IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

OLABODE DUROJOLA
21/06/2023

Outline

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

Executive Summary

In this project, Space X 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 Space X 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 will be achieved by using different machine learning classification algorithms.

The methodolgy followed will include Data Dollection, Data Wrangling and Preprocessing, Exploratory Data Analysis(EDA) with SQL, Data Visualization, building an interactive map with Folium, and 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 objective of this capstone project is to develop a predictive model that can determine the likelihood of a successful landing for the Falcon 9 first stage. SpaceX takes pride in its ability to reuse the first stage of its rockets, resulting in significant cost savings. In fact, they advertise on their website that their rocket launches cost $62 million, while other providers charge upwards of $165 million. The key to these savings lies in the reusability of the first stage. By accurately predicting whether the first stage will land successfully, we can estimate the cost of a launch. This information can be valuable for alternate companies interested in bidding against SpaceX for rocket launches.

Hence, the central question we aim to address is: Given a set of features related to a Falcon 9 rocket launch, can we determine the probability of a successful landing for the first stage?

Section 1

Methodology

Methodology

Executive Summary

The data for this project was gathered using two primary methods: retrieving data from the SpaceX API and performing web scraping on a Wikipedia page containing launch data. Python's pandas library was utilized for data wrangling, allowing for data transformation and cleaning.

Following data cleaning, exploratory data analysis (EDA) was conducted using visualization tools such as matplotlib and seaborn libraries in Python. Additionally, SQL queries were employed to address specific analytical questions. Interactive visualization packages in Python, including Folium for map creation and Plotly Dash for interactive data visualizations, were utilized to provide a more engaging analysis.

To perform predictive analysis, four distinct machine learning classification models were employed: logistic regression, support vector machines, k-nearest neighbor, and decision tree classifier. Each model underwent training, tuning, and evaluation in order to identify the optimal choice for predicting the successful landing of the Falcon 9 first stage.

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. Clean 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: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:Data wrangling

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 RGate* 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: EDA with Data Visualization

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: EDA with SQL

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: Interactive Map Analytics with Folium

Build a Dashboard with Plotly Dash

The dashboard application contains two charts:

- A pie chart 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 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:SpaceX 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:Machine Learning Prediction

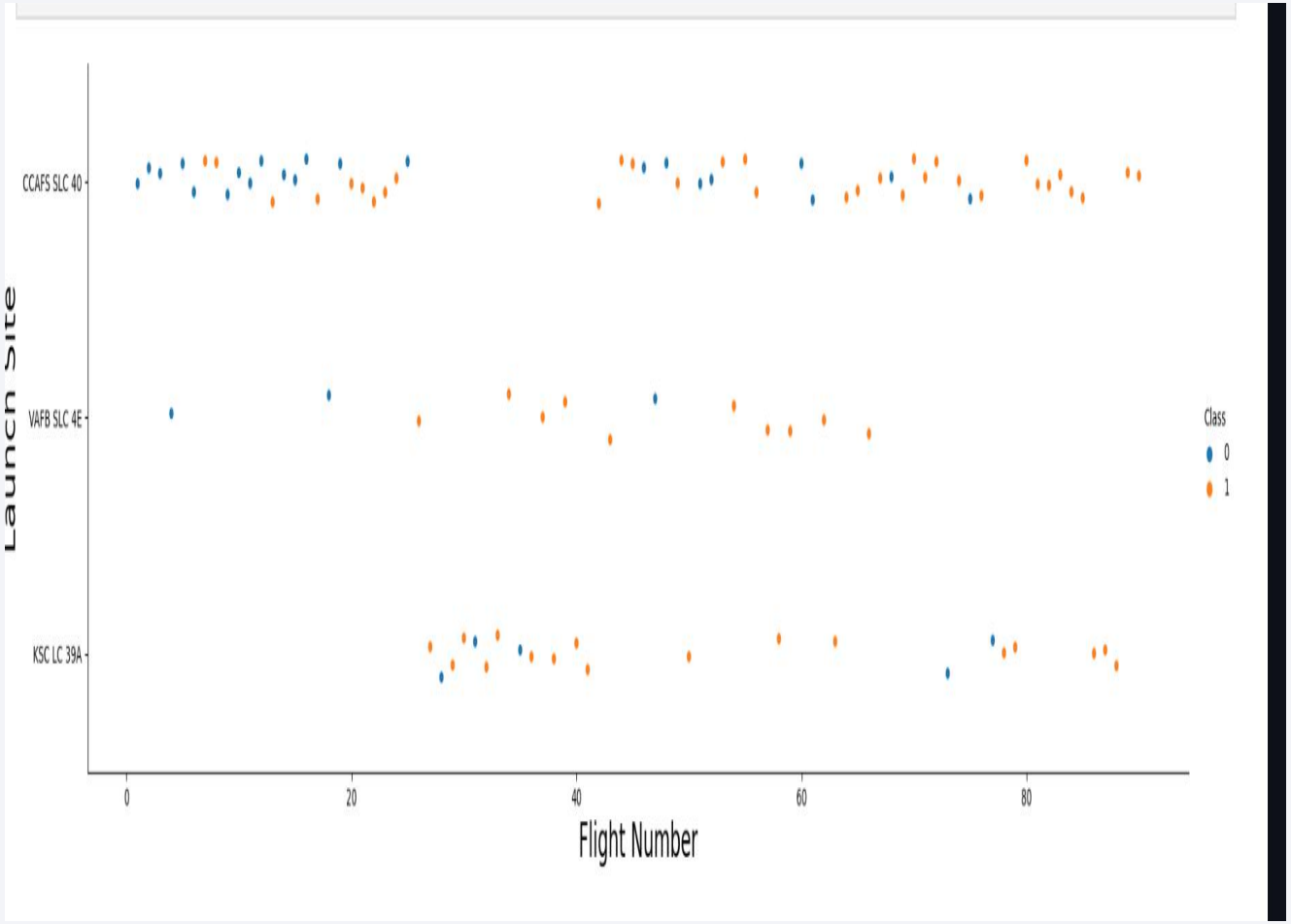
Results

- 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

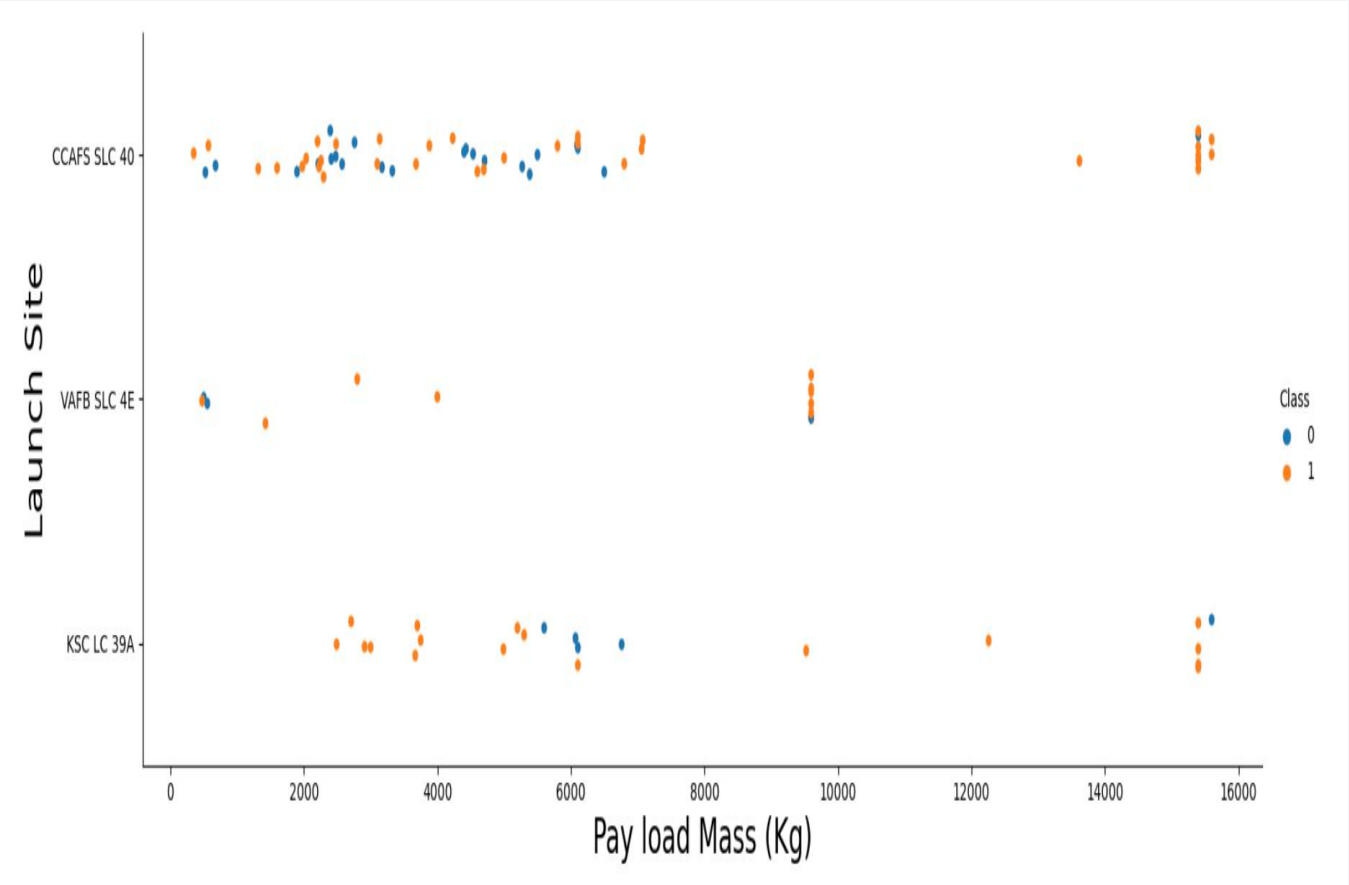
Flight Number vs. Launch Site



- 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.

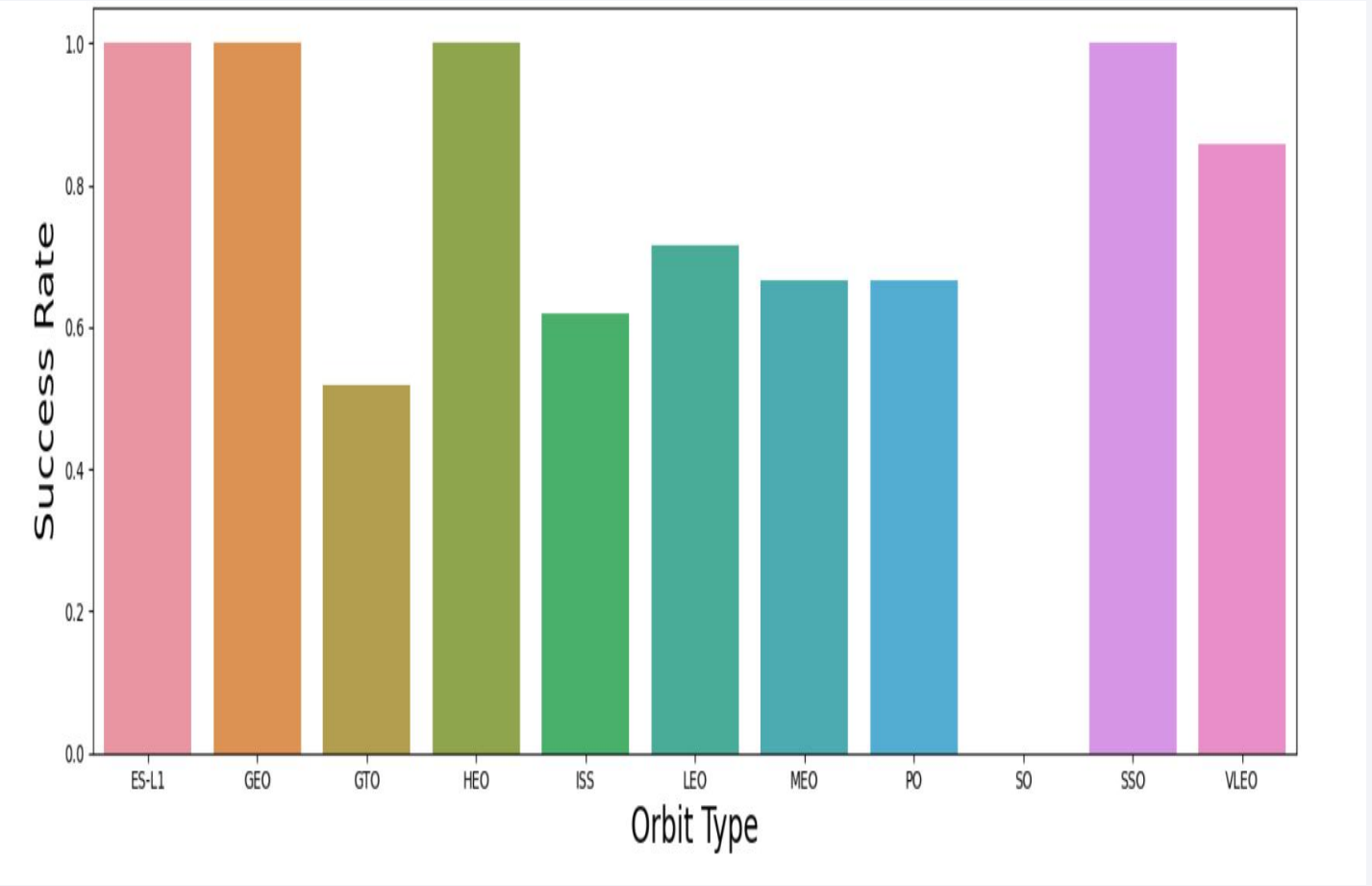
Payload vs. Launch Site

- 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.



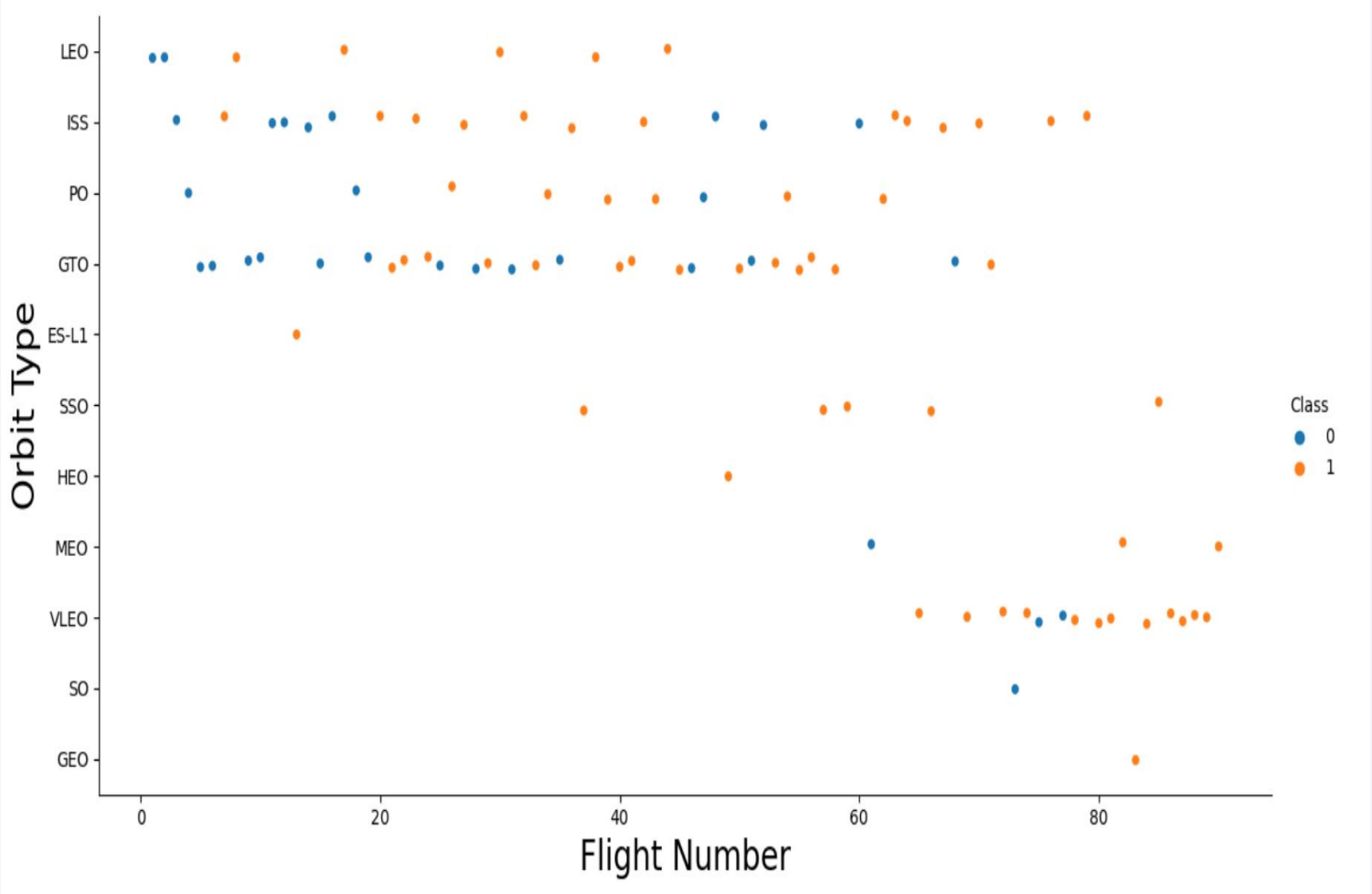
Success Rate vs. Orbit Type

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



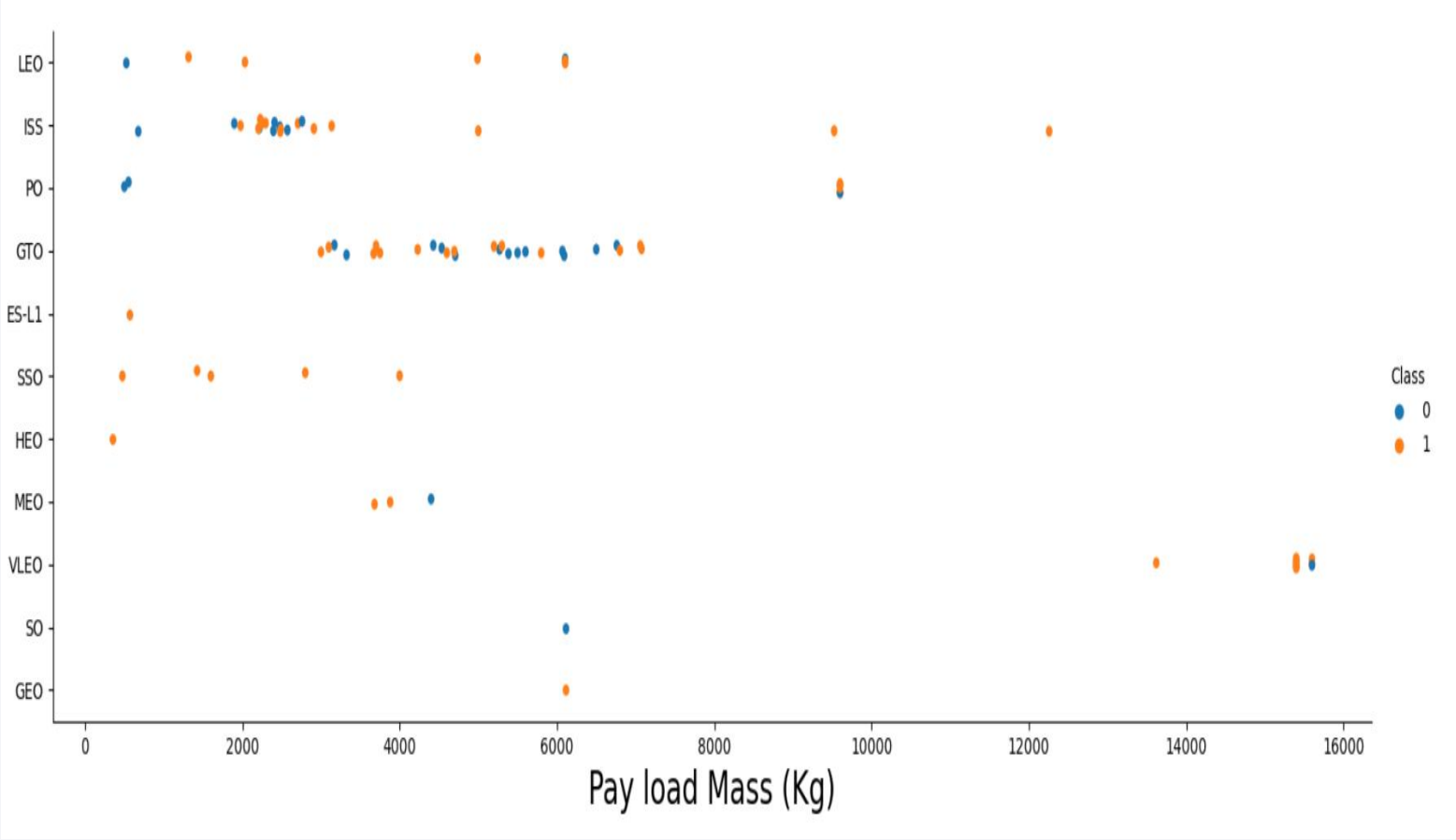
Flight Number vs. Orbit Type

- n 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.



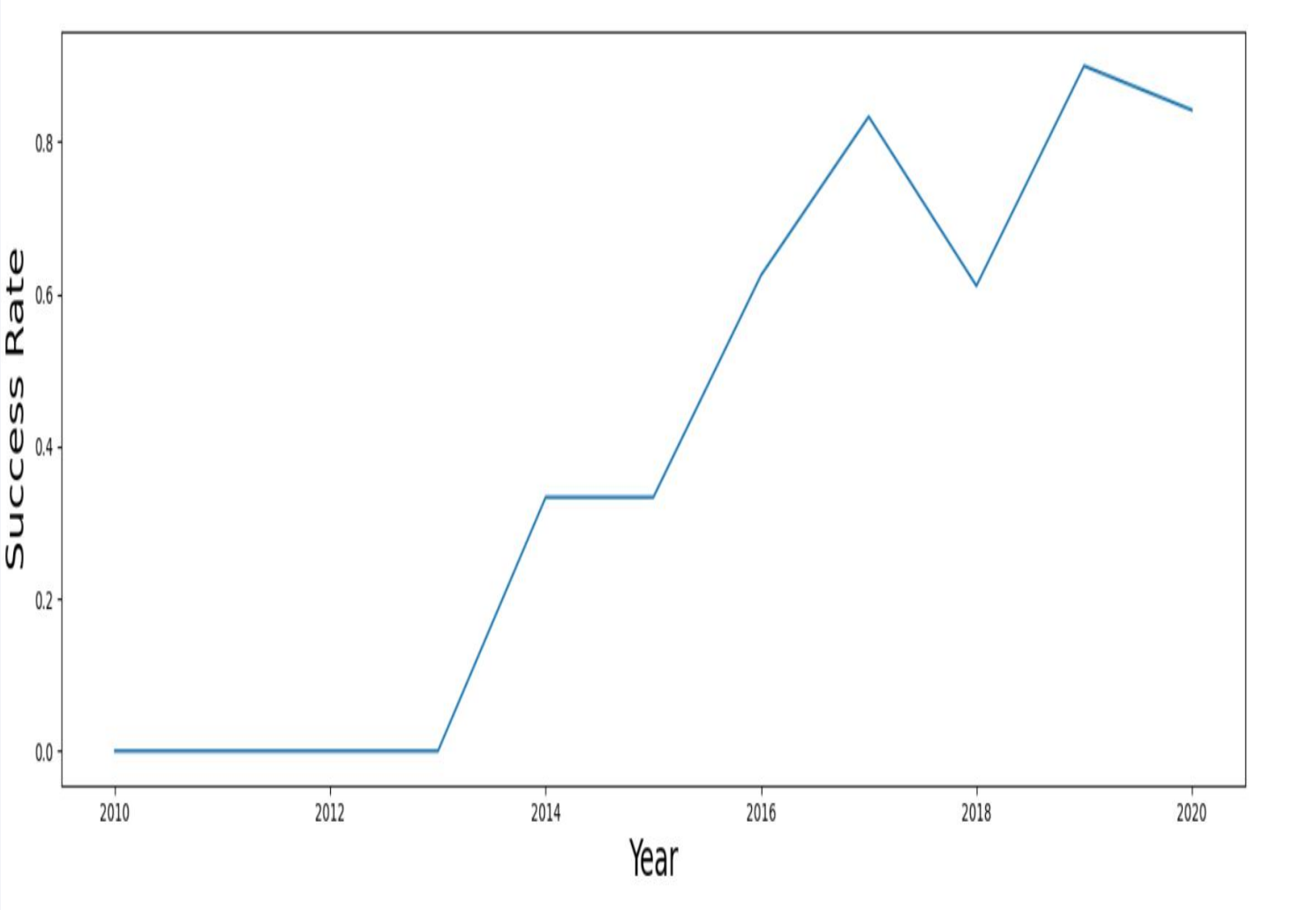
Payload vs. Orbit Type

- 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 presen



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
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

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	Lan
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Fai
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Fai
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

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
45596.0

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"

Average Payload mass
2928.4

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)'

min(Date)
01/08/2018

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 B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1

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	count(mission_outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
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	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0

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.

month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	01/10/2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14/04/2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

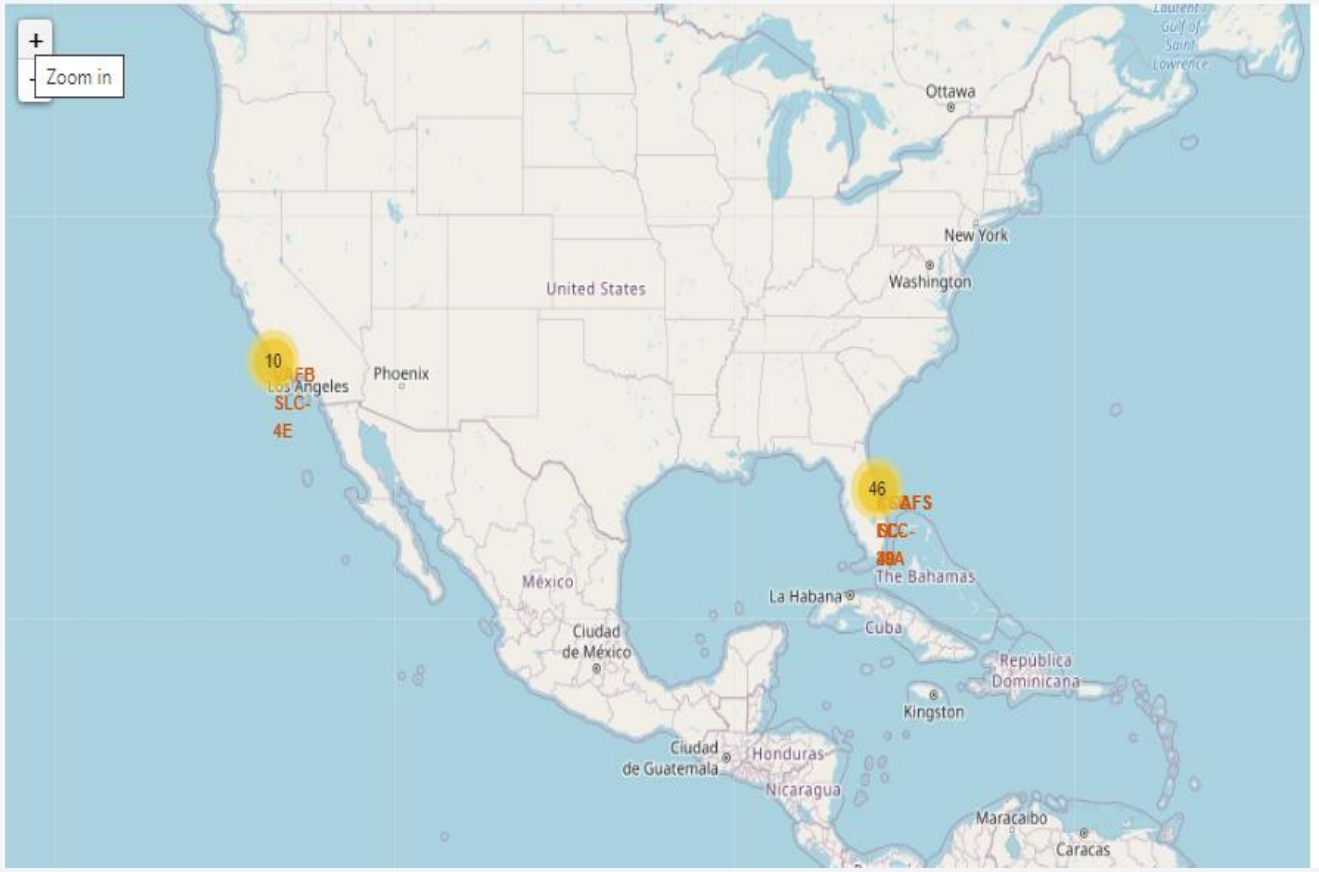
- COUNT() function was used to count the different *landing outcomes.* The WHERE and BETWEEN clauses filtered the results to only include results between 2010-06-04 and 2017-03-20.
- The GROUPBY clause ensure that the counts were grouped by their outcome. The ORDERBY and DESC clauses were used to sort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

Landing_Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1


Section 3
Launch Sites
Proximities Analysis

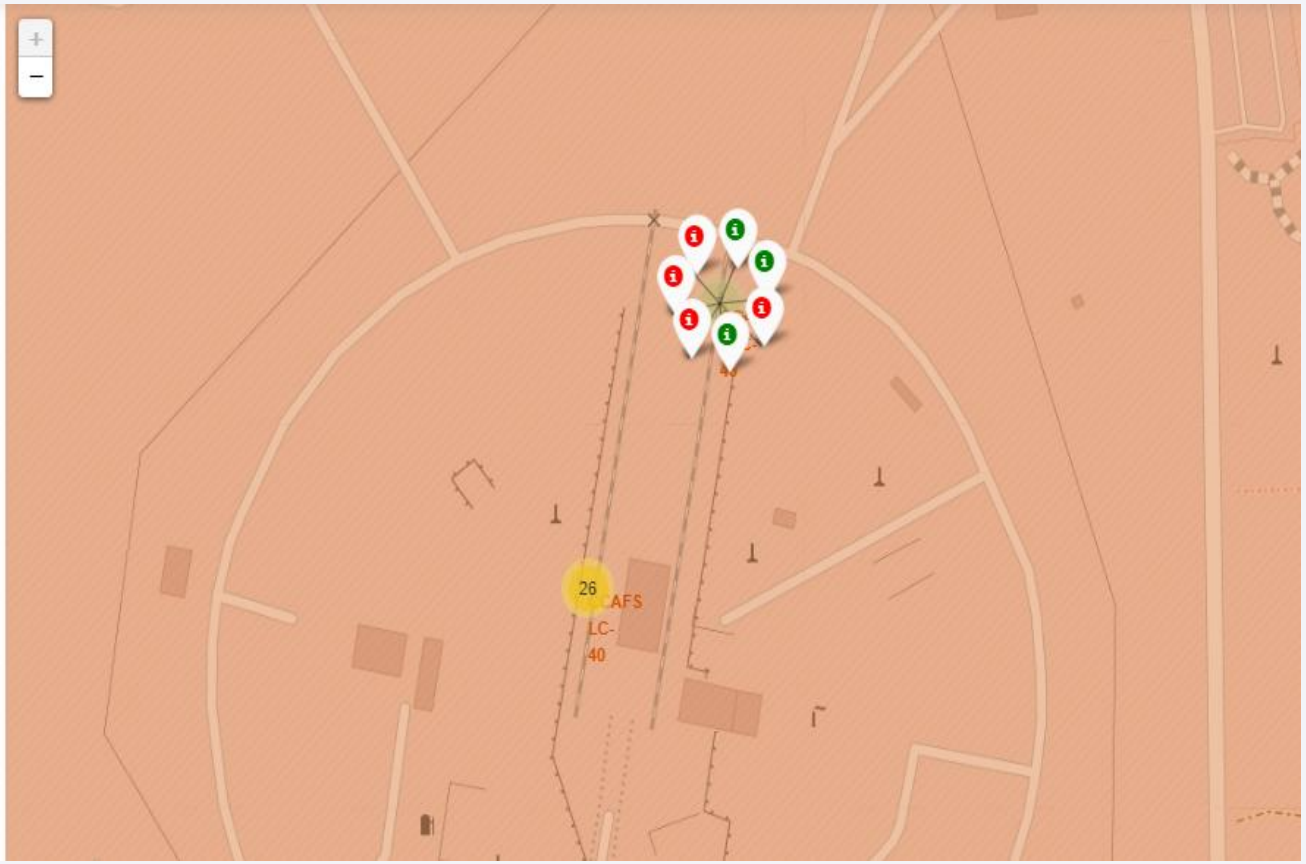
SpaceX Launch Sites Locations

- 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



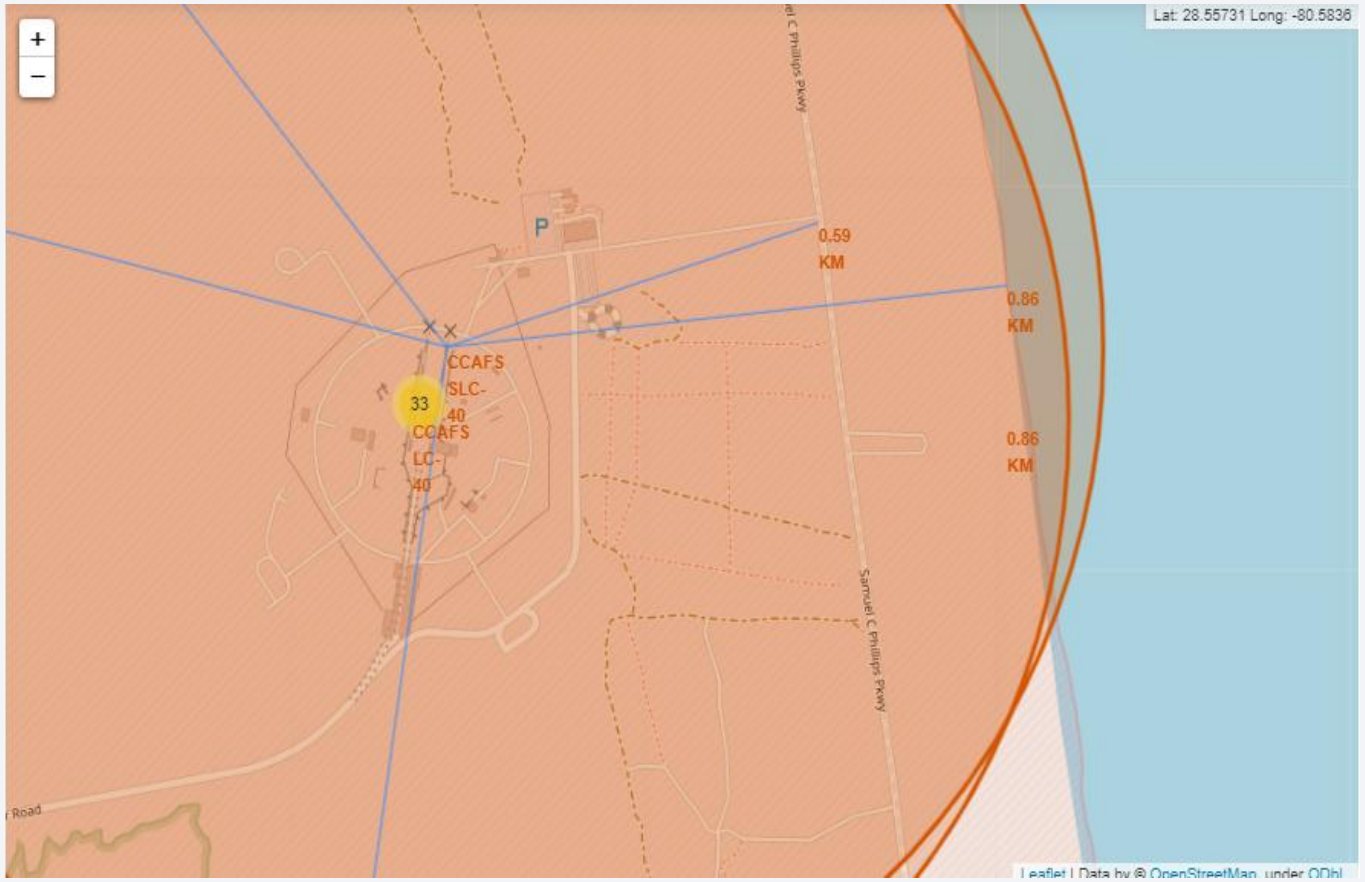
Success or Failure?

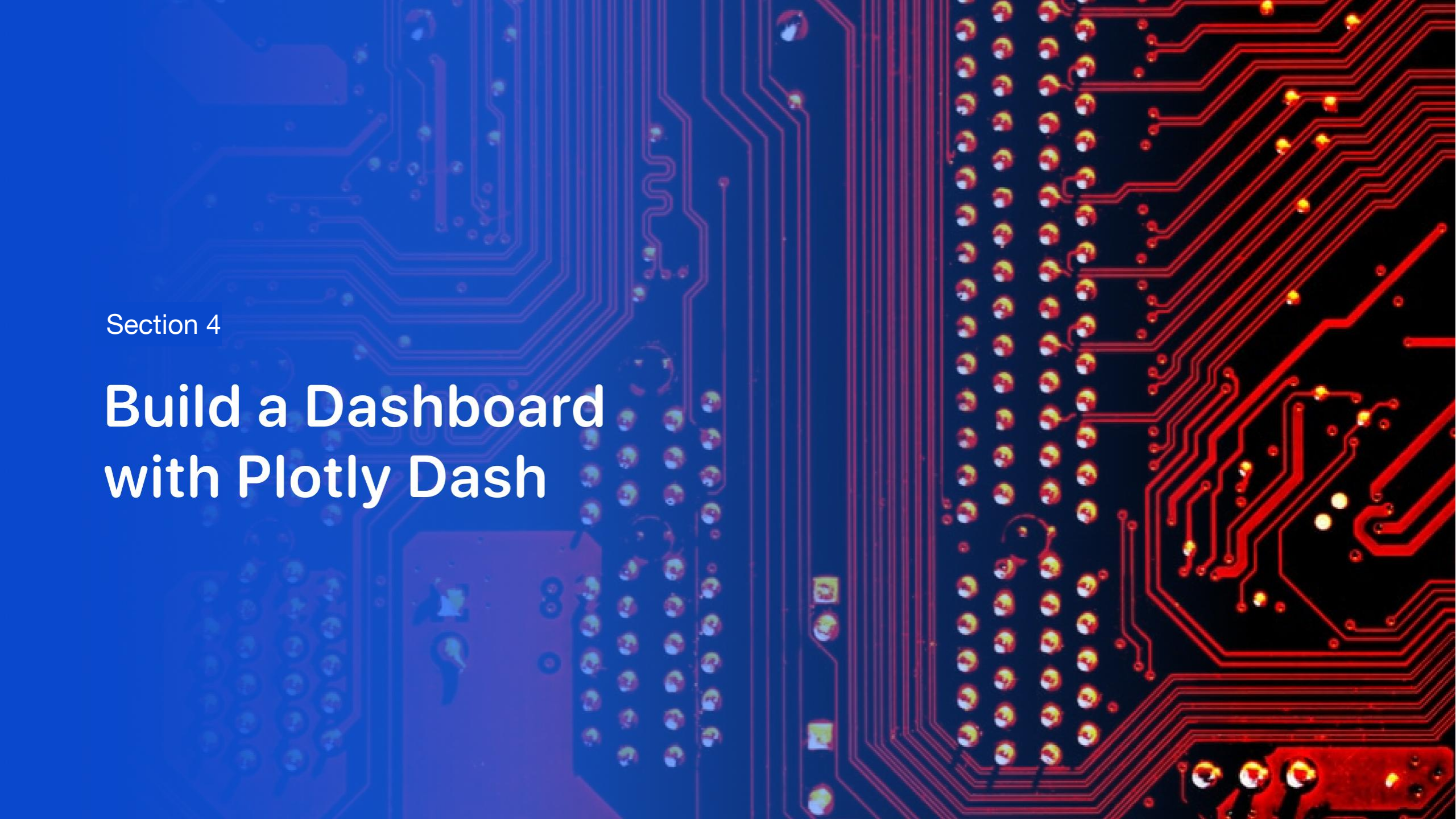
- When we zoom in on a launch site, we can click on the launch site which will display marker clusters of successful landings (green) or failed landing (red).



Launch Site Proximities

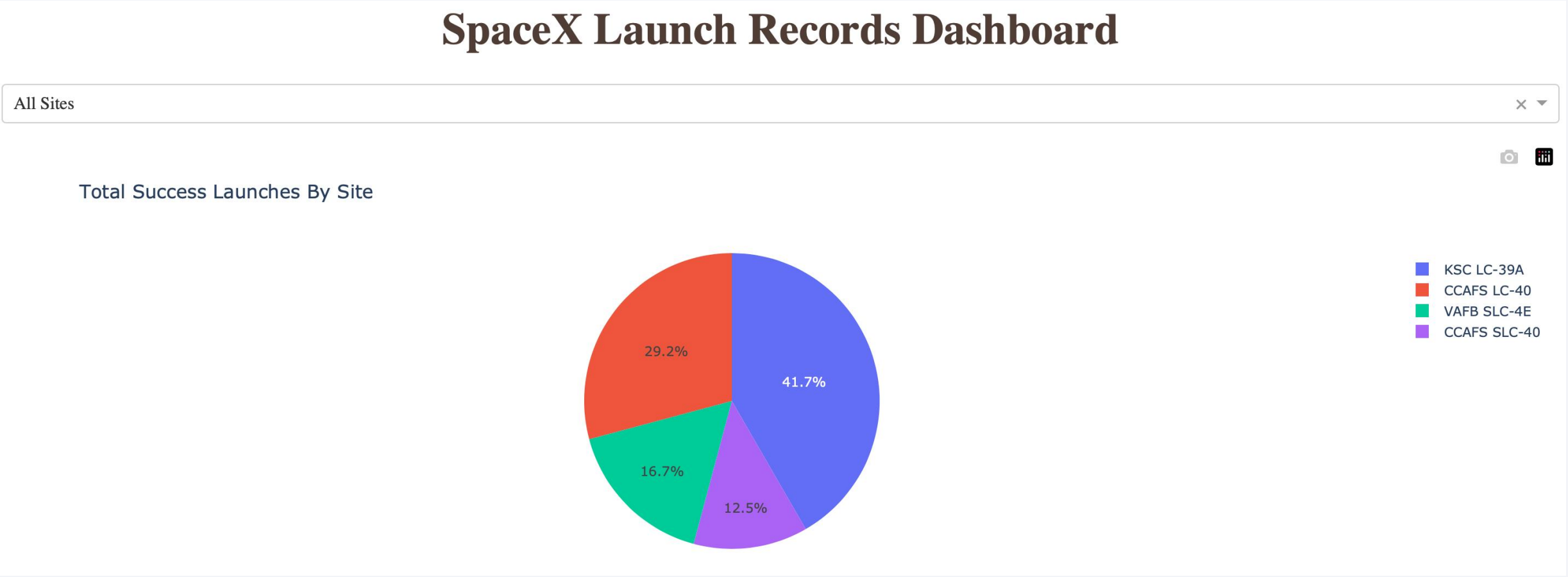
- The generated map shows that the selected launch site is close to a highway for transportation of personnel and equipment. The launch site is also close to the coastlines for launch failure testing.
- The launch sites also maintain a certain distance from the cities. (Can be viewed in notebook).




Section 4
Build a Dashboard with Plotly Dash

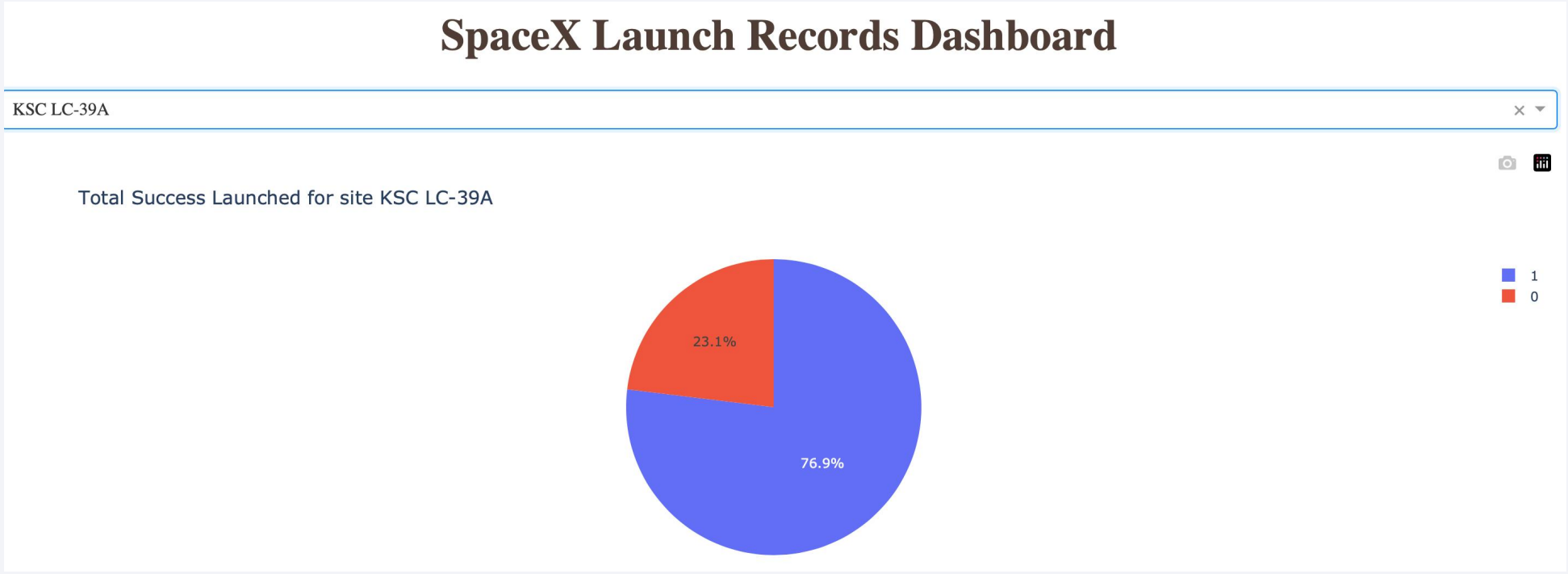
Total Successful Launches by Site

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



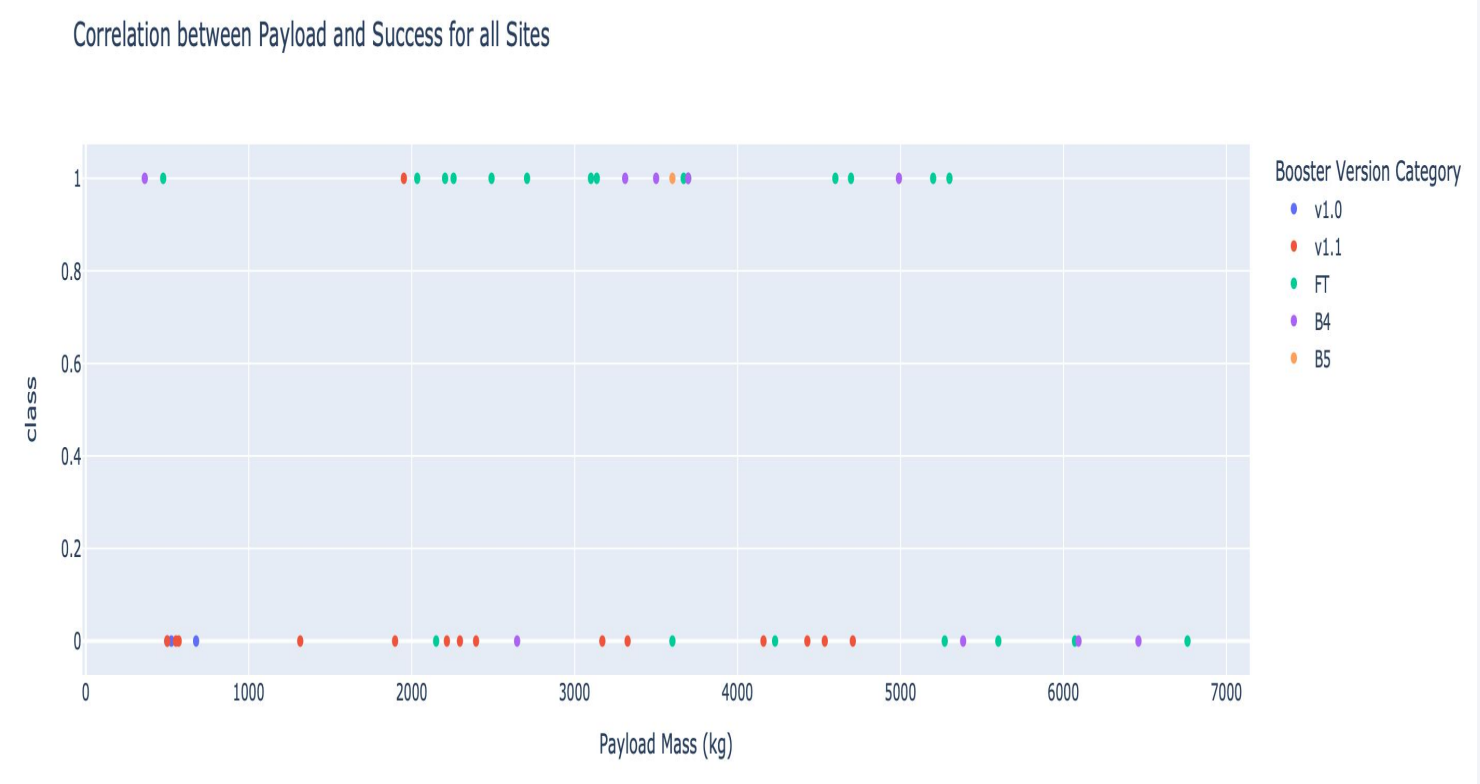
Launch Site With Highest Success Ratio

- 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*.

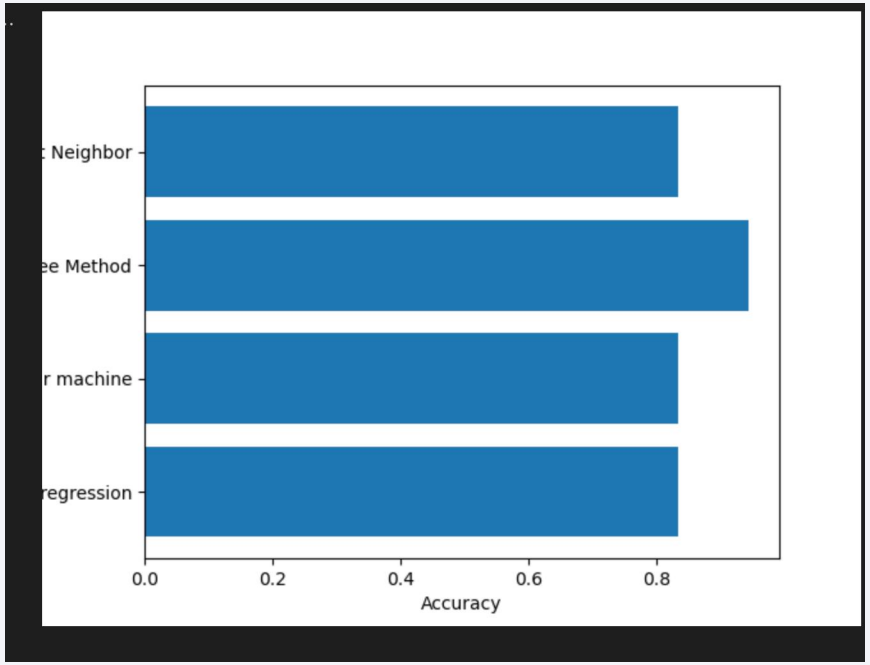


Section 5

Predictive Analysis (Classification)

Classification Accuracy

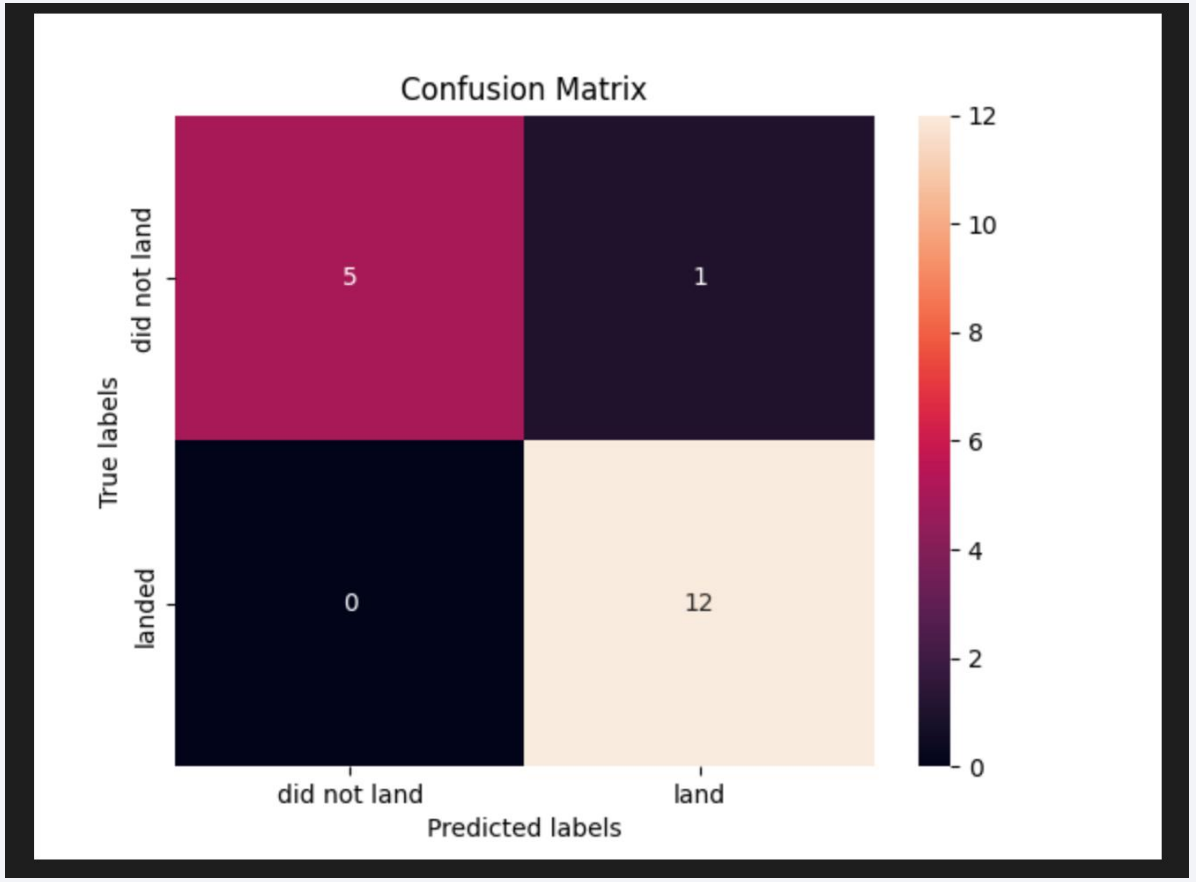
- The Decision Tree classifier had the best accuracy at 94%.



	method	accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision Tree Method	0.944444
3	K Nearest Neighbor	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

Cousera Project link: https://www.coursera.org/learn/applied-data-science-capstone/home/

Github Repository: https://github.com/Bondeystephen/projectSpaceX/tree/main

Thank you!