IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Konstantinos Grousouzakos
3/2/2024

Outline

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

Executive Summary

- Summary of methodologies

  i. Data Collection

  ii. Data Wrangling

  iii. Exploratory Data Analysis

  iv. Interactive Visualizations

  v. Predictive Analysis

- Summary of all results

After careful examination, specific machine learning methods were selected as the most suitable for the desired prediction task

Introduction

- **Background :** SpaceX advertises its Falcon 9 rocket launches for $62 million, much cheaper than other providers who charge over $165 million. This is because SpaceX can reuse the first stage of its rockets. Knowing if the first stage will land successfully helps estimate the launch cost. This is important for companies bidding against SpaceX. In this lab, we'll collect and format data from an API to predict if a launch will succeed or fail.
- **Goal** : Provide a model that predicts if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was gathered from both SPACEX API and through webscraping of the relevant Wikipedia page.
- Perform data wrangling:
  - The dataset underwent preprocessing to address any missing values and to prepare it for applications of machine learning methods in order to derive predictions.
- Perform exploratory data analysis (EDA) using visualization and SQL:
  - The data was subjected to exploration through analysis using charts, graphs and examination of specific metrics. This approach aimed to gain insights into the nature of the variables within the dataset.

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash:
    - Interactive analytics and dashboards were developed to further explore the dataset and facilitate dynamic insights.
- Perform predictive analysis using classification models:
    - Multiple machine learning models were deployed to generate predictions. These models were initially fine-tuned to enhance their performance. Subsequently, their outputs were compared and evaluated.

Data Collection

API

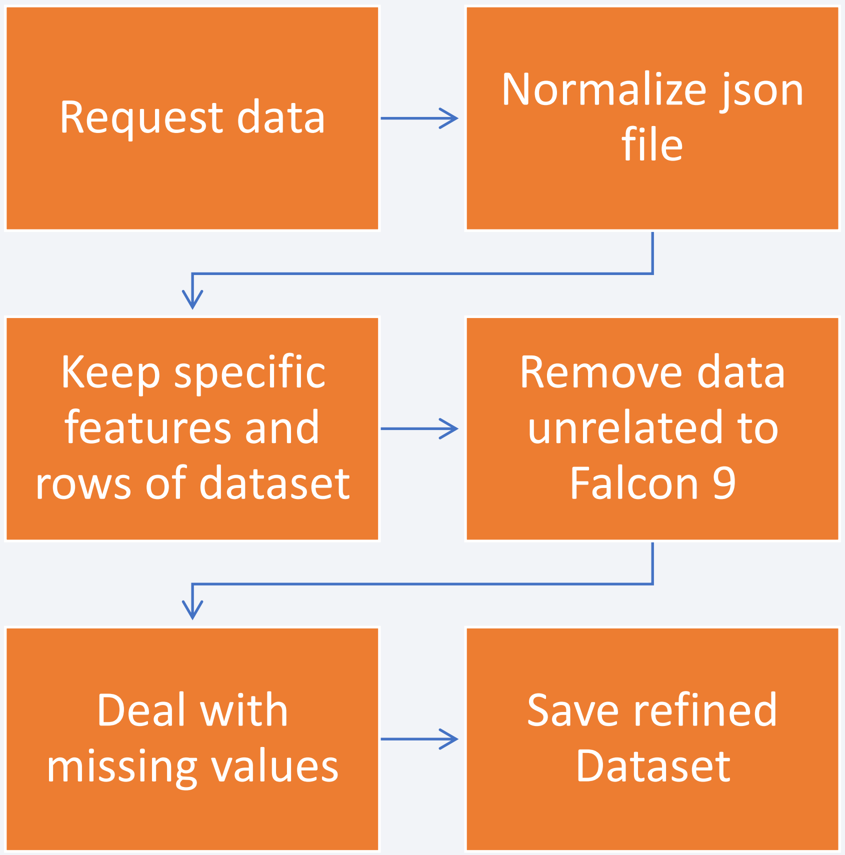
- Extract the data from API
- Perform Data Cleaning

Web Scraping

- Extract the data from the Wikipedia page
- Convert the data to a usable pandas datarame

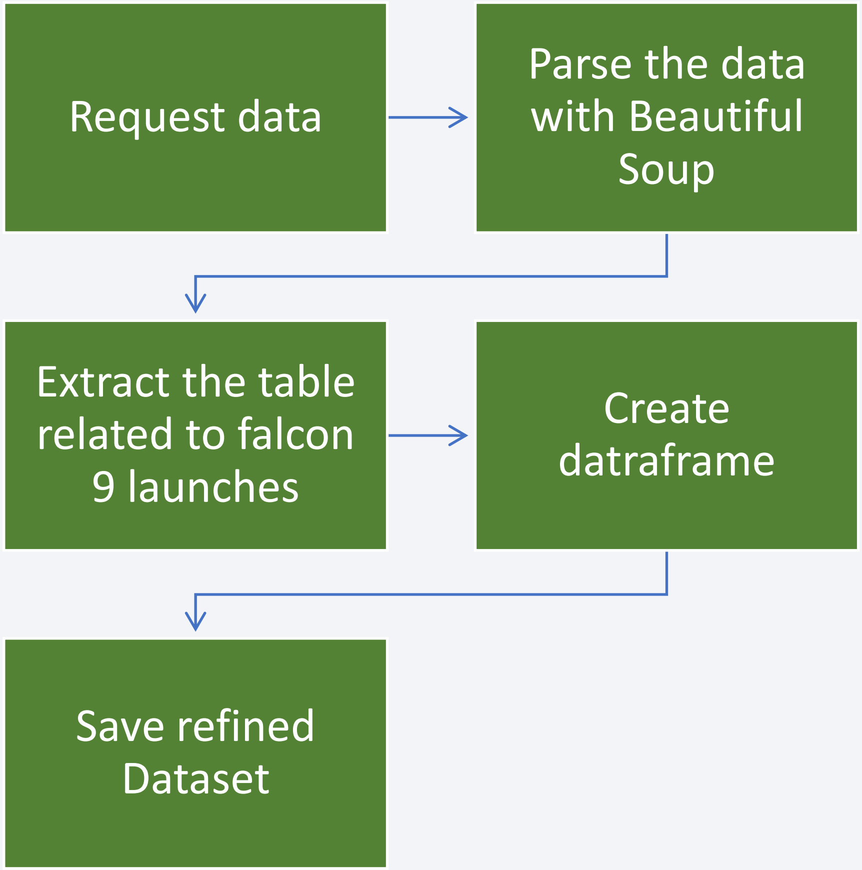
Data Collection – SpaceX API

- Request and parse the SpaceX launch data using the GET request
- Filter the dataframe to only include Falcon 9 launches
- Substitute missing values with mean of the corresponding column
- Save the dataset
- **GitHub URL :** https://github.com/Konstantinos-Grousouzakos/Portfolio/blob/main/Applied%20Data%20Science%20Capstone(IBM%20Data%20Science%20Professional%20Certificate)/1)%20jupyter-labs-spacex-data-collection-api.ipynb



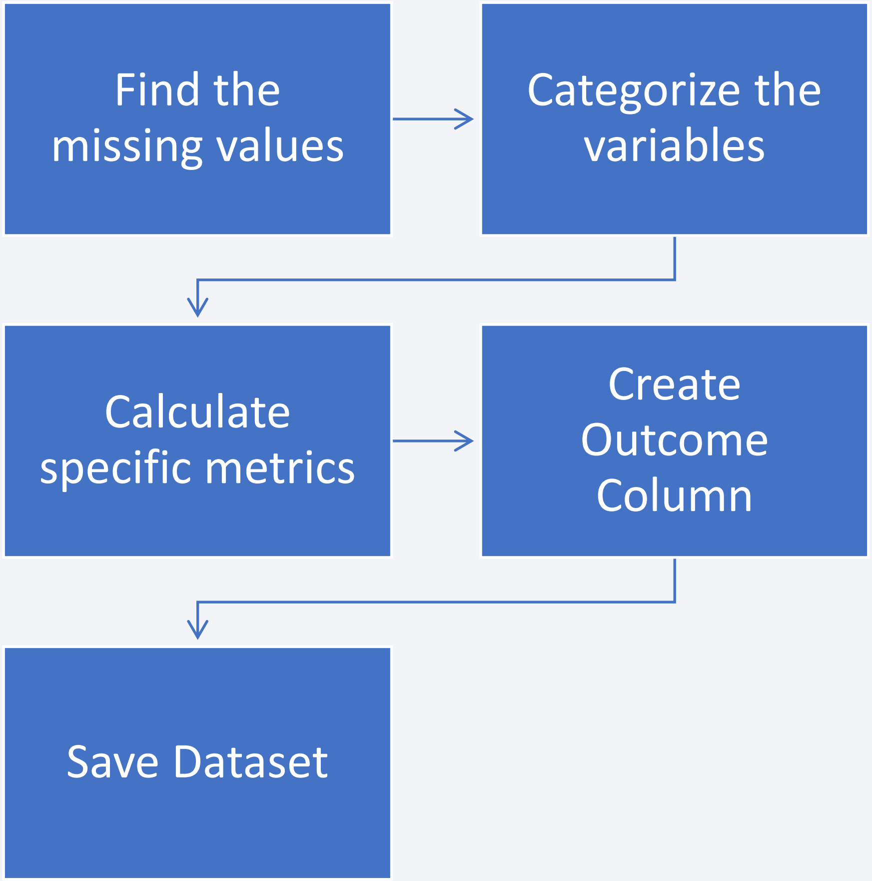
Data Collection - Scraping

- Request The Falcon 9 Launch Wiki page from its URL
- Extract all variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables
- Save the dataset
- **GitHub URL :** https://github.com/Konstantinos-Grousouzakos/Portfolio/blob/main/Applied%20Data%20Science%20Capstone(IBM%20Data%20Science%20Professional%20Certificate)/2)%20jupyter-labs-webscraping.ipynb



Data Wrangling

- Identify and calculate the percentage of the missing values in each attribute
- Identify which columns are numerical and categorical
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome of the orbits
- Assign the various outcomes of the launch into categories (0 for failure and 1 for success), thus creating the Outcome column.
- Save the dataset
- **GitHub URL :** https://github.com/Konstantinos-Grousouzakos/Portfolio/blob/main/Applied%20Data%20Science%20Capstone(IBM%20Data%20Science%20Professional%20Certificate)/3)%20labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Charts that we used include:
  1. A Cat plot for identifying correlation between variables 'Flight Number' and 'Payload Mass'
  2. A Cat plot for identifying correlation between variables 'Flight Number' and 'Launch Site'
  3. A Cat plot for identifying correlation between variables 'Launch Site' and 'Payload Mass'
  4. A Cat plot for identifying correlation between variables 'Flight Number' and 'Orbit Type'
  5. A Cat plot for identifying correlation between variables 'Orbit Type' and 'Payload Mass'
  6. A Bar plot in order to inspect the average success rate per orbit
  7. A Line plot to examine the yearly average success rate
- **GitHub URL :** https://github.com/Konstantinos-Grousouzakos/Portfolio/blob/main/Applied%20Data%20Science%20Capstone(IBM%20Data%20Science%20Professional%20Certificate)/5)%20jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries that were performed:
    1. Display the names of the unique launch sites
    2. Display 5 records where launch sites begin with the string ‘CCA’
    3. Display the total payload mass carried by boosters launched by ‘NASA (CRS)’
    4. Display the average payload mass carried by booster version ‘F9 v1.1’
    5. List the date when the first successful landing in ground pad was achieved
    6. List the names of the booster which have success in drone ship and payload mass greater than 4000 but less than 6000 kg
    7. List the total number of successful and failure mission outcomes
    8. List the names of the booster versions which have the maximum payload mass
    9. List the records displaying month names, failure landing outcomes in drone ship, booster version, launch sites for the months in year 2015
    10. Rank the count of landing outcomes between dates 2010-06-04 and 2017-03-20, in descending order
- **GitHub URL :** https://github.com/Konstantinos-Grousouzakos/Portfolio/blob/main/Applied%20Data%20Science%20Capstone(IBM%20Data%20Science%20Professional%20Certificate)/4)%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map objects that were added to the folium map:

    1. Circles and markers on the launch sites. Adding these elements aids in easily identifying the launch sites that are in close proximity to the equator line or to the coast.

    2. Color-coded Markers displaying the successor failure of launches for each site in order to facilitate the identification of launch with relatively high success rates.

    3. Lines connecting launch site to specific points in the map. These lines measure the distance of the launch site to these points, making it easier examine the selection of the location of the launch site.

- **GitHub URL :** https://github.com/Konstantinos-Grousouzakos/Portfolio/blob/main/Applied%20Data%20Science%20Capstone(IBM%20Data%20Science%20Professional%20Certificate)/6)%20lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Plots and Graphs that were added to the Dashboard:

  1. A pie chart displaying the Total Success Launches Per Site in order to identify the which sites produce successful launches more often.

  2. A pie chart displaying the Total Success Launches for each Site in order to compare the success and failure rate per site.

  3. A cat plot displaying the correlation between Mission Outcome and Payload mass per booster version. Helps identifying if a booster version is more successful than the rest.

- **GitHub URL :** https://github.com/Konstantinos-Grousouzakos/Portfolio/blob/main/Applied%20Data%20Science%20Capstone(IBM%20Data%20Science%20Professional%20Certificate)/7)%20spacex_launch_app.ipynb

Predictive Analysis (Classification)

- Process of building Plots and Graphs that were added to the Dashboard:

    1. Initially, the Y (target) variable is defined and the X (predictors) are scaled. Then the dataset is split into train and tests sets.

    2. The machine learning methods that were utilized are Logistic Regression, Support Vector Machines, Decision Tree and K Nearest Neighbors. For each model, a set of tuning hyperparameters was defined. Later each model was optimized via the GridSearchCV method, which iterates through all the different combination of the hyperparameters and selects the combination that produces the best accuracy score.

    3. The optimized models were trained using the training set. Then the performance of each model was assessed by inspecting the accuracy score on the test set and by examining the confusion matrix.

- **GitHub URL :** https://github.com/Konstantinos-Grousouzakos/Portfolio/blob/main/Applied%20Data%20Science%20Capstone(IBM%20Data%20Science%20Professional%20Certificate)/8)%20SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

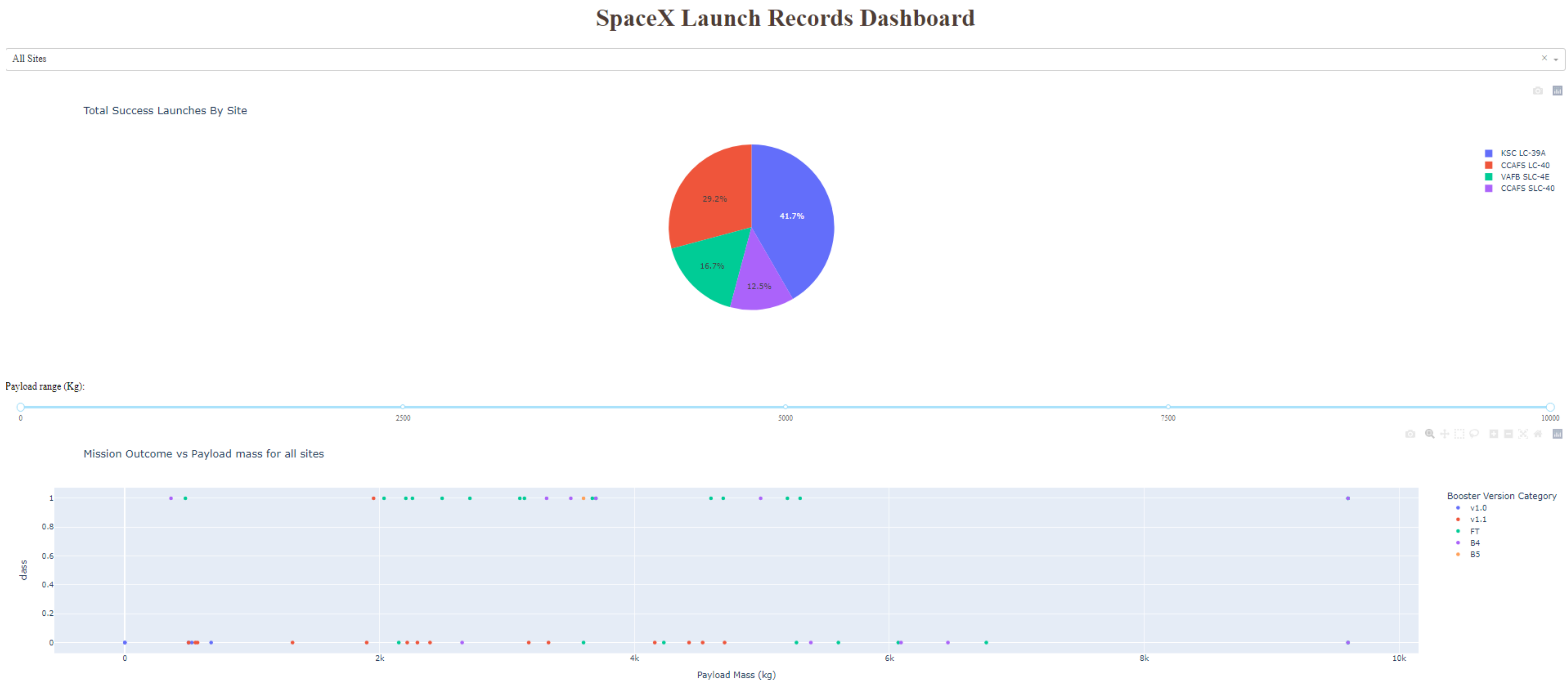
Results

- Exploratory data analysis results

1. There is a correlation between Flight number and success launches across all sites
2. Launch site “KSC LC 39A” and “VAFB SLC 4E” demonstrate a higher success rate.
3. Orbits “ES-L1”, “GEO”, “HEO” and “SSO” exhibit notably high average success rates
4. Payload 2000-6000 kg has the highest success rate.
5. Payload 6000-8000 kg has the lowest success rate.
6. From the F9 booster versions, “FT” has the highest success rate.

Results

- Interactive analytics demo in screenshots



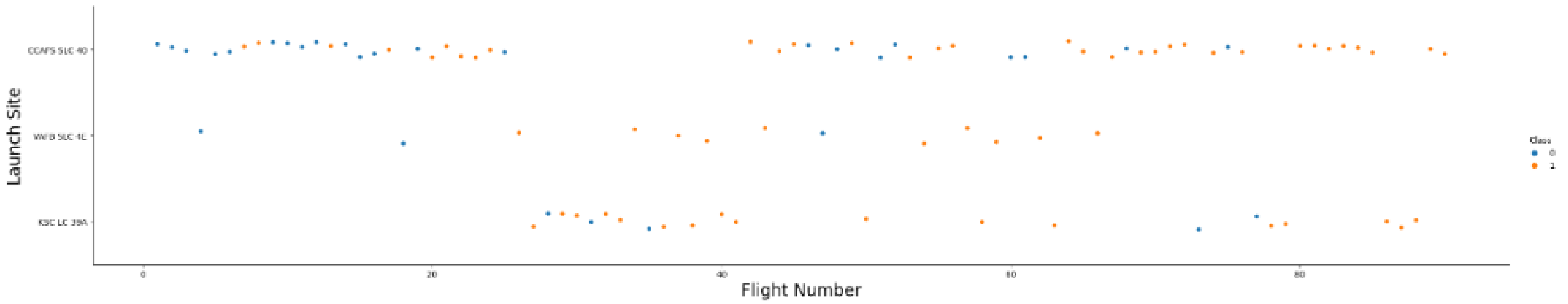
Results

- Predictive analysis results :
  - No method seems to be superior to the rest, since both the accuracy scores and the confusion matrices are relatively similar, if not exactly alike.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



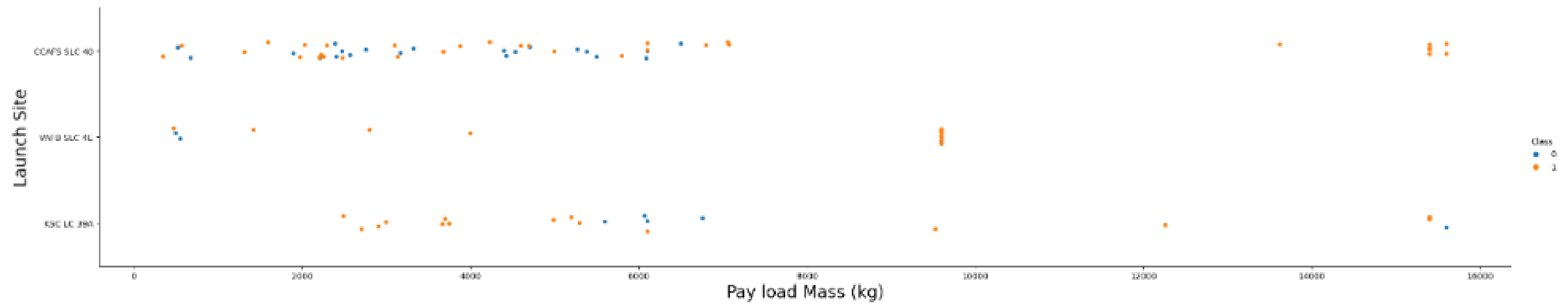
Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

Explanation :

1) There appears to be a correlation between the increase in flight numbers and the increase in successful launches across all launch sites.

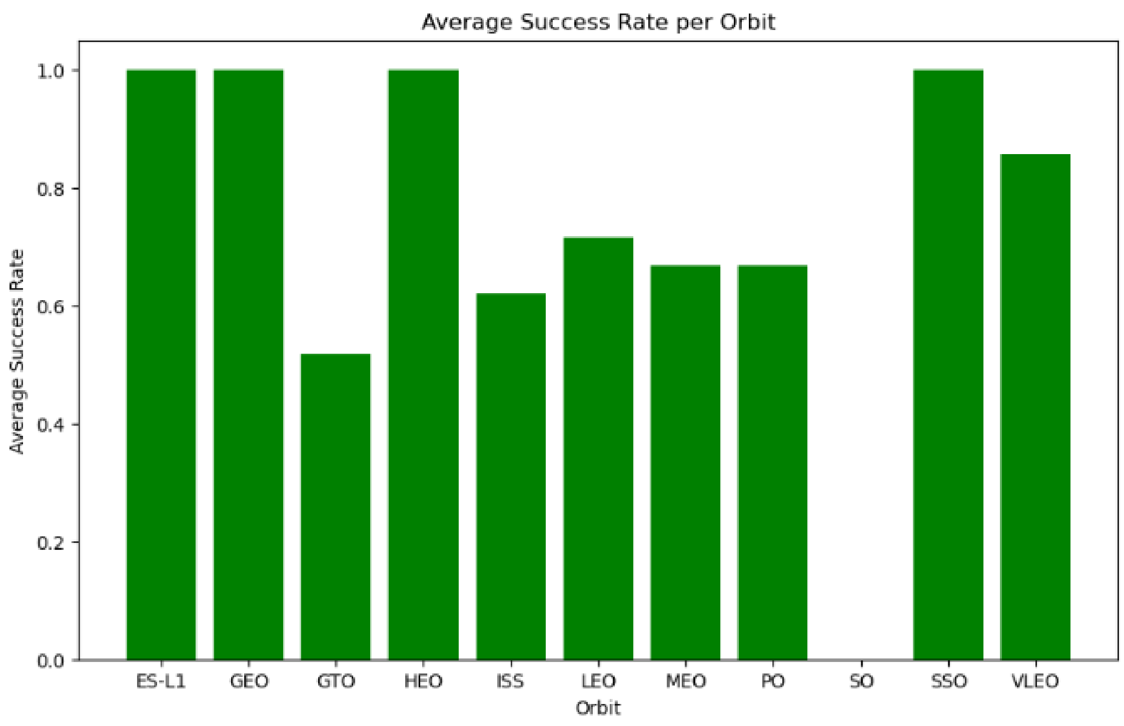
2) It appears that the launch sites "KSC LC 39A" and "VAFB SLC 4E" demonstrate a higher success rate compared to "CCAFS SLC 40".

Payload vs. Launch Site



Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



Analyze the ploted bar chart try to find which orbits have high sucess rate.

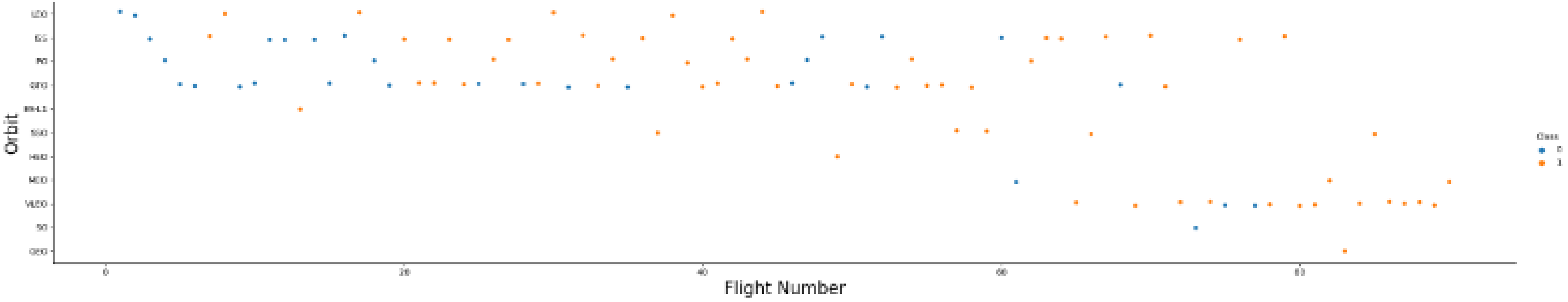
Explanation :

1) Orbits such as "ES-L1", "GEO", "HEO", and "SSO" exhibit notably high average success rates, approaching or even reaching 100%.

2) The orbit "VLEO" demonstrates a promising average success rate, slightly surpassing 80%.

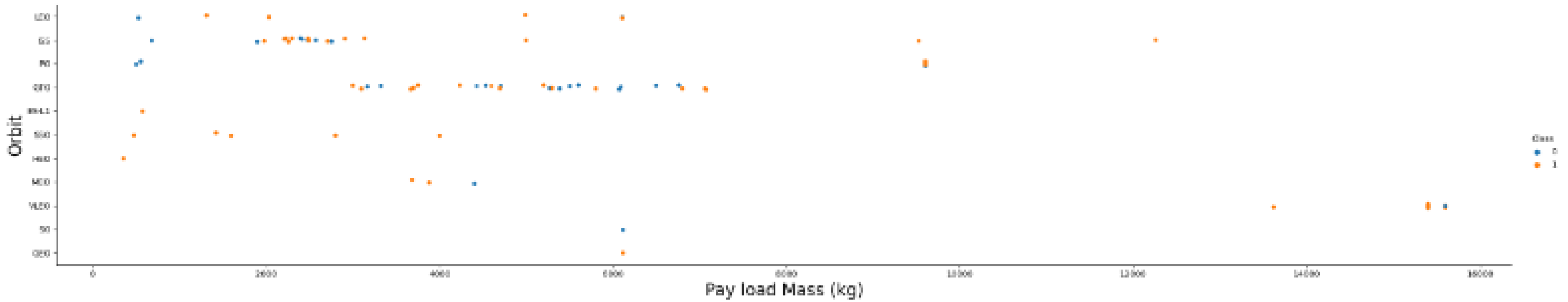
3) Conversely, the remaining orbits display average success rates below 70%, with "GTO" particularly notable for its average success rate hovering around 50%.

Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

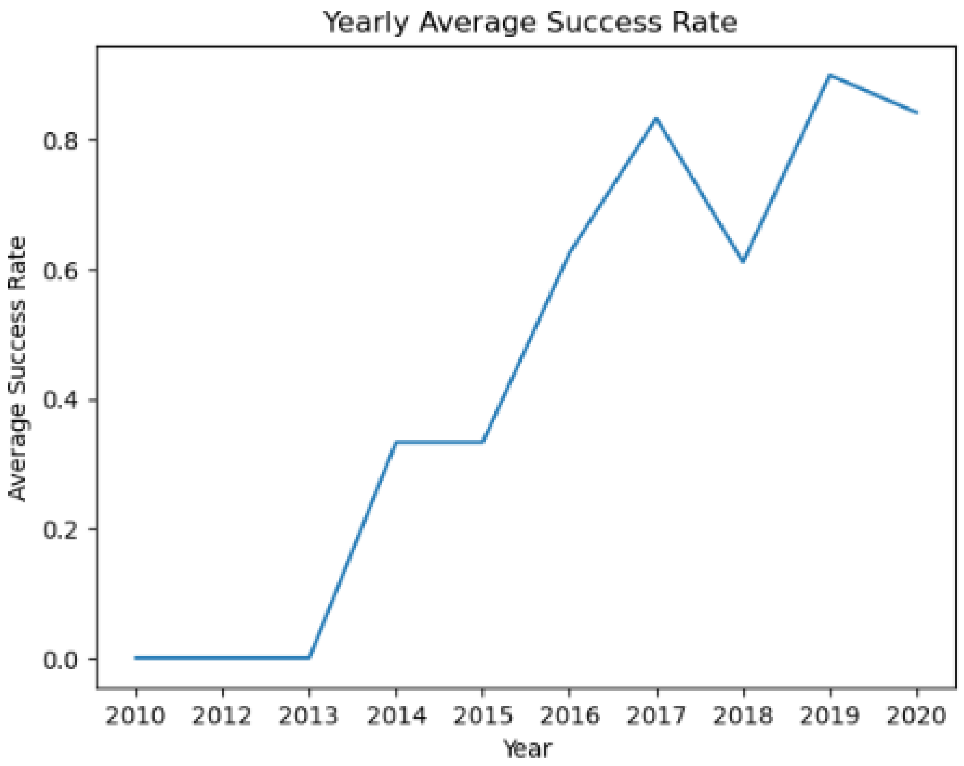


With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

```
Text(0, 0.5, 'Average Success Rate')
```



You can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

All Launch Site Names

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

```
%sql select distinct(Launch_Site) from spacextbl
```

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

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

4 different launch sites

Launch Site Names Begin with 'CCA'

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

```
%sql select * from spacextbl 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

The payload seems small

Total Payload Mass

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

```
%sql select sum(PAYLOAD_MASS__KG_) from spacextbl where Customer == 'NASA (CRS)';
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

```
%sql select AVG(PAYLOAD_MASS__KG_) from spacextbl where Booster_Version == 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

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)'
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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 distinct(booster_version) from spacextbl where (landing_outcome = '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

List the total number of successful and failure mission outcomes

```
%%sql

select count(Mission_Outcome) from spacextbl group by Mission_Outcome;
```

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

count(Mission_Outcome)
1
98
1
1

Boosters Carried Maximum Payload

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)
```

```
 * 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

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 substr(Date, 6,2) as month, landing_outcome, booster_version, launch_site from spacextbl where (
    landing_outcome == 'Failure (drone ship)' and substr(Date,0,5)='2015')
```

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

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

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

```
%%sql

select count(landing_outcome), landing_outcome from spacextbl where Date between '2010-06-04' and '2017-03-20'
group by landing_outcome
order by count(landing_outcome) Desc
```

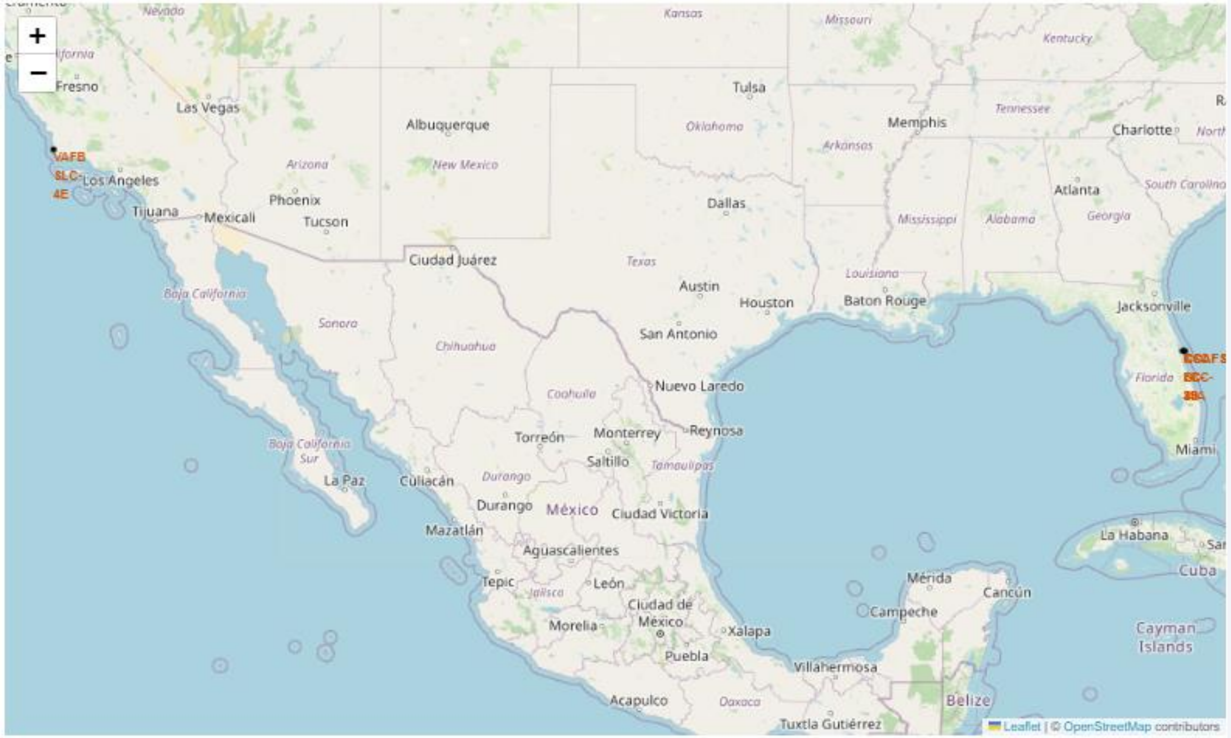
```
 * sqlite:///my_data1.db
Done.
```

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

Section 3

Launch Sites Proximities Analysis

Launch sites Locations

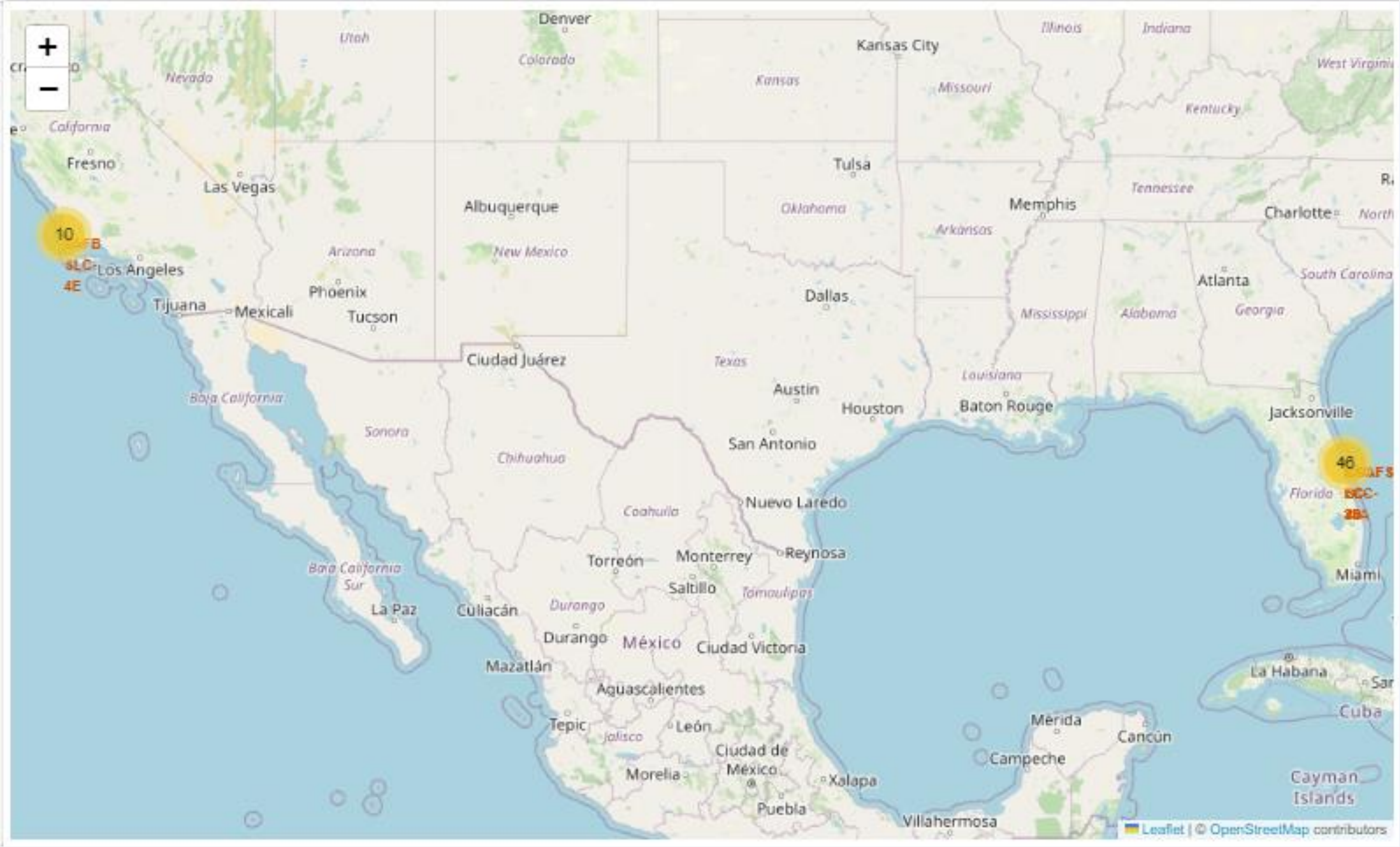


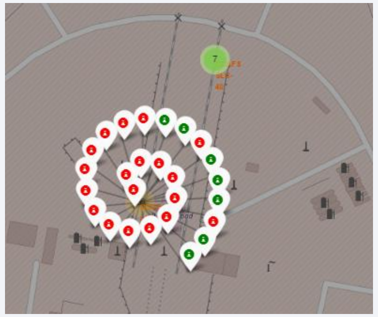
Explanation :

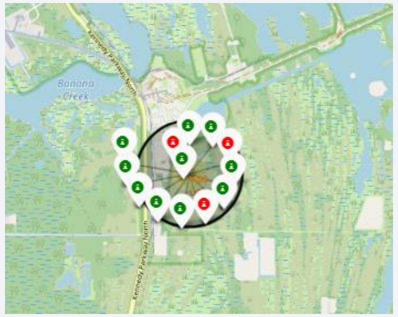
1) Launch sites are not all situated near the equator line because the United States is positioned slightly north of the equator, which is located in the northern part of Latin South America.

2) All launch sites are situated very close to coastlines. This proximity ensures safer experiments as they are conducte d away from populated areas. Additionally, it offers the opportunity for experiments involving water landings.

Successes and failures per launch site

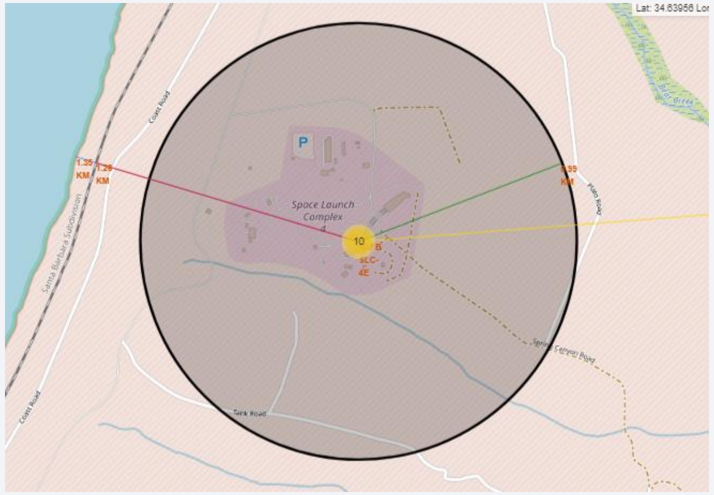






Explanation : We can visually inspect the success rates per site

Distance of Launch Site from points of interest



Explanation :

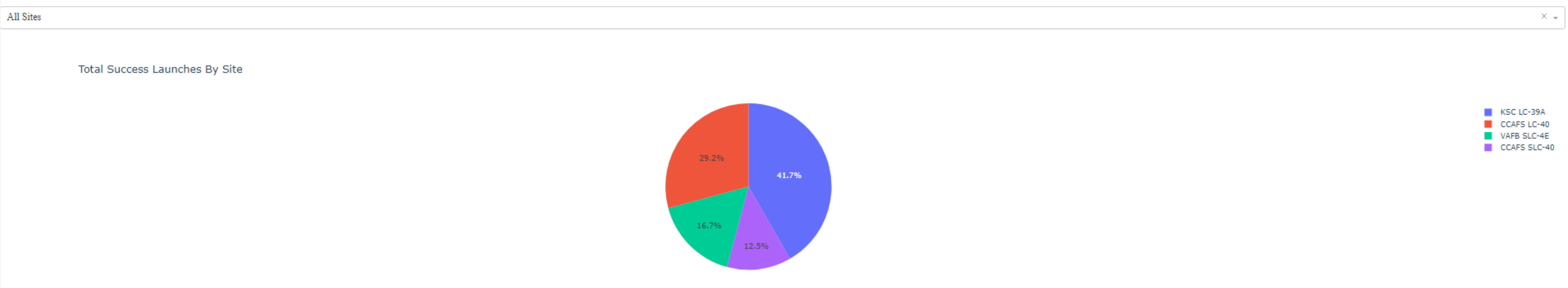
1) All launch sites are strategically located in close proximity to railroads, highways, and coastlines for various reasons. For instance, at the selected launch site "VAFB SLC-4E," the nearest railroad is approximately 1.26 km away, the closest highway is about 0.99 km away, and the nearest coastline is approximately 1.35 km away. The accessibility provided by these transportation routes is crucial for the smooth operation of the launch sites. Connected highways ensure that personnel working at these sites can easily commute to their workplace. Moreover, the presence of nearby railroads facilitates the transportation of equipment and other essential components required for launching rockets. Additionally, the necessity to be situated close to coastlines has been previously explained, providing opportunities for safer experiments and water landings.

2) All launch sites are intentionally located far from urban areas, ensuring that experiments take place away from densely populated regions, enhancing safety. At our selected site, "VAFB SLC-4E," the closest city, Lompoc, is situated approximately 14.03 km away, further underscoring the emphasis on conducting experiments away from urban centers.

Section 4

Build a Dashboard with Plotly Dash

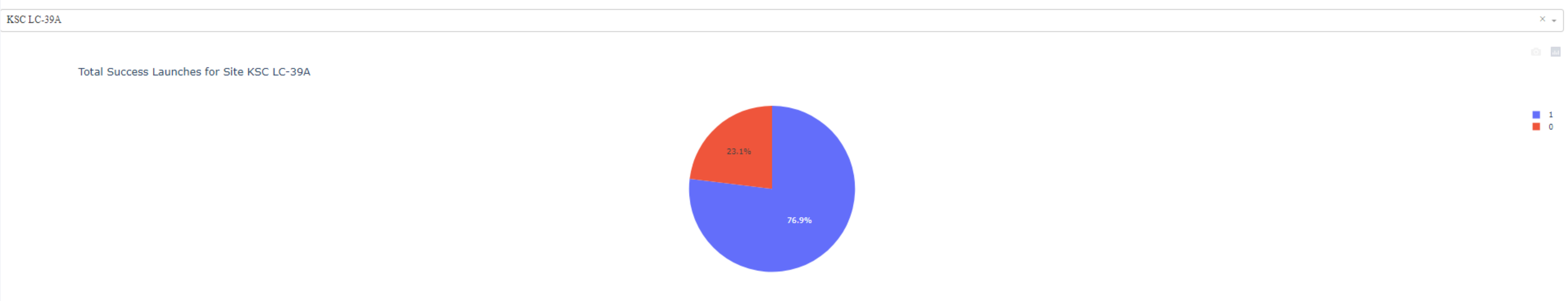
Total Success Launches by Site



2.Which site has the highest launch success rate?

2.KSC LC-394 with success rate equal to 76.9%.

Total Success Launches for Site KSC LC-39A




```
1.Which site has the largest successful launches?
```

```
1.VAFB SLC-4E : one successful launch with payload equal to 9600 kg.
```

Mission Outcome vs Payload mass for all sites

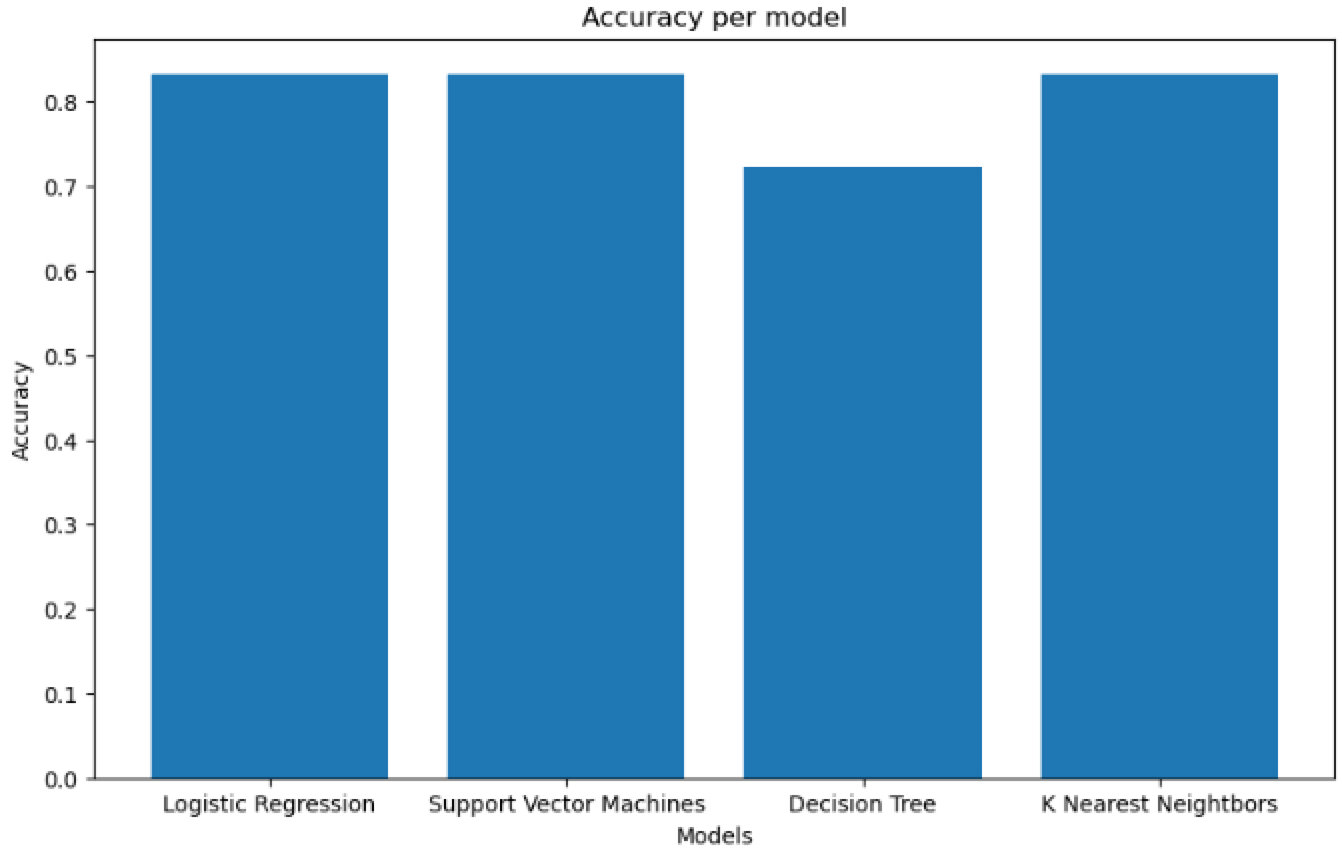

```
3.Which payload range(s) has the highest launch success rate?

4.Which payload range(s) has the lowest launch success rate?

5.Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?
```

```
3.2000 - 6000 kg.

4.6000 - 8000 kg.

5.FT
```

Section 5

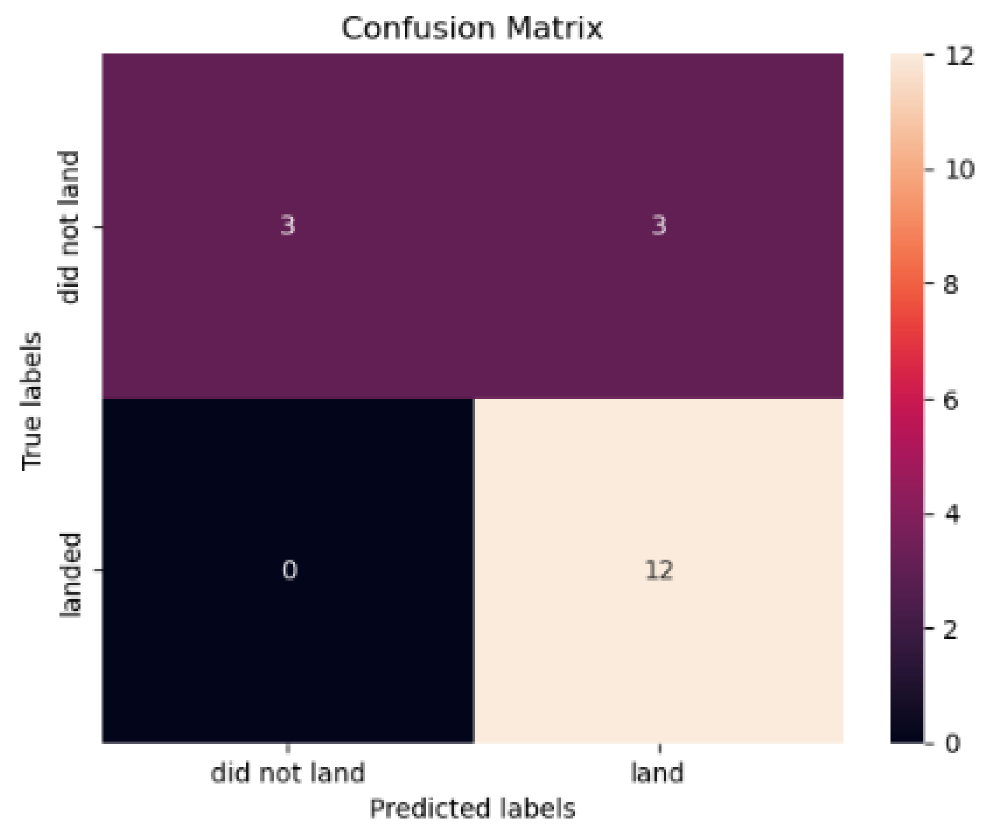
Predictive Analysis (Classification)

Classification Accuracy



Result : No method seems to be superior to the rest. Depending on the random_state of tree, the results differ.

Confusion Matrix



Explanation : This confusion matrix is the same for the models Logistic Regression, SVM and KNN.

Conclusions

- In the pursuit of achieving highly predictions regarding the landing of the first stage of Falcon 9, it is observed that Logistic Regression, SVM and KNN models yield similarly accurate predictions.
- There is a noticeable variability in the performance of the Decision Tree model used, which is influenced by the random seed used during its execution.
- Expanding the parameter grid for GridSearchCV with more comprehensive options could lead to discovering a different set of best parameters compared to the combination that was initially identified as the best.
- It is possible that with the expansion of the dataset in the future, the accuracy of the models may increase.

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

Thank you!