IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Akua K. Owusu
April 10, 2024

Outline

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

Executive Summary

- Summary of methodologies
- Data Collection via API, Web Scraping
- Exploratory Data Analysis(EDA) with Data Visualization
- EDA with SQL
- Interactive map with Folium
- Dashboards with Plotly Dash
- Predictive Analysis

- Summary of all results
- Exploratory Data Analysis
- Interactive maps and dashboard
- Predictive results

Introduction

- Project background and context

The aim of this project is to predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches 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. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

- **Problems you want to find answers**
- Will the Falcon 9 first stage land successfully?
- What factors are behind the failure of landing?

Section 1

Methodology

Methodology

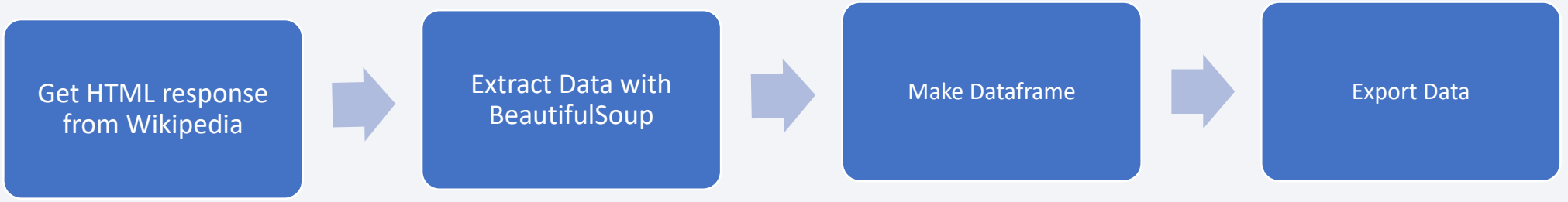
Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping from Wikipedia.
- Perform data wrangling
  - Data cleaning of null values and irrelevant columns
  - One hot encoding for classification models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - SVM, Classification Trees and Logistics Regression models were built and evaluated for the best classifier.

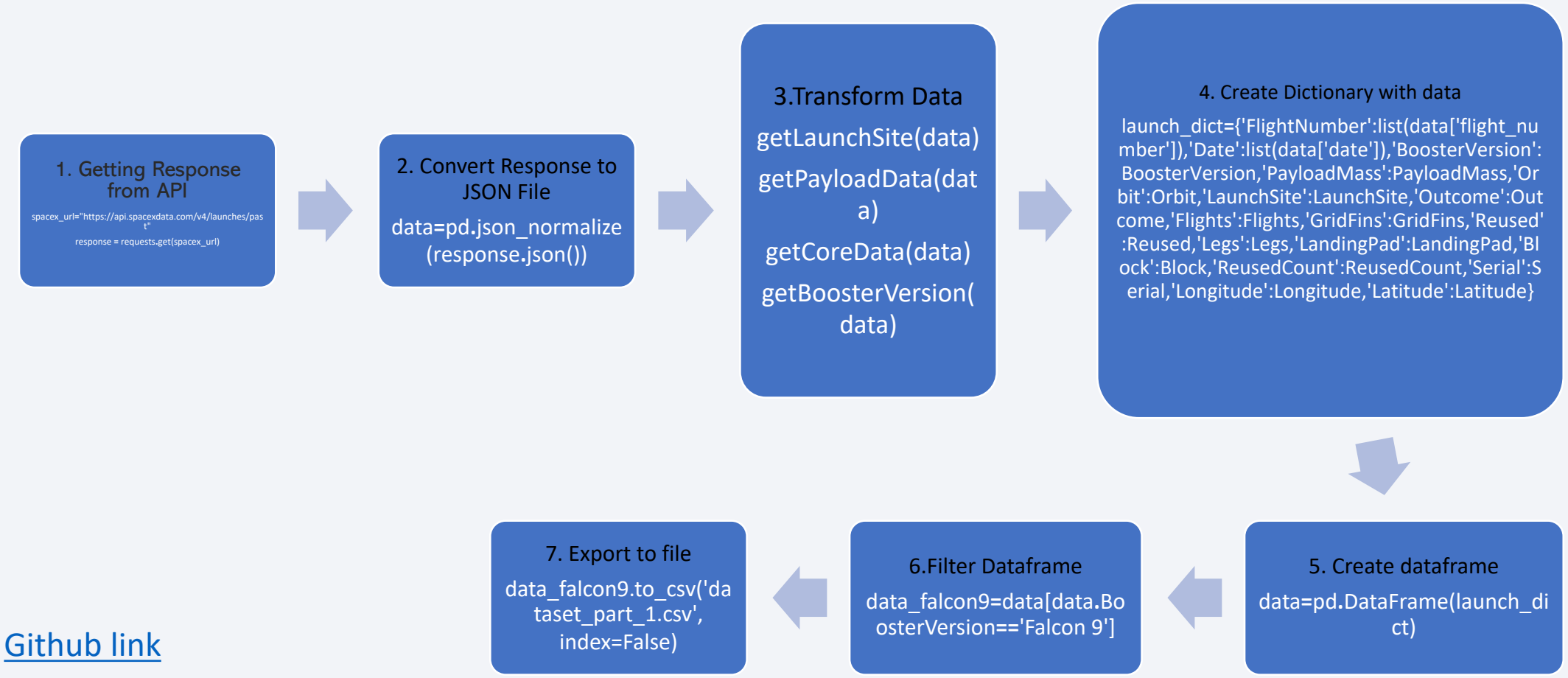
Data Collection

- Datasets are collected from Rest SpaceX API and webscrapping Wikipedia
  - The information obtained from the API are rocket, launches, payload information

  - The information obtained by webscrapping of Wikipedia are launches, landing and payload information.



Data Collection – SpaceX API



[Github link]

Data Collection - Scraping

Github link

1.Getting Response from HTML

```
r=requests.get(static_url)
```

2.Create BeautifulSoup Object

```
soup=BeautifulSoup(data,"html.parser")
```

3.Find all Tables

```
html_tables=soup.find_all('table')
```

4.Get Column names

```
forj,table_headerinenumerate(table_headers):name=extract_column_from_header(table_header)ifnameisnotNoneandlen(name)>0:column_names.append(name)
```

5.Create dictionary

```
# Let's initial the launch_dict with each value to be an empty listlaunch_dict['Flight No.']=[]launch_dict['Launch site']=[]launch_dict['Payload']=[]launch_dict['Payload mass']=[]launch_dict['Orbit']=[]launch_dict['Customer']=[]launch_dict['Launch outcome']=[]
```

6. Add data to keys

```
extracted_row=0#Extract each table
fortable_number,tableinenumerate(soup.find_all('table',"wikitable plainrowheaders collapsible")):# get table row
forrowsintable.find_all("tr"):
```

7. Create dataframe from dictionary

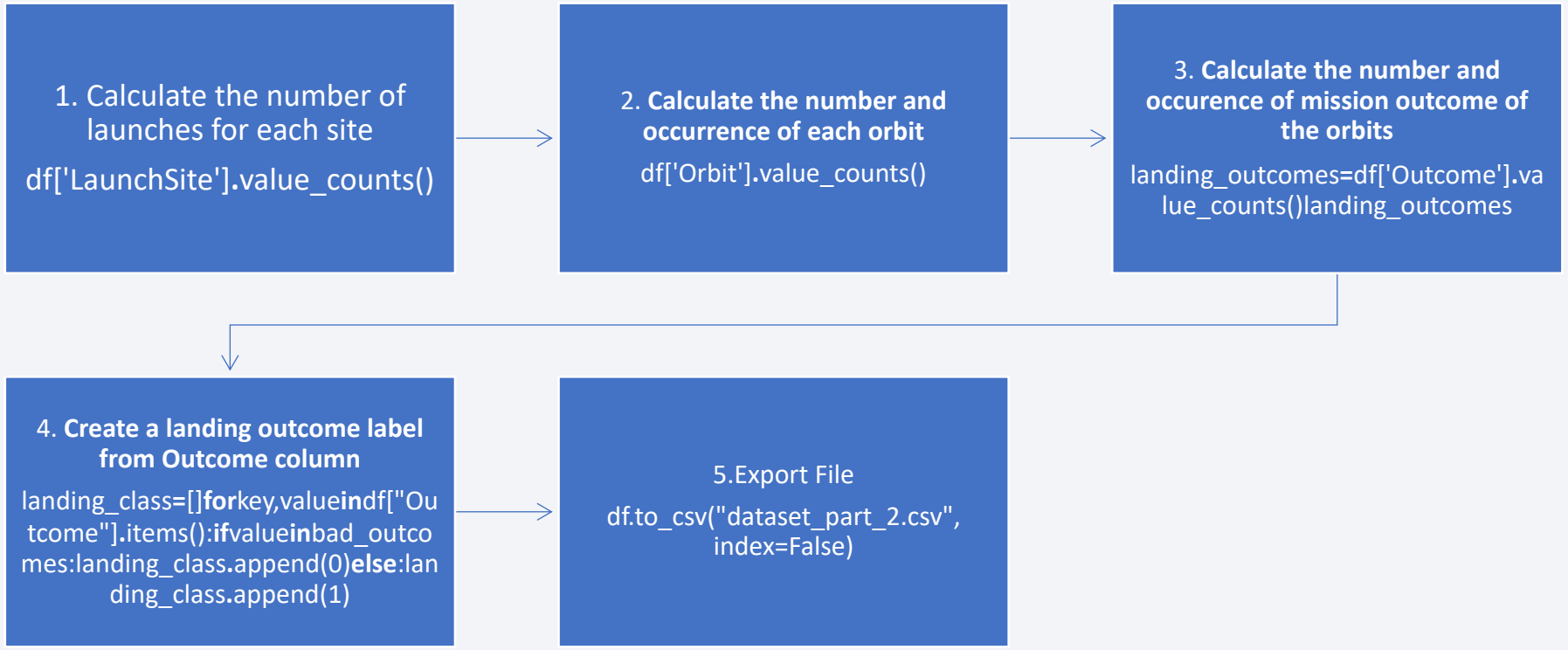
```
df=pd.DataFrame({key:pd.Series(value)forkey,valueinlaunch_dict.items()})
```

8. Export file

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

Data Wrangling

- In the dataset, we had some cases where the booster did not land successfully
  - True Ocean, True RTLS, True ASDS means the mission has been successful
  - False Ocean, False RTLs, False ASDS means the mission was a failure
  - We need to transform string variables into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure.



Github

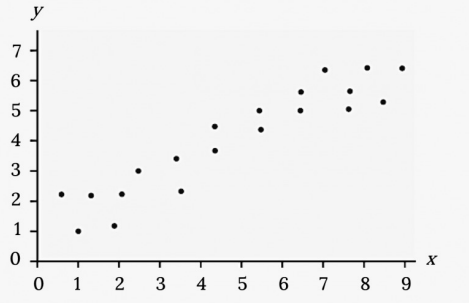
EDA with Data Visualization.

Github link

Scatter Plots

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload vs Launch Site
- Orbit vs Flight Number
- Payload vs Orbit type
- Orbit vs Payload Mass.

Scatter plots show relationship between variables.The relationship is called correlation



Bar Graph

- Success rate vs Orbit

Bar graphs show the relationship between numeric and categoric variables.

Line graph

- Success rate vs year

Line graphs show data variables and their trend.Line graphs can help to show global behavior and make prediction for unseen data.

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the 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 succesful landing outcome in ground pad was acheived.
  - List the names of the 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. Use a subquery
  - List the 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/akuakwartemaa/Data-Science-final-Capstone/blob/main/EDA SQL.ipynb

Build an Interactive Map with Folium

Github link

The Folium map object is a map centered on NASA Johnson Space Center at Houston, Texas:

- There's a red circle at NASA Johnson Space Center's coordinate with a label showing its name (utilizes folium.Circle, folium.map.Marker).
- Red circles placed at each launch site coordinates with a label displaying the launch site name (uses folium.Circle, folium.map.Marker, folium.features.DivIcon).
- Points are grouped in a cluster to exhibit multiple and different information for the same coordinates (implemented via folium.plugins.MarkerCluster).
- Markers are used to indicate successful and unsuccessful landings. Green markers denote successful landing and red markers for unsuccessful landing (employs folium.map.Marker, folium.Icon).
- Markers also show the distance between launch site to key locations like railway, highway, coastway, and city, and lines are plotted between them (uses folium.map.Marker, folium.PolyLine, folium.features.DivIcon).
- These objects are created to better understand the problem and the data. They help to easily display all launch sites, their surroundings, and the count of successful and unsuccessful landings.

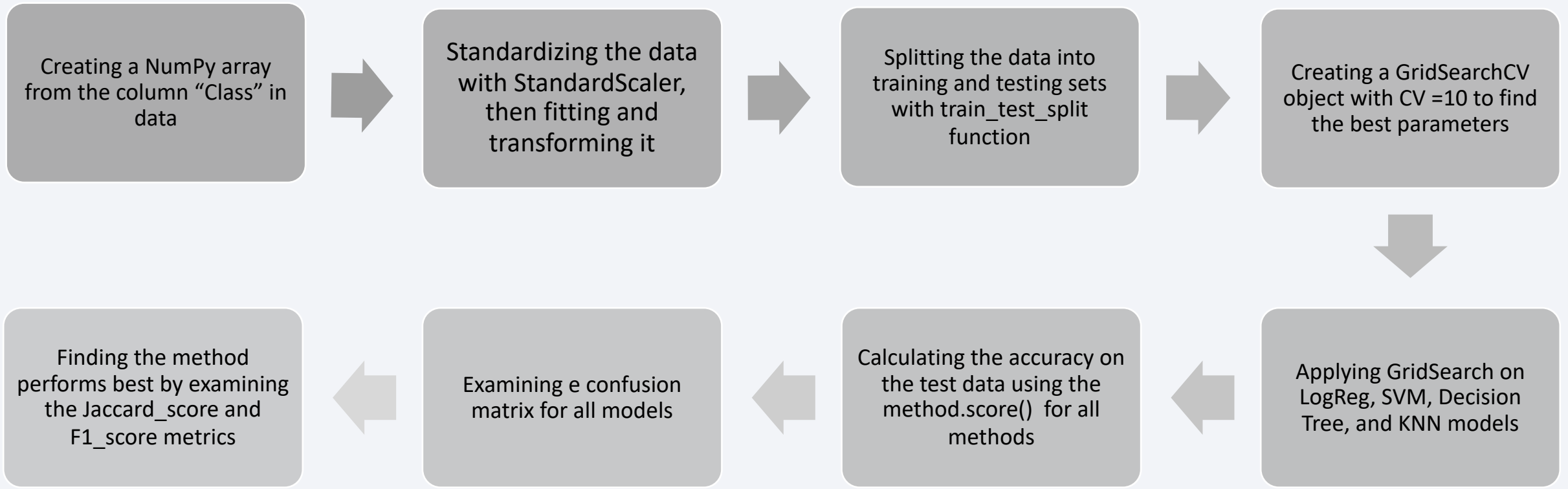
Build a Dashboard with Plotly Dash

Dashboard has dropdown, pie chart, rangeslider and scatter plot components

- Dropdown allows a user to choose the launch site or all launch sites (dash_core_components.Dropdown).
- Pie chart shows the total success and the total failure for the launch site chosen with the dropdown component (plotly.express.pie).
- Rangeslider allows a user to select a payload mass in a fixed range (dash_core_components.RangeSlider).
- Scatter chart shows the relationship between two variables, in particular Success vs Payload Mass (plotly.express.scatter)

Link to code

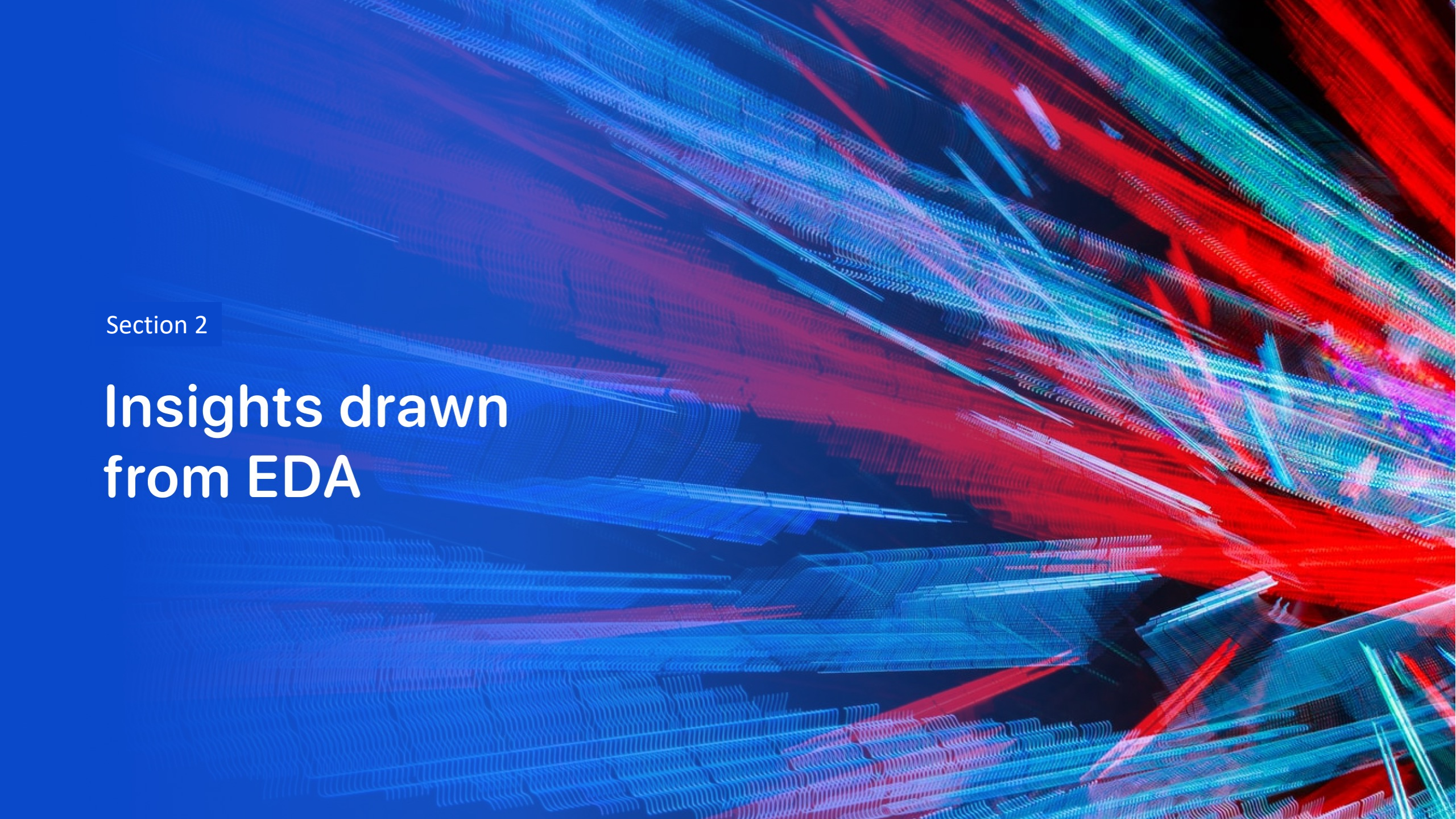
Predictive Analysis (Classification)



Github link

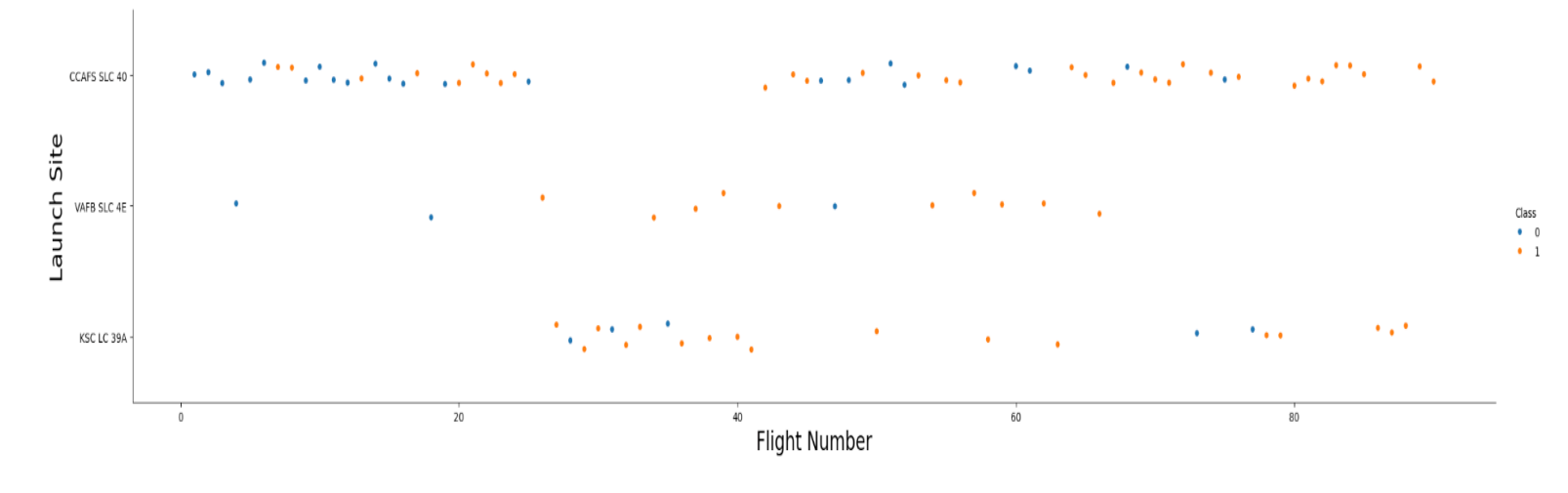
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

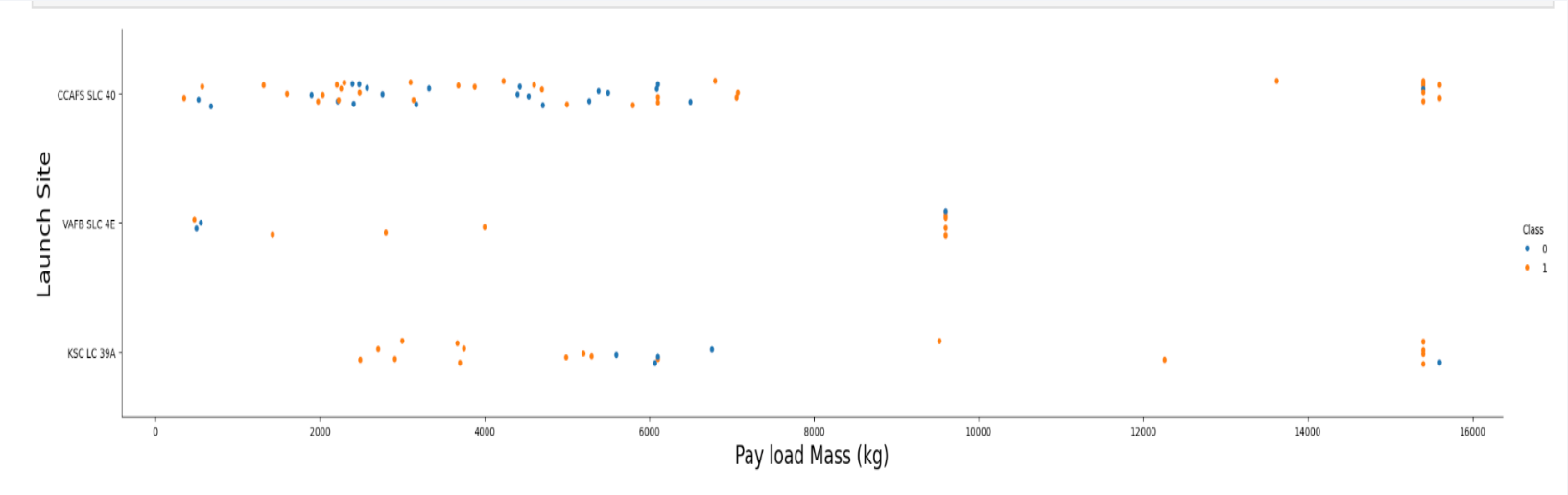
- Show a scatter plot of Flight Number vs. Launch Site



Launches from the site of CCAFS SLC 40 are significantly higher than launches from other sites

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

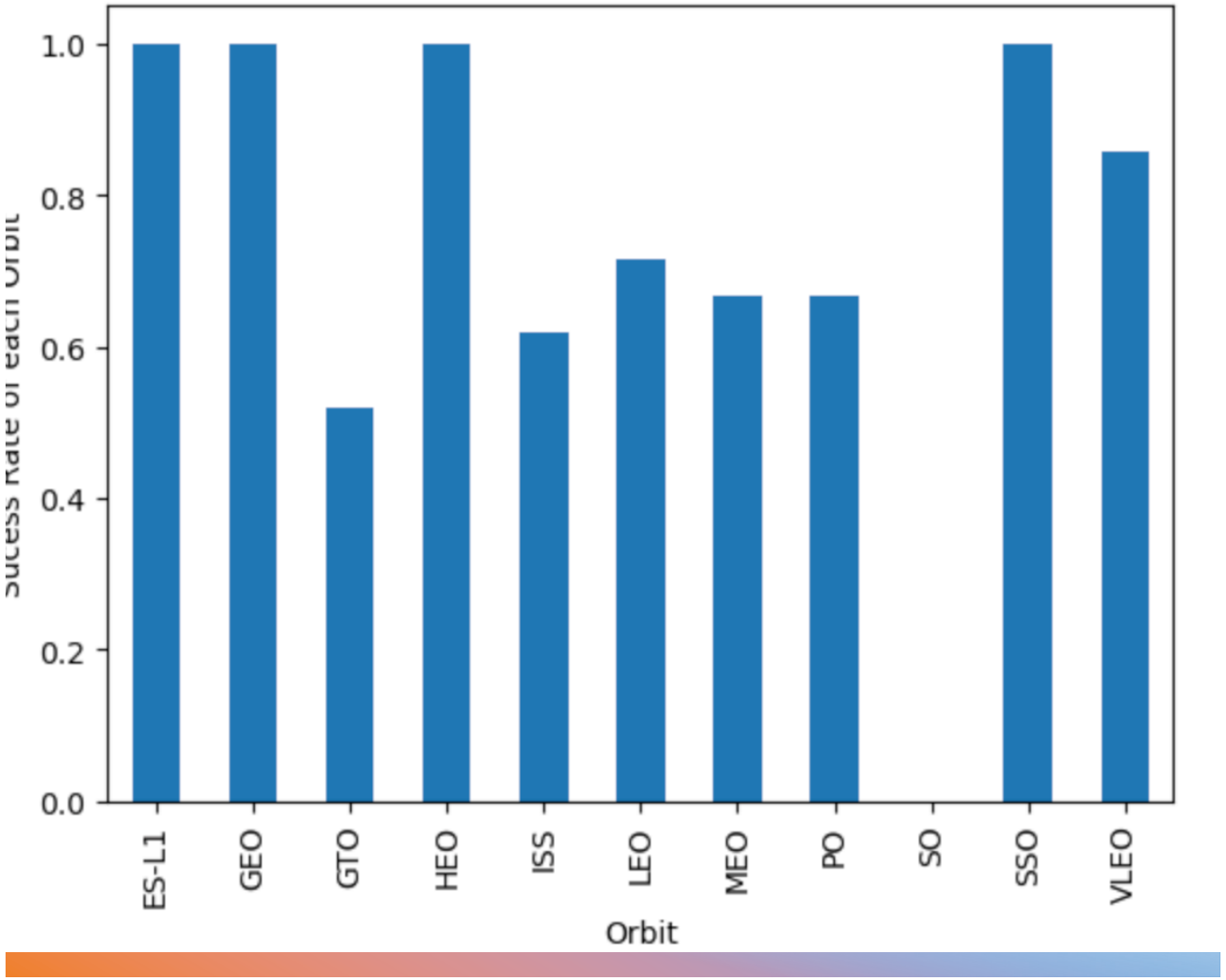


CCFAS SLC 40 had majority of the payload mass being low

Success Rate vs. Orbit Type

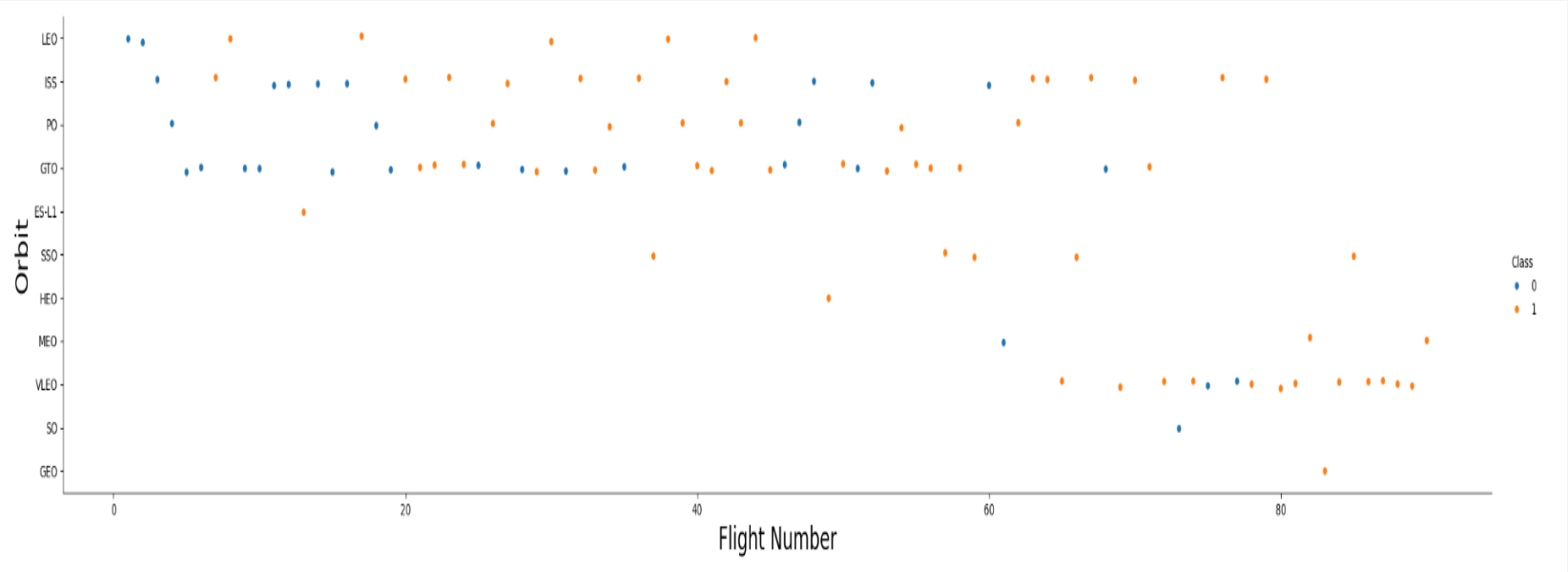
- Show a bar chart for the success rate of each orbit type

- The orbit types with the highest success rates were ES-L1, GEO, HEO AND SSO



Flight Number vs. Orbit Type

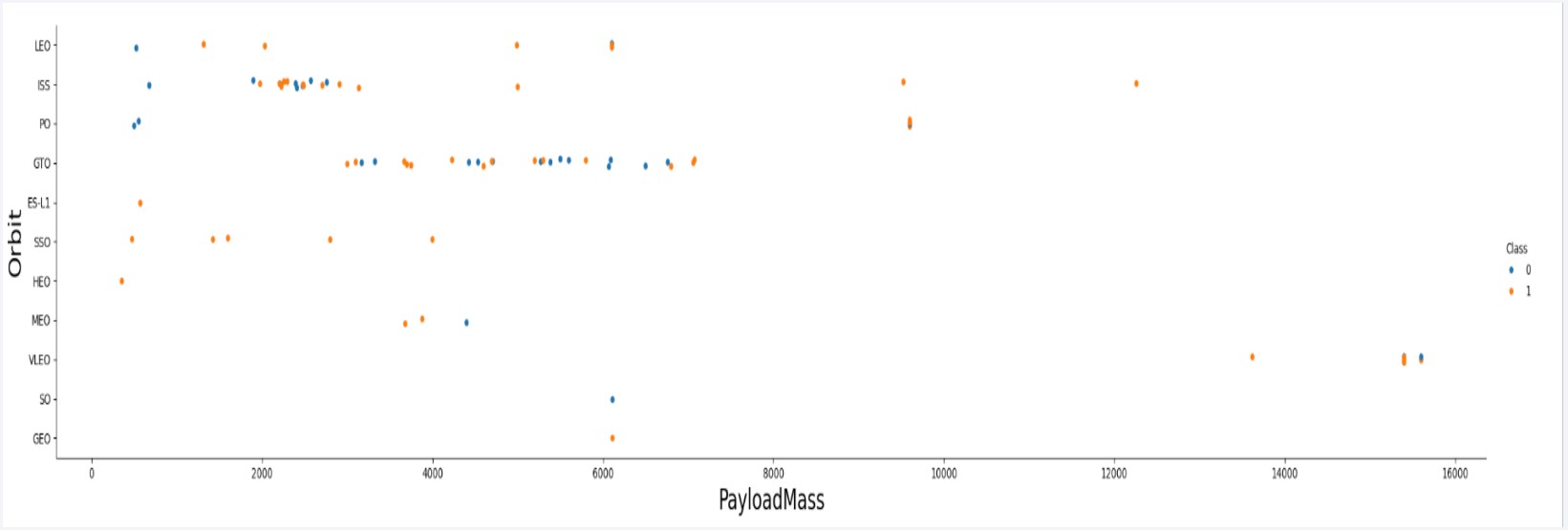
- Show a scatter point of Flight number vs. Orbit type



In 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

- Show a scatter point of payload vs. orbit type

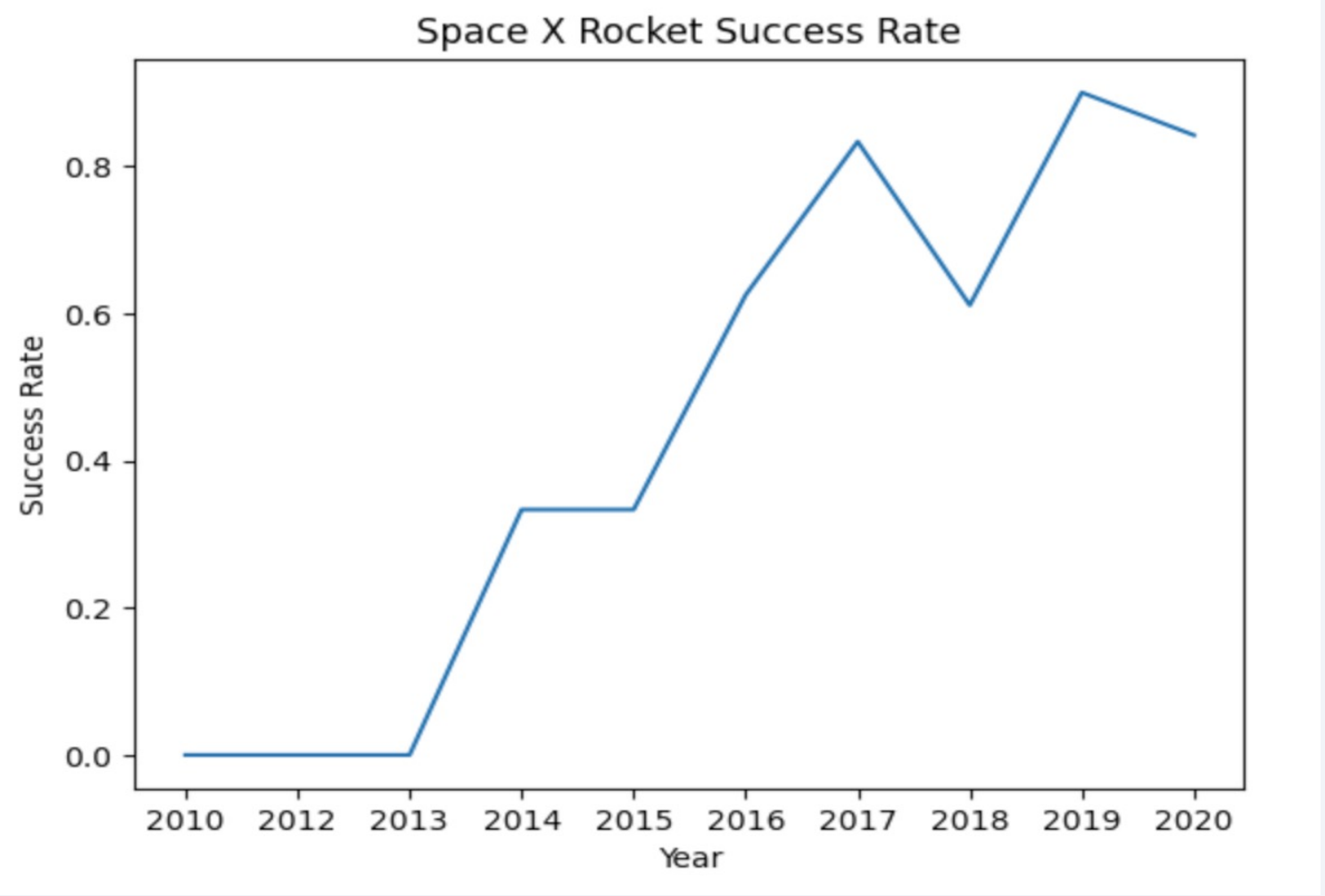


There is a stong correlation between GTO and the range of 4000-8000. There is also a strong correlation between ISS and payload range around 2000.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate

It can be observed that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites

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

- Present your query result with a short explanation here
- SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXTBL
- Explanation
- The use of DISTINCT in the query allows to remove duplicate

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

- Present your query result with a short explanation here

SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5

We can only get 5 rows by using “LIMIT”

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

SUM("PAYLOAD_MASS__KG_")
45596

- Present your query result with a short explanation here
- SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "CUSTOMER" = 'NASA (CRS)'
- We can get the sum of all values by using “SUM”

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

AVG("PAYLOAD_MASS__KG_")
2928.4

- Present your query result with a short explanation here
- %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" ='F9 v1.1'
- We can get the average values by using 'AVG'

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

MIN("DATE")
2015-12-22

- Present your query result with a short explanation here

- %sql SELECT MIN("DATE") FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'

- We are able to get the first successful data by using 'MIN' because the first date is the same with the minimum date.

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Present your query result with a short explanation here

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' \

AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

The payload mass was taken between 4000 and 6000 only and the landing outcome was determined to be “success drone ship”

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

[38]:

SUCCESS	FAILURE
100	1

- Present your query result with a short explanation here

- %sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, \
- (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS FAILURE

- We can get the number of successful and failure mission outcomes by using COUNT and LIKE “Success%” and COUNT and LIKE “Failure%” respectively

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
- %sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
- WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM S
- We can get the maximum payload masses by using “MAX”

]:

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

SQL QUERY

```
%sql SELECT substr("DATE",6, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM
SPACEXTBL \
WHERE "LANDING_OUTCOME" = 'Failure (drone ship)' AND substr("DATE",0,5) = '2015'
```

Results

Out[16]:

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

Explanation

This query result returns month, booster version, launch site where landing was unsuccessful and landing date took place in 2015. Substr function process date in order to take month or year. Substr(DATE, 6, 2) shows month. Substr(DATE,0, 5) shows year.

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

- **SQL QUERY**

%%sql select landing__outcome, **count**(*) **as** count_outcomes **from** SPACEXDATASET **where** date **between** '2010-06-04' **and** '2017-03-20' **group by** landing__outcome **order by** count_outcomes **desc**;

[50]:

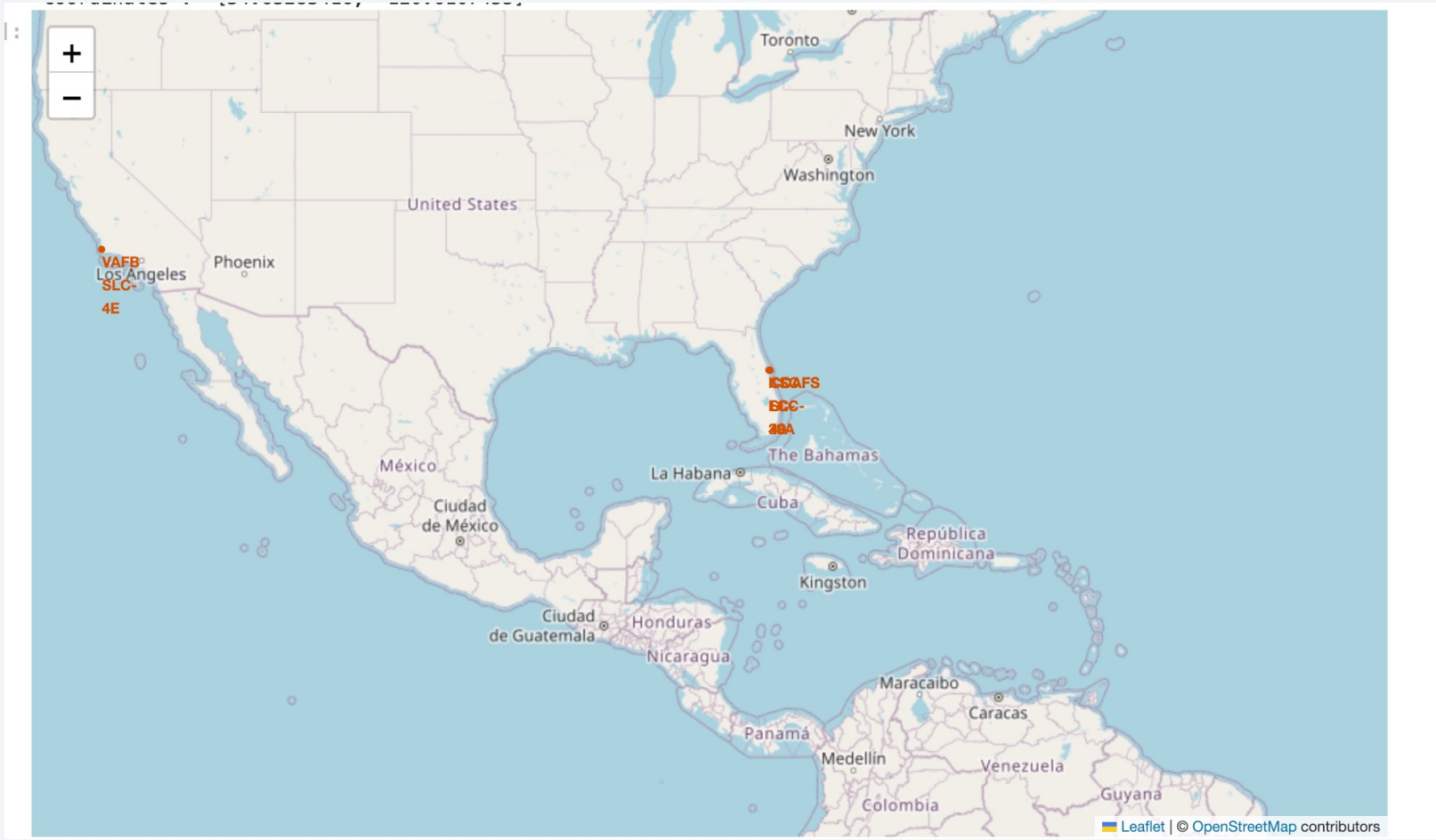
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Explanation:

Ranking 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.

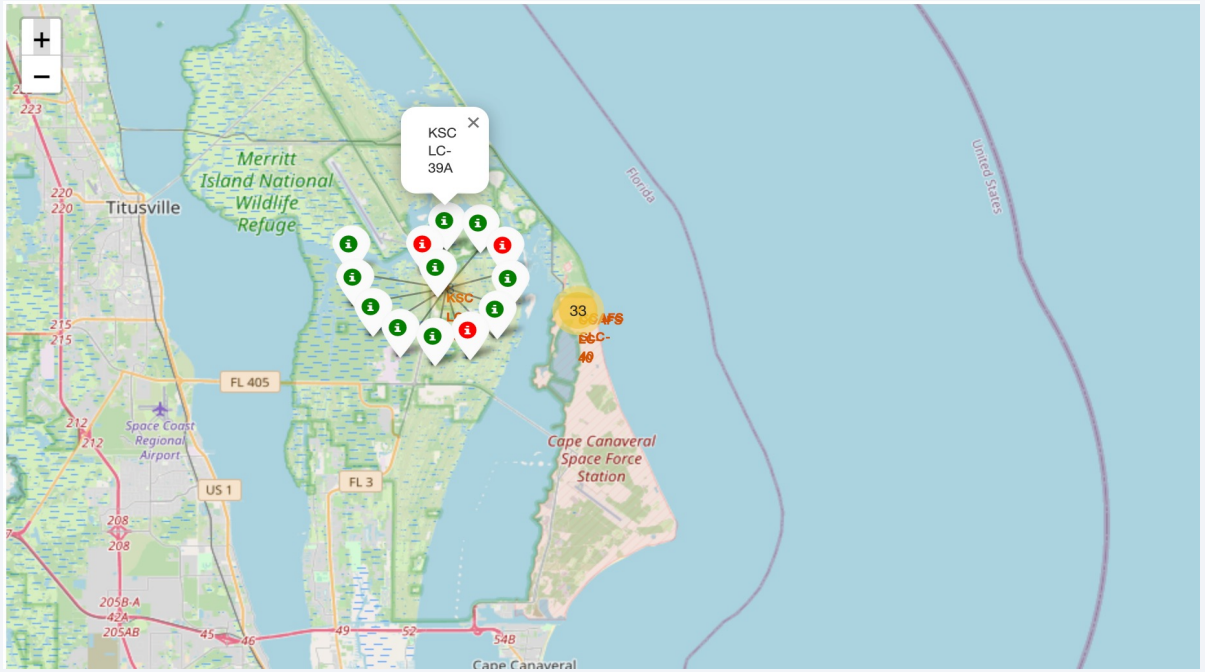

Section 3
Launch Sites
Proximities Analysis

Folium map – Ground stations



All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.

Folium map – Color Labeled Markers



Explanation:

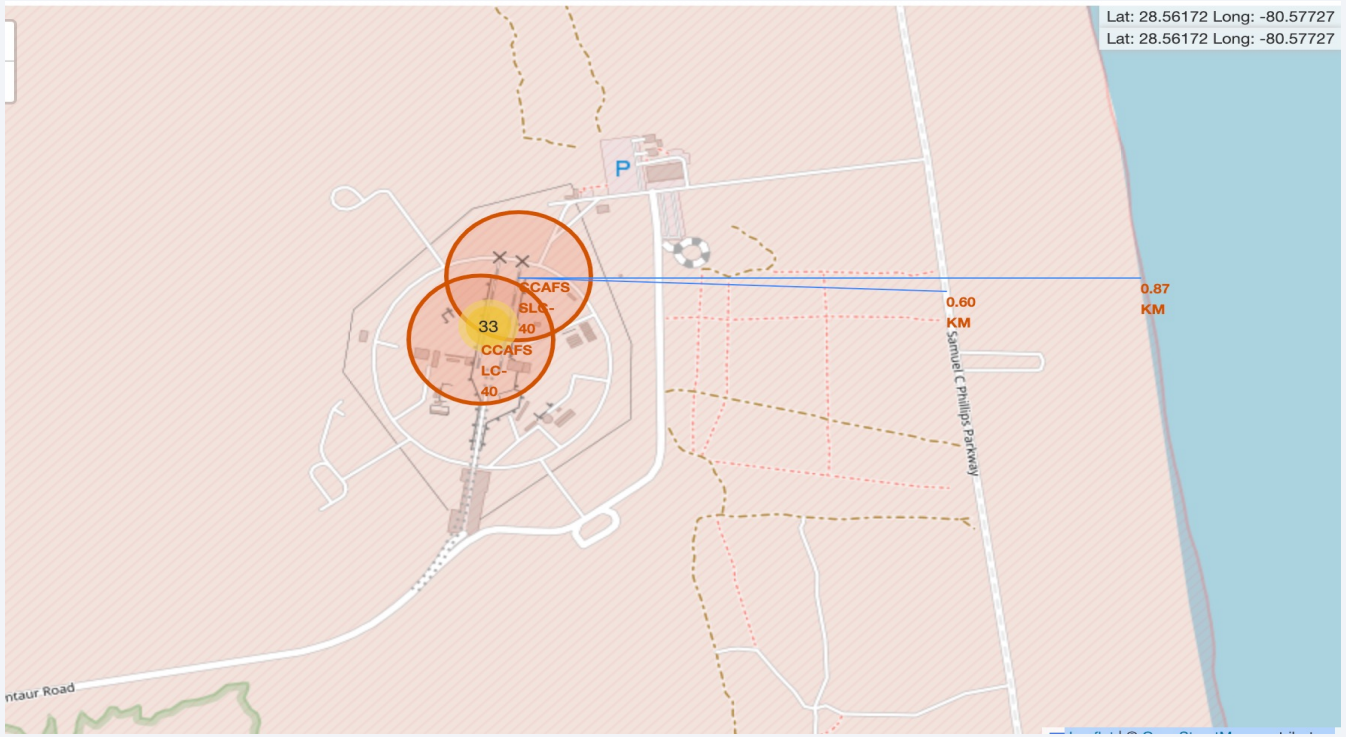
• From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

Green Marker = Successful Launch

Red Marker = Failed Launch

• Launch Site KSC LC-39A has a very high Success Rate.

Folium Map- Distances between CCAFS SLC-40 and its proximities

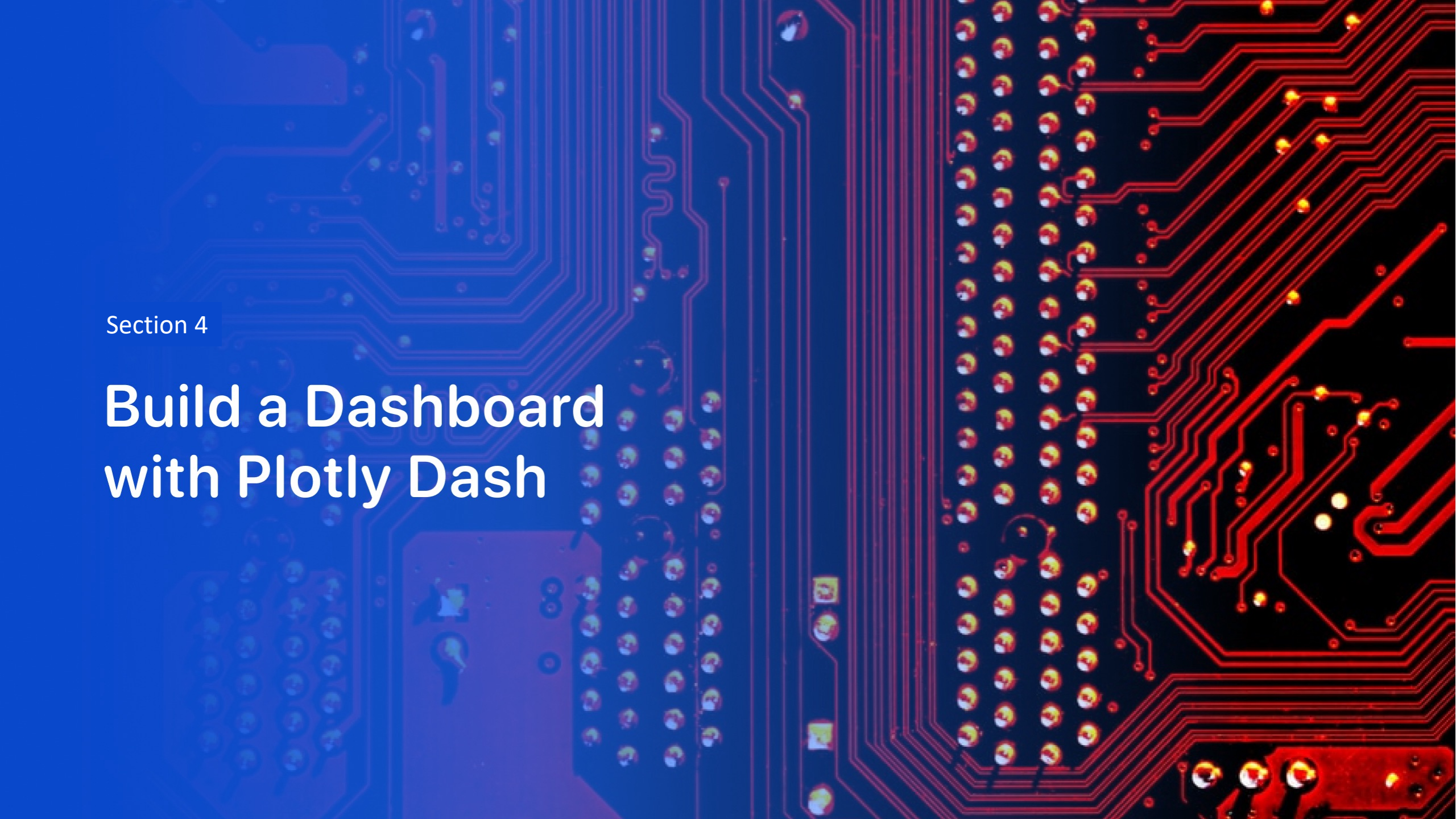


Is CCAFS SLC-40 in close proximity to railways ? Yes

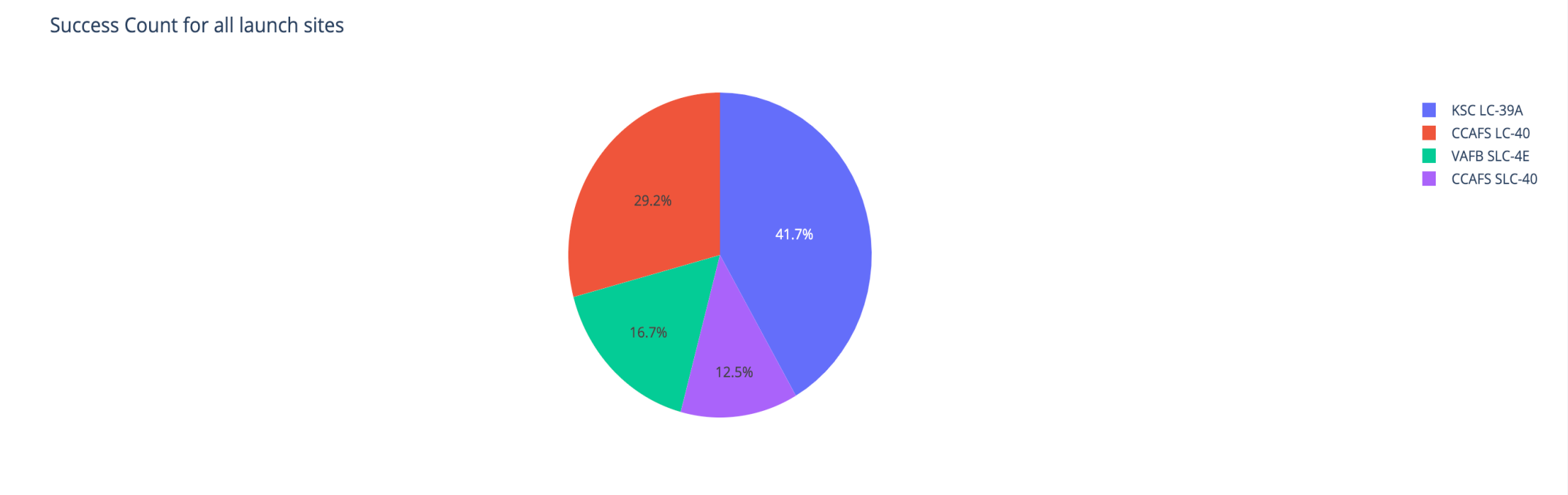
Is CCAFS SLC-40 in close proximity to highways ? Yes

Is CCAFS SLC-40 in close proximity to coastline ? Yes

Do CCAFS SLC-40 keeps certain distance away from cities ? No

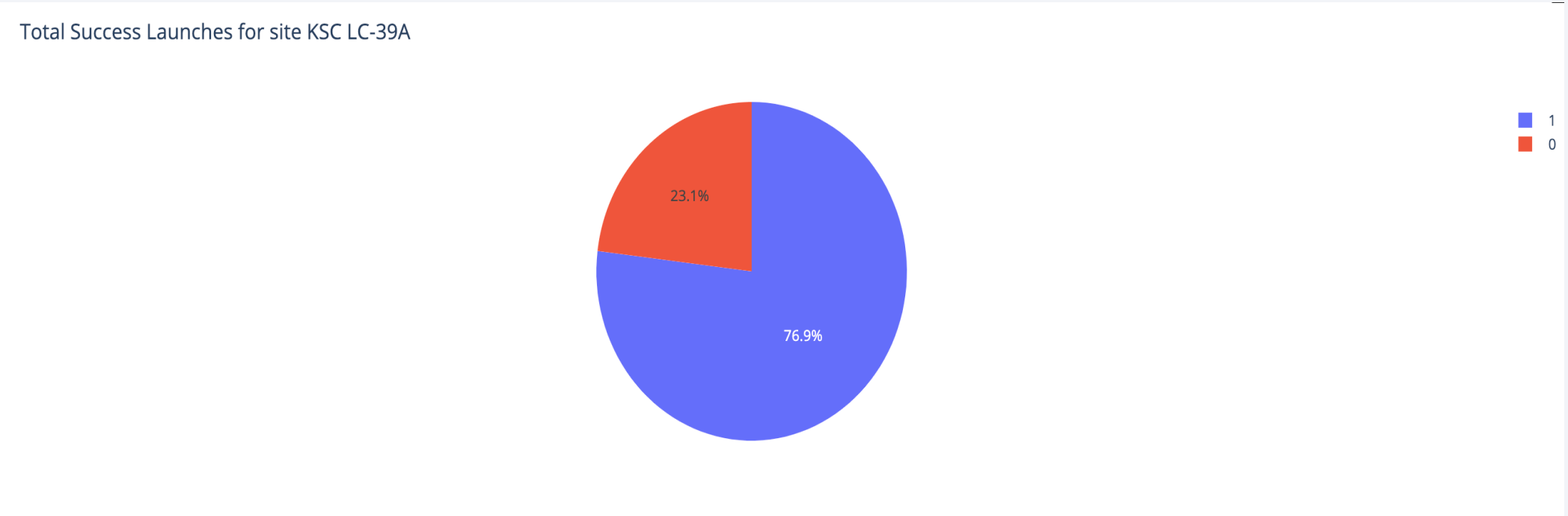

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



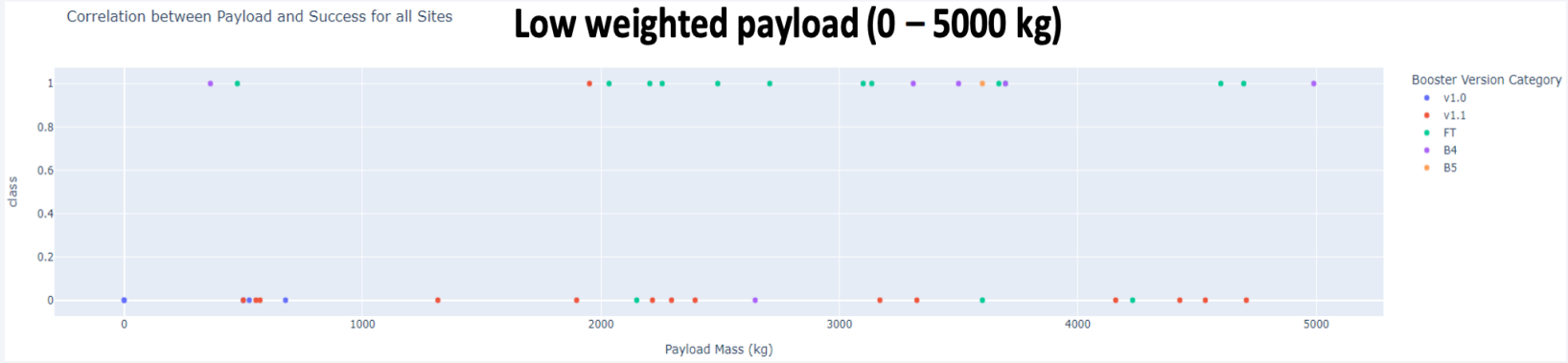
The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

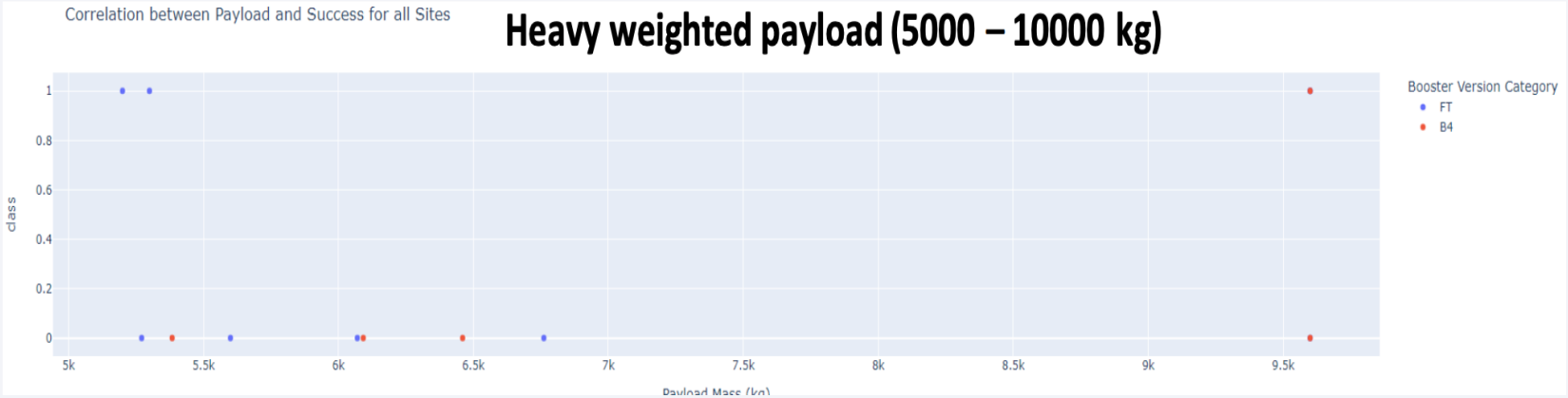
Launch site with highest launch success ratio



From the pie chart, KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate.

Payload mass vs Outcome for all sites with different payload mass selected



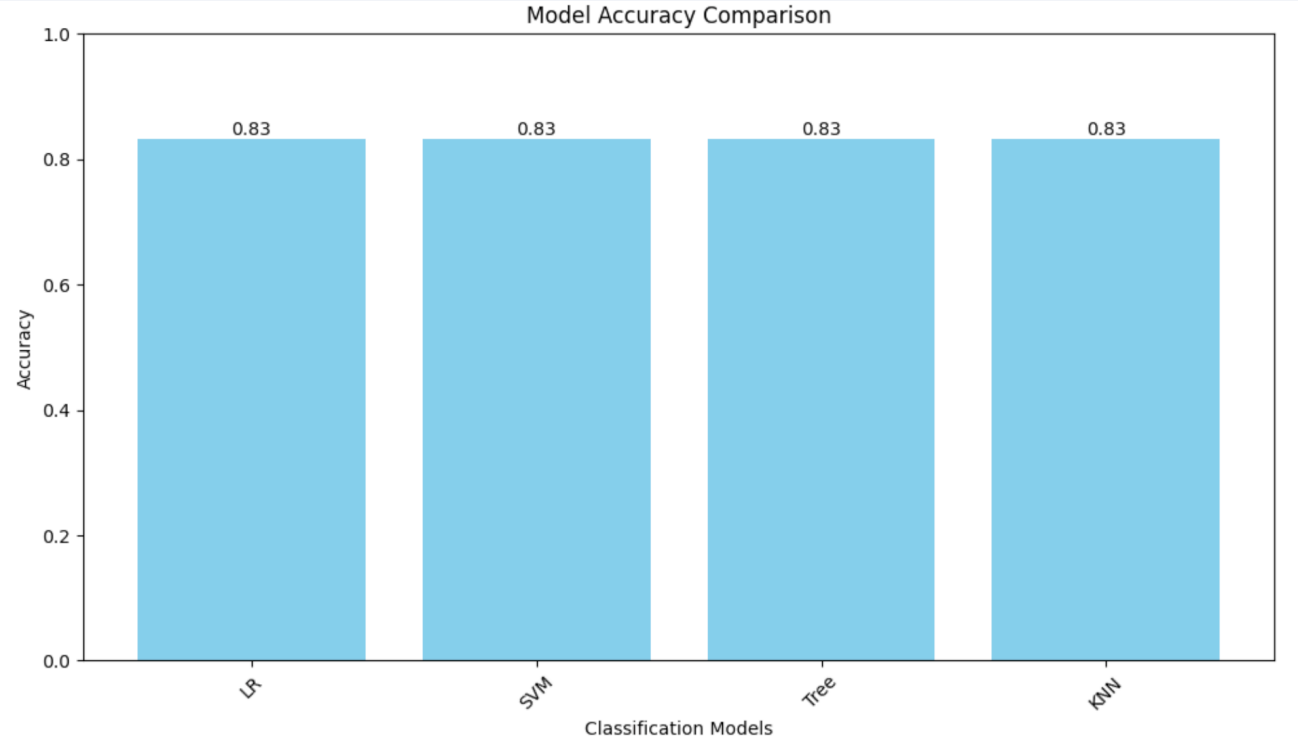


Low weighted payloads have a better success rate than the heavy weighted payloads.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

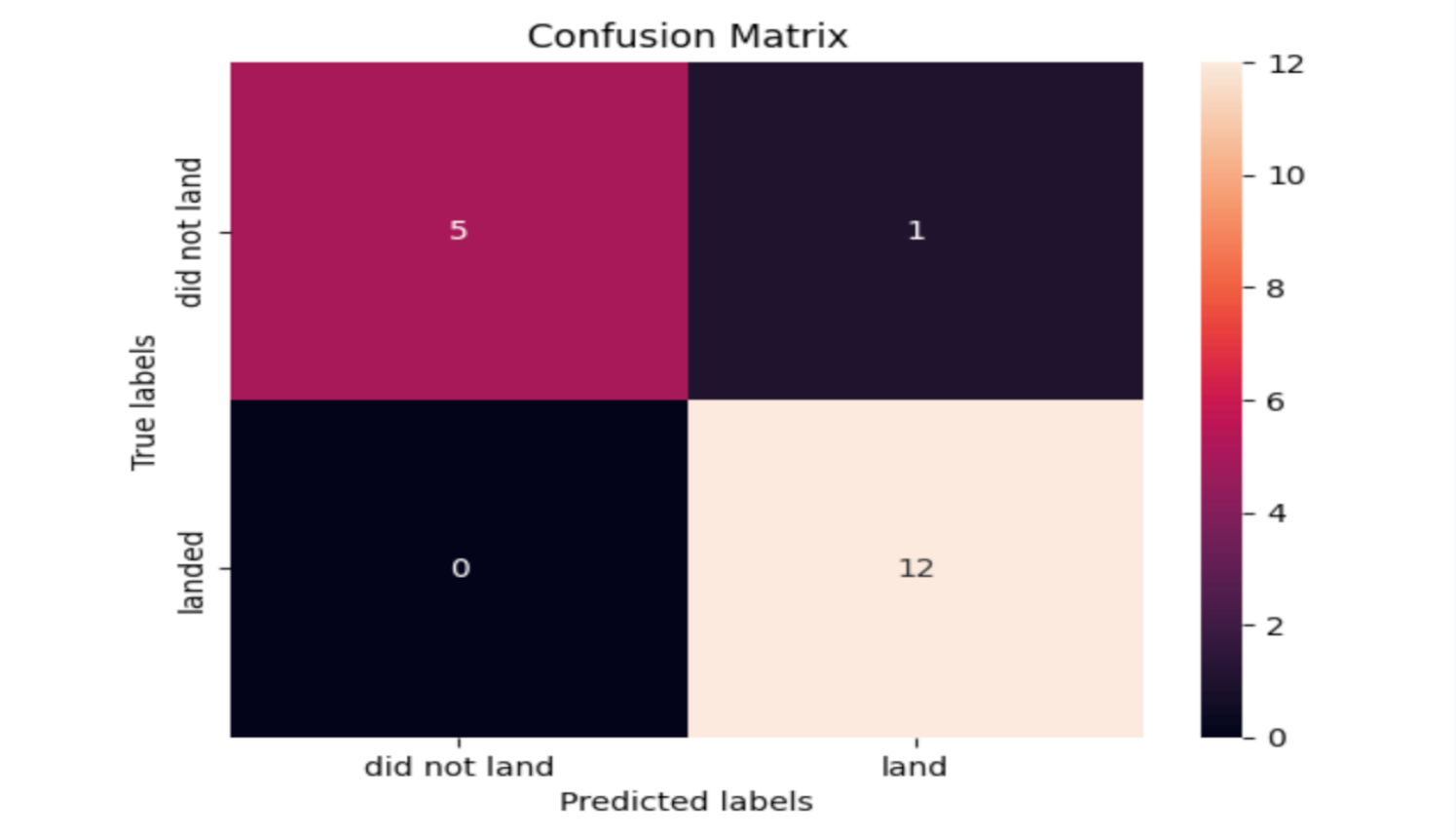


[41]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.952381	0.819444
F1_Score	0.909091	0.916031	0.975610	0.900763
Accuracy	0.866667	0.877778	0.966667	0.855556

- Based on the scores of the Test Set, we cannot confirm which method performs best.
- • Same Test Set scores may be due to the small test sample size (18 samples).
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Confusion Matrix



Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones. As the test accuracy are all equal, the confusion matrices are also identical. The main problem of these models are false positives.

Conclusions

- Certain orbits, specifically GEO, HEO, SSO, and ES-L1, consistently show higher success rates.
- The mass of the payload also plays a significant role in mission success, varying with the orbit type. Generally, missions with lighter payloads have higher success rates than those with heavier payloads.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Presently, the data does not conclusively explain why some launch sites outperform others, such as why KSC LC-39A is considered the premier site. Further investigation into atmospheric conditions or other pertinent factors might provide clarity.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- In analyzing the effectiveness of various predictive models with this dataset, the Decision Tree Algorithm was selected as the preferred model. Despite similar test accuracies across all models considered, the Decision Tree Algorithm demonstrated superior training accuracy, making it the model of choice.

Appendix

- Coursera
- Dash
- Machine learning

Thank you!