



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Erica Shackelford
May 25, 2024



Outline

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

Executive Summary

Summary of Methodologies

The research project aimed to identify the factors contributing to successful rocket landings by SpaceX. The methodologies involved were:

1. **Data Collection:** Gathered data from the SpaceX REST API and performed web scraping techniques.
2. **Data Wrangling:** Preprocessed the data to create a binary outcome variable indicating successful or failed landings.
3. **Exploratory Data Analysis (EDA):**
 - Explored the data using data visualization techniques, considering factors such as payload, launch site, flight number, and yearly trends.
 - Analyzed the data with SQL, calculating statistics like total payload, payload range for successful launches, and the total number of successful and failed outcomes.
 - Investigated launch site success rates and their proximity to geographical markers.
 - Visualized the launch sites with the highest success rates and successful payload ranges.
4. **Interactive Visualization:** Built an interactive map using the Folium library to visualize launch sites and related data.
5. **Dashboard Creation:** Developed an interactive dashboard using Plotly Dash for comprehensive data visualization and exploration.
6. **Predictive Analysis (Classification):** Built machine learning models using logistic regression, support vector machines (SVM), decision trees, and K-nearest neighbors (KNN) to predict landing outcomes based on the collected and preprocessed data.

Summary of all results

Exploratory Data Analysis

- Launch success improved over time
- KSC LC-39A had highest landing success rate
- Orbits ES-L1, GEO, HEO, SSO had 100% success

Visualization/Analytics

- Most launch sites near equator
- All launch sites near coasts

Predictive Analytics

- Decision tree model performed best
- All models showed similar predictive performance

Introduction

- **Project background and context**

SpaceX is revolutionizing space travel by reducing launch costs through reusable rocket technology. Their Falcon 9 launches cost \$62 million thanks to reusing the first stage, while competitors spend up to \$165 million per launch.

Accurately predicting first stage landing success is key to optimizing reusability and further driving down costs. By leveraging public data and machine learning models, we can forecast landing outcomes, unlocking significant cost savings for SpaceX and other providers through reusable rocket stages

- **Problems you want to find answers**

The project task is to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
 - Gathered data from the SpaceX REST API and performed web scraping techniques on Wikipedia.
- **Perform data wrangling**
 - Preprocessed the data by filtering, handling missing values, and applying one-hot encoding to prepare the data for analysis and modeling.
- **Perform exploratory data analysis (EDA) using visualization and SQL**
 - Explored the data using SQL queries and data visualization techniques to uncover insights and patterns.
- **Perform interactive visual analytics using Folium and Plotly Dash**
 - Utilized Folium and Plotly Dash libraries to create interactive visualizations and dashboards for in-depth data exploration.
- **Perform predictive analysis using classification models**
 - Built and evaluated classification models, including Logistic Regression, K-Nearest Neighbors, Support Vector Machines, and Decision Trees, to predict rocket landing outcomes. Tuned and compared the models to identify the best-performing one with optimal parameters.

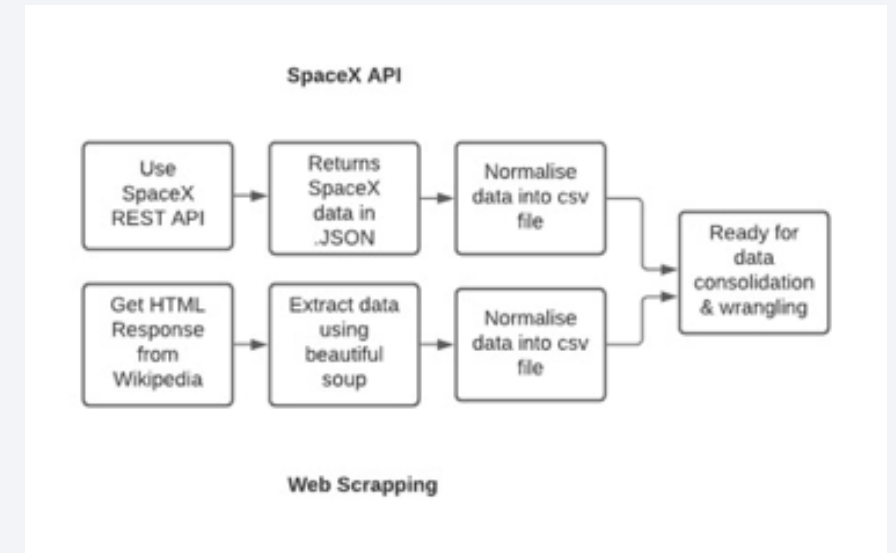
Data Collection

- Describe how data sets were collected.

The project utilized two main data sources:

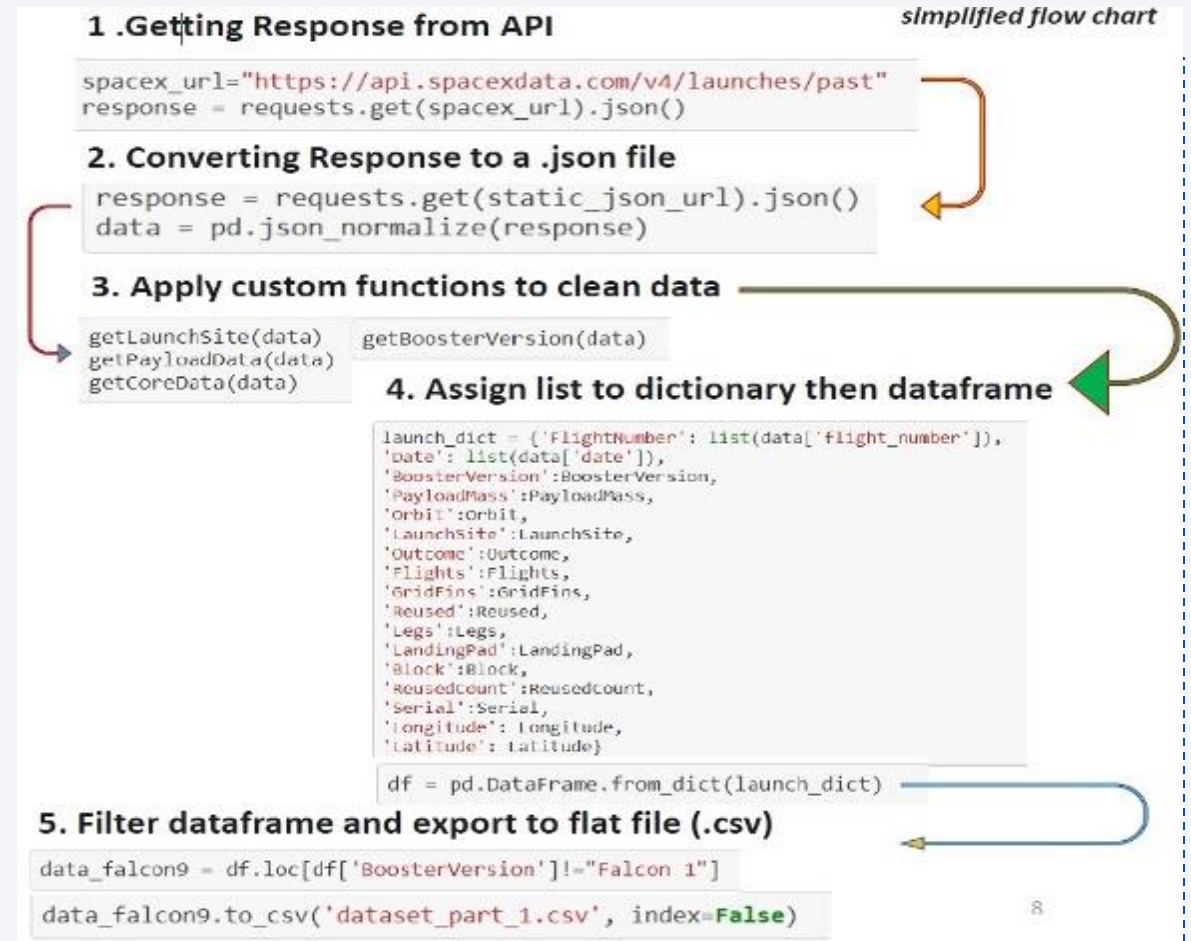
- SpaceX Launch Data: This comprehensive dataset was obtained through the SpaceX REST API (api.spacexdata.com/v4/). It provided detailed information about SpaceX launches, including rocket specifications, payload details, launch parameters, landing specifications, and landing outcomes.
- Falcon 9 Launch Data from Wikipedia: To complement the SpaceX API data, web scraping techniques were employed using the BeautifulSoup library to extract relevant Falcon 9 launch information from Wikipedia pages.

By combining these two data sources, a comprehensive and reliable dataset was constructed, serving as the foundation for further analysis and modeling.



Data Collection – SpaceX API

- Request data from the SpaceX API for rocket launch information.
- Decode the API response using `.json()` and convert it to a DataFrame using `.json_normalize()`.
- Use custom functions to request specific information about the launches from the SpaceX API.
- Create a dictionary from the obtained data.
- Convert the dictionary into a DataFrame.
- Filter the DataFrame to include only Falcon 9 launches.
- Replace missing values in the Payload Mass column with the calculated mean.
- Export the DataFrame to a CSV file.
- https://github.com/shackerica/IBM-Data-Science-SpaceX-Falcon9/blob/main/Notebooks/1_jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Request Falcon 9 launch data from Wikipedia.
- Create a BeautifulSoup object from the HTML response.
- Extract the column names from the HTML table header.
- Collect data by parsing the HTML tables.
- Create a dictionary from the parsed data.
- Convert the dictionary into a DataFrame.
- Export the DataFrame to a CSV file.
- https://github.com/shackerica/IBM-Data-Science-SpaceX-Falcon9/blob/main/Notebooks/2_jupyter-labs-webscraping.ipynb

1. Getting Response from HTML

```
page = requests.get(static_url)
```

2. Creating BeautifulSoup Object

```
soup = BeautifulSoup(page.text, 'html.parser')
```

3. Finding tables

```
html_tables = soup.find_all('table')
```

4. Getting column names

```
column_names = []
temp = soup.find_all('th')
for x in range(len(temp)):
    try:
        name = extract_column_from_header(temp[x])
        if (name is not None and len(name) > 0):
            column_names.append(name)
    except:
        pass
```

5. Creation of dictionary

```
launch_dict = dict.fromkeys(column_names)

# Remove an irrelevant column
del launch_dict['Date and time ( )']

launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
launch_dict['Version Booster'] = []
launch_dict['Booster landing'] = []
launch_dict['Date'] = []
launch_dict['Time'] = []
```

6. Appending data to keys (refer) to notebook block 12

```
In [12]: extracted_row = 0
#Extract each table
for table_number, table in enumerate(
    # get table row
    for rows in table.find_all("tr"):
        #check to see if first table
```

7. Converting dictionary to dataframe

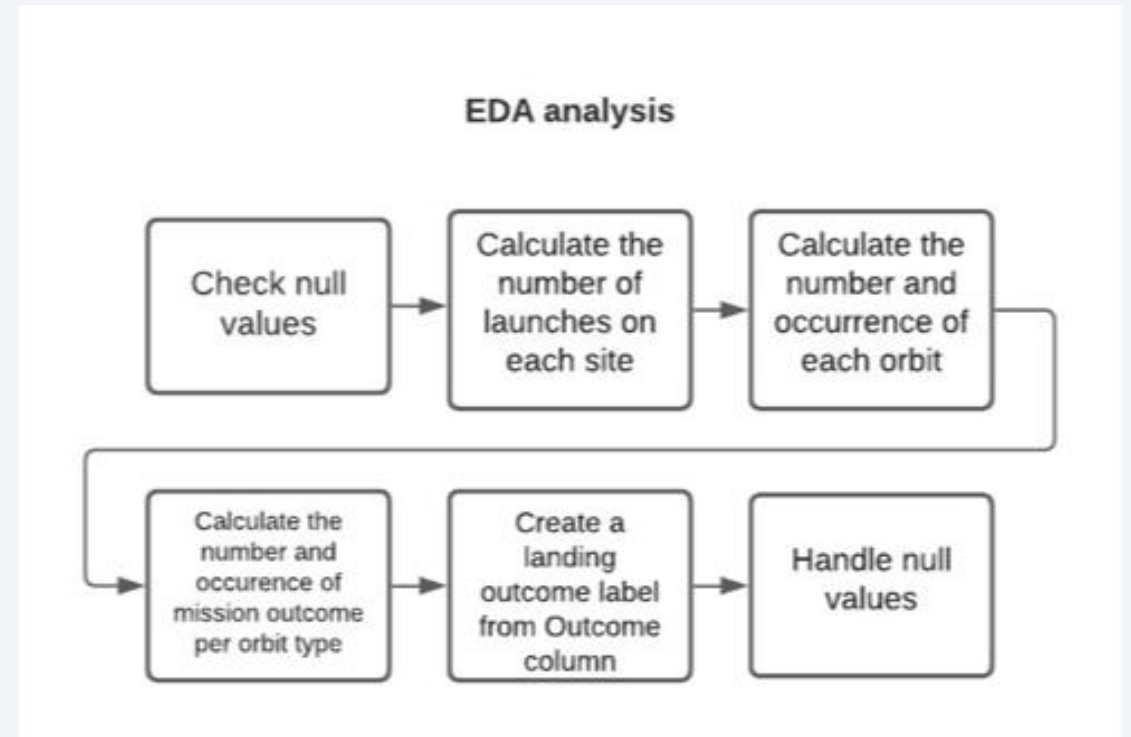
```
df = pd.DataFrame.from_dict(launch_dict)
```

8. Dataframe to .CSV

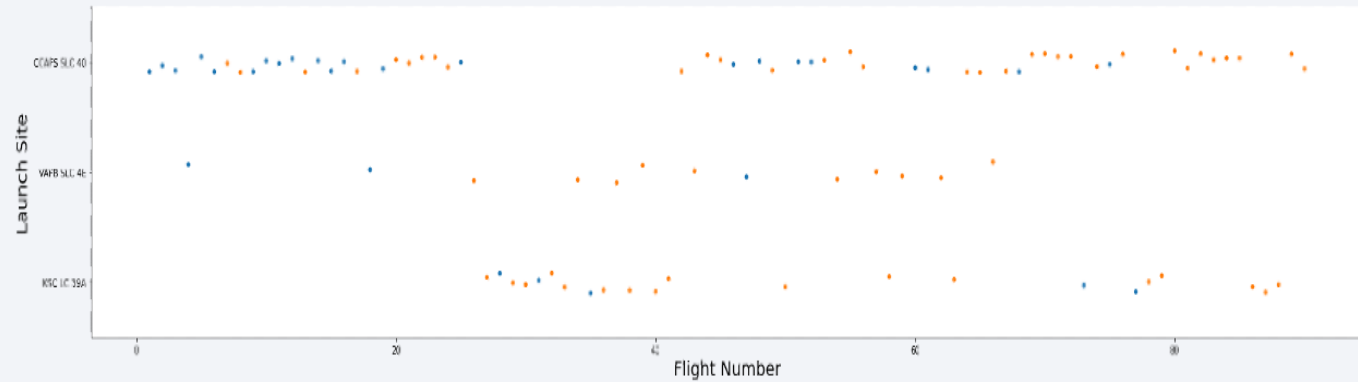
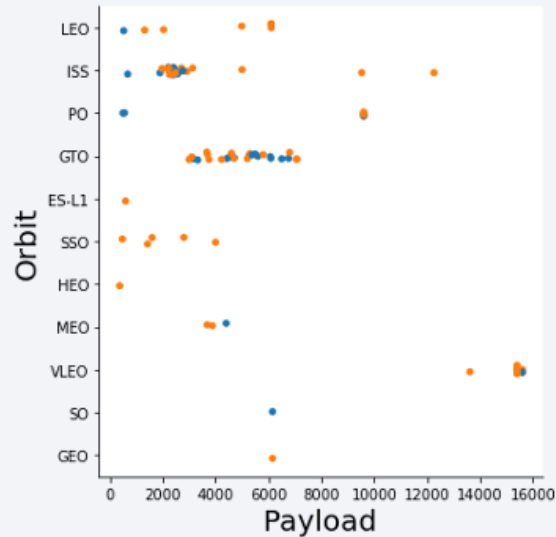
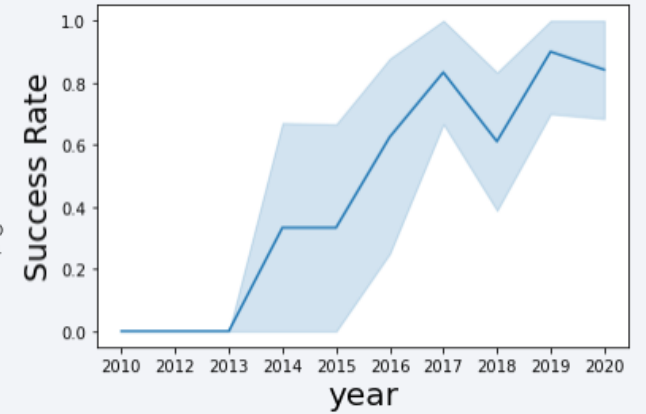
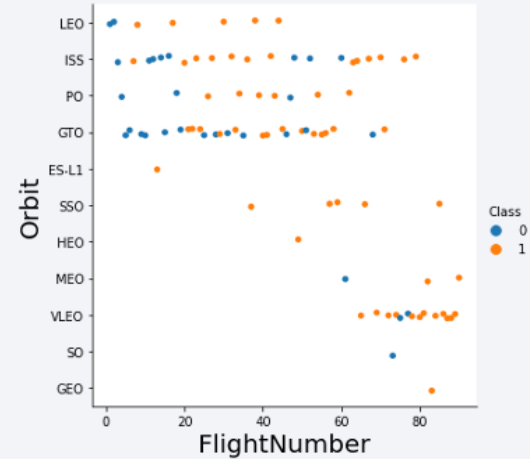
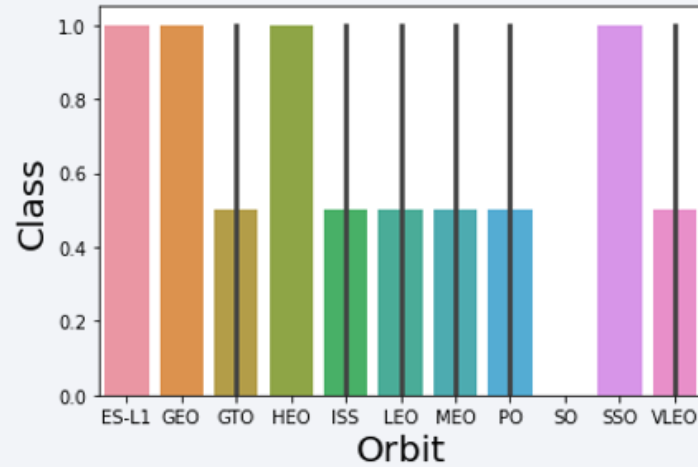
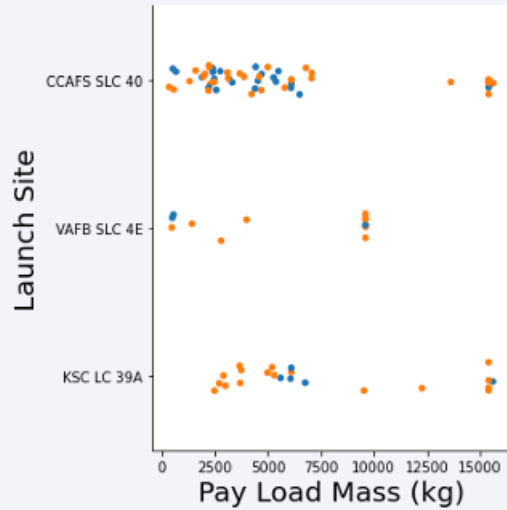
```
df.to_csv('spacex_web_scraped.csv', index=False)
```

Data Wrangling

- Perform Exploratory Data Analysis (EDA):
- Determine data labels.
- Calculate the number of launches for each site.
- Calculate the occurrence of each orbit.
- Calculate the occurrence of each mission outcome per orbit type.
- Create a Binary Landing Outcome Column:
- Add a binary column for landing outcomes (dependent variable).
- Export Data to CSV File:
- Save the processed DataFrame to a CSV file.
- https://github.com/shackerica/IBM-Data-Science-SpaceX-Falcon9/blob/main/Notebooks/3_labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

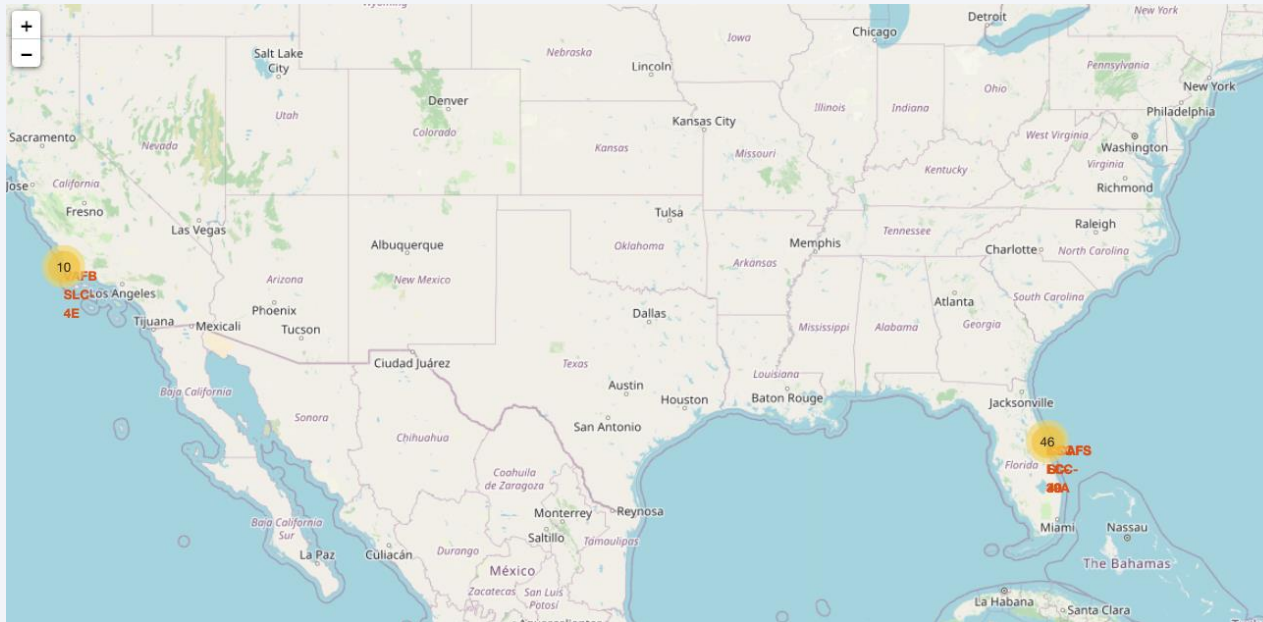


- https://github.com/shackerica/IBM-Data-Science-SpaceX-Falcon9/blob/main/Notebooks/5_jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Display unique launch sites.
- Display 5 records of launch sites starting with 'KSC'.
- Show total payload mass by NASA (CRS) launches.
- Calculate average payload mass of booster version F9 v1.1.
- List date of successful drone ship landings.
- List boosters succeeding in ground pad with payload mass 4000-6000.
- Show total count of successful and failed missions.
- Identify booster_versions with maximum payload mass.
- List records of successful ground pad landings, booster versions, and launch sites in 2017.
- Rank successful landings between June 4, 2010, and March 20, 2017, by count in descending order.
- https://github.com/shackerica/IBM-Data-Science-SpaceX-Falcon9/blob/main/Notebooks/4_jupyter-labs-eda-sql-coursera_sqlite.ipynb

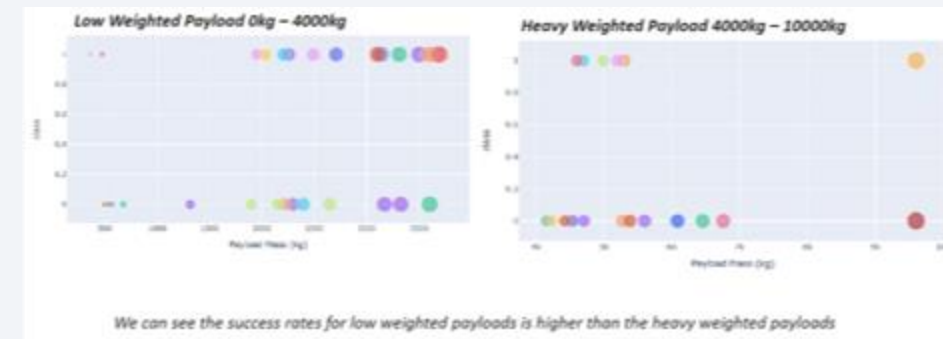
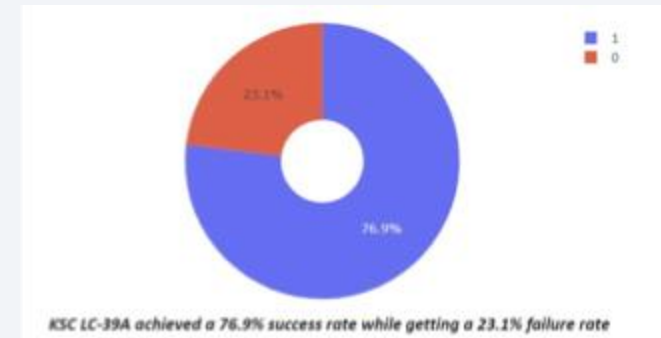
Build an Interactive Map with Folium



- Added a blue circle at the coordinate of NASA Johnson Space Center with a popup label displaying its name, using its latitude and longitude coordinates.
- Added red circles at all launch site coordinates with a popup label displaying their names, using their latitude and longitude coordinates.
- Included colored markers indicating successful (green) and unsuccessful (red) launches at each launch site to illustrate high success rates.
- Incorporated colored lines to indicate the distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city.
- https://github.com/shackerica/IBM-Data-Science-SpaceX-Falcon9/blob/main/Notebooks/6_lab_jupyter_launch_site_location.ipynb

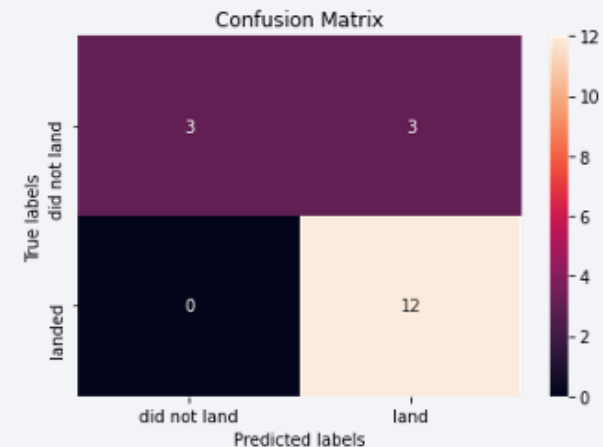
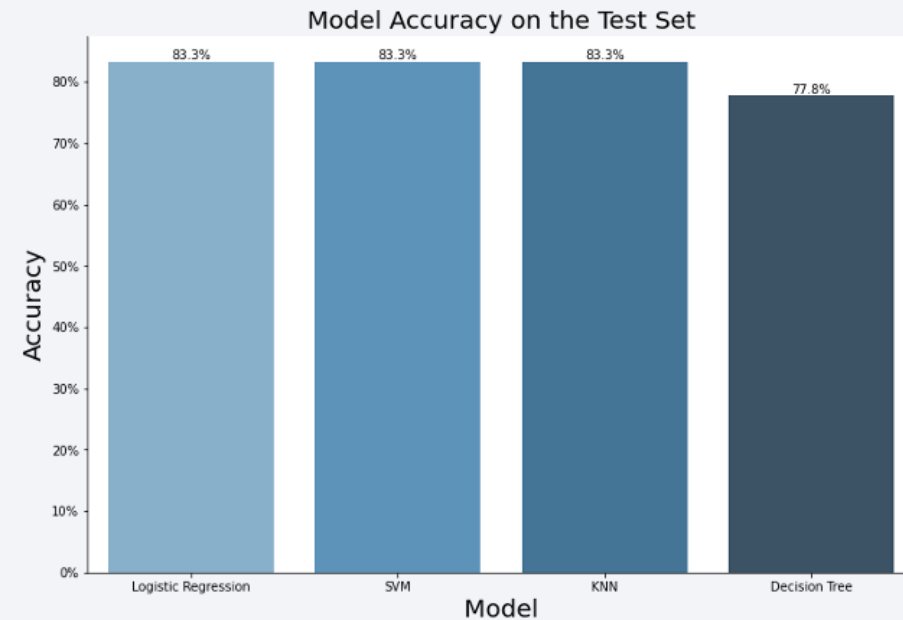
Build a Dashboard with Plotly Dash

- Dropdown List with Launch Sites: Provides users with a dropdown menu to select either all launch sites or a specific launch site from the available options.
- Pie Chart Showing Successful Launches: Presents a pie chart illustrating the percentage of successful and unsuccessful launches relative to the total number of launches, allowing users to visualize the success rate.
- Slider of Payload Mass Range: Offers users a slider interface to specify their desired range of payload masses, enabling them to filter and narrow down data based on payload mass criteria.
- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version: Displays a scatter chart depicting the relationship between payload mass and launch success rate, categorized by booster version. Users can observe any correlation between payload mass and launch success.
- https://github.com/shackerica/IBM-Data-Science-SpaceX-Falcon9/blob/main/Notebooks/7_SpaceX_Interactive_Visual_Analytics_Plotly.py



Predictive Analysis (Classification)

- Convert Class column to a NumPy array.
- Standardize data using StandardScaler.
- Split data using train_test_split.
- Perform GridSearchCV with cv=10 for parameter optimization.
- Apply GridSearchCV on logistic regression, support vector machine, decision tree, and K-Nearest Neighbor models.
- Calculate accuracy scores for all models.
- Assess confusion matrices for all models.
- Determine best model based on Jaccard Score, F1 Score, and Accuracy.
- The SVM, KNN, and Logistic Regression model achieved the highest accuracy at 83.3%, while the SVM performs the best in terms of Area Under the Curve at 0.958.
- https://github.com/shackerica/IBM-Data-Science-SpaceX-Falcon9/blob/main/Notebooks/8_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

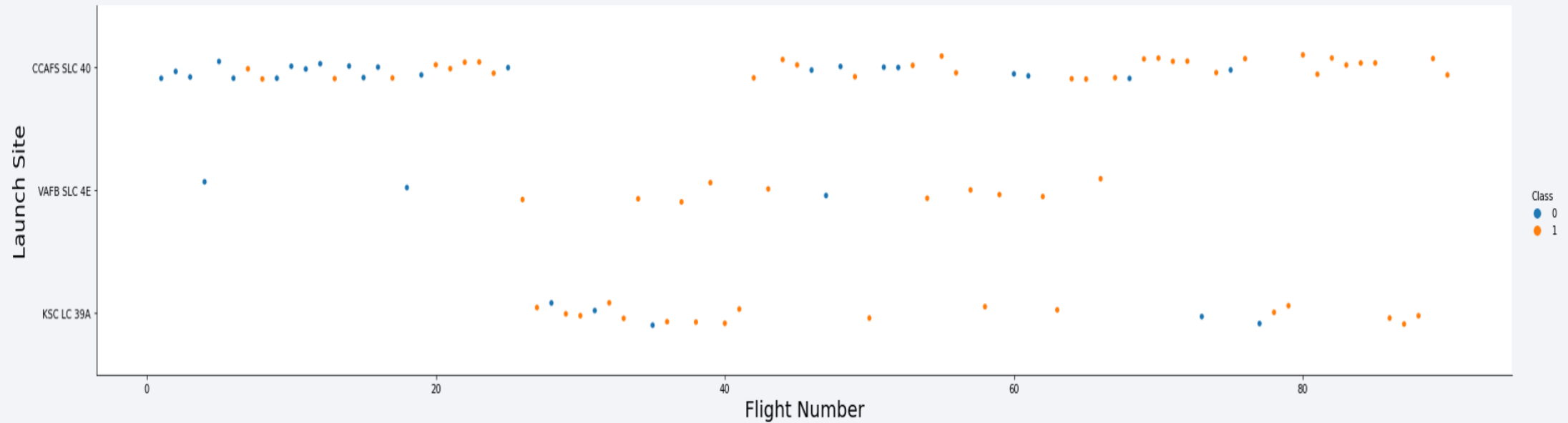
- Exploratory Data Analysis:
- Success rate of launches has increased over time.
- KSC LC-39A exhibits the highest success rate among landing sites.
- Orbits ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.
- Visual Analytics:
- Most launch sites are located near the equator and are coastal.
- Launch sites are strategically positioned away from populated areas and transportation routes to minimize potential damage from failed launches, while remaining accessible for logistical support.
- Predictive Analytics:
- The Decision Tree model demonstrates superior predictive performance for the dataset.

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of blue and red, creating a sense of motion or data flow. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is high-tech and digital.

Section 2

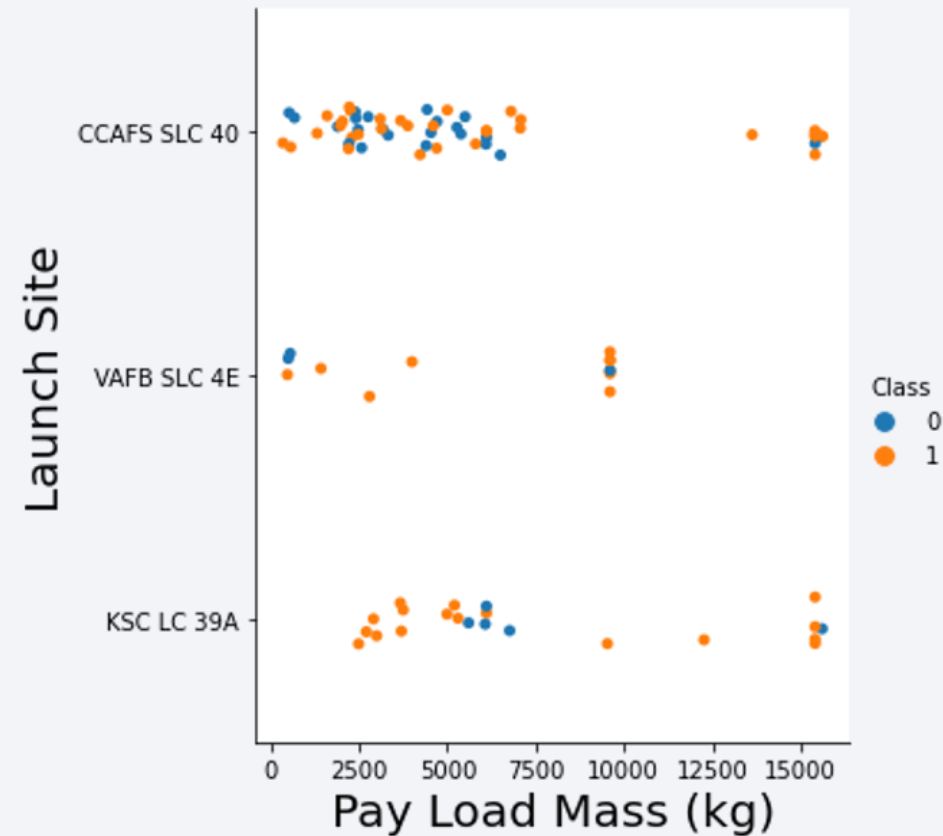
Insights drawn from EDA

Flight Number vs. Launch Site



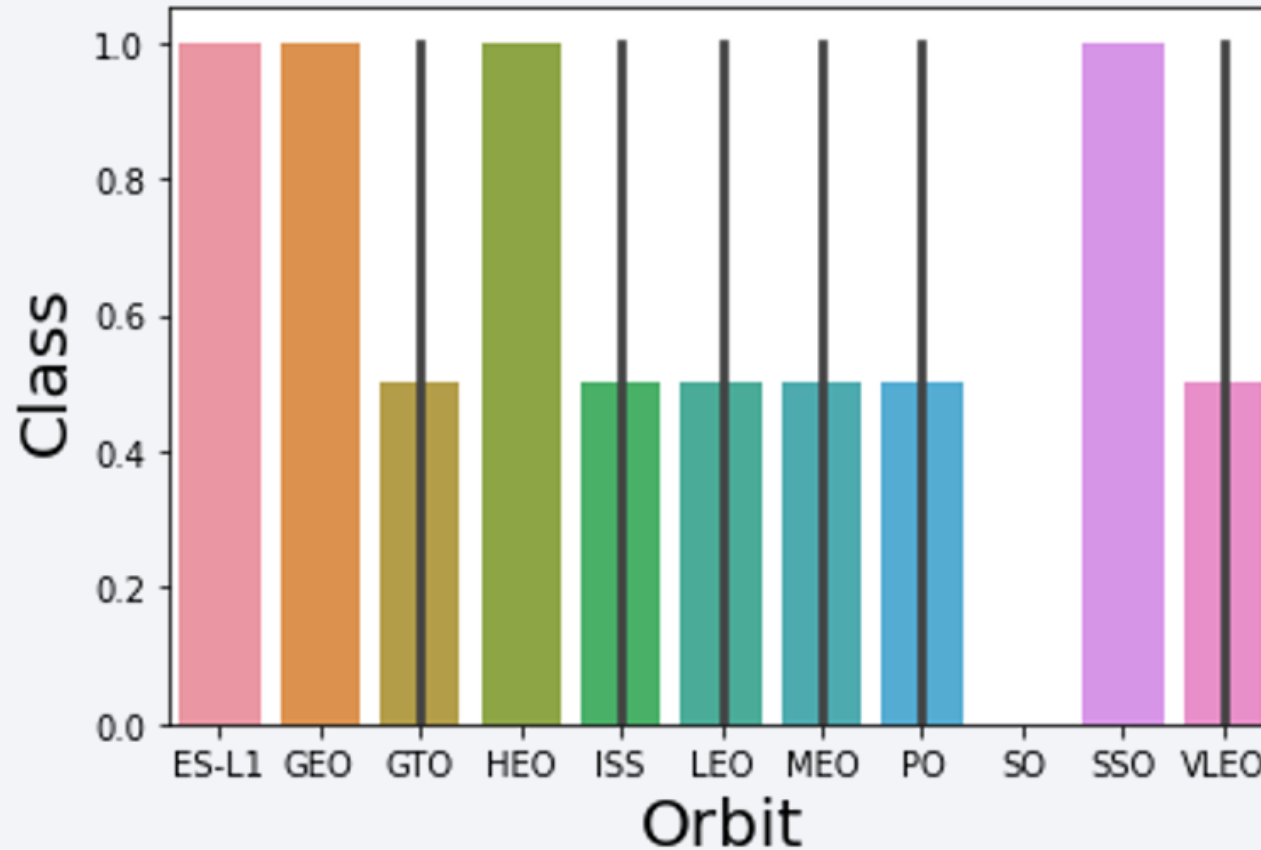
- Launches from the site of CCAFS SLC 40 are significantly higher than launches form other sites.

Payload vs. Launch Site



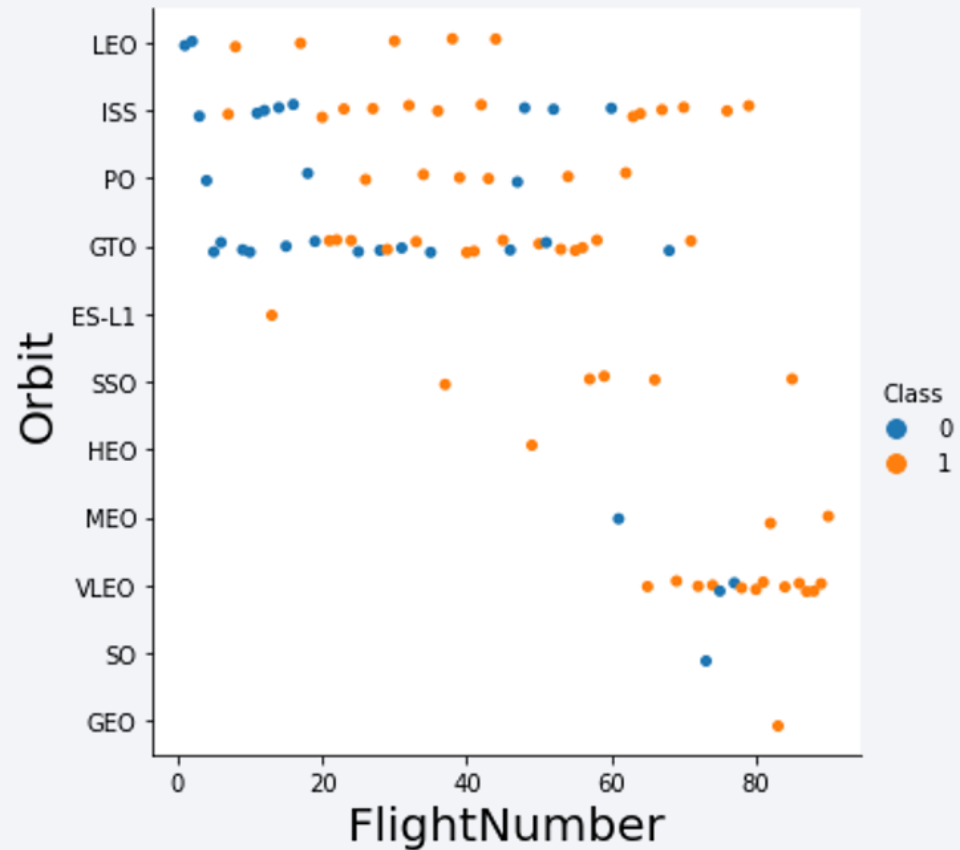
- The majority of IPay Loads with lower Mass have been launched from CCAFS SLC 40.

Success Rate vs. Orbit Type



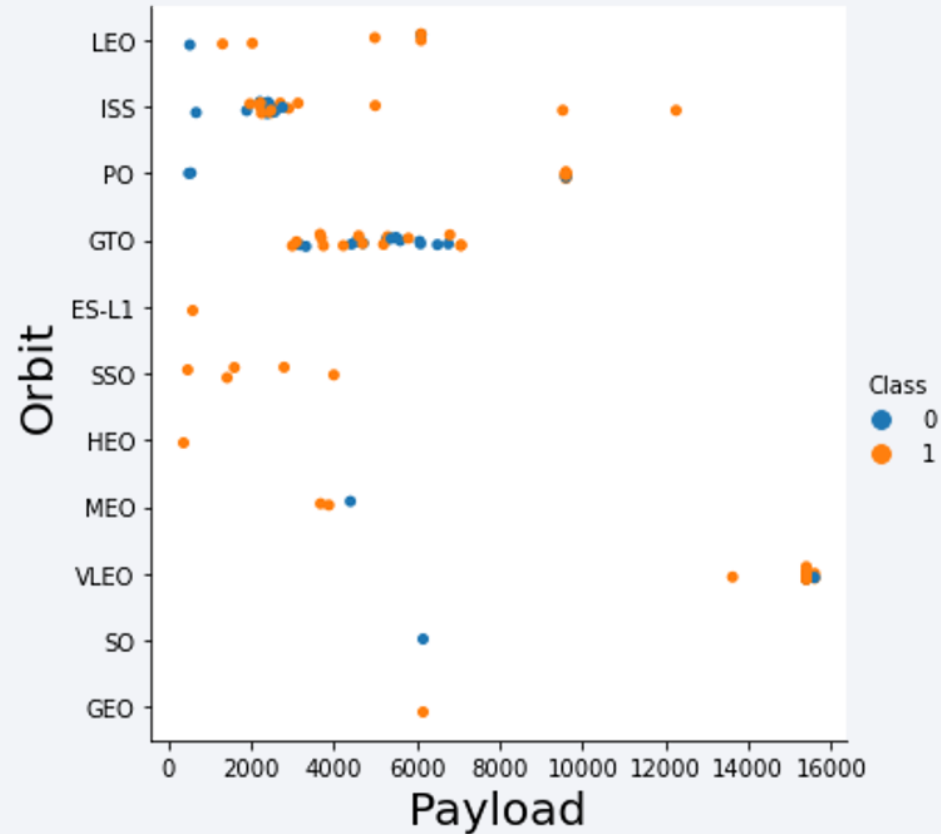
- The orbit types of ES-L1, GEO, HEO, SSO are among the highest success rate.

Flight Number vs. Orbit Type



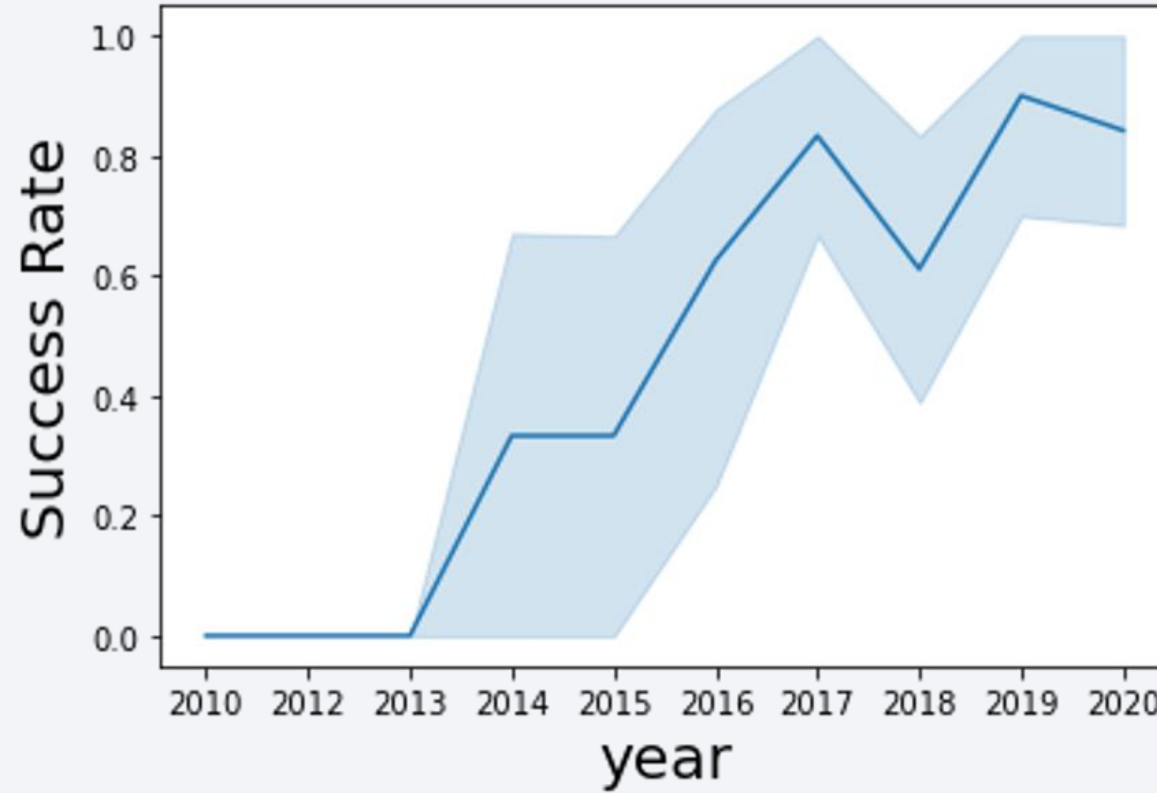
- A trend can be observed of shifting to VLEO launches in recent years.

Payload vs. Orbit Type



- There are strong correlation between ISS and Payload at the range around 2000, as well as between GTO and the range of 4000-8000.

Launch Success Yearly Trend



- Launch success rate has increased significantly since 2013 and has stabilised since 2019, potentially due to advance in technology and lessons learned.

All Launch Site Names

- %sql select distinct(LAUNCH_SITE) from SPACEXTBL

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

Launch Site Names Begin with 'CCA'

- `%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5`

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

- •%sql select sum(PAYLOAD_MASS KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'

45596

Average Payload Mass by F9 v1.1

- •%sql select avg(PAYLOAD_MASS KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'

2928.400000

First Successful Ground Landing Date

- %sql select min(DATE) from SPACEXTBL where Landing Outcome = 'Success (ground pad)'

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql select BOOSTER_VERSION from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS_KG_ > 4000 and PAYLOAD_MASS_KG_ < 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 count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'



100

Boosters Carried Maximum Payload

- %sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS_KG_ = (select max(PAYLOAD_MASS_KG_) from SPACEXTBL)

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 * from SPACEXTBL where Landing Outcome like 'Success%' and (DATE between '2015-01-01' and '2015-12-31') order by date desc

time_utc	booster_version	launch_site	payload	payload_mass_kg	orbit	customer	mission_outcome	landing_outcome
14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
05:24:00	F9 FT B1022.1	CCAFS LC-40	JCSAT-14	4700	GTO	SKY Perfect JSAT	Success	Success (drone ship)

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

- %sql select * from SPACEXTBL where Landing_Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc

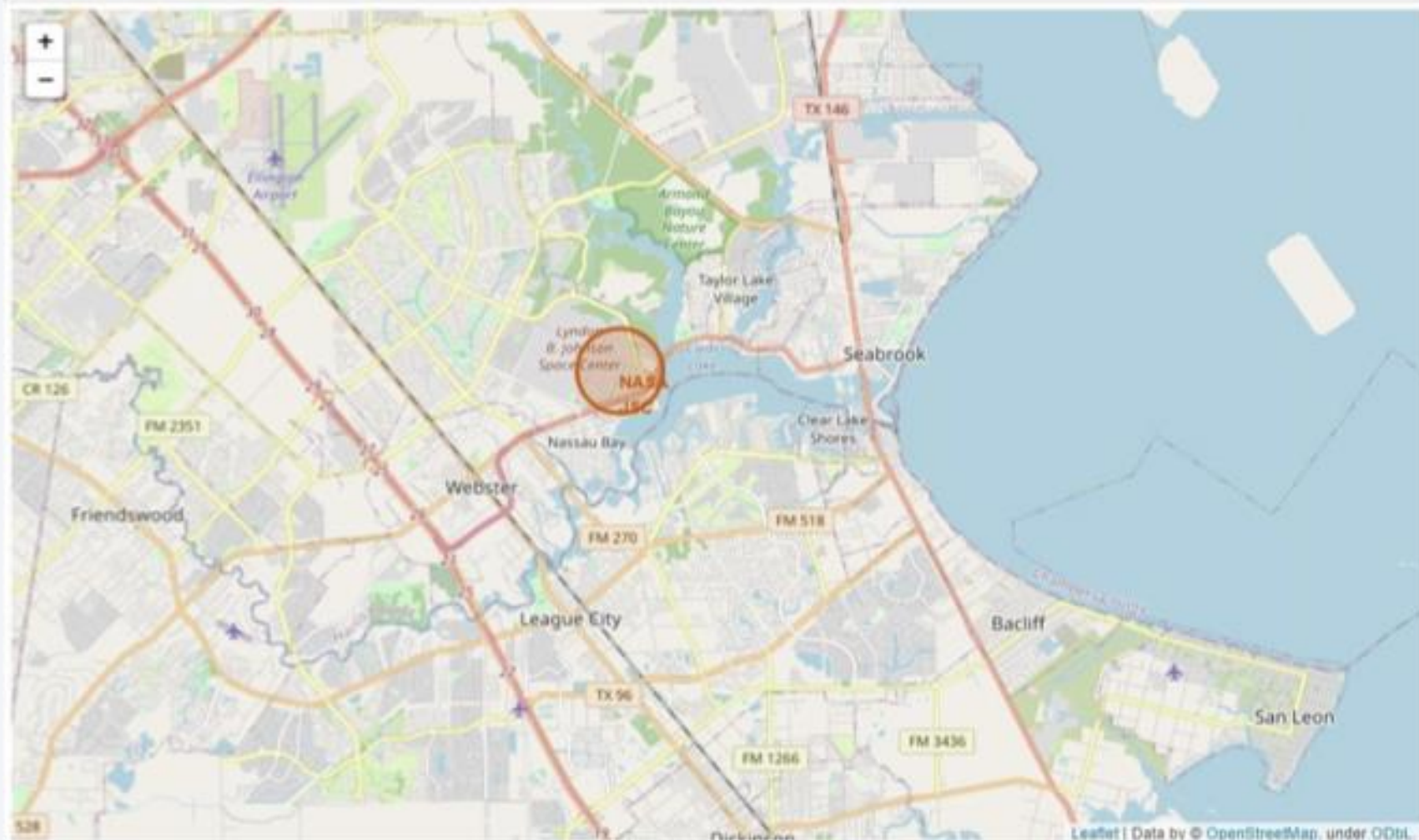
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

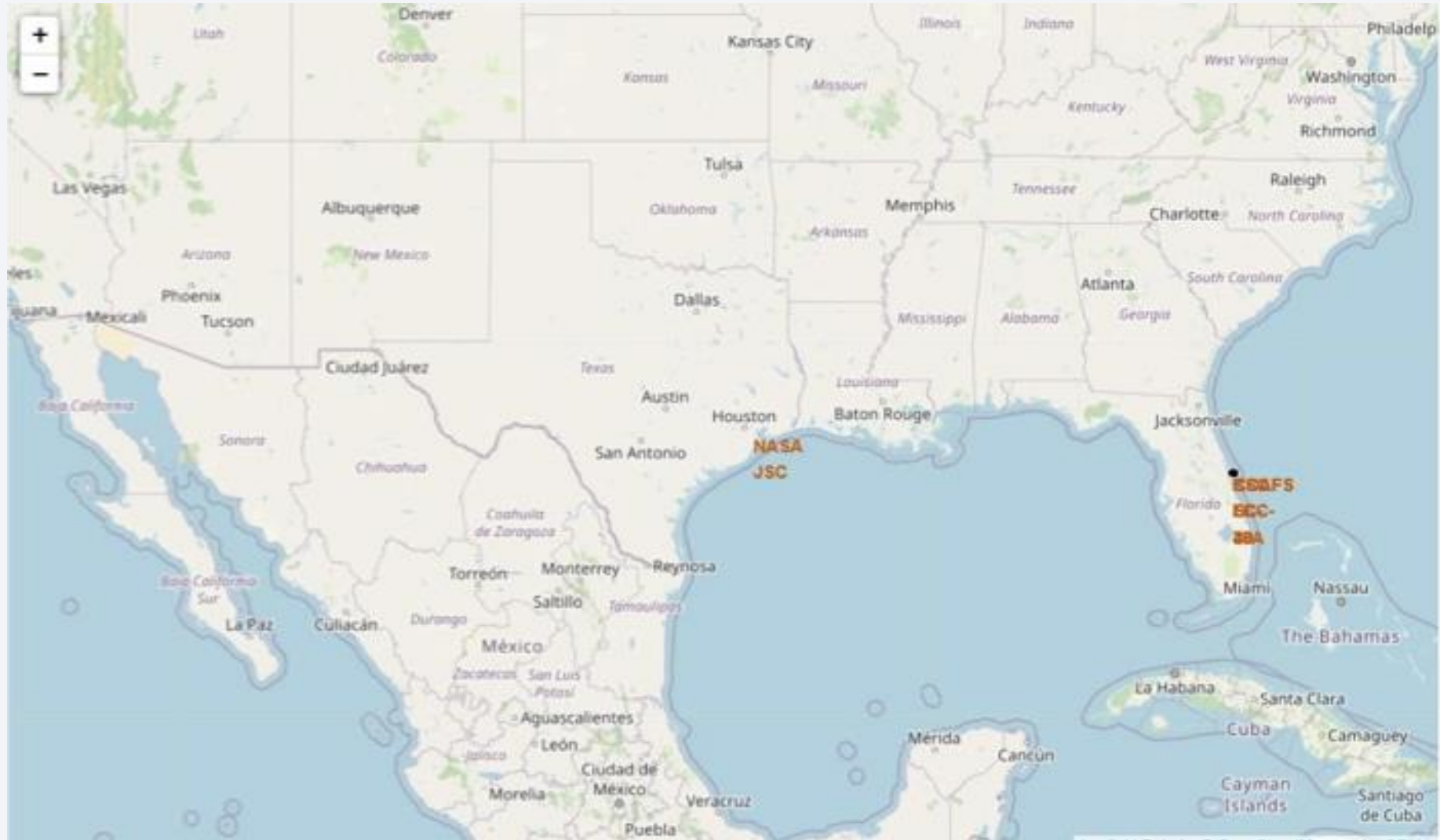
Section 3

Launch Sites Proximities Analysis

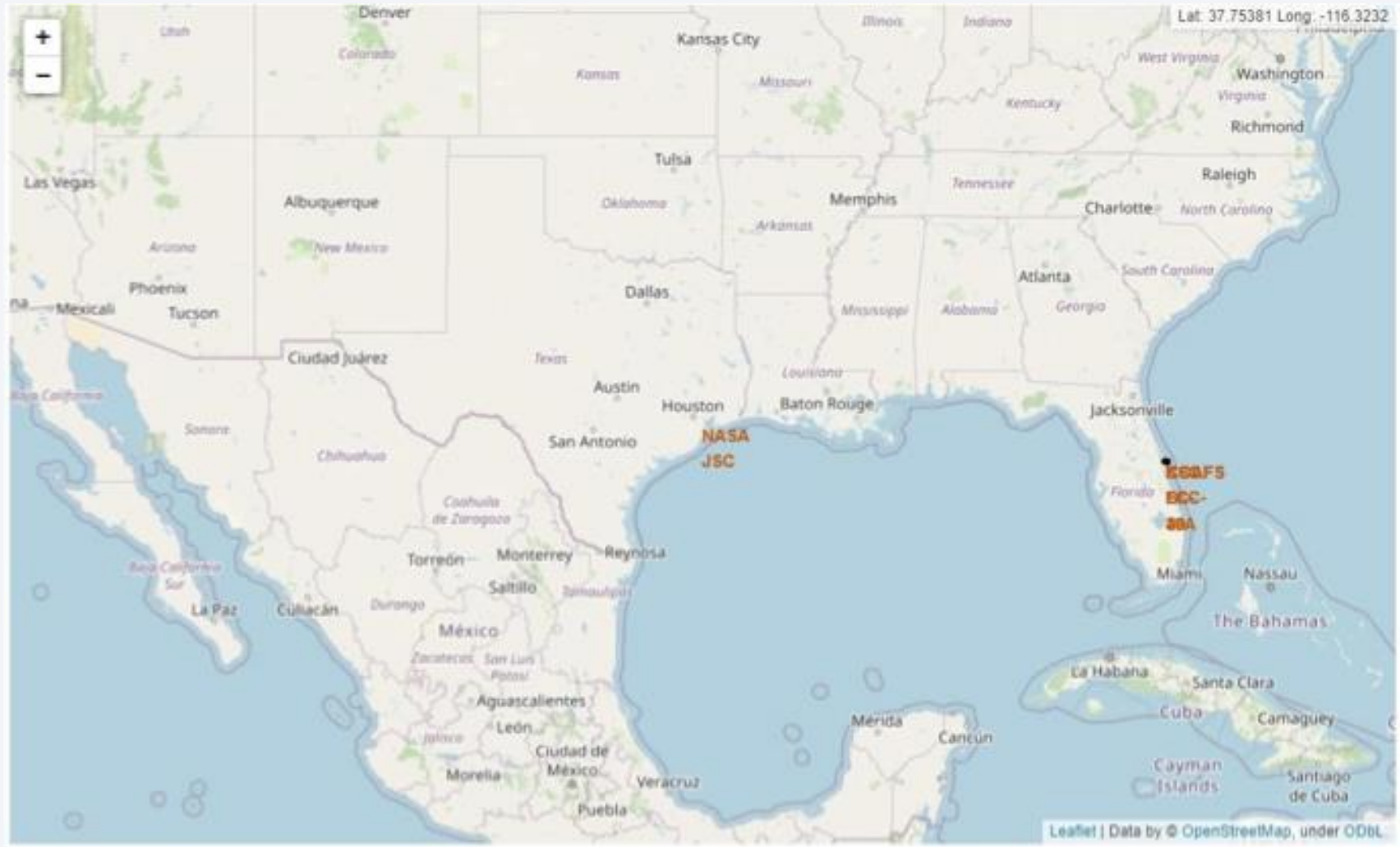
All launch sites marked on a map



Success/failed launches marked on the map



Distances between a launch site to its proximities



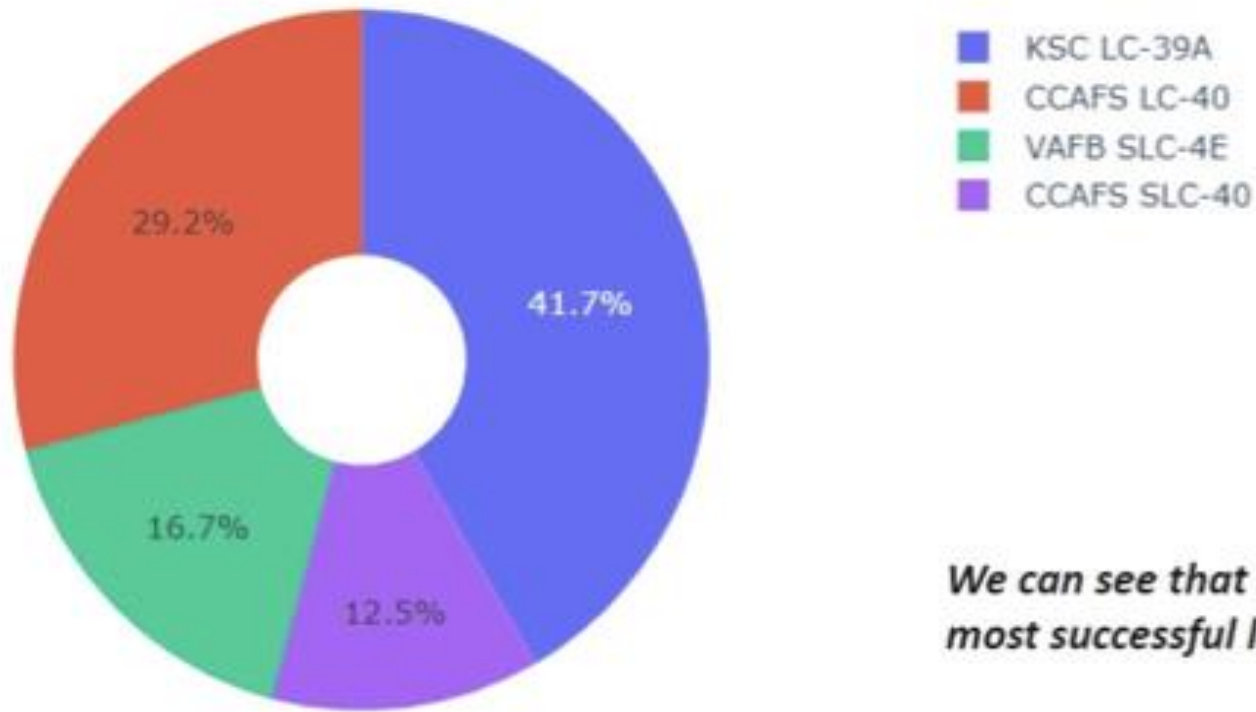


Section 4

Build a Dashboard with Plotly Dash

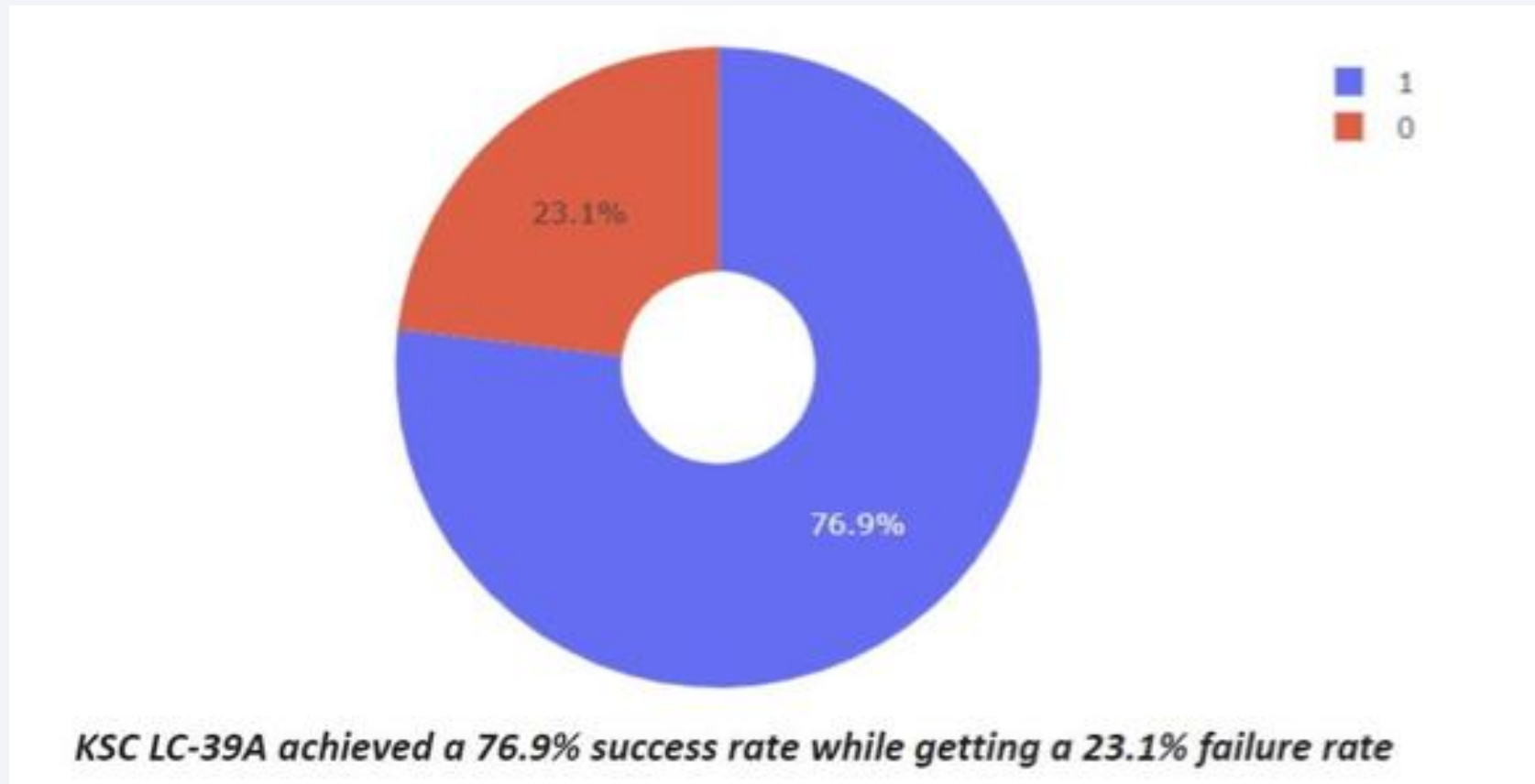
Total success launches by all sites

Total Success Launches By all sites

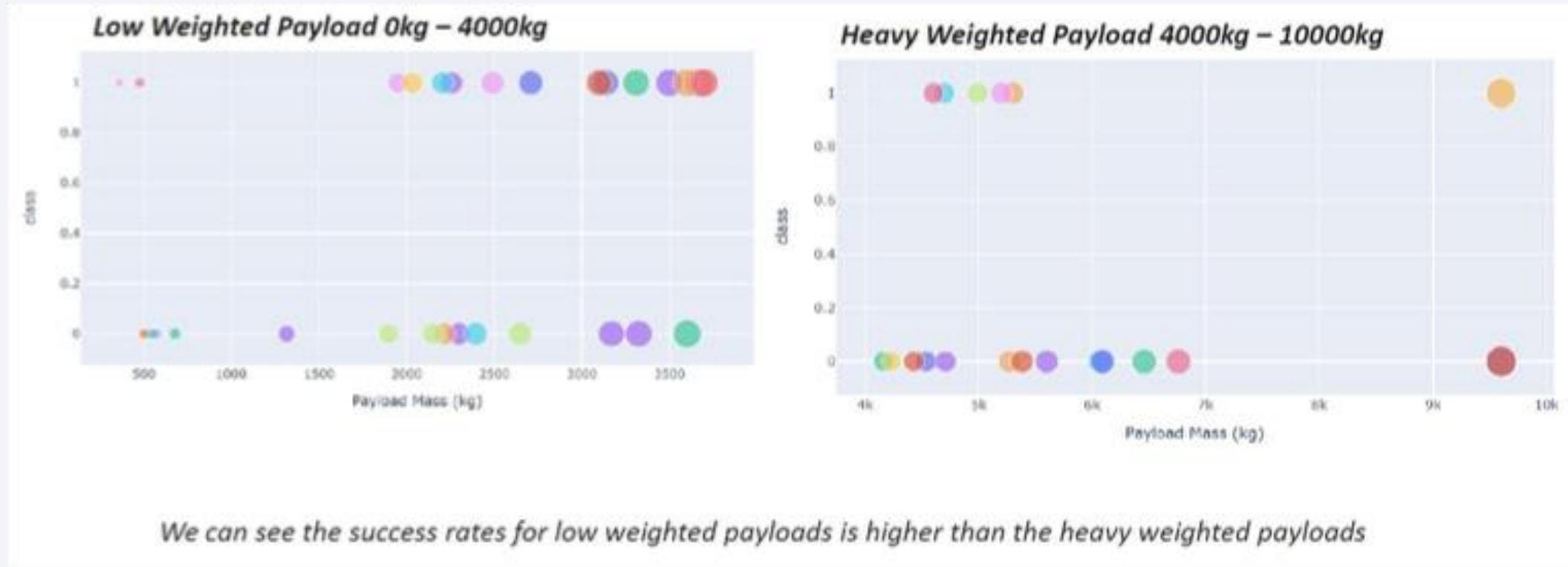


We can see that KSC LC-39A had the most successful launches from all the sites

Success rate by site



Payload vs launch outcome

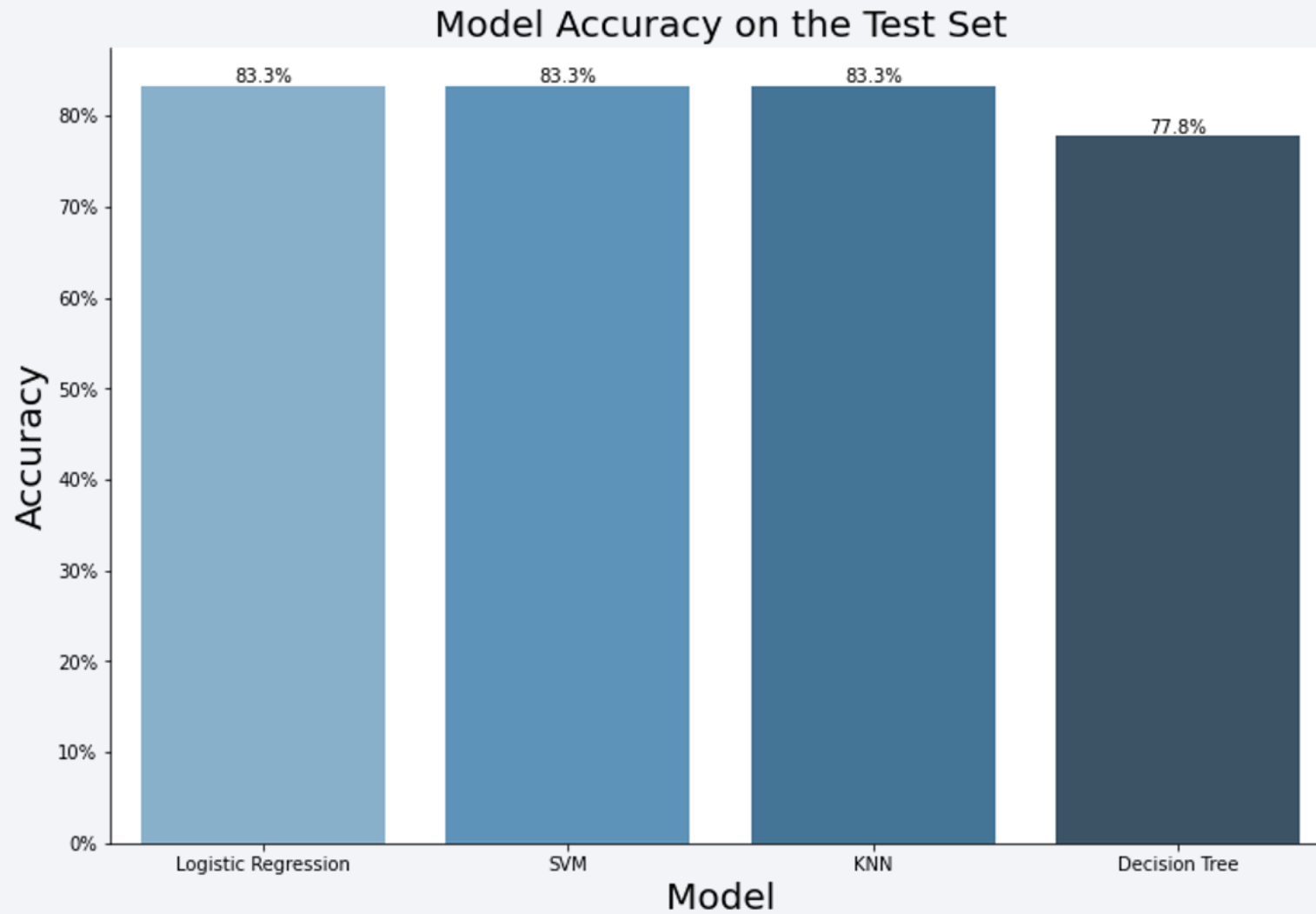




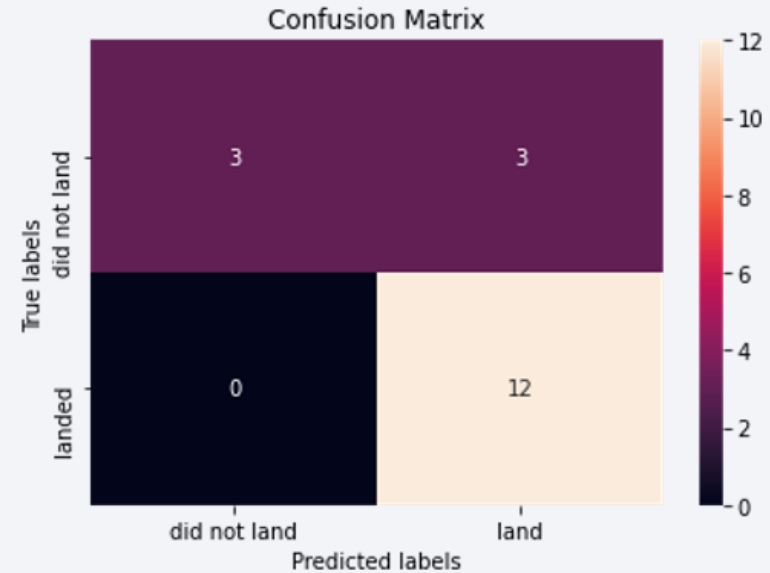
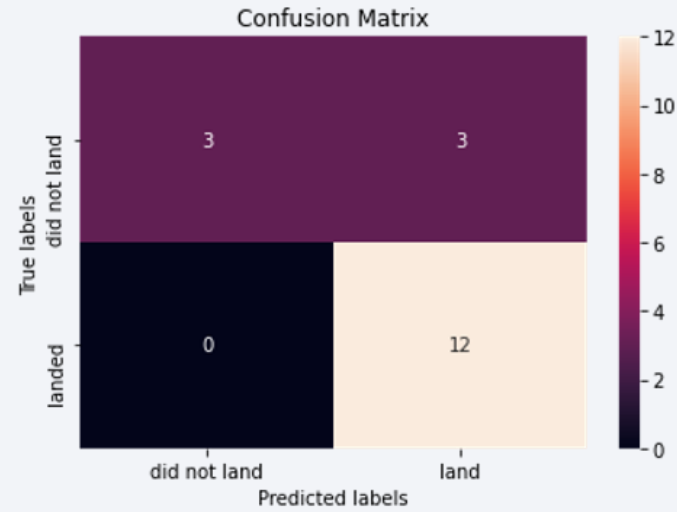
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Conclusions

- The SVM, KNN, and Logistic Regression models demonstrate superior prediction accuracy for this dataset.
- Lighter weighted payloads exhibit higher performance compared to heavier payloads.
- SpaceX's success rates increase proportionally with the passage of time, indicating an improving trend in launch capabilities over the years.
- KSC LC 39A stands out as the launch site with the highest success rate.
- Orbits GEO, HEO, SSO, and ES L1 display the highest success rates among all orbital paths.

Thank you!

