IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shivangi Verma
01-01-2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies

  Data Collection via API and web scraping

  Data Wrangling

  Exploratory Data Analysis using SQL

  Exploratory Data Analysis with data visualization library like Matplotlib, Folium

  Interactive Dashboard with plotly

  Model Building and training

- Summary of all results

  Useful insights from the data with the help of EDA

  Interactive visuals with the help of data visualization

  Predictive Analysis using Machine Learning

Introduction

- Project background and context

This project aims to predict whether the first stage in the rocket launching process will be successful or not. To achieve this, we predict the successful landing of Falcon 9 rockets. As per the company, SpaceX, the Falcon 9 rocket lunch caused 62 million dollars which is even lesser than the half of what other providers cost. By predicting the successful launch of this stage, We can determine the cost of a launch for SpaceX.

- Problems you want to find answers

How each feature(column) in the dataset is related with successful launching?

What causes failed landing and what causes successful landing?

What are the favourable factors that results in successful landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX RESTAPI
  - Web Scraping
- Perform data wrangling
  - Removed unnecessary rows and columns from the dataset
  - One hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Splited train and test data, Used different classification ML algorithm

Data Collection

The data collection was done from two ways

1. Some part of data was fetched using SpaceX Rest API. The ur is api.spacexdata.com/v4

SPACEX REST API call → JSON file returned by API → Convert Json to DataFrame → Clean data and export it

2. Some part of data was fetched by web scraping using Beautiful Soup Python library

HTML response from wikipedia → Extract data with BeautifulSoup → Create a Dataframe → Clean data and export it

Data Collection – SpaceX API

Step 1 : Getting Response from SpaceX API

Step 2 : Convert the response to json() and convert into dataframe using json_normalize()

Step 3: Applying some custom function to get more data

Step 4: Creating a dictionary from the gathered data

Step 5 : Convert the dictionary to pandas dataframe

Step 6 :Keep only Falcon 9 rocket launch data. Drop the other rows

Step 7: Replace null values of payload mass with its mean

Step 8 : Export to CSV file

GitHub URL : Data Collection API

Data Collection - Scraping

Step 1 : Getting response from HTML

Step 2 : Creating a BeautifulSoup Object

Step 3 : Getting Response from SpaceX API

Step 4 : Extracting column names from the HTML table

Step 5 : Create dictionary

Step 6 : Convert the dictionary into a dataframe

Step 7 : Export it to CSV

GitHub URL Web Scraping

Data Wrangling

The dataset has several cases where the booster did not land successfully. Sometimes a landing of the attempted but failed due to an accident. For example,

True Ocean, True RTLS, True ASDS means the mission has been successful.

False Ocean, False RTLS, False ASDS means the mission has been unsuccessful

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

[GitHub URL Data Wrangling]

Calculating number of launches for each site

Calculating number and occurrence of each orbit

Calculate number and occurrence of mission outcome per orbit type

Creating Landing outcome from outcome column

Export to CSV

EDA with Data Visualization

We have plotted the following charts.

Scatter Plot:

Scatter plot shows the correlation between two variables. We have plotted scatter plot for the below variables.

- 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

Bar Chart:

It shows the relationship between numeric and categoric values. We have plotted the bar graph for the following variable.

- Success rate vs. Orbit type

Line Graph:

Line graph is used to trend in the data.
We have plotted the line chart for the following column.

- Success rate vs. Year

[GitHub URL data Viz]

EDA with SQL

Performed following SQL queries to get better understanding of the data

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

GitHub URL EDA SQL

Build an Interactive Map with Folium

Markers of all Launch Sites: -

Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

GitHub URL Folium

Build a Dashboard with Plotly Dash

Dashboard has dropdown, pie chart, rangeslider 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).

GitHub URL SpaceX dash App

Predictive Analysis (Classification)

- Data Preparation

  Load data

  Normalize data

  Split the data into train and test

- Model Building and Training

  select classification models like Logistic Regression, Decision Tree, KNN

  Hypertune the model by finding the best estimator using GridSearchCV

  Train the Grid Search model with traiing data

- Model Evaluation

  Find Accuracy score for each Classification model

  Plot Confusion Matrix

- Model Comparison

  Compare the accuracy of the different models and choose the model with highest accuracy

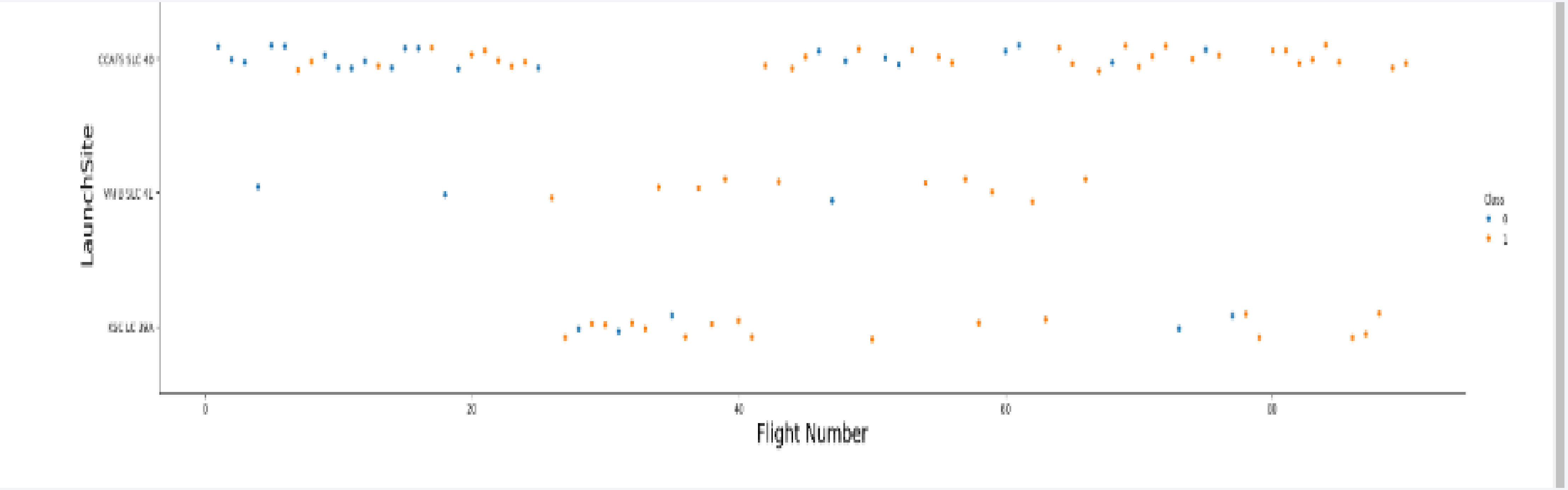
Results

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

Section 2

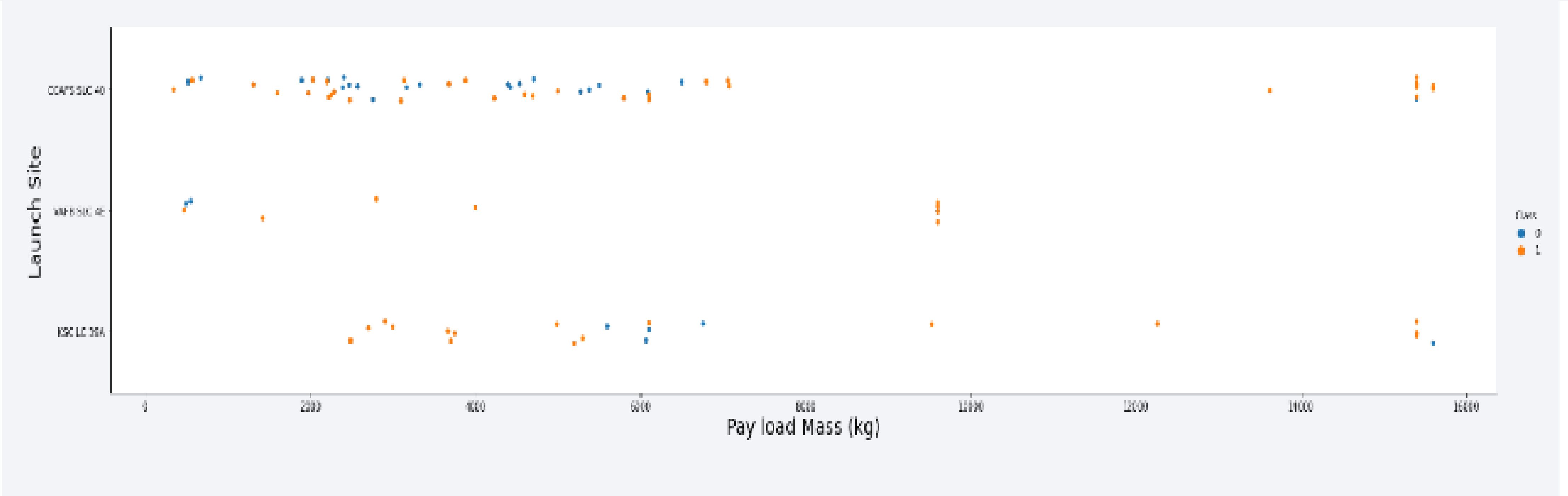
Insights drawn from EDA

Flight Number vs. Launch Site



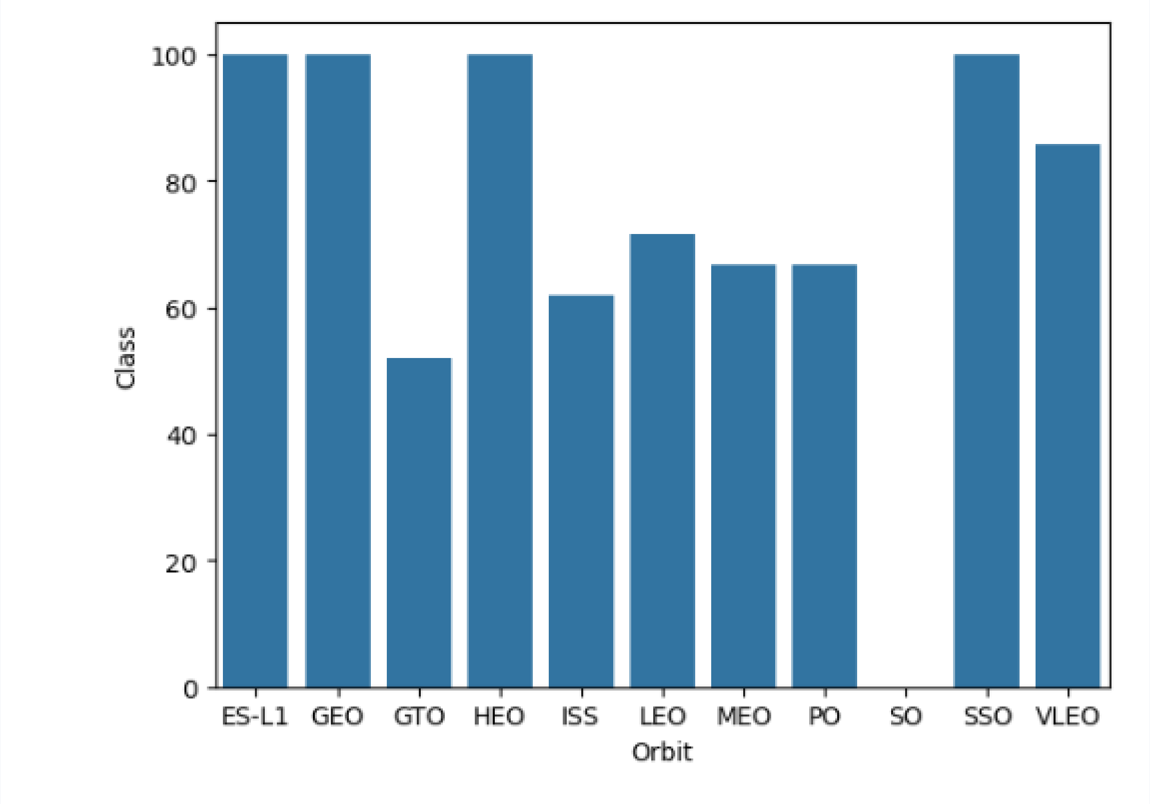
For each site, the success rate is increasing

Payload vs. Launch Site



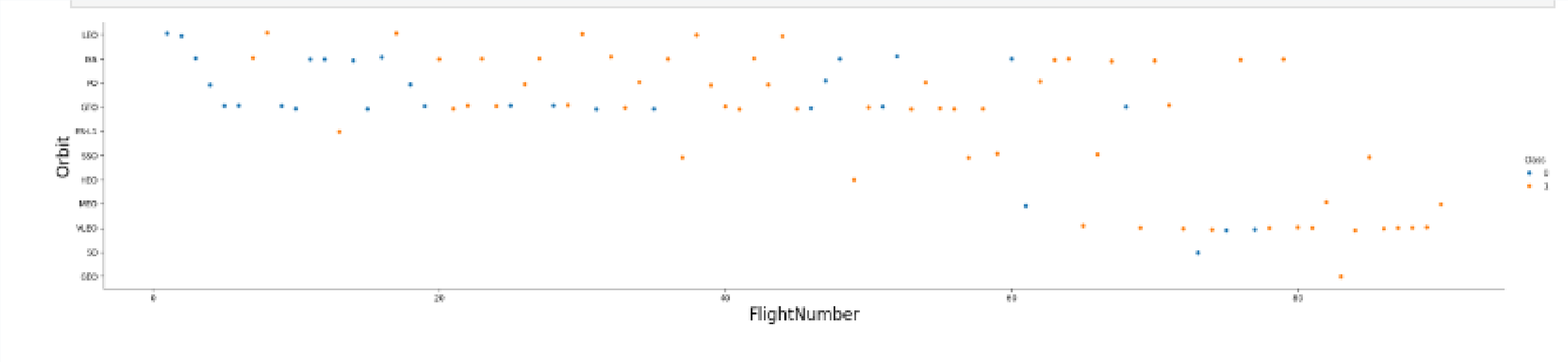
- Depending upon the launch site, a heavier payload may be a consideration for a successful landing. However, too much load can lead to a failed landing

Success Rate vs. Orbit Type



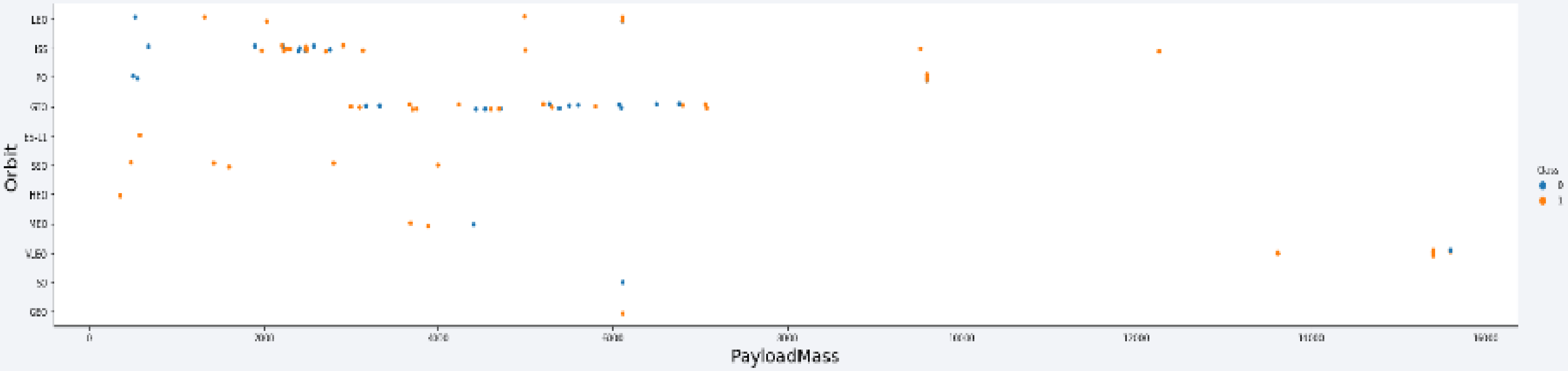
- From the above chart, ES-L1,GEO,HEO and SSO have best success rates.

Flight Number vs. Orbit Type



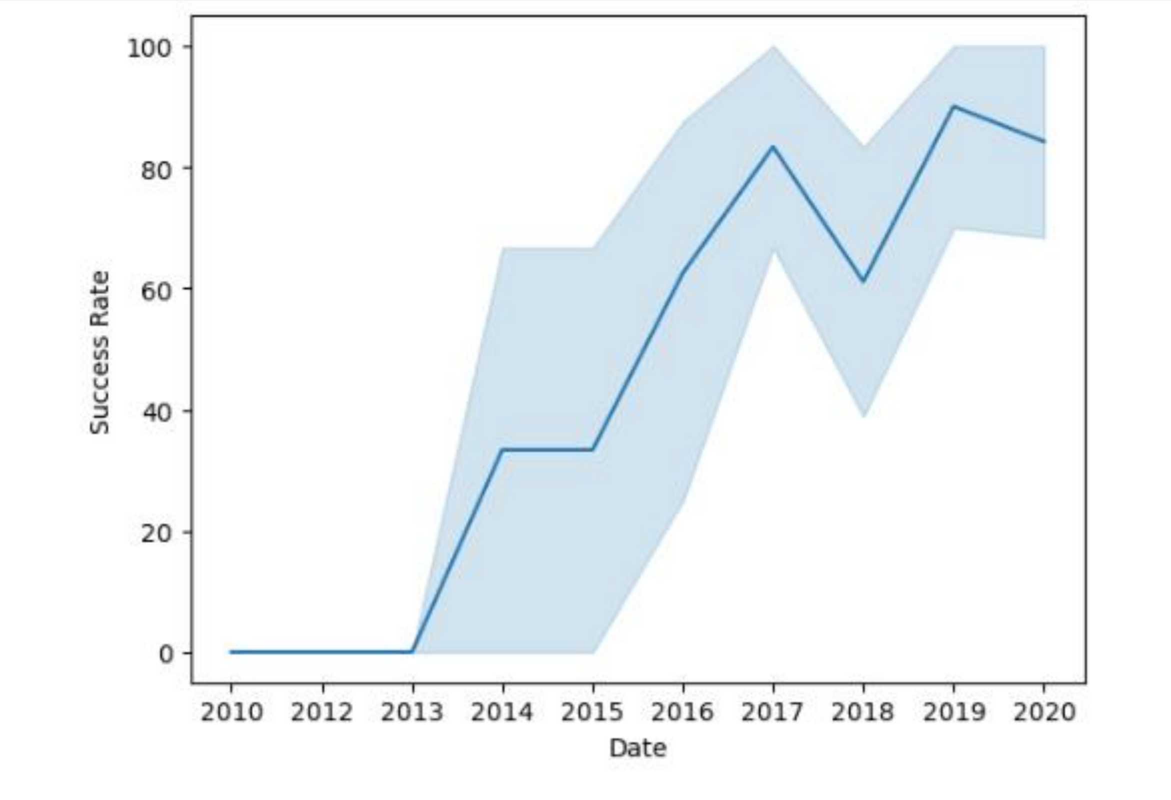
- Success rate increases with the number of flights in the LEO orbit. For some orbits like GTO, there is no relation between the success rates and number of flights.

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



- We can observe, the success rate kept increasing after 2013

All Launch Site Names

- Query

  Select distinct launch_site from SPACEXTABLE

Out[8]:

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

Distinct Keywords allows to fetch unique records only.

Launch Site Names Begin with 'CCA'

Query :
Select * from SPACEXTABLE where launch_site like ‘CCA%’ LIMIT 5

Out[9]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Where clause filters the dataset, Like clause helps to match pattern and LIMIT keyword limit the number of rows

Total Payload Mass

Query:

Select sum(payload_mass_kg_) from SPACEXTABLE where customer = ‘NASA (CRS)’

```
In [10]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where customer='NASA (CRS)'

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

Out[10]:

sum(PAYLOAD_MASS__KG_)
45596

SUM() is an aggregate function that returns the sum of a column.

Average Payload Mass by F9 v1.1

Query:

Select avg(payload_mass_kg_) from SPACEXTABLE where booster_version='F9 v1.1'

```
In [11]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where booster_version='F9 v1.1'

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

Out[11]:

avg(PAYLOAD_MASS__KG_)
2928.4

AVG() is an aggregate function that returns the average of a column.

First Successful Ground Landing Date

Query:

Select min(date) from SPACEXTABLE where mission_outcome=‘Success

```
In [12]: %sql select min(date) from SPACEXTABLE where mission_outcome ='Success'

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

Out[12]:

min(date)
2010-06-04

Explanation:

Min() is an aggregate function. It returns the minimum value present in the column

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query

select booster_version from SPACEXTABLE where (mission_outcome like ‘Success’) AND (payload_mass__kg_ between 4000 and 6000) AND (landing_outcome like ‘Success (drone ship)’)

Explanation:

Here, In this query we have used three conditions for mission outcome, payload mass kg and landing outcome. The AND keyword returns true if all the conditions are met

Out[13]:

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

Query:

Select mission_outcome, count(*) as count from spacextable GROUP BY mission_outcome ORDER BY mission_outcome

Done.

ut[50]:

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Explanation:

The data is first grouped by mission_outcome and then no of rows present in each mission_outcome is returned.

Boosters Carried Maximum Payload

Query:

select booster_version from SPACEXTABLE where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTABLE)

Explanation:

Here, we have used a subquery which returns the maximum payload mass. The query filters the data based on the payload massvalue returned by the subquery

Done.

Out[41]:

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

Query:

select substr(DATE,6,2) as Month, landing_outcome, booster_version, launch_site from SPACEXTABLE where landing_outcome = 'Failure (drone ship)'and substr(date,0,5)='2015'

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

Out[24]:

Month	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

Explanation:

To extract the month and year from the date, we have used substr() function. The query filters the data by the landing outcome and year

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

Query:

select landing_outcome, count(*) as count from SPACEXTABLE where Date **>=** '2010-06-04' AND Date **<=** '2017-03-20' GROUP by landing_outcome ORDER BY count Desc

Explanation:

The data is grouped by the landing_outcome to return the count of each landing outcome between the given date. At last it is sorted according to the count

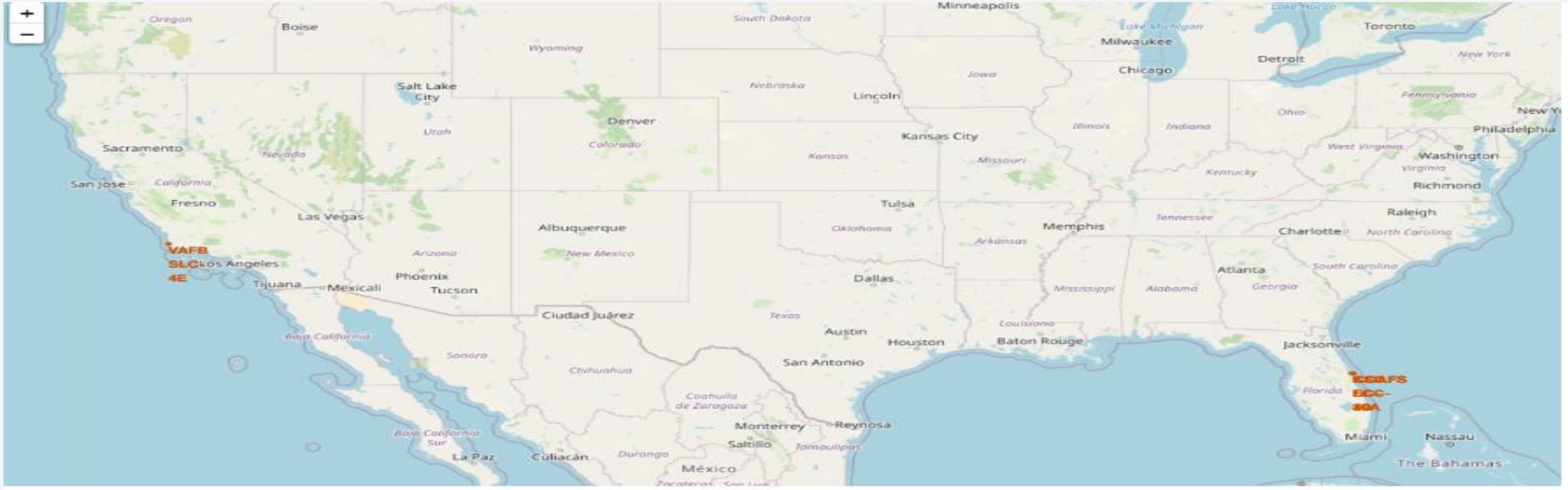
Out[61]:

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

Section 3

Launch Sites Proximities Analysis

Launch Site Location Marker

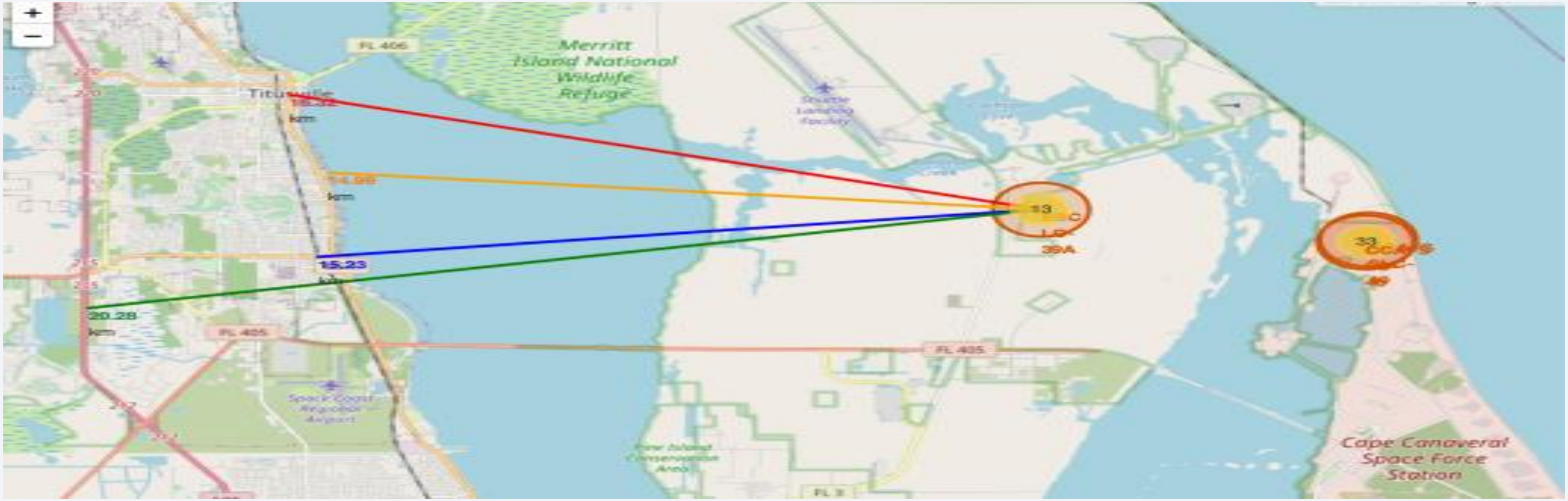


All of the launch sites are near the coastline in the United States.

Color-labeled launch outcomes

Green Mark – Indicates Successful launches
Red Mark – Indicates Unsuccessful launch
Launch station KSC LC-39 A has the highest number of successful launch

Distance from the launch site KSC LC-39A to its proximities

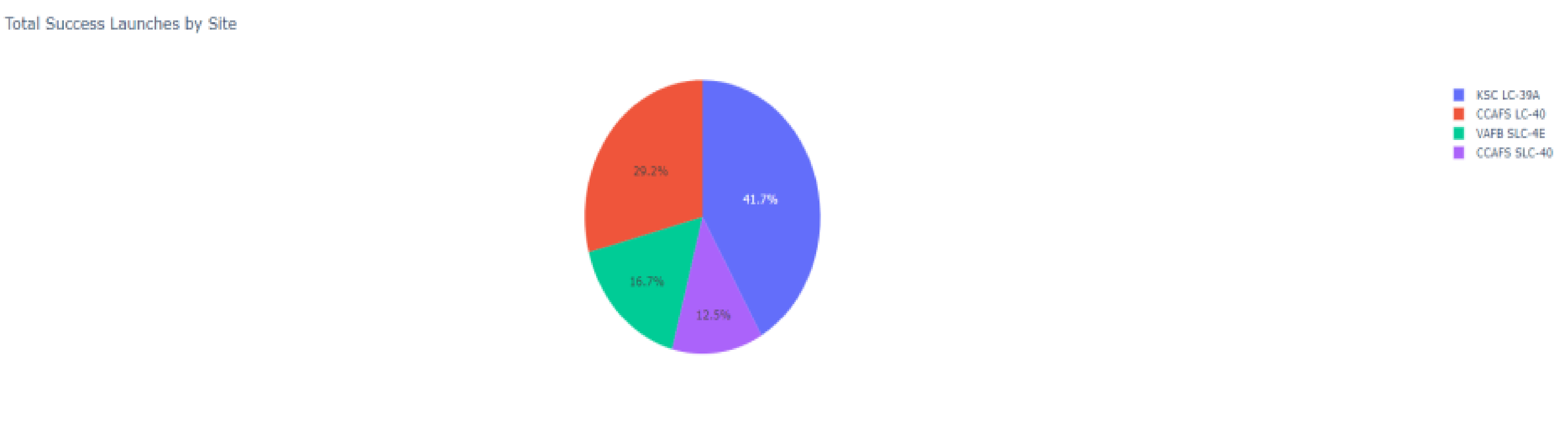


The Launch site is
relatively close to railway (15.23 km)
Relatively close to highway(20.28 km)
Relatively close to coastline (14.99 km)

Section 4

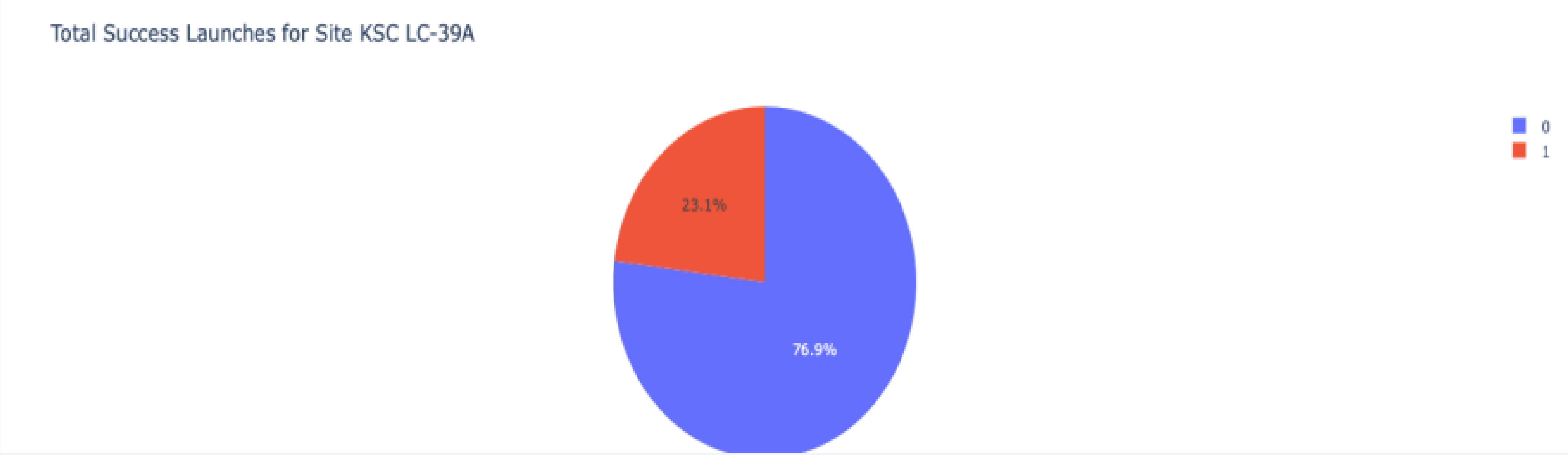
Build a Dashboard with Plotly Dash

Total Success Launches by site



We can see that KSC LC-39 A has the most successful launches

Success Ratio for KSC LC-39 A

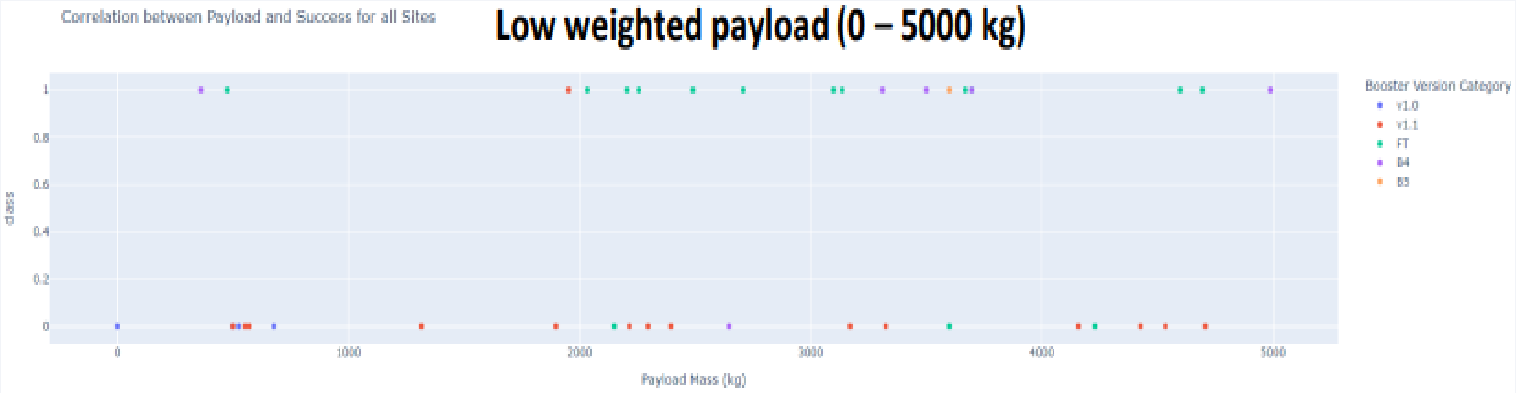


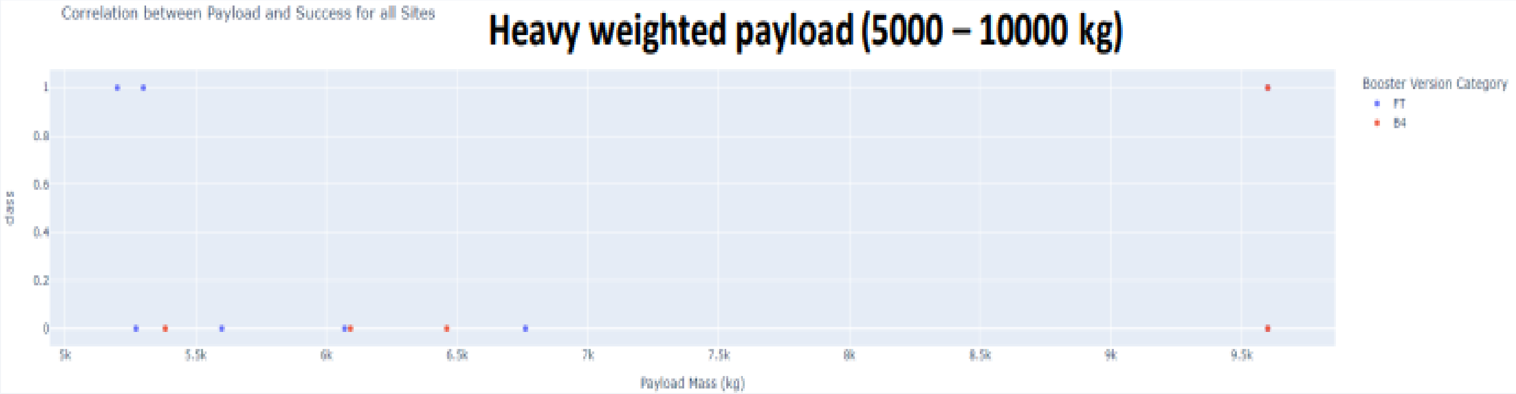
Only 23.1 % of the landings have failed from KSC LC-39 A launch site. Majority landings have been successful

Payload vs. Launch Outcome for all sites

The chart shows that too much heavy weighted payload leads to failed launching.

The payload mass between 2000 and 5000 kg have the highest success rate

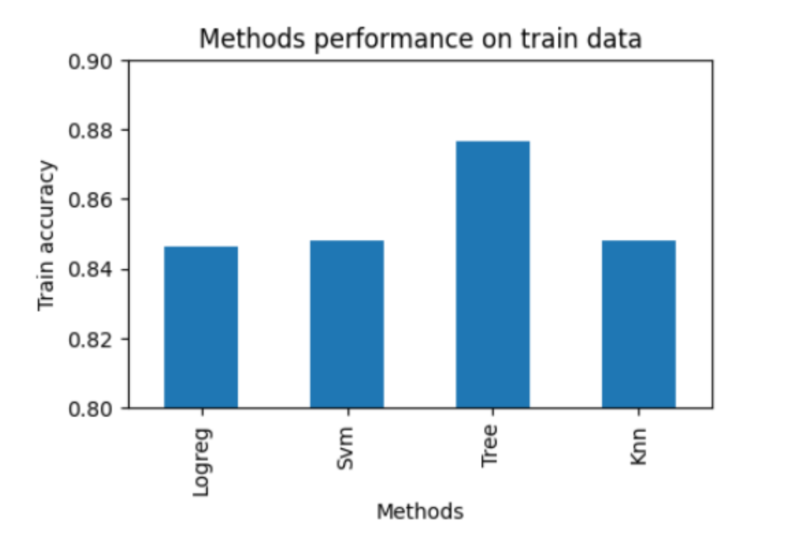


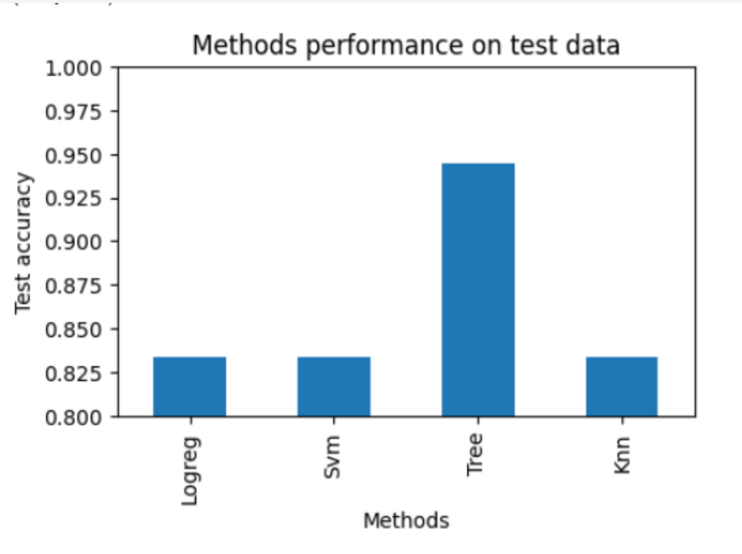


Section 5

Predictive Analysis (Classification)

Classification Accuracy



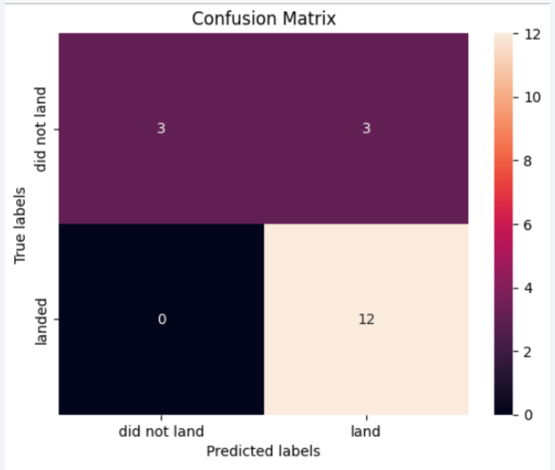


	Accuracy Train	Accuracy Test
Tree	0.876786	0.944444
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333

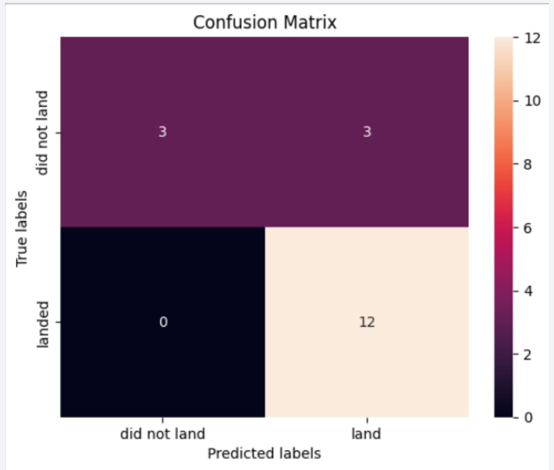
- Decision Tree gives better accuracy than the rest of the classification algorithm.

Confusion Matrix

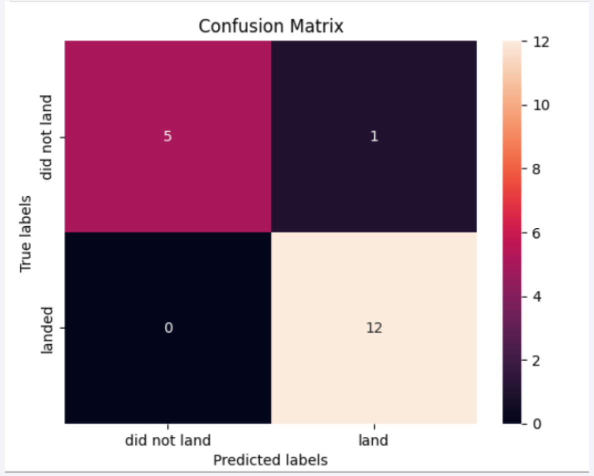
Logistic Regression



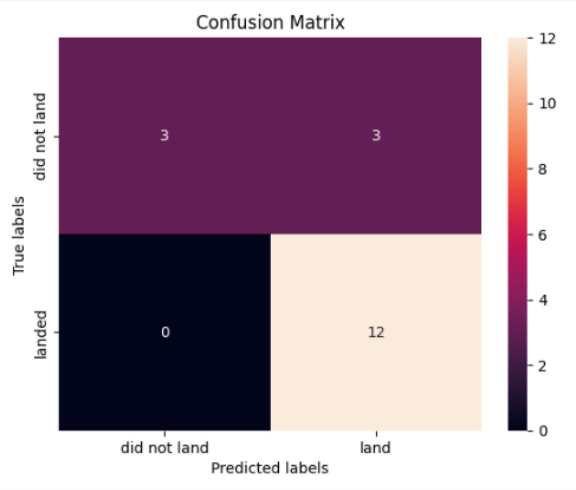
SVM



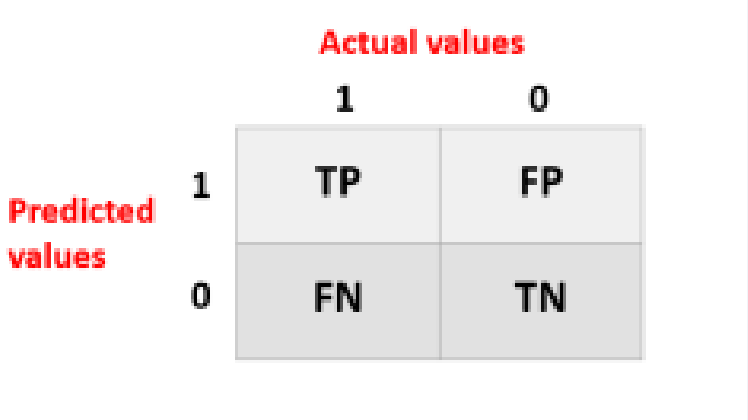
Decision Tree



KNN



Decision Tree performs better with least number of false positive.



Conclusions

- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.
- With the current data, we cannot explain why some launch sites are better than others (KSC LC-39A is the best launch site). To get an answer to this problem, we could obtain atmospheric or other relevant data.
- For this dataset, we choose the Decision Tree Algorithm as the best model even if the test accuracy between all the models used is identical. We choose Decision Tree Algorithm because it has a better train accuracy.

Thank you!