IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Amy E
Coursera – Applied Data Science Capstone

June 26, 2025

Outline

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

Executive Summary

- Leveraging data from the SpaceX Falcon 9 first stage, mission success was analyzed using several data science methodologies.
- The primary methodologies leveraged included:
  - Data collection, wrangling, and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- Results indicate that some features of the rocket launches are correlated to the success or failure of the launch.
- Decision tree predictions may provide the best outcome to to predict whether the Falcon 9 first stage will land successfully.

Introduction

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
- Determining if the first stage will land helps to inform the cost of a launch.
- This information can be used if an alternate company wants to bid against Space X for a rocket launch.

Section 1

Methodology

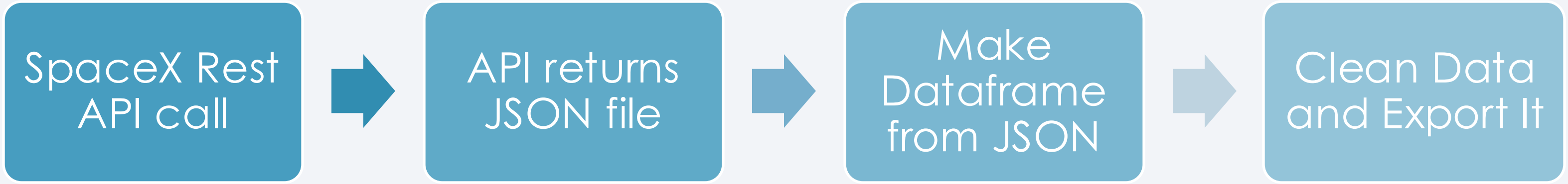
Methodology

Executive Summary

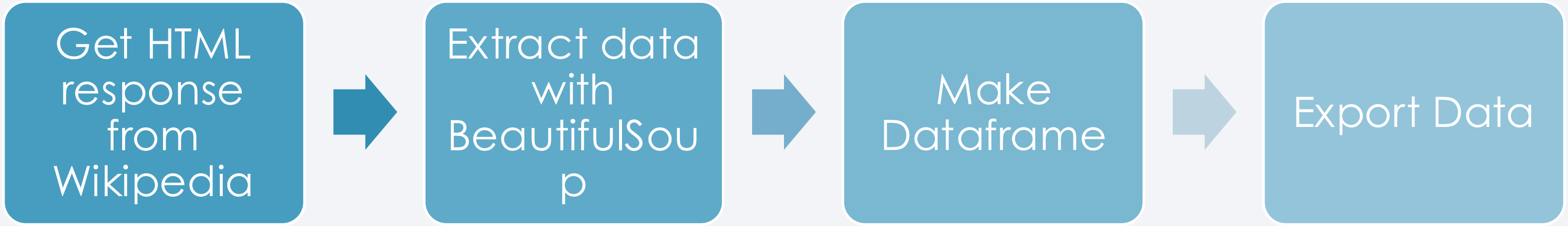
- Data collection methodology:
  - Leveraged via two public sources - SpaceX API and Web Scraping
- Perform data wrangling
  - Landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data into train and test set, use different classification algorithms, choose the best one

Data Collection

- Datasets are collected from Rest SpaceX API and web scrapping Wikipedia
  - The information obtained by the API are rocket, launches, payload information.
  - The SpaceX REST API URL is https://api.spacexdata.com/v4/launchpads/



- The information obtained by the webscrapping of Wikipedia are launches, landing and payload information
  - URL is https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922



Data Collection – SpaceX API

1. Getting Response from API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
```

2. Convert Response to JSON File

```
static_json_url = response.json()
data = pd.json_normalize(static_json_url)
```

3. Transform Data

```
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
getBoosterVersion(data)
```

4. Create dictionary with data

```
launch_dict = {'FlightNumber':
list(data['flight_number']), '
Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial, 'Longitude':
Longitude, 'Latitude': Latitude}
```

5. Create Dataframe

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

6. Filter Dataframe

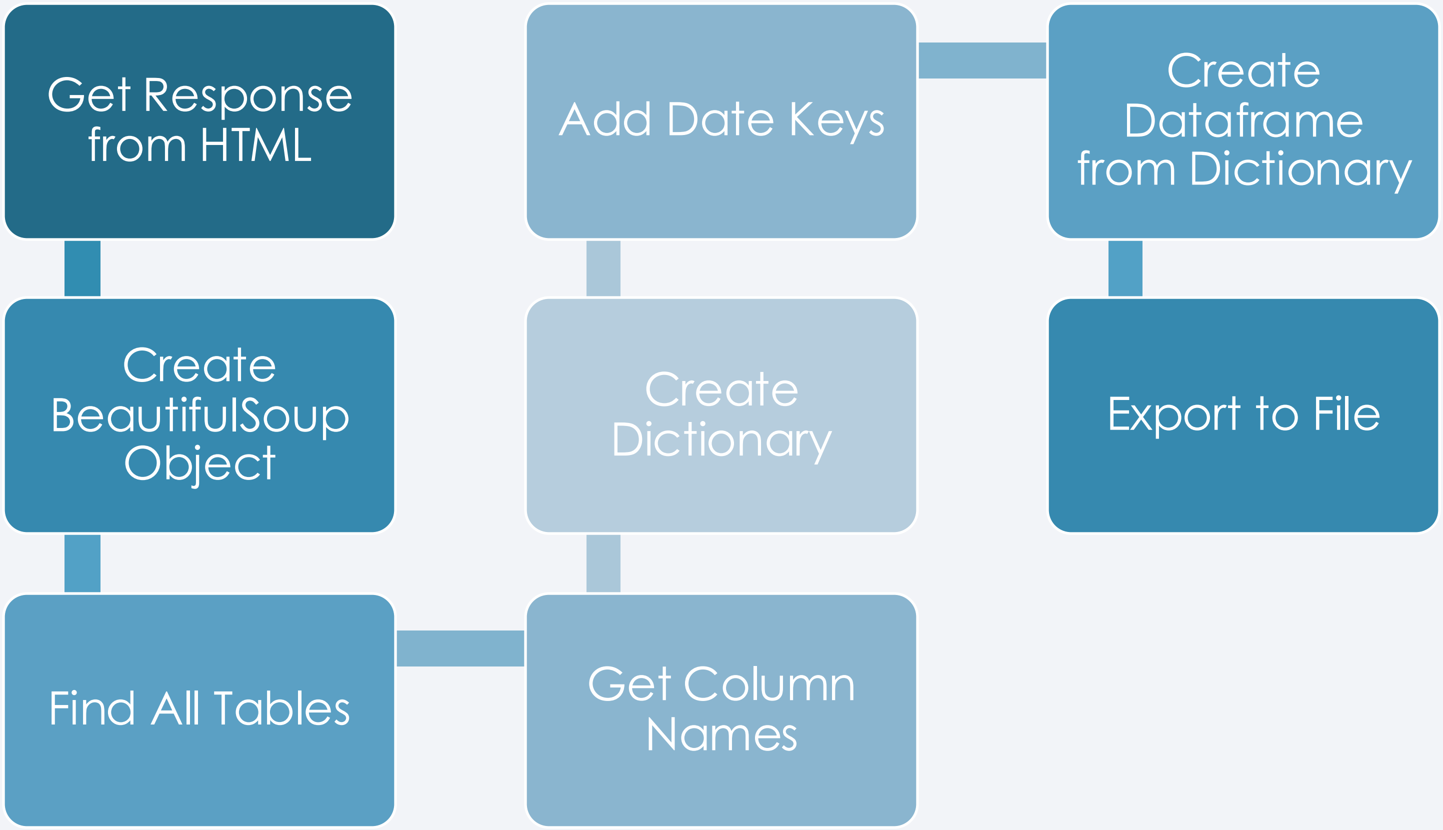
```
mask = df['BoosterVersion'] != 'Falcon 9'
data_falcon9 = df.drop(df[mask].index)
```

7. Export to file

```
data_falcon9.loc[:,'FlightNumber'] =
list(range(1, data_falcon9.shape[0]+1))
data_falcon9
```

Link to GitHub Data Collection

Data Collection - Scraping



[Link to GitHub Data Scraping]

Data Wrangling

In the dataset, there are several cases where the booster did not land successfully.

- True Ocean, True RTLS, True ASDS means the mission has been successful.
- False Ocean, False RTLS, False ASDS means the mission was a failure.

We need to transform string variables into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure.

Calculate launches number for each site → Calculate the number and occurence to each orbit → Calculate number and occurrence of mission outcome per orbit type → Create landing outcome label from outcome column → Export to file

[Link to GitHub Data Wrangling]

EDA with Data Visualization

Scatter Graphs

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload vs Launch Site
- Orbit vs Flight Number
- Payload vs Orbit Type
- Orbit vs Payload Mass

Scatter plots show relationship between variables. This relationship is called correlation

Bar Graph

- Success rate vs Orbit

Bar graphs show the relationship between numeric and categoric variables

Line Graph

- Success rate vs Year

Line graphs show data variables and their trends. Line graphs can help to show global behavior and make prediction for unseen data

[Link to GitHub Data Visualization]

EDA with SQL

By performing SQL queries to gather and understand data from dataset, the following was identified:

- Displaying the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List the names of the booster versions which have carried the maximum payload mass.
- List 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 successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

[Link to GitHub SQL]()

Build an Interactive Map with Folium

Folium map object is a map centered on NASA Johnson Space Center at Houson, Texas

- Red circle at NASA Johnson Space Center's coordinate with label showing its name (folium.Circle, folium.map.Marker).
- Red circles at each launch site coordinates with label showing launch site name (folium.Circle, folium.map.Marker, folium.features.DivIcon).
- The grouping of points in a cluster to display multiple and different information for the same coordinates (folium.plugins.MarkerCluster).
- Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing (folium.map.Marker, folium.Icon).
- Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them (folium.map.Marker, folium.PolyLine, folium.features.DivIcon)

These objects are created in order to understand better the problem and the data. We can show easily all launch sites, their surroundings and the number of successful and unsuccessful landings.

Link to GitHub Folium

Build a Dashboard with Plotly Dash

- Dashboard has dropdown, pie chart, range slider and scatter plot components
  - Dropdown allows a user to choose the launch site or all launch sites (dash_core_components.Dropdown)
  - Pie chart shows the total success and the total failure for the launch site chosen with the dropdown component (plotly.express.pie)
  - Rangeslider allows a user to select a payload mass in a fixed range (dash_core_components.RangeSlider)
  - Scatter chart shows the relationship between two variables, in particular Success vs Payload Mass (plotly.express.scatter)

Link to GitHub Plotly Dash

Predictive Analysis (Classification)

- Data preparation
  - Load dataset
  - Normalize data
  - Split data into training and test sets.
- Model preparation
  - Selection of machine learning algorithms
  - Set parameters for each algorithm to GridSearchCV
  - Training GridSearchModel models with training dataset
- Model evaluation
  - Get best hyperparameters for each type of model
  - Compute accuracy for each model with test dataset
  - Plot Confusion Matrix
- Model comparison
  - Comparison of models according to their accuracy
  - The model with the best accuracy will be chosen (see Notebook for result)

Link to GitHub Prediction Analysis

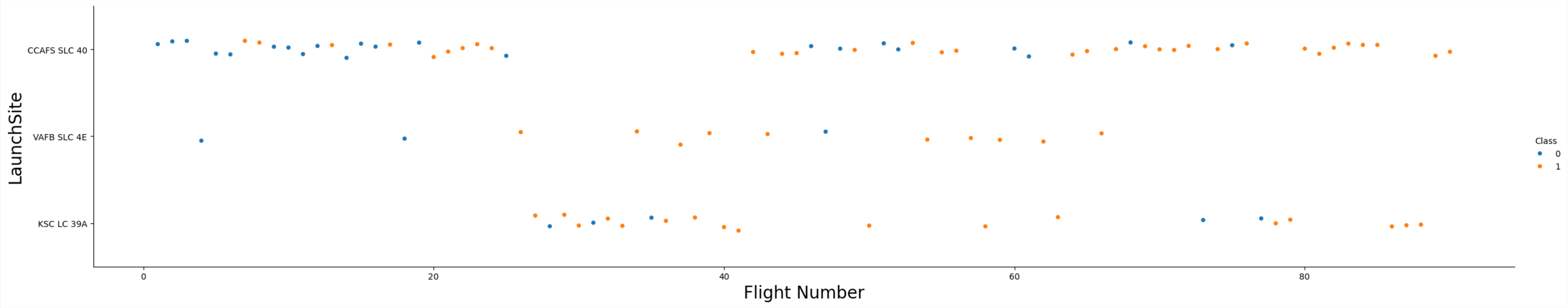
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

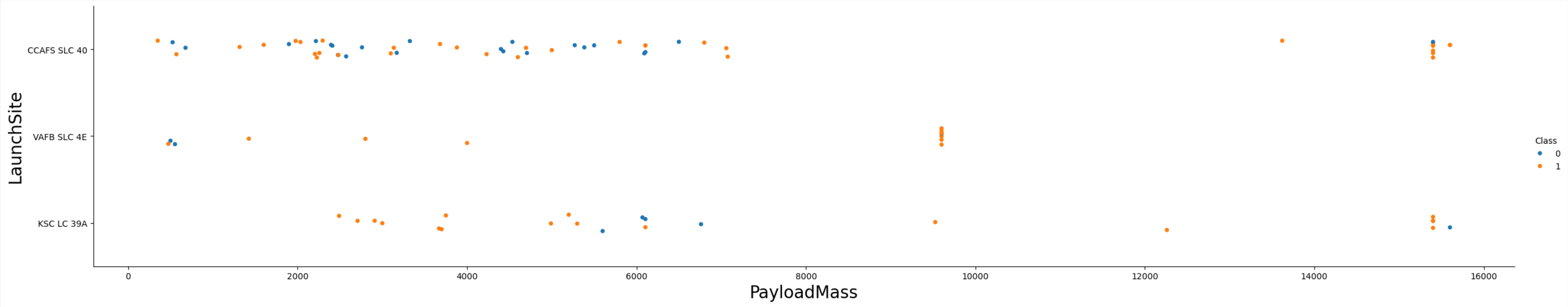
Insights drawn from EDA

Flight Number vs. Launch Site



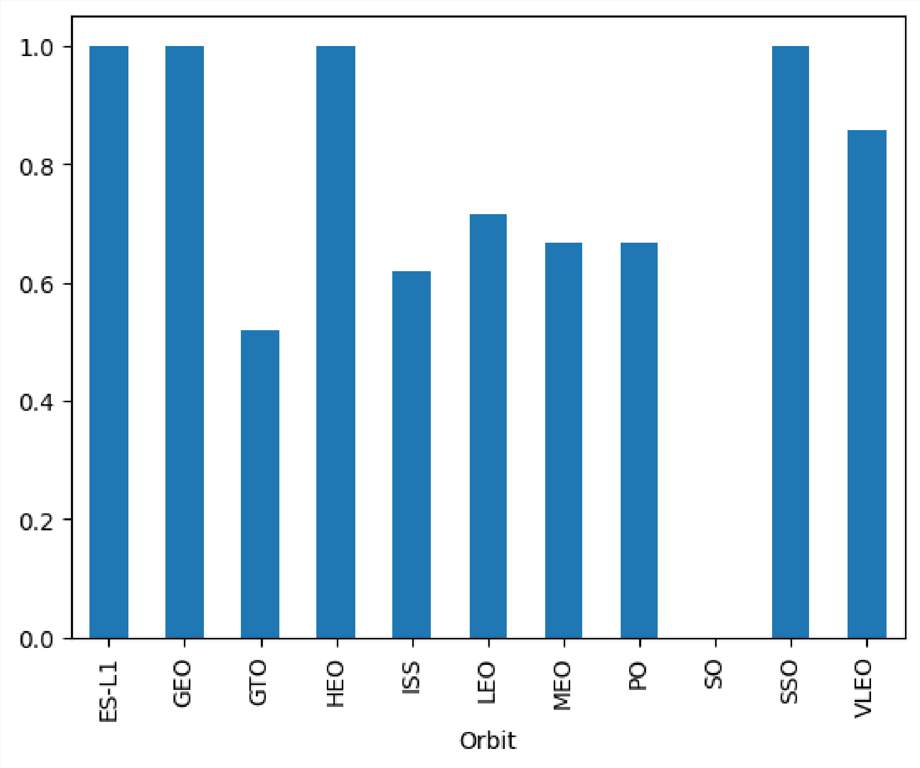
For each site, the success rate is increasing

Payload vs. Launch Site



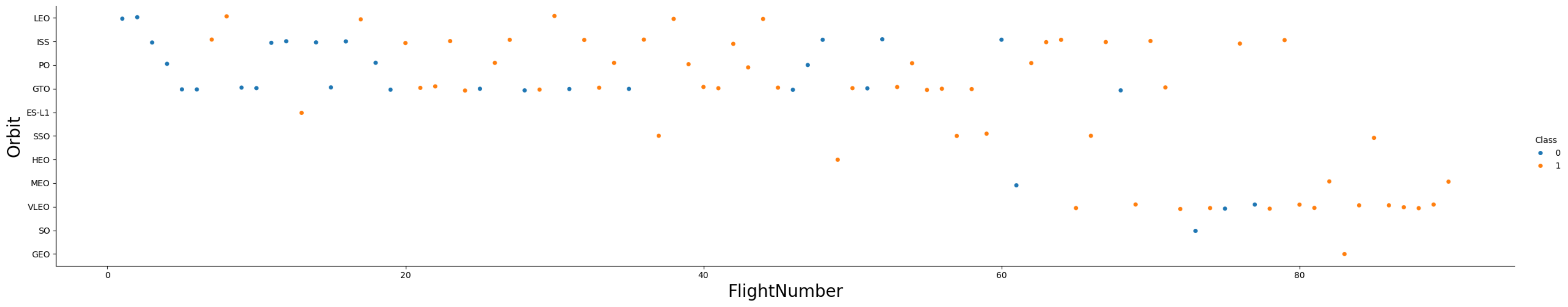
Depending on the launch site, a heavier payload may be a consideration for a successful landing. On the other hand, a too heavy payload can make a landing fall.

Success Rate vs. Orbit Type



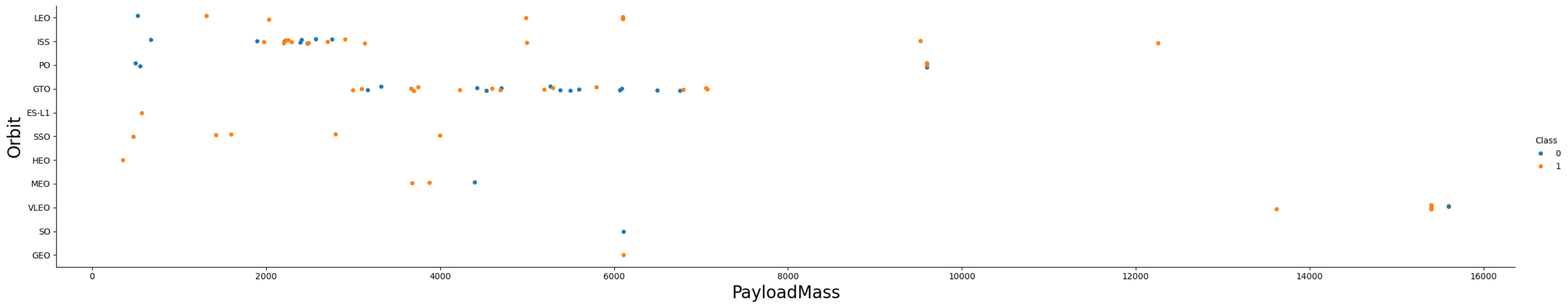
With this plot, we can see success rate for different orbit types. We note that ES-L1, GEO, HEO, SSO have the best success rate.

Flight Number vs. Orbit Type



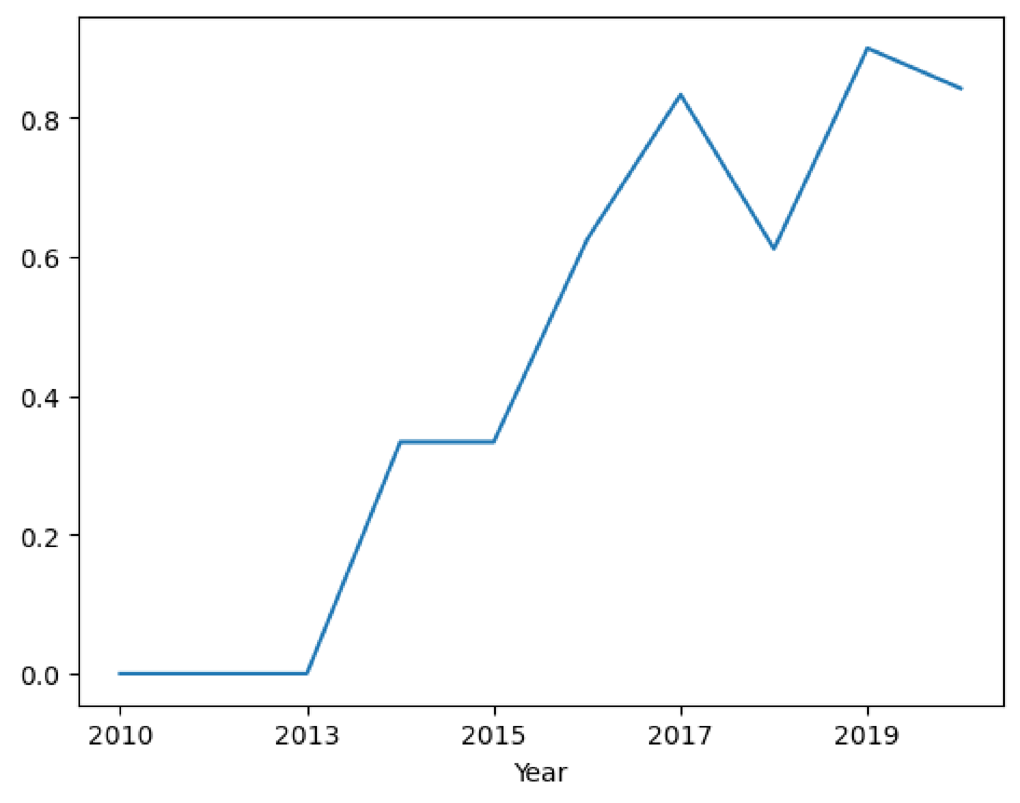
We notice that the success rate increases with the number of flights for the LEO orbit. For some orbits like GTO, there is no relation between the success rate and the number of flights. But we can suppose that the high success rate of some orbits like SSO or HEO is due to the knowledge learned during former launches for other orbits.

Payload vs. Orbit Type



The weight of the payloads can have a great influence on the success rate of the launches in certain orbits. For example, heavier payloads improve the success rate for the LEO orbit. Another finding is that decreasing the payload weight for a GTO orbit improves the success of a launch.

Launch Success Yearly Trend



Since 2013, we can see an increase in the Space X Rocket success rate.

All Launch Site Names

SQL Query

```
sql SELECT DISTINCT LAUNCH_SITE from SPACEXTBL order by 1
```

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

Explanation
The use of DISTINCT in the query allows to remove duplicate Launch_Site

Launch Site Names Begin with 'CCA'

SQL Query

SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5

Results

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
08-12-2010	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
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Explanation

The WHERE clause followed by LIKE clause filters launch sites that contain the substring CCA. LIMIT 5 shows 5 records from filtering.

Total Payload Mass

SQL Query

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE
WHERE "Customer" = 'NASA (CRS)';
```

Results

SUM("PAYLOAD_MASS__KG_")
45596

Explanation

This query returns the sum of all payload masses where the customer is NASA (CRS)

Average Payload Mass by F9 v1.1

SQL Query

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE
WHERE "Booster_Version" = 'F9 v1.1';
```

Results

AVG("PAYLOAD_MASS__KG_")
2928.4

Explanation

This query returns the average of all payload masses where the booster version contains the substring F9 v1.1

First Successful Ground Landing Date

SQL Query

```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE
"Landing_Outcome" = 'Success (ground pad)';
```

Results

MIN("Date")
2015-12-22

Explanation

With this query, the oldest successful landing is selected. The WHERE clause filters dataset in order to keep only records where landing was successful. With the MIN function, we select the record with the oldest date.

Successful Drone Ship Landing

with Payload between 4000 and 6000

SQL Query

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE
"Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" >
4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

Results

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

Explanation

This query returns the booster version where landing was successful and payload mass is between 4000 and 6000 kg. The WHERE and AND clauses filter the dataset.

Total Number of Successful and Failure Mission Outcomes

SQL Query

```
%sql SELECT "Mission_Outcome", COUNT(*) AS "Total" FROM
SPACEXTABLE WHERE "Mission_Outcome" IN ('Success', 'Failure')
GROUP BY "Mission_Outcome";
```

Results

Mission_Outcome	Total
Success	98

Explanation
With the first SELECT, we show the subqueries that return results. The first subquery counts the successful mission. The second subquery counts the unsuccessful mission. The WHERE clause followed by LIKE clause filters mission outcome. The COUNT function counts records filtered.

Boosters Carried Maximum Payload

SQL Query

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE
"PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM
SPACEXTABLE);
```

Explanation

We used a subquery to filter data by returning only the heaviest payload mass with MAX function. The main query uses subquery results and returns unique booster version (SELECT DISTINCT) with the heaviest payload mass.

Results

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 Query

```
%%sql

SELECT
    CASE
        WHEN substr("Date", 6, 2) = '01' THEN 'January'
        WHEN substr("Date", 6, 2) = '02' THEN 'February'
        WHEN substr("Date", 6, 2) = '03' THEN 'March'
        WHEN substr("Date", 6, 2) = '04' THEN 'April'
        WHEN substr("Date", 6, 2) = '05' THEN 'May'
        WHEN substr("Date", 6, 2) = '06' THEN 'June'
        WHEN substr("Date", 6, 2) = '07' THEN 'July'
        WHEN substr("Date", 6, 2) = '08' THEN 'August'
        WHEN substr("Date", 6, 2) = '09' THEN 'September'
        WHEN substr("Date", 6, 2) = '10' THEN 'October'
        WHEN substr("Date", 6, 2) = '11' THEN 'November'
        WHEN substr("Date", 6, 2) = '12' THEN 'December'
        ELSE 'Unknown'
    END AS "Month_Name",
    "Mission_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM
    SPACEXTABLE
WHERE
    substr("Date", 0, 5) = '2015';
```

Results

Month_Name	Mission_Outcome	Booster_Version	Launch_Site
January	Success	F9 v1.1 B1012	CCAFS LC-40
February	Success	F9 v1.1 B1013	CCAFS LC-40
March	Success	F9 v1.1 B1014	CCAFS LC-40
April	Success	F9 v1.1 B1015	CCAFS LC-40
April	Success	F9 v1.1 B1016	CCAFS LC-40
June	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
December	Success	F9 FT B1019	CCAFS LC-40

Explanation

This query returns month, booster version, launch site where landing was unsuccessful and landing date took place in 2015. Substr function process date in order to take month or year. Substr(DATE, 4, 2) shows month. Substr(DATE,7, 4) shows year.

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

SQL Query

```
%%sql

SELECT
    "Landing_Outcome",
    COUNT(*) AS "Count"
FROM
    SPACEXTABLE
WHERE
    "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY
    "Landing_Outcome"
ORDER BY
    COUNT(*) DESC;
```

Results

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

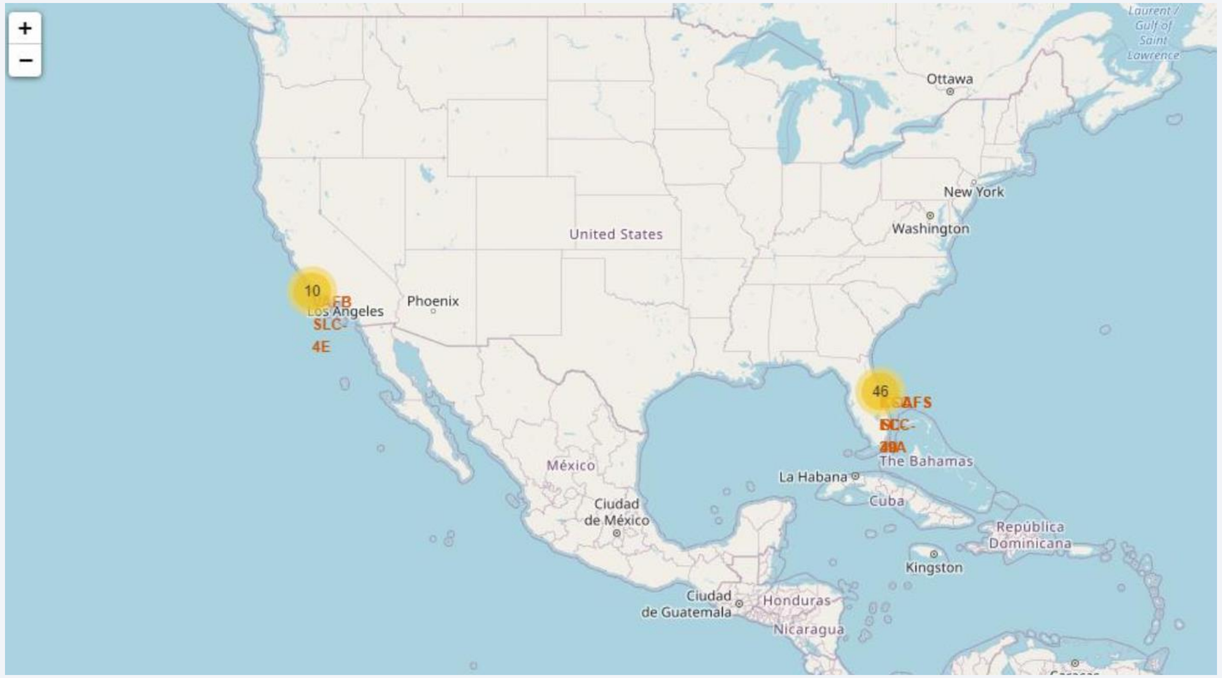
Explanation

This query returns landing outcomes and their count where mission was successful and date is between 04/06/2010 and 20/03/2017. The GROUP BY clause groups results by landing outcome and ORDER BY COUNT DESC shows results in decreasing order.

Section 3

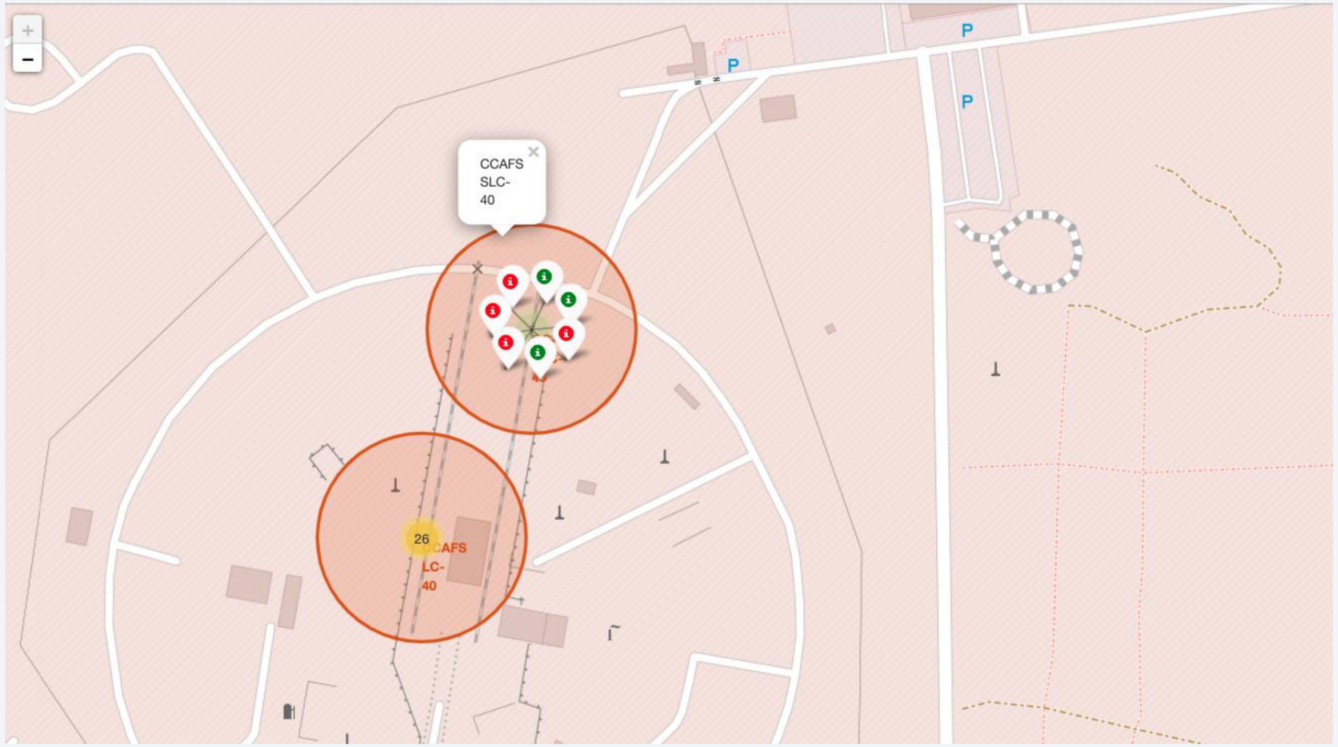
Launch Sites Proximities Analysis

Map | SpaceX Launch Sites



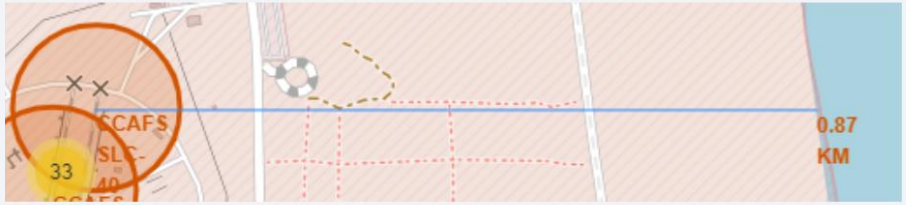
SpaceX launch sites are located on the US coasts

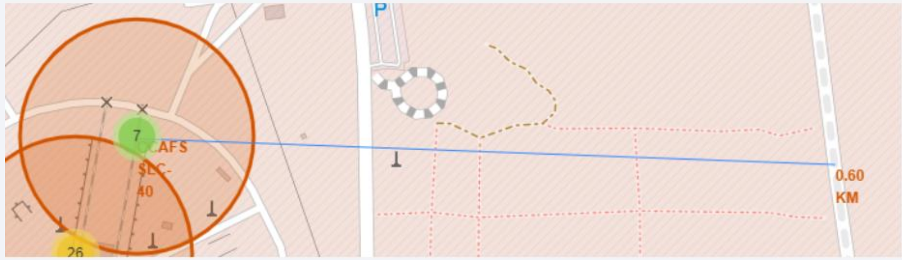
Map | Color Labeled Markers

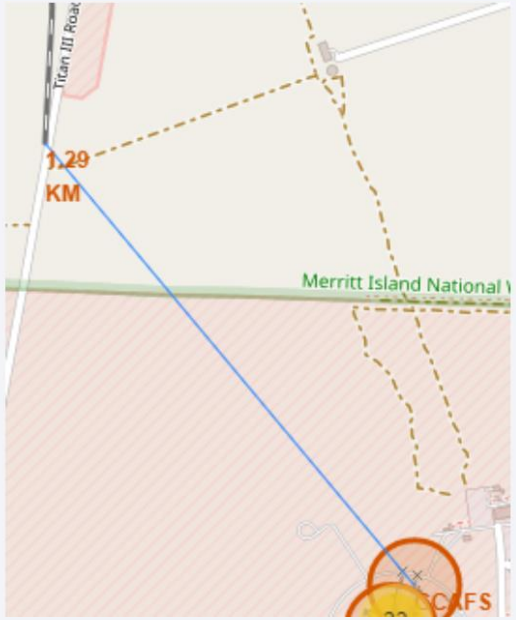


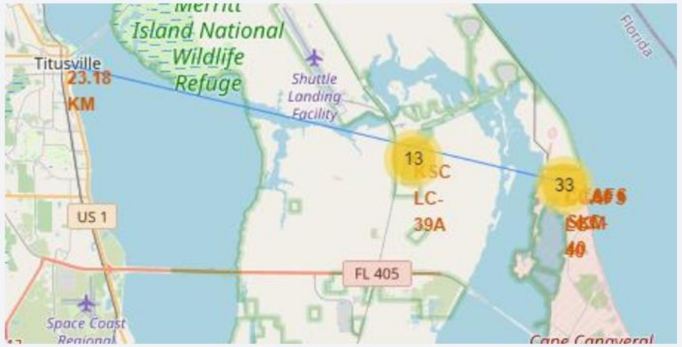
Green marker represents successful launches. Red marker represents unsuccessful launches. We note that KSC LC-39A has a higher launch success rate.

Map | Distances between CCAFS SLC-40 and its proximities







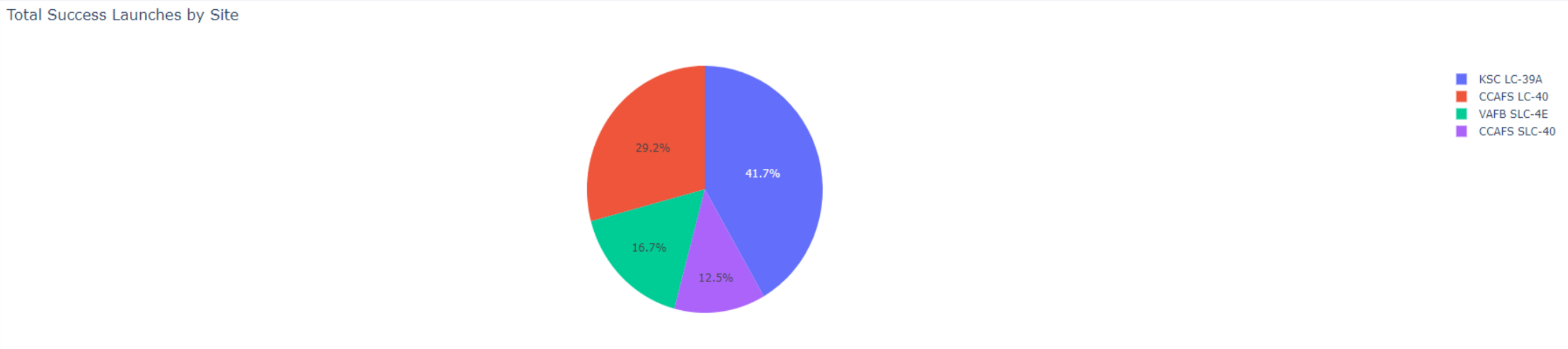


Is CCAFS SLC-40 in close proximity to railways ? Yes
Is CCAFS SLC-40 in close proximity to highways ? Yes
Is CCAFS SLC-40 in close proximity to coastline ? Yes
Do CCAFS SLC-40 keeps certain distance away from cities ? No

Section 4

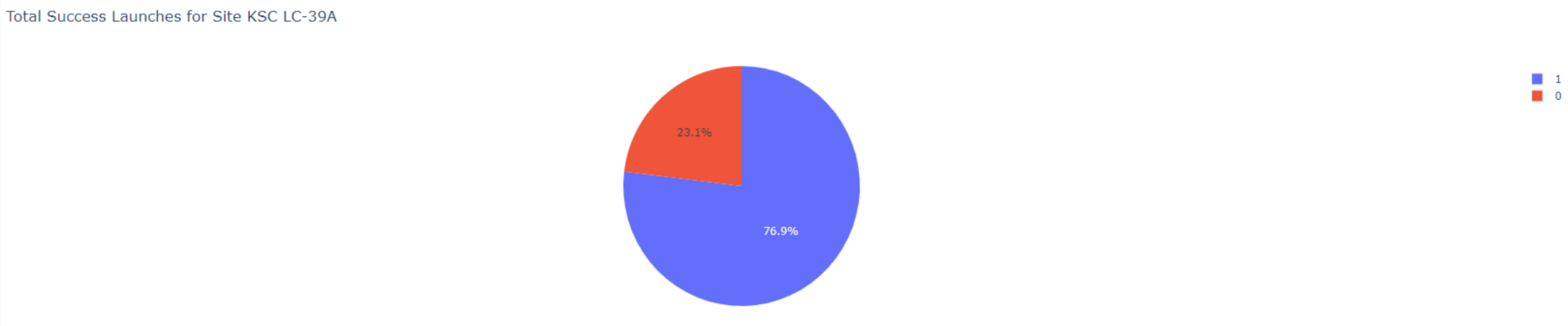
Build a Dashboard with Plotly Dash

Dashboard | Total Success by Site



KSC LC-39A has the best success rate of launches.

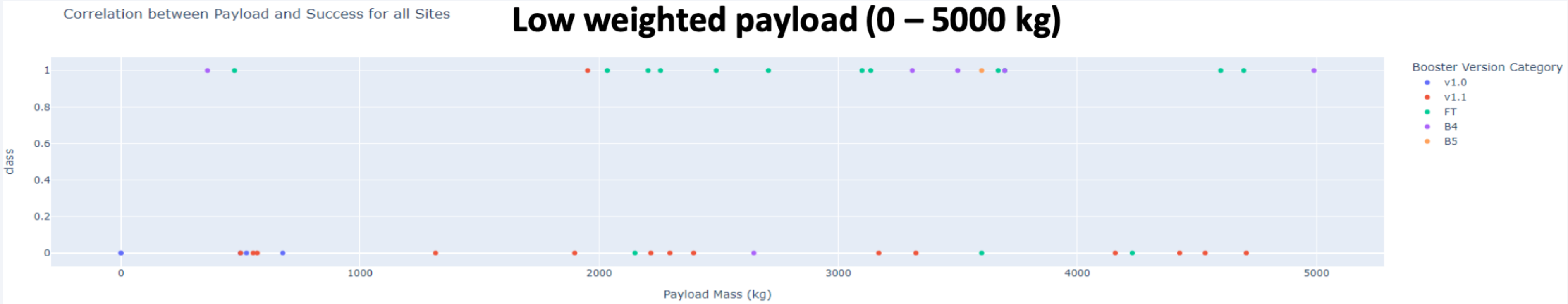
Dashboard | Total Success Launches for Site KSC LC-39A

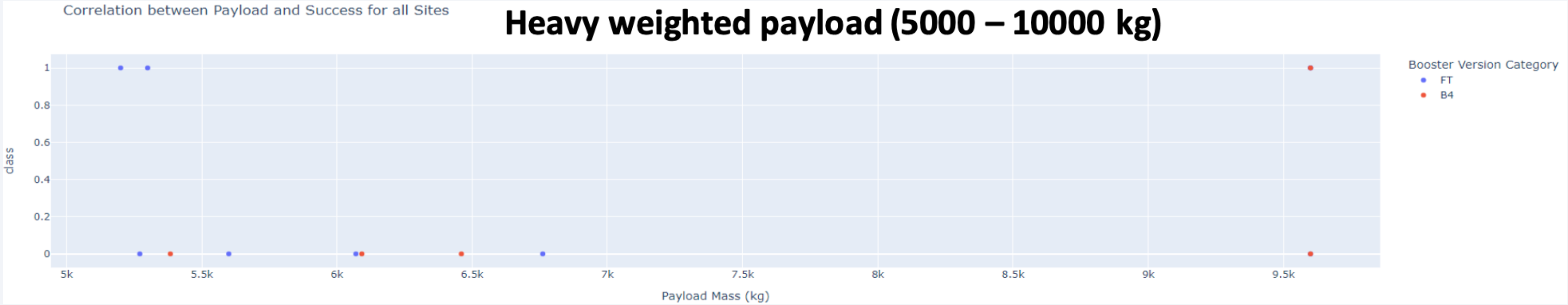


KSC LC-39A has achieved a 76.9% success rate while getting a 23.1% failure rate.

Dashboard | Payload Mass vs Outcome

for all sites with different payload mass selected





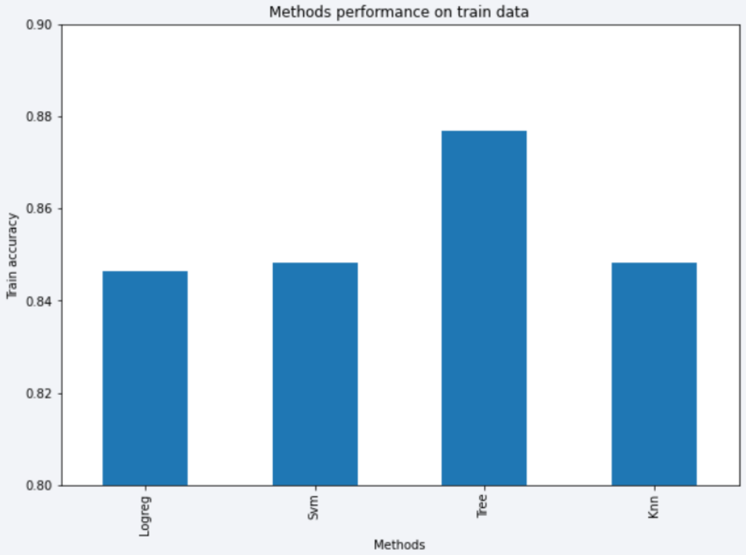
Low weighted payloads have a better success rate than the heavy weighted payloads.

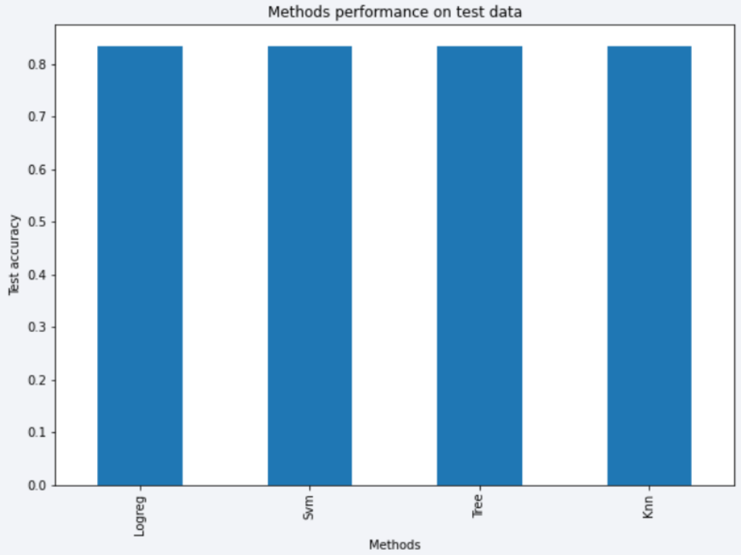
Section 5

Predictive Analysis (Classification)

Classification Accuracy

	Accuracy Train	Accuracy Test
Tree	0.876786	0.833333
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333





For accuracy test, all methods performed similar. More test data should be sought prior to selecting one, but if one must be selected now, then the decision tree should be selected.

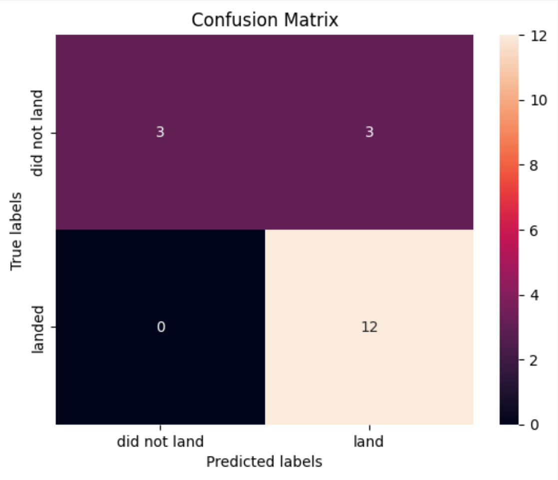
Decision tree best parameters

```
tuned hyperparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 2, 'splitter': 'random'}
```

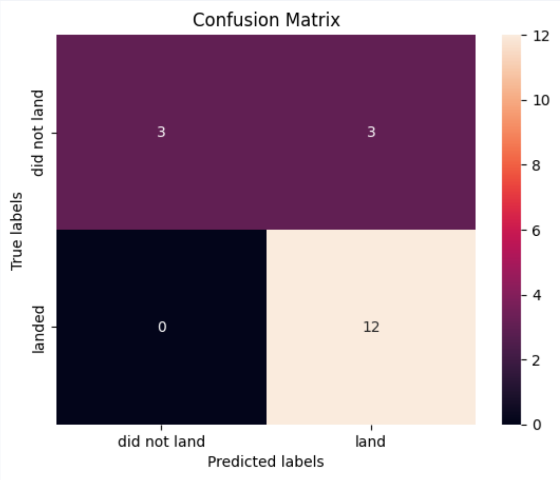
Confusion Matrix

As the test accuracy are all equal, the confusion matrices are also identical. The main problem of these models are false positives.

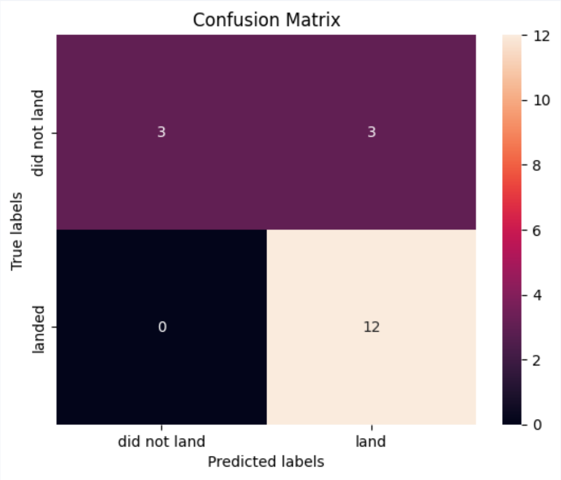
Logistic Regression



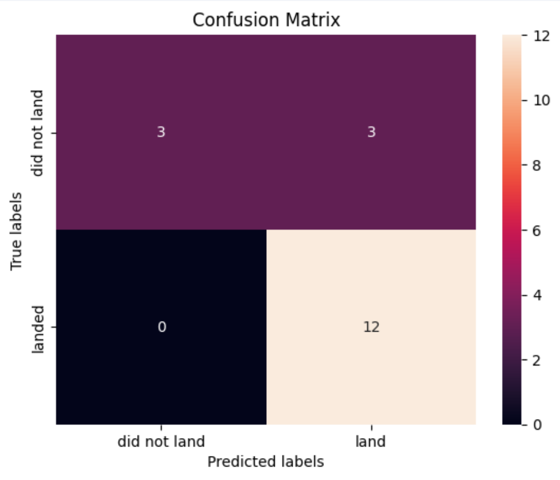
Decision Tree



kNN



SVM



Conclusions

- Mission success is influenced by several key factors, including the launch site, the target orbit, and most notably, the number of previous launches conducted.
- A pattern of improved outcomes suggests that knowledge and experience gained from earlier attempts contribute significantly to subsequent mission success.
- The orbits associated with the highest success rates in this dataset are Geostationary Earth Orbit (GEO), Highly Elliptical Orbit (HEO), Sun-Synchronous Orbit (SSO), and Earth-Sun Lagrange Point 1 (ES-L1).
- Payload mass appears to be a contributing factor to mission success, varying by orbit type; lighter payloads generally demonstrate higher success rates compared to heavier ones.
- Although KSC LC-39A emerges as the most successful launch site, current data does not provide sufficient explanation for site-specific performance differences; additional data—such as atmospheric conditions—may be necessary to clarify these trends.

Thank you!