IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

John Doe

Feb 1st 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context
  - SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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 SpaceX can reuse the first stage. Therefore, if we can determine if the irst stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.
- Problems you want to find answers
  - How do variables such as payload mass, launch site, number of lights, and orbits affect the success of the first stage landing? - Does the rate of successful landings increase over the years? - What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Using SpaceX Rest API
    - Using Web Scraping from Wikipedia
- Performed data wrangling
    - Filtering the data
    - Dealing with missing values
    - Using One Hot Encoding to prepare the data to a binary classification
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
    - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

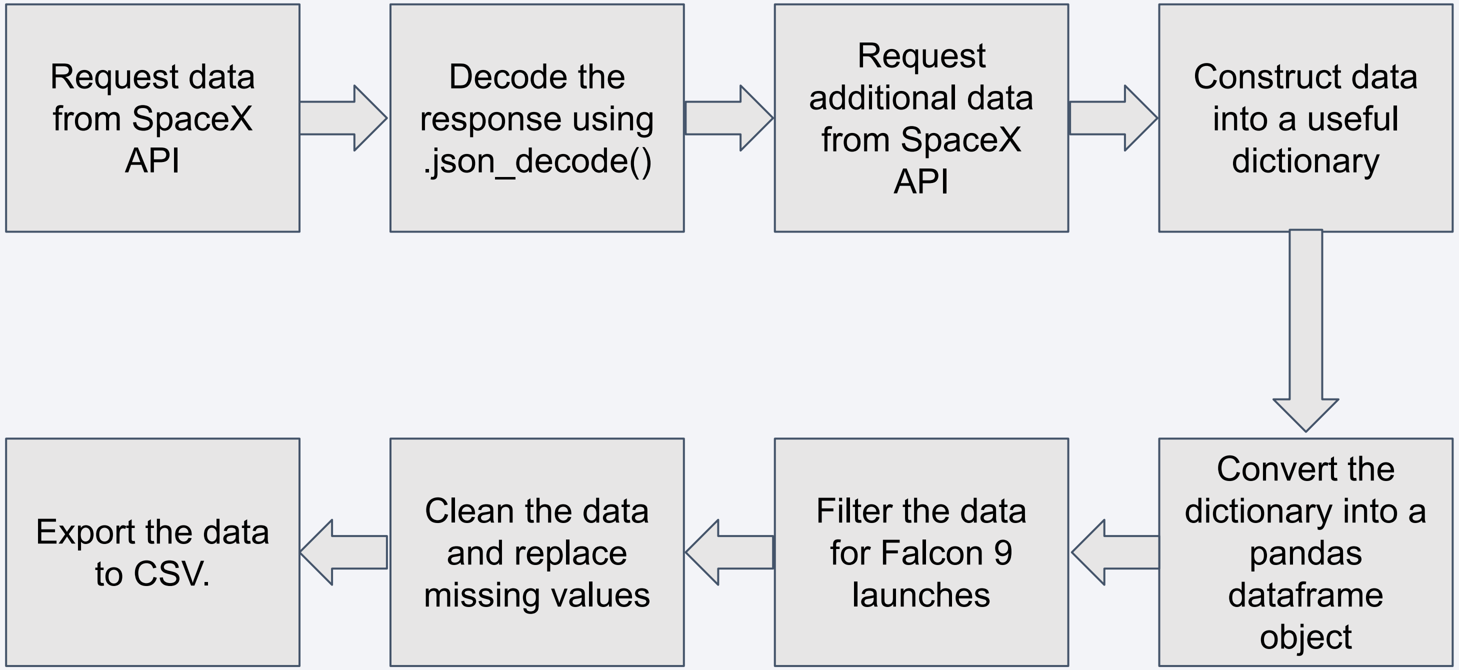
- Data Columns are obtained by using SpaceX REST API
    - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Web Scraping:
    - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

GitHub Link

- [jupyter-labs-spacex-data-collection-api-v2.ipynb]
- [dataset_part_1.csv]

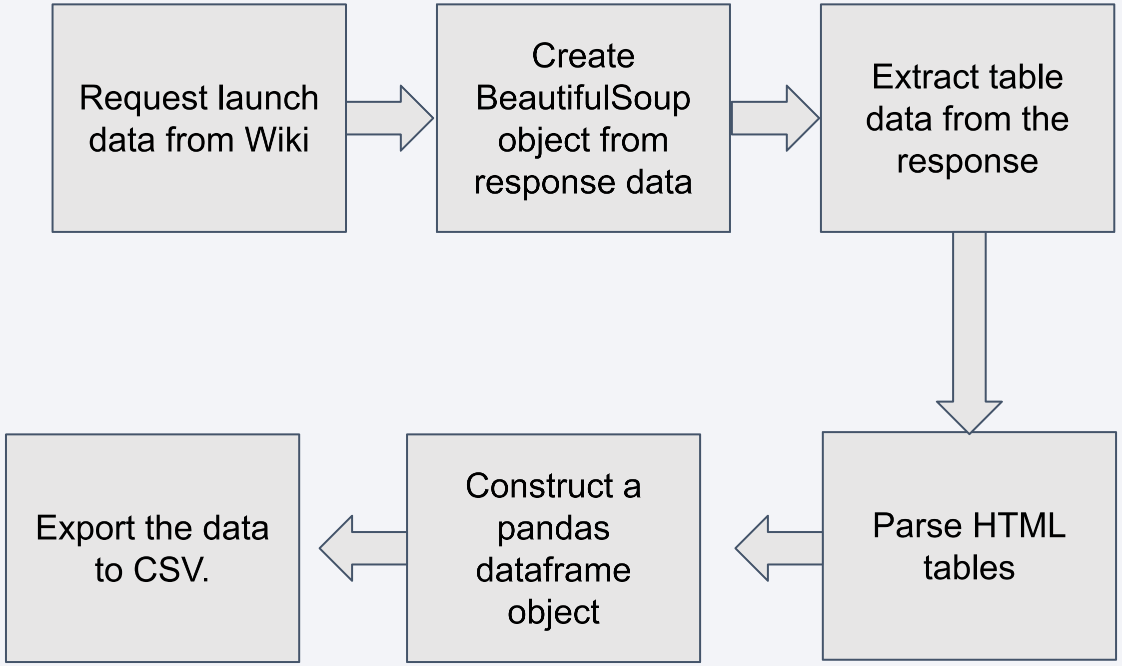


Data Collection – Scraping

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

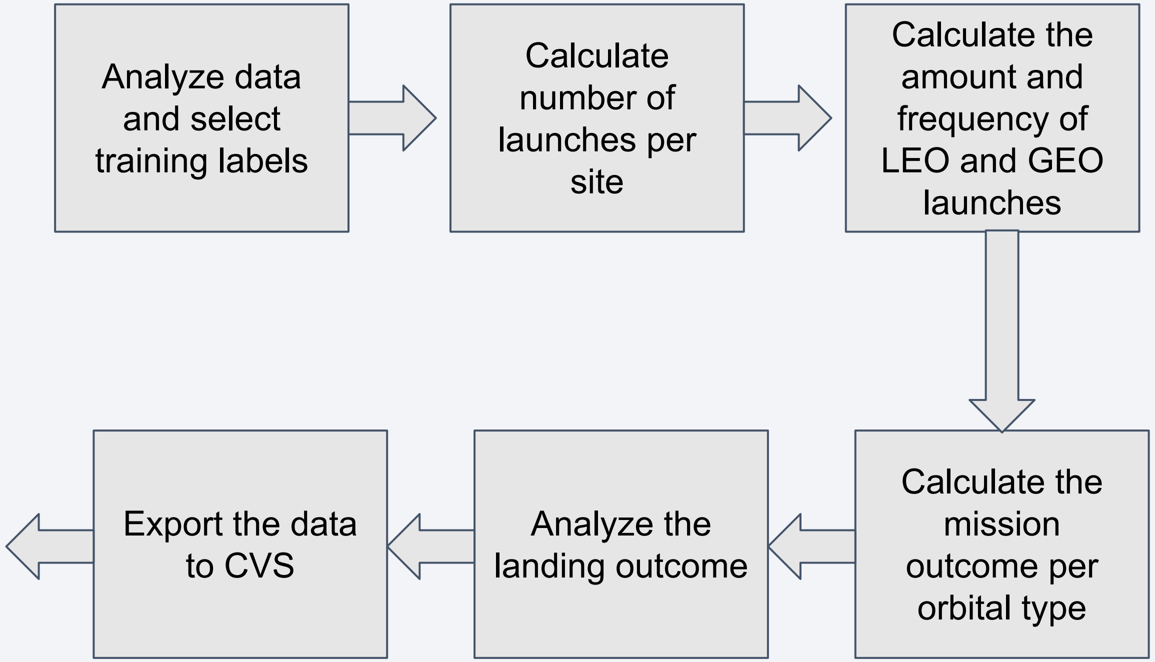
GitHub Link

- [jupyter-labs-webscraping.ipynb]
- [dataset_part_2.csv]



Data Wrangling

- The dataset includes cases where the booster failed to land successfully. Landings are categorized as follows: **True** (successful) or **False** (unsuccessful) for Ocean, RTLS (ground pad), and ASDS (drone ship). These outcomes are converted into training labels: **1** for a successful landing and **0** for an unsuccessful one.



GitHub Link

- labs-jupyter-spacex-Data wrangling-v2.ipynb
- dataset_part_3.csv

EDA with Data Visualization

- Charts Plotted
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend
- **Scatter plots** illustrate relationships between variables and can be useful for machine learning if a pattern exists.
- **Bar charts** compare discrete categories, highlighting relationships between categories and their measured values.
- **Line charts** display trends over time, making them ideal for time series analysis.
- jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- Performed SQL queries
    - 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 the months in year 2015
    - 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
- jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Markers for All Launch Sites:**
  - Placed markers with circles, popup labels, and text labels for all launch sites using their latitude and longitude coordinates to display their locations and proximity to the equator and coasts.
  - Added a marker for NASA Johnson Space Center with a circle, popup label, and text label as the starting location.
- **Colored Markers for Launch Outcomes:**
  - Used green markers for successful launches and red markers for failed launches.
  - Applied Marker Cluster to highlight launch sites with relatively high success rates.
- **Distances from a Launch Site to Nearby Features:**
  - Added colored lines to indicate distances from the KSC LC-39A launch site to nearby features such as railways, highways, coastlines, and the closest city.
- lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

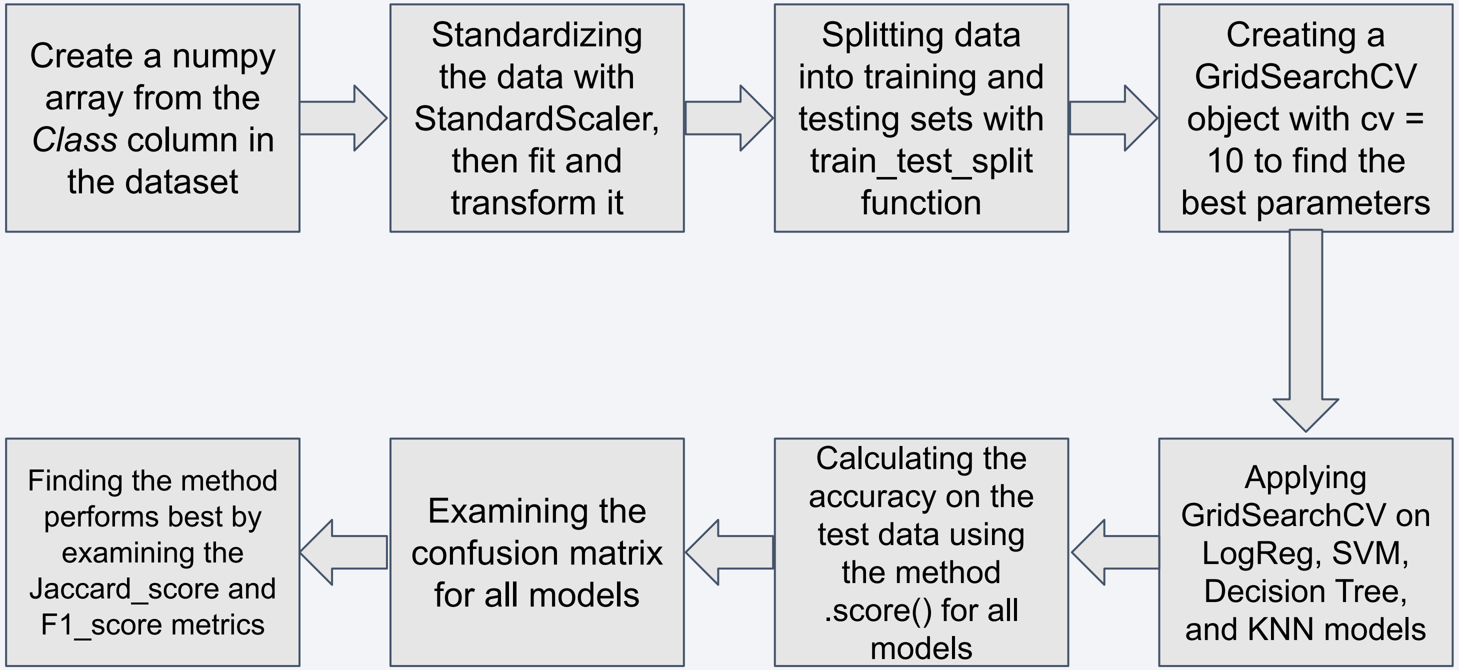
- **Launch Site Selection:**
    - Implemented a drop down menu to allow users to select a specific launch site.
- **Success Launches Pie Chart:**
    - Added a pie chart displaying the total successful launches across all sites.
    - When a specific launch site is selected, the chart shows the success vs. failure count for that site.
- **Payload Mass Range Slider:**
    - Integrated a slider to enable users to select a specific payload range.
- **Scatter Plot: Payload Mass vs. Success Rate:**
    - Created a scatter plot to illustrate the relationship between payload mass and launch success across different booster versions.
- capstone-spacex-dashboard-app.py

Predictive Analysis (Classification)

- Present your predictive analysis method with a flowchart.

GitHub Link

- [SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb]()



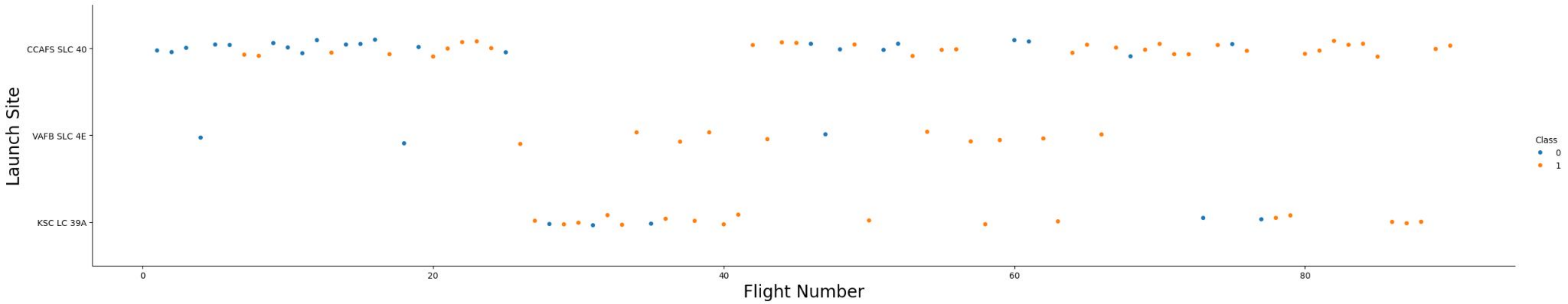
Results

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

Section 2

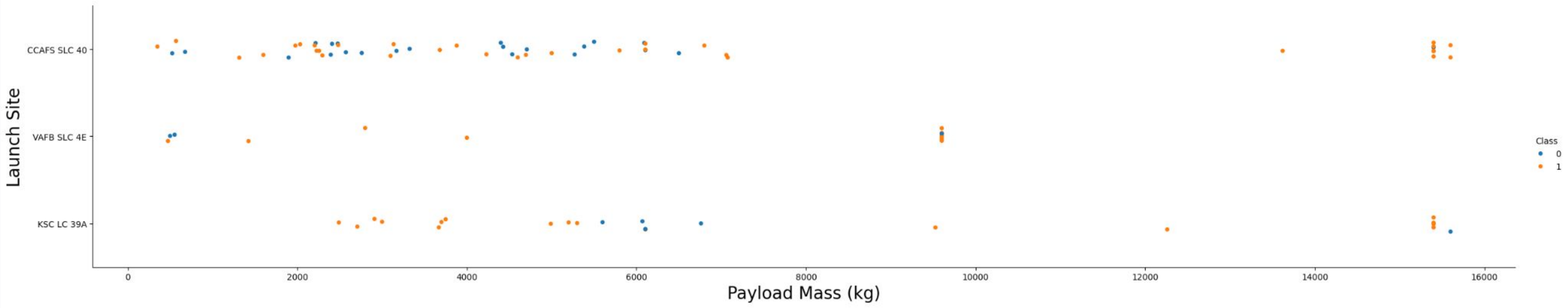
Insights drawn from EDA

Flight Number vs. Launch Site



- Explanation
  - The earliest lights all failed while the latest lights all succeeded.
  - The CCAFS SLC 40 launch site has about a half of all launches.
  - VAFB SLC 4E and KSC LC 39A have higher success rates.
  - It can be assumed that each new launch has a higher rate of success

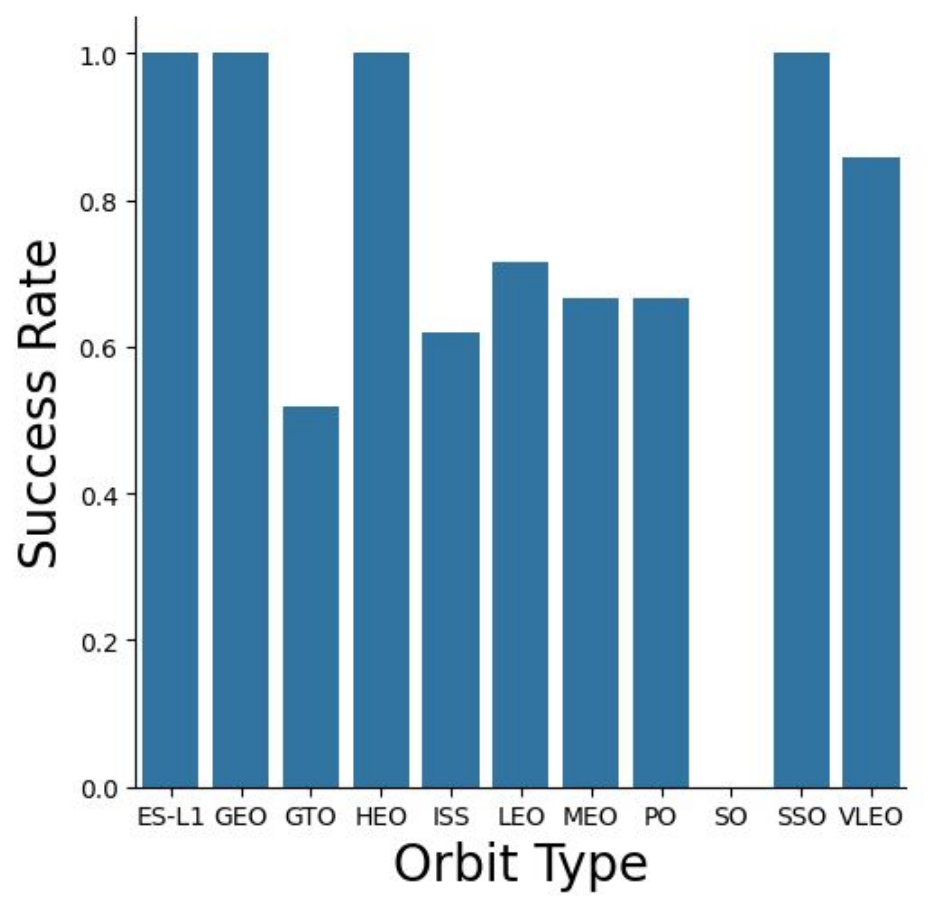
Payload vs. Launch Site



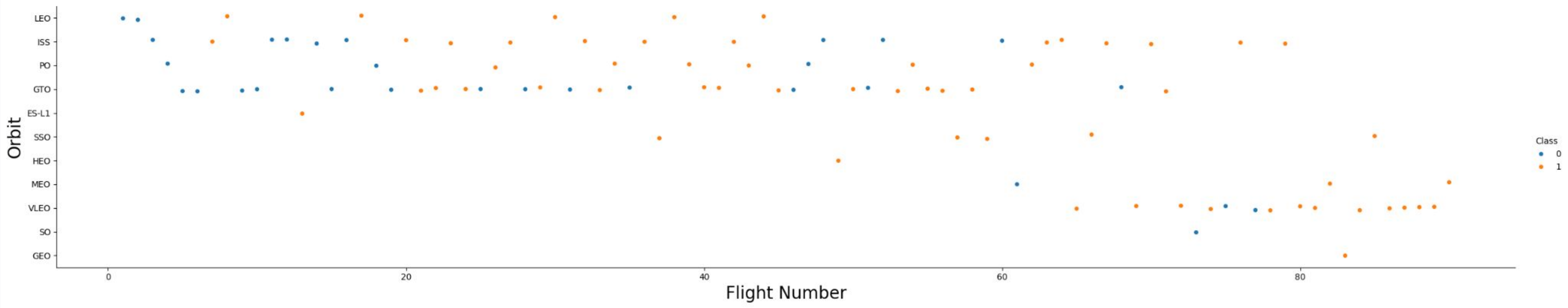
- Explanation
  - For every launch site the higher the payload mass, the higher the success rate.
  - Most of the launches with payload mass over 7000 kg were successful.
  - KSC LC 39A has a 100% success rate for payload mass under 5500 kg.

Success Rate vs. Orbit Type

- Explanation
  - Orbits with 100% success rate
    - ES-L1, GEO, HEO, SSO
  - Orbits with 0% success rate
    - SO
  - Orbits with success rate between 50% and 85%
    - GTO, ISS, LEO, MEO, PO

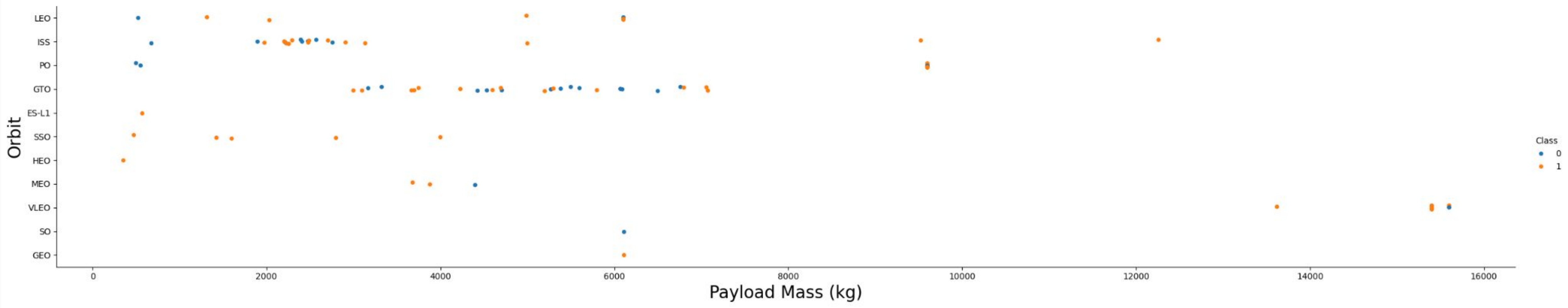


Flight Number vs. Orbit Type



- Explanation
  - In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number for the GTO orbit type.

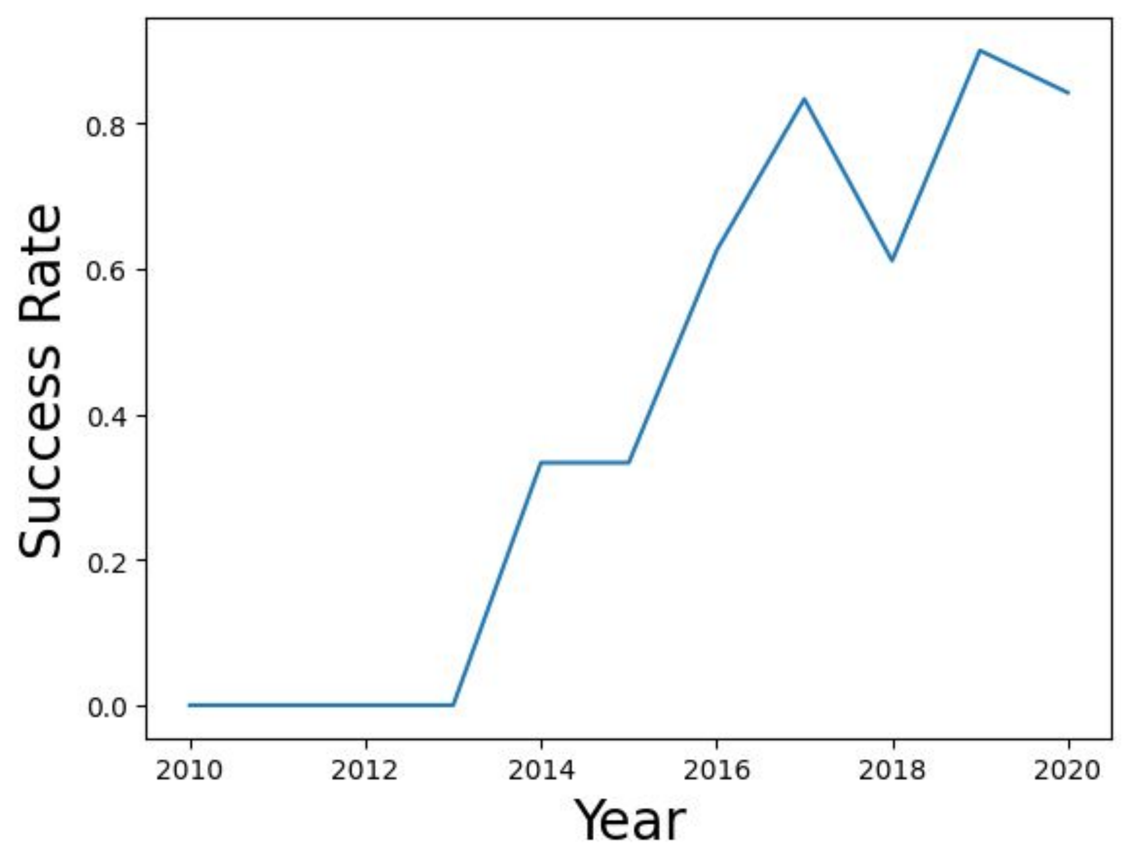
Payload vs. Orbit Type



- Explanation
  - Heavy payloads have a negative influence on GTO orbit types and positive on GTO, and Polar LEO (ISS) orbit types.

Launch Success Yearly Trend

- Explanation
  - The success rate since 2013 increased until 2018 where it saw a slight decline.
  - The success rate continues to improve from 2018 to 2020.



All Launch Site Names

- There are four total site names in the dataset.

```
%sql SELECT distinct launch_site from SPACEXTABLE;
```

[15]

```
* sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

[16] Python

* sqlite:///my_data1.db
Done.

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

This shows launch site names that begin with CCA and limits the results to 5 rows.

Total Payload Mass

```
[17]  %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';

 * sqlite:///my_data1.db
Done.
```

total_payload_mass
45596

- The total mass payload is 45,596 KG

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

[18]

```
* sqlite:///my_data1.db
Done.
```

average_payload_mass
2534.6666666666665

- The Average mass payload for Falcon 9 launches is 2,535 KG

First Successful Ground Landing Date

```
[20] %sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';

 * sqlite:///my_data1.db
Done.
```

first_successful_landing
2015-12-22

- The first successful ground landing was 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[22] %sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000;

 * sqlite:///my_data1.db
Done.
```

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

- This shows all booster versions with payloads between 4,000 KG and 6,000 KG wich successful landings.

Total Number of Successful and Failure Mission Outcomes

```
[23] %sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Summary totals from the dataset. The majority of mission outcomes were successful.

Boosters Carried Maximum Payload

```
[25]  %sql select booster_version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);

 * sqlite:///my_data1.db
Done.
```

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

- A full list of boosters that carried their maximum payload

2015 Launch Records

```
%%sql select SUBSTR(Date, 6, 2) as month, date, booster_version, launch_site, landing_outcome from SPACEXTABLE
        where landing_outcome = 'Failure (drone ship)' AND SUBSTR(Date, 1, 4) = '2015';
```

[34]

```
* sqlite:///my_data1.db
Done.
```

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

- Shows all failed drone to ship landing outcomes in 2015

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

- Ranks 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.

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
    where date between '2010-06-04' and '2017-03-20'
    group by landing_outcome
    order by count_outcomes desc;
```

[36]

```
 * sqlite:///my_data1.db
Done.
```

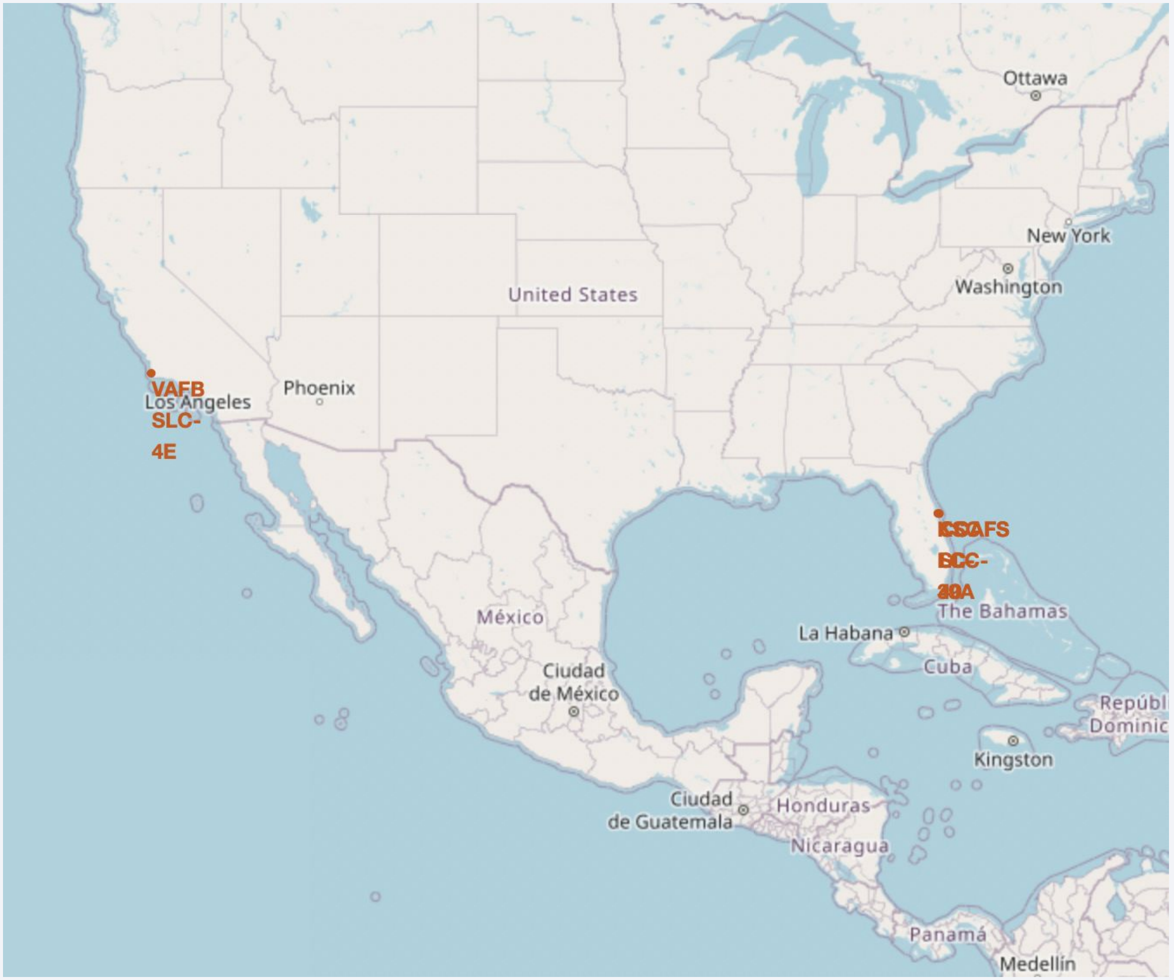
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

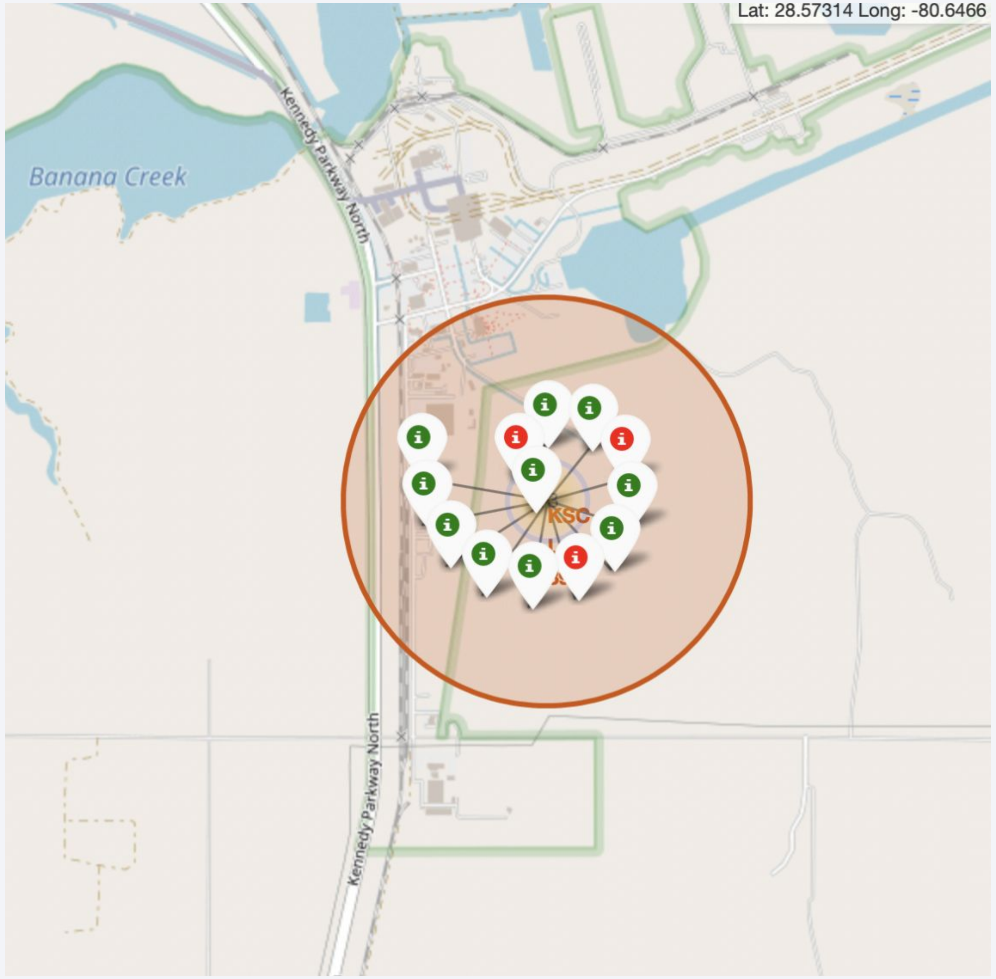
SpaceX Launch Sites

- **Proximity to the Equator**
  - Most launch sites are located near the equator, where Earth's rotation is fastest (1,670 km/h).
  - Launching from the equator provides an initial velocity advantage due to inertia, helping spacecraft achieve and maintain orbit efficiently.
- **Proximity to the Coast**
  - All launch sites are positioned close to the coast.
  - Launching rockets over the ocean reduces the risk of debris falling or explosions occurring near populated areas.



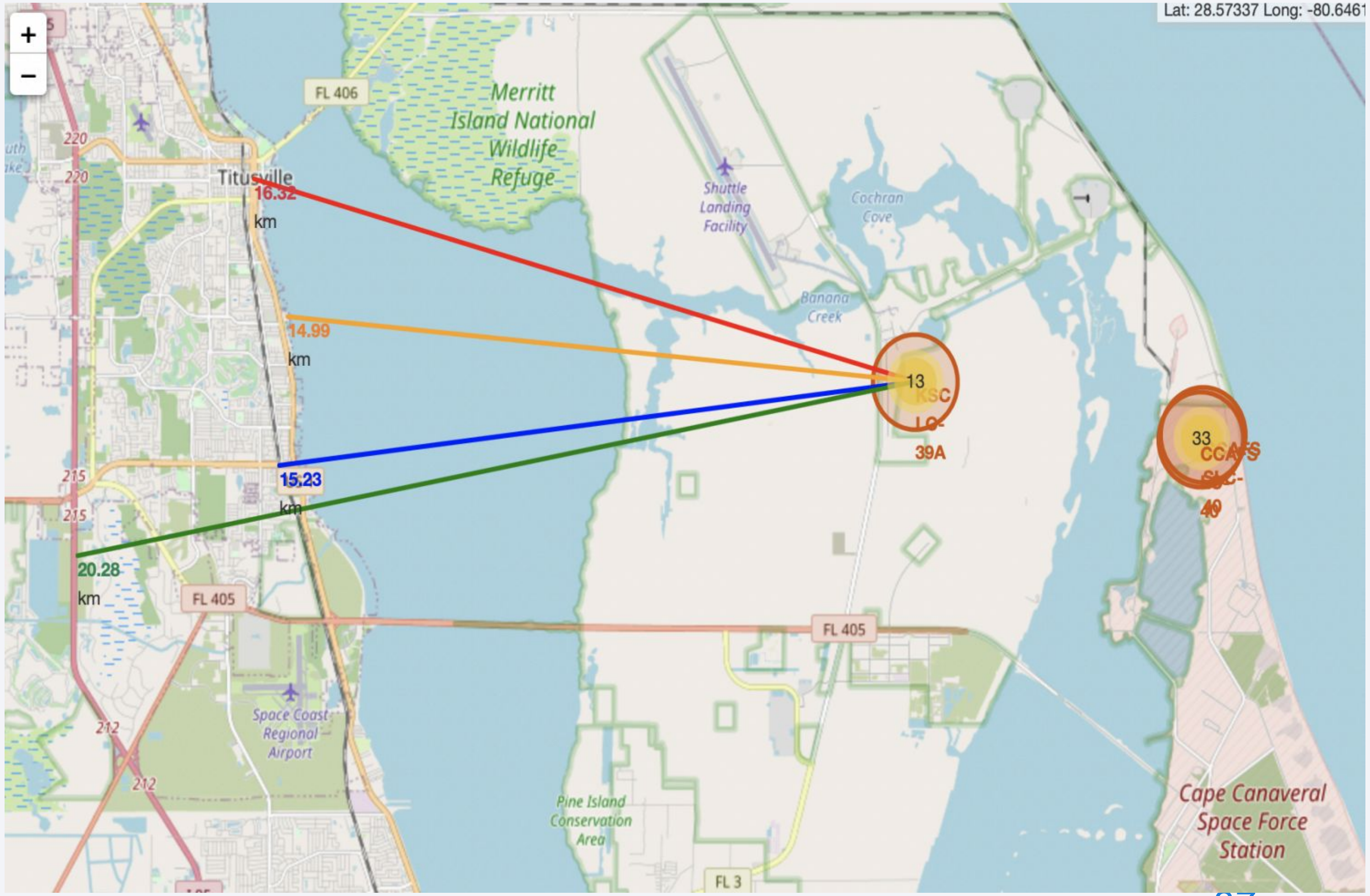
Launch Success Rate

- **Identifying Success Rates with Color-Labeled Markers**
  - Green Marker: Indicates a successful launch.
  - Red Marker: Indicates a failed launch.
  - These markers help visualize which launch sites have higher success rates.
- **High Success Rate at KSC LC-39A**
  - The launch site **KSC LC-39A** has a notably high success rate.



Proximity of Launch Site to Populous Areas

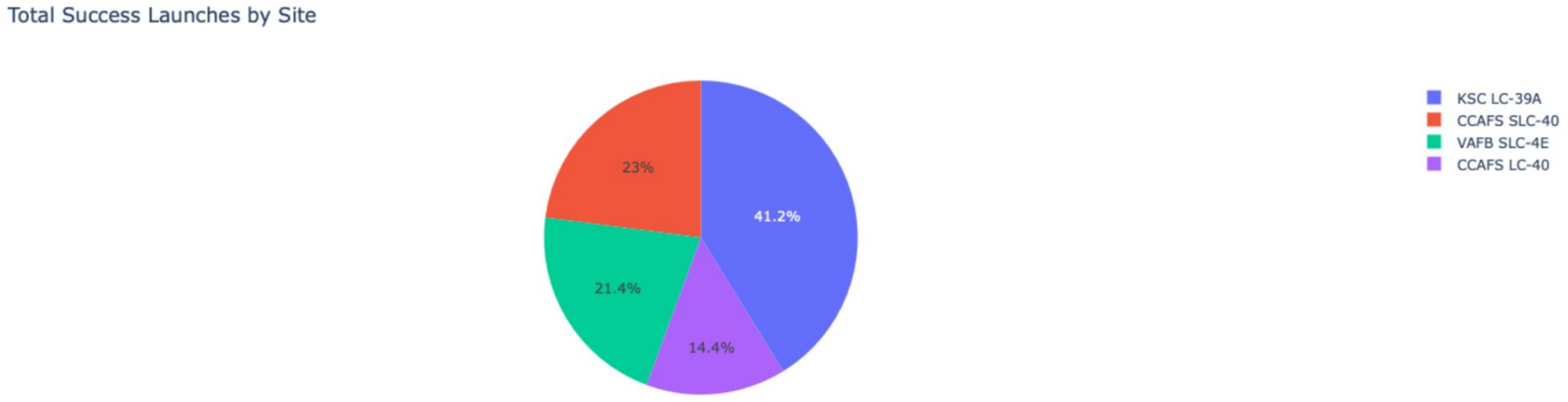
- **Proximity of Launch Site KSC LC-39A to Key Locations:**
  - Close to a railway: **15.23 km**
  - Close to a highway: **20.28 km**
  - Close to the coastline: **14.99 km**
  - Close to the nearest city, **Titusville**: **16.32 km**
- **Potential Risk of Failed Rockets:**
  - A failed rocket traveling at high speed can cover **15-20 km in seconds**, posing a potential danger to populated areas.



Section 4

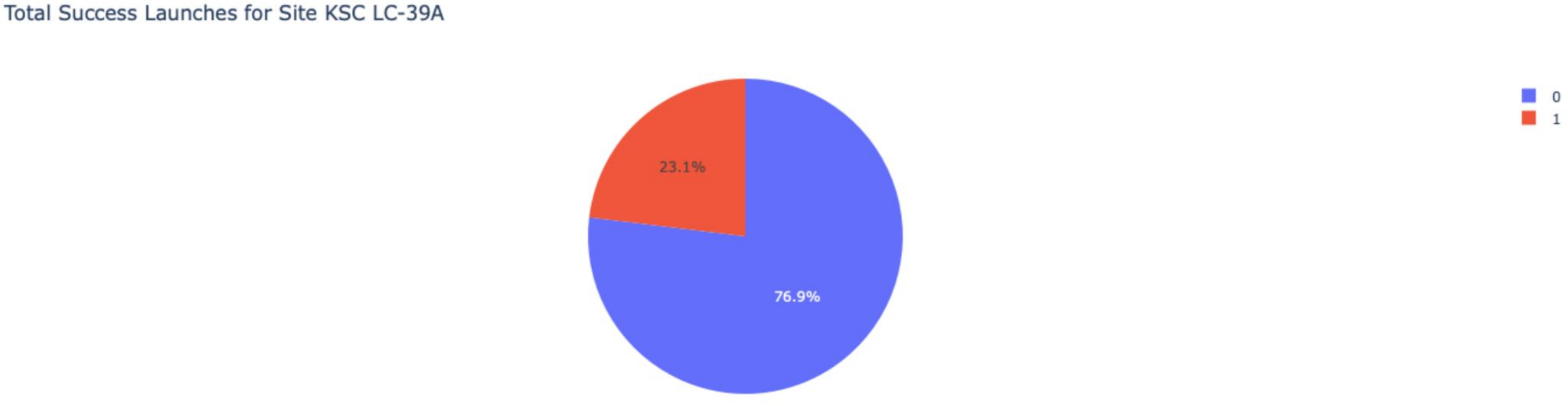
Build a Dashboard with Plotly Dash

Total Success Launches by Site



- This pie chart shows that the KSC LC-39A launch site has the highest number of successful launches.

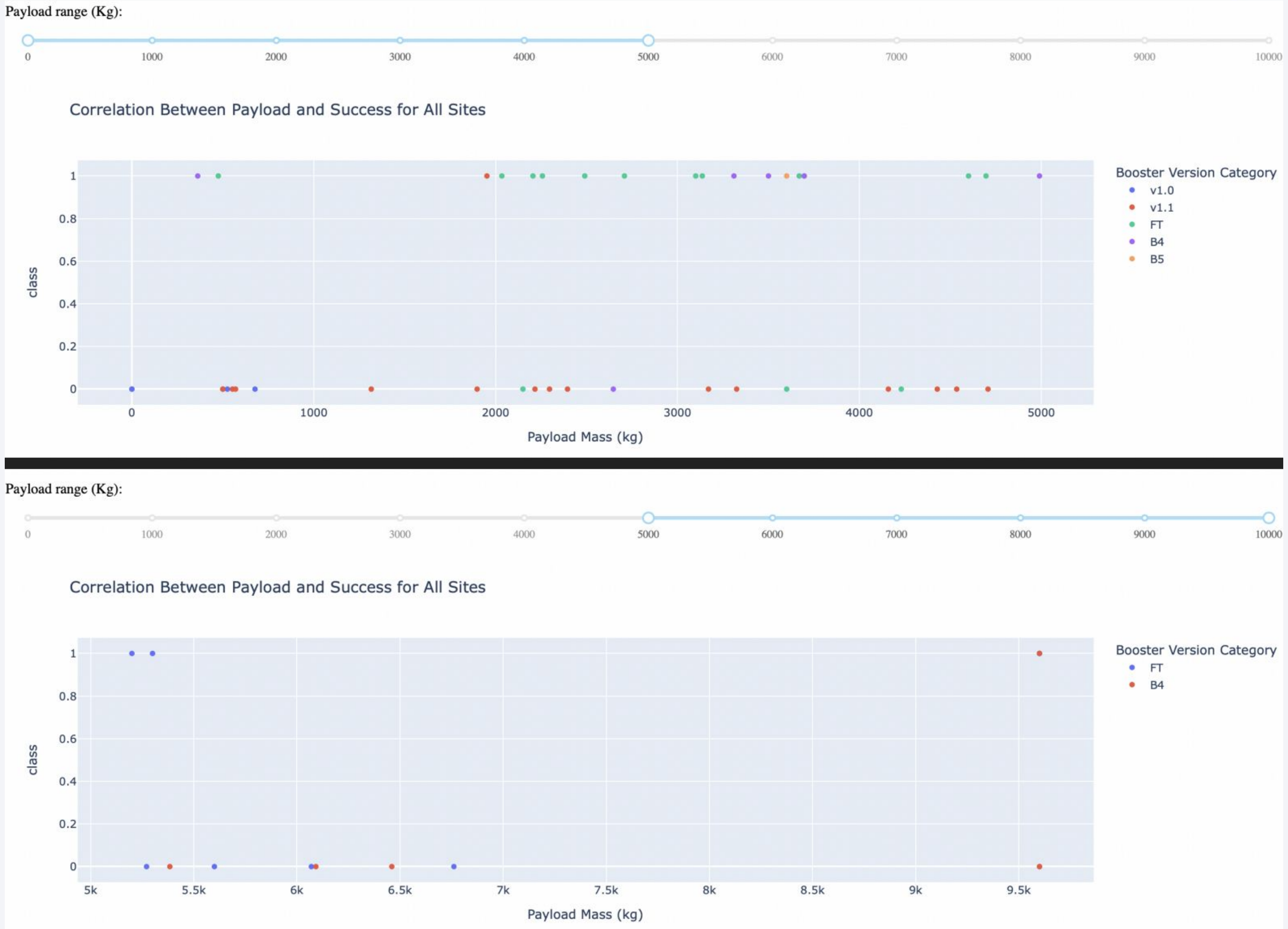
KSC LC-29A Launch Success



- This pie chart shows that 76.9% of launches from KSC LC-39A are successful.

Payload Mass vs Launch Success

- These plots show that launches with payloads between 2,000 KG and 5,500 KG are the most successful.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- **Uncertainty in Test Set Results:**
  - The test set scores do not clearly indicate which method performs best.
- **Impact of Small Test Sample Size:**
  - The test set contains only **18 samples**, which may lead to similar scores across methods.
  - To address this, all methods were tested using the **entire dataset**.
- **Decision Tree Model Performance:**
  - The **Decision Tree Model** achieved the **highest scores** and **best accuracy**, confirming it as the most effective model.

Scores from the test Dataset

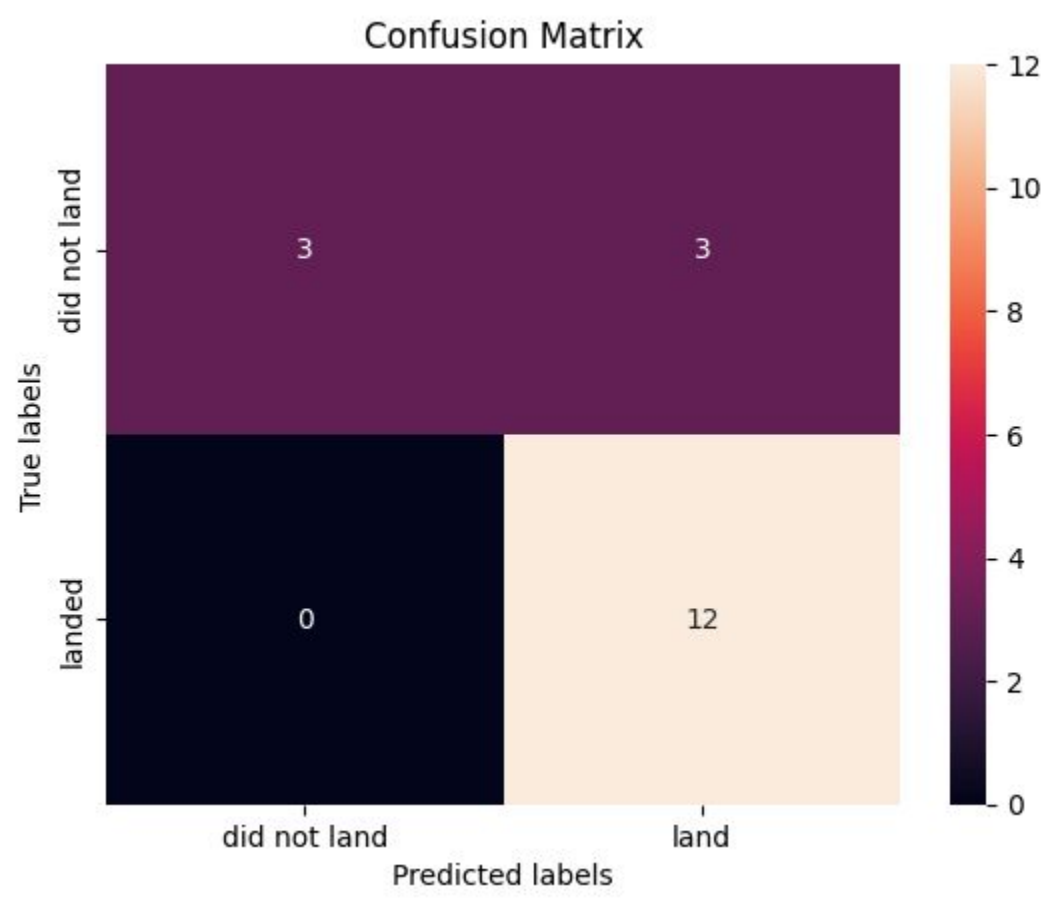
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.600000	0.800000
F1_Score	0.888889	0.888889	0.750000	0.888889
Accuracy	0.833333	0.833333	0.666667	0.833333

Scores from the whole Dataset

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.888889	0.819444
F1_Score	0.909091	0.916031	0.941176	0.900763
Accuracy	0.866667	0.877778	0.922222	0.855556

Confusion Matrix

- Explanation
    - Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- **Best Performing Model:**
  - The **Decision Tree Model** is the most effective algorithm for this dataset.
- **Influence of Payload Mass:**
  - Launches with **lower payload mass** tend to have higher success rates compared to those with **larger payload mass**.
- **Geographical Factors:**
  - Most launch sites are located **near the Equator**, and all are **close to the coast**.
- **Trends in Success Rate:**
  - The **success rate of launches has increased over the years**.
- **Top-Performing Launch Site:**
  - **KSC LC-39A** has the **highest success rate** among all launch sites.
- **Orbit Success Rates:**
  - **Orbits ES-L1, GEO, HEO, and SSO** have achieved a **100% success rate**.

Appendix

- GitHub Project
- Coursera Applied Data Science Capstone Course
- SpaceX.com

Thank you!