IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Suleyman Tankisi

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
- Interactive maps with Folium
- Dashboards with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive maps and dashboards
- Predictive analysis results

Introduction

Project background and context

- SpaceX is leading the race for commercial space travel. The main reason for this is that it is relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. we are going to predict if SpaceX will reuse the first stage in order to consult another company that is thinking of enter this space race.

Problems we want to find answers to

- What are the optimal conditions which will allow SpaceX to achieve the best landing success rate ?
- What is the total average cost of a space launch for SpaceX?

Section 1

Methodology

Methodology

- Data collection methodology:

    SpaceX Rest API

    Web Scrapping from Wikipedia

- Perform data wrangling

    Dealing with missing data

    One Hot Encoding for classification 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

    Building, tuning, and evaluating classification models

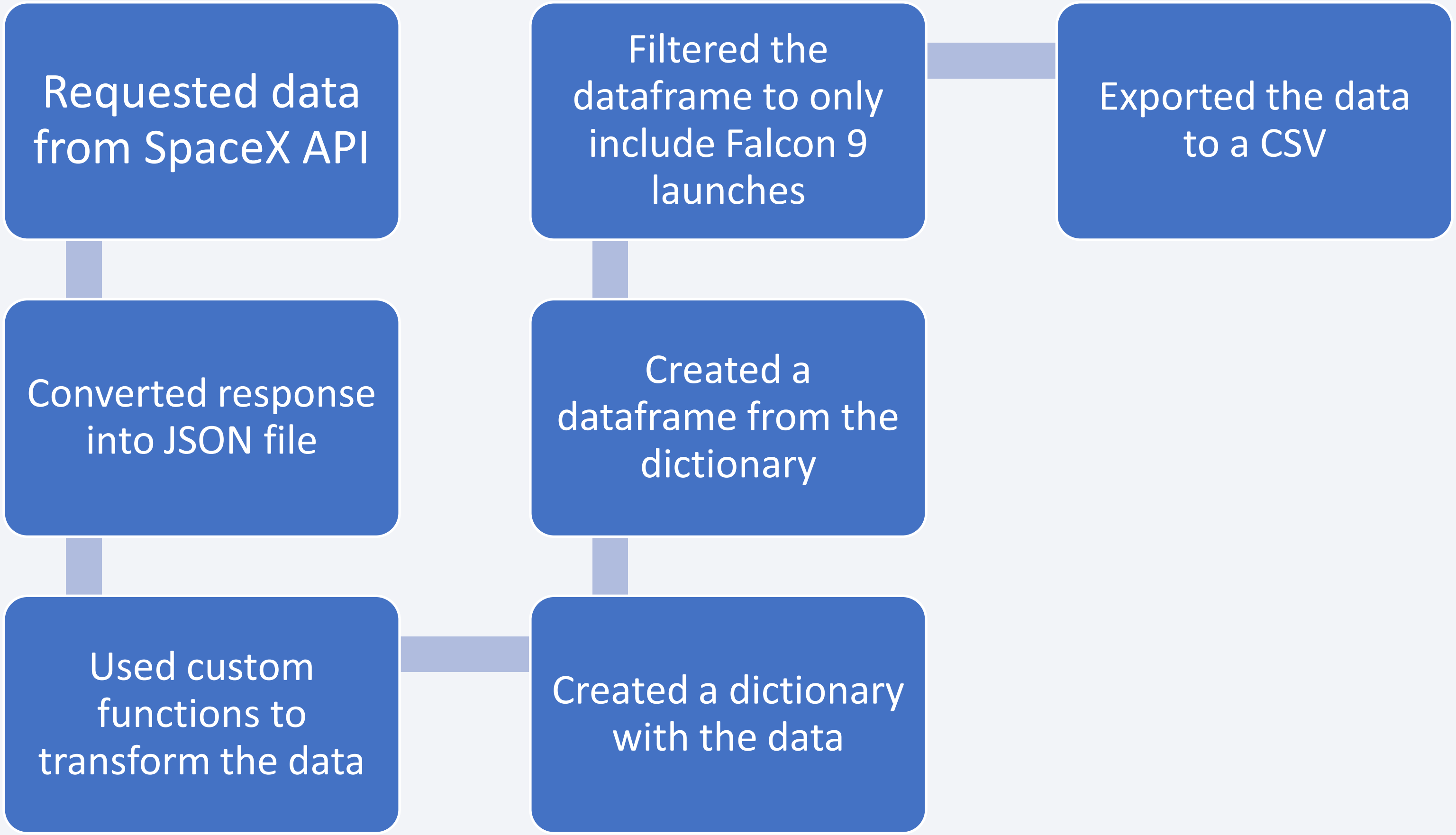
Data Collection

- I collected datasets from Rest SpaceX API and via Webscraping Wikipedia

  - The information obtained by the API is made up of Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude.

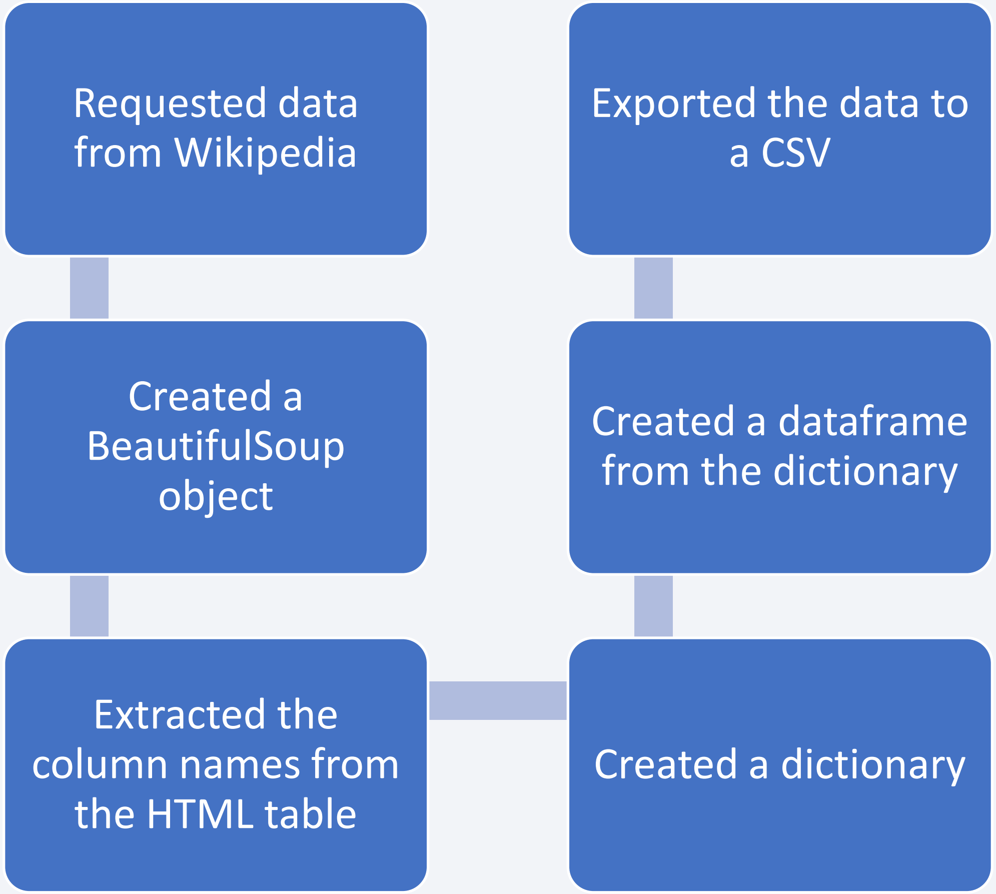
  - The information obtained by Webscraping Wikipedia is made up of Flight No., Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



[GitHub link: Data Collection - API's]

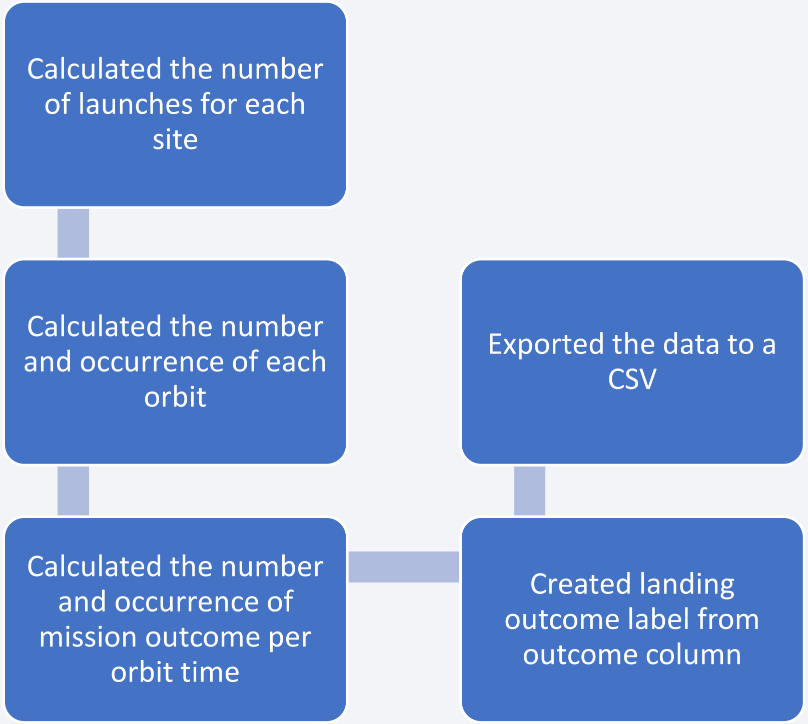
Data Collection - Scraping



[GitHub link: Data Collection - Webscraping]

Data Wrangling

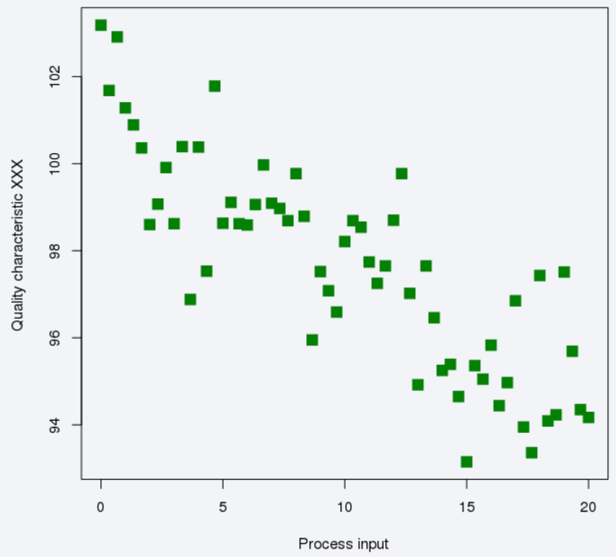
In the dataset, there are several cases where the booster did not land successfully. I transformed the string variables into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure.



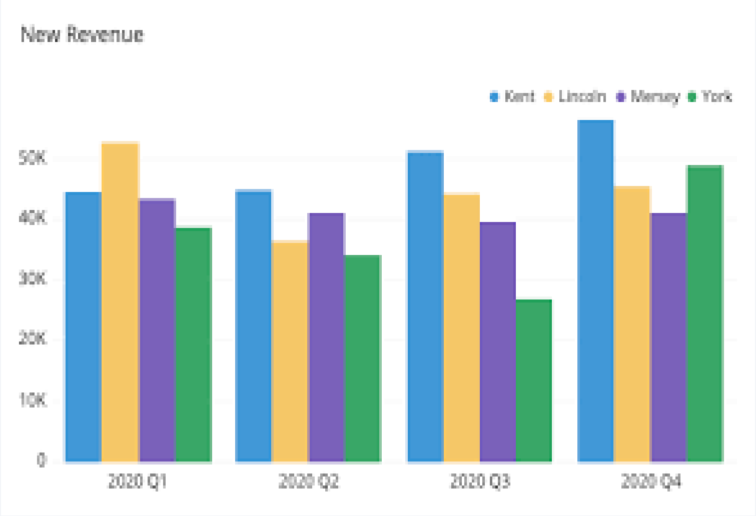
GitHub link: Data Wrangling

EDA with Data Visualization

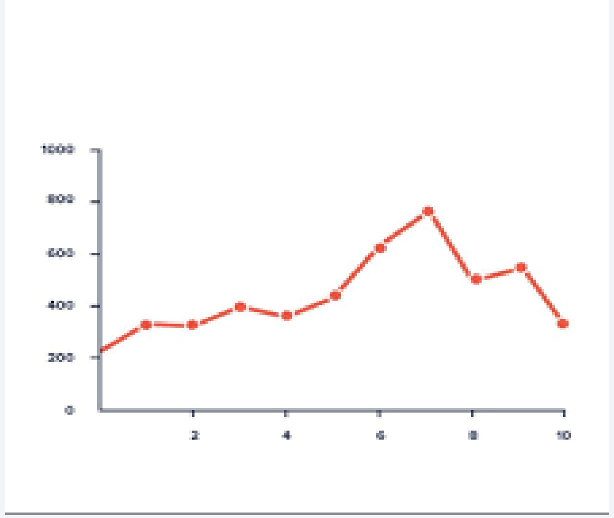
A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.



A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally.



A line chart or line graph or curve chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments.



GitHub link: EDA with visualisation

EDA with SQL

Performed SQL queries:

• Displayed the names of the unique launch sites in the space mission
• Displayed 5 records where launch sites begin with the string ‘CCA'
• Displayed the total payload mass carried by boosters launched by NASA (CRS)
• Displayed average payload mass carried by booster version F9 v1.1
• Listed the date when the first successful landing outcome in ground pad was achieved
• Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
• Listed the total number of successful and failure mission outcomes
• Listed the names of the booster versions which have carried the maximum payload mass
• Listed the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
• Ranked the count of landing outcomes between the dates of 04-06-2010 and 20-03-2017 in descending order

[GitHub link: EDA with SQL]

Build an Interactive Map with Folium

- Added marker at NASA Johnson Space Centre's coordinate with label showing its name
- Added marker at each launch site coordinates with label showing launch site name
- The grouping of points in a cluster to display multiple and different information for the same coordinates
- Added Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing.
- Added Markers to show distance between launch site to key locations and plot a line between them .

GitHub link: Visual analytics with Folium

Predictive Analysis (Classification)

Data preparation

•Loaded the dataset

•Normalised the data

•I Split data into training and test sets.

Model preparation

•4 different types of machine learning algorithms (LogReg, SVM, Decision Tree, and KNN models)

•Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Model evaluation

•I computed the accuracy for each model with a test dataset

•Plotted a Confusion Matrix for each model

Model comparison

•Compared the models based on their accuracy via tables

•Chose the model with the highest accuracy

[GitHub link: Machine Learning]()

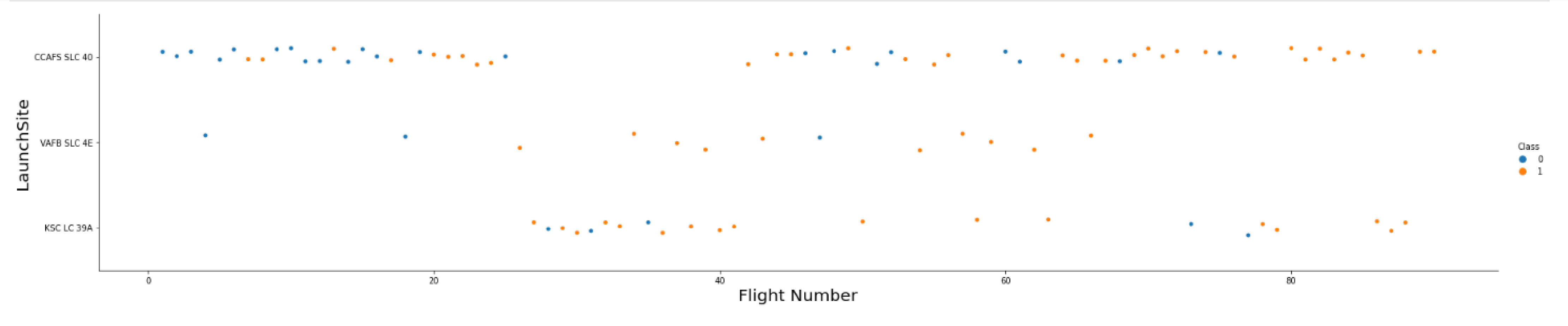
Results

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

Section 2

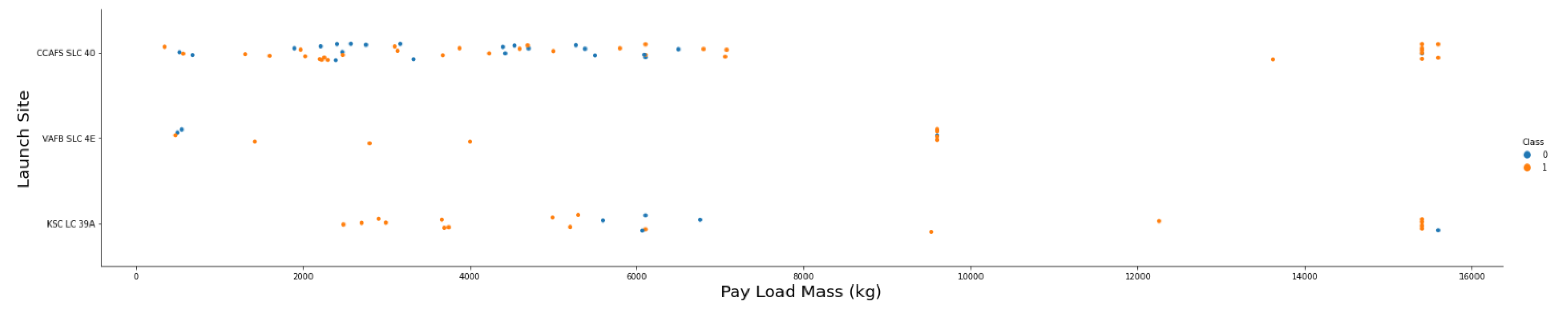
Insights drawn from EDA

Flight Number vs. Launch Site



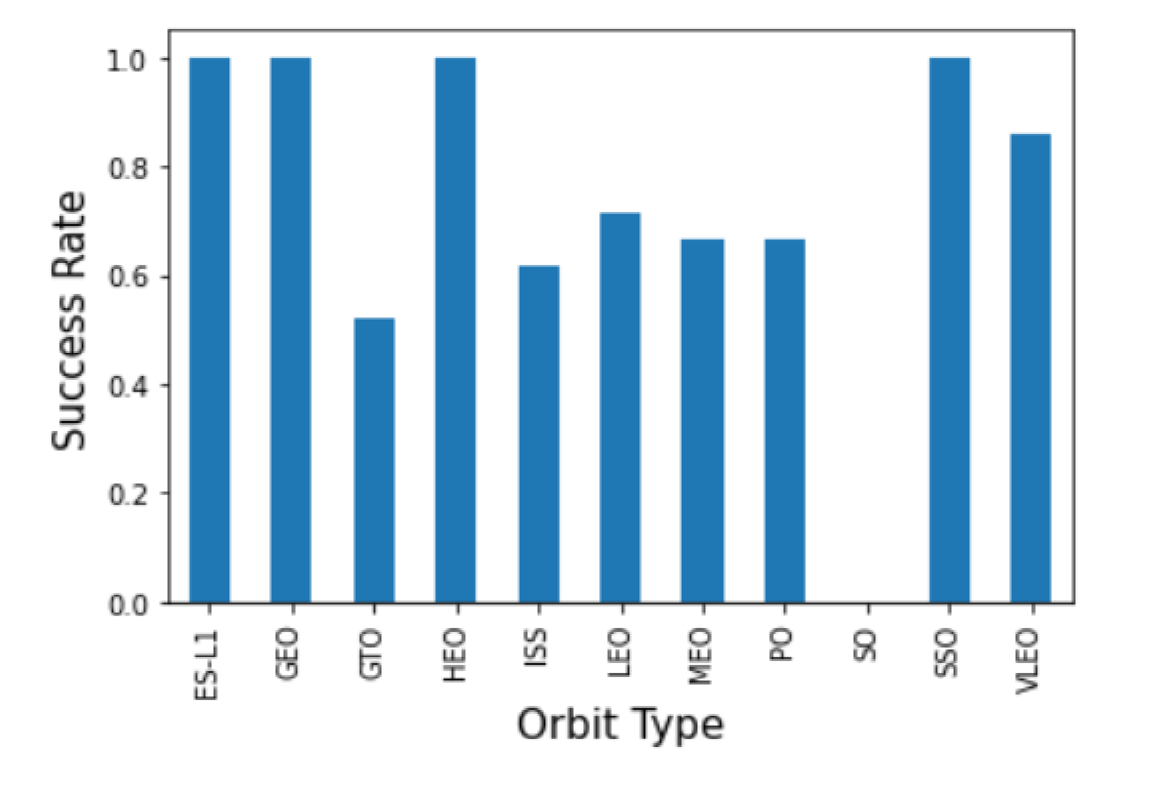
- The earlier flights failed while the later flights succeeded.
- For each launch site, the newer launches have a higher rate of success.

Payload vs. Launch Site



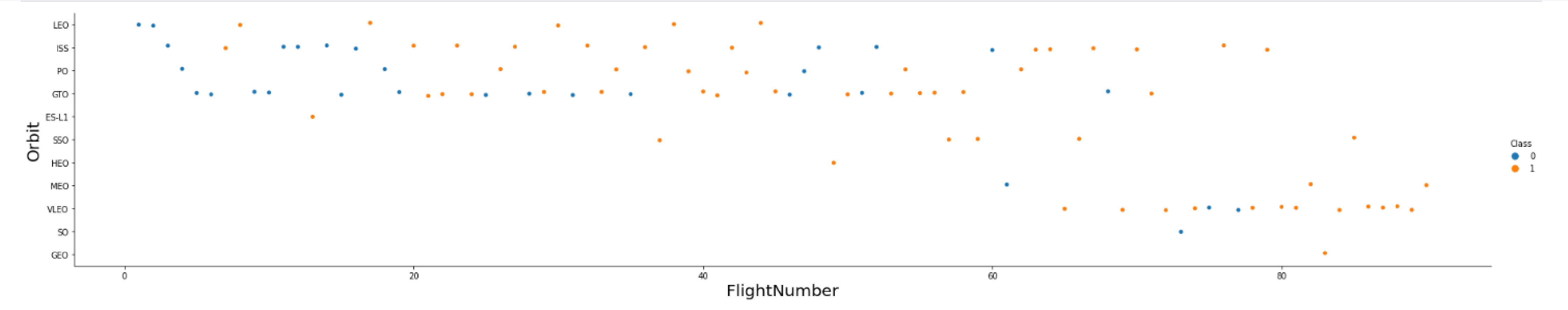
In general, the higher the payload mass, the higher the success rate. Although, if the payload mass is too heavy, this can lead to a failure.

Success Rate vs. Orbit Type



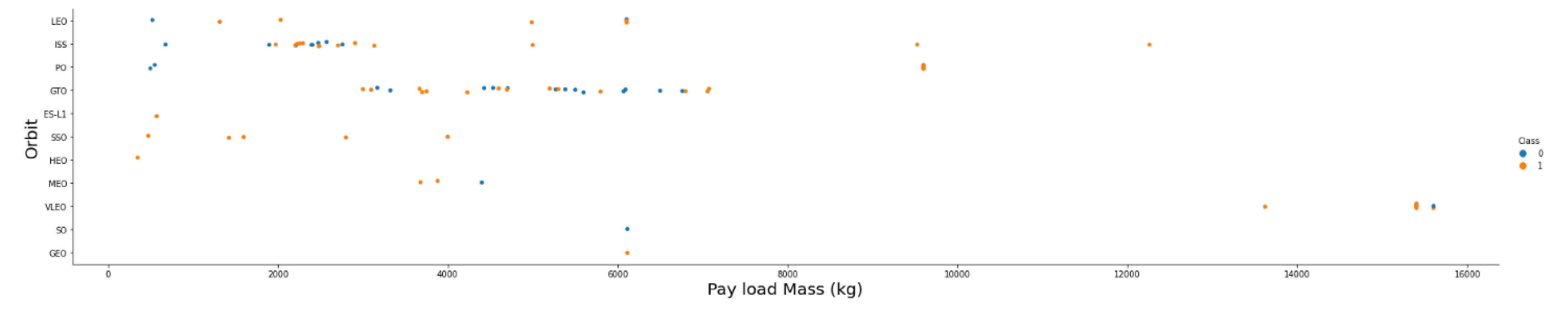
ES-L1, GEO, HEO, SSO, are all the winners with a 100% success rate.

Flight Number vs. Orbit Type



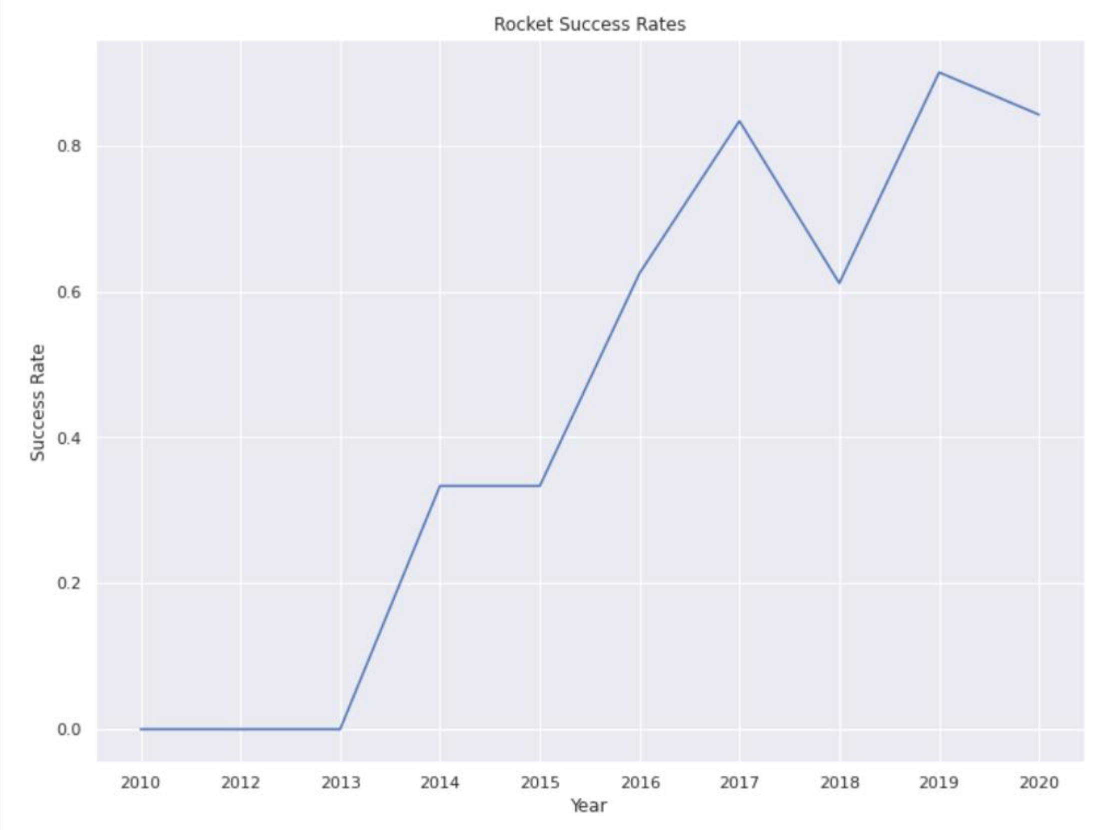
The 2 key takeaways are that the success rate increases with the number of flights for the LEO orbit and there is no relationship between flight number for the GTO orbit.

Payload vs. Orbit Type



The key takeaway is that the heavier payloads increase the success rate of the LEO orbit and a decreased payload increase the success rate of the GTO orbit launch.

Launch Success Yearly Trend



From 2013 onwards, the success rate increases

All Launch Site Names

Out[16]:

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

Displaying the names of the unique launch sites

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	None	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	None	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	None	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	None	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	None	677	LEO (ISS)	NASA (CRS)	Success	No attempt

5 records where launch sites begin with the string 'CCA'

Total Payload Mass

SUM("PAYLOAD_MASS__KG_")
45596

The total sum of the payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

The average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

MIN("DATE")
01-05-2017

The date when the first successful ground landing was achieved.

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

The names of the booster versions where landing was successful and payload mass is between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

SUCCESS	FAILURE
100	1

There were 100 successful launches and 1 failure.

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

The names of the booster versions which have carried the maximum payload mass.

2015 Launch Records

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

The failed landing outcomes for the year 2015

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

Landing Outcome	Total Count
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

Ranking the count of landing outcomes between the dates 2010-06-04 and 2017-03-20 in descending order

Section 3

Launch Sites Proximities Analysis

Folium Map – Location of the Launch Sites

The 2 launch sites are on the coast of the United States

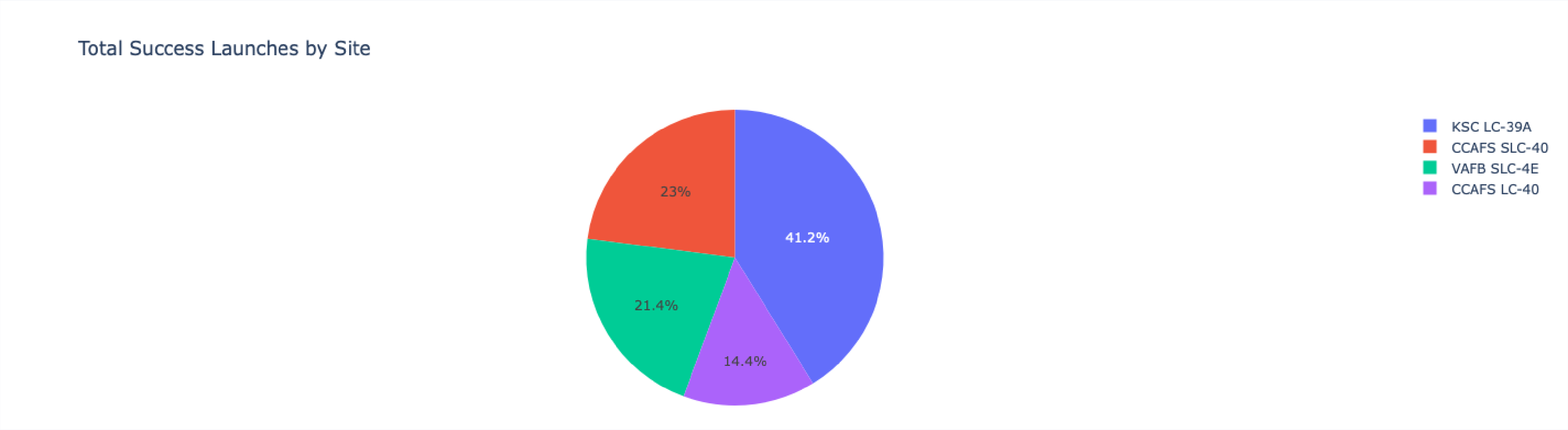
Folium Map – Colored Markers

Green markers represent successful launches and the Red markers represent failed launches

Section 4

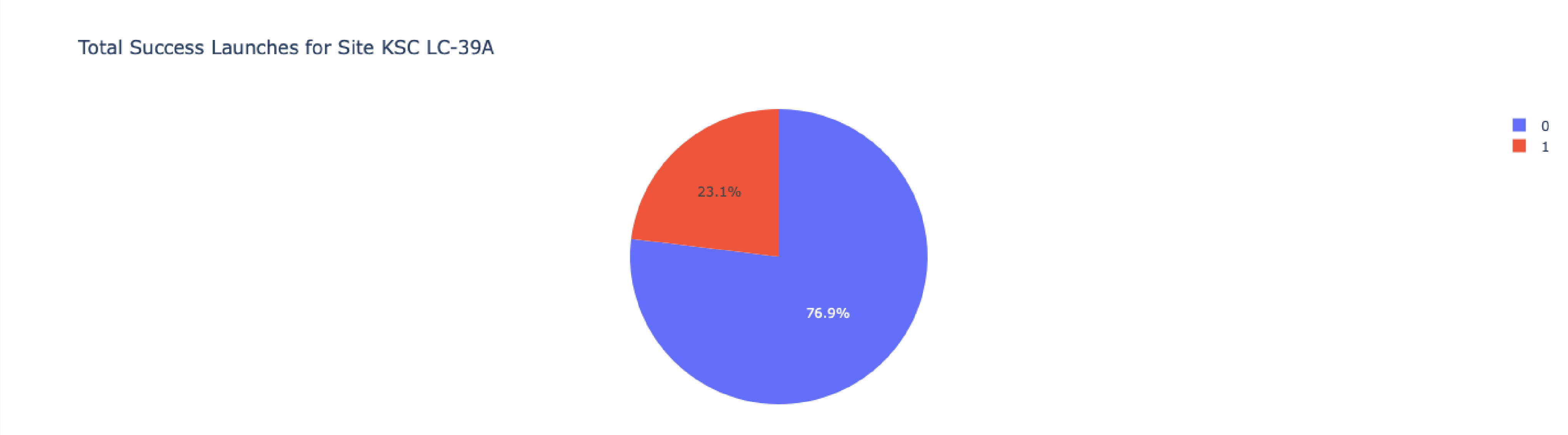
Build a Dashboard with Plotly Dash

Launch success rates



The KSC LC-39A launch site has the most successful launches.

Launch success rate for KSC LL-39A



KSC LL-39A is the most successful launch site and has achieved a 76.9% success rate while getting a 23.1% failure rate.

Section 5

Predictive Analysis (Classification)

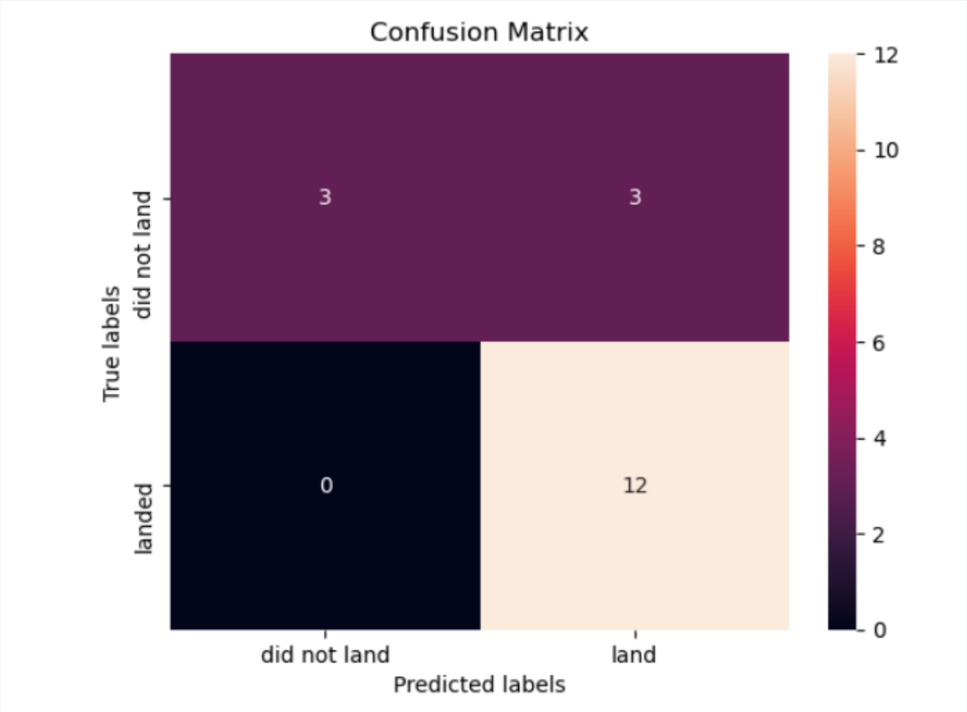
Classification Accuracy

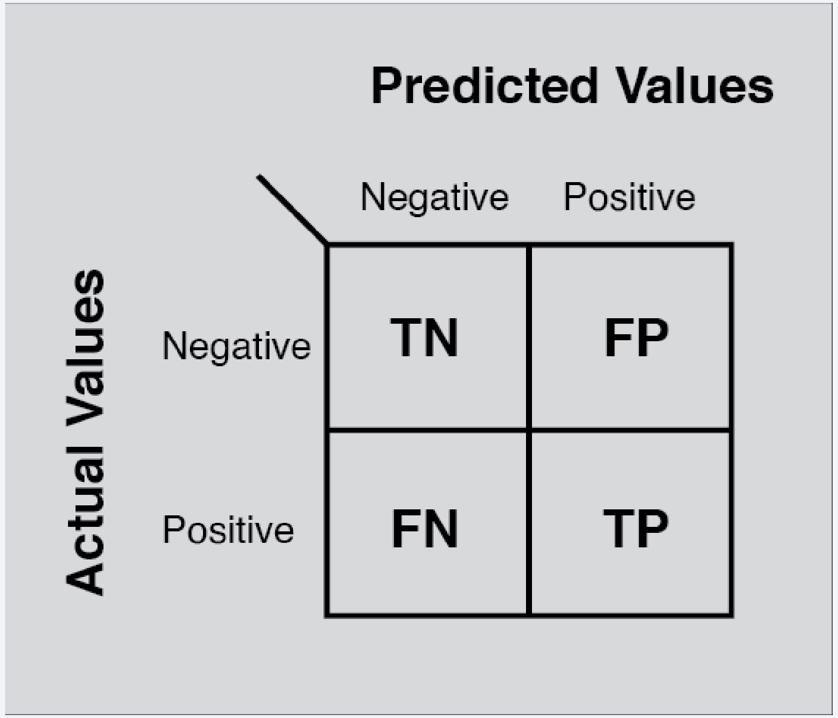
	Accuracy Train	Accuracy Test
Tree	0.888889	0.833333
Logreg	0.847222	0.833333
Svm	0.847222	0.833333
Knn	0.847222	0.833333

For the test sets, every method had the same result so we can not say which is best.

For the training sets, we can see that the Decision Tree model gives the best result.

Confusion Matrix





We can see that the decision tree classifier can distinguish between the different classes. The main problem is that we can get a lot of false positives.

Conclusions

- The success rate of launches increases over time
- KSC LC-39A has the highest success rate of the launches from all the sites.
- The orbits with the highest success rates are GEO, HEO, SSO, ES L1.
- The Tree Classifier Algorithm is the best, most accurate Machine Learning model to use for this dataset.
- We can successfully predict how much a launch will cost on average.

Thank you!