IBM

Applied Data Science Capstone

Venkata Pavani Perla

June02, 2024

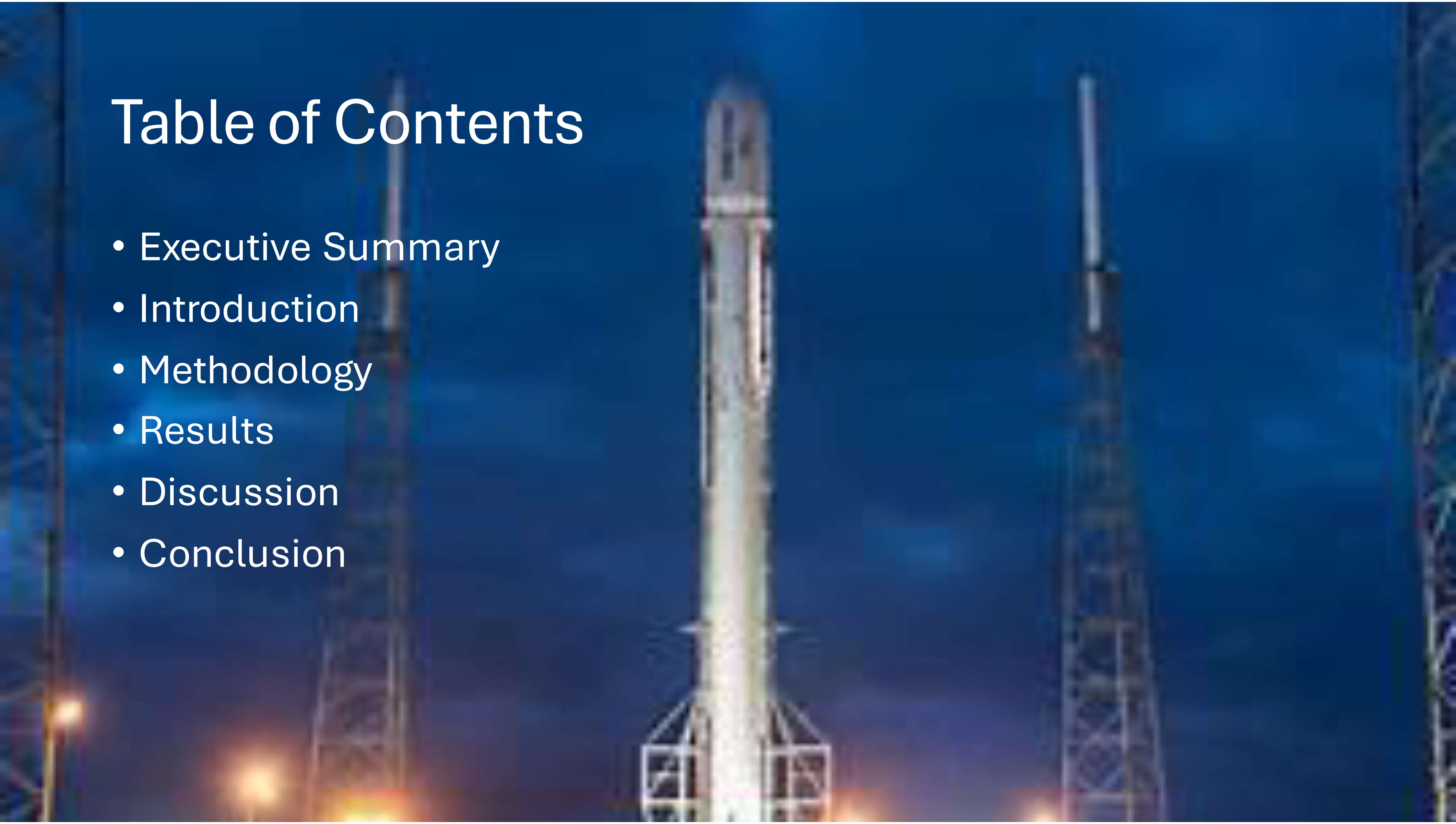


Table of Contents

- Executive Summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion

EXECUTIVE SUMMARY

- In this capstone, we will predict if the Falcon 9 first stage will land successfully using several machine learning classification algorithms.
- The main steps in this project include:
  - Data collection, Data wrangling and formatting.
  - Exploratory Data Analysis
  - Interactive Data Visualization
  - Machine Learning Prediction
- Our graphs show that some features of the rocket launches have correlation with the outcome of the launches i.e., success or failure.
- It is also concluded that decision tree may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully.

INTRODUCTION

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. 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 if we can determine if the first stage will 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.
- Most unsuccessful landings are planned. Sometimes, Space X will perform a controlled landing in the ocean.
- The main question we are trying to answer is , for a given set of features about a Falcon 9 rocket launch which include its payload mass, orbit type, launch site, and so on, will first stage of the rocket land successfully?

METHODOLOGY

The overall methodology includes:

1. Data collection, Data wrangling, and formatting, using:

SpaceX API

Web Scraping

2. Exploratory Data Analysis (EDA), using:

Pandas and Numpy

SQL

3. Data Visualization, using:

Matplotlib and Seaborn

Folium

Dash

4. Machine Learning prediction, using:

Logistic Regression

Support Vector Machine (SVM)

Decision Tree

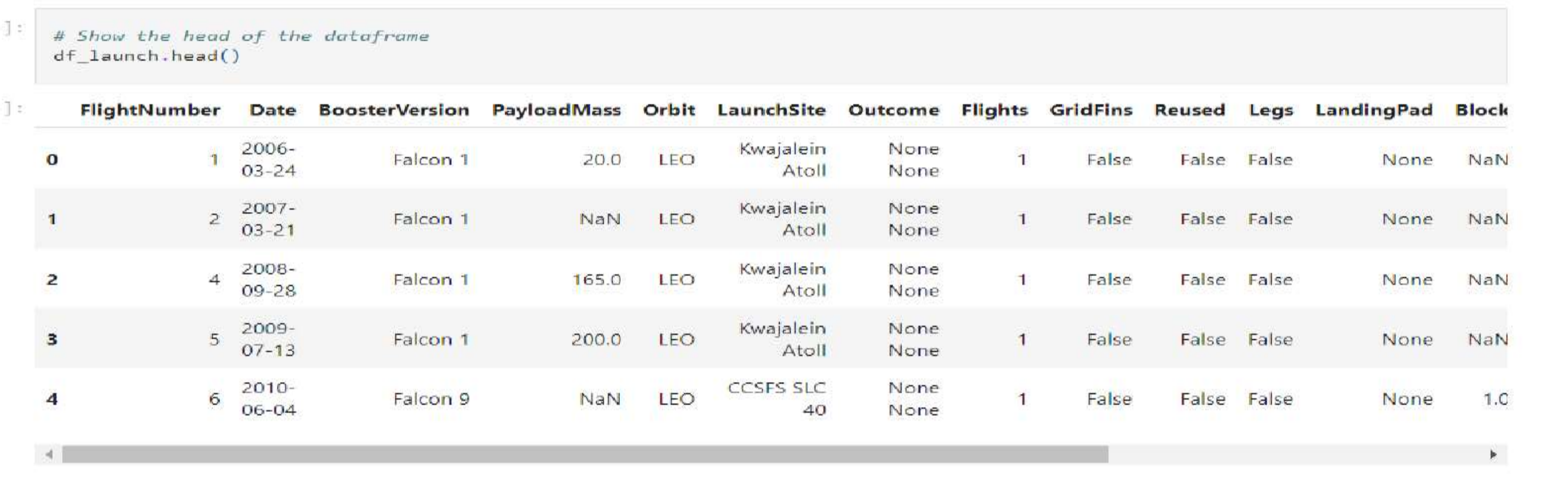
K-nearest neighbors (KNN)

METHODOLOGY

1. Data collection, Data Wrangling and Formatting

SpaceX API:

- The API used is "https://api.spacexdata.com/v4/launches/past"
- The API provides data about many types of rocket launches done by Space X, the data is therefore filtered to include only falcon 9 launches.
- We end up with 90 rows or instances and 17 columns or features. The below snapshot shows the first few rows of data:



```
# Show the head of the dataframe
df_launch.head()
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0

METHODOLOGY

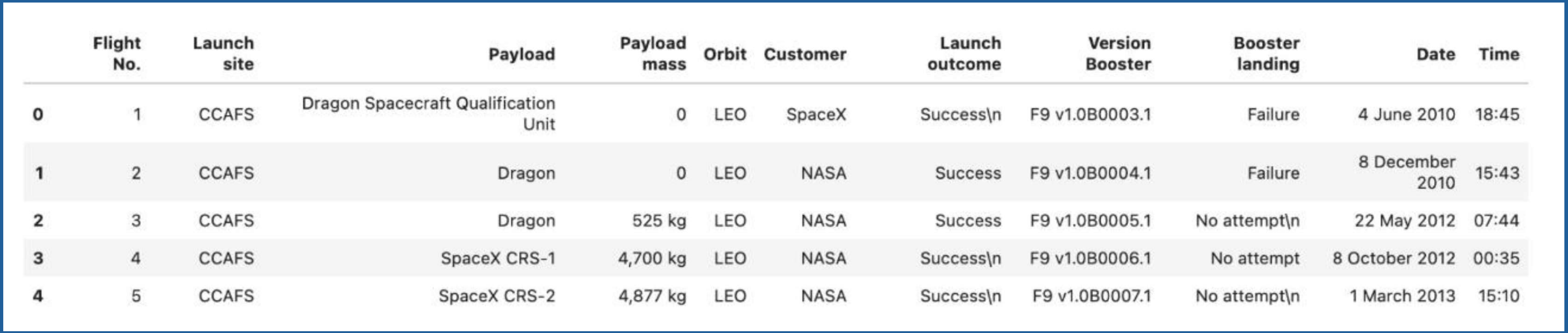
1. Data collection, Data Wrangling and Formatting

Web scraping

- The data is scraped from

"https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"

- The website contains only the data about Falcon 9 launches.
- We end up with 121 rows or instances and 11 column or features. The below snapshot shows the first few rows of the data:



	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

METHODOLOGY

1. Data collection, Data Wrangling and Formatting

- The data is later processed so that there is no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called 'class' is also added to the data frame. The column 'class' contains '0' if a given launch is failed and '1' if it is successful.
- In the end, we end up with 90 rows or instances and 83 columns or features.

METHODOLOGY

2. Exploratory Data Analysis (EDA)

Pandas and NumPy - Functions from the Pandas and NumPy libraries are used to derive basic information about the data collected, which includes:

- The number of launches in each site.
- The number of occurrences in each orbit.
- The number of occurrences of each mission outcome.

SQL- The data is queried using SQL to answer about the data such as:

- The names of the unique launch sites in the space mission.
- The average and total payload mass carried by boosters launched by NASA.

METHODOLOGY

3. Data Visualization

Matplotlib and Seaborn – Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts and line charts.

The plots and charts are used to understand more about the relationships between several features such as:

- The relationship between flight number and launch site.
- The relationship between payload mass and launch site.
- The relationship between success rate and orbit type.

Folium – Functions from the folium libraries are used to visualize the data through interactive maps.

The Folium library is used to:

- Mark all launch sites on a map.
- Mark the succeeded launches and failed launches for each site on a map.
- Mark the distances between a launch site to its proximities such as nearest city, railway or highway.

METHODOLOGY

3. Data Visualization

Dash – The functions from dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a ranger slider.

Using a pie chart and a scatter plot, the interactive site shows:

- The total success launches from each site.
- The correlation between payload mass and mission outcome (success or failure) for each launch site.

METHODOLOGY

4. Machine Learning Prediction

Functions from Scikit-learn library are used to create our machine learning models. The machine learning prediction phase include the following steps:

1. Standardizing the data
2. Splitting the data in to training and test data.
3. Creating machine learning models such as, Logistic Regression, Support Vector Machine, Decision tree, K nearest neighbors
4. Fit the models on the training set.
5. Find the best combination of hyper parameters for each model.
6. Evaluate the models based on their accuracy scores and confusion matrix.

RESULTS

The results are split in to 5 sections:

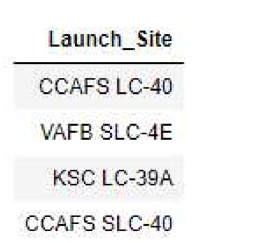
- SQL (EDA with SQL)
- Matplotlib and Seaborn
- Folium
- Dash
- Predictive Analysis

Note: In all the graphs that follows class '0' represents failed launch outcome and class '1' represents a successful launch outcome.

RESULTS

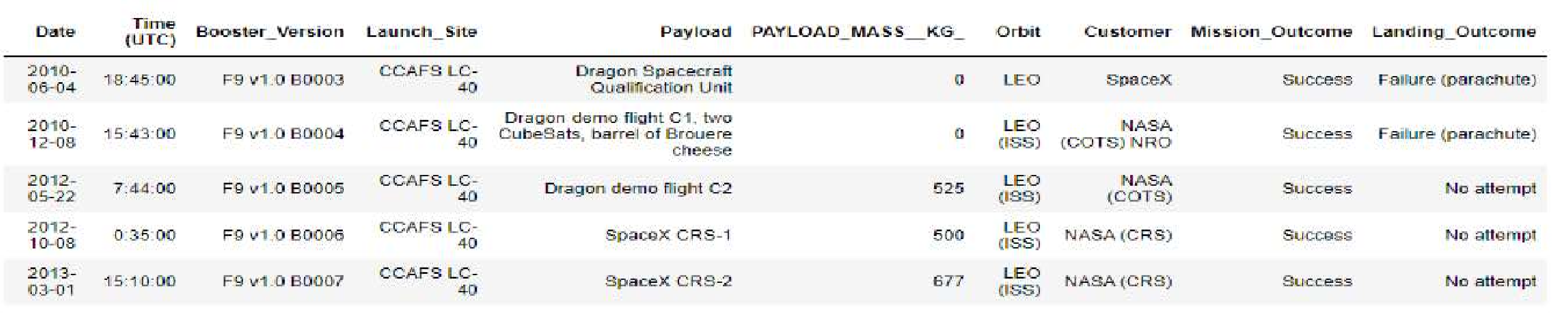
SQL (EDA with SQL)

- The names of the unique launch sites in the space mission.



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

- 5 Records where launch sites begin with 'CCA'

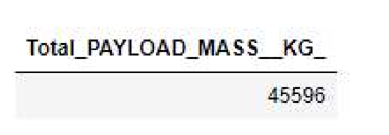


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

RESULTS

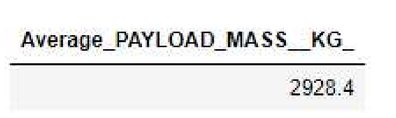
SQL (EDA with SQL)

- The total payload mass carried by boosters launched by NASA (CRS)



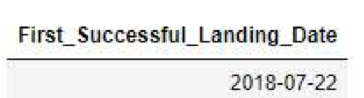
Total_PAYLOAD_MASS__KG_
45596

- The average payload mass carried by booster version F9v1.1



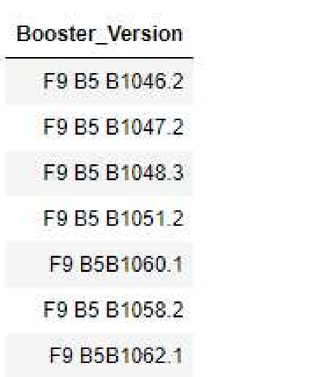
Average_PAYLOAD_MASS__KG_
2928.4

- The date when the first successful landing outcome in ground pad was achieved.



First_Successful_Landing_Date
2018-07-22

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

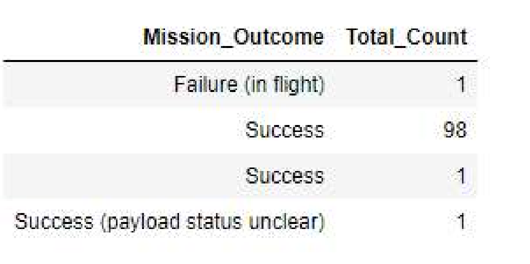


Booster_Version
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

RESULTS

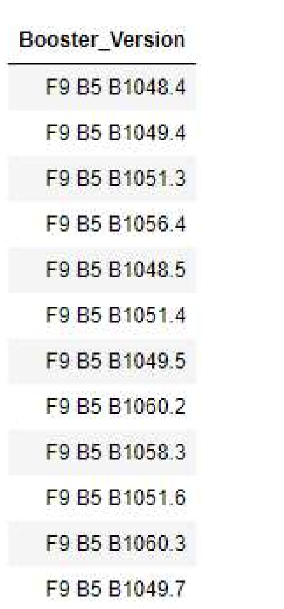
SQL (EDA with SQL)

- The total number of successful and failure mission outcomes.



Mission_Outcome	Total_Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- The names of the booster versions which have carried the maximum payload mass.

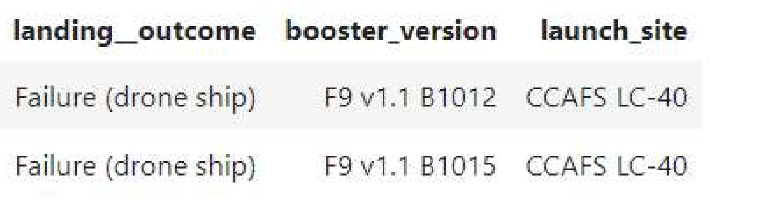


Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

RESULTS

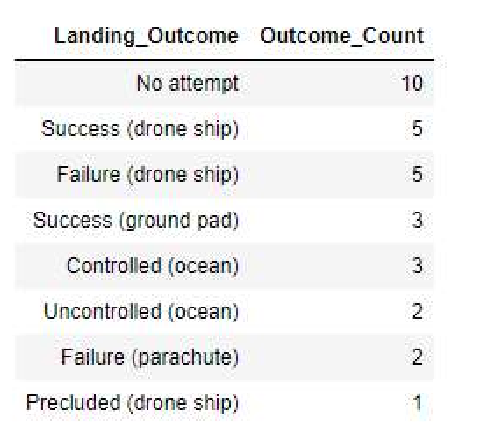
SQL (EDA with SQL)

- The failed landing outcomes in drone ship ,booster versions, and launch site names in year 2015



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

- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

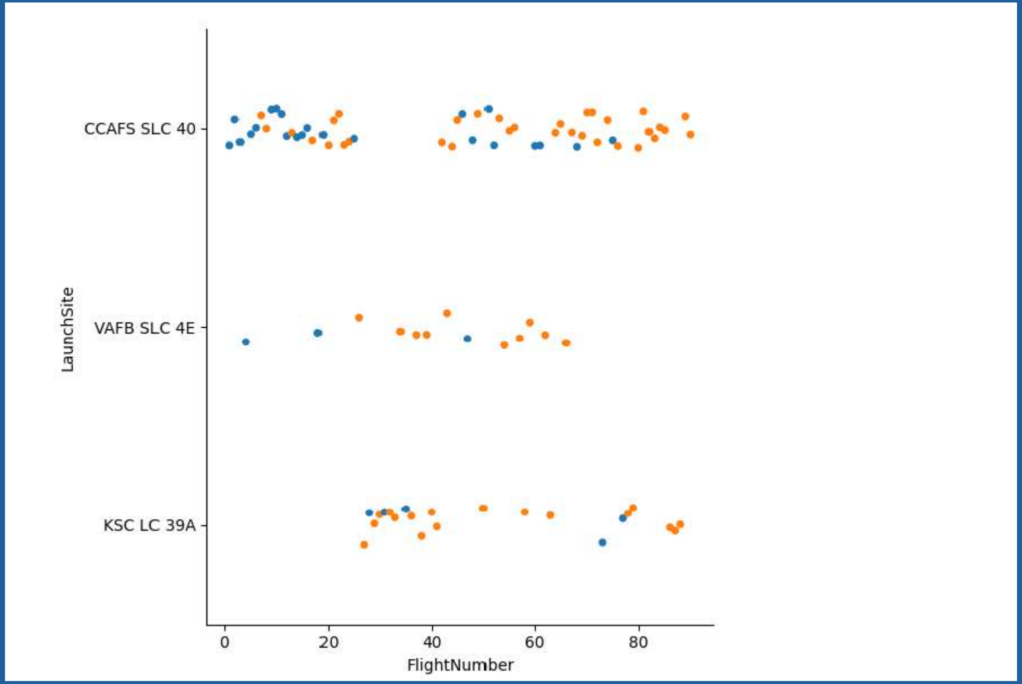


Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

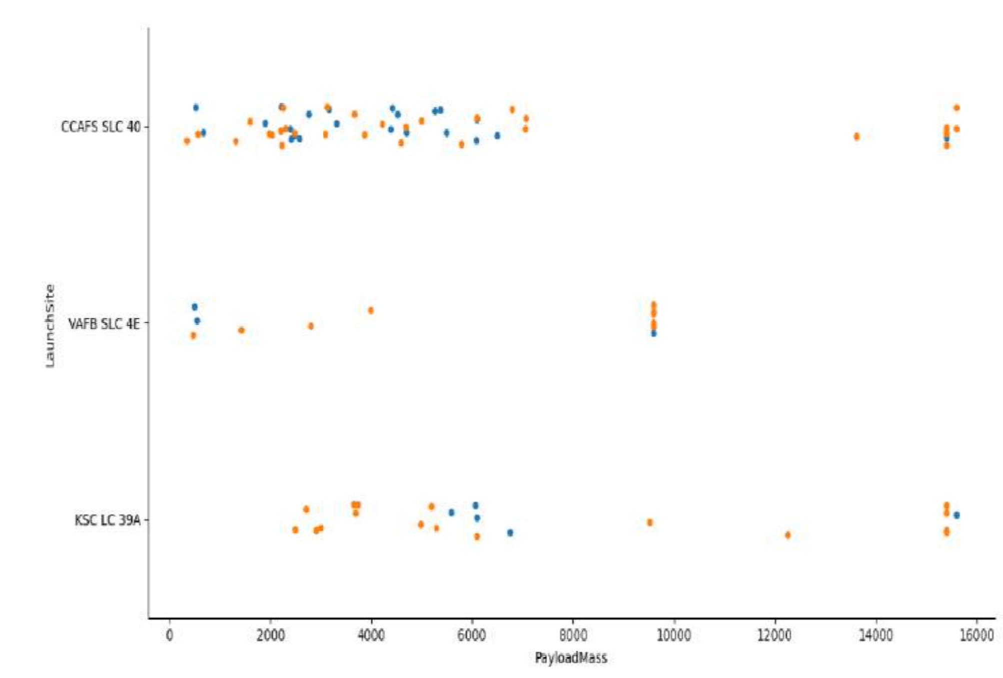
RESULTS

Matplotlib and Seaborn (EDA with Visualization)

- The relationship between flight number and launch site.



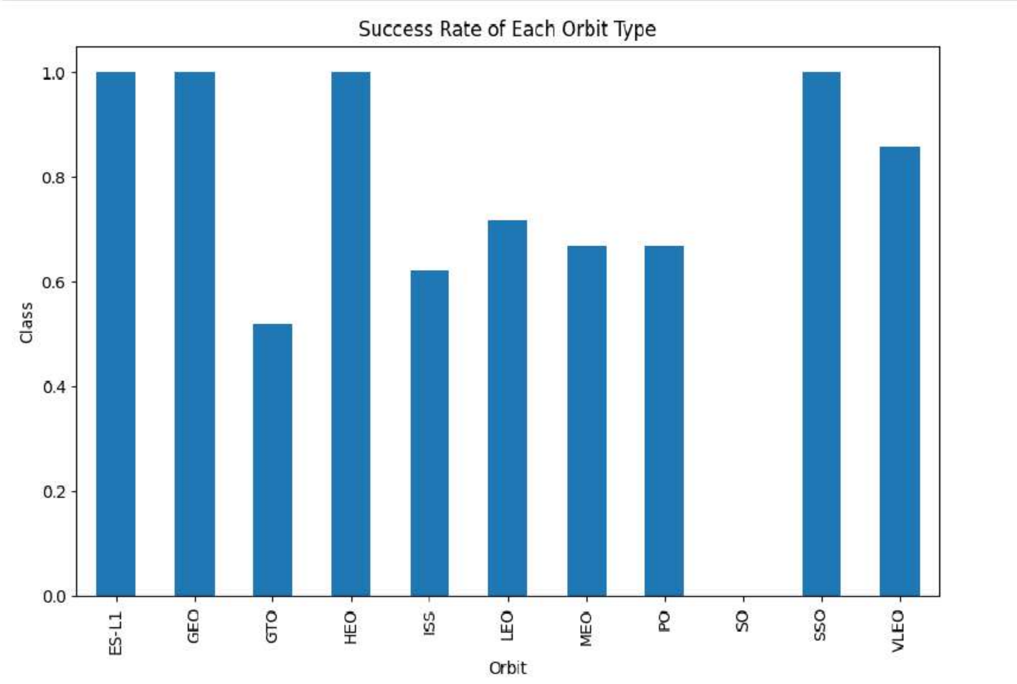
- The relationship between payload mass and launch site.



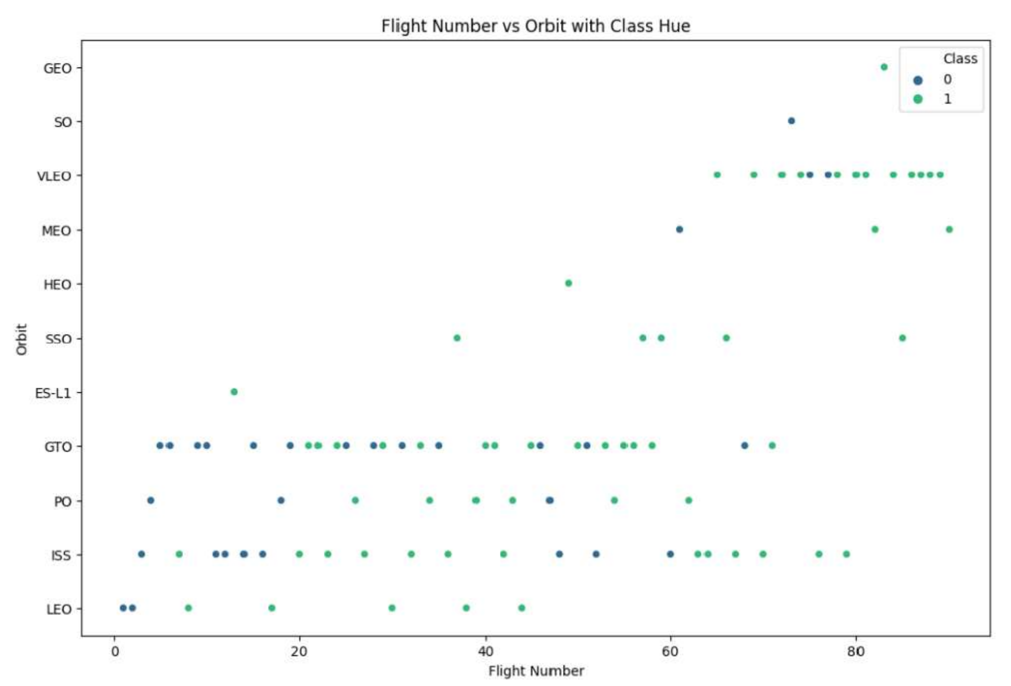
RESULTS

Matplotlib and Seaborn (EDA with Visualization)

- The relationship between success rate and orbit type.



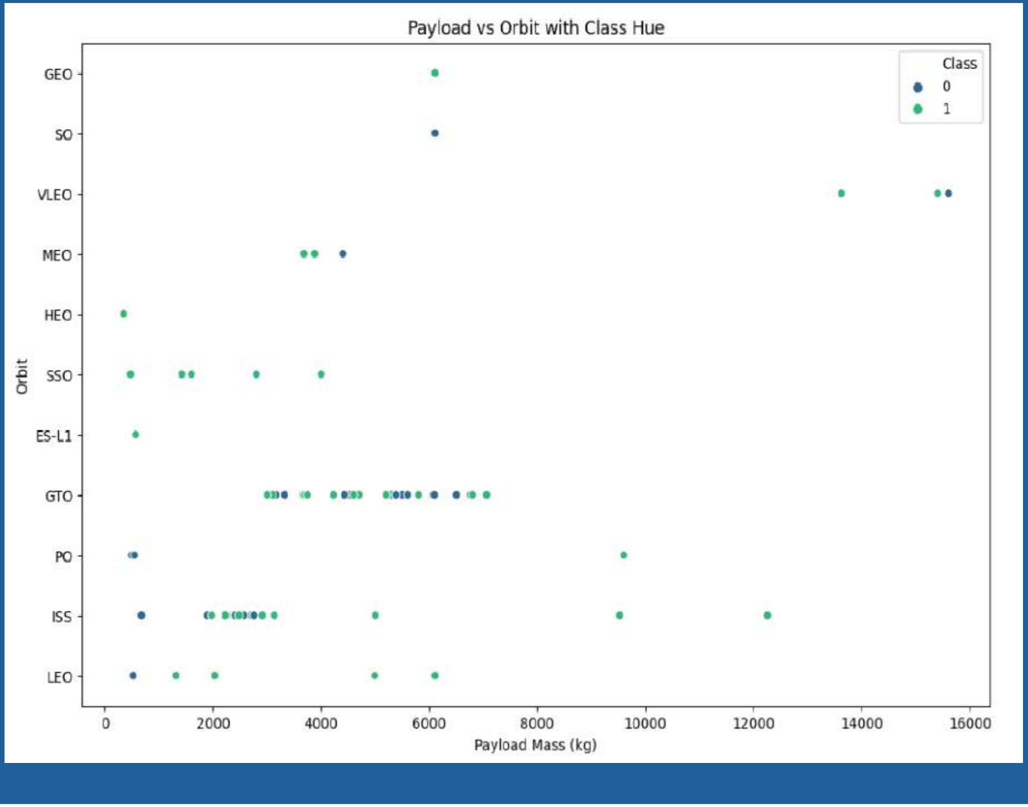
- The relationship between flight number and orbit type.



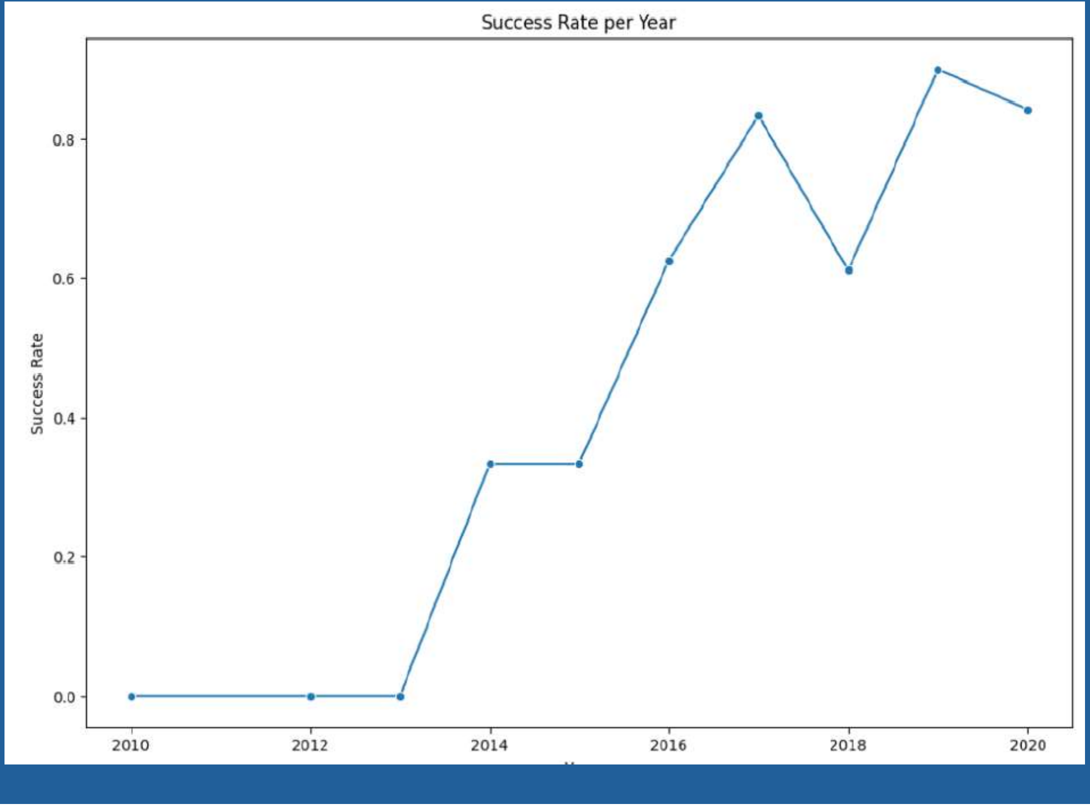
RESULTS

Matplotlib and Seaborn (EDA with Visualization)

- The relationship between payload mass and orbit type.



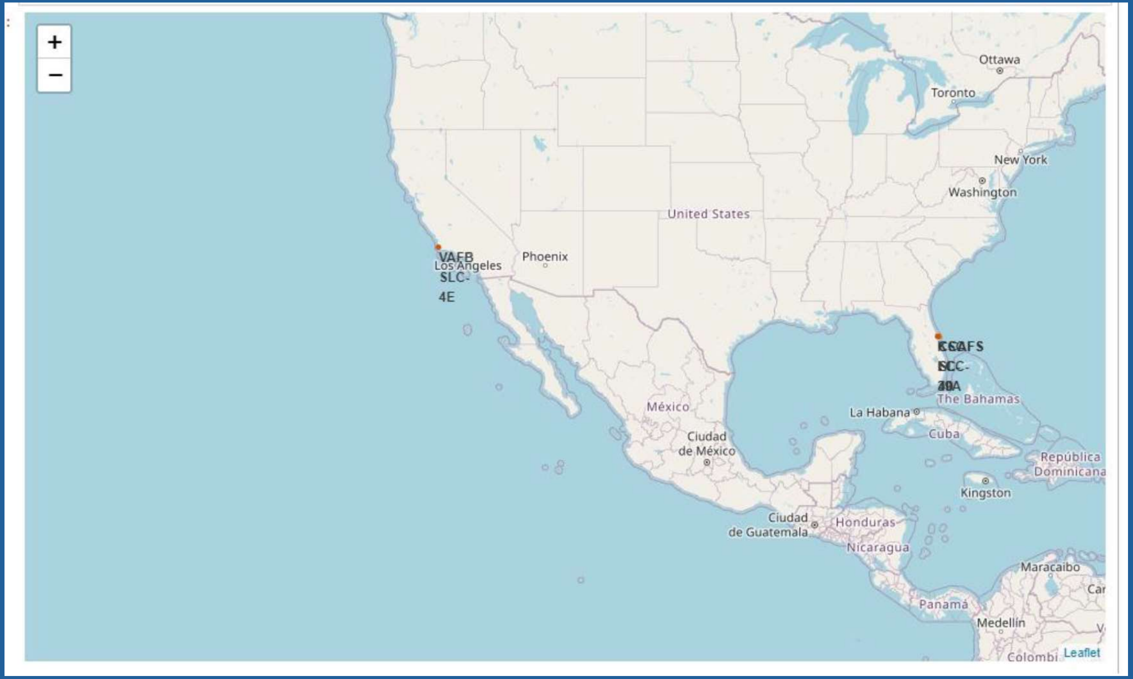
- The launch success yearly trend



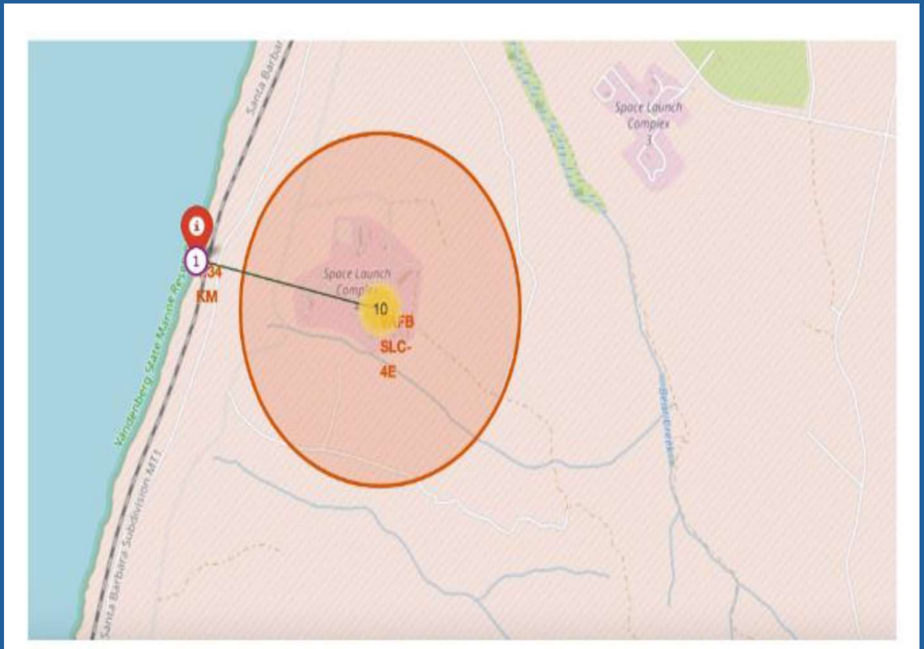
RESULTS

Folium

- All launch sites on map



- The distance between launch site to its proximities such as nearest city, railway or highway

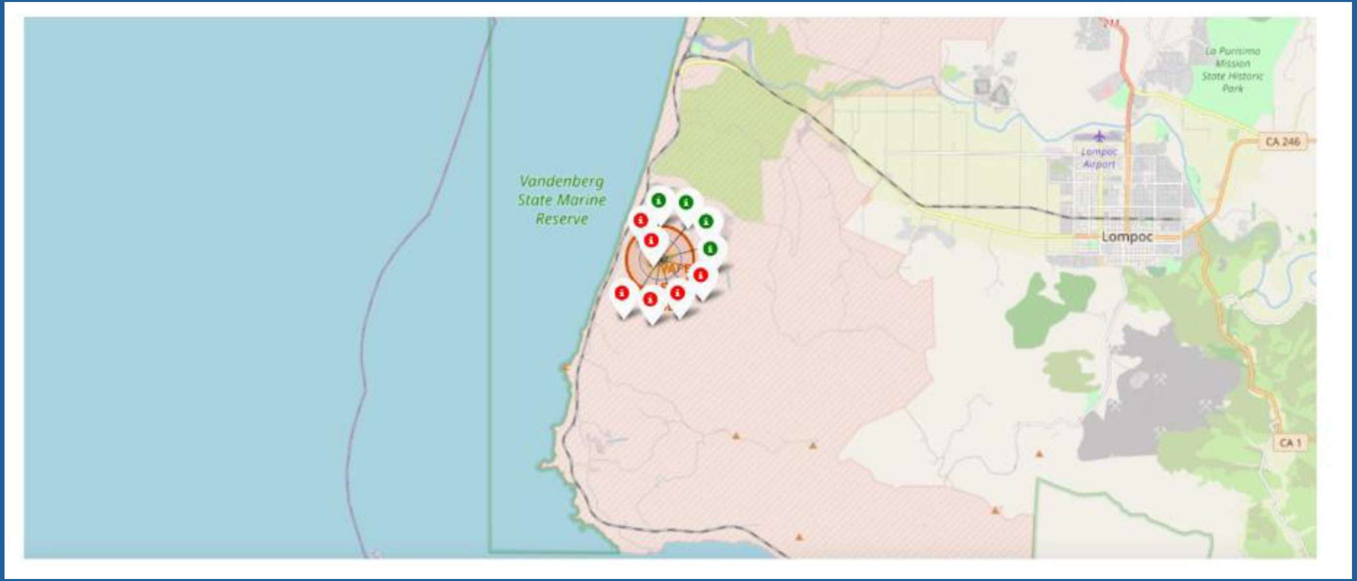


Space Launch Complex 3
Space Launch Complex
10
VAFB
SLC-
4E
1
34 KM
Vandenberg State Marine Reserve
Santa Barbara Subdivision MT1

RESULTS

Folium

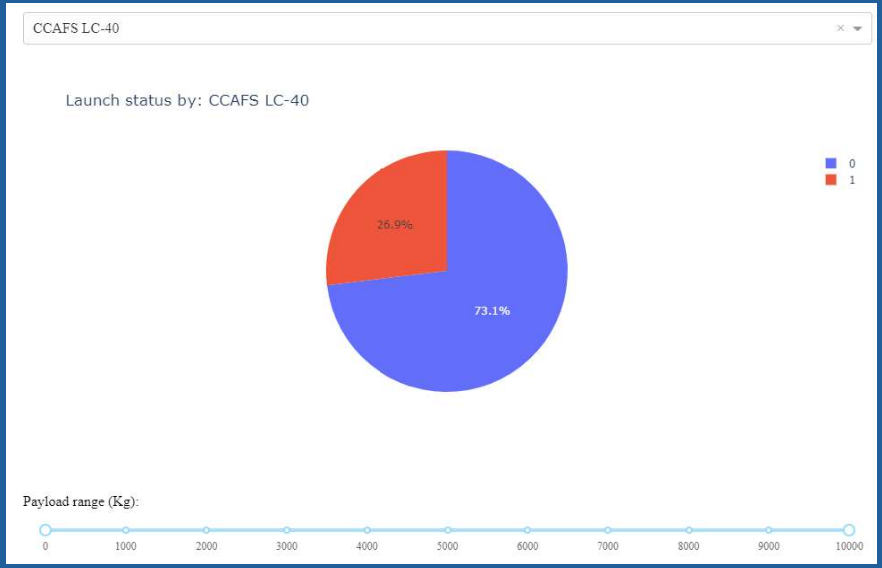
- The succeeded launches and failed launches for each site on map.


La Purisima Mission State Historic Park
CA 246
Lompoc Airport
Lompoc
Vandenberg State Marine Reserve
CA 1

RESULTS

Dash

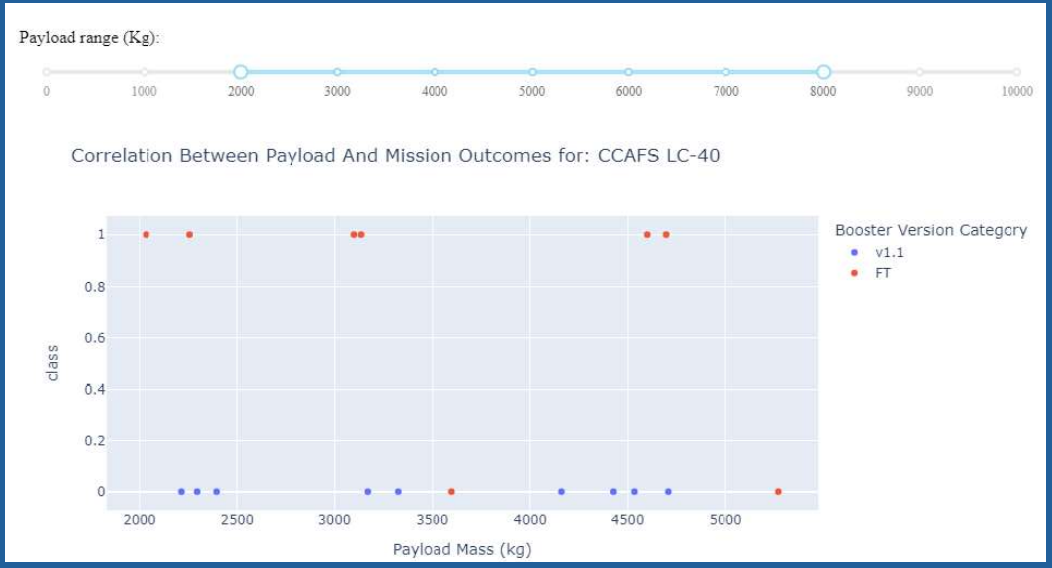
- The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.



CCAFS LC-40

Payload range (Kg):

- The picture below shows a scatter plot when the payload mass range is set to be from 2000 kg to 8000 kg



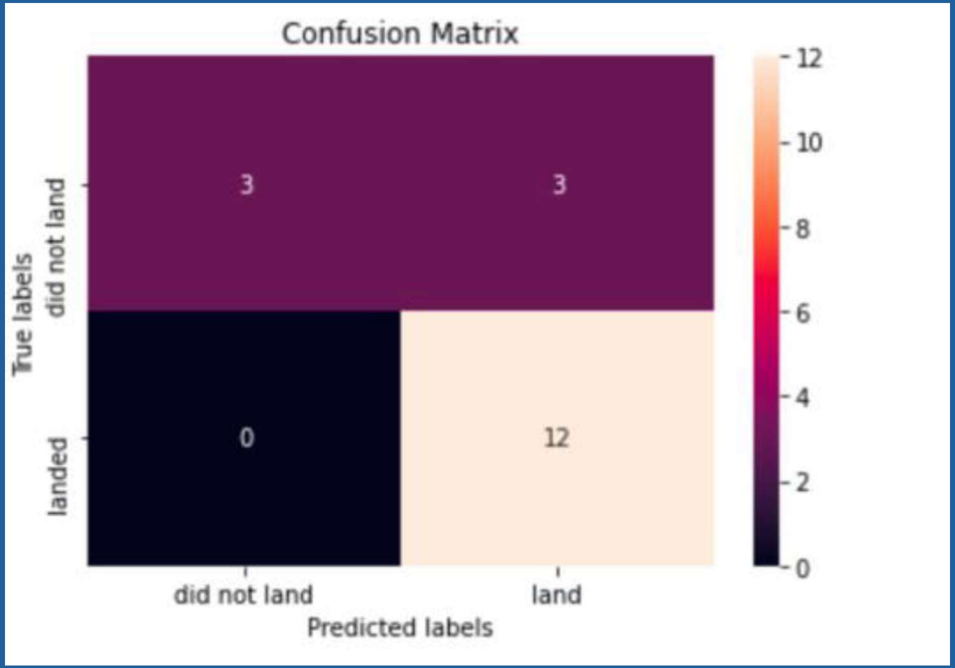
Payload range (Kg):

RESULTS

Predictive Analysis

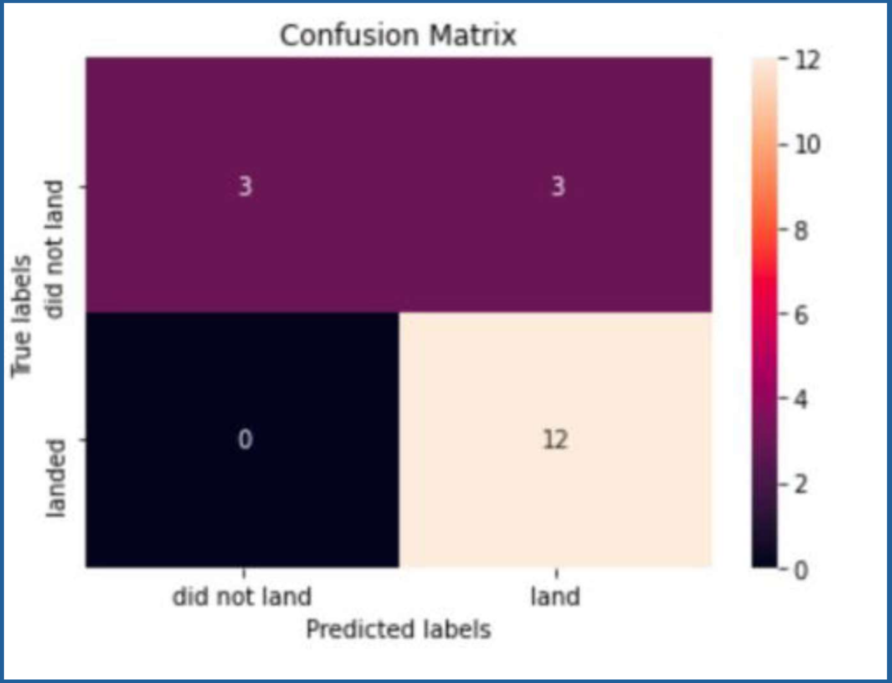
Logistic Regression

- GridsearchCV best score: 0.8464285714285713
- Accuracy score on test set: 0.8333333333333334
- Confusion Matrix:



Support Vector Machine (SVM)

- GridsearchCV best score: 0.8482142857142856
- Accuracy score on test set: 0.8333333333333334
- Confusion Matrix:

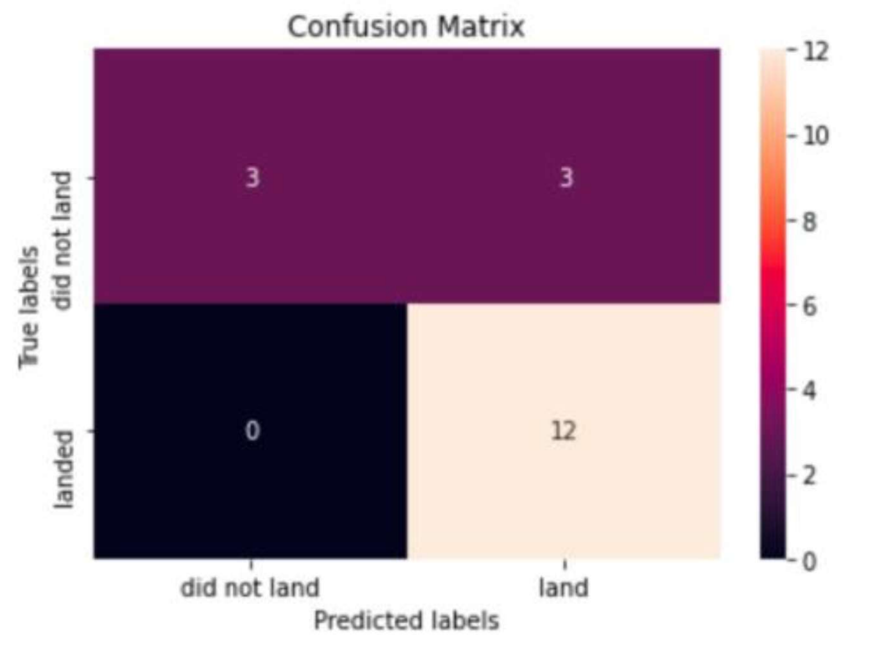


RESULTS

Predictive Analysis

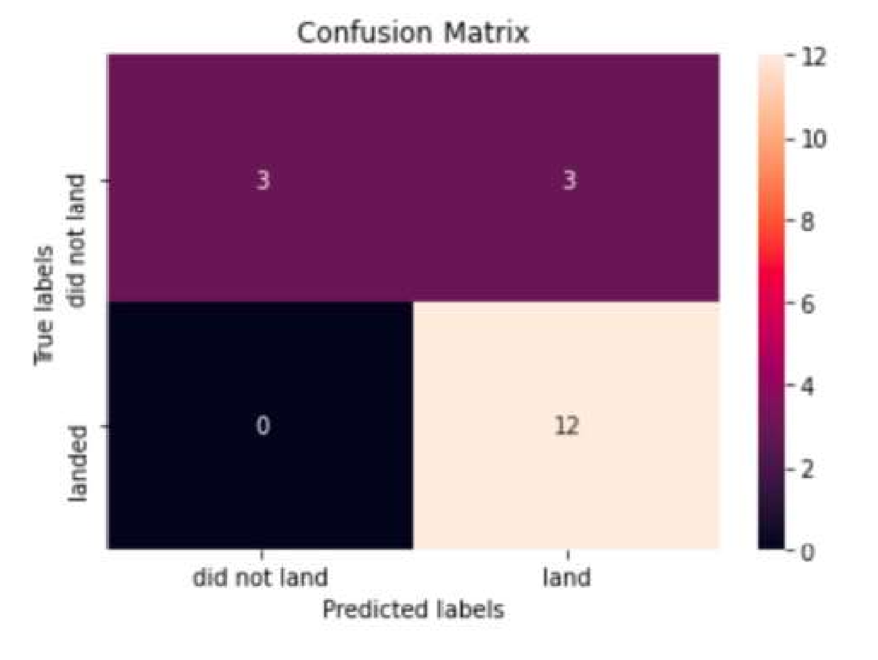
Decision Tree

- GridsearchCV best score: 0.8892857142857142
- Accuracy score on test set: 0.8333333333333334
- Confusion Matrix:



K Nearest neighbors

- GridsearchCV best score: 0.8482142857142858
- Accuracy score on test set: 0.8333333333333334
- Confusion Matrix:



RESULTS

Predictive Analysis

Putting the results of all the 4 models side by side, we can see that they all share the same accuracy and confusion matrix when tested on the test set.

Therefore, their GridsearchCV best scores are used to rank them instead. Based on the GridsearchCV best scores, the models are ranked in the following order with the first being the best and the last one being the worst:

1. Decision Tree: 0.8892857142857142
2. Logistic Regression: 0.8464285714285713
3. Support Vector Machine (SVM): 0.8482142857142856
4. K Nearest neighbors: 0.8482142857142858

DISCUSSION

- From the data visualization section, we can see that some features may have correlation with the mission outcome in several ways.
- Therefore, each feature may have a certain impact on the final mission outcome. The exact ways of how each of these features impact the mission outcome are difficult. However, we can use some machine learning algorithms to learn the pattern of the past data and predict whether a mission will be successful or not based on the given features.

CONCLUSION

- In this project, we try to predict if the first stage of a given Falcon 9 will land in order to determine the cost of a launch.
- Each feature of a Falcon 9launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way.
- Several machine learning algorithm are employed to learn the patterns of past Falcon 9 launch data to produce predictive models that can be used to predict the outcome of a Falcon 9 launch.
- The predictive model produced by decision tree algorithm performed the best among the 4 machine learning algorithms employed.