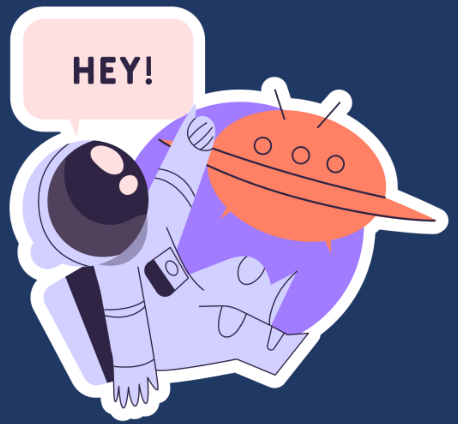


Winning Space Race with Data Science

David Adityo Marsono

3 November 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using SpaceX API and Web Scraping from Wikipedia;
  - Data Wrangling using Pandas and Numpy.
  - Exploratory Data Analysis (EDA) such as: Data Visualization (Seaborn & Matplotlib), Interactive Visual Analytics (Folium & Plotly).
  - Machine Learning Predictions (Logistics Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors).
- Summary of all results
  - Launch success has improved over time.
  - KSC LC 39A as launch site has the highest success rate than other sites.
  - All models of machine learning performed similarly on the test set, Decision Tree Classifier slightly outperformed.cc

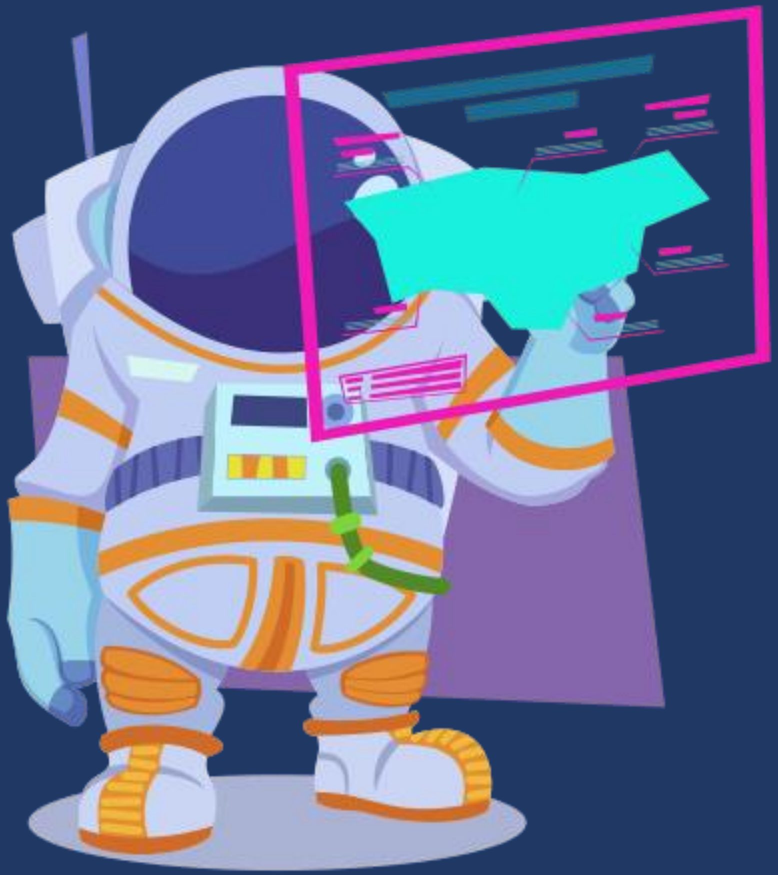
Introduction

Project background

The commercial space age has begun, many companies are making space travel affordable. SpaceX become a leader in the space industry. SpaceX advertise Falcon 9 rocket launches on its website with a cost of 62 million dollars when the others cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. 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 our company (SpaceY) can reuse the first stage.

Focus

- How independent variables (payload mass, launch site, number of flights, and orbits) affect first-stage landing success
- Rate of landings success over time.
- Best machine learning model (binary classification) to predict landing success



Methodology

- Data Collection using SpaceX REST API and web scraping from Wikipedia.
- Data Wrangling – filtering the data, dealing with missing values, and applying one hot encoding to prepare the data for analysis and modeling.
- Exploratory Data Analysis with Seaborn library and SQL.
- Data Visualization with Folium for geographic visualization and Plotly Dash for Interactive Dashboard.
- Machine Learning Model using classification models for predict landing outcomes. Tune and evaluate models to find best model and parameters.

Data Collection

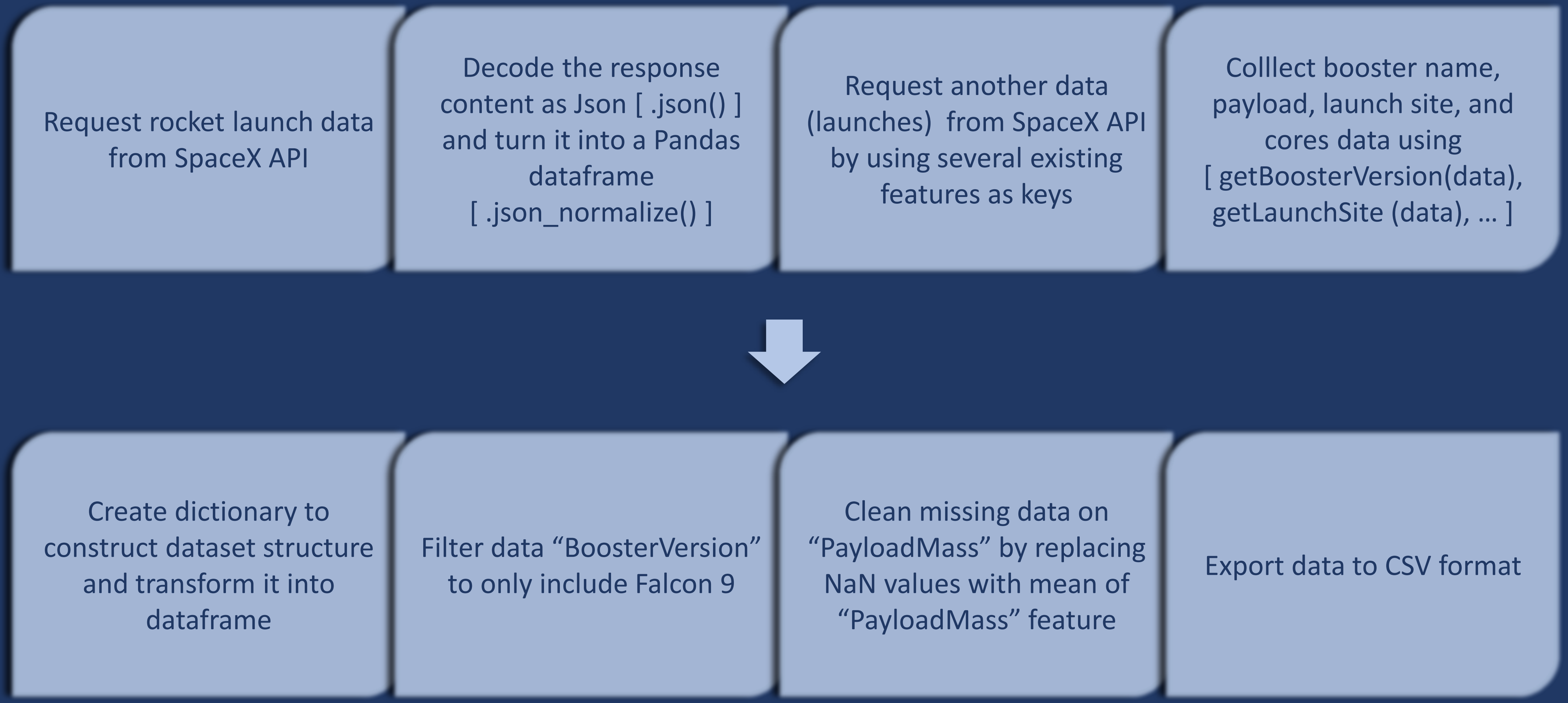
- Data collected from SpaceX REST API:

Date, BoosterVersion, PayloadMass, Orbit, Launchsite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longtitude, Latitude.

- Data collected from Wikipedia Web Scraping:

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

Data Collection – SpaceX REST API



Request rocket launch data from SpaceX API

Decode the response content as Json [.json()] and turn it into a Pandas dataframe [.json_normalize()]

Request another data (launches) from SpaceX API by using several existing features as keys

Colllect booster name, payload, launch site, and cores data using [getBoosterVersion(data), getLaunchSite (data), …]

Create dictionary to construct dataset structure and transform it into dataframe

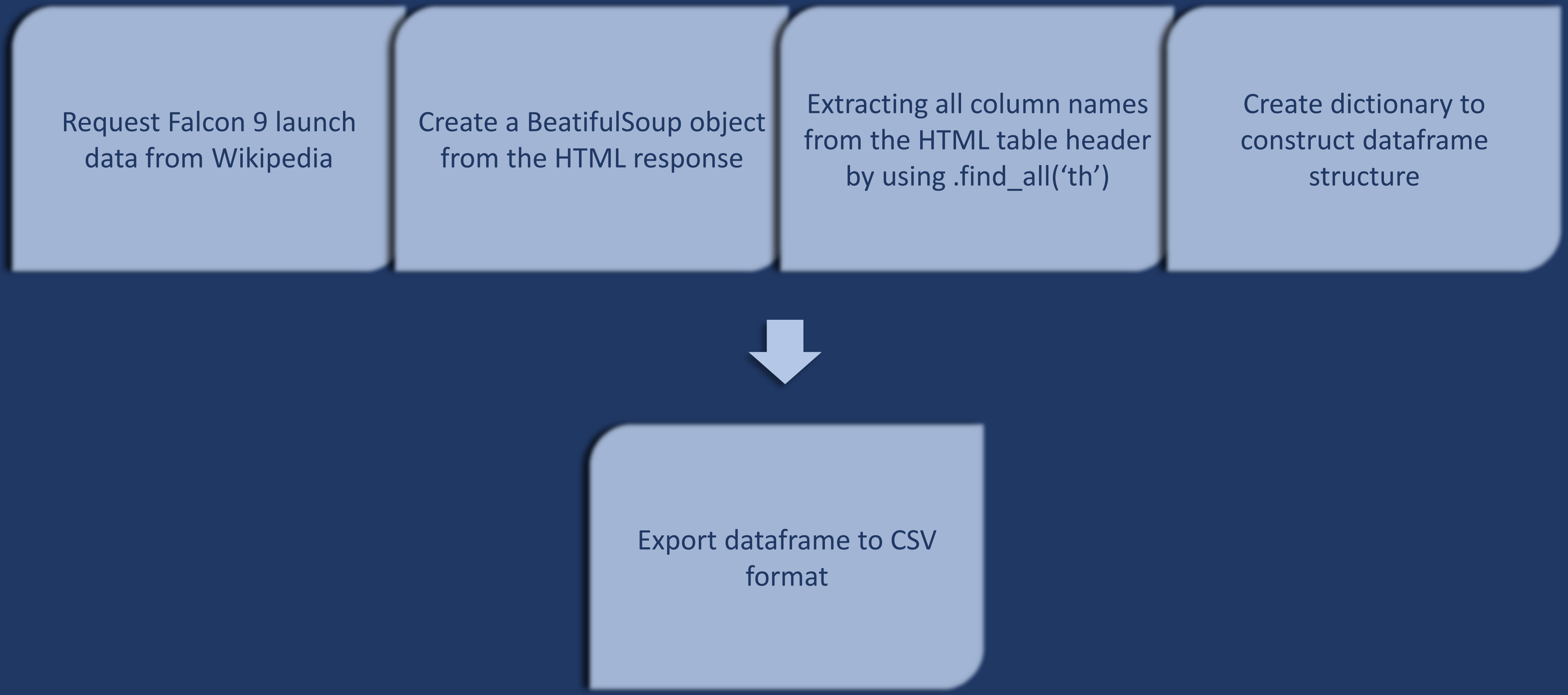
Filter data “BoosterVersion” to only include Falcon 9

Clean missing data on “PayloadMass” by replacing NaN values with mean of “PayloadMass” feature

Export data to CSV format

Github Url : Data Colletion API

Data Collection – Web Scraping



Request Falcon 9 launch data from Wikipedia

Create a BeatifulSoup object from the HTML response

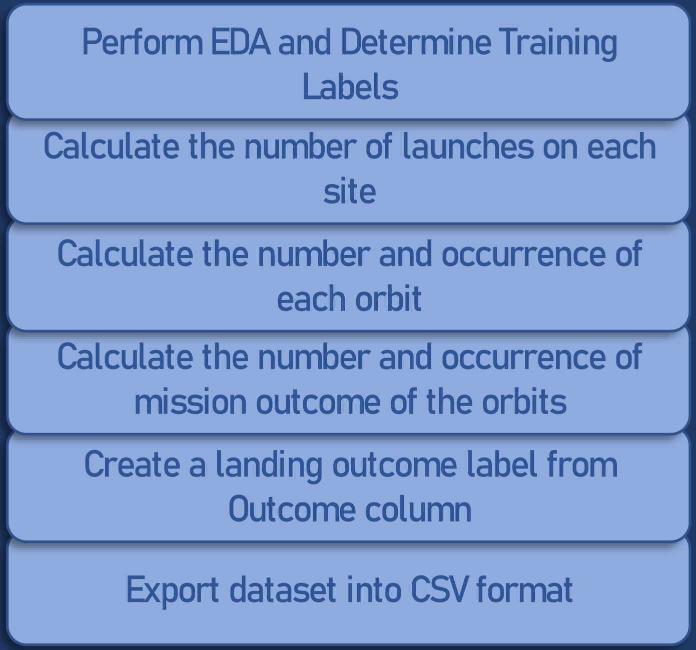
Extracting all column names from the HTML table header by using .find_all(‘th’)

Create dictionary to construct dataframe structure

Export dataframe to CSV format

Github Url : Data Collection Web Scrap

Data Wrangling



Perform EDA and Determine Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome of the orbits

Create a landing outcome label from Outcome column

Export dataset into CSV format

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad, True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. We mainly convert those outcomes into Training Labels with `1` means the booster successfully landed `0` means it was unsuccessful.

Github Url : Data Wrangling

EDA with Data Visualization

Charts created:

- Flight Number Vs Launch Site
- Payload Mass Vs Launch Site
- Success rate of each Orbit Type
- Flight Number Vs Orbit Type
- Payload Mass Vs Orbit Type
- Launch Success Yearly Trend

Explanation:

- Using Scatter Plot to view the relationship between variables. The variables could be useful for machine learning models if a relationship exists.
- Using Bar Chart to view comparisons among discrete categories. Bar Chart also show the relationship between the specific categories being compared and the value of categories are measured
- Using Line Chart to view data trends over time (time series)

Github Url : EDA with Seaborn

EDA with SQL

SQL queries:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters that have success in drone ship and have payload mass greater that 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions that have carried the maximum payload mass
- List the records that display the months names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in 2015
- Rank the count landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

Github Url : EDA with SQL

Interactive Map with Folium

Markers Indicating Launch Sites:

- Added Markers with Circle, Popup Label and Text label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Colored Markers of Launch Outcomes:

- Added colored markers of successful (green) and unsuccessful (red) launcehs at each launch sites to show which sites have high success rates.

Distances Between a Launch Site to Proximities:

- Added colored lines to show distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway highway, and city.

Github Url : Interactive Map Folium

Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site

Pie Chart Successful/Unsuccessful Launches:

- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range

- Allow user to select payload range mass

Scatter Chart Showing Payload Mass Vs Success Rate by Booster Version

- Allow user to see the correlation between Payload and Launch Success

Github Url : Interactive Dashboard Plotly Dash

Predictive Analysis (Classification)

- Create NumPy array from the Class column.
- Standardize the data with StandardScaler.Fit and transform the data.
- Split the data using train_test_split.
- Create a GridSearchCV object with cv=10 for parameter optimization.
- Apply GridSearchCV on different algorithms: logistic regression, support vector machine, decision tree, K-Nearest Neighbor.
- Calculate accuracy on the test data using .score() for all models.
- Assess the confusion matrix for all models.
- Identify the best model using Jaccard_Score, F1_Score, and Accuracy.

Github Url : Machine Learning Predictions

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.

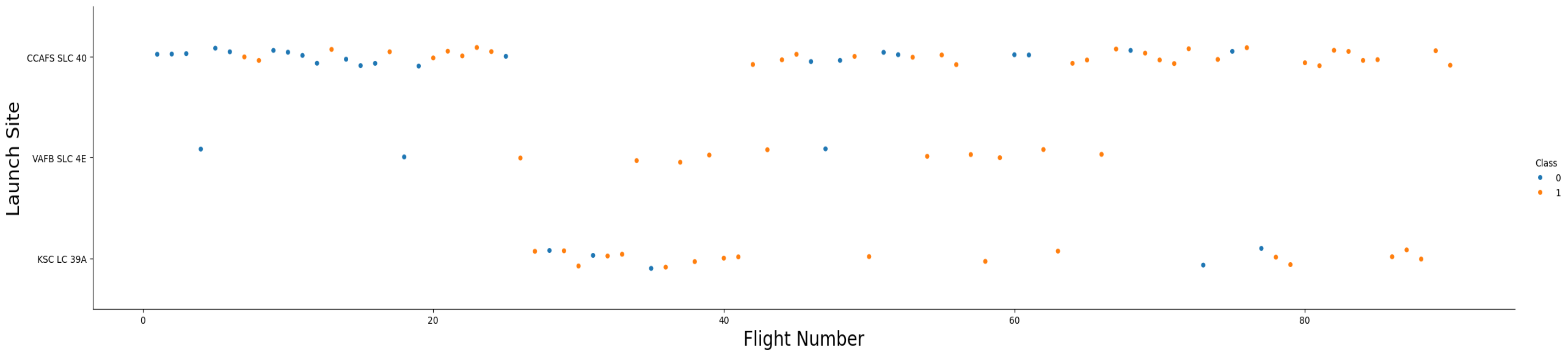
Visual Analysis:

- Most Launch sites are near the equator, and all are close to the coast.
- Launch sites are far enough from any proximities to avoid a failed launch damage (city, highway, railway).

Predictive Analysis:

- Decision Tree Model is the best predictive model for SpaceX dataset

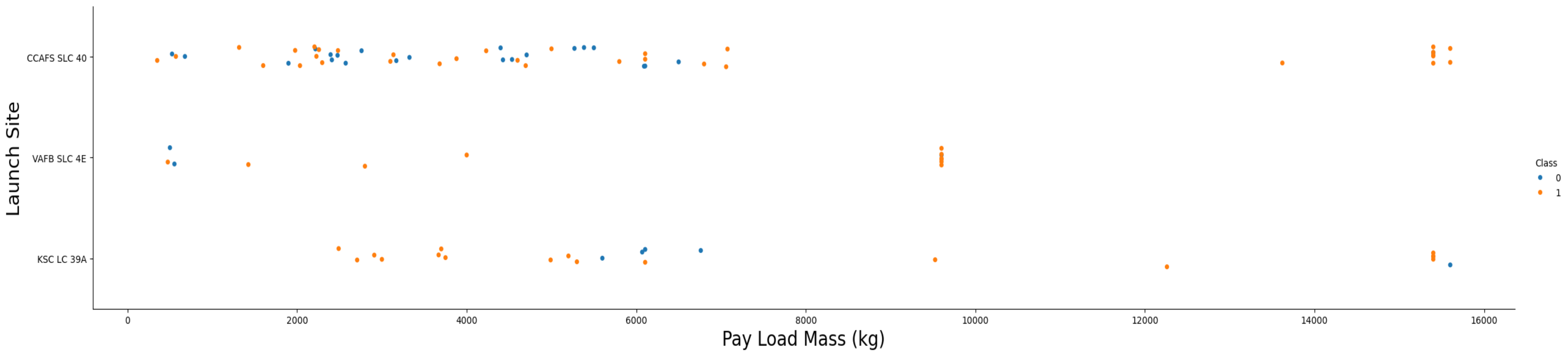
Flight Number vs. Launch Site



Explanation:

- Earlier flights (blue = fail) and Later flights (orange = success).
- The CCAFS SLC 40 launch site has around half of launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- We can assume that new launches have a higher success rate.

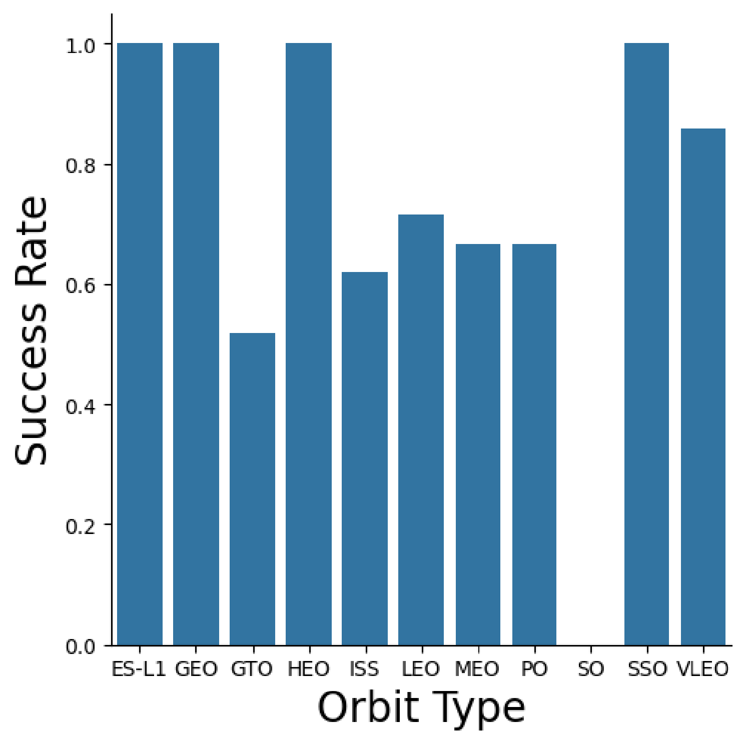
Payload vs. Launch Site



Explanation:

- We can see higher payload mass tends to result higher success rate.
- Most launches with > 7000kg payload were successful.
- KSC LC 39A has 100% success rate for <5500kg payload mass.

Success Rate by Orbit Type



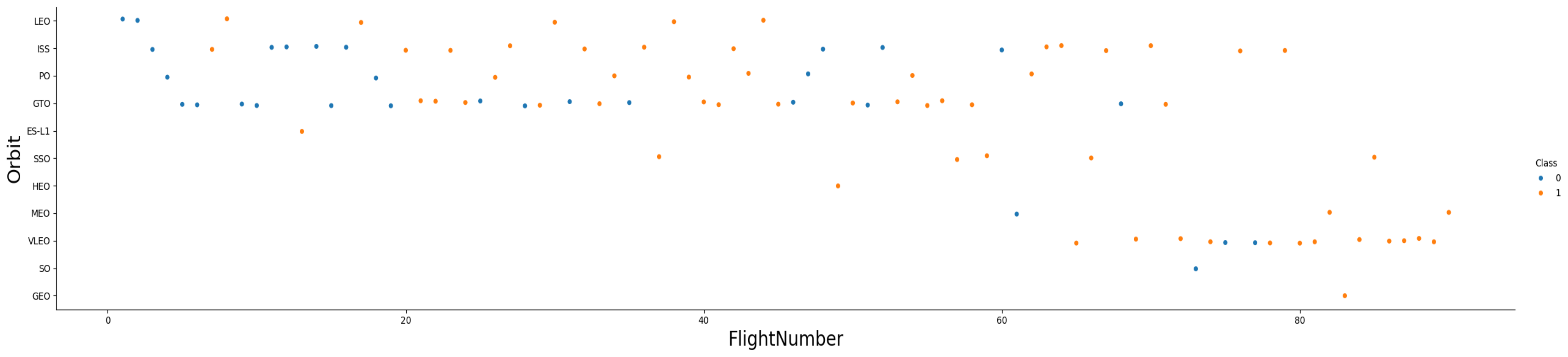
Explanation:

- 100% Orbits Success Rate: ES-L1, GEO, HEO, SSO
- 50% - 90% Orbits Success Rate: GTO, ISS, LEO, MEO, PO, VLEO
- 0% Orbits Success Rate: SO

Flight Number vs. Orbit Type

Explanation:

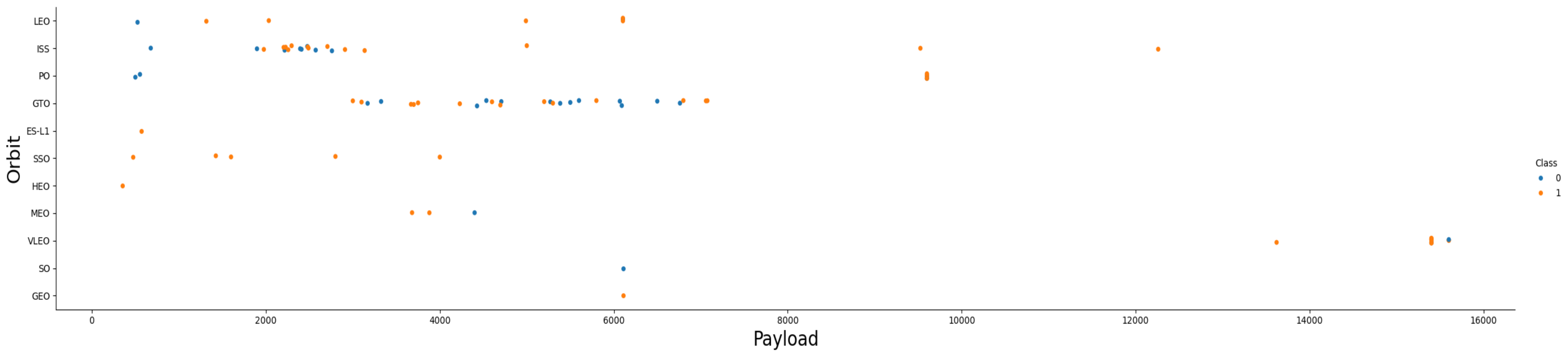
- We see unique phenomenon that the increases of flight numbers tend to result launch success for LEO Orbit.
- Otherwise for other orbits, the increases of flight numbers didn't affect launch success.



Payload Mass vs. Orbit Type

Explanation:

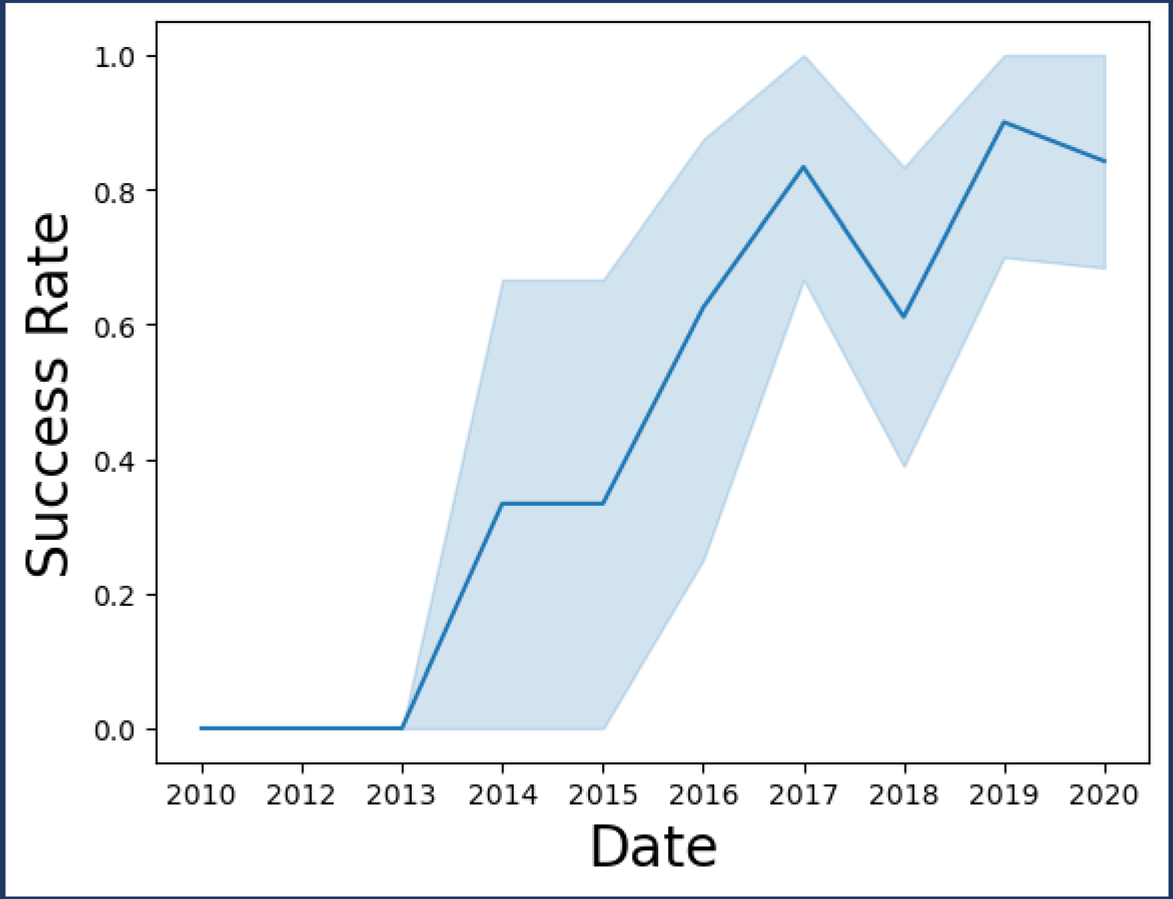
- Higher payload mass affect launch success better for LEO, ISS, and PO orbits.
- Higher payload mass didn't affect launch success for GTO, MEO, VLEO orbits



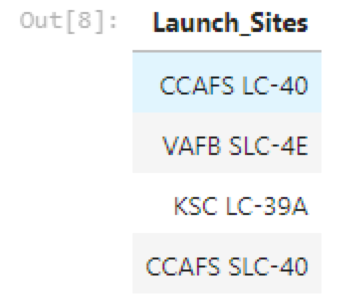
Launch Success Yearly Trend

Explanation:

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



Launch Site Information

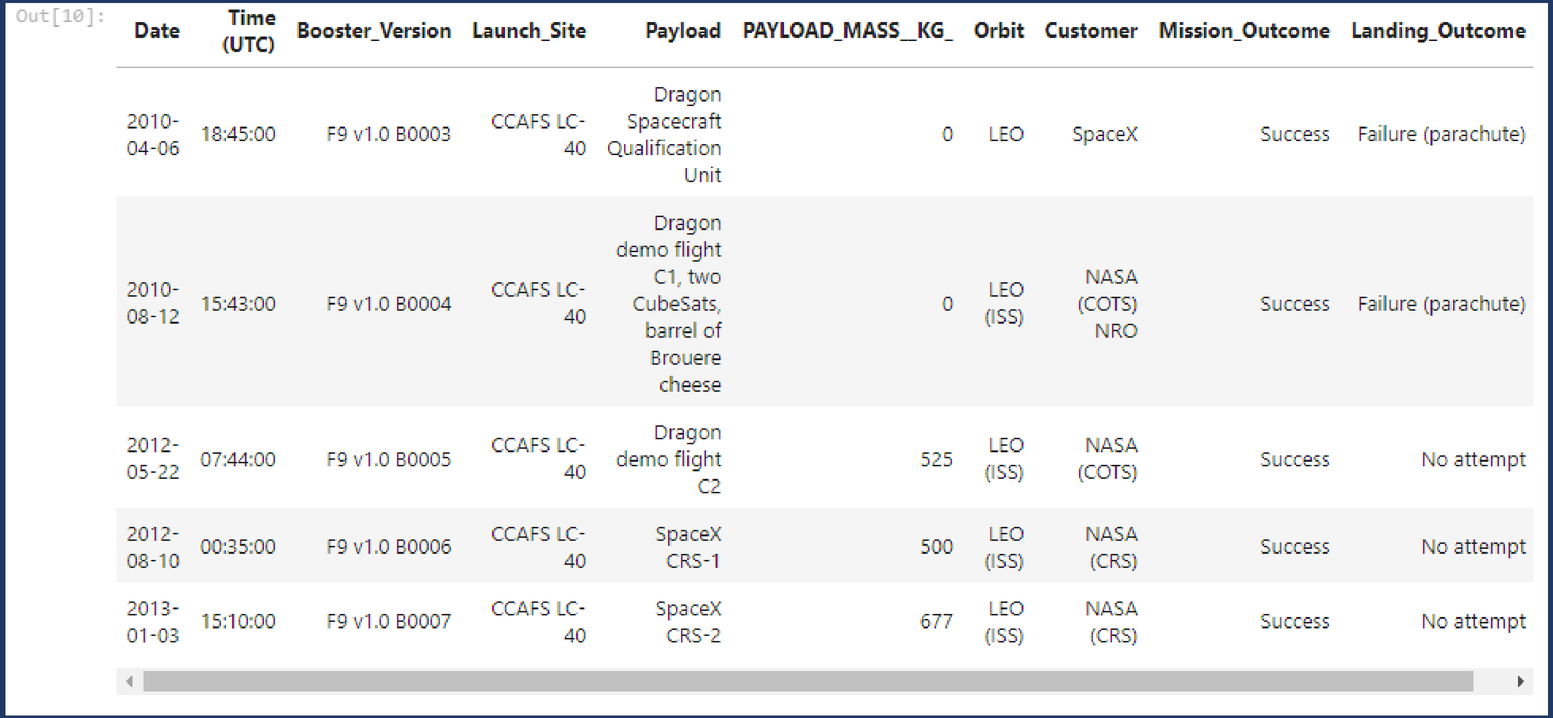


Out[8]:

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

The information contains launch site names.

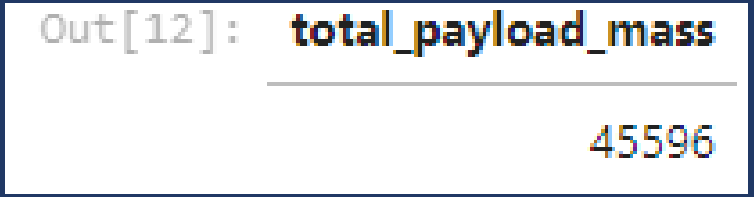
The Information contains records with launch site starting with CCA



Out[10]:

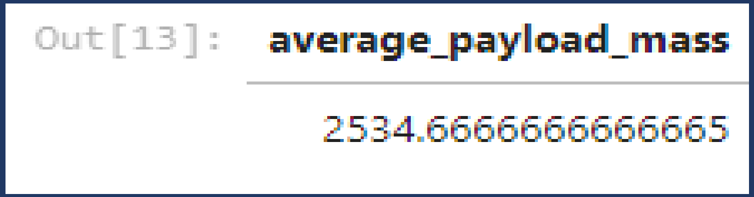
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Payload Mass



Out[12]:

total_payload_mass
45596



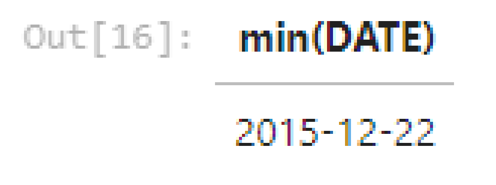
Out[13]:

average_payload_mass
2534.6666666666665

45.596 kg (Total) carried by boosters launched by NASA (CRS)

2.534 kg (Average) carried by booster version F9 v1.1

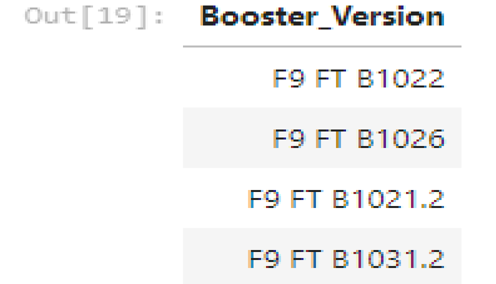
Landing & Mission Info



Out[16]:

min(DATE)
2015-12-22

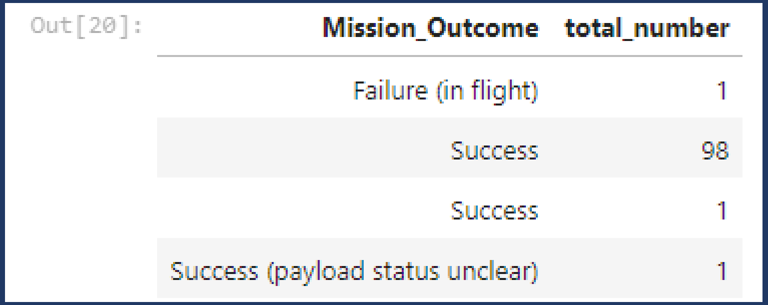
The first successful landing in ground pad occurred on 22 December 2015.



Out[19]:

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

The information contains booster version successful landing that carry payload between 4000 kg and 6000 kg.



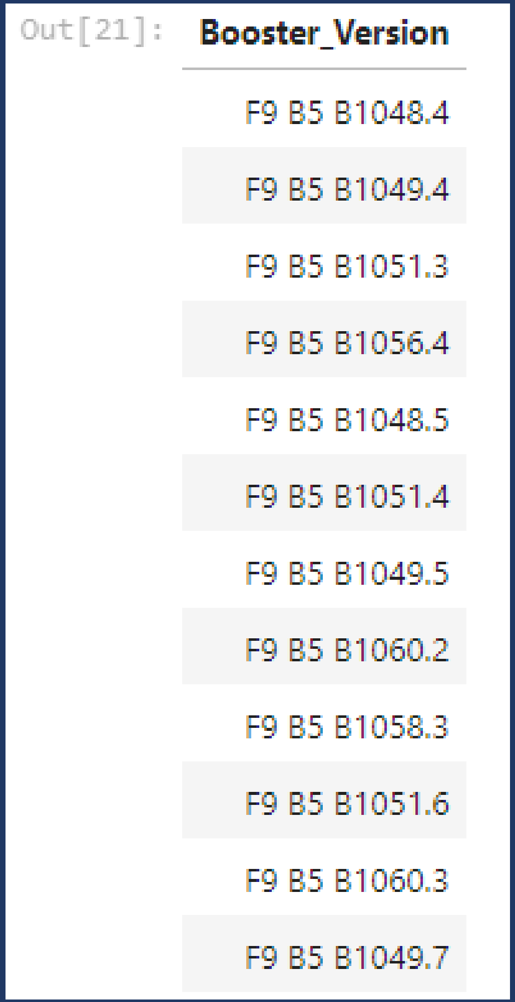
Out[20]:

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

There are 1 failure in flight and 99 success Mission.

1 Other success mission but payload status unclear.

Boosters Carried Maximum Payload

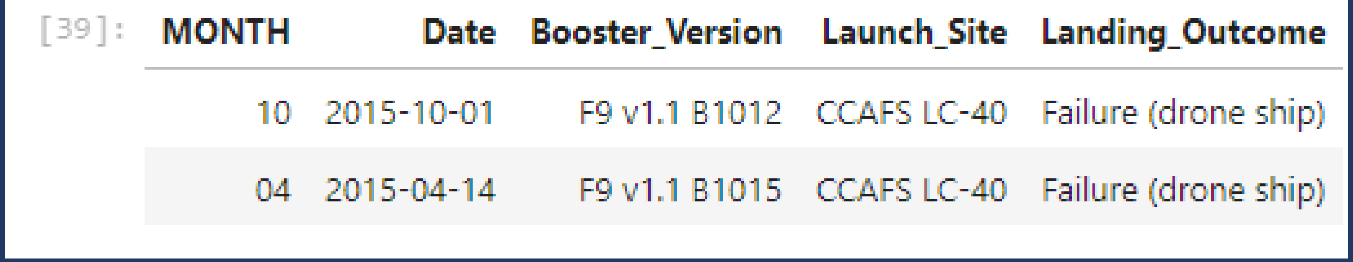


Out[21]:

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

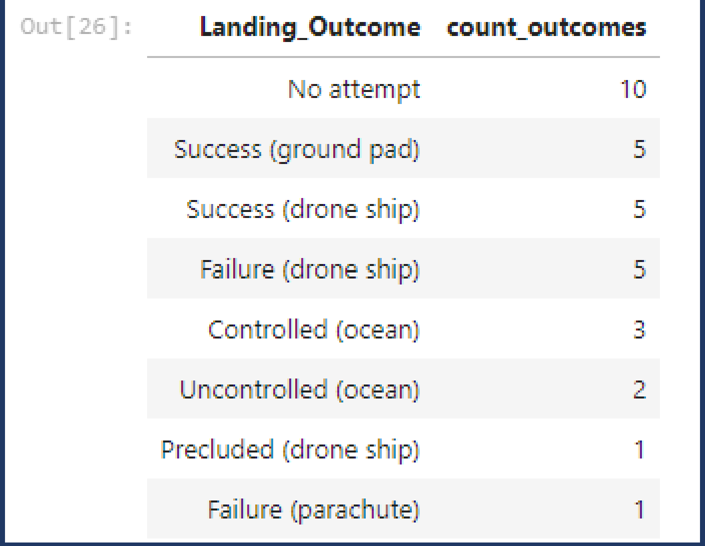
2015 Launch Records

List of failed landing outcomes in 2015



[39]:

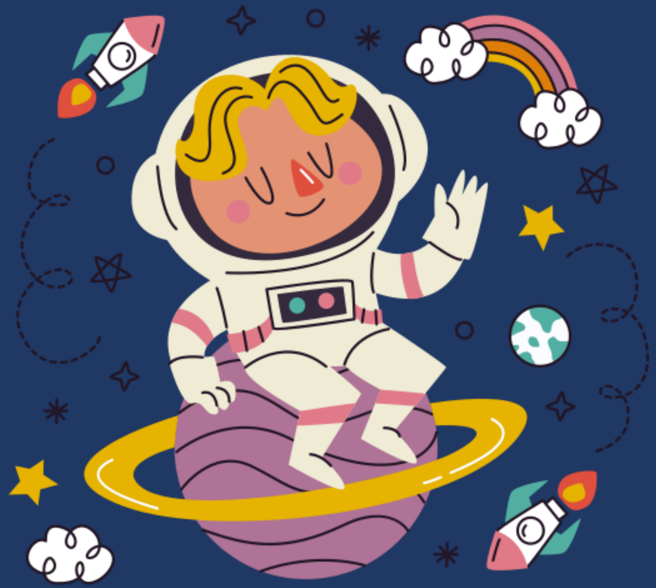
MONTH	Date	Booster_Version	Launch_Site	Landing_Outcome
10	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

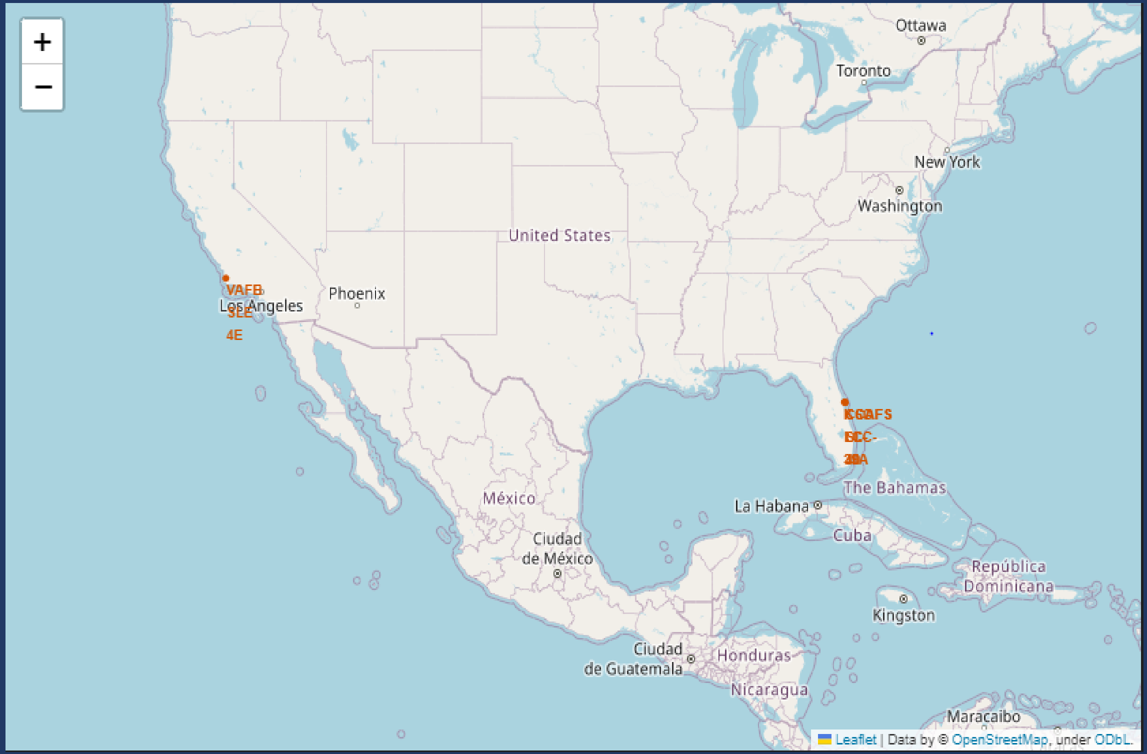


Out[26]:

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

The count of Landing Outcomes.



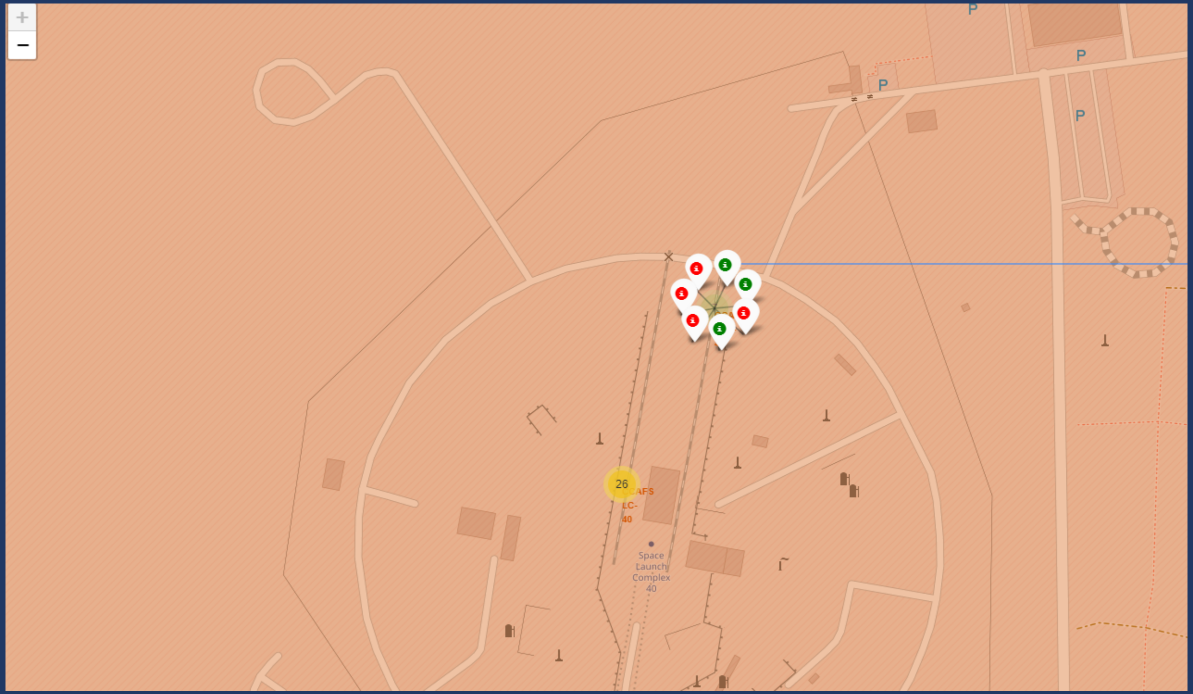


All Launch Sites

Explanation:

Most of Launch sites are in proximity to the Equator line. Rocket launches will be more cost-effective because the equator is the closest point on Earth to outer space orbit. Also all launch sites are very close to the coast, it’s minimizes the risk of launch failure that possible to impact on the surrounding environment.

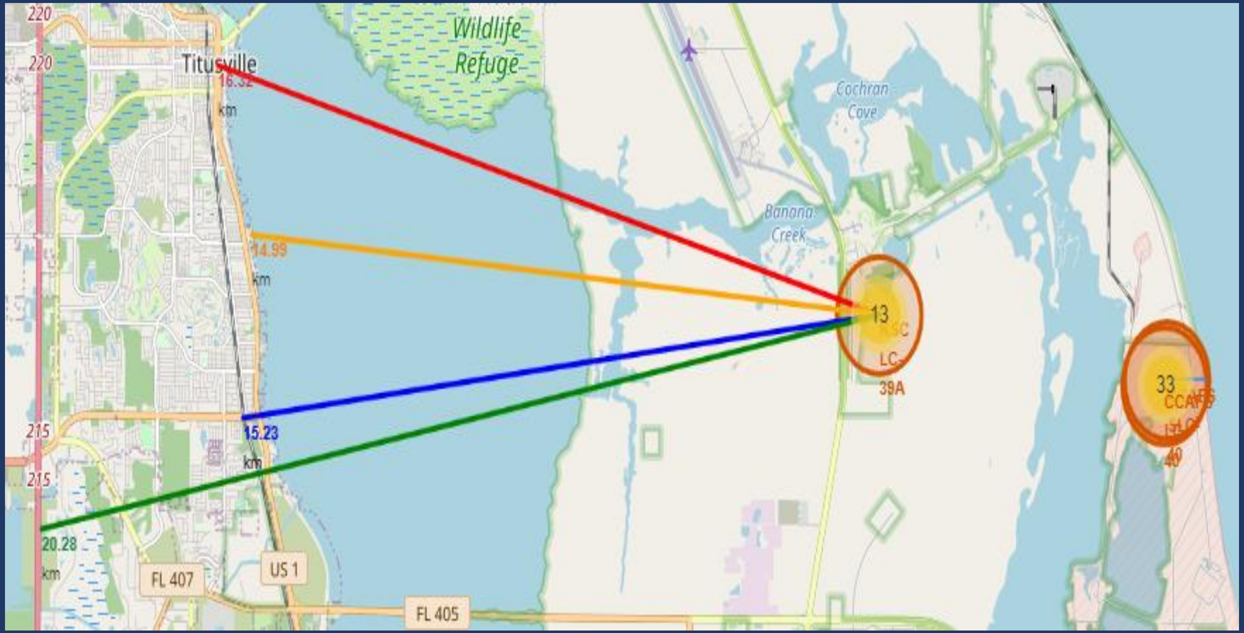
Launch Outcomes



Explanation:

- This is the example of visualization color-labelled marker for each launch sites (CCAFS SLC-40).
- Green markers are successful launches.
- Red markers are unsuccessful launches.

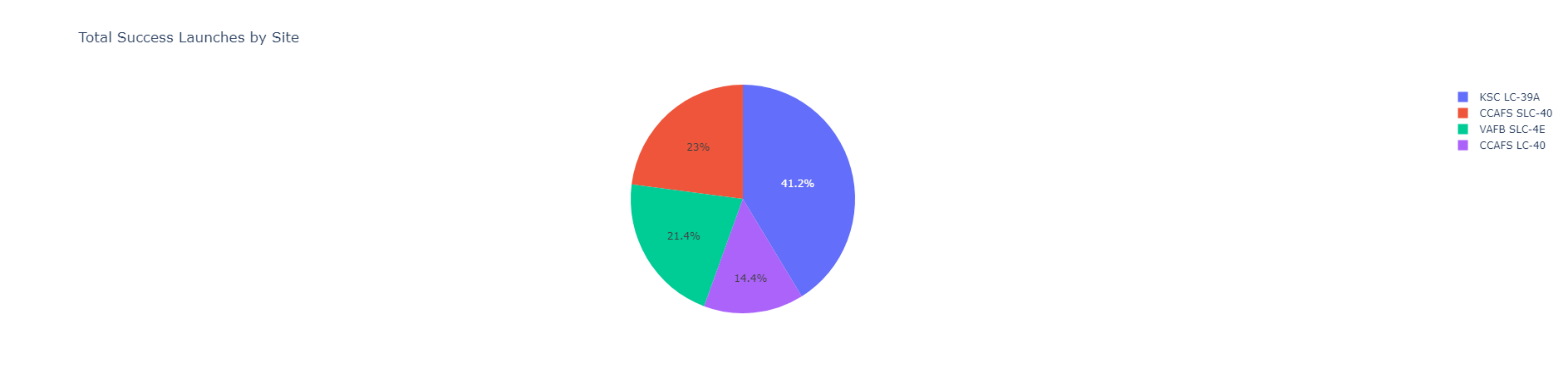
Distance to Proximities



KSC LC-39A Launch Site.

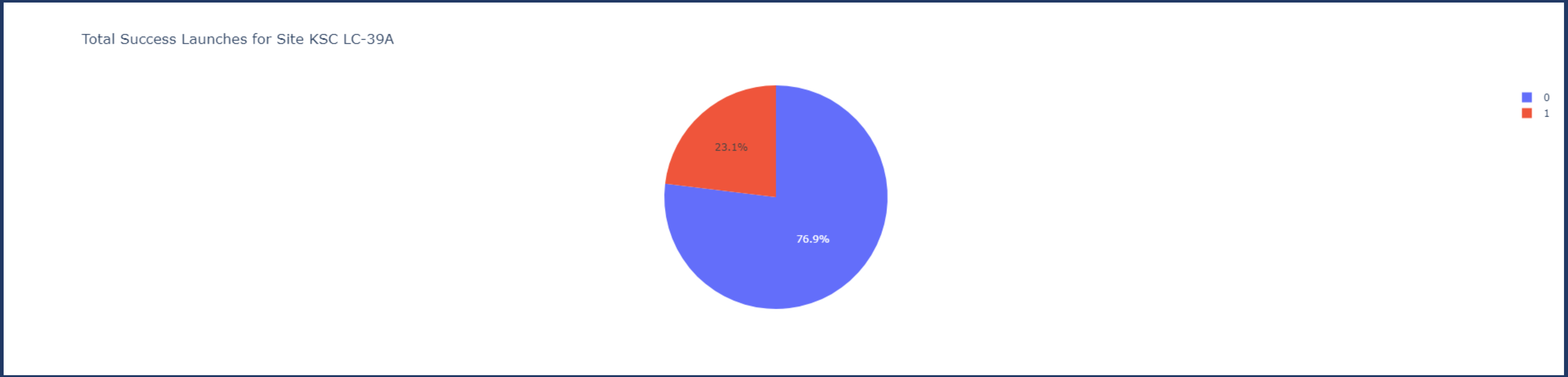
- Red line: distance between site and nearest city (Titusville 16,32 km)
- Orange line : distance between site and coastline (14,99 km)
- Blue line: distance between site and railway (15,23 km)
- Green line: distance between site and highway (20,28 km)

Launch Success by Site



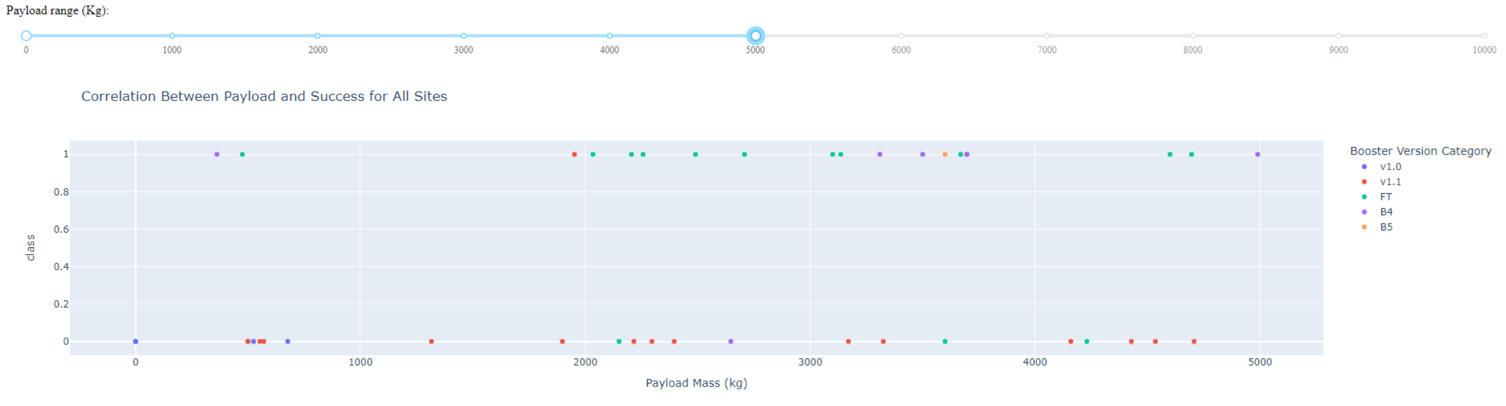
- KSC LC-39A has the most successful launches amongst launch sites by 41.2% from the all sites total launches.
- CCAFS LC-40 has the lowest successful launches by 14.4%.

Highest Launch Success (KSC LC-29A)



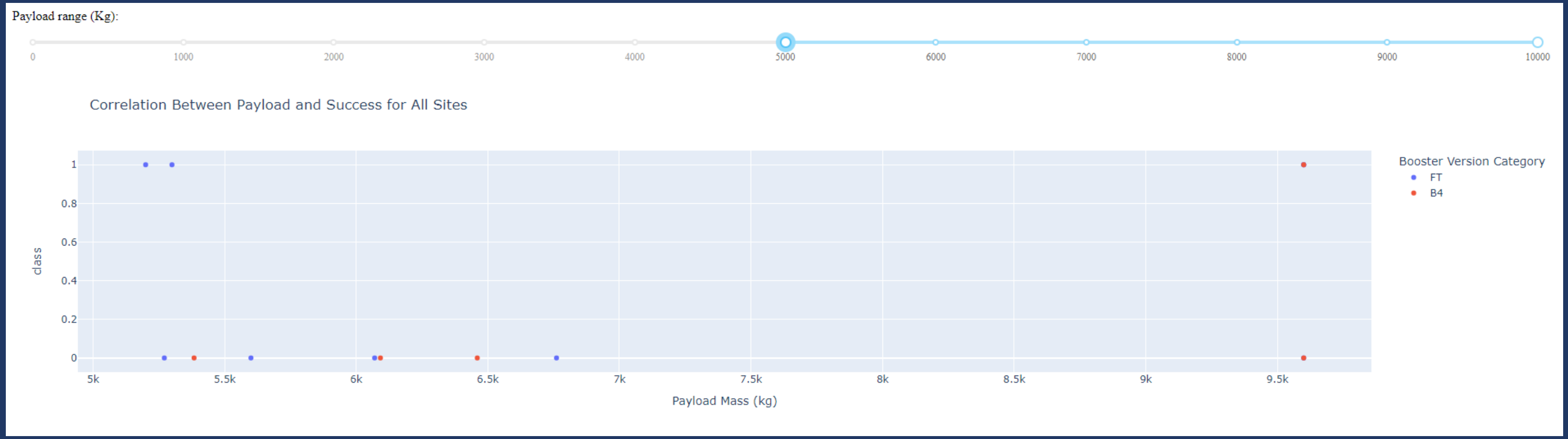
- KSC LC-39A has 76.9% launch success rate.
- 10 successful launches and 3 failed launches.

Payload Mass and Success

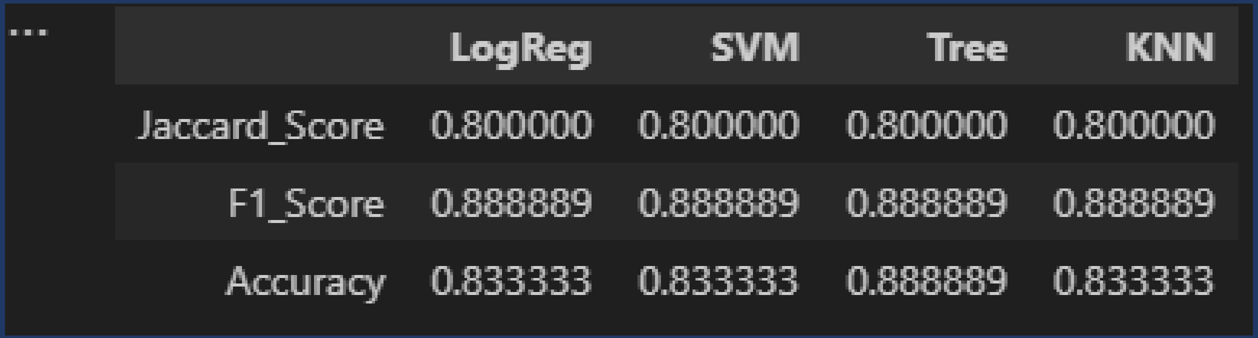


Explanation:

- 1 indicating successful outcome and 0 indicating unsuccessful outcome
- We can assume that Payload Mass between 2000 kg and 5000 kg have higher success rate then 5000 kg – 10000 kg.



Classification Analysis

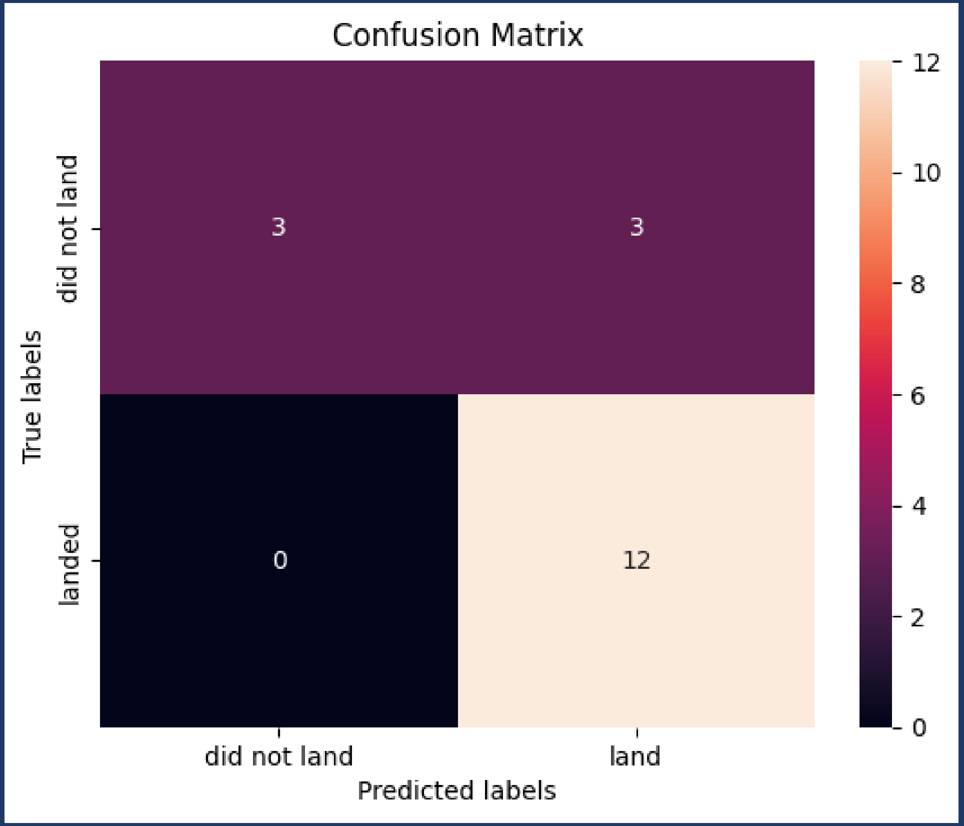


	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.888889	0.833333

Explanation:

- We can see all the models performances are similar. But the Tree Decision model has slightly outperformed the rest in Accuracy.
- In the future, we recommend to use Tree Decision model for predict launch outcomes.

Confusion Matrix

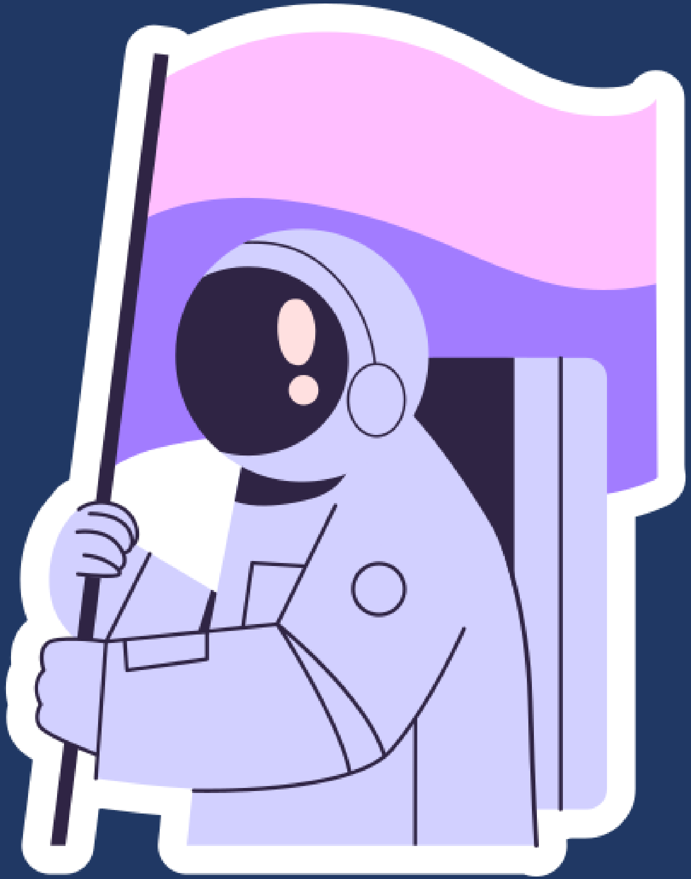


Explanation:

- All the models have identical confusion matrix.
- Confusion Matrix Outputs:
  - 12 True positive (TP)
  - 3 True negative (TN)
  - 3 False positive (FP)
  - 0 False negative (FN)
- Precision = TP/(TP+FP) = 12/15 = 0.80
- Recall = TP/(TP+FN) = 12/12 = 1
- F1 Score = 2 * (Precision * Recall)/(Precision + Recall) = 2 * (0.8 * 1) / (0.8 + 1) = 0.89
- Accuracy = (TP + TN) / (TP+ TN + FP + FN) = 0.833

Conclusion

- KSC LC-39A is the highest success rate launch site.
- ES-l1, GEO, HEO, and SSO are the orbits that have 100% success rate.
- Payload Mass has linear relationship with success rate, if payload mass is higher then success rate increases.
- Launch success increases over time.
- Most launch sites are in proximity to the Equator line and near the coast to minimize the impact of launch failure.
- Tree decision model is the best prediction model for this dataset.



Thank You:

Instructors

IBM

Coursera

Audience