IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Om Magdum
24.06.2024

Outline

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

Executive Summary

Methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with SQL
- EDA using Pandas and Matplotlib
- Interactive Map with Folium
- Interactive dashboard with Plotly Dash
- Predictive Analysis

Results

- EDA Results
- Interactive Analytics (Demo Screenshots)
- Predictive Analysis Results

Introduction

Project Background and Context

The commercial space age thrives on affordability. SpaceX's reusable Falcon 9 rockets, priced at $62 million compared to competitor costs exceeding $165 million, represent a significant breakthrough. This project harnesses machine learning to predict first-stage reusability, empowering more efficient launch cost estimations.

Questions To Be Answered

Can we leverage factors like payload weight, launch location, flight history, and target orbit to predict a successful first-stage landing for SpaceX Falcon 9 rockets?

Are SpaceX becoming more proficient at landing first stages? Analysing trends in successful landings over time.

Identifying the most effective machine learning model for classifying successful first-stage landings.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Rocket Launch Data through SpaceX REST API & Falcon 9 Data by Web Scraping Wikipedia
- Perform data wrangling
  - Identifying orbit frequencies, successful vs. failed landings, and labelling missions accordingly.
- 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

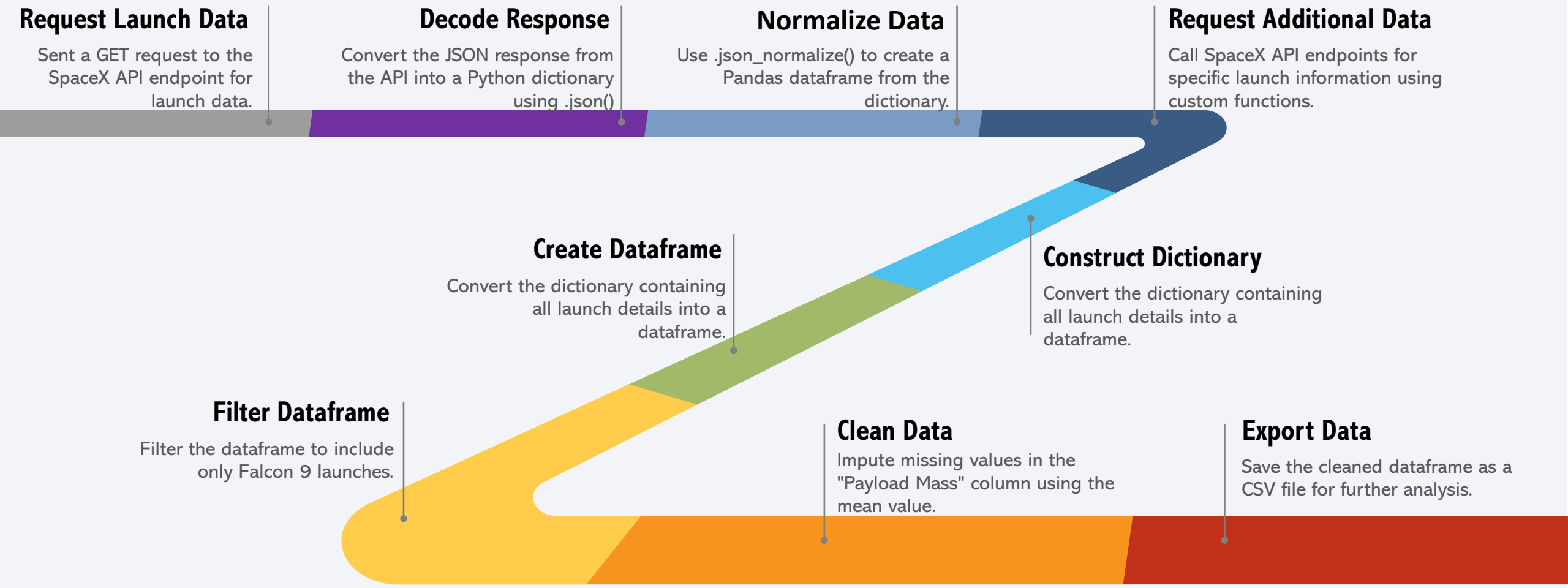
Data Collection

The data acquisition strategy incorporated using both the SpaceX REST API and Web Scraping comprehensive historical launch information about Falcon 9 from SpaceX's Wikipedia ensuring access to the latest launch information. All the obtained data columns by respective techniques are as follows :

Obtained Data Columns

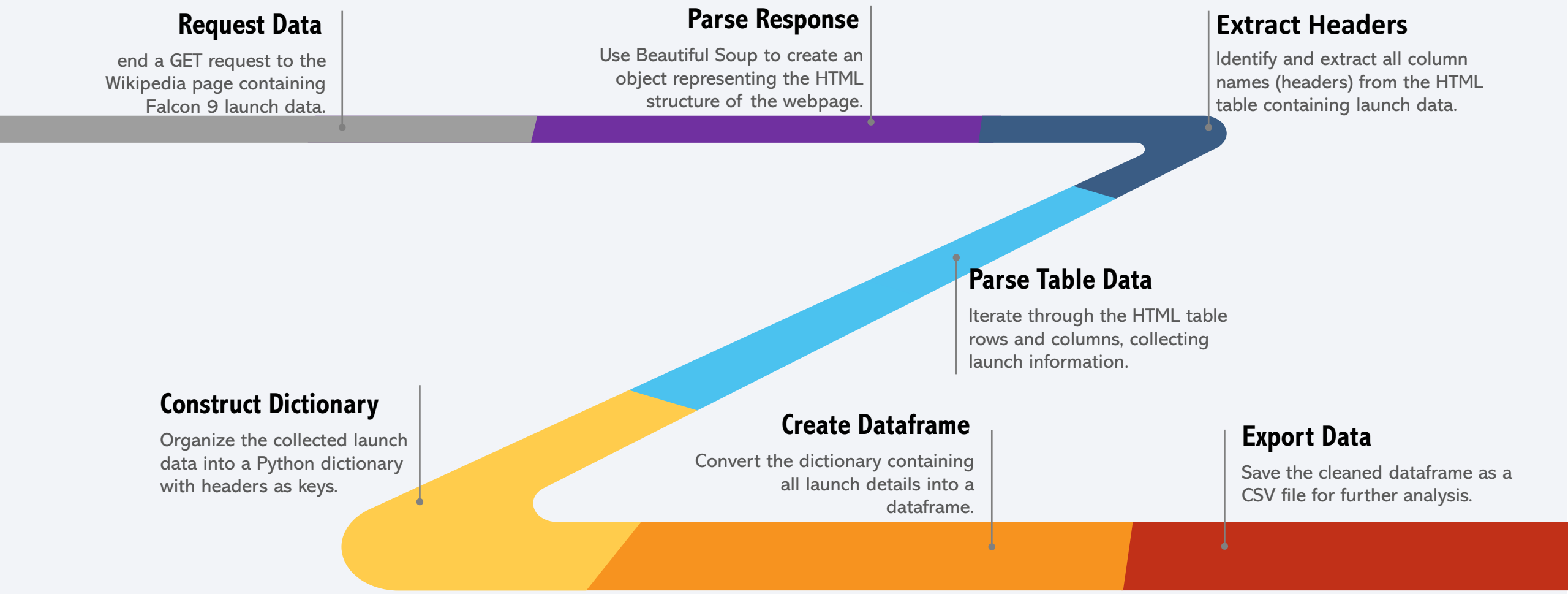
SpaceX REST API	Web Scraping
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude	Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub Link : Data Collection API

Data Collection - Scraping



GitHub Link : Data Collection Scraping

Data Wrangling

01

Import Libraries & Define Functions

- Import libraries like Pandas (data manipulation)
- Define functions for data cleaning/ transformation

02

Load SpaceX Dataset

Load the previously collected Falcon 9 launch data

03

Exploratory Data Analysis

Through EDA, we dissect SpaceX launch data: site frequency, popular orbits, and cracking the "Outcome" code (True/ False for landing) to define our model's success metric.

04

Create Training Labels

Transform the "Outcome" column into a binary "Landing_Label." Successful = 1, unsuccessful = 0.This step creates a crucial target variable for model training.

GitHub Link : Data Wrangling

EDA with Data Visualization

01 - Scatter Plot

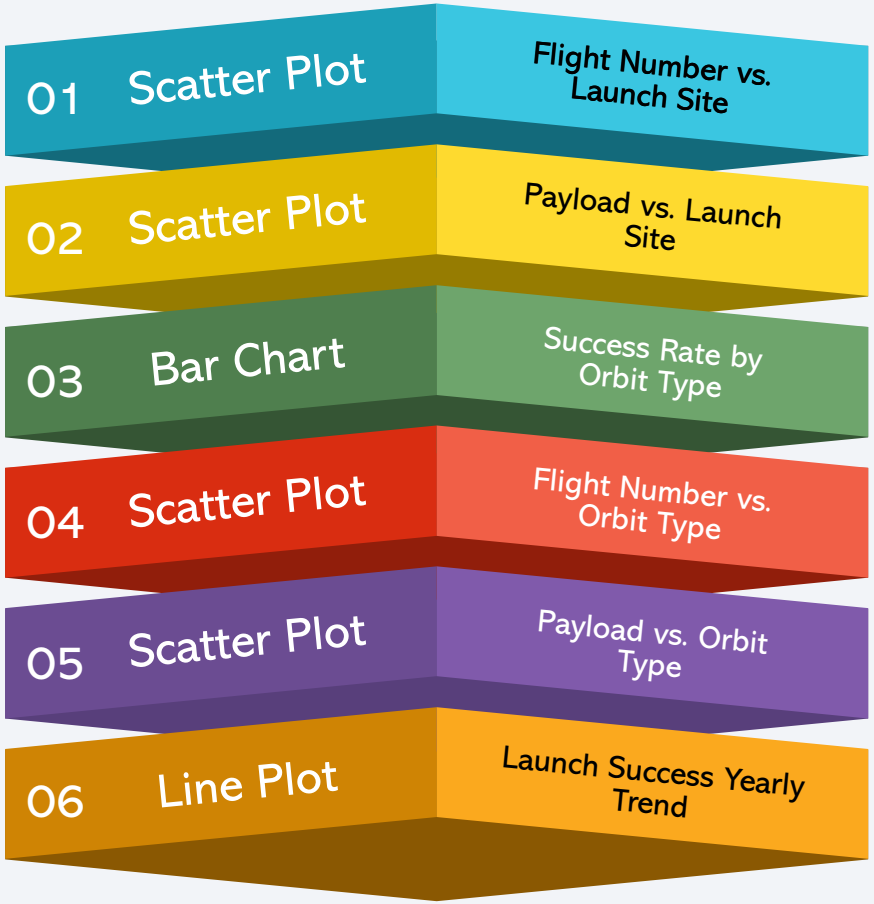
Identify any patterns between the order of launches (Flight Number) and the location from where they took off (Launch Site)

03 - Bar Chart

Visualize the success rate of missions for each specific orbit type (e.g., what percentage of launches achieved their intended orbit for each category)

05 - Scatter Plot

Explores any potential relationships between the cargo weight (Payload) and the type of orbit targeted by the launch (Orbit Type)



02 - Scatter Plot

Explore any potential relationships between the weight of the cargo carried (Payload) and the launch location (Launch Site)

04 - Scatter Plot

Uncover any patterns between the launch order (Flight Number) and the type of orbit the mission aimed to achieve (Orbit Type)

06 - Line Plot

Identify trends in launch success rates over time (yearly). It reveals whether the success rate is improving, declining, or staying consistent across years.

GitHub Link : EDA Data Visualisation

EDA with SQL

Performed SQL queries:

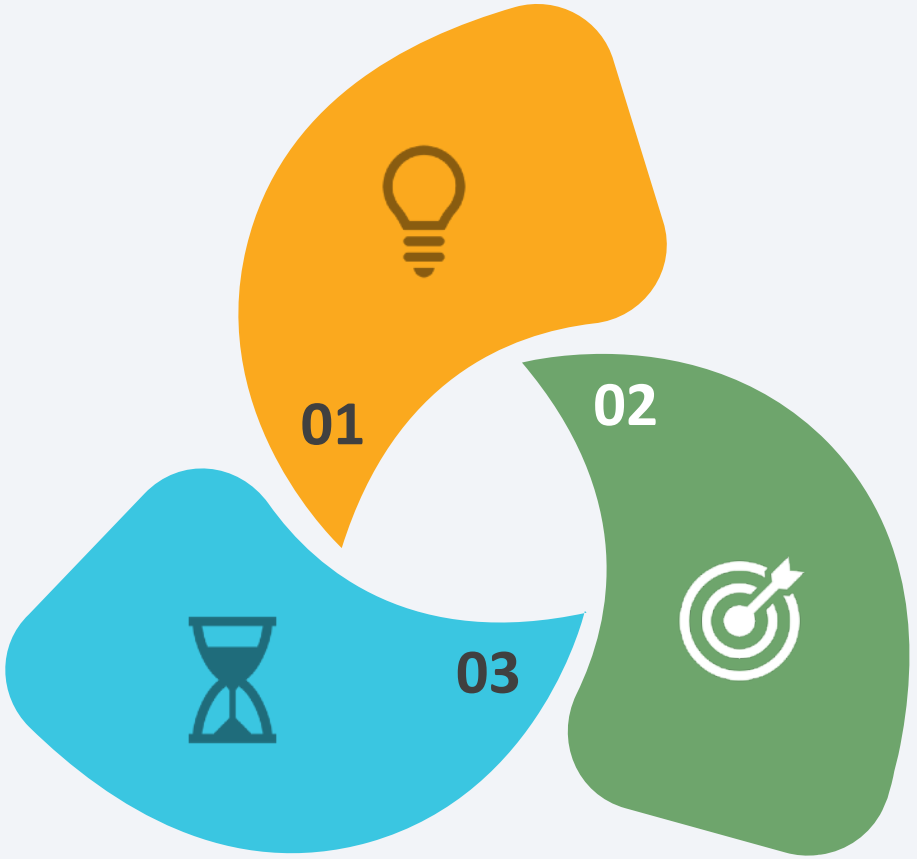
- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the 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
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names 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 : EDA With SQL

Build an Interactive Map with Folium

03 - Distances

Lines connect a specific launch site (e.g., KSC LC-39A) to nearby features like railways, highways, coastlines, and the closest city. These lines illustrate the proximity of launch sites to essential infrastructure and population centers.



01 - Launch Sites

Markers with circles highlight all launch site locations. Text labels provide additional information like "NASA Johnson Space Center." These markers showcase the geographical distribution of launch sites and their proximity to the equator and coastlines.

02 - Launch Outcomes

Green and red markers represent successful and failed launches, respectively. Marker clusters group these markers, allowing viewers to identify launch sites with high success rates at a glance.

GitHub Link : Interactive Map with Folium

Build a Dashboard with Plotly Dash

01. Dropdown List

This list allows users to filter data by a specific launch site. By selecting a site, the dashboard focuses on that location's launch information.

04. Scatter Chart

This scatter plot visualizes the relationship between payload mass (weight) and launch success rate for different booster versions. By filtering the data with the slider and potentially the dropdown, users can explore these correlations for specific launch sites or booster types.

02. Pie Chart

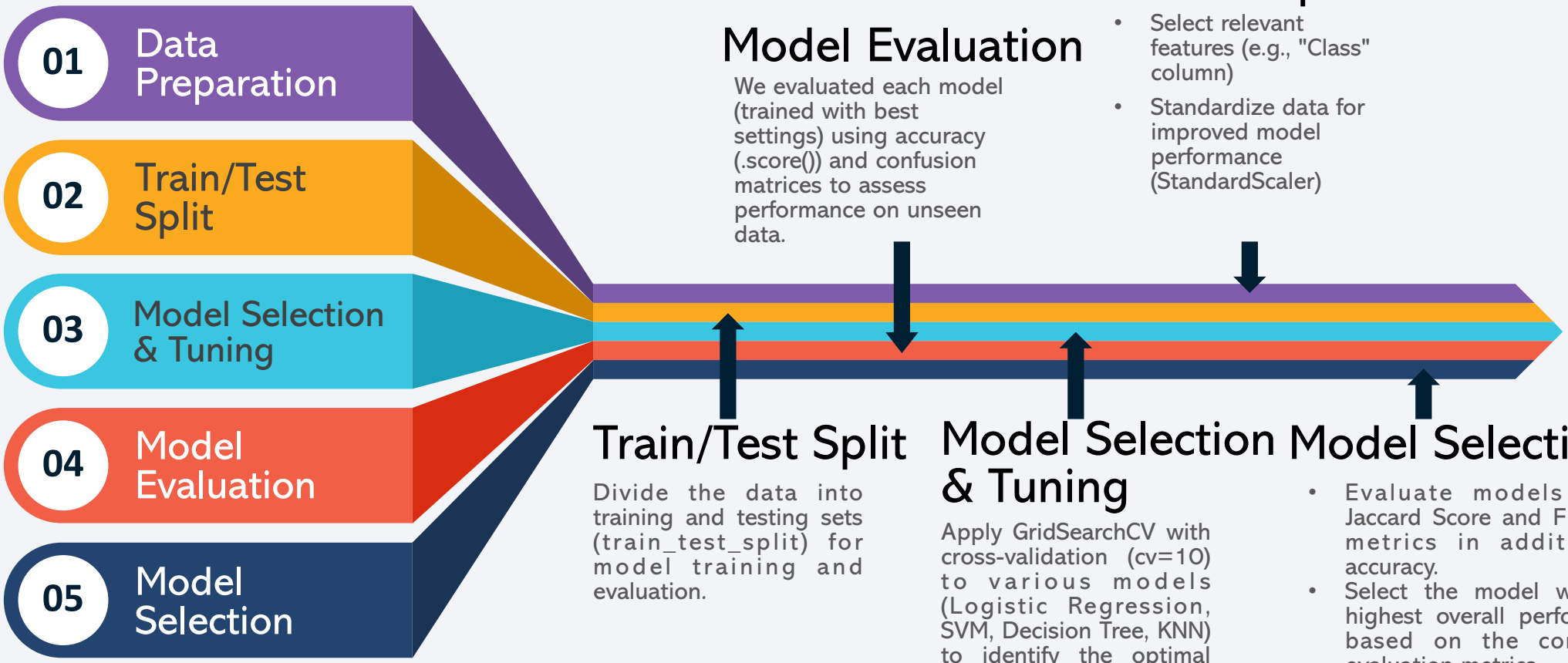
This pie chart displays the total number of successful launches across all sites. When a specific site is chosen from the dropdown, the pie chart dynamically updates to show the success and failure rates for that particular location, providing a focused view.

03. Slider

This slider enables users to select a specific range of payload mass. This allows for focused analysis on how launch success rates correlate with the weight of the cargo carried.

GitHub Link : Dashboard with Plotly Dash

Predictive Analysis (Classification)



Data Preparation

- Select relevant features (e.g., "Class" column)
- Standardize data for improved model performance (StandardScaler)

Model Evaluation

We evaluated each model (trained with best settings) using accuracy (.score()) and confusion matrices to assess performance on unseen data.

Train/Test Split

Divide the data into training and testing sets (train_test_split) for model training and evaluation.

Model Selection & Tuning

Apply GridSearchCV with cross-validation (cv=10) to various models (Logistic Regression, SVM, Decision Tree, KNN) to identify the optimal hyperparameter settings for each model.

Model Selection

- Evaluate models using Jaccard Score and F1 Score metrics in addition to accuracy.
- Select the model with the highest overall performance based on the combined evaluation metrics.

GitHub Link : Predictive Analysis

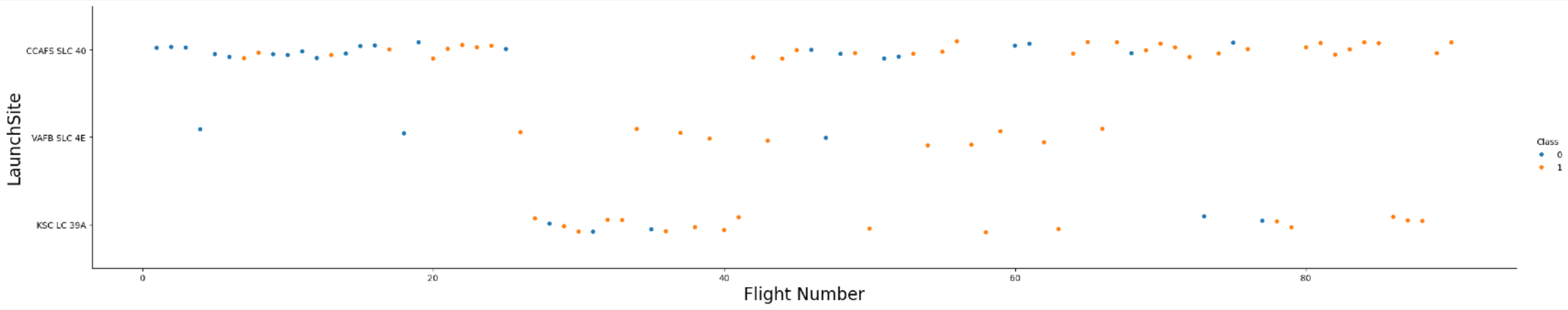
Results

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

Section 2

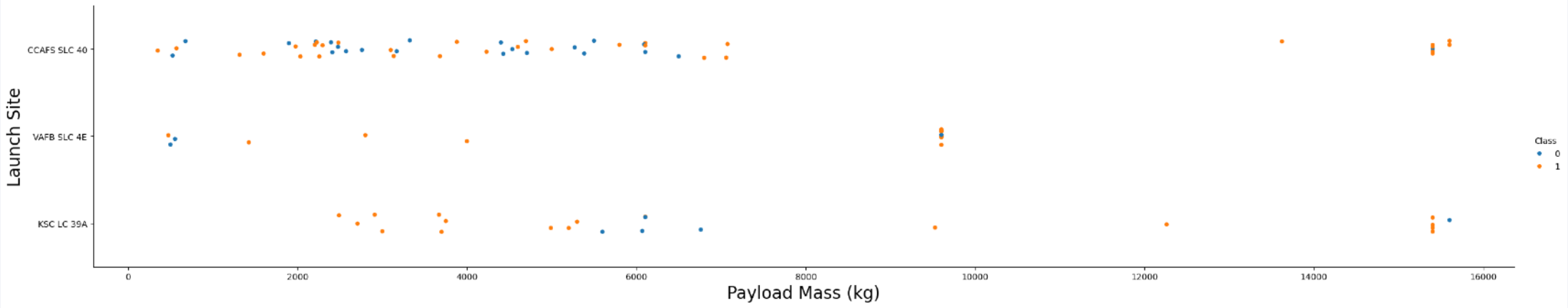
Insights drawn from EDA

Flight Number vs. Launch Site



Early launches struggled, recent ones thrived. CCAFS launches most, but VAFB & KSC have higher success. This suggests SpaceX is learning and improving over time.

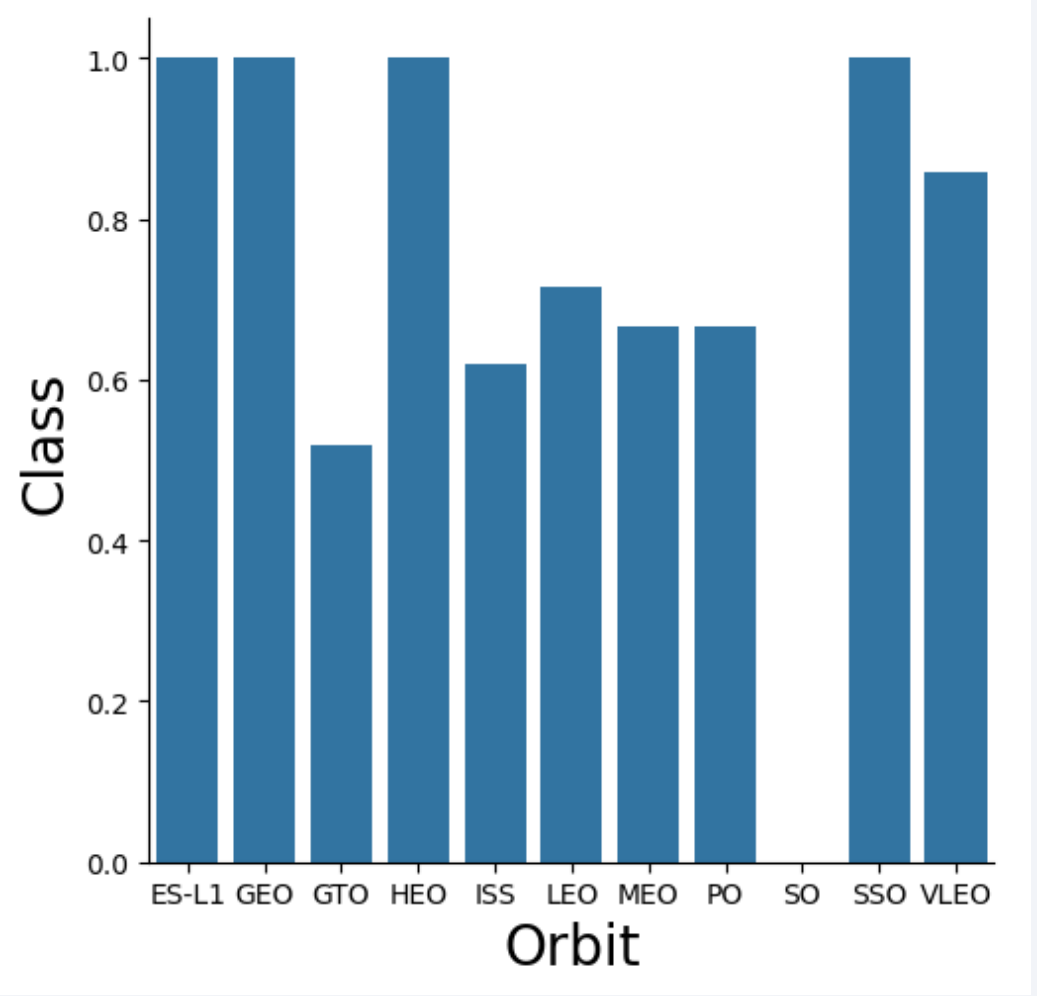
Payload vs. Launch Site



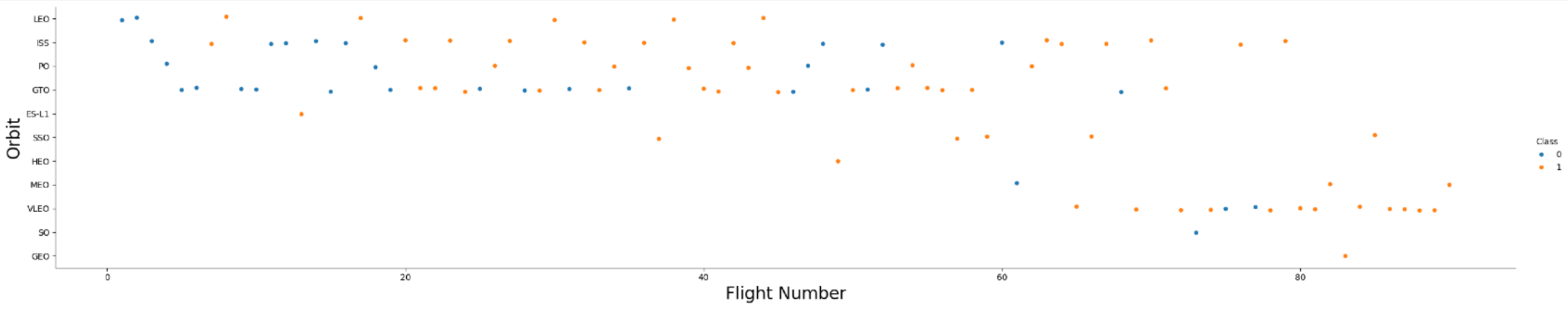
Early launches & lighter payloads struggled, but SpaceX is learning! Heavier payloads see higher success across sites, with KSC excelling for lighter ones.

Success Rate vs. Orbit Type

- According to the bar chart the following Orbits show a 100% success rate which include ES-L1, GEO, HEO and SSO.
- Where orbits GTO, ISS, LEO, MEO, PO and VLEO fall under 50-80% success rate.
- So on the other had showed 0% success rate.

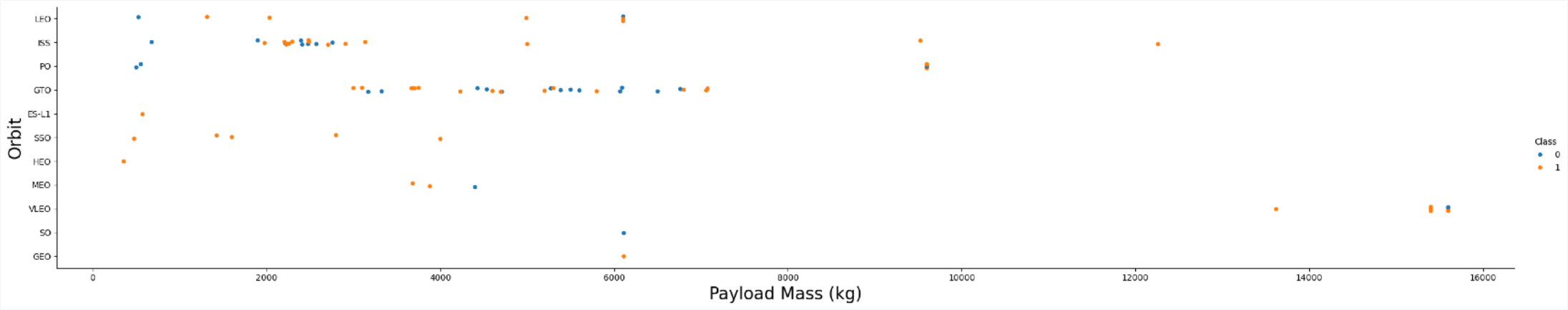


Flight Number vs. Orbit Type



According to the scatter plot 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.

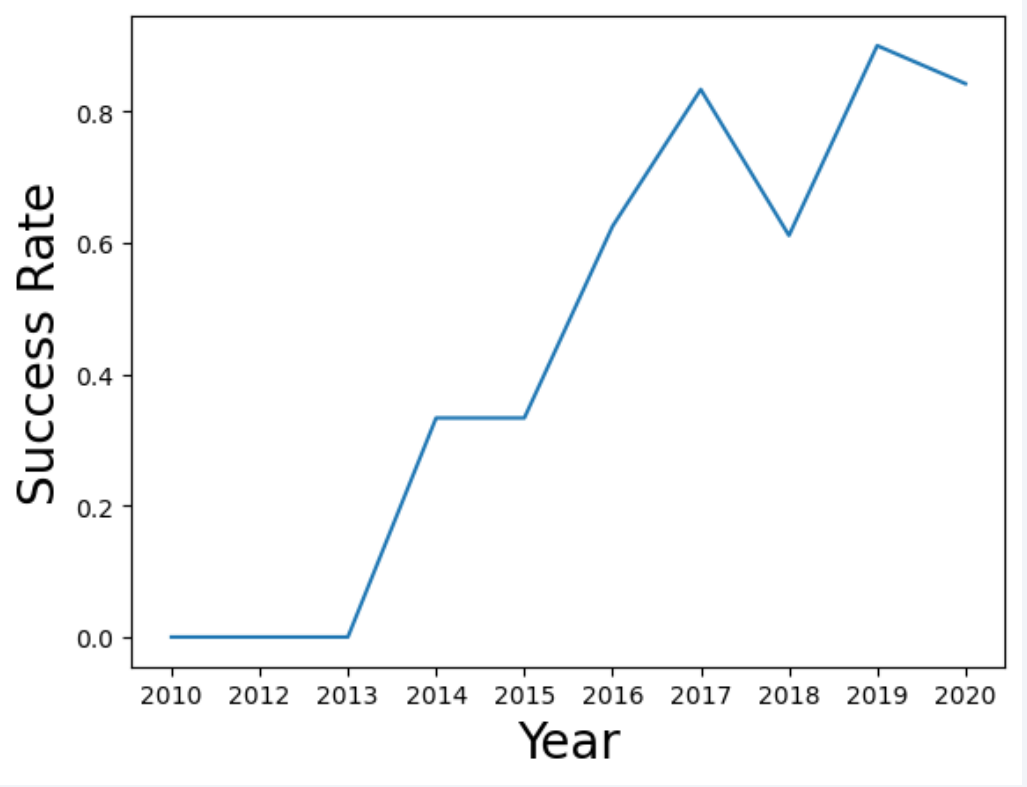
Payload vs. Orbit Type



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

According to the line chart the success rate kept increasing since 2013 to 2020



All Launch Site Names

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

✓ 0.0s Python

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

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

The following query displays the names of all launch sites used which are CCSFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

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

Python

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

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

The following query displays the records of launches done from all sites whose name begin with ‘CCA’

Total Payload Mass

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)';
```

Python

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

total_payload_mass
45596

The following query displays the total payload carried by the NASA (CRS) which is 45596 kg.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS average_payload_mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

Python

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

average_payload_mass
2928.4

The following query displays the average payload carried by the Booster Version F9 v1.1 which is 2928.4 .

First Successful Ground Landing Date

```
%sql SELECT MIN("DATE")AS first_ground_pad_success_date FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

Python

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

first_ground_pad_success_date
2015-12-22

The following query displays the first successful ground landing date which was 22nd December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000;
```

Python

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

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

The following query displays the Booster Versions which were successful in landing with payload between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT "Mission_Outcome", COUNT(*) AS total_number FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

Python

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

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

The following query groups and displays total number of mission outcomes be it success or failure.

Boosters Carried Maximum Payload

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

Python

```
 * 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

The following query gives a list of Booster Versions carrying maximum payload.

2015 Launch Records

```
%sql SELECT SUBSTR(DATE, 6, 2) AS month, DATE , "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE SUBSTR(DATE, 0, 5) = '2015' AND "Landing_Outcome" = 'Failure (drone ship)';
```

Python

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

month	Date	Booster_Version	Launch_Site
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40

The following query gives a list of Booster Versions and launch site for a launches in 2015 where the landing outcome was failure on drone ship.

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

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

The following query ranks landing outcome between specific dates and groups them by landing outcome with their counts.

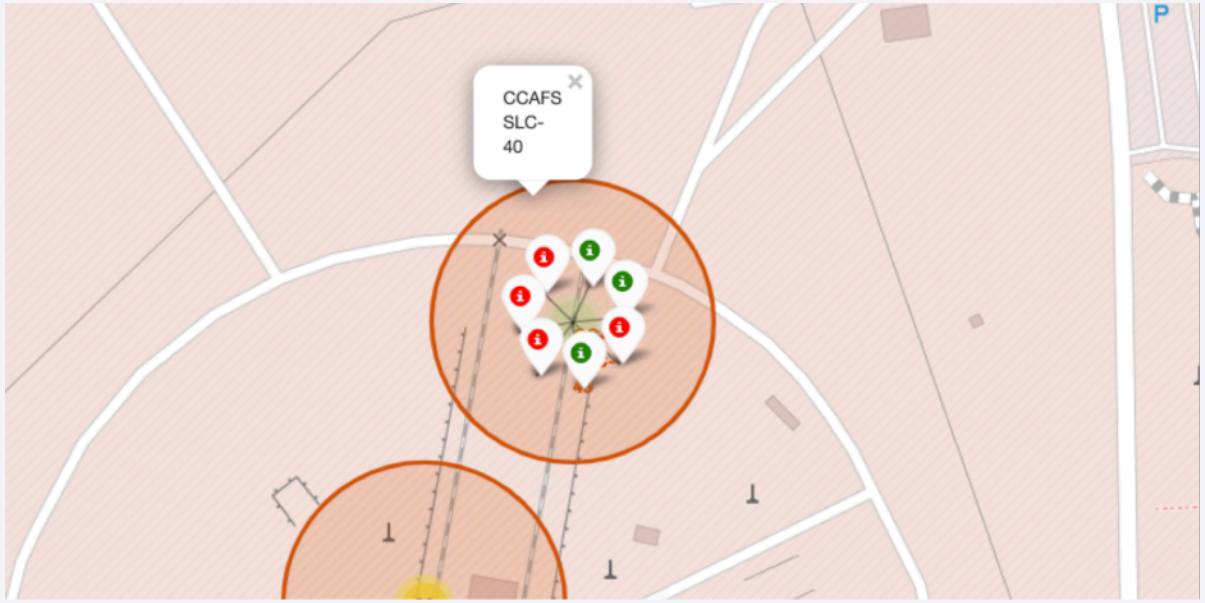
Section 3

Launch Sites Proximities Analysis

All Launch Sites' Location Markers On A Global Map

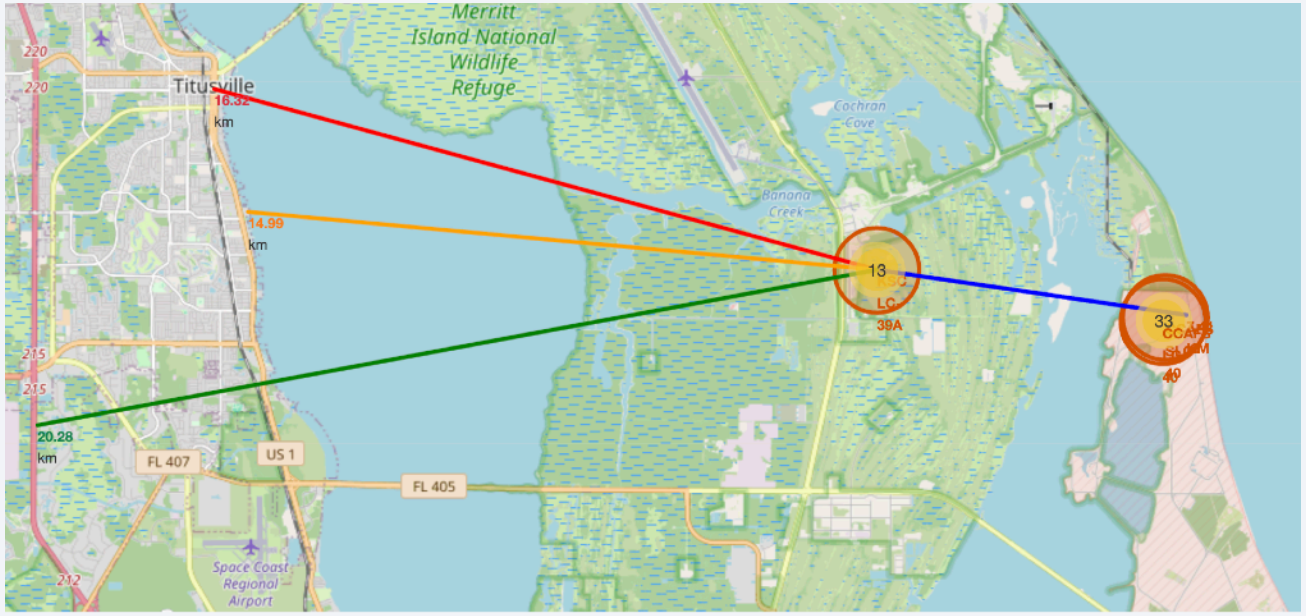
- Near oceans for debris mitigation.
- Equator proximity for Earth's rotational boost.

Color-Labeled Launch Outcomes



From the color-labeled markers in marker clusters, we can easily identify which launch sites have relatively high success rates.

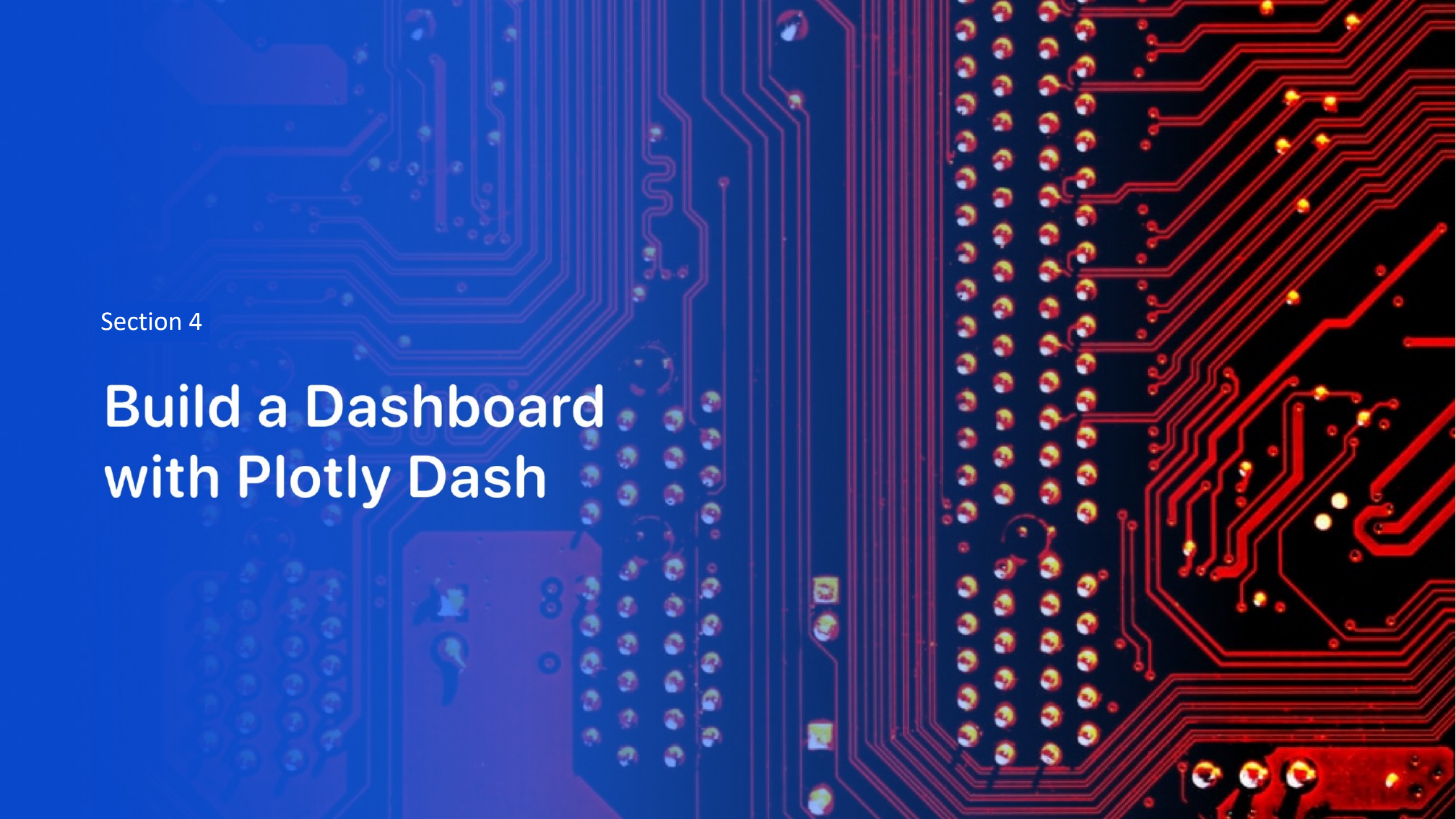
Proximities Distance From KSC LC-39A



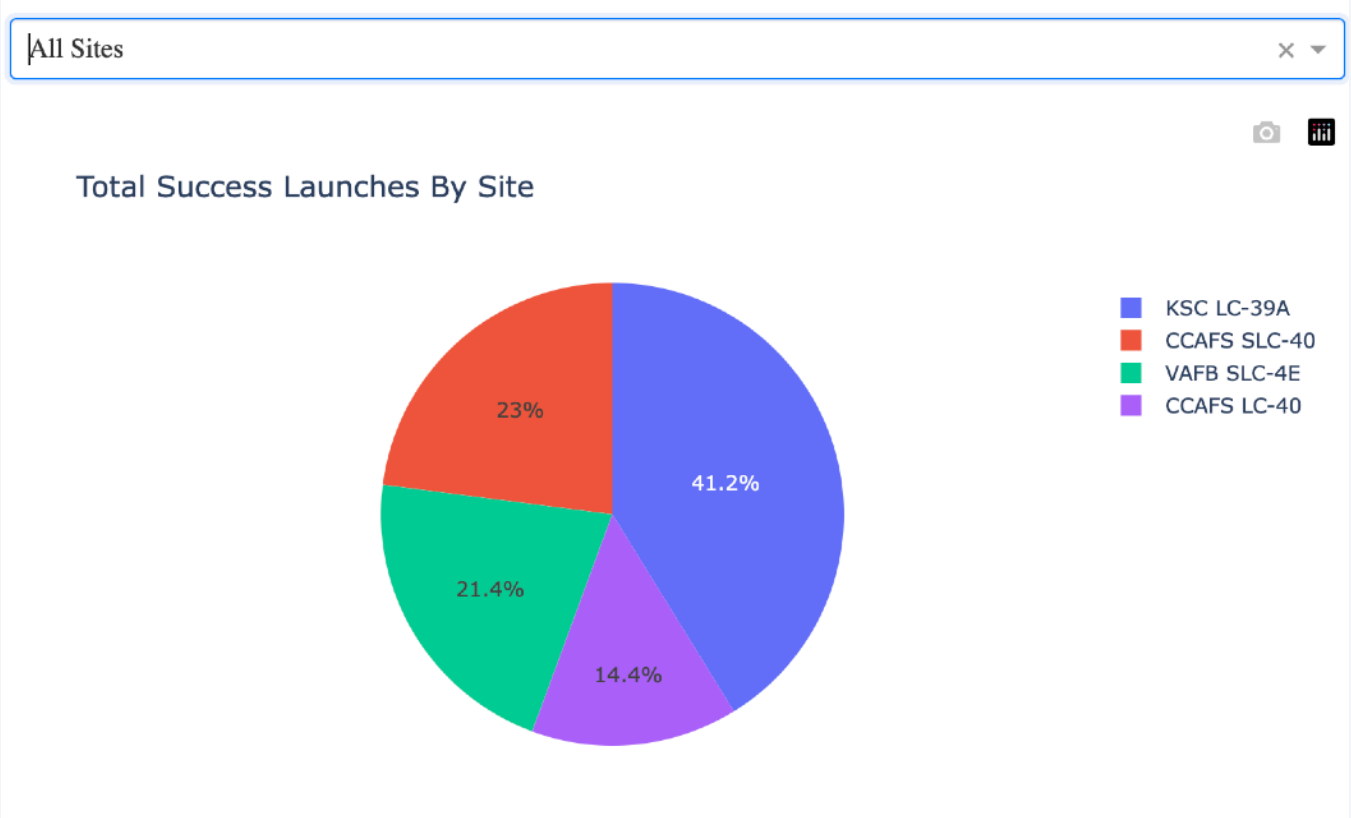
- The nearest railway is about 15 km away from the launch site.
- The nearest highway is about 20 km away from the launch site.
- The nearest coastline is about 6 km away from launch site.

Section 4

Build a Dashboard with Plotly Dash

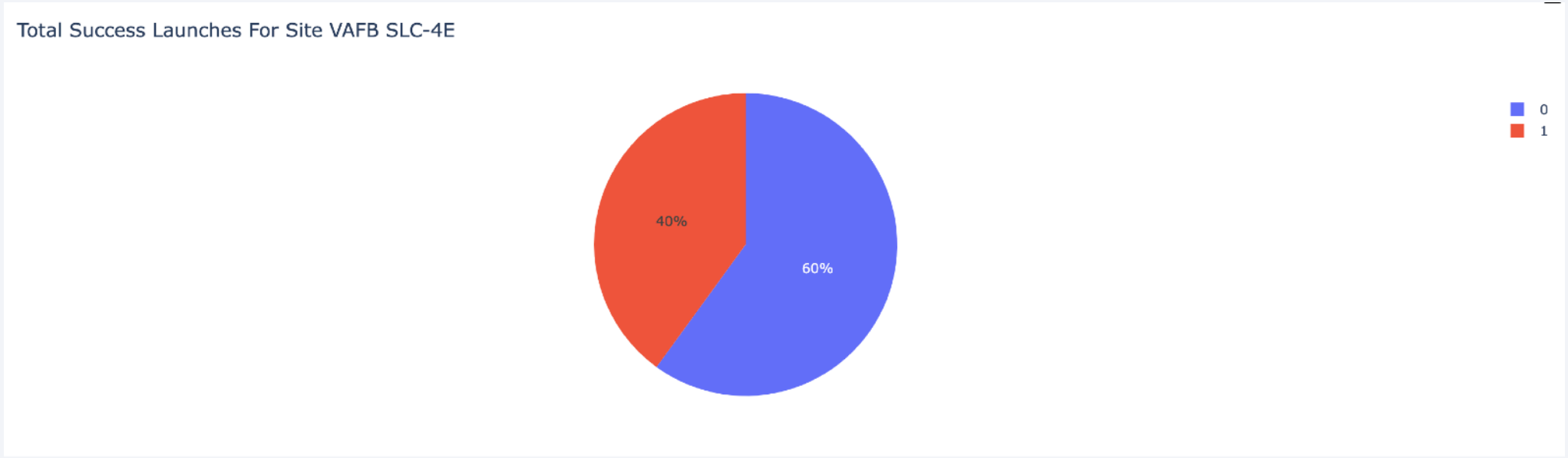


Total Success Launches By Site



According to the pie chart the KSC LC-39A contributes heavily for the most amount of successful launches among other sites for about 41.2% just above CCAFS SLC-40 which contribute around 23% and the rest as VAFB SLC-4E at 21.4% and at last CCAFS LC-40 with the lowest 14.4% successful launches.

Piechart For Launch Site With Highest Launch Success Ratio



The Pie chart represents that the highest launch success ratio is 60 : 40.

Payload vs. Launch Outcome For Different Payloads

The scatter plot represents launch outcome for payload mass between 0 to 10000 kg for various Booster Version Category.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Test Set

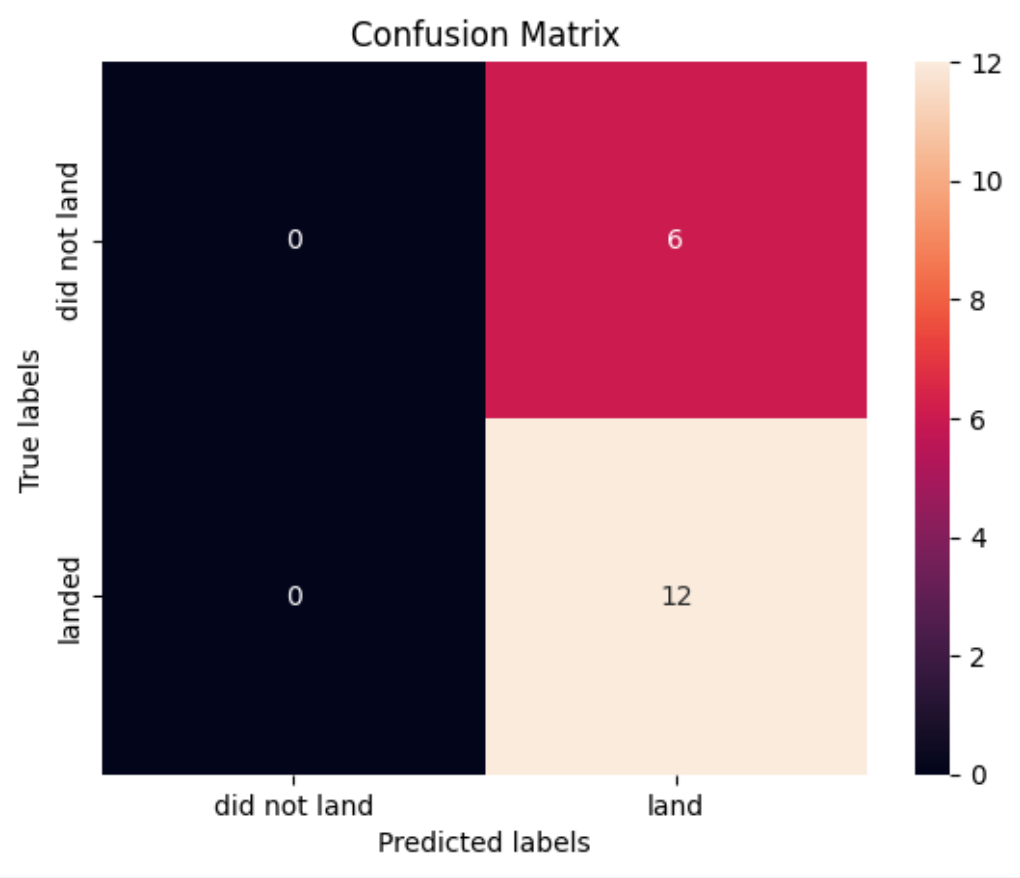
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Entire Data Set

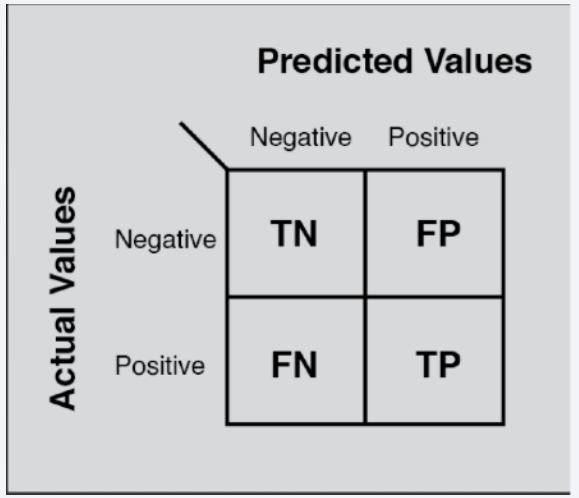
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

While initial tests on a smaller sample size were inconclusive, evaluating all models with the entire dataset revealed the Decision Tree model to be the clear winner. It achieved not only the highest overall scores but also boasted the best accuracy rate.

Confusion Matrix



The Decision Tree model outshines other options based on overall dataset scores and accuracy.



Conclusions

Success Trends:

- Early launches faced challenges, but recent ones show significant improvement, suggesting ongoing learning and refinement.
- Heavier payloads generally correlate with higher success rates across launch sites.
- Certain orbits (ES-L1, GEO, HEO, SSO) achieved a 100% success rate, while others (GTO, ISS, LEO, MEO, PO, VLEO) varied between 50-80%.
- The overall launch success rate has been steadily increasing since 2013.

Strategic Sites & Performance:

- Launch sites near oceans prioritize safety (debris mitigation) and leverage Earth's rotation (equator proximity) for an extra speed boost.
- KSC LC-39A leads in successful launches (41.2%), followed by CCAFS SLC-40 (23%), VAFB SLC-4E (21.4%), and CCAFS SLC-41 (14.4%).

Data Exploration & Modeling:

- Interactive maps helped visualize launch site locations and proximity to infrastructure.
- The Decision Tree model emerged as the most effective for predicting launch success based on comprehensive dataset analysis.

Appendix

All external link to Python code snippets, SQL queries, charts, Notebook outputs, or data sets used in the project is provided here.

Special Thanks to all the Instructors, Coursera and IBM.

Thank you!