IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ivy Odoemenam
21st July, 2024

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory data analysis using visualization and SQL
- Perform interactive visual analytics with Folium and Plotly Dash
- Perform predictive analysis using classification models:
  - LR, SVM, DT, KNN models

Summary of all results

- EDA results
- Interactive analysis results
- Predictive analysis

Introduction

Project background and context

- SpaceX advertised Falcon 9 rocket launches on their website which costs several millions of dollars. They can save a lot of the first stage of the Falcon 9 will land successfully.

Problems you want to find answers

- Will the first stage of the SpaceX Falcon 9 rocket land successfully?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- SpaceX Rest API
- Web scraping from wikipedia

Perform data wrangling

- One Hot Encoding categorical data
- Data cleaning of null data and selecting only relevant columns

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Use GridSearchCV to build, tune, evaluate the classification models.

Data Collection

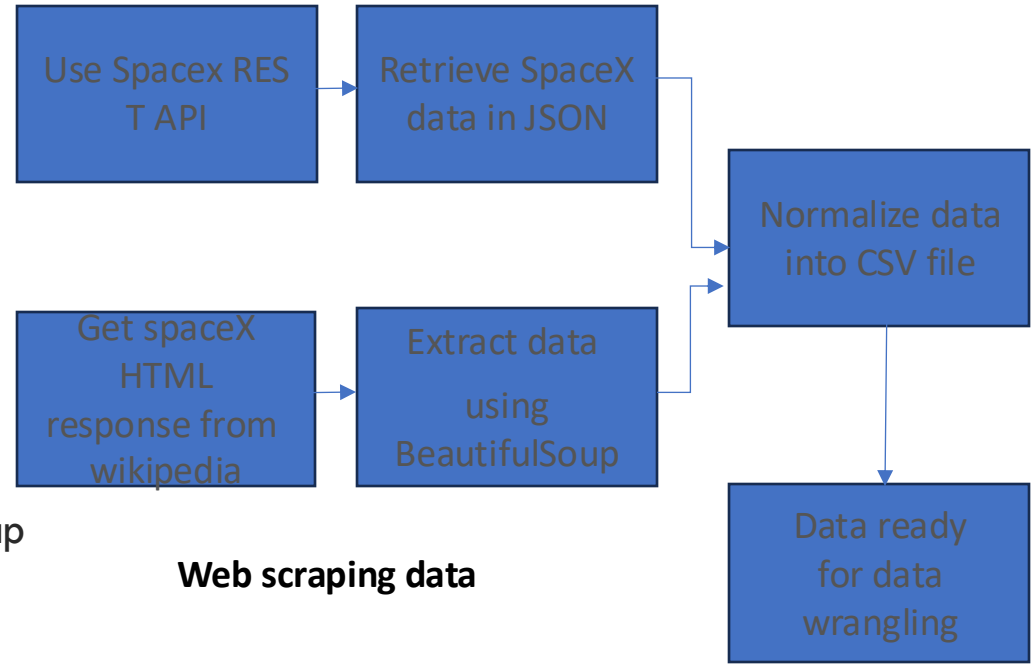
SpaceX REST API

- Retrieve spacex JSON file using the spaceX API that starts with api.spacexdata.com/v4
- Normalize the json file into csv file.

Web scraping

- Get spaceX HTML response from wikipedia.
- Extract relevant data using BeautifulSoup object.
- Normalize data into csv file

SpaceX REST API



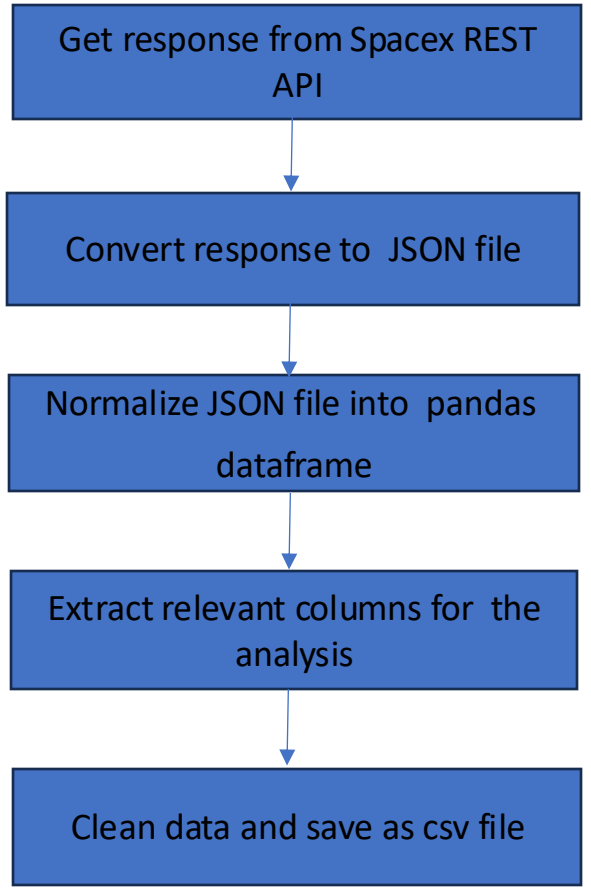
Web scraping data

Data Collection – SpaceX API

SpaceX REST API

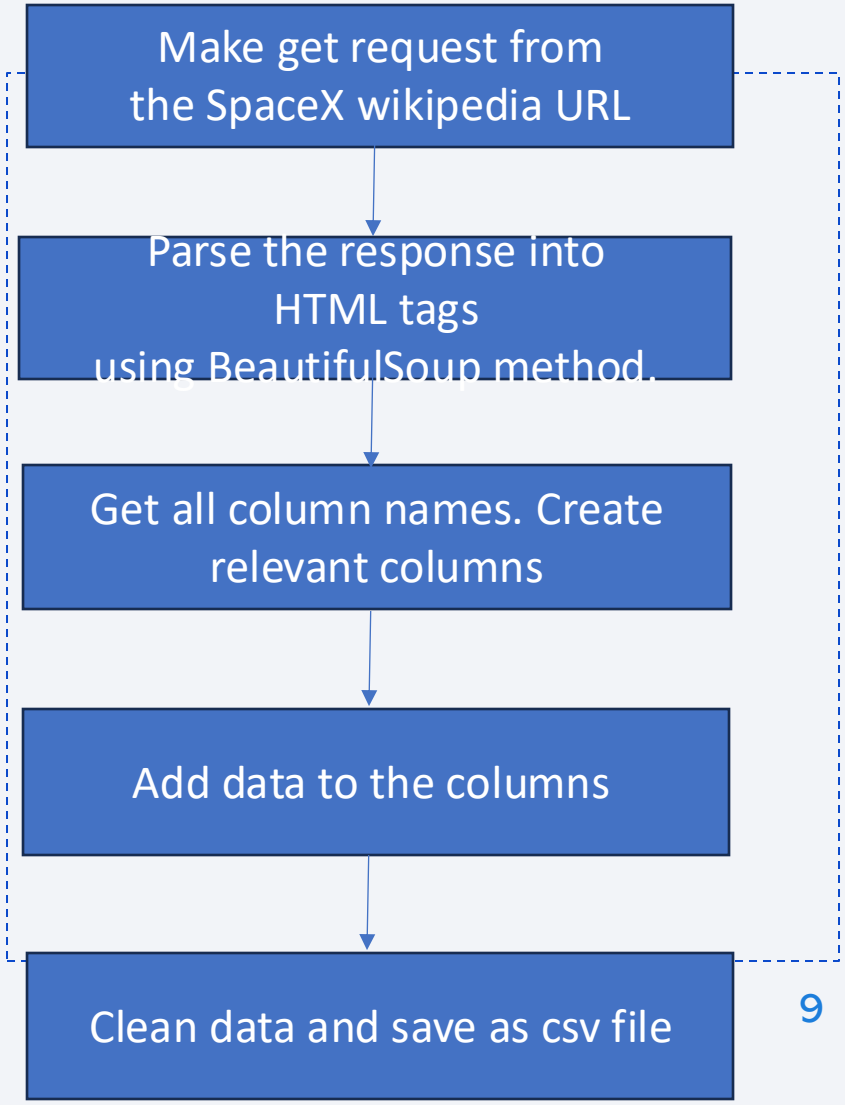
- API: api.spacexdata.com/v4
- Retrieve spacex data in JSON using the API
- Normalize the JSONfile into a pandas dataframe
- Data is ready for data wrangling

https://github.com/Ivy-Odoemenam/IBM-Datascience-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

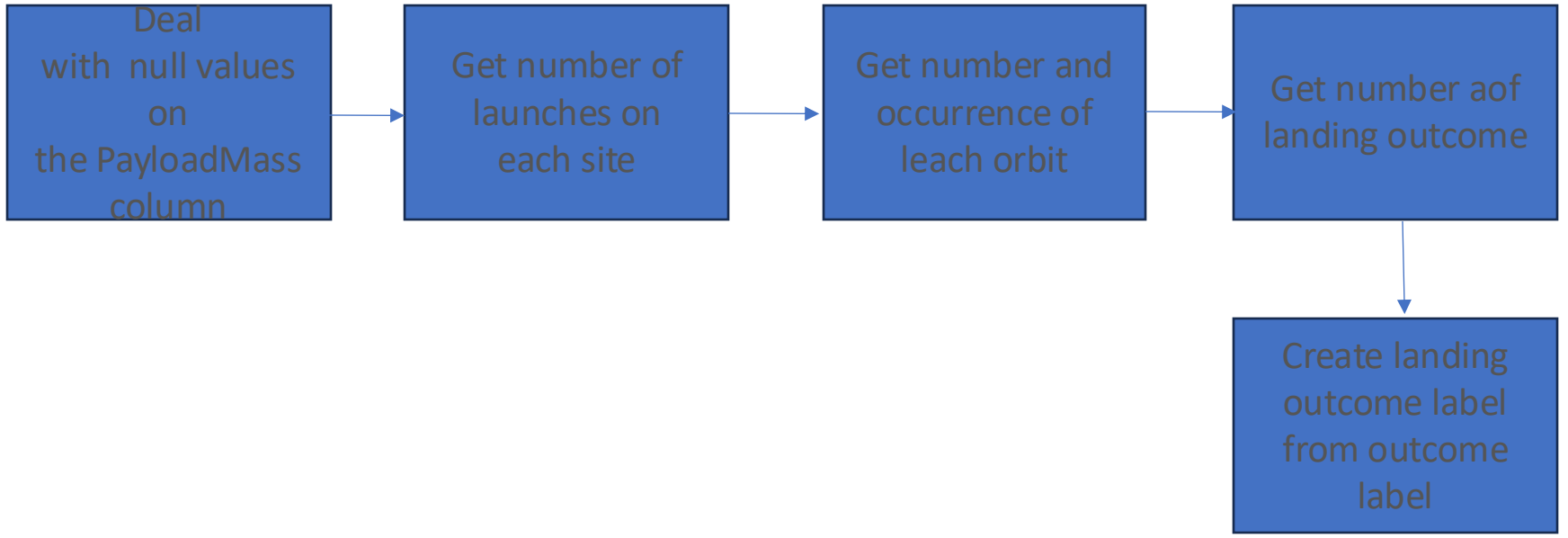


Data Collection - Scraping

- Make get request from the SpaceX wikipedia URL.
- Parse the response into HTML tags using BeautifulSoup method.
- Get all column names, remove irrelevant column and create relevant columns
- Extract data from the HTML table and append to columns as keys.
- Save the dataframe as csv file
- https://github.com/Ivy-Odoemenam/IBM-Datascience-Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling



https://github.com/Ivy-Odoemenam/IBM-Datascience-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- https://github.com/Ivy-Odoemenam/IBM-Datascience-Capstone-Project/blob/main/edadataviz.ipynb

EDA with SQL

SQL queries you performed:

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

https://github.com/Ivy-Odoemenam/IBM-Datascience-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Create a map with a central location
- Add circles and markers of all launch sites
- Add lines such as distance from a launch site to the closest coastline, rail, highway, and city.

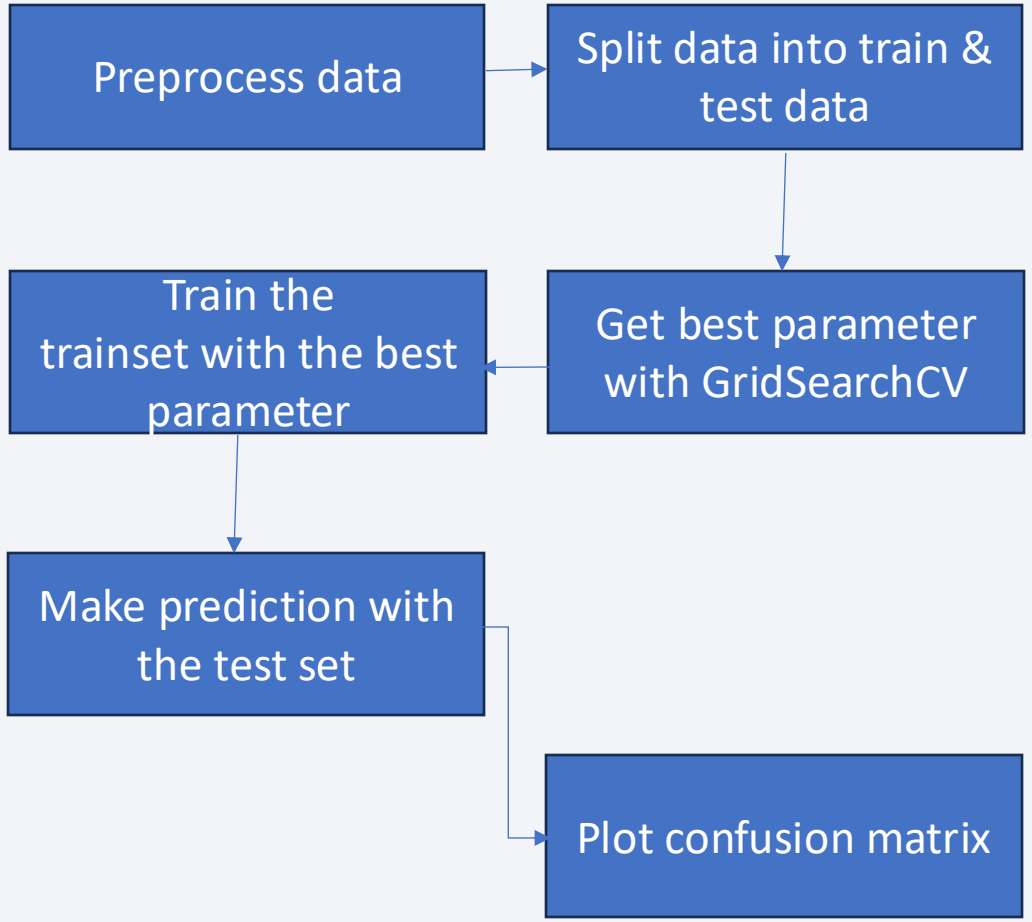
- https://github.com/Ivy-Odoemenam/IBM-Datascience-Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Add a pie chart to dashboard to display the proportion of successful landing for each site when all is selected; or
- Display proportion of successful and unsuccessful landing for each selected site
- Add a slider to select different range of payload mass in kg
- Add a scatter plot to display the relationship between the selected payload mass and the outcome class
- https://github.com/Ivy-Odoemenam/IBM-Datascience-Capstone-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

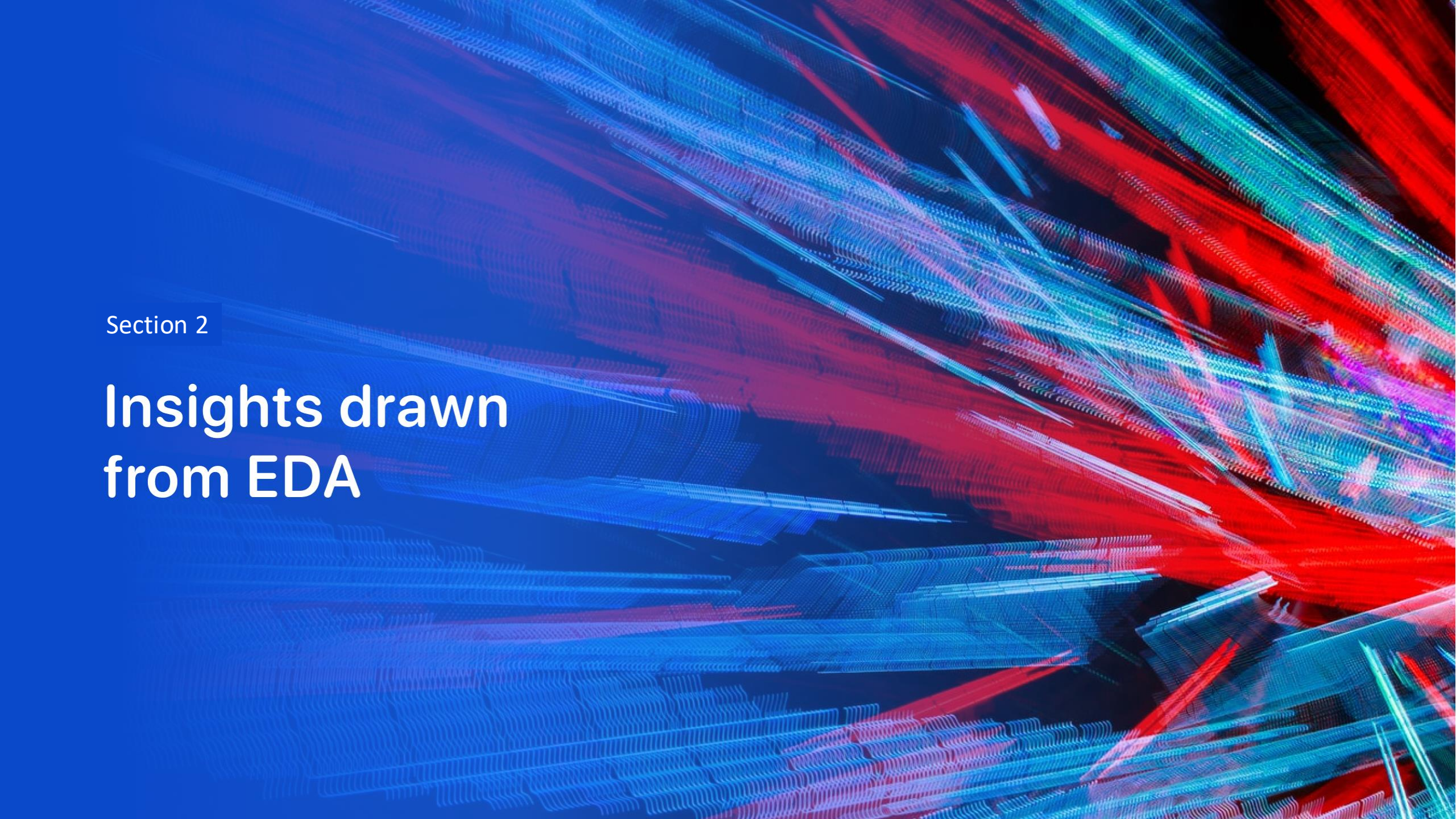
- Create the model object for logistic regression, support vector machine, decision tree classifier and K neighbors classifier.
- Pass a list of relevant parameters to GridSearchCV method
- Fit the train dataset to find the best performing parameters.
- Get the score of the model using the test datasets
- Make a prediction using the test dataset
- Plot the performance metrics using confusion matrix
- Compare the accuracy scores between all 4 models to decide the best performing classification model

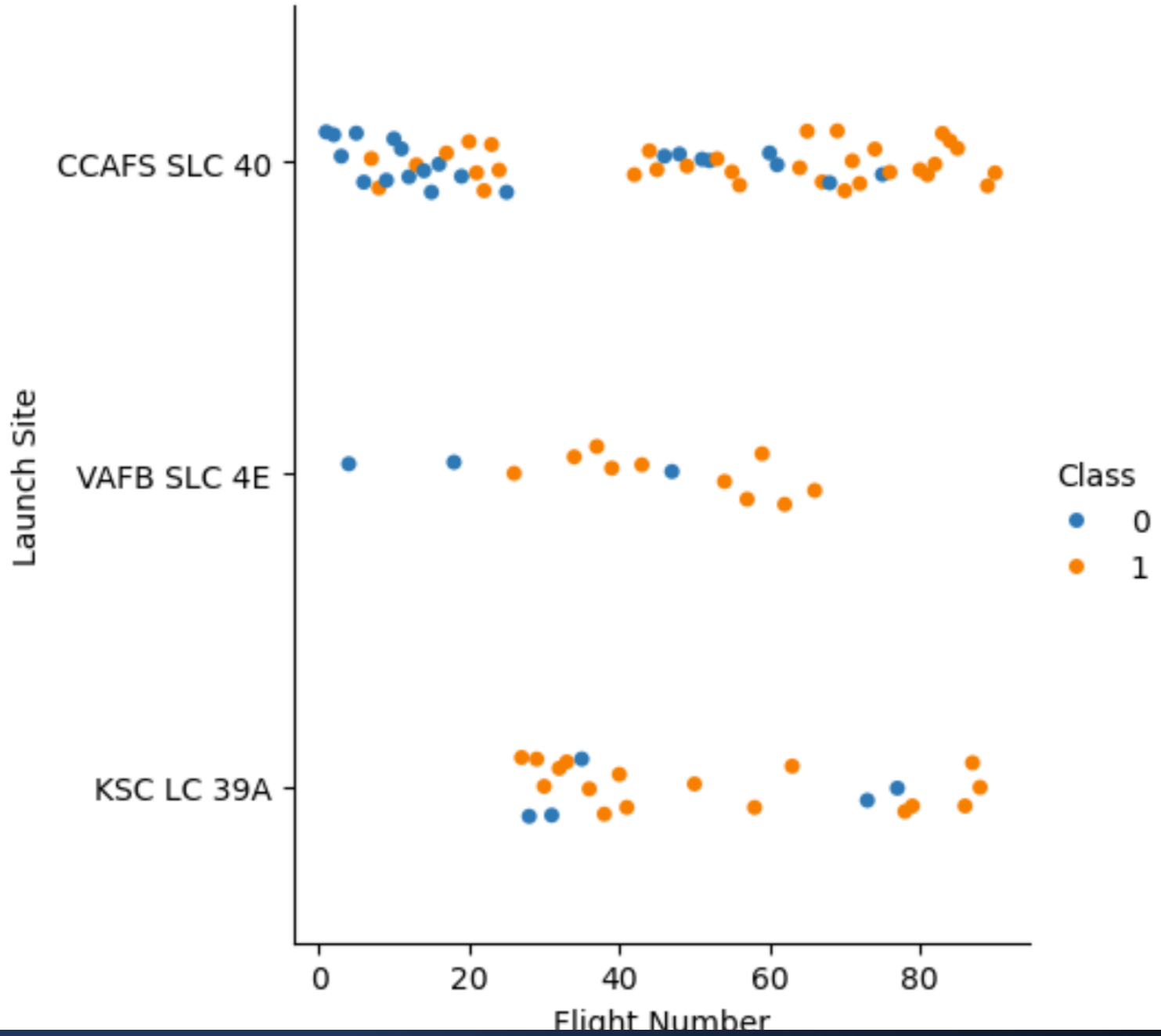


- https://github.com/Ivy-Odoemenam/IBM-Datascience-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

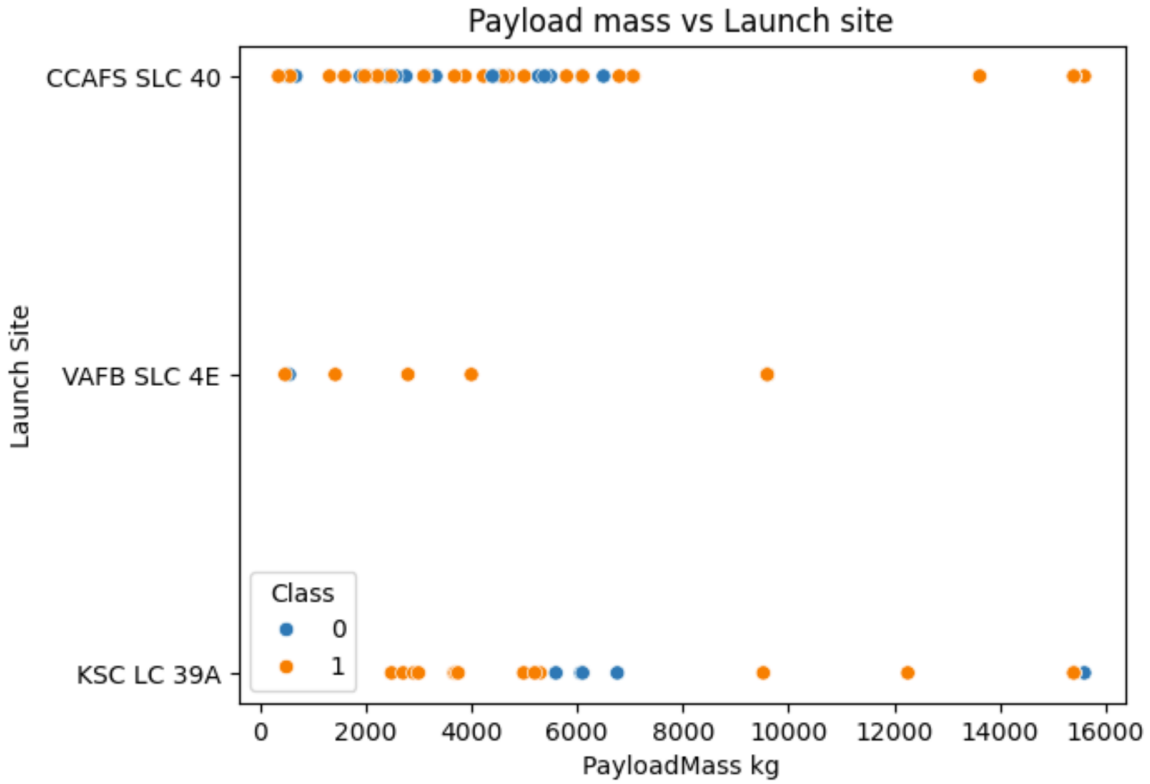

Section 2
Insights drawn from EDA



Flight Number vs. Launch Site

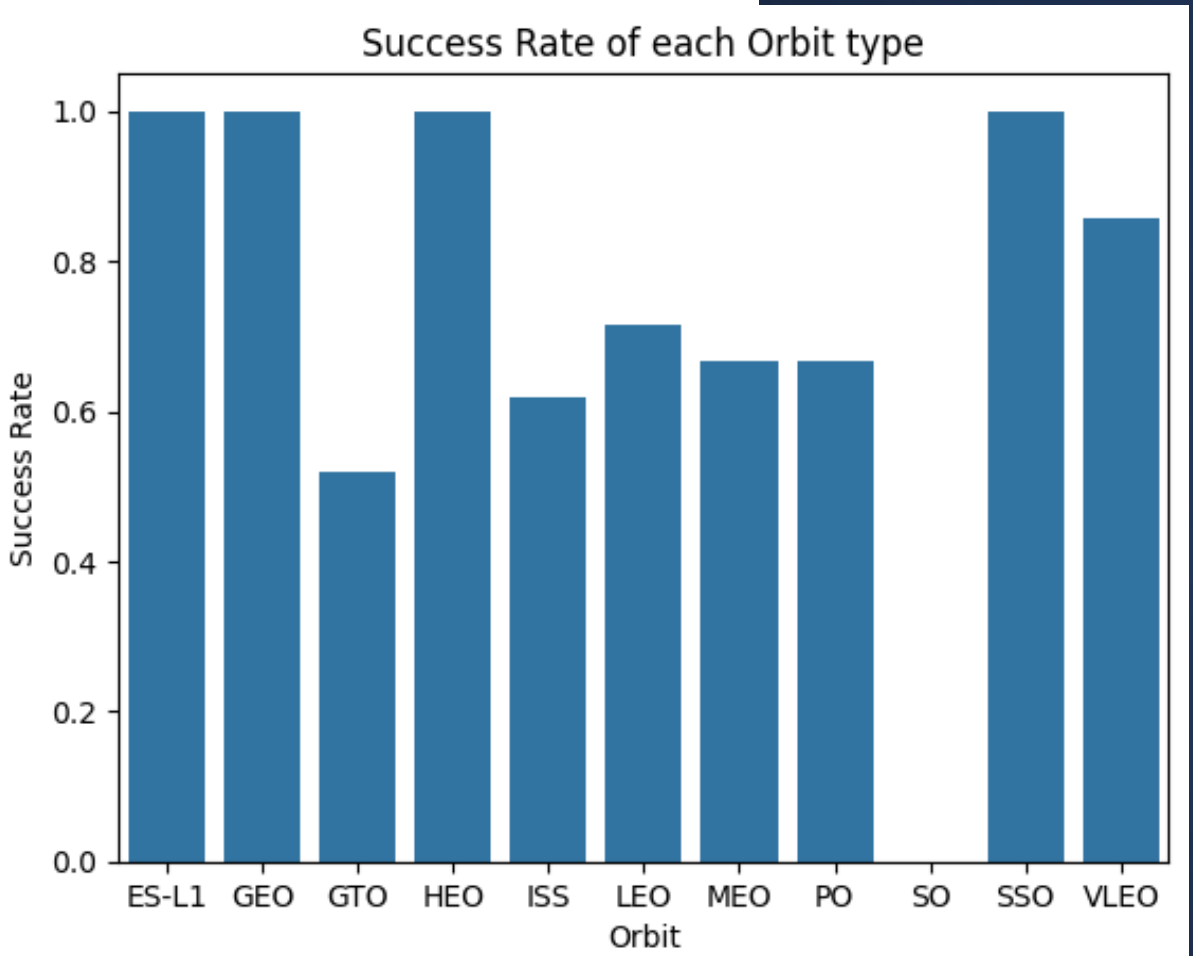
- Launches from the site of CCAFS SLC 40 are seen to be higher than launches from the other two sites.

Payload vs. Launch Site



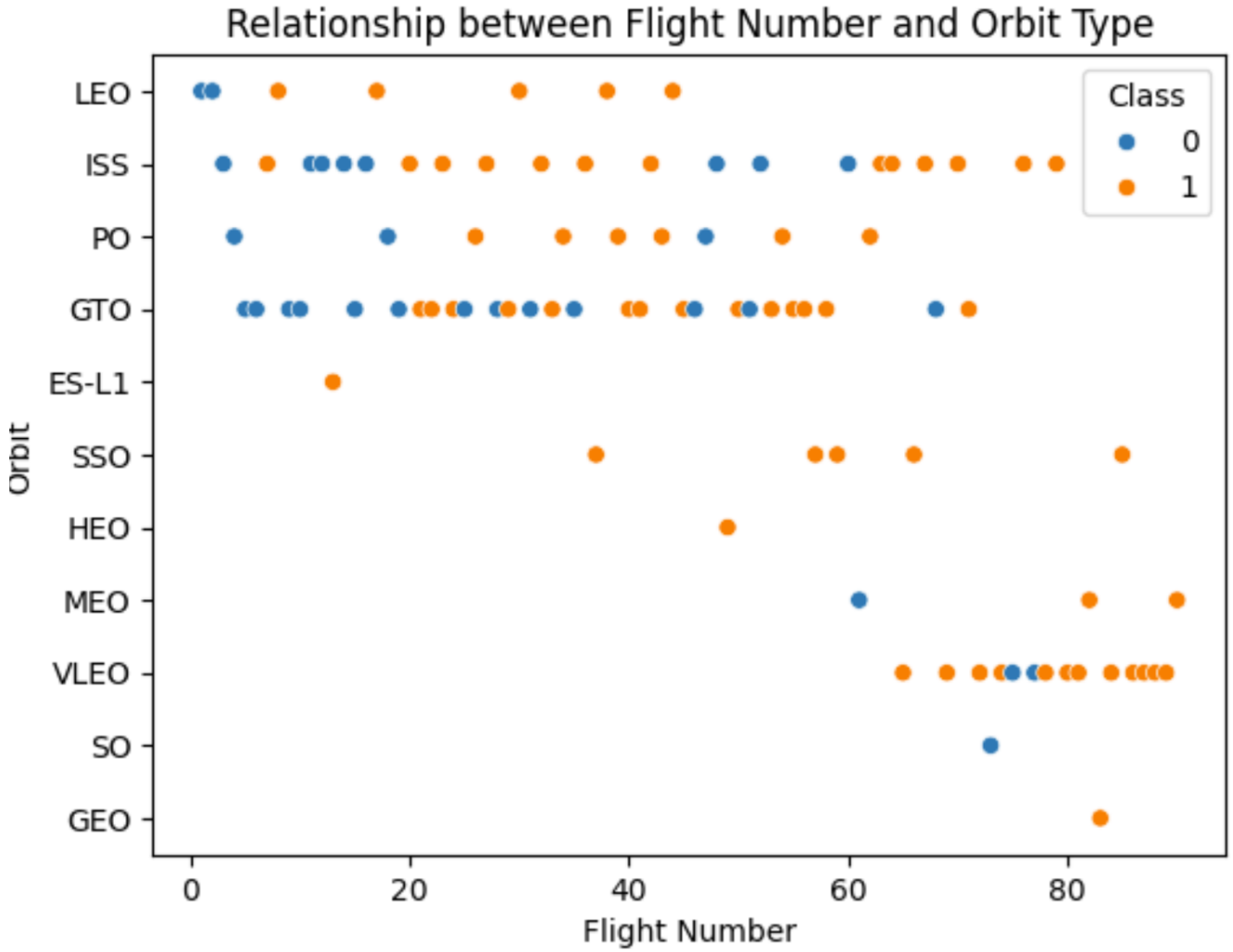
- Lower payload mass was launched from CCAFS SLC 40 and KSC LC 39A

Success Rate vs. Orbit Type



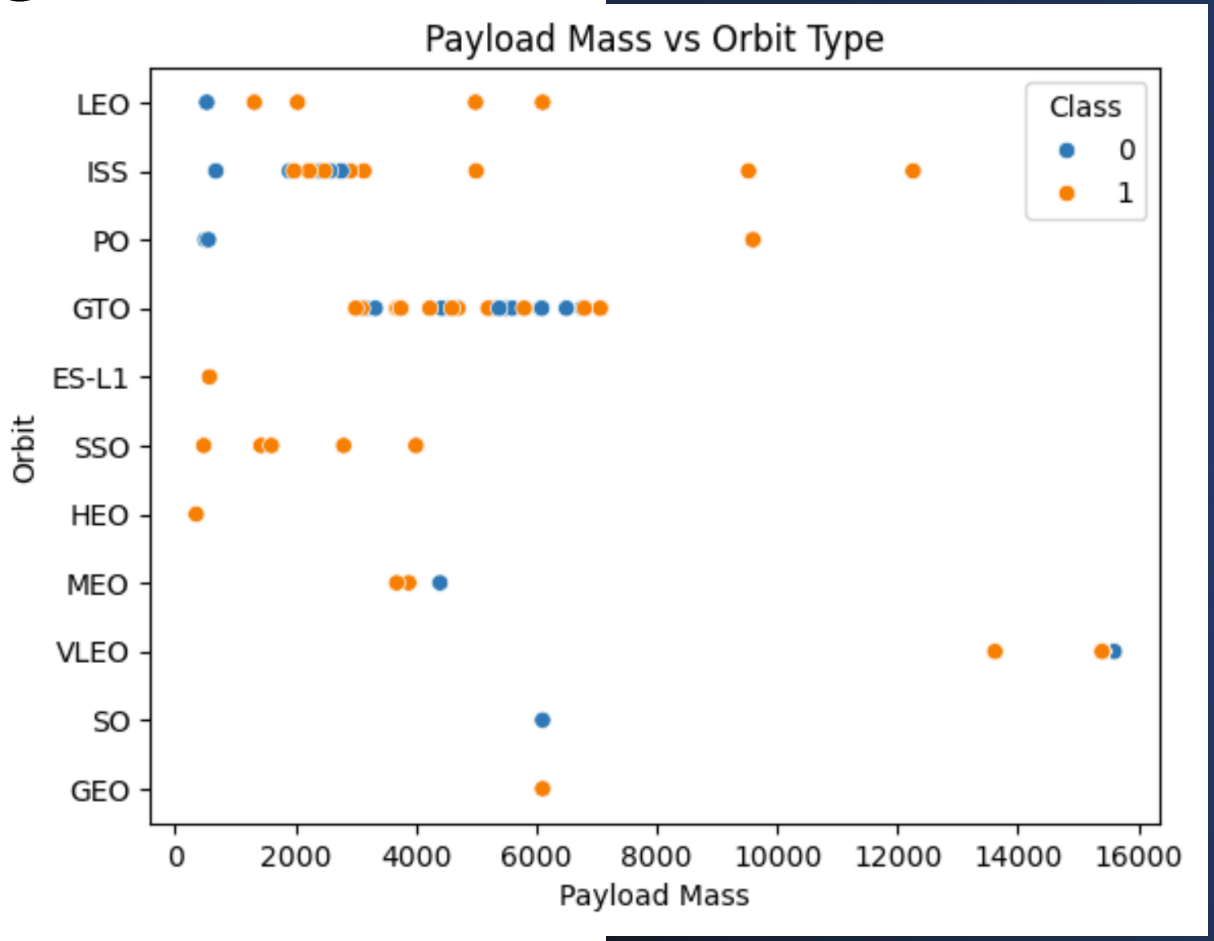
- Orbits ES-L1, GEO, SSO and VLEO made the highest number of successful launches

Flight Number vs. Orbit Type



- VLEO had more successful flights

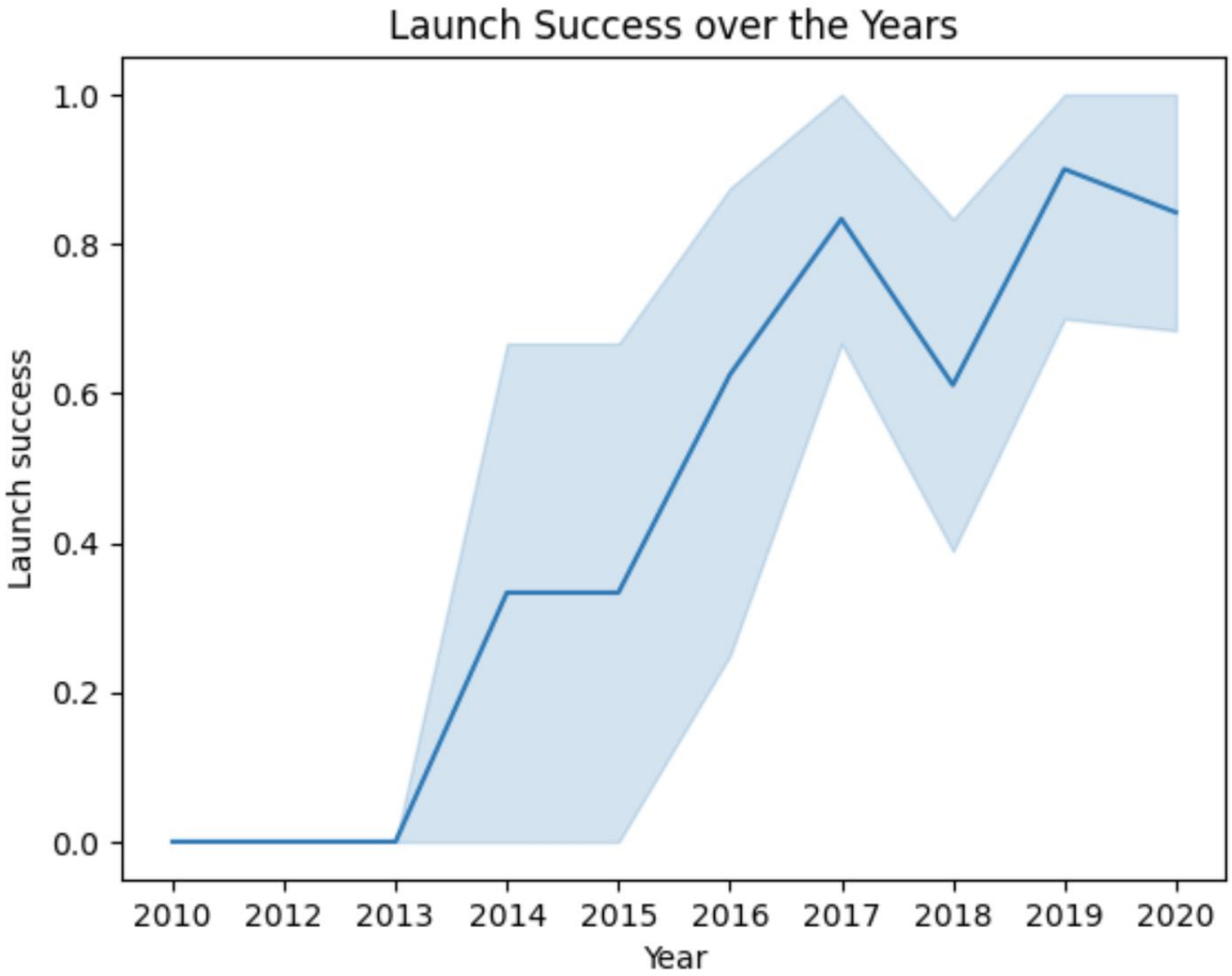
Payload vs. Orbit Type



- Low weighted payload performs better than heavier payloads

There is a significant increase in the rate of successful launches from 2013 to 2020

Launch Success Yearly Trend



All Launch Site Names

- %%sql select distinct "launch_site" from SPACEXTABLE

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Launch Site Names Begin with 'CCA'

- %%sql select * from SPACEXTABLE where Launch_site like 'CCA%' limit 5

Total Payload Mass

	sum(PAYLOAD_MASS__KG_)
0	45596

- %%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'

Average Payload Mass by F9 v1.1

	avg(PAYLOAD_MASS__KG_)
0	2928.4

- %%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'

First Successful Ground Landing Date

	min(date)
0	2015-12-22

- %%sql select min(date) from SPACEXTABLE where Landing_Outcome = "Success (ground pad)"

Successful Drone Ship Landing with Payload between 4000 and 6000

- %%sql select Booster_Version from SPACEXTABLE where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000

	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

	Mission_Outcome	count(*)
0	Failure	1
1	Success	100

- %%sql select substr(Mission_Outcome,1,7) as Mission_Outcome, count(*) from spacextable group by 1

Boosters Carried Maximum Payload

- %%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

2015 Launch Records

- %%sql select substr(Date, 6, 2) as month, Landing_Outcome, Booster_Version, Launch_Site from spacextable where Landing_Outcome='Failure (drone ship)' and substr(Date, 0, 5) = '2015'

	month	Landing_Outcome	Booster_Version	Launch_Site
0	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- **There are 2 failed drone ship landing outcome from launch site CCAFS LC-40 in the year 2015**

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

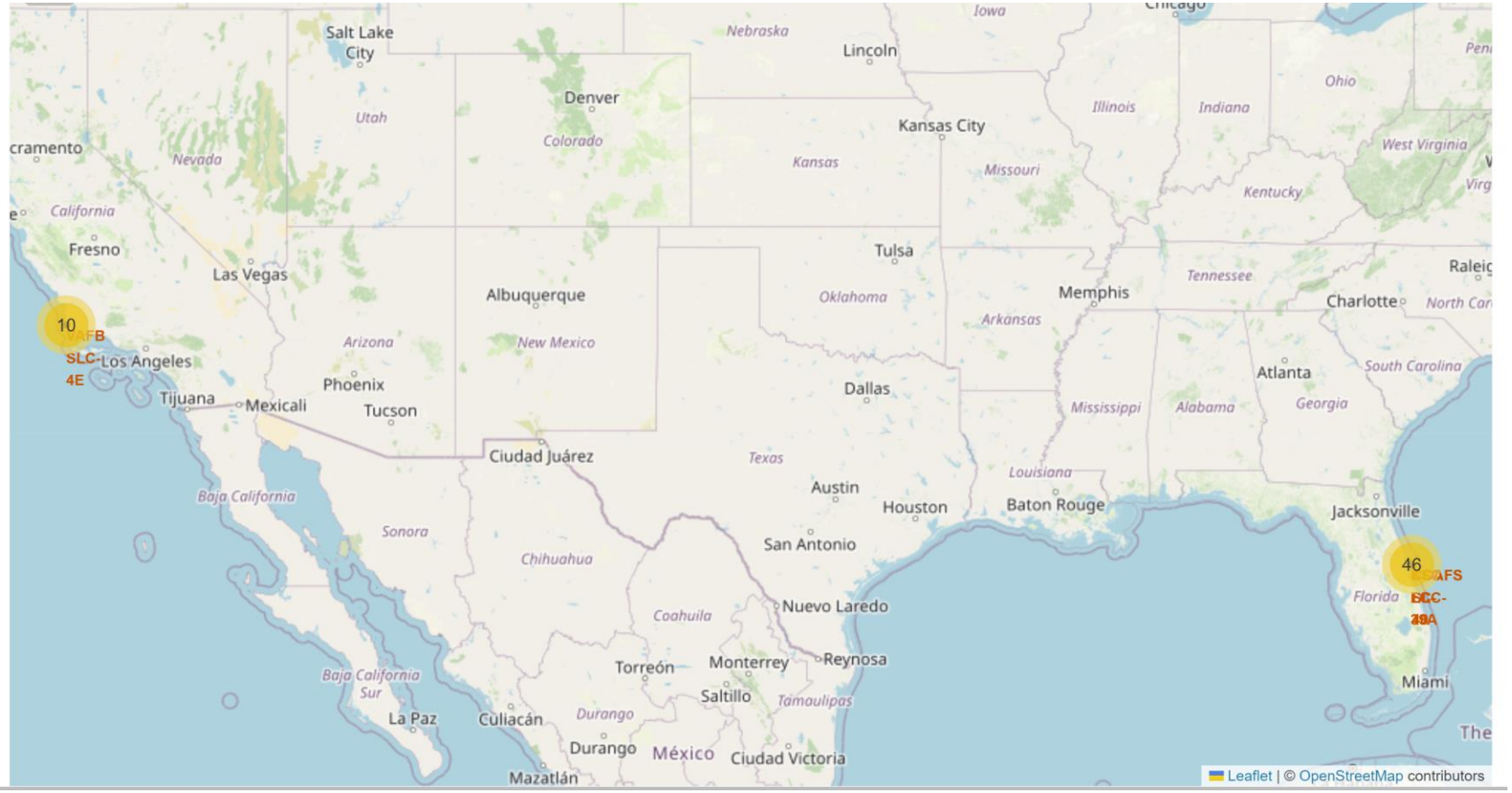
- %%sql select Landing_Outcome, count(*) as 'Count' from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Count desc

	Landing_Outcome	Count
0	No attempt	10
1	Success (drone ship)	5
2	Failure (drone ship)	5
3	Success (ground pad)	3
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Failure (parachute)	2
7	Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

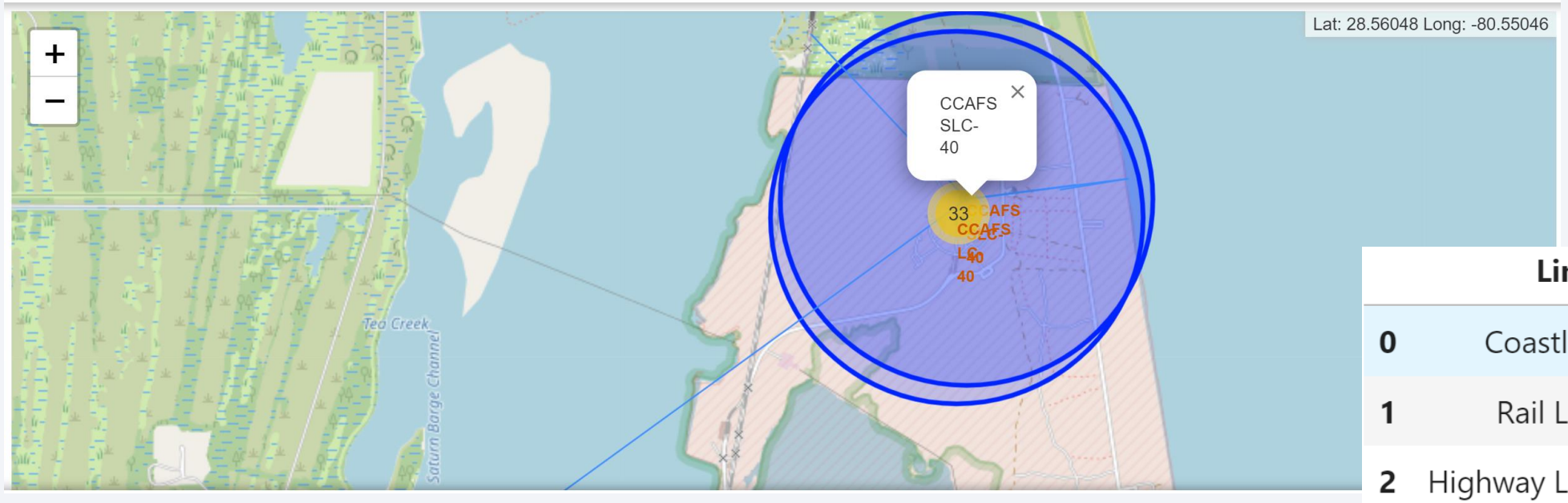
All Launch Sites Location Markers on Global map



Color-Labeled Launch outcome from CCAFS LC 40

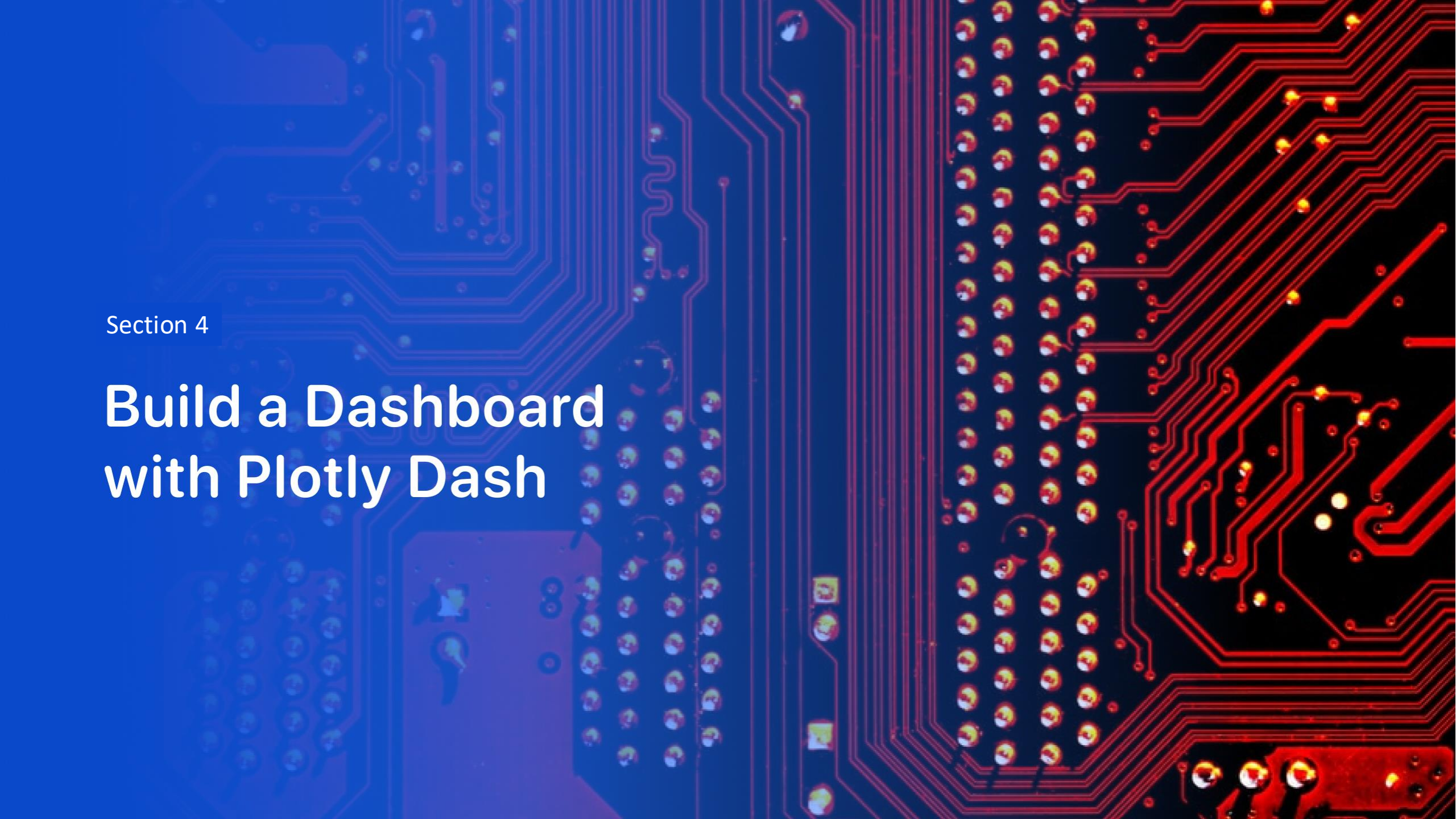
- There are a total of 26 launches from launch-site CCAFS LC 40, with 7 successful launches and 19 failed launches.

Launch Site CCAFS SLC 40 to its proximities

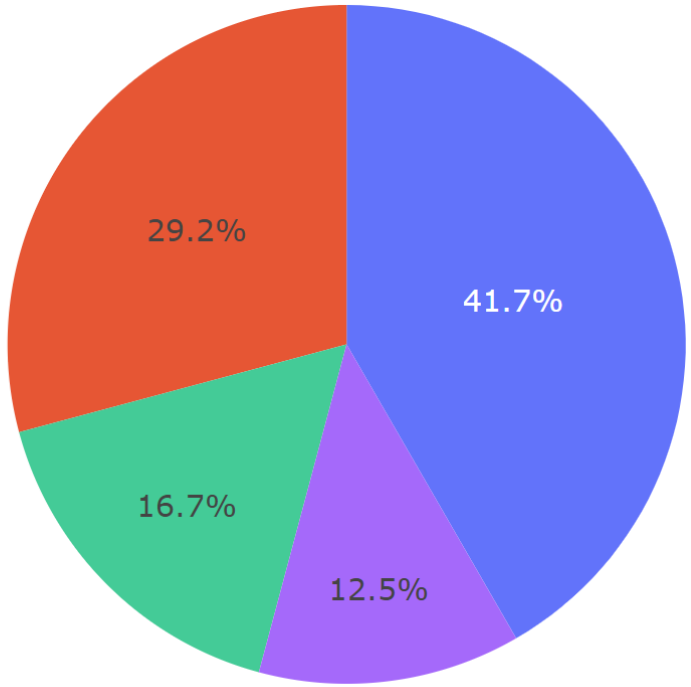


	Lines	Distance
0	Coastline	0.869065
1	Rail Line	1.212606
2	Highway Line	0.589073
3	City Line	8.738924

- CCAFS SLC 40 is closest to the highway, compared to other proximities
- https://github.com/Ivy-Odoemenam/IBM-Datascience-Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb


Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Site Success Distribution (All Sites)



Interactive Dashboard: all Sites Successful Launches

- **KSC LC-39A had the highest proportion of successful launches at 41.7%, while CCAFS SLC-40 has the least proportion of successful launches at 12.5%**

SpaceX Success Rate of KSC LC-39A Launch Site

Launch Site with Highest Launch Success

- **KSC LC 39A recorded the most successful launches from all sites**

All sites Payload vs. Launch Outcome Scatter plot

- Booster version FT has the most success rate with lower payload mass that ranges from 0kg to 3000kg, while booster version v1.1 has the most unsuccessful given the same payload mass.

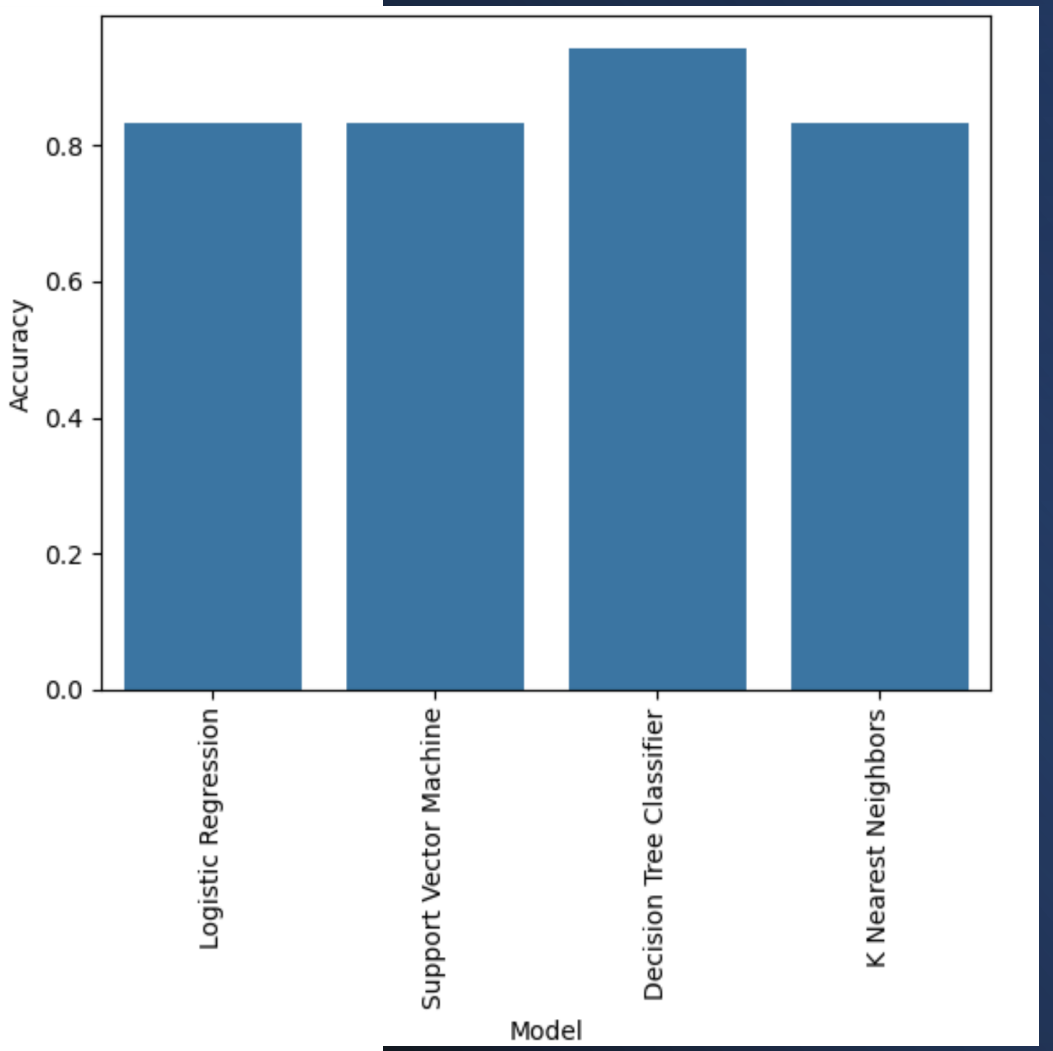
- Booster version FT has the most success rate with lower payload mass that ranges from 4000kg to 9000kg

Section 5

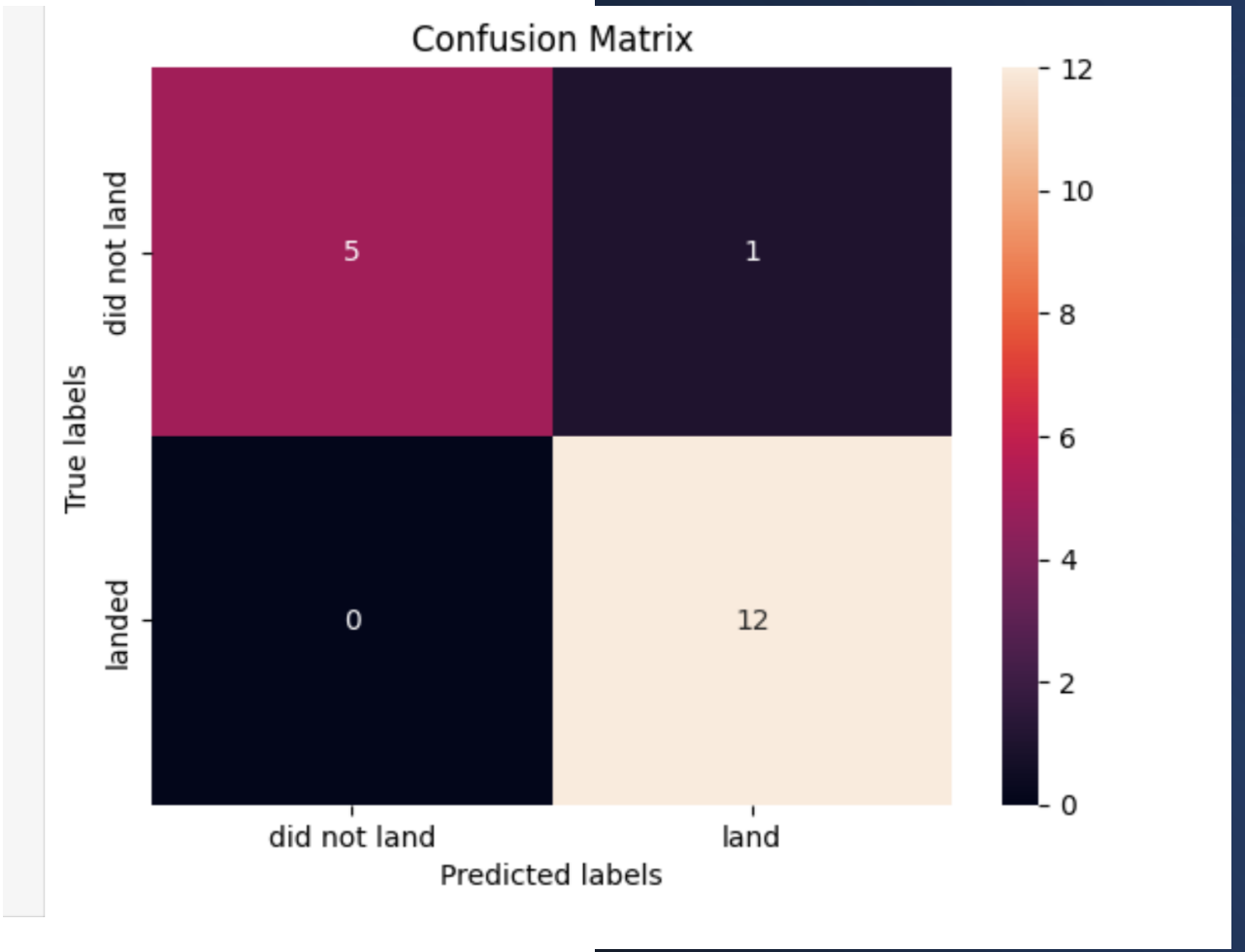
Predictive Analysis (Classification)

Classification Accuracy

- The decision tree classification model performs better with the spaceX dataset prediction.



Confusion Matrix



- Confusion matrix of the decision tree classifier model

Conclusions

The decision tree classifier matrix is the best performing model in terms of prediction accuracy for this dataset.

KSC LC 39A recorded the most successful launches from all sites.

Low weighted payload performs better than heavier payloads.

Orbits ES-L1, GEO, SSO and VLEO made the highest number of successful launches.

More successful launches over the years.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!