IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Grace Aliko

7th July, 2024

Table Of Contents

Executive Summary

In this capstone project, we aim to predict the successful landing of the Falcon 9 first stage. We collect data using the SpaceXdata API and web scraping from Wikipedia. After gathering the data, we clean it by renaming columns, replacing missing values, classifying some columns, and conducting basic exploratory data analysis (EDA).

Following the initial data cleaning, we perform EDA using SQL and further analyze the data with visualization techniques, creating various plots and preparing the data for modeling. We also use the Folium library for visual analytics to observe physical and environmental factors, and create a dashboard to display key results.

In the final stage, we model the preprocessed and cleaned data using various classification models with hyperparameter tuning. Our analysis revealed that the success rate increased with the flight number. Additionally, we observed that launch sites are located near highways and coasts. Among the models tested, the Decision Tree Classifier provided the highest accuracy on the test data.

GitHub Link for the project repo: SpaceX-Falcon9-Landing-Prediction-Capstone-Project

Introduction

- **Project background and context:**
- SpaceX advertises Falcon 9 rocket launches on its website at a cost of 62 million dollars, whereas other providers charge upwards of 165 million dollars per launch. A significant portion of SpaceX's savings comes from the ability to reuse the first stage of the rocket. Therefore, if we can predict whether the first stage will land successfully, we can estimate the cost of a launch. This information would be valuable for an alternative company looking to compete with SpaceX for a rocket launch contract. The goal of this project is to train a machine learning model to predict the successful landing of the first stage, thereby estimating the total mission cost.
- **Questions to which answers are to be found:**
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a

successful landing?

- What operating conditions needs to be in place to ensure a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - One of the dataset used was extracted by using SpaceXData API
  - The other dataset was webscraped using BeautifulSoup from wikipedia
- Perform data wrangling
  - Performed basic EDA
  - Determined training labels
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Visualized and analyzed data using scatterplots
- Perform interactive visual analytics using Folium and Plotly Dash
  - Built a dashboard to view pie-chart and scatter plots according to each site
- Perform predictive analysis using classification models
  - Trained classification models and determined the best model, and also plotted confusion matrix for each model.

Data Collection

The objective was to collect spacex launch data so that we can clean, analyse, visualize and model data to predict if a launch is successful or not.

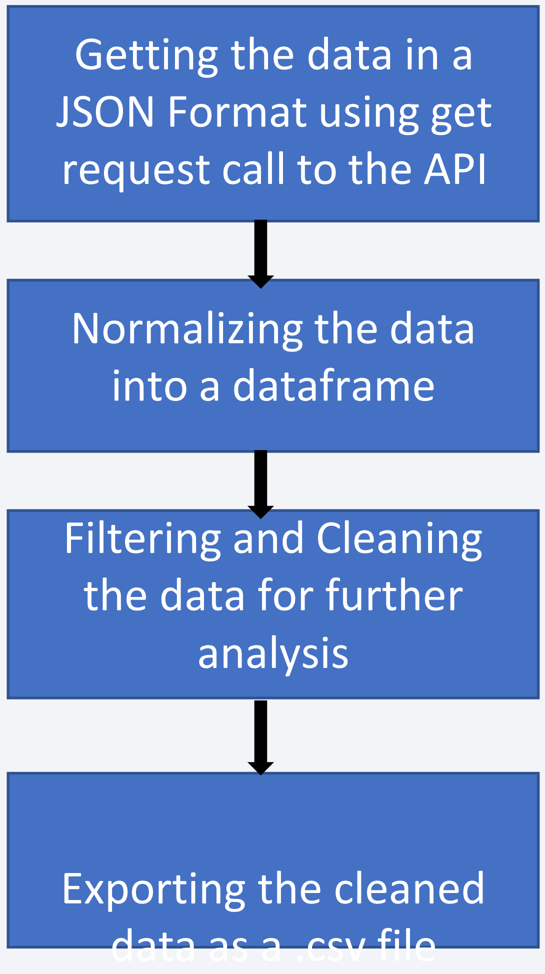
- One of the data set was collected by using Space X API and then the json response was normalized to a dataframe and then further preprocessing and analysis was done
- Another dataset containing Launch dates etc was webscraped from Falcon9 Wikipedia page using using BeautifulSoup package.
- Relevant data was extracted from tables and put into a dataframe for further analysis

Data Collection – SpaceX API

The first dataset is collected using the **SpaceX API, using get requests in Python**. Then we normalize the json contents into a dataframe and then using functions and pandas we extract relevant information, clean the data, and export the cleaned data.

The basic flowchart of the process is shown:

GitHub URL of the Notebook: [Data Collection Using API]

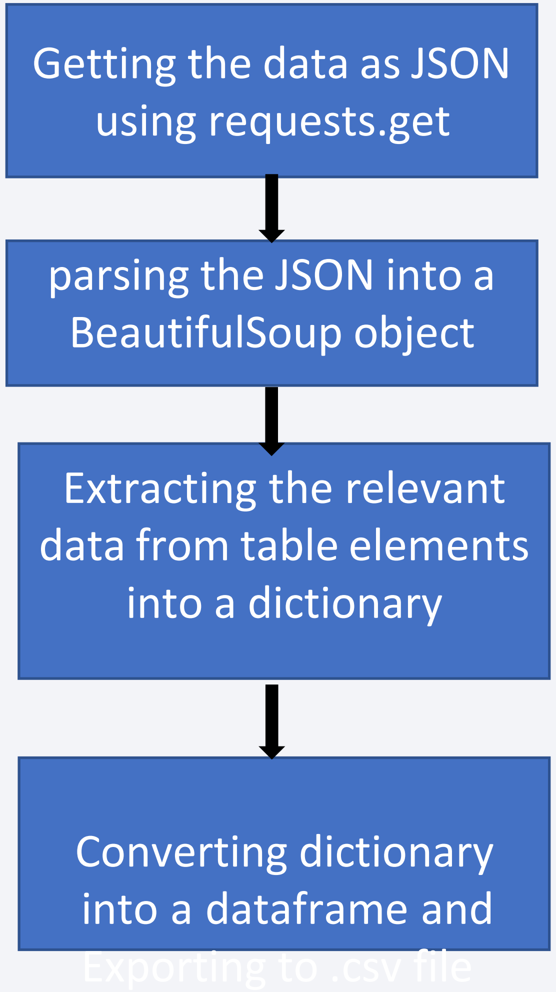


Data Collection - Scraping

Performed Web Scraping to collect Falcon 9 historical launch records from a Wikipedia page titled ‘List of Falcon 9 and Falcon Heavy launches’

We requested the page using response.get() method and then parsed the data as a BeautifulSoup object. Then we extracted all column/variable names from the HTML table header, and converted it into a datframe.

GitHub URL of the Notebook: [Data Collection with WebScraping]



Data Wrangling

Exploratory Data Analysis (EDA) was performed on the dataset before wrangling.

These tasks were performed:

- Calculated the number of launches at each site.
- Calculated the number and occurrence of each orbits on which each launch was aimed to.
- Calculated the number and occurence of mission outcome per orbit type

Finally, the landing outcome label was created from Outcome column, class 0 denoted failure and 1 denoted successful landing.

The cleaned data was exported to a .csv file.

GitHub URL of the notebook: Data Wrangling and EDA

EDA with Data Visualization

Visualized the relationship between Flight Number and Launch Site, Payload and Launch Site, FlightNumber and Orbit type, Payload and Orbit type using scatterplots as we need to observe the trend and overall relationship between the variables.

Visualized the success rate of each orbit type using barchart so that side by side comparison could be done

Visualizing the launch success yearly trend using a line plot as it best depicts a trend, whether it is increasing or not.

Applied One hot Encoding to the features to prepare for data modelling.

GitHub URL: EDA using Data Visualization

EDA with SQL

We loaded the dataset in IBM Db2 and connected to the database via ibm db2 API.

We performed EDA with SQL by executing queries to get insights from the data.

- unique launch sites
- launch sites beginning with CCA
- total payload mass carried by boosters launched by NASA (CRS)
- average payload mass carried by booster version F9 v1.1
- date when the first successful landing outcome in ground pad was acheived.
- The total no of successful and failed outcomes
- names of the booster_versions which have carried the maximum payload mass
- Rank of the count of landing outcomes between a specified date.

GitHub URL of the notebook: [EDA using SQL]

Build an Interactive Map with Folium

Folium Circles, Markers, Marker clusters, Mouseposition and polyline objects were added.

- **Circles:** To add a highlighted circle area with a text label on a specific site.
- **Markers:** To mark the site
- **Marker Clusters:** To simplify the map as it contained any markers having the same coordinate
- **Mouse Position:** to get the coordinates for the position the mouse points on the map.
- **Polyline:** to draw a line from a site to the nearest coast, city, highway, etc.

GitHub URL of the notebook: Visual Anlaytics using Folium

Build a Dashboard with Plotly Dash

Built an interactive Dashboard using Plotly Dash having the following features:

Added a dropdown for the site input, and plotted a pie chart to view the relative success and failures in launches for each site

Added a Range Slider for the Payload range to view the scatterplot between Payload and Class, according to the input site and within the provided payload range.

These plots and interactions allowed to quickly visualize and analyze the relation between payloads and launch sites, which helped to find the best site for launching Falcon 9

GitHub URL of python file: SpaceX_Dash_Application

Predictive Analysis (Classification)

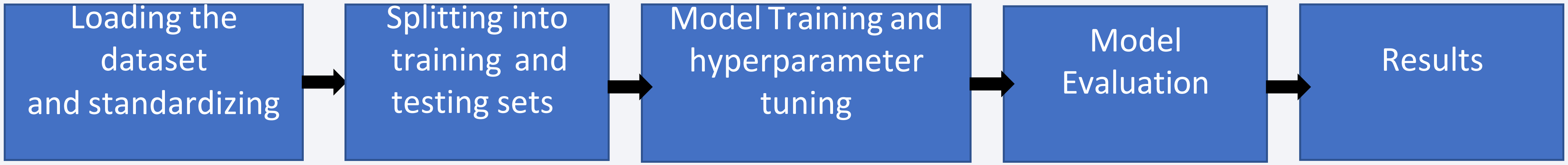
The features/predictor variables and the label/target variable were loaded into respective dataframes. The predictordata was standardized using standar scalar.

The data was split into training and test data. The labels are the data that we want to predict. The labels are in the column labeled "class". The features are all the other columns.

For the problem statement, classification models such as logistic regression, SVM, Decision Trees, and KNN were used. We determined the best hyperparameters using GridSearchCV with 10 fold cross validation We also plotted the confusion matrices for each model.

We will use the test data to determine which machine learning model performs the best using confusion matrices and scores, then compile the results.

GitHub Notebook Link: [Predictive Analysis using ML]



Results

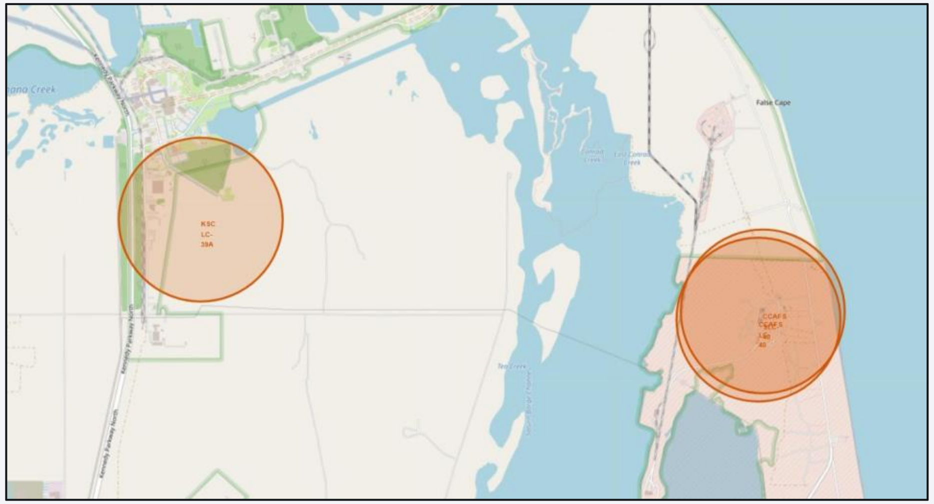
Exploratory data analysis results

- Newer rockets could carry more payload
- Payloads over 8000kg have high success rate
- the sucess rate since 2013 kept increasing till 2020
- 2010-2013 period had no success rate
- Space X uses 4 different launch sites;
- VLEO orbit has 14 launches and 85% success rate
- The first successful landing outcome was in 2015, five years after the first launch.
- With booster F9, almost every mission outcome was successful.
- around 70 landing outcomes were successful, while there were 22 no attempts, and around 10 failed.
- With time, the success increased mostly due to advancement in technology.

Results (Contd..)

Interactive analytics results:

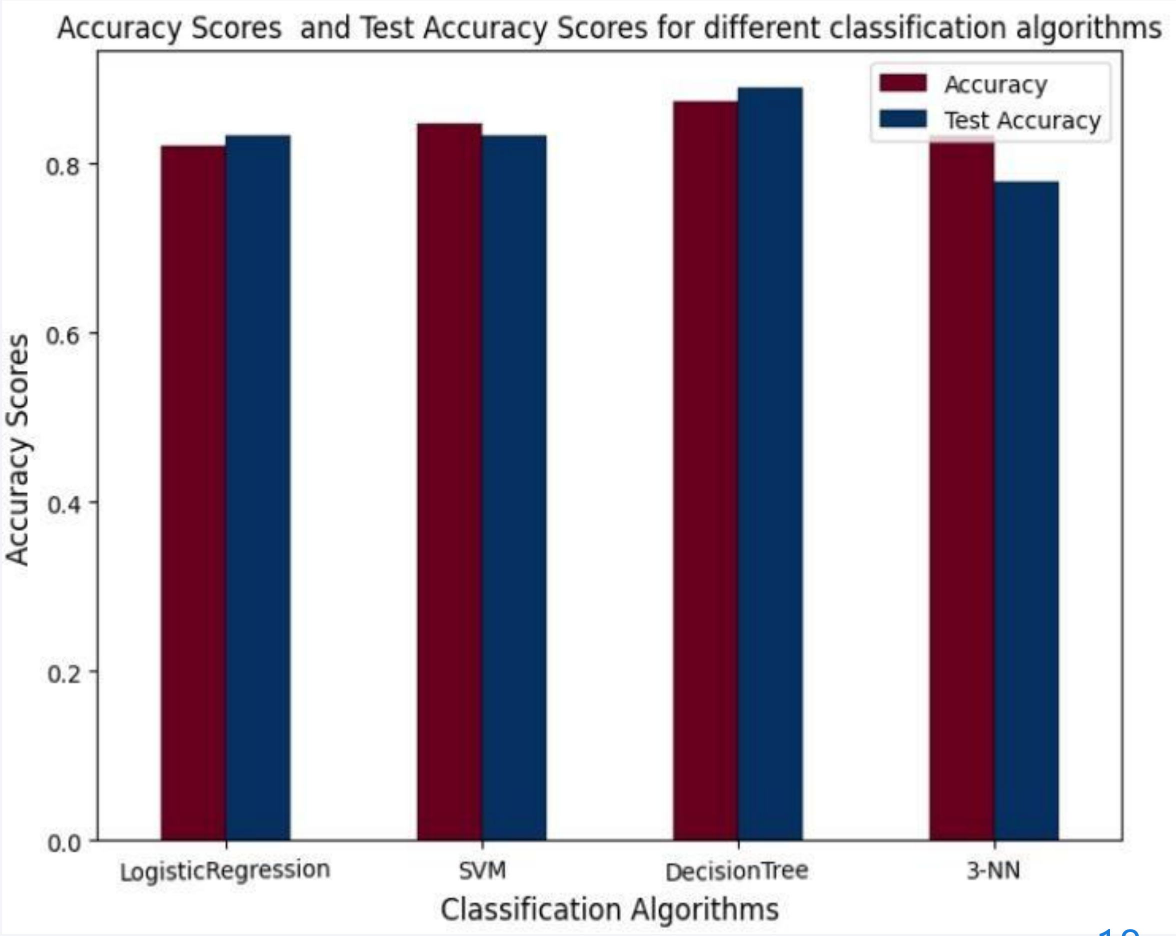
- launch sites are close to the equator
- Most launches happens at east cost launch sites
- launch sites are in close proximity to railways
- Launch sites are in close proximity to highways
- Launch sites in close proximity to coastline
- Launch sites keep certain distance away from cities



Results (Contd..)

Predictive analysis results:

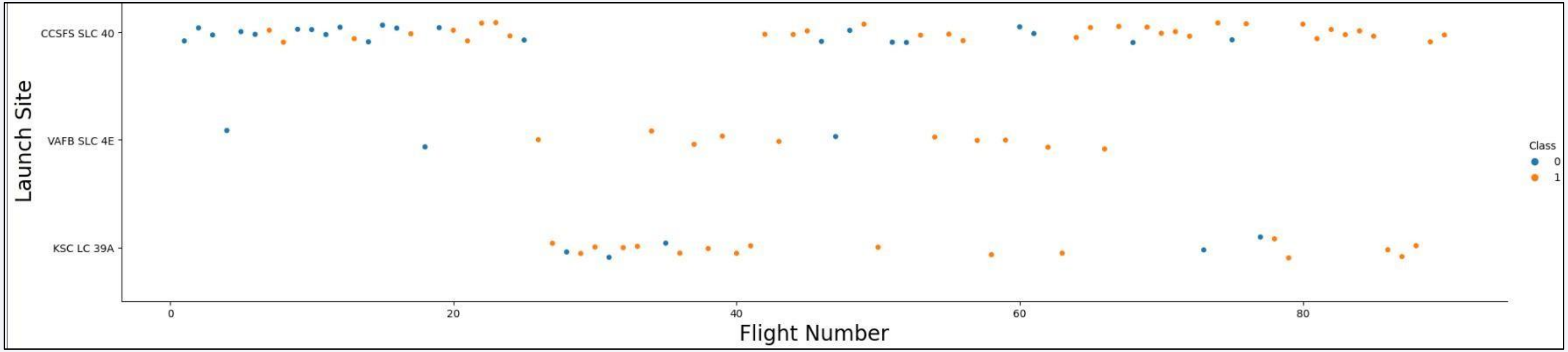
- All models, except KNN had almost the same accuracy for the test data
- The best model is the Decision tree Classifier, having approximately 87.5% accuracy on training data and 88.9% test accuracy
- The worst model is the KNN, having approximately 80% mean accuracy



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

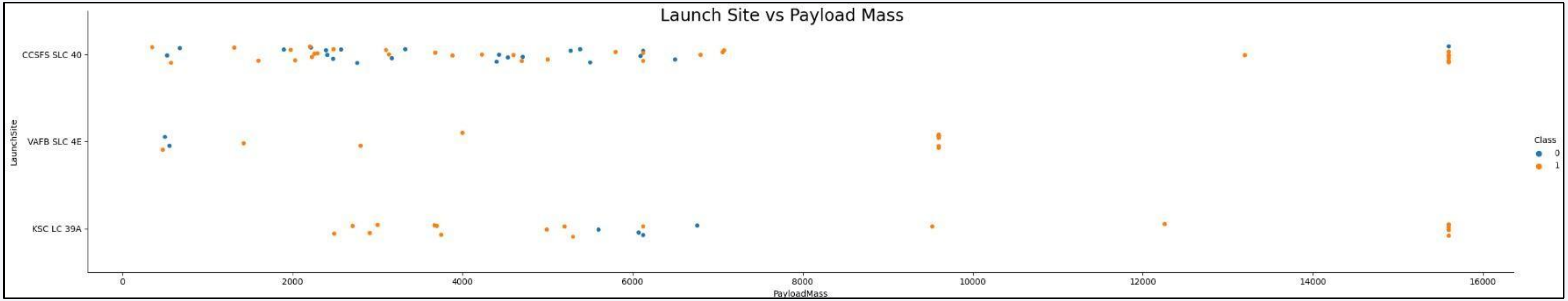


Scatter plot of Flight Number vs. Launch Site

Observations:

- CCAFS LC-40 has overall lower success rate than other two as it failed a lot during initial flights
- KSC LC-39A and VAFB SLC 4E have almost same success rate, and they have a relatively higher flight number so failure rate is low.
- Best Site is CCAFS SLC-40 as it has very high success rate in recent times.

Payload vs. Launch Site

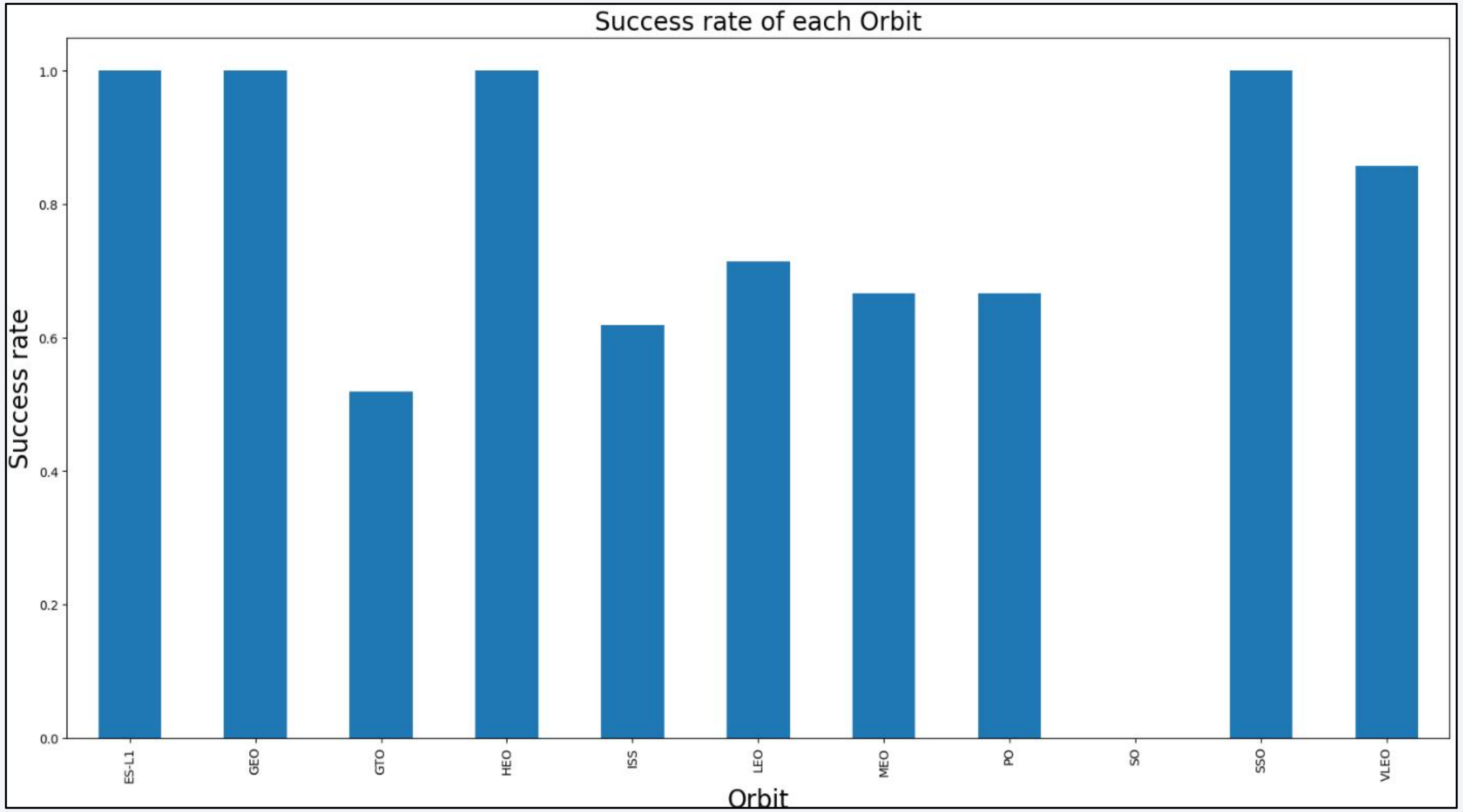


Observations:

- There are no rockets launched for heavy payload mass (greater than 10000) for the VAFB-SLC launchsiteA
- Payloads over 8000kg have high success rate
- Payloads less than 6000kg have high failure rate for the CCAFS SLC-40 launch site

Success Rate vs. Orbit Type

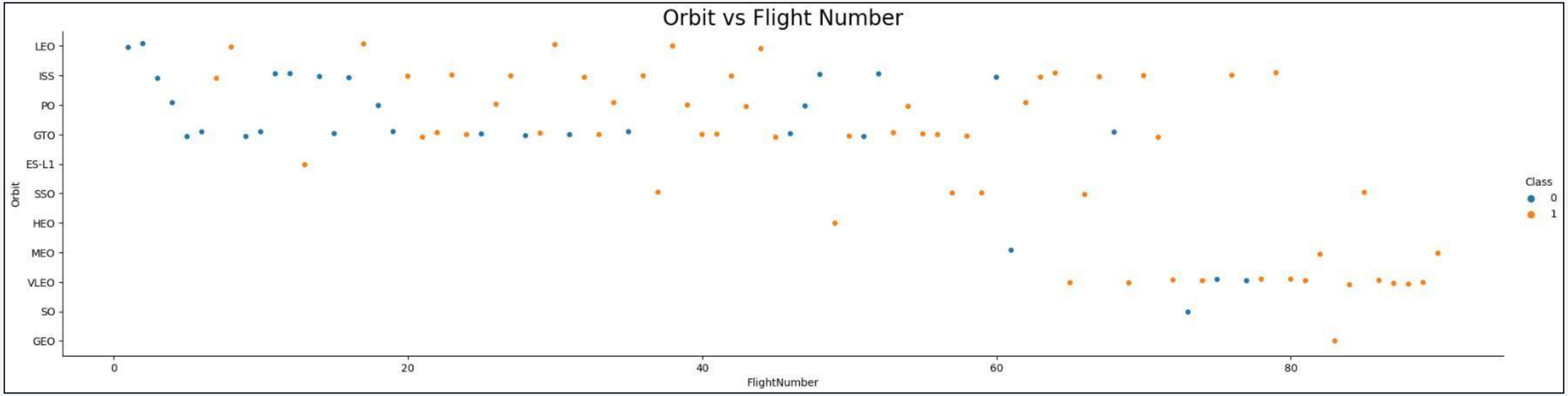
```
Orbit
ES-L1      1
GEO        1
GTO       27
HEO        1
ISS       21
LEO        7
MEO        3
PO         9
SO         1
SSO        5
VLEO      14
Name: Orbit, dtype: int64
```



Observations from data and bar chart:

- GEO, HEO, ES-L1, SSO have 1 launches and 100% success rate
- SO has 1 launch and 0% succes rate
- ISS has 21 launches 61% success rate
- VLEO has 14 launches and 85% success rate
- PO has 9 launches and ~65% success rate

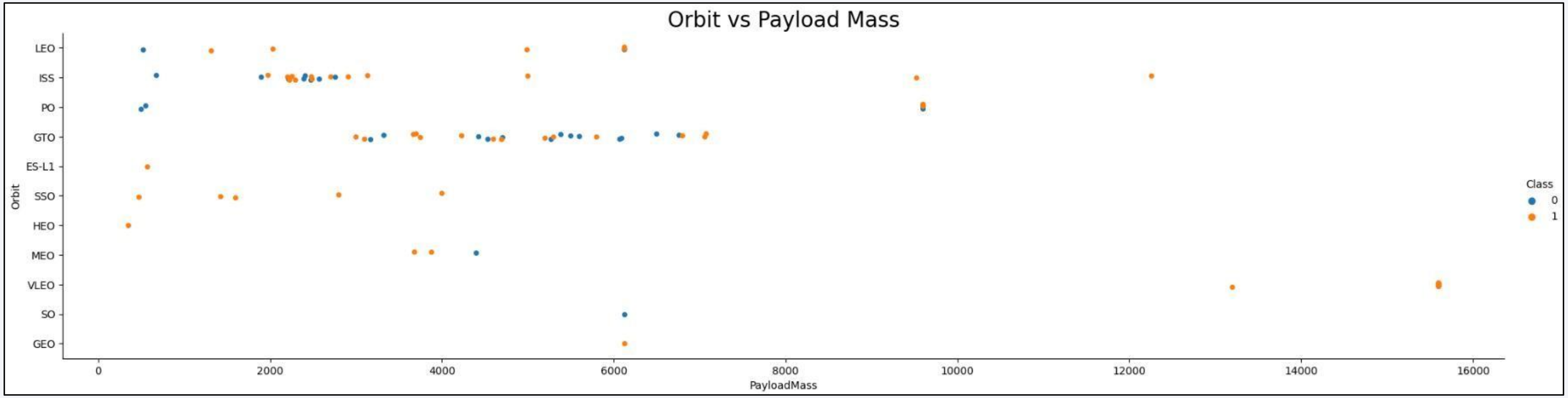
Flight Number vs. Orbit Type



Observations:

- We can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- Success rate is low for the first flight and it increases as the flight number increases

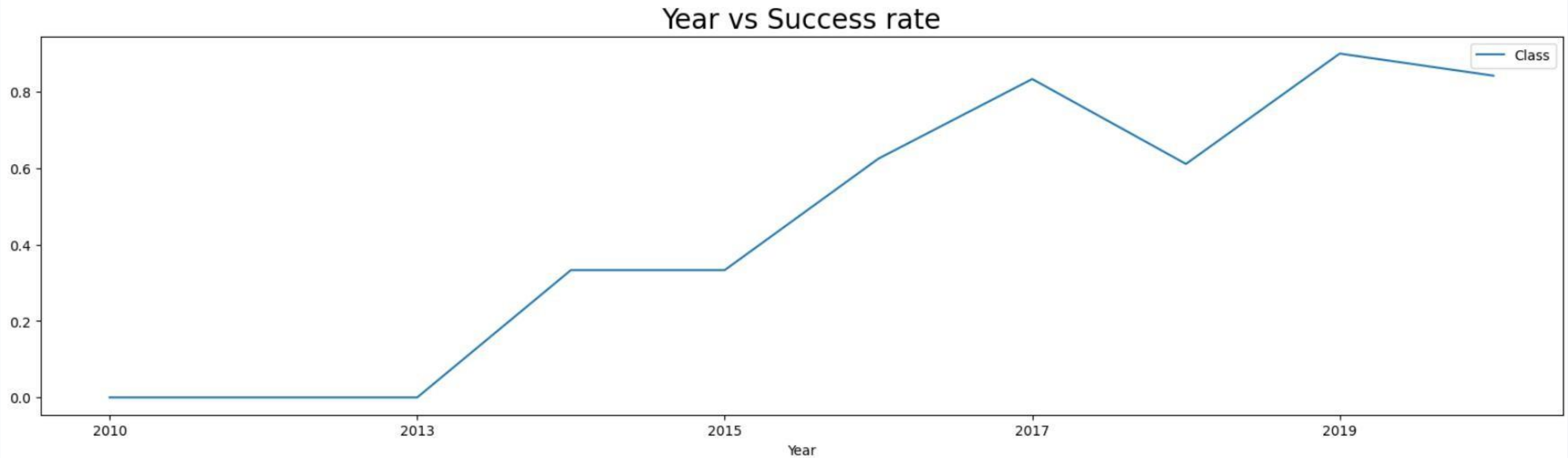
Payload vs. Orbit Type



Observations:

- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS orbits.
- 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



Observations:
- The sucess rate since 2013 kept increasing till 2020
- 2010-2013 period had no success rate

All Launch Site Names

These are the site names queried from the database using "Distinct" Keyword

```
%sql select distinct launch_site from spacex
✓ 0.8s

 * ibm_db_sa://hnr90643:***@21fecfd8-47b7-4937-840d-d79
Done.
```

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

Launch Site Names Beginning with 'CCA'

```
%sql select * from spacex where launch_site like 'CCA%' limit 5
```

✓ 0.4s Python

```
* ibm_db_sa://hnr90643:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
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

Used the like operator and limit keyword to display only 5 records, of site names beginning with CCA

Total Payload Mass by NASA (CRS)

```
%sql select sum(payload_mass_kg_) as SUM from spacex where customer like 'NASA (CRS)'
✓ 0.5s

 * ibm_db_sa://hnr90643:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databa
Done.
```

SUM
45596

The total payload carried by boosters from NASA (CRS) using the like operation. can also use '=' operator.

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass_kg_) as AVG from spacex where booster_version = 'F9 v1.1'
✓ 0.6s

 * ibm_db_sa://hnr90643:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.a
Done.
```

AVG
2928

The average payload mass carried by booster version F9 v1.1 is found out using the '=' operator and avg() aggregate function as shown above.

First Successful Ground Landing Date

```
%sql select min(date) as DATE from spacex where landing_outcome like '%ground pad%'
```

✓ 0.5s

```
* ibm_db_sa://hnr90643:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.a
Done.
```

DATE
2015-12-22

The date of the first successful landing outcome on ground pad is found out using the min() aggregate function and like operator

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
select booster_version as name from spacex
where landing_outcome like '%drone%' and payload_mass_kg_ > 4000 and payload_mass_kg_ < 6000
```

✓ 0.8s

```
* ibm_db_sa://hnr90643:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.app
Done.
```

name
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

We used the where clause to filter for boosters which have successfully landed on drone ship and applied the and condition to find the query result.

Total Number of Successful and Failed Mission Outcomes

used count() aggregate function and where clause to find total numer of successful and failed outcomes

```
%%sql
select count(*) as successful_launches from spacex where mission_outcome like '%Success%';
```

✓ 0.5s

```
* ibm_db_sa://hnr90643:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.ap
Done.
```

successful_launches
100

```
%sql select count(*) as failed_launches from spacex where mission_outcome like '%Failure%'
```

✓ 0.5s

```
* ibm_db_sa://hnr90643:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.ap
Done.
```

failed_launches
1

Boosters Carried Maximum Payload

Used a subquery to calculate the max payload and then used = operator to filter the queries

```
%%sql
select booster_version as MAX_PAYLOAD_BOOSTERS from
        spacex where payload_mass_kg_ =
        (select max(payload_mass_kg_) from spacex)
```

✓ 0.6s

```
 * ibm_db_sa://hnr90643:***@21fecfd8-47b7-4937-840d-d791d0218
Done.
```

max_payload_boosters
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 booster_version, launch_site, landing_outcome from spacex where Year(date) = 2015 and landing_outcome like '%Failure%drone%'
```

✓ 0.5s

```
* ibm_db_sa://hnr90643:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/bludb
Done.
```

booster_version	launch_site	landing_outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 is found out using the ‘Year()’ function, ‘and’ condition and ‘like’ operator

Rank Landing Outcomes Between dates

The query to 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 is shown beside:

```
%%sql
select landing_outcome, count(landing_outcome) as COUNT from spacex
        group by landing_outcome order by count(landing_outcome) desc
```

✓ 0.5s

```
* ibm_db_sa://hnr90643:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08k
Done.
```

landing_outcome	COUNT
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

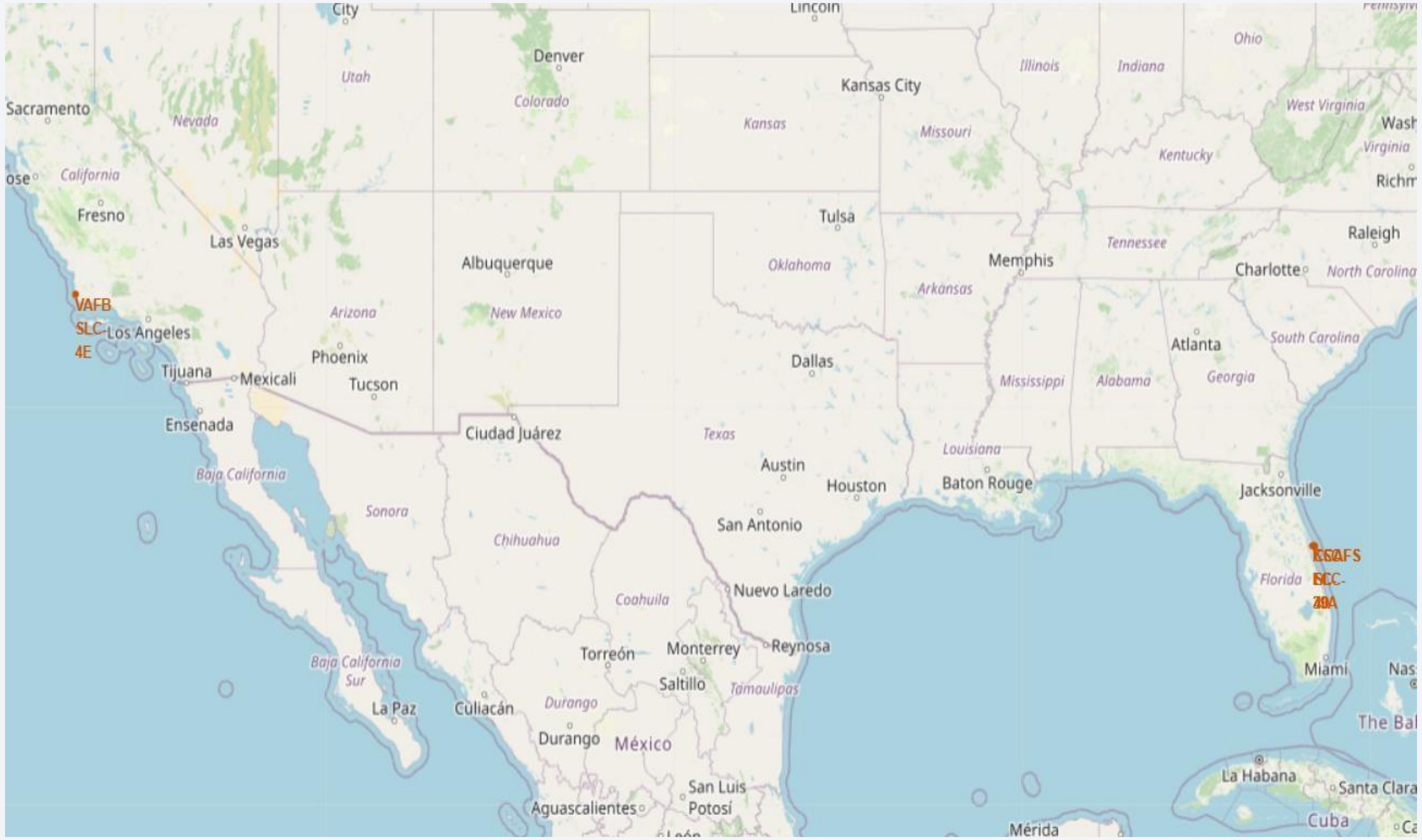
Section 3

Launch Sites Proximities Analysis

Map showing All Launch Sites

The map shows all the sites of Falcon 9 launch.

- We can see that the sites are located very close to the coastline as the failure rates of rockets are high (about 5-10%) and civilian areas must be avoided.
- 3 launch sites are located on the east coast of the US and 1 on the west coast
- all the sites are located close to the equator as it takes less fuel to launch a rocket from the equator

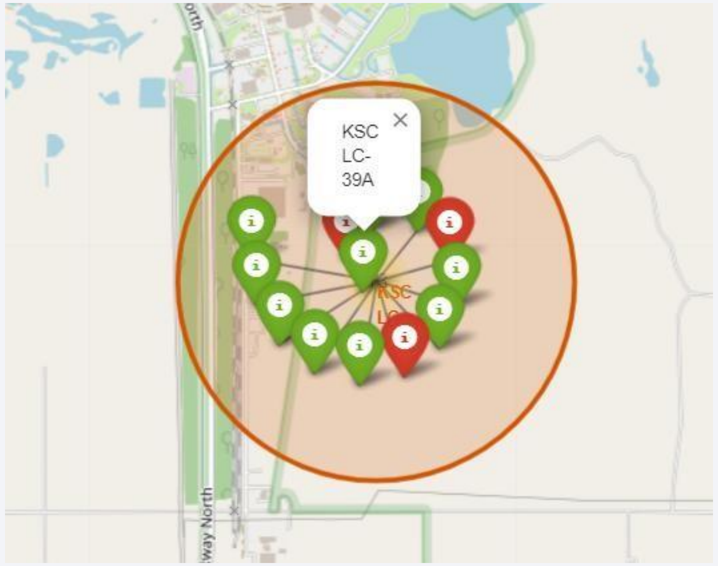


Map Showing the no. of launches for each site

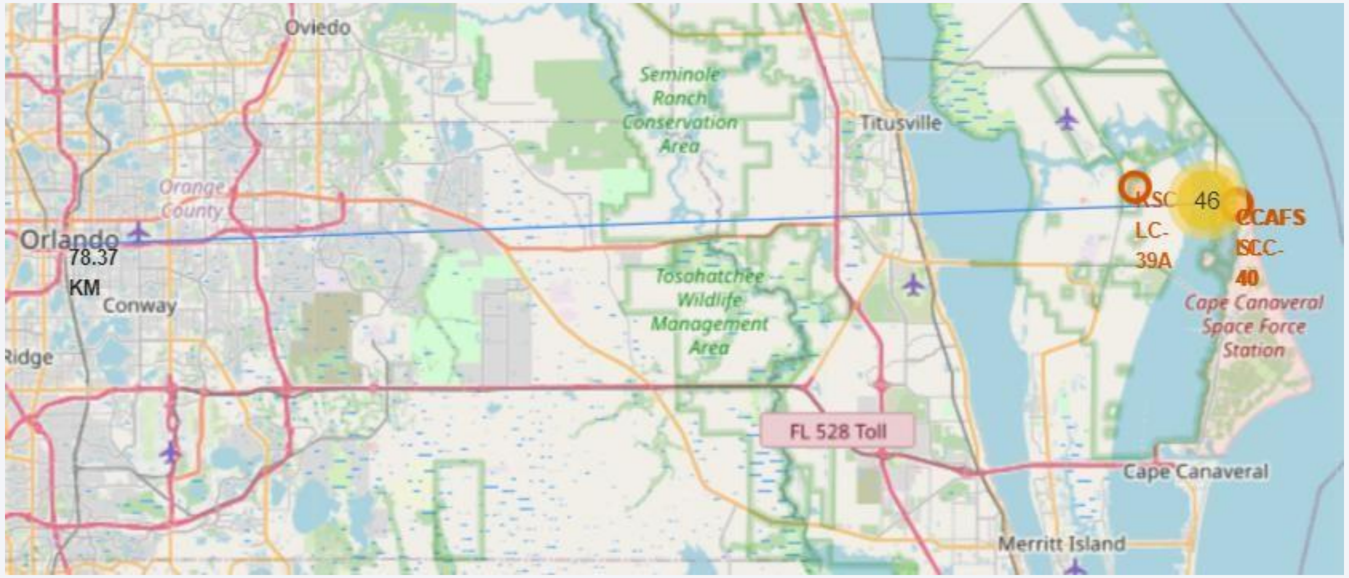
From the color-labeled markers in marker clusters, it can be easily identified which launch sites have relatively high success rates.

Thhe green markers show successful launches and red markers show unsuccessful ones.

Most launches happen on the east coast



Map showing distance of sites from nearby physical features



Marked the distance of site with the west coast, and calculated the distance of site from Orlando

We can observe that:

- launch sites are in close proximity to railways, highways, coastline
- Launch sites keep certain distance away from cities

Section 4

Build a Dashboard with Plotly Dash

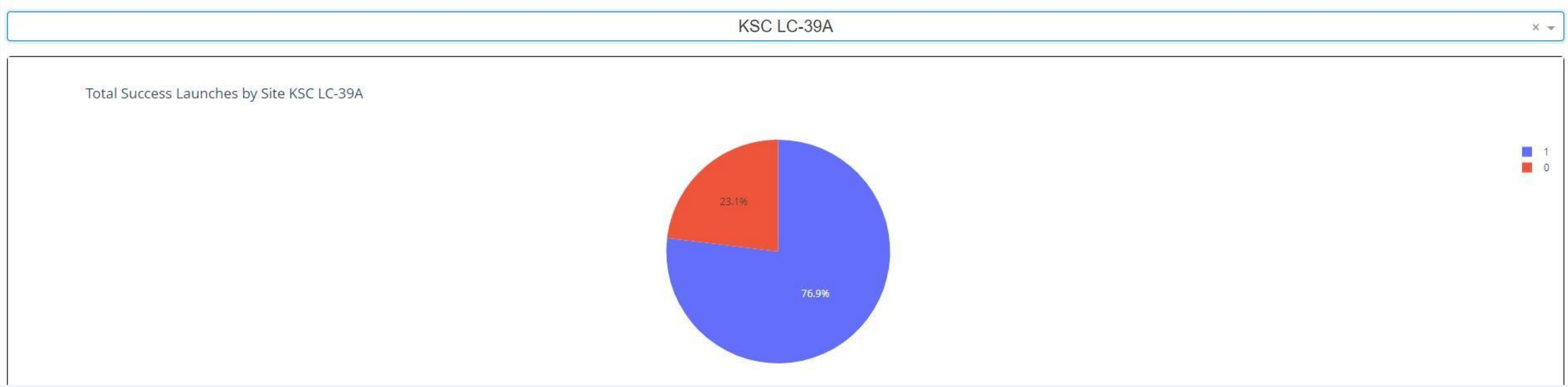
Launch Success Count Ratio For all Sites

The piechart shows the ratio of the count of successful launches for all sites.

We can see that KSC LC-39A has the highest no of successful launches of all sites.

CCAFS SLC-40 is the runner up in terms of successful launces

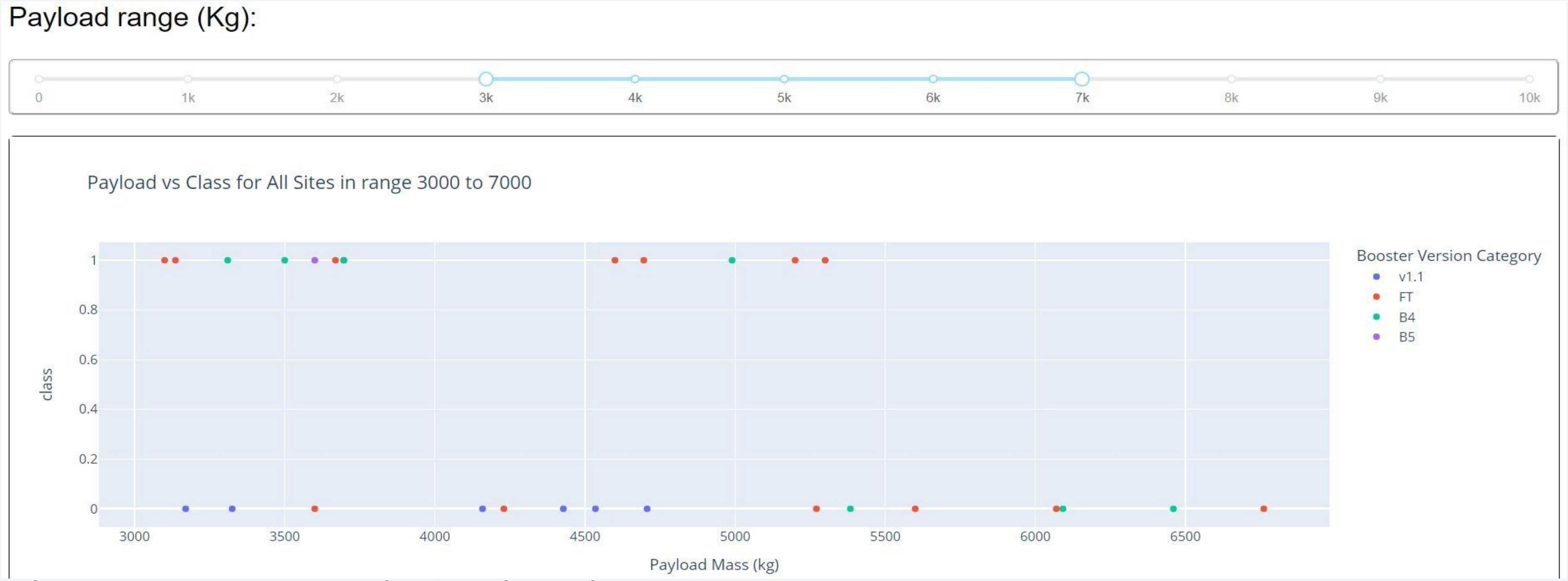
Launch site with highest launch success ratio



The KSC LC-39A has the highest no of successful launches of all sites, and this pie chart shows the success and failure share.

We can see that it has 76.9% success rate.

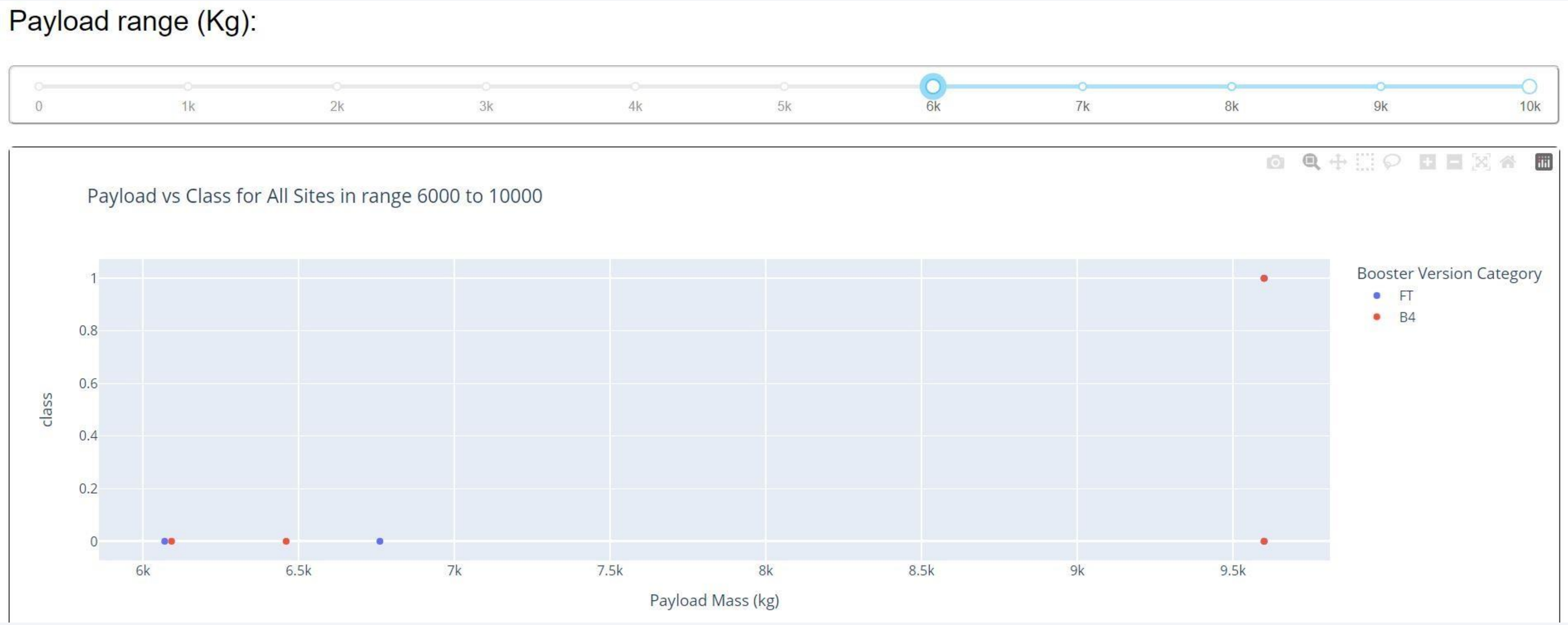
Payload vs Launch Outcome For All Sites



Payload Range 3000 to 7000 kg is selected:

- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.
- B4 booster has high success rate
- FT booster has average success rate while having a high failure rate for high payload.

Payload vs Launch Outcome For All Sites (Contd..)



Payload Range 6000 to 10000 kg is selected:

- There is a high failure rate for this payload range

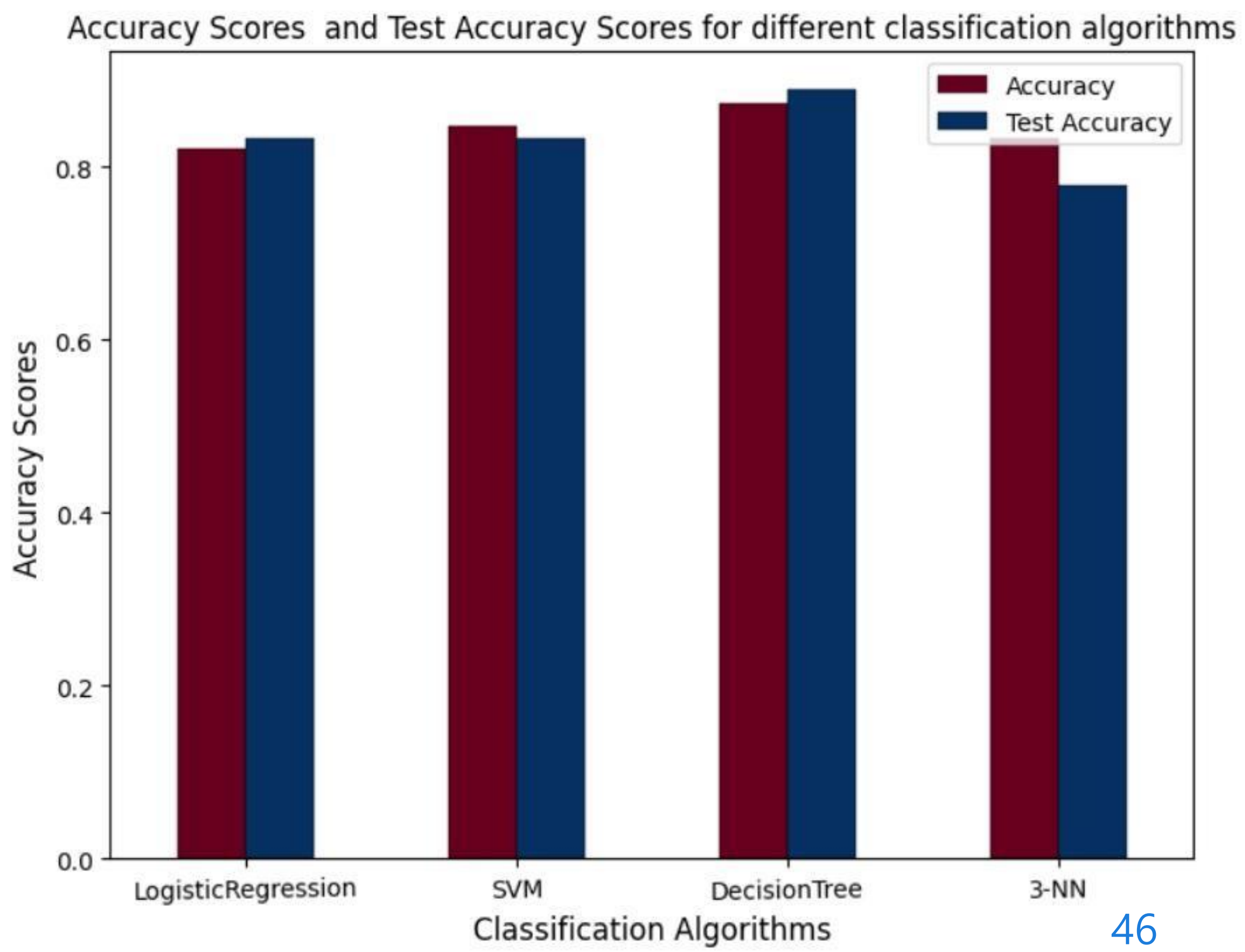
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Visualized the accuracy of all classification models used in the project in a bar chart, as in the figure beside:

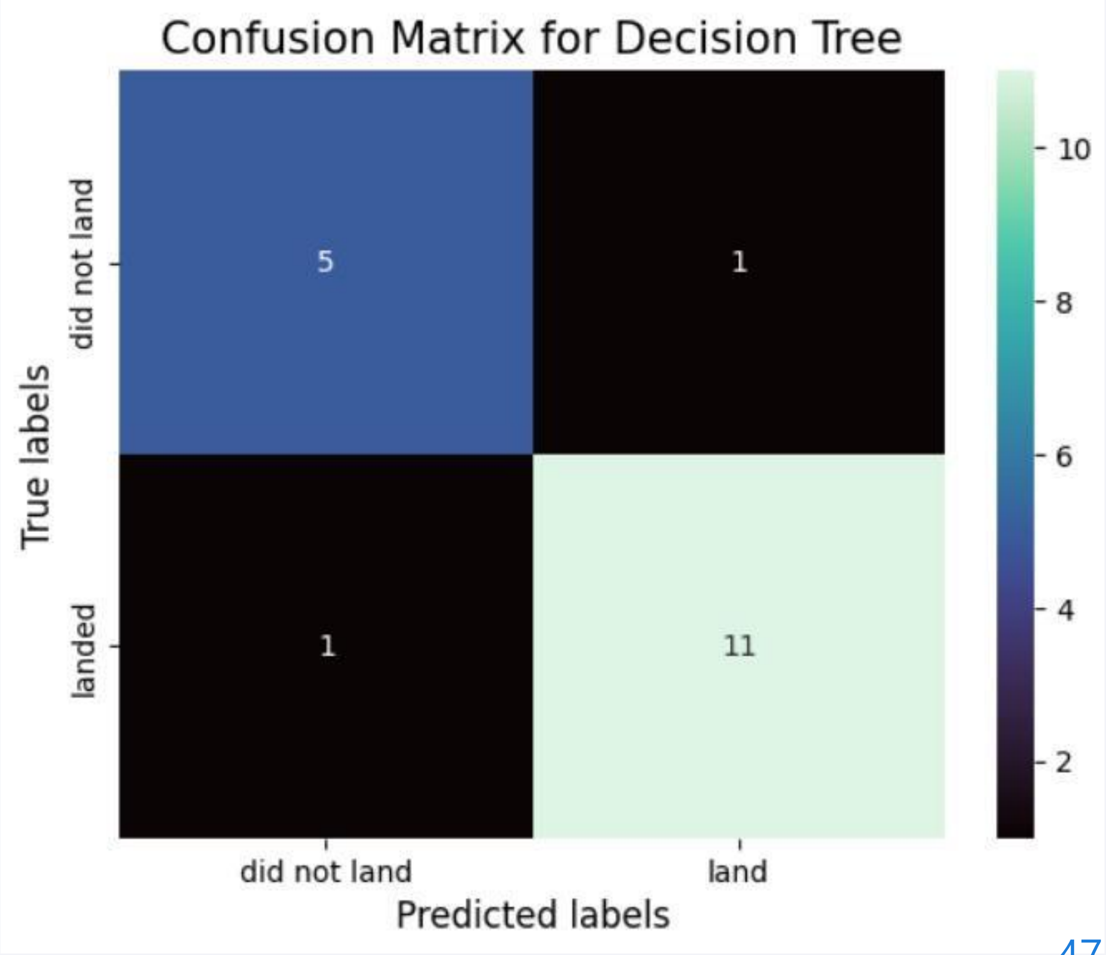
We can see that Decision Tree Classifier has the highest accuracy, while KNN with K=3 has the lowest accuracy score.



Confusion Matrix

The confusion matrix shows high number of true positive and true negative compared to the false ones.

The Decision Tree correctly classified 16 test points and misclassified only 2 test data points.



Conclusion

The conclusions drawn from the project are:

- The best launch site is KSC LC-39A
- Launches with payloads over 8000kg have high success rate
- VLEO orbit is overall a good choice for launch as it has high success rate for high number of launches
- Failure rate of new launches are low.
- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS orbits.
- Launch sites are located close to the equator, and in close proximity to the coast, railway lines and highways.
- Decision Tree Classifier is the best model for the problem and can be used to predict the success or failure of upcoming launches.

Thank you!