IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bhaveshkumar Jadav
11/11/2023

Outline

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

Executive Summary

Accuracy of result depends on data and source of data.

Payload Mass having in range 2k-4k has higher success ratio.

KSC LC-39A site has highest success ratio.

FT booster version is most successful version having highest success rate.

Introduction

- SpaceX's Falcon 9, renowned for its cost-effectiveness driven by first stage reusability, stands as a pioneering example in the aerospace industry.
- This capstone project aims to leverage predictive modeling to determine the success of Falcon 9 first stage landings, a crucial factor influencing the overall cost of a launch. By analyzing historical data encompassing various launch parameters, mission outcomes, and contextual features, the project seeks to build a robust machine learning model capable of forecasting the likelihood of a successful landing.
- This capstone project encompasses a comprehensive methodology, including data collection, preprocessing, feature engineering, model selection, and evaluation.
- The significance of this endeavor lies in its potential impact on cost competitiveness within the space launch industry. Accurate predictions regarding the Falcon 9's first stage landing success not only contribute to mission planning and execution but also empower competing entities to strategically position themselves in bidding scenarios. With SpaceX's reusability driving down launch costs, an effective predictive model can assist alternate companies in formulating competitive bids, fostering a dynamic and cost-efficient environment in the space exploration domain.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

Data collected using SpaceX API and Web Scrapping from Wikipedia information.

Perform data wrangling

Replacing missing values with Mean value of column(e.g. column: PayloadMass)

Perform exploratory data analysis (EDA) using visualization and SQL

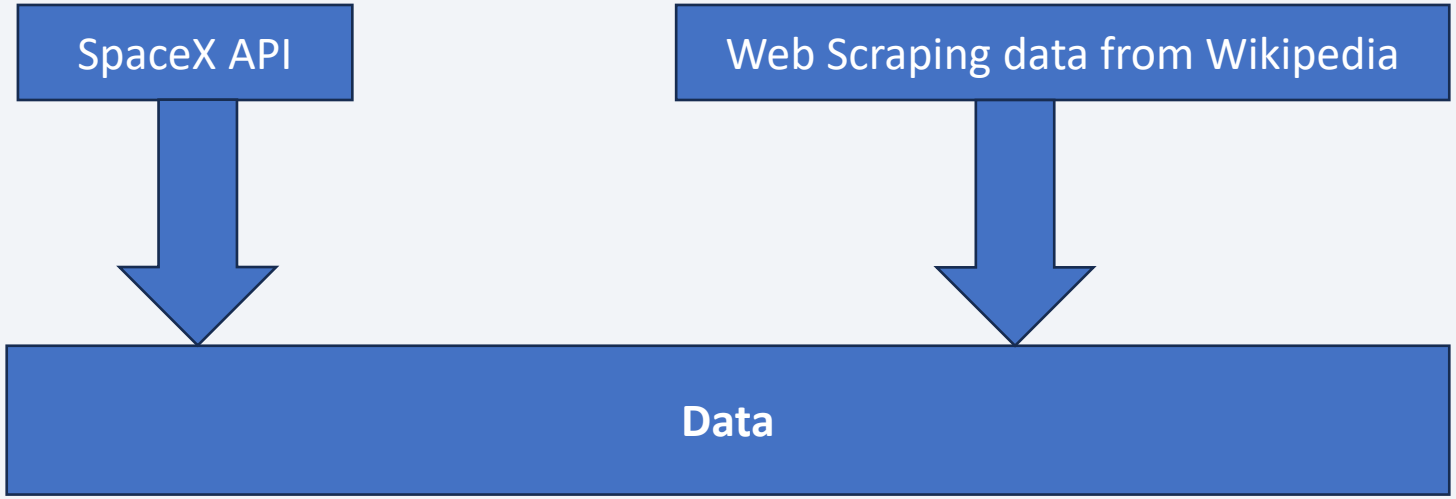
Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

Select different classification models and find best parameters for tunning and last evaluate accuracy and precision

Data Collection

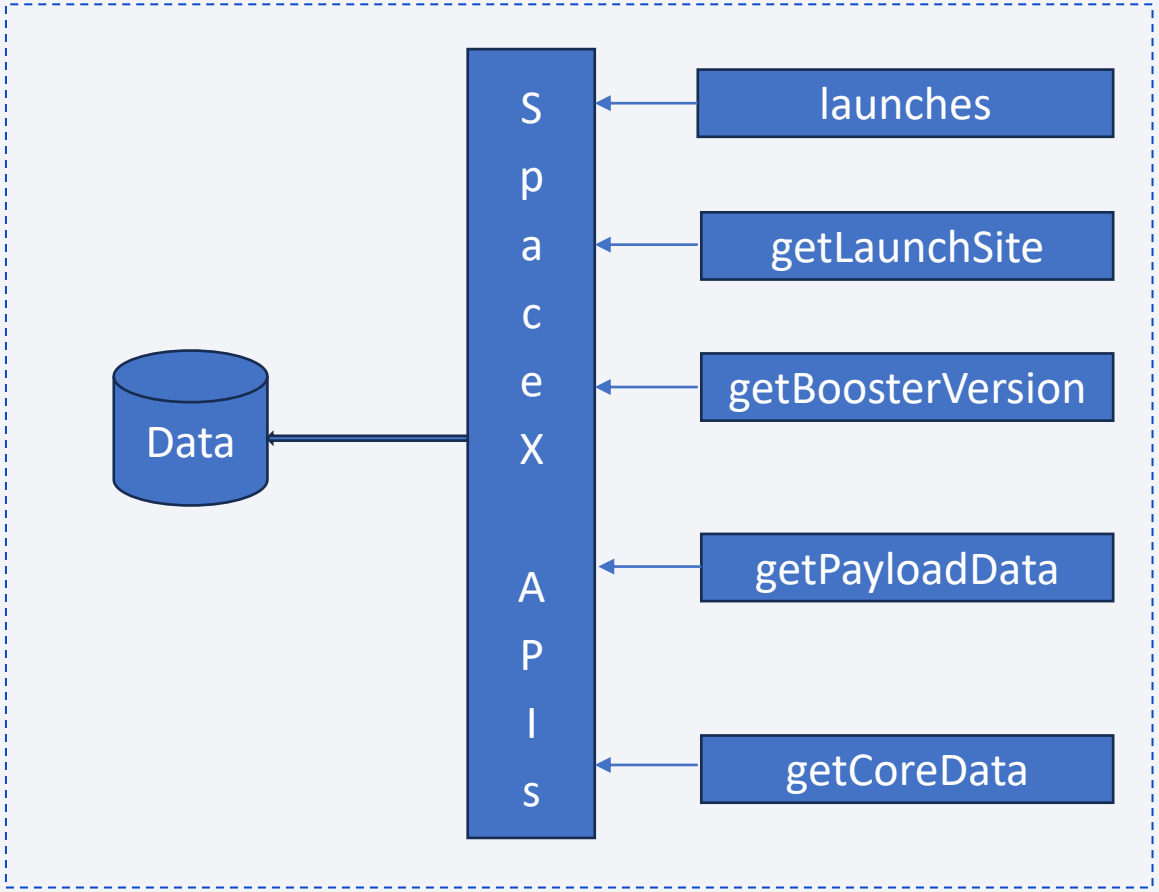
- Identified different source of data such as SpaceX data API and Web Scraping methods And collected data from sources.



Data Collection – SpaceX API

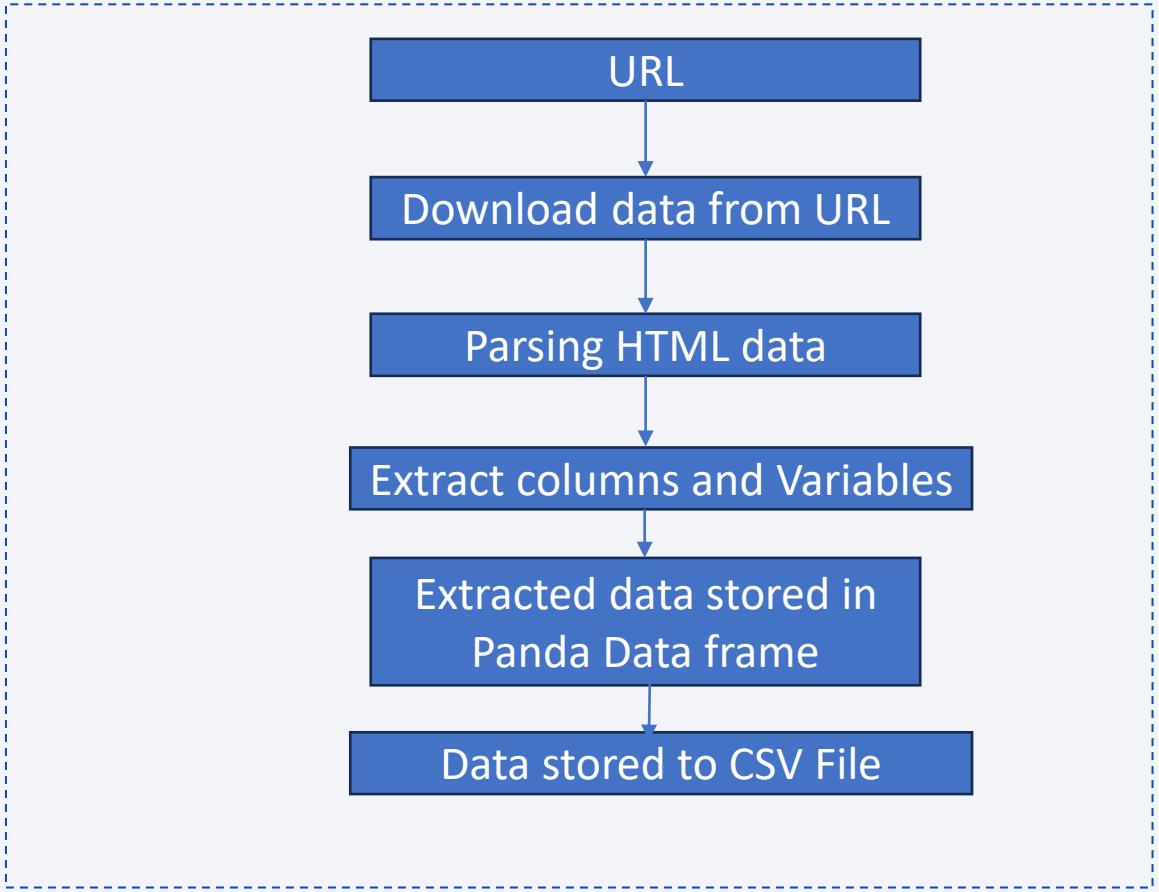
- Couple of API used to collect data as shown in right section.

- https://github.com/radhe004/DataScienceCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Collected data from Wikipedia page by web scraping using BeautifulSoup.
- Wiki URL: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- https://github.com/radhe004/DataScienceCapstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Cleaned the data
- Organizing raw data into a format suitable for analysis
- https://github.com/radhe004/DataScienceCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- I have scatter plot chat Because it help to visualize the relationship between two variables.
- I plotted a bar chart So that we can do comparison.
- I also plotted line chart to visual launch success yearly trend
- Following Chart Plotted:
  - Flight Number vs Payload Mass, Flight Number vs Launch Site, Payload vs Launch Site, Flight Number vs Orbit type, Payload vs Orbit type, Success rate of each orbit type, Launch success yearly trend
- https://github.com/radhe004/DataScienceCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Summarization of the SQL performed queries
  - **Unique launch site names in the space mission**
  - **5 records where launch sites begin with the string 'CCA'**
  - **The total payload mass carried by boosters launched by NASA (CRS)**
  - **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 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**
  - **The names of the booster versions which have carried the maximum payload mass using sub query**
  - **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.**
- https://github.com/radhe004/DataScienceCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

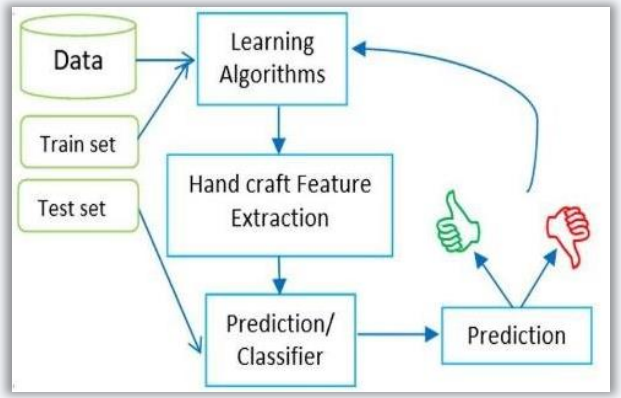
- I have used Markers, Circles, Lines and following is the detail for being used these objects:
  - Marker - Used to pinpoint specific locations on the map
  - Circle – To draw a circle around a specified point on the map
  - Line – To connect multiple points on the map and draw line for showing distance
- https://github.com/radhe004/DataScienceCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- I have created dashboard application with the Python Plotly Dash Package
- This dashboard application contains input components such as a dropdown list(location site selection) and a range slider(Payload selection) to interact with a pie chart and a scatter point chart
- You can analysis data using dashboard that success launch ratio of sites; also to visualize payload mass for booster version category and its success.
- https://github.com/radhe004/DataScienceCapstone/blob/main/spacex_dash_app.py

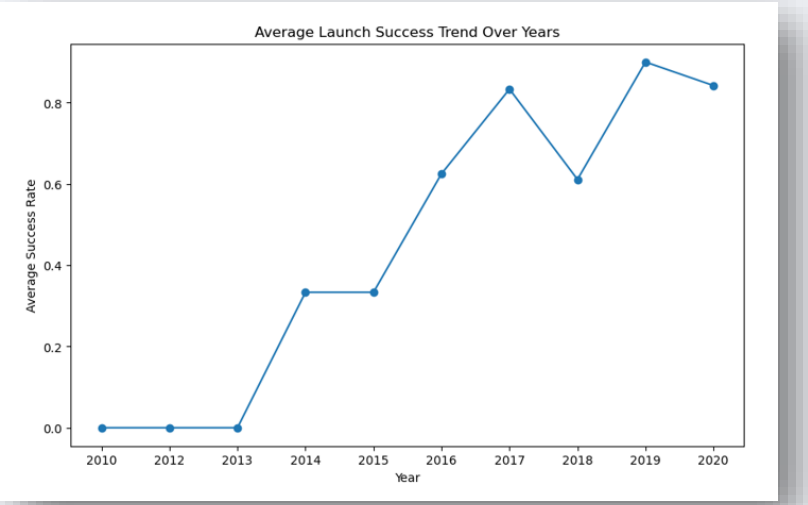
Predictive Analysis (Classification)

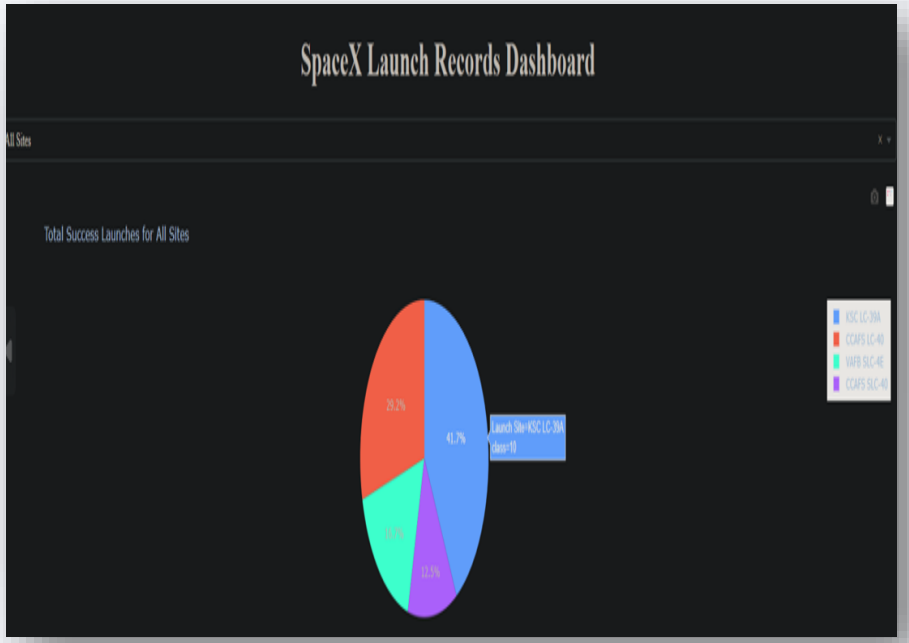
- Created a class column(output variable)
- Standardize the data, Created training and test data. here test size used 0.2.
- Used different Classification Models such as SVM, Classification Trees and Logistic Regression.
- Used Grid Search to find the best tuned parameters for prediction
- Check Accuracy and Confusion matrix for validating model
- https://github.com/radhe004/DataScienceCapstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

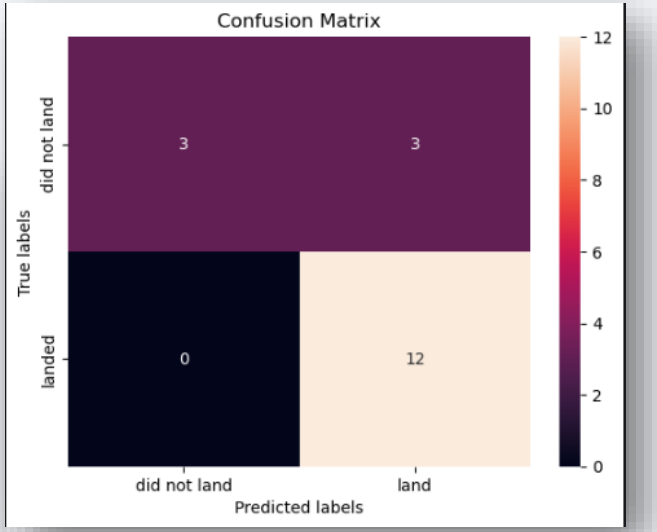


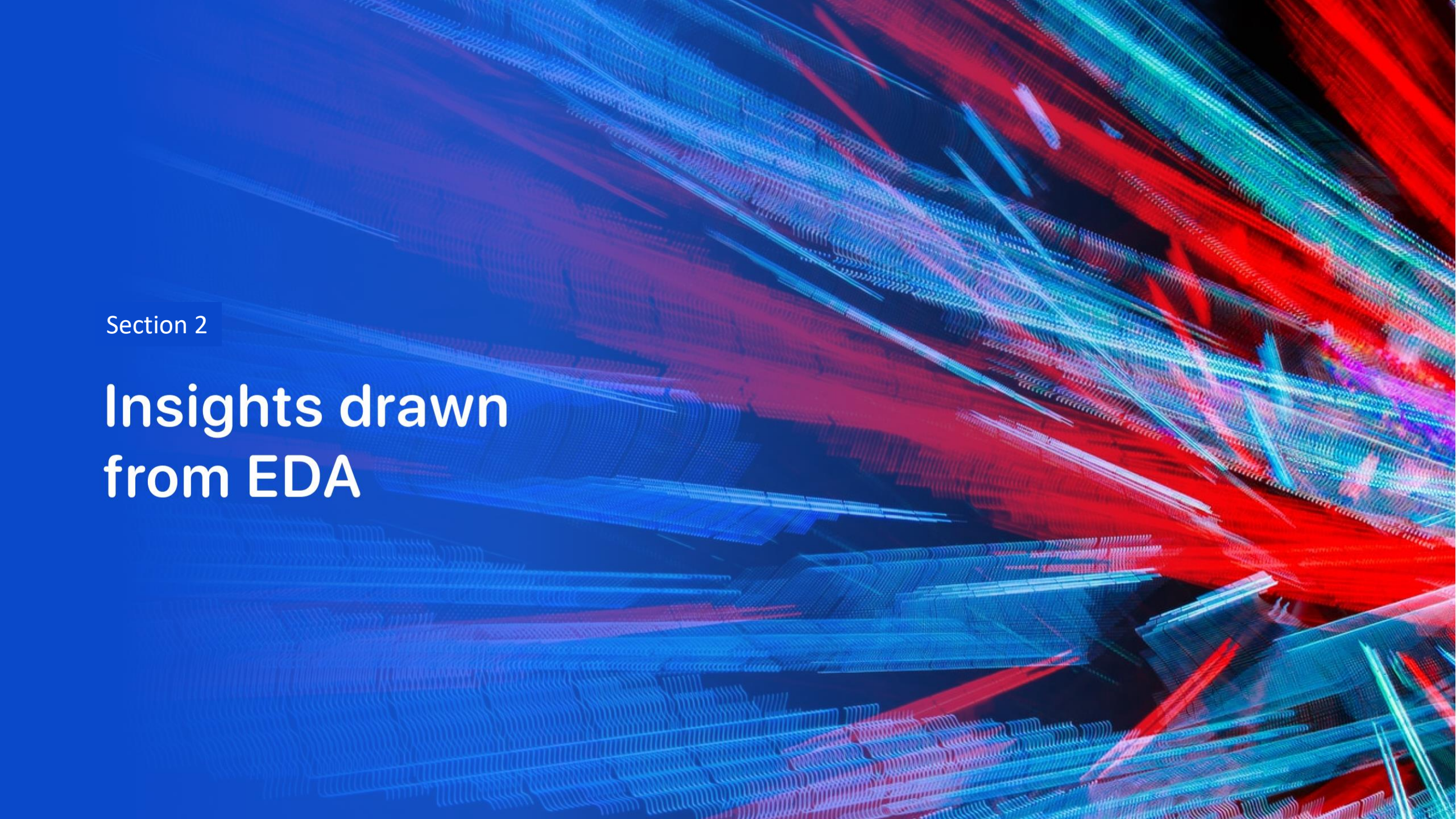
Results

- Exploratory data analysis results - The success rate since 2013 kept increasing till 2020
- Interactive analytics demo in screenshots - KSC LC-39A site has highest success ratio. 2-4k Payload mass range having good success ratio.
- Predictive analysis results – Best Model Confusion Matrix shown.



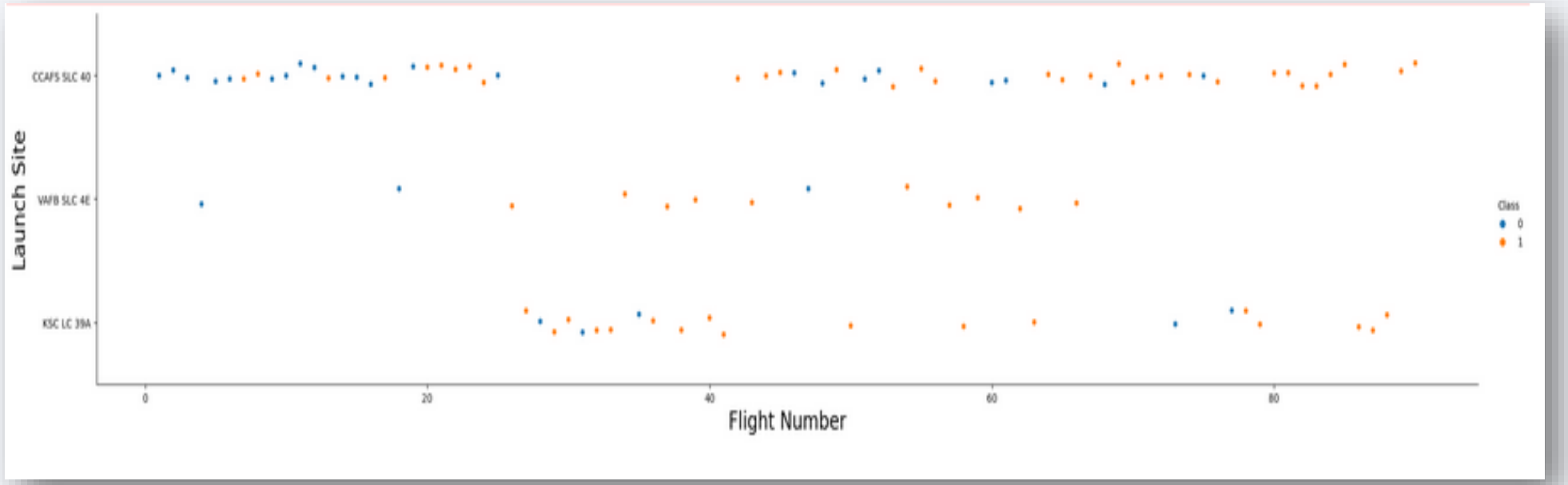





Section 2
Insights drawn from EDA

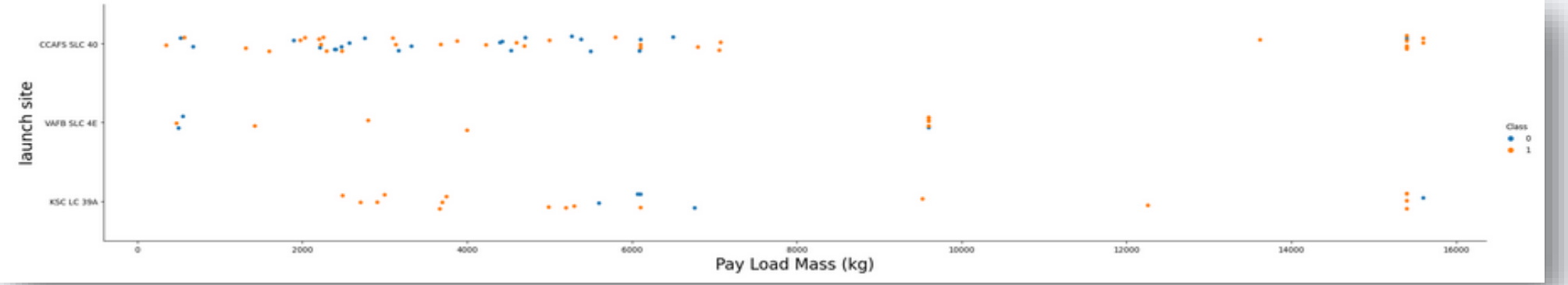
Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site
- Success ratio is increasing along with flight number



Payload vs. Launch Site

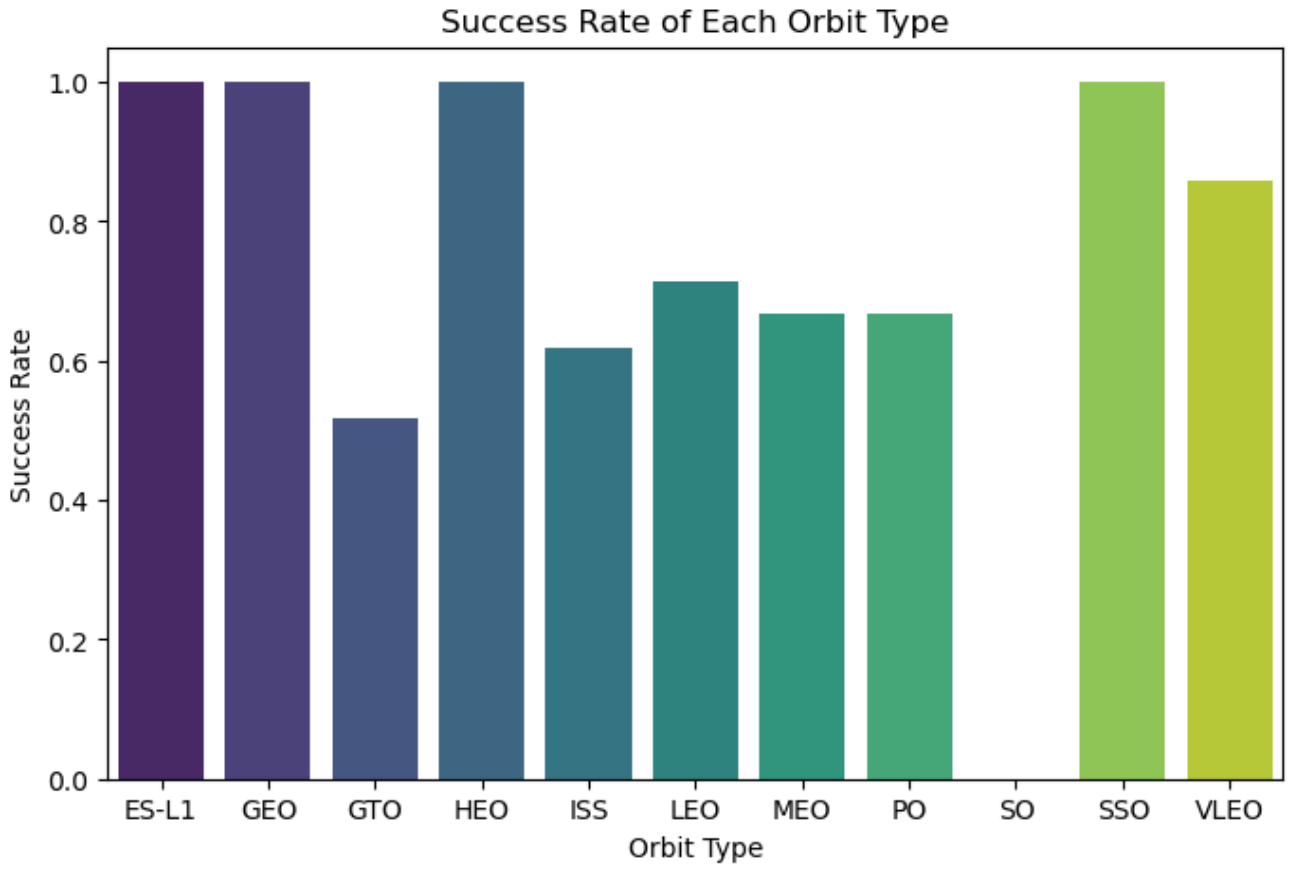
- Show a scatter plot of Payload vs. Launch Site
- VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).



Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

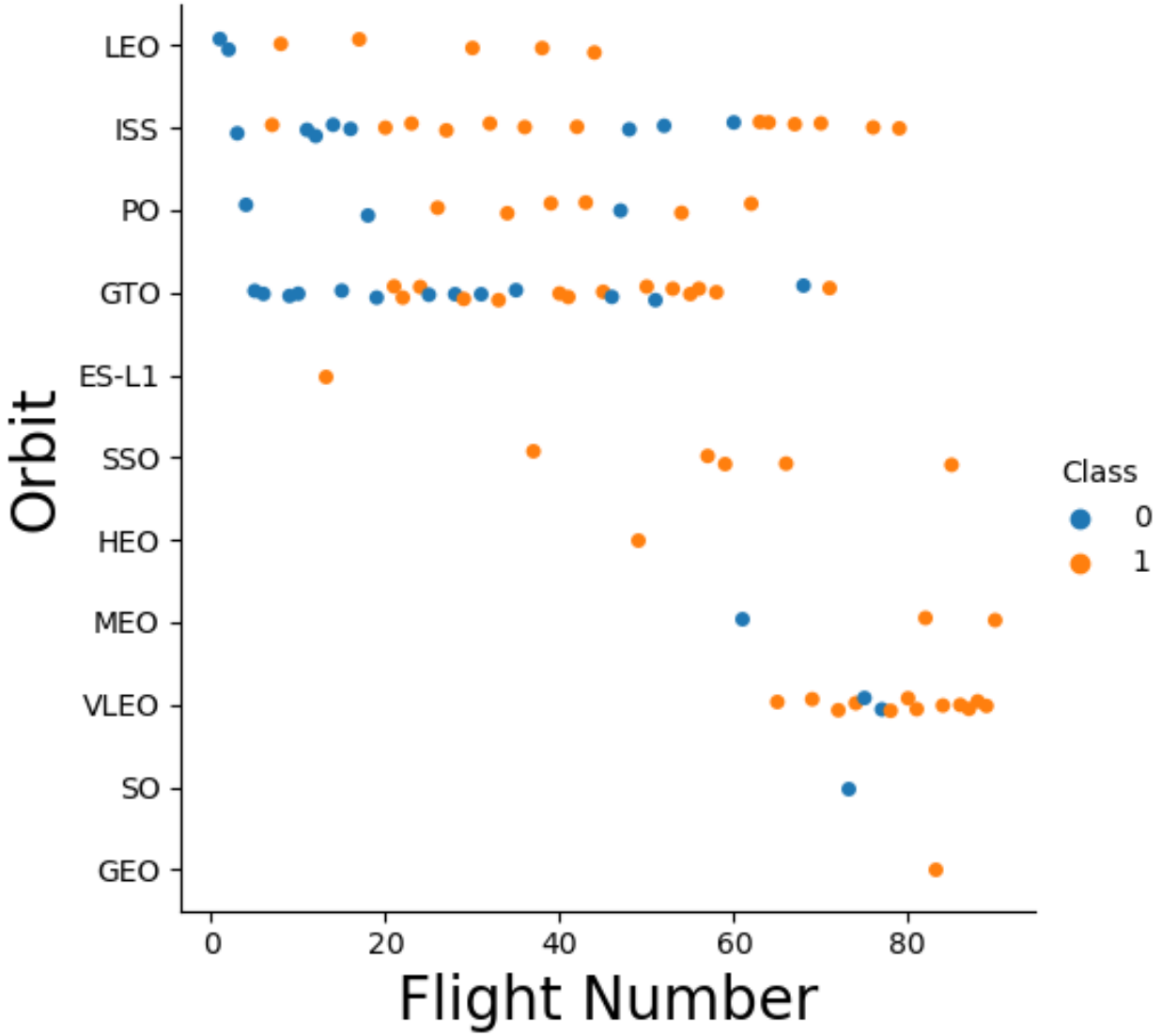
- ES-L1, GEO, HEO, SSO having higher success ration as compared to others.



Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

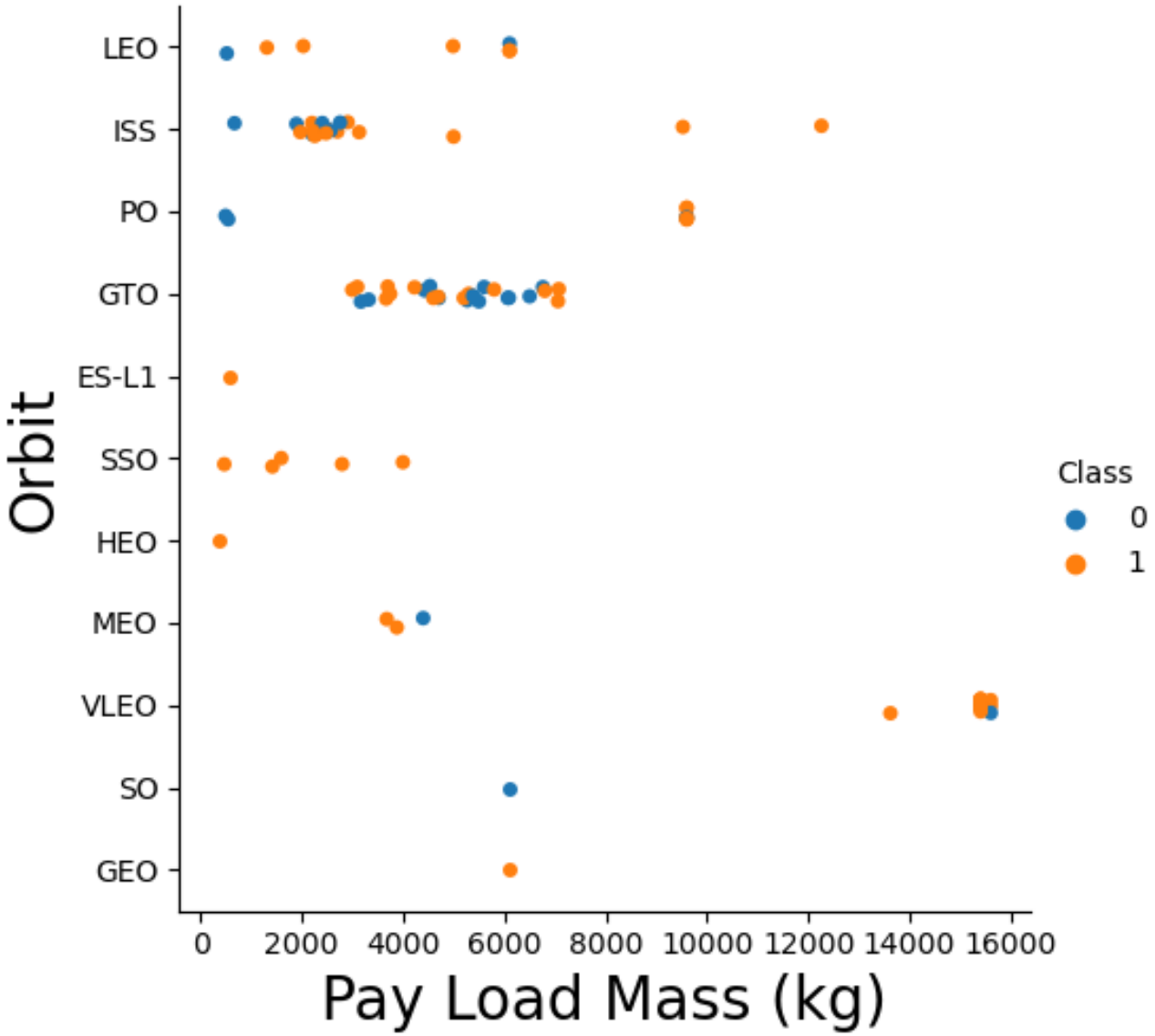
- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

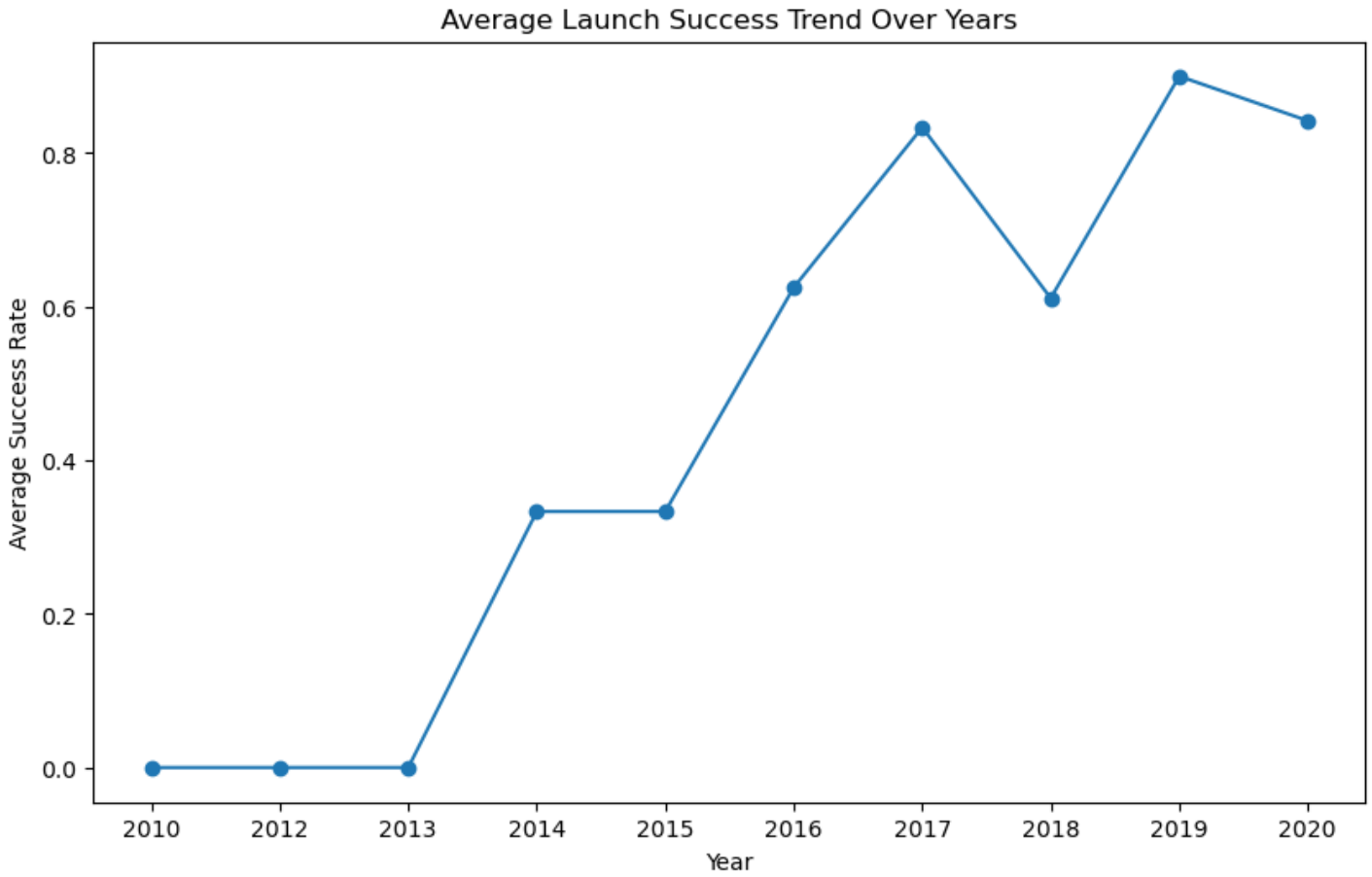
- GTO we cannot distinguish as we have both values.
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS



Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- The success rate since 2013 kept increasing till 2020



All Launch Site Names

- Unique launch sites
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- This give unique launch sitename from spacextable. Disticnt keywork gives unique values.

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
6/4/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
12/8/2010	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)
12/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
10/8/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
3/1/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Limit function used to get 5 records.

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' limit 5;
```

Total Payload Mass

- Total Payload Mass carried by boosters launched by NASA

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

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

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Query use **AVG** SQL function to calculate average

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

```
%%sql SELECT AVG(payload_mass__kg_) AS average_payload_mass
FROM SPACEXTBL
WHERE booster_version = 'F9 v1.1'
```

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

average_payload_mass
2928.4

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad as shown below

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%sql

SELECT Date as first_successful_landing_date
--min(Date) as first_successful_landing_date
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)' limit 1;
```

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

first_successful_landing_date
22/12/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2
- Used multiple and clauses

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes
- Used aggregate functions.

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome,count(1) from SPACEXTBL group by Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- SQL Sub query is used.

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

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Used **substr** and **LENGTH** function to get Year

```
%%sql
SELECT substr(Date, LENGTH(Date) - 3) year, Date,Landing_Outcome,Booster_Version, Launch_Site
FROM SPACEXTBL
WHERE substr(Date, LENGTH(Date) - 3)='2015' and Landing_Outcome = 'Failure (drone ship)'
```

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

year	Date	Landing_Outcome	Booster_Version	Launch_Site
2015	1/10/2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	14/04/2015	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
- Use many SQL function such like BETWEEN, COUNT, GROUP BY, DENSE_RANK…

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.

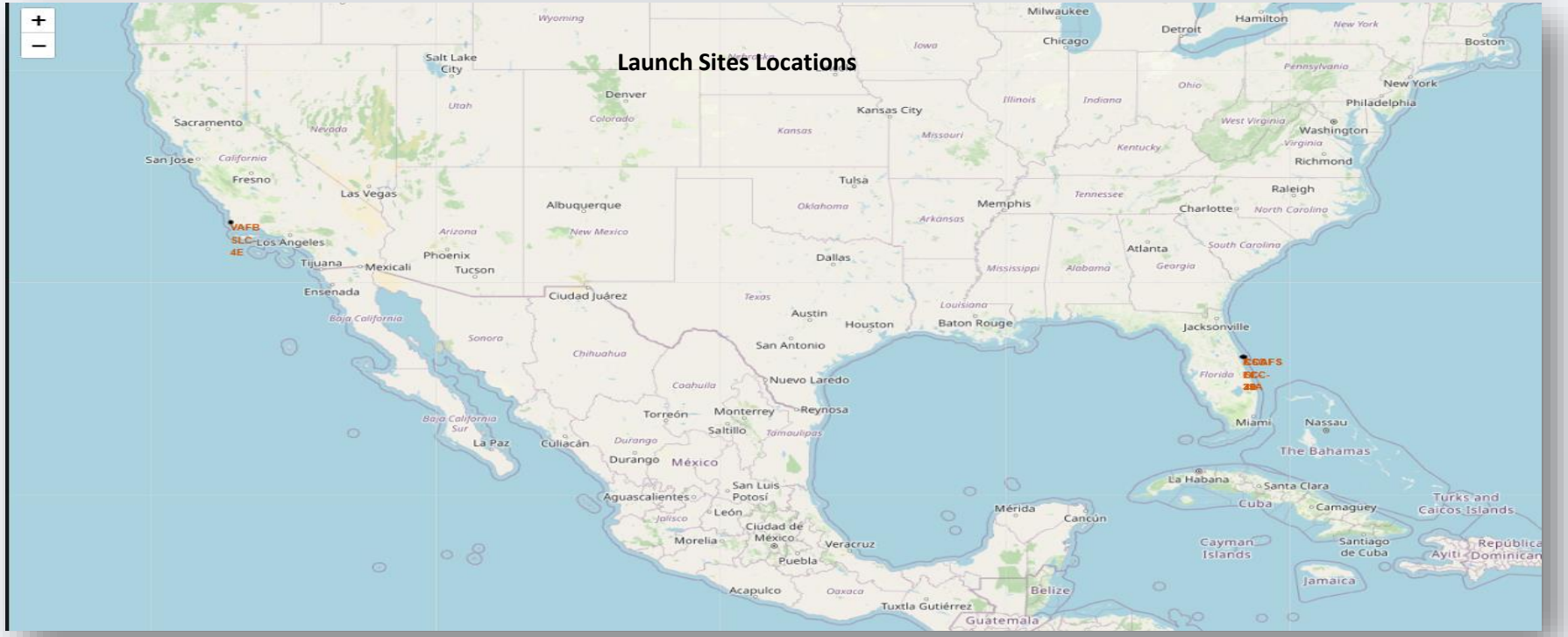
```
%%sql
WITH OutcomeCounts AS (
    SELECT
        Landing_Outcome,
        COUNT(*) AS outcome_count
    FROM
        SPACEXTBL
    WHERE
        Date BETWEEN '2010-06-04' AND '2017-03-20'
    GROUP BY
        Landing_Outcome
)

SELECT
    Landing_Outcome,
    DENSE_RANK() OVER (ORDER BY outcome_count DESC) AS ranking
FROM
    OutcomeCounts
ORDER BY
    ranking;
```

Section 3

Launch Sites Proximities Analysis

SS1: ALL LAUNCH SITES LOCATIONS LOCATED

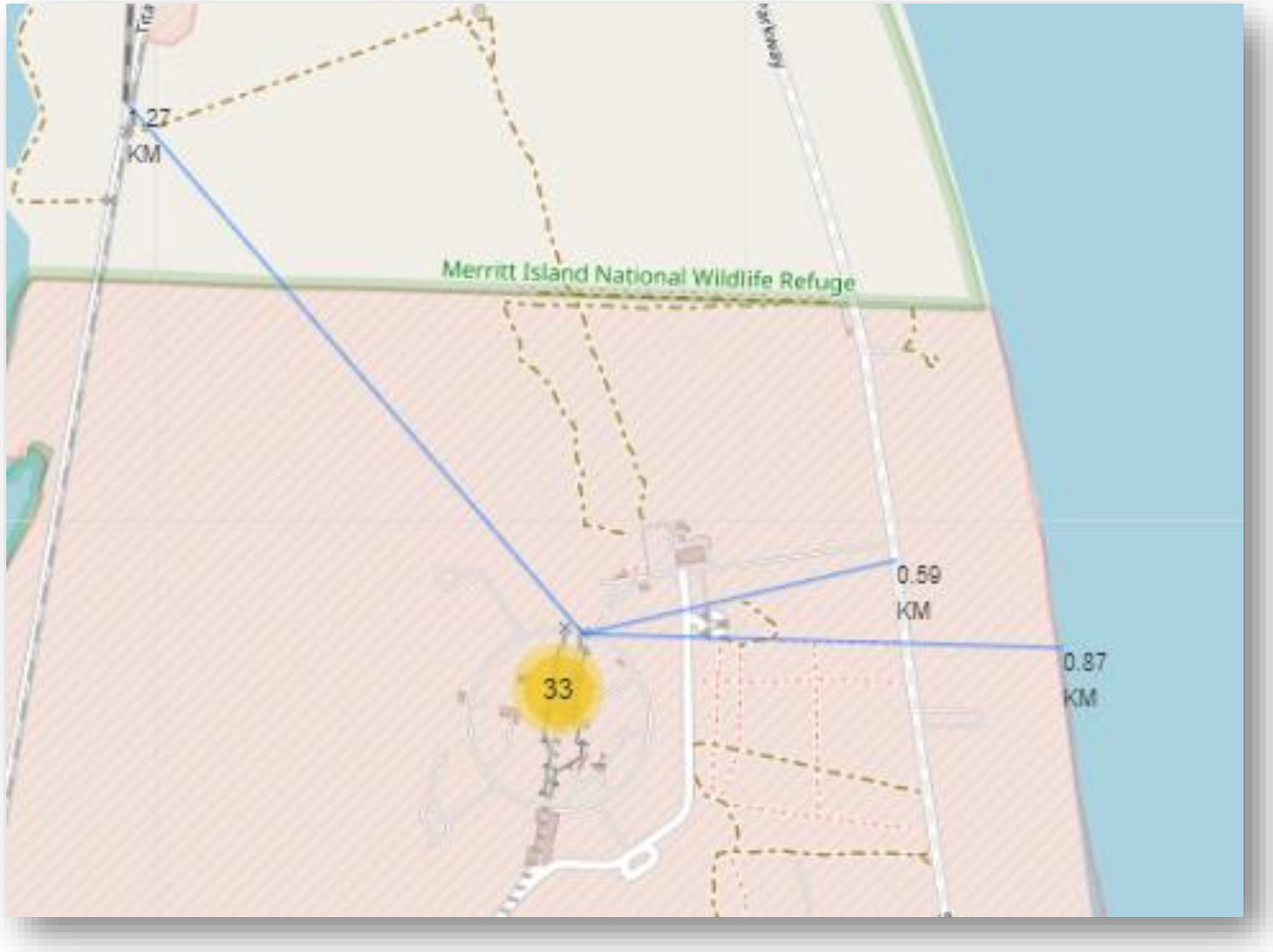


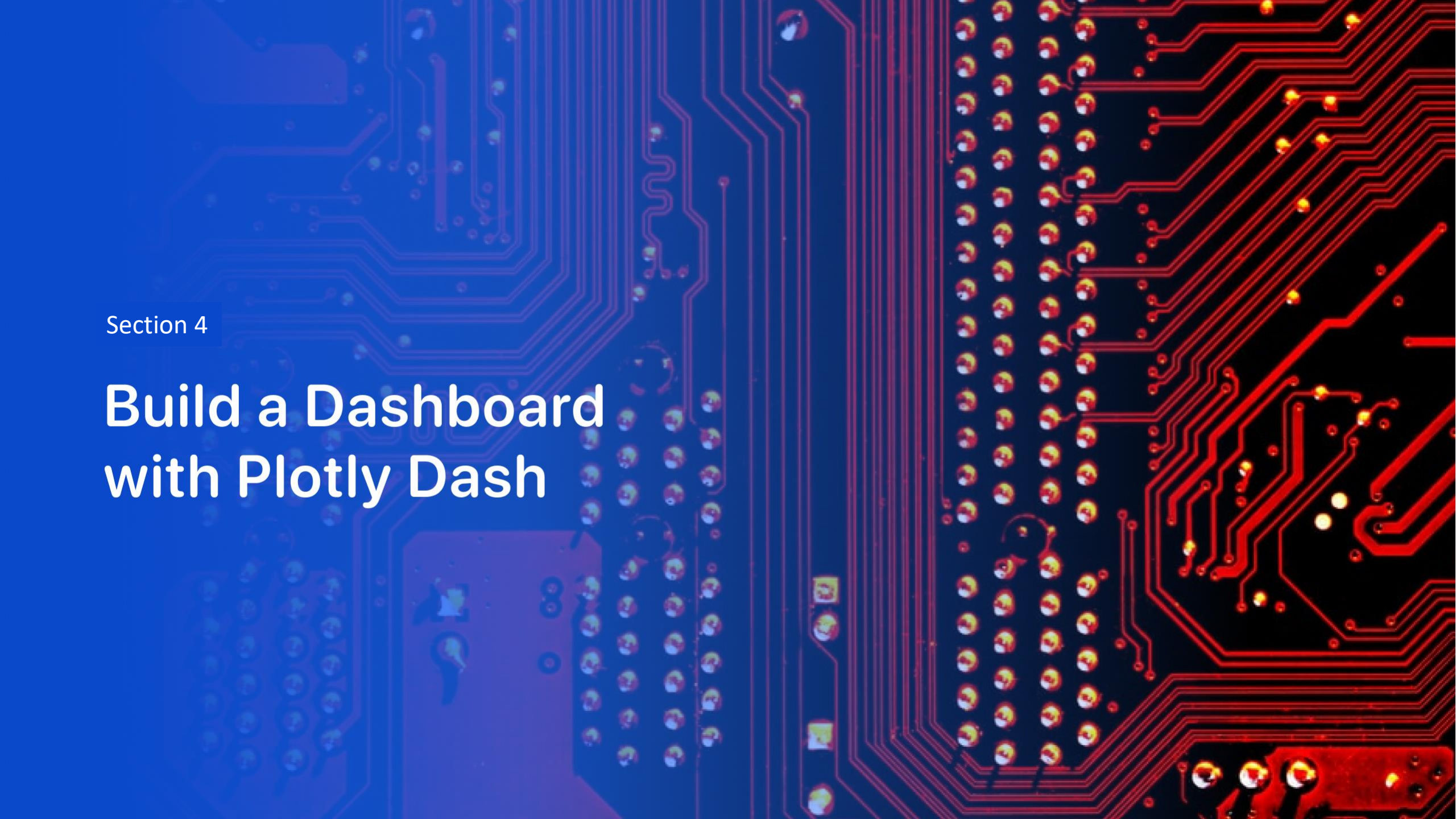
Site SLC 40 Launch Success/Failure

- Below screenshot shows green are successful launched and red are failure.

Launch site to its proximities such as railway, highway, coastline, with distance

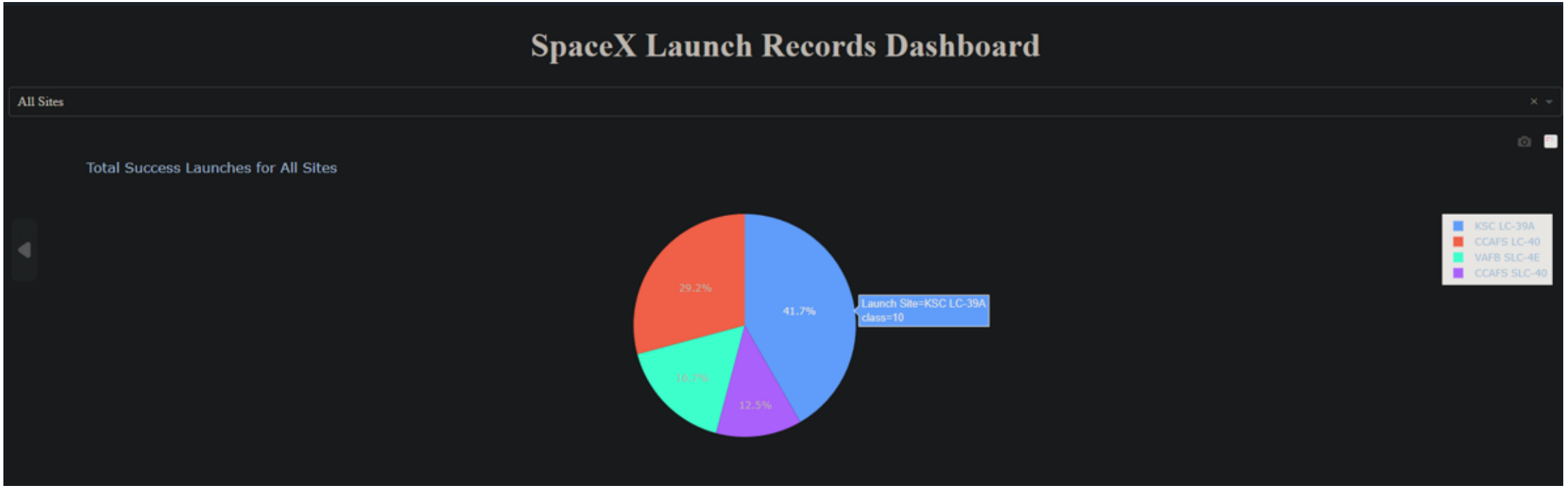
- Site and Railway distance is 1.27 KM
- Site and Coastline distance is 0.87 KM
- Site and Highway distance is 0.59 KM




Section 4
Build a Dashboard with Plotly Dash

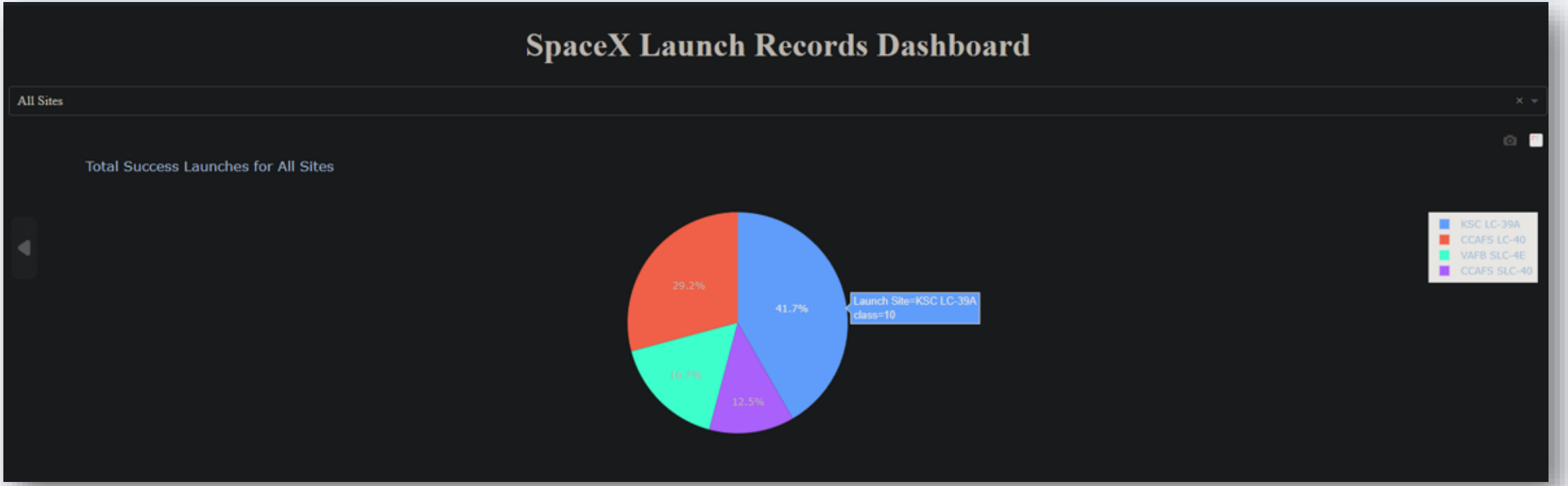
Launch success count for all sites – Pie Chart

- Four sites success launch percentage are shown.
- KSC LC-39A having highest percentage.



KSC LC-39A having highest success ratio.

- It has 41.7% success ratio.



Corelation between Payload and Success for all sites

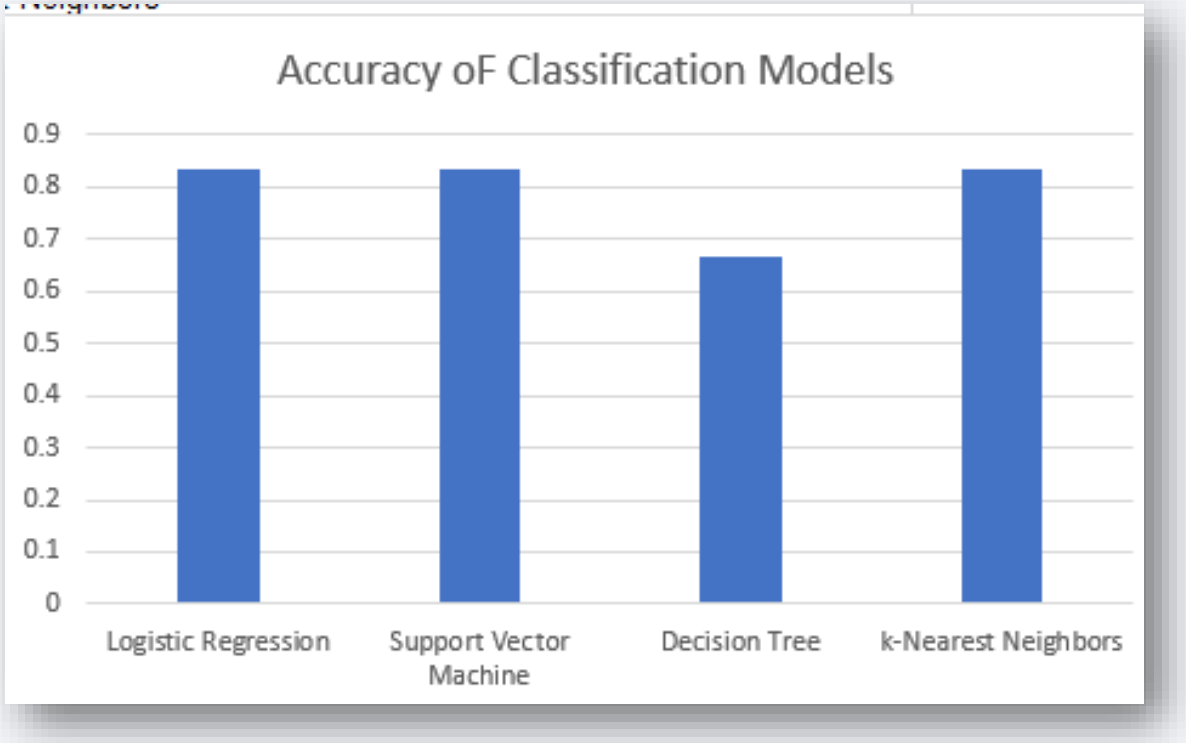
- Shown screenshot of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- FT booster version has highest success, second highest success booster version is B4
- Payload Range 2000-4000 having highest success ratio.

Section 5

Predictive Analysis (Classification)

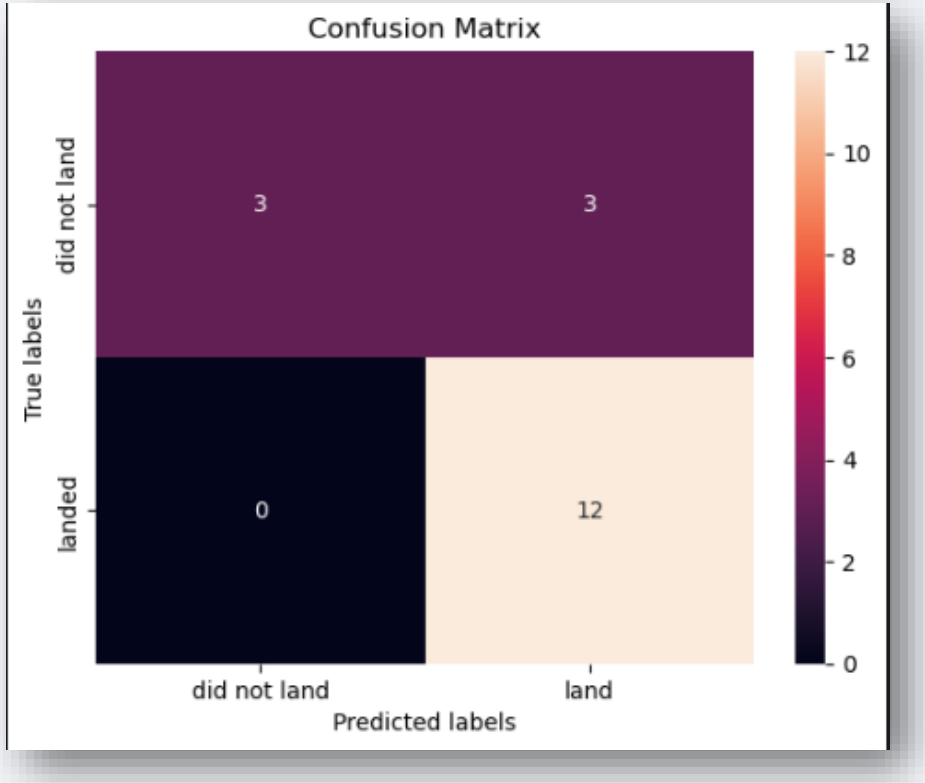
Classification Accuracy

- Logistic Regression, Support Vector Machine, KNN has highest accuracy 0.8333333333333334



Confusion Matrix

- Show the confusion matrix of the best performing mode.
- It has highest sum of true positives and true negatives so This is good model.



Conclusions

- Reusable falcon 9 rocket reduce much cost.
- Payload Mass having in range 2k-3k has higher success ration.
- KSC LC-39A site has highest success ratio.
- FT booster version is most successful version having highest success rate.

Appendix

- Capstone Project Files: https://github.com/radhe004/DataScienceCapstone

Innovative Insights

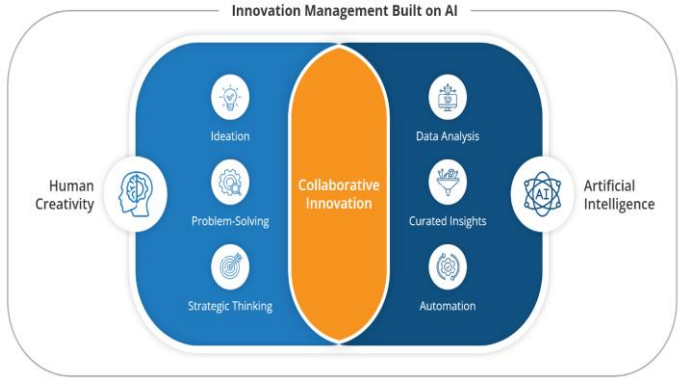
1. **Explainable AI (XAI)**: Increasing focus on making AI systems transparent and understandable, particularly in sensitive domains.

AI for Generative Creativity: Most Utilization of AI in creative fields such as art, music, and literature for generating original content.

AI for Generative Creativity: Exploration of the intersection between quantum computing and AI for more efficient solutions to complex problem

AI-driven Drug Discovery: Acceleration of drug discovery processes through AI analysis of large datasets to identify potential drug candidates.

AI in Healthcare: Integration of AI for diagnostics, personalized medicine, and drug discovery, with applications in medical image analysis and treatment optimization.



Thank you!