Winning Space Race with Data Science

LEONG YEE FAI
10/02/2025

Outline

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

Executive Summary

- Summary of methodologies
  - ❖ Data of Falcon 9 first stage landings from 2010 to 2020 from a public API (https://api.spacexdata.com/) unaffiliated with SpaceX and from the publicly available data on Wikipedia (https:// en.wikipedia.org/wiki/SpaceX) were collected by web scrapping method. Furthermore, additional data sets were provided with the course.
  - ❖ Data cleaning / wrangling included extracting landing outcome data to serve as the dependent variable for the machine learning models.
  - ❖ SQL queries and data visualizations, including static plots, interactive maps, and an interactive dashboard, were used to discover insights about the dataset.
  - ❖ Predictive analysis was performed using the following machine learning models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbors (KNN). Its then all the model were done with comparison on accuracy.
- Summary of all results
  - ❖ The data of the SpaceX Falcon 9 first stage landings include the flight number, date of launch, payload mass, orbit type, launch site, and mission outcome.
  - ❖ Decision Tree had the highest accuracy while Logistic Regression, SVM, and KNN all performed equally well on this dataset for predictive purposes.

Introduction

Background

- SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, by determine if the first stage will successfully land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Business Problem

- What is the nature and extent of the available data about 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 future Falcon 9 first stage landing be 100% successful?

Section 1

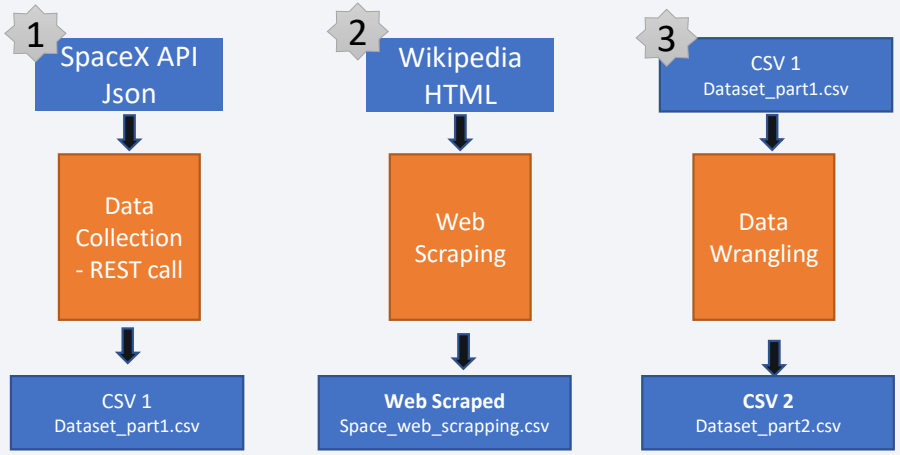
Methodology

Methodology

- First stage data of SpaceX Falcon 9 landings was collected from multiple source as public API, unaffiliated with SpaceX as well as Wikipedia article. Additional data sets were provided with the course in CSV file format.
- Data was prepared by wrangled and cleaned for visualizations, queries, and machine learning model training
- Exploratory Data Analysis (EDA) was performed using data visualizations and SQL.
- Interactive data visualizations were created using Folium and Plotly Dash.
- Predictive analysis using classification models was done using machine learning models

Data Collection

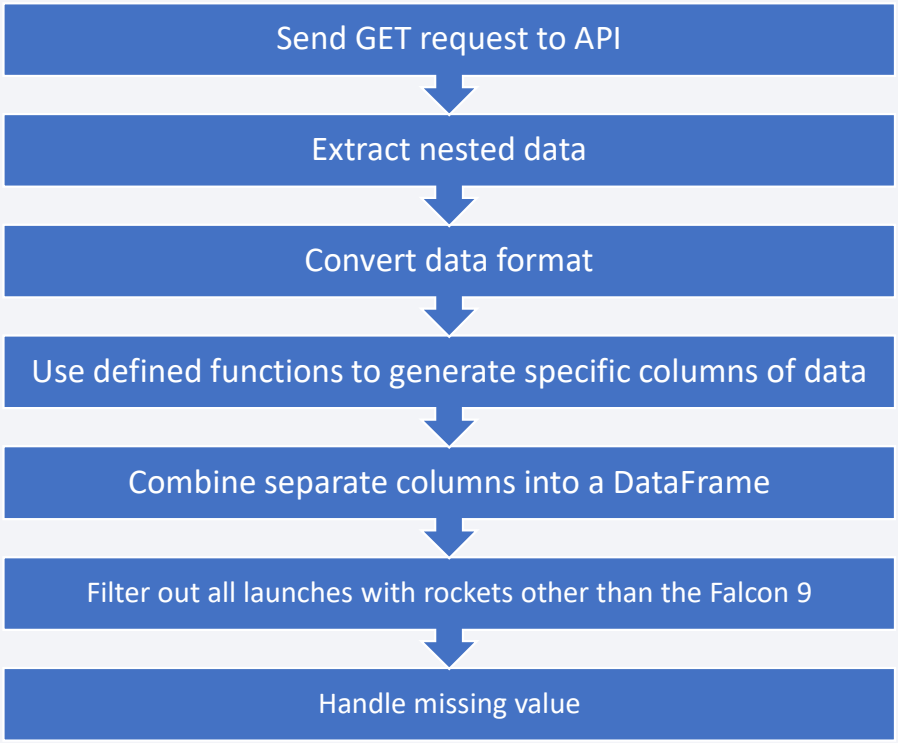
- 3 type of data source, SpaceX API, Wikipedia and additional data provided by the course.
- Each of the data format had undergo different method of data cleaning and data restructure by data collection , web scraping and data wrangling method to prepare for data visualization.



Data Collection – SpaceX API

- SpaceX data are available at the API endpoint: https://api.spacexdata.com/
- Data was extracted from the response from the API and loaded into a pandas dataframe for later analysis.

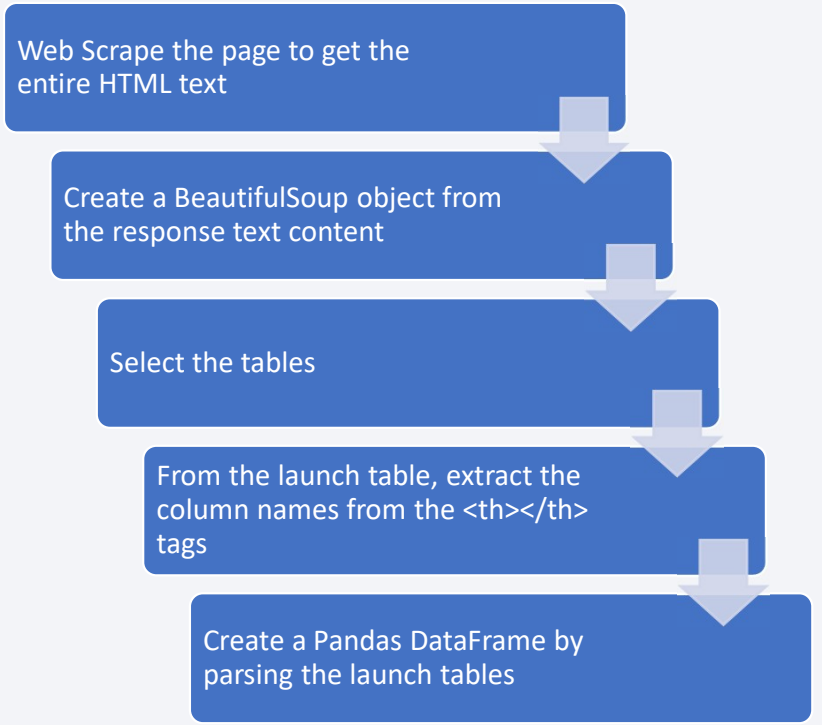
Github : https://github.com/yeefai96/Applied-data-Science-Capstone/blob/c142a60f2062371b9e2fb5cc09ee889476525be3/jupyter-labs-spacex-data-collection-api-bak-2025-02-07-10-34-18Z.ipynb



Data Collection - Scraping

- SpaceX launch data was scraped from HTML tables on a permanently-linked copy of the SpaceX Wikipedia webpage (https://en.Wikipedia.org/wiki/SpaceX)

Github: https://github.com/yeefai96/Applied-data-Science-Capstone/blob/c142a60f2062371b9e2fb5cc09ee889476525be3/jupyter-labs-webscraping.ipynb

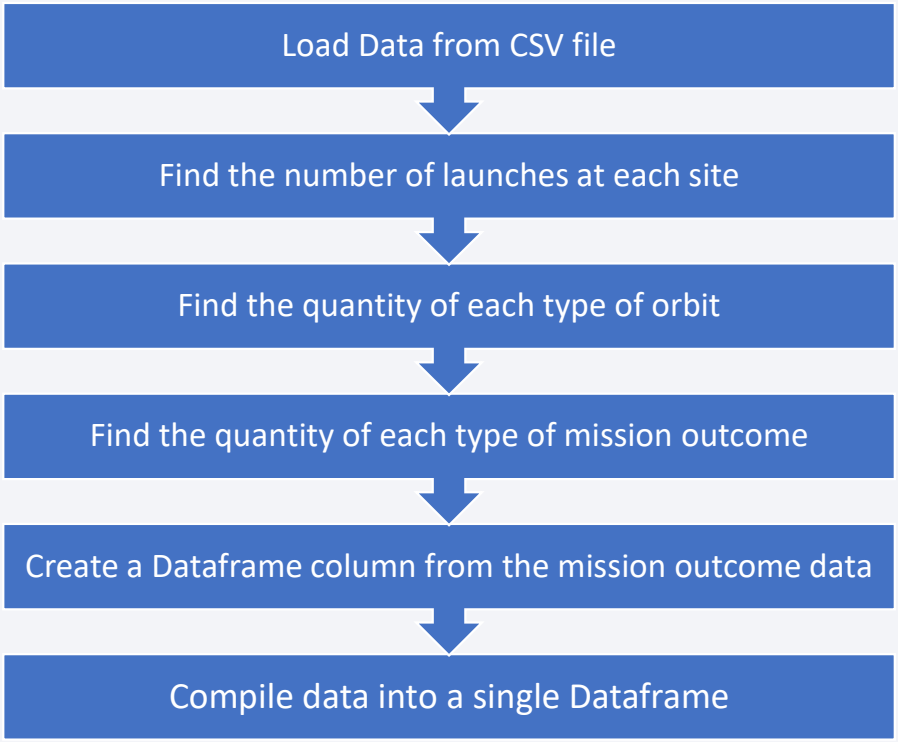


Data Wrangling

- The CSV file from the data collection consist of other unwanted data.
- The launch sites, orbit types and mission outcomes were filter and reformatted.
- The mission outcome types were converted to a binary classification (one-hot encoding) where 1 represented the Falcon 9 first stage landing being a success and 0 represented a failure.
- The new mission outcome classification column was added to the DataFrame.

- GitHub URL (Data Wrangling):

https://github.com/yeefai96/Applied-data-Science-Capstone/blob/9fb5e6ad7353ecc7a22c4d2c6671975b200f55a4/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with SQL

- Necessary information such as launch sites, payload mass, dates, booster types, as well as mission outcomes were extracted by using SQL queries

- GitHub:

https://github.com/yeefai96/Applied-data-Science-Capstone/blob/38c1bcb89c9c5076c6348402461845fdf53aaf87/jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with Data Visualization

- Relationship between flight number and payload with Trends of launch site and orbit type were compared corresponding by virtualized in scatterplot and bar chart to see the outcome.
- Lastly line plot were used to compare on mission outcome trend by year.

- Github:

https://github.com/yeefai96/Applied-data-Science-Capstone/blob/38c1bcb89c9c5076c6348402461845fdf53aaf87/edadataviz.ipynb

Build an Interactive Map with Folium

- The data were created with map object 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 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:

https://github.com/yeefai96/Applied-data-Science-Capstone/blob/18038bdcdc8edf728e12571f79061ab01d4a25e6/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The Plotly Dash dashboard included a dropdown input to select data from ‘one’ or ‘all’ launch sites to display on the pie chart and scatterplot.
- **For ‘one’ launch site**, the pie chart displayed the distribution of successful and failed Falcon 9 first stage landings for that site.
- **For ‘all’ launch sites**, the pie chart displayed the distribution of successful Falcon 9 first stage landings between the sites.
- **The input slider** is used to filter the payload masses for the scatterplot.
- **The scatterplot** displayed the distribution of Falcon 9 first stage landings split by payload mass, mission outcome and by booster version category.

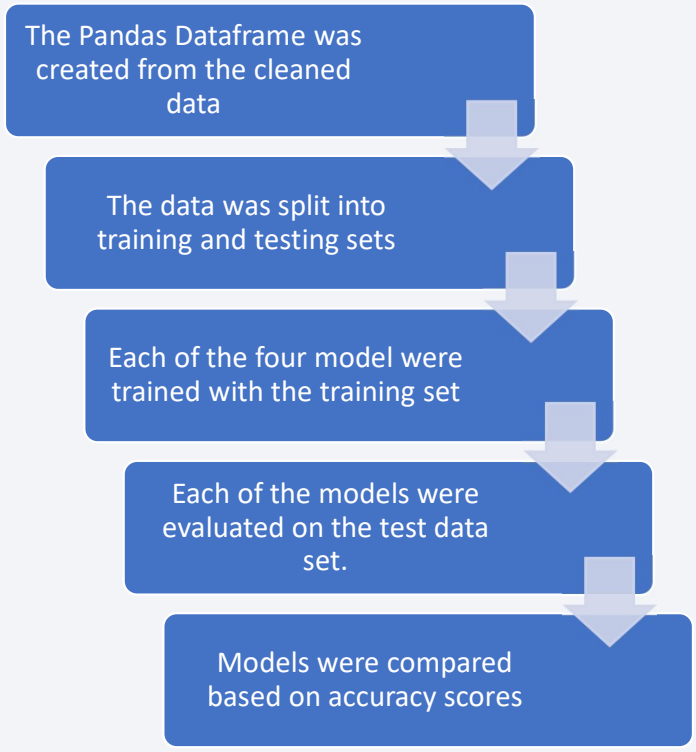
Github:

https://github.com/yeefai96/Applied-data-Science-Capstone/blob/d258264459fc8a5ad4ef4426683ac29dc9a2bf49/spacex_dash_app.py

Predictive Analysis (Classification)

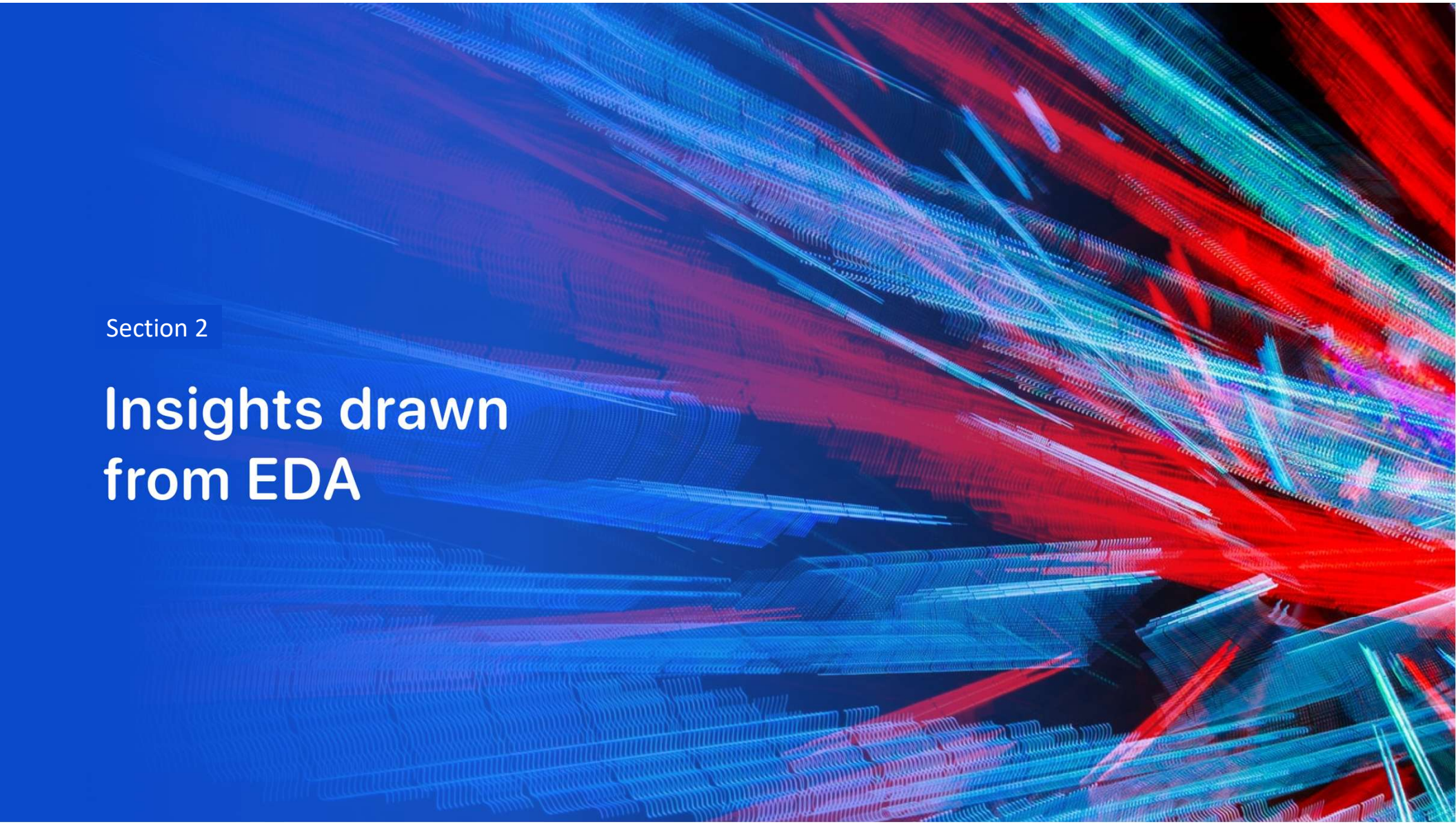
- The dataset was split into training and testing sets.
- The following machine learning models were trained on the training data set:
  - Logistic Regression
  - SVM (Support Vector Machine)
  - Decision Tree
  - KNN (k-Nearest Neighbors)
- Hyper-parameters were evaluated using GridSearchCV() and the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

https://github.com/yeefai96/Applied-data-Science-Capstone/blob/36fc2afdd3fbc7005686a3c586a92a2ee1f2ae15/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

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

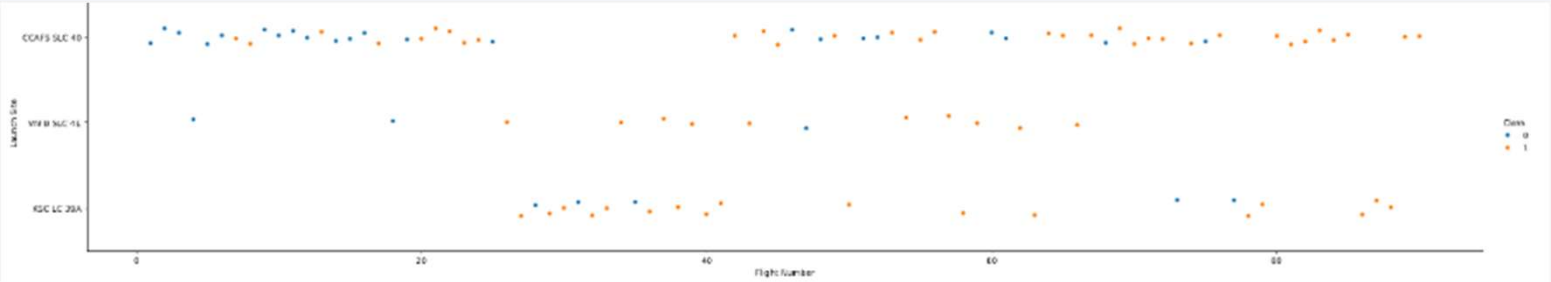


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

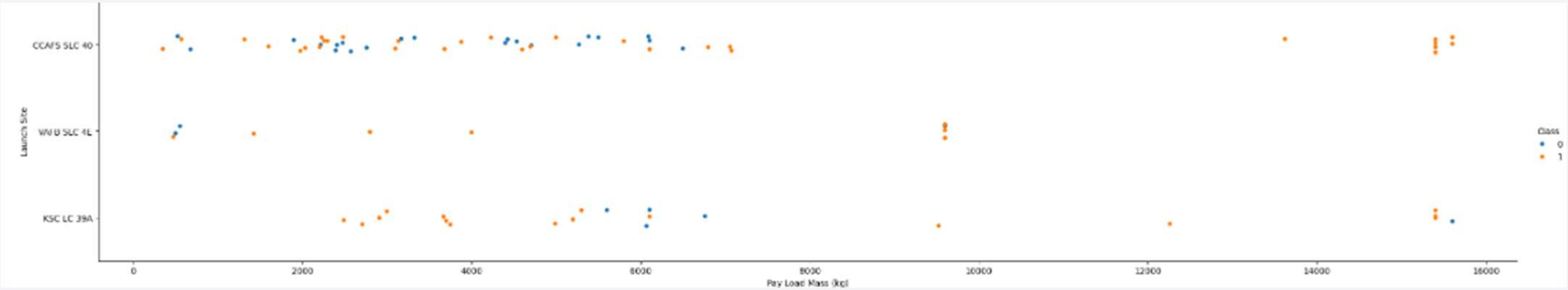
Scatter plot below show the plot of flight number vs launch sites:



- Total number of launches from CCAFS SLC 40 were the highest among the other sites
- Falcon 9 first stage launch in CCAFS SLC had achieved the greatest number of continuously success among the sites.

Payload vs. Launch Site

Scatter plot below show the plot of Payload vs launch sites:

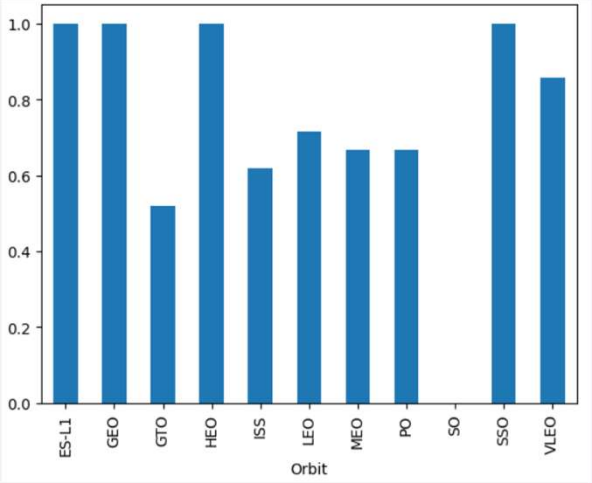


Success rate increase along with payload mass, the successful stage 1 landing rate increase whe the payload increases over 8000 kg and before 16000 kg

Success Rate vs. Orbit Type

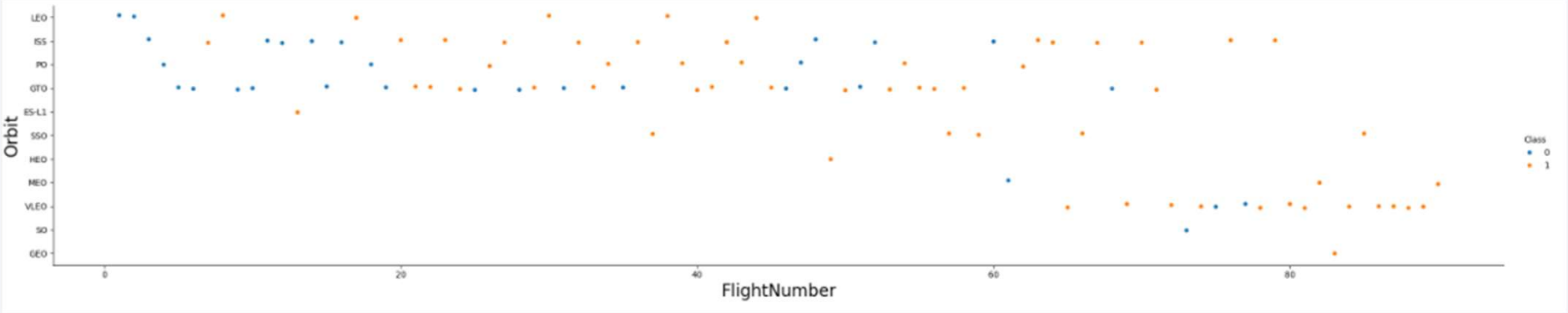
- ES-L1, GEO, HEO and SSO have the 100% success rate of landing of stage 1

- Meanwhile, SO has no successful first stage landing



Flight Number vs. Orbit Type

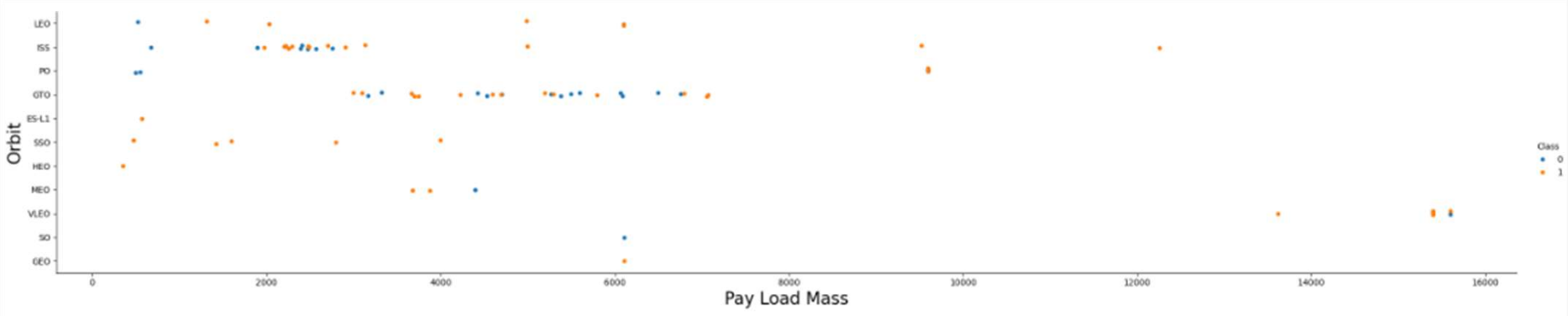
Scatter plot below show the plot of flight number vs orbit type:



The larger the flight number the higher the success rate of successful stage 1 landing

Payload vs. Orbit Type

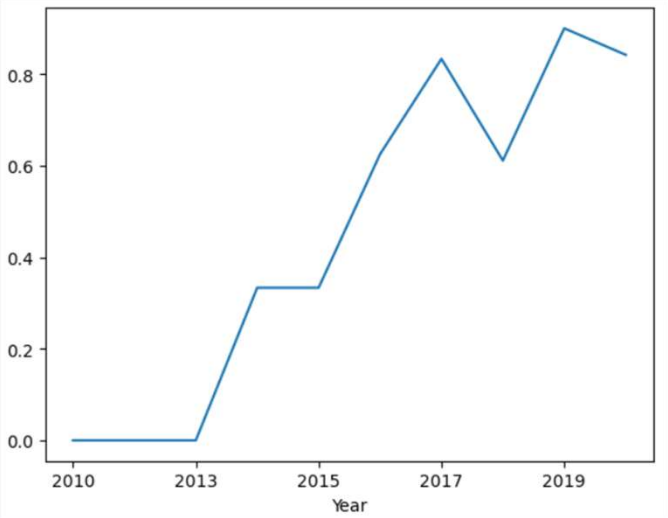
Scatter plot below show the plot of payload vs orbit type:



Success rate appeared to have no obvious correlation with payload mass.

Launch Success Yearly Trend

Scatter plot below show the plot of launch success vs yearly trend:



The success rate of the Falcon 9 first stage landings has increased significantly over the selected interval of years.

All Launch Site Names

- **Question :** Find the names of the unique launch sites
- **Query :** %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
- **Explanation:** Filter the unique launch sites, and there are only 4 sites
- **Result:**

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

Launch Site Names Begin with 'CCA'

- **Question :** Find 5 records where launch sites begin with `CCA`
- **Query :** %sql SELECT * FROM 'SPACEXTBL' WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- **Explanation:** Filter data without knowing the full naming of the names
- **Result:**

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcom
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 attemp
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attemp
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Total Payload Mass

- **Question :** Calculate the total payload carried by boosters from NASA
- **Query :** %sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
- **Explanation:** filter data with specific customer and total payload
- **Result:**

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- **Question :** Calculate the average payload mass carried by booster version F9 v1.1
- **Query :** %sql SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1%';
- **Explain**: Using AVG function to calculate the average
- **Result:**

Payload Mass Kgs	Customer	Booster_Version
2534.6666666666665	MDA	F9 v1.1 B1003

First Successful Ground Landing Date

- **Question:** Find the dates of the first successful landing outcome on ground pad
- **Query:** SELECT min(`DATE`) AS "First Successful Landing Outcome Date" FROM `SPACEXDATASET` WHERE `landing__outcome` LIKE 'Success (ground pad)'
- **Explanation:** using min(date) to find the earliest date among the data.
- **Result:**

First Successful Landing Outcome Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Question :** List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- **Query :** %sql SELECT DISTINCT "Booster_Version", "Payload" FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (drone ship)" AND "PAYLOAD_MASS__KG_"> 4000 AND "PAYLOAD_MASS__KG_"<6000;
- **Explanation:** using “and” function to filter the in between value.
- **Result:**

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- **Question:** Calculate the total number of successful and failure mission outcomes
- **Query:** %sql 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%';
- **Explanation:** Present your query result with a short explanation here
- **Result:**

Success	Failure
61	40

Boosters Carried Maximum Payload

- **Question:** List the names of the booster which have carried the maximum payload mass
- **Query:** %sql SELECT "Booster_Version","Payload", "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
- **Explanation:** using “MAX” Function to select the maximum value
- **Result:**

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

2015 Launch Records

- **Question:** List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- **Query:** %sql SELECT MONTHNAME(DATE) AS "Month", landing__outcome, booster_version, launch_site FROM SPACEXDATASET WHERE YEAR(DATE) = 2015;
- **Explanation:** by using "=" function to filter the exact value
- **Result:**

Month	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
February	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
March	No attempt	F9 v1.1 B1014	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
April	No attempt	F9 v1.1 B1016	CCAFS LC-40
June	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
December	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

- **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 :** %sql 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

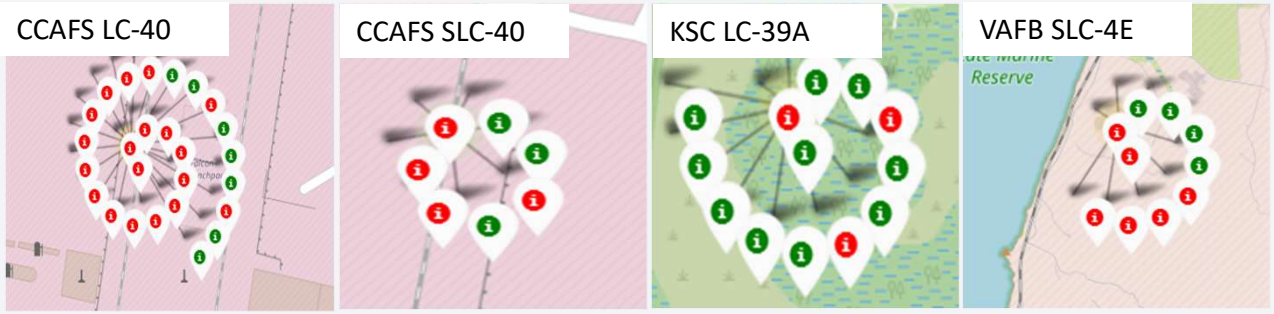
Section 3

Launch Sites Proximities Analysis

Falcon 9 Launch Sites

All launch location are pin with name on the map

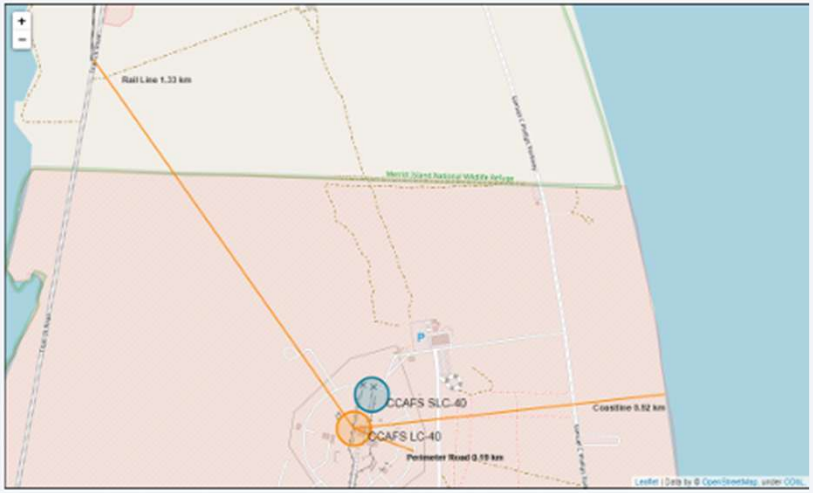
Success/Failure Attempt at Each Site

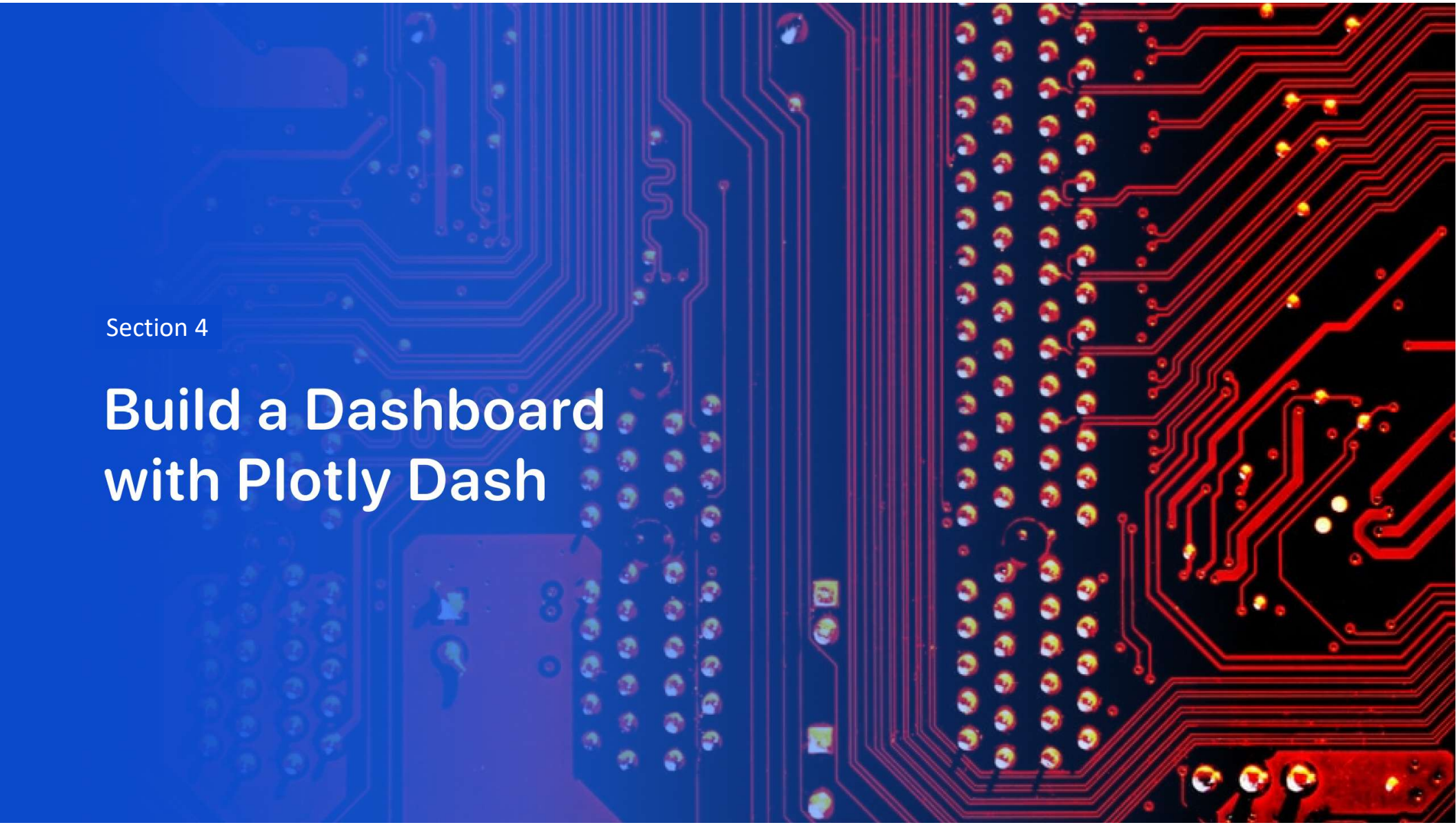


- The markers display the mission outcomes (Success/Failure) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site
- A sense of a launch site's success rate for Falcon 9 first stage landings can be gleaned from the relative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

Relation between Launch Site and 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

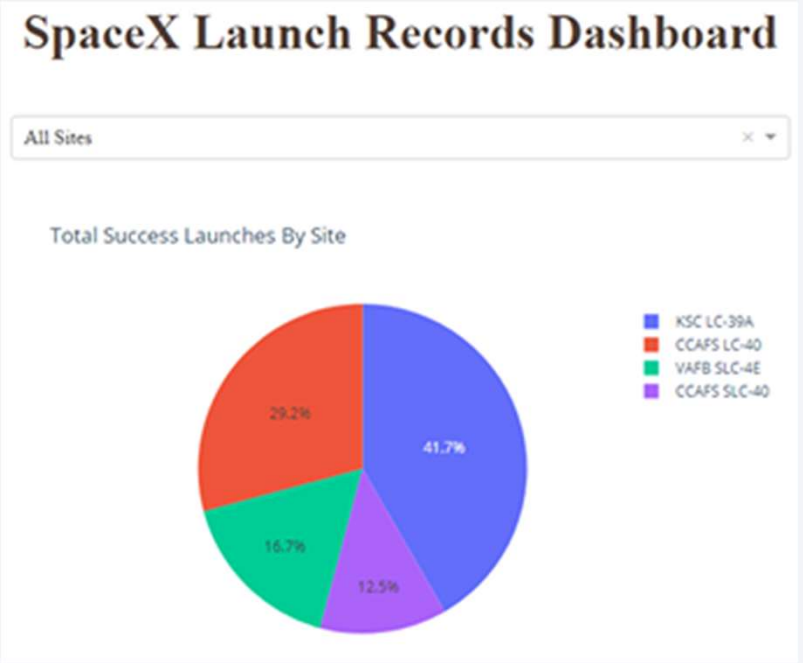
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard

Figure shown the records of successful launches at all sites by pie chart

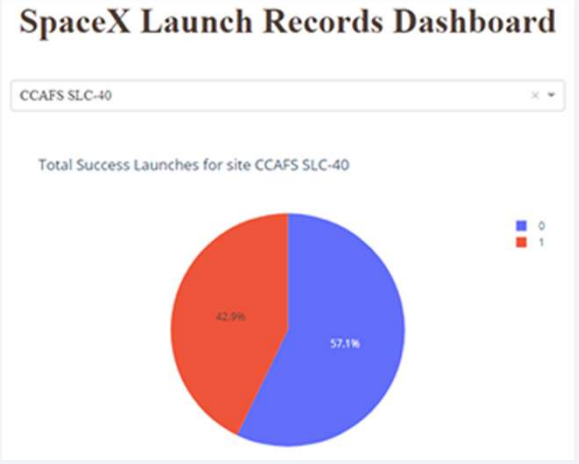
KSC LC-39A had the highest successful launches among all the sites which had the 41.7% success rate.

CCAFS SLC-40 had the lowest success rate among the sites, which only had 12.5%



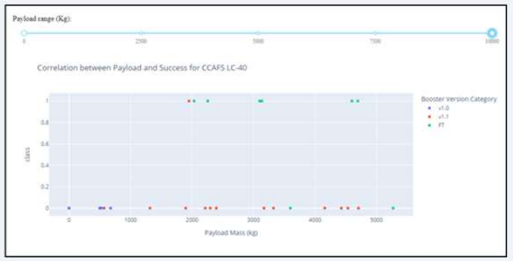
Highest Launch success ratio

- 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%).

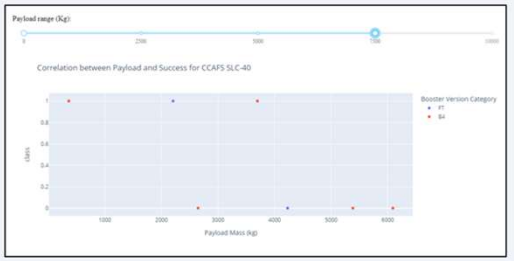


Payload vs Outcome

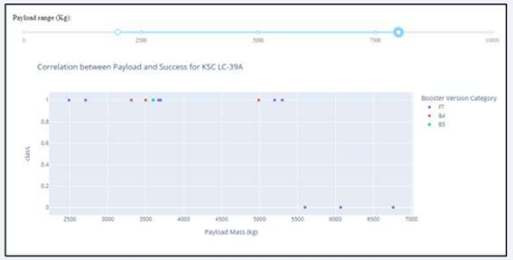
- These screenshots are of the Payload vs. Launch Outcome scatter plots for all sites, with different payload selected in the range slider.
- The payload range from about 2,000 kg to 5,000 kg has the largest success rate.
- The ‘FT’ booster version category has the largest success rate



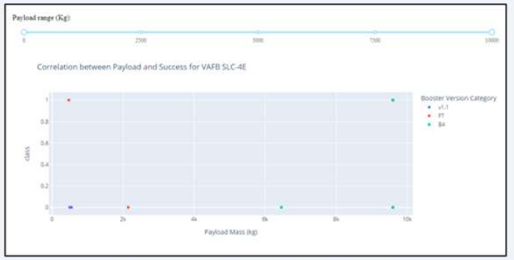
CCAFS LC-40



CCAFS SLC-40



KSC LC-39A



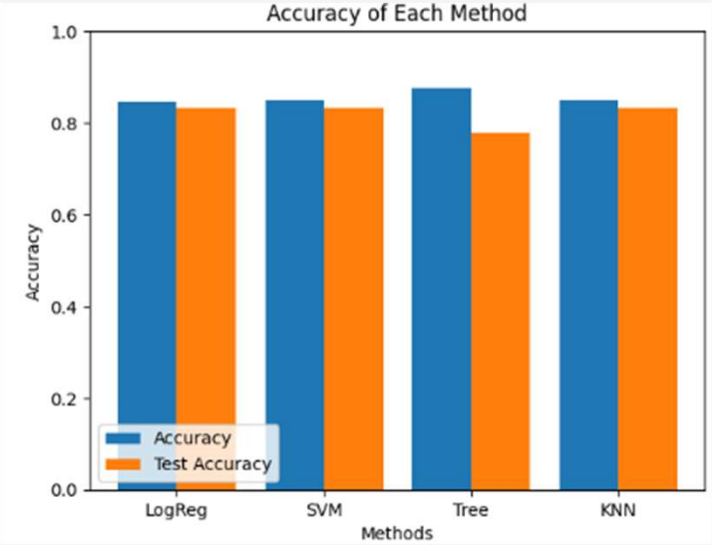
VAFB SLC-4E

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- 3 model LogReg , SVM and KNN model had highest and equally accuracy. On the other hand, Decision Tree model had performed the worst accuracy.

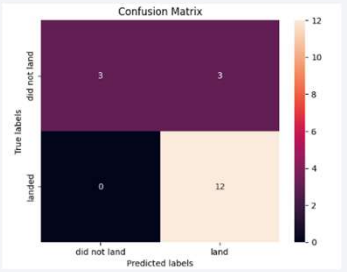


Confusion Matrix

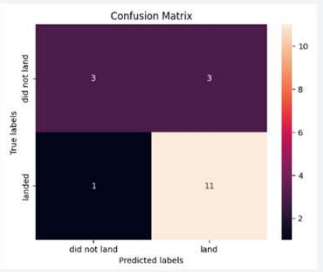
- Shown here is the confusion matrix for the Logistic Regression model.
- Confusion matrices can be read as:

True Negative	False Positive
False Negative	True Positve

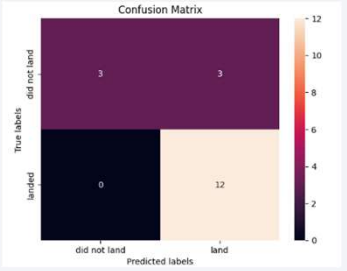
- Based on 4 confusion matrix, tree has the greatest number of false negative among model.



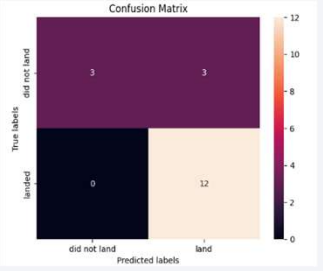
LogReg



Tree



SVM



KNN

Conclusions

- LR, SVM, KNN are top-performing models for forecasting outcomes in this data
- Lighter payloads have a higher performance compared to heavier ones.
- The likelihood of a SpaceX launch succeeding increases with the number of years of experience, suggesting a trend towards flawless launches over time.
- Launch Complex 39A at Kennedy Space Center has the highest number of successful launches compared to other launch sites.
- GEO,HEO,SSO,ES L1 orbit types exhibit the highest rates of successful launches

Thank you!