IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lim Soon Tat
26 Oct 2023

Outline

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

Executive Summary

The objective of this project is to predict the success of SpaceX first stage landed successfully. The prediction will help in determining the cost of each launch of Falcon 9 rocket by SpaceX. With the result, company Y will make decision whether to compete with SpaceX in any bidding.

- Methodologies summary:
  - Data collection through SpaceX REST API and Wikipedia
  - Created output labels with 1 presenting first stage landed successfully and 0 representing first stage did not land successfully.
  - Perform exploratory data analysis (EDA) using visualization and SQL
  - Perform interactive visual analytics using Folium and Plotly Dash
  - Perform predictive analysis using classification models

- Result summary:

This report included the results of:

  - EDA of visualization and SQL
  - Interactive visuals of geospatial map and dashboard
  - Predictive model results

Introduction

- SpaceX is capable of launching Falcon 9 rocket with relatively lower cost of 62 million dollars compared to industry cost of 162 million dollars. The great saving is due to SpaceX can reuse the first stage of the rocket. This provides SpaceX great advantage against its competitors.

- There are chances where SpaceX first stage of the rocket does not land successfully. The objective here is to predict the success of SpaceX first stage landed successfully, the prediction will help in determining the cost of each launch of Falcon 9 rocket by SpaceX. With the result, company Y will make decision whether to compete with SpaceX in any bidding.

Section 1

Methodology

Methodology

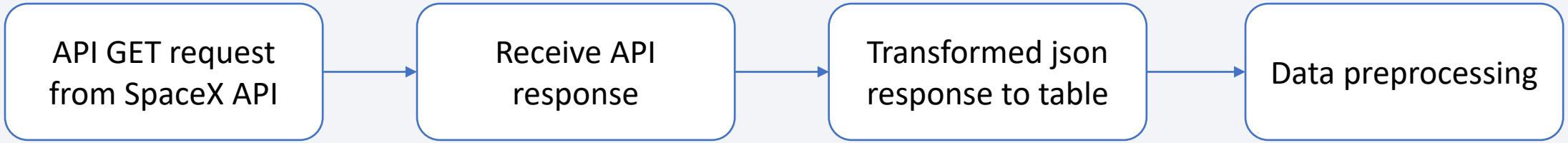
Executive Summary

- Data collection methodology:
  - SpaceX launch data is collected from SpaceX REST API
  - Web scraping Falcon 9 and Falcon Heavy launches records from Wikipedia
- Perform data wrangling
  - Created output labels with 1 presenting first stage landed successfully and 0 representing first stage did not land successfully.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Trained Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors models and perform Grid Search, select best model with best accuracy.

Data Collection – Wikipedia

In this project, we accessed SpaceX's REST API via GET requests to obtain data in JSON format. We transformed this JSON data into a structured table, and filter to focus on Falcon 9 launches only. Subsequently, we handled the missing data to ensure the integrity of analysis.

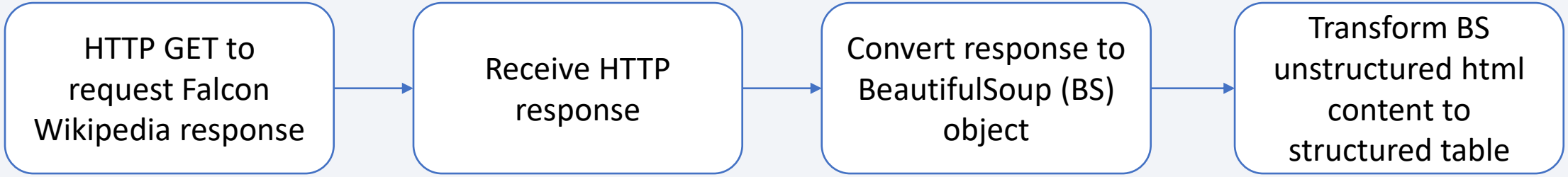
GitHub Reference



Data Collection – Falcon 9 and Falcon Heavy Launches Records from Wikipedia

In this project, Python BeautifulSoup package has been used to web scrape HTML tables from Wikipedia that contain Falcon 9 launch records. The data has been parsed from those tables and converted them Pandas data frame for further visualization and analysis.

GitHub Reference



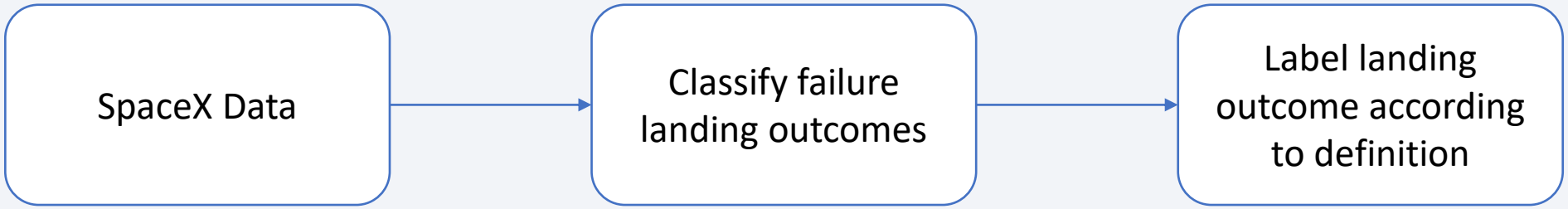
Data Wrangling

SpaceX data contains various mission outcomes, below definitions were used to classify the failure and successful landing:

- True Ocean means the mission outcome was successfully landed to a specific region of the ocean
- False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean
- True RTLS means the mission outcome was successfully landed to a ground pad
- False RTLS means the mission outcome was unsuccessfully landed to a ground pad
- True ASDS means the mission outcome was successfully landed to a drone ship
- False ASDS means the mission outcome was unsuccessfully landed to a drone ship
- None ASDS and None None these represent a failure to land

Based on the definition, created output labels with 1 presenting first stage landed successfully and 0 representing first stage did not land successfully.

GitHub Reference



EDA with Data Visualization

- The following charts were plotted to perform Exploratory Data Analysis (EDA):
    1. Scatter plot:
        - Visualize the relationship between Flight Number and Launch Site
        - Visualize the relationship between Payload and Launch Site
        - Visualize the relationship between FlightNumber and Orbit type
        - Visualize the relationship between Payload and Orbit type
    2. Bar chart
        - Visualize the relationship between success rate of each orbit type
    3. Line chart
        - Visualize the average launch success yearly trend

GitHub Reference

EDA with SQL

Further Exploratory Data Analysis (EDA) was conducted using SQL on SpaceX data:

- 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 succesful landing outcome in ground 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 names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the names of the booster_versions which have carried the maximum payload mass.
- 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.
- 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.

GitHub Reference

Build an Interactive Map with Folium

The utilization of an interactive Folium map enhances the capacity to analyze geospatial data and improve understanding on the impact of location and proximity of launch site on launch success rate.

The below map objects were created and added to the map:

- Folium.Circle and Folium.Marker: Add a highlighted circle area with a text label on a specific coordinate.
- MarkerCluster: Create green marker for success launch red marker for a failed launch for all launch records.
- To calculate distance between launch site and coastline/railroad/highway/city:
    - MousePosition: To get coordinate for a mouse over a point of interests on the map.
    - Folium.Marker: To display distance between launch site and a point on the map in KM
    - Folium.PolyLine: To draw a line between a launch site to the selected point.

GitHub Reference

Build a Dashboard with Plotly Dash

A dashboard has been created using Dash application users to perform interactive visual analytics on SpaceX launch data in real-time. This dashboard application contains

1. Dropdown list: As input component to filter the dashboard by launch site.
2. Pie chart: To display total success launch rate for launch site.
3. Range slider: As input component to filter scatter plot by payload mass (KG) range.
4. Scatter plot: To display correlation of payload mass (KG) correlated with mission outcomes for selected launch site.

The dashboard has helped to answer:

- Which site has the largest successful launches? KSC LC-39A with 10 successful launches
- Which site has the highest launch success rate? KSC LC-39A with 76.9% success rate
- Which payload range(s) has the highest launch success rate? 2,000 – 4,000 KG
- Which payload range(s) has the lowest launch success rate? > 4,000 KG
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate? FT

GitHub Reference

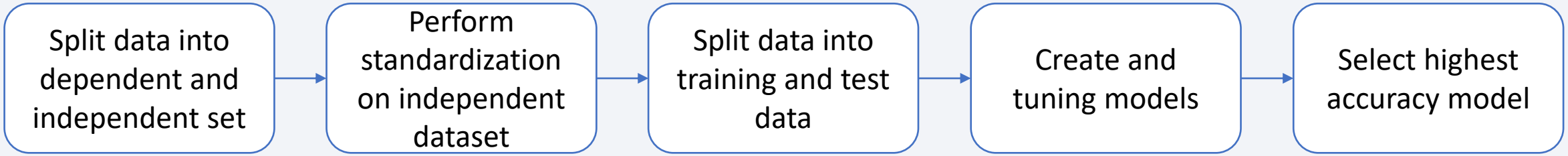
Predictive Analysis (Classification)

The data has been separated into 2 datasets:

1. Dataset with dependent variable - Y
2. Dataset with independent variables - X

The data within dataset X has undergone standardization process. After that, both X and Y have been split into training and test data, subsequently trained 4 models: Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors models and perform Grid Search for model tuning. Finally, select best model with best accuracy.

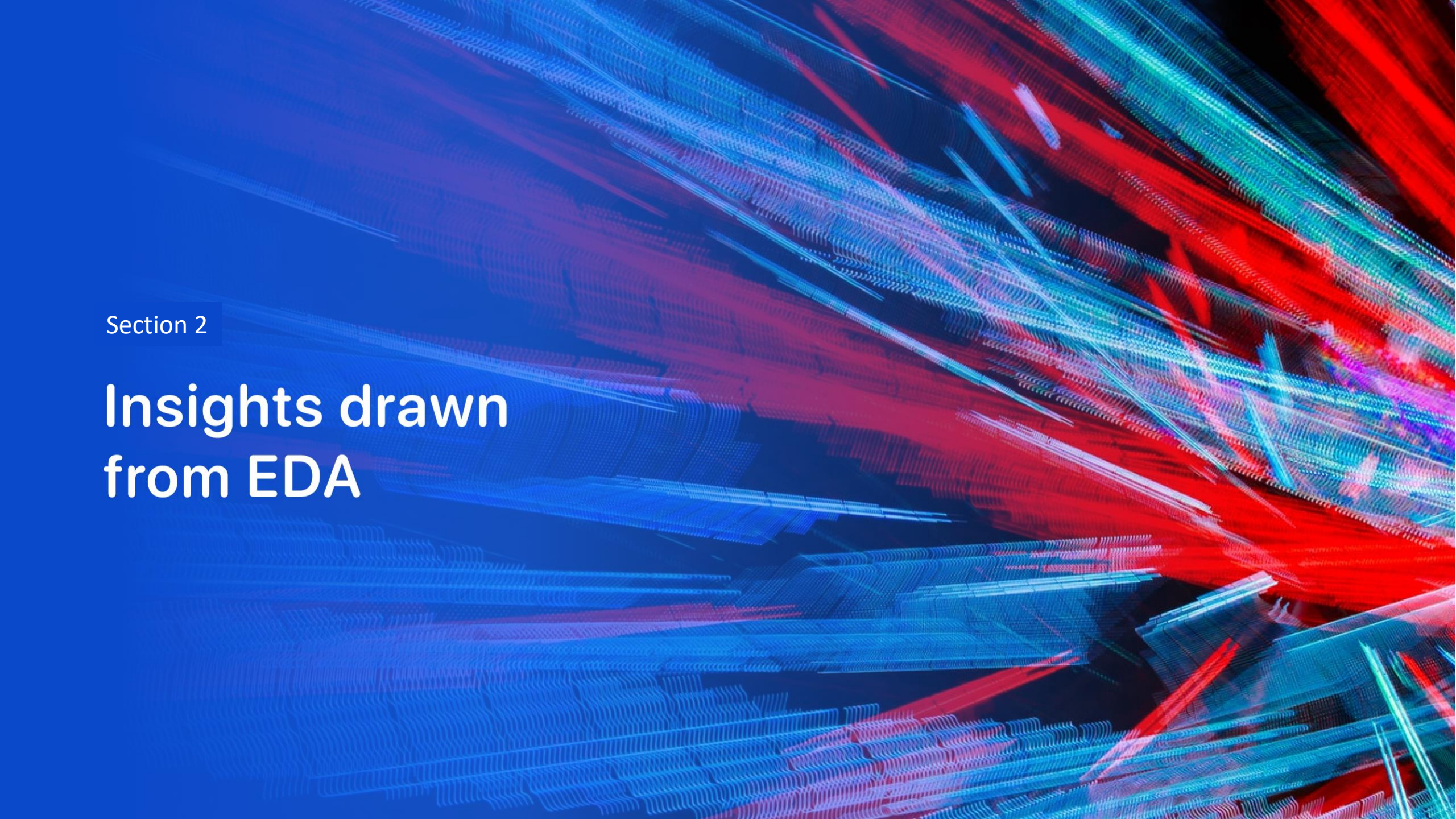
GitHub Reference



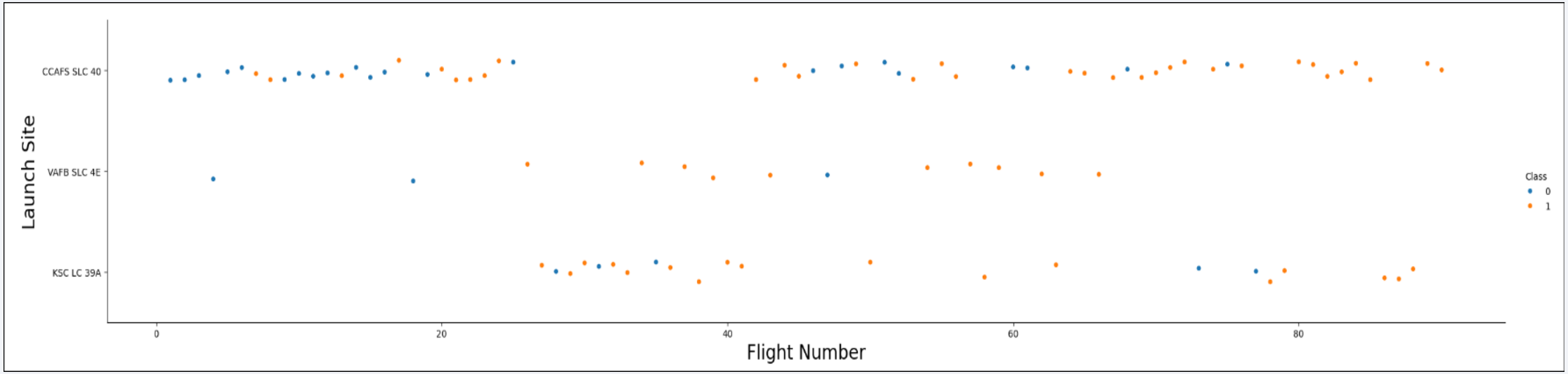
Results

This section explains the results of:

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

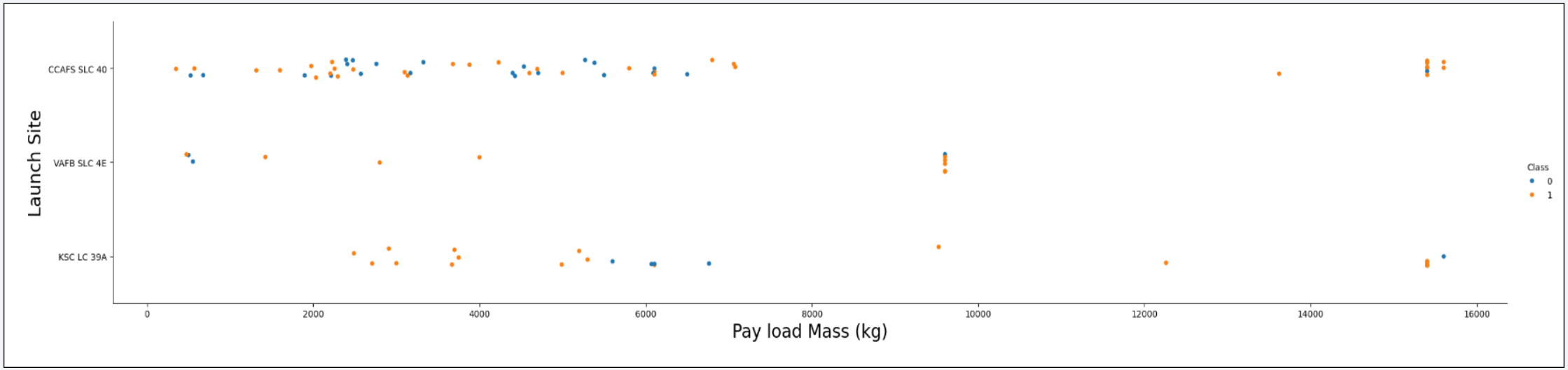

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



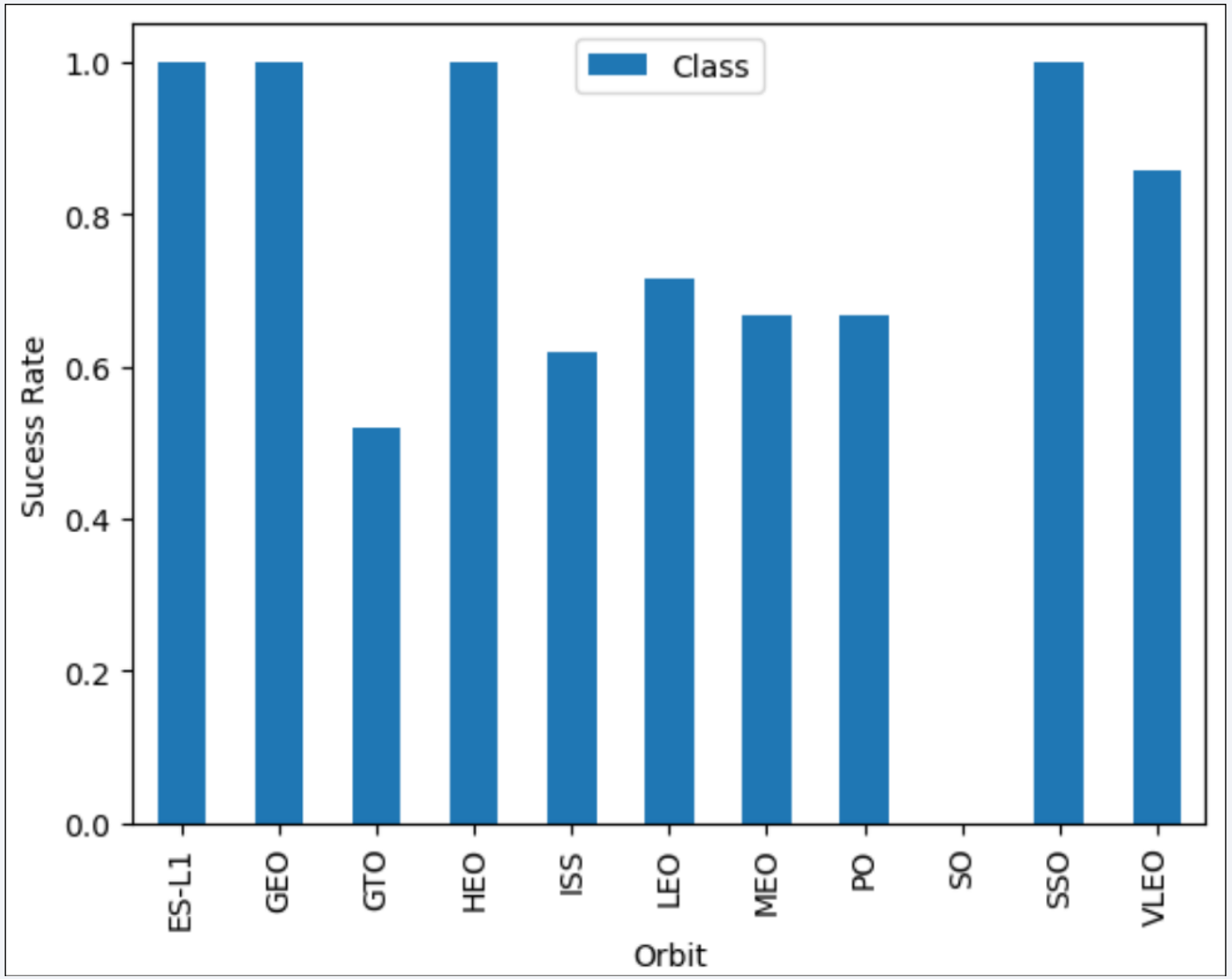
Based on observations from the scatter plot, the chance of success launch increased as the number of flight increased.

Payload vs. Launch Site



Based on observations from the Payload Vs. Launch Site scatter point chart, VAFB-SLC 4E launch site has no rockets launched for heavy payload mass(greater than 10,000).

Success Rate vs. Orbit Type

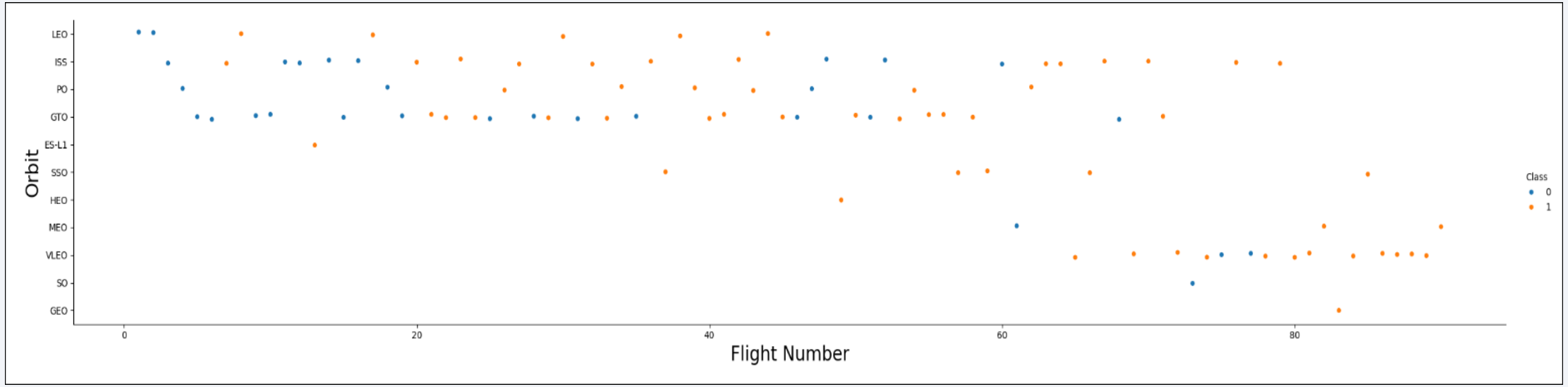


Based on the observations from the bar chart, there are 4 orbit have highest success rate:

- ES-L1
- GEO
- HEO
- SSO

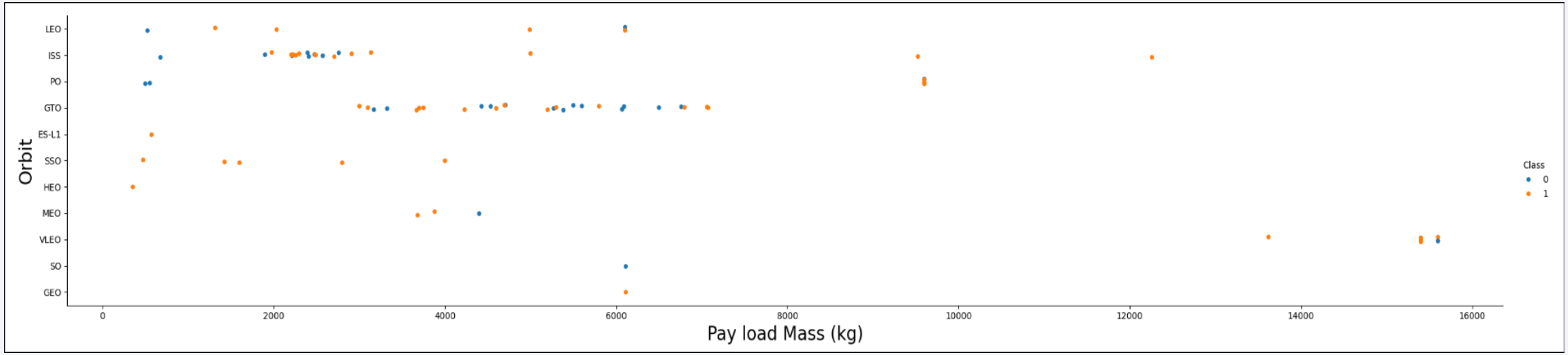
SO orbit has the lowest success rate.

Flight Number vs. Orbit Type



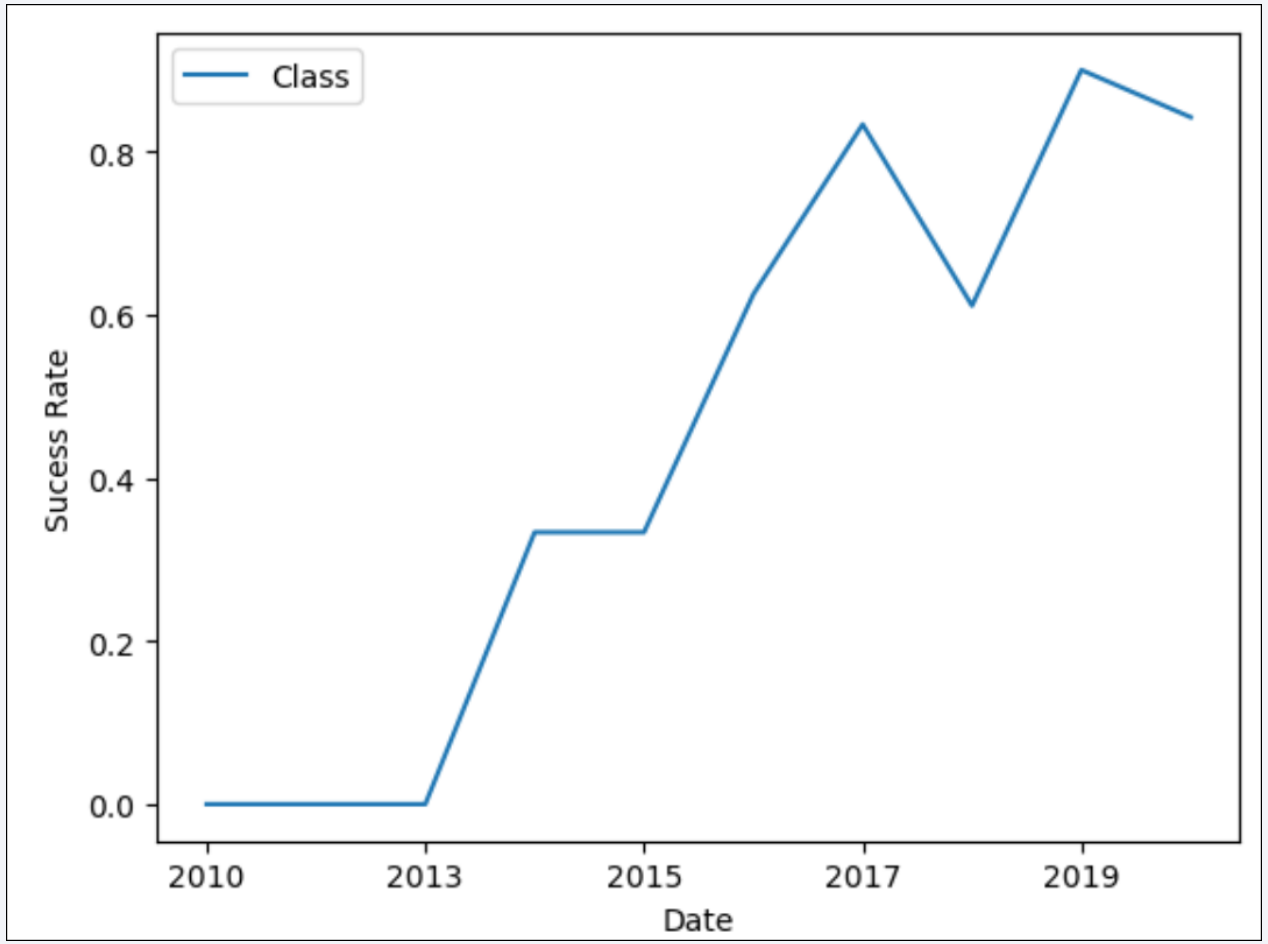
Based on observations from the above scatter point chart, LEO orbit the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



Based on observations from the above scatter point chart, with heavy payloads the successful landing or positive landing rate are more for Polar, VLEO, 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



Based on the observations from the line chart, the success rate increase since 2013 kept increasing till 2020 despite a drop in year 2018.

All Launch Site Names

- Query

```
%sql select distinct Launch_Site from SPACEXTABLE
```

- Description

Select unique launch sites in the space mission using DISTINCT statement. 4 records returned.

- Result:

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

Launch Site Names Begin with 'CCA'

- Query

```
%sql select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5
```

- Description

Select records where launch sites begin with the string 'CCA' using LIKE statement and restrict result to top 5 records using LIMIT statement.

- Result

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Query

```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTABLE where Customer = "NASA (CRS)"
```

- Description

Calculate the total payload mass carried by boosters launched by NASA (CRS) with SUM and WHERE statement. The result returned is 45,596 KG.

- Result

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Query

```
%sql select avg(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTABLE where Booster_Version = "F9 v1.1"
```

- Description

Calculate the average payload mass carried by booster version F9 v1.1 with AVG and where statement.
The result returned is 2,928.4 KG.

- Result

total_payload_mass
2928.4

First Successful Ground Landing Date

- Query

```
%sql select min(date) from SPACEXTABLE where Landing_Outcome = "Success (ground pad)"
```

- Description

List when the first successful landing outcome in ground pad was achieved with MIN and WHERE statement. The result returned is 22 Dec 2015.

- Result

min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4,000 and 6,000

- Query

```
%sql select Booster_Version from SPACEXTABLE where (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000) and Landing_Outcome = "Success (drone ship)"
```

- Description

List the names of the boosters which have success in drone ship and have payload mass greater than 4,000 but less than 6,000, using WHERE and AND statement. 4 records returned.

- Result

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

Total Number of Successful and Failure Mission Outcomes

- Query

```
%sql select Landing_Outcome, count(*) from SPACEXTABLE group by Landing_Outcome having Landing_Outcome in ("Success", "Failure")
```

- Description

List the total number of successful and failure mission outcomes using COUNT, GROUPBY, and HAVING statements. Result shows 3 failure and 38 success mission outcomes.

- Result

Landing_Outcome	count(*)
Failure	3
Success	38

Boosters Carried Maximum Payload

- Query

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

- Description

List the names of the booster version which have carried the maximum payload mass, using WHERE and MAX statements, also subquery method. 12 records returned.

- Result

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

2015 Launch Records

- Query

```
%sql select substr(Date, 6,2) as month, Landing_Outcome, Booster_Version, launch_site from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome = "Failure (drone ship)"
```

- Description

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. Included SUBSTR, WHERE and AND statements, 2 records returned.

- Result

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

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

- Query

%sql select Landing_Outcome, count(*) as landing_cnt, rank() over (order by count(*) desc) as RANK from SPACEXTABLE where date >= "2010-06-04" and date <= "2017-03-20" group by Landing_Outcome order by landing_cnt desc

- Description

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. Included COUNT, RANK, ORDER BY, WHERE, GROUPBY and AND statements, 2 records returned.

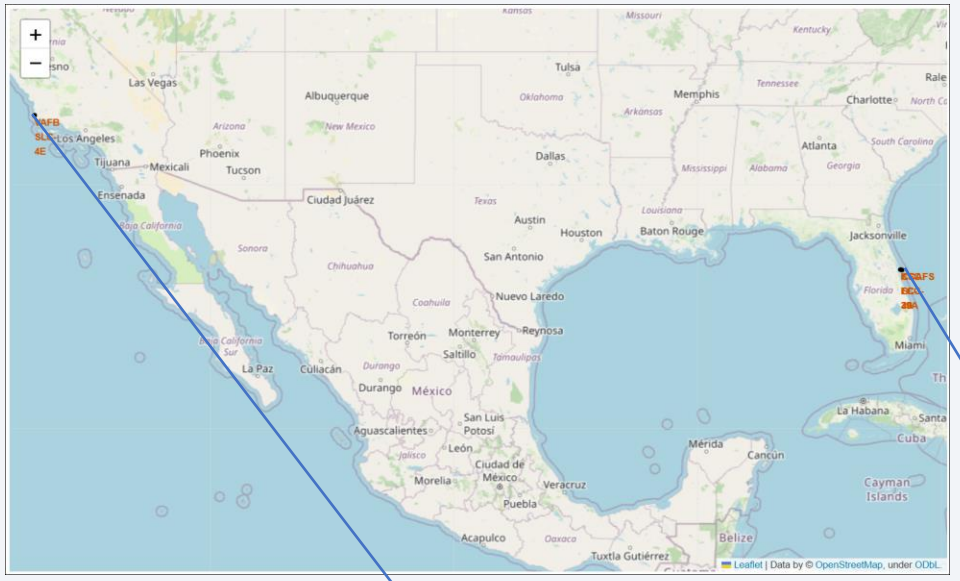
- Result

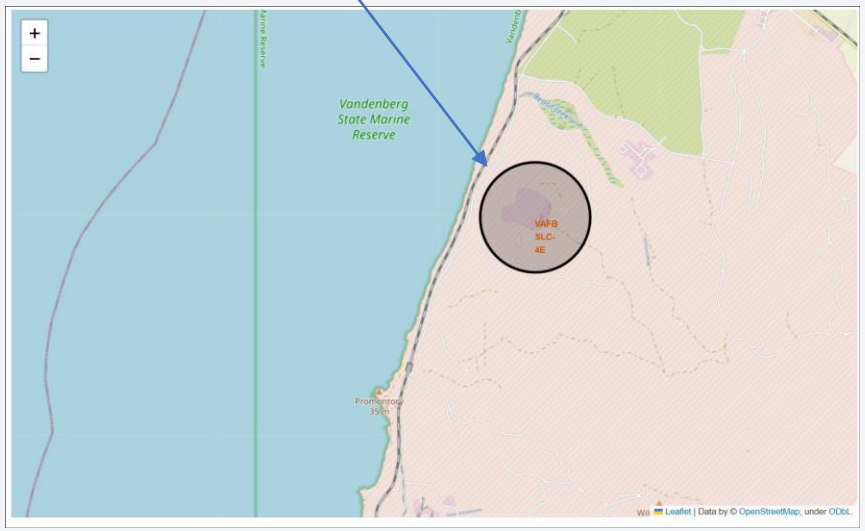
Landing_Outcome	landing_cnt	RANK
No attempt	10	1
Success (ground pad)	5	2
Success (drone ship)	5	2
Failure (drone ship)	5	2
Controlled (ocean)	3	5
Uncontrolled (ocean)	2	6
Precluded (drone ship)	1	7
Failure (parachute)	1	7

Section 3

Launch Sites Proximities Analysis

SpaceX Falcon 9 – Launch Sites Map

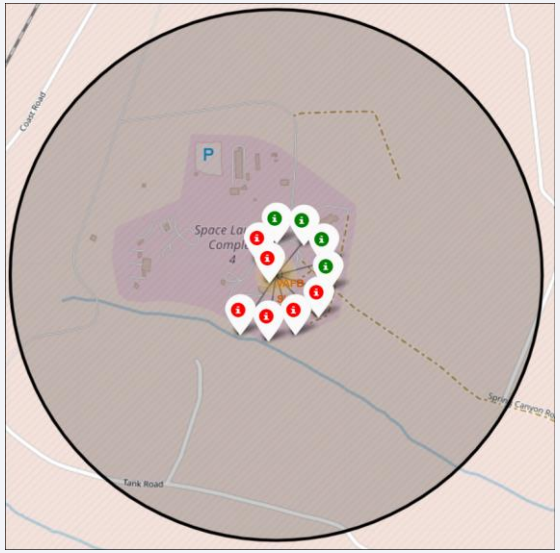




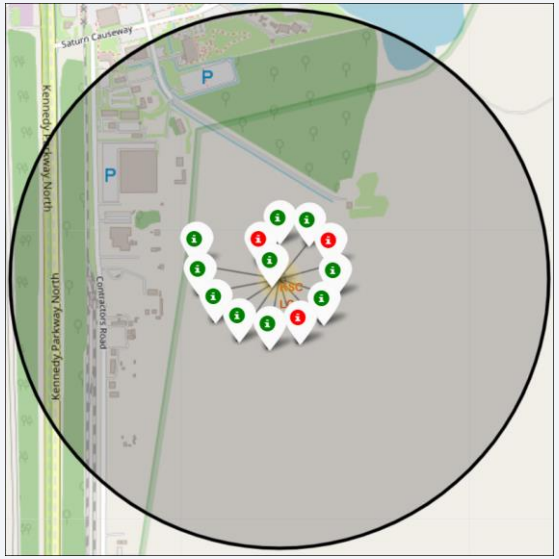
From these 3 Folium maps, we observed that:

1. SpaceX has setup launch sites at 2 main states: California and Florida
2. 3 launch sites in Florida and 1 launch site at California
3. All the sites are near to coast

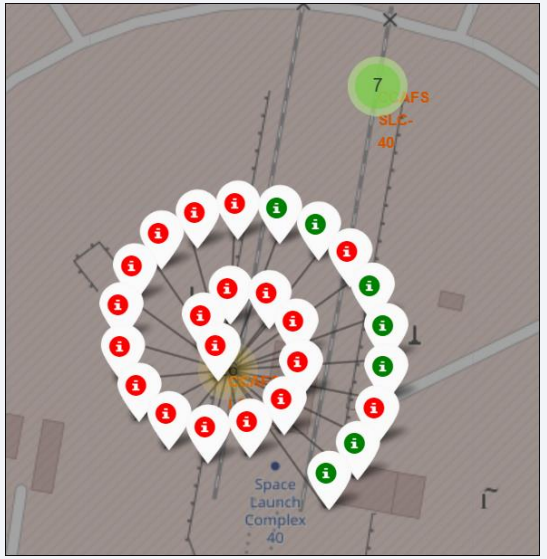
SpaceX Falon9 - All Launches Records



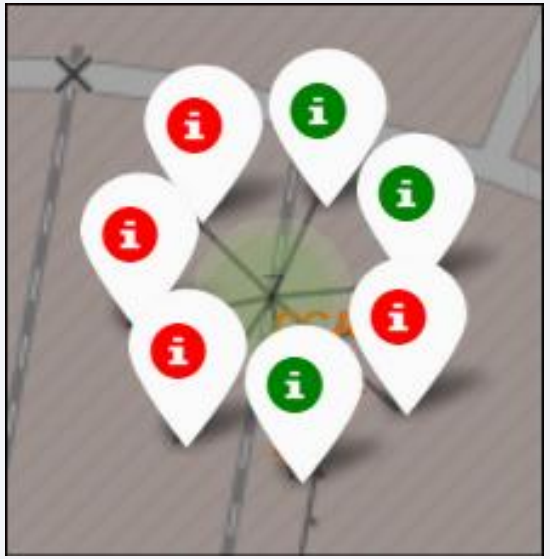
VAFB SLC-4E



KSC LC-39A



CCAFS SLC-40



CCAFS LC-40

Above 4 Folium maps zoom into each launch site and displays all launches with green marker indicating success launch and red marker indicating failed launch.

We observed that KSC LK-39A launch site has highest success launch rate.

SpaceX Falcon9 - Proximity Distance Map

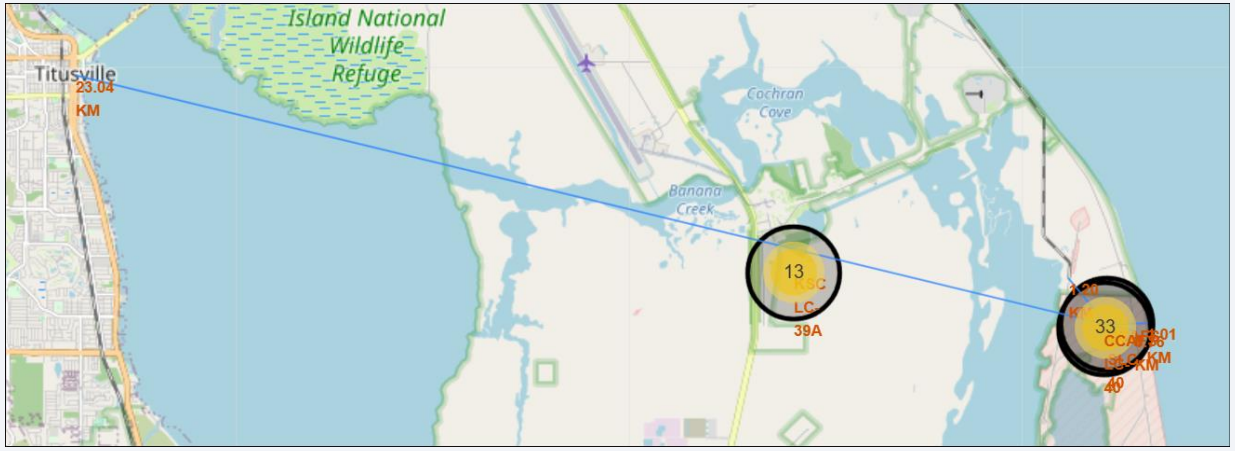
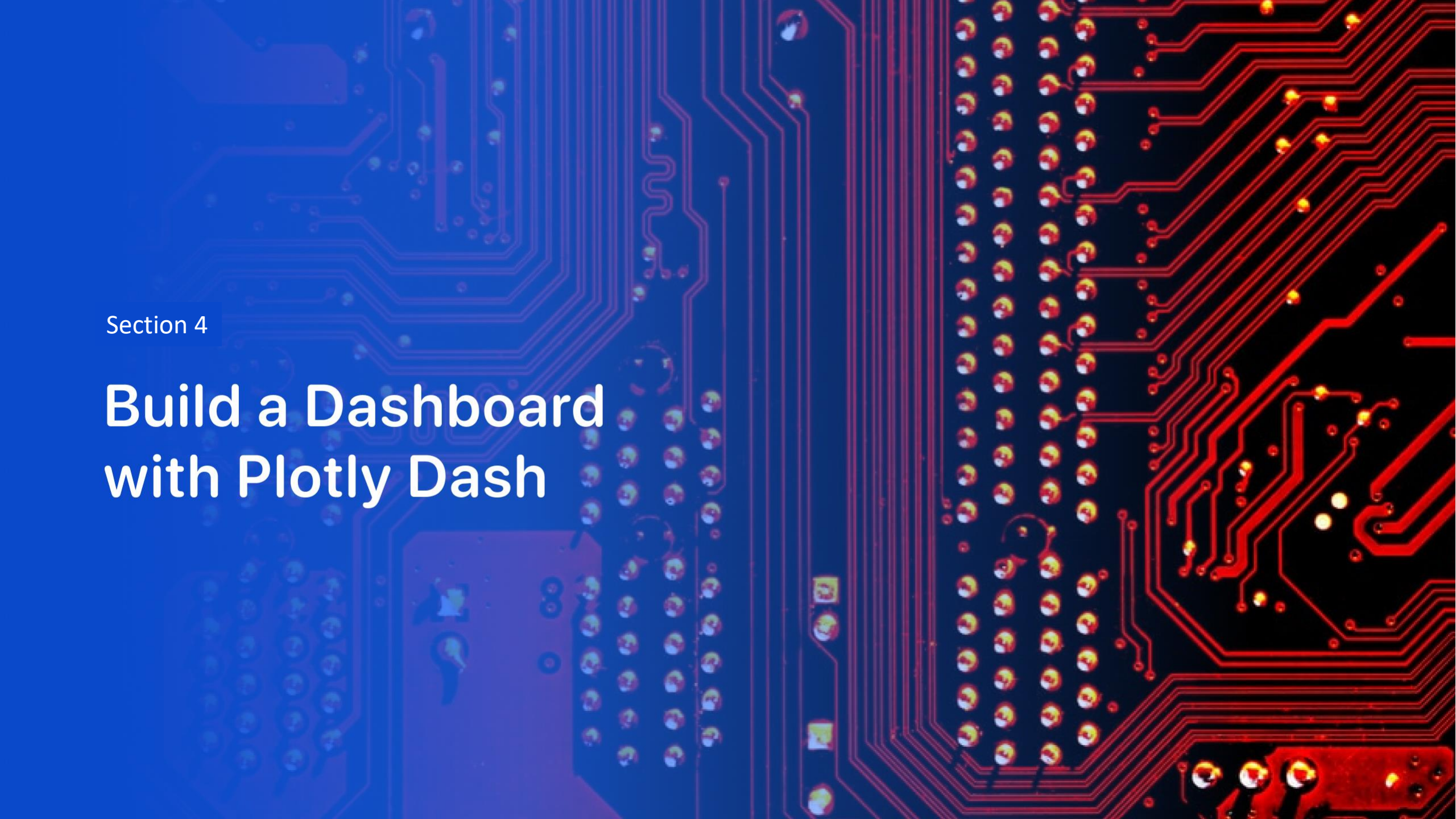


Figure 1: Distance of CCAFS LC-40 to closest city

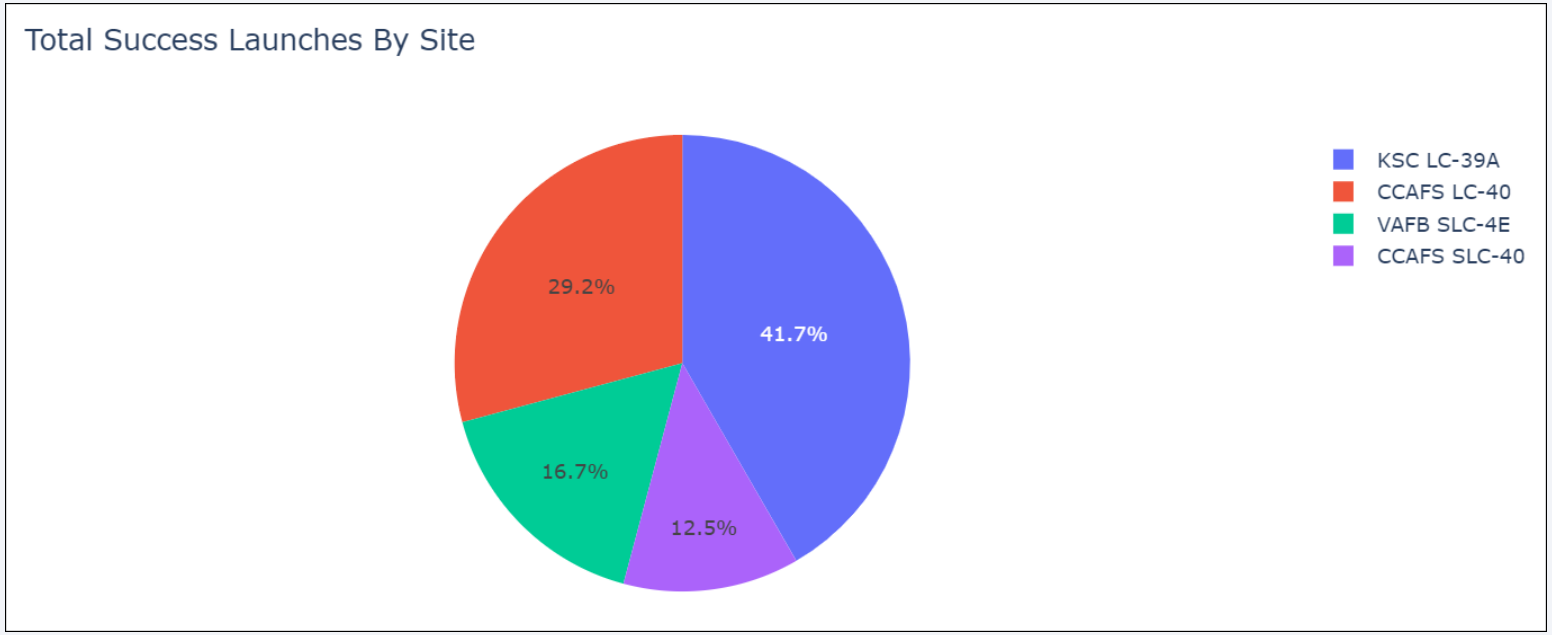
Figure 2: Distance of CCAFS LC-40 to closest coastline, highway and railroad

Figure 1 display the distance between CCAFS LC-40 to closest city Titusville, approximately 23.04KM. Figure 2 display the distance between CCAFS LC-40 to closest coastline, highway and railroad, which are approximately 1.01KM, 0.76KM and 1.20KM respectively.

In general, launch sites are located further from cities than coastlines, railroads, and highways to minimize the potential adverse effects on the public.


Section 4
Build a Dashboard with Plotly Dash

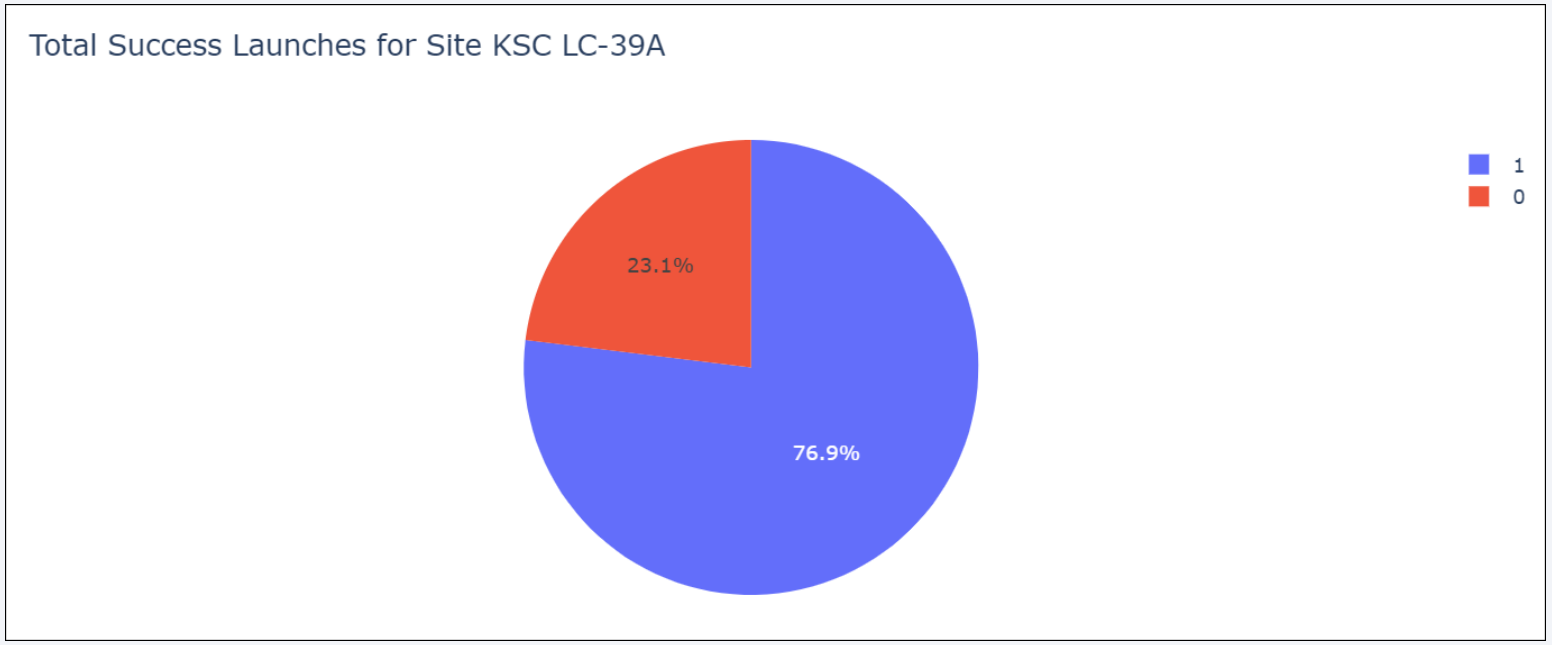
Total Success Launches by All Launch Site



The above pie chart illustrates the distribution of total success launches based on launch site.
Observed that:

1. Launch site KSC LC-39A has the highest launch success rate.
2. Launch site CCAFS SKC-40 has the lowest launch success rate.

Launch Site with Highest Launch Success Ratio



KSC LC-39A is the launch site with highest launch success ratio, the above pie chart shows that its success ratio is up to 76.9%.

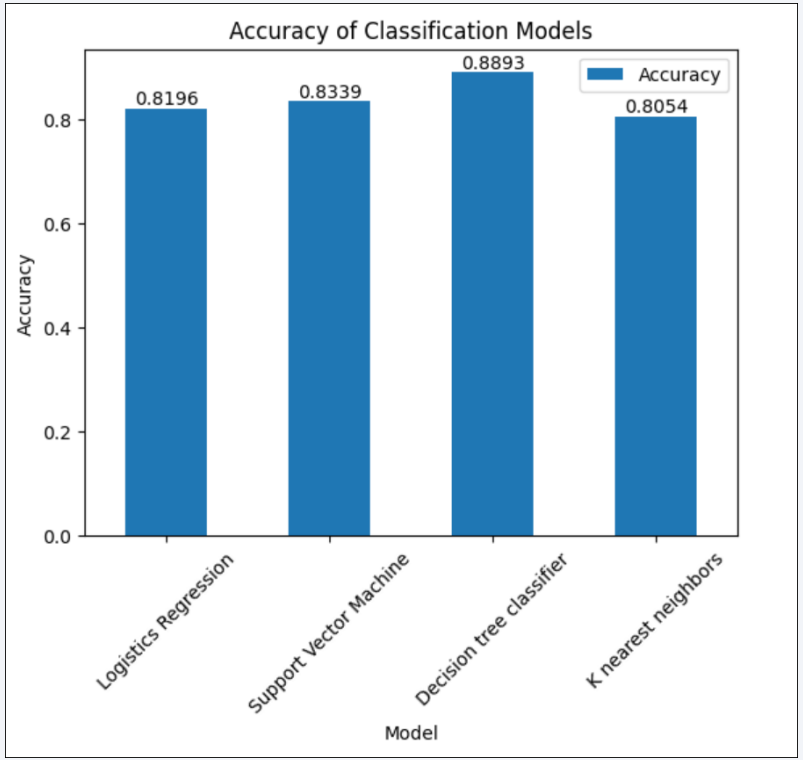
Payload vs. Launch Outcome Scatter Plot For All Sites

Most of the success launch carry payload mass range from 2,000 KG to 6,000 KG. Within this range FT booster version has the highest success rate out of all booster version.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

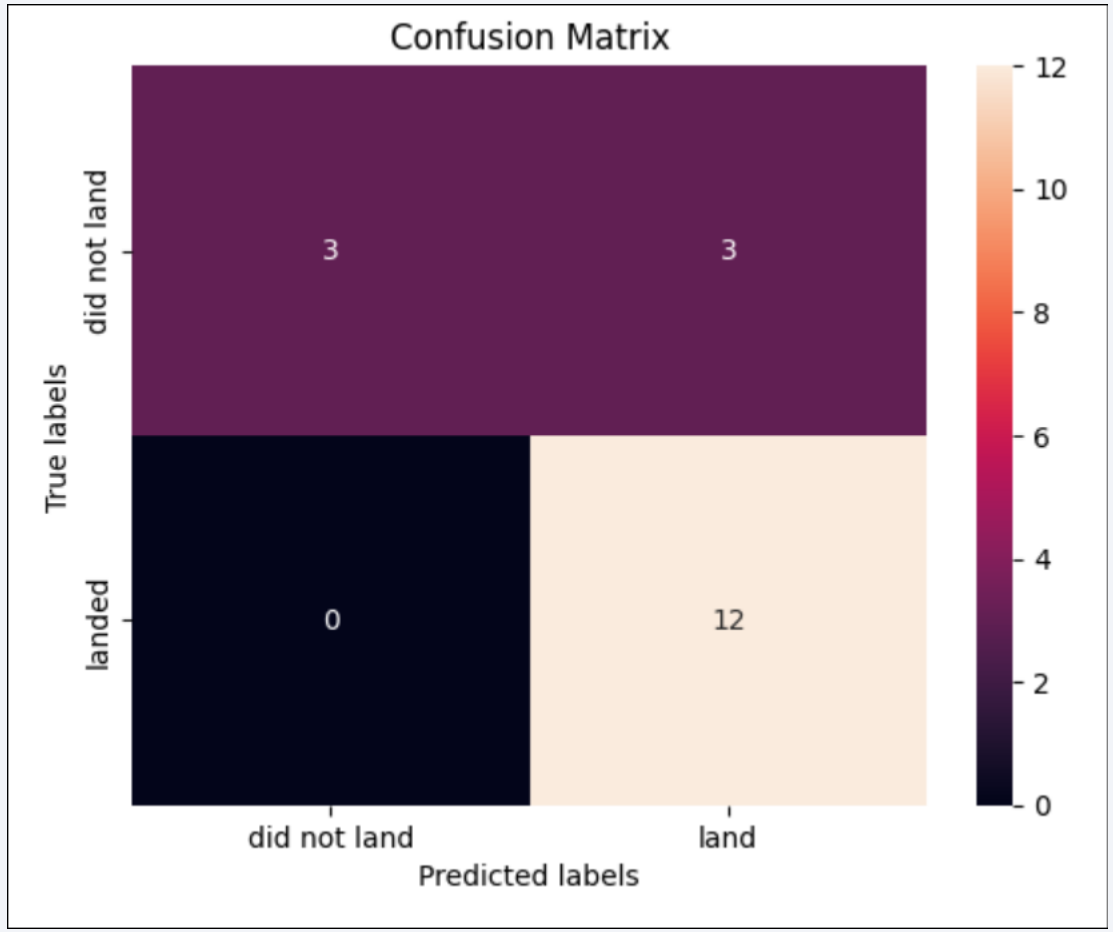


	Model Name	Accuracy	Test Data Accuracy
0	Logistics Regression	0.819643	0.833333
1	Support Vector Machine	0.833929	0.833333
2	Decision tree classifier	0.889286	0.833333
3	K nearest neighbors	0.805357	0.722222

Decision tree classifier has the highest accuracy score with value 0.8893 based on the bar chart. It also has accuracy score of 0.8333 when performing model evaluation using test data.

These evidences made decision tree classifier the best performing model.

Confusion Matrix



This confusion matrix of decision tree classifier shows that the model has performed 18 predictions:

- 12 predictions were predicted correctly for success landing – True positive
- 3 predictions were predicted correctly for failed landing – True negative
- 3 predictions were predicted to be landed successfully however they did not land successfully – False positive

Conclusions

- The chance of success launch increased as the number of flight increased.
- There are 4 orbit have highest success rate: ES-L1, GEO, HEO, and SSO
- Based on the observations from the line chart, the success rate increase since 2013 kept increasing till 2020 despite a drop in year 2018.
- In general, launch sites are located further from cities than coastlines, railroads, and highways to minimize the potential adverse effects on the public.
- Launch site KSC LC-39A has the highest launch success rate and launch site CCAFS SKC-40 has the lowest launch success rate.
- Most of the success launch carry payload mass range from 2,000 KG to 6,000 KG. Within this range FT booster version has the highest success rate out of all booster version.
- Decision tree classifier is the best performing model as it has the highest accuracy score with value 0.8893 and accuracy score of 0.8333 when performing model evaluation using test data.

Appendix

- All relevant materials of this project has been uploaded to GitHub.

Thank you!