Outline

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

Executive Summary

Methodologies Applied for Data Analysis:

- Data collection using web scraping and the SpaceX API.
- Exploratory Data Analysis (EDA), encompassing data cleaning, visualization, and interactive analytics.
- Machine learning for predictive modeling.

Summary of Results:

- Successfully gathered valuable data from public sources.
- EDA highlighted the key features most predictive of launch success.
- Machine learning identified the optimal model for determining the most critical factors driving this opportunity, utilizing all the collected data.

Introduction

Project Background and Context

SpaceX has become a leader in the commercial space industry by significantly lowering space travel costs. The company offers Falcon 9 rocket launches for $62 million, far less than competitors who charge over $165 million, largely due to SpaceX's ability to reuse the first stage of the rocket. This project will leverage public data and machine learning models to predict whether the first stage will successfully land, helping to estimate launch costs.

Questions to be Addressed

- How do factors like payload mass, launch site, number of flights, and orbital parameters influence the success of the first stage landing?
- Has the success rate of first stage landings improved over time?
- Which algorithm is most effective for binary classification in predicting first stage landing success?

Section 1

Methodology

Methodology

Data Collection Methodology:
•Utilized SpaceX REST API.
•Employed web scraping from Wikipedia.

Data Wrangling:
•Filtered and cleaned the data.
•Handled missing values.
•Applied One Hot Encoding to prepare the data for binary classification.

Exploratory Data Analysis (EDA):
•Conducted EDA using visualization tools and SQL queries.
•Created interactive visual analytics with Folium and Plotly Dash.

Predictive Analysis:
•Developed, tuned, and evaluated classification models to optimize performance and ensure the best predictive outcomes.

Data Collection

The data collection process used both the SpaceX REST API and web scraping from Wikipedia to gather complete launch details for analysis.

Data from SpaceX REST API:

•FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

Data from Wikipedia:

•Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Booster Version, Booster landing, Date, Time.

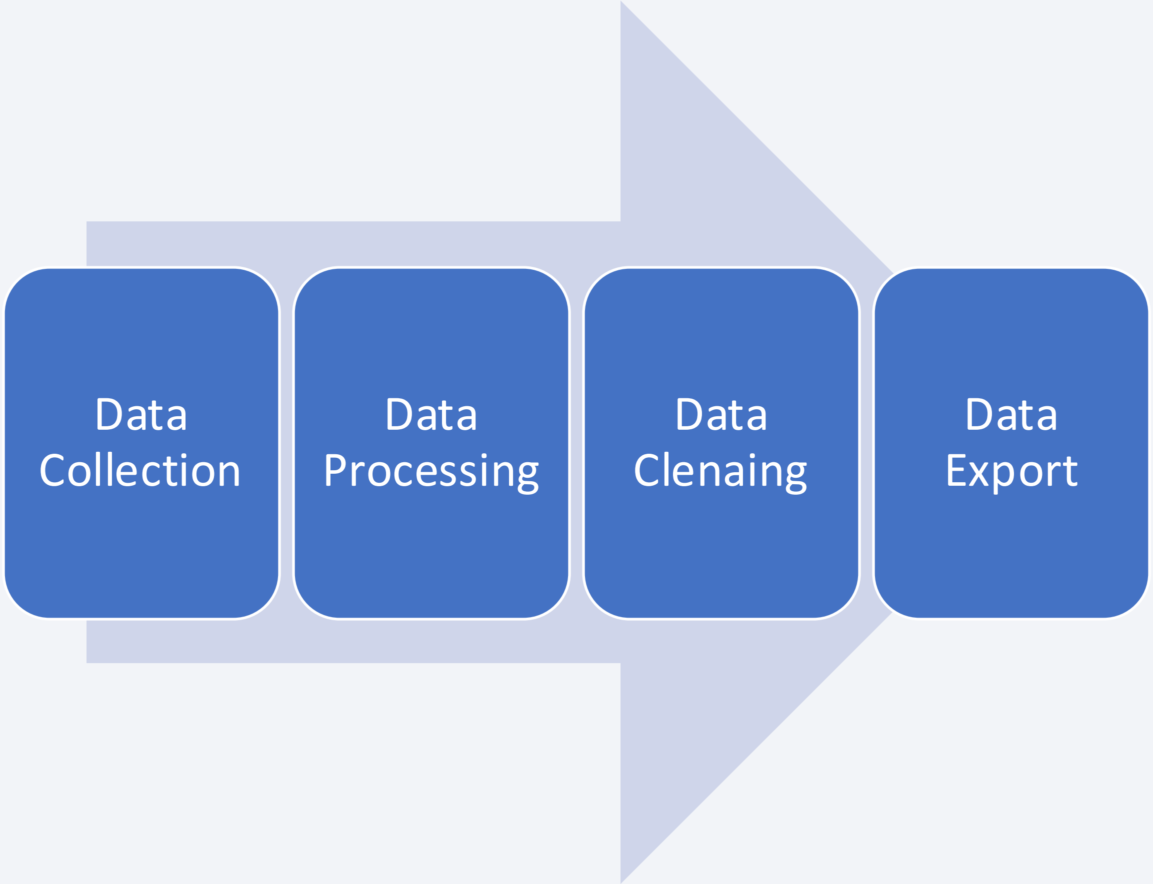
Data Collection – SpaceX API

Data Collection: Request rocket launch data from the SpaceX API and extract the response into a DataFrame using .json_normalize().

Data Processing: Apply custom functions to retrieve specific launch information and filter the DataFrame to include only Falcon 9 launches.

Data Cleaning: Replace missing values in the Payload Mass column with the calculated mean.

Data Export: Convert the cleaned data into a dictionary and export it to a CSV file.



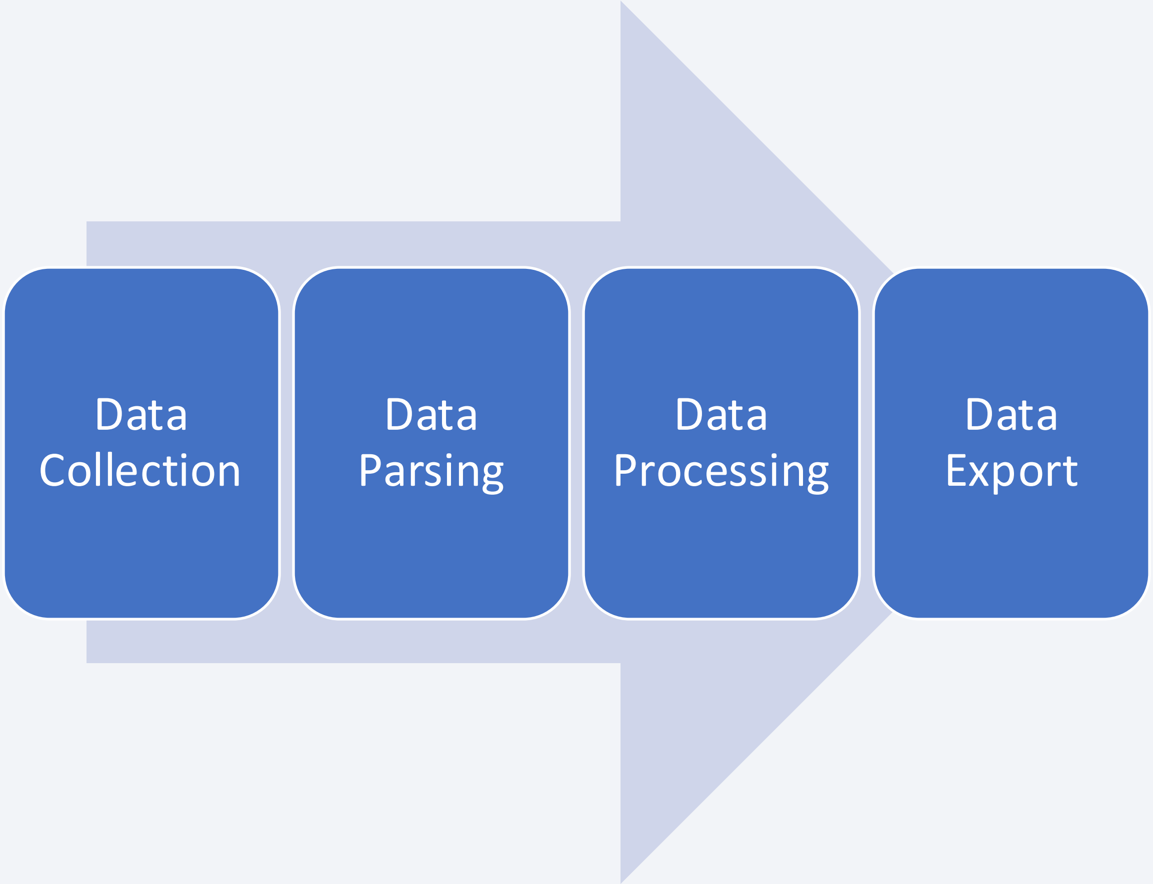
Data Collection - Scraping

1.Data Collection: Request Falcon 9 launch data from Wikipedia and create a BeautifulSoup object to parse the HTML response.

2.Data Parsing: Extract column names and collect data by parsing the HTML tables.

3.Data Processing: Construct the parsed data into a dictionary and convert it into a DataFrame.

4.Data Export: Export the final DataFrame to a CSV file.



Data Wrangling

The data set includes various outcomes for booster landings, with some attempts failing due to accidents. The outcomes are categorized by location and success:

- **True Ocean/False Ocean**: Landing succeeded/failed in a specific ocean region.
- **True RTLS/False RTLS**: Landing succeeded/failed on a ground pad.
- **True ASDS/False ASDS**: Landing succeeded/failed on a drone ship.

These outcomes were converted into binary training labels, where "1" indicates a successful landing and "0" indicates an unsuccessful one.

EDA with Data Visualization

Charts were created to analyze various relationships, including:

- **Scatter plots**: Examined relationships between variables to identify potential features for machine learning models.
- **Bar charts**: Compared discrete categories to highlight relationships between categories and measured values.
- **Line charts**: Illustrated trends in data over time, particularly in time series analysis.

EDA with SQL

SQL queries were performed for Exploratory Data Analysis (EDA), focusing on:

•Identifying unique launch sites and specific records based on launch site prefixes.
•Calculating total and average payload mass for specific missions and booster versions.
•Listing key events like the first successful ground pad landing and successful drone ship landings with specific payloads.
•Counting successful and failed mission outcomes.
•Identifying boosters with the maximum payload and failed drone ship landings in 2015.
•Ranking landing outcomes by frequency within a specified date range.

Build an Interactive Map with Folium

Key mapping features included:

- **Launch Site Markers**: Added markers with circles, popup labels, and text labels for NASA Johnson Space Center and all launch sites to display their geographic locations and proximity to the equator and coasts.
- **Outcome Markers**: Used colored markers (green for success, red for failure) and Marker Clusters to highlight launch site success rates.
- **Distance Visualization**: Added colored lines to show distances from the KSC LC-39A launch site to nearby features such as railways, highways, coastlines, and the closest city.

Build a Dashboard with Plotly Dash

Key interactive features included:

•**Launch Site Dropdown**: Added a dropdown menu for selecting specific launch sites.
•**Success Rate Pie Chart**: Added a pie chart to display the total successful launches for all sites, and success vs. failure counts for a selected site.
•**Payload Mass Slider**: Included a slider to filter by payload mass range.
•**Payload vs. Success Scatter Chart**: Added a scatter chart to illustrate the correlation between payload mass and launch success for different booster versions.

Predictive Analysis (Classification)

Data Preparation: Converted the "Class" column into a NumPy array and standardized the data using StandardScaler.
Data Splitting: Divided the data into training and testing sets using the train_test_split function.
Model Tuning: Created a GridSearchCV object with cross-validation (cv = 10) to find the best parameters, then applied it to various models including Logistic Regression, SVM, Decision Tree, and KNN.
Model Evaluation: Calculated accuracy on the test data using .score(), examined confusion matrices, and identified the best-performing method using the Jaccard and F1 scores.

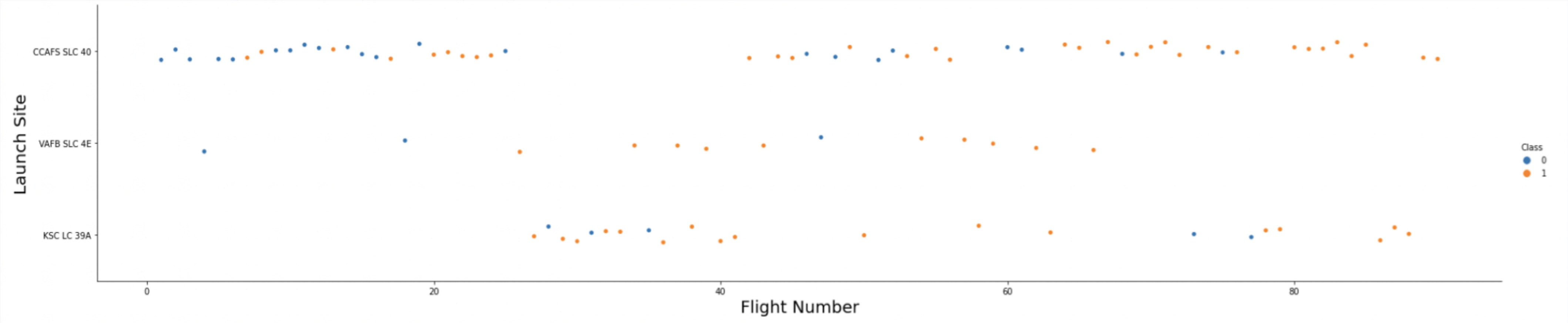
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



Key observations include:

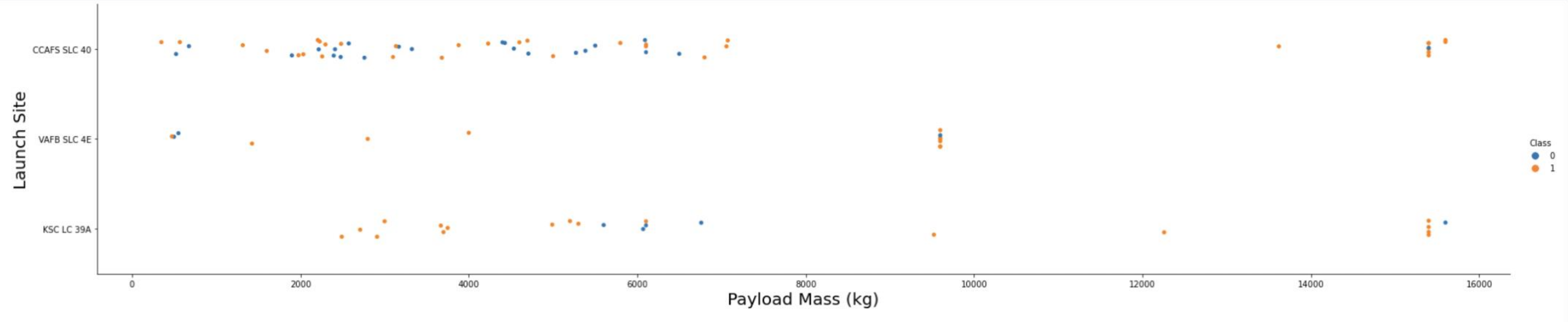
•Early flights mostly failed, while recent flights have all succeeded.

•The CCAFS SLC 40 launch site accounts for about half of all launches.

•VAFB SLC 4E and KSC LC 39A sites have higher success rates.

•There's a trend suggesting that newer launches tend to have a higher success rate.

Payload vs. Launch Site



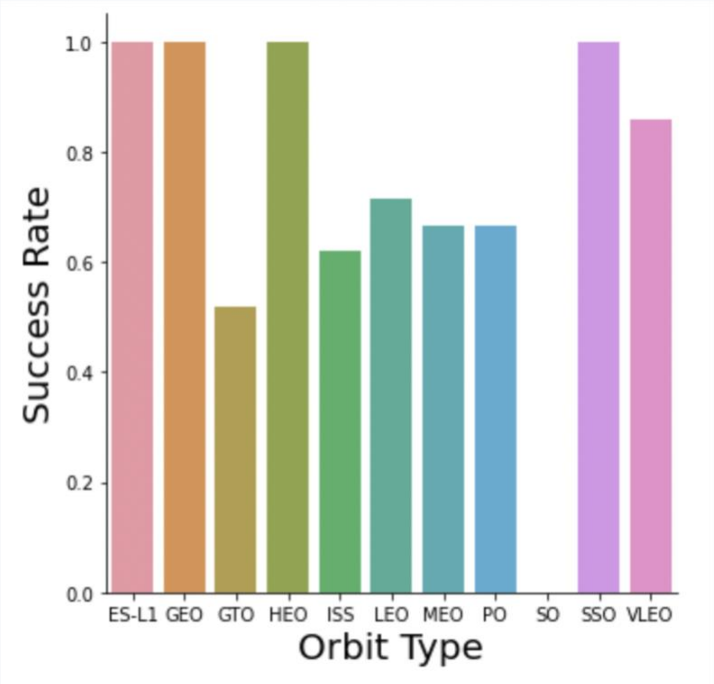
Key findings include:

•Higher payload mass is generally associated with a higher success rate across all launch sites.

•Most launches with a payload mass over 7,000 kg were successful.

•KSC LC 39A has a 100% success rate for payloads under 5,500 kg.

Success Rate vs. Orbit Type



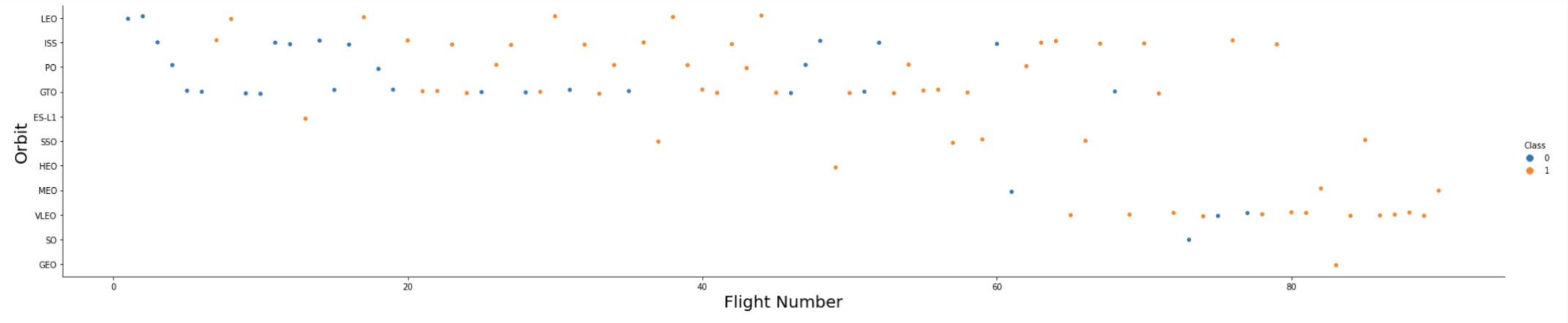
Summary of orbit success rates:

•**100% Success Rate**: ES-L1, GEO, HEO, SSO.

•**0% Success Rate**: SO.

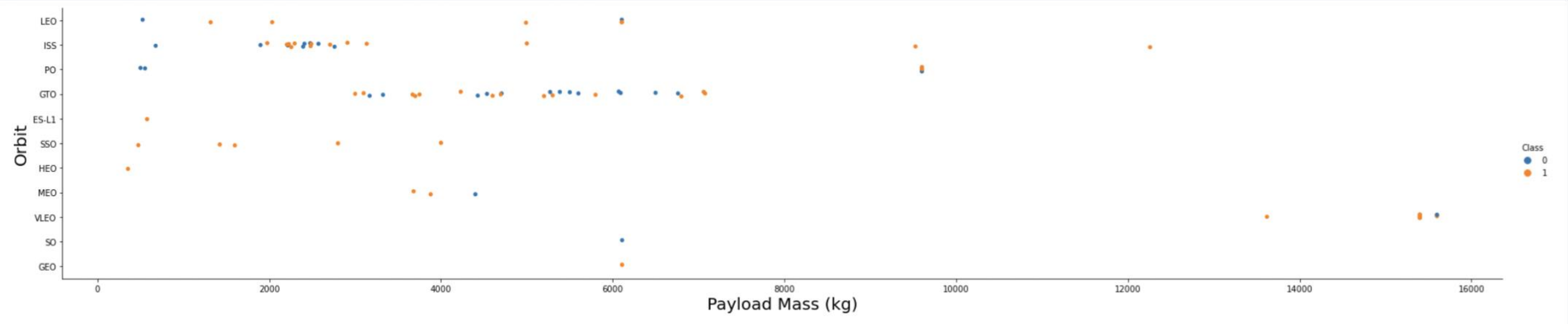
•**50% to 85% Success Rate**: GTO, ISS, LEO, MEO, PO.

Flight Number vs. Orbit Type



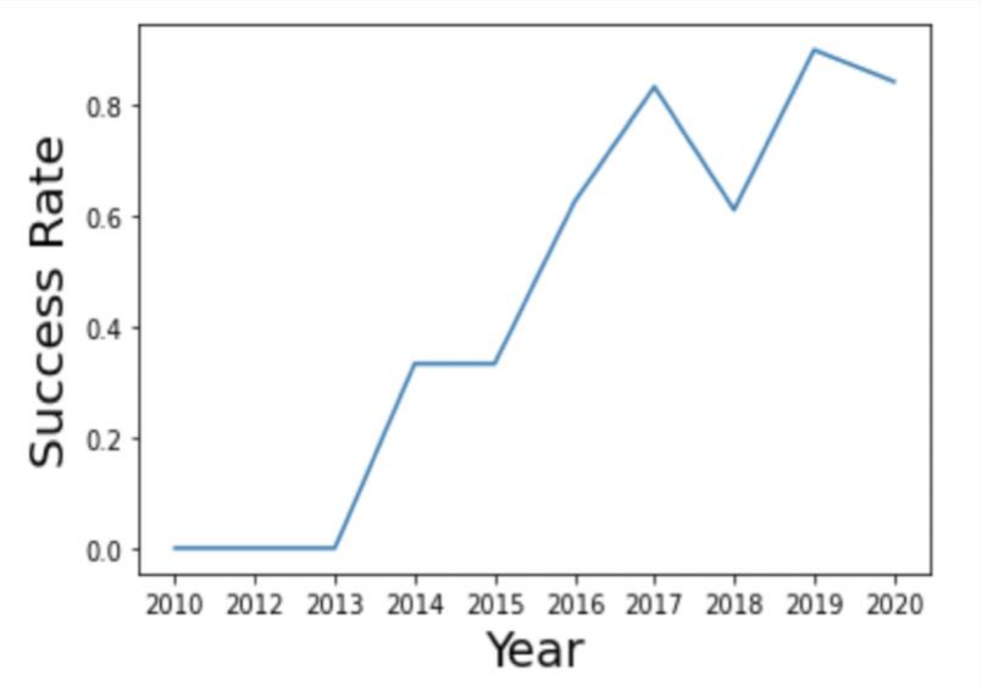
In the LEO orbit, success is linked to the number of flights, whereas in the GTO orbit, there is no apparent relationship between success and flight number.

Payload vs. Orbit Type



Heavy payloads negatively affect GTO orbits but have a positive impact on Polar LEO (ISS) orbits.

Launch Success Yearly Trend



The success rate consistently improved from 2013 to 2020.

All Launch Site Names

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

Launch Site Names Begin 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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

total_payload_mass
45596

First Successful Ground Landing Date

first_successful_landing
2015-12-22

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

Total Number of Successful and Failure Mission Outcomes

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

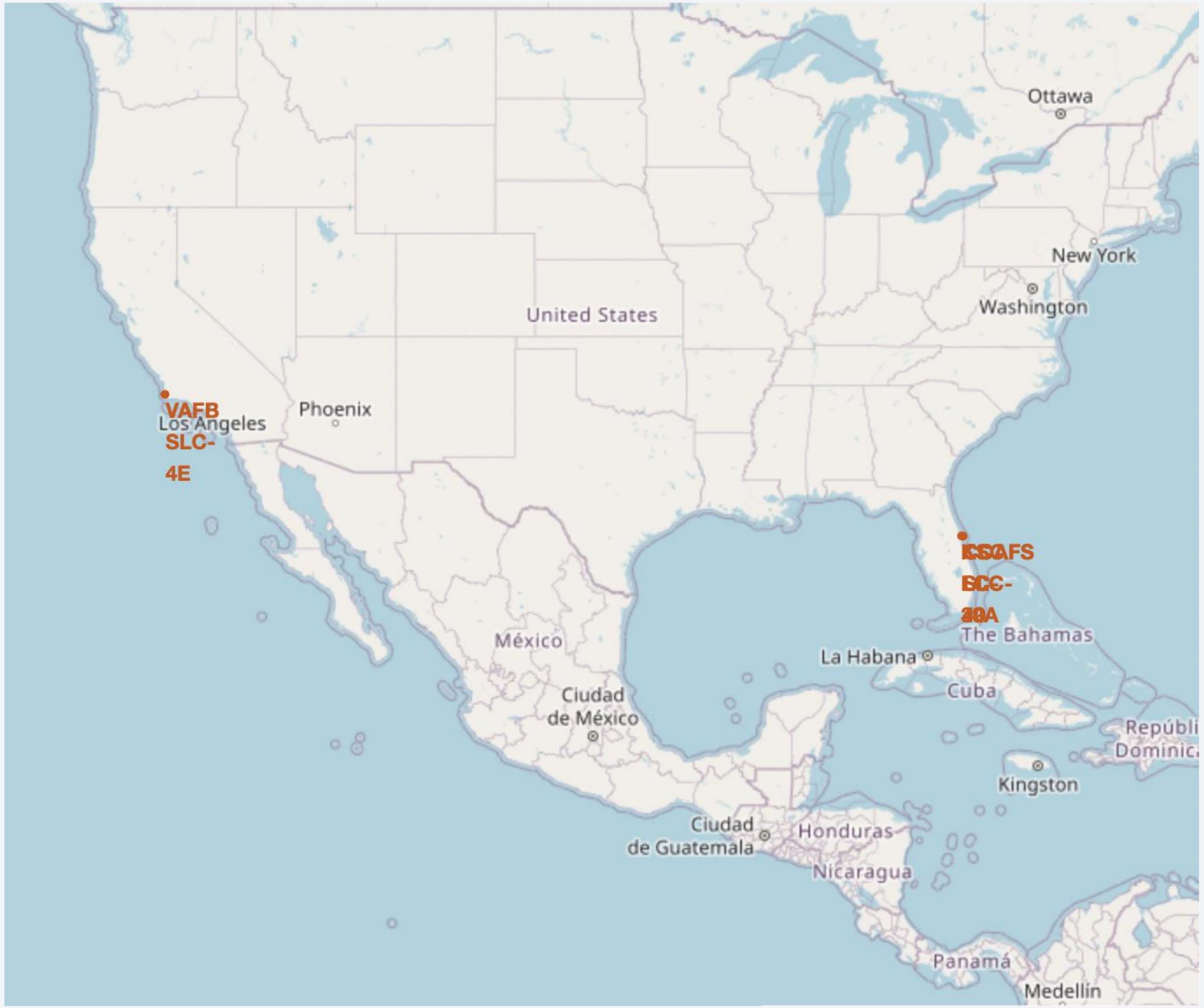
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

landing__outcome	count_outcomes
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

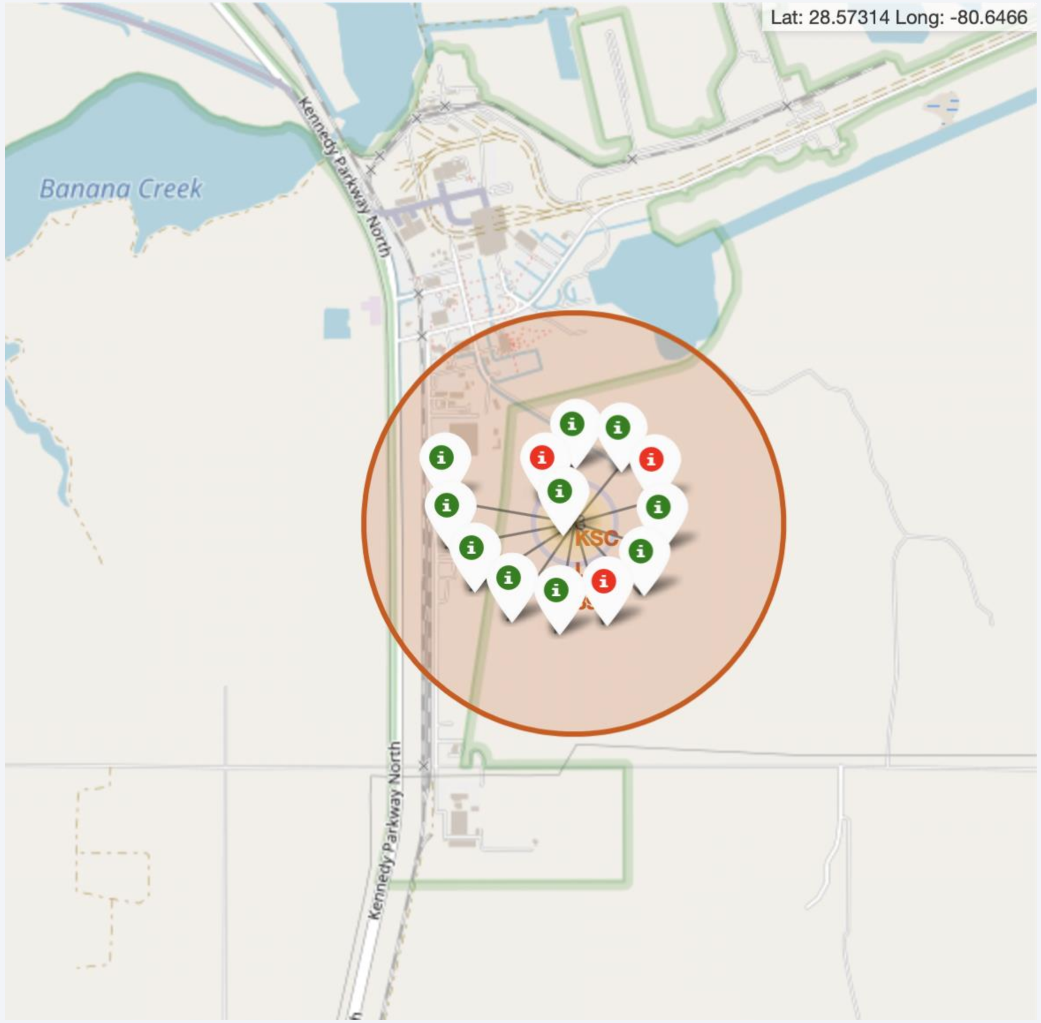
Section 3

Launch Sites Proximities Analysis

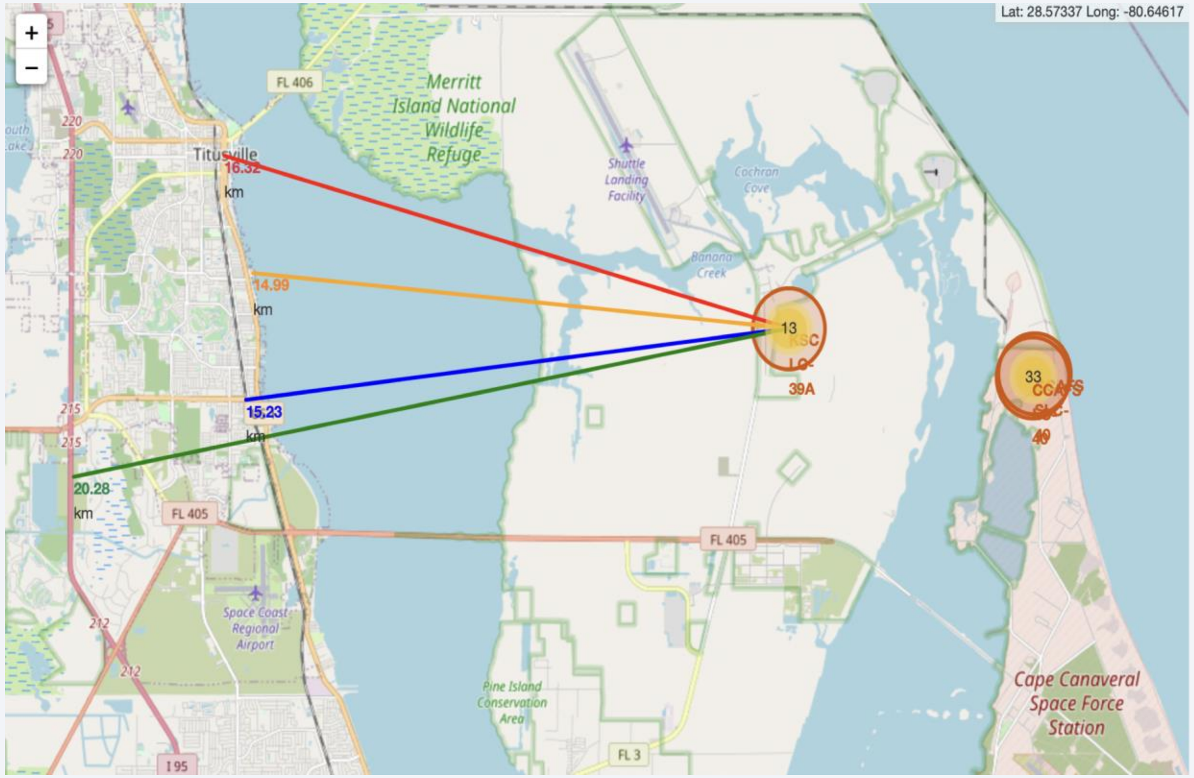
<Folium Map Screenshot 1>



<Folium Map Screenshot 2>



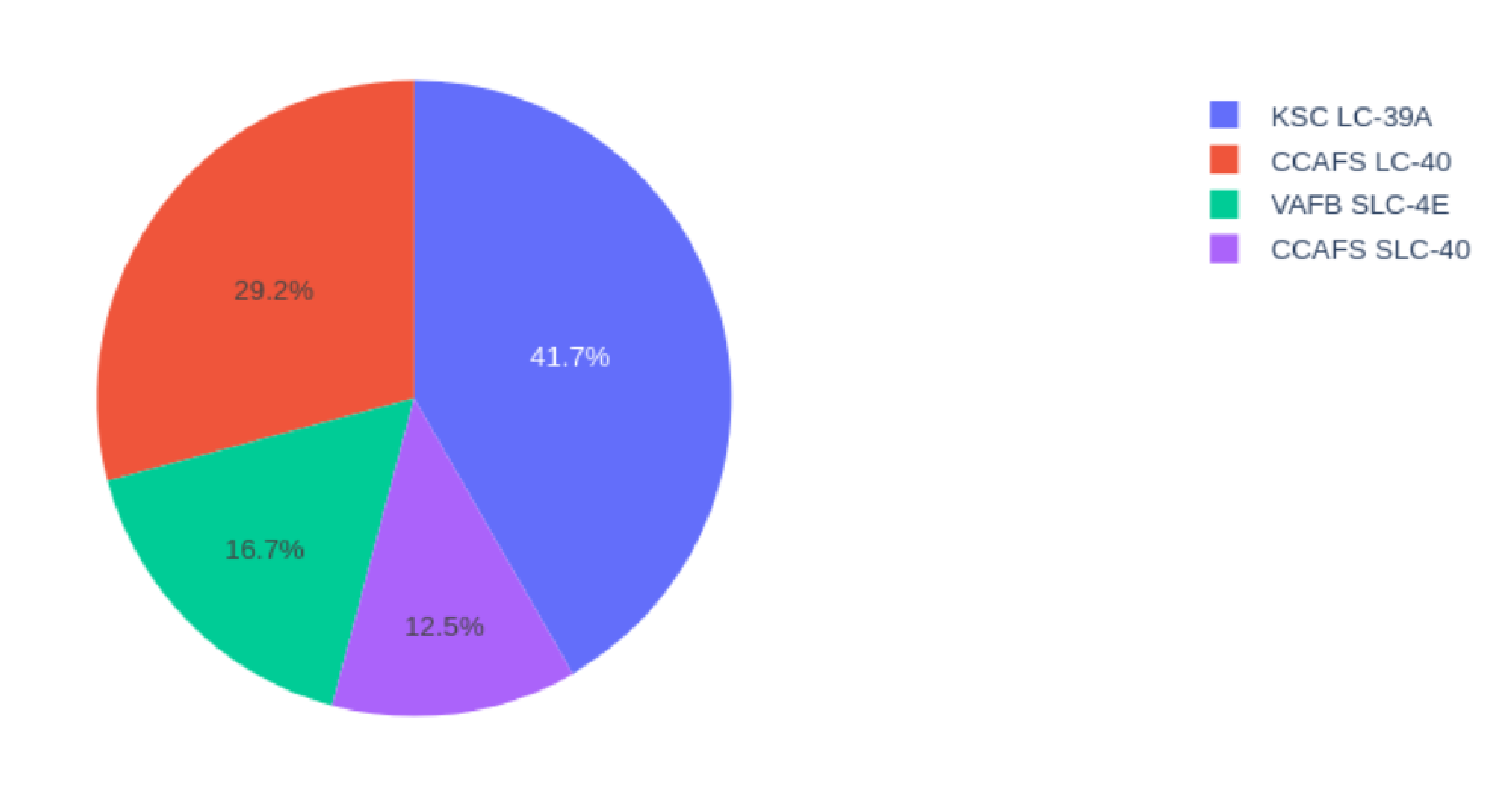
<Folium Map Screenshot 3>



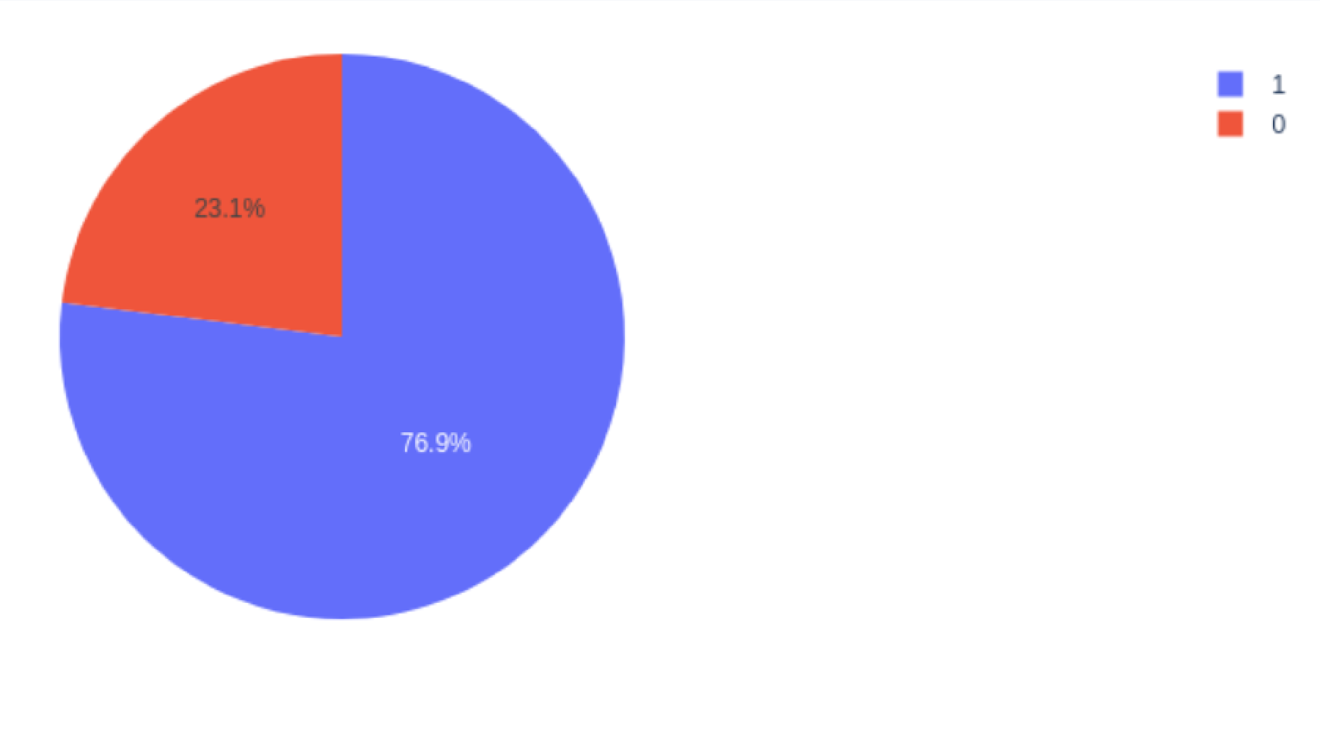
Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



<Dashboard Screenshot 2>



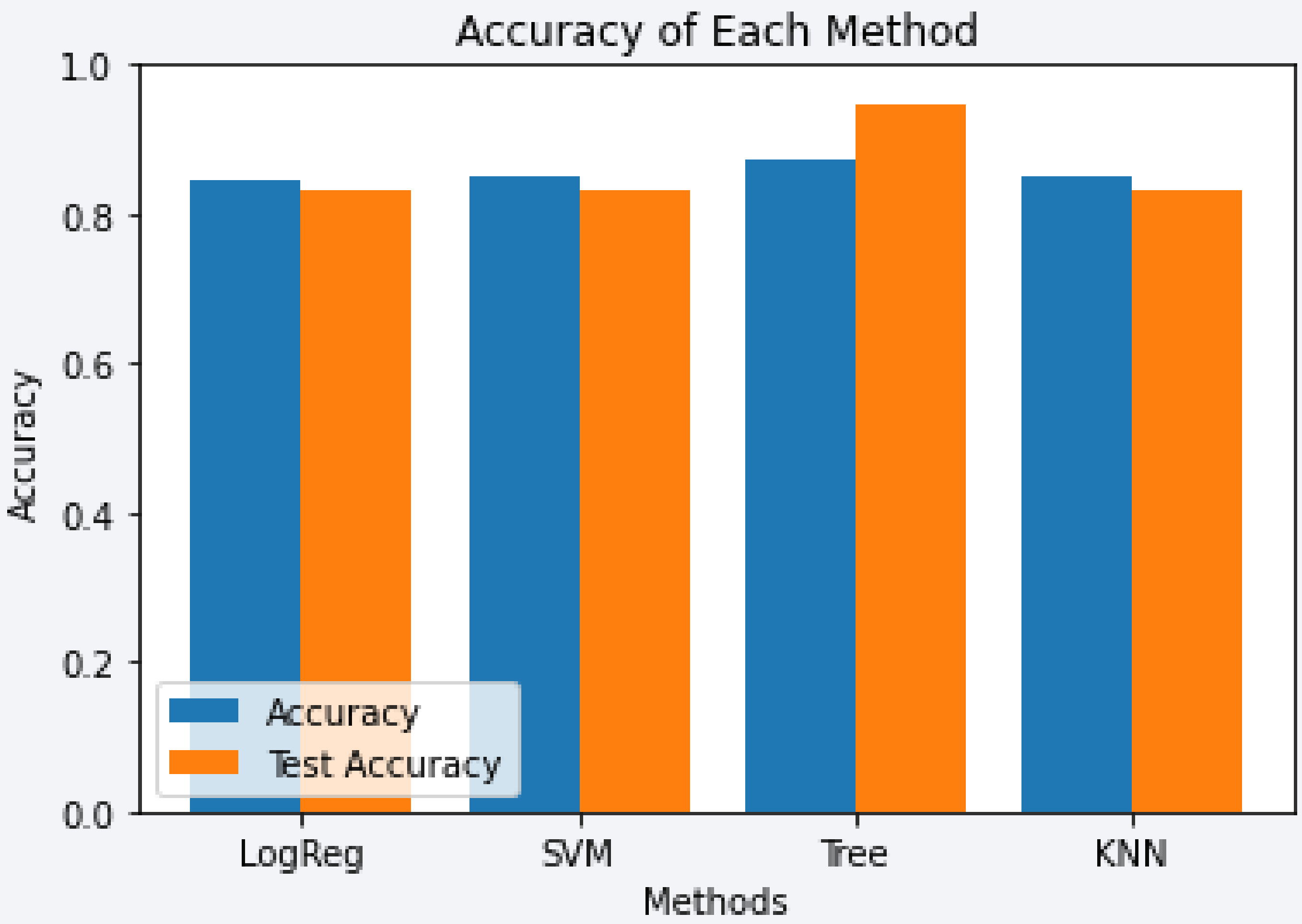
<Dashboard Screenshot 3>

My code didn’t work out for some reason.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

Conclusions

- KSC LC-39A is identified as the best launch site.
- Launches over 7,000 kg are less risky.
- Successful landing outcomes have improved over time due to advancements in processes and rocket technology.
- The Decision Tree Classifier is effective for predicting successful landings and can help increase profits.

Thank you!