IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ronald Mbanga

19 Jan 2023

Outline

Executive Summary

Summary of methodologies

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Summary of all results

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

Introduction

Project background and context

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.

Problems you want to find answers

What are the factors that determine if the rocket will land successfully?

Will the rocket land successfully in the first stage given a set of features?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Data was collected using SpaceX API and web scraping from Wikipedia.

- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API

- Request and parse the SpaceX launch data using the GET request
- Normalize JSON response into a DataFrame
- Extract only useful columns using auxiliary functions
- Create new pandas DataFrame from dictionary
- Filter DataFrame to only include Falcon 9 launches
- Handle missing values
- Export to CSV file

1. https://github.com/rmbanga/Space-X-Landing-Success-Prediction/blob/main/1.%20SpaceX%20Data%20Collection%20API.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

https://github.com/rmbanga/Space-X-Landing-Success-Prediction/blob/main/2.%20Data%20Collection%20with%20Web%20Scraping.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 occurrence of mission outcome per orbit type
4. Create a landing outcome label from Outcome column using one-hot encoding
5. Export to CSV

EDA with Data Visualization

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.
- 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.*

https://github.com/rmbanga/Space-X-Landing-Success-

Prediction/blob/main/5.%20EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:
  - The names of unique launch sites in the space mission.
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.

https://github.com/rmbanga/Space-X-Landing-Success-Prediction/blob/main/4.%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities. We answered some question for instance:
  - Are launch sites near railways, highways and coastlines.
  - Do launch sites keep certain distance away from cities.

https://github.com/rmbanga/Space-X-Landing-Success-Prediction/blob/main/6.%20Interactive%20Map%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

We built an interactive dashboard with Plotly dash:

- 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,

https://github.com/rmbanga/Space-X-Landing-Success-Prediction/blob/main/7.%20Space-X%20Dashboard.ipynb

Predictive Analysis (Classification)

1. We loaded the data using NumPy and Pandas, transformed the data, split our data into training and testing.
2. We built different machine learning models and tune different hyperparameters
3. Find best Hyperparameter for SVM, Decision Trees, K-Nearest Neighbours and Logistic Regression.
4. Use test data to evaluate models based on their accuracy scores and confusion matrix and we found the best performing classification model

https://github.com/rmbanga/Space-X-Landing-Success-Prediction/blob/main/8.%20Space-X%20Machine%20Learning%20Prediction.ipynb

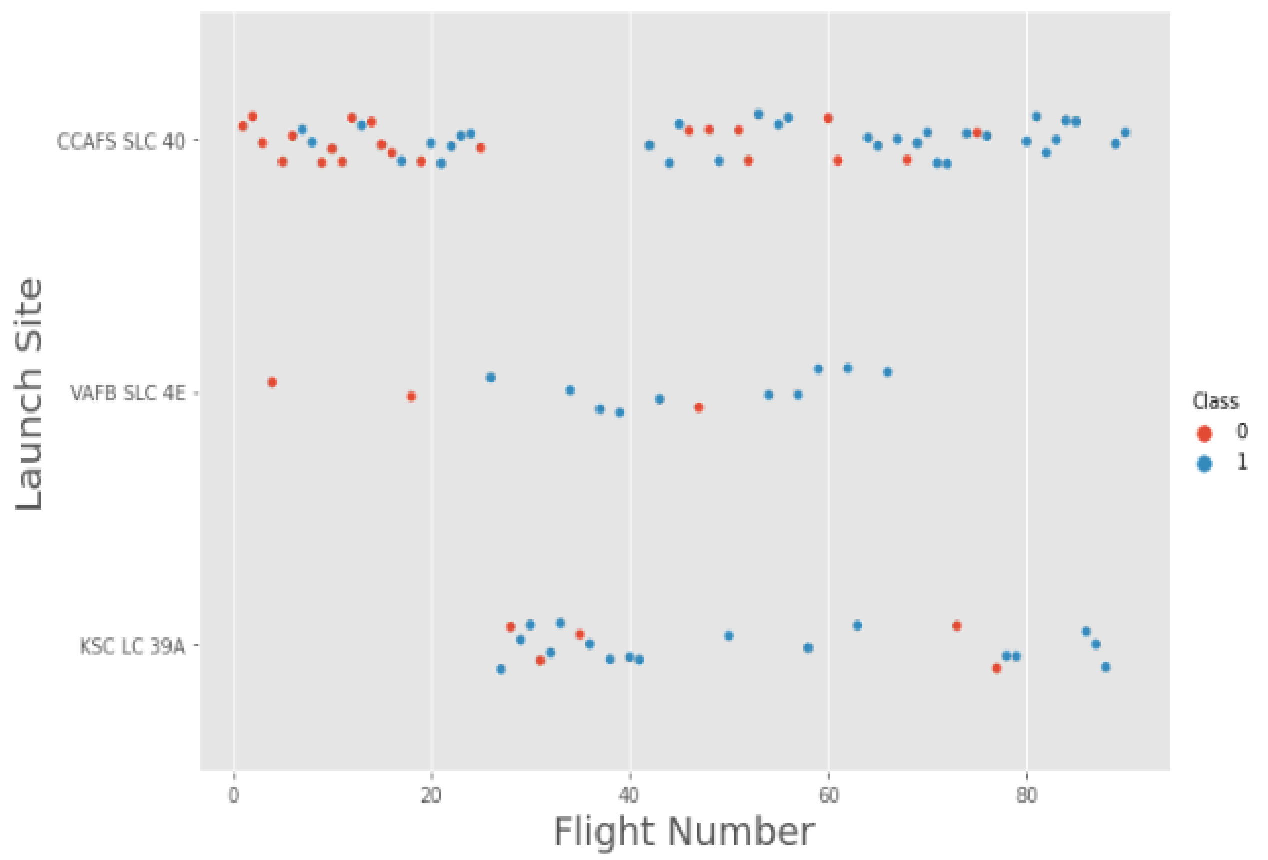
Results

- 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

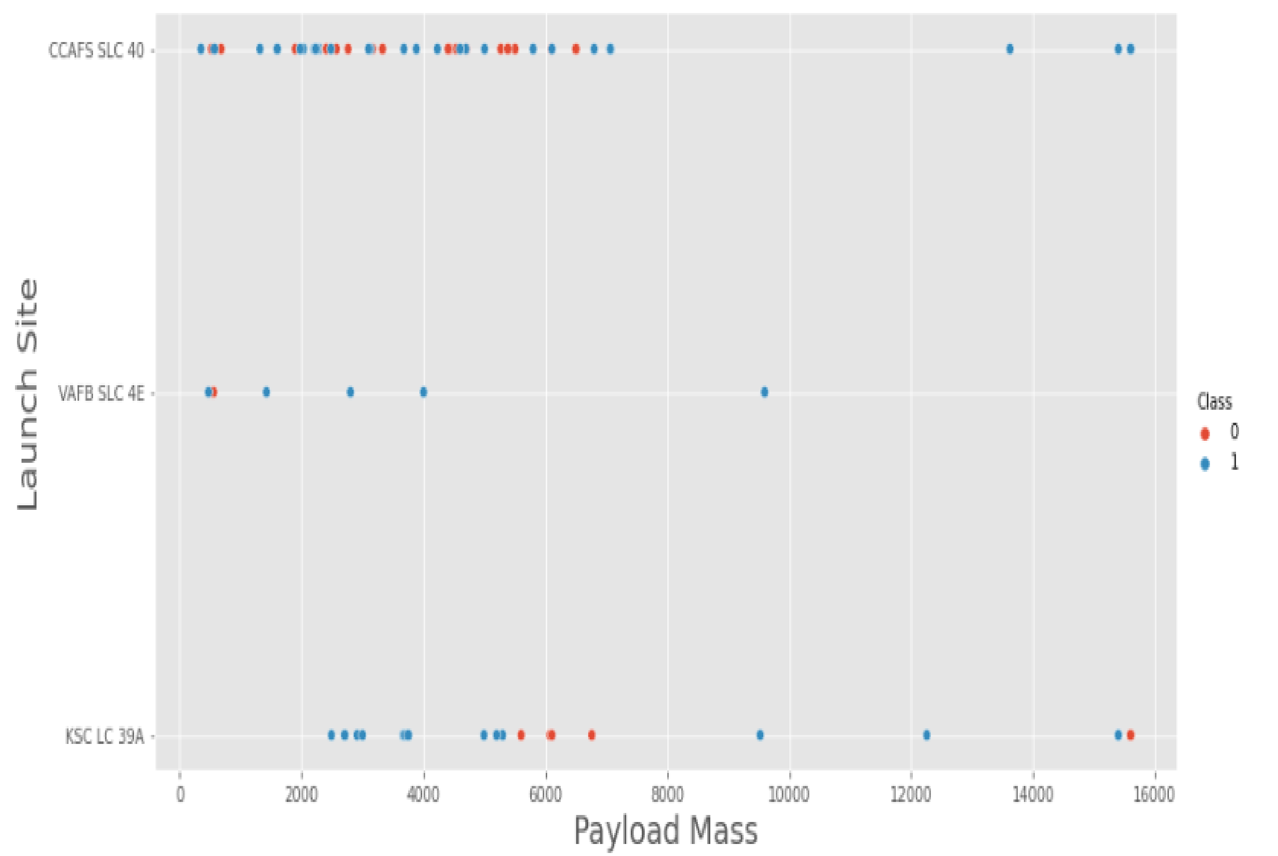
Insights drawn from EDA

Flight Number vs. Launch Site



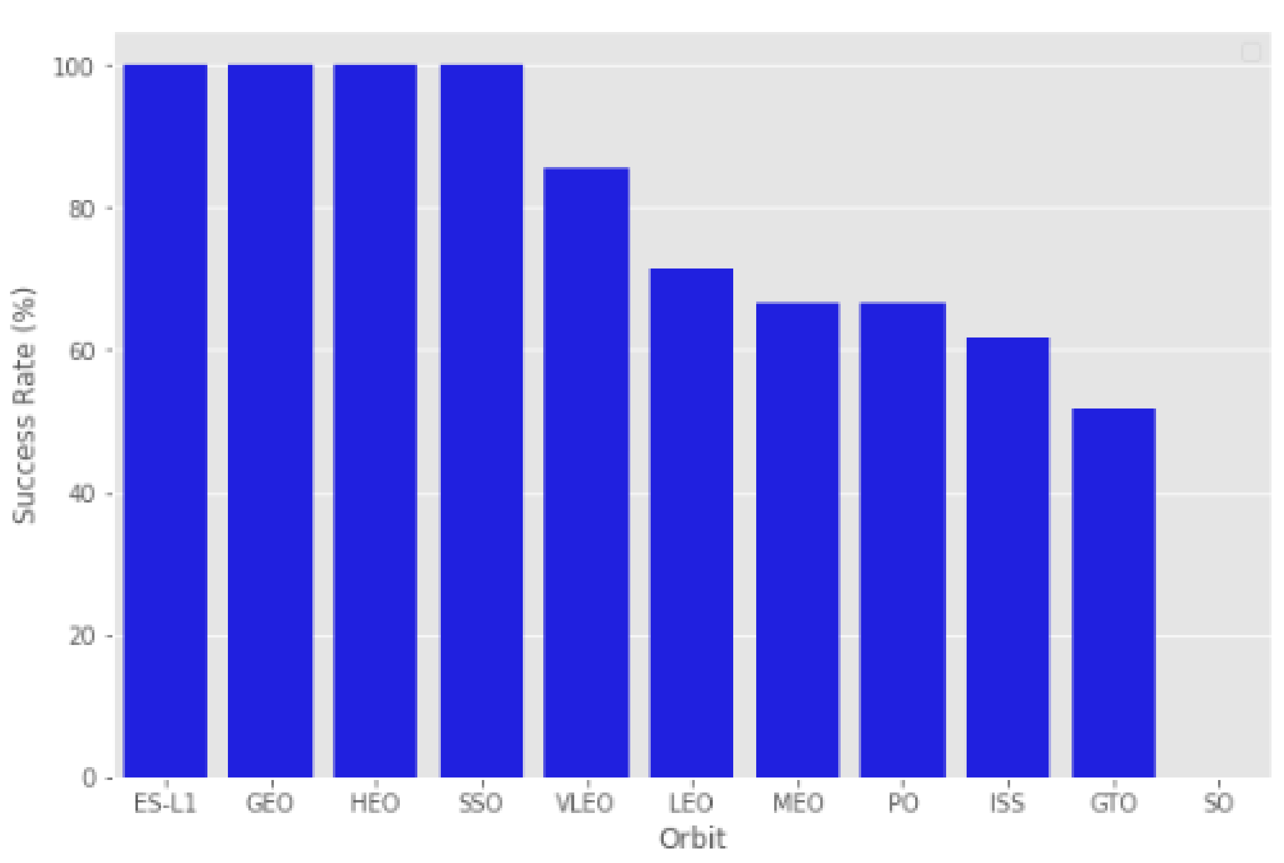
- The blue dots represent the successful launches while the red dots represent unsuccessful launches.
- The success of landing increased as flight number increased.

Payload vs. Launch Site



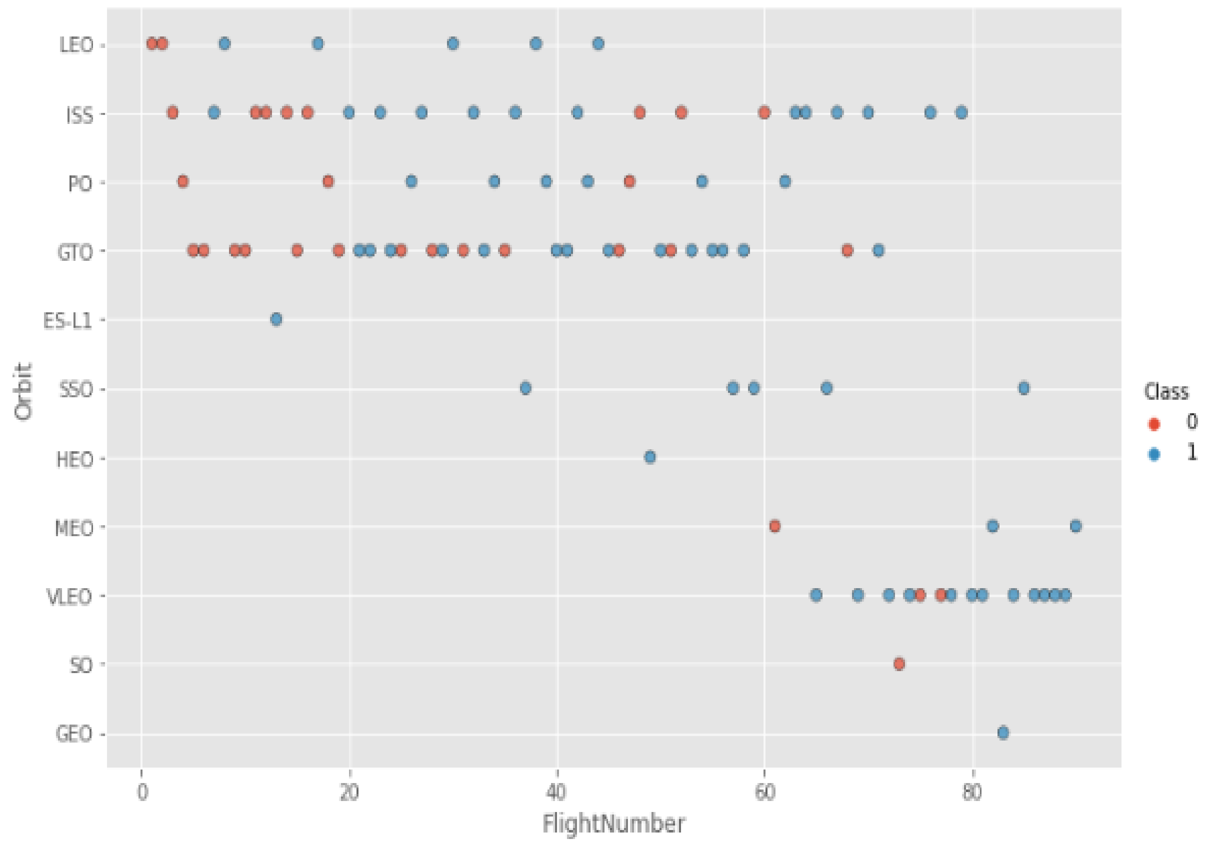
- Most of the launches were carried out at CCSFS SLC-40, followed by KSC LC-39A and least launches were carried out at VAFB SLC 4E

Success Rate vs. Orbit Type



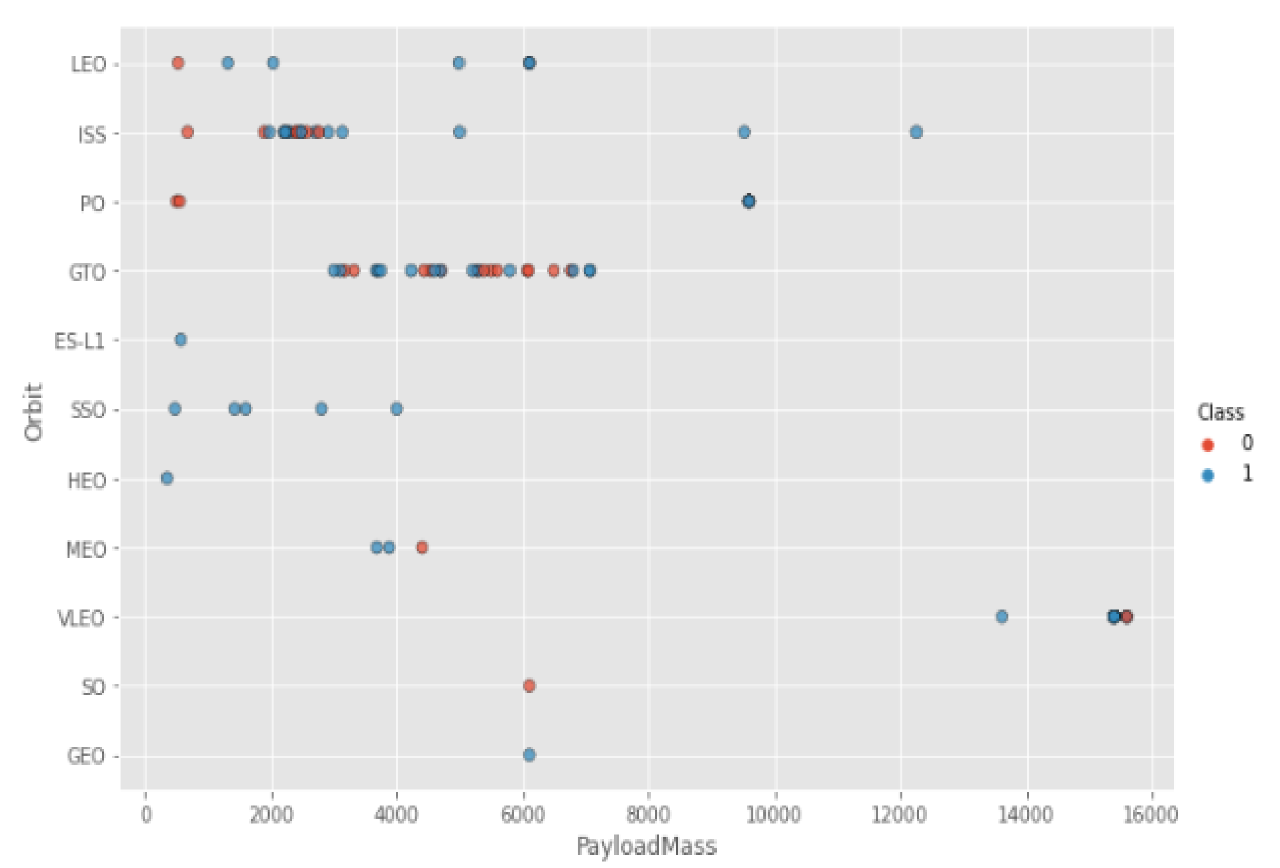
- Orbits SSO, HEO, GEO, and ES-L1 have 100% success rates.
- SO orbit did not have any successful launches with a 0% success rate.
- Other orbits such as VLEO, LEO, MEO, PO, ISS and GTO recorded success rates above 50%

Flight Number vs. Orbit Type



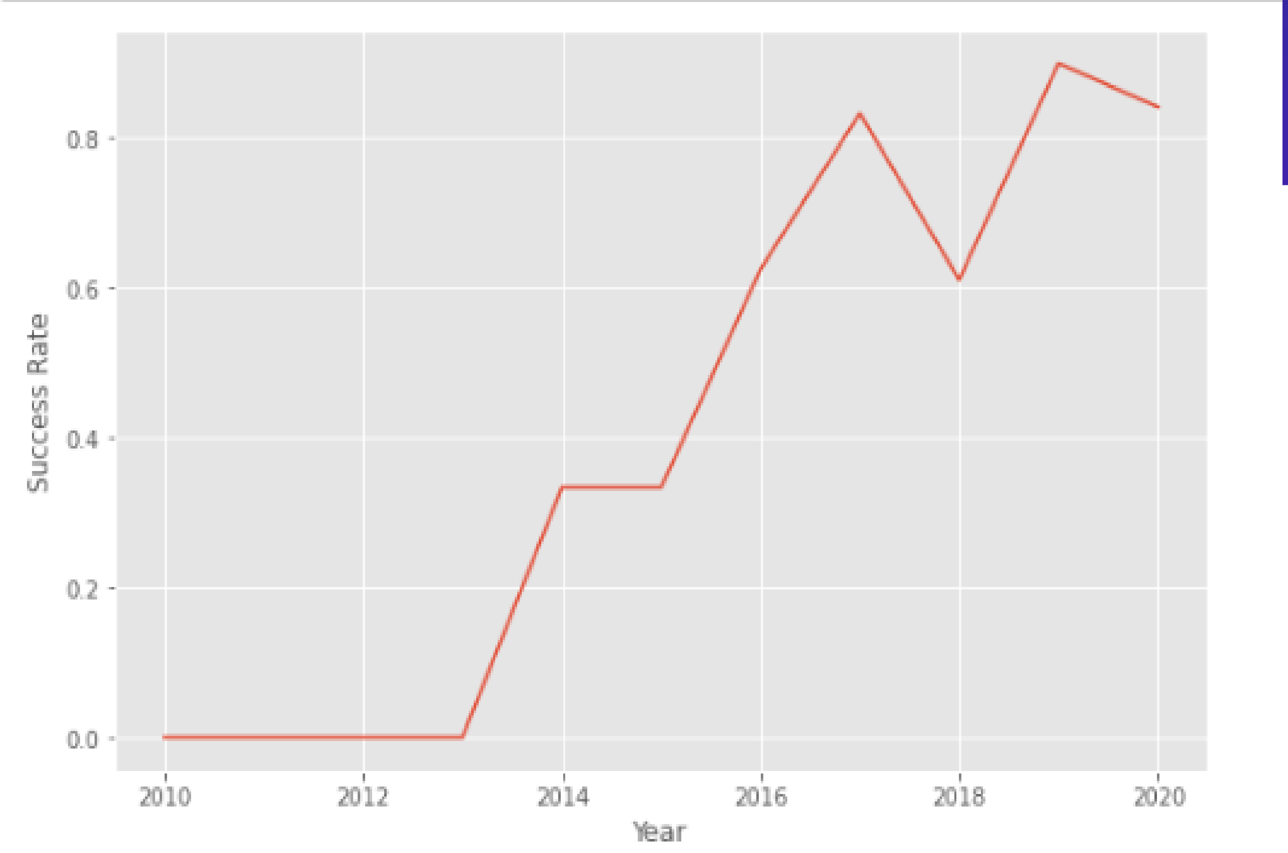
- There seems to be no relationship between flight number in the GTO orbit.
- In the LEO orbit, the success is positively correlated to the the number of flights.

Payload vs. Orbit Type



- We observe that Heavy payloads have a negative influence on GTO orbits and positive on Polar LEO and ISS orbits.
- There seems to be no direct correlation between orbit type and payload mass for GTO orbit as both successful and failed launches are equally present

Launch Success Yearly Trend



- We can observe that the success rate since 2013 kept increasing till 2020
- There was however recorded failure rates 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 occurrences 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 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

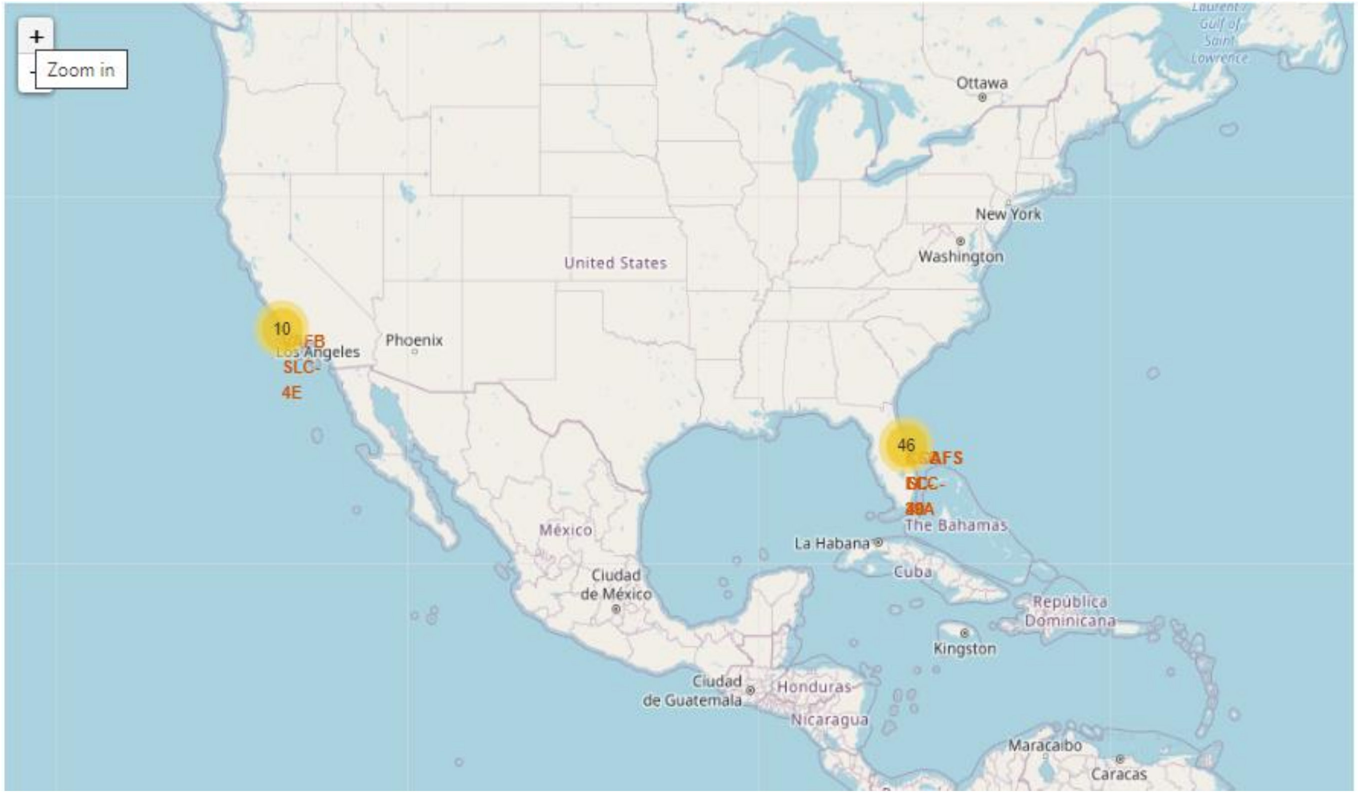
- COUNT() function was used to count the different *landing outcomes.* The WHERE and BETWEEN clauses filtered the results to only include results between 2010-06-04 and 2017-03-20. The GROUPBY clause ensure that the counts were grouped by their outcome. The ORDERBY and DESC clauses were used to sort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

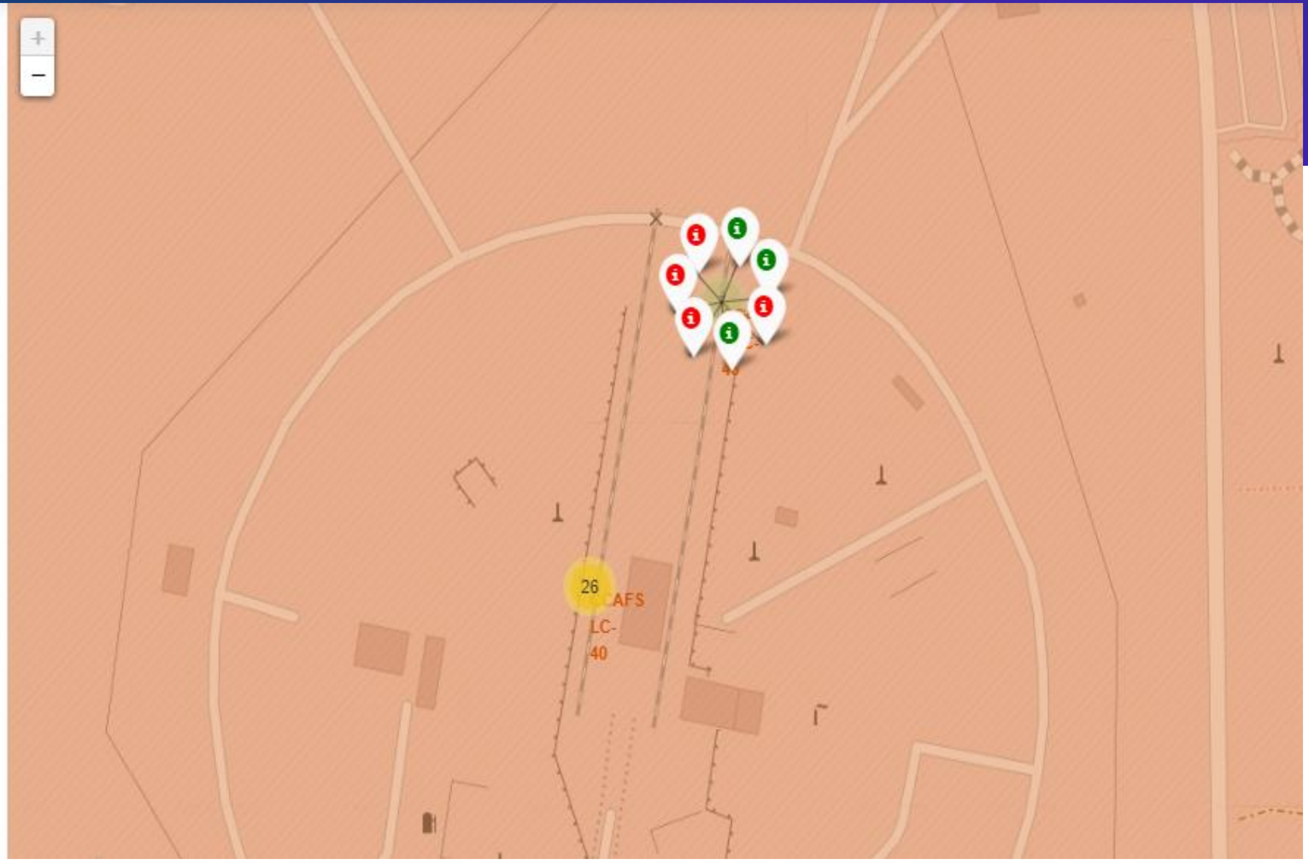
Launch Sites Proximities Analysis

SpaceX Launch Sites Locations



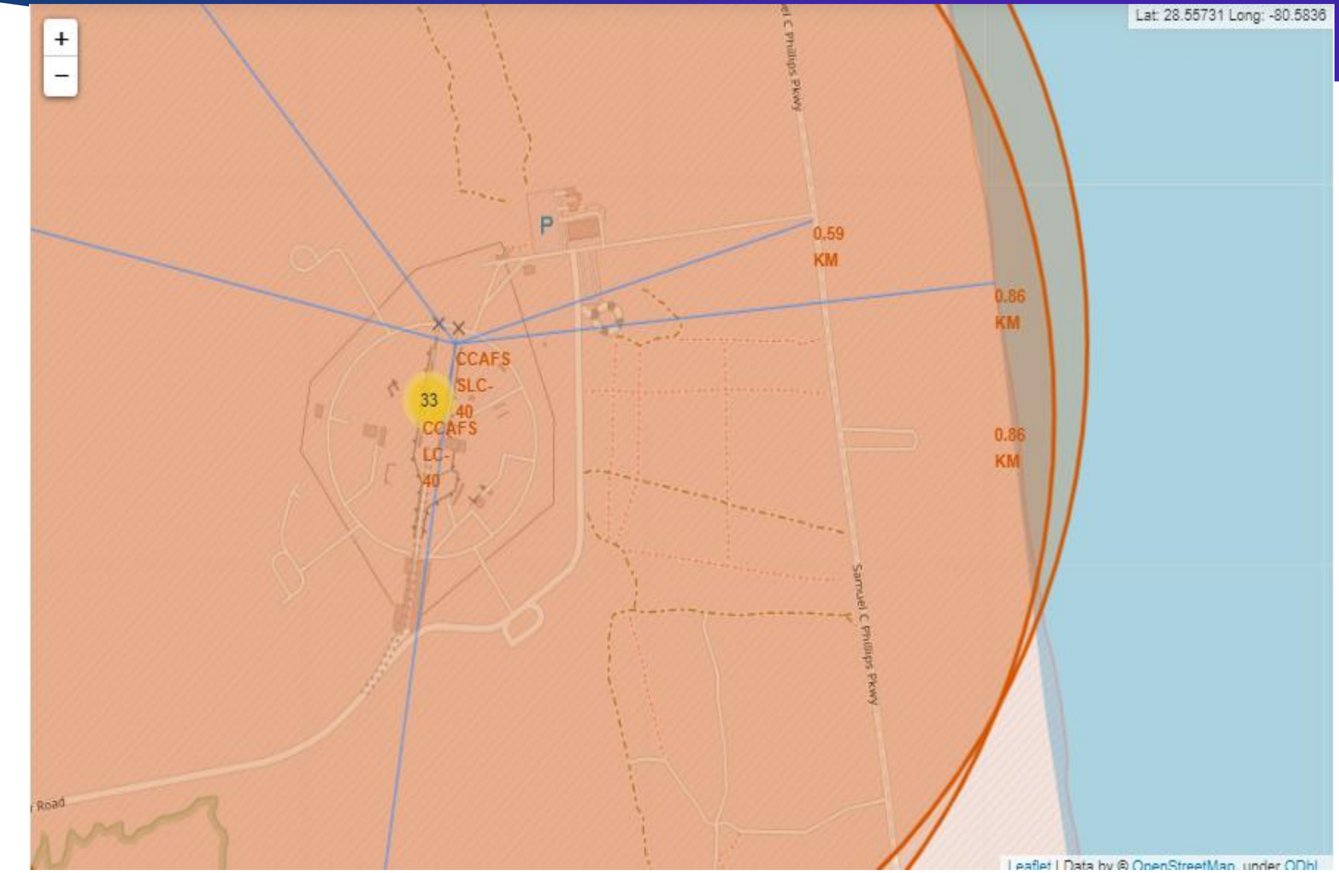
- The yellow markers are indicators of where the locations of all the SpaceX launch sites are situated in the US.
- The launch sites have been strategically placed near the coast

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 markers) or failed landing (red markers).

Launch Site Distance to Landmarks



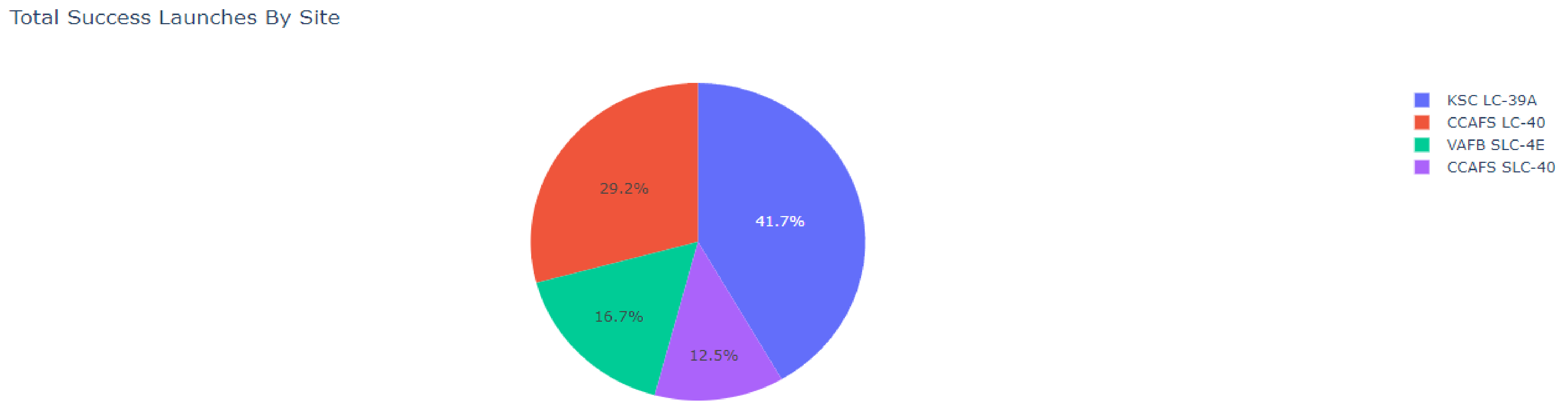
- The generated map shows that the selected launch site is close to a highway for transportation of personnel and equipment. The launch site is also close to the coastlines for launch failure testing.

Section 4

Build a Dashboard with Plotly Dash

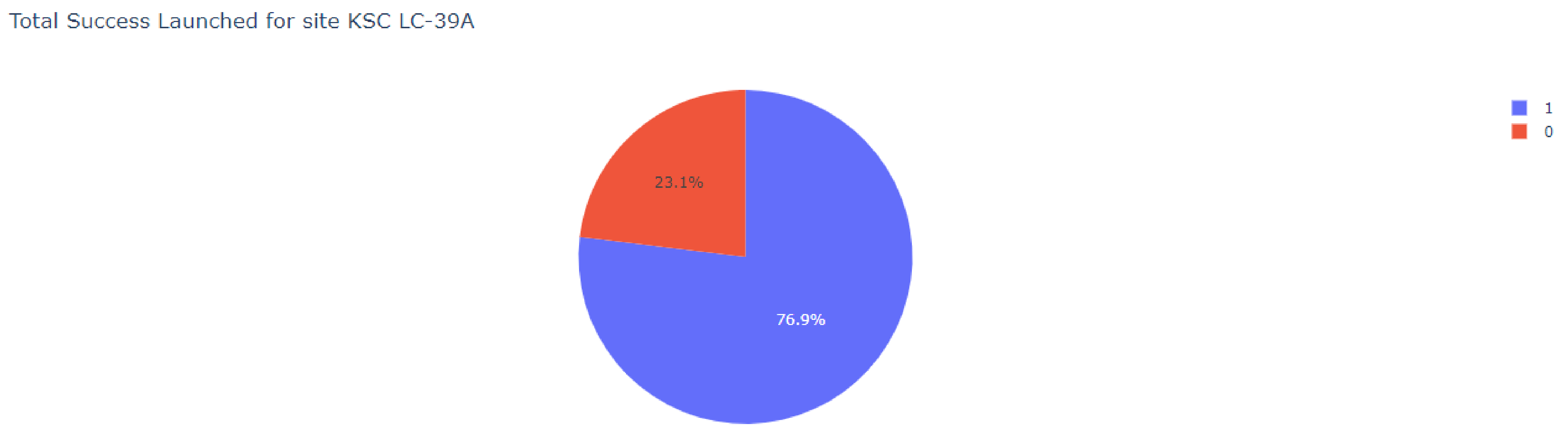
Total Successful Launches By Site

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



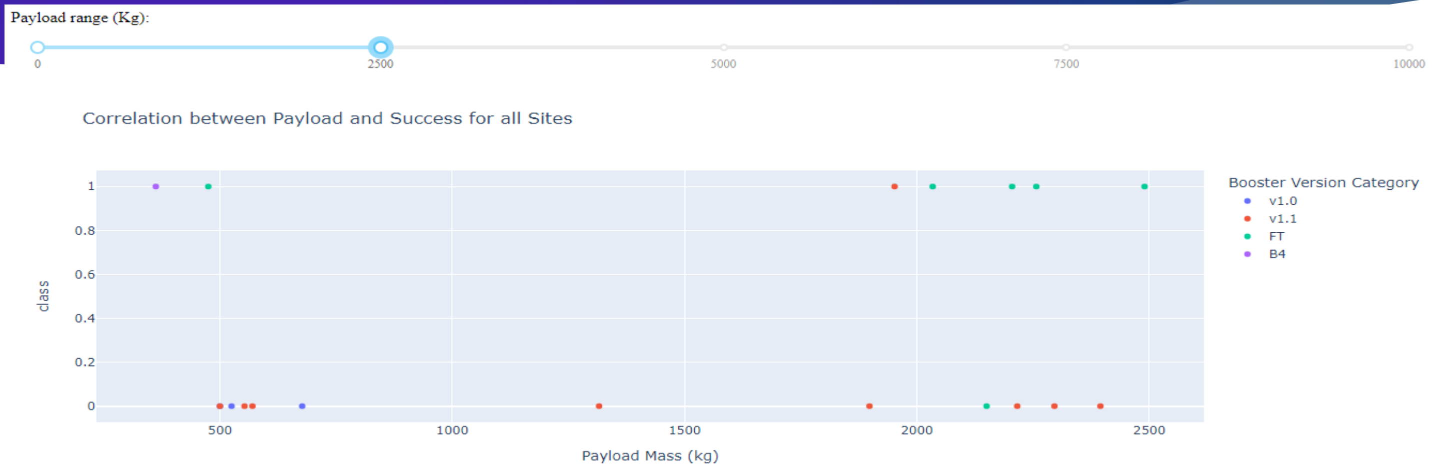
Launch Site With Highest Success Ratio

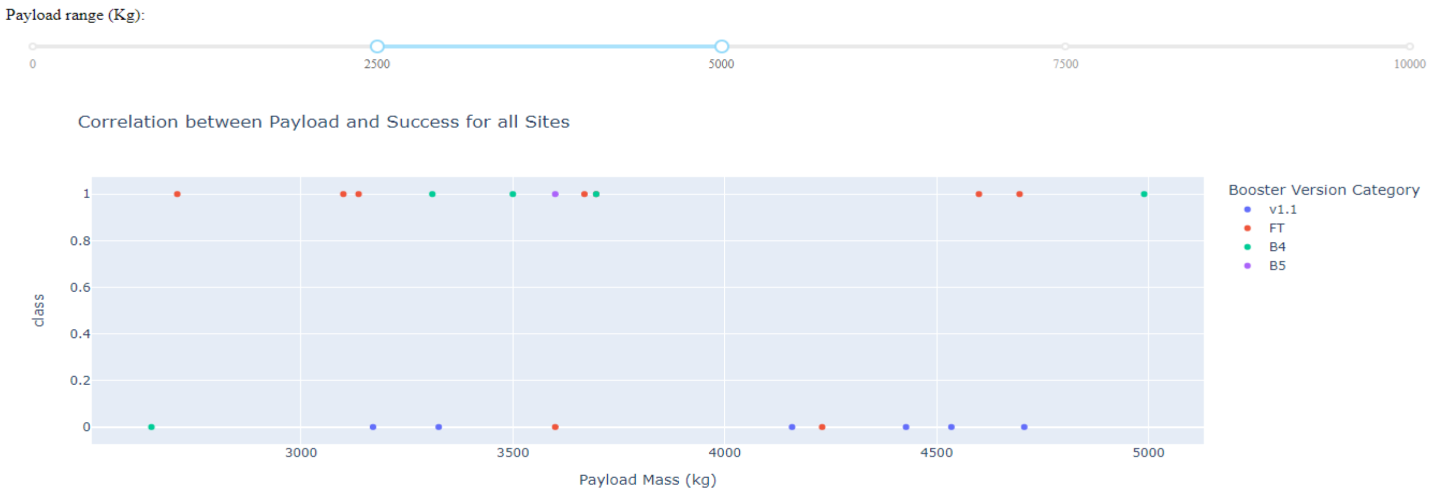
- The KSLC-39A has the highest success rate with 76.9%.



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*.



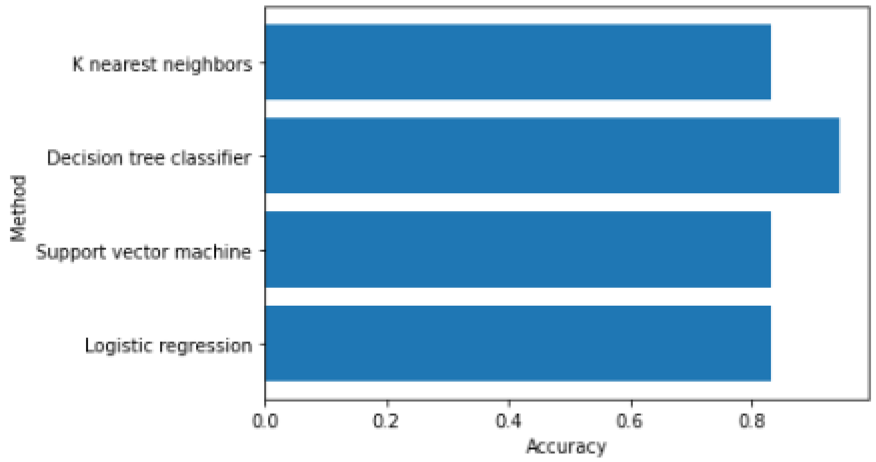


Section 5

Predictive Analysis (Classification)

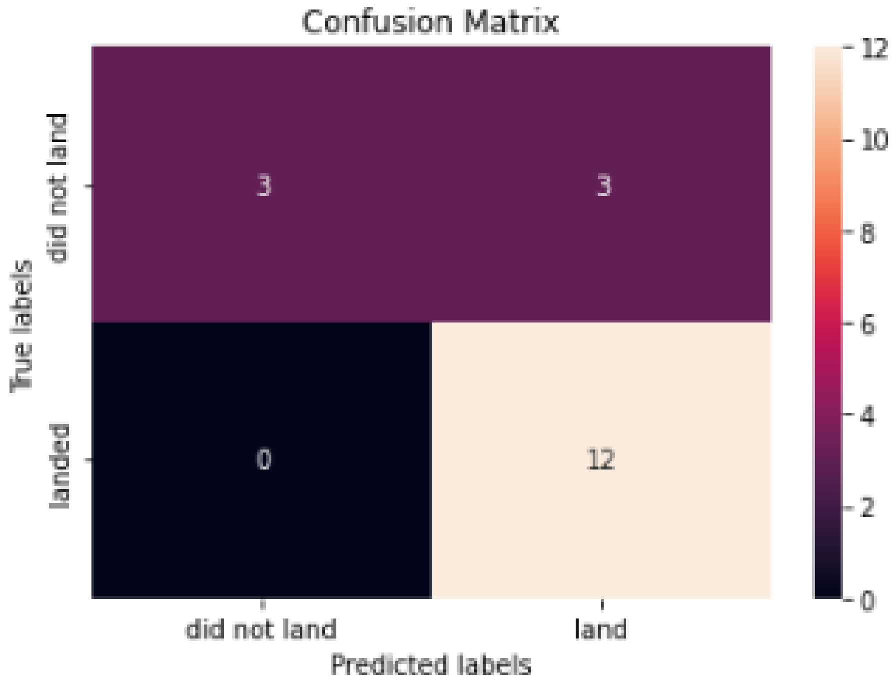
Classification Accuracy

- 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 larger the flight amount at a launch site, the greater the success rate at a launch site
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- 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 personnel and cargo, but also far away from cities for safety.
- The Decision tree classifier is the best machine learning algorithm for this task
- Launch success rate started to increase in 2013 till 2020

Appendix

GitHub Repository:

https://github.com/rmbanga/Space-X-Landing-Success-Prediction

Thank you!