IBM Developer
SKILLS NETWORK

Winning the Space Race with Data Science

Sérgio Amaro
September 2024

Outline

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

Executive Summary

SUMMARY OF METHODOLOGIES

- Data collection via API and Web Scraping
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis with classification models

SUMMARY OF ALL RESULTS

- Exploratory Data Analysis results
- Interactive analytics via dashboards
- Predictive analysis results

Introduction

PROJECT BACKGROUND AND CONTEXT

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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 SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

PROBLEMS WE WANT TO FIND ANSWERS

- What are the main characteristics of a successful or failed landing?
- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- What are the conditions which will allow SpaceX to achieve the best landing success rate?

SECTION 1
Methodology

Methodology

Executive Summary

Data collection methodology:

- SpaceX REST API
- Web Scrapping from Wikipedia

Perform data wrangling:

- Dropping unnecessary columns
- One Hot Encoding for classification models

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Building, tuning and evaluation of classification models to ensure the best results

Data Collection

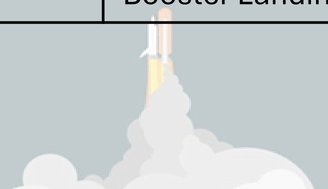
- Data collection process involved a combination of API requests from **SpaceX REST API** and **Web Scraping** data from a table in SpaceX's Wikipedia entry.

- Data obtained from SpaceX REST API:

Flight Number	LaunchSite	Legs	Longitude
Date	Outcome	Landing Pad	Latitude
Booster Version	Flights	Block	
Payload Mass	Grid Fins	Reused Count	
Orbit	Reused	Serial	

- Data Columns are obtained by using Wikipedia Web Scraping:

Flight Number	Payload Mass	Launch Outcome	Date
Launch Site	Orbit	Version Booster	Time
Payload	Customer	Booster Landing	



Data Collection – SpaceX API

1. Request rocket launch data from SpaceX API.

2. Decode the response content using .json() and turn it into a dataframe using .json_normalize().

3. Retrieve information about the launches from SpaceX API by applying custom functions.

4. Construct the retrieved data into a dictionary.

5. Create a dataframe from the dictionary.

6. Filter the dataframe only to include data from Falcon 9 launches.

7. Replace missing values of Payload Mass column with the .mean() for this column.

8. Export data to CSV.

GitHub link: M1 - 1 Collecting the Data

Data Collection - Scraping

1. Request Falcon 9 launch data from Wikipedia.
2. Create a BeautifulSoup object from the HTML response.
3. Extract all column names from the HTML table header.
4. Collect the data by parsing HTML tables.
5. Construct a dictionary with the obtained data.
6. Create a dataframe from the dictionary.
7. Export data to CSV.

🔗 GitHub link: M1 - 2 Webscraping

Data Wrangling

1. Calculate number of launches for each site.

2. Identify the diferente orbits and respective occurences.

3. Calculate number and occurrence of mission outcome for each orbit type.

4. Create a landing outcome label from Outcome column.

5. Export data to CSV.

GitHub link: M1 - 3 Data Wrangling

EDA with SQL

SQL queries performed in this Lab:

- Display 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 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.

🔗 GitHub link: M2 - 1 EDA with SQL

EDA with Data Visualization

The following charts were plotted:

Flight Number vs. Payload Mass	Flight Number vs. Orbit Type
Flight Number vs. Launch Site	Payload Mass vs Orbit Type
Payload Mass vs. Launch Site	Success Rate Yearly Trend
Orbit Type vs. Success Rate	

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model. Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value. Line charts show trends in data over time (time series).

🔗 GitHub link: M2 - 2 Exploring and Preparing Data

Build an Interactive Map with Folium

The following objects were added to the map, centered on NASA Johnson Space Center at Houson, Texas:

- Red circle on NASA Johnson Space Center's coordinates, with label showing its name (folium.Circle, folium.map.Marker).
- Red circles on 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 different informations 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 a plot line between them (folium.map.Marker, folium.PolyLine, folium.features.DivIcon).

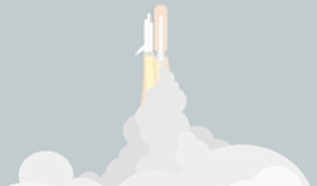
These objects were selected to add context to the data gathered. They show launch sites, points of interest in their surroundings, and the number of successful and unsuccessful landings.

🔗 GitHub link: M3 - 1 Analysis with Folium

Build a Dashboard with Plotly Dash

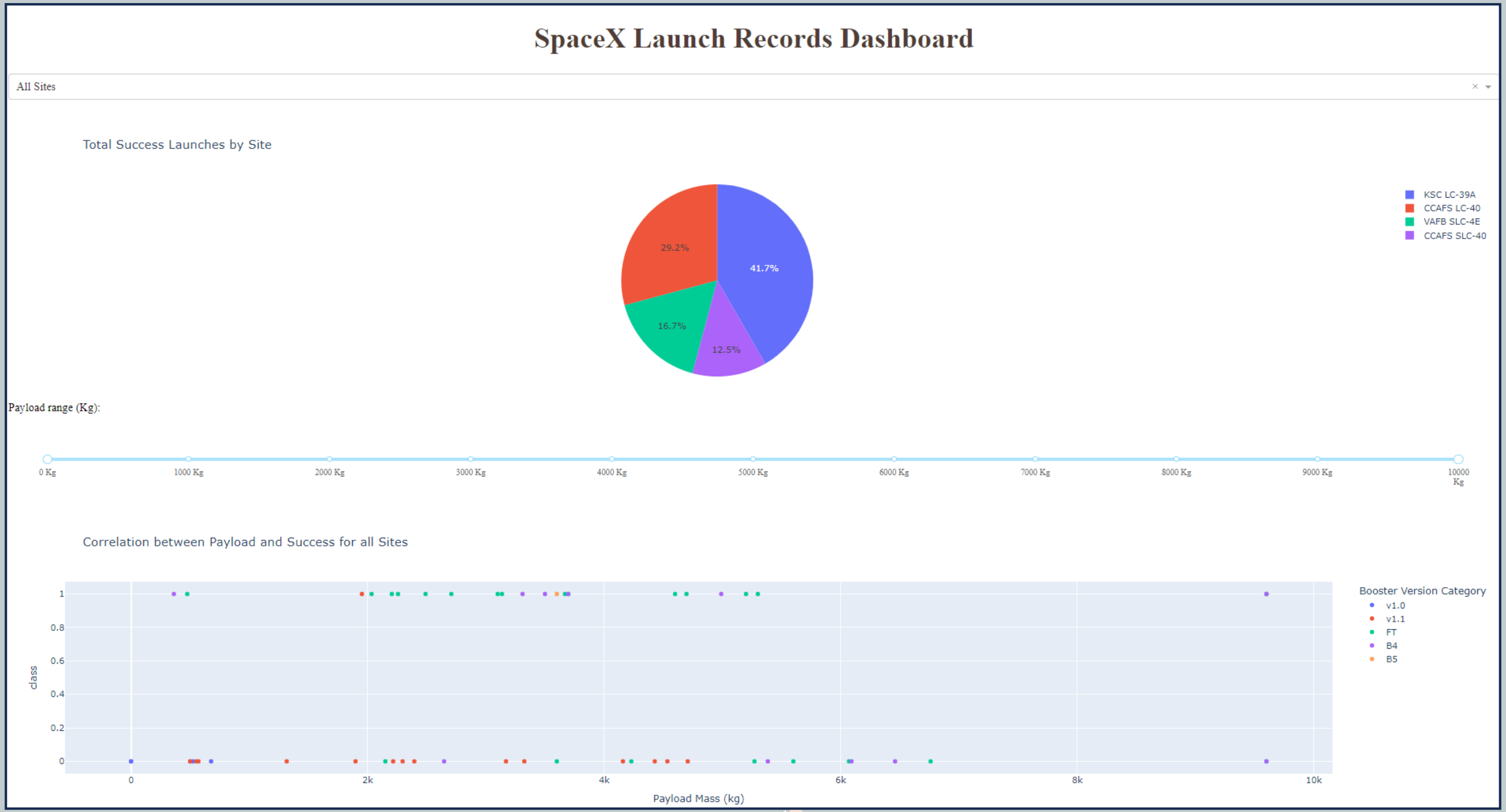
The dashboard has the following components:

Dropdown menu	Allow the user to choose between the different launch sites or view the data for all of them.
Pie Chart	Shows the total success and the total failure for the launch site chosen on the dropdown menu.
Rangeslider	Allows a user to select a payload mass in a fixed range.
Scatterplot	Shows the relationship between Success vs. Payload Mass.



Build a Dashboard with Plotly Dash

Dashboard printscreen:



🔗 [GitHub link: M3 - 2 Interactive Dashboard with Ploty Dashs]

Predictive Analysis (Classification)

1. Create a NumPy array from the column "Class" in data.

2. Standardize the data with StandardScaler, then fitting it and transforming it.

3. Split the data into training and testing sets with the train_test_split function.

4. Create a GridSearchCV object with cv = 10 to find the best parameters.

5. Apply GridSearchCV on LogReg, SVM, Decision Tree, and KNN models.

6. Calculate the accuracy on the test data using the method .score() for all models.

7. Examine the confusion matrix for all models.

8. Find the method that performs the best by examining the Jaccard_score and F1_score metrics.

🔗 [GitHub link: M4 - 1 Machine Learning Prediction]

Results

EXPLORATORY DATA ANALYSIS RESULTS:

- Lighter payloads perform better when compared to heavier ones.
- The lunch success is increasing with the number of years of experience, demonstrating a positive trend over time.
- The Launch Complex 39A at Kennedy Space Center has the highest number of successful launches of the 4 sites analyzed.
- GEO, HEO, SSO and ES L1 are the orbit types with the highest rate of successful launches.

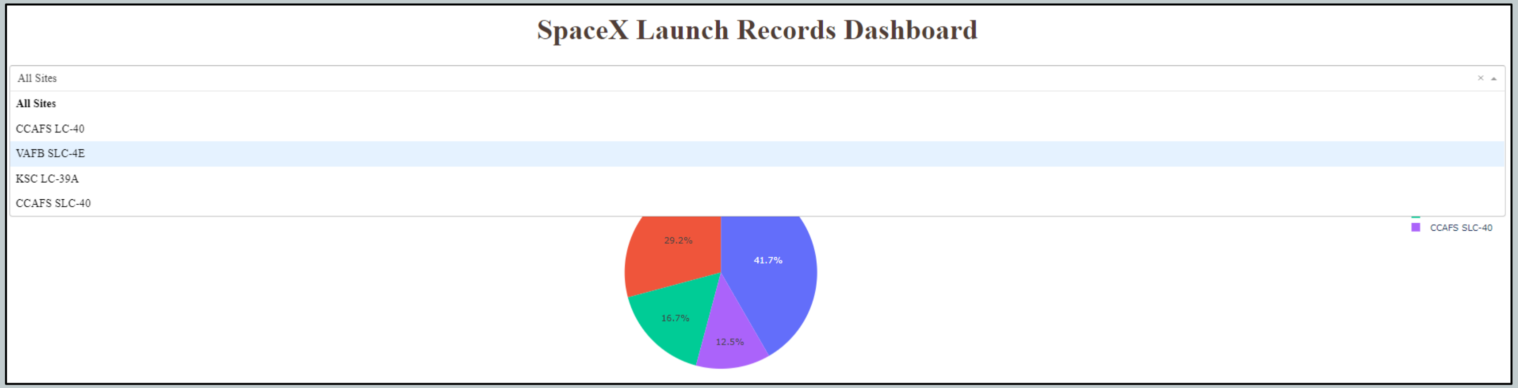
PREDICTIVE ANALYSIS RESULTS:

```
Accuracy of GridSeachCV: 0.8625
Accuracy of SVM score: 0.7777
Accuracy of Decision Tree: 0.90357
Accuracy of KNN model: 0.8767
Best Parameters are:Decision Tree with a score of 0.90357
Best Parameters are:: {'criterion': 'entropy', 'max_depth': 16, 'max_features': 'sqrt',
                       'min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'random'}
```

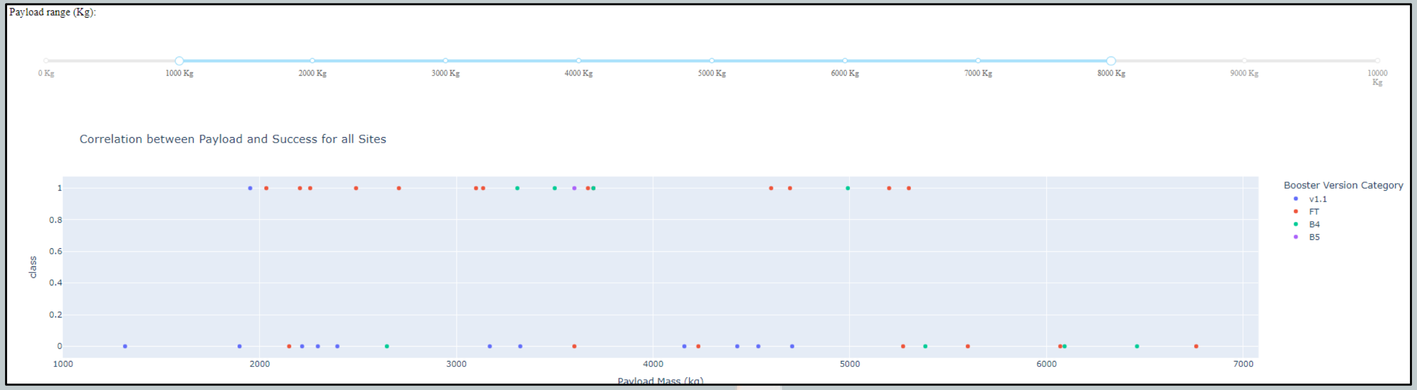
Results

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

Dropdown menu to select launch sites in order to show in the pie chart the total success launches:

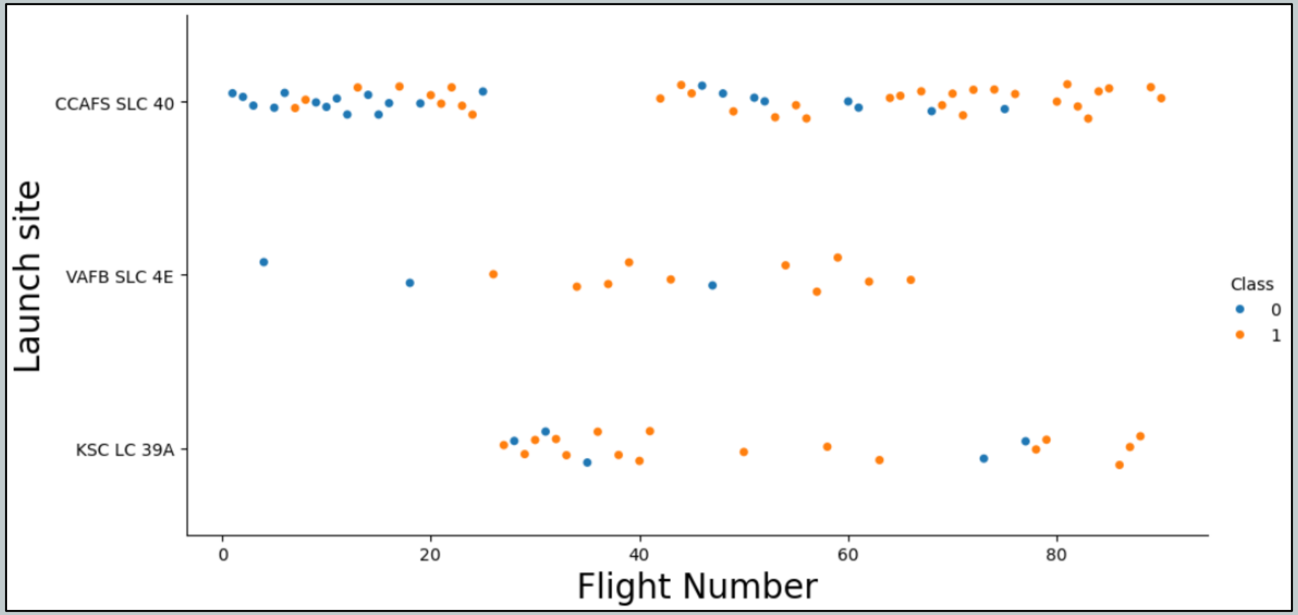


Scatterplot representing correlation with payload and success for all sites, with range slider to select payload:



SECTION 2
Insights drawn from EDA

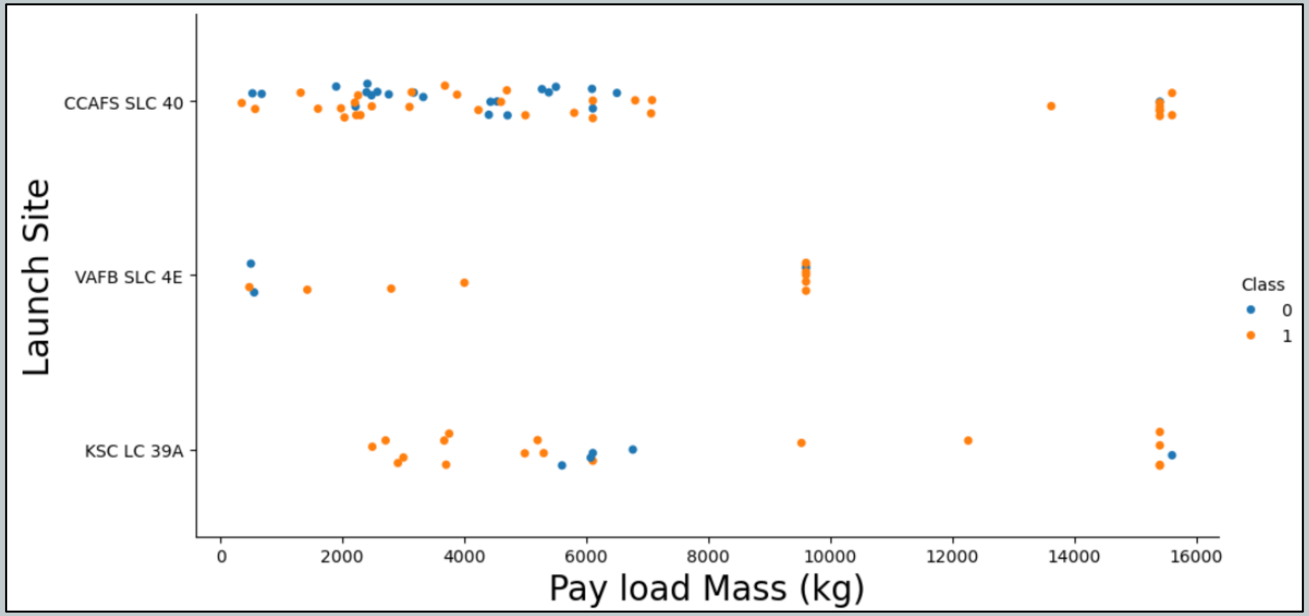
Flight Number vs. Launch Site



OBSERVATIONS:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

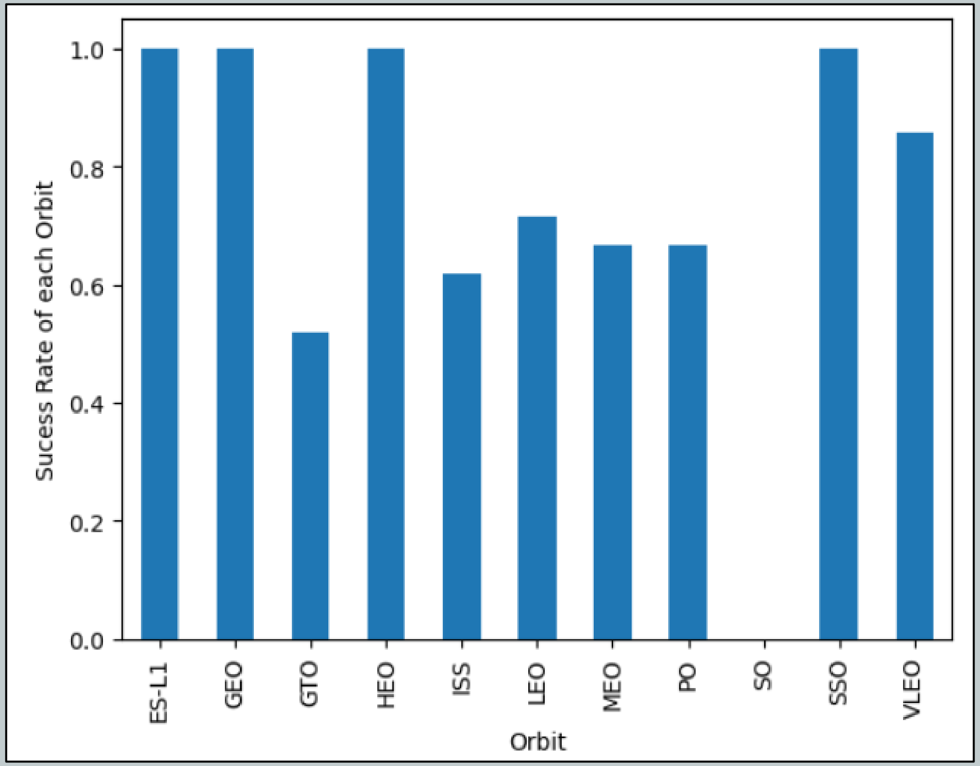
Payload vs. Launch Site



OBSERVATIONS:

- For every launch site, the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg.

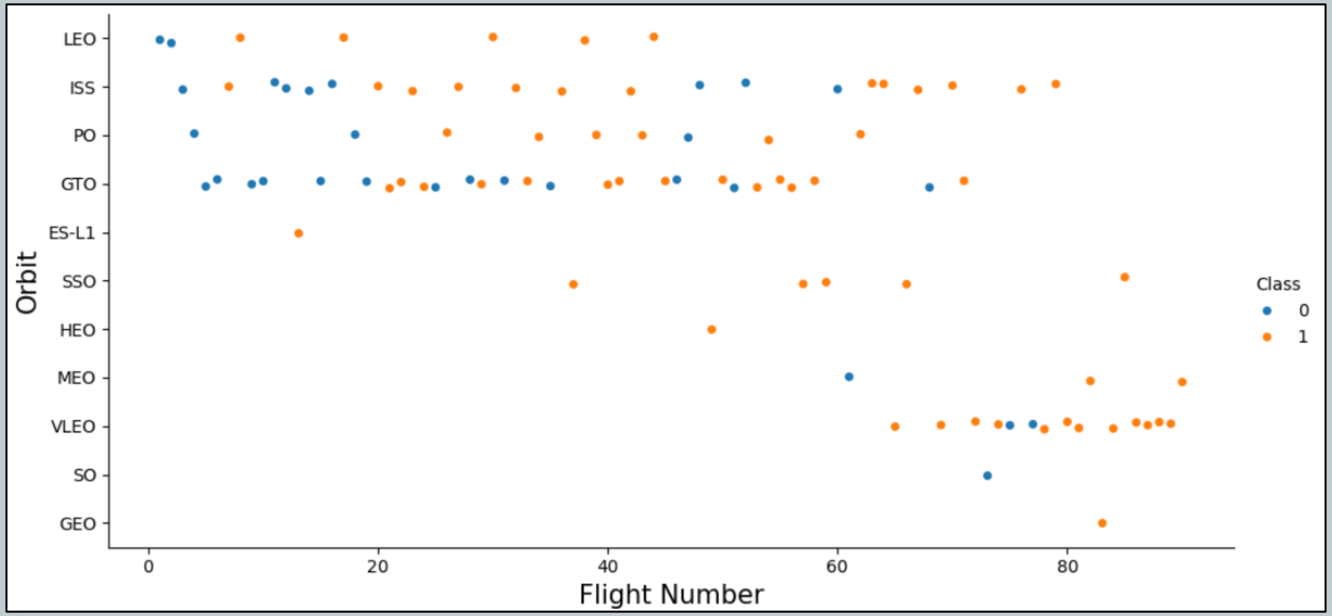
Success Rate vs. Orbit Type



OBSERVATIONS:

- ES-L1, GEO, HEO and SSO have 100% success rate.
- SO as a 0% success rate.
- GTO, ISS, LEO, MEO and PO have a success rate between 50% an 85%.

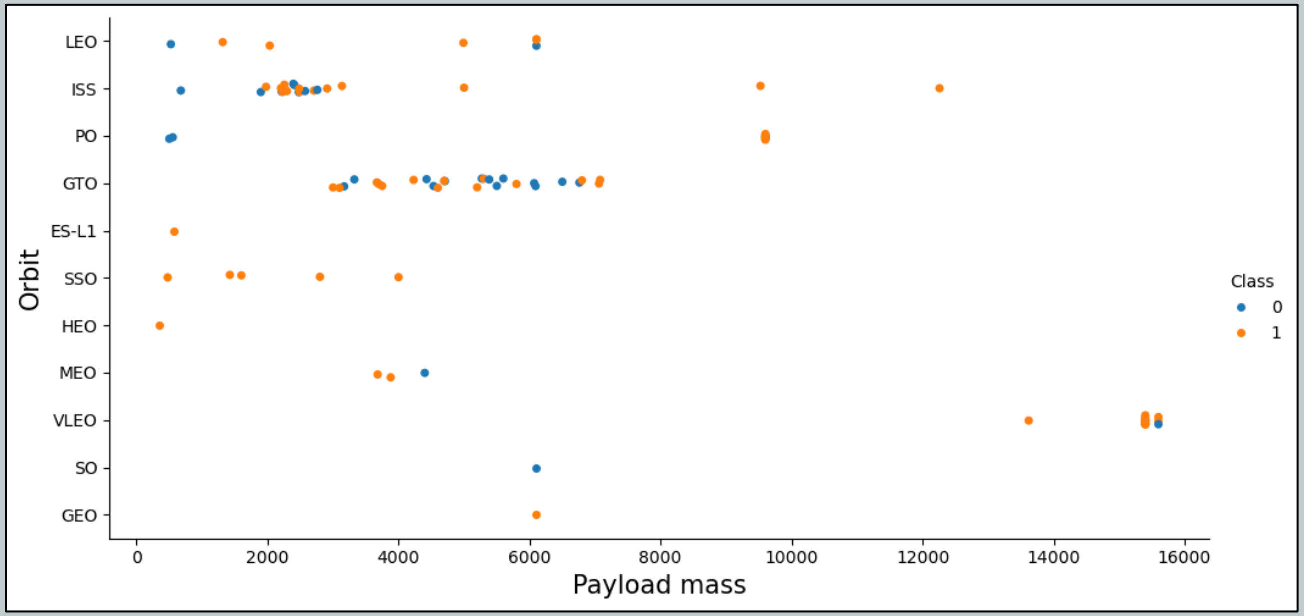
Flight Number vs. Orbit Type



OBSERVATIONS:

- 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.
- High success rates from orbits SSO or HEO may be due to the knowledge acquired from previous launches for other orbits.

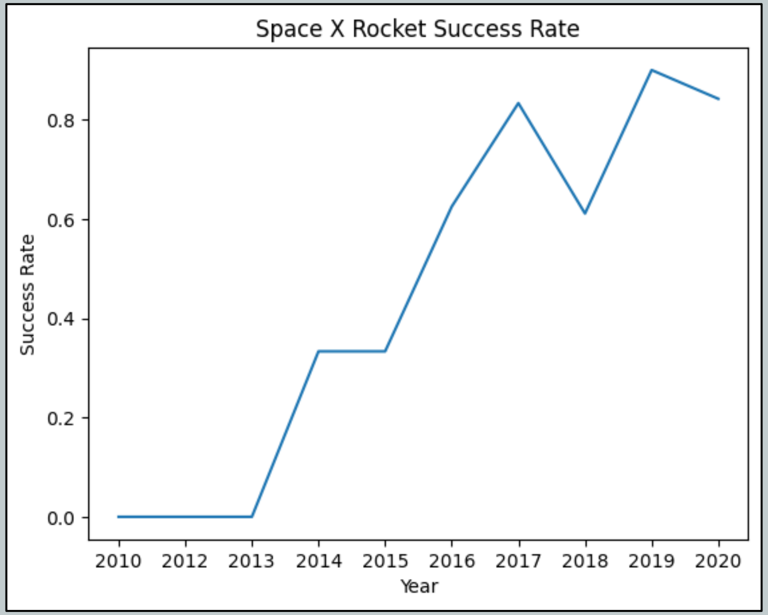
Payload vs. Orbit Type



OBSERVATIONS:

- Heavy payloads show a negative influence on GTO orbits and a positive influence on GTO and Polar LEO (ISS) orbits.
- Lower payloads for a GTO orbit improves the success of a launch.

Launch Success Yearly Trend



OBSERVATIONS:

- After the first 3 years of unsuccessful launches, the success rate has been increasing up to 2020.

All Launch Site Names

QUERY:

Task 1

Display the names of the unique launch sites in the space mission

[13]:
```
%sql select distinct launch_site from SPACEXTABLE;
```

RESULT:

[13]:

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

EXPLANATION:

Use of DISTINCT in the query allows to remove duplicate LAUNCH_SITE entries.

Launch Site Names Begin with 'CCA'

QUERY:

Task 2

Display 5 records where launch sites begin with the string 'CCA'

[14]:
```
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

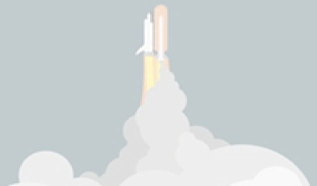
RESULT:

[14]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

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

QUERY:

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

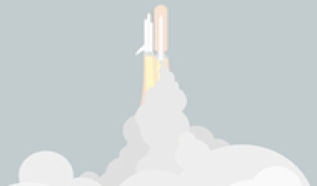
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

RESULT:

SUM(PAYLOAD_MASS__KG_)
45596

EXPLANATION:

- The query sums all the payload masses where the customer is NASA (CRS).



Average Payload Mass by F9 v1.1

QUERY:

Task 4

Display average payload mass carried by booster version F9 v1.1

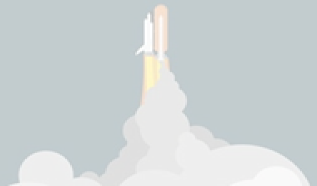
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.0%';
```

RESULT:

AVG(PAYLOAD_MASS__KG_)
340.4

EXPLANATION:

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



First Successful Ground Landing Date

QUERY:

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

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

RESULT:

MIN(Date)
2015-12-22

EXPLANATION:

- The WHERE clause filters the dataset in order to keep only records where landing was successful. Then, with the MIN function, the record with the earliest date is selected.

Successful Drone Ship Landing with Payload between 4000 and 6000

QUERY:

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

RESULT:

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

EXPLANATION:

- The query returns the booster versions 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

QUERY:

Task 7

List the total number of successful and failure mission outcomes

In [44]:

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

RESULT:

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

EXPLANATION:

- The different values of the Mission Outcome column were counted and presented on a table as total number of occurrences.

Boosters Carried Maximum Payload

QUERY:

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

RESULT:

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

EXPLANATION:

- A subquery was used to filter data by returning only the heaviest payload mass with MAX function. The main query uses subquery results and returns unique booster version with the heaviest payload mass.

2015 Launch Records

QUERY:

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

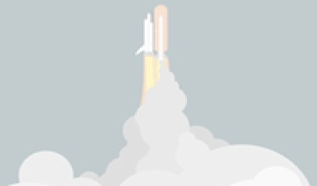
```
%%sql
SELECT Landing_Outcome, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)'
    AND year(DATE) = 2015;
```

RESULT:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

EXPLANATION:

- The query returns landing outcome, booster version and launch site where landing was unsuccessful, and landing date was the year 2015.



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

QUERY:

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY TOTAL_NUMBER DESC
```

RESULT:

Landing_Outcome	TOTAL_NUMBER
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

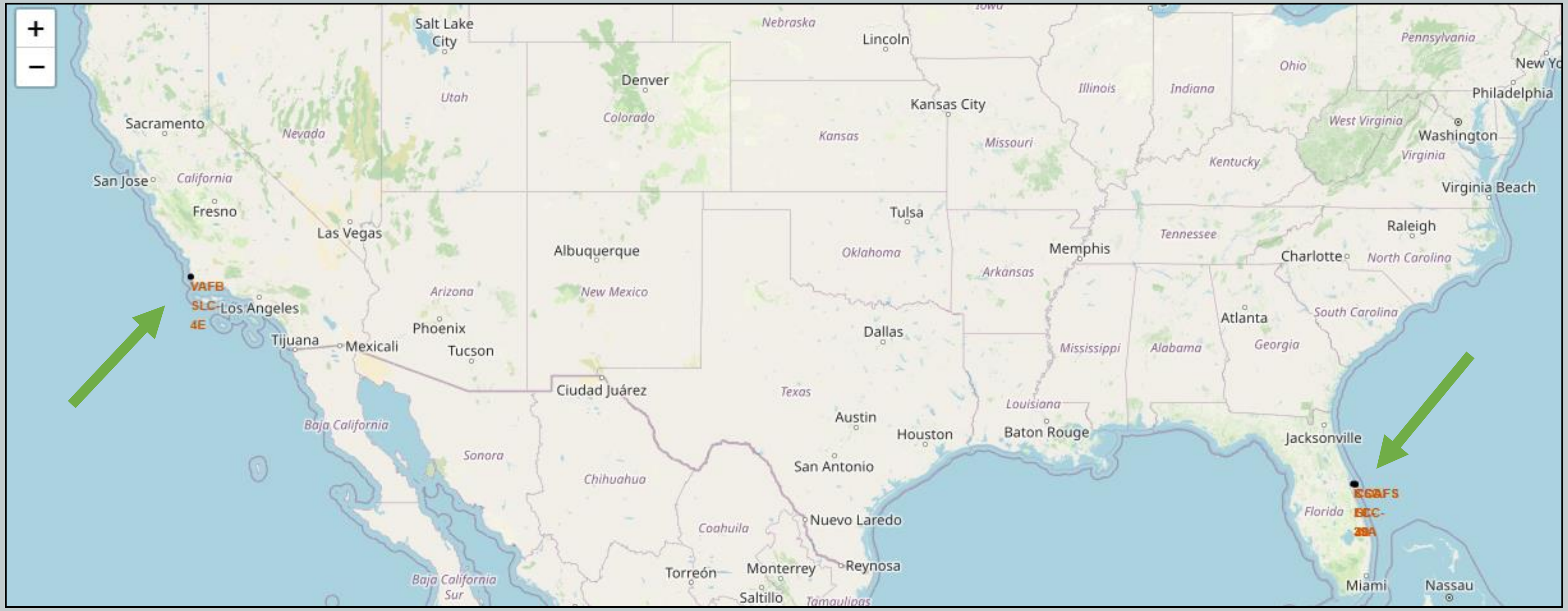
EXPLANATION:

- The query counts all the landing outcomes instances thar occurred between 2010-06-04 and 2017-03-20 and sets up the results in descending order.

SECTION 3

Launch Sites Proximities Analysis

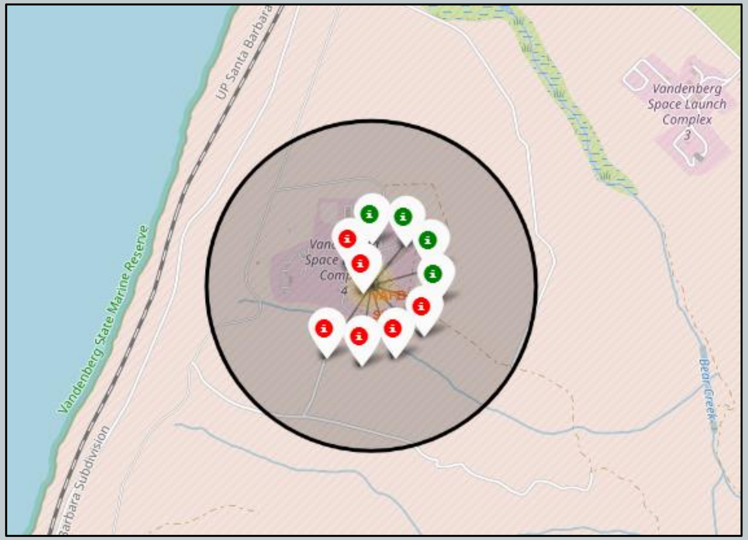
Ground Stations

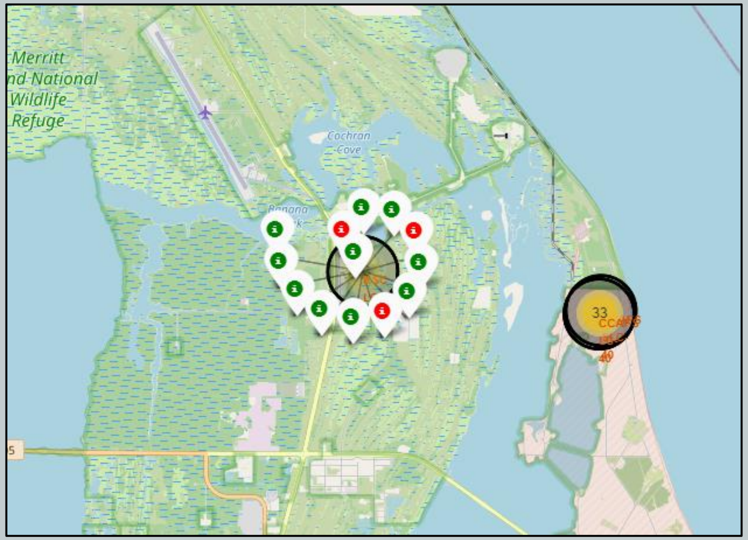


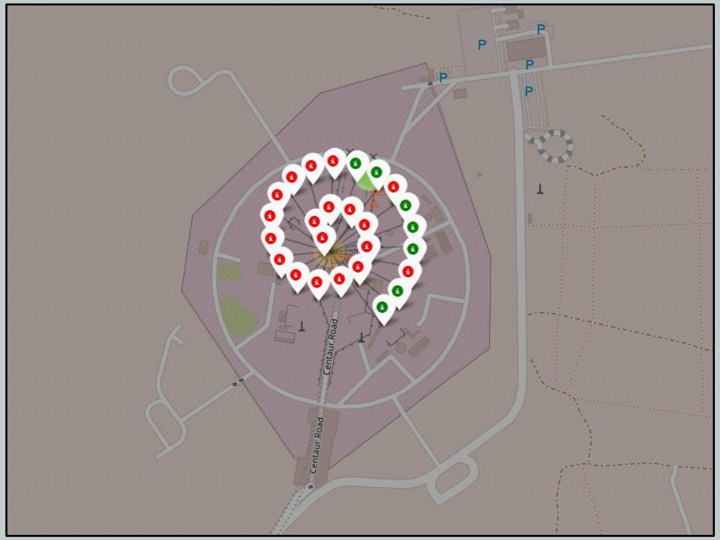
OBSERVATIONS:

- The map shows the launch locations on the US map, both on the east and west coast.
- The ocean proximity reduces risk to human populations in case of launch failures and also provides a safer trajectory for the rocket stages that fall back to Earth after separation.

Color-labeled launch



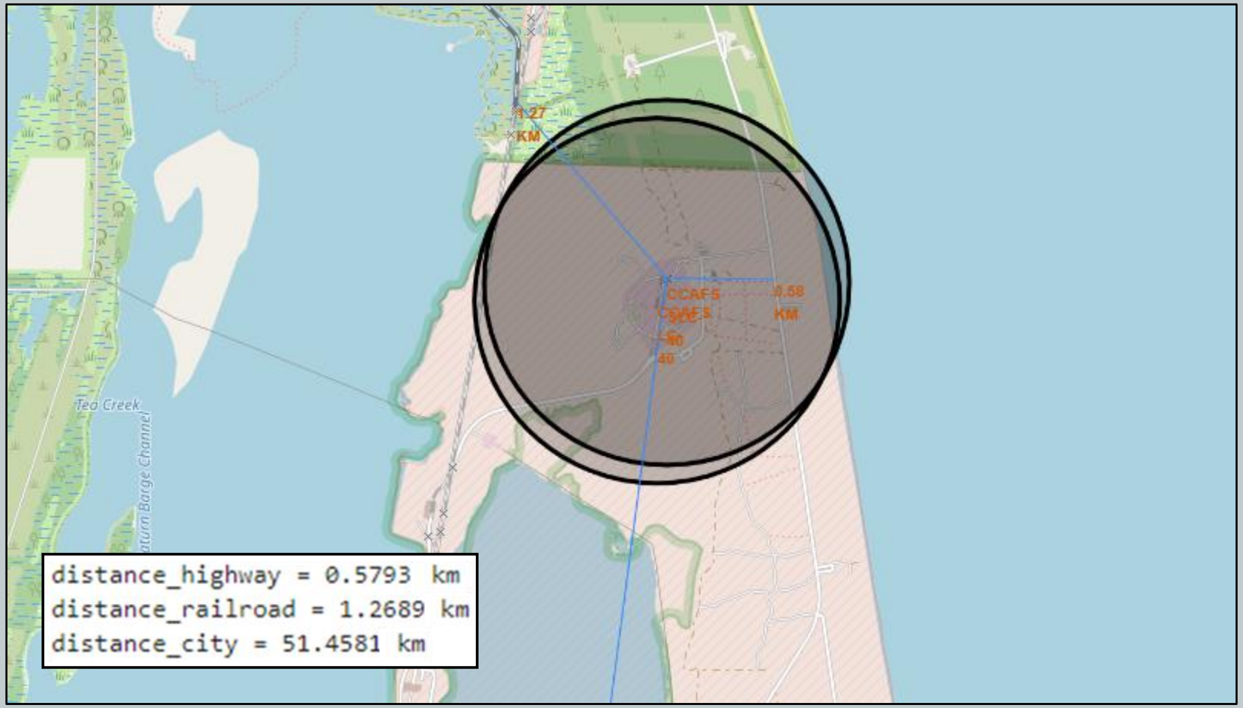




OBSERVATIONS:

- From the color-labeled markers it's easy to identify which launch sites have relatively high success rates.
- Launch Site KSC LC-39A has a very high success rate.

Distances from CCAFS-SLC4



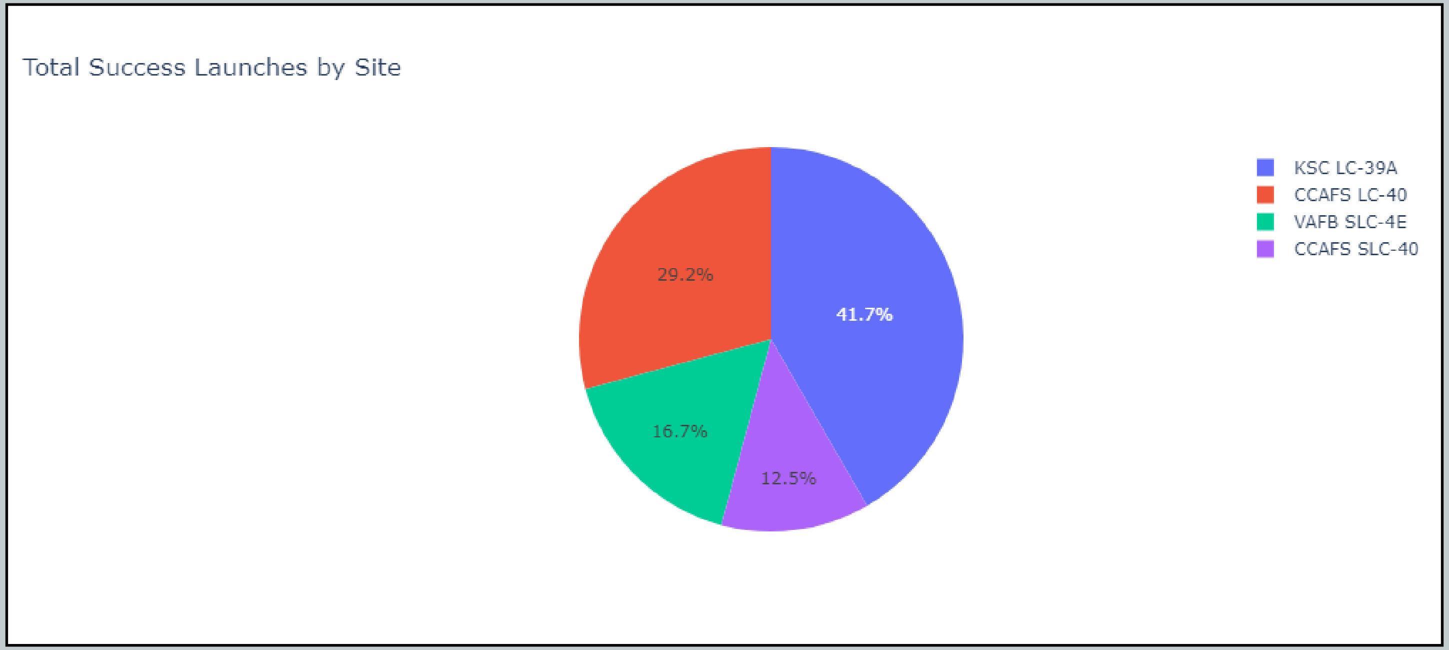
OBSERVATIONS:

- The map shows CCAFS-SLC4 benefits from being in the vicinity of the coastline, a higway and a railroad.
- The nearest city is located about 51km away.

SECTION 4

Dashboard with Plotly Dash

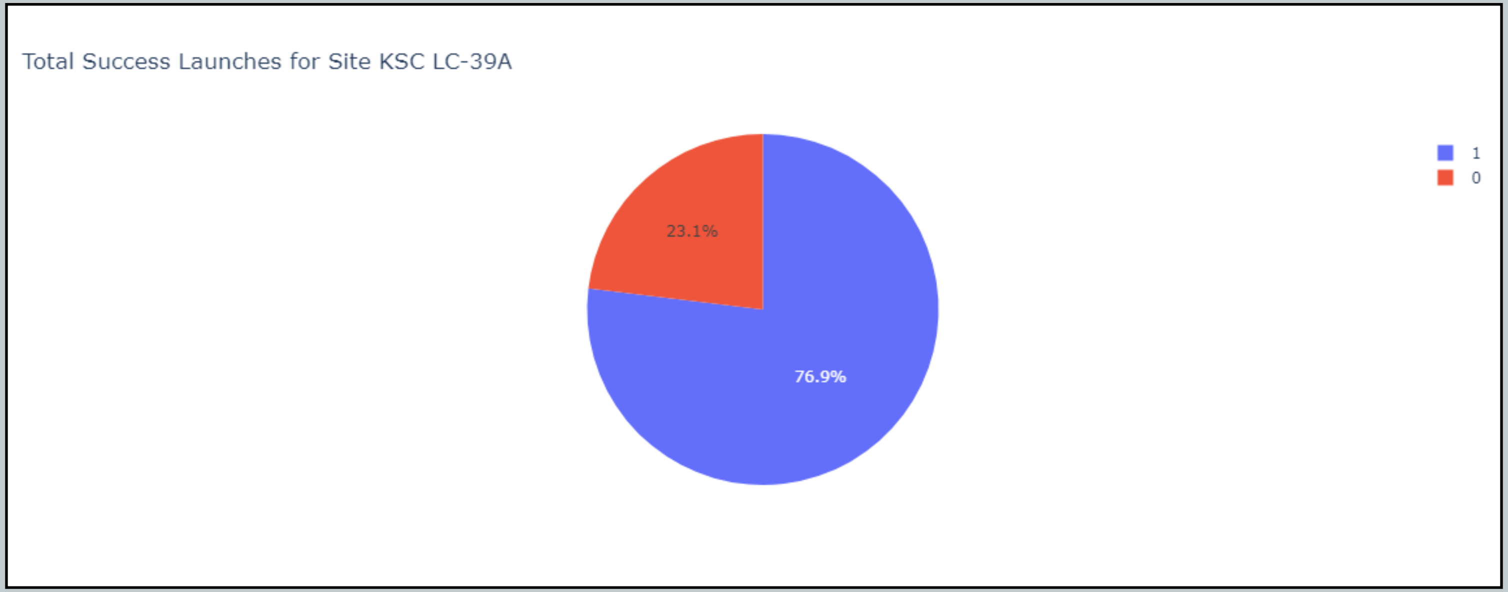
Dashboard – Total success by Site



OBSERVATIONS:

- The pie chart shows the total successful launches count for all sites.
- The KSC LC-39A launch site has the best success rate of launches with 41,7%.

Dashboard – Total success launches for KSC LC-39A

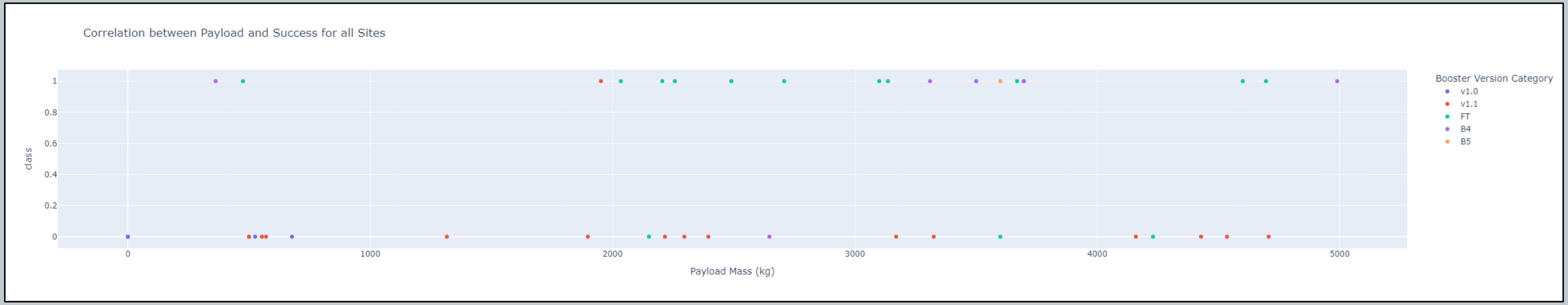


OBSERVATIONS:

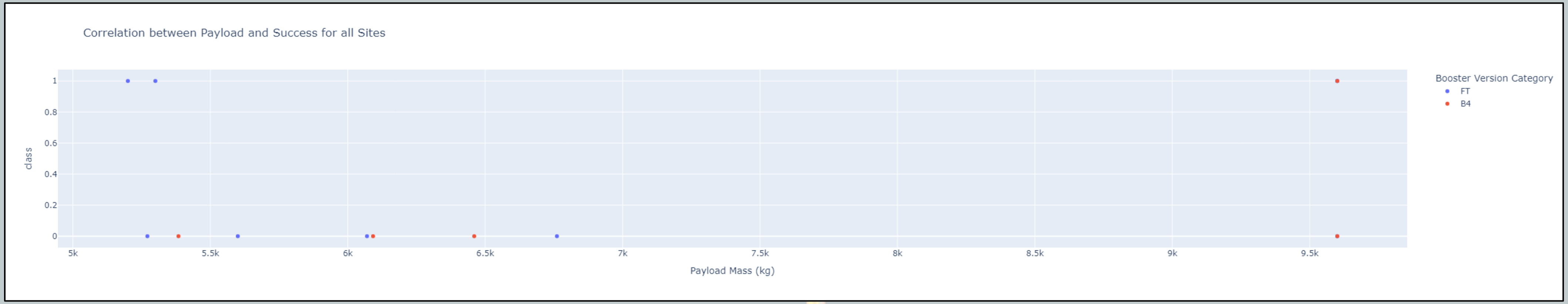
- The KSC LC-39A launch site has a 76.9% success rate and a 23.1% failure for the total launches performed at the site.

Dashboard – Payload Mass vs Outcome

Payload Mass vs. Outcome for payload masses between 0 and 5000kg:

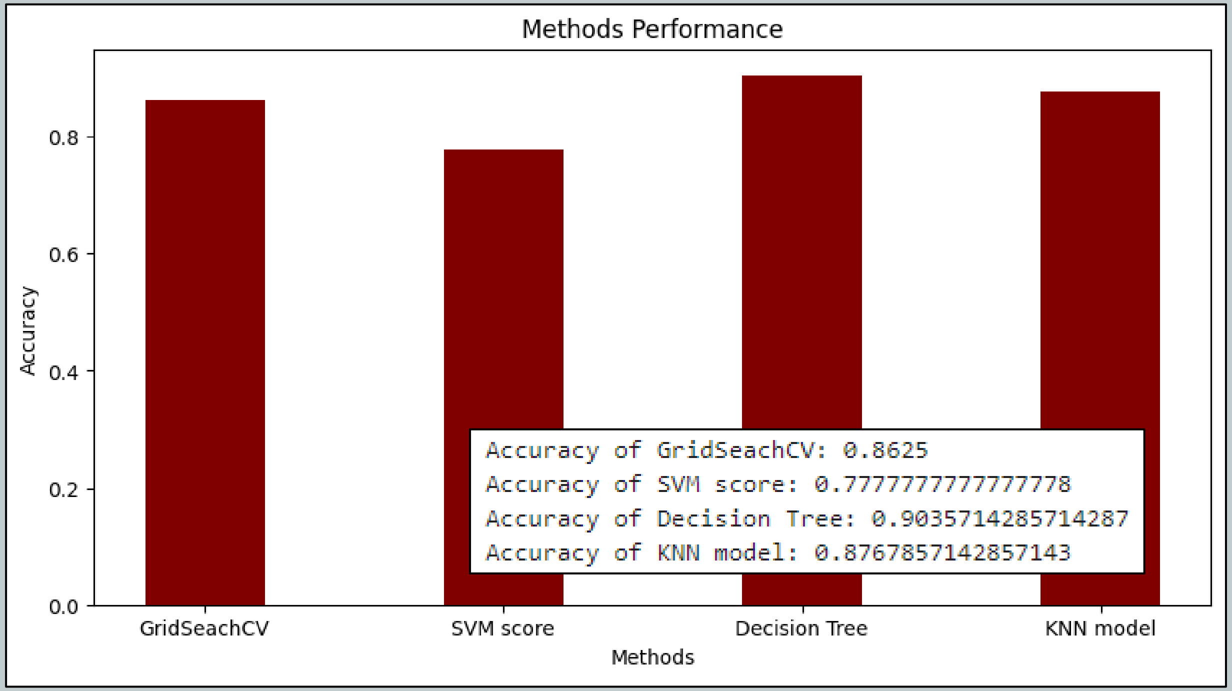


Payload Mass vs. Outcome for payload masses between 5000 and 5000kg:



SECTION 5
Predictive Analysis

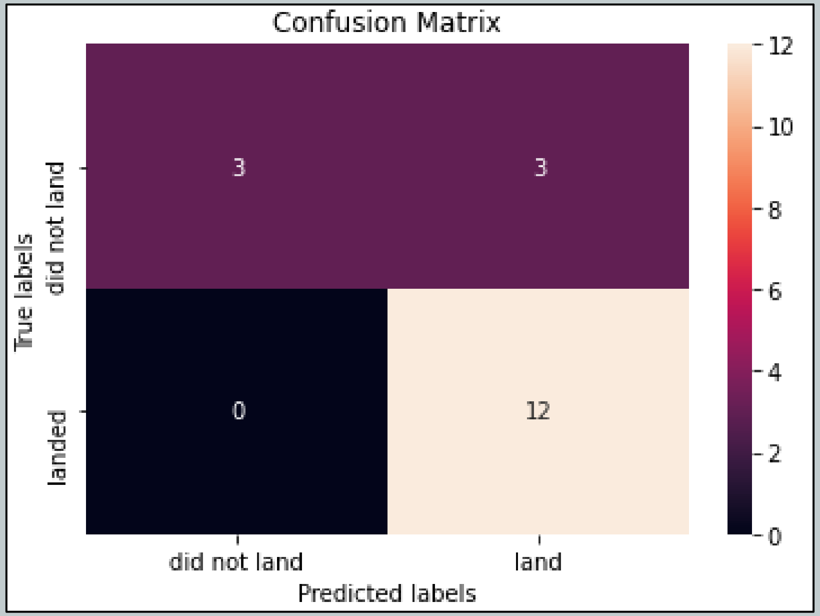
Classification Accuracy



OBSERVATIONS:

- The Decision Tree has the highest accuracy score with a value of 90,4%.
- SVM has the lowest accuracy score, with a value of 77,8%.

Confusion Matrix – Decision Tree Model



OBSERVATIONS:

- The confusion matrix predicts 12 true positives, 3 false positives, 3 true positive, and 0 false negative.
- The false positives may indicate the model is over-predicting the positive class.

Conclusions

- A successful mission is explained relies on a set of factors: launch site, orbit and especially the number of previous launches. It’s observable that the more recent the data, the most success launches are.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- Low weight payloads show a better performance than heavy weighted payloads. When analysing the different orbits, it’s also possible to conclude that the payload can be a factor for the success of the mission.
- Launch site KSC LC-39A has the best success rate when compared to all the sites analysed.
- The Decision Tree Algorithm proved to be best model, with the highest train accuracy.

Appendix

🔗 GitHub Complete Repository

THANK YOU