IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

ELINA RZAYEVA

10.20.2023

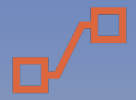
OUTLINE

Executive Summary

Introduction

Methodology

Results



Conclusion

Appendix

EXECUTIVE SUMMARY

Summary of methodologies

Data Collection through API and Web Scraping

Data Wrangling

Exploratory Data Analysis with SQL

Exploratory Data Analysis with Data Visualization

Interactive Maps with Folium

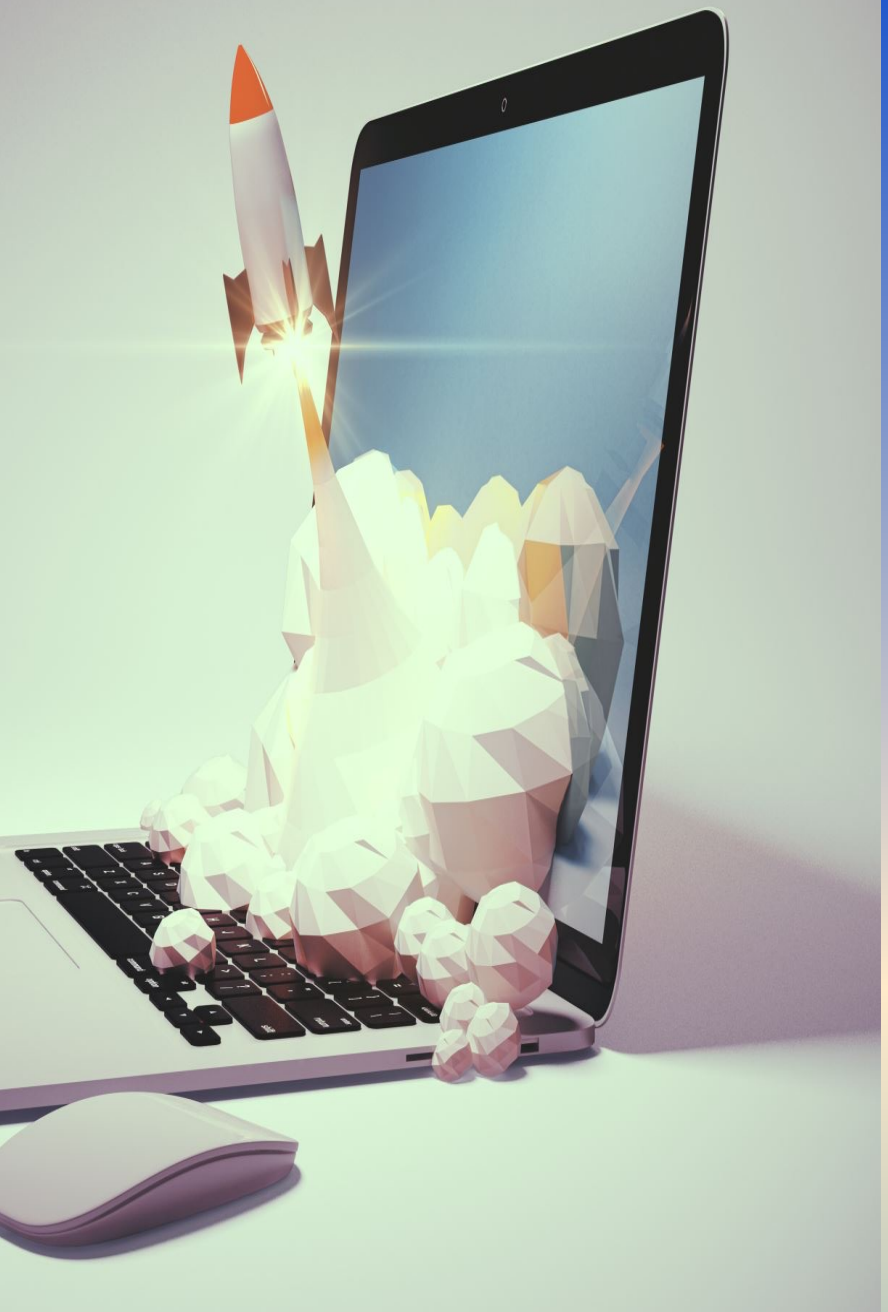
Predictive Analysis (Classification)

Summary of all results

Exploratory Data Analysis results

Interactive analytics in screenshots

Predictive analysis result



INTRODUCTION

Project background and context

Nowadays commercial space age is real. Therefore, companies are trying to make travel to space affordable for mankind. Today one of the most successful companies in this field is SpaceX. One reason we believe SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars however, other providers cost upwards of 165 million dollars each. Therefore, if we can determine the first stage successfully landing, we can determine the cost of a launch as well. Moreover, Spaces X's Falcon 9 could launch like a regular rockets. So, the goal of the project is to create a machine learning model to predict if the SpaceX reuse of the first stage and its land successfully.

Problems we wish to find Answers:

- Our goal is to use this data to predict whether SpaceX will successfully land on its first stage or not?

- How does variables such as payload mass, landing pad, number of flights and launch site affect to the success rate of the first stage landing?

- what is the best algorithm that can be used for classification factors in this situation?

Section 1

Methodology

METHODOLOGY

Executive Summary

Data collection methodology:

- via SpaceX Rest API
- via Web Scraping from Wikipedia

Performed data wrangling

- Filtering the data, dealing with missing values
- One-hot encoding was applied to categorical features

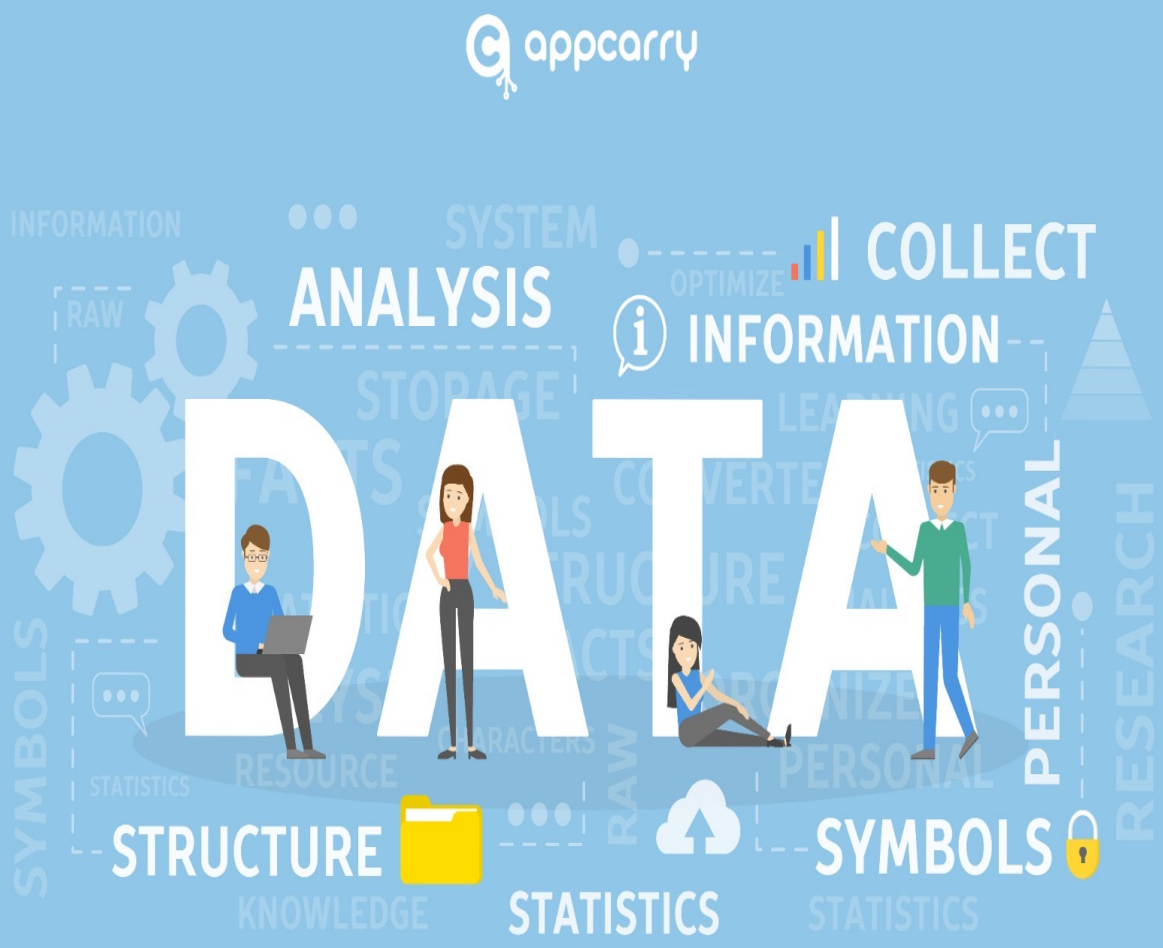
Performed exploratory data analysis (EDA) using visualization and SQL

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models

- Building, tuning and evaluating classification models to secure best outcome

DATA COLLECTION



Data sets were collected by implementing various methods such as:

First, there were used “get request” to the SpaceX Rest API and Web Scraping data

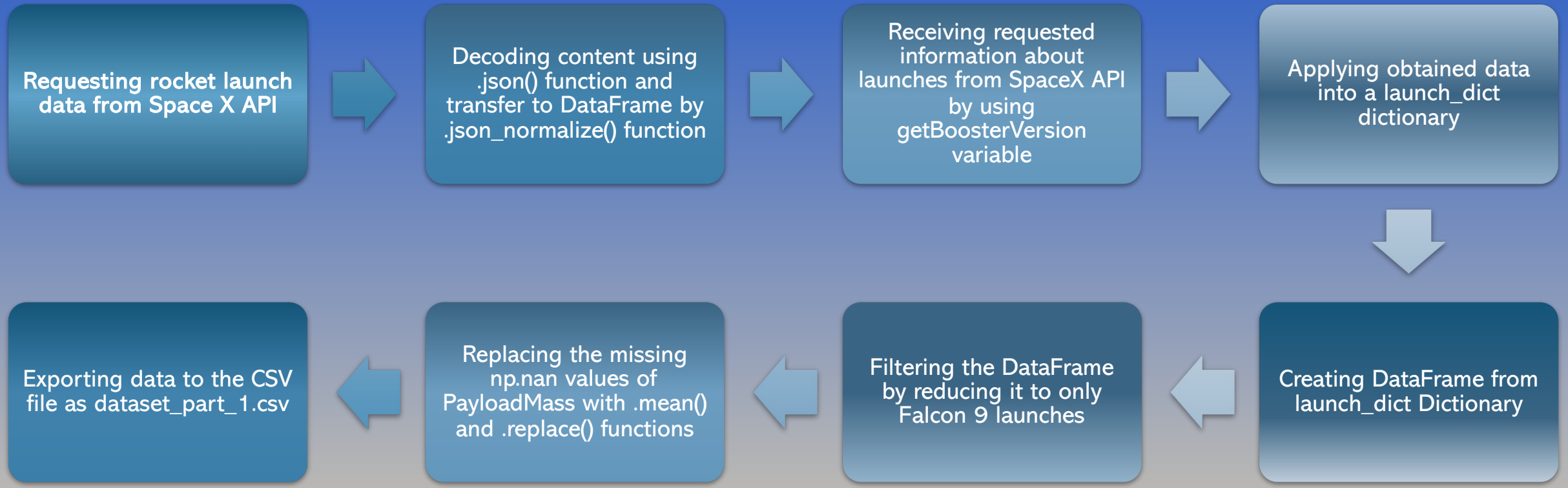
Then, response content was decoded Json using .json() function call and turn the data into a pandas dataframe using .json normalize()

After, cleaning data, the missing values were checked and filled into missing points by necessity

In addition, web scraping from Wikipedia for Falcon 9 launch recors with BeautifulSoup was performed

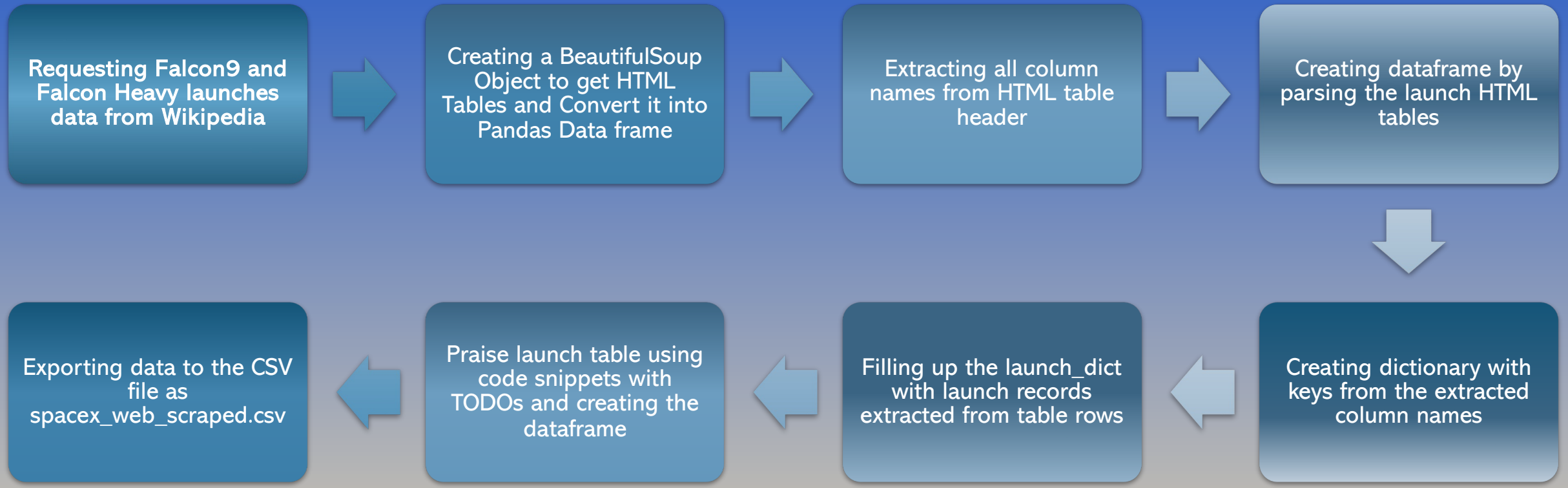
The objectives was to extract the launch records as HTML table and parse the table and convert it into a pandas DataFrame for further detailed analysis

DATA COLLECTION – SPACEX API



SpaceX Data Collection with API URL

DATA COLLECTION – WEB SCRAPING



SpaceX Data Collection with Web Scraping URL

DATA WRANGLING

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

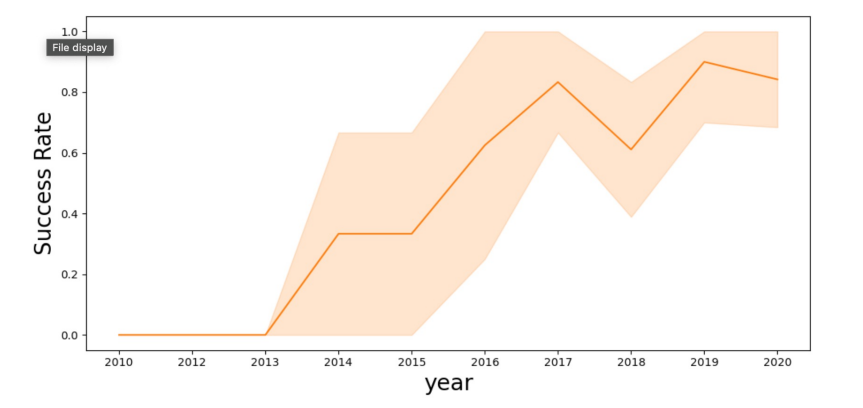
In this lab we will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

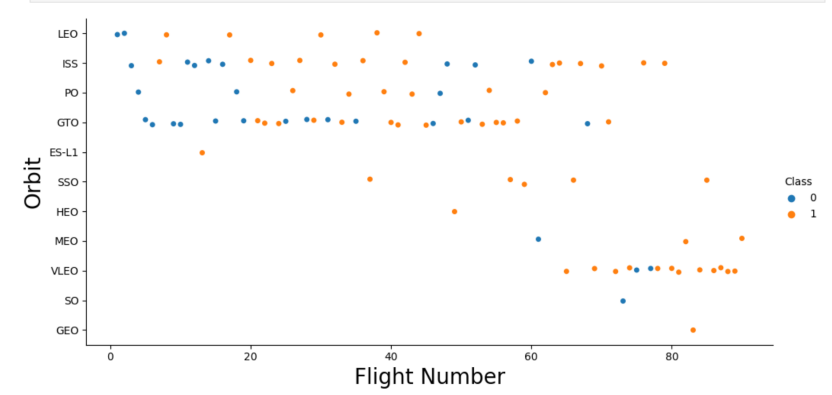
Falcon 9 first stage will land successfully.

- Performing Exploratory Data Analysis and Determining the Training Labels
- Calculating the number of launches at each site and orbit
- Calculating the number and occurrence of each orbit
- Calculating the number and occurrence of missing outcome per orbit type
- Creating a landing outcome label from Outcome column
- Exporting the date to CSV file

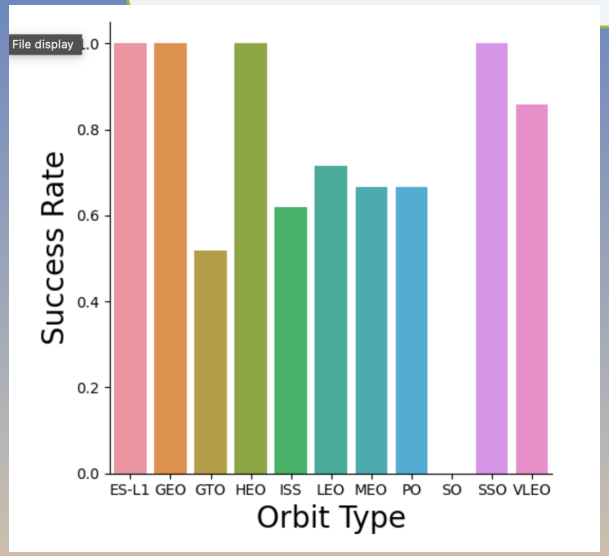
Data Wrangling URL

EDA WITH DATA VISUALIZATION





The following mentioned charts were plotted to calculate and visualize different launch sites like CCAFS LC-40, KSC LC-39A and VAFB SLC 4E and their success rates which was observed increasing since 2013 till 2020. We see that as the flight number increases, the first stage is more likely to land successfully.



Payload Mass vs Flight Number, Launch Site vs Flight Number, Launch Site vs Payload Mass, Success Rate vs Orbit Type, Orbit vs Flight Number, Orbit vs Payload Mass, Class vs Date, Success Rate vs Year

Additionally,
function get_dummies and features dataframe was applied OneHotEncoder to the columns Orbits, LaunchSite, LandingPad and Serial where several values was assigned to the variables features_one_hot dataframe which only contained numbers cast the entire dataframe to variable type float64.

EDA WITH DATA VISUALIZATION URL

EDA WITH SQL

Summation of the performed SQL queries:

- Display of the unique launch sites names in the space mission
- Display of 5 records where launch sites begin with the string 'CCA'
- Display of the total payload mass carried by boosters launched by NASA (CRS)
- Display of average payload mass carried by booster version F9 v1.1
- List of the date when the first successful landing outcome in ground pad was achieved by using 'min function'
- List of the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List of the total number of successful and failure mission outcomes
- List of the booster version names which have carried the maximum payload mass by using 'subquery'
- List of the records which displays the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015
- Ranking of the successful landing outcomes counts like drone ship failure or ground pad success between the date 04-06-2010 and 20-03-2017 in descending order

EDA WITH SQL URL

BUILD AN INTERACTIVE MAP WITH FOLIUM

In Folium map all launch sites were marked and pinned on map with markers, circles, lines, pop-up objects and text labels. Thus, We assigned the dataframe launch_outcomes (failures, successes) to classes 0 and 1 with Green and Red markers on the map in a Marker Cluster ( )

With the help of those objects, we could visualize success and failure of launches for each site on the folium map.

The benefits of marker clusters are ability to simplify a map containing many markers which are having the same coordinates. For this we can use Marker Cluster object

Mouse Position is another object by which we can easily identify coordinates of any points of interests such as railway, highway, coastline, etc. Moreover, we could find how close this launch sites were to above mentioned proximities

Interactive Visual Analytics with Folium

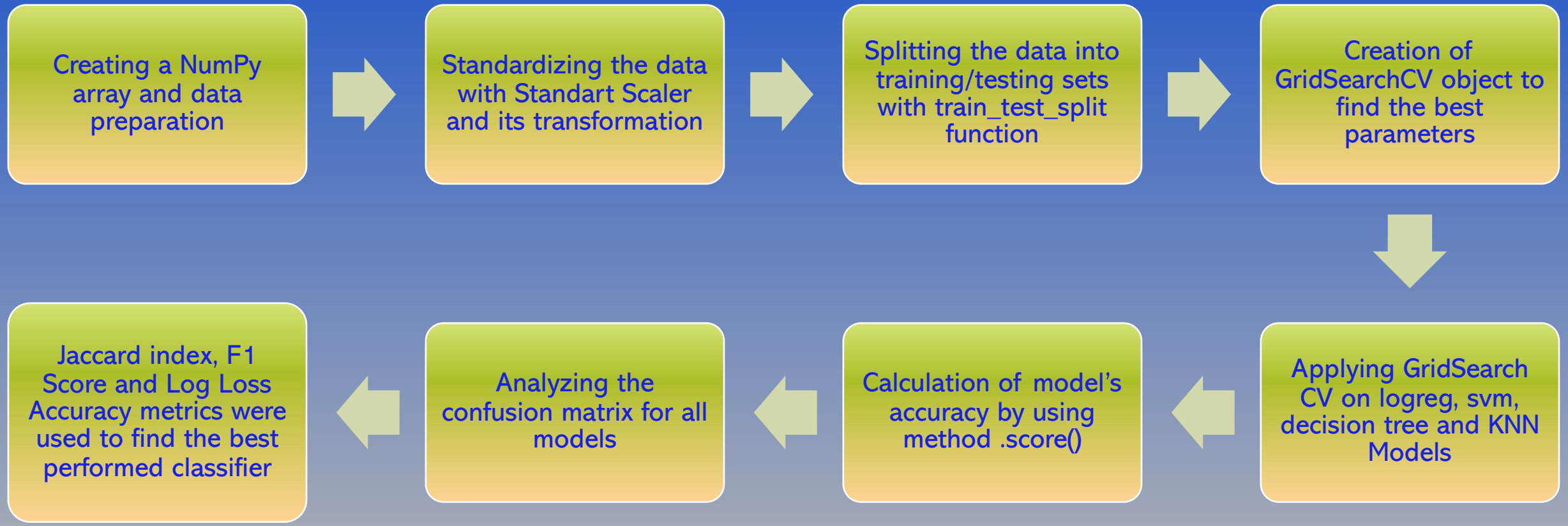
BUILD A DASHBOARD WITH PLOTLY DASH

Launch Sites Dropdown list, Pie chart which shows Successful Launches of all or certain sites, slider for Payload Mass Range and Scatter Chart for Payload Mass vs. Success Rate were added to the dashboard.

- By adding Dropdown list, we could enable launch Site selection
- Added Pie chart showed Success/Fail counts for selected Launch Site
- Added Slider helped to select specify Payload range
- Added Scatter chart showed the correlation between Payload and F9 Booster version highest and lowest Launch Success Rate.

Dashboard App

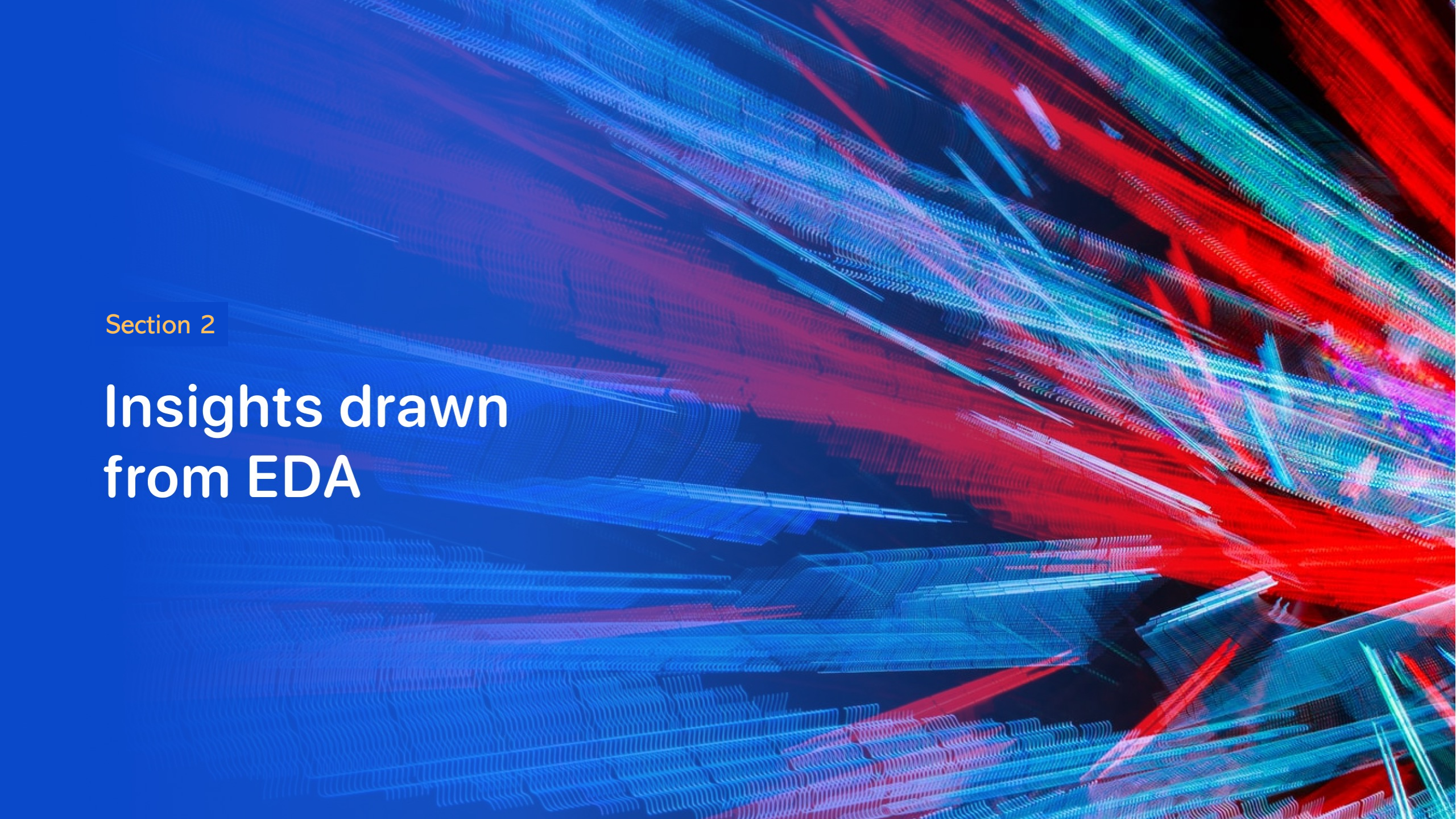
Predictive Analysis (Classification)



Machine Learning Analysis

RESULTS

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



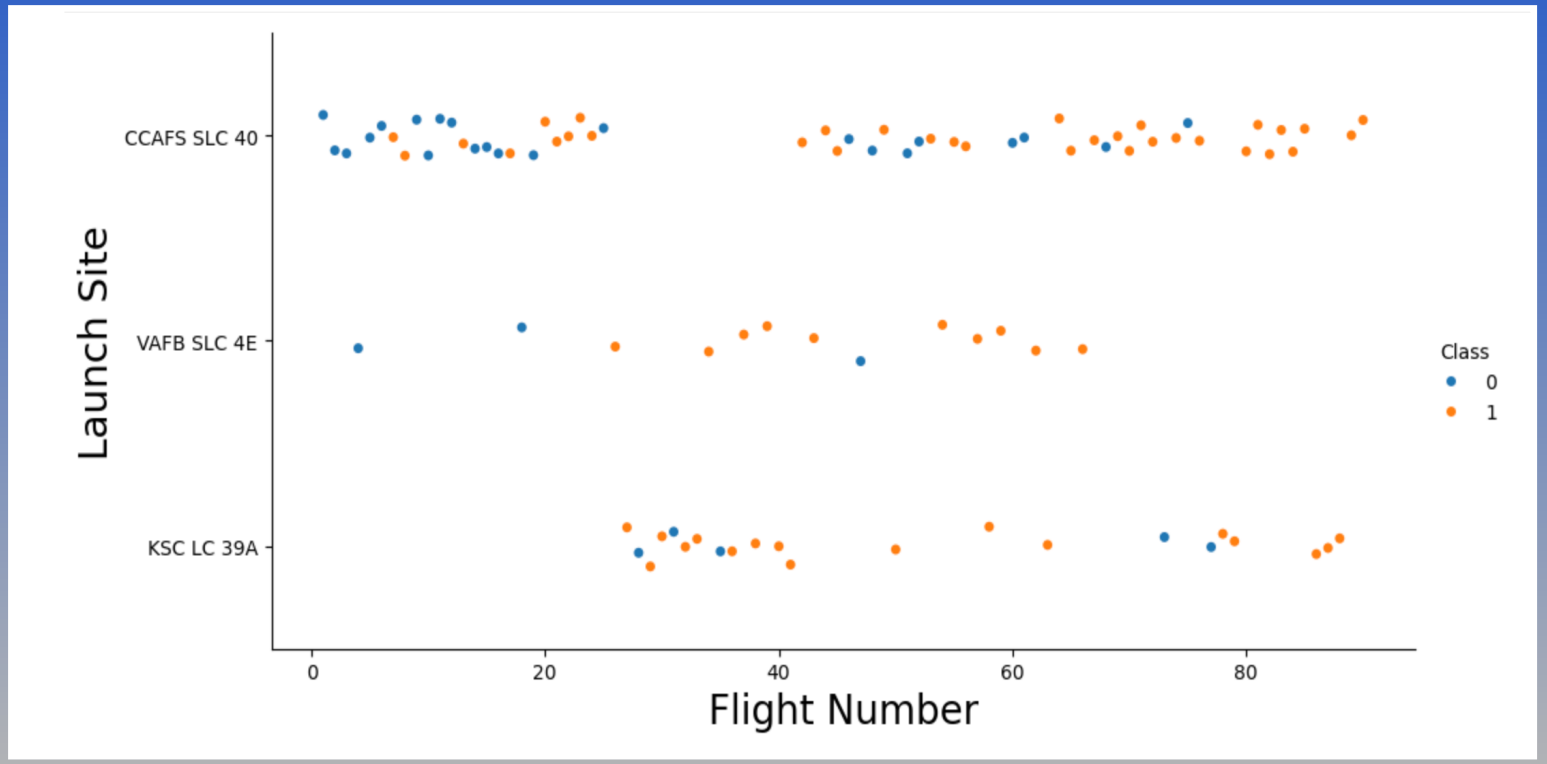
Section 2

Insights drawn from EDA

EDA WITH VISUALIZATION

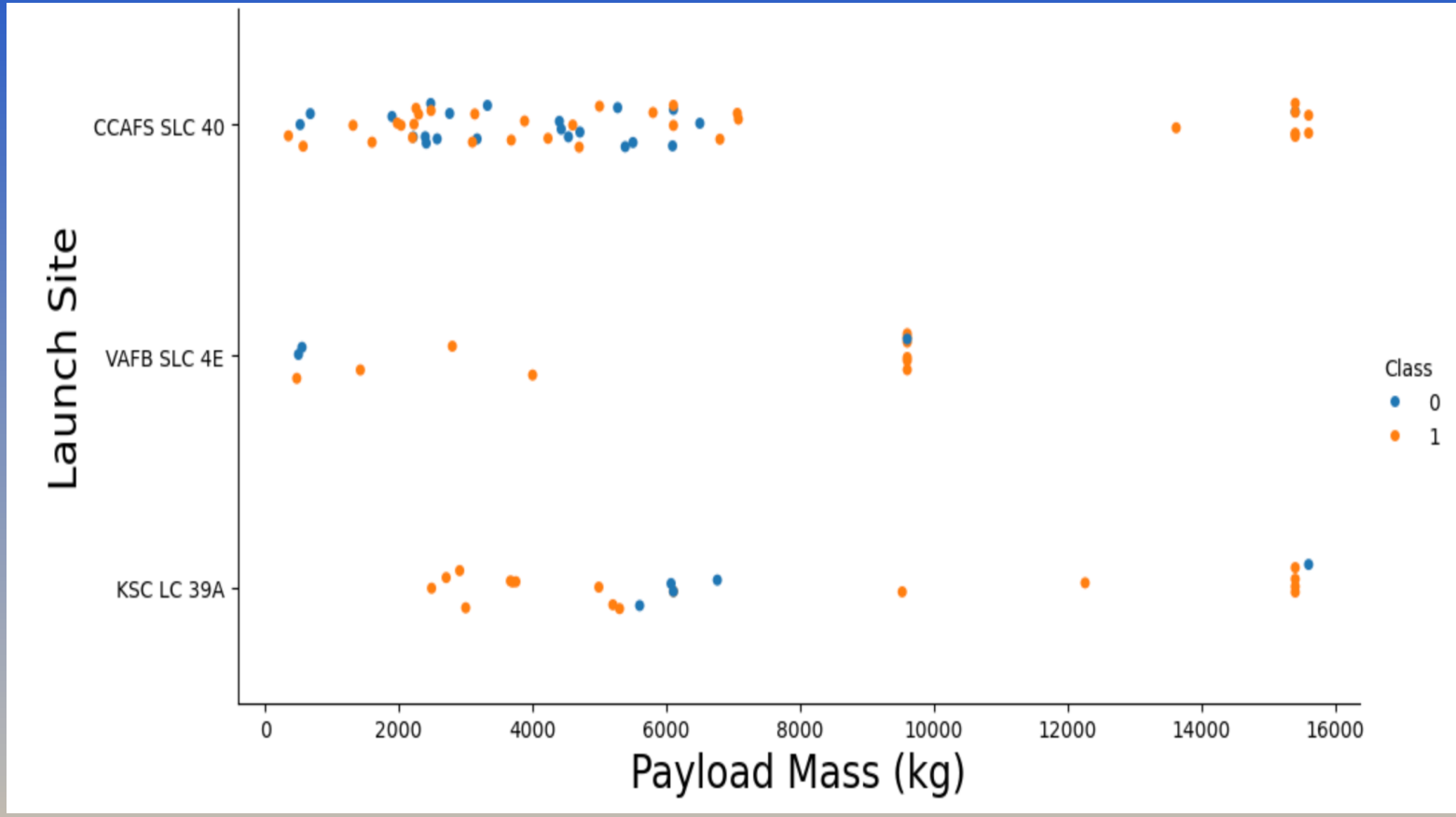
FLIGHT NUMBER VS. LAUNCH SITE

- According to the given screenshot we can identify that the best launch site is CCAFS SLC 40, where most of the launches was successful.
- After CCAFS SLC 40 in second place for best launch results stays CAFB SLC 4E
- And finally in last place is KSC LC 39A
- We can also see the improvement of general success rate by the time passes. Therefore, it can be assumed that each new launch has a higher rate of success.



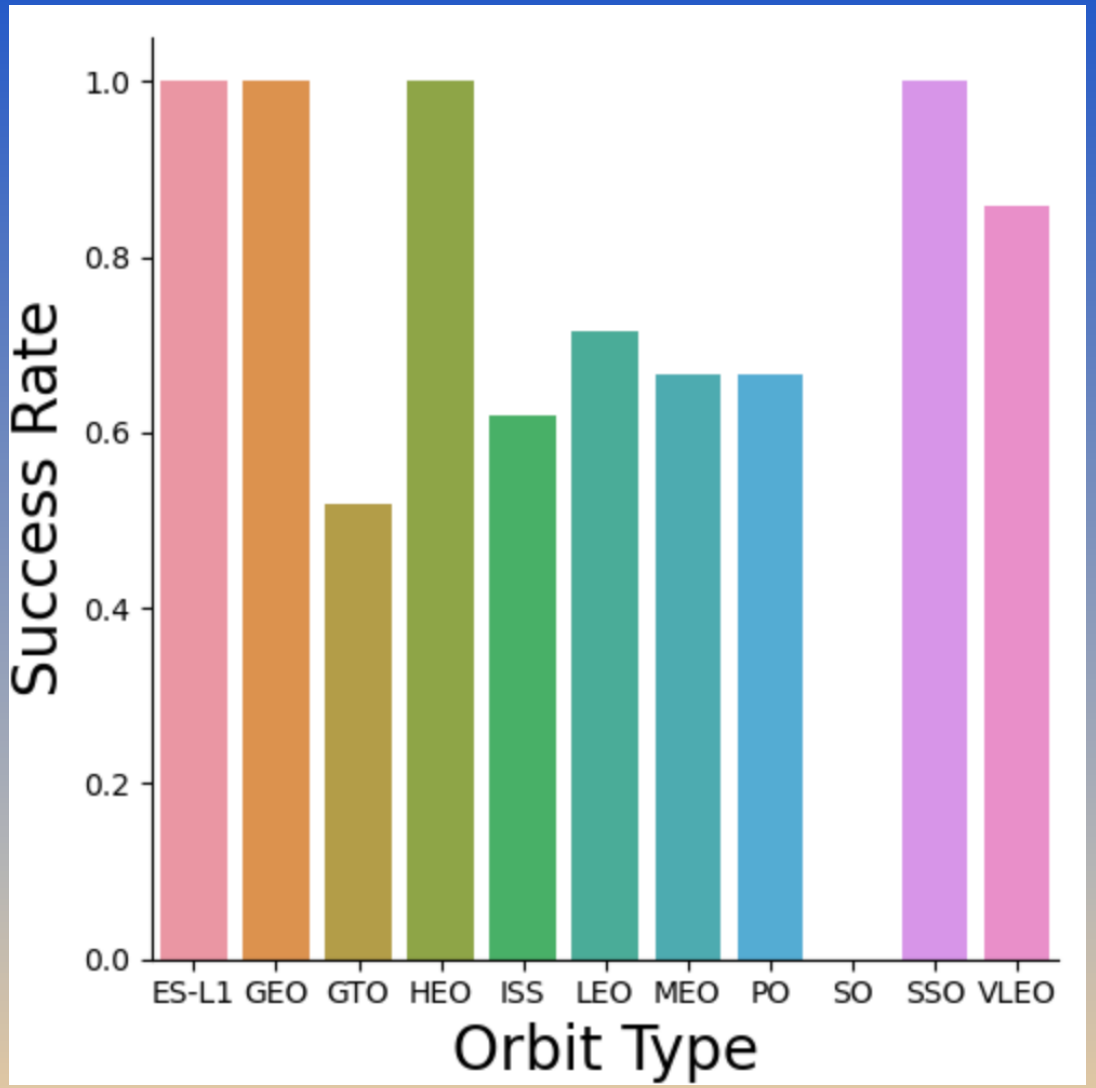
PAYLOAD VS. LAUNCH SITE

- According to the given screenshot we can identify that the higher is the payload mass, the higher is the success rate
- Most of the successful launches were till to 8000 kg payload mass
- CCAFS SLC 40 has the most successful outcome among the other launching sites.
- In comparison to less heavy launches Payloads with over 12000 kg has less success rate on launches.



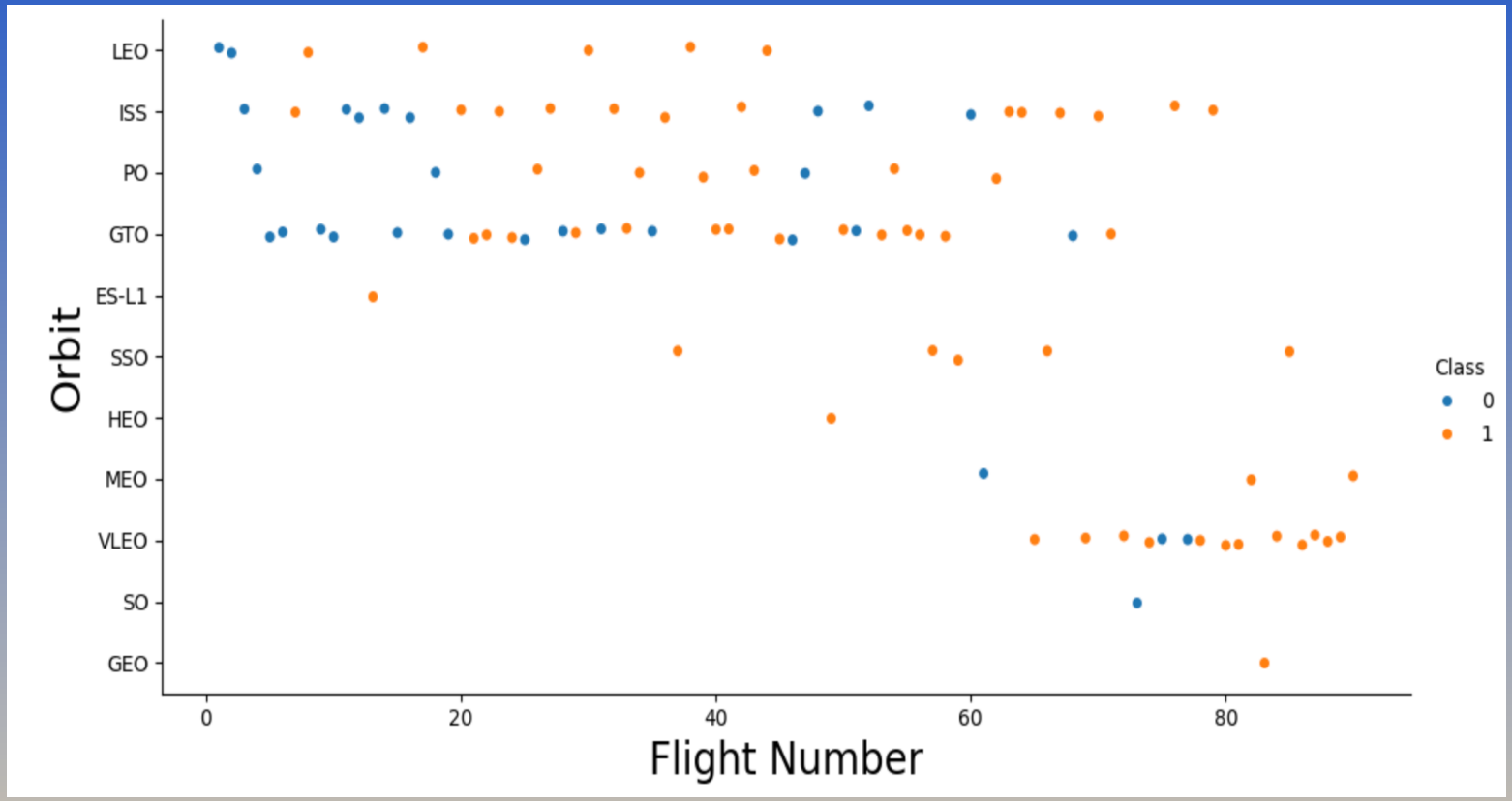
SUCCESS RATE VS. ORBIT TYPE

- The orbits which has total success rate are ES-L1, GEO, HEO, SSO
- However, SO is the orbit which has zero success rate
- The rest orbits (GTO, ISS, LEO, MEO, PO and VLEO) which we may observe from there photo have the middle success rate



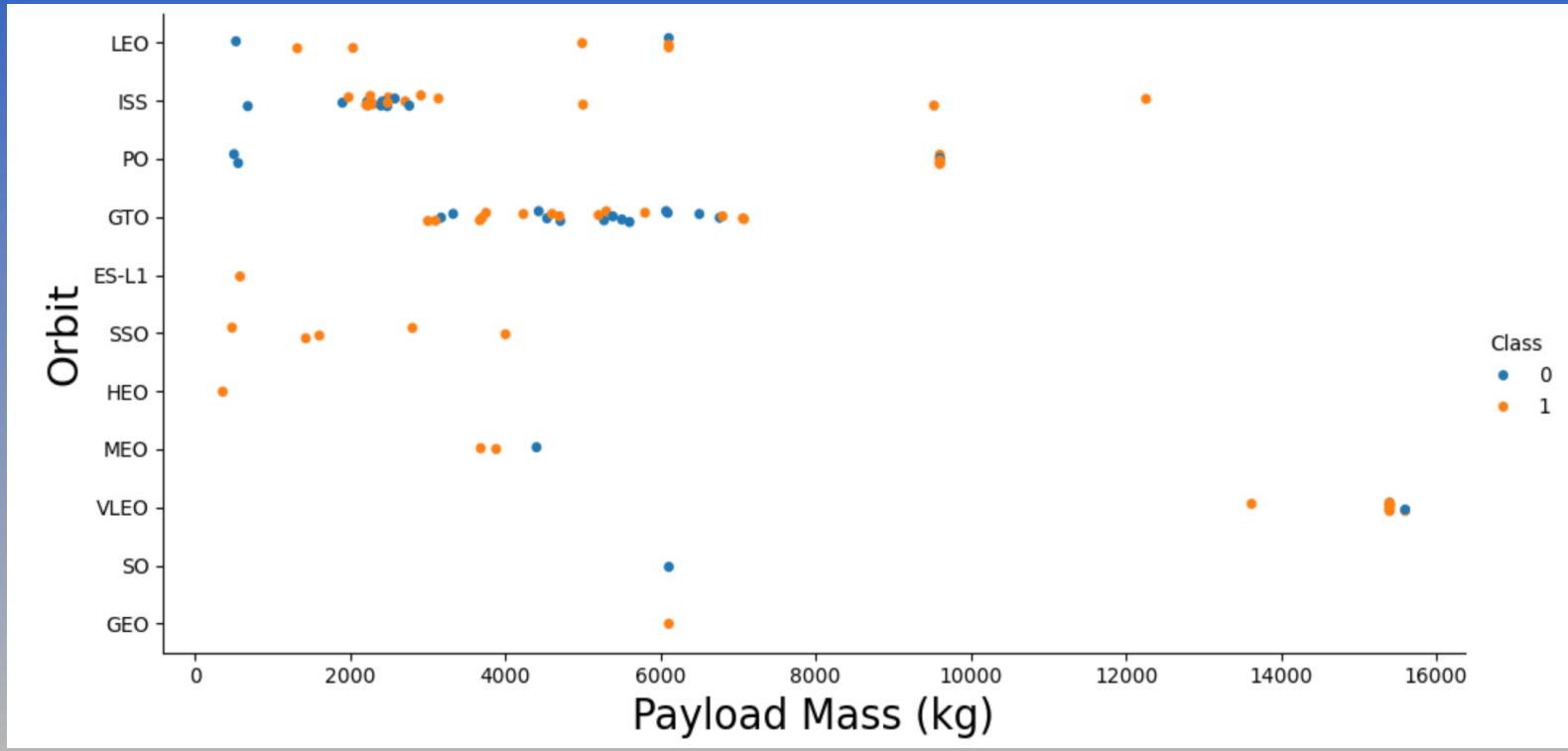
FLIGHT NUMBER VS. ORBIT TYPE

- The plot in the given photo shows the relationship between flight numbers vs orbit type. From the plot we can observe that the LEO orbits success rate is related to the number or flights.
- However, GTO orbit does not show the connection to any flight numbers

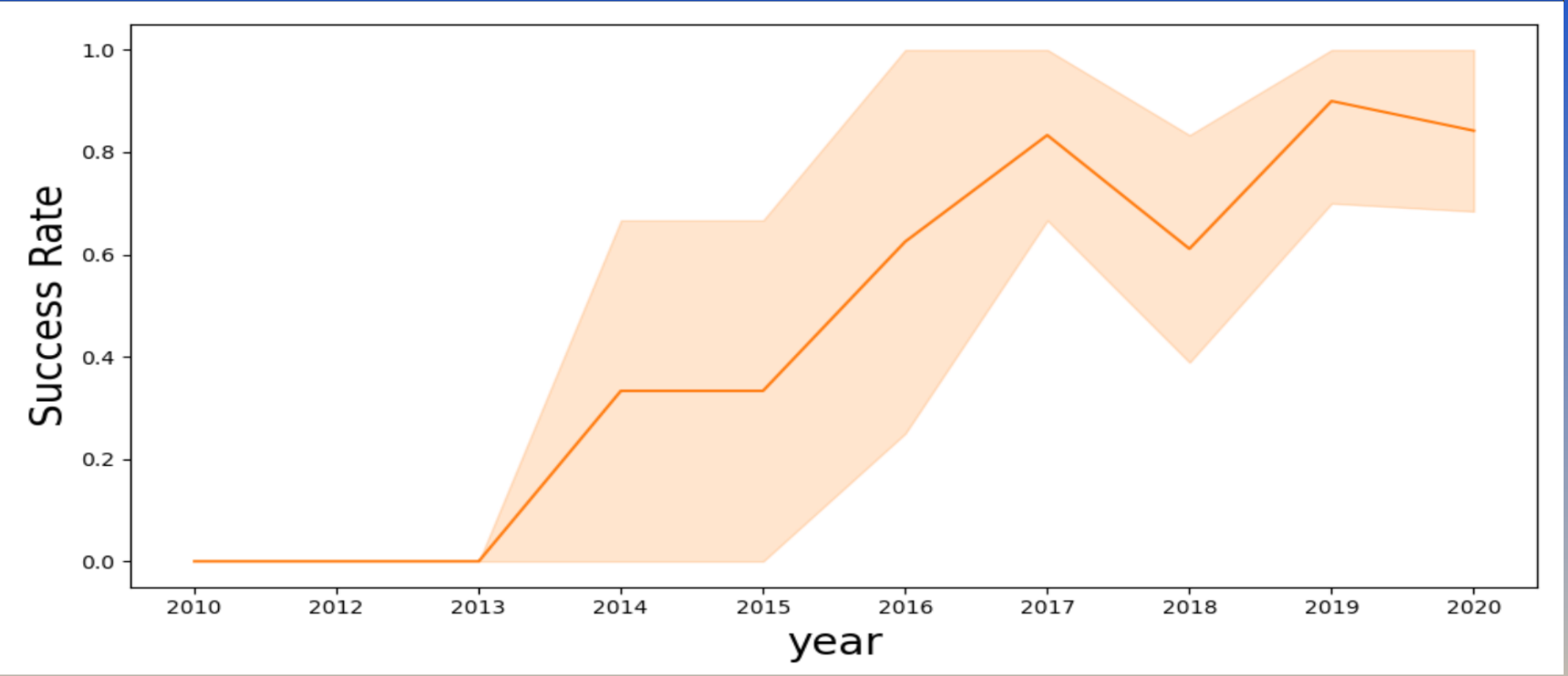


PAYLOAD VS. ORBIT TYPE

- The plot in the given photo shows the relationship between payload vs orbit type. From the plot we can observe that heavy payloads have negative effect over the GTO orbit and positive effect ISS with higher rate of success
- We can also observe very less but important launches from other orbits like SO and GEO.



LAUNCH SUCCESS YEARLY TREND



The success rate continues to increase starting from 2013 till 2017

After decreasing between 2017-2019 years, it again increases


EDA WITH SQL

All Launch Site Names

```
[19]: %sql SELECT LAUNCH_SITE, COUNT(LAUNCH_SITE) AS LS_COUNT FROM SPACEXTBL GROUP BY LAUNCH_SITE;
 * sqlite:///my_data1.db
Done.
```

[19]:

Launch_Site	LS_COUNT
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

We selected “Launch_Site” values to show only unique launch sites from the SpaceX data

Launch Site Names Begin with 'CCA'

```
[20]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

[20]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

We used query to displayed five records of launch site names which begin with ‘CCA’

Total Payload Mass

```
%sql SELECT sum(payload_mass__kg_) as sum_payload from SPACEXTBL where (customer) = 'NASA (CRS)'
```

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

sum_payload
45596

We calculated the total payload carried by boosters from NASA (CRS) by SUM function and presented the query result as 45596.

Average Payload Mass by F9 v1.1

```
%sql SELECT avg(payload_mass__kg_) as average_payload from SPACEXTBL where (booster_version) = 'F9 v1.1'
```

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

average_payload
2928.4

We calculated the average payload mass carried by booster version F9 v1.1 and presented the query result as 2928.4

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) AS FIRST_DATE_SUCCESS_GROUND_PAD FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

FIRST_DATE_SUCCESS_GROUND_PAD
2015-12-22

We found the dates of the first successful landing outcome on ground pad and presented the query result as 22nd December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION, LANDING_OUTCOME, PAYLOAD_MASS__KG_ FROM SPACEXTBL\
WHERE LANDING_OUTCOME='Success (drone ship)' AND (PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000);
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

In query we used WHERE clause to filter the booster versions which have successfully landed on drone ship and had payload mass between 4000 and 6000 KG

Total Number of Successful and Failure Mission Outcomes

```
%%sql SELECT MISSION_OUTCOME, COUNT(*) AS TOTAL_NUMBER FROM SPACEXTBL
WHERE MISSION_OUTCOME LIKE 'Failure%' OR MISSION_OUTCOME LIKE 'Success%' GROUP BY MISSION_OUTCOME;
```

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

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

We calculated the total number of successful and failure mission outcomes and presented the query result as 100 Successful and 1 Failure Mission Outcome

Boosters Carried Maximum Payload

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

We listed the names of the boosters which have carried the maximum payload mass and presented the 12 query results

2015 Launch Records

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
FROM SPACEXTBL \
where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

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

We listed the failed landing outcomes in drone ship, their booster versions and launch site names for a year 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.

```
%%sql SELECT * FROM (SELECT LANDING_OUTCOME, DATE FROM SPACEXTBL WHERE LANDING_OUTCOME='Success (ground pad)'
                     OR LANDING_OUTCOME='Failure (drone ship)')
WHERE DATE > '2010-06-04' AND DATE < '2017-03-20' ORDER BY DATE DESC;
```

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

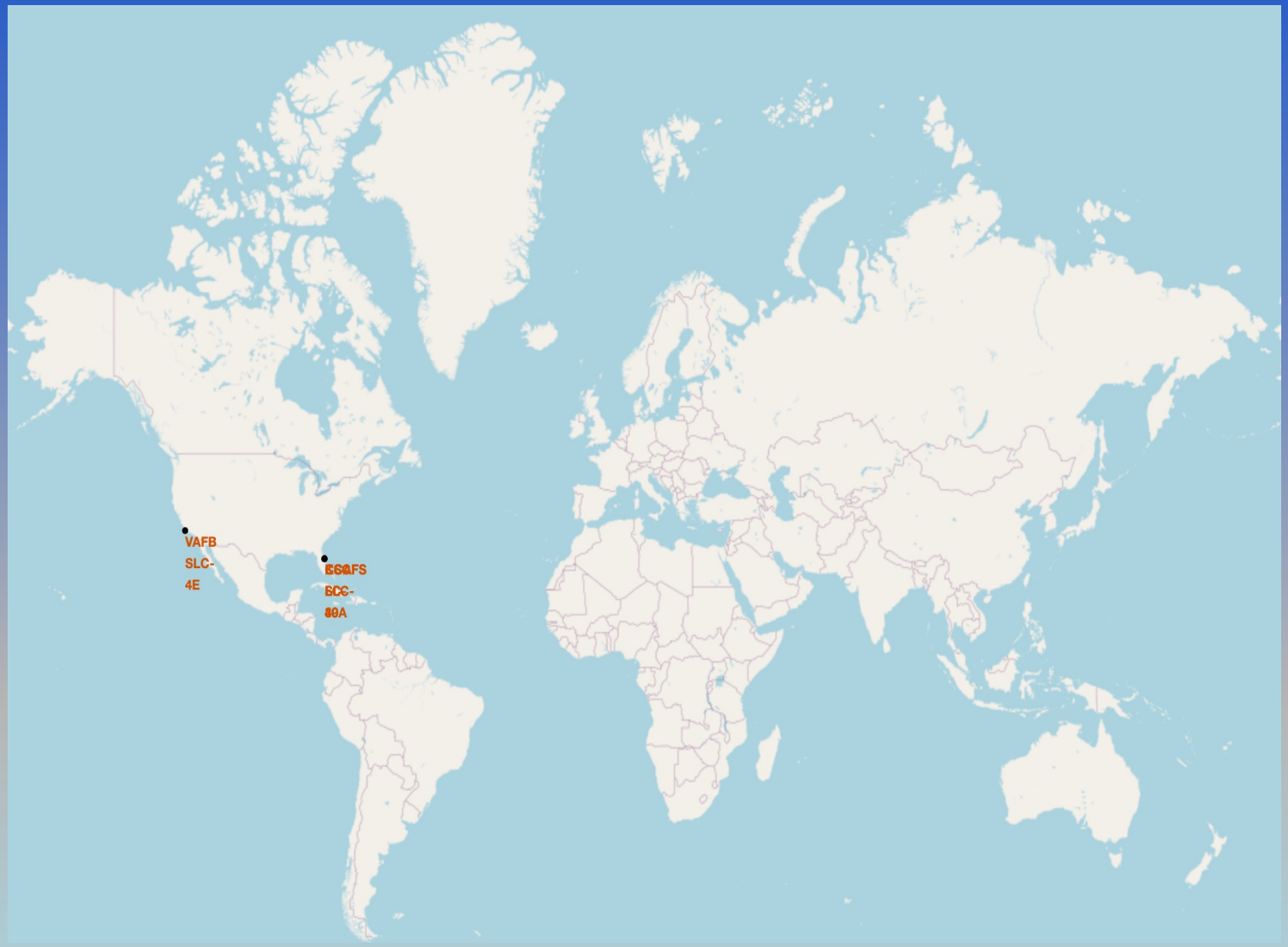
LANDING_OUTCOME	DATE
Success (ground pad)	2017-03-06
Success (ground pad)	2017-02-19
Success (ground pad)	2017-01-05
Success (ground pad)	2016-07-18
Failure (drone ship)	2016-06-15
Failure (drone ship)	2016-04-03
Failure (drone ship)	2016-01-17
Success (ground pad)	2015-12-22
Failure (drone ship)	2015-10-01
Failure (drone ship)	2015-04-14

We ranked the counts of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the 2010-06-04 and 2017-03-20 dates in descending order

Section 3

Launch Sites Proximities Analysis

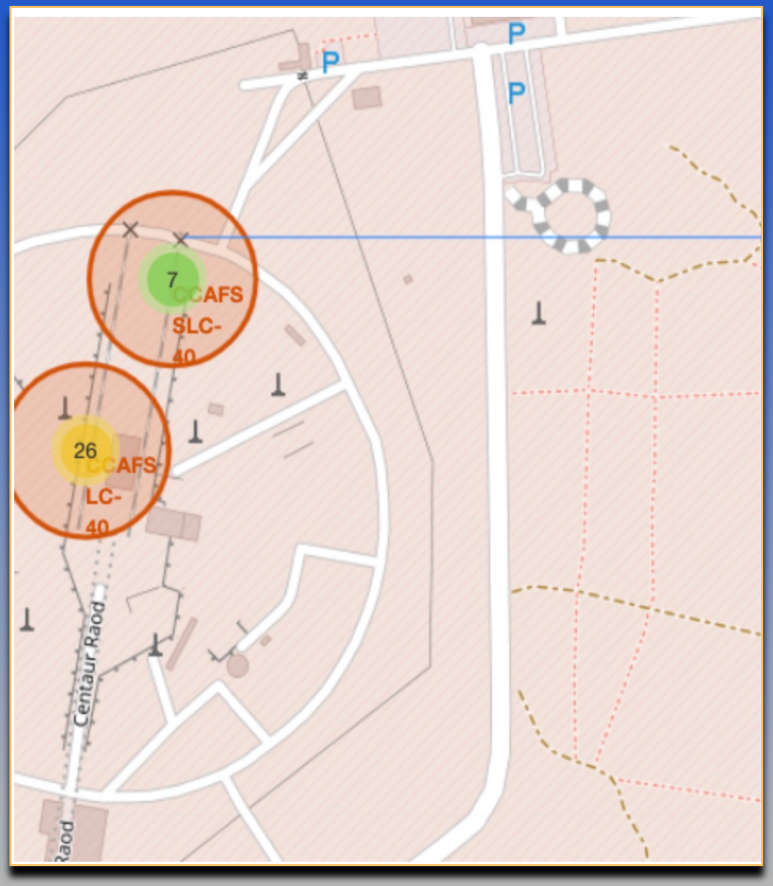
All Launch Sites on a global map

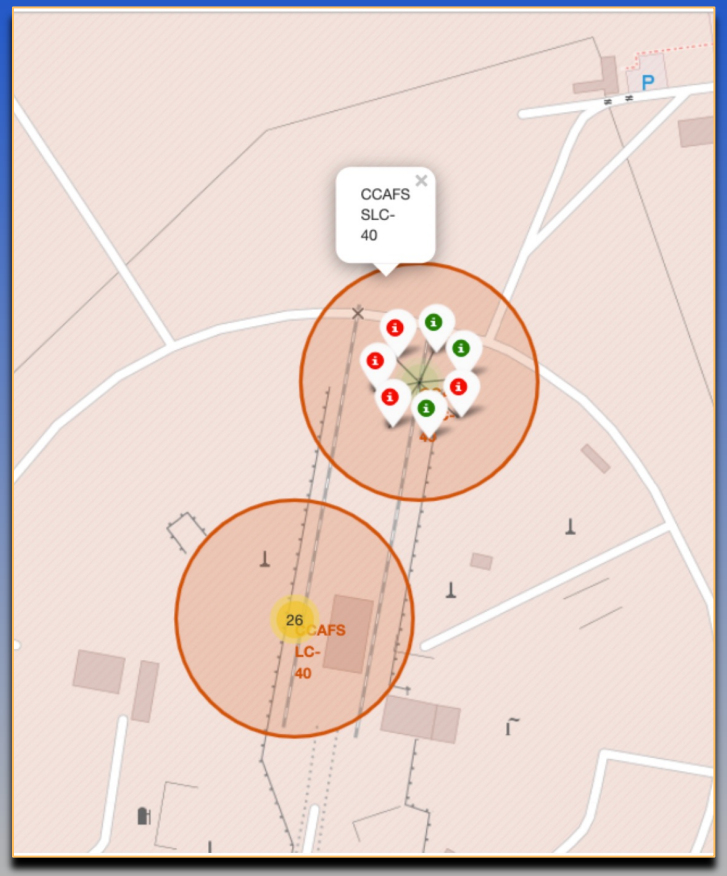


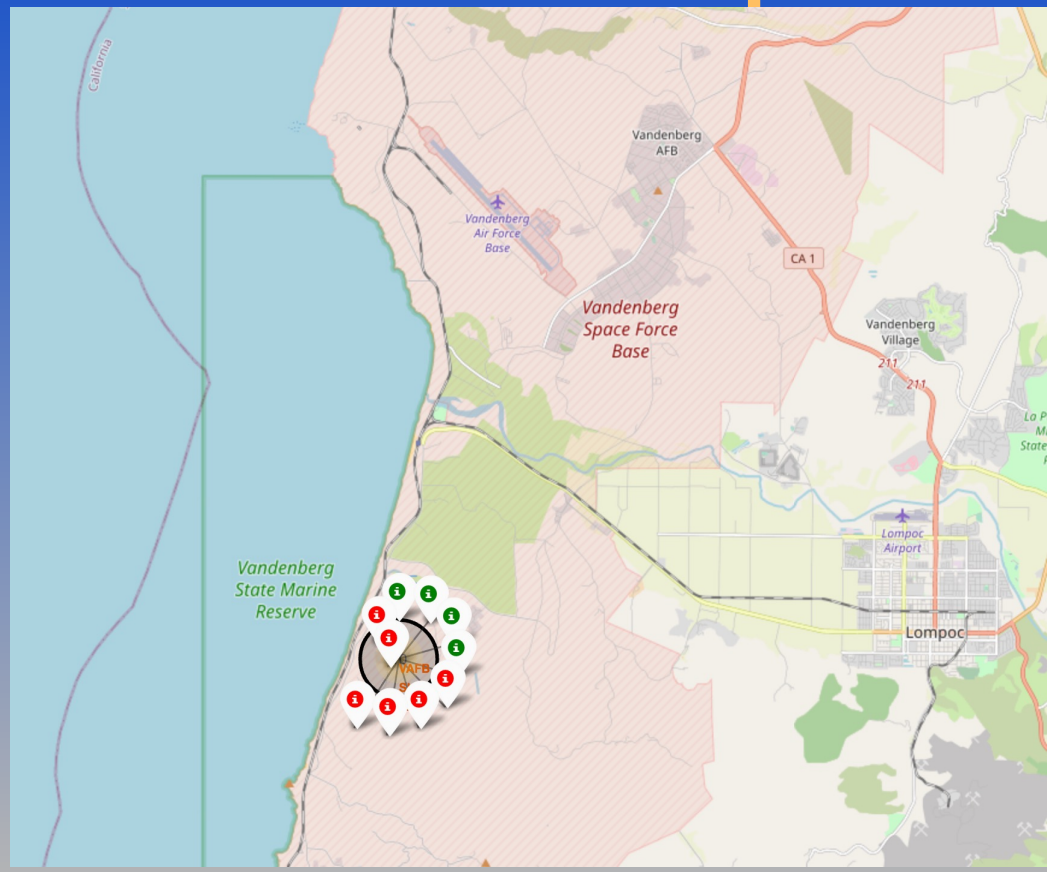
We explored the generated folium map and made a screenshot which include all launch sites' location markers on a global map

We can also observe that all SpaceX launches happened in the United States of America's Florida and California coasts which makes it safe for people in the case of ships exploding risks

Launch records with color labels on the map

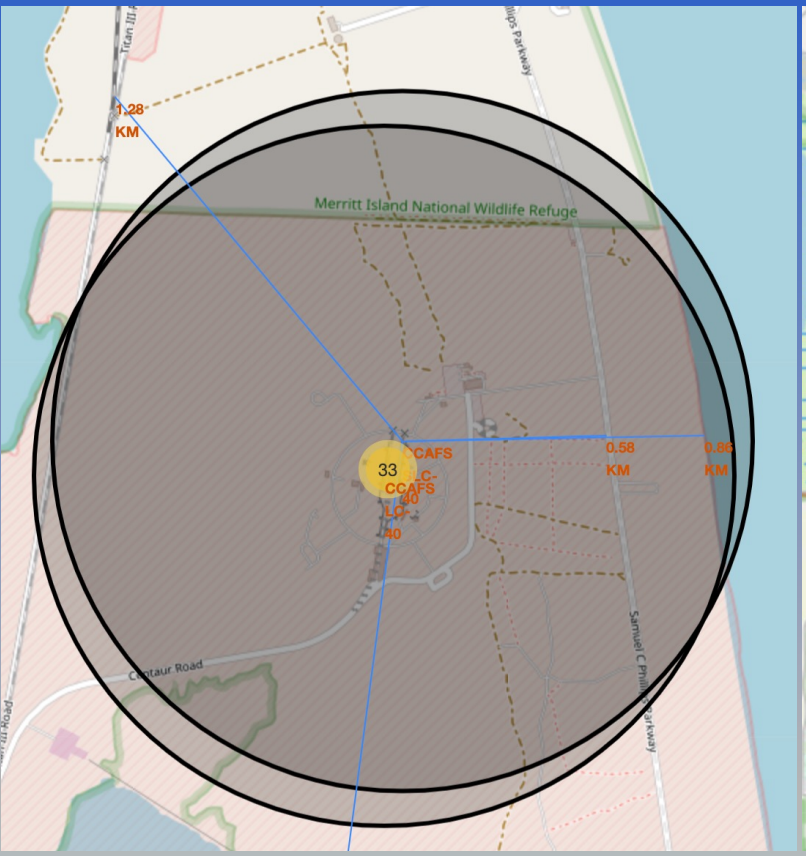


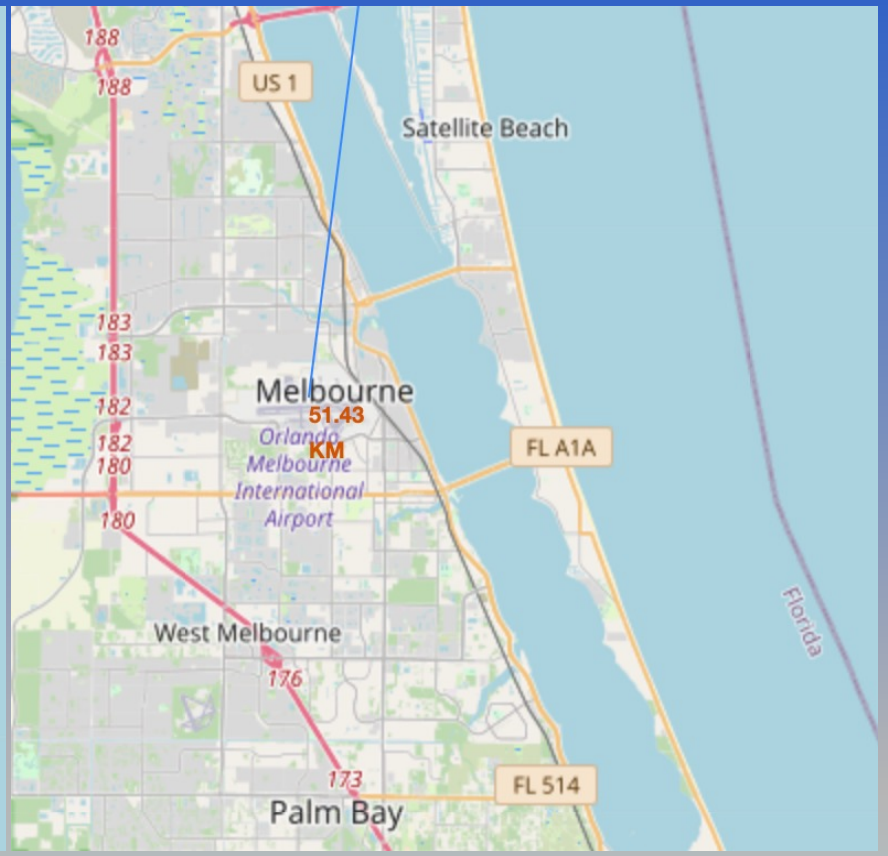




In the given folium map we showed the color-labeled CCAFS SLC-40 launch site in Florida and VAFB SLC-4E in California where **Green Markers** showed Successful launches and **Red Markers** showed Failures.

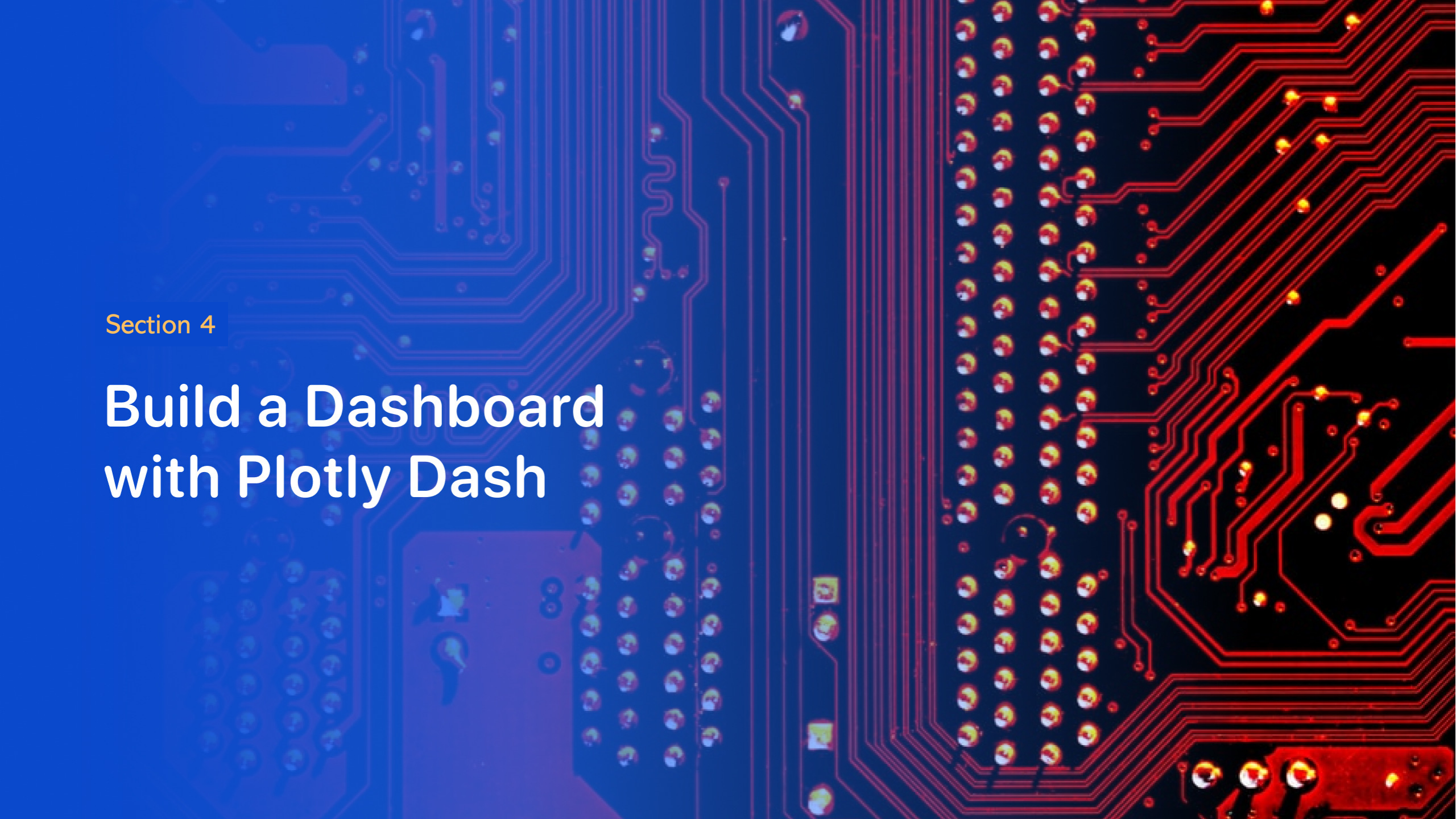
Distance from the CCAFS SLC-40 Lunch Site to its proximities





There was generated folium map which shows selected launch site CCAFS SLC-40 to its proximities as following:

- 1.28 km to the railway
- 0.58 km to the highway
- 0.86 km to the coastline
- 51.43 km km to the nearest city Melbourne


Section 4
Build a Dashboard with Plotly Dash

Total Success Launches count for all sites

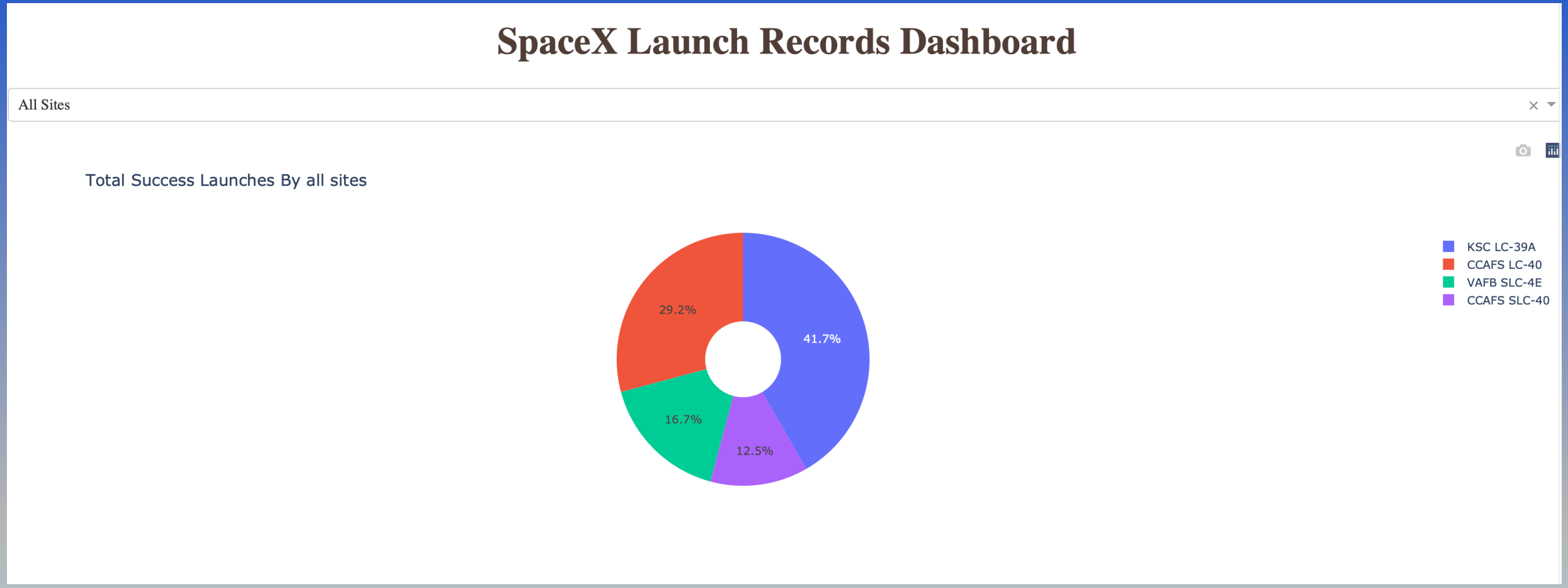
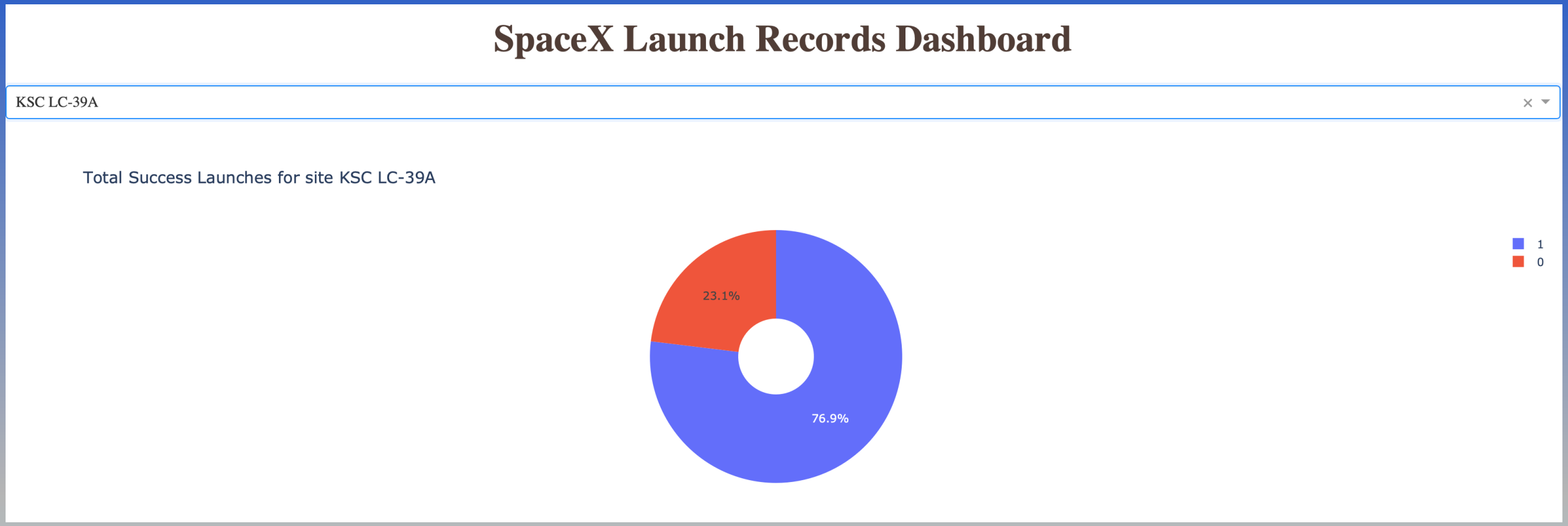


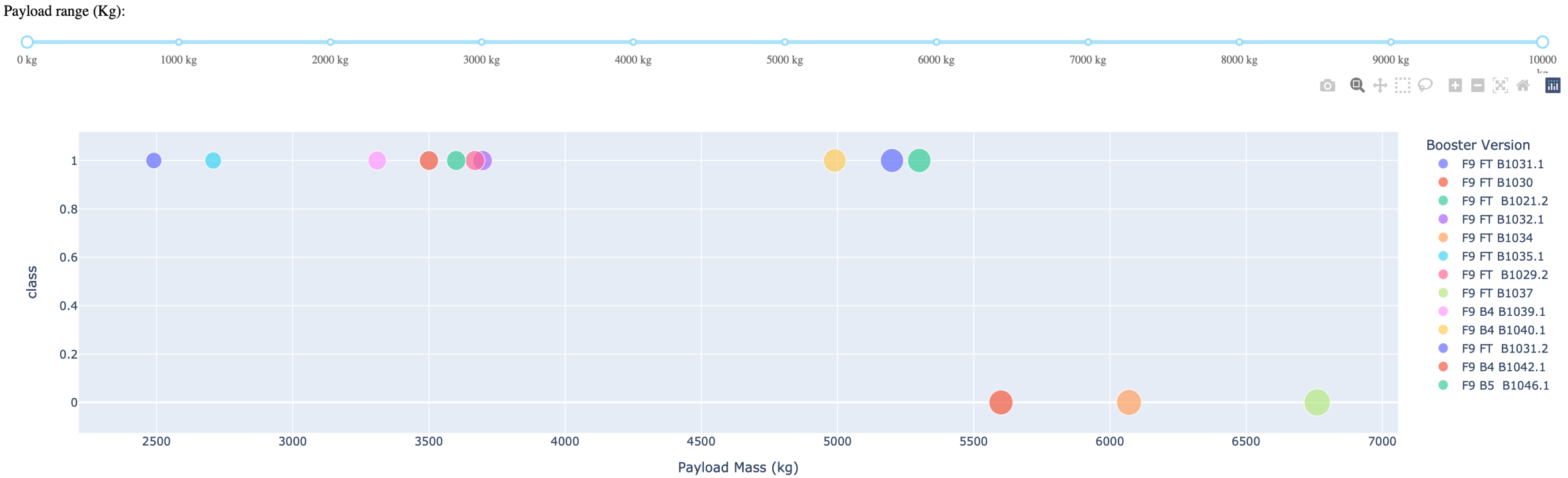
Chart represents the distribution of success rate among the launch sites where the biggest part is contained by KSC LC- 39A

Launch Site with the highest launch success ratio



KSC LC-39A has the highest launch success ratio with 76.9% and with 10 successful and only 3 failed landings

Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



The chart shows that the payloads between 2500 kg - 5800 kg along with the FT boosters have the highest success rate

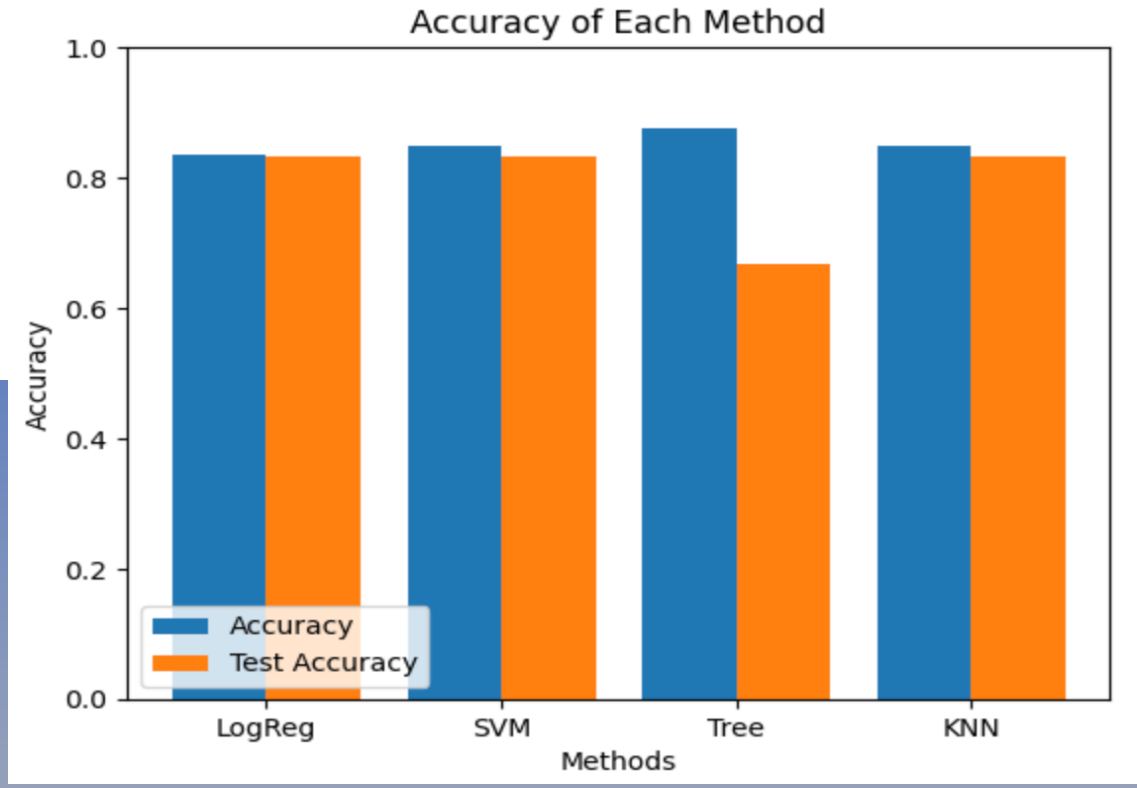
Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
Best Algorithm is Tree with a score of 0.8752380952380954
Best Params is : {'criterion': 'entropy', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'random'}
```

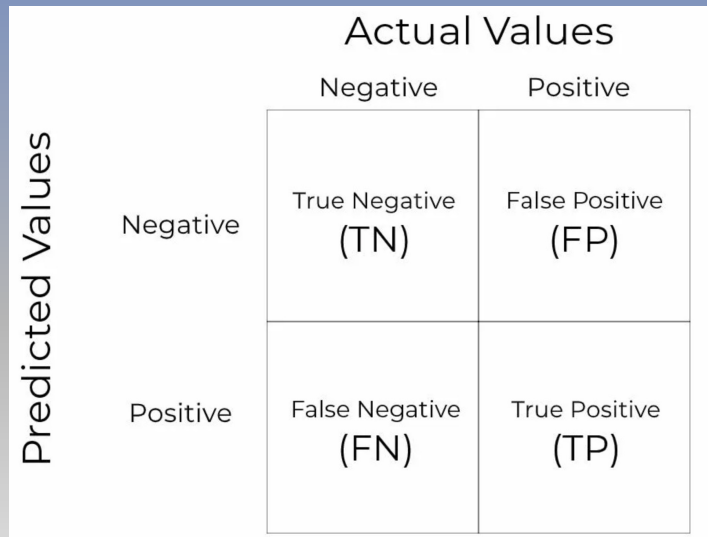
Model	Accuracy	TestAccuracy
LogReg	0.83429	0.83333
SVM	0.84821	0.83333
Tree	0.87524	0.66667
KNN	0.84821	0.83333

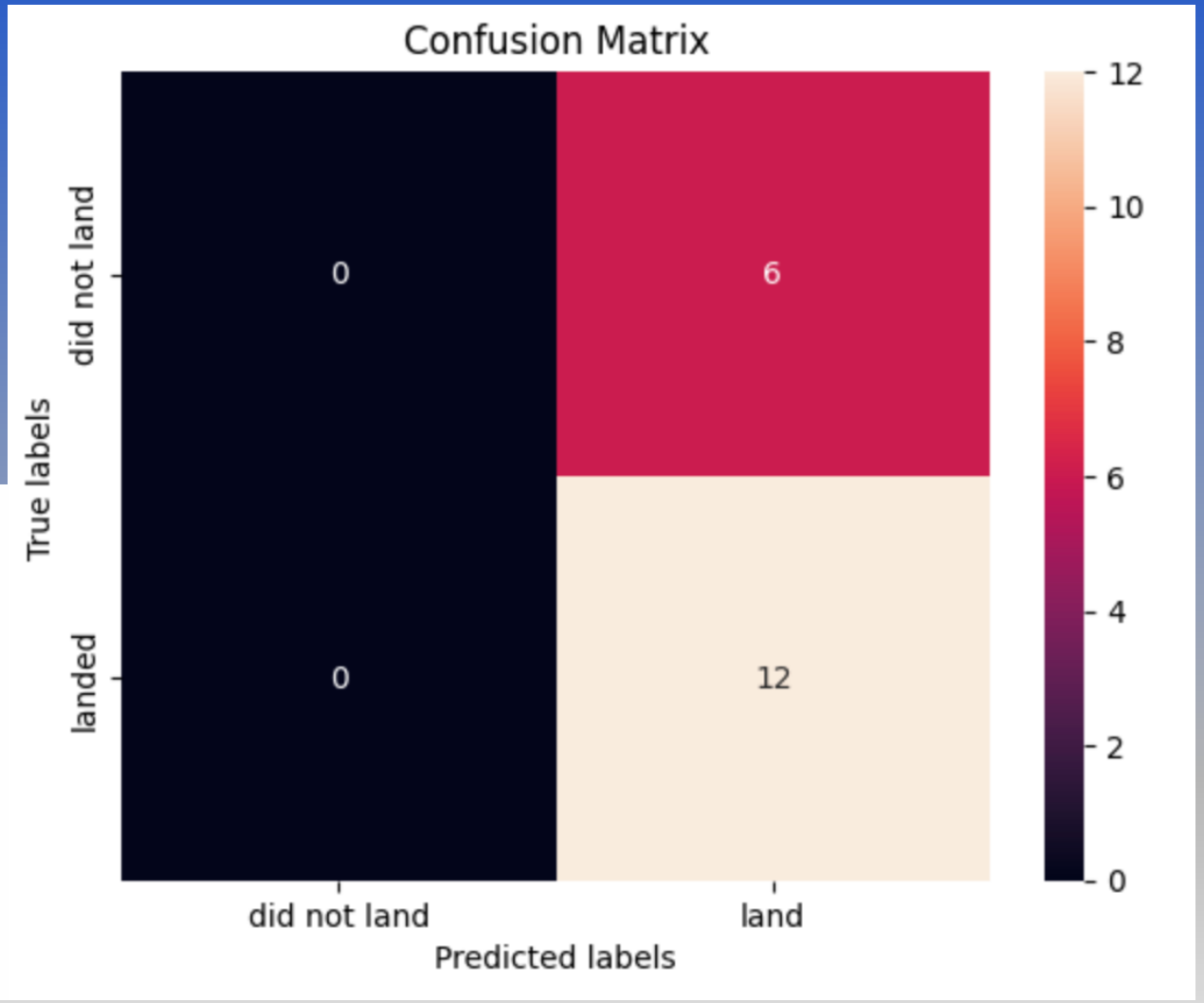


Based on the calculation of each model we can say that tree algorithm has the highest classification accuracy in contrast to the LogReg, SVM or KNN classification models

Confusion Matrix

The confusion matrix for the decision tree classifier shows that it can distinguish between different classes. Moreover, as a major problem we can determine the False Positive according the confusion matrix understanding table below





Conclusions

Different data sources were analyzed, and conclusion of the project is set as follows:

The larger the flight amount at a launch site, the greater the success rate at a launch site;

Launches with less payload mass performs better results rather than heavy payload mass launches;

Success rate of all launches had increased over the passed years;

Orbits ES-L1, GEO, HEO and SSO had the highest success rate;

KSC LC-39A had the best Launch site among the rest of all;

Launches above 7000 kg are less risky than others;

Decision Tree Classifier can be accepted as a best predicting ML algorithm for this dataset.

Appendix

Python Cheat Sheet: The Basics

Working with Data in Python Cheat Sheet

Pandas Documentation

NumPy Documentation

Web Scraping

API's and Data Collection Cheat Sheet

Data Preprocessing Tasks in Pandas & Plot Libraries Cheat Sheet

Plotting with Matplotlib using Pandas Cheat Sheet

Maps, Waffles, WordCloud and Seaborn Cheat Sheet

Dashboarding tools, Dash Python User Guide, **Plotly and Dash Cheat Sheet**

Thank you!