IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ajibose G Okanlawon
21/03/2023

Outline

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

Executive Summary

We will forecast whether the SpaceX Falcon 9 first stage will successfully land in this capstone assignment. The price of a launch can be calculated if we can know if the first stage will land.

Several 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 included in the methodology.

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

Hence, we draw the conclusion that the Decision Tree may be the most effective machine learning approach for this issue.

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.

This gets us to the key inquiry we are attempting to address: Under the specified conditions of a Falcon 9 rocket launch, will the rocket's first stage successfully land or not?

Section 1

Methodology

Methodology

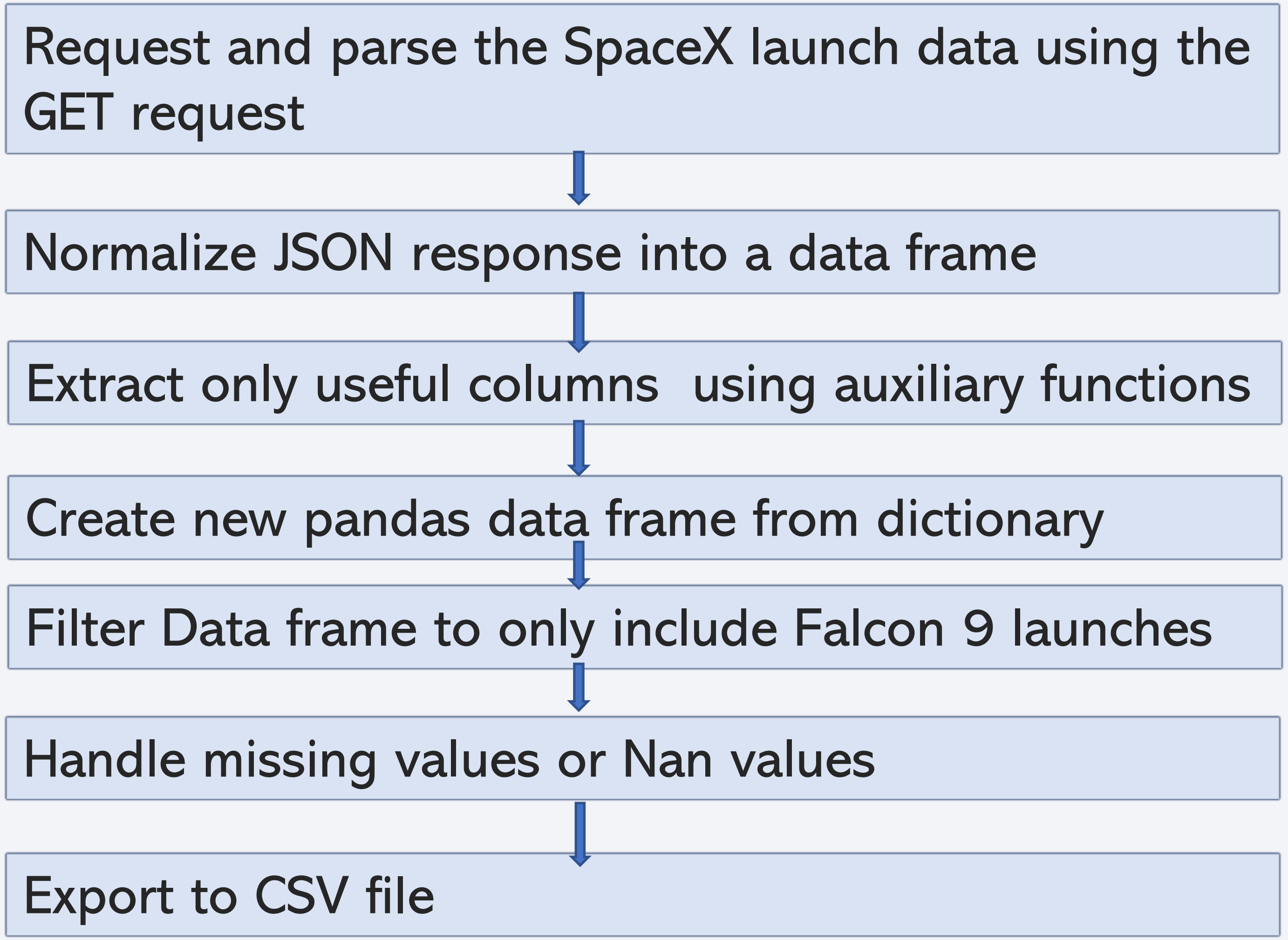
Executive Summary

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 some 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. Support vector machines, logistic regression, k-nearest neighbour, and decision tree classifier are the models that were applied. Each model was trained and evaluated.

Data Collection- Using SpaceX API



Request and parse the SpaceX launch data using the GET request

Normalize JSON response into a data frame

Extract only useful columns using auxiliary functions

Create new pandas data frame from dictionary

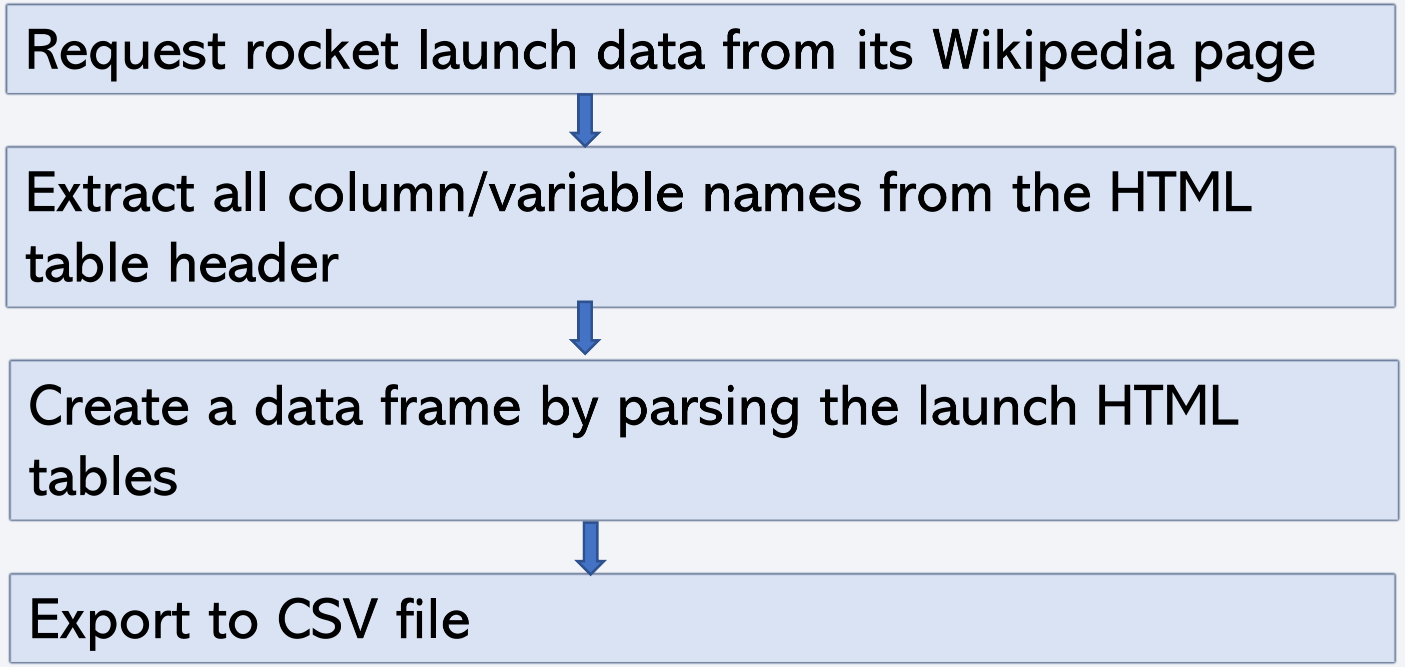
Filter Data frame to only include Falcon 9 launches

Handle missing values or Nan values

Export to CSV file

GitHub Link: Data Collection- Using SpaceX API

Data Collection- Using Web Scraping



Request rocket launch data from its Wikipedia page

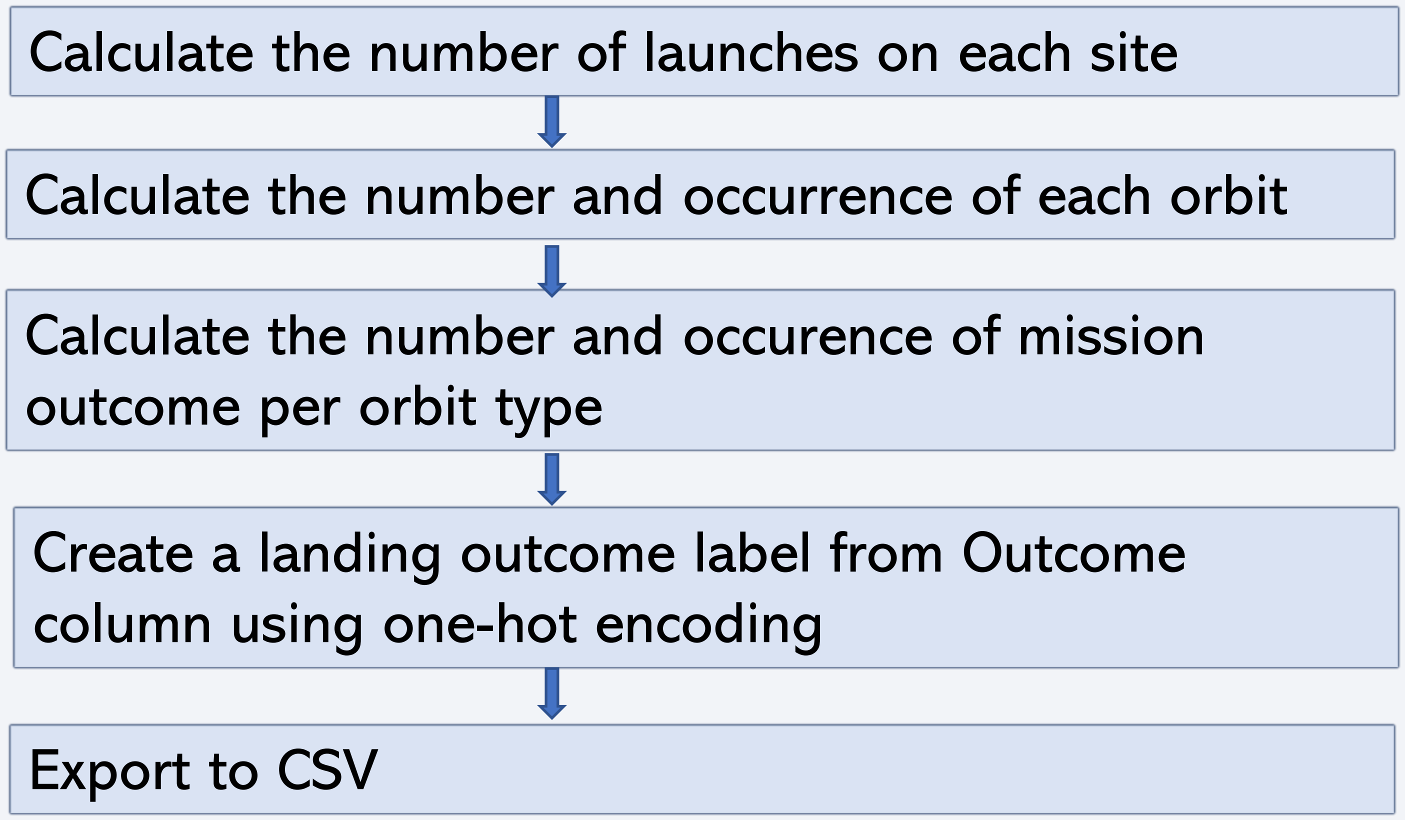
Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Export to CSV file

GitHub Link: Data Collection-Using Web Scraping

Data Wrangling



Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome per orbit type

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

Export to CSV

GitHub Link: Data Wrangling

EDA with Data Visualization

- The link between two variables was represented using scatter plots. varying sets of features The link between two variables was represented using scatter plots. As examples of comparisons, Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type, and Payload vs. Orbit Type were made.
- Bar chart: The usage of bar charts allows for quick comparison of values between various groups. The y-axis indicates a discrete number, while the x-axis represents a category. The Success Rate for various Orbit Types was compared using bar charts.
- Line graphs are useful for displaying data trends across time. The Success Rate over several years was displayed using a line chart.
- GitHub Link: EDA with Data Visualization

EDA with SQL

The following is a list of some of the SQL queries that were run on the dataset:

- Displaying the names of the distinct launch sites used in the space mission; Displaying 5 records where the first letter of the launch site is "CCA“; Displaying the total mass of payloads carried by NASA-launched boosters (CRS); Displaying the average mass of payloads carried by booster version F9 v1.1.

- The names of the boosters that have succeeded in drone ships and have 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 have carried the maximum payload mass; the date when the first successful landing outcome in a ground pad was achieved; a list of some of the SQL operations made on the database Listing the drone ship unsuccessful landing outcomes for the year 2015, their booster versions, and launch locations —- Ranking the number of landing outcomes between 2010-06-04 and 2017-03-20, from most to least successful.

- GitHub Link: 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.
- The following geographic patterns about launch places are discovered by including these objects:

✓ The train lines close to the launch sites.

✓ The motorways are nearby the launch sites

✓ Launch sites are near to a shore

✓ Launch sites maintain a specific distance from cities

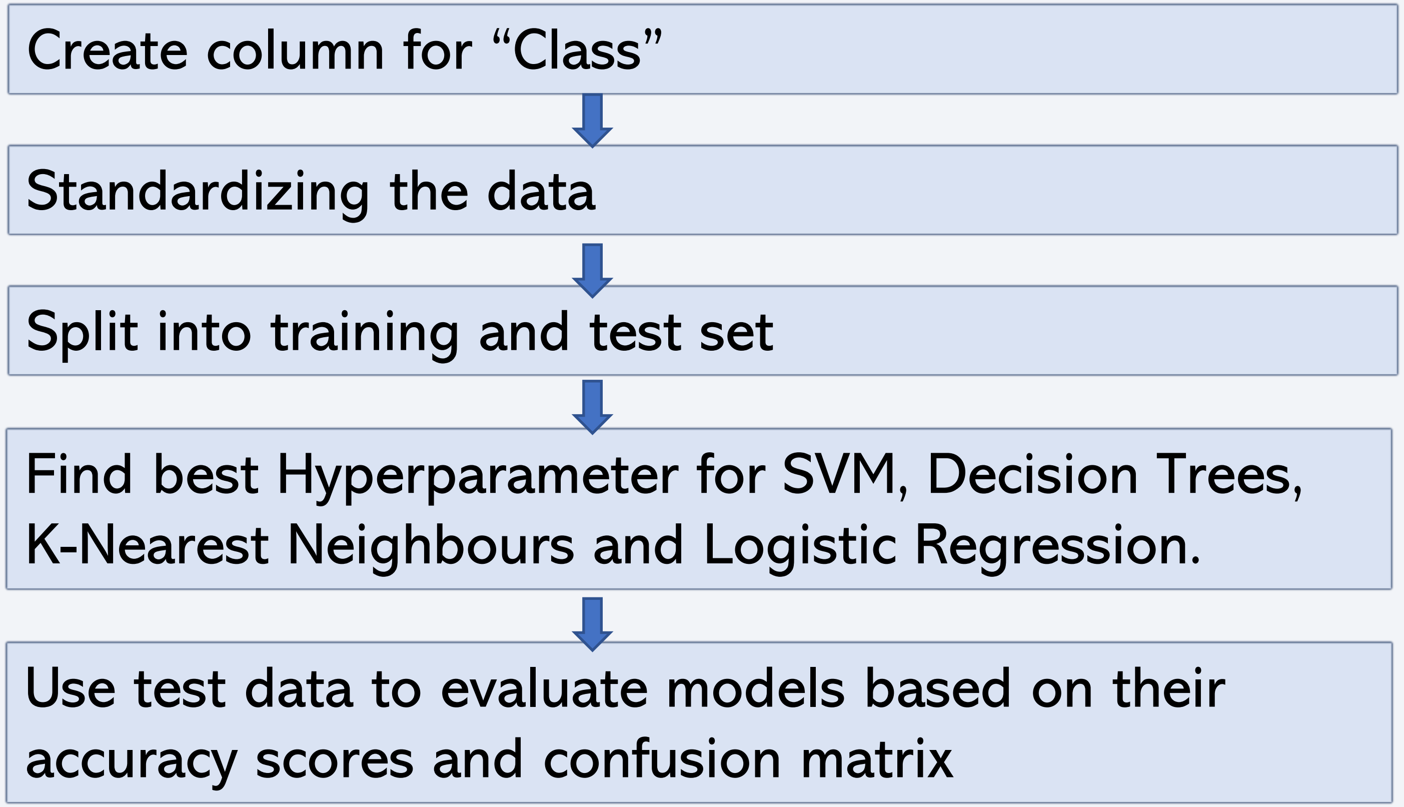
- GitHub Link: Building an Interactive Map with Folium

Build a Dashboard with Plotly Dash

Two graphs are included in the dashboard application:

- A pie chart displaying each site's successful launch. This graph is helpful since it allows you to display the success rate of launches on specific sites or visualise the distribution of landing outcomes 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 Link: Building a Dashboard with Plotly Dash

Predictive Analysis (Classification)



Create column for “Class”

Standardizing the data

Split into training and test set

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

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

- GitHub Link: Space-X Machine Learning Prediction

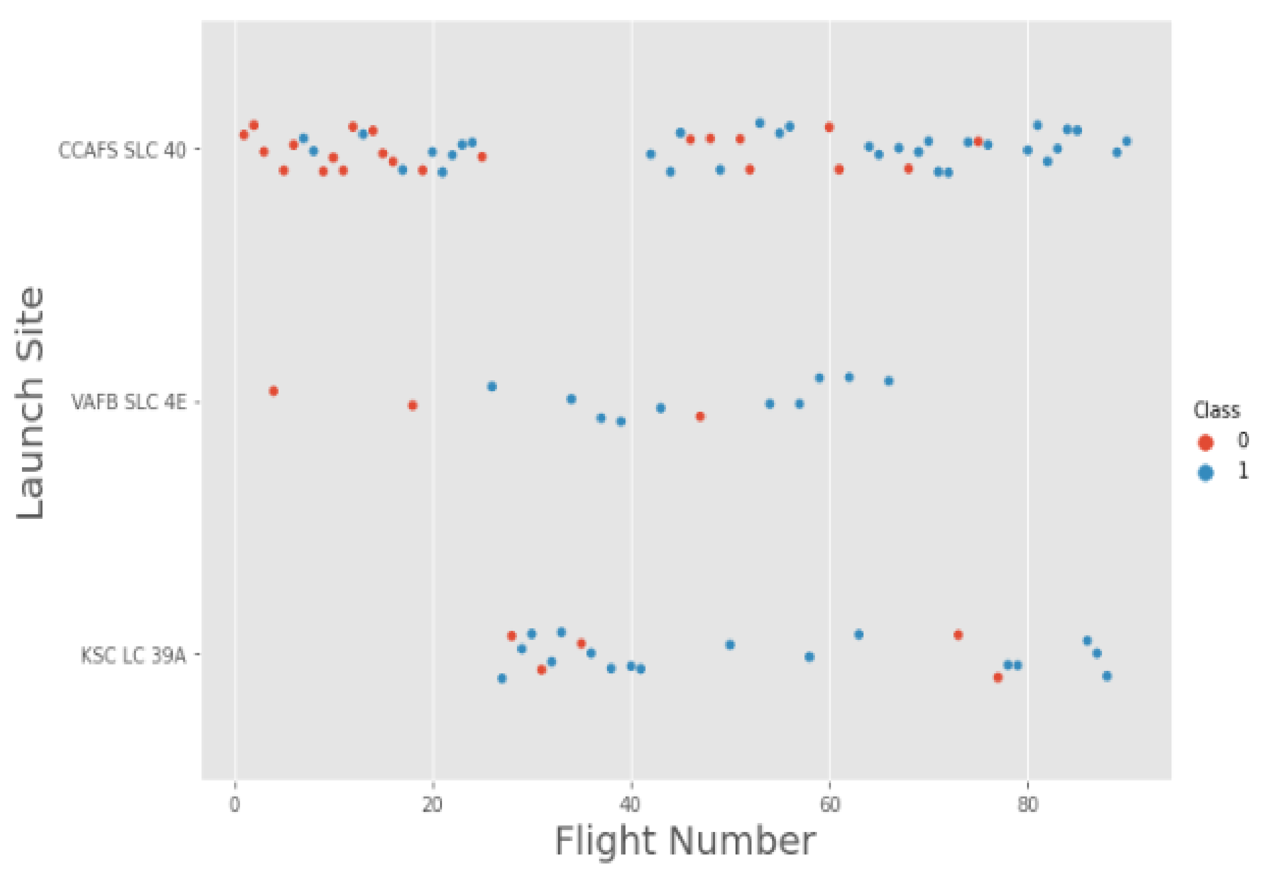
Results

- The outcomes of the exploratory data analysis showed that 66.66% of Falcon 9 landing attempts were successful.
- The Decision Tree algorithm was the best classification technique, with a 94% accuracy rate, according to the results of the predictive study.

Section 2

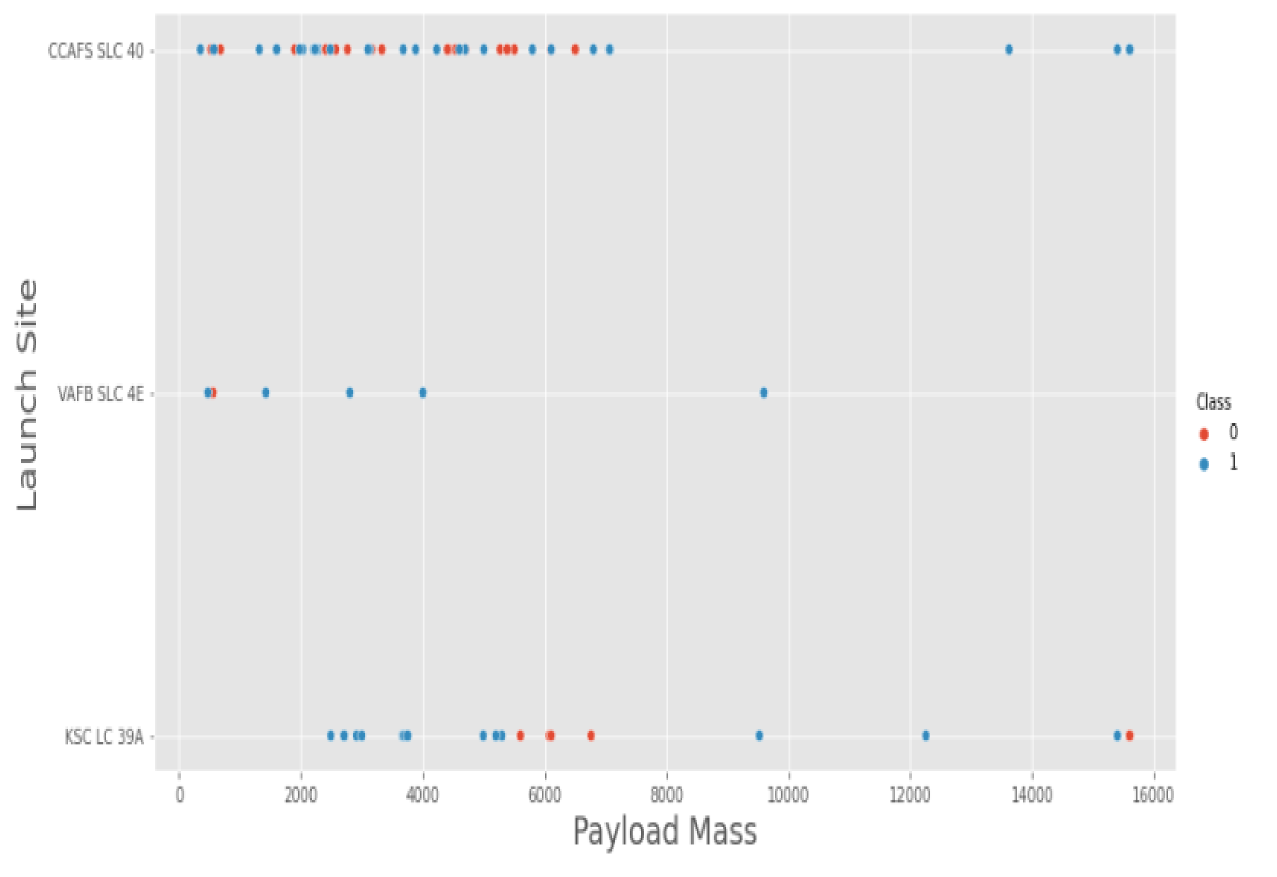
Insights drawn from EDA

Flight Number vs. Launch Site



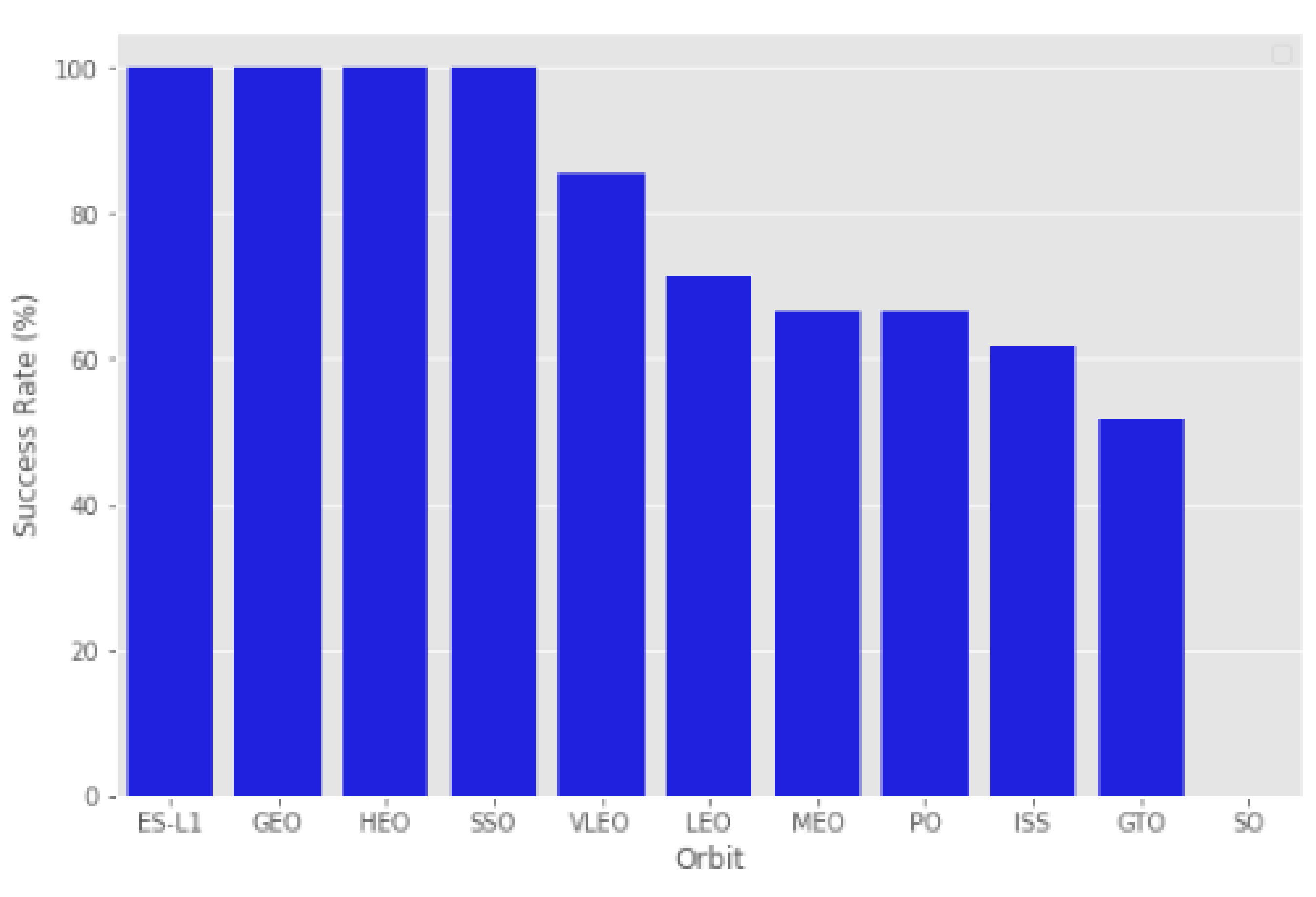
- The exploratory data analysis findings showed by the graph demonstrates that as the number of flights increased, so did the success rate.
- The successful launches are shown by blue dots, whereas the failed launches are represented by red dots.
- After the 40th launch, there seems to be an uptick in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

Payload vs. Launch Site



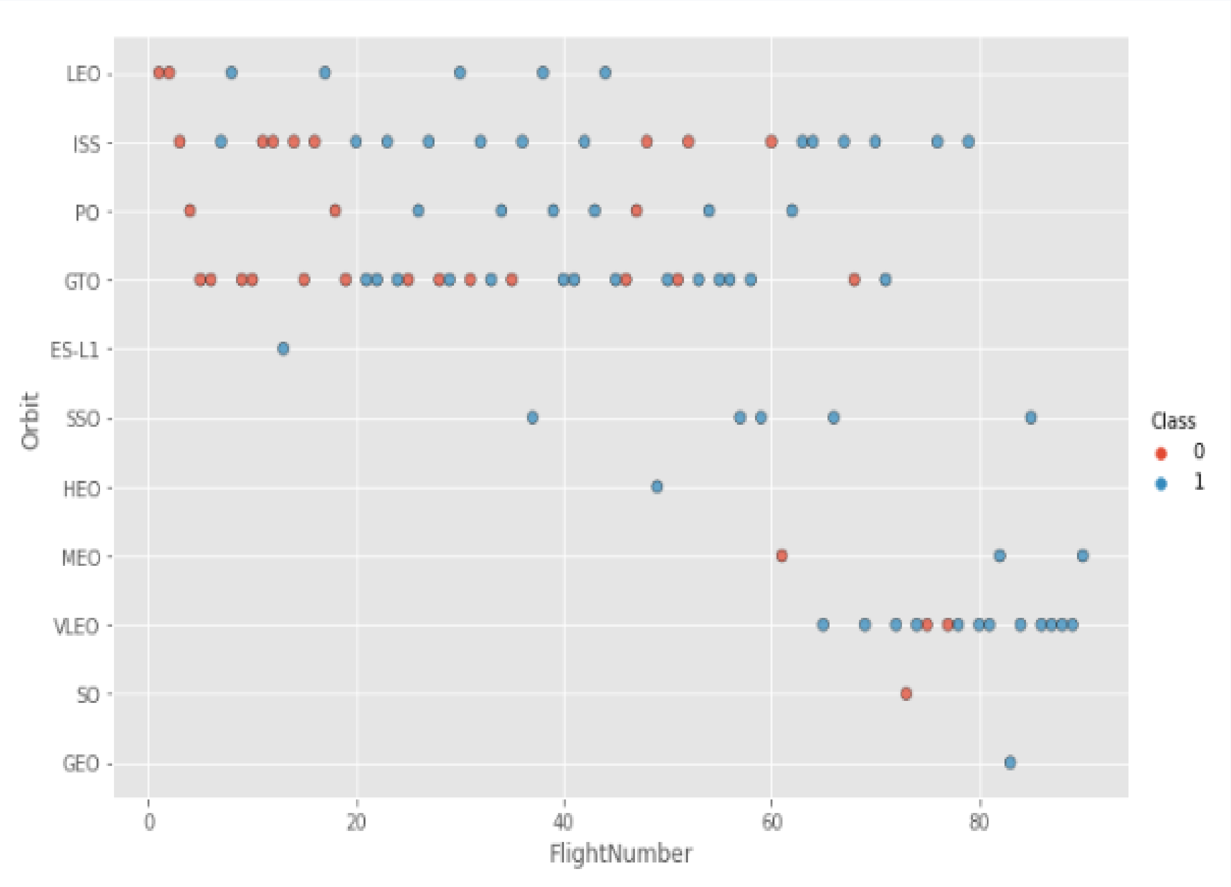
- According to the findings of the exploratory data analysis, successful launches are represented by blue dots, whereas failure launches are represented by red dots.
- There are no rockets launched for heavy payload mass from the VAFB-SLC launch site.
- Decisions cannot be made using this metric because there appears to be a poor association between Payload and Launch Location.

Success Rate vs. Orbit Type



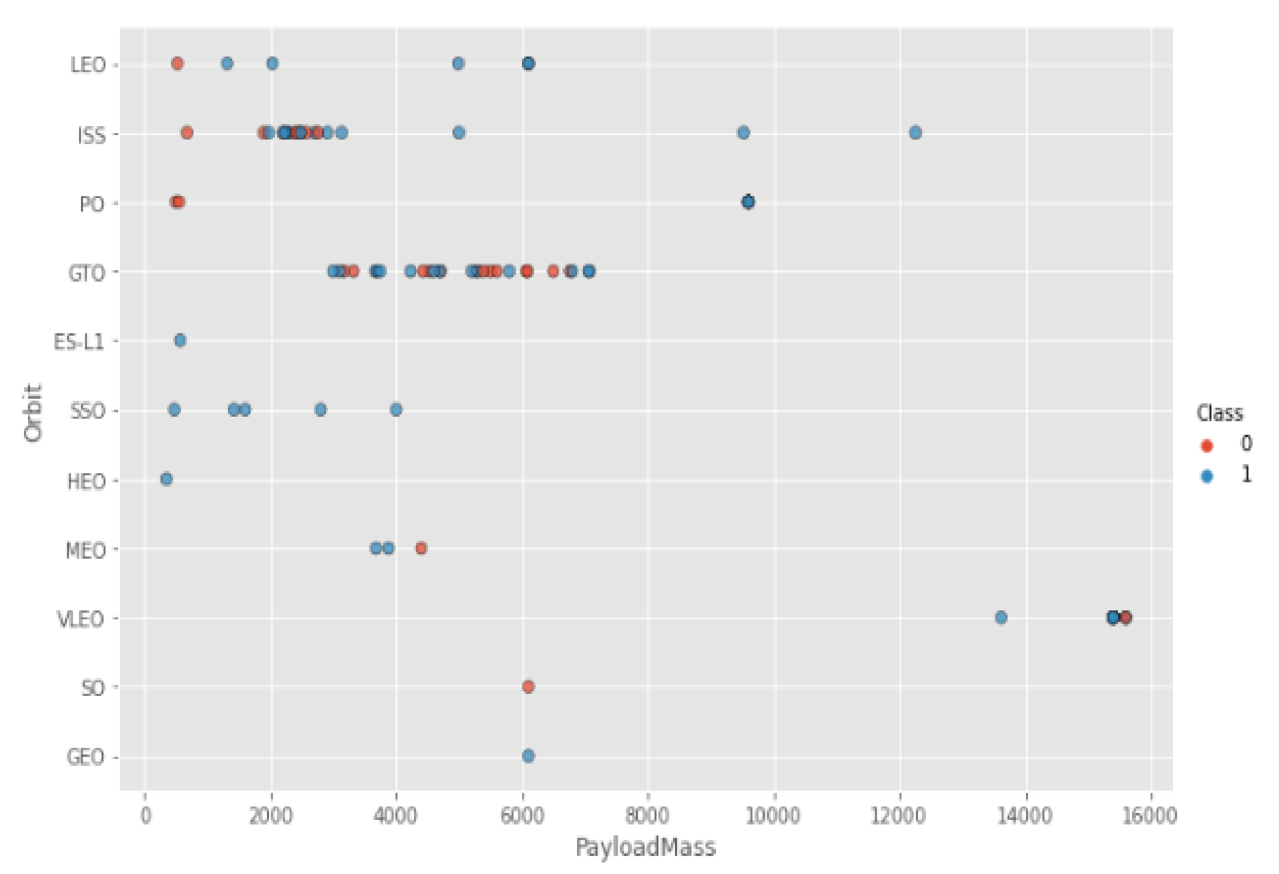
- The findings of the exploratory data analysis showed that the success rates for the Orbits SSO, HEO, GEO, and ES-L1 are all 100%. The findings of the exploratory data analysis showed that the success rates for the Orbits SSO, HEO, GEO, and ES-L1 are all 100%.
- With a 0% success rate, SO orbit had no launches that were successful.

Flight Number vs. Orbit Type



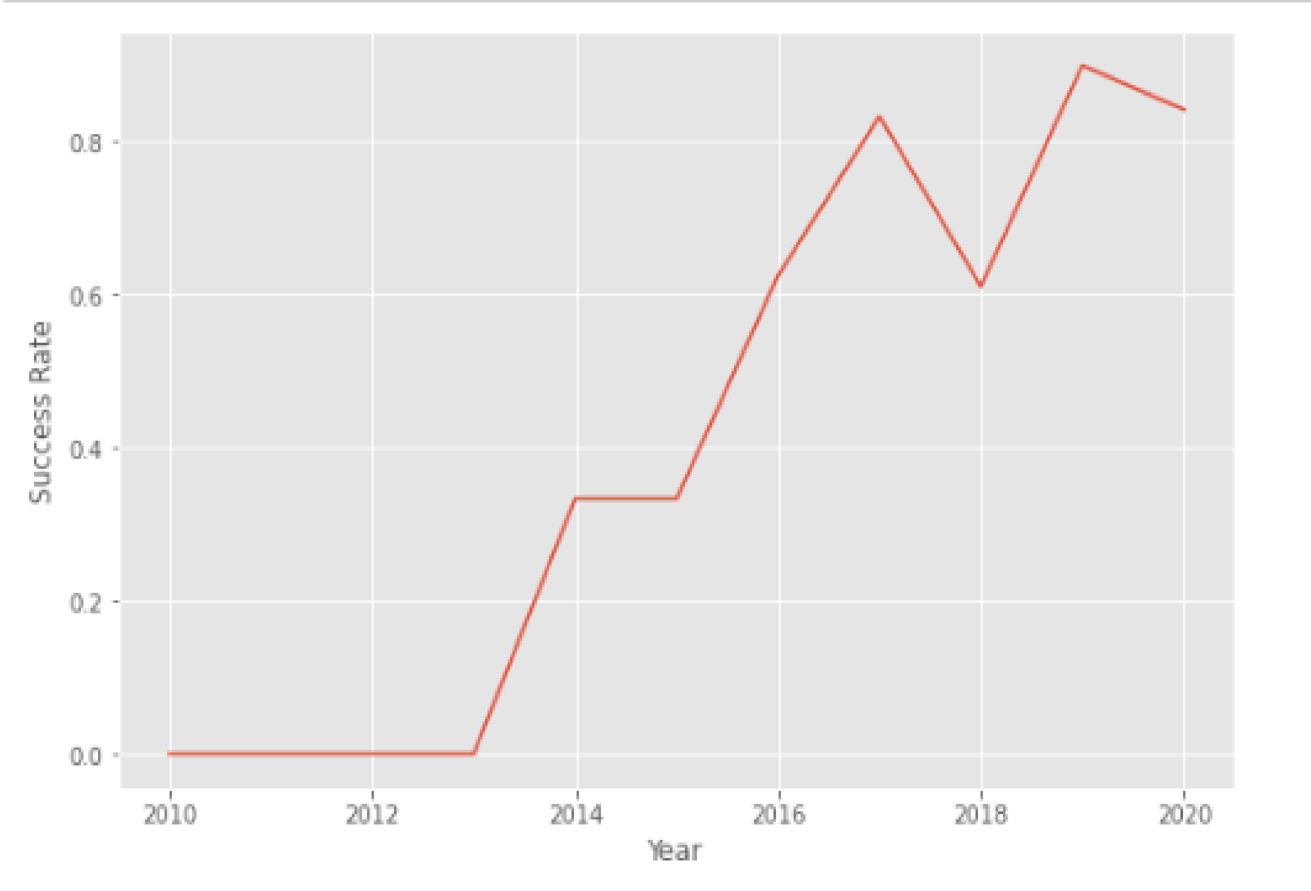
- The number of trips is strongly connected with success in the LEO orbit.
- In the GTO orbit, there does not appear to be a correlation between flight numbers.
- The SSO orbit has fewer flights than the other orbits but a success record of 100%
- .The success rate of flights with numbers over 40 is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

Payload vs. Orbit Type



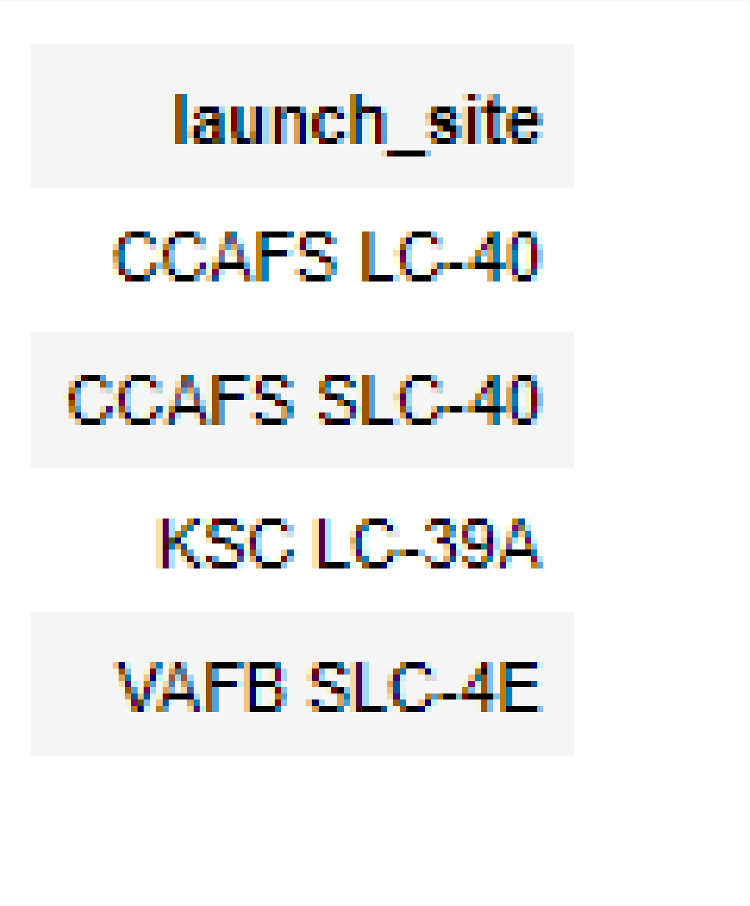
- The success rate rises in the PO, SSO, LEO, and ISS orbits as payload weight increases.
- While both successful and unsuccessful launches are frequently observed, there does not appear to be a direct association between orbit type and payload mass for GTO orbit.

Launch Success Yearly Trend



- There seems to be a sharp increase in success rate from 0% from 2013.
- The general trend of the chart shows an increase in landing success rate as the years pass. There is however a dip in 2018 as well as in 2020.

All Launch Site Names

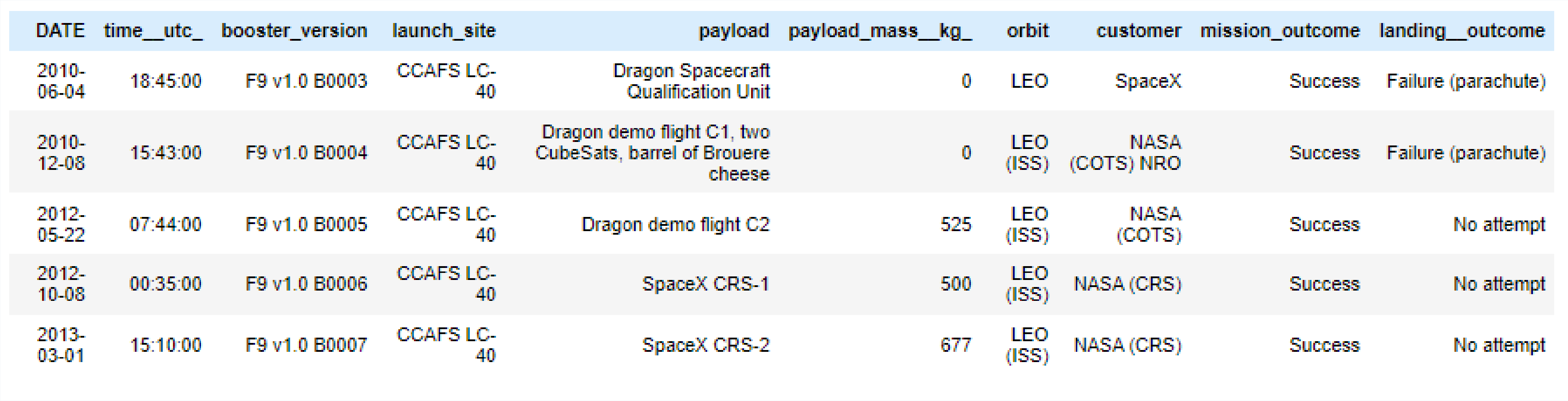


launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- The DISTINCT clause was used to return only the unique rows from the *launch_site* column.
- The names of the launch sites are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E .

Launch Site Names Begin with 'CCA'

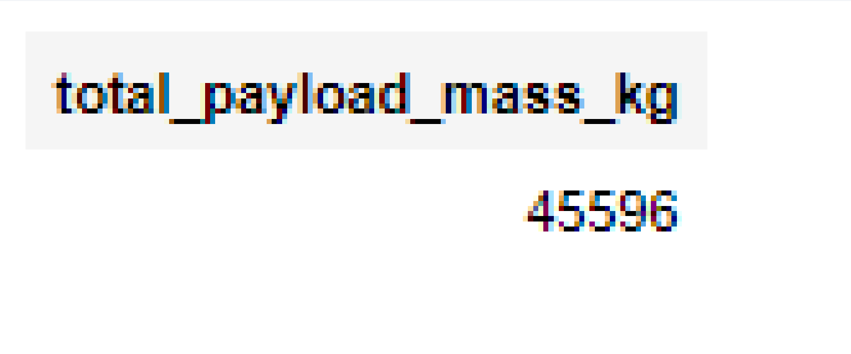
- The LIMIT and LIKE clauses were used to display only the top five results where the *launch_site* name starts with 'CCA'



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	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

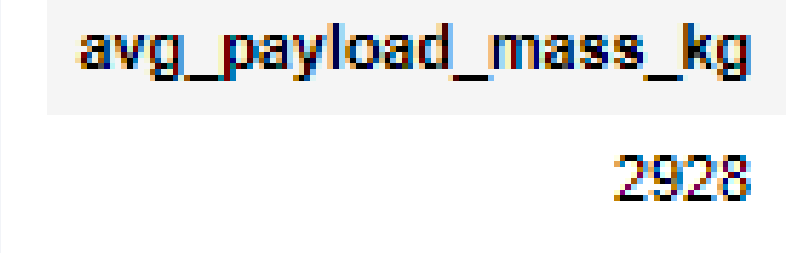
- The SUM() function was used to the calculate the total payload carried by boosters from NASA from the *payload_mass__kg* column.



total_payload_mass_kg
45596

Average Payload Mass by F9 v1.1

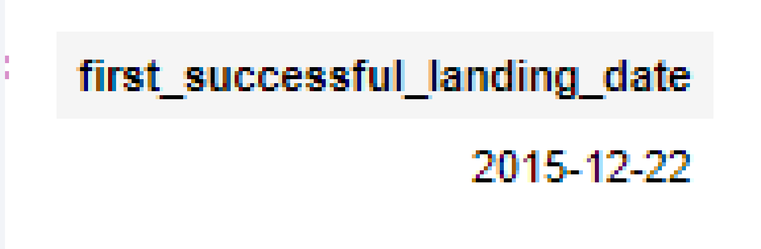
- The AVG() function was used to the calculate the average payload the average payload mass carried by booster version F9 v1.1
- The WHERE clause was used to filter results so that the calculations were only performed on *booster_versions* only if they were named “F9 v1.1”



avg_payload_mass_kg
2928

First Successful Ground Landing Date

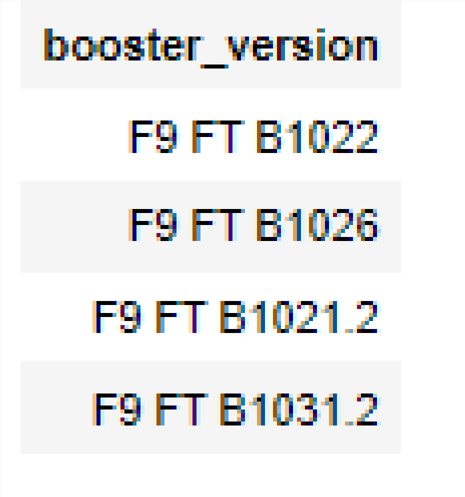
- The MIN(DATE) function was used to find the date of the first successful landing outcome on ground pad
- The WHERE clause ensured that the results were filtered to match only when the *'landing*_outcome' column is 'Success (ground pad)'



first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The BETWEEN clause was used to retrieve only those results of payload mass greater than 4000 but less than 6000. The WHERE clause filtered the results to include only boosters which successfully landed on drone ship.

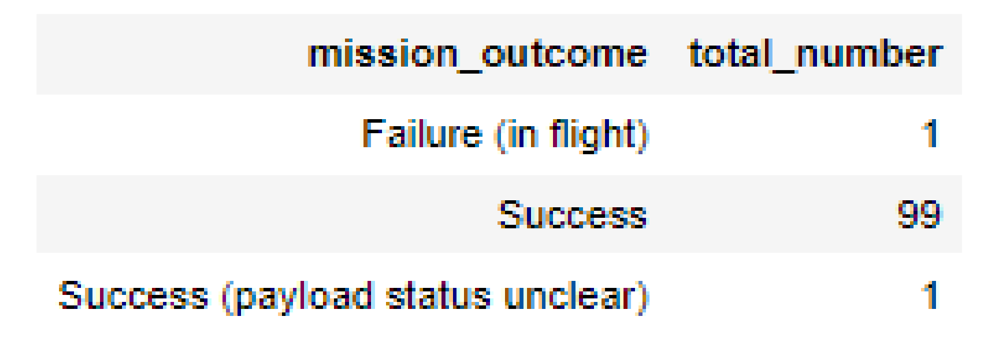


booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The COUNT() function is used to count the number of occurences of different mission outcomes with the help of the GROUPBY clause applied to the '*mission_outcome*' column. A list of the total number of successful and failure mission outcomes was returned.

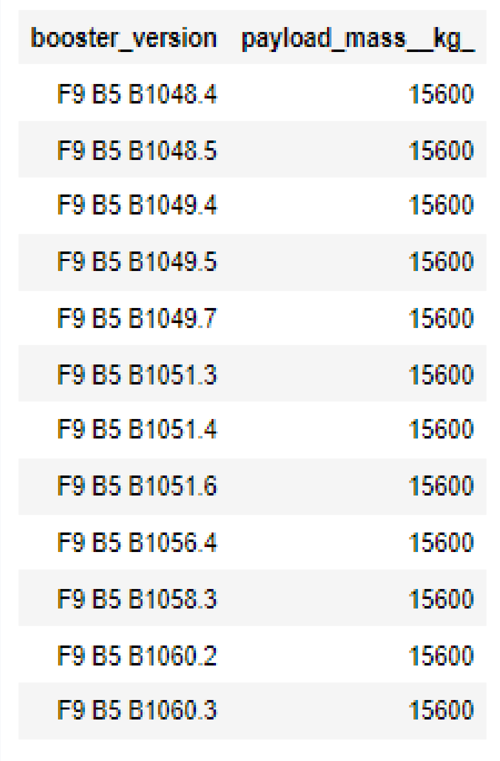
- There have been 99 successful mission outcomes out of 101 missions.



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

Boosters Carried Maximum Payload

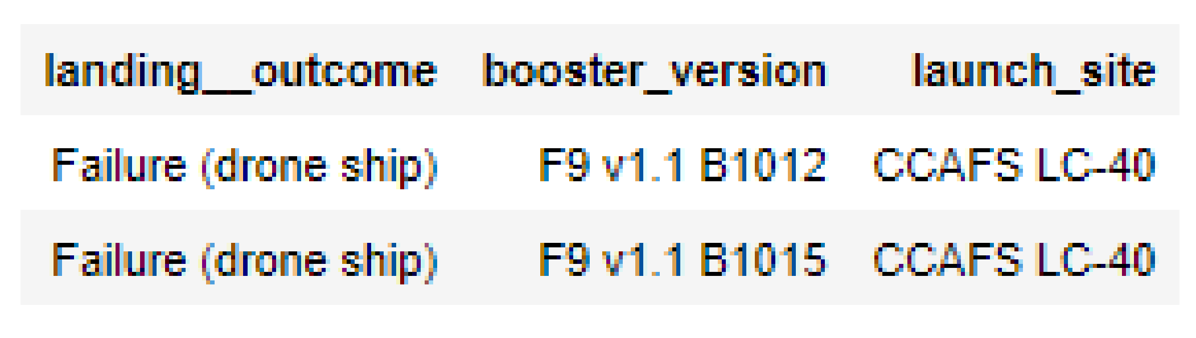
- The MAX() function was used in a subquery to retrieve a list of boosters which have carried the maximum payload mass.



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

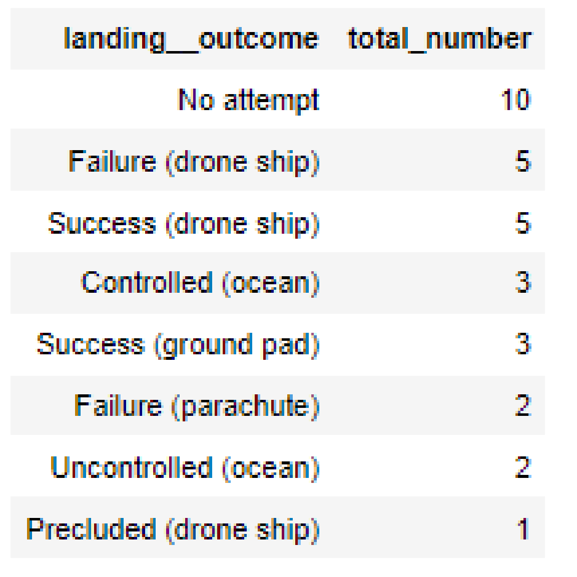
- The SELECT statement was used to retrieve multiple columns from the table. The YEAR(DATE) function was used to retrieve only those rows with a 2015 launch date.



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

- COUNT() function was used to count the different *landing outcomes.*The WHERE and BETWEEN clauses filtered the results to only include results between 2010-06-04 and 2017-03-20. The GROUPBY clause ensure that the counts were grouped by their outcome. The ORDERBY and DESC clauses were used to sort the results by descending order.

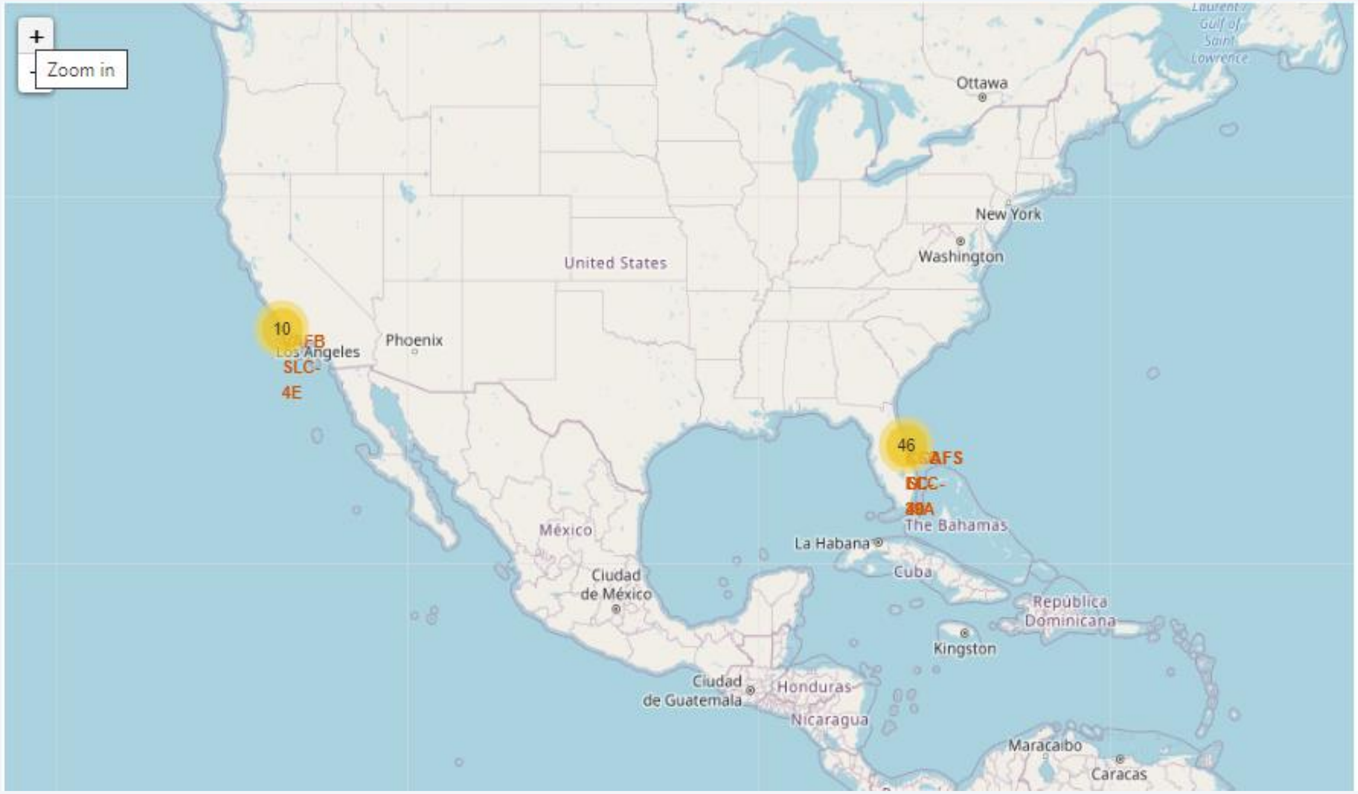


landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

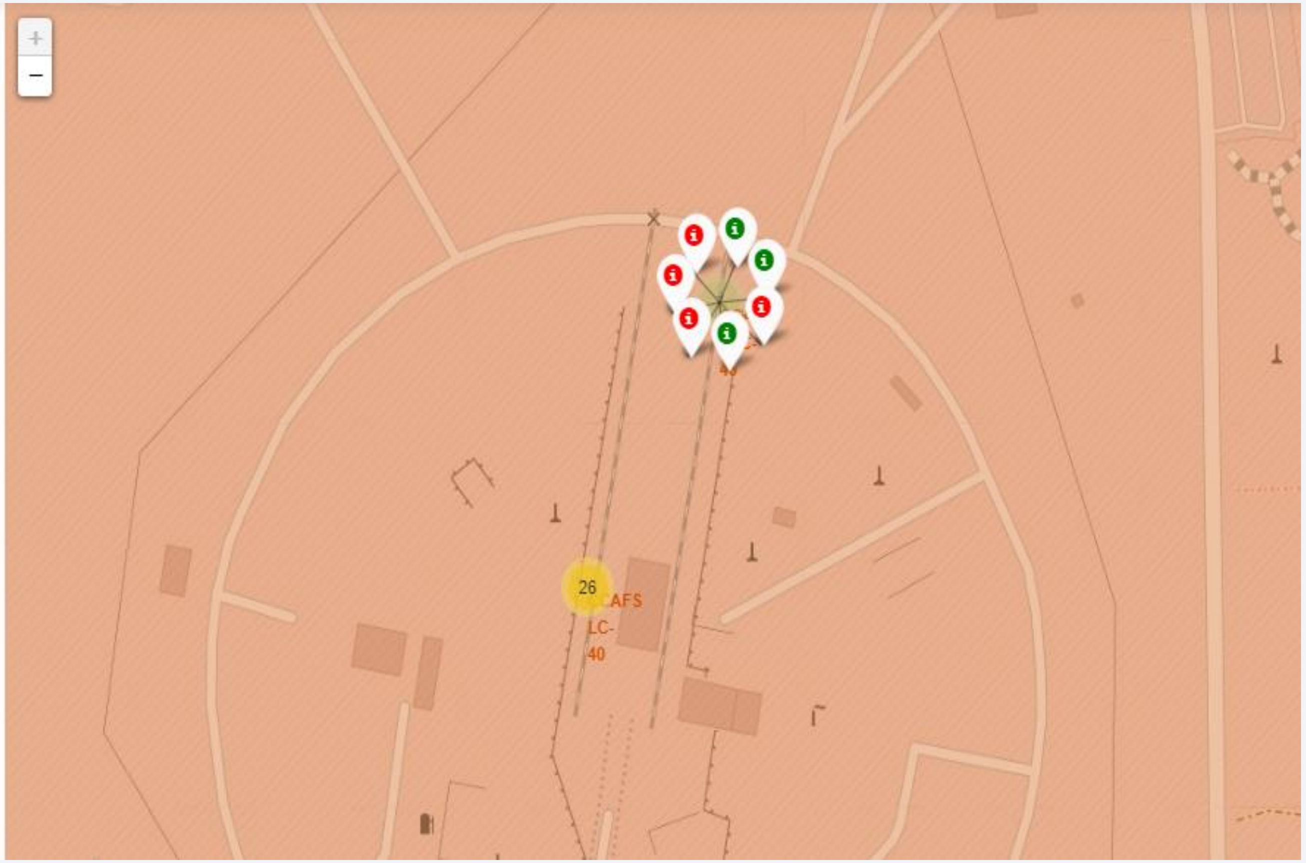
Launch Sites Proximities Analysis

SpaceX Launch Sites Locations



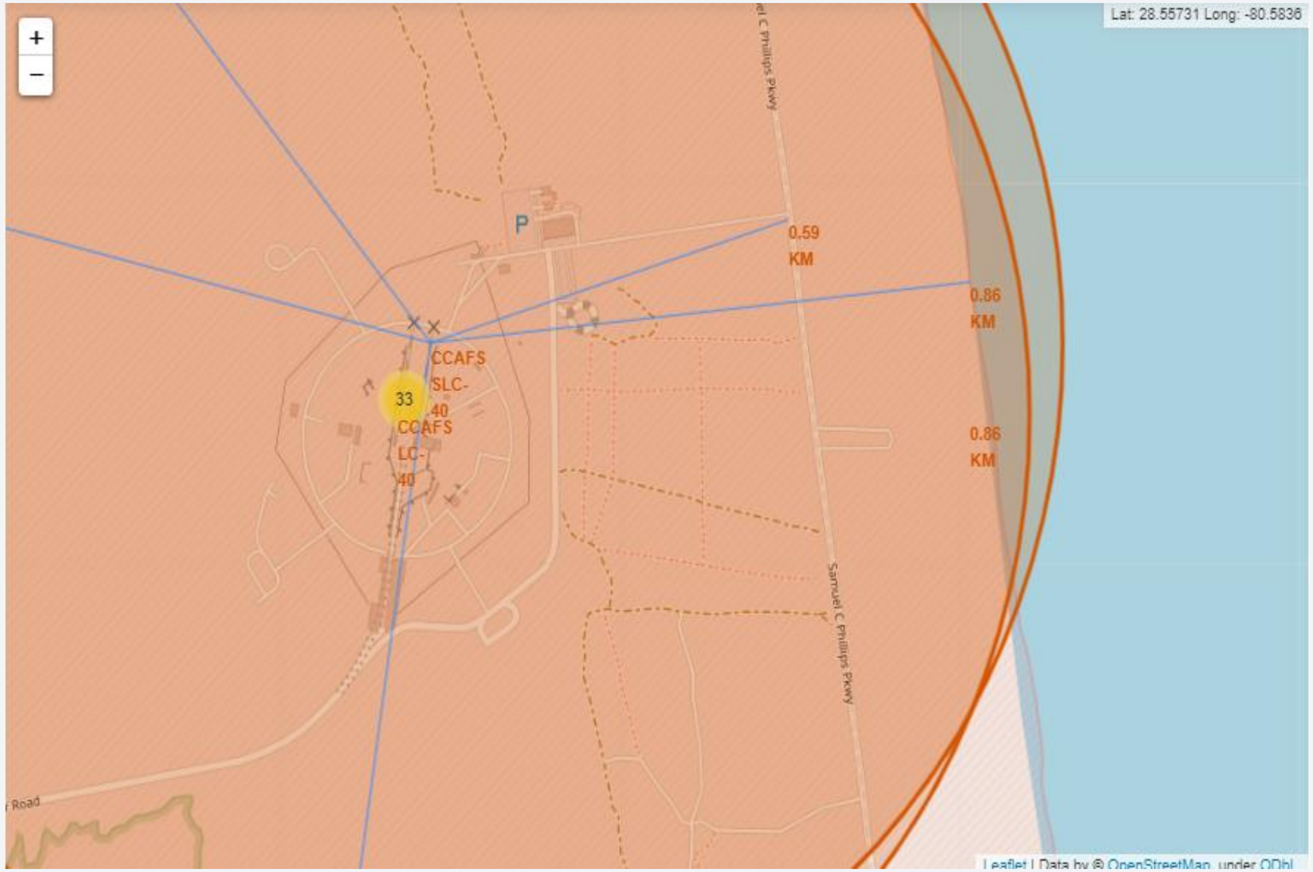
- The locations of all the SpaceX launch sites in the US are indicated by the yellow markers.
- The launch pads have been situated in a suitable location close to the coast.

Success and Failure on Launch Site

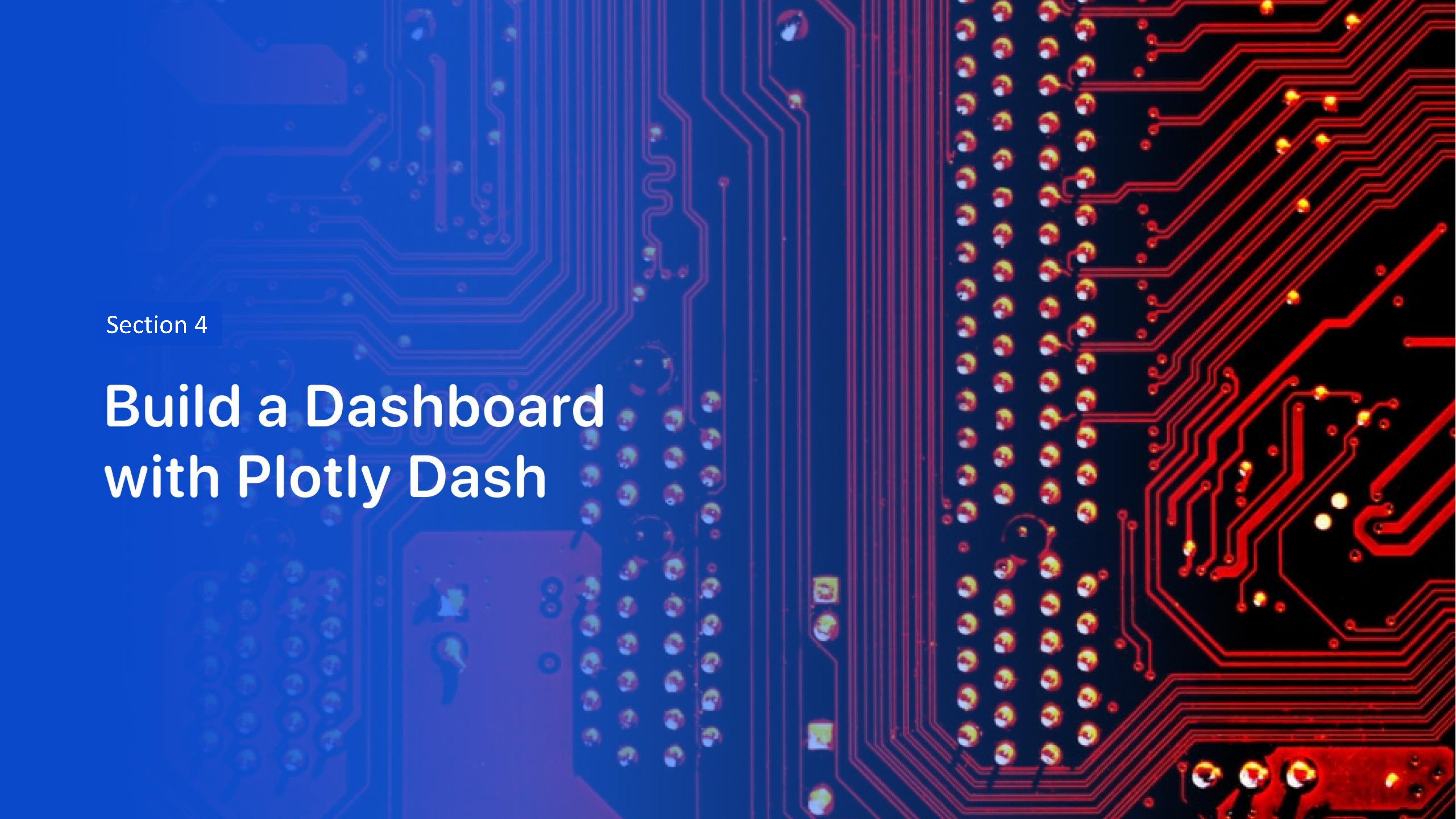


- The launch site will show marker clusters of successful landings (green) or failed landings as we zoom in on a launch location (red).

Launch Site Proximities To Resources



- The generated map reveals that the chosen launch point is close to a highway for people and equipment transportation.
- For launch failure testing, the launch site is also close to the coast.
- Also, the launch locations keep a specific distance from the cities.

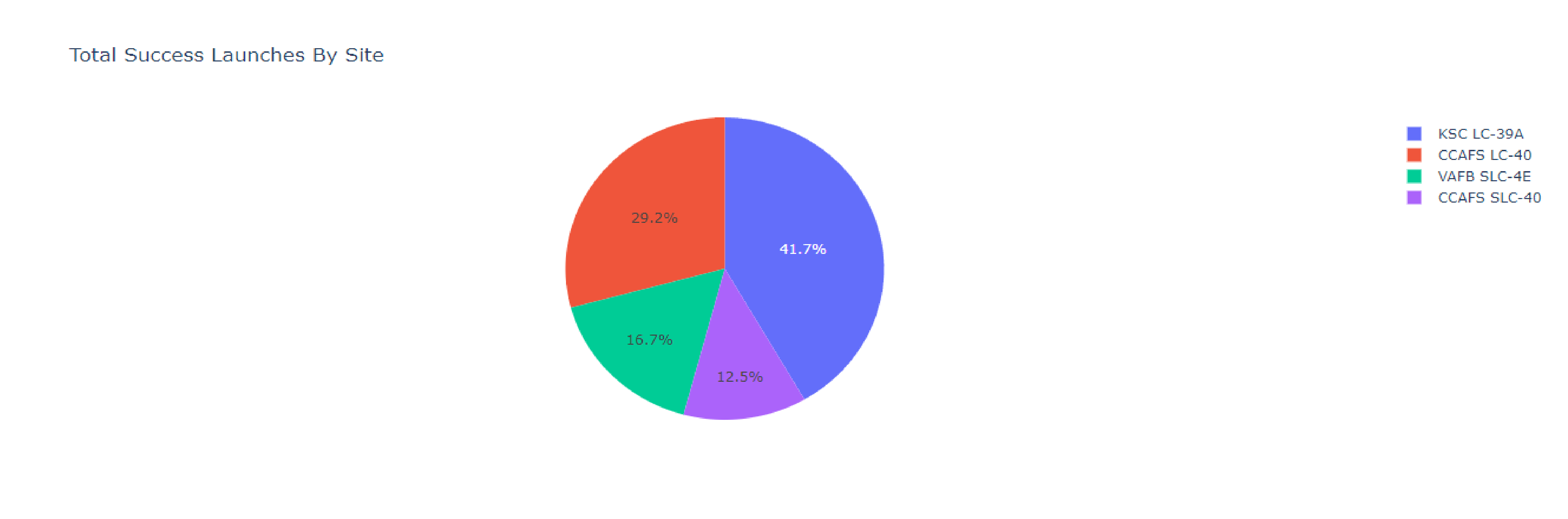


Section 4

Build a Dashboard with Plotly Dash

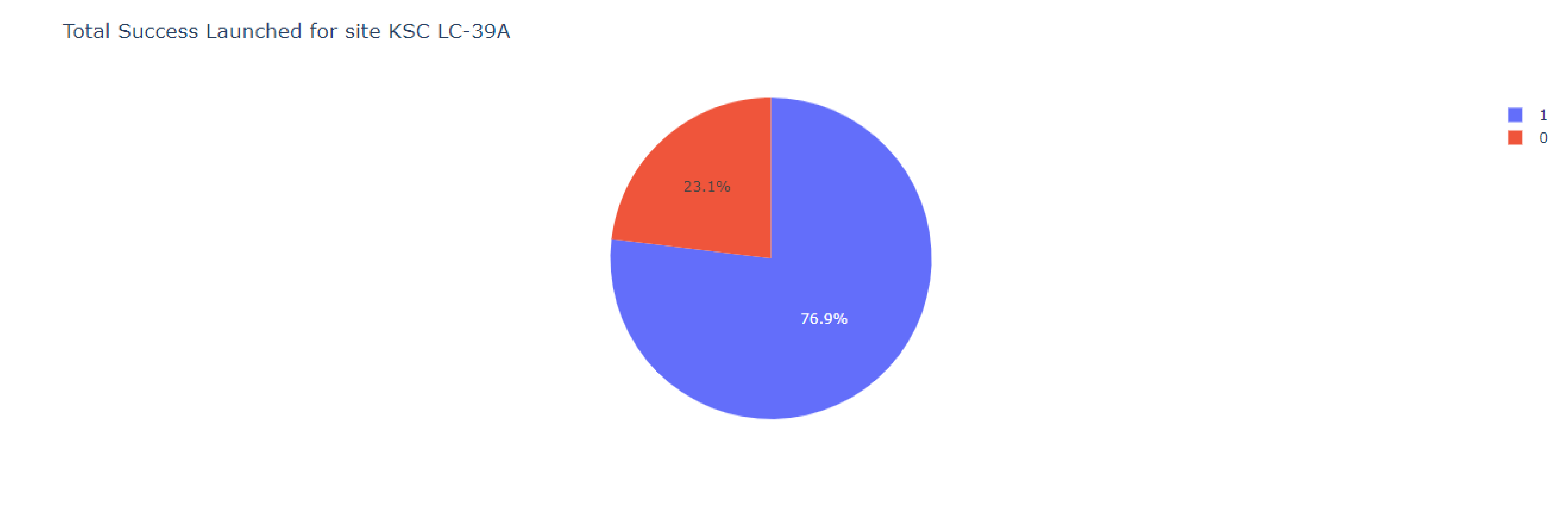
Total Successful Launches By Site

- With a total of 10, the KSC LC-39A Launch site has had the most successful launches.

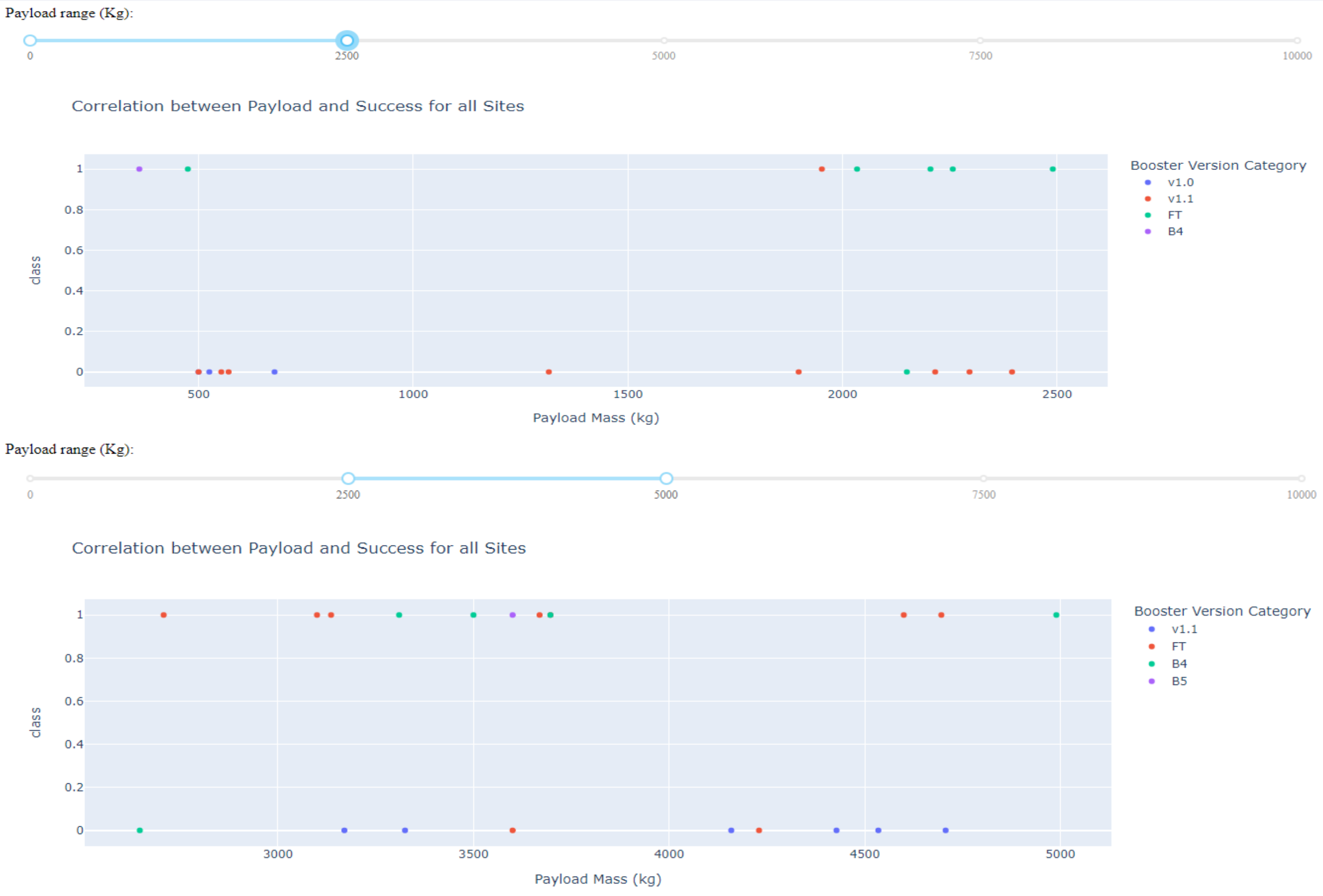


Launch Site With Highest Success Ratio

- With a success rate of 76.9%, the KSLC-39A leads the pack.



Payloads vs Launch Outcome



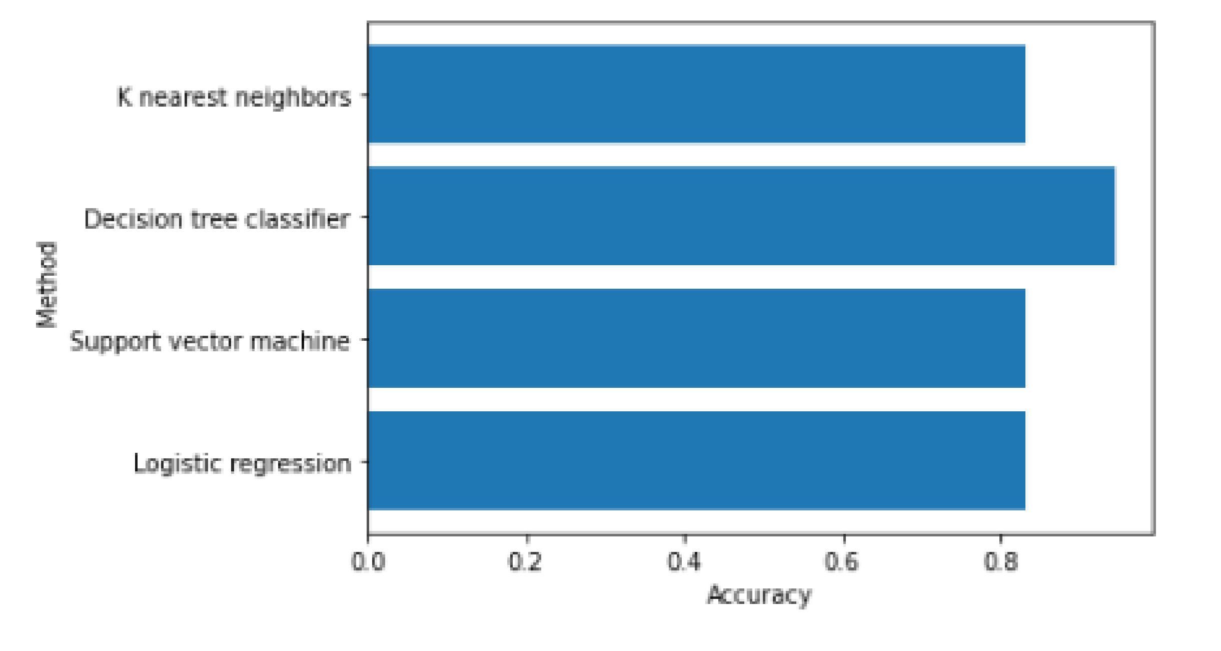
- Payloads between 0 and 2500 kg have a somewhat lower launch success percentage than payloads between 250 and 5000 kg. Really, there isn't much of a distinction between the two.
- In both weight ranges, the v1.1 booster version had the highest effectiveness rate.

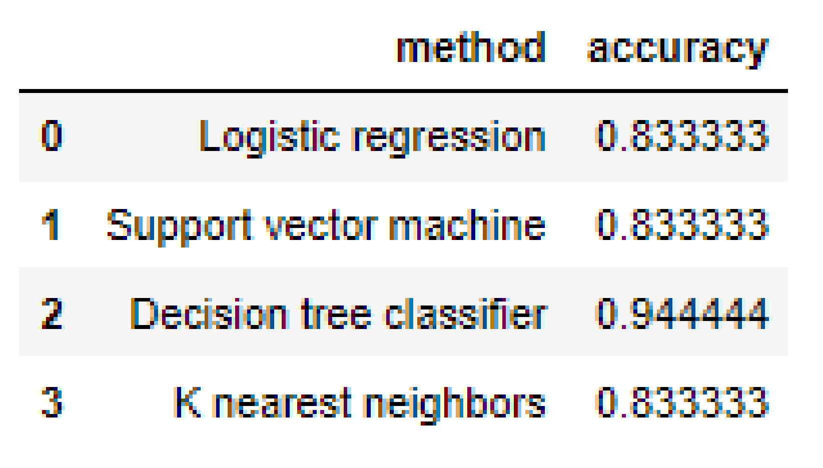
Section 5

Predictive Analysis (Classification)

Classification Accuracy

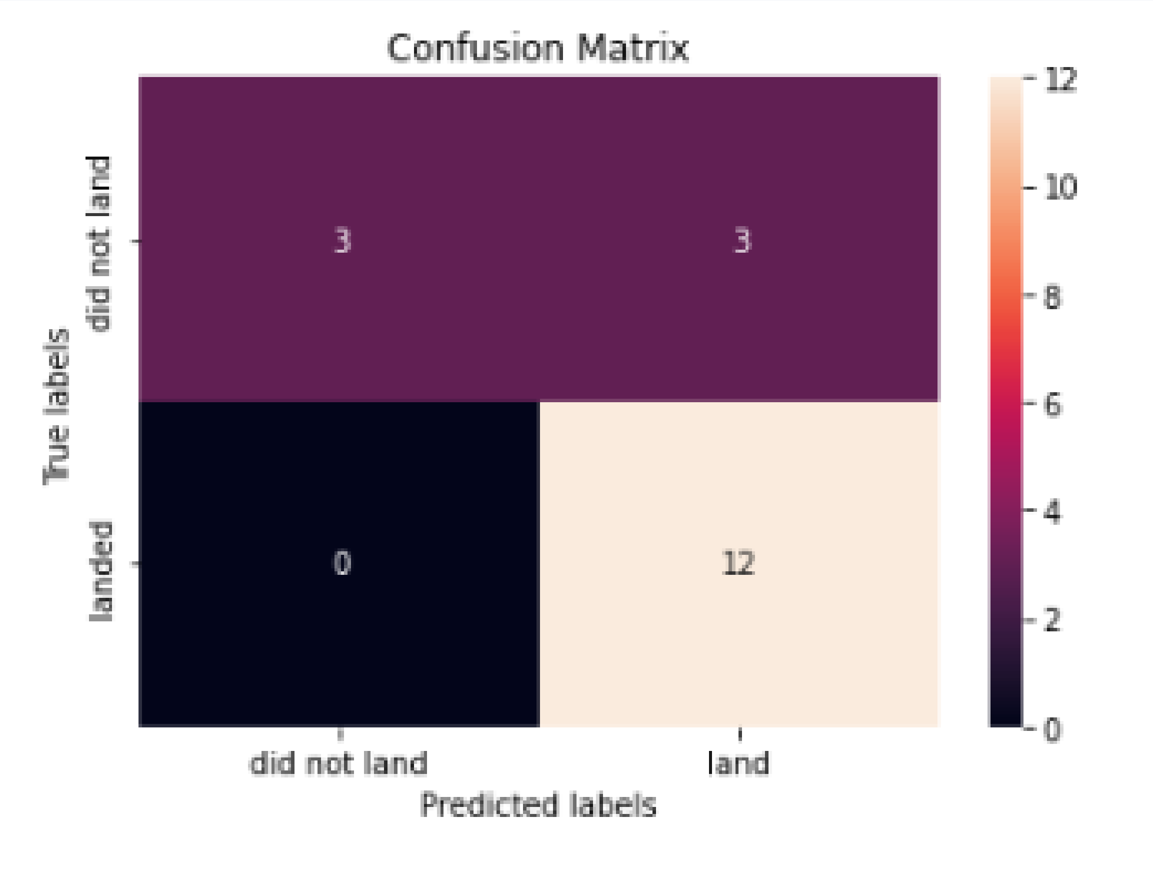
- The Decision tree classifier had the best accuracy at 94%.





	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 it was failure, it predicted 3 unsuccessful landings (True Negative).
- When the True label was an unsuccessful landing, the algorithm also predicted three successful landings (False Positive).
- 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 most accurate predictive model for this dataset, having a 94% accuracy rate.

Appendix

- Coursera Project Link: https://www.coursera.org/learn/applied-data-science-capstone/home/welcome
- GitHub Repository: https://github.com/Ajay0-data/IBM-Data-Science-Capstone-Project

Thank you!