IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Muaadh Shamhan
04/06/2023

Outline

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

Executive Summary

Summary of methodologies

- The data was collected from SpaceX API and SpaceX Wikipedia page using web scraping focusing on Falcon 9 rockets.
- Data wrangling and Exploratory Data Analysis (EDA) including SQL analysis, visualizations and interactive visual analysis.
- Machine learning to classify landing success or failure.

Summary of all results

- The EDA identified the importance of characteristics such as launch site and payload.
- The prediction score for the machine learning models is 83%.

Introduction

Space Y aims to compete with SpaceX, hence data gathering, analysis and machine learning predictions are needed to determine the characteristics of success.

Objectives

- To determine the costs of launching.
- To predict successful landings based on historical data.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from two sources, SpaceX API and SpaceX Wikipedia page using web scraping.
- Perform data wrangling
  - Identified the data types and missing values, as well as generating labels for the outcome.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Multiple ML models were used. GridSearch was used to tune the model, while score was used to evaluate the prediction.

Data Collection

The data was collected from two sources:

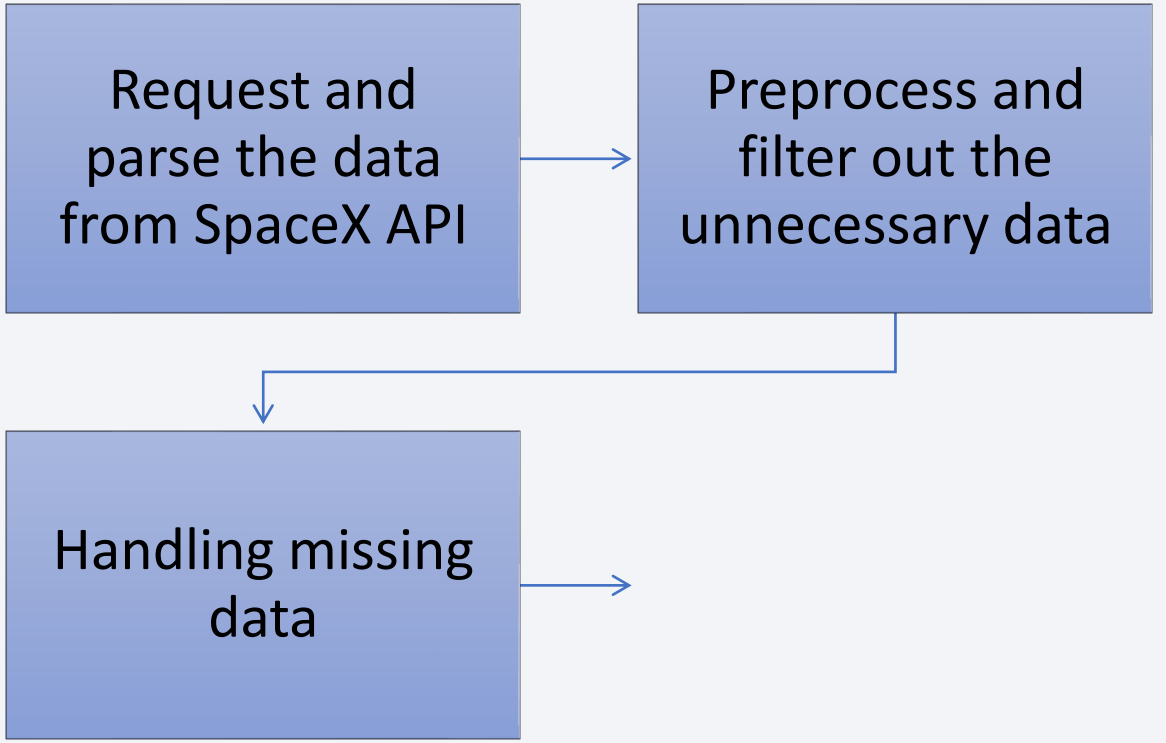
- Source 1: SpaceX API (https://api.spacexdata.com/v4/rockets/)

  Get request method was used to extract the data in .json format.

- Source 2: SpaceX Wikipedia page (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

  Web scraping was carried out using BeautifulSoup package.

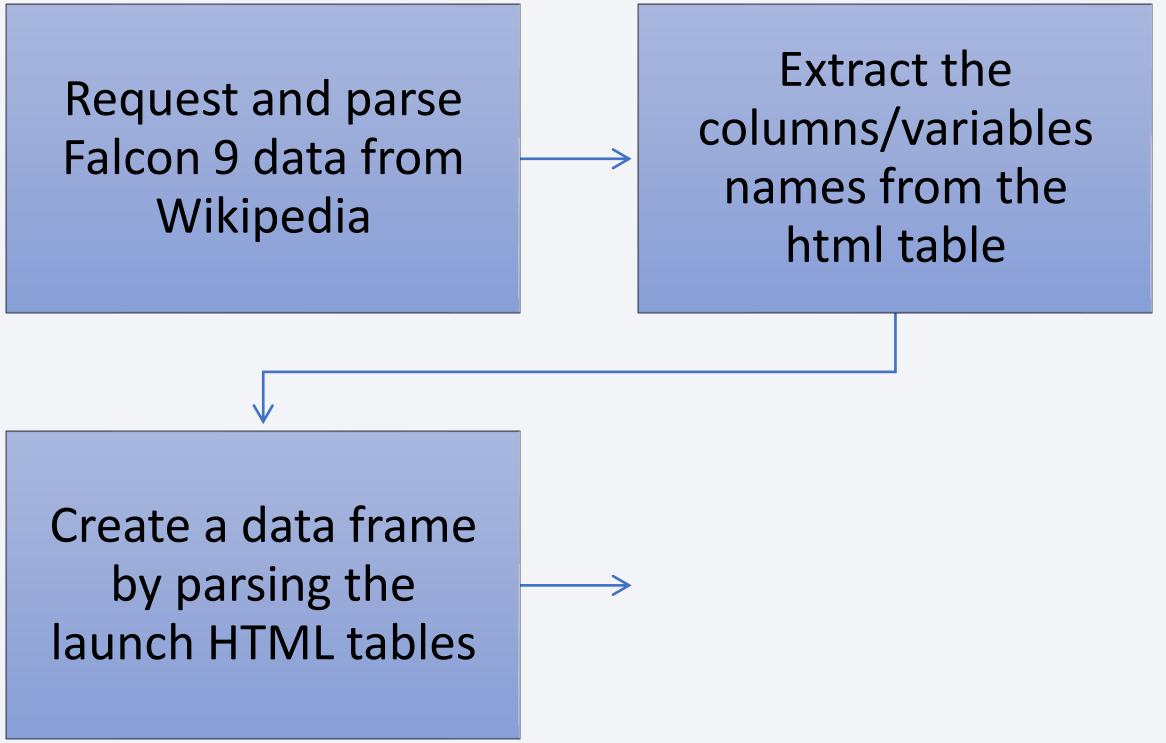
- After collection, the data was saved in a csv file.


Request and parse the data from SpaceX API
Preprocess and filter out the unnecessary data
Handling missing data

Data Collection - Scraping

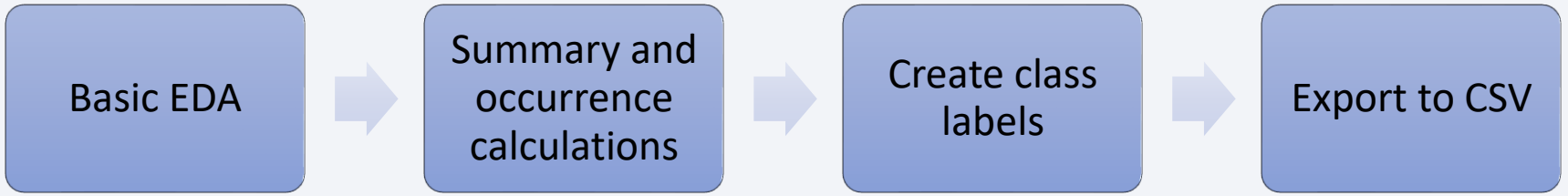
- The data was obtained from Wikipedia using web scraping. BeautifulSoup was used to parse the content using html.parser

- GitHub URL: https://github.com/MuaadhS/IBM_Applied_Data_Science_Capstone/blob/main/Week%201/jupyter-labs-webscraping.ipynb



Data Wrangling

- Initially, basic EDA was performed to understand the dataset, Where summaries of the launch sites, occurrence of orbits and mission outcomes were generated.
- Numerical labels were created to indicate success and failure.
- GitHub URL: https://github.com/MuaadhS/IBM_Applied_Data_Science_Capstone/blob/main/Week%201/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- EDA and data visualization were used to explore the relationships between payloads, launch sites, orbit types and flight numbers highlighting successful and failed landing attempts.

- GitHub URL: https://github.com/MuaadhS/IBM_Applied_Data_Science_Capstone/blob/main/Week%202/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

SQL queries were used to explore the following:

- Explore the unique launch sites.
- Calculate the total payload mass launched by NASA CRS and the average payload mass carried by F9 v1.1 boosters.
- Find the first date when the first successful landing was achieved.
- List the successful boosters for payload masses between 4000 and 6000, then find the maximum payloads for each booster version.
- List the successful and failed mission outcomes and their count as well as launching date.
- GitHub URL: https://github.com/MuaadhS/IBM_Applied_Data_Science_Capstone/blob/main/Week%202/jupyter_labs_eda_sql_coursera_sqllite.ipynb

Build an Interactive Map with Folium

Folium was used to visualize launch sites and success/failure on an interactive map.

- Markers were used to indicate launch sites.
- Circles were used to highlight areas around launch sites.
- Marker clusters to group success and failure in each site.
- Line to indicate distance between coordinates.

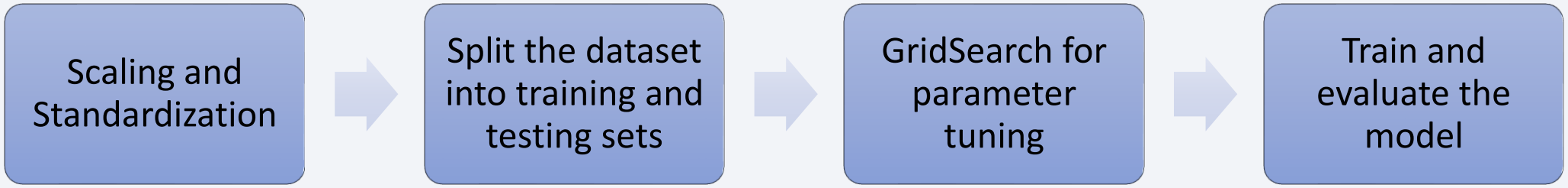
- GitHub URL: https://github.com/MuaadhS/IBM_Applied_Data_Science_Capstone/blob/main/Week%203/_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The dashboard consists of pie charts to visualize the percentage of successful attempts per launching site.
- A slider was used to set the payload mass.
- A scatter plot was used to visualize mission outcomes for each site.
- GitHub URL: https://github.com/MuaadhS/IBM_Applied_Data_Science_Capstone/blob/main/Week%203/_jupyter_launch_site_location.ipynb

Predictive Analysis (Classification)

- Multiple classification models were tested that are Logistic Regression, SVM, Decision Trees and KNN.
- GitHub URL: https://github.com/MuaadhS/IBM_Applied_Data_Science_Capstone/blob/main/Week%204/SpaceX_Machine_Learning_Prediction_Part_5.ipynb

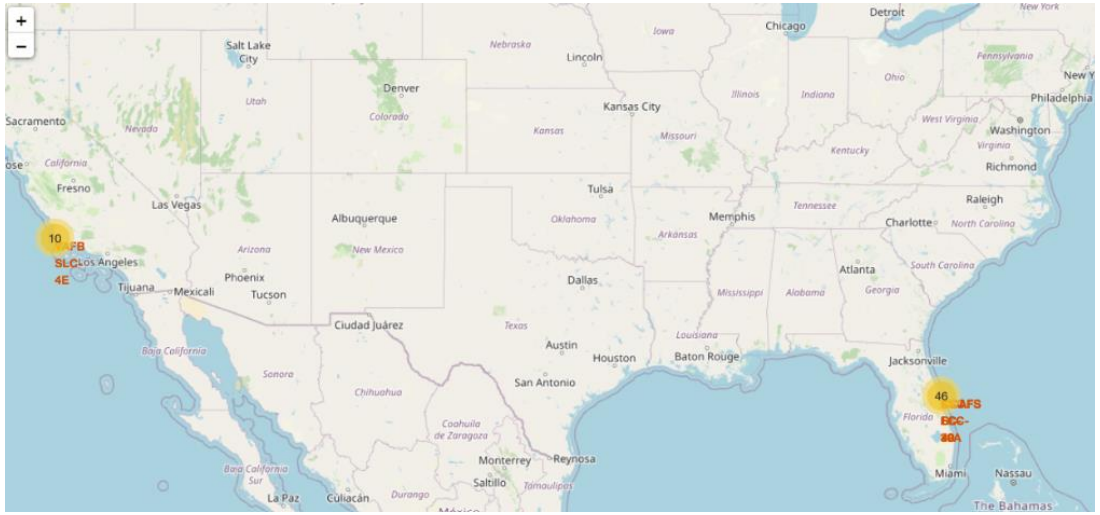


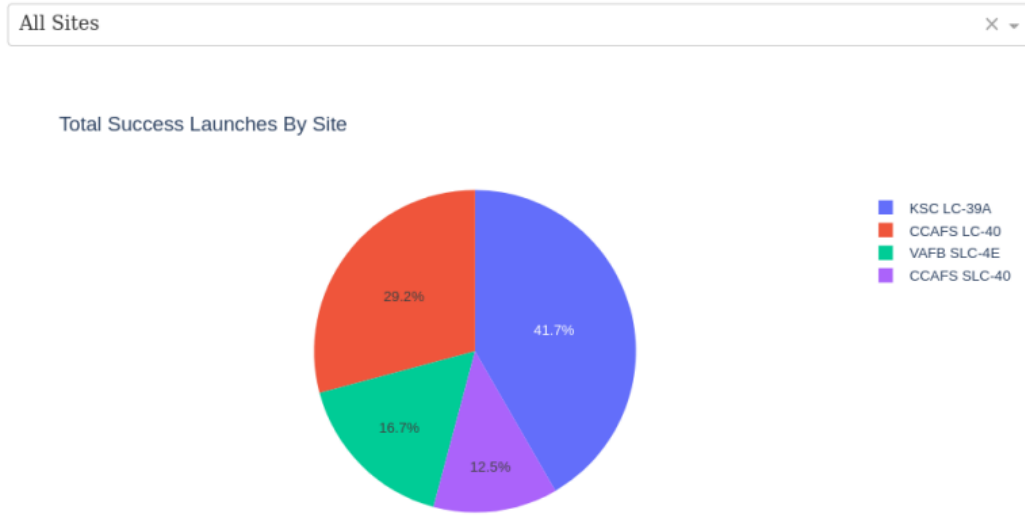
Results

- Exploratory data analysis results
  - Successful landing depends on multiple factors such as launch site, payload and orbit.
  - Success rate increased by approximately 80% between 2010 and 2020.
  - SpaceX have used 4 launch sites, CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.
  - The total payload launched by NASA boosters is 48213 kg.
  - F9 v1.1 boosters have an average payload mass of 2534.67 kg.
  - The first successful landing was recorded on 01-08-2018.
  - The boosters F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2 have higher success rate with payloads between 4000 and 6000 kg.
  - 99 successful landings are recorded in the dataset.
  - The maximum launched payload is 15600 kg that was recorded with multiple boosters.
  - In 2015, two booster versions failed at landing in drone ships that are F9 v1.1 B1012 and F9 v1.1 B1015.
  - Between the date 04-06-2010 and 20-03-2017, all landing attempts were successful.

Results

- Interactive analytics
  - 46 launches were from east coast, while 10 launches were from west coast. All launches are performed near sea.
  - 42% of the successful landing attempts were at KSC LC-39A launch site.





Results

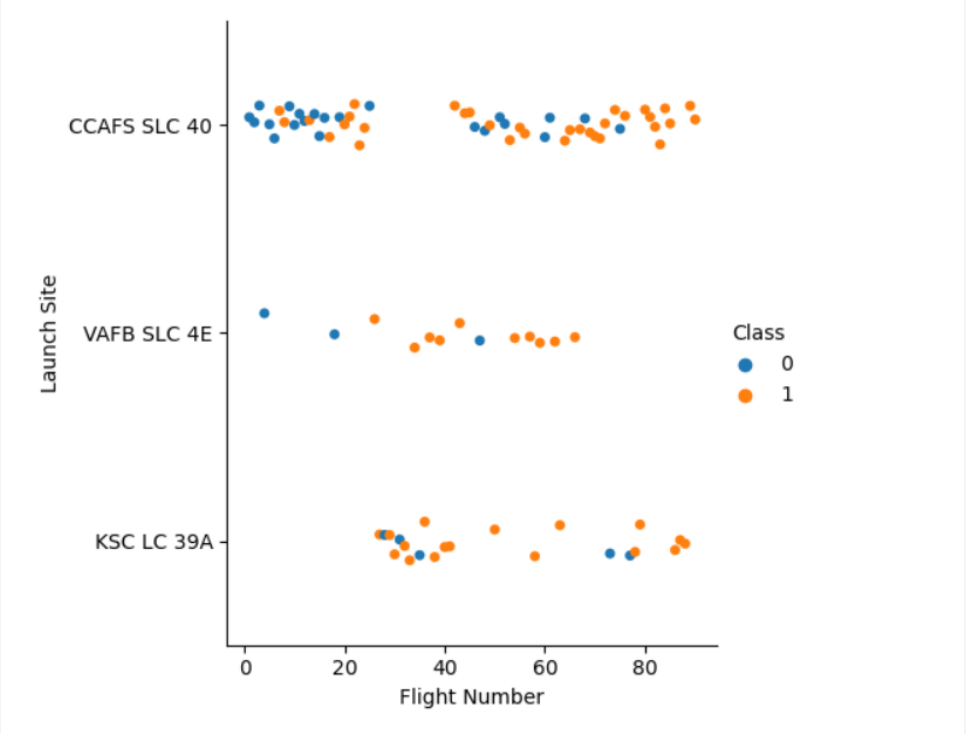
- Predictive analysis results
  - All models achieved 83% accuracy on the testing set.
  - The best model is the Decision tree classifier. It achieved a ~90% accuracy which is the highest accuracy achieved on the training set.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

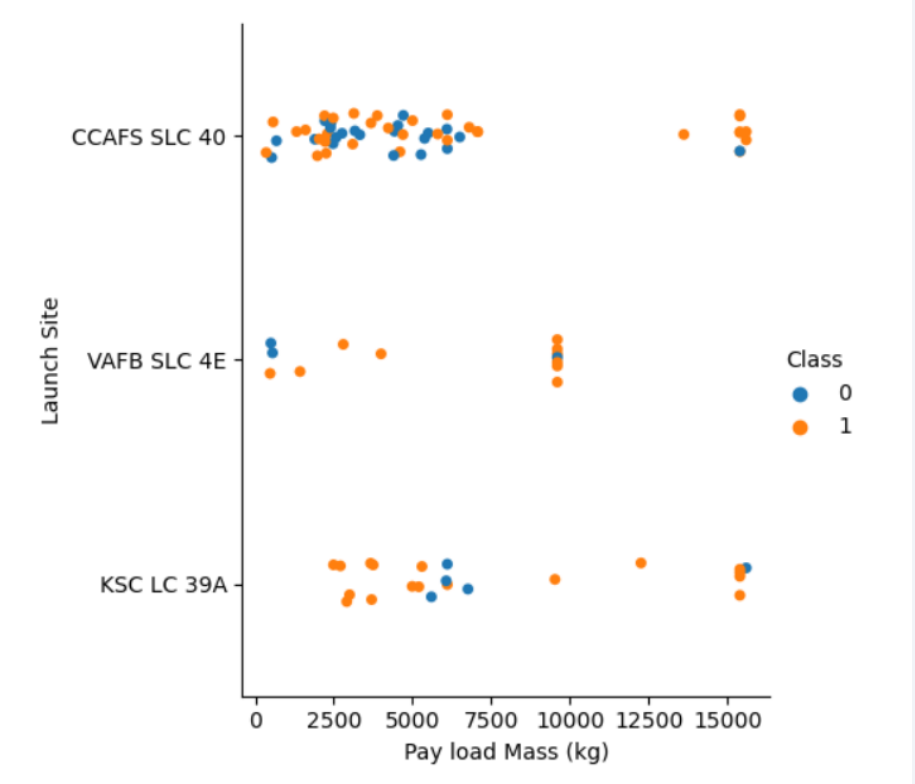
- For all launching sites, the latest landing attempts have been successful.
- CCAFS SLC 40 has the most successful landings in recent attempts.



Payload vs. Launch Site

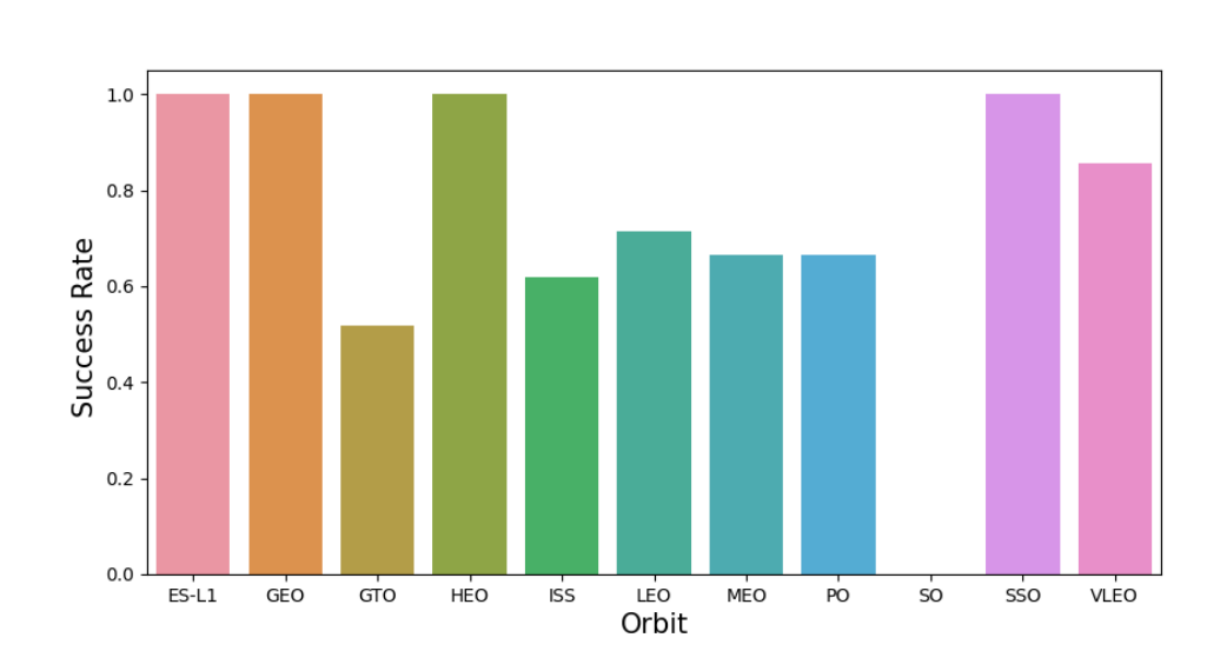
- The best success rate is recorded at payload masses between 7500kg and 10000kg.

- Payloads more than 10000kg are launched from CCAFS SLC 40 and KSC LC 39A launch sites.



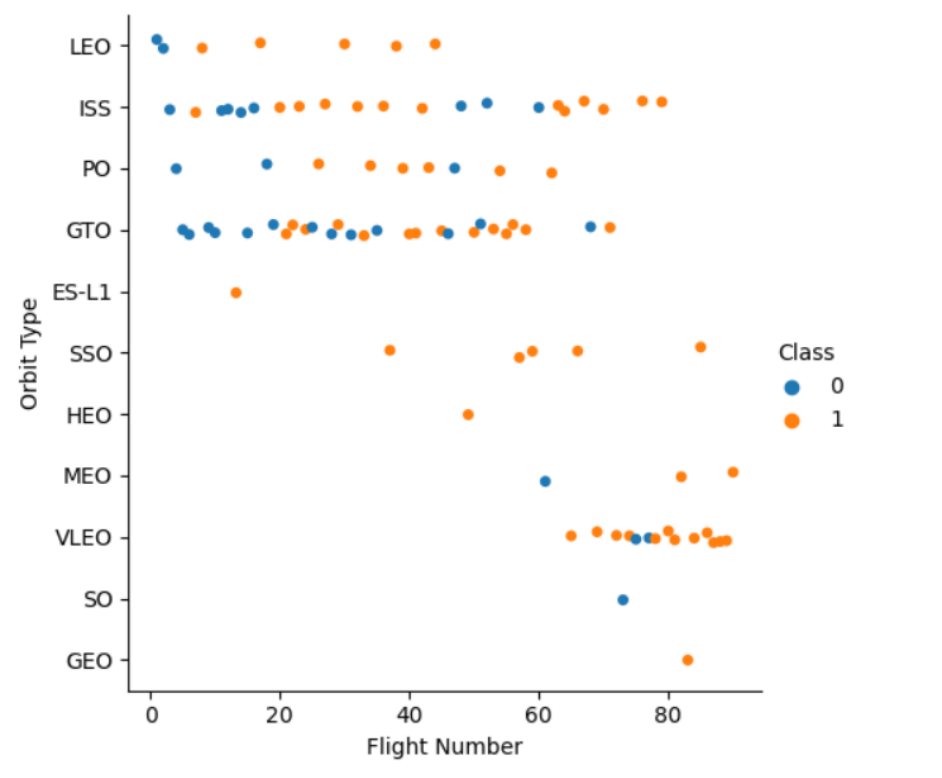
Success Rate vs. Orbit Type

- The best success rates are recorded at the orbits ES-L1, GEO, HEO and SSO.



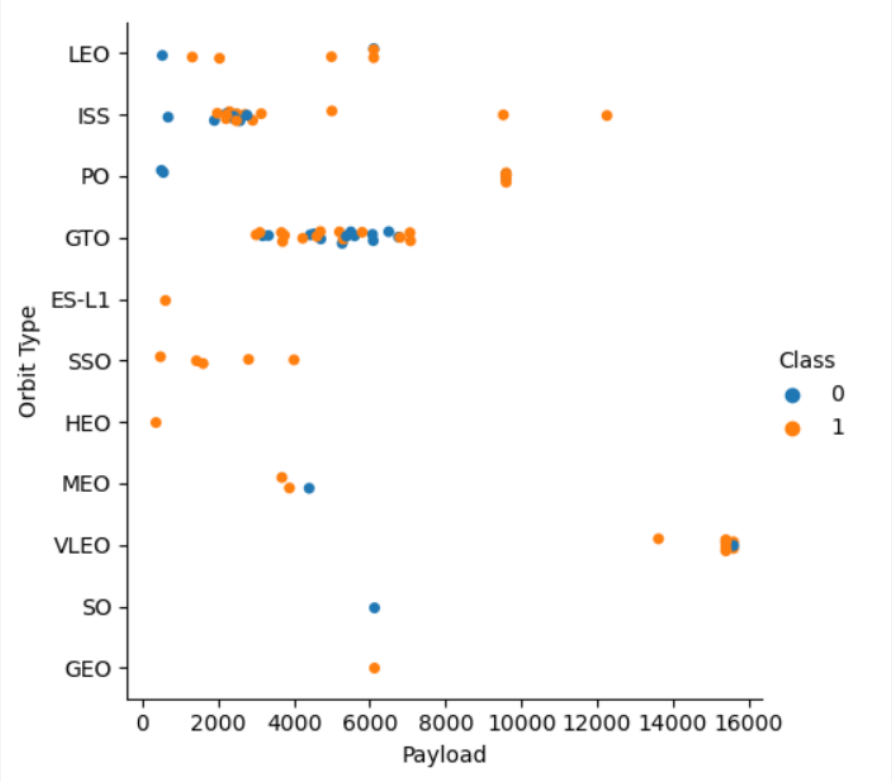
Flight Number vs. Orbit Type

- Recent flights have been successful for all orbit types.
- VLEO is the most targeted orbit in recent flights with a high success rate.



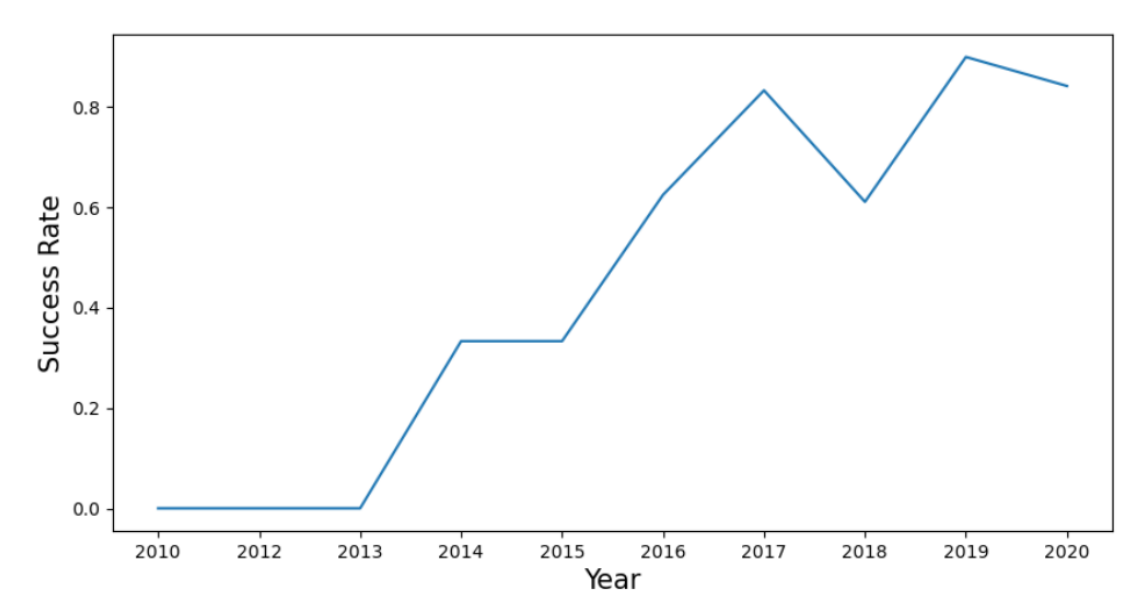
Payload vs. Orbit Type

- For payloads less than 5000kg. All attempts are successful for SSO orbit type.
- LEO has the best success rate between 5000kg and 8000kg payloads.
- For more than 8000kg, ISS, PO and VLEO have good success rates.



Launch Success Yearly Trend

- Success rate increased by approximately 80% between 2010 and 2020.



All Launch Site Names

- Four launch sites are found in the dataset as shown below.

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

SQL query: SELECT DISTINCT(launch_site) FROM SPACEXTBL;

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
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	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No att
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No att
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No att

SQL query: SELECT * FROM SPACEXTBL WHERE launch_site LIKE('CCA%') LIMIT 5;

Total Payload Mass

- The total payload carried by boosters from NASA is as shown below:

SUM(payload_mass__kg_)
48213.0

SQL query: SELECT SUM(payload_mass__kg_) FROM SPACEXTBL WHERE customer LIKE('%NASA (CRS)%');

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1:

AVG(payload_mass__kg_)
2534.6666666666665

SQL query: SELECT AVG(payload_mass__kg_) FROM SPACEXTBL WHERE booster_version LIKE('F9 v1.1%');

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad:

min(date)
01/08/2018

SQL query: SELECT min(date) FROM SPACEXTBL WHERE landing_outcome = 'Success (ground pad)';

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

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

SQL query: SELECT booster_version FROM SPACEXTBL WHERE payload_mass__kg_ BETWEEN 4000 AND 6000 AND landing_outcome = 'Success (drone ship)';

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes, where 99 attempts have successful outcomes.

Mission_Outcome	COUNT(*)
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

SQL query: SELECT mission_outcome, COUNT(*) FROM SPACEXTBL GROUP BY mission_outcome;

Boosters Carried Maximum Payload

- The names of the booster 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
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

SQL query: SELECT booster_version, payload_mass__kg_ FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

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

SQL query: SELECT substr(date, 4, 2) AS month, substr(date, 7, 4) AS year, landing_outcome, booster_version, launch_site FROM SPACEXTBL WHERE landing_outcome = 'Failure (drone ship)' AND substr(date, 7, 4) = '2015';

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

- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:

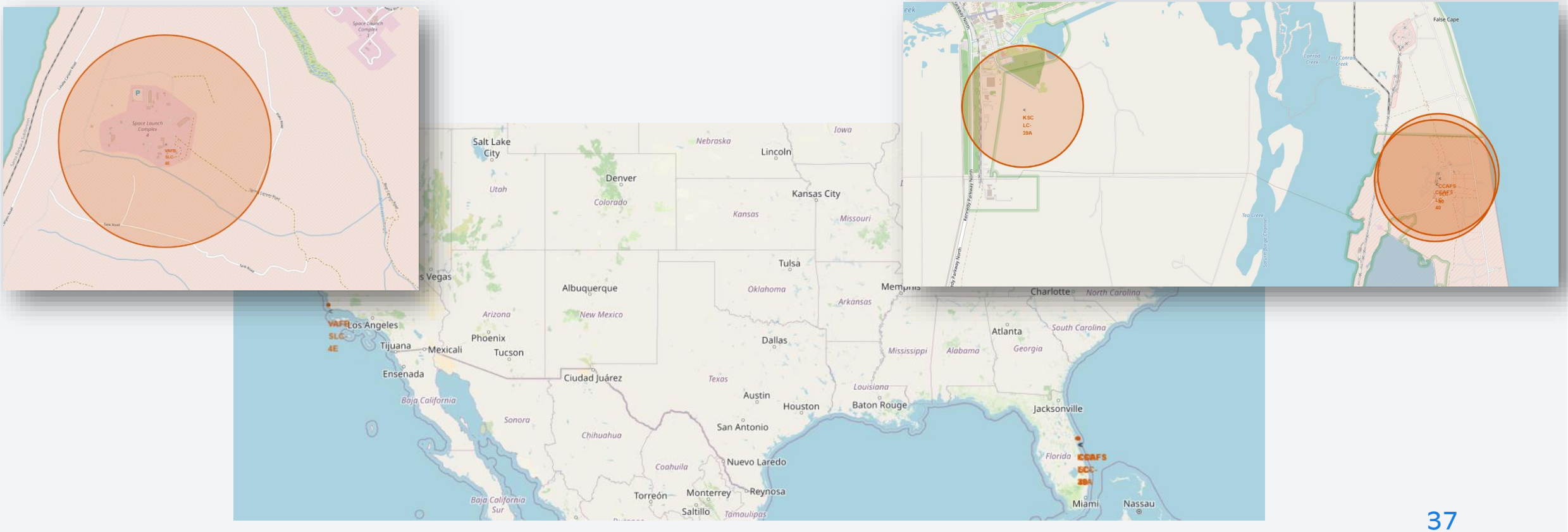
Landing_Outcome	success_count
Success	20
Success (ground pad)	0
Success (drone ship)	0
No attempt	0
No attempt	0
Failure (parachute)	0
Failure (drone ship)	0
Failure	0
Controlled (ocean)	0

SQL query: SELECT landing_outcome, COUNT(CASE WHEN landing_outcome = 'Success' THEN 1 END) AS success_count FROM SPACEXTBL WHERE date BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY landing_outcome ORDER BY success_count DESC;


Section 3
Launch Sites
Proximities Analysis

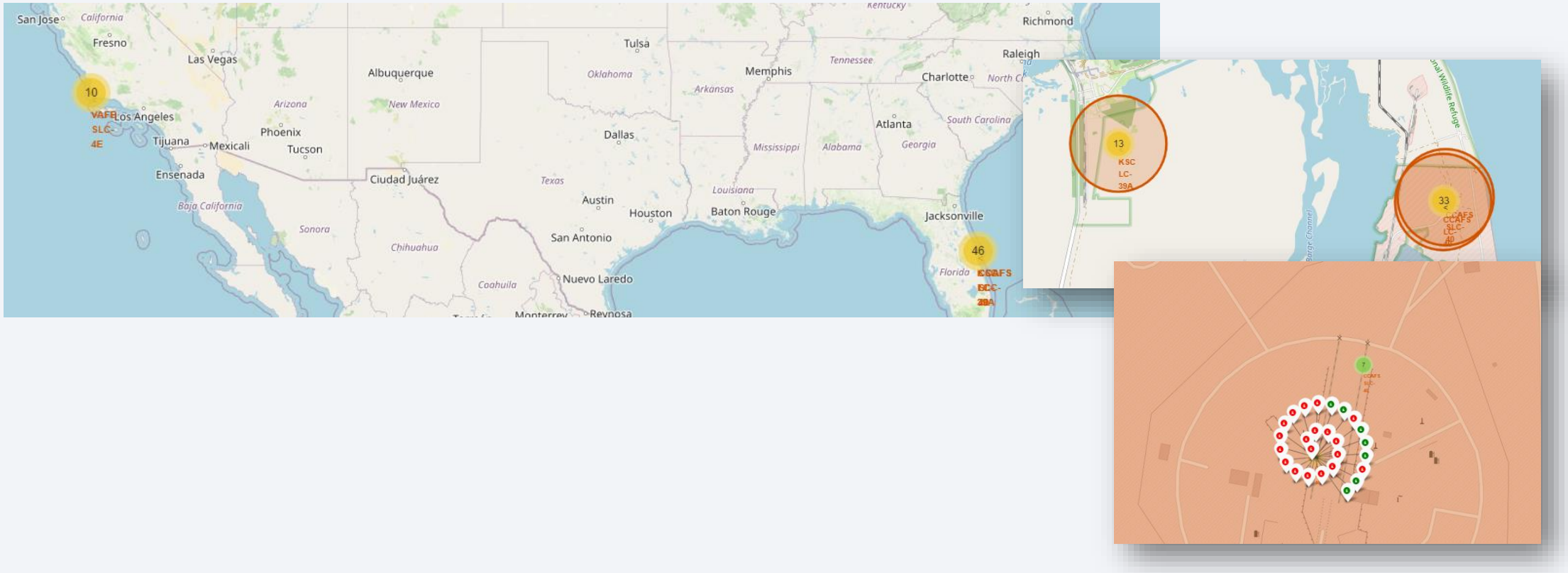
Launch Sites

- Three launch sites are in east coast while only one launch site is in west coast.



Missions Outcome by Launch Sites

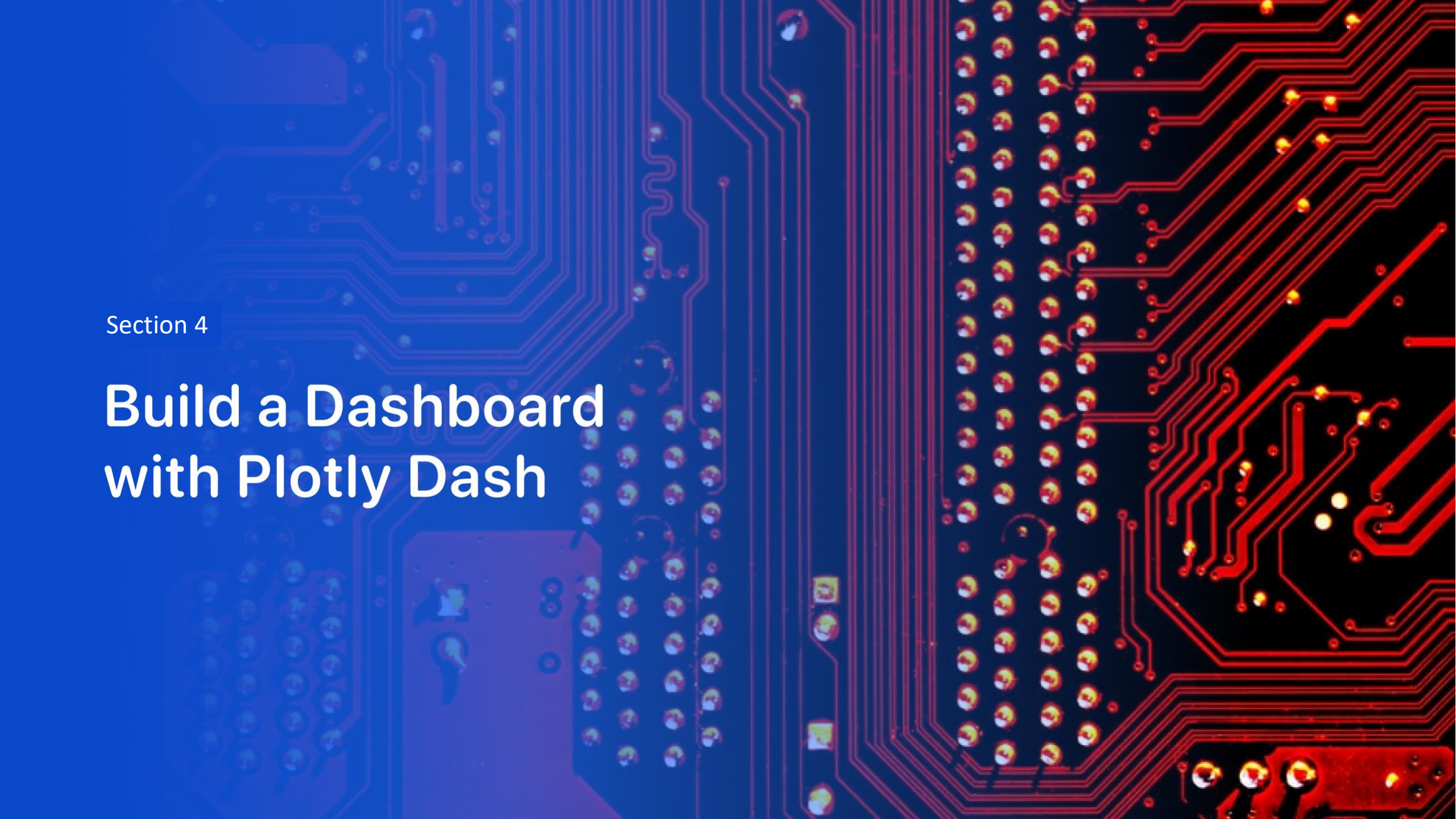
- The count of landing attempts is displayed on each site.
- Green and red markers are used to indicate success and failure respectively.



Launch Sites Logistics

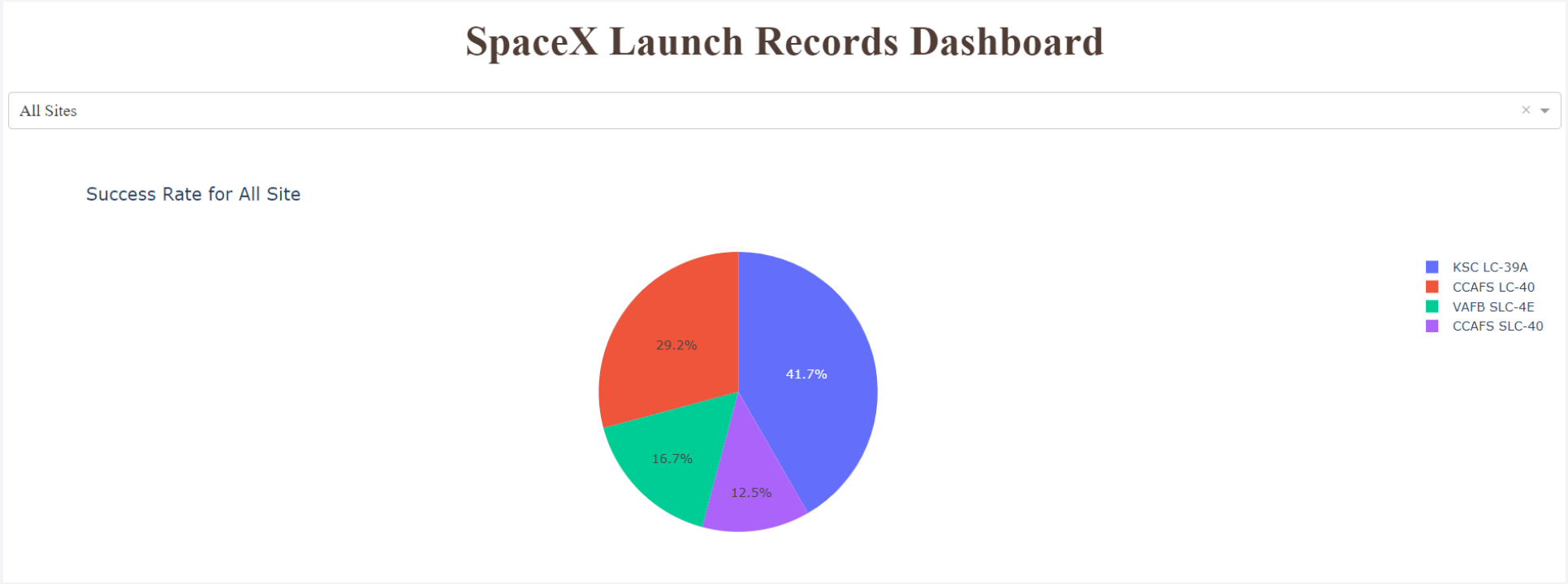
- The distance between the launch site and the coast is ~0.88km.

- The distance between the launch site and the nearest city (Orlando) is ~77km.


Section 4
Build a Dashboard with Plotly Dash

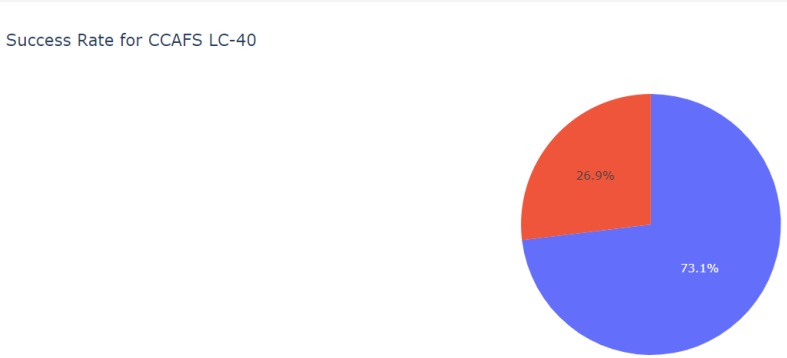
Success Rate for All Sites

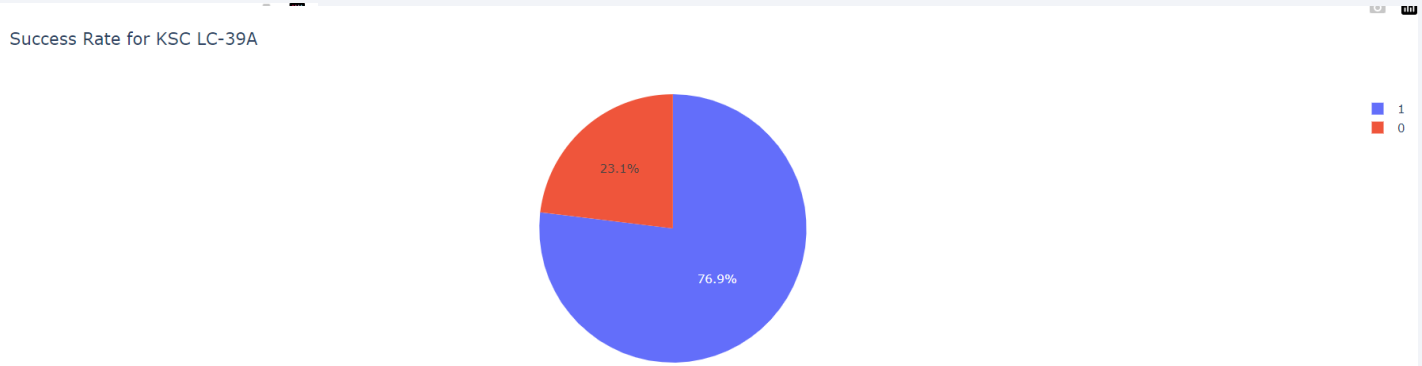
- The highest success rate is recorded at KSC LC-39A launch site.

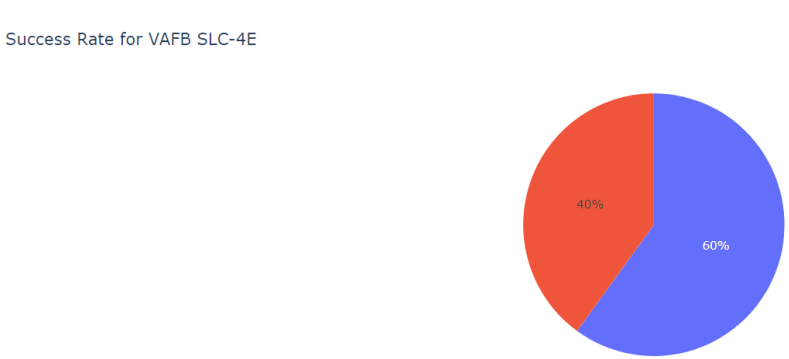


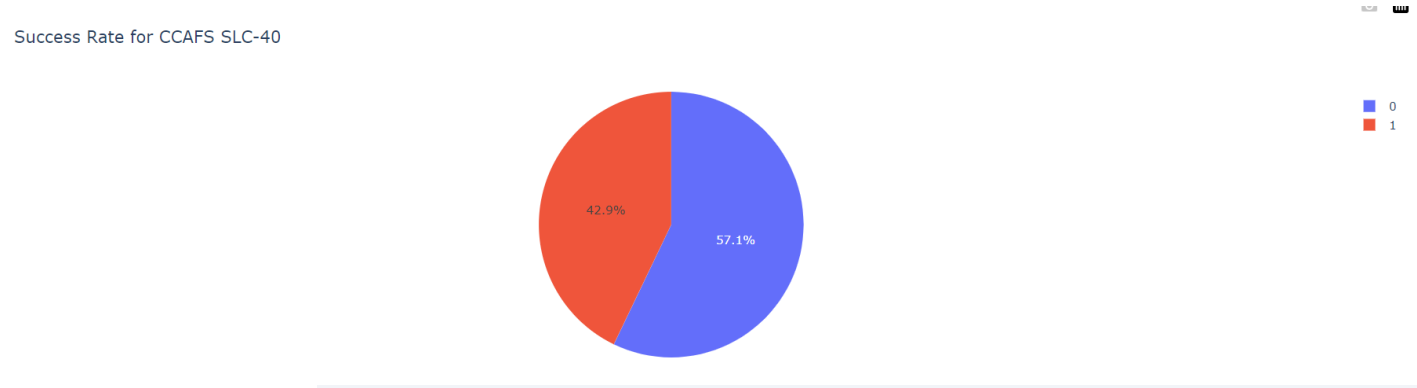
Success Rate for Individual Sites

- Class 0 indicates Failure while class 1 indicate success.





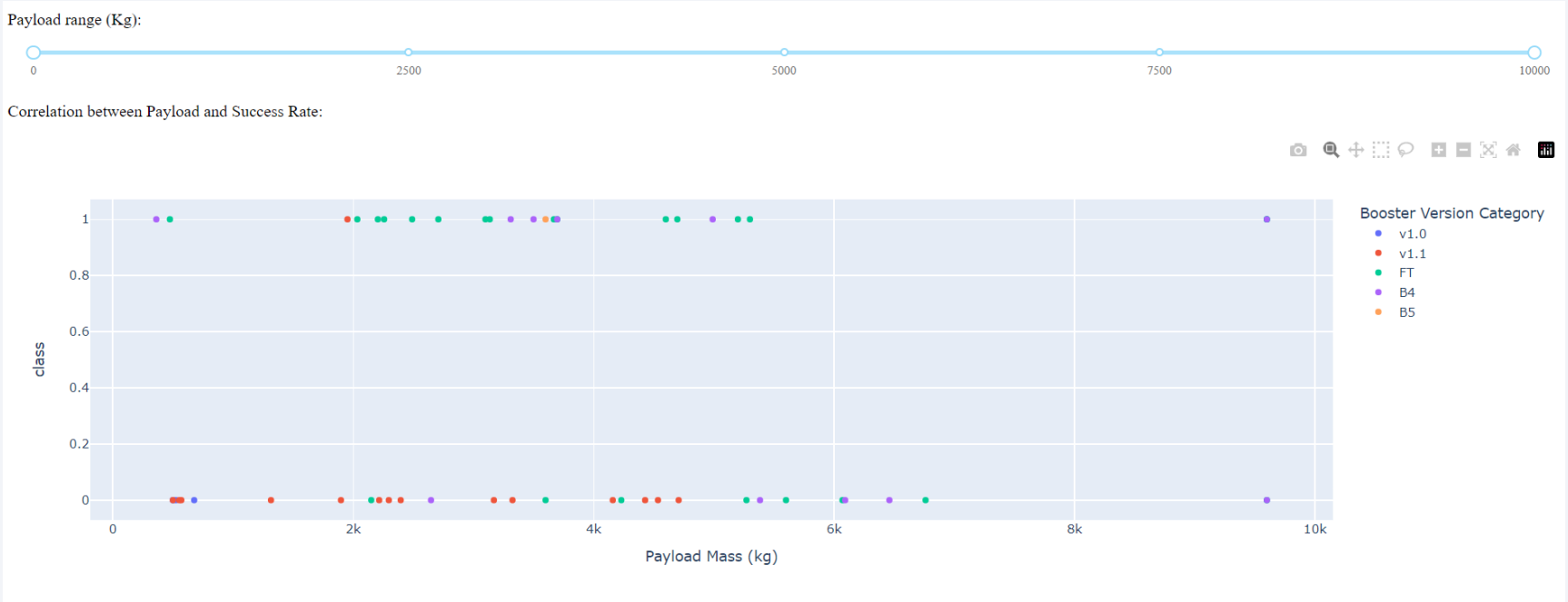




The Highest success rate is at KSC LC-39A at ~77%, while the lowest rate is at CCAFS SLC-40 at ~57%

Payload vs. Mission Outcomes

- The payload is controlled by a slider, the values range between 0kg and 10000kg.
- The colors indicate the booster version, while class 1 indicates success and class 0 indicates failure.
- FT boosters with payloads less than 6000kg have the most successful records.

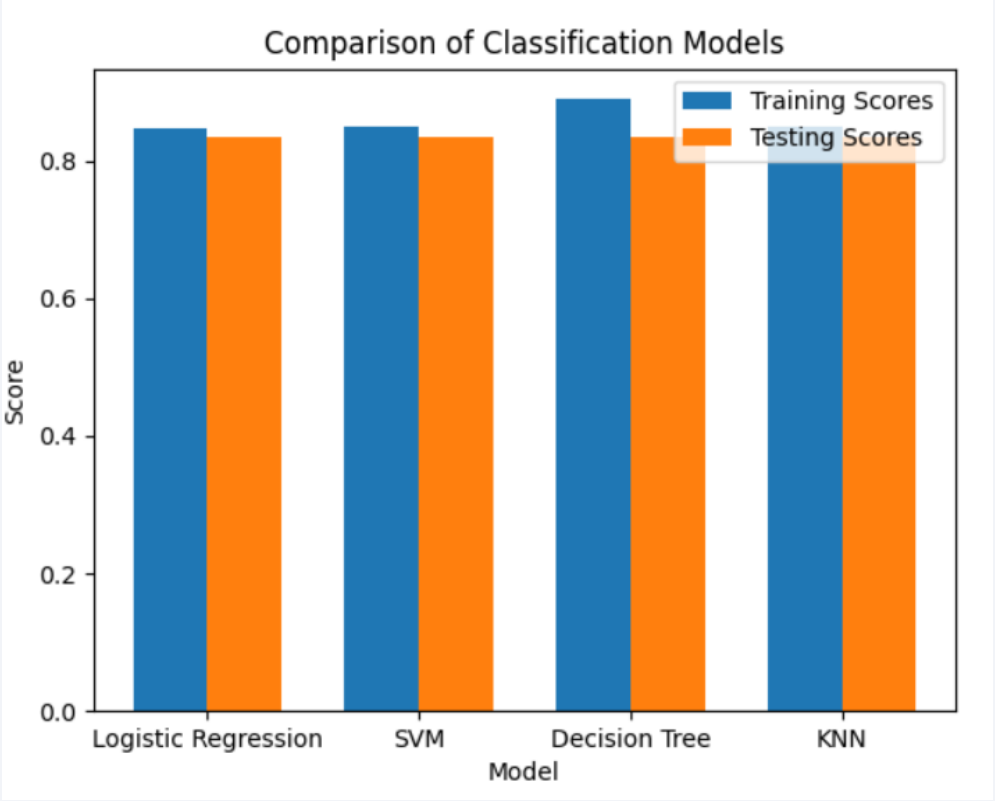


Section 5

Predictive Analysis (Classification)

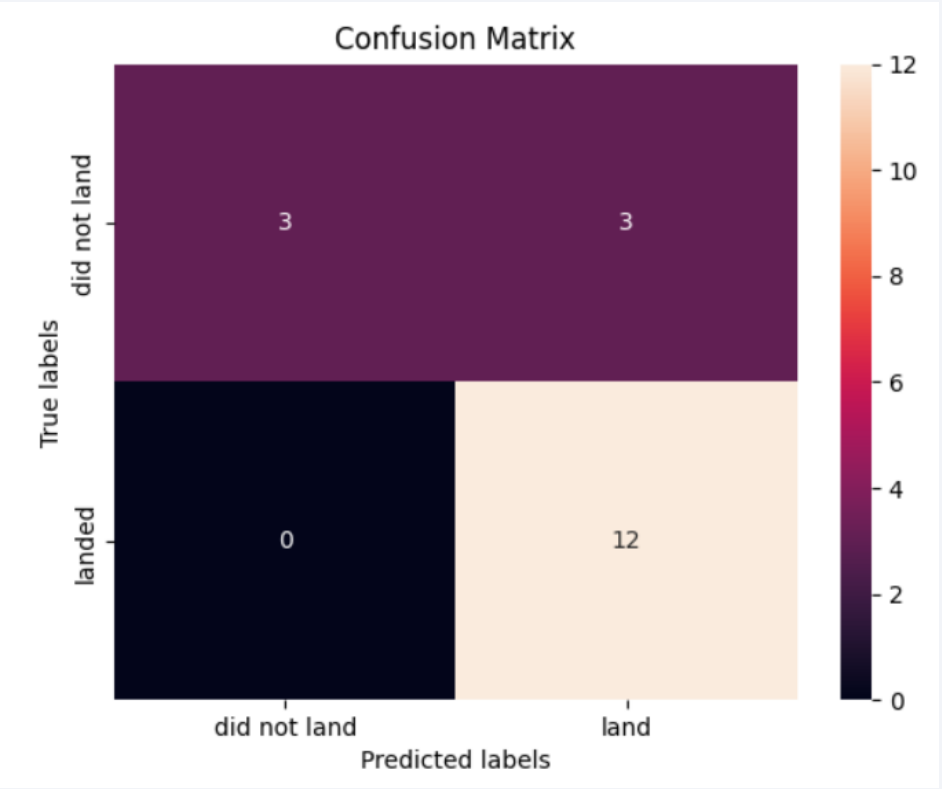
Classification Accuracy

- The highest classification accuracy is achieved with Decision Tree classifier at ~89% accuracy for the training set and ~83% accuracy for the testing set.



Confusion Matrix

- The confusion matrix of the Decision Tree classifier reflects a good score based on the achieved true positives and true negatives.



Conclusions

- In this project, Falcon 9 data was collected and processed.
- KSC LC-39A launch site has the best success rate.
- Payload, orbit type and booster version are correlated with successful outcomes.
- Rockets with payloads less than 6000kg are more likely to succeed.
- Successful landing has an incremental trend over time.
- Decision tree classifier can help predict success and failure based on the conditions and characteristics of historical data.

Appendix

- All codes are available on my GitHub repository: https://github.com/MuaadhS/IBM_Applied_Data_Science_Capstone

- For folium and dash files, the output is not displayed on GitHub and may require running the codes to view the plots.

Thank you!