IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ana Beatriz Rodrigues Floriano dos Santos
21/09/2024

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

- Summary of methodologies
  - In this project I applied techniques such as API, web scraping for initial data collection, data wrangling for restructuring the data, EDA with visualization to get hidden insights, patterns and trends. Finally, I preprocessed the data using feature engineering before applying four different machine learning algorithms for predictive modeling to know whether the first stage of the Falcon 9 landing will be successful.

- Summary of all results
  - The project is about predicting if SpaceX's Falcon 9 rockets will land successfully in the first stage. SpaceX's provides it rockets at a much lower cost of 62 million dollars as compared to other providers which is 165 million dollars, this is because SpaceX's can reuse the first stage - IBM.

Introduction

- **Project background and context**
  - SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.
- **Problems you want to find answers**
  - How payload mass, launch site, number of flights, and orbits affect first-stage landing success
  - Rate of successful landings over time
  - Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology
With Rest API and Web Scrapping

Perform data wrangling
Using pandas, data was cleaned and transformed to be applied latter to the Machine Learning models

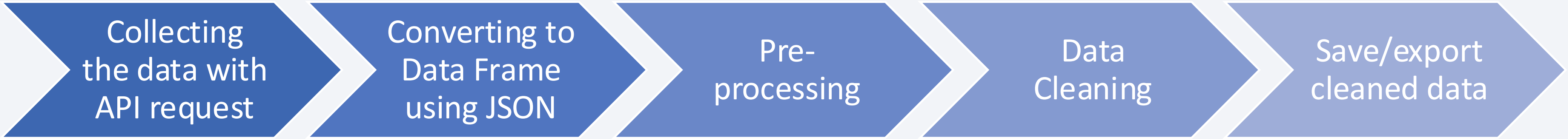
Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

Data Collection

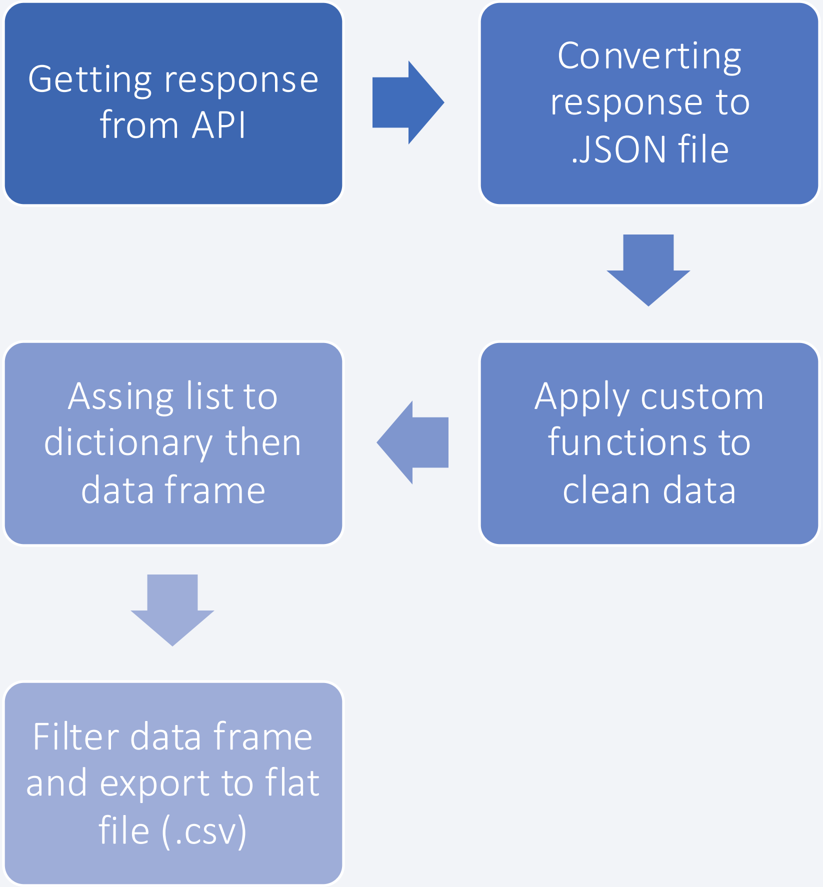
- Comprehensive research on SpaceX was conducted, and their publicly available API was utilized to retrieve data in JSON format using the HTTP Python library. The JSON data was then converted into a data frame to extract relevant information for analysis. Data related to 'Falcon 1' was filtered out, retaining only 'Falcon 9' data due to its successful landing history over recent years. Missing values were addressed during the data cleaning process, with the 'LaunchPad' field left null to indicate cases where no launch pad was used.



Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained
- The API was used according to the flowchart

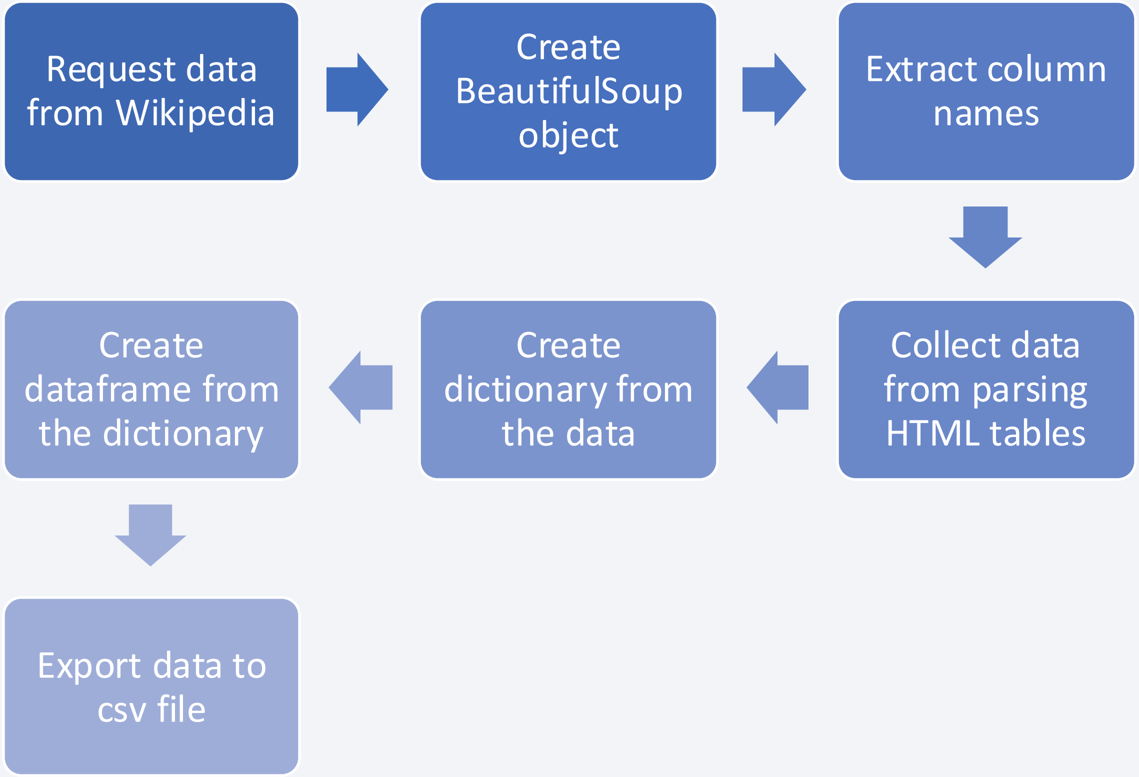
Collecting the data



Data Collection - Scraping

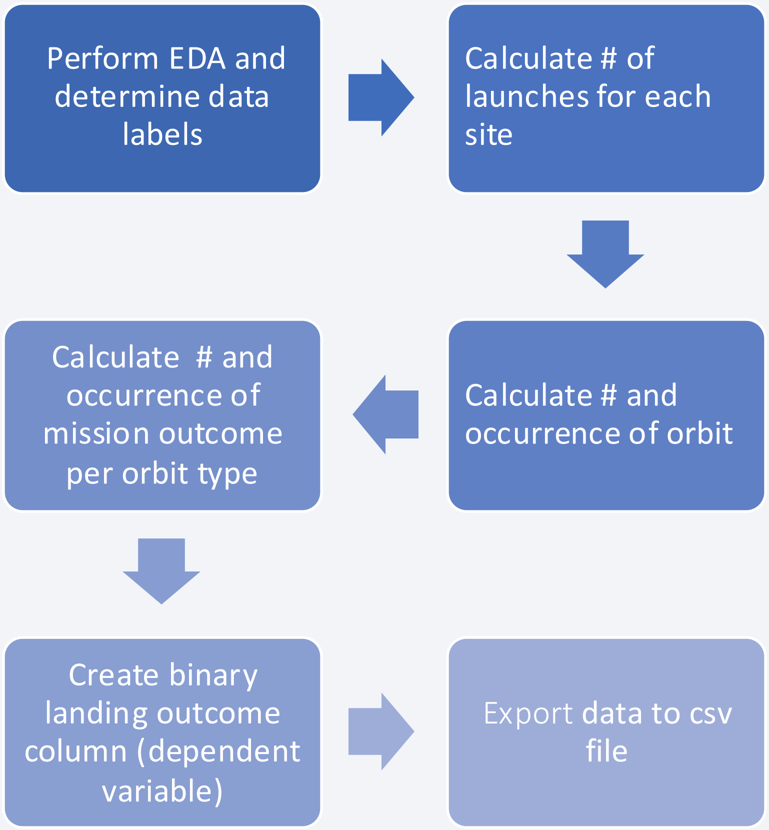
- Data from SpaceX launches can also be found on Wikipedia and were used according to the flowchart

[Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia]



Data Wrangling

- Initially some Exploratory Data Analysis was performed on the dataset
- Then the summuries launches per site, occurrences of each orbit and occurences of missions outcome per orbit type were calculated
- Finally, the Landing Outcome label was created from the Outcome Column

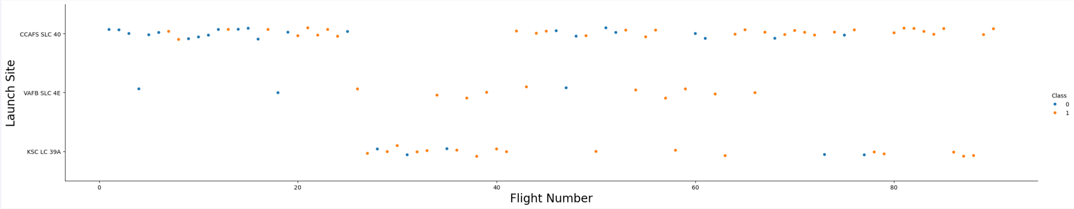


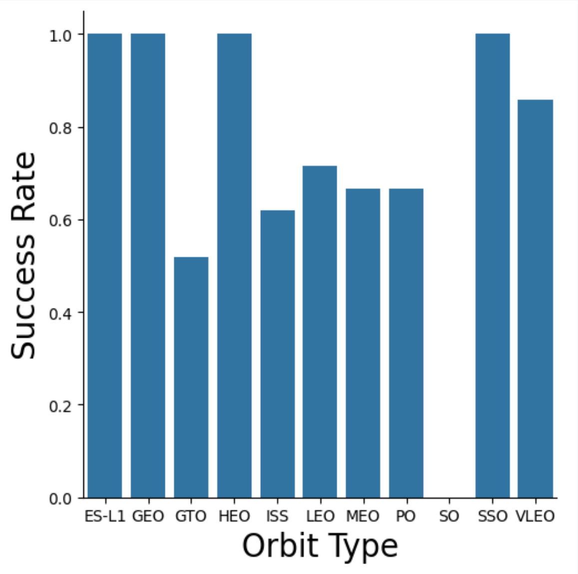
[Data wrangling]

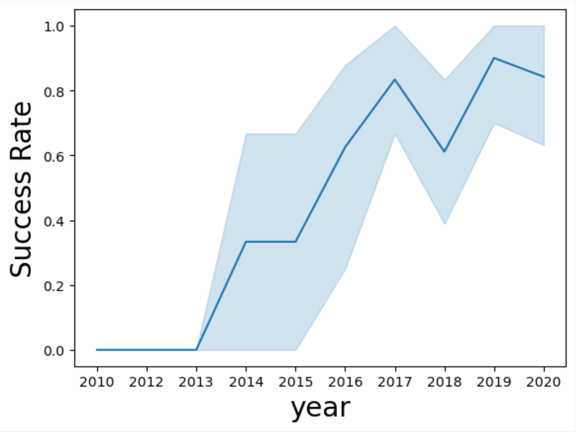
EDA with Data Visualization

- To explore the data, scatterplots, barplots and line charts were used to visualize the relationship between pair of features such as:
  - Flight Number vs. Payload
  - Flight Number vs. Launch Site
  - Payload Mass (kg) vs. Launch Site
  - Payload Mass (kg) vs. Orbit type

[EDA with Visualization]







EDA with SQL

- Display:
  - Names of unique launch sites
  - 5 records where launch site begins with 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1.
- List:
  - Date of first successful landing on ground pad
  - Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
  - Total number of successful and failed missions
  - Names of booster versions which have carried the max payload
  - Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
  - Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

[EDA with SQL]()

Build an Interactive Map with Folium

- **Markers Indicating Launch Sites**
  - Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
  - Added red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates
- **Colored Markers of Launch Outcomes**
  - Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates
- **Distances Between a Launch Site to Proximities**
  - Added colored lines to show distance between launch site CCAFS SLC40 and its proximity to the nearest coastline, railway, highway, and city

Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

- Dropdown List with Launch Sites
  - Allow user to select all launch sites or a certain launch site
- Slider of Payload Mass Range
  - Allow user to select payload mass range
- Pie Chart Showing Successful Launches
  - Allow user to see successful and unsuccessful launches as a percent of the total
- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version
  - Allow user to see the correlation between Payload and Launch Success

[Interactive Dashboard with Ploty Dash]

Predictive Analysis (Classification)

- The cleaned data was imported, with features assigned to the variable 'X' and the target to 'Y'. The features were scaled using sklearn's StandardScaler object and split using the built-in train_test_split function, with 20% of the data allocated for testing.
- Models were instantiated, and hyperparameters were tuned using GridSearchCV to identify the optimal performance parameters. The models were fitted on the training set and evaluated on the unseen test set. The DecisionTreeClassifier model achieved the best results, with an accuracy of 88.88% and an F1-score of 88.21%, utilizing GridSearchCV for hyperparameter optimization.

[Machine Learning Prediction]()

Import data → Assign (features and Target) → Feature engineered → Split (train-test) → Model Evaluate

Results

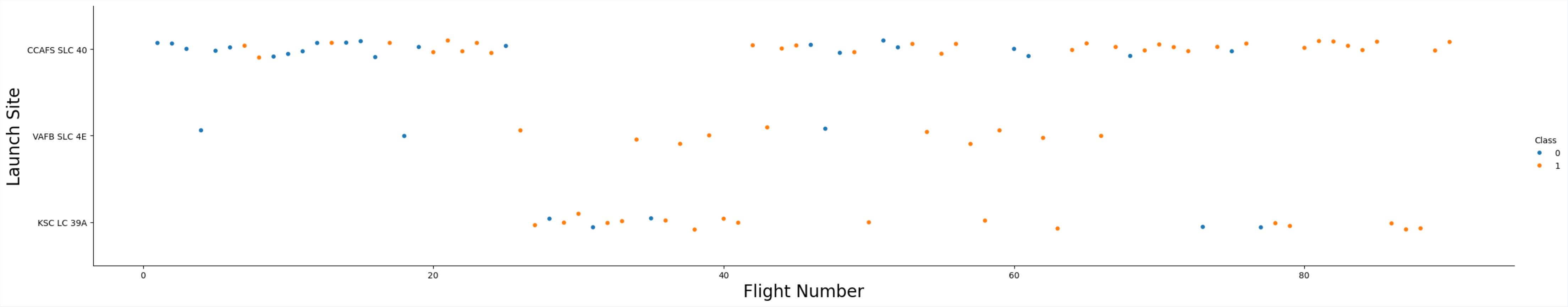
- Exploratory Data Analysis
  - Launch success has improved over time
  - KSC LC-39A has the highest success rate among landing sites
  - Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- Visual Analytics
  - Most launch sites are near the equator, and all are close to the coast
  - Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities
- Predictive Analytics
  - All nut the Decision Tree model showed equal predictive model for the dataset

Section 2

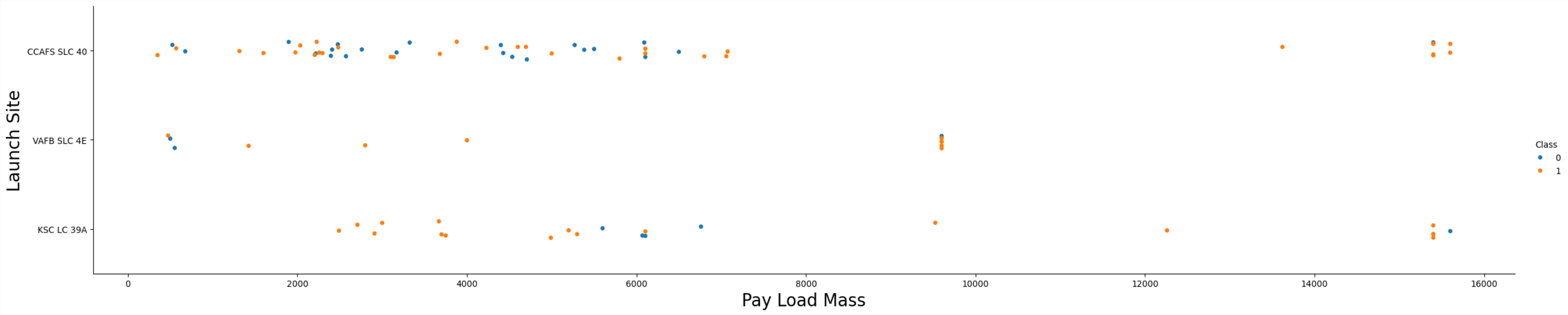
Insights drawn from EDA

Flight Number vs. Launch Site

As shown below the relationship between Flight Number and LaunchSite, we can see that with lower flight number (20) there's no failure or success metric for "KSC LC 39A", but two failures for the "VAFB SLC 4E". But as the flight number increases past 80, there are more success "CCAFS SLC 40" and none for "VAFB SLC 4E".



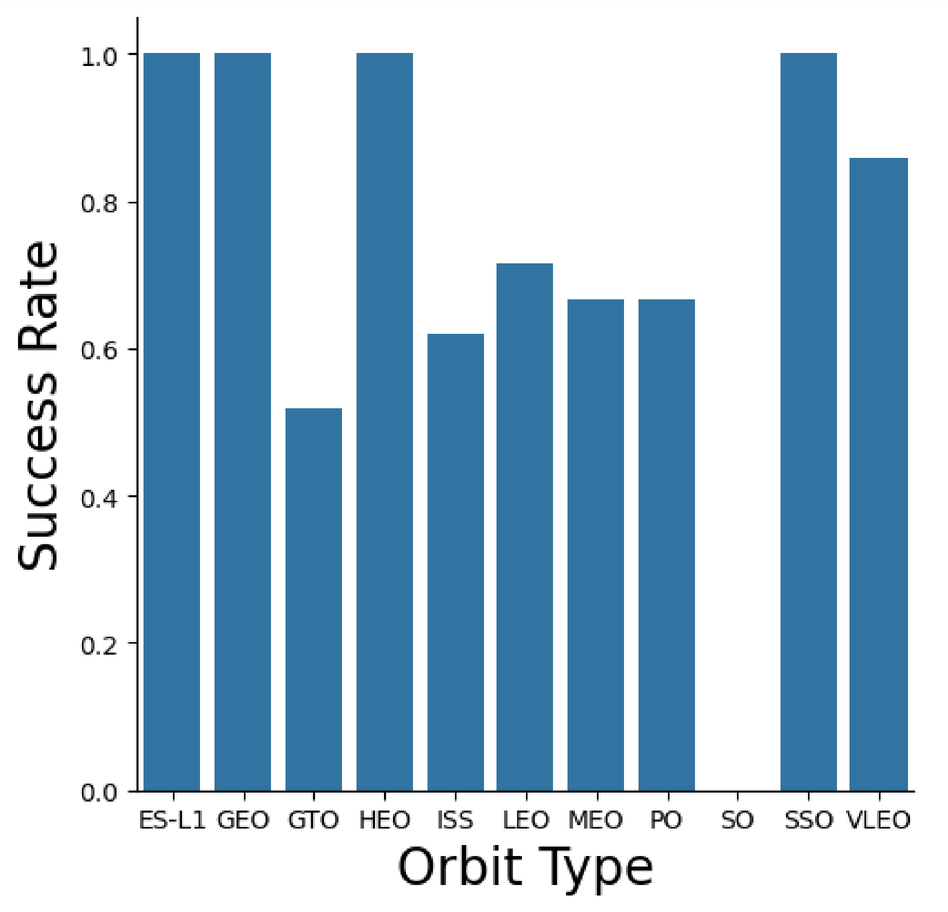
Payload vs. Launch Site



As shown above the relationship between Flight Number and LaunchSite, we can see that with lower flight number (20) there's no failure or success metric for "KSC LC 39A", but two failures for the "VAFB SLC 4E" and more failures than success for "CCAFS SLC 40". But as the flight number increases past 80, there are more success "CCAFS SLC 40" and none for "VAFB SLC 4E".

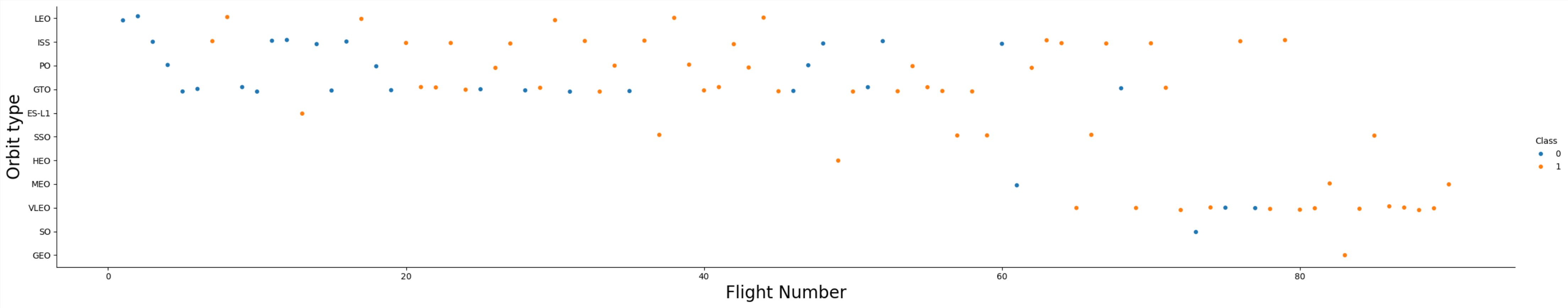
Success Rate vs. Orbit Type

- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



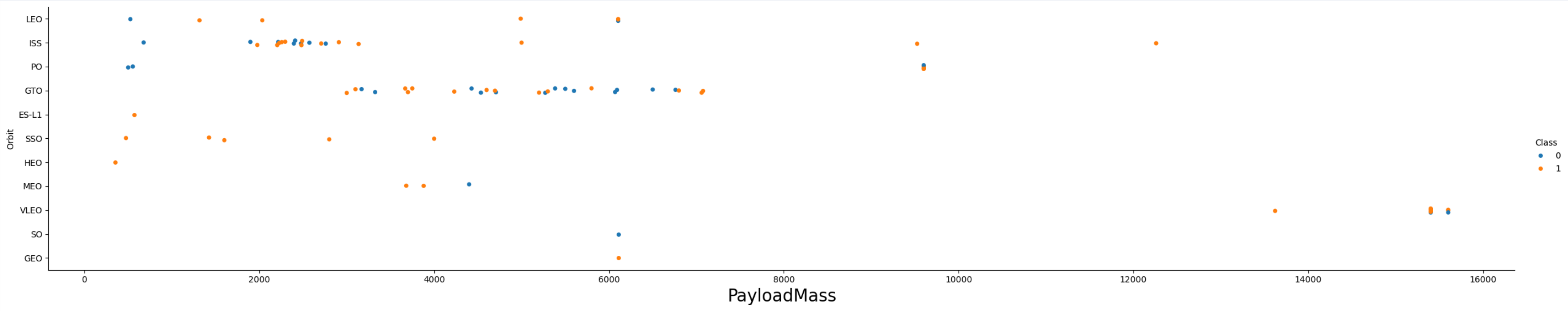
Flight Number vs. Orbit Type

You can see that in the VLEO orbit there seem to be success rate only when the flight number is greater than 60. While there seem to be no relationship in the GTO.



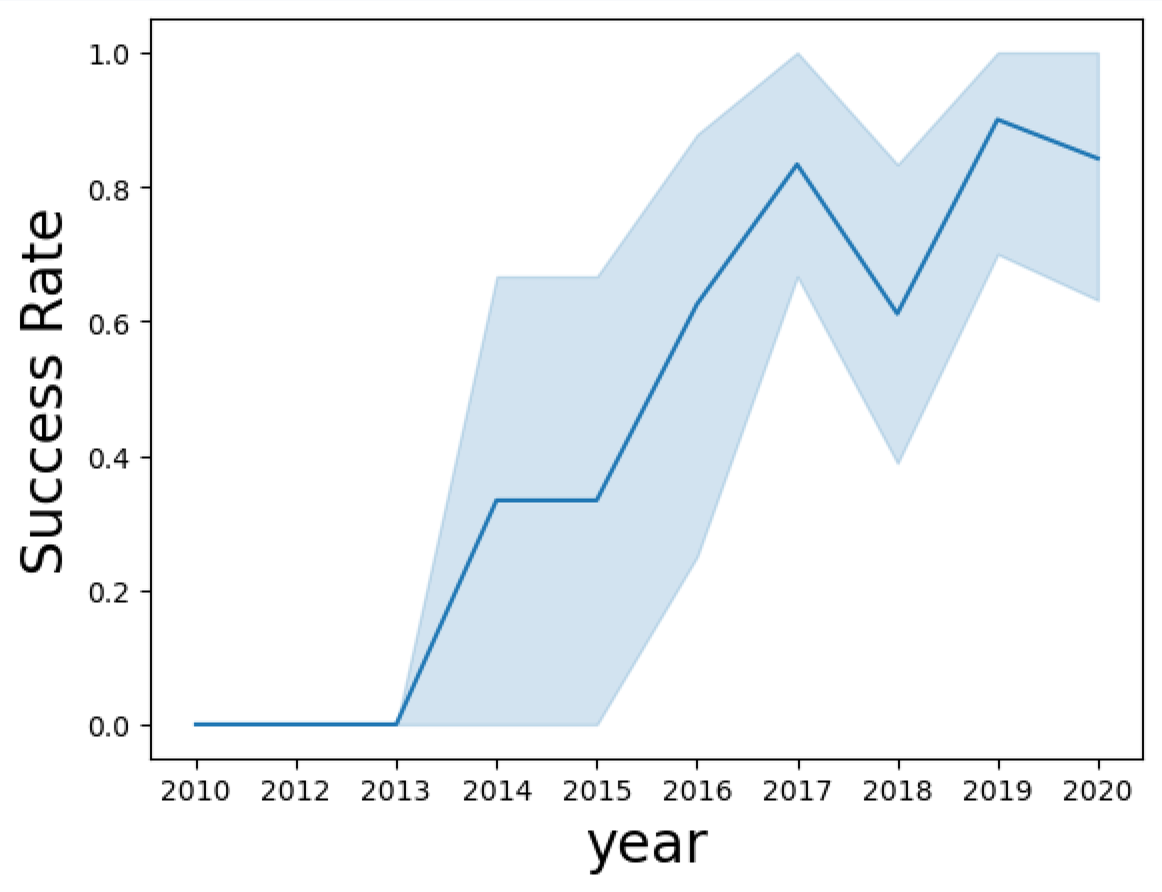
Payload vs. Orbit Type

There is more positive landing rate with heavy payloads for Polar, LEO and ISS, however for GTO there’s the presence of both positive and negative landing.



Launch Success Yearly Trend

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



All Launch Site Names

- **CCAFS LC-40**: Space Launch Complex 40 formerly Launch Complex 40 (LC-40) is an orbital launch pad located in northern Cape Canaveral , Florida. (Wikipedia)
- **VAFB SLC-4E**: Vandenberg AFB Space Launch Complex 4 is a launch and landing site at Vandenberg Space Force Base, California, U.S. It has two pads, both of which are used by SpaceX for Falcon 9, one for launch operations, and other as Landing Zone 4 for SpaceX landings. (Wikipedia)
- **KSC LC-39A**: Kennedy Space Center Launch Complex 39A Launch Complex 39A is the first of Launch Complex 39's three launch pads, located at NASA's Kennedy Space Center in Merritt Island, Florida. (Wikipedia)
- **CCAFS SLC-40**: Space Launch Complex 40 (SLC-40), sometimes referred to as "Slick Forty," is a launch pad located at Cape Canaveral Space Force Station in Florida. Initially opened as Launch Complex 40 (LC-40) and used by the United States Air Force for 55 launches of rockets from the Titan family between 1965 and 2005. In 2007, SpaceX acquired a lease for SLC-40 and has since transformed the complex into a high-volume launch site for the Falcon 9 rocket. (Wikipedia)

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

Launch Site Names Begin with 'CCA'

- Here we can see five samples of Cape Canaveral launches

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

Total Payload Mass

- Total Payload Mass
  - 45,596 kg (total) carried by boosters launched by NASA (CRS)

Customer	Total_NASA_CRS_mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Average Payload Mass
  - 2,928 kg (average) carried by booster version F9 v1.1 A

Booster_Version	avg_Booster_versionF9_v1_1
F9 v1.1	2928.4

First Successful Ground Landing Date

- 1st Successful Landing in Ground Pad
  - 12/22/2015

Mission_Outcome	Date_First_Succ_Land
Success	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Booster mass greater than 4,000 but less than 6,000
  - F9 FT B1022 -- 4696KG
  - F9 FT B1026 -- 4600KG
  - F9 FT B1021.2 -- 5300KG
  - F9 FT B1031.2 -- 5200KG

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

Mission_Outcome	Total (Success or failure)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- 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

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

2015 Launch Records

- F9 v1.1 B1012
  - 10/01/2015
- F9 v1.1 B1015
  - 14/04/2015

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

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

- Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

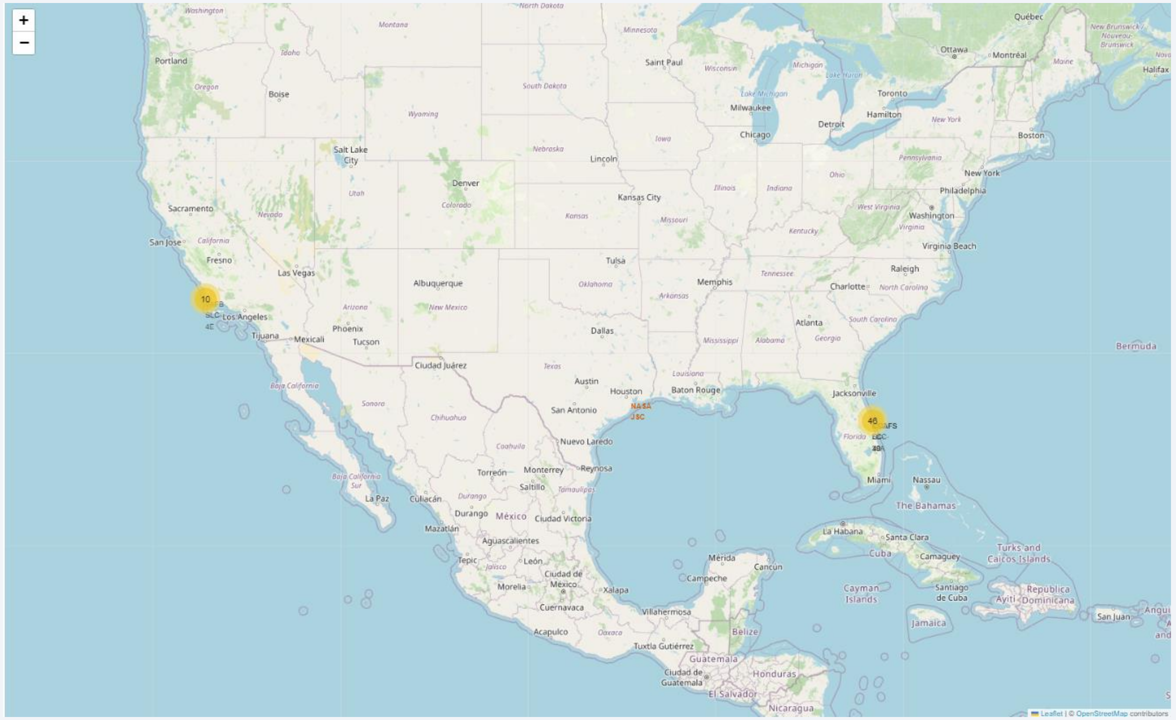
Landing_Outcome	Count_Outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

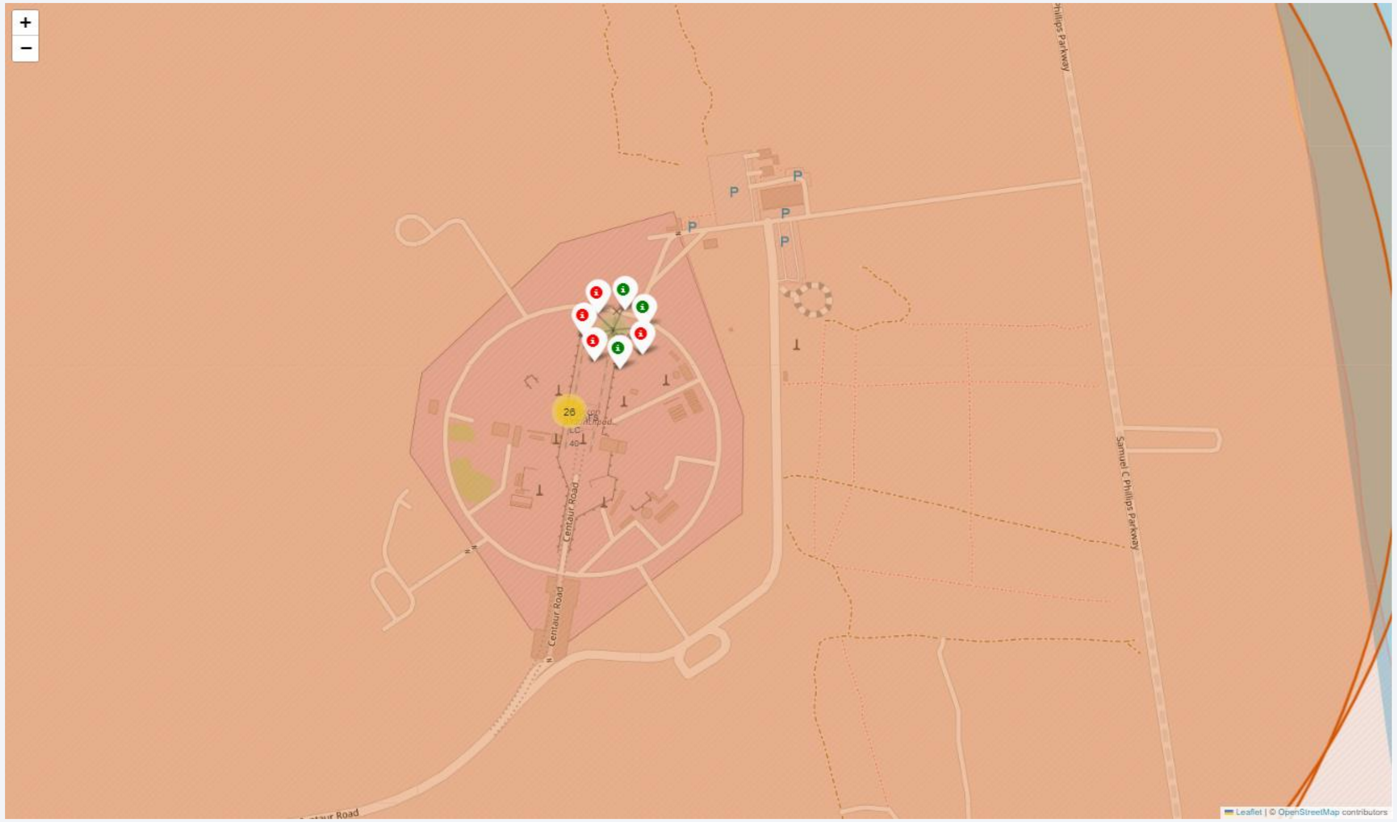
Launch Sites

- Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.



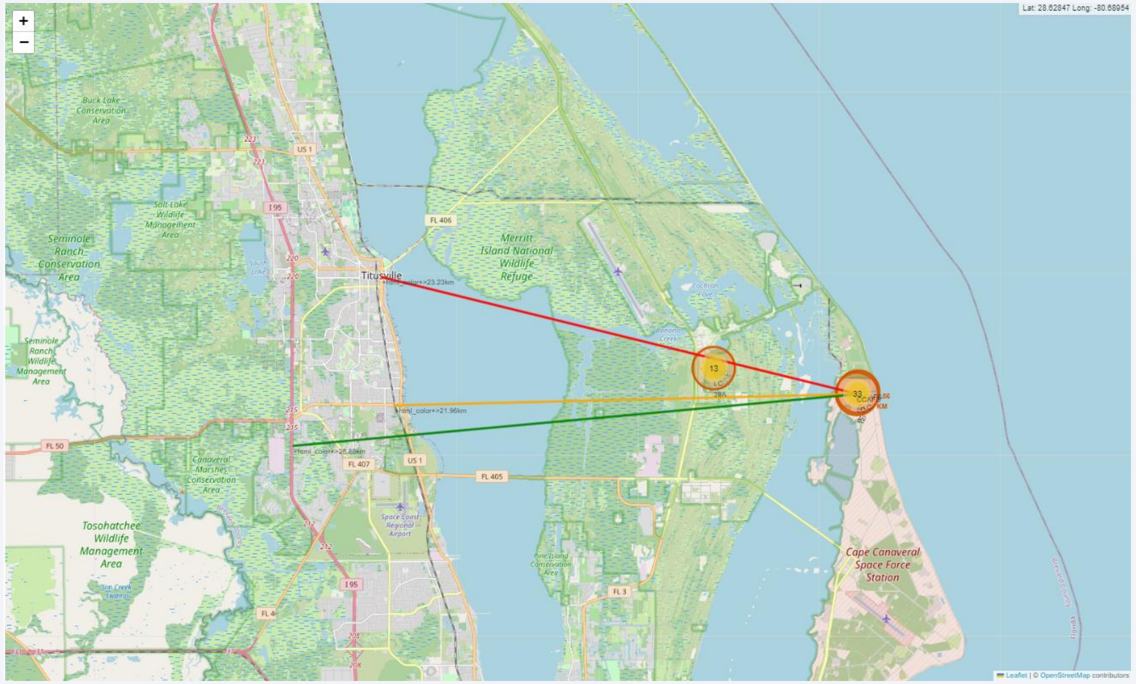
Launch Outcomes

- At Each Launch Site
- **Outcomes**:
  - Green markers for successful launches
  - Red markers for unsuccessful launches
  - Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)



Distance to Proximities

- CCAFS SLC-40
  - .86 km from nearest coastline
  - 21.96 km from nearest railway
  - 23.23 km from nearest city
  - 26.88 km from nearest highway



- **Coasts:** help ensure that spent stages dropped along the launch path or failed launches don't fall on people or property.
- **Safety / Security:** needs to be an exclusion zone around the launch site to keep unauthorized people away and keep people safe.
- **Transportation/Infrastructure and Cities:** need to be away from anything a failed launch can damage, but still close enough to roads/rails/docks to be able to bring people and material to or from it in support of launch activities.

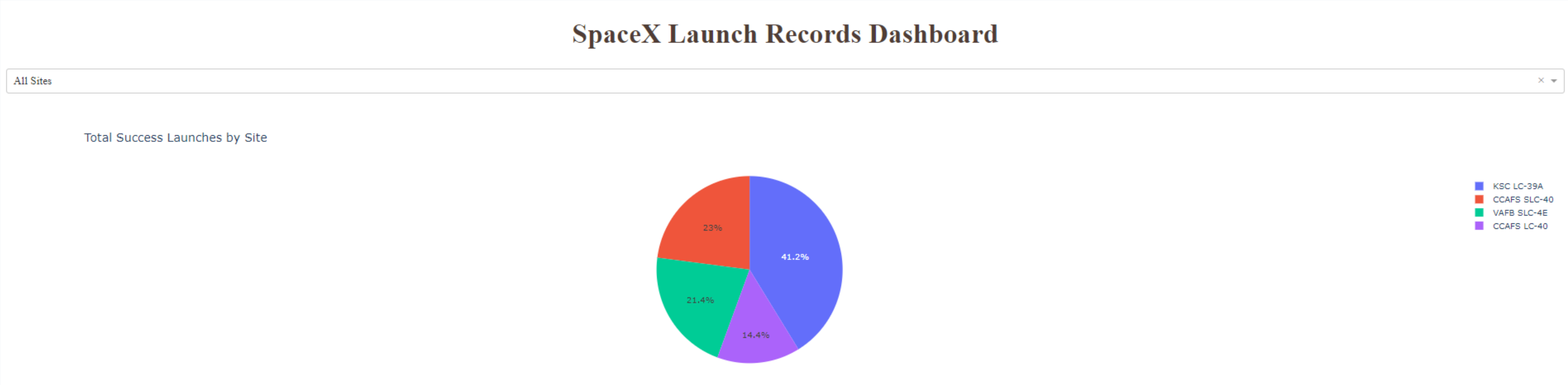
Section 4

Build a Dashboard with Plotly Dash

Launch Success by Site

Success as Percent of Total

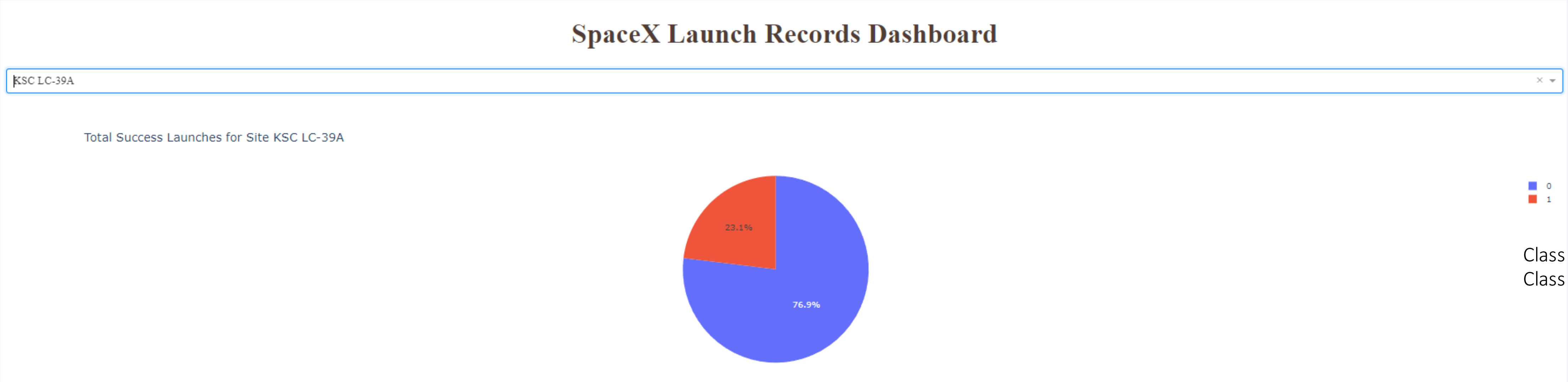
- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



Launch Success (KSC LC-29A)

Success as Percent of Total

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches

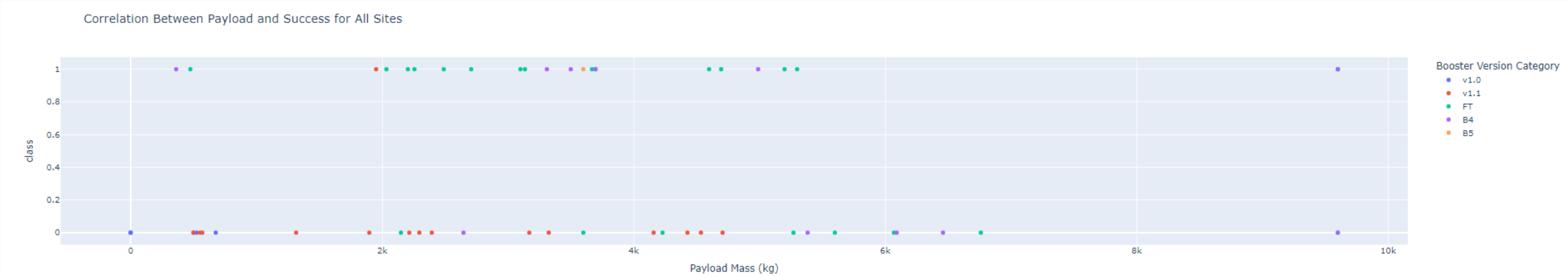


Class 0 = Fail
Class 1 = Success

Payload Mass and Success

By Booster Version

- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome

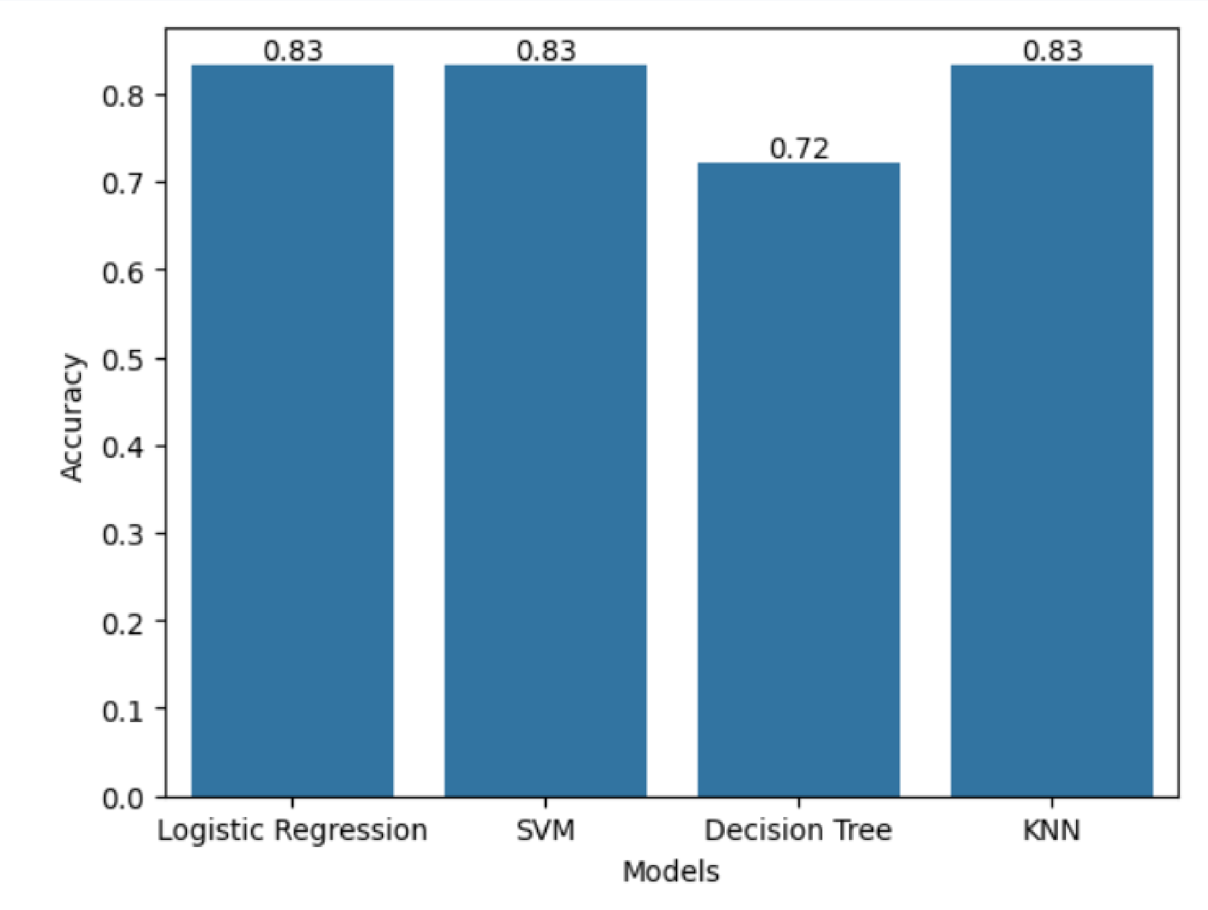


Section 5

Predictive Analysis (Classification)

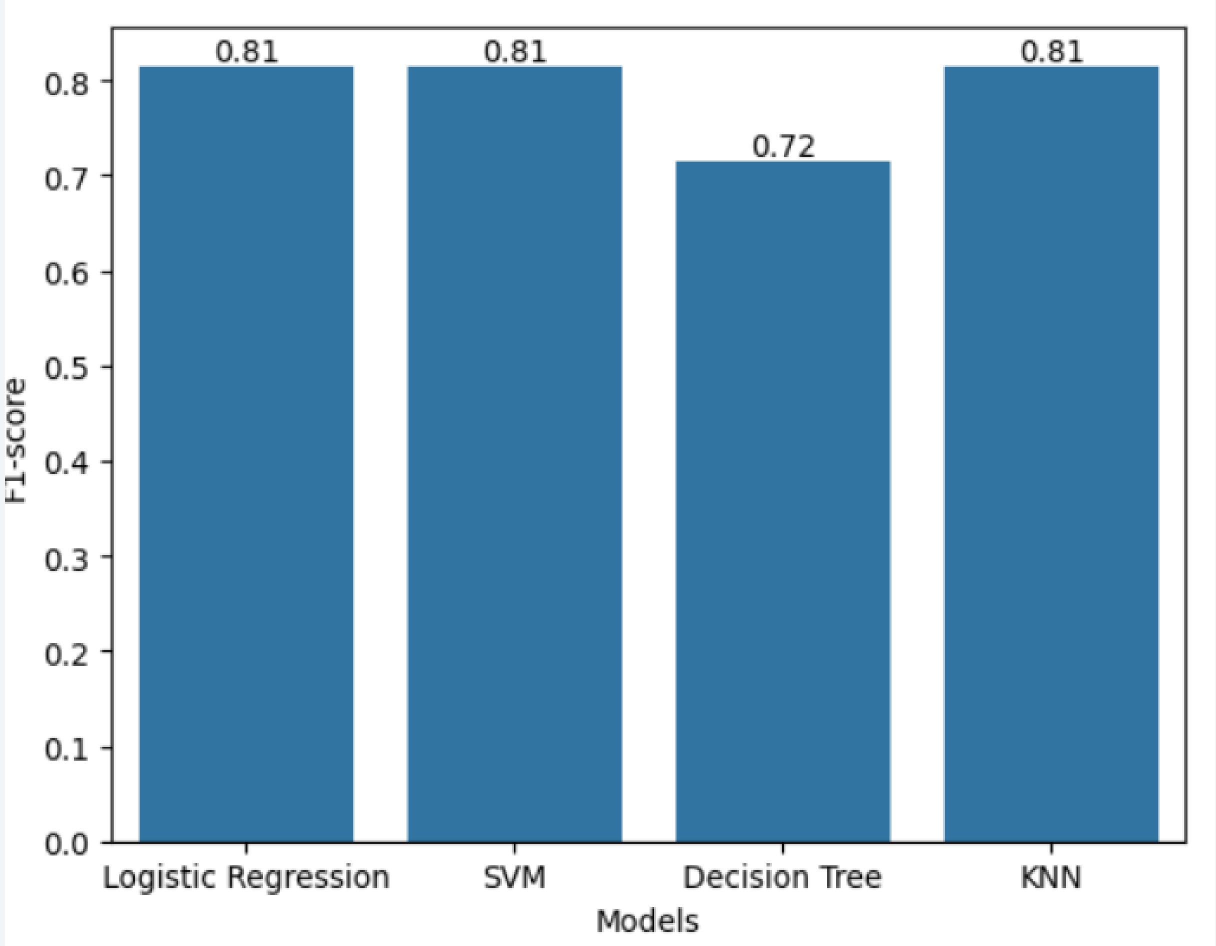
Classification Accuracy

- All but Decision Tree had the same Accuracy (0.83)



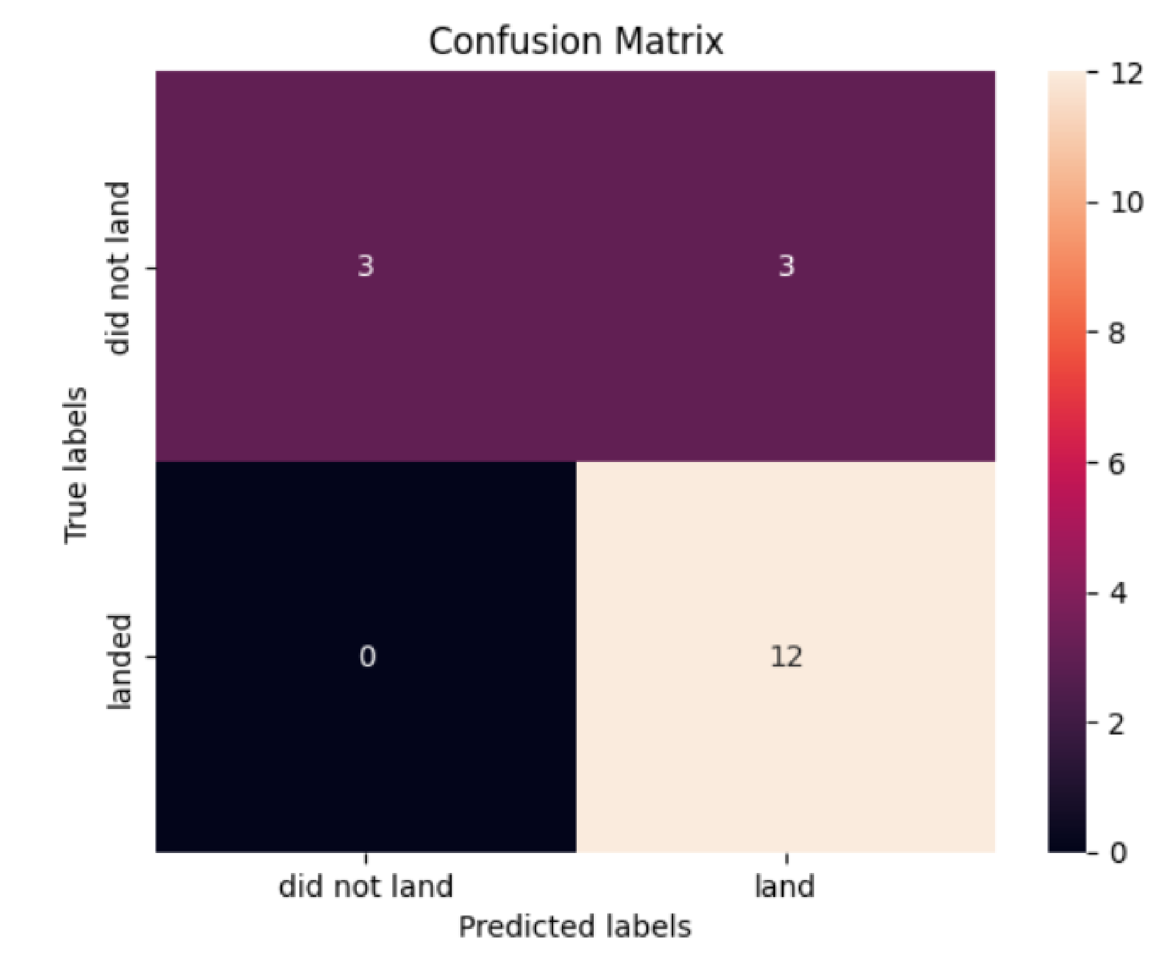
Classification F1-Score

- All but Decision Tree had the same F1-score (0.81)



Confusion Matrix

- All but the Decision tree model has 3 false-positive and no false negative and achieved a precision accuracy of 80.00% while recall is 100%.



Conclusions

- **Model Performance:** The models performed similarly on the test set with the decision tree model slightly underperforming
- **Equator:** Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast:** All the launch sites are close to the coast
- **Launch Success:** Increases over time
- **KSC LC-39A:** Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits:** ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass:** Across all launch sites, the higher the payload mass (kg), the higher the success rate

Appendix

All relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, and data sets included in this presentation can be found on my GitHub.

Thank you!