Winning Space Race with Data Science

Sujal Donga
November, 2023

Outline

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

Executive Summary

- I collected data from the public SpaceX API and the SpaceX Wikipedia page. Then, I added a column called 'class' to classify successful landings. To analyze the data, I used SQL, visualizations, folium maps, and dashboards. I selected the relevant columns as features for further analysis.
- Next, I converted all the categorical variables into binary using a technique called one hot encoding. I standardized the data and used GridSearchCV to find the best parameters for the machine learning models. Finally, I visualized the accuracy scores of all the models.
- I created four machine learning models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. Surprisingly, all the models produced similar results, with an accuracy rate of around 83.33%. However, it's important to note that all the models tended to over-predict successful landings.
- To improve the model's accuracy and make more accurate predictions, we need to gather more data.

Introduction

- Background
  - Commercial travels to Space is being more affordable.
  - Space X has best pricing ($62 million vs. $165 million USD)
  - The reason space travel is so expensive is because the materials used are not reusable. Space Y aims to compete with Space X by finding ways to make space travel more affordable through reusing the stage 1.

Problem

- Space Y aims to develop the capability to predict whether or not the first stage of a rocket can be successfully recovered.

Section 1

Methodology

Methodology

Executive Summary

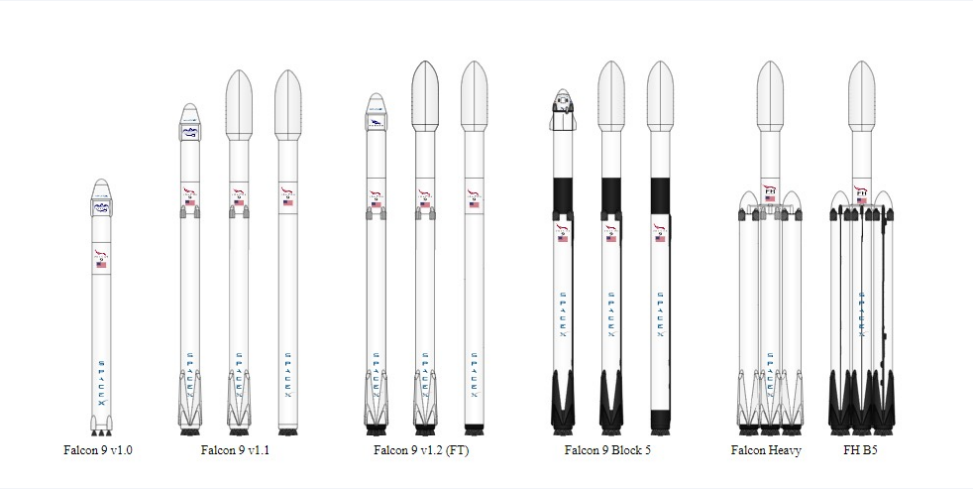
- Data collection methodology:
  - Data was collected from two sources : SpaceX public API and Wikipedia.
- Perform data wrangling
  - Data wrangling by landing success.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The object of model was created, then the hyperparameters were evaluated and selected based on r2 score and result was compared to other ML models.

Data Collection

- To gather the data, I used a combination of two methods. First, I made requests to Space X's public API to get some information. Second, I scraped data from a table on Space X's Wikipedia page.
- In the API data, we have different categories such as Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, GridFins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, and Latitude.
- The Wikipedia data that we gathered through web scraping includes Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Booster Version, Booster Landing, Date, and Time.
- In the next slide, you will see a flowchart that explains how we collected data from the API, and in the slide after that, you will see a flowchart showing how we collected data through web scraping.

Data Collection – SpaceX API

- Data collection -API
- [GitHub URL]



Extract data from Space X API > Json file

From the Json extracted was transformed into Dataframe

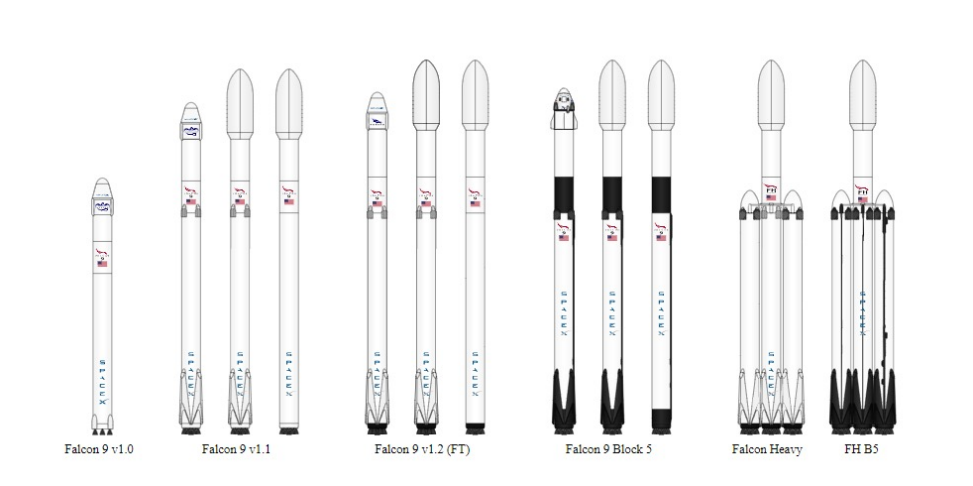
Relevant Dictionary was extracted from the previous dataframe

The important keys and information was selected and saved into another df

Data quality was checked by replacing the missing values with the mean.

Data Collection - Scraping

- GitHub URL



Request website Wikipedia by using Beautiful Soup.

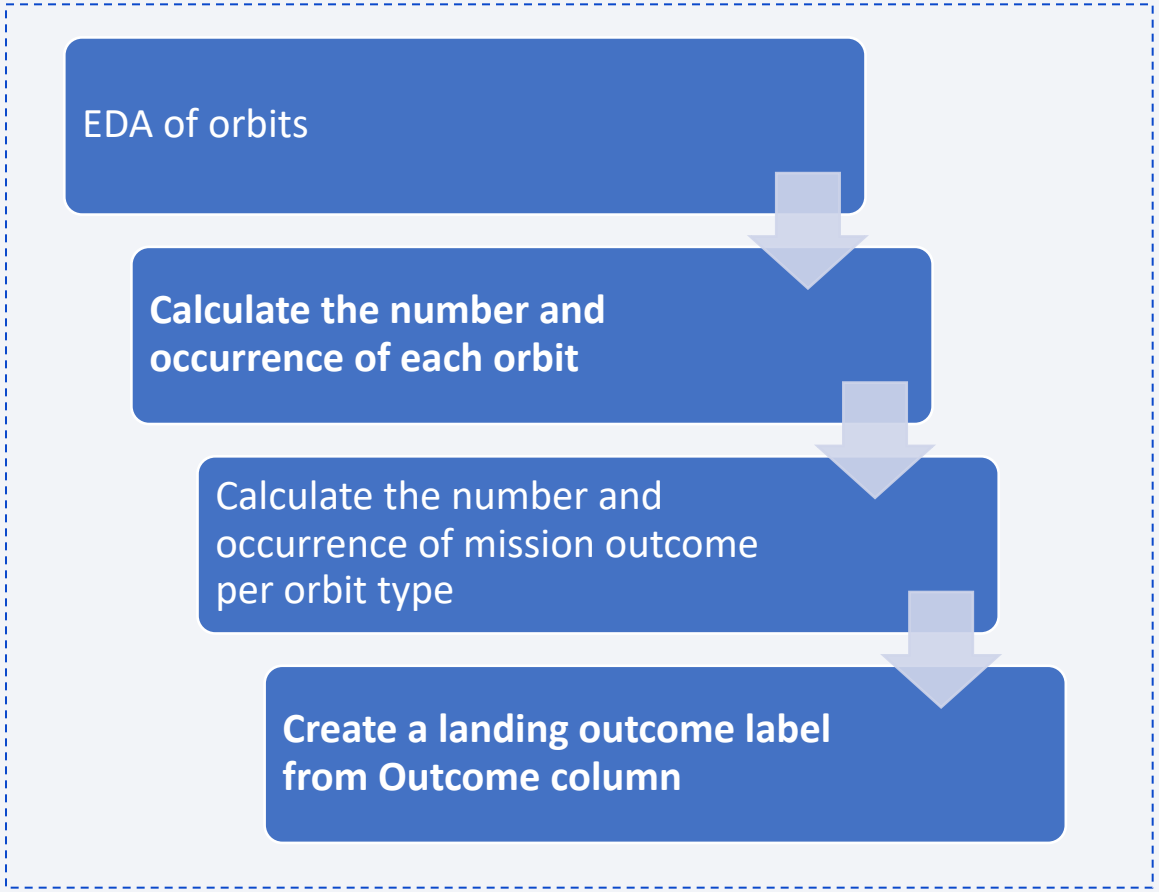
Find the proper html table to get the data needed.

Create dict with all values needed form the target table

Cast the dictionary to dataframe

Data Wrangling

- To create a training label indicating the landing outcomes, I assigned a value of 1 for successful landings and 0 for failures. The outcome column consists of two components: "Mission Outcome" and "Landing Location".
- I created a new training label column called "class". If the "Mission Outcome" is True, I set the value in the "class" column to 1. For the specific landing locations, I map the following values:
  - If the "Landing Location" is True for ASDS (Autonomous Spaceport Drone Ship), RTLS (Return to Launch Site), or Ocean, we will set the "class" value to 1.
  - If the "Landing Location" is None for both components or if "Mission Outcome" is False for ASDS, None for ASDS, False for Ocean, or False for RTLS, I set the "class" value to 0.



GitHub URL

EDA with Data Visualization

- `FlightNumber vs. PayloadMass`
  - **Catplot was used to see the changes of pay loads per flight number. Both variables are numeric. Catplot make easier to see all the values of x axis vs y axis if we compare with scatterplot.**
- `Visualize the relationship between Flight Number and Launch Site`
  - **Barplot was used to see to plot one numerical and categorical variable.**
- `Visualize Flight Number and the launch site.`
  - **Catplot was used to visualize the where the flight was launched over the time(flight number).**
- `Visualize the relationship between Payload , PayloadMass and Launch Site`
  - **Catplot was used to verify if the payload change with the launch site**
- `Success over the time`
  - **Lineplot is used to show trends.**

GitHub URL

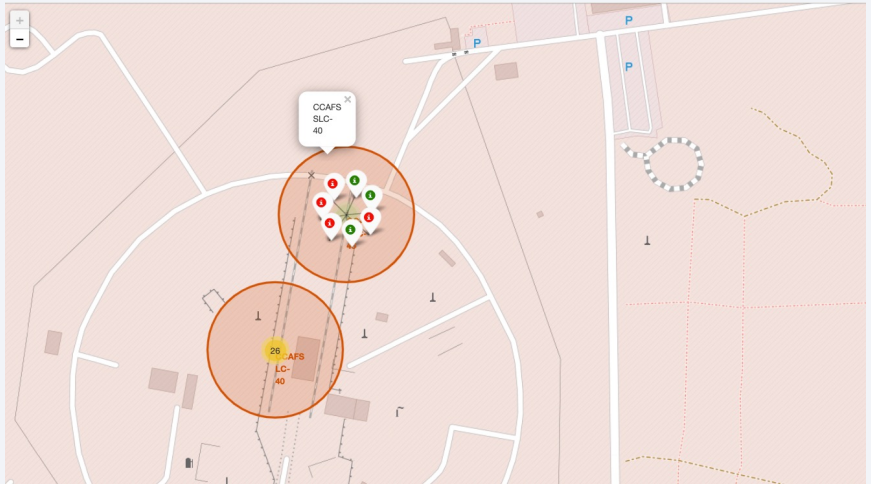
EDA with SQL

- 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.

GitHub URL

Build an Interactive Map with Folium

- The map shows unsuccessful and successful landings in different locations by showing circles and markers.
- Proximity is showed by a line.
- In the map includes markers, lines and circles.
- Folium maps depict Launch Sites, both successful and unsuccessful landings, and demonstrate their proximity to important locations such as Railway, Highway, Coast, and City. This enables us to and visually analyze the distribution of successfucomprehend the rationale behind the selection of launch sites l landings in relation to their geographical placement.



GitHub URL

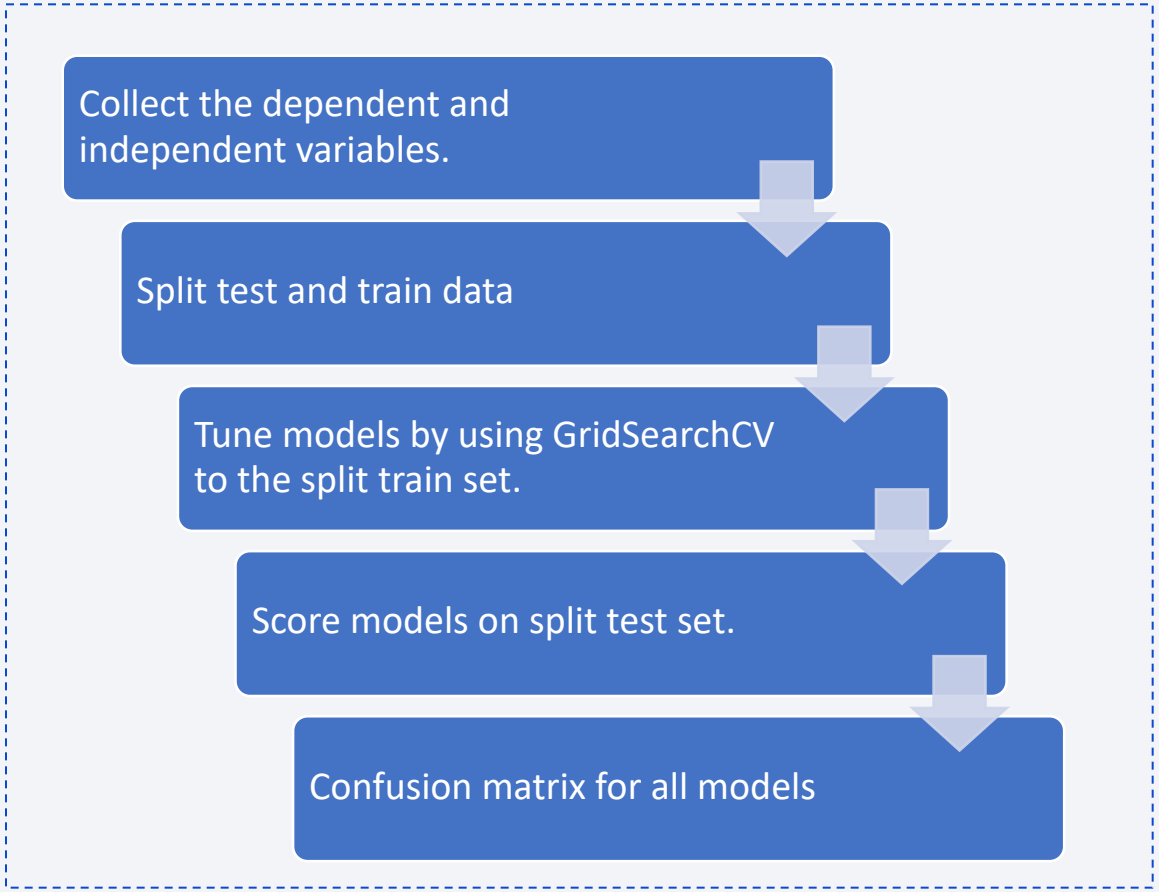
Build a Dashboard with Plotly Dash

- The dashboard consists of two visualizations: a pie chart and a scatter plot.
  - The pie chart provides options to display the distribution of successful landings across all launch sites or the success rates of individual launch sites. By selecting the pie chart, I can visualize how successful landings are distributed among different launch sites.
  - The scatter plot allows to explore the relationship between two variables. I can choose to view data for all launch sites or select an individual launch site. Additionally, I could adjust the payload mass using a slider ranging from 0 to 10000 kg. The scatter plot helps us analyze how the success of launches varies based on launch sites, payload mass, and the booster version category.

GitHub URL

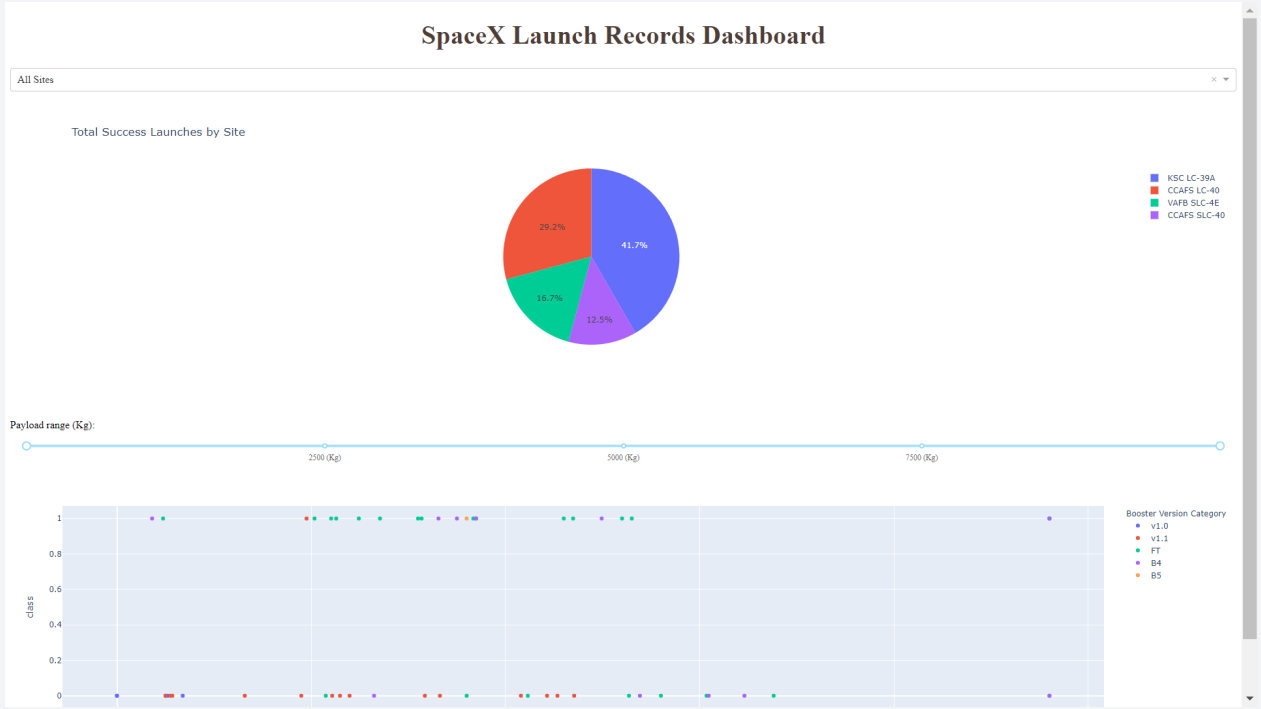
Predictive Analysis (Classification)

- Two dataset were extracted: One contains the class column considerer as dependent variable. Other dataset contains the independent variables.
- A split is being executed taking 20% of all rows as test data.
- Train set was used to gridsearch and tune the hyperparameters.
- The model is being evaluated by comparing the prediction with the real values of the test dataset.
- Finally, the scores are being compare within the machine learning algorithm to conclude the best one.
- As conclusion all model performed properly with similar score (0.83) except for decision tree classifier.

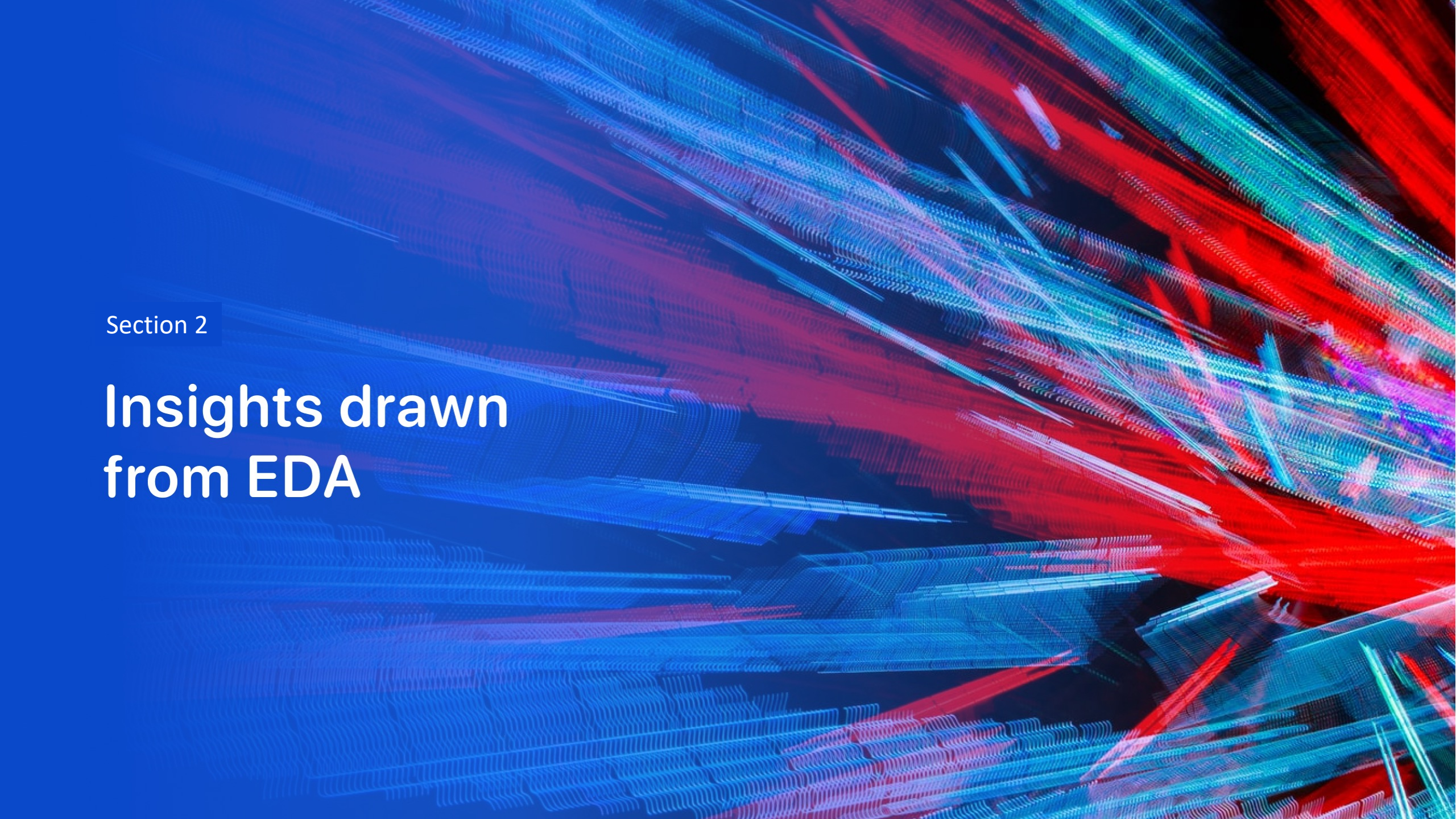


GitHub URL

Results

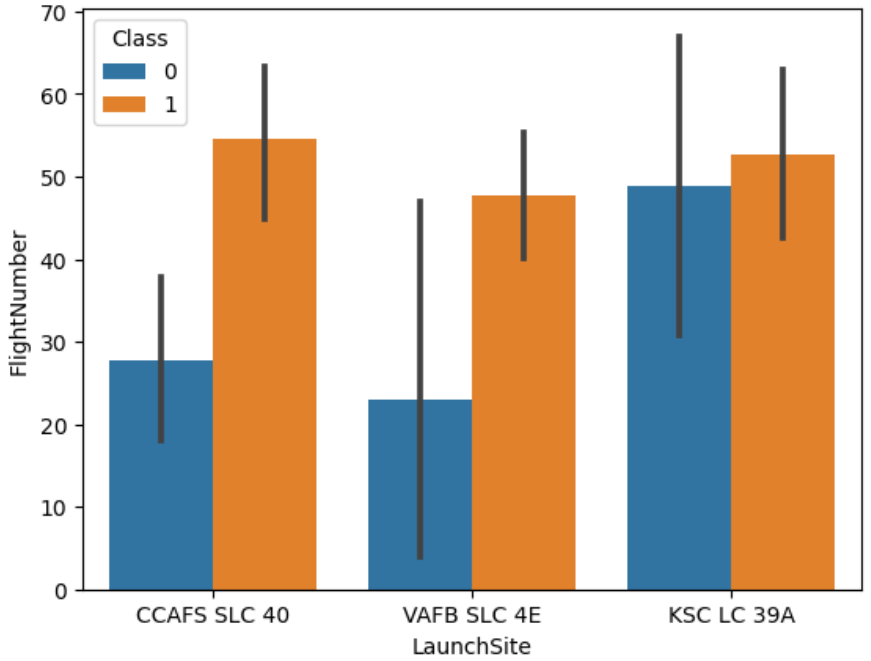


Here is a sneak peek of the dashboard. The upcoming slides will present the results of our Exploratory Data Analysis (EDA) through visualizations, EDA with SQL, an Interactive Map created using Folium, and lastly, the outcomes of our model with an impressive accuracy of approximately 83%.


Section 2
Insights drawn from EDA

EDA with Data Visualization

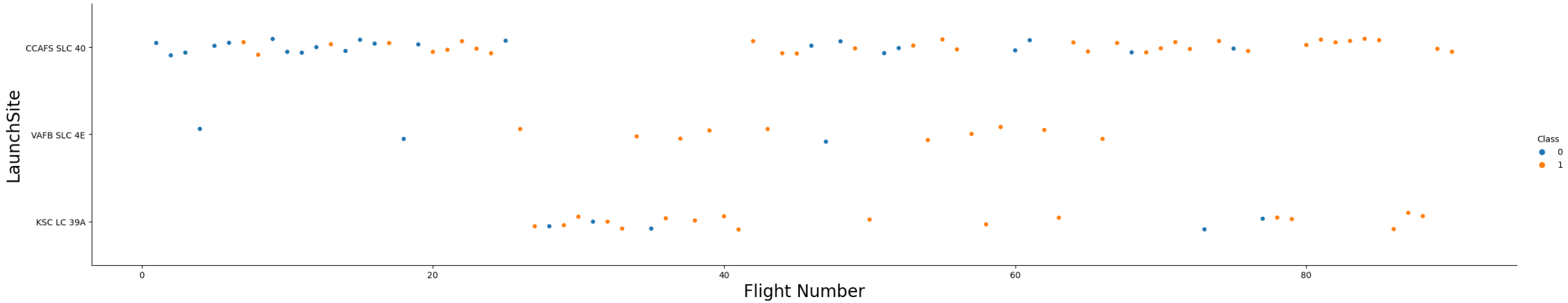
- **Visualize the relationship between Flight Number and Launch Site**
  - **Barplot was used to see to plot one numerical and categorical variable.**



GitHub URL

EDA with Data Visualization

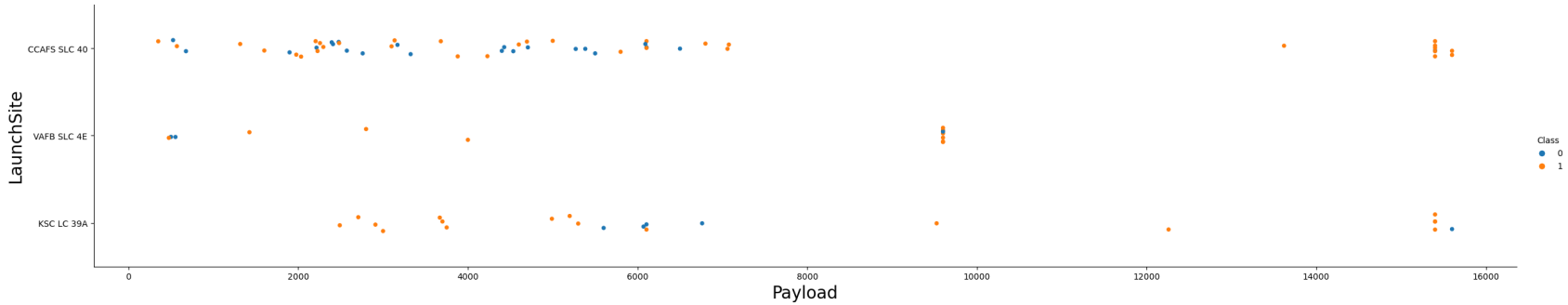
- **Visualize Flight Number and the launch site.**
  - **Catplot was used to visualize the where the flight was launched over the time(flight number).**

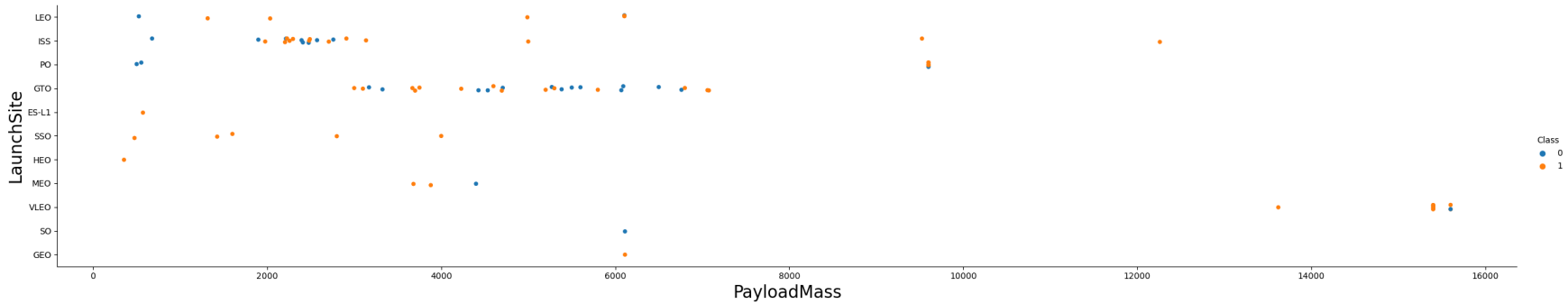


GitHub URL

EDA with Data Visualization

- **Visualize the relationship between Payload , PayloadMass and Launch Site**
  - **Catplot was used to verify if the payload change with the launch site.**

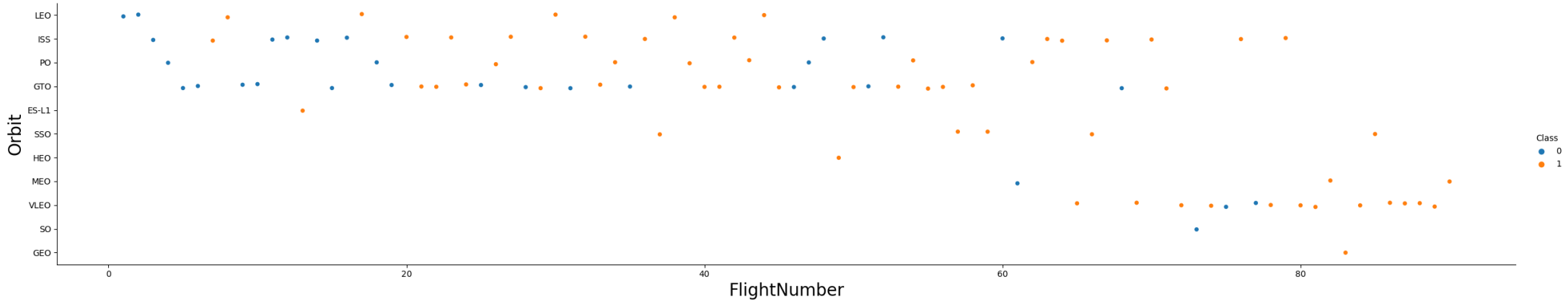




GitHub URL

EDA with Data Visualization

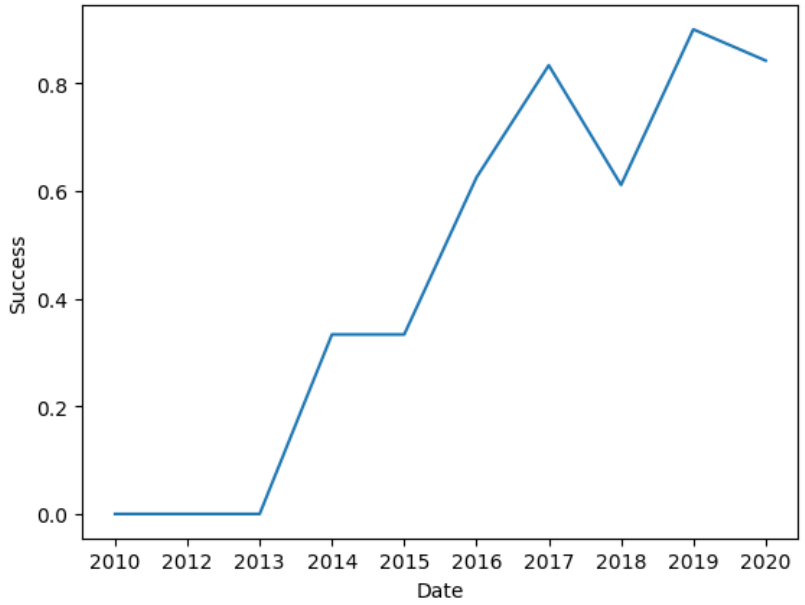
- **Orbit vs flight number**
  - **Catplot was used to verify the orbit changes over the time (flight number)**



GitHub URL

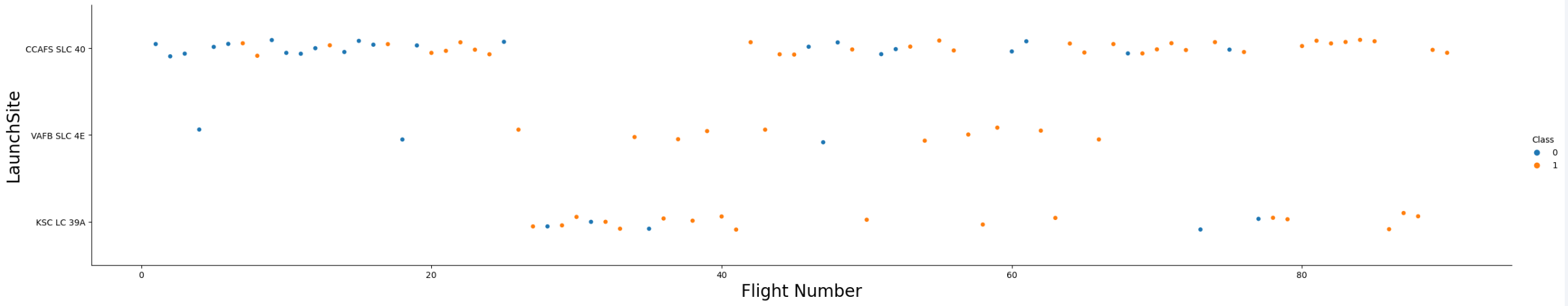
EDA with Data Visualization

- **Success over the time**
  - **Lineplot is used to show trends over the time from 2010 to 2020. After 2013, the trend shows as the time goes the rate of success increase more than 80% of change to land successfully .**



GitHub URL

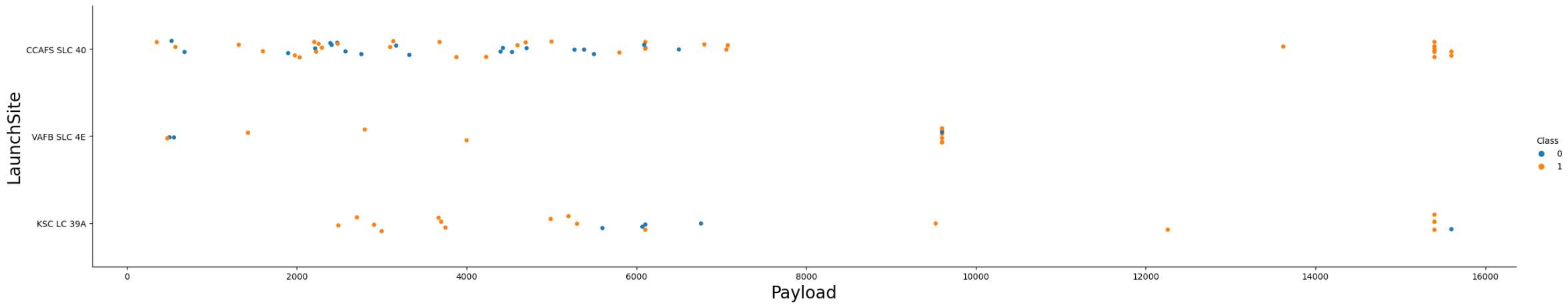
Flight Number vs. Launch Site



- The colors refers whether the landing was successfully or not. The y-axis shows the 3 launch sites and finally the x-axis describe the fight number which is a direct variable to measure the time too. As the flight number increase the tries to land a rocket raise as well.

GitHub URL

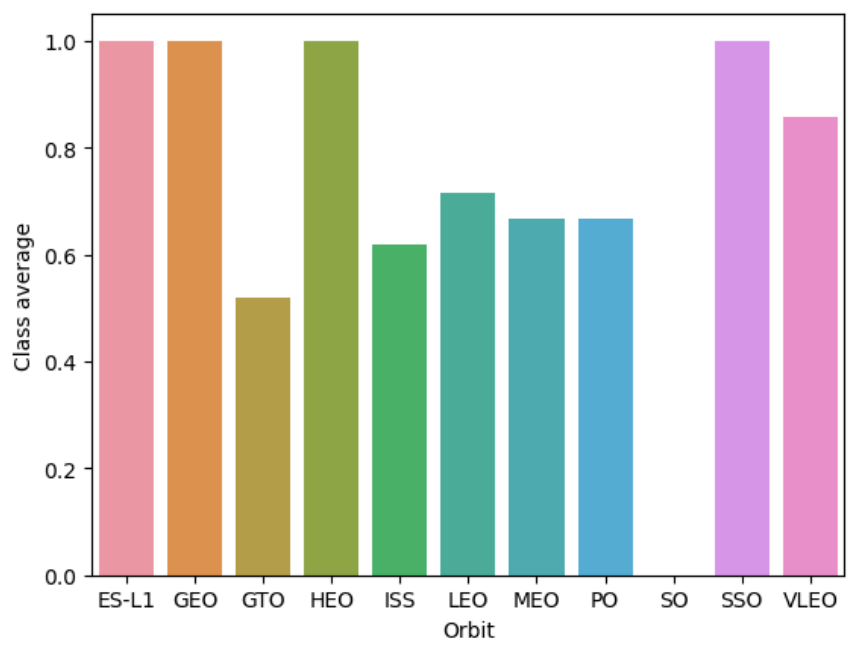
Payload vs. Launch Site



- Orange indicate successful launch, while Blue unsuccessful launch.
- The graphic indicates a noticeable improvement in the success rate over time, as indicated by the Flight Number. There seems to be a significant breakthrough around flight number 20, which resulted in a substantial increase in the success rate.
- Additionally, the data suggests that the Cape Canaveral Space Force Station (CCAFS) is the primary launch site, as it shows the highest volume of launches compared to other sites.

GitHub URL

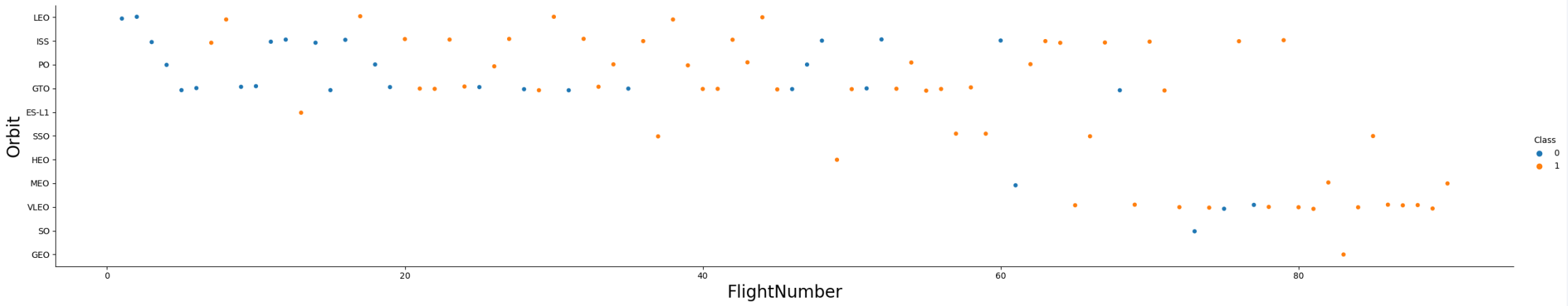
Success Rate vs. Orbit Type



- 100% success rate can be achieved reaching the orbits ES-L1,GEO,HEO and SSO.
- 80% of success rate can be done reaching to the orbit VLEO.
- 70% of success to land the stage 1 when reaching to the orbits:
  - GTO,ISS,LEO,MEO and PO
- 0% of success to reach the orbit SO.

GitHub URL

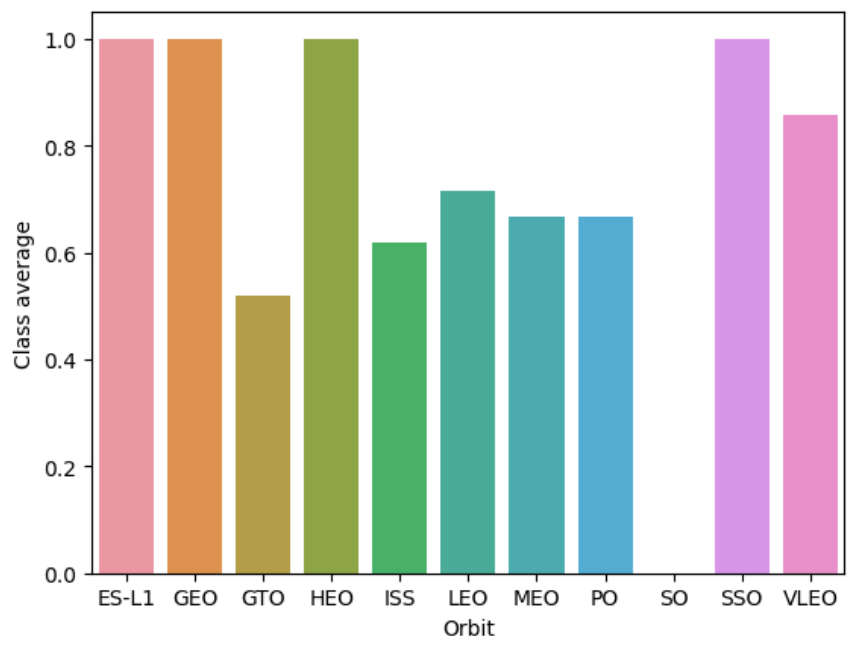
Flight Number vs. Orbit Type



- The colors refers whether the landing was successfully or not. The y-axis shows the Orbits in which the rocket was aim to reach and finally the x-axis describe the fight number which is a direct variable to measure the time too. As the flight number increase the tries to land a rocket raise as well. Mainly, The latest launches were aimed to reach the orbit VLEO and some of them were ISS, in which most of the 80% of landings were successfull.

GitHub URL

Payload vs. Orbit Type

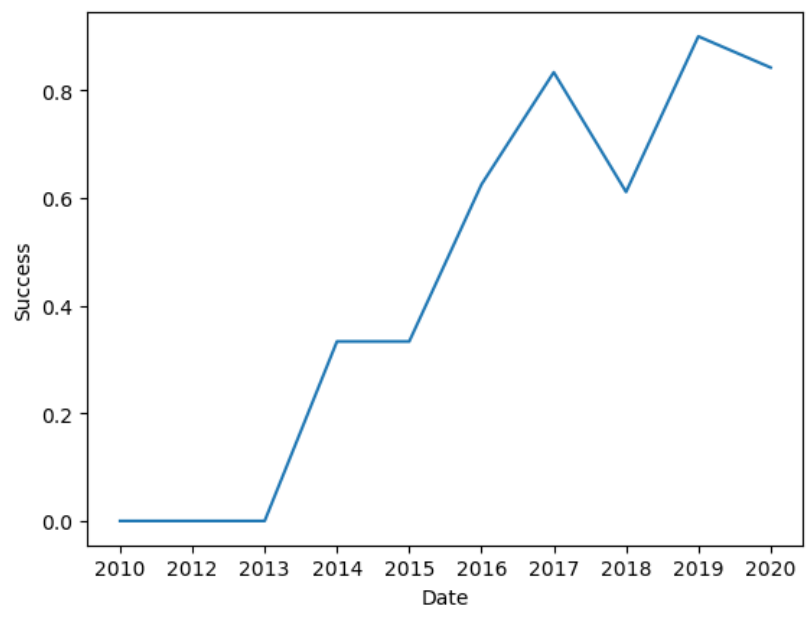


- This graph shows the average of success of landing of the booster. We can conclude it is more feasible to recover the booster when the launch is aimed to reach the orbits ES-L1, GEO, GTO, HEO and SSO.

GitHub URL

Launch Success Yearly Trend

- Lineplot is used to show trends over the time from 2010 to 2020. After 2013, the trend shows as the time goes the rate of success increase more than 80% of change to land successfully .



GitHub URL

All Launch Site Names

- After the execution of the launch sites, we find 4 all them:
  - CCAFS LC-40: This refers to Cape Canaveral Air Force Station Launch Complex 40, located in Florida, USA. It is one of the primary launch sites used by SpaceX for launching Falcon 9 and Falcon Heavy rockets.
  - VAFB SLC-4E: This stands for Vandenberg Air Force Base Space Launch Complex 4E, situated in California, USA. It is another launch site used by SpaceX for launching primarily polar orbit missions.
  - KSC LC-39A: This denotes Kennedy Space Center Launch Complex 39A, located in Florida, USA. It is a historic launch site that has been used for numerous crewed missions, including SpaceX's Falcon 9 launches and the launch of the Crew Dragon spacecraft.
  - CCAFS SLC-40: This is another reference to Cape Canaveral Air Force Station Launch Complex 40, which is the same as the first entry mentioned. It indicates the same launch site in Florida used by SpaceX for Falcon 9 and Falcon Heavy launches.

```
[14]  1  %sql select distinct Launch_Site from SPACEXTABLE
```

```
 * 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'

```
[25]  1  %sql select  * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

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

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

- These are the first 5 entries of launch Sites that begins with CCA.

Total Payload Mass

- This query sums the total payload mass in kg where NASA (CRS) is the customer.

```
[26]  1  %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer=='NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

```
32]  1  %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version=='F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- The query compute the average of payload mass in kg where the rocket use booster with version F9 v1.1.

First Successful Ground Landing Date

- The way to determine the first successful ground landing date is computing the minimum date where the mission outcome was success, which was 2010-04-06.

```
37]  1  %sql SELECT MIN(DATE) FROM SPACEXTABLE WHERE Mission_Outcome=='Success'
```

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

MIN(DATE)
2010-04-06

Successful Drone Ship Landing with Payload between 4000 and 6000

- This shows 4 boosters versions which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

-

```
]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome=='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND
 * sqlite:///my_data1.db
Done.
]:
```

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

Total Number of Successful and Failure Mission Outcomes

- 99% of the mission outcomes were successfully while the one launch has unclear payload and other failed in fight.

```
1  %sql SELECT Mission_Outcome , COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- The list of names of the booster version are listed in the following query. The highest payload was 15600 Kg. The booster's version F9 B5 B10xx.x variation were used to carry the heaviest payloads.

```
[41]  1  %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_==(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * 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

2015 Launch Records

- In April and October, landing_outcomes failed in drone ship, their booster versions were similar, and launch site names was CCAFS LC-40 for in year 2015

```
20]: l SELECT substr(Date, 6, 2) AS MONTH,Landing_Outcome ,Booster_Version,Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' a
```

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

20]:

MONTH	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- There were 20 landings between the date 2010-06-04 and 2017-03-20, there were 10 of them were successfull (5 drone ship and 5 ground pad) and 9 Failures (5 drone ships, 2 parachute and 1 complete failure).

```
1  %sql SELECT distinct Landing_Outcome, count(Landing_Outcome)as count FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' and Landing_Outcome  like 'Succ%'   or Landing_Outcome  like '
```

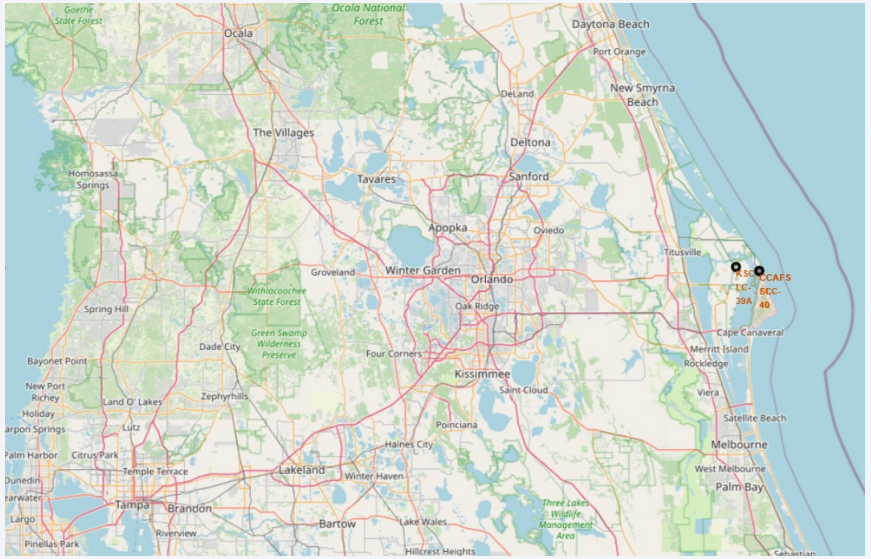
Python

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

Landing_Outcome	count
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2


Section 3
Launch Sites
Proximities Analysis

Total launch sites locations



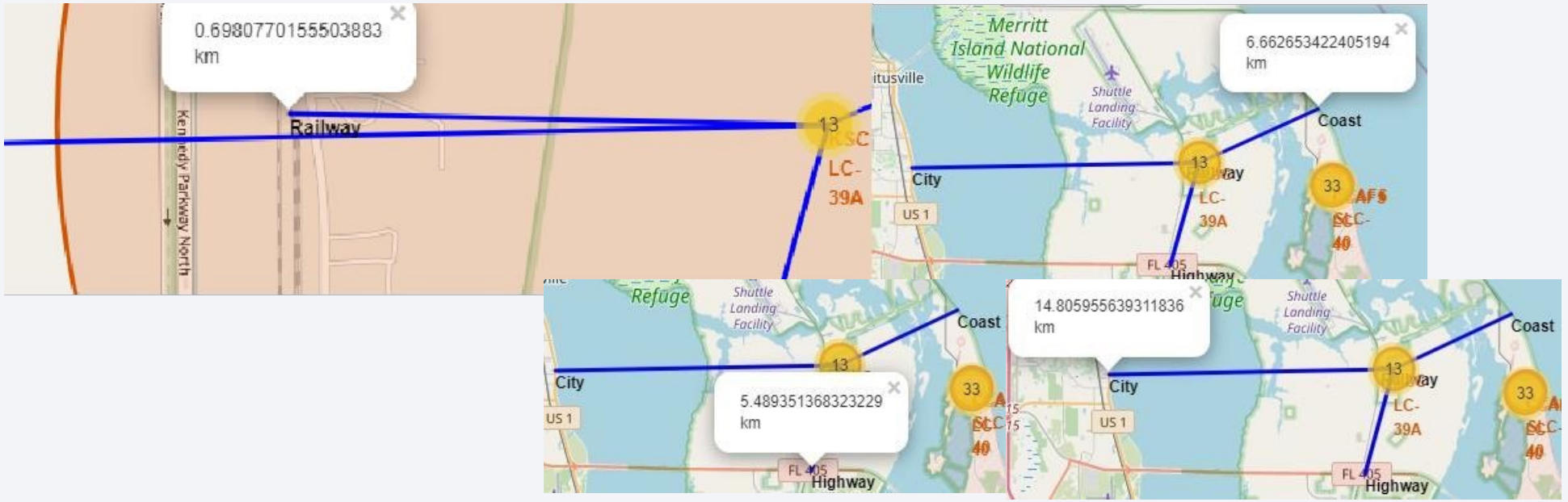
- Right map: all location in US
- Left map: Launch sites location in Florida.

- Florida has 3 launch sites Locations. KSC LC39A, CCAF SLC40 and CCAFS LC 40

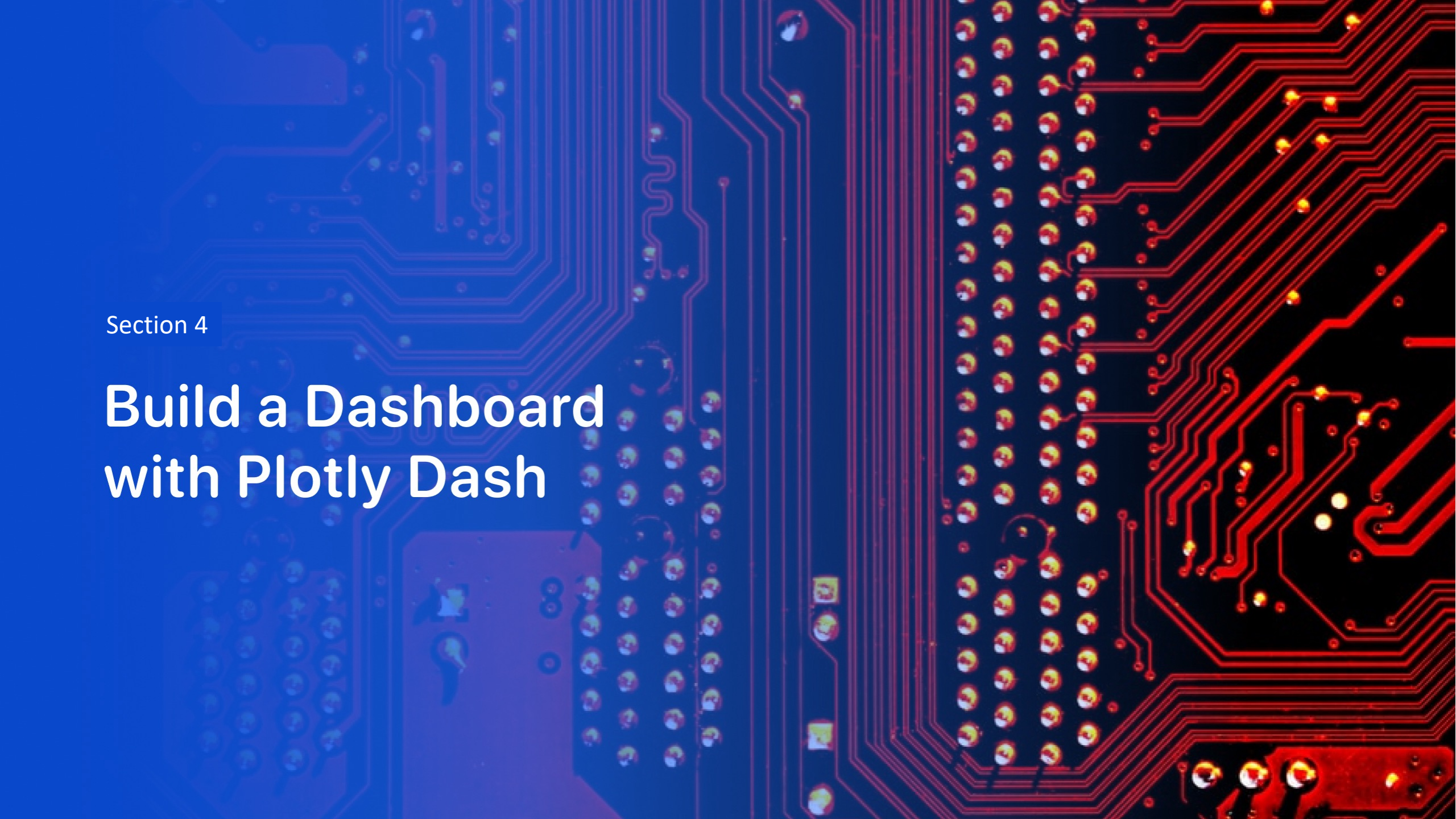
Outcomes of launch (color coded)

- 26 launches were made in the location CCAFS LC 40; 7 were successfully and 19 Failed. The last 10 launches 80% were success.

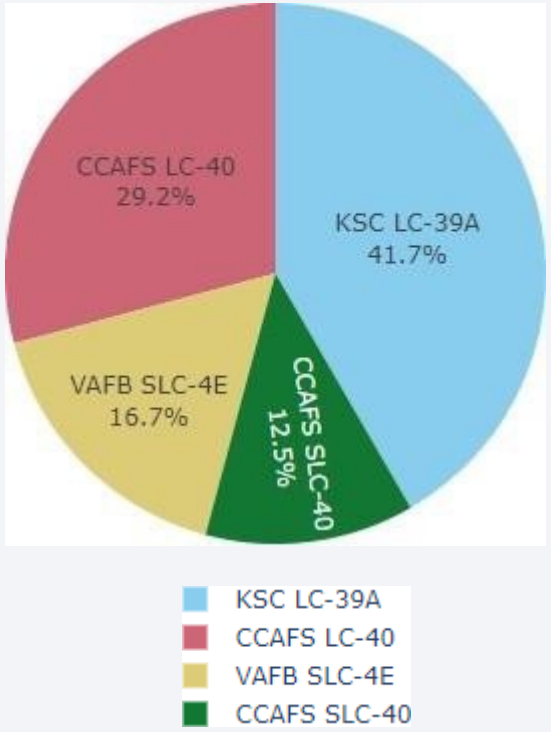
Folium Map Screenshot 3



- Taking KSC LC-39A as an illustration, launch sites are strategically situated in close proximity to railways to facilitate efficient transportation of large components and supplies. They are also located near highways to ensure convenient transportation for personnel and logistical needs. Additionally, launch sites are positioned near coastlines and away from densely populated cities. This positioning serves the purpose of directing launch failures towards the sea, minimizing the risk of rockets falling on densely populated areas and prioritizing safety.


Section 4
Build a Dashboard with Plotly Dash

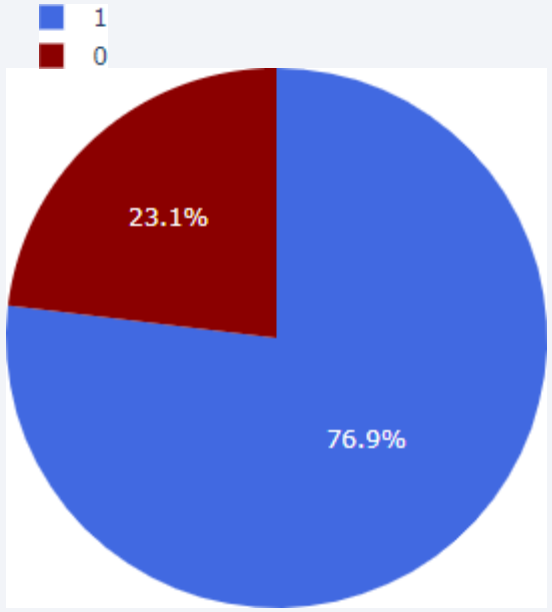
Successful launches across launch sites



The distribution of successful landings across all launch sites reveals some interesting patterns. Firstly, it is important to note that CCAFS LC-40 and CCAFS SLC-40 are essentially the same launch site, with CCAFS LC-40 being the old name. As a result, CCAFS and KSC have an equal number of successful landings. However, it is worth mentioning that a majority of these successful landings occurred prior to the name change.

On the other hand, VAFB has the smallest share of successful landings. This could be attributed to a smaller sample size and potentially increased difficulty in launching from the west coast. Launching from Vandenberg Air Force Base (VAFB) on the west coast may present different challenges and operational considerations compared to the Florida-based launch sites.

<Dashboard Screenshot 2>



- Among the launch sites, KSC LC-39A boasts the highest success rate, having achieved 10 successful landings and experiencing only 3 instances of failed landings.

<Dashboard Screenshot 3>

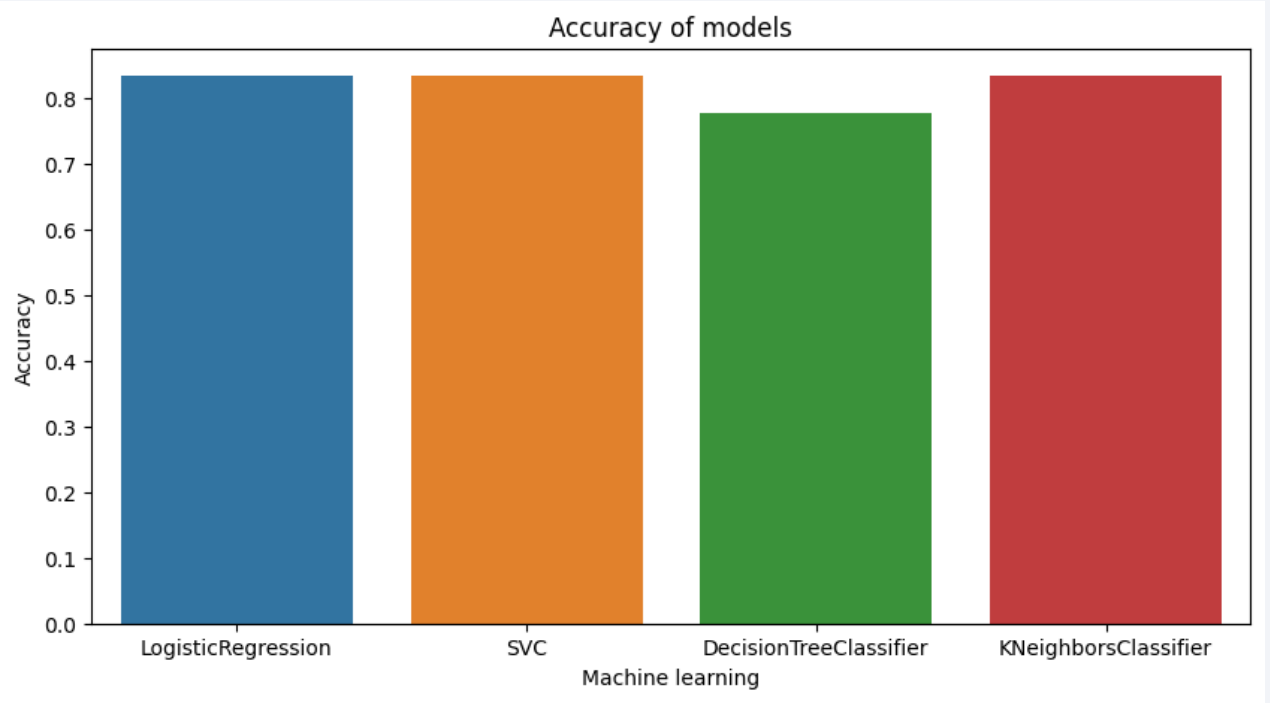
- The Plotly dashboard includes a Payload range selector, but it is currently set from 0 to 10,000 instead of the maximum payload value of 15,600. The class variable indicates a value of 1 for successful landings and 0 for failures. The scatter plot incorporates the booster version category through color coding and represents the number of launches with varying point sizes.
- Interestingly, within the specific payload range of 0 to 6,000, there are two failed landings that stand out. These failures have payloads recorded as zero kilograms, which adds a unique aspect to the data analysis.

Section 5

Predictive Analysis (Classification)

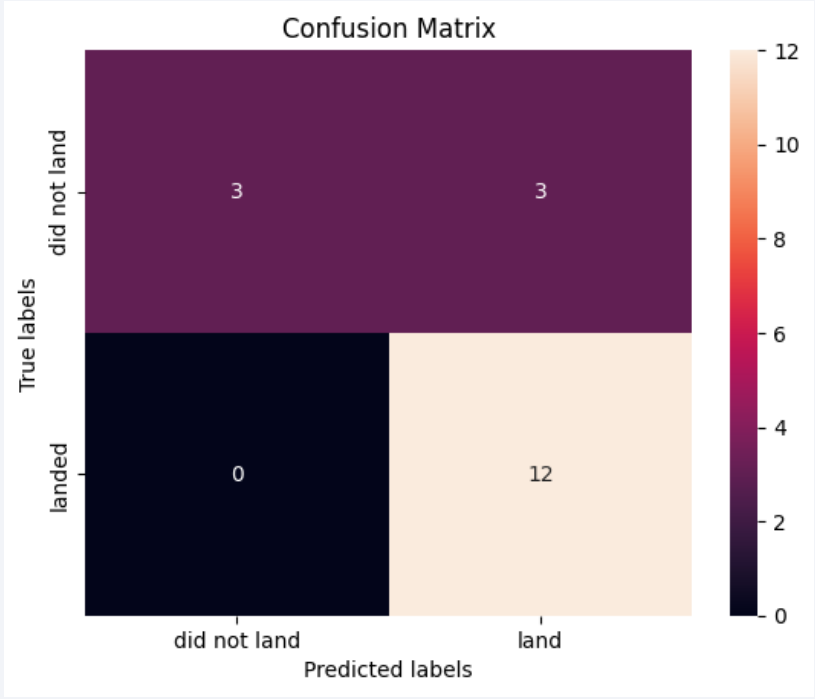
Classification Accuracy

- The accuracy of all models on the test set was nearly identical, with a shared accuracy rate of 83.33%. However, it is important to consider that the test size is relatively small, consisting of only 18 samples. This limited sample size can lead to significant variance in accuracy results, as seen in the case of the Decision Tree Classifier model during repeated runs.



Confusion Matrix

- Across all models, the performance on the test set was consistent, resulting in identical confusion matrices. The models correctly predicted 12 instances of successful landings when the true label indicated a successful landing. Additionally, they correctly identified 3 instances of unsuccessful landings.
- However, it is worth noting that the models also made some misclassifications. Specifically, they incorrectly predicted 3 instances of successful landings when the true label indicated unsuccessful landings, leading to false positives. This indicates that the models tend to overpredict successful landings.
- This observation highlights a tendency of the models to overestimate the occurrence of successful landings, which could be an area for further investigation and adjustment to improve the accuracy of predictions.



Conclusions

- Developed a machine learning model for Space Y to compete with SpaceX
- Objective: Predict successful Stage 1 landings to save ~$100 million USD
- Utilized data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data in a DB2 SQL database
- Developed a visualization dashboard for data analysis
- Achieved 83% accuracy with the machine learning model
- Allon Mask and SpaceY can use the model to predict successful Stage 1 landings before launch
- Model aids in determining whether a launch should proceed based on predicted landing success
- Collecting more data can further improve the accuracy and identify the best machine learning model for the task.

Appendix

- GitHub repository:

https://github.com/xrayid/IBM-Applied-Data-Science-Capstone

Thank you!