IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Le Cong Binh
27/01/2023

Outline

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

Executive Summary

- Summary of methodologies
- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building interactive map with Folium
- Building Dashboard with Plotly Dash
- Predict analysis (Classification)
- Summary of all results
- Exploratory Data Analysis Results
- Interactive analytics demo in screenshots
- Predictions data analysis results

Introduction

- **Project background and context**

The SpaceX is the most successful company of the commerical space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with cost 62 millions dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Therefore, if can determine if the first stage will land, we can determine the cost of launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- **Problems you want to find answers**

- *How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the stage landing?*
- *Does the rate of successful landings increase over the years?*
- *What is the best algorithm that can be used for binary classification in this case?*

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
- Perform data wrangling
  - Using web Scraping from Wikipedia Performed data wrangling – Filtering the data
  - Using One-hot-encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- Building, Tuning and evaluation classification models to ensure the best results

Data Collection

- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.
- **Data columns are obtained by using SpaceX REST API:**
- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, Block, Serial, Longitude, Latitide,Block, ReusedCount, Legs, GridFins
- **Data columns are obtained by using Wikipedia :**

- Web Scraping: Flights No, LaunchSite, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

1.Requesting rocket launch data from SpaceX API

2. Decoding the response content using .json() and turning it into a dataframe using .json_normalize()

3. Requesting needed information about the launches from SpaceX API by applying custom function

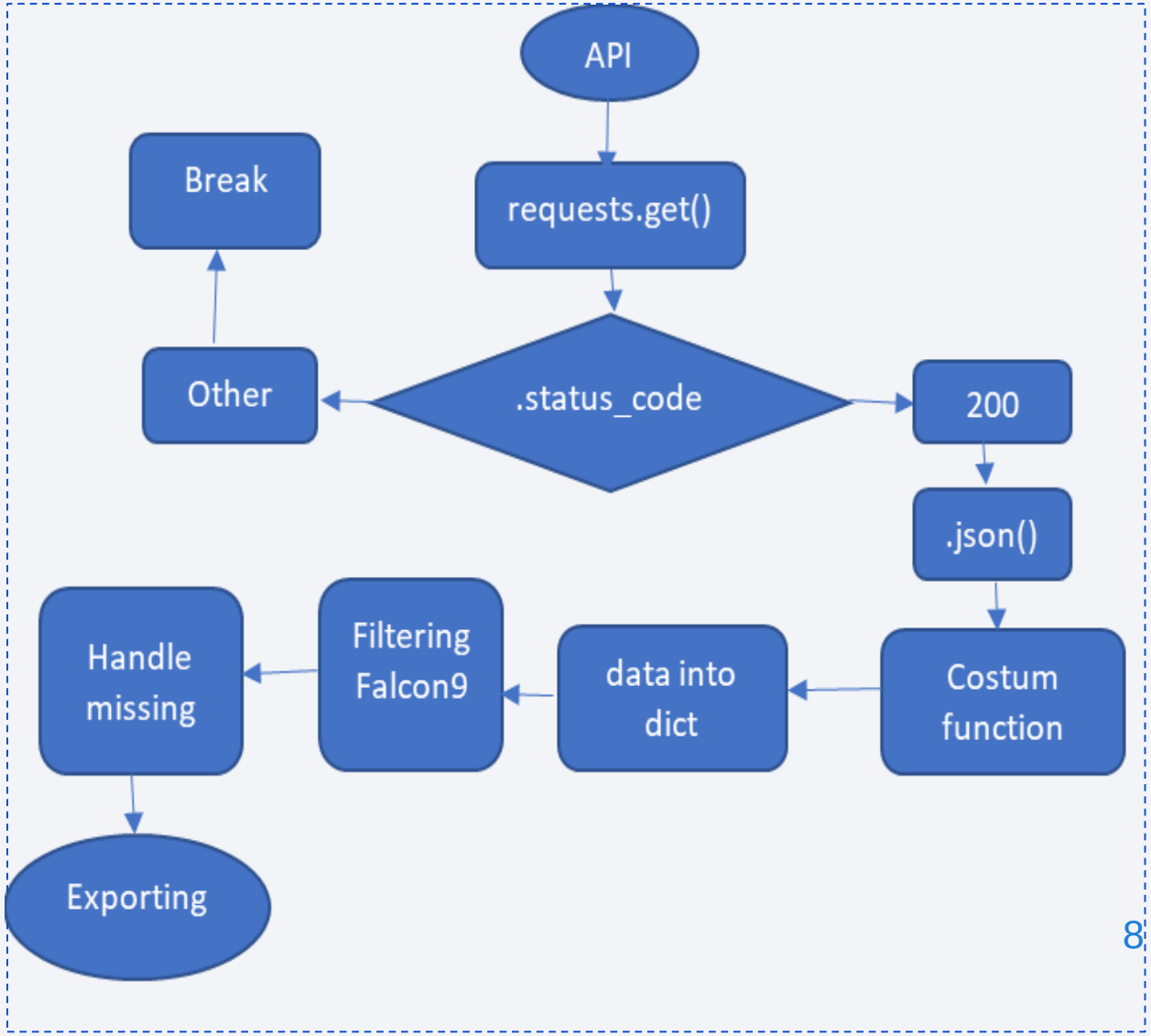
4. Constructing data we have obtained into a dictionary

5. Flitering the Dataframe to only include Falcon 9 launches

6. Replacing missing values of Payload Mass column with calculated .mean() for this column

7. Exporting the data to CSV

- https://github.com/BWBLF/DS_Captones



Data Collection - Scraping

1.Requesting Falcon 9 launch data from Wikipedia

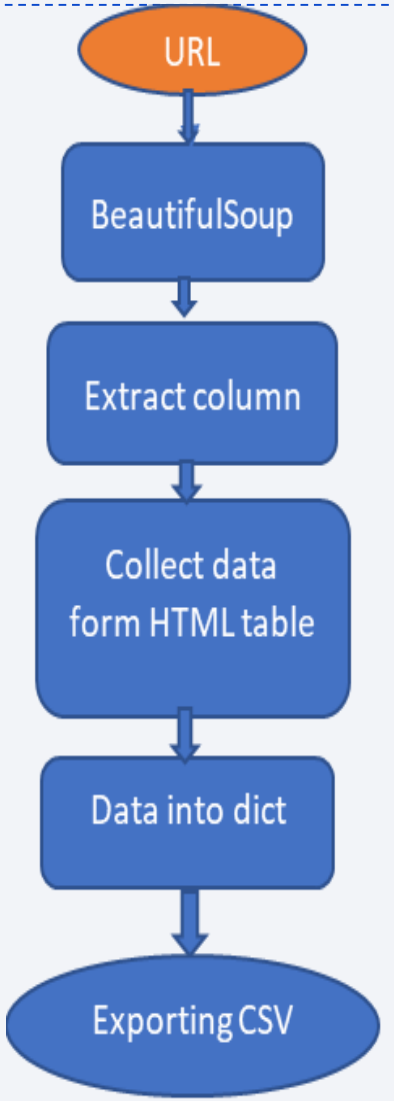
2. Creating a BeautifulSoup object from HTML response

3. Extracting all column names from HTML table header

4. Collecting the data by parsing HTML tables

5. Constructing data we have obtained into a dictionary

6. Creating a dataframe from dictionary



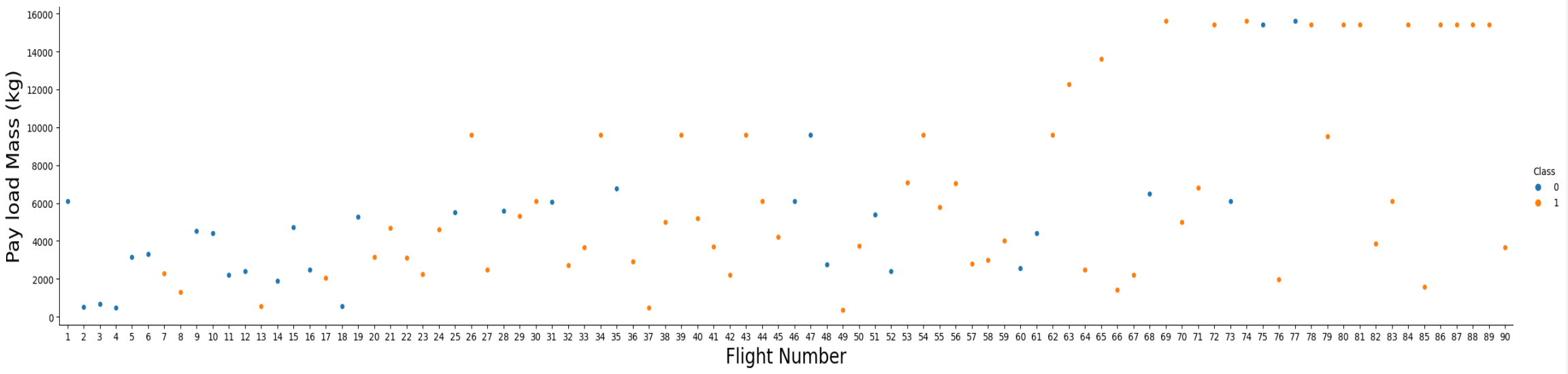
- https://github.com/BWBLF/DS_Captones/blob/master/web_scraping.ipynb

Data Wrangling

- In the dataset, there are several differnet cases where the booster did not land successful. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome ws successfully landed to a specific region of the ocean while False Ocean means the missions outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was unseccessfully landed to a ground pad. We mainly convert those outcomes into Training Labels with "1" means the booster successfully landed, "0" means it was unseccessfully.

- https://github.com/BWBLF/DS_Captones/blob/master/Wrang_ling.ipynb

EDA with Data Visualization



- We can see the relationship between Number Flights and PayloadMass. In the early stages of Falcon 9, the success rate is almost always very low, as the number of flights increases, the number of successes also improves. And Payload is also important, the increased payload means the Falcon 9's return is lower.
- https://github.com/BWBLF/DS_Captones/blob/master/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Query 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 successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- https://github.com/BWBLF/DS_Captones/blob/master/jupyter-labs-eda-sql-coursera_sqllite.ipynb

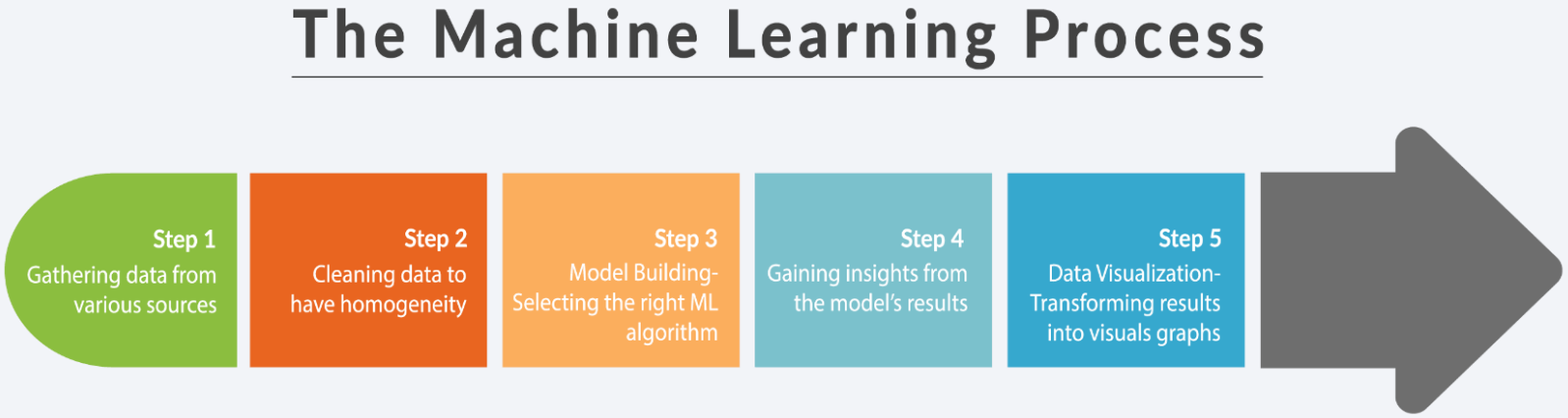
Build an Interactive Map with Folium

- Markers of all Launch Sites:
- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Coloured Markers of the launch outcomes for each Launch Site:
- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a Launch Site to its proximities:
- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its procimities like Railway, Highway, Coastline and Closest City.

Build a Dashboard with Plotly Dash

- Pie charts are added to the Dash Board to show and compare the success rates between Launch Sites, and see which Launch sites have the highest success rates.
- While Scatter shows the relationship between Payload mass and Version with success rate. Show us quickly and intuitively how Payload affects launch success rate.
- https://github.com/BWBLF/DS_Captones/blob/master/dash_interactivity.py

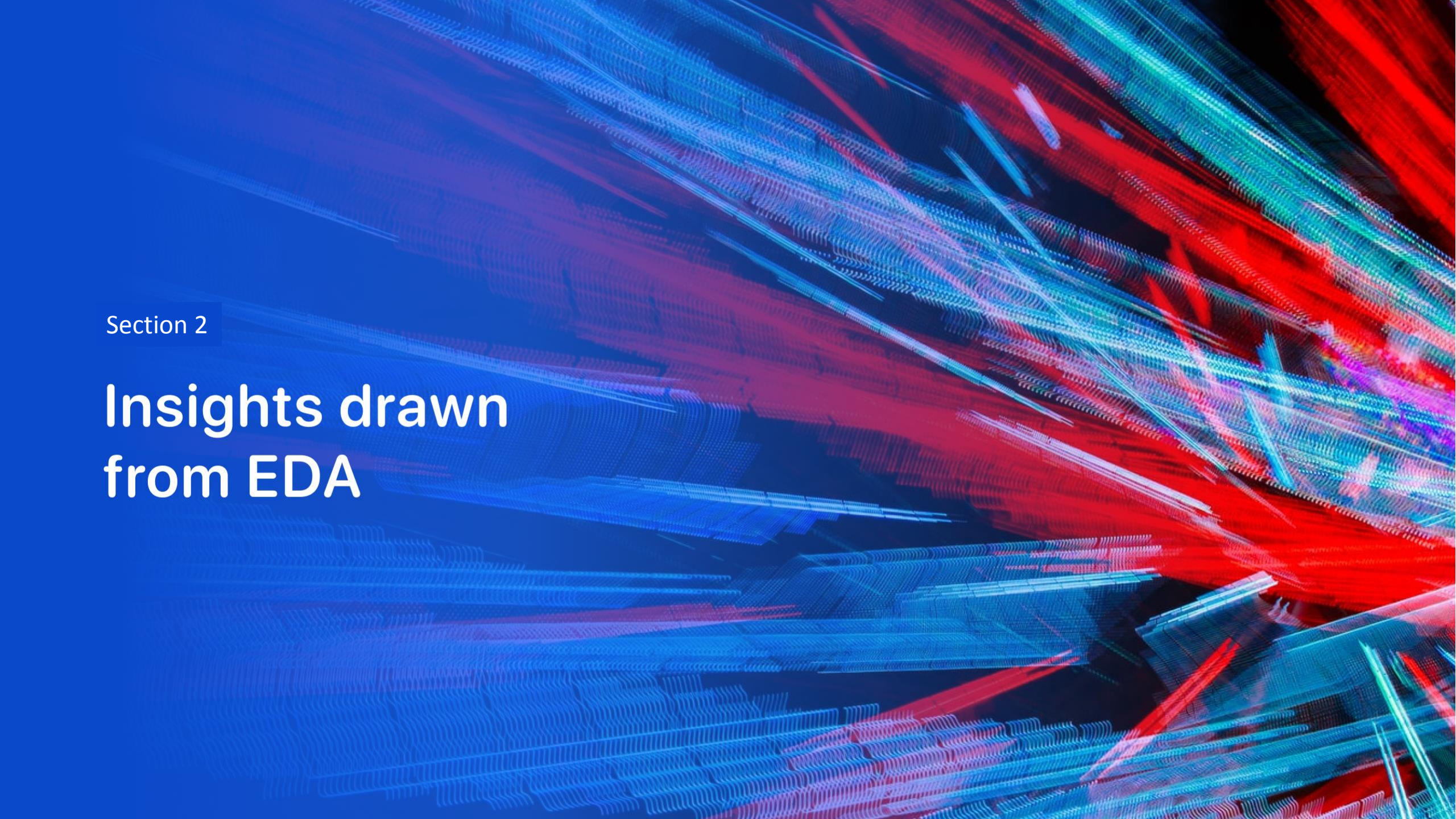
Predictive Analysis (Classification)



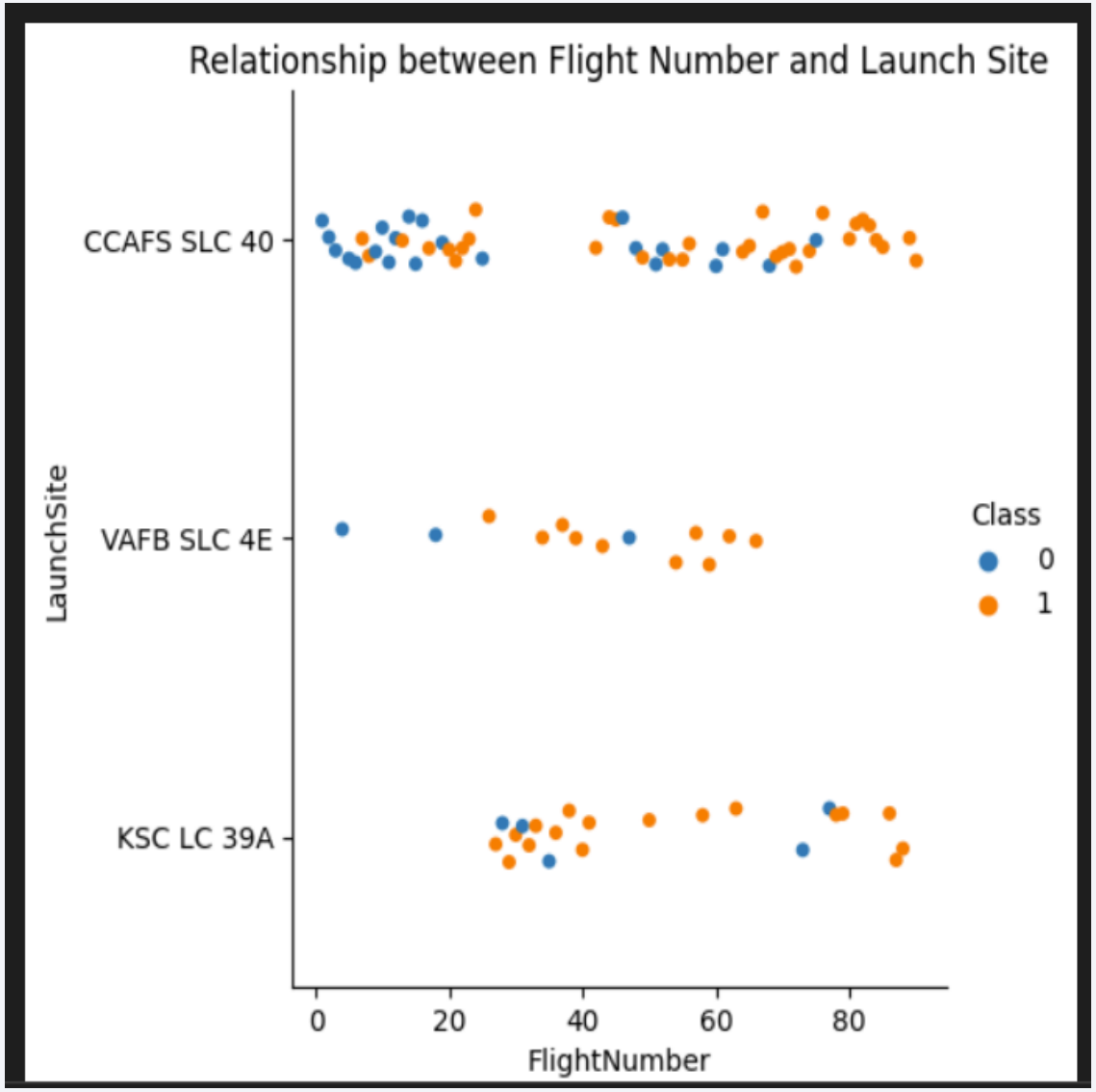
- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- https://github.com/BWBLF/DS_Captones/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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


Section 2
Insights drawn from EDA

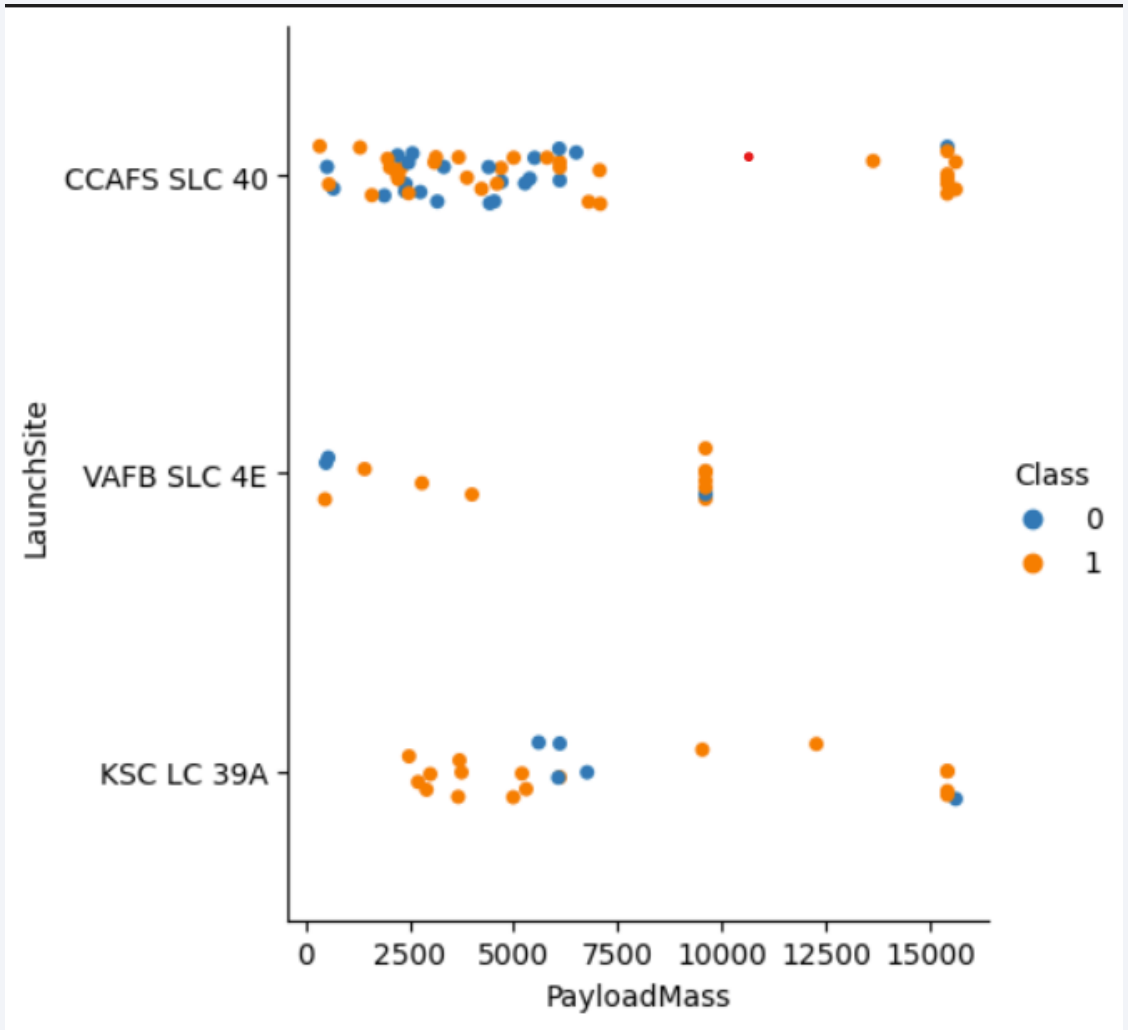
Flight Number vs. Launch Site



- All the earliest launches failed, while all the last launches were successful.
- The CCAFS SLC-40 launch site accounts for more than 50% of the launches here.
- At the first stage, in the first 20 launches, only 2 were performed at VAFB SLC 4E, the rest SpaceX mainly used CCAFS LC-40. And the graph also shows that as the number of launches increases, the success rate also improves.
- VAFB SLC 4E and KSC LC 39A have higher success rates.

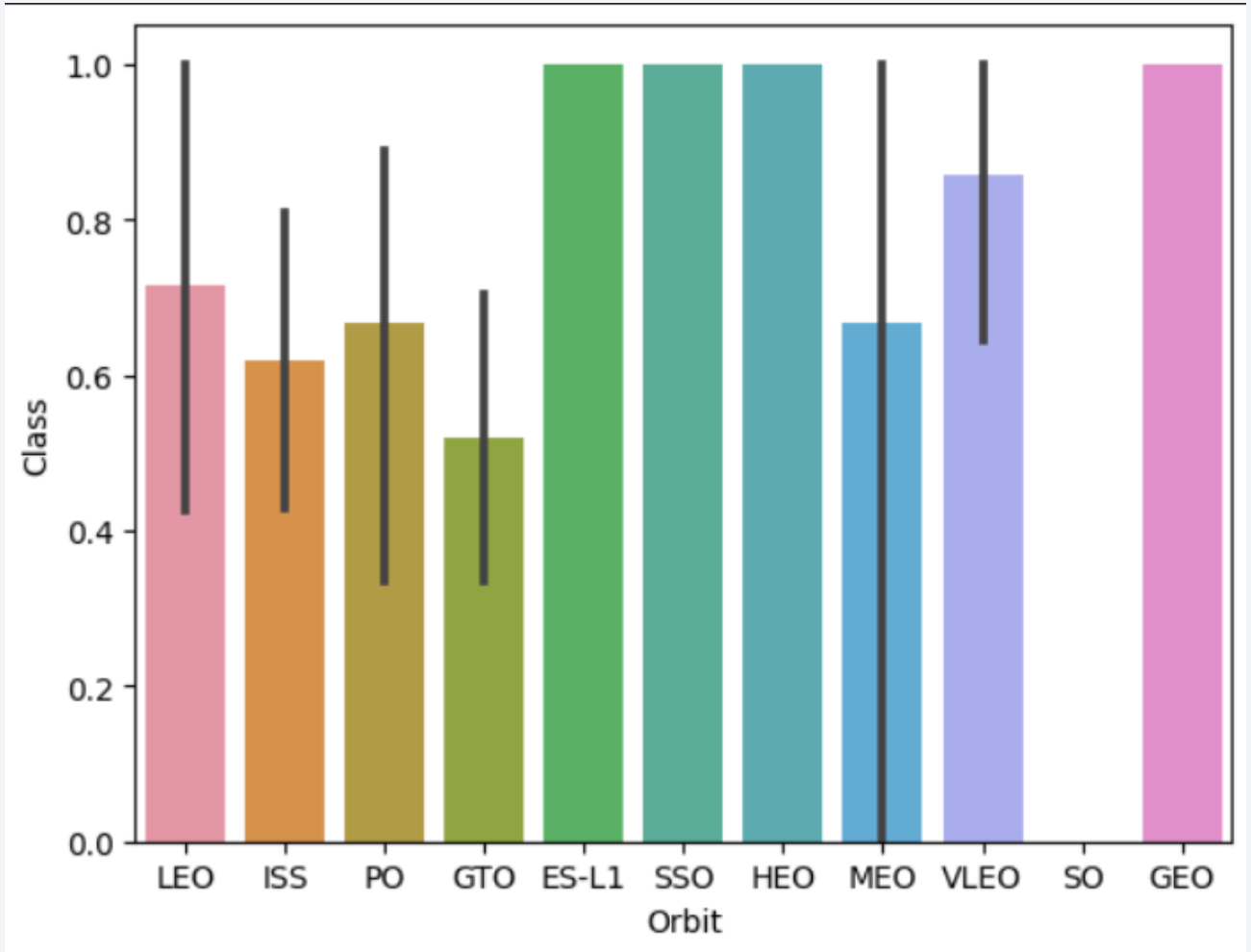
Payload vs. Launch Site

- The graph shows the relationship between Payload vs. Launch Site, most Payload is mainly under 7500kg. And an increased Payload shows a higher success rate than a low Payload.
- Most ofthe launches with payload ma ss over 7000kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500kg.



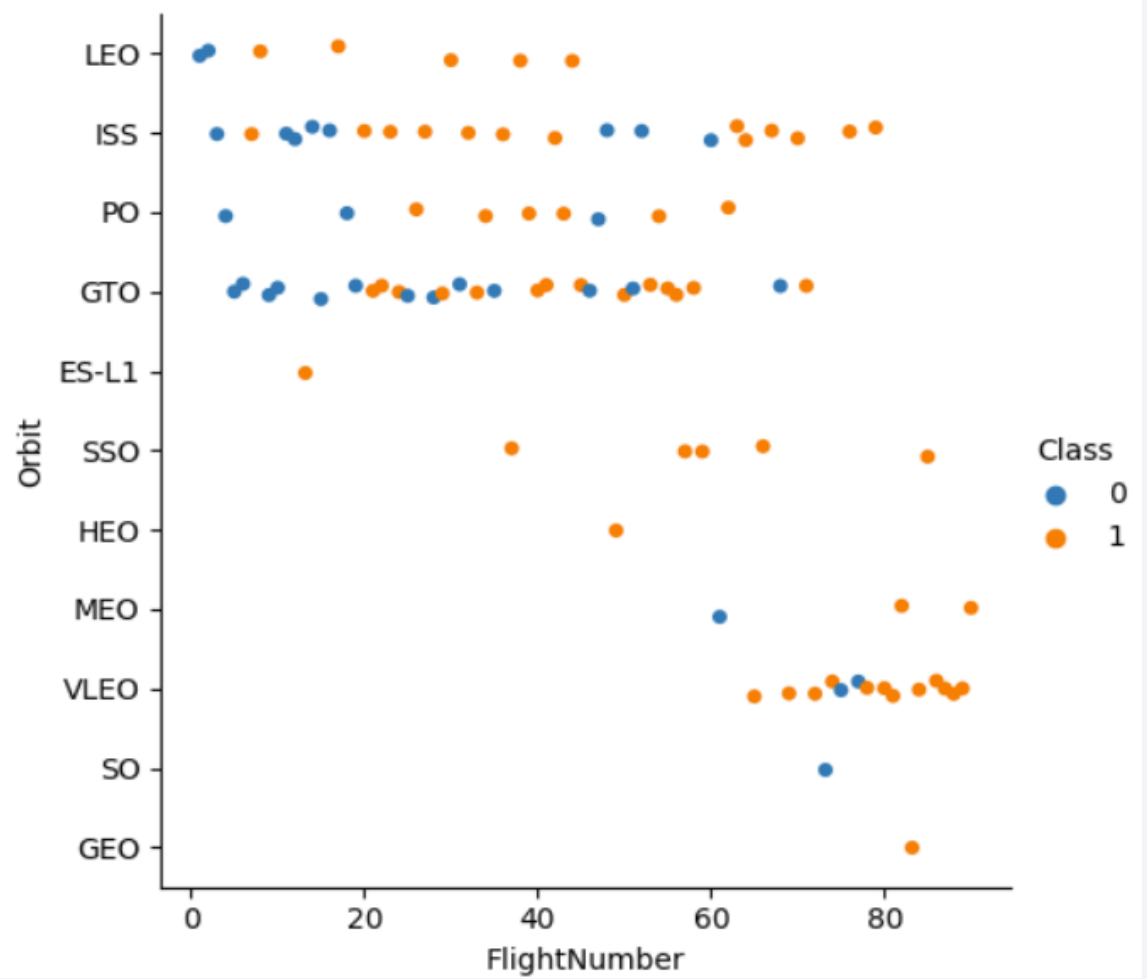
Success Rate vs. Orbit Type

- The success rate at Orbit ES-L1, SSO, HEO, GEO is 100%, while the success rate at SO is 0%. The remaining Orbits have a success rate of 50%-80%.



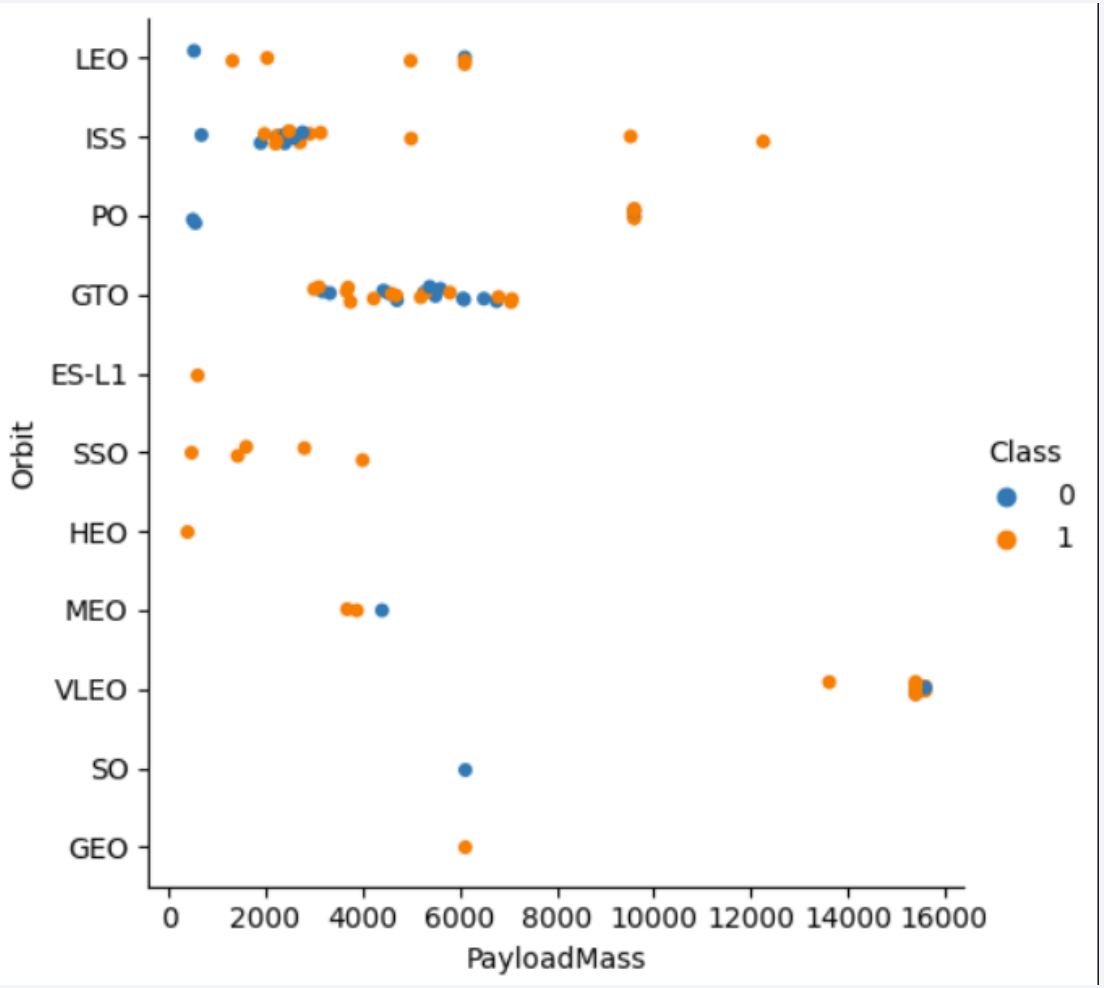
Flight Number vs. Orbit Type

The success rate at Orbit ES-L1, SSO, HEO, GEO is 100% but the number of launches in these Orbits is only from 1 to 5 launches. The launches are mainly in Orbits with a success rate of 50% to 80% (LEO, ISS, PO, GTO, VLEO Orbits)



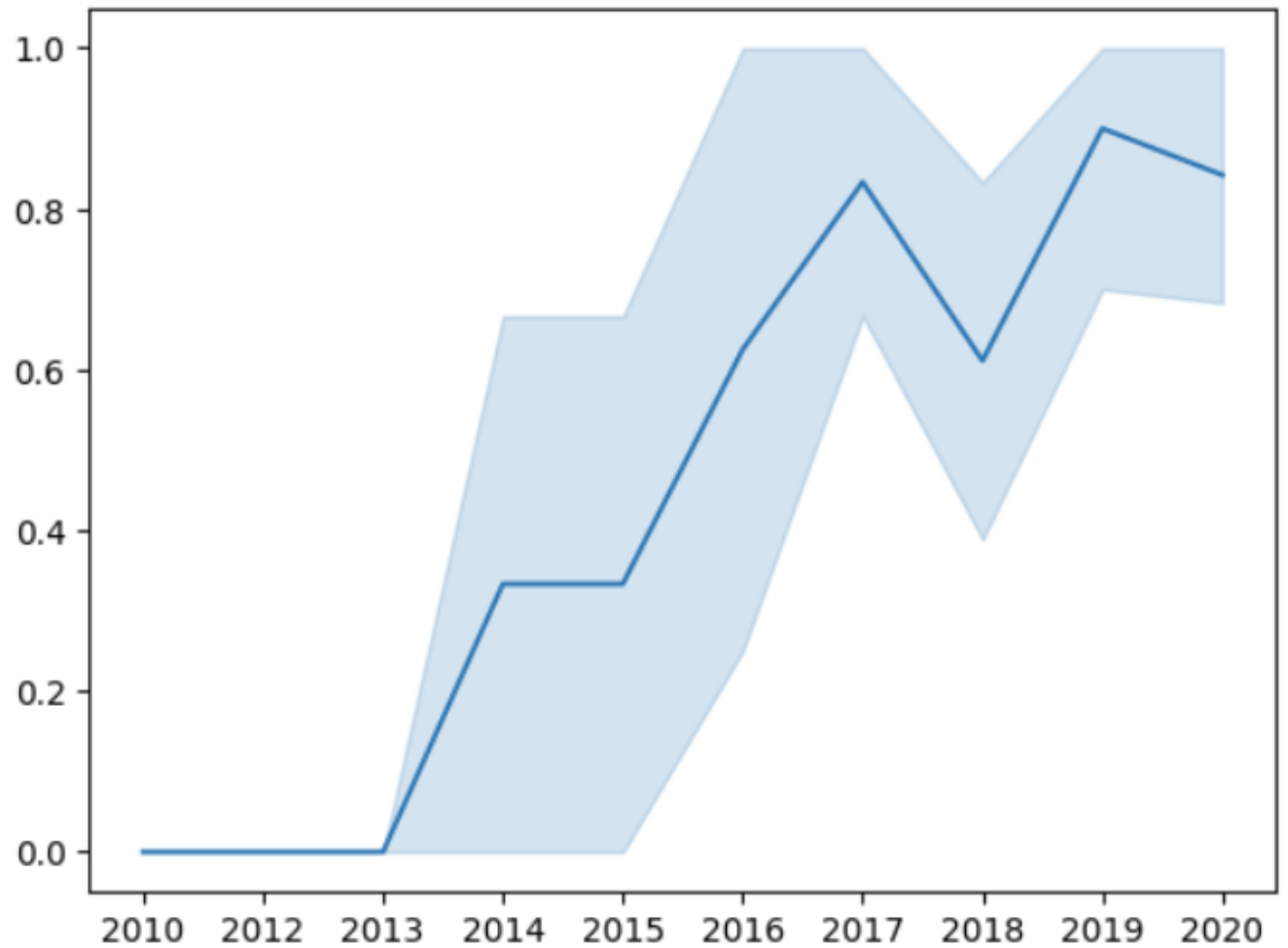
Payload vs. Orbit Type

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



Launch Success Yearly Trend

- From 2010 to 2013 there were no successful tests. Since 2013, the success rate of the years has improved over the years by more than 30% in 2014 and 80% in 2020. The highest in 2019 when the success rate is nearly 90%.



All Launch Site Names

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;
```

Python

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

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL LIMIT(5);
```

✓ 0.3s Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Present your query result with a short explanation here

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer =='NASA (CRS)';
```

✓ 0.8s Python

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

SUM(PAYLOAD_MASS__KG_)
45596

- Present your query result with a short explanation here

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE  SPACEXTBL.BOOSTER_VERSION like'%F9 v1.1%';
✓ 0.8s                                                                                    Python
 * sqlite:///my_data1.db
Done.

 AVG(PAYLOAD_MASS__KG_)
        2534.6666666666665
```

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

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) as First_successful_landing FROM SPACEXTBL WHERE SPACEXTBL.LANDING__OUTCOME='Success (ground pad)';
```

✓ 0.7s Python

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

First_successful_landing
01-05-2017

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME='Success (drone ship)' AND (PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000);
```

✓ 0.1s Python

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

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

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME,COUNT(*) as TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME ;
```

✓ 0.9s Python

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

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

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

Boosters Carried Maximum Payload

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

✓ 0.1s Python

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

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

- List the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
             where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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

```
%sql SELECT COUNT(*) as COUNT_OUTCOMES, LANDING__OUTCOME FROM SPACEXTBL WHERE  DATE BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY LANDING__OUTCOME ORDER BY CO
```

0.7s Python

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

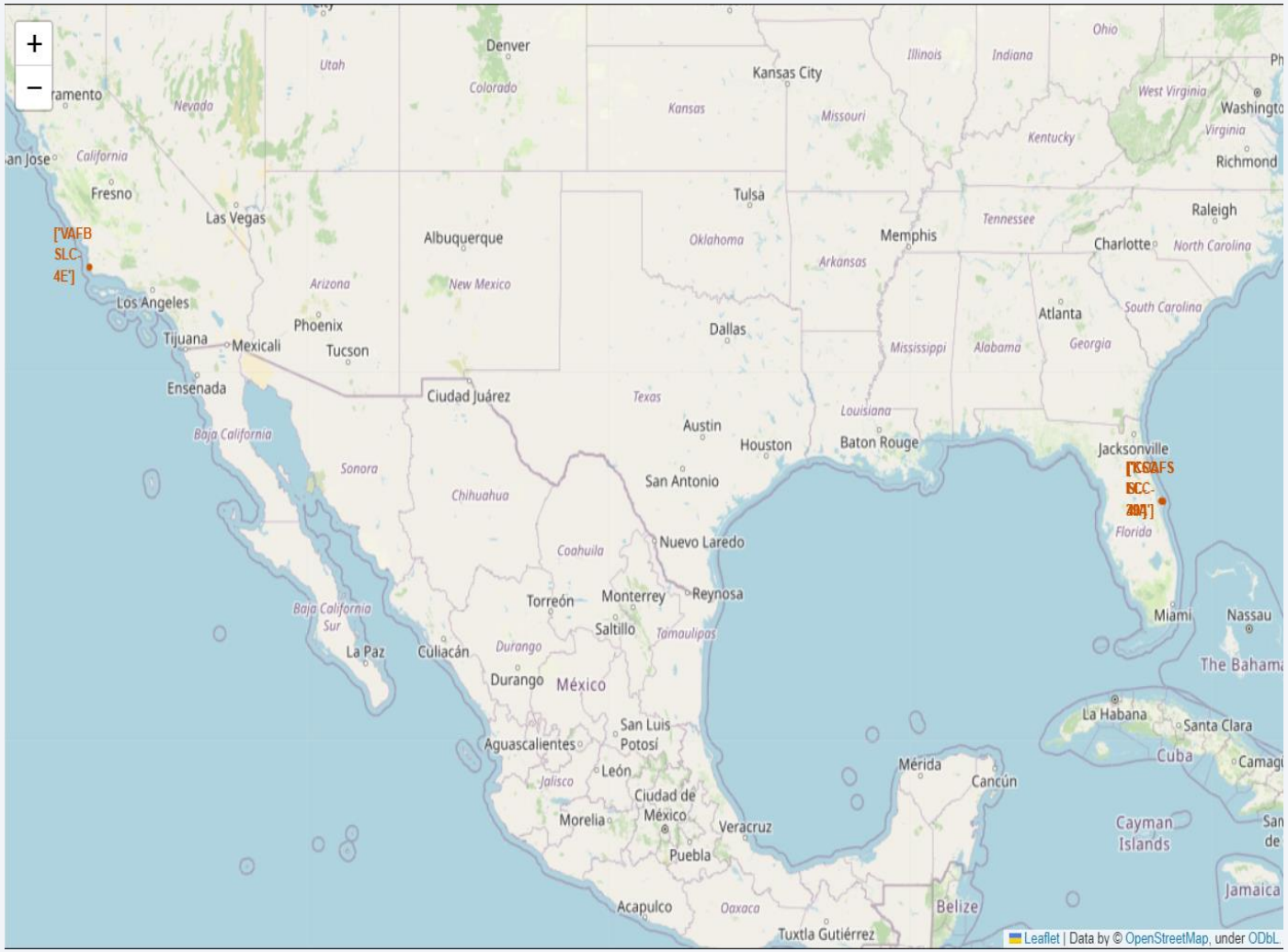
COUNT_OUTCOMES	LANDING__OUTCOME
20	Success
10	No attempt
8	Success (drone ship)
6	Success (ground pad)
4	Failure (drone ship)
3	Failure
3	Controlled (ocean)
2	Failure (parachute)
1	No attempt

- 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


Section 3
Launch Sites
Proximities Analysis

All Launch site's location markers on a global map



- Most of Launch sites are in proximity to the Equator line
- The land is moving faster at the equator than any other place on the surface of the Earth.
- All launch sites are very close proximity to the coast , while lauching rockets towards the ocean it minizes the risk of having debris dropping or exploding near people.

Colour-labeled launch records on the map

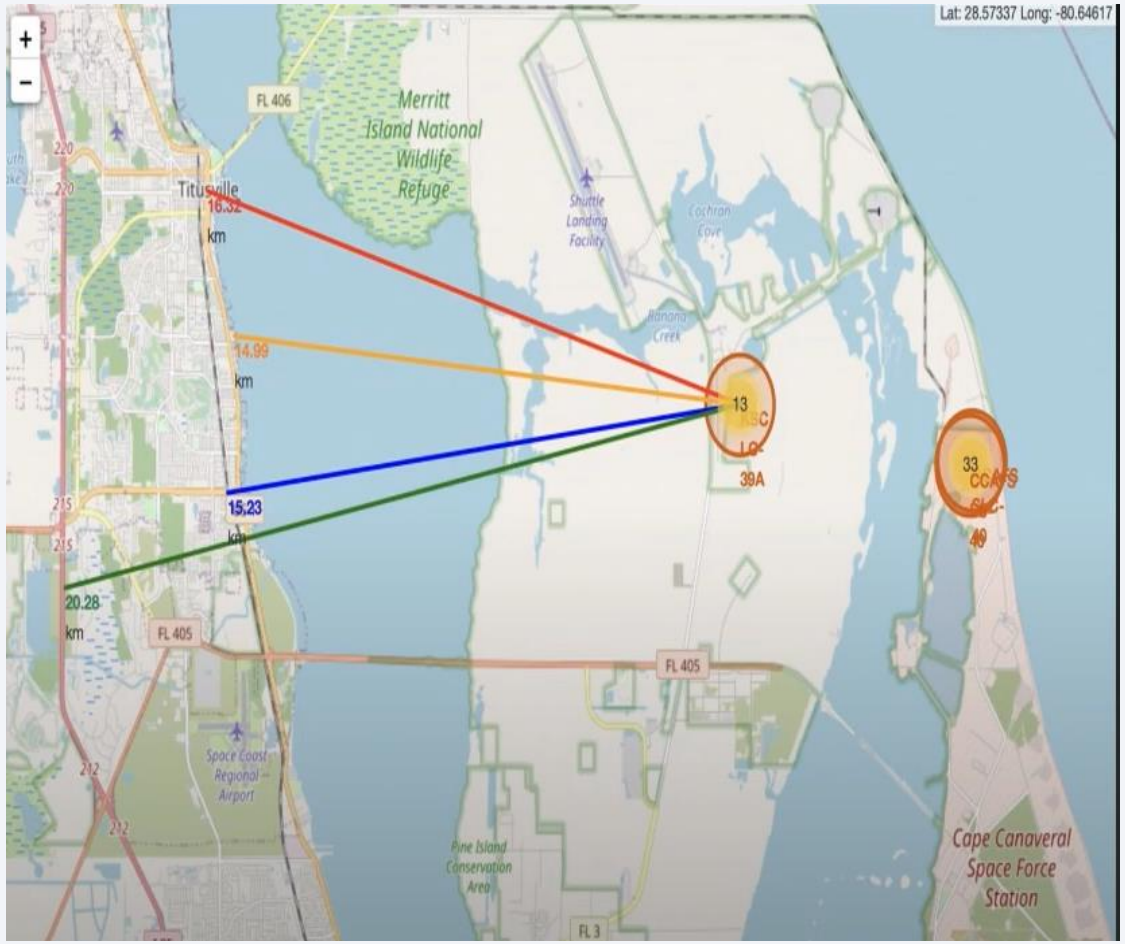
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

- Green Marker = Successful Launch

- Red Marker = Failed Launch

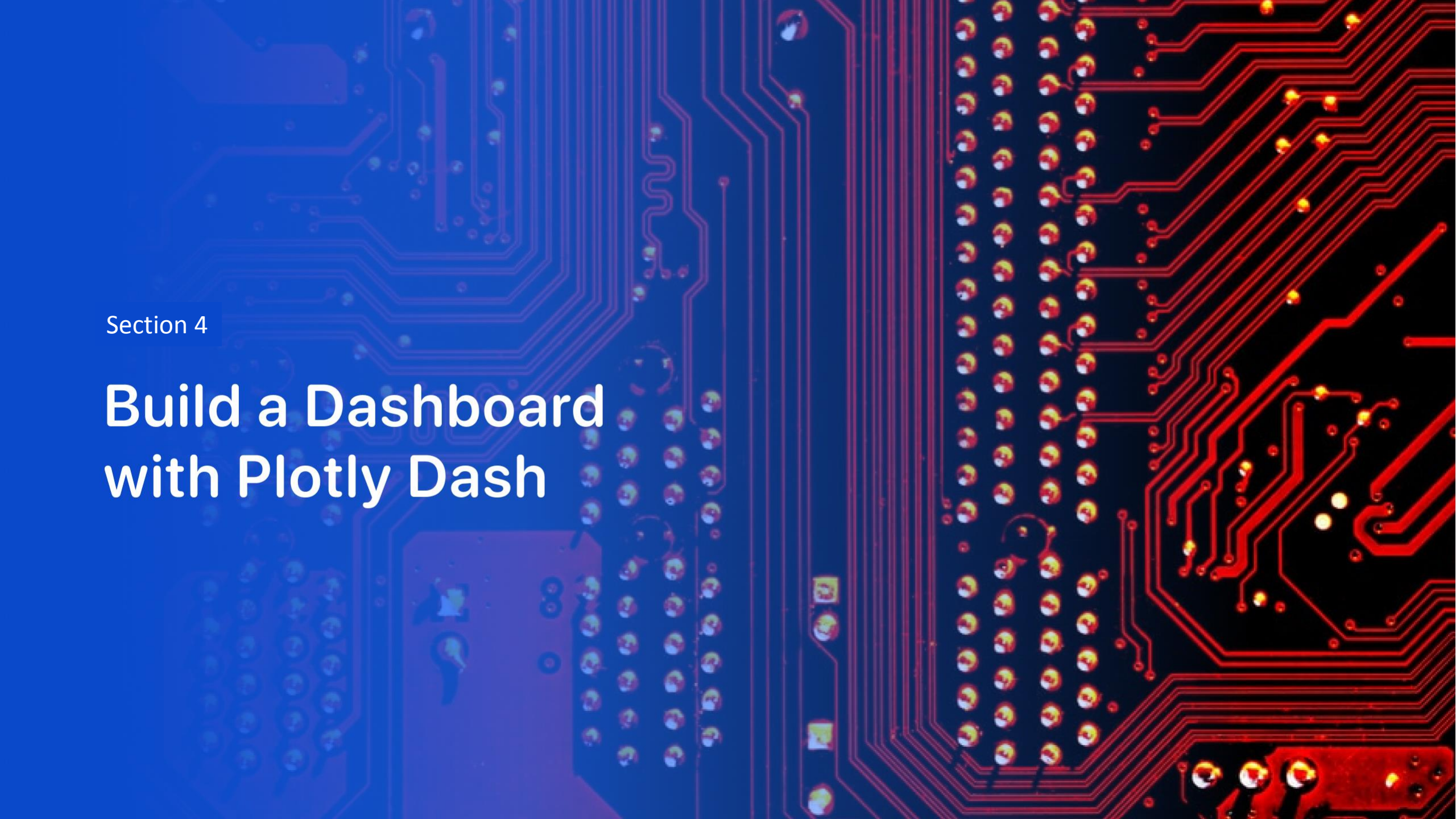
- Launch Site KSC LC-39A has a very high Success Rate.

Distance from the launch site KSC LC-39A to its proximities

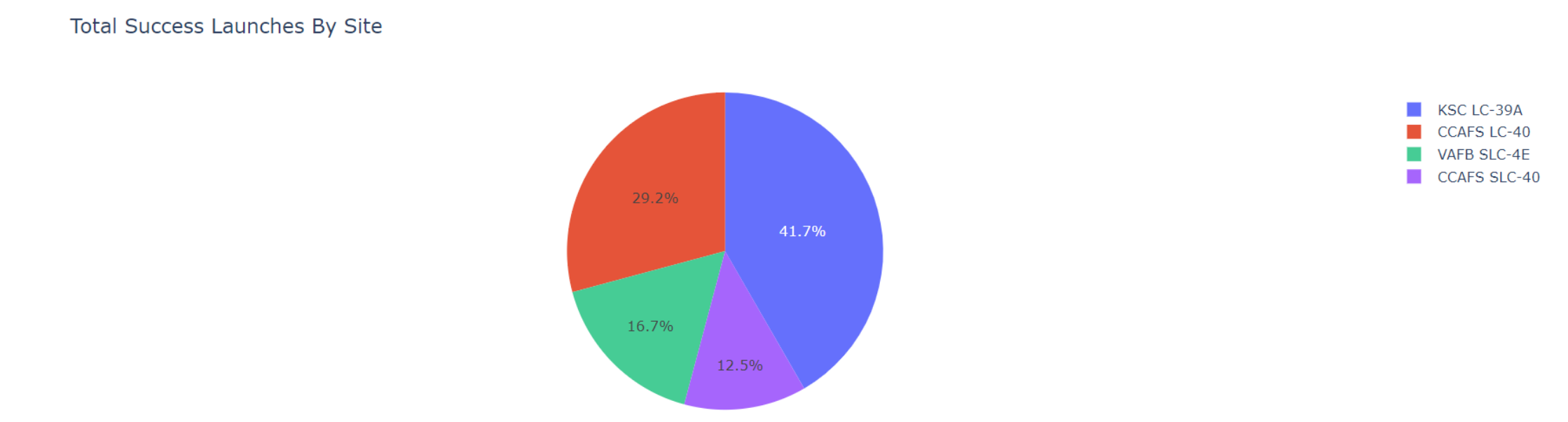


- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

- Relative close railway : 15.23km
- Relative close to highway: 20.28km
- Relative close to coastline: 14.99km
- Failed rocket with its high speed can cover distances like 15-20km in few seconds. It could be potentially dangerous to populated areas.

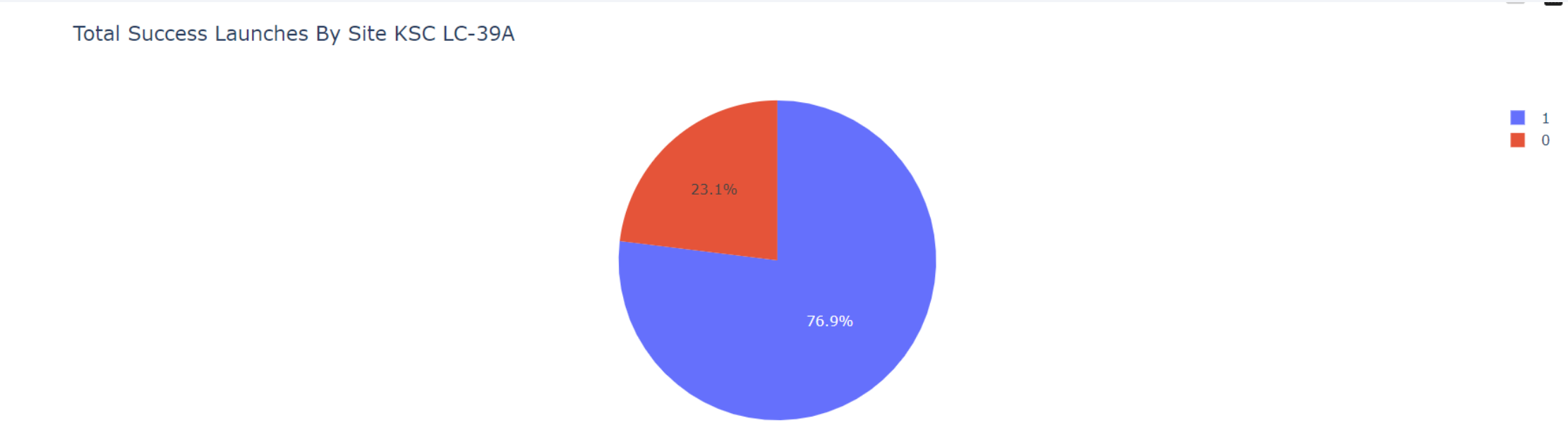

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches is 41.7%.

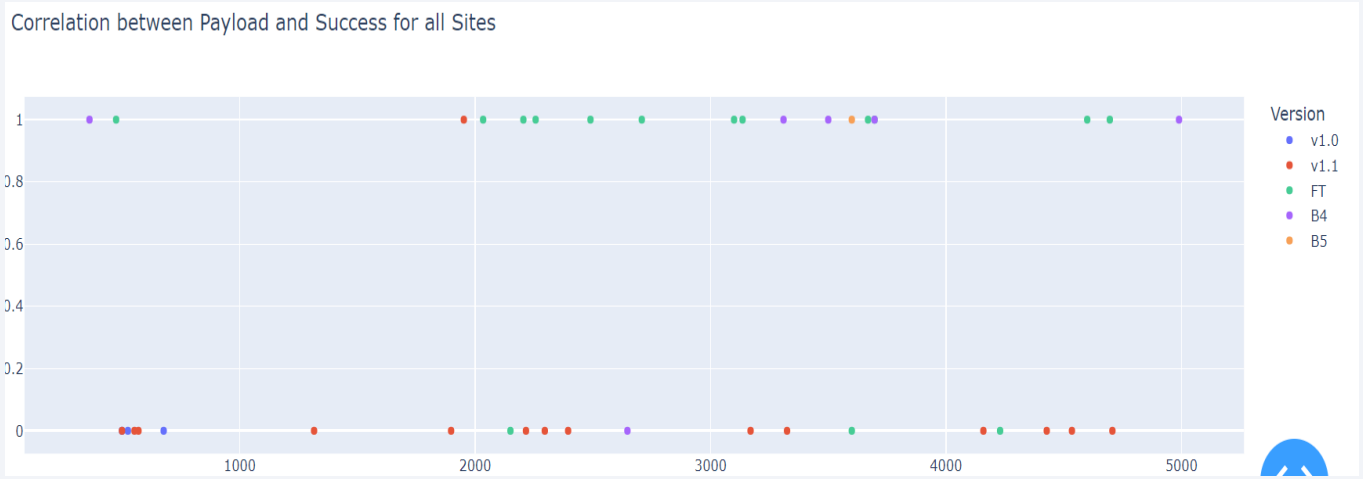
Launch site with the highest successful launch rate

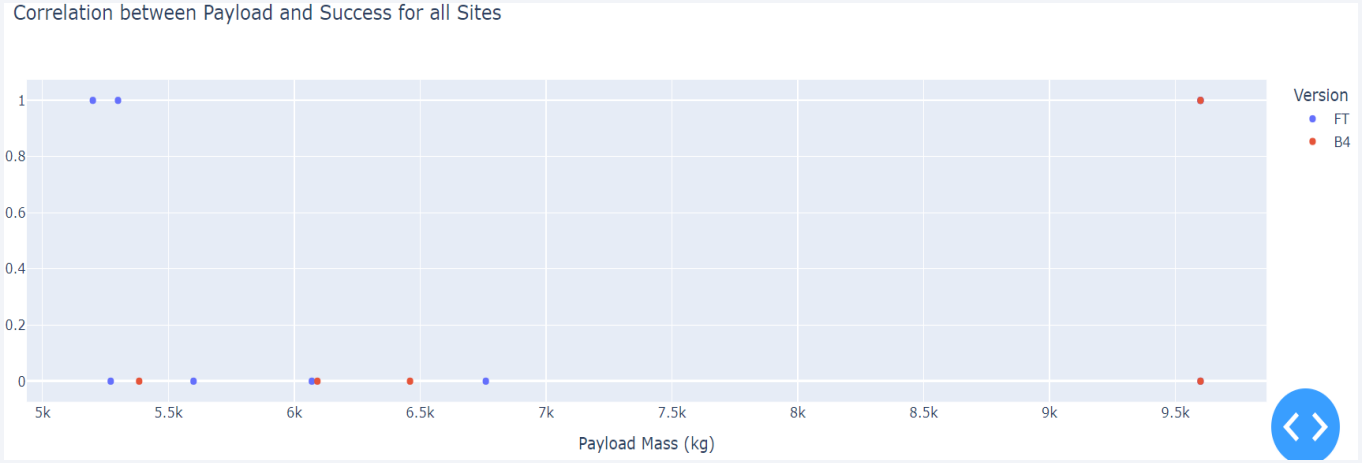


- KSC LC-39A with the highest success rate with 76.9% in 10 tests here

The relationship between Payload and Version affects the success rate

- Successful launch rate at payload from 0 to 5000kg is higher than from 5000kg to 10000kg.
- The highest attack rate is payload at 2000kg to 4000kg.
- Payload over 5000kg only 2 versions are FT and B4





Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Accuracy on testset shows no difference between models.
- Accuracy on the entire dataset shows that Decision Tree has the highest ratio and F1-Score is also the highest, showing that the model's performance is superior to that of other models.

Scores and Accuracy of the Test Set

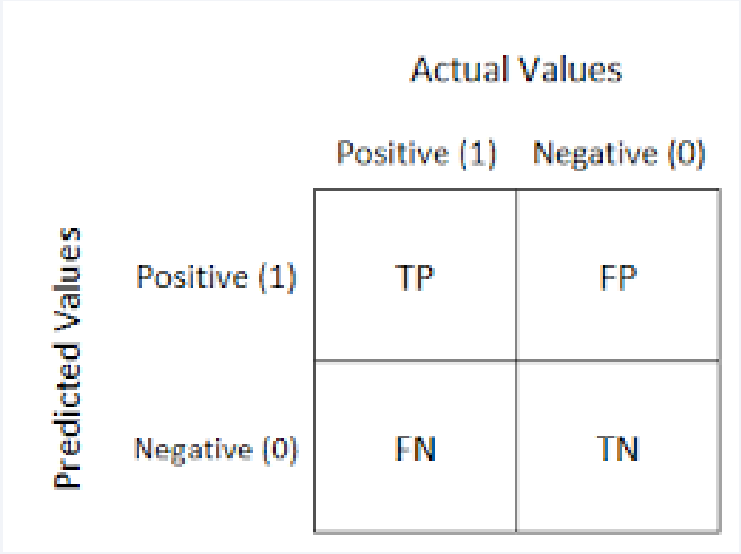
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

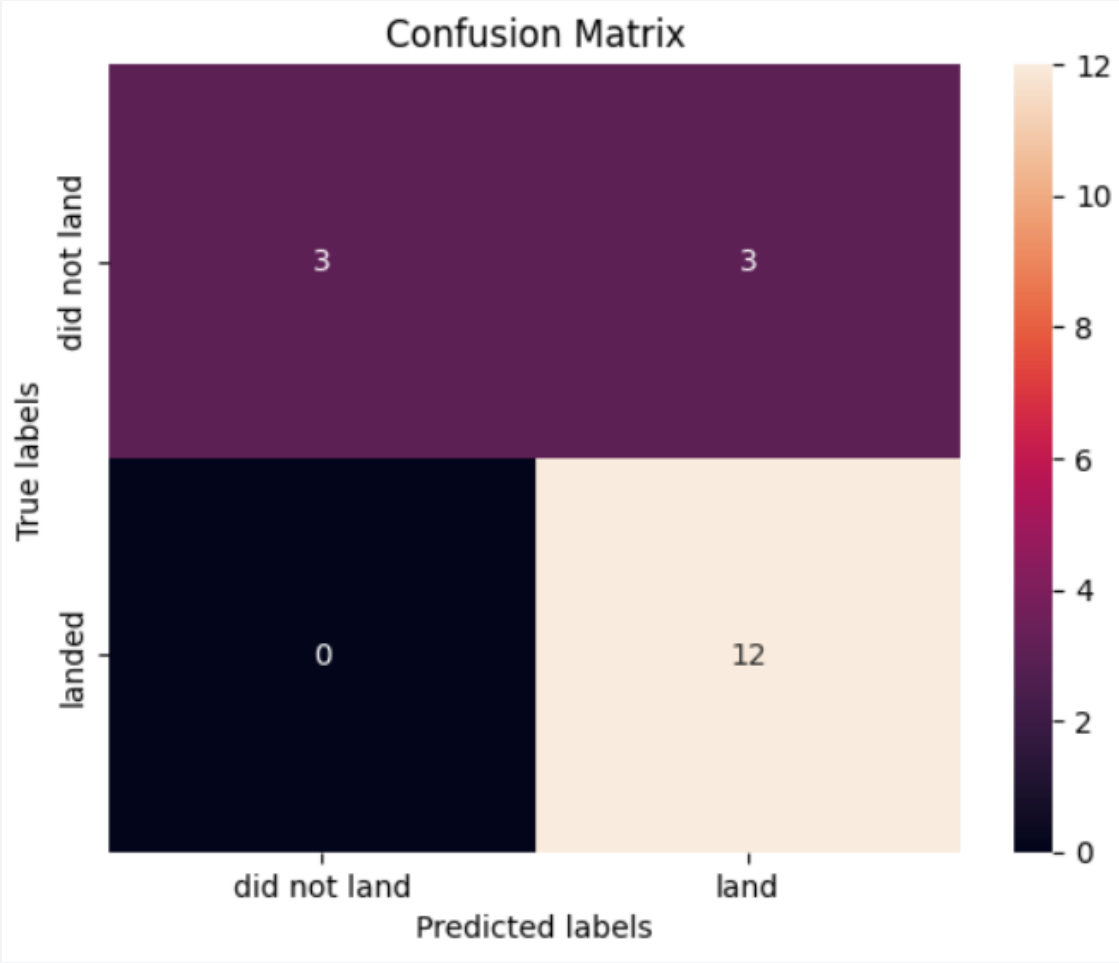
Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

- We can see here is the confusion matrix of Logistic-Regression with 15 correct classifiers and only 3 wrong classifiers.
- https://en.wikipedia.org/wiki/Confusion_matrix





Conclusions

- Decision Tree is the algorithm with the highest accuracy rate and F1-Score on the entire data file at 93.7%.
- KSC LC-39A Launch Site has the highest success rate of 76.9%.
- All Launch Sites are located close to the coastline and maximize distance from residential areas and highways to avoid the risk of rockets falling.
- The success rate has improved over the years, and peaked at 90% in 2019.
- Payload mass below 5000kg shows a superior success rate compared to Payload over 5000kg. and Payload between 2000kg and 4000kg shows the highest success rate.
- The success rate at Orbit ES-L1, SSO, HEO, GEO reaches 100%.

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!