IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ivan Stepanov
01/06/2024

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection and Wrangling
  - Exploratory Data Analysis (EDA)
  - Interactive Dashboard
  - Predictive Analysis
- Summary of all results:
  - Data Insights
  - Model Performance
  - Dashboard Utility

Introduction

Project Background and Context:

- SpaceX has revolutionized the space industry by significantly lowering launch costs through the reuse of the Falcon 9 first stage. While SpaceX offers launches for about $62 million, other providers charge around $165 million, mainly due to the inability to reuse rocket stages. This project aims to leverage data science to predict the landing success of Falcon 9's first stage, providing insights for companies looking to compete with SpaceX.

Problems to Address:

- Can we predict the success of Falcon 9's first stage landing?
- What factors most influence landing success?
- How can these predictions inform competitive strategies?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from SpaceX API and web scraped from wikipedia
- Perform data wrangling
  - Transforming the raw data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Implemented four machine learning classification algorithms: Logistic Regression, Decision Trees, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)

Data Collection

- The data for this project was collected from two sources: the SpaceX API and web scraping from Wikipedia.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

SpaceX provides a publicly accessible interface, known as an API (Application Programming Interface), that allows users to retrieve data for their own use.

- Send GET Request to SpaceX API
- Decoding the response content and turning it into a dataframe
- Constructing data into a dictionary
- Exporting the data to CSV file
- Filtering dataframe, and replacing missing values
- Creating dataframe from the dictionary

GitHub

Data Collection - Scraping

Data from Wikipedia site,extracted from html

GitHub 

- Requesting Falcon 9 launch data from Wikipedia
- Creating a Beautiful Soup object from the HTML response
- Parsing HTML tables to collect the data
- Exporting the data to CSV file
- Creating a dataframe from the dictionary
- Constructing data we have obtained into a dictionary

Data Wrangling

- Perform EDA and determine training labels
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Exporting the data to CSV file
- Create a landing outcome label from Outcome column
- Calculate the number and occurence of mission outcome of the orbits

GitHub

EDA with Data Visualization

The charts were selected to visually explore different aspects of the SpaceX launch data and derive insights.

1. Flight Number vs. Payload Mass: To observe trends over time, distinguishing between successful and unsuccessful flights.

2. Flight Number vs. Launch Site: To compare launch distributions across sites, highlighting success outcomes.

3. Payload vs. Launch Site: To analyze payload distribution and its impact on launch outcomes across sites.

4. Success Rate of Each Orbit: To compare success rates across different orbit types.

5. Flight Number vs. Orbit: To examine orbit success rates over time.

6. Payload Mass vs. Orbit: To explore the relationship between payload mass and orbit success.

7. Mean Success Rate by Year: To visualize trends in launch success rates over time.

GitHub

EDA with SQL

- There are 4 distinct landing sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
- The total payload mass carried by boosters launched by NASA (CRS): 45596 kg
- An average payload mass carried by booster version F9 v1.1: 2534.6667 kg
- The date when the first successful landing outcome in ground pad was achieved: 2015-12-22
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2
- There are 100 successful mission outcomes and 1 failure
- There are 12 booster versions which have carried the maximum payload mass 15600 kg.
- There are 2 records which have failure (drone ship) landing outcome in 2015
- In the period from 2010-06-04 to 2017-03-20, there were 9 successful landings on the ground and 5 unsuccessful landings on the drone ship

GitHub

Build an Interactive Map with Folium

In the data visualization process using Folium, various map objects such as markers, circles, and lines were created and added to the map to provide a clear and informative visualization of SpaceX launch data. Here is a summary of each object:

- Markers for launch sites:

- Added markers for each SpaceX launch site to clearly indicate the geographical locations of all launch sites on the map, allowing for easy identification and comparison.

- Marked the success/failed launches for each site:

- Added markers for each launch at the respective launch sites, color-coded by success (e.g., green for success, red for failure) to provide a visual representation of the outcome of launches at each site, making it easy to see patterns of success and failure geographically.

- Calculated the distances between a launch site to its proximities:

- Added circles around each launch site to indicate proximity ranges to visually represent areas around each launch site, helping to understand the geographical spread and potential impact areas.

- Added lines between the launch sites and the nearest highway, railway, city have been added to indicate distances to provide a visual assessment of the distances from the launch sites to significant objects,

GitHub

Build a Dashboard with Plotly Dash

- Added dropdown menu for "All sites" and individual launch sites to provide flexibility in the data being viewed, allowing users to focus on overall trends or drill down into specific launch sites for detailed analysis.
- ● Added a pie chart for the selected launch site (there is an option for all sites) to give a clear visual representation of the success ratio, helping users quickly grasp the performance of SpaceX launches either in aggregate or at specific sites.
- ● Payload Mass Range Slider: Added a range slider for payload mass from 0 to 10,000 kg to enable detailed analysis of how payload mass affects launch success, allowing users to investigate specific payload ranges and their corresponding outcomes.
- ● Added a scatter plot for Payload Mass vs Success that updates dynamically based on the selected launch site and payload mass range, with different colors for different booster versions. This helps in visually exploring the relationship between payload mass and launch success, and understanding how different factors influence the success of launches.

GitHub

Predictive Analysis (Classification)

Creating a Numpy array from the "Class" column

Standardizing the data with StandardScaler

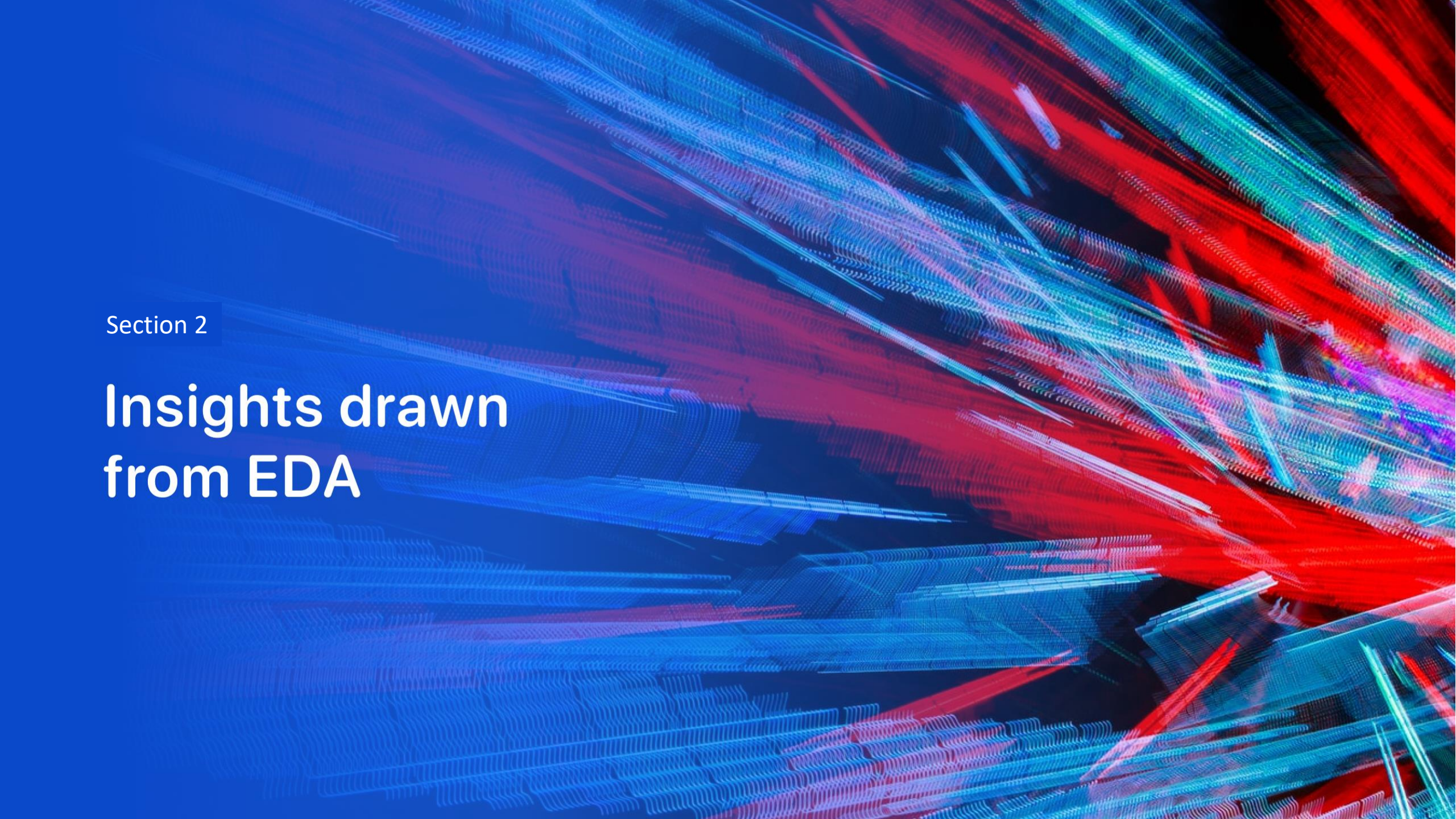
Splitting data into training and testing sets

Finding the model performs best with a highest performance across all evaluation metrics

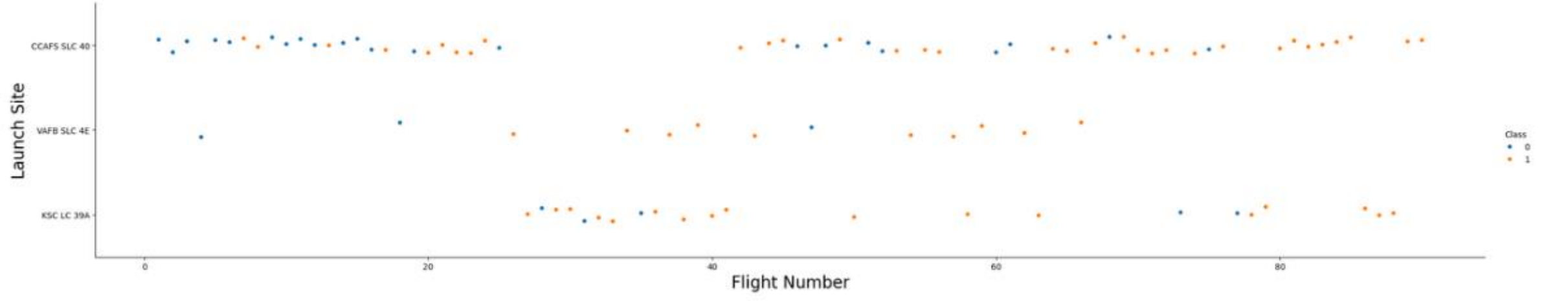
Developing 4 machine learning classification models: Logistic Regression, SVM, Decision Tree, KNN

Applying GridSearchCV for finding the best parameters of each model

GitHub


Section 2
Insights drawn from EDA

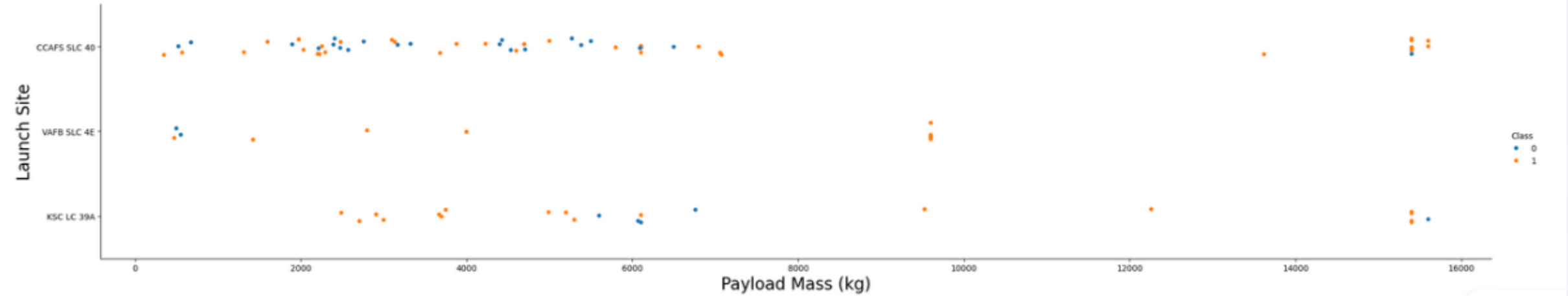
Flight Number vs. Launch Site



Explanation:

- The first launches were unsuccessful, while the latest ones are more successful
- The CCAFS SLC 40 launch site has the most launches
- However, VAFB SLC 4E and KSC LC39A have higher success rates

Payload vs. Launch Site



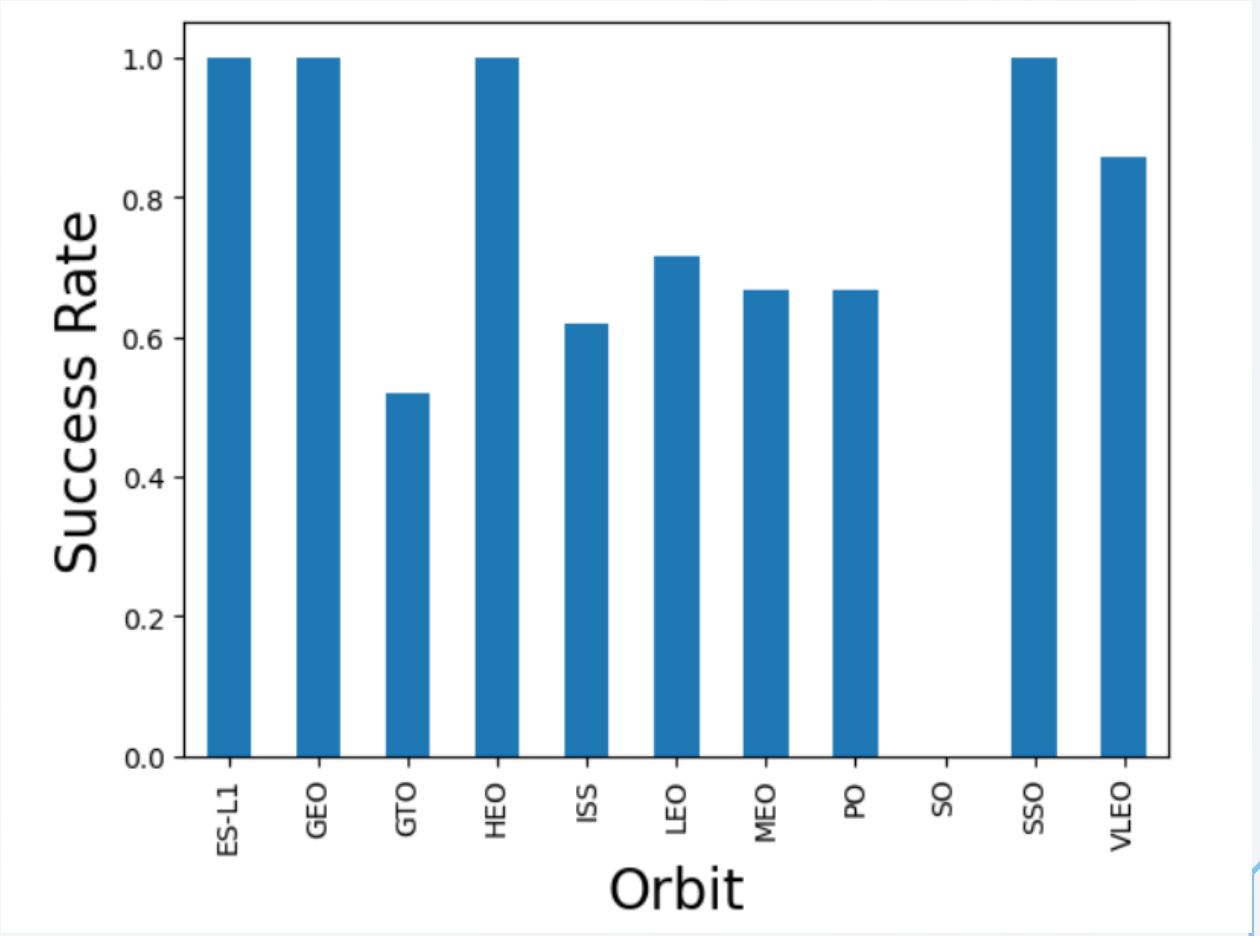
Explanation:

- For each launch site the higher payload mass gives higher success rate
- Most launches with payload mass more than 7500 kg were successful
- KSC LC 39A has also high success rate for payload mass less than 5000 kg

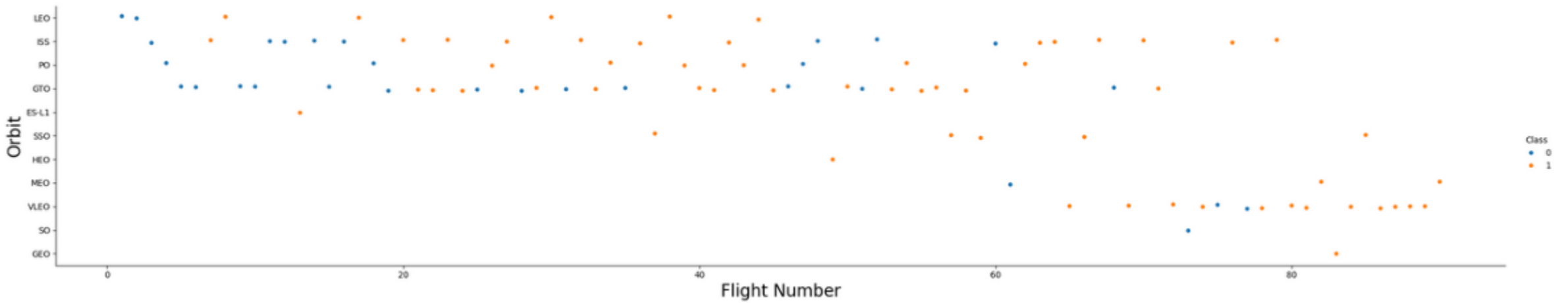
Success Rate vs. Orbit Type

Explanation:

- There are 4 orbits with maximum success rate: ES-L1, GEO, HEO, SSO
- SO orbit has 0% success rate
- GTO, ISS, LEO, MEO, PO and VLEO orbits have success rate between 50% and 85%



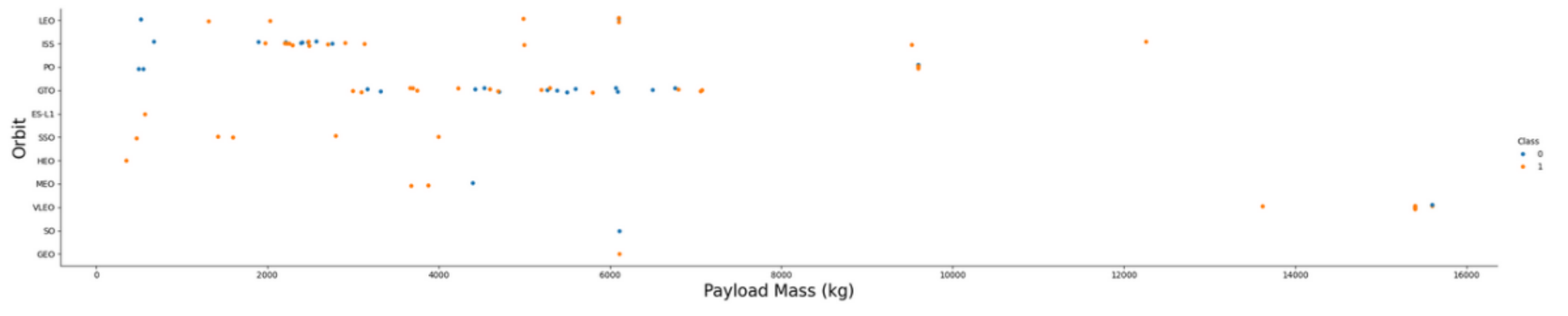
Flight Number vs. Orbit Type



Explanation:

- For LEO orbit the success appears related to the number of flights
- However, for GTO orbit there isn't no relationship between flight number and success
- We can again see 100% success rate for 4 orbits

Payload vs. Orbit Type



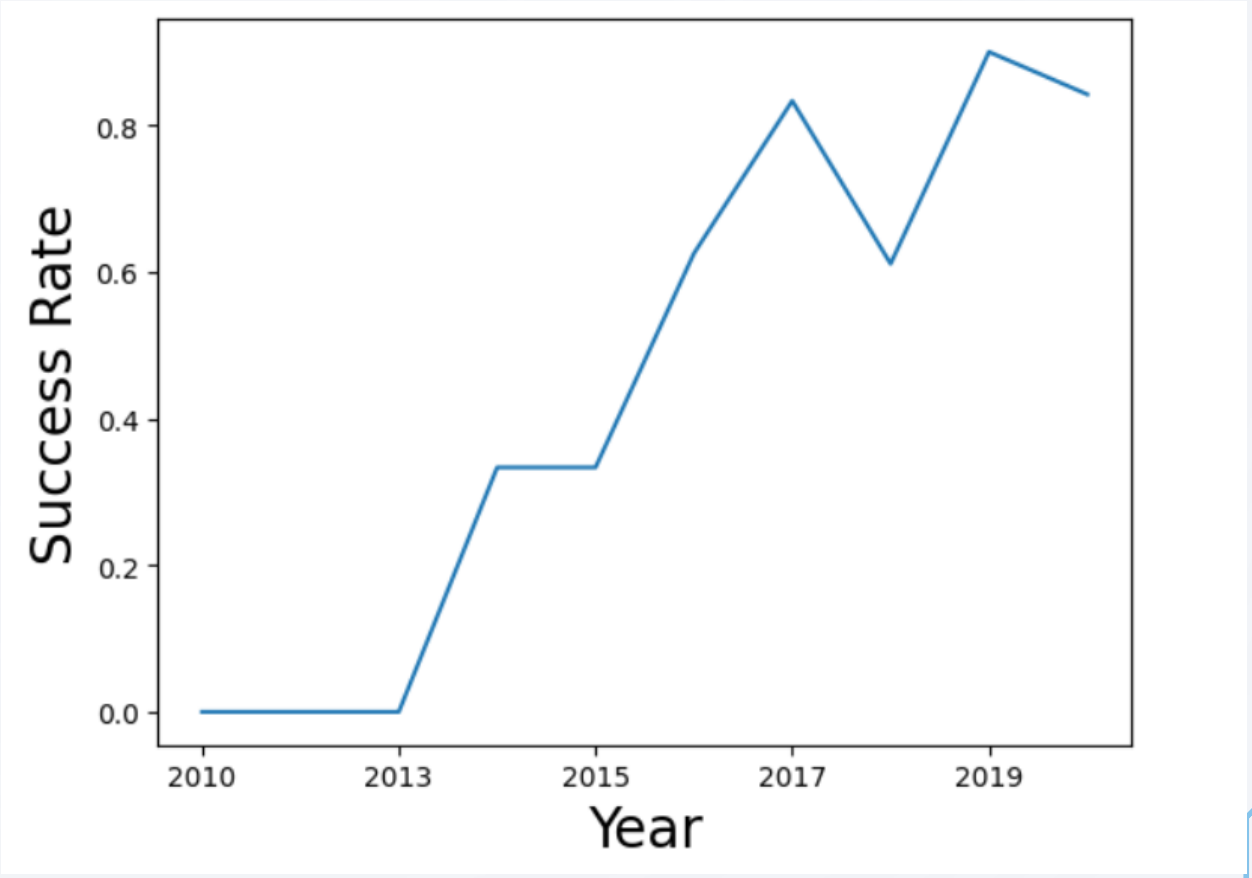
Explanation:

- Heavy payloads have a negative influence on GTO orbit and positive on ISS

Launch Success Yearly Trend

Explanation:

- The success rate 2013 kept increasing till 2020
- There is decline in 2018
- From 2010 to 2013 there is 0% success rate



All Launch Site Names

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

Explanation:

- We see the names of the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (paracl
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 (paracl
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atte
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atte
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atte

Explanation:

- We see first 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

sum(PAYLOAD_MASS__KG_)
45596

Explanation:

- We see the total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

Explanation:

- We see an average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

min(Date)
2015-12-22

Explanation:

- We see the date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Explanation:

- We see the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Explanation:

- We see the total number of successful and failure mission outcomes

Landing_Outcome	Total
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

Explanation:

- We see the names of the boosters which have carried the maximum payload mass

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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

Explanation:

- We see the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

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

Landing_Outcome	Count
Success (ground pad)	9
Failure (drone ship)	5

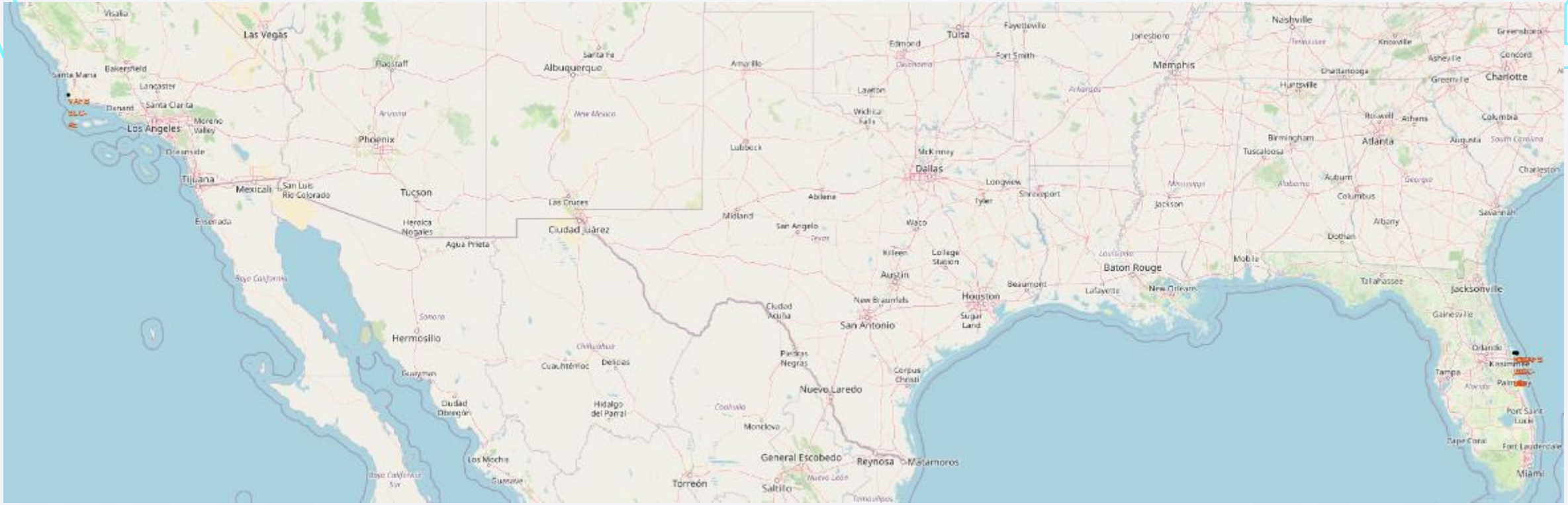
Explanation:

- We see 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.

Section 3

Launch Sites Proximities Analysis

Map marked with launch sites



Explanation:

- All launch sites in proximity to the Equator line, because the Earth is moving faster at the equator than any other place on the surface. This speed helps the rockets keep up a good speed to stay in orbit
- All launch sites are very close to the coast, because it minimizes the risk of exploding near people

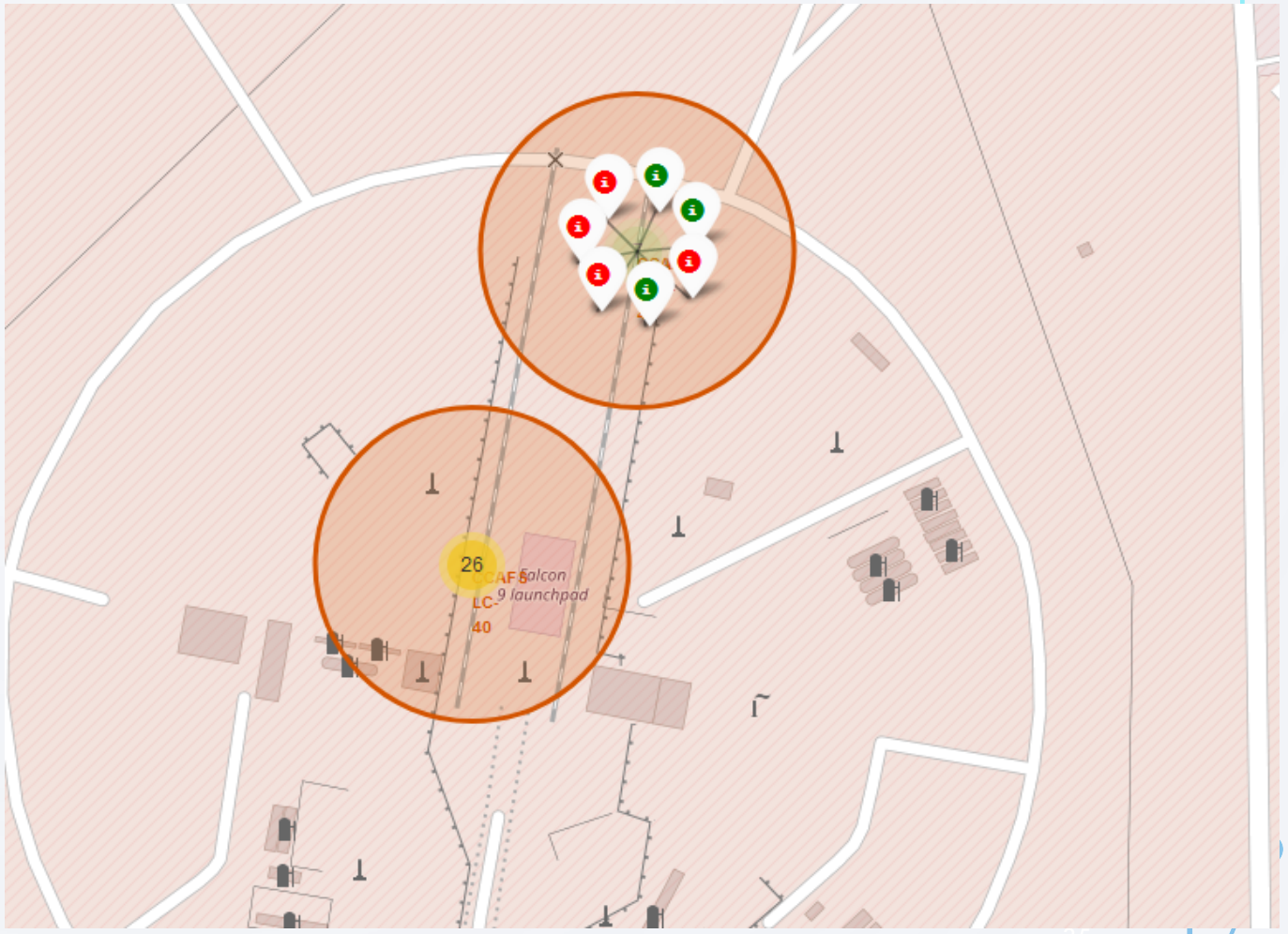
Color-labeled launches on the map

Explanation:

- All launch records are labeled into green or red

(where green is successful launch and red - failure) markers, so we can easy identify which launch sites have high success rates

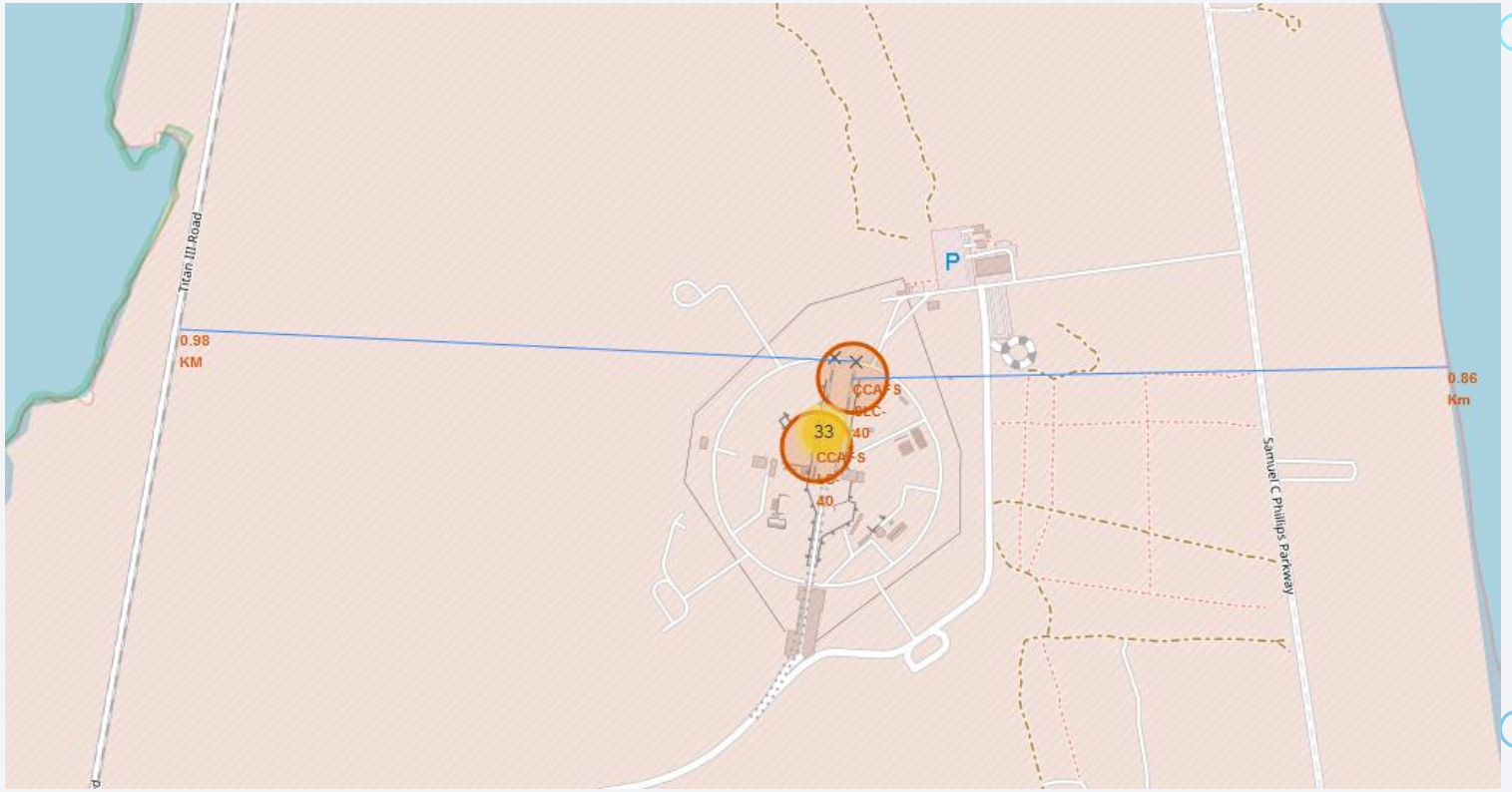
- The map also shows the number of all launches from each launch site

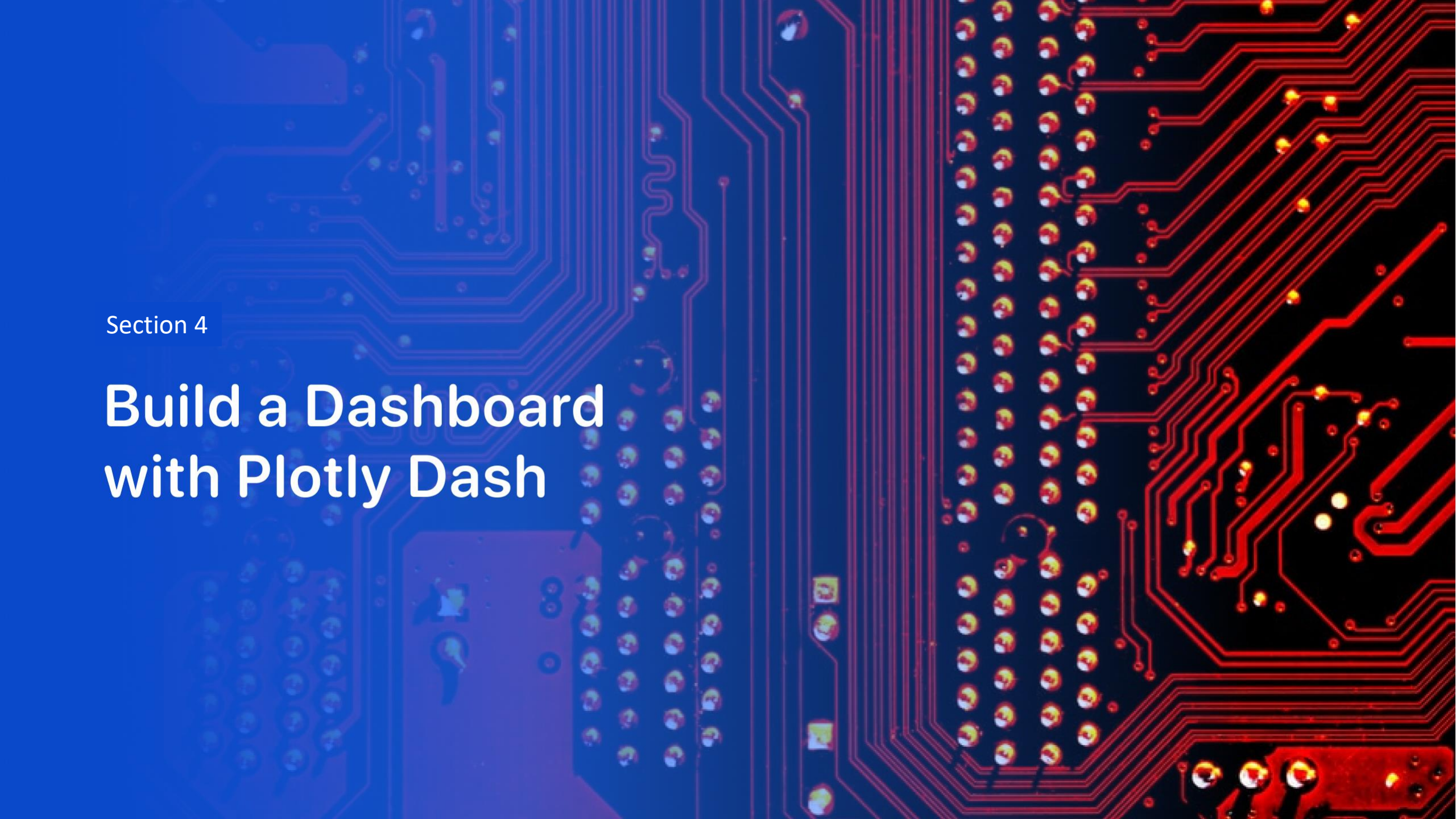


The distances between a launch site to its proximities

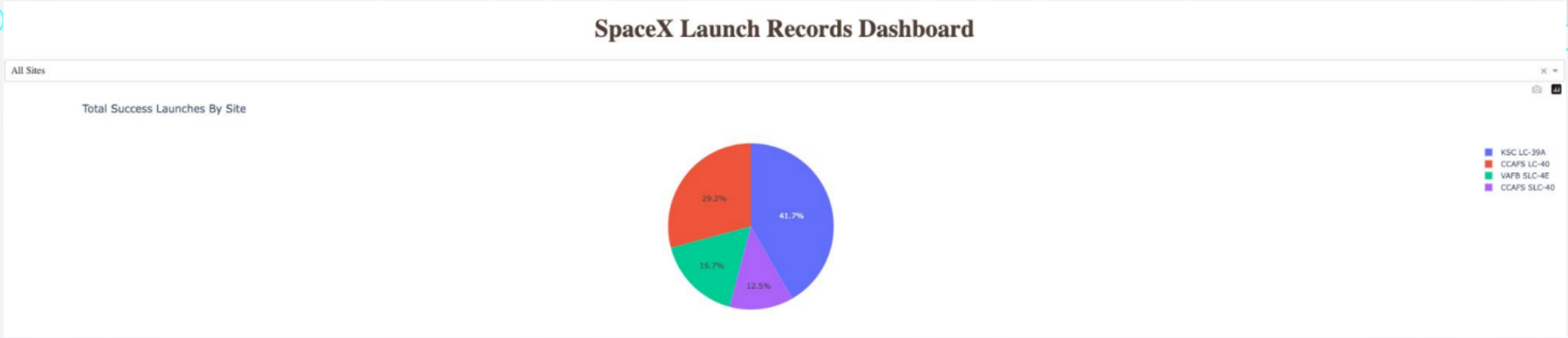
Explanation:

- For the launch site CCAFS SLC-40 we can see that:
- distance to the nearest railway (0.98 km)
- distance to the nearest coastline (0.86 km)
- So this launch site is close to highway, railway and coastline but the distance to the nearest city is quite large




Section 4
Build a Dashboard with Plotly Dash

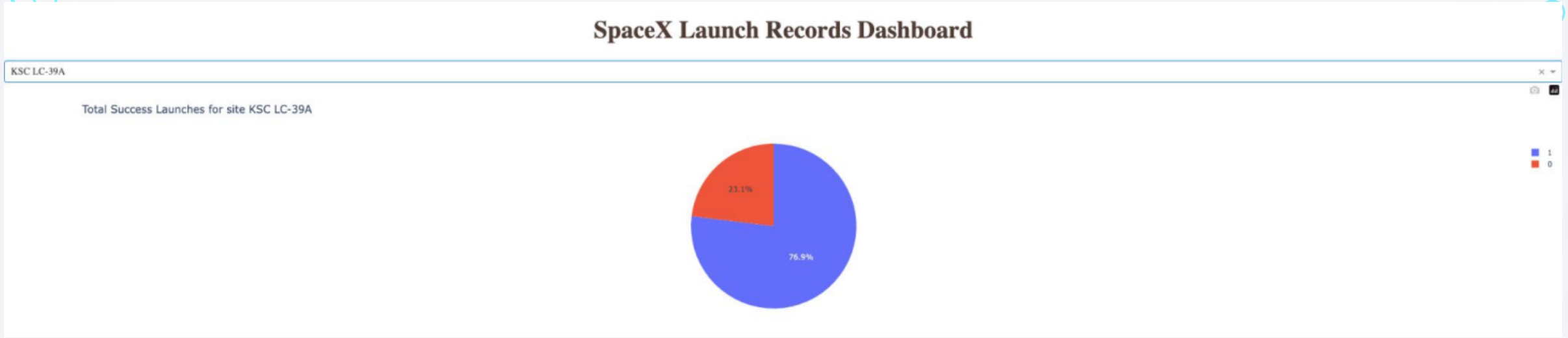
All sites launch success rate



Explanation:

- From the chart above we can get which launch sites are more successful
- So, we can see, that KSC LC-39A launch site has the highest number of successful launches (41.7%)

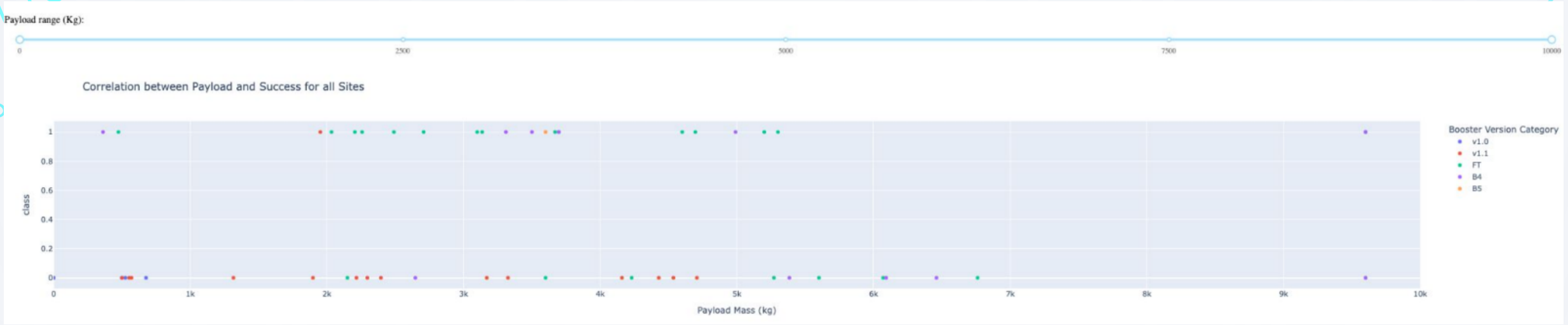
The highest success rate launch site



Explanation:

- The KSC LC-39A had a total of 13 launches, 10 were successful (76.9%) and only 3 failed (23.1%)

All sites Launch success vs Payload mass



Explanation:

- From the scatter charts we can easily identify that payloads between 2000 and 5500 kg have the highest success rate
- We also see which versions of boosters are more successful

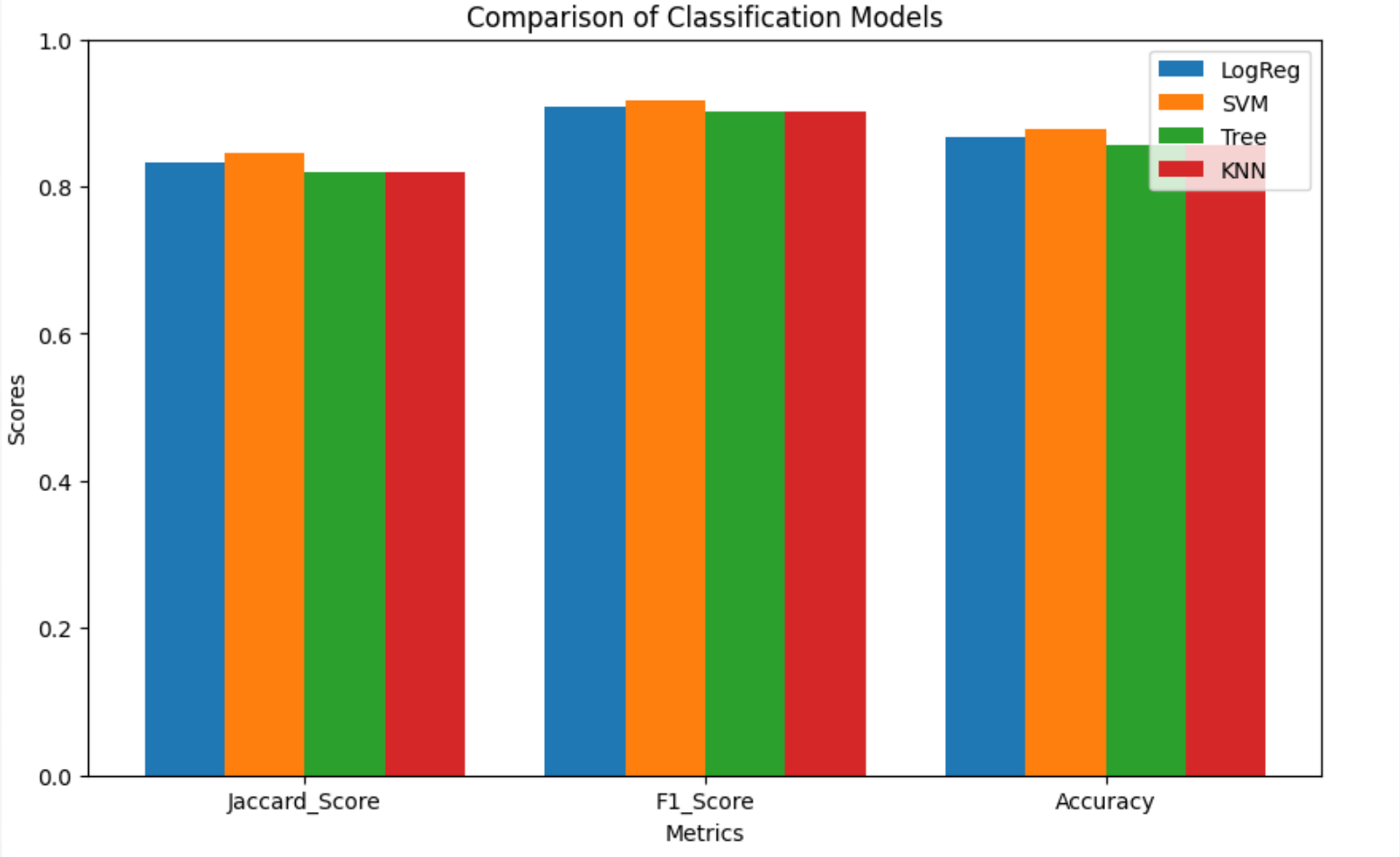
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

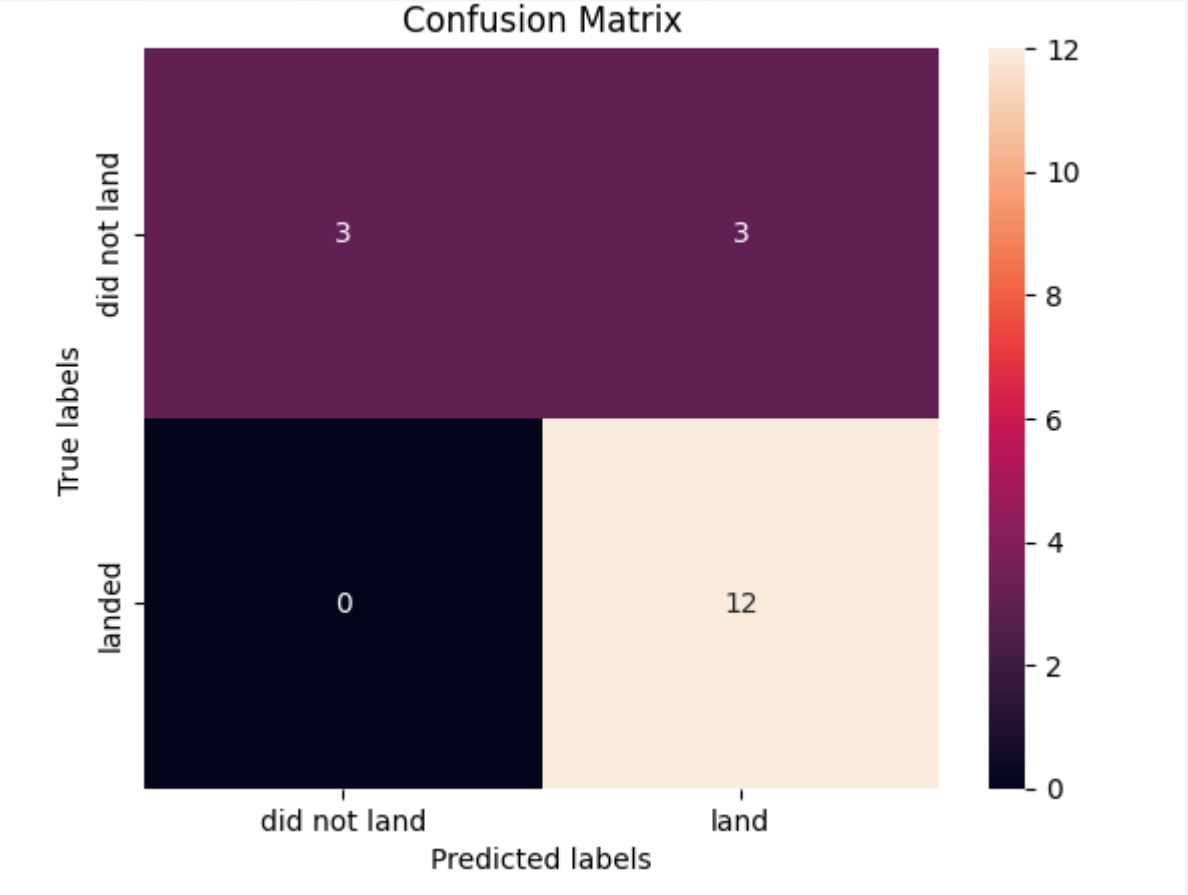
- On the bar chart we can see that SVM model (colored in orange) is the best model with highest Jaccard score, F1 score and Accuracy
- This scores is calculated for the whole dataset, as the accuracy of each classification model is the same on the test set (0.8333)



Confusion Matrix

Explanation:

- Examining the confusion matrix, we see that Support Vector Machine (SVM) model can distinguish between the different classes. We see that the major problem is false positives (FP)



Conclusions

• The Support Vector Machine (SVM) algorithm is the most effective machine learning classification method for this dataset. The model's accuracy and scores are high, allowing us to reliably predict the success of Falcon 9's first stage landing.

• Key factors influencing the success of Falcon 9's first stage landing include:

- Payload Mass (kg): Launches with a lower payload mass achieve better outcomes
- Orbit Type: Orbits such as ES-L1, GEO, HEO, and SSO boast a 100% success rate
- Launch Site: KSC LC-39A exhibits the highest launch success rate

Appendix

A. Detailed Methods and Libraries

Algorithms and Models:

Support Vector Machine (SVM): Utilized with a linear kernel. The model showed the highest accuracy and prediction reliability for Falcon 9's first stage landing success.

Logistic Regression: Employed to predict binary outcomes. This model provided competitive but slightly lower performance compared to SVM.

Decision Tree: Used for its interpretability and ease of implementation. Though not the top performer, it offered insights into feature importance.

K-Nearest Neighbors (KNN): Selected for its simplicity. While effective, it was slightly less accurate than SVM and Logistic Regression.

Libraries:

Scikit-Learn: Version 0.24.2
Pandas: Version 1.3.3
NumPy: Version 1.21.2
Matplotlib: Version 3.4.3

Parameters:

For SVM, the regularization parameter C was set to 1.0.
For Logistic Regression, default parameters were used.
For Decision Tree, the max_depth parameter was set to 5.
For KNN, n_neighbors was set to 5.

B. Data Collection and Preprocessing

Data Sources:

The dataset was obtained from SpaceX's open data portal, including details on launch outcomes, payload mass, orbit type, and launch site.

Data Cleaning:

Missing values were handled using imputation techniques. Specifically, mean imputation was applied to numerical features.

Categorical variables were encoded using one-hot encoding to ensure compatibility with machine learning algorithms.

Feature Engineering:

Created new features such as Payload_Mass_Binned to categorize payload mass into discrete bins.

Encoded orbit types and launch sites into numerical values for model training.

C. Additional Tables and Visualizations

Accuracy Metrics:

The following table shows the accuracy, Jaccard score, and F1 score for each model:

Metric	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.819444	0.819444
F1_Score	0.909091	0.916031	0.900763	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

Feature Importance:

Decision Tree feature importances highlighted Payload_Mass and Orbit_Type as critical features affecting launch success.

D. Additional Results and Discussion

Sensitivity Analysis:

Varying the C parameter in the SVM model showed minimal impact on accuracy, indicating robustness of the chosen parameter value.

Testing with different depths for the Decision Tree model confirmed that a maximum depth of 5 provided the best balance between bias and variance.

Alternative Models:

Attempts to use ensemble methods like Random Forest and Gradient Boosting showed comparable but not superior performance to SVM.

E. Technical Details

Software and Hardware:

Analysis performed on a machine with an Intel i7 processor and 16GB RAM.

All scripts were executed in a Jupyter Notebook environment running Python 3.8.

Thank you!