IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Archismita Ghosh

12/10/2024

Outline

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

Executive Summary

Summary of Methodologies :

The project aims to predict if the Falcon 9 first stage will land successfully and how various factors influence it. In order to achieve this goal, the following methodologies have been used :

- **Data Collection** using SpaceX REST API and web scrapping technique.
- **Data Wrangling** to create success/fail outcome variable.
- **Exploratory Data analysis with SQL** for calculating various statistics like total payload, total number of successful and failure mission outcomes, etc.
- **Exploratory Data Analysis with data visualization** techniques considering various factors.
- **Interactive visualization of the data with Folium** to mark launch sites, success/failed launches for each site and calculate distances between launch site to its proximities.
- **Build Interactive Dashboard** to analyze launch records.
- **Build Models** to predict landing outcomes using Logistic Regression, Support Vector Machine (SVM), Decision Tree and K-Nearest Neighbor (KNN).

Executive Summary

Summary of Results:

Exploratory Data Analysis :

- The success rate of launches have increased over the years.
- The orbits ES-L1, GEO, HEO, SSO have a 100% success rate.

Visualisation/Analytics :

- Most launch sites are in close proximity to coastlines, railways and highways but are far from cities.

Predictive Analysis :

- All models performed similarly on the test set. But, the Decision Tree performed the best among them.

Introduction

Project background and context

In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Problems you want to find answers

- How factors like payload mass, launch site, number of flights, orbit type and so on, affect first-stage landing success
- Rate of successful landings over time and determining best factors to ensure successful launch
- Best predictive model that can be used to predict successful landing

Section 1

Methodology

Methodology

1. Data collection methodology:

The data was collected using SpaceX REST API and by web scraping on Wikipedia webpages.

2. Perform data wrangling

The data was processed using Pandas and NumPy, and some of the main methods used are : Removal of unnecessary columns, OneHot encoding, data normalization and standardization.

3. Perform exploratory data analysis (EDA) using visualization and SQL

Libraries like Matplotlib and Seaborn were used for data visualization and SQL for querying data

4. Perform interactive visual analytics using Folium and Plotly Dash

Folium and Plotly Dash were used for Interactive visual analytics and building dashboard

5. Perform predictive analysis using classification models

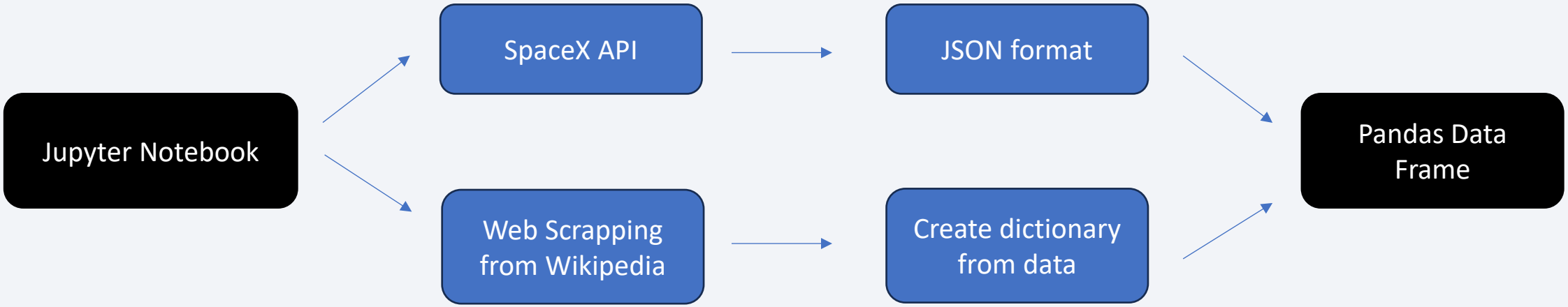
Data was split into train and test sets, followed by identifying the best algorithm and parameters through hyperparameters tuning using Grid Search. Finally the best algorithm and parameters were used for model deployment.

Data Collection

Data has been collected from two main sources :

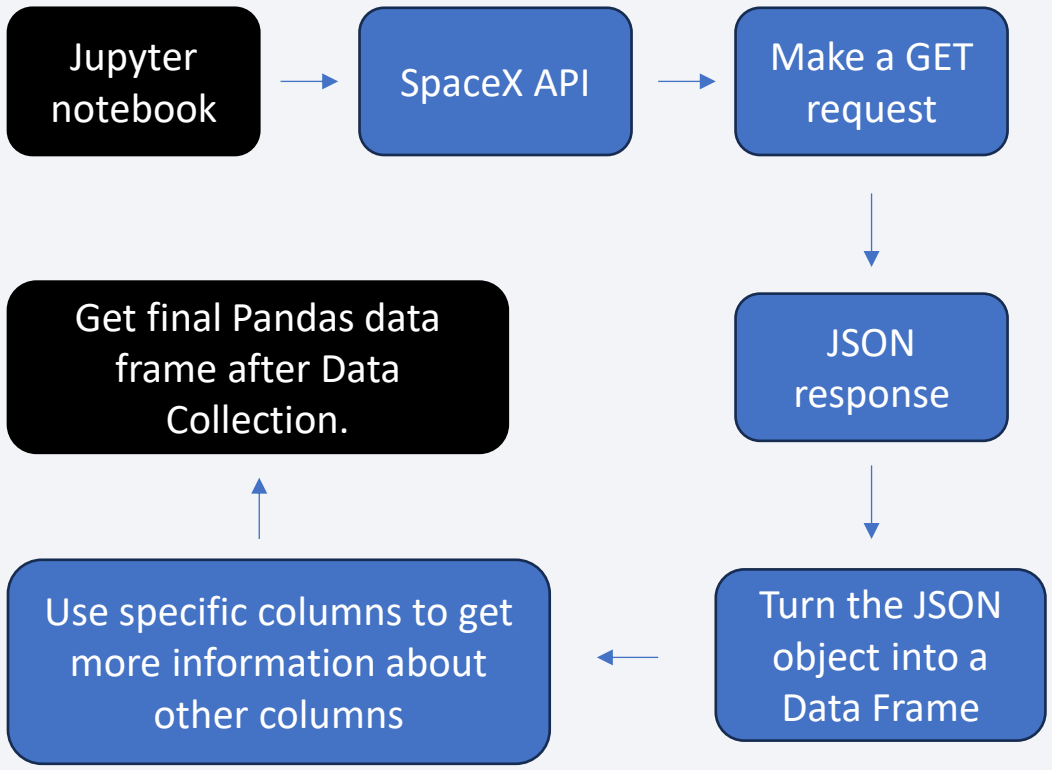
1. SpaceX REST API : Open source REST API for rocket launch, core, payload, launchpad data
2. Wikipedia webpage

Data collection process flowchart :



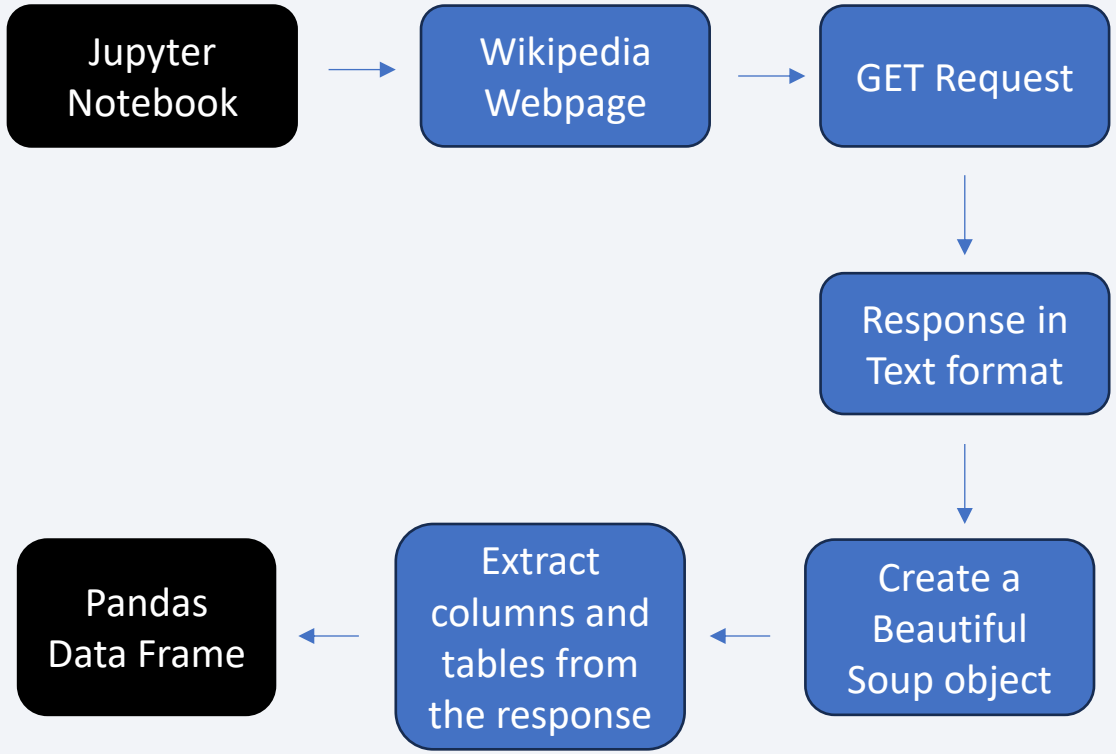
Data Collection – SpaceX API

- Required Libraries like Requests, Pandas, NumPy were imported and a GET request was made to the API. The requested JSON results was turned into a Pandas data frame using .json_normalize(). Some columns like rocket, payload, launchpad, cores were used to get more information and finally another new data frame was created with all the information.
- GitHub URL : Link



Data Collection - Scraping

- Import BeautifulSoup, Pandas and Request libraries. Here, a Wikipedia page has been used as source where again GET request is made. Beautiful soup was used to extract tables and columns from the response (in text format) and finally converted to a Pandas Data frame
- GitHub URL : Link



Data Wrangling

- Pandas and NumPy libraries were imported. The data collected from the previous stage is loaded to clean the data and perform exploratory data analysis.
- GitHub URL : Link

1. Load the collected data into a Pandas Data Frame
2. Identify and calculate the percentage of missing values in each attribute
3. Identify which columns are numerical and categorical
4. Calculate the number of launches on each site
5. Calculate the number and occurrence of each orbit
6. Create a landing outcome label from Outcome column
7. Determine the success rate

EDA with Data Visualization

- In this stage, the aim was to find the relationship between the features and target and visualize their relationship using Seaborn and Matplotlib. Lastly, feature engineering has also been performed by converting categorical values to dummy values and casting the numerical columns.
- GitHub URL : Link

1. Visualise the relationship between Flight Number and Launch Site
2. Visualise the relationship between Payload and Launch Site
3. Visualise the relationship between success rate of each orbit type
4. Visualise the relationship between Flight Number and Orbit type
5. Visualise the relationship between Payload and Orbit type
6. Visualise the launch success yearly trend

EDA with SQL

- In this stage, further analysis of the data is done with SQL. The important SQL queries done here have been listed on the right
- GitHub URL : Link

SQL Queries

- 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 carries by boosters launched by NASA (CRS)
- Display the average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in grand pad was achieved
- 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
- List the records which will display month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

Build an Interactive Map with Folium

- Here, the Folium library was imported and used to visualize the geospatial data in the form of interactive maps, and draw markers, circles and lines to represent other important landmarks and information on the map.
- GitHub URL : Link

Tasks performed with the Interactive Map with Folium

- The 4 launch sites belonging to Falcon 9 rocket launches have been marked with 4 Circles on the map. The information of the launch sites is as given below :

Launch Site	Latitude	Longitude
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610746

- The successful/failed outcomes on each site ahs been marked using marker objects (MarkerCluster) assigned specific colours based on 'class' attribute.
- The distances between the launch site to its proximities closest city, railways, highways, coastline was calculated and Lines were drawn to represent these distances using PolyLine object.

Build a Dashboard with Plotly Dash

In this stage, the following tasks have been performed :

- A dropdown list was added to select Launch Site with the following options :

All sites, CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A

- A Pie chart was added to show the total successful and failed launches of the launch sites based on the selection
- A Range slider was added to select Payload Range between 0 – 10000
- A Scatter Plot was added to show the correlation between Payload and Launch success at different sites.

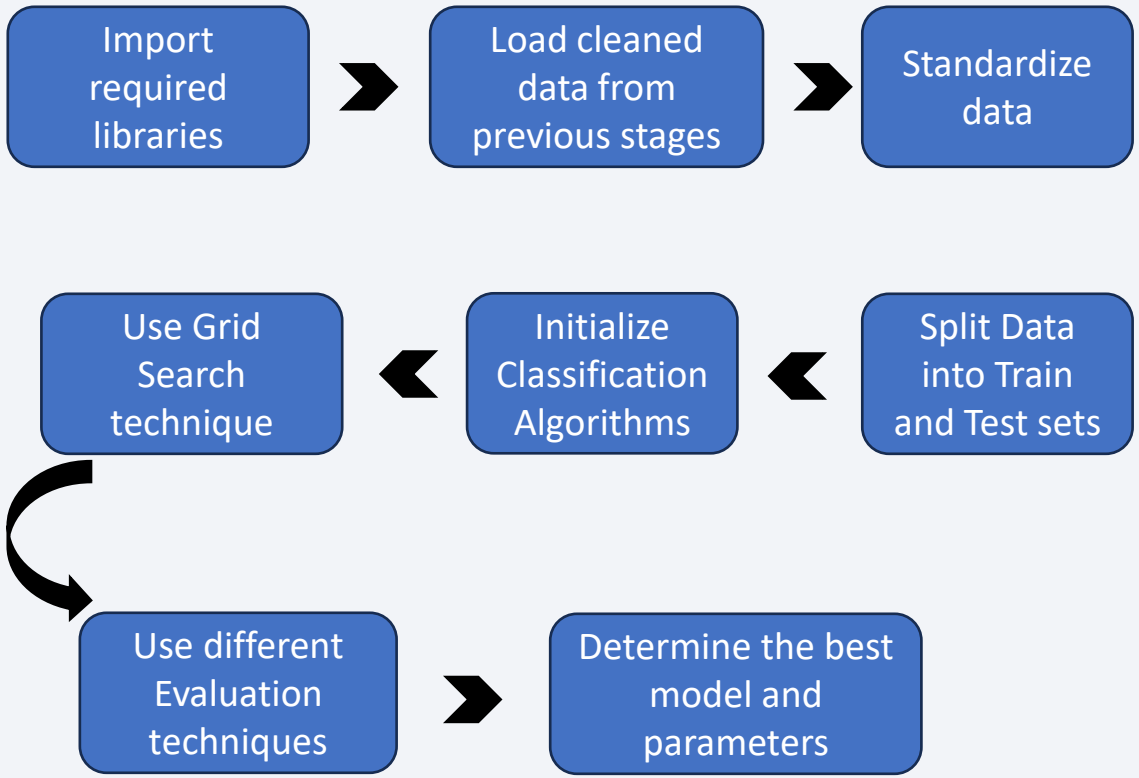
- GitHub URL : Link

Predictive Analysis (Classification)

Stages :

1. Import all required libraries for this stage
2. Load the cleaned data obtained from previous stages
3. Standardize the data and then split the data with 20% for testing set and rest 80% for training set
4. Initialize the 4 different classification algorithms :
   - Logistic Regression (LR)
   - Support Vector Machine (SVM)
   - Decision Tree
   - K Nearest Neighbors (KNN)
5. Use Grid Search technique to determine the best parameters
6. Use Confusion Matrix, F1 score, Jaccard score to evaluate and find the best model among all the algorithms that have been tried.

GitHub URL : Link

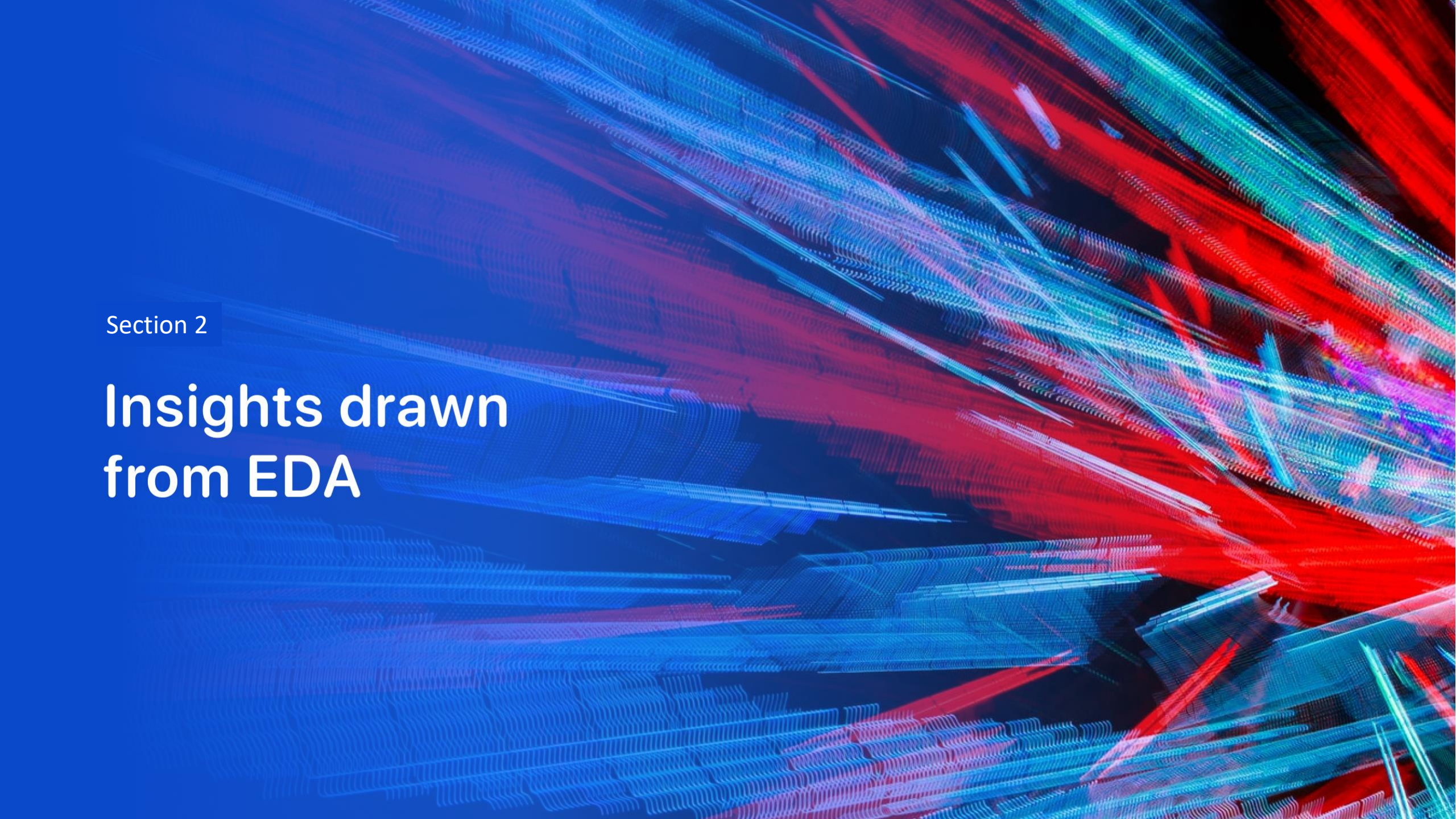


Results

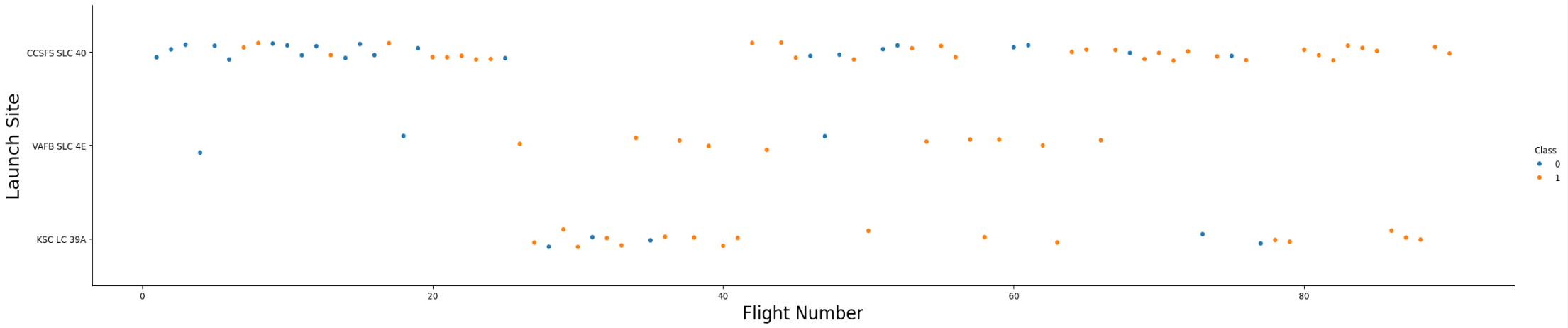
- Exploratory data analysis results

- Interactive analytics demo in screenshots

- Predictive analysis results

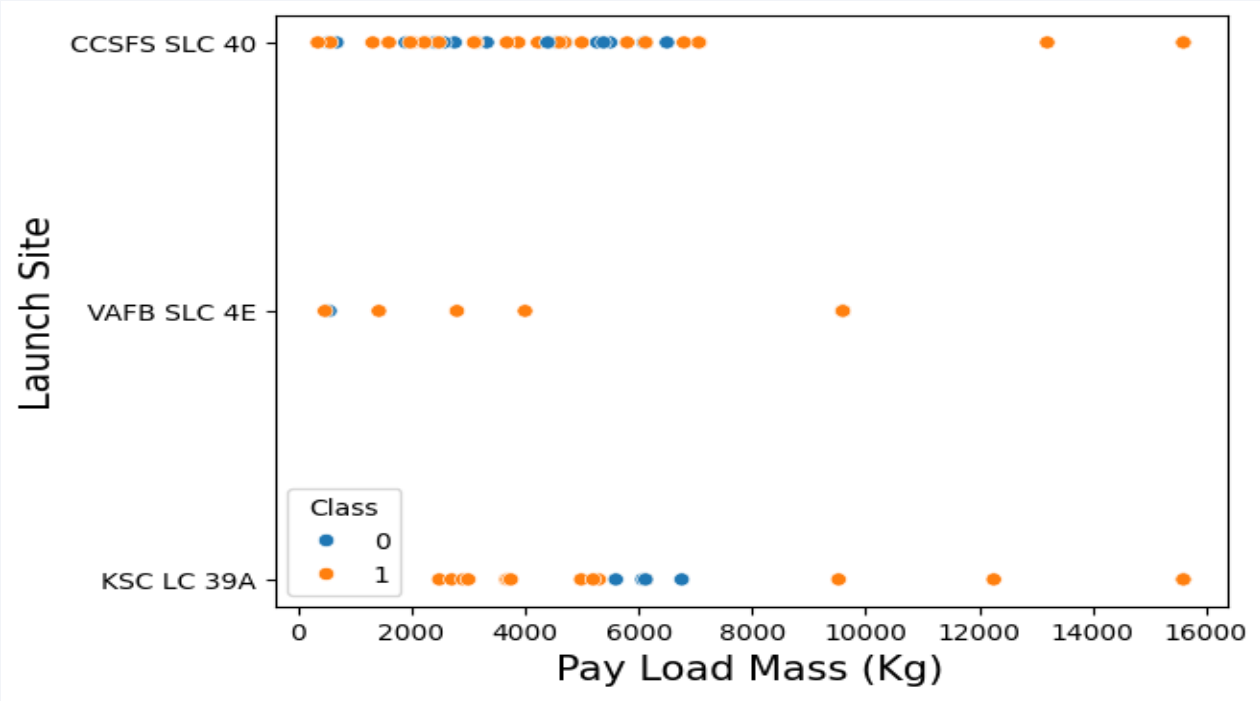

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



1. **CCAFS SLC-40 :** Most used site having most no. of launch trails. It has 55 trials, of which 33 were successful and 22 failed. Hence, Success rate is 60%.
2. **VAFB SLC 4E :** Least used site with only 13 launch trials. 10 of the trials were successful and 3 failed. Hence, Success rate is 77%.
3. **KSC LC 39A :** Moderately used site, with 22 launch trials. 17 of the trials were successful and 5 failed. Hence, Success rate is 77%

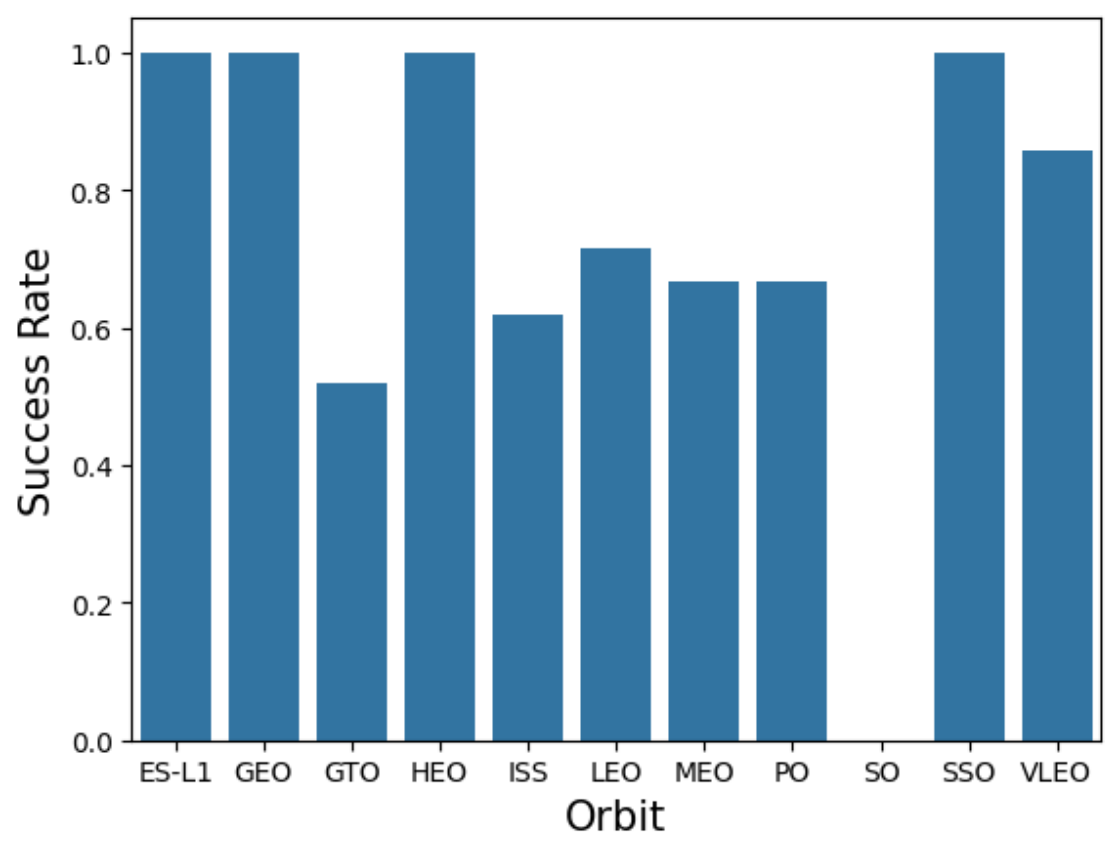
Payload vs. Launch Site



From the plot obtained, we observe that :

1. There is no strong relationship between Payload Mass and Success/Fail outcome at the Launch site.
2. For the VAFB SLC 4E launch site, no rockets were launched for heavypayload mass (greater than 10000)

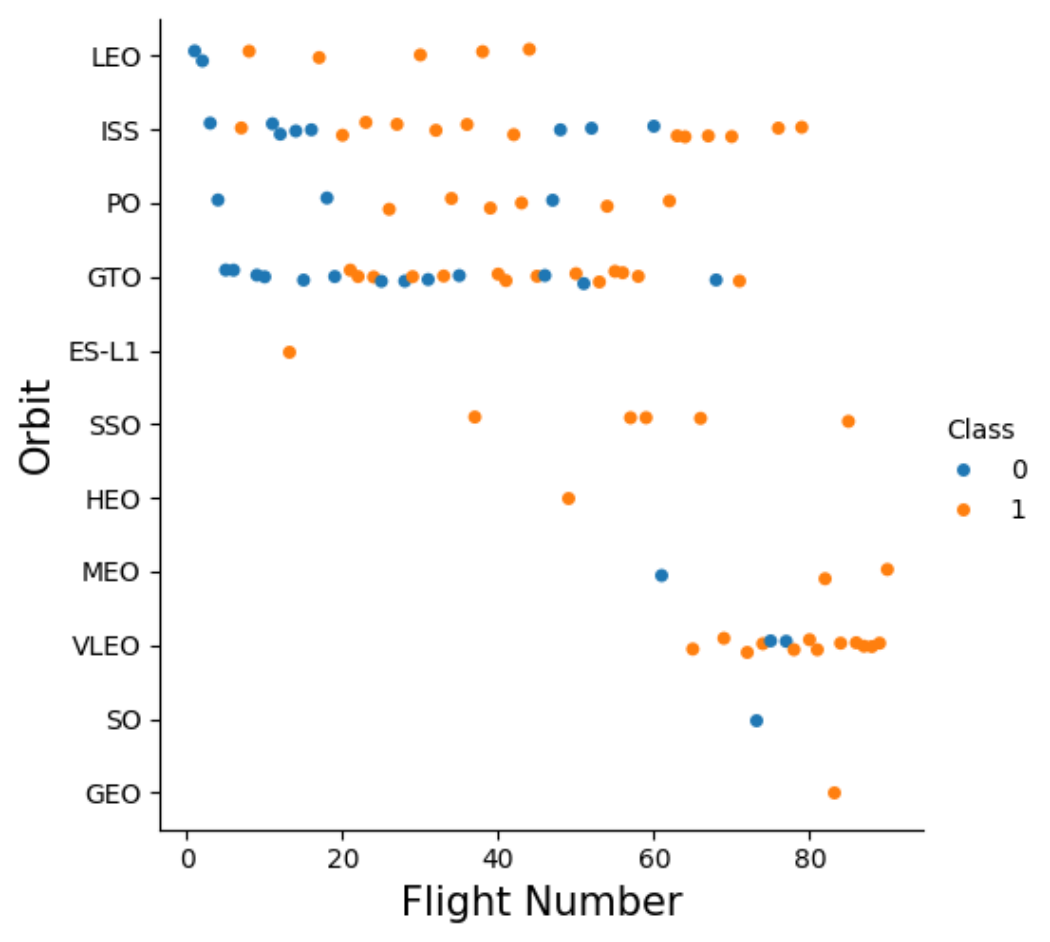
Success Rate vs. Orbit Type



From the bar plot obtained, we observe that :

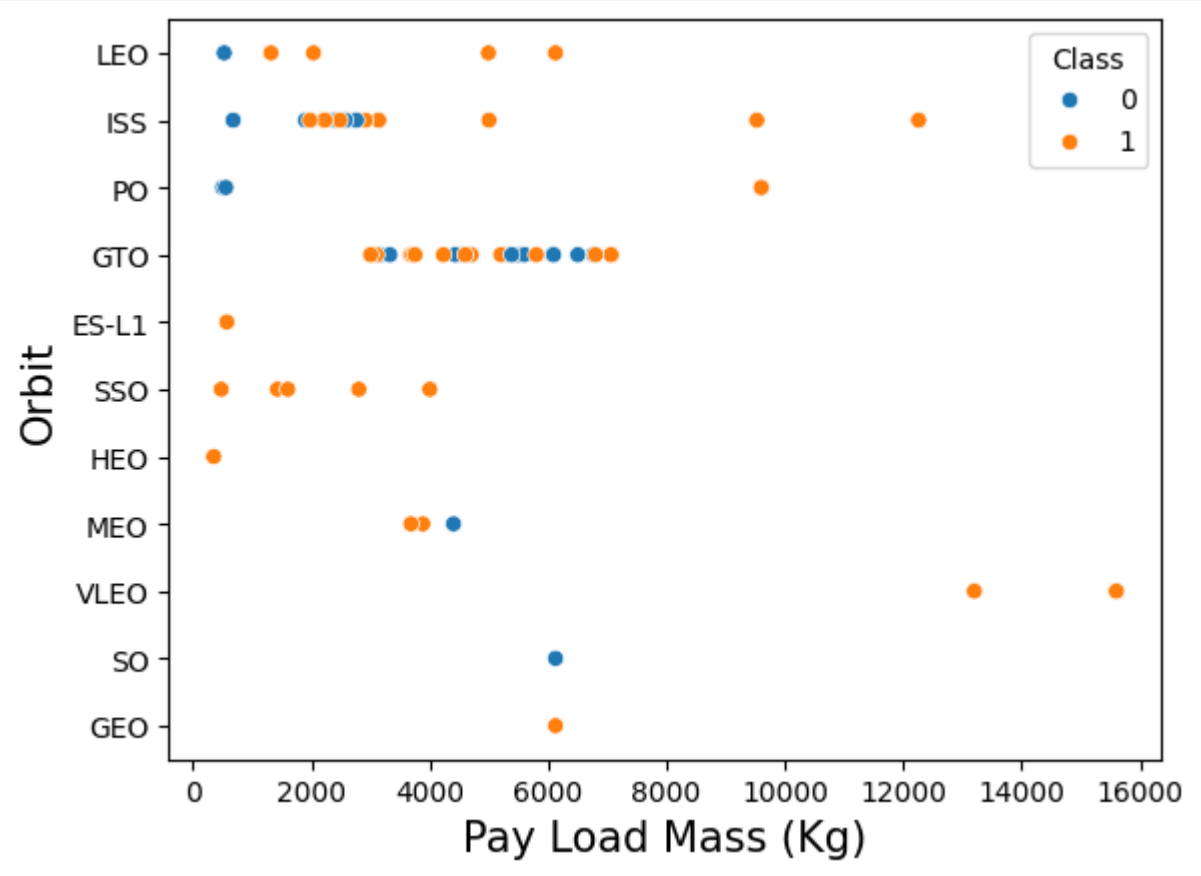
1. The orbits ES-L1, GEO, HEO, SSO are the orbits with most success (100% success rate)
2. The worst performing orbits are found to be SO and GTO, with least success rate among all orbits

Flight Number vs. Orbit Type



From the plot, we observe that in LEO orbit, Success appears to be related to the number of Flights. On the other hand, there seems to be no relationship between Flight number when in GTO orbit.

Payload vs. Orbit Type

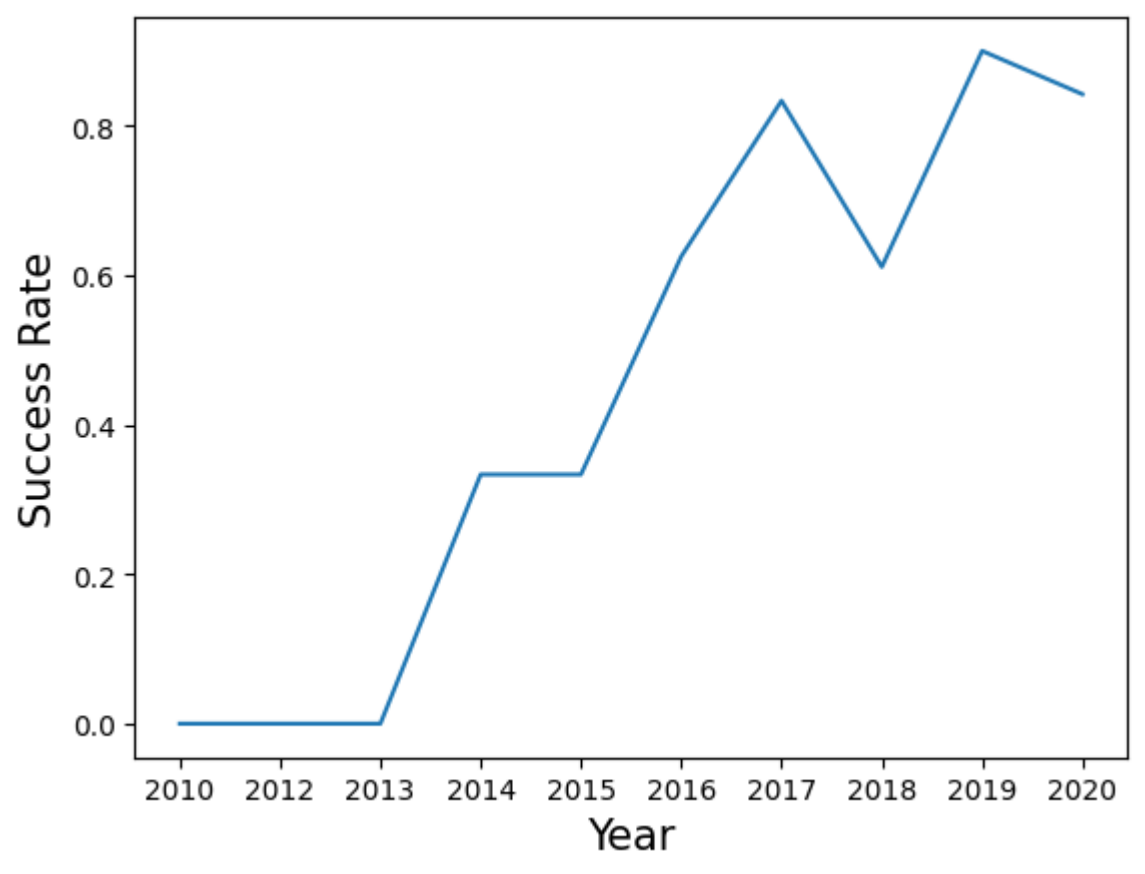


According to the plot obtained, it is observed that :

1. With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

2. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



From the plot, it is understandable that :

The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

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

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

From the results obtained, we have **4 unique launch sites** as given in the table on the right.

Launch Site Names Begin with 'CCA'

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

```
%%sql
SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5
```

Python

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

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

In the table above, 5 records of the query has been shown wit launch site names beginning with ‘CCA’. Here, the first 5 records are those from launch site ‘CCAFS LC-40’.

Total Payload Mass

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

```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") AS "TOTAL_PAYLOAD_MASS" FROM SPACEXTABLE WHERE "Customer" LIKE "NASA (CRS)"
```

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

TOTAL_PAYLOAD_MASS
45596

The Total Payload mass carried by Boosters launched by NASA (CRS) equals to **45596 Kg**

Average Payload Mass by F9 v1.1

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

```
%%sql
SELECT AVG("PAYLOAD_MASS__KG_") AS "AVG_PAYLOAD_MASS" FROM SPACEXTABLE WHERE "Booster_Version" LIKE "F9 v1.1"
```

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

AVG_PAYLOAD_MASS
2928.4

The Average Payload Mass carried by Booster version F9 v1.1 is **2928.4 Kg**.

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

```
%%sql
SELECT MIN("Date") AS "First successful landing date in ground pad" FROM SPACEXTABLE
WHERE "Landing_Outcome" LIKE "Success (ground pad)"
```

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

First successful landing date in ground pad
2015-12-22

The date when the first successful landing outcome in ground pad was achieved is : **22 – 12 - 2015**

Successful Drone Ship Landing with Payload between 4000 and 6000

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 SPACEXTABLE WHERE
"Landing_Outcome" LIKE "Success (drone ship)" AND
"PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000
```

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

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

The boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are :

1. F9 FT B1022
2. F9 FT B1026
3. F9 FT B1021.2
4. F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%%sql
SELECT COUNT("Mission_Outcome") AS "SUCCESSFUL" FROM SPACEXTABLE WHERE
"Mission_Outcome" LIKE "Success%"
```

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

SUCCESSFUL
100

```
%%sql
SELECT COUNT("Mission_Outcome") AS "FAILURE" FROM SPACEXTABLE WHERE
"Mission_Outcome" LIKE "Failure%"
```

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

FAILURE
1

It has been concluded that there are **100 successful outcomes and 1 failure**. So, the success rate of the mission outcomes is quite dominant.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
SELECT "Booster_Version" FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_"= (
    SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE
)
```

```
 * sqlite:///database.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

The booster versions which have carried the maximum payload mass are shown in the table on the left.

2015 Launch Records

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.

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", "Booster_Version", "Launch_Site", "Landing_Outcome" FROM SPACEXTABLE
WHERE "Landing_Outcome" LIKE "Failure (drone ship)" AND substr("Date", 0, 5) = '2015'
```

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

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

We have 2 failed outcomes in 2015 in drone ship, both on the same launch site ‘CCAFS LC-40’ and with Booster versions F9 v1.1 B1012 and F9 v1.1 B1015.

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 "Landing_Outcome", COUNT("Landing_Outcome") AS "COUNT" FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_OutCome" ORDER BY "COUNT" DESC
```

Python

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

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

The Landing outcomes have been ranked based on their count as shown in the table above.

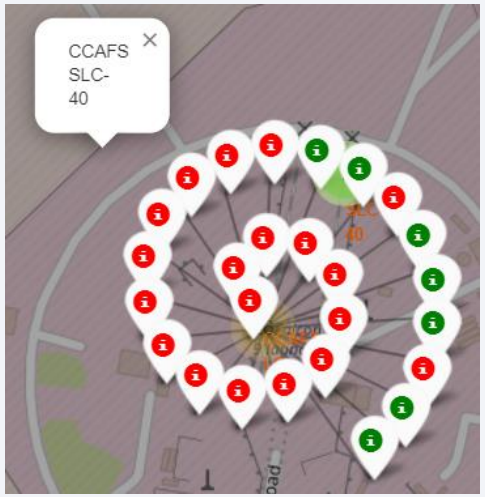
Section 3

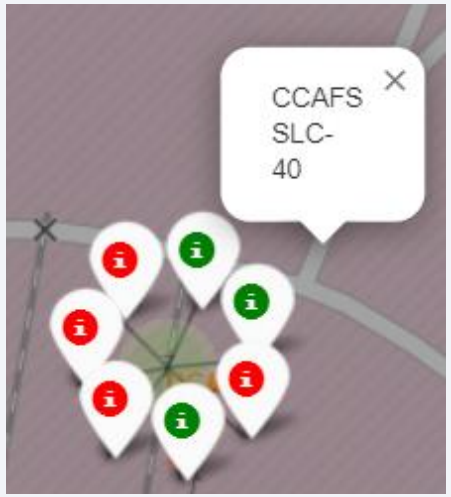
Launch Sites Proximities Analysis

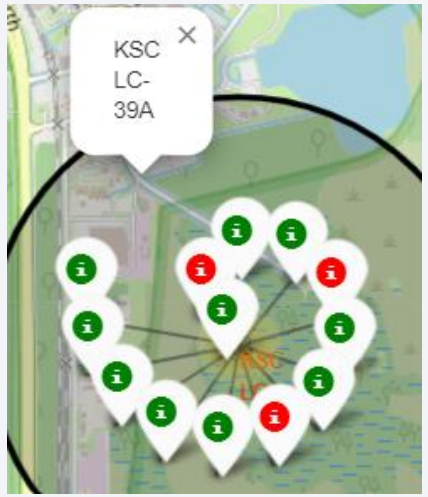
Folium Map: Launch Sites

All launch sites are near the Equator line and are close to the coast as well. The launch sites are in two states : California and Florida

Folium Map: Success/Failed launches for each Launch site





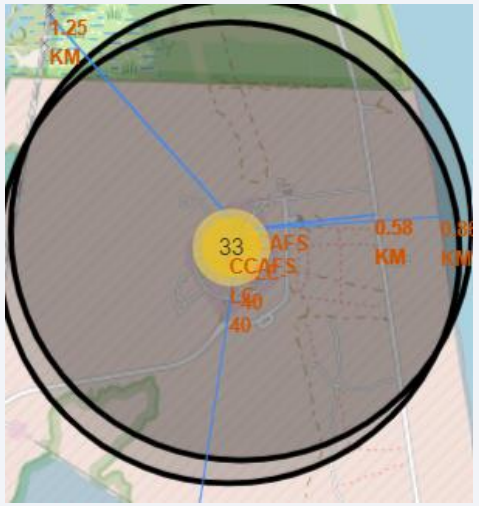


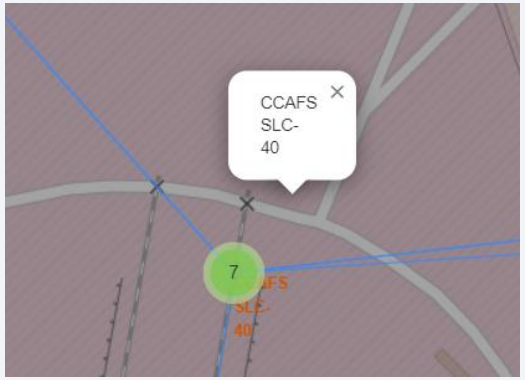
From the colour-labelled markers in Marker Clusters, we can identify success/failed launches at different launch sites.

Green Marker = Successful

Red Marker = Failure

Folium Map : Closest Proximities to CCAFS SLC-40





We have calculated the distance of CCAFS SLC-40 launch site from its closest coastline, city, railway and highway and have also marked the distance via Polylines.

Proximities Coordinates :

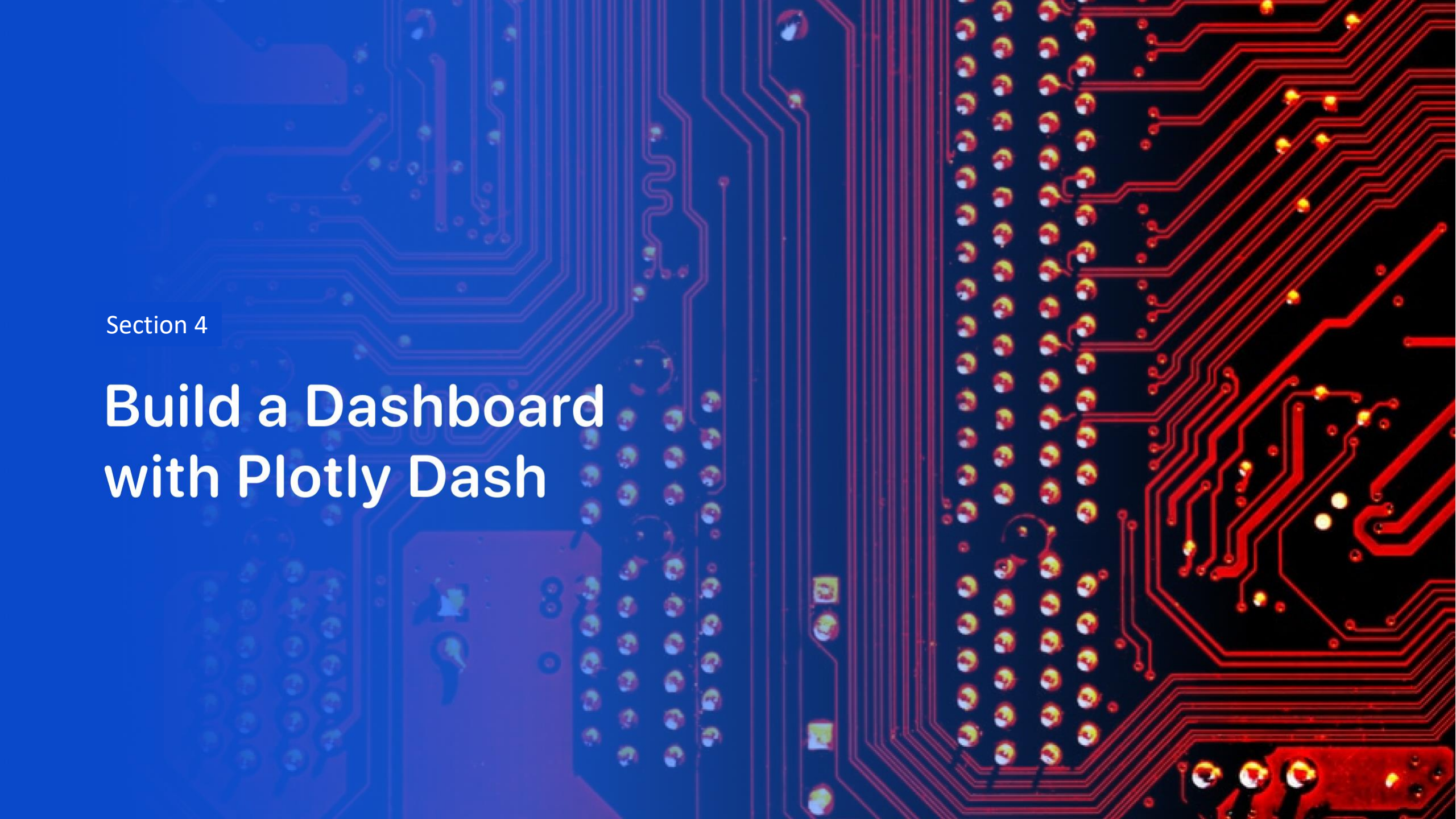
Location	Latitude	Longitude
Coast	28.5637	-80.56802
City	28.39683	-80.60565
Railway	28.57156	-80.58535
Highway	28.56379	-80.57087

Nearest Coastline distance = 0.86 km

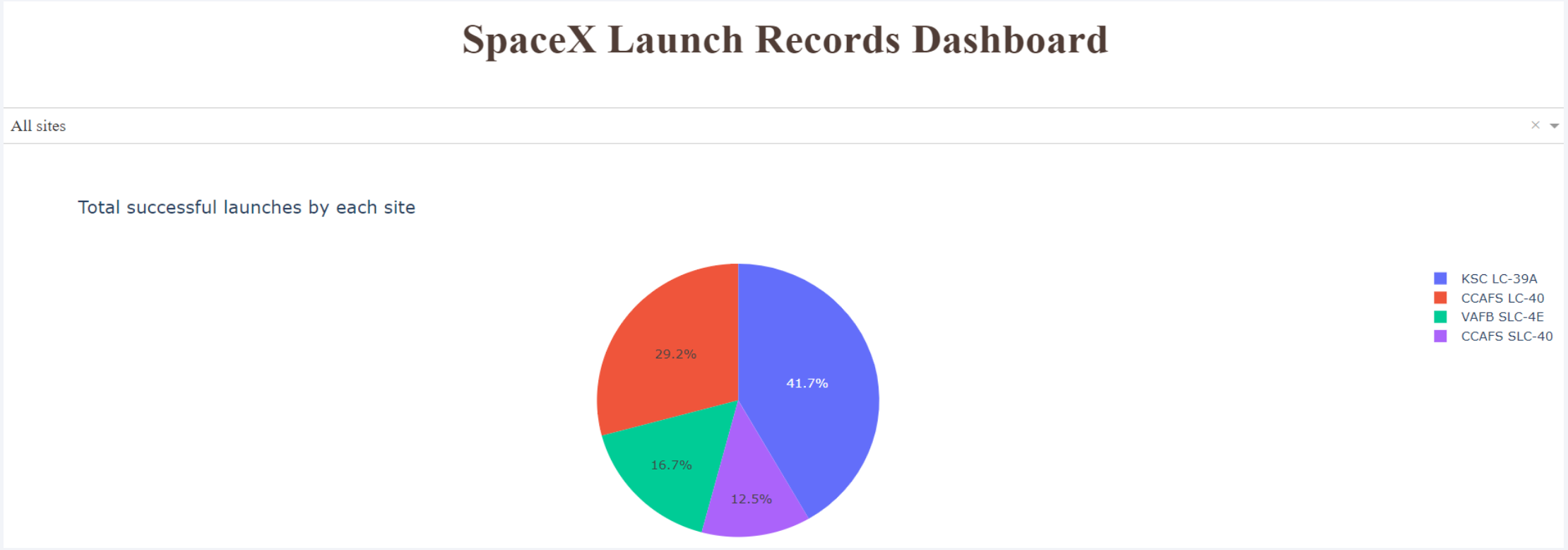
Nearest city distance (Cape Canaveral) = 18.72 km

Nearest railway distance = 1.25 km

Nearest highway distance = 0.58 km


Section 4
Build a Dashboard with Plotly Dash

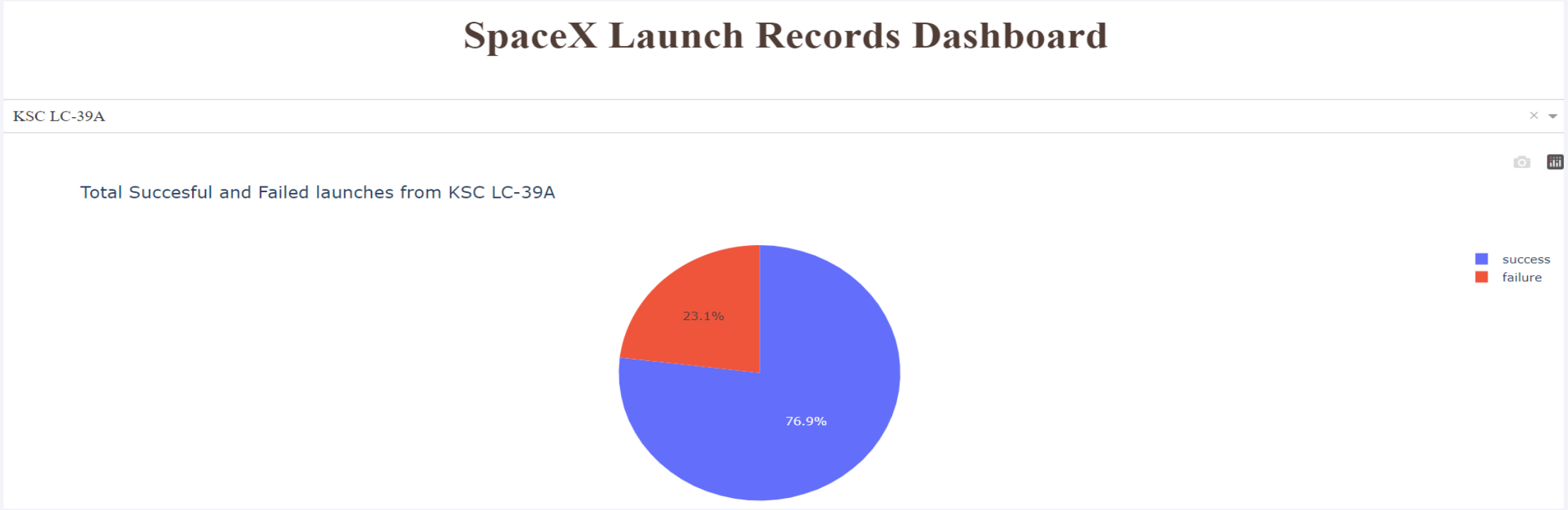
Dashboard : Launch success count for all sites



The pie chart shows the success percentage for each site :

- The best launch site is KSC LC-39A with 41.7% of total successful launches among all sites.
- The site with the least successful launches is CCAFS SLC-40 with only 12.5%

Dashboard : Launch site for KSC LC-39A



Launch success of KSC LC-39A, the site with highest success rate :

- 76.9% of launches were successful from KSC LC-39A
- Only 23.1% of launches failed from KSC LC-39A

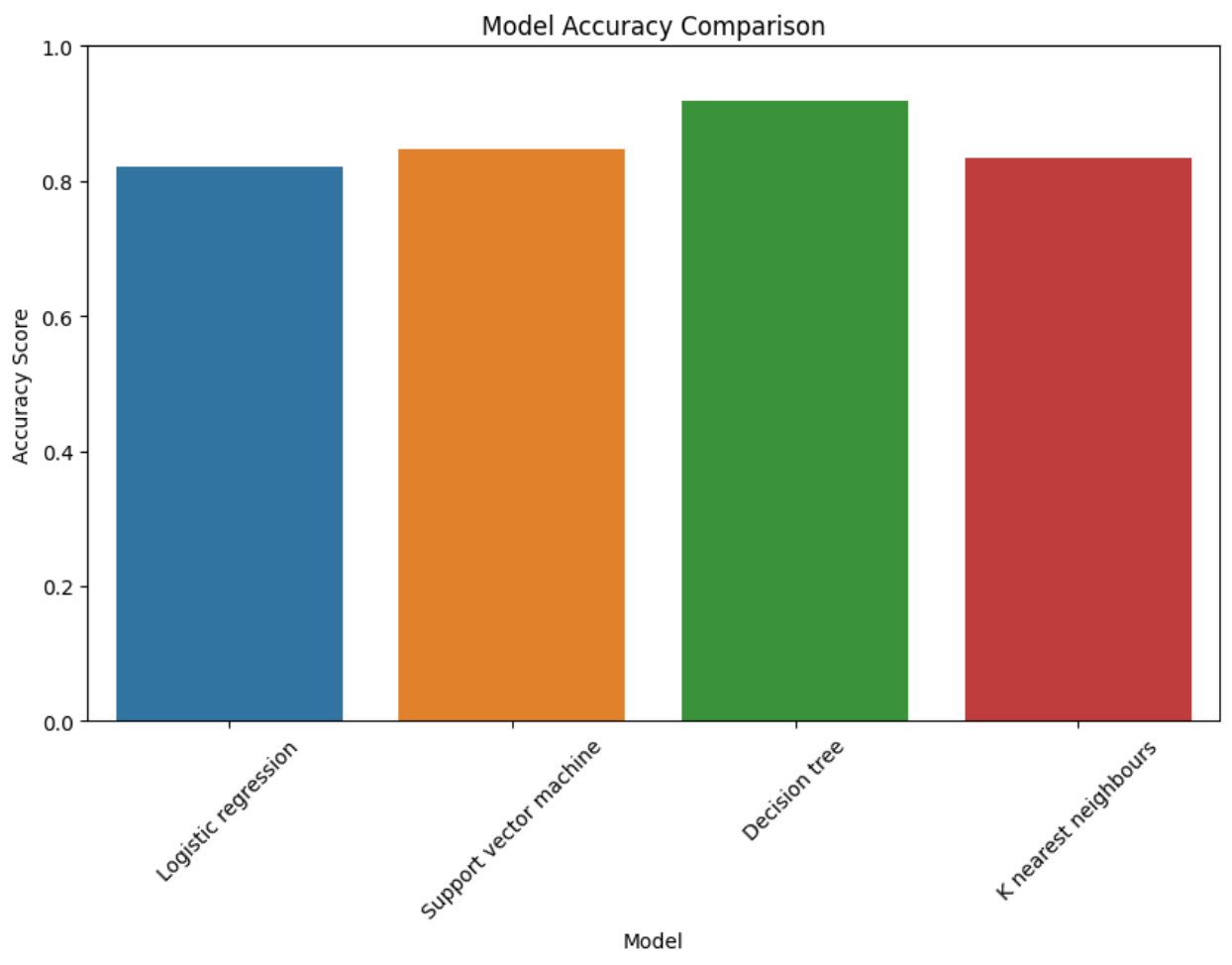
Dashboard : Payload vs. Launch Outcome for all sites

The scatter plot shows the relationship of the Launch Outcome with the Payload mass and by varying the payload range slider, we can analyse how the payload mass affects the outcome of the mission.

Section 5

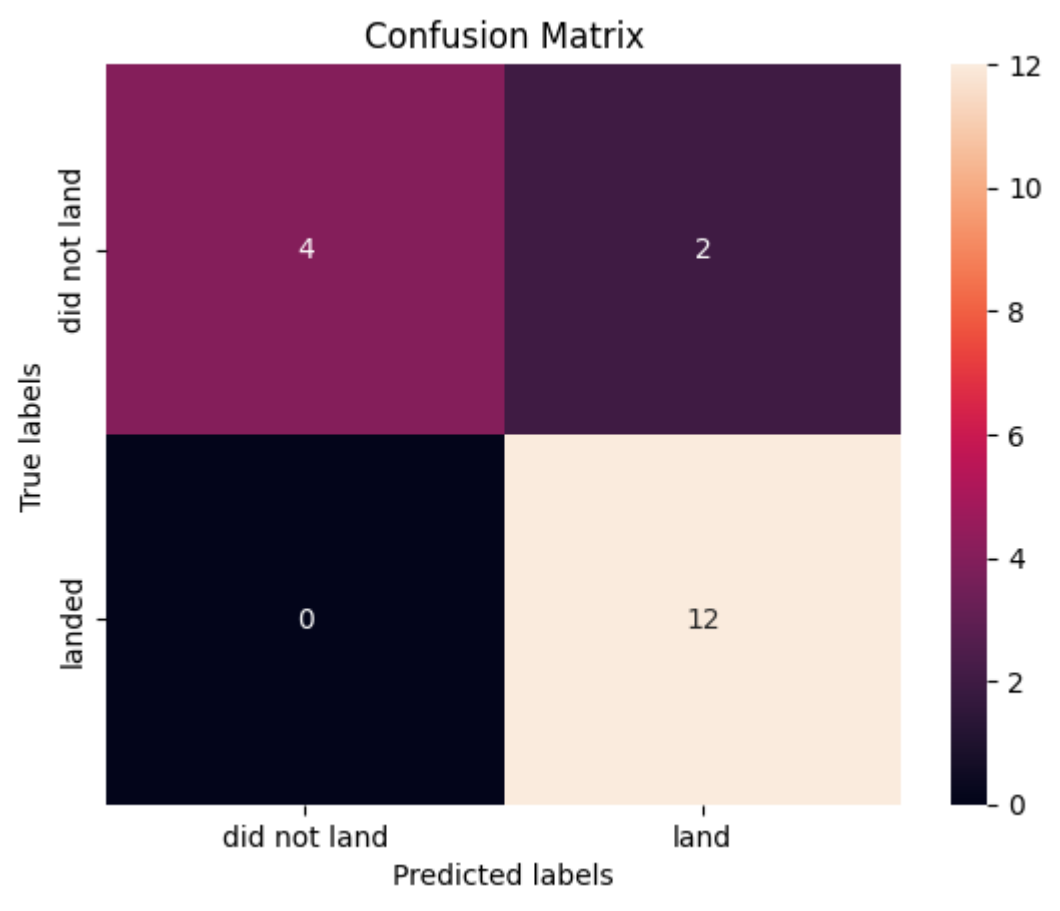
Predictive Analysis (Classification)

Classification Accuracy



The Decision Tree has performed the best of all the models namely Logistic Regression, Support Vector Machine, K Nearest Neighbours which have a similar performance.

Confusion Matrix



The Decision Tree performs the best among all the models with Test Accuracy : 0.91786

The results as shown in the confusion matrix of the Decision Tree is :

- True positive = 12
- False Positive = 2
- True Negative = 4
- False Negative = 0

Conclusions

- A successful outcome in the first stage leads to huge savings in terms of launch cost for the company
- Various attributes contribute to the outcome of a successful first stage return. In the model developed, 80 attributes were taken into consideration as features
- The launch sites are found to be close to the coast, railway and highway but are relatively much far from cities. The proximity to coast, railway and highway helps a lot in cutting transportation cost
- The success rates have increased over the years, and the KSC LC-39A is the launch site with highest success rate
- The orbits ES-L1, GEO, HEO, SSO are the orbits with most success.

Appendix

- Data collected from API : dataset_1.csv
- Data collected by Web scraping : data_webscrapped.csv
- Data after Data Wrangling : dataset_2.csv
- Dataset used for performing SQL queries : data_sql.csv
- Data after EDA and Feature Engineering : dataset_3.csv
- Data used for Visualization by Folium : data_location_folium.csv
- Data used for preparing Dashboard : data_dash.csv
- Dashboard plots screenshots : dashboard plots

Thank you!