IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Edgar Gonzalez
26 – 01 – 2024

Table of contents

Executive Summary

- Introduction
- Methodology
  - Data collection
    - SpaceX API
    - Scraping
  - Data Wrangling
  - EDA
    - With Data Visualization
    - With SQL
  - Build an Interactive Map with Folium
  - Build a Dashboard with Plotly Dash
  - Predictive Analysis
- Results
- Conclusion
- Appendix

Executive Summary

Space X – Site Launch Analysis

The document take you through a travel about the methodologies used to gather all relevant information present in several platform as web pages, APIs, CSV files, SQL tables, also shows how to give shape to that information to have better understanding of the data therefore at the same time the outcomes are showed utilizing dashboards and plotted graphics to present the result as comprehensible as possible.

- The methodologies to explore are

Connection to API, Web scrapping and data wrangling, Exploratory analysis with plots and Sql Queries and show results with Plotly building dashboards to find relevant information and insights. Some information will be presented over maps to have better understanding where the launches are located using folium.

The predictive analysis realized is presented in a bar graphic and a confusion matrix was used to know the reliability and performance of the model for this scenario.

- At the end the results showed make emphasis in the next

The KSC LC-39 is the site launch with the major success rate than other sites and the CCA FS-SLC40 has the most launches, The Orbits with major success rate are ES-L1, GEO, HEO, SSO, just two booster could load above 6000kg FT. For the predictive analysis the model with the best accuracy is "Method Decision Tree Classifier" and it has a accuracy above 0.8 as the others predictive analysis.

Introduction

Project background and context

- SpaceX aims to revolutionize space technology with the ultimate goal of enabling people to live on other planets.
- The company's business model focuses on reducing the costs associated with space logistics by *developing reusable rockets*.
- The company generates revenue by launching objects and people into space.
- Falcon 9 rocket launches is advertised on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Problems to be addressed

- Predict if the Falcon 9 first stage will land successfully.
- Determine if the first stage will land and therefore to determine the launch cost, beside this information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Source: Introduction - How Does SpaceX Make Money? Dissecting Its Business Model (2024)

Section 1

Methodology

Methodology

Data collection methodology:

- API Requests

The dataset was collect using the API from spacex that contains all the information form the last launches realized by the company, therefore analysis of .json documents got it from the same web page.

- Webscrapping using BeautifulSoup

Information gathered from available web pages i.e. wikipedia

Perform data wrangling

- Numpy used to identify null values, types, counts values, means.
- Seaborn used to figure out graphically relations between concepts i.e Launch sites, payloads etc.

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

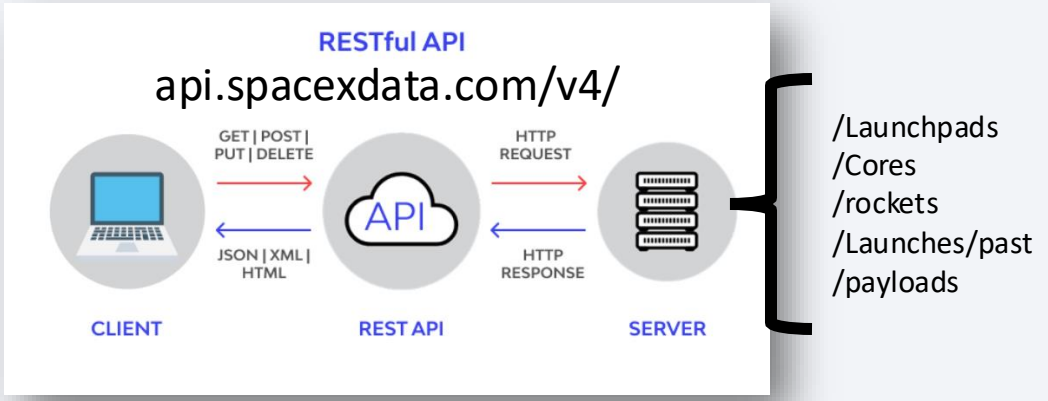
Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

Data collection API

https://github.com/Ragde26/CapstoneSpaceX/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

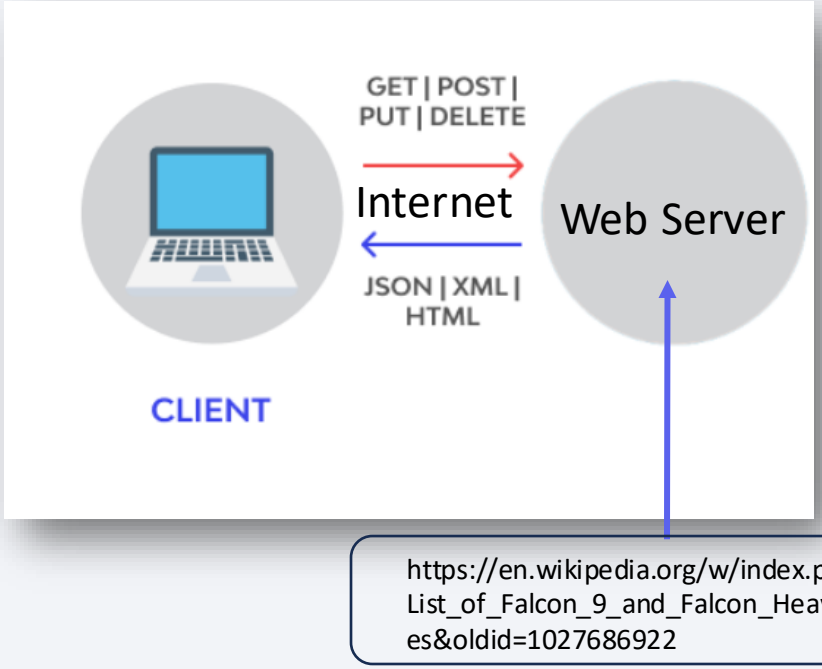


Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

Webscraping - Wikipedia
https://github.com/Ragde26/CapstoneSpaceX/blob/main/jupyter-labs-webscraping.ipynb

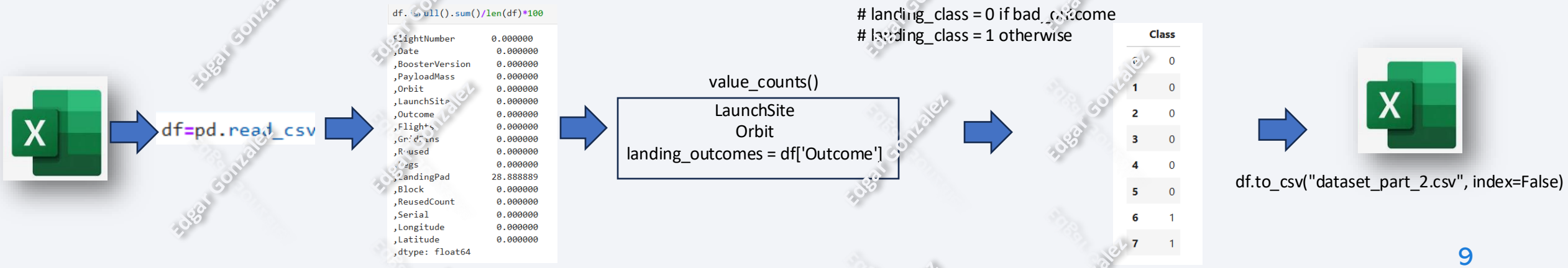


Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

Data_Wrangling

https://github.com/Raggle26/CapstoneSpaceX/blob/main/labs-jupyter-spacex-Data wrangling.ipynb



EDA with Data Visualization

Summarize what charts were plotted and why you used those charts

- Category plot - To see how the FlightNumber (indicating the continuous launch attempts.) and Payload variables would affect the launch outcome.
- Category Plot - For visualize the relationship between Flight Number and Launch Site
- Category PLot - For visualize the relationship between Payload Mass and Launch Site
- Bar Chart - For visualize the relationship between success rate of each orbit type¶
- Category Plot - For Visualize the relationship between FlightNumber and Orbit type
- Category Plot - For Visualize the relationship between Payload Mass and Orbit type
- Line Plot - For Visualize the launch success yearly trend

Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

https://github.com/Ragde26/CapstoneSpaceX/blob/main/edadataviz.ipynb

EDA with SQL

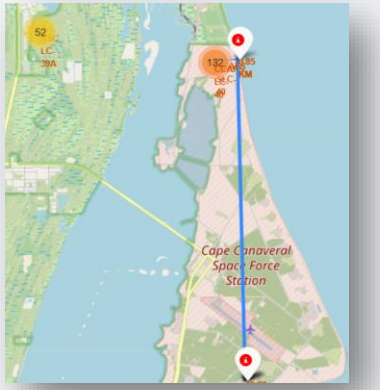
- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

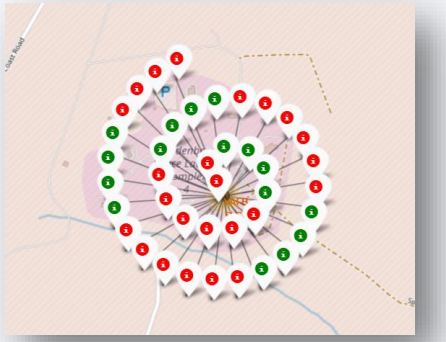
https://github.com/Ragde26/CapstoneSpaceX/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- con = sqlite3.connect("my_data1.db")
- cur = con.cursor()
- %sql sqlite:///my_data1.db
- %sql DROP TABLE IF EXISTS SPACEXTABLE;
- %sql create table SPACEXTABLE as select * from SPACEXTBL where Date is not null
- %sql select DISTINCT Launch_Site from SPACEXTABLE
- %sql select * from SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' limit 25
- %sql select SUM(PAYLOAD_MASS__KG_) as Total_PAYLOAD_KG from SPACEXTABLE WHERE Customer='NASA (CRS)'
- %sql select AVG(PAYLOAD_MASS__KG_) as Average_Payload from SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
- %sql select min(Date) from SPACEXTABLE where Landing_Outcome = "Success (ground pad)"
- %sql select Booster_Version from SPACEXTABLE where Landing_Outcome = "Success (drone ship)" and (4000<PAYLOAD_MASS__KG_'<6000)
- %sql select Mission_Outcome, count(*) as Total from SPACEXTABLE group by Mission_Outcome
- %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
- %sql select "Landing_Outcome", "Booster_Version", "Launch_Site", substr(Date, 6, 2) as 'month', substr(Date,0,5) as 'year' from SPACEXTABLE where year = '2015' and Landing_Outcome = "Failure (drone ship)"
- %sql select Landing_Outcome, count(*) as Total from SPACEXTABLE where Date > '2010-06-04' and Date < '2017-03-20' group by Landing_Outcome order by Total desc

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects





site_map.add_child(circle) - to highlight and mark all launch sites on the map.
site_map.add_child(marker) - to mark the exact position of each launch site.
site_map2.add_child(marker_cluster2) to enhance the map by adding the launch outcomes for each site and have mayor visibility for each launch.
site_map2.add_child(mouse_position) - It was used to know the coordinates of an interest point
site_map2.add_child(distance_marker)- It was used to obtain the distance between points
site_map2.add_child(lines) - It was used to trace a line between two point of interest.

- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

https://github.com/Ragde26/CapstoneSpaceX/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions

Drop-down input Component - To choose the Launch site or select all of them.

Pie Chart – To visualize the success for each launch site or all of them at the time depending of the drop down input.

Range Slider - To select a range Payload in specific.

Scatter chart - To render the success payload in accordance to the range slider

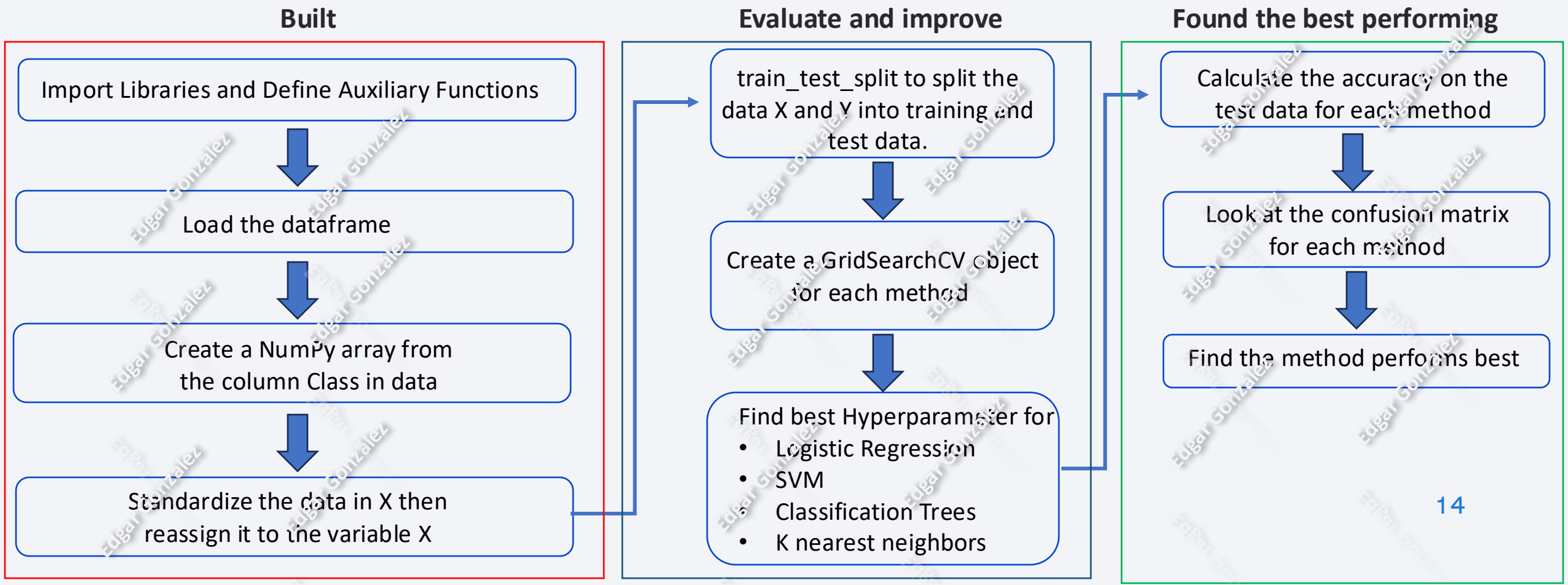
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

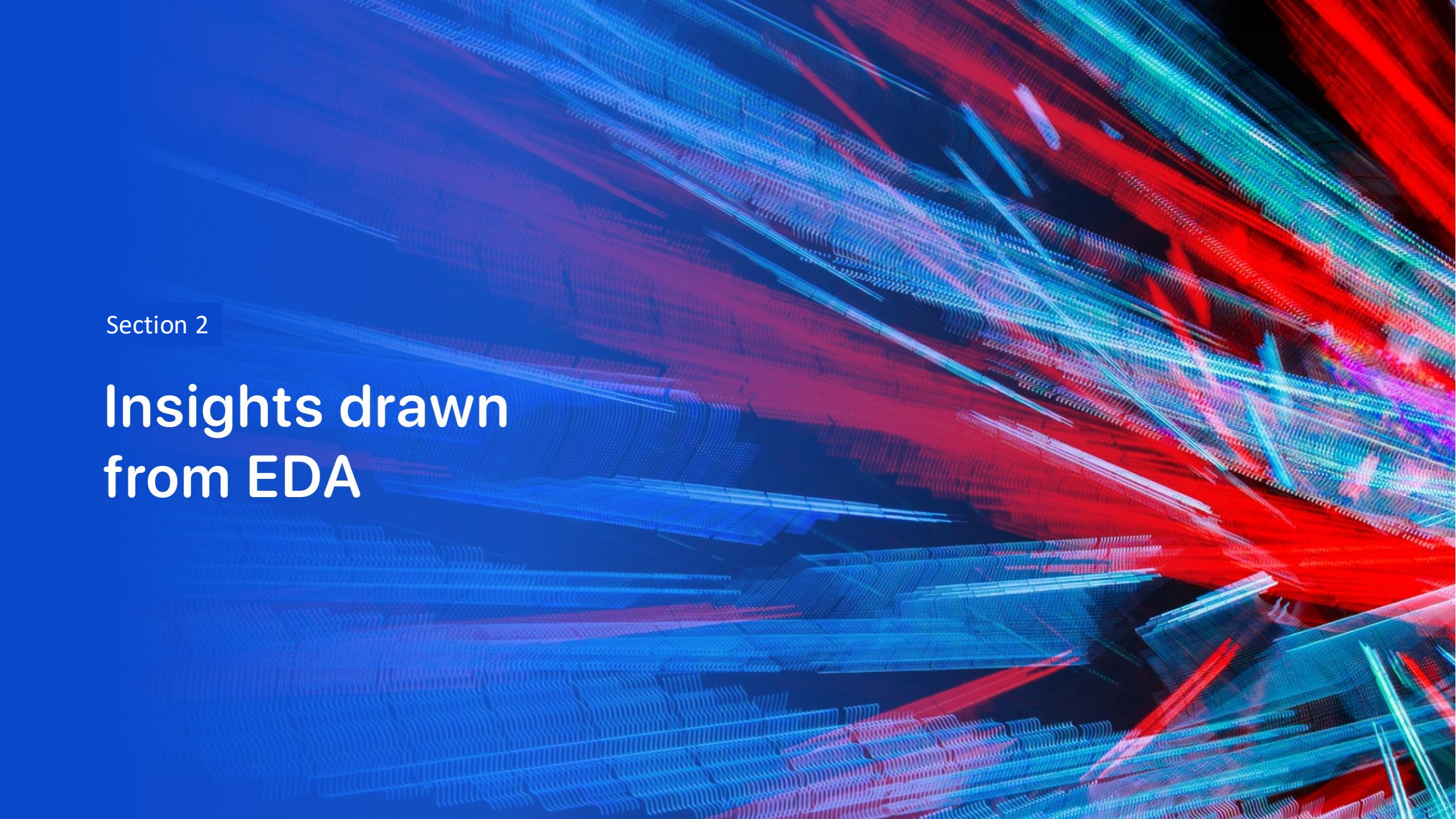
https://github.com/Ragde26/CapstoneSpaceX/blob/main/spacex_dash_app%20final.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

https://github.com/Ragde26/CapstoneSpaceX/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

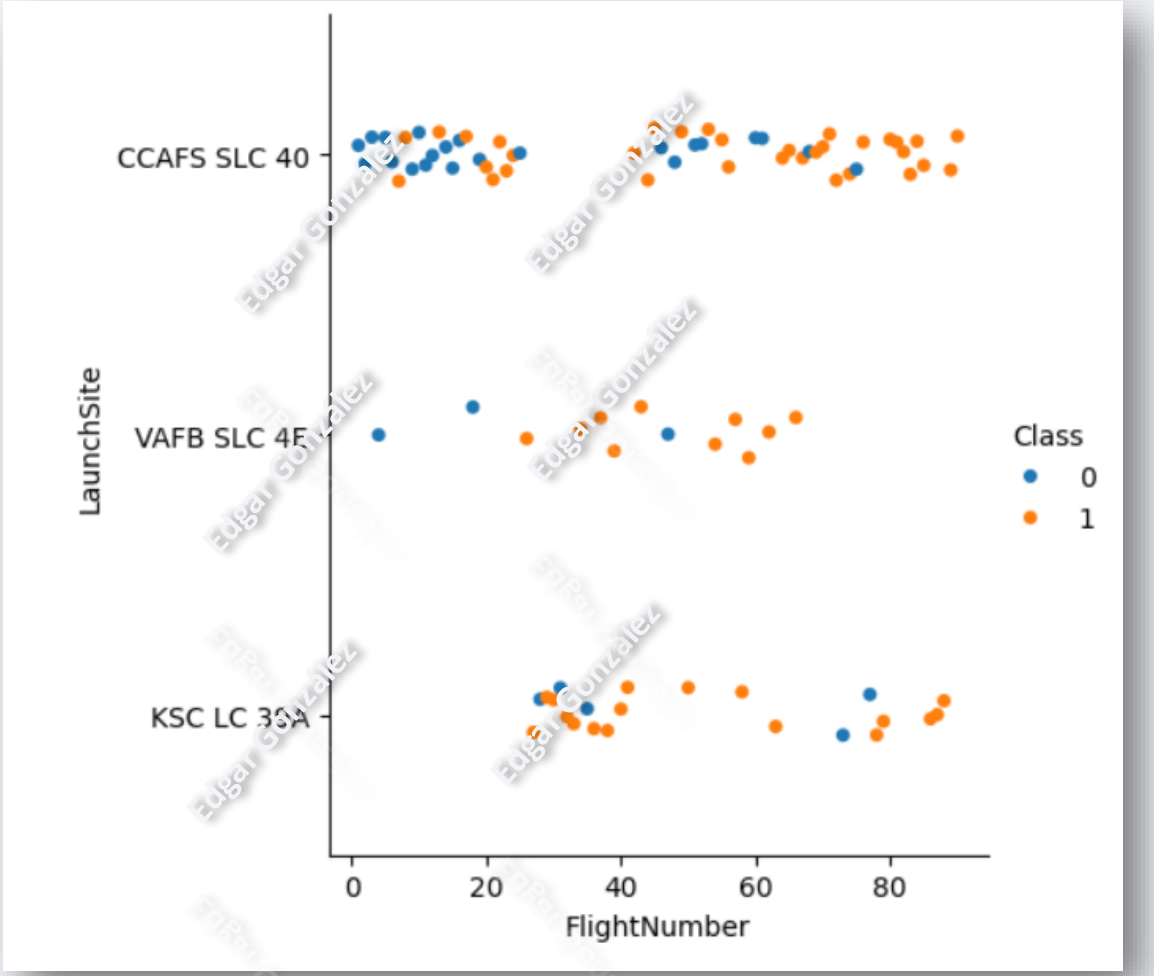



Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site
- Show the screenshot of the scatter plot with explanations

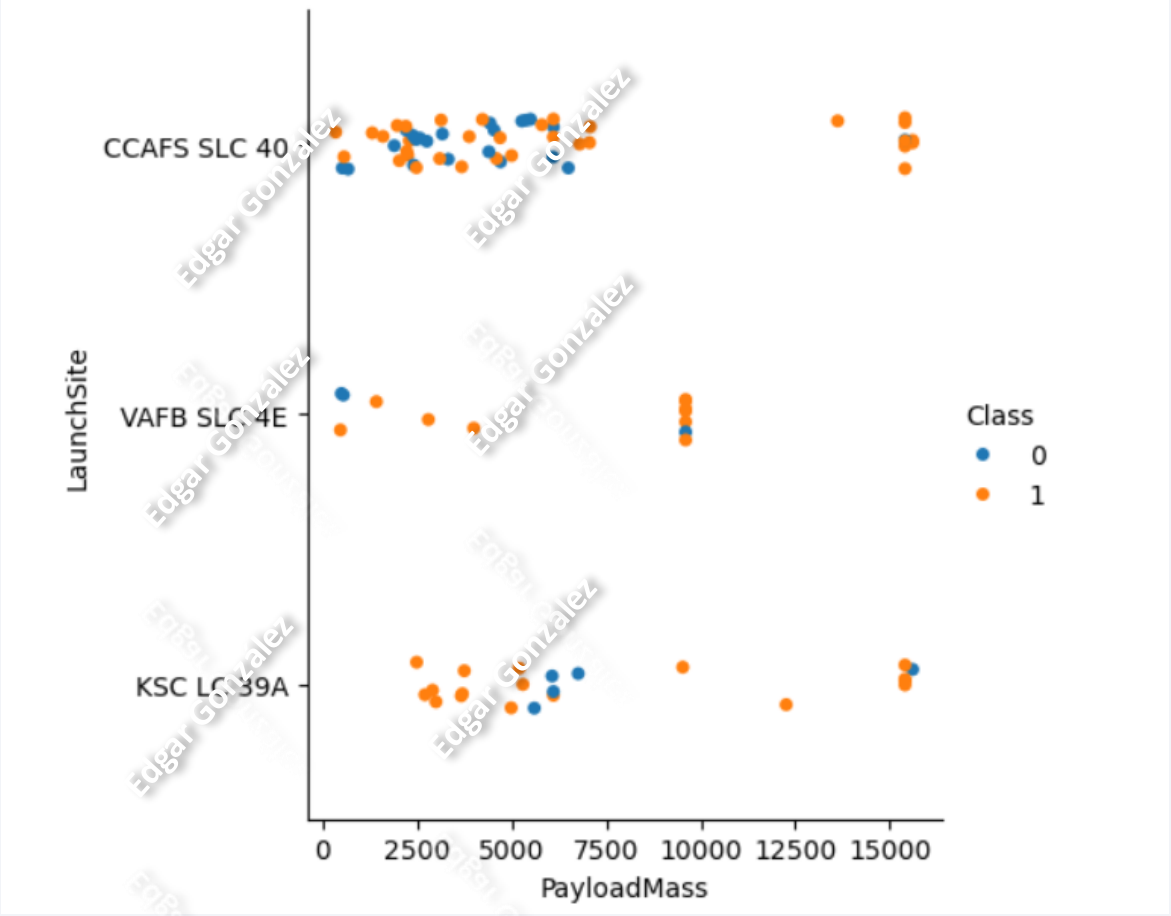
The launch site VAFB SLC4E have more success rate after flight number 20 and the launch site has major success rate after flight number 70



Payload vs. Launch Site

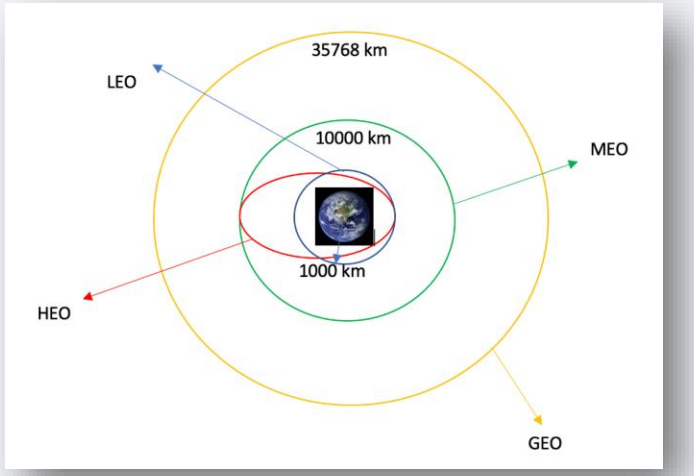
The VAFB-SLC launchsite there are no rockets launched for heavy payload mass the maximum is 10000kg

- Show a scatter plot of Payload vs. Launch Site
- Show the screenshot of the scatter plot with explanations

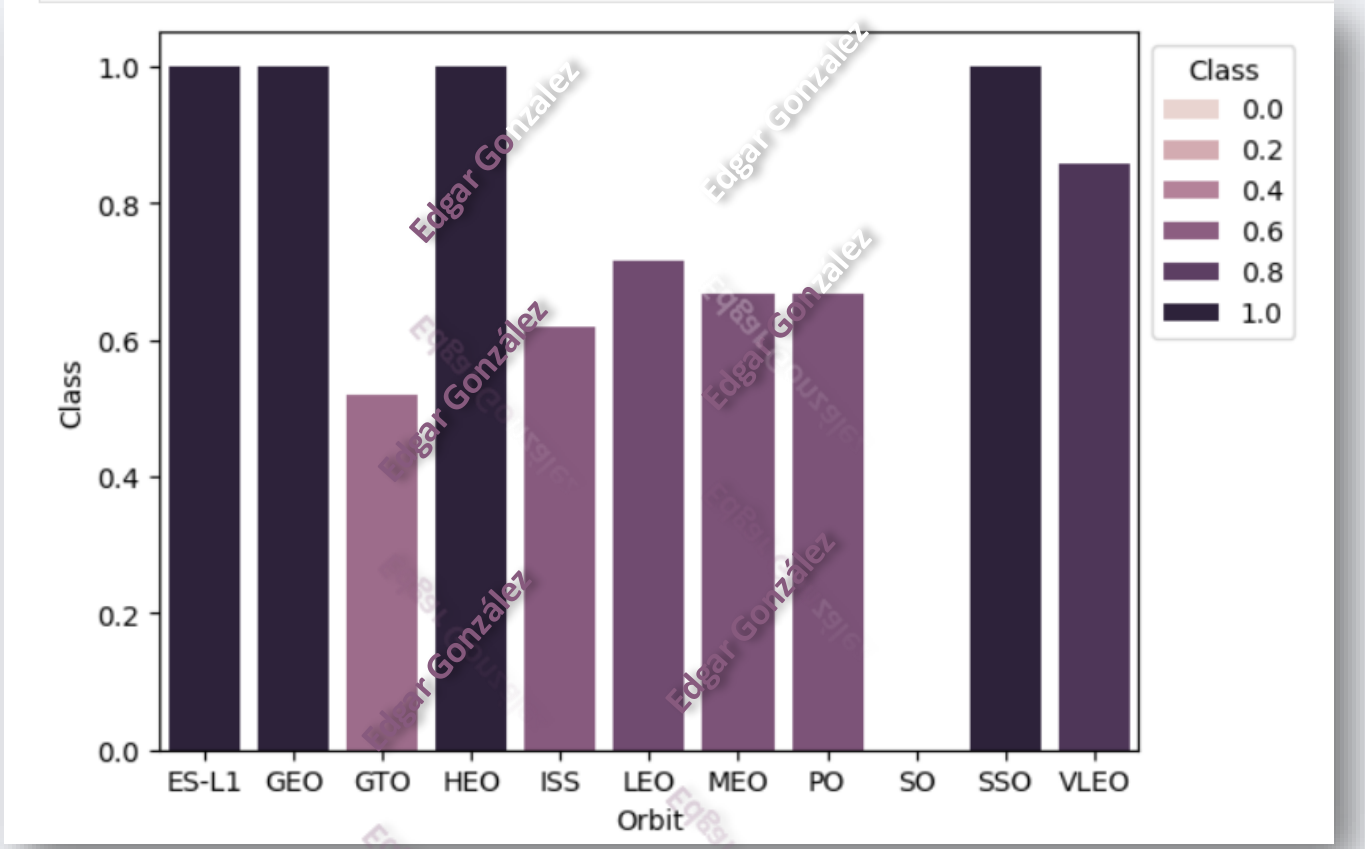


Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the scatter plot with explanations



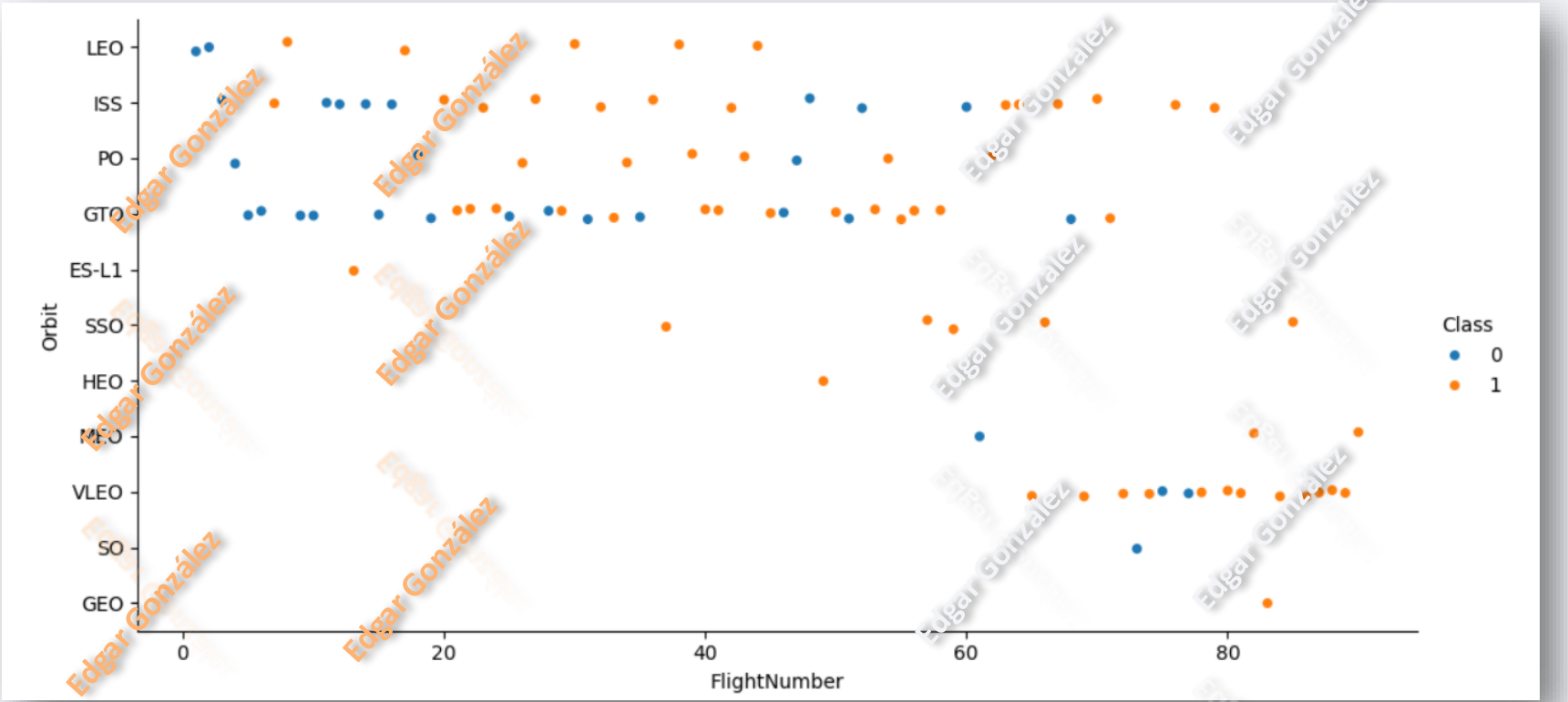
The Orbits with major sucess rate are ES-L1, GEO, HEO, SSO



Flight Number vs. Orbit Type

The SSO Orbit has high rate of success, In the LEO orbit, success is related to the number of flights. In the GTO orbit, there appears to be no relationship between flight number and success.

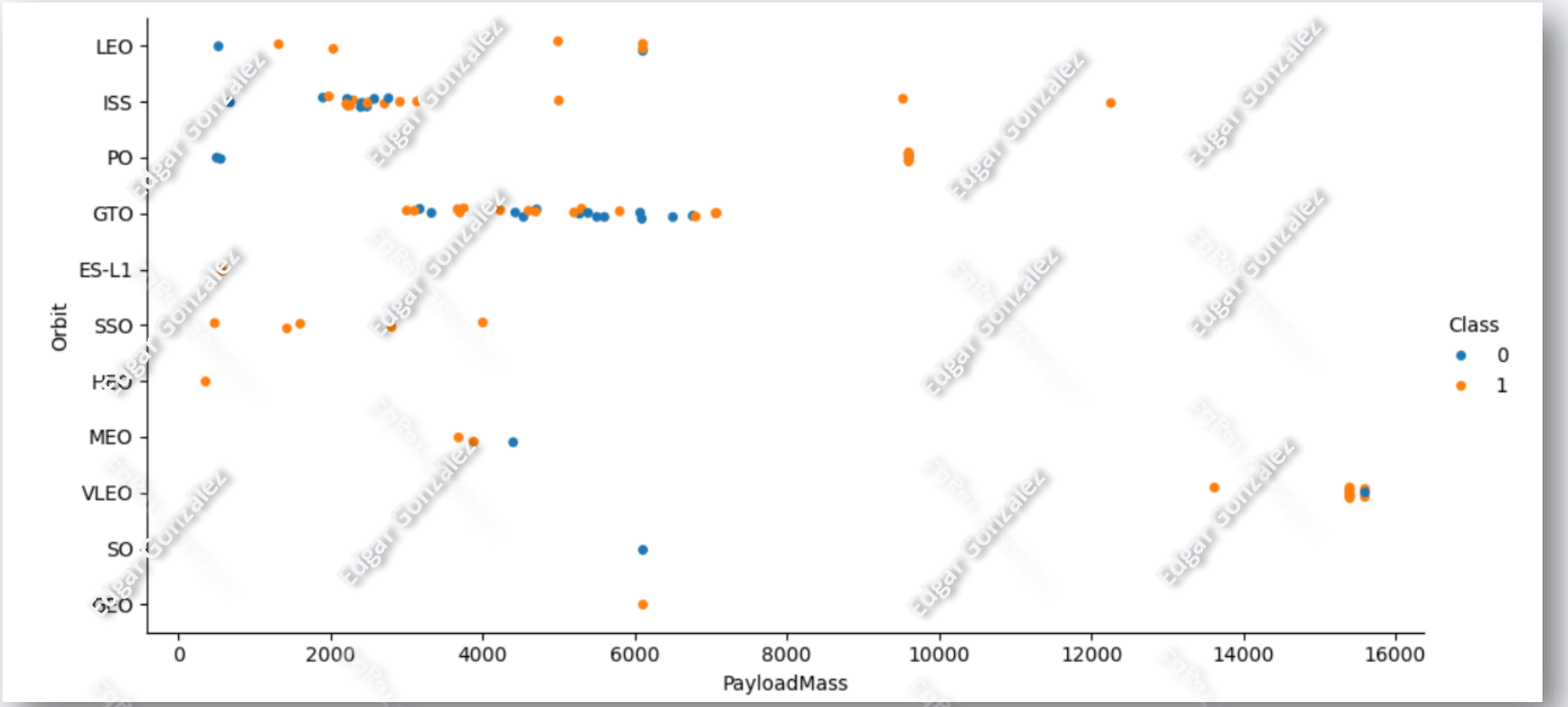
- Show a scatter point of Flight number vs. Orbit type
- Show the screenshot of the scatter plot with explanations



Payload vs. Orbit Type

With heavy payloads the successful landing or positive landing rate are more for Polar, VLEO and ISS.
However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

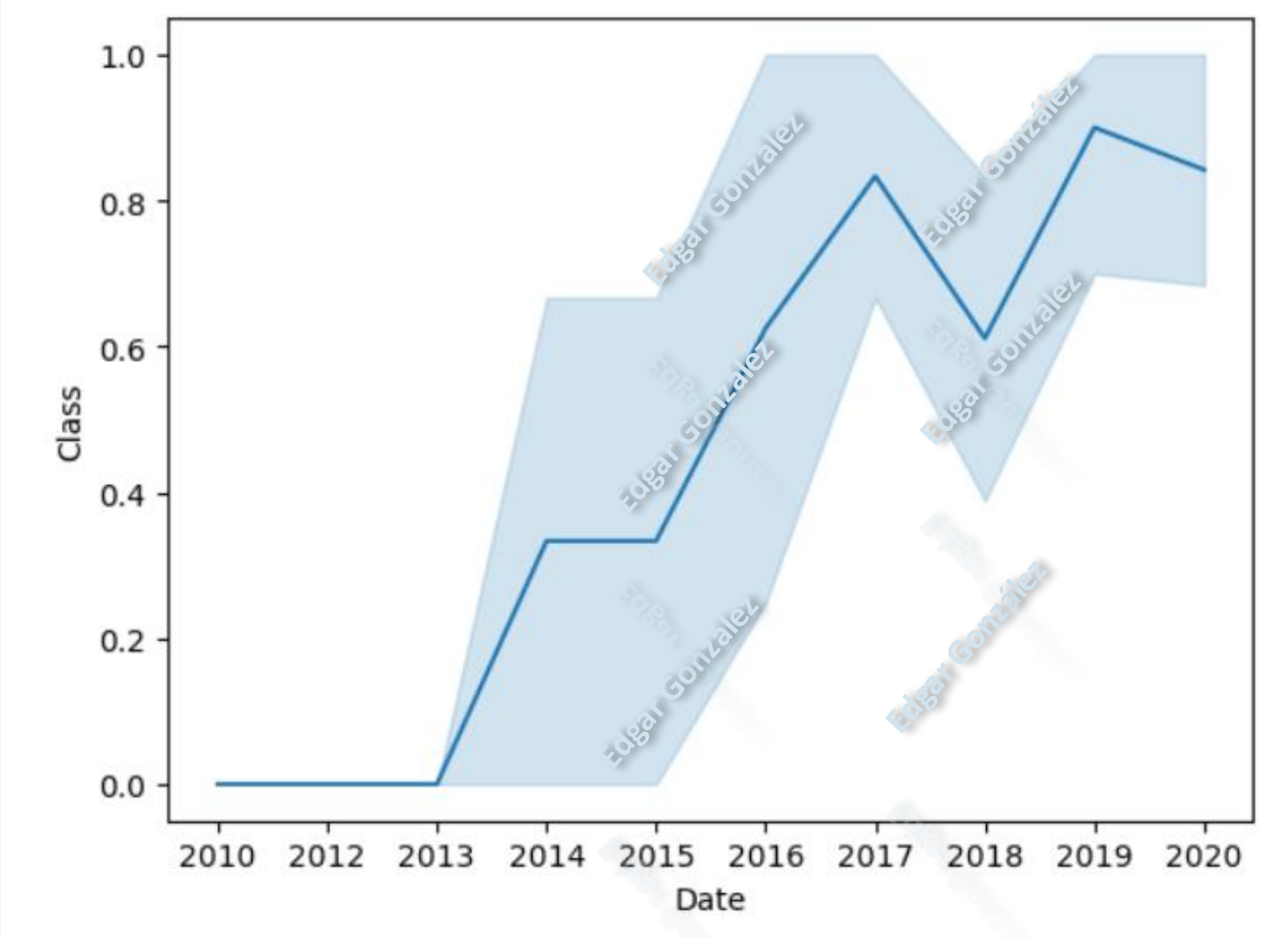
- Show a scatter point of payload vs. orbit type
- Show the screenshot of the scatter plot with explanations



Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Show the screenshot of the scatter plot with explanations

Since 2013 in ahead there was a increased success rate



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

%sql select DISTINCT Launch_Site from SPACEXTABLE

The SELECT DISTINCT statement is used to return only different values in the column

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`
- Present your query result with a short explanation here

%sql select * from SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' limit 5

The **WHERE** clause is used to filter records. It is used to extract only those records that fulfill a specified condition.
The **LIKE** operator is used in a WHERE clause to search for a specified pattern in a column.
There are two wildcards often used in conjunction with the LIKE operator:

- The percent sign % represents zero, one, or multiple characters

LIMIT is used to limit the number of rows returned by a SELECT statement,

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
- Present your query result with a short explanation here

It add the total of payload from the customer nasa and shows the result in a new column named total_payload_kg

```
%sql select SUM(PAYLOAD_MASS__KG_) as Total_PAYLOAD_KG from SPACEXTABLE WHERE Customer='NASA (CRS)'
```

Total_PAYLOAD_KG
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

From all the rows containing F9 v1.1 is getting the average from the data in the column Payload_mass_kg and a new name is added for the new column

```
%sql select AVG(PAYLOAD_MASS__KG_) as Average_Payload from SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```

Average_Payload
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

The min() function returns the smallest value of the selected column.

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome = "Success (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
- Present your query result with a short explanation here

The and operator is used to filter records based on more than one condition

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

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

The Count() function returns the number of rows that matches a specified criteria and The group by statement groups row that have the same values into summary rows,

%sql select Mission_Outcome, count(*) as Total from SPACEXTABLE group by Mission_Outcome

Mission_Outcome	Total
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
- Present your query result with a short explanation here

Subquery used to get the maximum value in the column payload_mass and it is used to filter the column booster_version with those that make match

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

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
- Present your query result with a short explanation here

The query selects the rows with year 2015 and only failure landings to work only with those rows, the columns required are choosen to be display in the table.

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

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

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
- Present your query result with a short explanation here

The element "where" choose only the rows in the range of date desired, the group by gather the similar concepts inisde the column landing_outcome that is displayed and "count" review how many times that concept appear and the outcome is under the column named Total.

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

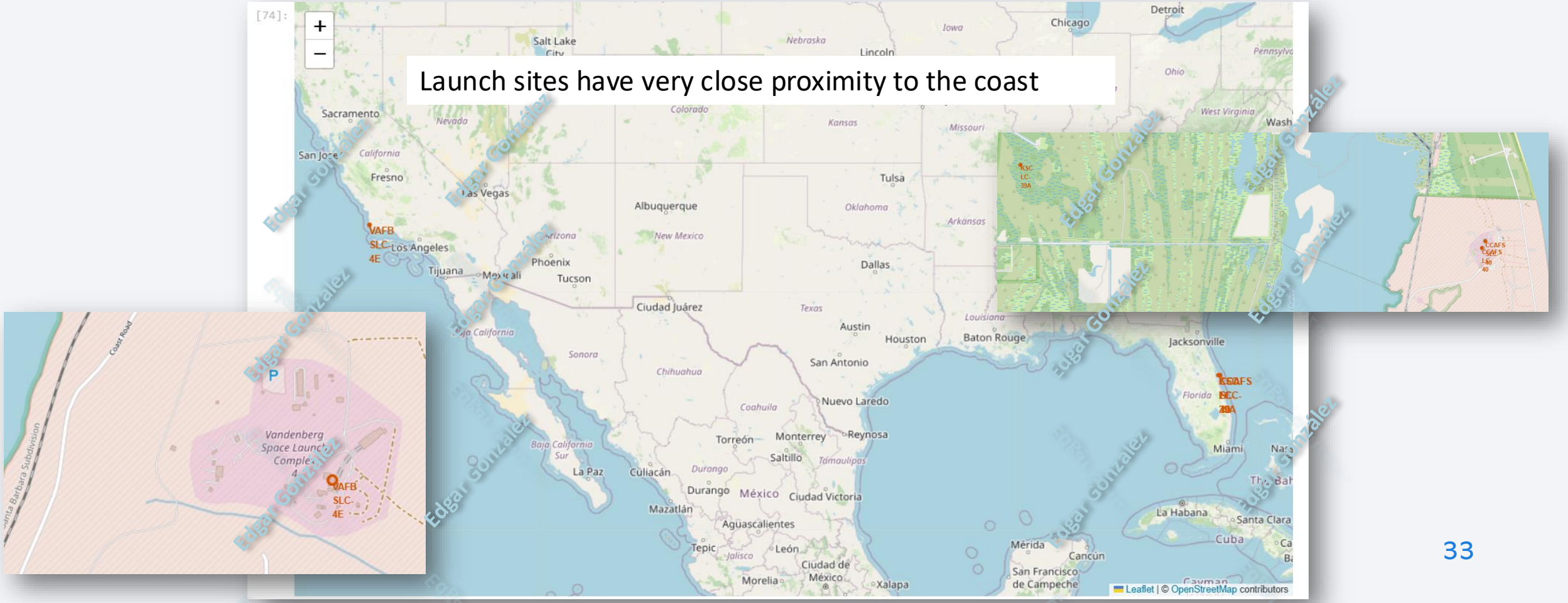
Landing_Outcome	Total
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

SpaceX Launch sites across United States

- Replace <Folium map screenshot 1> title with an appropriate title
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot



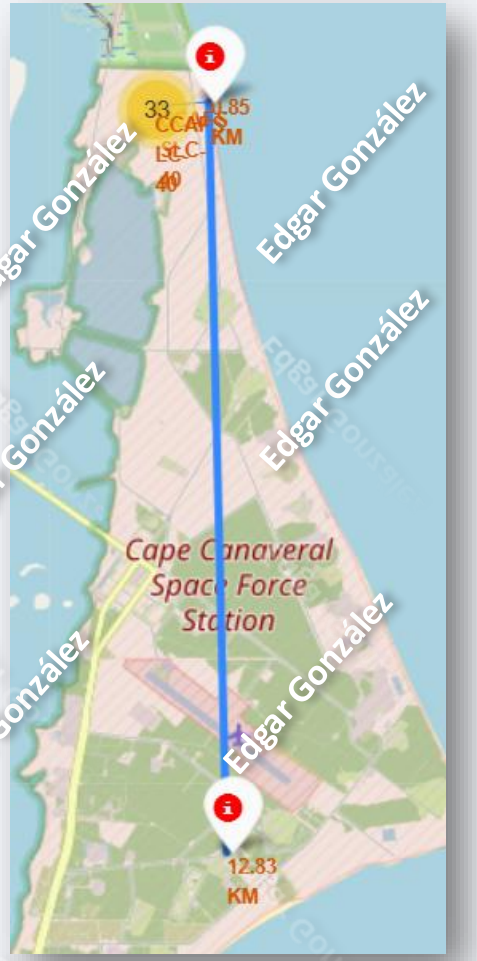
Launch Site Outcomes

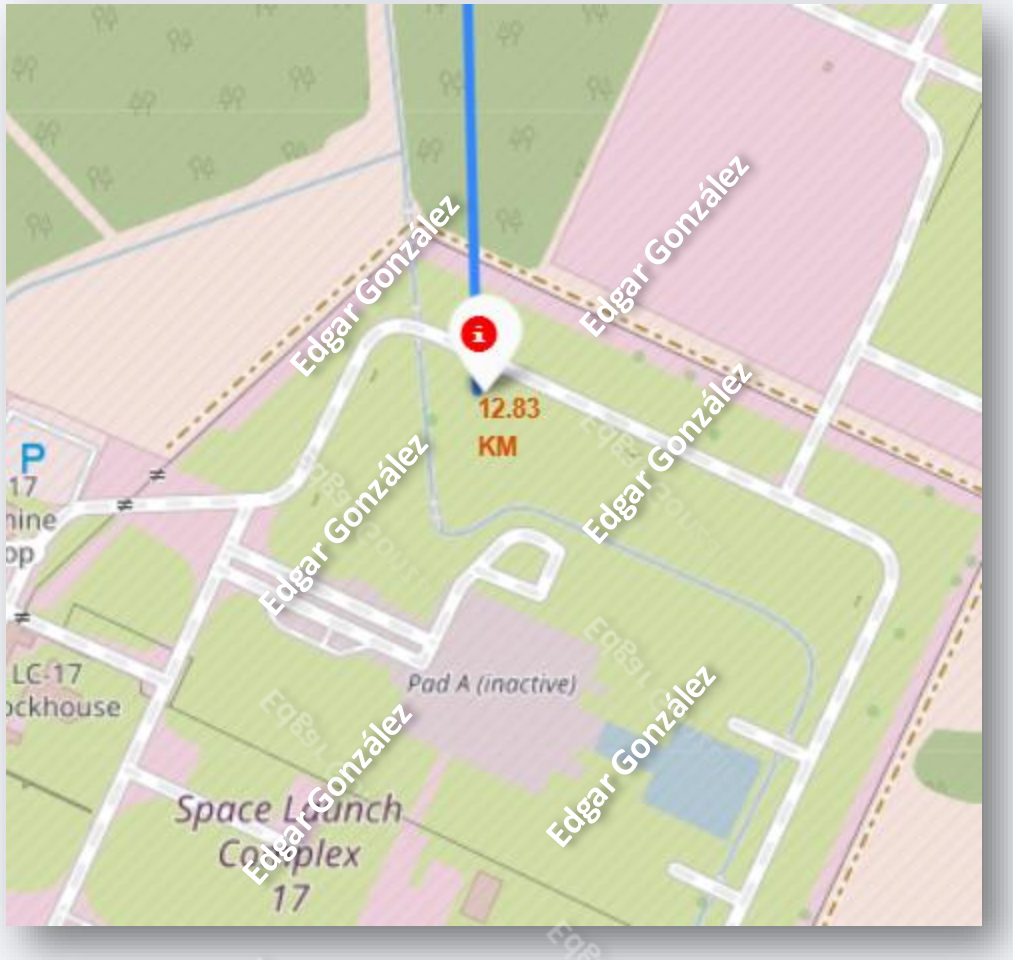
- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

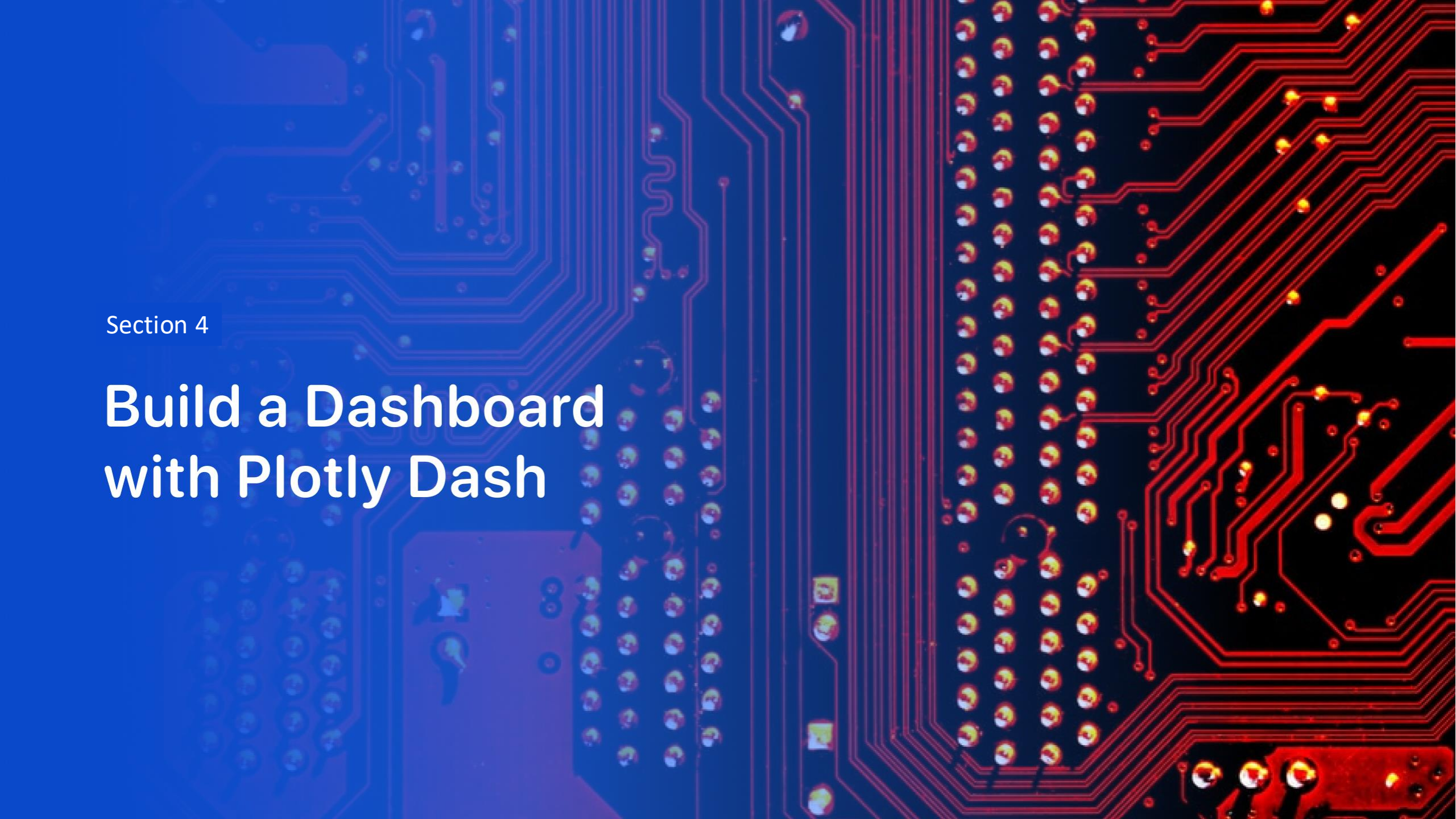
We can observe that the KSC LC-39 has success rate better than other sites and the CCA FS-SLC40 has the most launches

Distance from active Launch site to Pad inactive

- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot

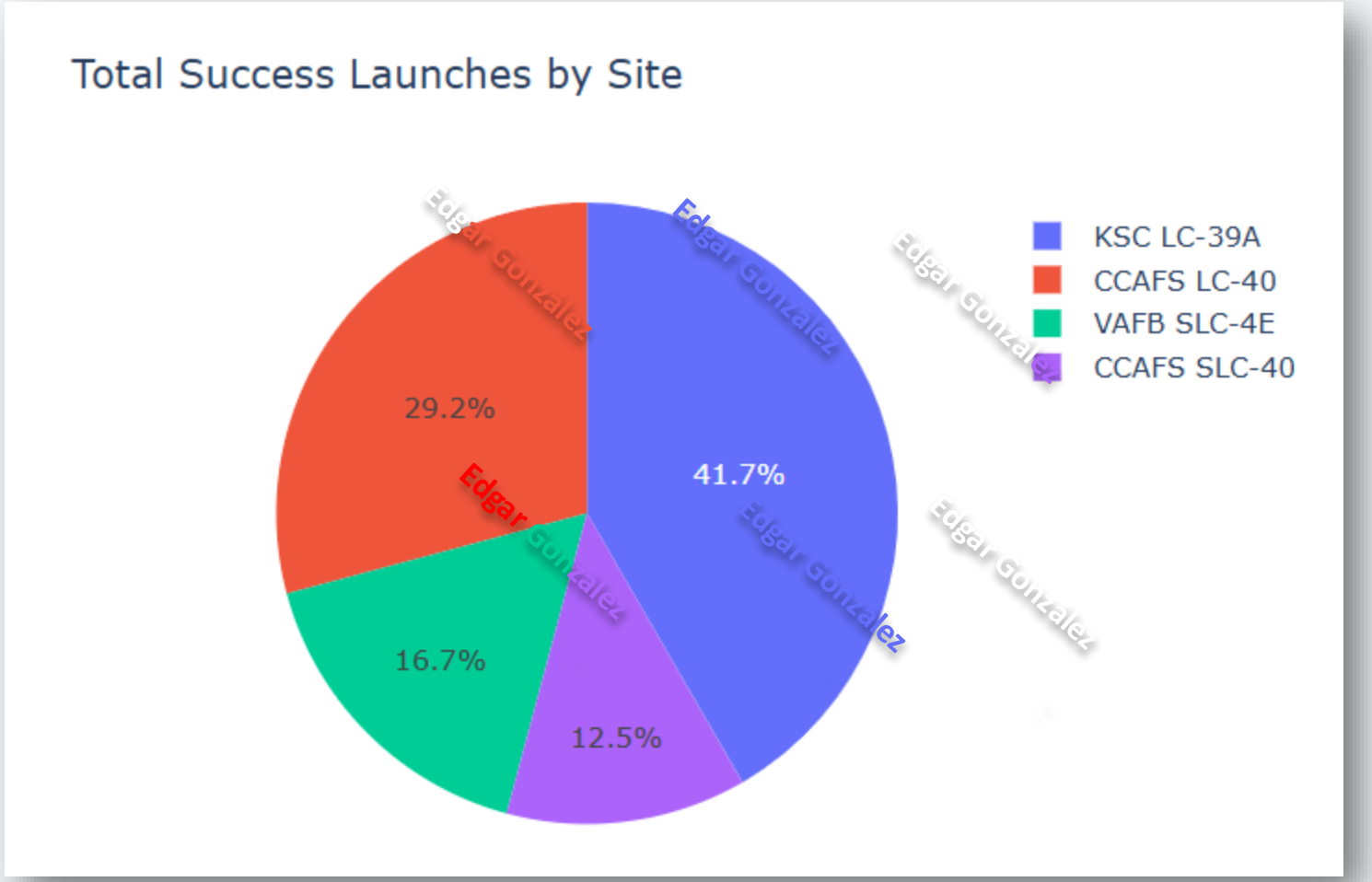





Section 4
Build a Dashboard with Plotly Dash

SpaceX - Success Launches by site

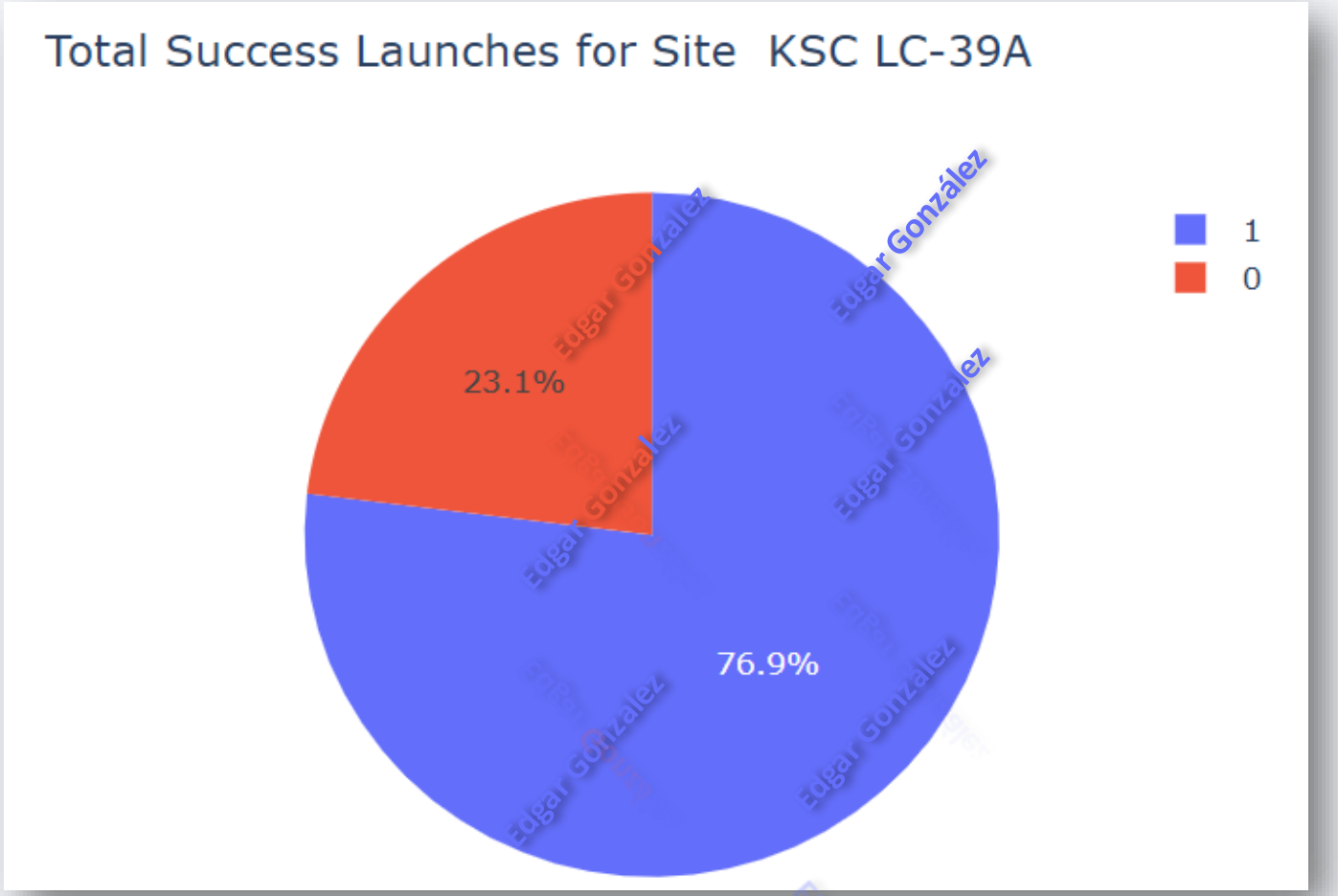
- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot



The Site KSC LC-39A has the mayor percentage of success rate, comparing VAFB vs CCAFS it have almost the same success rate.

Highest Rate - KSC LC-39A

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

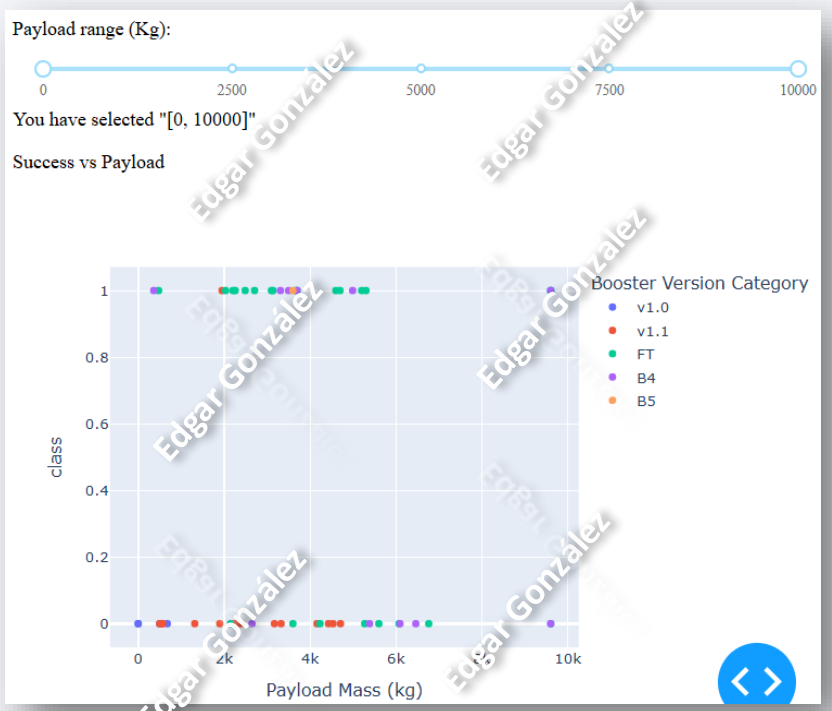


This launch site has a 23.1% of nonsuccess launches versus 76.9% of success in the launches

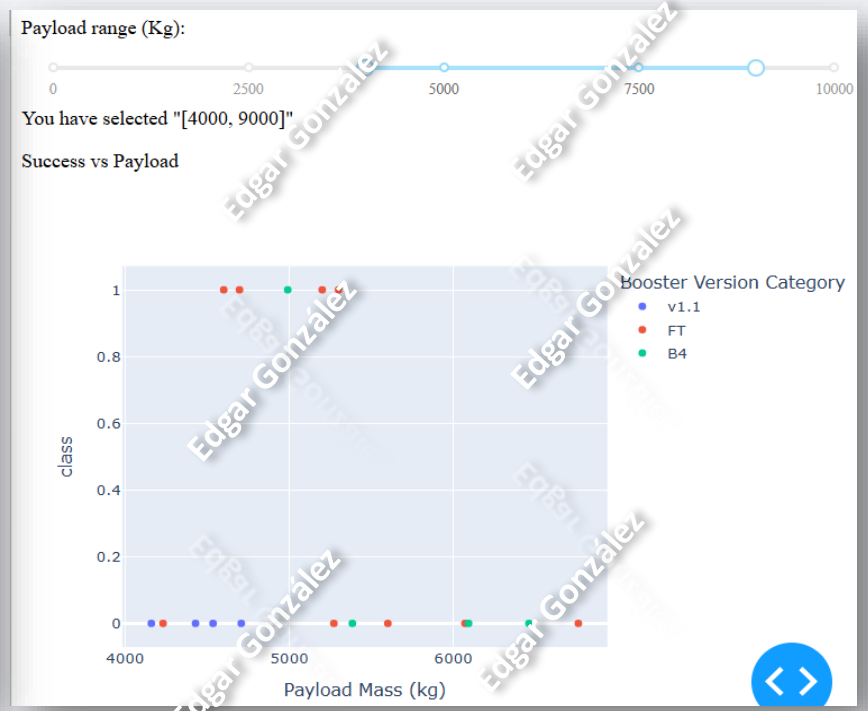
Booster Version Outcome

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Whole payload range 0 to 10000 kg



From 4000 in ahead just three models are above this capacity



From 6000 in ahead just two models of boosters are above this capacity FT and B4

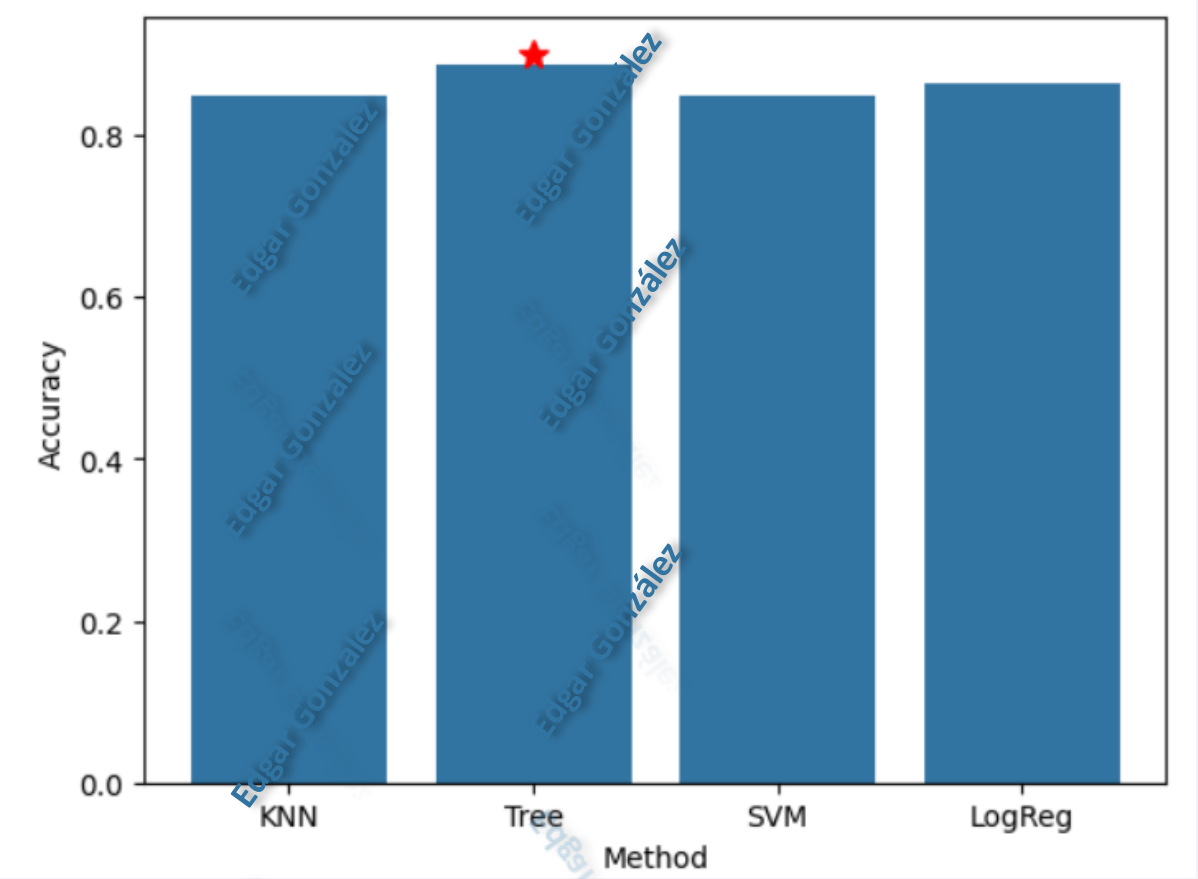
Section 5

Predictive Analysis (Classification)

Classification Accuracy

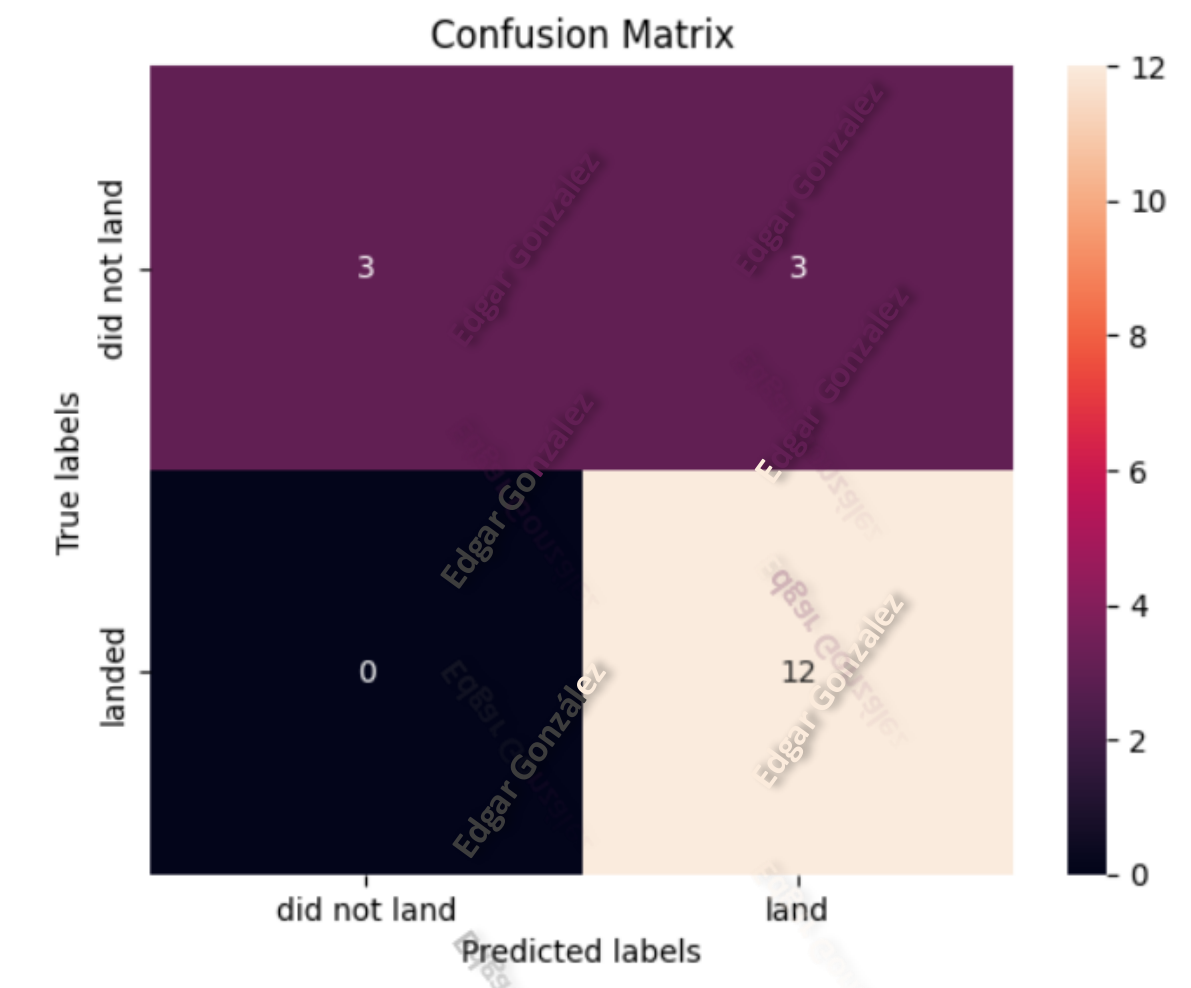
- Visualize the built model accuracy for all built classification models, in a bar chart
- Find which model has the highest classification accuracy

Method Decision Tree Classifier



Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation



True Positive lower right corner- 12 (True label is land, Predicted label is also land)

True negative upper left corner – 3 (True label is did not land and predicted is did not land)

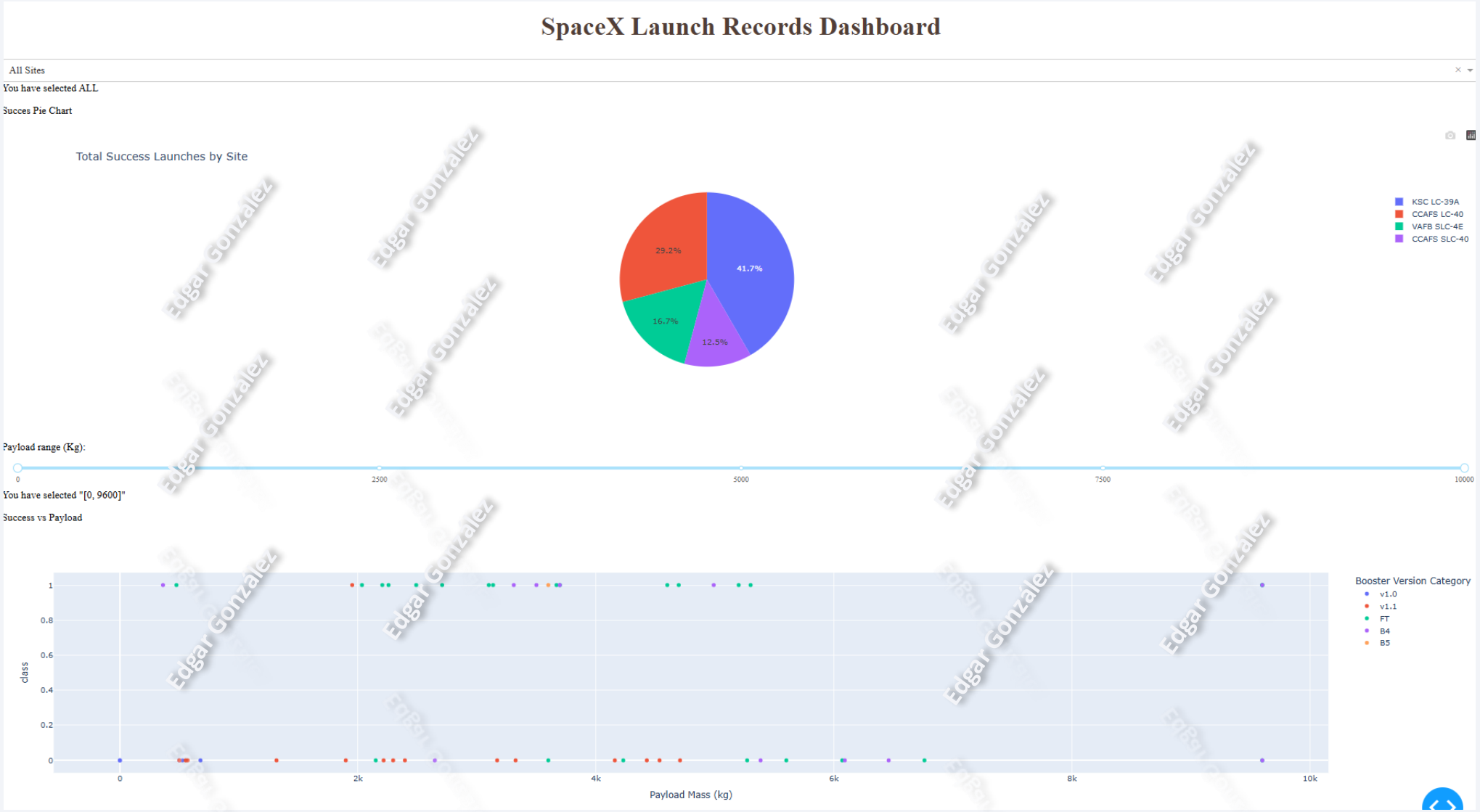
False Positive upper right corner- 3 (True label is not landed, Predicted label is landed)

High valuss in the True Positive and True Negative cells is better

Conclusions

- The KSC LC-39 is the site launch with the major success rate than other sites.
- The CCA FS-SLC40 has the most launches.
- The Orbits with major success rate are ES-L1, GEO, HEO, SSO.
- Just two booster could load above 6000kg FT and B4.
- KSC LC-39 launch site has a 23.1% of nonsuccess launches versus 76.9% of success in the launches.
- Launch sites have very close proximity to the coast.
- Since 2013 in ahead there was a increased success rate.
- The VAFB-SLC launchsite there are no rockets launched for heavy payload mass the maximum is 10000.
- For the predictive analysis the model with the best accuracy is Method Decision Tree Classifier and it has a accuracy above 0.8 as the other predictive analysis.

Appendix


SpaceX Launch Records Dashboard
All Sites
You have selected ALL
Succes Pie Chart
Total Success Launches by Site
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
0
2500
5000
7500
10000
You have selected "[0, 9600]"
Success vs Payload
class
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5

Thank you!