IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Josiah Nyongesa
February 17, 2025

Outline

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

Executive Summary

- Summary of methodologies
    - Data Collection
    - Data Wrangling
    - EDA with Data Visualization
    - EDA with SQL
    - Interactive Visualization with Folium
    - Dashboard with Plotly Dash
    - Predictive analysis using classification models
- Summary of all results
    - EDA Results
    - Interactive Analytics
    - Predictive Analysis

Introduction

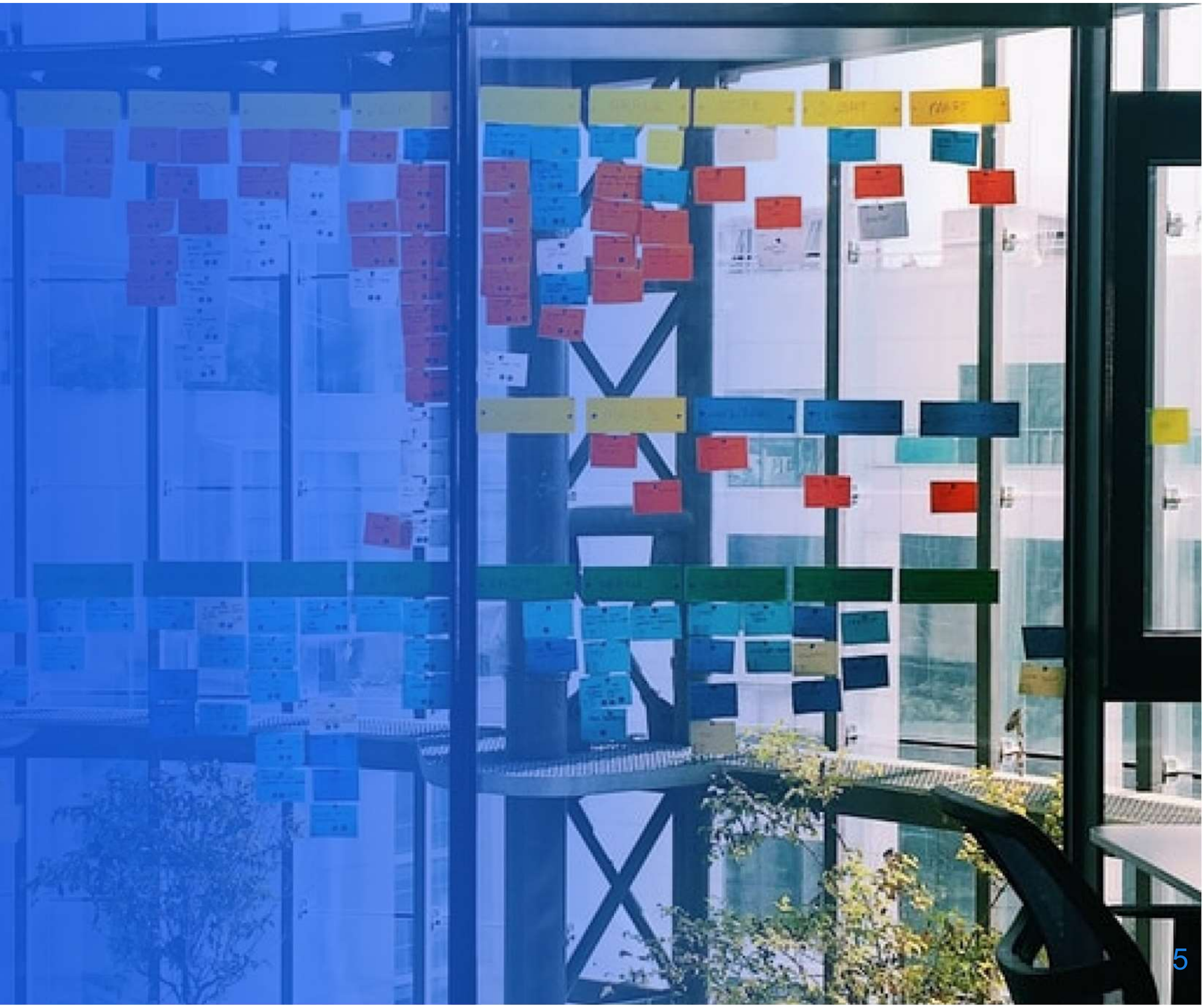
The space industry is evolving rapidly, with SpaceX leading the way in reducing launch costs through reusable rocket technology. As a new competitor, SpaceY aims to optimize its rocket landings using data science. This project applies machine learning and predictive analytics to analyze past launches, identify key success factors, and develop models to improve booster recovery rates.

Business Problem

Rocket reusability is key to lowering launch costs. While SpaceX has cut costs from $150M to $62M per launch, not all boosters are recoverable due to payload weight, orbit type, and landing conditions. To compete, SpaceY must predict booster landings accurately. This project develops a predictive model to enhance cost efficiency and mission success, positioning SpaceY as a strong industry contender.

Section 1

Methodology



Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using web scraping.
- Perform data wrangling
  - Exploratory Data Analysis to find some patterns in the data and determine what would be the label for training supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Exploratory Data Analysis using python library (Matplotlib, Seaborn) and using SQL , Feature Engineering
- Perform interactive visual analytics using Folium and Plotly Dash
  - Use Folium to view previously observed correlations
- Perform predictive analysis using classification models
  - Create a machine learning pipeline to predict if the first stage will land given the data from the preceding observations

Data Collection

Import Libraries and Define Auxiliary Functions → **Request and parse the SpaceX launch data using the GET request** → **Filter the dataframe to only include Falcon 9 launches**

Libraries:	**Functions**:
Request	getBoosterVersion
Pandas	getLaunchSite
Numpy	getPayloadData
Datetime	getCoreData

Decode the response content as a Json using and turn it into a Pandas dataframe

Remove the Falcon 1 launches keeping only the Falcon 9 launches.

Data Collection – SpaceX API

Request rocket launch data from SpaceX API using URL

→

Decode the response content as a Json using and convert it into a pandas dataframe

→

Fixed information in correct columns (using previously defined functions)

→

Construct dataset using the data we have obtained

GITHUB URL : [Data Collection]

Data Collection - Scraping

Request the Falcon9 Launch Wiki page from its URL

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

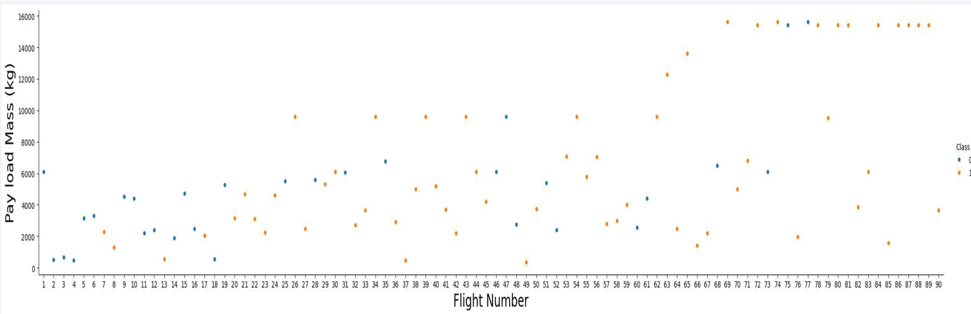
GITHUB URL : [Data Collection]

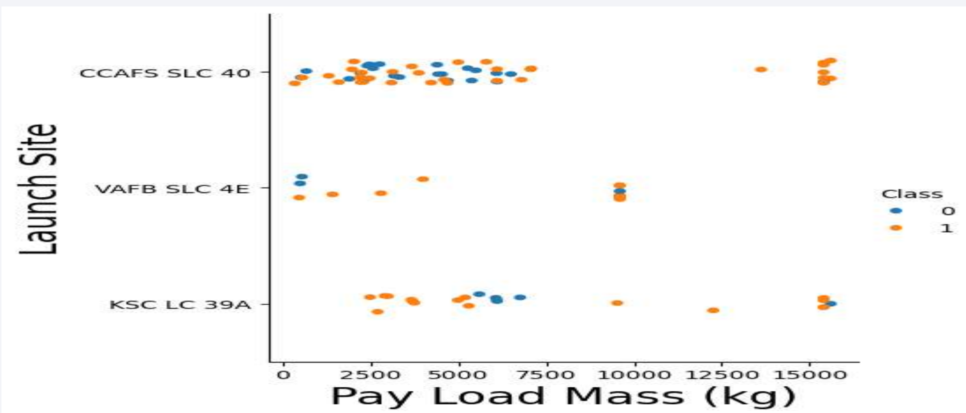
Data Wrangling

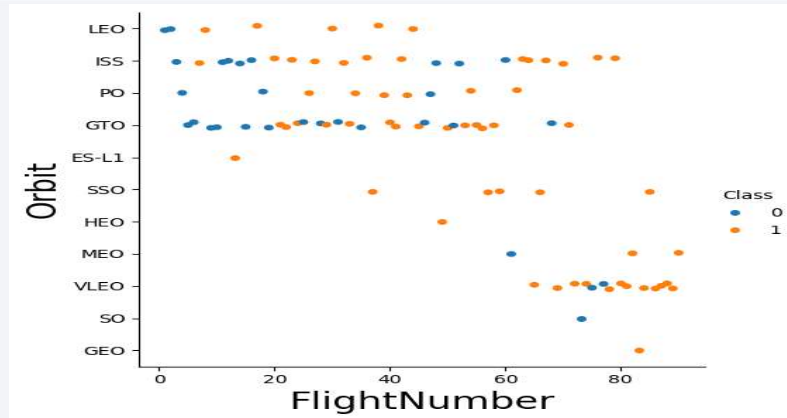
Dealing with Missing Values → Calculate the number of launches on each site → Calculate the number and occurrence of each orbit → Calculate the number and occurence of mission outcome per orbit type → Create a landing outcome label from Outcome column

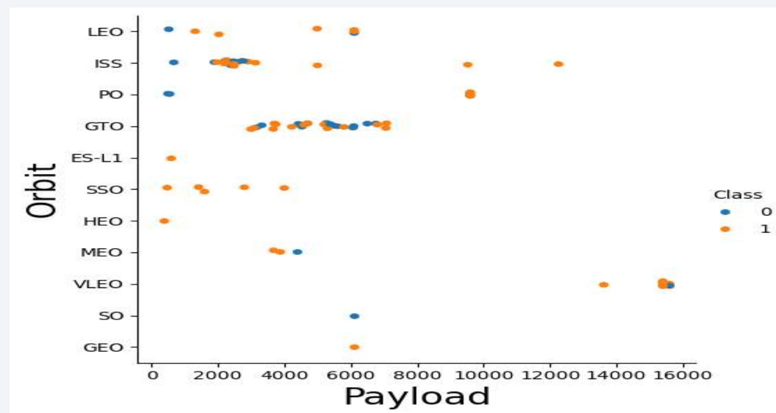
GITHUB URL:[Data Wrangling]

EDA with Data Visualization





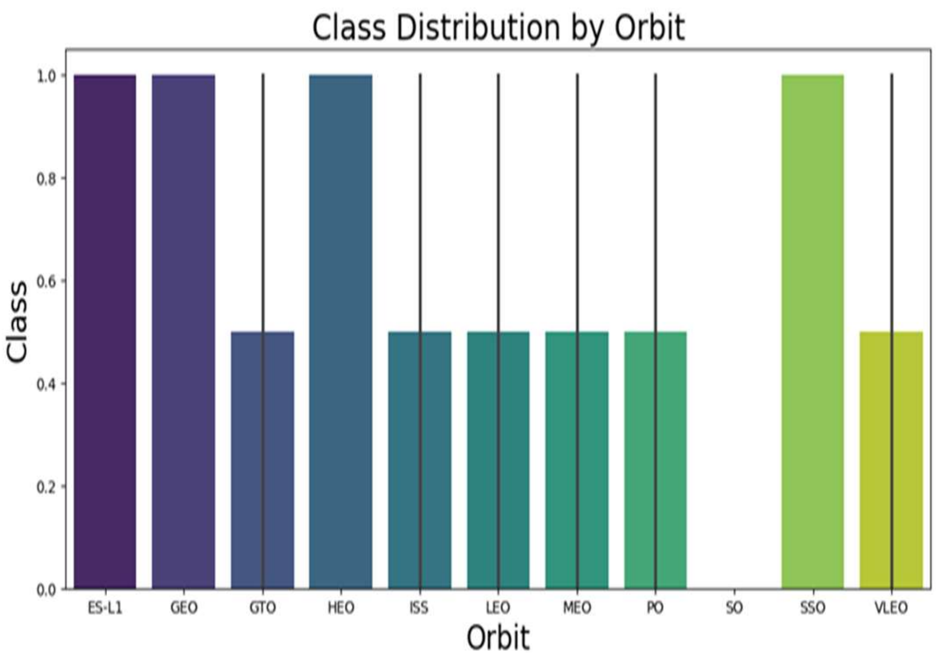


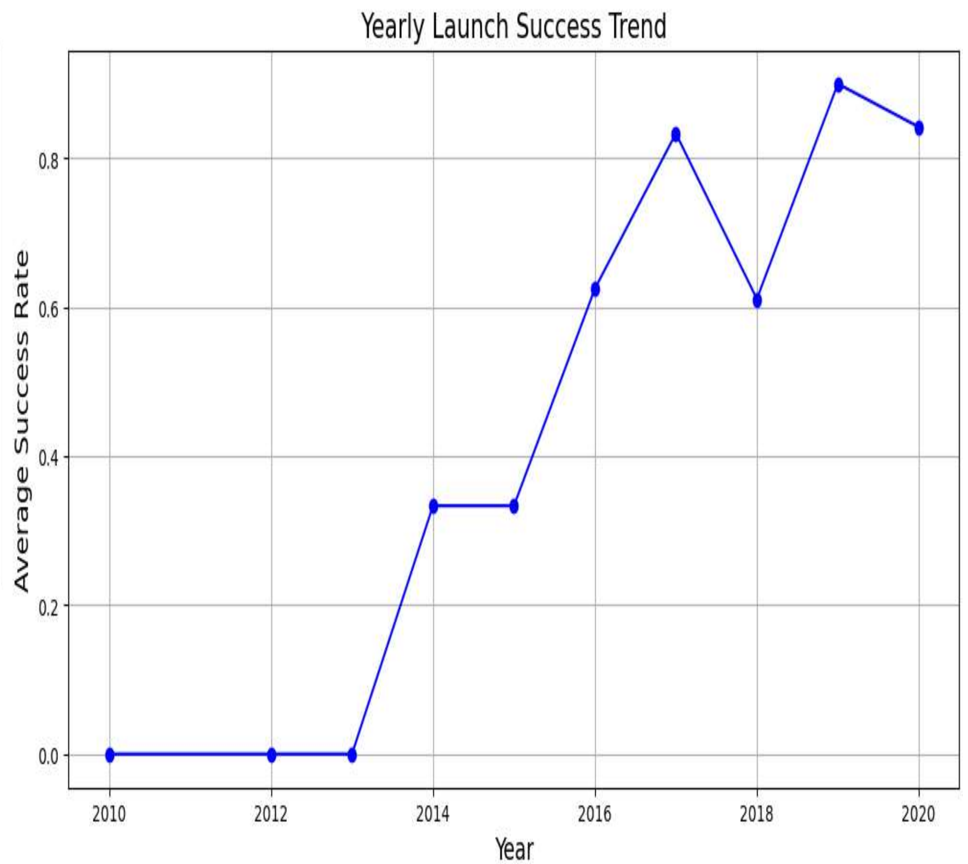


Scatter point charts are used to visualize the relationship between (1) Flight Number and Launch Site, (2) Payload and Launch Site, (3) Flight Number and Orbit type, (4) Payload and Orbit type.

GITHUB URL: [EDA Data Viz]

EDA Viz Cont’d





Bar chart was used to visualize the relationship between success rate of each orbit type; Line chart was used to visualize the launch success yearly trend.

GITHUB URL: [EDA Data Viz]

EDA with SQL

- **SQL queries performed**

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

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

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

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

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

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

7. List the total number of successful and failure mission outcomes

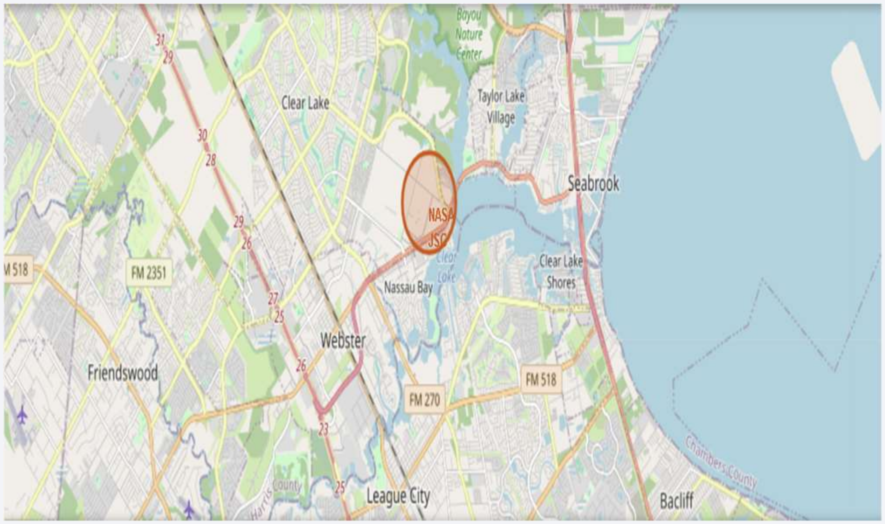
8. List the names of the booster versions which have carried the maximum payload mass using a subquery

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.

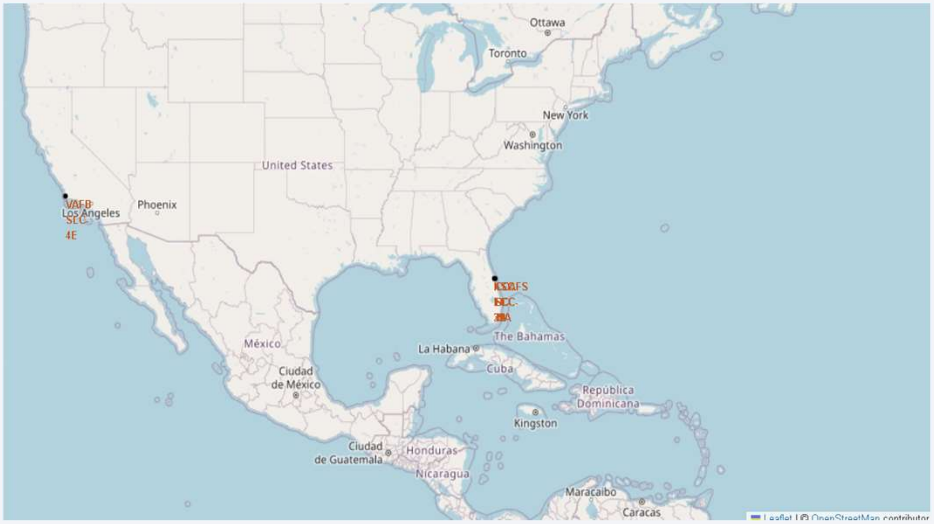
10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

GITHUB URL: [EDA Data Viz]

Build an Interactive Map with Folium



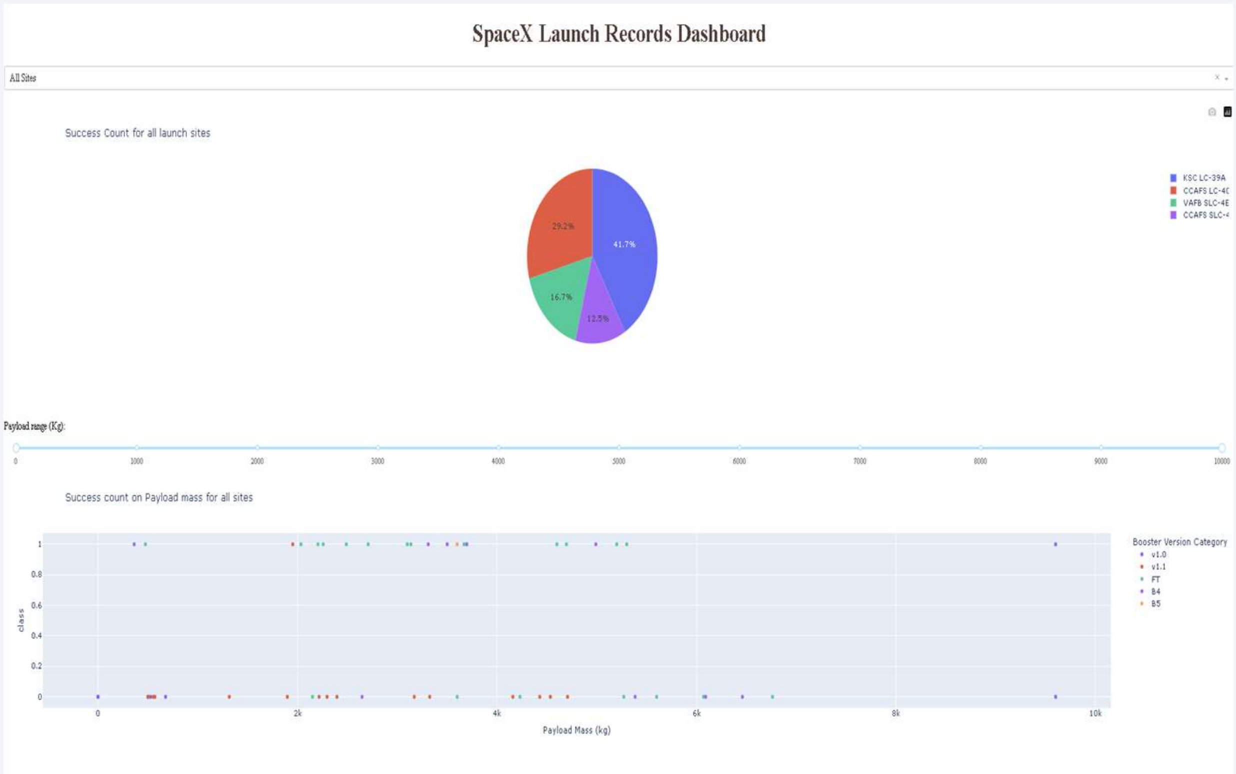
A circle marker was created to show NASA Johnson Space Center's coordinate.



Distance marker was created to show distances between a launch site to its proximities.

GITHUB URL: [Launch site map Visuals]()

Build a Dashboard with Plotly Dash

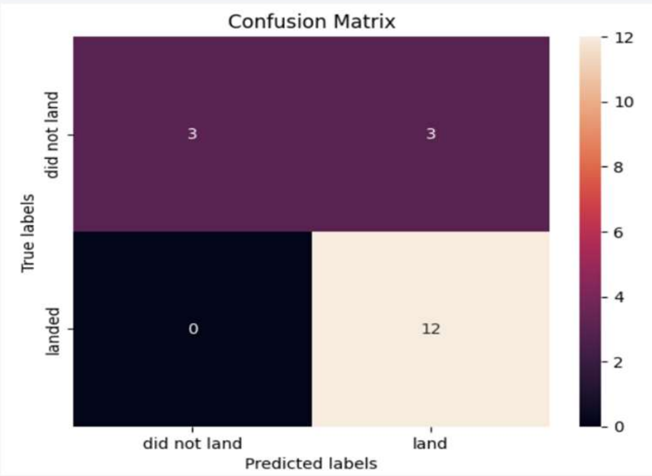


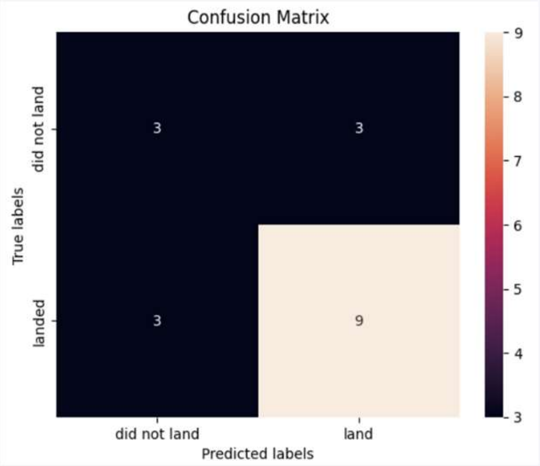
Dropdown option of pie chart created to show the success launches of all / each site

Scatter plot with payload range slider to show the success launches of all / each site by payload mass

GitHub URL: [SpaceX Launch Records Dashboard]()

Predictive Analysis (Classification)



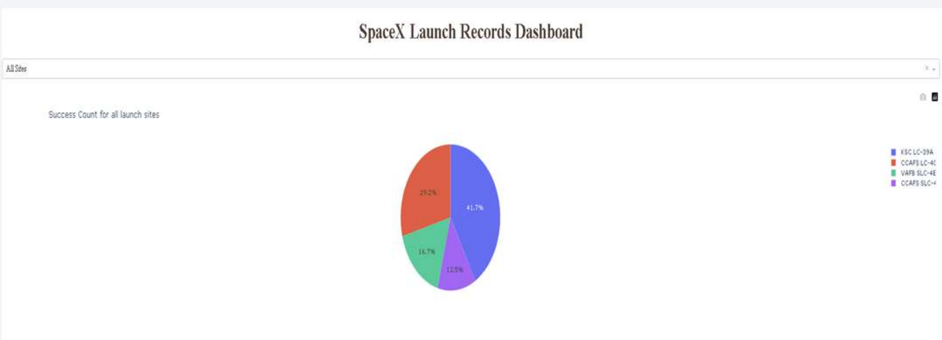


LR, SVM, Decision Tree and KNN models were trained using GridSearchCV to optimize parameters before fitting them on a training set. The confusion matrices for LR, SVM and KNN were identical, while the Decision Tree showed variance.
Model accuracy on test data was evaluated, revealing that LR, SVM and KNN performed best each achieving 83.33% accuracy, whereas the Decision Tree had a lower accuracy of 66.67%.

GITHUB URL: [ML prediction]

Results

- LR, SVM, KNN are top-performing models for forecasting outcomes in this data.
- Lighter payloads have a higher performance compared to heavier ones.
- The likelihood of a SpaceX launch succeeding increases with the number of years of experience, suggesting a trend towards flawless launches over time.
- Launch Complex 39A at Kennedy Space Center has the highest number of successful launches compared to other launch sites.
- GEO,HEO,SSO,ES L1 orbit types exhibit the highest rates of successful launches.

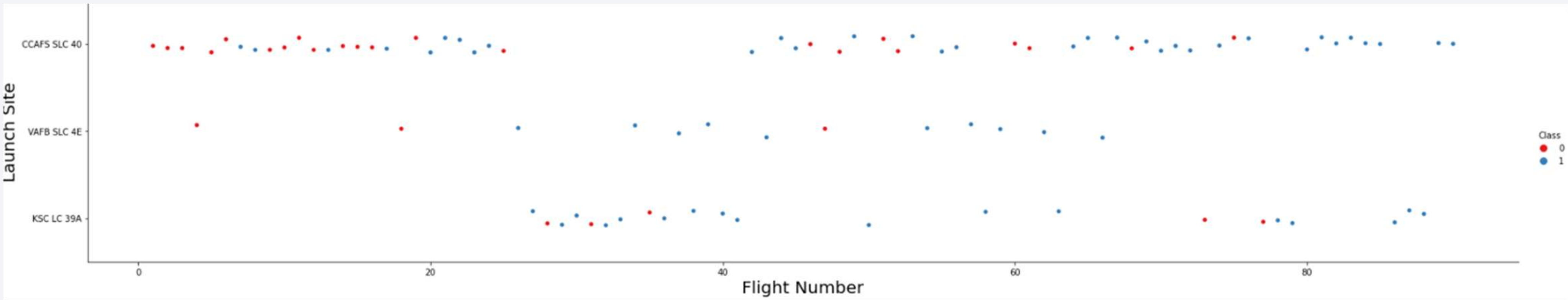


KSC LC-39A has the most successful launches overall

Section 2

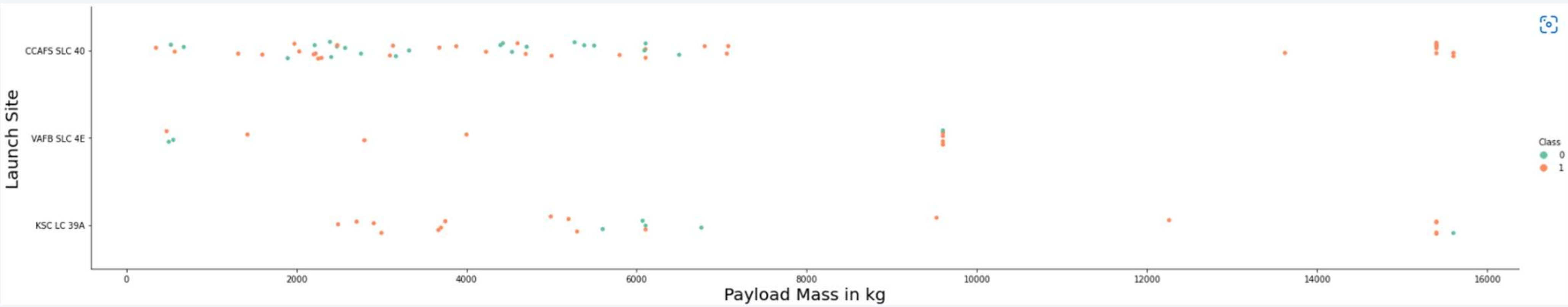
Insights drawn from EDA

Flight Number vs. Launch Site



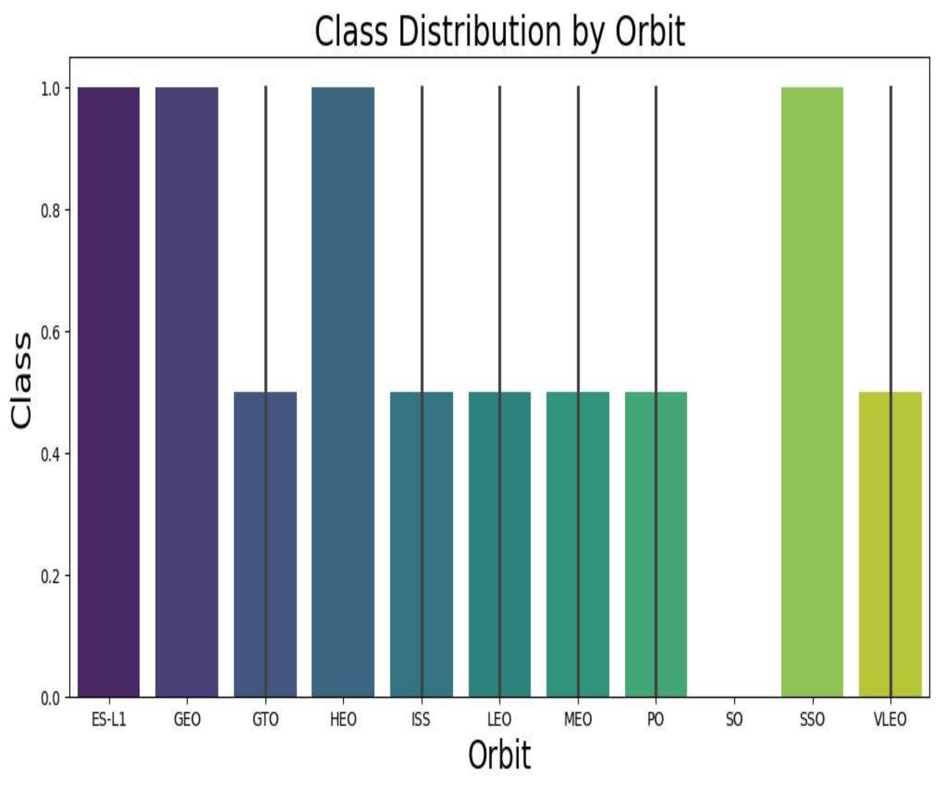
- For the CCFS SLC 40 category there seems to be a higher concentration than class 1 for flights with Flight Number high.

Payload vs. Launch Site



- In this case there does not appear to be strong correlations

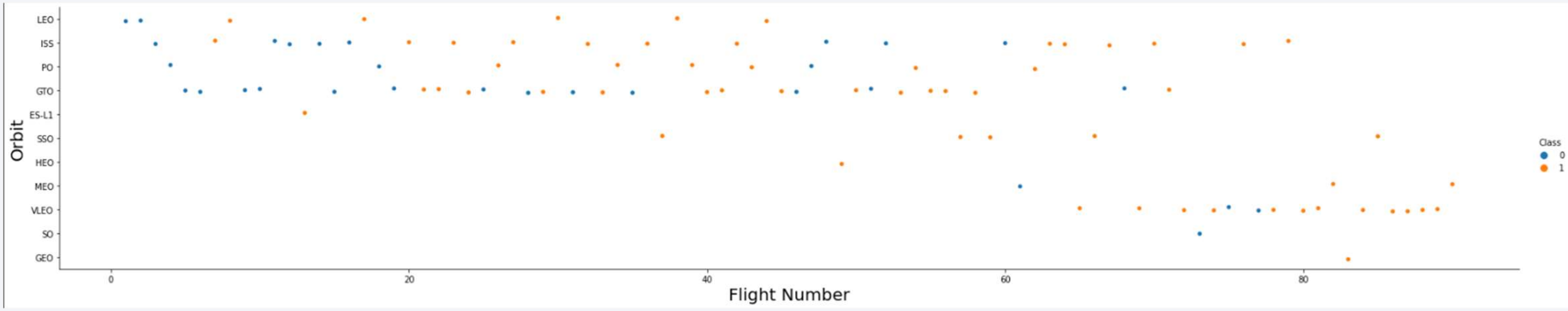
Success Rate vs. Orbit Type



- We can see that the orbits with the highest success rate are:SSO, HEO, GEO, ES-L1

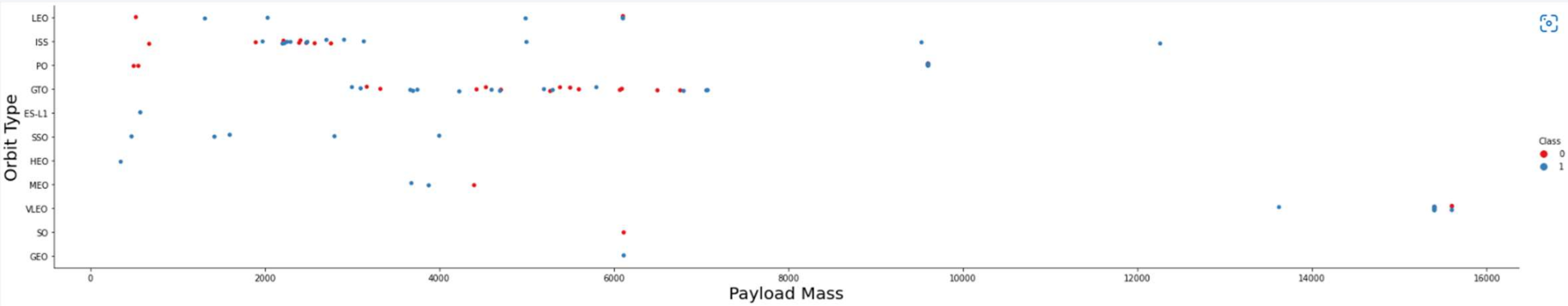
While the GTO Orbit it is the one with lowest rate

Flight Number vs. Orbit Type



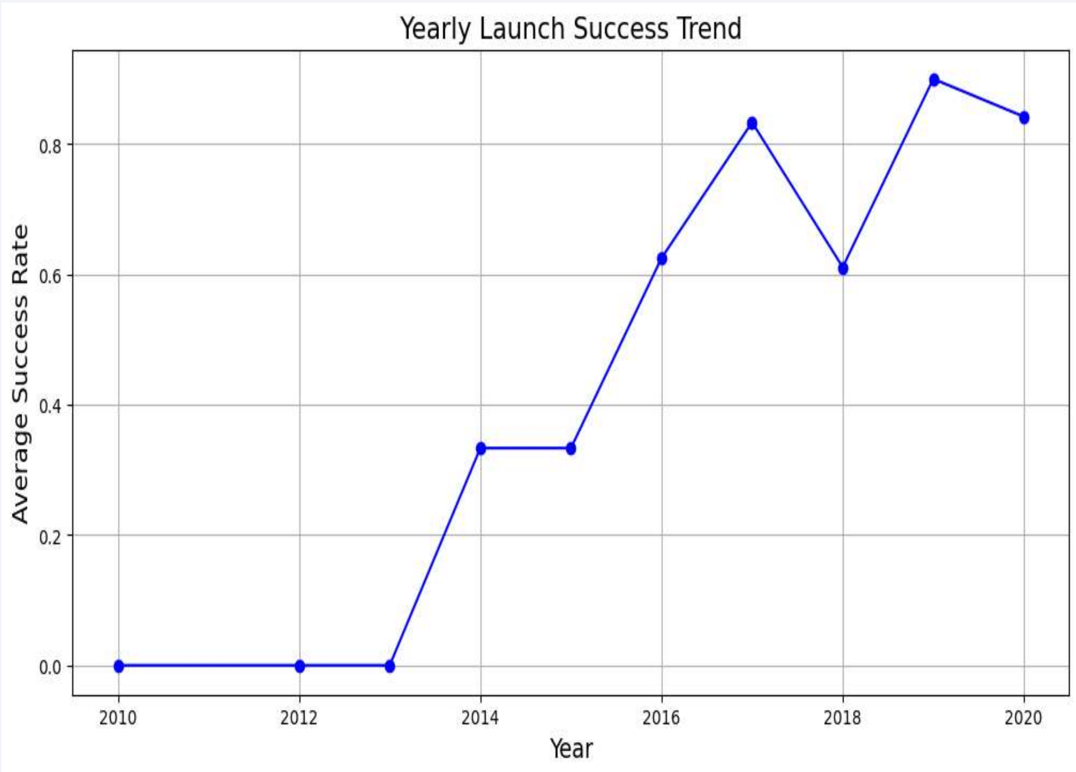
- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand , there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



- You can see that the success rate since 2013 kept increasing till 2020.

All Launch Site Names

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

- This four launch sites are present in the database.

Launch Site Names Begin with 'CCA'

- Records where launch sites begin with `CCA`

DATE	time_utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	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 is 45596.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928

First Successful Ground Landing Date

- First successful landing outcome on ground pad is 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

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

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

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

- This is the names of the booster which have carried the maximum payload mass

2015 Launch Records

landing_outcome	launch_site	booster_version
Failure (drone ship)	CCAFS LC-40	F9 v1.1 B1012
Failure (drone ship)	CCAFS LC-40	F9 v1.1 B1015

- There are two failed landing_outcomes in drone ship, in 2015, and they have the same launch site.

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

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

- We can see that the most numerous are <<No attempt>>,followed by failure/success (drone ship)

Section 3

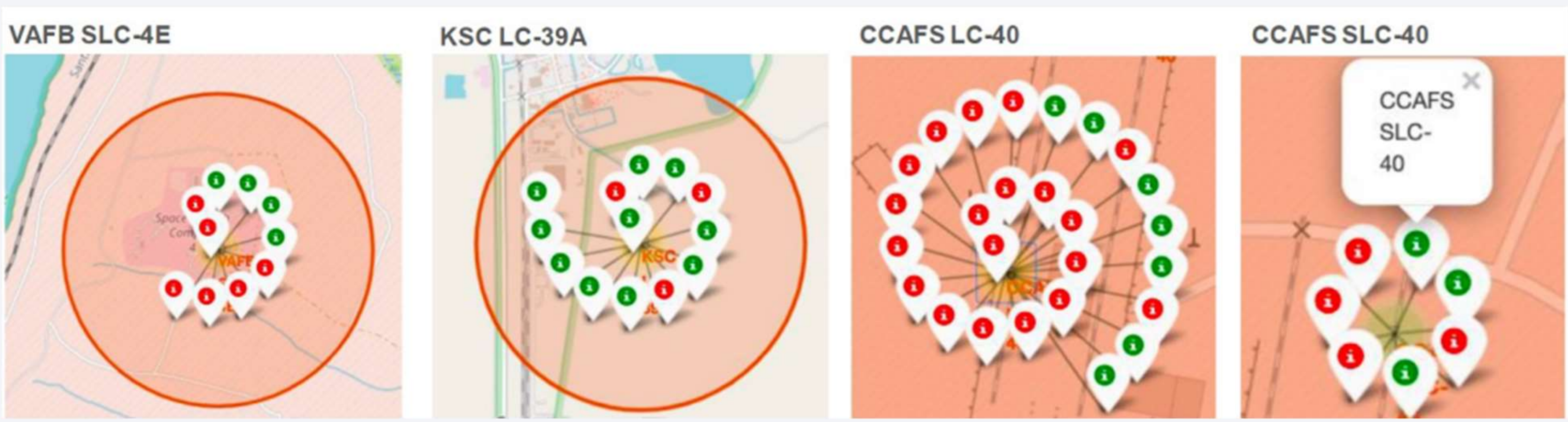
Launch Sites Proximities Analysis

Launch Site

- Both ground and sea surface sites are necessary also south areas maybe a proper areas.
- The transportation base chosen maybe important

Success of launch

- Color icons is impressive way to show the rate of success. For example, in KSC there is the best rate of success.

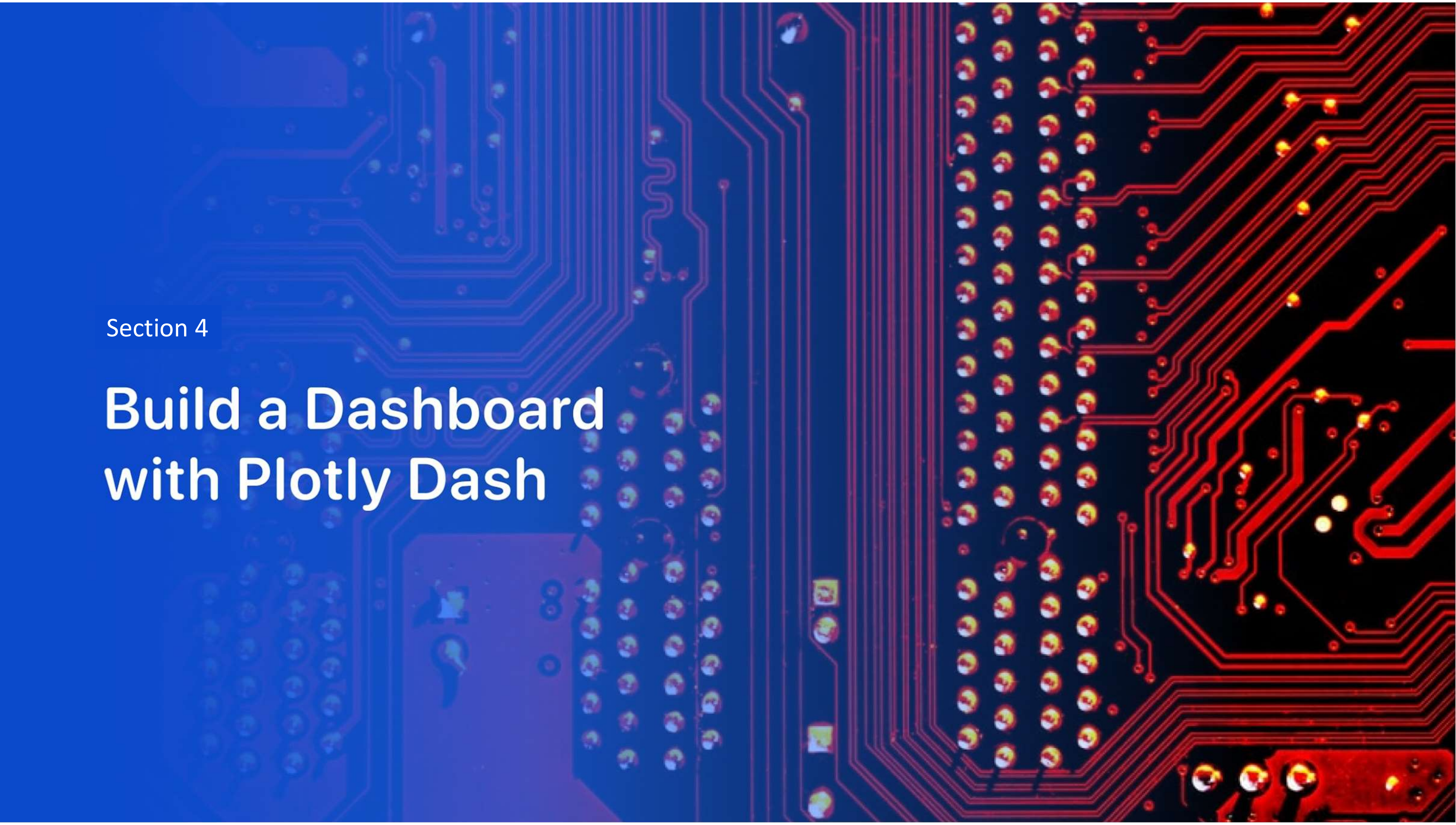


Distance

- This map shows the distance from points of interest. In this case the distance from railway, coast or port may be importance

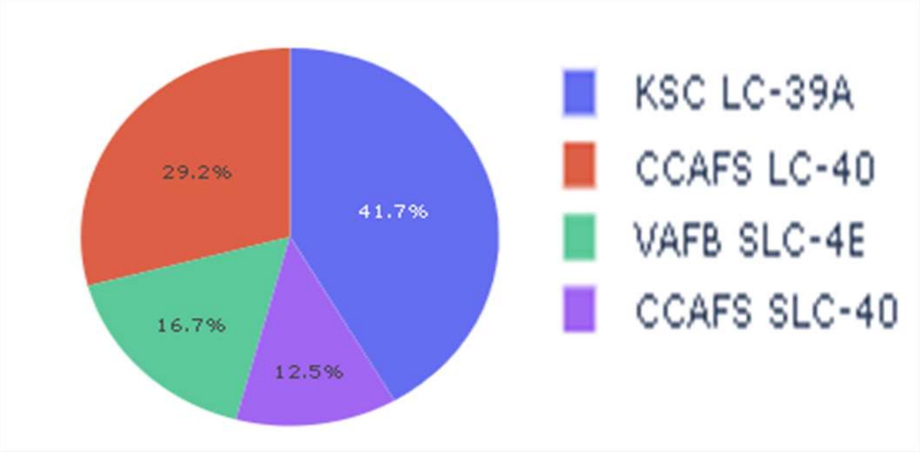
Section 4

Build a Dashboard with Plotly Dash

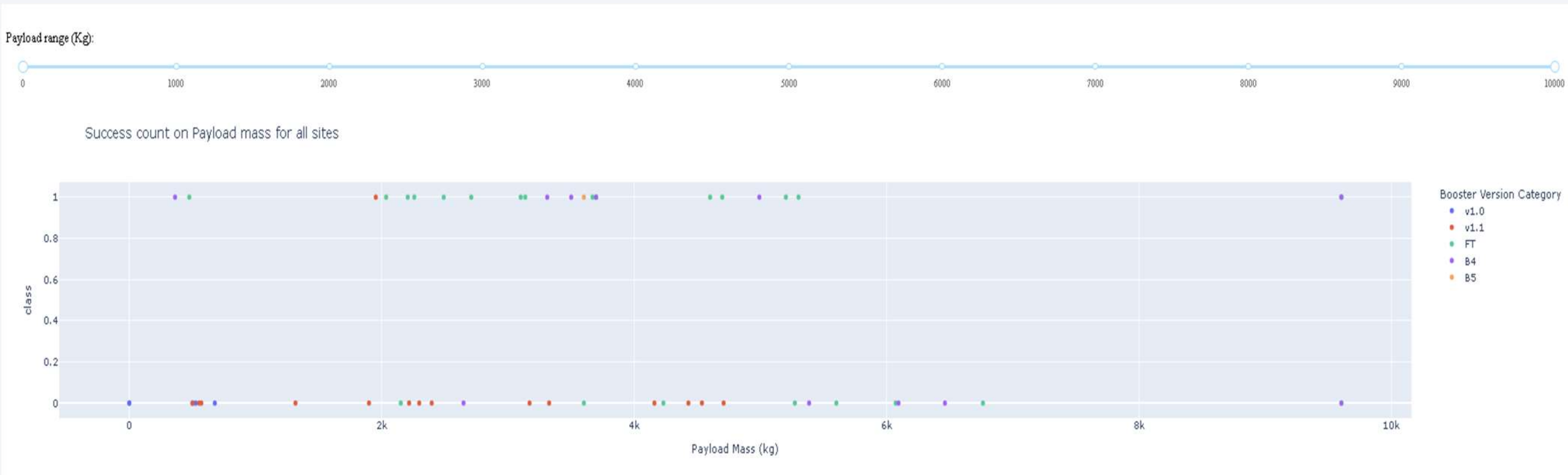


The most successful launch site

- We can see that CCAFS SLC-40 had the most successful launches from all the sites.



Payload and Success launch



• The payload range that has the highest success launches is between 2,000 to 4,000 kg, which can be seen by the most number of plots in that range, followed by the payload range of 4,000 to 6,000 kg, with the second most number of plots.
• Booster version FT (green spots) has the highest success launches, followed by B4 (purple spots) with the second highest success launches, among all booster versions.

Section 5

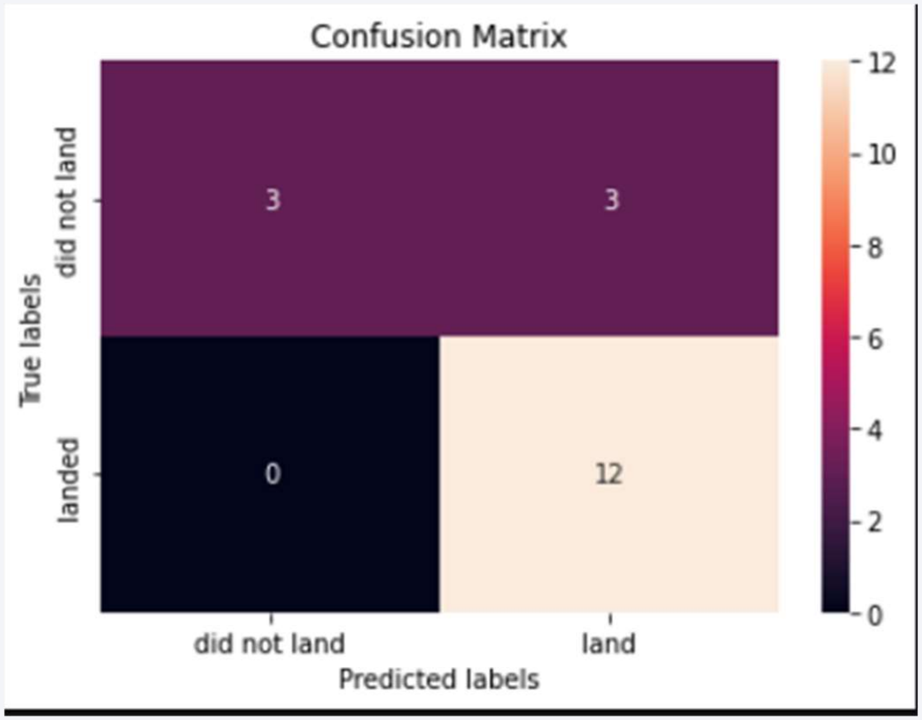
Predictive Analysis (Classification)

Classification Accuracy

- Model accuracy on test data was evaluated, revealing that LR, SVM, and KNN performed best, each achieving 83.33% accuracy, whereas the Decision Tree had a lower accuracy of 66.67%.

```
LR Accuracy: 83.33%
SVM Accuracy: 83.33%
Decision Tree Accuracy: 66.67%
KNN Accuracy: 83.33%
```

Confusion Matrix



Conclusions

This project demonstrates the power of data science and machine learning in optimizing rocket reusability. By analyzing past SpaceX launches, we developed predictive models to determine booster landing success, with Logistic Regression, SVM, and KNN achieving the highest accuracy (83.33%).

Takeaways include:

- Lighter payloads have a higher success rate.
- Launch success have improved over time
- Certain orbit types(GEO, HEO, SSO, ES-L1) show higher landing success.

By leveraging these insights, SpaceY can enhance launch efficiency, minimize costs, and compete effectively in the commercial space industry. Future improvements could include incorporating real-time launch conditions and deep learning models for even more accurate predictions.

.

Appendix

- GitHub link to all the codes, queries and images used. Click here


Thank you!