IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mave
13.06.2024

Outline

Executive Summary

- Data collection methodology:
    - Request to the SpaceX API
    - Web Scrapping
- Data wrangling
- Exploratory data analysis (EDA)
    - Visualization
    - SQL
- Interactive visual analytics
    - Folium
    - Plotly Dash
- Predictive analysis using classification models
    - Build, tune, evaluate classification models

Introduction

- Project background and context
    - Spacex rockets have a low cost depends on success launch rate
    - Determine success landing rate is crucial for determining cost
- Problems you want to find answers
    - Determine key features of rockets launching that affect the launch success rate
    - Predict the launch success rate of rockets launch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

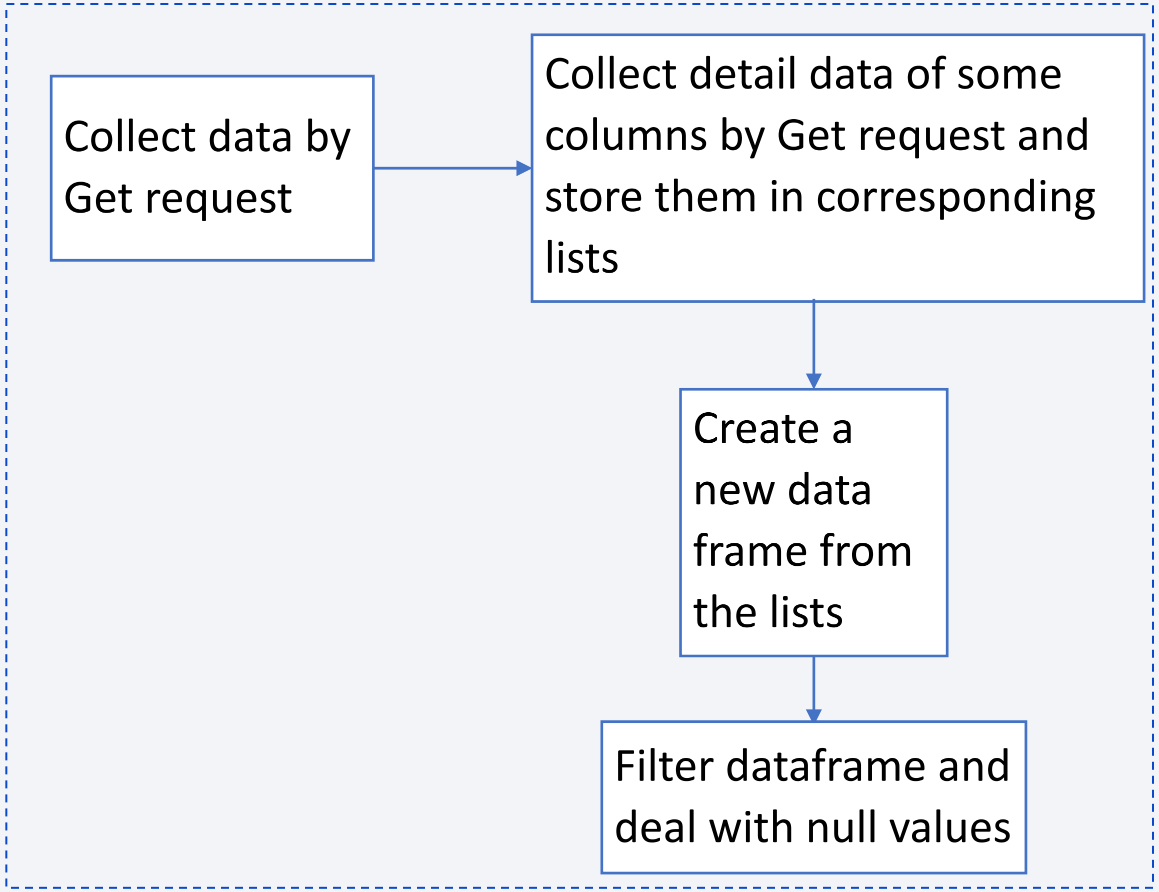
Data Collection

- In this project, datasets are collected in two ways:
  - Request to the SpaceX API
  - Web Scrapping

Data Collection – SpaceX API

- Data collection with SpaceX REST calls flowcharts

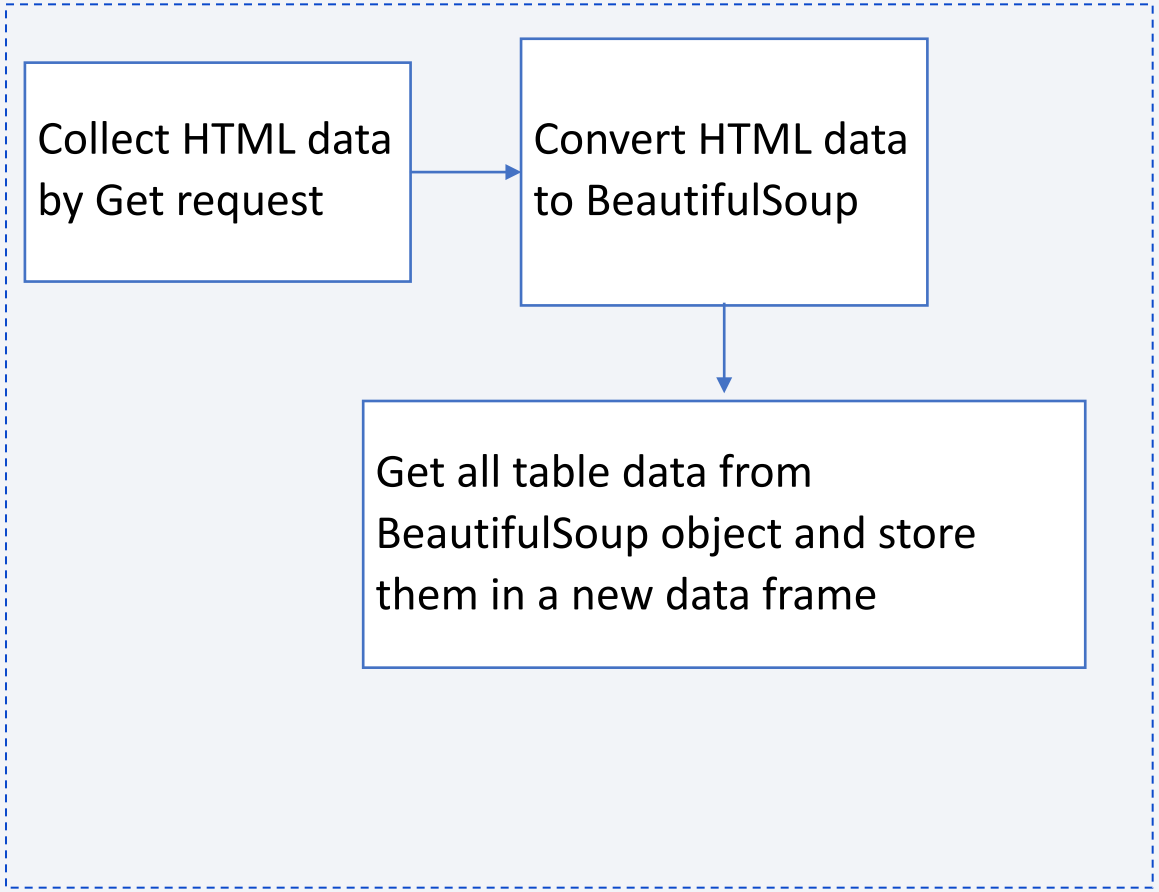
- https://github.com/schickwu/ds-capstone-template-coursera/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web scraping process using key flowcharts

- https://github.com/schickwu/ds-capstone-template-coursera/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Checking empty values and different data types for all columns.
- Calculating some statistical character of data:
    - The number of launches on each site
    - The number of launches on each orbit type
    - The occurrence of mission outcome
- Assign binary column indicating each outcome result
- https://github.com/schickwu/ds-capstone-template-coursera/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Find relationship between number of flights and orbit types & launch sites
  - Scatter plot of flight number and orbit types
  - Scatter plot of flight number and launch sites
- Find relationship between payload mass and orbit types & launch sites
  - Scatter plot of payload mass and orbit types
  - Scatter plot of payload mass and launch sites
- Bar plot of success rate and orbit types
- https://github.com/schickwu/ds-capstone-template-coursera/blob/main/edadataviz.ipynb

EDA with SQL

- Launch sites under different conditions
- Payload mass under different conditions
- Booster version under different conditions
- Outcomes under different conditions
- https://github.com/schickwu/ds-capstone-template-coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Circles and markers for all launch sites
- Icons and markers for all outcomes on each launch sites
- Lines and markers indicating the distances between specific launch site and its proximities
- We add these objects to find out the geographic characters of each launch site
- https://github.com/schickwu/ds-capstone-template-coursera/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie plots of outcome on each launch site and all launch sites
  - The launch sites is intractable and changeable
- Scatter plot of outcome among different payload mass under different launch sites
  - Payload mass and the launch sites are intractable and changeable
- We want to find out the characters that affecting the outcomes of launches
- https://github.com/schickwu/ds-capstone-template-coursera/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Logistic regression, SVM, Tree, KNN
- GridSearchCV method to train and evaluate parameters for each model, which has the best performance on the train dataset
- Test each best trained model with test dataset with accuracies and confusion matrixes
- https://github.com/schickwu/ds-capstone-template-coursera/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

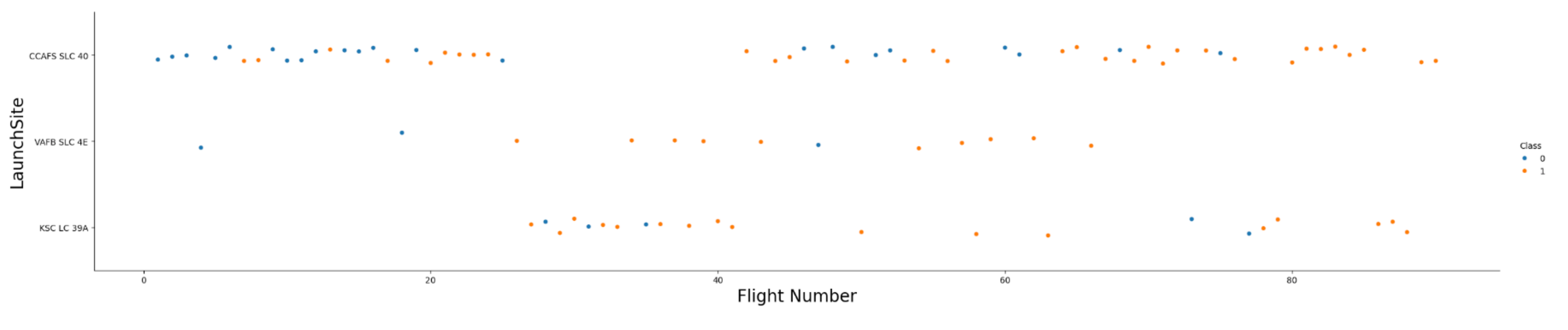
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

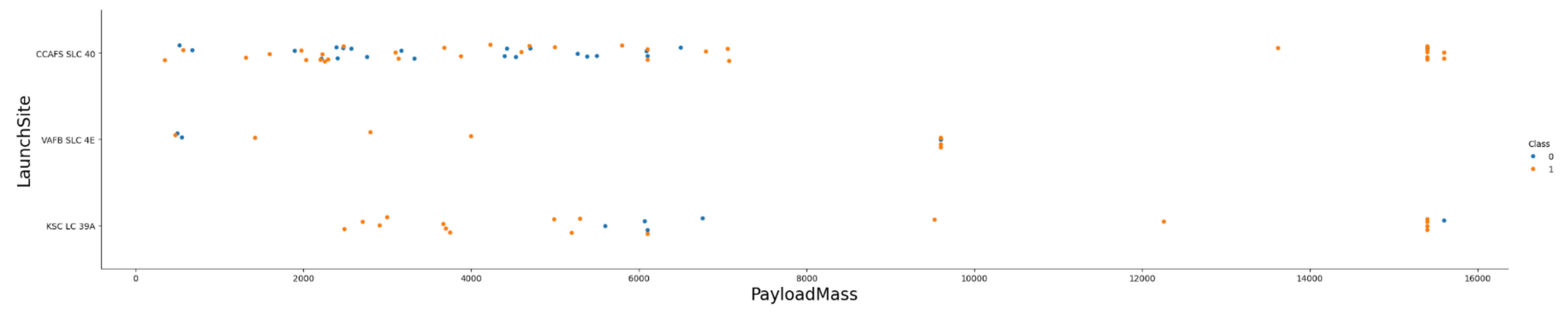
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site
- There are more rockets launched from CCAFS SLC and KSC LC launch sites than launched from the VAFB-SLC launch site.



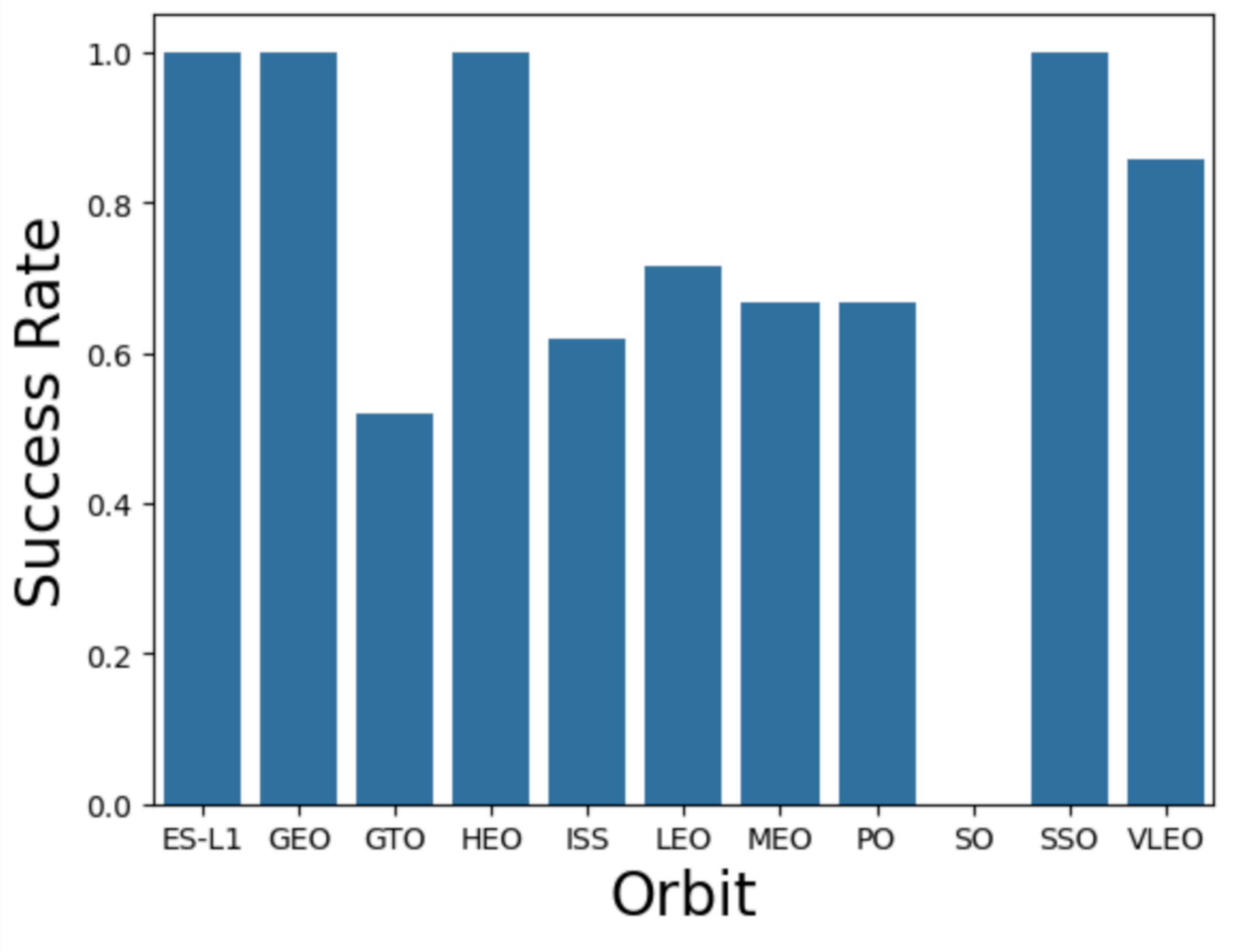
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000);
- Most of rockets have a payload mass between 0 and 8000.
- For heavy payload rockets, the successful rate is much higher than other rockets.



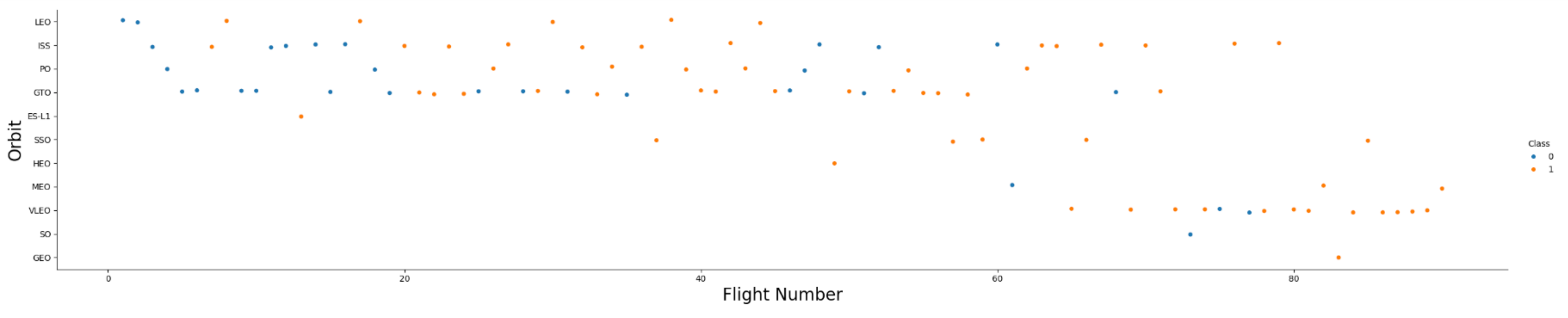
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type
- Rockets with orbit type: ES-L1, GEO, HEO, SSO, and VLEO have clear higher success rate compares to other orbit types.
- There are no success launches for rockets with SO as orbit type.



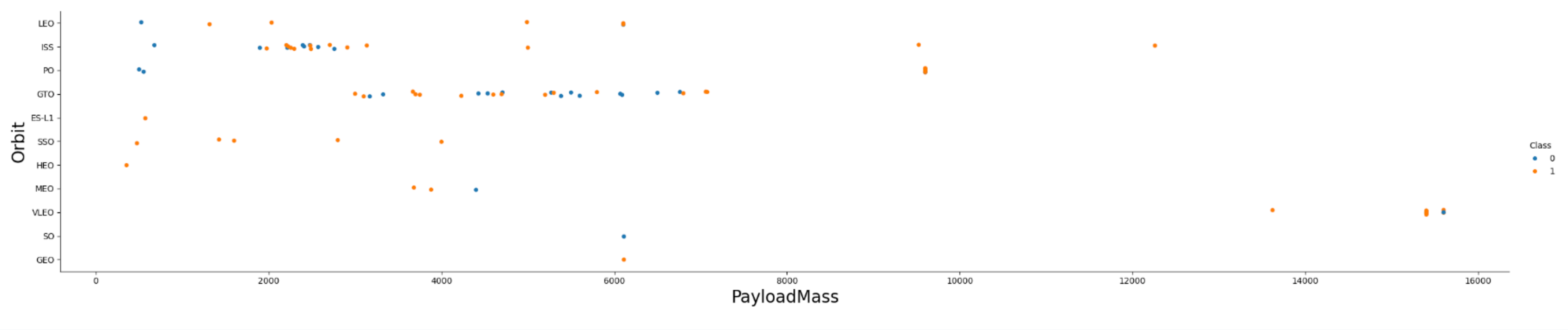
Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type
- The majority of rocket launches are in the orbit: LEO, ISS, PO, GTO.
- Rockets with LEO and VLEO orbit type have a much higher successful landing rate.



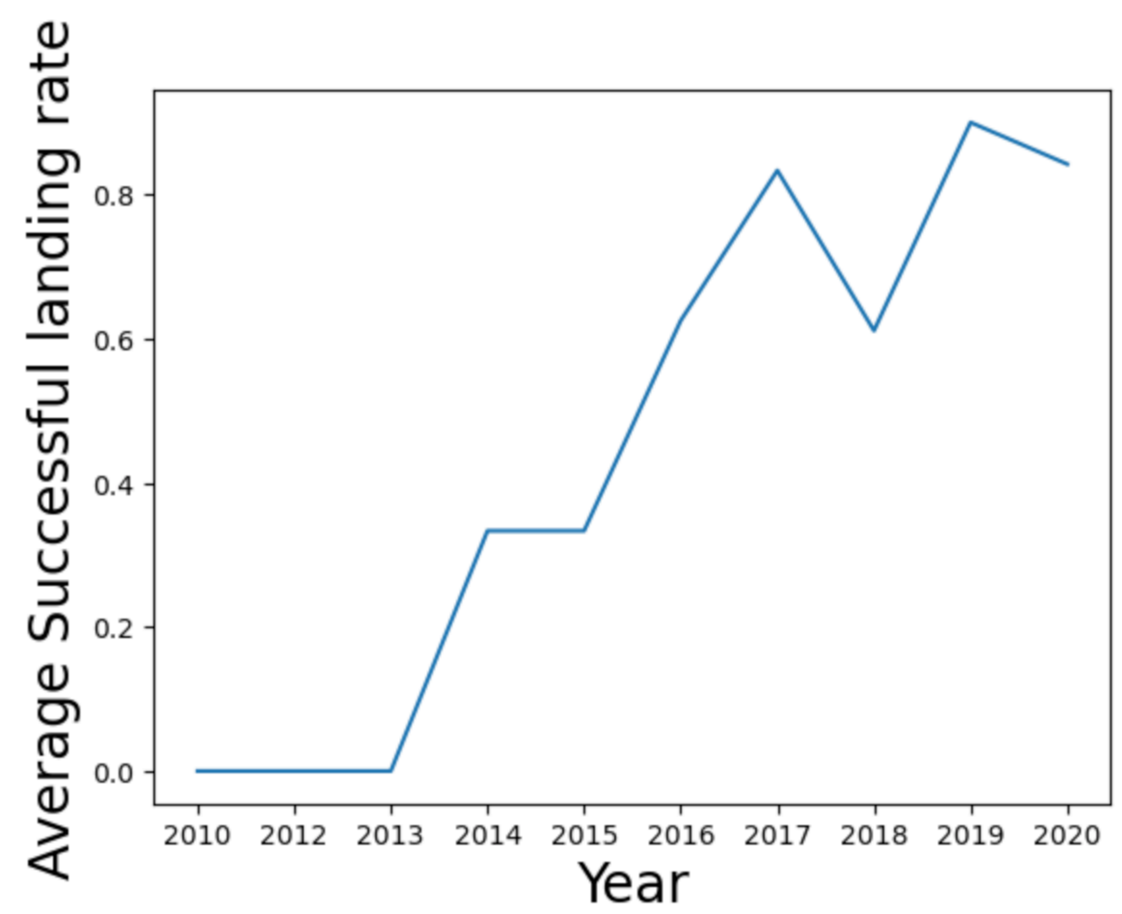
Payload vs. Orbit Type

- Scatter point of payload vs. orbit type
- For Polar,LEO and ISS orbit types, the successful landing rate are higher With heavy payloads.
- Only for ISS, VLEO orbit type there are some heavy payloads rockets launches(greater than 10000)



Launch Success Yearly Trend

- Line chart of yearly average success rate
- The success rate since 2013 kept increasing till 2020
- Only in 2018 there is a clear decreasing



All Launch Site Names

- Names of the unique launch sites as shown in the figure
- There are in total 4 unique launch site in the space mission

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` as shown in the figure below
- Here are first 5 records listed, after we filtered out the records whose launch site is not 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

- The total payload carried by boosters from NASA
- Here we filtered all records whose customer is NASA (CRS) from the SPACEXTABLE, and calculated total payload of these records.

total_payload_carry
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1
- First, we filtered records whose Booster_Version starts with 'F9 v1.1’, then we calculated the average PAYLOAD_MASS.

Average_Payload_Mass
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- We first filtered the success ground landing records, then we find the minimal date of these records as the first successful landing date

First_Successful_Ground_Landing_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- We have filtered records which have payload mass greater than 4000 but less than 6000, and have successfully landed on drone ship.

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
- We group the records into two groups according to the successful and failure mission outcomes, and counted the number of records of each group.

outcome	number
Failure	10
Success	61

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass
- First we use a subquery to find the maximum payload mass, and then we select the records which have the corresponding 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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- We filtered records in year 2015 and have failed drone ship landing outcomes.

month	Booster_Version	Landing_Outcome	Launch_Site
01	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	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

- Filter records between the date 2010-06-04 and 2017-03-20, group these records by different Landing outcomes and count the number of records belong to each group, finally order these groups by their count in descending order.

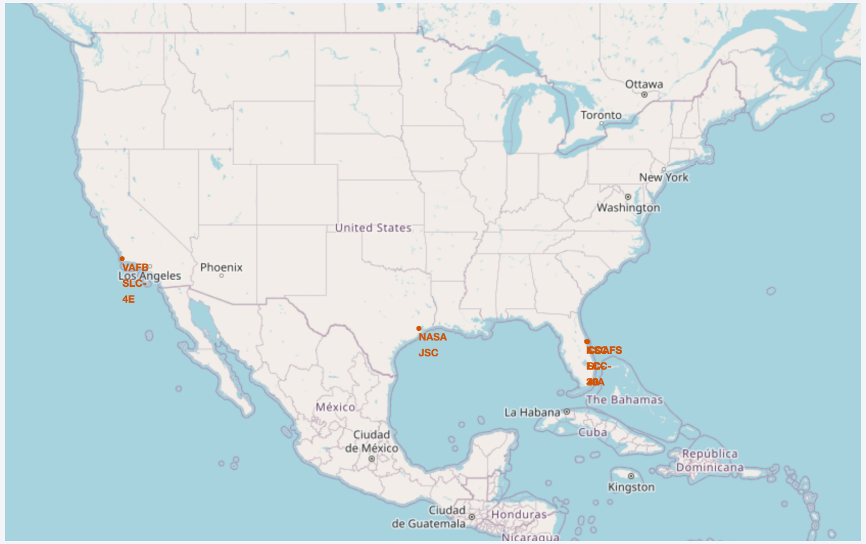
Landing_Outcome	number
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

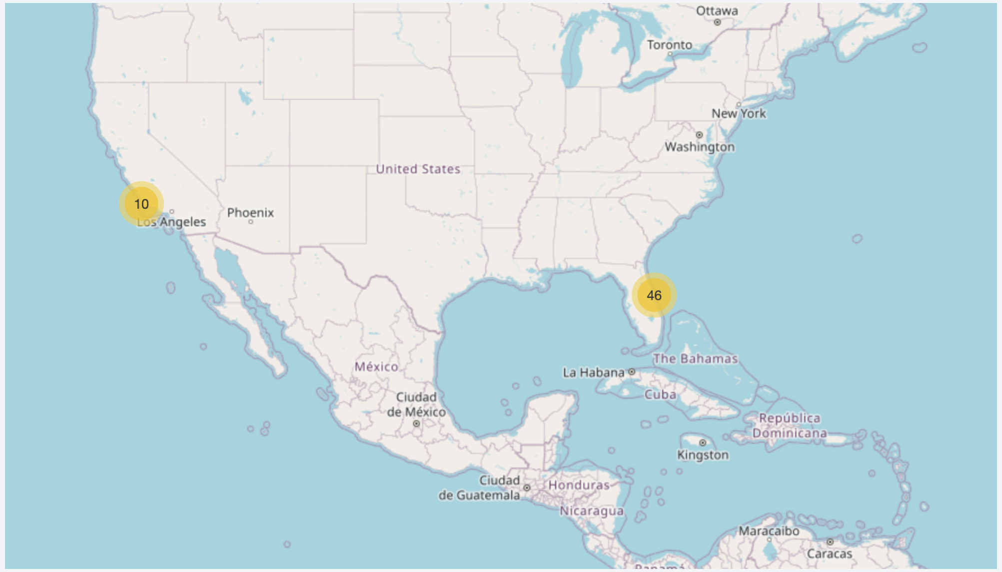
All launch sites' location markers on a global map

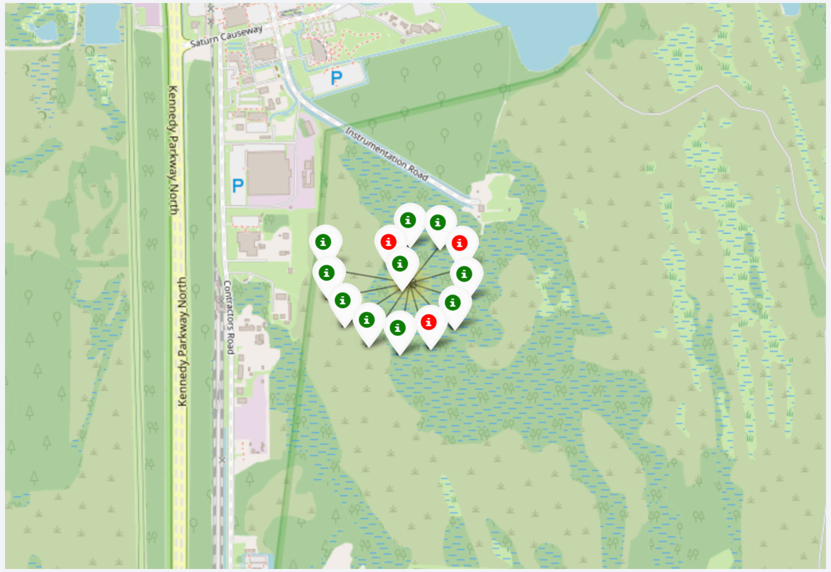
- There are in total 4 launch sites' location, and all these locations are very close to the coasts.



Color-labeled launch outcomes on the map

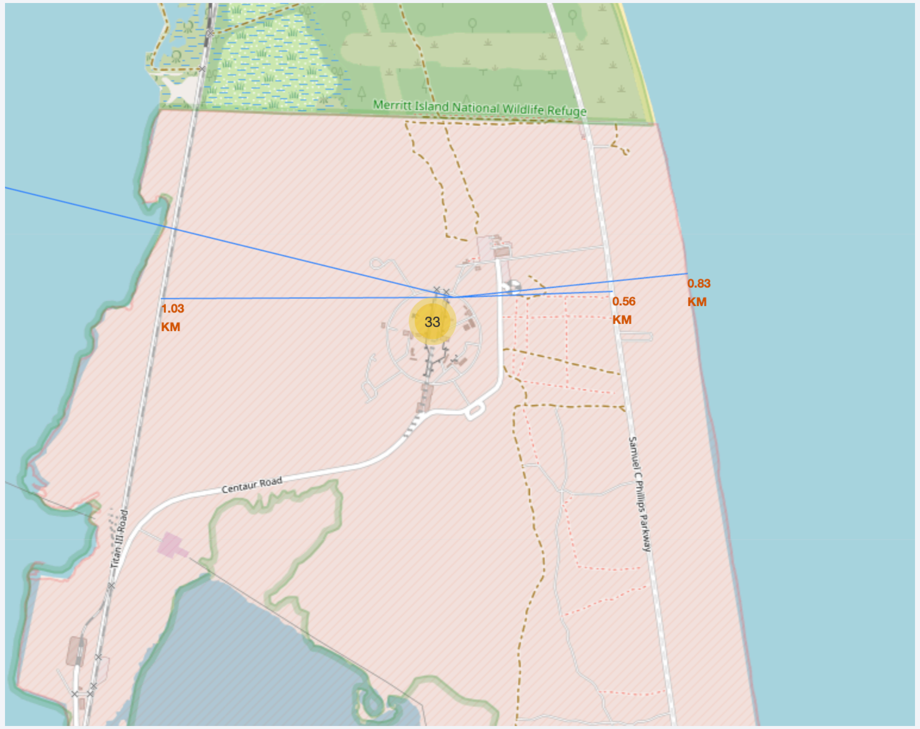
- There are more launch attempts on the launch site near east coast.
- We can get success rate for each launch site from the map.





Selected launch site to its proximities (Distances)

- The selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- These launch sites are close to coastline, railway and highway, but far away from cities.
- Geographic characters guarantee the safety of rocket launches.

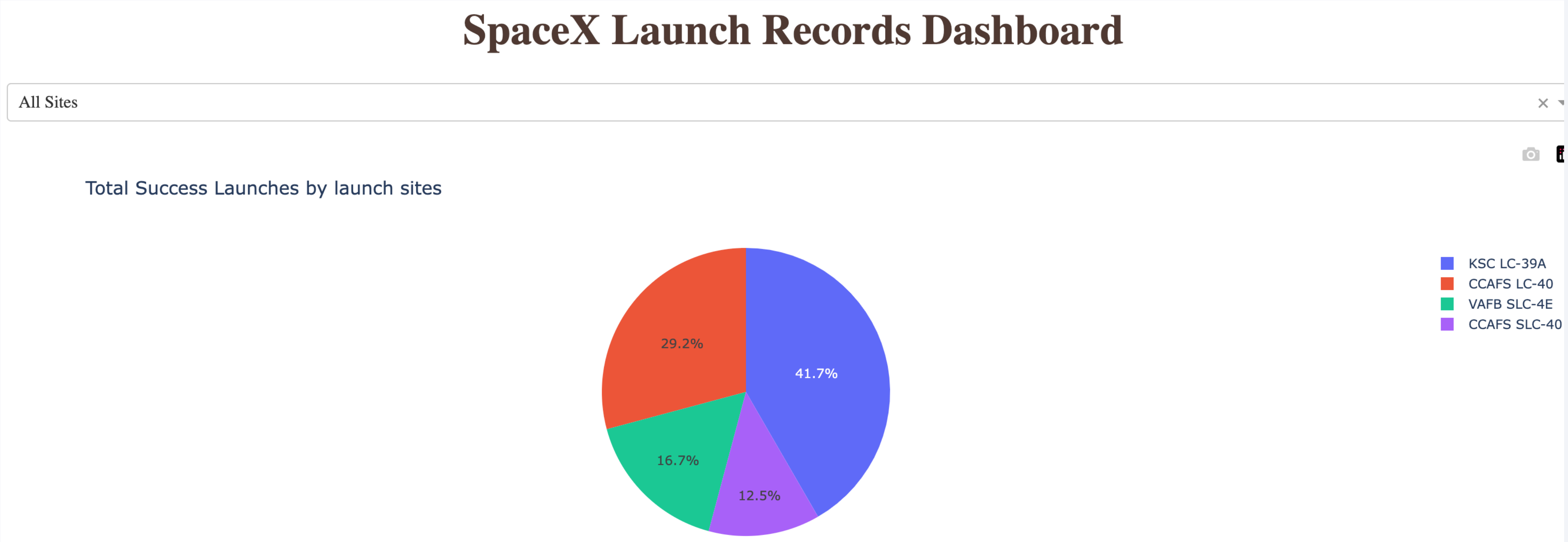


Section 4

Build a Dashboard with Plotly Dash

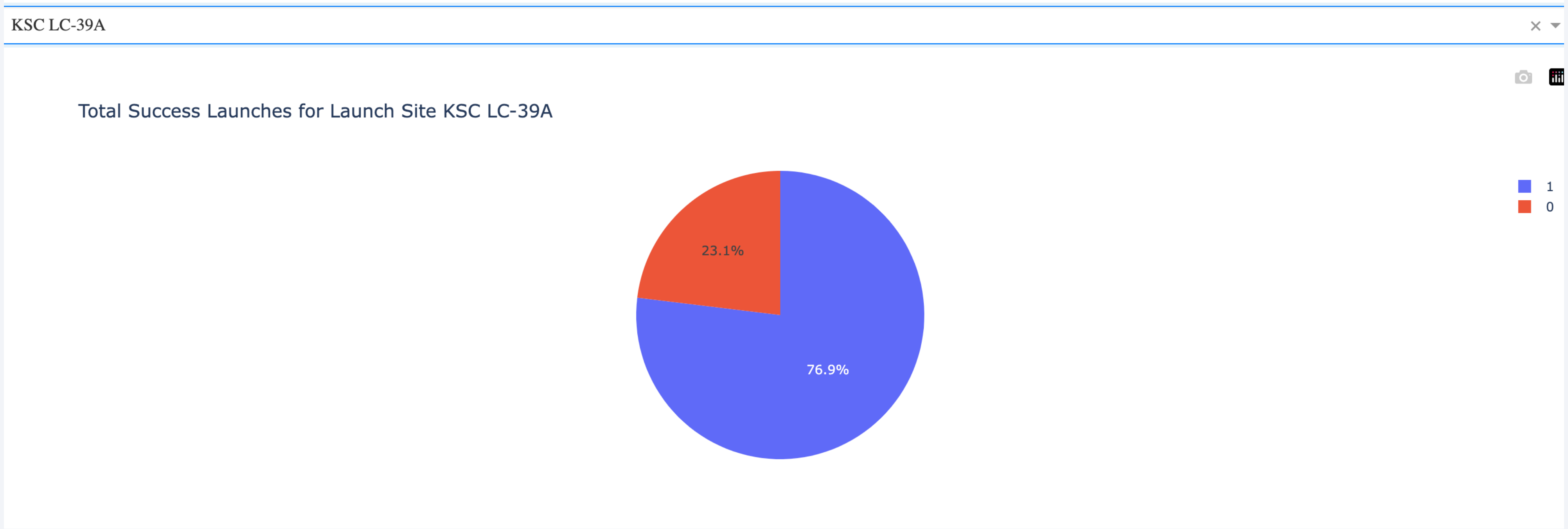
Total Success Launches by Launch Sites

- The KSC LC-49A launch site has the best success rate
- The CCAFS SLC-40 launch site has the worst success rate



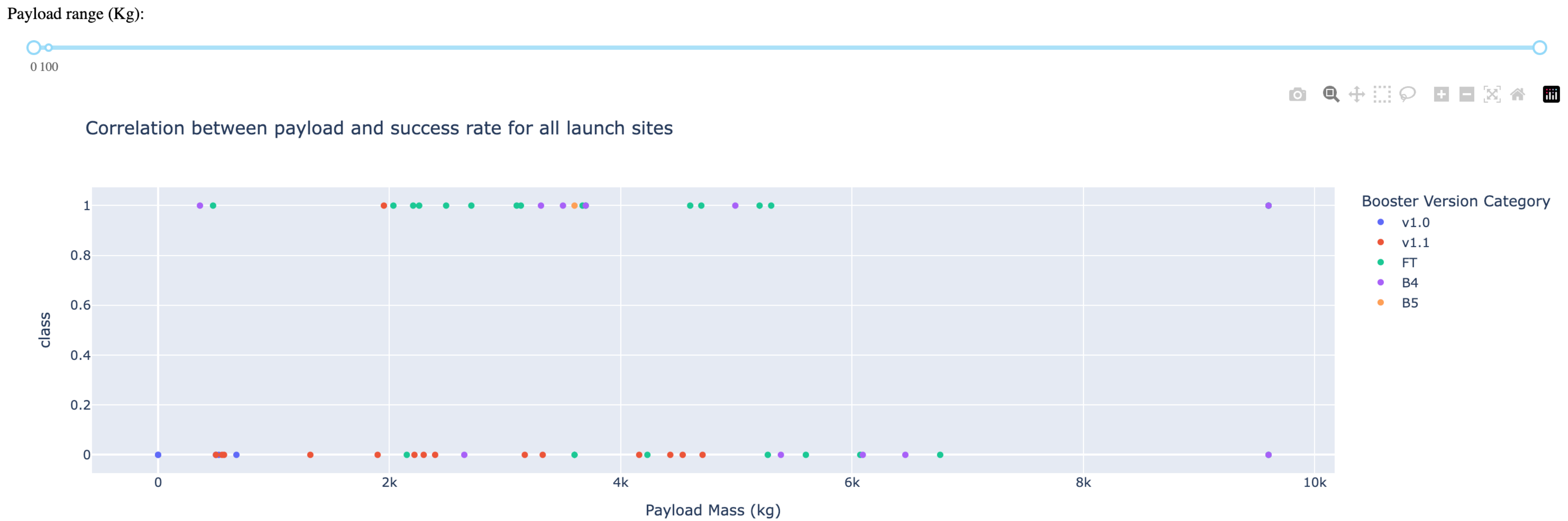
Total Success Launches for Launch Site KSC LC-39A

- Over 70 percent of all launches on launch site KSC LC-39A are successful landing



Payload vs. Launch Outcome Scatter Plot for All Sites

- Booster version FT has the best success landing rate
- V1.1 has the worst success landing rate



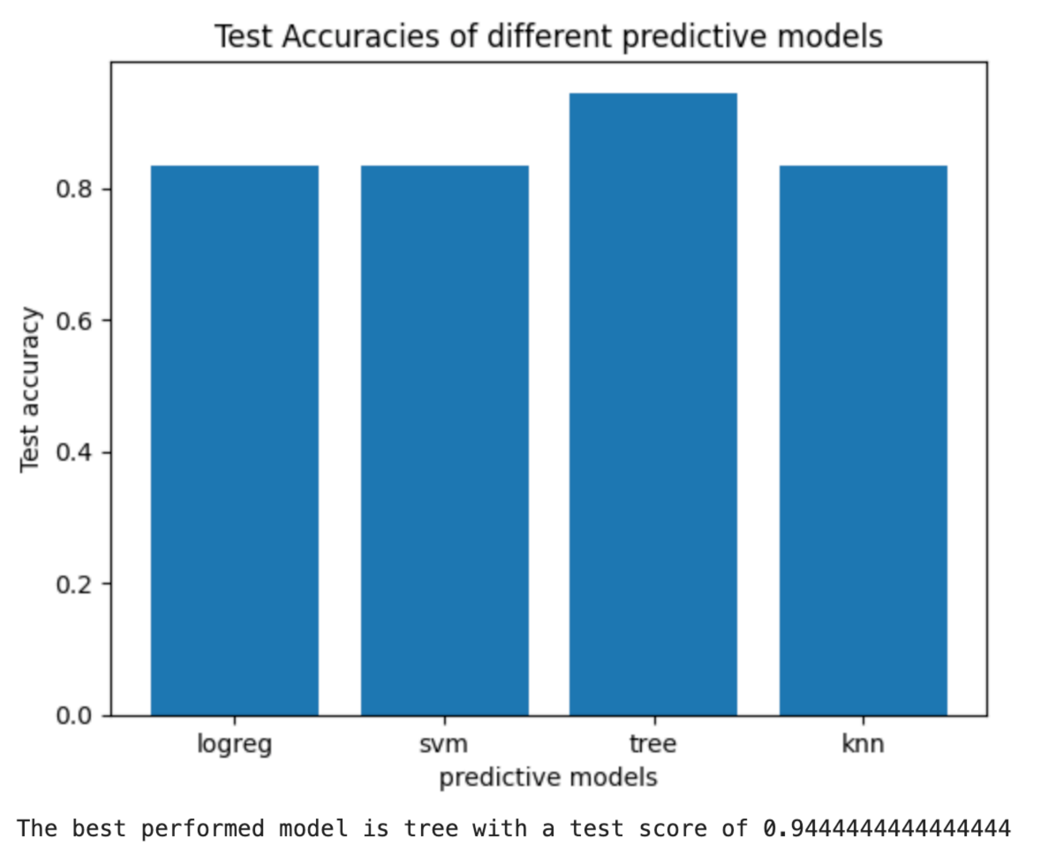
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Bar chart of model test accuracy for all built classification models

- The best model is the Decision Tree model with 0.944 test accuracy



Confusion Matrix

- For the Decision Tree model, only one launch that is actually failed was misclassified to success, there were no success launch have been misclassified

For this tree model, only one launch that is actually failed was classified to success, there were no success landing have been misclassified.

Conclusions

- Orbit type and payload mass are key features that affecting the outcome of each launch
- Geographic characters of launch sites guarantee the safety of rocket launches.
- The Decision Tree model has the best prediction performance.

Appendix

- All relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets:
  - https://github.com/schickwu/ds-capstone-template-coursera/tree/main

Thank you!