IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alejandro Monroy-Azpeitia
April 3rd, 2024

Outline

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

Executive Summary

Summary of methodologies:

- Data used for this work was collected using public data from SpaceX API and Wikipedia, then Data wrangling was doing using Pandas and SQL, Exploratory Data Analysis was doing using Pandas.
- Data visualizations were created using Folium, Matplotlib and Seaborn, interactive dashboard were created using Dash and Plotly
- Using and comparing Machine Learning techniques for classification as Logistic Regression, Support Vector Machine, Decision Trees, k-Nearest Neighbors, the final model was created.

Summary of all results:

- Informative visualizations to understand the data was created.
- Interactive dashboards and analytics were created.
- The best model is Decision Tree

Introduction

Project background and context:

A rocket launch company wants to know wich launch characteristics has the best chance to be successful, instead of making a hard and specific model that considers all the variables, implement a Data-Science approach where the data available from SpaceX's launch aliments a Machine Learning Model for prediction (classification)

Problems you want to find answers

How is the Data available and what does it tell us about the failed or succeed of a launch? What is the best machine learning model for prediction? Will a new launch of this new company be successful?

Section 1

Methodology

Methodology

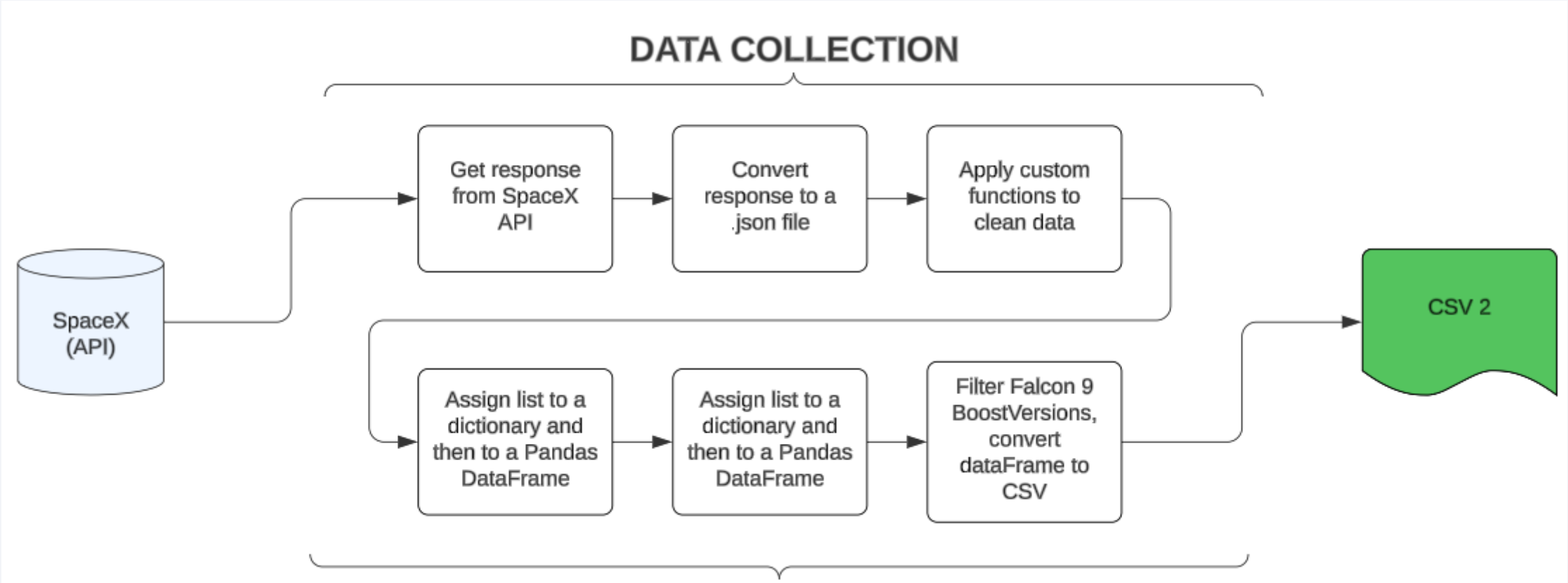
Executive Summary

- Data collection methodology: collected using public data from SpaceX API and Wikipedia
  - SpaceX Api (Python)
  - Web-Scrapping from the corresponded Wikipedia page
- Perform data wrangling
  - One-hot-encoding for categorical variables.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Making a train-test split, then train the models (Logistic Regression, SVM, KNN, Decision Tree) and finally evaluate it and comparing them

Data Collection

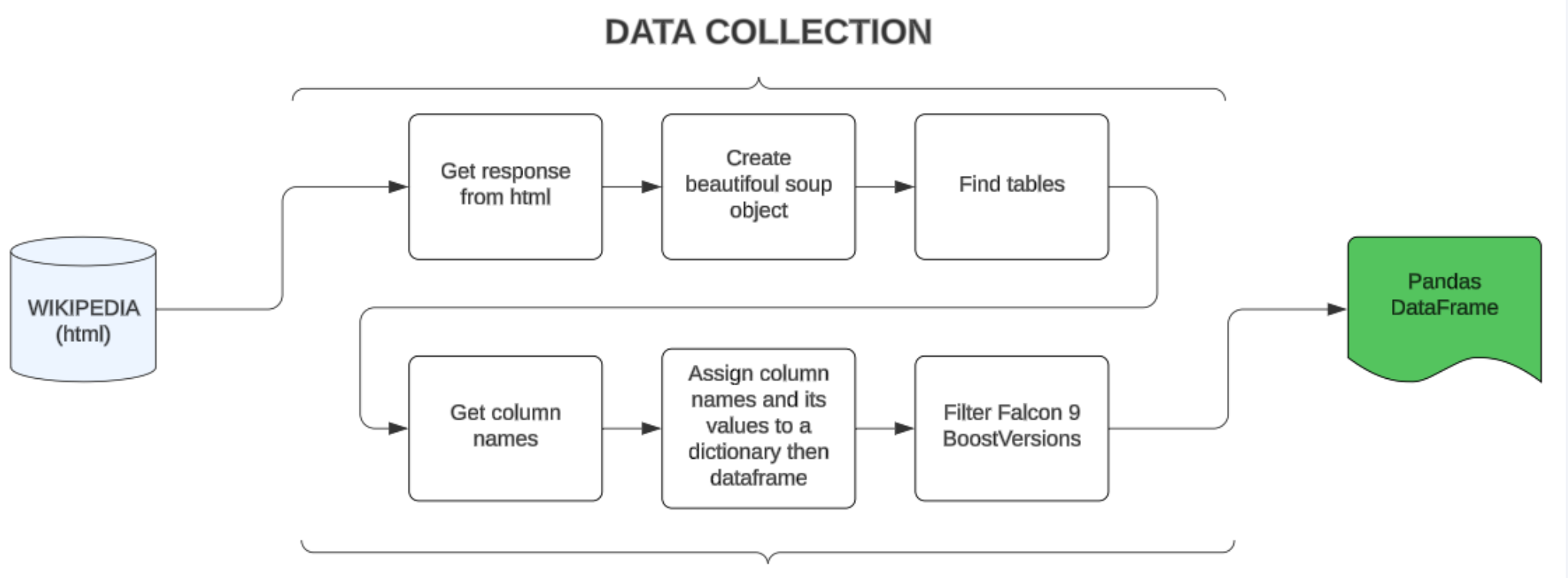
- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API



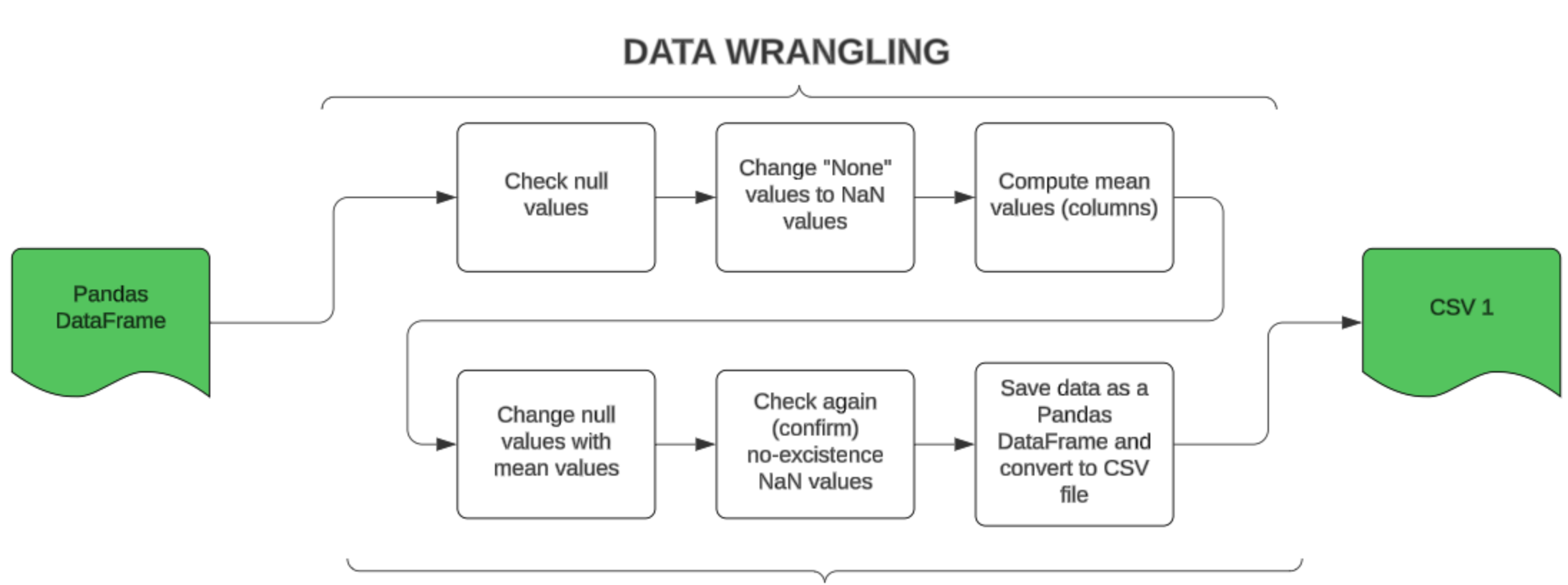
https://github.com/AzpMon/IBM-Final-course-DS-/blob/main/Spacex_data_collection_api.ipynb

Data Collection – Scraping (1/2)



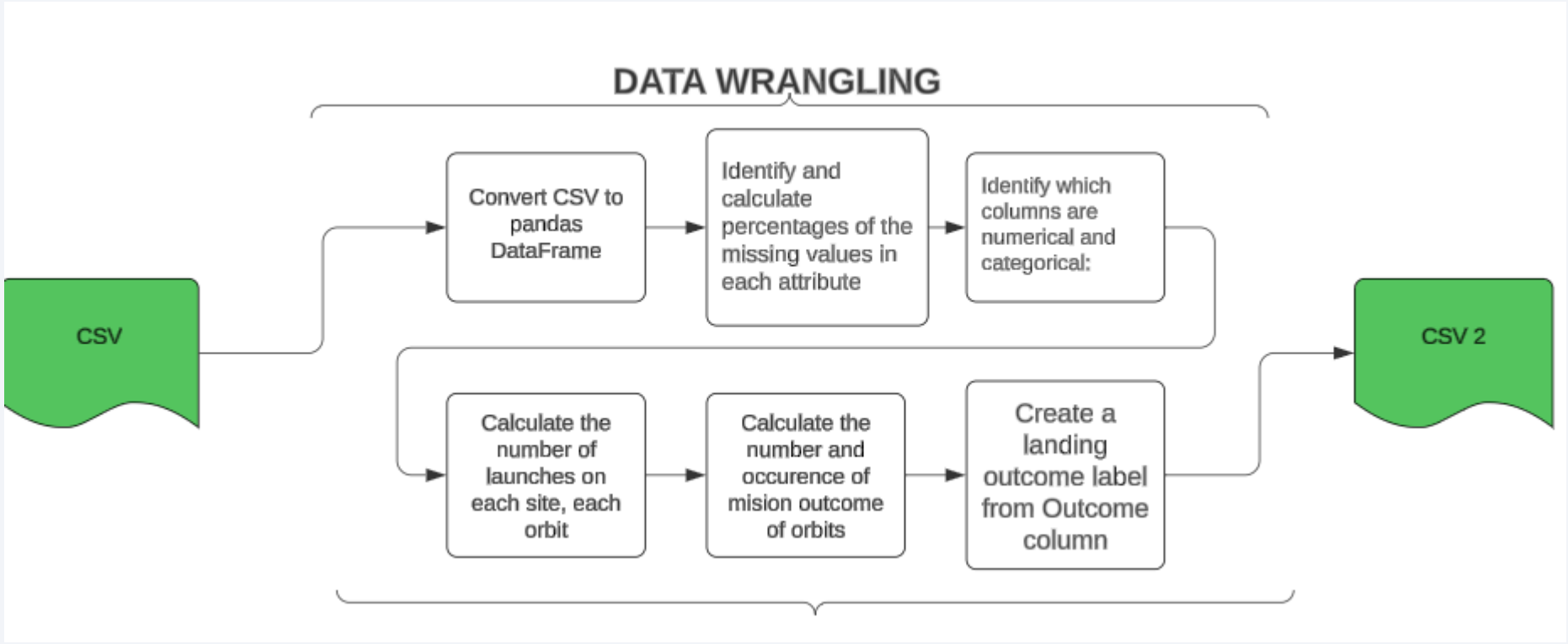
https://github.com/AzpMon/IBM-Final-course-DS-/blob/main/Spacex_data_webscraping.ipynb

Data Collection – Scraping (Wrangling part) (2/2)



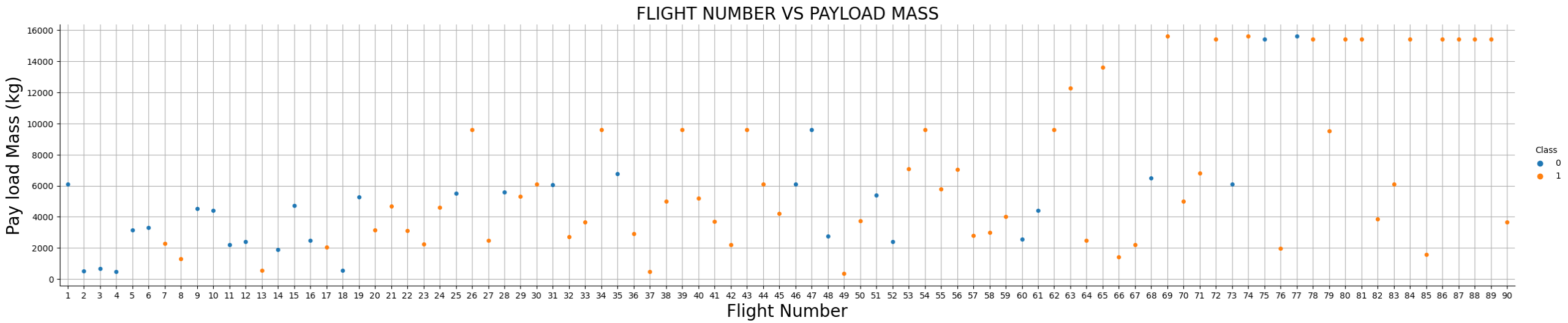
https://github.com/AzpMon/IBM-Final-course-DS-/blob/main/Spacex_data_webscraping.ipynb

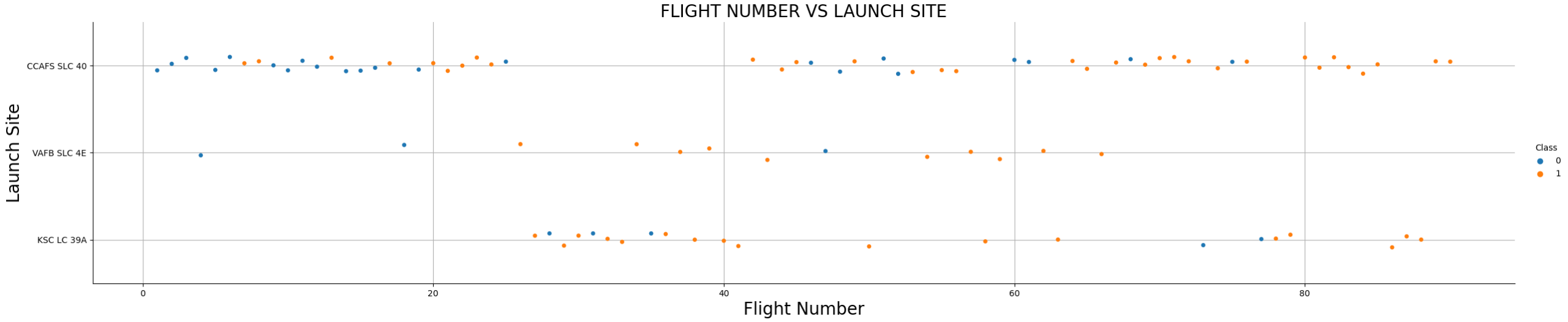
Data Wrangling



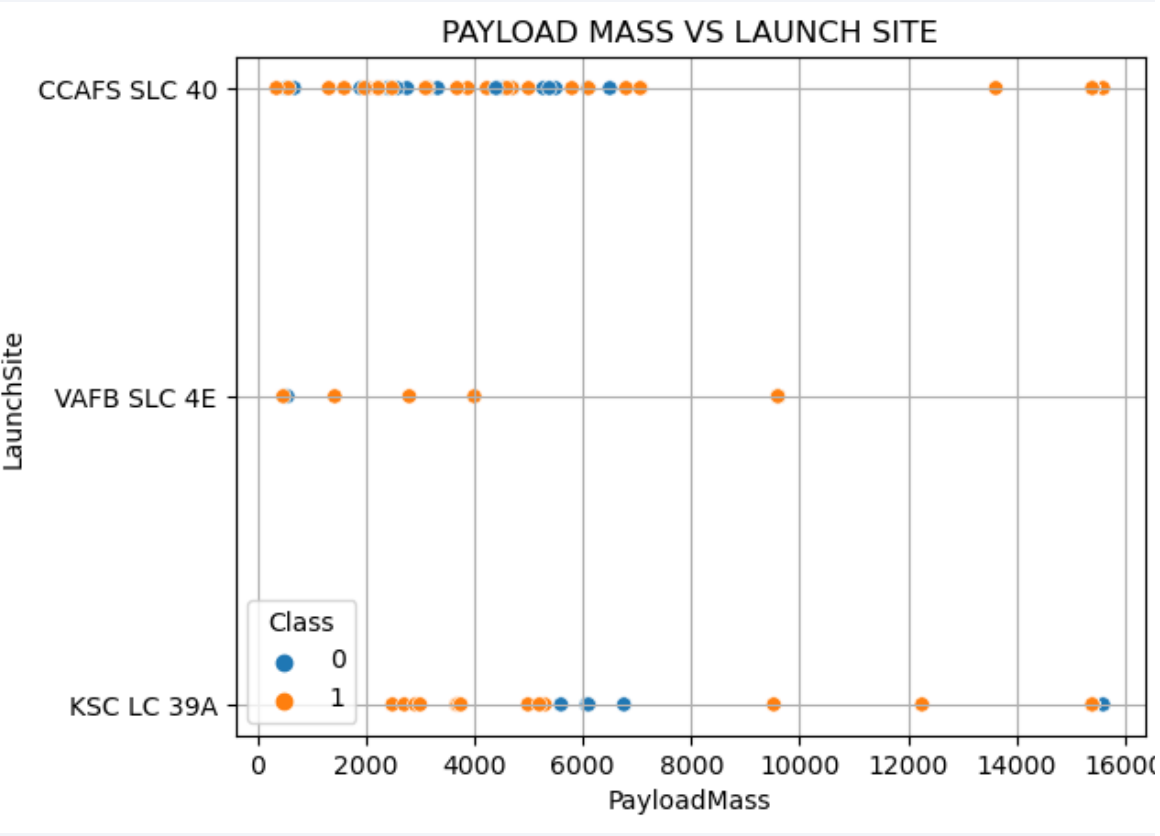
https://github.com/AzpMon/IBM-Final-course-DS-/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

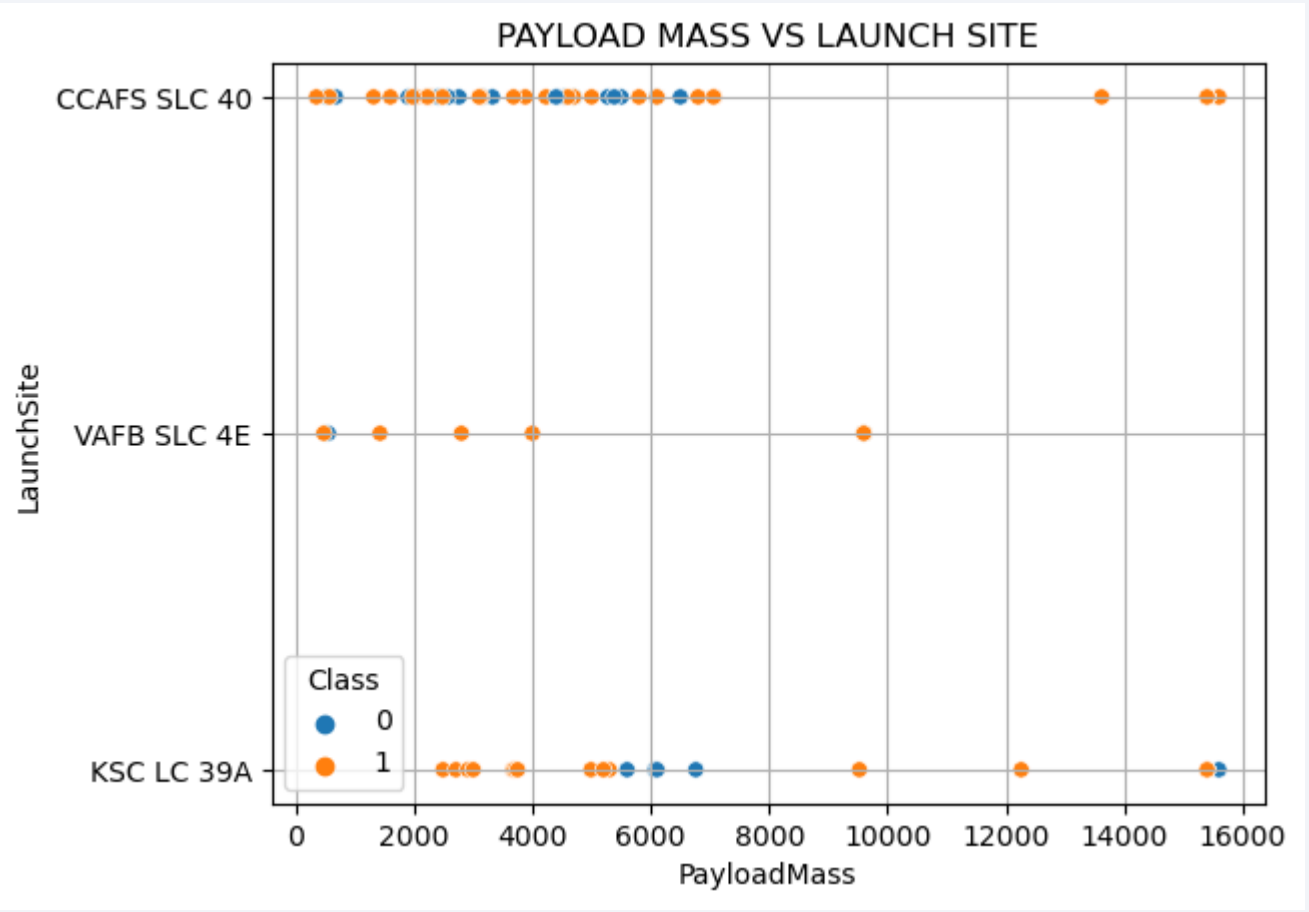
EDA with Data Visualization (1/3)



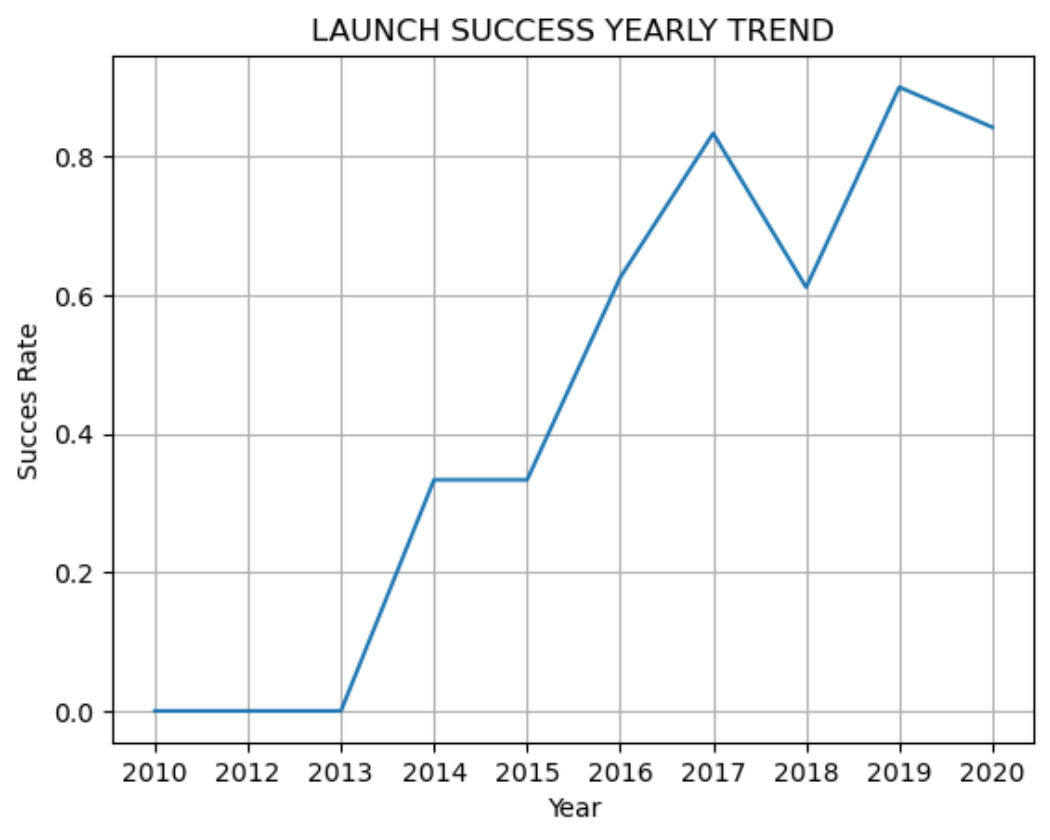


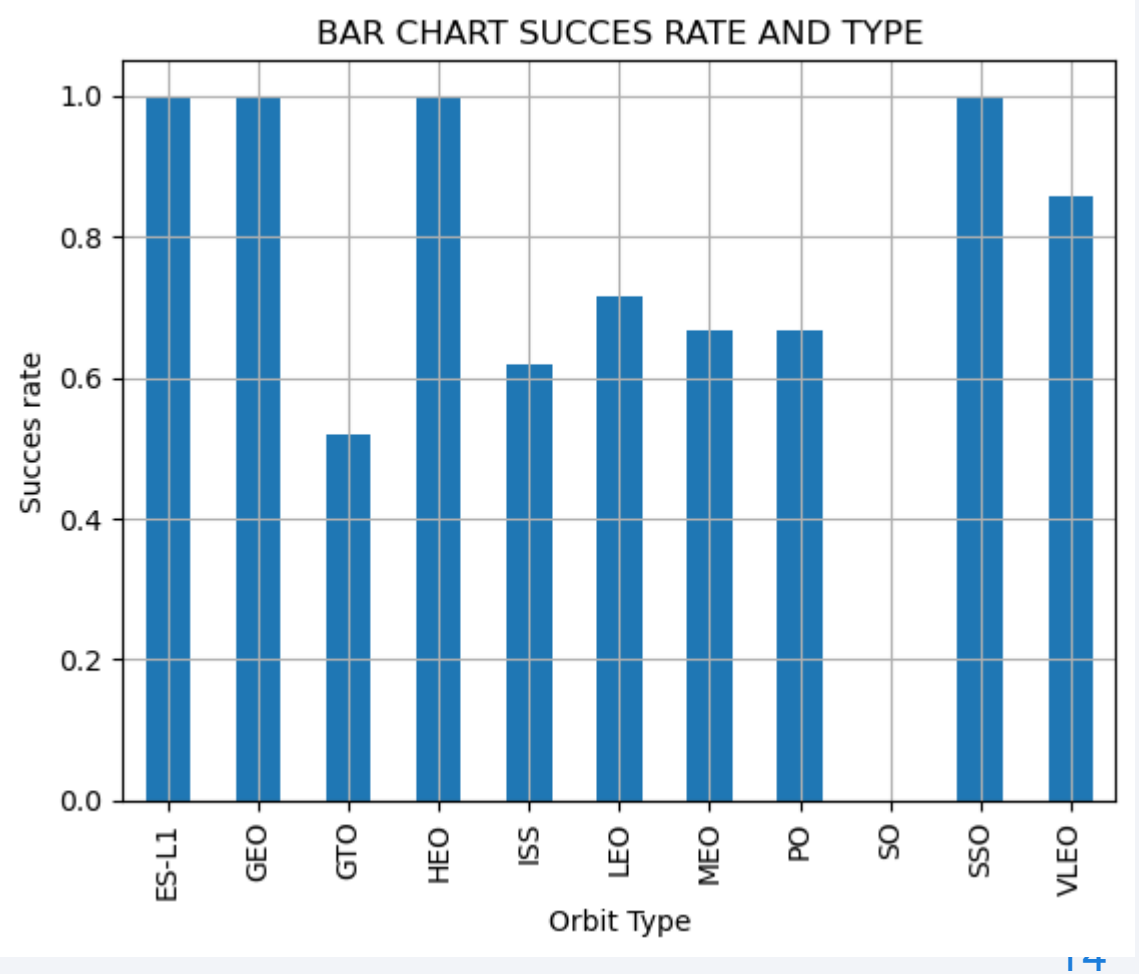
EDA with Data Visualization (2/3)





EDA with Data Visualization (3/3)





EDA with SQL

Performed SQL queries were for display:

- Names of the unique launch sites
- 5 records where launch sites begins with “CCA”
- Total payload mass carried by boosters by NASA (CRS)
- When the first successful landing outcome in ground pad was achieved.
- Date when the 1st successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
- 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.

https://github.com/AzpMon/IBM-Final-course-DS-/blob/main/sql_SpaceX.ipynb

Build an Interactive Map with Folium

Objects created and added to the folium map and its explanation:

- Markers were added for launch sites and for the NASA Johnson Space Center
- Circles were added for the launch sites.
- Lines were added to show the distance to nearby features (Distance from CCAFS LC-40 to the coastline, Distance from CCAFS LC-40 to the rail line , Distance from CCAFS LC-40 to the perimeter road)

https://github.com/AzpMon/IBM-Final-course-DS-/blob/main/lab_jupyter_launch_site_location.ipynb

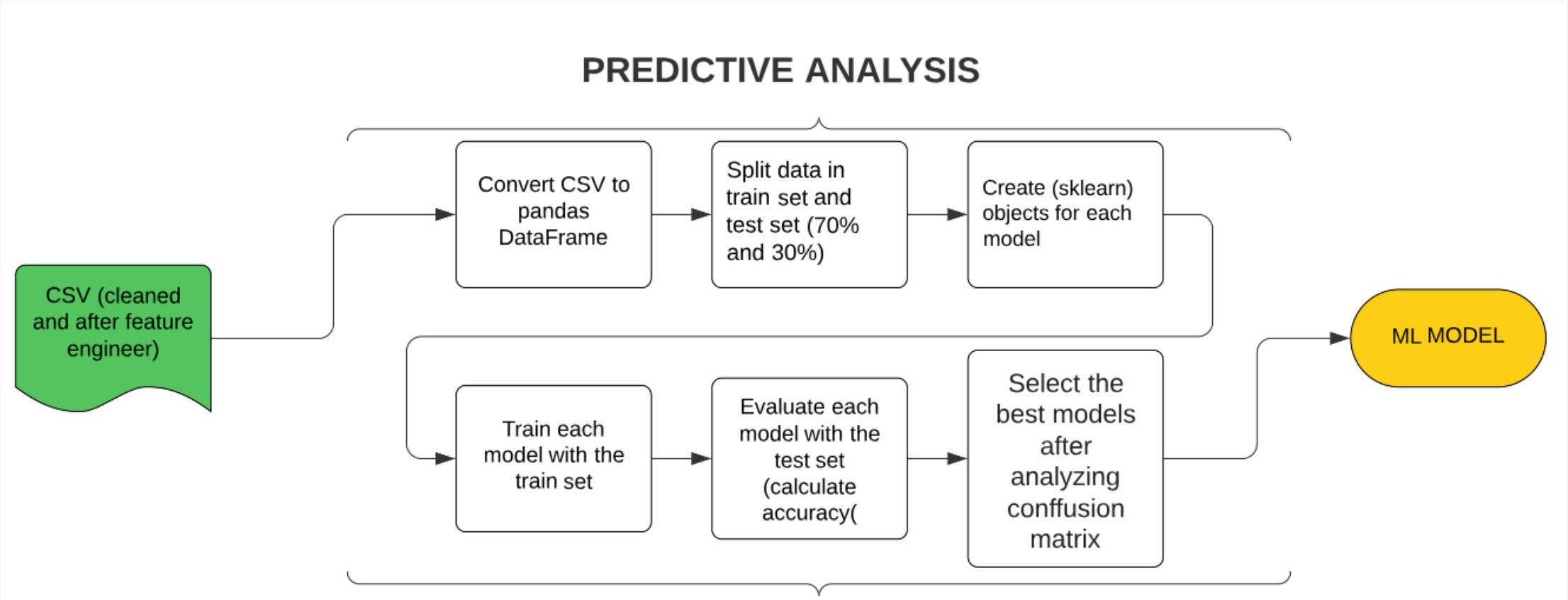
Build a Dashboard with Plotly Dash

Plots/graphs and interaction added to the dashboard for the Falcon 9 were:

- Pie char created to show the distribution of successful launches (all/each site)
- Scatter plot for the distribution of successful and failed 1st stage landings

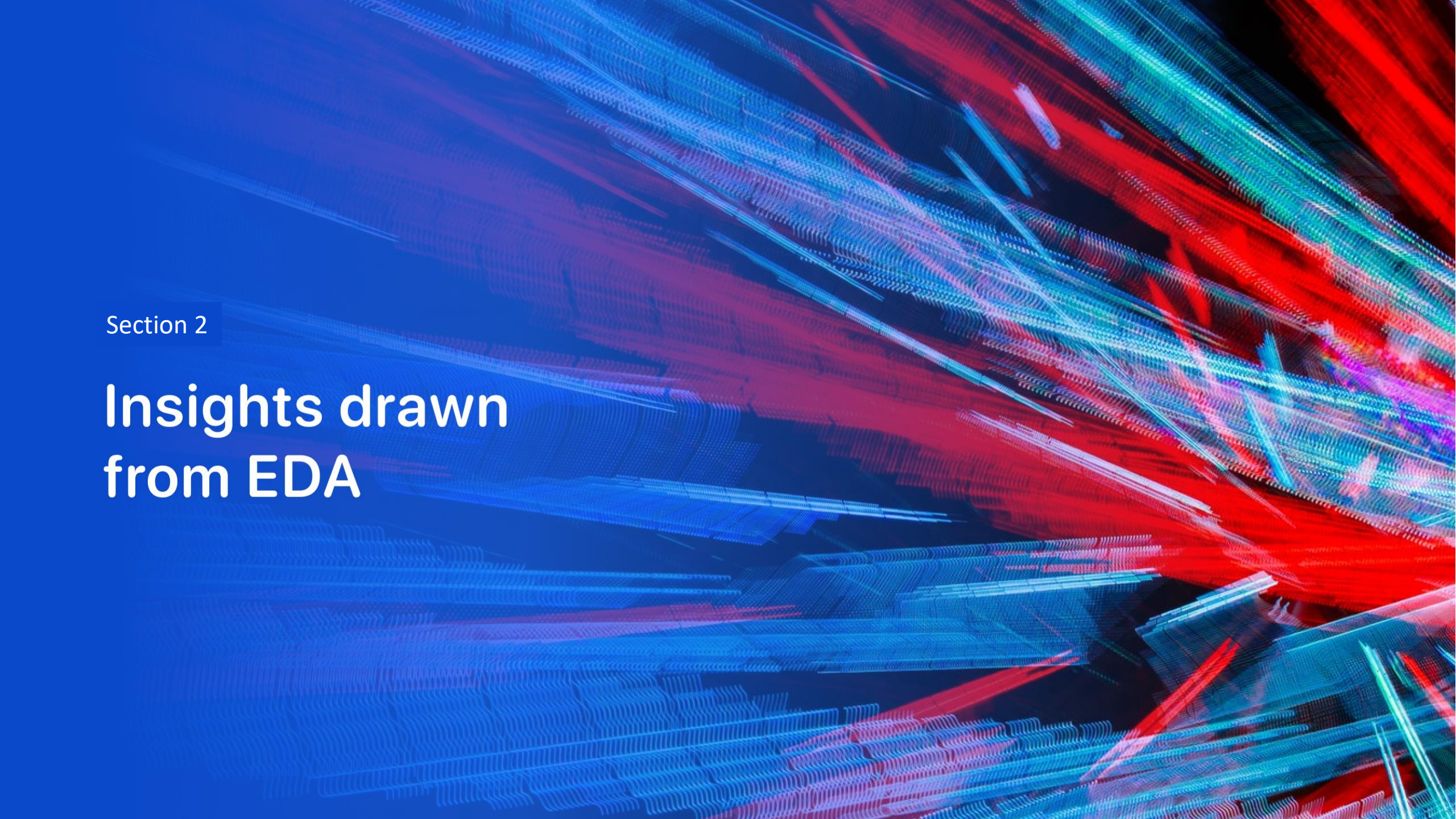
https://github.com/AzpMon/IBM-Final-course-DS-/blob/main/dash_final.py

Predictive Analysis (Classification)

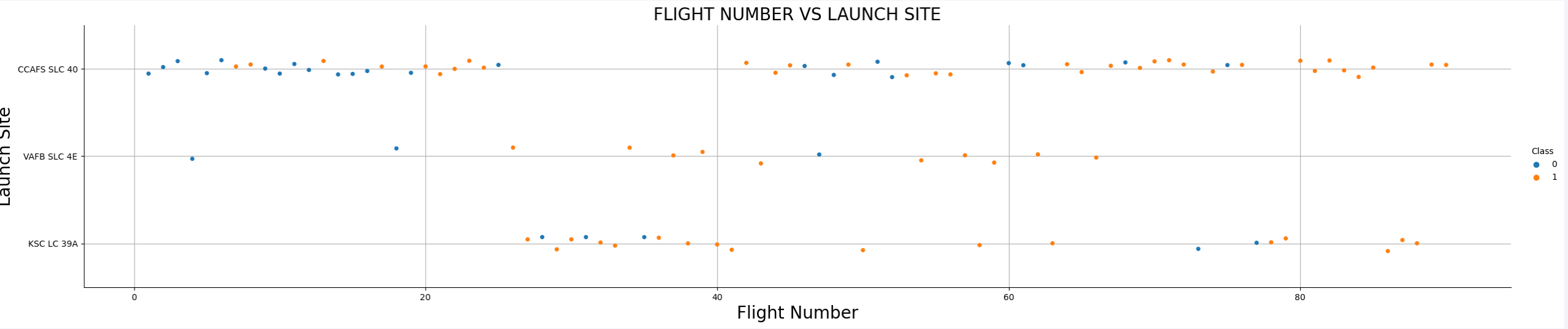


Results

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


Section 2
Insights drawn from EDA

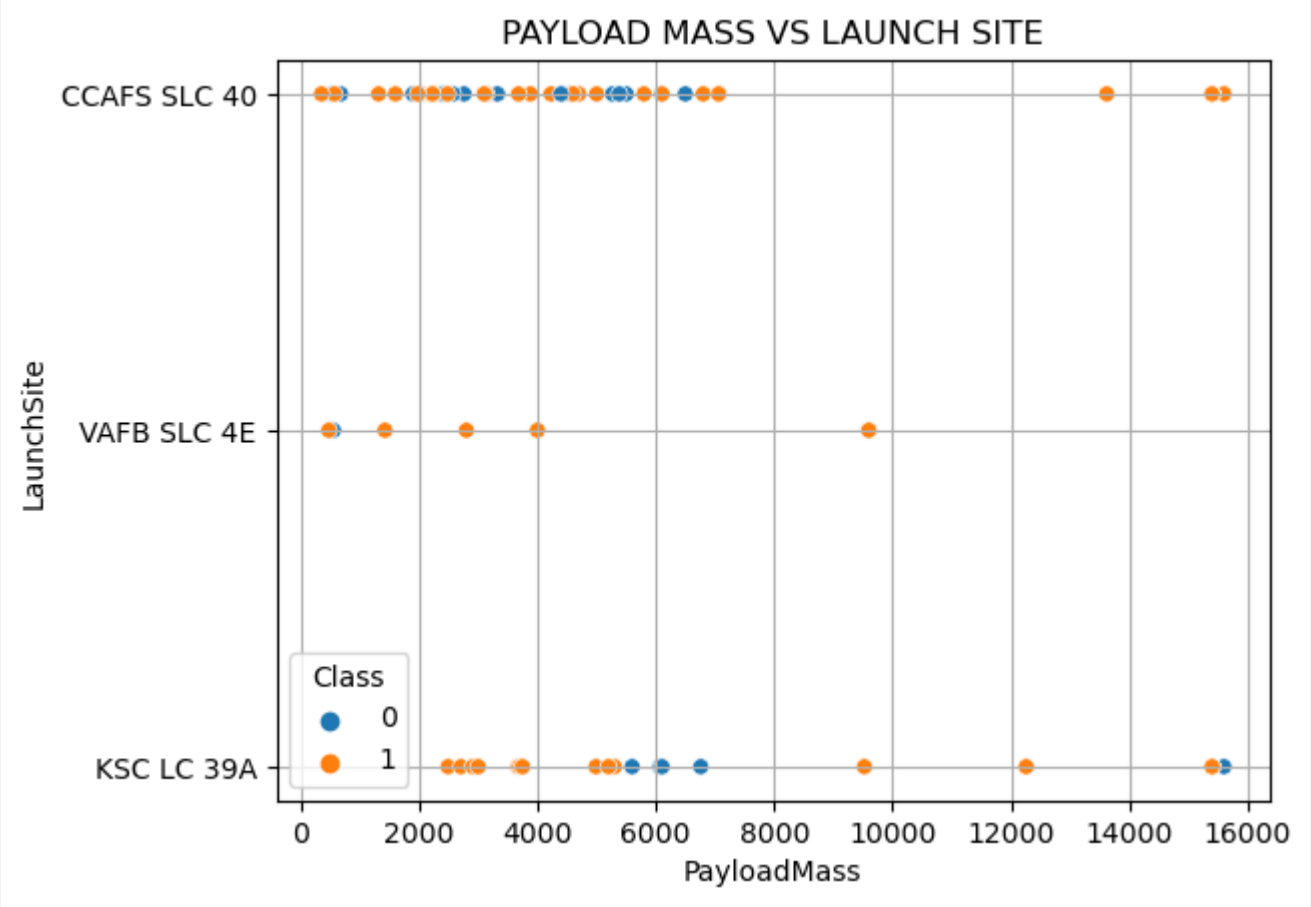
Flight Number vs. Launch Site



Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots:

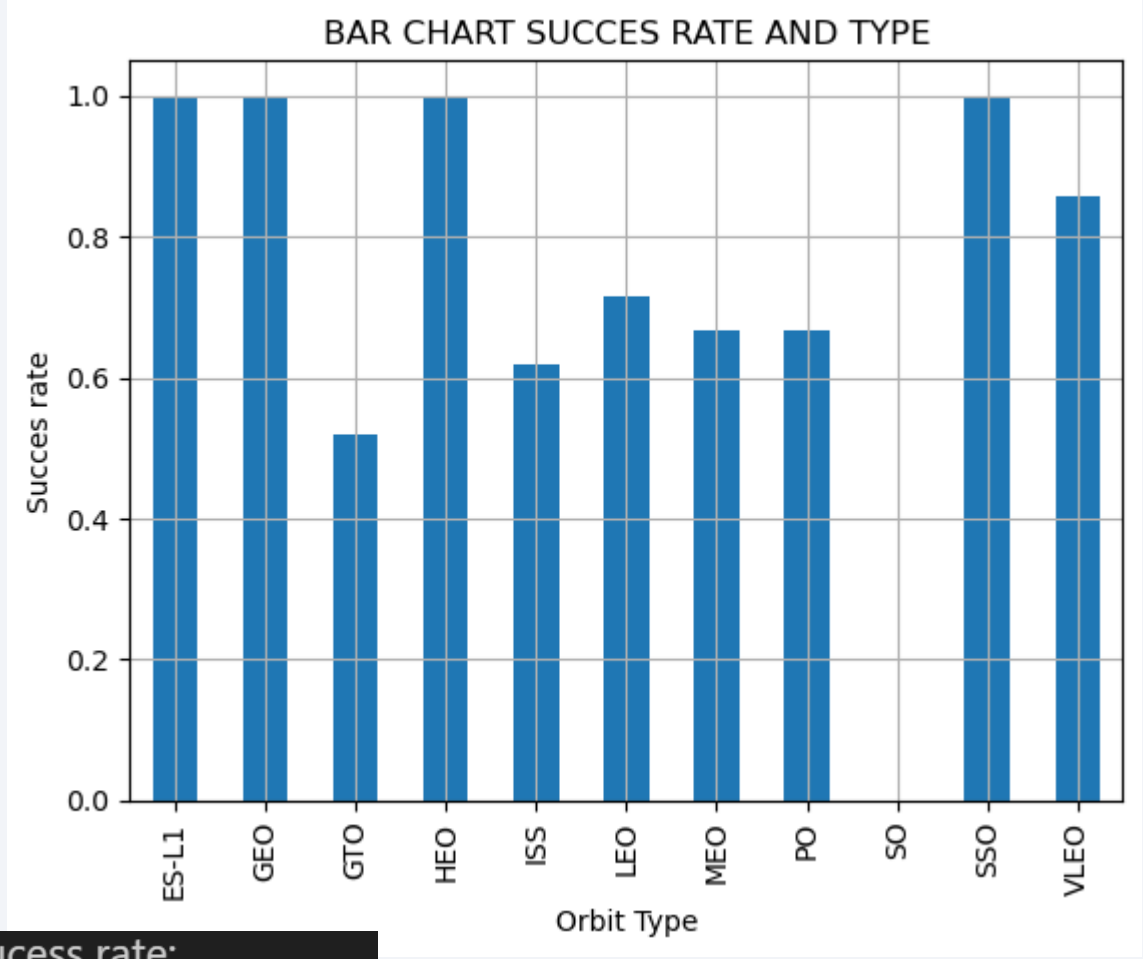
1. ***For ubication CCAFS SLC-40, as the flight number increase the launch tends to be succeed and the fists Flight numbers (over 70) tends to failed***
2. ***For VAFB SLC 4E, ocurrs something similar: flight number > 50 corresponds to a succesed launch***
3. ***For KSC LC 39A, flight number, there's not enough information to conclude that as the flight number increas the launch succeed.***

Payload vs. Launch Site



For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

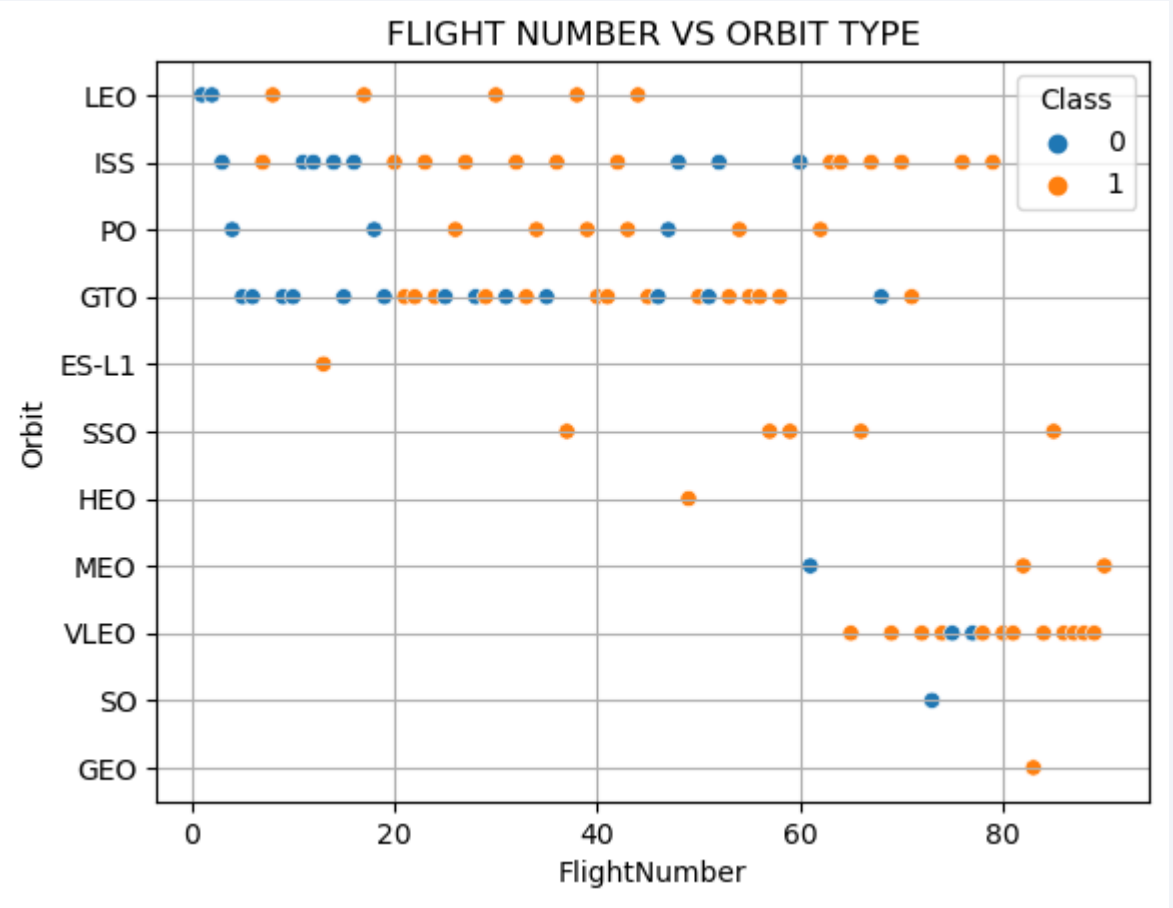


Analyze the ploted bar chart try to find which orbits have high sucess rate:

1. ***The orbits type ES-L1, GEO, HEO and SSO have a 100% succes rate***
2. ***For the orbit GTO,the worst with 50% succes rate, is the same as toss a coin***

Flight Number vs. Orbit Type

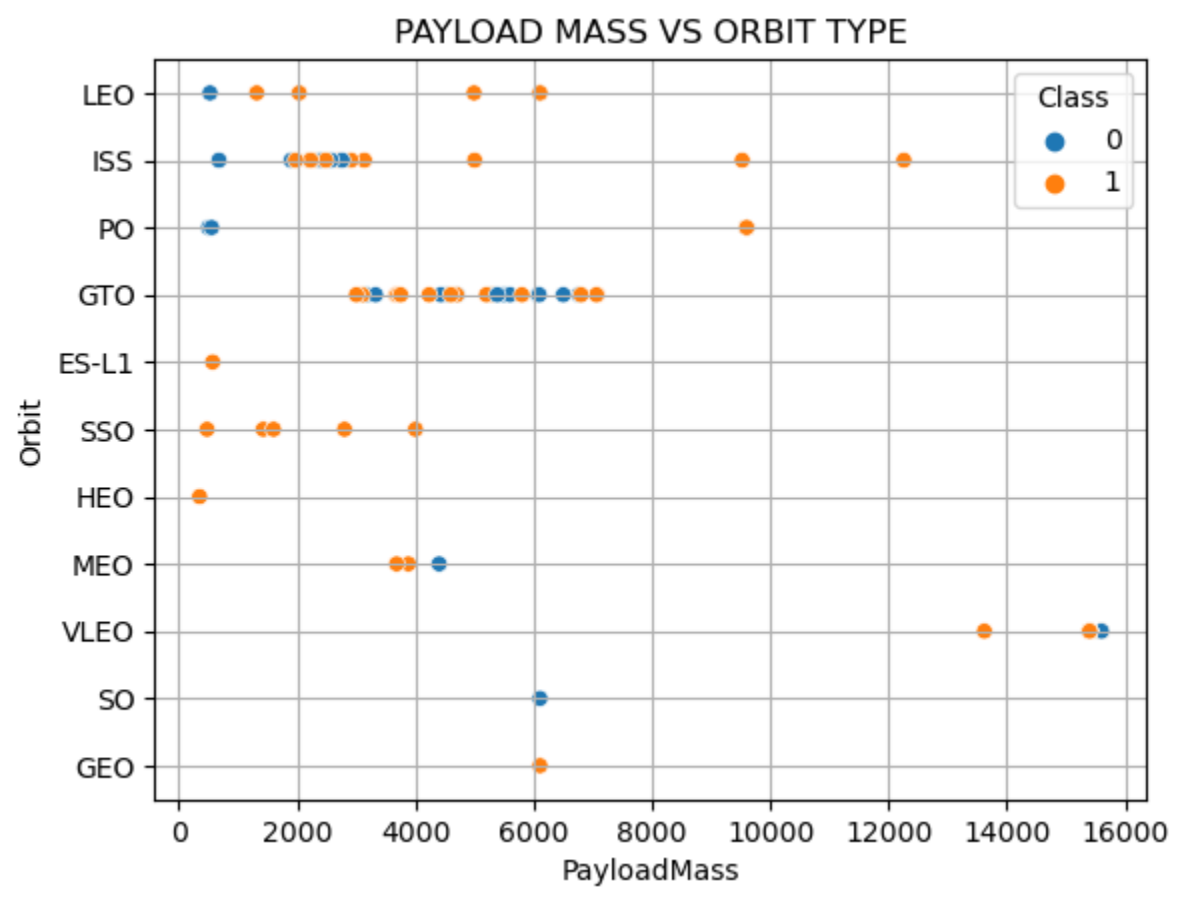
The LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

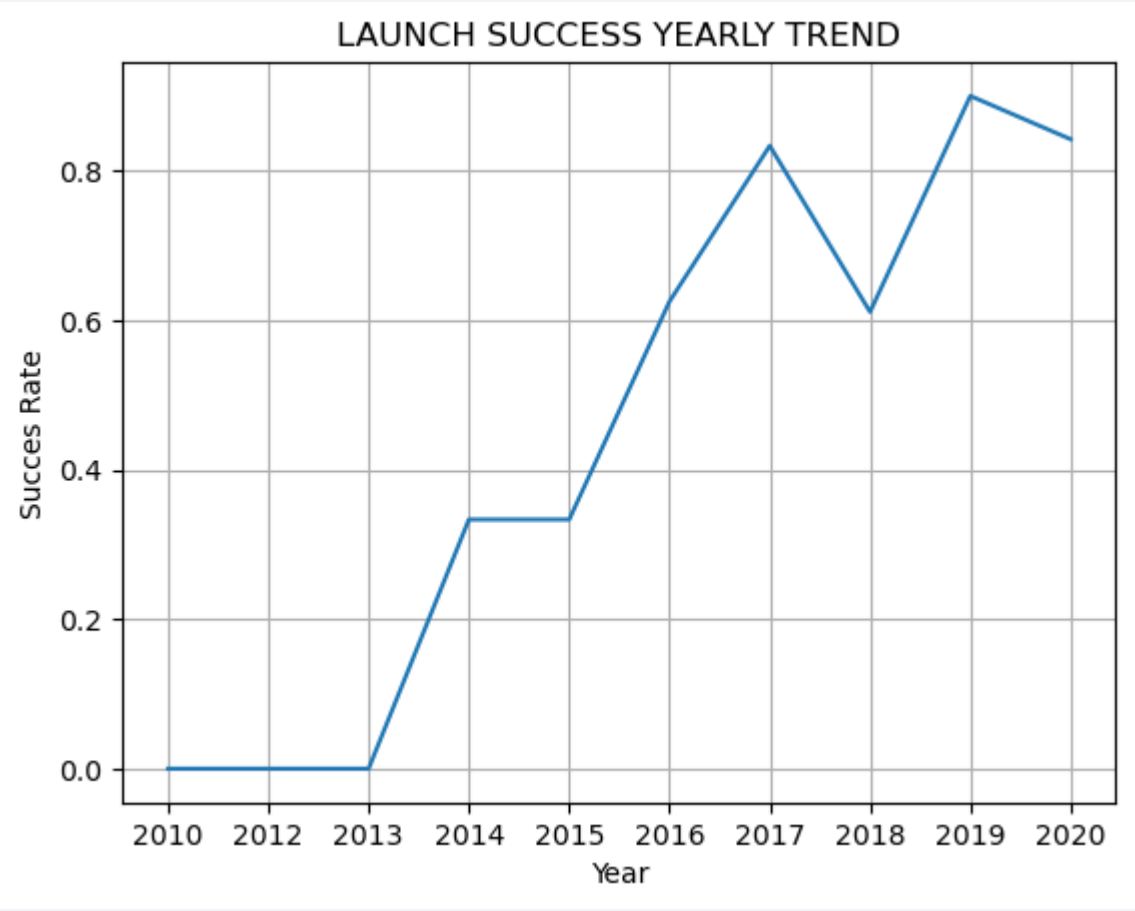
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here



Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

```
%sql SELECT DISTINCT "Launch_Site" FROM spacextable
```

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

Using the magic command %sql, from the spacextable we select the distinct Launch Sites

Launch Site Names Begin with 'CCA'

```
%sql SELECT* FROM spacextable WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

Using the magic cell %sql, and select all data where Launch Site starts with ‘CCA’ ; select only 5 rows

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) from spacextable
```

```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
619967

Using the magic command %sql, using the SUM method, compute the total payload mass from the spacextable

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM spacextable
```

```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
6138.287128712871

Using the magic command %sql, from spacextable, select the average payload mass

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM spacextable WHERE "Landing_Outcome" = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

MIN(Date)
2015-12-22

Using the %sql magin command, from the spacextable, select the first date where Landing Outcome is ‘Succes (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
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Using the magic command %sql%, from the spacextable we select the successful Dron Ship Landing with Payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome,  COUNT(Mission_Outcome) FROM spacextable GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

Using the magic command %sql, count the total number of successful and failure mission outcome from spacextable

Boosters Carried Maximum Payload

```
%sql SELECT booster_Version FROM spacextable ORDER BY PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) from spacextable)
```

Using the %sql magin command, select all the booster version carried maximum payload from spacextable and order by payload mass

2015 Launch Records

```
%sql SELECT substr(Date, 6, 2) as month, date, Landing_Outcome,
Booster_Version, Launch_Site from spacextable WHERE Landing_Outcome = 'Failure (drone ship)' and substr(Date, 0, 5) = '2015'
```

```
 * sqlite:///my_data1.db
Done.
```

month	Date	Landing_Outcome	Booster_Version	Launch_Site
01	2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Using the magic command %sql, from spacextable, select the months in 2015 where Landing Outcome is Failure for drone ship

Section 3

Launch Sites Proximities Analysis

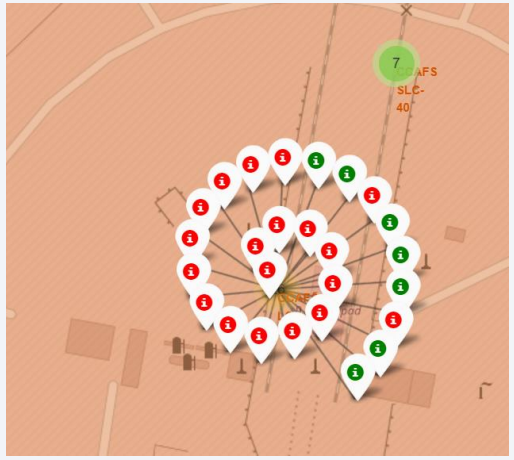
Falcon 9 Launch Site Locations (Falcon 9)

A folium marker and a folium Circle is associated to each Launch Location for Falcon 9 rocket.

Success/Failed Landings

VAFB SLC-4E

KSC LC-39A

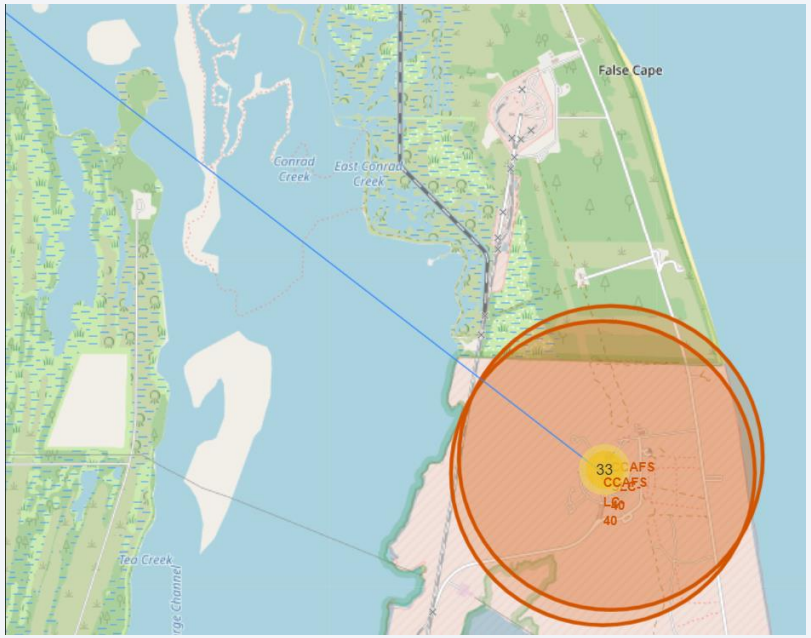


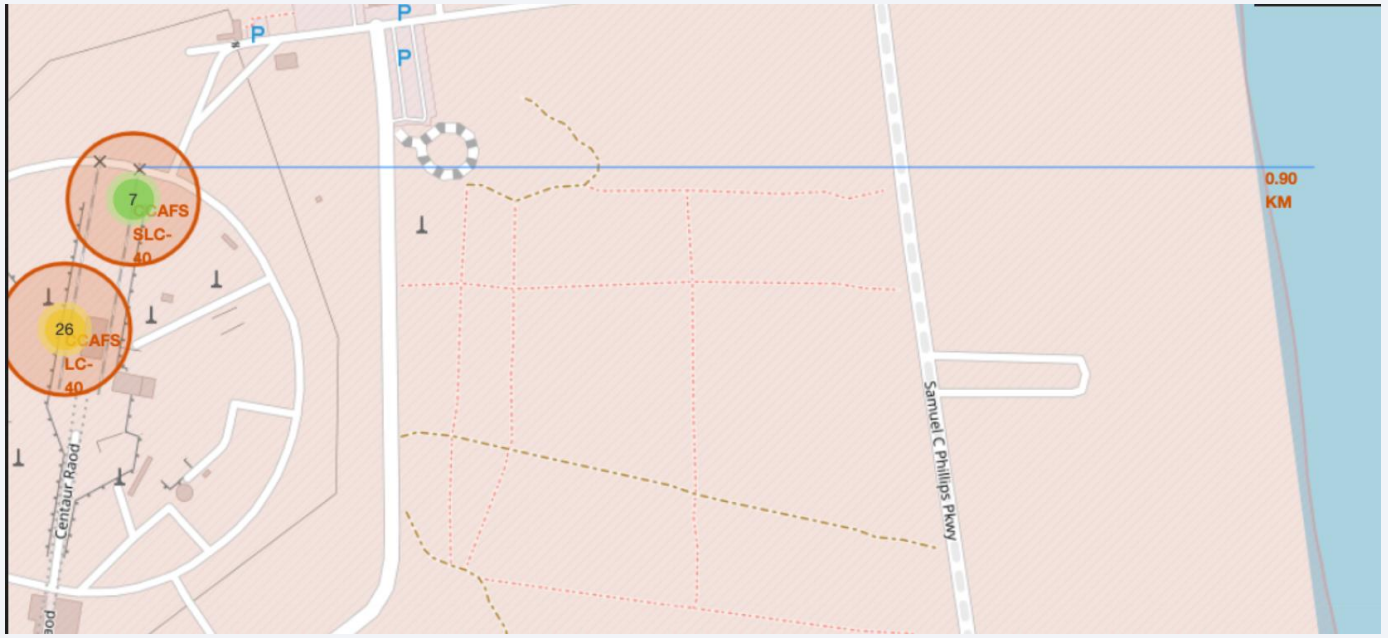
CCAFS LC-40

CCAFS SLC-40

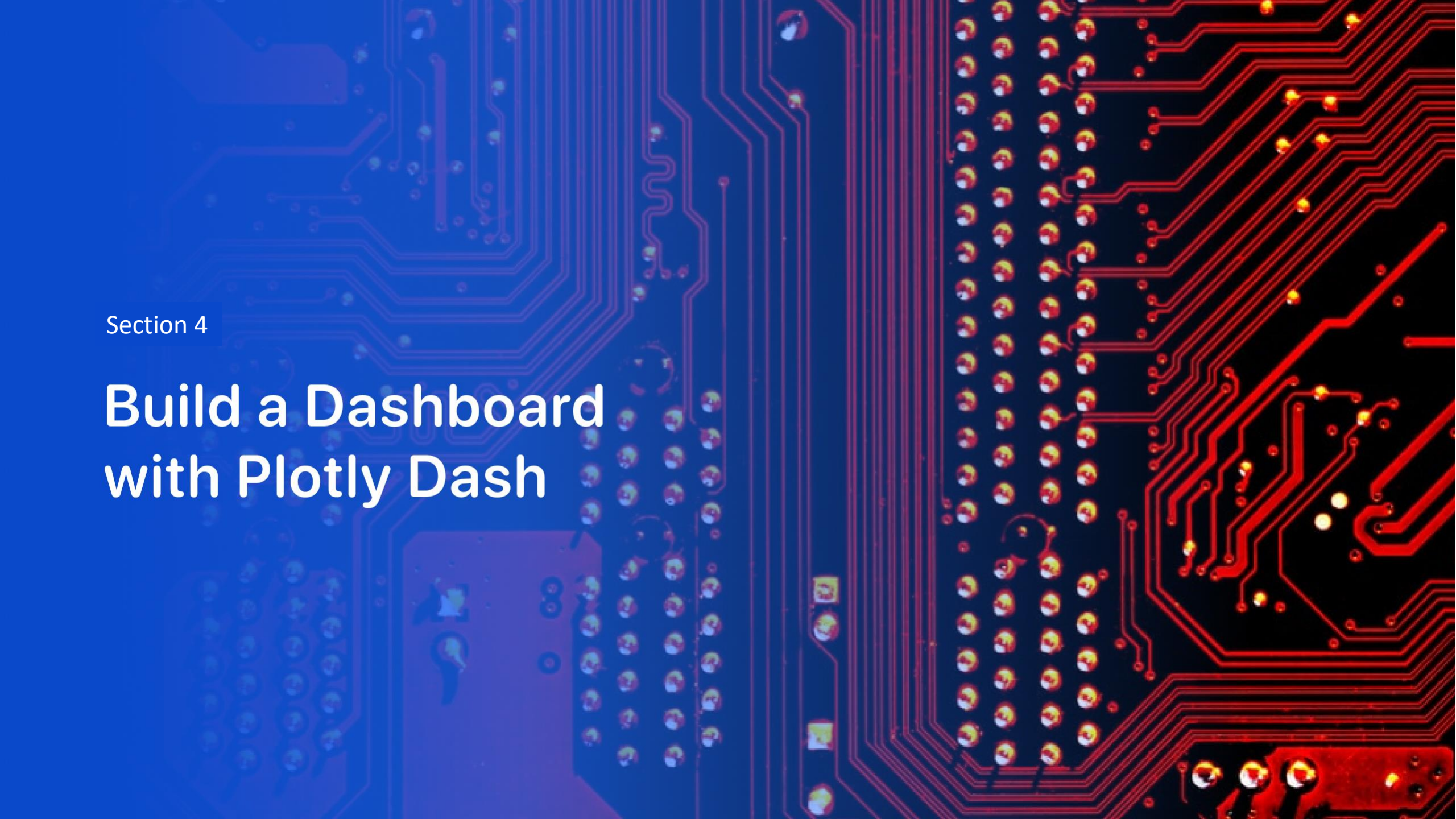
A folium marker and a folium cluster is associated to each location with the consideration of Launches outcome (success or fail)

Distance from Launch site to proximities





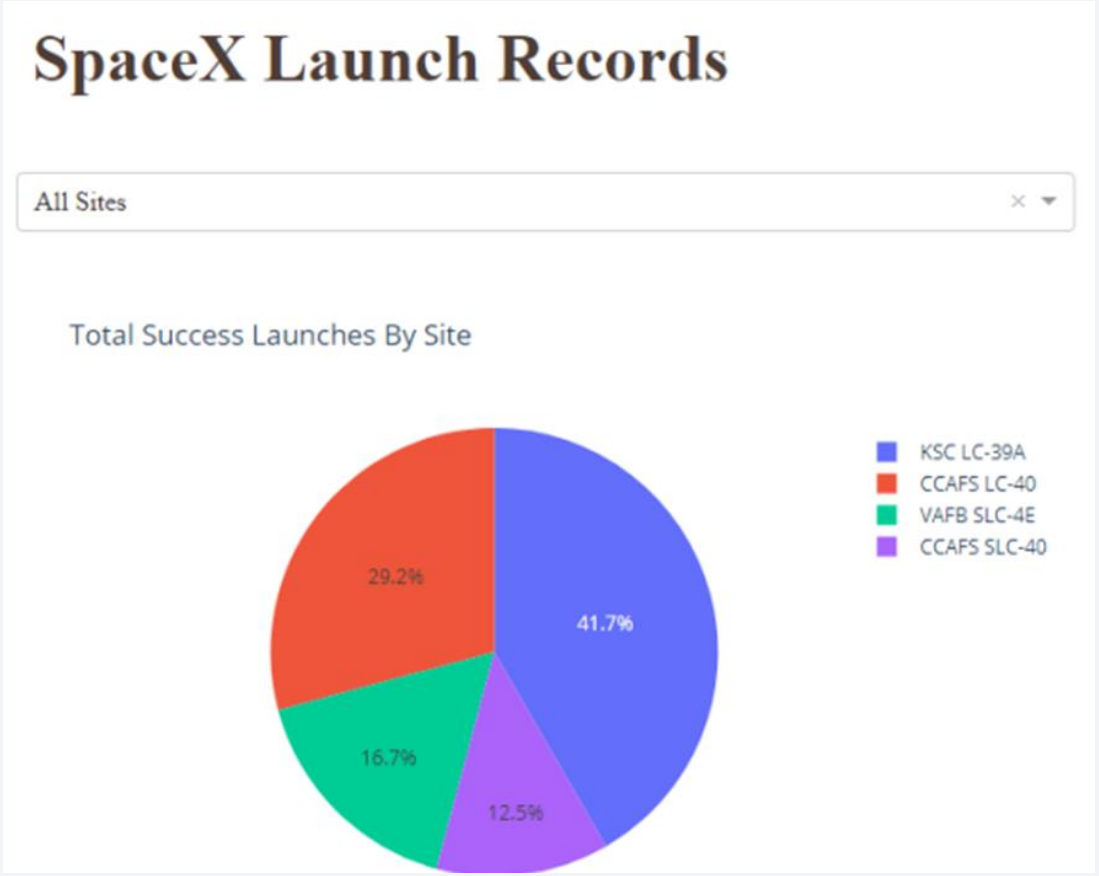
Distances from launch ubications to its proximities


Section 4
Build a Dashboard with Plotly Dash

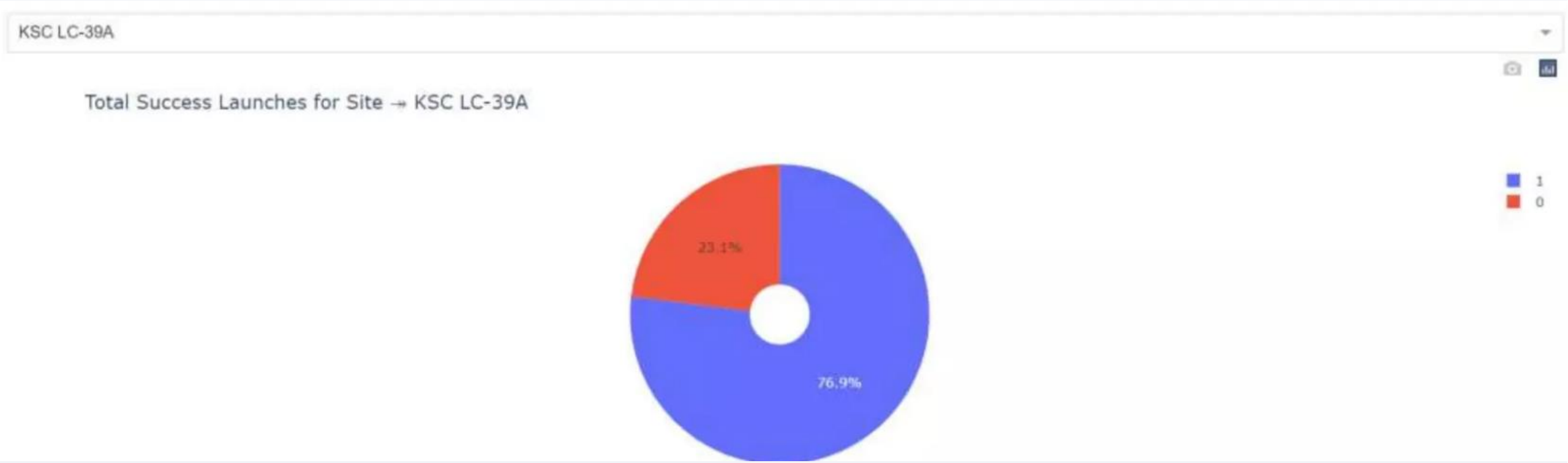
Launch success (all Sites)

The pie char looks at the percentage of total success launches of Falcon 9 first stage:

- KSC LC-39A has the biggest percentage of succeed (47.7%)
- VAFP SLC-4E has the smallest percentage of succeed (12.5%)



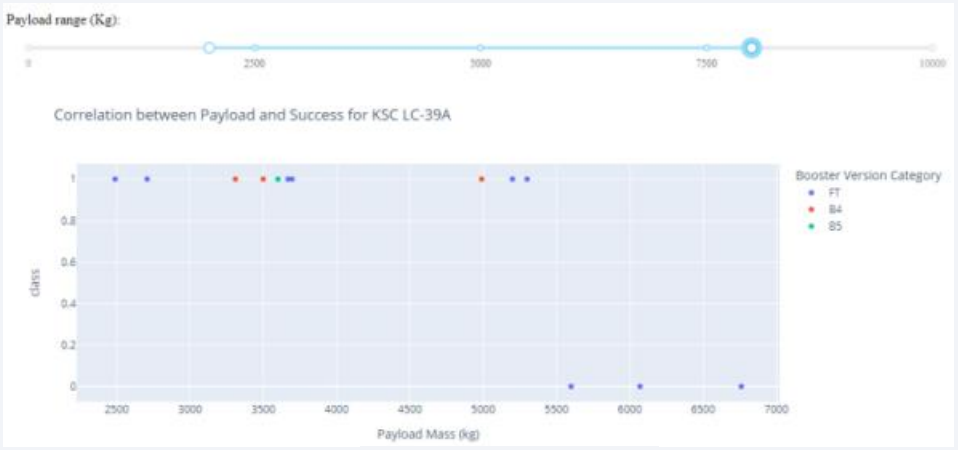
Launch site with highest launch success ratio



KSC LC-39A failed landings (blue label) and success landing (red label) for the Falcon 9 1st stage.

Launch Outcome in function of Payload mass (all sites)

CCAFS LC-40



KSC LC-39A

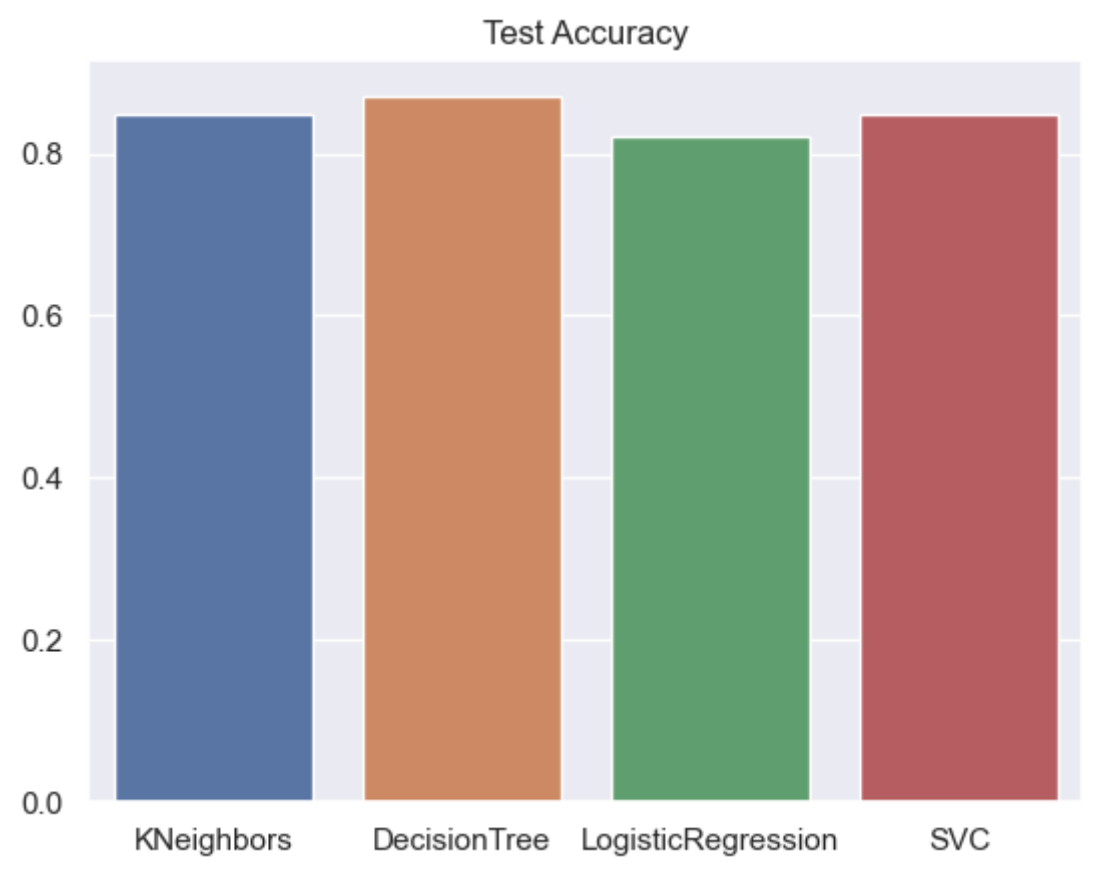
CCAFS SLC-40

VAFB SLC-4E

Section 5

Predictive Analysis (Classification)

Classification Accuracy



Test accuracy for the fitted models: In that sense the best model is the DecisionTree

Confusion Matrix (Decision Tree)

The confusion Matrix looks at:

- True Positives = 11

  11 of the predicted success launches were really success

- True Negatives = 5

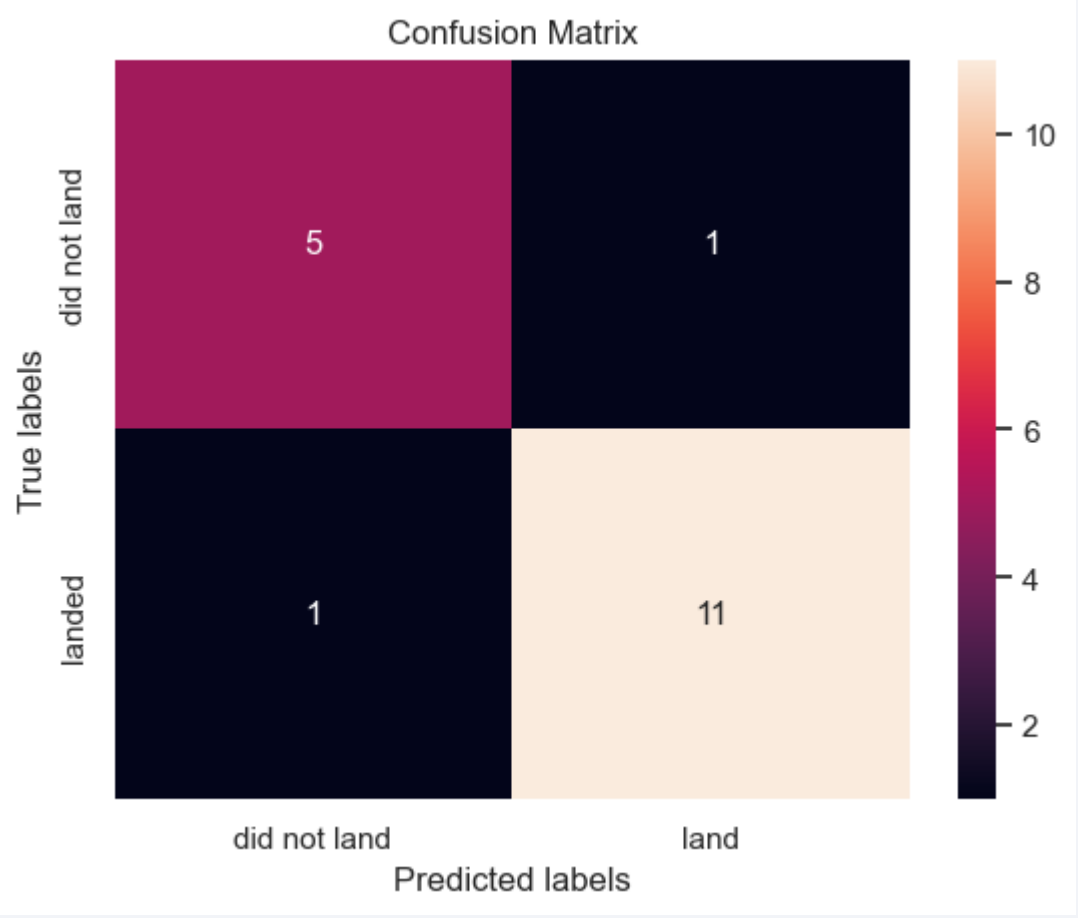
  5 of the predicted failed launches were really failed”

- False Positives = 1

  1 of the predicted succussed launches weren’t really success”

- False Positives = 1

  1 of the predicted succussed launches weren’t really success”



Conclusions

- The decision tree is the best model in terms of accuracy
- KSC LC-39A (location) has the biggest percentage of succeed launches
- As the number of flight increase the success increment to.
- The best orbits for success are ES-L1, GEO, HEO and SSO
- For the Leo Orbit, as the number of flight increase the increase too, and for heavy payloads the success landing are more for Polar, LEO and ISS.

Appendix

All the code, specific charts and data are available in

https://github.com/AzpMon/IBM-Final-course-DS-

Thank you!