Winning The Space Race with Data Science

Paul Fagan
25/01/24

Outline

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

Executive Summary

Summary of methodologies

- Collect launch data from SpaceX Rest API and web scrape launch records from Wikipedia
- Wrangle the data by transforming JSON objects into a clean dataset (dataframe), filter data, using one-hot encoding, convert outcomes into classes
- Explore the data with (EDA) using visualization and SQL
- Visualise using Folium and Plotly Dash
- Predict landing outcomes by build, tune, evaluate classification models

Summary of all results

- We found 3 models that are all accurate at determining first stage will land. Finding these models means we can determine the cost of a launch.

“We live in the commercial space where companies are now competing to provide affordable space travel. SpaceX is the most successful with its Falcon 9 rocket launches at a cost of 62 million dollars vs 165 million dollars. Much of the savings is because SpaceX can reuse the first stage. SpaceY is evaluating how best to compete with SpaceX therefore if we can determine if the first stage will land, we can determine the cost of a launch.”

- Alon Mask, ‘visionary’ billionaire

Methodology

Methodology

Executive Summary

- Collect launch data from SpaceX Rest API and web scrape launch records from Wikipedia
- Wrangle the data by transforming JSON objects into a clean dataset (dataframe), filter data, using one-hot encoding, convert outcomes into classes
- Explore the data with (EDA) using visualization and SQL
- Visual using Folium and Plotly Dash
- Predict landing outcomes by build, tune, evaluate classification models

Data Collection – SpaceX API

GITHUB LINK

This step involves data retrieval, manipulation and cleaning to prepare the dataset for further analysis

1. **Request and parse the SpaceX launch data**:
    a. Make a GET request to the SpaceX API
2. **Subset and clean the data:**
    a. Extract key columns e.g. ‘Rocket’, ‘Payloads, ‘launchpad’ and define date ranges
3. **Construct dataset:**
    a. Combine extracted information into a Panda’s dataframe
    b. Filter for Falcon 9 launches
    c. Deal with missing values

Data Collection - Web Scraping

GITHUB LINK

Parsing Falcon 9 launch records from Wikipedia using web scraping techniques

1. **Request the Falcon 9 Launch Wiki page:**
    a. Use requests.get() to request the HTML page
2. **Create beautifulsoup object**
3. **Extract Column/variable names from HTML table header and parse Launch HTML tables**
    a. e.g. flight number, date, time, booster version, launch site, payload, payload mass, orbit, customer, launch outcome, and booster landing status.
4. **Create a dictionary for the launch records using the extracted info**

Data Wrangling

GITHUB LINK

Data wrangling plays a crucial role in preparing and transforming the raw dataset into a format suitable for further analysis and model training.

- **Task 1: Calculate Launch Site Success Rates:**
  - Use value_counts() to determine the number of launches at each site.
  - Calculate success rates for each launch site.
- **Task 2: Calculate Number and Occurrence of Orbits:**
  - Use value_counts() to determine the number and occurrence of each orbit**.**
- **Task 3: Calculate Number and Occurrence of Mission Outcomes:**
  - Use value_counts() to determine the number of landing outcomes.
- **Task 4: Create Landing Outcome Label:**
  - Create a list of labels (0 or 1) based on successful or unsuccessful landing outcomes.

EDA with Data Visualization

GITHUB LINK

We created a series of visualisations:

- **FlightNumber vs. PayloadMass Scatter Plot:**
  - To observe the relationship between the flight number and payload mass concerning the success or failure of the launch.
- **FlightNumber vs. LaunchSite Scatter Plot:**
  - To explore the relationship between the flight number and launch site in terms of the launch outcome.
- **PayloadMass vs. LaunchSite Scatter Plot:**
  - To observe the relationship between payload mass and launch site in the context of the launch outcome.
- **Orbit Class Success Rate Bar Chart:**
  - To visualize the success rate of each orbit type.

EDA with Data Visualization (cont…)

GITHUB LINK

- **Orbit Class Success Rate Bar Chart:**
  - To visualize the success rate of each orbit type.
- **FlightNumber vs. Orbit Scatter Plot:**
  - To examine the relationship between the flight number and orbit type concerning the launch outcome.
- **PayloadMass vs. Orbit Scatter Plot:**
  - To visualize the relationship between payload mass and orbit type in terms of the launch outcome.
- **Average Success Rate Over Years Line Chart:**
  - To observe the trend in average success rates over the years.

EDA with SQL

GITHUB LINK

Load the SpaceX dataset into a table in a Db2 database and execute the following queries

Display:

- the names of unique launch sites in the space mission.
- 5 records where launch sites begin with the string 'CCA'.
- the total payload mass carried by boosters launched by NASA (CRS).
- the average payload mass carried by booster version F9 v1.1.
- records showing month names, failure landing outcomes on the drone ship, booster versions, and launch sites for the months in the year 2015.

List:

- the date when the first successful landing outcome on the ground pad was achieved.
- the names of boosters with success in a drone ship and payload mass between 4000 and 6000.
- the total number of successful and failure mission outcomes.
- the names of booster versions that carried the maximum payload mass using a subquery.

Rank:

- the count of landing outcomes between the dates 2010-06-04 and 2017-03-20 in descending order.

Build an Interactive Map with Folium

GITHUB LINK

- **In Task 1, a folium map was created to mark all launch sites**.
  - Various map objects such as circles and markers were added to represent each launch site. The circles serve as visual indicators of the general area of the launch site, and markers provide specific locations on the map. The map was centered around NASA Johnson Space Center in Houston, Texas.
- **In Task 2, additional markers were added to represent the success or failure of each launch at the respective launch sites.**
  - Green markers indicate successful launches, while red markers indicate failed launches. Marker clusters were used to handle multiple markers at the same location, providing a clearer visualization of success rates.
- **In Task 3, distances between a launch site and specific proximities (coastline and city) were calculated and visualized.**
  - Markers with distance labels were added, and polyline objects were used to draw lines connecting launch sites to their closest points of interest.

Build a Dashboard with Plotly Dash

GITHUB LINK

Dropdown List for Launch Site Selection (Task 1):

- A dropdown list (dcc.Dropdown) was added to enable the selection of a specific launch site. The default value is set to 'ALL', representing all launch sites. This allows users to filter data based on the launch site.

Pie Chart for Total Successful Launches (Task 2):

- A pie chart (dcc.Graph) was added to show the total number of successful launches for all sites or a selected site. The chart dynamically updates based on the chosen launch site in the dropdown. This provides an overview of the success distribution among different launch sites.

Payload Range Slider (Task 3):

- A range slider (dcc.RangeSlider) was introduced to enable the selection of payload mass ranges. Users can slide to choose a specific payload mass range, and the selected range is used for filtering data in the scatter chart.

Scatter Chart for Payload vs. Launch Success (Task 4):

- A scatter chart (dcc.Graph) was added to display the correlation between payload mass and launch success. The chart includes a color dimension for the Booster Version Category, providing additional insights. Users can interactively choose a launch site and payload range to visualize specific subsets of the data.

Predictive Analysis (Classification)

GITHUB LINK

The objective is to predict whether the first stage of SpaceX Falcon 9 rocket will land successfully, based on preceding data.

1. **Data Exploration and Labeling:**
   a. Explored and analyzed data.
   b. Created a column 'Class' to represent the target variable.
2. **Data Standardization:**
   a. Standardized the features in the dataset using the Standard Scaler.
3. **Data Splitting:**
   a. Split the data into training and test sets using a train-test split ratio of 80:20.
4. **Model building (see across)**
5. **Performance Evaluation:**
   a. Plotted confusion matrices for each model to visualize true positives, true negatives, false positives, and false negatives

- **Logistic Regression Model:**
  - Created a Logistic Regression model.
  - Utilized GridSearchCV to find the best hyperparameters.
  - Obtained accuracy on the test data.
- **Support Vector Machine (SVM) Model:**
  - Created an SVM model.
  - Used GridSearchCV to find the best hyperparameters.
  - Calculated accuracy on the test data.
- **Decision Tree Model:**
  - Constructed a Decision Tree classifier.
  - Applied GridSearchCV to determine the best hyperparameters.
  - Evaluated accuracy on the test data.
- **K Nearest Neighbors (KNN) Model:**
  - Implemented a KNN classifier.
  - Employed GridSearchCV to find optimal hyperparameters.
  - Assessed accuracy on the test data.

Predictive Analysis (Cont…)

1. **Data Results:**
   a. The table shows the results from Logistic Regression, Support Vector Machine, Decision tree and K-Nearest Neighbors models

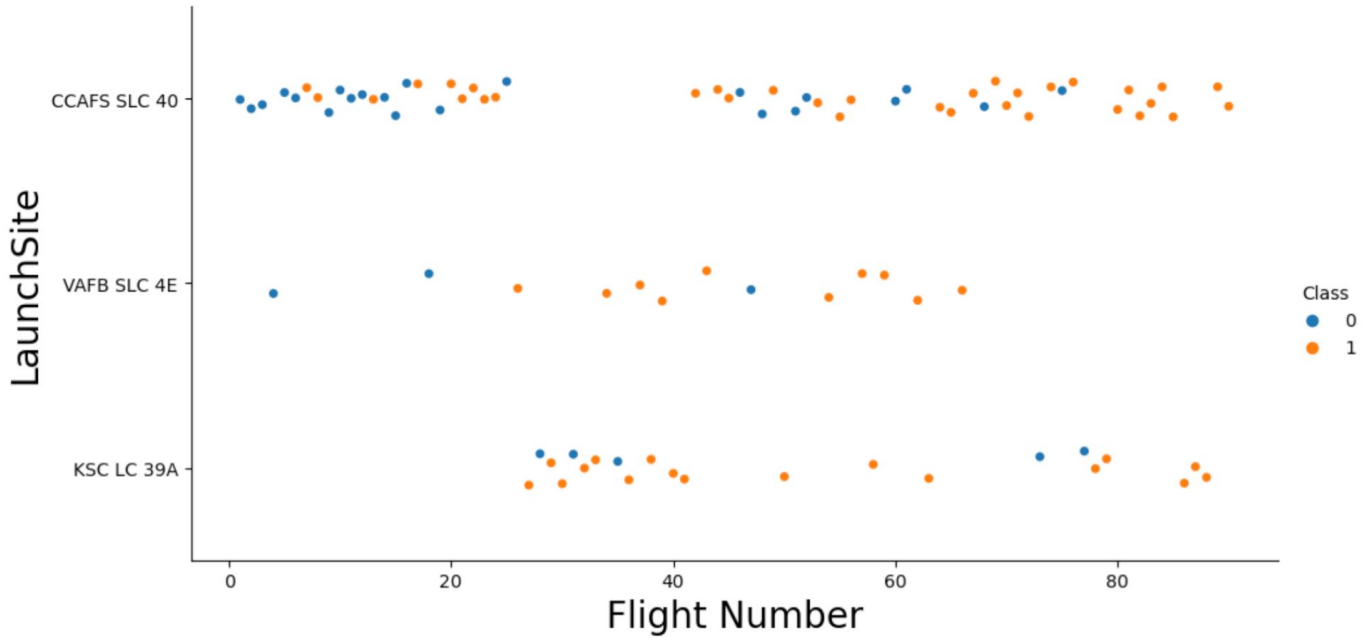
Model	Accuracy
Logistic Regression	0.8333333333333334
Support Vector Machine	0.8333333333333334
Decision Tree	0.7222222222222222
K-Nearest Neighbors	0.8333333333333334

Insight Drawn from EDA

Different launch sites have different success rates.

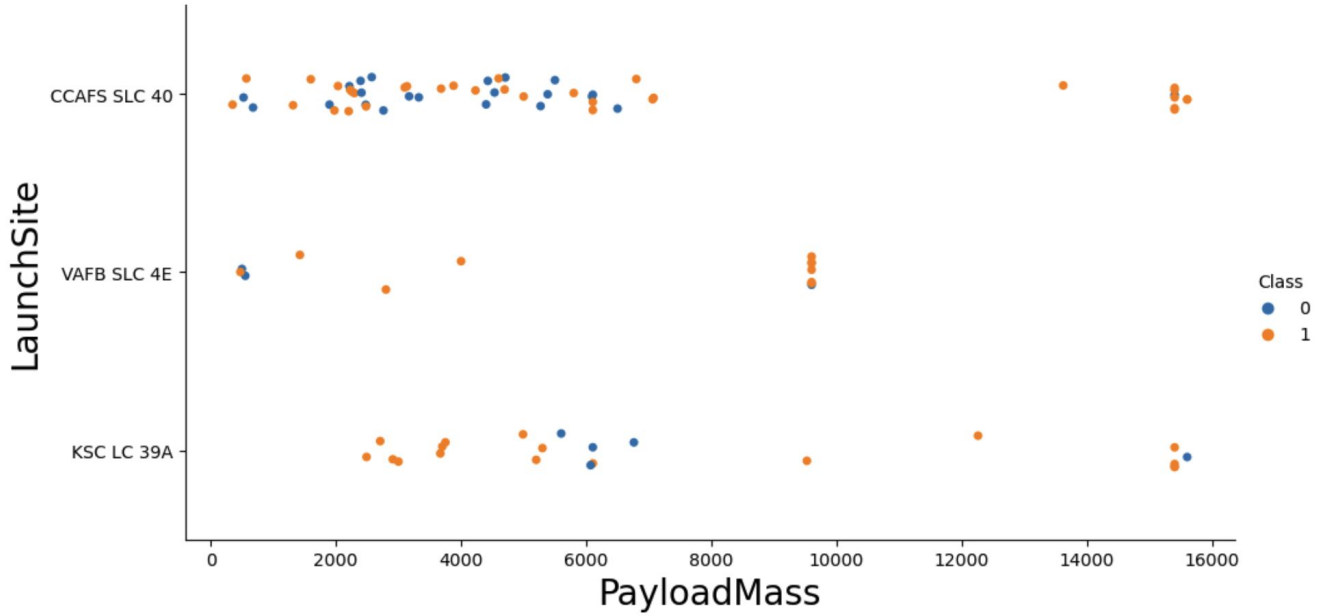
For example, CCAFS LC-40 has a success rate of 60%, while KSC LC-39A success rate is 24% and VAFB SLC 4E 14%.

Flight Number vs. Launch Site



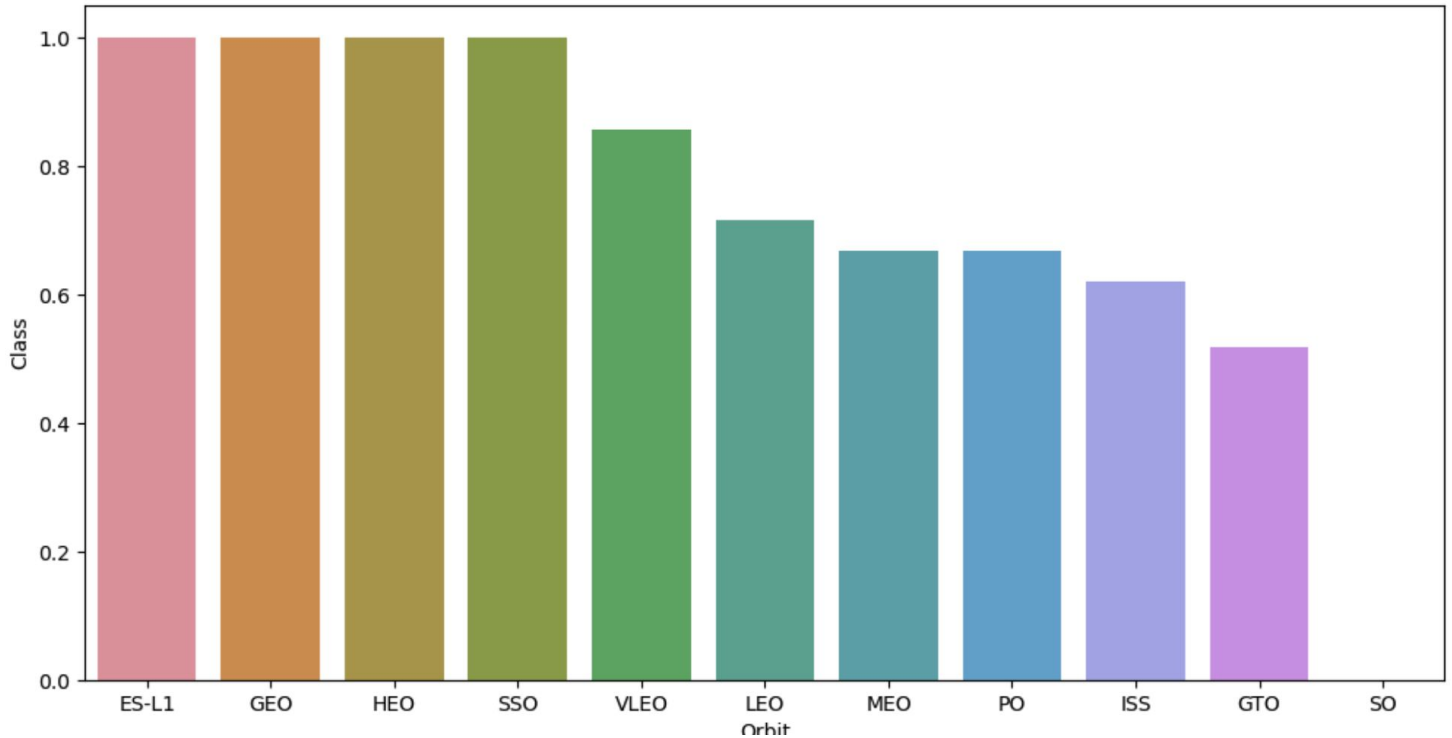
Payload vs. Launch Site

There is a positive correlation between payload mass and launch success rate



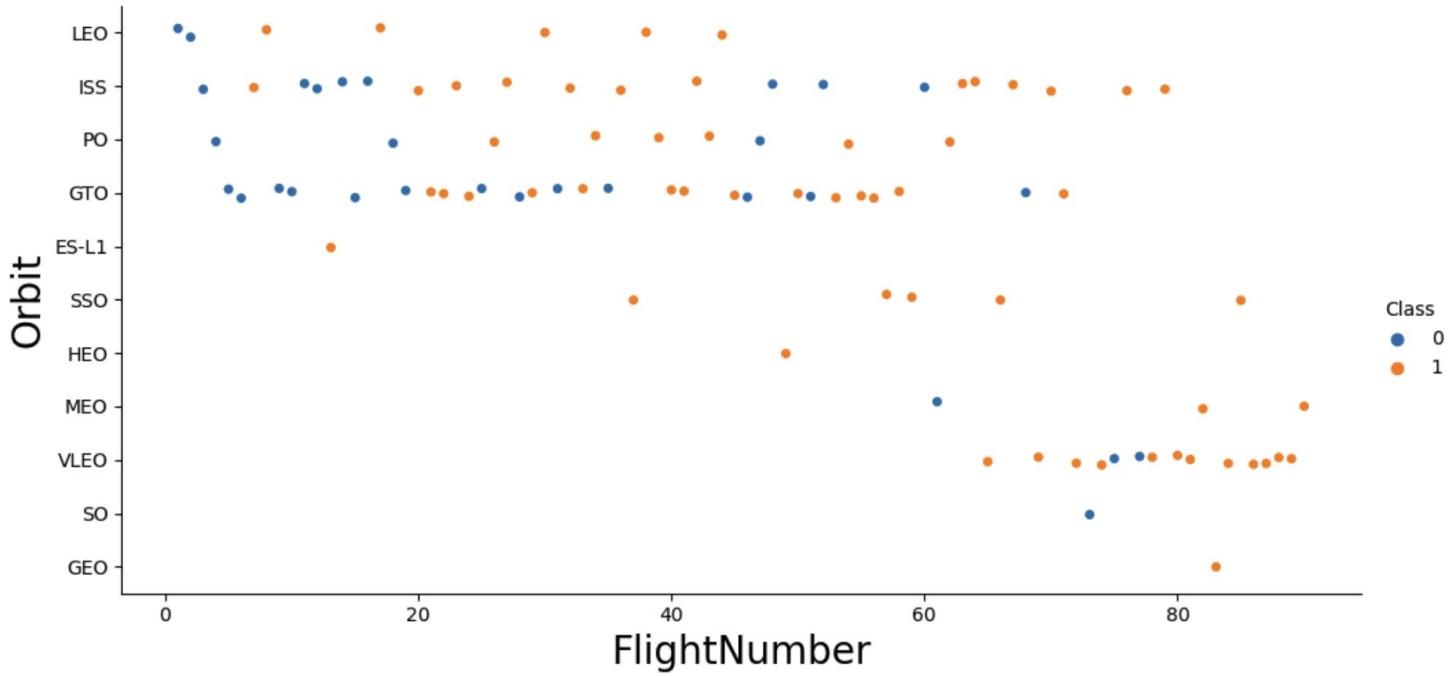
Orbits like Polar, LEO, and ISS have higher success rates compared to GTO.

Success rate vs. Orbit type



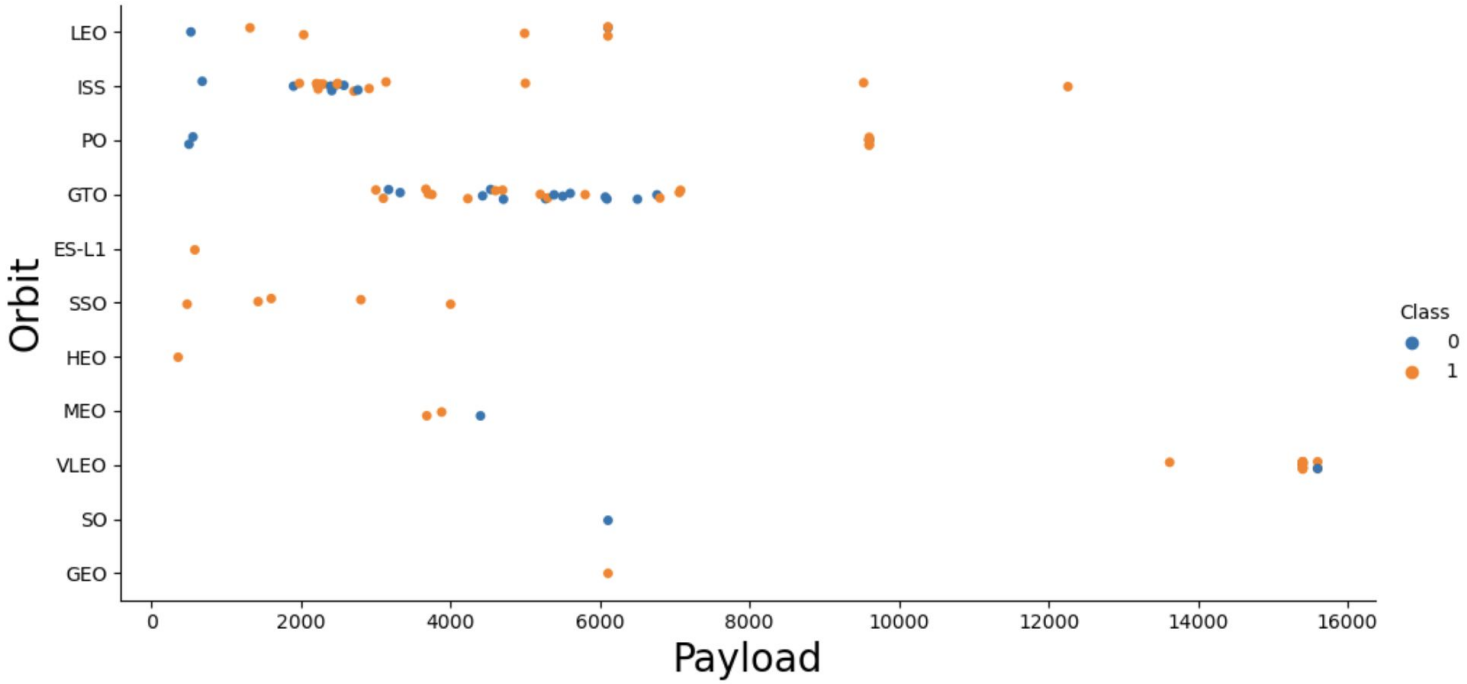
In the LEO orbit, success appears related to the number of flights, while there seems to be no relationship between flight number and GTO orbit success.

Flight Number vs. Orbit Type



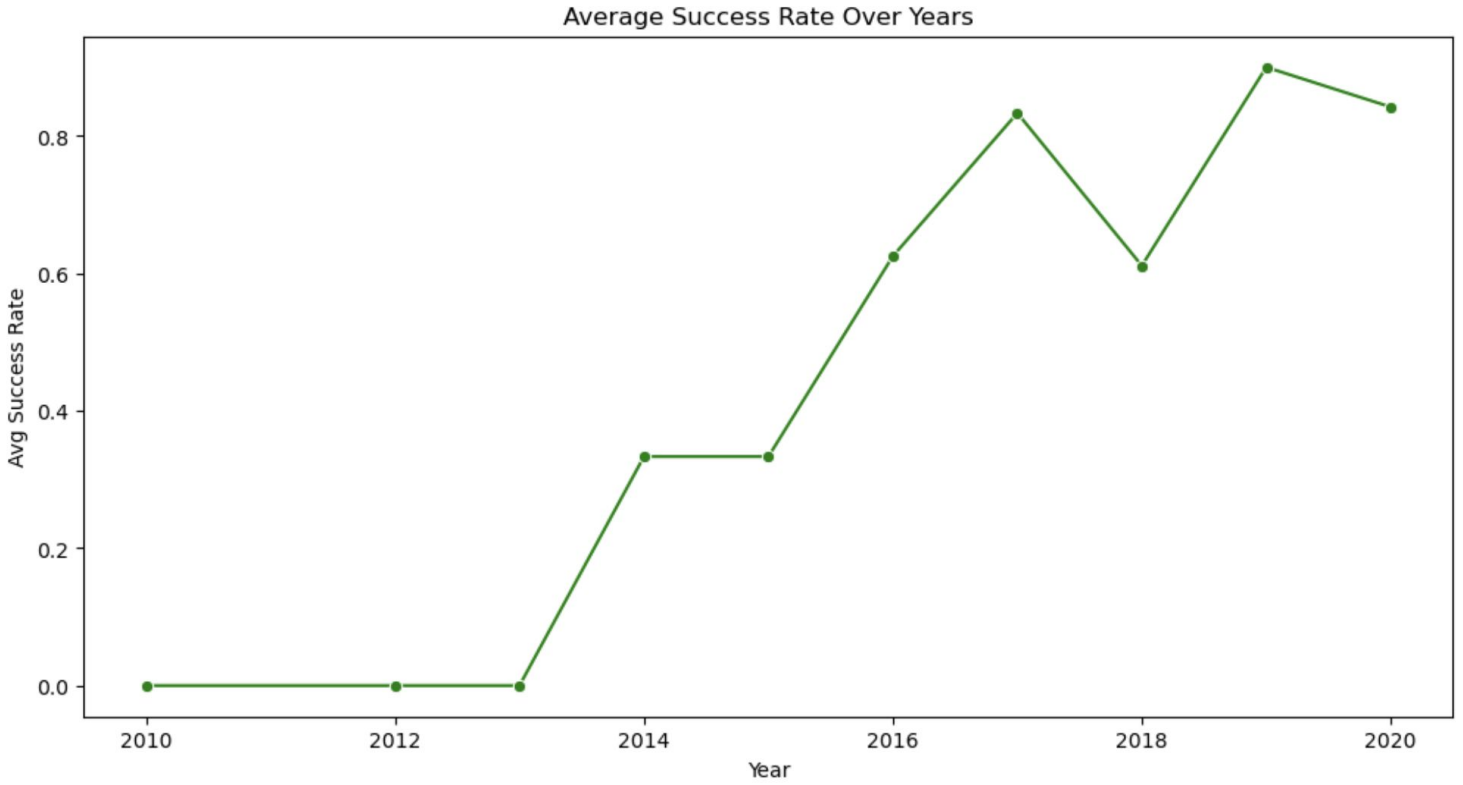
Heavy payloads show higher success rates for Polar, LEO, and ISS orbits. GTO does not distinguish well between positive and negative landing rates.

Payload vs. Orbit Type



Average success trend

The success rate has been increasing since 2013, with a stable period in 2014 and a significant increase after 2015



All Launch Site Names

- **CCAFS LC-40:** Cape Canaveral Air Force Station Launch Complex 40. It is located in Florida, USA.
- **VAFB SLC-4E:** Vandenberg Air Force Base Space Launch Complex 4E. It is located in California, USA.
- **KSC LC-39A:** Kennedy Space Center Launch Complex 39A. It is located in Florida, USA.
- **CCAFS SLC-40:** Another representation of Cape Canaveral Air Force Station Launch Complex 40.

```
%sql select distinct "Launch_Site" from SPACEXTBL
```

Q. Site Names Begin with 'CCA'

Launch_Site
CCAFS LC-40
CCAFS SLC-40

```
%sql select distinct "Launch_Site" from SPACEXTBL where "Launch_Site" LIKE 'CCA%' LIMIT 5
```

Q. Total Payload Mass

Total_Payload_Mass
45596

```
%sql select sum ("PAYLOAD_MASS__KG_") AS Total_Payload_Mass from SPACEXTBL where "Customer" = 'NASA (CRS)'
```

Q. Average Payload Mass by F9 v1.1

Avg_Payload_Mass
2928.4

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

Q. First Successful Ground Landing Date

MIN("Date")
2015-12-22

```
%sql select MIN("Date") from SPACEXTBL where "Landing_Outcome" = 'Success (ground pad)'
```

Q. Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

```
%sql select "Booster_Version" from SPACEXTBL where "Landing_Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_
```

Q. Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	Count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

```
%sql select "Mission_Outcome", Count(*) from SPACEXTBL Group by "Mission_Outcome"
```

Q. Boosters Carried Maximum Payload

Booster_Version
F9 B4 B1043.1

```
%sql select "Booster_Version" from SPACEXTBL where Payload = (select max("Payload") from SPACEXTBL )
```

Q. 2015 Launch Records

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

```
%sql select substr(Date, 6,2) AS Month, Booster_Version, launch_site \
from SPACEXTBL where  substr(Date,0,5)='2015' and "landing_Outcome" LIKE 'Failure (drone ship)%'
```

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

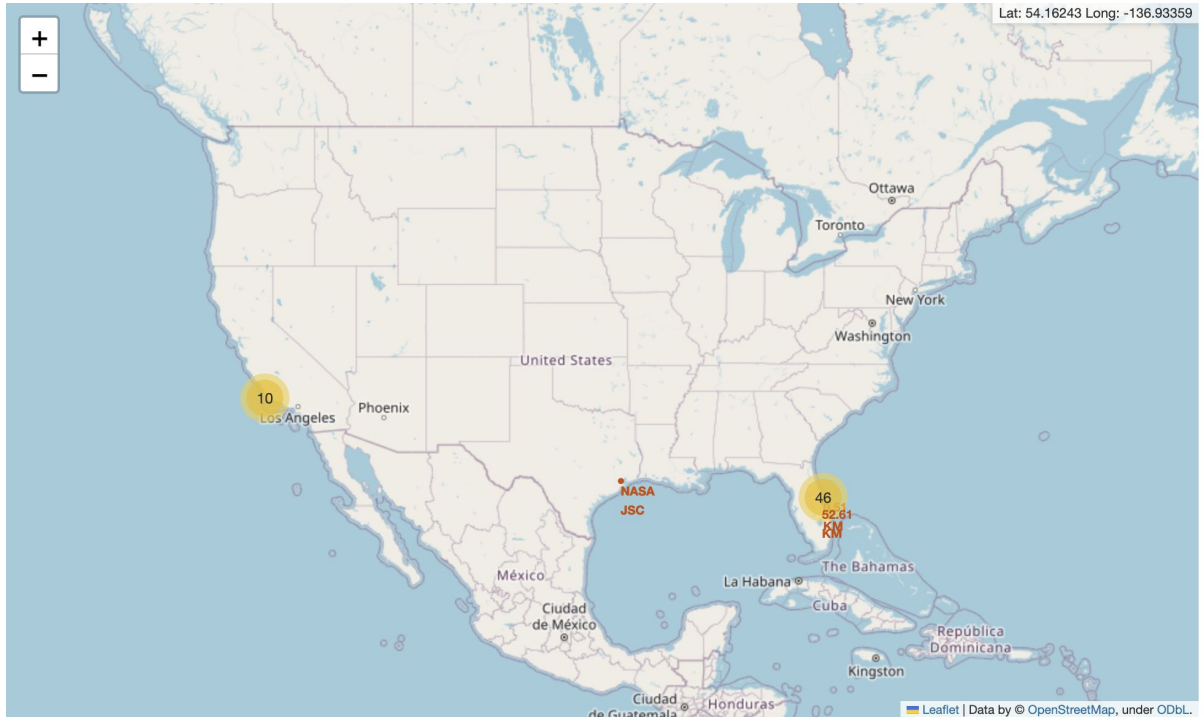
Landing_Outcome	OutcomeCount
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

```
%sql SELECT Landing_Outcome, count(*) as OutcomeCount \
from SPACEXTBL where DATE between '2010-06-04' and '2017-03-20' \
group by Landing_Outcome order by OutcomeCount DESC
```

Launch Sites Proximities Analysis

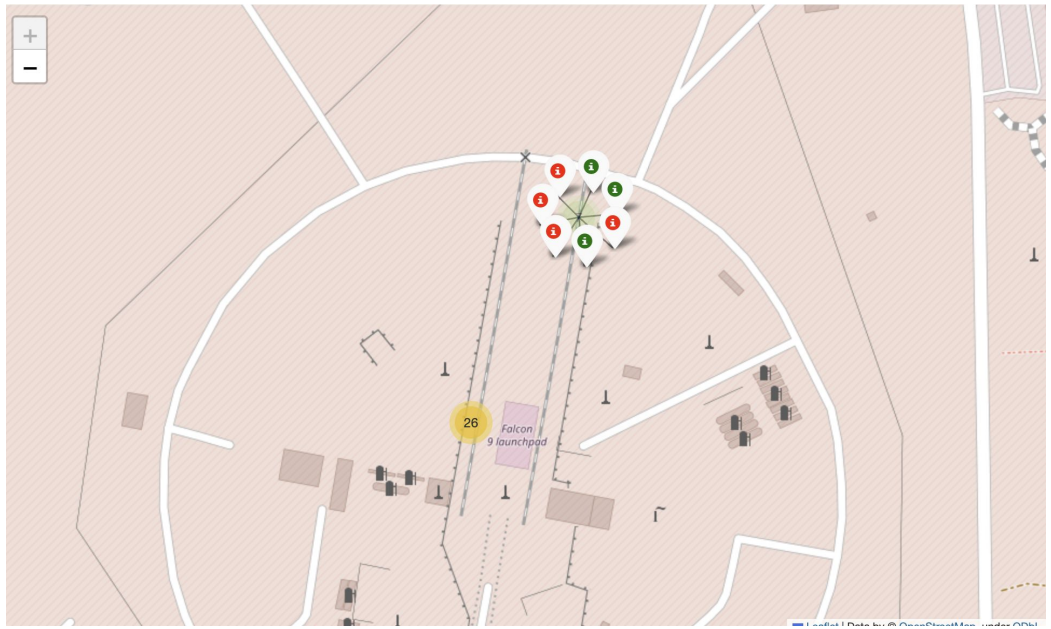
All SpaceX launch sites in the US

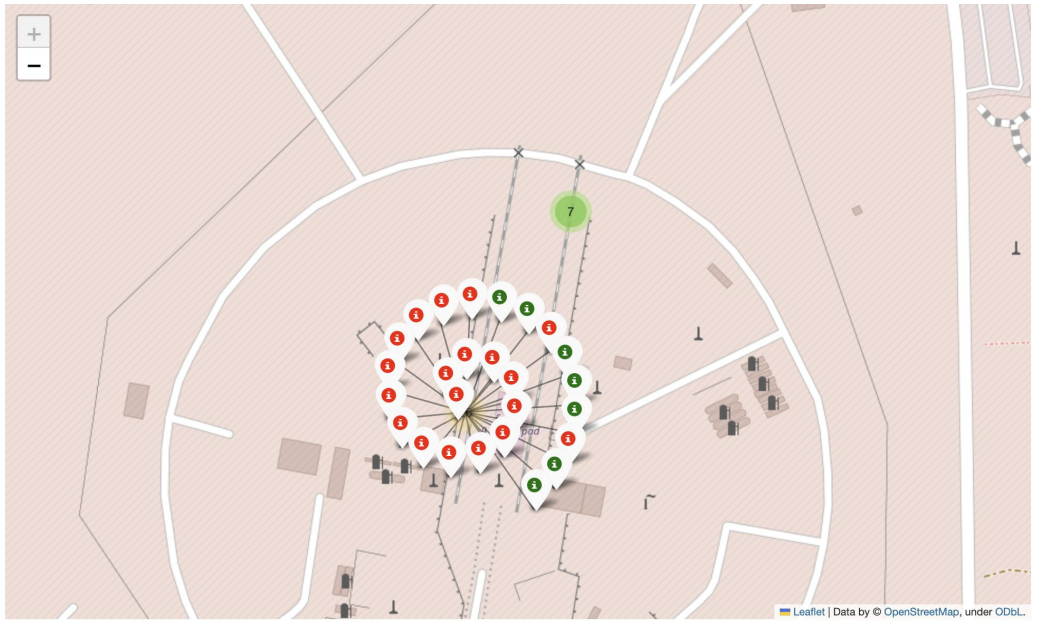
- There were:
  - 10 launches at VAFB SLC-4E
  - 46 KSC L-39A & CCAFS LC-40/SLC-40



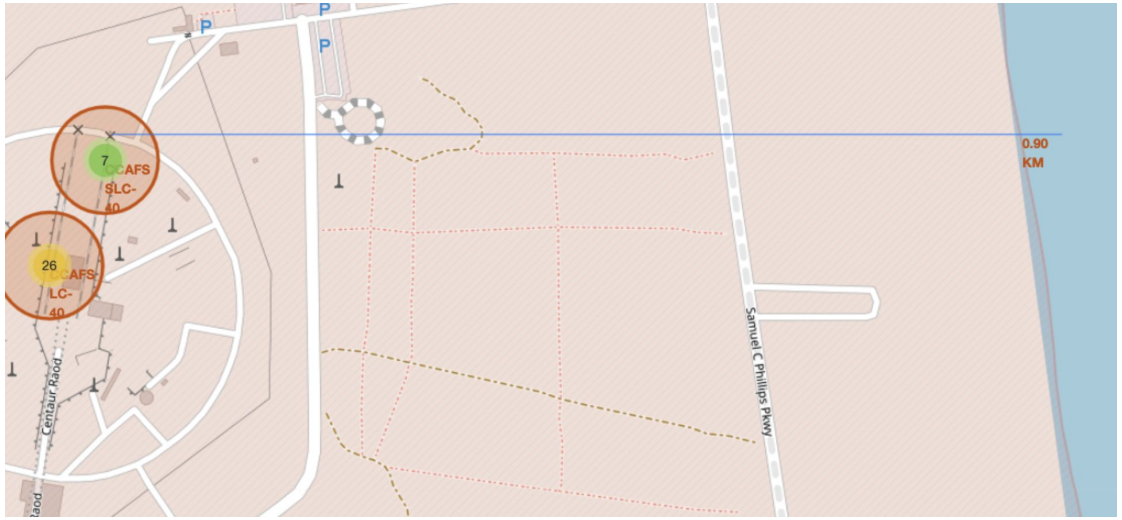
The success/failed launches for each SpaceX site

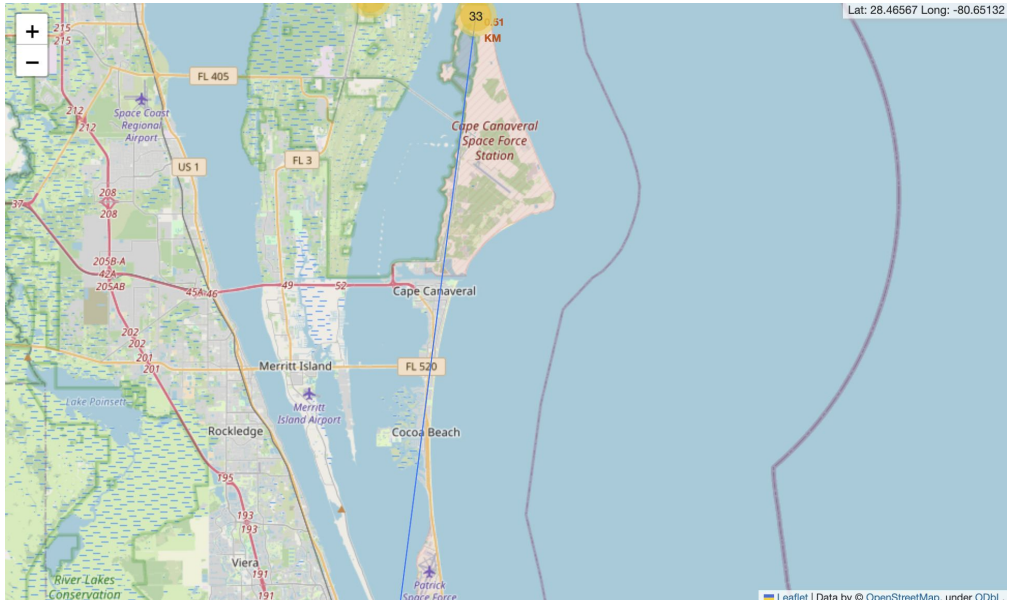
We've adding the launch outcomes for each site to see which sites have high success rates.





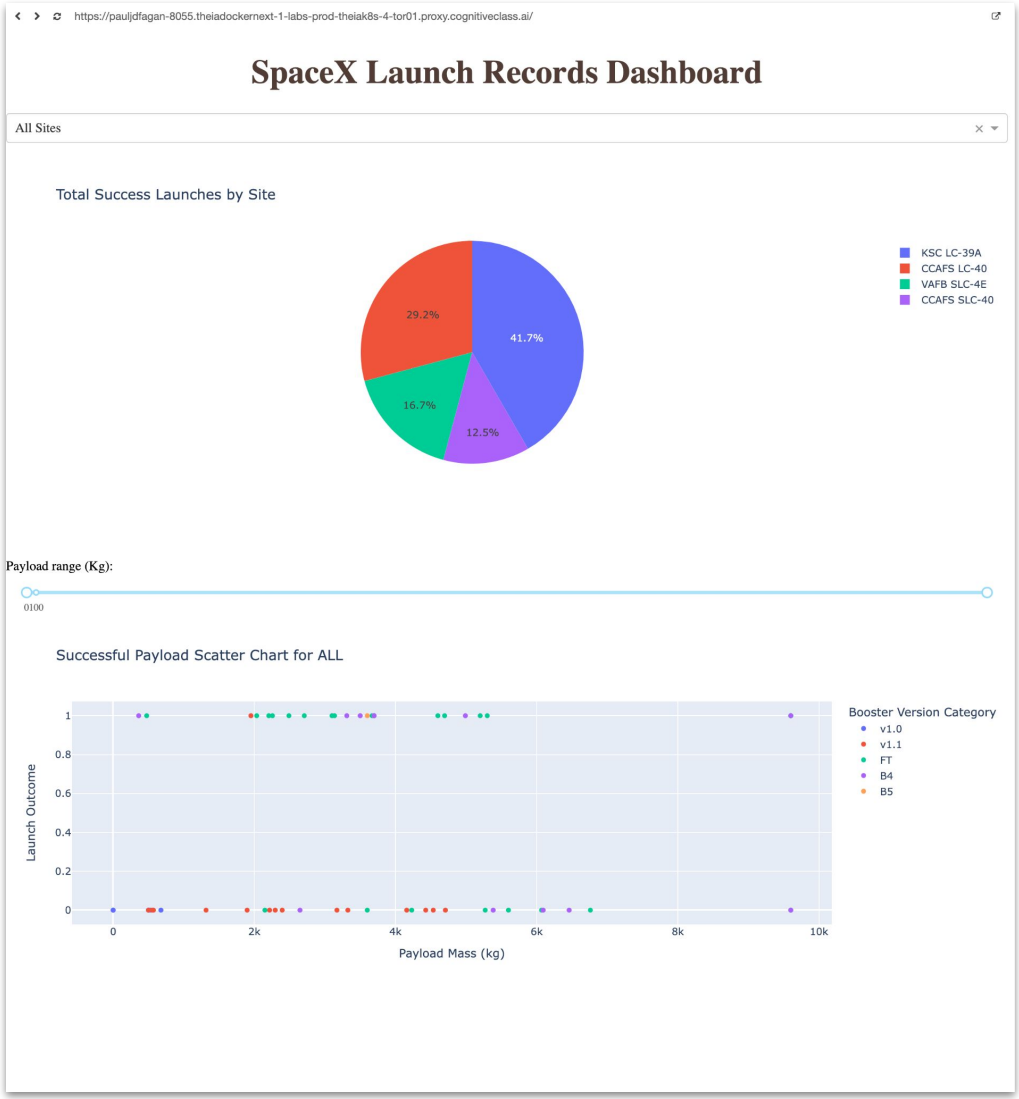
Calculate the distances between a launch site to its proximities



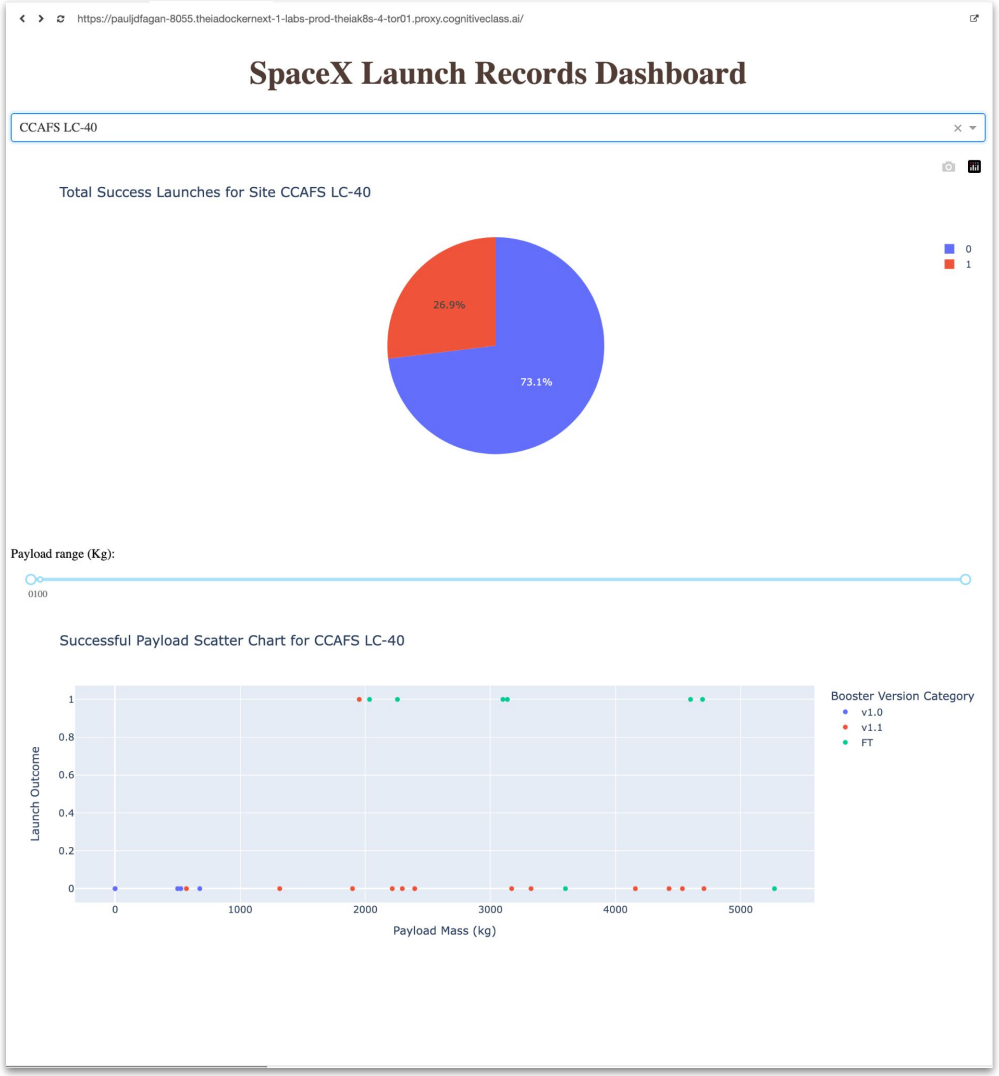


Build a Dashboard with Plotly Dash

Total successful launches for each SpaceX site

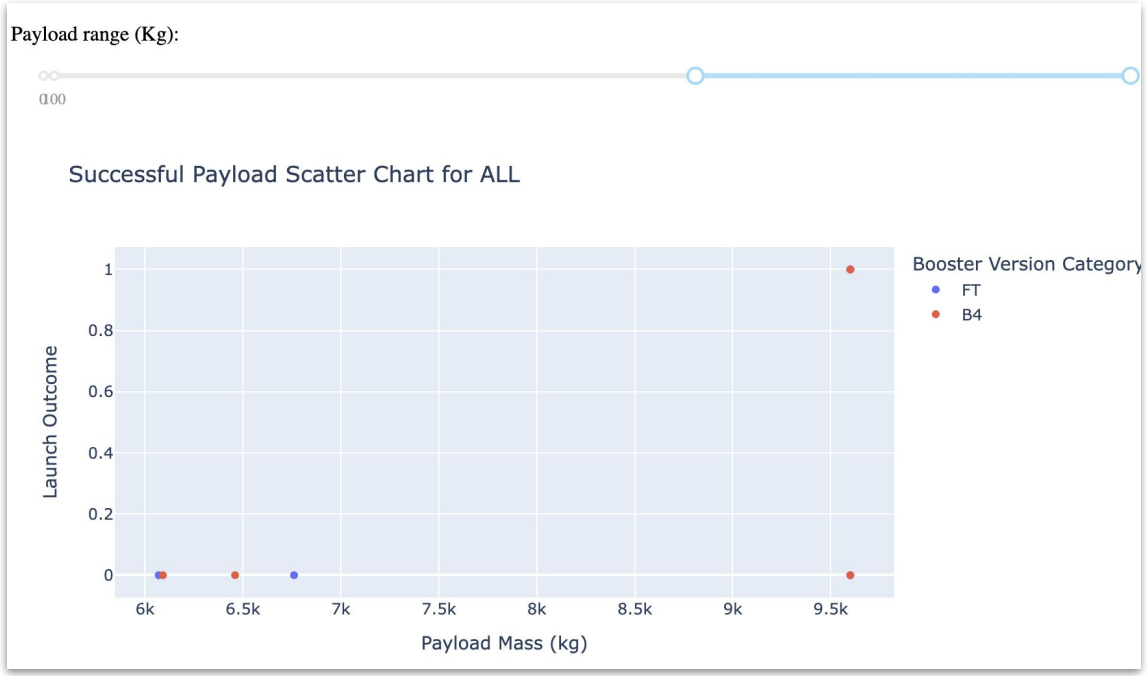


The launch site with highest launch success ratio

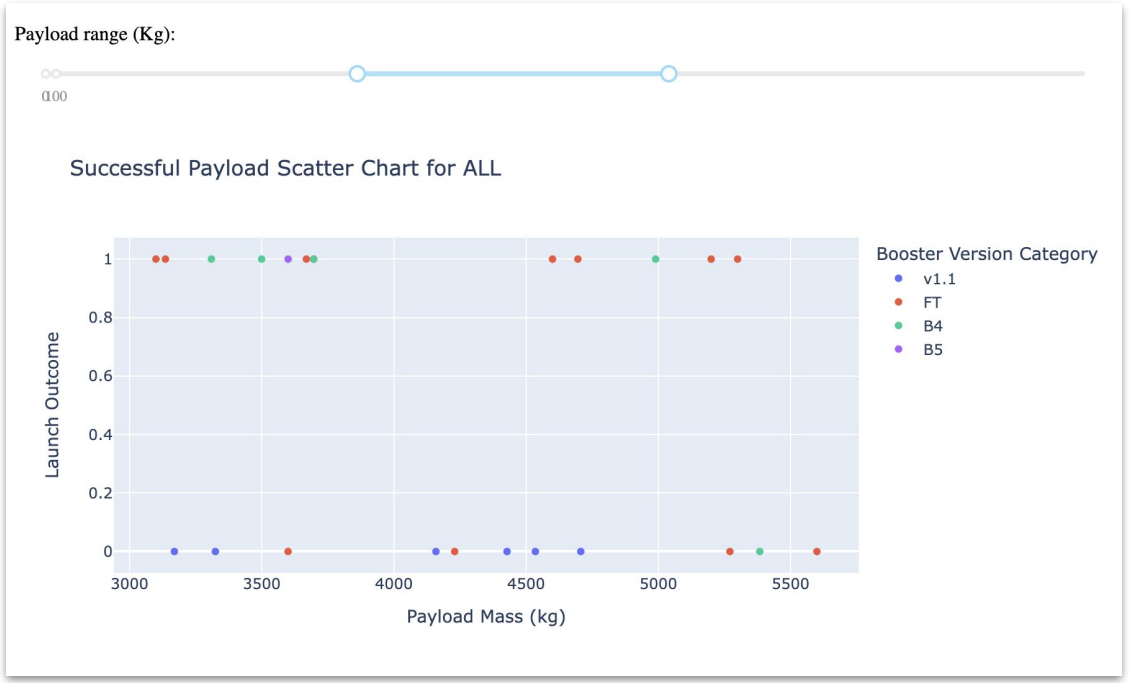


Successful payload scatter for all sites

The B4 booster is more successful for larger payloads of over 9.5kg.



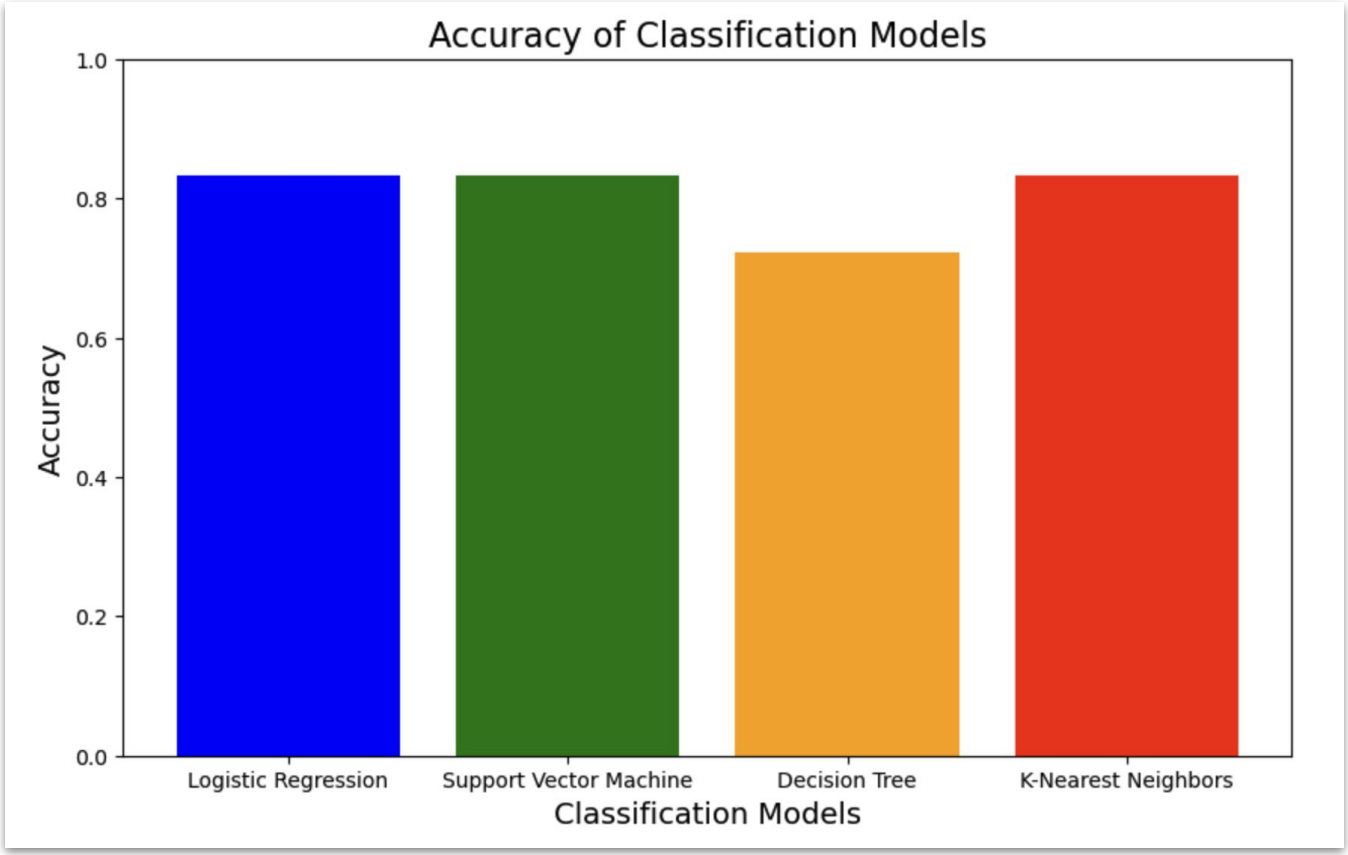
The FT is an effective booster for 4.5-5.5 kg loads



Predictive Analysis

Classification Accuracy

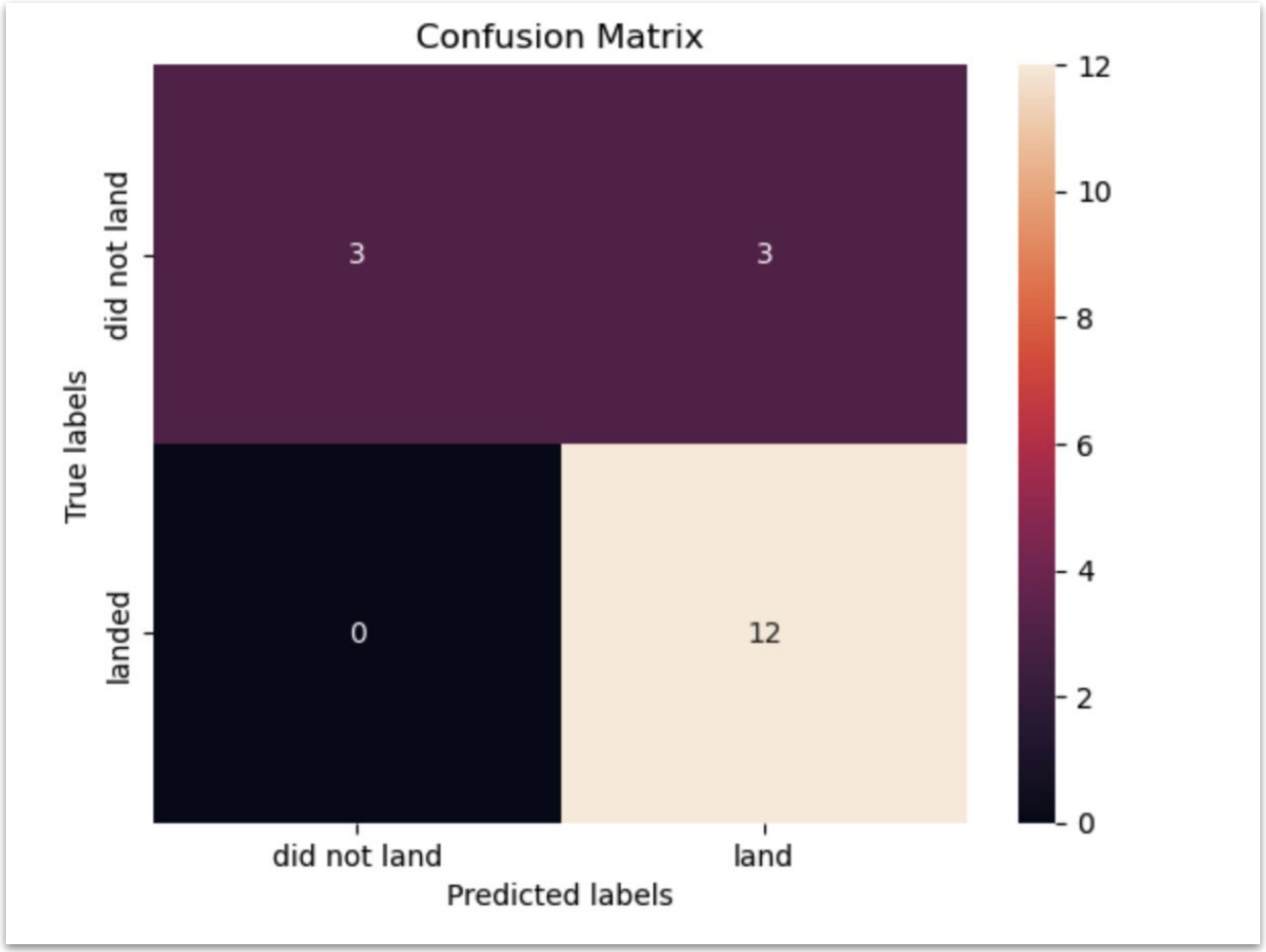
Logistic Regression, Support Vector Machine, and K-Nearest Neighbors have the highest accuracy of 0.8333333333333334.



Confusion Matrix

Given ¾ of the models performed equally, I've added confusion matrix of the KNN

- 0 False Negatives
- 3 False Positives



Conclusions

1. We found 3 models that are all accurate at determining first stage will land
2. Finding these models means we can determine the cost of a launch.
3. We must be careful before putting the models into production. Each of the models shows 3 False Positives i.e. classifying a successful landing when the ship did not land

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thanks!