IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adam Zacpal
03/16/2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Data collection involved SpaceX API usage and web scraping.
- Data was then wrangled, cleaned, and enriched.
- Exploratory Data Analysis (EDA) utilized visualization and SQL.
- Interactive visual analytics were conducted with Folium and Plotly Dash.
- Predictive analysis employed classification models, including building, tuning, and evaluating them for accuracy and performance improvement.
- Decision tree performed the best with accuracy of 0.9444

Introduction

- This was a final project of the IBM Data Science Professional Certificate on Coursera learning website
- We wanted to find out what is the connection between launch sites and success of rocket booster landings of SpaceX

Section 1

Methodology

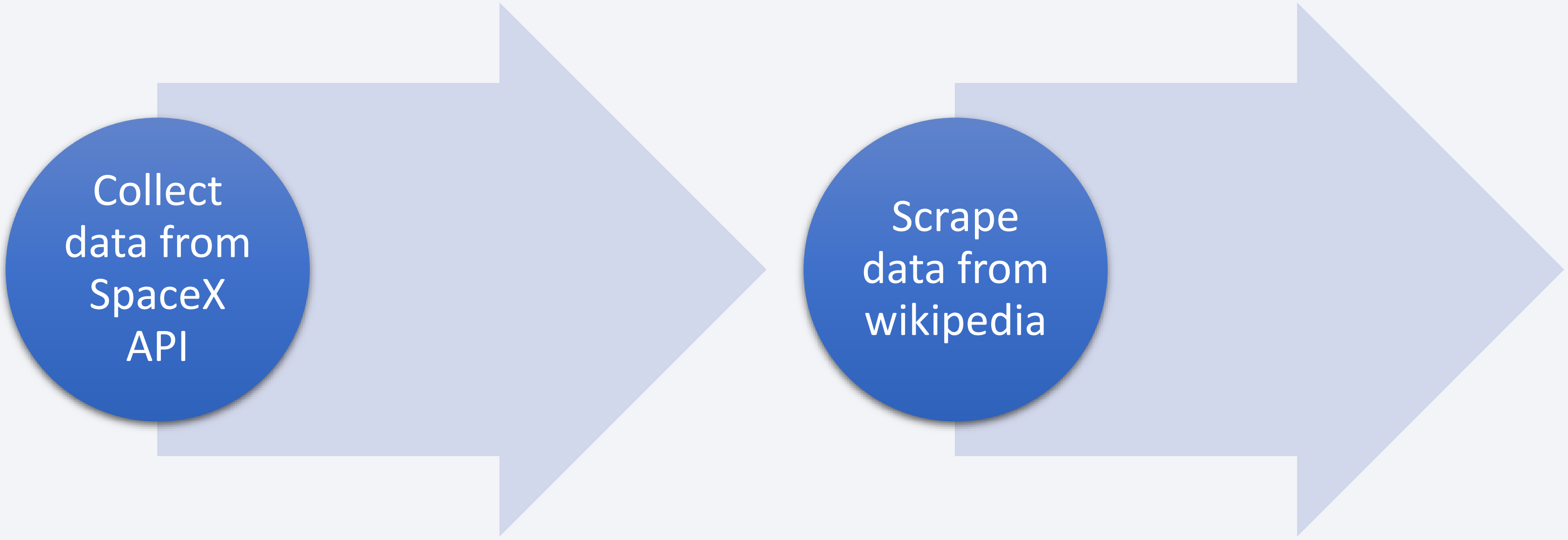
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping techniques
- Perform data wrangling
  - Data was cleaned, filtered, missing values were replaced and new column was added
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

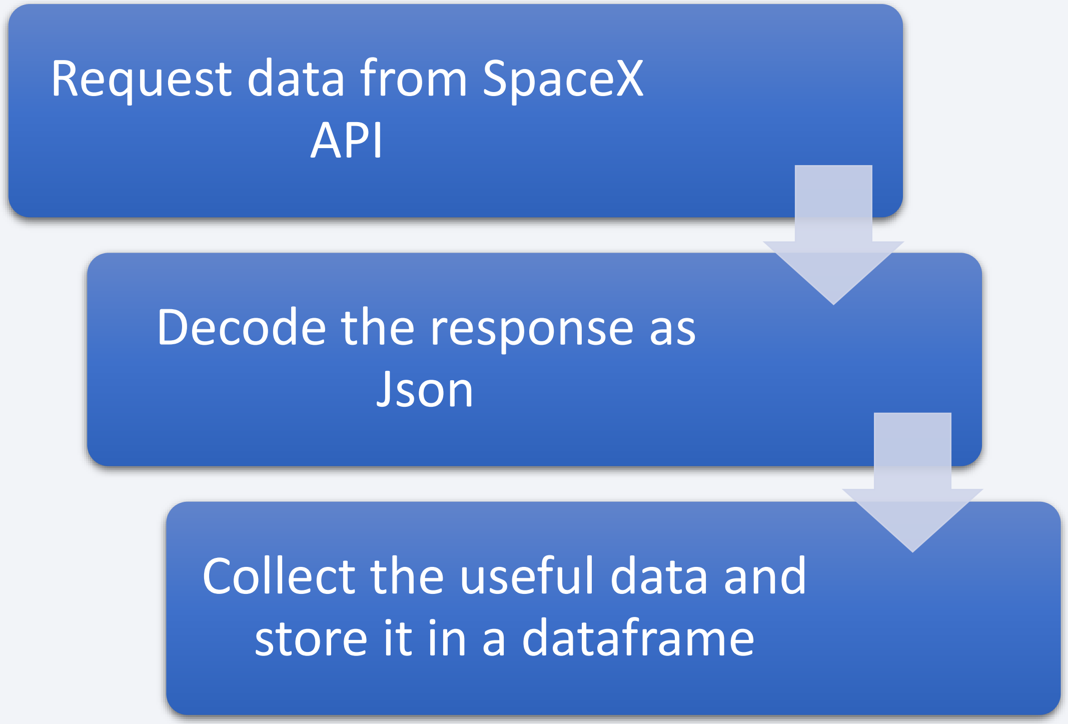
Data Collection

- Data was collected using SpaceX API and web scraping techniques from wikipedia



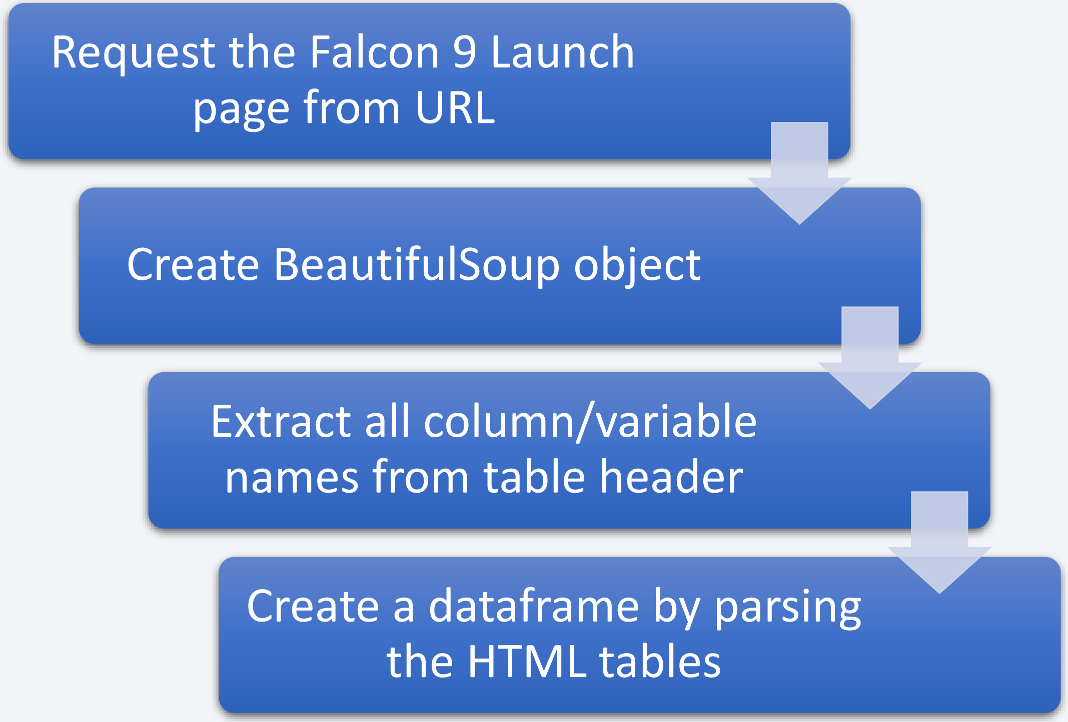
Data Collection – SpaceX API

- In this part of the project data was collected from SpaceX API
- First we requested the data, then decoded the response as Json and finally collected useful data and stored it in a dataframe
- https://github.com/Adazac/spaceY/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- In this part of the lab data was scraped from wikipedia
- BeautifulSoup library was used to parse the data and store it to a dataframe
- https://github.com/Adazac/spaceY/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data was cleaned to extract only useful information
- Filtering was made to only include data from Falcon 9 launches
- Payload Mass had 5 null values, which were replaced by mean of the data
- New column „Class“ was added to address the outcome of landing

- https://github.com/Adazac/spaceY/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- FlightNumber vs PayloadMass catplot with class of the Outcome – to see how success of launches with different payload developed as flight number increased
- FlightNumber vs Launch Site catplot with class of the Outcome – to see how success of launches from different launch sites developed as flight number increased
- Launch Site vs Payload Mass scatterplot with class of the Outcome
- Flight number vs Orbit scatter plot with class of the Outcome
- Bar chart of Success rates of launches to different Orbits
- Payload Mass vs Orbit scatter plot with class of the Outcome
- Visualized the launch success yearly trend to see if success rate improved over the years
- https://github.com/Adazac/spaceY/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Displayed the names of the unique launch sites in the space mission
  - %sql select distinct("Launch_Site") from SPACEXTABLE
- Displayed 5 records where launch sites begin with the string 'CCA'
  - %sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
  - %sql select sum("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Customer"=='NASA (CRS)‘
- Displayed average payload mass carried by booster version F9 v1.1
  - %sql select AVG("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version"=='F9 v1.1'
- Listed the date when the first succesful landing outcome in ground pad was acheived
  - %sql select MIN("Date") from SPACEXTABLE where "Landing_Outcome"=='Success (ground pad)'

https://github.com/Adazac/spaceY/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with SQL

- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - %sql select "Booster_Version" from SPACEXTABLE where "Landing_Outcome"=='Success (drone ship)' and "PAYLOAD_MASS__KG_" >4000 and "PAYLOAD_MASS__KG_" < 6000
- Listed the total number of successful and failure mission outcomes
  - %sql select "Mission_Outcome", count(*) from SPACEXTABLE group by "Mission_Outcome"
- Listed the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - %sql select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" == (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)
- Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
  - %sql select substr("Date", 6,2), "Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where substr("Date", 0,5) = "2015" and "Landing_Outcome" like "Failure (drone ship)"
- Ranked 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
  - %sql select "Landing_Outcome", count(*) as "count_of_outcomes" from SPACEXTABLE where "Date" between "2010-06-04" and "2017-03-20" group by "Landing_Outcome" order by "count_of_outcomes" desc

https://github.com/Adazac/spaceY/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

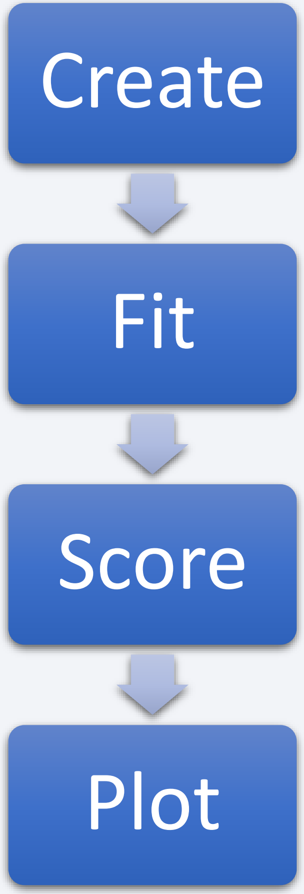
- Marked all launch sites on a map
- Marked the success/failed launches for each site on a map
- Calculate the distances between a launch site to its proximities
- Objects were added to better understand the positioning of launch sites and their success rate
- https://github.com/Adazac/spaceY/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Launch Site Drop-down Input Component
- Success pie chart based on selected site dropdown
- Range Slider to Select Payload
- Success vs payload scatter plot
- The dashboard was built to answer these questins:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?
- https://github.com/Adazac/spaceY/blob/main/spacex_dash_app.py

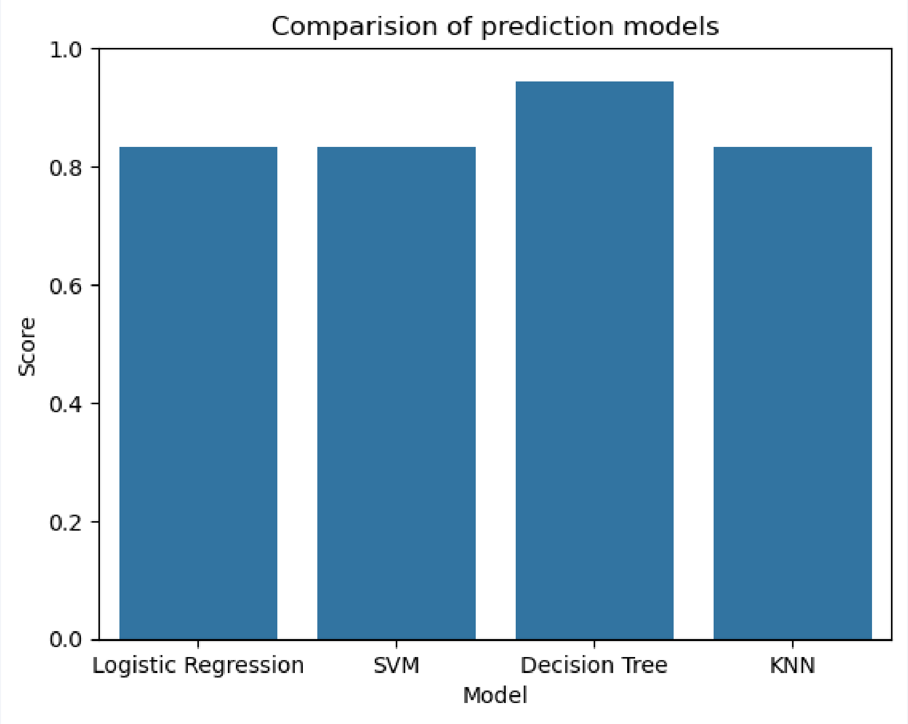
Predictive Analysis (Classification)

- We started by creating a NumPy array from the column Class and assigning it to variable Y
- Standardized the data in a dataframe with independent variables and assigned it to variable X
- Splitted data into train and test data
- Performed a series of prediction models which consisted of:
    - Creating the model
    - Fitting the model
    - Scoring the model
    - Plotting a confusion matrix of the model
- Compared which model performs best
- https://github.com/Adazac/spaceY/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

- From EDA we found out that success of launches with different payload increased as flight number increased
- We also found out that success rate varied based on the Orbit rocket was launched to
- In predictive analysis we found out that Decision tree performer the best



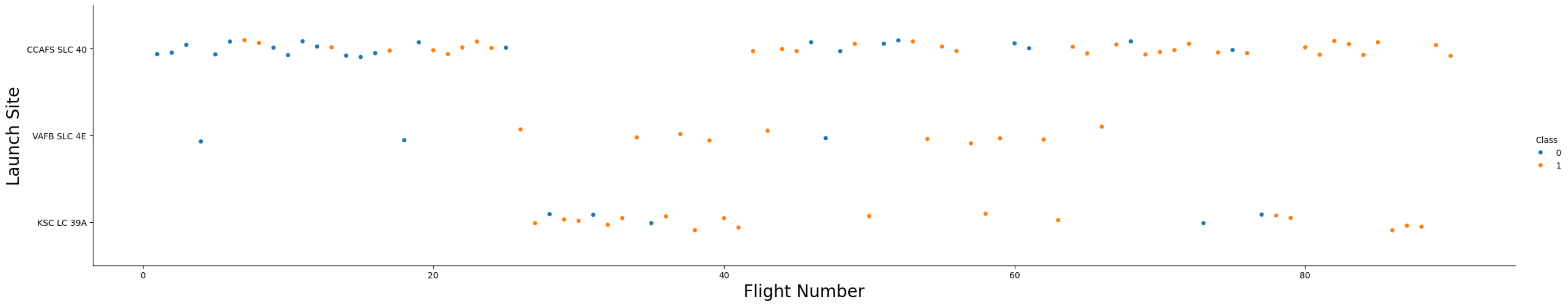
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- A scatter plot of Flight Number vs. Launch Site

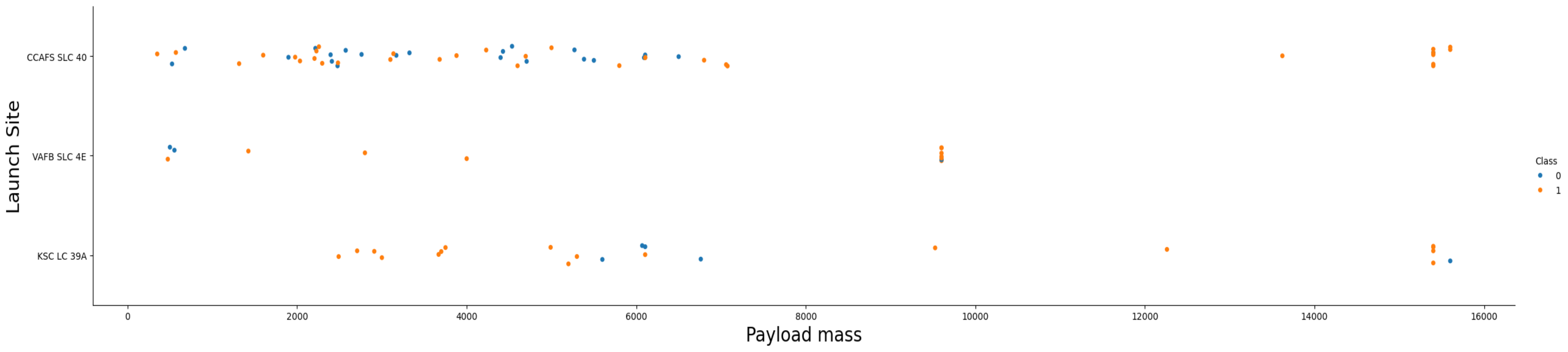
- We can see that as flight number increased the outcome was more likely to be success



Payload vs. Launch Site

- A scatter plot of Payload vs. Launch Site

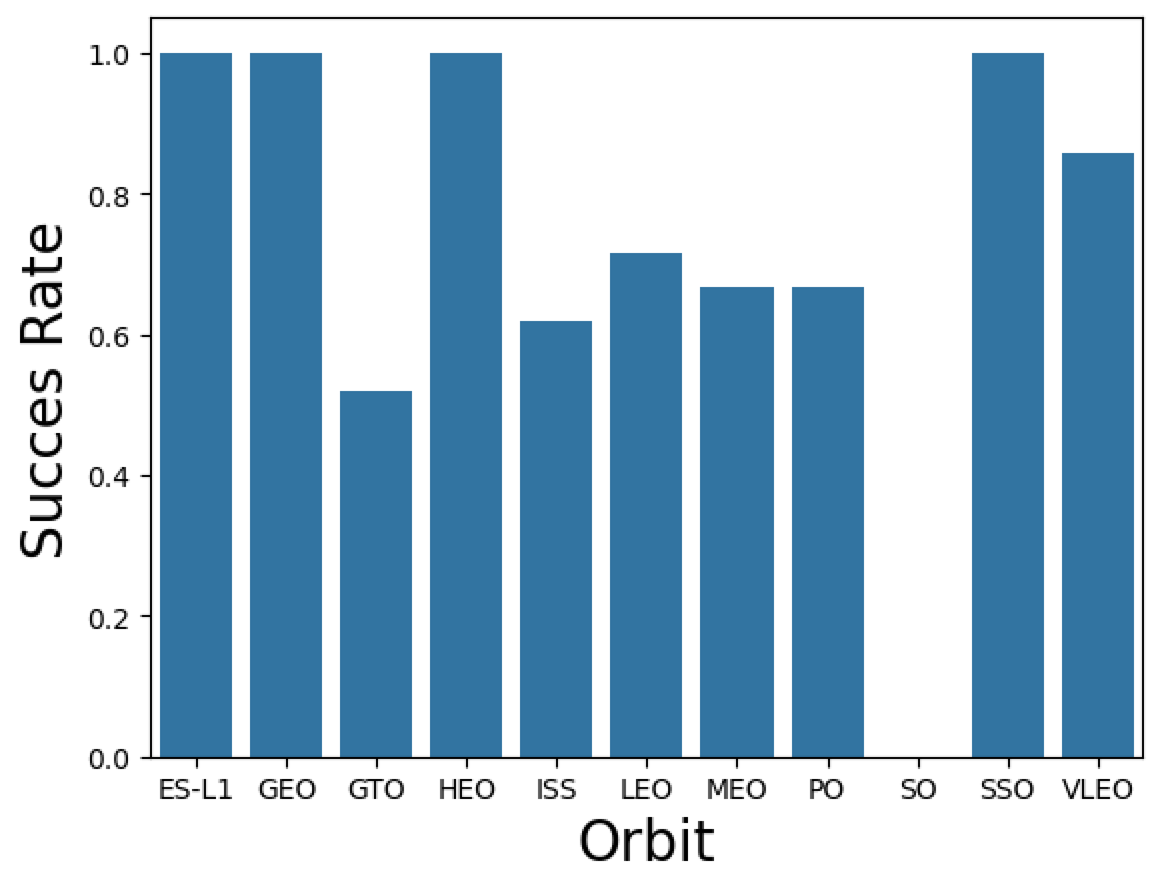
- We can see that as Payload mass increased the outcome was more likely to be success



Success Rate vs. Orbit Type

- A bar chart for the success rate of each orbit type

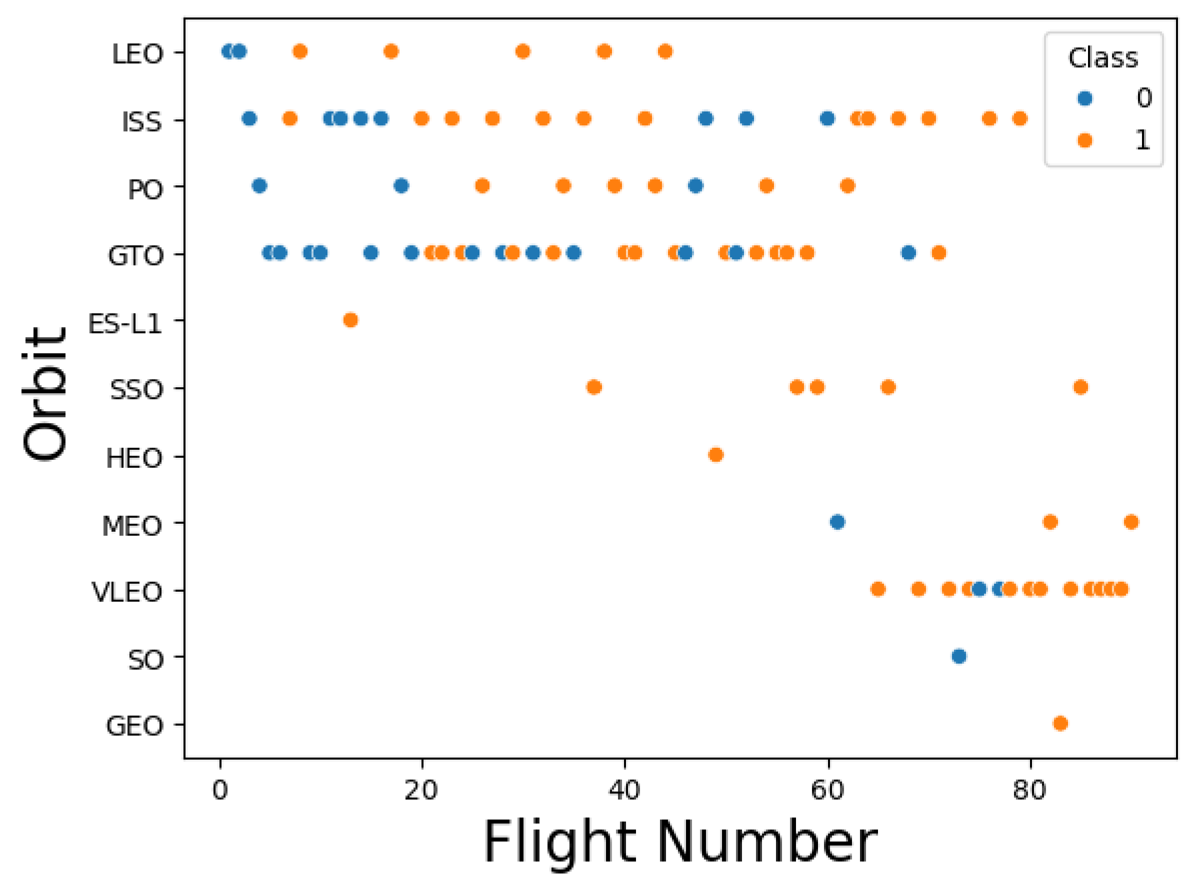
- We can see that Success rates vary based on Orbit type



Flight Number vs. Orbit Type

- A scatter plot of Flight number vs. Orbit type

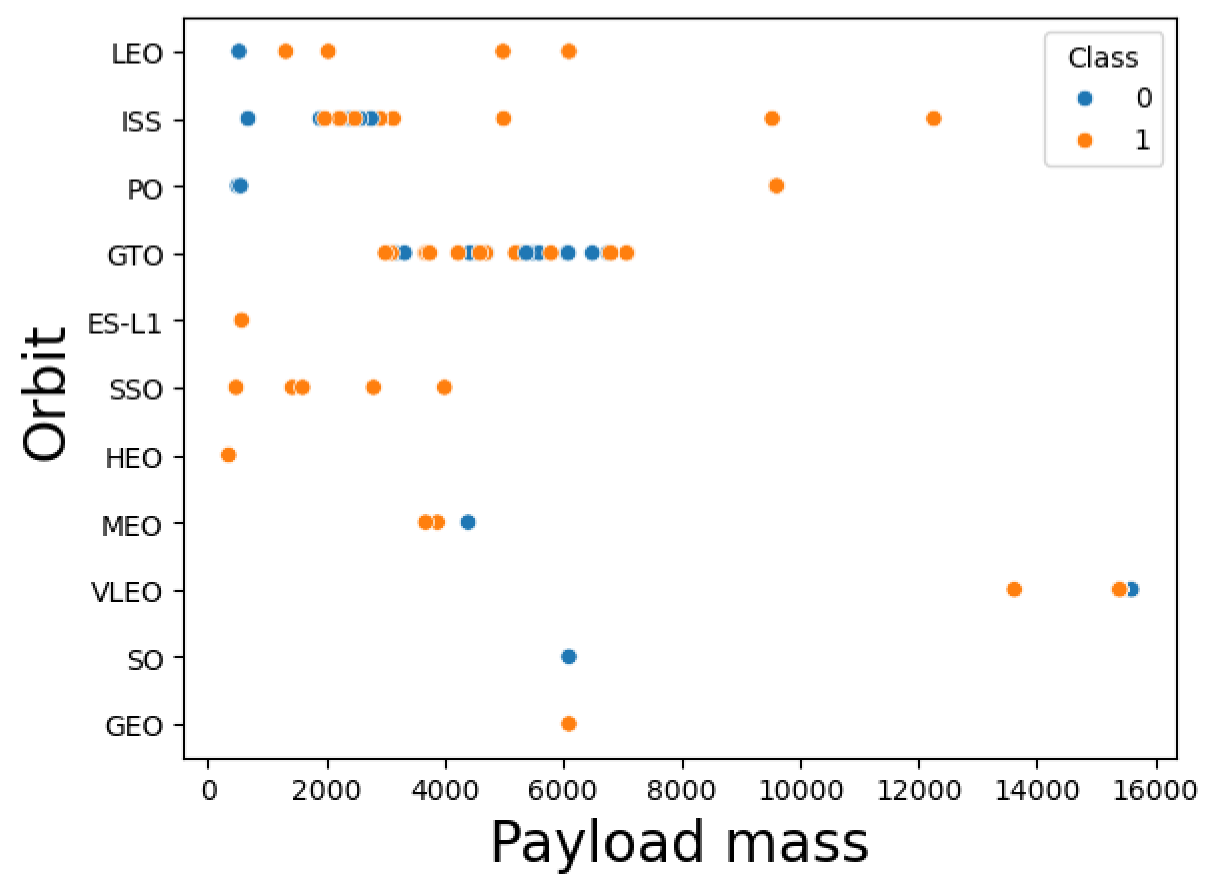
- We can see that flights to orbit with low flight numbers were likely to be unsuccessful



Payload vs. Orbit Type

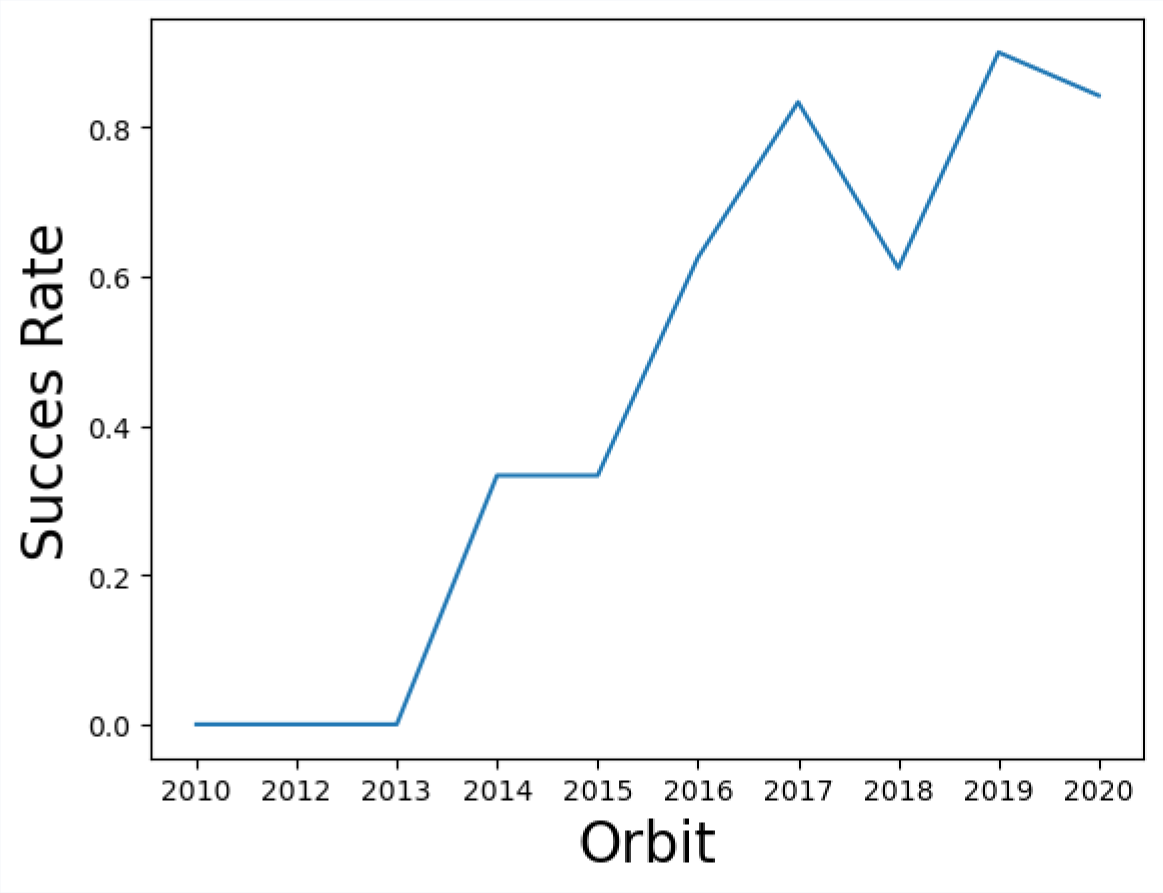
- A scatter point of payload vs. orbit type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS



Launch Success Yearly Trend

- A line chart of yearly average success rate
- We can observe that the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites

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

Launch Site Names Begin with 'CCA'

- Find 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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Total_Mass
45596

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

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

MIN(Date)
2015-12-22

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

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

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

2015 Launch Records

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

substr("Date", 6,2)	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

- 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_of_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 map in Folium

- We can see that all launch sites are located on the coastline on the south of United States

Launch site and its successful/unsuccessful launches in Folium

- We can see that launches on this site were mostly unsuccessful

Launch site with its proximities in Folium

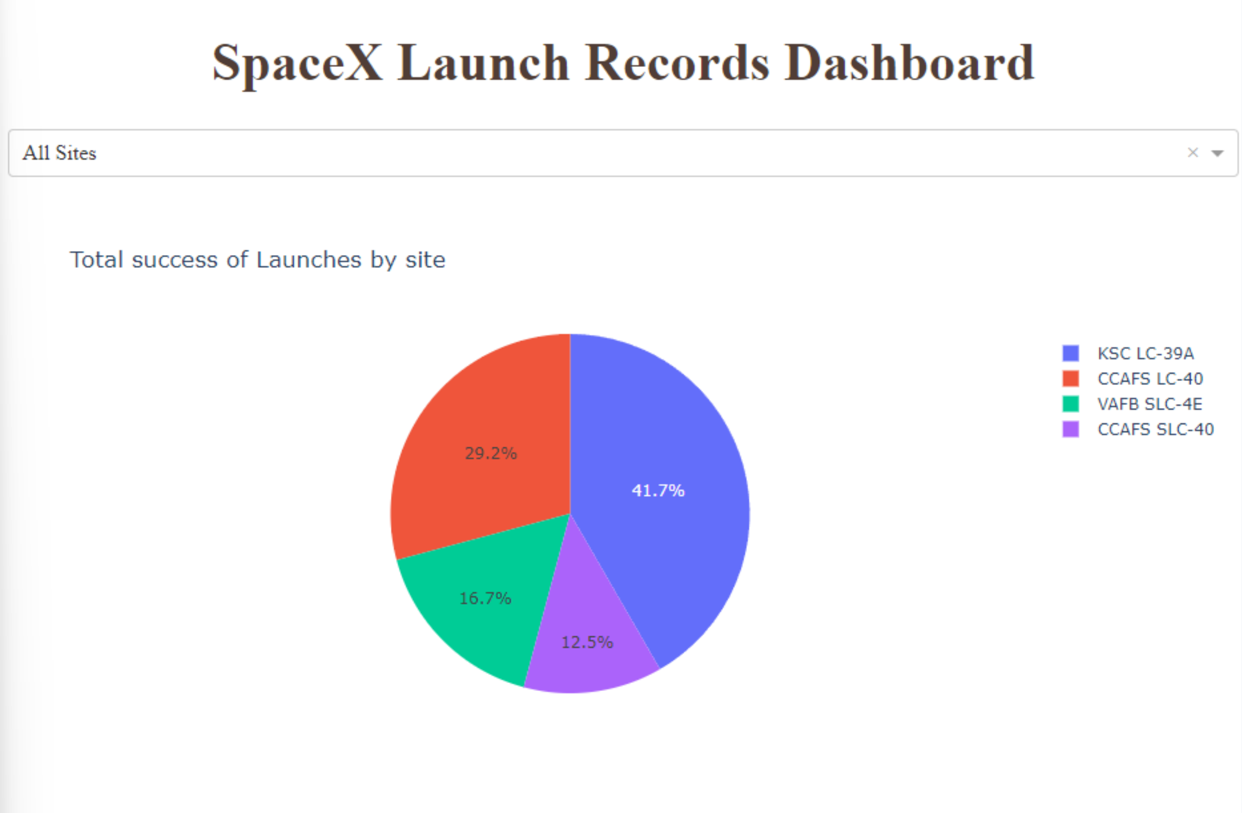
- Selected launch site with a line to a coastline, with distance calculated and displayed
- We can see that the launch site is less than 1 kilometer away from a costline

Section 4

Build a Dashboard with Plotly Dash

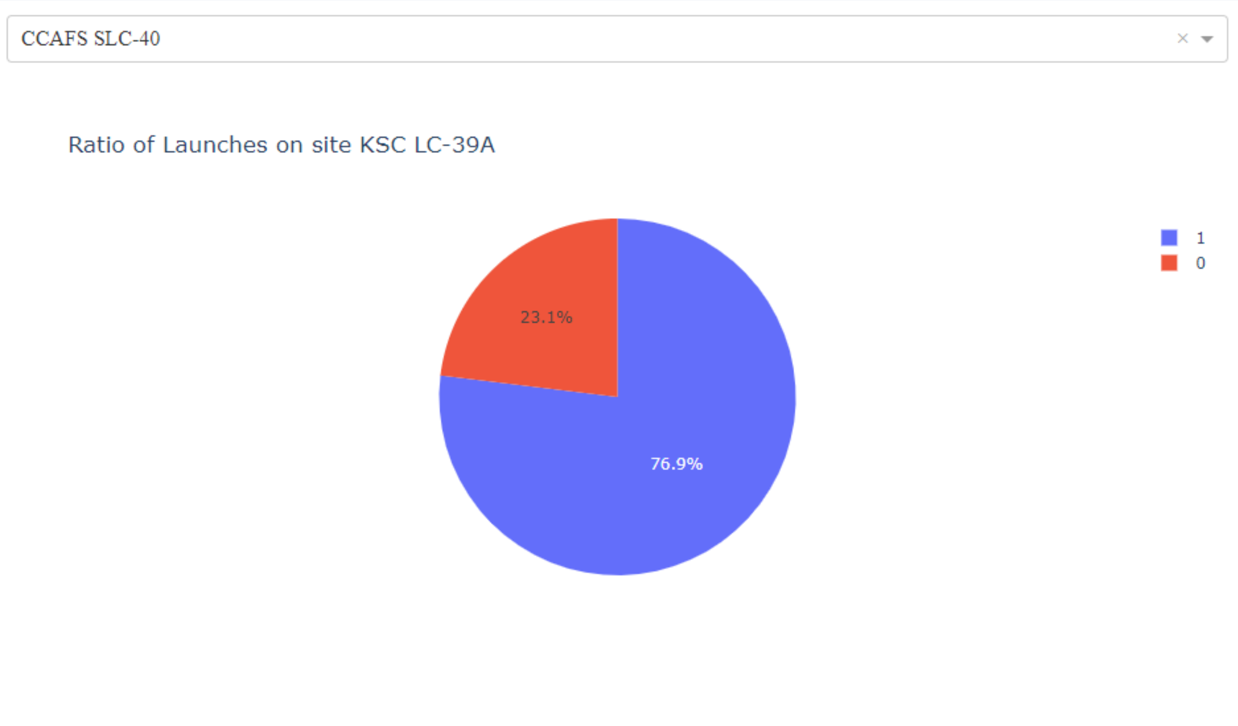
Launch success count for all sites Dashboard

- From this pie chart, we can see that KSC LC-39A site is responsible for almost half of the successful launches



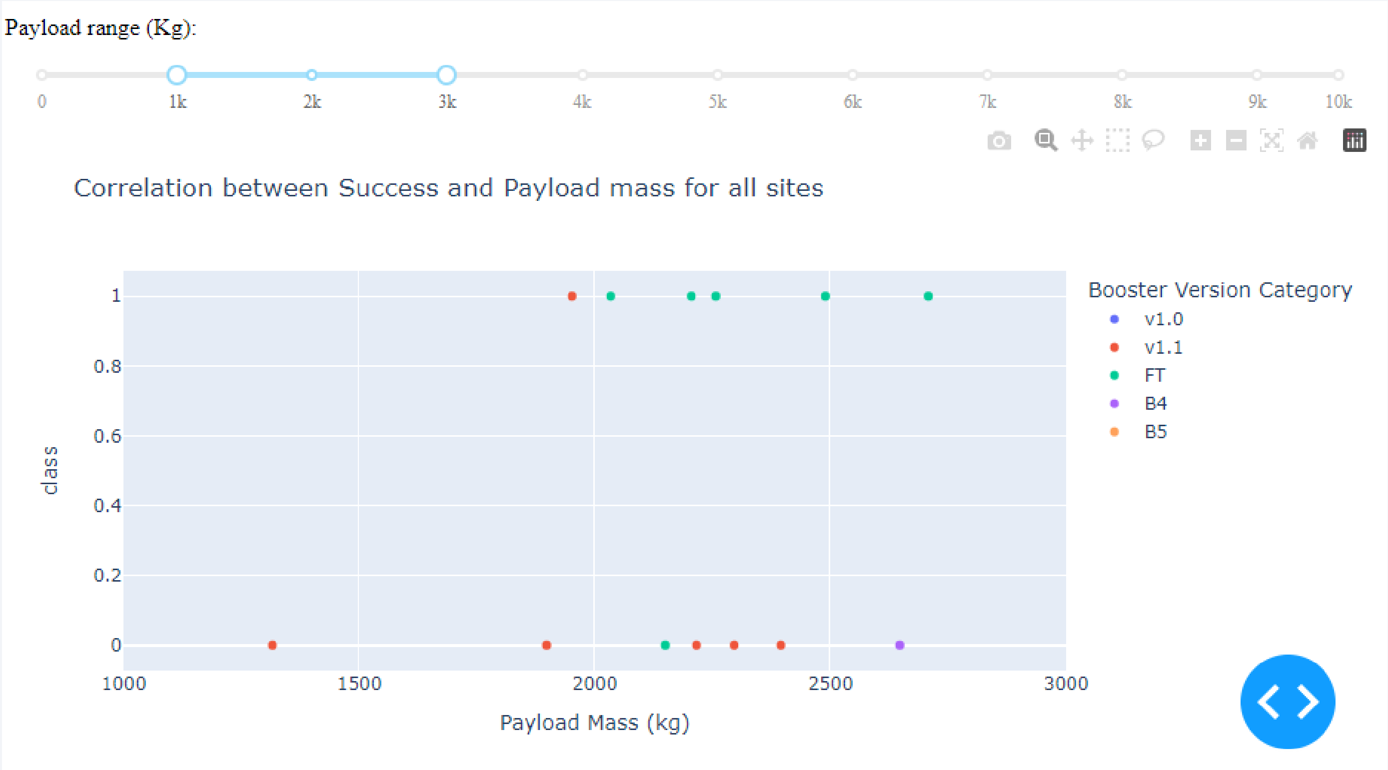
Highest launch success ratio Dashboard

- We can see that the launch site with the highest launch success ratio has 76.9% success rate



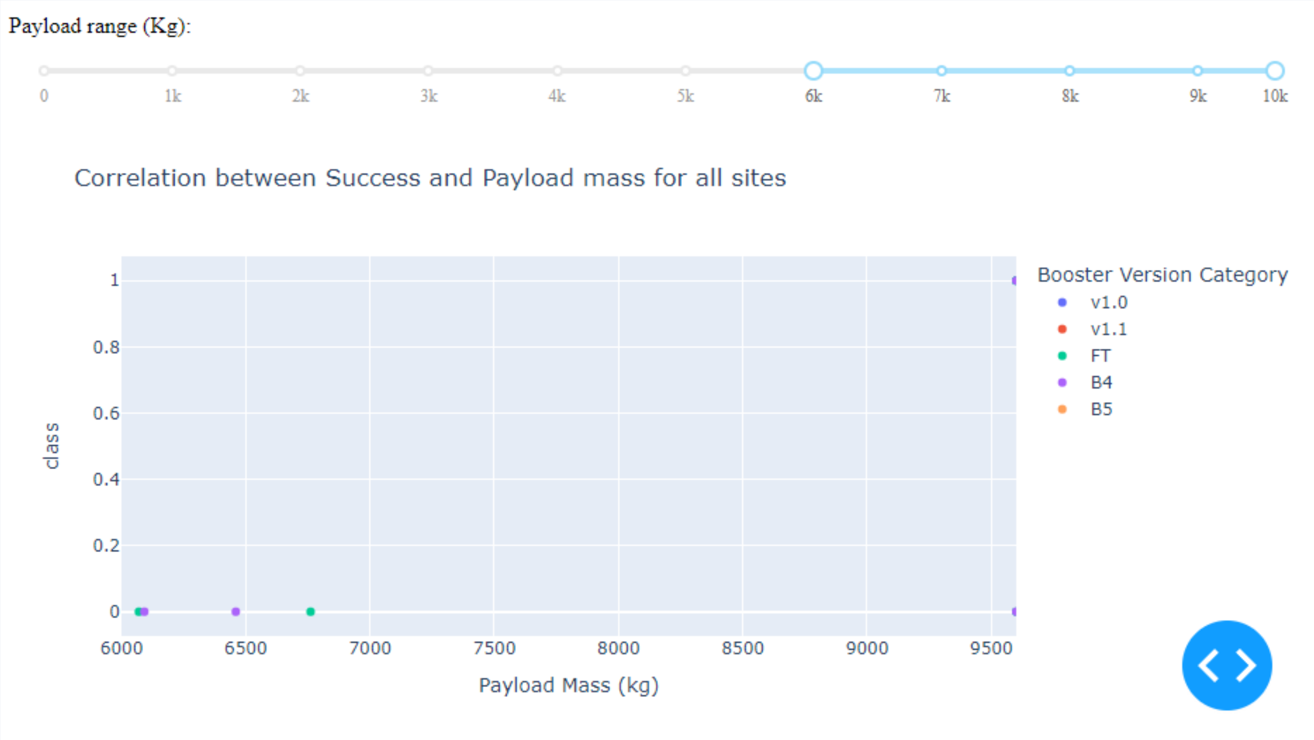
Payload vs Launch Outcome Dashboard 1

- We can see that in range from 1k to 3k Kg the success rate of booster FT is very high in contrast to booster v1.1 which is very low



Payload vs Launch Outcome Dashboard 2

- We can see that in range from 6k to 10k Kg there were no successful booster landings

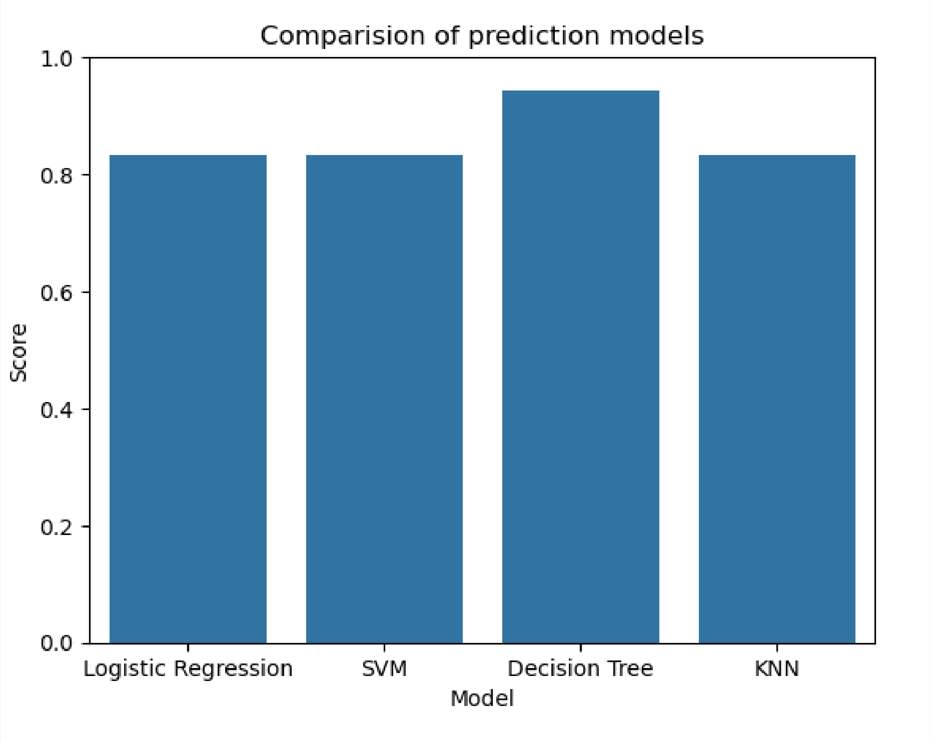


Section 5

Predictive Analysis (Classification)

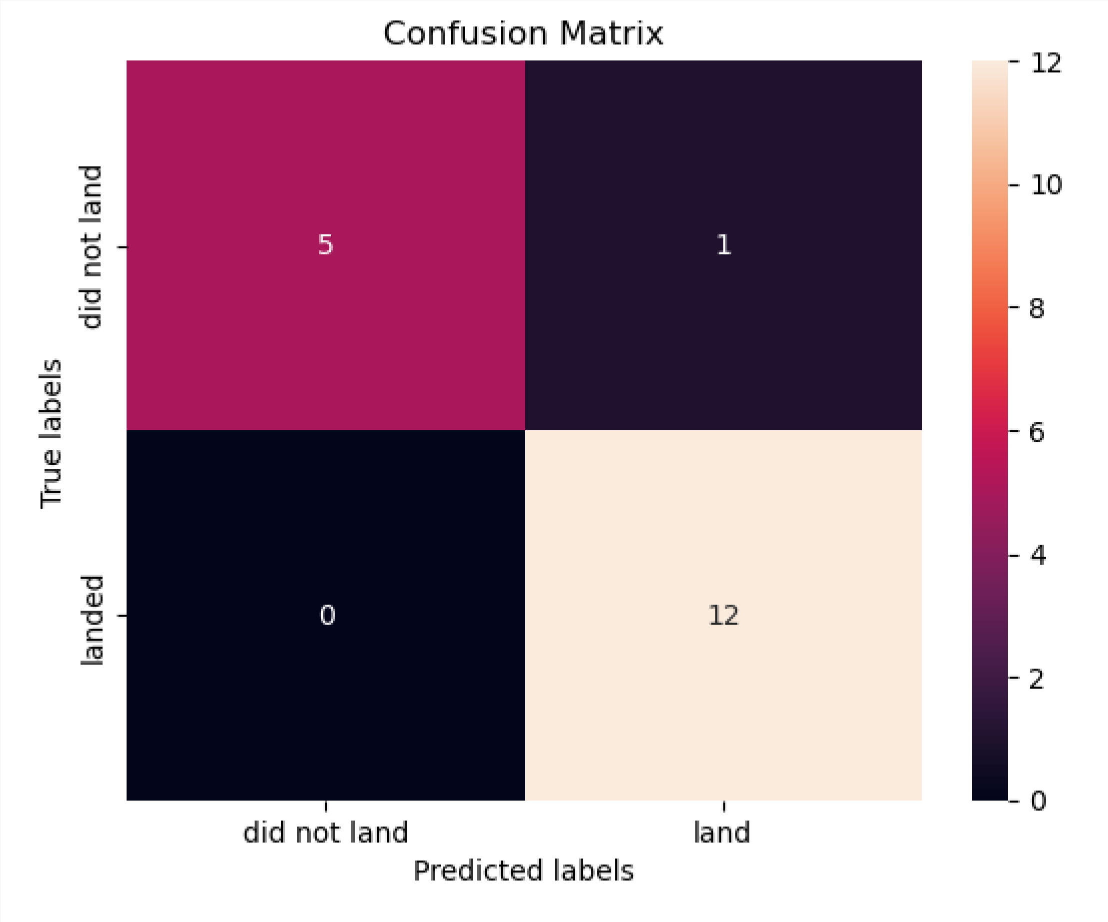
Classification Accuracy

- Decision tree model has the highest classification accuracy of 0.944444



Confusion Matrix

- From the confusion matrix of the best model, which was decision tree, we can see that it predicted 17 out of 18 values right and only 1 wrong
- Of 13 that were predicted to land 12 truly landed and only 1 did not
- All 5 that were predicted to not land did not land



Conclusions

- All 4 models performed reasonably good
- The best model was a Decision tree model
- Only 1 outcome was predicted wrong

Thank you!