IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yacine Bouzid
27/10/2023

Outline

Introduction

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.

So, the objective it’s too find out interesting insights from the SpaceX data to predict if the first stage will land to determine the cost of a launch.

Executive Summary

Summary of methodologies:

- Data Collection
- Data Wrangling
- Data Analysis
- Data Visualization
- Modeling

Summary of all results

- Exploratory data analysis results
- Interactive analytics results
- Predictive analysis results

Section 1

Methodology

Executive Summary

Data collection methodology:

- Retrieve the data from the SPACEX REST API
- Retrieve the data by web scraping some HTML Table from the Wikipedia pages using Python Beautiful soup

Perform data wrangling

- Extract some information from the data to have a better understanding of the data
- Convert landing outcomes columns (the information we want to predict) to class (0 or 1)

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- To build the model, we have created a NumPy array with the class columns, standardized the data, split the data into training and test data and create the classification models
- To tune the model, we have used the GridsearchCV object to find the best parameters and we have fitted the model with the best parameters
- To evaluate the classification models, we have calculated the score and created a confusion matrix to find the best model

Data Collection – SpaceX API

- The objective was to retrieve into a data frame the following data:

  Booster Version , Payload Mass , Orbit , Launch Site , Outcome , Flights , Grid Fins , Reused , Legs , Landing Pad , Block , Reused Count , Serial , Longitude , Latitude.

- We have retrieved the data using the following URL: https://api.spacexdata.com/v4/launches/past

- GitHub URL: https://github.com/Yacine13012/Data-science-final-exam/blob/main/Notebook/Lab1-Spacex-data_collection_api.ipynb

Step to collect the data from SpaceX API:

Getting the data using get method from the request library

↓

Decode the data using the json method

↓

Convert the json result into a dataframe using the json_normalize method from the pandas library

Data Collection - Scraping

- The objective was to retrieve into a data frame the following data:

  Flight No. , Launch site , Payload , Payload mass , Orbit , Customer , Launch outcome , Version Booster , Booster landing , Date , Time.

- We have retrieved the data using the following URL:

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- GitHub URL: https://github.com/Yacine13012/Data-science-final-exam/blob/main/Notebook/Lab2-Spacex-data_webscraping.ipynb

Step to collect the data from Wiki page:

Getting the data using beautiful soup

↓

Extract the information using find and find_all method

↓

Create a dataframe using the DataFrame method from pandas

Data Wrangling

- Beter understanding of the data
  - Calculate the percentage of missing values in each attribute.
  - Identify the types or each columns.
  - 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.

- Converting landing outcomes to class (0 or 1) with 0 when the landing failed and 1 when the landing succeed:
  - Visualization of the different values of outcome (True ASDS , False ASDS …)
  - Creation of a set of values where the mission outcome did not land
  - Append 0 in a list when the value is in the set of bad outcomes else Append 1

- GitHub URL: https://github.com/Yacine13012/Data-science-final-exam/blob/main/Notebook/Lab3-Spacex-data_wrangling.ipynb

EDA with Data Visualization

- Charts plotted:
  - Flight Number vs. Launch Site.
  - Payload vs Launch Site.
  - Rate vs Orbit Type.
  - Payload vs Orbit type.
  - Yearly average launch success rate.
- Purpose of the first 4 plots:

Determine how this attribute are correlated and to determine which attribute are correlated with successful landing while the fifth plot is to check how the launch success rate evolute through the year.

- GitHub URL: https://github.com/Yacine13012/Data-science-final-exam/blob/main/Notebook/Lab5-jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- SQL queries performed:
  - Display the unique launch site name.
  - Display 5 records where launch sites begin with `CCA`.
  - Calculate the total payload carried by boosters from NASA.
  - Calculate the average payload mass carried by booster version F9 v1.1.
  - Find the dates of the first successful landing outcome on ground pad.
  - List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

EDA with SQL

- SQL queries performed:
  - Calculate the total number of successful and failure mission outcomes.
  - List the names of the booster which have carried the maximum payload mass.
  - List the failed landing outcomes in drone ship, their booster versions, and launch site names for 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 URL : https://github.com/Yacine13012/Data-science-final-exam/blob/main/Notebook/Lab4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Objects created and added to a folium map:
    - Circle with pop up and marker to indicated where is located the NASA.
    - Circle with pop up and marker to indicated where are located the launch sites.
    - Marker Cluster to add the load outcomes for each site with green marker if the launch was successful and red marker in the launch was failed.
    - Line between a site and the nearest coastline, railway, highway and city with a distance marker to visualize how far the site is from these points of interest.
    - I have created this object to have a better understanding on how the site have been chosen and to where the site with the most successful rate is located to find some correlation between site location and launch successful rate.
- Github URL: https://github.com/Yacine13012/Data-science-final-exam/blob/main/Notebook/lab6_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to a dashboard
  - Drop-down to select different launch site (All, CCAFS LC-40, VAFB SLC-4 , KSC LC-39A, CAFS SLC-40).
  - Pie chart to show the total succesfull launches for the selected site.
  - Range Slider to Select a range of Payload.
  - Scatter plot to show the correlation between payload and successful launch grouped by booster category for the selected site and selected range.

Build a Dashboard with Plotly Dash

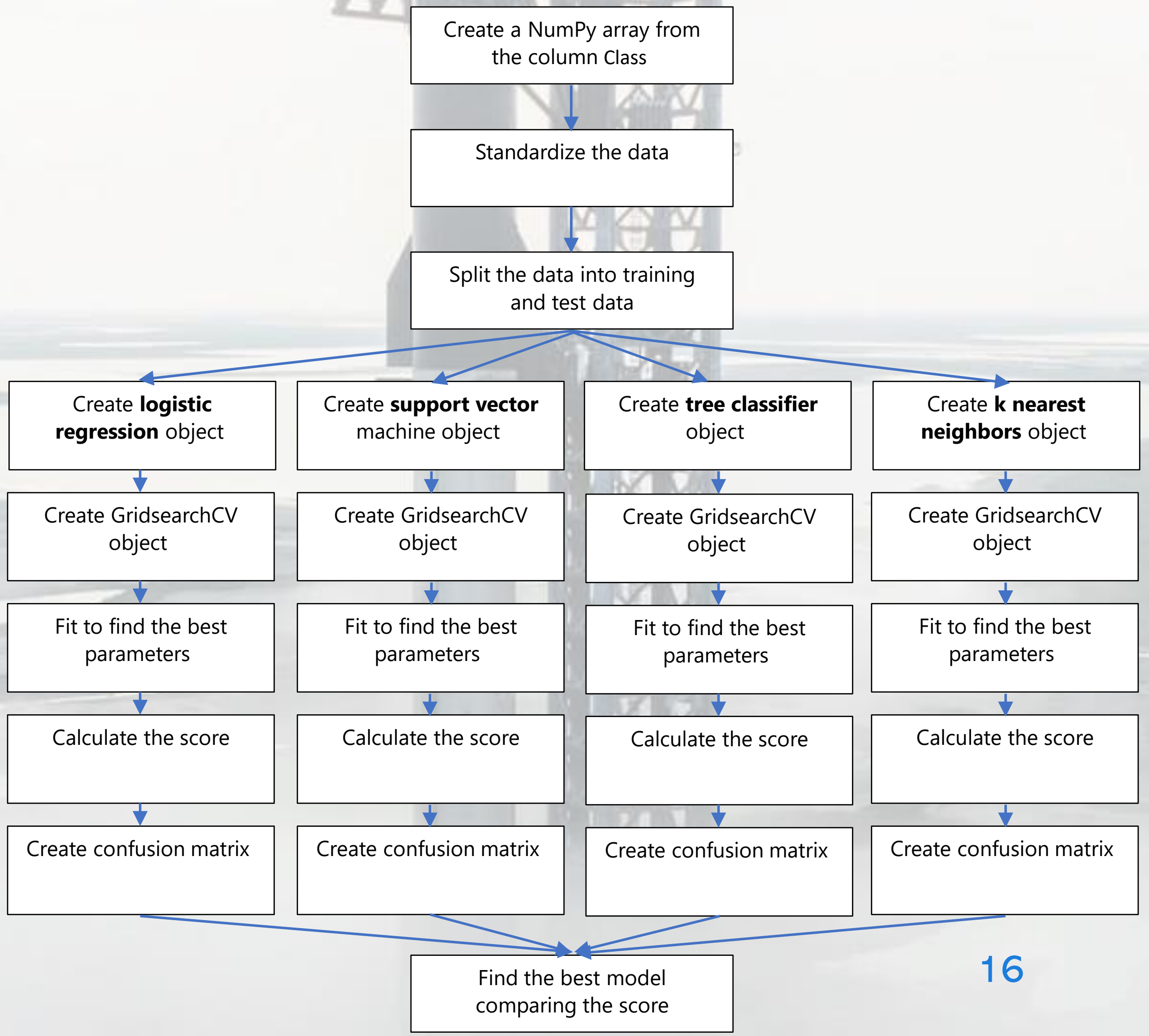
- Adding these plots and interactions help us to find visually:
  - Which site has the largest successful launches.
  - Which site has the highest launch success rate.
  - Which payload range has the highest launch success rate.
  - Which payload range has the lowest launch success rate.
  - Which F9 Booster version has the highest launch success rate.

- GitHub URL: https://github.com/Yacine13012/Data-science-final-exam/blob/main/Dashboard/spacex_dash_app.py

Predictive Analysis (Classification)

- To built, evaluated, improved, and found the best performing classification model
  - Create a NumPy array from the column Class in data, by applying the method to_numpy().
  - Standardize the data preprocessing.StandardScaler().
  - Use the function train_test_split to split the data X and Y into training and test data (80% train data vs 20% test data).
  - For logistic regression, support vector machine, tree classifier and k nearest neighbors models:
    - Create model object.
    - Create a GridsearchCV object with the model object and some parameters to find the best parameters.
    - Calculate the score.
    - Create confusion matrix.
  - Compare the different score to find the best model.

- GitHub URL: https://github.com/Yacine13012/Data-science-final-exam/blob/main/Notebook/Lab7_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

- Exploratory data analysis results:
  - The first 25 launches was mostly launched from the CCAF5 SLC 40 with a low successful rate.
  - The heavy payload mass launched by CCAFS SLC 40 and KSC LC 39A.
  - The orbit with more successful are ES-L1, GEO,HEO.
  - The success rate kept increasing since 2013.
- Interactive analytics demo in screenshots results:
  - The site with the largest successful launches.
  - The site has the highest launch success rate.
  - The payload range 2,000 to 6,000kg has the highest launch success rate.
  - The payload range 0 to 2,000kg has the lowest launch success rate.
  - FT and B5 Booster versions has the highest launch success rate.
- Predictive analysis results:

  The best predictive model is k nearest neighbors with a score of 0,94 each.

Section 2

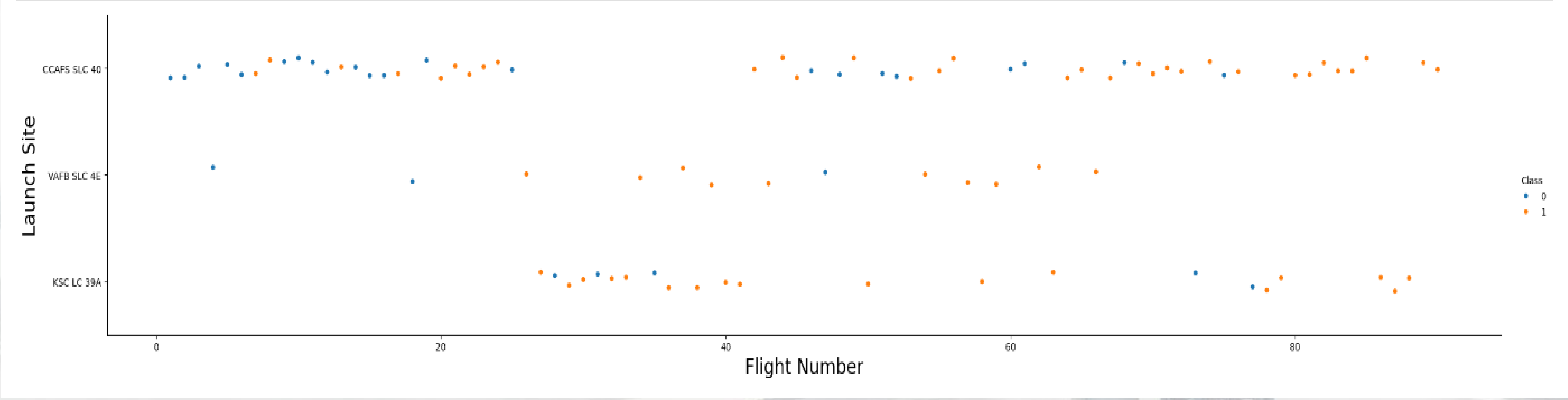
Insights drawn from EDA

Flight Number vs. Launch Site

- Code to show a scatter plot of Flight Number vs. Launch Site:

```
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)

plt.xlabel("Flight Number",fontsize=20)

plt.ylabel("Launch Site",fontsize=20)

plt.show()
```

Flight Number vs. Launch Site



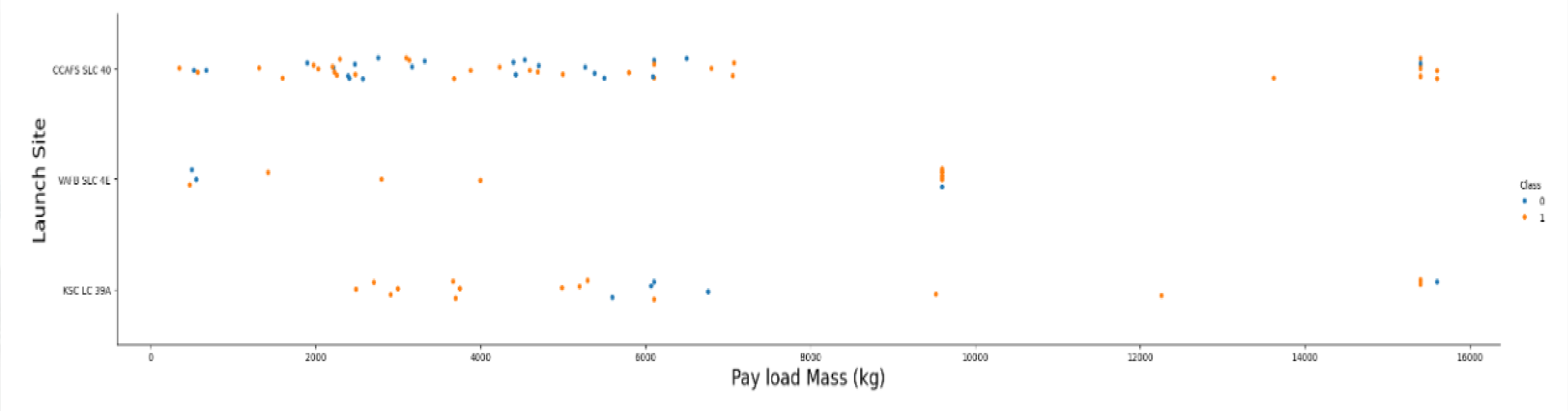
We can see from this graph that the first 25 launches was mostly launched from the CCAF5 SLC 40 site with a high failure rate then the launch successful rate increase.

Payload vs. Launch Site

- Code to show a scatter plot of Payload vs Launch Site

```
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

Payload vs. Launch Site



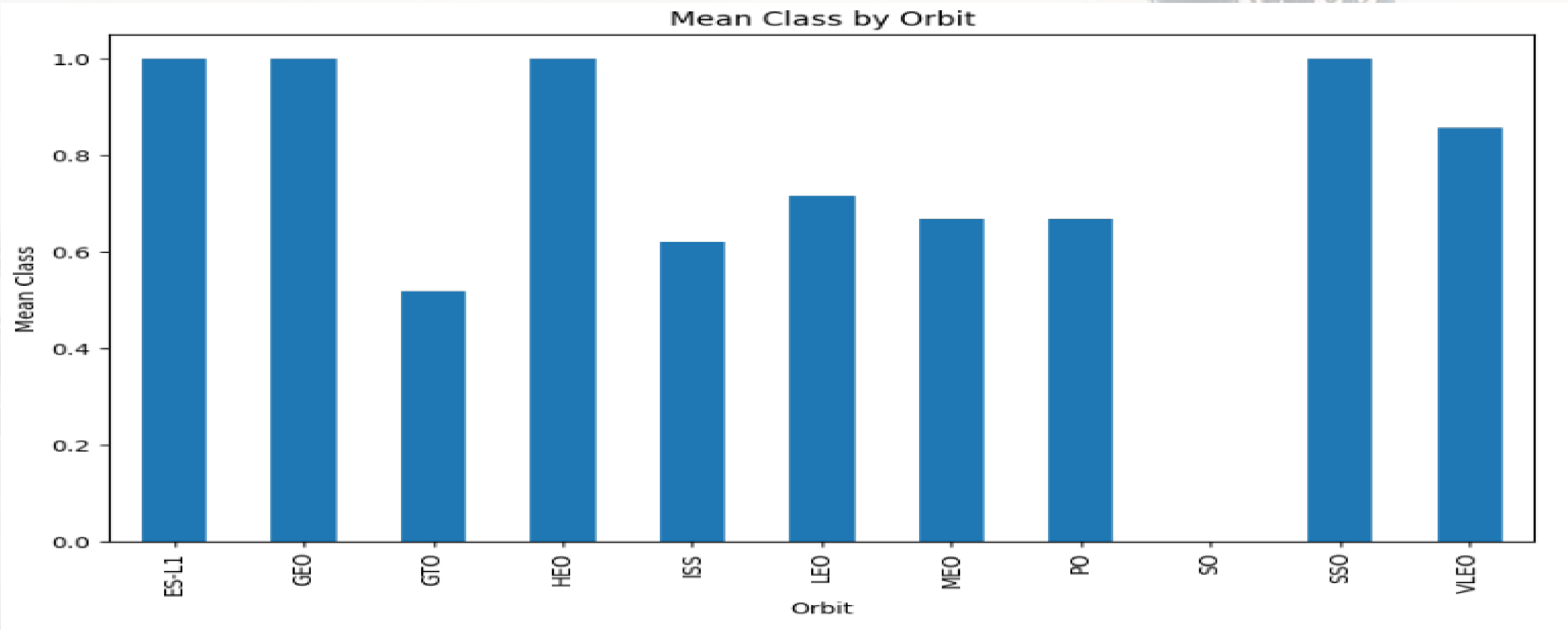
We can see from this graph that the heavy payload mass are launched by CCAFS SLC 40 and KSC LC 39A launch sites, there are no heavy rockets launched by VAFB SLC 4E.

Success Rate vs. Orbit Type

- Code to show a bar chart for the success rate of each orbit type:

```
df_orbit = df.groupby('Orbit', axis=0).mean()['Class']


df_orbit.plot(kind='bar', figsize=(10, 6))


plt.xlabel('Orbit') # add to x-label to the plot
plt.ylabel('Mean Class') # add y-label to the plot
plt.title('Mean Class by Orbit') # add title to the plot
plt.show()
```

Flight Number vs. Orbit Type



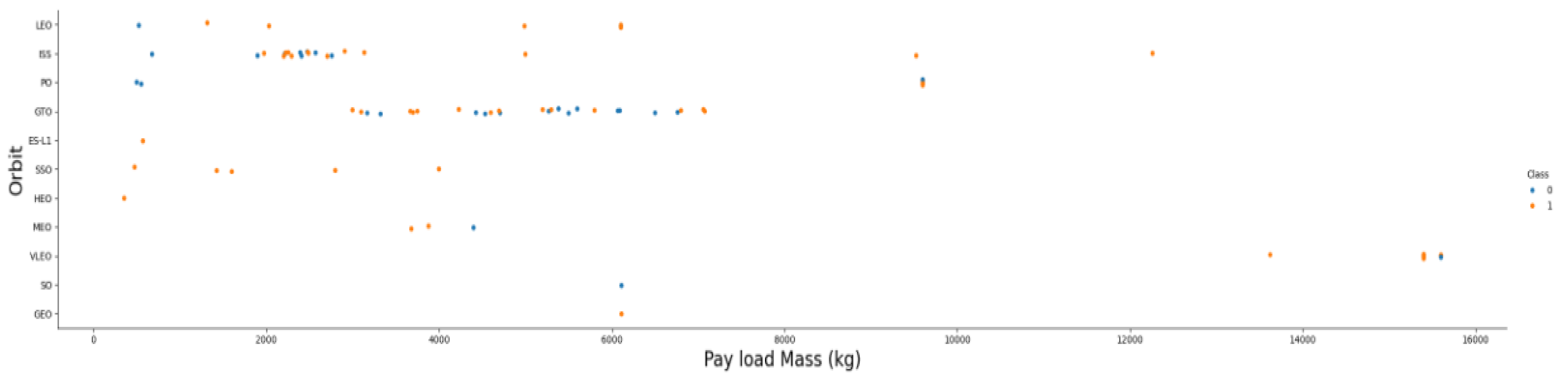
We can see from this graph that ES-L1, GEO,HEO,SSO are 100% of success and VLEO more the 80% of success , the other orbits are a lower success.

Payload vs. Orbit Type

- Code to show a scatter point of payload vs orbit type:

```
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)

plt.xlabel("Pay load Mass (kg)",fontsize=20)

plt.ylabel("Orbit",fontsize=20)

plt.show()
```

Payload vs. Orbit Type



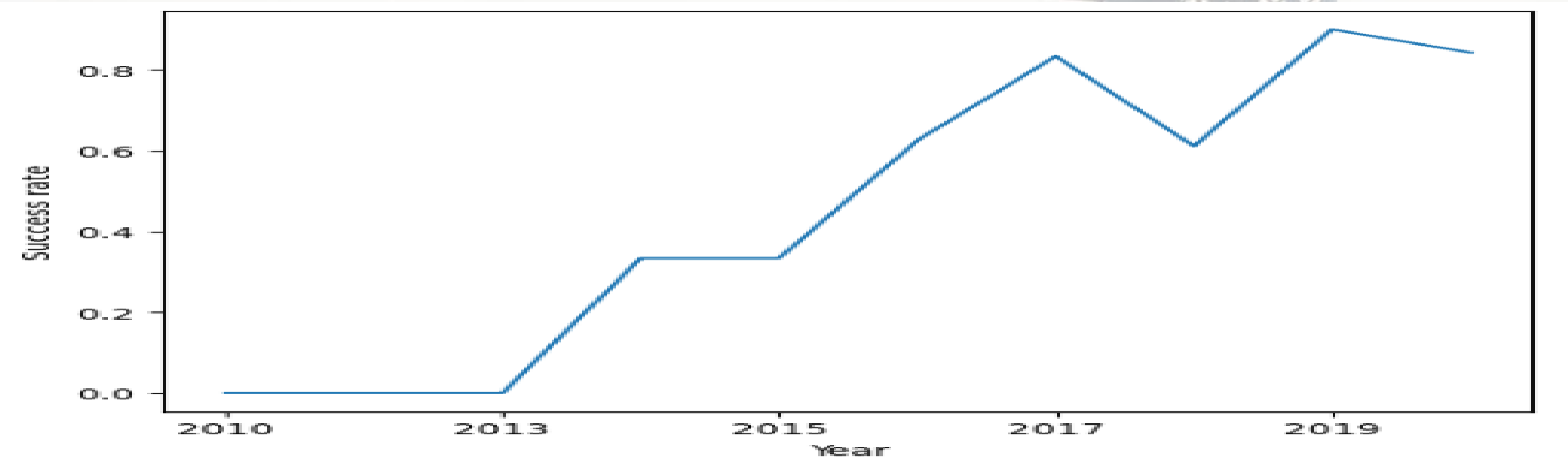
We can see from this graph that the heavy payload mass are launched essentially on Polar, LEO and ISS. However, for GTO both positive landing rate and negative landing (unsuccessful mission) are both here we can say that there seems to be no relationship between Payload when in GTO orbit.

Launch Success Yearly Trend

- Code to show a line chart of yearly average success rate

```
df1 = df.groupby('Date', axis=0).mean()['Class']

df1.head()

df1.set_index("Date",inplace=True)

df1.plot()

plt.xlabel("Year",fontsize=10)

plt.ylabel("Success rate",fontsize=10)

plt.show()
```

Launch Success Yearly Trend



We can easily see from this trend that the success rate kept increasing since 2013.

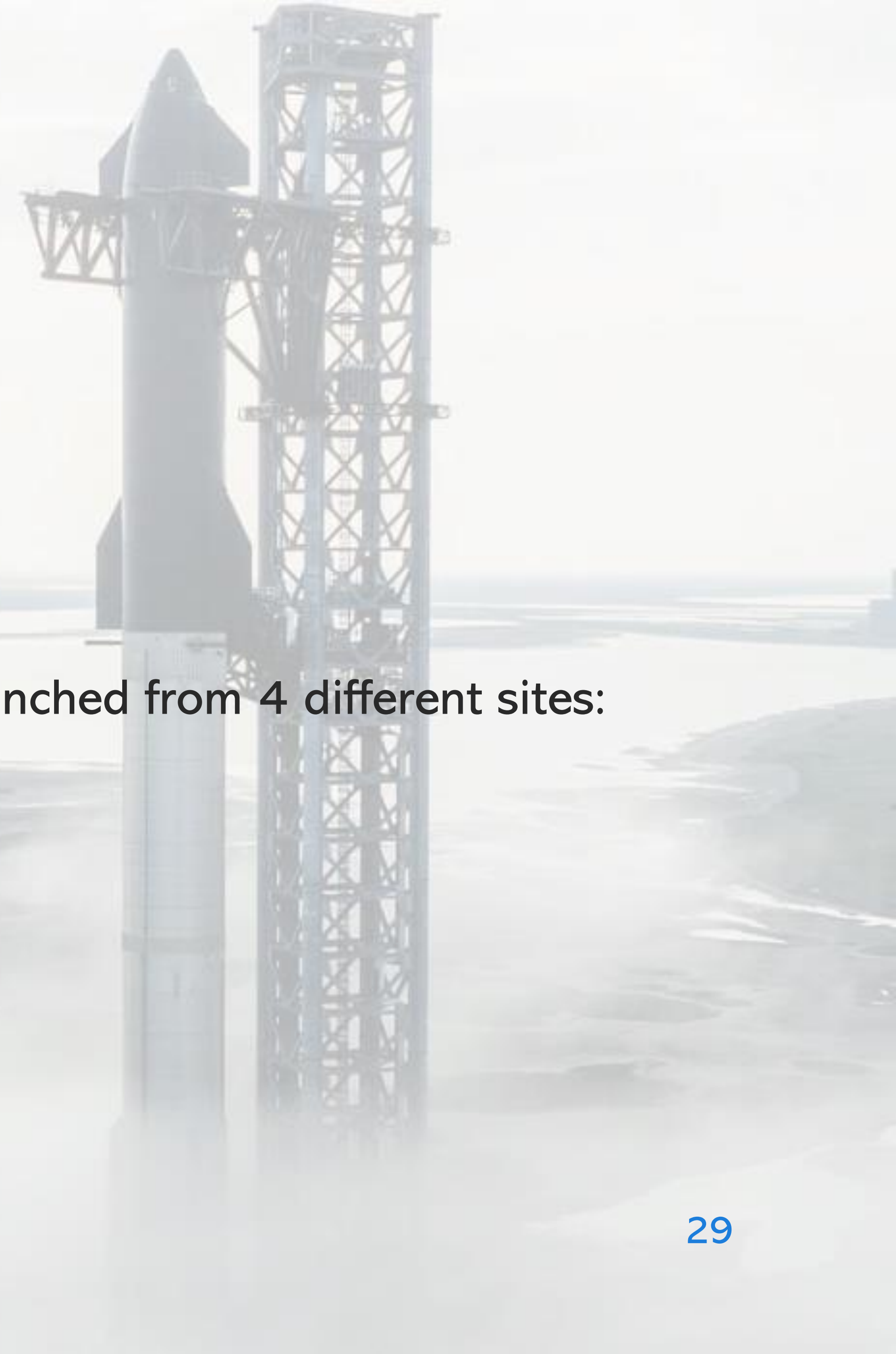
All Launch Site Names

- SQL query to display the unique launch site name:

```
%sql select distinct "Launch_Site" from SPACEXTBL;
```

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

- The Space X Falcon 9 rockets were launched from 4 different sites:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40



Launch Site Names Begin with 'CCA'

- SQL query to display 5 records where launch sites begin with `CCA`:

%sql select * from SPACEXTBL where "Launch_Site" LIKE "CCA%" LIMIT 5

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- We can see the 5 first records where launch sites begin with `CCA`. All the 5 launches are launched from CCAFS LC-40 to LEO and failed to landed.

Total Payload Mass

- SQL Query to calculate the total payload carried by boosters from NASA:

%sql select "Booster_Version" ,

SUM("PAYLOAD_MASS__KG_") AS TOTAL_PAYLOAD_MASS,

"customer" from SPACEXTBL

where "customer" = "NASA (CRS)“

GROUP BY "Booster_Version";

- We can see the total payload carried by boosters from NASA with the total payload from 500kg (F9 v1.0 B0006) to 3310kg (F9 B4 B1039.1).

Booster_Version	TOTAL_PAYLOAD_MASS	Customer
F9 B4 B1039.2	2647	NASA (CRS)
F9 B4 B1039.1	3310	NASA (CRS)
F9 B4 B1045.2	2697	NASA (CRS)
F9 B5 B1056.2	2268	NASA (CRS)
F9 B5 B1058.4	2972	NASA (CRS)
F9 B5 B1059.2	1977	NASA (CRS)
F9 B5B1050	2500	NASA (CRS)
F9 B5B1056.1	2495	NASA (CRS)
F9 FT B1035.2	2205	NASA (CRS)
F9 FT B1021.1	3136	NASA (CRS)
F9 FT B1025.1	2257	NASA (CRS)
F9 FT B1031.1	2490	NASA (CRS)
F9 FT B1035.1	2708	NASA (CRS)
F9 v1.0 B0006	500	NASA (CRS)
F9 v1.0 B0007	677	NASA (CRS)
F9 v1.1	2296	NASA (CRS)
F9 v1.1 B1010	2216	NASA (CRS)
F9 v1.1 B1012	2395	NASA (CRS)
F9 v1.1 B1015	1898	NASA (CRS)
F9 v1.1 B1018	1952	NASA (CRS)

Average Payload Mass by F9 v1.1

- Query to calculate the average payload mass carried by booster version F9 v1.1:

```
%sql select "Date","Landing_Outcome" As "First Landing Success"
 from SPACEXTBL
 where "Landing_Outcome" = "Success";
```

Booster_Version	AVG_PAYLOAD_MASS
F9 v1.1	2928.4

- The average payload carried by booster version F9 v1.1 is 2928.4kg.

First Successful Ground Landing Date

- SQL Query to find the dates of the first successful landing outcome on ground pad:

```
%sql select min("Date") As "Date",

"Booster_Version", "customer","Landing_Outcome" from SPACEXTBL

 where "Landing_Outcome" = "Success (ground pad)";
```

Date	Booster_Version	Customer	Landing_Outcome
2015-12-22	F9 FT B1019	Orbcomm	Success (ground pad)

- The booster which success first to land on ground pad is the F9 FT B1019 launched by Orbcomm on 22th of December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL Query to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

```
%sql select "Booster_Version" , "PAYLOAD_MASS__KG_", "Landing_Outcome“
 from SPACEXTBL where "Landing_Outcome" = "Success (drone ship)" and
 "PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000
```

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

- There are 4 boosters which success to land in drop ship with a payload between 4000 and 6000kg

Total Number of Successful and Failure Mission Outcomes

- SQL Query to calculate the total number of successful and failure mission outcomes:

```
%sql select "Mission_Outcome", COUNT("Mission_Outcome") AS
 "Total Mission Outcome" from SPACEXTBL GROUP BY "Mission_Outcome";
```

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

- There are 100 missions that succeed which include one with the payload status is unclear. However, there is one mission that failed.

Boosters Carried Maximum Payload

- SQL Query to list the names of the booster which have carried the maximum payload mass:

%sql select "Booster_Version" , "PAYLOAD_MASS__KG_" from SPACEXTBL

where "PAYLOAD_MASS__KG_" = (select MAX("PAYLOAD_MASS__KG_")

from SPACEXTBL)

- There are 12 different boosters which have carried the maximum payload mass (15600kg).

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- SQL Query to list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

```
%sql select "Date", 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'
```

Date	Month	Landing_Outcome	Booster_Version	Launch_Site
2015-10-01	10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- In 2015, there are 2 boosters (F9 v1.1 B1012 on November, F9 v1.1 B1015 on April) that failed to land in a drop ship.

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

- SQL 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:

```
select "Landing_Outcome", COUNT("Landing_Outcome") AS
Total from SPACEXTBL where "Date" > 2010-06-04 and
"Date" > 2017-03-20 GROUP BY "Landing_Outcome“
ORDER BY Total Desc
```

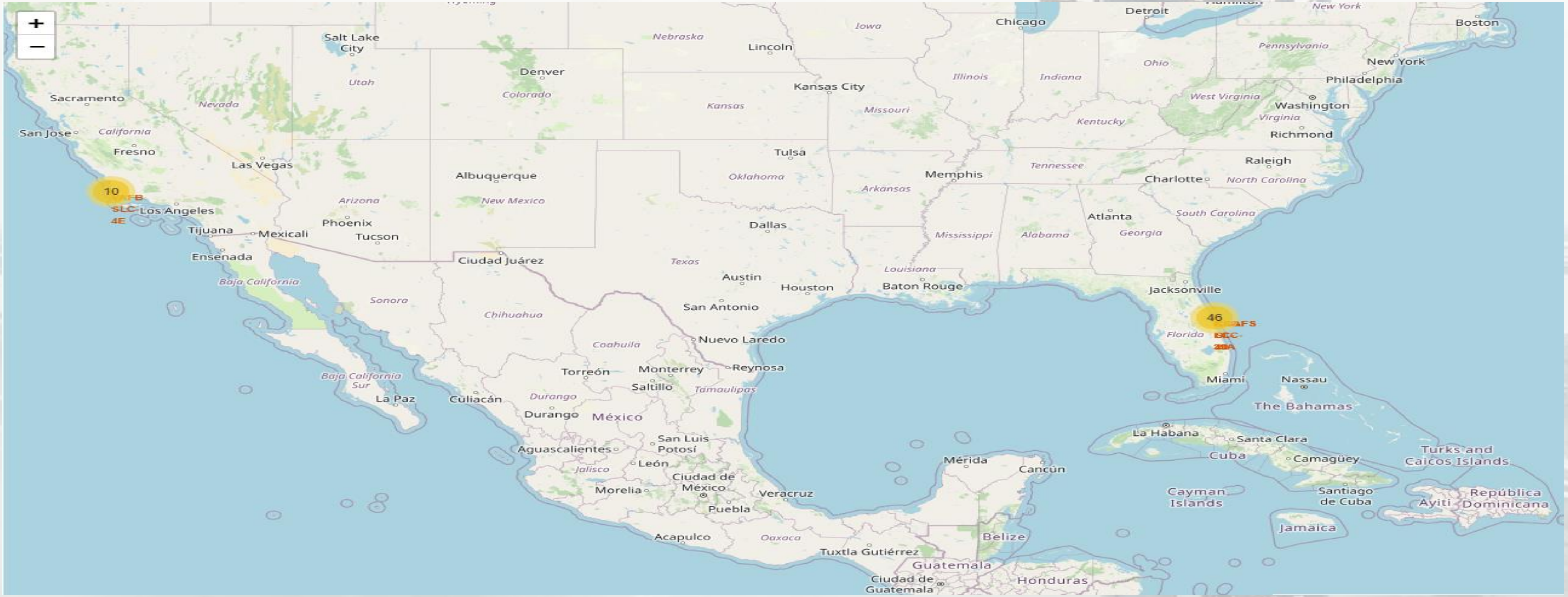
- There are 15 launches that succeed vs 5 that failed in drop ship. However, there are 9 launches that succeed with 0 failure in ground pad.

Landing_Outcome	Total
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

Section 3

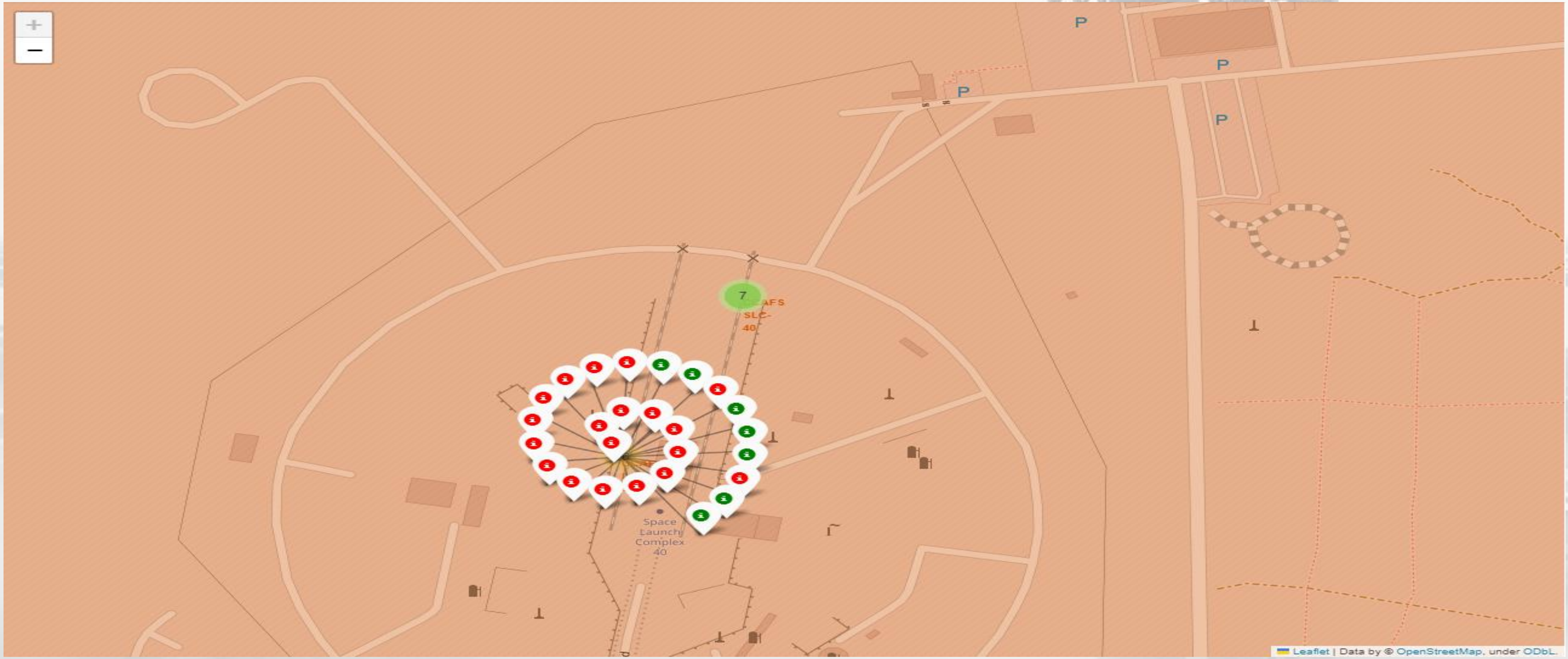
Launch Sites Proximities Analysis

Launch site location map



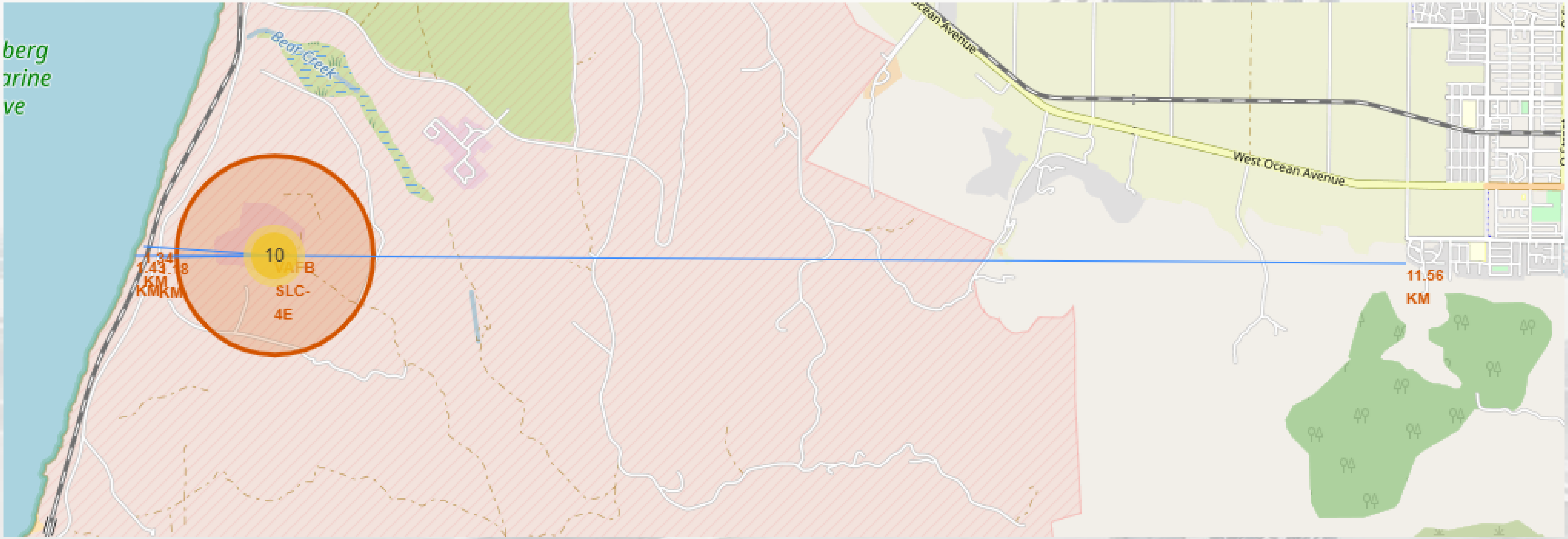
We can see from this map that there are 4,6 times more booster launched from the east coast nearby the Nasa site (46) than from the west coast (10).

Launch site location map with color-labeled launch outcomes



With the color labeled we can easily see that there are more failed than success launches from this location.

Distance between launch site and points of interests

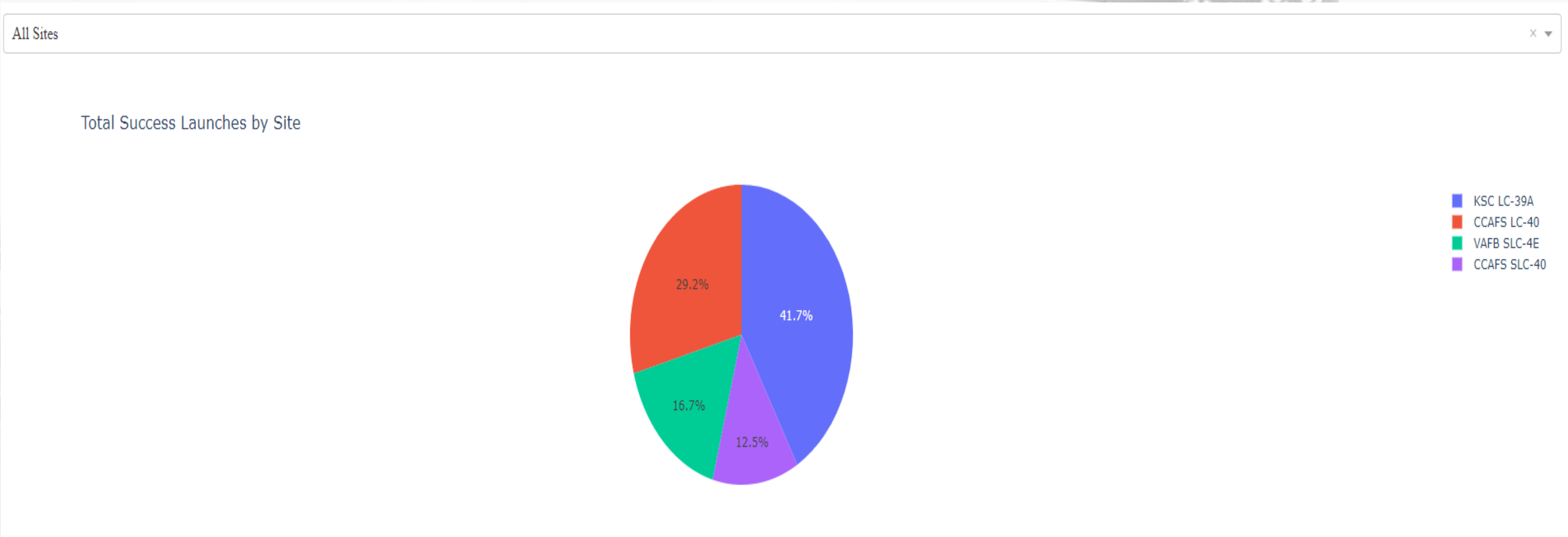


- The launch site is far from the nearer city and nearby the coastline for the security of people or material especially when the launch fail.
- The launch site is nearby railway and highway for the transportation of people and materials into the site.

Section 4

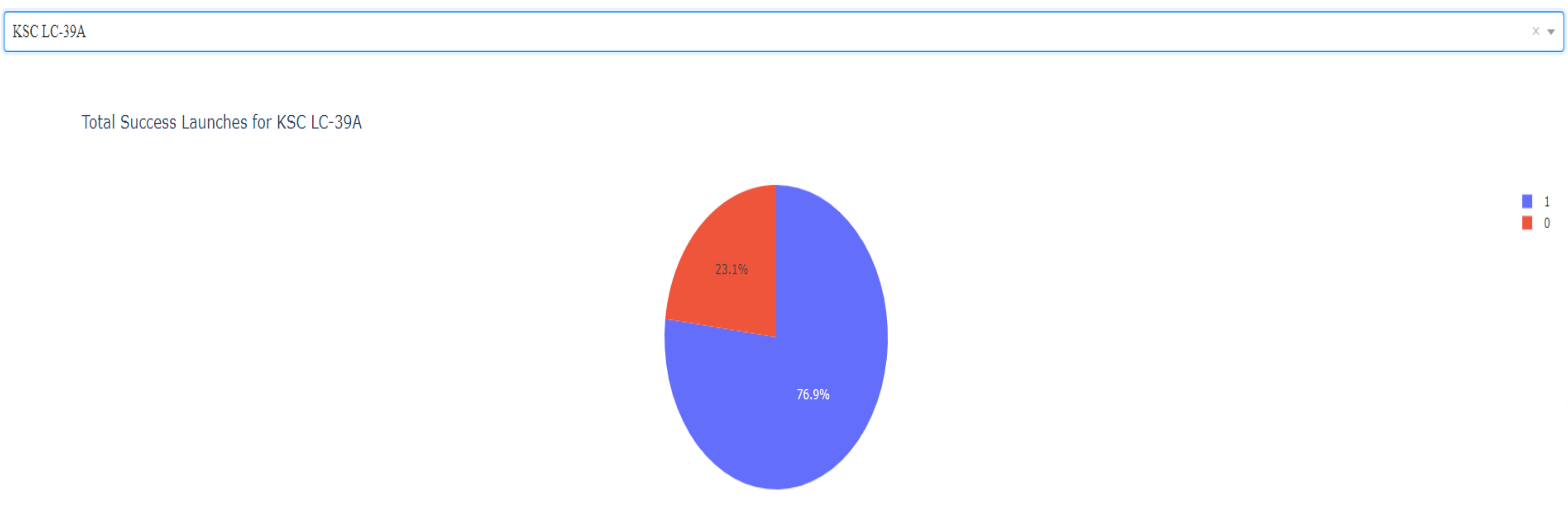
Build a Dashboard with Plotly Dash

Total success launches by site



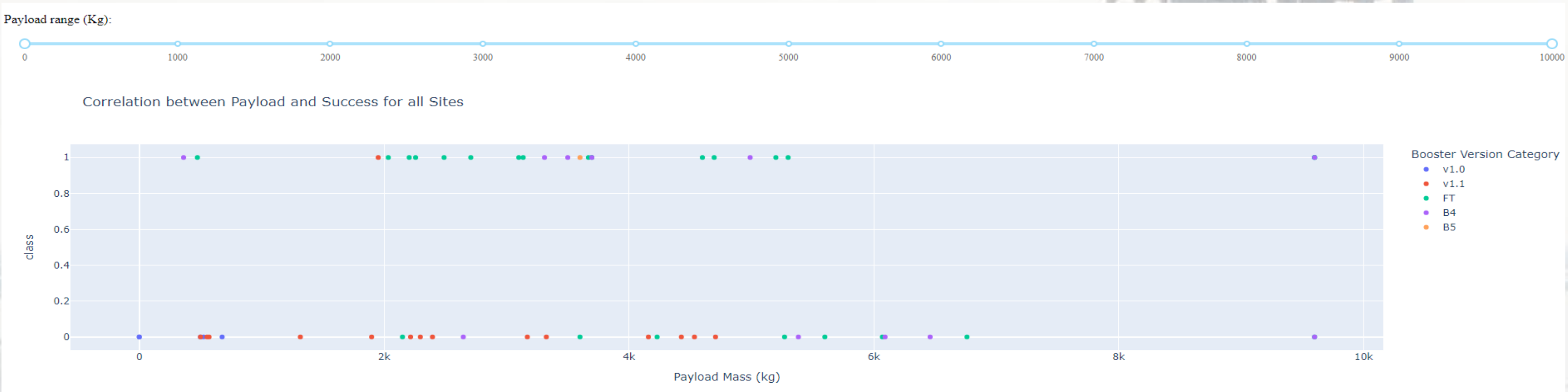
We can see from this chart that the site KSC LC-39A has the largest successful rate with 41,7% and the site CCAFS SLC-40 has the smallest successful rate with 12,5%.

Launch site with highest launch success ratio

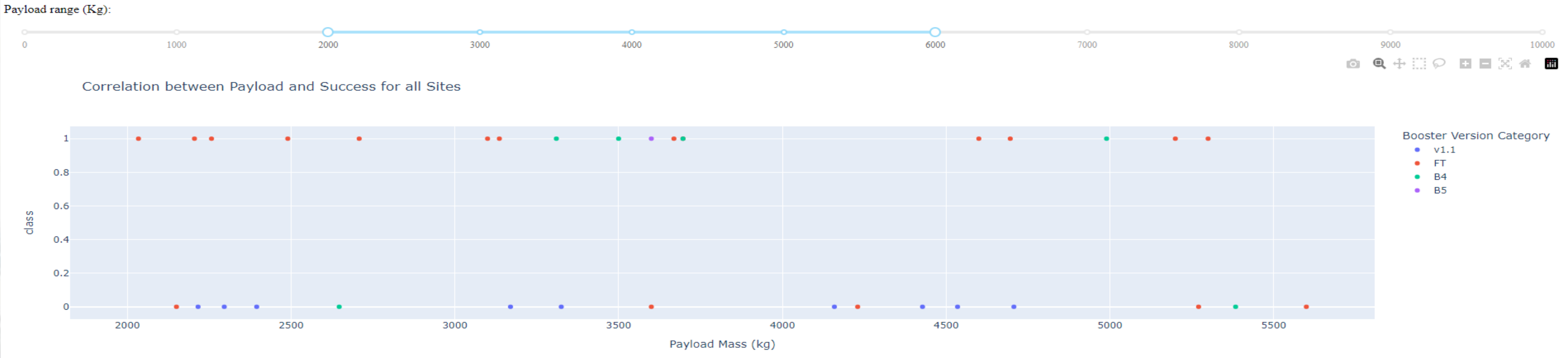


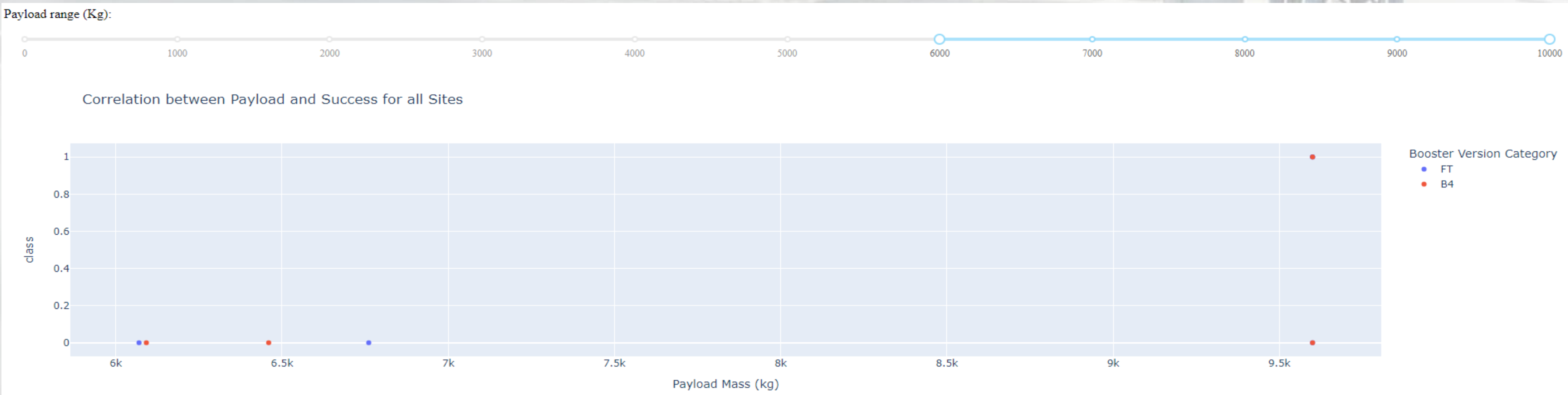
The KSC LC-39A is the launch site with the highest launch ratio, it has 76.9% of its launches which succeed (10 success vs 3 Fail). This is much better than the second highest with 42.9%.

Payload vs success for all sites



Payload vs success for all sites





Payload vs success for all sites

We can see from the previous charts:

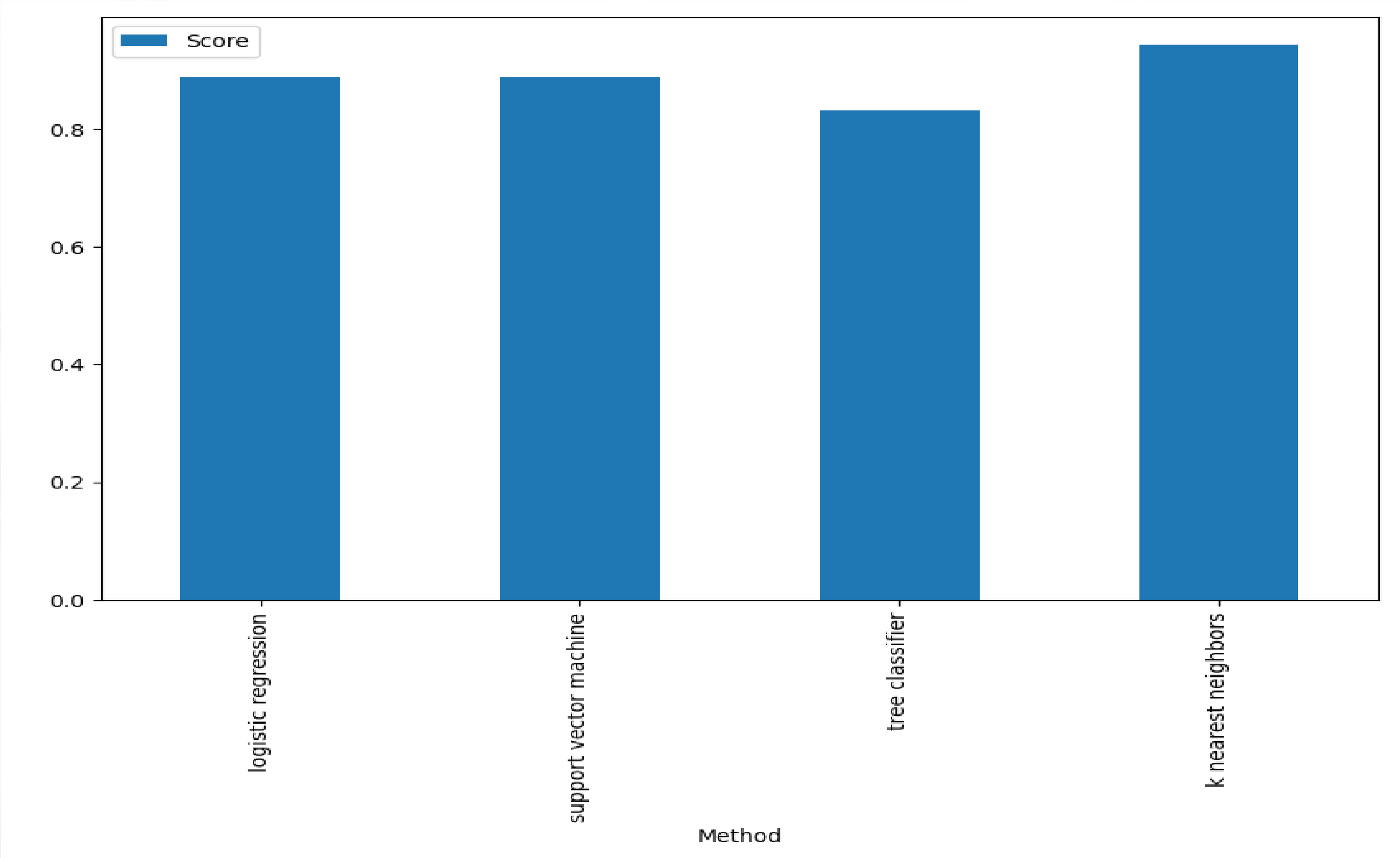
- The payload range 2,000 to 6,000kg has the highest launch successful rate whereas the payload range 0 to 2,000kg has the lowest launch successful rate.

- FT and B5 booster version have the highest launch successful rate (but B5 has only one record) whereas V1.0 and V1.1 booster version have the lowest successful rate.

Section 5

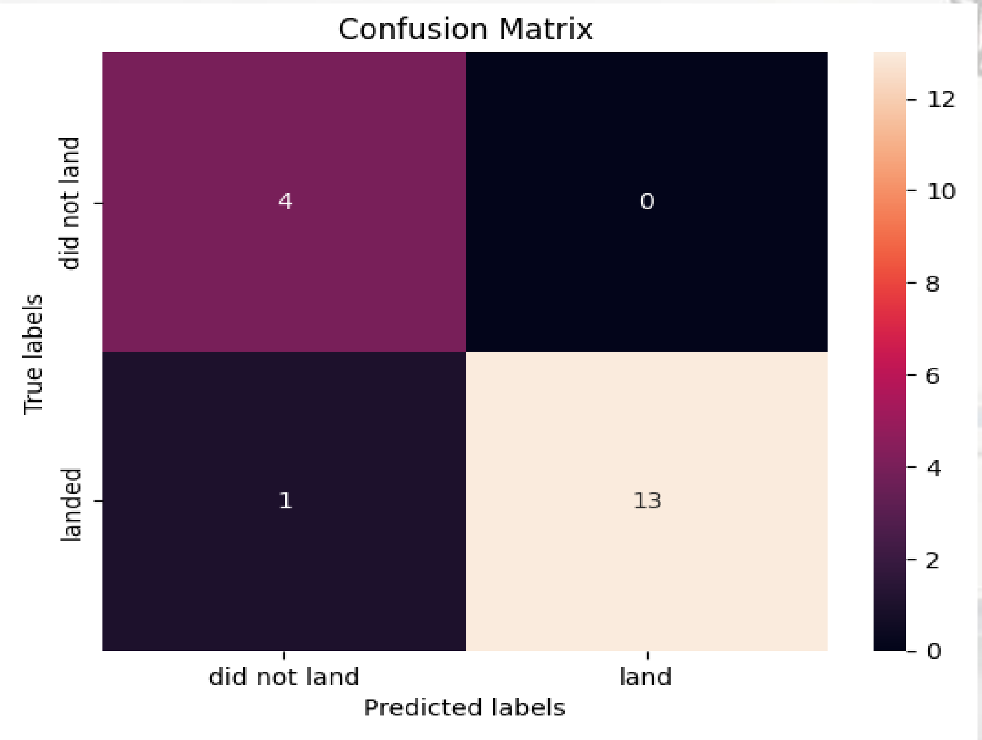
Predictive Analysis (Classification)

Classification Accuracy



- The model has the highest classification accuracy is K nearest neighbors

Confusion Matrix



- We can see from the confusion Matrix:
  - Out of 4 cases where the stage 1 did not land the classifier correctly predicted 4 of them.
  - Out of 14 cases where the stage 1 did land the classifier correctly predicted 13 of them.

Conclusions

By analyzing SpaceX's data, we were able to determine which are the attributes correlated with a successful landing:

- Payload
- Orbit Type
- Launch site and its localization.

And to build a model allowing us to predict with a good reliability 94%, if the first stage will land.

This project includes only the data from SpaceX.

We may find new attributes correlated with successful landing by analyzing competitor's data in future study.

The results of this study can be used to determine the cost of a launch and proposed the best offer to the clients.

Appendix

GitHub link with all the Notebook, Dashboard, Database and CSV files created for this project:

https://github.com/Yacine13012/Data-science-final-exam/tree/main

Thank you!