IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chee Wei Han
4th March 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection through API & Web scraping
  - Data Wrangling
  - EDA with SQL
  - EDA with visualization
  - Interactive Visual Analytics with Folium and Dash
  - Machine Learning Prediction
- Summary of all results
  - EDA Results
  - Interactive Analytics Result
  - Predictive Analysis Result

Introduction

- **Project background and context**

Reducing cost for launching satellites is crucial to sustain the rocket launch company. For example, SpaceX has achieved low cost if they reuse the first stage where the parts of rocket land back after launching the satellites. Hence, predicting if the first stage will land successfully will helps our company Space Y by reducing the cost.

- **Problems want to find answers**

With the data extracted from SPACE X info,

  - Determine the price of each launch.
  - Determine what features make the first stage successfully land back.

Section 1

Methodology

Methodology

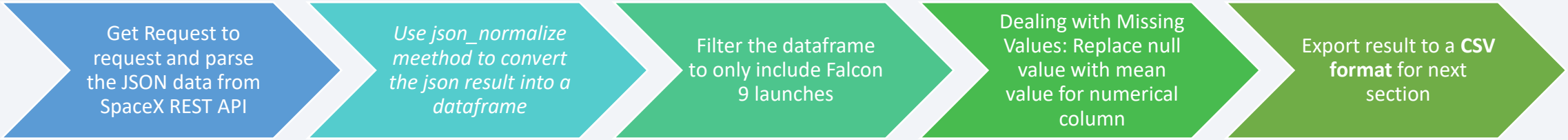
Executive Summary

- Data collection methodology:
  - Get request from SpaceX REST API and JSON
  - Perform Web scraping using BeautifulSoup function
- Perform data wrangling
  - Data is processed by feature engineering, handling imputation with numerical and categorical data to make sure the data is clean and good to be processed and analyzed.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building different models with different parameters and obtain the best accuracy.

Data Collection

- The SpaceX dataset was collected by getting request from SpaceX REST API and also from wiki page extracted using web scraping using Beautiful Soup function. Then these data will be stored into data frame using python.pandas package for ease on data processing.

Data Collection – SpaceX API



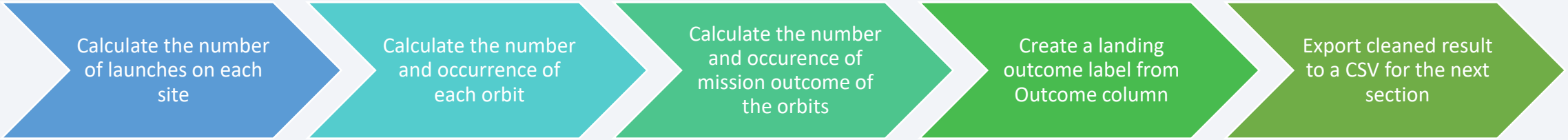
Link: SpaceX API

Data Collection - Scraping

Link : Web Scraping

Data Wrangling

Data Wrangling in this project includes exploratory data analysis (EDA) on the attributes of the column and labelling the outcome to 1 for the data showed success landing and 0 for fail landing.



Link : Data Wrangling

EDA with Data Visualization

- The charts were plotted to get the insight how these attributes will affect the outcome for reference. Charts includes:
  - FlightNumber vs. Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Orbit type vs Success Rate
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
  - Success Rate year wise

Link : EDA with Data Visualization

EDA with SQL

SQL queries performed includes:

- 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 achieved.
- List the boosters which have success in drone ship and have payload mass between 4000 and 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass.
- List the records failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

Link : EDA with SQL

Build an Interactive Map with Folium

Perform the visualization more interactive and detail, includes:

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

Link : Interactive Map with Folium

Build a Dashboard with Plotly Dash

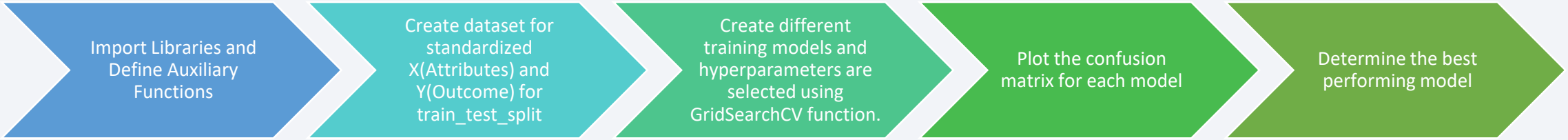
Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.

- Add a Launch Site Drop-down Input Component
- Add a callback function to render success-pie-chart based on selected site dropdown
- Add a Range Slider to Select Payload
- Add a callback function to render the success-payload-scatter-chart scatter plot

Link : SpaceX Dash App

Predictive Analysis (Classification)

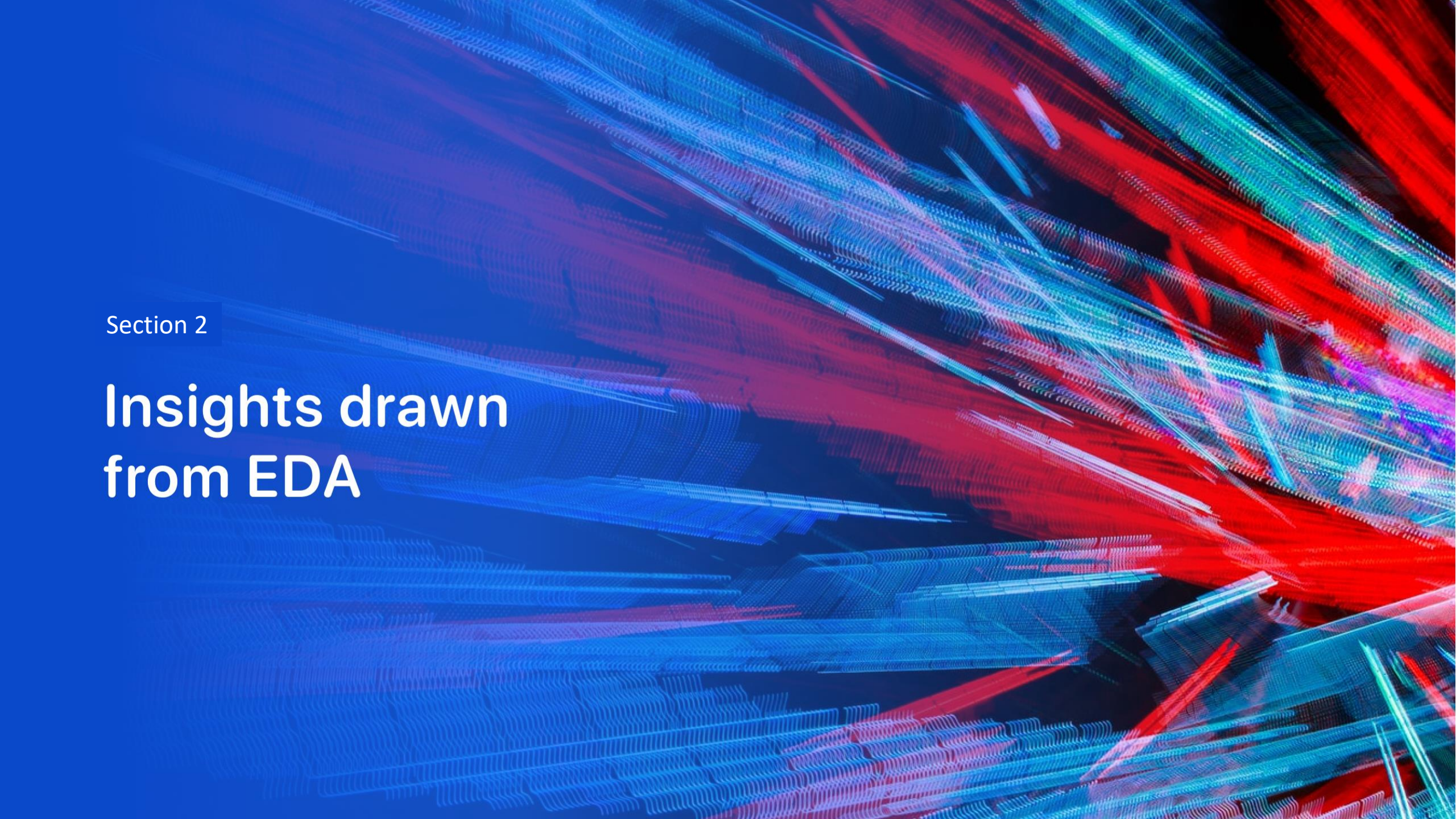
- Built, evaluated, improved, and found the best performing classification model
- Classification models that were used are Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbor Method. Each models are repeated trained with different parameters to get the best accuracy itself.



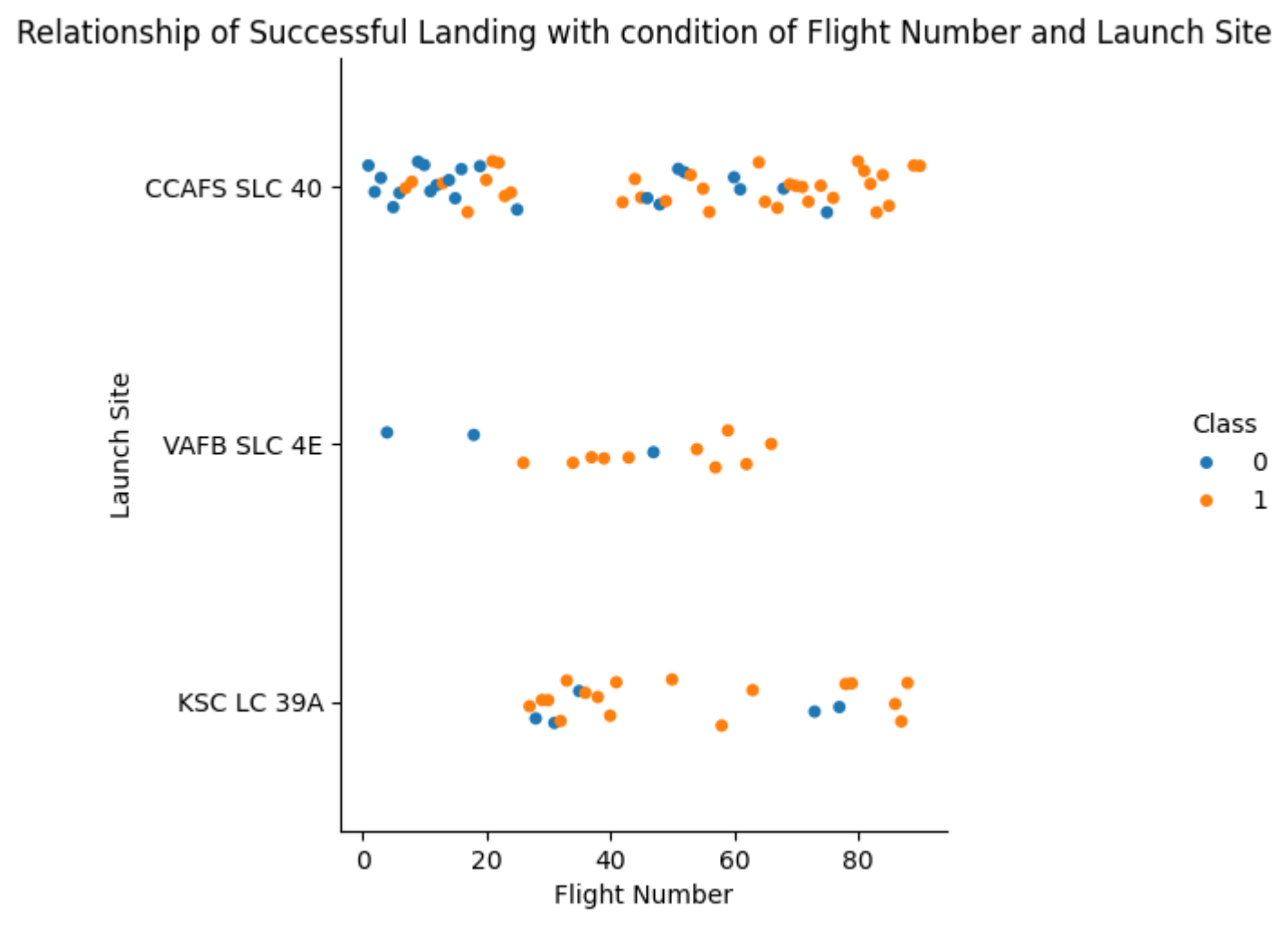
Link : Predictive Analysis (Classification)

Results

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

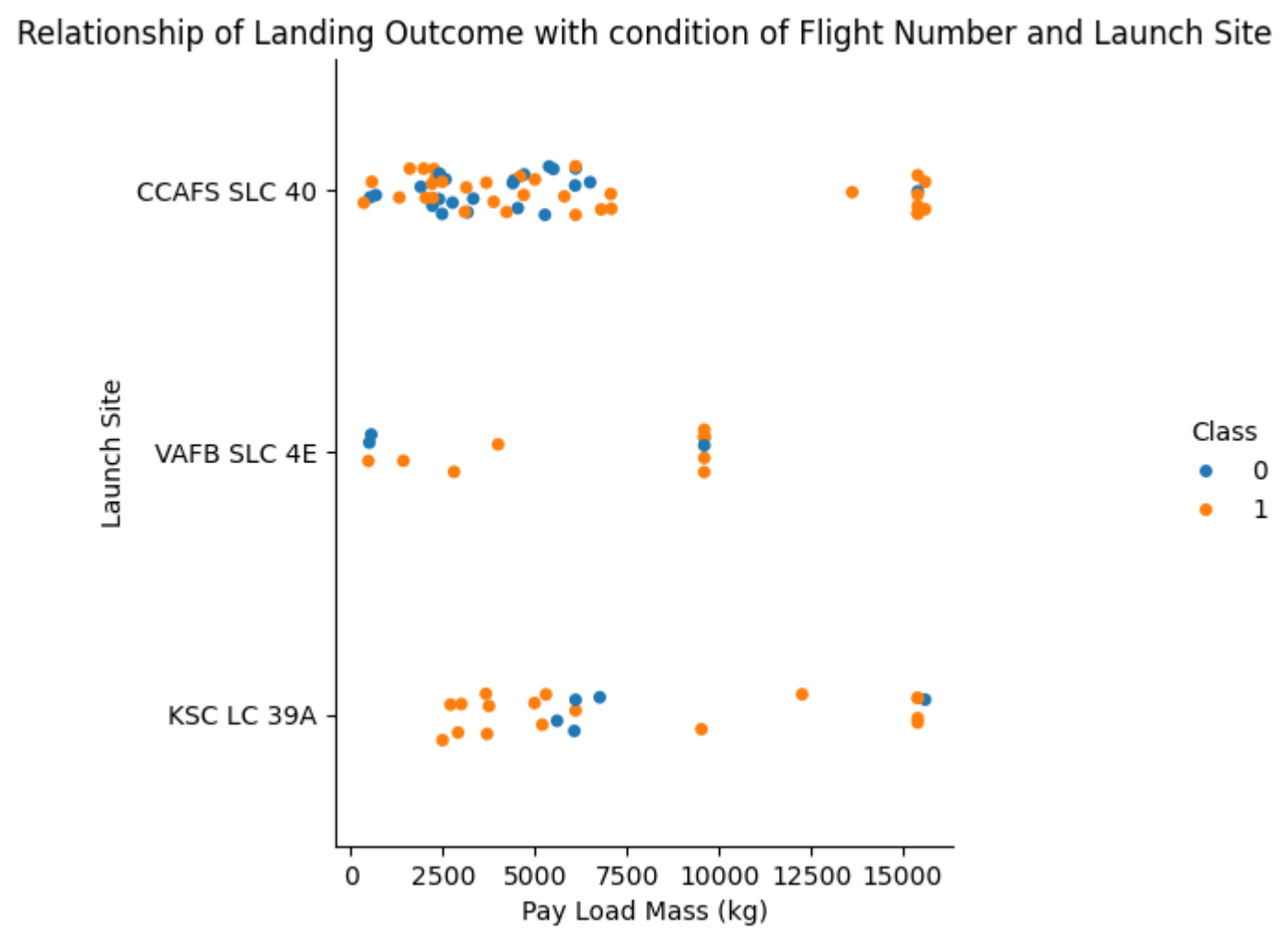

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



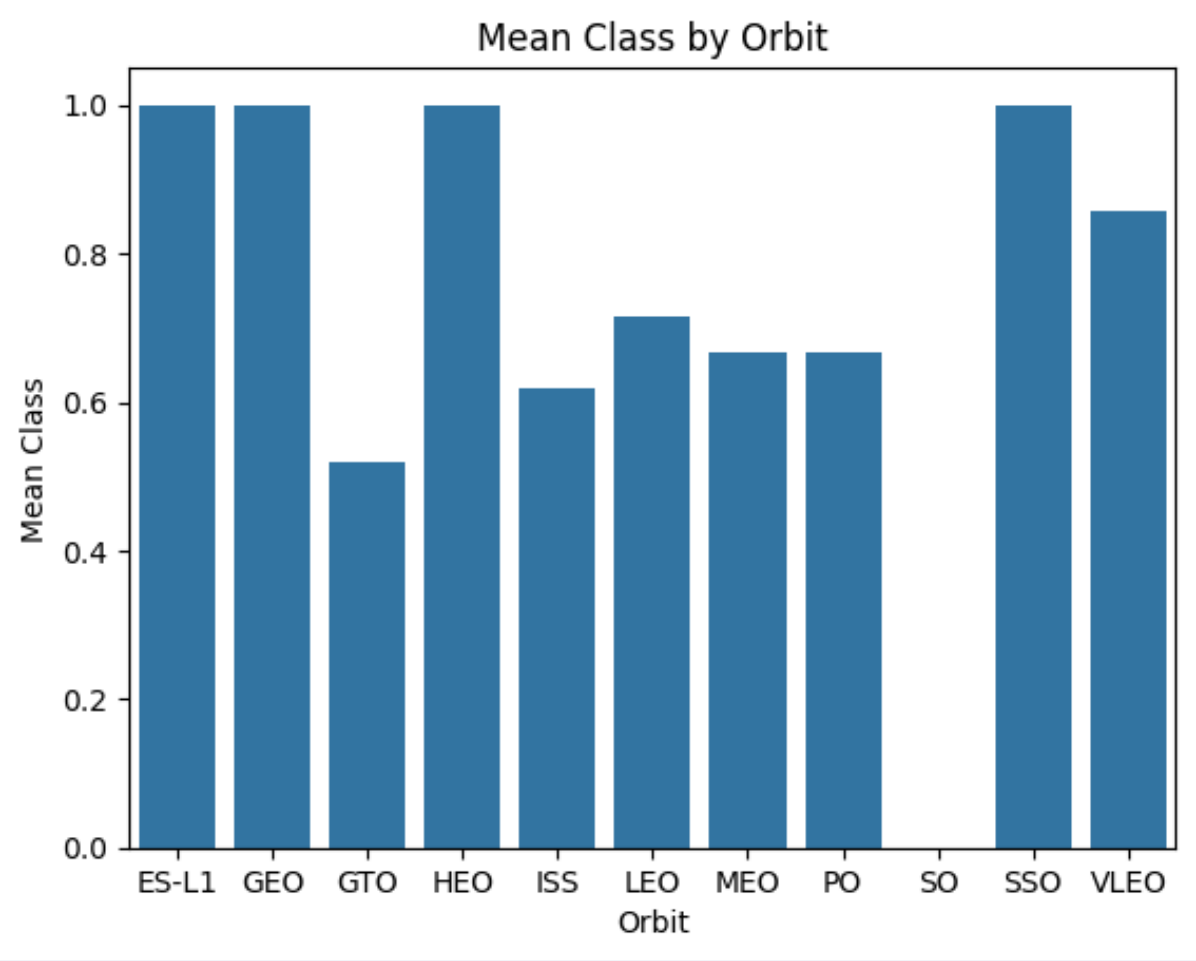
- The early flight number in each site has fail for landing, but success for higher flight number.
- CCAFS SLC 40 has highest number of launches.

Payload vs. Launch Site



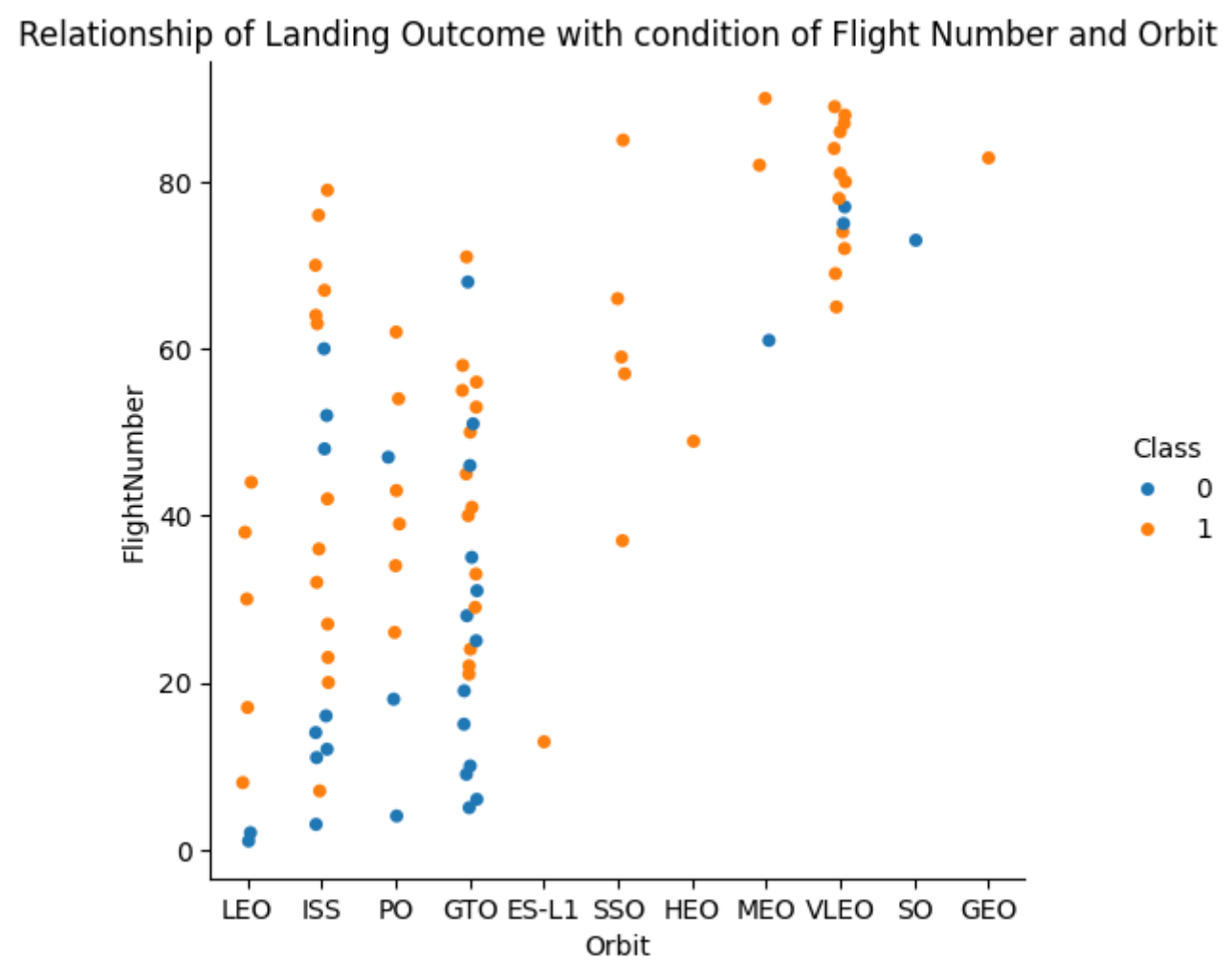
- The early flight number in each site has fail for landing, but success for higher flight number.
- CCAFS SLC 40 has highest number of launches.

Success Rate vs. Orbit Type



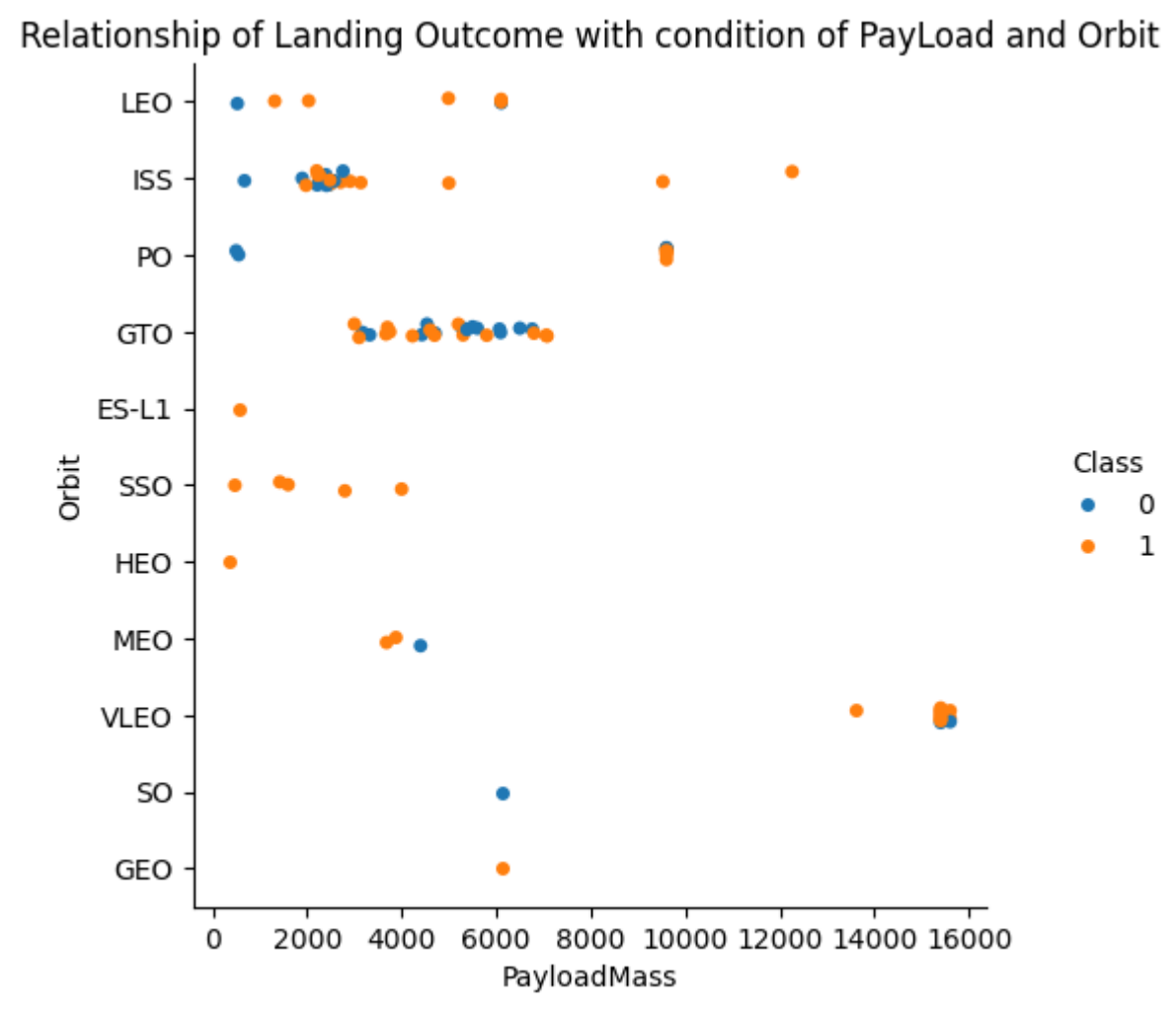
- Highest Success Rate for the orbit type are ES-L1, GEO, HEO and SSO

Flight Number vs. Orbit Type



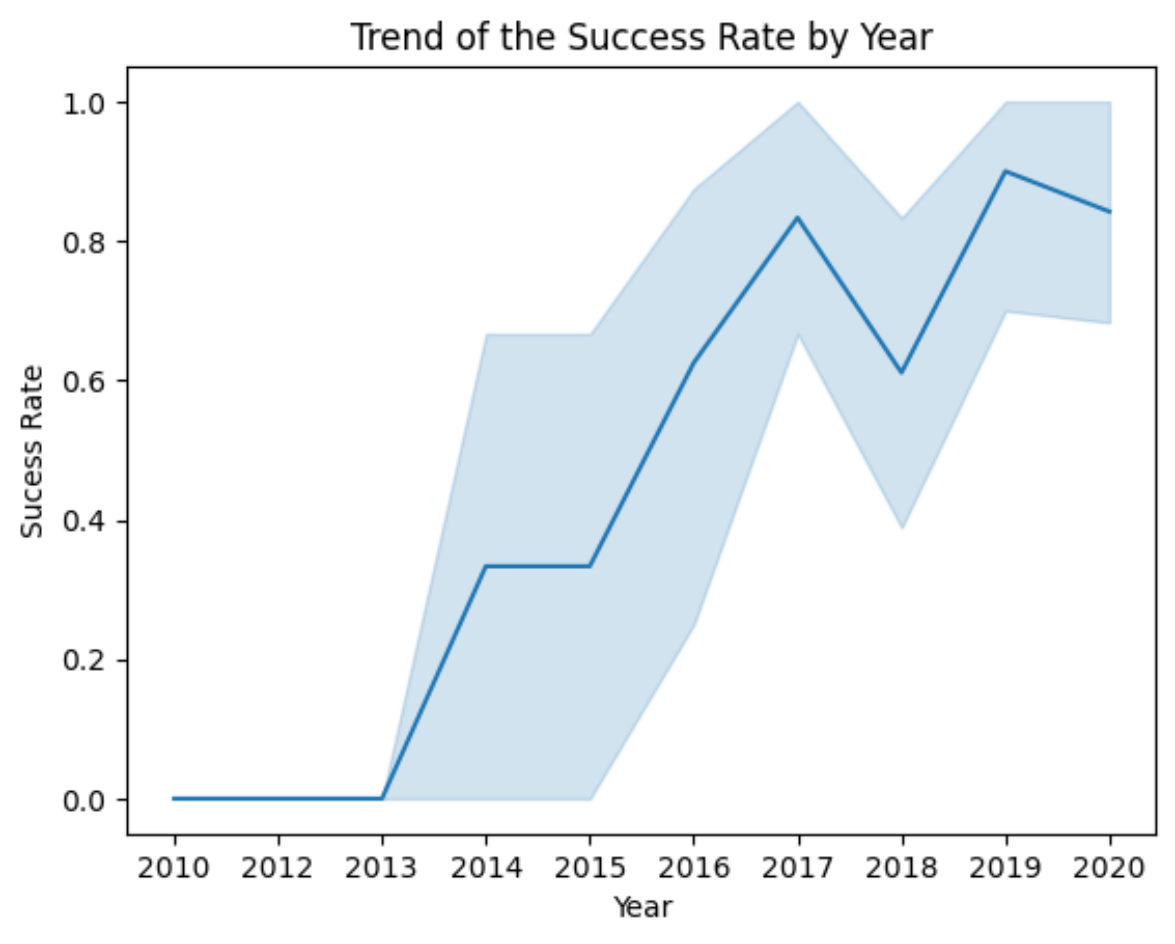
- When flight number increase the success landing in each orbit also increase.
- SSO have highest success rate in each flight number.

Payload vs. Orbit Type



- Majority Pay Load Mass is below 8000kg.
- SSO orbit has achieved all success that below 6000kg of pay load.

Launch Success Yearly Trend



- The Success Rate keep increasing since 2013 year.

All Launch Site Names

Names of the unique launch sites: (CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40)

Task 1

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

```
%sql SELECT distinct(Launch_Site) FROM SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

Task 2

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

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

Python

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total payload carried by boosters from NASA

Task 3

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

```
%sql SELECT Customer, SUM(PAYLOAD_MASS__KG_)  AS Total_Payload_MASS FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

Customer	Total_Payload_MASS
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

Task 4

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

```
%sql SELECT Booster_Version, avg(PAYLOAD_MASS__KG_) AS Average_Payload_MASS FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	Average_Payload_MASS
F9 v1.1	2928.4

First Successful Ground Landing Date

Dates of the first successful landing outcome on ground pad

```
[42] %sql SELECT * FROM SPACEXTABLE WHERE Date = (SELECT min(Date)  FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)')
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

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

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT Booster_Version,PAYLOAD_MASS__KG_,Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' and (PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000)
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome)FROM SPACEXTABLE GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Names of the booster which have carried the maximum payload mass

Task 8

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

```
[51] %sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_ ) FROM SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
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

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

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT substr(Date,6,2) AS Months,substr(Date,0,5) AS Year, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' and Landing_Outcome = '
```

Python

```
 * sqlite:///my_data1.db
Done.
```

Months	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Task 10

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
%sql SELECT Landing_Outcome, COUNT(*) AS Outcomes_Number FROM SPACEXTABLE WHERE (Date BETWEEN '2010-06-04' and '2017-03-20') GROUP BY Landing_Outcome ORDER BY Outcomes_Number DESC
```

Python

```
 * sqlite:///my_data1.db
Done.
```

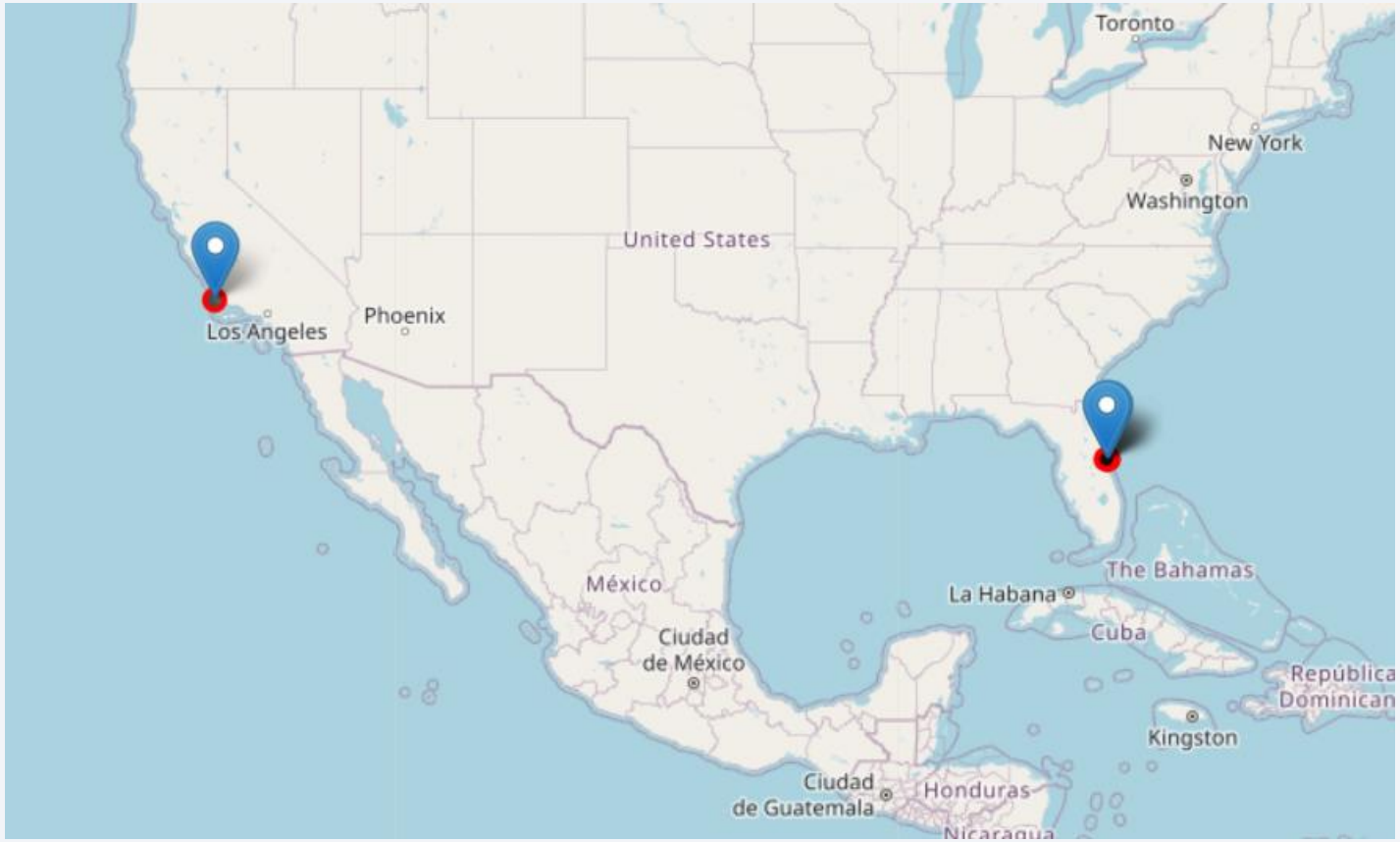
Landing_Outcome	Outcomes_Number
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

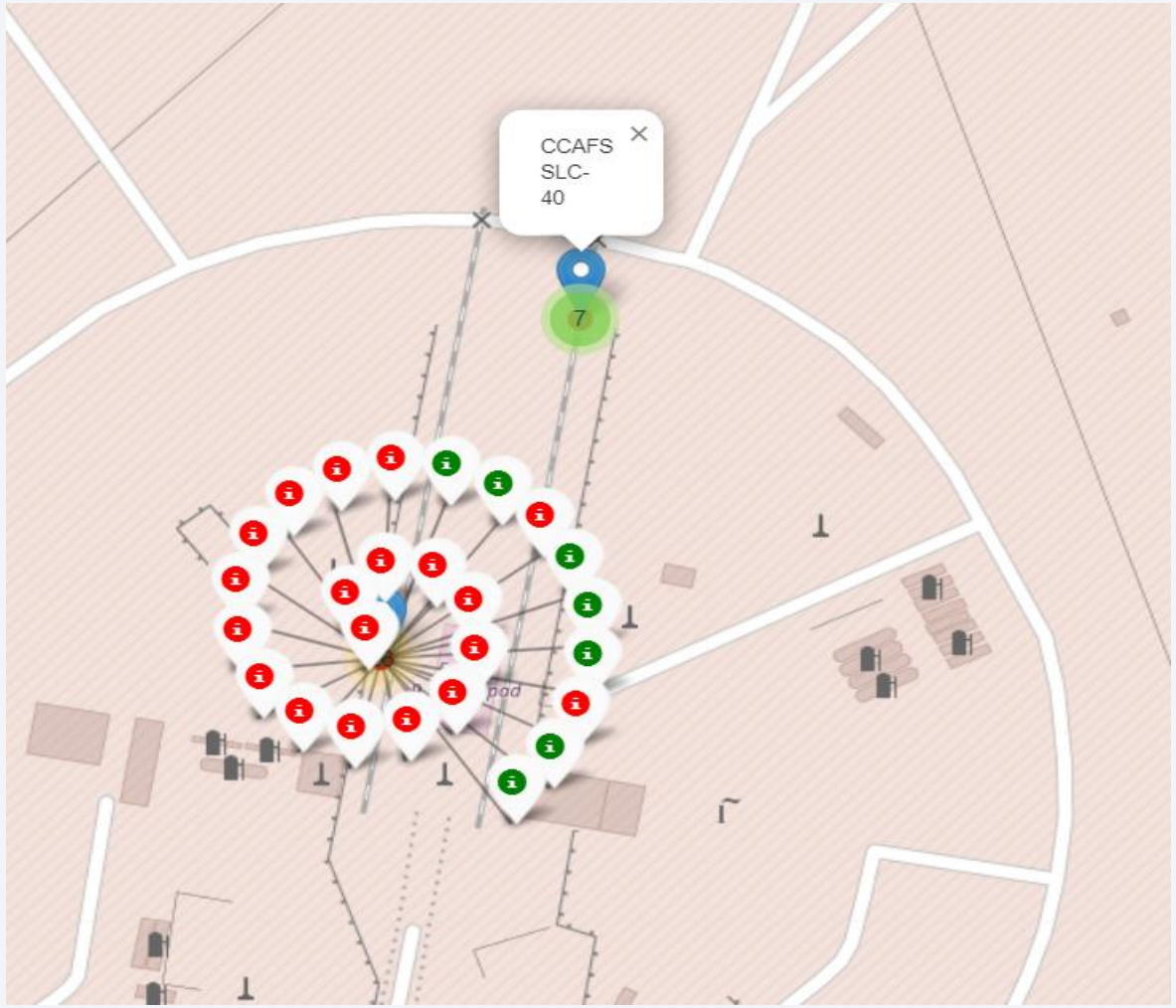
<Folium Map Launch Sites on map>

- All launch sites' location is marked on a global map



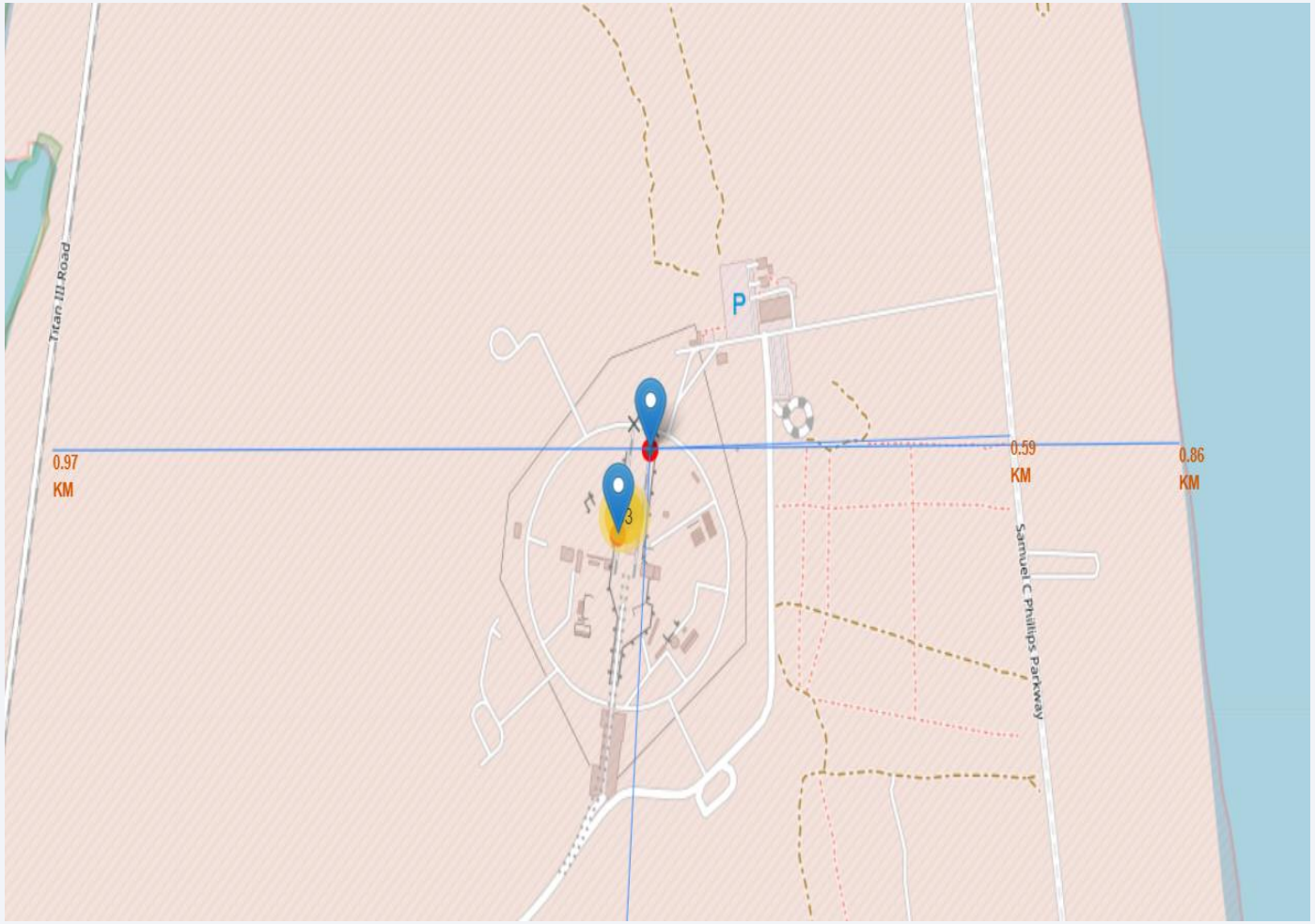
<Folium Map Success/Failed Launches for each site>

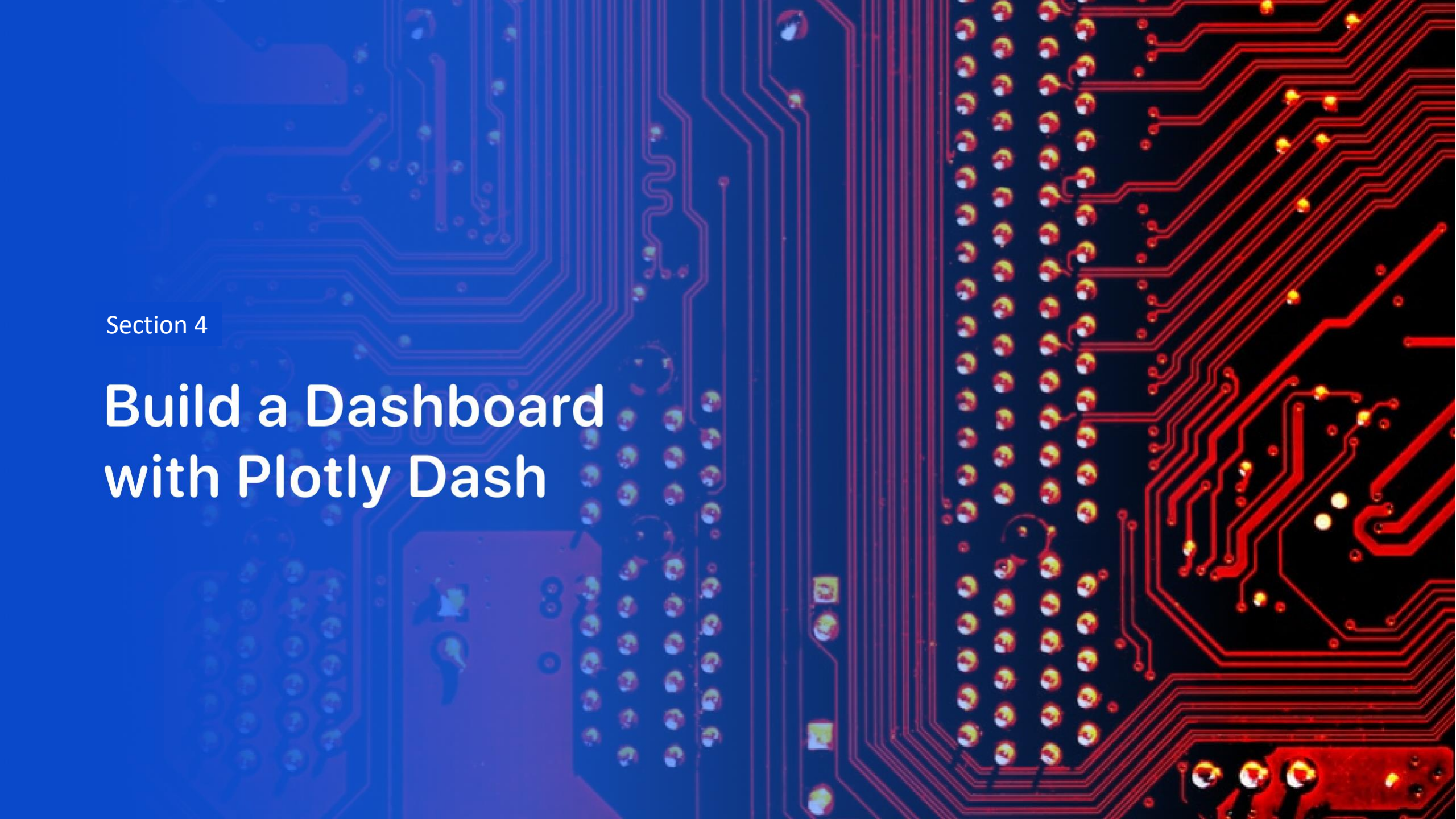
- The green label is the success launch and red for the failed launched on each site.



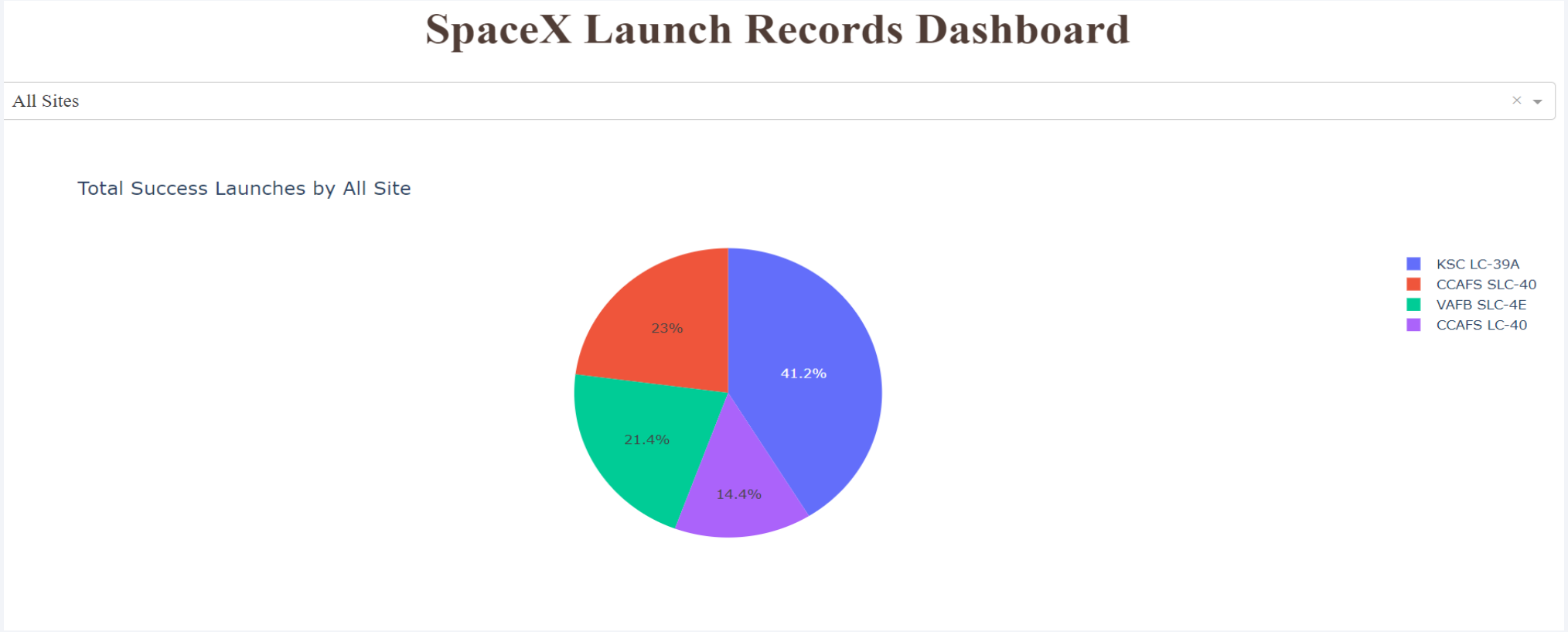
<Folium Map Distance between launch sites and proximities>

- Selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed



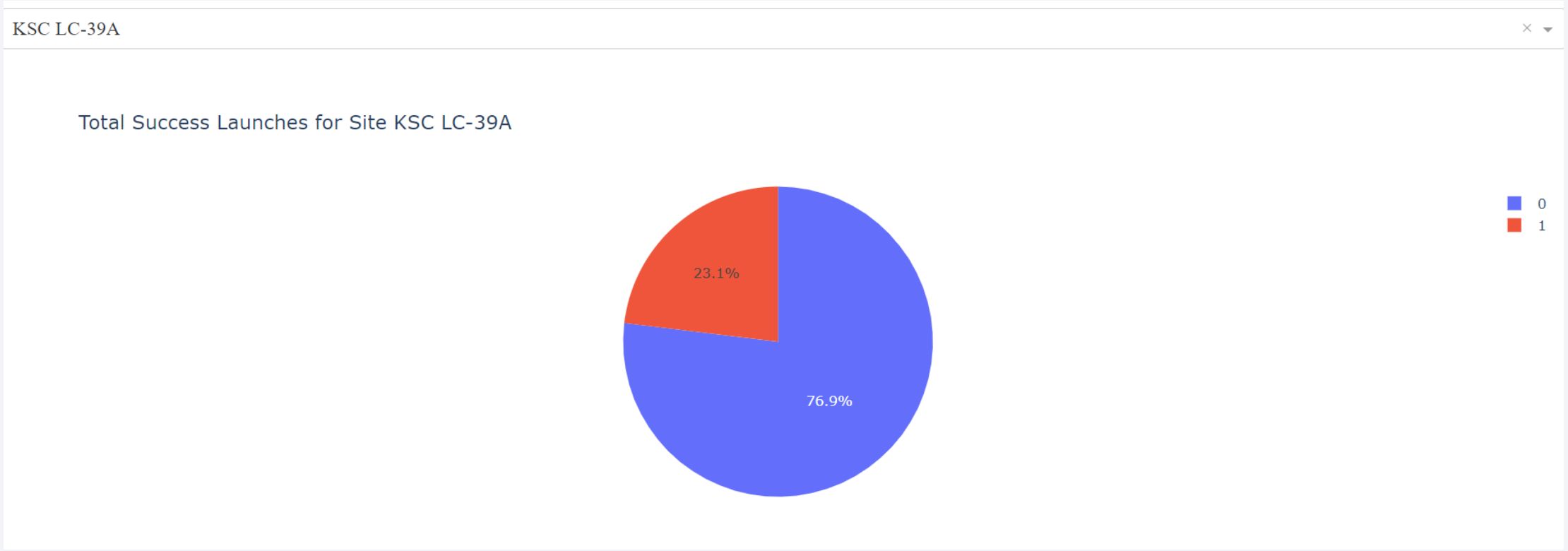

Section 4
Build a Dashboard with Plotly Dash

<SpaceX Launch Records All Sites>



- KSC IC-39A Site has highest success launches.

<Launch Site for Highest Success Ratio>



- Site KSL LC-39A has highest launch success ratio (76.9% for success launch and 23.1% failed launch)

<Payload vs. Launch Outcome for All Sites>

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

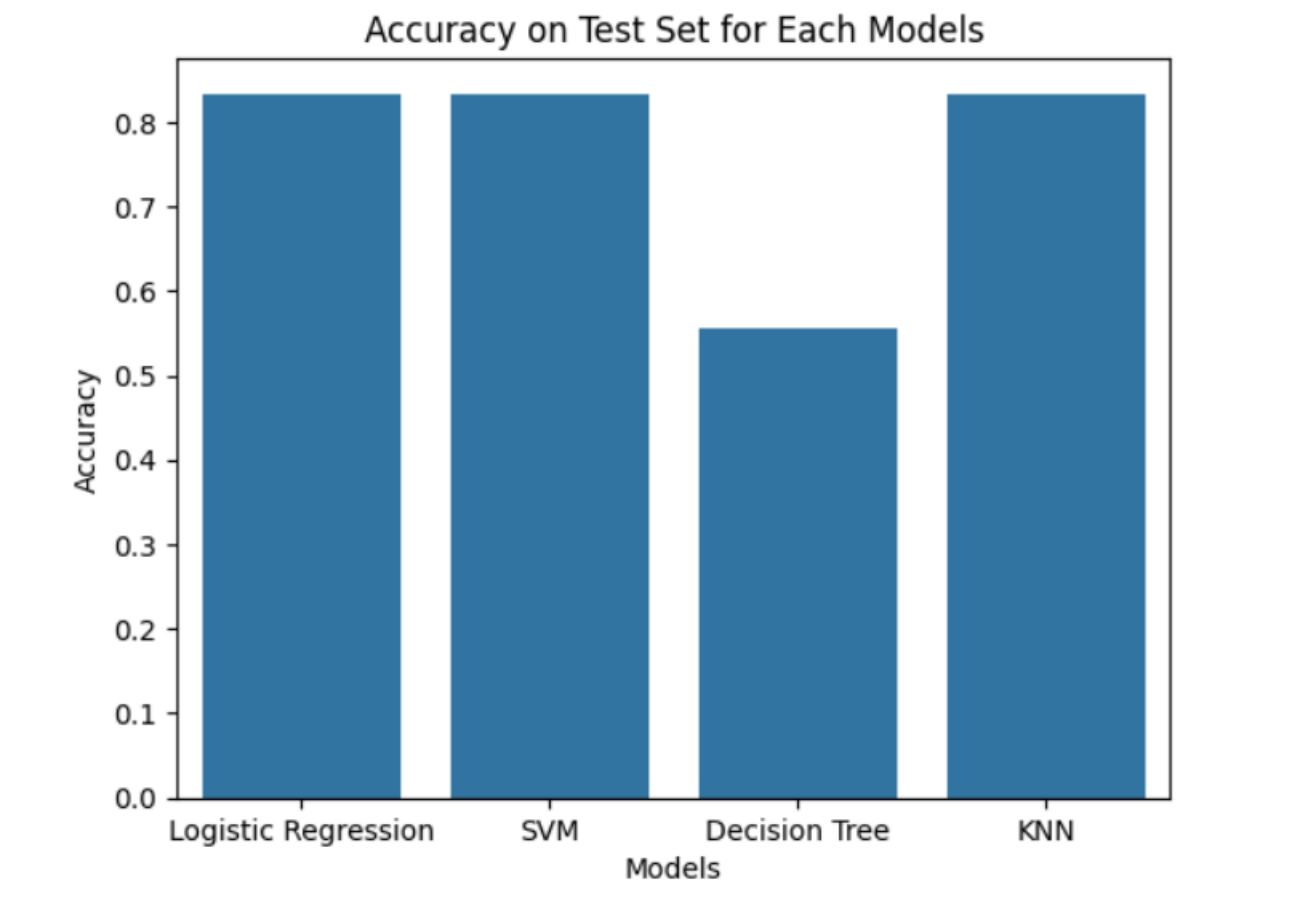
Section 5

Predictive Analysis (Classification)

Classification Accuracy

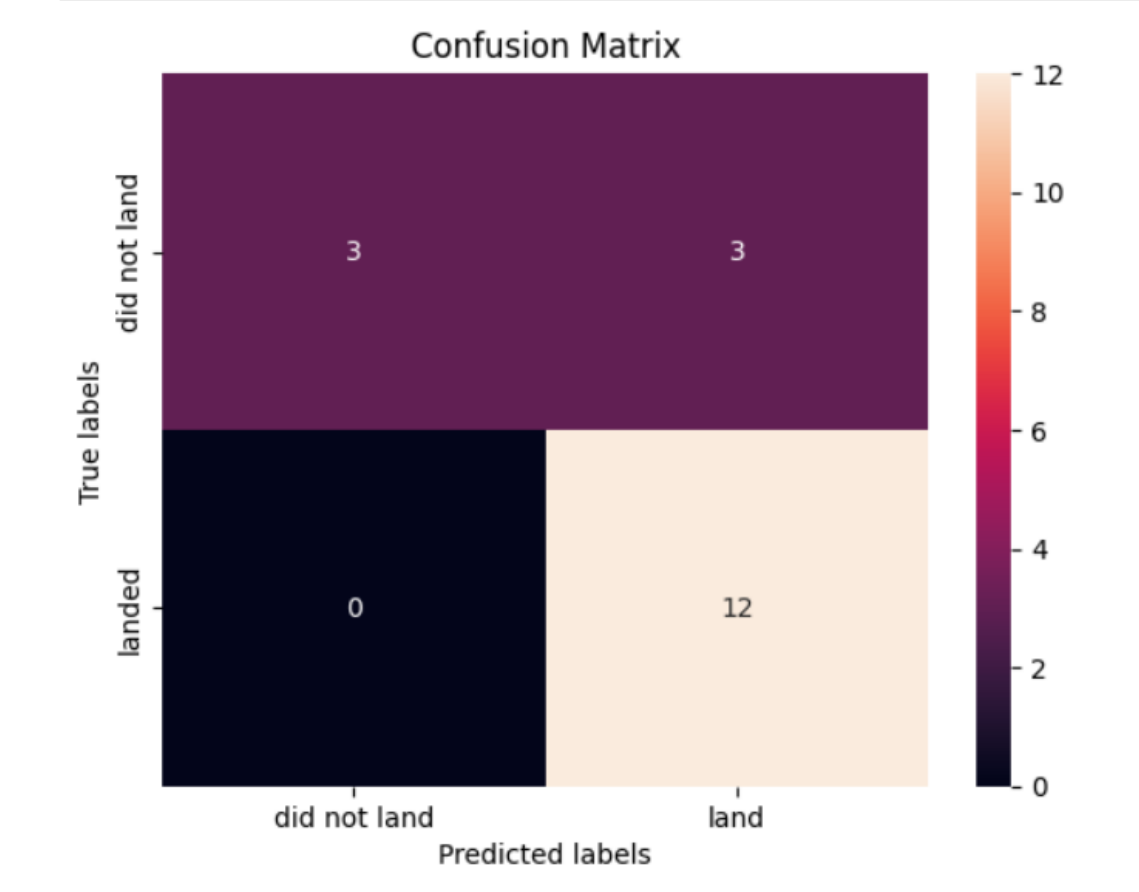
- Visualize the built model accuracy for all built classification models, in a bar chart

- All models has the highest classification accuracy except for the Decision Tree Models.



Confusion Matrix

The confusion matrix of the best performing model- Logistic Regression Model



Conclusions

- CCAFS SLC 40 has highest number of launches.
- Highest Success Rate for the orbit type are ES-L1, GEO, HEO and SSO
- SSO orbit can be considered for launching as it has 100% success rate and below 6000kg Pay Load Mass.
- The Success Rate keep increasing since 2013 year
- All models has the highest classification accuracy except for the Decision Tree Models. Hence, with the simple model, Logistic Regression can be selected.

Thank you!