IBM Developer SKILLS NETWORK

Winning Space Race with Data Science

Ana Ranđelov
18.11.2024.

Outline

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

Executive Summary

- Summary of methodologies:
    - Data collection and wrangling
    - Exploratory data analysis
    - Interactive visual analytics and dashboard
    - Predictive analysis (Classification)

- Summary of all results

Introduction

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- We want to find out:
  - What are the factors that influence the success of landing
  - What are the relationships between these factors

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Get request to the SpaceX API and webscraping from a Wikipedia page
- Perform data wrangling
  - Calculated number of launches and orbits, occurences of mission outcomes, and used one – hot encoding for the mission outcomes to determine training labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was standardized and split into a test and training sets, and then then the performance of various models was tested to determine the best model

Data Collection

- SpaceX launch data was collected using the GET request on a json file, and the response was decoded using .json() method and turned into a Pandas dataframe using the .json_normalize() method. This step was necessary because working with DataFrames is essential for cleaning data, handling missing values etc.
- Using BeautifulSoup module webscraping was performed – from a Wikipedia page we parsed the data using a html parser and created a soup object. Then we extracted the data from the chosen tables from the Wikipedia page.

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- https://github.com/arandelov/test_repo/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb

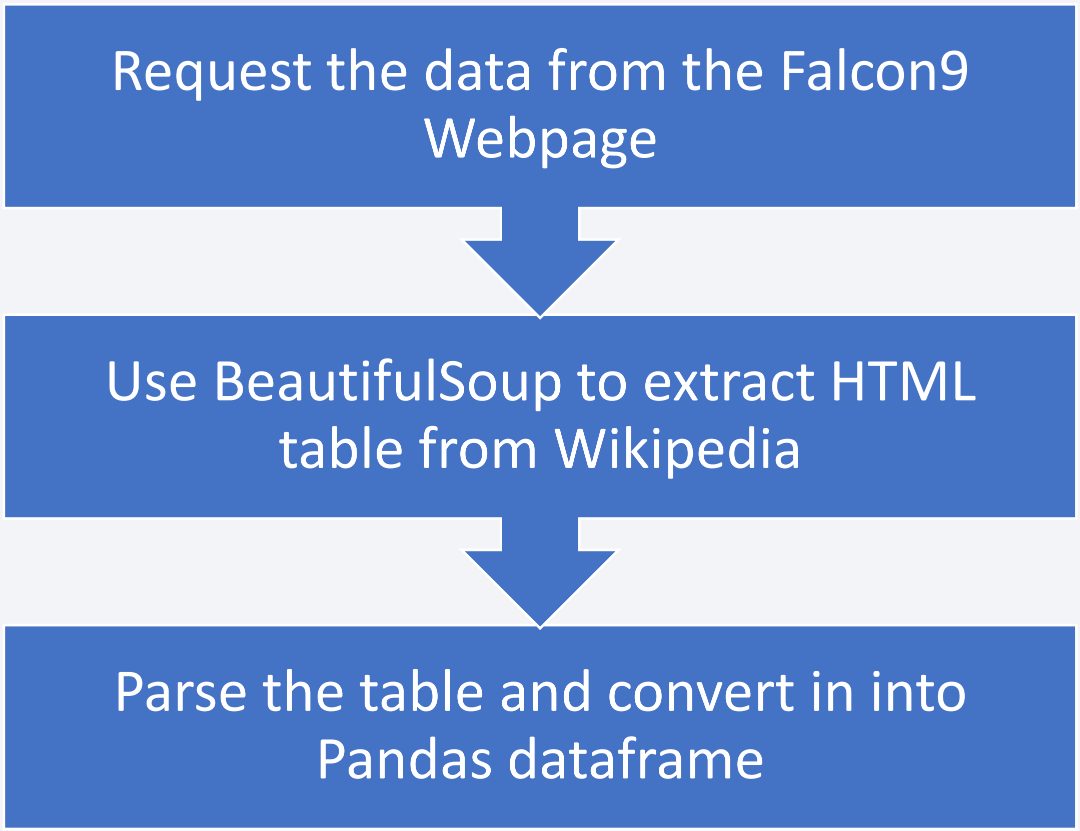
API
- Get request on a JSON file

JSON to DF
- Converted JSON object to Pandas dataframe using the json_normalize() method

Data cleaning
- Filled in the missing values with the average

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts
- https://github.com/arandelov/test_repo/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- The number of ocurences of unique values are counted for several different columns, and then the data is split based on the outcome (True/False) of the landing, with the goal of creating outcome labels we will need for the following tasks
- You need to present your data wrangling process using key phrases and flowcharts
- https://github.com/arandelov/test_repo/blob/main/labs-jupyter-spacex-Data%20wrangling-v2.ipynb

EDA with Data Visualization

- The following charts were plotted:
  - ❖ Flight Number vs Launch Site: catplot
  - ❖ Payload Mass vs Launch Site: scatterplot
  - ❖ Success rate of each orbit type: barplot
  - ❖ Flight Number vs Orbit Type: scatterplot
  - ❖ Payload mass vs Orbit type: scatterplot
  - ❖ Launch success yearly trend: line plot
- https://github.com/arandelov/test_repo/blob/main/edadataviz.ipynb

EDA with SQL

- The following SQL queries were performed:
  - ❖ Displayed the names of unique launch sites (used DISTINCT)
  - ❖ Displayed 5 sites beginning with “CCA” (used WHERE … LIKE ‘CCA%’)
  - ❖ Displayed total payload mass carried by boosters (used SUM())
  - ❖ Displayed average payload mass carried by boosters (used AVG ())
  - ❖ Listed boosters having payload mass between 40 000 and 60 000 (used < AND >, alternatively could use BETWEEN)
  - ❖ Listed grouped mission outcomes by success/failure (used GROUP BY)
  - ❖ Listed the maximum payload mass booster carriers using a subquery (used MAX () and a subquery to select the value for which we will take the maximum)
  - ❖ Ranked the count of landing outcomes for a specific date (using WHERE, GROUP BY and ORDER BY)
- https://github.com/arandelov/test_repo/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Marked all launch sites, then the success and failed launches and calculated the distances between a launch site and its proximities. We used longitude and latitude coordinates for calculation, class columns for outcomes and the column containing the sites names.
- Outcomes are assigned red or green circle for aesthetically pleasing visualization. We wanted to find a geographical relationship between sites and answer the following 2 questions:
  - ❖ Are the launch sites close to the Equator line?
  - ❖ Are the launch sites in the close proximity to the coast?
- Haversine formula is used to calculate the distance between two points on the map based on its coordinates
- https://github.com/arandelov/test_repo/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The following graphs are added:
  - ❖ Success launches by site: pie chart
  - ❖ Success vs payload mass for all sites: scatter plot
- We wanted to answer the following questions:
  - ❖ Site with largest successful launches?
  - ❖ Which site has the largest success rate?
  - ❖ Which payload range(s) has the highest launch success rate?
  - ❖ Which payload range(s) has the lowest launch success rate?
  - ❖ Which F9 booster version has the highest launch success rate?
- https://github.com/arandelov/test_repo/blob/main/data_science_capstone_dash.ipynb

Predictive Analysis (Classification)

- After loading the data, we split it into training and test set, and performed a grid search after deciding which model to use. Then we calculated the accuracy for each model, and based on obtained hyperparameters chose the best model and plotted the confusion matrix.
- You need present your model development process using key phrases and flowchart
- https://github.com/arandelov/test_repo/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

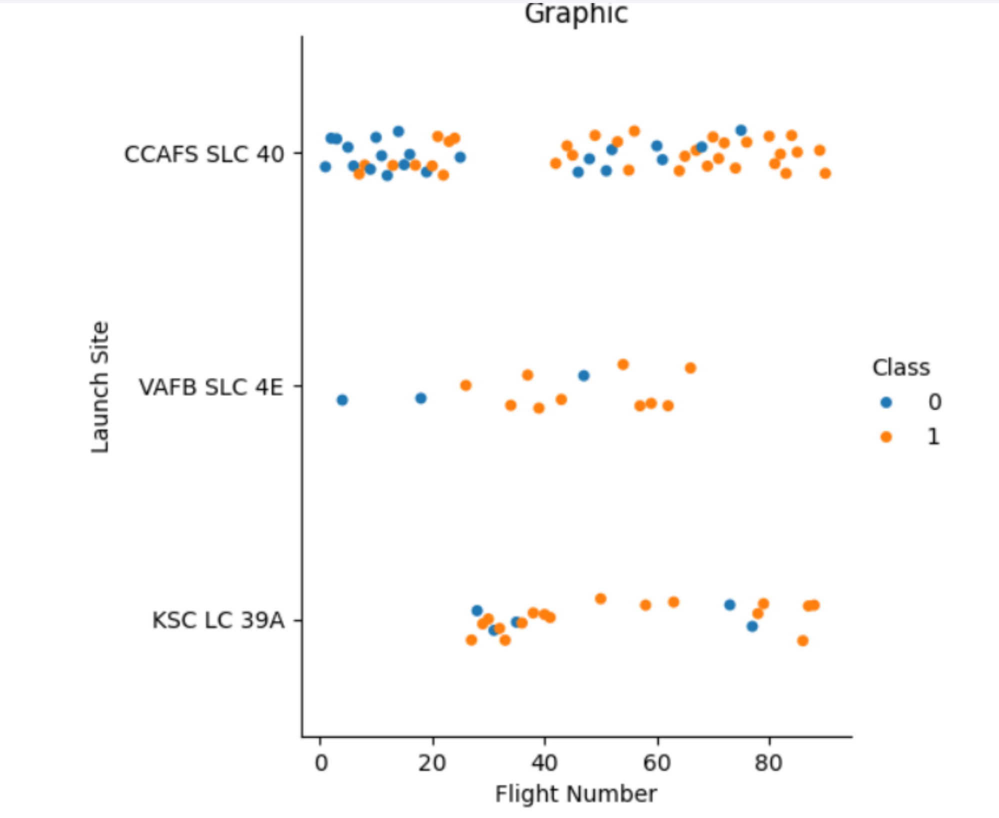
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

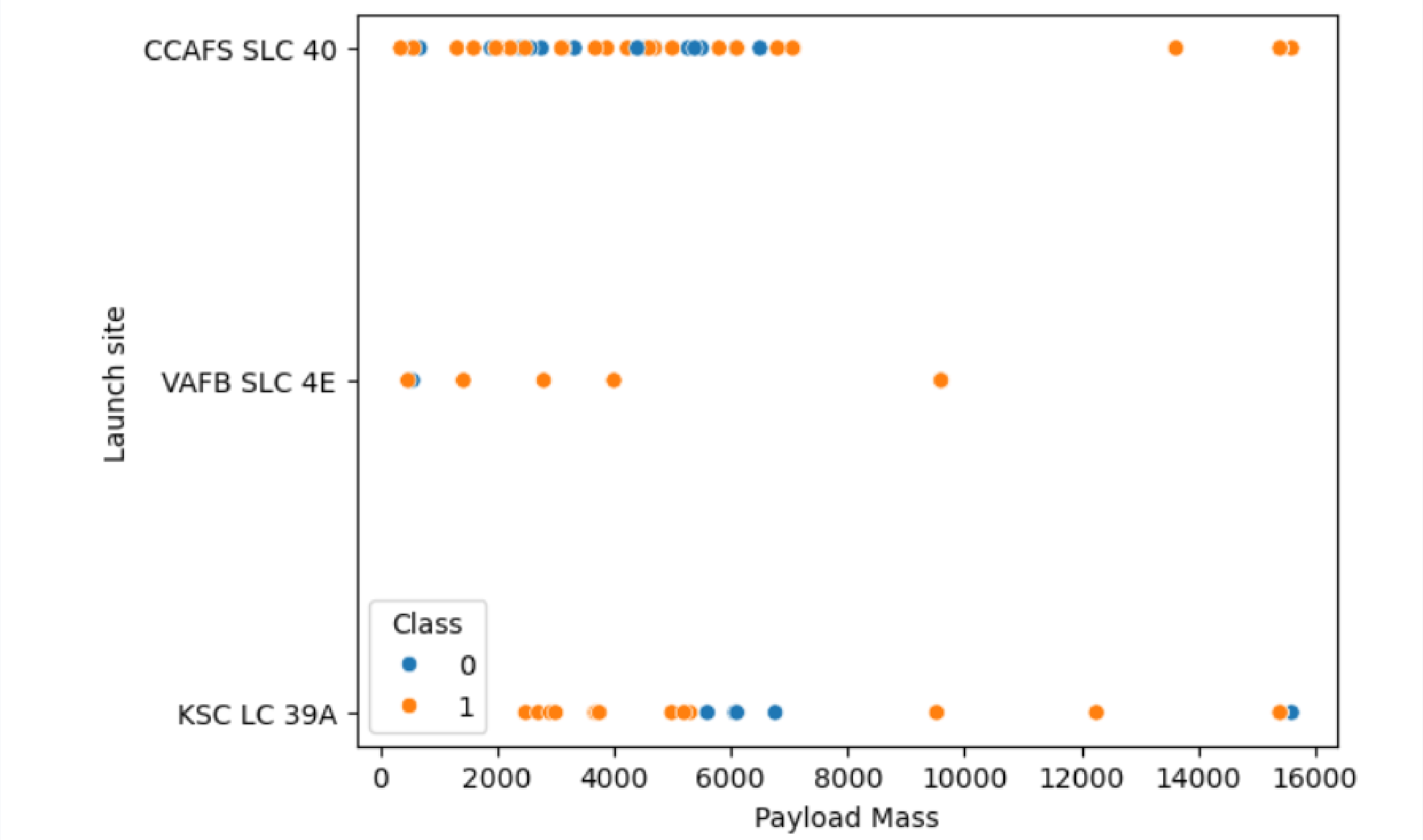
- Show a scatter plot of Flight Number vs. Launch Site
- The plot shows that the more flights we have (flight number higher), the greater success on the launch site will be.



Payload vs. Launch Site

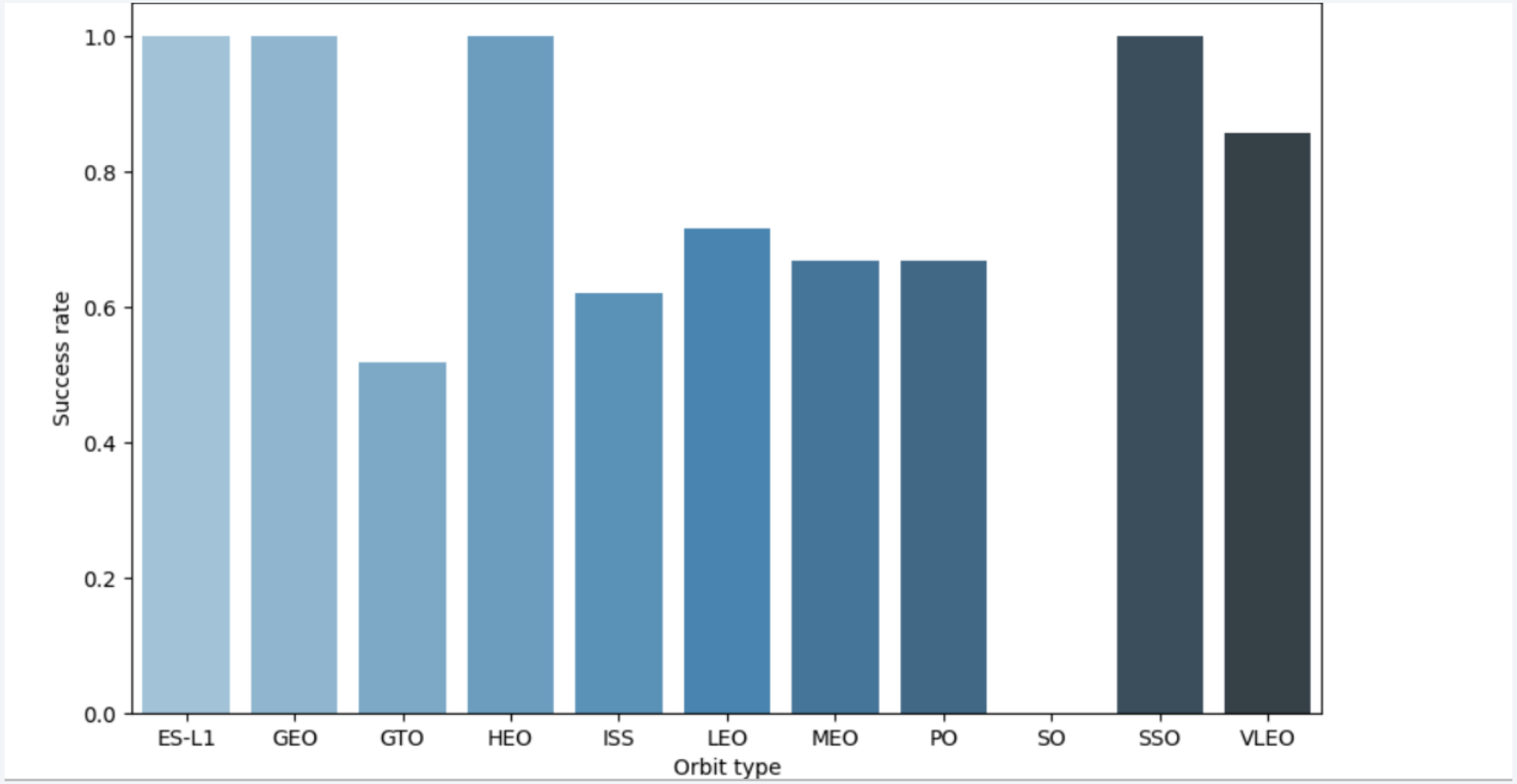
- Show a scatter plot of Payload vs. Launch Site

- The success rate is higher for lower payload mass in general for all 3 sites, but there is no definite pattern other than that. Also, the success is high for higher masses with some exceptions. For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).



Success Rate vs. Orbit Type

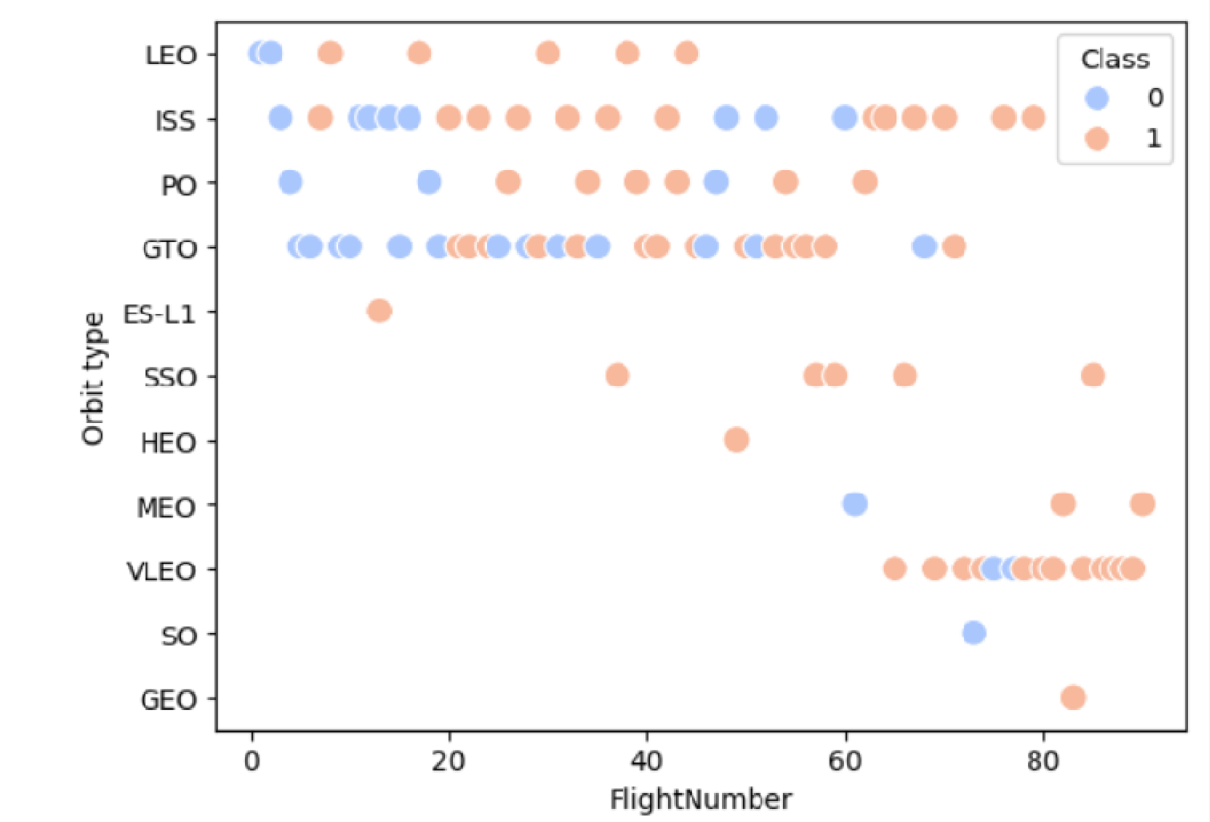
- Show a bar chart for the success rate of each orbit type
- For ES-L1, GEO, HEO and SSO the success rate is 100%, while SO orbit had a 0% success rate. Other orbits had successes in between these values, greater than 50% in general. However, for some sites there was only a single launch (hence the success of 100%), so the data is not adequate for drawing such success conclusions prematurely.



Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

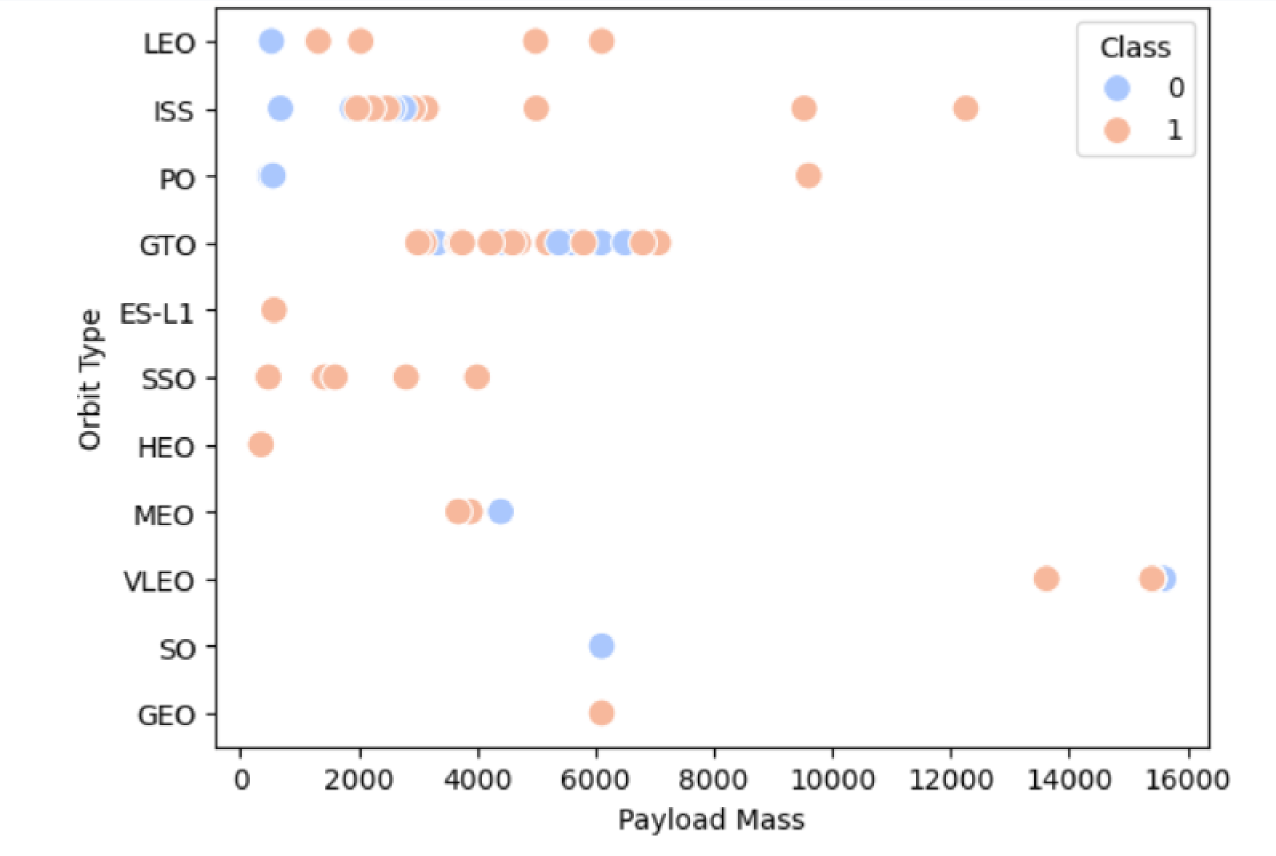
- The more flights on each orbit, the greater the success rate overall. For GTO, LEO and ISS the success is found also for a smaller flight number compared to other orbits.



Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

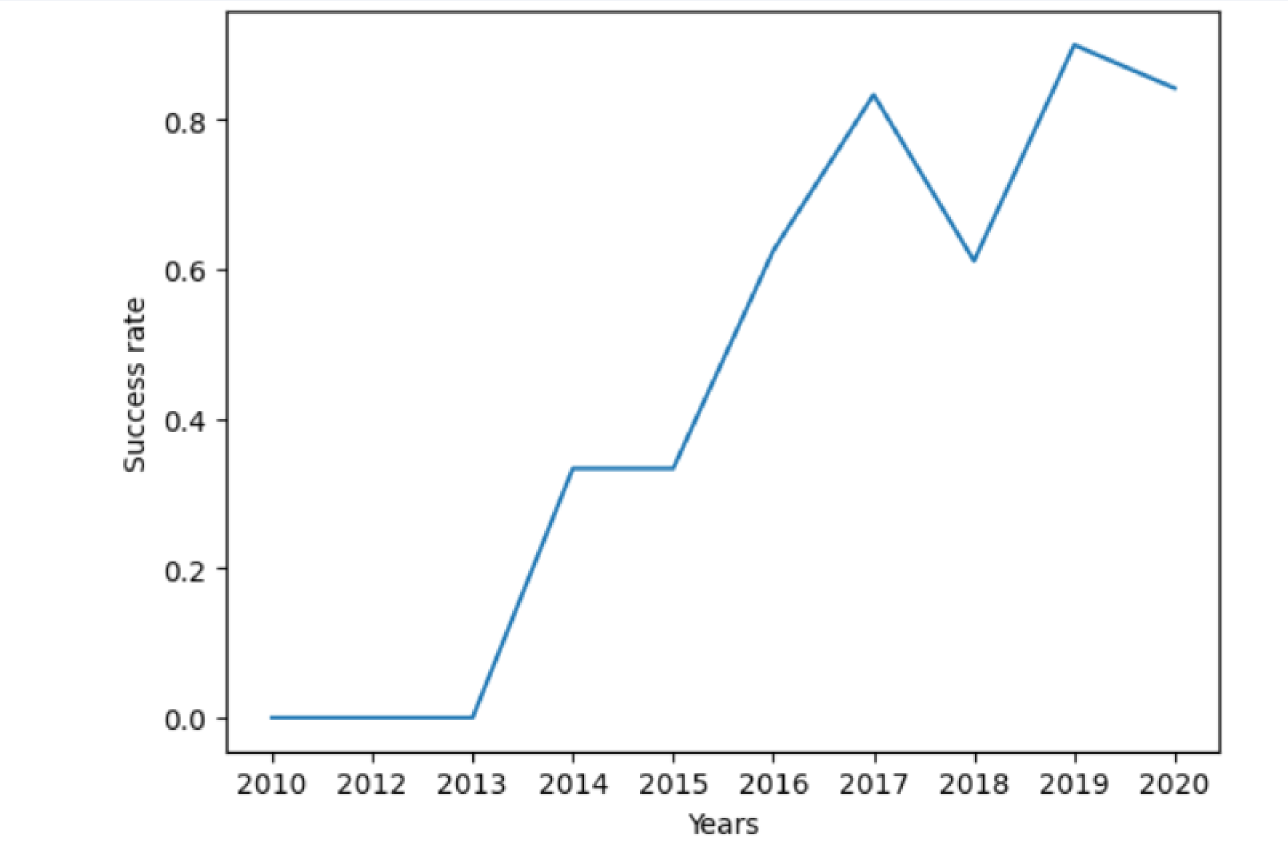
- For ES-L1, SSO and HEO, less payload mass leads to greater success. For GTO, the conclusion cannot be drawn. For LEO and ISS more payload mass leads to better outcome.



Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- There is an upward general trend: from 2010 to 2013 launches weren't successful at all, then the figure witnessed a steady rise from 2013 to 2020, experiencing a slight drop In 2018.



All Launch Site Names

- Unique launch sites were found using the DISTINCT method in an sql query

Task 1

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

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

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

Launch Site Names Begin with 'CCA'

- Displayed 5 records where launch sites begin with `CCA`

Task 2

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

Total Payload Mass

- Calculated the total payload carried by boosters from NASA

Task 3

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

```
print(df.columns)

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

```
Index(['Date', 'Time (UTC)', 'Booster_Version', 'Launch_Site', 'Payload',
       'PAYLOAD_MASS__KG_', 'Orbit', 'Customer', 'Mission_Outcome',
       'Landing_Outcome'],
      dtype='object')
 * sqlite:///my_data1.db
Done.
```

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- Calculated the average payload mass carried by booster version F9 v1

Task 4

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

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

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

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- Found the dates of the first successful landing outcome on ground pad

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 SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (ground pad)%';
```

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

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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 SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (drone ship)%'
AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" <6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculated the total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%%sql
SELECT "Mission_Outcome", COUNT(*) AS "Total"
FROM SPACEXTABLE
GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

- Listed the names of the booster which have carried the maximum payload mass

Task 8

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

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

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

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

- Listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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
    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",
    "Booster_Version",
    "Launch_Site",
    "Landing_Outcome"
FROM SPACEXTABLE
WHERE substr("Date", 0, 5) = '2015'
  AND "Landing_Outcome" LIKE 'Failure (drone ship)%';
```

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

Month_Name	Booster_Version	Launch_Site	Landing_Outcome
January	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Rank 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

Task 10

Rank 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.

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

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

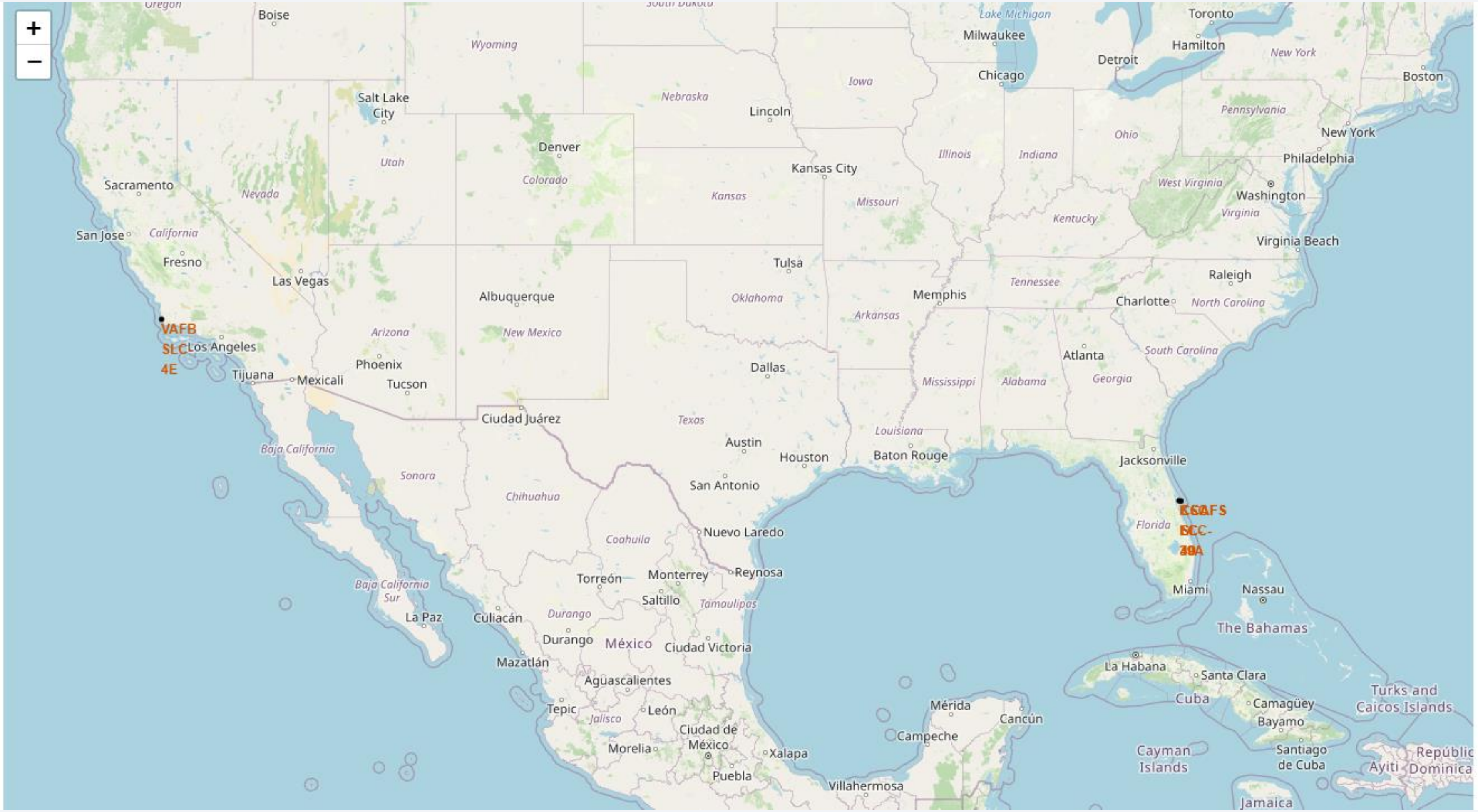
Landing_Outcome	Date	Landing_Count
No attempt	2012-05-22	10
Success (drone ship)	2016-04-08	5
Failure (drone ship)	2015-01-10	5
Success (ground pad)	2015-12-22	3
Controlled (ocean)	2014-04-18	3
Uncontrolled (ocean)	2013-09-29	2
Failure (parachute)	2010-06-04	2
Precluded (drone ship)	2015-06-28	1

Section 3

Launch Sites Proximities Analysis

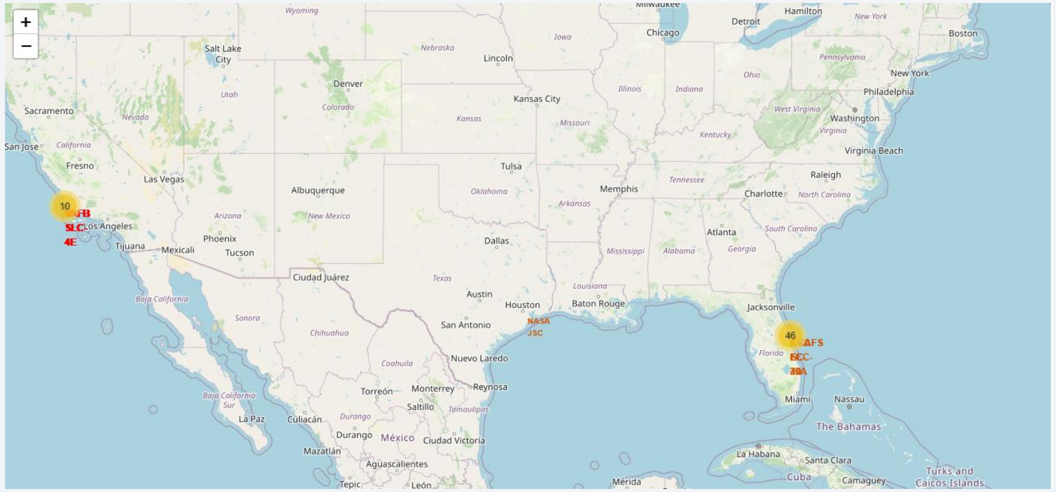
Markers for launch sites on a world map

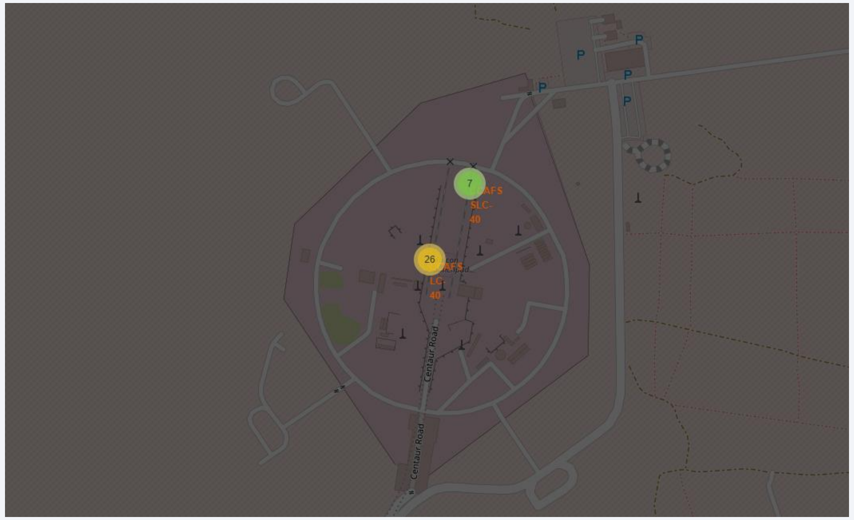
- The map with marked launched sites

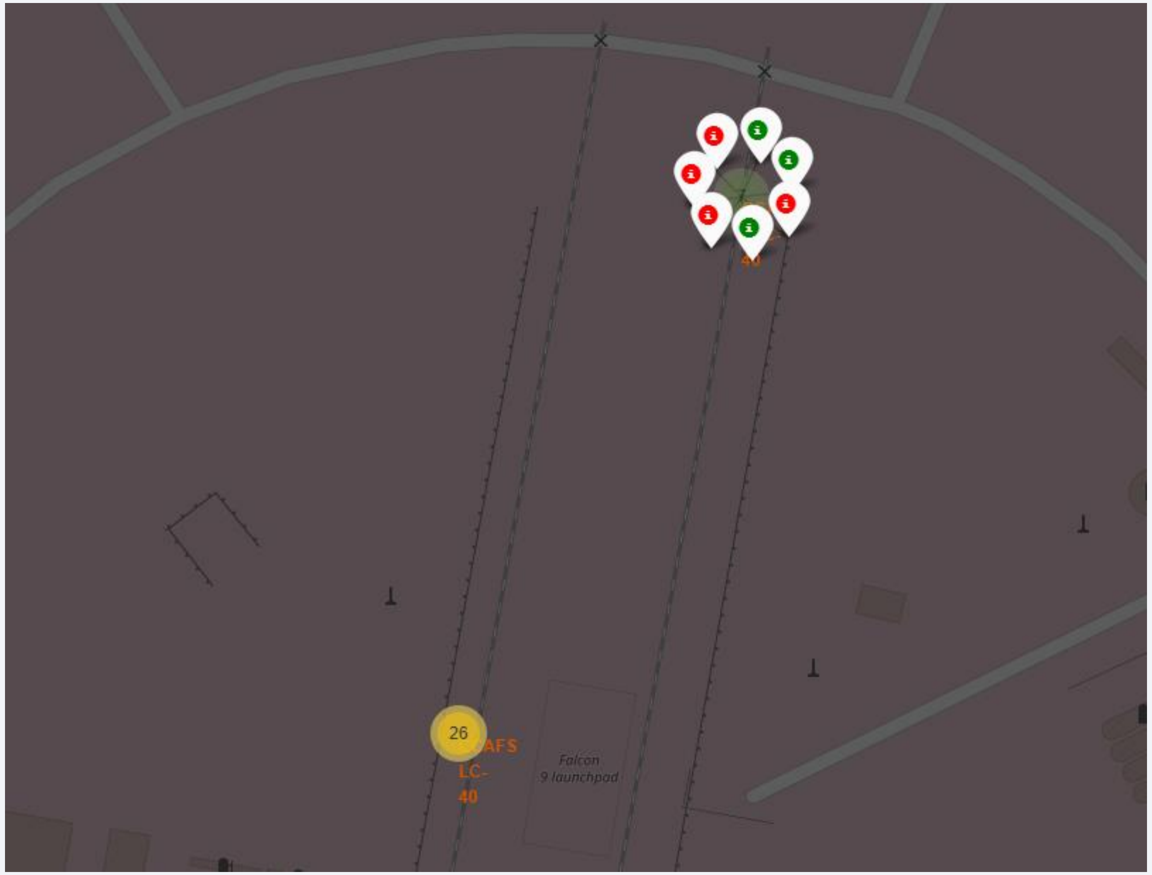


Map with a number of successful launches

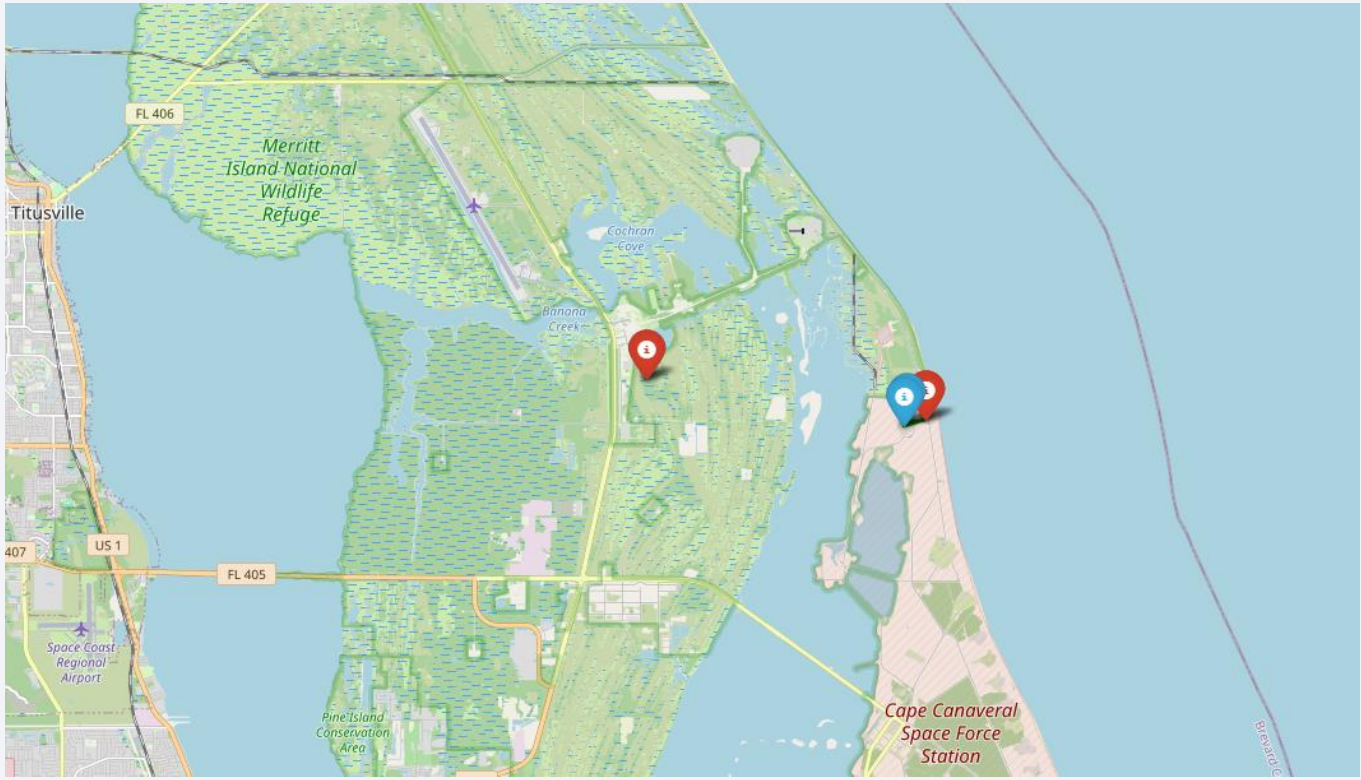
- For each launch site on the map the number of successful launches is calculated and marked







Proximity of launch sites to other sites or coastlines

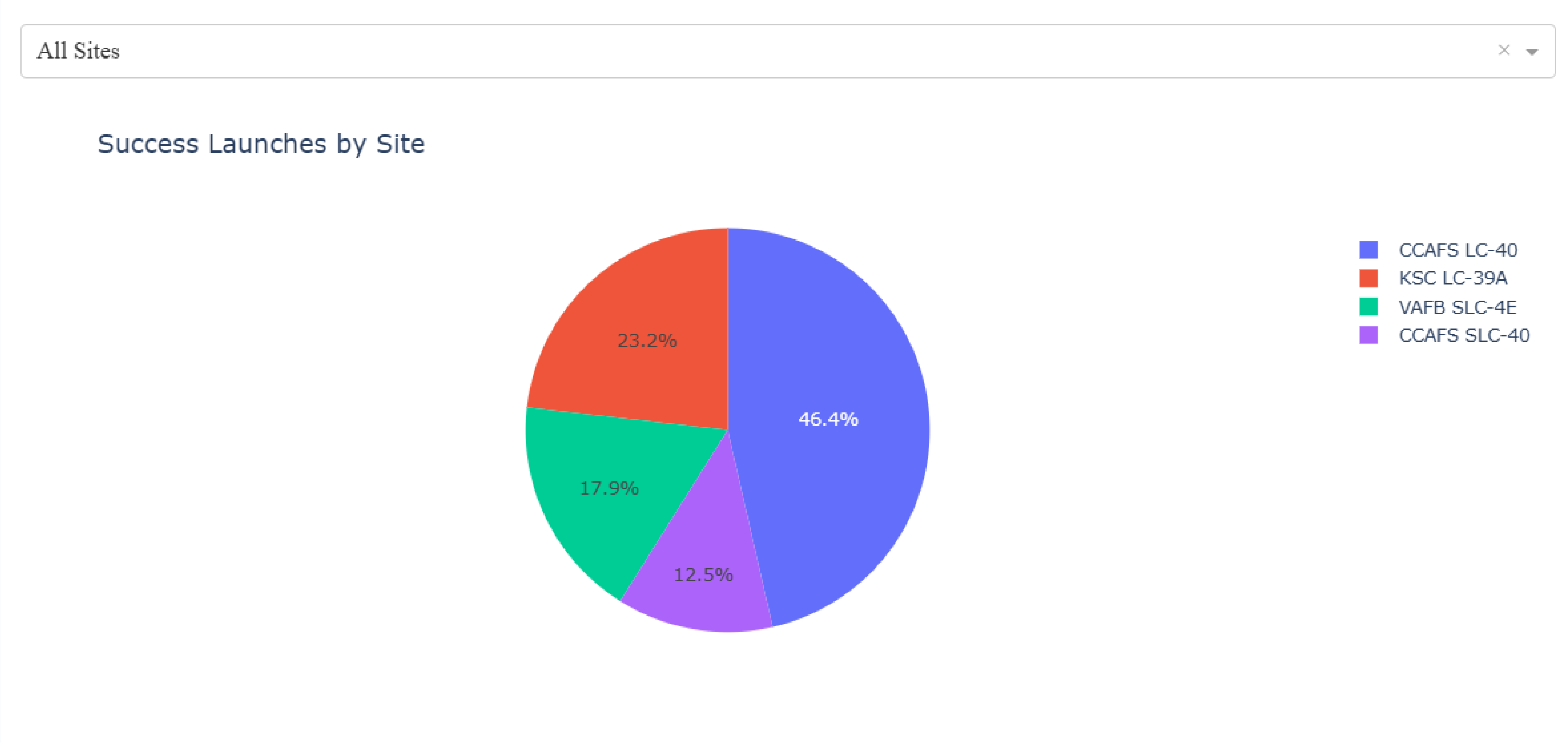


Section 4

Build a Dashboard with Plotly Dash

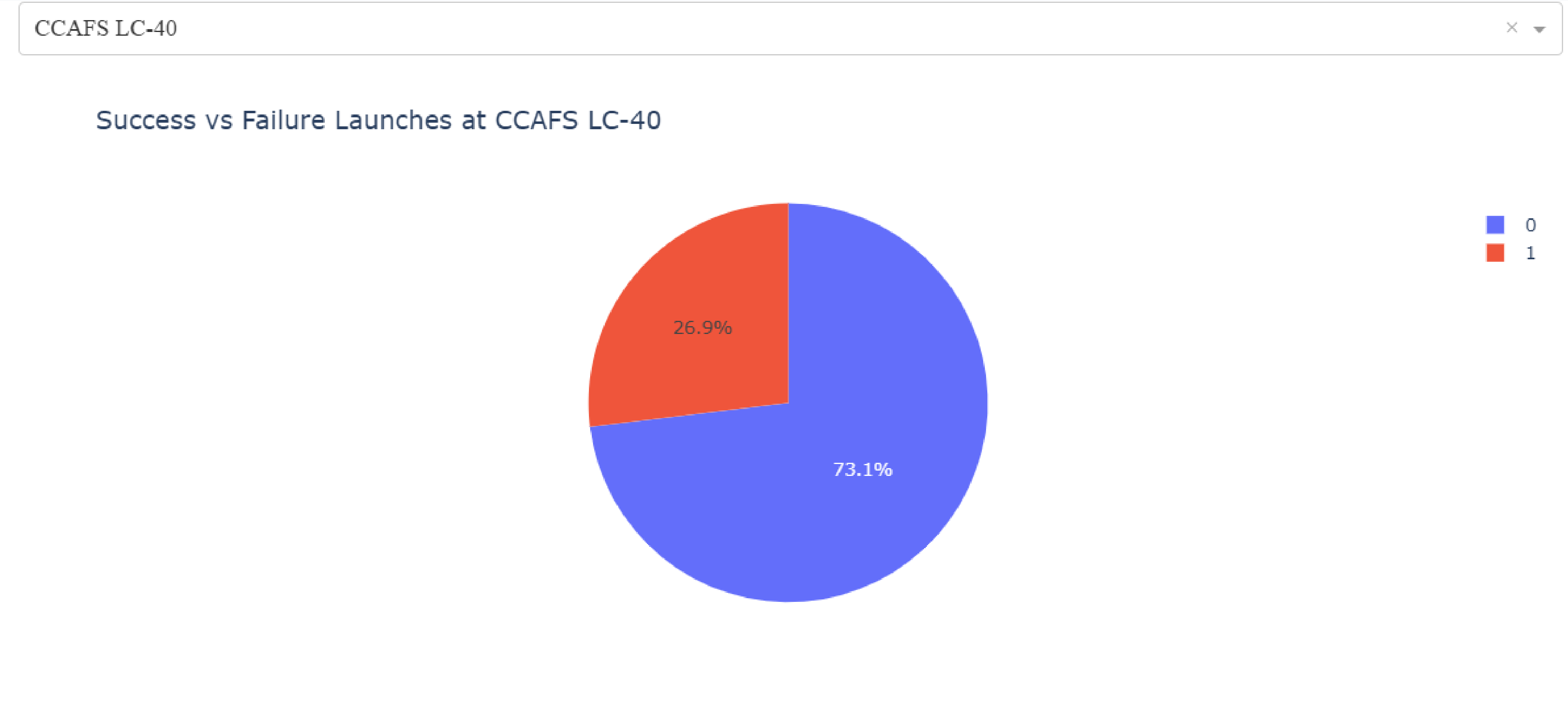
Success launches by site

- CCAFS LC-40 has the highest number of launches, 47% of the total, followed by KSC LC-39A and VAFB SLC – 4E with 23% and 17% respectively, while CCAFS LC-40 has the least percentage of launches of 12%



Success vs failure of launch for the highest launch number site

- CCAFS LC-40 has 73% successful and 27% unsuccessful launches



Success vs payload for all launch sites

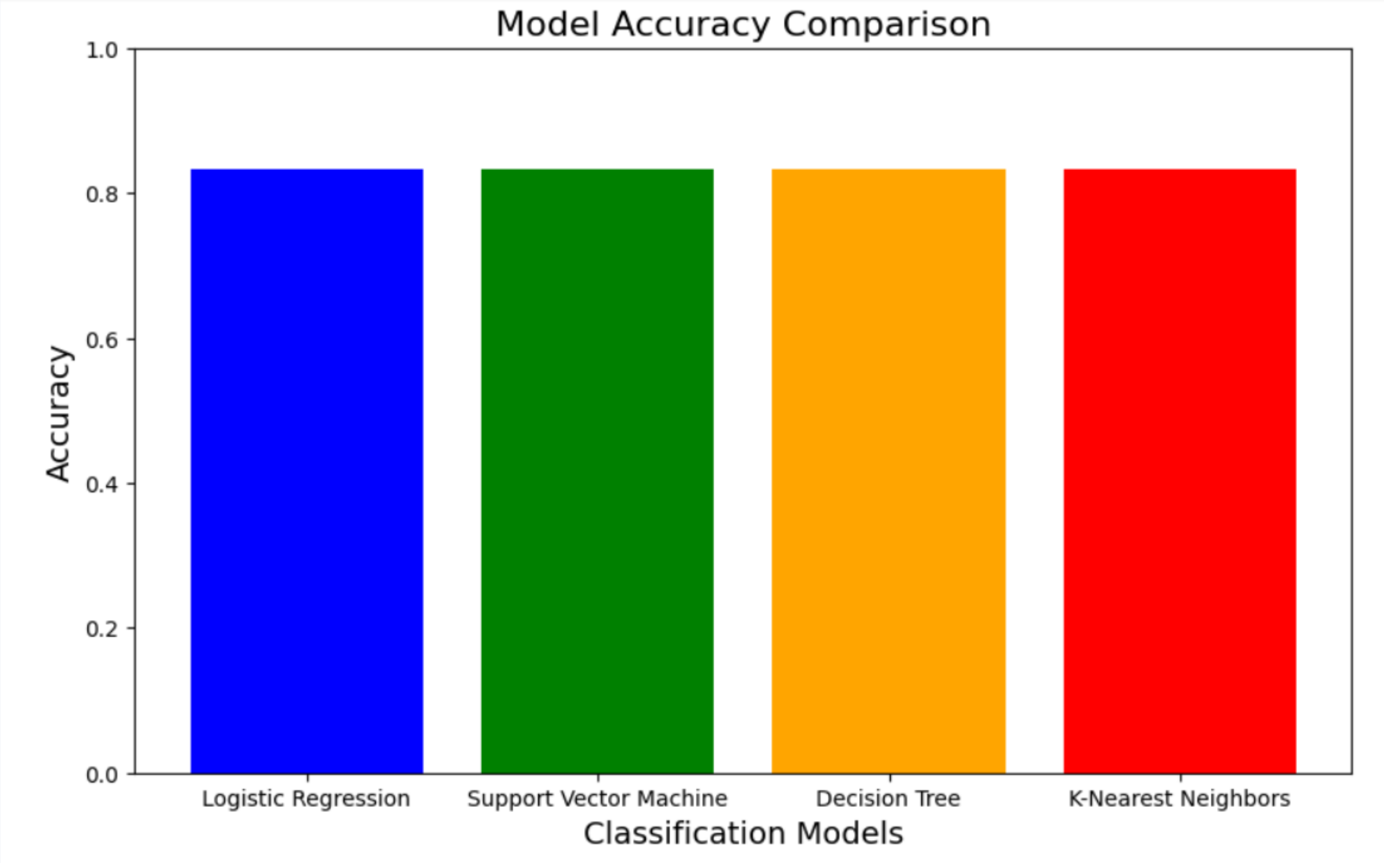
- Success rate is higher for lower payloads as opposed to the higher ones

Section 5

Predictive Analysis (Classification)

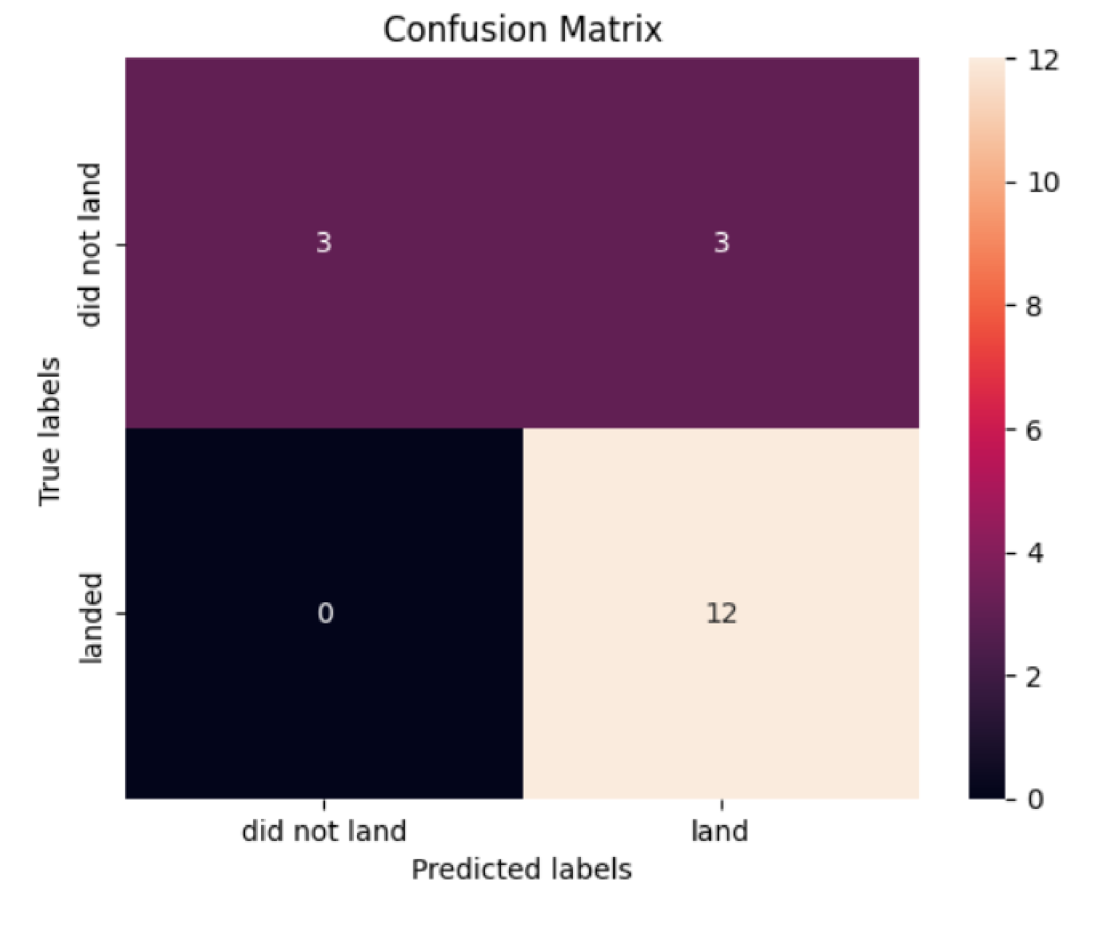
Classification Accuracy

- Visualized the built model accuracy for all built classification models, in a bar chart
- All models have equal classification accuracy



Confusion Matrix

- Confusion matrix shows that classifier can distinguish different classes for every model



Conclusions

- The larger the number of flights, the greater the success rate
- The success rate of the launches increases overall throughout the surveyed period (2010 – 2020)
- All models have equal accuracy fro this task

Thank you!