IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Name: Shweta JS

Date: 12-05-2023

Outline

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

Executive Summary

In this capstone project, we will make a prediction on how well the SpaceX Falcon 9 first stage will land. The price of a launch can be calculated if we can know if the first stage will land. Different machine learning classification techniques will be used to achieve this.

Data collection, data wrangling and preprocessing, exploratory data analysis, data visualisation, and machine learning prediction will all be part of the methodology used.

The findings of our inquiry and analysis suggest that there are specific characteristics of rocket launches that are correlated with successful or unsuccessful launches.

The Decision Tree may be the best machine learning algorithm for this task, we conclude.

Introduction

This capstone project's major objective is to foretell if the Falcon 9 first stage will successfully land. SpaceX advertises on their website that their rocket launches cost 62 million while other providers charge upwards of 165 million because they take great pride in being able to reuse the first stage of a rocket launch. The reuse of the first stage is largely responsible for these cost savings. The price of a launch can be calculated if we can know if the first stage will land. If a different business want to compete with SpaceX for a rocket launch, it may use the information provided here.

Section
1

Methodology

Methodology

Executive Summary

Data was gathered using two techniques: web scraping launch information from a Wikipedia article and requesting information from the SpaceX API. The data was then transformed and cleaned using the pandas module in Python.

Exploratory data analysis (EDA) was done on the clean data utilising visualisation tools including Python's matplotlib and seaborn packages, as well as SQL queries to provide answers. In order to respond to various analytical queries, interactive visualisation packages in Python were employed. Maps were produced using Folium, and interactive data visualisations with Plotly Dash.

For the prediction study, four alternative machine learning classification models were employed. The models utilised were decision tree classifier, logistic regression, support vector machines, and k-nearest neighbour. To choose the best model, each one was trained, adjusted, and tested.

Data Collection – SpaceX API

1. Request and parse the SpaceX launch data using the GET request

2. Normalize JSON response into a dataframe

3. Extract only useful columns using auxilary functions

4. Create new pandas dataframe from dictionary

5. Filter dataframe to only include Falcon 9 launches

6. Handle missing values

7. Export to CSV file

- GitHub URL: Data Collection API

Data Collection - Scraping

1. Request rocket launch data from its Wikipedia page

2. Extract all column/variable names from the HTML table header

3. Create a data frame by parsing the launch HTML tables

4. Export to CSV file

- GitHub URL: Data Collection With Web Scraping

Data Wrangling

1. Calculate the number of launches on each site

2. Calculate the number and occurrence of each orbit

3. Calculate the number and occurence of mission outcome per orbit type

4. Create a landing outcome label from Outcome column using one-hot encoding

5. Export to CSV

- GitHub URL: Exploratory Data Analysis

EDA with Data Visualization

- Scatter plots: To depict the relationship between two variables, scatter plots were utilised. There were comparisons between various feature sets, including Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type, and Payload vs. Orbit Type.

- Bar chart: The use of bar charts makes it simple to quickly compare values between several groupings. A discrete value is represented by the y-axis, while a category is represented by the x-axis. To compare the Success Rate for various Orbit Types, bar charts were employed.

- Line chart: Data patterns over time can be shown using line charts. To display the success rate over a certain number of years, a line chart was employed.

- GitHub URL: EDA with Data Visualisation

EDA with SQL

A list of some of the SQL queries performed on the dataset is listed below:

- Displaying the names of the distinctive launch sites in the space mission --- Displaying 5 records where launch sites start with the string "CCA" --- Displaying the overall payload mass carried by boosters launched by NASA (CRS) --- Displaying the average payload mass carried by booster version F9 v1.1,
- The names of the boosters with successful landing outcomes in drone ships and payload masses greater than 4000 but less than 6000, the total number of successful and unsuccessful mission outcomes, the names of the booster versions that carried the maximum payload mass, the failed landing outcomes in drone ships, their booster versions, and the launch site na decreasing order of the number of landing outcomes between 2010-06-04 and 2017-03-20.
- GitHub URL: EDA with SQL

Build an Interactive Map with Folium

The creation and addition of objects to a Folium map. All launch sites, as well as the successful and unsuccessful launches for each site, were displayed on a map using marker objects. The distances between a launch location and its environs were calculated using line objects.

- By adding these objects, following geographical patterns about launch sites are found:
    - Are launch sites in close proximity to railways? Yes
    - Are launch sites in close proximity to highways? Yes
    - Are launch sites in close proximity to coastline? Yes
    - Do launch sites keep certain distance away from cities? Yes
- GitHub URL : Interactive Map Analytics with Folium

Build a Dashboard with Plotly Dash

The dashboard application contains two charts:

- a pie chart displaying each site's successful launch. This graph is helpful since it allows you to see the success rate of launches on specific sites or visualise the distribution of landing results across all launch locations.

- a scatter diagram illustrating the relationship between landing success and the mass of various boosters. The site(s) and payload mass are the dashboard's two inputs. This chart is helpful since it allows you to see how different factors influence the results of the landing.

- GitHub URL : Space-X Dashboard

Predictive Analysis (Classification)

1. Create column for “Class”

2. Standardizing the data

3. Split ito training and test set

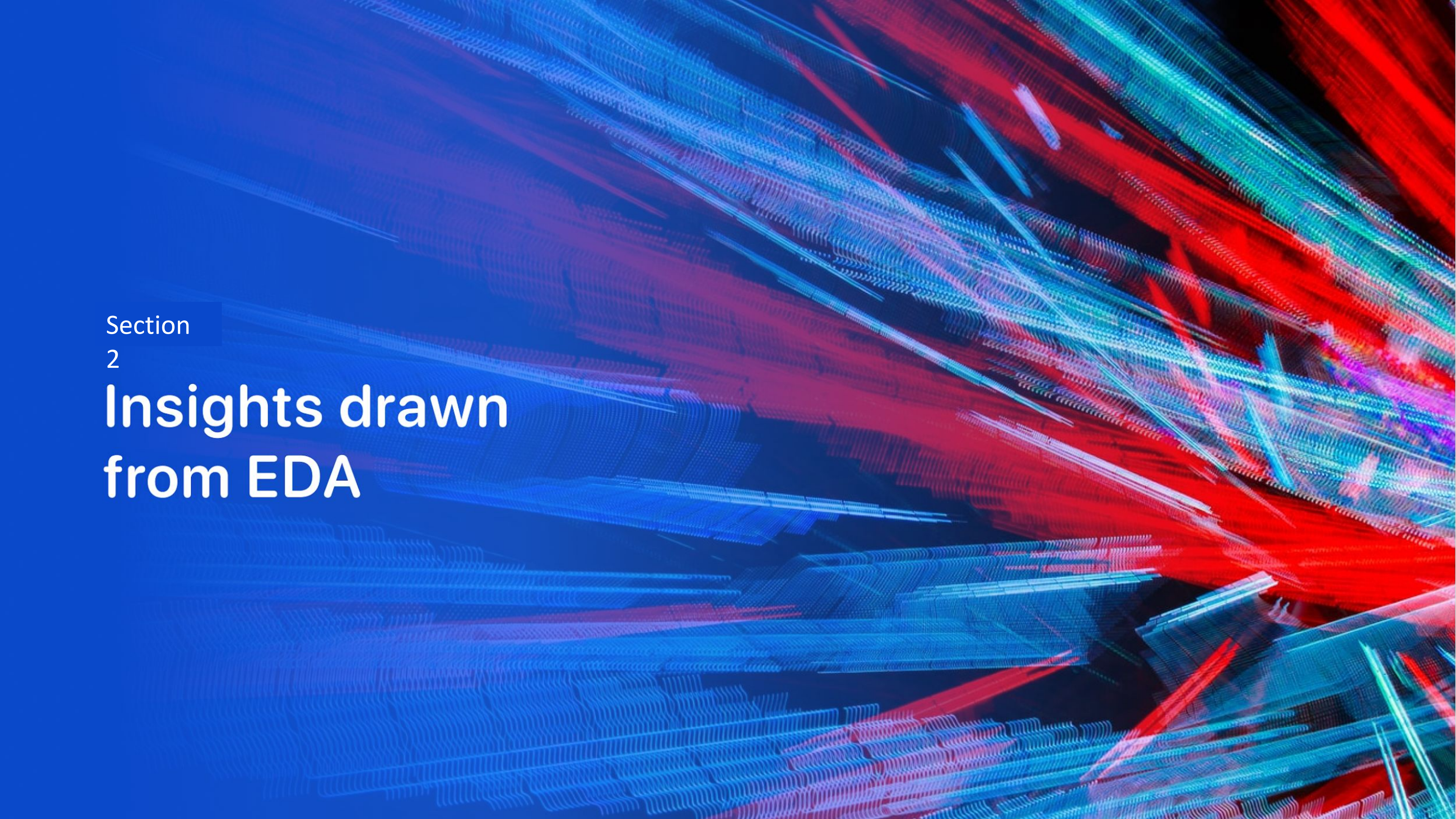
4. Find best Hyperparameter for SVM, Decision Trees, K-Nearest Neighbours and Logistic Regression.

5. Use test data to evaluate models based on their accuracy scores and confusion matrix

- GitHub URL: Space-X Machine Learning Prediction

Results

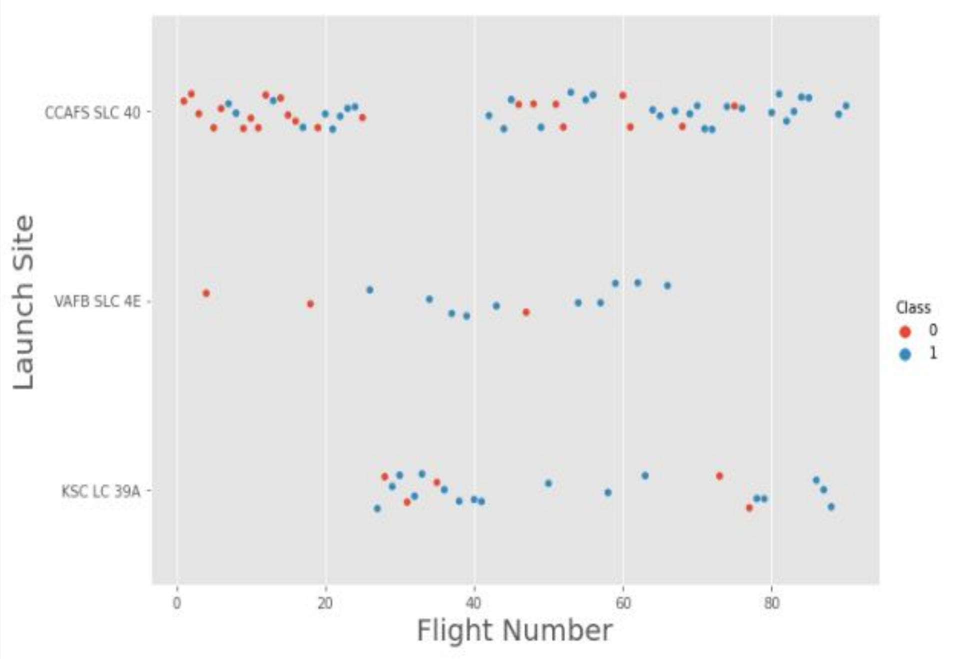
- The success rate of the Falcon 9 landings was 66.66%, according to the findings of the exploratory data analysis.
- The Decision Tree algorithm, which had a 94% accuracy rate, was the best classification technique, according to the results of the predictive analysis.



Section 2

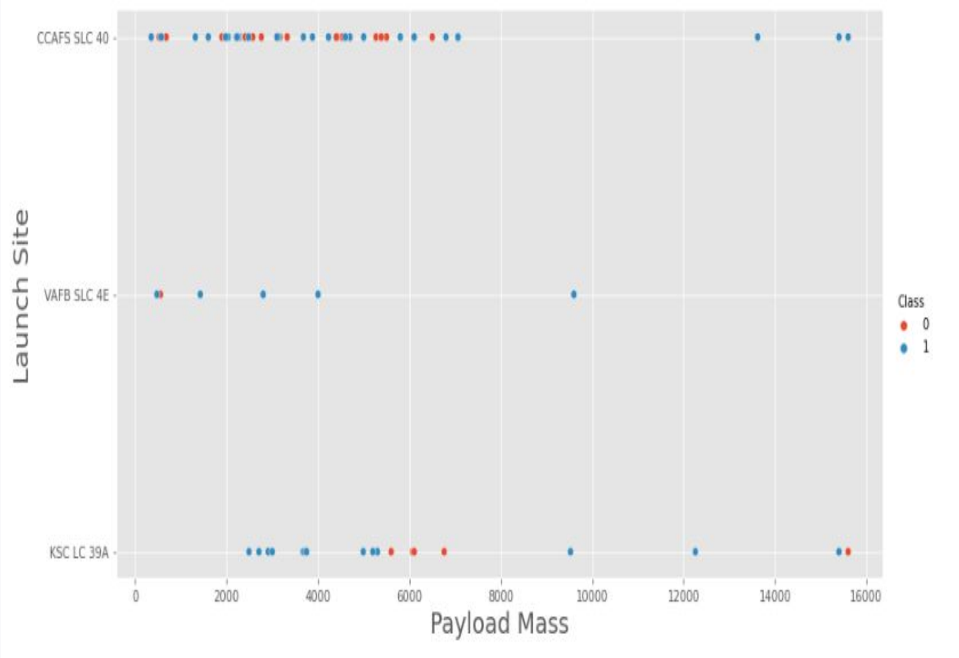
Insights drawn from EDA

Flight Number vs. Launch Site



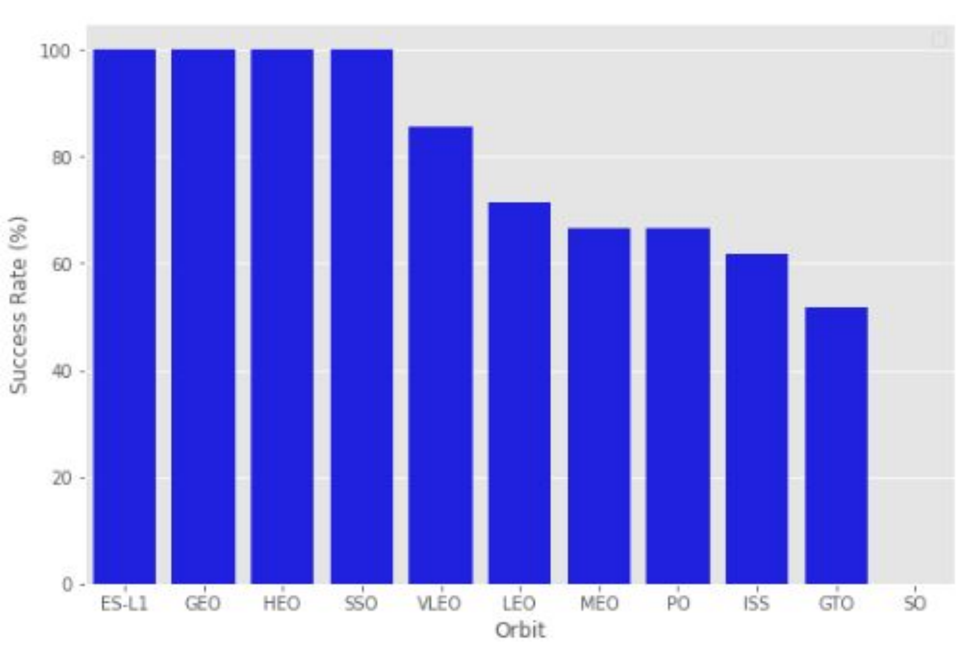
- This graph demonstrates that as the number of flights increased, so did the success rate.
- The successful launches are shown by blue dots, whereas the failed launches are represented by red dots.
- After the 40th launch, there seems to be an uptick in successful flights.

Payload vs. Launch Site



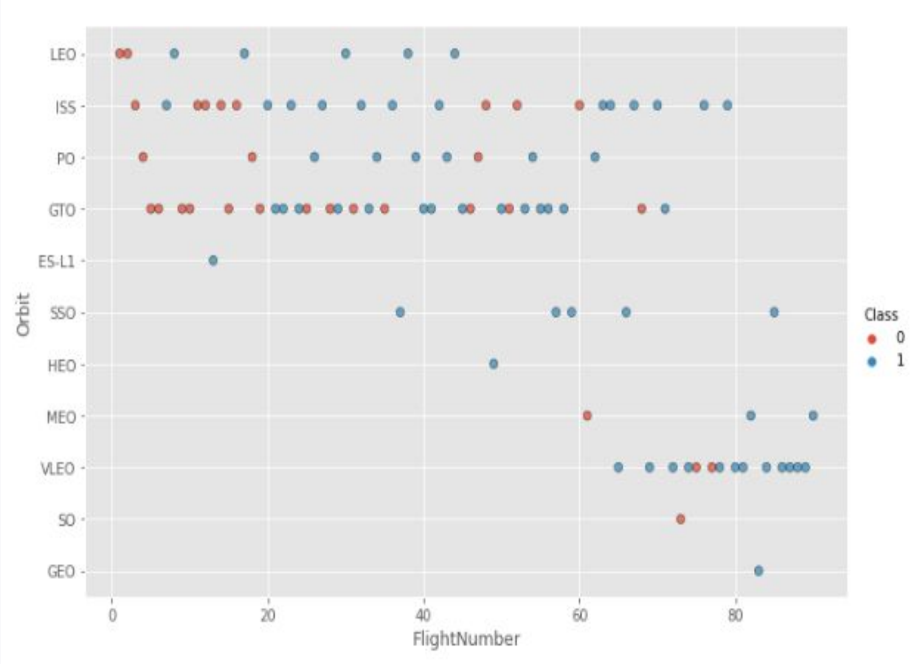
- The successful launches are shown by blue dots, whereas the failed launches are represented by red dots.
- There are no rockets launched for heavy payload mass from the VAFB-SLC launch site.
- Decisions cannot be made using this metric because there appears to be a poor association between Payload and Launch Site.

Success Rate vs. Orbit Type



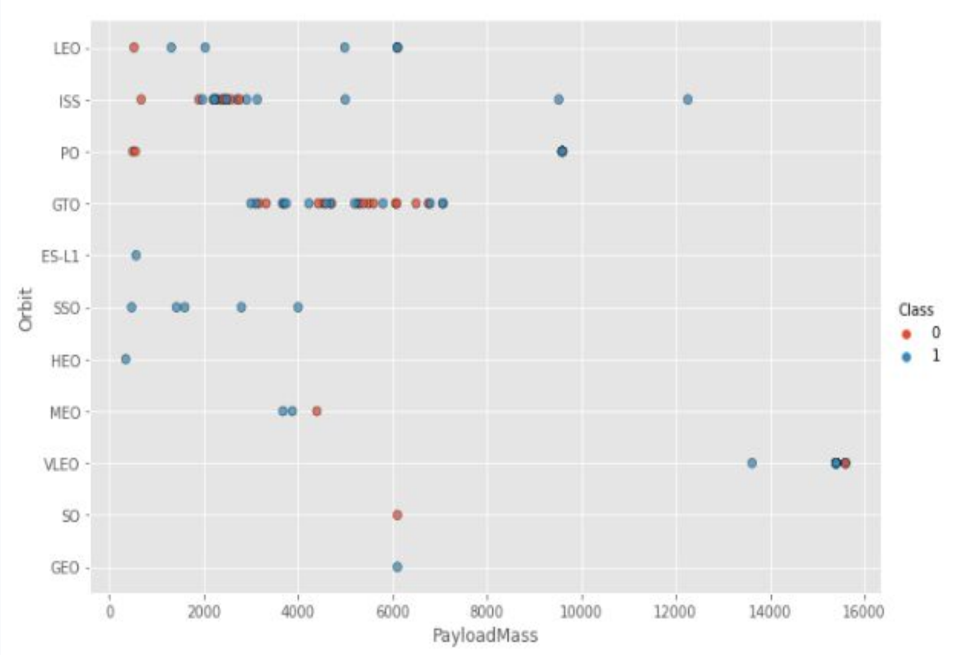
- SSO, HEO, GEO, and ES-L1 orbits have never failed.
- With a 0% success rate, SO orbit had no launches that were successful.

Flight Number vs. Orbit Type



- The number of trips is strongly connected with success in the LEO orbit.
- In the GTO orbit, there does not appear to be a correlation between flight numbers.
- The SSO orbit has fewer flights than the other orbits but a success record of 100%.
- The success rate of flights with numbers over 40 is higher than that of flights with numbers between 0 and 40.

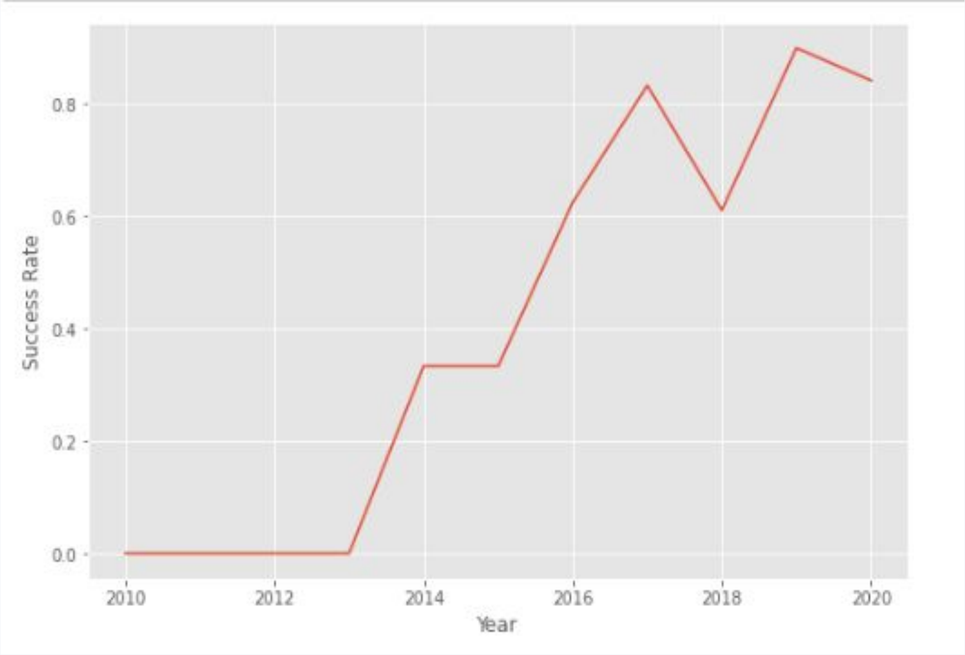
Payload vs. Orbit Type



- As the payloads get heavier, the success rate increases in the PO, SSO, LEO and ISS orbits.

- There seems to be no direct correlation between obrit type and payload mass for GTO orbit as both successful and failed launches are eually present

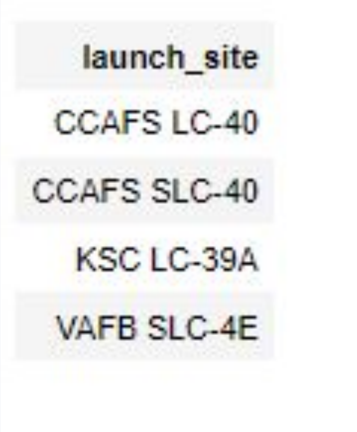
Launch Success Yearly Trend



- As the years go by, the chart's overall trend indicates an increase in landing success rate. However, both 2018 and 2020 see a decline.

All Launch Site Names

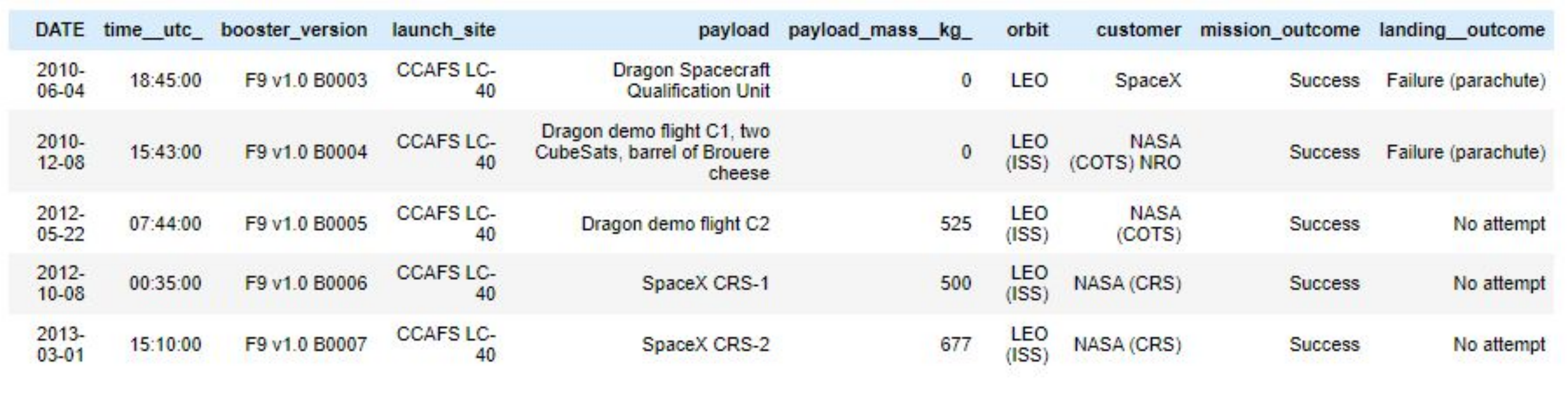
- The DISTINCT clause was used to return only the unique rows from the launch_site column.
- The launch locations are known by the designations CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.



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

Launch Site Names Begin with 'CCA'

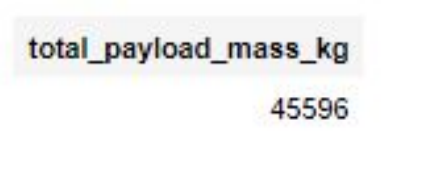
- Only the first five entries where the launch_site name begins with "CCA" were shown using the LIMIT and LIKE clauses.



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

Total Payload Mass

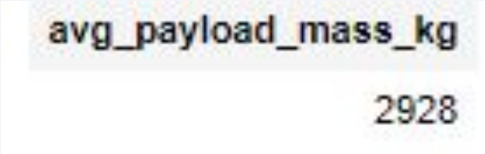
- The payload_mass__kg field was used to compute the overall payload that NASA boosters carried using the SUM() algorithm.



total_payload_mass_kg
45596

Average Payload Mass by F9 v1.1

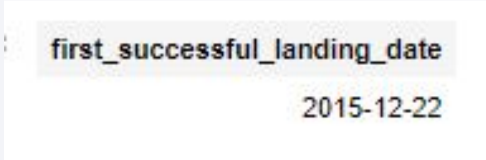
- The average payload mass carried by booster version F9 v1.1 was calculated using the AVG() tool.
- The computations were limited to booster_versions that were named "F9 v1.1" by using the WHERE clause to filter the results.



avg_payload_mass_kg
2928

First Successful Ground Landing Date

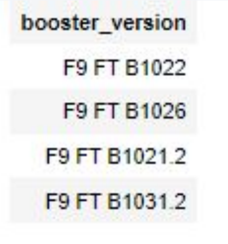
- The date of the first successful landing attempt on the ground pad was determined using the MIN(DATE) method.
- The WHERE clause made sure that the results were filtered to only match when the 'landing_outcome' column was set to 'Success (ground pad)'



first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

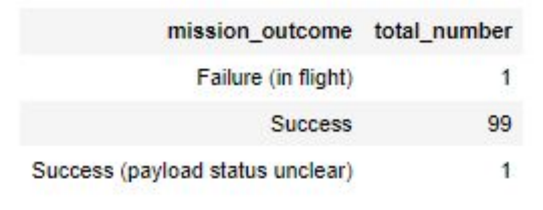
- Using the BETWEEN clause, only results with a payload mass of at least 4,000 but not more than 6,000 were returned. Only rockets that successfully landed on the drone ship were included in the results after being filtered by the WHERE clause.



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

Total Number of Successful and Failure Mission Outcomes

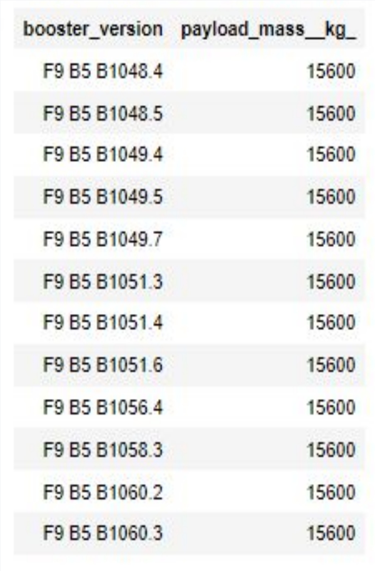
- With the use of the GROUPBY clause applied to the'mission_outcome' column, the COUNT() function is used to count the number of instances of various mission outcomes. The total number of mission outcomes—both successful and unsuccessful—is returned.
- Out of 101 missions, 99 have resulted in successful missions.



mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

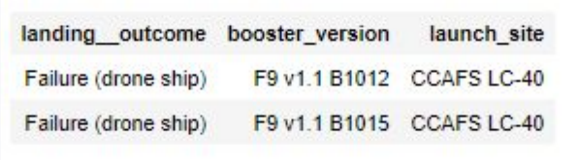
- A list of boosters that have transported the largest payload mass was retrieved using the MAX() function in a subquery.



booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

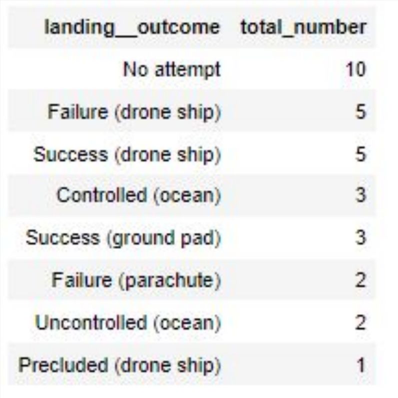
- The table's many columns were retrieved using the SELECT query. The ONLY entries with a launch date of 2015 were retrieved using the YEAR(DATE) method.



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

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

- To tally the various landing results, the tally() method was utilised.Only results between 2010-06-04 and 2017-03-20 were included in the results thanks to the WHERE and BETWEEN clauses. The GROUPBY phrase made sure the counts were grouped according to how they turned out. The results were arranged in descending order using the ORDERBY and DESC clauses.

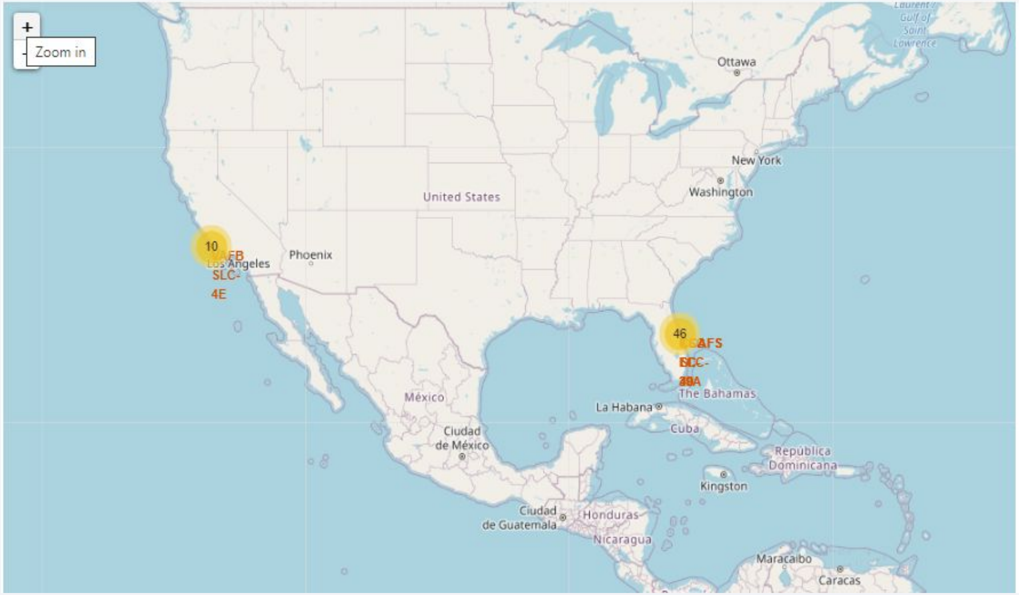


landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

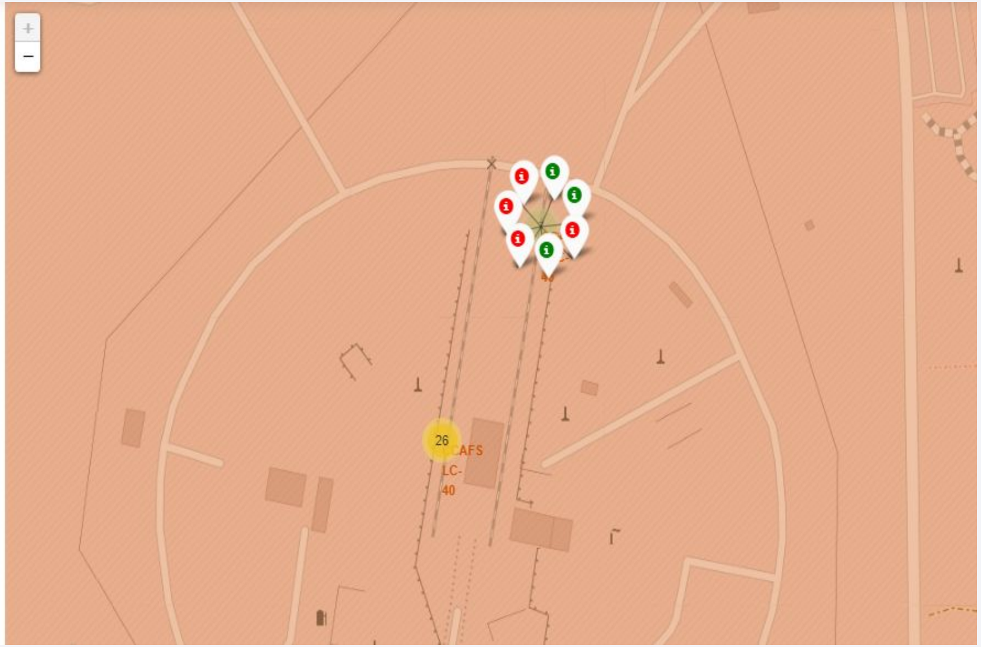
Launch Sites Proximities Analysis

SpaceX Launch Sites Locations



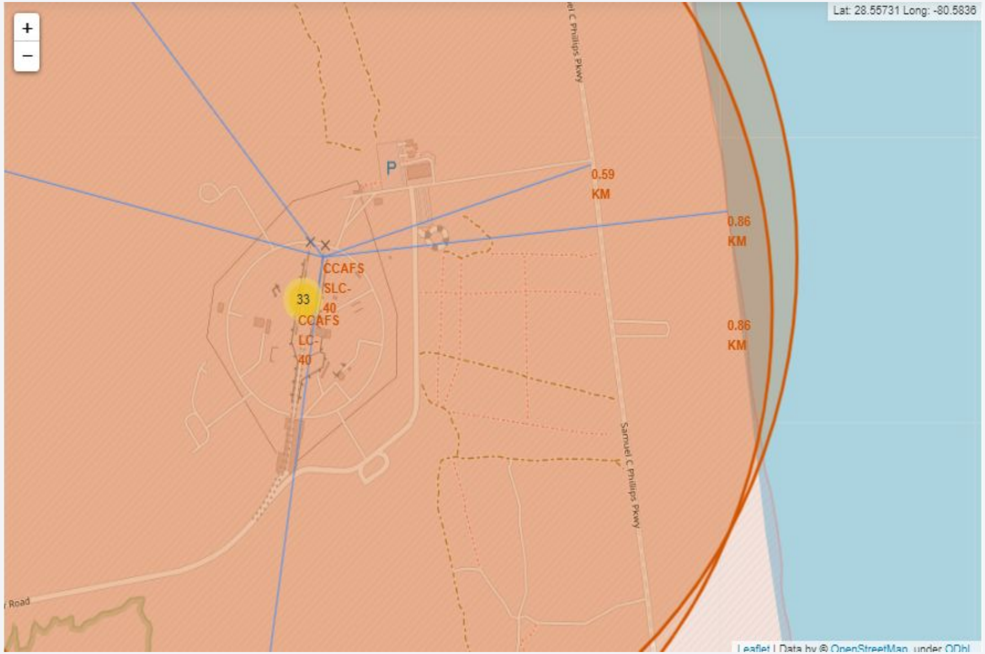
- The locations of all the SpaceX launch sites in the US are indicated by the yellow markers.
- The launch pads have been situated in a suitable location close to the coast.

Success or Failure?



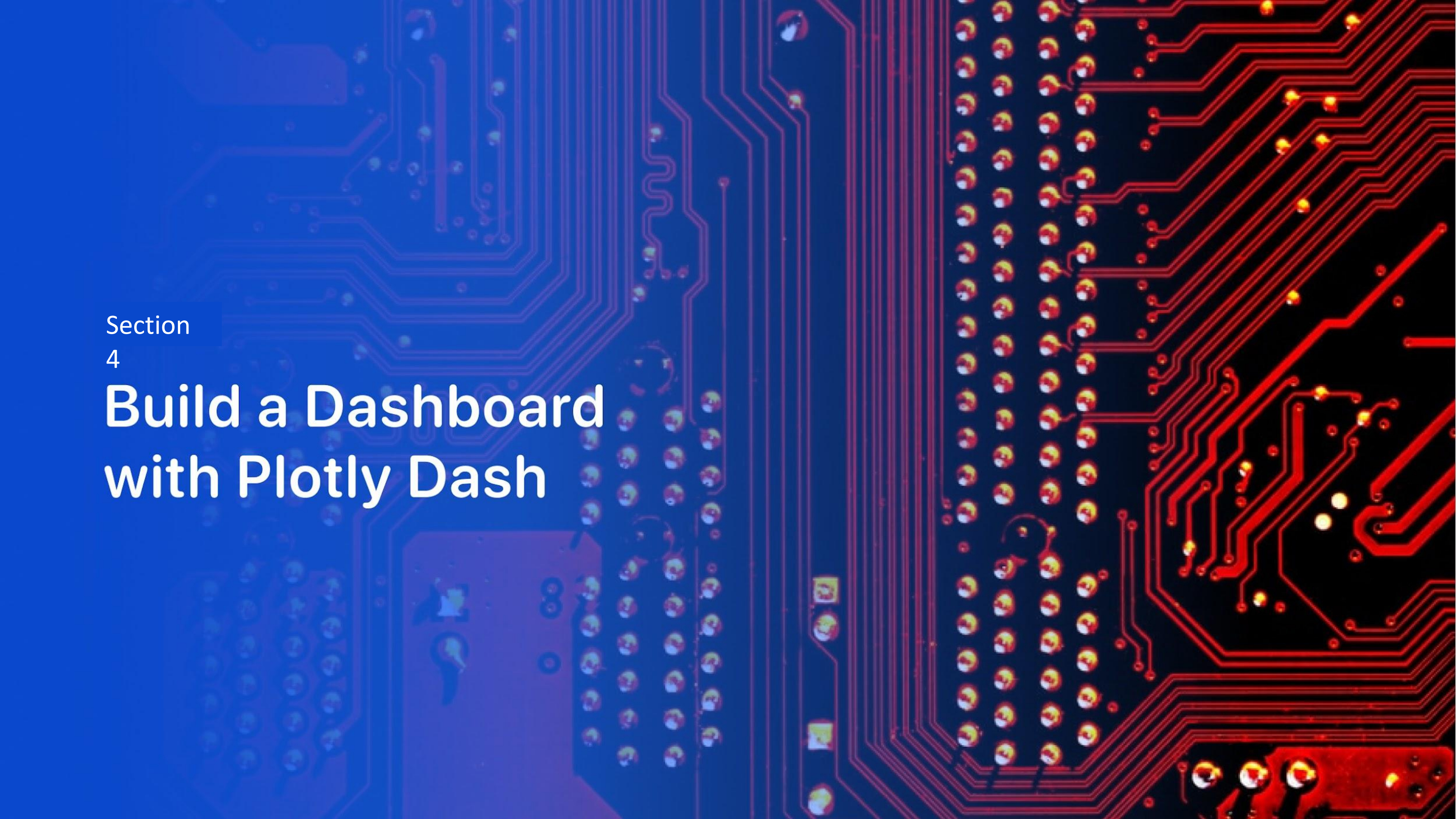
When we zoom in on a launch point, we can click there to bring up marker clusters of successful or unsuccessful landings (green or red, respectively).

Launch Site Proximities



- The generated map shows that the selected launch site is close to a highway for transportation of personnel and equipment. The launch site is also close to the coastlines for launch failure testing.

- The launch sites also maintain a certain distance from the cities. (Can be viewed in notebook).

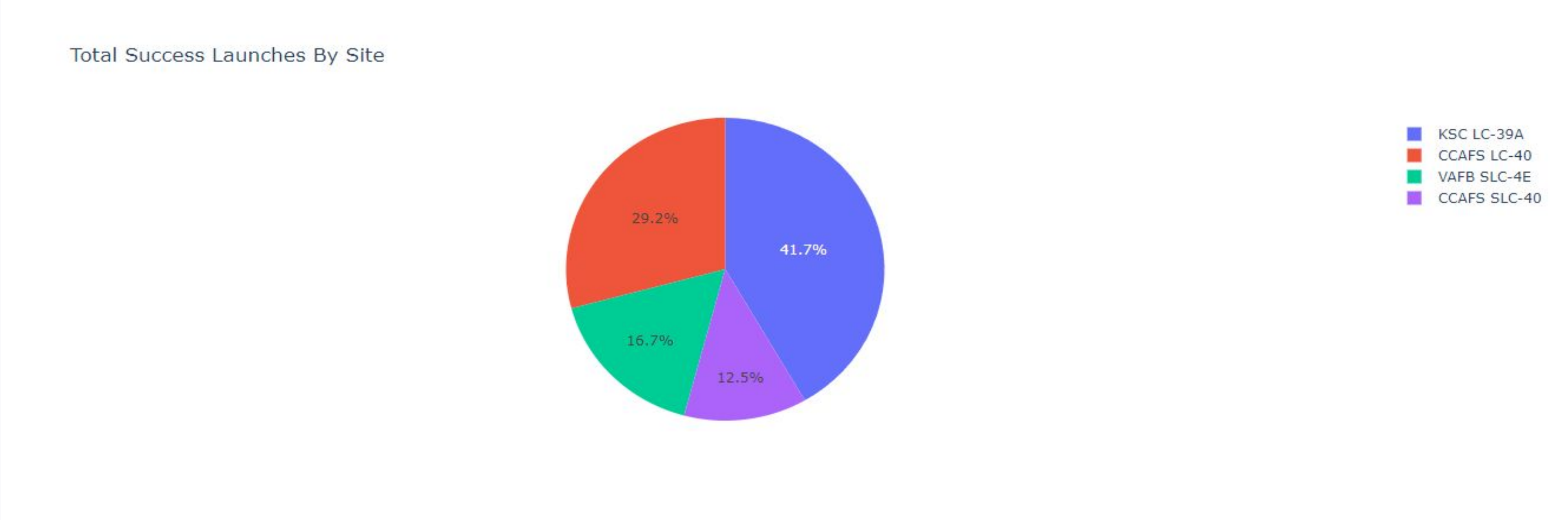


Section
4

Build a Dashboard with Plotly Dash

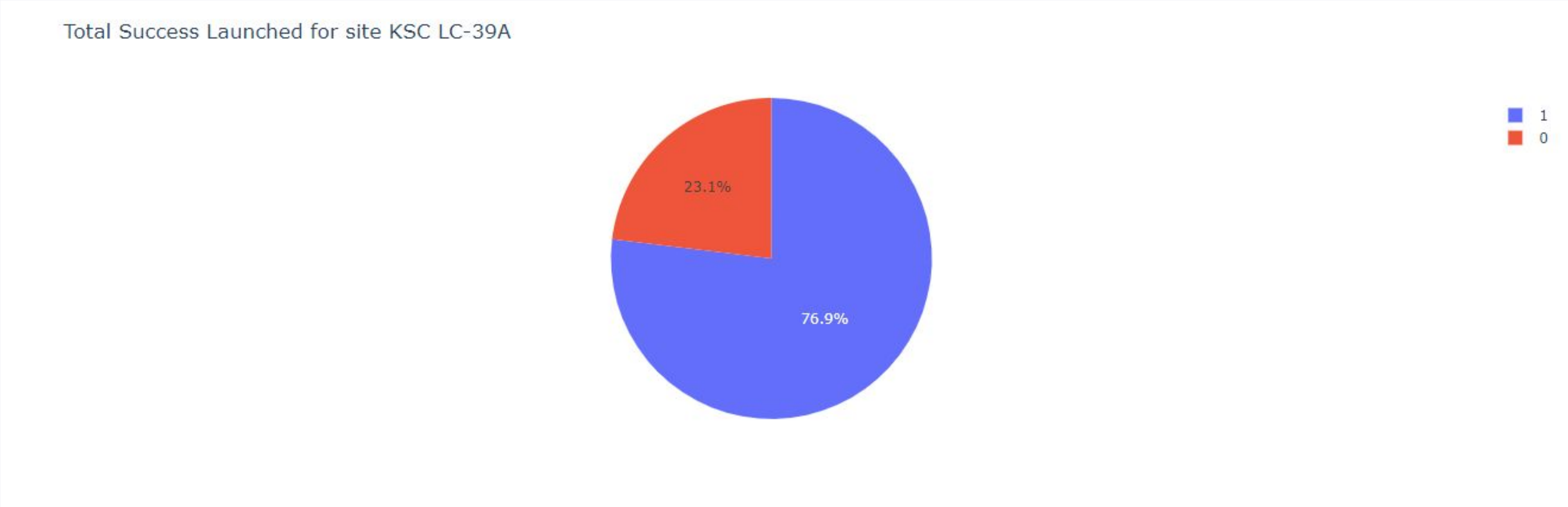
Total Successful Launches By Site

- The KSC LC-39A Launch site has the most successful launches with 10 in total.



Launch Site With Highest Success Ratio

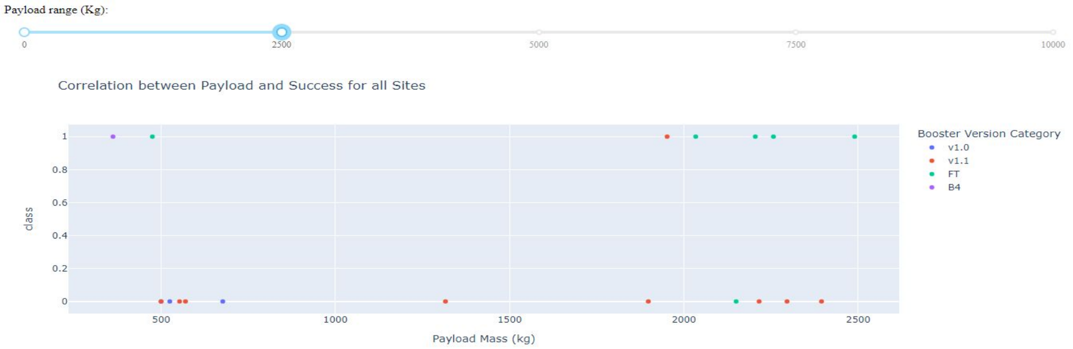
- The KSLC-39A has the highest success rate with 76.9%.

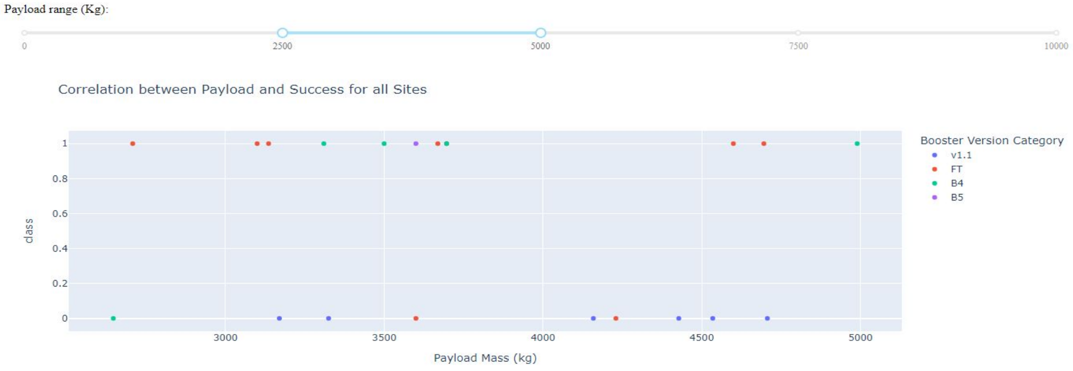


Payloads vs Launch Outcome

- Payloads between 0 and 2500 kg have a somewhat lower launch success percentage than payloads between 250 and 5000 kg. Actually, there isn't much of a distinction between the two.

- In both weight ranges, the v1.1 booster version had the highest effectiveness rate.



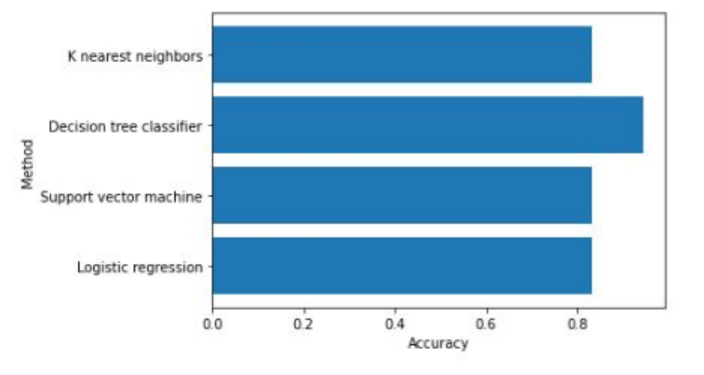


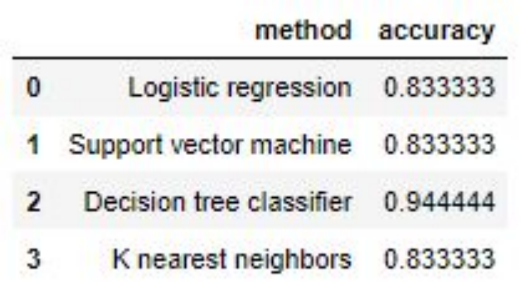
Section
5

Predictive Analysis (Classification)

Classification Accuracy

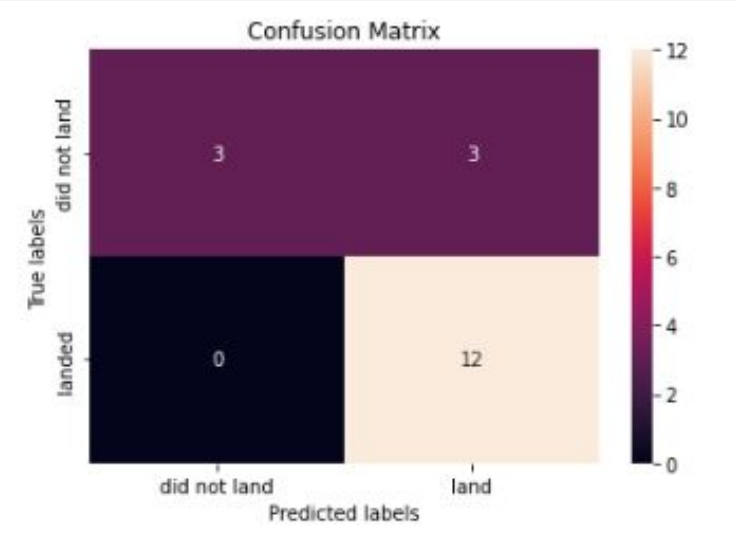
- The Decision Tree classifier had the best accuracy at 94%.





	method	accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.944444
3	K nearest neighbors	0.833333

Confusion Matrix



- When the True label was success (True Positive), the model predicted 12 successful landings, and when the True label was failure (True Negative), it predicted 3 unsuccessful landings.
- When the True label indicated a failed landing (False Positive), the model also predicted three successful landings.
- Successful landings were frequently predicted by the model.

Conclusions

- The analysis revealed that as the success rate has increased over time, there is a positive correlation between the number of flights and success rate.
- The most successful launches occurred in orbits like SSO, HEO, GEO, and ES-L1.
- Payload mass can be related to success rate since lighter payloads have typically had better results than bigger payloads.
- The launch sites are placed in a safe distance from cities but strategically close to roads and railroads for the transit of people and goods.
- The Decision Tree Classifier is the best prediction model to employ for this dataset as it has the highest accuracy (94%)..

Appendix

- Coursera Project Link: https://www.coursera.org/learn/applied-data-science-capstone/home/welcome
- GitHub Repository: https://github.com/shweta-js/ibm_ds_capstone

Thank you!