Space Race with Data Science

IBM Developer
SKILLS NETWORK

OUTLINE

- Executive summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Data collection and data wrangling methodology related
- EDA and interactive visual analytics methodology related
- Predictive analysis methodology related
- EDA with visualization result
- EDA with SQL result
- Interactive map with Folium result
- Plotly Dash dashboard result
- Predictive analysis (classification) results
- Conclusion
- Innovative insights

INTRODUCTION

- Project Background
  - In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. This course allows learners to apply data science techniques to real-world problems, providing a portfolio-worthy project that demonstrates their ability to solve complex data challenges.

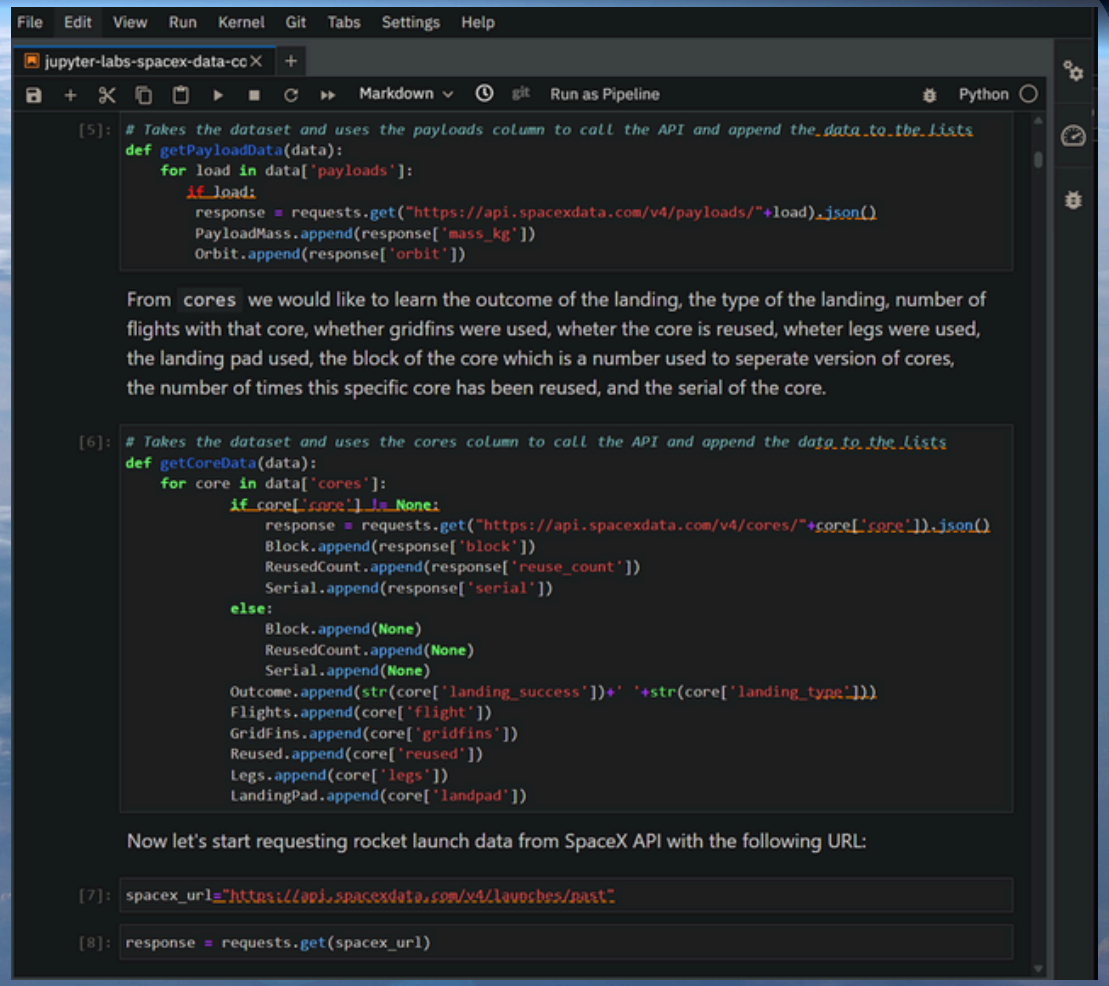
METHODOLOGY

DATA COLLECTION AND DATA WRANGLING METHODOLOGY

- Data Collection
  - Identify the dataset and related data.
  - Requesting API from SpaceX.
  - Ensuring the data format, structure, and data collected are accurate.
-
- Data Wrangling
  - Exploratory Data Analysis
  - Determine Training Labels

DATA COLLECTION AND DATA WRANGLING METHODOLOGY

- Data Collection
  - Identify the dataset and related data.
  - Requesting API from SpaceX.



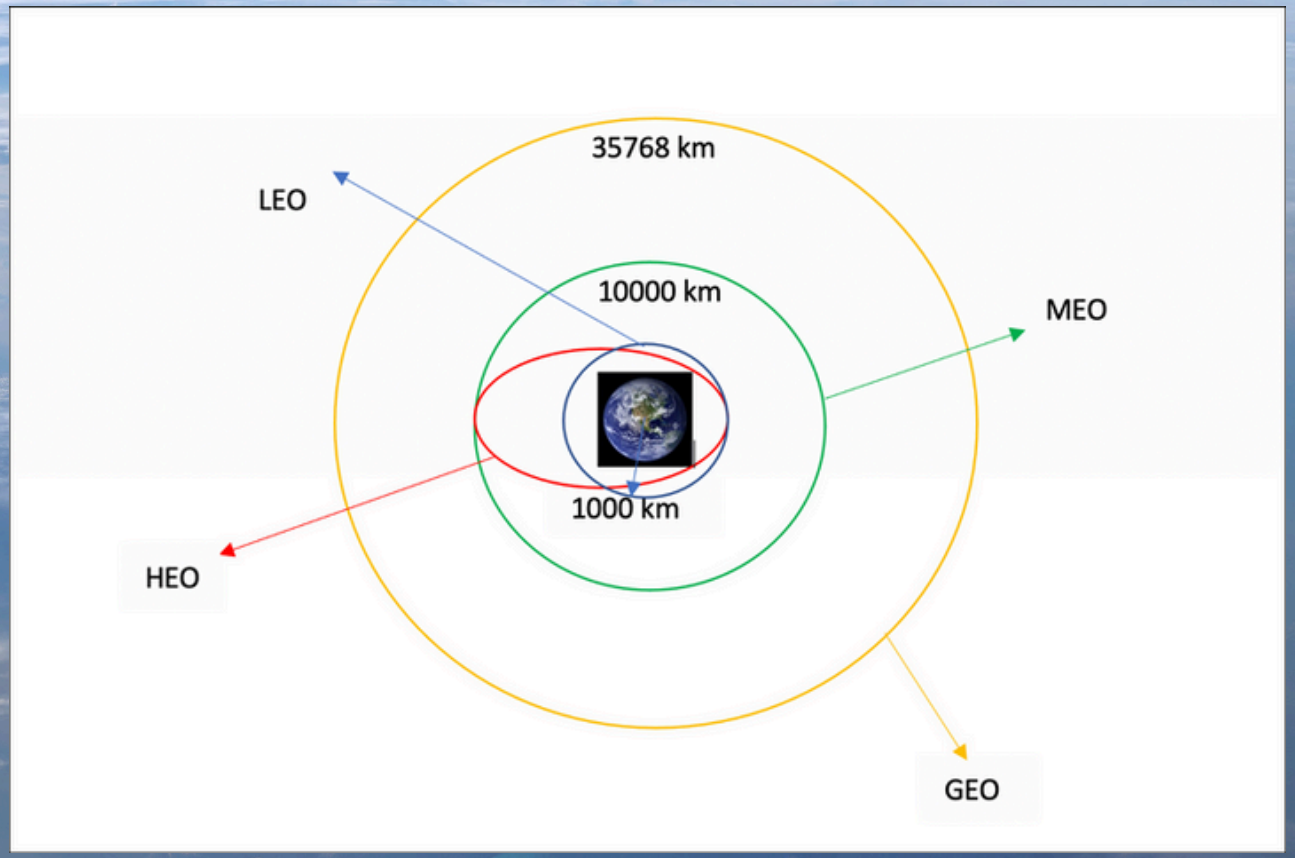
DATA COLLECTION AND DATA WRANGLING METHODOLOGY

- Data Collection
  - Ensuring the data format, structure, and data collected are accurate.

```
# Lets take a subset of our dataframe keeping only the features we want and
the flight number, and date_utc.
data = data[['rocket', 'payloads', 'launchpad', 'cores', 'flight_number',
'date_utc']]

# We will remove rows with multiple cores because those are falcon rockets
with 2 extra rocket boosters and rows that have multiple payloads in a single
rocket.
data = data[data['cores'].map(len)==1]
data = data[data['payloads'].map(len)==1]

# Since payloads and cores are lists of size 1 we will also extract the single
value in the list and replace the feature.
data['cores'] = data['cores'].map(lambda x : x[0])
data['payloads'] = data['payloads'].map(lambda x : x[0])

# We also want to convert the date_utc to a datetime datatype and then
extracting the date leaving the time
data['date'] = pd.to_datetime(data['date_utc']).dt.date

# Using the date we will restrict the dates of the launches
data = data[data['date'] <= datetime.date(2020, 11, 13)]
```

DATA COLLECTION AND DATA WRANGLING METHODOLOGY

- Data Wrangling
  - Exploratory Data Analysis
  - Determine Training Labels



EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY

Exploratory Data Analysis is the first step of any data science project. In the first lab, we will perform some Exploratory Data Analysis using a database. In the second lab, we will see if the data can be used to automatically determine if the Falcon 9's first stage will land.

- Predictive Analysis Methodology
- EDA with Visualization
- Complete EDA with SQL

EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY

- **Predictive Analysis Methodology**

Exploratory Data Analysis

First, let's read the SpaceX dataset into a Pandas dataframe and print its summary

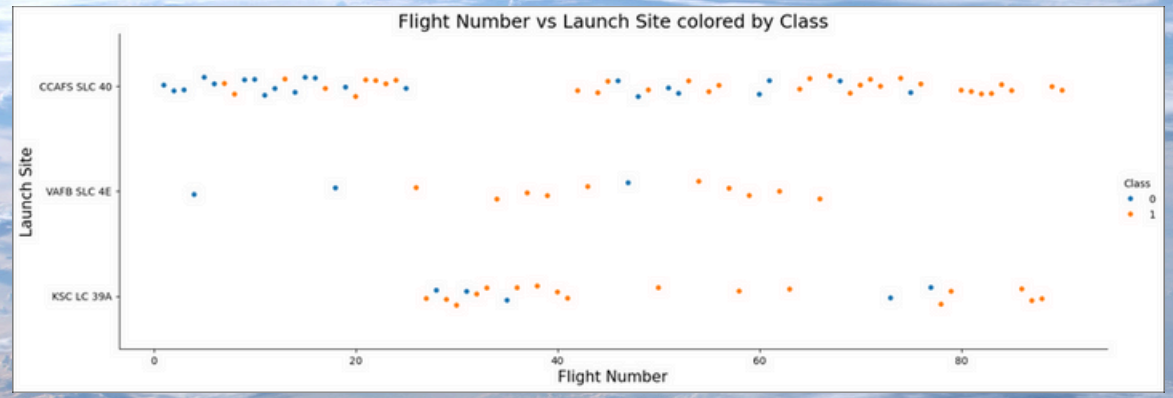
```
from js import fetch
import io

URL = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_2.csv"
resp = await fetch(URL)
dataset_part_2_csv = io.BytesIO((await resp.arrayBuffer()).to_py())
df=pd.read_csv(dataset_part_2_csv)
df.head(5)
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

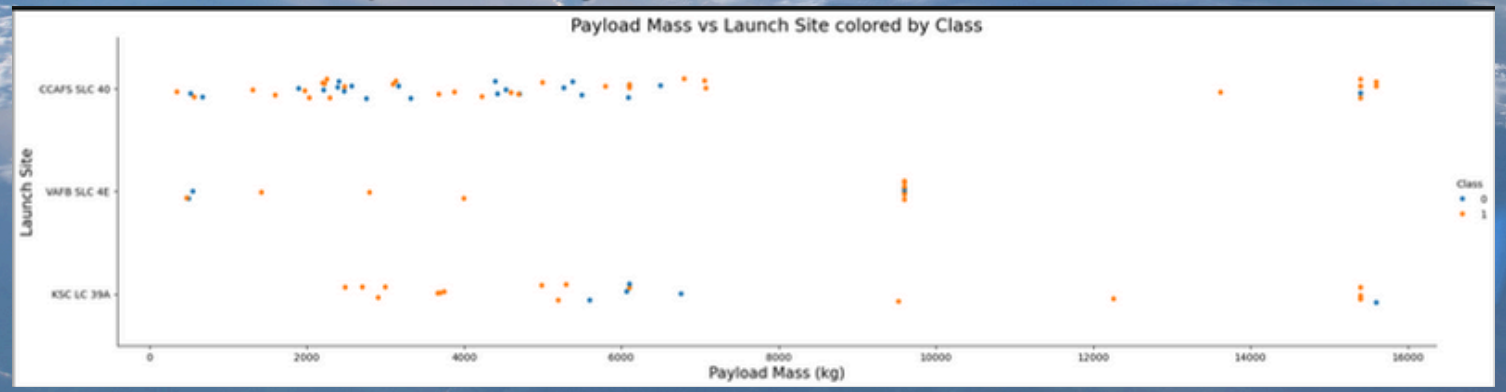
EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY

- ## EDA with Visualization

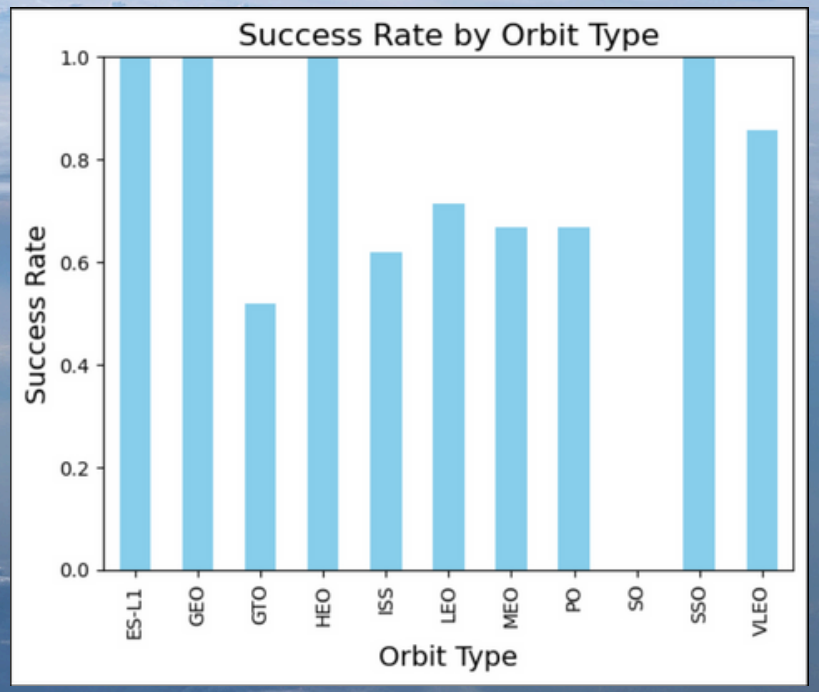
  - Visualize the relationship between Flight Number and Launch Site



  - Visualize the relationship between Payload Mass and Launch Site



  - Visualize the relationship between success rate of each orbit type

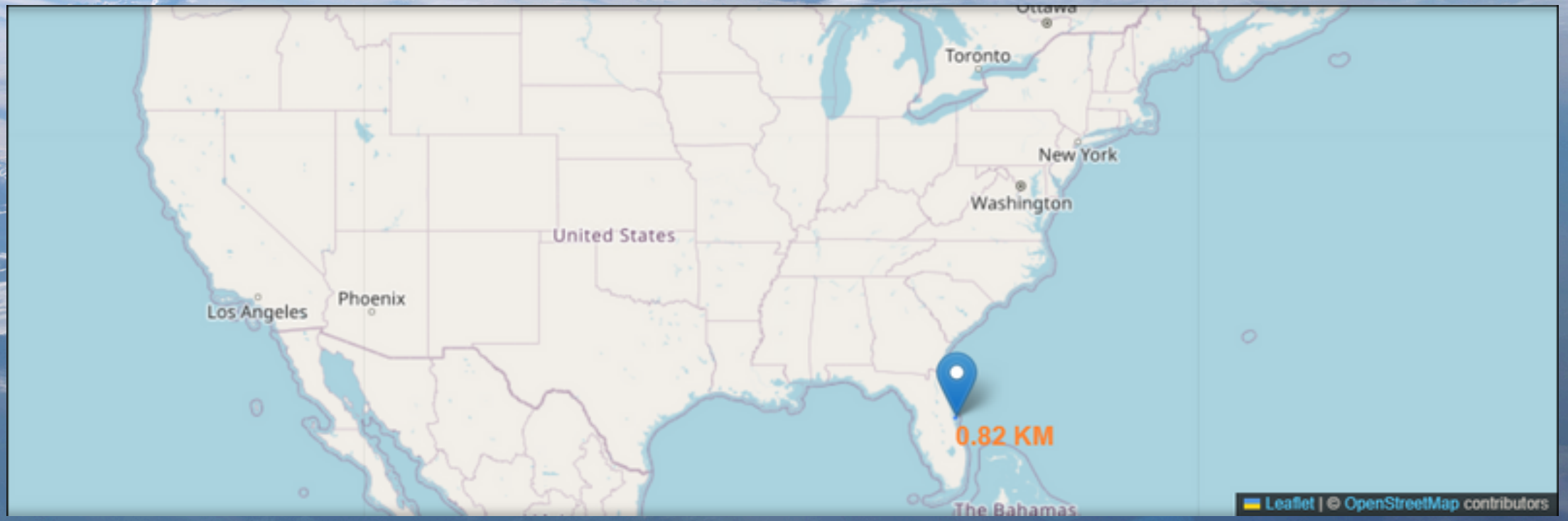


EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY

- Complete EDA with SQL
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - 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
  - List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.
  - 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.

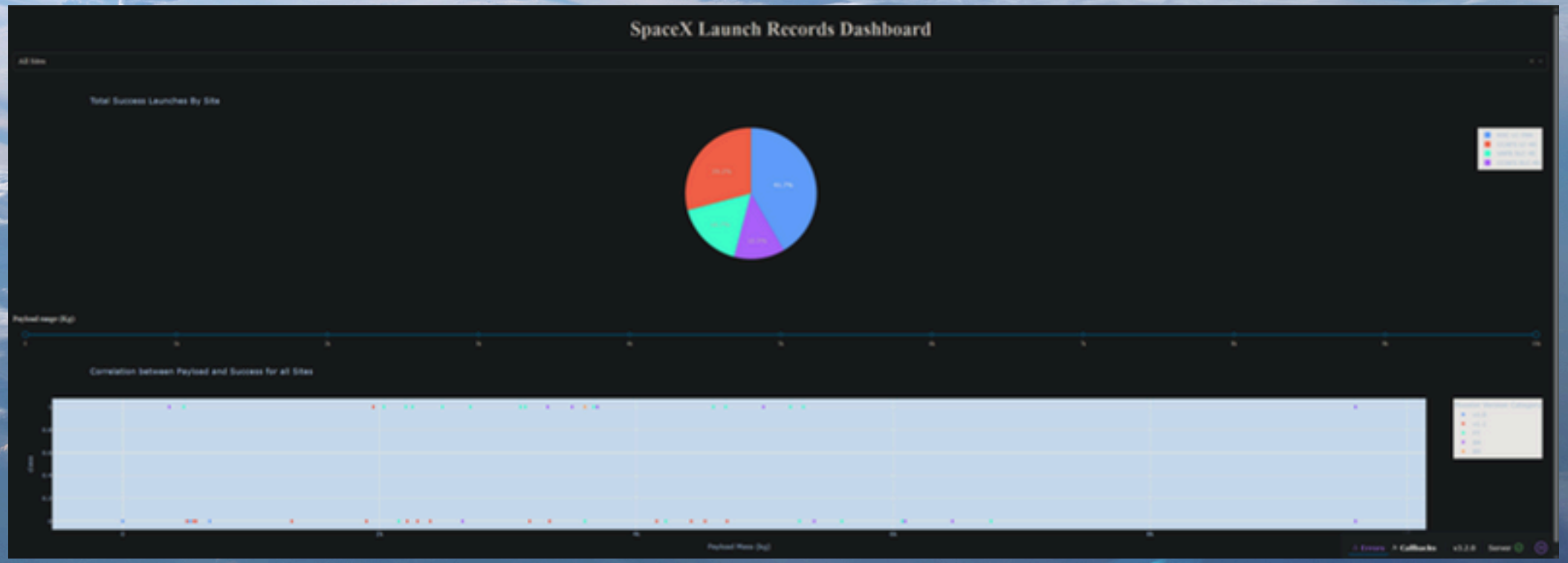
INTERACTIVE MAP WITH FOLIUM

The launch success rate may depend on many factors such as payload mass, orbit type, etc. Finding an optimal location for building a launch site involves many factors. In this lab, we conduct exploratory data analysis using Folium.



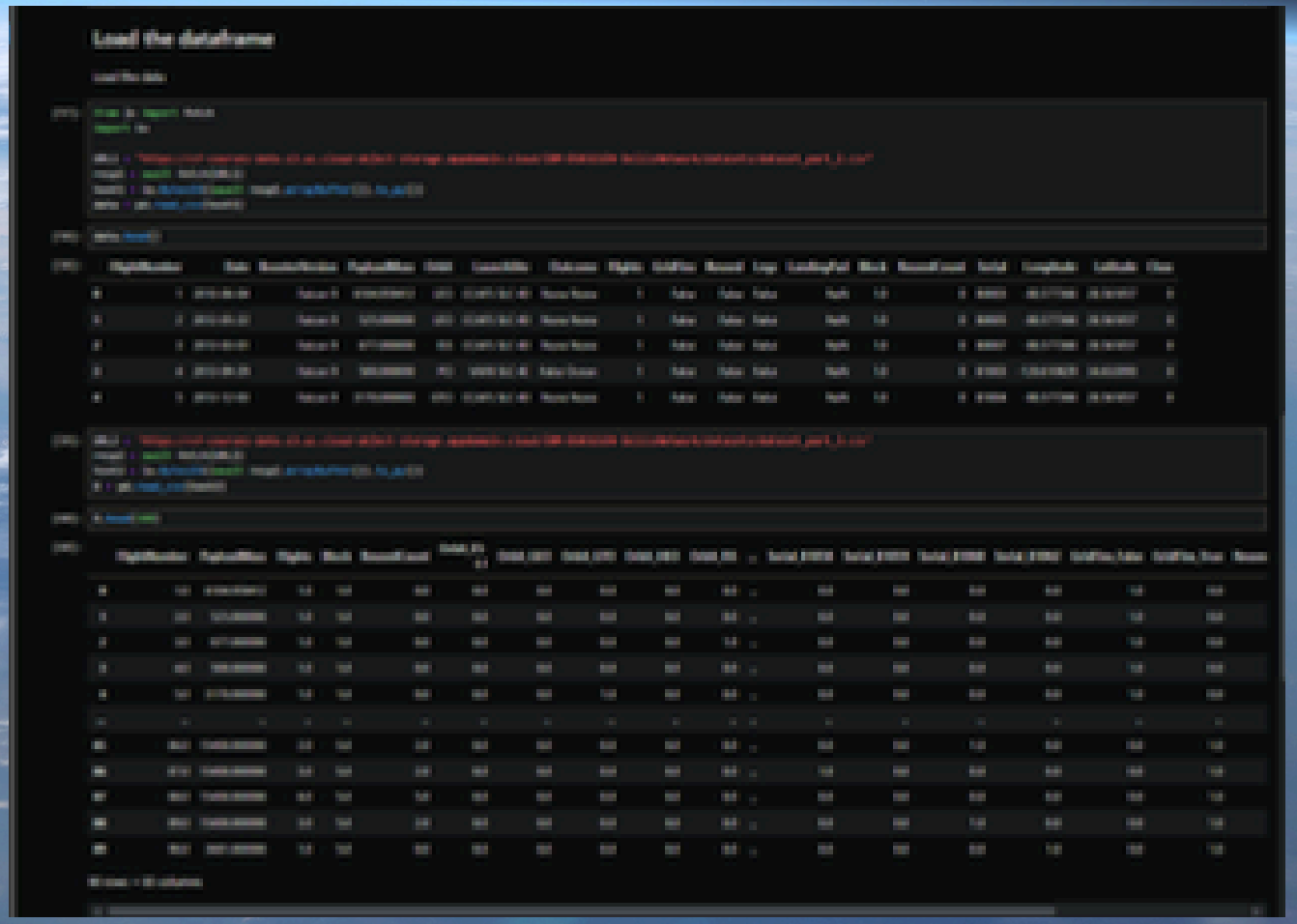
PLOTLY DASH DASHBOARD

In this lab, we build a Plotly Dash application to perform interactive visual analytics on SpaceX launch data in real-time. This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.



PREDICTIVE ANALYSIS (CLASSIFICATION)

In this lab, we build a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully. This will include: Preprocessing, allowing us to standardize our data, and Train_test_split, allowing us to split our data into training and testing data, We will train the model and perform Grid Search, allowing us to find the hyperparameters that allow a given algorithm to perform best.



CONCLUSION

Base on this course, we can conclude that:

- Data understanding and collection are crucial in every data science project. Without clean, structured data, even the best machine learning models would not perform effectively.
- Building accurate models requires careful consideration of the features involved and the choice of algorithms.
- Model evaluation is as important as building the model itself. The effectiveness of a model should be tested rigorously, and the parameters should be tuned to ensure the model performs optimally.
- Deployment is the stage where a model is made usable for the target audience or system. While building and evaluating the model is important, deployment ensures that the results of the model can be applied to real-world data and scenarios.
- The ultimate goal of data science is not only to build a high-performing model but to derive actionable insights that can benefit real-world decision-making. Communicating these insights effectively is just as important as the technical implementation of the model.

INNOVATIVE INSIGHTS

The IBM Applied Data Science Capstone course offers opportunities to innovate by using real-time data for faster decision-making through predictive models. Automated hyperparameter tuning and AI tools can streamline the process of optimizing models and make machine learning accessible to non-experts. By improving model transparency with explainable AI, we can build trust in critical sectors like healthcare and finance. Real-time data pipelines and cross-industry collaborations can enhance predictions by combining data from different fields. AI-powered tools for data cleaning and wrangling can save time and improve model quality. Lastly, focusing on ethics and bias detection in models ensures fairer and more reliable AI systems, making data science more socially responsible.