IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Name: Haseeb Khan

Date: 17-March-2023

Outline

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

Executive Summary

In this capstone project, we will predict whether the SpaceX Falcon 9 first stage will land successfully. If we can determine if the first stage will land, we can determine the cost of a launch. This will be achieved with the use of different machine learning classification algorithms.

The methodolgy followed will include Data Dollection, Data Wrangling and Preprocessing, Exploratory Data Analysis, Data Visualization and finally, Machine Learning Prediction.

During our investigation, the results of our analysis indicate that there are some features of rocket launches that have a correlation with the success or failure launches.

In the end we conclude that the Decision Tree may be the best machine learning algorithm to for this problem.

Introduction

The main goal of this capstone project is to predict whether the Falcon 9 first stage will land successfully. SpaceX prides itself in being able to reuse the first stage of a rocket launch so much so that they avertise on their website that their rocket launches cost 62 million while other provides cost upward 165 million. Much of these savings are down to the first stage's reusability. 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.

This brings us to our main question that we are trying to answer : For a given set of features
about a Falcon 9 rocket launch, will the first stage of the rocket land successfully?

Section 1

Methodology

Methodology

Executive Summary

Data was collected through two methods: requesting data from the SpaceX API and web scraping launch data from a Wikipedia page. Data wrangling was then performed to transform and clean the data using Python's pandas library.

With the clean data, exploratory data analysis (EDA) was performed using visualization tools such as Python's matplotlib and seaborn libraries, as well as answering questions using SQL queries. Python's interactive visualization packages were used to answer some analytical questions. Folium was used for creating maps while Plotly Dash was used to create interactive data visualizations.

Four different machine learning classification models were used for the predictive analysis. The models that were used are logistic regression, support vector machines, k-nearest neighbour and decision tree classifier. Each model was trained, tuned and evaluated to find the best one.

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 data-frame from dictionary
5. Filter data-frame to only include Falcon 9 launches
6. Handle missing values
7. Export to CSV file

- GitHub URL: https://github.com/Niaziz007/Data-Science-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

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: https://github.com/Niaziz007/Data-Science-Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb

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: https://github.com/Niaziz007/Data-Science-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plots: Scatter plots were used to represent the relationship between two variables. Different sets of features were compared such as *Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type and Payload vs. Orbit Type.*
- Bar chart: Bar charts were used makes it easy to compare values between multiple groups at a glance. The x-axis represents a category and the y-axis represents a discrete value. Bar charts were used to compare the *Success Rate* for different *Orbit Types*
- Line chart: Line charts are useful for showing data trends over time. A line chart was used to show *Success Rate* over a certain number of *Years.*
- GitHub URL: https://github.com/Niaziz007/Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

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

- Displaying the names of the unique launch sites in the space mission -Displaying 5 records where launch sites begin with the string 'CCA' - Displaying the total payload mass carried by boosters launched by NASA (CRS) - Displaying average payload mass carried by booster version F9 v1.1,
- Listing the date when the first successful landing outcome in ground pad was achieved - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 - Listing the total number of successful and failure mission outcomes - Listing the names of the booster versions which have carried the maximum payload mass - Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 - Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.
- GitHub URL: https://github.com/Niaziz007/Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Objects were created and added to a Folium map. Marker objects were used to show all launch sites on a map as well as the successful/failed launches for each site on the map. Line objects were used to calculate the distances between a launch site to its proximities

- 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 : https://github.com/Niaziz007/Data-Science-Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

The dashboard application contains two charts:

- A pie chart using Dash that shows the successful launch by each site. This chart is useful as you can visualize the distribution of landing outcomes across all launch sites or show the success rate of launches on individual sites.
- A scatter chart using Dash that shows the relationship between landing outcomes an the payload mass of different boosters. The dashboard takes two inputs, namely the site(s) and payload mass. This chart is useful as you can visualize how different variables affect the landing outcomes,
- GitHub URL : https://github.com/Niaziz007/Data-Science-Capstone-Project/blob/main/Plotly.ipynb

Predictive Analysis (Classification)

1. Create column for "Class"

2. Standardizing the data

3. Split ito training and test set

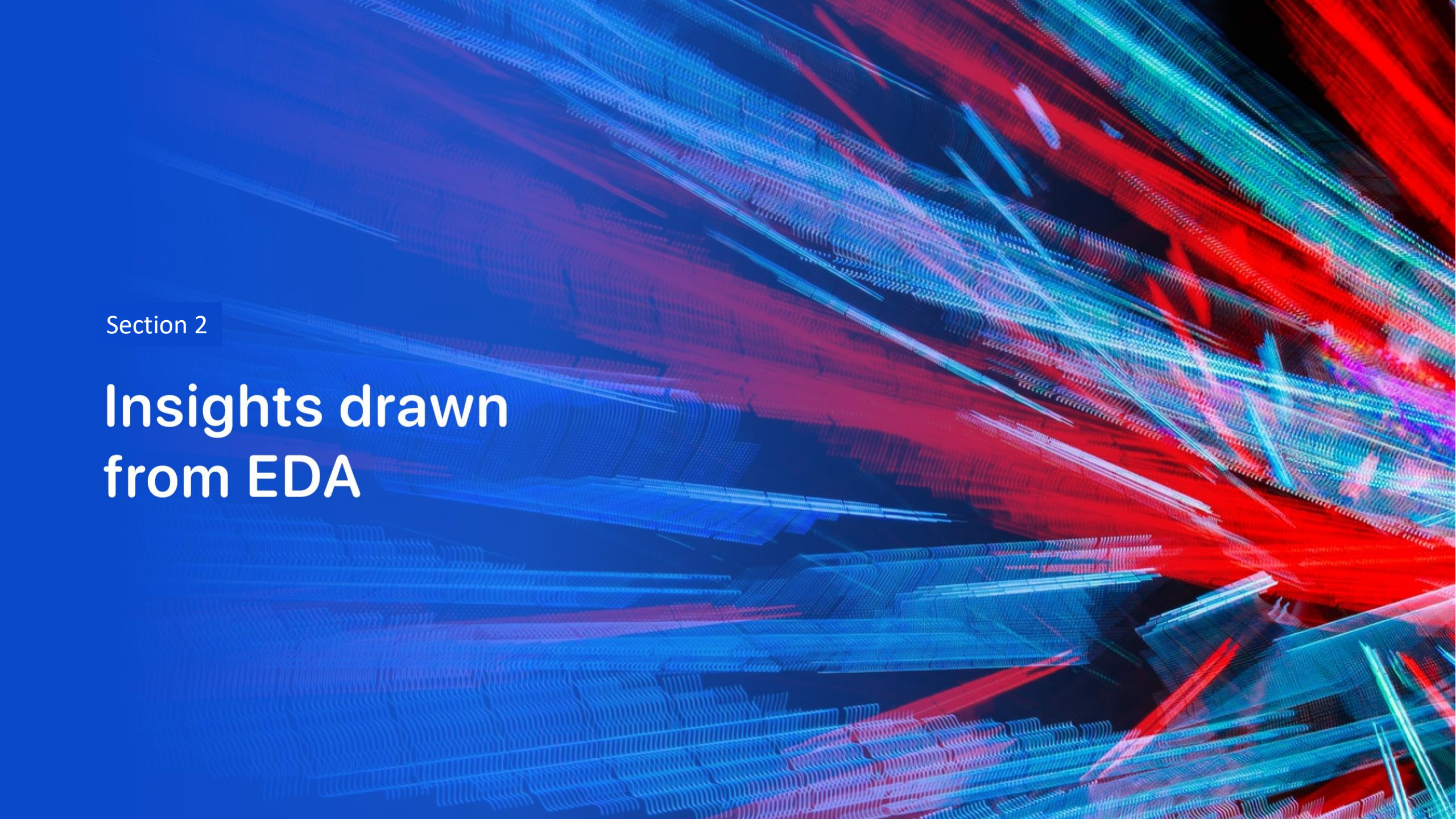
4. Find best Hyperparameter for SVM, Decision Trees, K-Nearest Neighbours and Logistic Regression.

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

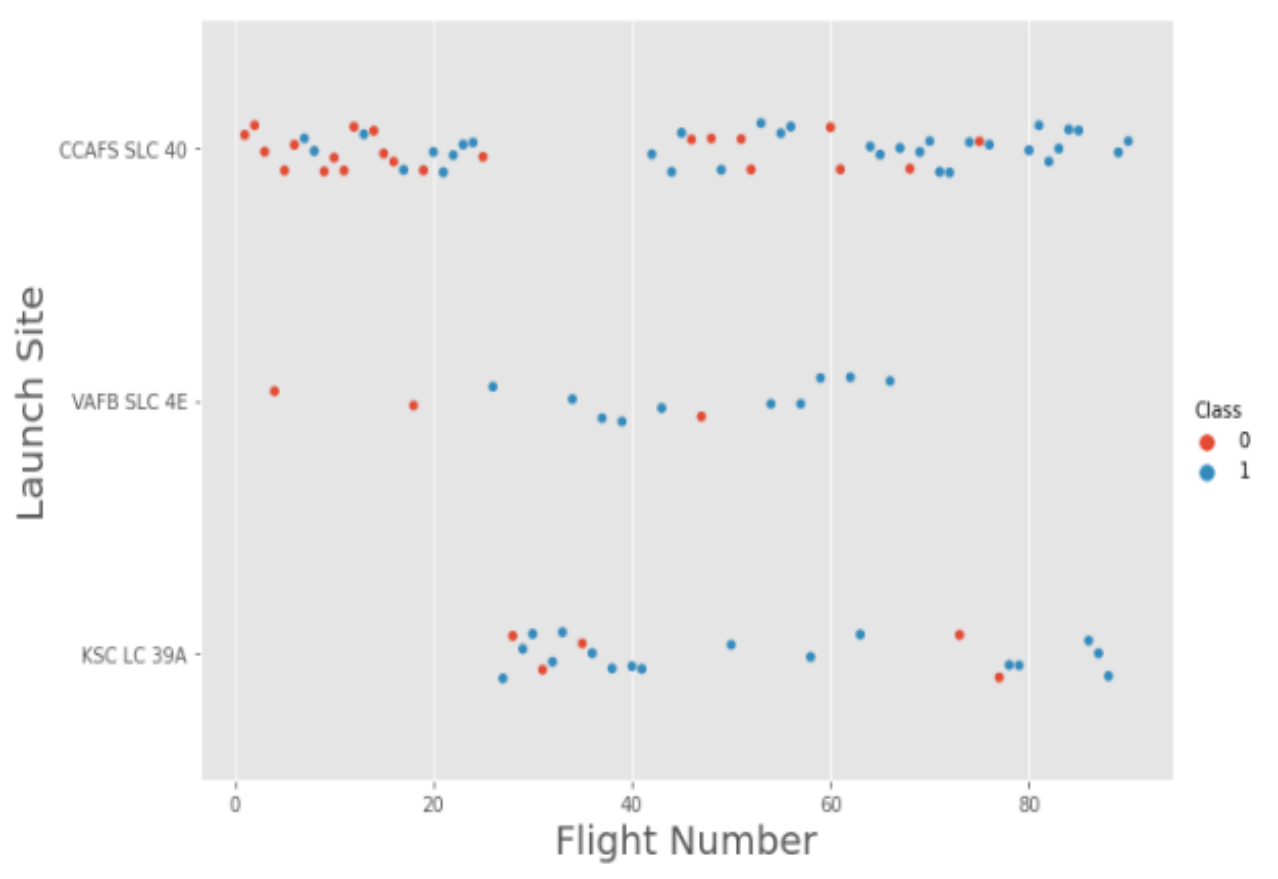
- GitHub URL: https://github.com/Niaziz007/Data-Science-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- The results of the exploratory data analysis revealed that the success rate of the Falcon 9 landings was 66.66%
- The predictive analysis results showed that the Decision Tree algorithm was the best classification method with an accuracy of 94%

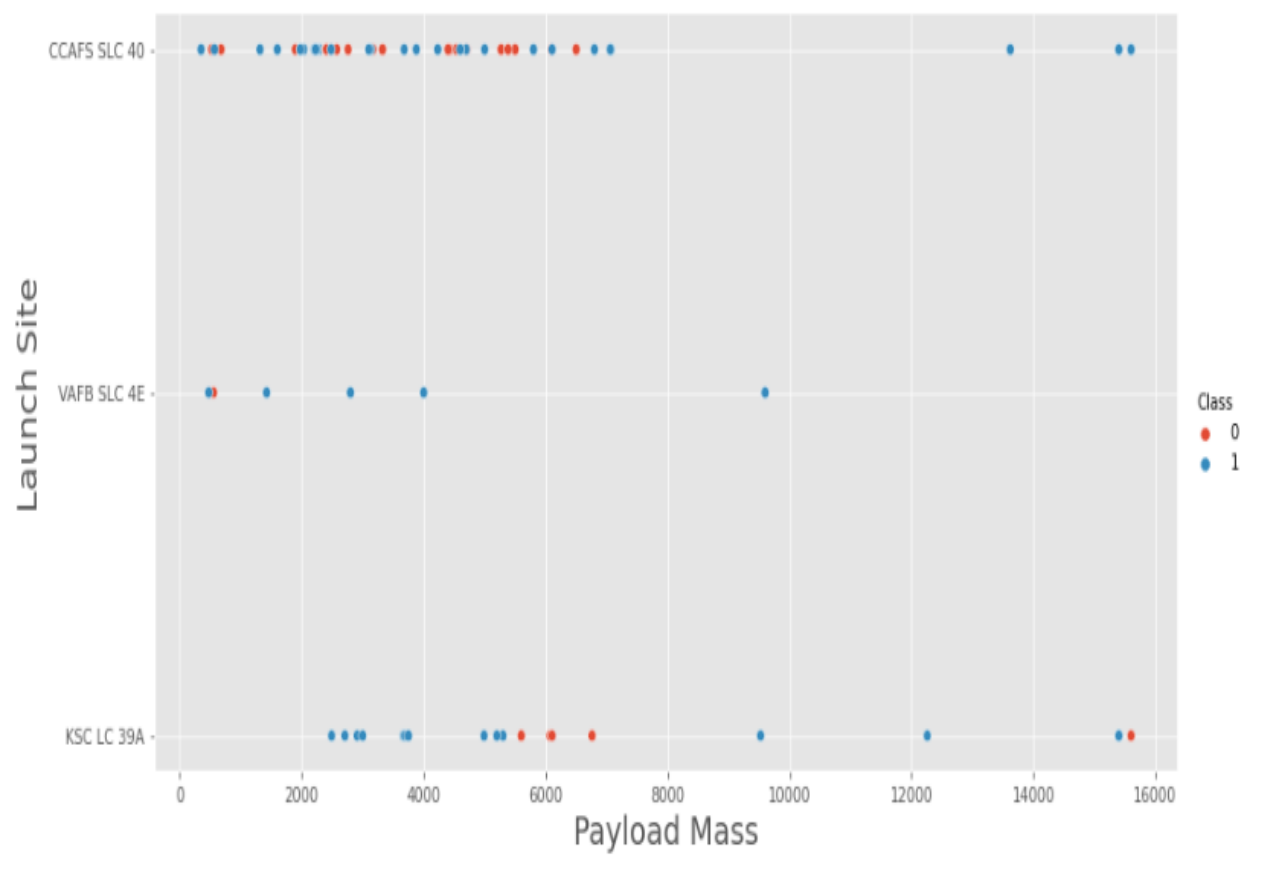

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



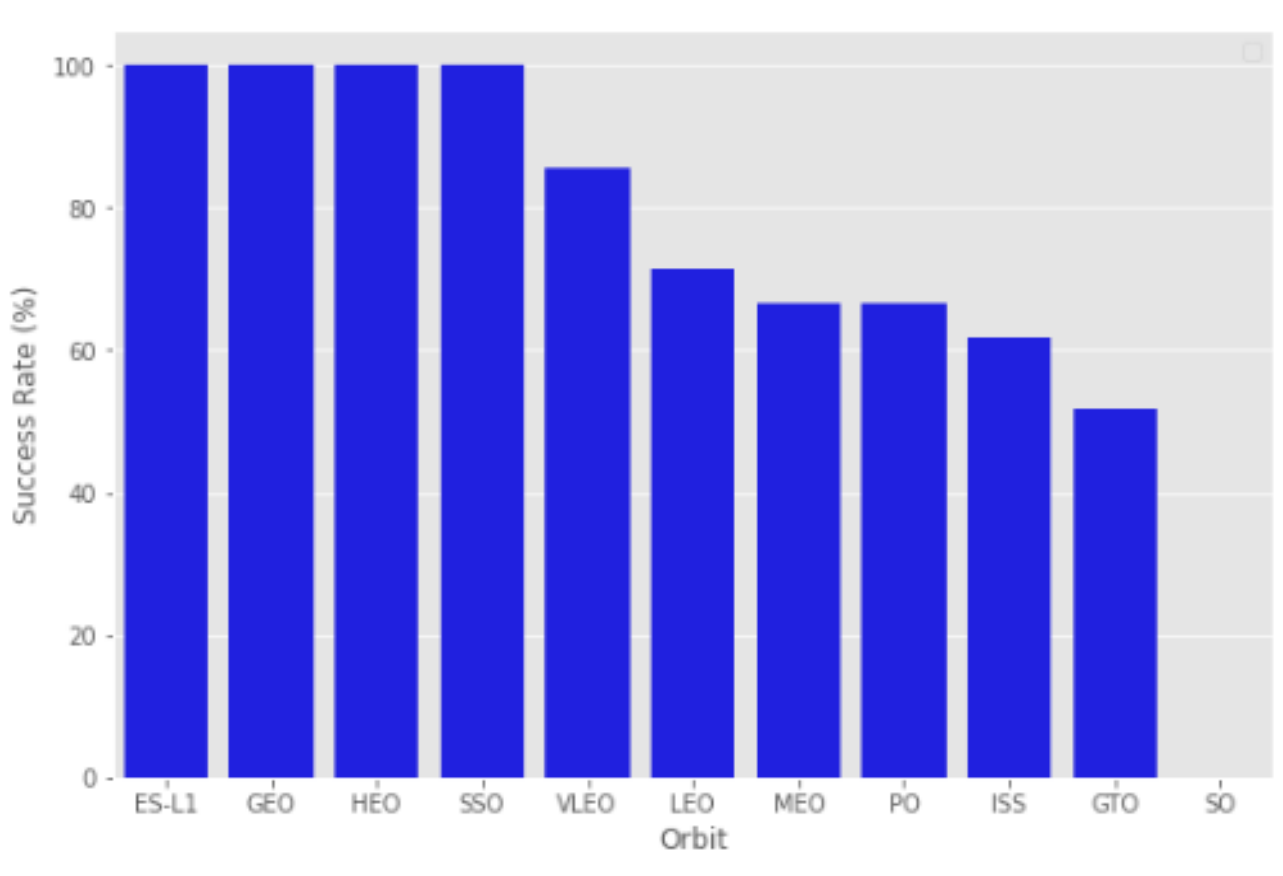
- This figure shows that the success rate increased as the number of flights increased.
- The blue dots represent the successful launches while the red dot represent unsuccessful luanches.
- There seems to be an increase in successful flights after the 40th launch.

Payload vs. Launch Site



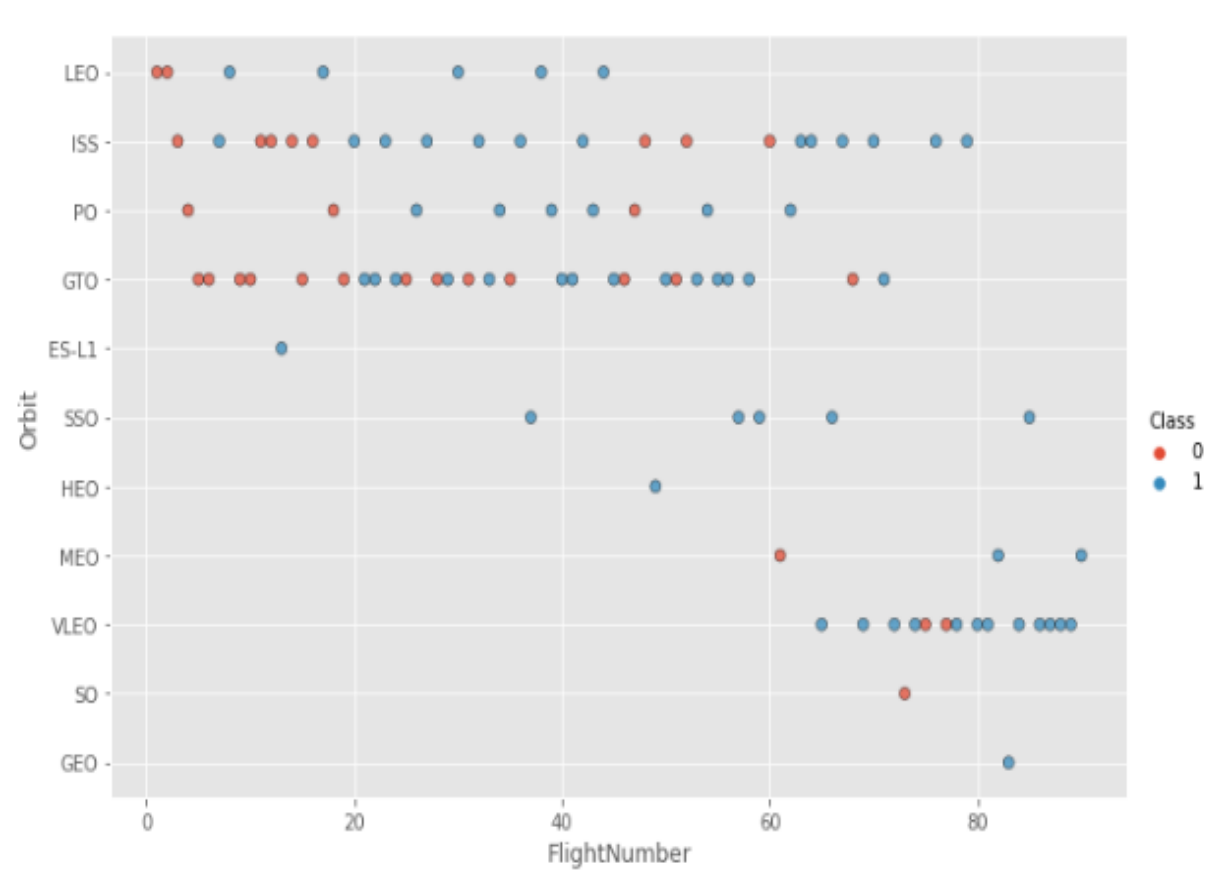
- The blue dots represent the successful launches while the red dots represent unsuccessful launches.
- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass
- There seems to be a weak correlation between Payload and Launch Site and therefore decisions cannot be made using this metric.

Success Rate vs. Orbit Type



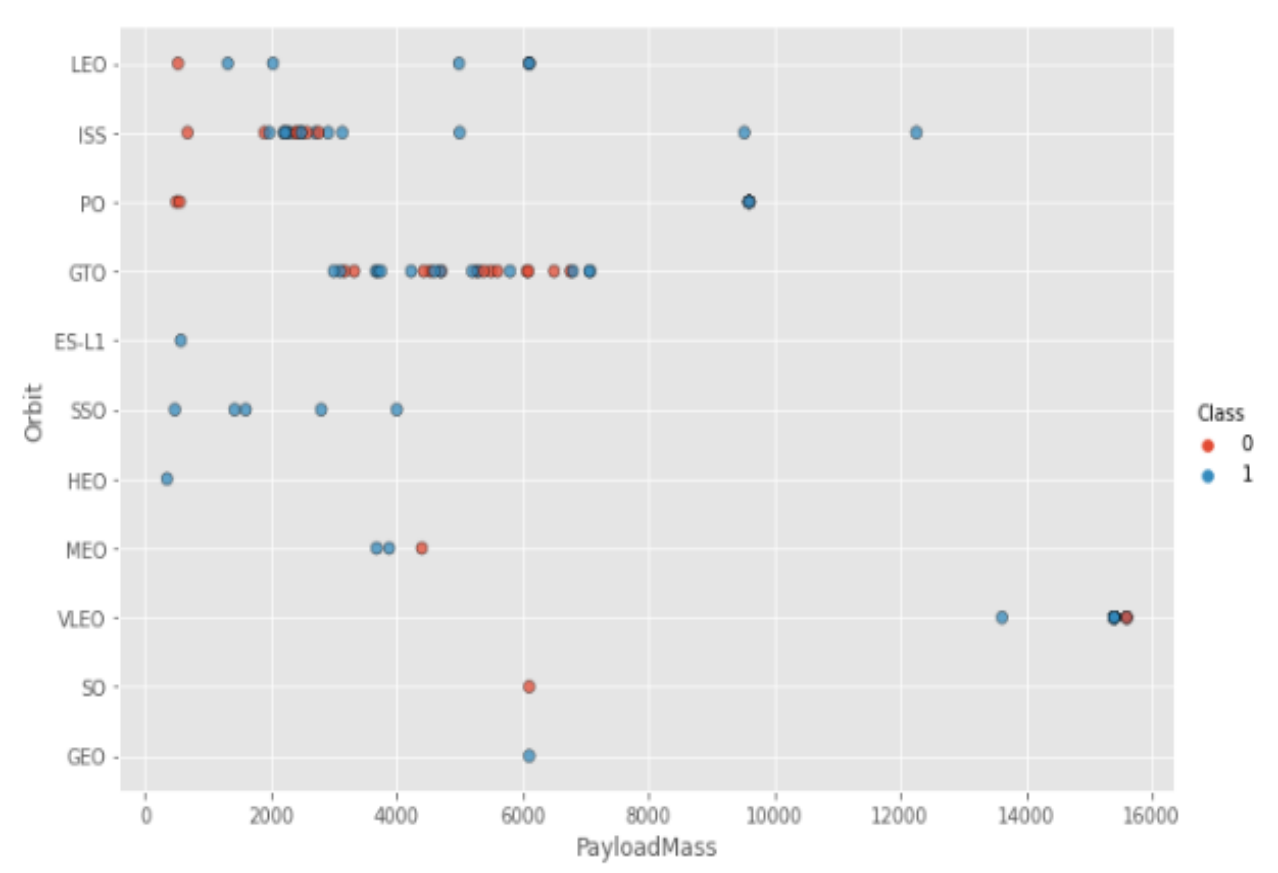
- Orbits SSO, HEO, GEO, and ES-L1 have 100% success rates.
- SO orbit did not have any successful launches with a 0% success rate.

Flight Number vs. Orbit Type



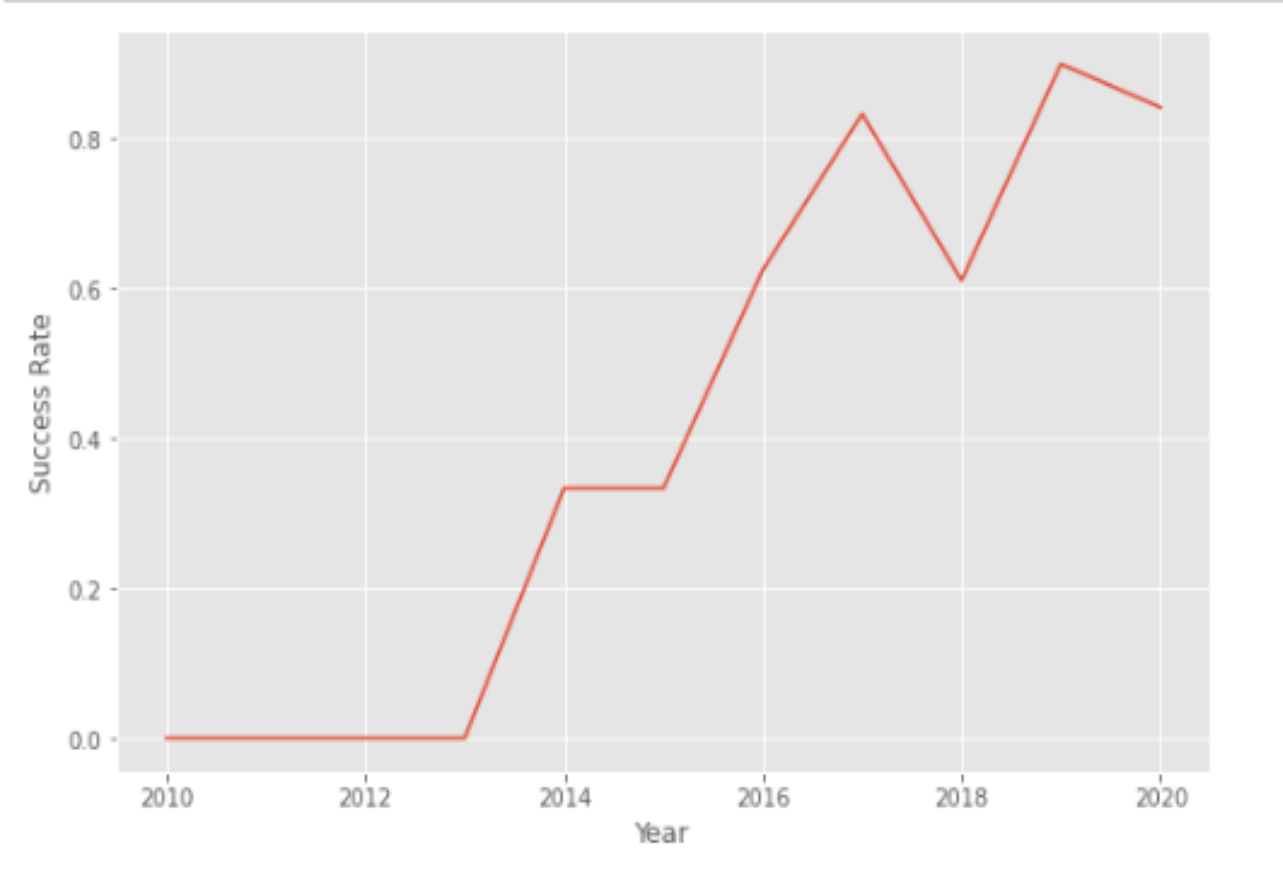
- In the LEO orbit, the success is positively correlated to the the number of flights.
- There seems to be no relationship between flight number in the GTO orbit.
- The SSO orbit has a 100% success rate however with fewer flights than the other orbits
- FLights numbers greater than 40 have a higher success rate than flight numbers between 0-40.

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



- 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

- 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
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 os 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

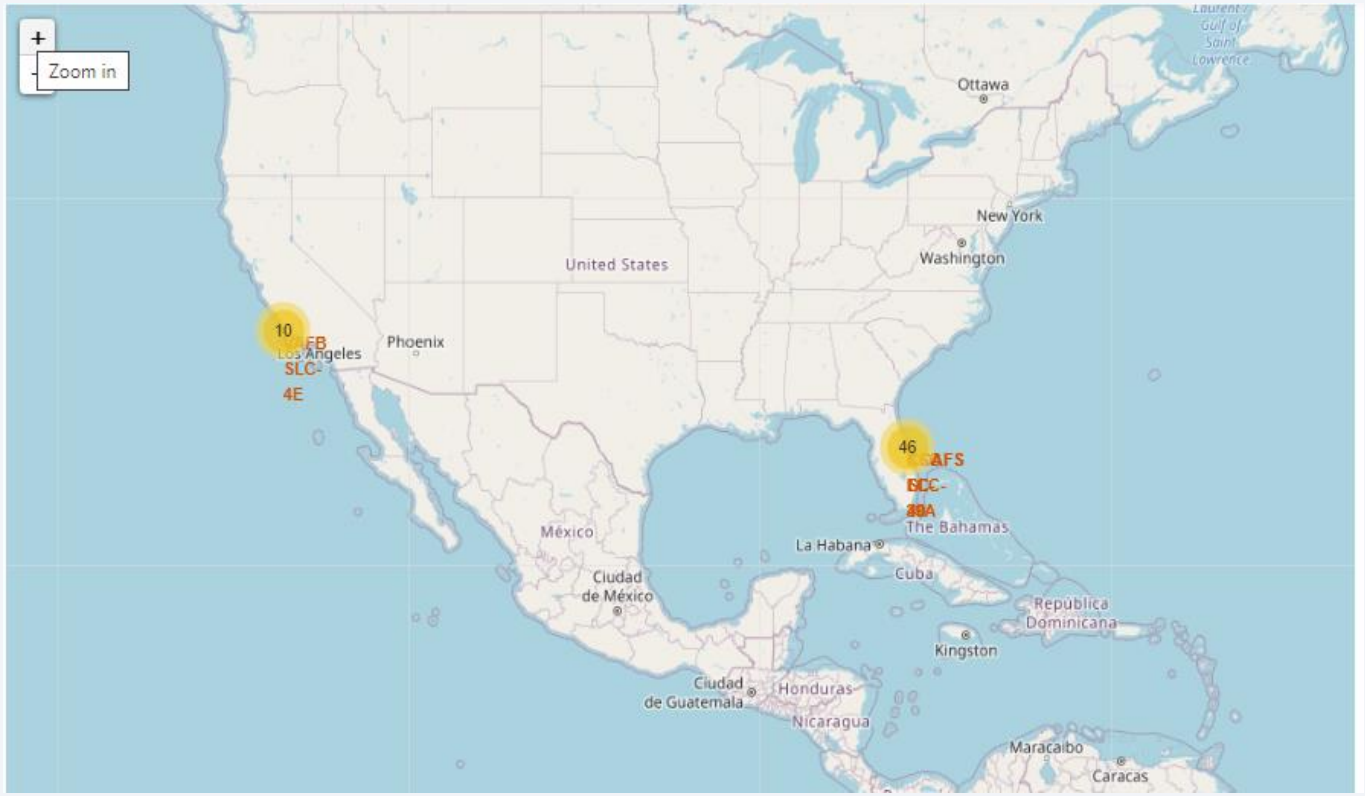
- COUNT() function was used to count the different *landing outcomes.* The WHERE and BETWEEN clauses filtered the results to only include results between 2010-06-04 and 2017-03-20. The GROUPBY clause ensure that the counts were grouped by their outcome. The ORDERBY and DESC clauses were used to sort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

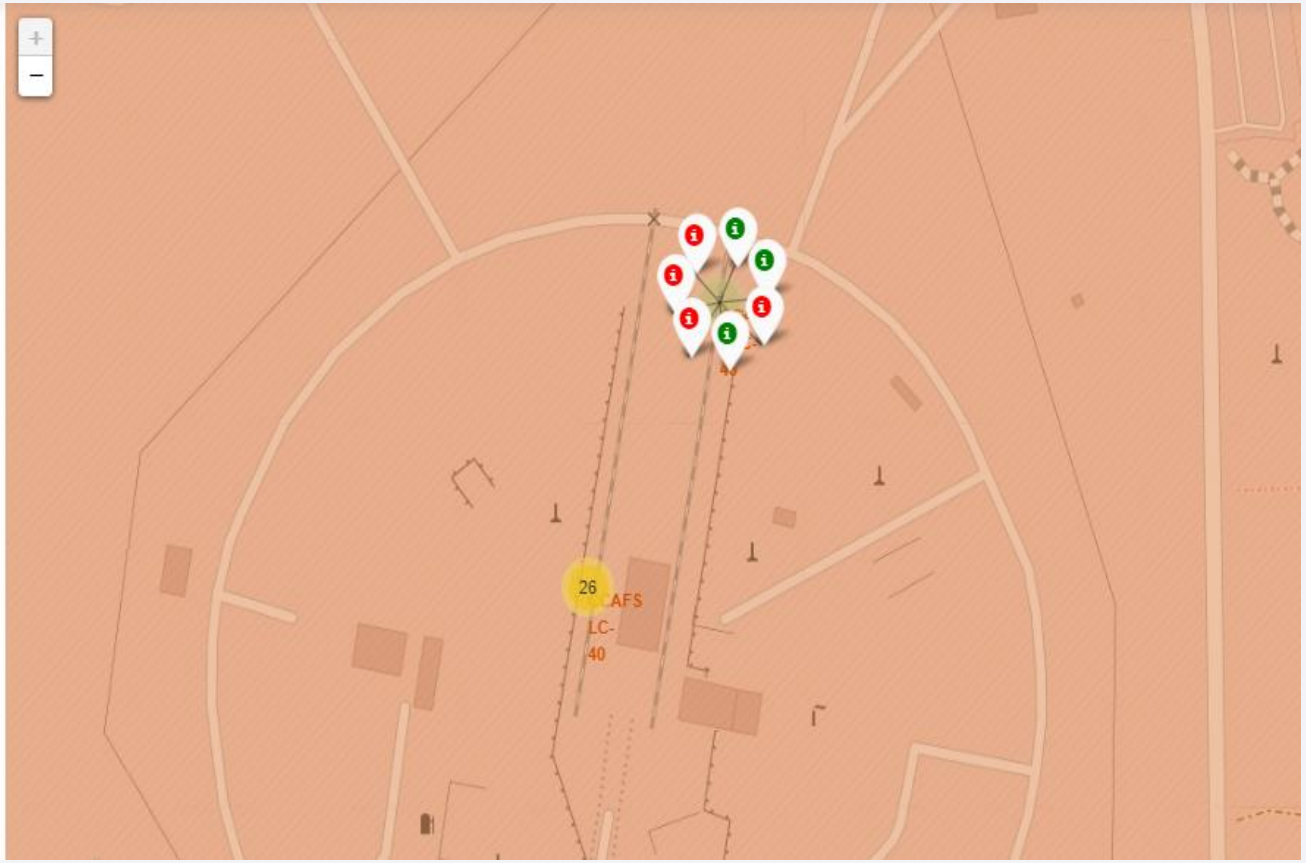
Launch Sites Proximities Analysis

SpaceX Launch Sites Locations



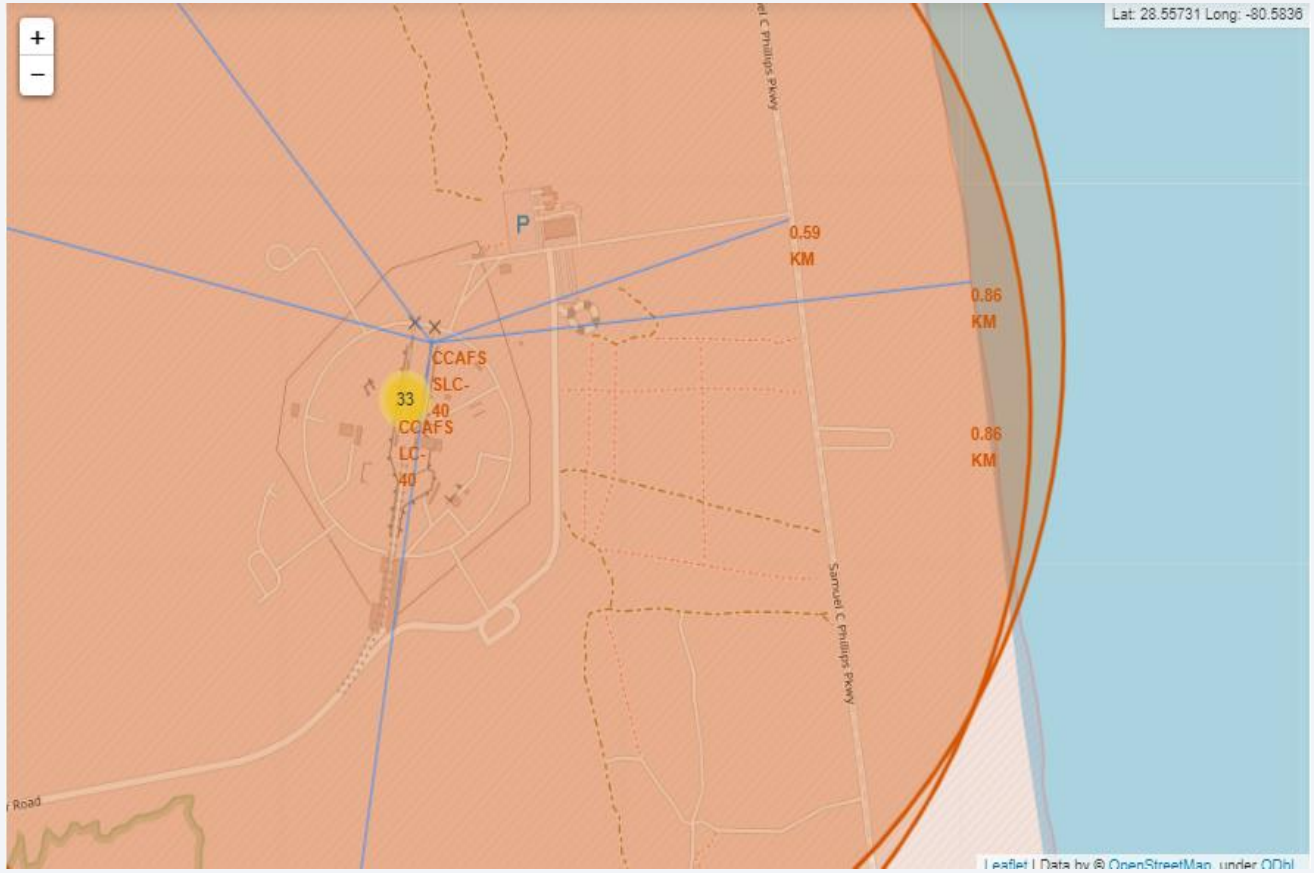
- The yellow markers are indicators of where the locations of all the SpaceX launch sites are situated in the US.
- The launch sites have been strategically placed near the coast

Success or Failure?

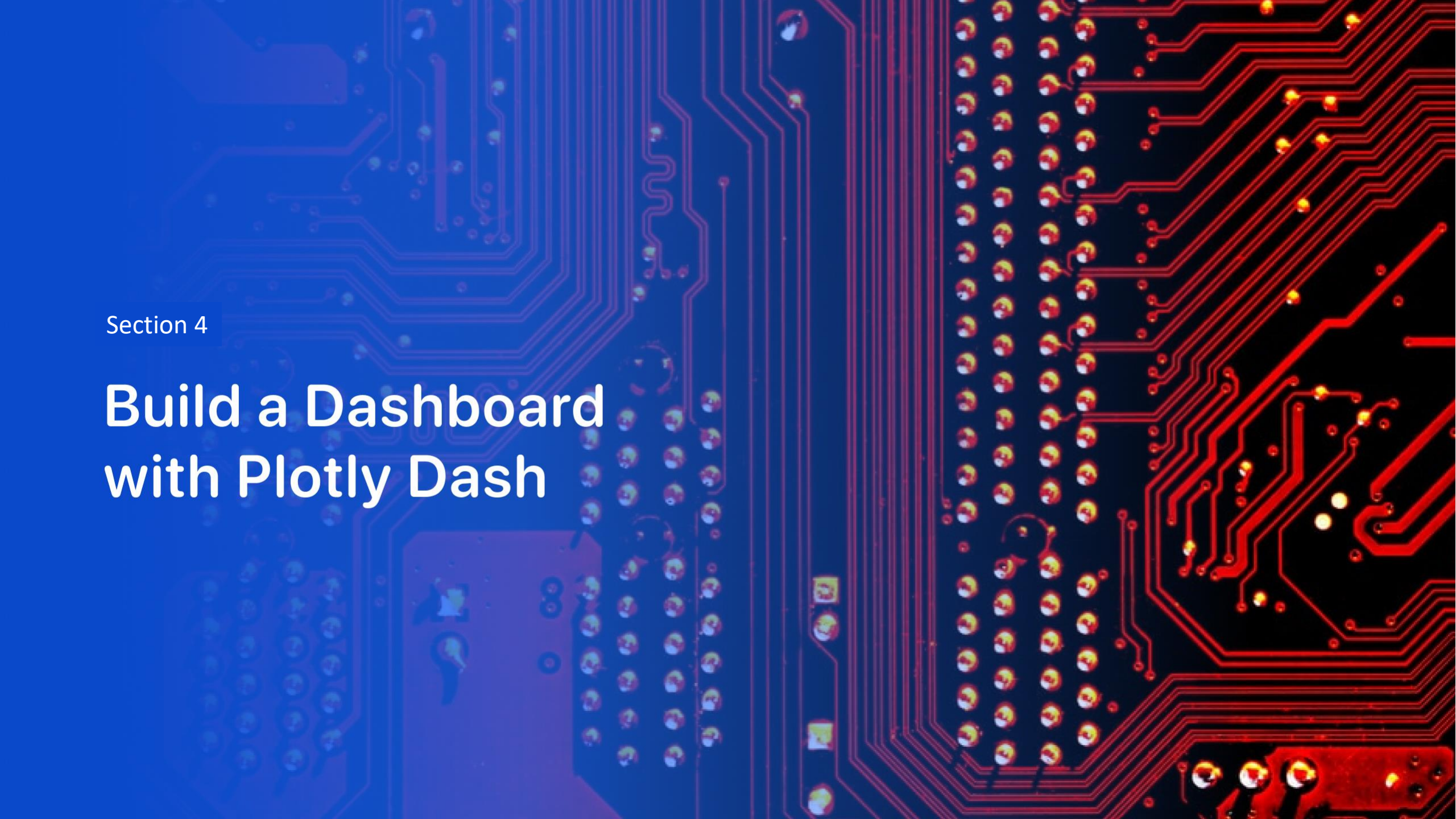


- When we zoom in on a launch site, we can click on the launch site which will display marker clusters of successful landings (green) or failed landing (red).

Launch Site Proximities

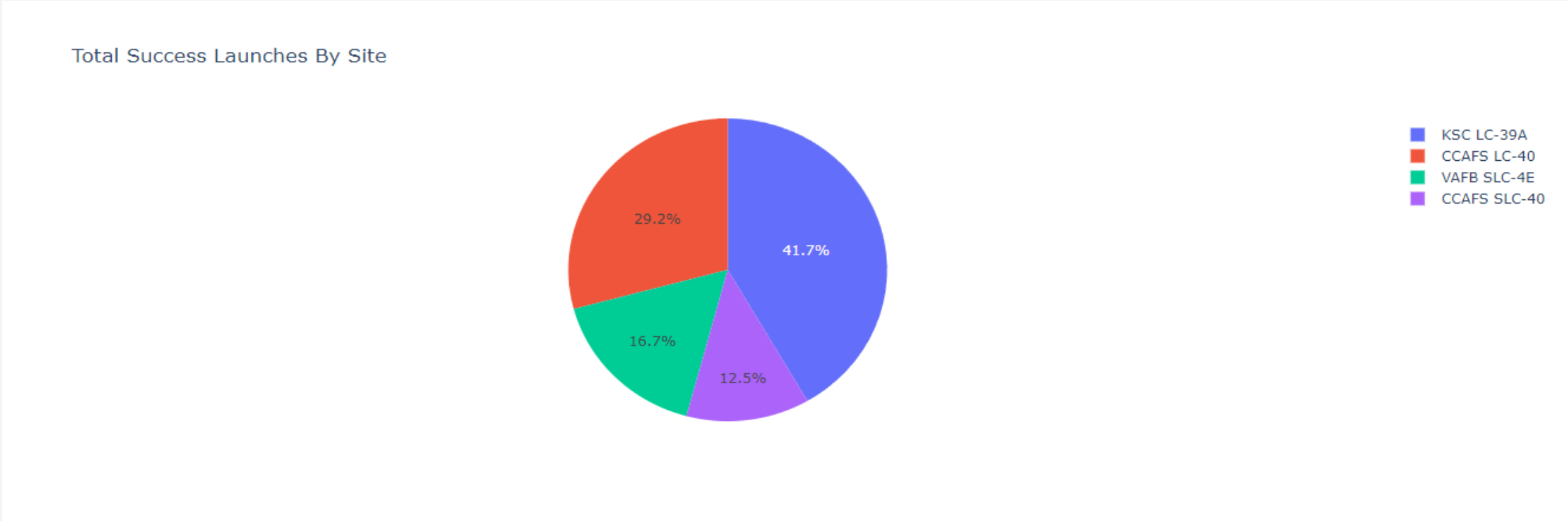


- The generated map shows that the selected launch site is close to a highway for transportation of personnel and equipment. The launch site is also close to the coastlines for launch failure testing.
- The launch sites also maintain a certain distance from the cities. (Can be viewed in notebook).


Section 4
Build a Dashboard with Plotly Dash

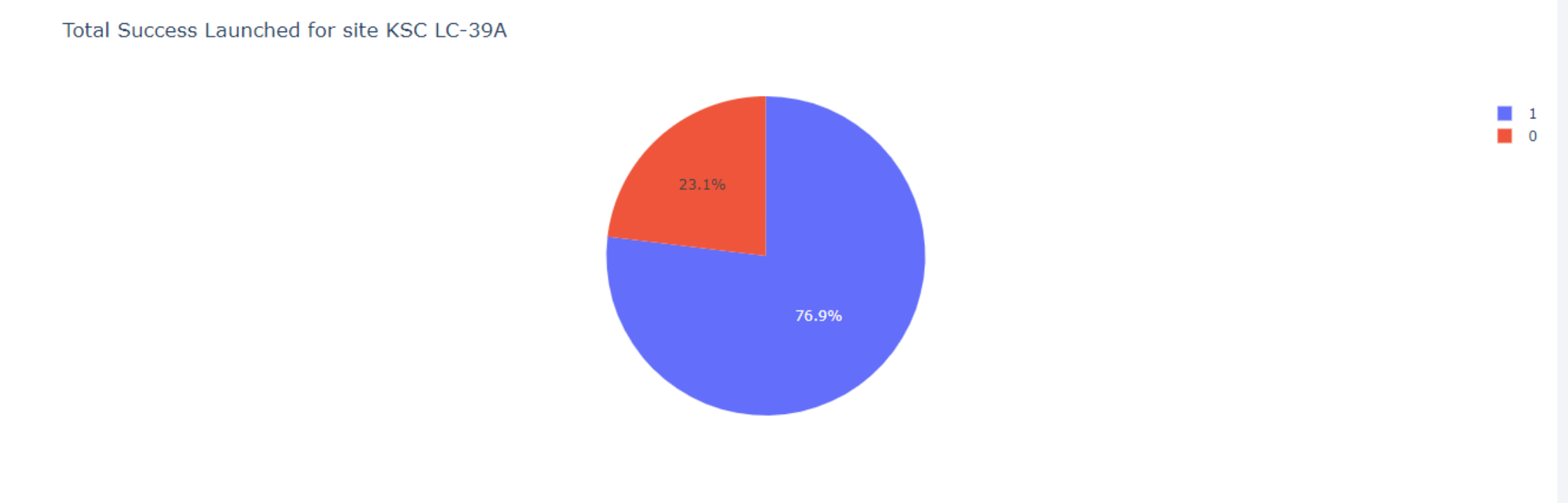
Total Successful Launches By Site

- 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

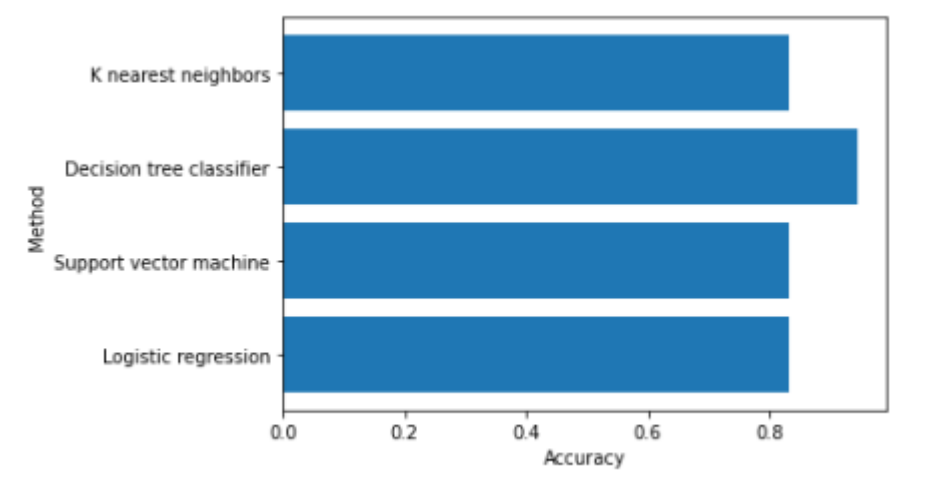
- The launch success rate for payloads 0-2500 kg is slightly lower than that of payloads 2500-5000 kg. There is in fact not much difference between the two.
- The booster version that has the largest success rate, in both weight ranges is the *v1.1*.

Section 5

Predictive Analysis (Classification)

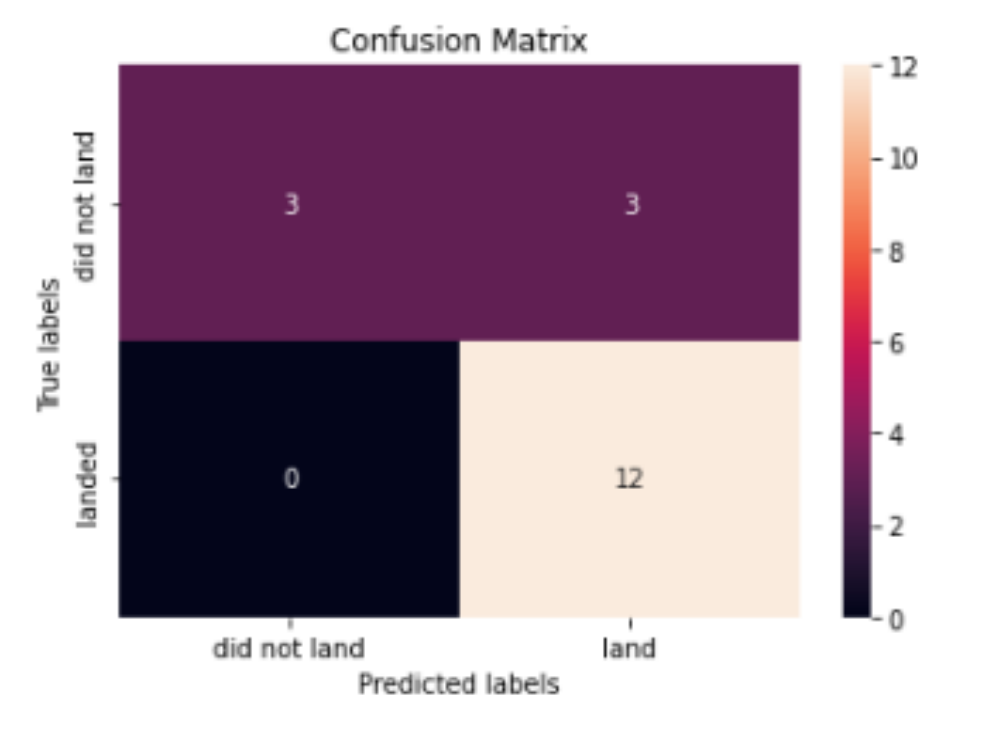
Classification Accuracy

- 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



- The model predicted 12 successful landings when the True label was successful (True Positive) and 3 unsuccessful landings when the True label was failure (True Negative).
- The model also predicted 3 successful landings when the True label was unsuccessful landing (False Positive).
- The model generally predicted successful landings.

Conclusions

- The analysis showed that there is a positive correlation between number of flights and success rate as the success rate has improved over the years.
- There are certain orbits like SSO, HEO, GEO, and ES-L1 where launches were the most successful.
- Success rate can be linked to payload mass as the lighter payloads generally proved to be more successful than the heavier payloads.
- The launch sites are strategically located near highways and railways for transportation of personel and cargo, but also far away from cities for safety.
- The best predictive model to use for this dataset is the Decision Tree Classifier as it had the highest accuracy with 94%.

Appendix

- Coursera Project Link: https://www.coursera.org/learn/applied-data-science-capstone/home/welcome
- GitHub Repository: https://github.com/Niaziz007/Data-Science-Capstone-Project

Thank you!