Winning Space Race with Data Science

Tomáz Giansante
D September 2022

Outline

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

Executive Summary

- Build a Dataset with SpaceX API
- Machine Learning Models with Sklearn
- Landing Outcome correctly predicted with good accuracy

Introduction

- SpaceX is one of the biggest private rocket manufactures on the world.
- To create a competitive bussiness against SpaceX, unsderstanding the weakest and strongest points of it’s service is essencial.
- How can we predict wich rockets succede their landing based on their caracteristics ?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We collect the data with the SpaceX API
- Perform data wrangling
  - Using Pandas DataFrame functions to do Features Engineering
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build Machine Learning Models using Sklearn Package

Data Collection

Throught the use of the Spacex API (https://api.spacexdata.com/v3) we built a diverse dataset with caracteristics such as:

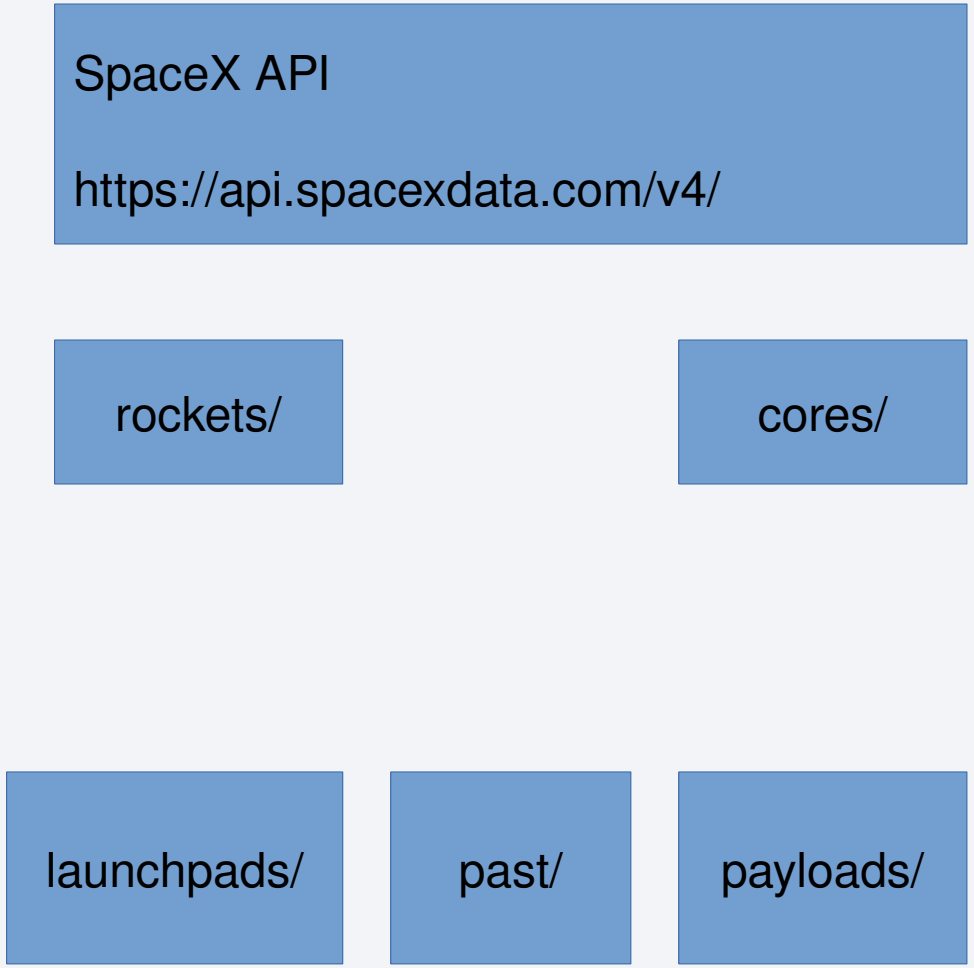
Flight Number, Date, Time (UTC), Rocket Booster Version, Launch Site e etc.

After the REST calls, the request is turned into a json file and then using the Pandas package we can extract the data into a Pandas DataFrame with the function pd.json_normalize()

Data Collection – SpaceX API

- The following flowchart has the Url and the keys used to scrap the SpaceX data.

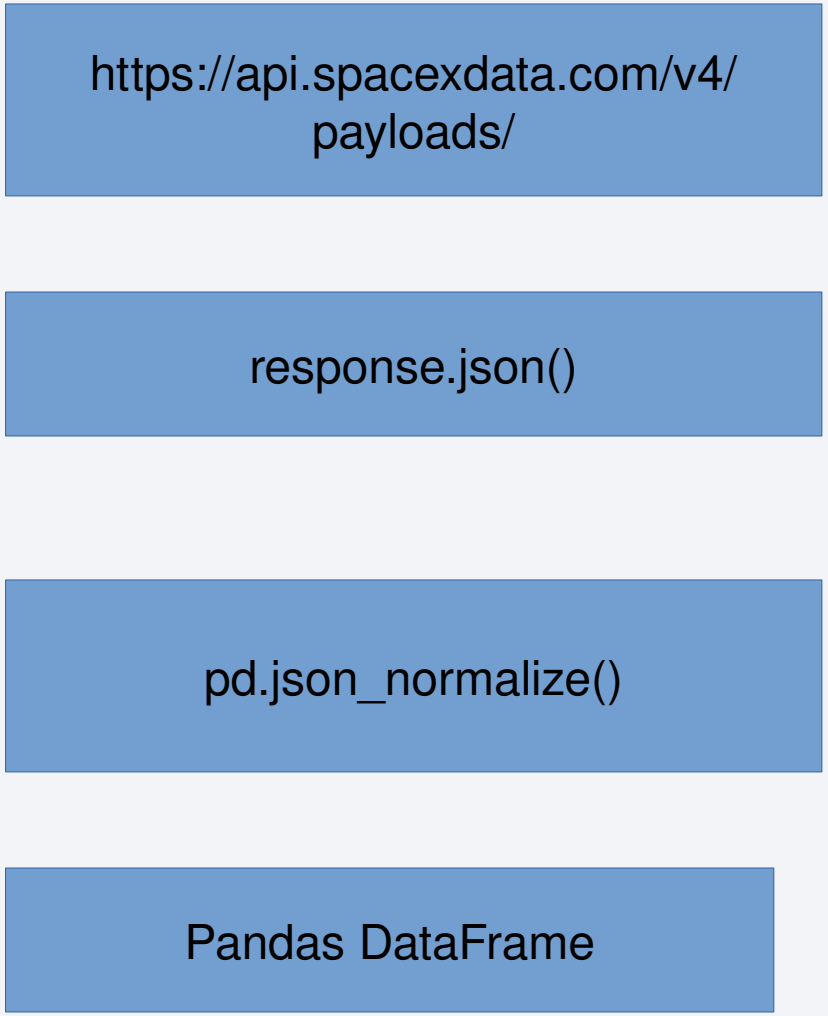
https://gitlab.com/mazeeqe/ibmdatascience_public/-/blob/main/Applied%20Data%20Science%20Capstone/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- After getting the response using the request package, we will extract it's features using the functions described in the flowchart.

https://gitlab.com/mazeeqe/ibmdatascience_public/-/blob/main/Applied%20Data%20Science%20Capstone/jupyter-labs-webscraping.ipynb



Data Wrangling

In the dataset built some values of the Payload Mass and LandingPad as null values.
For Payload mass we substitute the null values with the mean of all the others launchs with the use of the use of .mean() and .replace() from the numpy package.
For LandingPad null values, they actually represent when landing pads were not used.
A new column Class will be created that represents if the land was successful (1) or a failure (0).

https://gitlab.com/mazeeqe/ibmdatascience_public/-/blob/main/Applied%20Data%20Science%20Capstone/labs-jupyter-spacex-Data_wrangling.ipynb

EDA with Data Visualization

To get a better understanding of our data, plotting some graphs will help give in new insights, so that when we build our Machine Learning Models, the best features used as inputs will be determined.

Important features are: Payload Mass, Launch Site, Orbit type, Flight Number.

Such features will be plotted against the Success Rate using the Seaborn package.

https://gitlab.com/mazeeqe/ibmdatascience_public/-/blob/main/Applied%20Data%20Science%20Capstone/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Using SQL we can explore our dataset, discover basic informations about it:

- The Launch Sites used;
- Average Payload Mass;
- The number of successful landings;
- The booster version of each rockets;
- The most common customer.

https://gitlab.com/mazeeqe/ibmdatascience_public/-/blob/main/Applied%20Data%20Science%20Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

The Folium package can help a person understand better the data through the construction of interactive maps.

Every launch has a latitude and longitude coordinate, with thoose coordinates it’s possible to plot into the map the especific location of each launch.

The MarkerCluster() object makes it possible to cluster launchs that where close to each other, if the user decides to zoom in, the cluster will divide into smaller cluster to a easier visualization.

https://gitlab.com/mazeeqe/ibmdatascience_public/-/blob/main/Applied%20Data%20Science%20Capstone/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

The Plotly Dash package can build interactive graphs to facilitade the understanding of the data being analysed.

Building a Pie Chart to visualize wich of the launch sites has the best success rate or choose a indivudual site to see it’s induvidual success rate.

To Build a Scatter Plot where the X axis is the Payload Mass (Kg) and the Y axis is the landing success variable and all data points color coded based on their Booster Version Category. This will help understantd wich Payload Mass range and Booster Version Category is the most successful.

Predictive Analysis (Classification)

The Sklearn package was used to deploy numerous classification models with the goal of finding the best predictor.

Logistic Regression, Support Vector Machine, Decision Trees, K Nearest Neighboors.

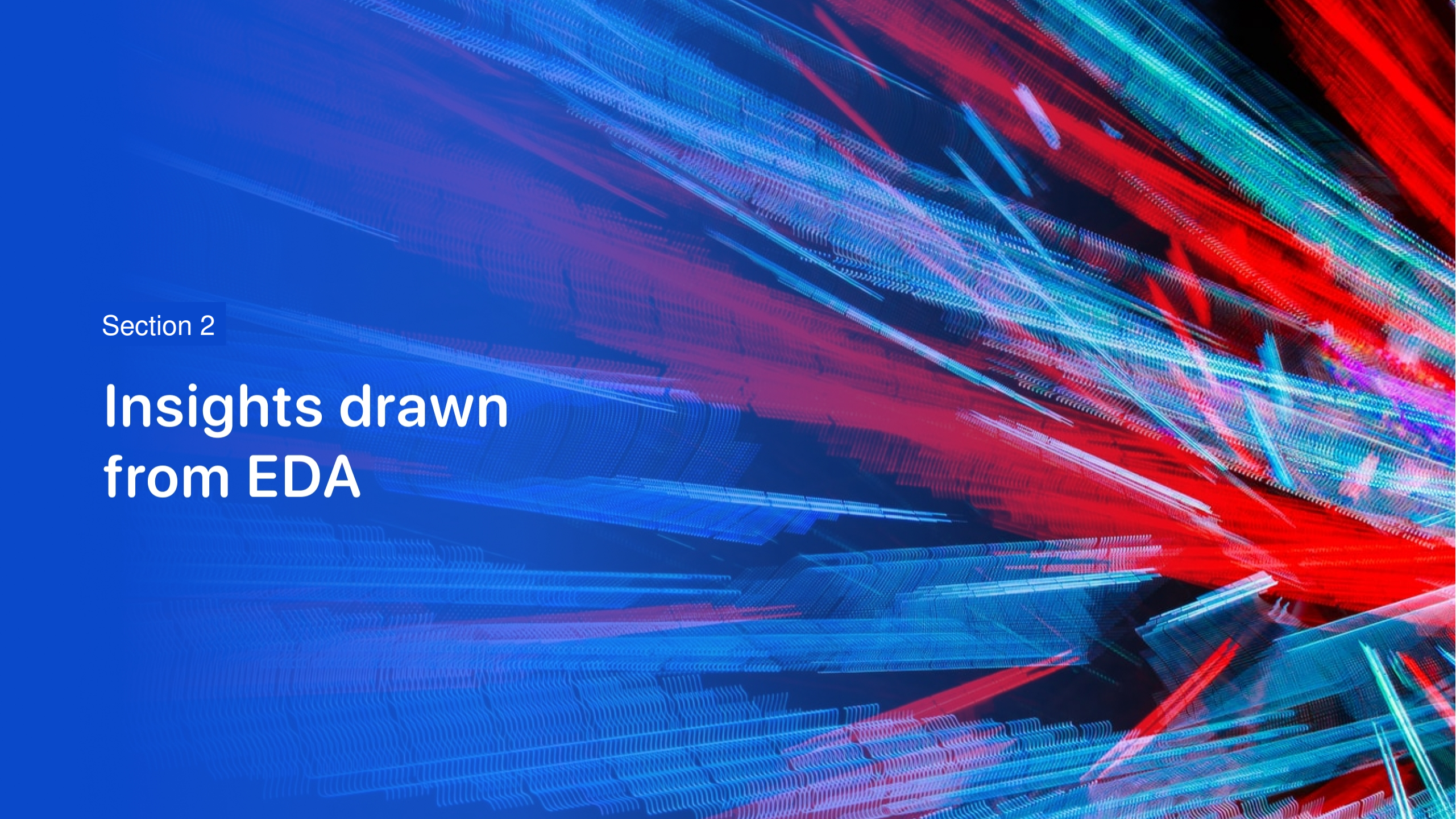
The object GridSearchCV can perform the learning process for each model with a multitude of parameters to find out the best performing parameters and score using the respective functions:

GridSearchCV.best_params_ , GridSearchCV.best_score_

https://gitlab.com/mazeeqe/ibmdatascience_public/-/blob/main/Applied%20Data%20Science%20Capstone/SpaceX_Machine_Learning_Prediction_Part_5.ipynb

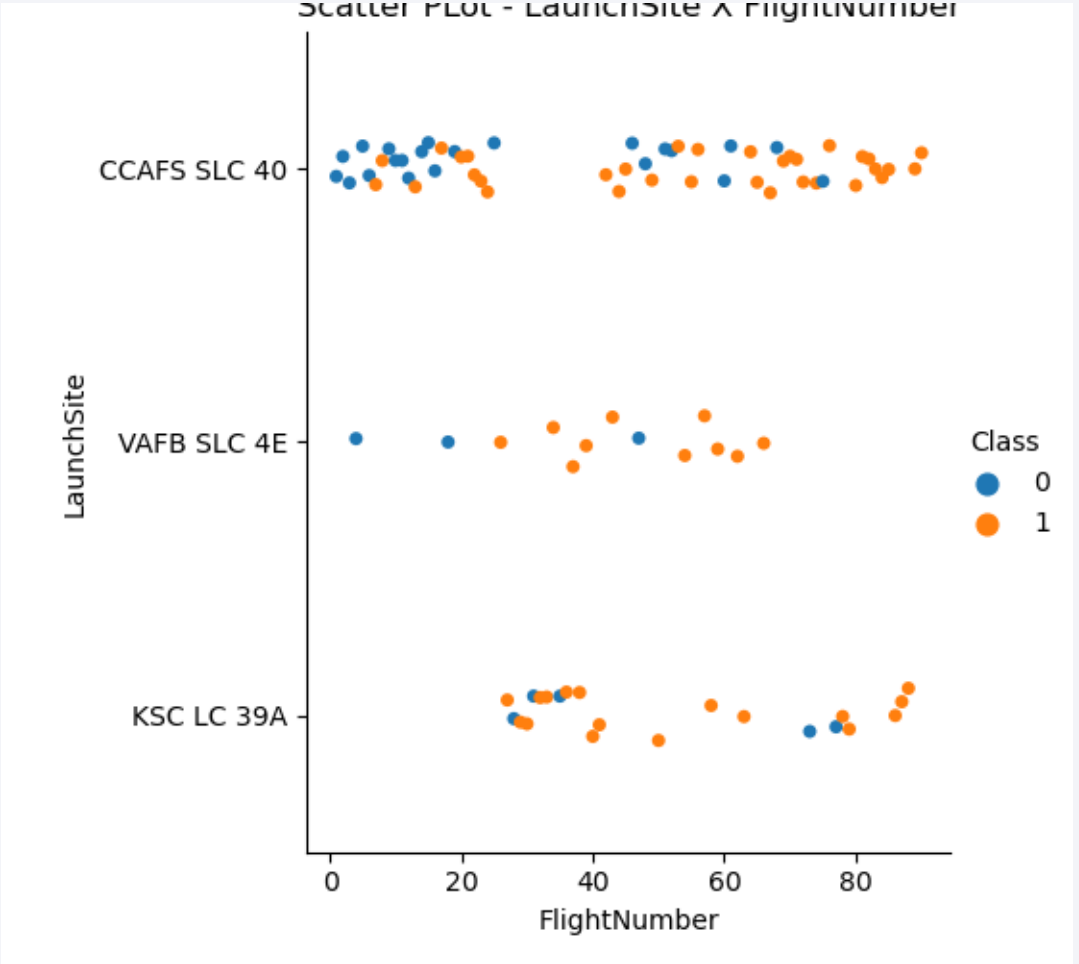
Results

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


Section 2
Insights drawn from EDA

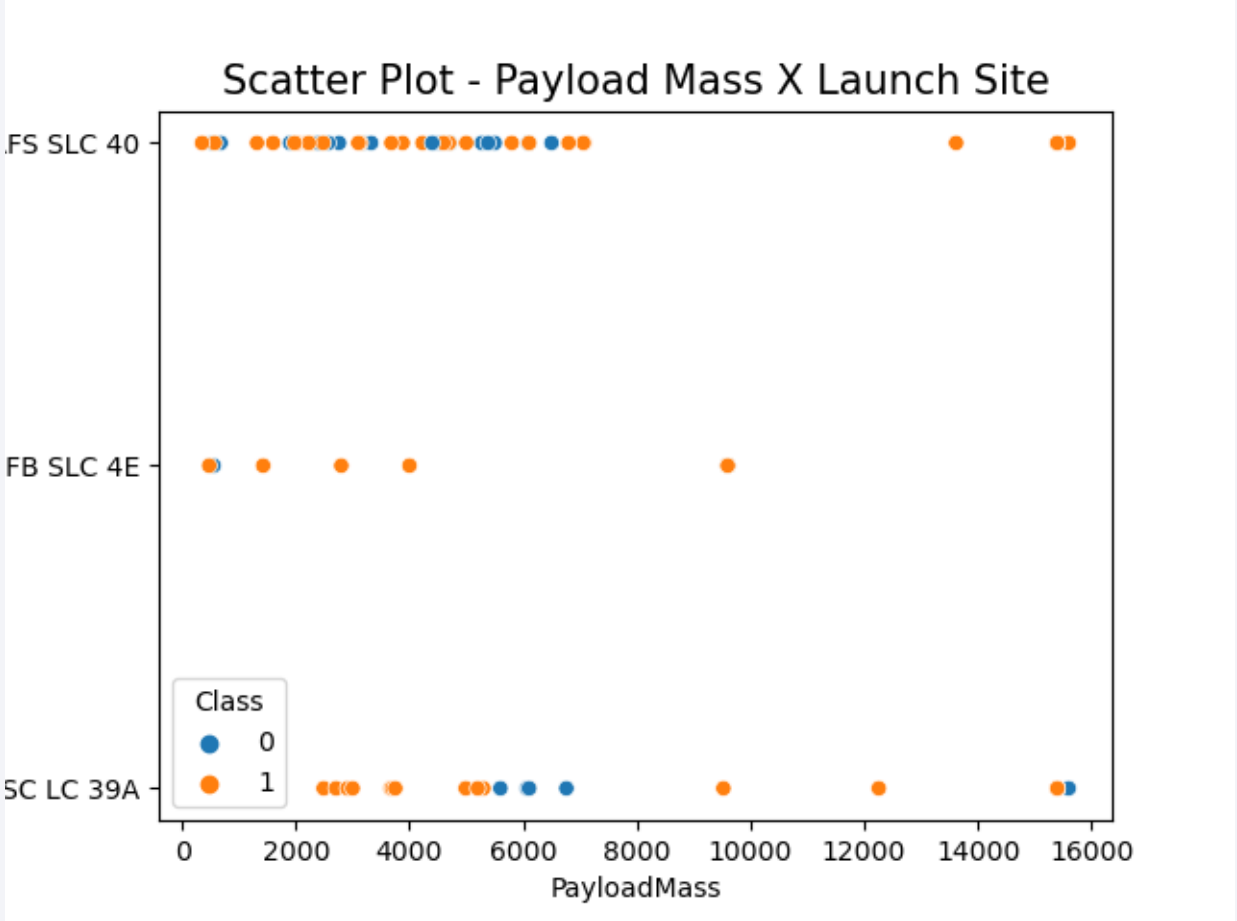
Flight Number vs. Launch Site

- As number of flights increases, the success rate also increases (orange dots).
- Launch Site KSC LC 39A has the highest success rate.



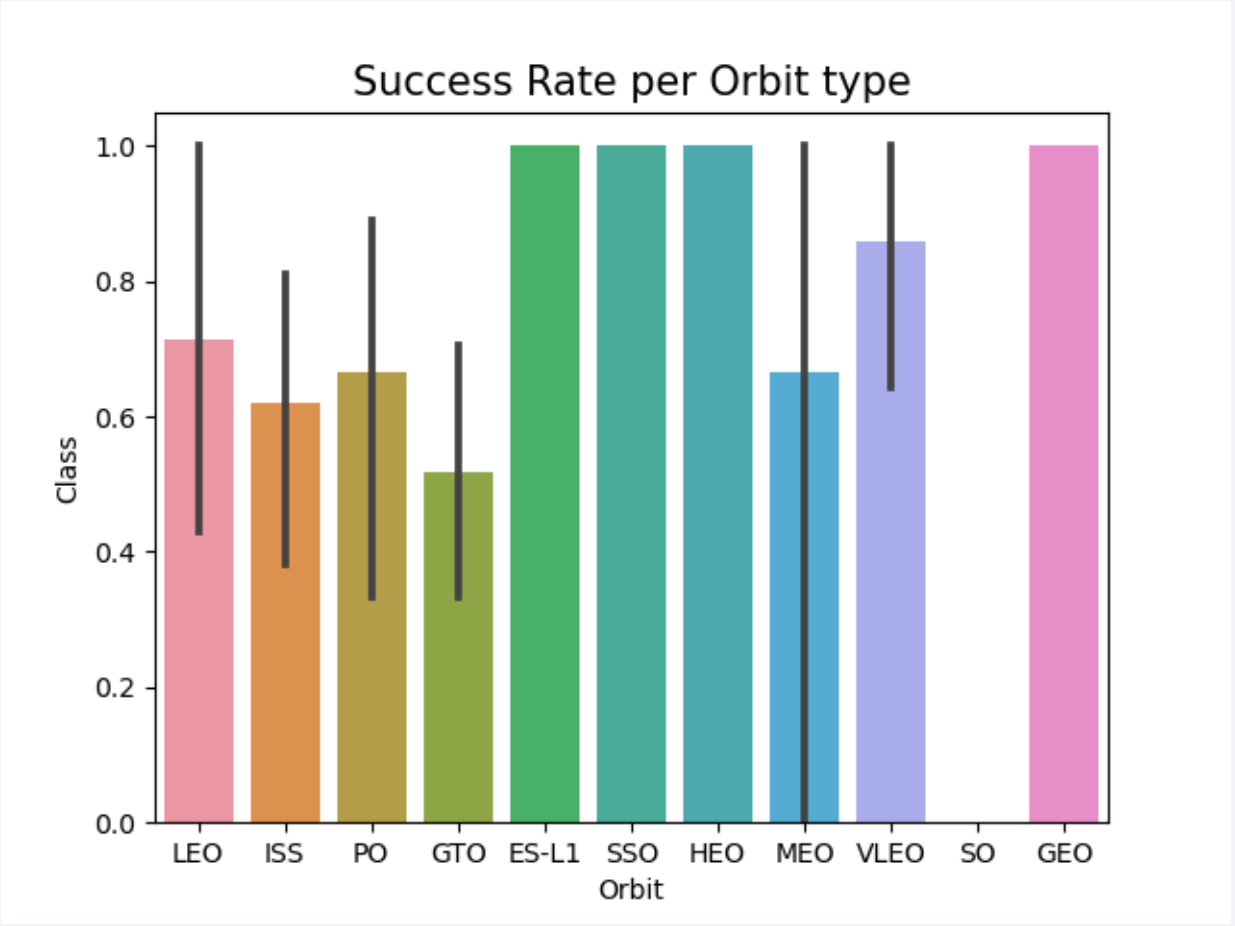
Payload vs. Launch Site

- Success rate increases with payload mass.
- There are no rockets launched for heavypayload mass (greater than 10000 Kg) on launch site VAFB-SLC.



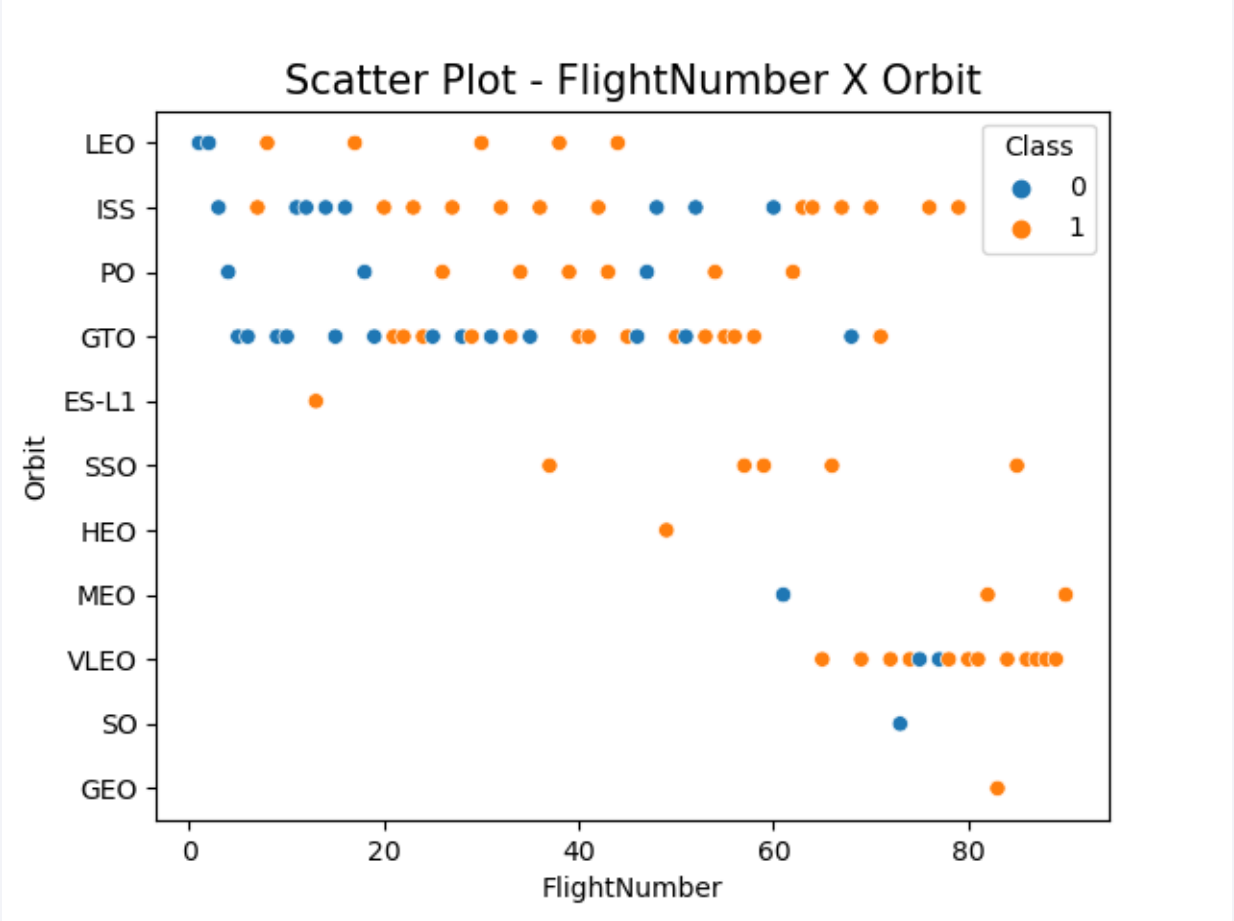
Success Rate vs. Orbit Type

- ES-L1, SSO, HEO, GEO has the highest rates, however they don't possess error bars due the lack of data.
- VLEO has the highest rate given the boundaries of the error bar.



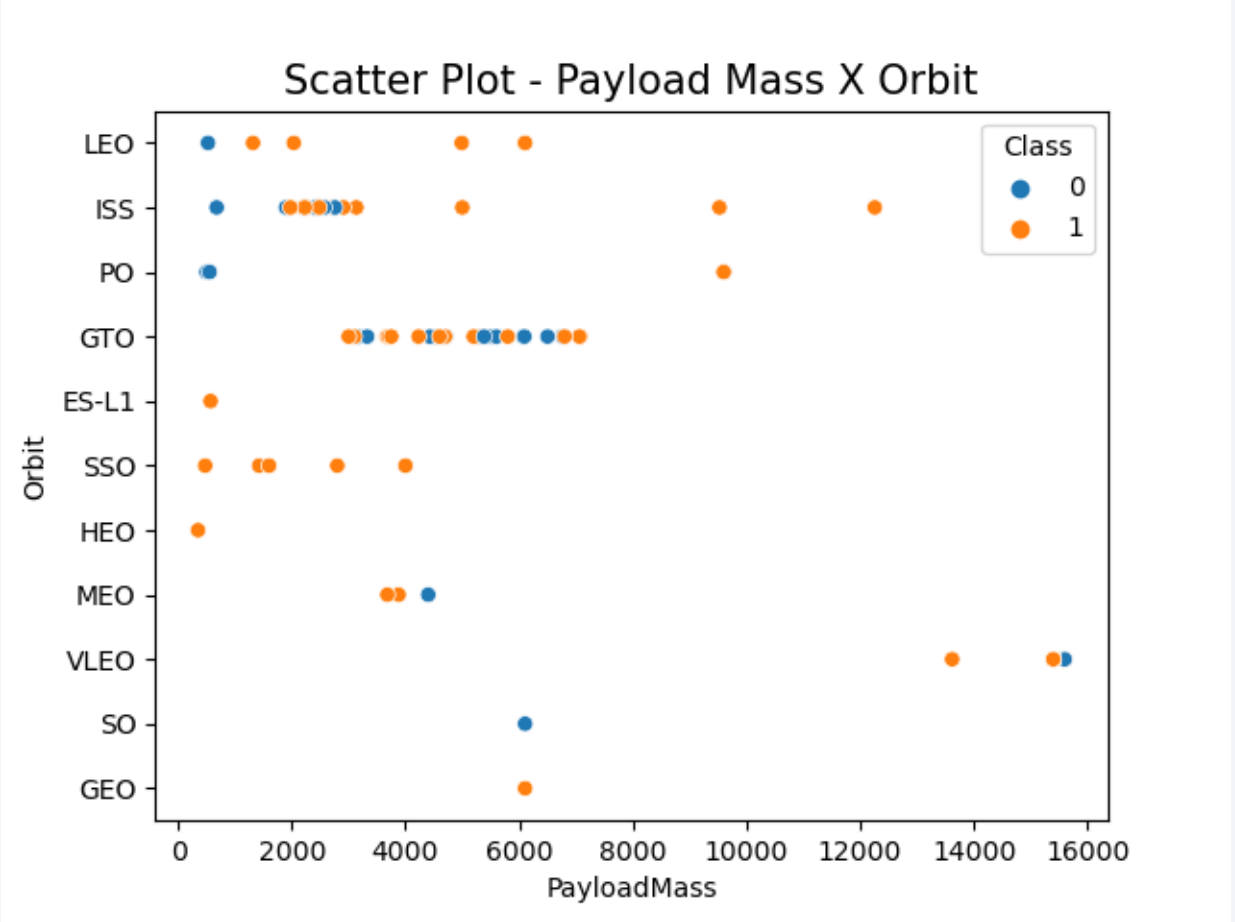
Flight Number vs. Orbit Type

- ES-L1, HEO, SO and GEO only has one data point and will be hard to estimate.
- ISS, GTO and VLEO have the most number of flights, making it easier to predict.



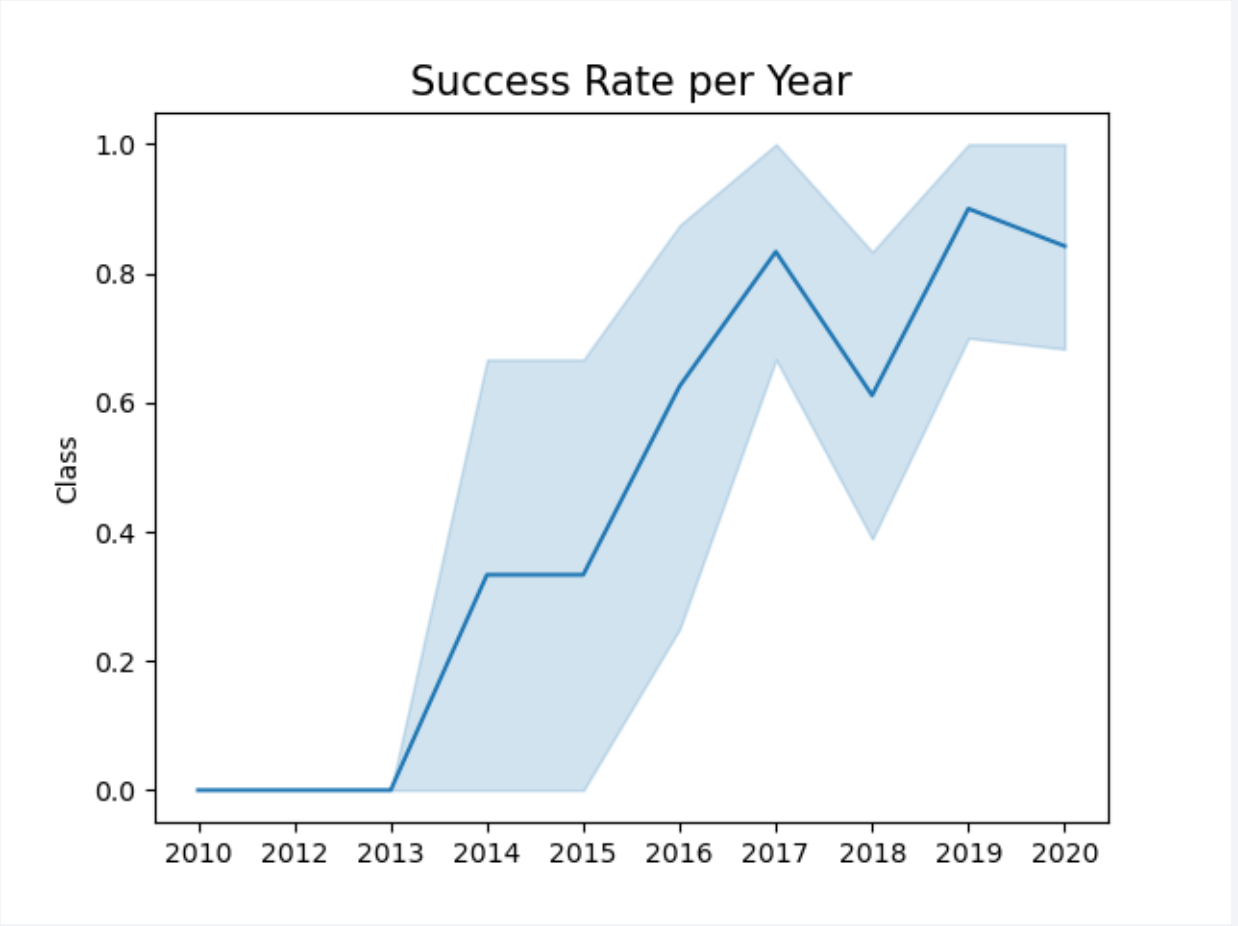
Payload vs. Orbit Type

- For ISS rockets the Payload Mass is between 2000 and 4000 Kg.
- For GTO orbits the mass is between 2000 and 8000 Kg.



Launch Success Yearly Trend

- Clearly as time progresses SpaceX is getting better at the production of rockets
- An sharp decline in the success rate in 2018.



All Launch Site Names

Find the names of the unique launch sites

```
[11]: %%sql

      SELECT DISTINCT Launch_Site FROM SPACEXTBL;

       * sqlite:///my_data1.db
      Done.
[11]:   Launch_Site
        CCAFS LC-40
        VAFB SLC-4E
        KSC LC-39A
        CCAFS SLC-40
```

Using DISTINCT function to find All Launch Site Names

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Calculate the total payload carried by boosters from NASA

```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayload FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)';
```

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

TotalPayload
45596

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1

```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) AS AveragePayloadMass FROM SPACEXTBL
WHERE Booster_Version='F9 v1.1';
```

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

AveragePayloadMass
2928.4

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

```
%%sql

SELECT MIN(Date) FROM SPACEXTBL
WHERE Mission_Outcome='Success';
```

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

MIN(Date)
01-03-2013

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 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 Booster_Version FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

```
%%sql

SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Result FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

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

```
%%sql

SELECT 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

2015 Launch Records

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

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, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%%sql

select substr(Date,4,2) AS Month, Landing_Outcome, Booster_Version, Launch_site from SPACEXTBL
WHERE substr(Date,7,4)='2015' AND Landing_Outcome ='Failure (drone ship)';
```

```
 * 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

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

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql

SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTBL
WHERE Date BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC;
```

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

Landing_Outcome	COUNT(Landing_Outcome)
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

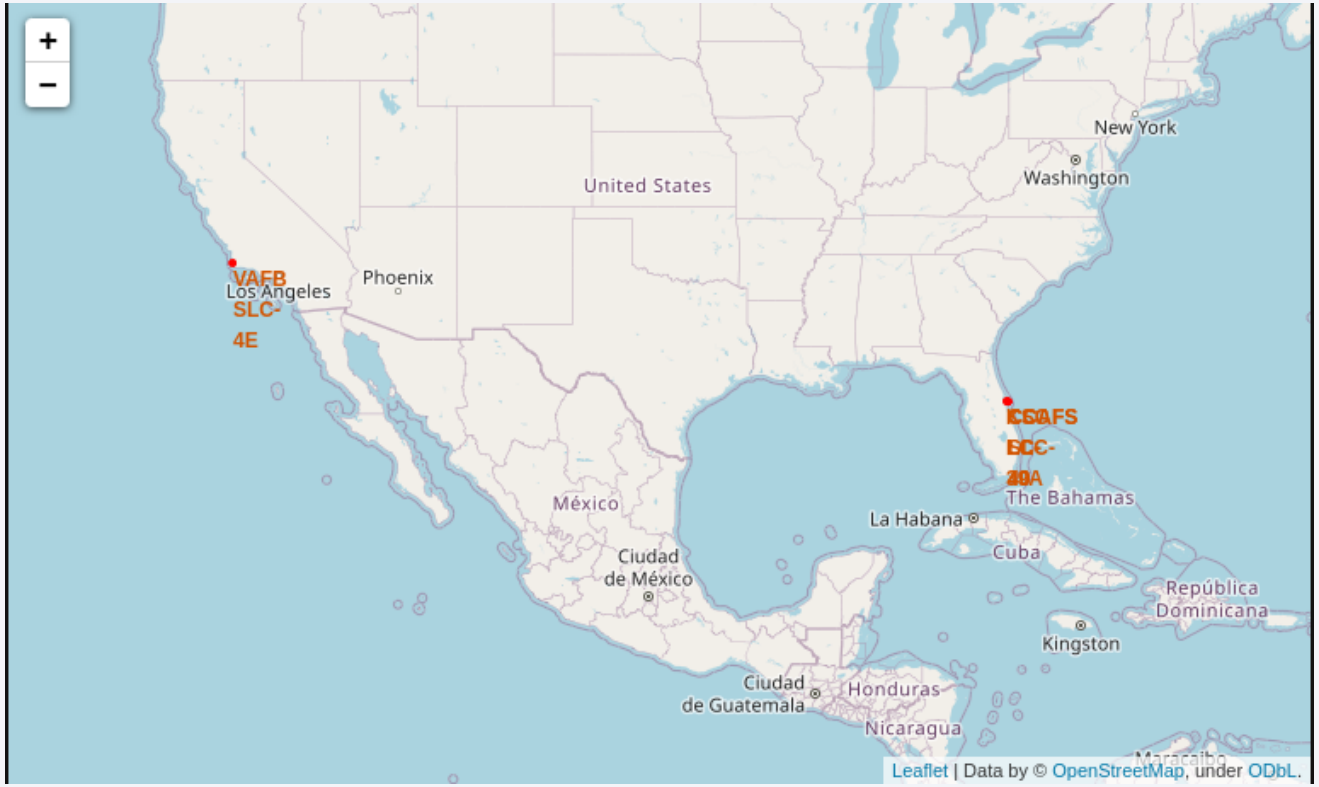
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

Launch Sites Locations.

VAFB SLC-4E is located on California because is close to the SpaceX HeadQuarters.

Launch Sites KSC LC-39A, CCAFS LC-40, CCAFS SLC-40 are located in Cape Canaveral, the usual place where NASA launch it's rockets.



<Folium Map Screenshot 2>

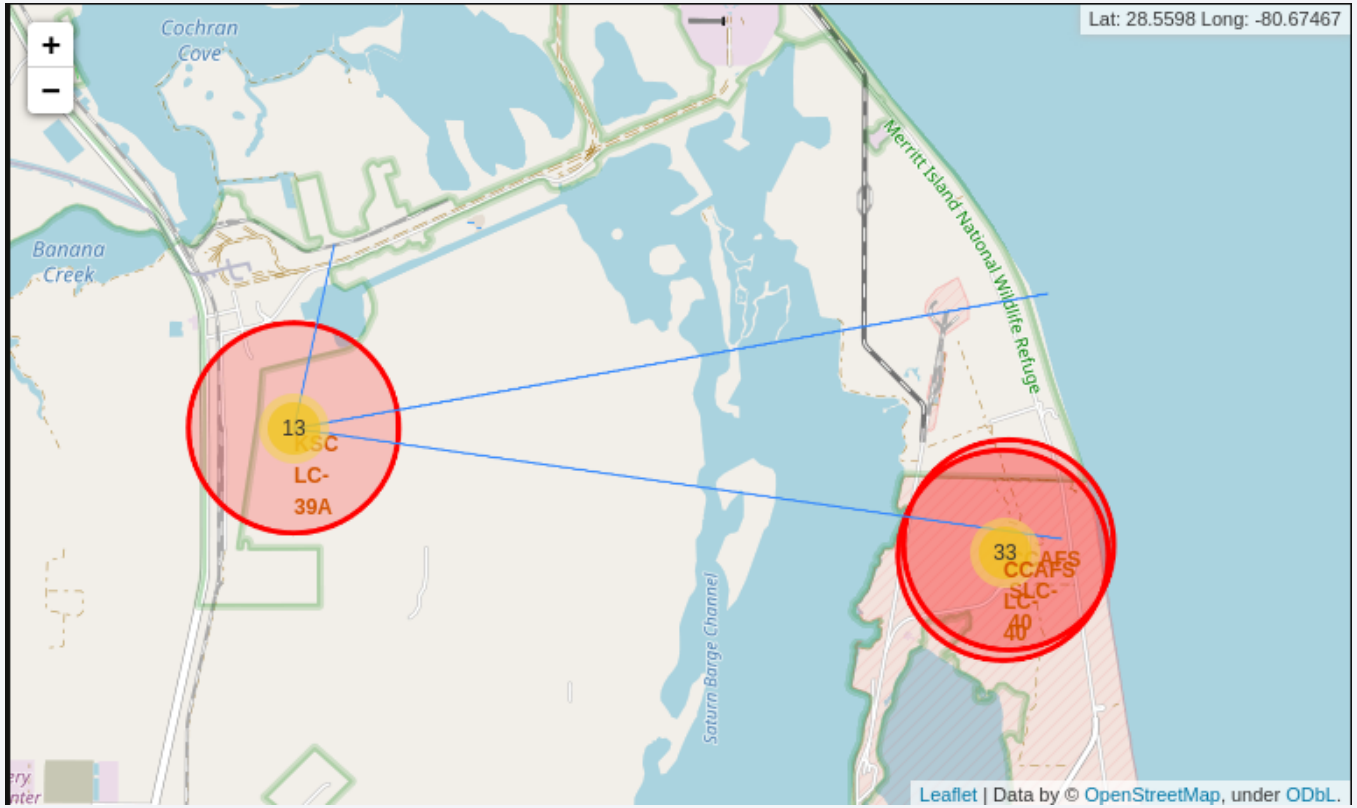
Number of rockets launched in clusters.

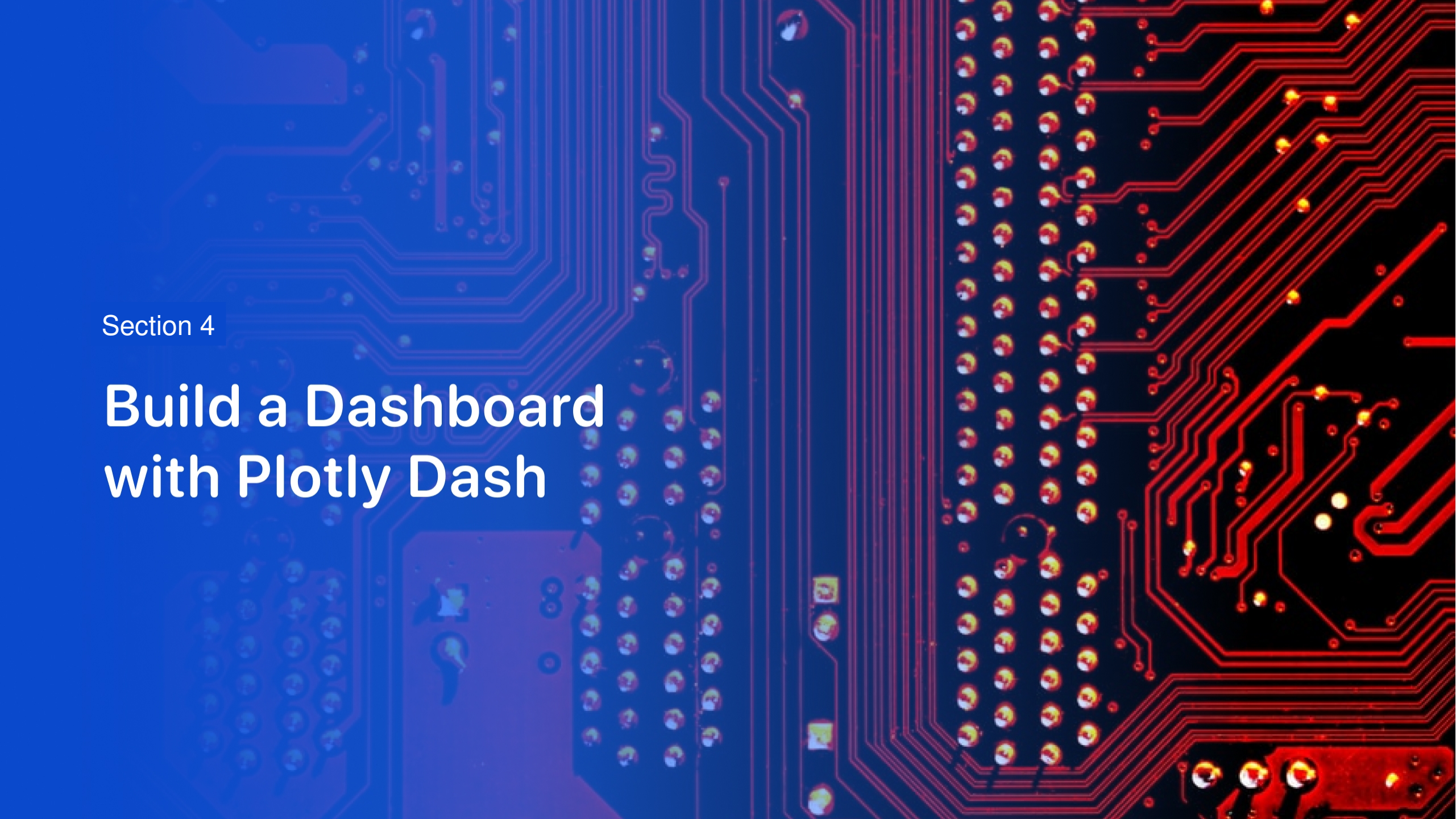
The rockets has been put in cluster based on their proximity and zoom on the map.

We can see the majority of the rockets are launched in Florida due to the fact that NASA is located there.

<Folium Map Screenshot 3>

Distance between Launch Site KSC LC-39A and a highway, the coastline and a railway.

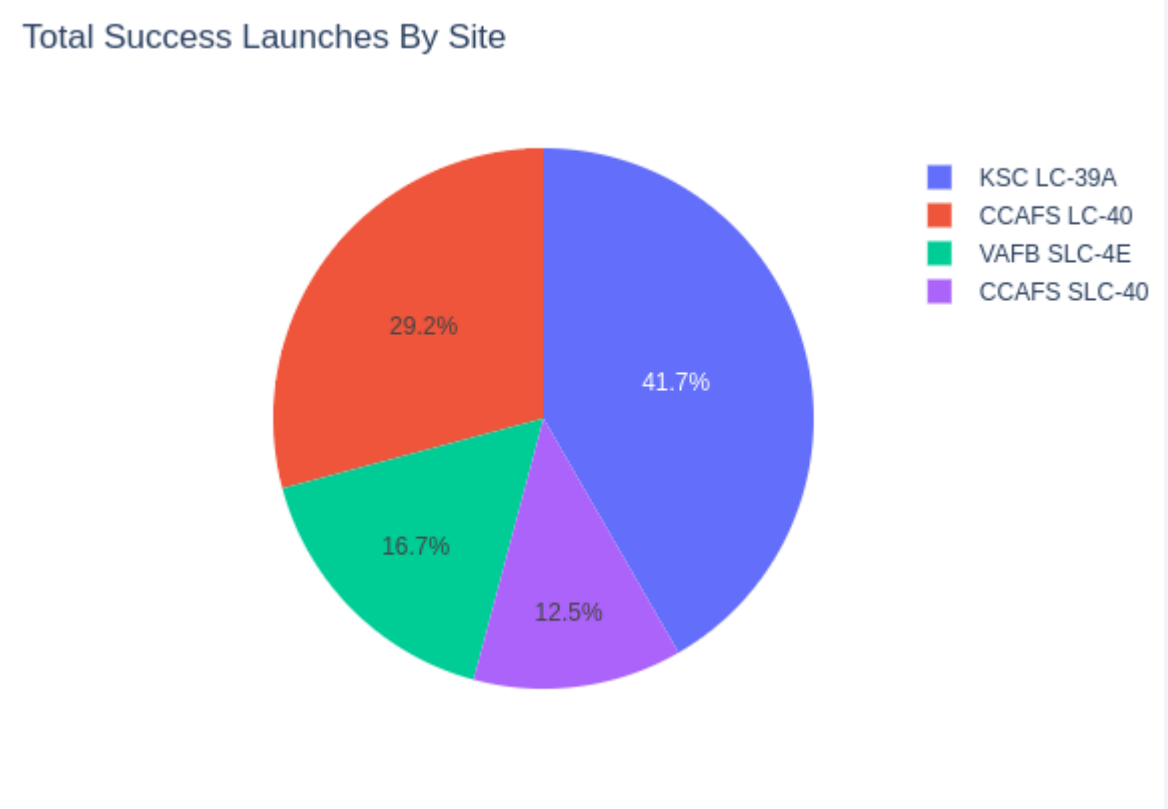



Section 4
Build a Dashboard with Plotly Dash

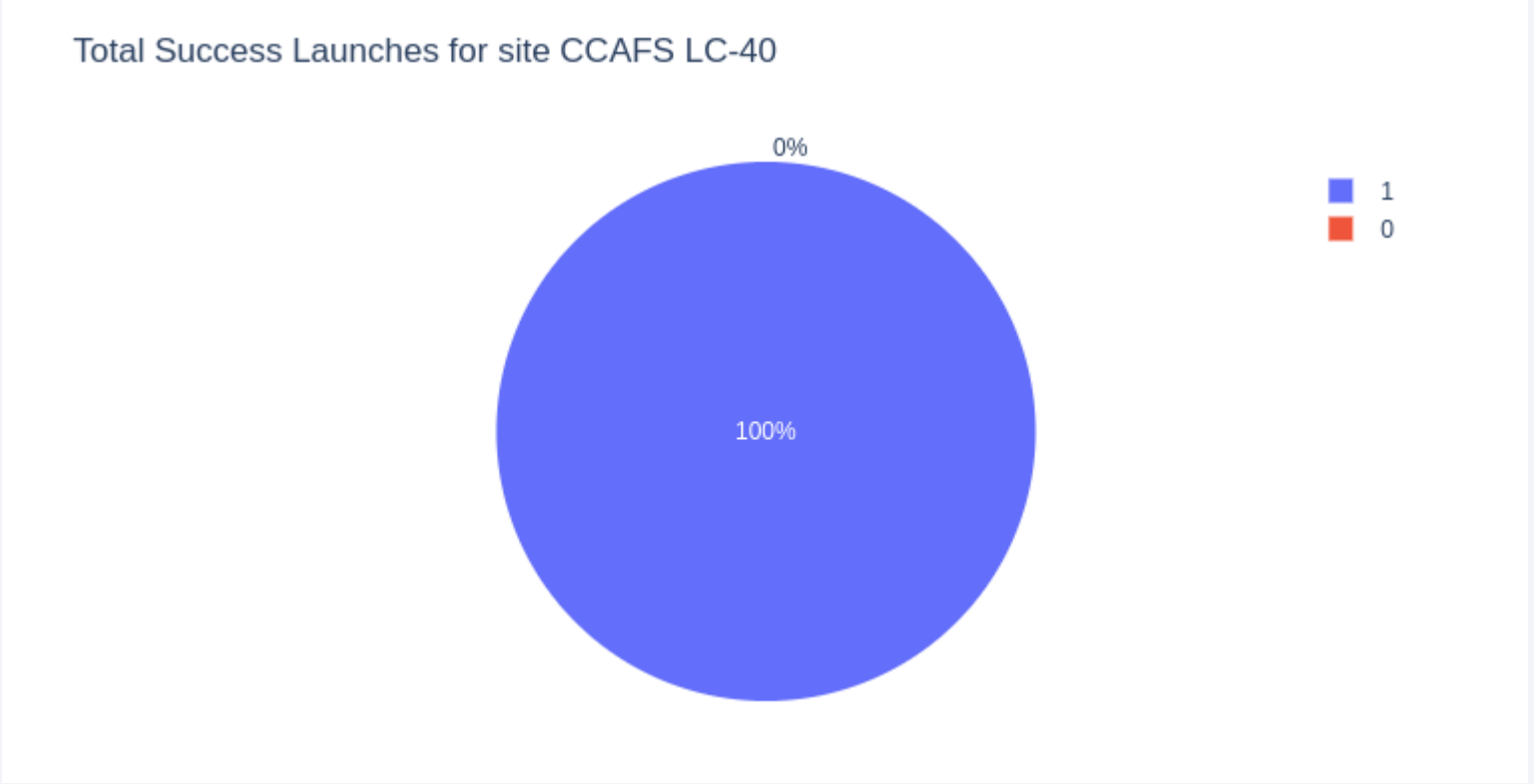
<Dashboard Screenshot 1>

The majority of success launches occur on the Launch Site KSC LC-39A

While the site with the lowest success is CCAFS SLC-40.



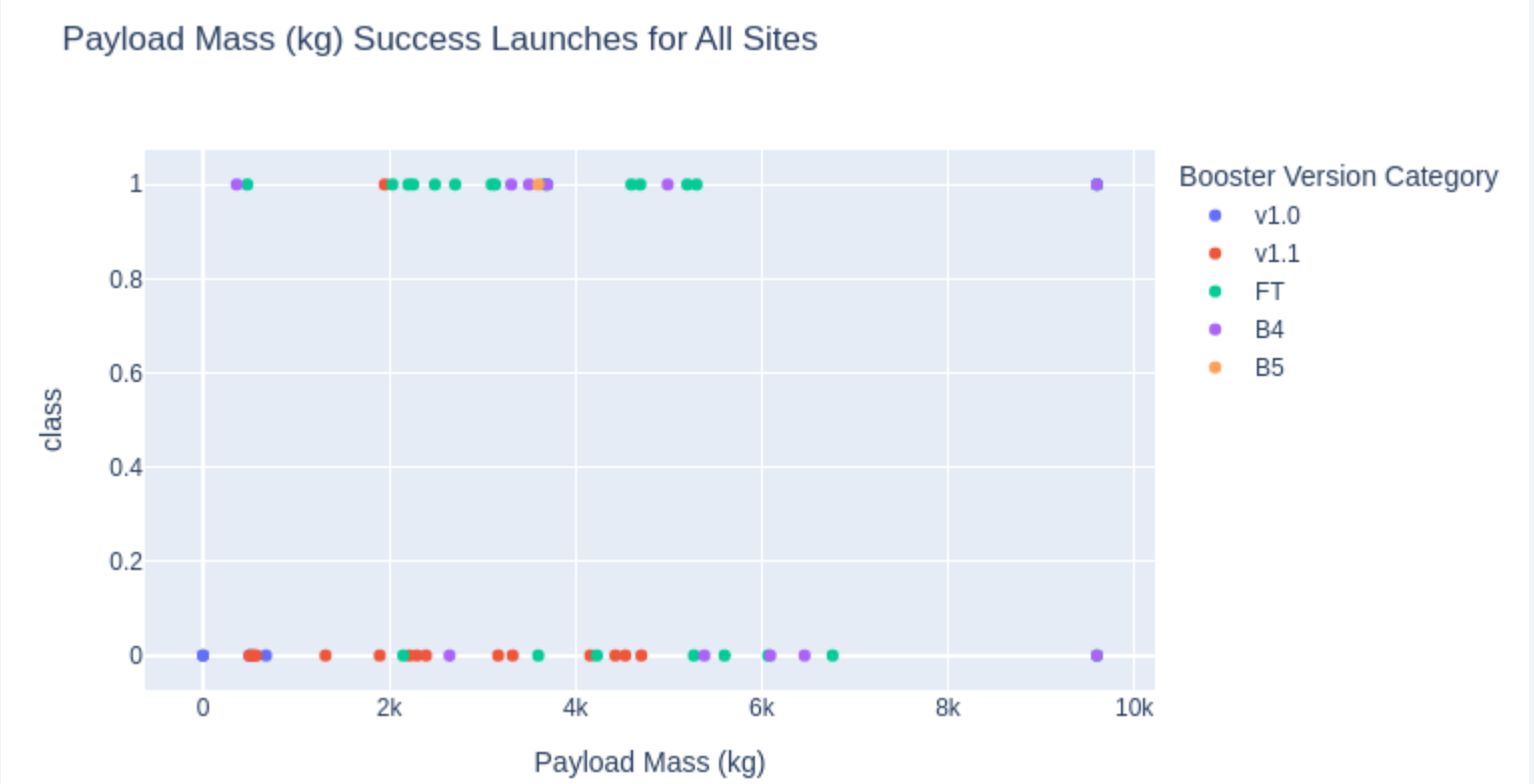
<Dashboard Screenshot 2>



<Dashboard Screenshot 3>

From the image we can see that for Payload Mass between 2000 and 4000 Kg there is the highest success rate.

The colours indicate that for Booster Version Category, the FT category has the highest success rate.

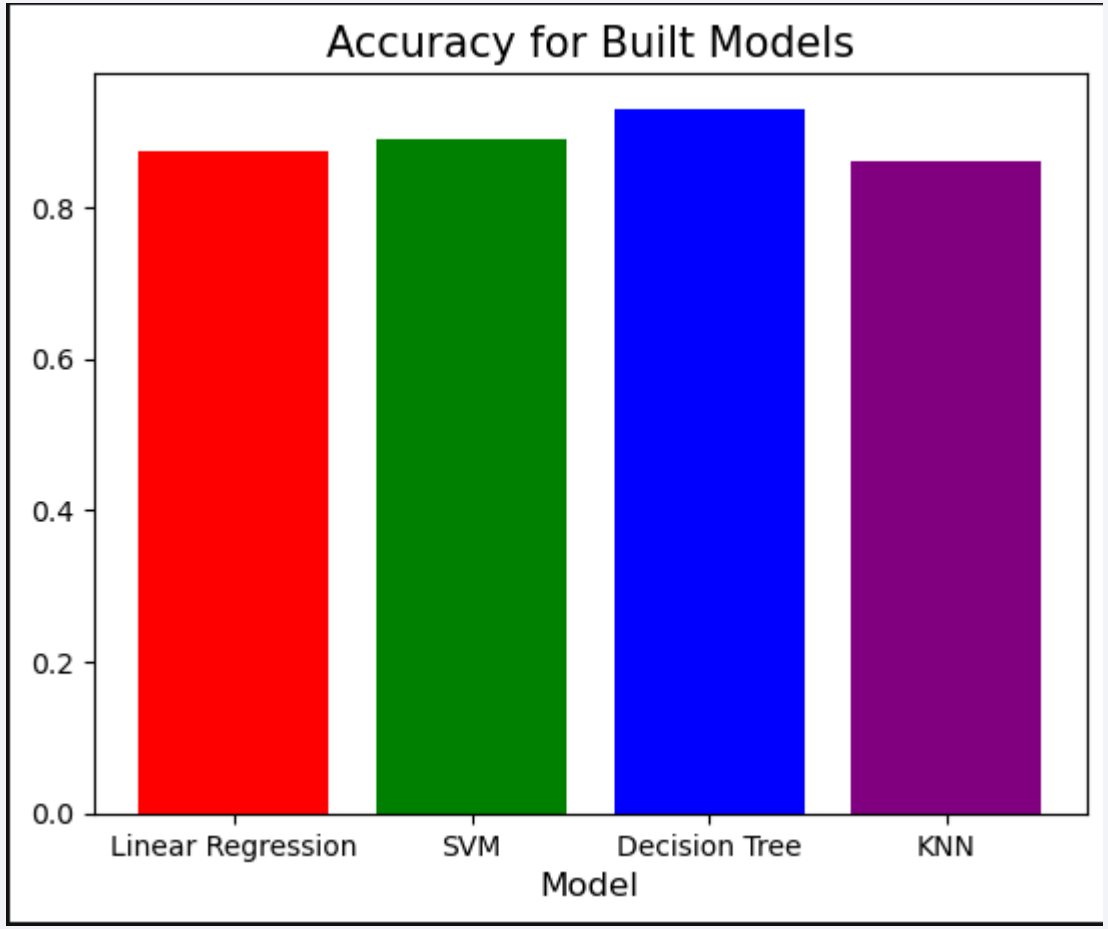


Section 5

Predictive Analysis (Classification)

Classification Accuracy

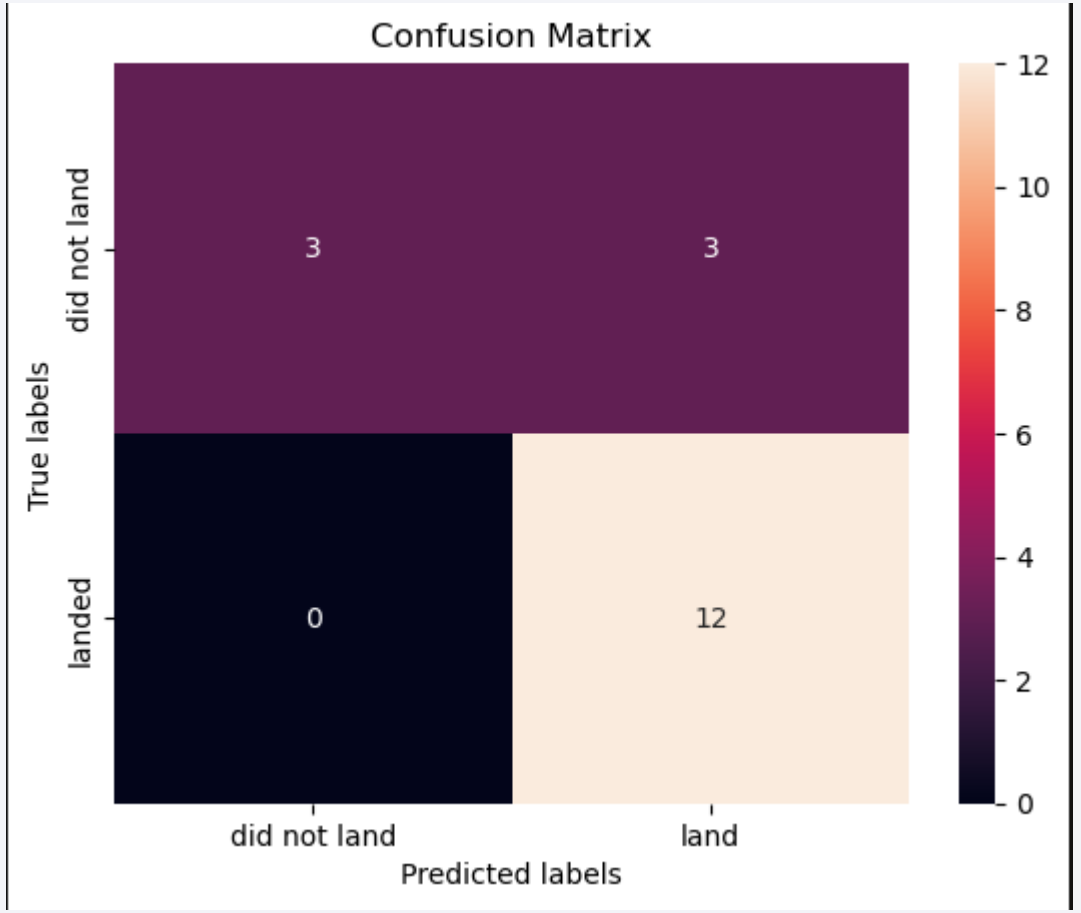
Plotting the accuracy of each model built in a Bar Plot we can see that the Decision Tree model resulted in the highest accuracy.



Confusion Matrix

To understand the Confusion Matrix, the numbers on the main diagonal are the labels correctly predicted and the labels on the secondary diagonal has been wrongly predicted.

The Decision Tree model only wrongly predicted 3 labels as a False Positive and no False Positives, in the end we got a good model.



Conclusions

Due the small size of the dataset, all the models built predicted igually, resulting in the same confusion matrix.

It’s possible to predict the results of a landing without the need to use any Rocket Engineering

With the increase on the success rate through the years, old data might not be as relevant

Appendix

SpaceX Wikipedia Page:

https://en.wikipedia.org/wiki/Spacex

Thank you!