IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Muhammad Farhan
Naveed
06-07-2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA (Exploratory Data Analysis) with data visualization
  - EDA with SQL
  - Interactive map with Folium
  - Interactive dashboard with plotly dash
  - Machine Learning Predictive (Classification) analysis
- Summary of all results
  - EDA results obtained
  - Interactive analytics built
  - Predictive analysis result obtained

Introduction

Project background and context

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 are 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.

In this project, we will take the role of a data scientist working for a new rocket company and try to predict if the first stage will land successfully or not taking different factors into account and using different methodologies/techniques of data analysis, data visualization, and machine learning.

Problems you want to find answers

- On what factors the successful landing of the rocket depends on?
- Effect of various factors on each other that determine the successful landing of rocket.
- What conditions should be met for the best result in the landing of the rocket and the safety of the crew.

Section 1

Methodology

Methodology

Executive Summary

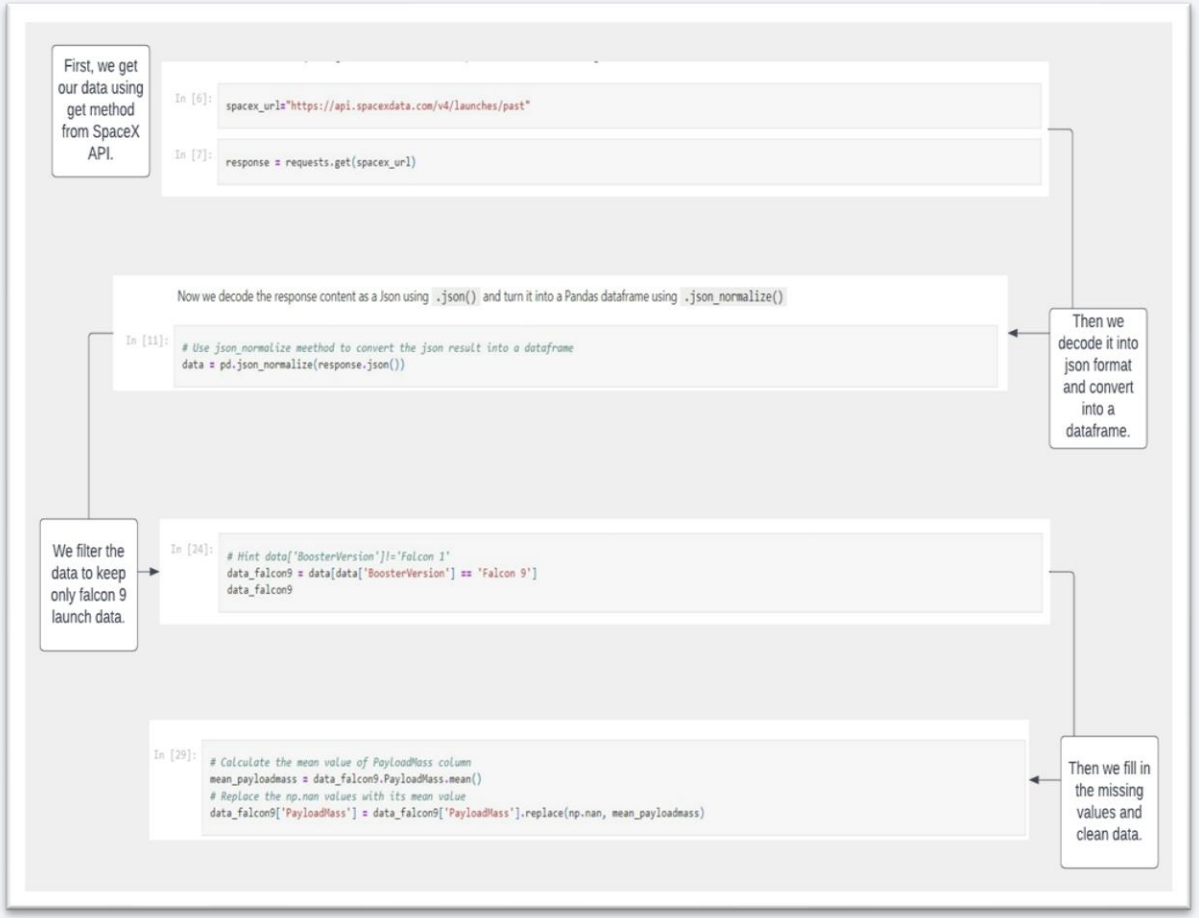
- Data collection methodology:
  - Data was collected using SpaceX API and by web scrapping from Wikipedia.
- Performed data wrangling
  - One-hot encoding was applied on data fields for machine learning and null and irrelevant values and columns were removed.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Logistic Regression, K-Nearest Neighbor (KNN), Decision Tree and Support Vector Machine (SVM) models were built and evaluated to find the best classification model.

Data Collection

- Data was collected using SpaceX API using the get method.
- Then, it was decoded as Json using json() function and turned into a pandas dataframe using json_normalize() function.
- We filtered the data keep only Falcon 9 launch data and also filled the missing values and cleaned the data.
- We also performed web-scrapping to get Falcon 9 launch data, using BeautifulSoup, from Wikipedia.
- We extracted HTML table containing the Falcon 9 data, parsed the table and converted it into a pandas dataframe for analysis.

Data Collection – SpaceX API

- We used the get function to collect data from SpaceX API, convert it into a dataframe, filter it and fill in the missing values.
- https://github.com/MFN-998/IBM_Applied_Data_Science_Capstone/blob/main/(1st%20lab)%20Data%20Collection%20(API).ipynb



Data Collection - Scraping

- We used BeautifulSoup to get data from Wikipedia by web scrapping.
- From all the data of web page, we extracted the required data table and stored the data in our own dictionary.
- Then we used the dictionary, containing the data, to create a dataframe for further use.
- https://github.com/MFN-998/IBM_Applied_Data_Science_Capstone/blob/main/(2nd%20lab)%20Data%20Collection%20with%20web%20scrapping.ipynb

Data Wrangling

- In this section we performed Exploratory Data Analysis on the dataset. We calculated the number of launches at each site, the number and occurrences of every orbit for launch, number and occurrences of mission outcome per orbit. We also created a landing outcome label from outcome column and work out success rate for every landing.
- https://github.com/MFN-998/IBM_Applied_Data_Science_Capstone/blob/main/(3rd%20lab)%20EDA%20lab%20(Data%20Wrangling).ipynb

EDA with Data Visualization

- We plotted the following graphs for EDA:
  - Scatter graph to show relation between following: (Flight Number VS. Payload Mass) (Flight Number VS. Launch Site) (Payload VS. Launch Site) (Orbit VS. Flight Number) (Payload VS. Orbit Type) (Orbit VS. Payload Mass).
  - Bar graph to show the success rate of each orbit.
  - Line graph to show the average launch success trend.
- https://github.com/MFN-998/IBM_Applied_Data_Science_Capstone/blob/main/(5th%20lab)%20EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- We performed SQL queries to get the following information from the dataset:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

EDA with SQL

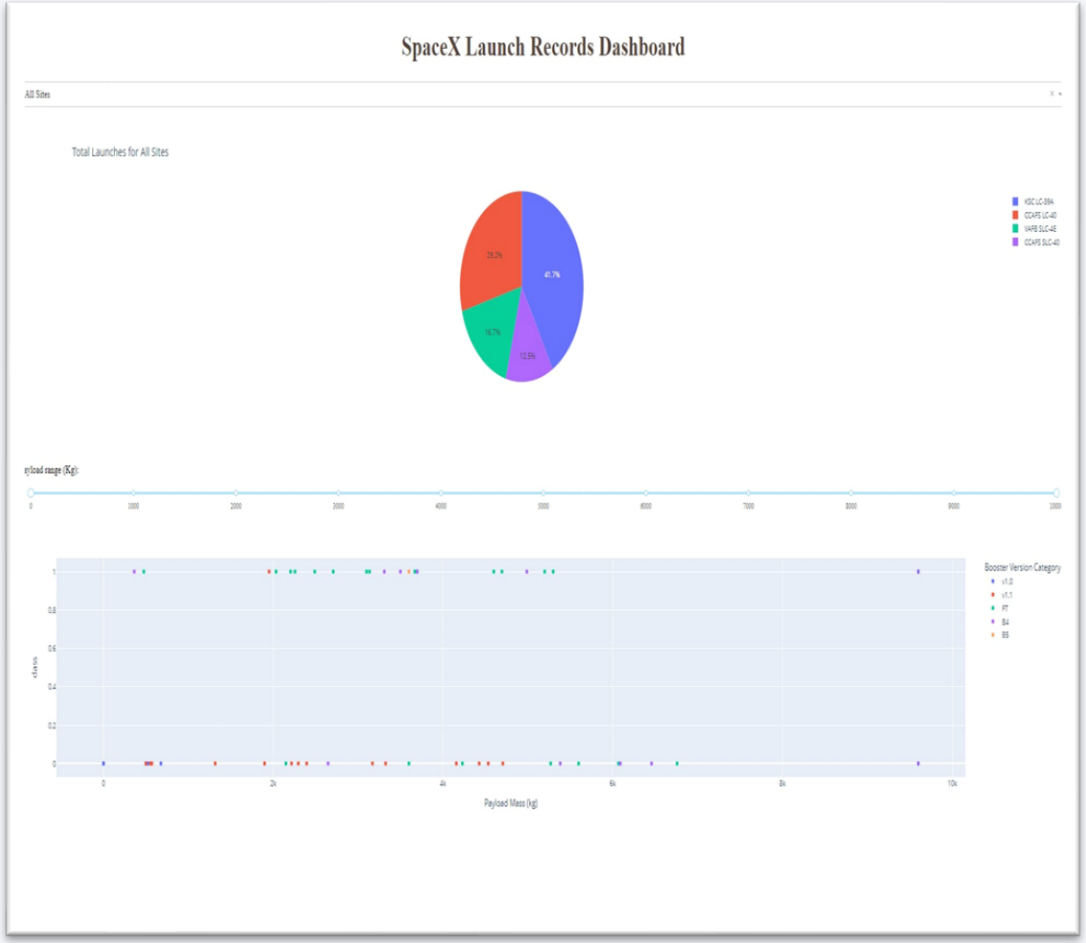
- https://github.com/MFN-998/IBM_Applied_Data_Science_Capstone/blob/main/(4th%20lab)%20EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

- We built an interactive map to visualize the launch data. We took longitude and latitude of each launch site and added a circle marker around the launch site with a label of the launch site name. We then assigned the outcomes of launches with 0 and 1 classes in a class column, along with green and red markers on map, for successful and failed launches respectively.
- We also measured the distance of launch site to various landmarks like coastline, city, railway, highway etc. to find what is around a launch site. We draw lines on the map to show the distance between the launch site and these landmarks.
- https://github.com/MFN-998/IBM_Applied_Data_Science_Capstone/blob/main/(6th%20lab)%20Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

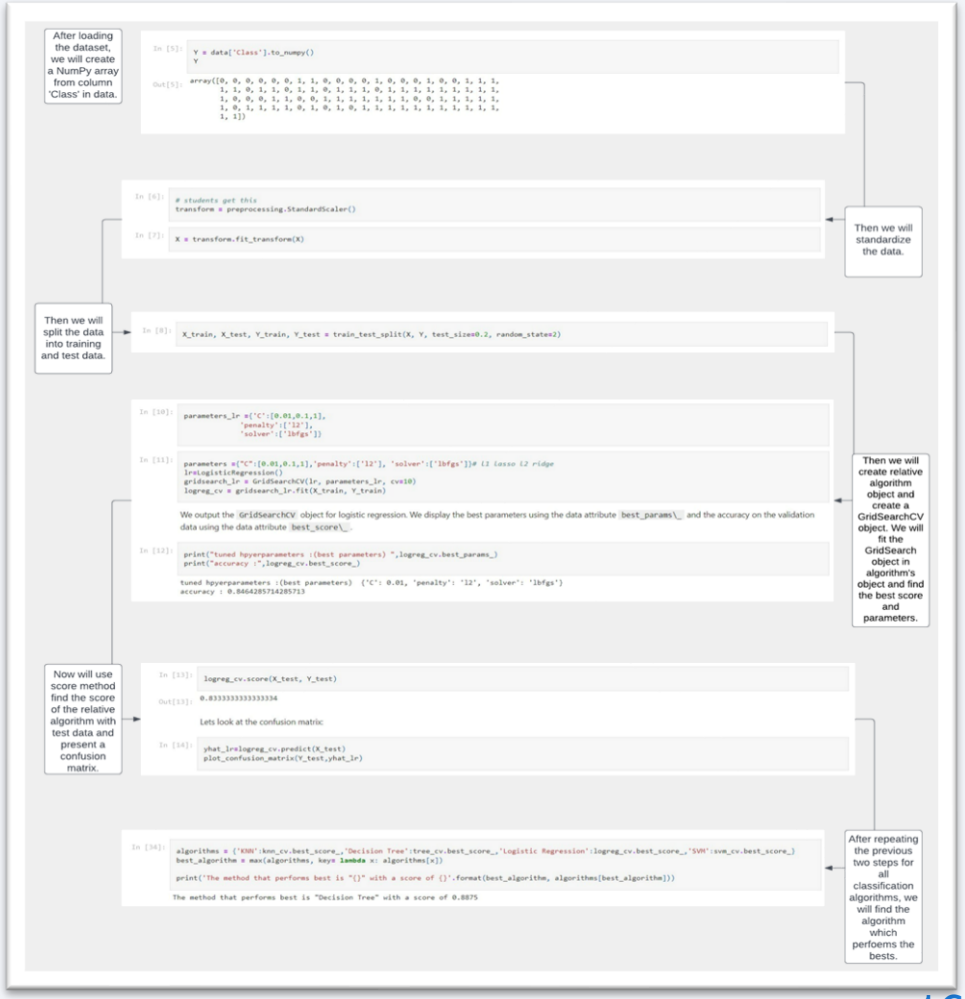
Build a Dashboard with Plotly Dash

- We built a interactive dashboard using plotly and dash. Following graphs were added to the dashboard to visualize data as stated below:
  - ➢ Pie chart: It is used to show total number of launches by all sites or by a certain site.
  - ➢ Scatter graph: This graph shows relation between different Payload Mass (Kg) and Outcome of launch for different booster versions.
- https://github.com/MFN-998/IBM_Applied_Data_Science_Capstone/blob/main/(7th%20lab)%20An%20Interactive%20Dashboard%20with%20Ploty%20Dash)spacex_dash_app.py



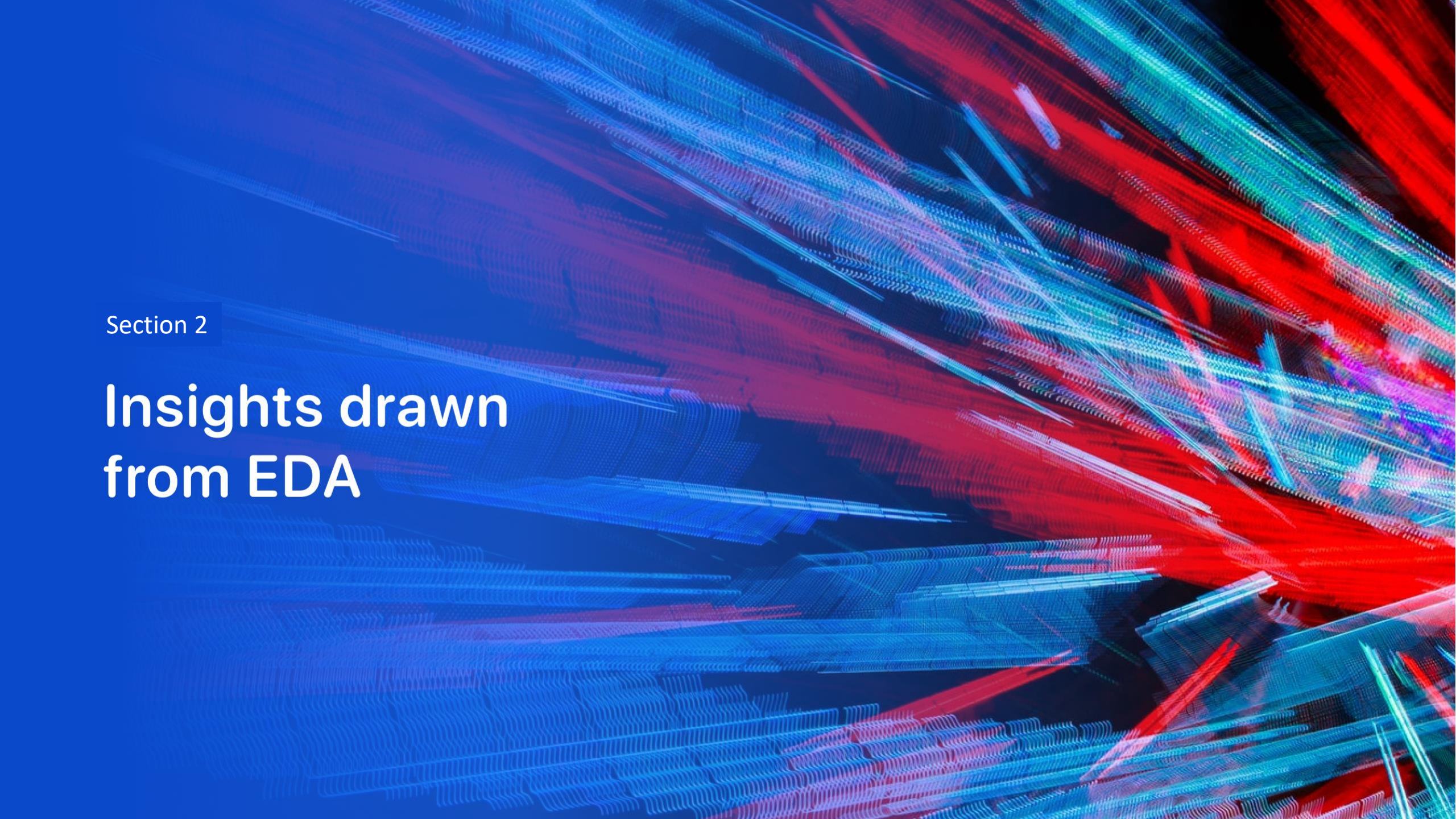
Predictive Analysis (Classification)

- First we loaded our data. Then transformed it and split it into training and test data sets.
- Then we built different machine learning classification (LR, KNN, SVM, DT) models. For each model we set our parameters and algorithms to GridSearchCV and fit our datasets into GridSearchCV object and train out dataset.
- We checked different parameters for each model, plotted confusion matrix for every model and found best performing model using best_score_ method to find their accuracy.
- https://github.com/MFN-998/IBM_Applied_Data_Science_Capstone/blob/main/(8th%20lab)Machine%20Learning%20Prediction%20lab.ipynb



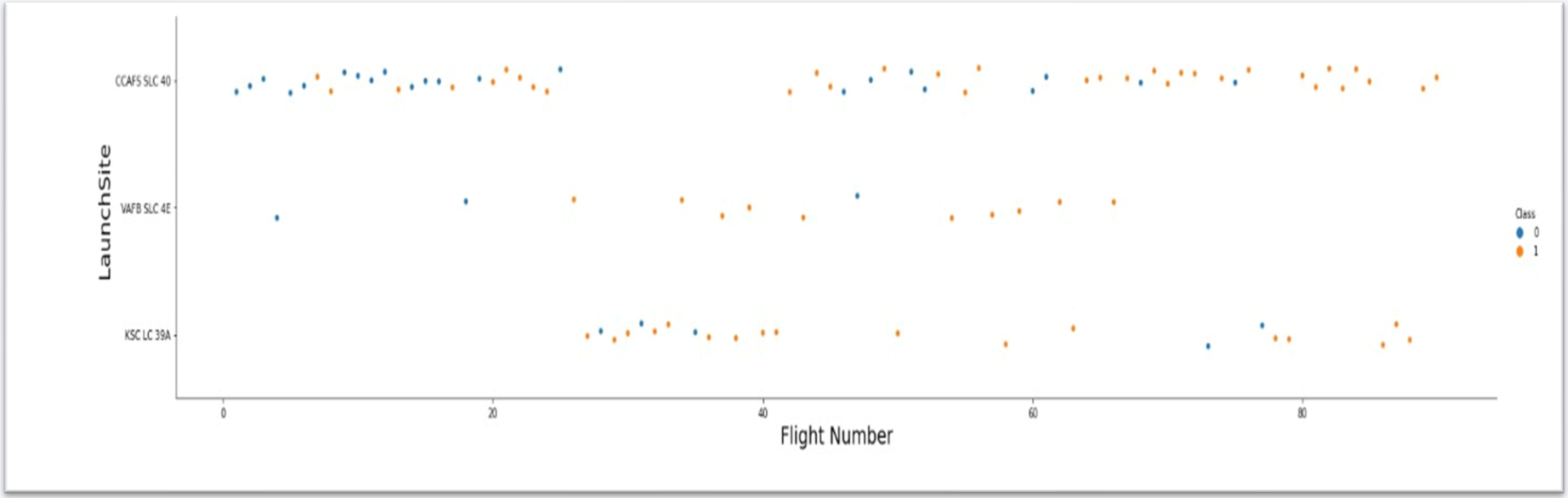
Results

- Decision Tree gives the highest score among all classification models.


Section 2
Insights drawn from EDA

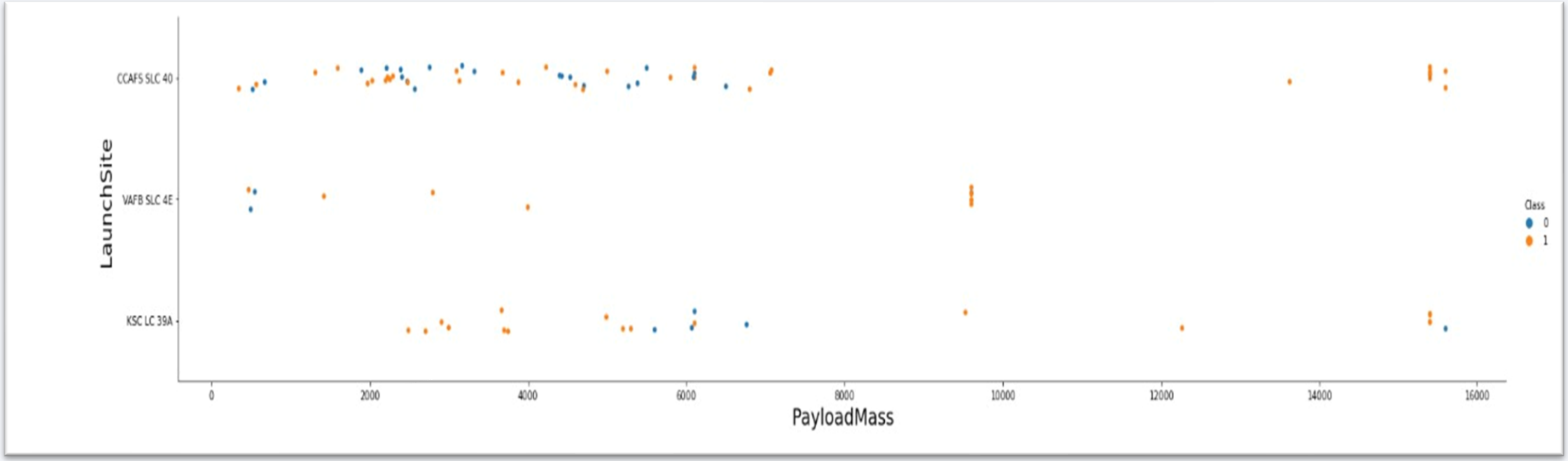
Flight Number vs. Launch Site

- This shows that as the number of flights increases at a launch site the success rate increases as well for that site.



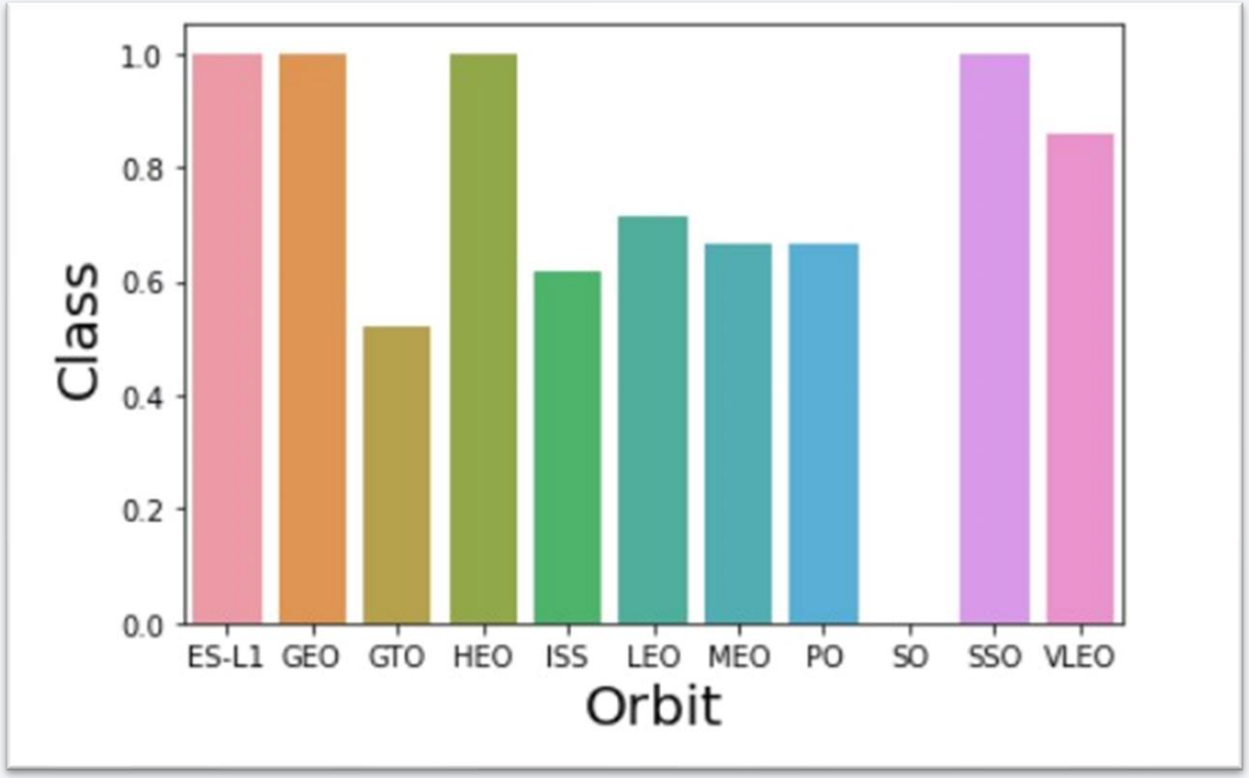
Payload vs. Launch Site

- As the payload mass increases the success rate is also increasing for CCAFS SLC 40. But there is no clear pattern to show that launch site is dependent on payload mass for successful launch



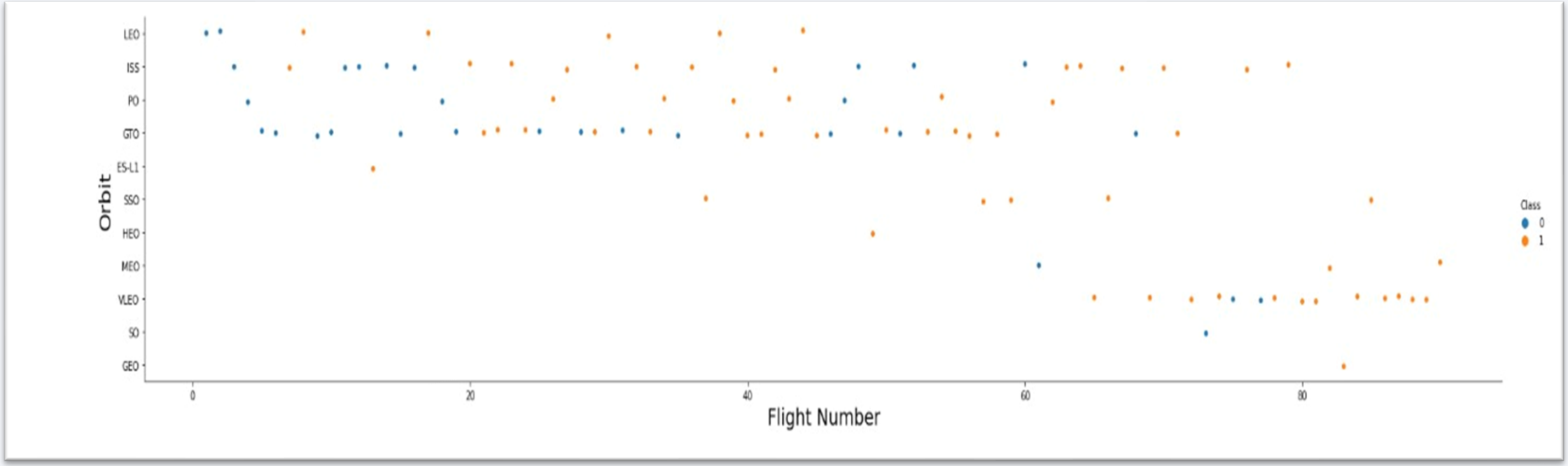
Success Rate vs. Orbit Type

- From the bar graph we can see that the orbits ES-L1, GEO, HEO, SSO and VLEO has more successful rate than other orbits.



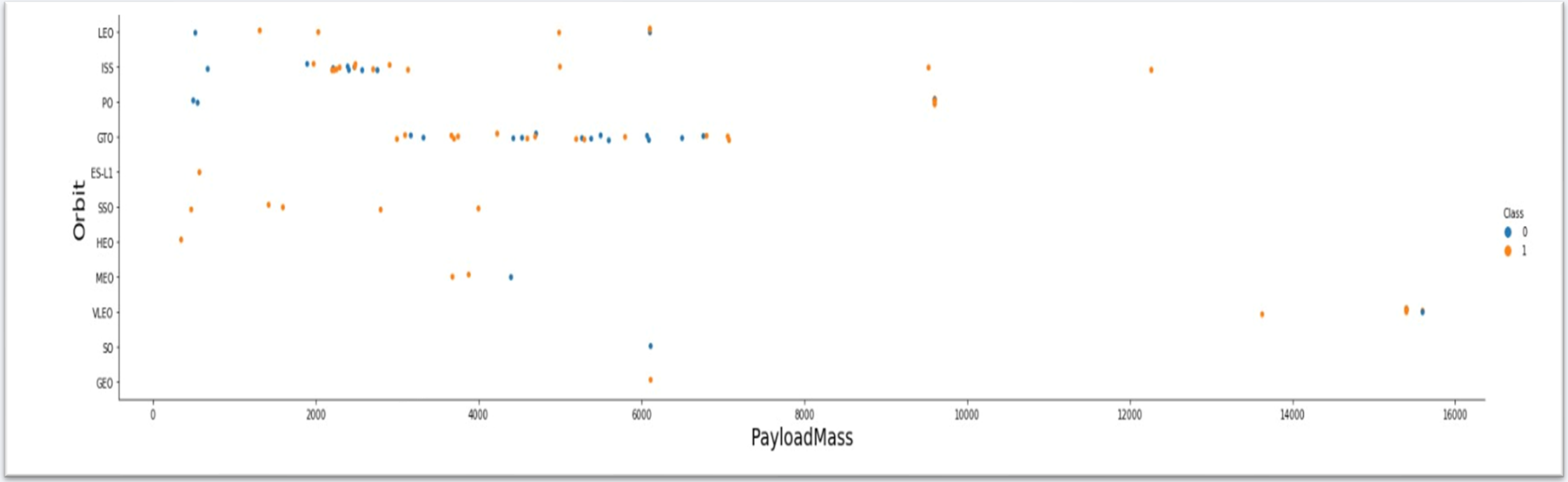
Flight Number vs. Orbit Type

- It is difficult to say anything with surety, but we can see that there is no actual relation between flight number and GTO. Also, as the number of flights increases in LEO orbit successful flights also increase.



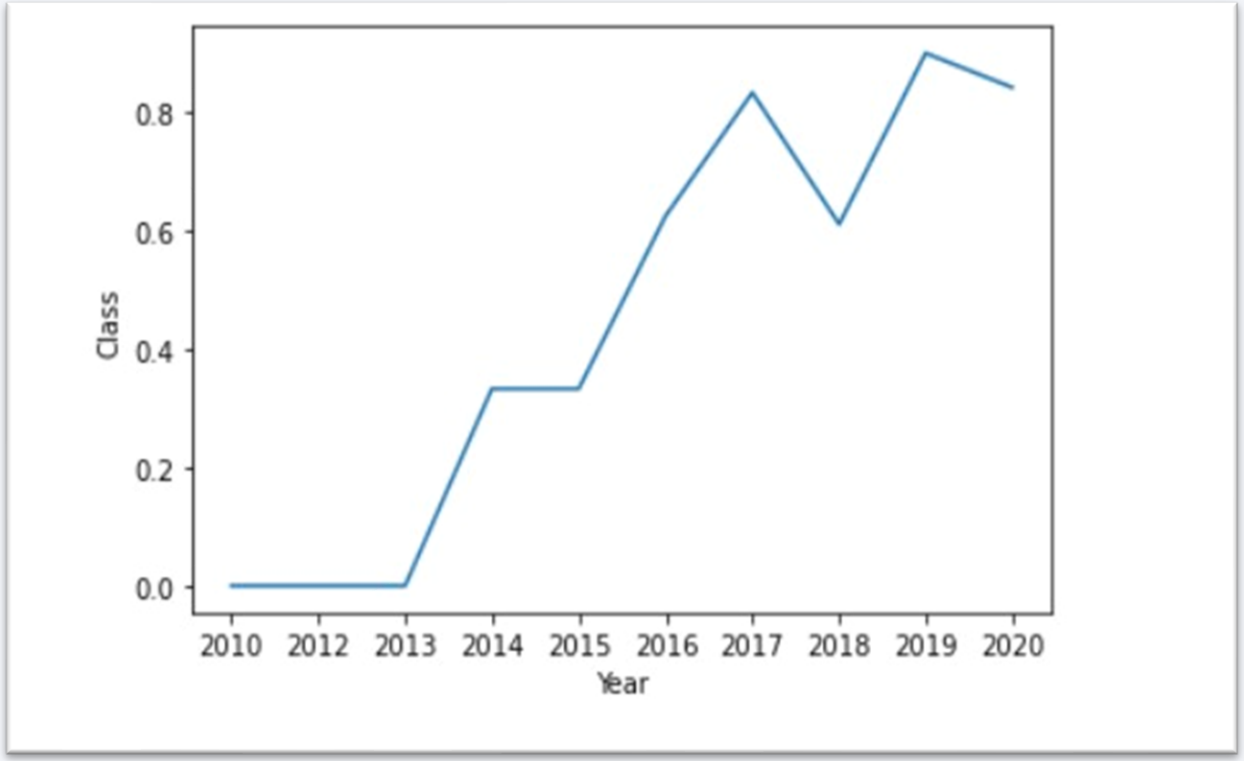
Payload vs. Orbit Type

- 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 both there here.



Launch Success Yearly Trend

- We can see that from year 2013 the success rate increases.



All Launch Site Names

- Names of the unique launch sites

```
%sql SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXDATASET;
```

Done.

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

Launch Site Names Begin with 'CCA'

- Five records where launch sites begin with `CCA`.

```
%sql SELECT * FROM SPACEXDATASET WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

- Total payload carried by boosters from NASA (CRS).

```
%sql SELECT sum(PAYLOAD_MASS__KG_) as TotalPayloadMass__NASA_CRS FROM SPACEXDATASET WHERE CUSTOMER = 'NASA (CRS)';
```

Done.

totalpayloadmass__nasa_crs
45596

Average Payload Mass by F9 v1.1

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

```
%sql SELECT avg(PAYLOAD_MASS__KG_) as Average_PayloadMass__F9_V1 FROM SPACEXDATASET WHERE BOOSTER_VERSION = 'F9 v1.1';
```

Done.

average_payloadmass__f9_v1
2928

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad

```
%sql SELECT min(DATE) FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

```
%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXDATASET WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone ship)';
```

Done.

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

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes.

```
%sql SELECT MISSION_OUTCOME, COUNT(*) FROM SPACEXDATASET GROUP BY MISSION_OUTCOME;
```

Done.

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass.

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEXDATASET WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

Done.

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

```
%sql SELECT DATE, LAUNCH_SITE, BOOSTER_VERSION FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE LIKE '2015%';
```

Done.

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

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

- Ranked 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(*) as quantity FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER
```

Done.

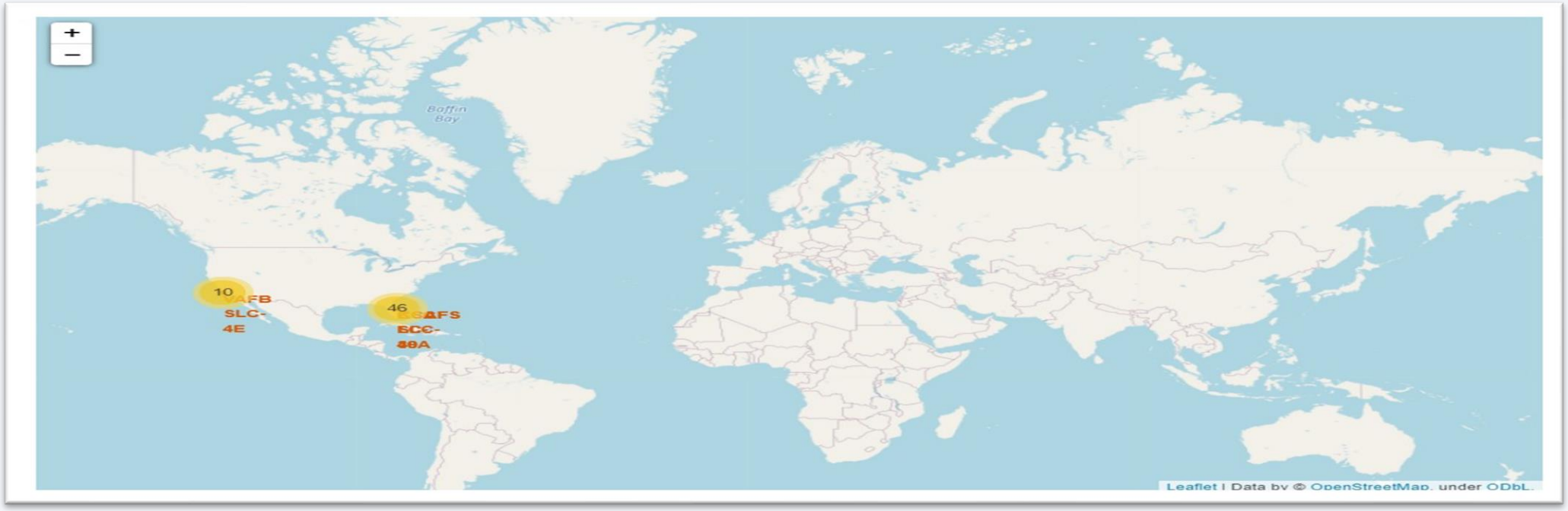
landing__outcome	quantity
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

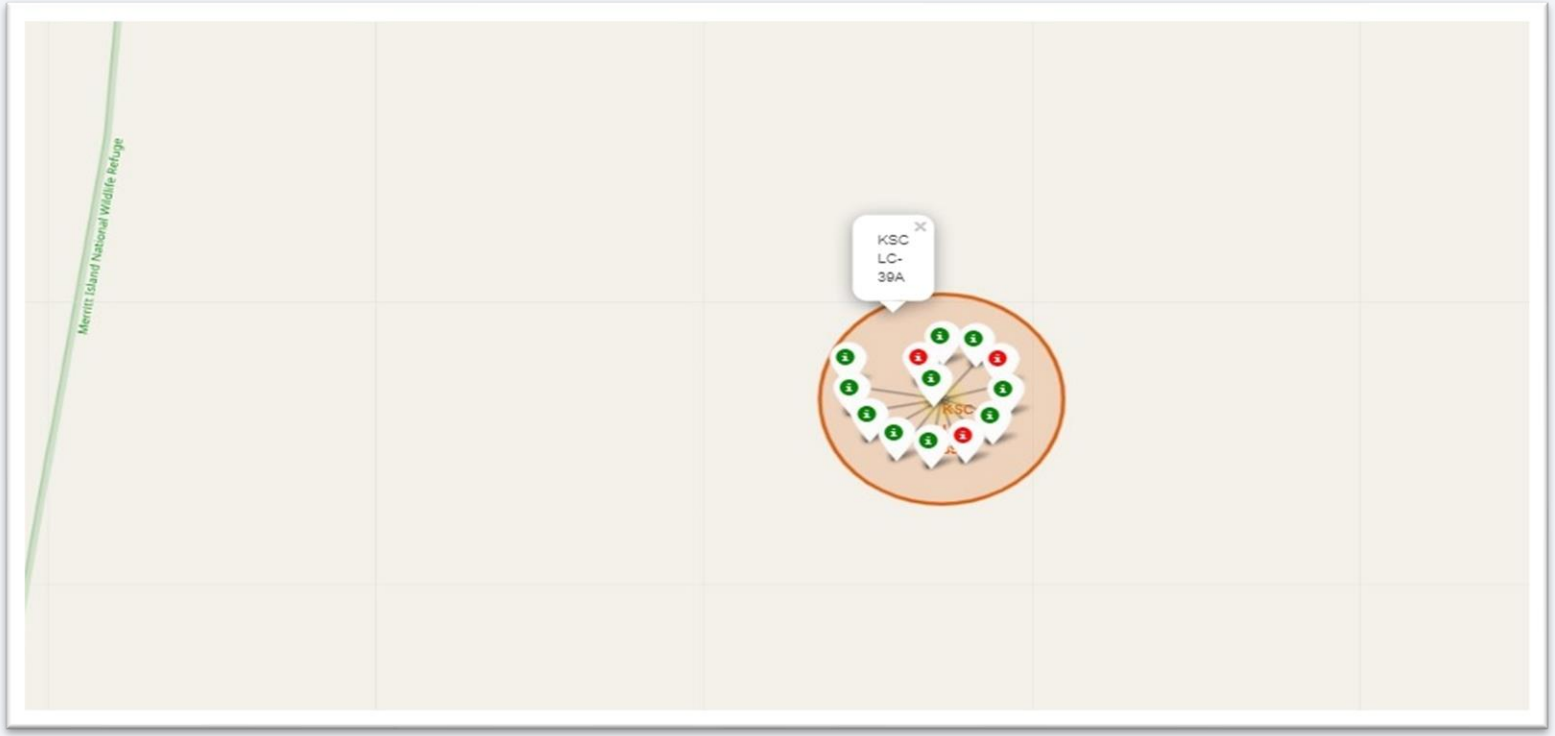
Launch Site Locations on Map

- The map shows that the all launch sites are near the coasts of United States of America



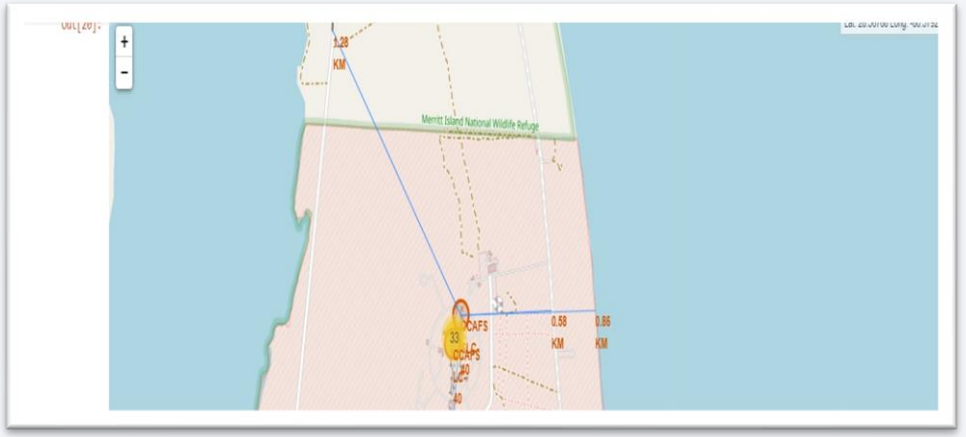
Color-Labeled Markers

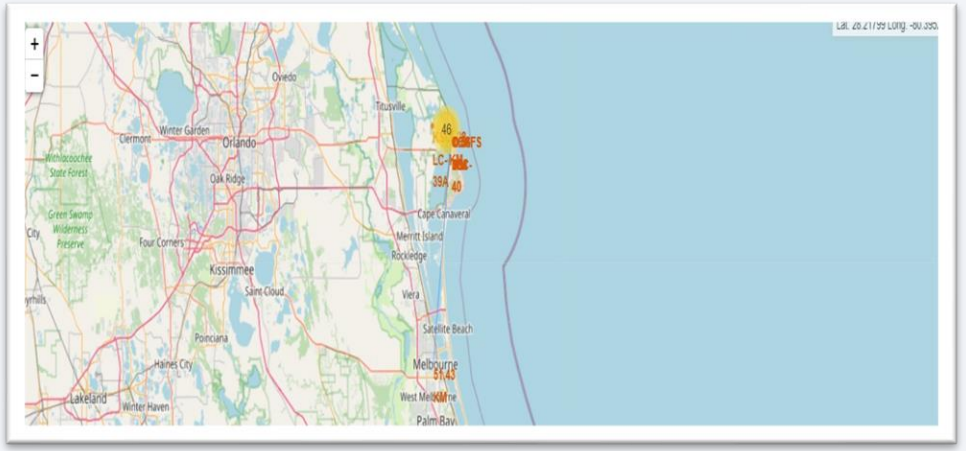
- The picture shows the launch site KSC LC-29A. Here, the green markers represent successful launches and red represent failed launches.

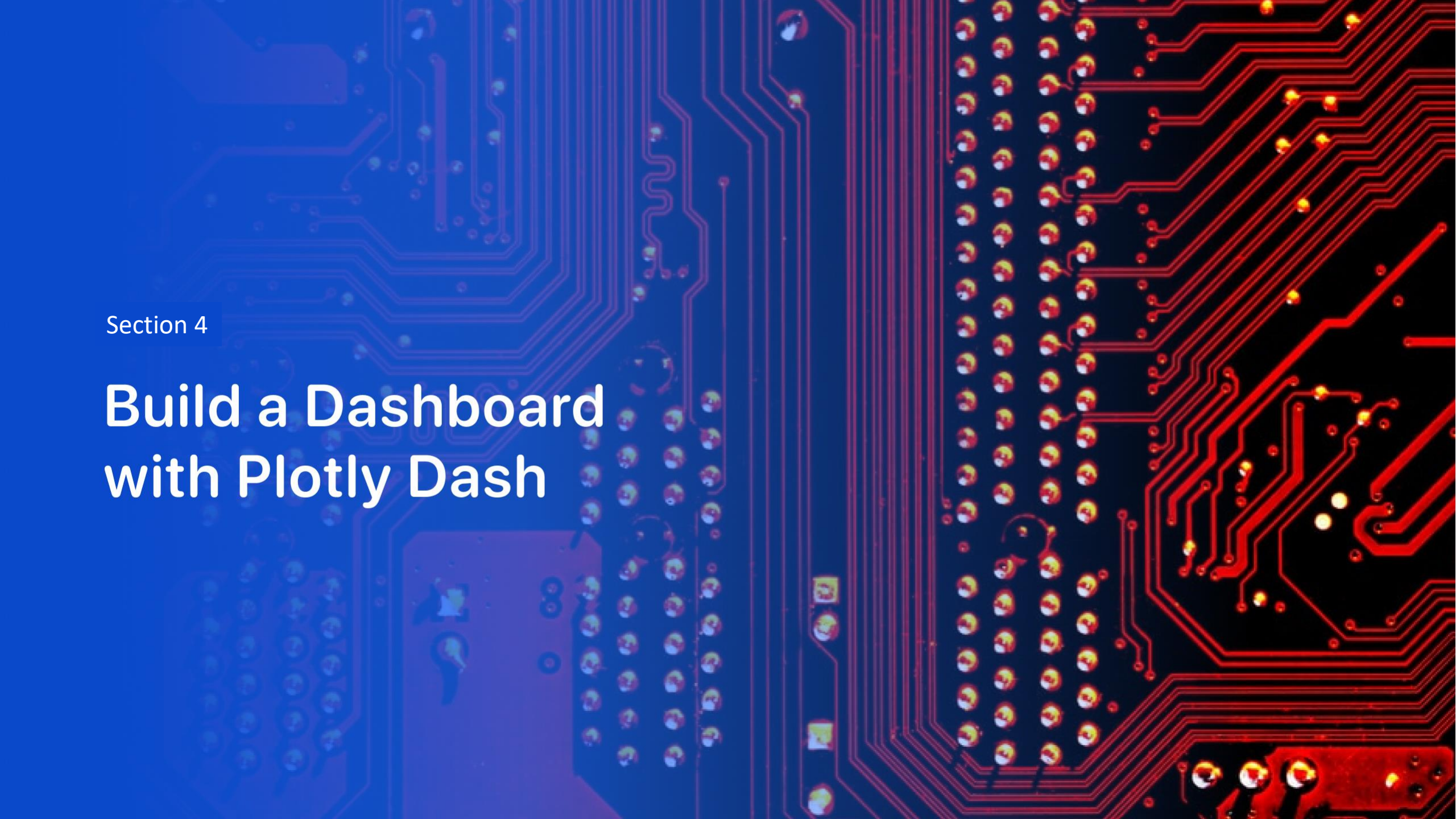


Distance from different landmarks

- In the pictures we measure the distance of landmarks like railway, highway, city etc. from CCAFS SLC-40 launch site.
- The launch site (CCAFS SLC-40) is in close proximity(approx. 1.28 Km) to railways, this helps us in transporting heavy cargo and machinery/equipment to the launch site.
- The launch site (CCAFS SLC-40) is in close proximity(approx. 0.58 Km) to highways, which allows for easy travel for people to and from the launch site.
- The launch site (CCAFS SLC-40) is in close proximity(approx. 0.86 Km) to coastline. This allows crew to abort launch and attempt water landing in case something goes wrong and minimize people and property at risk from falling debris.
- The launch site (CCAFS SLC-40) is at certain distance(approx. 51.43 Km) from the Melbourne City, which minimizes danger to population in the city area.

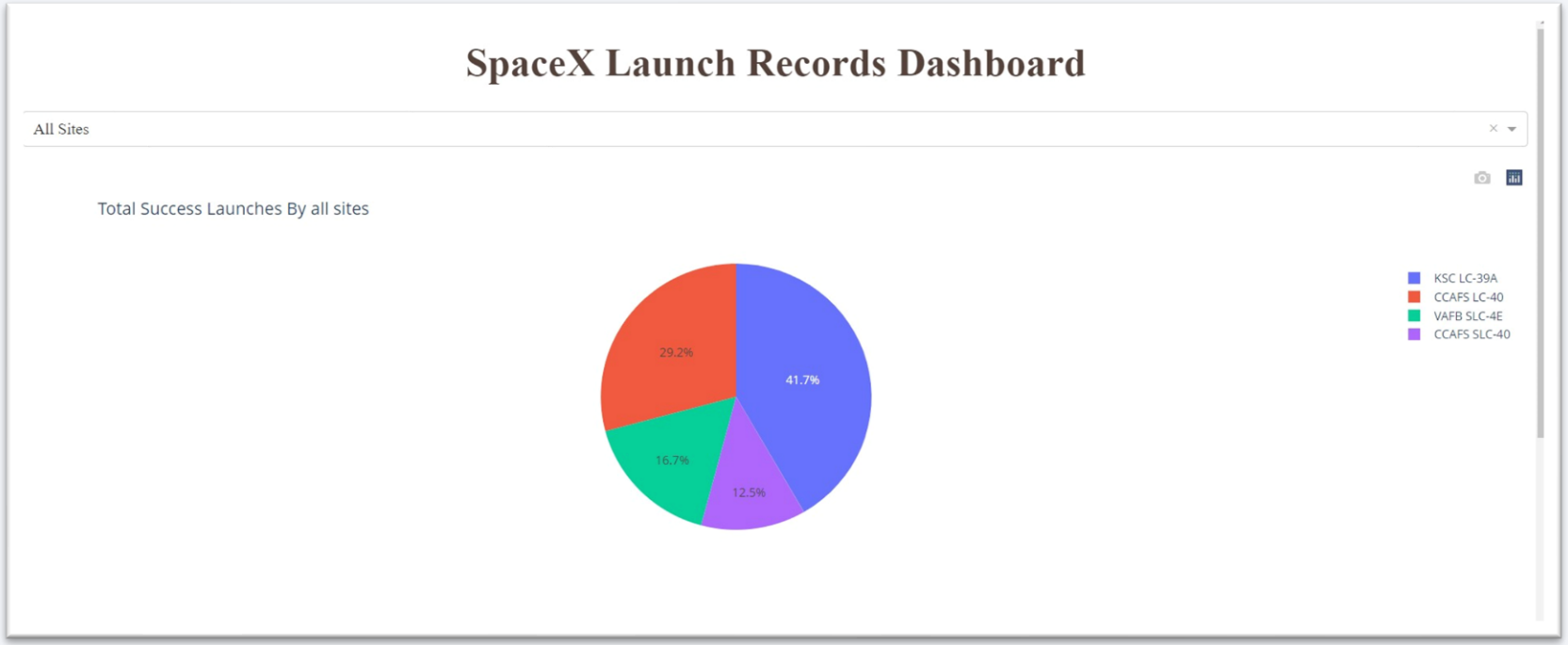





Section 4
Build a Dashboard with Plotly Dash

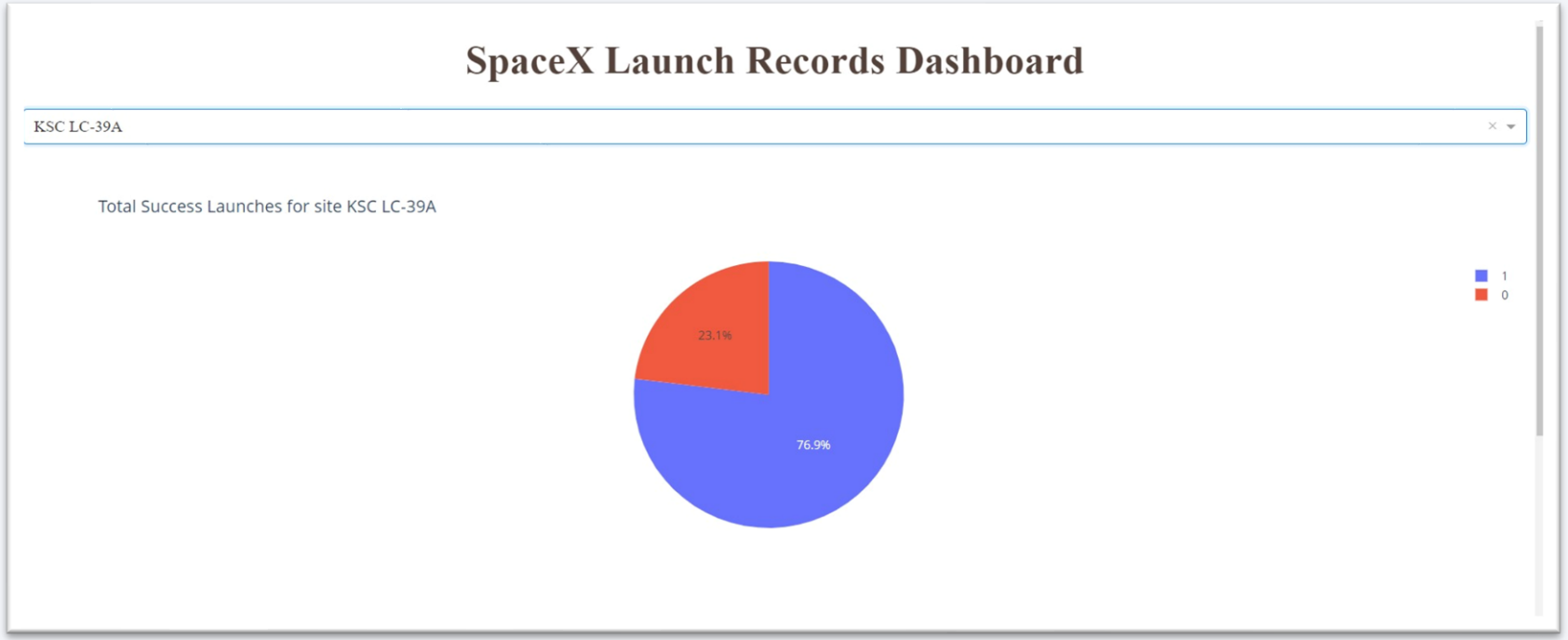
Successful Launch of All Sites

- The pie chart shows the percentage of successful launches of all sites. KSC LC-39A had the most successful launches



Highest Success Ratio Site

- The pie chart shows the success to failure ratio of KSC LC-39A launch site which is 76.9% to 23.1% respectively.



Payload vs Launch Outcome Scatter Plot with Payload Mass Slider

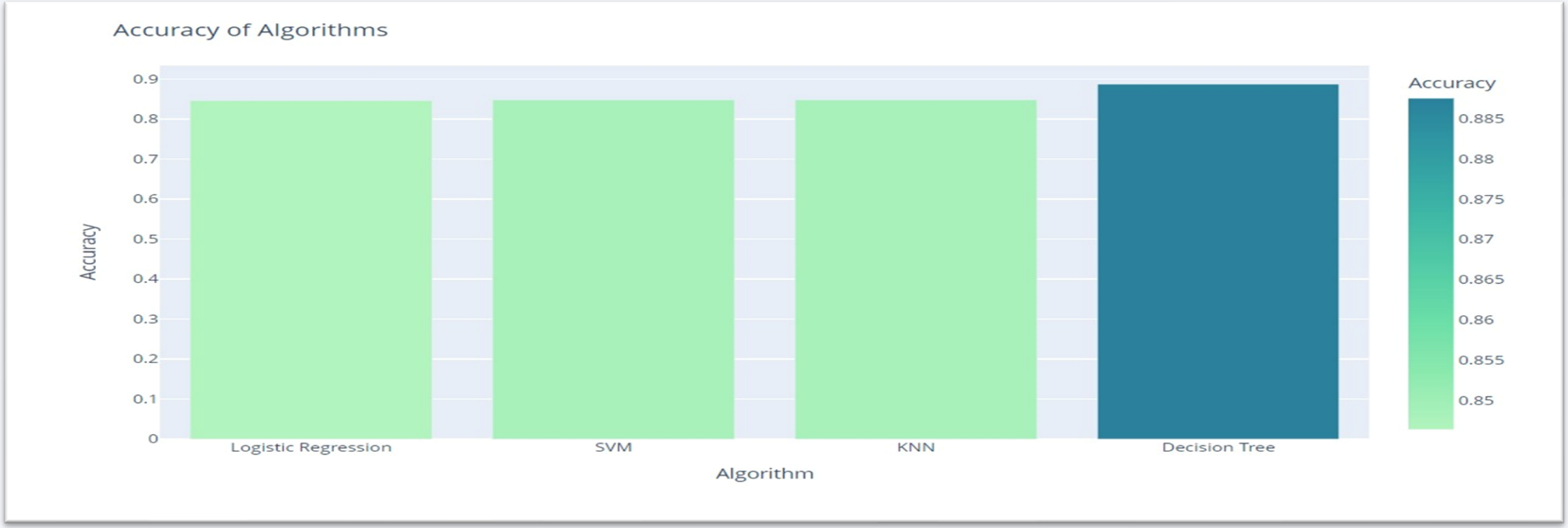
- The following screenshots show outcome of different payload masses (kg).
- In the first screenshot (Top to Bottom) payload mass is 1000 kg, in second it is 4000 kg and in last it is 6000 kg.
- We can see that as the payload mass increases the success rate decreases.

Section 5

Predictive Analysis (Classification)

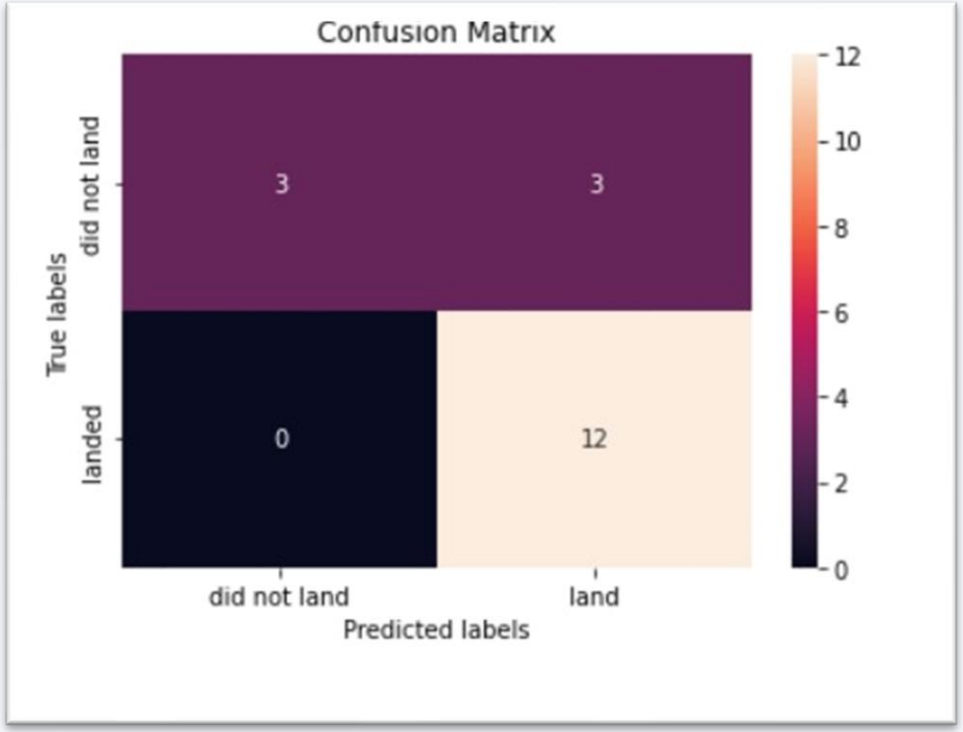
Classification Accuracy

- The following bar chart shows that decision tree has the highest accuracy out of all four.



Confusion Matrix

- After seeing the confusion matrix of decision tree, we see that it can distinguish between different classes, except false positive are a problem (Unsuccessful landing marked as successful landing by decision tree.)



Conclusions

- After going through different methodologies, we have following conclusion:
  - ➢ Decision Tree is the best Machine Learning Classification Model for this project.
  - ➢ As the payload mass increases the success rate decreases.
  - ➢ The KSC LC-39A launch site has the highest number of successful launches.
  - ➢ GEO, HEO, SSO, ES-L1 orbits has the highest success rate.
  - ➢ Launch sites need to be away from cites and near coast lines for safe launch environment.

Appendix

- GitHub link for all code files: https://github.com/MFN-998/IBM_Applied_Data_Science_Capstone

Thank you!