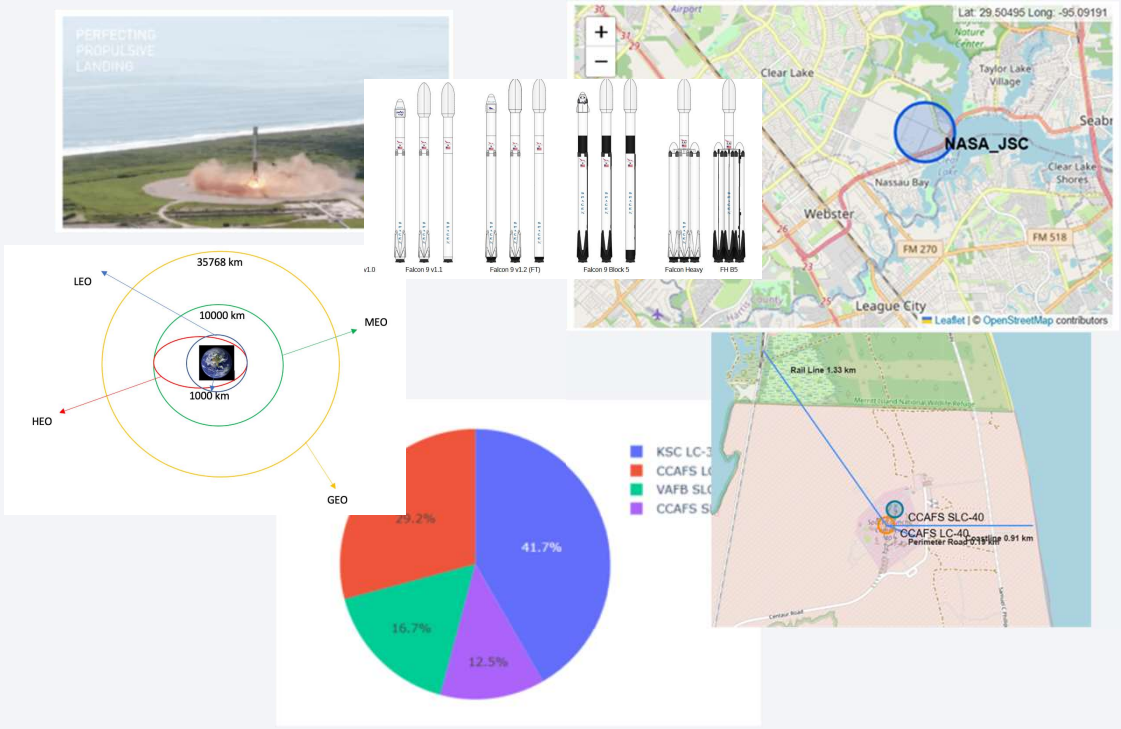
Winning Space Race with Data Science

Kashfia Haque
24/12/24

Outline

- Project Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix



Project Summary

- Methodology Summary
  - Data for the first stage landing of Falcon 9 was collected from 2010 to 2020 from a public API (https://api.spacexdata.com/) and publicly available information in Wikipedia (https://en.wikipedia.org/wiki/SpaceX). Further information was provided in the course.
  - Data cleaning/ wrangling was used to extract data for landing outcome which was used as the dependent variable for the machine learning tools.
  - SQL queries and data visualizations, including static plots, interactive maps and interactive dashboard was used to analyze the data to find different answers.
  - Machine learning models: Logistic Regression, Support Vector Machine (SVM), Decision Tree and k-Nearest Neighbors (KNN) were used for predictive analysis.
- Result Summary
  - EDA Results
  - Interactive Visual analytics and Dashboard
  - Predictive Analysis

Introduction

- SpaceX is a commercial company that offers manned mission for common individuals that is relatively cheaper. They can offer a lower price as they can reuse their first stage unlike other companies.
- A competitor, SpaceY, is trying to determine the price of each launch. If we can determine if the first stage will land, we can determine the cost of a launch. In this project a machine learning model will be trained using public information available about different attributes of SpaceX ability to use the first stage.
- Questions to answer:
  - What kind and how much data is available about the SpaceX Falcon 9 first stage landings publicly?
  - What machine learning tools should be used to predict the best possible outcome for future landing?
  - What attributes effects the successful landing of Falcon 9?

Section 1

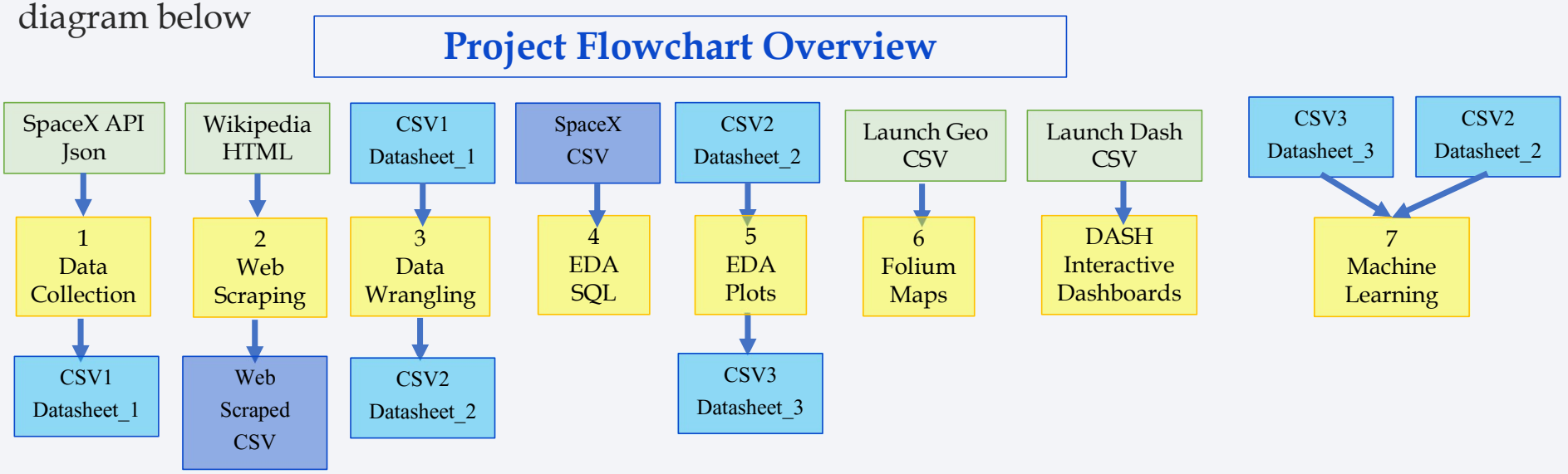
Methodology

Methodology

- Data for the first stage landing of Falcon 9 was collected from 2010 to 2020 from a public API and publicly available information in Wikipedia. Additional Data was provided in CSV format in the course.
- Data cleaning/ wrangling was used to extract data for landing outcome which was used as the dependent variable for the machine learning tools.
- Exploratory Data Analysis (EDA) was performed using SQL queries and data visualizations.
- Folium and Plotty Dash were used for Interactive Data visualization.
- Machine learning models: Logistic Regression, Support Vector Machine (SVM), Decision Tree and k- Nearest Neighbors (KNN) were used for predictive analysis.

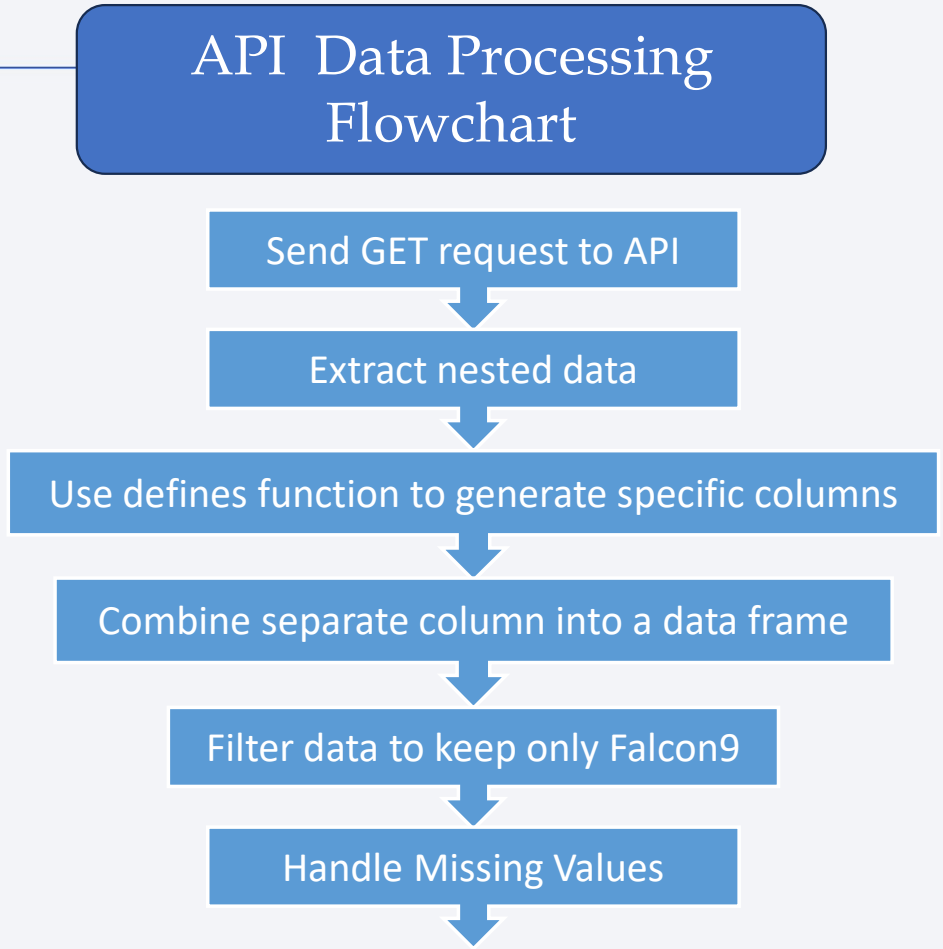
Data Collection

- The data sets were collected from
  - An IBM copy of a response from a publicly accessible API with launch data in JSON format
  - A Wikipedia page with launch data in HTML tables.
  - Any additional datasets were provided with the course in CSV format shown in the blue box in diagram below



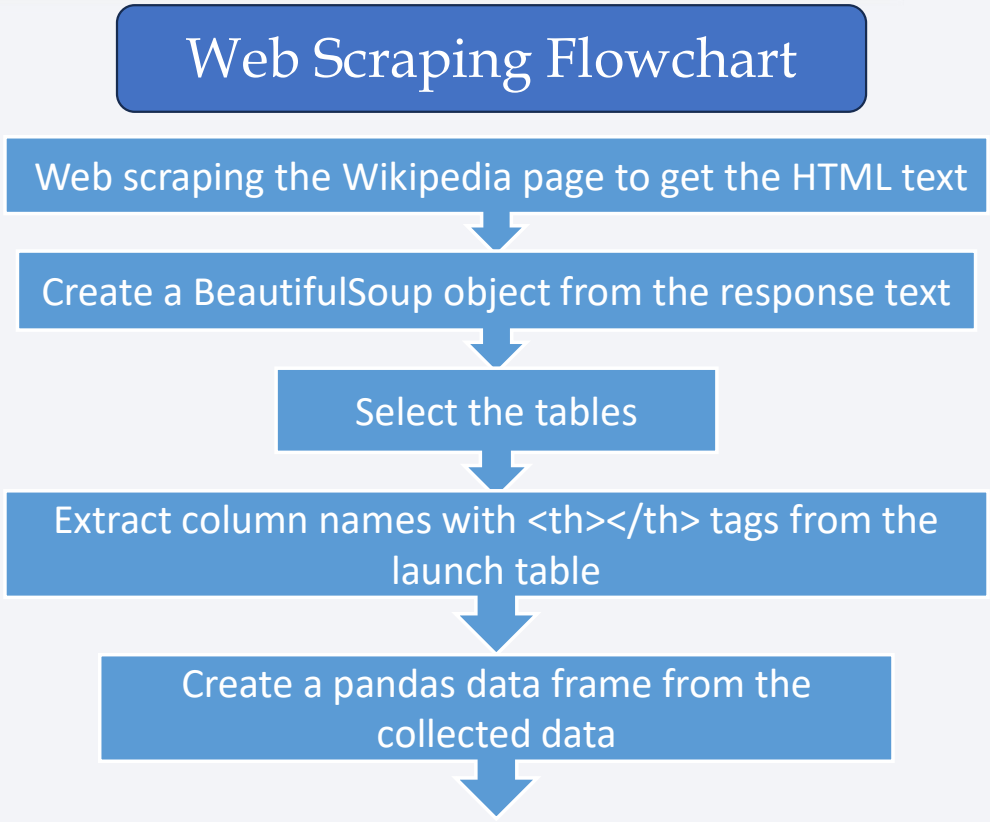
Data Collection – SpaceX API

- SpaceX Data is Publicly at the API: https://api.spacexdata.com/.
- A copy of this API was made available for this project.
- Data was extracted from the API and loaded into a Pandas DataFrame for further analysis.
- https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab1A_collecting%20the%20data.ipynb



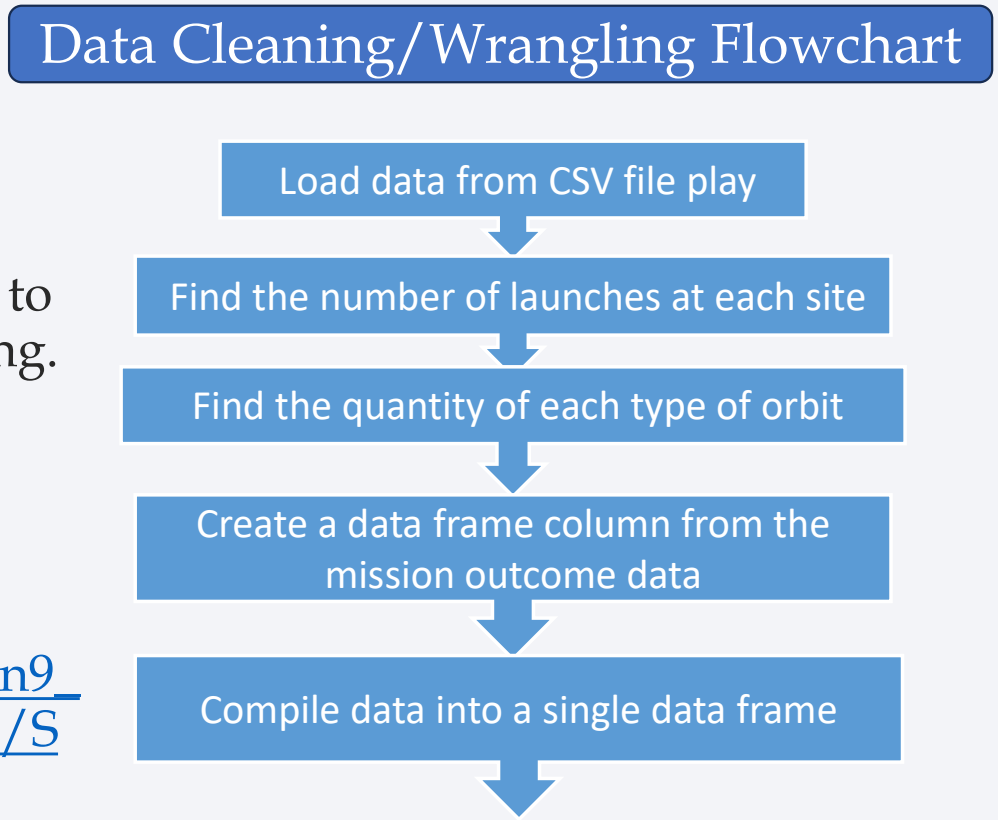
Data Collection – Scraping Wikipedia web page

- SpaceX launch data was scrapped from HTML tables on a permanently linked copy of the SpaceX Wikipedia web page.
- Launch Data was extracted from these tables and loaded into a pandas data frame for further analysis
- https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab1B_webscraping%20from%20wikipedia.ipynb



Data Wrangling

- The CSV file needed cleaning /wrangling.
- The launch sites orbit types and mission outcomes were processed and.
- The mission outcome types were converted to a binary classification using one hot encoding. Falcon 9 first stage landing success was encoded 1 and 0 represented a failure.
- The new mission outcome classification column was added to the DataFrame.
- https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab2%20Data%20wrangling.ipynb



EDA with Data Visualization

- Data analysis was performed using pandas and matplotlib
- Scatter plot was used to see mission outcome relationship split by
  - Launch sight Vs flight number.
  - Launch sight Vs payload mass.
  - Payload mass Vs flight number.
- Bar chart was used to visualize the relationship between success rate of each orbit types
- Line plot was used to display launch success trend by the year.
- https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab3B_Exploring%20and%20preparing%20data.ipynb

EDA with SQL

- SQL queries were performed to extract information about:
  - The names of unique Launch sites are :

    1. CCAFS LC-40 2. CCAFS SLC-40 3. KSC LC-39A 4.VAFB SLC-4E

  - The total payload mass carried by boosters launched by NASA (CRS) is 619967 Kg.
  - The average Payload mass caried by booster version F9v1.1 is 6138 kg.
  - The date when the first successful landing outcome in ground pad was achieved is 2010-06-04.

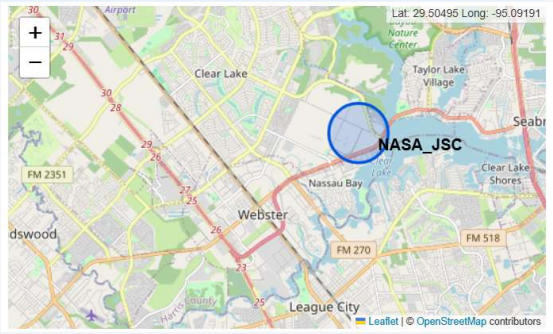
EDA with SQL (continued)

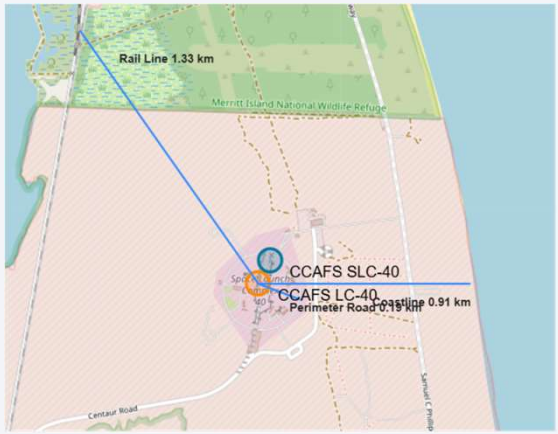
- The following records displays the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_lab3A_EDA%20with%20SQL.ipynb

1	mission_outcome	booster_version	launch_site
1	Success	F9 v1.1 B1012	CCAFS LC-40
2	Success	F9 v1.1 B1013	CCAFS LC-40
3	Success	F9 v1.1 B1014	CCAFS LC-40
4	Success	F9 v1.1 B1015	CCAFS LC-40
4	Success	F9 v1.1 B1016	CCAFS LC-40
6	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

Build an Interactive Map with Folium

- Map objects were created and added to the 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 proximities
    - distance from a long site to the coastline is 0.91 km
    - distance from a long side to the rail line is 1.33 km
    - distance from a launch site to the perimeter Rd is 0.19 km
- https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab4A_Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb





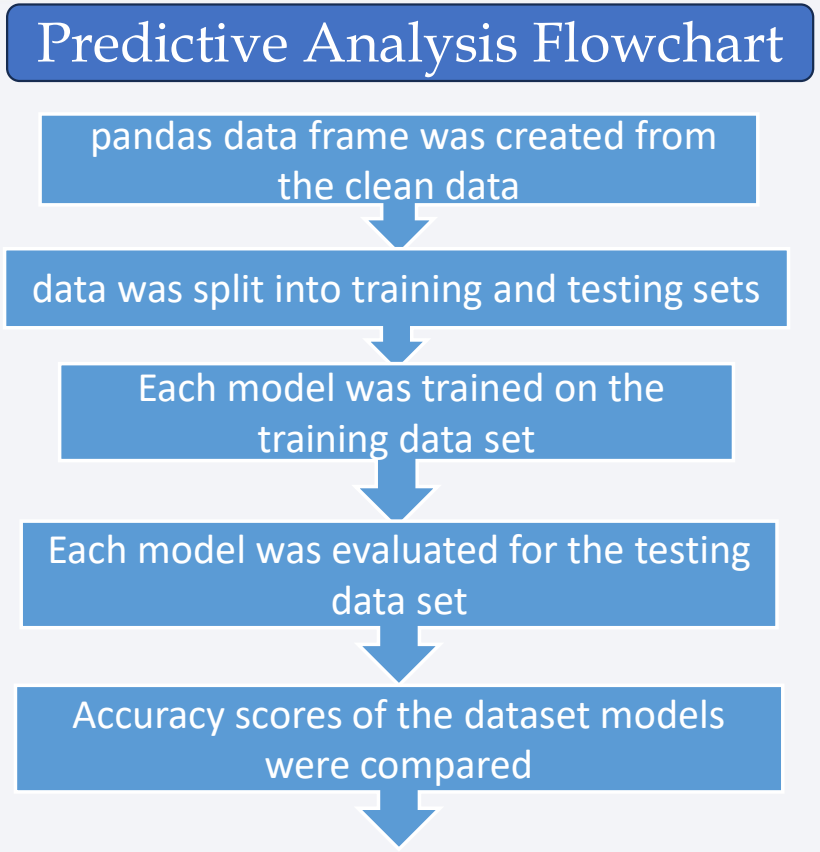
Build a Dashboard with Plotty Dash

- The Plotty dashboard included a dropdown input that can select data from “one” to “all” launch sites to be displayed on the pie chart or the scatterplot.
- For single launch site, the pie chart displayed the distribution of successful Falcon 9 first stage landing between the sites slide.
- For all launch sites the pie chart displayed the distribution of successful Falcon nine first stage landing between the sites
- The input slider is used to filter the payload masses for the scatter plot
- The scatter plot displayed the distribution of Falcon nine first stage landing split by payload mass mission outcome and by booster version category
- https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab4B_Interactive%20plotty%20dash.py

Predictive Analysis (Classification)

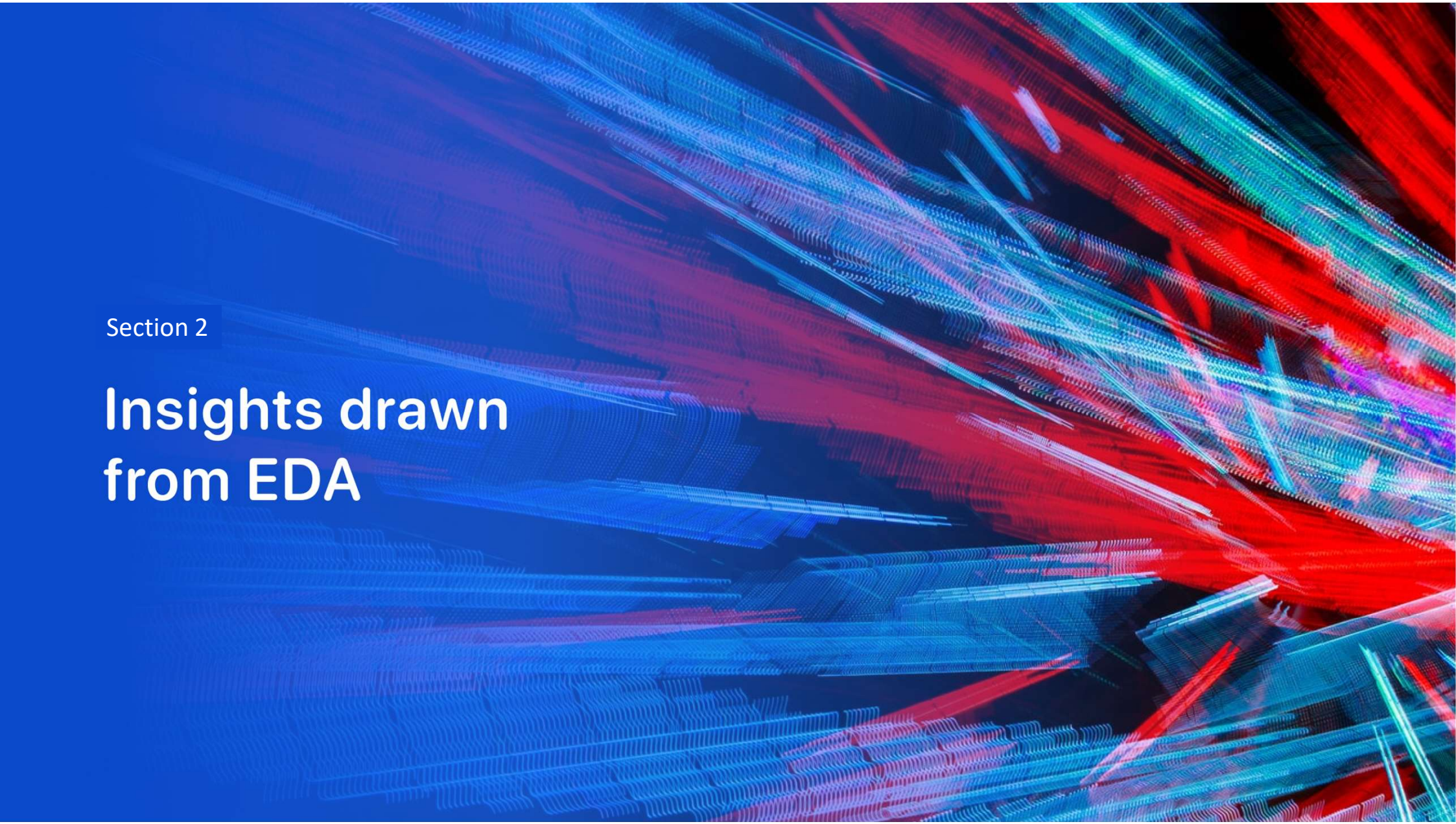
- The data set was split into training and testing sets.
- The following machine learning models were trained on the training data set:
  - logistic regression
  - SVM (support vector machine)
  - decision tree
  - KNN (k-Nearest Neighbors)
- Hyper-parameters were evaluated using GridSearchCV()and the best was selected using the best_ params method.
- Each of the four models were scored on accuracy by using hyperparameters comparing the testing datasets.

- https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab5_Machine%20Learning%20Prediction_Part_5.ipynb



Results

- Exploratory data analysis results
  - Data visualization
  - SQL queries
- Interactive analytics demo in screenshots
  - Interactive folium maps
- Will the dashboard with plotty dash
  - interactive plotting dash dashboard
- Predictive analysis results
  - predictive analysis using machine learning


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Success rate varied noticeably for different launch sites
- Falcon 9 first stage landing success appear to become more dominant as the flight number increases;

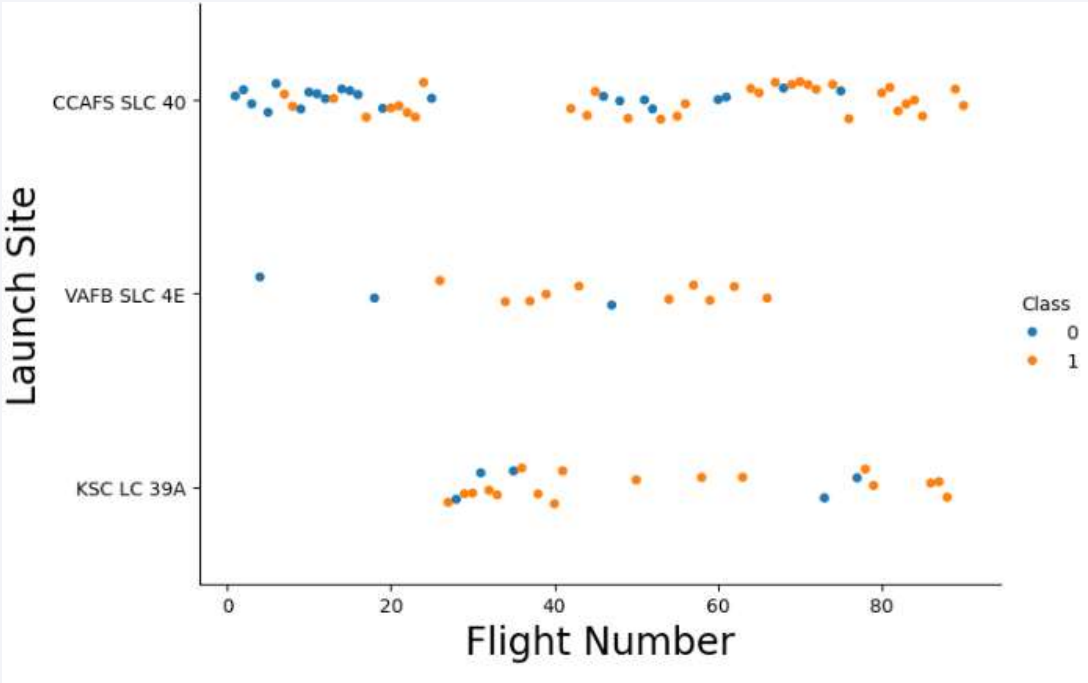


Fig: Scatter plot of Flight Number vs. Launch Site. The failed landings are marked blue, and the successful landings are marked orange

Payload vs. Launch Site

- For the CCAFS SLC 40 launch site, the payload mass and the landing outcome the correlations appears to be weak.
- The failed landings at the KSC LC 39A launch site are mostly grouped around a narrow band of payload masses.

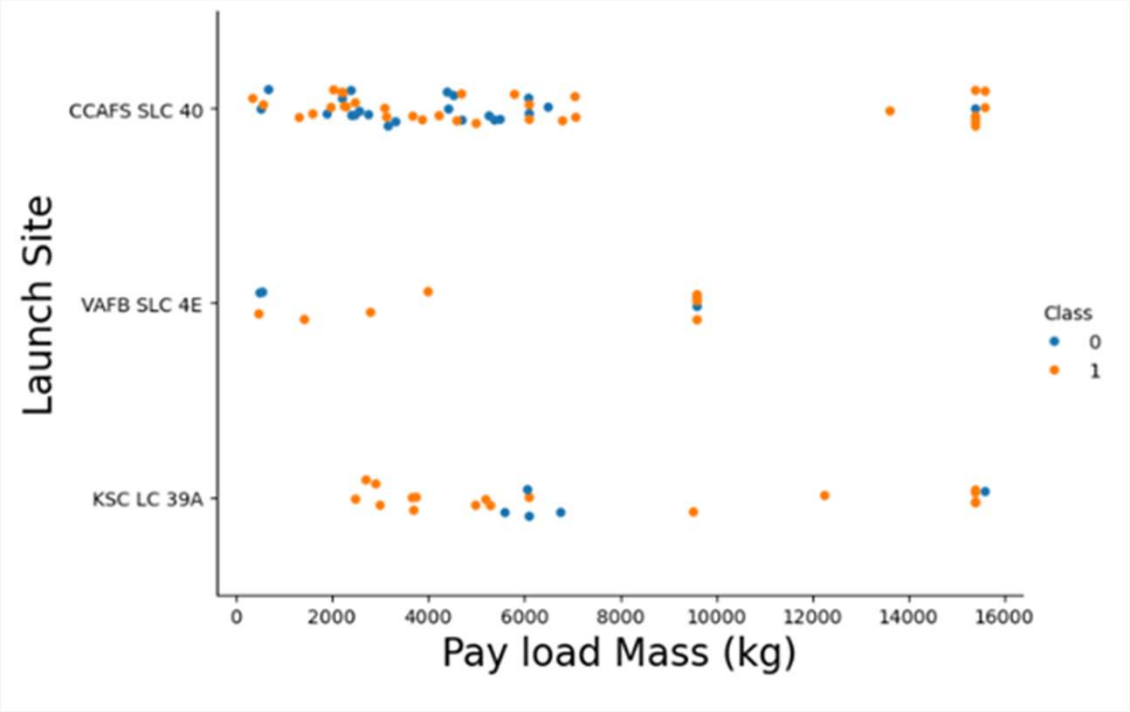


Fig: Scatter plot of Payload vs. Launch Site. The failed landings are marked blue, and the successful landings are marked orange.

Success Rate vs. Orbit Type

- ES-L1, SSO, HEO and GEO orbits have 100 successful first stage landings.
- The landing success rate gradually decreases from VLEO following LEO, MEO, PO, ISS to GTO.
- SO orbit have no successful first stage landing.

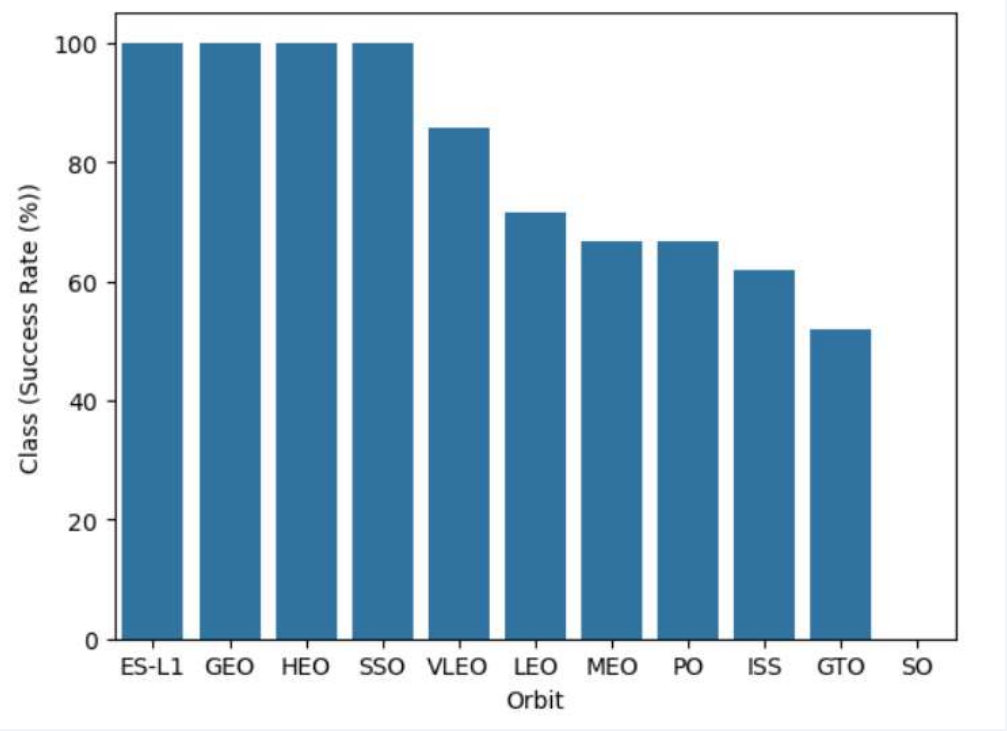


Fig: Bar chart for the success rate of each orbit type.

Flight Number vs. Orbit Type

- Flat number and success rate appears to show a positive correlation.

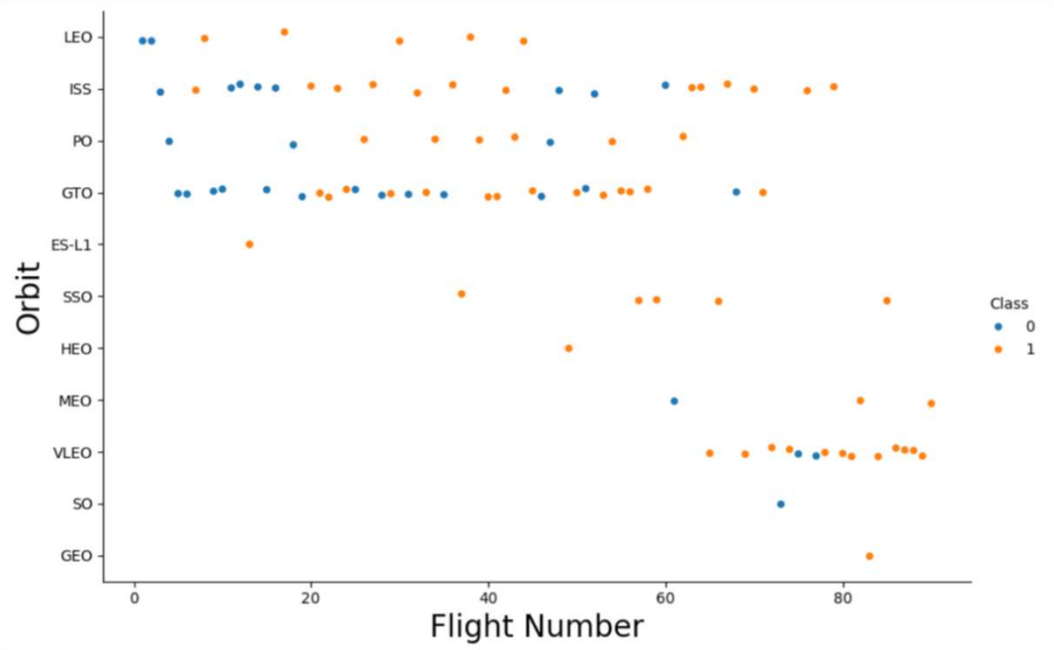


Fig: Scatter plot of Flight number vs. Orbit type. The failed landings are marked blue, and the successful landings are marked orange.

Payload vs. Orbit Type

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

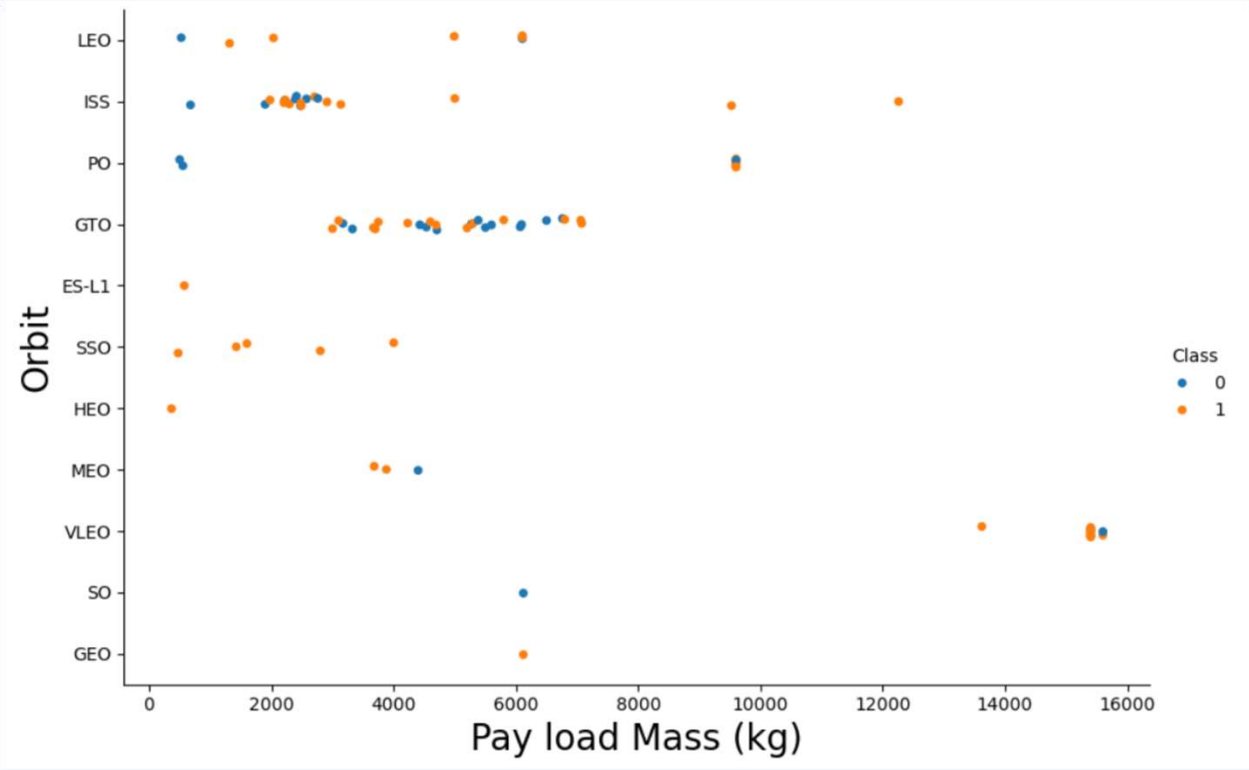


Fig: Scatter plot of payload vs. Orbit type. The failed landings are marked blue, and the successful landings are marked orange.

Launch Success Yearly Trend

- The success rate of the Falcon 9 first stage landings has increased significantly over the interval from 2010 to 2020.

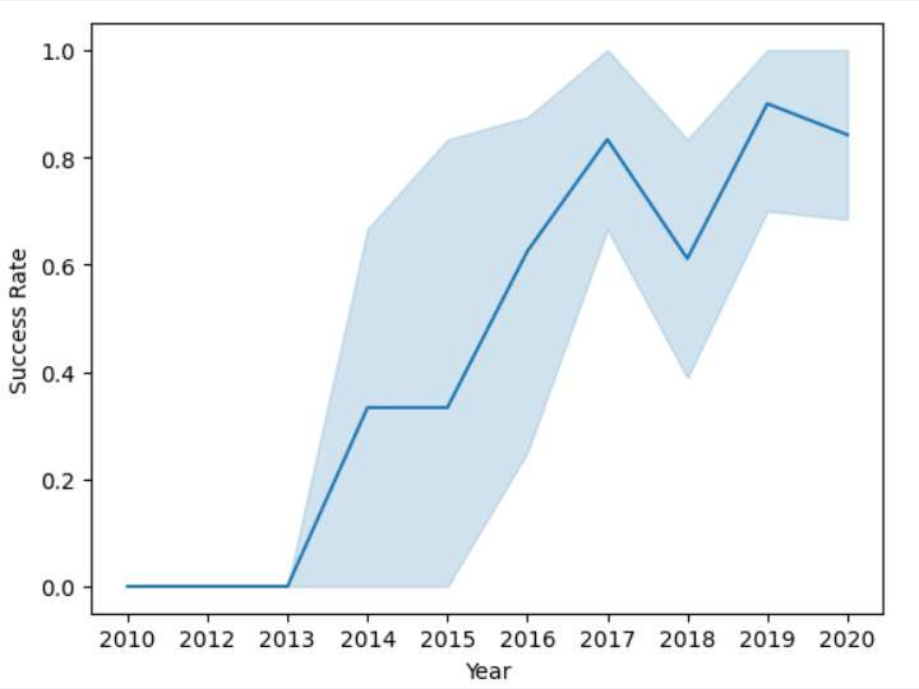


Fig: Line chart of yearly average success rate

All Launch Site Names

- The names of the unique launch sites were found using the following query
- %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX;
- query result is shown here

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

- There are 4 unique launch sites

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` was found using the following query:
- %sql SELECT LAUNCH_SITE from SPACEX where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

- The total payload carried by boosters from NASA was calculated using the following query:
- %sql SELECT sum(payload_mass__kg_) AS "Total Payload Mass (kg)" FROM SPACEX WHERE customer LIKE '%NASA (CRS)%';
- query result:

Total Payload Mass (kg)
48213

- The total payload carried by boosters from NASA is 48213 kg.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was calculated using the following query:
- %sql SELECT sum(payload_mass__kg_) / count(payload_mass__kg_) AS "Average Payload Mass (kg)" FROM SPACEX WHERE booster_version LIKE 'F9 v1.1';
- query result:

Average Payload Mass (kg)
2928

- The average payload mass carried by booster version F9 v1.1 was 2928 kg.

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad was found using the following query:
- %sql SELECT min(DATE) AS "First Successful Landing Outcome Date" FROM SPACEX WHERE landing_outcome LIKE 'Success (ground pad)';
- The result :

First Successful Landing Outcome Date
2015-12-22

- The first successful landing date was 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 was found with the query:
- %sql SELECT DISTINCT booster_version FROM SPACEX WHERE landing_outcome = 'Success (drone ship)' and payload_mass__kg_ BETWEEN 4000 and 6000;
- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1021.2, F9 FT B1031.2, F9 FT B1022 and F9 FT B1026.

- Answer:

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes were calculated using the following query
- %sql SELECT (SELECT count(*) FROM SPACEX WHERE lcase(landing_outcome) LIKE '%success%') AS "Success", count(*) AS "Failure" FROM SPACEX WHERE lcase(landing_outcome) NOT LIKE '%success%';
- Answer:

Success	Failure
61	40

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass was listed.
- Query: %sql SELECT max(payload_mass__kg_) AS "Max Payload Mass (kg)" FROM SPACEX;
- Answer:

Max Payload Mass (kg)
15600

- The maximum pay load mass is 15600 kg

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 was listed using following query.
- %sql SELECT DATE, YEAR(DATE) AS "Year", MONTH(DATE) AS "Month", DAY(DATE) AS "Day", DAYOFWEEK(DATE) AS "day_of_week" FROM SPACEX LIMIT 5;

Answer:

DATE	Year	Month	Day	day_of_week
2010-06-04	2010	6	4	6
2010-12-08	2010	12	8	4
2012-05-22	2012	5	22	3
2012-10-08	2012	10	8	2
2013-03-01	2013	3	1	6

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

- 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 was ranked using the query:
- %sql SELECT landing_outcome, count(landing_outcome) AS "Count" FROM SPACEX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing_outcome ORDER BY count(landing_outcome) DESC;

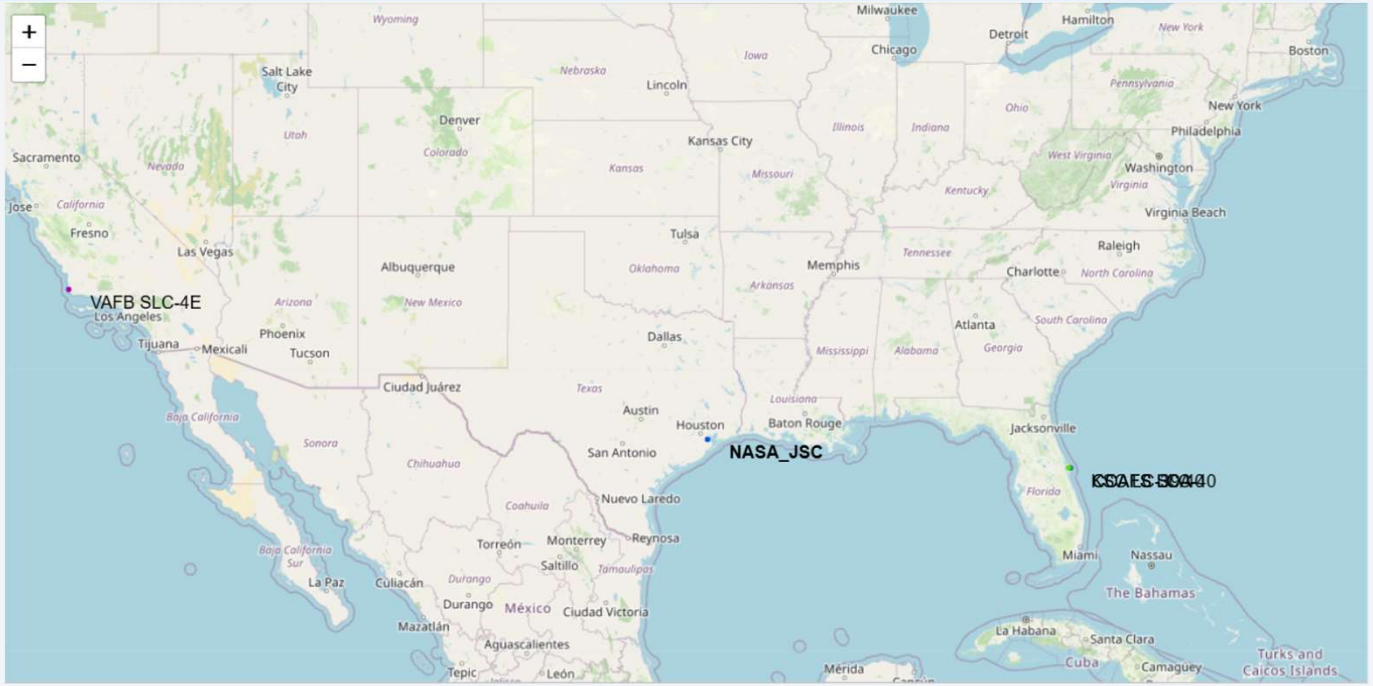
Answer

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


Section 3
Launch Sites
Proximities Analysis

Falcon 9 launch site locations

- California, USA
  - VFP SLC- 4E
- Florida, USA
  - KSC LC-39A
  - CCAFS LC- 40
  - CCAFS LC- 40



- Fig: Folium map view of the Falcon 9 Launch site locations

Map markers with landing outcome

- The markers display the mission outcomes For Falcon nine first stage landings. They are grouped on the map associated with their geographical coordinates for the launch site.

- Fig a shows the map with markers with the number of landings.
- Fig b and c is a zoom in of CCAFS SLC-40 and VAFB SLC 4E showing how different markers can be selected to see failure or success.

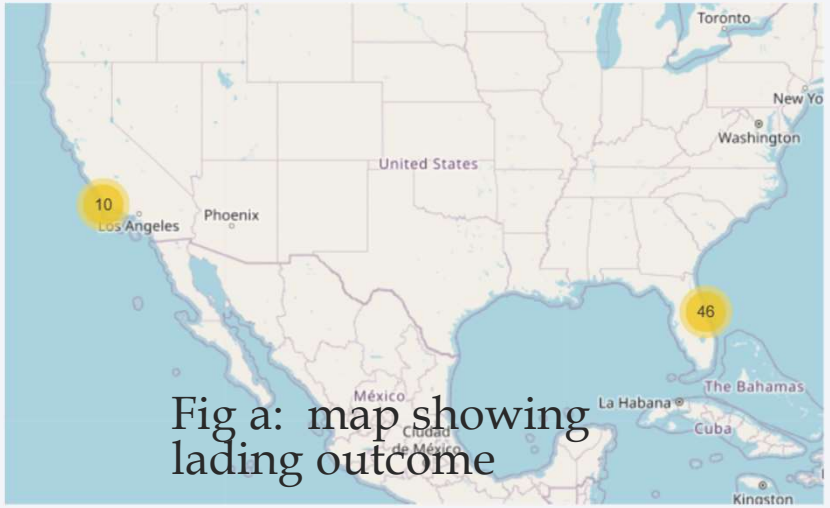


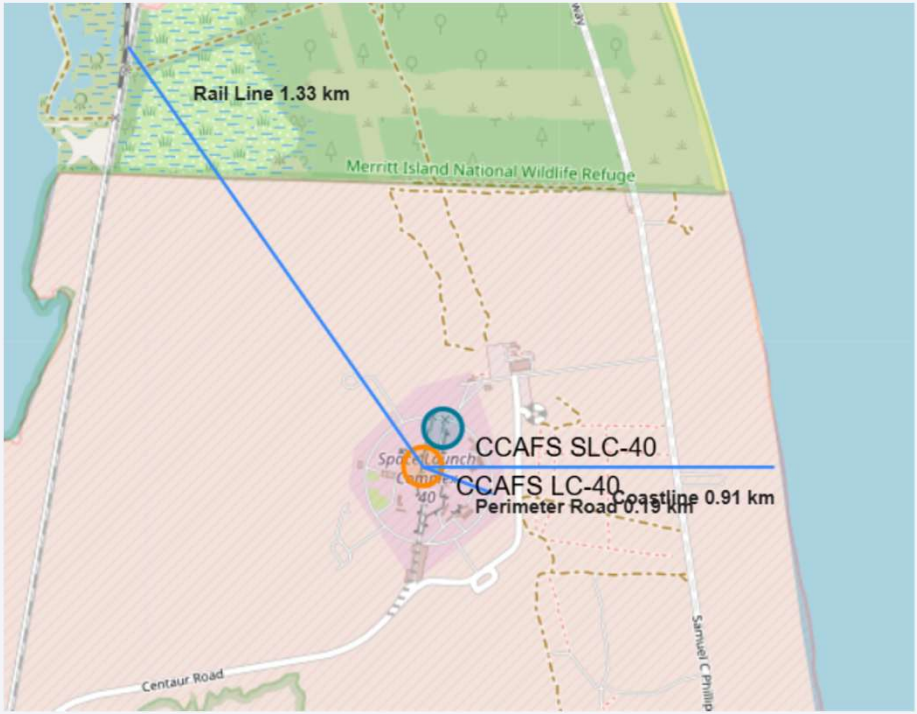
Fig a: map showing lading outcome

Fig b: CCAFS SLC-40

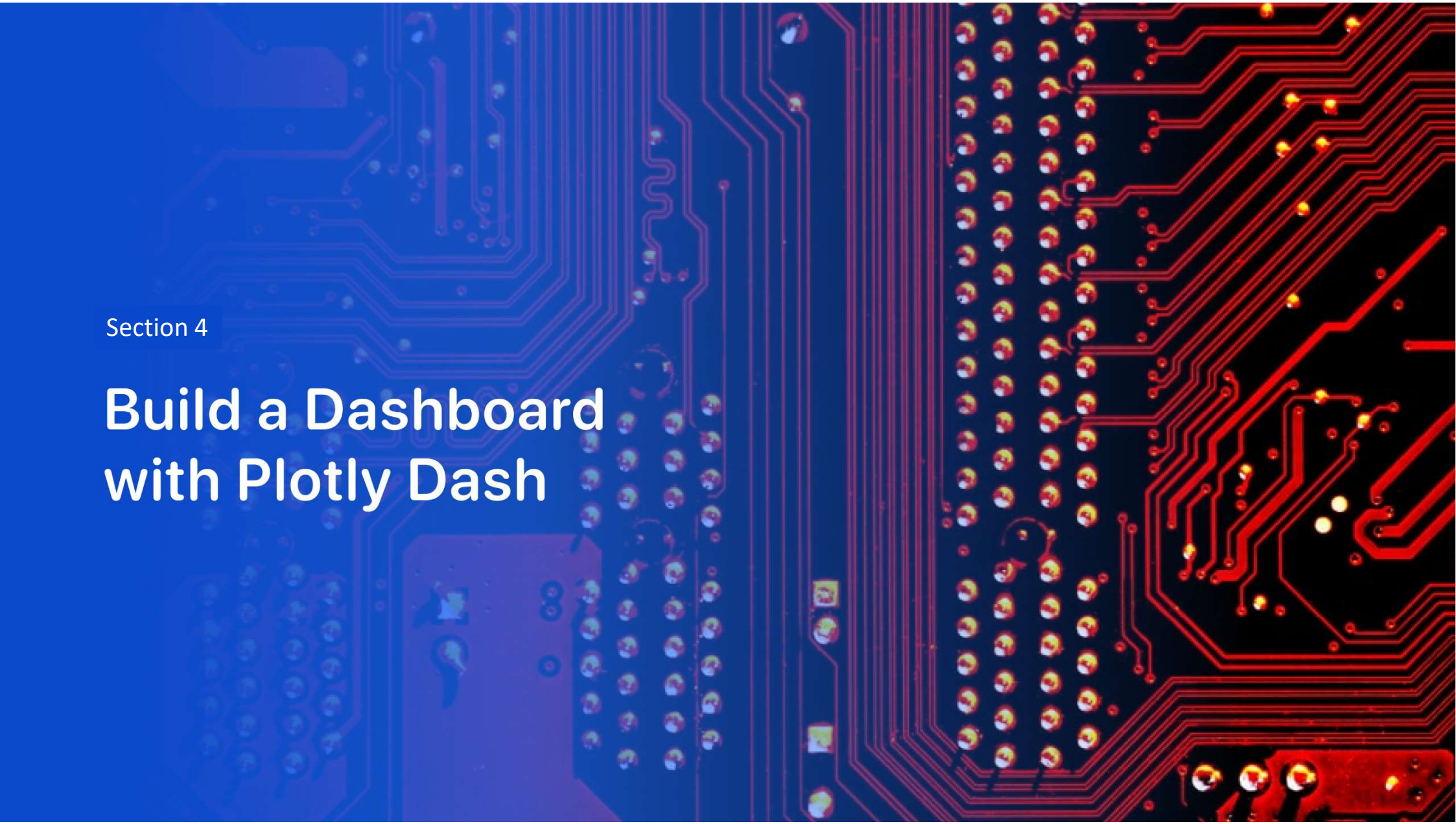
Fig c: VAFB SLC 4E

Distance from Launch site to proximities

- The CCAFS SLC 40 and CCAFS LC 40 launch sites have coordinates that are close to each other
- The perimeter road is 0.19 kilometer away from the launch site CCAFS LC 40.
- The coastline is 0.92 kilometer away from CCAFS LC 40.
- The rail li
- CCAFS LC 40 ne is 1.33 kilometer away from

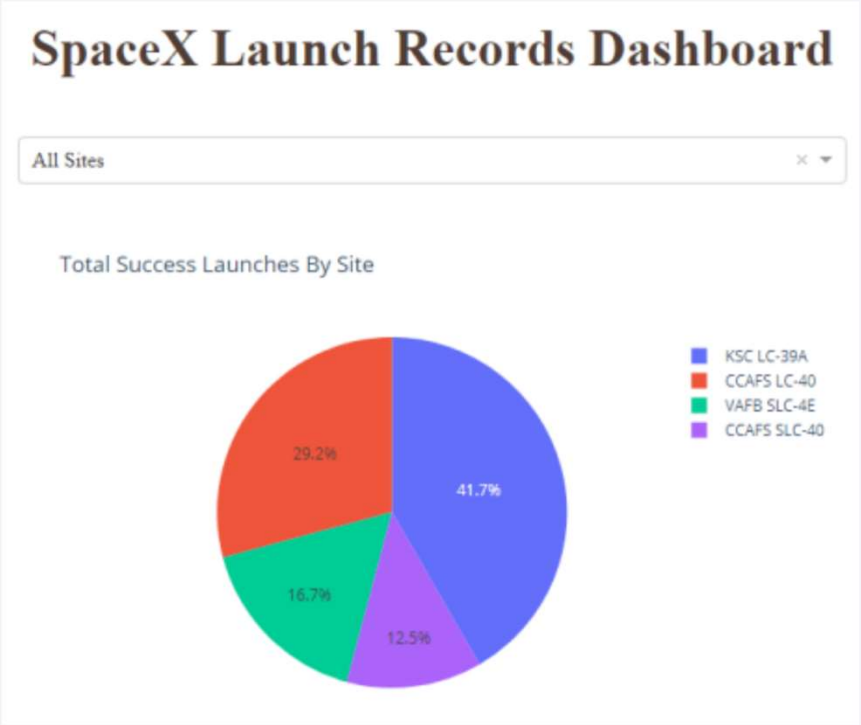


- Fig: Proximity distances from the launch site CCAFS LC 40


Section 4
Build a Dashboard with Plotly Dash

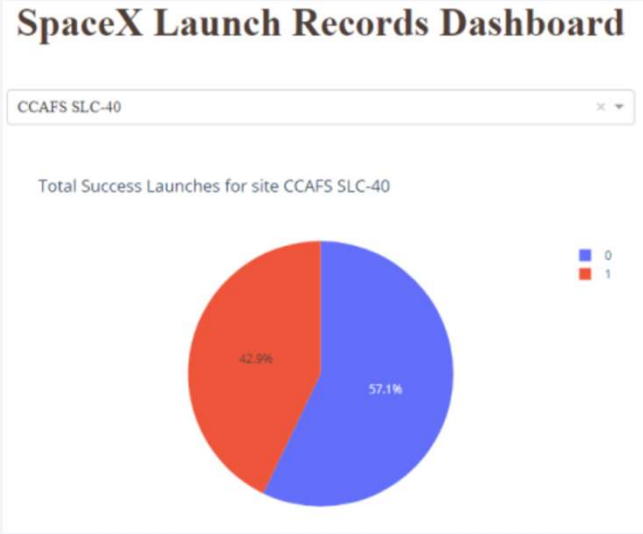
Launch success data

- The dropdown menu allows to select one of the launch sites.
- When all launch sites are selected, the pie shows the distribution of successful Falcon 9 first stage landing outcomes between different launch sites.
- KSC LC-39A has the greatest success of first stage Falcon 9 landing outcomes (41.7%).



Launch site with highest success rate

- CCAFS SLC-40 is the launch site with the highest success rate with a success rate of 42.9%.
- The pie chart shows the landings outcomes for Falcon 9 for CCAFS SLC-40.
- Blue indicates the failure and red indicates the success rate.



Payload vs. launch outcome

- The figures on the right shows the payload vs launch outcome scatter plots for all sites individually
- The payload range from 2000 kg to 5000 kg has the largest success rate,

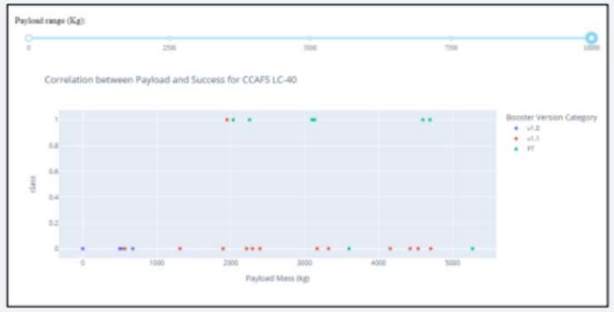


Fig: CCAFS LC-40

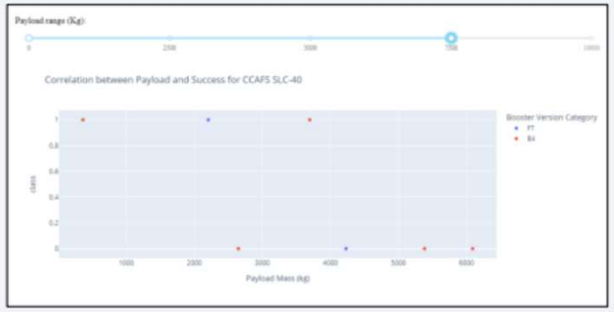


Fig: CCAFS SLC-40

Fig: KSC LC-39A

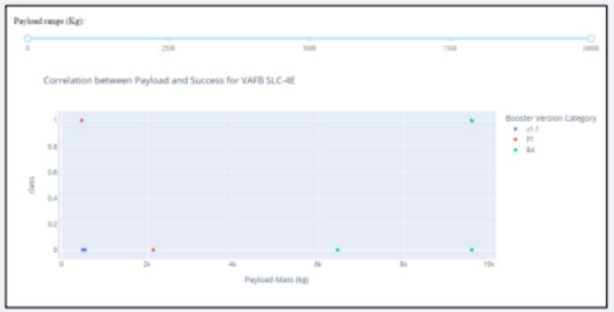


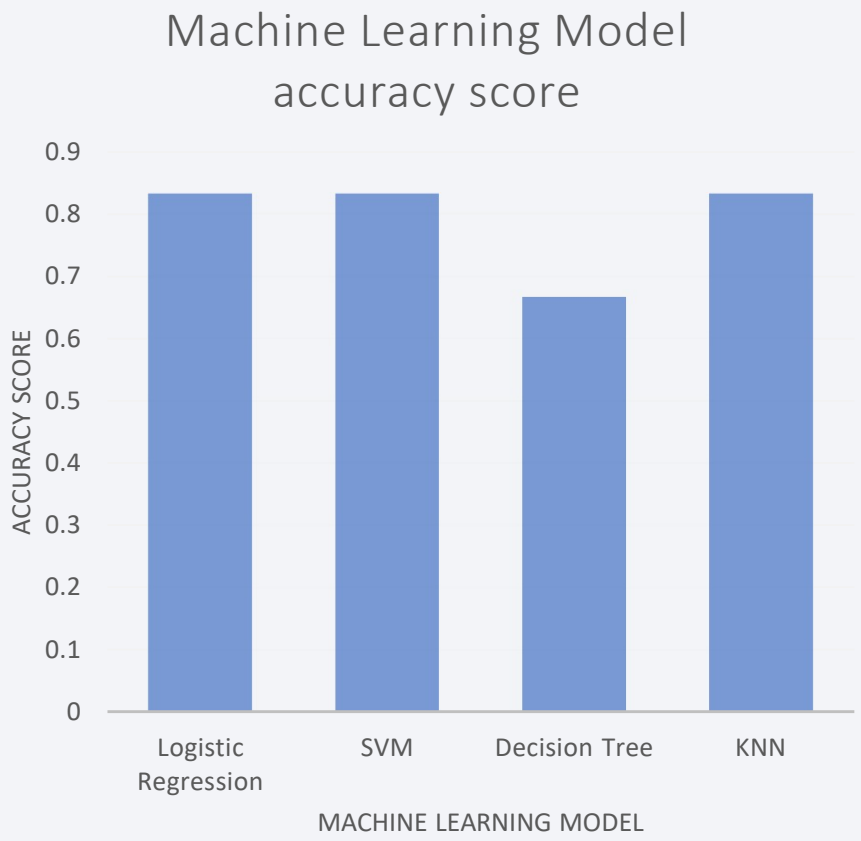
Fig: VAFB SLC-4E

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All model performed equally well except for the decision tree model which perform poorly relative to the other models.

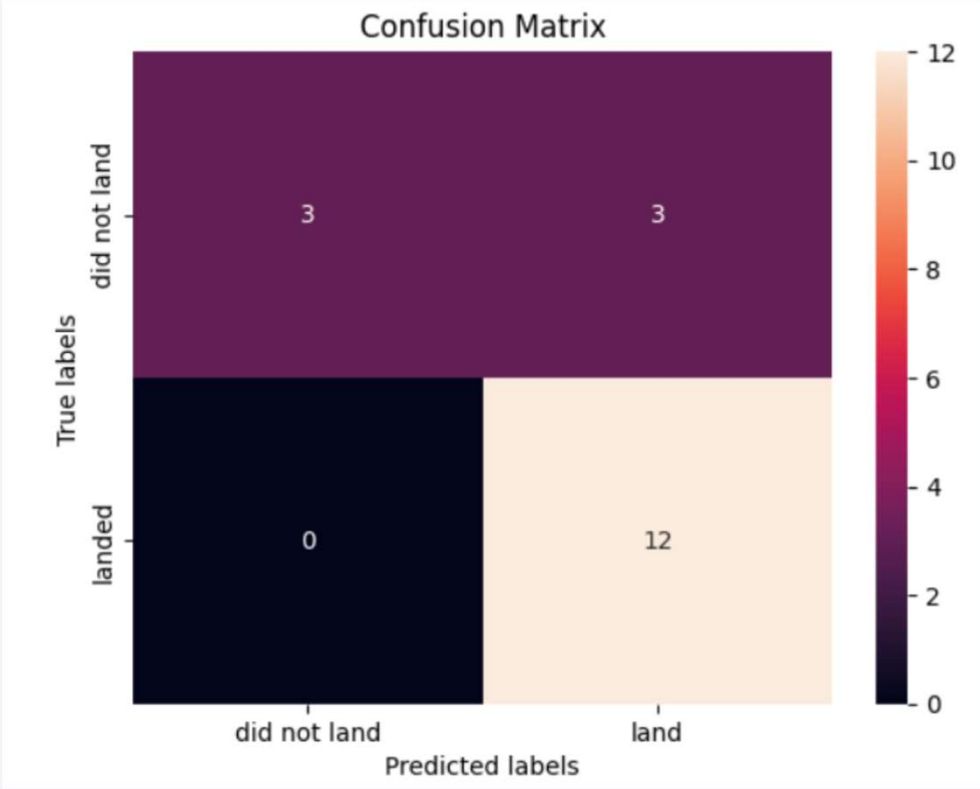


Confusion Matrix

- The figure shown on the right shows the confusion metrics for the logistic regression model,
- Confusion matrix can be read as

True Negative	False Positive
False Negative	True Positive

- Prediction breakdown:
  - 12 true positive and 3 true negatives
  - 3 false positive and 0 false negatives



Conclusions

- SpaceX record for Falcon nine first stage landing outcome has improved from 2010 to 2020.
- The trend is toward better performance and greater success with more launches being made.
- The machine learning models can be used by SpaceY as a benchmark to persue their interest.

Github link for the project

- **Complete the Data Collection API Lab**

https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab1A_collecting%20the%20data.ipynb

- **Complete the Data Collection with Web Scraping lab**

  https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab1B_webscraping%20from%20wikipedia.ipynb

- **Data Wrangling:**

  https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab2%20Data%20wrangling.ipynb

- **Complete the EDA with SQL**

  https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_lab3A_EDA%20with%20SQL.ipynb

Github link for the project

- **EDA with Visualization Lab:**

https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab3B_Exploring%20and%20preparing%20data.ipynb

- **Interactive Visual Analytics with Folium lab:**
  https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab4A_Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

- **Build an Interactive Dashboard with Ploty Dash ;**
  https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab4B_Interactive%20plotty%20dash.py

- **Complete the Machine Learning Prediction lab:**
  https://github.com/Kashex/SpaceX_Falcon9_first_stage_Landing_Prediction/blob/main/SpaceX_Lab5_Machine%20Learning%20Prediction_Part_5.ipynb

External Data sources

- Wikipedia page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- SpaceX API (JSON): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- • 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

Thank you!