Data science capstone project

Tanish Gupta

07/14/2024

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

- **Methodologies: -**
  - **Data Collection**
  - **Data Preparation**
  - **Exploratory Data Analysis (EDA)**
  - **Interactive Visual Analytics**
  - **Machine Learning Prediction**
  - **Results Compilation**
- **Key Findings and Model Performance:**
  - Developed four machine learning models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K-Nearest Neighbors, achieving an average accuracy rate of approximately 83.33%.
  - All models consistently over-predicted successful landings, indicating a potential bias towards positive outcomes.
  - The analysis revealed that while current models perform adequately, acquiring more diverse data could further enhance model accuracy and reliability.

Introduction

Project background:-

- SpaceX, leveraging the reusability of Falcon 9's first stage, offers rocket launches at $62 million—significantly cheaper than competitors' rates of $165 million. The capability to predict the successful landing of this first stage can directly influence cost estimations for launches. This project aims to develop a machine learning pipeline to forecast the likelihood of the first stage landing successfully, providing crucial competitive insights for companies looking to challenge SpaceX in the market

Problems:-

- What are the critical factors that influence the successful landing of the rocket's first stage?
- How do different variables interact to affect the success rate of the rocket's landing?
- What specific operating conditions are essential to ensure a successful landing program?

Methodology

Methodology

- **Data Collection and Preparation:** Data was sourced from the SpaceX API and supplemented through web scraping from Wikipedia. We conducted thorough data wrangling and applied one-hot encoding to transform categorical variables for analysis.
- **Exploratory and Interactive Analysis:** Utilized SQL for exploratory data analysis and employed Folium and Plotly Dash for interactive visual analytics to derive deeper insights.
- **Predictive Modeling:** Developed and fine-tuned various classification models to predict the success of rocket landings. The process involved building, tuning, and evaluating each model to optimize performance

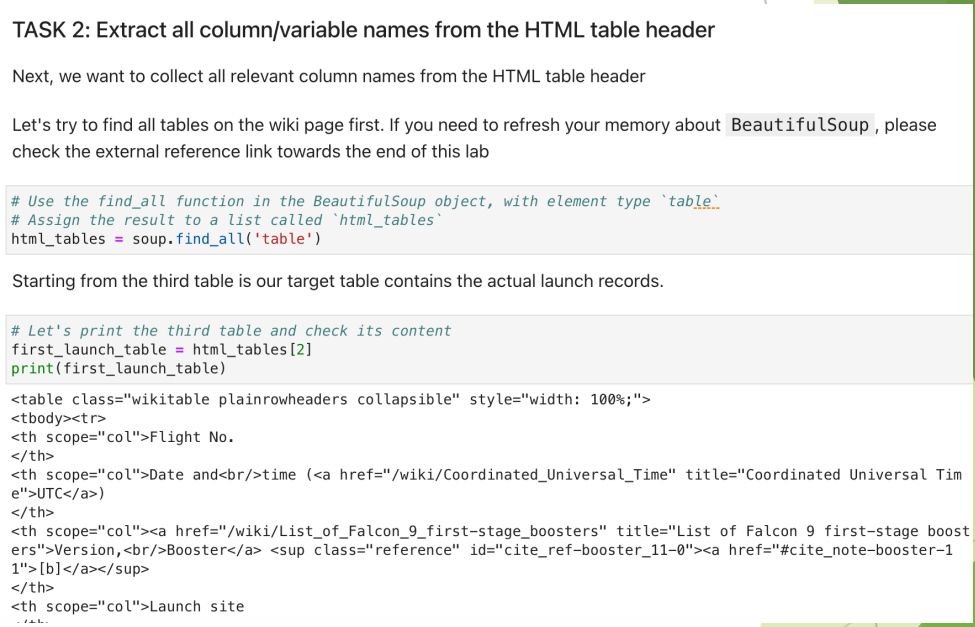
Data Collection

Data was gathered through multiple methods:

- Utilized GET requests to access the SpaceX API.
- The response content was then converted into a JSON format using the .json() function, and subsequently transformed into a pandas dataframe with the help of .json_normalize().
- The dataset underwent cleaning processes, including identifying and addressing missing values.
- Additionally, web scraping was conducted on Wikipedia to obtain Falcon 9 launch records using BeautifulSoup.
- This involved extracting launch records presented in HTML tables, parsing these tables, and converting them into a pandas dataframe for subsequent analysis.

Data Collection - Scraping

- We utilized web scraping to extract Falcon 9 launch records using BeautifulSoup.
- We processed the table and transformed it into a pandas dataframe.



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

Let's try to find all tables on the wiki page first. If you need to refresh your memory about `BeautifulSoup`, please check the external reference link towards the end of this lab

```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables = soup.find_all('table')
```

Starting from the third table is our target table contains the actual launch records.

```
# Let's print the third table and check its content
first_launch_table = html_tables[2]
print(first_launch_table)
```

```
<table class="wikitable plainrowheaders collapsible" style="width: 100%;">
<tbody><tr>
<th scope="col">Flight No.
</th>
<th scope="col">Date and<br/>time (<a href="/wiki/Coordinated_Universal_Time" title="Coordinated Universal Tim
e">UTC</a>)
</th>
<th scope="col"><a href="/wiki/List_of_Falcon_9_first-stage_boosters" title="List of Falcon 9 first-stage boost
ers">Version,<br/>Booster</a> <sup class="reference" id="cite_ref-booster_11-0"><a href="#cite_note-booster-1
1">[b]</a></sup>
</th>
<th scope="col">Launch site
```

Data Wrangling

- We conducted exploratory data analysis and established the training labels. We computed the total launches at each site and analyzed the frequency of each orbit. Additionally, we generated a landing outcome label from the outcome column and exported the results to a CSV file.

TASK 2: Calculate the number and occurrence of each orbit

Use the method `.value_counts()` to determine the number and occurrence of each orbit in the column `Orbit`

```
# Apply value_counts on Orbit column
df['Orbit'].value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

TASK 3: Calculate the number and occurence of mission outcome of the orbits

Use the method `.value_counts()` on the column `Outcome` to determine the number of `landing_outcomes`.Then assign it to a variable landing_outcomes.

```
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
```

`True Ocean` means the mission outcome was successfully landed to a specific region of the ocean while `False Ocean` means the mission outcome was unsuccessfully landed to a specific region of the ocean. `True RTLS`

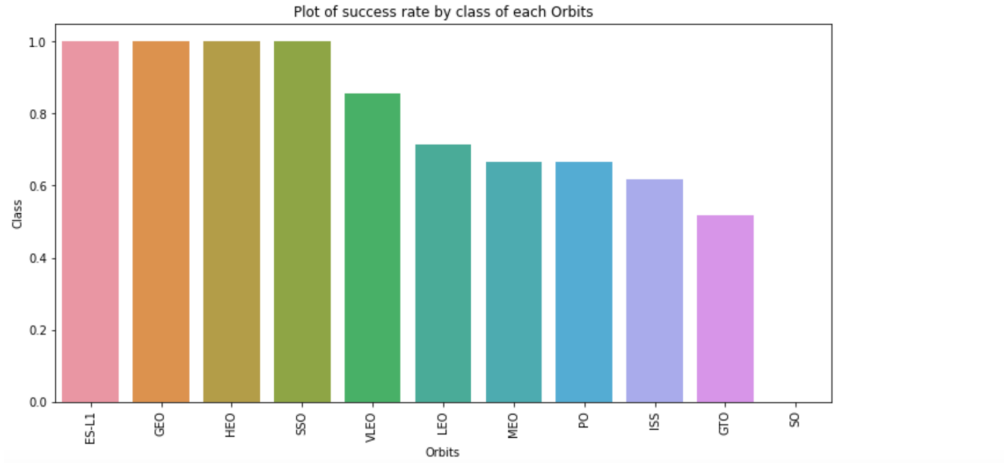
EDA with Data Visualization

Plots Used:

- Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend
- Scatter plots, line charts, and bar plots were used to compare relationships between variables to
- decide if a relationship exists so that they could be used in training the machine learning model
- We analyzed the data by examining the correlations between various parameters. Specifically, we visualized the relationships between flight number and launch site, payload mass and launch site, and success rates across different orbit types. Additionally, we explored connections between flight number and orbit type and tracked the yearly trends in launch success.

File display

```
groupby method on Orbit column and get the mean of Class column
grouped_orbits = df.groupby(by=['Orbit'])['Class'].mean().sort_values(ascending=False).reset_index()
fig, ax=plt.subplots(figsize=(12,6))
ax = sns.barplot(x = 'Orbit', y = 'Class', data=grouped_orbits)
ax.set_title('Plot of success rate by class of each Orbits', fontdict={'size':12})
ax.set_ylabel('Class', fontsize = 10)
ax.set_xlabel('Orbits', fontsize = 10)
ax.set_xticklabels(ax.get_xticklabels(), fontsize = 10, rotation=90);
```



EDA With sql

- Loaded data set into IBM DB2 Database.
- Queried using SQL Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes
- We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:
  - 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 of notebook:- https://github.com/tanishtg/Data-Science/blob/main/Applied_Data_Science_Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We plotted all launch sites on a Folium map, incorporating map objects such as markers, circles, and lines to denote the success or failure of launches at each location.
- We categorized launch outcomes into two classes: 0 for failure and 1 for success.
- By using color-labeled marker clusters, we identified which launch sites exhibited relatively high success rates.
- We also measured the distances from each launch site to nearby features, addressing questions such as: Are launch sites close to railways, highways, and coastlines? and Do launch sites maintain a certain distance from cities?

Build a Dashboard with Plotly Dash

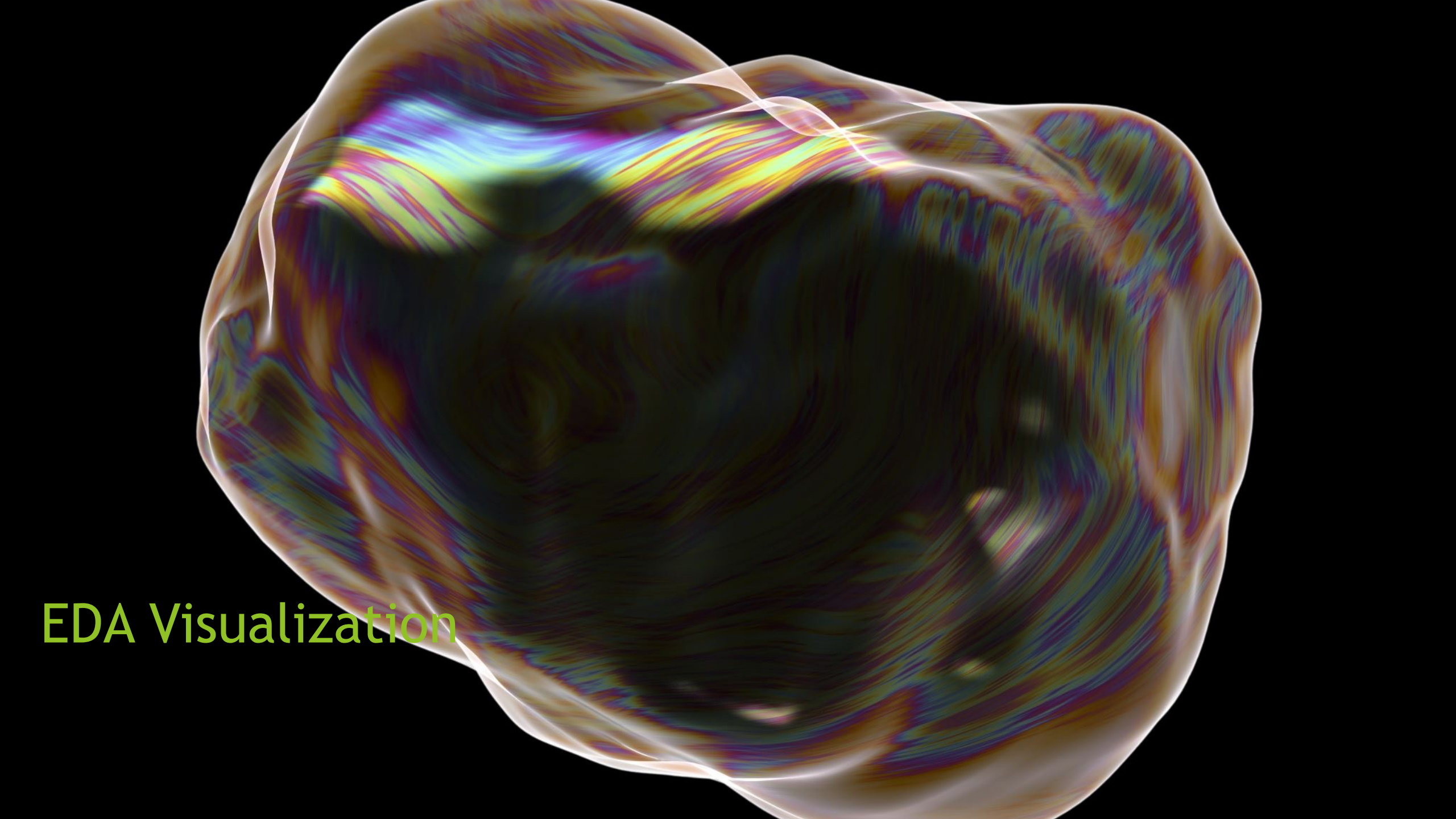
- Dashboard features a pie chart and a scatter plot.
- The pie chart provides options to view the distribution of successful landings both collectively across all launch sites and individually by site.
- The scatter plot allows selections between all sites or specific sites and includes a slider to adjust the payload mass range from 0 to 10,000 kg.
- This pie chart effectively illustrates the success rates at each launch site, while the scatter plot aids in examining how success varies with launch site, payload mass, and booster version.

Predictive Analysis (Classification)

- We imported the data using numpy and pandas, processed the data, and divided it into training and testing sets.
- We constructed various machine learning models and optimized numerous hyperparameters via GridSearchCV.
- Accuracy served as the evaluation metric for our models, which we enhanced through feature engineering and algorithm optimization.
- Ultimately, we identified the highest-performing classification model.

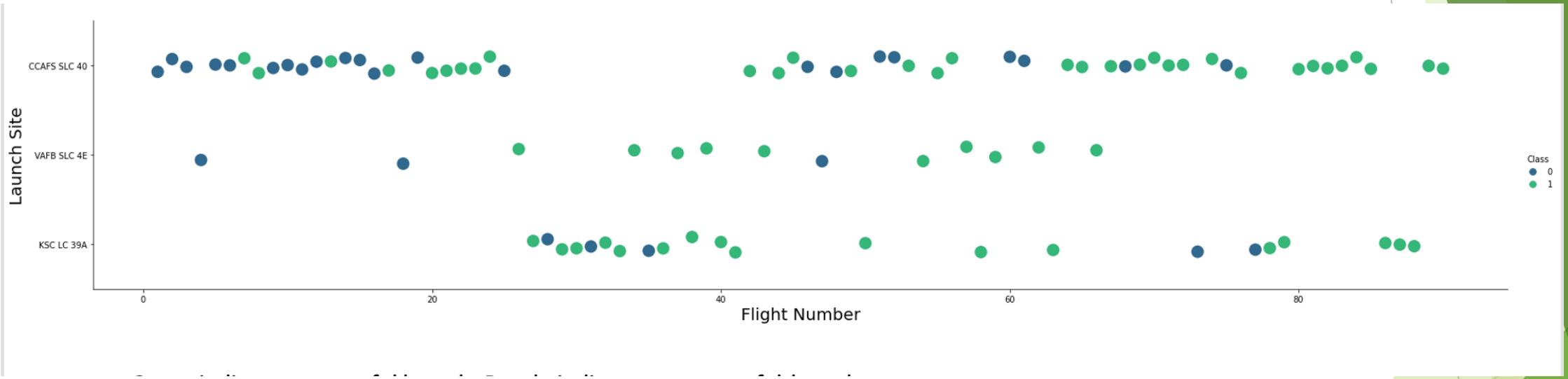
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results
- Preview of the Plotly dashboard.
- The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.



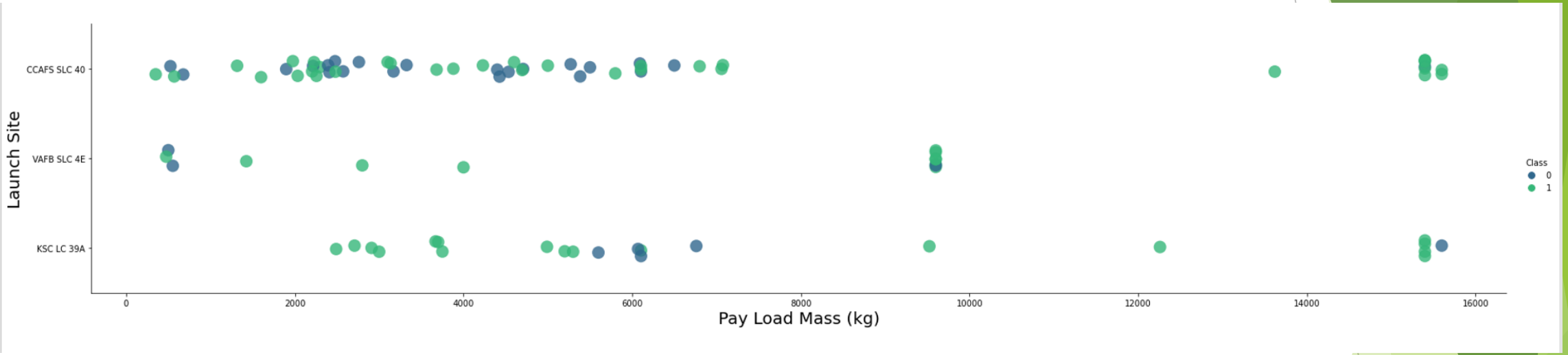
EDA Visualization

Flight Number vs. Launch Site



The Graph revealed that higher flight volumes at a launch site are associated with increased success rates at that site

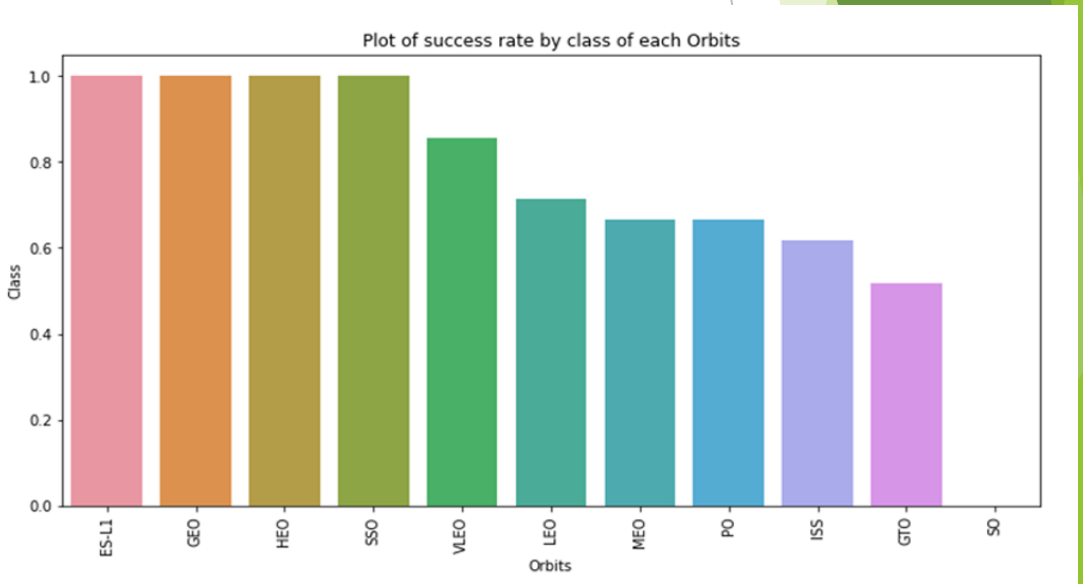
Payload vs. Launch Site



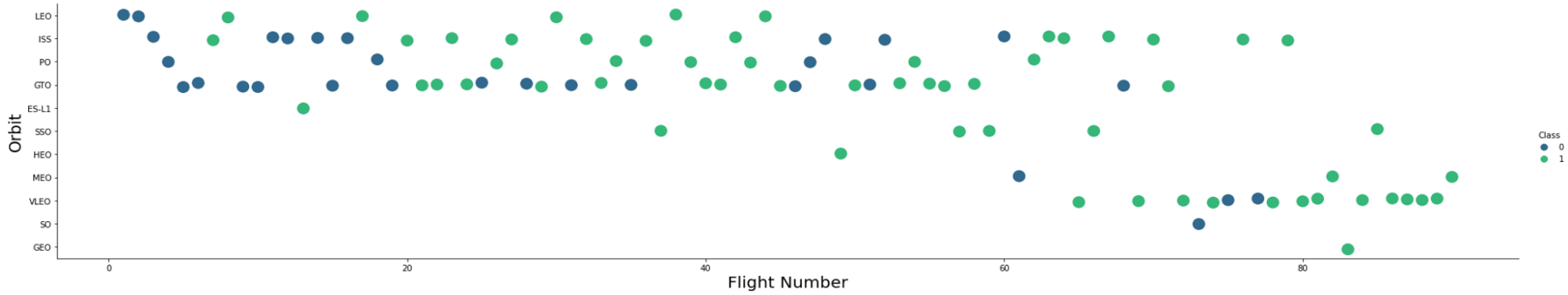
Greater the payload mass for launch site the higher the rate of success

Success Rate vs. Orbit Type

- This graph shows the success rate of satellite launches into different orbits, with ES-L1 and GEO orbits having the highest success rates. The graph can help in understanding which orbits have historically been more reliable for successful missions

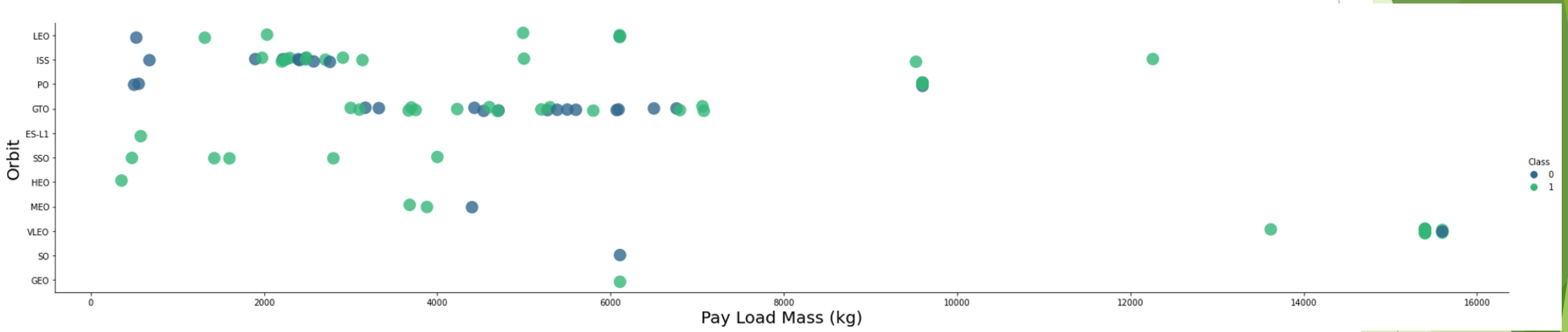


Flight Number vs. Orbit Type



- The scatter plot displays SpaceX flight numbers against various orbit types, color-coded by class—green for success (1) and blue for failure (0). It shows a general trend of increasing success rates as flight numbers progress, particularly noticeable in the LEO and ISS orbits where early flights exhibit more failures and recent flights are predominantly successful

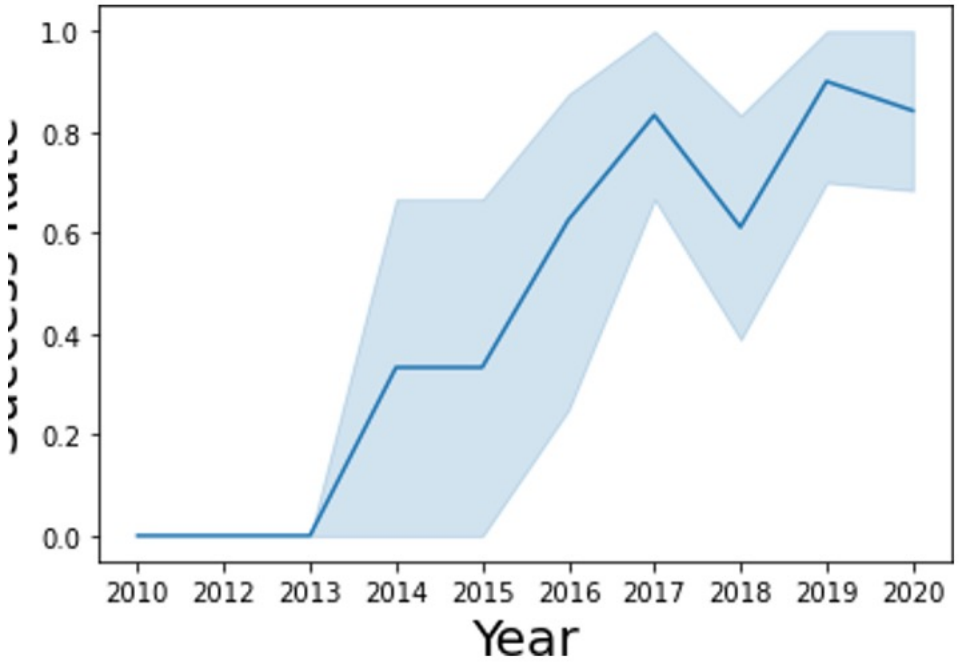
Payload vs orbit type



The scatter plot correlates payload mass with orbit types for SpaceX launches, showing a trend where higher payload masses generally have higher success rates (green), especially noticeable in orbits like VLEO and ES-L1.

Launch Success Yearly Trend

- The line graph illustrates a significant upward trend in SpaceX's launch success rate from 2010 to 2019, peaking near 2018 before a slight decline into 2020.



EAD with sql


mirror_mod.mirror_object
operation == "MIRROR_X":
mirror_mod.use_x = True
mirror_mod.use_y = False
mirror_mod.use_z = False
operation == "MIRROR_Y":
mirror_mod.use_x = False
mirror_mod.use_y = True
mirror_mod.use_z = False
operation == "MIRROR_Z":
mirror_mod.use_x = False
mirror_mod.use_y = False
mirror_mod.use_z = True
selection at the end -add
ob.select= 1
er_ob.select=1
ntext.scene.objects.acti
("Selected" + str(modifier
irror_ob.select = 0
bpy.context.selected_ob
ata.objects[one.name].se
int("please select exactly
OPERATOR CLASSES
types.Operator):
X mirror to the selected
ject.mirror_mirror_x"
ror X"

All Launch Site Names

CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same launch site with data entry errors

only 3 unique launch_site values:
CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

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

Launch Site Names Begin with 'CCA'

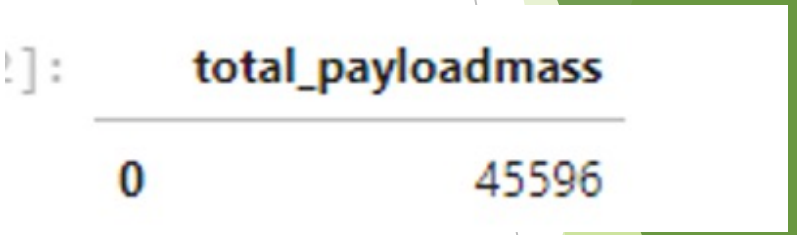
	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

the first five records from the database where the Launch Site name starts with "CCA."

Total payload mass

After running the query for sum of the total payload mass in kg where NASA was the customer.

CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).



	total_payloadmass
0	45596

average payload mass by f9v1.1

After running the query to calculate average payload mass or launches which used booster version f9 v1.1

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

First successful landing date on ground was Dec 22, 2015

	firstsuccessfull_landing_date
0	2015-12-22

Successful drone ship landing with payload between 4000 and 6000

After running the query we got four booster version that successful drone ship landing which has payload mass between 4000 and 6000.

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total number of successful and failure mission outcomes

- SpaceX successfully completes its missions nearly 99% of the time, suggesting that most landing failures are planned.
- Notably, one launch has an ambiguous payload status, and regrettably, another experienced an in-flight failure.

Boosters Carried Maximum Payload

After running the query we get the booster versions that carried the highest payload (1500 kg)

The version of the booster are similar and all are F9 B5

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

Failed 2015 Drone ship Landing

We run the query which return the landing outcome, booster version, payload mass and launch site and we got 2 value I result.

```
create_pandas_df(task_9, database=conn)
```

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

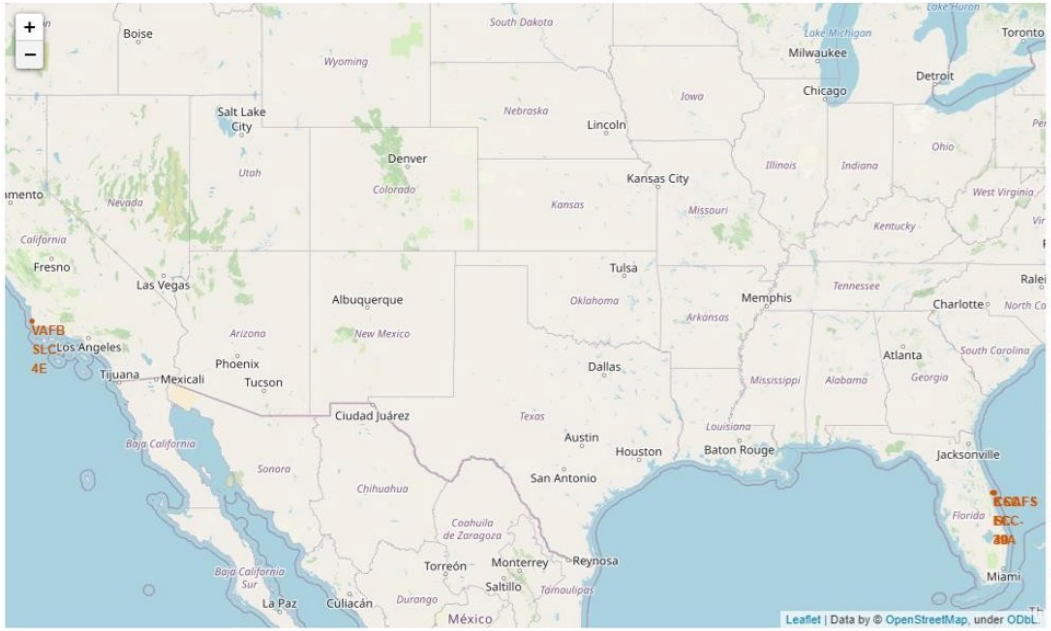
```
create_pandas_df(task_10, database=conn)
```

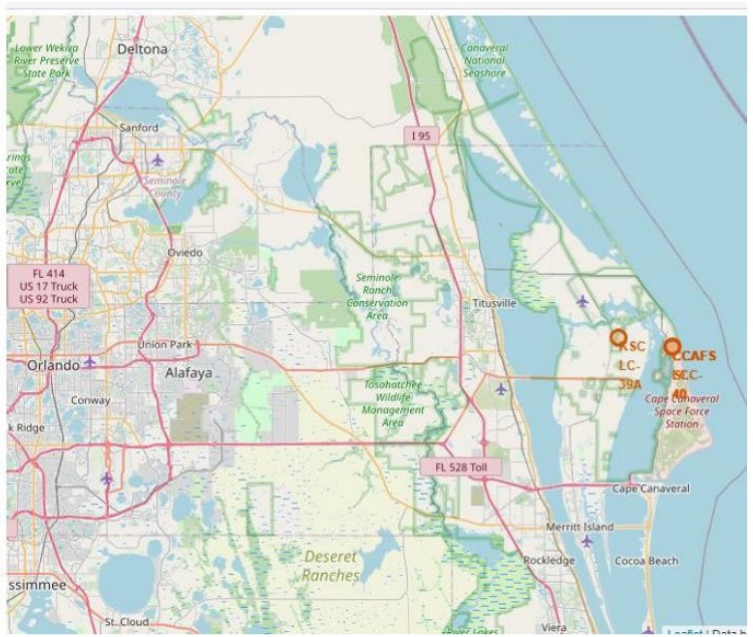
After selecting landing outcomes and their counts, applying a WHERE clause to filter records from March 20, 2010, to June 4, 2010. Then grouped the data by landing outcomes using the GROUP BY clause and sorted the groups in descending order with the ORDER BY clause

	LandingOutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

Launch Sites Analysis

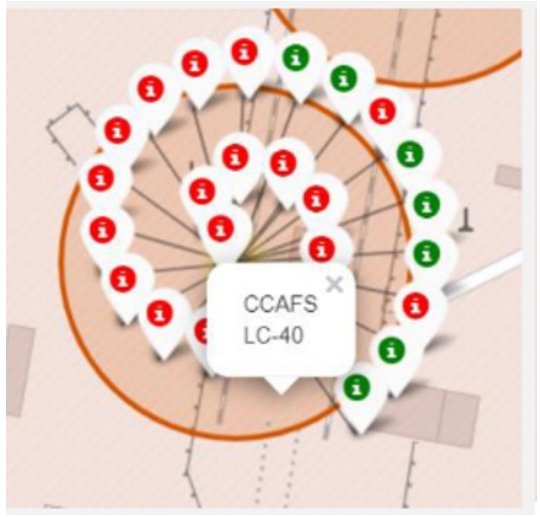
Launch site location

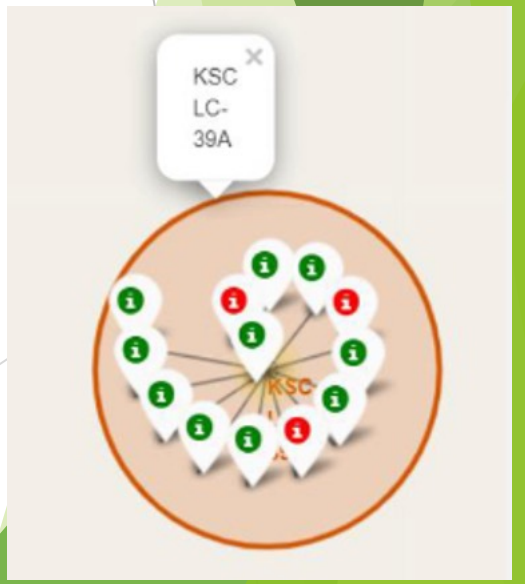




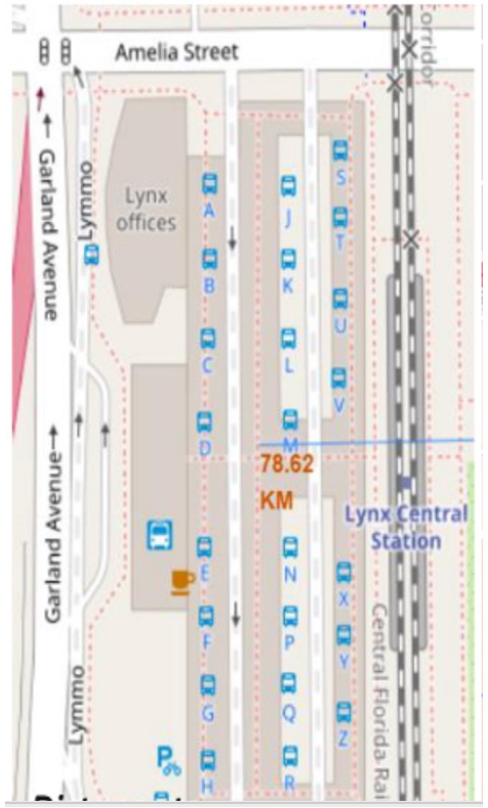
We can see the launch sites. All sites are near ocean

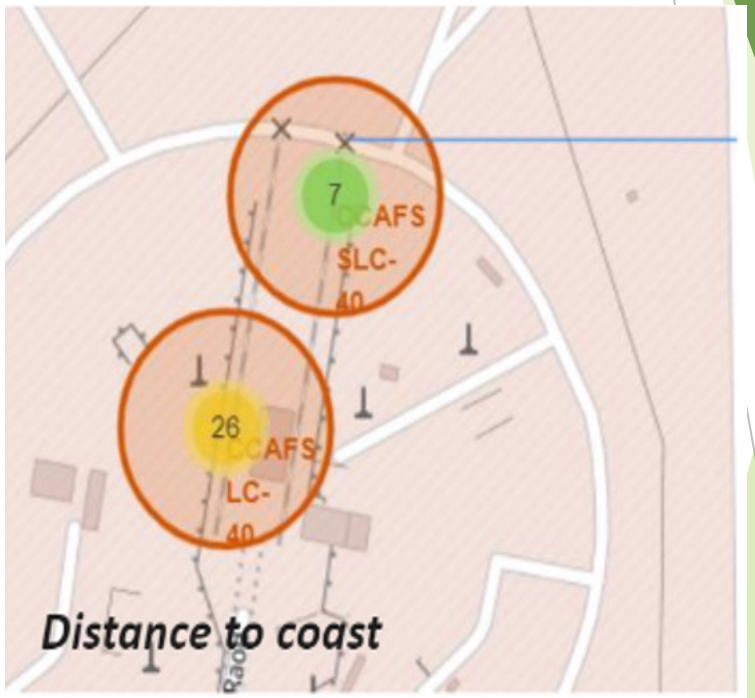
Launch site with color label





Launch site



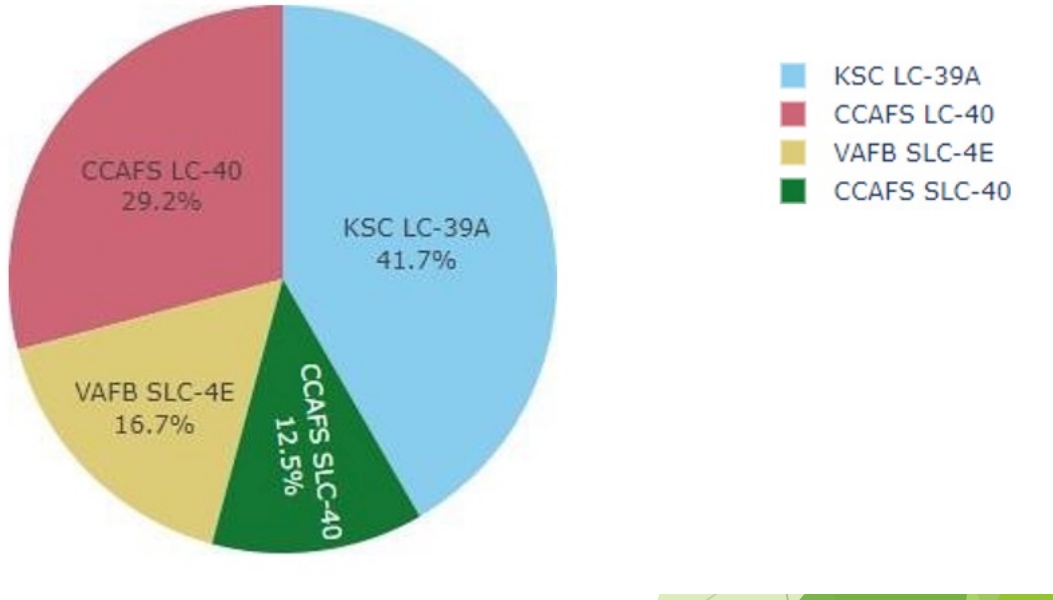


We can see the lunch sites are near coastline

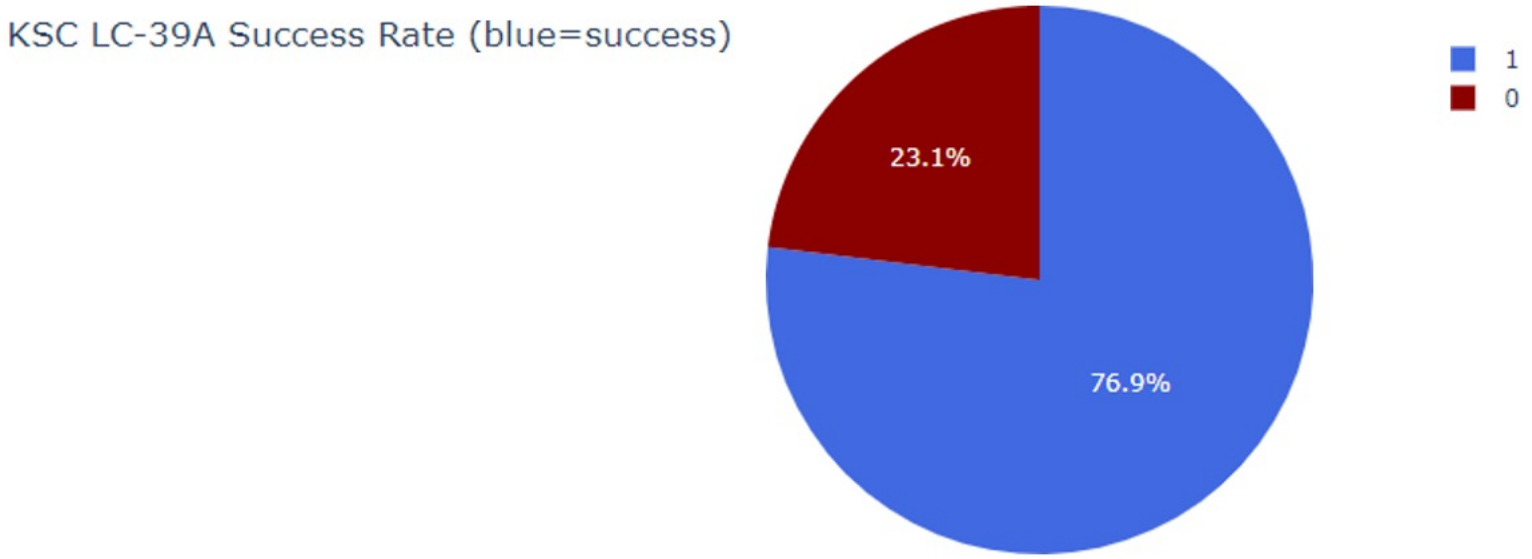

Build dashboard with
ploty dash

The chart showing the percentage achieved by each launch site

We can see the KSC LC-39 has most successful launches followed by CCAFS LCC-40 site



Highest success rate launch site



KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings

Payload vs Launch Outcome for all sites, with different payload selected in the range slider

Predictive analysis (classification)

Classification accuracy

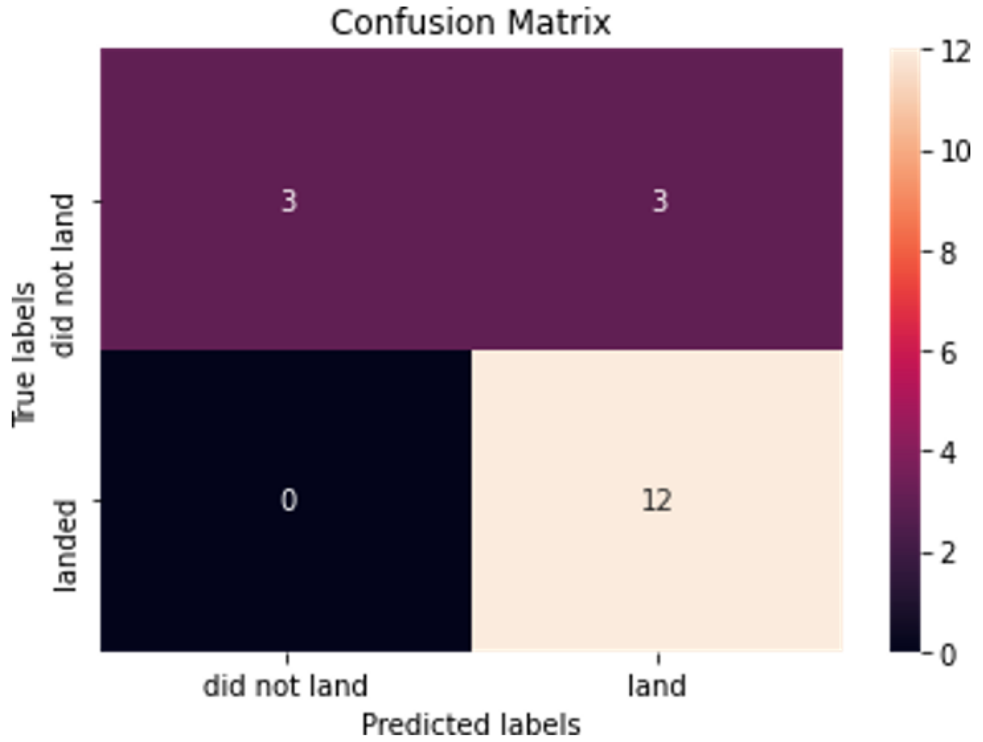
- All models achieved a similar accuracy of 83.33% on the test set, which is notable given the small sample size of only 18.
- This limited size can lead to significant fluctuations in accuracy results, as observed with the Decision Tree Classifier across multiple runs.
- More data is needed to reliably determine the most effective model.

```
print('Accuracy on test data is: {:.3f}'
```

```
Accuracy on test data is: 0.833
```

Confusion matrix

The confusion matrix for the decision tree classifier indicates it distinguishes between classes with a significant issue of false positives, where unsuccessful landings are incorrectly marked as successful. Across all models, the matrix remains consistent, predicting 12 true successful landings, 3 true unsuccessful landings, and 3 unsuccessful landings falsely labeled as successful, suggesting a tendency of the models to over-predict successful outcomes.



conclusion

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Analysis also reveals a direct correlation between the number of flights at a launch site and its success rate, with sites like KSC LC-39A exhibiting the highest number of successful launches.
- The decision tree classifier emerged as the most effective algorithm for this dataset, achieving an accuracy of 83% in predicting the successful landing of Stage 1.
- Certain orbits, specifically ES-L1, GEO, HEO, SSO, and VLEO, consistently show higher success rates, which could guide future launch planning.