IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

EROTHI ALEKYA

Outline

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

Executive Summary

- Data Collection
- Data Collection With Web Scraping
- Data Wrangling
- Exploratory Data analysis with SQL(EDA)
- Exploratory Data Analytics with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Introduction

- The project background context is to determine the price of each launch of SPACEX.
- In this I'm gathering information about SpaceX and creating dashboards for the team, and determine if SpaceX will reuse the first stage.
- To determine if the stage will land successfully, you will train aa machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API and Web Scrapping
- Perform data wrangling
  - Data were transformed and one hot encoded to be apply later on the Machine Learning models.
- Perform exploratory data analysis (EDA) using visualization and SQL

  Discovering new patterns in the data with visualization techniques such as scatter plots.
- Perform interactive visual analytics using Folium and Plotly Dash

  Dash and Folium were used to achieve this goal
- Perform predictive analysis using classification model

  Creating best machine Learning Model

Data Collection

- Data sets are collected by SpaceX API request and Web Scraping.

Enter the URL page you want to analyse for the project

Request and parse the SpaceX launch data using the GET request

Convert the data into JSON file and convert into Pandas data frame

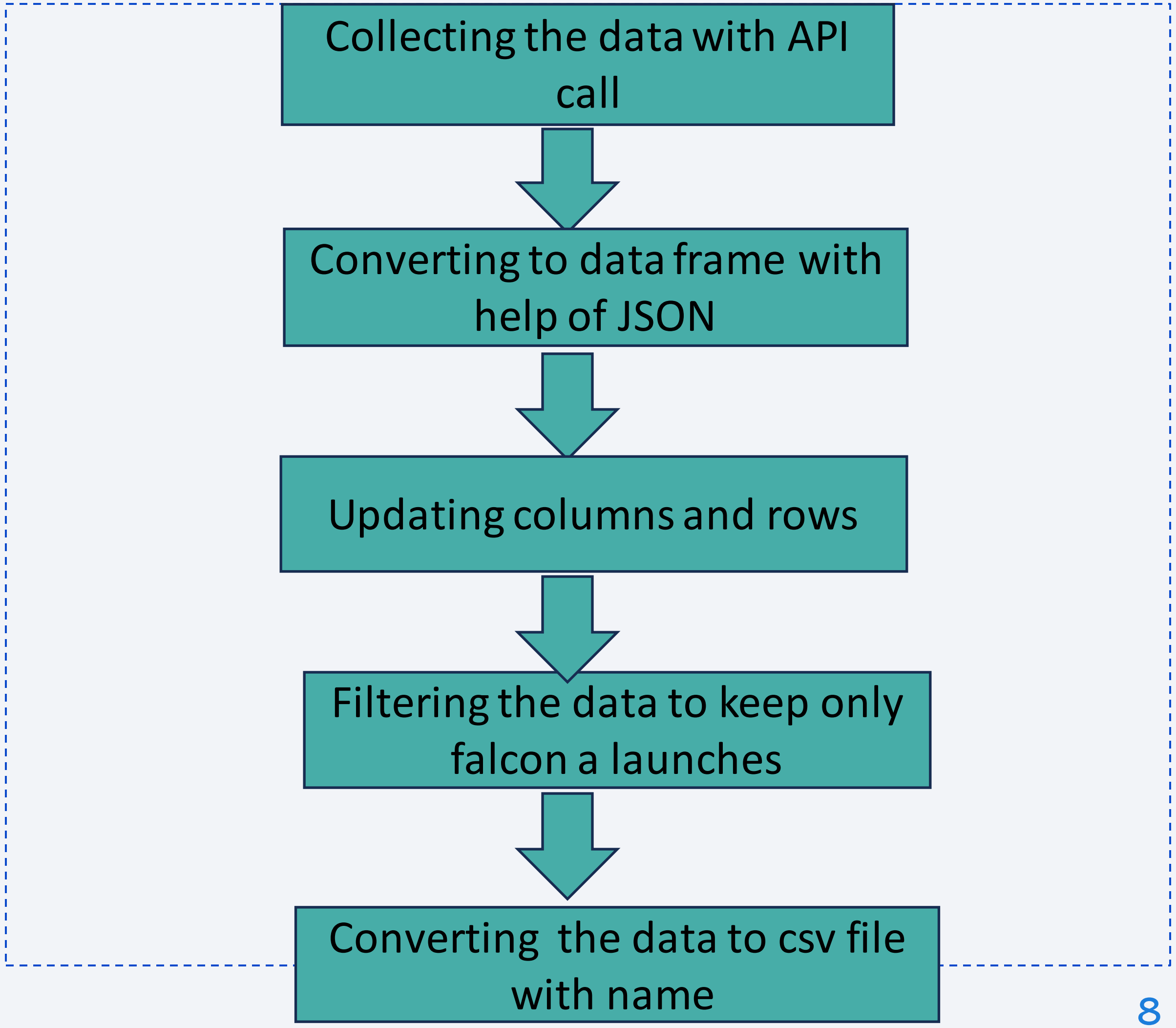
Now use the API again to get information about the launches using the IDs given for each launch

Filter the data frame to only include Falcon 9 launches and replace null values and get required output

Data Collection – SpaceX API

- Present my data collection with SpaceX REST calls using key phrases and flowcharts.

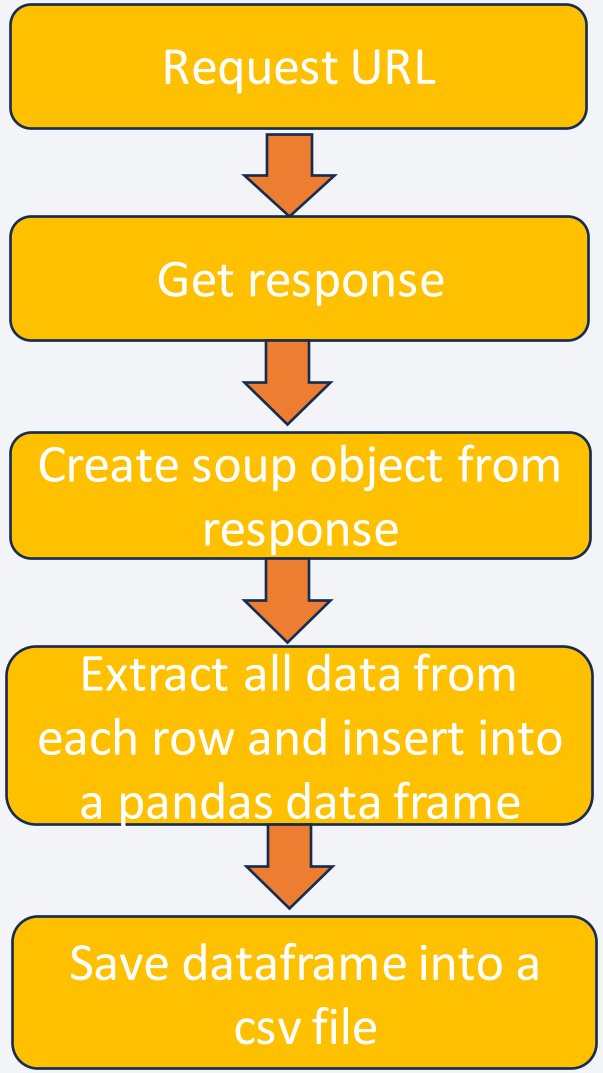
- Git Hub Link https://github.com/alekya8096/testrepo/blob/de4a7008369bf54aeff7b58ea9cd3ceaf7b8d83f/Complete%20the%20Data%20Collection%20API%20Lab



Data Collection - Scraping

- Used the request library to scrape data from URL link
- Used beautiful soup to parse the content returned in the response
- The parsed data was added to a pandas data frame and then exported to a CSV file
- Githublink:

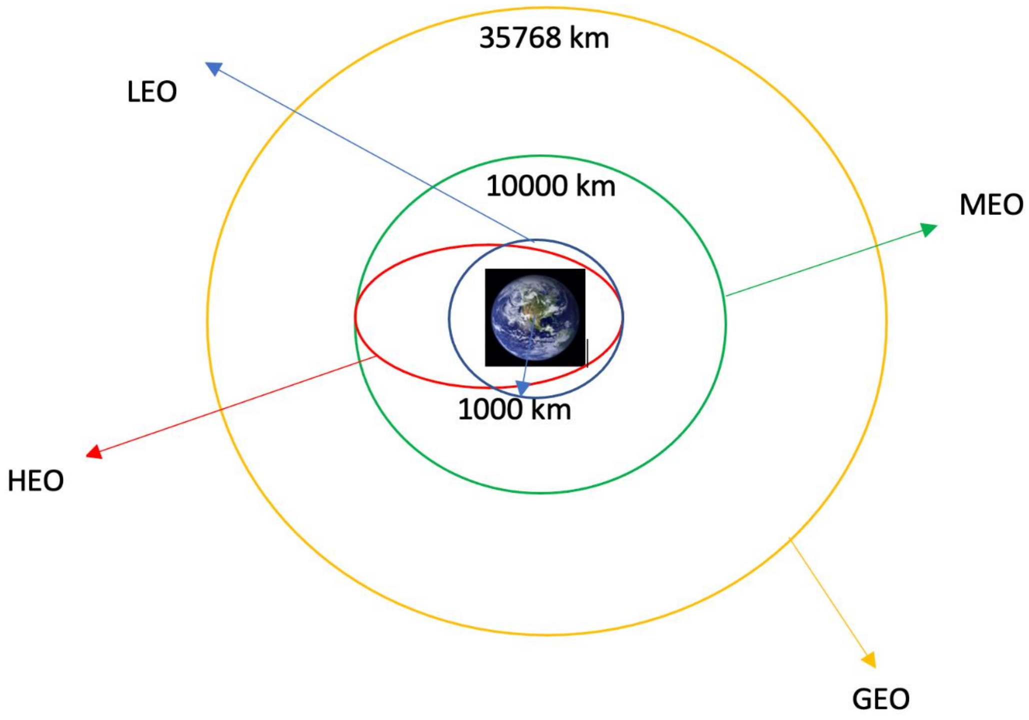
https://github.com/alekya8096/testrepo/blob/de4a7008369bf54aeff7b58ea9cd3ceaf7b8d83f/datacollection%20with%20webscra



Data Wrangling

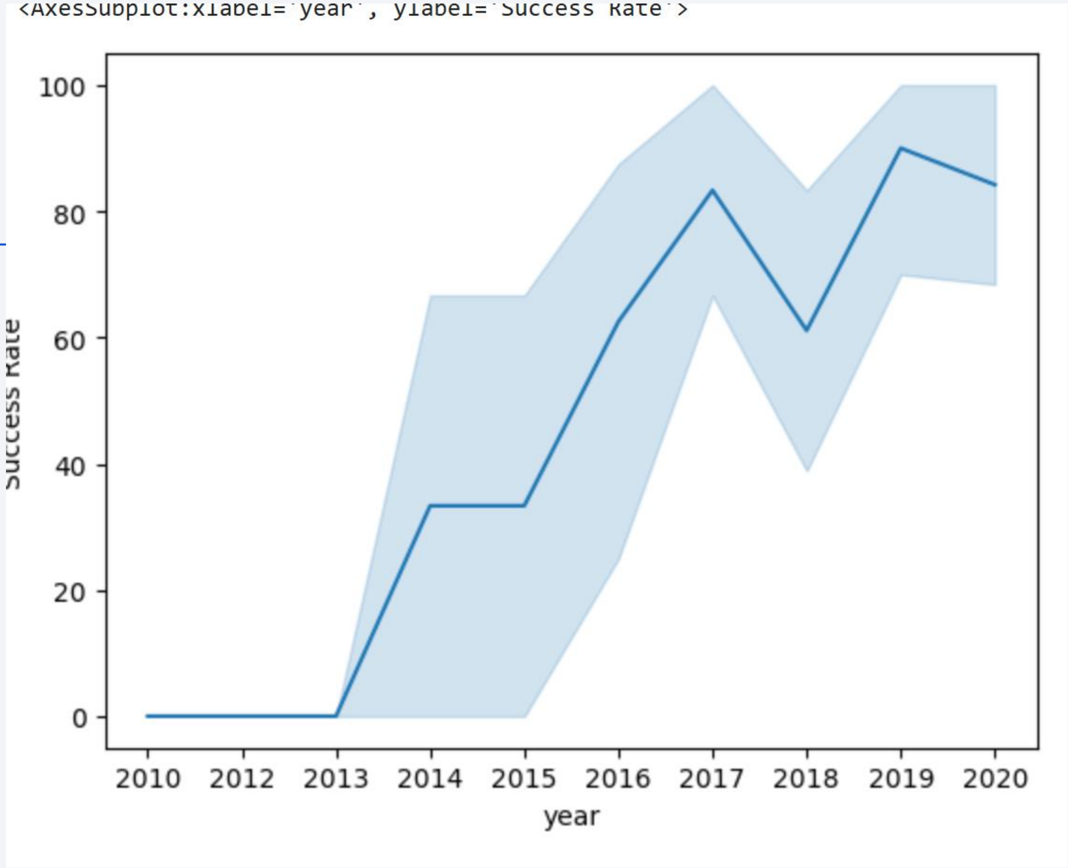
- We perform some Exploratory data analysis to find some patterns in the data and determine what would be the label for training supervised models.
- We calculate the number of launches on each site, and calculate the number and occurrence of each orbit.
- We created landing outcome label from outcome column and exported into csv file.
- Git hub link:

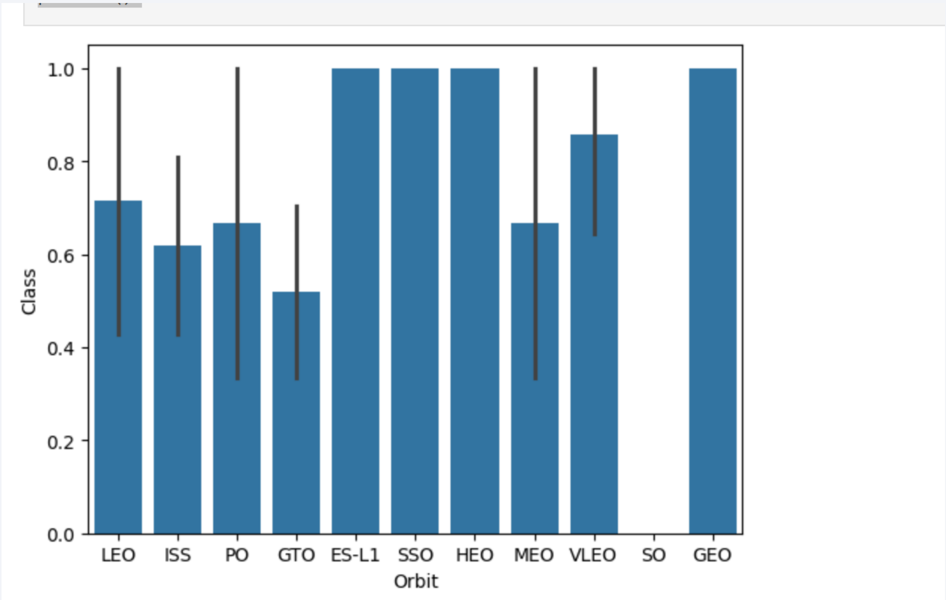
https://github.com/alekya8096/testrepo/blob/de4a7008369bf54aeff7b58ea9cd3ceaf7b8d83f/Data%20wrangling.ipynb



EDA with Data Visualization

- We exploring the data by visualizing, the bar chart represent the relationship between success rate and orbit type and the line represent the relation between launch success yearly trend.





https://github.com/alekya8096/testrepo/blob/5bfc143ff7a88d1244f3b94b730fadfeb0f6c4b4/edadataviz.ipynb

EDA with SQL

- We wrote queries to find out for insance :

1. the names of the unique launch sites in the space mission.

2. the 5 records where launch sites begin with the string 'CCA'.

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

4. the average payload mass carried by booster version F9 v1.1.

5. the date when the first successful landing outcome in ground pad was acheived.

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

Git hub link https://github.com/alekya8096/testrepo/blob/2e60ef450897cefb2a05ad2c071ee127d6788335/EDA%20with%20sql.ipynb

Build an Interactive Map with Folium

- Markers and circles were added to interactive folium map indicate launch site locations.
- Marker cluster is used for visualize the outcomes of each launch site.
- A line was drawn between a launch site and the coast and label with the distance was added to show the proximity of the two. Additional lines help users visualize distances to other important features that you want to avoid, such as roads, railways and towns/cities etc.
- Git hub link : https://github.com/alekya8096/testrepo/blob/2e60ef450897cefb2a05ad2c071ee127d6788335/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard was created to allow users to investigate the effects of launch site, payload mass and booster type on the launch outcome (good or bad).
- Launch site was selectable from a drop down menu and a range of payload masses could be selected using a slider control.
- A pie chart showed either the successful outcome for all launch sites or the proportion of good and bad outcomes for any one selected launch site.
- A scatter chart showed how the launch outcome varied by the selected site and payload range and the data points were color-coded by booster type.
- Git hub link :https://github.com/alekya8096/testrepo/blob/2e60ef450897cefb2a05ad2c071ee127d6788335/spacex_dash_app1.py

Predictive Analysis (Classification)

- The following classification algorithms:

  1.logistics regression

  2.Support vector machines

  3.decesion tree classification

  4.k nearest neighbors

- Git hub link:

https://github.com/alekya8096/testrepo/blob/2e60ef450897cefb2a05ad2c071ee127d6788335/SpaceX_Machine%20Learning%20Prediction_Part_5.backup.ipynb

Results

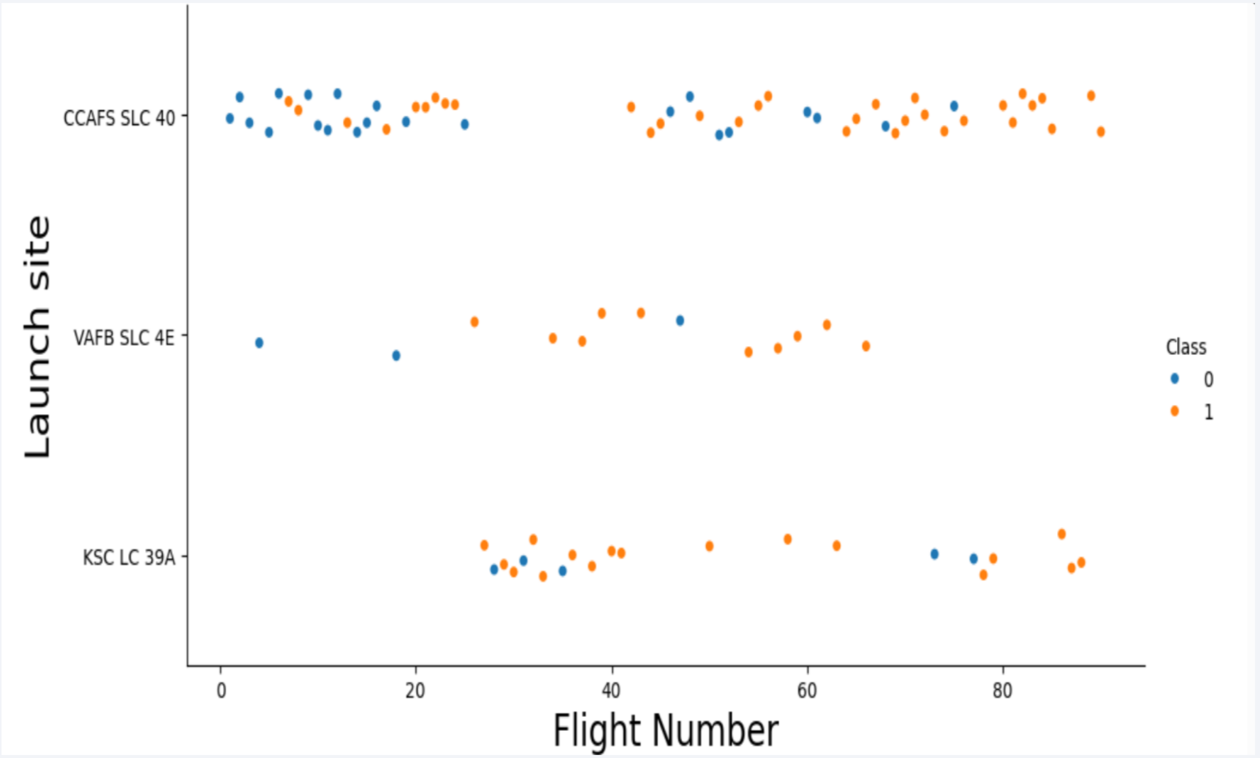
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

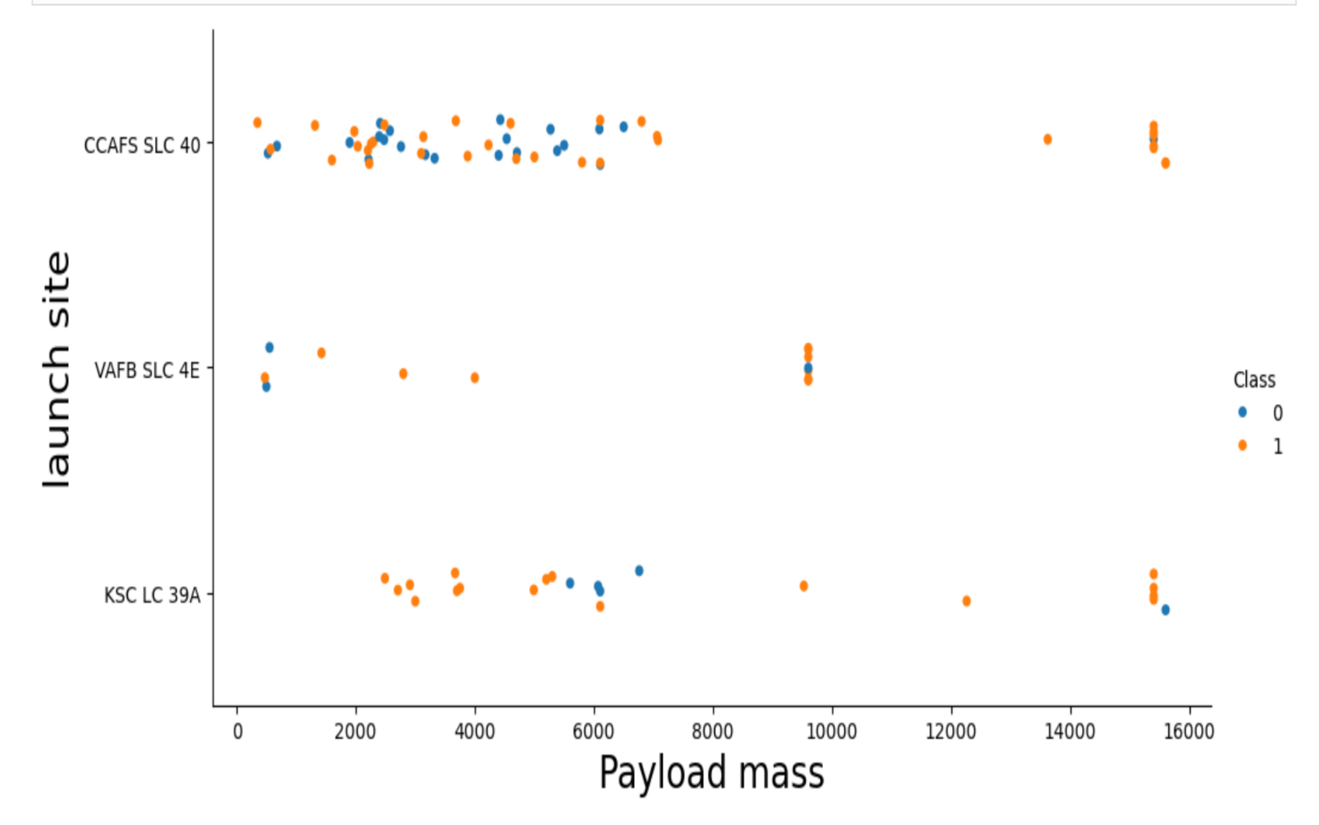
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site outcomes by colored
- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.



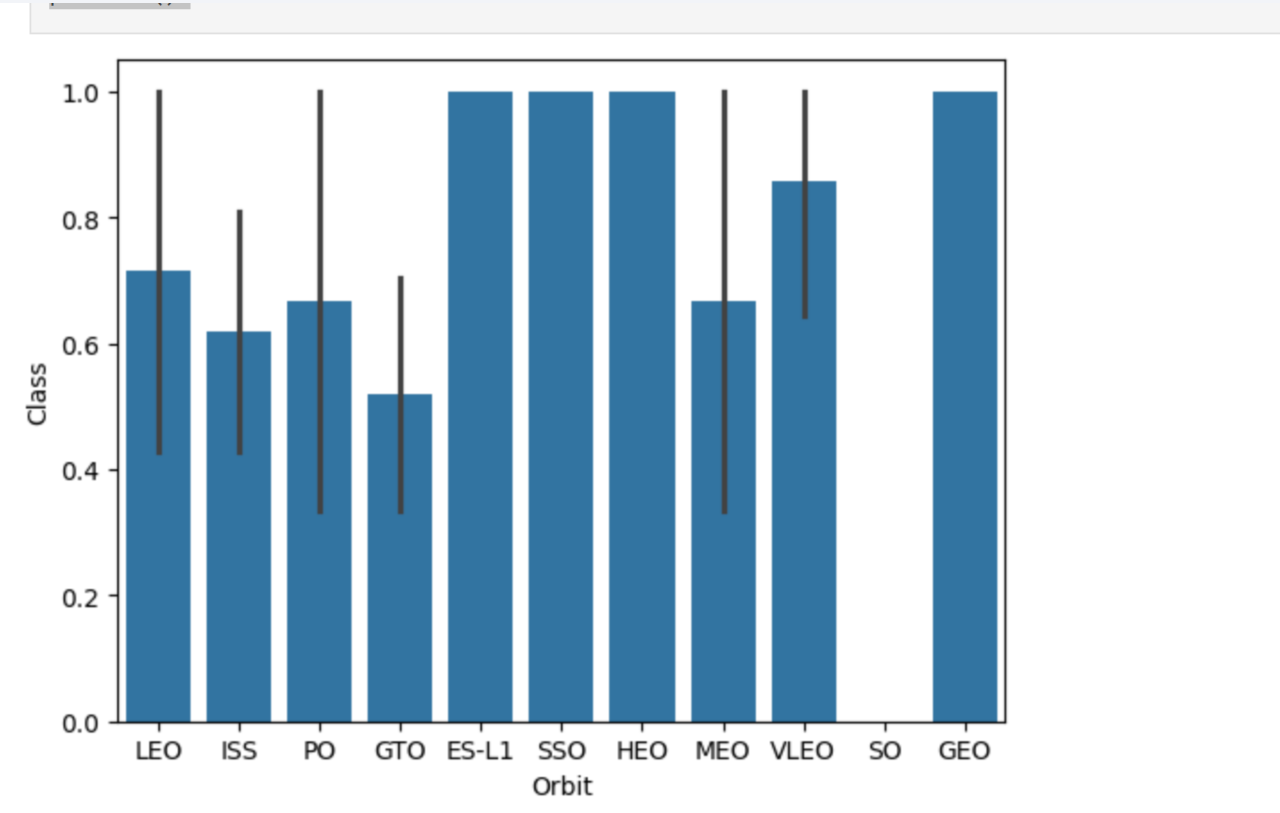
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site outcomes by colored.
- if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).



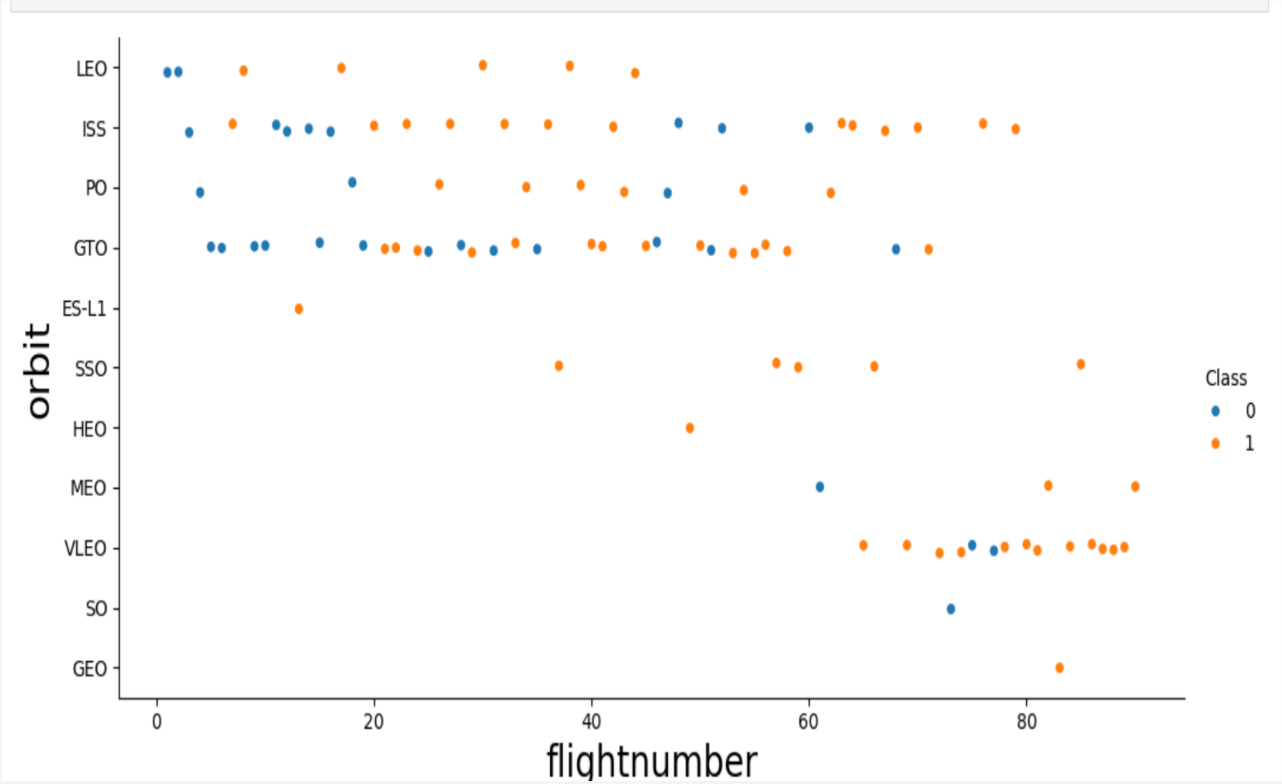
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type outcomes.
- ES-L1,HEO and GEO are the highest success rate .



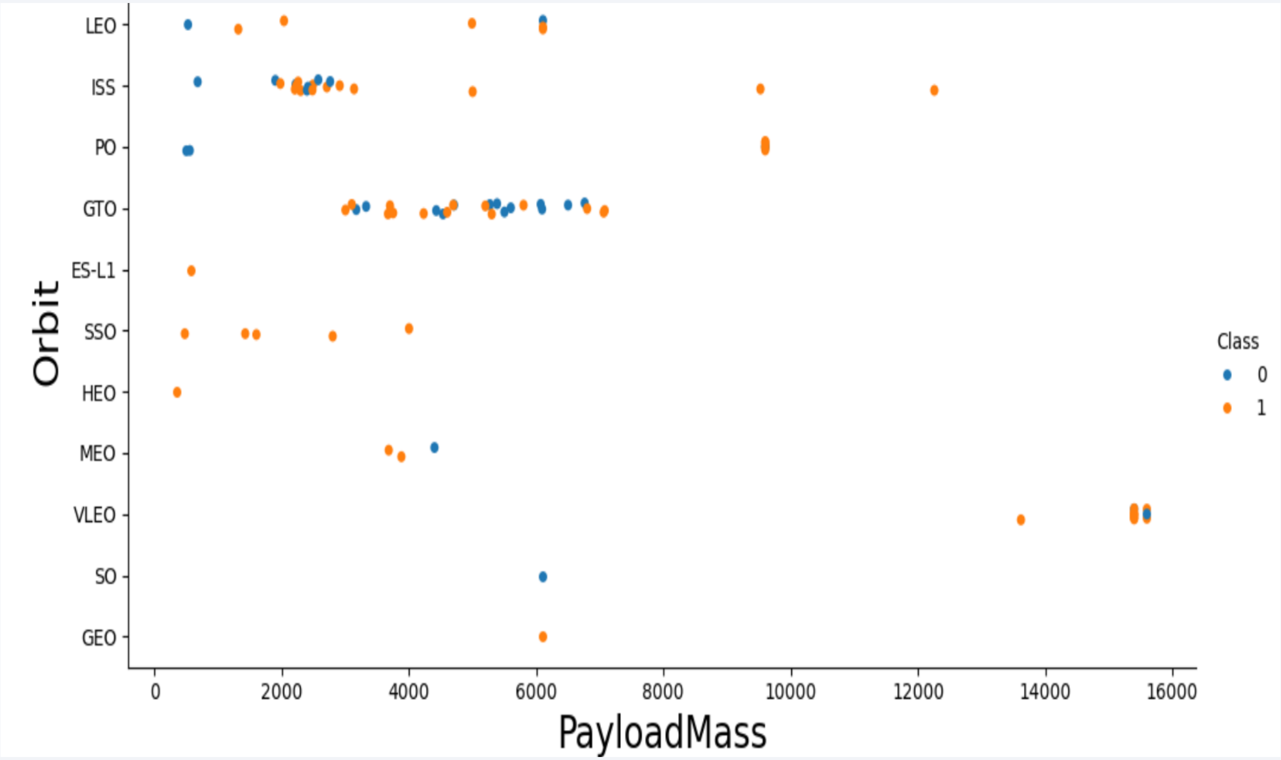
Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type outcomes by colored.
- You should see that in the 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.



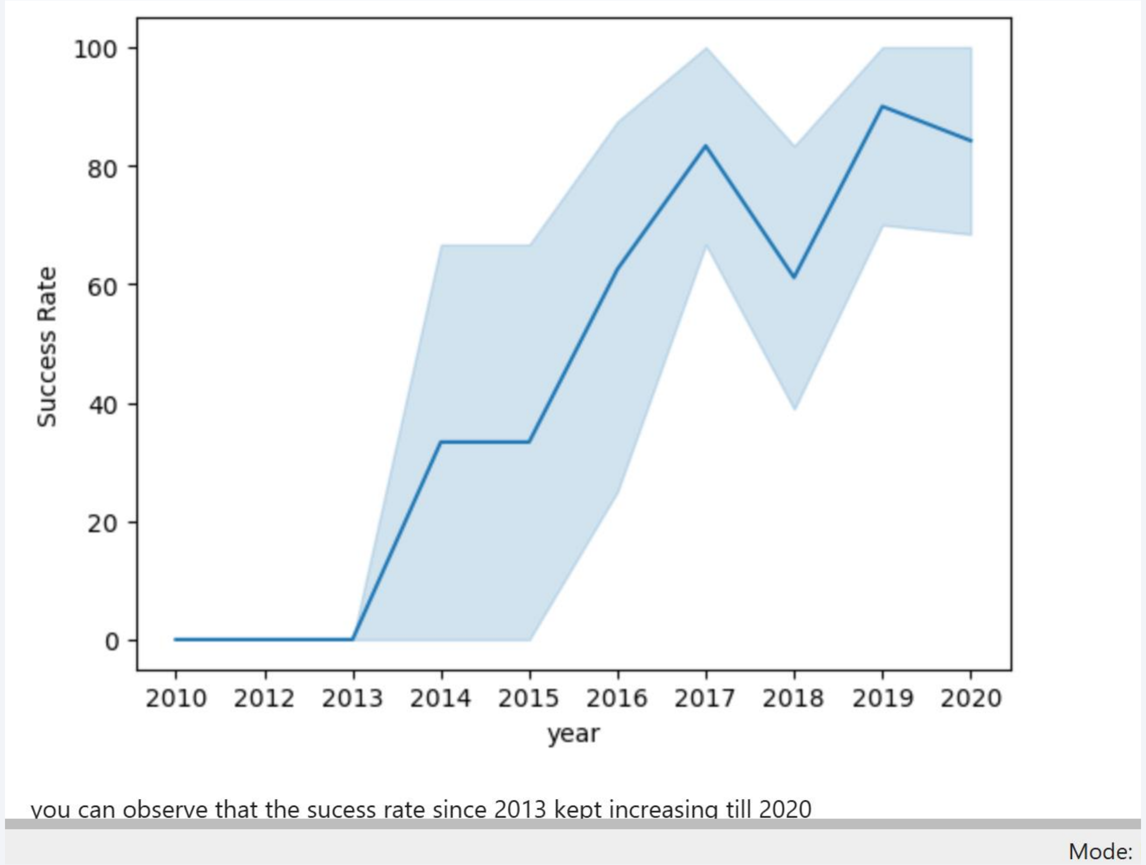
Payload vs. Orbit Type

- Scatter point of payload vs. orbit type outcomes by colored.
- 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

- Line chart of yearly average success rate outcomes
- you can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- Names of the unique launch sites query :

  SELECT DISTINCT("LAUNCH_SITE")
FROM SPACEXTABLE:

- This query gives you all launch site names and the DISTINCT keyword gives you unique values in specific column.

]:

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`

Select * from SPACEXTABLE where Launch_Site LIKE 'CCA%' LIMIT 5;

- In this WHERE key word is used select specific column and LIKE is select specific name selected by first or last alphabet ,and LIMIT gives you how many rows we want.

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

- The total payload carried by boosters from NASA

Select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer='NASA (CRS)';

- In this sSUM function is used gives you total sum value of payload_mass_kg column where customer NASA (CRS)

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1:

Select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version ='F9v1.1';

- In this the AVG function is used for gives you total average of payload_mass_kg column.

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad

select min(Date) from SPACEXTABLE where Landing_Outcome ='Success (ground pad)';

- In this the MIN function used for gives you minimum value of particular column or row.

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000

- The less than and greater than symbols are used to filtering the records by some range of values within a column

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

Select Mission_Outcome,count(*) from SPACEXTABLE group by Mission_Outcome;

- In this GROUP BY aggregates function used the data according to the specified variable.

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

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

Select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);

- A subquery is used to find the maximum payload mass across all records.

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

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

select substr(Date,6,2) as Monthname,Landing_Outcome,Booster_Version,Launch_Site from SPACEXTABLE where Landing_Outcome ='Failure (droneship)' and substr(Date,0,5)='2015';

- The AND condition is used to create two filters with each one working one value in a specific column

YEAR	MONTH	Landing _Outcome	Booster_Version	Launch_Site
2015	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- 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

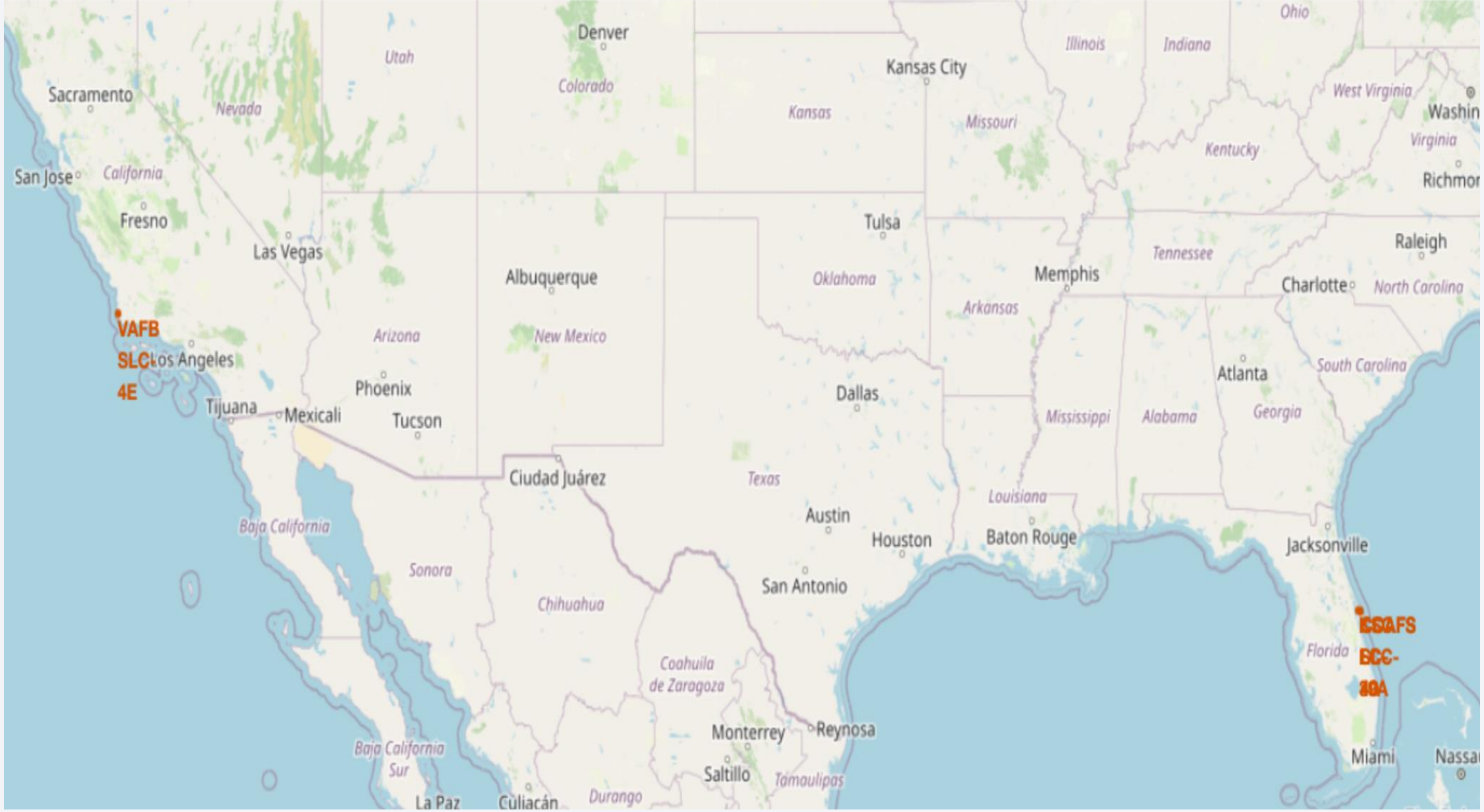
```
select count(Landing_Outcome) from SPACEXTABLE where Landing_Outcome='Success (ground pad)' or Landing_Outcome='Failure (drone ship)'  AND (Date between '2010-06-04' and '2017-03-20');
```

count(Landing_Outcome)
14

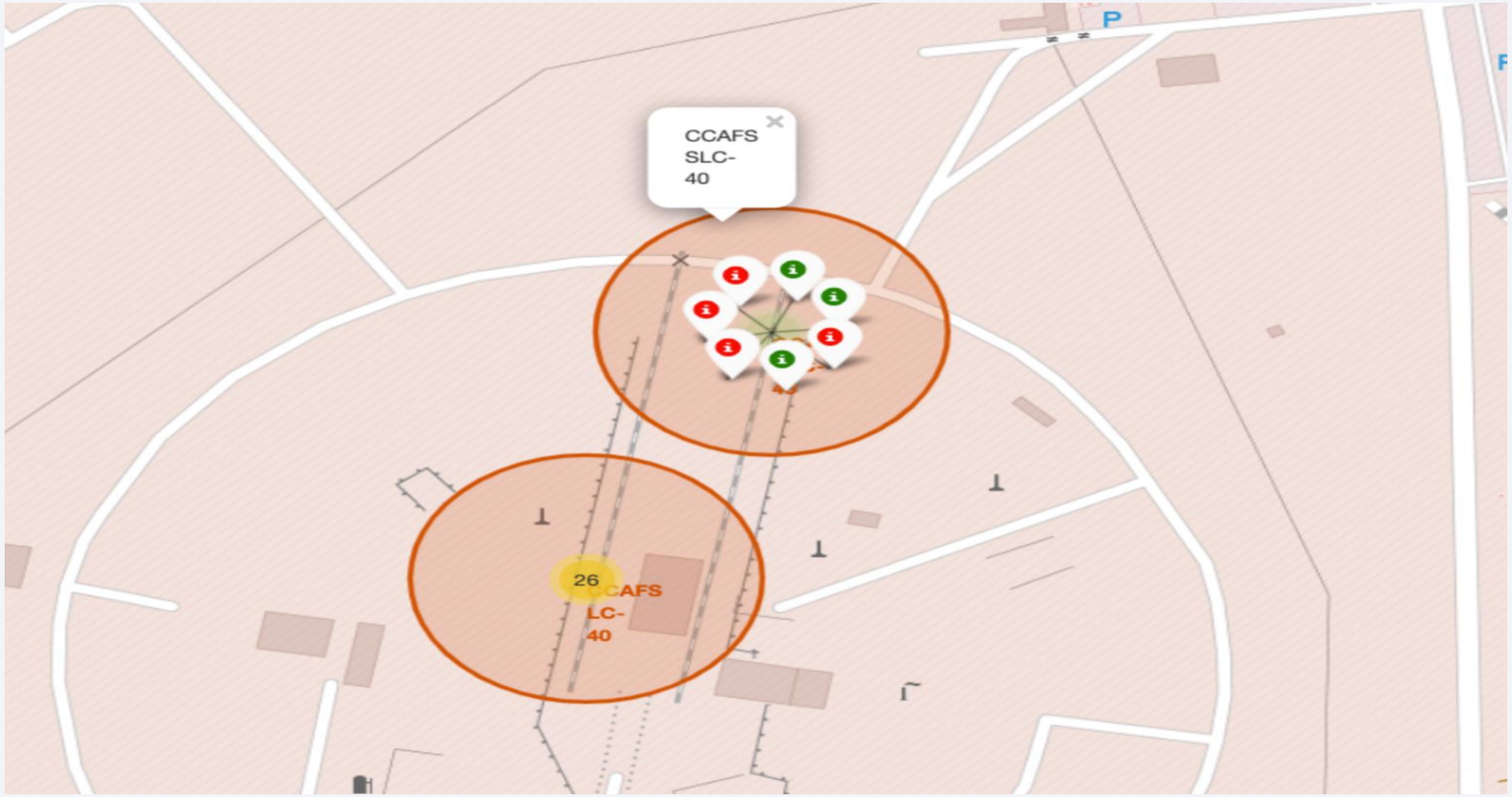
Section 3

Launch Sites Proximities Analysis

Launch sites loactions:

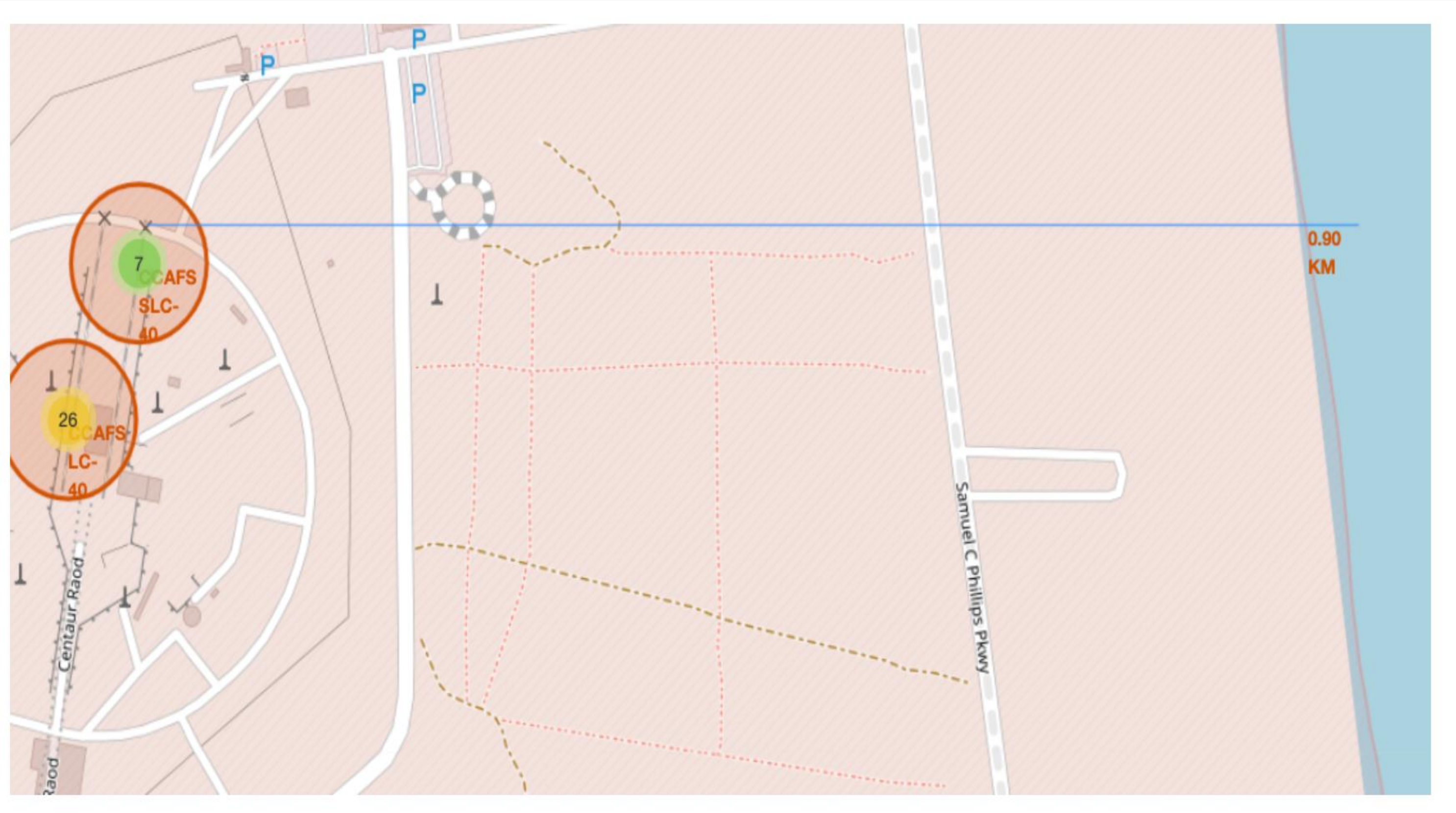


Color labeled launch outcomes



- The color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

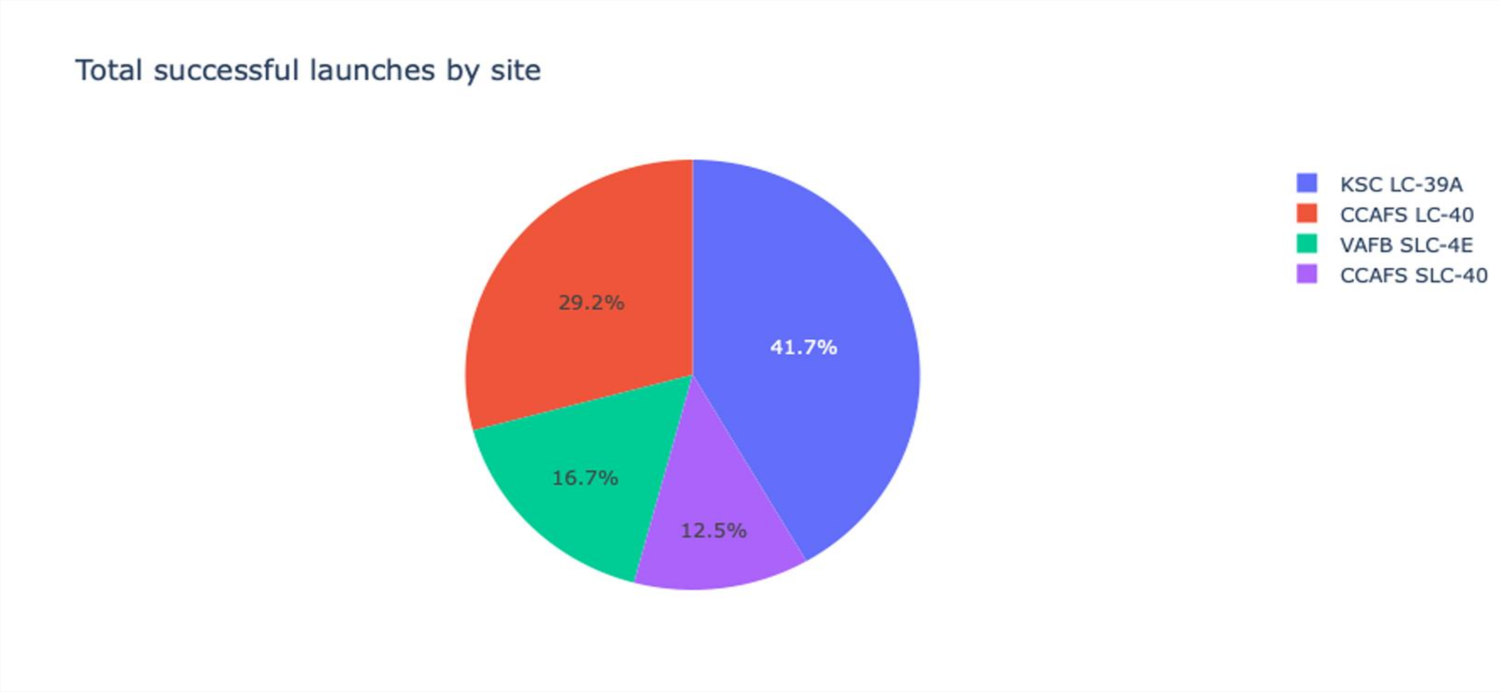
Selected launch site its proximities



Section 4

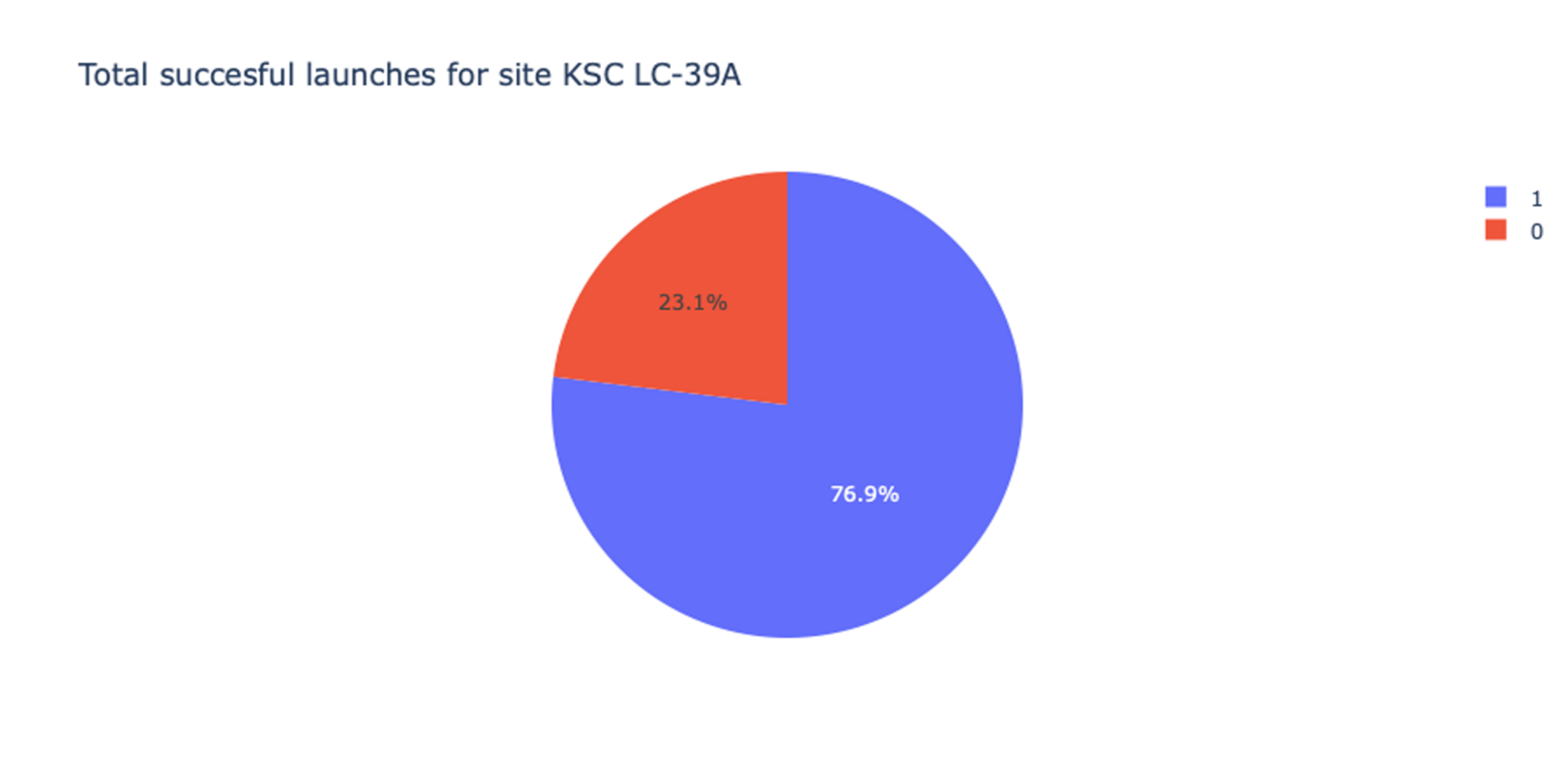
Build a Dashboard with Plotly Dash

Launch success count for all sites

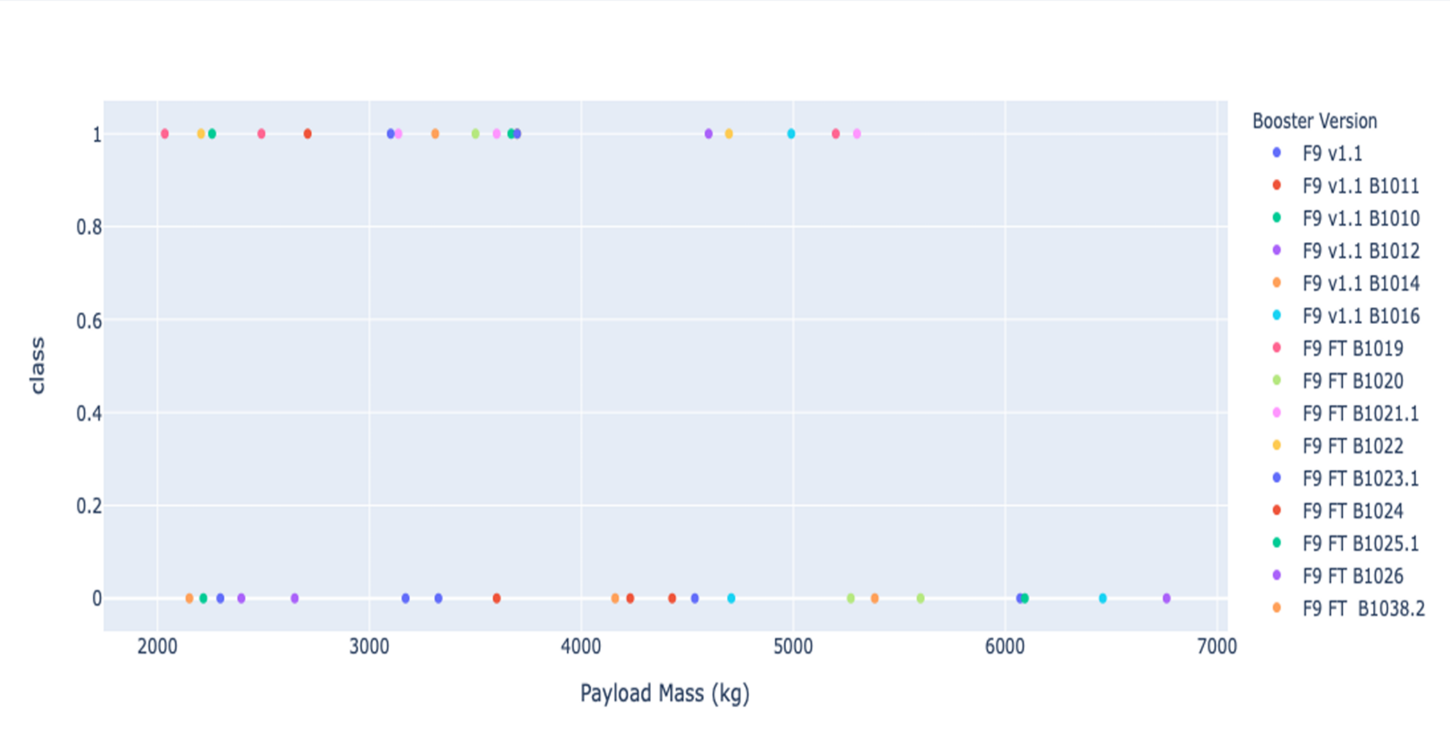


Highest success launches in KSC LC-39A

Launch site with highest launch success ratio



Payload Vs. Launch Outcome scatter plot for all sites

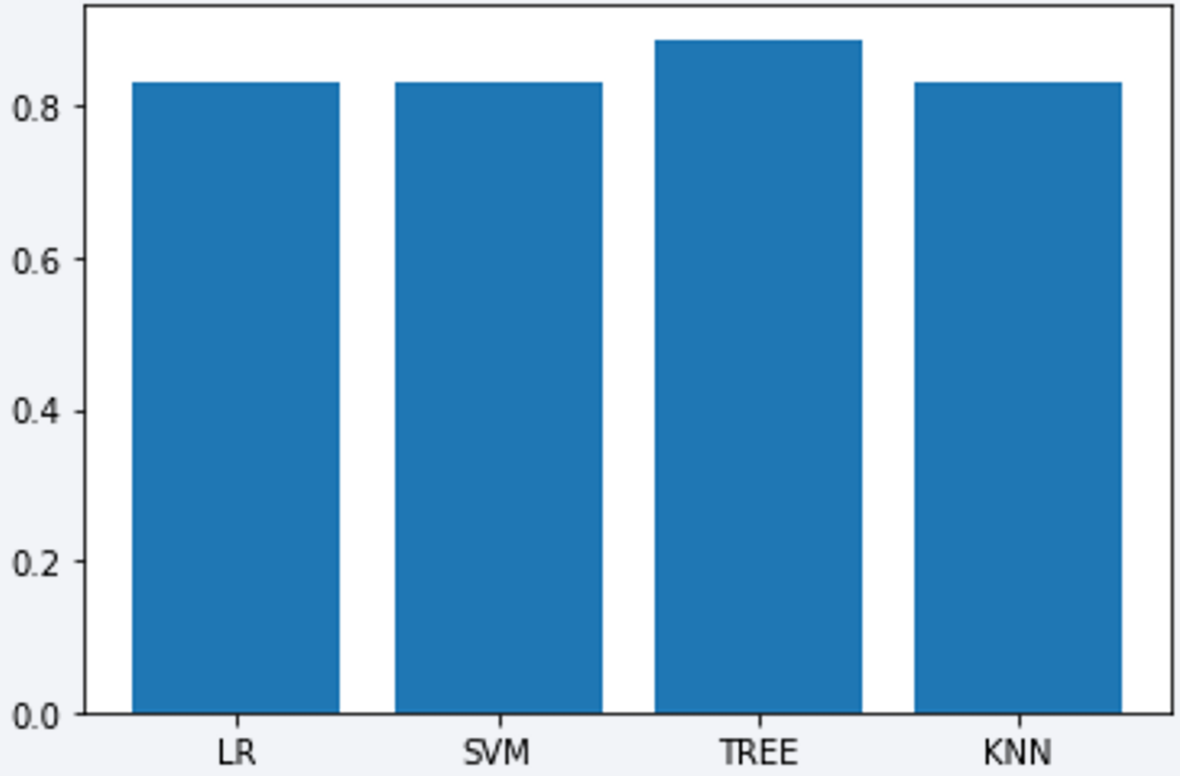


Section 5

Predictive Analysis (Classification)

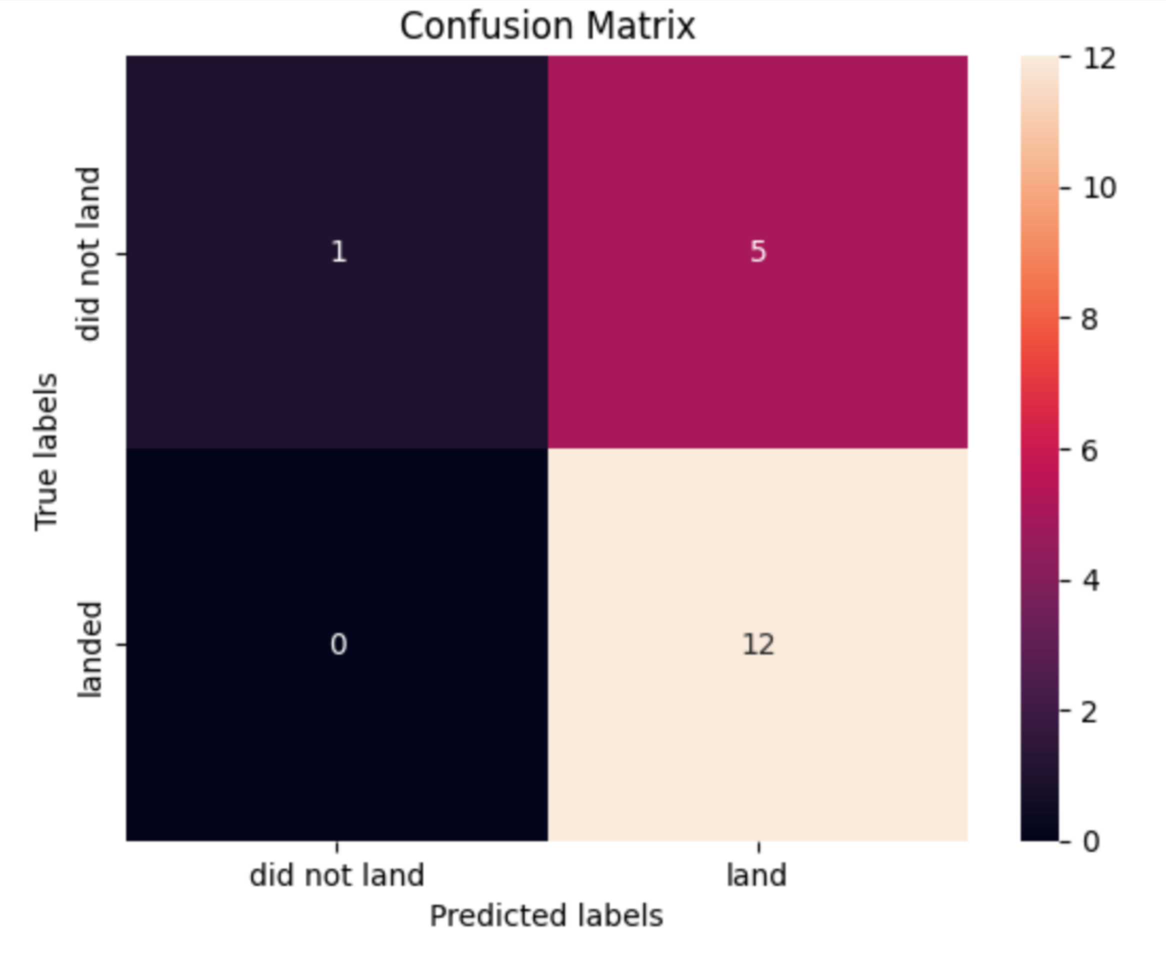
Classification Accuracy

- Built model accuracy for all built classification models,
- The TREE model has the highest classification accuracy.



Confusion Matrix

- This is the best confusion matrix it indicates the tree classification model.



Conclusions

- Highest success launches are KSC
- A decision tree classification model is the best model for predicting.

Appendix

- There is nothing to include here , all git hub links, and graphs or charts are provided in this presentation, thank you.

Thank you!