Winning Space Race with Data Science

<Elton Trindade>
<October 18th>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context
- Problems you want to find answers

Introduction

- Project background and context

  Space X is a company that aims to popularize space tourism and operations, with measures to lower the high costs inherent of a launch. However, one of the main aspects that lower the costs are the reuse of the first stage of the launching equipment, whereas in other launches this equipment cannot be recovered and must be built from scratch. Since the landing is not always successful, a feasibility study is advisable to measure if this reuse approach is indeed more profitable. This project conducts experiments to ponder the launches, predict the expected mean outcome and whether it is profitable on average.

- Problems you want to find answers

  - Which variables are relevant to the landing outcome (successful or failed) and which ones are more meaningful;

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collectes from SpaceX database, using their API
- Perform data wrangling
  - Labels where binary assigned based on the outcome of the landing (bad and successful)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Several models were trained on the same dataset (after appropriate data filtering, cleaning and feature engineering) and compared based on the appropriate metrics)

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

- 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

Place your flowchart of SpaceX API calls here

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

Place your flowchart of web scraping here

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

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

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

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

Build a Dashboard with Plotly Dash

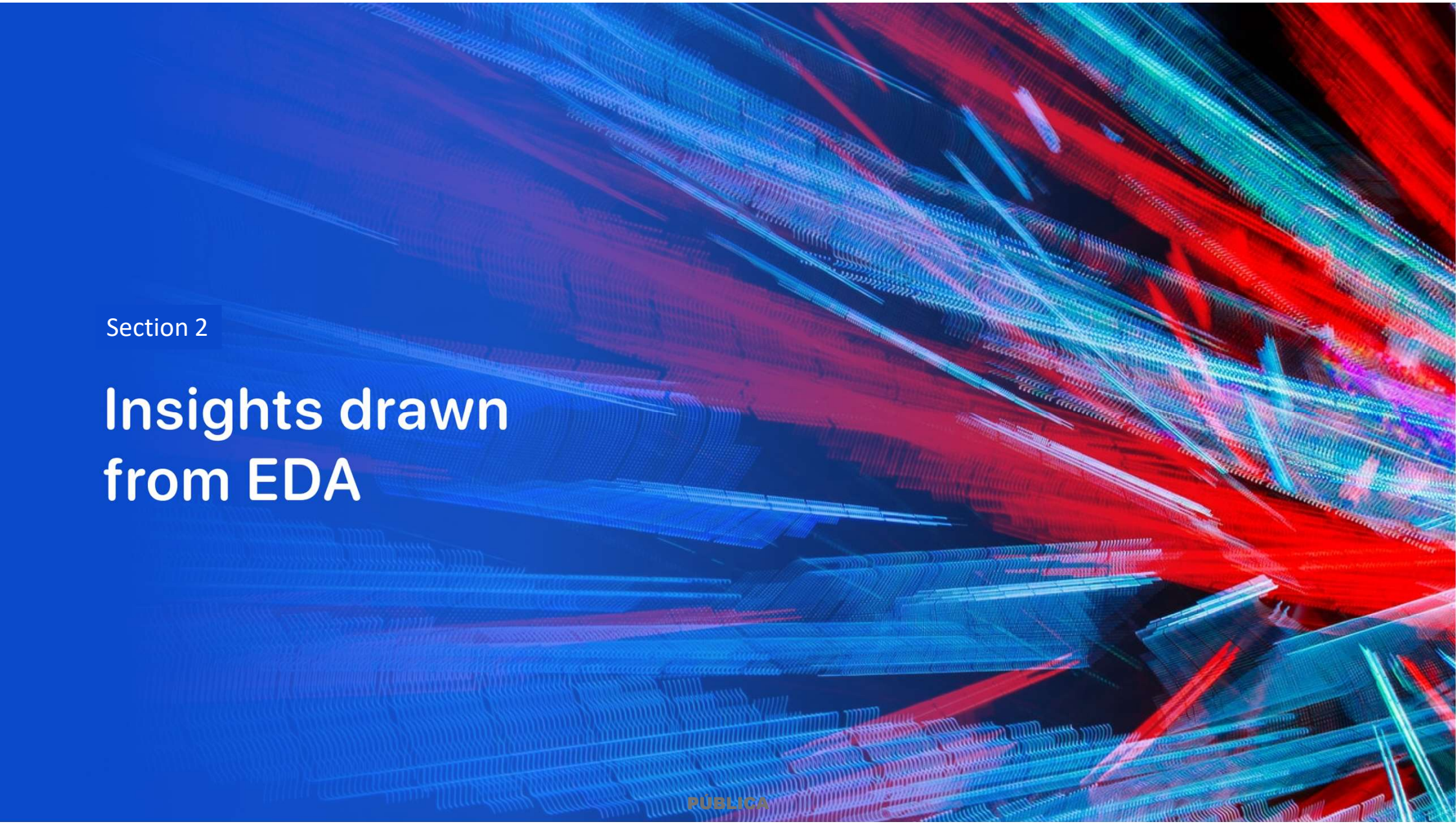
- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

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

Results

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

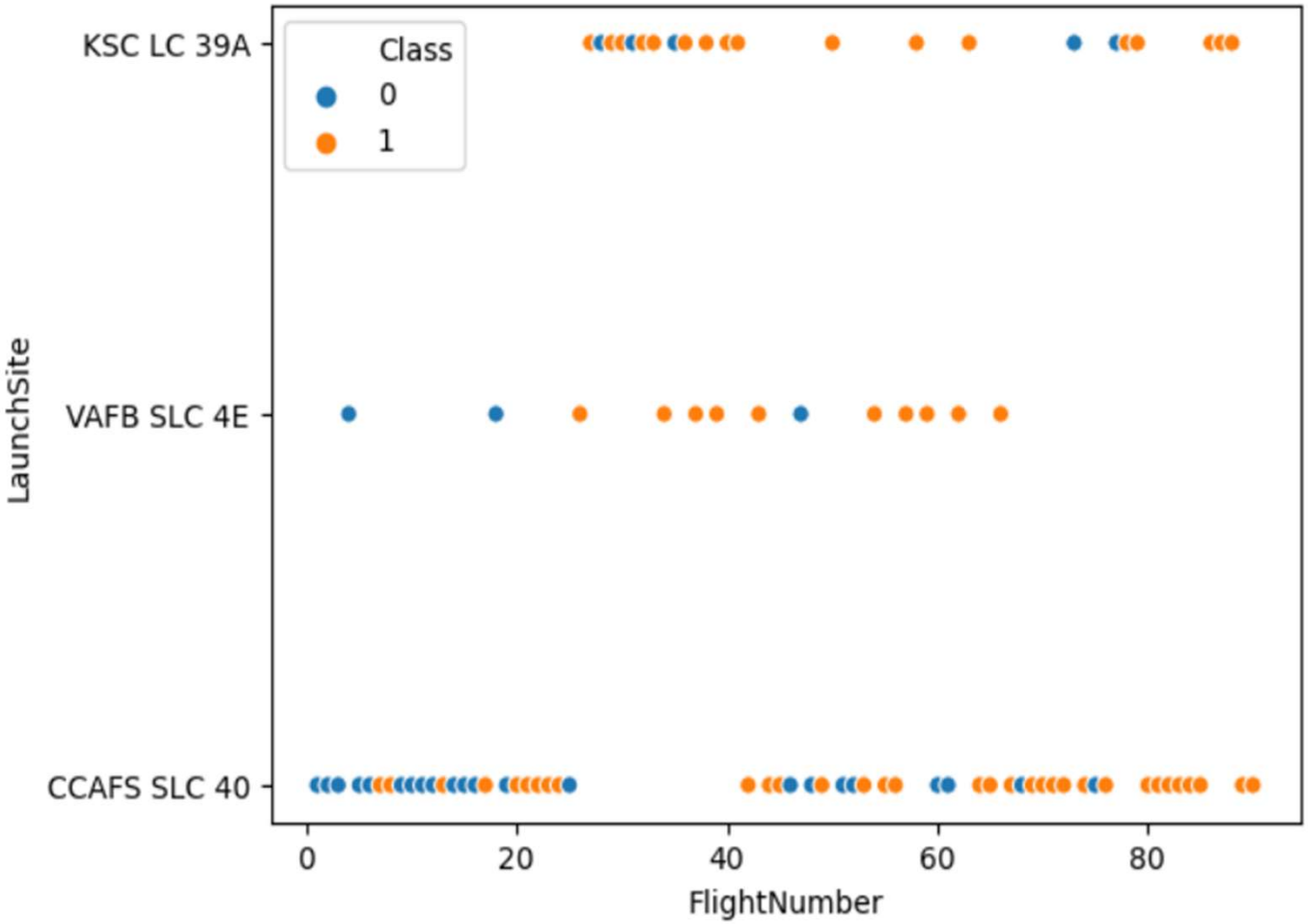

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site

- Show the screenshot of the scatter plot with explanations

Flight Number vs. Launch Site



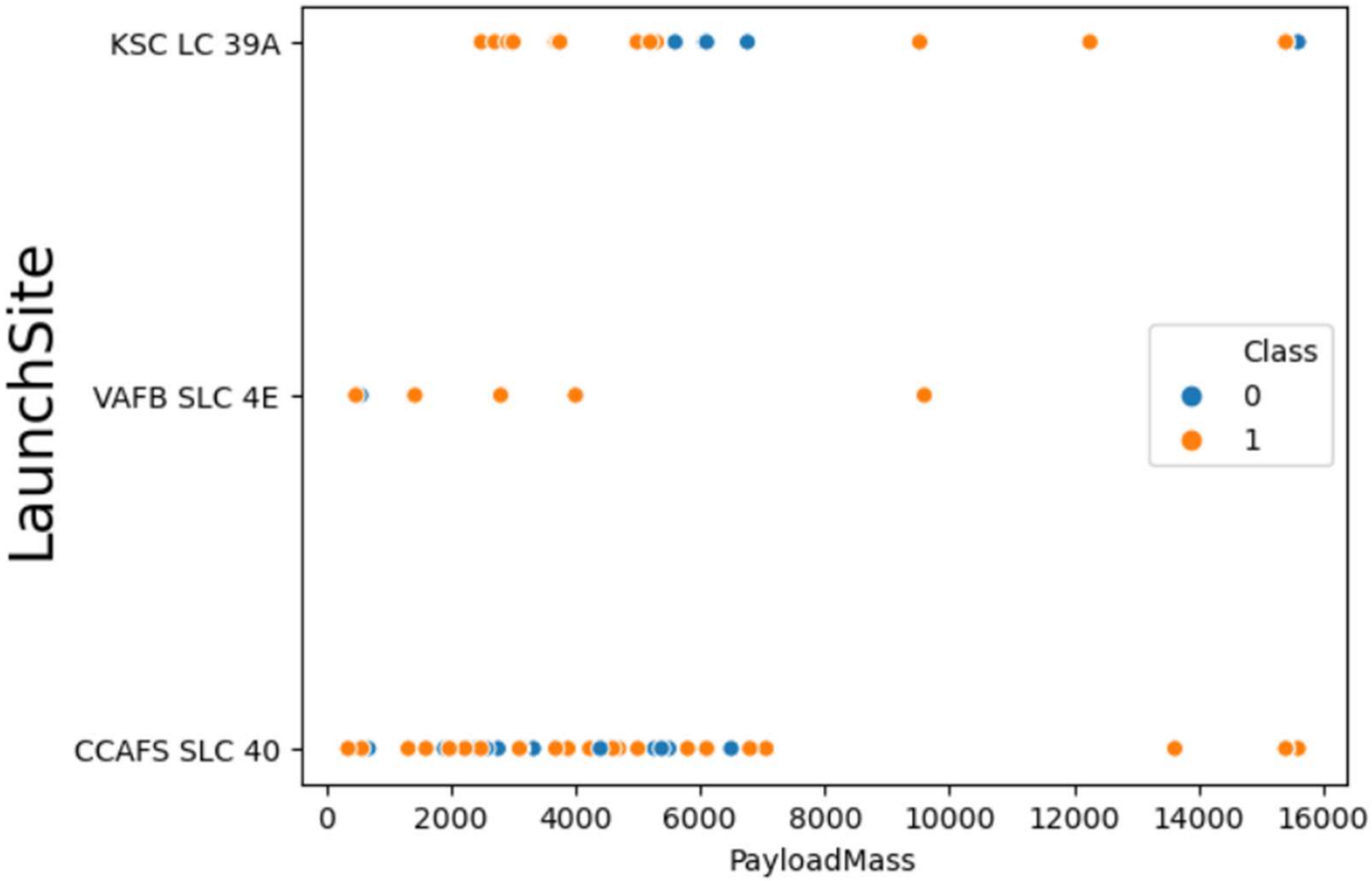
- Show a scatter plot of Flight Number vs. Launch Site
- Show the screenshot of the scatter plot with explanations: launchSites generally have a higher success rate with larger FlightNumber. LAunchSite “CCAFS SLC 40” has a much lower success rate with lower ‘FlightNumber’

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

- Show the screenshot of the scatter plot with explanations

Payload vs. Launch Site



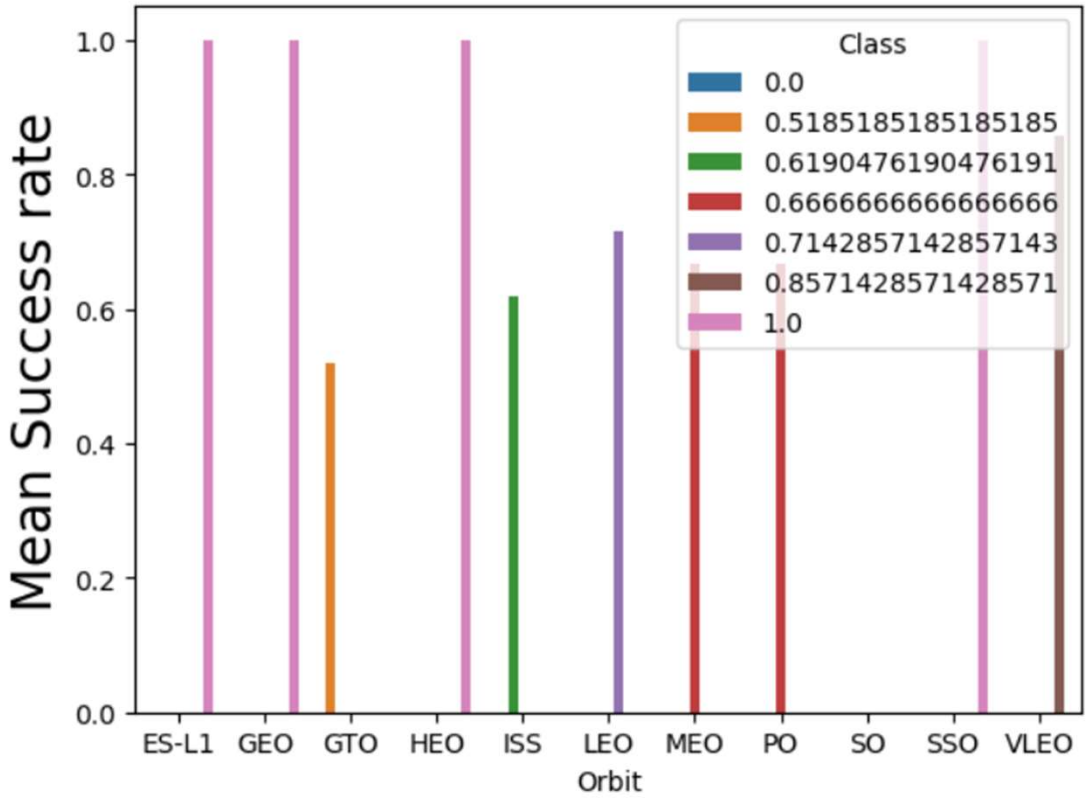
- Show a scatter plot of Payload vs. Launch Site
- Show the screenshot of the scatter plot with explanations: for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000).

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

Success Rate vs. Orbit Type



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

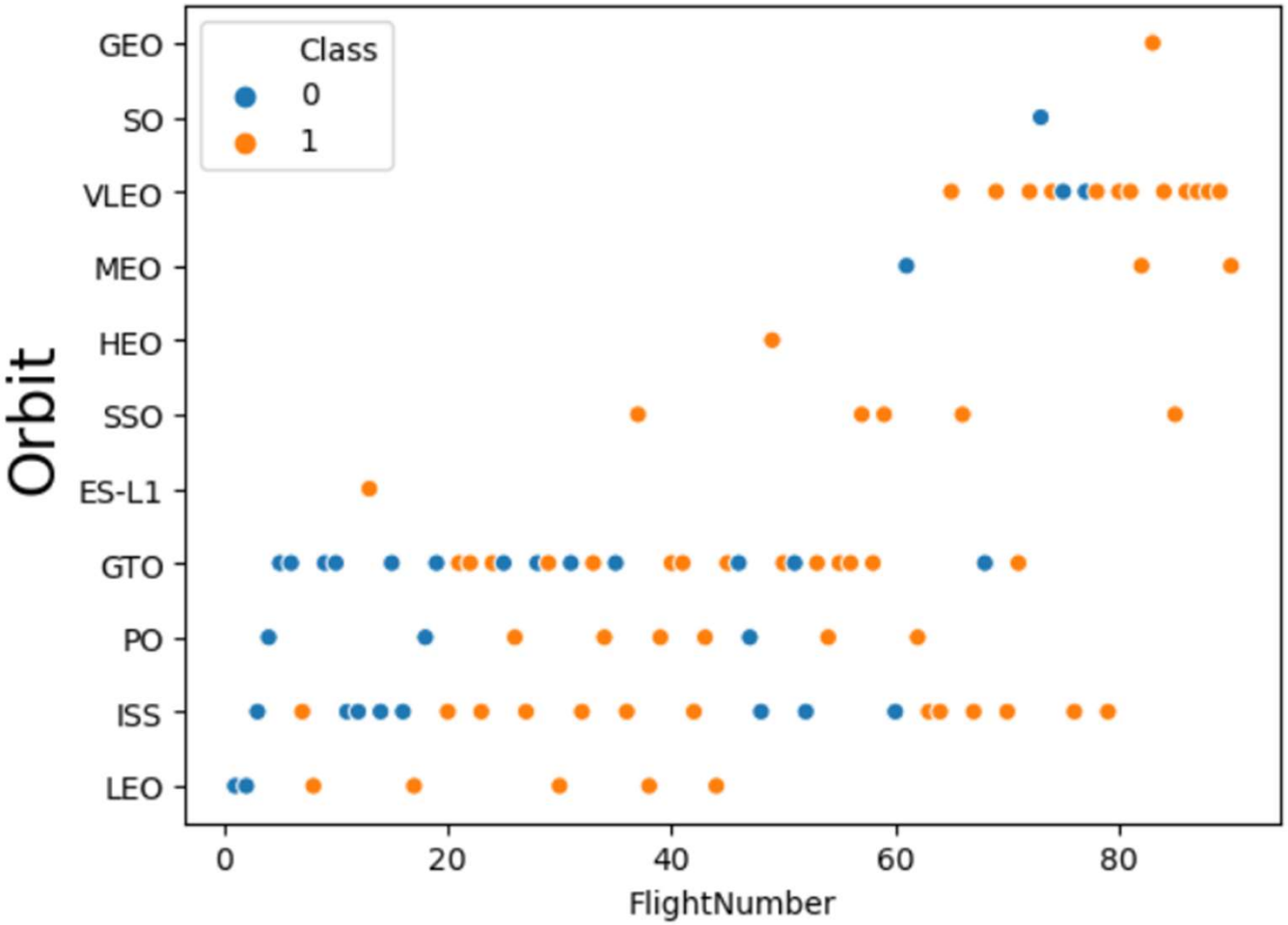
- Orbits “ES-L1’, ‘GEO’, ‘HEO’, ‘SSO’ have very high success rate, whereas ‘SO’ has the worst.

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

- Show the screenshot of the scatter plot with explanations

Flight Number vs. Orbit Type



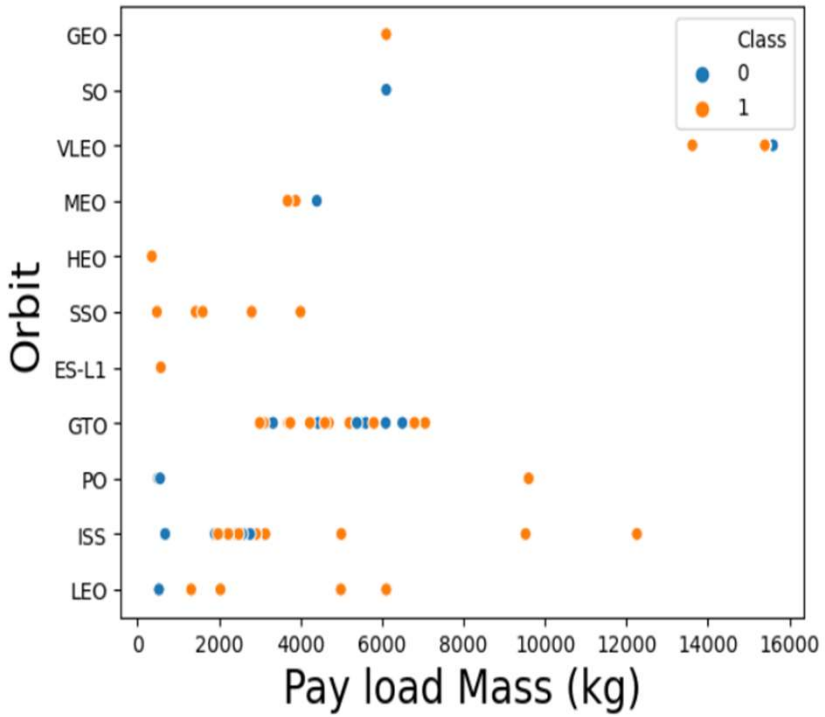
- LEO orbit Success has strong correlation with Flight Number;
- The aforementioned 4 orbits have 100% success rate (GEO, HEO, ES-LI, SSO)

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

- Show the screenshot of the scatter plot with explanations

Payload vs. Orbit Type



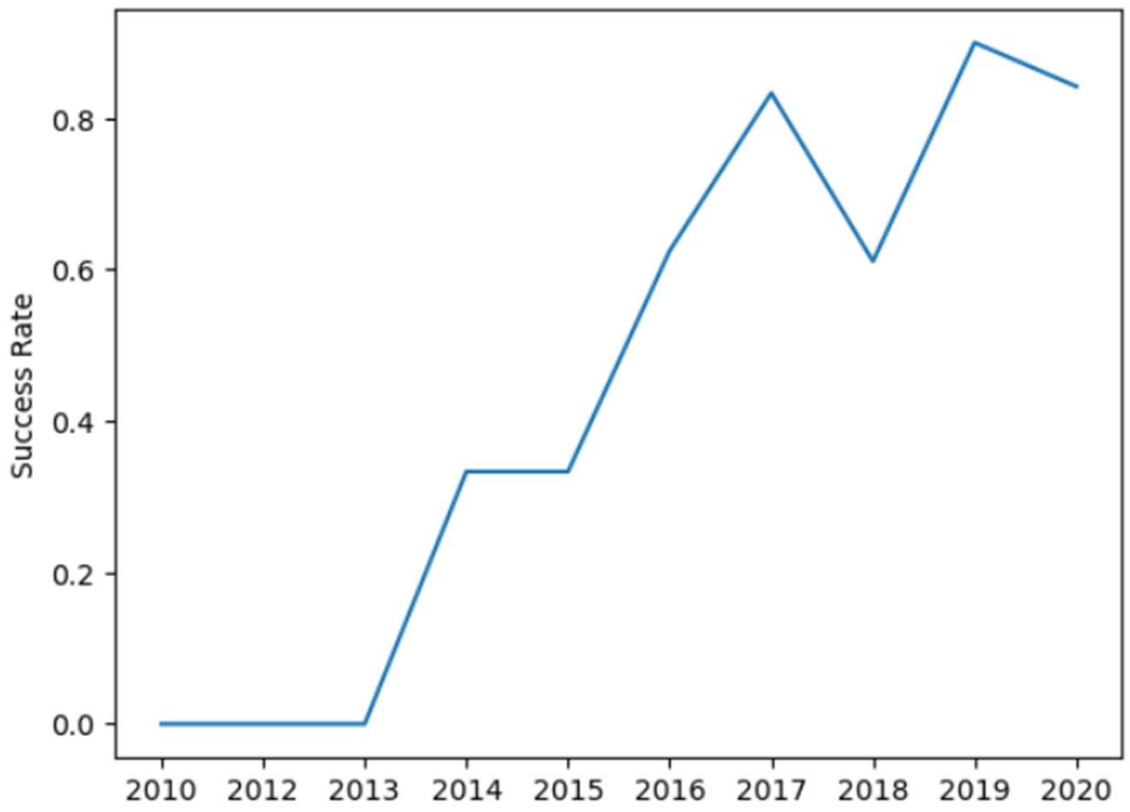
- Show a scatter point of payload vs. orbit type
- Show the screenshot of the scatter plot with explanations

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

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



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

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

There are only 4 distinct LaunchSites in the dataset

```
[5]: df['LaunchSite'].unique()

[5]: array(['CCAFS SLC 40', 'VAFB SLC 4E', 'KSC LC 39A'], dtype=object)
```

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here:
- The code makes a search for the string ‘CCA’ in the column ‘Launch Site’, limiting to the first 5 occurrences

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [6]:
```
%sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Done.

Out[6]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here:
- **The query searches for the key 'NASA (CRS)' in the column "COSTUMER", and sums the column 'PAYLOAD_MASS_KG' from the rows**

Display the total payload mass carried by boosters launched by NASA (CRS)

In [12]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

Out[12]:

total_payload_mass
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

Display average payload mass carried by booster version F9 v1.1

In [13]:
```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

Out[13]:

	avg_payloadmass
0	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

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [15]:
```
%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTBL \
    WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

Out[15]:

first_successful_ground_landing
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

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [36]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL \
    WHERE (LANDING__OUTCOME = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

Out[36]:

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
- Present your query result with a short explanation here

List the total number of successful and failure mission outcomes

In [49]:
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

Out[49]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
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

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [17]:
```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

Out[17]:

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

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 expression “Where” e “AND” results in a query with an inner join (intersection) in the database

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [58]:
```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL \
    WHERE (LANDING__OUTCOME = 'Failure (drone ship)') AND (EXTRACT(YEAR FROM DATE) = '2015');
```

Done.

Out[58]:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	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

- Present your query result with a short explanation here

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

In [66]:
```
%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL \
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
    GROUP BY LANDING__OUTCOME \
    ORDER BY TOTAL_NUMBER DESC;
```

Done.

Out[66]:

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

Section 3

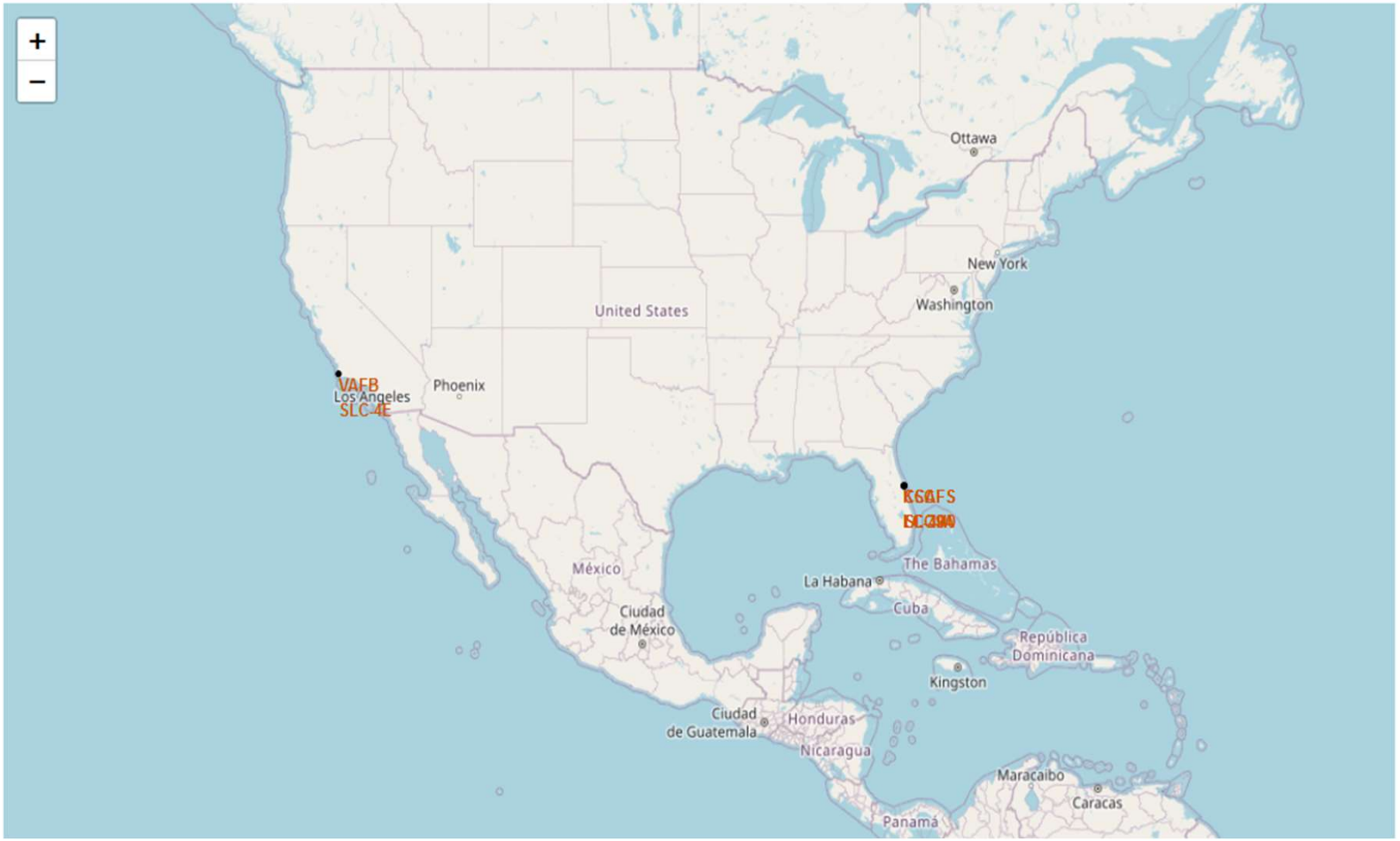
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- 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 SITES

[154]:



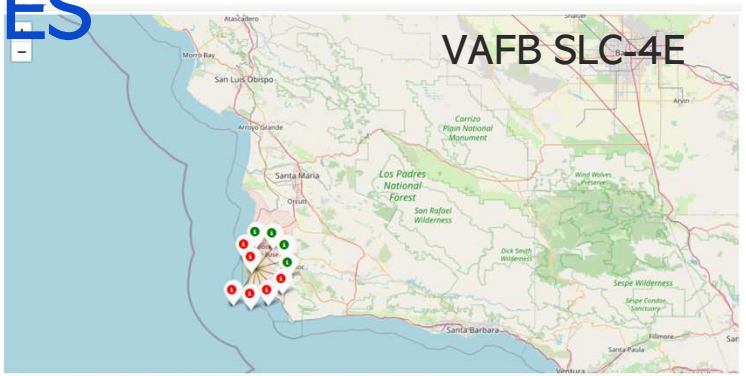
- The launchsites are located as far south as possible, since it's more feasible (and cheaper) to launch from lower latitudes (closer distance).Only one of them is not located in Florida (VAFB SLC-4E). The other 3 sites are so close to each other that are overlayed by each other in the map.

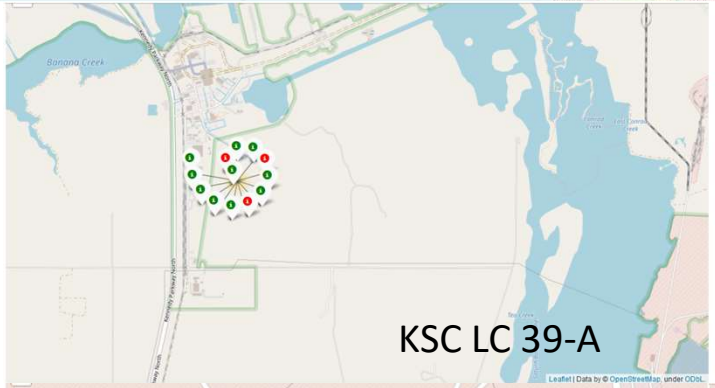
<Folium Map Screenshot 2>

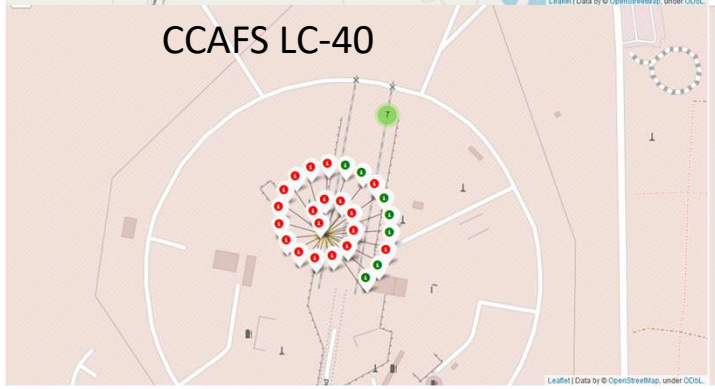
- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

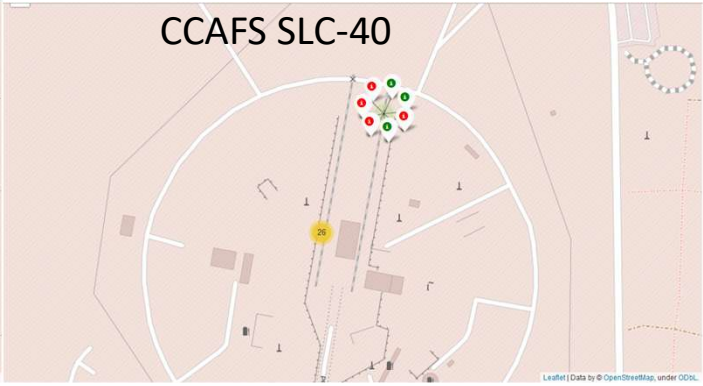
LAUNCH OUTCOMES

- Only launchSite KSC LC 39-A has a success rate above 50%, even though it is quite close to 2 other LaunchSites (CCAFS LC-40 and CCAFS SLC-40)





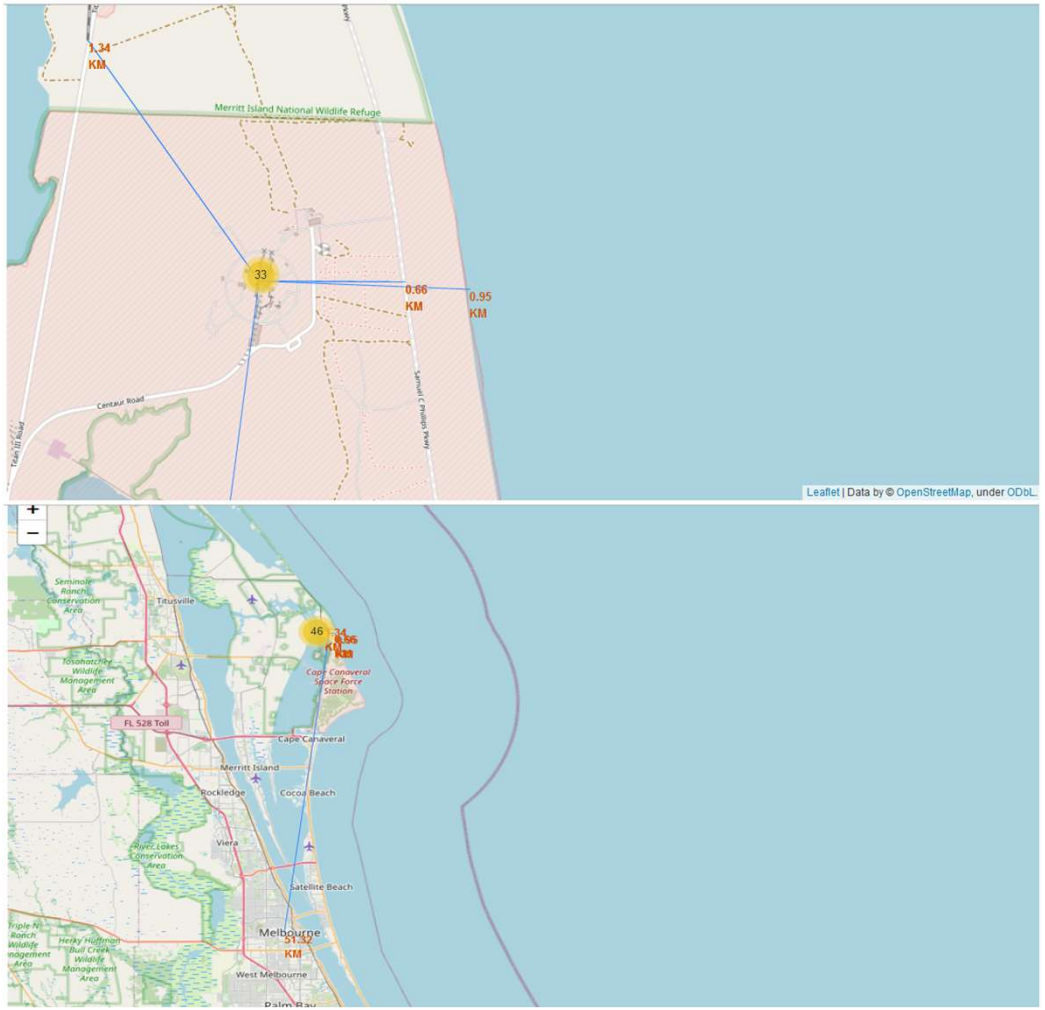




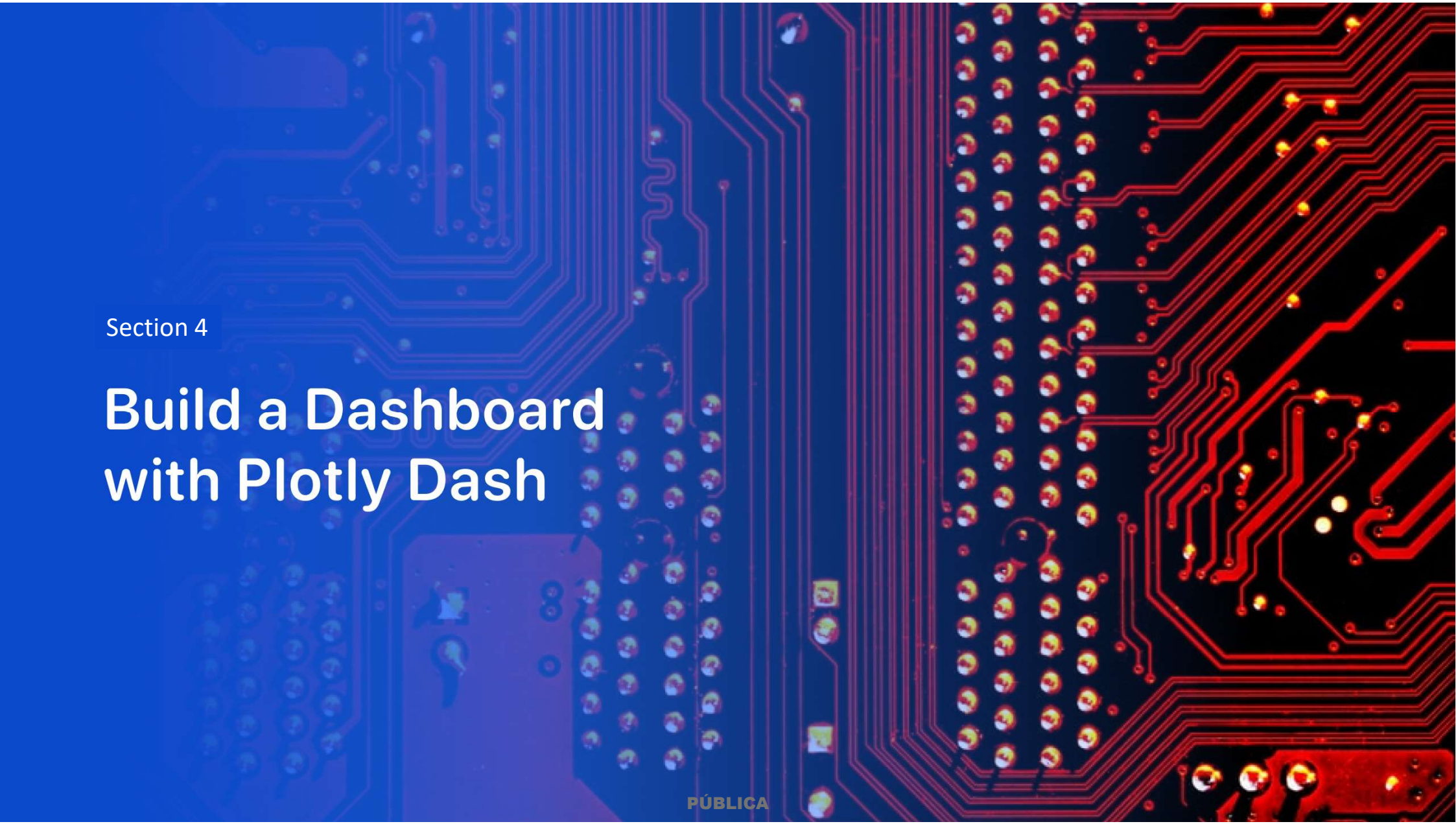
<Folium Map Screenshot 3>

- 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

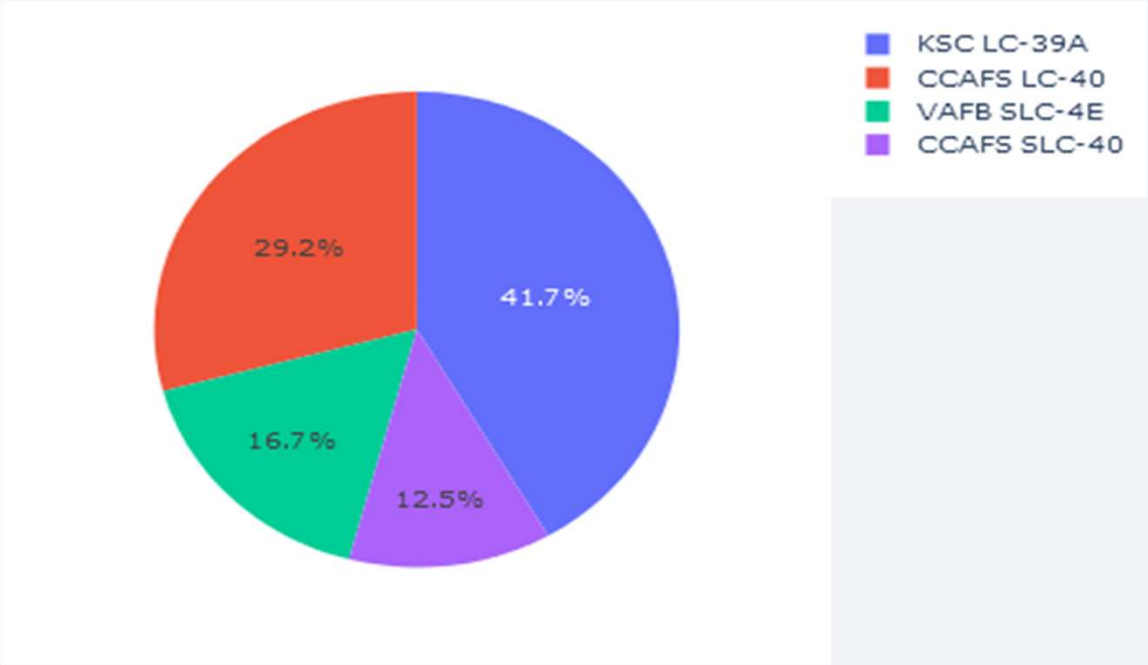
LAUNCH SITES PROXIMITIES



- Roads and railways are less than a mile away from the launchsite, which makes sense for logistic porpuses (transportation of heavy and technological equipments);
- The closest city is located over 50 km from the launchsite, decreasing the risk of any hazard to civilians during the operation;



Success Rate for all sites

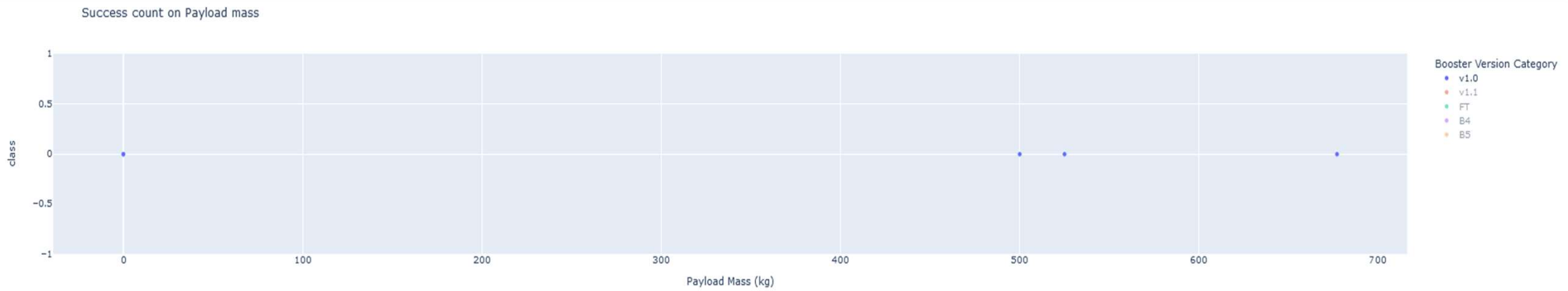


- The launchSite “KSC LC-39A” is the most successful one, responsible for almost half of the successful launches;
- The launches in florida are reportedly more successful than the ones in California;
- Although really close to each other, the launchSite CCAFS LC-40 is twice more successful than CCAFS SLC-40;

Success Rate of the Best Launch Site

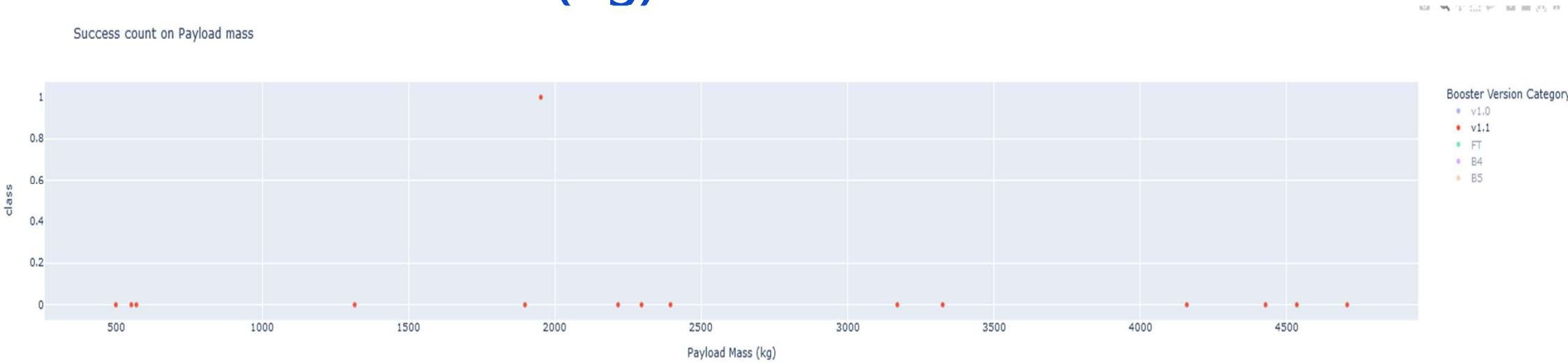
- Launch Site “KSC LC-39A” has a great success rate (above 76%);

PAYLOAD MASS (Kg) X LAUNCH OUTCOME



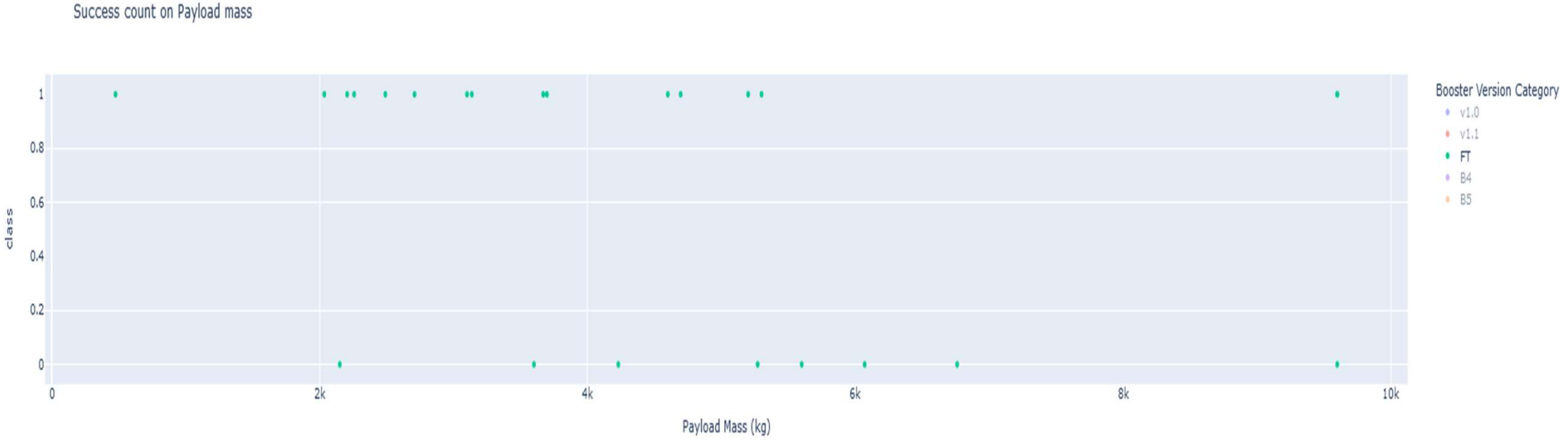
Booster v1.0 Has a very high failure rate, with failures in every single launch.

PAYLOAD MASS (Kg) X LAUNCH OUTCOME



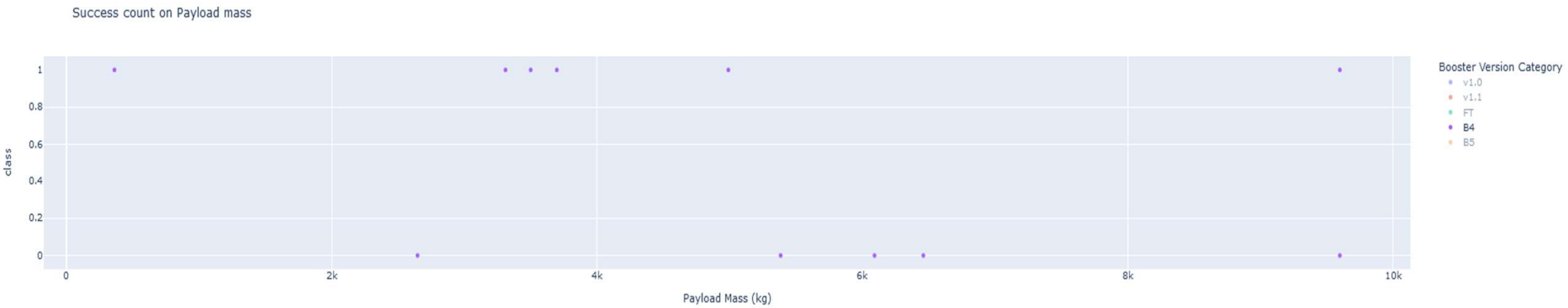
- Booster v1.1 (probably a proposed update from the previous one) did not fare much better than v1.0. Apparently increasing the payloadMsss did not change de outcome(failure), with one success at around 200kg of Payload Mass

PAYLOAD MASS (Kg) X LAUNCH OUTCOME



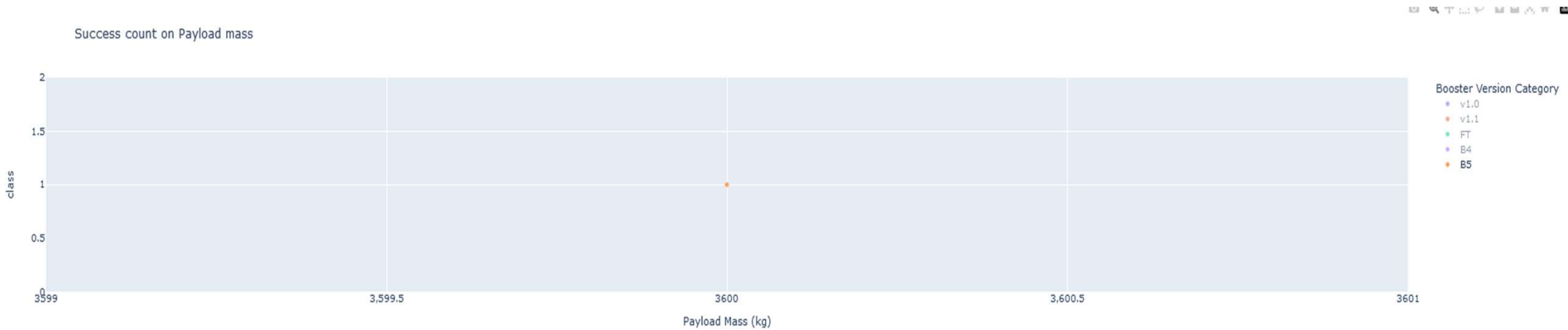
- Booster FT has the highest success rate in the data, with an optimal range of PayloadMass between 0 and 5000 kg ;

PAYLOAD MASS (Kg) X LAUNCH OUTCOME



- Booster B4 has a success rate above 50%, although with a smaller sampe size compared to the aforementioned Booster FT, and has the same pattern of higher success in payload mass inferior to 5000 kg.

PAYLOAD MASS (Kg) X LAUNCH OUTCOME



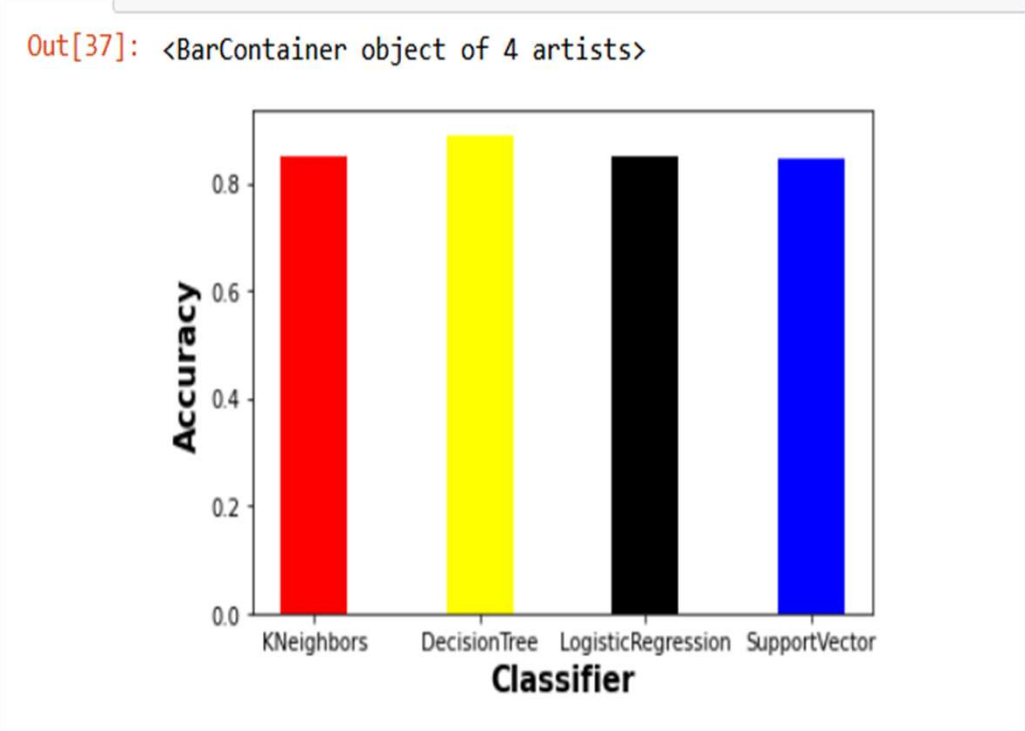
- Although mathematically booster B5 has the highest success rate (100%), it does not possess enough statistical relevance to be considered the best booster, since only one launch is registered;

Section 5

Predictive Analysis (Classification)

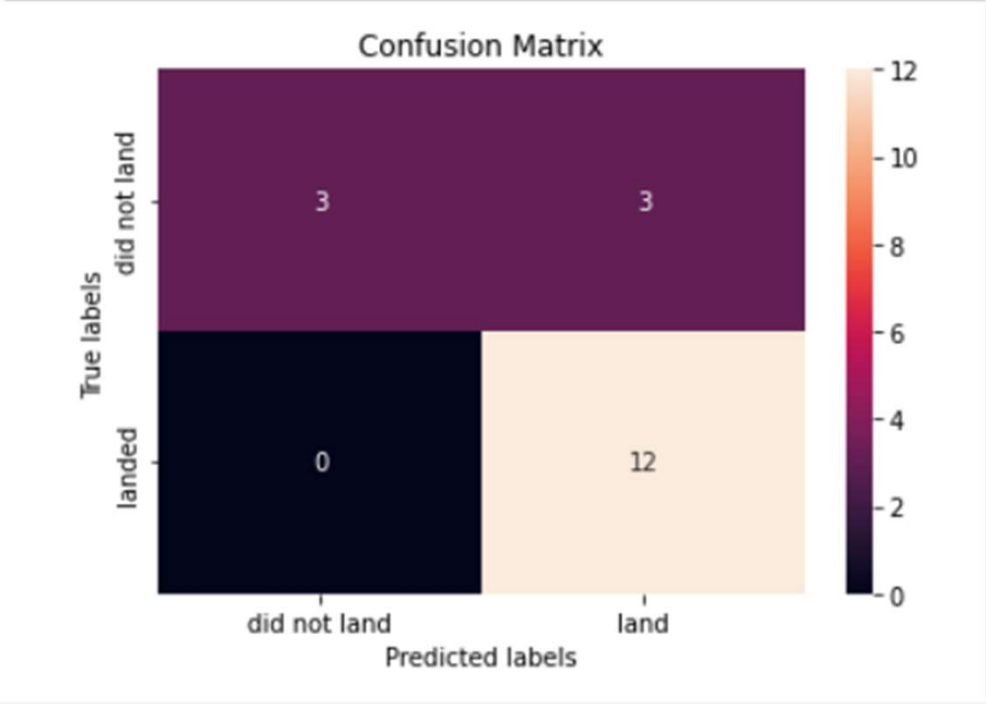
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- Find which model has the highest classification accuracy
- **The decision Tree Algorithm has the best accuracy of the 4 models in the experiment**



Confusion Matrix

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



The model has an accuracy of ~84% and has a better recall than precision, since it did not mistake a failed landing with a true one. Perhaps the data is more unbalanced towards succesfull landings.

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!