IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cristhian Ramirez
04/26/24

GitHub Repository: https://github.com/CristhianRam/IBM-DataScience-Final-Project.git

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Interactive map with Folium
  - Interactive Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  SpaceX, the most successful company of the commercial space, advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars while 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. So, our goal is, using Machine Learning Classification Models, determinate if the Falcon 9 first stage will land successfully,.

- Problems you want to find answers

  1. What kind of factors can affect the success of the Falcon 9 first stage landing.
  2. What is the best ML algorithm that can be used for the binary classification required in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using the SpaceX API and Web Scraping from Wikipedia.
- Perform data wrangling
  - Missing values handling.
  - Filtering and Formatting Data.
- 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 evaluating the perform of classification models.

Data Collection

In order to collect the required data for our SpaceX Falcon 9 launches data set, we perform two main techniques:

- Using SpaceX API:

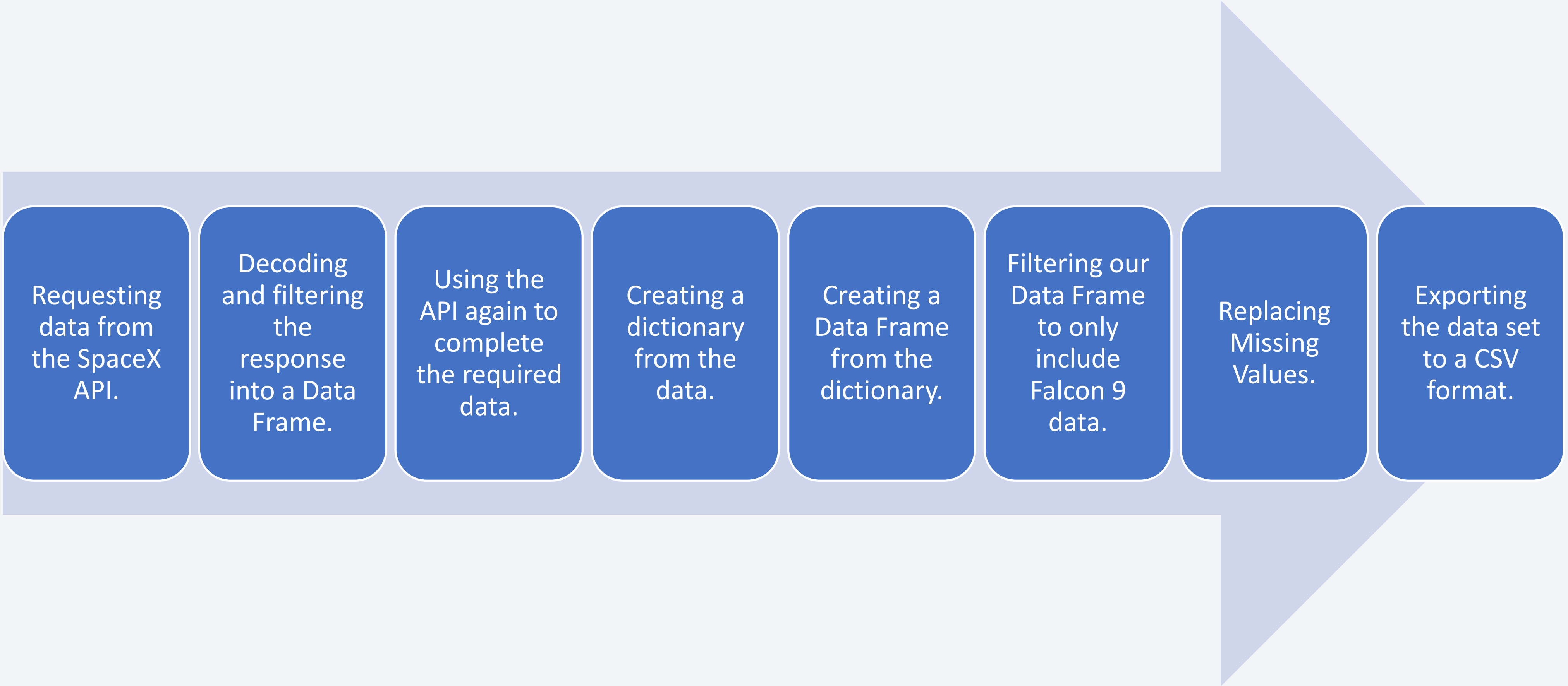
  The objective was making a request to the SpaceX API in order to get the needed data in a JSON format to then decode it into a Data Frame and export it in a CSV format.

- Using Web Scraping from Wikipedia:

  The objective was to extract the launch records as HTML table, then parse the table and

  convert it to a Pandas Data Frame to be able to also export it in a CSV format.

Data Collection – SpaceX API



GitHub Link: https://github.com/CristhianRam/IBM-DataScience-Final-Project/blob/101c9d53b4b23e271f28f742500cebbb19a4fd7e/spacex-data-collection-api.ipynb

Data Collection - Scraping

1. Requesting Falcon 9 launch data from Wikipedia.
2. Creating Beautiful Soup object from the response.
3. Extracting the column names from the HTML table header.
4. Collecting data by parsing the launch HTML tables.
5. Constructing a dictionary from the collected data.
6. Creating a Data Frame from the dictionary
7. Exporting the data set to a CSV format.

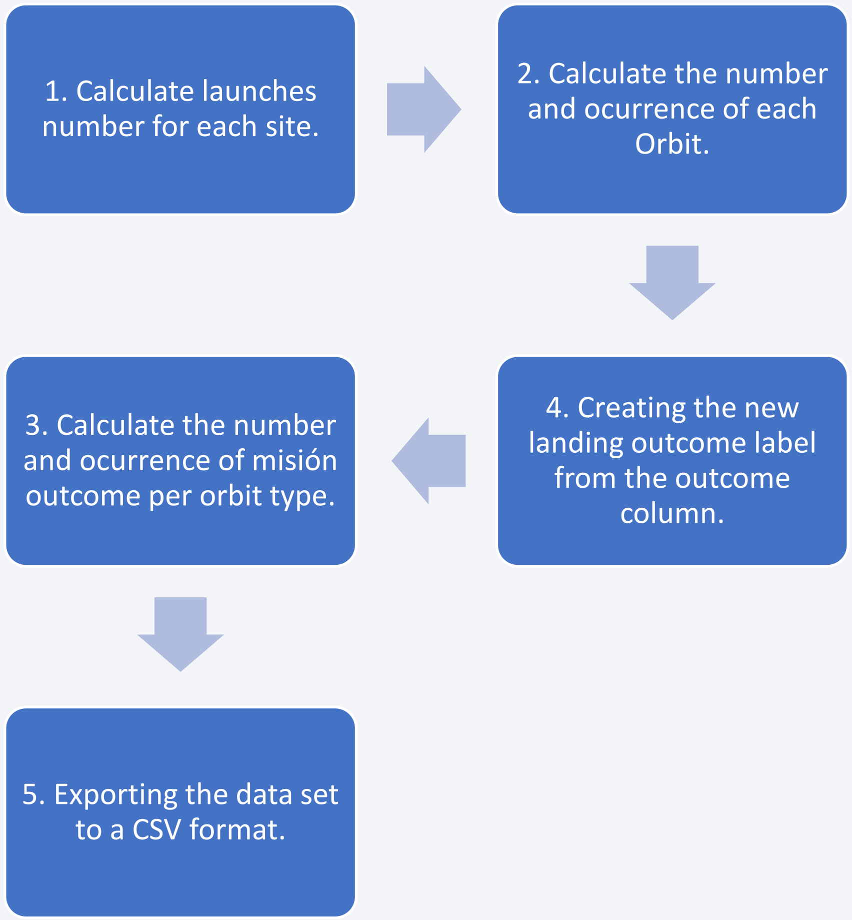
GitHub Link: https://github.com/CristhianRam/IBM-DataScience-Final-Project/blob/101c9d53b4b23e271f28f742500cebbb19a4fd7e/jupyter-labs-webscraping.ipynb

Data Wrangling

At this point in the data set, we have 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.

So, we needed to convert those categorical outcomes into numerical ones, in this case we assigned "1" to all the cases where the booster successfully landed, and "0" to the rest.



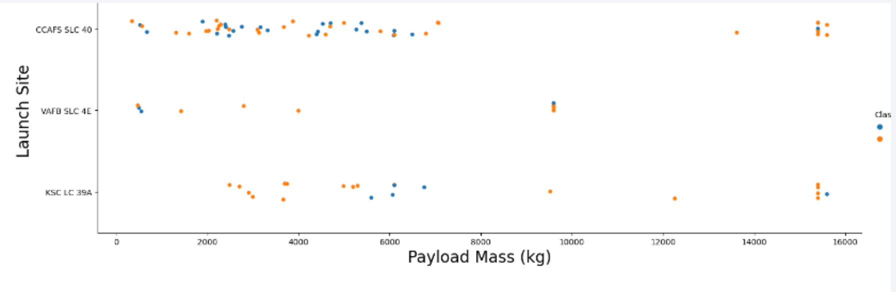
GitHub Link: https://github.com/CristhianRam/IBM-DataScience-Final-Project/blob/f5e099908fb7a3b7bdf1418e18070beab8f9e176/spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Scatter Point Chart

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

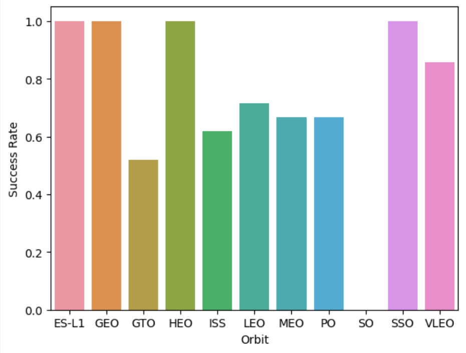
This kind of plot was used to show relationship between variables.



Bar Chart

- Success rate of each orbit type

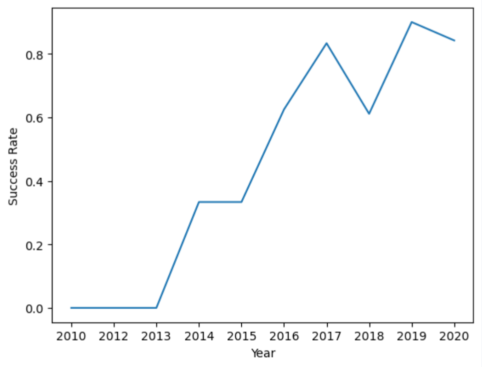
This kind of plot was used to visualize the relationship between numeric and categoric variables.



Line Chart

- Success rate per year

This kind of plot was used to show trends in data over time.



GitHub Link: https://github.com/CristhianRam/IBM-DataScience-Final-Project/blob/6e51d3b0a9bcdd2288e6713dd3c2c82dadccd36b/eda-dataviz.ipynb

EDA with SQL

- **We performed the following SQL queries:**
  - 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 succesful landing outcome in ground pad was acheived.
  - 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 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 Link: https://github.com/CristhianRam/IBM-DataScience-Final-Project/blob/76030b88db73e2bc6fca728ff7e85fb9fed2e4c3/eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers for all launch sites
  - Red circle at NASA Johnson Space Center's coordinate with Popup label and text label showing its name.
  - Red circles at each launch site coordinates with Popup label and text label showing launch site name.
- Colored Markers for launch outcomes
  - Colored markers to show success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Markers for distance
  - Lines to show distance between a selected launch site to key locations such as railway, highway, coastline and the closest city.

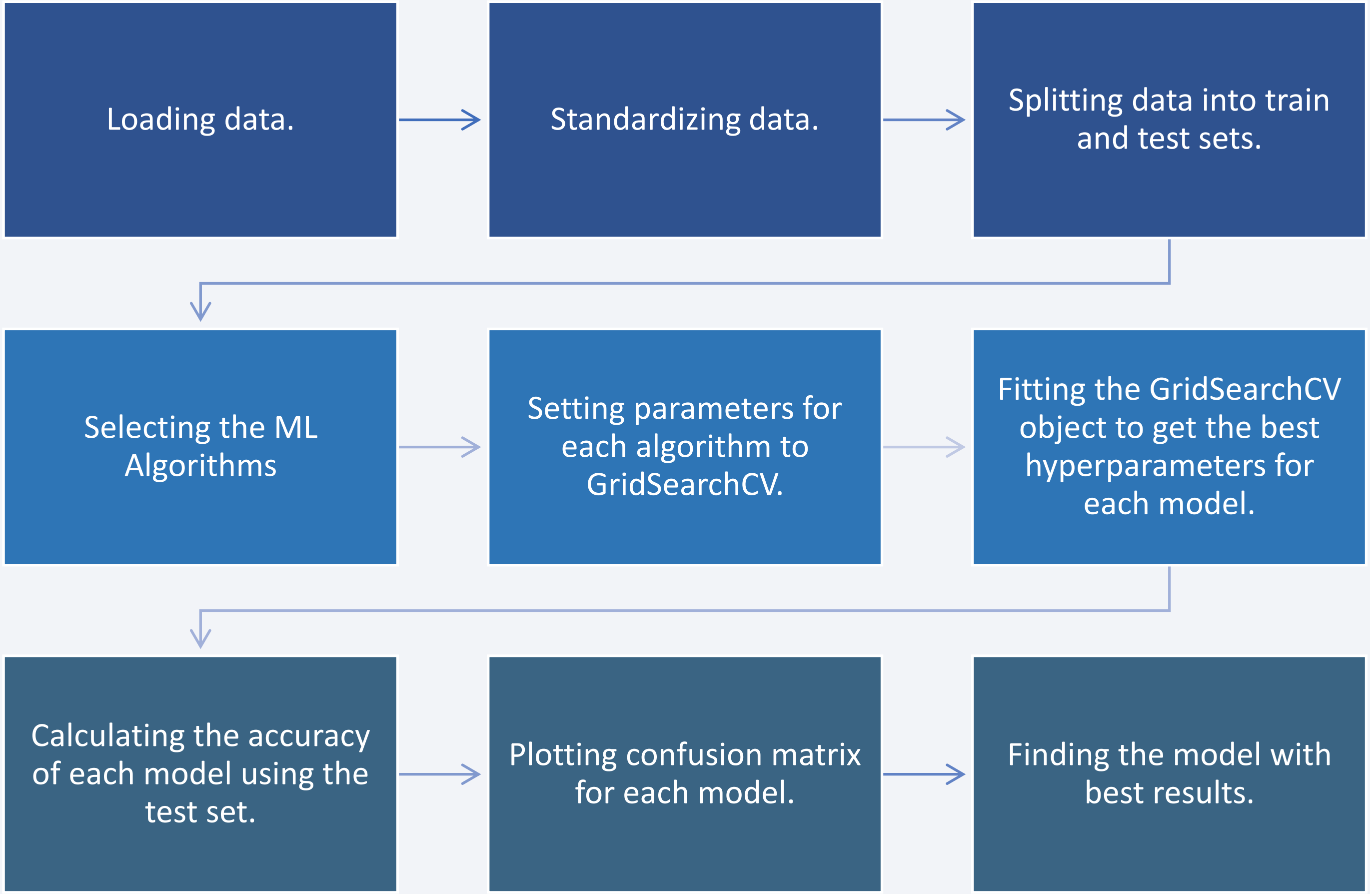
GitHub Link: https://github.com/CristhianRam/IBM-DataScience-Final-Project/blob/6e51d3b0a9bcdd2288e6713dd3c2c82dadccd36b/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Dropdown list
  - Used to select a launch site.
- Pie Chart
  - Used to show the total successful launches count for all sites, or If a specific launch site was selected, show the Success vs. Failed counts for that site.
- Range Slider
  - Used to select an interval of payload mass.
- Scatter Chart
  - Used to show the correlation between payload and launch success.

GitHub Link: https://github.com/CristhianRam/IBM-DataScience-Final-Project/blob/6e51d3b0a9bcdd2288e6713dd3c2c82dadccd36b/spacex_dash_app.py

Predictive Analysis (Classification)



GitHub Link: https://github.com/CristhianRam/IBM-DataScience-Final-Project/blob/6e51d3b0a9bcdd2288e6713dd3c2c82dadccd36b/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

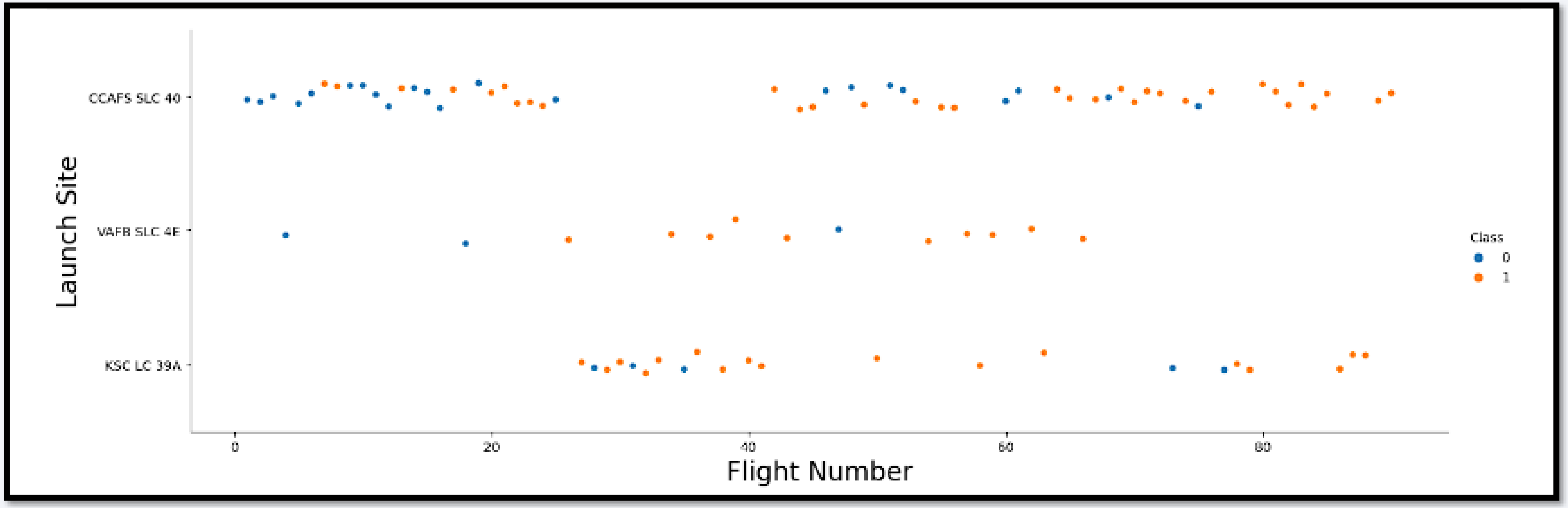
Results

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

Section 2

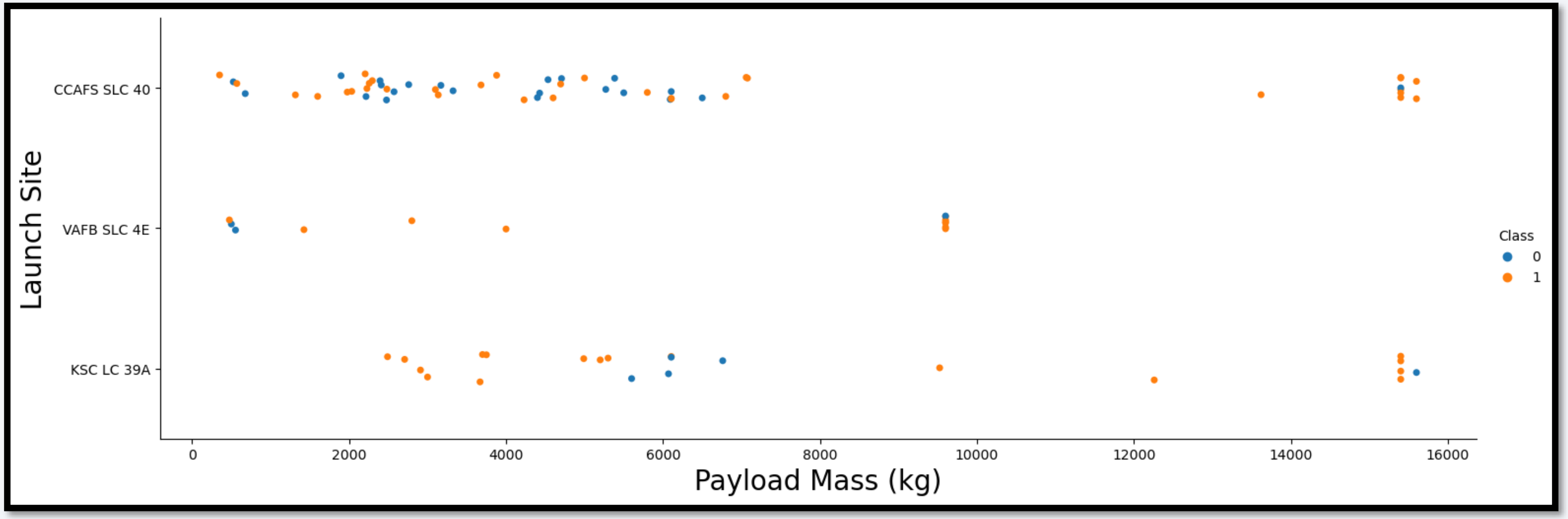
Insights drawn from EDA

Flight Number vs. Launch Site



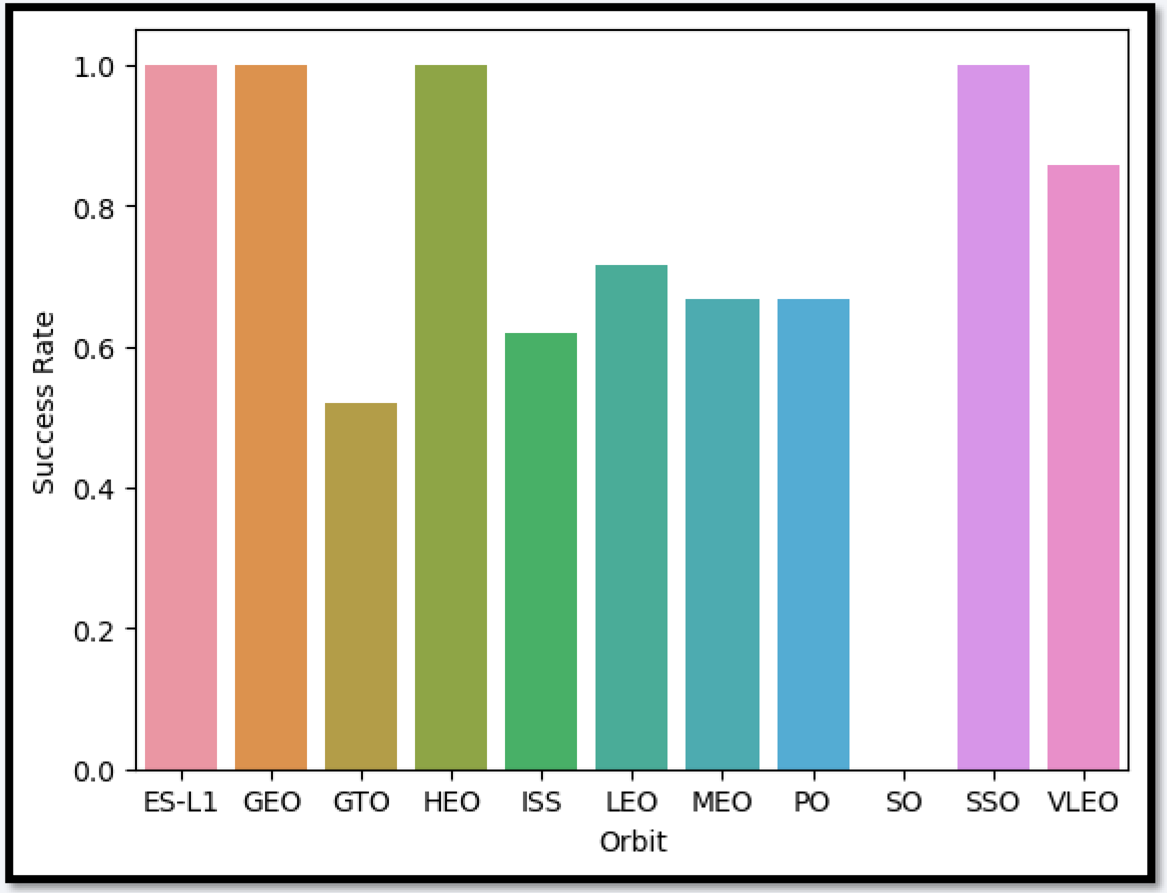
- We can observe that in most of the cases, for each Launch Site, as the flight number increases, the first stage is more likely to land successfully.
- And we can see that the sites KSC-LC 39A and VAFB-SLC 4E have the highest success ratio.

Payload vs. Launch Site



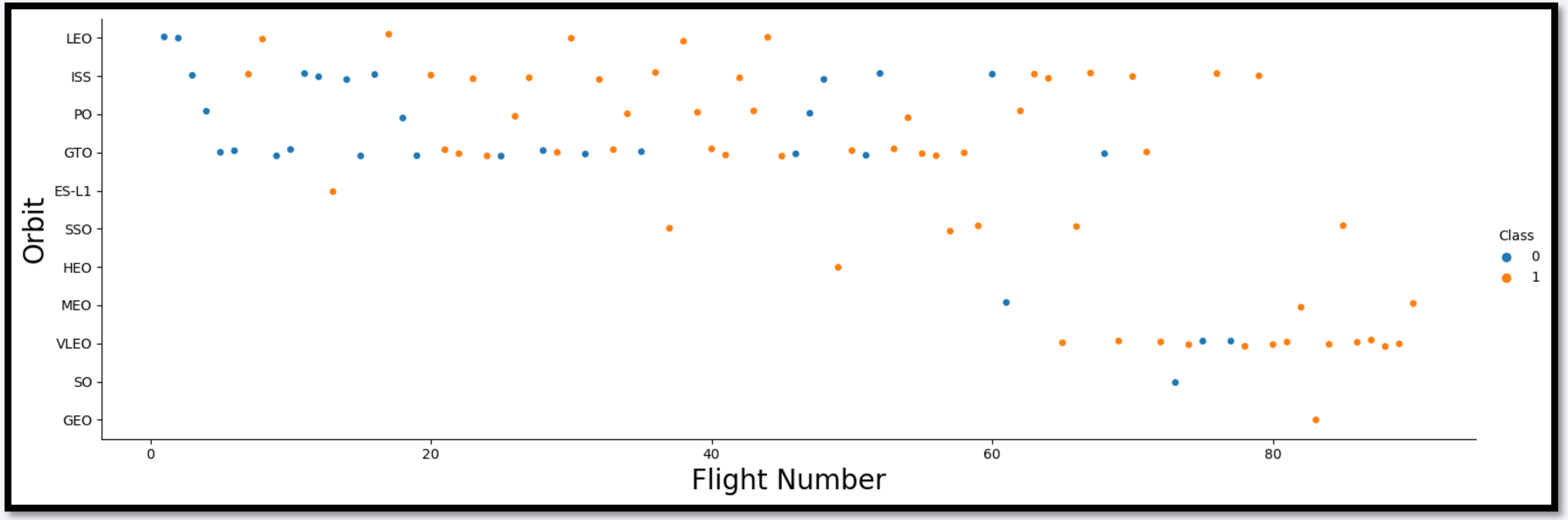
- We can find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type



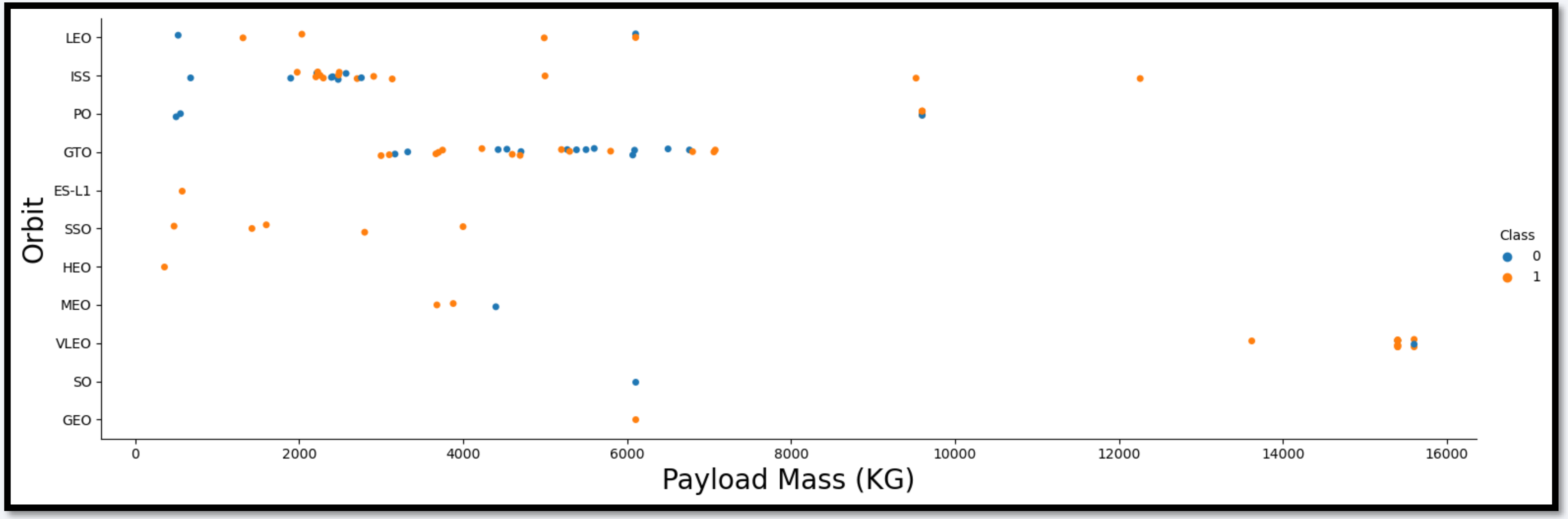
- We can see that the Orbits with the best success rate (100%) are:
  - ES-L1, GEO, HEO and SSO
- Also, the Orbit with the lowest success rate (0%) was SO.

Flight Number vs. Orbit Type



- This scatter plot shows that in most of the cases, a larger the flight number on each orbits, tends to mean more chances of succes (especially LEO orbit).

Payload vs. Orbit Type

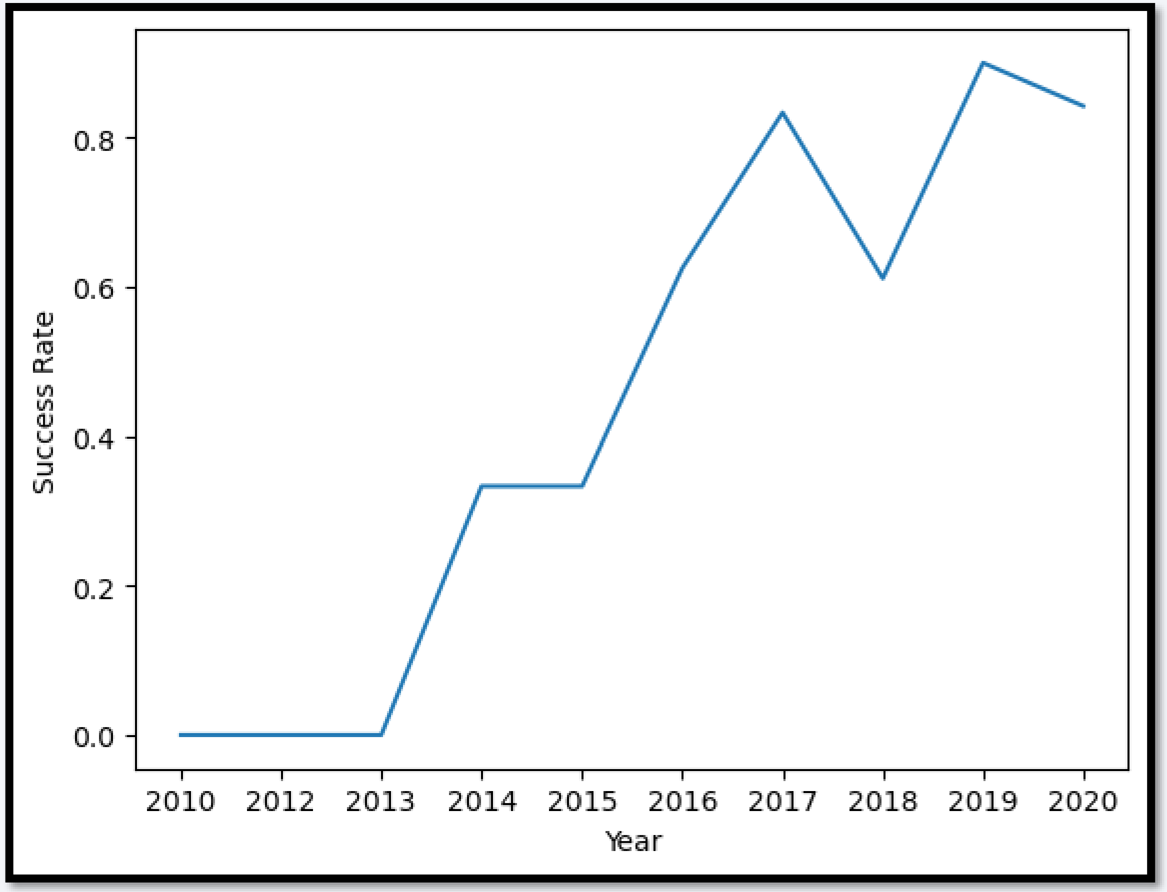


- We can see that heavy payloads have a negative influence on MEO and VLEO orbits and positive on LEO, PO and ISS orbits.

Launch Success Yearly Trend

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

- In general, the success rate was increasing in most of the time.



All Launch Site Names

In [10]:

```
%sql SELECT DISTINCT("Launch_Site") FROM SPACEXTABLE
```

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

Out[10]:

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

- The DISTINCT query in SQL allows us to remove duplicates in the column “Launch_Site”, displaying the unique Launch Sites.

Launch Site Names Begin with 'CCA'

In [12]:
```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

Out[12]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	M
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
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	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

- We make use of the WHERE and LIKE commands to filter the entries where launch sites begin with the string 'CCA'.

  And used the LIMIT 5 command to only display 5 entries.

Total Payload Mass

```
In [16]: %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE Customer=='NASA (CRS)'

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

Out[16]:

SUM(PAYLOAD_MASS__KG_)
45596

- We use this query to display the only the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
In [17]: %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE Booster_Version=='F9 v1.1'

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

Out[17]:

AVG(PAYLOAD_MASS__KG_)
2928.4

- We use this query to return the average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

In [21]:
```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome=='Success (ground pad)'
```

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

Out[21]:

MIN(Date)
2015-12-22

- We use this query with the MIN function to return the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

- We use this query to return the list with the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

In [45]:
```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") FROM SPACEXTABLE GROUP BY "Mission_Outcome"
```

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

Out[45]:

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

- We use this query to list the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
In [47]: %sql SELECT DISTINCT(Booster_Version) FROM SPACEXTABLE \
         WHERE PAYLOAD_MASS__KG_==(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

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

Out[47]:

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

- We used this query to list the names of the booster_versions which have carried the maximum payload mass, using a subquery to calculate the maximum payload mass.

2015 Launch Records

In [55]:

```
%sql 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[55]:

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

- We used this query to list the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE \
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.
```

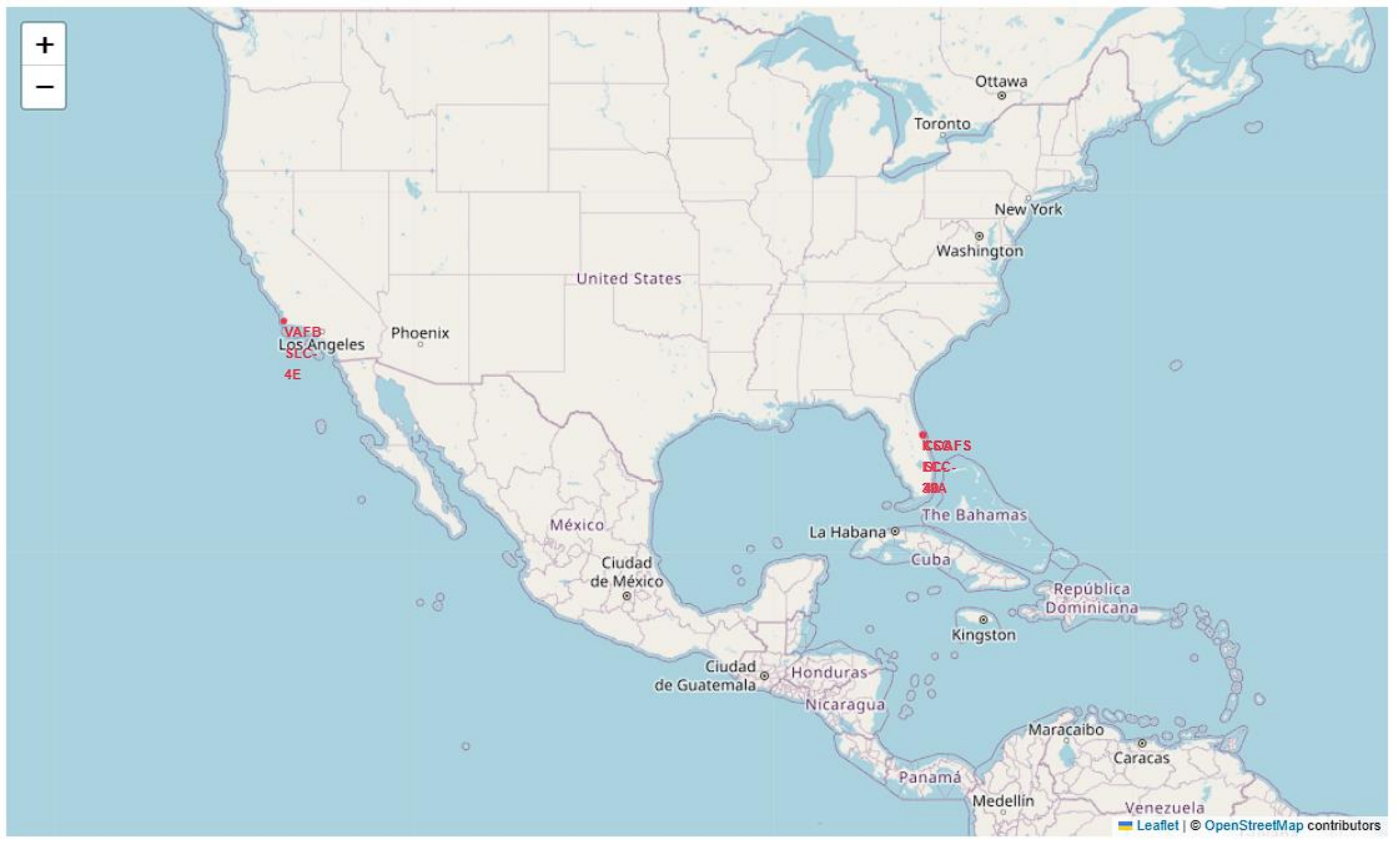
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- We use this query to 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.

Section 3

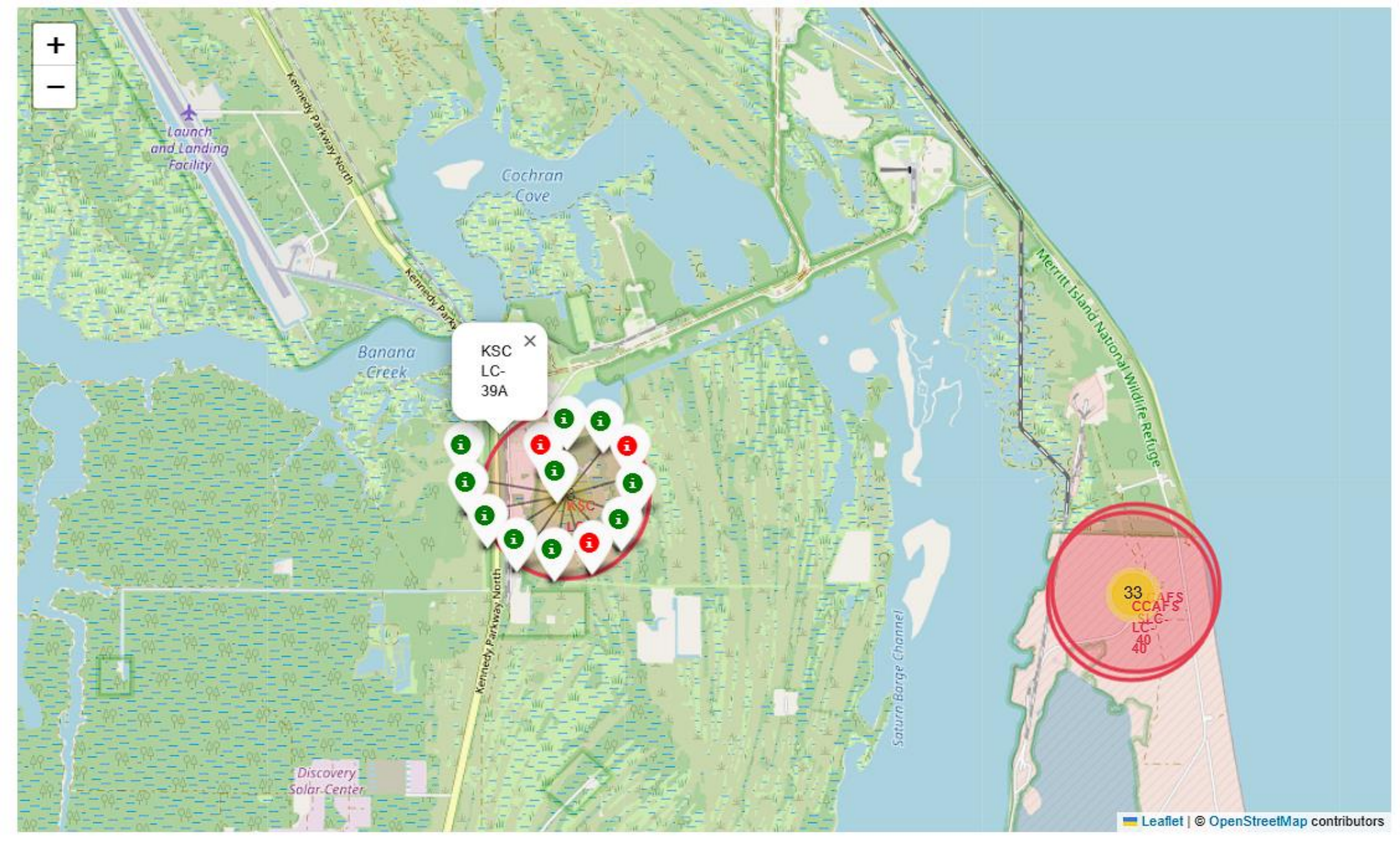
Launch Sites Proximities Analysis

Location of All Launch Sites



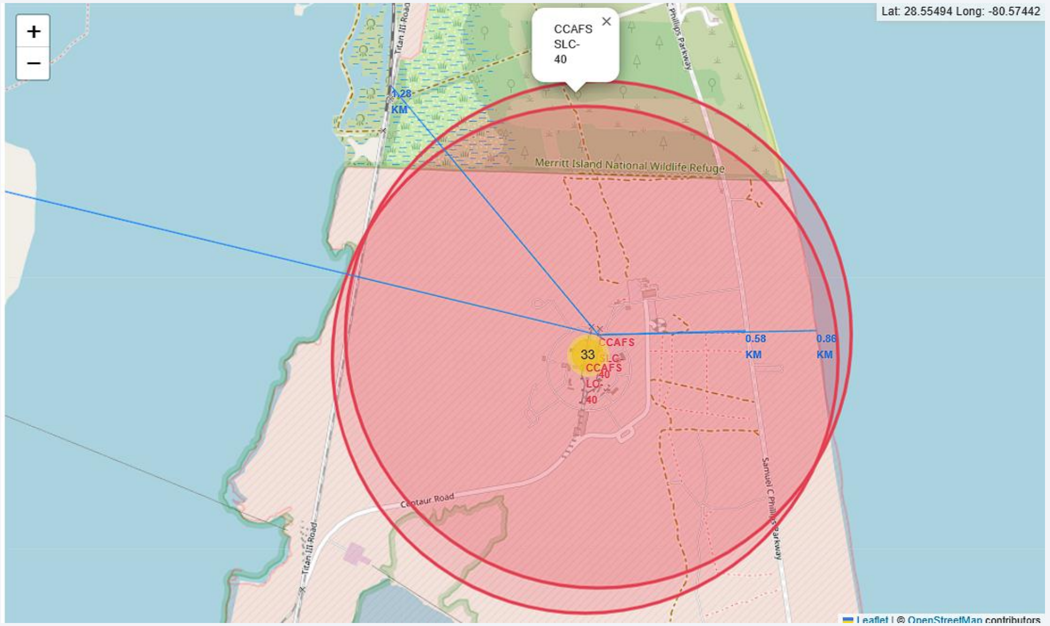
- We can see that all the launch sites are located on the coast of the United States.

Colored markers in the map



- With the colored markers we can easily identify which Launch Sites have a high success rate, such as the KSC LC-39A launch site.
- Green marker = successful launch.
- Red marker = failed launch.

Distance from Launch Site to Key Proximities



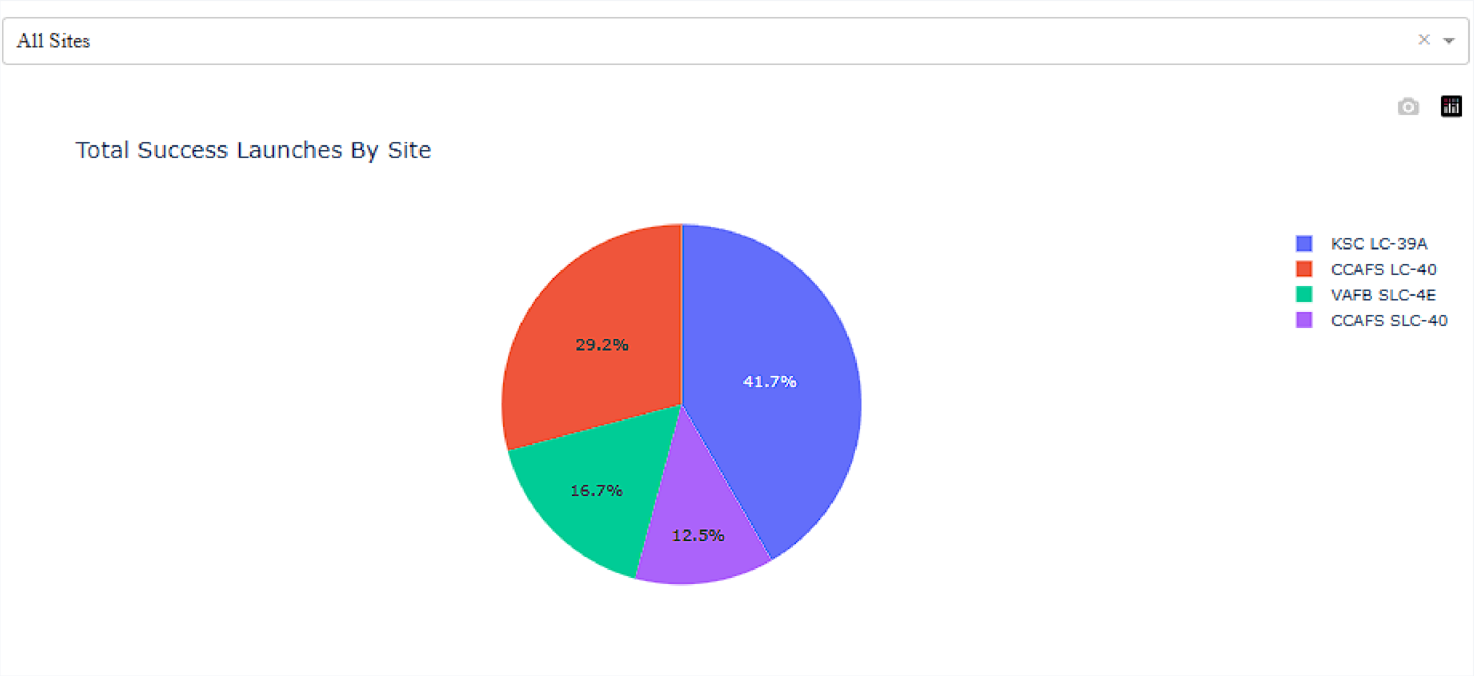
The CCAFS SLC-40 launch site is close to the Highway, the coastline and the railroad.

And it is relatively close to Titusville City.

Section 4

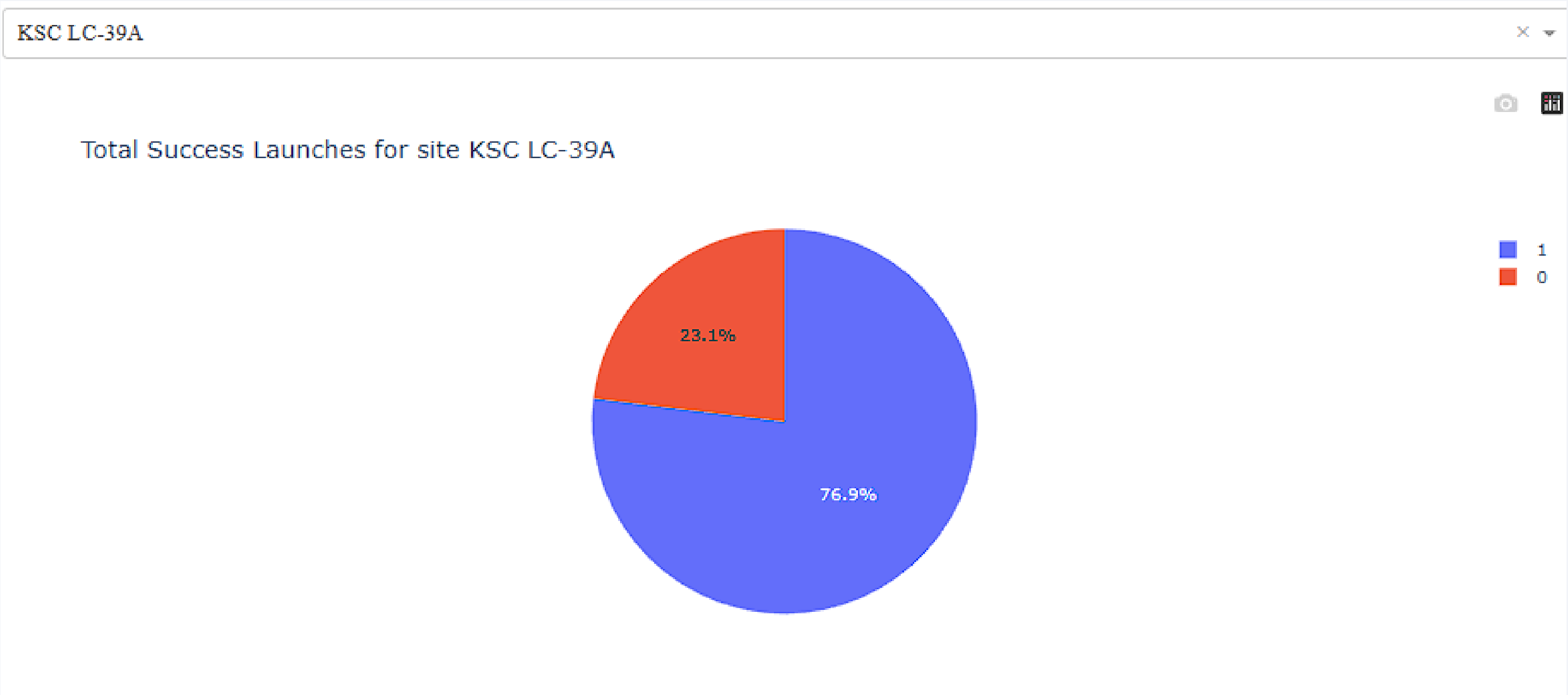
Build a Dashboard with Plotly Dash

Launch Success Count for all Launch Sites



- We can find that the site KSC LC-39A has the highest count of Successful Launches, and VAFB SLC-4E has the lowest.

Launch Site with the Highest Success Ratio



- The Launch Site KSC LC-39A has the highest success ratio (76.9%), with 10 successful launches and only 3 failed launches.

Payload vs Launch Outcome for all sites.

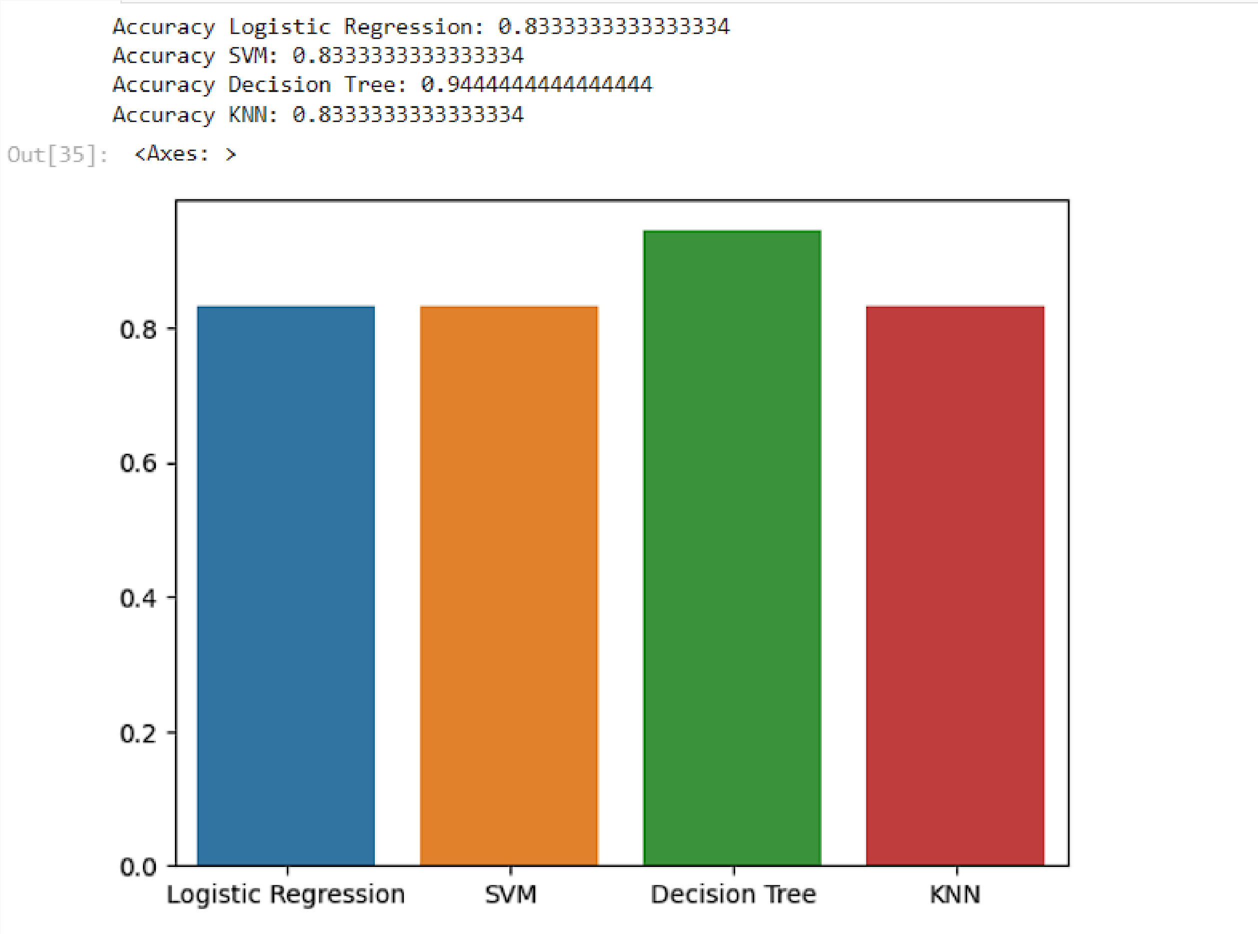
- We can see that the success rate is higher for low weighted payload mass (<5000 KG) compared to the success rate for heavy weighted payload (>5000 KG).

Section 5

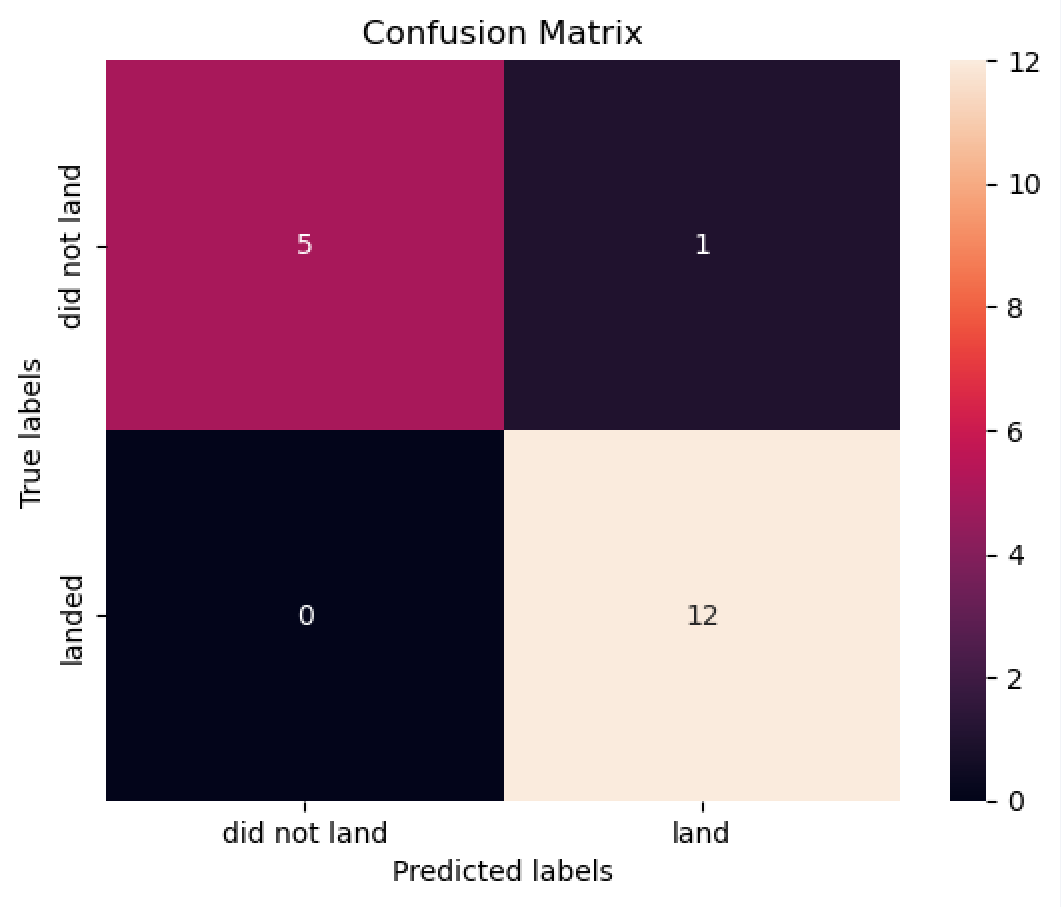
Predictive Analysis (Classification)

Classification Accuracy

- As wee can see in the image, 3 of the 4 methods gave us the same result for accuracy in the test set (83.3%).
- The best method was the Decision Tree method, with 94.4% of accuracy in the test set.



Confusion Matrix



Confusion Matrix of the Decision Tree model.

- With the confusion matrix of the Decision Tree Classifier, we can see that the only problem of this method is a single False Positive.

Conclusions

- Decision Tree Model is the best algorithm for this data set with an accuracy of 94.4%.
- The low weighted payloads (<5000 KG) have a better chance to be successful than the heavy weighted payloads (>5000 KG).
- The success rate is most of the time increasing over the years.
- Most of the Launch Sites are close to the coastline.
- There are 4 orbits with 100% success rate: ES-L1, GEO, HEO and SSO
- The Launch Site KSC LC-39A has the highest success ratio (76.9%), and highest count of Successful Launches with 10 successful launches.

Thank you!