IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Summary of methodologies:
  - Data collection -> Data wrangling -> EDA -> Predictive Analysis
- Summary of all results
  - Identified key success factors
  - Found correlations between payload and success
  - Built a predictive classification model with high accuracy

Introduction

- This project, developed as part of the Data Science Capstone Course, involved working as a Data Scientist for a private space launch company.
- The goal was to analyse historical SpaceX launch data to identify patterns, uncover insights, and develop predictive models to enhance the company's decision-making and launch success rate.

Section 1

Methodology

Methodology

Executive Summary

- **Data collection:** Collected SpaceX launch data using REST API and web scraping, converted JSON to DataFrame, and cleaned data by handling NULL values and filtering Falcon 1 launches.
- **Data Wrangling:** Processed key attributes and converted landing outcomes into binary classes (0 = Fail, 1 = Success)..
- **Exploratory Data Analysis**: Analyzed success rates by site and payload mass, identified correlations, and applied one-hot encoding.
- **Data Visualization:** Built an interactive dashboard with Folium and Plotly Dash to visualize launch patterns and success factors.Perform predictive analysis using classification models
- **Predictive Modeling:** Trained ML models (Logistic Regression, SVM, Decision Tree, KNN), optimized hyperparameters with Grid Search, and evaluated with a confusion matrix

Data Collection

- Gathered SpaceX launch data using SpaceX REST API and web scraping (BeautifulSoup).
- Converted JSON data to a DataFrame using json_normalize.
- Cleaned data by filtering out Falcon 1 launches and handling NULL values using the mean.

Data Collection – SpaceX API

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

- Data Collection - SpaceX API

Place your flowchart of SpaceX API calls here

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- Data Collection - Scrapping

Place your flowchart of web scraping here

Data Wrangling

- Processed key attributes: Flight Number, Date, Booster Version, Payload Mass, Launch Site, and Outcome.
- Converted categorical landing outcomes to binary classes (0 = Fail, 1 = Success)
- Data Wrangling

EDA with Data Visualization

- Scatter Plot: To show the relationship between two numerical variables.
- Bar Chart: To compare categories and highlight differences.
- Line Plot: To display trends over time or continuous data.
- EDA with Data Visualization

EDA with SQL

- Unique Launch Sites
- Launch Sites Starting with 'CCA'Total Payload Mass (NASA)
- Average Payload Mass (F9 v1.1)
- First Successful Ground Pad Landing
- Successful Drone Ship Landings (4000-6000 kg)
- Mission Outcome Count
- Maximum Payload Mass
- Failed Drone Ship Landings (2015)
- Rank Landing Outcomes
- [EDA with SQL]

Build an Interactive Map with Folium

- Marker: Added to show the location of launche sites, the closest coastline, and the closest city with the distance labelled.
- Line: Added to visually connect the launch site to the closest city and the closest coastline, highlighting the proximity.
- Interactive Map with Folium

Build a Dashboard with Plotly Dash

- Dropdown (Launch Site): Added to allow users to select a specific launch site or view data for all sites. Helps identify success rates across different sites.
- Pie Chart: Added to visualize the proportion of successful versus failed launches for the selected site or all sites. Helps analyze the success rate of each site.
- Range Slider (Payload): Added to filter data by payload range. Helps explore how payload mass affects launch success.
- Scatter Plot: Added to display the relationship between payload mass and launch outcome, with color labels for different booster versions. Helps identify patterns and correlations between payload, booster version, and success rates.
- Dashboard with Plotly Dash

Predictive Analysis (Classification)

- [Predictive Analysis]

Results

- Launch site with the highest success rate: KSC LC-39A
- Payload mass range with highest success rate: Between 2,000 kg and 5,000 kg
- Booster version with highest success rate: Falcon 9 Block 5
- Predictive model accuracy: 85%

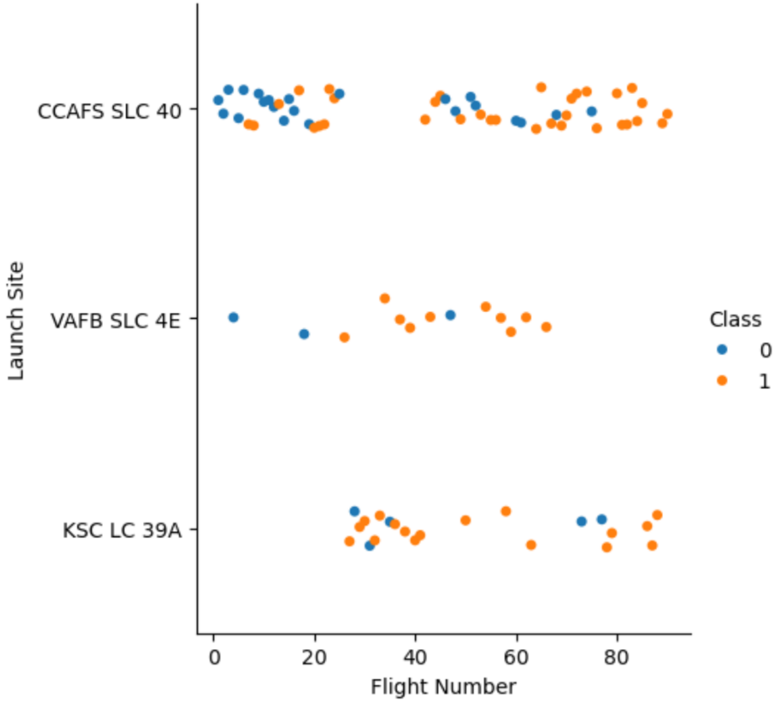
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Launches with higher flight numbers are more likely to succeed. This presents a consistent improvement throughout SpaceX's developing process

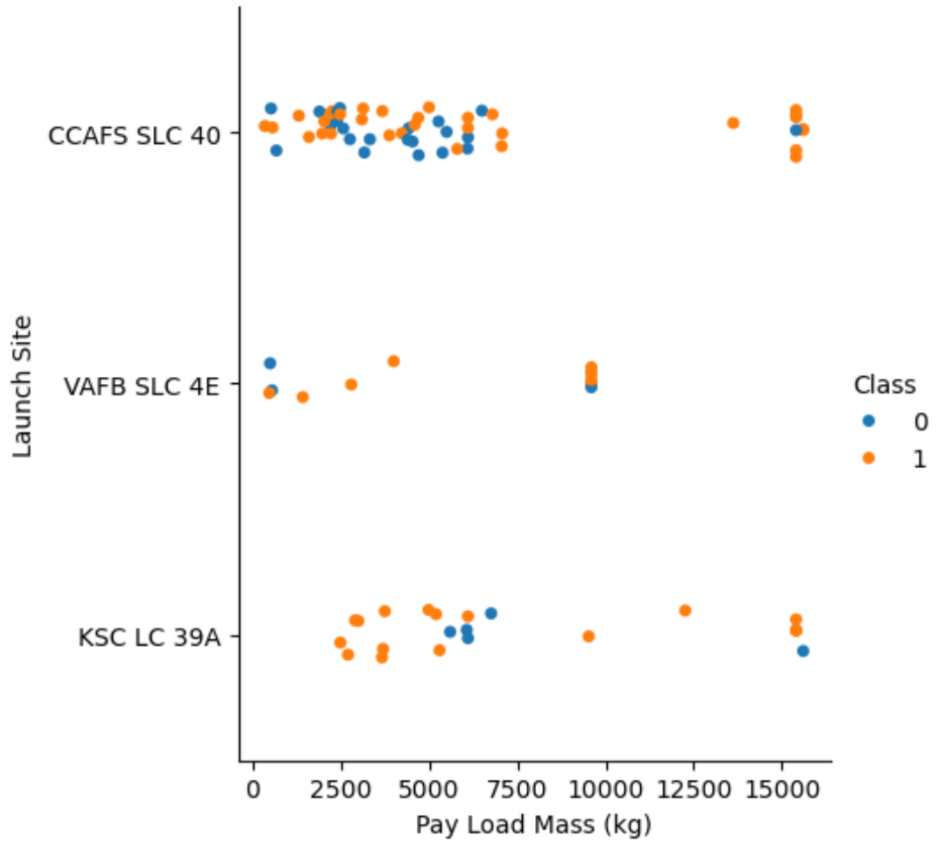
```
In [6]: # Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site,
sns.catplot(x="FlightNumber", y ="LaunchSite", hue="Class", data=df)
plt.xlabel('Flight Number')
plt.ylabel('Launch Site')
plt.show()
```



Payload vs. Launch Site

- There is no rocket launched for payload mass greater than 10000kg
- There is no observable correlation between payload, launch site and successful rate

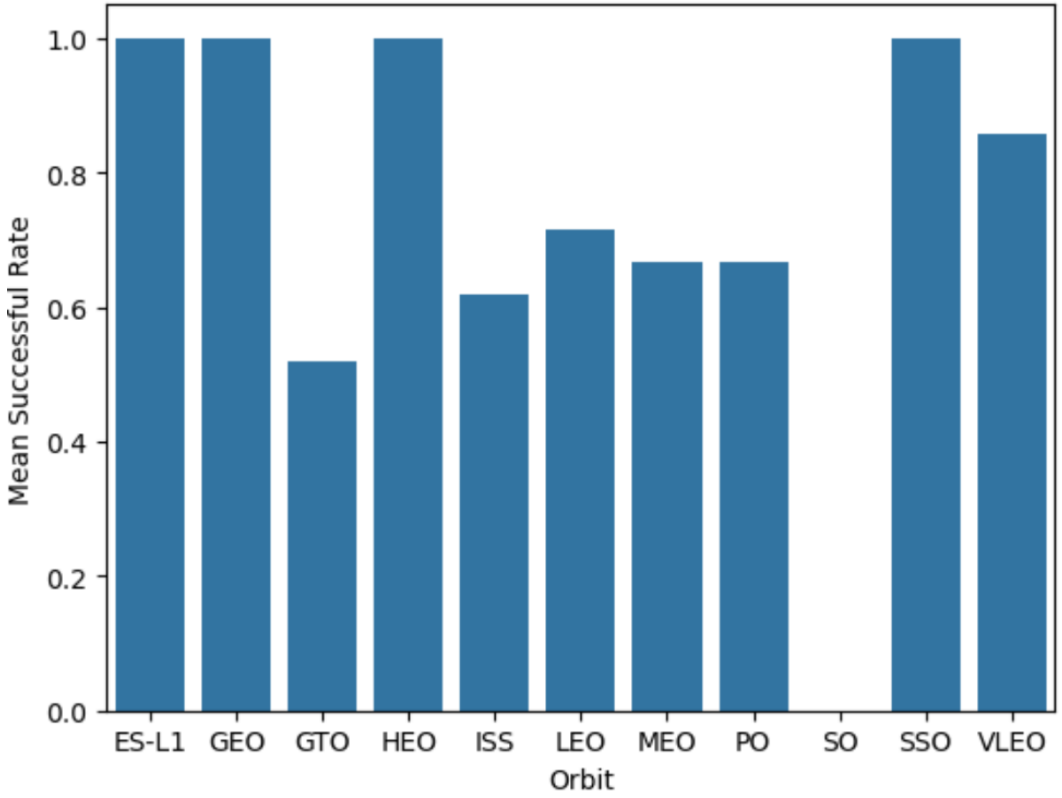
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the L
sns.catplot(x='PayloadMass', y='LaunchSite', data=df, hue='Class')
plt.xlabel('Pay Load Mass (kg)')
plt.ylabel('Launch Site')
plt.show()
```



Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO have the highest success rates (between 95%-100%)
- GTO, ISS, LEO, MEO, and PO have lower success rates falling between 50%-70%
- SO has the lowest success rate at 0%

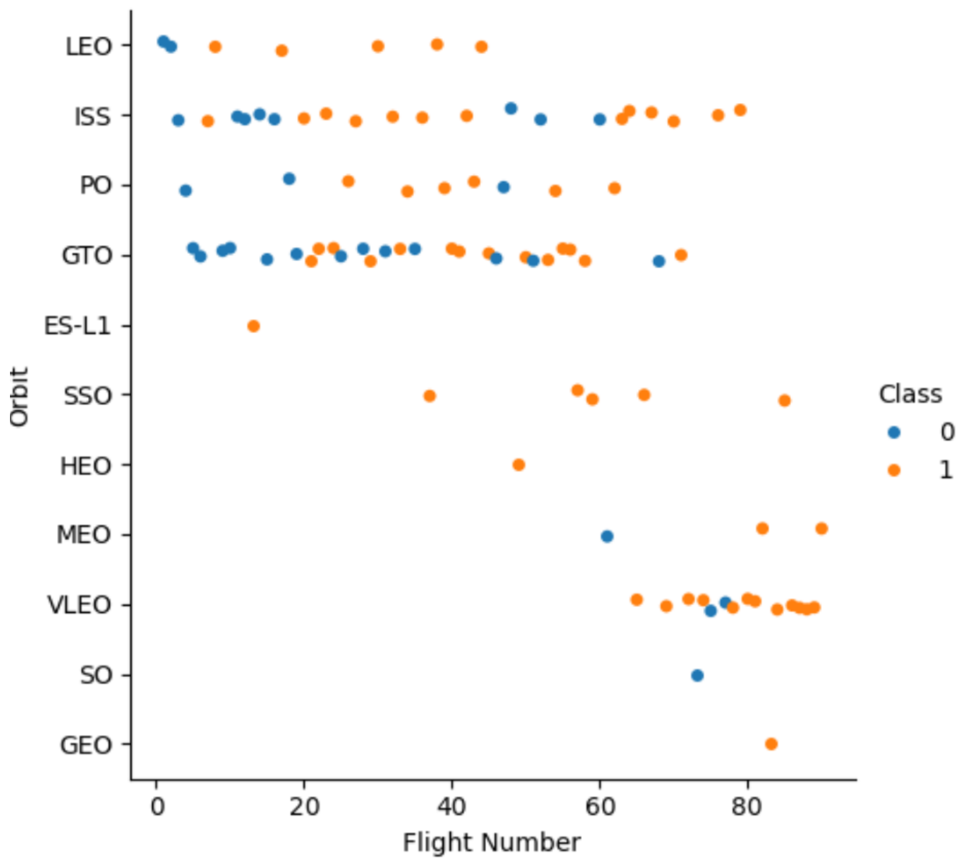
```
# HINT use groupby method on Orbit column and get the mean of Class column
orbit_success_rates = df.groupby('Orbit')['Class'].mean()

sns.barplot(x=orbit_success_rates.index, y = orbit_success_rates.values)
plt.xlabel('Orbit')
plt.ylabel('Mean Successful Rate')
plt.show()
```



Flight Number vs. Orbit Type

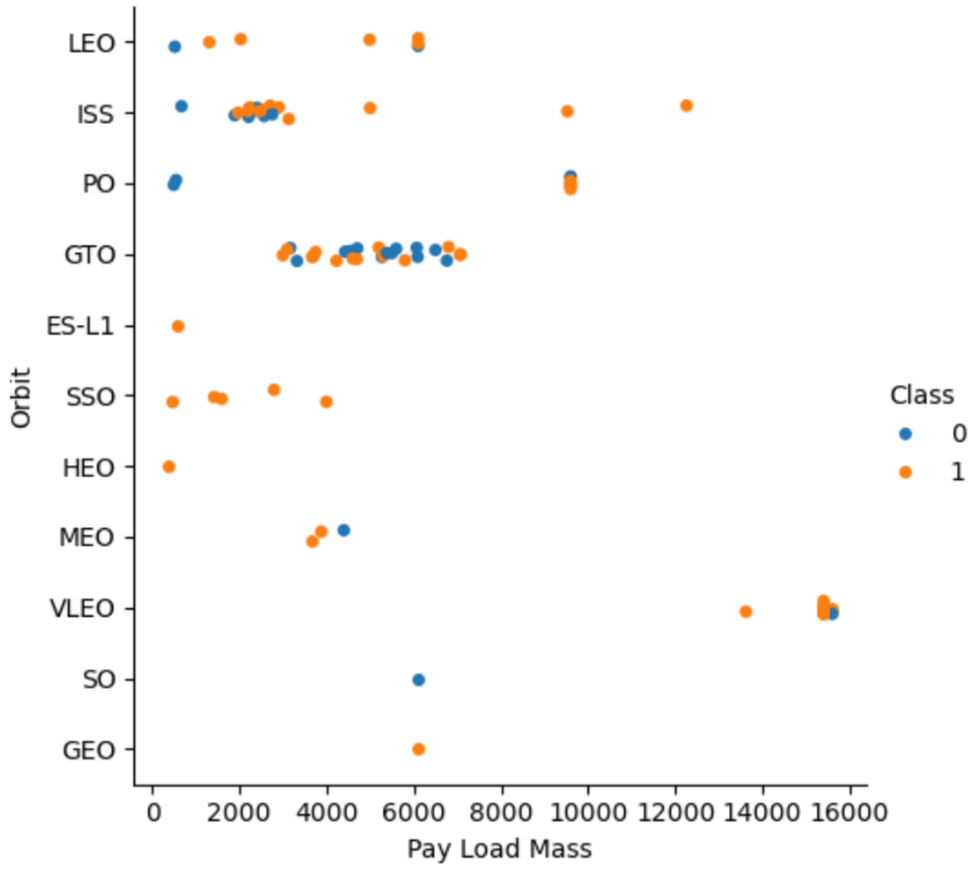
- It is observable that in the LEO orbit, success seems to be related to the number of flights.
- Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and
sns.catplot(x='FlightNumber', y='Orbit', data=df, hue='Class')
plt.xlabel('Flight Number')
plt.ylabel('Orbit')
plt.show()
```



Payload vs. Orbit Type

```
# Plot a scatter point chart with x axis to be Payload Mass and y axis to be the Orbit, a
sns.catplot(x='PayloadMass', y='Orbit', data=df, hue='Class')
plt.xlabel('Pay Load Mass')
plt.ylabel('Orbit')
plt.show()
```

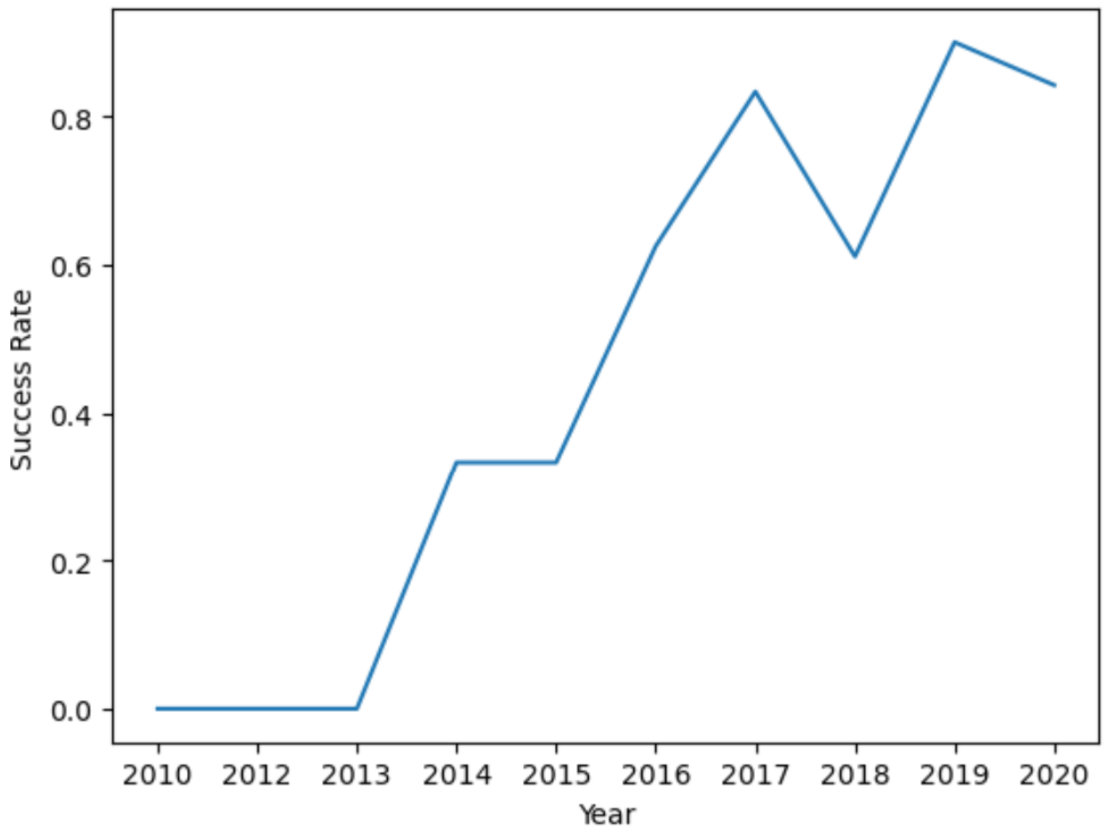


- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- Overall, the success rate increase over the period between 2010 and 2020

```
# Plot a line chart with x axis to be the extracted year and y axis to be the succ
yearly_success_rates = df.groupby('Date')['Class'].mean()
sns.lineplot(x=yearly_success_rates.index, y=yearly_success_rates.values)
plt.xlabel('Year')
plt.ylabel('Success Rate')
plt.show()
```



All Launch Site Names

- All launch site names are:

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

- Present your query result with a short explanation here

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

Launch Site Names Begin with 'CCA'

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer == 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_version LIKE 'F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome =='Success (ground pad)'
```

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome == 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 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

```
%sql SELECT Mission_Outcome, COUNT(Date) FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

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

```
%sql SELECT substr(Date, 2,6), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE

WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date, 0,5) ='2015'
```

substr(Date, 2,6)	Landing_Outcome	Booster_Version	Launch_Site
015-01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
015-04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- %sql SELECT Landing_Outcome, COUNT(Date) FROM SPACEXTABLE GROUP BY Landing_Outcome ORDER BY COUNT(DATE) DESC

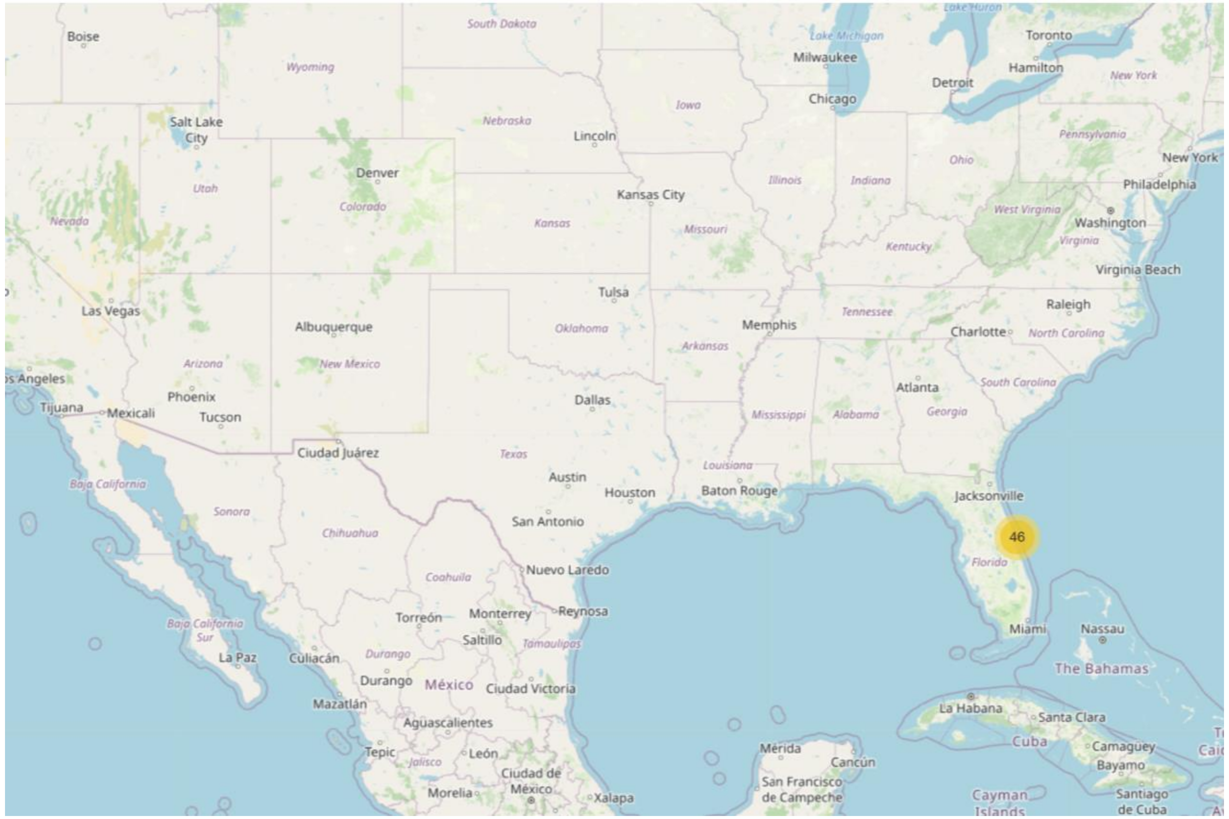
Landing_Outcome	COUNT(Date)
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

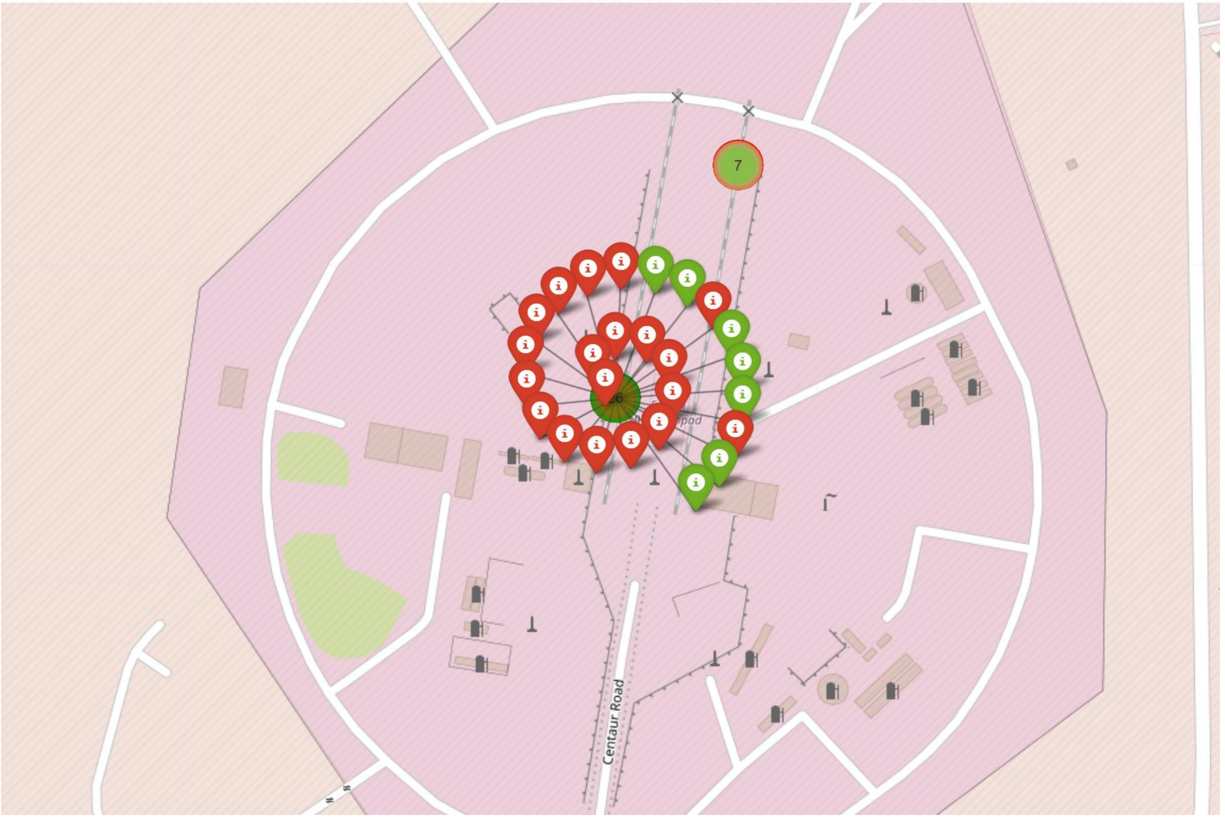
Section 3

Launch Sites Proximities Analysis

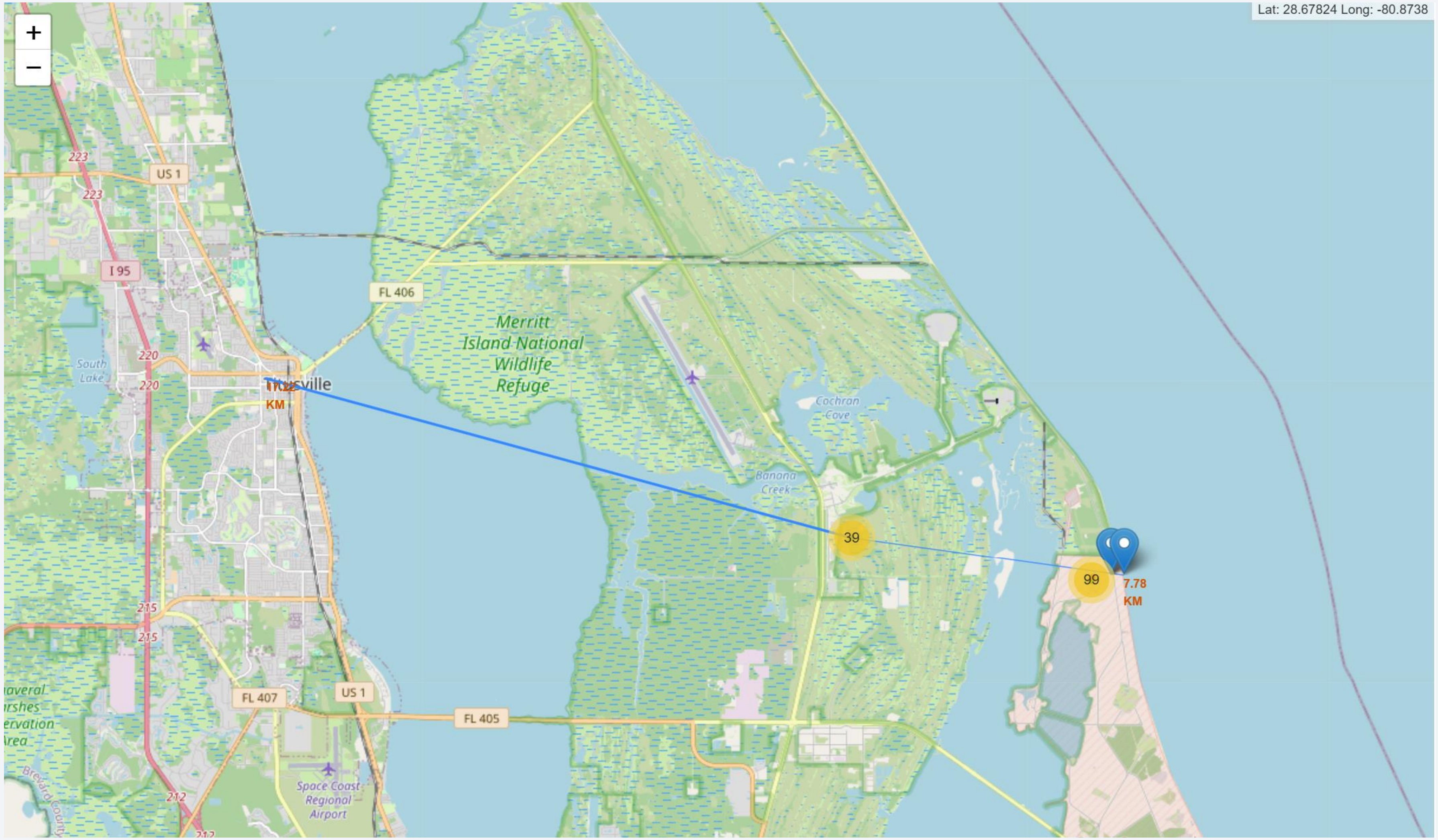
All Launch Sites

Launch Site Clusters





Distances

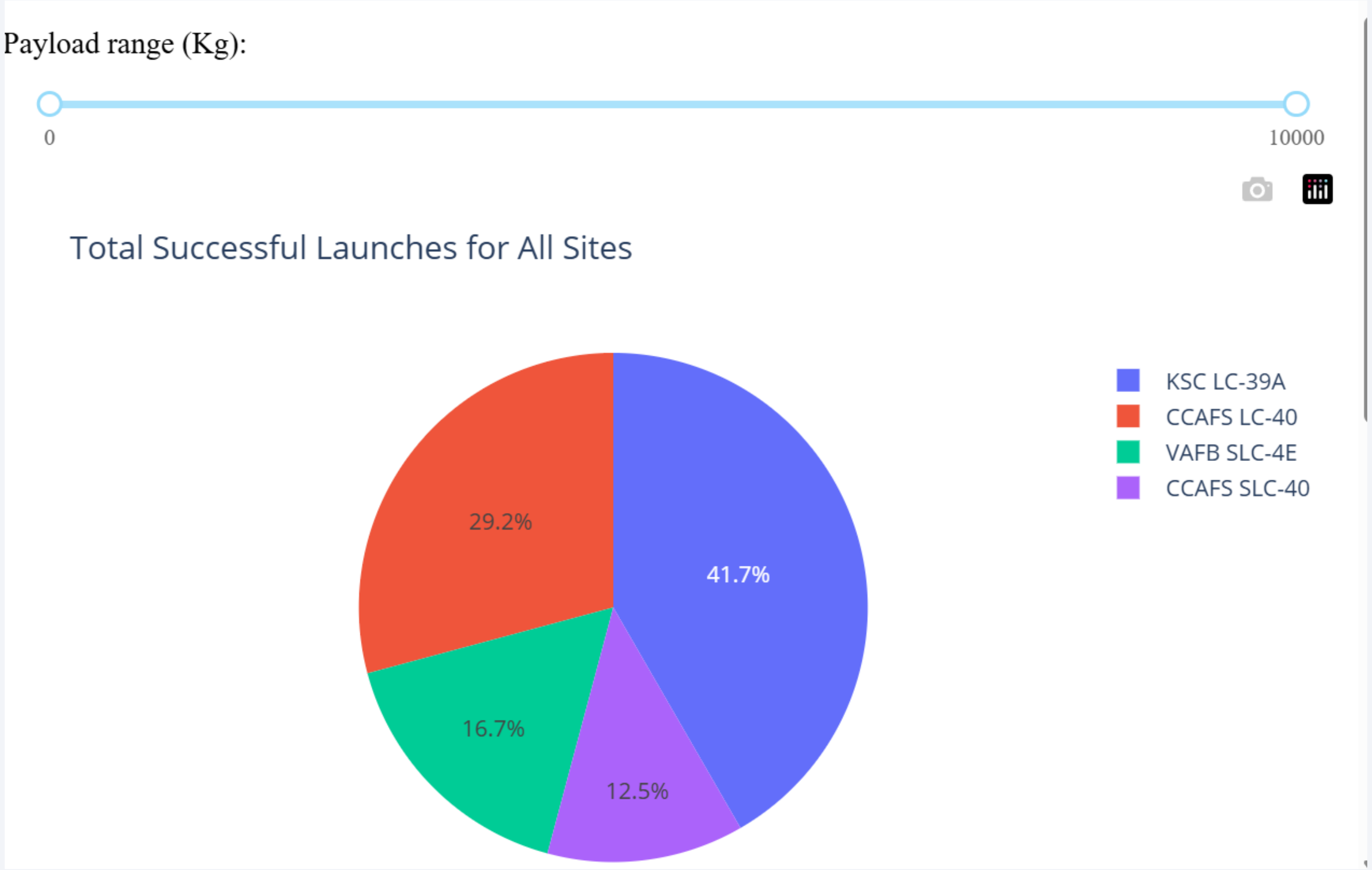


Lat: 28.67824 Long: -80.8738
Merritt Island National Wildlife Refuge
Titusville
KM
US 1
I 95
FL 406
FL 407
FL 405
Cochran Cove
Banana Creek
South Lake
Space Coast Regional Airport
39
99
7.78 KM

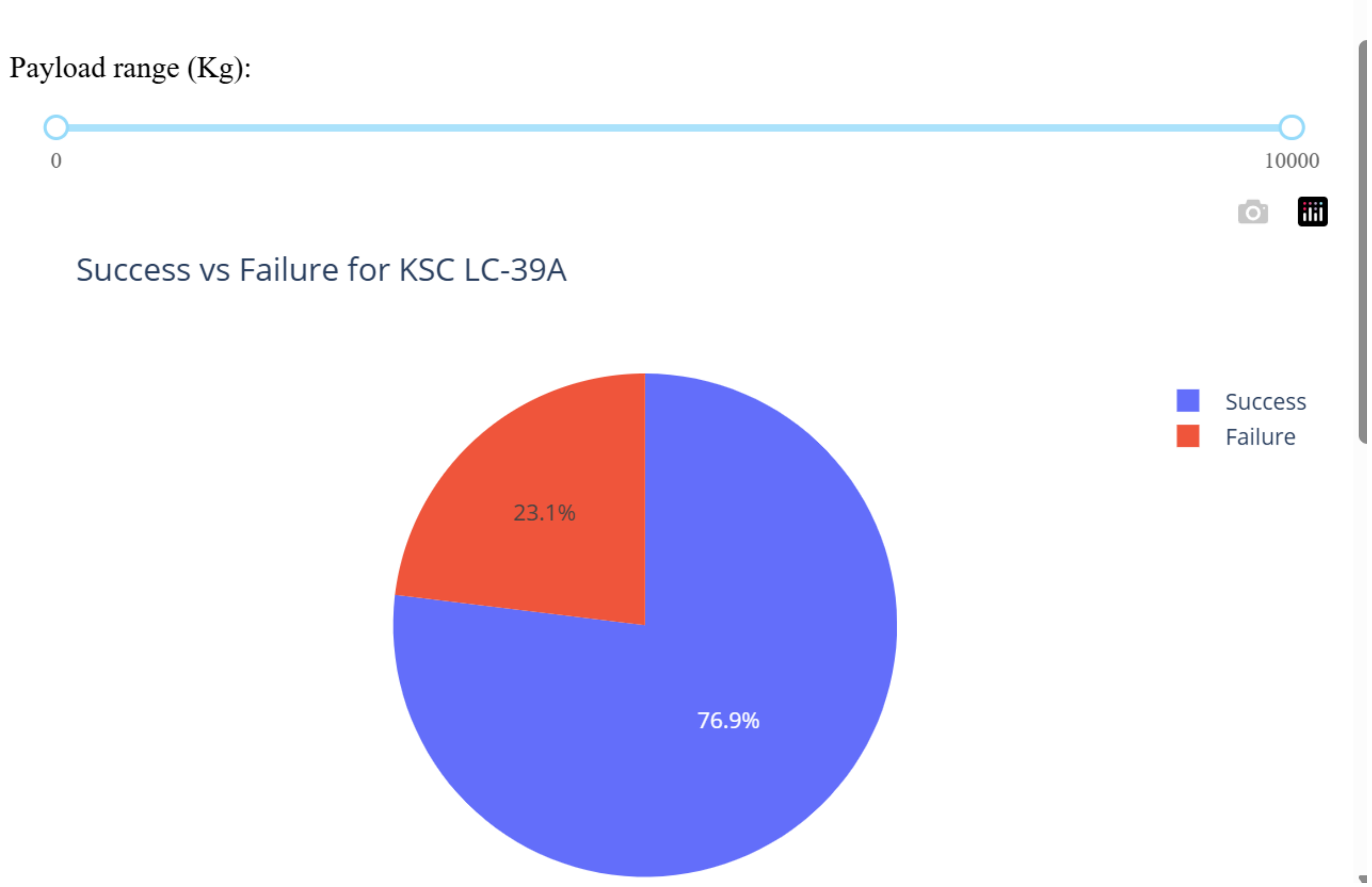
Section 4

Build a Dashboard with Plotly Dash

All Site Successful Launches



Highest Launch Success Ratio



Payload vs. Launch Outcome (Payload between 0 and 7000)

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

- Find which model has the highest classification accuracy

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation

Conclusions

In this project, I have

- Performed exploratory data analysis (EDA) to identify patterns and trends in SpaceX launch data.
- Applied feature engineering using one-hot encoding to prepare data for machine learning.
- Developed an interactive Plotly Dash dashboard for real-time analysis of launch success rates by site and payload range.
- Built and tuned predictive classification models using GridSearchCV and identify the best performer.
- Evaluated the model using accuracy scores and a confusion matrix to identify the best-performing configuration.
- Provided actionable insights to improve future SpaceX launch success rates.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!