IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sheetal Sattiraju
1/10/2024

Outline

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

Executive Summary

- Methodologies Used:
  - Data Collection, Normalization & Cleaning
  - Data Web Scraping using BeautifulSoup
  - Exploratory Data Analysis using SQL
  - Exploratory Data Analysis using Python Visualization
  - Interactive Map using Folium
  - Interactive Analytics using Plotly Dash
  - Predictive Analytics using Machine Learning

- Summary of Results:
  - EDA Results: Discovering Important Features
  - Interactive Analytics: Displaying EDA Results
  - Predictive Analytics Results: Model Accuracy

Introduction

- Project background:
  - SpaceX is a successful spacecraft manufacturer & launcher service provider. SpaceX advertises Falcon9 Rocket launches, which cost 62 million dollars, while others cost upwards of 165 million dollars each. This is because SpaceX can reuse the first stage; thus if we can calculate if the first stage will land, we can determine the cost of a launch.
  - The objective of this project is to find the right price & evaluate if SpaceY can compete with SpaceX

- Questions:
  - What variables affect the success of first stage landing?
  - What are the variables relationships to each other?
  - What is the best algorithm to determine if SpaceX will reuse the first stage?

Section
1
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collection from SpaceX Rest API & Wikipedia
- Perform data wrangling
  - Filtered data, Cleaning Missing Values, One-Hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Pre-processing, divided data into train/test sets, evaluated 4 models & tested accuracy

Data Collection

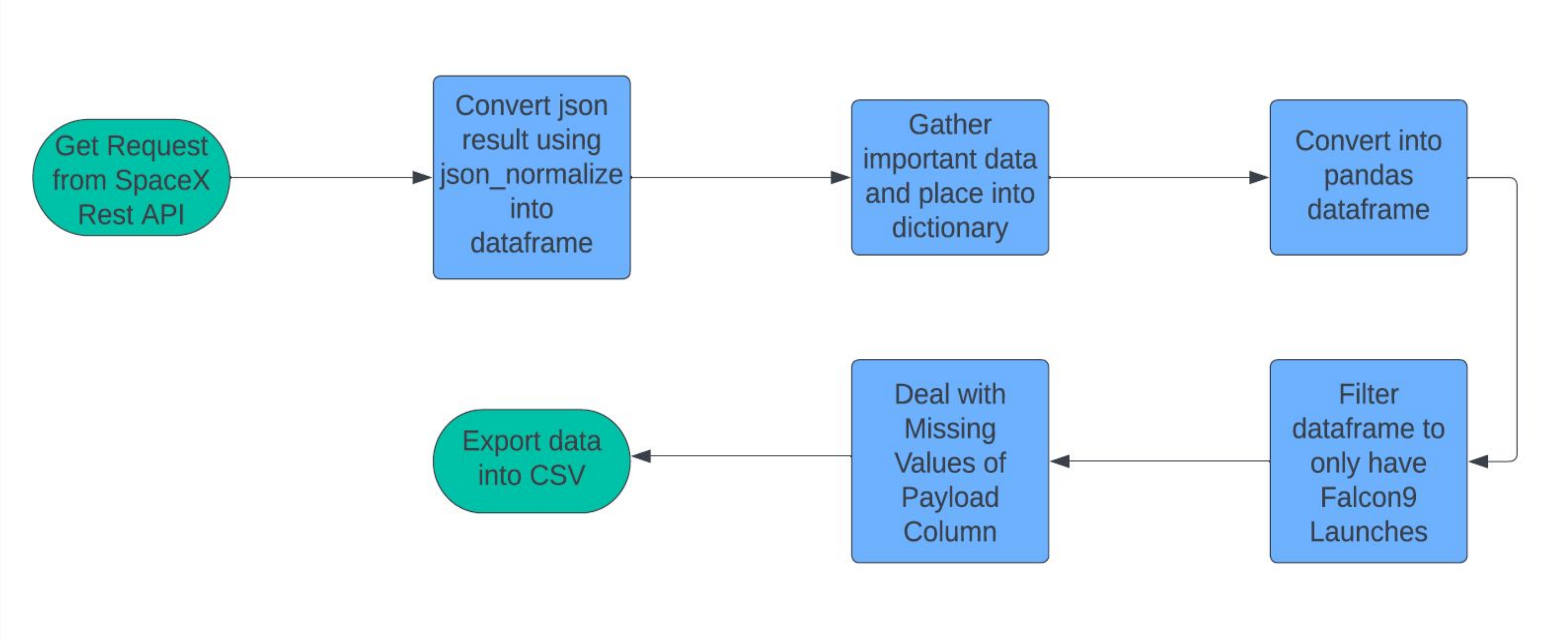
- The Data Collection Process included gathering data from SpaceX Rest API & Wikipedia.

- Data from SpaceX Rest API was collected by a 'get' request. With json_normalize, we can normalize the response & convert the data into a pandas dataframe

- Data from Wikipedia was collected using web scraping. With BeautifulSoup, we can extract the launch records as a HTML table & convert the data into a pandas dataframe

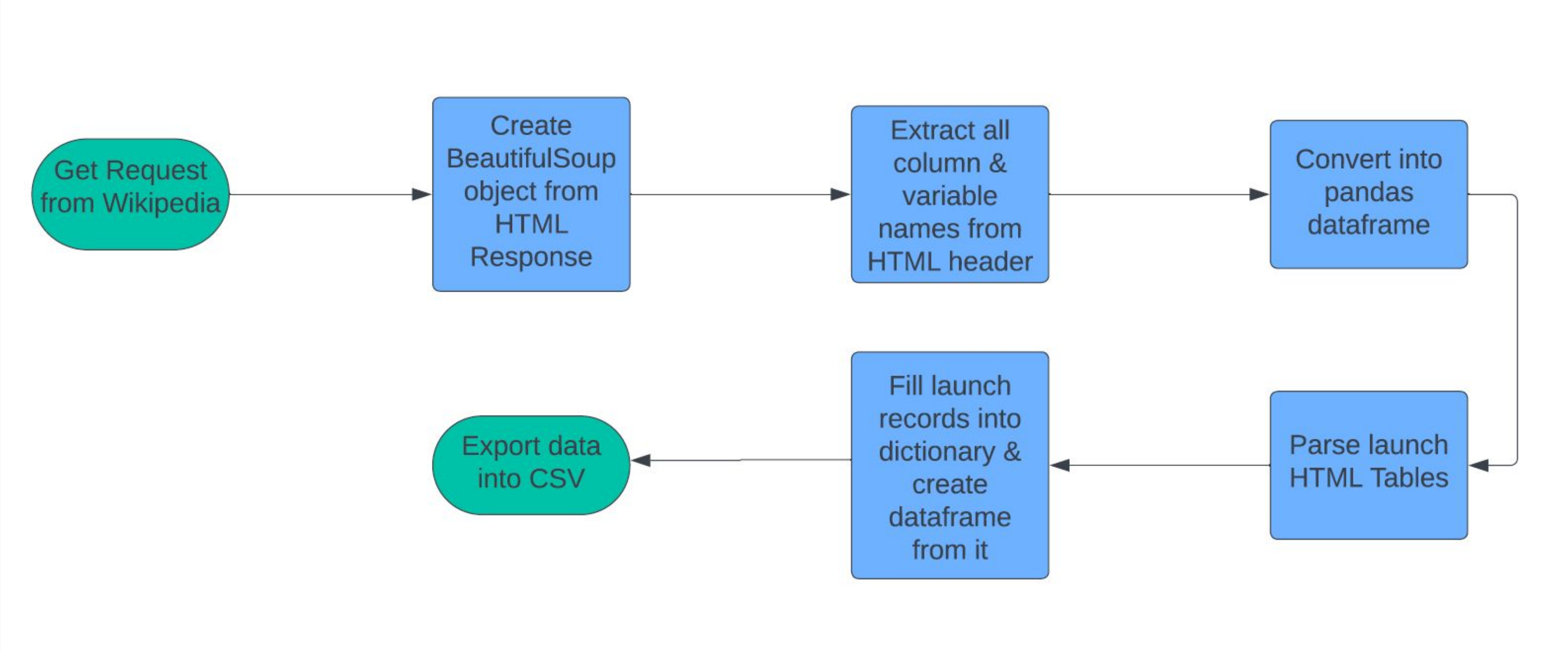
- SpaceX Rest API: https://api.spacexdata.com/v4/
- Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API



Github

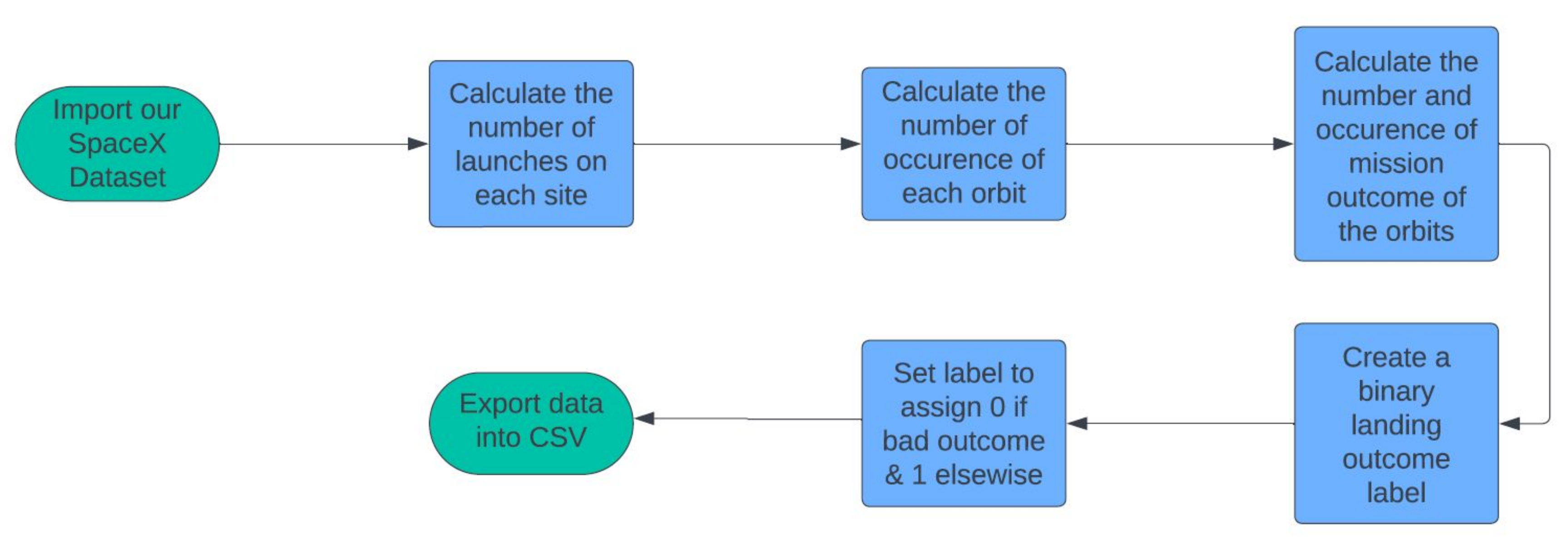
Data Collection - Scraping



Github

Data Wrangling

- In this project, the data wrangling process consisted of performing some EDA to find some patterns in the data & determine labels.



[Github]

EDA with Data Visualization

- Scatterplots Plotted:
  - Flight Number vs PayloadMass
  - Flight Number vs Launch Site
  - PayloadMass and Launch Site
  - Flight Number vs Orbit
  - PayloadMass vs Orbit
- BarChart & Line Chart Plotted:
  - Orbit vs Success Rate
  - SpaceX Rocket Success Rate over the years
- Through all of these Charts, we can see key trends & help us answer our business questions

* see charts in GitHub

Github

EDA with SQL

- SQL Queries performed:
    - Displaying the names of launch sites
    - Displaying 5 launch sites that start with ‘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 of the first successful landing outcome in ground pad
    - Listing the names of the boosters with success in drone ship & payloadmass greater than 4000 but less than 6000
    - Listing the total number of ‘successful’ & ‘failure’ mission outcomes
    - Listing the names of the booster_versions that have carried the max payloadmass
    - List the records of ‘failure’ outcomes in the year 2015 (month, date, booster_version, launch_site included)
    - Rank the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

[Github]

Build an Interactive Map with Folium

- Map Objects created:
  - Markers: to highlight all launch sites
  - Circles: to highlight coordinates & text, like NASA Johnson Space Center
  - Colored Markers: to depict success (green) & failure (red) launches
  - Lines: to depict distances between different coordinates
- All of these objects provide important value to the interactive map. We can zoom in/out, click for information on the launch sites and see key information that we are looking for

Github

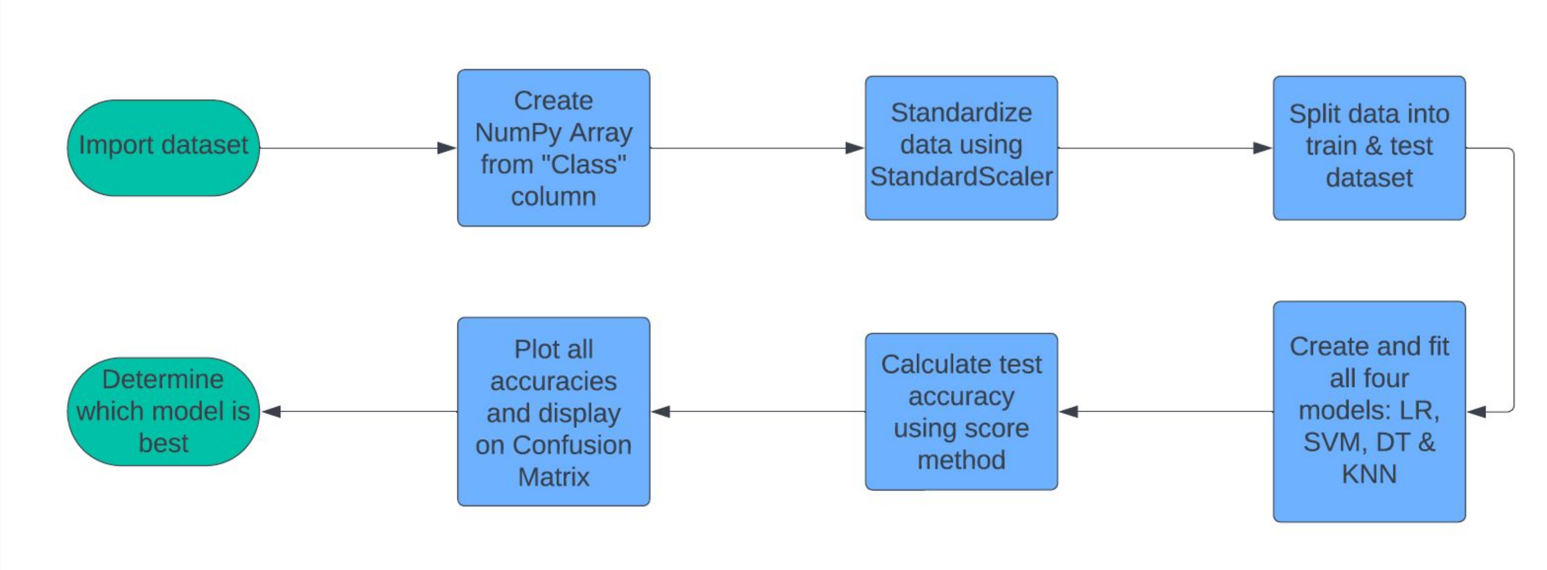
Build a Dashboard with Plotly Dash

- Drop down List:
  - Allows user to pick launch site or see all sites
- Pie chart
  - Shows total successful launches for all sites
  - Shows Success vs Failed counts for site picked
- Scatter plot
  - Shows correlation between PayloadMass and Site
- Slider
  - Allows user to select Payload range

Github

Predictive Analysis (Classification)

- Models used:
    - Logistic regression, SVM, Decision Tree & KNN



[Github]

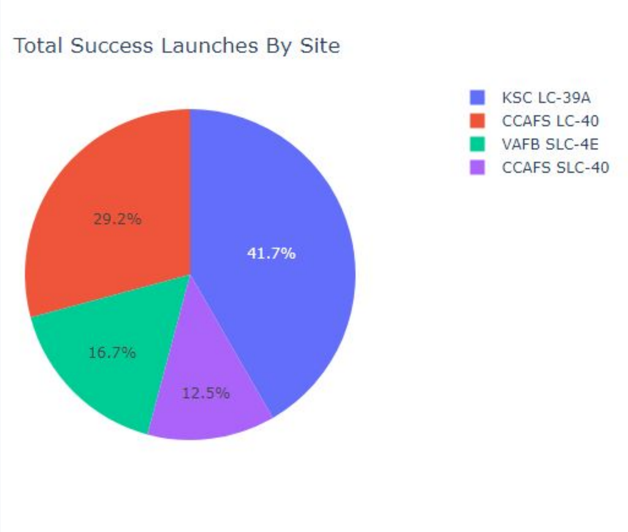
Results

- Some Exploratory data analysis results:
  - SpaceX has 4 launch sites
  - As the number of flights increased, the higher chance of first stage landing
  - SpaceX Rocket Success Rate increased over time until 2020
  - Different launch sites had varying success rates
- Predictive analysis results
  - Almost all the models had a similar accuracy and confusion matrix, except for the DT

Model	TestAccuracy
LogReg	0.833
SVM	0.833
Tree	0.944
KNN	0.833

Results Part 2

- Interactive analytics:



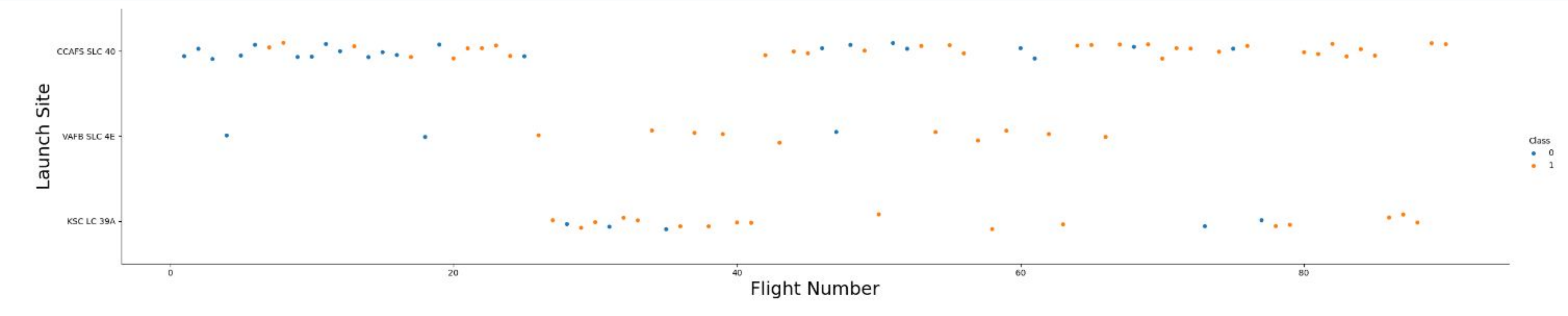
We can see what sites have better success.

We can see which launches were successful.

Section
2

Insights drawn from EDA

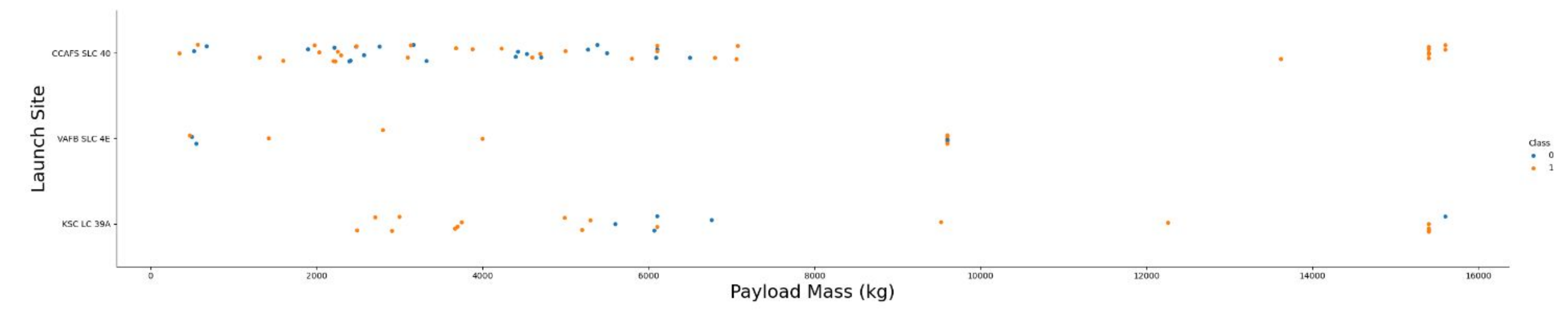
Flight Number vs. Launch Site



From this scatter plot, we can see that KSC LC 39-A had the most success overall. We can also note that CCAFS SLC 40 has had the best recent success.

The overall trend of all launch sites is that with more flights, the average success increases.

Payload vs. Launch Site



From this scatter plot, we can see that most flights have a payload mass between 0-7000 kg.

We can also note that any payload mass over 10,000 kg has only successfully been done at CCAFS SLC 40 & KSC LC 39 A.

Success Rate vs. Orbit Type

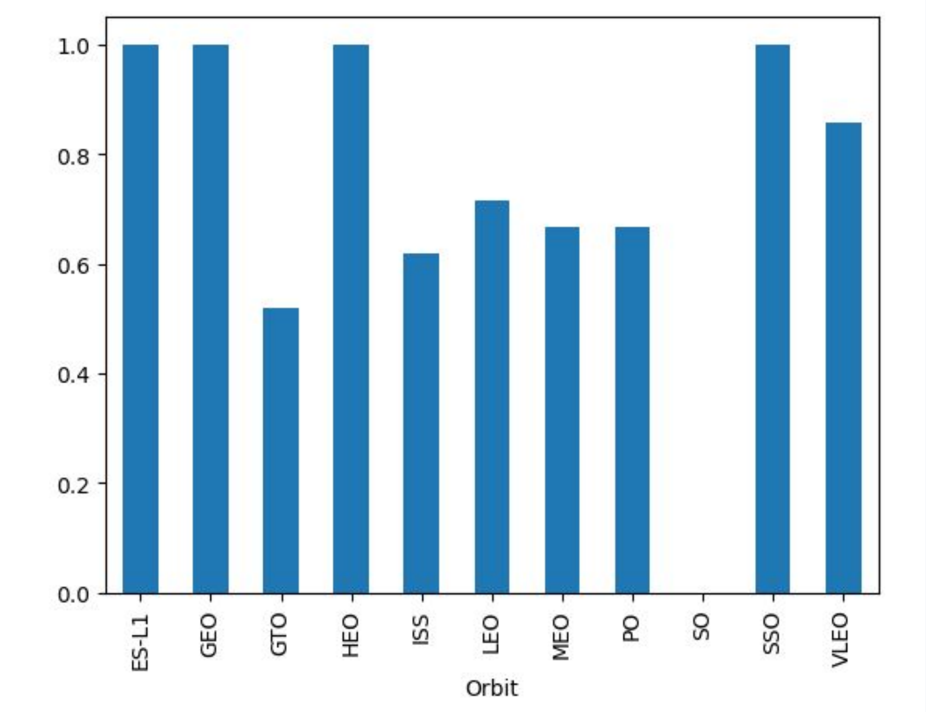
We can see that the best Orbit types are:

- ES-L1, GEO, HEO & SSO

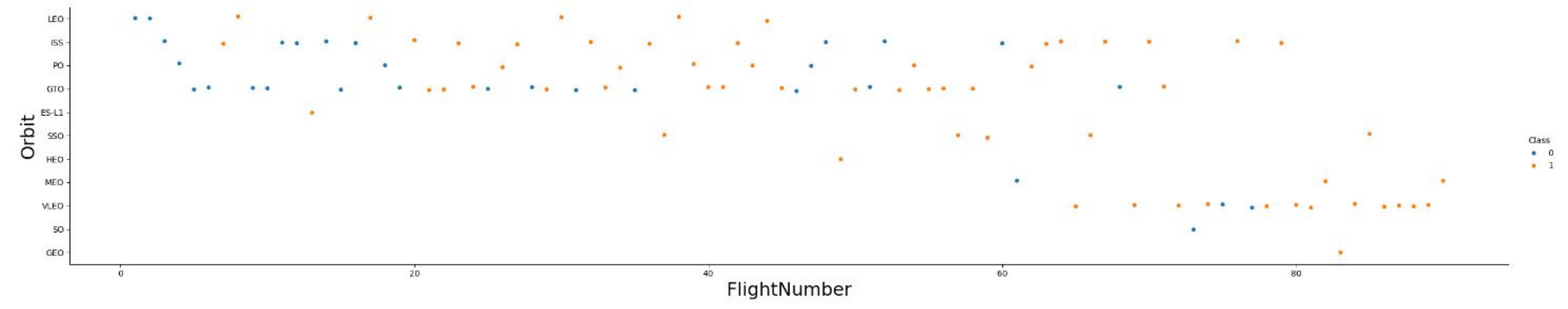
We can also see that VLEO was a close second. The next set of orbit types with similar success rates are:

- ISS, LEO, MEO, PO & GTO

The worst orbit type was SO at a 0% success rate.



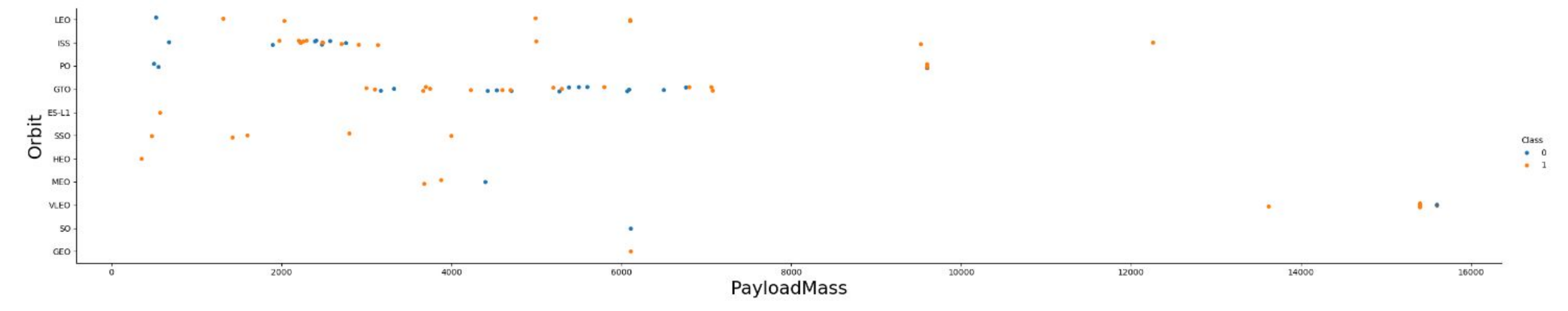
Flight Number vs. Orbit Type



We can see that overtime, the success rate of orbits improved as the number of flight's went up. (exception SO)

We can also not that VLEO has had a recent high success rate; while GTO had a low success rate start.

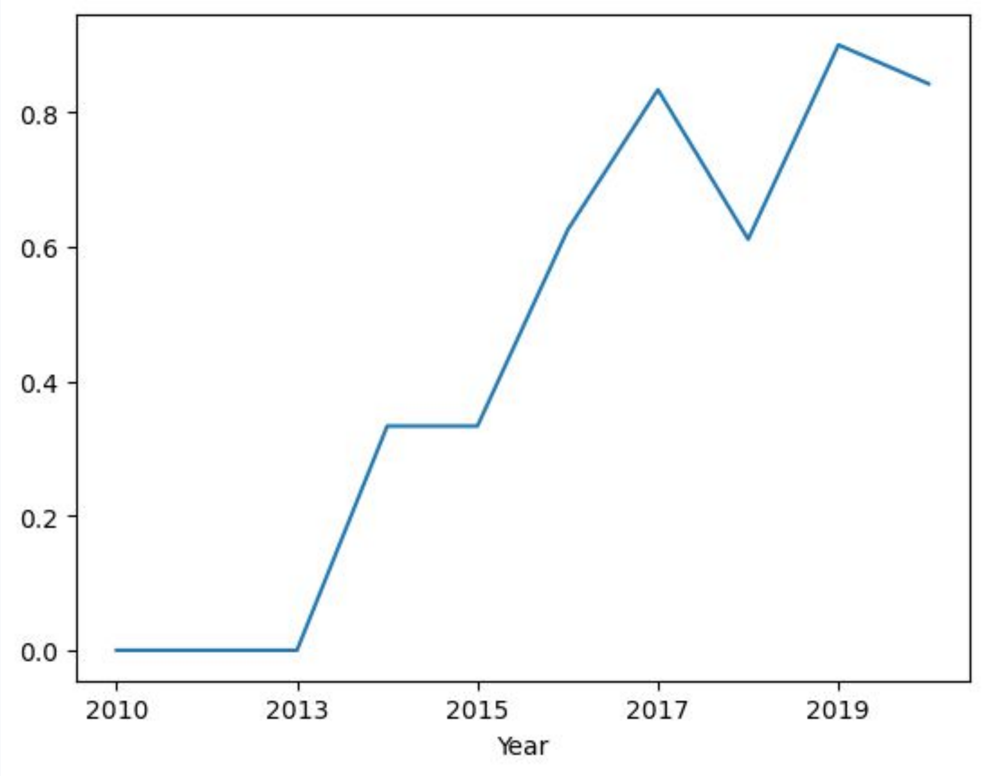
Payload vs. Orbit Type



We can see that for LEO, ISS & POLAR, the heavier the payload, the higher success rate.

However, for GTO, we cannot see any such relationship between these attributes.

Launch Success Yearly Trend



We can see that overall, the success rate went up.

We can also note that from 2010 to 2013, the success rate was 0% → likely to be a preliminary testing period.

We can see after 2020, there is a decrease but still maintained around an 80% success rate.

All Launch Site Names

- 4 Unique Launch Sites:

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

By selecting distinct launch_site from the spacetable, we can see this information.

Launch Site Names Begin with 'CCA'

- 5 records where launch sites 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	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

By setting the condition: LIKE ‘CCA’ limit 5, we can display this information.

Total Payload Mass

- Total payload carried by boosters from NASA

SUM(PAYLOAD_MASS__KG_)
45596

By selecting the sum of payload_mass__kg from the space table, we can see this information.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1:

AVG(PAYLOAD_MASS__KG_)
2928.4

By selecting the average of payload_mass__kg from space table where the boost version was F9 v1.1, we can see this information.

First Successful Ground Landing Date

- First successful landing outcome on ground pad (Date):

MIN(DATE)
2015-12-22

By selecting the minimum date from space table where the landing outcome was success (ground pad), we can see this information.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

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

By selecting the booster version from the space table where the payload mass is greater than 4000 but less than 6000, we can see this information.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes:

COUNT(*)
101

By selecting the count from space table where the outcome is success or failure, we can see this information.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass:

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

To find the max payload mass, we have to make use of an aggregate function. On the outside of this function, we simply select the booster version from the space table & connect these clauses with a ‘where’ clause.

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

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)

There were only two occurrences of failure (drone ship) in 2015.

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

- Rank 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:

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

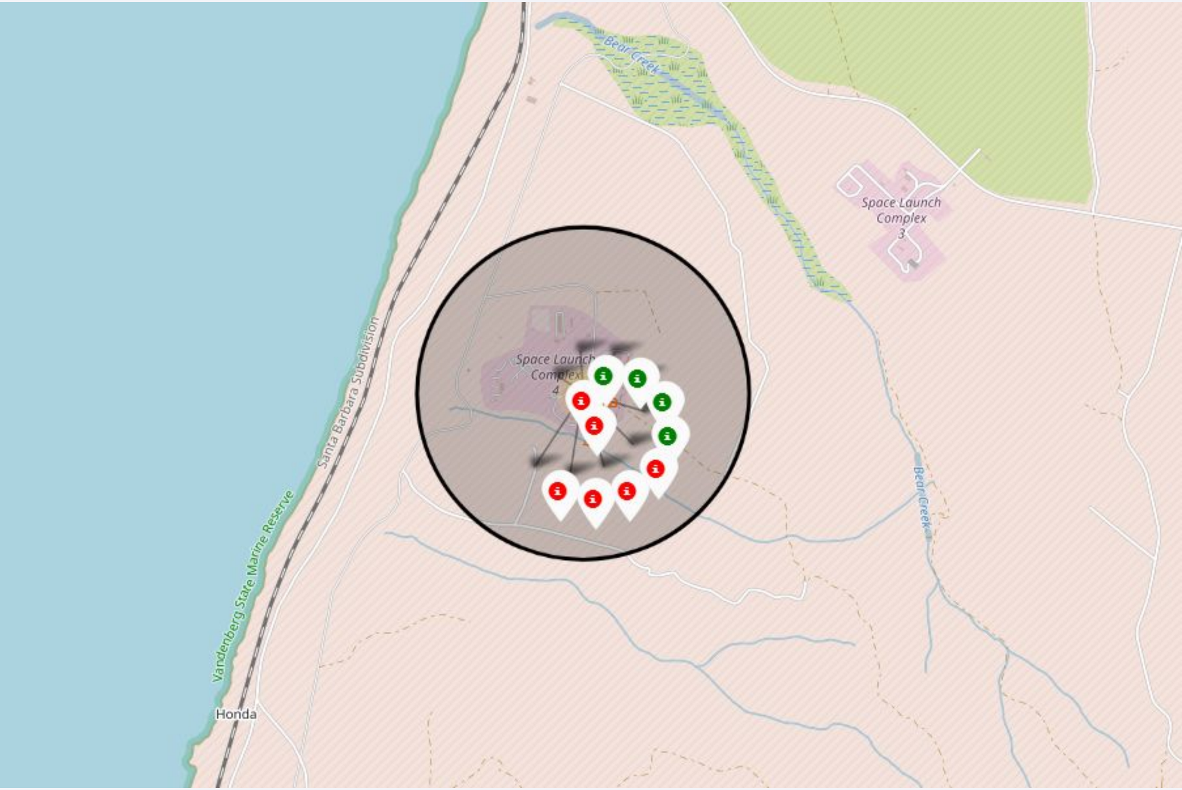
Launch Sites Proximities Analysis

All Launch Site Locations

As we can see, the site locations are coastal → no risk in case of failure, since it will fall into the ocean

We can also note that the launch sites are near the equator → speed boost from Earth's rotation

Launch Outcomes



Example: VAFB SLC-4E

As we can see, red markers indicate failure & green markers indicate success

This particular site has a lower success rate comparatively.

Launch Sites to Proximities

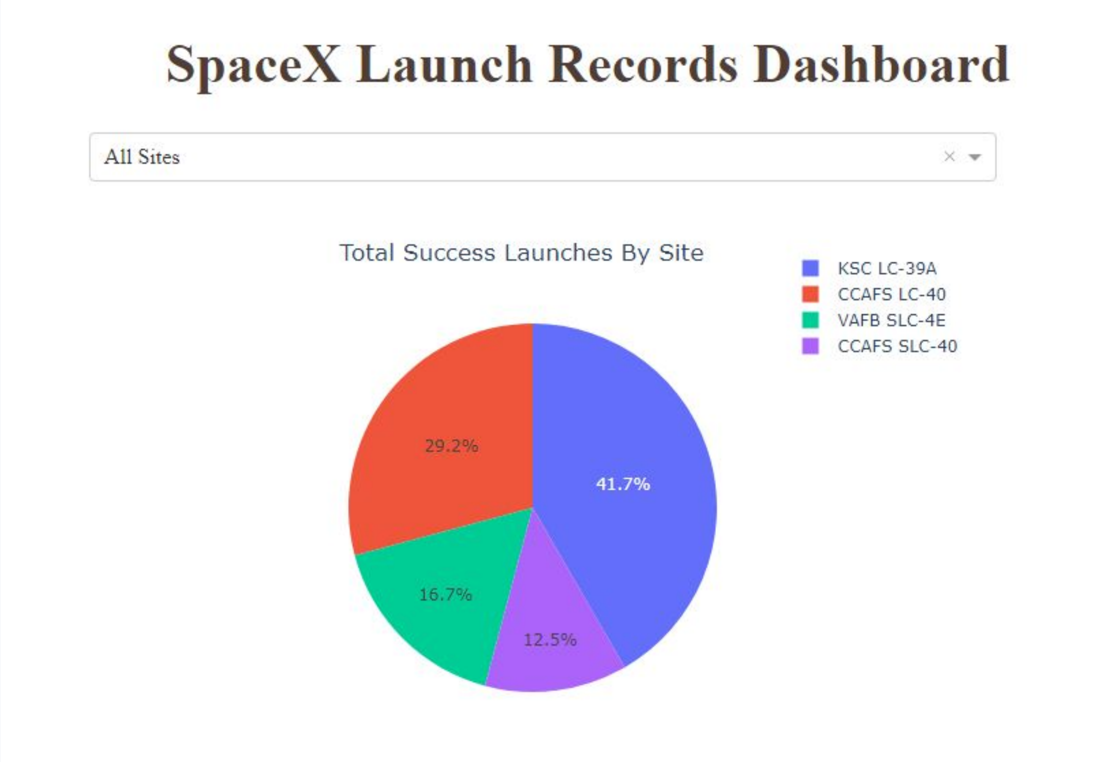
Example: KSC LC-39A Region

We can see the location's proximity to the highway, railroad, coastline & distance from cities.

Section
4

Build a Dashboard with Plotly Dash

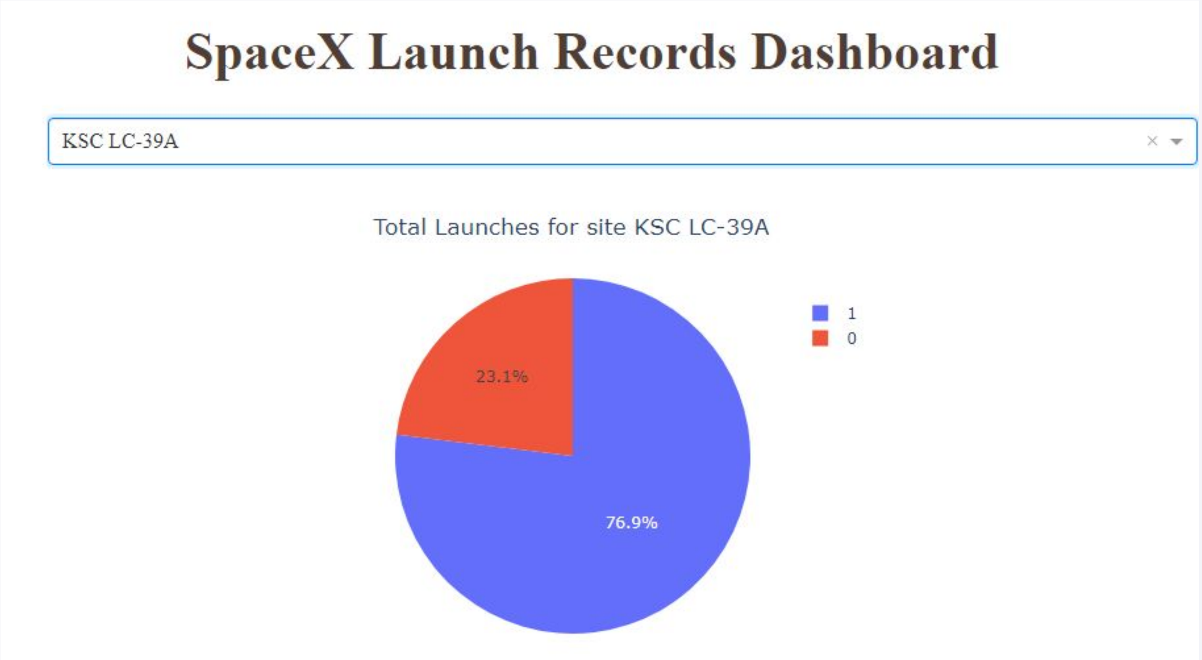
Launch Success Count



We can see the total success launches by site in this pie chart.

We can note that KSC LC-39A has most total success launch by site (41.7%), while CCAFS SLC-40 has the worst (12.5%).

Launch Site for KSC LC-39A

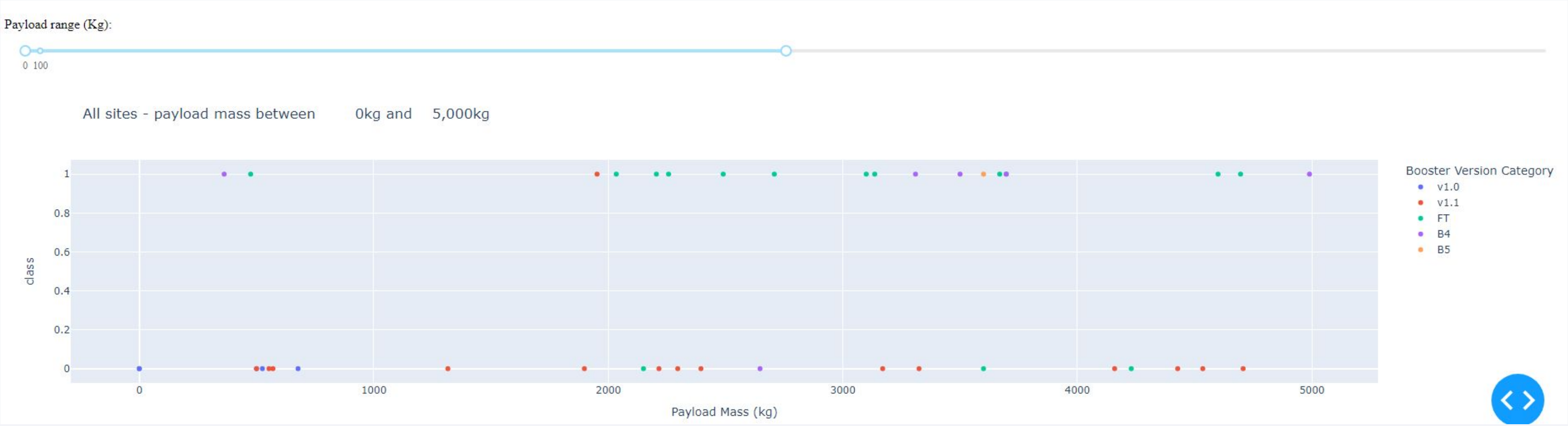


We can see the total success launches vs failure for KSC LC-39A.

Note that 1 is success & 0 is failure.

We can see that 76.9% launches at this site are successful, beating other sites by a pretty large margin.

Payload vs Launch Outcome



We can see the Payloadmass vs Class for ranges 0-5000 kg

We can see most of the successes for this range were of FT booster version. The next runner up was B4 booster version.

The worst was v1.1 booster version.

Payload vs Launch Outcome

We can see the Payloadmass vs Class for ranges 5000-10,000 kg

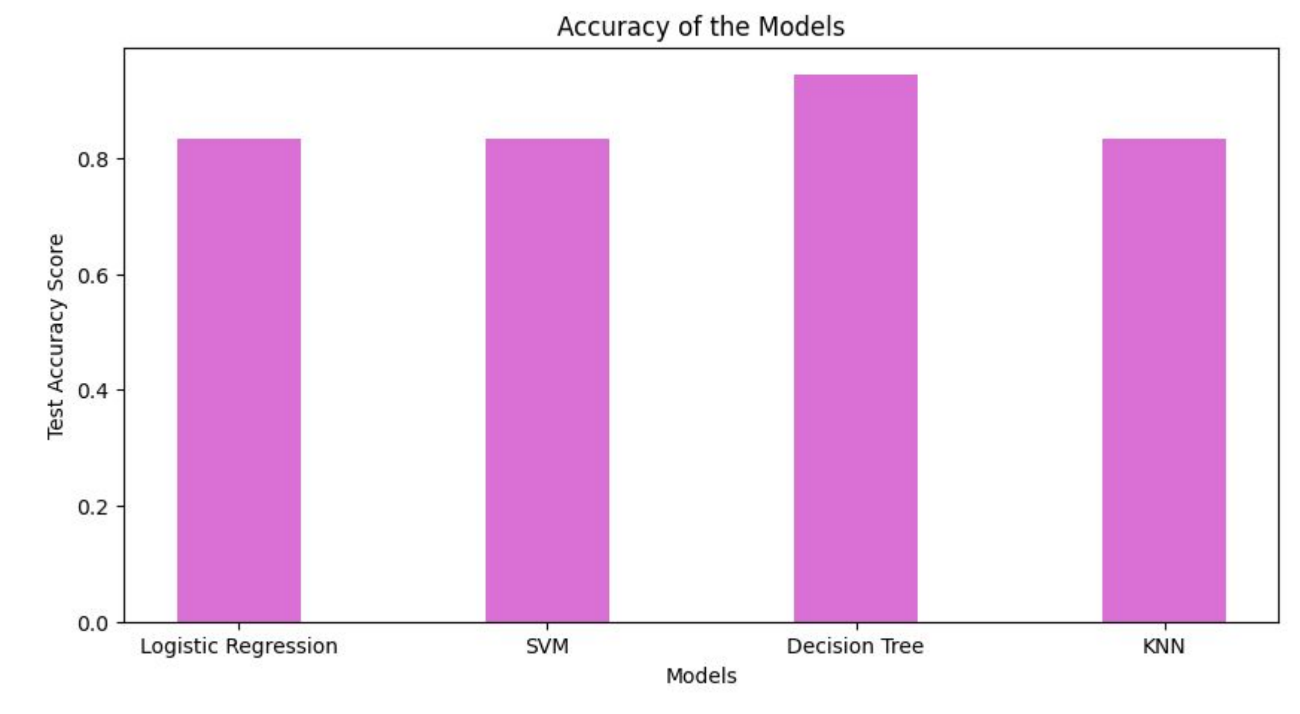
We can see that between 5000-7000, FT Booster was better, but most cases were failures.

However, after 7000kg, we do not have enough data to make an assessment of what booster version is best

Section 5

Predictive Analysis (Classification)

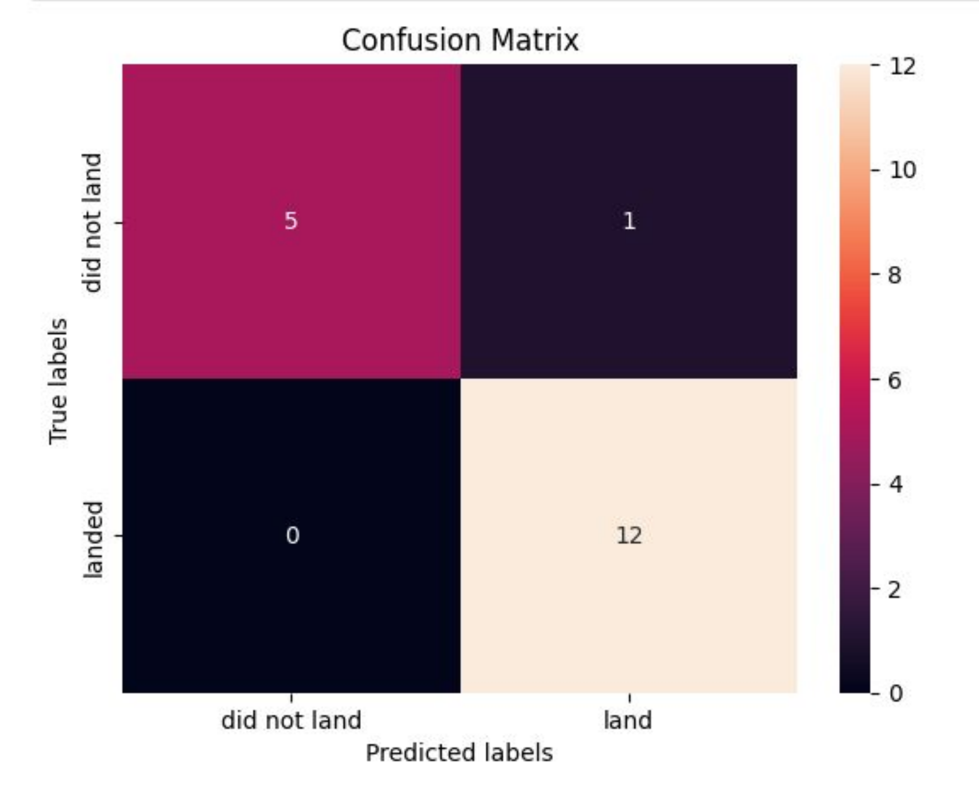
Classification Accuracy



We can see that on average, all of the models had an accuracy above 83.33%.

However, Decision Tree was most accurate at 94.44%

Confusion Matrix



We can see through the confusion matrix how the Decision Tree classifier labelled the data.

We can see that the model correctly predicted 12 that landed correctly, and 1 case where landed but predicted incorrectly.

We can also see that the model accurately predicted all the ‘did not land’ cases.

Conclusions

- Overall conclusions we can draw are:
    - The best launch site was KSC LC 39-A, with a 76.9% success rate.
    - As the number of flights increase, so did the success rate.
    - The heavier payload, the higher success rate.
    - The best orbit types are: ES-L1, GEO, HEO & SSO.
    - Decision Tree Classifier was the most accurate predictor for this dataset.

Appendix

- I had to set a random seed for my Decision Tree Classifier, because it was outputting different accuracy results based on the seed it picked. (np.random.seed(1))

Thank you!