IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kiki Ayu W.
December, 30 2022

Outline

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

Introduction

- Project background and context

In this capstone, we will predict if the Falcon 9 first stage will land successfully. 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 is 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.

- Problems you want to find answers

1. Determine the features/ data that affect success rate of landing.
2. Determine the characteristic of success landing in terms of launch site, orbit type, payload mass and booster version based on SpaceX past launches data history
3. Determine the best model to build predictive analysis to predict the first stage of rocket will successfully land. It can minimize the cost of second stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Get Request from the SpaceX REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Drop unnecessary column
  - One Hot Encoding for classification models
  - Adding new column showing the success landing infromation
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

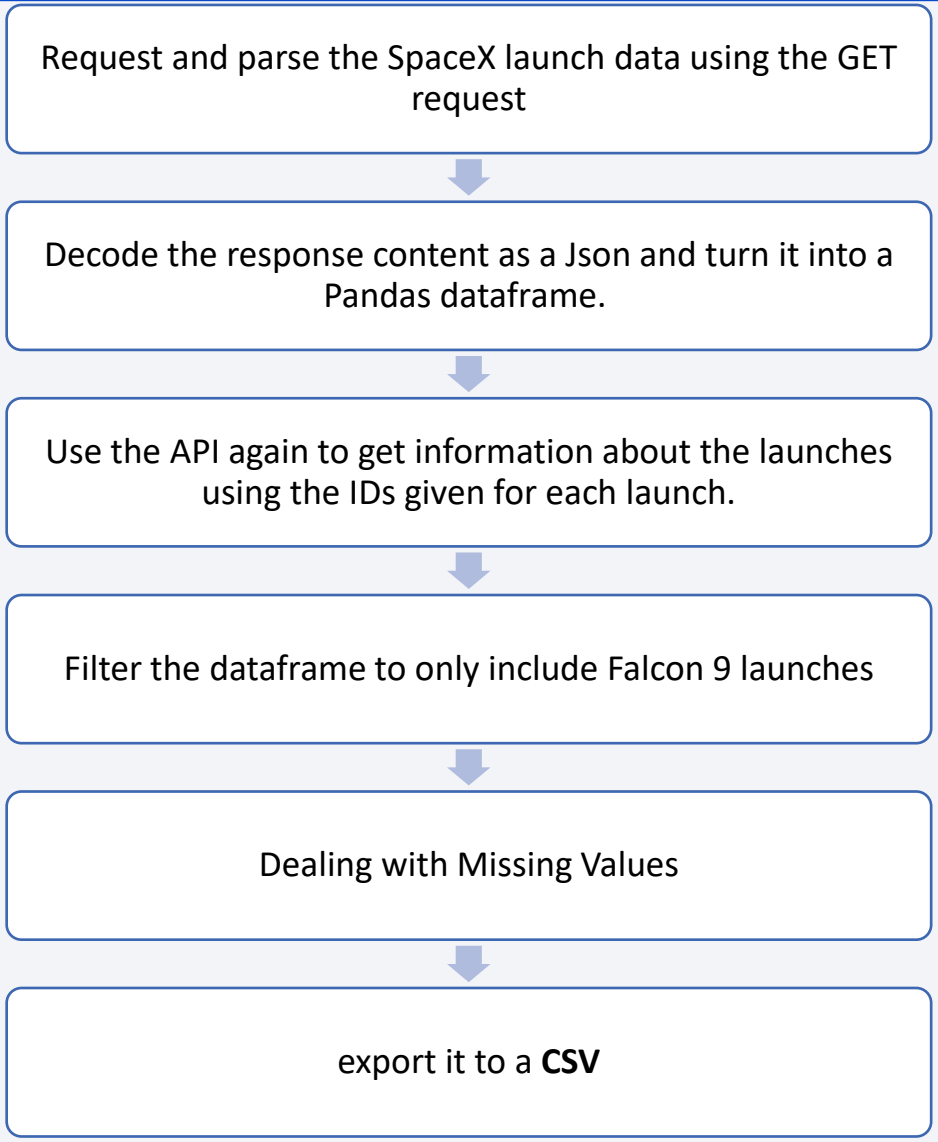
Data Collection

1. The data are collected from REST API SpaceX and SpaceX Wikipedia combined:
   - Data from Rest API SpaceX : Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
   - Data from Wikipeda (List of Falcon 9 and Falcon Heavy launches) : Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome

     Added some new columns: Version Booster, Booster landing, Date, Time
2. The data from Rest API source and Wikipedia source then filtered to only Falcon 9
3. Dealing with missing values
4. Export the data to CSV files

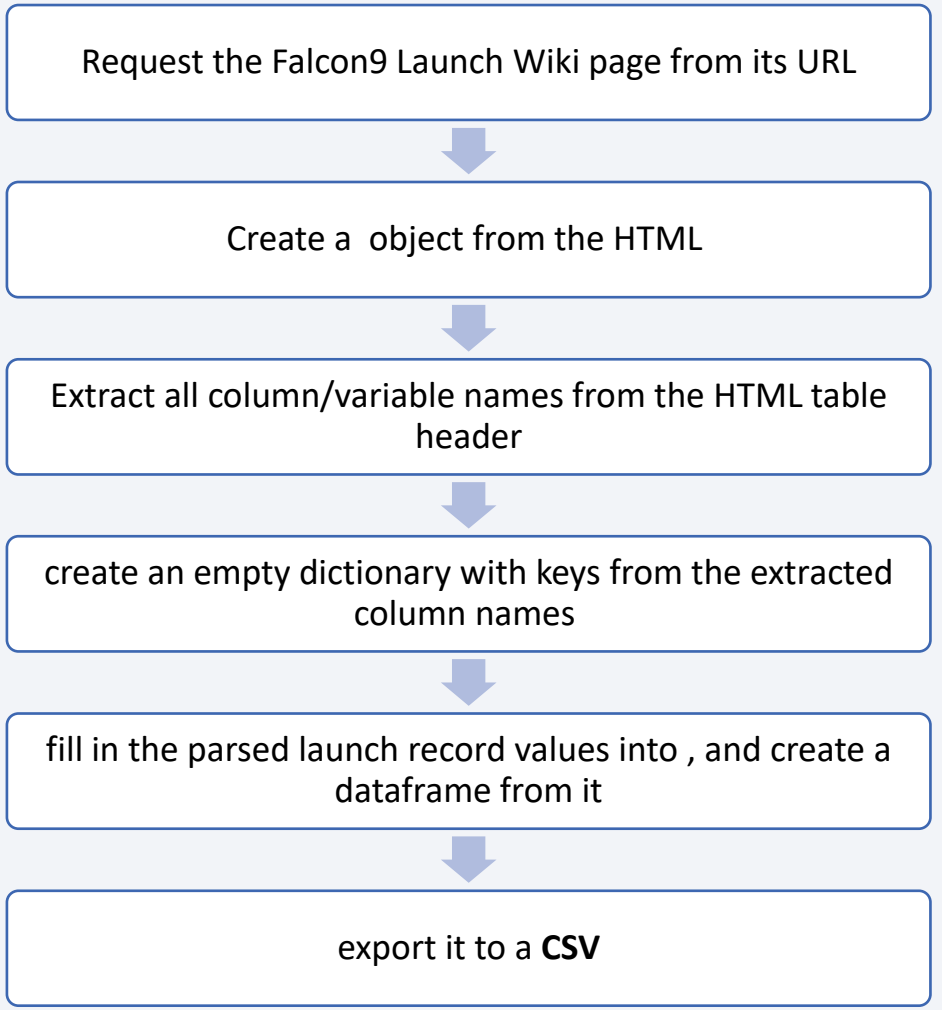
Data Collection – SpaceX API

GitHub URL of the completed SPACEX REST API.



Data Collection - Scraping

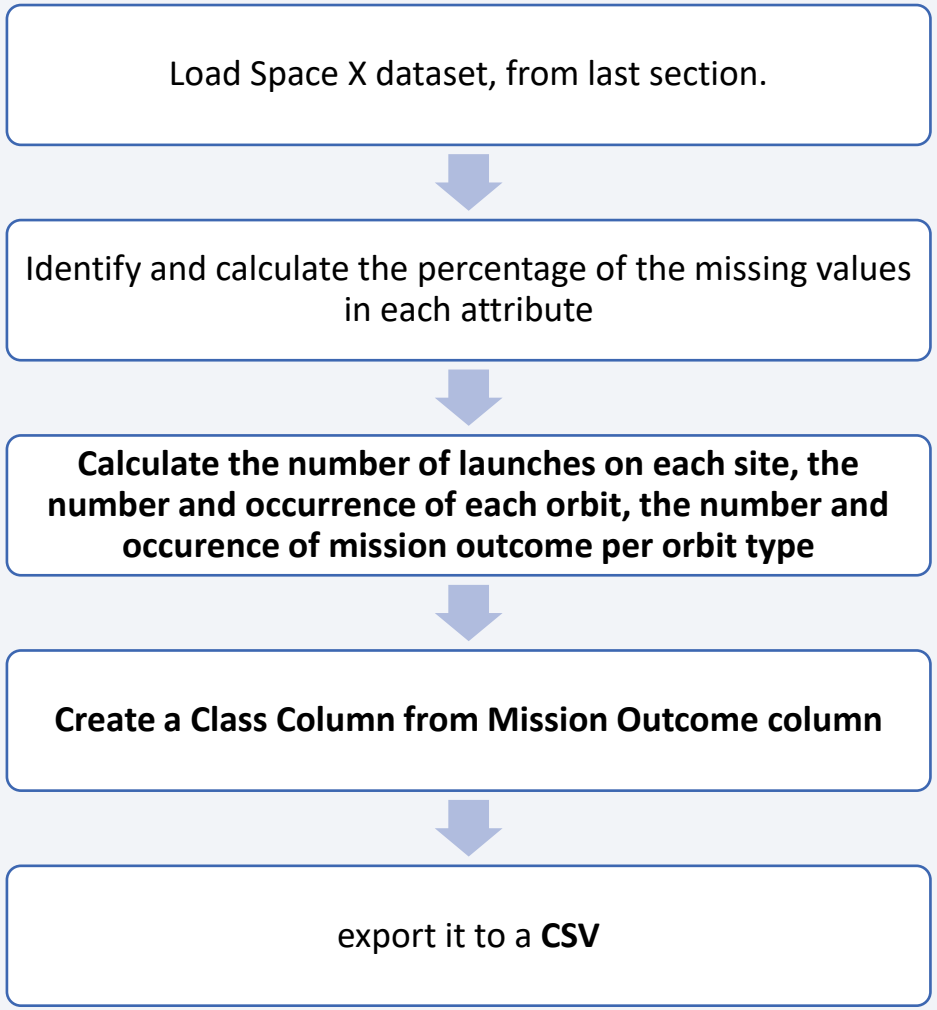
GitHub URL of the completed web scraping notebook



Data Wrangling

- The dataset from last section is loaded for this section and identify the missing values percentage for each attribute. Result : the mission values is 0% for each sttribute, except for LandingPad (type object)
- Count the mission outcome column and assign value 0 for failed outcome, 1 for success outcome per orbit type.
  - True ASDS (1), None None (0), True RLTS(1), False ASDS(0), True Ocean (1), False Ocean(0), None ASDS (0), False RTLS(1)
- Add class column into dataframe and assign it with value from mission outcome column above.

[GitHub URL of completed data wrangling]



EDA with Data Visualization

Charts were plotted in this project

1. Scatter Chart to see the relationship between: FlightNumber Vs PaylodMass, FlightNumber Vs LaunchSite, LaunchSite Vs PayloadMass, FlightNumber Vs Orbit type, Payload vs. Orbit

   with hue = class (0 : failed, 1: success).

2. Bar Chart to compare the success rate between each orbit type

3. Line chart to see the success rates trend every year

GitHub URL of completed EDA

EDA with SQL

- 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 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 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 successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub URL of completed EDA with SQL

Build an Interactive Map with Folium

Map Objects Used:

- create a folium `Map` object, with an initial center location to be NASA Johnson Space Center at Houston, Texas.
- Use `folium.Circle` to add a highlighted circle area with a text label on a specific coordinate
- Use folium.Marker to Create an icon as a text label
- Use object MarkerClusters () to simplify a map containing many markers having the same coordinate.
- Use object `MousePosition` to get coordinate for a mouse over a point on the map
- Draw a `PolyLine` between a launch site to the selected coastline point

GitHub URL of completed interactive map with Folium map

Build a Dashboard with Plotly Dash

Plots/graphs and interactions added to a dashboard:

1. dcc.Dropdown: to select which Launch Sites we want to see (All sites, CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40)
2. Pie chart: visualizing launch success counts when receiving input frim dcc.dropdown/site-dropdown
3. dcc.RangeSlider : to to be able to easily select different payload range and see if we can identify some visual patterns.
4. A scatter plot : with the x axis to be the payload and the y axis to be the launch outcome we can visually observe how payload may be correlated with mission outcomes for selected site(s)
5. Dash Callback function: receiving an input form dcc.dropdown and range sluder and giving a pie chart and scatter chart output

GitHub URL of completed Plotly Dash lab

data = pd.read_csv(text1)

Predictive Analysis (Classification)

Create a NumPy array from the column Class in Data and assign it as target value (Y)

Standardize the data in then reassign it to predictor value (X)

Use the function train_test_split to split the data X and Y into training and test data.

Create a logistic regression, support vector machine object, decision tree classifier object, k nearest neighbors object

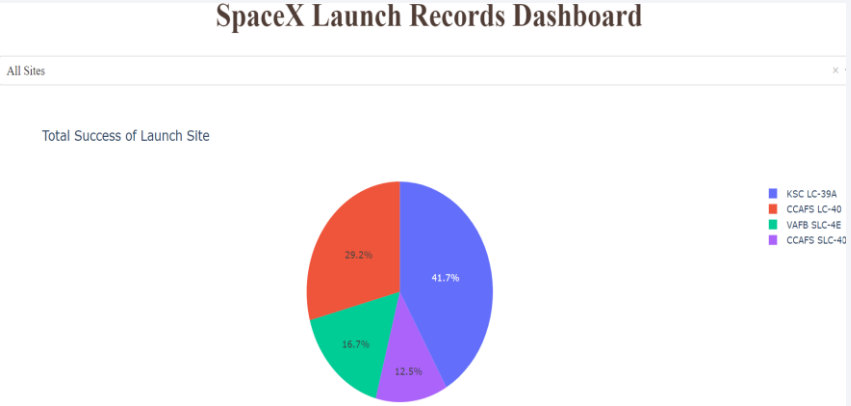
Create GridSearch for each object to find the best parameters

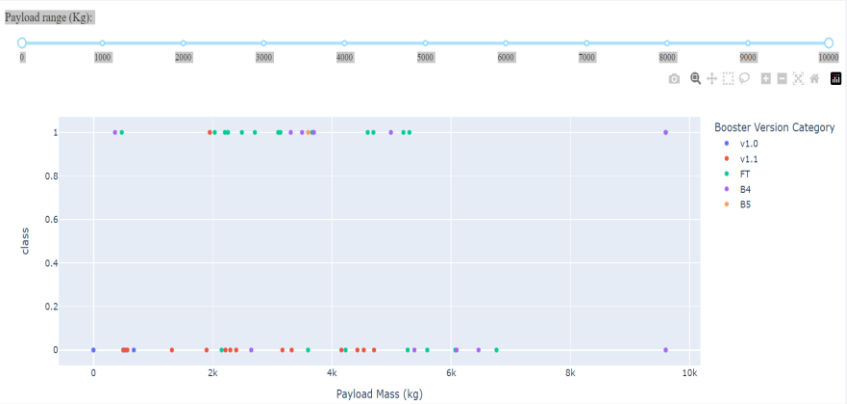
Calculate the accuracy of each object on the test data using the method score() and compare them. Find the best model with highest accuracy

GitHub URL of completed predictive analysis lab

Results

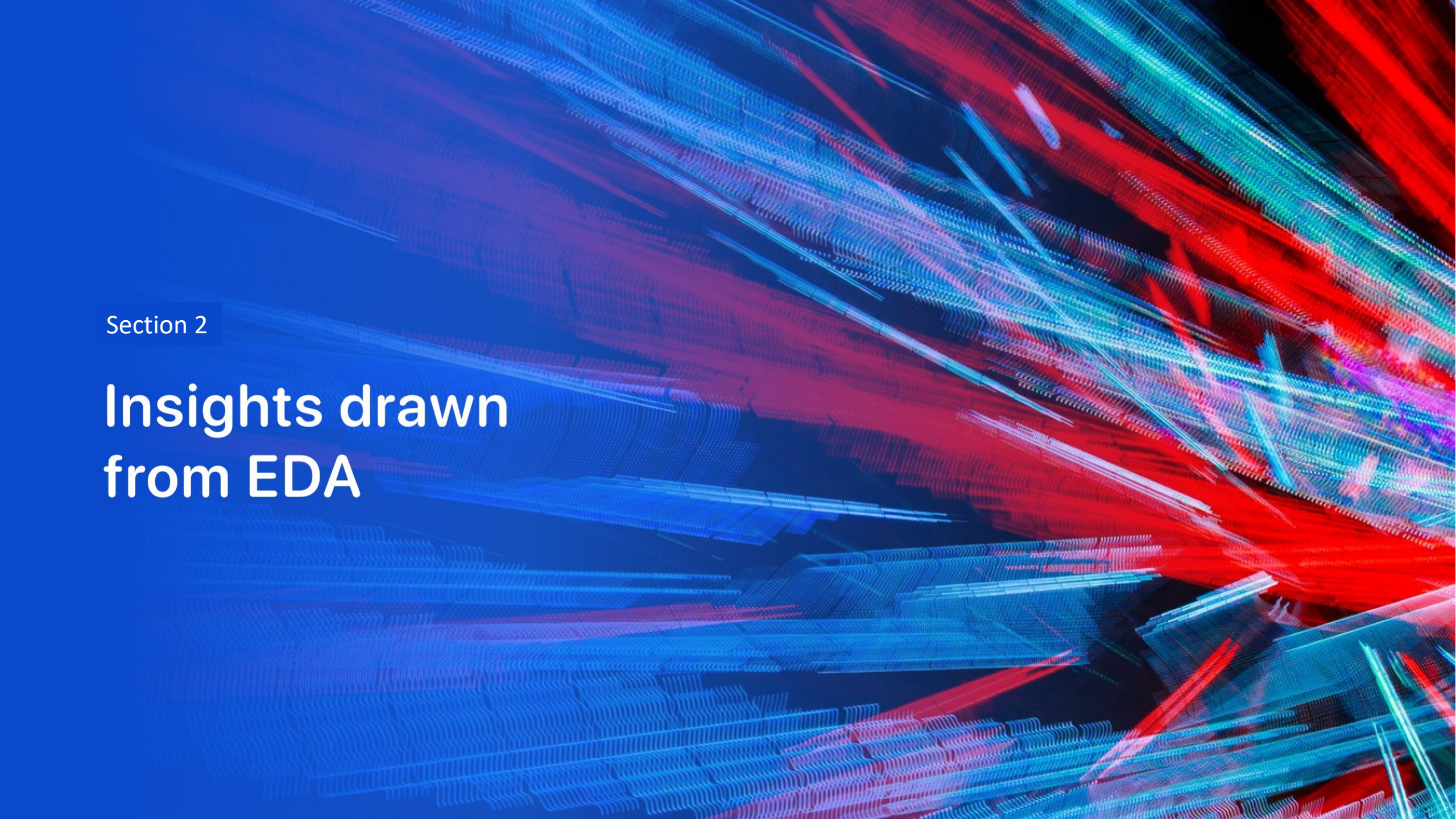
- Exploratory data analysis results
- Interactive analytics demo in screenshots





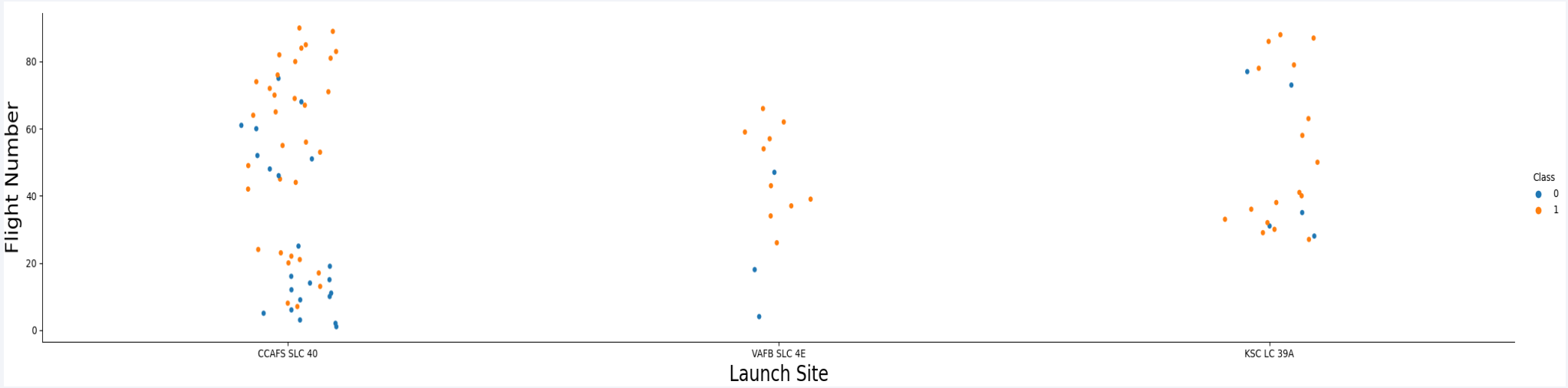
- Predictive analysis results

	score	best score
logreg	0.833333	0.846429
svm	0.833333	0.848214
tree	0.500000	0.875000
knn	0.833333	0.848214


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

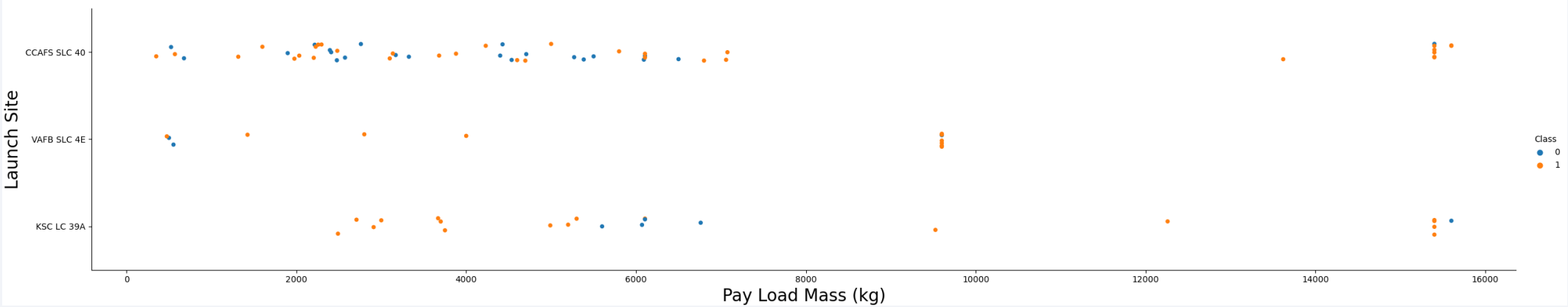
- Scatter plot of Flight Number vs. Launch Site



- CCAFS SLC 40 has more launches compare to other Launch Sites with flight number range from 0 to 80

Payload vs. Launch Site

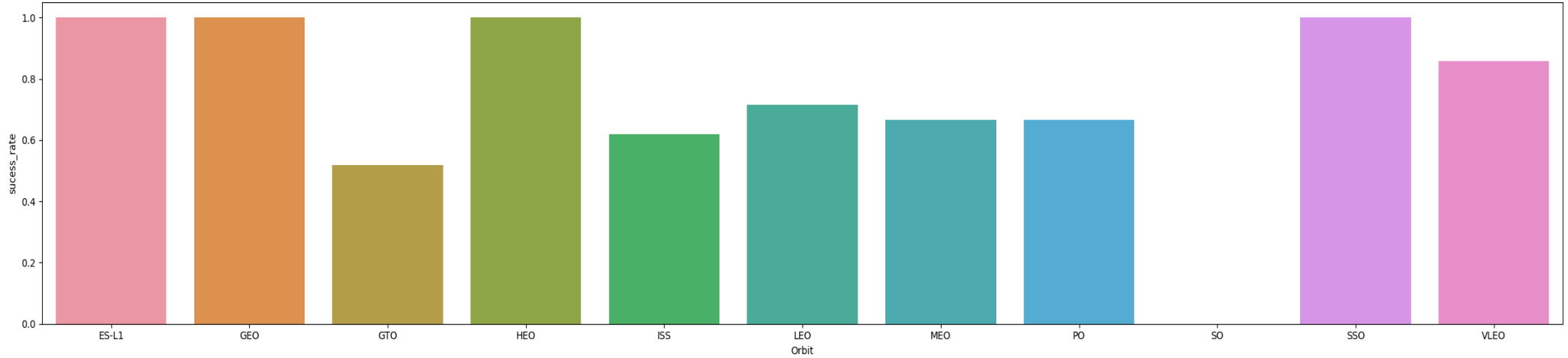
- Scatter plot of Payload vs. Launch Site



- VAFB-SLC launchsite has no rockets launched for payload mass greater than 10000 kg

Success Rate vs. Orbit Type

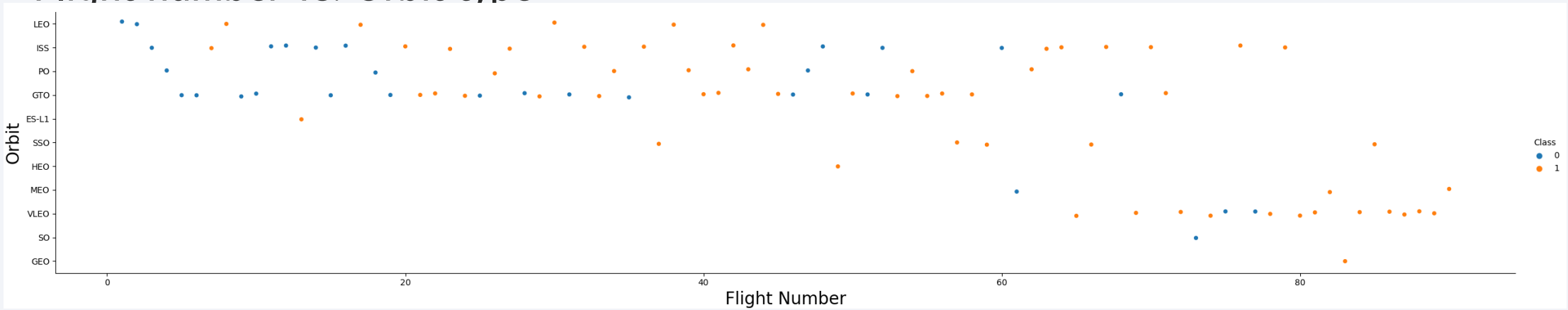
- Bar chart for the success rate of each orbit type



- Orbit type ES-L1, GEO and SSO have highest success Rate (1), while GTO has lowest success rate < 0.6

Flight Number vs. Orbit Type

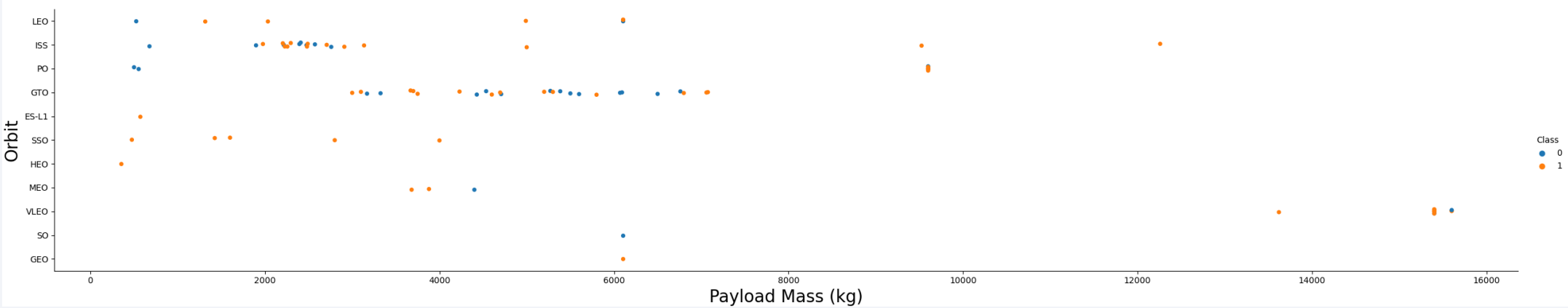
- Flight number vs. Orbit type



- LEO, ISS, PO, GTO orbit the Succes appears with the flight number between 0 -20
- With flight number range between 20-40 th GTO is only success orbit

Payload vs. Orbit Type

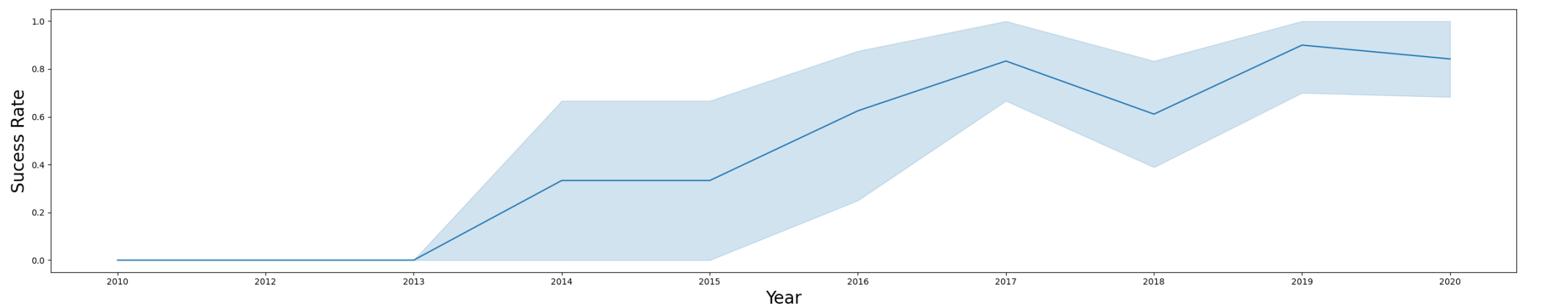
- Scatter point of payload vs. orbit type



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

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



- Sucess rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql Select Distinct Launch_Site From SPACEXTBL
```

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

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

- Displaying the launch Sites with 4 Unique Name in the space mission

Launch Site Names Begin with 'CCA'

In [120…

```
%sql Select * from SPACEXTBL Where Launch_Site Like '%CCA%' limit 5
```

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

Out[120…

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Display 5 records where launch sites begin with `CCA`

Total Payload Mass

```
In [47]: %sql select sum(PAYLOAD_MASS__KG_) as payloadmass_NASA_CRS_KG from SPACEXTBL WHERE CUSTOMER='NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[47]:
```

payloadmass_NASA_CRS_KG
45596

- Display the total payload mass carried by boosters launched by NASA (CRS) = 45596 kg

Average Payload Mass by F9 v1.1

```
In [48]: %sql select avg(PAYLOAD_MASS__KG_) as payloadmass_F9v11 from SPACEXTBL WHERE Booster_Version like '%F9 v1.1%'
 * sqlite:///my_data1.db
Done.
```

Out[48]:

payloadmass_F9v11
2534.6666666666665

Display the average payload mass carried by booster version F9 v1.1 = 2534 kg

First Successful Ground Landing Date

```
[13]: %sql select min(date) from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
[13]: min(date)
      01-05-2017
```

- The dates of the first successful landing outcome on ground pad is in 2017

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
[15]: %%sql select distinct Booster_Version from SPACEXTBL where "Landing _Outcome" like '%Success (drone ship)%'
and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

[15]:

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

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000kg but less than 6000kg, namely: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

```
[17]: %sql select Mission_Outcome, COUNT (*) as Total_Outcome from SPACEXTBL group by Mission_Outcome
 * sqlite:///my_data1.db
Done.
```

[17]:

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

- Total Number of failure Mission Outcome is 1, Success is 99 and Success (payload status unclear) is 1

Boosters Carried Maximum Payload

```
[109]: %sql select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_= (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

[109]:

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

There are 12 names of the booster which have carried the maximum payload mass

2015 Launch Records

```
[125]: %%sql select substr(Date, 4, 2) as Month_2015, "Landing _Outcome", Booster_Version, Launch_Site from SPACEXTBL
       where "Landing _Outcome"='Failure (drone ship)' and substr(Date,7,4)='2015'

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

[125]:

Month_2015	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Failed landing_outcomes in drone ship in year 2015 are from CCAFS LC-40.

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

```
[22]: %%sql SELECT ("LANDING _OUTCOME"), count (*) as "count_outcomes" FROM SPACEXTBL where
"Date" between '04-06-2010' and '20-03-2017'
group by "Landing _Outcome"
order by "count_outcomes" desc
```

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

[22]:

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

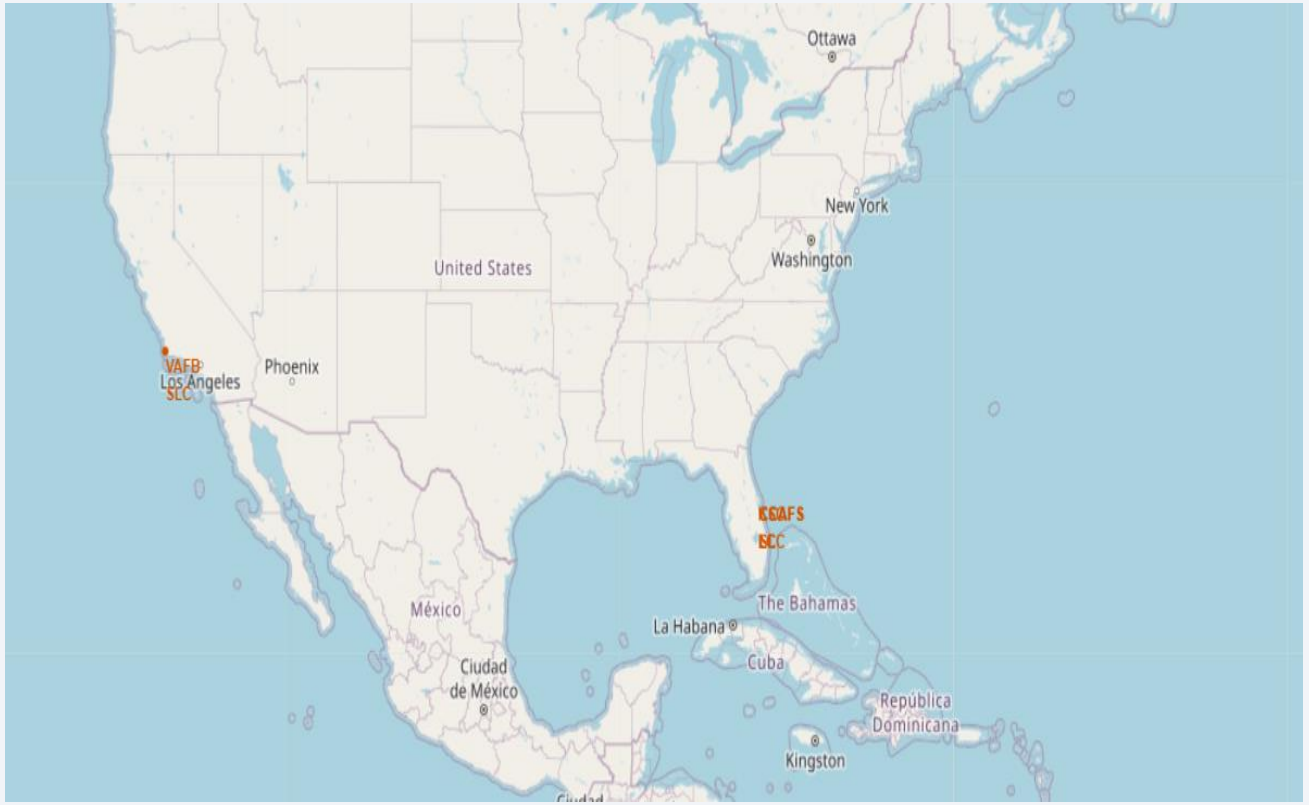
Explanation : Displaying Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

Section 3

Launch Sites Proximities Analysis

All Launch Sites in Global Map

- All launch site have close proximity to the coastline to minimize the rocket debris affect populated area



- Explain the important elements and findings on the screenshot

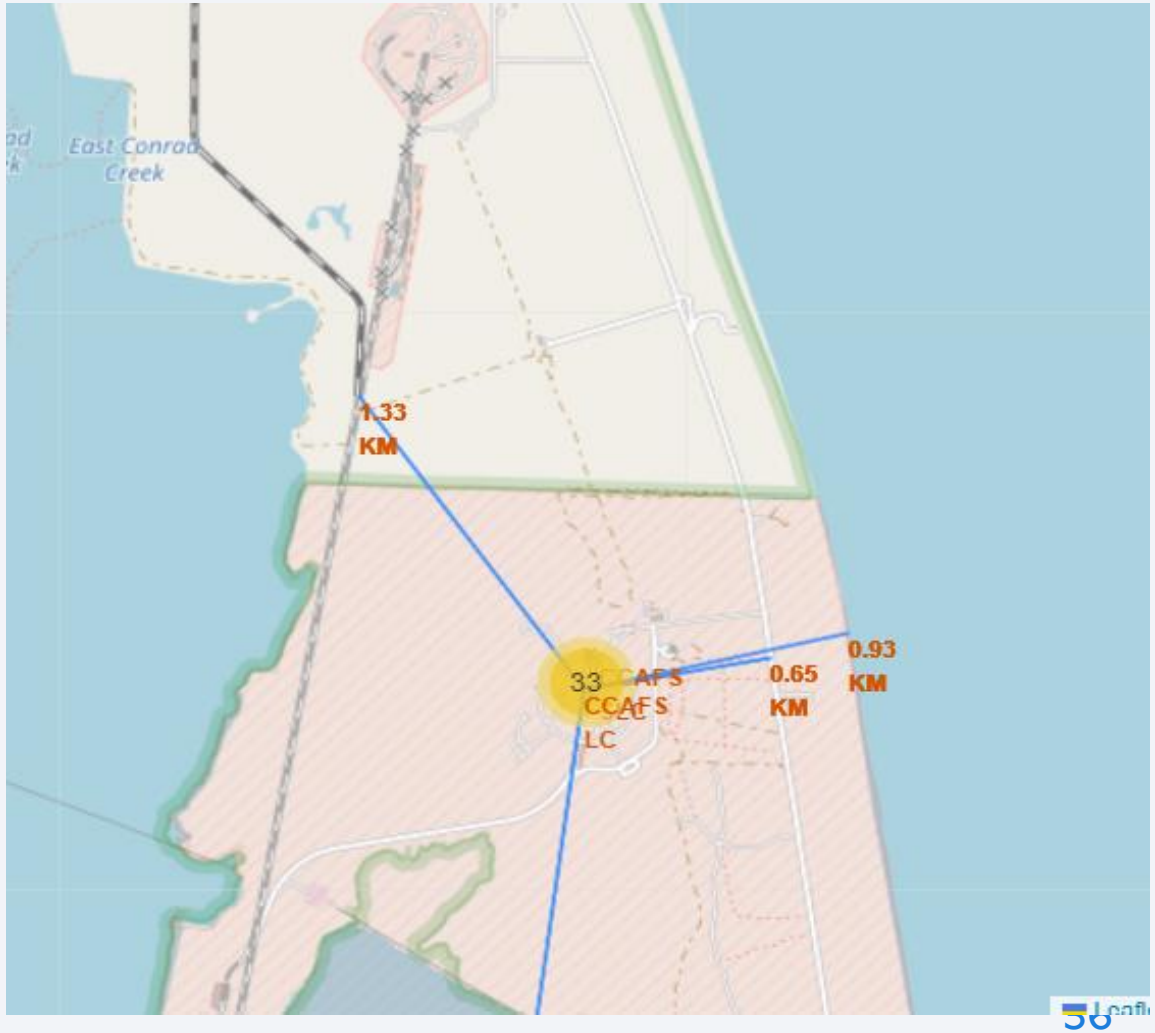
Color Labeled Launch Records

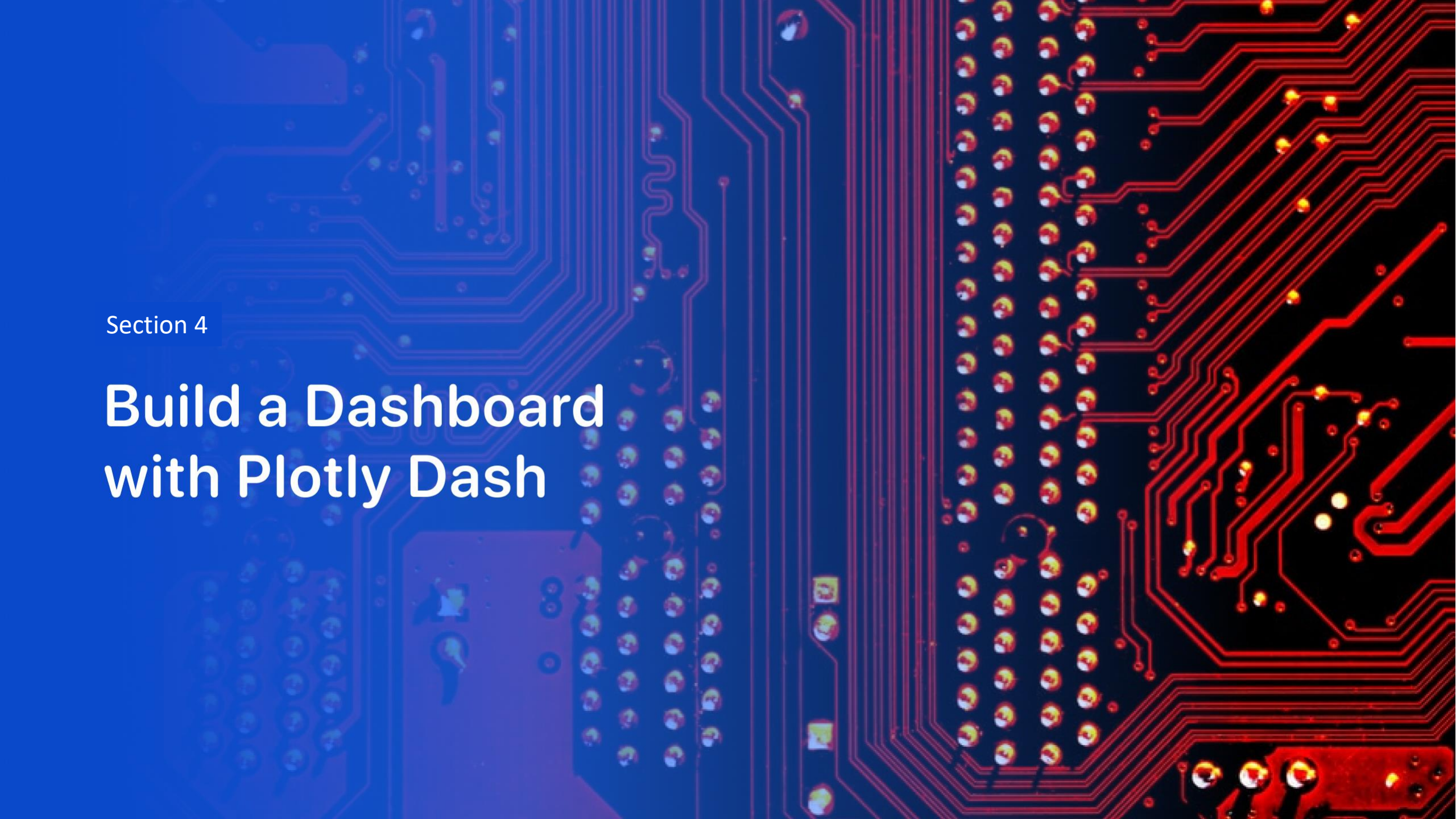
- We can identify the number of successful and unsuccessful launch for each site by the color:
  - Green Marker: Successful Launch
  - Red Marker : Unsucessful Launch
- For example, based on the map CCAF – LC has 26 total launches with 7 of them are sucessful

distance_highway = 0.6461880763256931 km distance_railroad = 1.3314378832303038 km distance_city = 51.32864612692812 km

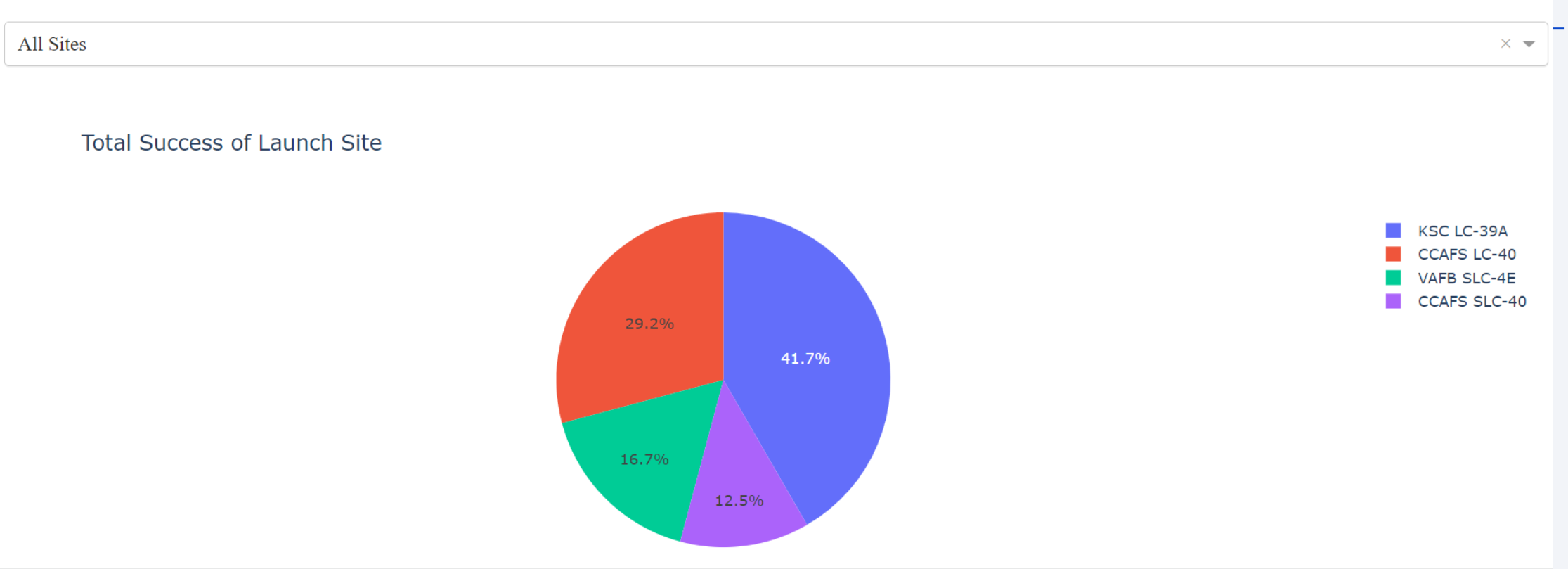
<Folium Map Screenshot 3>

- Closest distance_highway = 0.64 km
- Closest distance_railroad = 1.33 km
- Closest distance city = 51.3 km
- Closes distance costline = 0.93 km



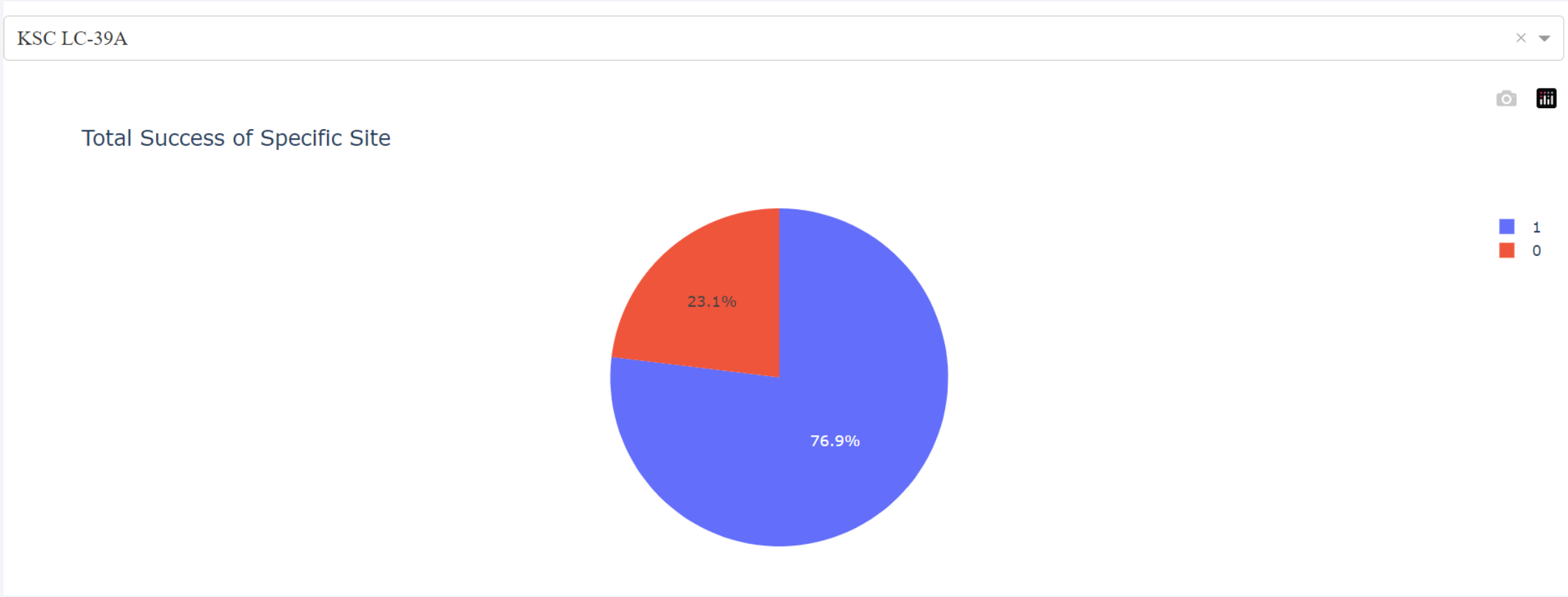

Section 4
Build a Dashboard with Plotly Dash

Total Success of All Launch Sites



- The Majority of successful launch Sites come from KSC-LC-39A with 41.7%, followed by CCAFS LC-40 29.2% and VAFB SLC-40 16.7%
- The least successful Launch Sites come from CCAFS-SLC-40 with 12.5%

Total Success of Specific Sites



- KSC-LC-39 has 76.9% success launch

Payload Mass Vs Launch Outcome for ALL Sites

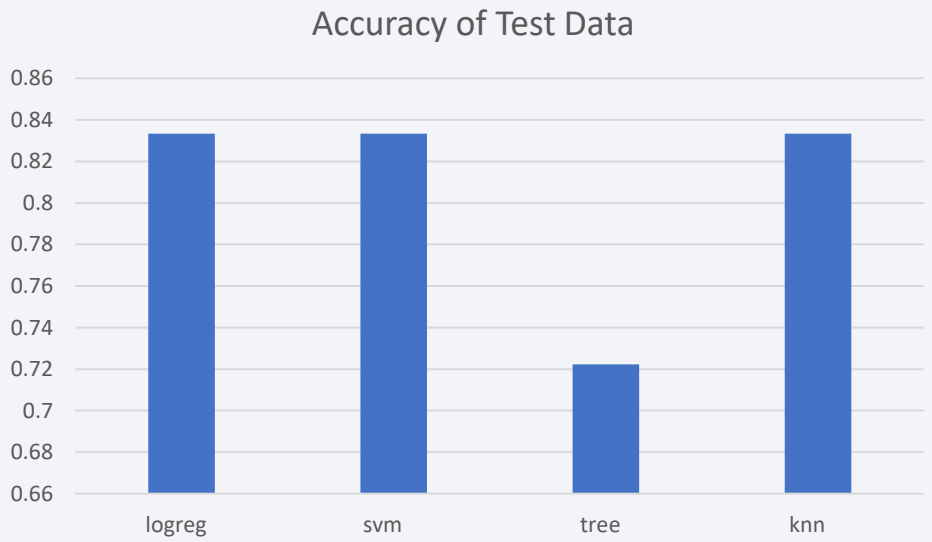
- The successful launch has payload mass range from 2000 kg to 5500 kg with FT booster version category

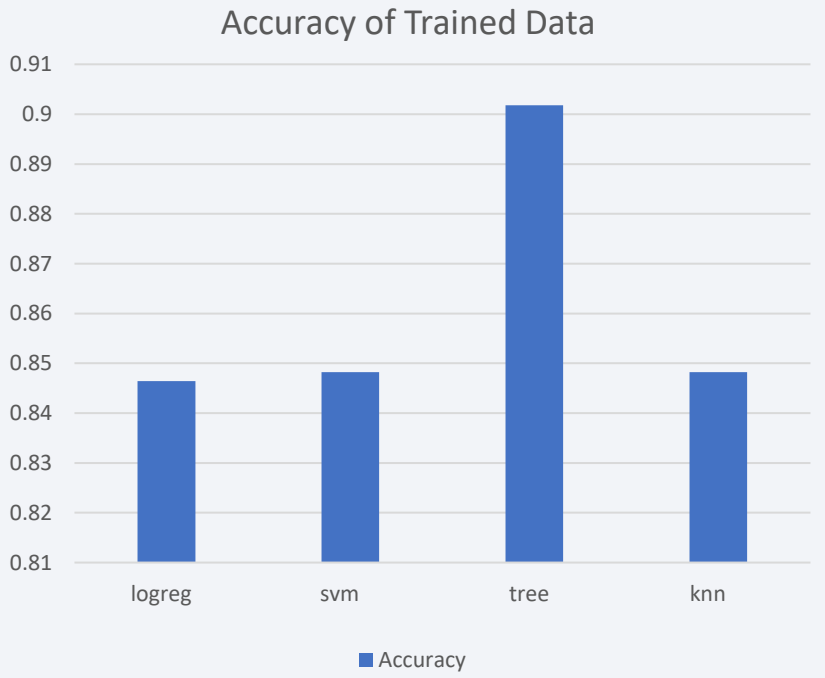
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree has lowest accuracy of test data but have highest accuracy for train data.
- The model can be considered the best model since for train accuracy it has significantly different with other model



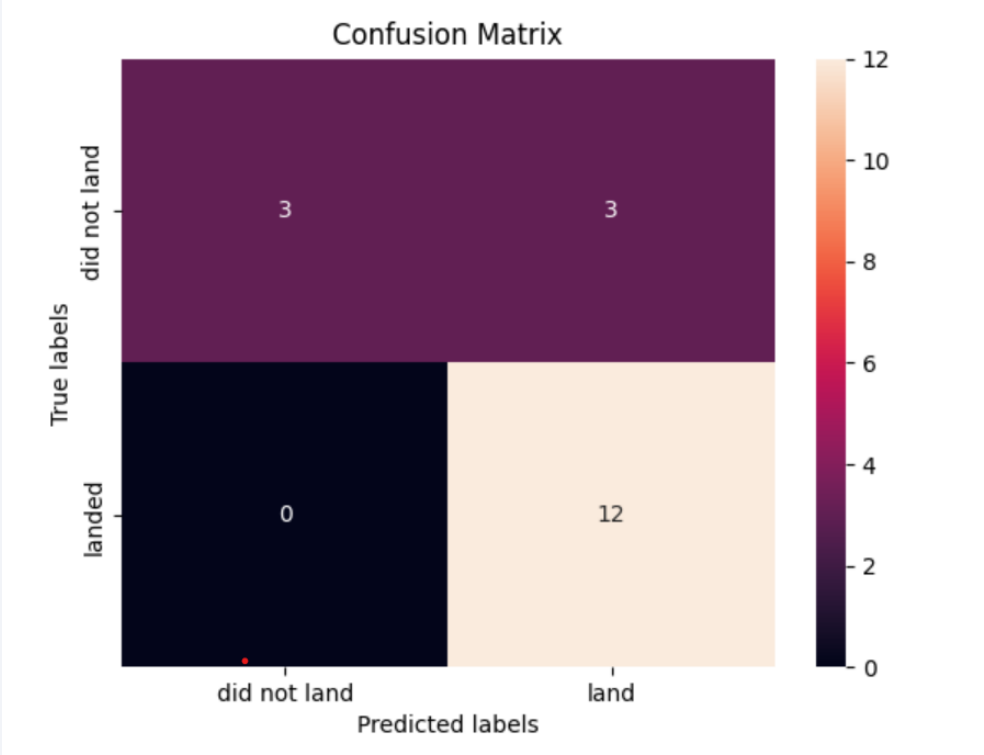



```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 4, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitte
r': 'best'}
accuracy : 0.9017857142857142
```

Confusion Matrix

- All the test accuracy for all model are equal
- Based on confusion matrix, the main problem is false positive

True Labels		
	TP	FP
	FN	TN
	Predicted values	



Conclusions

- Based on scatter plot CCAFS SLC 40 has more launches compare to other Launch Sites.
- However, the pie chart display the highest success rates among the others comes from KSC-LC-39A.
- The type of orbit also shows correlation for the success rates. Orbit type ES-L1, GEO and SSO has 100% success rate.
- The payloadMass also has positive impact towards success rate. The highest success rate is for the payloadMass between 2000 kg and 5500 kg with FT booster version category
- All launch site have close proximity to the coastline to minimize the rocket debris affect populated area
- The best model for predictive analysis is Decision Tree with accuracy of trained data is around 0.9

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!