IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Author: Brian Njoroge
Date : 15/03/2023

Outline

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

Executive Summary

- The goal of this project is to determine the cost of each launch.
- We do this by determining whether SpaceX will reuse its first stage by successfully landing it after launch.
- Data used in this exercise is gathered from the SpaceX REST API and from scraping html tables.
- We clean and organize the collected data into a usable format after which we perform some exploratory data analysis, build charts, graphs and an interactive dashboard to visualize the findings.
- Next we proceed to building different classification machine learning models to predict whether the first stage will land, analyze the results of the models and determine which one performs best.

Introduction

- We are a data scientist working for SpaceY, a space exploration company that aims to compete with SpaceX.
- We want to determine the price of each rocket launch.
- To do this, we need to establish whether the first stage is reusable by determining whether it will land successfully or not.
- All this is done using historical SpaceX data.

Section 1

Methodology

Methodology

Executive Summary

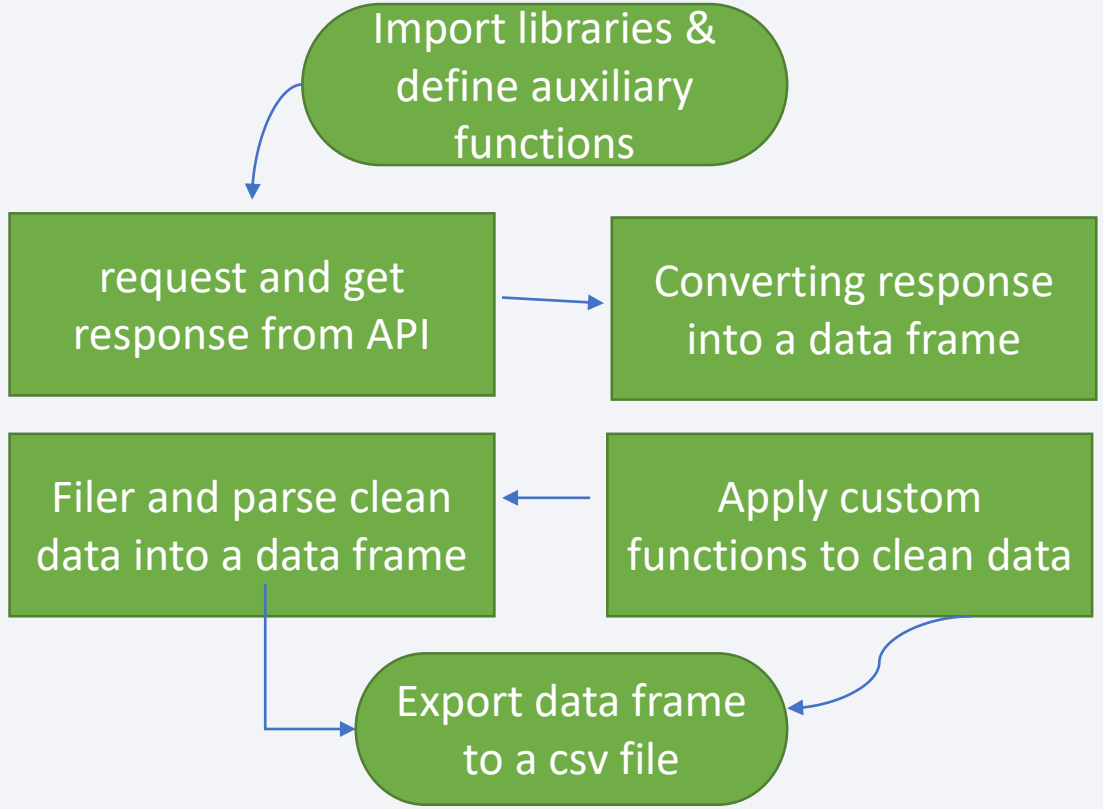
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- We made a get request to the SpaceX API in order the collect the first data set.
- We also scraped Wikipedia to get HTML tables containing Falcon 9 and Falcon Heavy Launches records.
- The collected tables were then converted into a Pandas data frame.

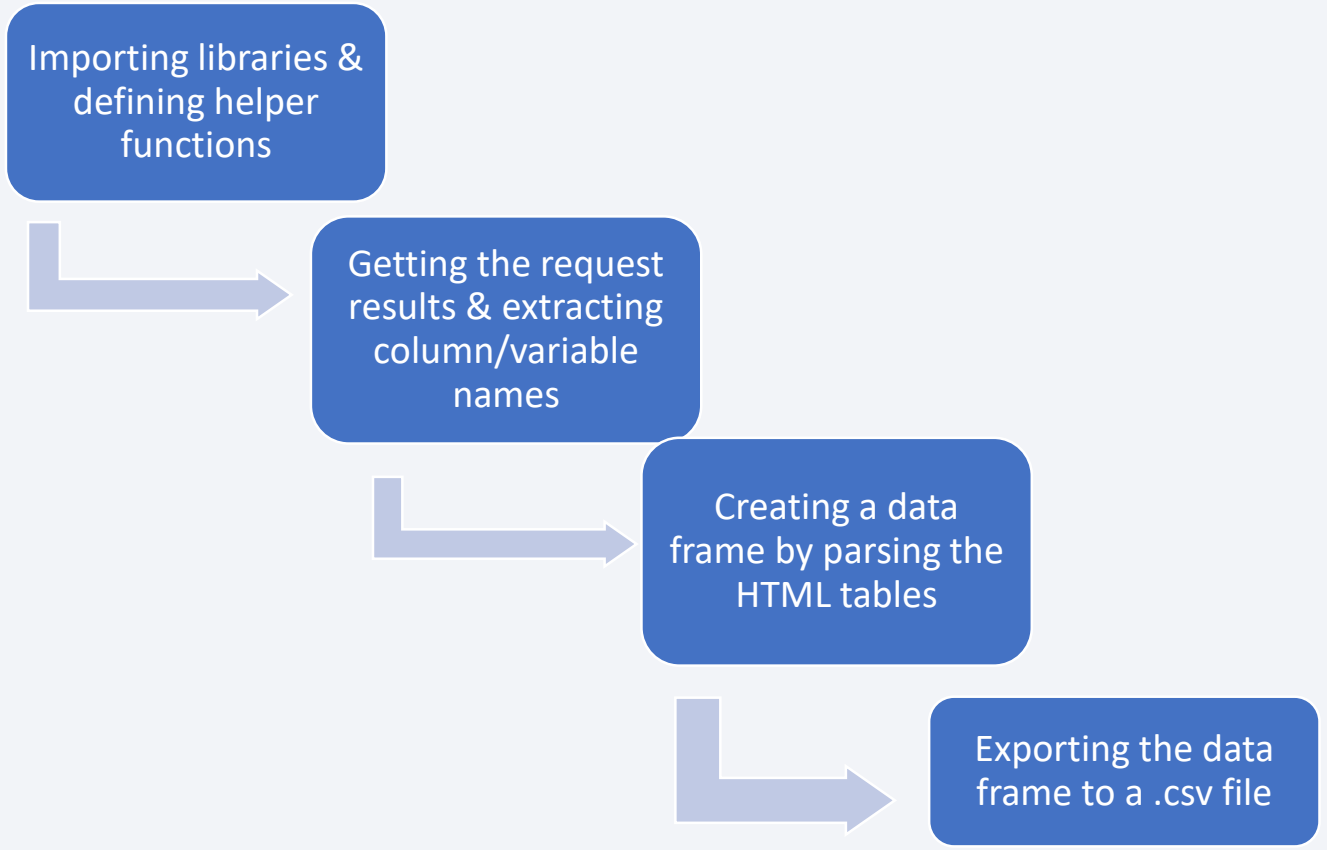
Data Collection – SpaceX API

- https://github.com/Sabacon/IBM-Data-Science-Professional-Cert/blob/main/Capstone%20Project/1.%20Data%20Collection/jupyter-labs-spacex-data-collection-api.ipynb



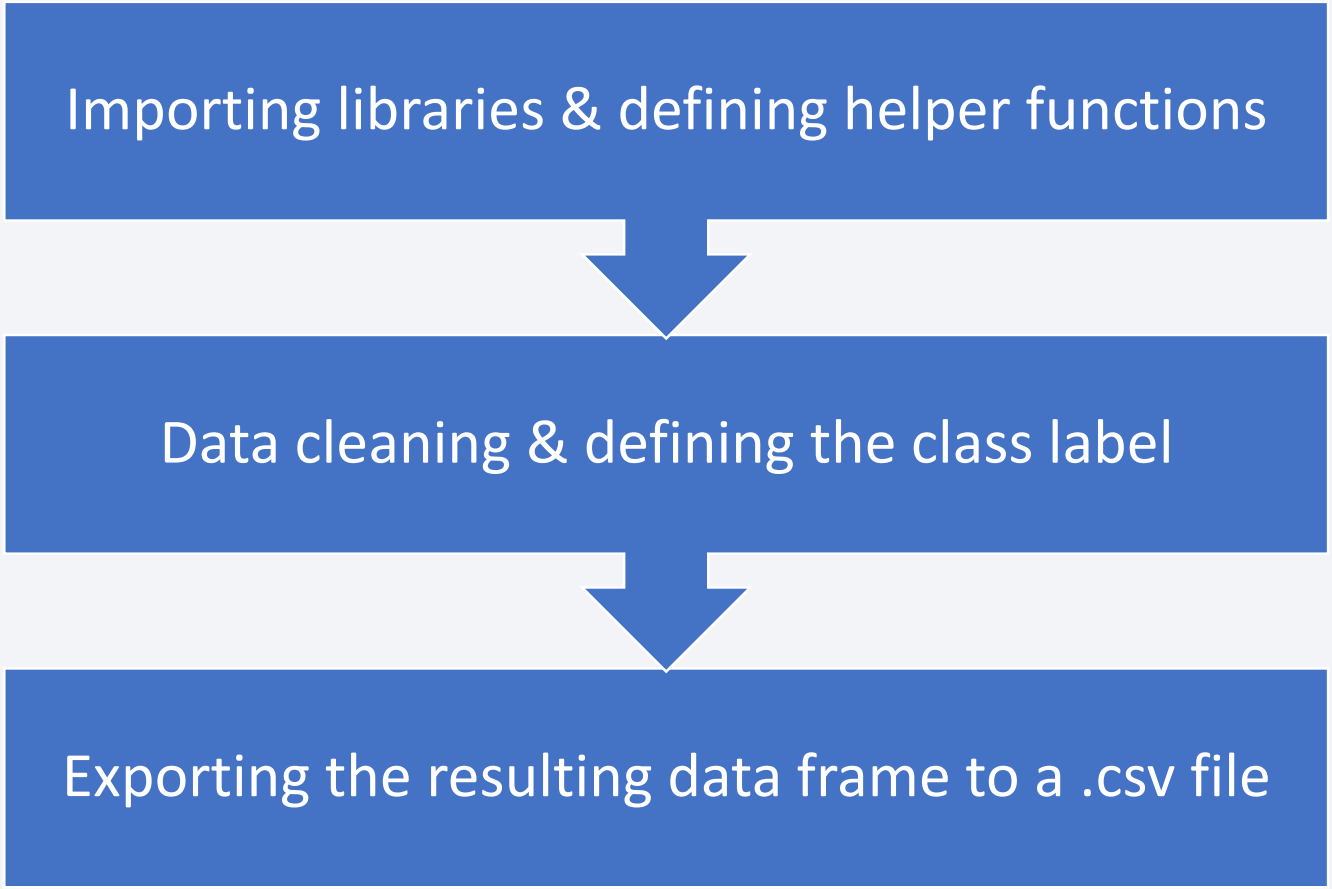
Data Collection - Scraping

- https://github.com/Sabacon/IBM-Data-Science-Professional-Cert/blob/main/Capstone%20Project/1.%20Data%20Collection/jupyter-labs-webscraping.ipynb



Data Wrangling

- We read in the data & calculated the percentage of missing values for each feature.
- We calculated the number of launches on each site and the number and occurrence of each orbit.
- We calculated the occurrence of each mission outcome.
- We created a class column(labels) from landing outcomes with 0 & 1 representing bad & good outcomes respectively.
- We exported the resulting data frame to a csv file.
- https://github.com/Sabacon/IBM-Data-Science-Professional-Cert/blob/main/Capstone%20Project/2.%20Data%20Wrangling/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- We used cat plots and scatterplots colored by the class label. These kinds of plots are great at visualizing the relationship between two features.
- We created a bar plot to visualize the success rate per orbit type.
- We rounded up our EDA with visualization by using a line chart to how the success rate changed in the decade between 2010 and 2020.
- https://github.com/Sabacon/IBM-Data-Science-Professional-Cert/blob/main/Capstone%20Project/3.%20EDA%20with%20SQL%20and%20Python/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- We began with listing the unique launch sites by using the SELECT DISTINCT statement on the SPACEXTBL table.
- We then showed the first 5 launch sites that began with the string ‘CCA’.
- We used the SELECT and WHERE statement with the SUM aggregate function to return the total payload mass carried by boosters launched by NASA(CRS).
- We used the SELECT and WHERE statement with the AVG aggregate function to return the average payload mass carried by booster version F9 v1.1.
- We used the min function to return the date when the first successful landing outcome in ground pad was achieved.
- We used the GROUP BY statement alongside the SUBSTR() & COUNT() functions to return the total number of successful and failure outcomes
- We used a subquery to list the names of the booster versions which have carried the maximum payload mass.
- To conclude, we used the COUNT and BETWEEN functions to rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- https://github.com/Sabacon/IBM-Data-Science-Professional-Cert/blob/main/Capstone%20Project/3.%20EDA%20with%20SQL%20and%20Python/jupyter-labs-eda-sql-coursera_sqllite.ipynb

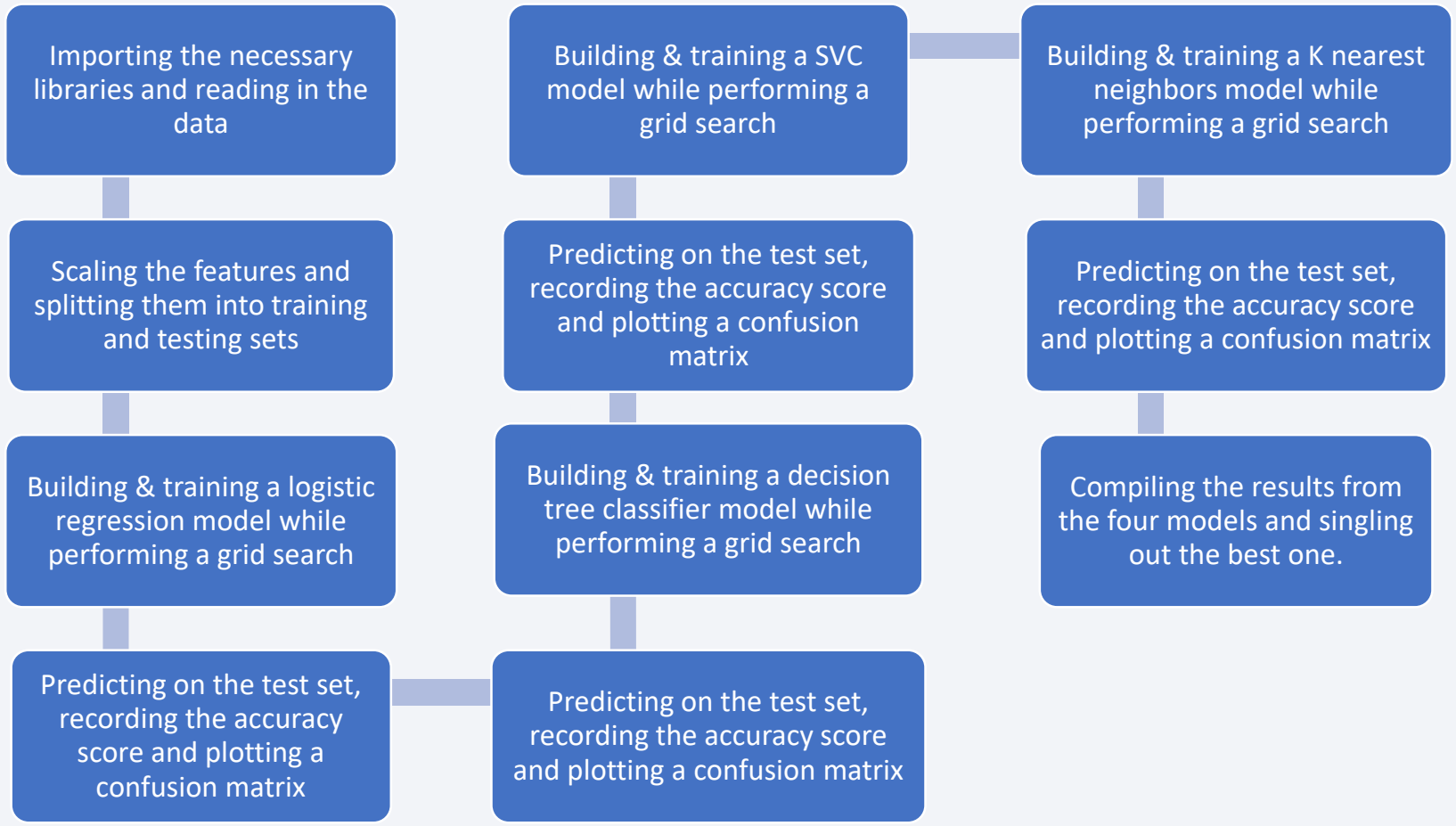
Build an Interactive Map with Folium

- We created and added folium.Circle and folium.Marker objects to mark each launch site on the site map.
- Then we marked the success/failed launches for each site.
- We added a MousePosition and used it alongside a marker to show the distance between the coastline point and launch site.
- We drew a PolyLine between the launch site and the selected coastline point.
- Finally we drew a line between a launch site to its closest city.
- https://github.com/Sabacon/IBM-Data-Science-Professional-Cert/blob/main/Capstone%20Project/4.%20Interactive%20Visual%20Analytics%20and%20Dashboard/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- We began by adding a dropdown list to enable launch site selection.
- Then we added a pie chart to show the total successful launches for all sites.
- We added a slider to select a payload range.
- Finally we added a scatterplot to show the correlation between payload and launch success.
- https://github.com/Sabacon/IBM-Data-Science-Professional-Cert/blob/main/Capstone%20Project/4.%20Interactive%20Visual%20Analytics%20and%20Dashboard/spacex_dash_app.py

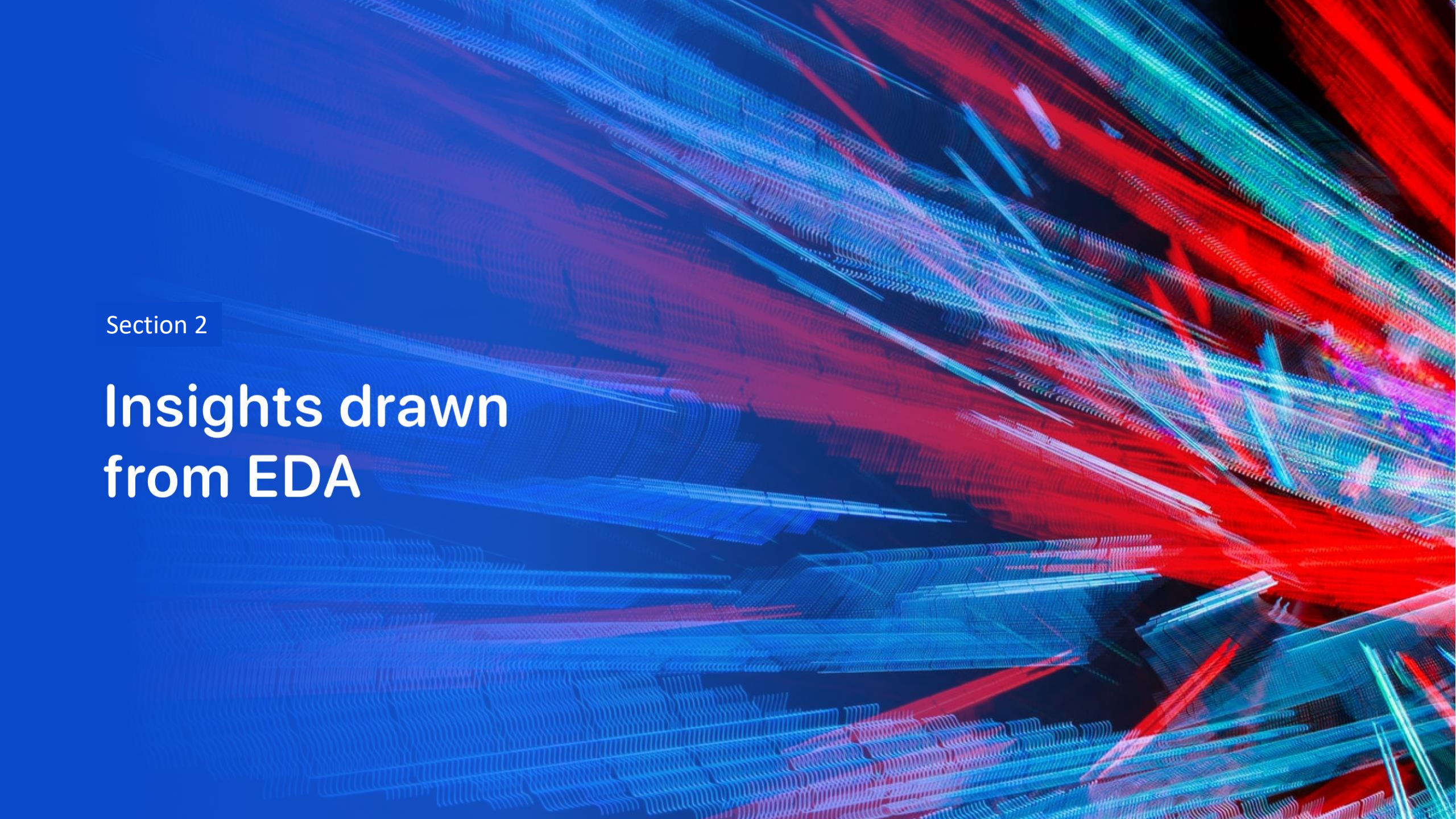
Predictive Analysis (Classification)



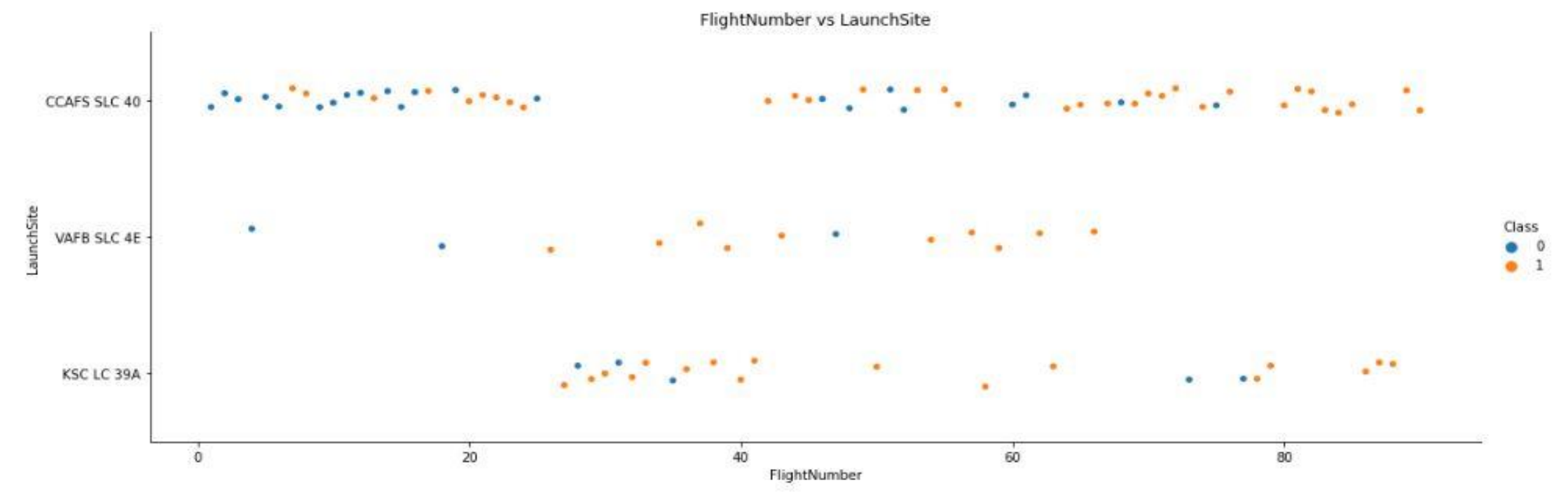
https://github.com/Sabacon/IBM-Data-Science-Professional-Cert/blob/main/Capstone%20Project/5.%20Machine%20Learning/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

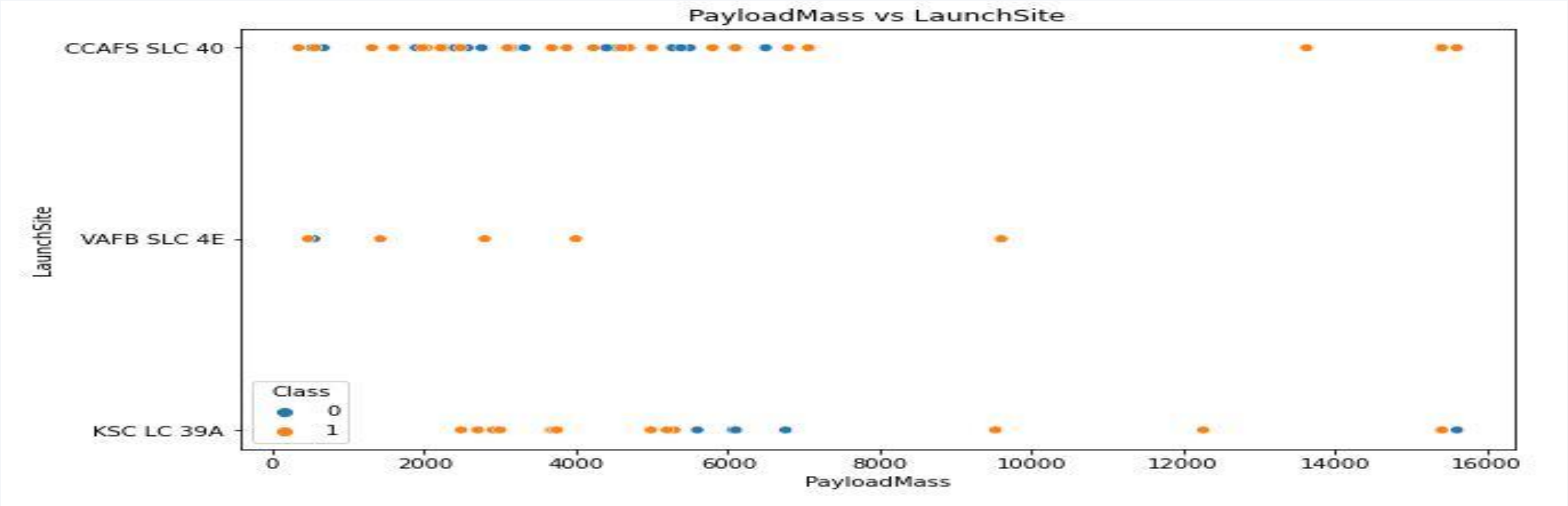

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



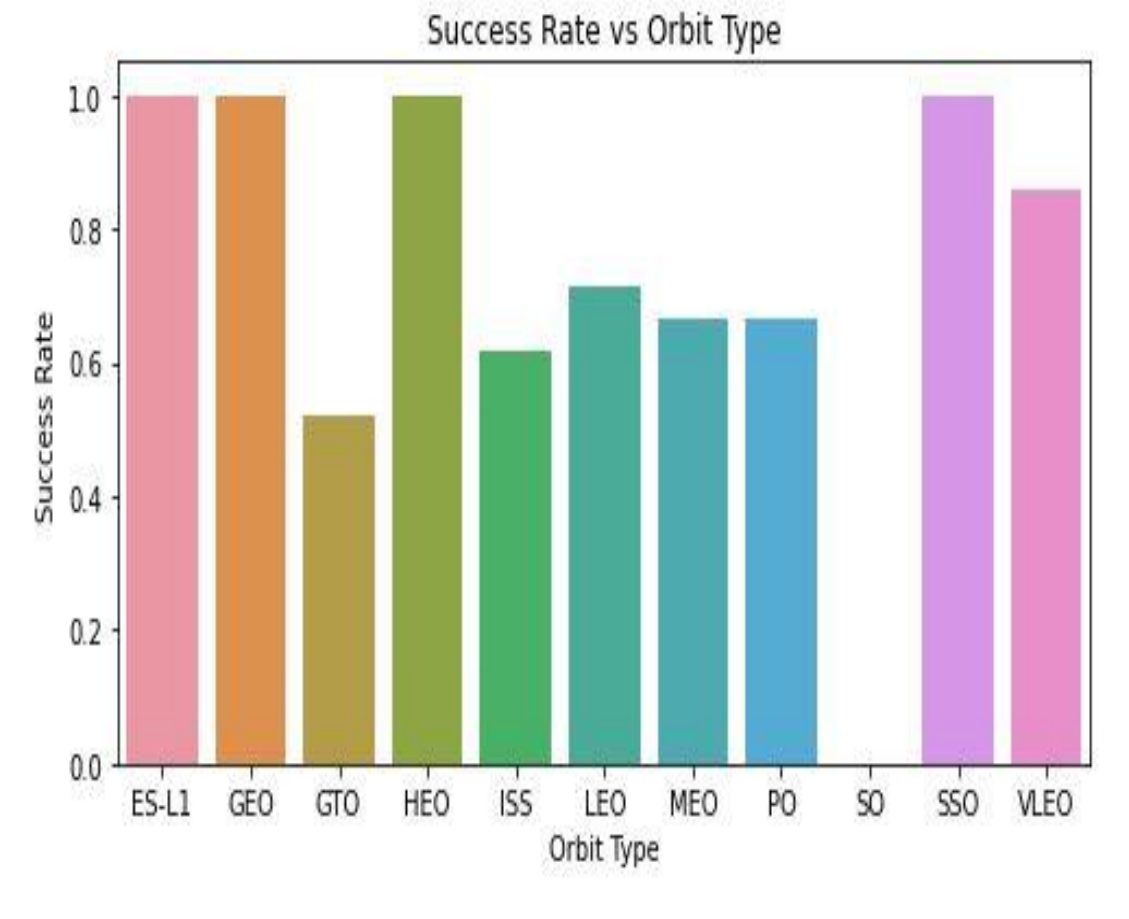
- As the flight number increases the more likely it is for each launch site to register a successful outcome.
- Launch site VAFB SLC 4E has significantly less flight numbers and a very high success rate.

Payload vs. Launch Site



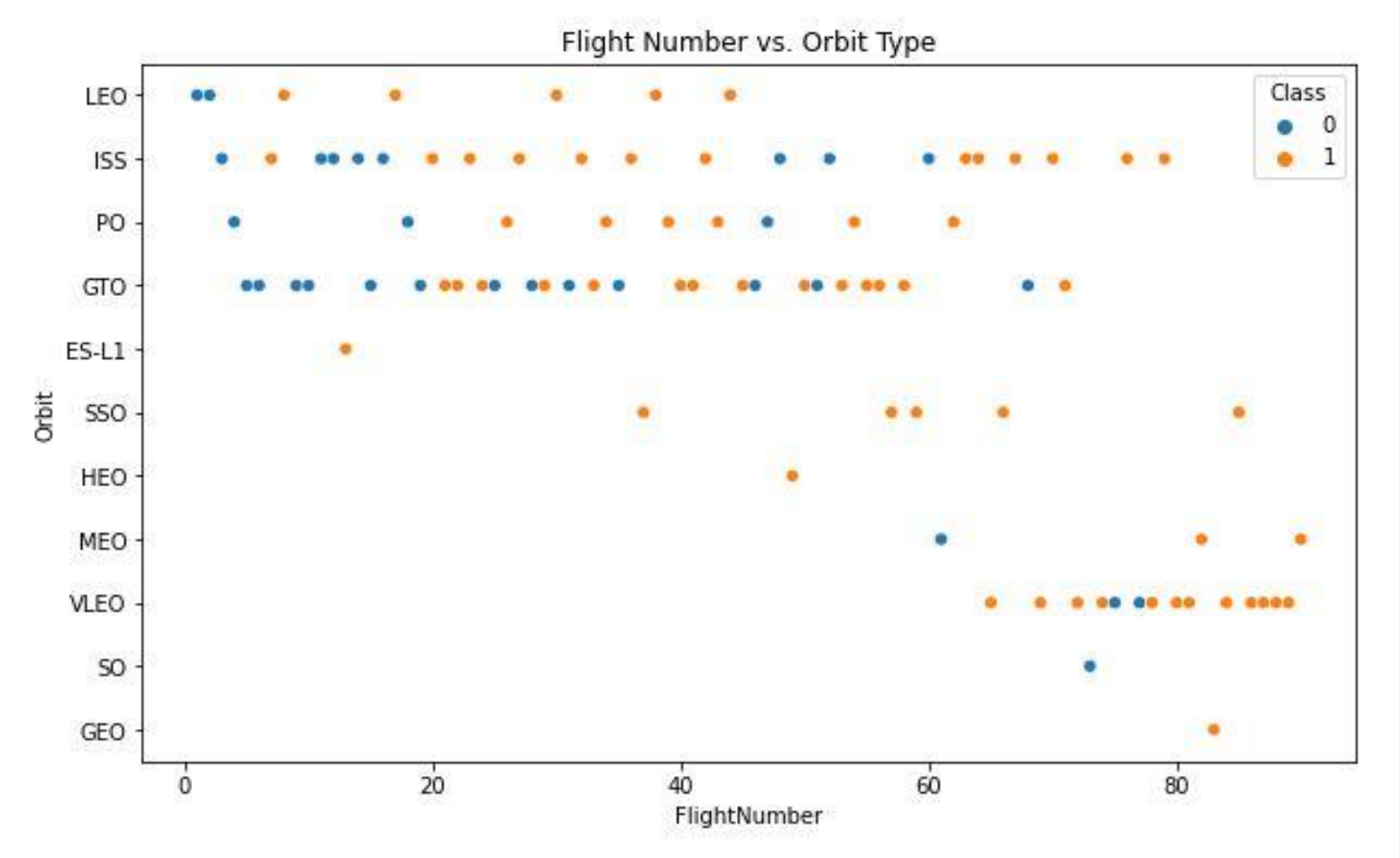
- Launch site VAFB SLC 4E didn't launch any rockets with a payload mas greater than 10000.
- Most of the rockets launched by each site had a payload mass less than 8000 kg.

Success Rate vs. Orbit Type



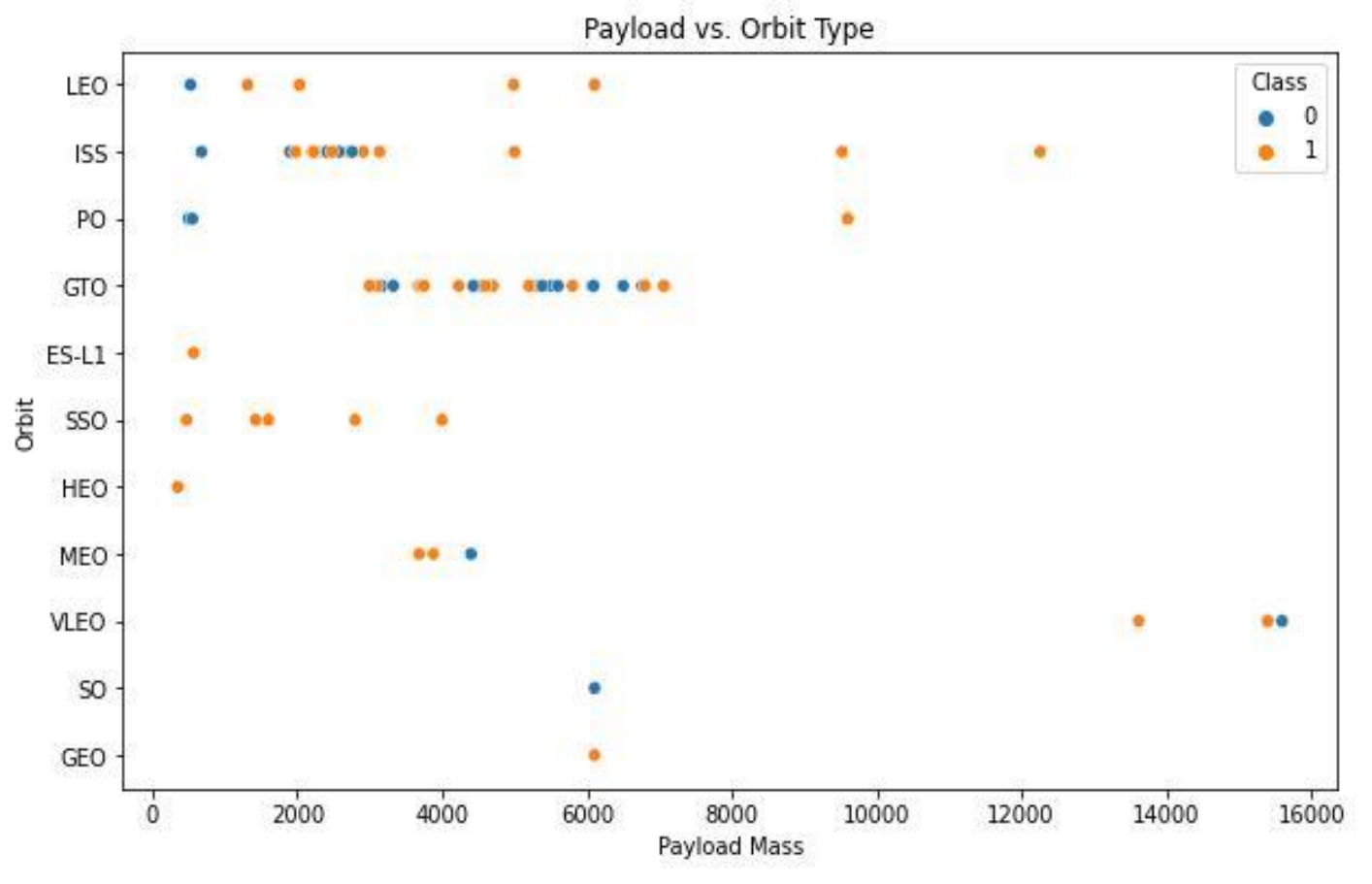
- Orbit type SO has a success rate of 0.
- Five of the eleven orbit types has a success rate greater than 80% with 4 having 100%.

Flight Number vs. Orbit Type



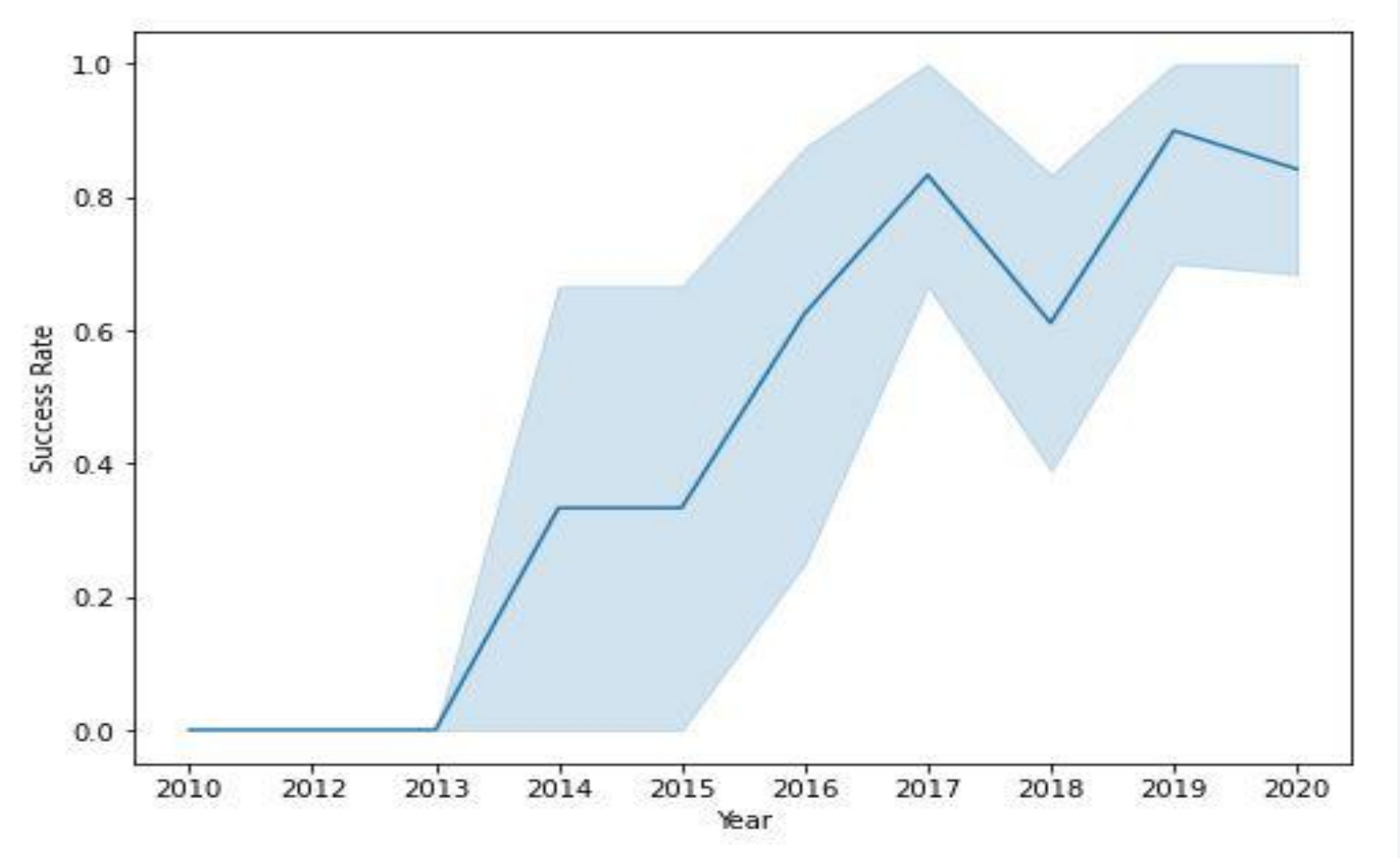
- The greater the flight number the more likely it is to record a successful outcome.
- Orbits SO and GEO only has one flight number.

Payload vs. Orbit Type



- Most rockets launched had a payload mass less than 8000kg.
- The likelihood of a successful outcome increased with an increase in payload mass.

Launch Success Yearly Trend



- The first stage did not land successfully for any of the first three years.
- The success rate's general trend was up for the subsequent years with a decrease in 2018.

All Launch Site Names

- A query of the SpaceX table revealed four unique launch sites as shown on the graphic.

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

Launch Site Names Begin with 'CCA'

	Date	Time_(UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing__Outcome
0	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	08-12-2010	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	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
3	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Upon querying the SpaceX table, the first five launch sites with names beginning with ‘CCA’ are shown above.

Total Payload Mass

- Boosters from NASA carried a total payload mass of 45596 kg.

	SUM(PAYLOAD_MASS__KG_)
0	45596

Average Payload Mass by F9 v1.1

- Booster version F9 v1.1 carried an average payload mass of 2928.4 kg.

	AVG(PAYLOAD_MASS__KG_)
0	2928.4

First Successful Ground Landing Date

- The first successful ground landing was recorded on the first day of May the year 2017.

	min(Date)
0	01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- The query results shown represent the boosters with successful drone ship landings that had a payload mass greater than 4000 kg but less than 6000 kg.

Total Number of Successful and Failure Mission Outcomes

- The ratio of missions that were successful to the ones that failed is 100:1.

	Outcome	Number
0	Failure	1
1	Success	100

Boosters Carried Maximum Payload

- We used a subquery to find the booster versions that carried the maximum payload mass.
- The results are shown in the graphic.

	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

	Year	Month	Landing__Outcome	Booster_Version	Launch_Site
0	2015	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- This represents the list of landing outcomes that were a failure in drone ships in 2015.
- The list also includes their booster versions, launch sites and the exact months.

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

	Landing__Outcome	Number
0	Success	20
1	Success (drone ship)	8
2	Success (ground pad)	6

- Landing outcomes between 2010-06-04 and 2017-03-20 ranked in descending order.

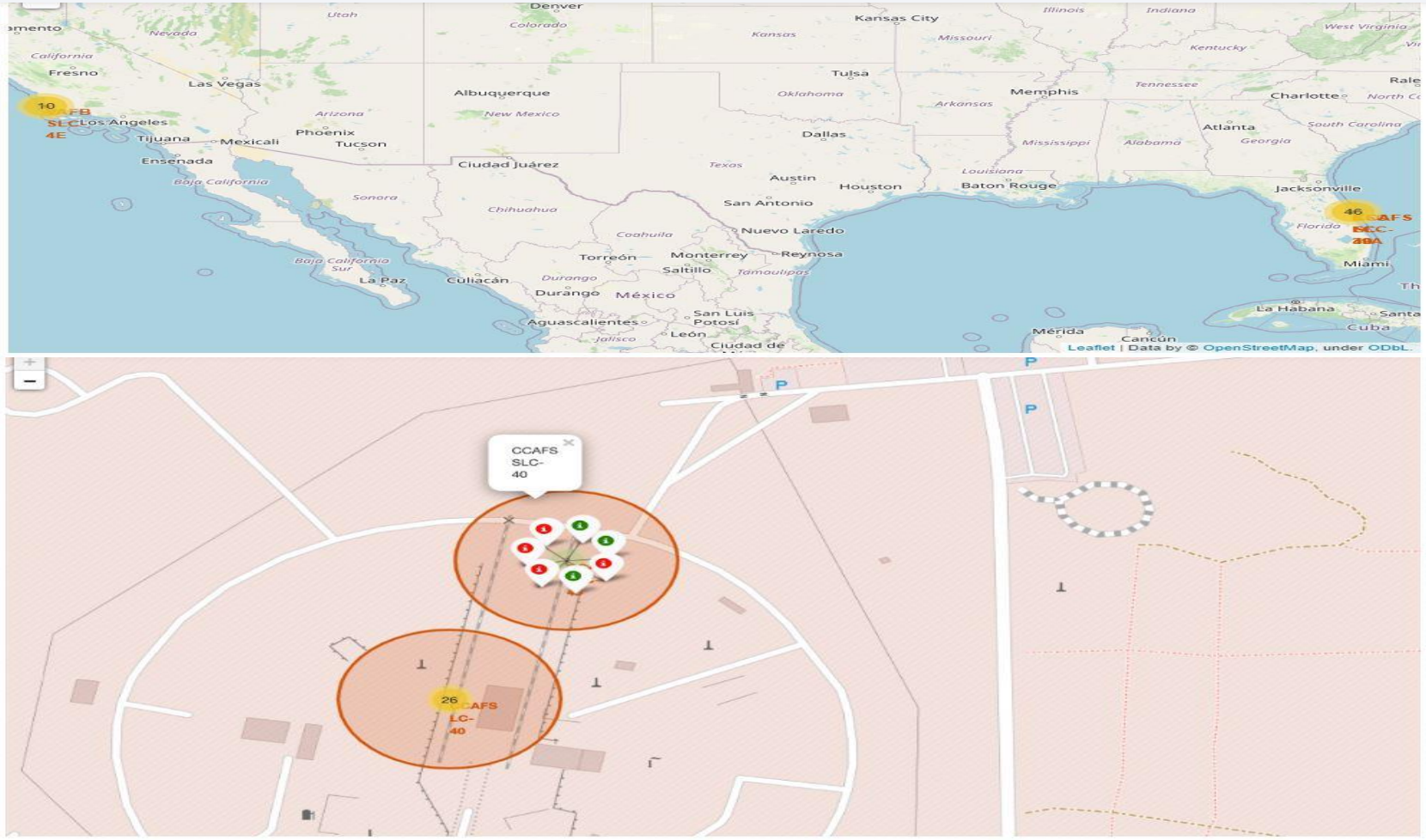
Section 3

Launch Sites Proximities Analysis

All Launch Sites Marked

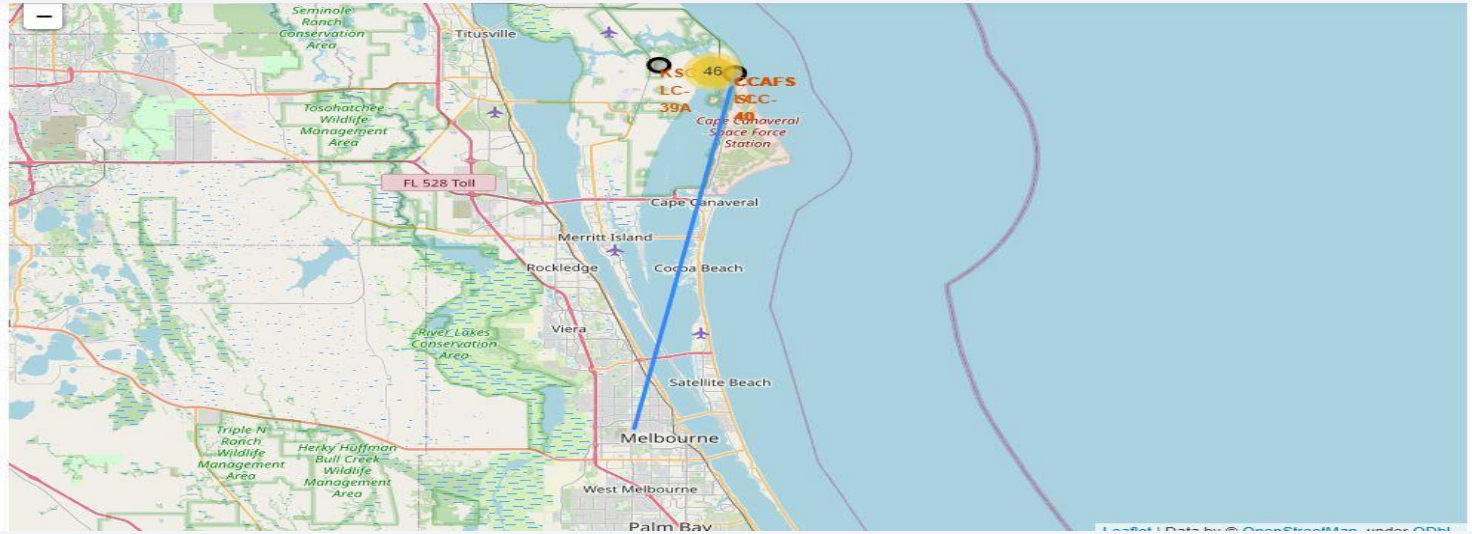
- All the launch sites are located along a coastline.

Launch outcomes for each site

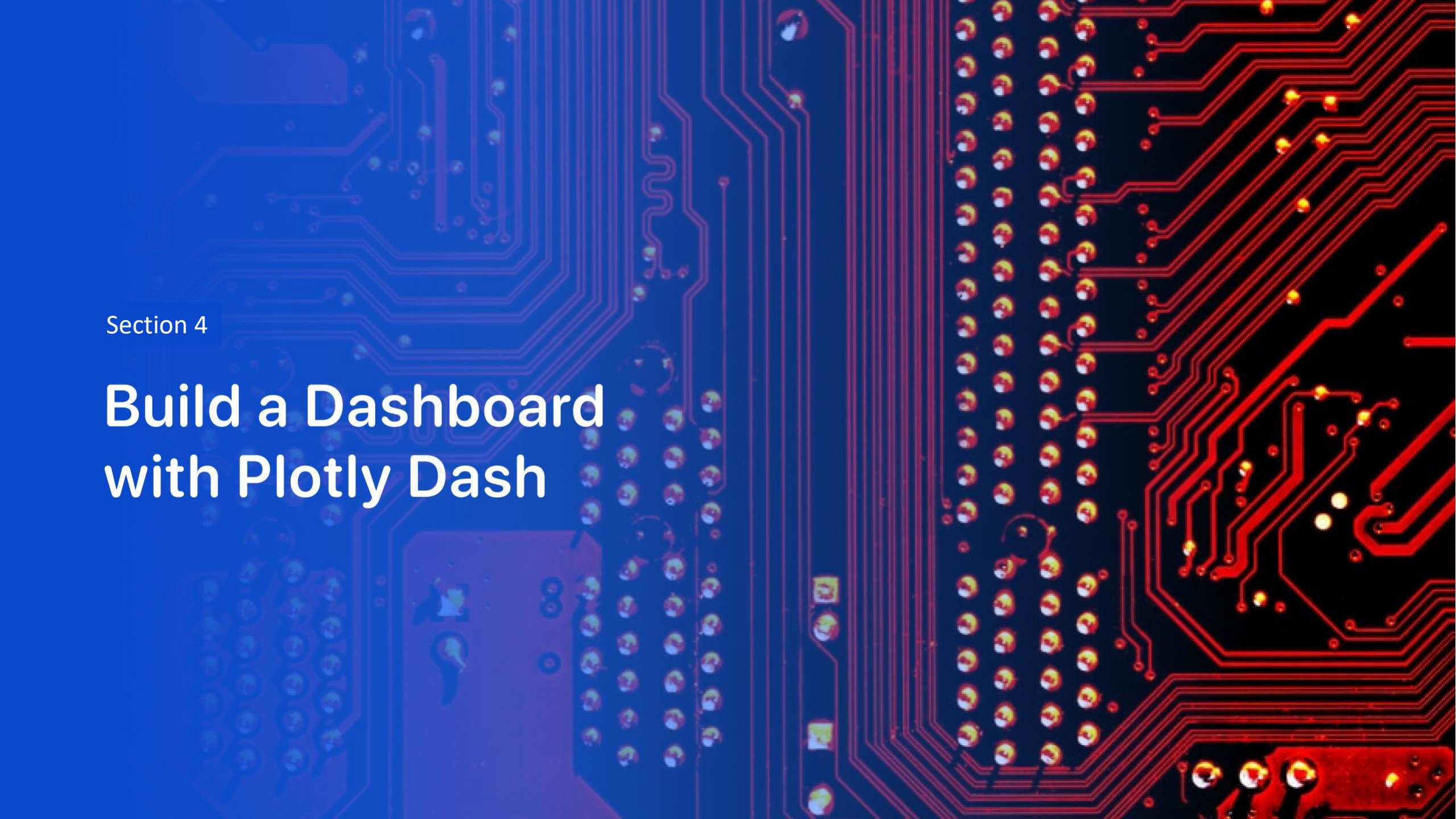


- The East coast has 46 outcomes while the West has 10.
- Successful outcomes are shown in red; failed ones in blue.

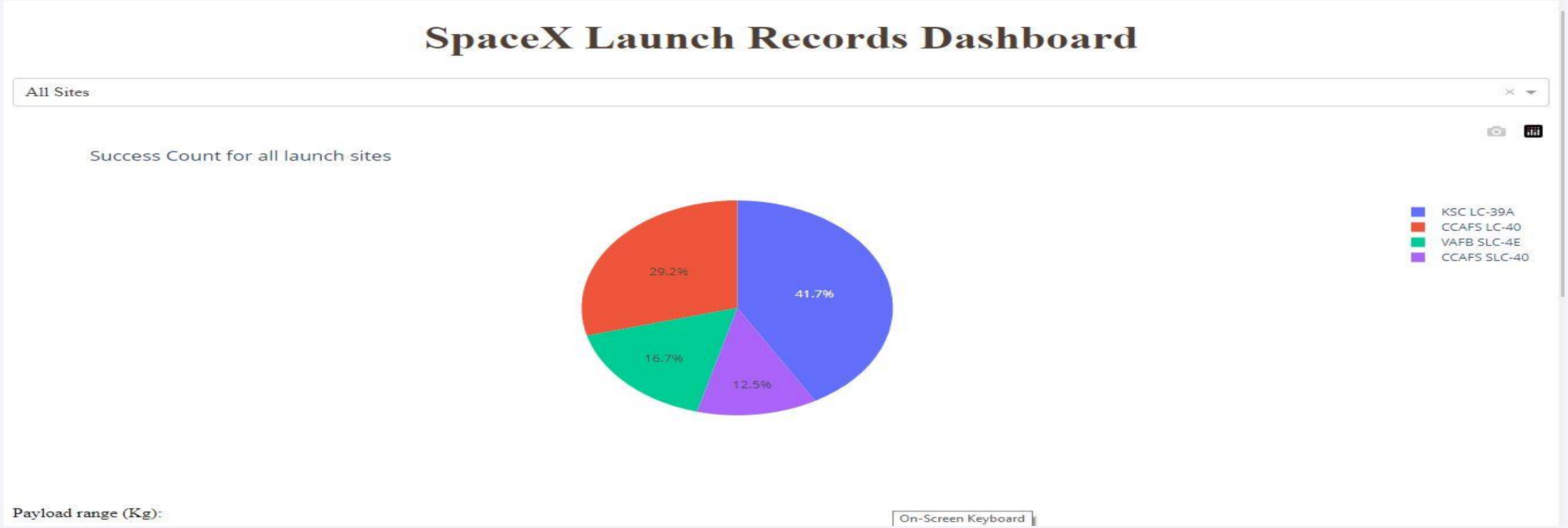
Distance from a Launch Site to its Proximities



- First graphic shows distance between a launch site to its nearest coastline.
- Graphic 2 shows a launch site to its nearest town.

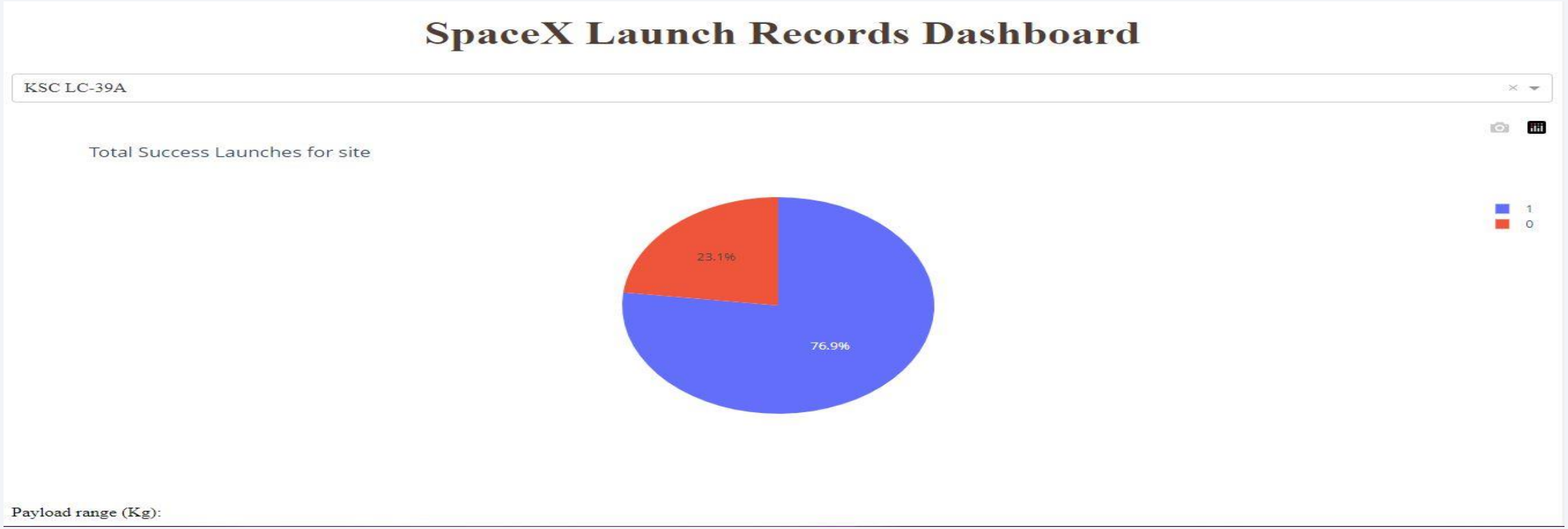

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



- KSC LC-39A had the most successful outcomes; CCAFS SLC-40 with the least.

Site with Highest Success Ratio



- Over three quarters of the launches from KSC LC-39A were successful.

Payload v Launch Outcome

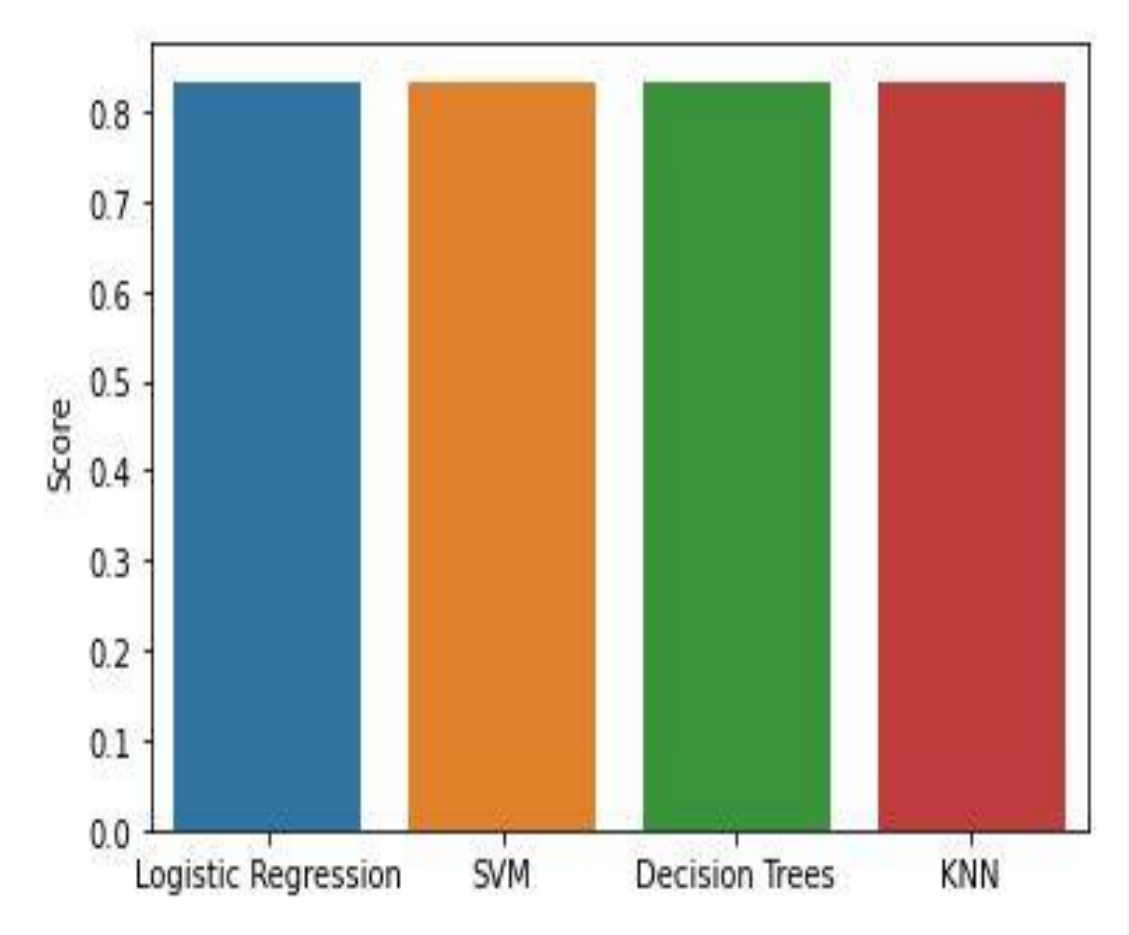
- With a payload range up to 7000 kg, FT had the most successful outcomes with v1.1 showing the most failed outcomes.

Section 5

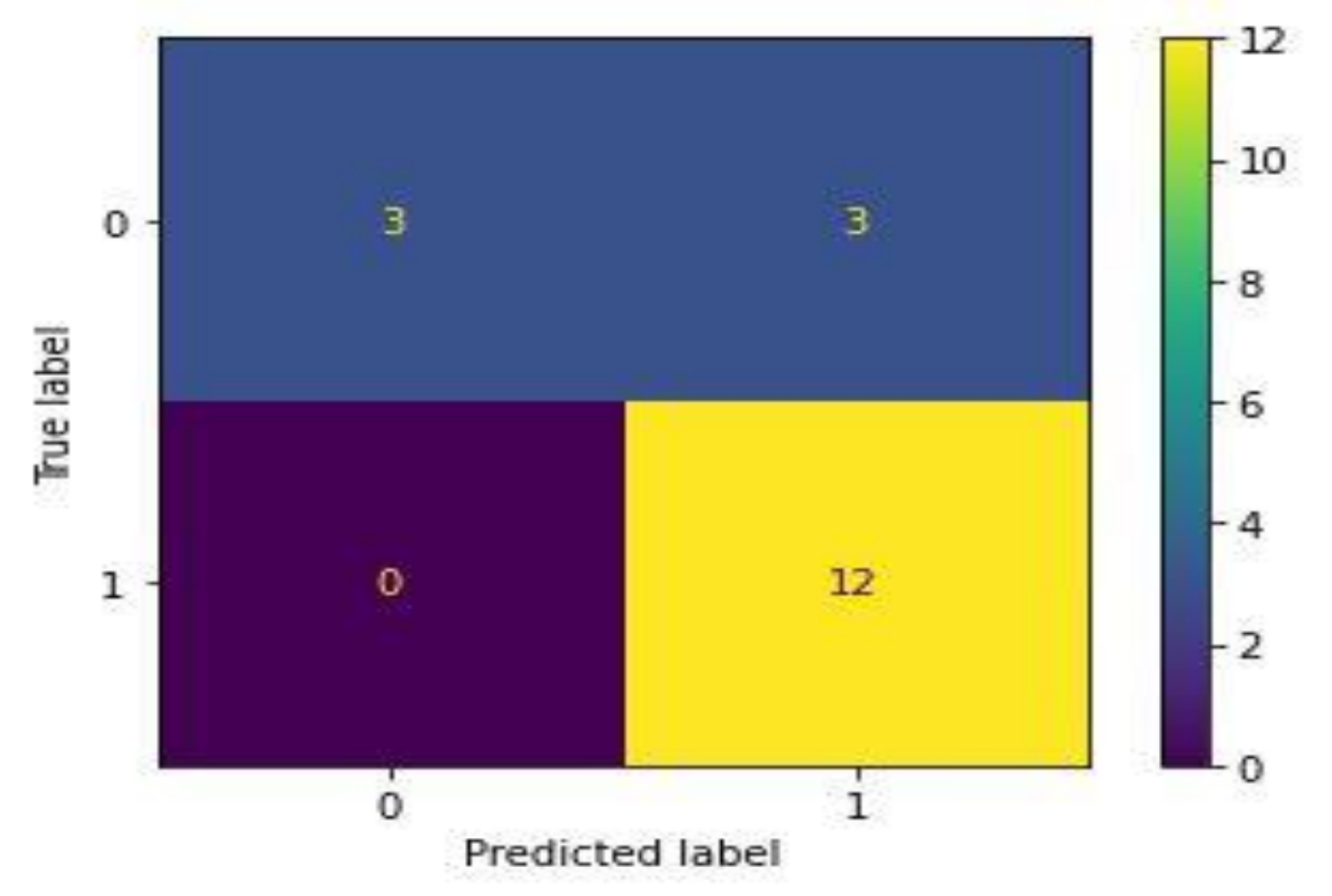
Predictive Analysis (Classification)

Classification Accuracy

- The four classification models we build performed exactly the same with an accuracy of 83.33%.
- This is surprising but not unheard of if the data set is small and has class labels that can be very easily distinguished.



Confusion Matrix



- Out of the 18 observations from the test set only 3 were misclassified.
- The model predicted 3 outcomes that were actually failed, as successful outcomes.

Conclusions

- The vast majority of landing outcomes were successful.
- There has been a steady increase in the success rate after the first 3 years.
- To increase the likelihood of getting a successful outcome, we need to have a high flight number and a high payload mass.
- We need to collect more data

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!