IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Issei Kuzuki
01/09/2022

Outline

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

Executive Summary

- Collected data from SpaceX REST API and by web scraping from Wikipedia pages and using the Beautiful soup library
- Cleaned the data and included a 'class' column to indicate success of landings
- Performed exploratory data analysis EDA using visualization libraries matplotlib, seaborn and by using SQL queries with magic SQL to get useful insights from the data
- Created interactive visual analytic maps using Folium and a dashboard using Plotly dash to help visualize the data and geographic information
- Built classification machine learning models using the scikit-learn library to perform predictive analysis

Introduction

- The aim of this project is to determine the price of each SpaceX rocket launch
- SpaceX is a successful company focused on spaceflight
- SpaceX's Falcon 9 rocket launches can land and reuse their first stage which saves huge costs for each launch
- However, the first stage does not always land due to accidents and more often intentional sacrifices of the first stage
- Using data science techniques, we want to build models and visualizations which will allow us to predict whether a launch will land its first stage
- This information can subsequently be used to estimate the price of launches

Section 1

Methodology

Methodology

- Data collection methodology:
  - API and web scraping to collect data
- Perform data wrangling
  - Sampling, cleaning data and dealing with nulls
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build classification models and fit them to training data to optimize parameters

Data Collection

- Data was collected from 2 sources, the SpaceX REST API and Wikipedia pages
- The SpaceX REST API contains data about launches including rocket information, payload delivered as well as landing outcome
- Data was scraped from the Wikipedia pages which contains information on Falcon 9 Launch data
- The raw data extracted was wrangled to clean and transform it into a useful form in order to perform analysis on it at a later stage

Data Collection – SpaceX API

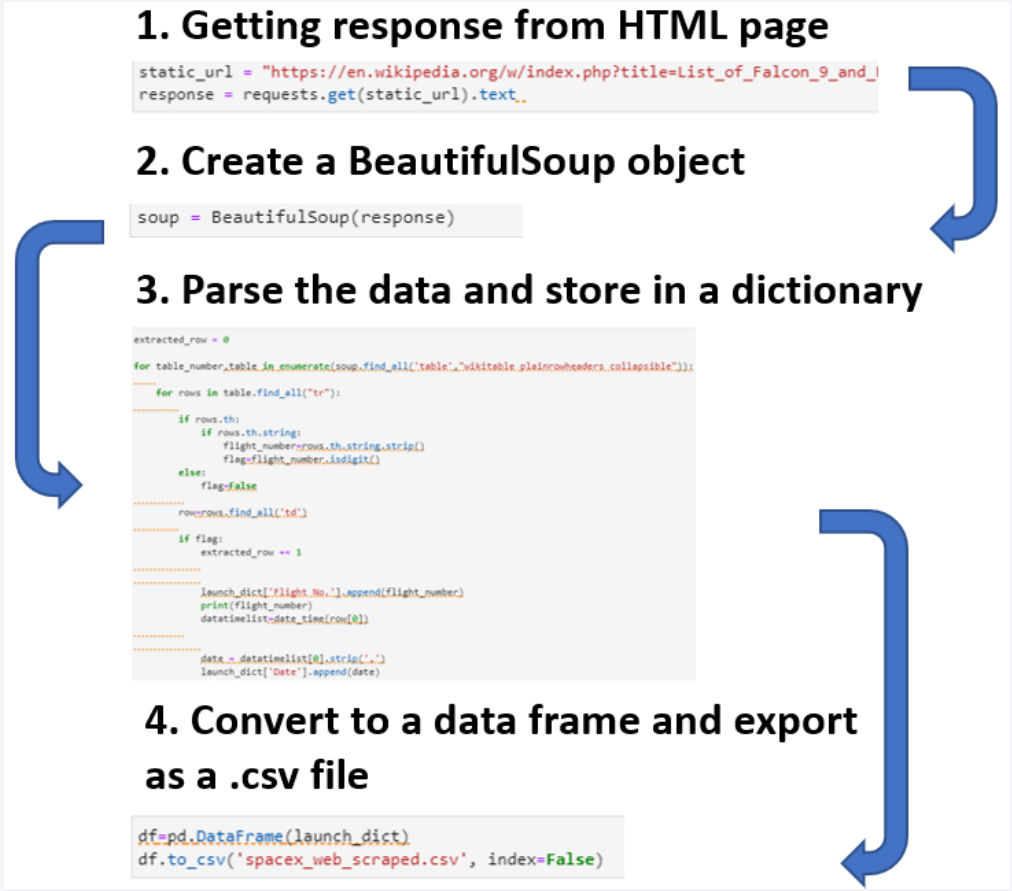
- Data was requested from the SpaceX API
- Custom functions were used to extract key data after cleaning

https://github.com/ikuzuki/IBM-Data-science-capstone/blob/main/Week%201%20tasks/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Data was scraped from a Wikipedia webpage
- Key information was extracted from HTML tables
- The data was stored in a dictionary which was converted to a .csv file

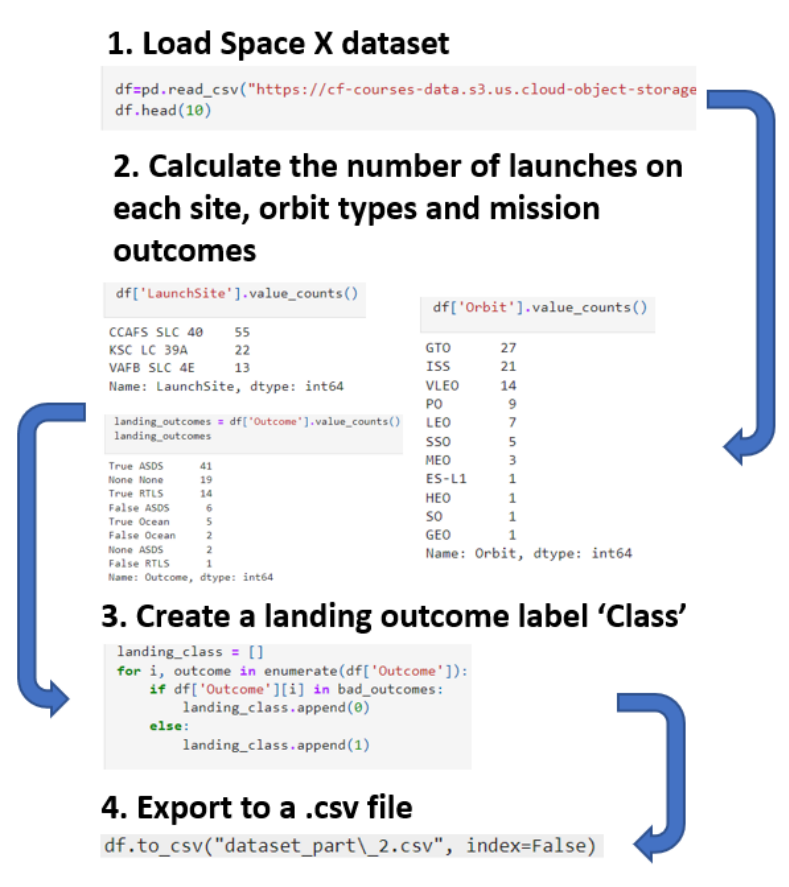
https://github.com/ikuzuki/IBM-Data-science-capstone/blob/main/Week%201%20tasks/jupyter-labs-webscraping%20(2).ipynb



Data Wrangling

- .value_counts() was used to group mission outcomes
- A new column ‘Class’ was created, a binary variable indicating whether the second stage landed successfully
- Additionally, in the data collection section the Payload mass for rows with no data (Nulls) were replaced with the mean using .mean() and .replace()

https://github.com/ikuzuki/IBM-Data-science-capstone/blob/main/Week%201%20tasks/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Exploratory data analysis was performed to spot patterns for the success of landings
- Scatter plots were used to visualize how flight number, launch site and payload mass influenced success rate
- A bar chart was used to visualize success rates for different orbit types
- A line chart was used to visualize the trend of average success rate with years past

https://github.com/ikuzuki/IBM-Data-science-capstone/blob/main/Week%202%20tasks/jupyter-labs-eda-dataviz.ipynb

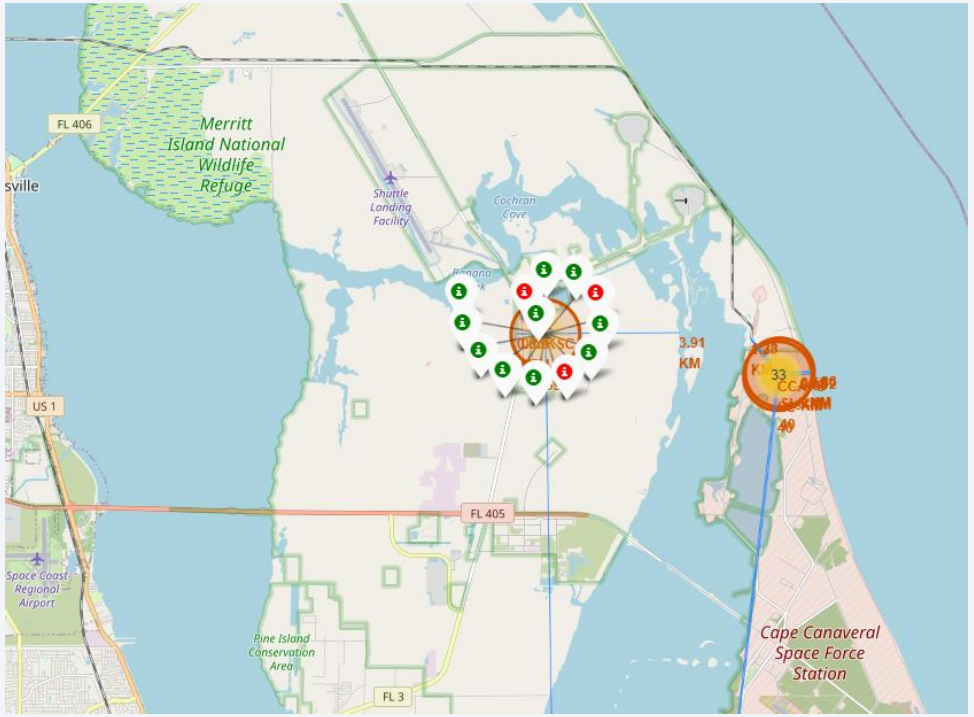
EDA with SQL

- The SELECT and FROM SQL query was used to extract certain parts of the dataset
- The WHERE clause followed by statements such as LIKE, LIMIT were used to set conditions on the data required
- Other statements which were used include GROUP BY (group column by categories), SUM (sum columns), COUNT (show frequencies)
- Sub queries were used when the MAX statement was needed in the WHERE clause

https://github.com/ikuzuki/IBM-Data-science-capstone/blob/main/Week%202%20tasks/jupyter-labs-eda-sql-coursera_sqllite%20(2).ipynb

Build an Interactive Map with Folium

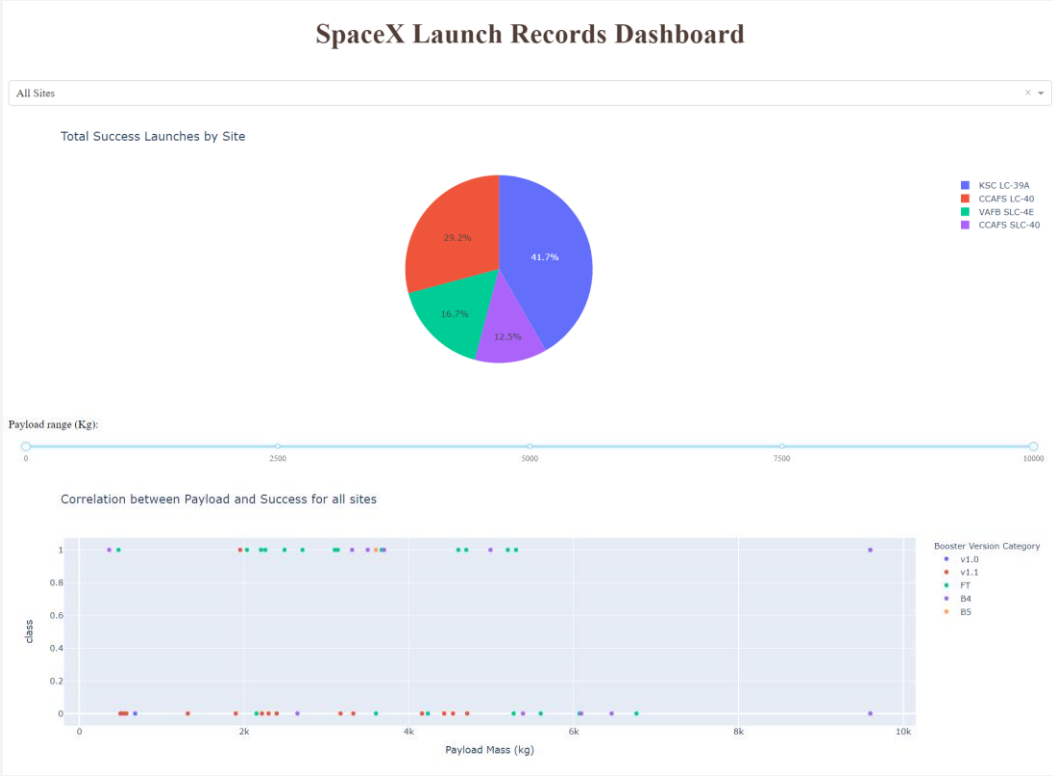
- A Folium map was used to give a visual insight into the launches from different sites
- A marker and a circle were added to the Folium map at the location of each launch site
- A Marker Cluster object was created containing colored markers indicating successful/failed landings at a particular launch site
- Lines and distance markers were added to show the closest distance from each launch site to railways, highways, coastlines and cities



https://github.com/ikuzuki/IBM-Data-science-capstone/blob/main/Week%203%20tasks/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- A dropdown was added to select a certain launch site and on option for ‘all sites’
- A pie chart was added which shows the success rate of launches at a particular site or the total successful launches by site for the ‘all sites’ selection
- A scatter plot was added plotting ‘class’ (success or fail) against payload mass with colored labels for each site
- A range slider was added to select the payload mass range, this helps to visualize which ranges have the best success rate for launches

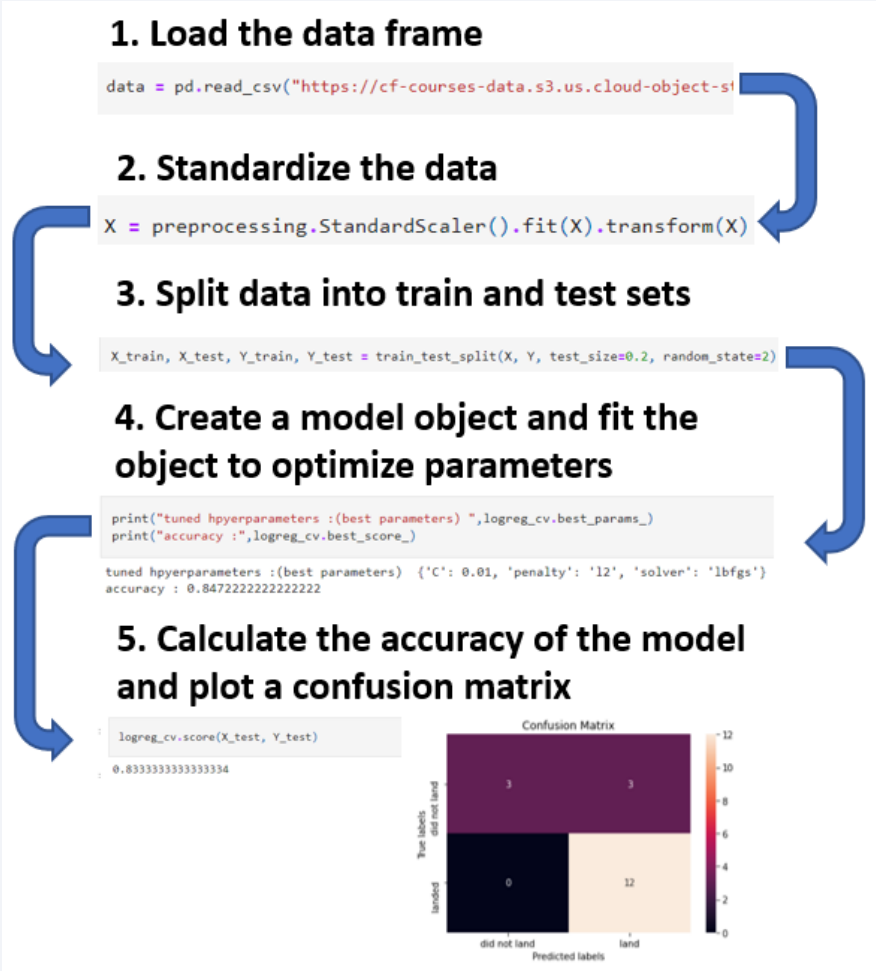


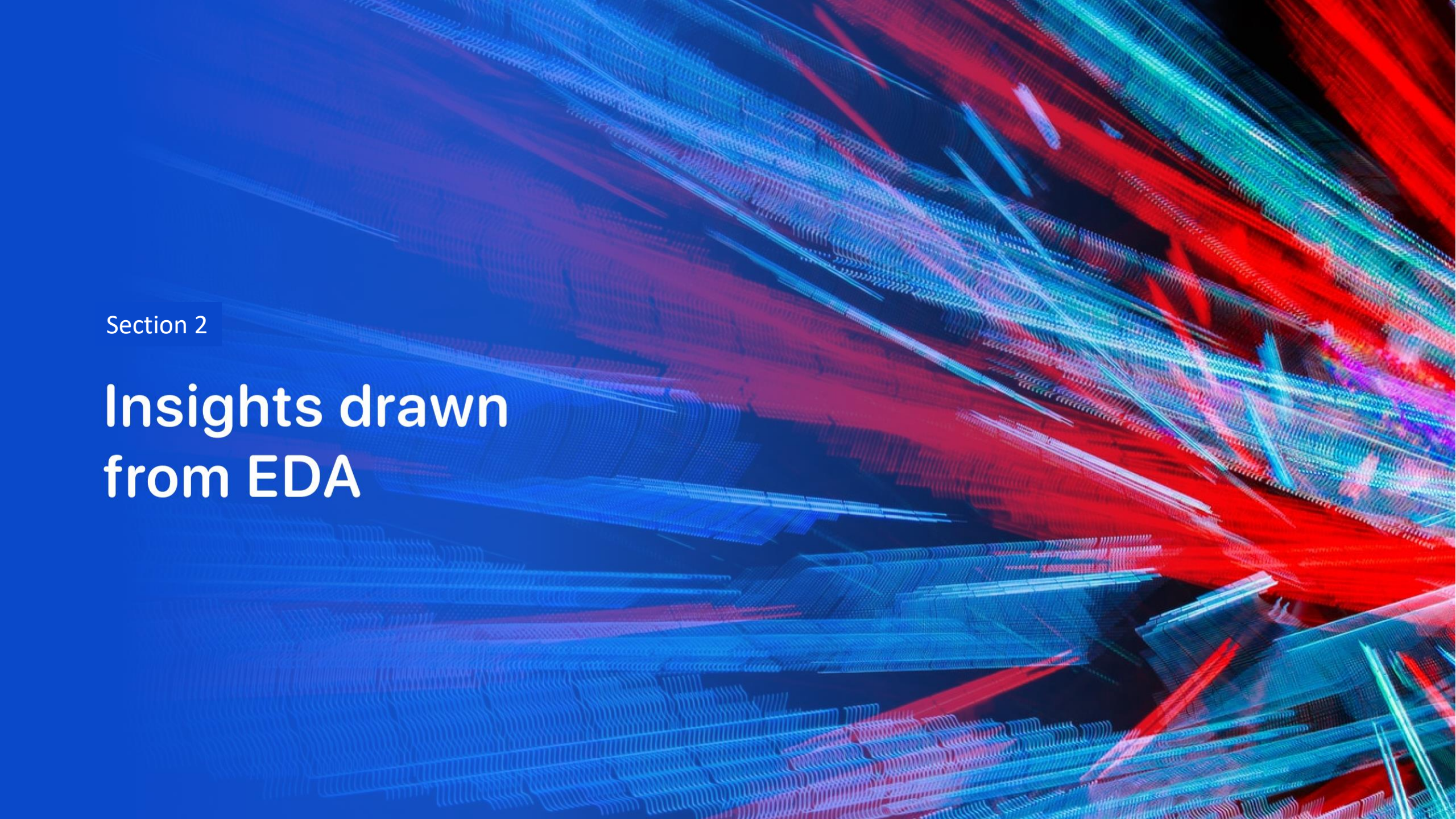
https://github.com/ikuzuki/IBM-Data-science-capstone/blob/main/Week%203%20tasks/capstone_dash.py

Predictive Analysis (Classification)

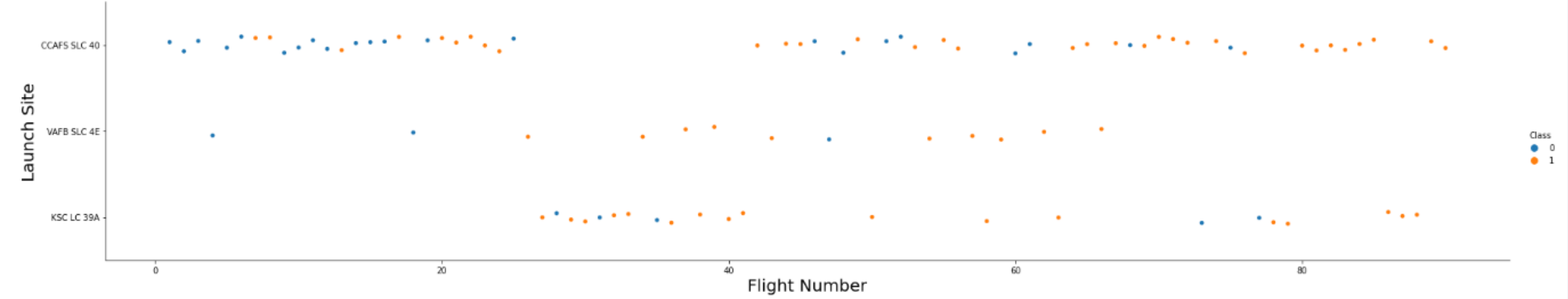
- 4 different models were built: logistic regression, support vector machine, decision tree classifier and k nearest neighbors
- Fitting the model gave the accuracy of the training set for the best parameters for the model
- The confusion matrix can be used to visualize whether the model has problems with false positives or negatives

https://github.com/ikuzuki/IBM-Data-science-capstone/blob/main/Week%204%20tasks/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb



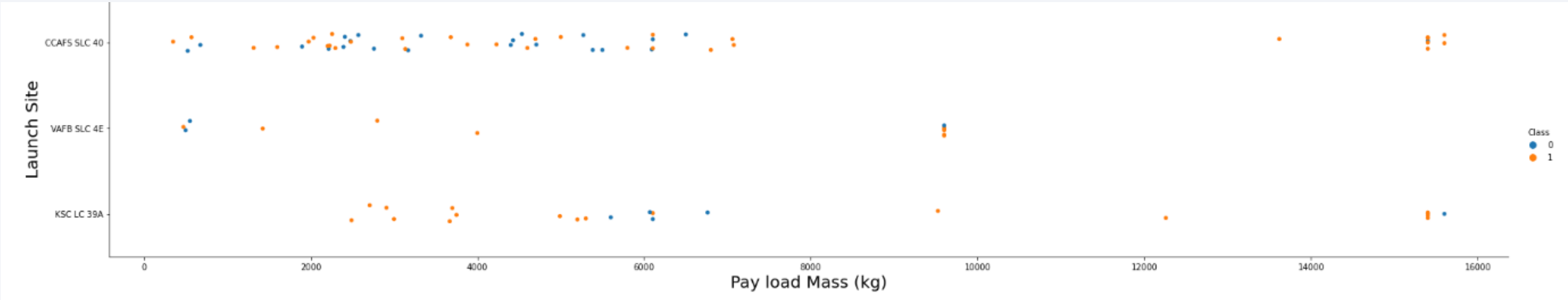

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



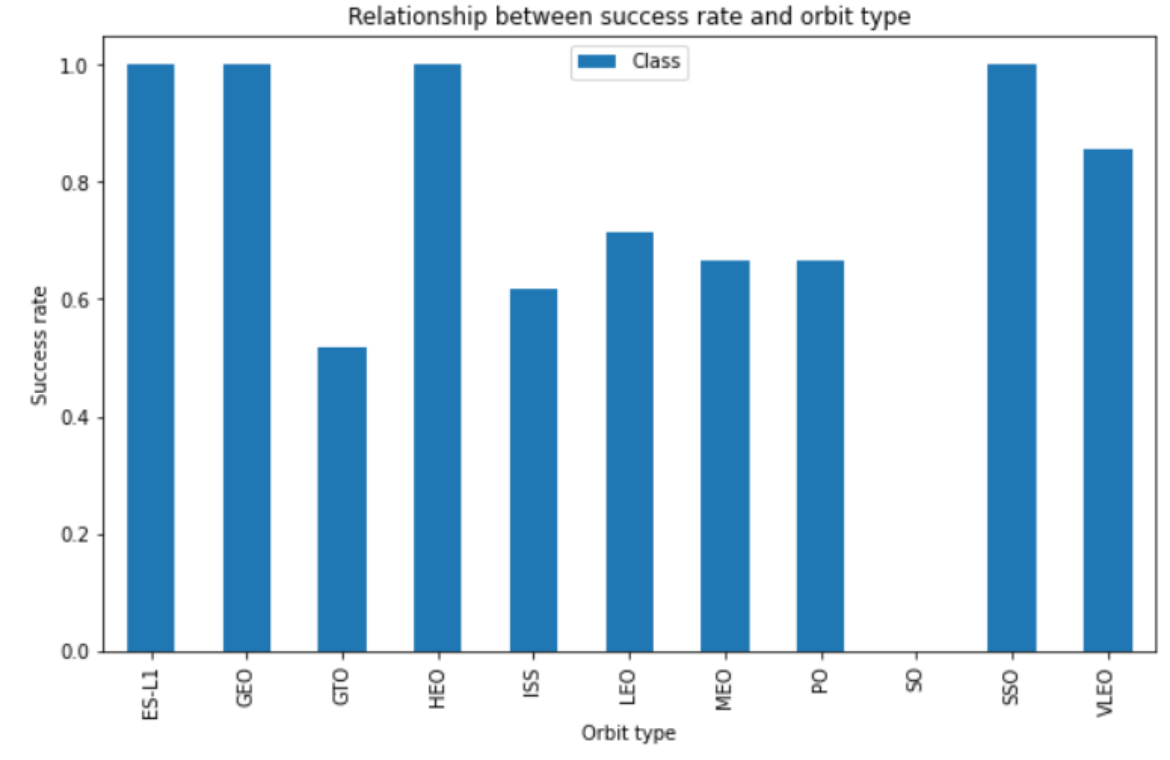
- Initially, mostly the CCAFS SLC 40 site was used with a lot of failures
- However, for later flight numbers, all launch sites are used with similar success rates

Payload vs. Launch Site



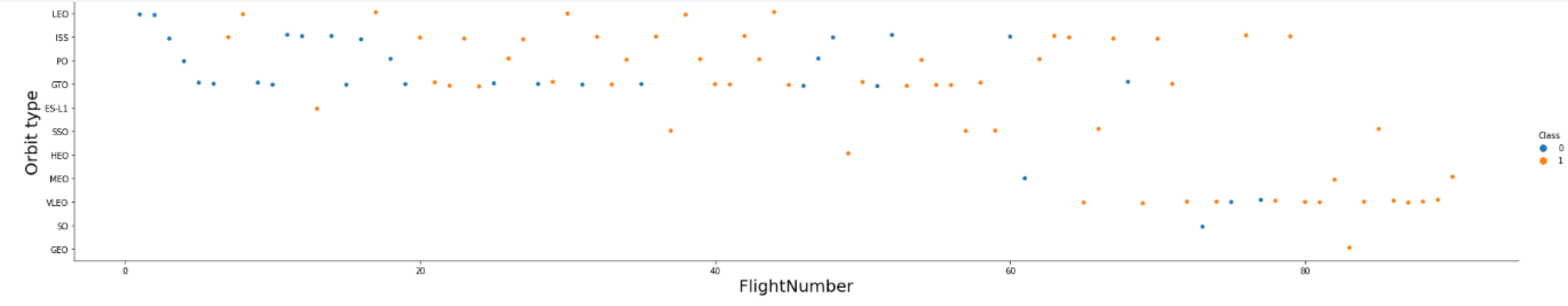
- The VAFB SLC 4E site has no rockets launched for heavy payload mass (greater than 10000)
- The KSC LC 39A site has no rockets launched for very light payload mass (less than 2000)

Success Rate vs. Orbit Type



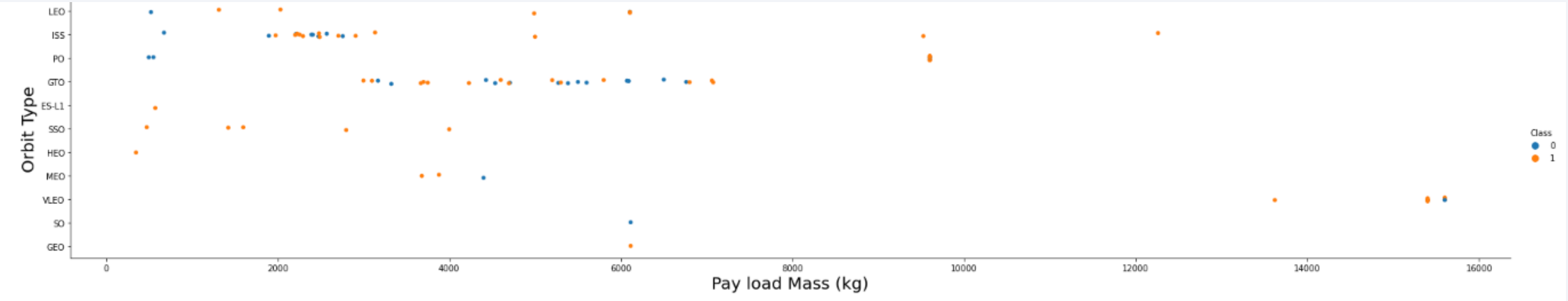
- The orbit types ESL1, GEO, HEO and SSO all have a 100% success rate
- SO has a 0% success rate

Flight Number vs. Orbit Type



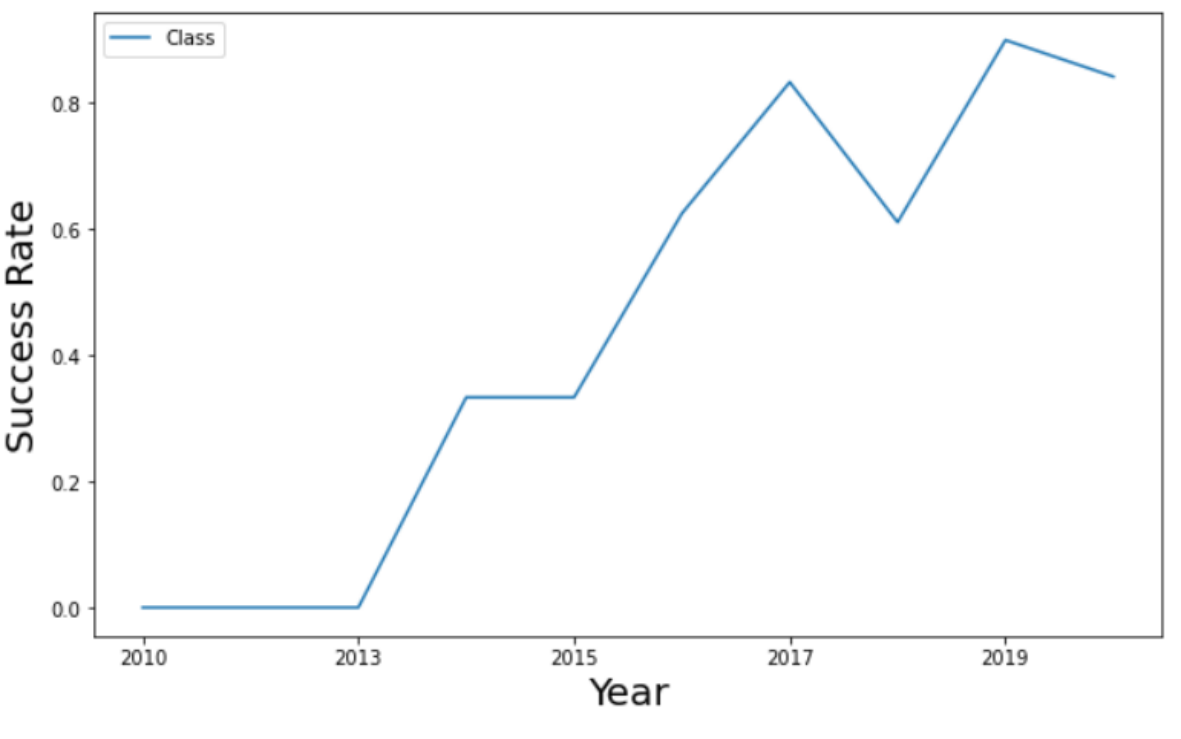
- For LEO orbit, the success rate appears to improve with flight number
- For most other orbit types there is no clear pattern, especially the GTO orbit

Payload vs. Orbit Type



- With heavy payloads the landing rate is high for Polar, LEO and ISS
- There is no clear pattern with GTO

Launch Success Yearly Trend



- Barring a few blips, the average success rate shows a generally increasing trend from 2013 to 2020

All Launch Site Names

- There are 4 unique launch sites

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- There are 2 different launch site names which begin with ‘CCA’
- 5 different records are shown

Total Payload Mass

- The total payload mass for boosters launched by NASA (CRS) is displayed

Total payload mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by the booster version F9 v1.1 was calculated

Average payload mass
2928.4

First Successful Ground Landing Date

- This is the date of the first successful landing outcome on a ground pad

Date of first successful landing
20151222

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are 4 boosters which have achieved a successful drone ship landing and have a payload mass between 4000 and 6000
- The names of these boosters are shown

Booster Name
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- The different mission outcomes and their corresponding frequencies are listed
- There is only 1 failed mission out of 101 total missions

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

Boosters Carried Maximum Payload

- The maximum payload is 15600
- Listed are the names of the booster versions which have carried this 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

2015 Launch Records

- The records for failed landing outcomes in drone ship are listed
- The records are restricted to the year 2015

Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

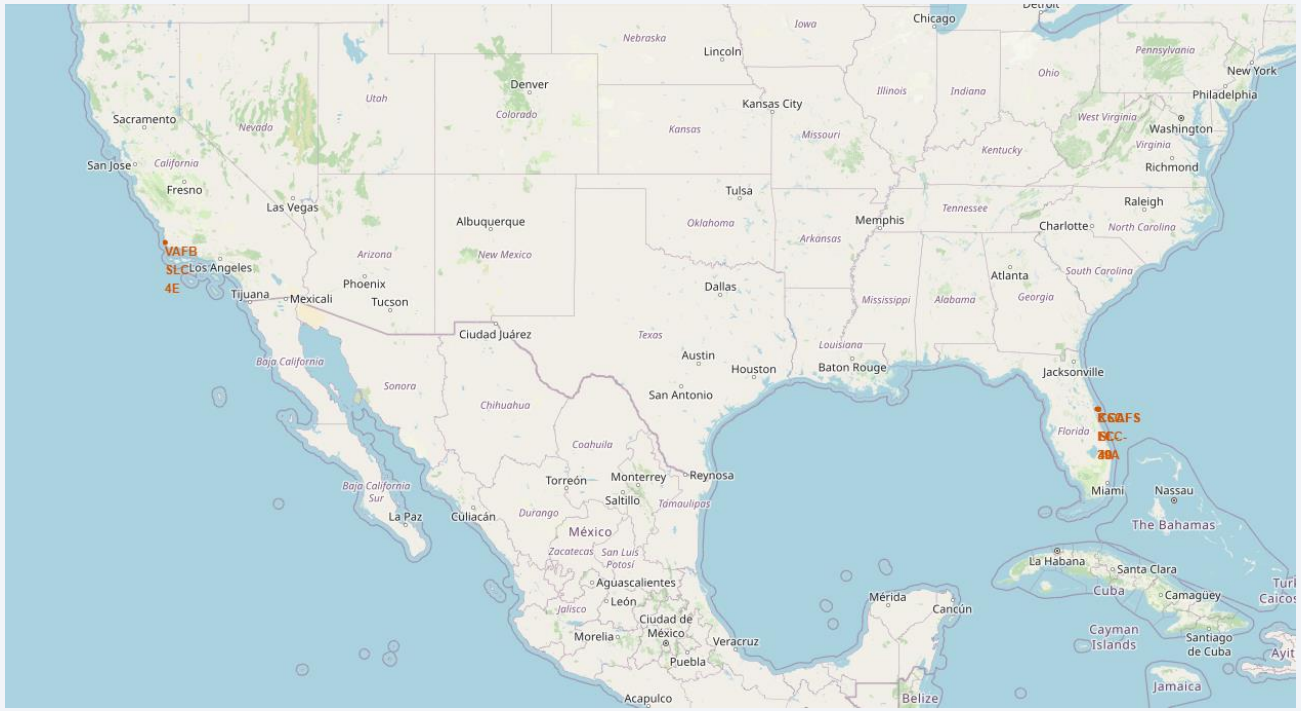
- The count of successful landing outcomes that have occurred between the selected dates have been ranked in descending order
- There are only 2 types of successful landing outcomes: drone ship and ground pad

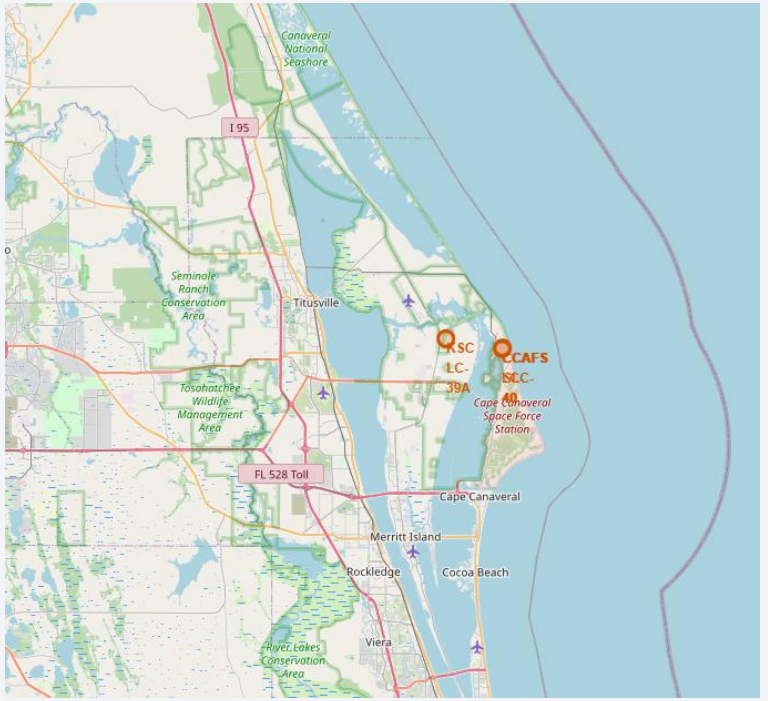
Landing Outcome	Count
Success (drone ship)	5
Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

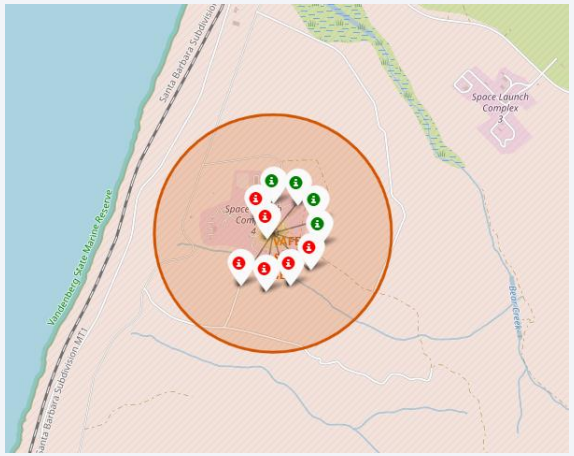
Global map of launch sites



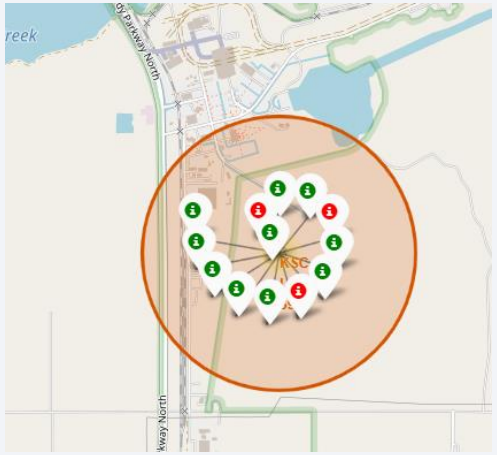


- All the launch sites are near the equator line, this is because land at the equator has extra rotational speed which gives the rocket a speed boost on launch
- All the launch sites are near the coast, this is because in case of intentional or unintentional explosions, the ocean is largely uninhabited so is much safer for debris to fly off into

Success/failed landings at launch sites

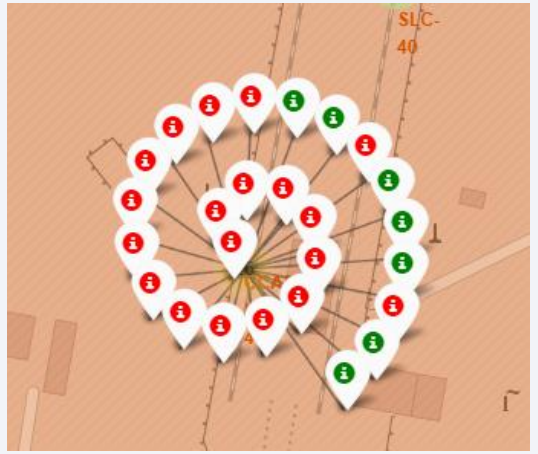


VAFB SLC-4E



KSC LC-39A

CCAFS SLC-40

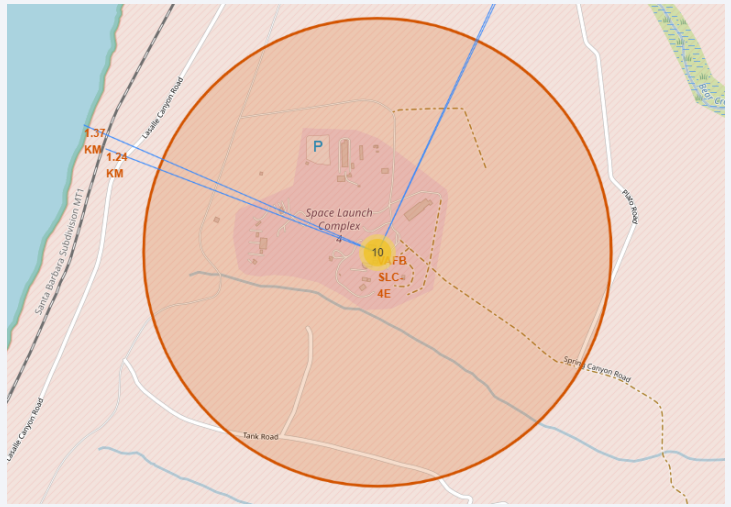


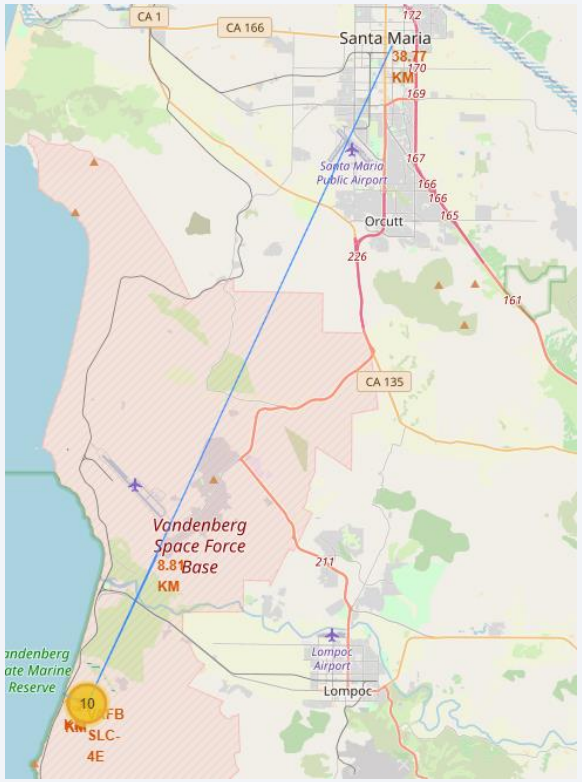
CCAFS LC-40

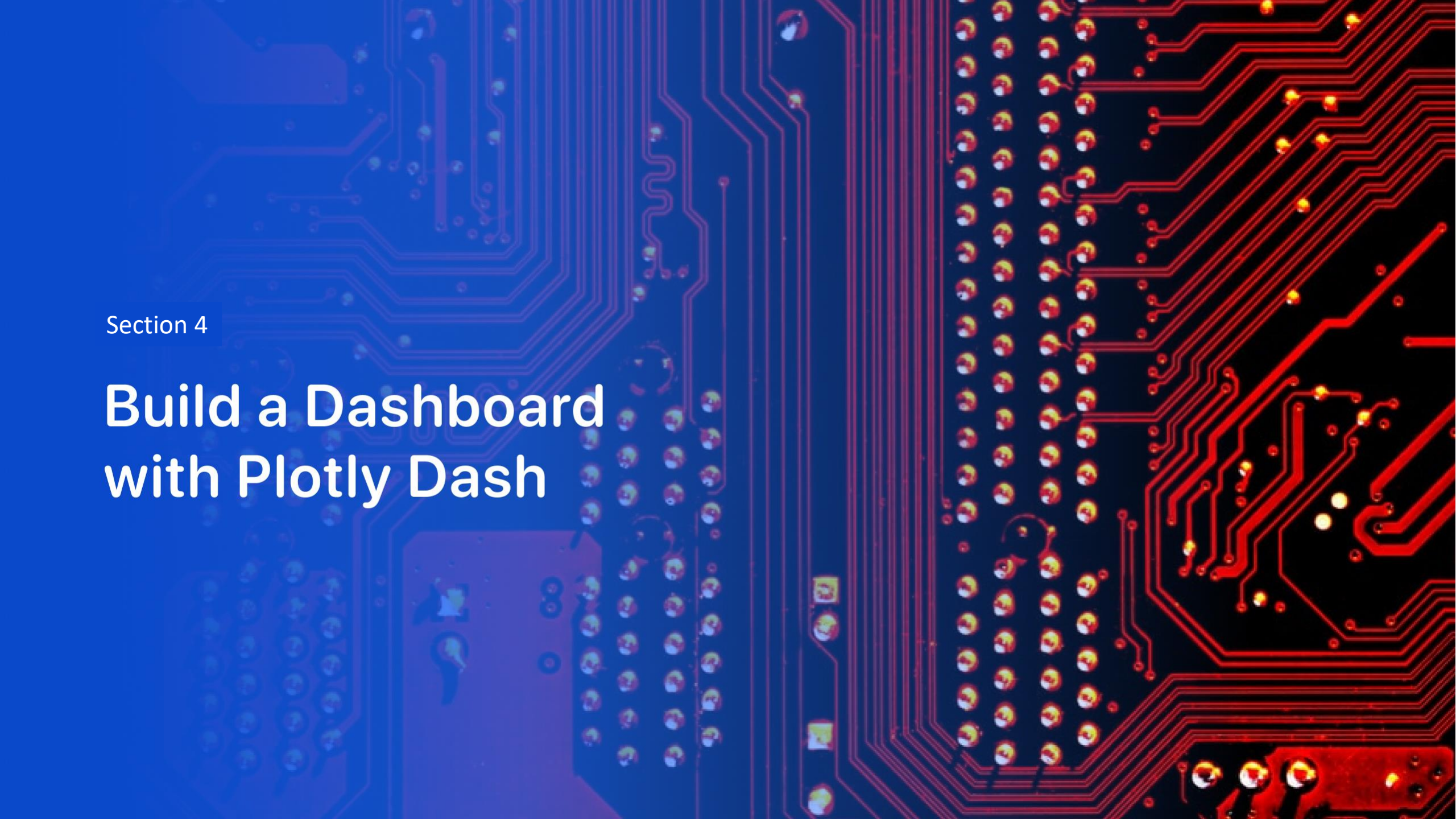
- The green markers indicate successful landings while the red markers indicate failures
- KSC LC-39A has the best success rate for landings as well as the highest count of successful landings
- CCAFS LC-40 has the most failed landings

Proximity of launch site to key features

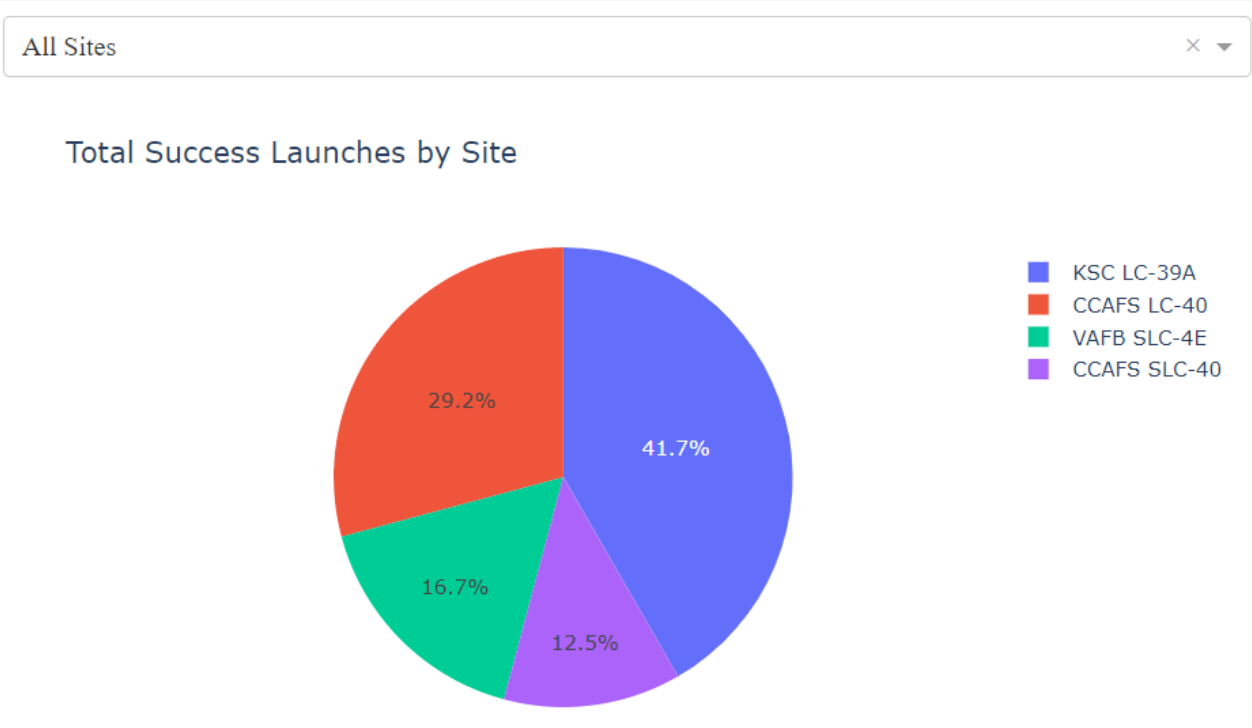
- The lines and labels indicate the distance from the launch site VAFB SLC-4E to the nearest railway, highway, coastline and city
- Launch sites are in close proximity to highways and railways in order to easily transport workers, heavy parts and equipment
- They are also located far from cities in order to minimize the risk to the population in case of failures





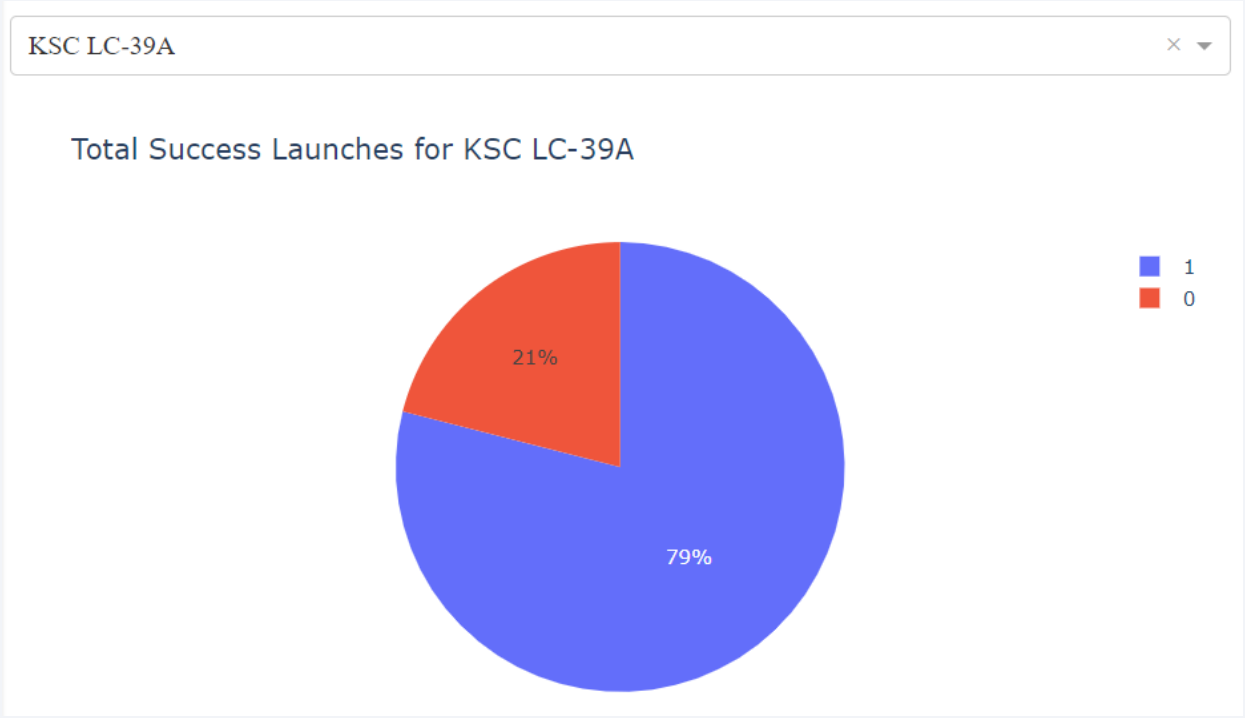

Section 4
Build a Dashboard with Plotly Dash

Successful landings by launch site



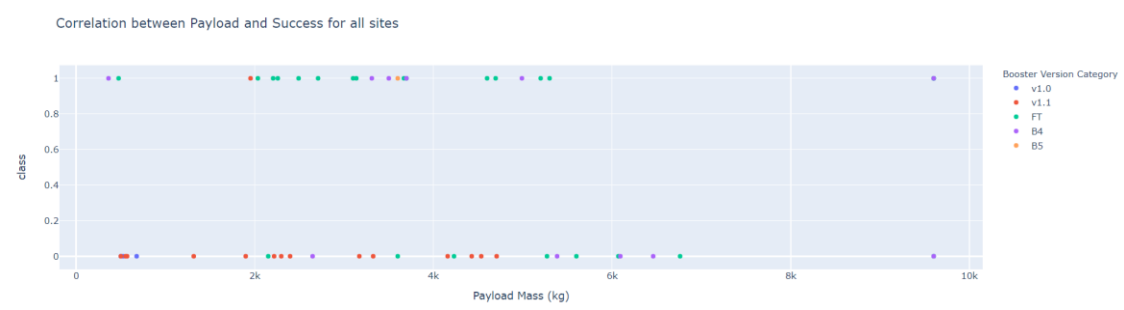
- Launches from KSC LC-39A have the most successful landings with 41.7% of the total
- CCAFS LC-40 also has a high proportion of the successful landings with 29.2%

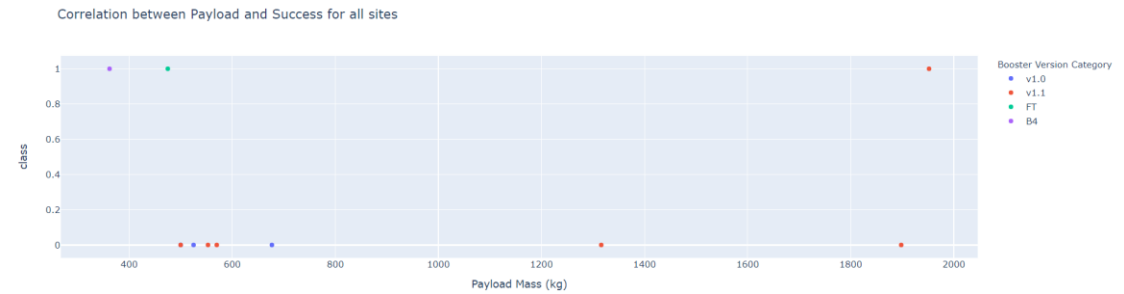
Launch success for KSC LC-39A

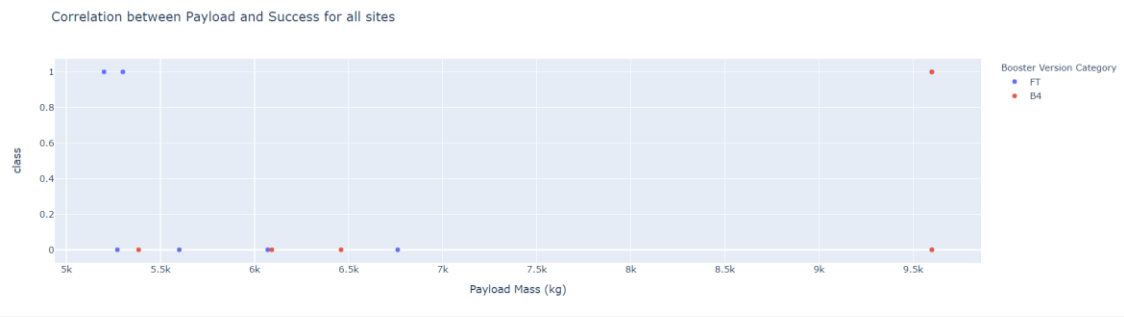


- Class 1 represents success while 0 represents failed landings
- Launches from KSC LC-39A have the best success rate with 79% of launches landing successfully

Payload vs. Launch Outcome





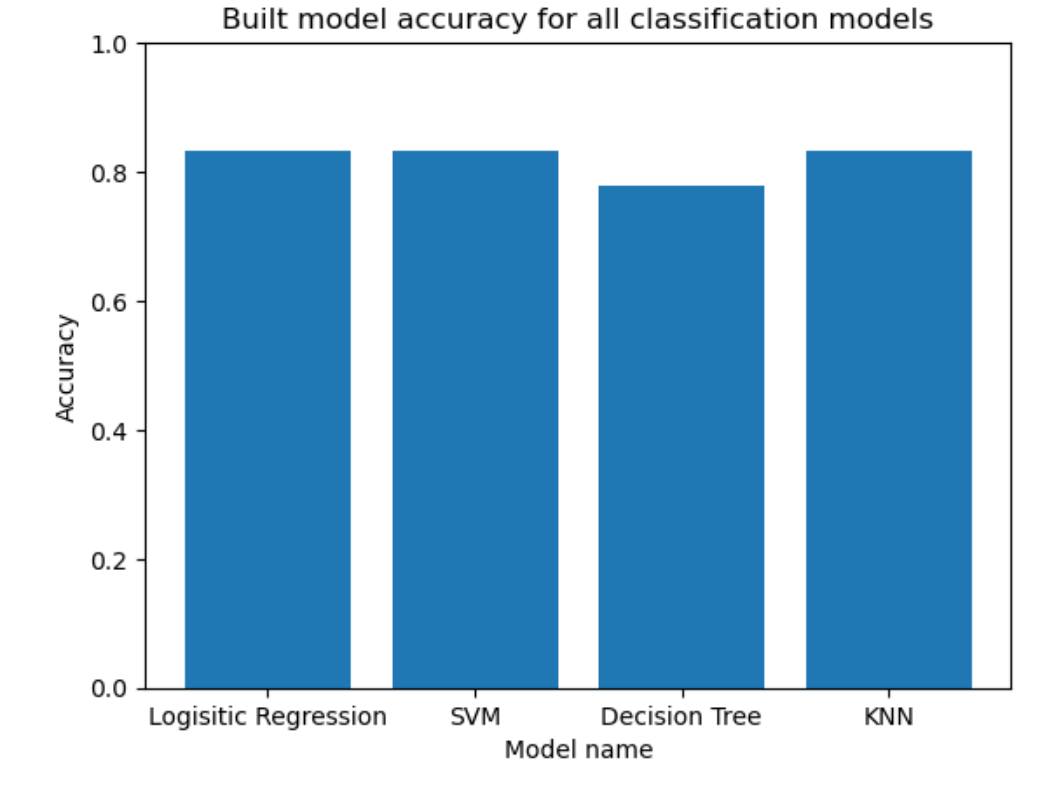


- The payload range from 2k to 5.5k has the most launches and the success rate is close to 50%
- The ranges 500 to 1.9k and 5.4k to 10k have low success rates
- Booster versions v1.0 and v1.1 have a very low success rate
- The booster versions FT and B4 have a higher success rate
- The booster version B5 has a 100% success rate although it has only been used once

Section 5

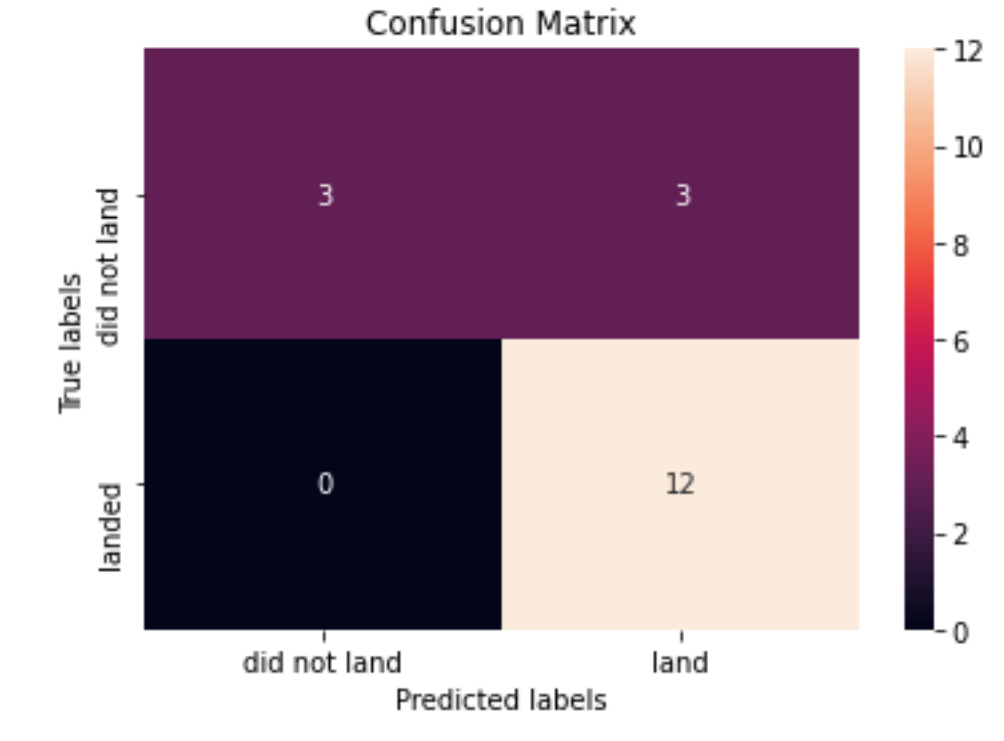
Predictive Analysis (Classification)

Classification Accuracy



- All 4 models have a similar classification accuracy
- The decision tree classifier has a slightly worse accuracy than the others (0.778 vs 0.833)

Confusion Matrix



- The confusion matrix was identical for all 4 built models
- The matrix shows there were 12 true positives, 3 false positives, 3 true negatives and 0 false negatives
- This means the main issue with the model's accuracy is false positives, it predicts some launches to land despite failing to land in reality

Conclusions

- Almost all failed landings are intentional which suggests it should be possible to predict future landing outcomes with launch data and without diving into the rocket science
- The key indicators which seem to determine whether a landing will succeed are flight number, date of launch, orbit type, payload mass and the launch site
- Launch sites are typically near the equator and in close proximity to coastlines, highways and railways but they are located far from cities
- Our machine learning models can predict whether a landing will occur with 83.3% accuracy, false positives are the main source of errors
- All four models perform very similarly although the decision tree classifier showed a slightly worse accuracy compared to the rest

Thank you!