IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gaganageethika Seerlancha
4th July,2024

Outline

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

Executive Summary

- ***Summary of methodologies:***
  - *Data Collection using API and Web Scraping*
  - *Data Wrangling*
  - *Exploratory Data Analysis using SQL, Pandas and Matplotlib*
  - *Interactive maps with Folium*
  - *Predictive Analysis for each Classification Model*
- ***Summary of all results:***
  - *Analyzing data through Interactive Visuals*
  - *Best Model for Predictive Analysis*

Introduction

- ***Project background and context***

-Predicting if Falcon9 first stage will land successfully.

-SpaceX advertises Falcon9 rocket launches on its website, with a cost of $62,000,000;

other providers cost upward of $165,000,000 each, much of the savings is because SpaceX can reuse the first stage.

-Therefore, if we can determine if the first stage will land successfully. This information can be used if an alternative company wants to bid against SpaceX for a rocket launch.

- ***Problems you want to find answers:***

- With what factors, the rocket will land successfully?

- Conditions which will aid to SpaceX to achieve the best results.

- Dependency of the landing outcomes on various variables.

Section 1

Methodology

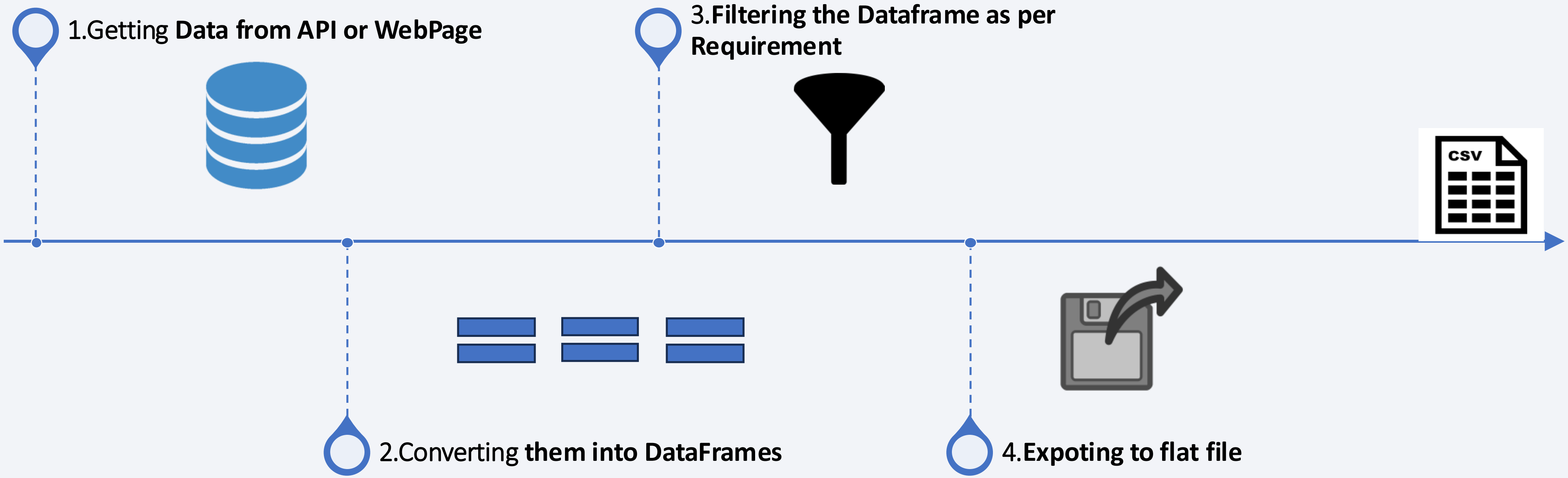
Methodology

Executive Summary

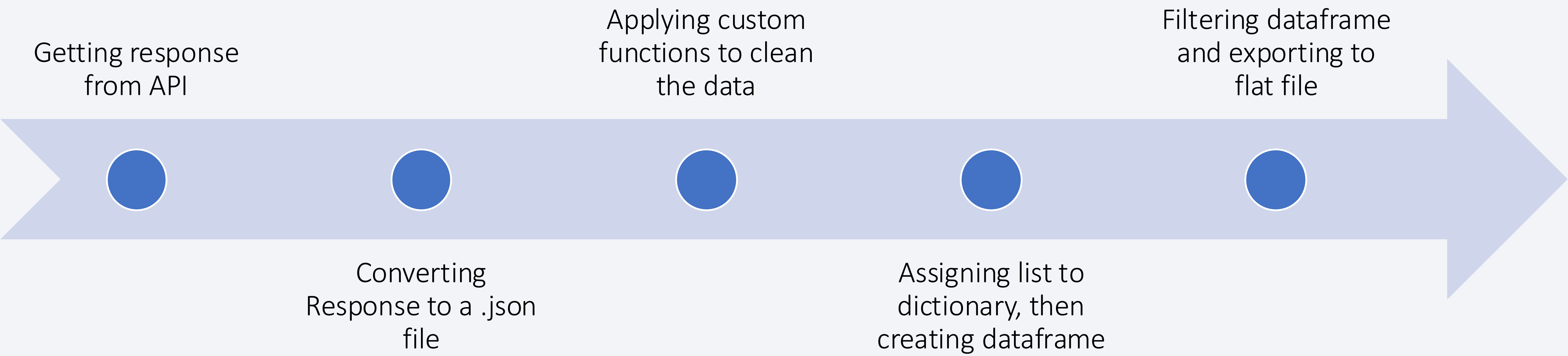
- *Data collection methodology*
    - *SpaceX REST API*
    - *Web Scraping from Wikipedia*
- *Perform data wrangling*
    - *One hot encoding data fields for machine learning and dropping irrelevant columns(Transforming data for the Analysis)*
- *Perform exploratory data analysis (EDA) using visualization and SQL*
- *Perform interactive visual analytics using Folium and Plotly Dash*
- *Perform predictive analysis using classification models*
    - *Building and evaluating classification models for predictive analysis*

Data Collection

- *The process of collecting and evaluating information or data from multiple sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.*
- *It is an essential phase in all kinds of research, analysis and decision-making.*



Data Collection – SpaceX API



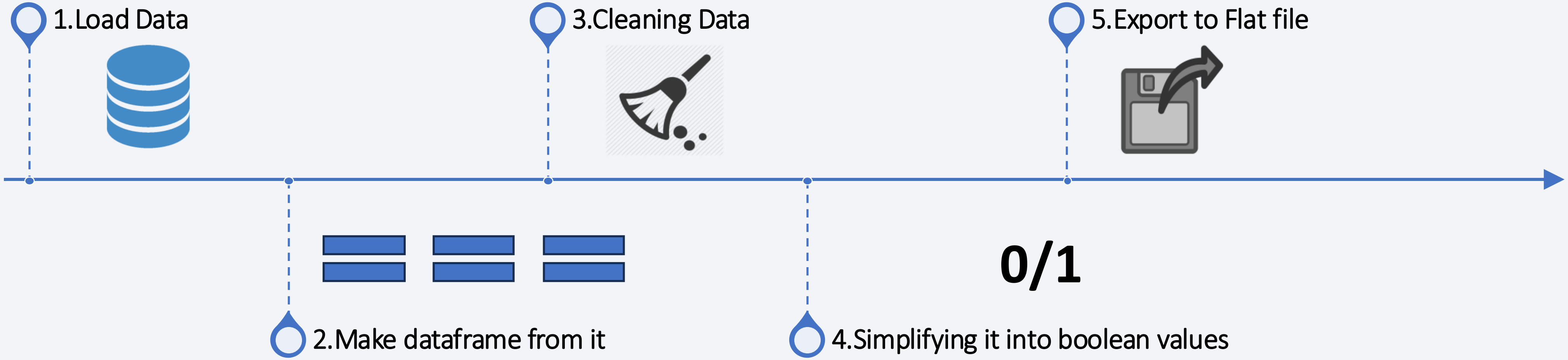
GitHub link for Data Collection using SpaceX API: [GitHub URL]

Data Collection - Scraping

GitHub link for Data Collection using Web Scraping: GitHub URL

Data Wrangling

- *Data wrangling enables you to gather data from multiple sources into a central spot.*
- *Cleaning and converting data into a standard format enables you to perform cross-data set analytics.*
- *Data wrangling prepares data by removing flawed and missing elements, readying it for data mining, and empowering businesses to make concrete, data-driven decisions.*

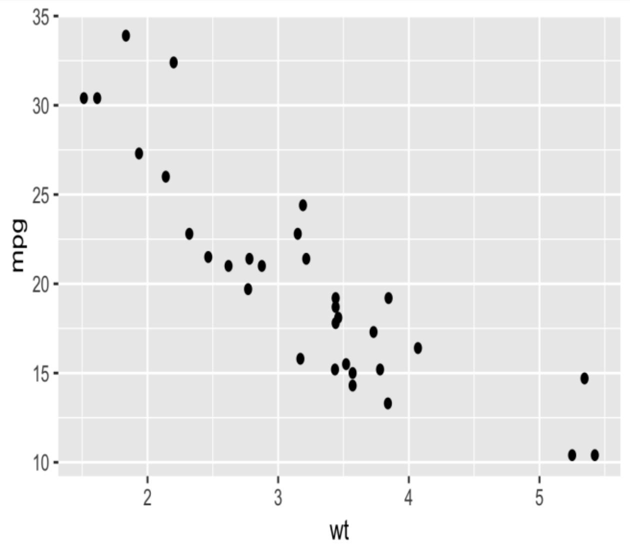


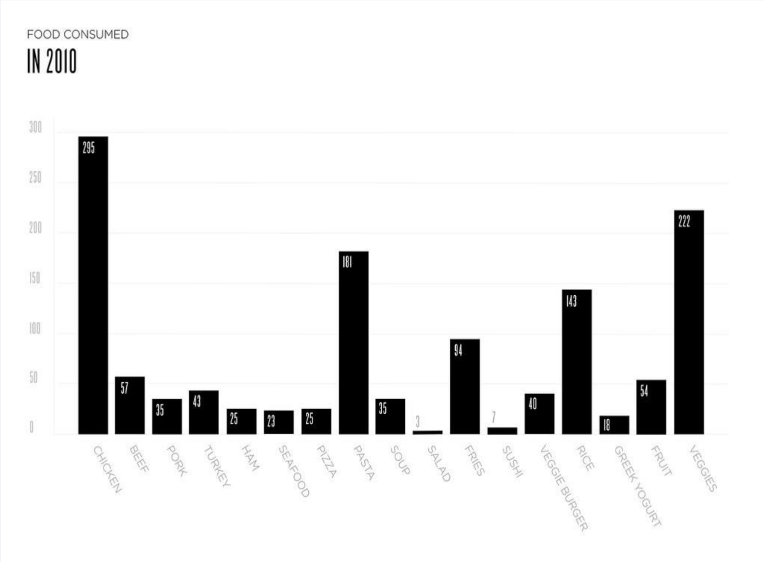
GitHub link for Data Wrangling: GitHub URL

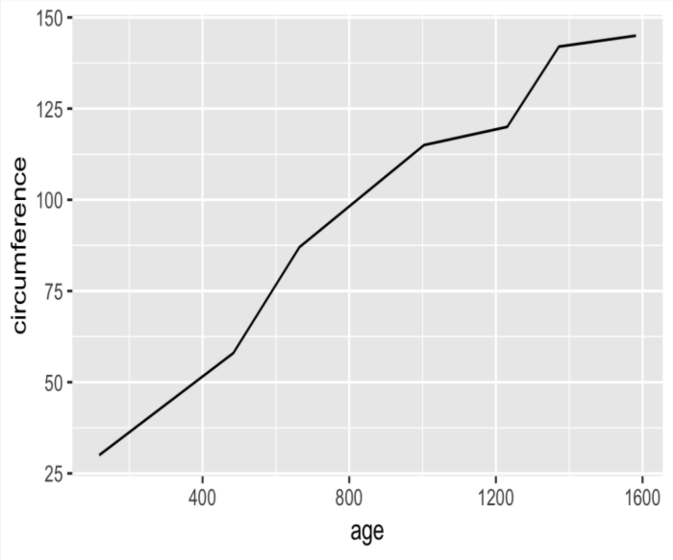
EDA with Data Visualization

Exploratory Data Analysis is an approach of analyzing data sets to summarize their main characteristics, using statistical graphs and other visualization methods

- **Scatter Plots** *show dependency of attributes on each other*
- **Bar Graphs** *are easiest to interpret a relationship between attributes*
- **Line Graphs** *are useful in showing trends clearly and aiding in predictions for future*







GitHub link for EDA with Data Visualization: [GitHub URL]

EDA with SQL

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass using a subquery
- Listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- 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.

GitHub link for EDA with SQL: GitHub URL

Building an Interactive Map with Folium

***Folium** is a powerful Python library that helps you create several types of Leaflet maps.*

Map Objects	Code	Use
Map Marker	folium.Marker()	Map object to make a mark on map
Icon Marker	folium.Icon()	Create an icon on map
Circle Marker	folium.Circle()	Create a circle where marker is being placed
PolyLine	folium.PolyLine()	Create a line between points
Marker Cluster Object	MarkerCluster()	Simplify a map containing many markers with same coordinates
AntPath	folium.plugins.AntPath()	Create an animated line between points

GitHub link for Interactive Maps with Folium: GitHub URL

Build a Dashboard with Plotly Dash

- **Pie Chart** *showing the total success for all launch sites or by a certain launch site.*
    - *Percentage of success in relation to launch site*
- **Scatter Plot** *showing the correlation between Payload and Success for all sites or by certain launch site*
    - *It shows relationship between SuccessRate and BoosterVersion category*

Map Object	Code	Use
Dash and its components	*import* dash *import* dash_html_components as html *import* dash_core_components as dcc *from* dash_dependencies *import* Input, Output	*The Dash core component library contains a set of higher level components like sliders, graphs, tables and more. Dash provides all the available HTML tags as user-friendly Python classes*
Pandas	*import* pandas as pd	*Fetching values from CSV and creating a dataframe*
Plotly	*import* plotly.express as px	*Plot the graphs with interactive plotly library*
Dropdown	dcc.Dropdown()	*Create a dropdown for launch sites*
Rangeslider	dcc.RangeSlider()	*Create a rangeslider for Payload mass range selection*
Pie Chart	px.pie()	*Create a pie graph for success percentage display*
Scatter Plot	px.scatter	*Create a scatter plot for correwlation display*

GitHub link for Dashboard: [GitHub URL]

Predictive Analysis (Classification)

Building Model

- Load our feature engineered data into data frame
- Transform it into NumPy arrays
- Standardize and transform data
- Split data into training and test data sets
- Check how many test samples have been created
- List down Machine Learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our data sets into the GridsearchCV objects and train our model

Evaluating Model

- Check accuracy for each model
- Get best hyperparameters for each type of algorithm
- Plot Confusion Matrix

Finding best performing Classification Model

- The model with best accuracy score wins the best performing model

GitHub link for Predictive Analysis: GitHub URL

Results

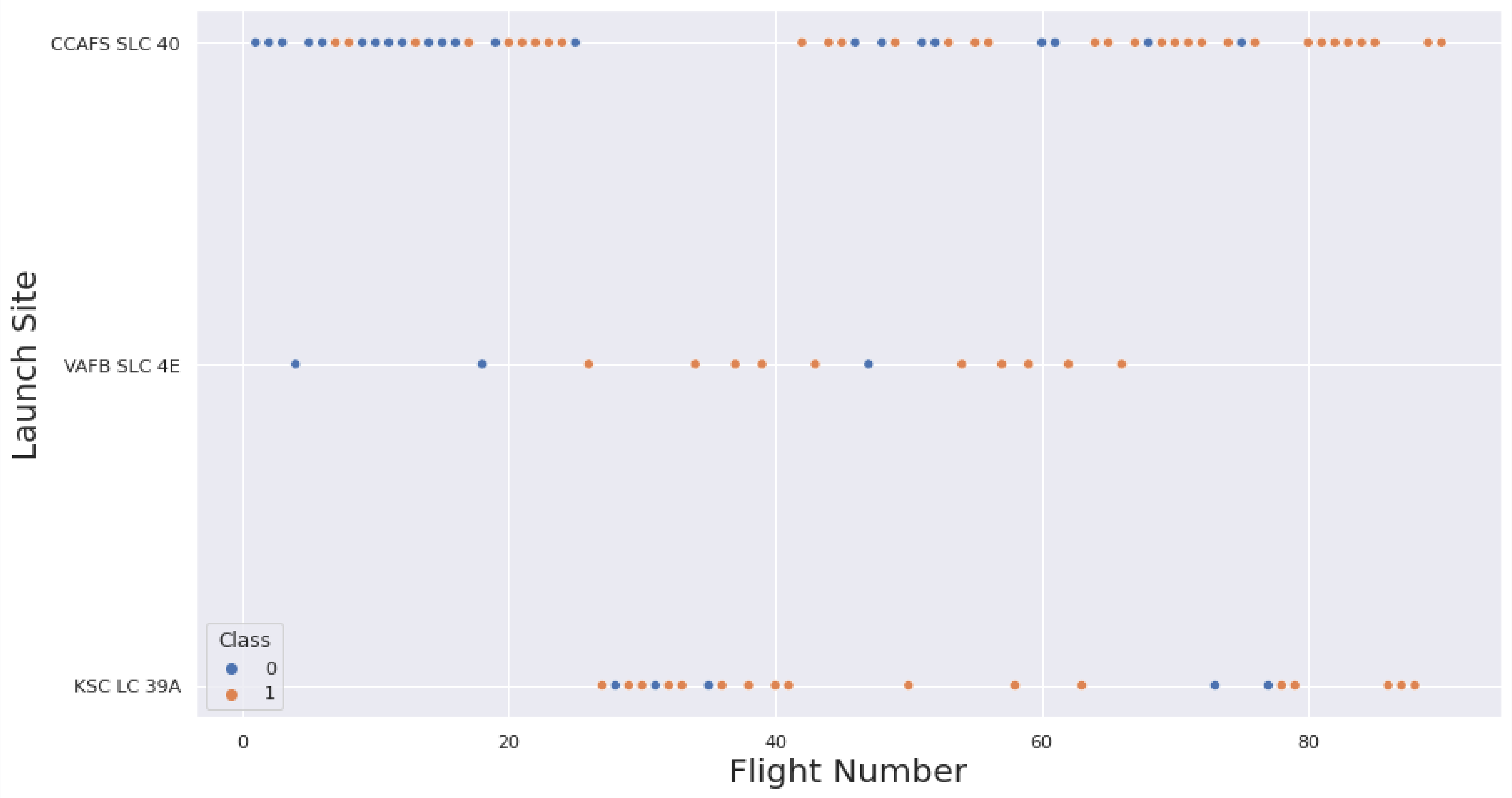
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

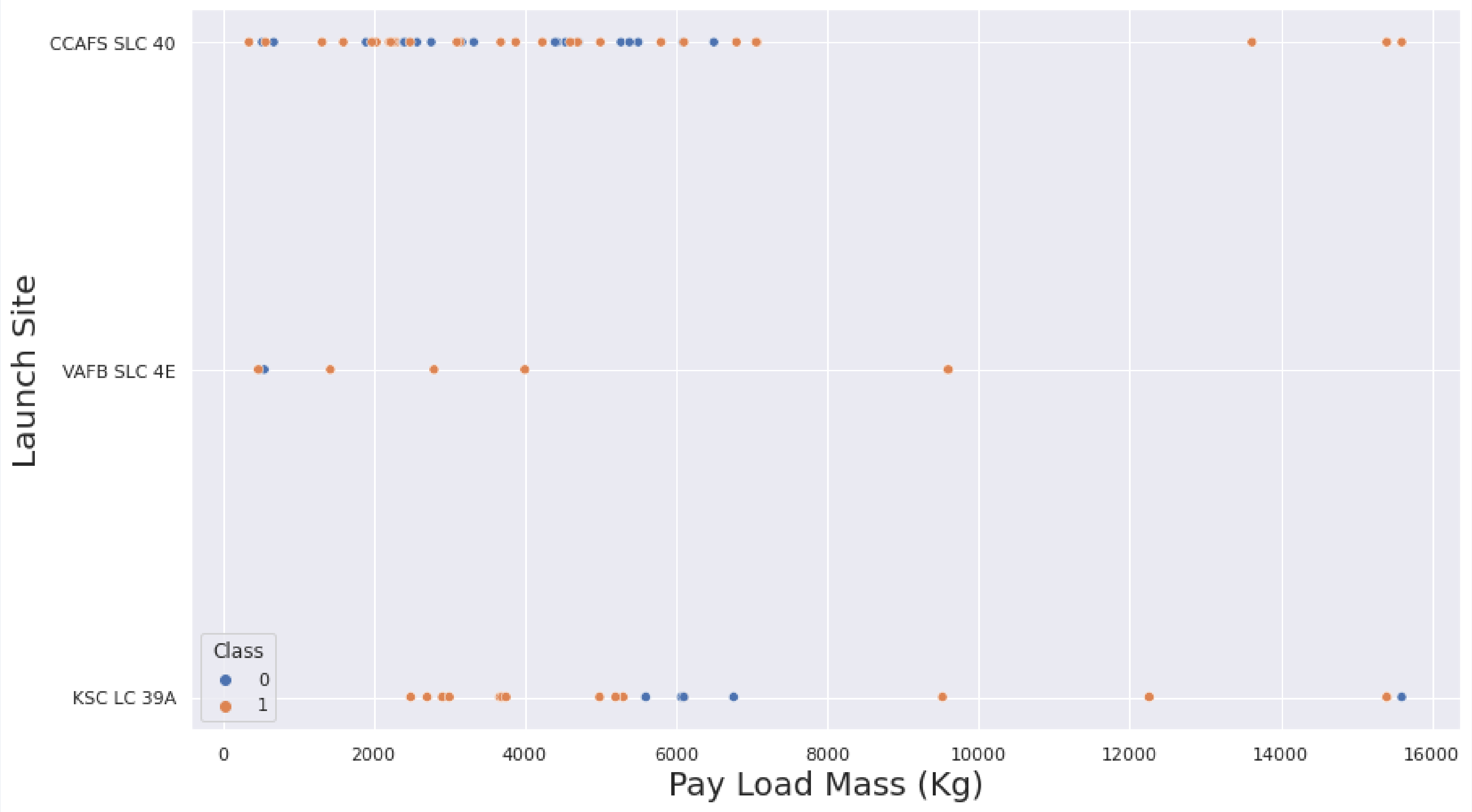
Flight Number vs. Launch Site

- *With higher Flight Numbers(greater than 30) the success rate for rocket is increasing.*



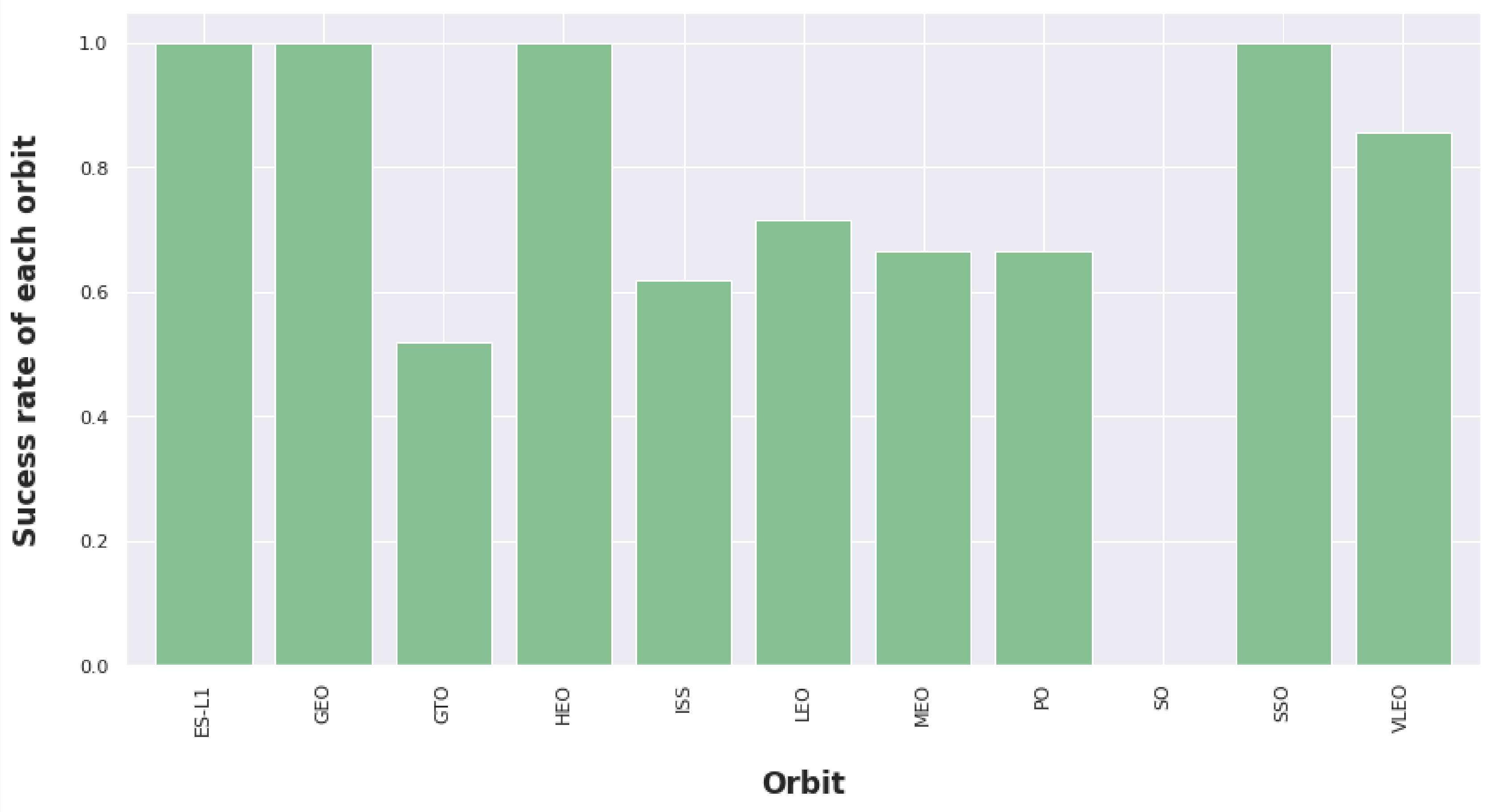
Payload vs. Launch Site

- *The greater the Pay Load mass(greater than 7000 kg) higher the success rate for the rocket. But there's no clear pattern to take a decision, if the launch site is dependent on Pay Load mass for a successful launch*



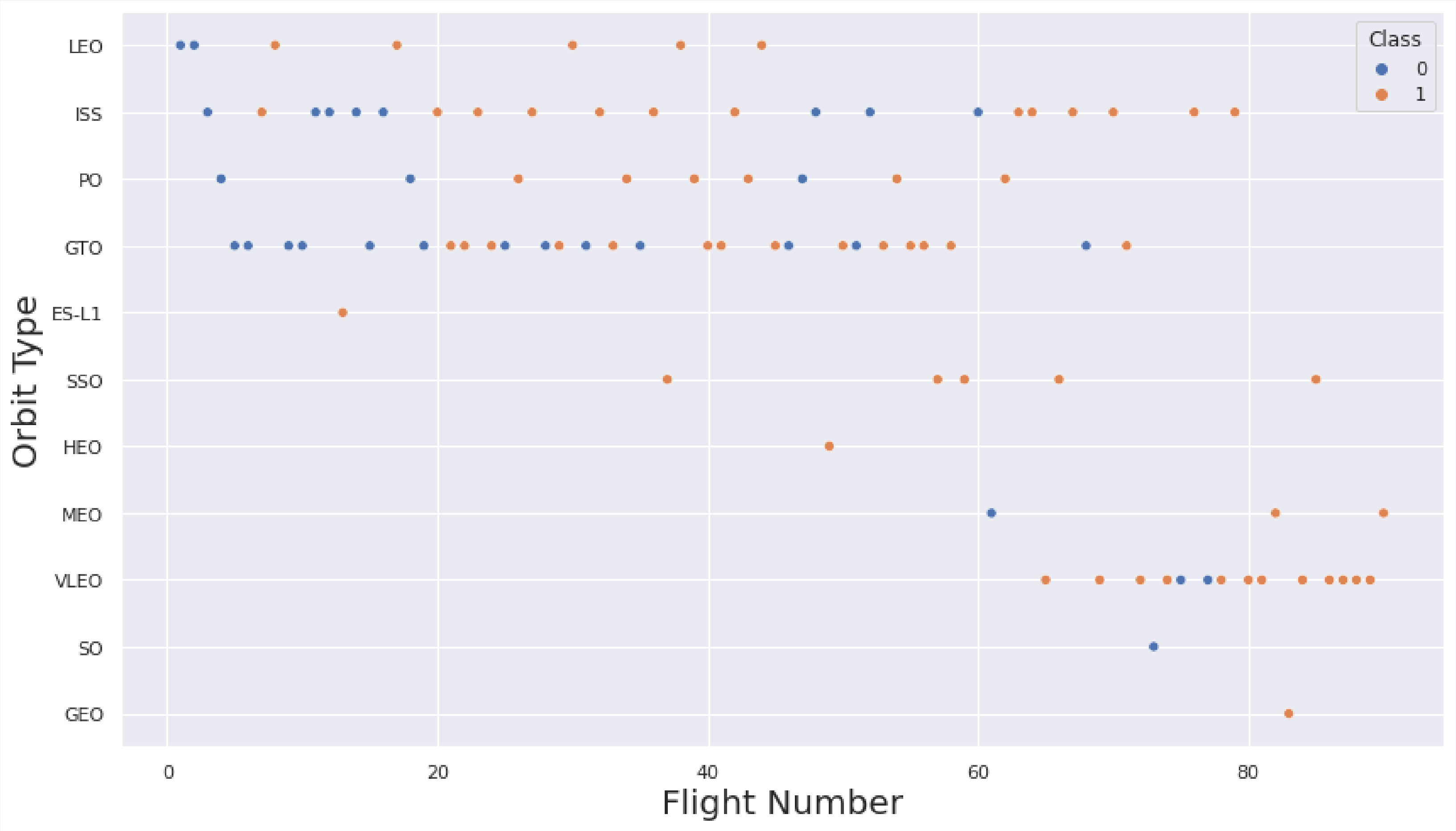
Success Rate vs. Orbit Type

- *The Orbit Types ES-L1, GEO, Heo and SSO have the Highest Success Rates*



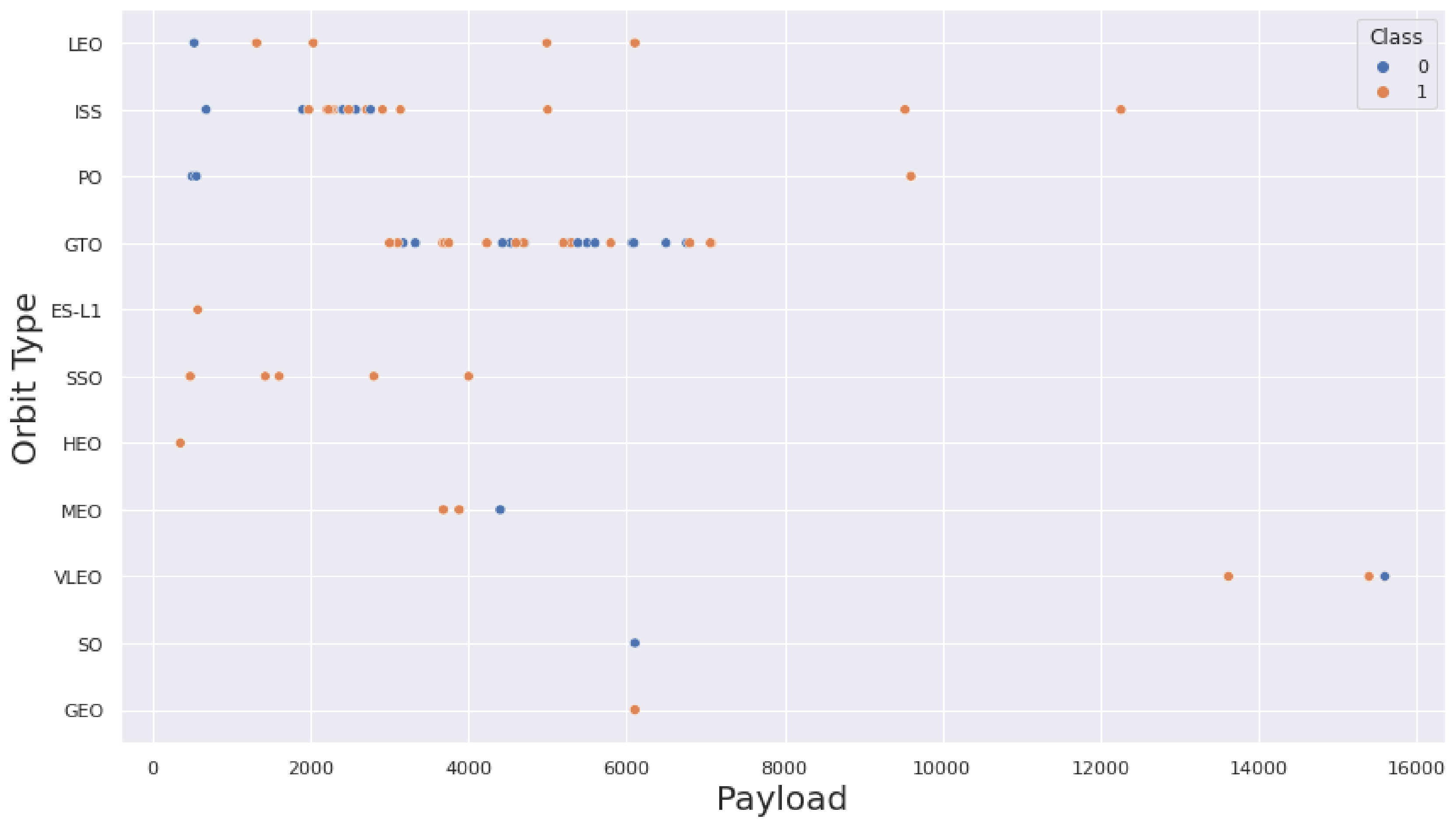
Flight Number vs. Orbit Type

- *We see that for LEO Orbit, the success increases with the number of flights*
- *On the other hand, there seems to be no relationship between Flight Number and the GTO Orbit*



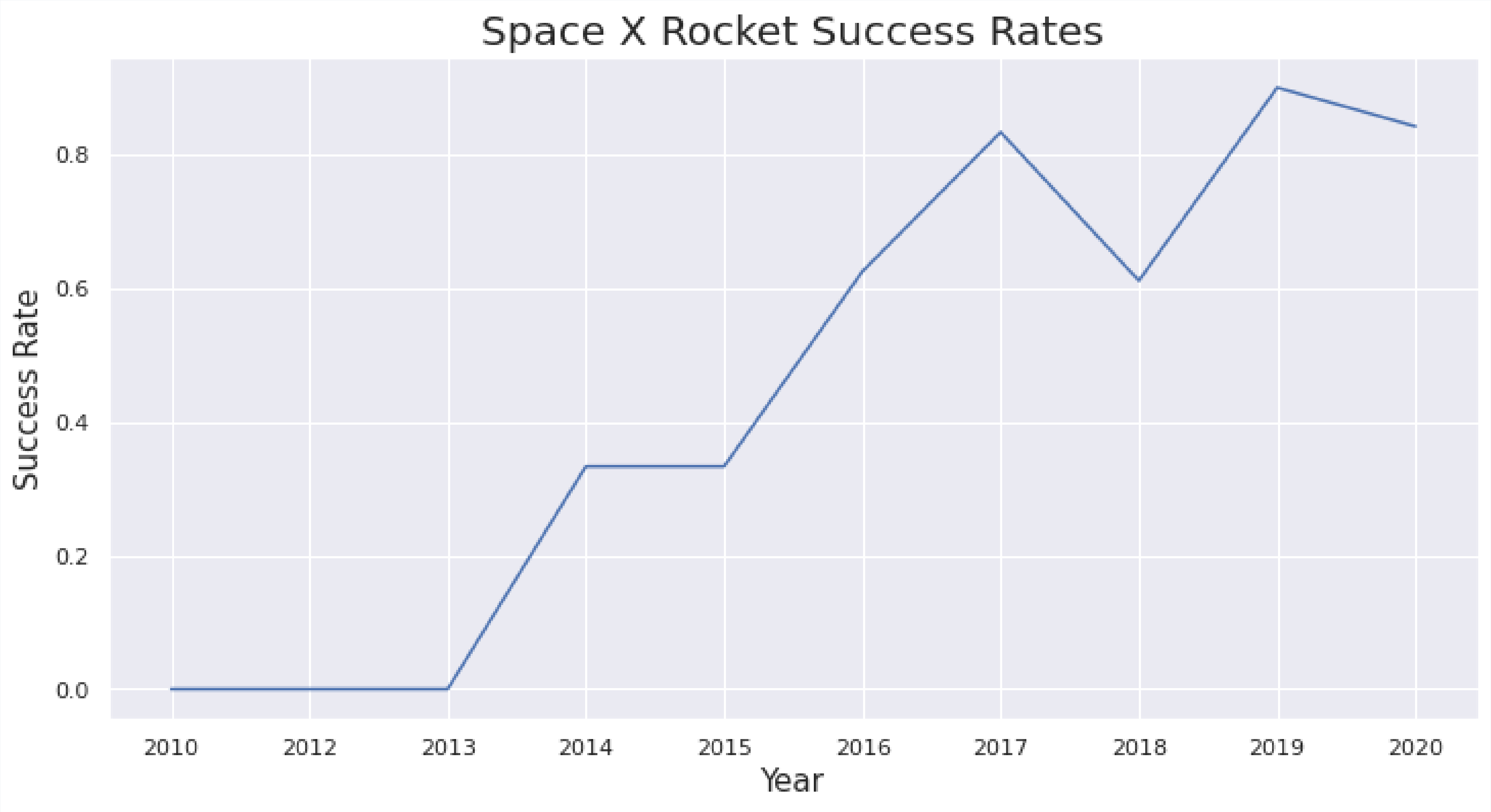
Payload vs. Orbit Type

- *We see that heavy payloads have negative influence on MEO, GTO and VLEO Orbits, but positive influence on LEO and ISS Orbits*



Launch Success Yearly Trend

- *We can see that the success rate since 2013 kept increasing relatively though there is a slight dip after 2019*



All Launch Site Names

- ***Query:*** `%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;`
- ***Explanation:*** *Using the word DISTINCT, we pull unique values for Launch_Site column from the table SPACEXTBL*
- ***Result:***

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

Launch Site Names Begin with 'CCA'

- ***Query:*** `%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;`
- ***Explanation:*** *Using the keyword LIMIT 5, we fetch 5 records from the table SPACEXTBL and with the condition LIKE keyword with wildcard 'CCA%'.*
- ***Result:***

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

- ***Query:*** `%sql SELECT SUM (PAYLOAD_MASS__kg_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA(CRS)' ;`
- ***Explanation:*** *Displaying the total payload mass carried by boosters launched by NASA (CRS)*
- ***Result:***

SUM (PAYLOAD_MASS__kg_)
None

Average Payload Mass by F9 v1.1

- ***Query:*** `%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version='F9 v1.1';`
- ***Explanation:*** Displaying average payload mass carried by booster version F9 v1.1
- ***Result:***

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- ***Query:*** `%sql SELECT MIN (DATE) AS "First Successful Landing" FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';`
- ***Explanation:*** *Listing the date when the first succesful landing outcome in ground pad was acheived.*
- ***Result:***

First Successful Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- ***Query:***

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone
ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

- ***Explanation:*** *Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000*

- ***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 sum(case when MISSION_OUTCOME LIKE '%Success%' then 1 else 0 end) AS "Successful
Mission", \

    sum(case when MISSION_OUTCOME LIKE '%Failure%' then 1 else 0 end) AS "Failure Mission" \

FROM SPACEXTBL;
```

- ***Explanation:*** *Listing the date when the first successful landing outcome in ground pad was achieved*

- ***Result:***

Successful Mission	Failure Mission
100	1

Boosters Carried Maximum Payload

- ***Query:***

```
%sql SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass"
FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

- ***Explanation:*** *Listing the names of the booster_versions which have carried the maximum payload mass.*

- ***Result:***

Booster Versions which carried the Maximum Payload Mass
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
SPACEXTBL WHERE DATE LIKE '2015-%' AND LANDING_OUTCOME = 'Failure (drone ship)';
```

- ***Explanation:*** Listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

- ***Result:***

MONTH	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

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

- ***Query:***

```
%sql SELECT LANDING_OUTCOME as "Landing Outcome", COUNT(LANDING_OUTCOME) AS "Total Count" FROM SPACEXTBL \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY  LANDING_OUTCOME \
ORDER BY COUNT(LANDING_OUTCOME) DESC ;
```

- ***Explanation:***
Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) … 10-06-04 and 2017-03-20, in descending order.

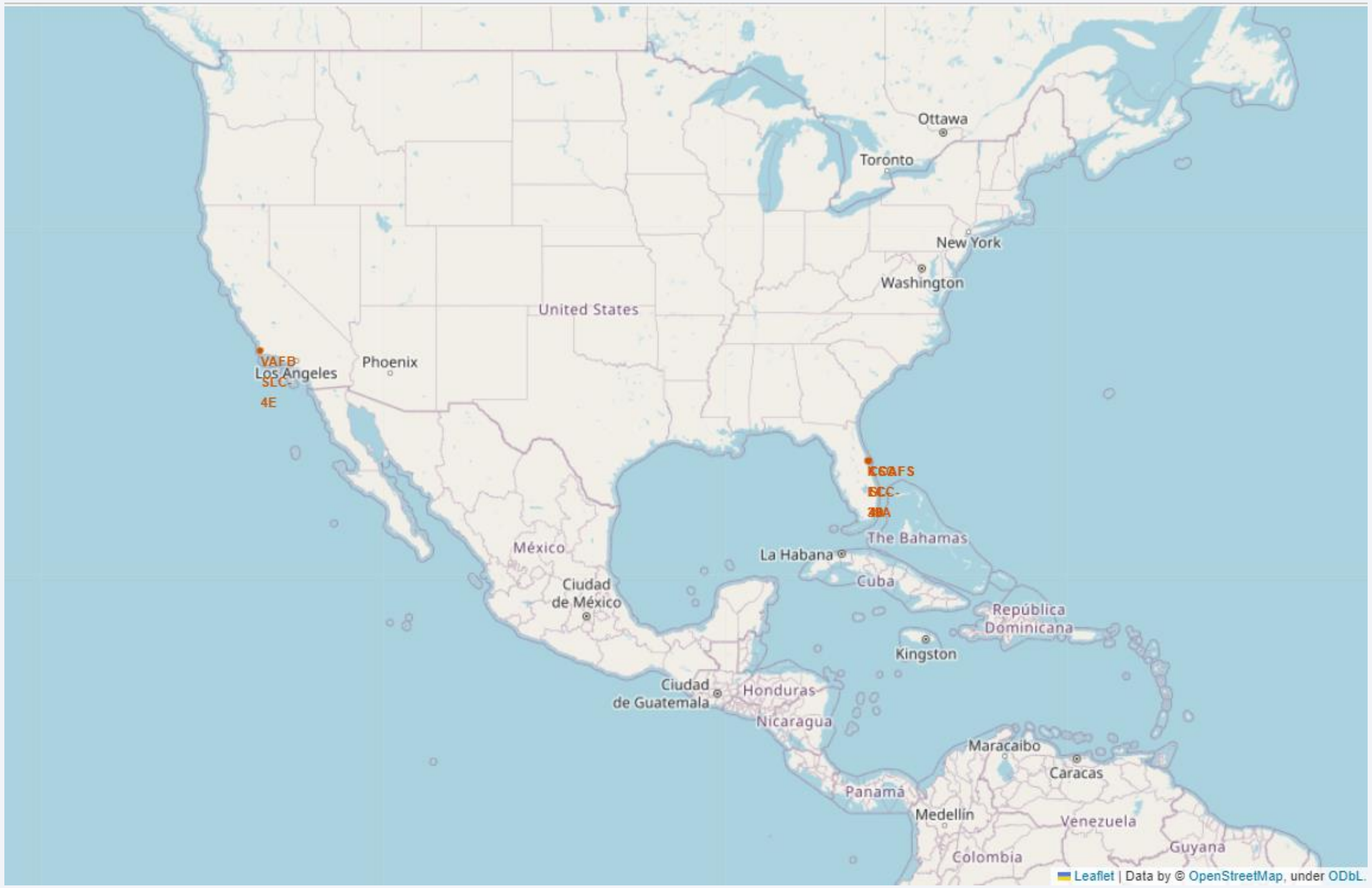
- ***Result:***

Landing Outcome	Total Count
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

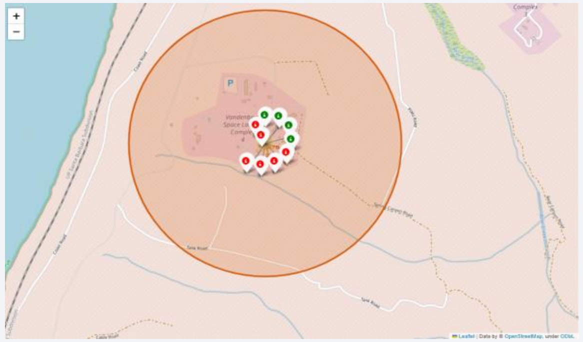
Folium Map: All Launch Sites



We can see that the SpaceX launch sites are near to the USA coast lines, i.e., Florida and California regions

Folium Map: Color-labeled Launch Outcomes

CCAFS SLC-40



VAFB SLC-4E

CCAFS LC-40

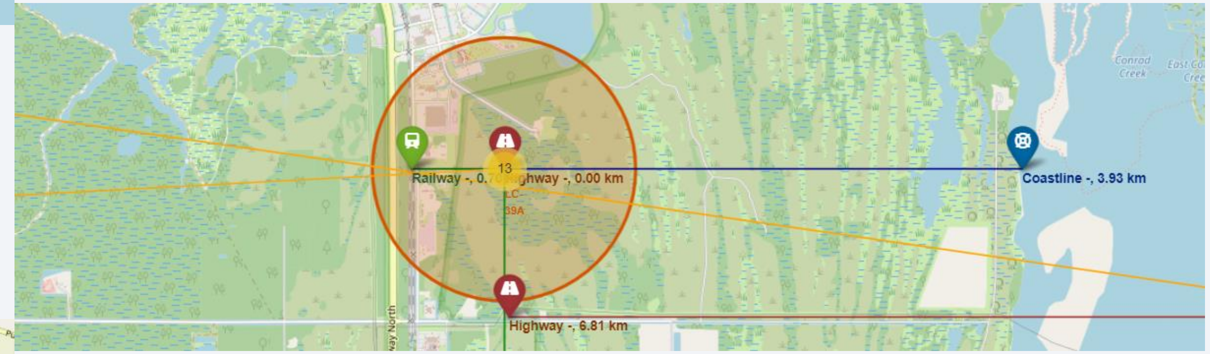
KSC LC-39A

- *Green Marker indicates successful launches whereas the Red Marker indicates all the Failure outcomes.*

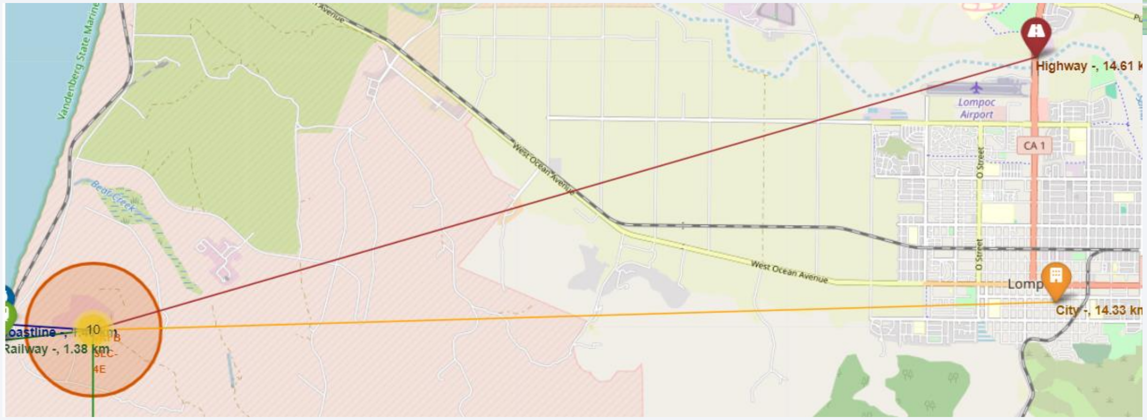
- *From these pictures, we can say that KSC LC-39A has the maximum probability of success .*

Folium Map: Launch Site Proximities

Distance for all launch sites are less than 4km from the Coastlines.



Distance for all launch sites from railway tracks are greater than 0.7km, i.e., they are not far away from the railway tracks.

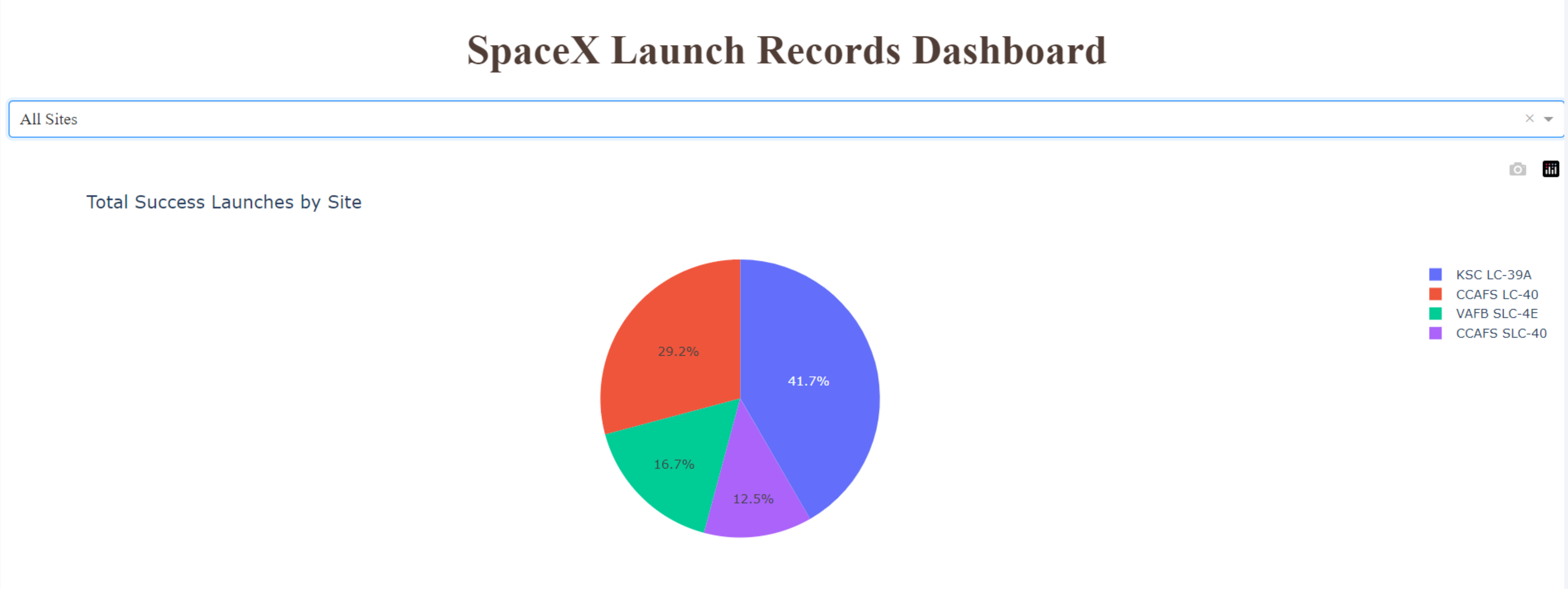


Distance for all launch sites are greater than 5km from the Highways and greater than 14km from the cities, therefore, relatively farther away.

Section 4

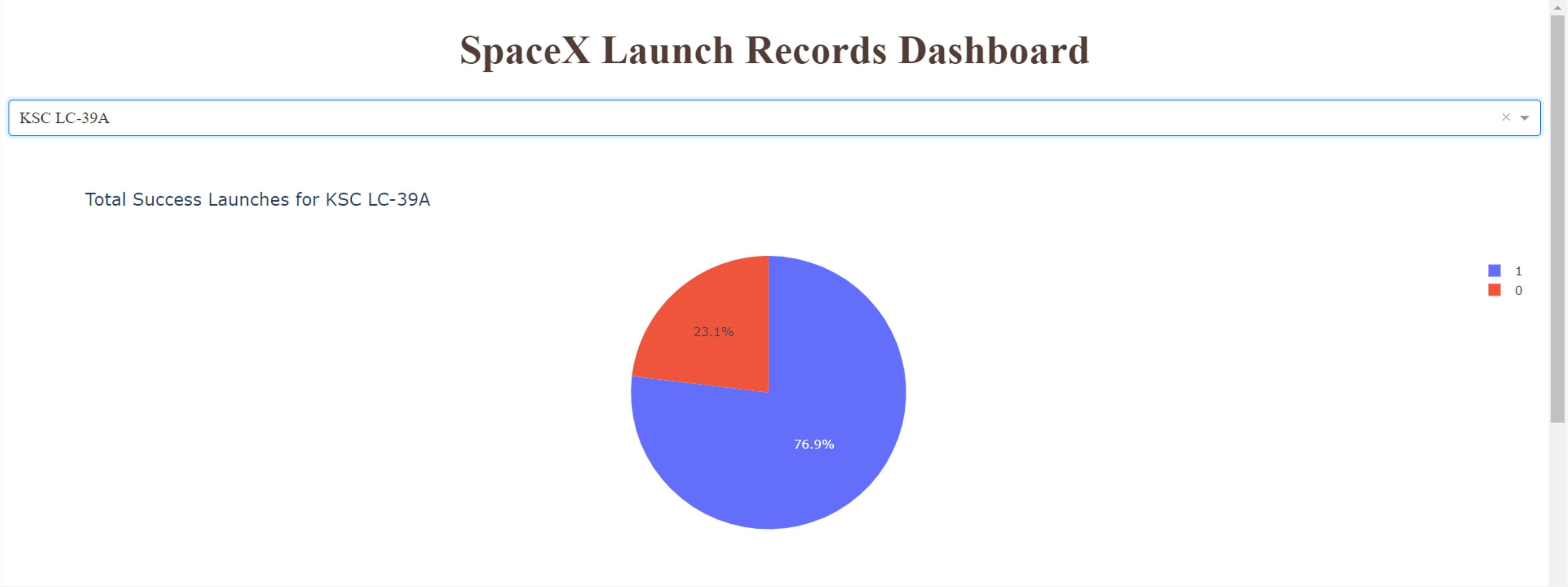
Build a Dashboard with Plotly Dash

Successful Launches Count for All Launch Sites



Findings: *KSC LC-39A has the most number of successful launches (blue color in the pie chart)*

Launch Site with Highest Launch Success Ratio



<u>Findings</u>: *76.9% of launches at KSC LC-39A have been successful.*

Payload vs. Launch Outcome Scatter Plot

=> *Low-weighted Payload(0-4000 kgs)*

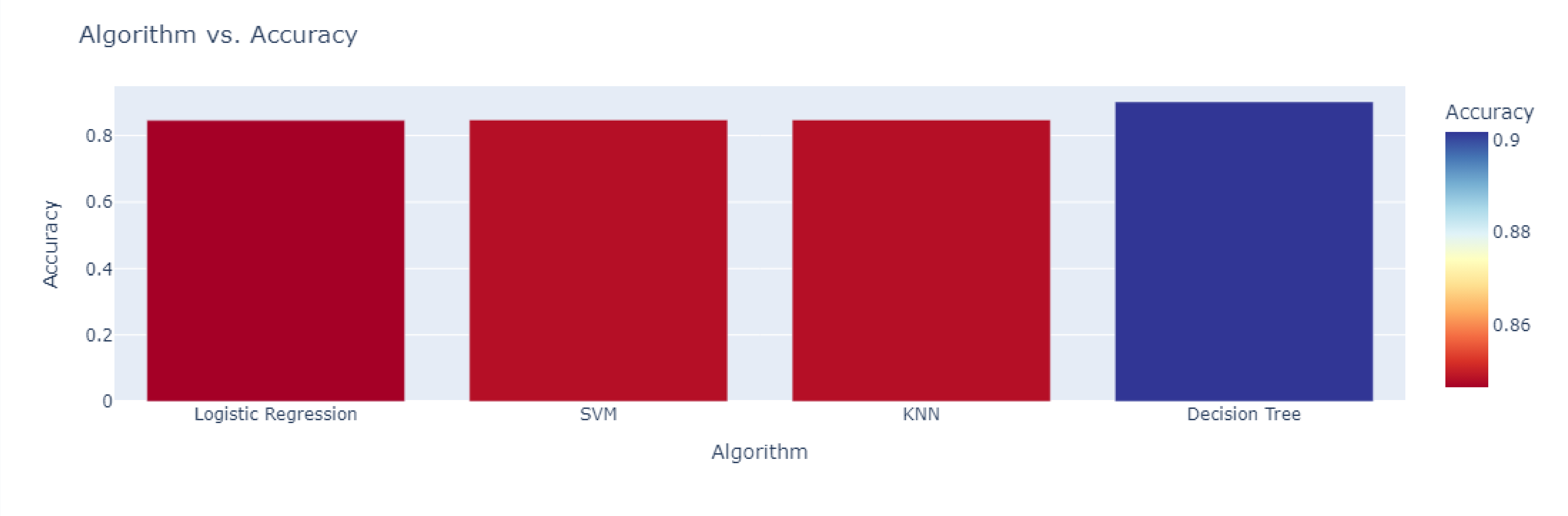
=>*Heavy-weighted Payload(4k-10k kgs)*

Findings: *Success rate for low-weighted Payloads is higher than heavy-weighted Payloads.*

Section 5

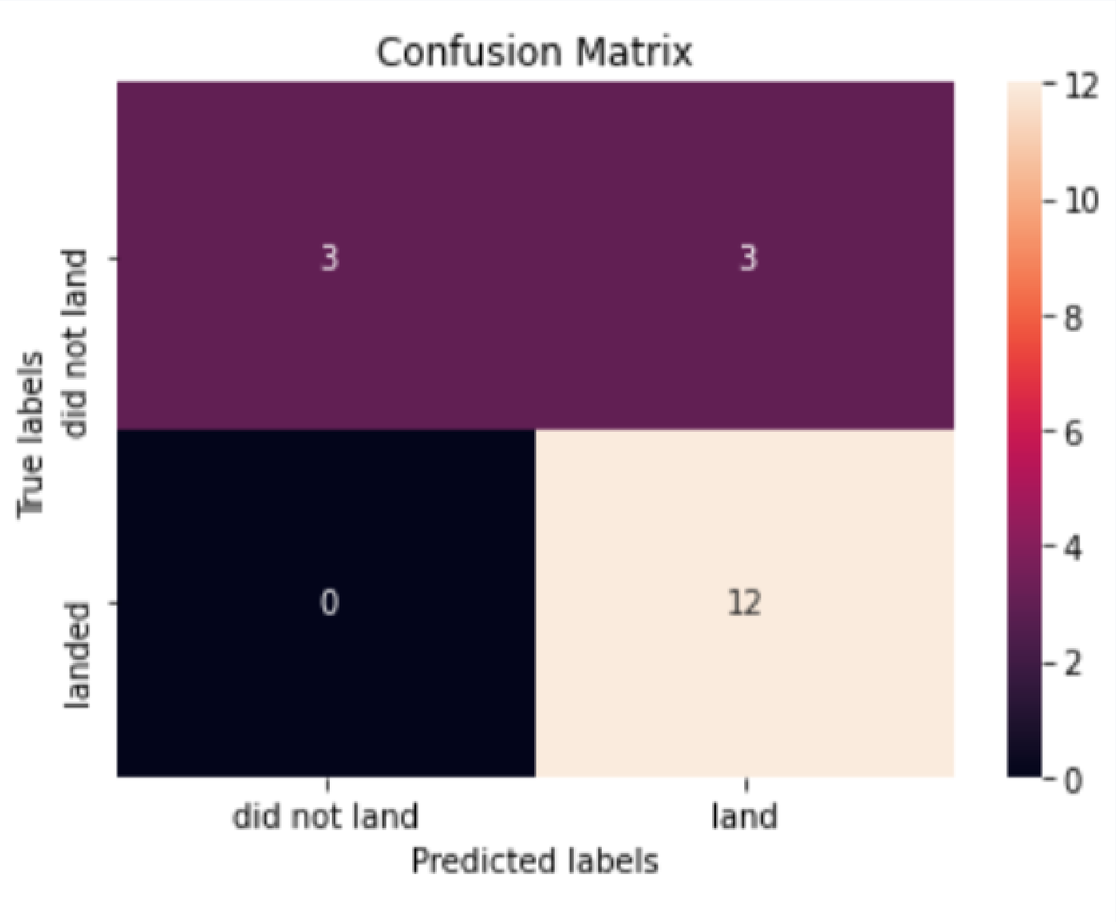
Predictive Analysis (Classification)

Classification Accuracy



The highest accuracy is given by the ***Decision Tree*** *Classification Model.*

Confusion Matrix



- *Unfortunately, all the classification models have the same confusion matrix.*
- ***Accuracy:*** *(TP+TN)/Total=(12+3)/18=0.83333*
- ***Misclassification Rate:****(FP+FN)/Total=(3+0)/18=0167*
- ***True Positive Rate:*** *TP/Actual Yes=12/12=1*
- ***False Positive Rate:*** *FP/Actaul No=3/6=0.5*
- ***True Negative Rate:*** *TN/Actual No=3/6=0.5*
- ***Precision:*** *TP/Predicted Yes=12/15=0.8*
- ***Prevalence****: Actual Yes/Total=12/18=0.6667*

Conclusions

Orbits with Highest Success Rates

- Orbits ES-L1, HEO, GEO, and SSo have highest success rates.

Most Successful Launch Site

- KSC LC-39A had the most successful launches but, increasing payload mass seems to have negative impact on success

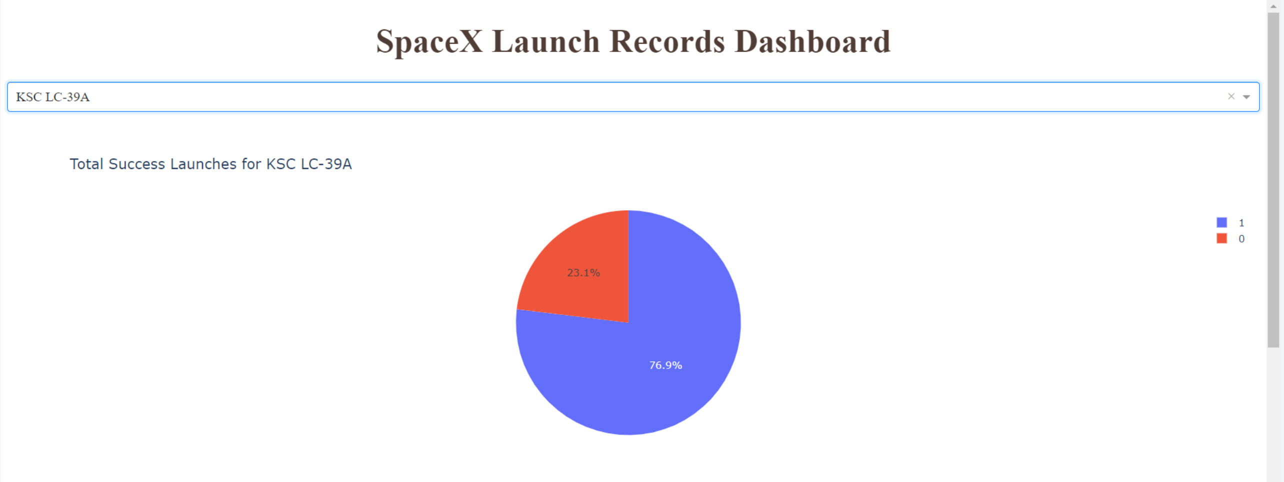
Best Classification Model

- Decision Tree Classifier Algorithm is the best Machine Learning Model for the provided dataset.

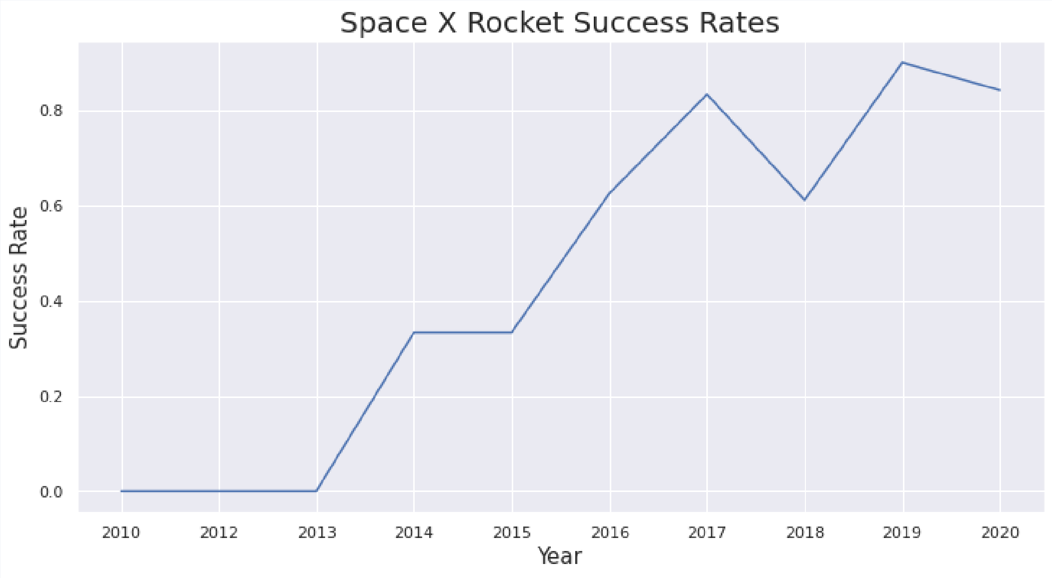
Success Rate

- Success rates for SpaceX launches have been increasing relatively with time and it looks like soon they will reach the required target.

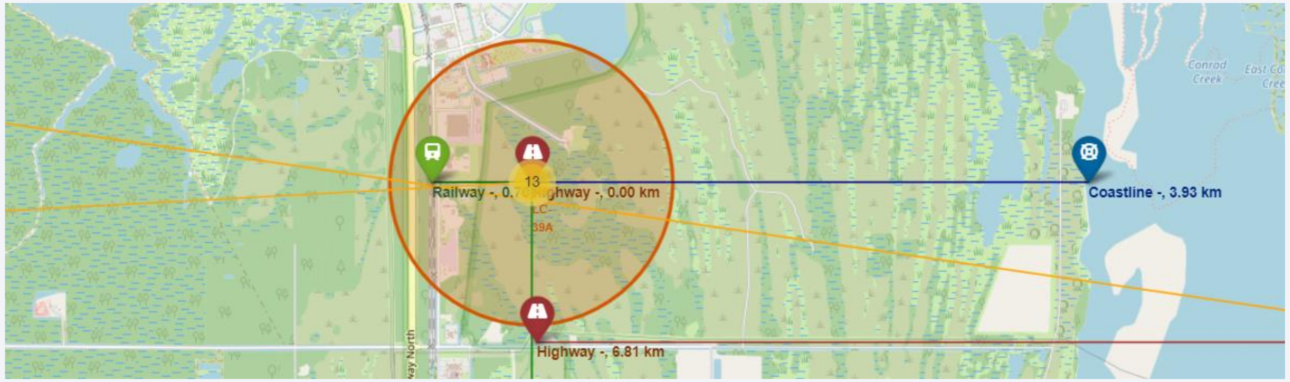
Appendix



Plotly Interactive Dashboard



Visualization for Yearly Trends



Folium Map to know the Launch Site Proximities

Thank you!