RIPHAH INTERNATIONAL UNIVERSITY

Winning Space Race with Data Science

HAZRAT ABBAS KHAN

abbaskhan0345060@gamail.com

abbaskhan0345 (GitHub)

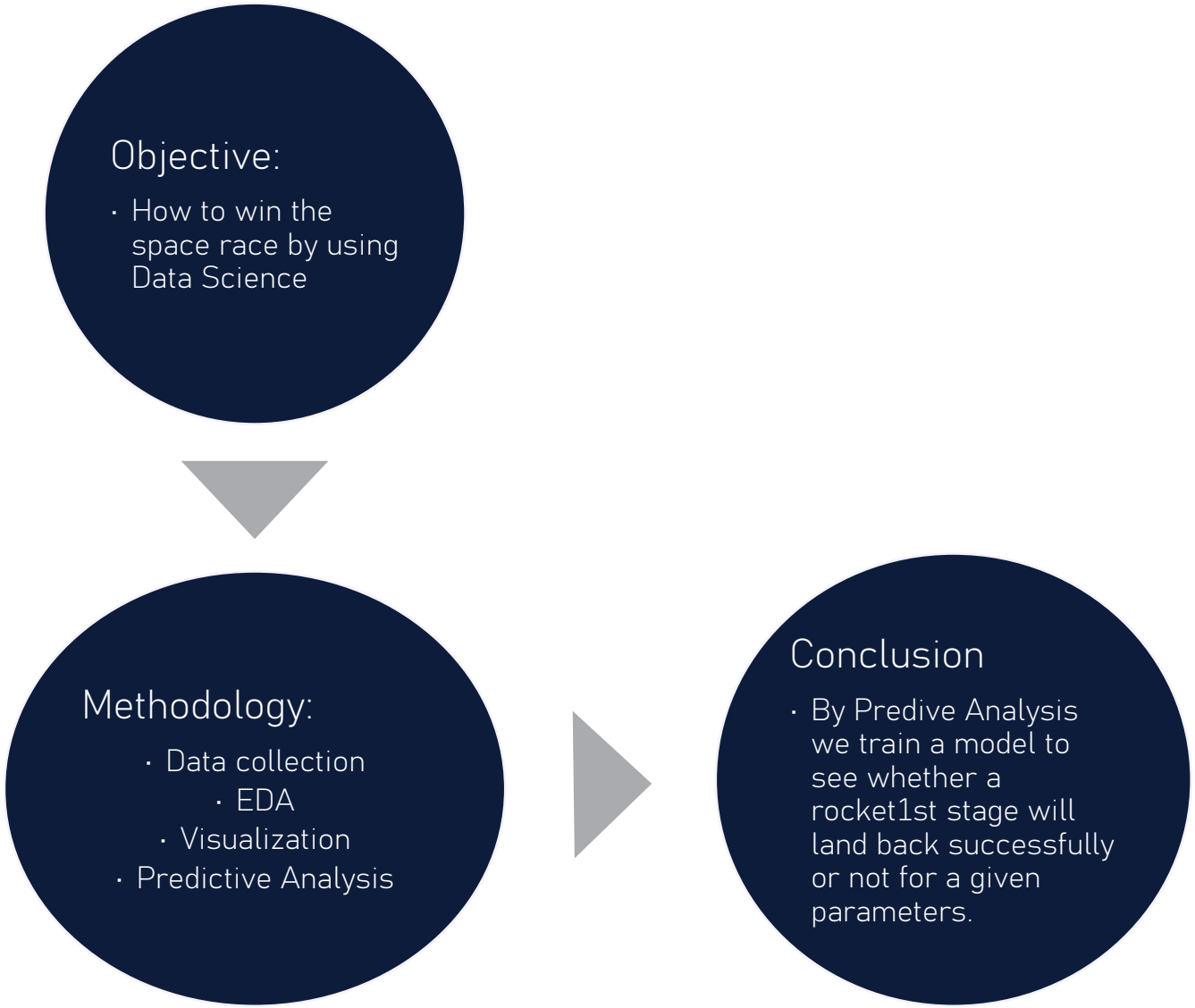
Riphah Institute of Informatics

IBM Developer SKILLS NETWORK

Table of Content

Executive Summary

Executive Summary



INTRODUCTION AND BACKGROUND

IBM Developer
SKILLS NETWORK

INTRODUCTION AND BACKGROUND

- The traditional space race is about to launch a rocket and send it to space with no hope of reusability.
- Us Space-Y and our competitor Space-X took initiative to change that approach with the reusability of rocket's first stage.
- Rocket is launch deliver the payload and the fist stage come back successfully. Which save cost and recycling of that Rocket.
- The Rocket type that is used in this specific task is Falcon9.

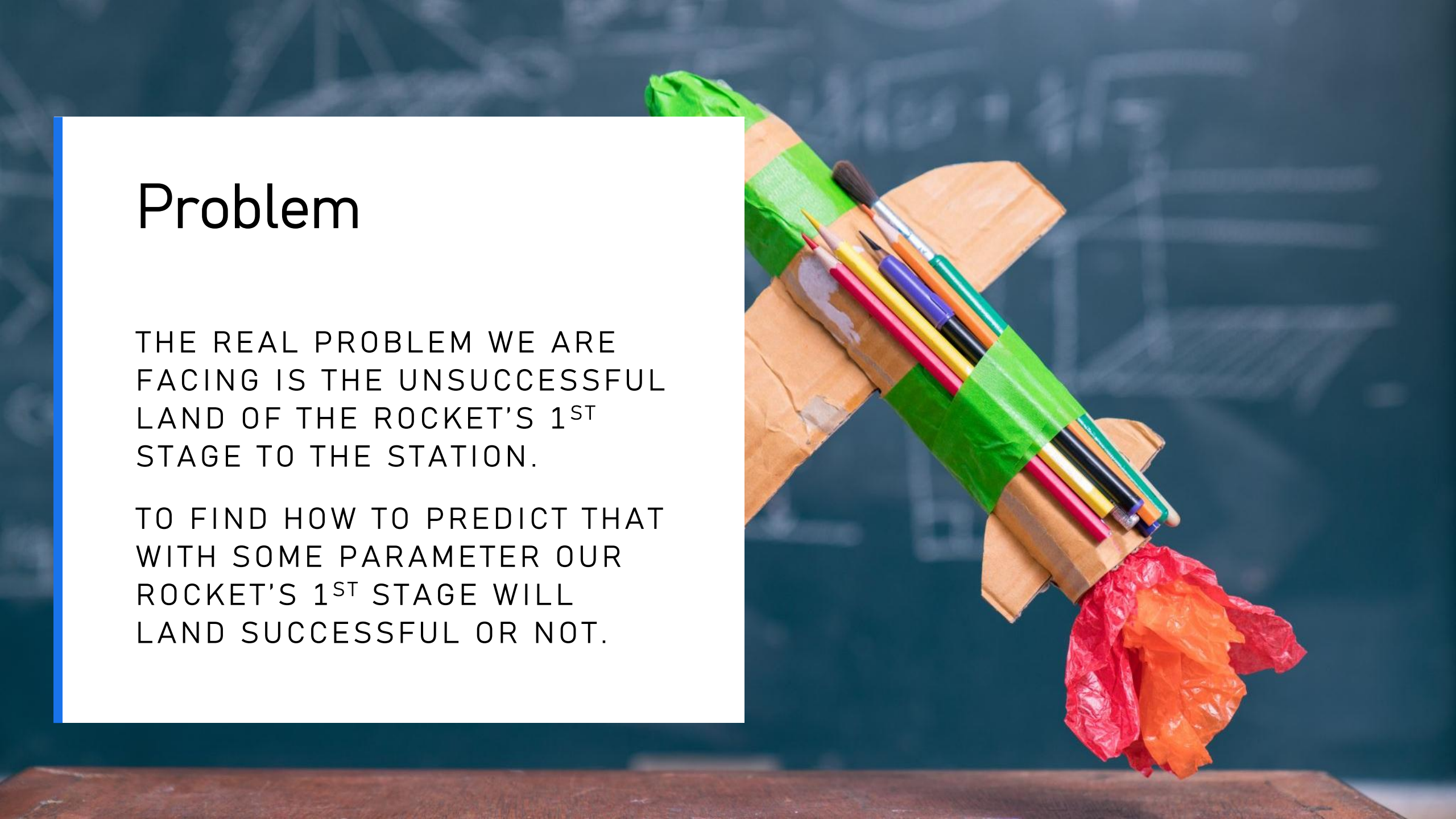
Section 1

Problem Statement

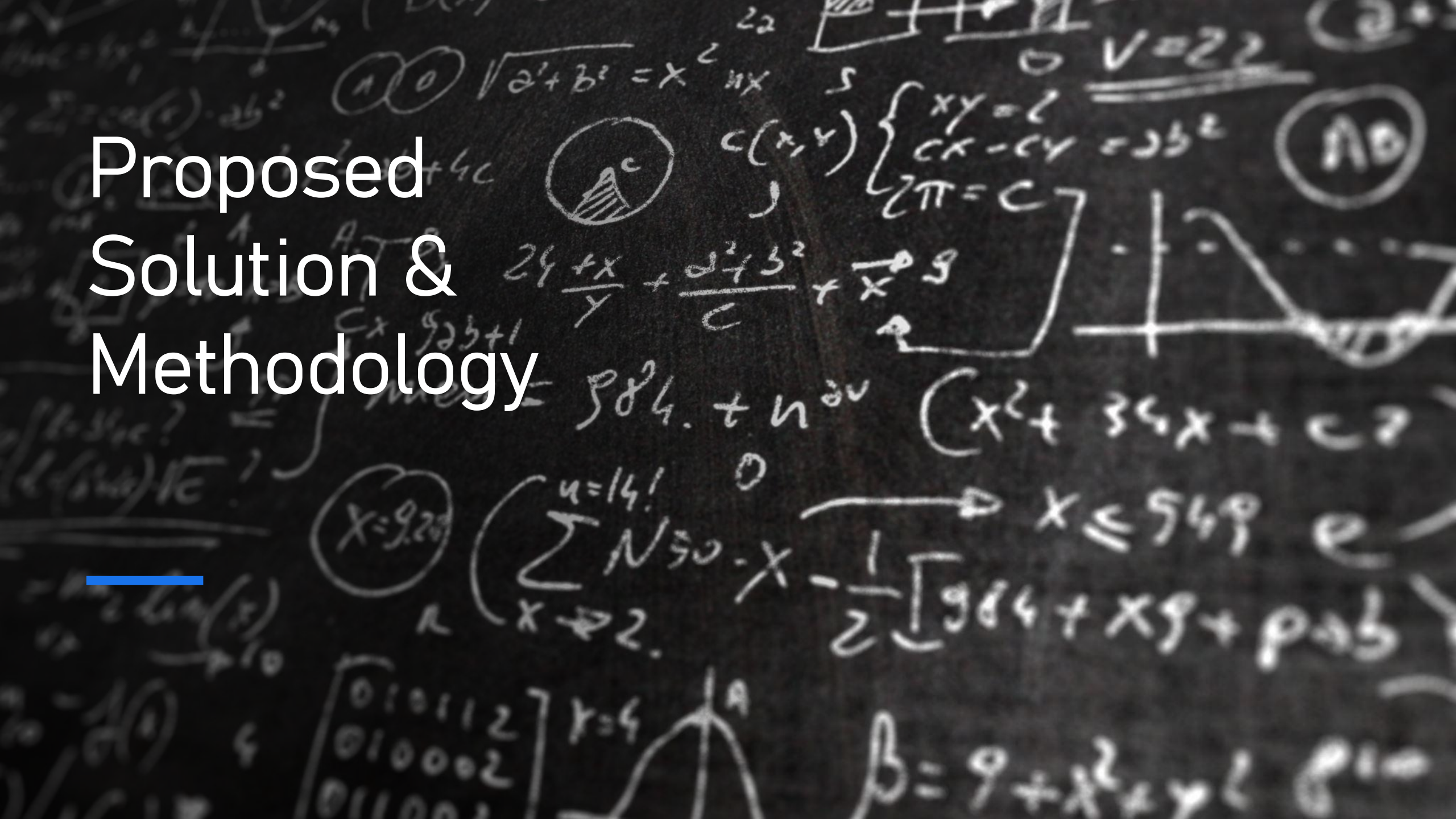
Problem

THE REAL PROBLEM WE ARE FACING IS THE UNSUCCESSFUL LAND OF THE ROCKET'S 1ST STAGE TO THE STATION.

TO FIND HOW TO PREDICT THAT WITH SOME PARAMETER OUR ROCKET'S 1ST STAGE WILL LAND SUCCESSFUL OR NOT.



Proposed Solution & Methodology



Solution

- A predative machine learning model which is train on past data from Space-X.
- The model will inform us with the probability of success or not.

Methodology (Executive Summary)

Data collection	Data Wrangling	EDA	Interactive dashboards	Predictive Modeling
From Space-X open API.	Data fetched from Wikipedia page and prosecced.	Performed the EDA on collected data by the help of SQL queries and Visualization.	Perform interactive visual analytics using Folium on the launch sites and Plotly dash for interactive data visualization based on time.	Train and evaluate the different classification models after the preprocessing of data.

Data Collection

- Web scraping data from Wiki Pedia
- Data collection is the process of gathering data from available sources. This data can be
- structured, unstructured, or semi-structured. For this project, data was collected via
- SpaceX API and Web scrapping Wiki pages for relevant launch data.

Data Collection SpaceX API

The data collection process began with api request call from https://api.spacexdata.com/v4

Where we collected relevant information such as:

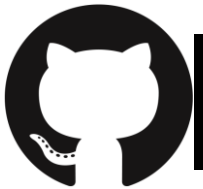
All the data received is then being saved to a csv file.

From the **rocket** column we obtained the **booster name**.

From the **launchpad** we were like to know the name of the **launch site** being used, the **logitude, and the latitude**.

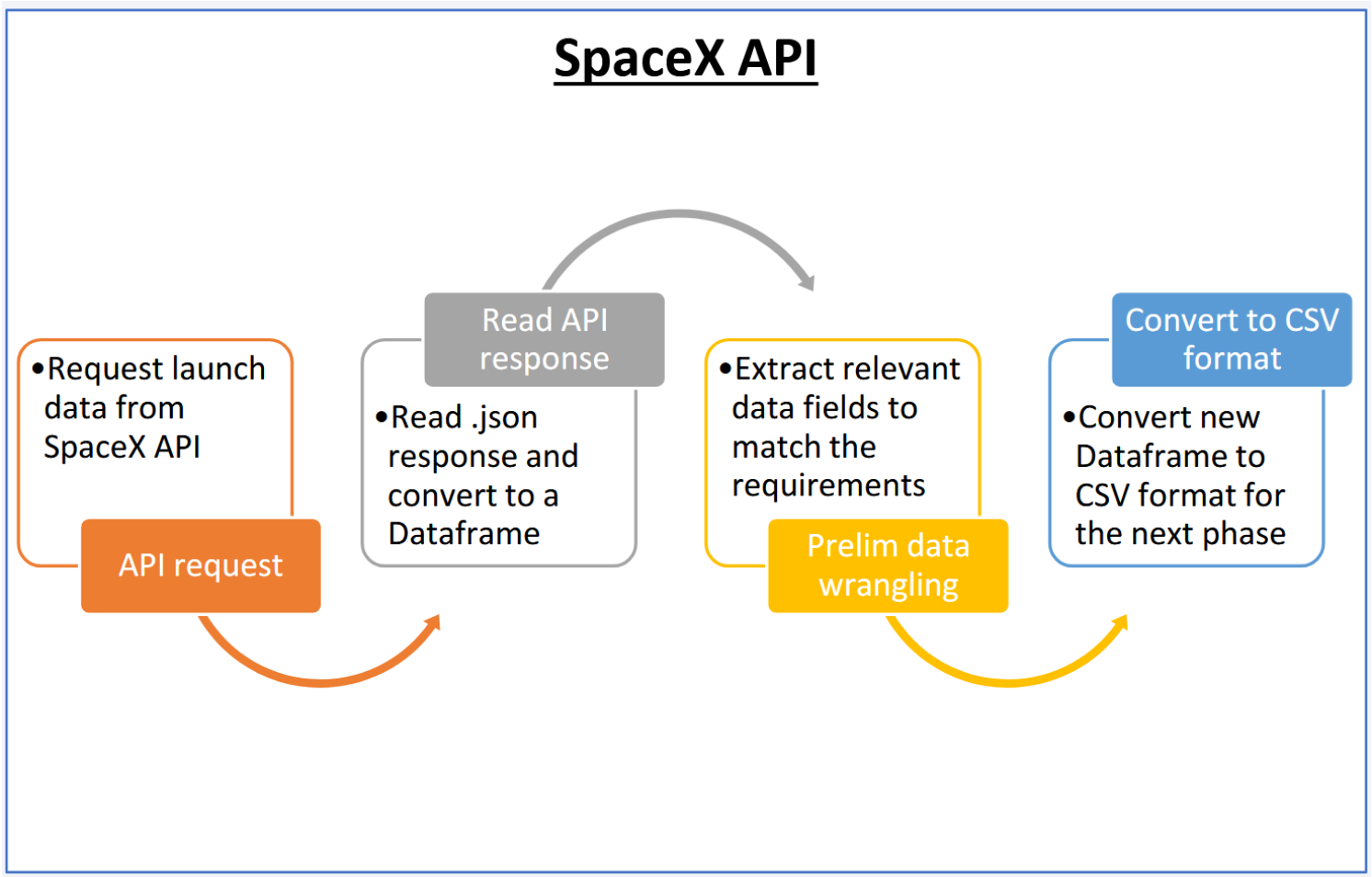
From the **payload** we gain the info of the **mass of the payload** and the **orbit** that it is going to.

From **cores** we get to learn the **outcome of the landing**, the **type of the landing**, **number of flights** with that core, whether gridfins were used, wheter the core is reused, wheter legs were used, the landing pad used, the **block of the core** which is a number used to seperate **version of cores**, the number of times this specific core has been **reused**, and the **serial of the core**.



Data-Science-Specialization-Coursera-/Applied Data Science Capstone/jupyter-labs-spacex-data-collection-api (1).ipynb at main · abbaskhan0345/Data-Science-Specialization-Coursera-

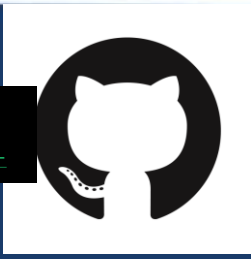
Data Collection SpaceX API



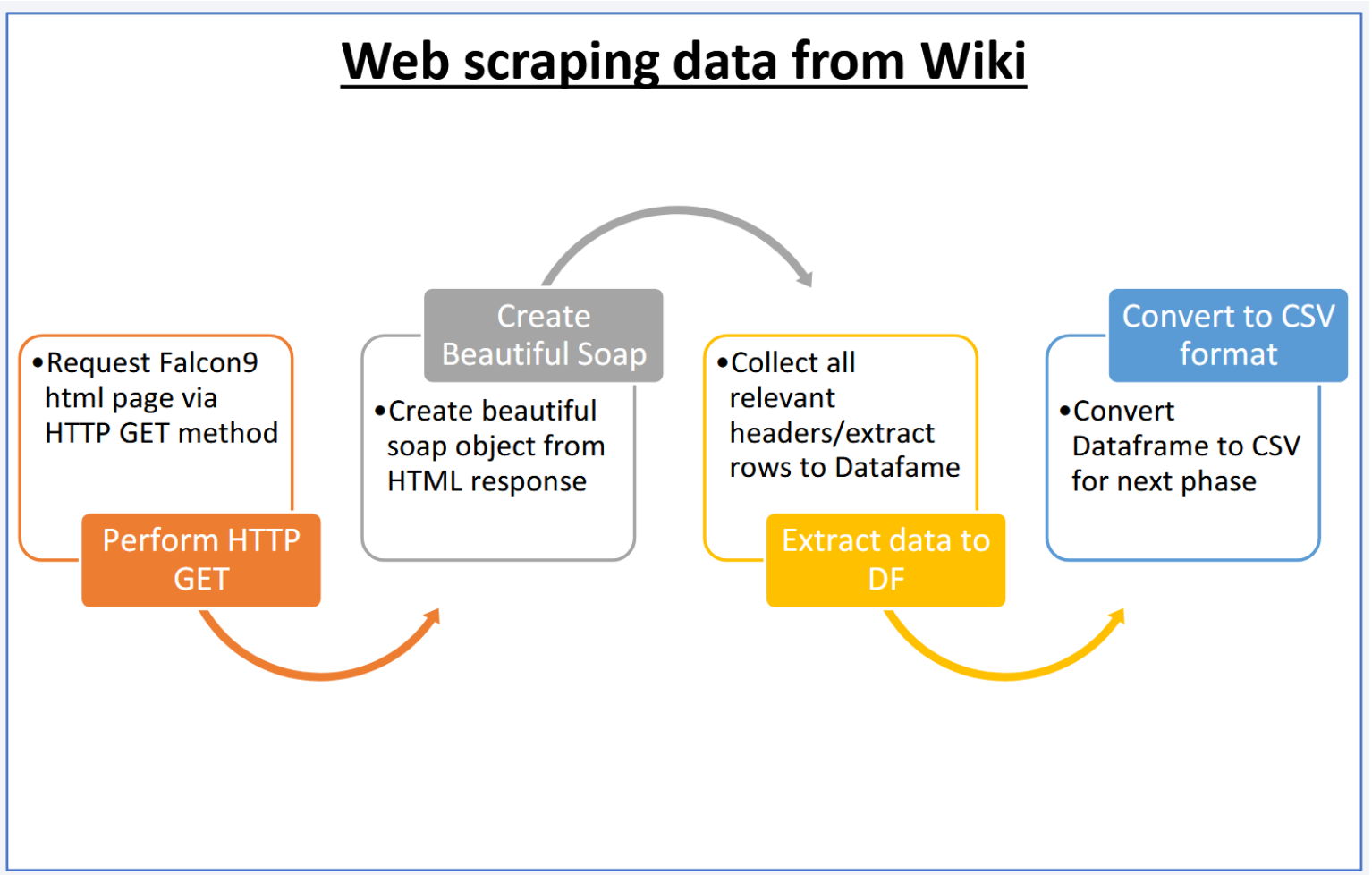
Data Collection - Scraping

- Web scraping are done where we retrieved Falcon 9 and Falcon Heavy Launches Records from Wikipedia.
- Web scraped Falcon 9 launch records with BeautifulSoup:
  - To Extract a Falcon 9 launch records HTML table from Wikipedia
  - To Parse the table and convert it into a Pandas data frame
- Then saved the collected data into CSV

Data-Science-Specialization-Coursera-/Applied Data Science Capstone/jupyter-labs-webscraping (1).ipynb at main · abbaskhan0345/Data-Science-Specialization-Coursera-



Data Collection - Scraping



Data Wrangling

- In summary I Performed exploratory Data Analysis and determine Training Labels
  - Exploratory Data Analysis
  - Determine Training Labels

Data-Science-Specialization-Coursera-/Applied Data Science Capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · abbaskhan0345/Data-Science-Specialization-Coursera-

Data Wrangling

Performed Data Analysis where:

- Loaded the Space X dataset, from last section.
- Identified null values
- Identified and calculated the percentage of the missing values in each attribute
- Identified which columns are numerical and categorical

Calculated the number of launches on each site

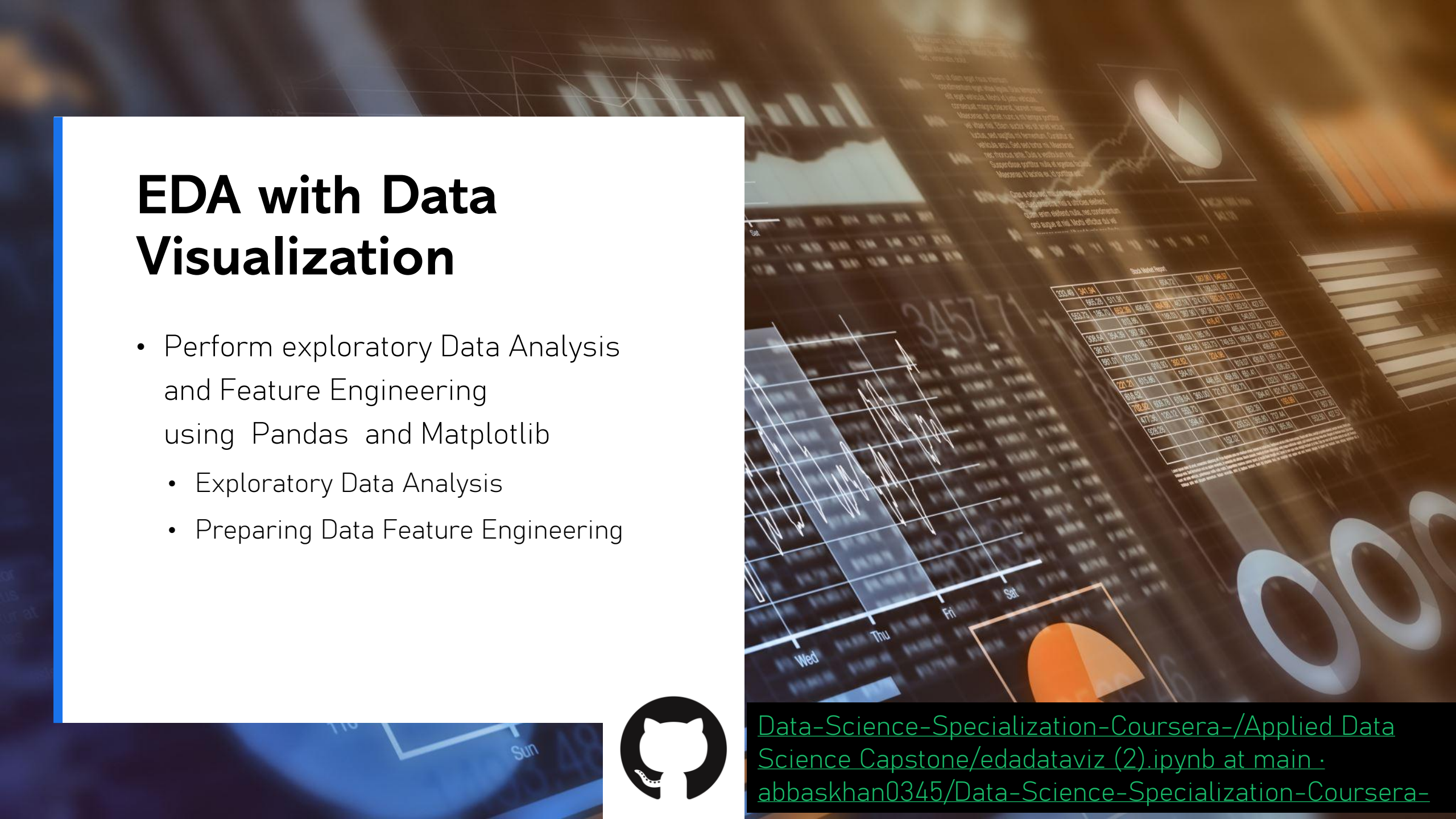
Calculate the number and occurrence of each orbit

Calculated the number and occurence of mission outcome of the orbits

Create a landing outcome label from Outcome column

EDA with Data Visualization

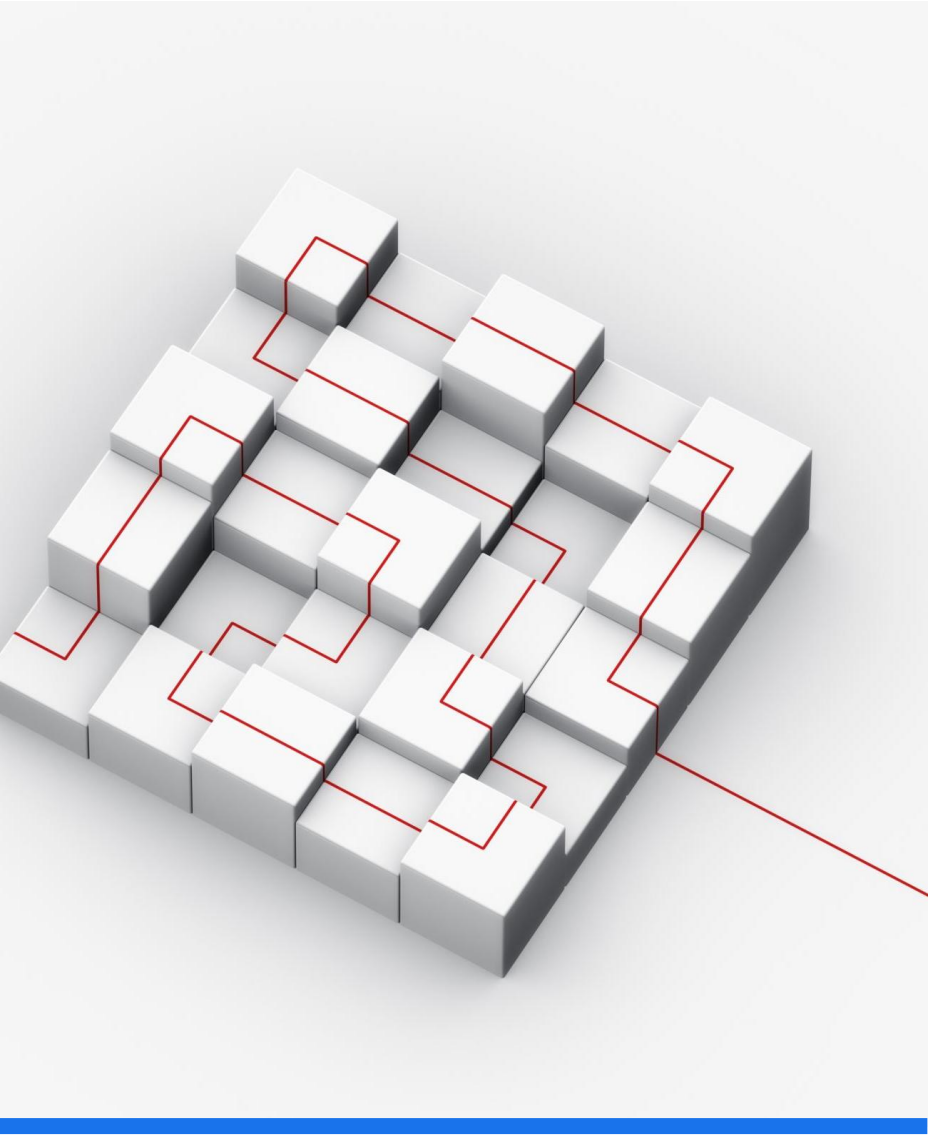
- Perform exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib
  - Exploratory Data Analysis
  - Preparing Data Feature Engineering



Data-Science-Specialization-Coursera-/Applied Data Science Capstone/edadataviz (2).ipynb at main · abbaskhan0345/Data-Science-Specialization-Coursera-

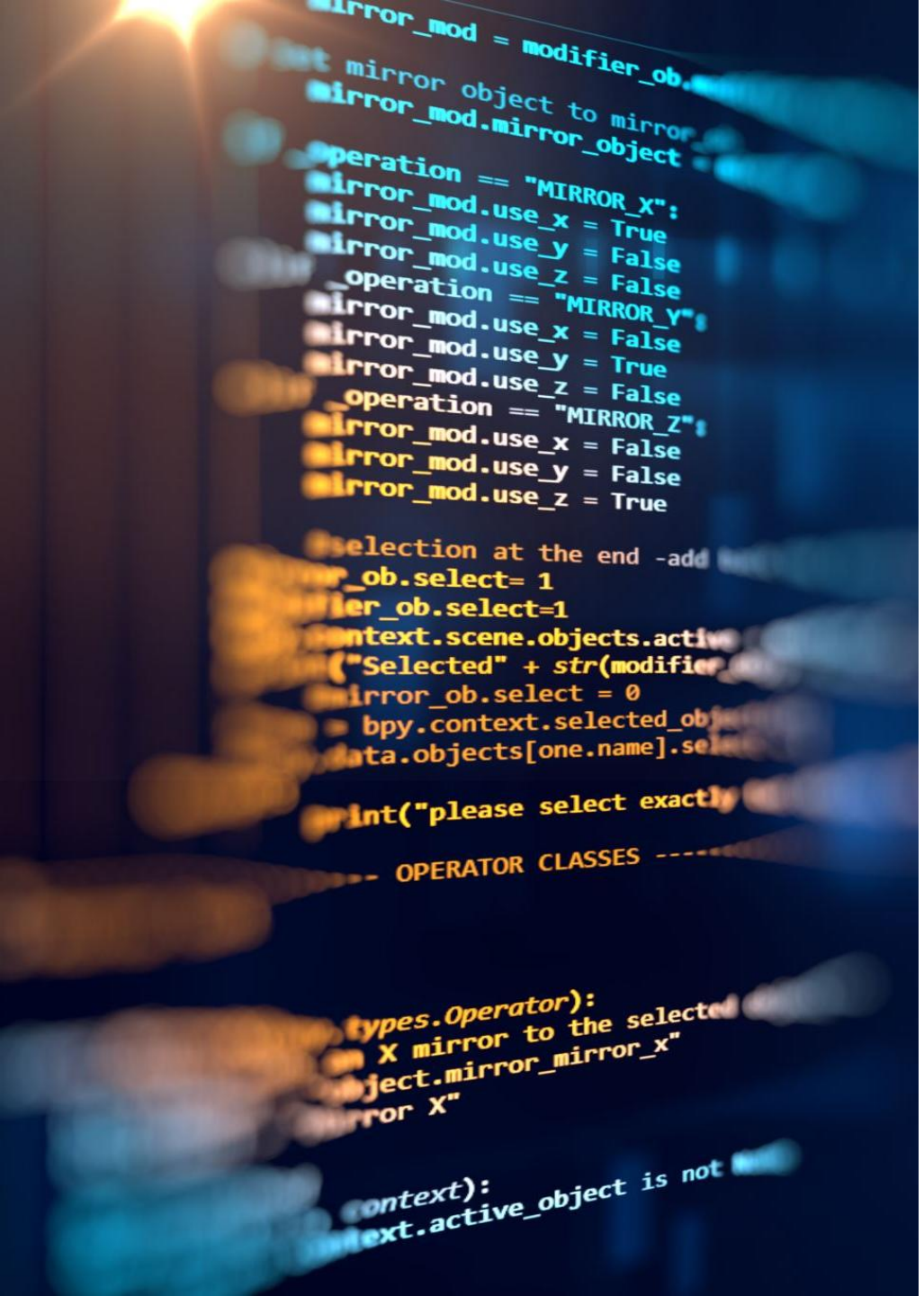
EDA with Data Visualization (Visualization)

- Visualized the catplot for the relationship between Flight Number and Launch Site
- Plotted the scatterplot to Visualize the relationship between Payload Mass and Launch Site
- Bar chart were plotted to Visualize the relationship between success rate of each orbit type
- Another scatter plot was used to Visualize the relationship between FlightNumber and Orbit type
- 3rd scatter plot for the Visualization of relationship between Payload Mass and Orbit type was used.
- A line chart abilities was utilized to Visualize the launch success yearly trend.



EDA with Data Visualization (Feature engineering)

- We obtain some preliminary insights about how each important variable would affect the success rate, we selected the features that will be used in success prediction in the future module.
  - Create dummy variables to categorical columns
  - Cast all numeric columns to float64
  - Feature engineered dataframe was saved to csv file

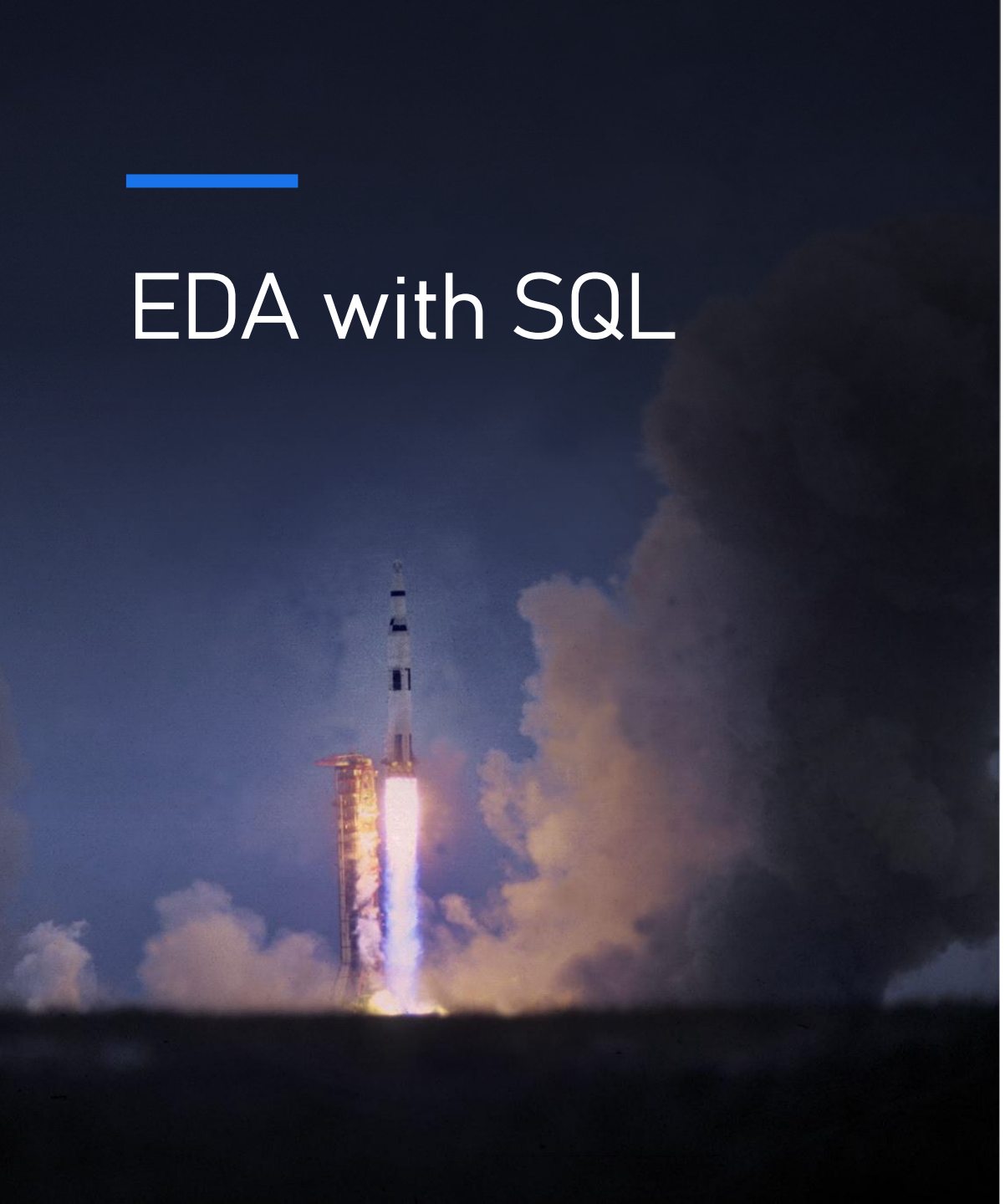


EDA with SQL

- In EDA with sql I:
  - Understand the Spacex DataSet
  - Loaded the dataset into the corresponding table in a Db2 database
  - Executed SQL queries to answer relative questions about data

Data-Science-Specialization-Coursera-/Applied Data Science Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · abbaskhan0345/Data-Science-Specialization-Coursera-

EDA with SQL



- The following steps were performed in EDA via SQL:
    1. Downloaded the datasets
    2. Connected to the database with sqlite3
    3. Displayed the names of the unique launch sites in the space mission
    4. Displayed 5 records where launch sites begin with the string 'CCA'
    5. Displayed the total payload mass carried by boosters launched by NASA (CRS)
    6. Displayed average payload mass carried by booster version F9 v1.1
    7. Listed the date when the first succesful landing outcome in ground pad was acheived.
    8. Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    9. Listed the total number of successful and failure mission outcomes
    10. Listed all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.
    11. 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.
    12. 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.

Build an Interactive Map with Folium

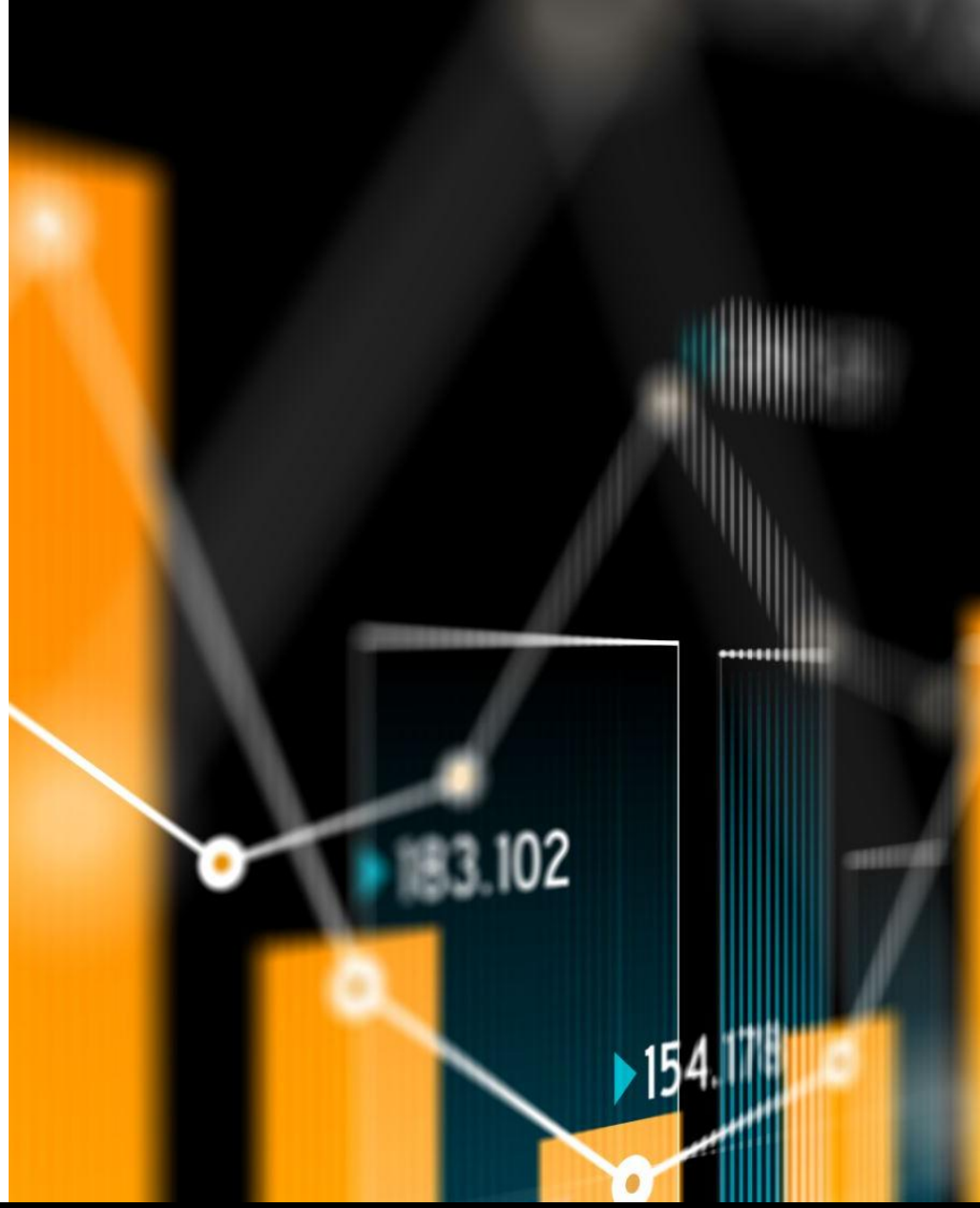
- This lab contains the following tasks we have performed:
  - TASK 1: Mark all launch sites on a map
  - TASK 2: Mark the success/failed launches for each site on the map
  - TASK 3: Calculate the distances between a launch site to its proximities
- After plotting the distance lines to the proximities, we answer the following questions easily:
  - Are launch sites in close proximity to railways? YES
  - Are launch sites in close proximity to highways? YES
  - Are launch sites in close proximity to coastline? YES
  - Do launch sites keep certain distance away from cities? YES

Data-Science-Specialization-Coursera-/Applied Data Science Capstone/lab_jupyter_launch_site_location.ipynb at main · abbaskhan0345/Data-Science-Specialization-Coursera-

Build a Dashboard with Plotly Dash

· Built a Plotly Dash web application to perform interactive visual analytics on SpaceX launch

- data in real-time. Added Launch Site Drop-down, Pie Chart, Payload range slide, and a
- Scatter chart to the Dashboard.
- 1. Added a Launch Site Drop-down Input component to the dashboard to provide an ability to filter
- Dashboard visual by all launch sites or a particular launch site
- 2. Added a Pie Chart to the Dashboard to show total success launches when 'All Sites' is selected and
- show success and failed counts when a particular site is selected
- 3. Added a Payload range slider to the Dashboard to easily select different payload ranges to identify
- visual patterns
- 4. Added a Scatter chart to observe how payload may be correlated with mission outcomes for selected
- site(s). The color-label Booster version on each scatter point provided missions outcomes with
- different boosters



Data-Science-Specialization-Coursera-/Applied Data Science Capstone/spacex-dash-app.py at main · abbaskhan0345/Data-Science-Specialization-Coursera-

Build a Dashboard with Plotly Dash

· Dashboard helped answer following questions:

- 1. Which site has the largest successful launches? KSC LC-39A with 10
- 2. Which site has the highest launch success rate? KSC LC-39A with 76.9% success
- 3. Which payload range(s) has the highest launch success rate? 2000 – 5000 kg
- 4. Which payload range(s) has the lowest launch success rate? 0-2000 and 5500 - 7000
- 5. Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate? FT

Predictive Analysis (Classification)

Performed exploratory Data Analysis and determine Training Labels

- created a column for the class
- Standardized the data
- Split into training data and test data

-Finded best Hyperparameter for SVM, Classification Trees and Logistic Regression

Finded the method performs best using test data

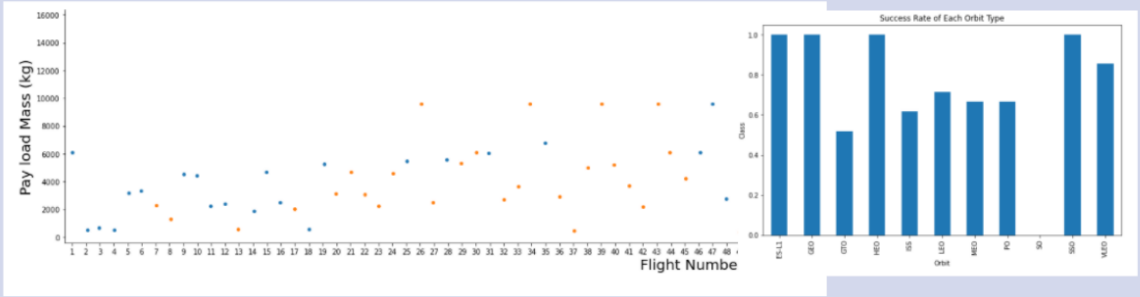
Data-Science-Specialization-Coursera-/Applied Data Science Capstone/SpaceX_Machine Learning

Results

Following sections and slides explain results for:

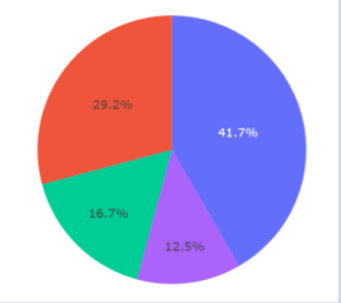
Exploratory data analysis results

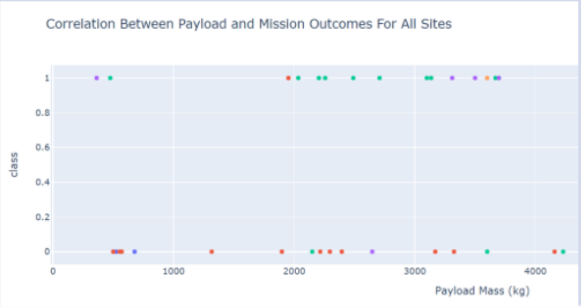
- Samples:



Interactive analytics demo in screenshots

- Samples

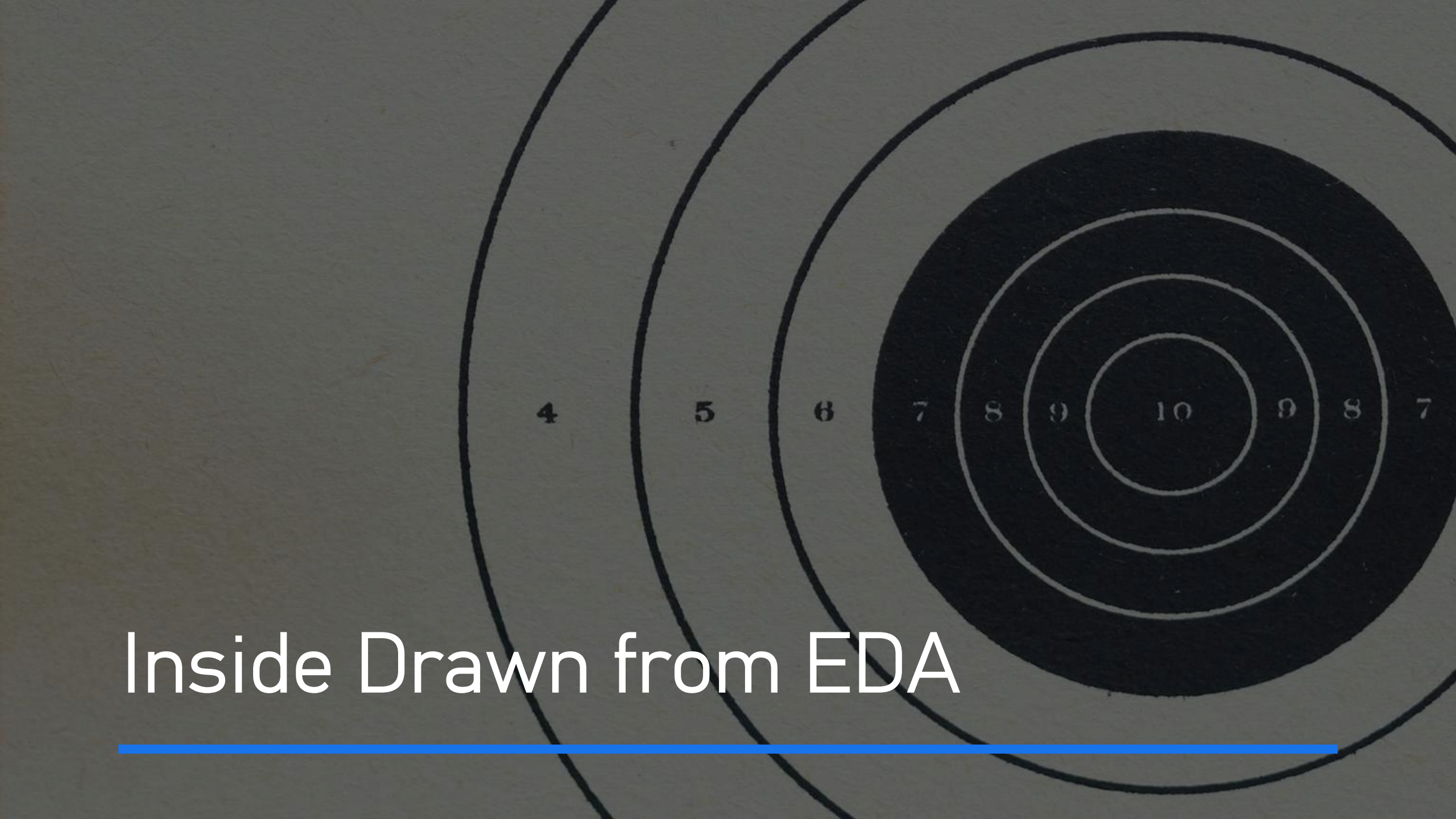




Predictive analysis results

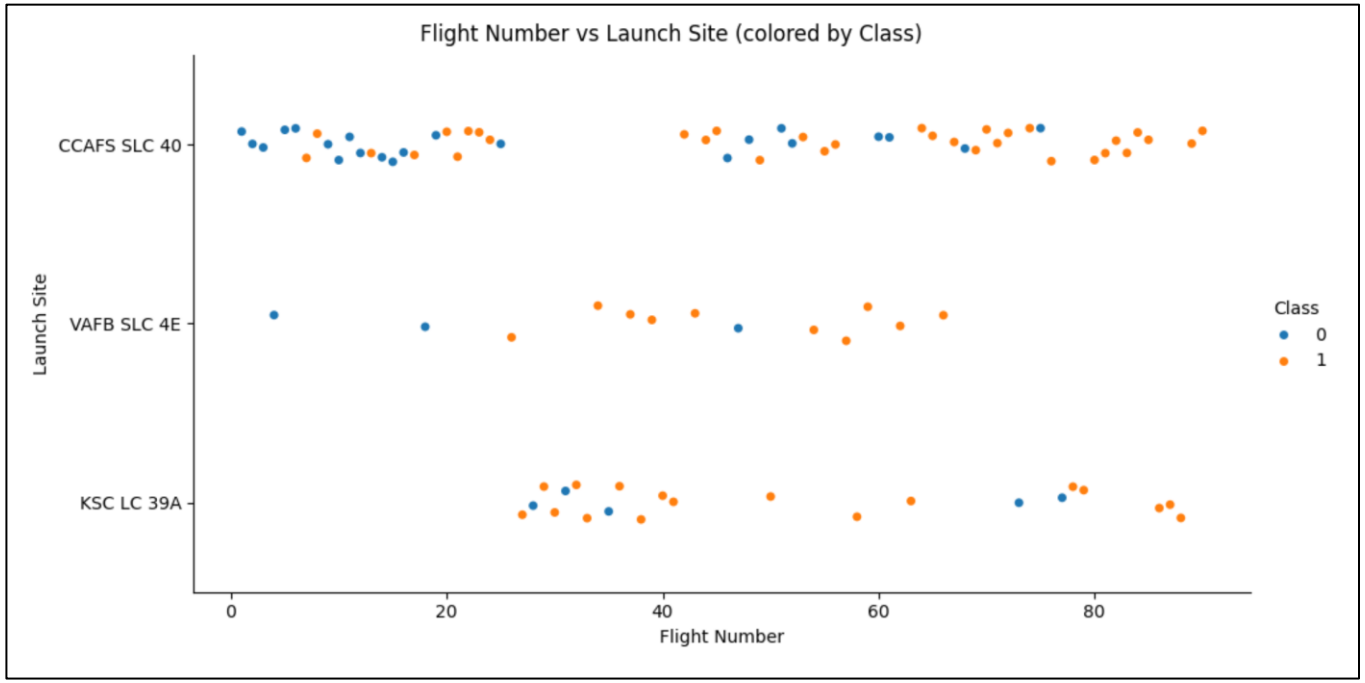
- Samples

	Algo Type	Accuracy Score
2	Decision Tree	0.903571
3	KNN	0.848214
1	SVM	0.848214
0	Logistic Regression	0.846429

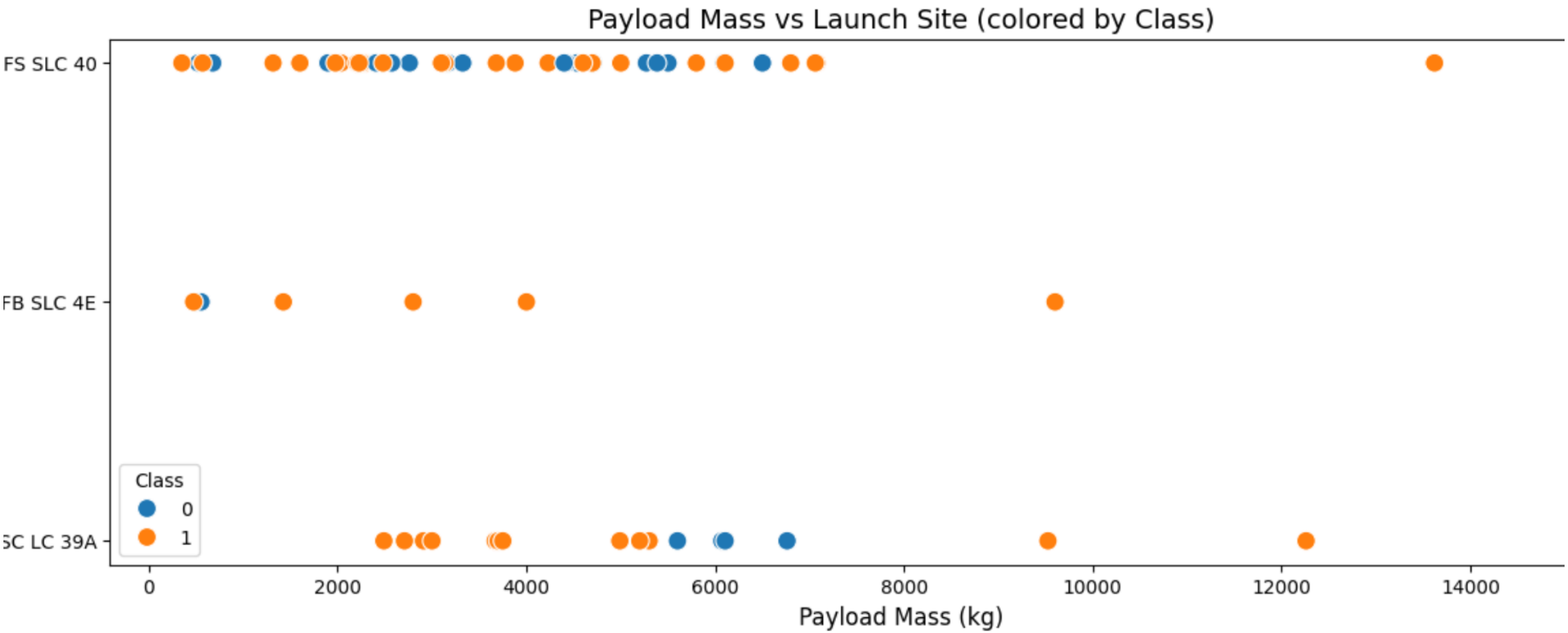


Inside Drawn from EDA

Flight Number vs. Launch Site



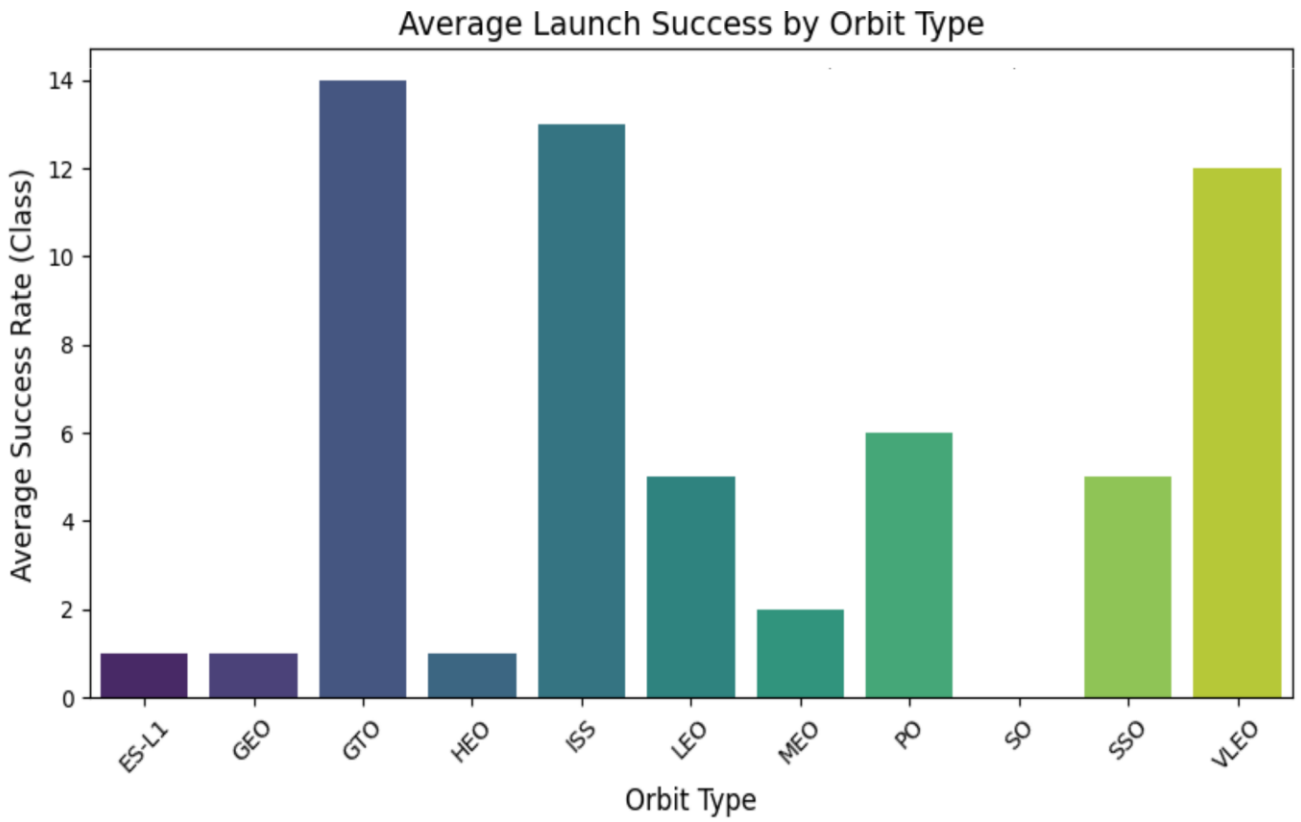
- As we can see most of the flight are done from CCAFS SLC 40. With the approximately equal number of class 0 and 1.
- For the least number of flight that are launch from VAFB SLC 4E. With only 3 class 0 and 9 class 1.
- With the most successful the KSC LC 39A class 1 and less equal to 5 class 0.



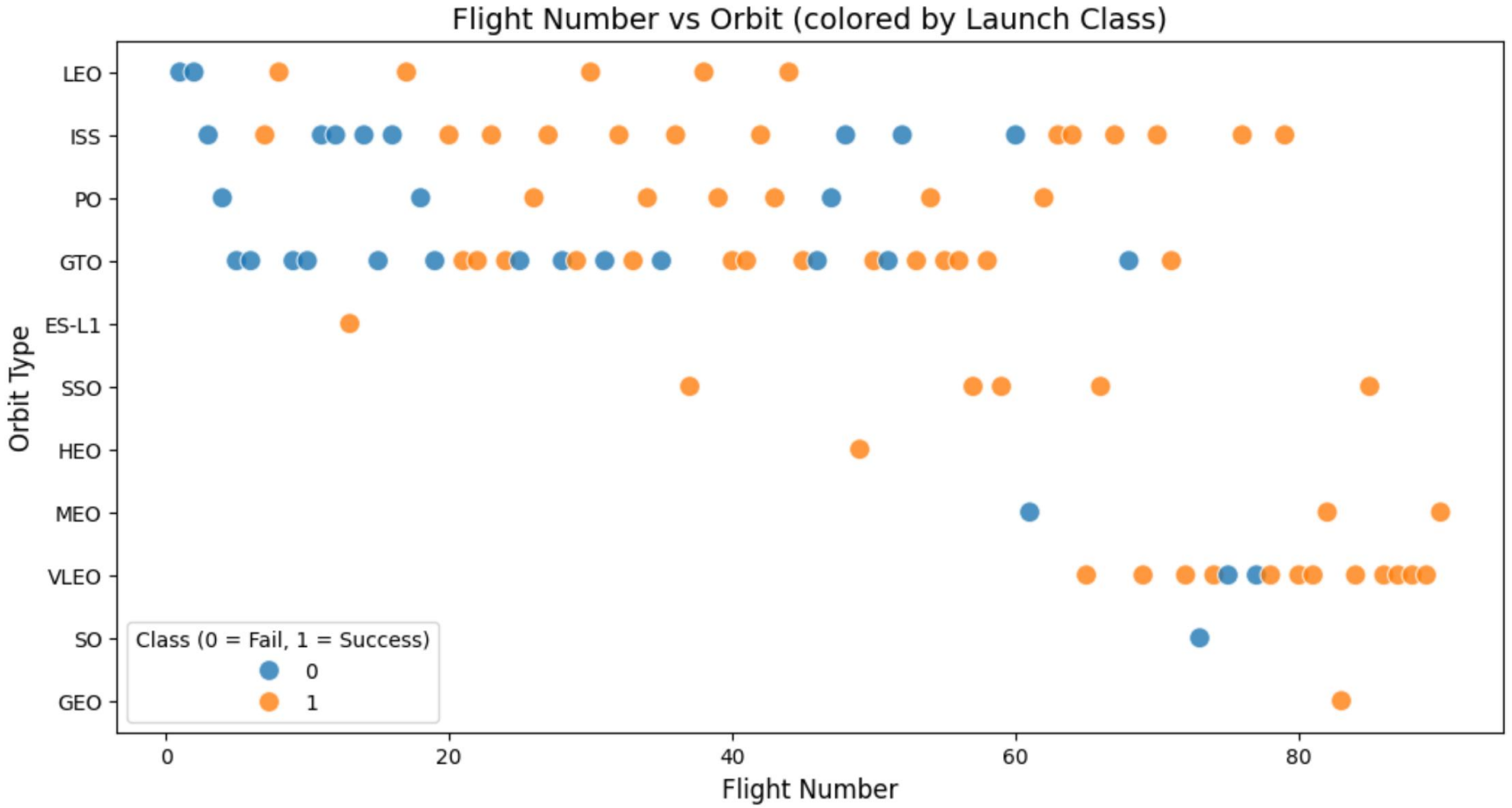
Payload vs. Launch Site

- When the payload mass increases the chances of failure relatively increases. But that is not always the case.

Success Rate vs. Orbit Type



- The orbit also effect the success of the launch as we can see the GTO, ISS and VLEO has highest success rate while the ES-L1, GEO, HEO and SO has very little success rate.

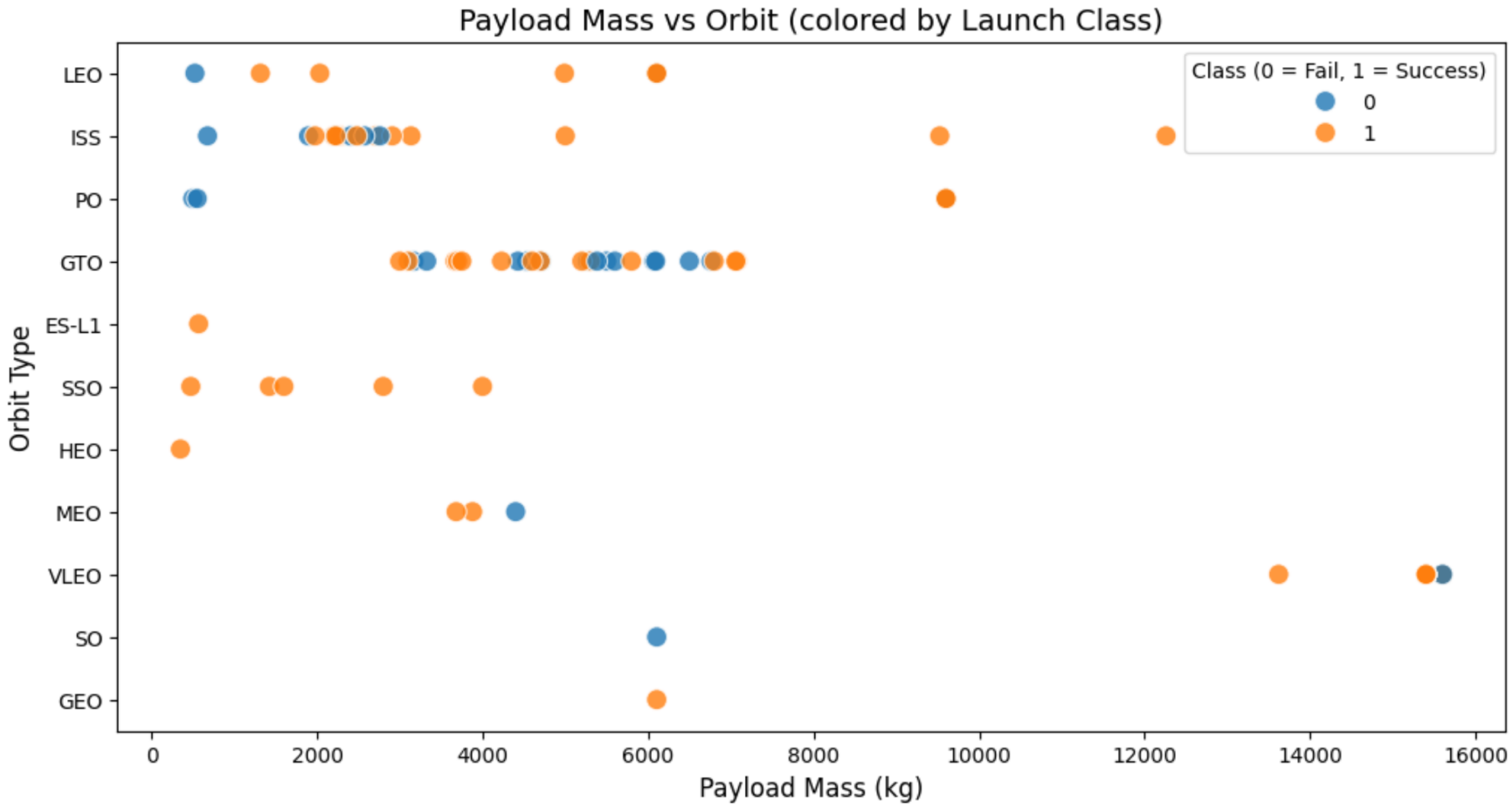


Flight Number vs. Orbit Type

THE SAME EFFECT OF ORBIT TYPE IS CAN BE SEEN THE HERE IN THE FOLLOWING CLASS.

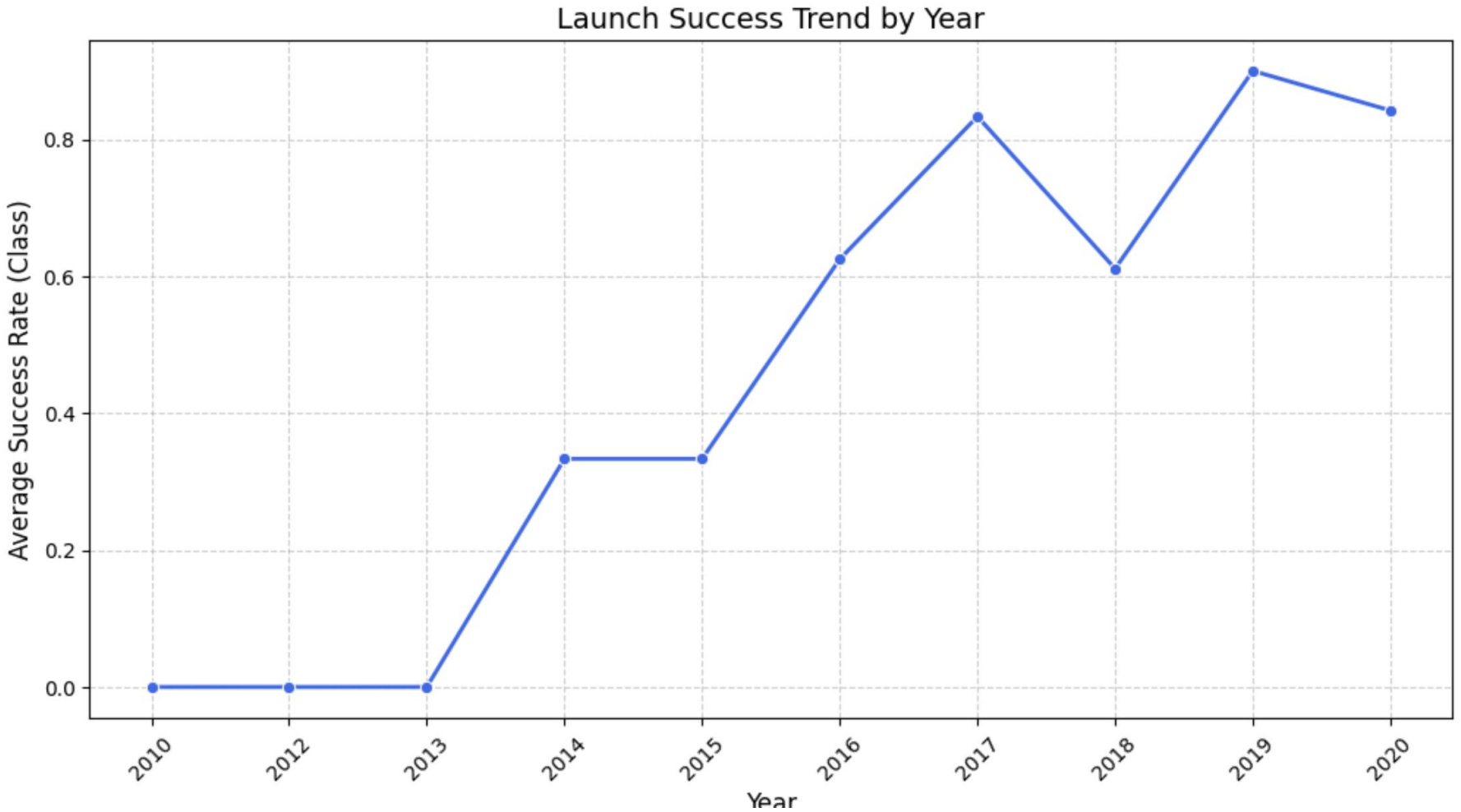
Payload vs. Orbit Type

THE MASS IS DIFFERENT FOR EVERY TYPE OF ORBIT FOR IT SUCCESSFUL.



Launch Success Yearly Trend

AS WE CAN SEE THAT THE SUCCESS TREND IS GOING HIGH WITH TIME.



All Launch Site Names

```
In [13]: %sql select distinct(Launch_Site) from SPACEXTBL
```

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

Out[13]:

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

Launch Site Names Begin with 'CCA'

Task 2

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

In [21]:
```
%sql select * from SPACEXTBL where Launch_Site LIKE '%CCA%' limit 5;
```

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

Out[21]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcom
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachut
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	Failure (parachut
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

Total Payload Mass

- %sql select sum(PAYLOAD_MASS__KG_) as ' Payload mass carried by boosters launched by NASA (CRS)' from SPACEXTBL where Customer = 'NASA (CRS)';

Task 3

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

In [26]:
```
%sql select sum(PAYLOAD_MASS__KG_) as ' Payload mass carried by boosters launched by NASA (CRS)' from SPACEXTBL  where Cus
```

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

Out[26]:

payload mass carried by boosters launched by NASA (CRS)
45596

Task 4

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

```
%%sql select avg(PAYLOAD_MASS__KG_) as 'Average payload mass carried by boo
            from SPACEXTBL where Booster_Version like '%F9 v1.0%'
```

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

Average payload mass carried by booster version F9 v1.1
340.4

Average Payload Mass by F9 v1.1

First Successful Ground Landing Date

```
%sql select min(Date), [Time (UTC)] from SPACEXTBL where Landing_Outcome like '%Success%';
```

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

min(Date)	Time (UTC)
2015-12-22	1:29:00

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

```
%%sql SELECT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)' and Booster_Version in
(SELECT Booster_Version
      FROM SPACEXTBL where PAYLOAD_MASS__KG_ > 4000
        AND PAYLOAD_MASS__KG_ < 6000 );
```

```
 * 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

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

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

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

Boosters Carried Maximum Payload

```
%%sql select distinct(Booster_Version) as 'List all the booster_versions that have carried the maximum payload mass'
from SPACEXTBL
    where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

List all the booster_versions that have carried the maximum payload mass
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 select substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL
    where Date like '%2015%' and Landing_Outcome = (select Landing_Outcome from SPACEXTBL where Landing_Outcome = 'Failure
```

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

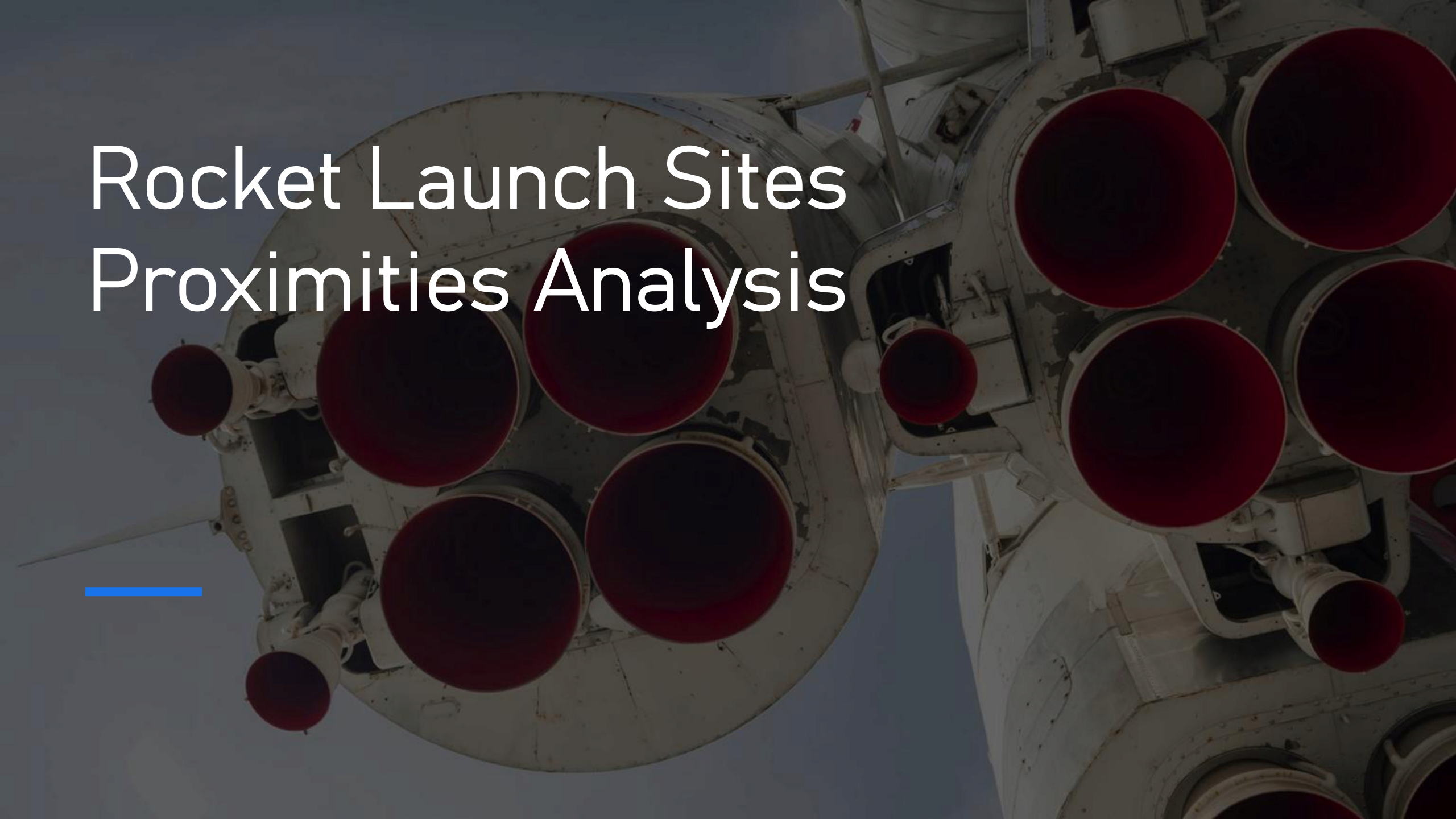
Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Total
FROM SPACEXTBL
WHERE Landing_Outcome IN (
    SELECT DISTINCT Landing_Outcome
    FROM SPACEXTBL
    WHERE Date >= '2010-06-04' AND Date <= '2017-03-20'
)
GROUP BY Landing_Outcome
ORDER BY Total DESC;
```

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

Landing_Outcome	Total
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1



Rocket Launch Sites Proximities Analysis

SpaceX Falcon9 - Launch Sites Map

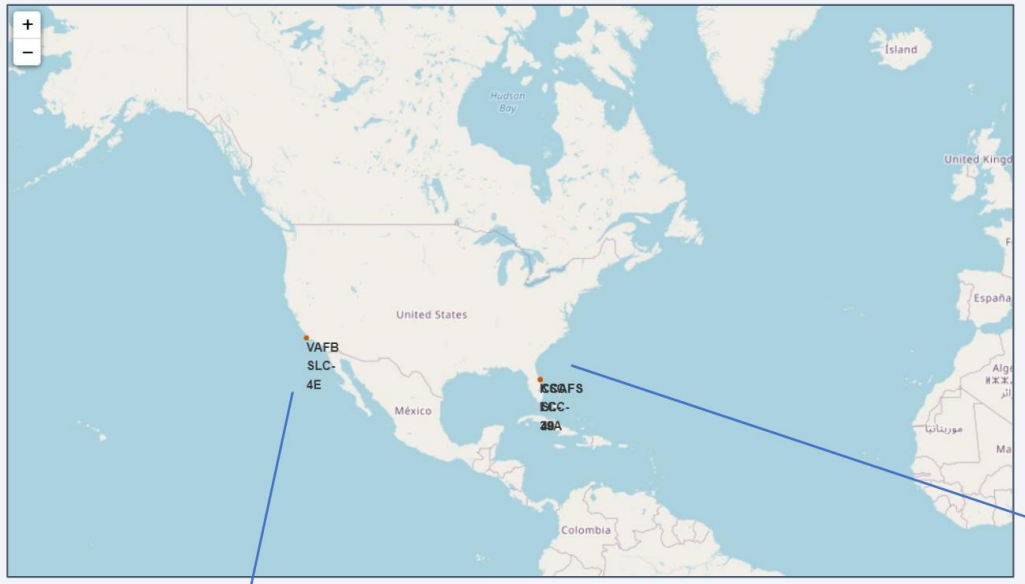


Fig 1 – Global Map

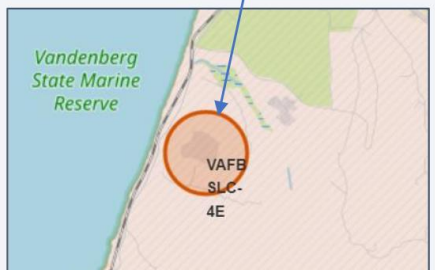


Figure 1 on left displays the Global map with Falcon 9 launch sites that are located in the United States (in California and Florida). Each launch site contains a circle, label, and a popup to highlight the location and the name of the launch site. It is also evident that all launch sites are near the coast.

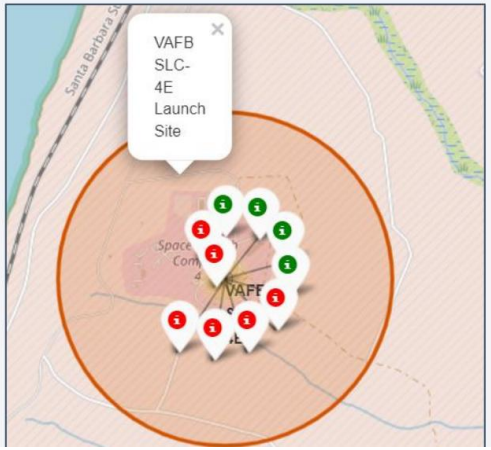
Figure 2 and Figure 3 zoom in to the launch sites to display 4 launch sites:

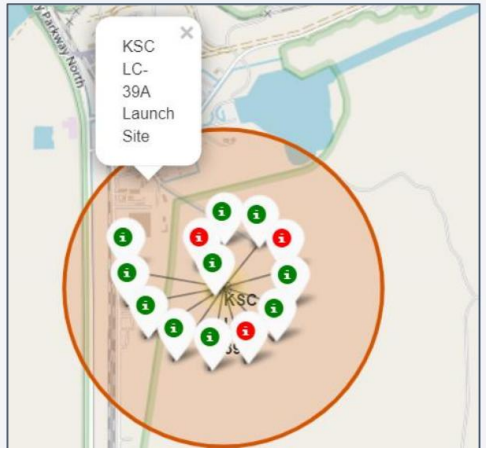
- VAFB SLC-4E (CA)
- CCAFS LC-40 (FL)
- KSC LC-39A (FL)
- CCAFS SLC-40 (FL)

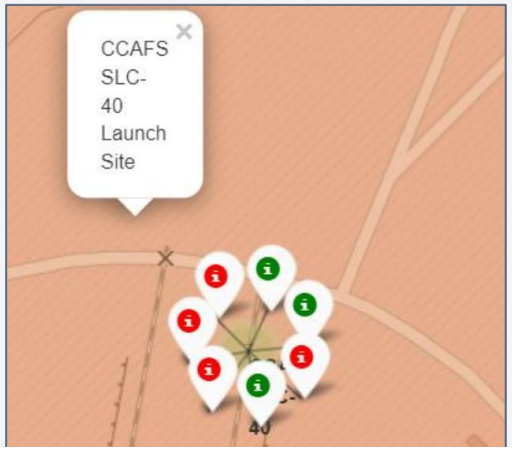
SpaceX Falon9 – Success/Failed Launch Map for all Launch Sites

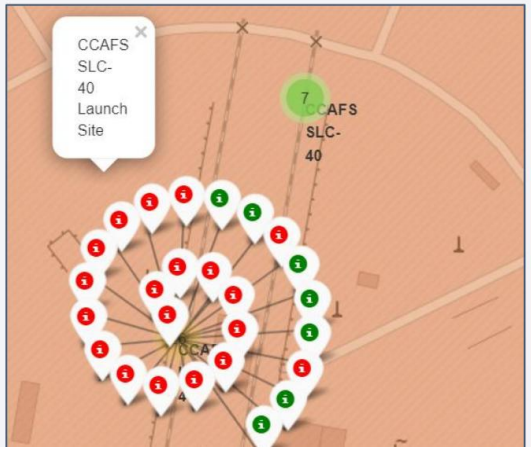
Fig 1 – US map with all Launch Sites

- Figure 1 is the US map with all the Launch Sites. The numbers on each site depict the total number of successful and failed launches
- Figure 2, 3, 4, and 5 zoom in to each site and displays the success/fail markers with green as success and red as failed
- By looking at each site map, KSC LC-39A Launch Site has the greatest number of successful launches









SpaceX Falcon9 – Launch Site to proximity Distance Map

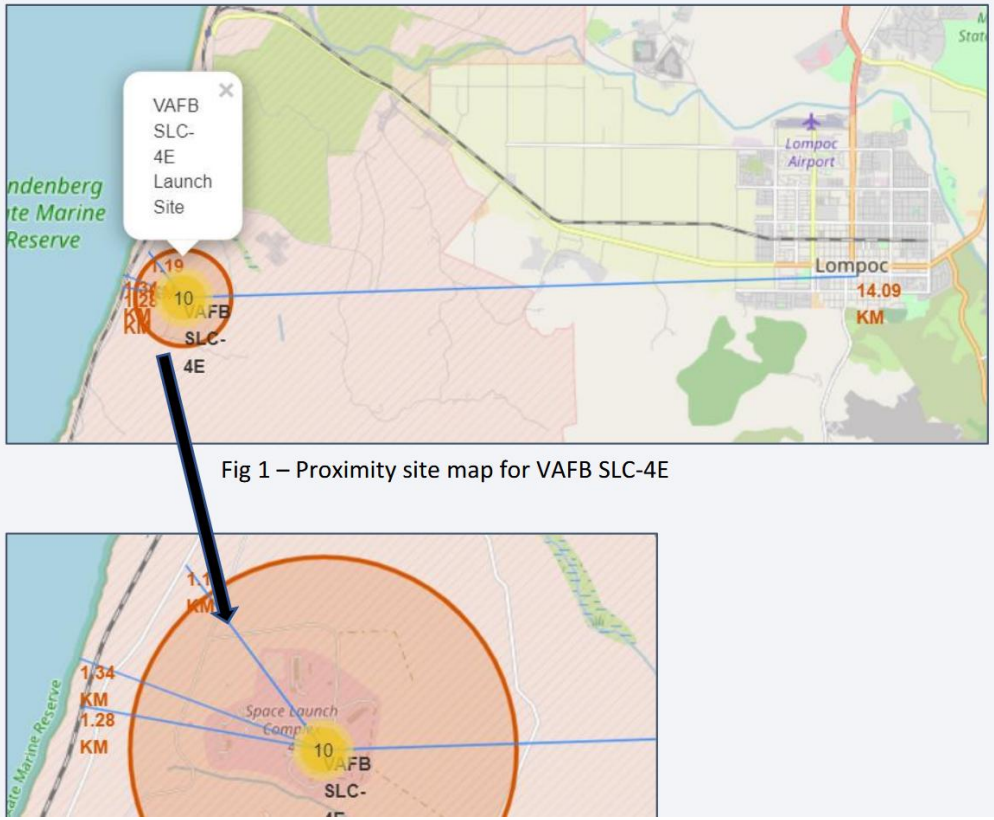


Fig 1 – Proximity site map for VAFB SLC-4E

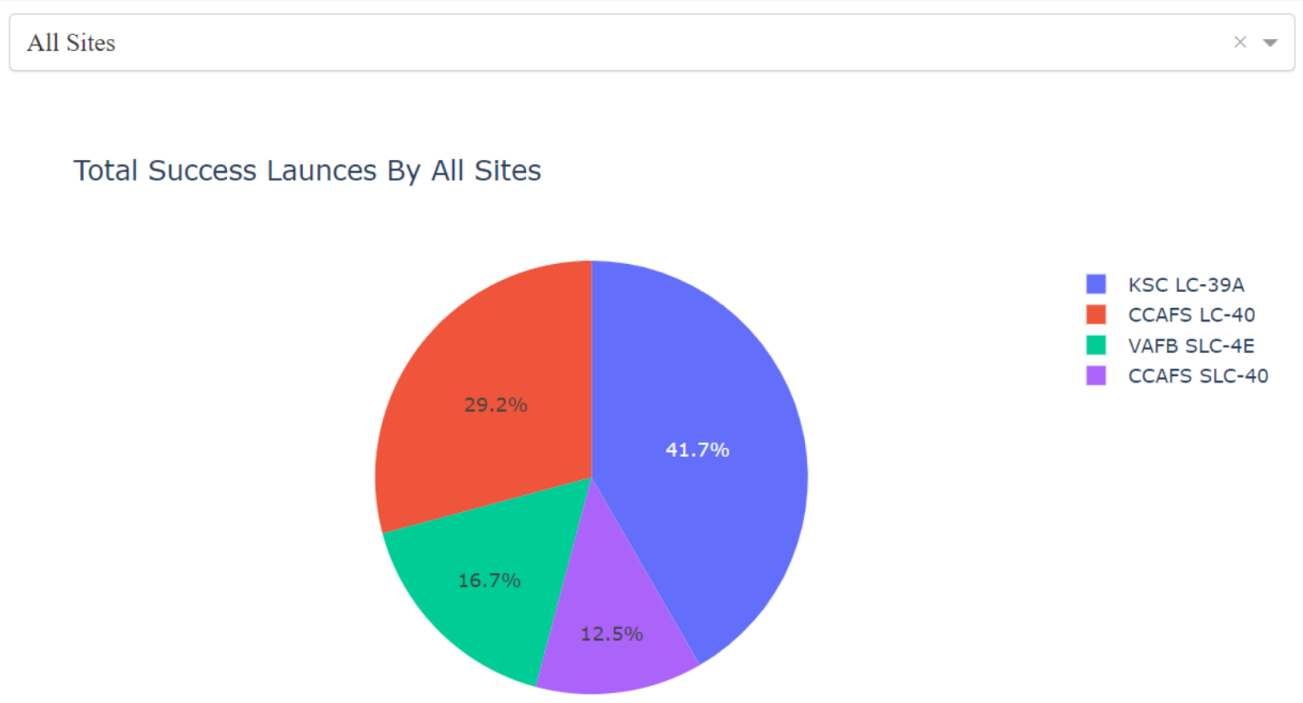
Figure 1 displays all the proximity sites marked on the map for Launch Site VAFB SLC-4E. City Lompoc is located further away from Launch Site compared to other proximities such as coastline, railroad, highway, etc. The map also displays a marker with city distance from the Launch Site (14.09 km)

Figure 2 provides a zoom in view into other proximities such as coastline, railroad, and highway with respective distances from the Launch Site

In general, cities are located away from the Launch Sites to minimize impacts of any accidental impacts to the general public and infrastructure. Launch Sites are strategically located near the coastline, railroad, and highways to provide easy access to resources.

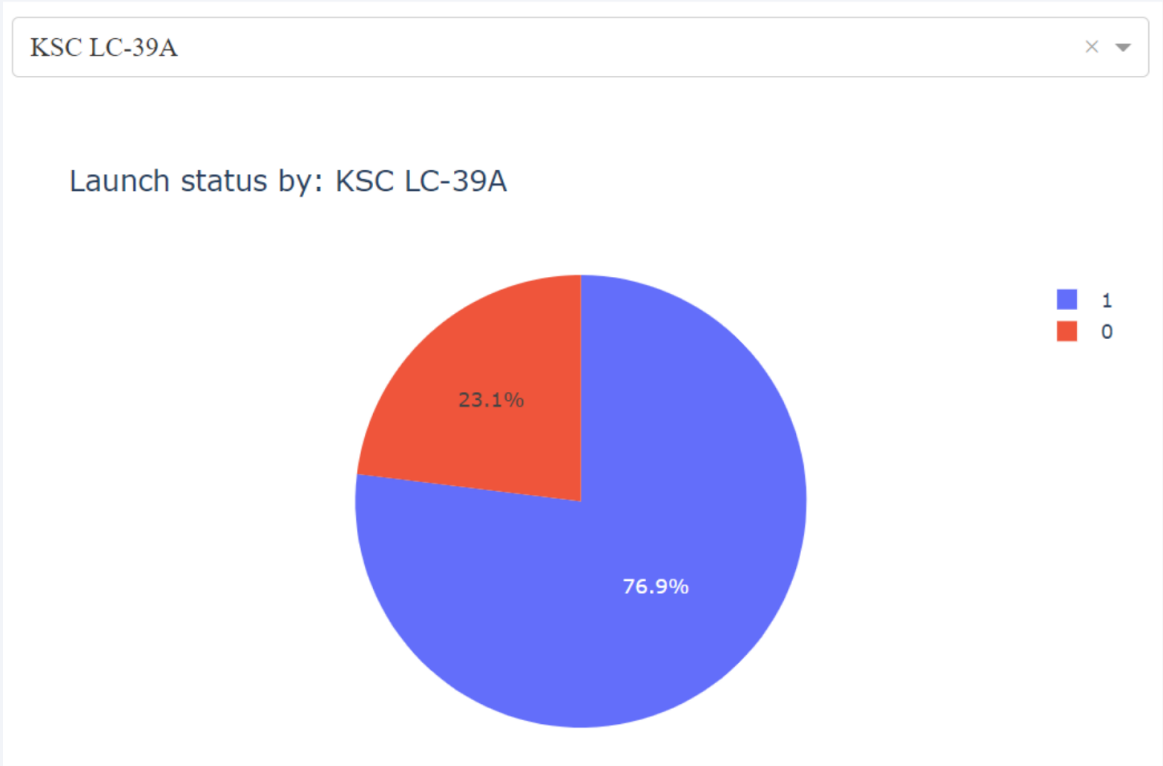
Build a Dashboard with Plotly Dash

Launch Success Counts For All Sites



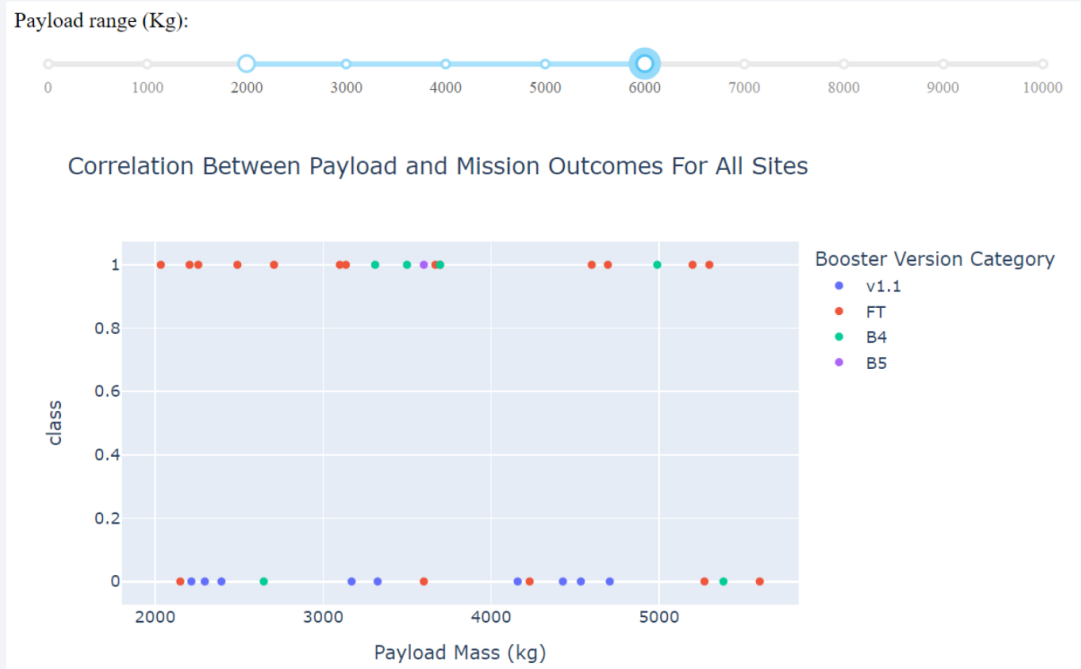
- Launch Site ‘KSC LC-39A’ has the highest launch success rate
- Launch Site ‘CCAFS SLC-40’ has the lowest launch success rate

Launch Site with Highest Launch Success Ratio



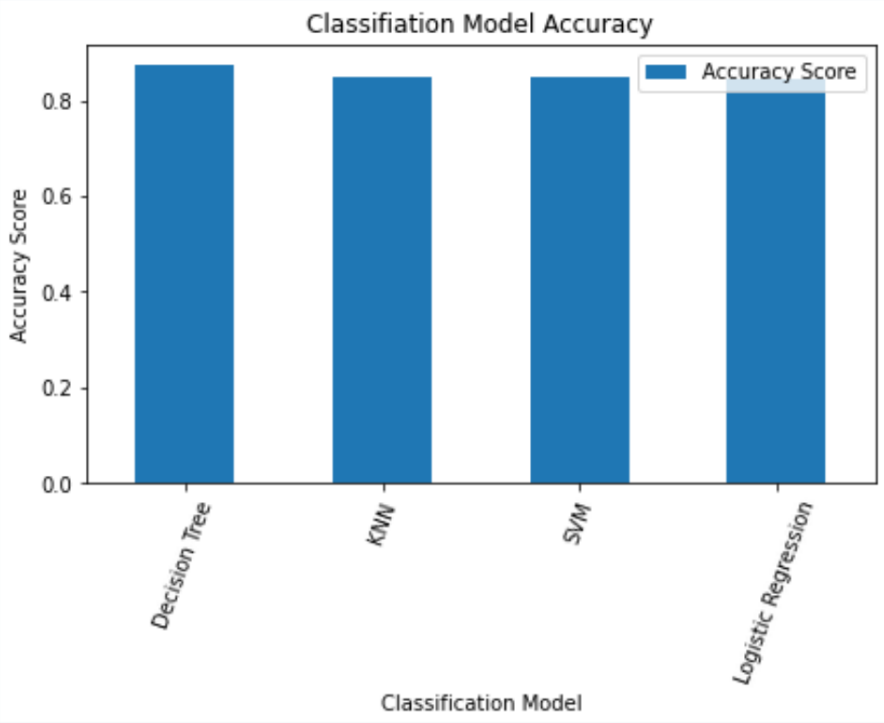
- KSC LC-39A Launch Site has the highest launch success rate and count
- Launch success rate is 76.9%
- Launch success failure rate is 23.1%

Payload vs. Launch Outcome Scatter Plot for All Sites



- Most successful launches are in the payload range from 2000 to about 5500
- Booster version category ‘FT’ has the most successful launches
- Only booster with a success launch when payload is greater than 6k is ‘B4’

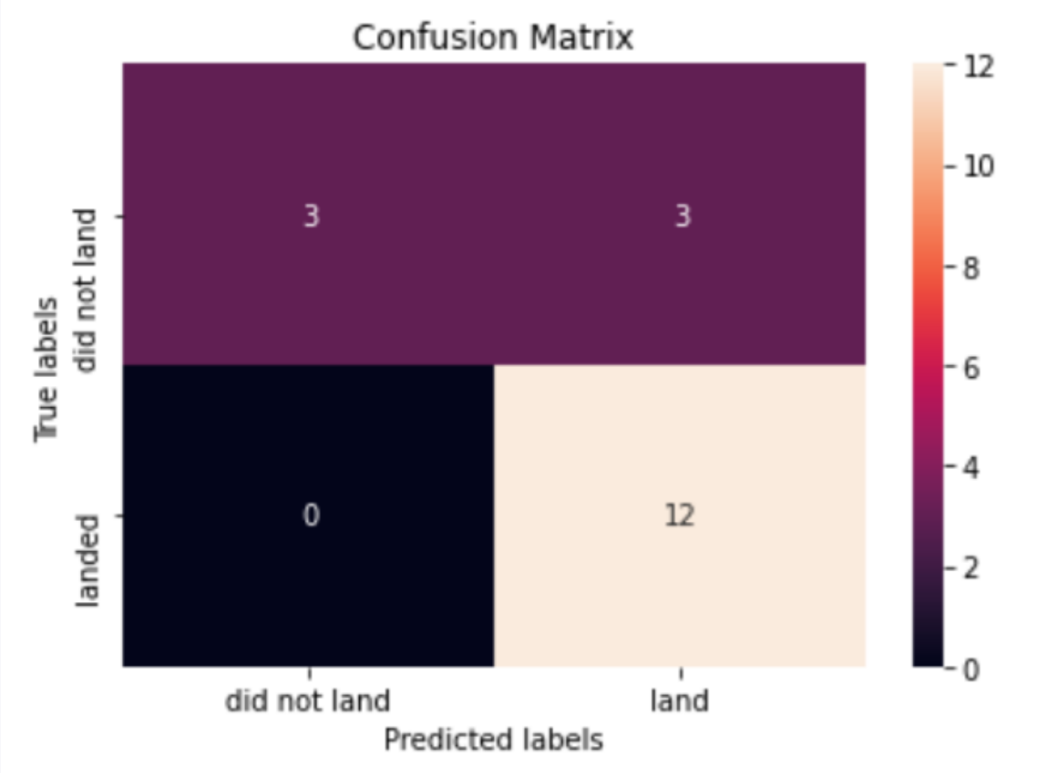
Classification Accuracy



	Algo Type	Accuracy Score	Test Data Accuracy Score
2	Decision Tree	0.875000	0.833333
3	KNN	0.848214	0.833333
1	SVM	0.848214	0.833333
0	Logistic Regression	0.846429	0.833333

- Based on the Accuracy scores and as also evident from the bar chart, Decision Tree algorithm has the highest classification score with a value of .8750
- Accuracy Score on the test data is the same for all the classification algorithms based on the data set with a value of .8333
- Given that the Accuracy scores for Classfication algorithms are very close and the test scores are the same, we may need a broader data set to further tune the models

Confusion Matrix



- The confusion matrix is same for all the models (LR, SVM, Decision Tree, KNN)
- Per the confusion matrix, the classifier made 18 predictions
- 12 scenarios were predicted Yes for landing, and they did land successfully (True positive)
- 3 scenarios (top left) were predicted No for landing, and they did not land (True negative)
- 3 scenarios (top right) were predicted Yes for landing, but they did not land successfully (False positive)
- Overall, the classifier is correct about 83% of the time ((TP + TN) / Total) with a misclassification or error rate ((FP + FN) / Total) of about 16.5%

Conclusions

- As the numbers of flights increase, the first stage is more likely to land successfully
- Success rates appear go up as Payload increases but there is no clear correlation between Payload mass and success rates
- Launch success rate increased by about 80% from 2013 to 2020
- Launch Site ‘KSC LC-39A’ has the highest launch success rate and Launch Site ‘CCAFS SLC-40’ has the lowest launch success rate
- Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rates and orbit GTO the lowest
- Lunch sites are located strategically away from the cities and closer to coastline, railroads, and highways
- The best performing Machine Learning Classfication Model is the Decision Tree with an accuracy of about 87.5%. When the models were scored on the test data, the accuracy score was about 83% for all models. More data may be needed to further tune the models and find a potential better fit.

Thank you All for your Time.

Connect With me:

Linkedin:
linkedin.com/in/hazrat-abbas-khan-113136329

Github:
github.com/abbaskhan0345

Gmail:
abbaskhan0345060@gmail.com