

#### OUTLINE

- Executive Summary
- Introduction
- Methodology
- Data Collection
- Data Wrangling
- Exploratory Data Analysis (Python)
- Exploratory Data Analysis (SQL)
- Interactive Maps and Dashboards
- Predictive Analysis
- Conclusion

## EXECUTIVE SUMMARY

#### Summary of Methodologies Used

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Interactive Visual Analytics with Folium
- Machine Learning based Predictive Analysis

#### Summary of Results Shown

- Screenshots and Results of Exploratory Data Analysis
- Screenshots and Results of Interactive Analytics
- Screenshots and Results of Predictive Analytics

#### INTRODUCTION

#### Project background:

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars while other providers cost upward of 165 million dollars each, much of the savings is because Space X 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 space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

#### Problem statements:

- What factors determine a successful rocket landing?
- How do the various features determine the success rate of a successful landing?
- What operating conditions need to be ensured for a successful landing program?

### METHODOLOGY

#### Data collection methodology:

- 1. Data was collected using SpaceX API and web scraping from Wikipedia.
- 2. Data Wrangling was performed and One-hot encoding was applied to categorical features
- 3. Performed exploratory data analysis (EDA) using python visualization and SQL
- 4. Performed interactive visual analytics using Folium and Plotly Dash
- 5. Performed ML-based predictive analysis using classification models



### DATA COLLECTION

#### The data was collected using various methods

- Data collection was executed using 'get requests' to the SpaceX API.
- The response content was decoded as a Json using .json() function call and converted into a pandas DataFrame.
- Data cleaning was performed by checking for missing values and replacing them.
- In addition, web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup was performed.
- The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas DataFrame for future analysis.

## DATA COLLECTION – SPACEX API

- Used the 'get request' to the SpaceX
   API to collect data, clean the
   requested data and perform some
   basic data formatting.
- The link to the notebook is
   https://github.com/omsh97/IBM Capstone-Project/blob/main/jupyter labs-spacex-data-collection-api.ipynb

```
Now let's start requesting rocket launch data from SpaceX API with the following URL
     response = requests.get(spacex url)
Check the content of the response
     print(response.content)
 b'[{"fairings":{"reused":false, "recovery_attempt":false, "recovered":false, "ships":[]}, "links":{"patch":{"small":"https://images2.imgbox.com/94/f2/NN6Ph
You should see the response contains massive information about SpaceX launches. Next, let's try to discover some more relevant information for this project.
Task 1: Request and parse the SpaceX launch data using the GET request
To make the requested JSON results more consistent, we will use the following static response object for this project:
     static json url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API call spacex api.json'
We should see that the request was successfull with the 200 status response code
     response=requests.get(static json url)
    response.status code
```

## DATA COLLECTION – SCRAPING

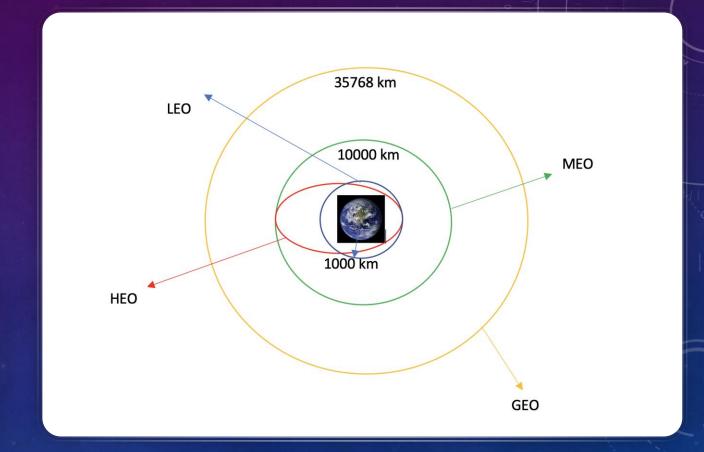
- Conducted web scrapping with BeautifulSoup to collect Falcon 9 launch records
- Parsed the data and converted it into a pandas dataframe.
- The link to the notebook is <u>https://github.com/omsh97/IBM-</u> <u>Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb</u>

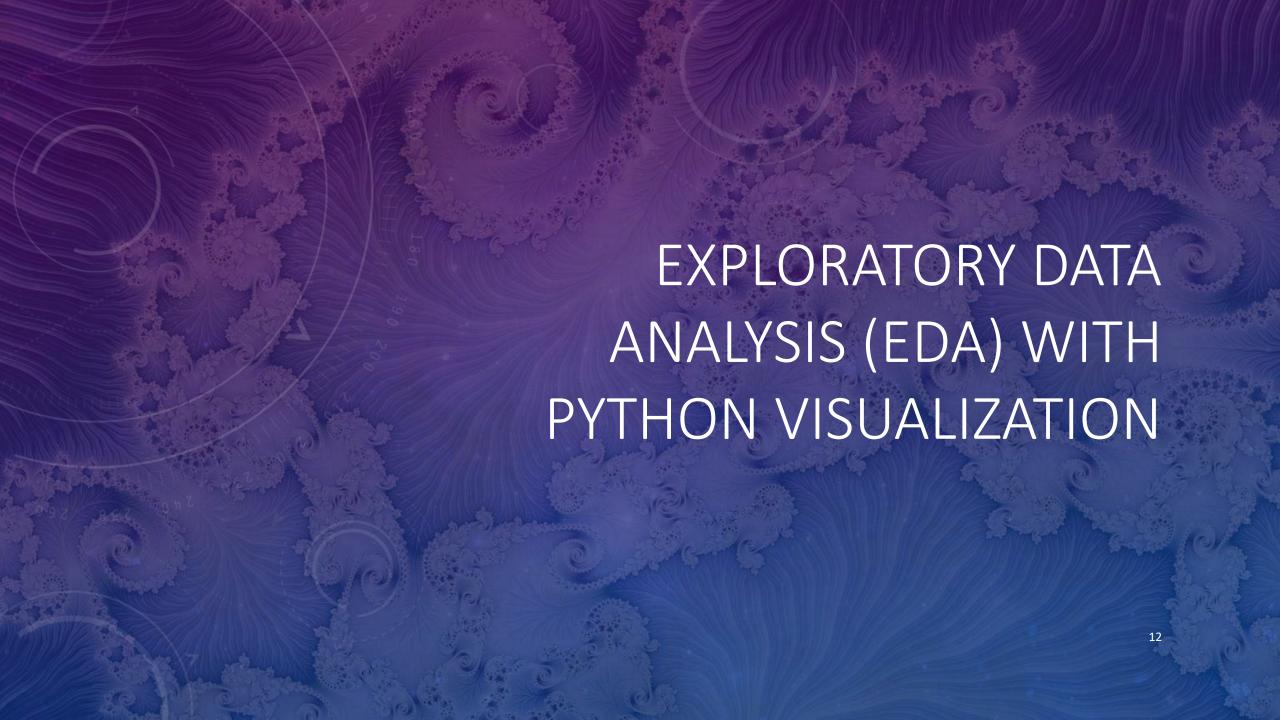
```
static url = "https://en.wikipedia.org/w/index.php?title=List of Falcon 9 and Falcon Heavy launches&oldid=1027686922"
Next, request the HTML page from the above URL and get a response object
TASK 1: Request the Falcon9 Launch Wiki page from its URL
First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response
    headers = {'User-Agent': 'Mozilla/5.0'}
    response = requests.get(static_url, headers = headers)
Create a BeautifulSoup object from the HTML response
     soup = BeautifulSoup(response.text, "html.parser")
Print the page title to verify if the BeautifulSoup object was created properly
    print(soup.title.string)
 <title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
 List of Falcon 9 and Falcon Heavy launches - Wikipedia
TASK 2: Extract all column/variable names from the HTML table header
Next, we want to collect all relevant column names from the HTML table header
```



#### DATA WRANGLING

- Tried to find some patterns in the data.
- Calculated the number of launches at each site, and the number and occurrence of each orbits.
- Created landing outcome labels from outcome column and exported the results to a csv file.
- The link to the notebook is <u>https://github.com/omsh97/IBM-</u> <u>Capstone-Project/blob/main/labs-</u> <u>jupyter-spacex-Data%20wrangling.ipynb</u>



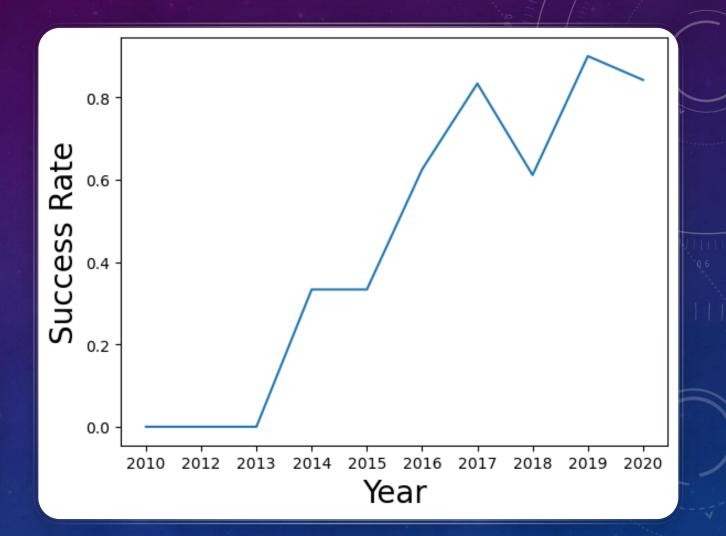


# EDA WITH DATA VISUALIZATION

Performed several types of EDA plotting

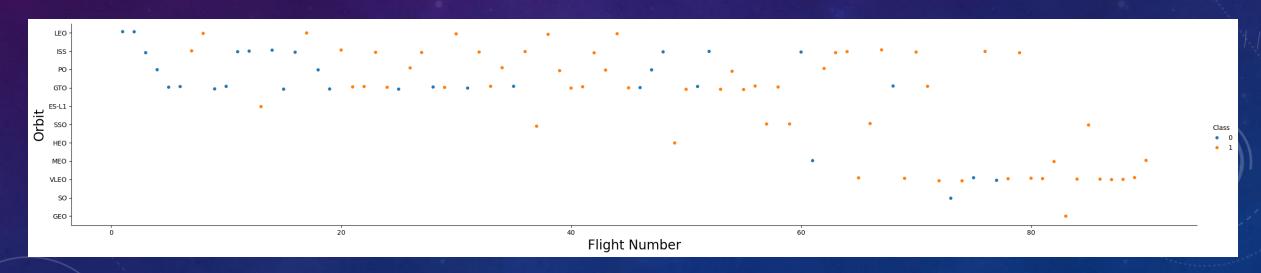
The link to the notebook is <a href="https://github.com/omsh97/IBM-Capstone-">https://github.com/omsh97/IBM-Capstone-</a>

Project/blob/main/edadataviz.ipynb



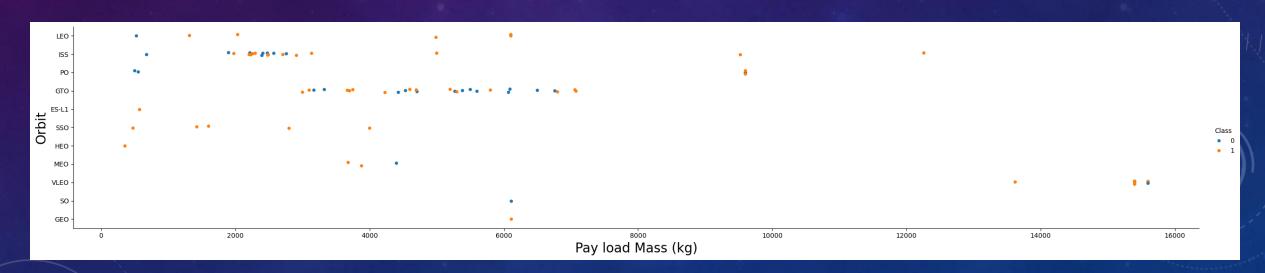
# VISUALIZE THE RELATIONSHIP BETWEEN FLIGHT NUMBER AND ORBIT TYPE

The plot below shows the Flight Number vs. Orbit type. It is observed that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



# VISUALIZE THE RELATIONSHIP BETWEEN PAYLOAD MASS AND ORBIT TYPE

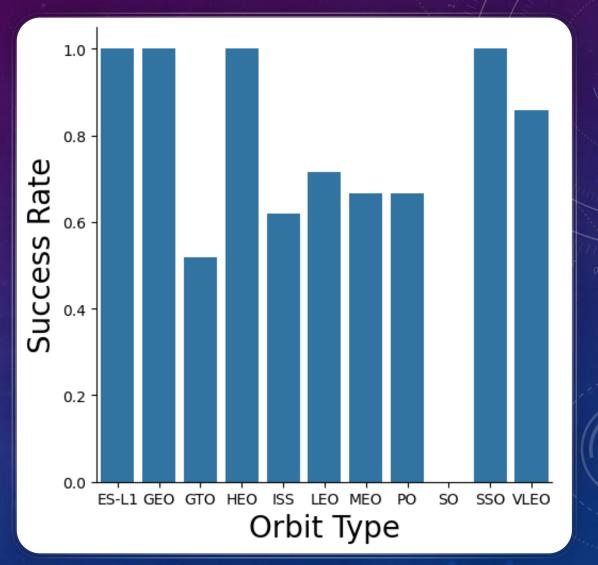
PO, LEO and ISS orbits have achieved more successful landings with a heavier payload



### VISUALIZATION OF SUCCESS RATES FOR EACH ORBIT TYPE

Orbits with highest success rates (100%):

- ES-L1
- GEO
- HEO
- SSO





### EDA WITH SQL

- The SpaceX dataset was loaded into an SQLite database.
- EDA with SQL was used to get insights from the data.
- Some of the queries used are:
  - The names of unique launch sites in the space mission.
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.
- The link to the notebook is <a href="https://github.com/omsh97/IBM-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera sqllite.ipynb">https://github.com/omsh97/IBM-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera sqllite.ipynb</a>

# ALL LAUNCH SITE NAMES

Used the key word **DISTINCT** to show only unique launch sites from the SpaceX data.

#### Task 1

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

%sql select distinct launch\_site from SPACEXTABLE;

\* sqlite:///my\_data1.db

Done.

Launch\_Site

CCAFS LC-40

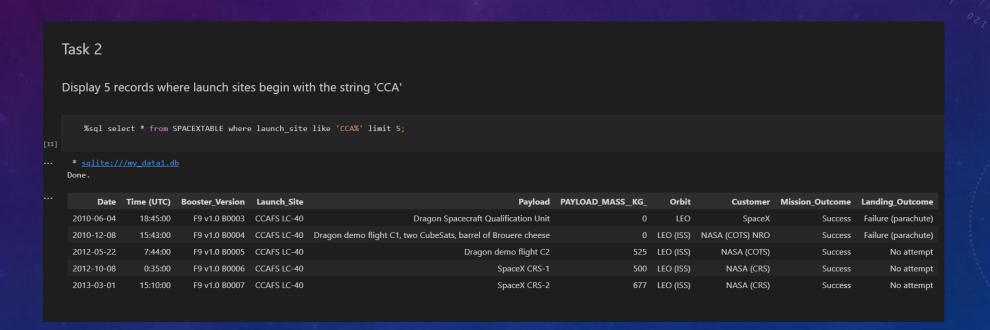
VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

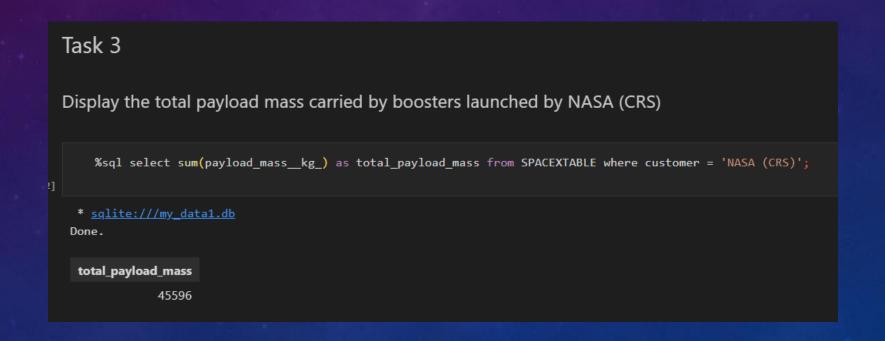
# EDA (SQL) - LAUNCH SITE NAMES BEGINNING WITH 'CCA'

Used the query below to display 5 records where launch sites begin with `CCA`



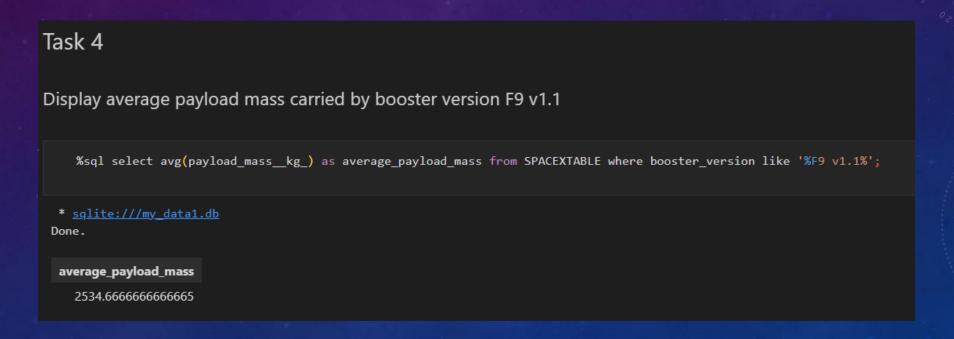
# EDA (SQL) - TOTAL PAYLOAD MASS CARRIED BY BOOSTERS LAUNCHED BY NASA (CRS)

The total payload carried by boosters from NASA is 45596 kg



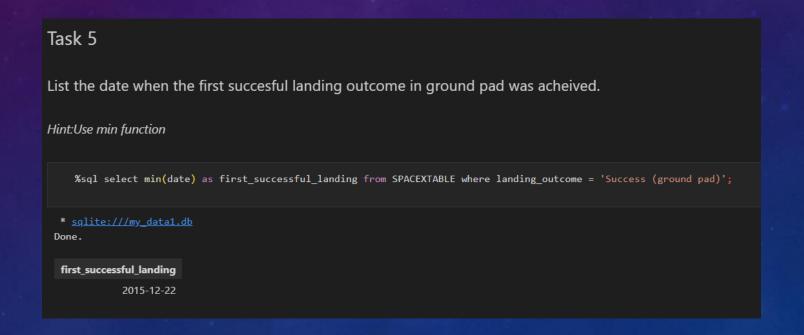
# EDA (SQL) - AVERAGE PAYLOAD MASS CARRIED BY BOOSTER VERSION F9 V1.1

The average payload mass carried by booster version F9 v1.1 is 2534.67 kg



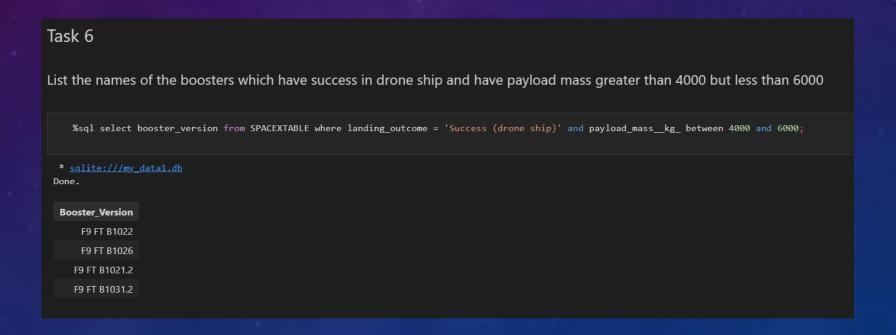
# EDA (SQL) - DATE OF FIRST SUCCESSFUL LANDING OUTCOME IN GROUND PAD WAS ACHIEVED

The date of the first successful landing outcome on ground pad was 22nd December 2015



# EDA (SQL) - NAMES OF SUCCESSFUL BOOSTERS WITH PAYLOAD MASS BETWEEN 4000 AND 6000

Used **WHERE** clause to filter for boosters which have successfully landed on drone ship, applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000



# EDA (SQL) - LIST THE NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

Used **GROUP BY** to show mission outcomes

Task 7				
List the total number of successful and failure mission outcomes				
%sql select mission_outco	ome, count(*)	as total_number from SPACEXTABLE group by mission_outcome;		
* <u>sqlite:///my_data1.db</u> Done.				
Mission_Outcome	total_number			
Failure (in flight)	1			
Success	98			
Success	1			
Success (payload status unclear)	1			

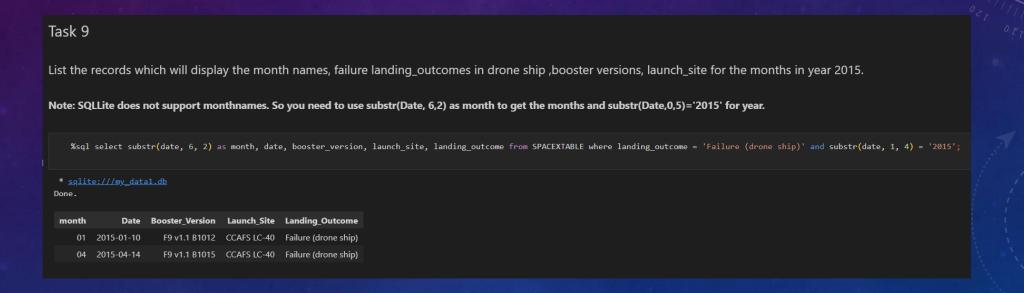
# EDA (SQL) - LIST ALL THE BOOSTER VERSIONS THAT HAVE CARRIED THE MAXIMUM PAYLOAD MASS

Determined the booster versions that carried the maximum payload using a **subquery** in the **WHERE** clause and the **MAX()** function.

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.  #sql select booster_version from SPACEXTABLE where payload_mass_kg_ = (select max(payload_mass_kg_) from SPACEXTABLE);  * sqlite://my_datal.db Done.  Booster_Version F9 B5 B1048.4 F9 B5 B1048.4 F9 B5 B1051.3 F9 B5 B1056.4 F9 B5 B1051.4 F9 B5 B1051.4 F9 B5 B1049.5
* salite://my_datal.db Done.  Booster_Version F9 B5 B1048.4 F9 B5 B1051.3 F9 B5 B1056.4 F9 B5 B1051.4
Booster Version F9 B5 B1048.4 F9 B5 B1049.4 F9 B5 B1051.3 F9 B5 B1056.4 F9 B5 B1051.4
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 B1049.4 F9 B5 B1051.3 F9 B5 B1056.4 F9 B5 B1048.5 F9 B5 B1051.4
F9 B5 B1051.3 F9 B5 B1048.5 F9 B5 B1051.4
F9 B5 B1056.4 F9 B5 B1048.5 F9 B5 B1051.4
F9 B5 B1051.4
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

### EDA (SQL) - LIST THE RECORDS FOR YEAR 2015

Used a combination of the WHERE clause, FROM, AS, and SUBSTR to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015



# EDA (SQL) - RANK THE COUNT OF LANDING OUTCOMES BETWEEN GIVEN DATES

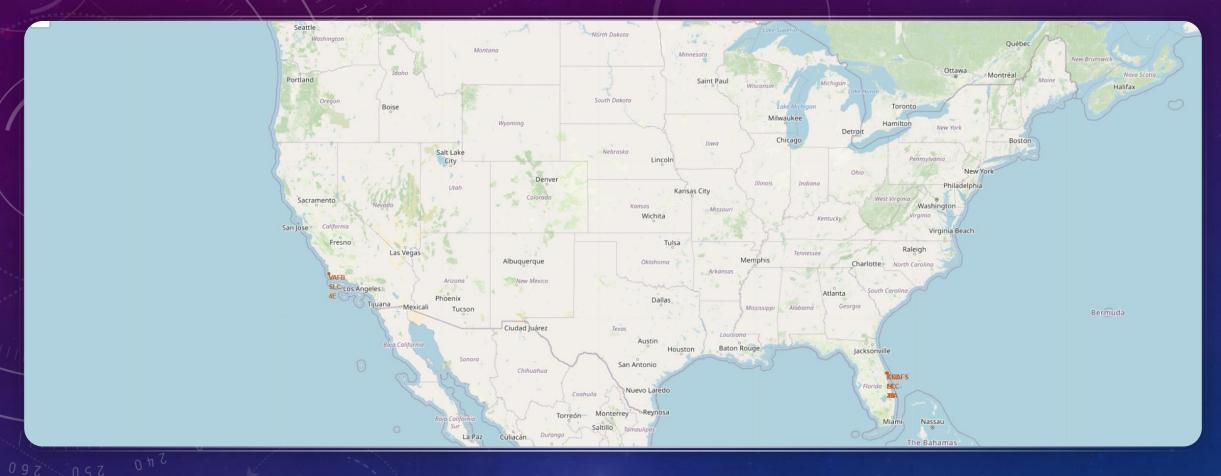
Selected Landing outcomes and the **COUNT** of landing outcomes from the data and used the **WHERE** clause to filter for landing outcomes **BETWEEN 2010-06-04 to 2010-03-20**, applied the **GROUP BY** clause to group the landing outcomes and the **ORDER BY** clause to order the grouped landing outcome in descending order.

Task 10		
Rank the count of land	outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending	g order.
%sql select landing_c	ne, count(*) as count_outcomes from SPACEXTABLE where date between '2010-06-04' and '2017-03-20' group by landing_outcome order by count_outcomes	desc;
* <u>sqlite:///my_data1.db</u> Done.		
Landing_Outcome cou	tcomes	
No attempt	10	
Success (drone ship)		
Failure (drone ship)		
Success (ground pad)		
Controlled (ocean)		
Uncontrolled (ocean)		
Failure (parachute)		
Precluded (drone ship)		



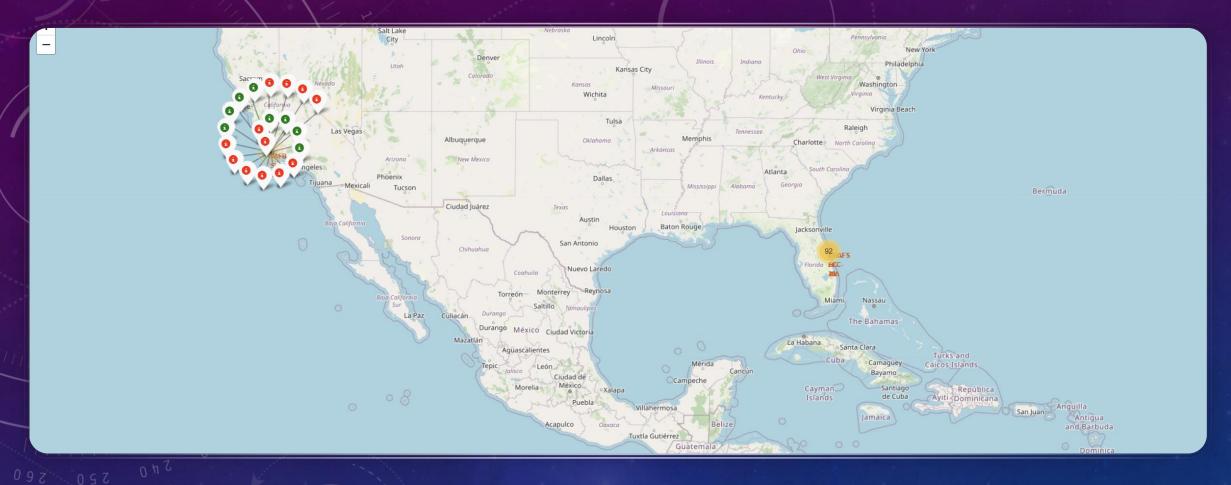
### BUILD AN INTERACTIVE MAP WITH FOLIUM

- Marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- Assigned the feature launch outcomes (failure or success) to class 0 for failure, and class 1 for success.
- Used the color-labeled marker clusters to identify which launch sites have a relatively high success rate.
- Calculated the distances between a launch site to its proximities.
- The link to the notebook is <a href="https://github.com/omsh97/IBM-Capstone-Project/blob/main/lab\_jupyter\_launch\_site\_location.ipynb">https://github.com/omsh97/IBM-Capstone-Project/blob/main/lab\_jupyter\_launch\_site\_location.ipynb</a>



### LAUNCH SITES GLOBAL MAP MARKERS

WE CAN SEE THE SPACEX LAUNCH SITES AT THE COAST OF FLORIDA AND CALIFORNIA



### LAUNCH SITE MARKERS WITH COLOR LABELS

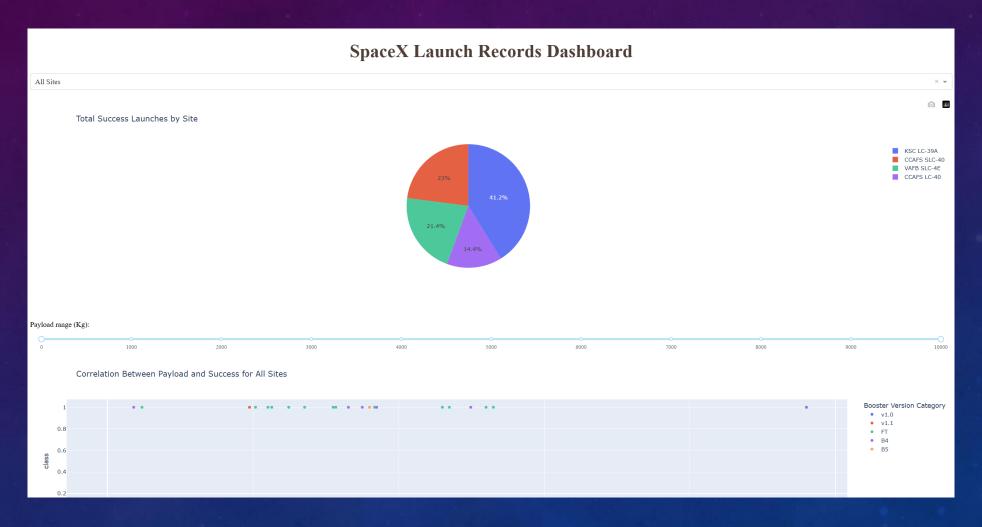
GREEN MARKERS – SUCCESSFUL LAUNCHES | RED MARKERS – FAILED LAUNCHES



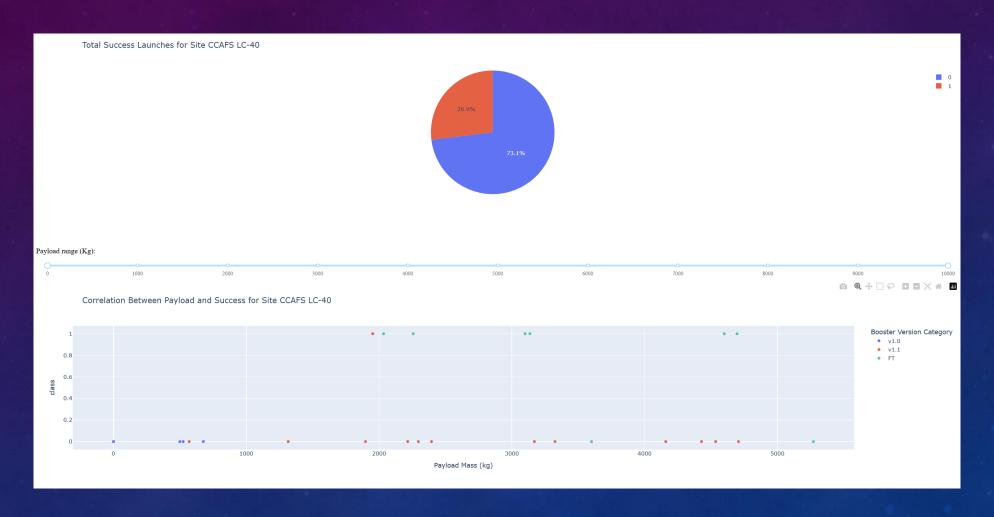
### BUILD A DASHBOARD WITH PLOTLY DASH

- Built an interactive dashboard with Plotly dash using IBM Skills Network Labs
- Implemented dynamic pie charts plots showcasing the total launches by the selected sites
- Implemented dynamic scatter graphs with a slider showing the relationship with Outcome and Payload Mass (kg) for the different booster version.
- The link to the python app code is <a href="https://github.com/omsh97/IBM-Capstone-Project/blob/main/spacex-dash-app.py">https://github.com/omsh97/IBM-Capstone-Project/blob/main/spacex-dash-app.py</a>

### PLOTLY DASHBOARD: HOME



## PLOTLY DASHBOARD: CCAFS LC-40





## PREDICTIVE ANALYSIS (CLASSIFICATION)

- Loaded the data using Numpy and Pandas, transformed the data and split it into training and testing data sets.
- Built different machine learning models and tuned different hyperparameters using GridSearchCV.
- Accuracy was used as the metric for selecting the best model.
- All models performed equally on the Test Data.
- The link to the notebook is <a href="https://github.com/omsh97/IBM-Capstone-">https://github.com/omsh97/IBM-Capstone-</a>
   Project/blob/main/SpaceX Machine%20Learning%20Prediction Part 5.ipynb

#### **CLASSIFICATION ACCURACY**

#### Surprisingly, all models return the same accuracy with Test Data

```
TASK 12

Find the method performs best:

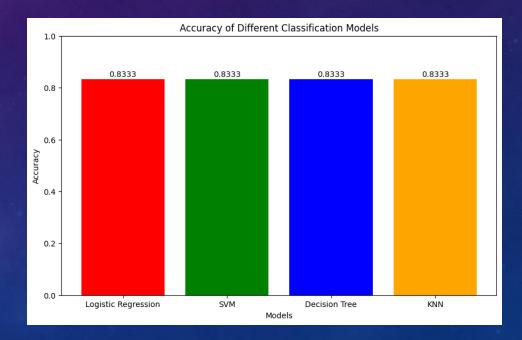
models = ['Logistic Regression', 'SVM', 'Decision Tree', 'KNN']
accuracies = [logreg_accuracy, svm_accuracy, tree_accuracy, knn_accuracy]

plt.figure(figsize=(10, 6))
plt.bar(models, accuracies, color=['red', 'green', 'blue', 'orange'])

plt.title('Accuracy of Different Classification Models')
plt.xlabel('Models')
plt.ylabel('Accuracy')
plt.ylabel('Accuracy')
plt.ylim(0, 1) # Setting the y-axis range from 0 to 1

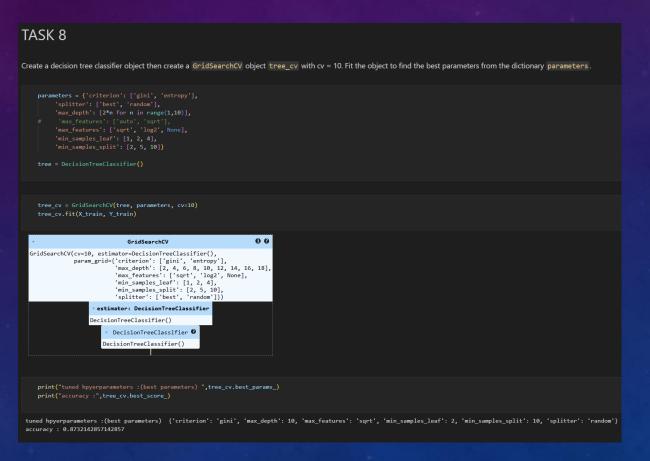
for i, accuracy in enumerate(accuracies):
    plt.text(i, accuracy + 0.01, f'{accuracy:.4f}', ha='center')

plt.show()
```



#### CLASSIFICATION ACCURACY

Decision Tree Classifier, however, returned the best results with Training Data





#### CONCLUSION

- As the number of launches goes up, the rate of failure and therefore the probability of failure seems to decrease
- Launch success rate started to increase from 2013 until 2020.
- Launch Site CCAFS SLC 40 has the highest number of launches and the maximum number of failures as well.
- Orbits ES-L1, GEO, HEO, SSO had a success rate of 100%.
- KSC LC-39A had the most successful launches of any sites.
- All classifiers perform equally with current data.
- The Decision tree classifier might be the best machine learning algorithm for this task w.r.t training data accuracy.

