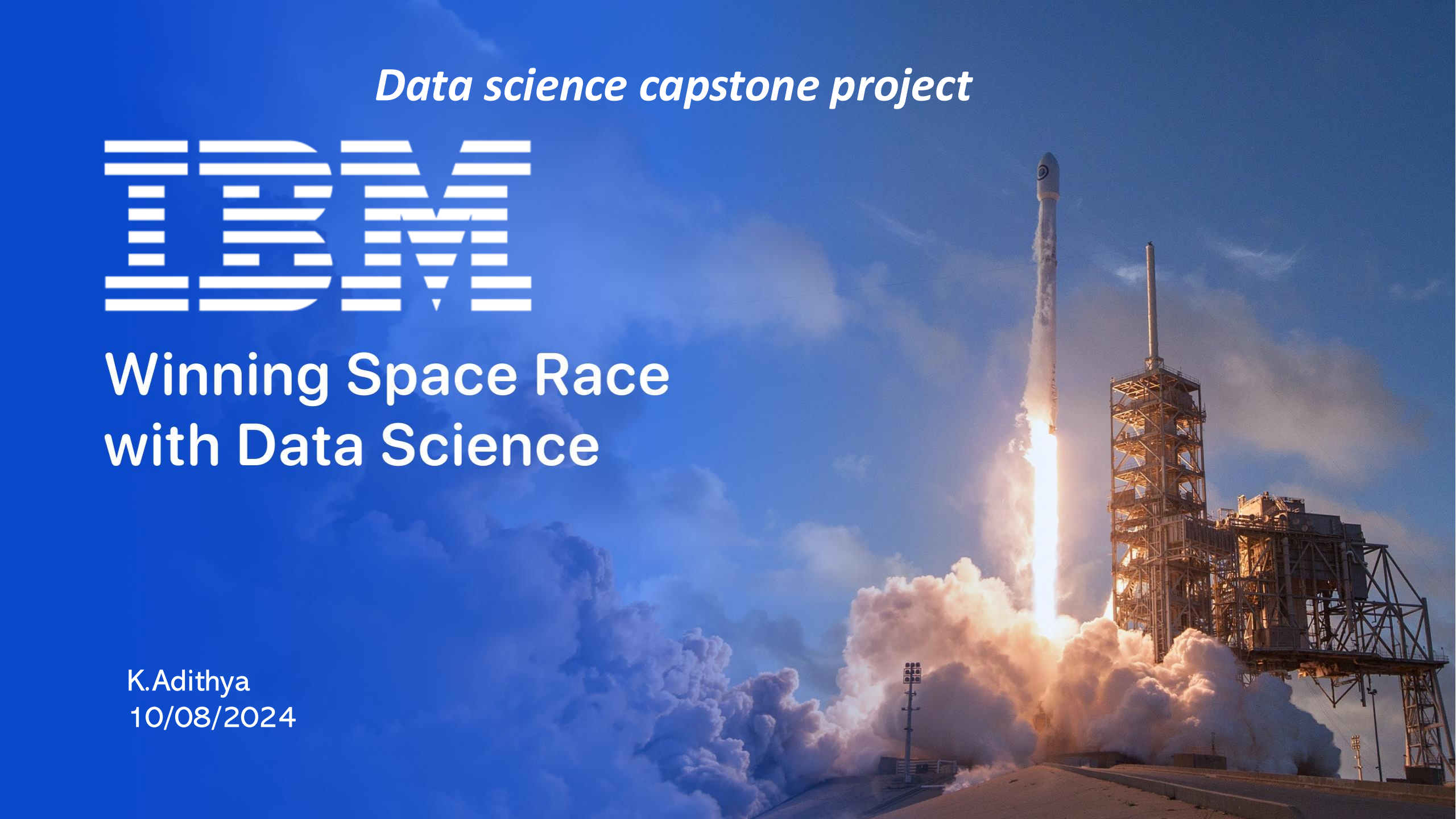
Data science capstone project



Winning Space Race with Data Science

K.Adithya

10/08/2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Discussion

Executive Summary

- The intention behind this capstone project was to predict if a SpaceX Falcon 9 first stage will land successfully or not using different machine learning classification algorithms.
- The general view of this project goes like this :
  - Data collection, wrangling, and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- According to our plots, a couple of features of the rocket launches are slightly correlated with the launch outcome, that is, as success or failure.
- It therefore leads to the conclusion that a decision tree model might be the most appropriate of the machine learning algorithms that can predict whether the Falcon 9 performs a successful first stage.

Introduction

- In this capstone, we will make a prediction of whether the Falcon 9 first stage will land successfully. SpaceX, on its webpage, advertises Falcon 9 rocket launches for 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings because SpaceX can reuse the first stage. Therefore, if we can determine whether or not the first stage can land, then we can estimate the cost for launching. That information comes in handy when some other alternate company might want to bid against SpaceX on a rocket launch.
- Most landings that fail are planned. Sometimes SpaceX will do a controlled landing in the ocean.
- The guiding question is this: Given a set of features regarding a Falcon 9 rocket launch, which includes payload mass, orbit type, launch site, and so on, the first stage of the rocket will land successfully.

Section 1

Methodology

Methodology

- The overall methodology includes:

1. Data collection, wrangling, and formatting, using:
   - SpaceX API
   - Web scraping
2. Exploratory data analysis (EDA), using:
   - Pandas and NumPy
   - SQL
3. Data visualization, using:
   - Matplotlib and Seaborn
   - Folium
   - Dash
4. Machine learning prediction, using
   - Logistic regression
   - Support vector machine (SVM)
   - Decision tree
   - K-nearest neighbors (KNN)

Data Collection

- SpaceX API
  - The API used is https://api.spacexdata.com/v4/rockets/.
  - The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.
  - Every missing value in the data is replaced the mean the column that the missing value belongs to.
  - We end up with 90 rows or instances and 17 columns or features. The picture below shows the first few rows of the data:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedC
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	

Data Collection - Scraping

Web scraping

- The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The website contains only the data about Falcon 9 launches.
- We end up with 121 rows or instances and 11 columns or features. The picture below shows the first few rows of the data:

[14]:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA (COTS)\nNRO	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

- The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.
- In the end, we end up with 90 rows or instances and 83 columns or features.

[19]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	Land
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	0	1	False	False	False	
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	0	1	False	False	False	
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	0	1	False	False	False	
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	0	1	False	False	False	
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	0	1	False	False	False	

EDA with Data Visualization

Pandas and NumPy

- Functions from the Pandas and NumPy libraries are used to derive basic information about the data collected, which includes:
  - The number of launches on each launch site
  - The number of occurrence of each orbit
  - The number and occurrence of each mission outcome

SQL

- The data is queried using SQL to answer several questions about the data such as:
  - The names of the unique launch sites in the space mission
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1

Build an Interactive Map with Folium and Seaborn

Matplotlib and Seaborn

- Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts.
- The plots and charts are used to understand more about the relationships between several features, such as:
  - The relationship between flight number and launch site
  - The relationship between payload mass and launch site
  - The relationship between success rate and orbit type

Folium

- Functions from the Folium libraries are used to visualize the data through interactive maps.
- The Folium library is used to:
  - Mark all launch sites on a map
  - Mark the succeeded launches and failed launches for each site on the map
  - Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

Build a Dashboard with Plotly Dash

Dash

- Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.

- Using a pie chart and a scatterplot, the interactive site shows:

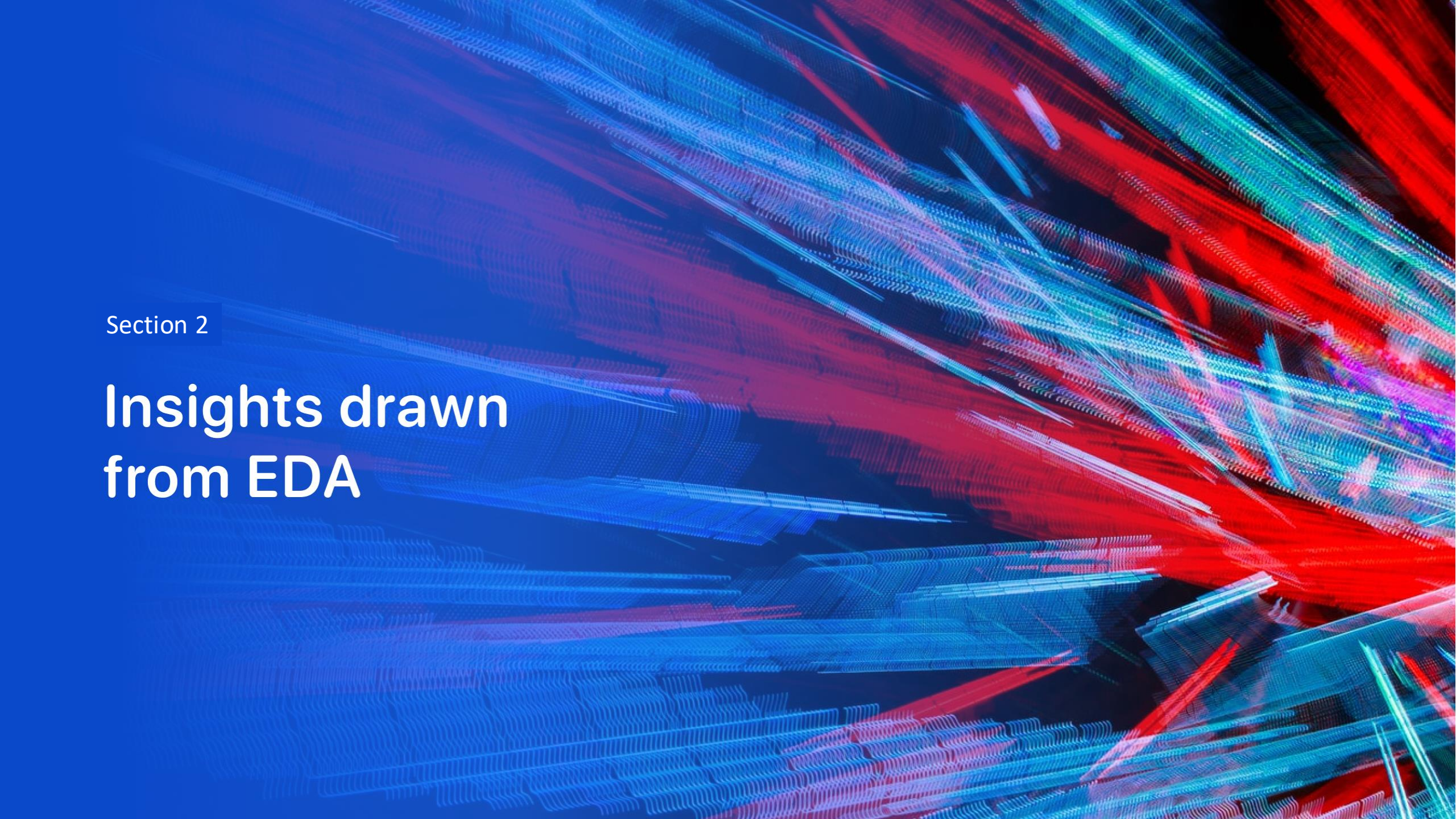
    - The total success launches from each launch site
    - The correlation between payload mass and mission outcome (success or failure) for each launch site

Predictive Analysis (Classification)

- Functions from the Scikit-learn library are used to create our machine learning models.
- The machine learning prediction phase include the following steps:
  - Standardizing the data
  - Splitting the data into training and test data
  - Creating machine learning models, which include:
    - Logistic regression
    - Support vector machine (SVM)
    - Decision tree
    - K nearest neighbors (KNN)
- Fit the models on the training set
- Find the best combination of hyperparameters for each model
- Evaluate the models based on their accuracy scores and confusion matrix

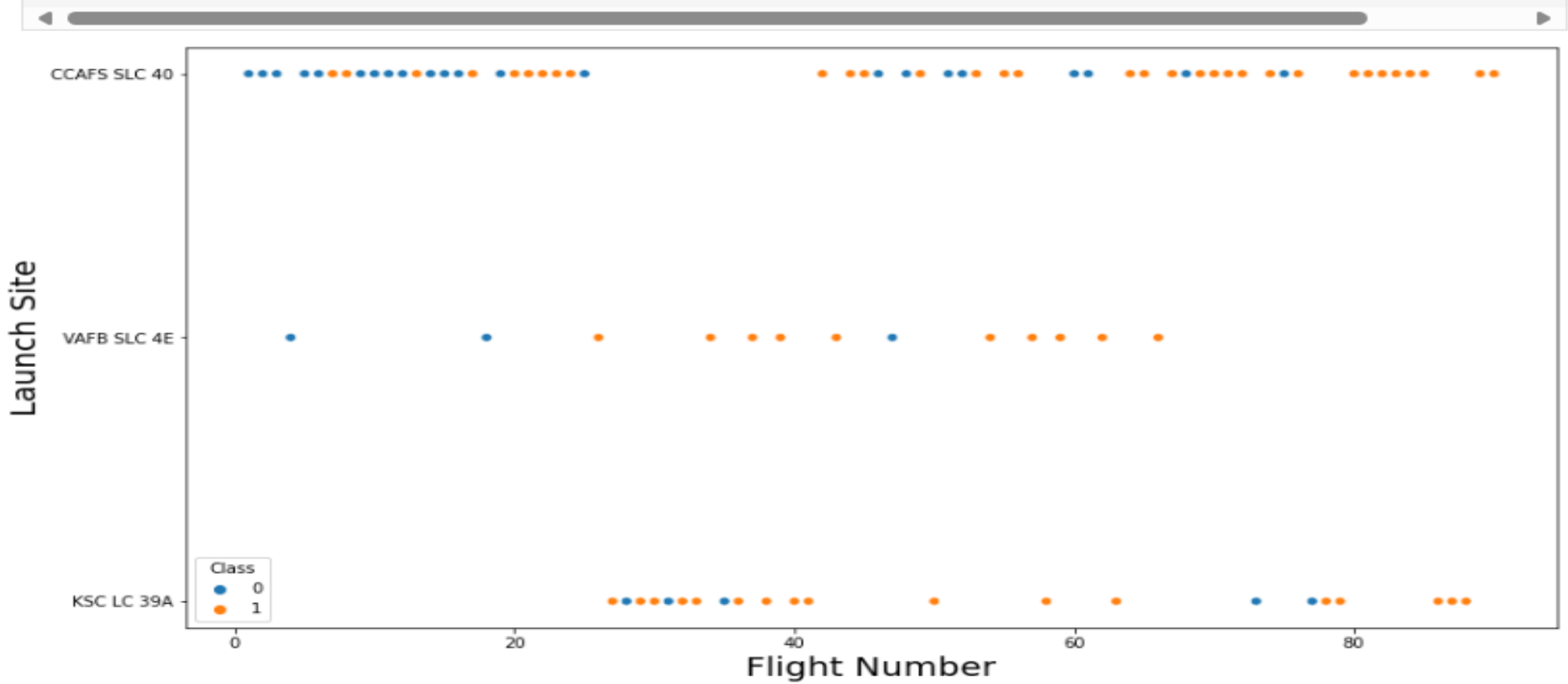
Results

- The results are split into 5 sections:
  - SQL (EDA with SQL)
  - Matplotlib and Seaborn (EDA with Visualization)
  - Folium
  - Dash
  - Predictive Analysis
- In all of the graphs that follow, class 0 represents a failed launch outcome while class 1 represents a successful launch outcome.


Section 2
Insights drawn from EDA

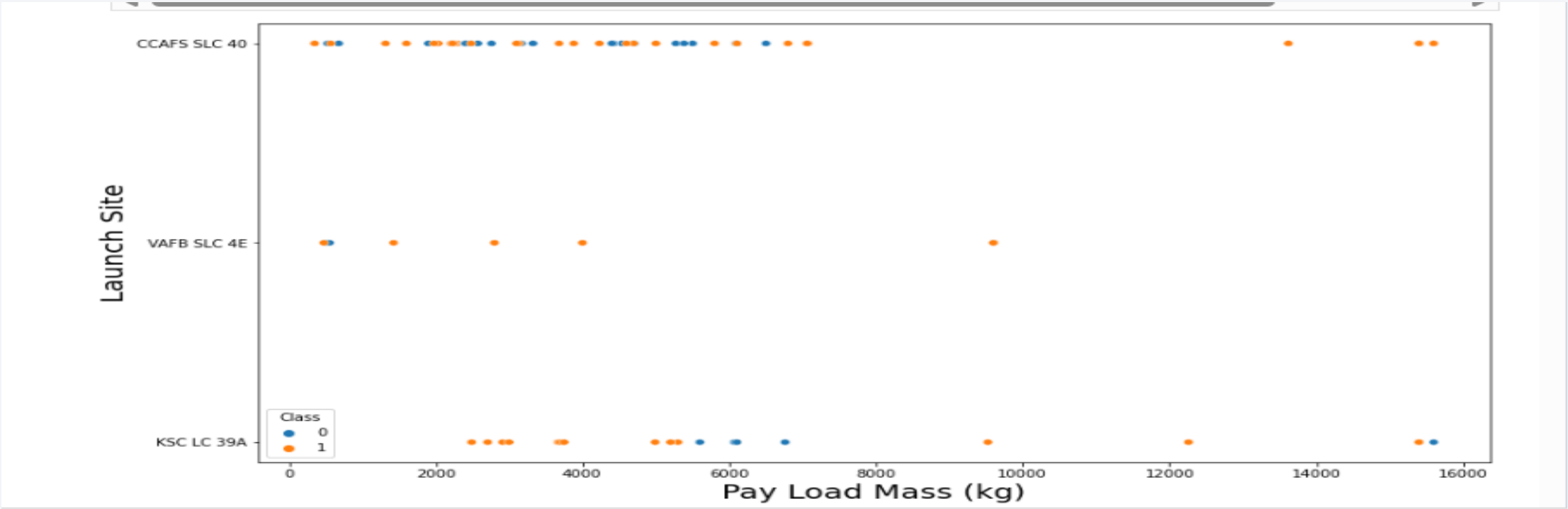
Flight Number vs. Launch Site

- The below figure shows the Flight Number of rockets and the base from they are launched



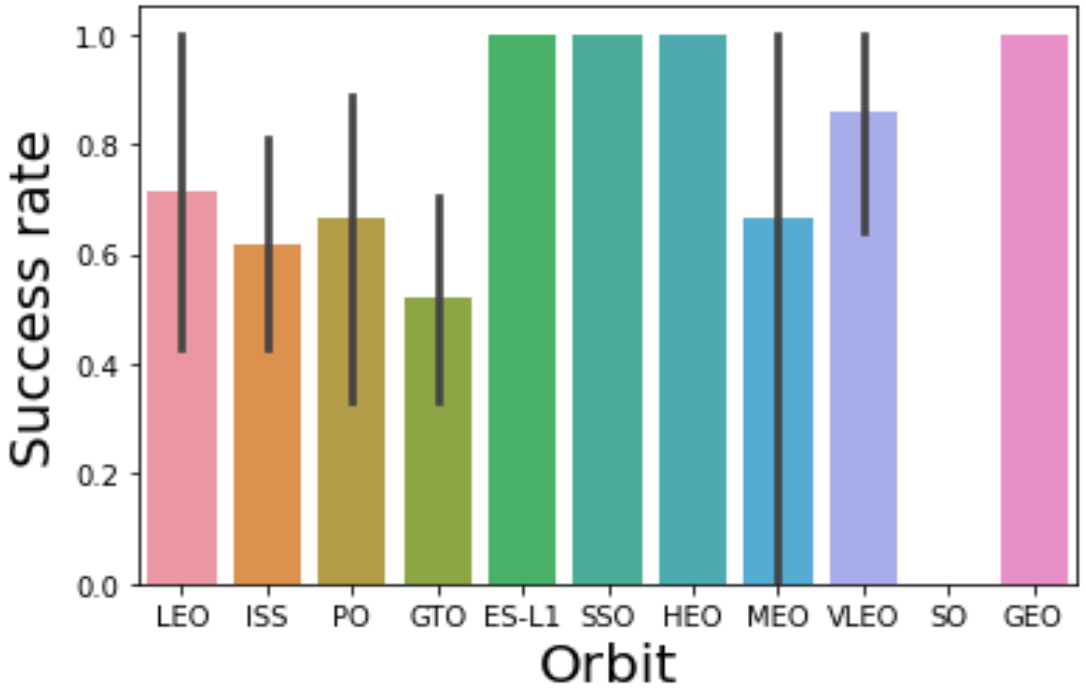
Payload vs. Launch Site

- This figure shows relation between payload and Launch site. Depending on payload what site was chosen for what payload category



Success Rate vs. Orbit Type

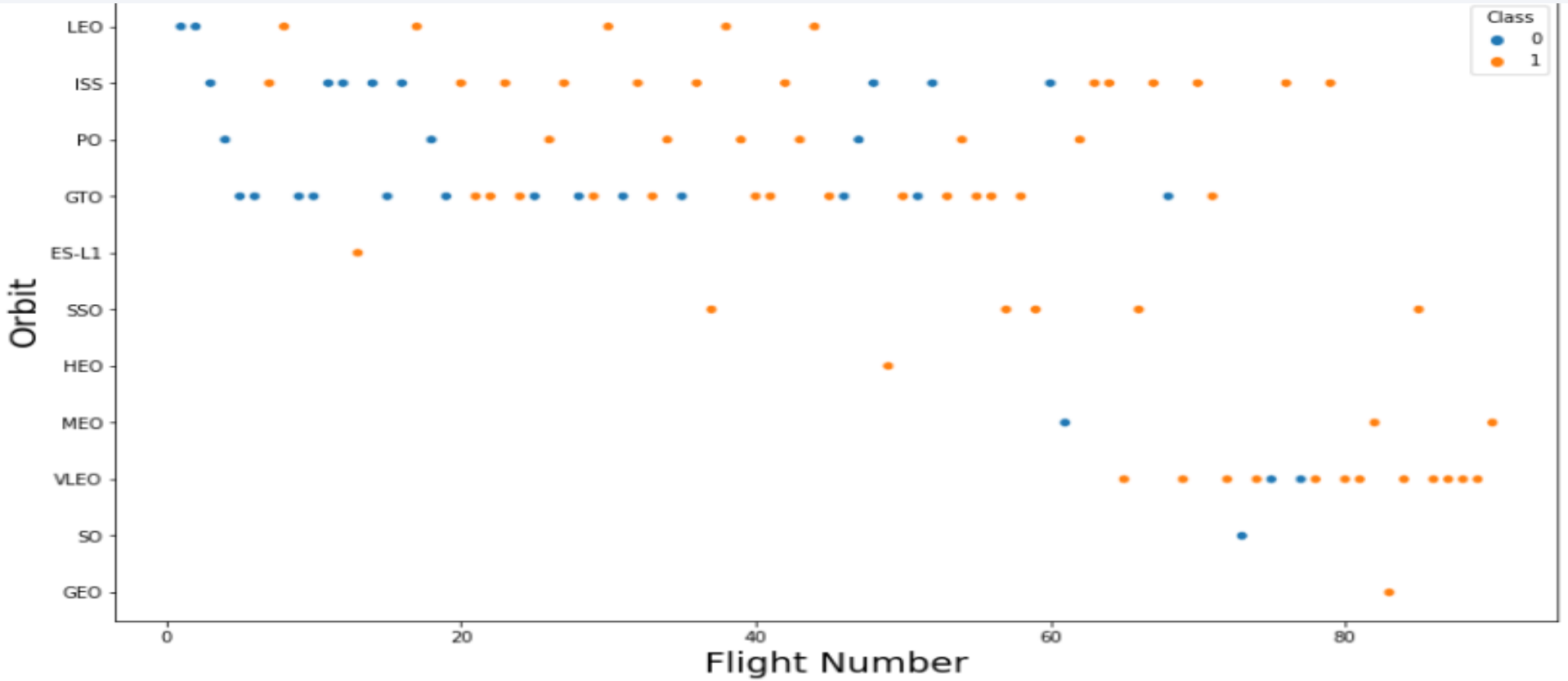
This figure shows the relation between orbital type and success rate of rockets launched to those orbits



Analyze the ploted bar chart try to find which orbits have high sucess rate.

Flight Number vs. Orbit Type

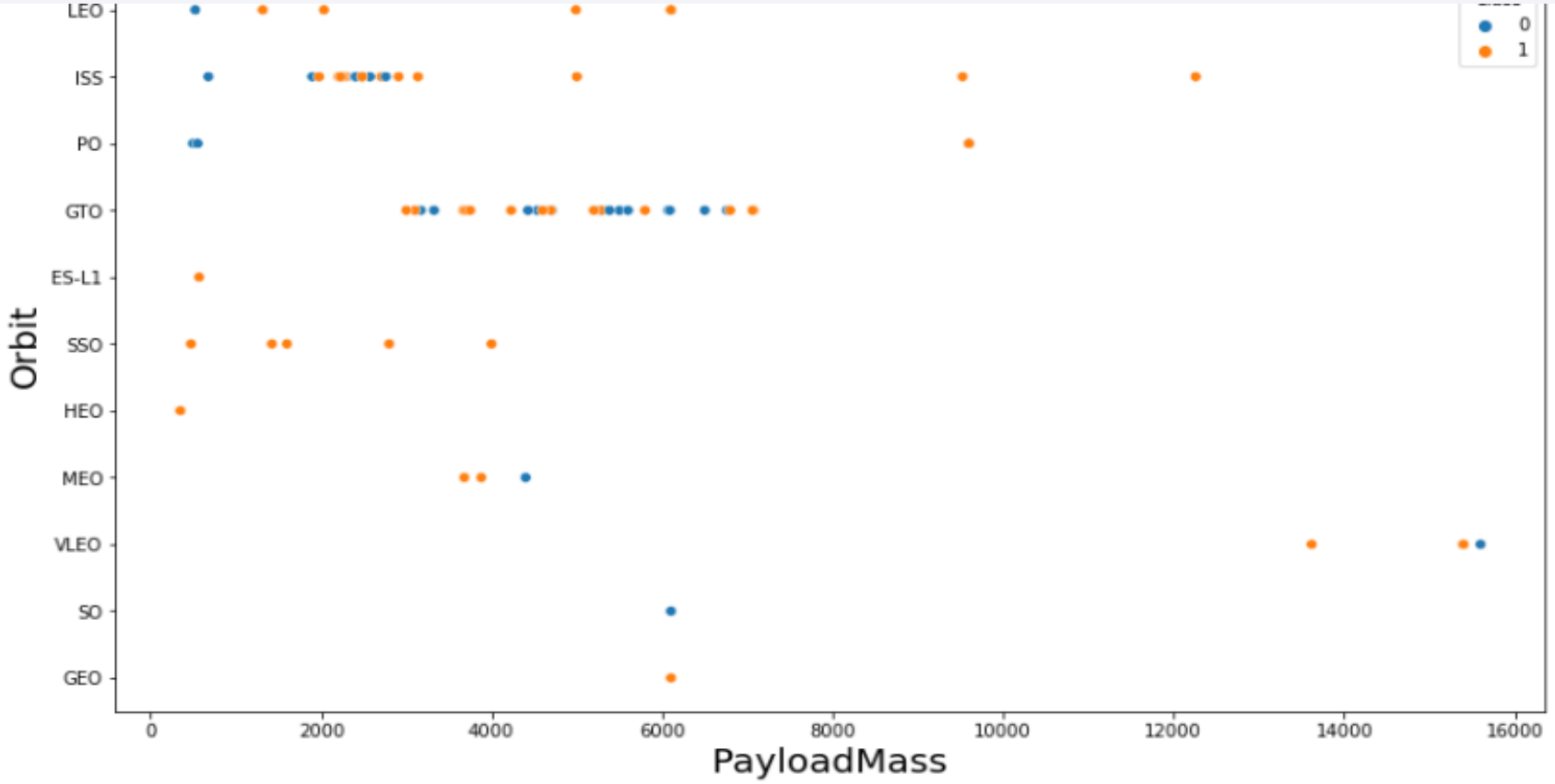
- This figure shows the flight numbers and their orbital types



You should 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.

Payload vs. Orbit Type

This figure shows the type of payload is used for various orbital levels

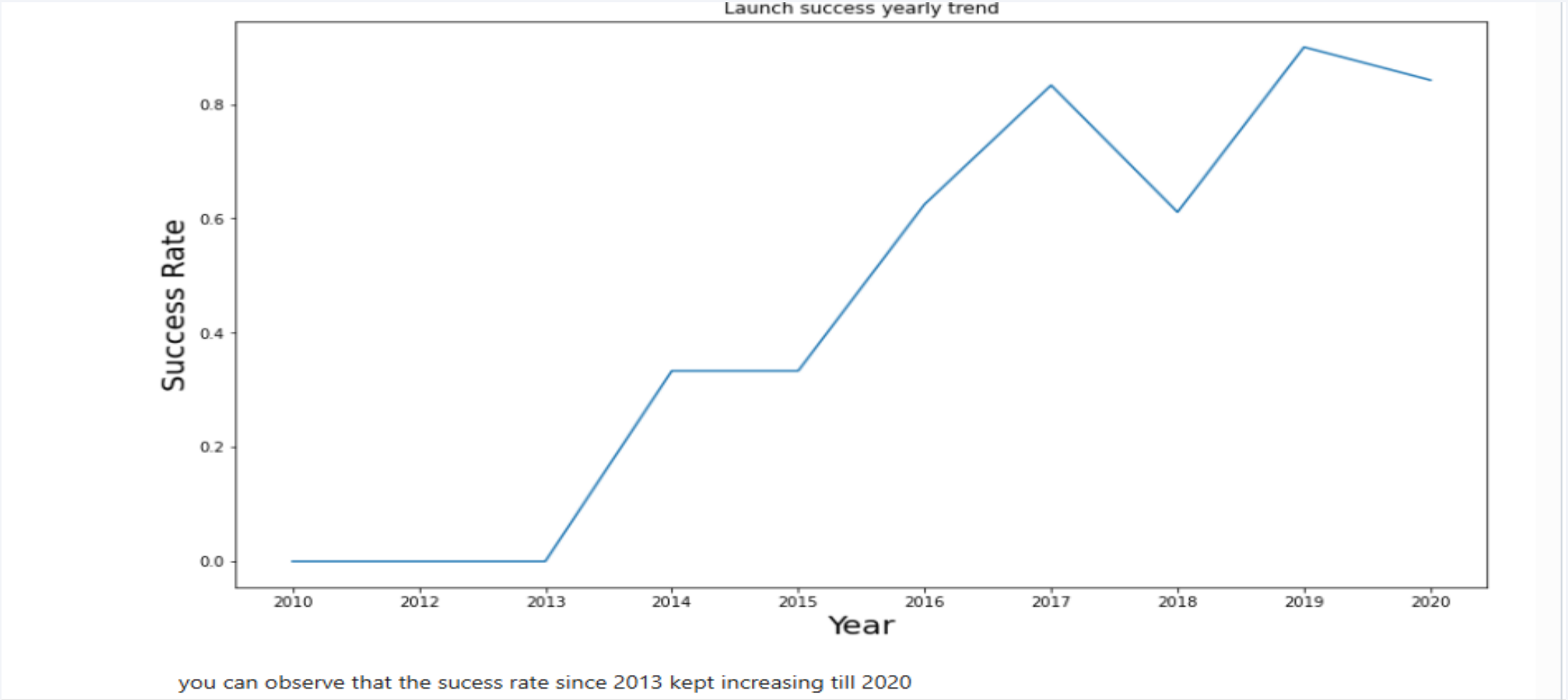


With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

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

Launch Success Yearly Trend

- This figure shows the successful trend of rocket launches.



All Launch Site Names

- The below figure shows all lauch site names

```
In [16]: %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

```
 * ibm_db_sa://dgy37633:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

Out[16]:

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

Launch Site Names Begin with 'CCA'

This figure shows 5 records where launch sites begin with `CCA`

Out[17]:

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landin
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure
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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass and Average Payload Mass by F9 v1.1

- The total payload carried by boosters from NASA

Total payload mass by NASA (CRS)
45596

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

Average payload mass by Booster Version F9 v1.1
2928

First Successful Ground Landing Date and Successful Drone Ship Landing with Payload between 4000 and 6000

The first successful landing outcome on ground pad

Date of first successful landing outcome in ground pad
2015-12-22

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

Done.

Out[21]:

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

Total Number of Successful and Failure Mission Outcomes and Boosters Carried Maximum Payload

The total number of successful and failure mission outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

The names of the booster which have carried the maximum payload mass

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

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

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

- Ranking 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.

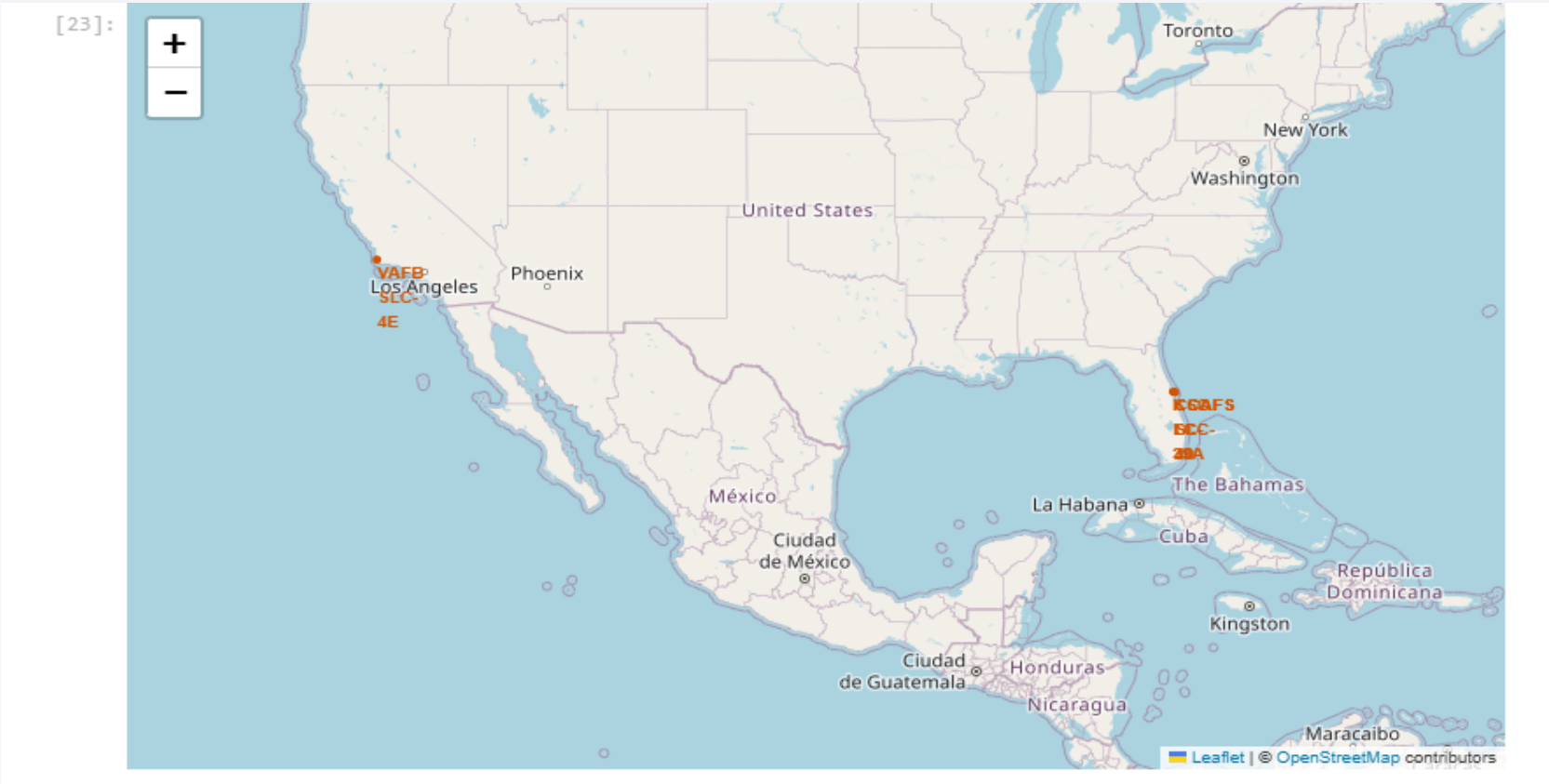
landing_outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

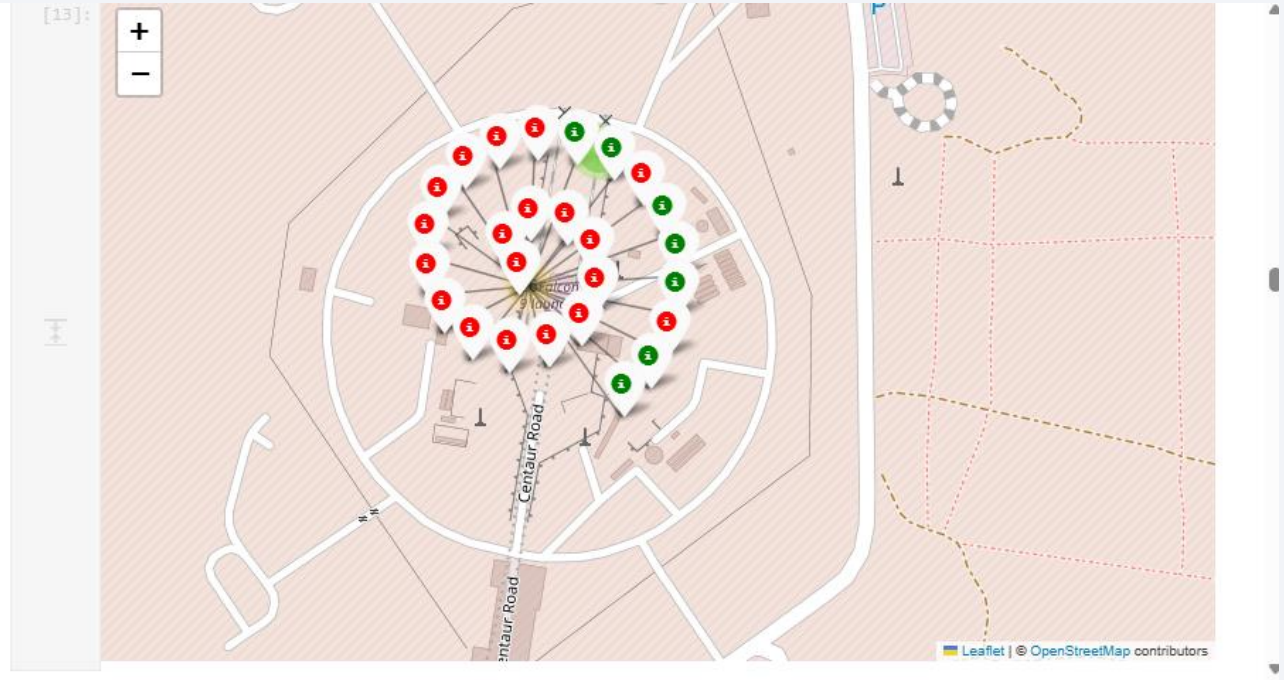
All launch sites on a map with folium

Exploring the generated folium map and all launch sites' location markers on a global map and they are CCAFS LC-40,CCAFS SLC-40,KSC LC-39A ,VAFB SLC-4E



Successful launch outcomes at a site

Exploring the folium map to show the color-labeled launch outcomes on the map. Site : CCAFS LC-40 . Red tag shows it fail and green mean success



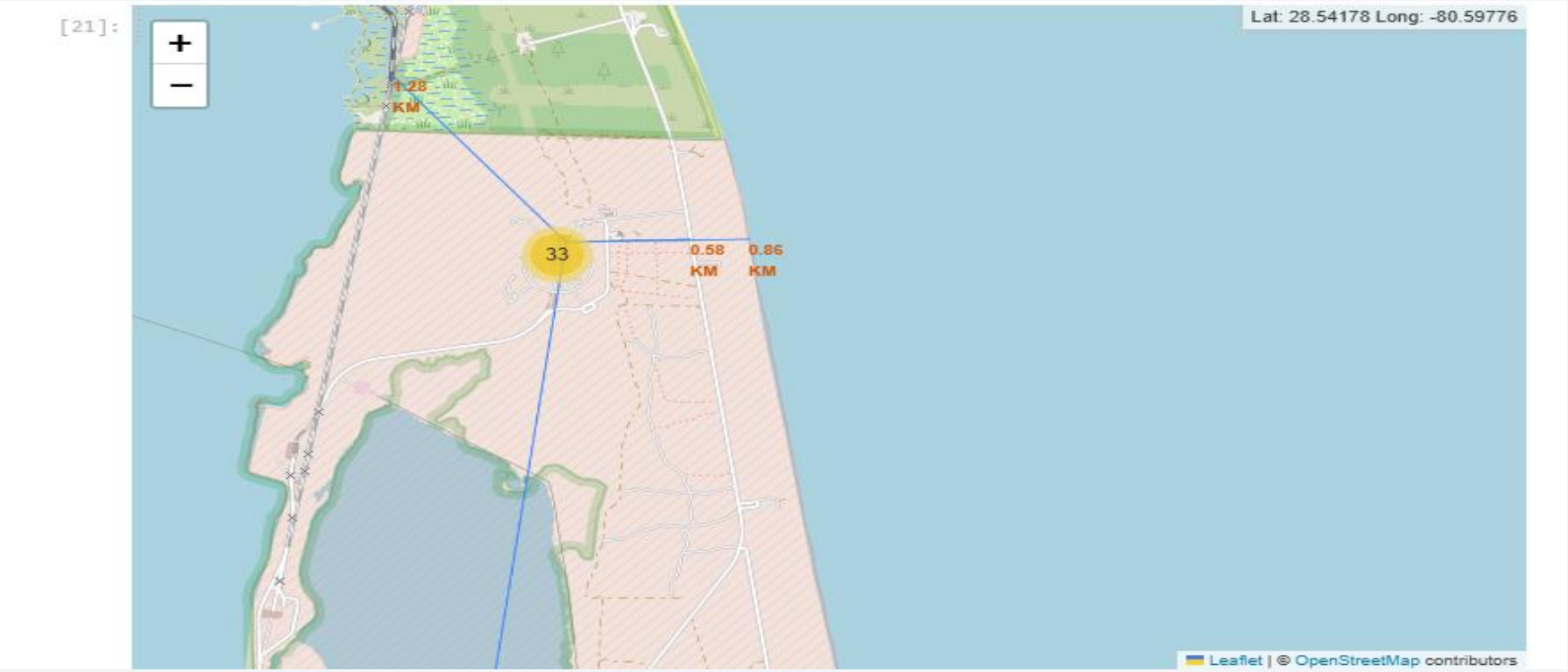
Infrastructure distance from launch site

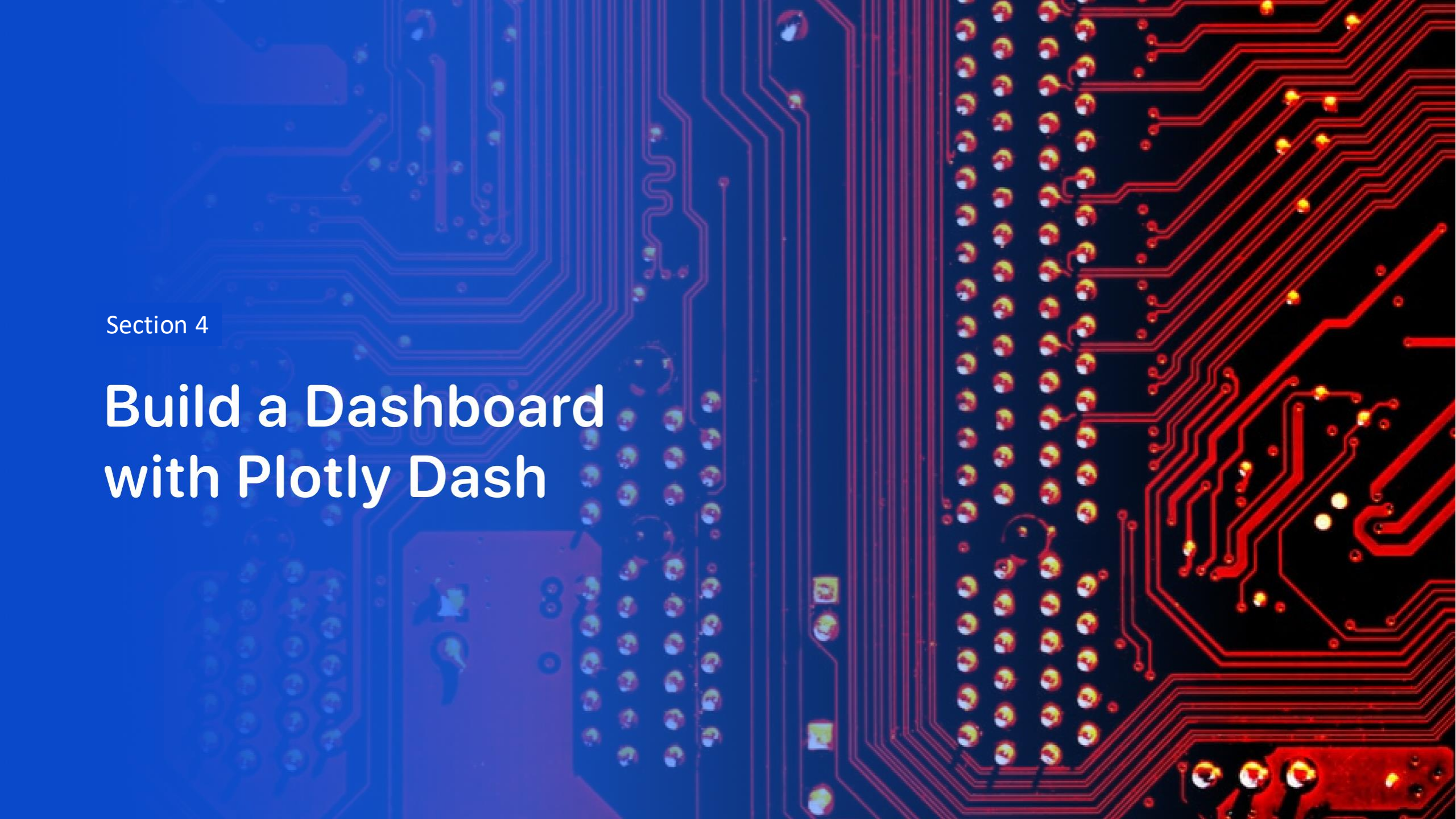
Exploring the generated folium map of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed.

distance_highway = 0.5834695366934144 km

distance_railroad = 1.2845344718142522 km

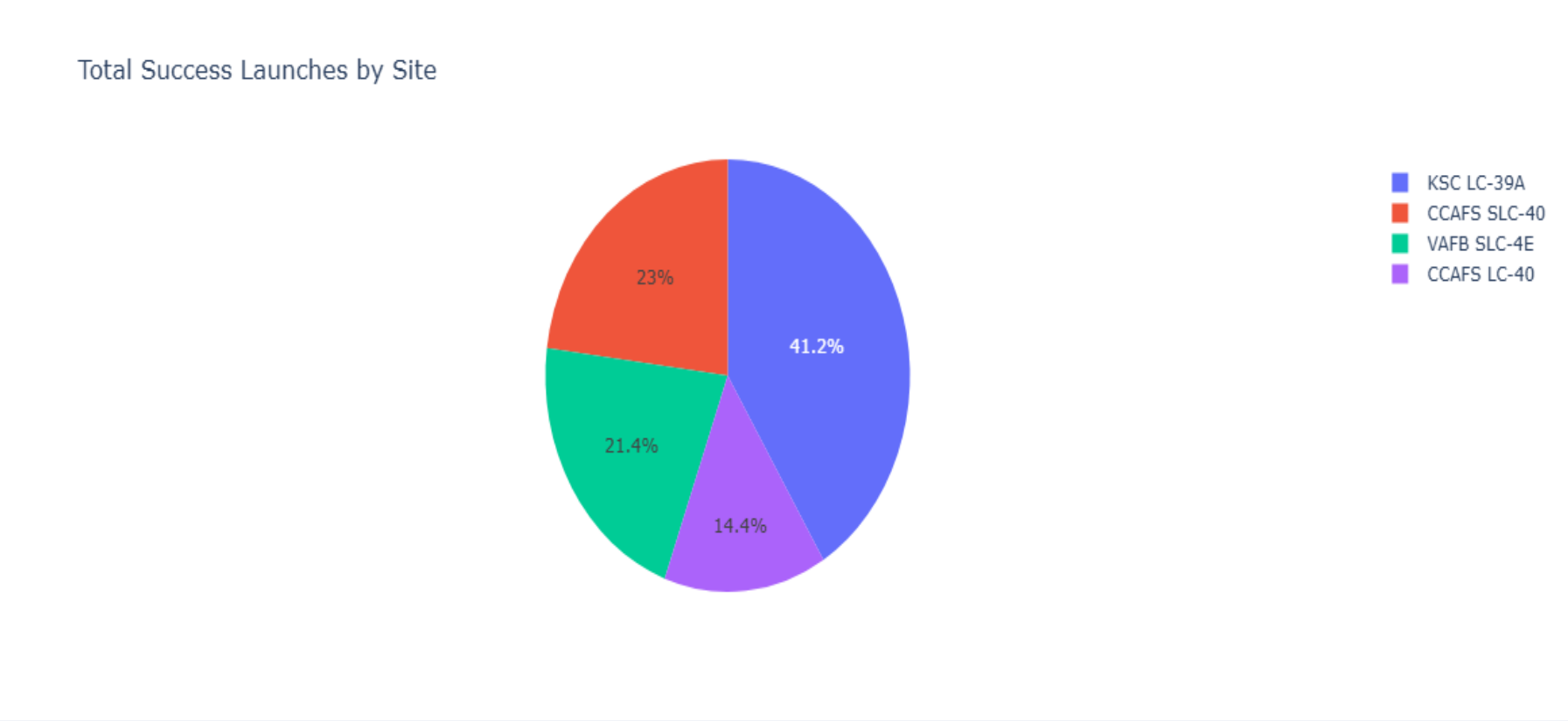
distance_city = 51.434169995172326 km




Section 4
Build a Dashboard with Plotly Dash

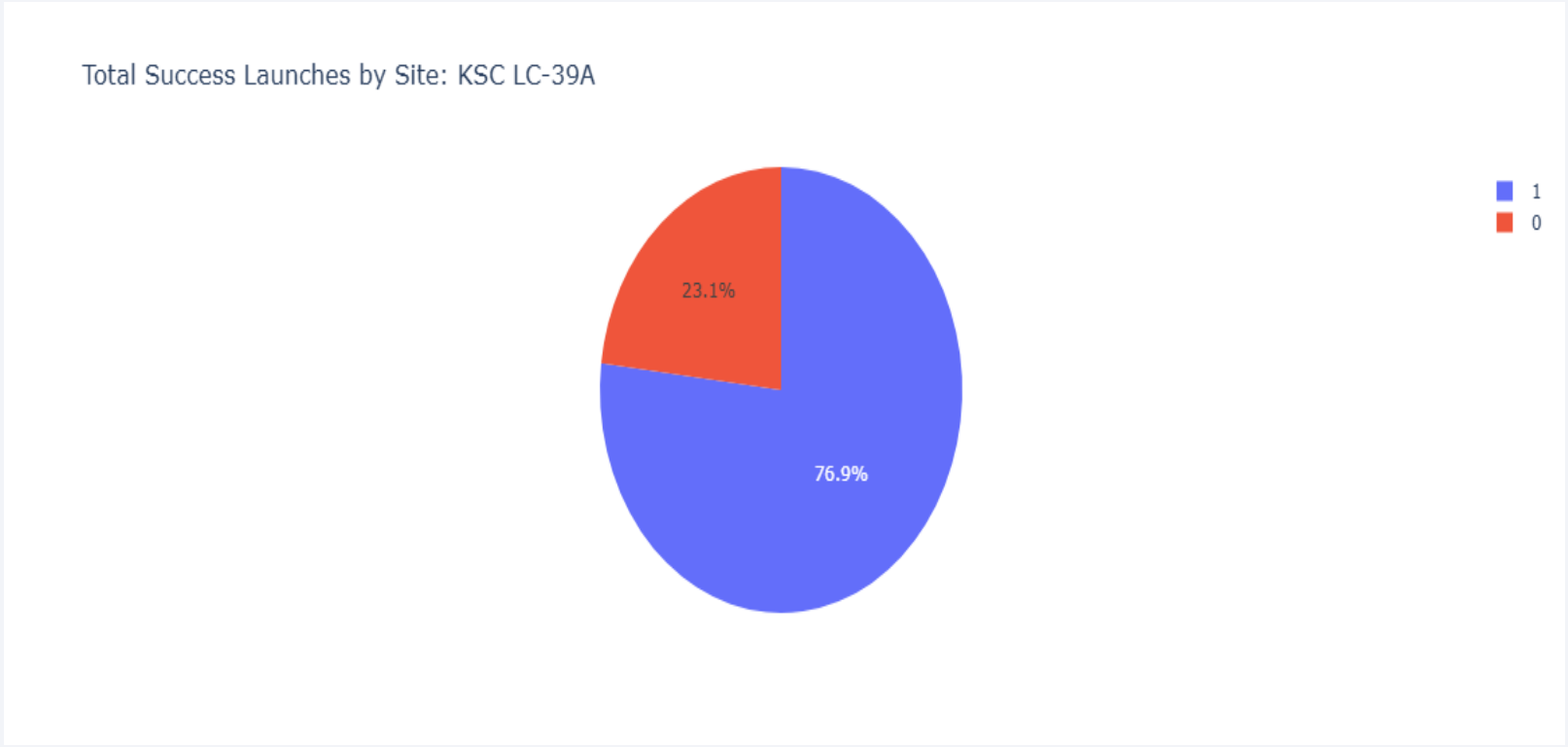
Success rate of each launch site

This below figure shows launch success count for all sites, in a pie-chart.



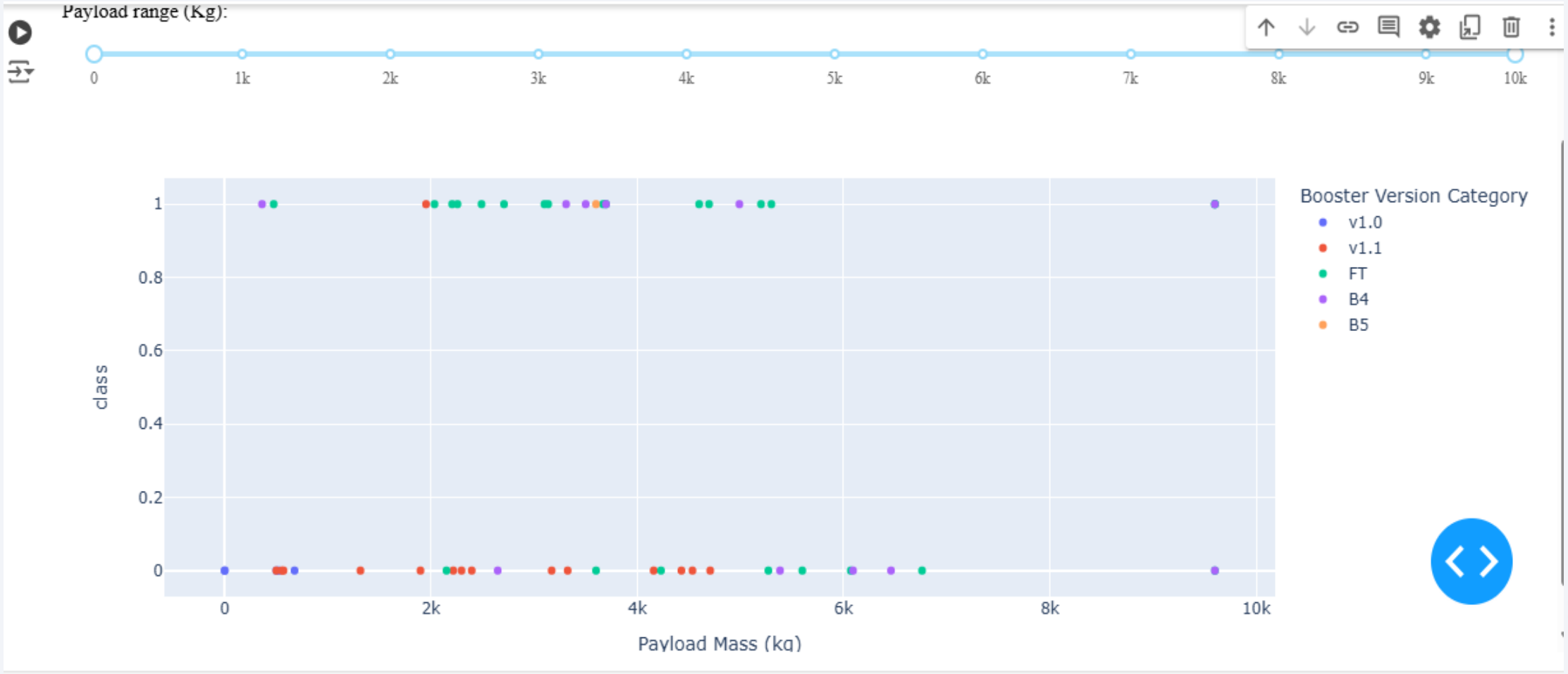
Launch site with highest success ratio

The pie-chart for the launch site with highest launch success ratio is KSC LC-39A with a success rate of 76.9%



Payload vs Booster version

Payload vs. Booster version scatter plot for all sites, with different payload selected in the range slider. Class 0 shows failed and class one shows success.



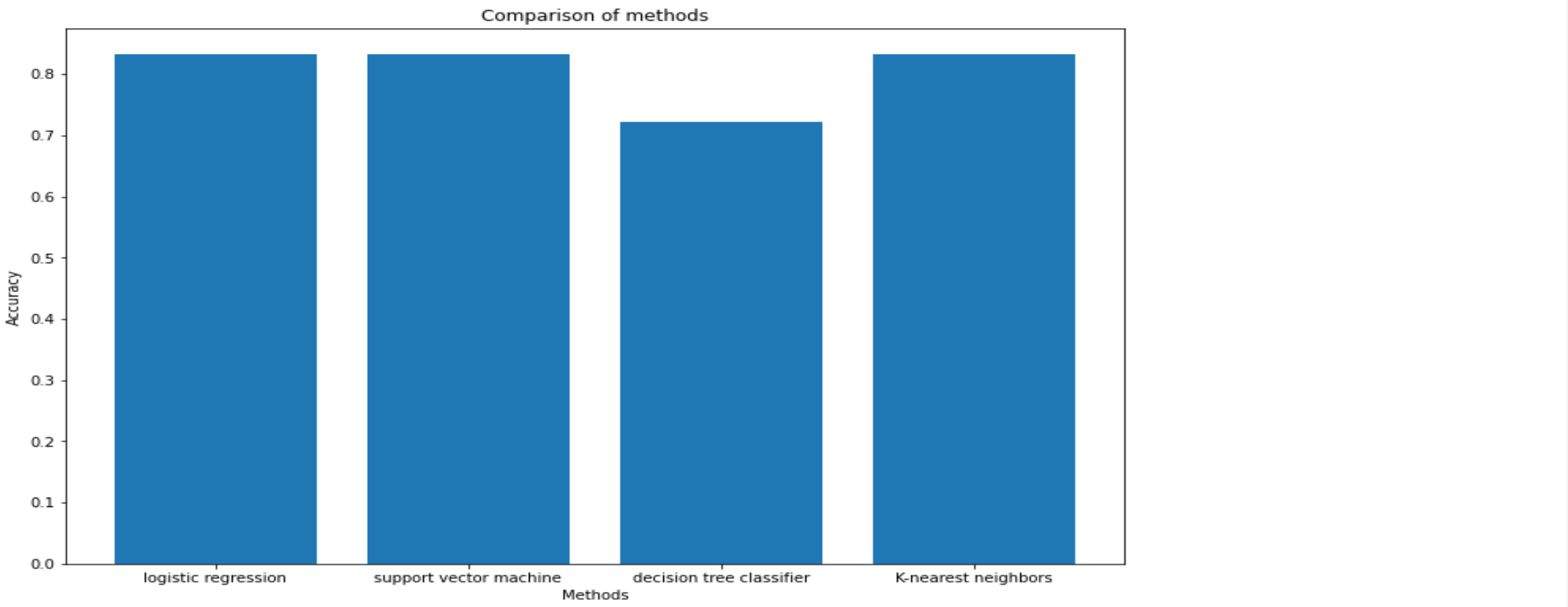
Section 5

Predictive Analysis (Classification)

Classification Accuracy

This below chart Visualizes the built model accuracy for all built classification models, in a bar chart.

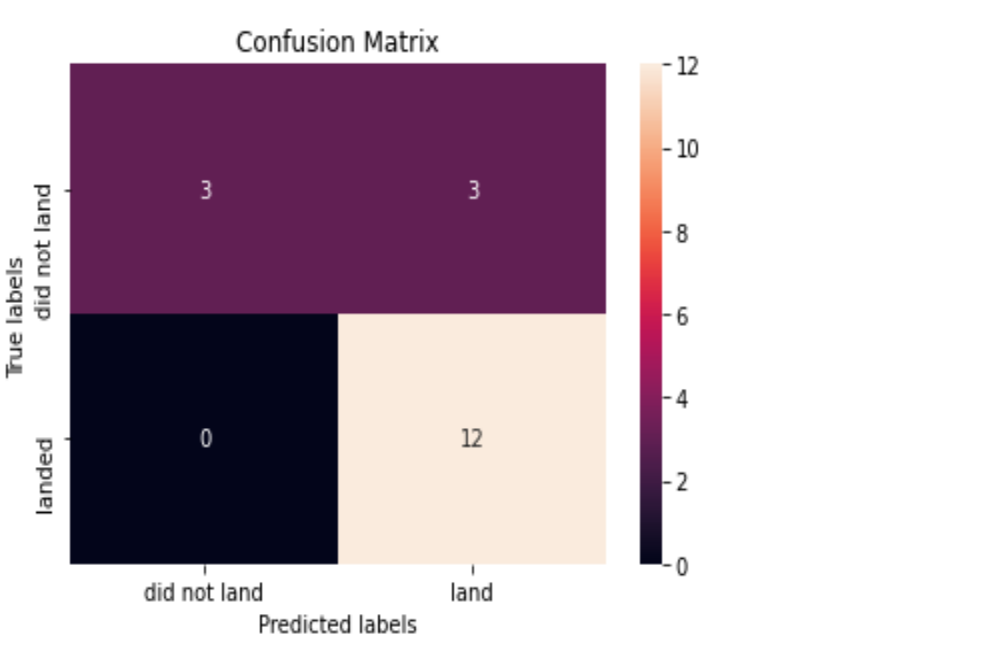
Logistic regression, Support vector machine and K-nearest neighbors, all of these three methods give the best performance, with accuracy of 0.8333333333333334.

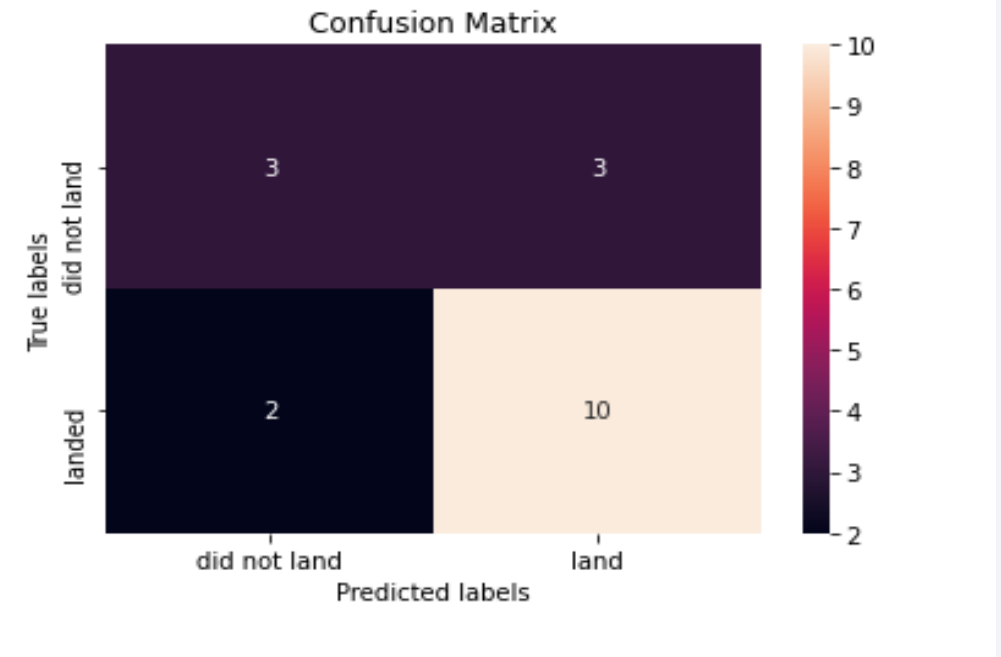


Confusion Matrix

a. Putting the results of all 4 models on a bar chart, we can see that except decision tree model all share the same accuracy score and confusion matrix when tested on the test set.

b. Fig 1 shows confusion matrix of all three that is logistic, K means, SVM and fig 2 shows confusion matrix of decision tree





Conclusions

- Therefore, their GridSearchCV best scores on a training data set and it is used to rank them instead. Based on the GridSearchCV best scores, the models are ranked in the following order with the first being the best and the last one being the worst:
  - Decision tree (GridSearchCV best score: 0.8892857142857142)
  - K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
  - Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
  - Logistic regression (GridSearchCV best score: 0.8464285714285713)

- Based on test set only Decision Tree set performs worst it shows overfitting with accuracy of 0.72 on test data and other models perform similarly on test data with an accuracy of 0.83

Discussion

- From the data visualization section, we can see that some features may have correlation with the mission outcome in several ways. For example, with heavy payloads the successful landing or positive landing rate are more for orbit types Polar, LEO and ISS. However, for GTO, we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
- Therefore, each feature may have a certain impact on the final mission outcome. The exact ways of how each of these features impact the mission outcome are difficult to decipher. However, we can use some machine learning algorithms to learn the pattern of the past data and predict whether a mission will be successful or not based on the given features.

Thank you!