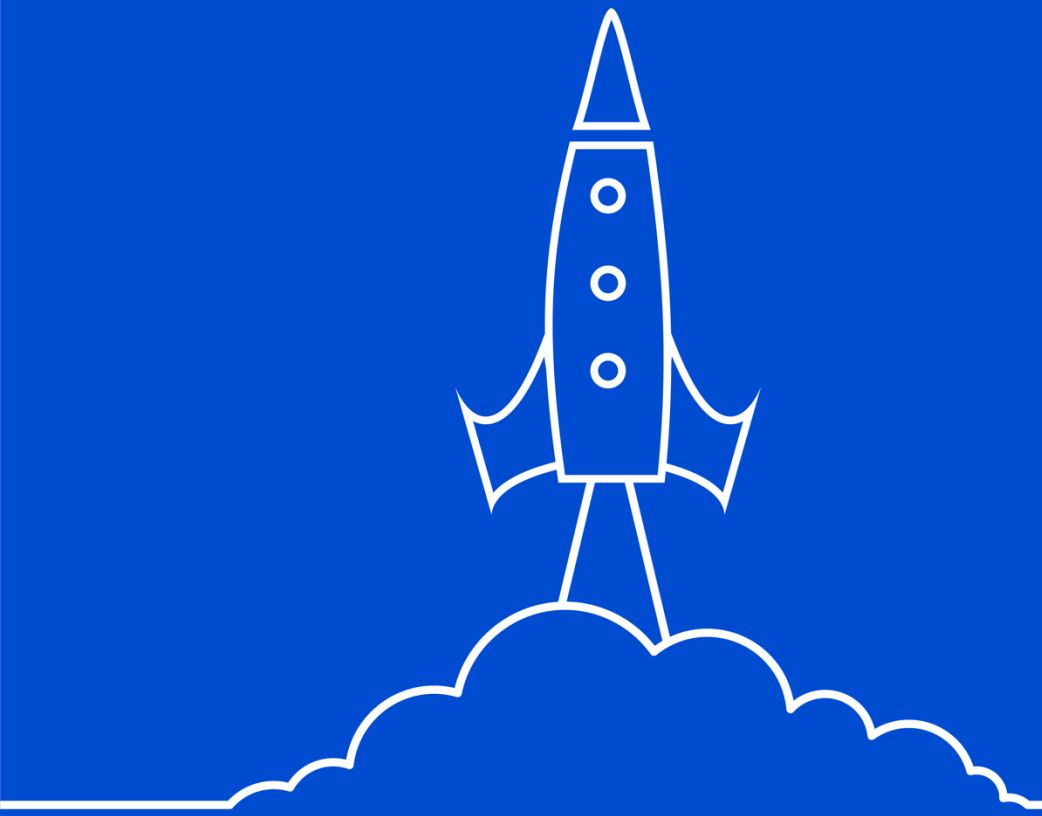


IBM Capstone Project

Mana Nomura
2024-12-16

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

Executive Summary

Summary of Methodology

Data Collection and Preparation

- **API**: SpaceX API data was processed into dataframes, with missing Payload Mass values replaced by the mean.
- **Web Scraping:** Falcon 9 data was extracted from Wikipedia using BeautifulSoup and exported to CSV.
- **Data Wrangling**: Landing outcomes were encoded as binary values (1 = success, 0 = failure).

Exploration and Visualization

- **EDA:** Relationships between variables were analyzed using charts, SQL queries, and insights into payload success rates.
- **Mapping & Dashboards:** Folium maps highlighted launch site proximities, while Plotly Dash visualized success metrics.

Predictive Modeling

- **Approach:** Data was standardized, split into training/testing sets, and optimized using GridSearchCV.
- **Models Tested:** Logistic Regression, SVM, Decision Trees, and KNN, evaluated using metrics like Jaccard, F1, and Accuracy.

Summary of Results

Success rates for rocket launches have improved significantly over time, with the KSC LC-39A site achieving the highest success rate among launch sites. Certain orbits, such as ES-L1, GEO, HEO, and SSO, consistently achieve 100% success. Geographically, launch sites are strategically located near the equator to leverage Earth's rotational speed for cost efficiency and are all coastal to simplify rocket retrieval. In predictive modeling, all algorithms performed similarly, but the decision tree model showed slightly better performance.

Introduction

Introduction

- **Overview**

- The commercial space industry is thriving, with companies like SpaceX, Virgin Galactic, and Blue Origin making space travel more accessible. SpaceX stands out for its cost-effective Falcon 9 rocket, which reduces launch expenses to $62M through first-stage reusability. This stage, responsible for most of the rocket's work, is critical but sometimes unrecoverable due to mission requirements. In this capstone, we will act as a data scientist for Space Y, using machine learning and public data to predict SpaceX's first-stage reuse and estimate launch costs, enabling informed decisions.

- **Problems**

- How do payload mass, launch sites, number of flights, and orbits influence first-stage landing success?
- How have success rates changed over time?
- Which predictive model best classifies successful landings?

Methodology

Methodology

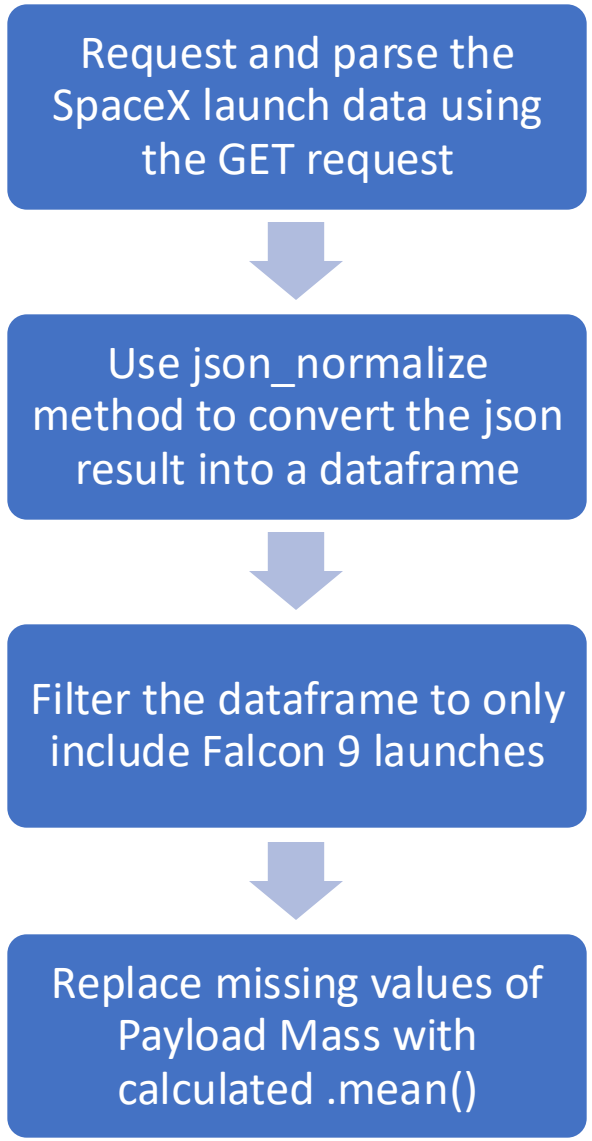
Executive Summary

- Data collection methodology:
  - Collect data using API and Web Scaping
- Perform data wrangling
  - Filter, and handle missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune, evaluate classification models using GridSearchCV

Data Collection – SpaceX API

- **Request to the SpaceX API**
  - Request rocket launch data from SpaceX API with the URL
  - Decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()
  - Apply getBoosterVersion function method to get the booster version
- **Clean the requested data**
  - Filter the data dataframe using the BoosterVersion column to only keep the Falcon 9 launches.
  - Calculate the mean for the PayloadMass using the .mean().
  - Use the mean and the .replace() function to replace np.nan values in the data with the mean

URL
https://github.com/mananomura/IBM_Data_Science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

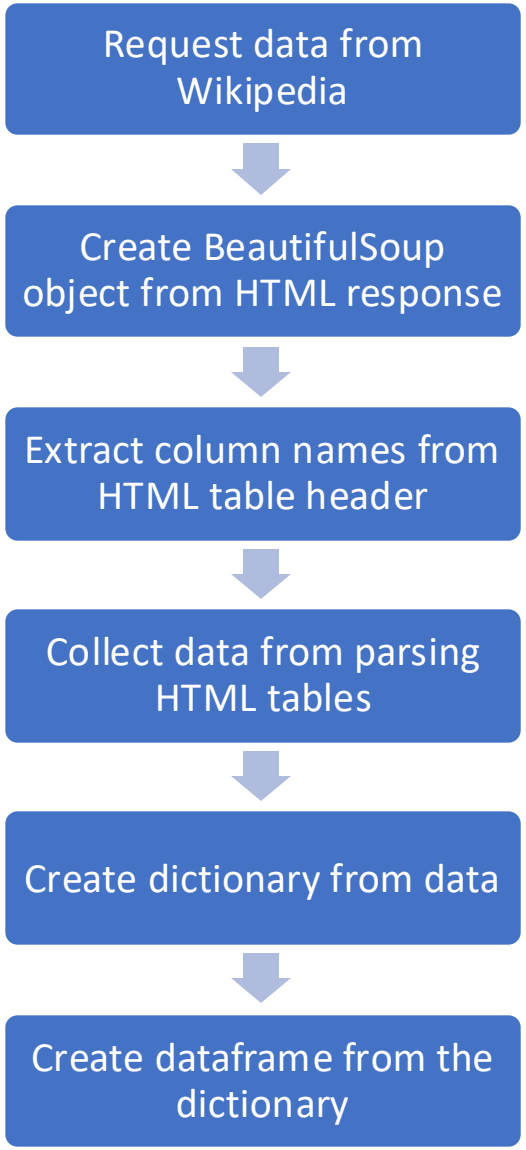


Data Collection – Web Scraping

- **Extract a Falcon 9 launch records HTML table from Wikipedia**
  - HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response
  - Create a BeautifulSoup object from the HTML response
  - Iterate through the <th> elements and apply the provided extract_column_from_header() to extract column name one by one

- **Parse the table and convert it into a Pandas data frame**

URL
https://github.com/mananomura/IBM_Data_Science/blob/main/jupyter-labs-webscraping%20(1).ipynb

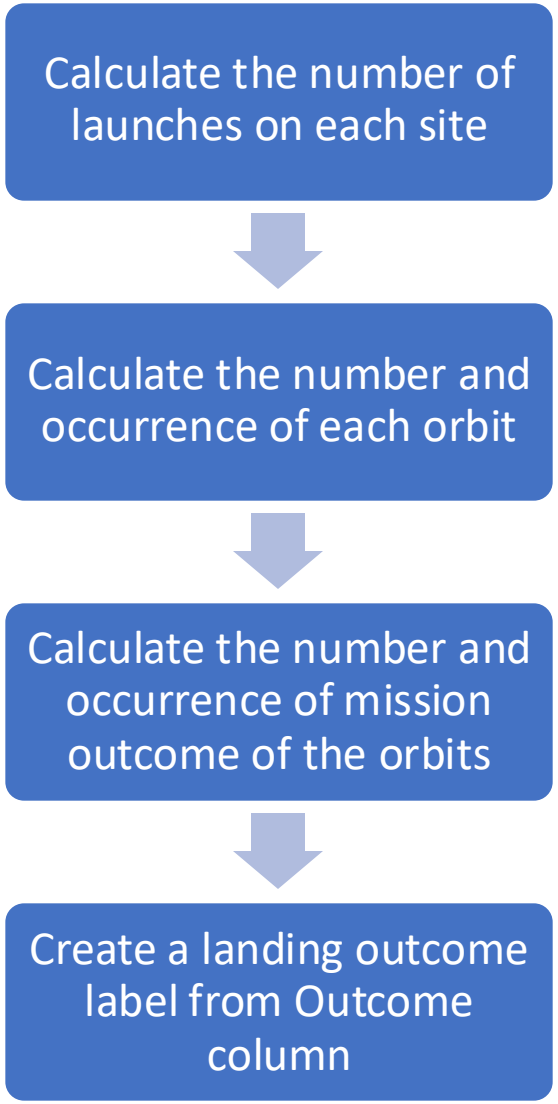


Data Wrangling

- **Exploratory Data Analysis**
  - Use the method value_counts() on the column LaunchSite to determine the number of launches on each site
  - Use the method .value_counts() to determine the number and occurrence of each orbit in the column Orbit
  - Use the method .value_counts() on the column Outcome to determine the number of landing_outcomes.
- **Determine Training Labels**
  - Using the Outcome, create a list where the element is zero if the corresponding row in Outcome is in the set bad_outcome; otherwise, it's one. Then assign it to the variable landing_class

URL

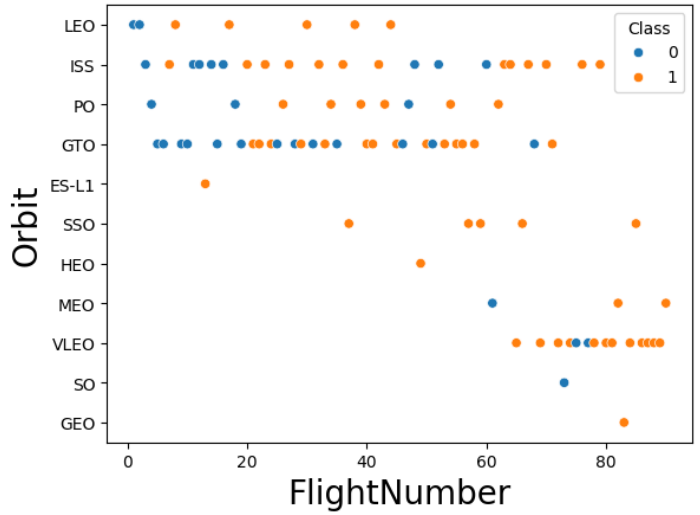
https://github.com/mananomura/IBM_Data_Science/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

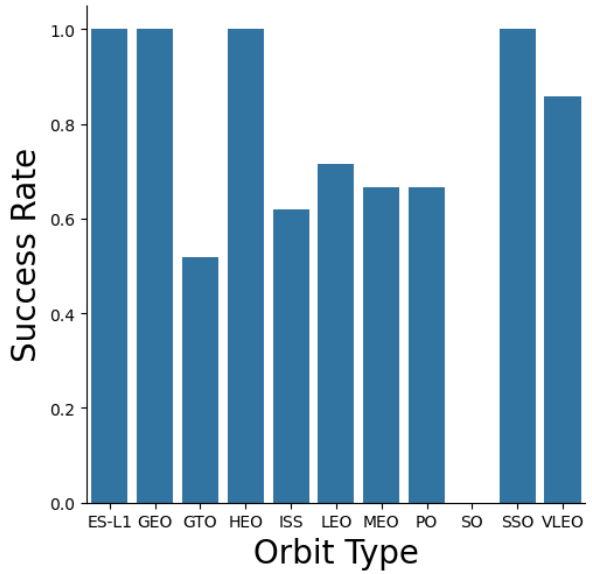
- **Scatter plots**
  - Flight Number vs. Payload
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Payload Mass vs. Orbit Type

  → the variables could be useful for machine learning if a relationship exists



- **Bar Charts**
  - Success Rate vs. Orbit Type

  → show the relationships among the categories and a measured value



URL
https://github.com/mananomura/IBM_Data_Science/blob/main/edadataviz.ipynb

EDA with SQL

Display

- Names of unique launch sites
- 5 records where launch site begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA(CRS)
- Average payload mass carried by booster version F9 v1.1

List

- The date when the first successful landing outcome in ground pad was achieved
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes
- The names of the booster versions which have carried the maximum payload mass.
- 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 between the date 2010-06-04 and 2017-03-20, in descending order.

URL
https://github.com/mananomura/IBM_Data_Science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- Add **blue** circles at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
- Add **red** circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates

Colored Markers of Launch Outcomes

- Add colored markers of successful (**green**) and unsuccessful (**red**) launches at each launch site to show which launch sites have high success rates

Distances Between a Launch Site to Proximities

- Add colored line to show distance between site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city

URL
https://github.com/mananomura/IBM_Data_Science/blob/main/lab_jupyter_launch_site_location%20(3).ipynb

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites
- Allow user to select all launch sites or a specific one

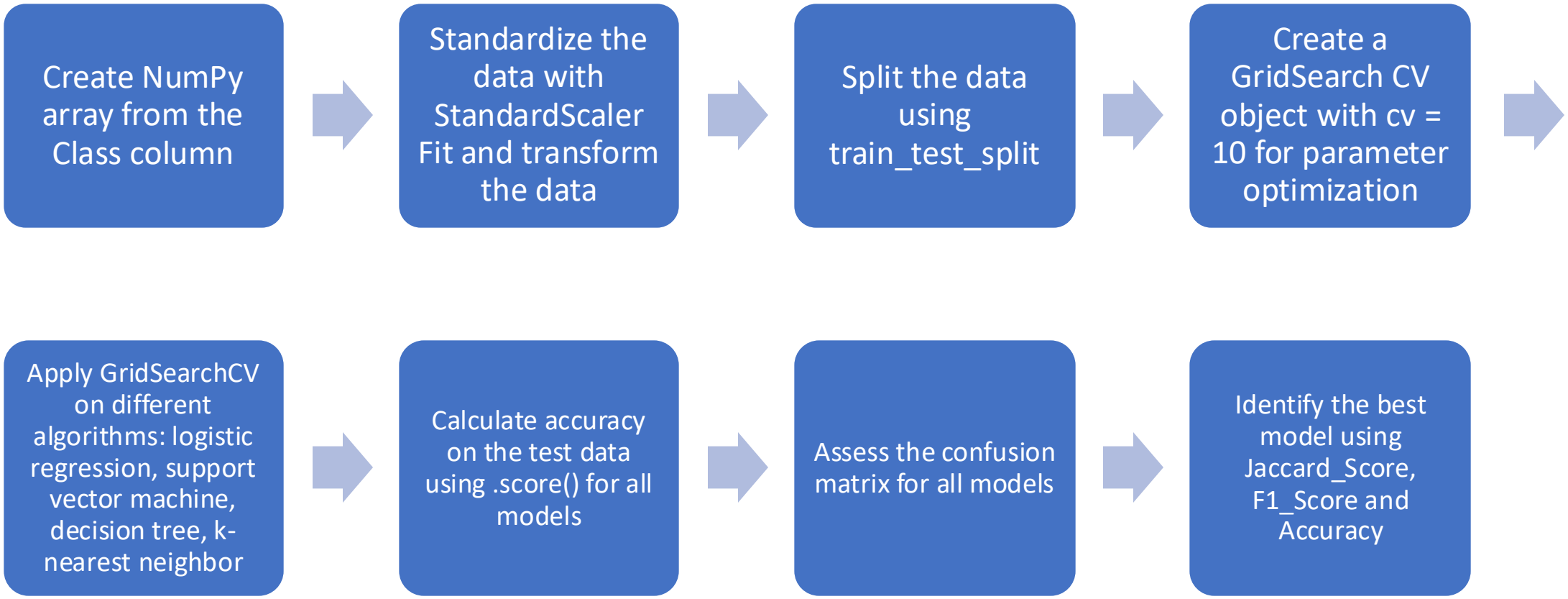
Pie Chart Showing Successful Launches
- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass range
- Allow user to select payload mass range

Scatter Chart Showing Payload Mass vs. Success rate by Booster Version
- Allow user to see the correlation between Payload and launch success

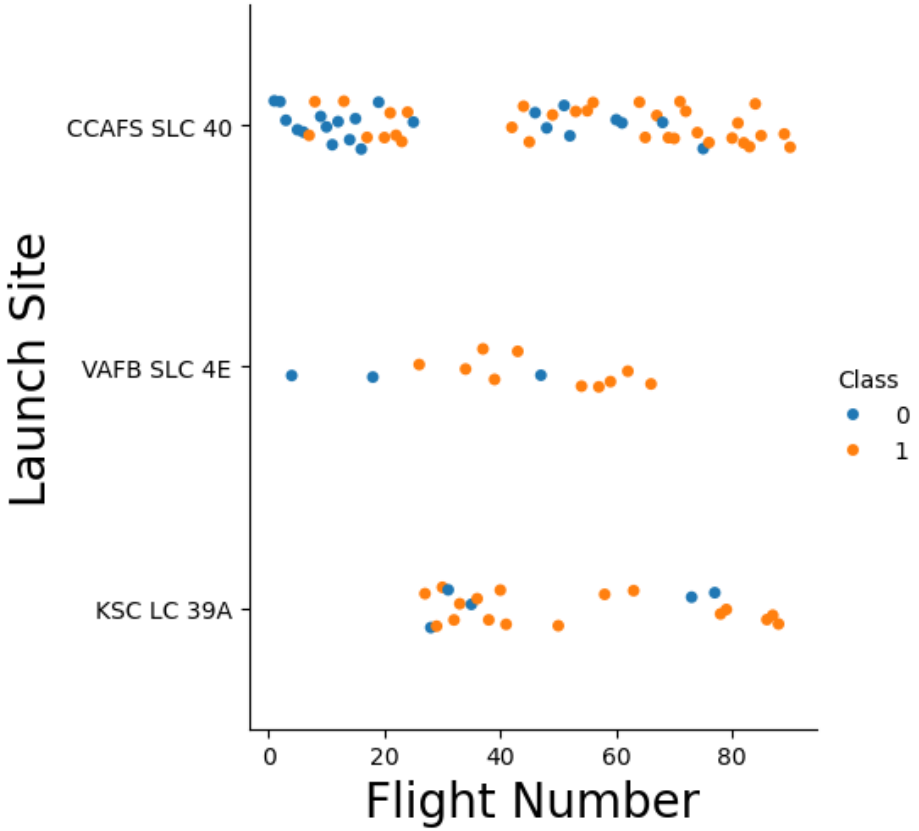
URL
https://github.com/mananomura/IBM_Data_Science/blob/main/plotly.py

Predictive Analysis (Classification)



Results

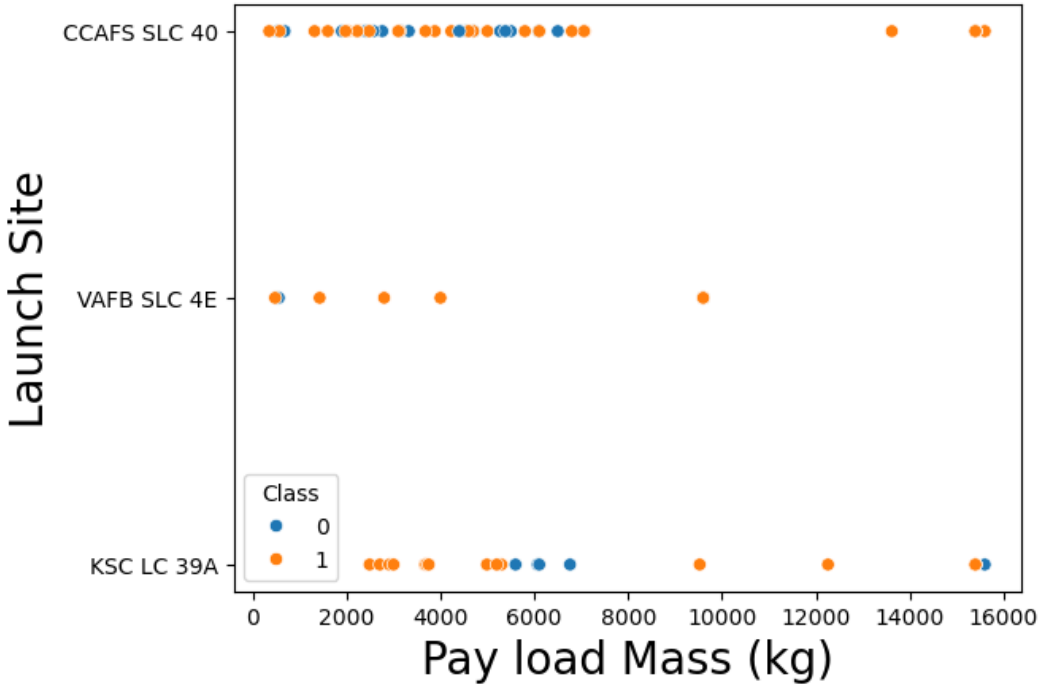
Flight Number vs. Launch Site



Exploratory Data Analysis

- Earlier flights had a lower success rate (blue=fail)
- Later flights had a higher success rate (orange=success)
- Around half of launches were from CCAFS SLC 40
- New launches have a higher success rate

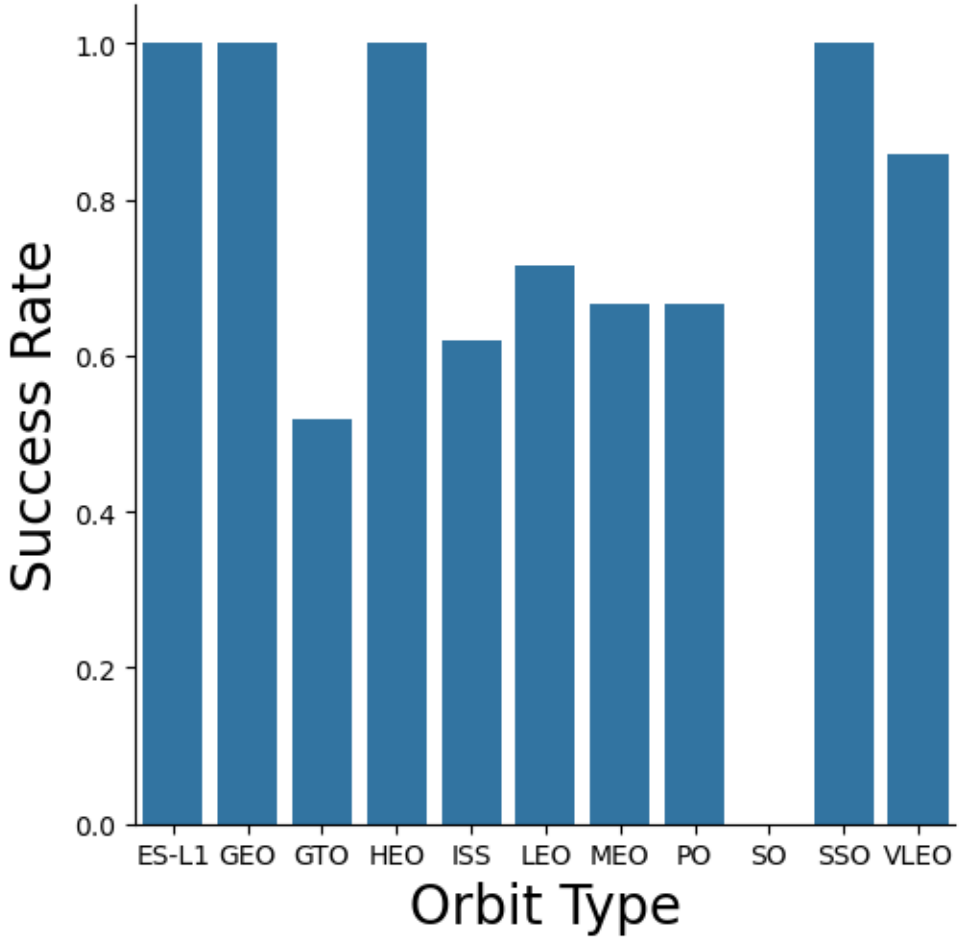
Payload vs. Launch Site



Exploratory Data Analysis

- Higher the payload mass, the higher the success rate
- Most launches with a payload greater than 8000 kg were successful
- KSC LC 39A has a 100% success rate less than 5500 kg

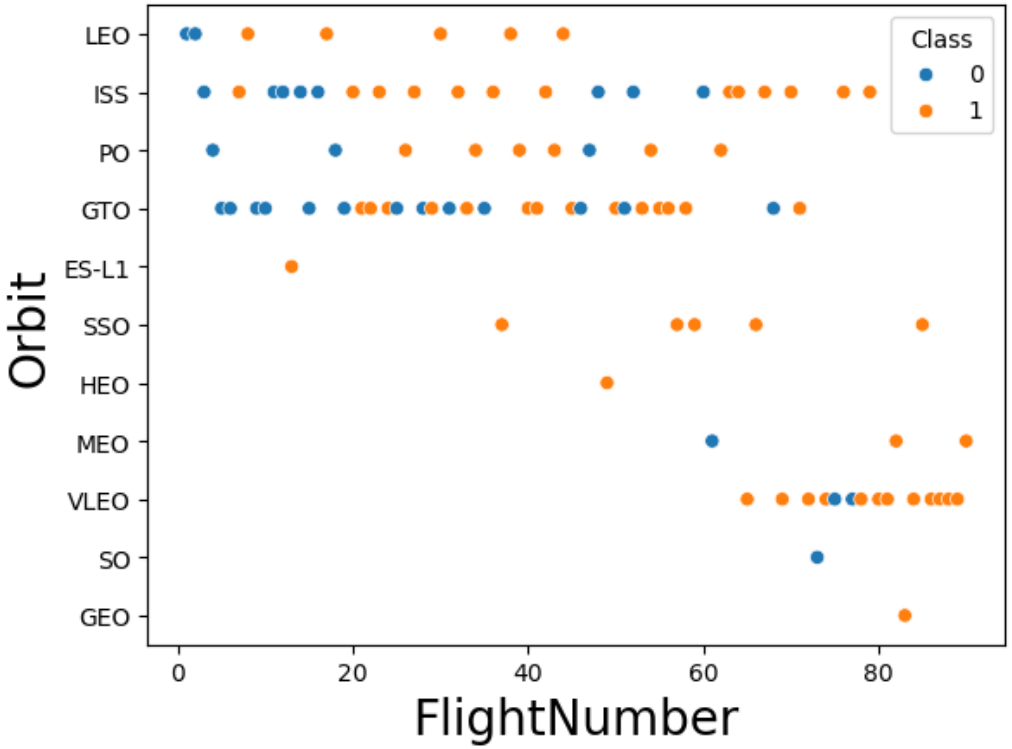
Success Rate vs. Orbit Type



Exploratory Data Analysis

- 100% success rate: ES-L1, GEO, HEO and SSO
- 50-80% success rate: GTO, ISS, LEO, MEO, PO
- 0% success rate: SO

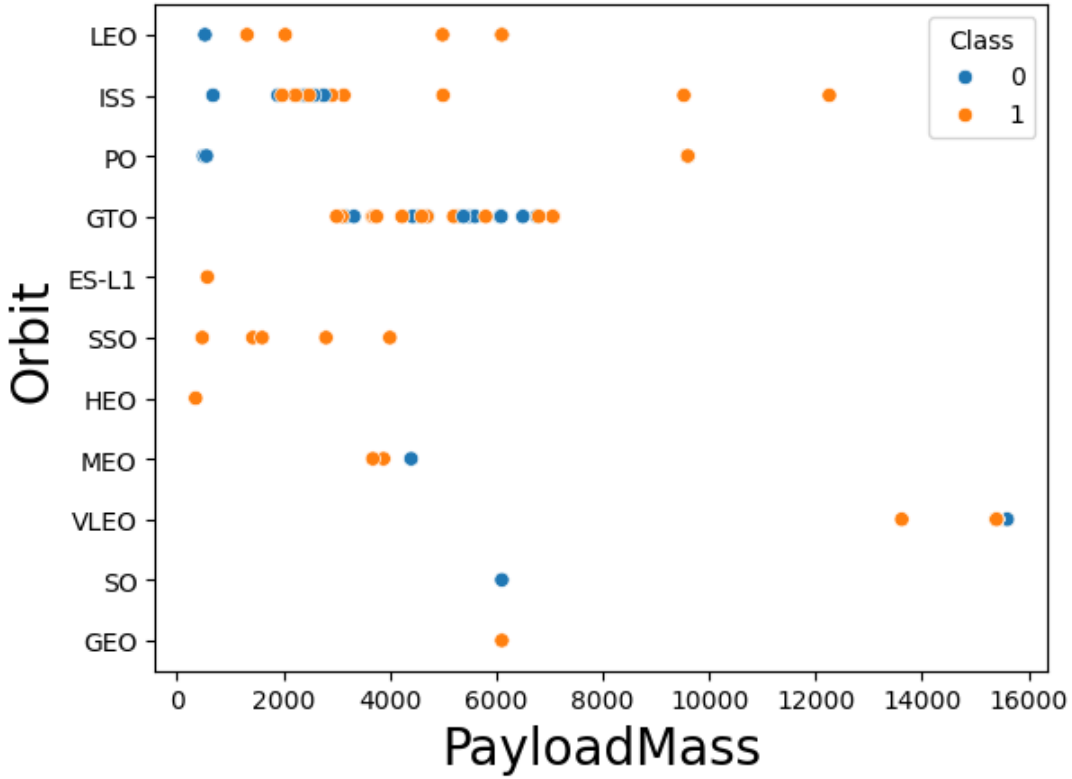
Flight Number vs. Orbit Type



Exploratory Data Analysis

- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO does not follow the trend

Payload vs. Orbit Type



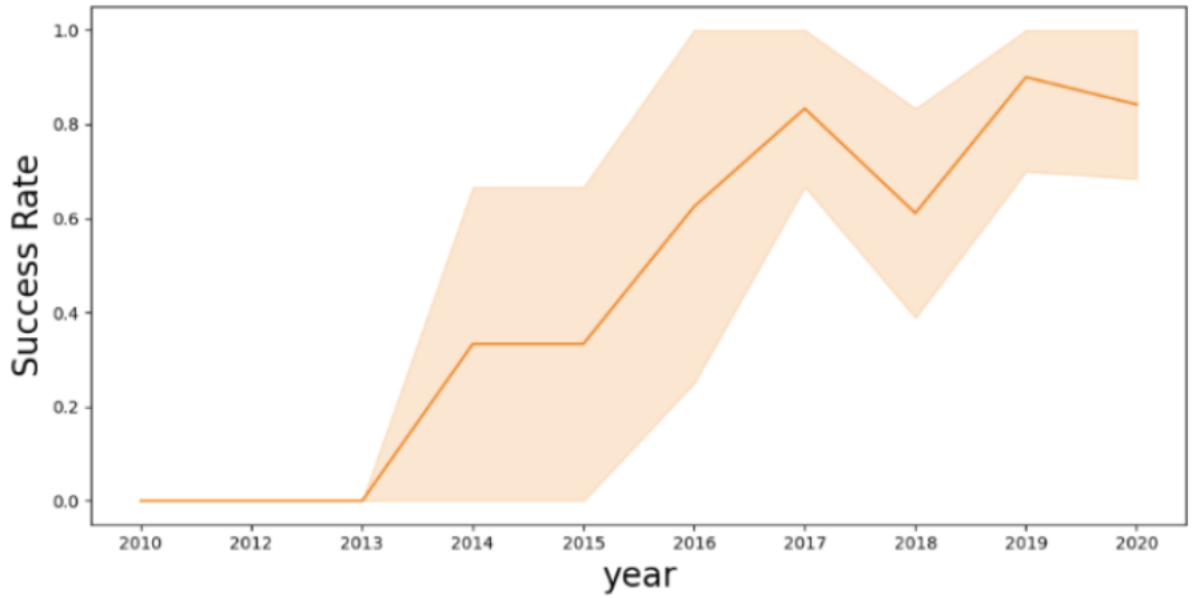
Exploratory Data Analysis

- Heavy payloads are better with LEO, ISS, and PO orbits
- The GTO orbit has mixed success with heavier payloads

Launch Success Yearly Trend

Exploratory Data Analysis

- The success rate improved from 2013-2017 and 2018-2019
- Overall, the success rate has improved since 2013



Launch Site Information

All Launch Site Names

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

Records with Launch Site Starting with CCA

```
%sql select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Payload Mass

Total Payload Mass

- **45,496** (kg) carried by boosters by NASA (CRS)

```
%sql select sum(PAYLOAD_MASS__KG_) as sum from SPACEXTABLE where Customer = "NASA (CRS)"
```

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

sum
45596

Average Payload Mass

- **2,928.4** (kg) carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) as avg from SPACEXTABLE where Booster_Version = "F9 v1.1"
```

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

avg
2928.4

Landing and Mission Information

1st Successful Landing in Ground Pad

- 12/22/2015

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

date
2015-12-22

Total Number of Successful and Failed Mission Outcomes

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

Success_Count	Failure_Count
100	1

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	
F9 FT B1021.1	F9 FT B1036.1
F9 FT B1022	F9 FT B1038.1
F9 FT B1023.1	F9 B4 B1041.1
F9 FT B1026	F9 FT B1031.2
F9 FT B1029.1	F9 B4 B1042.1
F9 FT B1021.2	F9 B4 B1045.1
F9 FT B1029.2	F9 B5 B1046.1

Launch Records

Boosters Carried Maximum Payload

Booster_Version	
F9 B5 B1048.4	F9 B5 B1049.5
F9 B5 B1049.4	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1058.3
F9 B5 B1056.4	F9 B5 B1051.6
F9 B5 B1048.5	F9 B5 B1060.3
F9 B5 B1051.4	F9 B5 B1049.7

2015 Launch Records

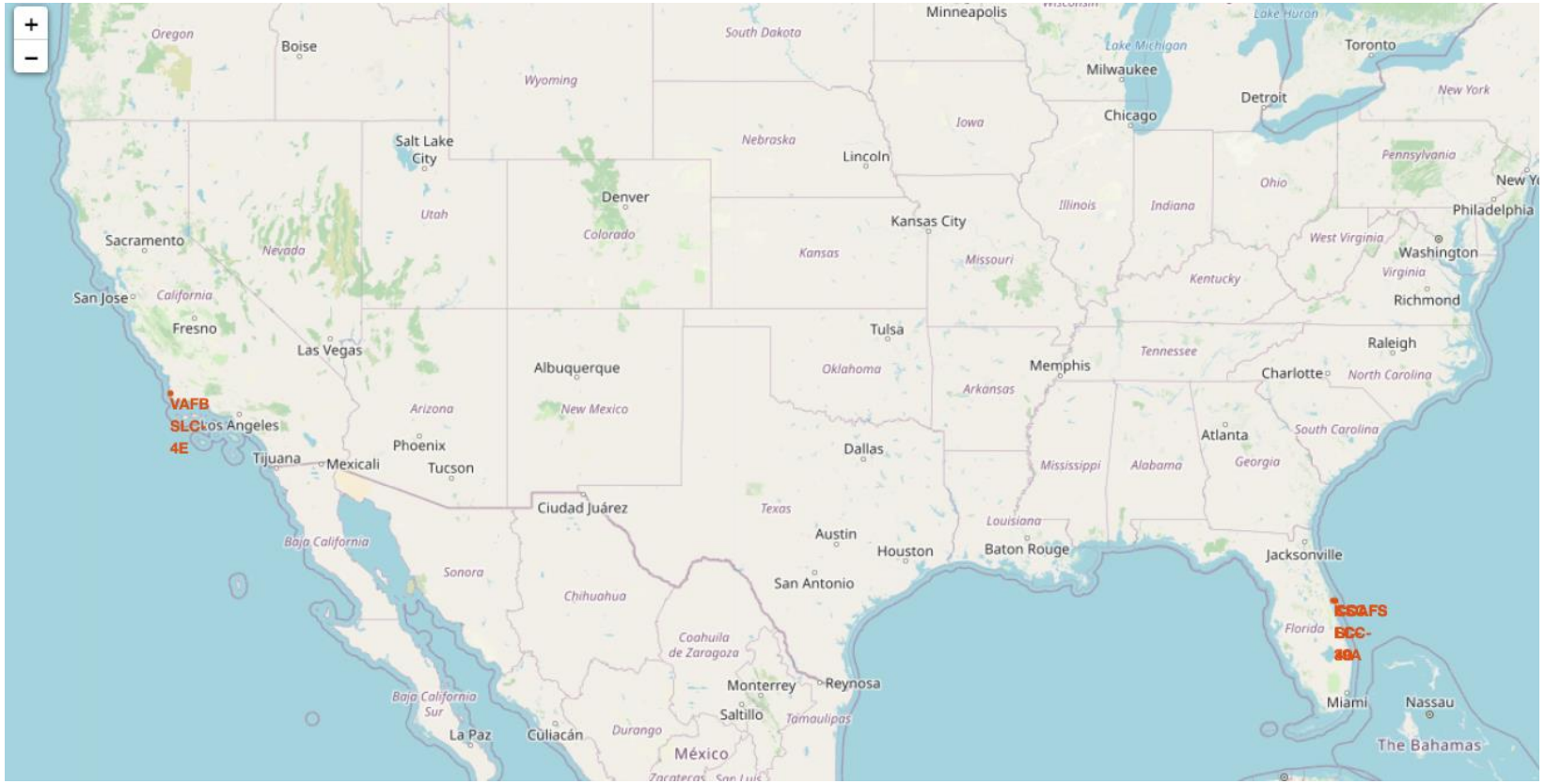
month_names	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

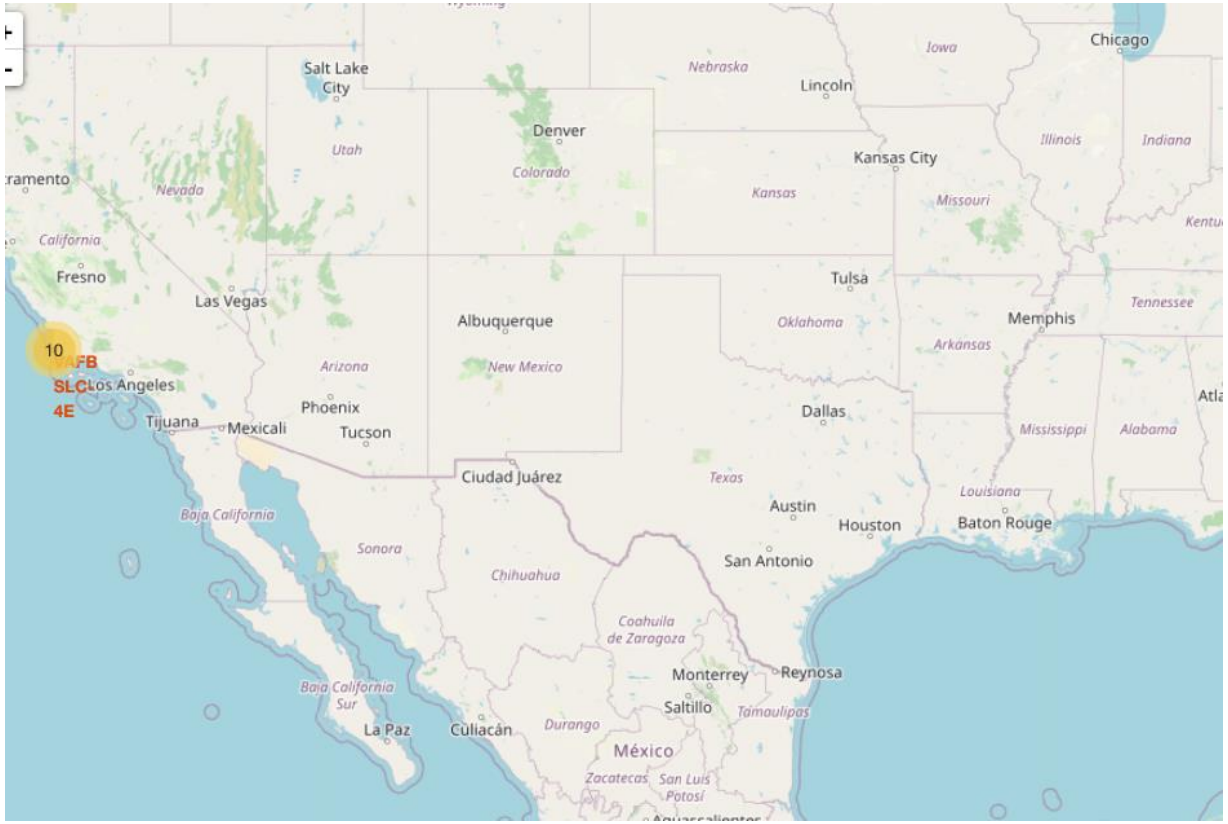
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

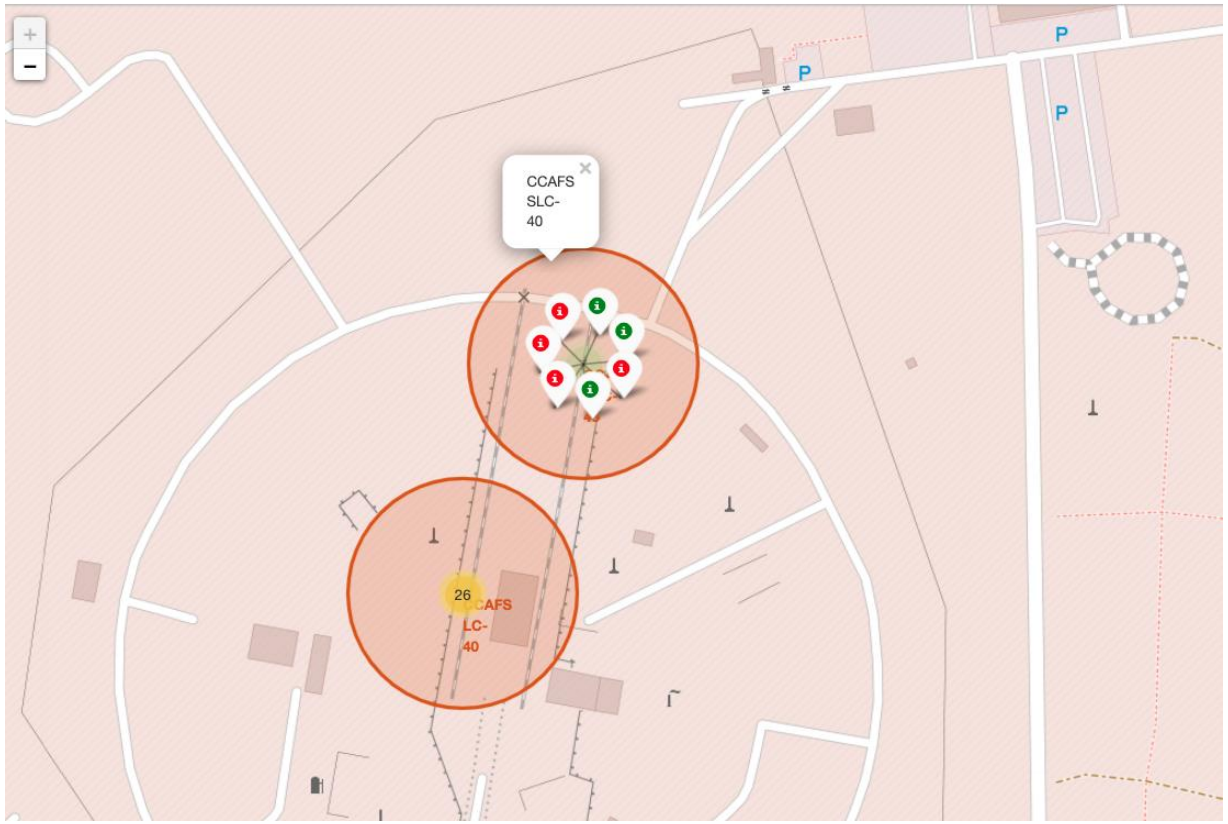
Folium Map

All Launch Site Location Markers

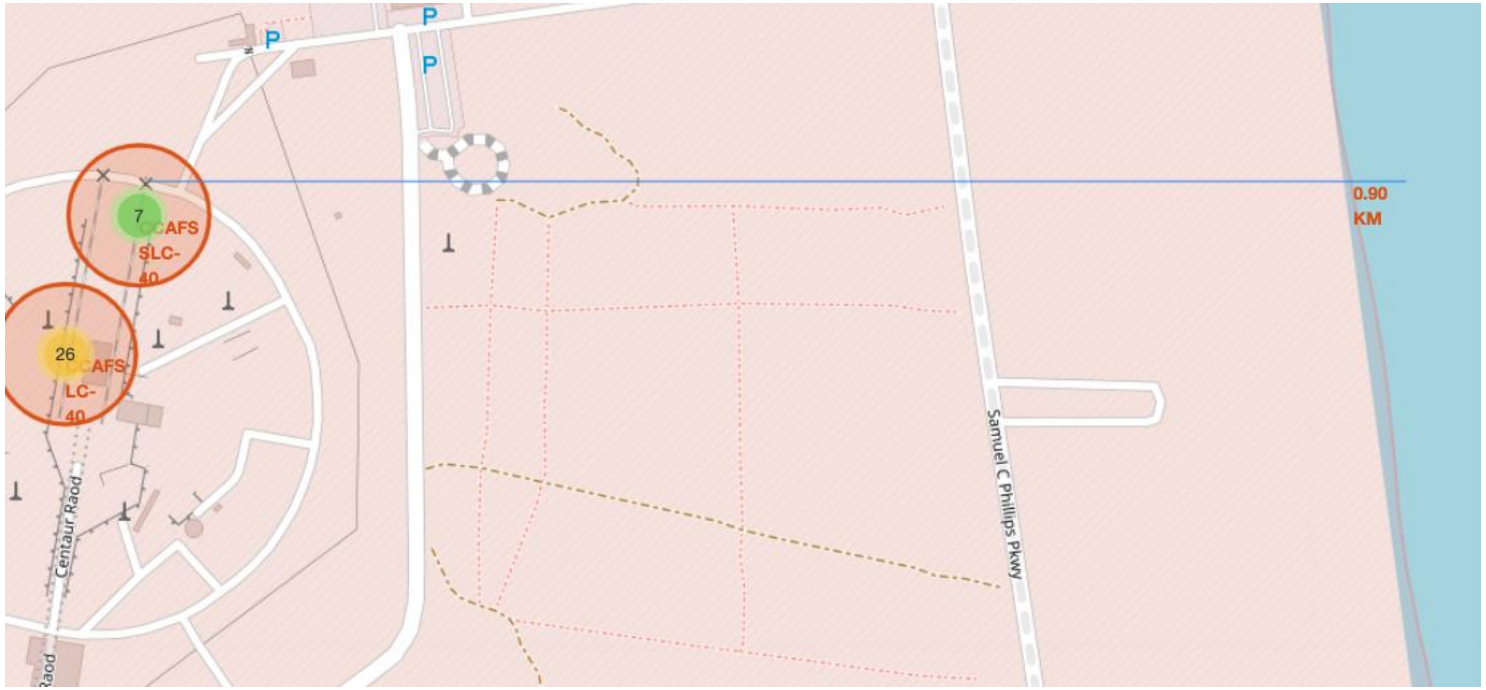


Success/Failed Launches for each site





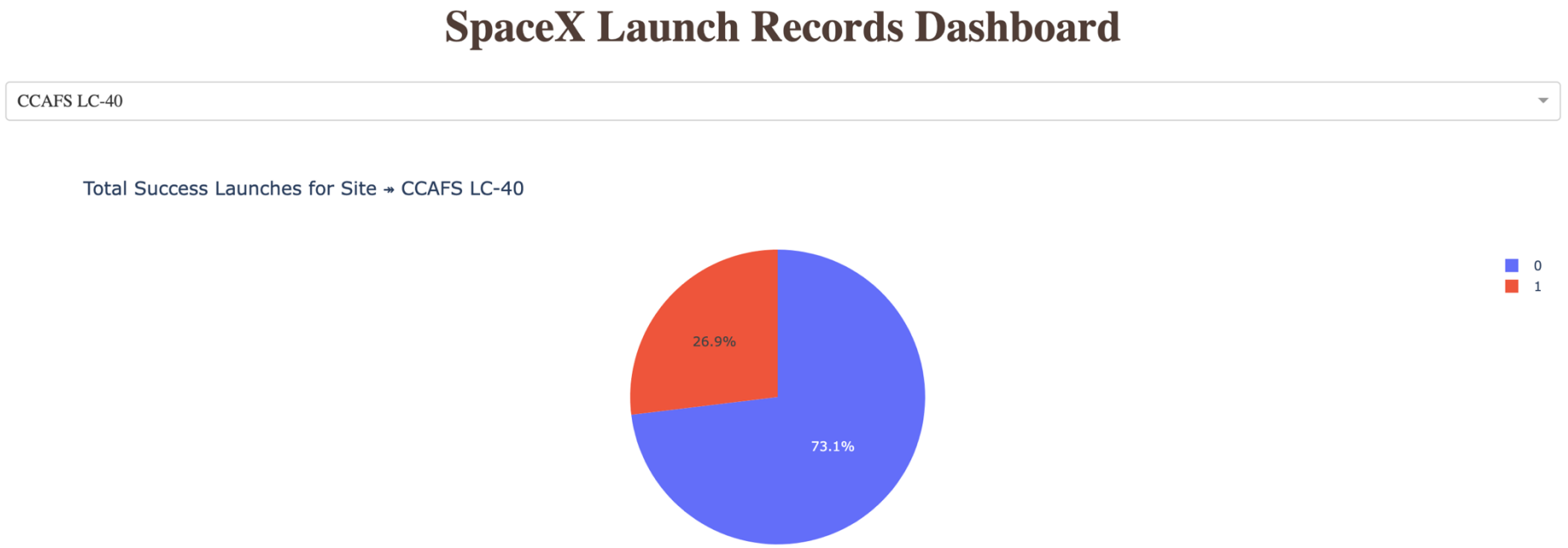
Calculate the Distances between launch sites



Dashboard

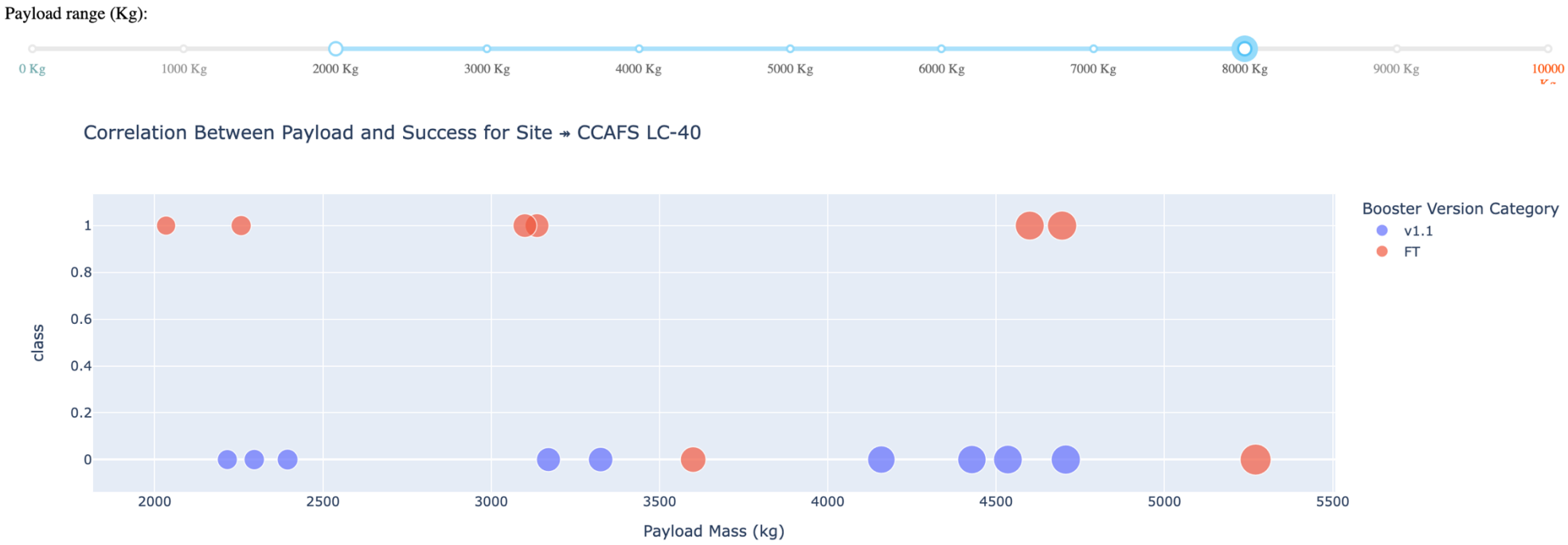
Dashboard

- The picture shows a pie chart when launch site CCAFS LC-40 is chosen.
- 0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.



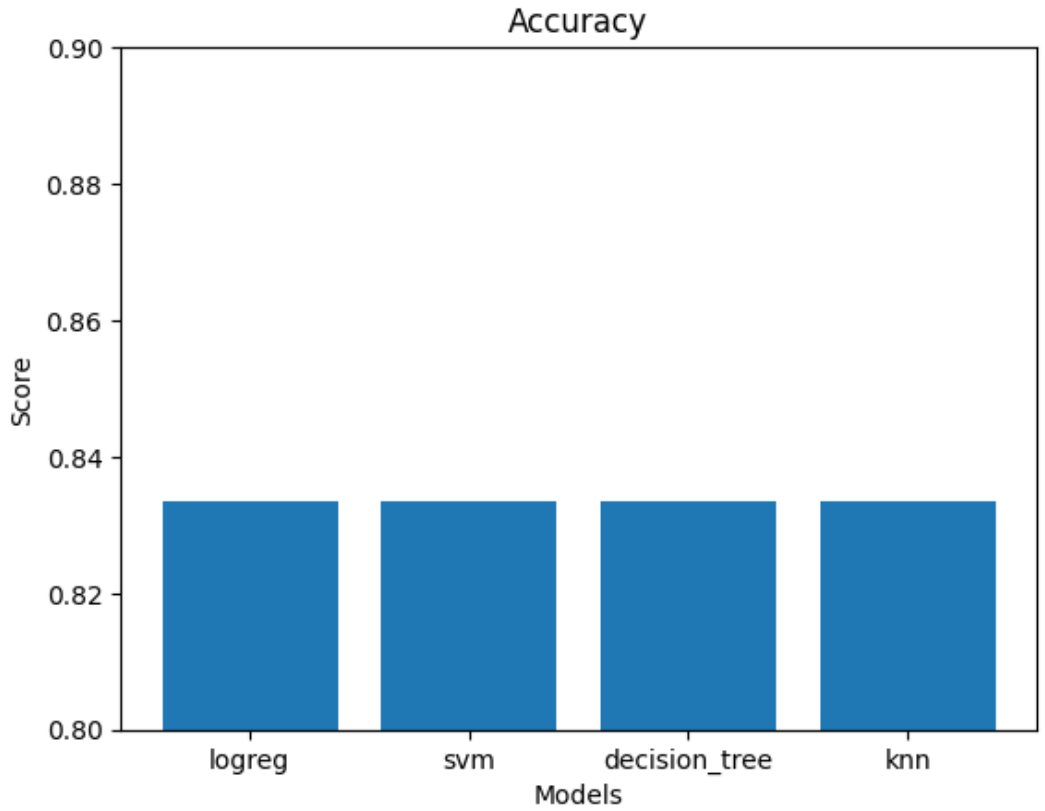
Dashboard

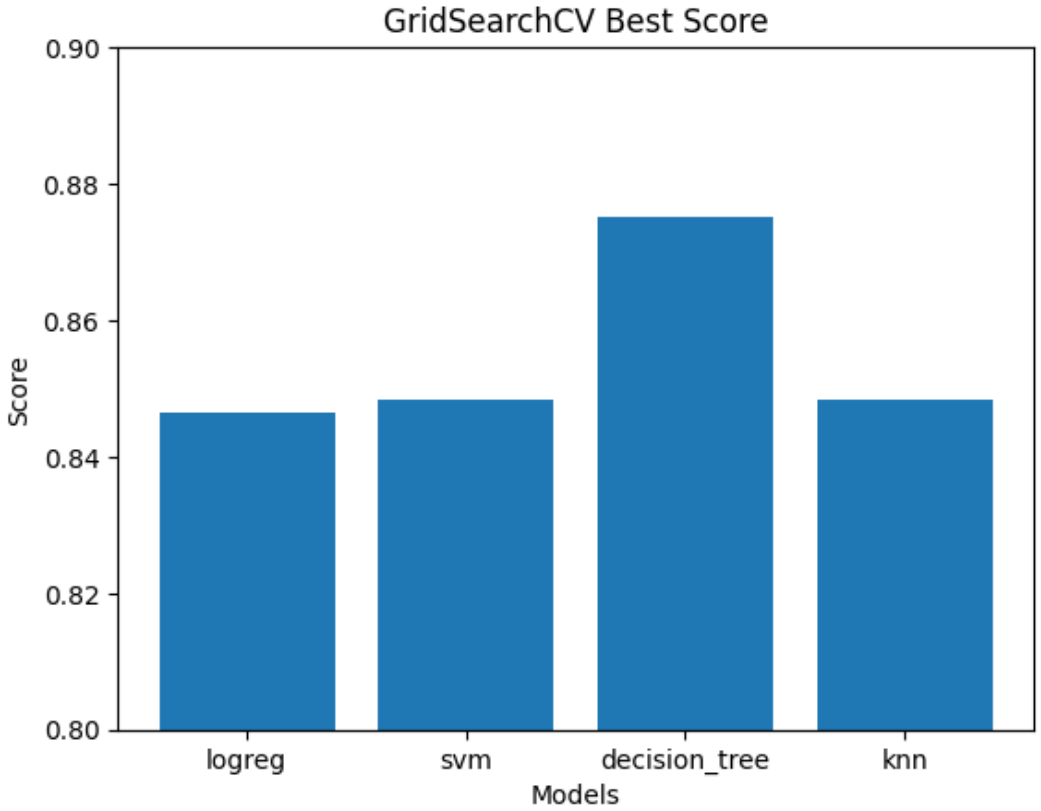
- The picture shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
- Class 0 represents failed launches while class 1 represents successful launches.



Conclusion

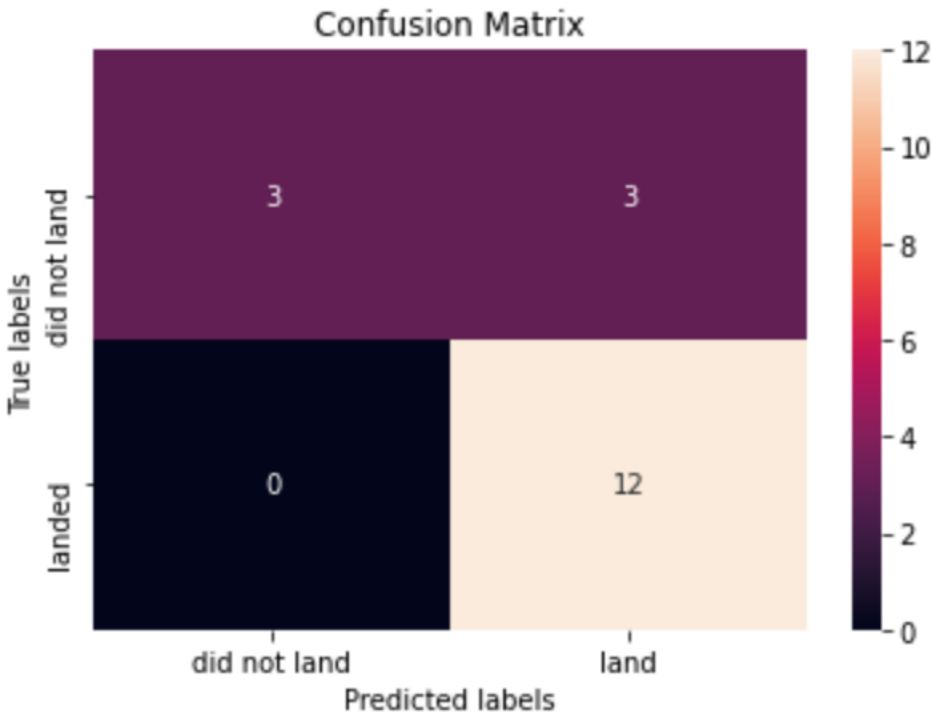
Classification Accuracy





Putting the results of all 4 models side by side, we can see that they all share the same accuracy score. Therefore, their GridSearchCV best scores are used to rank them instead. Based on the GridSearchCV best scores, decision tree model performs the best.

Confusion Matrix



- Decision tree
  - GridSearchCV best score: 0.8892857142857142
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:

Conclusions

- In this project, we try to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch.
- Each feature of a Falcon 9 launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way.
- Several machine learning algorithms are employed to learn the patterns of past Falcon 9 launch data to produce predictive models that can be used to predict the outcome of a Falcon 9 launch.
- The predictive model produced by decision tree algorithm performed the best among the 4 machine learning algorithms employed.

Appendix

```
import matplotlib.pyplot as plt

# Data
data = [0.8464285714285713, 0.8482142857142856, 0.875, 0.8482142857142858]
labels = ['logreg', 'svm', 'decision_tree', 'knn'] # Labels for the bars

# Create the bar graph
plt.bar(labels, data)

# Add labels and title
plt.xlabel('Models')
plt.ylabel('Score')
plt.title('GridSearchCV Best Score')

plt.ylim(0.80, 0.90)
# Display the graph
plt.show()
```