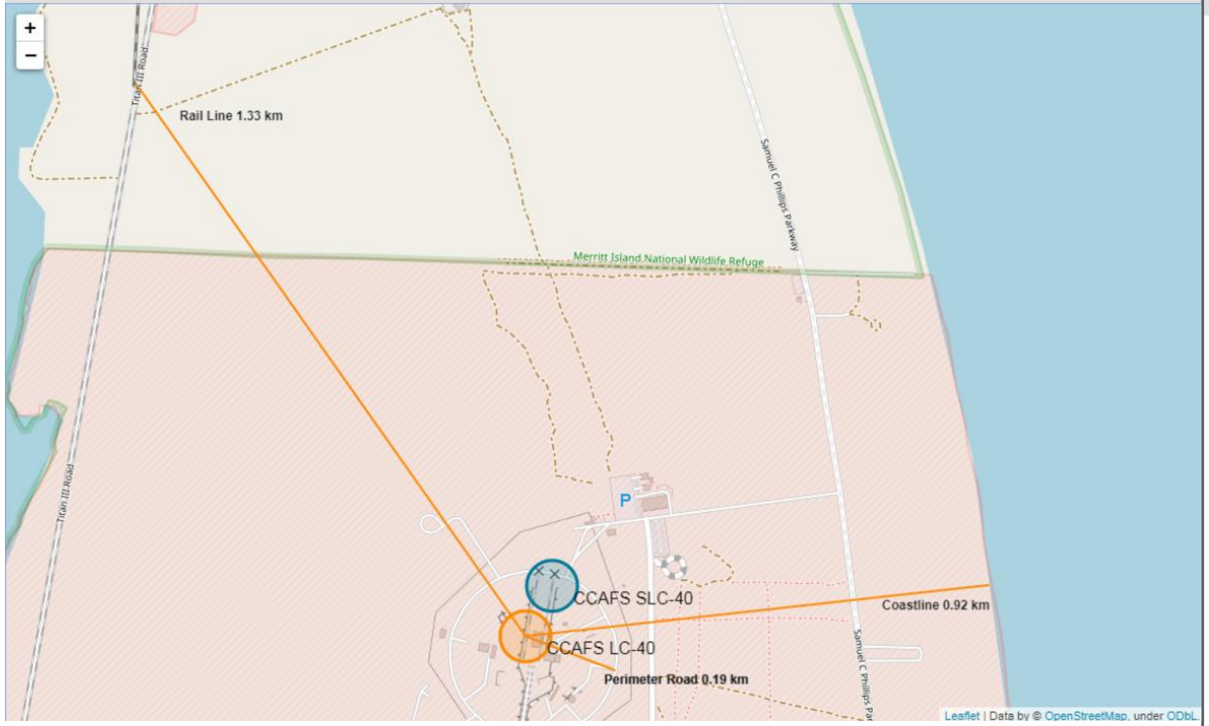
Success/ Failure Landings Markers

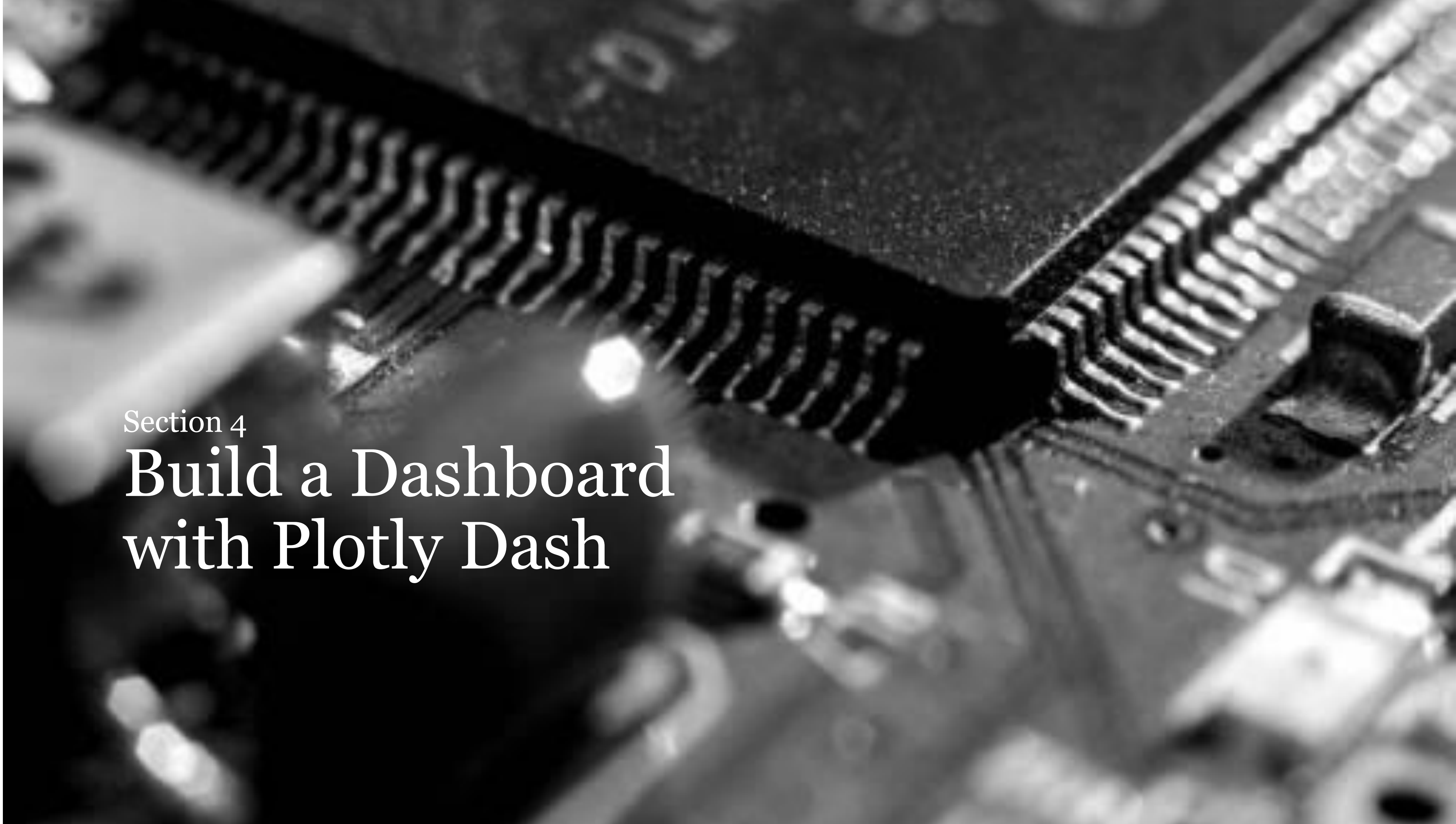
- The markers display the mission outcomes (Success/Failure) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site.
- Launch site's success rate for Falcon 9 first stage landings is derived from the relative number of green success markers to red failure markers.



Distance of Launch Sites 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.



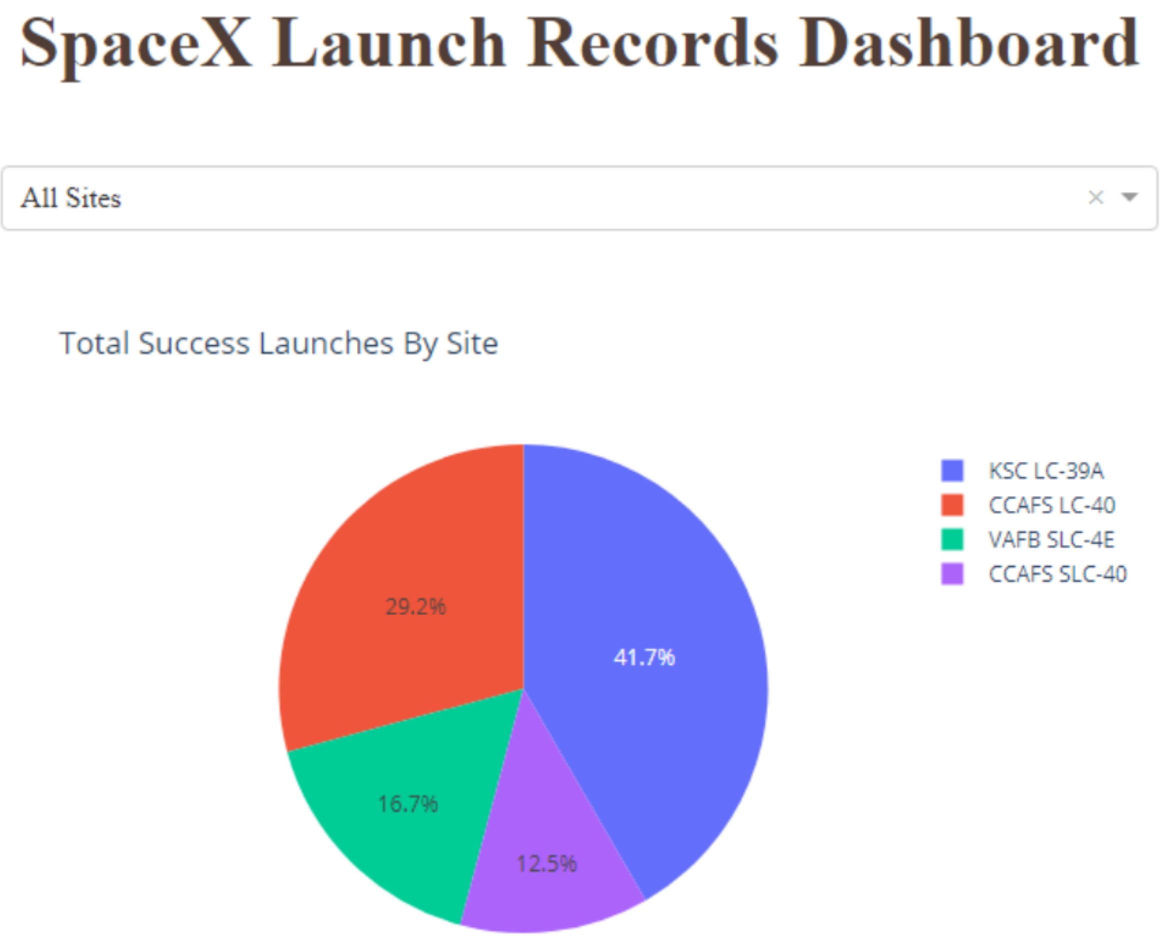


Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

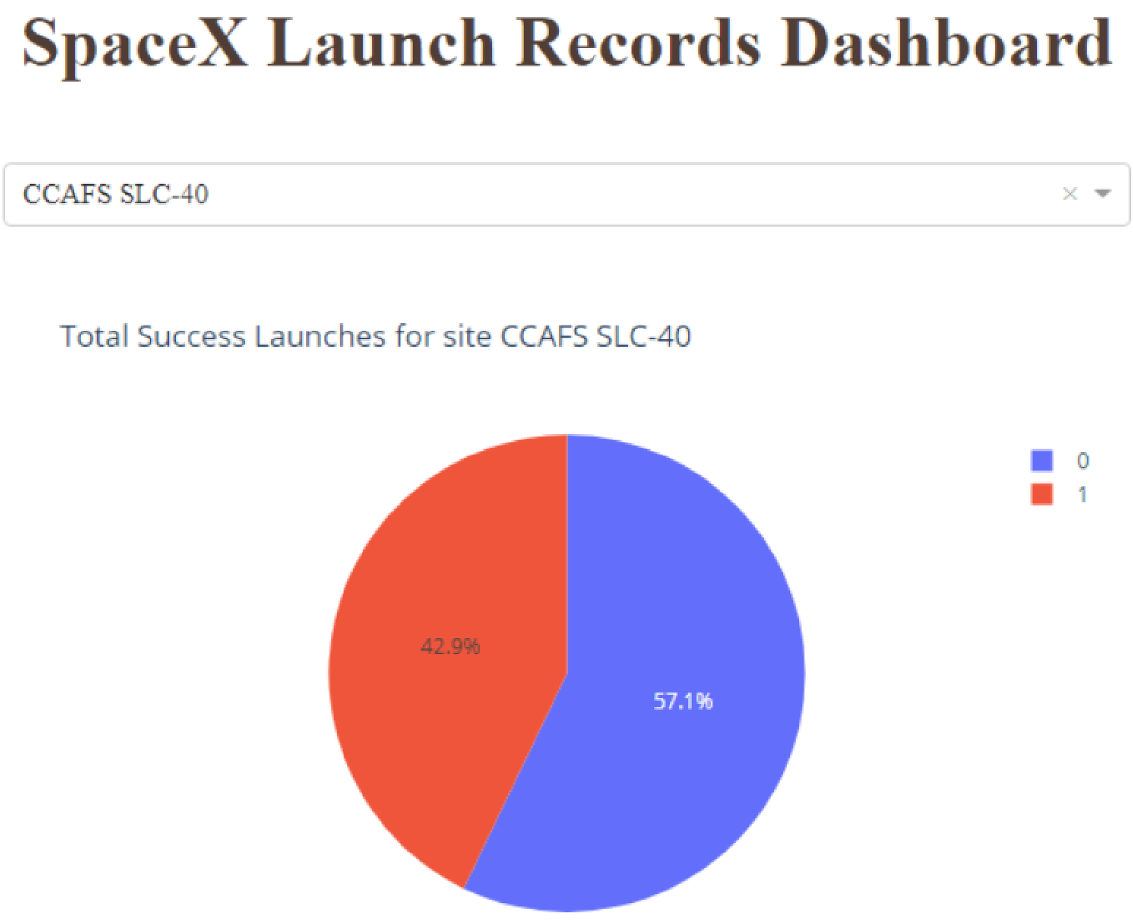
- The dropdown menu allows the selection of one or all launch sites.
- With all launch sites selected, the pie chart displays the distribution of successful Falcon 9 first stage landing outcomes between the different launch sites.
- The greatest share of successful Falcon 9 first stage landing outcomes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



Launch Site with the 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%**).

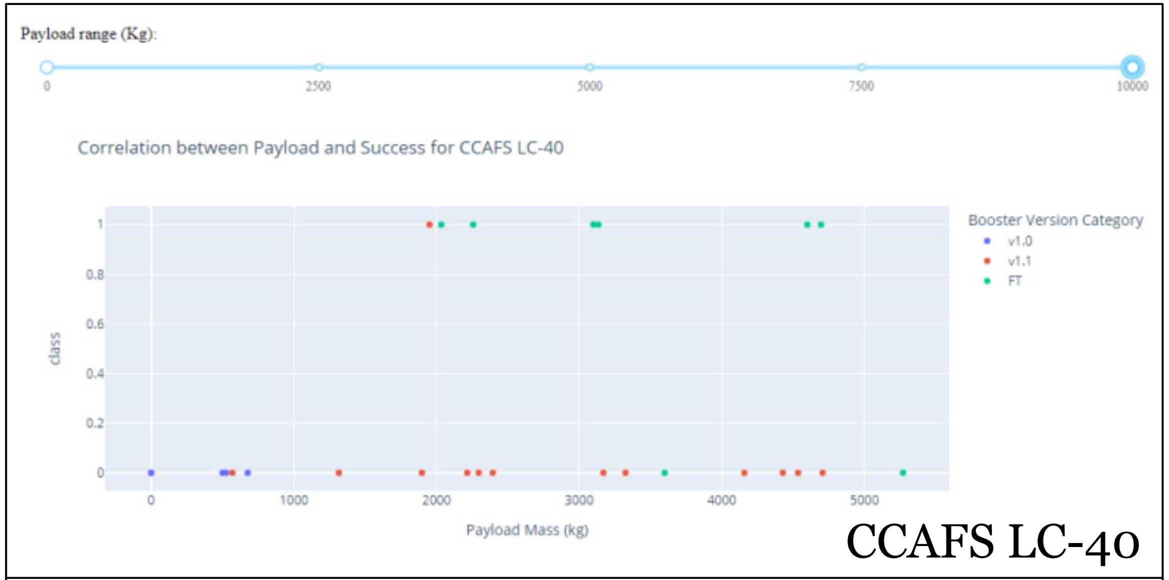


Payload vs. Launch Outcome

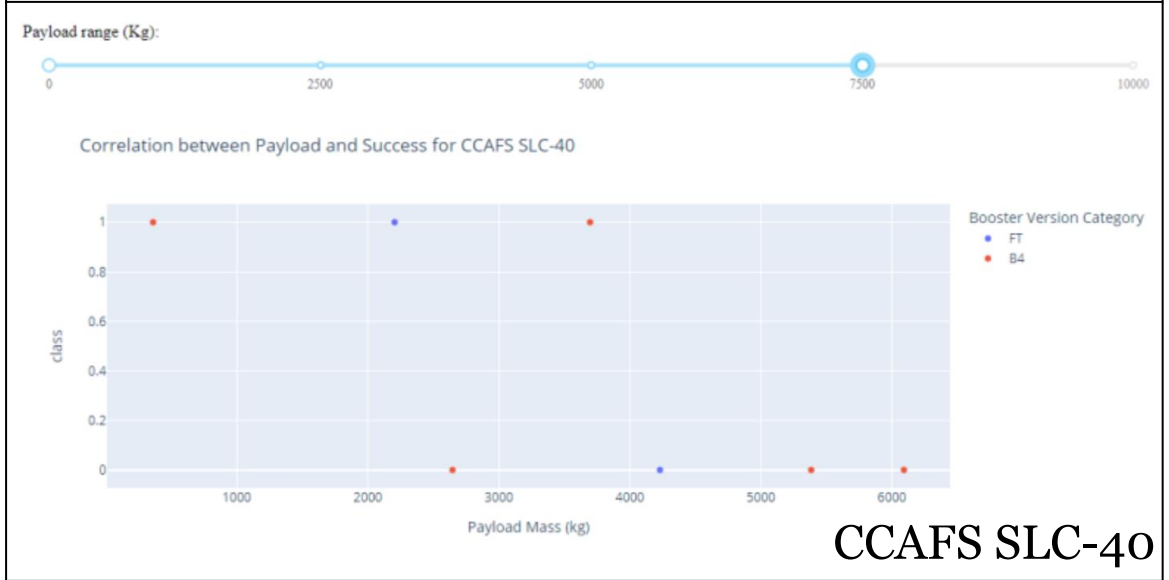
- These screenshots are of the Payload vs. Launch Outcome scatter plots for all sites, with different payload selected in the range slider.

- 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



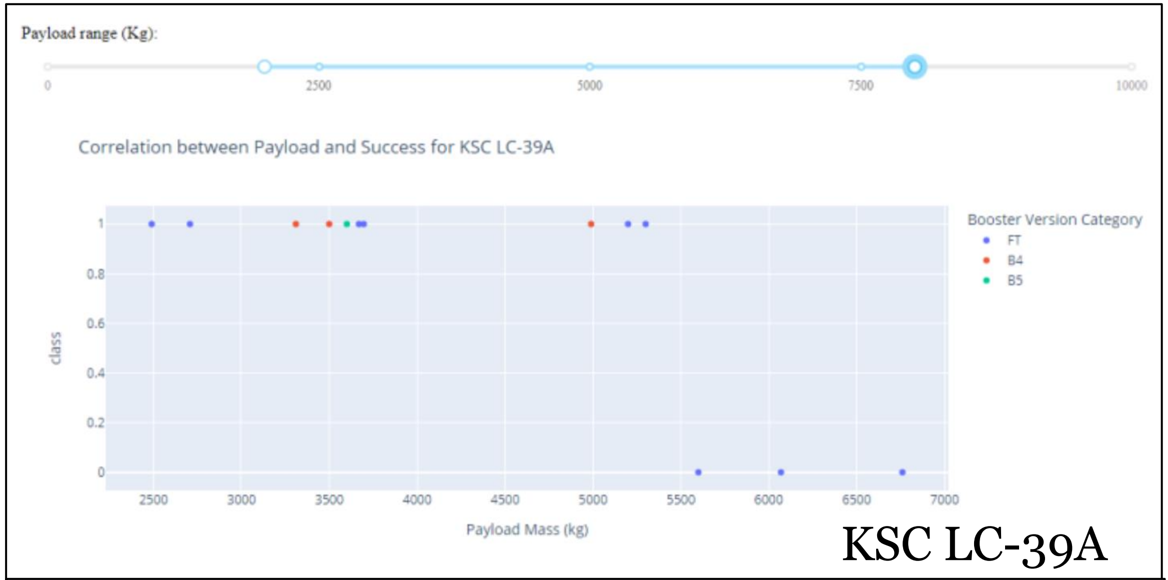
CCAFS SLC-40

Payload vs. Launch Outcome

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

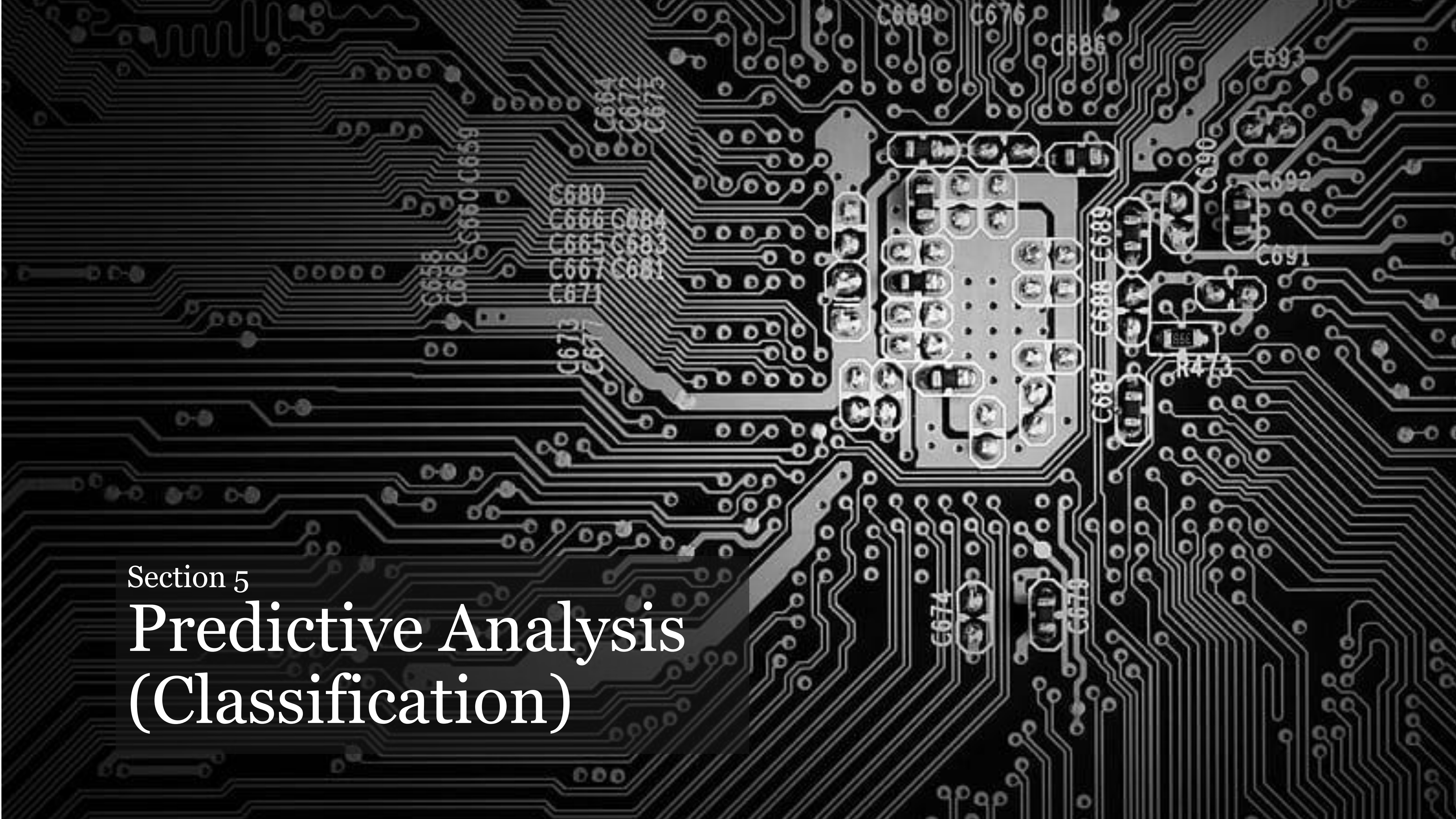
- The payload range from about 2,000 kg to 5,000 kg has the largest success rate.

- The 'FT' booster version category has the largest success rate.



KSC LC-39A

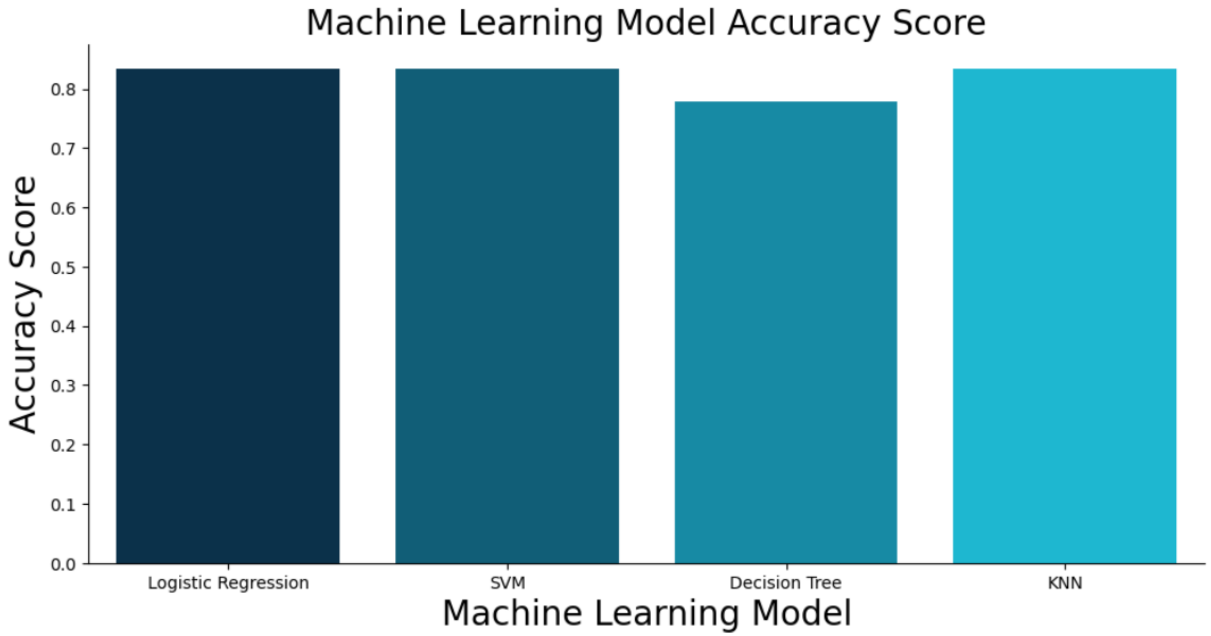
VAFB SLC-4E



Section 5

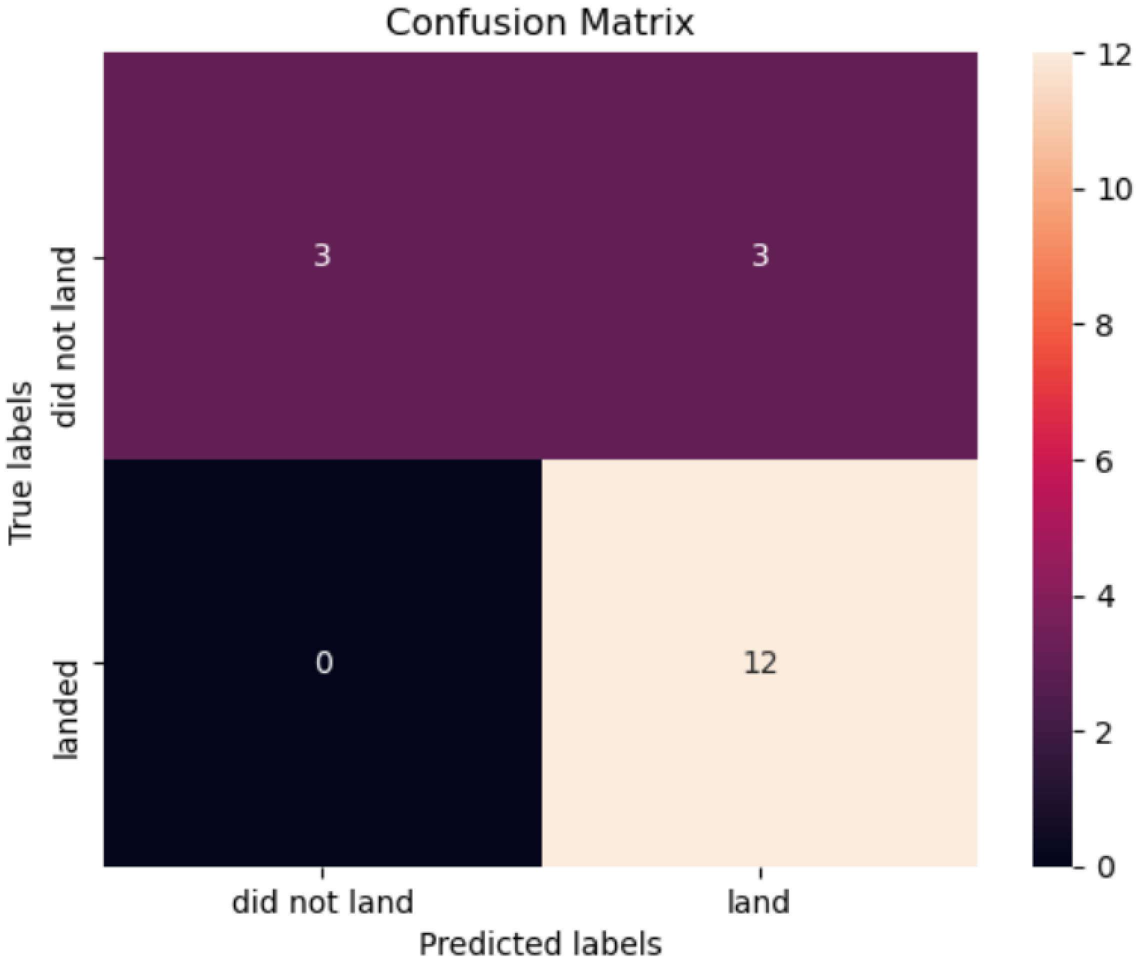
Predictive Analysis (Classification)

Classification Accuracy



- 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.
- 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

Conclusions

- SpaceX has shown significant improvement of Falcon 9 first stage landing outcome over the years. This goes along with frequent launches, as the saying goes “practice makes perfect”.

- Falcon 9 has never had a perfect track record of performing stage landing outcomes.

- Machine learning models are very useful when it concerns first stage landing. It can be further explored to predict future SpaceX Falcon 9 or any rockets’ first stage landing outcomes.

Appendix

Initial Data Sets

- Wikipedia (Webpage): https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- SpaceX API (JSON): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- 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
- https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%203(B)Plotly%20Dash.py

Appendix

Jupyter Notebooks and Dashboard Python File

- GitHub URL (Data Collection): https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%201(A)%20Data%20collection%20API%20checkpoint.ipynb
- GitHub URL (Web Scraping): https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%201(B)%20Data%20Collection%20with%20Web%20Scraping%20lab.ipynb
- GitHub URL (Data Wrangling): https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%201(C)%20Data%20Wrangling.ipynb
- GitHub URL (EDA with SQL): https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%202(A)%20EDA%20with%20SQL%20(2).ipynb
- GitHub URL (EDA with Data Visualization): https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%202(B)%20EDA%20with%20Visualisation%20lab.pdf
- GitHub URL (Folium Maps): https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%203A%20Interactive%20Visual%20Analytics%20with%20Folium%20lab.pdf
- GitHub URL (Dashboad File): https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%203(B)Plotly%20Dash.py
- GitHub URL (Machine Learning): https://github.com/snajiajamil/SpaceX-Data-Science-Project/blob/main/Week%204%20Machine%20Learning%20Prediction%20lab.pdf


Thank you!