IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eng. Tania Perez Ramirez
30-04-2023

Outline

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

Executive Summary

- Summary of methodologies:
  - This report presents the results of the SpaceX Falcon 9 first stage Landing Prediction. The goal is predict if the Falcon 9 first stage will land successfully, defined as a classification problem. The data was collected from the SpaceX public API and the SpaceX Wikipedia page with Web Scraping. Exploring Data Analysis with SQL, Data Visualization, Folium Maps and Building a Dashboard with Plotly Dash.

- Summary of all results:
  - Were found the best Hyperparameters for SVM, Classification Trees and Logistic Regression with GridSearchCV. Predicting with a accuracy around the 83%.

Introduction

- Project background and context:
  - SpaceX is a private aerospace company founded by Elon Musk in 2002 with the goal of reducing the cost of space exploration and eventually colonizing Mars. One of SpaceX's most significant achievements has been the development of reusable rockets, specifically the first stage of the Falcon 9 rocket. 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 due to the fact that 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.

- Problems you want to find answers:
  - How the features such payload mass, launch site, number of flights, and others affect the success of the first stage?
  - How the correlation between features affect the outcome?
  - What is the best algorithm for predict the success or not of the first stage?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Used SpaceX REST API.
  - Used Web Scrapping from Wikipedia with BeautifulSoup()
- Perform data wrangling:
  - Dropped unnecessary columns.
  - Applied One Hot Encoding for categorical features.
- Perform exploratory data analysis (EDA) using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash.

Methodology

- Perform predictive analysis using classification models:
  - Used StandardScaler() to standardize the data.
  - Split the data into train data and test data.
  - Used GridSearchCV to find the best Hyperparameters for SVM, Classification Trees and Logistic Regression.
  - Evaluated the prediction with score function from GridSearchCV for the test data.

Data Collection

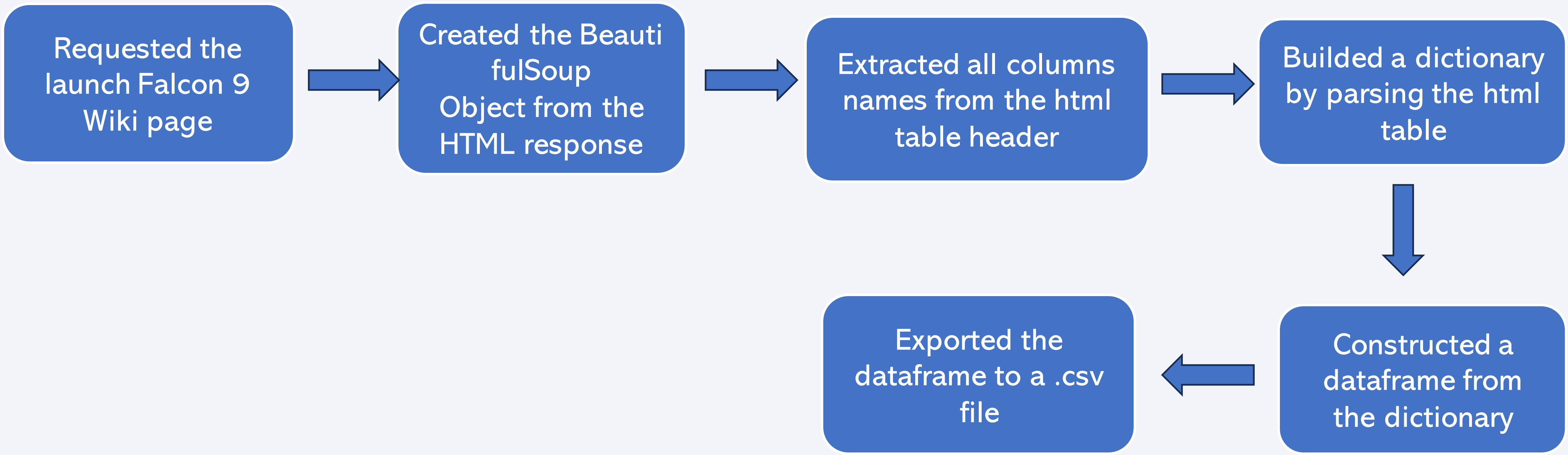
- The data was collect from 2 source: SpaceX public API and the SpaceX Wikipedia page. In order to obtain more information to have the most complete analysis possible.
- Space X API Data Columns Collected:
  - BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- SpaceX Wikipedia page Columns Collected:
  - Flight No., Date and time ( ), Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome.

Data Collection – SpaceX API

1. Requested to rocket launch data from SpaceX API
2. Decoded the response content with .json()
3. Saved into a dataframe with .json_normalize()
4. Constructed a dictionary with the data collected
5. Created a dataframe with the dictionary
6. Filtered the dataframe to only include Falcon 9 launches
7. Replaced missing values in the Column PayloadMass Using the mean()

[Data Collection API.ipynb]

Data Collection - Scraping



[Data Collection with Web Scraping.ipynb]

Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.
- Was converted those outcomes into Training Labels with "1" means the booster successfully landed, "0" means it was unsuccessful.

GitHub [Lab 2_Data wrangling.ipynb]

Calculated the number of launches on each site

↓

Calculated the number and occurrence of each Orbit

↓

Calculated the number and occurrence of mission outcome per orbit type

↓

Create a landing outcome (binary, landing success will be 1 otherwise will 0) label from Outcome column

EDA with Data Visualization

- Charts were plotted:
  - Catplot to show the distribution between a categorical variable and continuo variable: FlightNumber vs PayloadMass, FlightNumber vs LaunchSite.
  - Scatter plot to show the relationship between variables: FlightNumber vs LaunchSite, PayloadMass vs LaunchSite, FlightNumber vs Orbit, PayloadMass vs Orbit.
  - Bar plot to display and compare the frequency, count or proportion of discrete categories or groups in a dataset: Orbit vs Class.
  - Line Plot to display and visualize trends and patterns over time or across continuous data: Date vs Class.

GitHub [EDA with Data Visualization.ipynb]

EDA with SQL

- Displayed the names of the unique launch sites in the space mission.
- Displayed 5 records where launch sites begin with the string 'CCA'.
- Displayed the total payload mass carried by boosters launched by NASA (CRS).
- Displayed average payload mass carried by booster version F9 v1.1.
- Listed the date when the first successful landing outcome in ground pad was achieved.
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listed the total number of successful and failure mission outcomes.
- Listed the names of the booster_versions which have carried the maximum payload mass.
- Listed the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015.
- Ranked the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub

EDA with SQL.ipynb

Build an Interactive Map with Folium

- Marked all launch sites on a map. For answer the questions: Are all launch sites in proximity to the Equator line?, Are all launch sites in very close proximity to the coast?
- Marked the success/failed launches for each site on the map. Added color Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Calculated the distances between a launch site to its proximities. For answer the questions: Are launch sites in close proximity to railways?, Are launch sites in close proximity to highways?, Are launch sites in close proximity to coastline?, Do launch sites keep certain distance away from cities?

GitHub

[Interactive Visual Analytics with Folium lab.ipynb]

Build a Dashboard with Plotly Dash

- A Dropdown with a list to select one of the Launch Sites.
- A Pie Char to show the total success and the total failure for the launch site for the Launch Site selected in the Dropdown.
- A RangeSlider to select the range of the Payload Mass.
- A Scatter Char to show the relationship Payload Mass (kg) vs Class for the Launch Site selected in the Dropdown and between the range selected in the RangeSlider.

GitHub

SpaceX-Dash_App.py

Predictive Analysis (Classification)

1. Created a numpy array Y from the column Class in the data
2. Standarized the data with StandardScaler()
3. Splited the data into X_train, X_test, Y_train, Y_test with train_test_split()
4. Created a GridSearchCV() with cv=10 to find the best hyperparameters
5. Calculated GridSearchCV() with differents models: LogisticRegression(), SVM(), DecisionTreeClassifier() and KNeighborsClassifier()
6. Calculated the accuracy for each of the models with the data test
7. Generated the confusion matrix for all models
8. Printed the models with the max accuracy

GitHub

[Machine Learning Prediction.ipynb]

Results

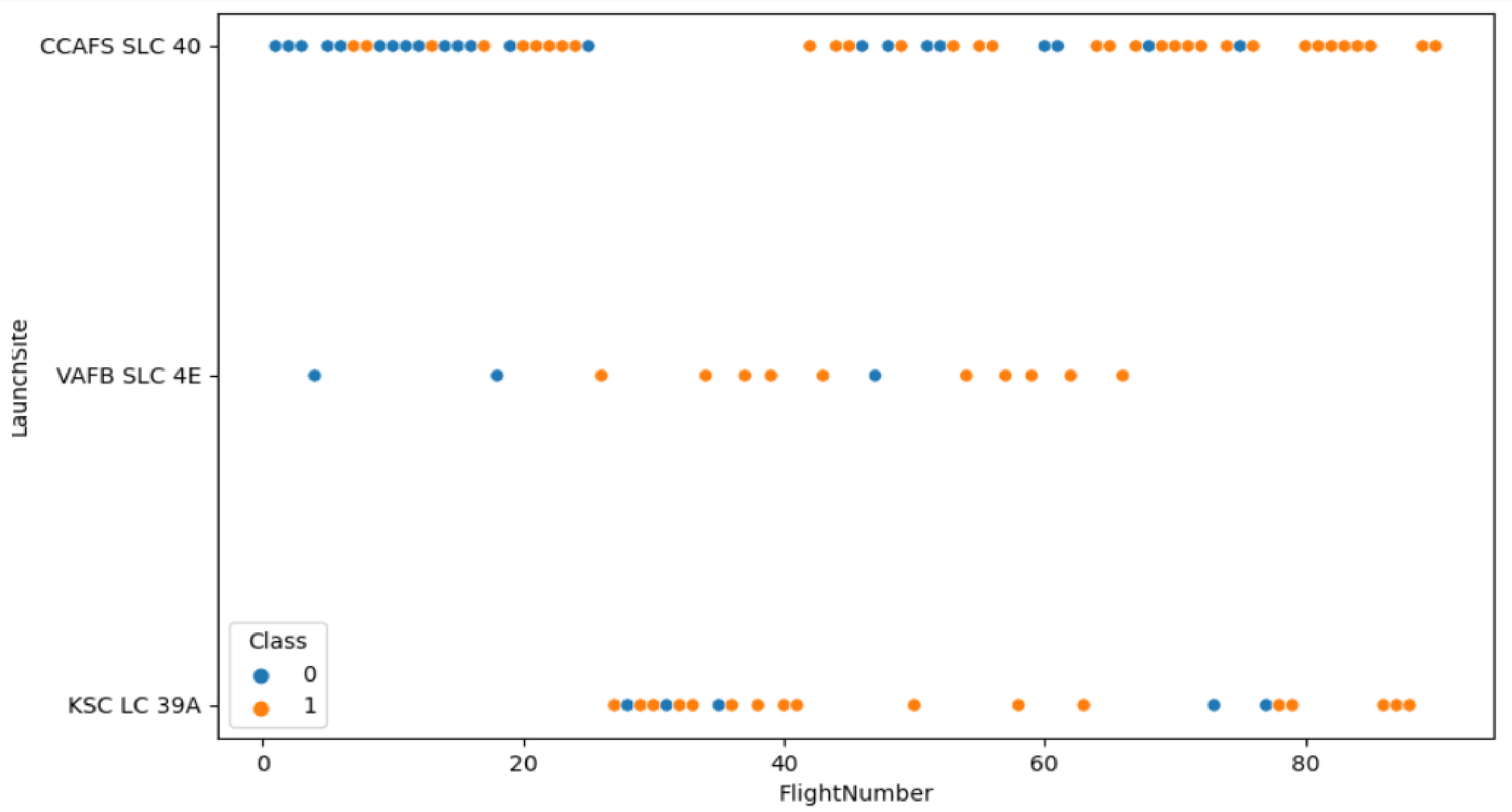
- Exploratory data analysis results.
- Interactive analytics demo in screenshots.
- Predictive analysis results.

Section 2

Insights drawn from EDA

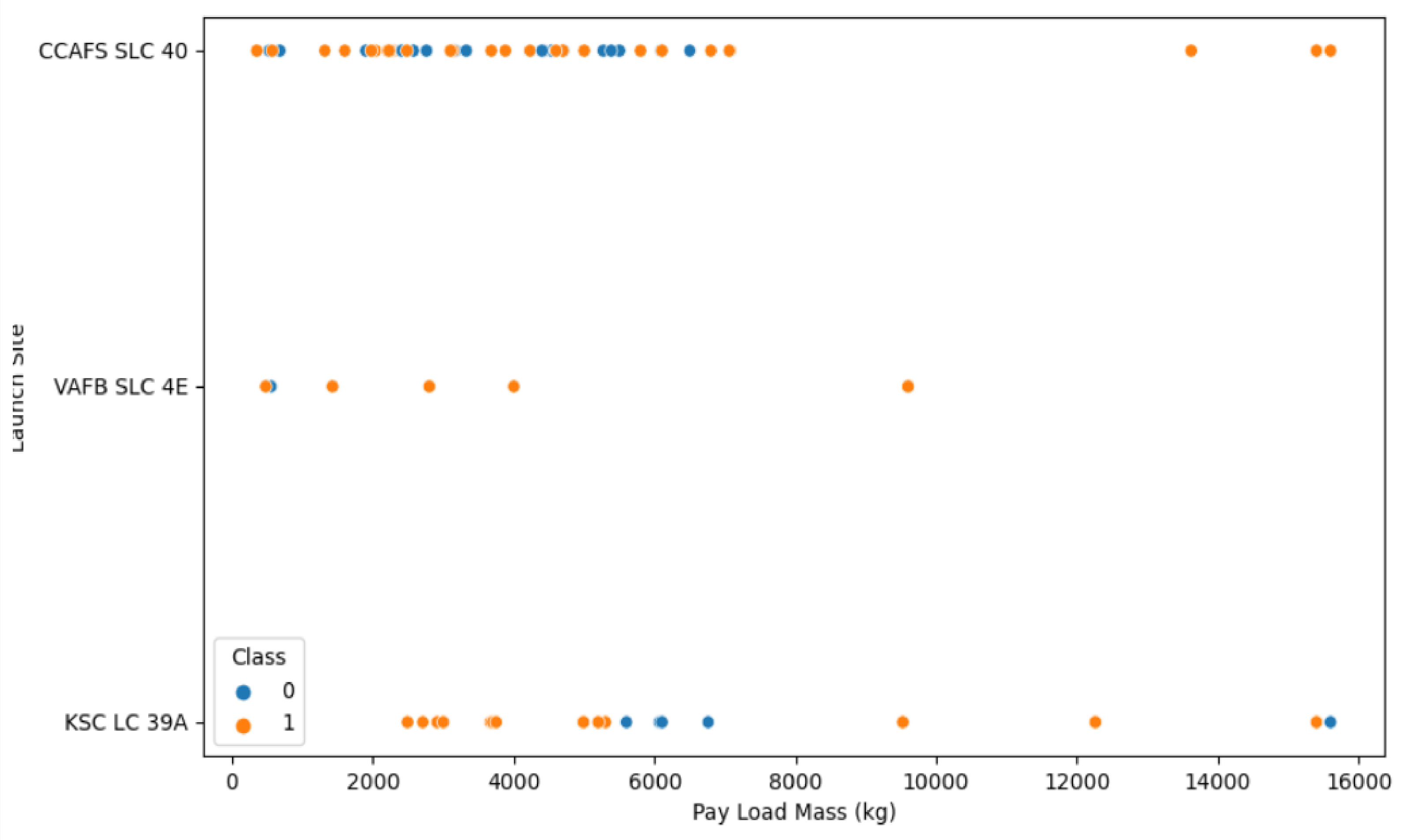
2.1 EDA with Visualization

Flight Number vs. Launch Site



The scatter char shows the relationship between LaunchSite and the FlightNumber. It reveal that when the Number of Flights increase, the success lands, also increase for all of Launch Site.

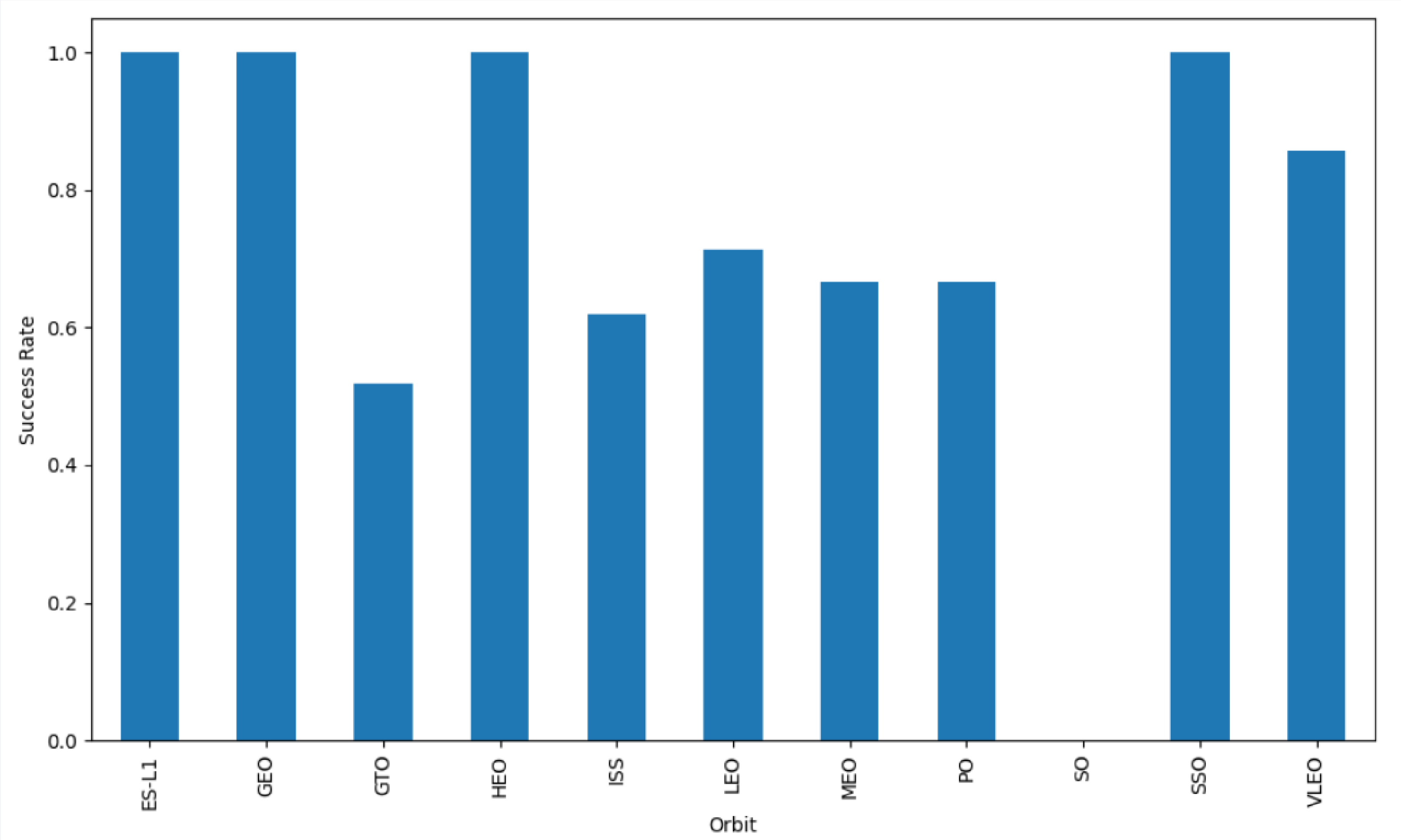
Payload vs. Launch Site



The scatter char shows the relationship between LaunchSite and Pay Load Mass (Kg). It reveal:

- For the Launch Site CCAFS SLC 40, when the Payload Mass is more than 12000 the success of the land tends to increase. However is not a complete conclusion because only 3 launch.
- For the Launch Site VAFB SLC 4E, with the Payload Mass range between 2000 and 10000 Kg was successful. However there are no rockets launched for heavy payload mass(greater than 10000).
- For the Launch Site KSC LC 39A, the more successful launch was in the Payload range between 2000 and round the 5000. After that range and round the 7000 Payload Mass the launch were failed. Round 10000 and 12000 were successful. However round the 16000 the Payload Mass there aren't different.

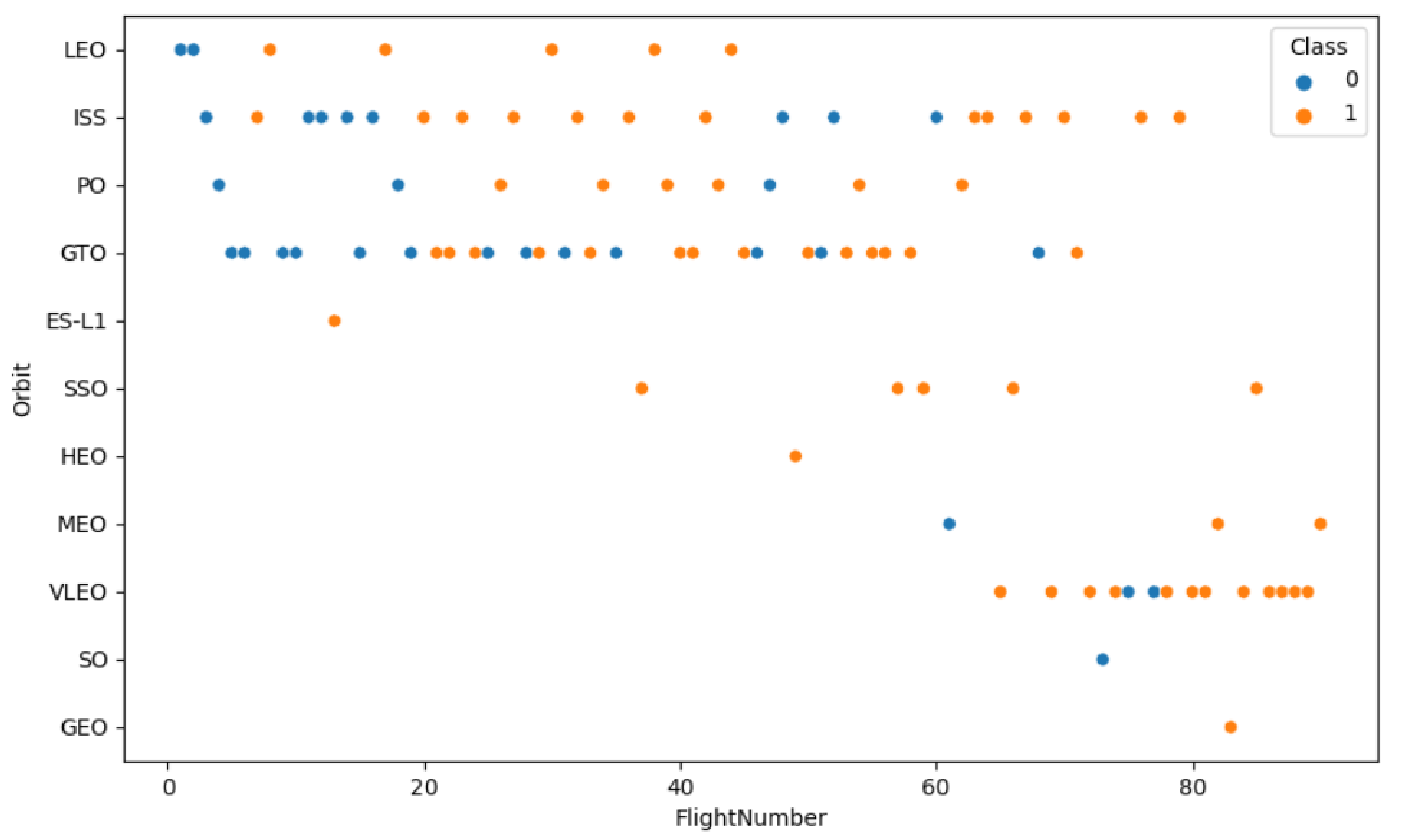
Success Rate vs. Orbit Type



The bar char shows the relationship between Success Rate and Orbit. It reveal:

- The Orbit ES-L1, GEO, HEO, SSO have 100% of success rate.
- The Orbit SO has 0% of success rate.
- The other Orbits have a success rate between 50% and 80%.

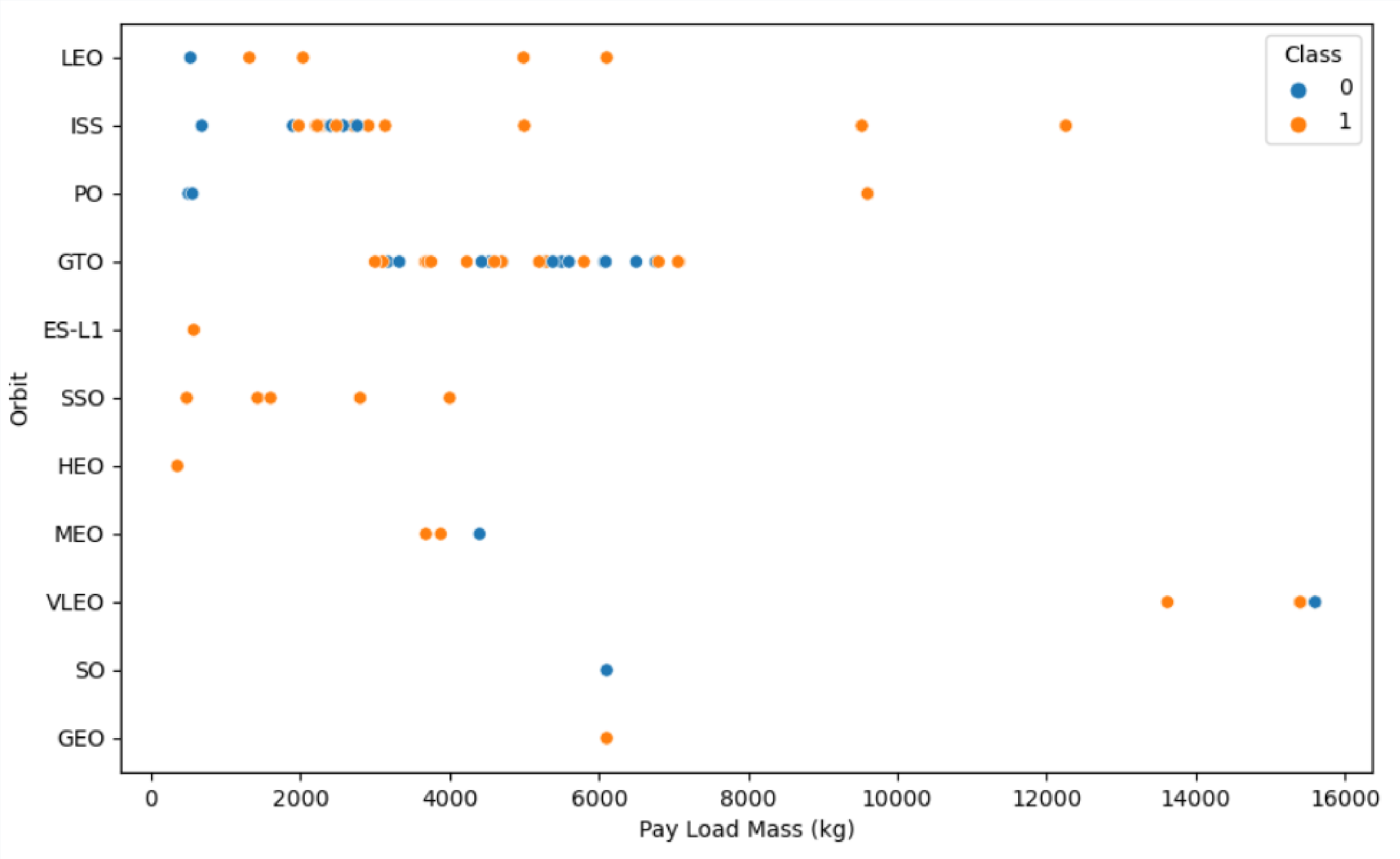
Flight Number vs. Orbit Type



The scatter char shows the relationship between Orbit and the FlightNumber. It reveal:

- For the Orbits LEO, ISS, VLEO while the Flight Number increase, the launch will be successful.
- For the Orbits ISS, PO, the range of Flight Number between 20 and 40 the launch was successful.

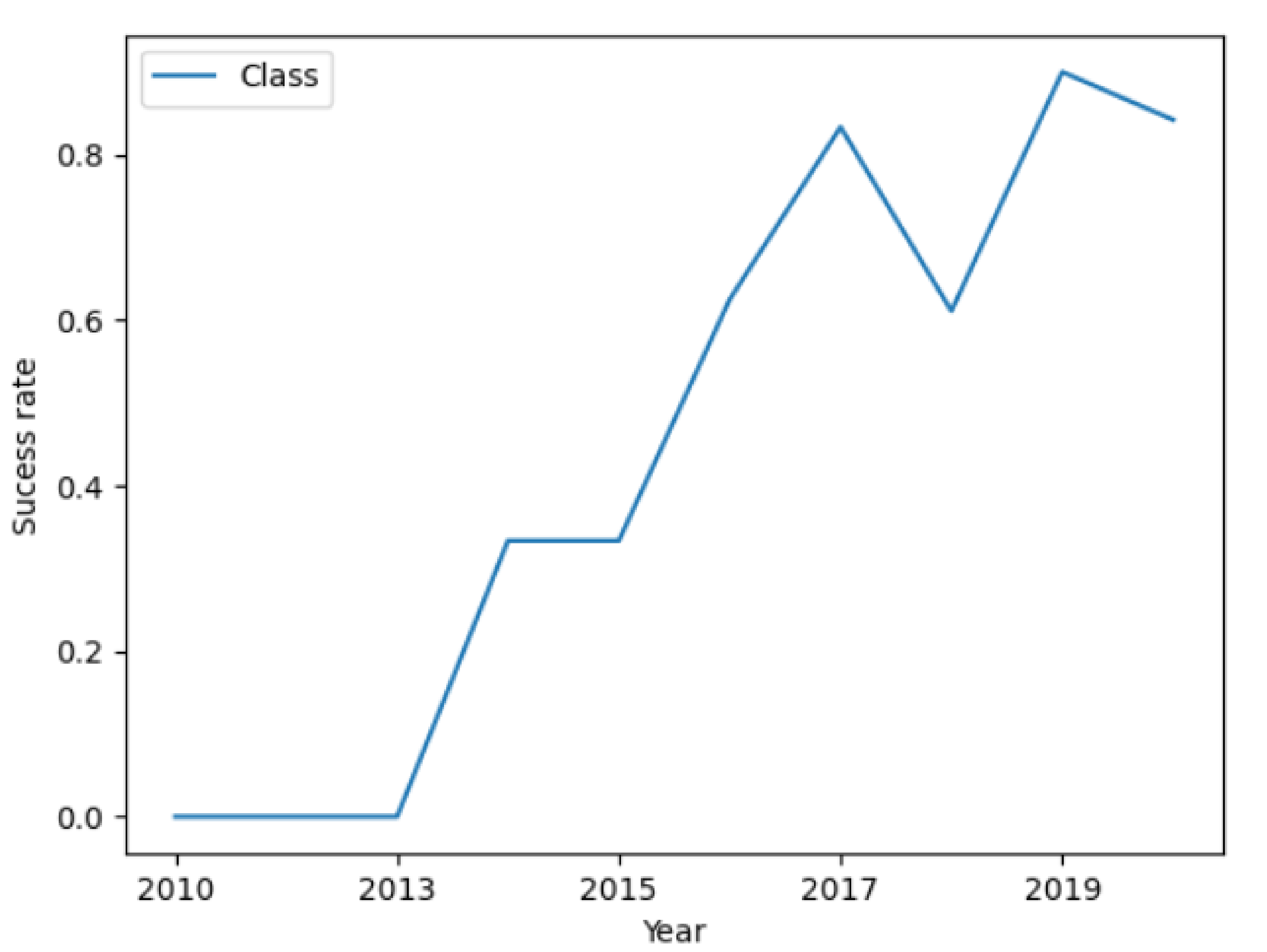
Payload vs. Orbit Type



The scatter char shows the relationship between Orbit and the Payload Mass. It reveal:

- For the Orbits LEO, ISS, PO while the Payload Mass increase, the launch will be successful.
- For the Orbit GTO, the 2 launch with more Payload have been successful. However we cannot infer that with more Payload the launch for this Orbit will be successful. Because in the other launch while the Payload increase the launch has been successful and others launch has not.
- For the Orbits ES-L1, SSO, and HEO not matter the Payload the launch has been successful.
- For the Orbit MEO while the Payload increase the successful of the launch decrease.

Launch Success Yearly Trend



The line char shows the relationship between Success Rate and the Years. It reveal that the success rate since 2013 kept increasing till 2020.

Section 2

Insights drawn from EDA

2.2 EDA with SQL

All Launch Site Names

```
%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

It displays the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

It displays 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
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

It displays the total payload mass carried by boosters launched by NASA (CRS).

```
%sql select SUM(PAYLOAD_MASS__KG_) as 'Total Payload Mass launched by NASA (CRS)' from SPACEXTBL where Customer = 'NASA (CRS)'
```

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

Total Payload Mass launched by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

It displays average payload mass carried by booster version F9 v1.1.

```
%sql select AVG(PAYLOAD_MASS__KG_) as 'Average Payload Mass by booster version F9 v1.1' from SPACEXTBL where Booster_Version like 'F9 v1.1%'
```

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

Average Payload Mass by booster version F9 v1.1
2534.6666666666665

First Successful Ground Landing Date

It lists the date when the first successful landing outcome on ground pad was achieved.

```
%sql select min(Date) as 'First Successful Landing on ground pad' from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)'
```

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

First Successful Landing on ground pad
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

It lists 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 ("Landing _Outcome" = 'Success (drone ship)') and (PAYLOAD_MASS__KG_ between 4000 and 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

It lists the total number of successful and failure mission outcomes.

```
%%sql select T.Total as 'Total of Mission Outcome', Sum(S.Success) as 'Success Mission', Sum(F.Failure) as 'Failure Mission' from
    (select count(Mission_Outcome) as Success from SPACEXTBL where Mission_Outcome like 'Success%' ) as S,
    (select count(Mission_Outcome) as Failure from SPACEXTBL where Mission_Outcome like 'Failure%' ) as F,
    (select Count(Mission_Outcome) as Total from SPACEXTBL) as T
```

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

Total of Mission Outcome	Success Mission	Failure Mission
101	100	1

Boosters Carried Maximum Payload

```
%sql select DISTINCT Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL )
```

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

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

It lists the names of the booster_versions which have carried the maximum payload mass.

2015 Launch Records

```
%%sql select CASE substr(Date, 4, 2)
    WHEN '01' THEN 'January'
    WHEN '02' THEN 'February'
    WHEN '03' THEN 'March'
    WHEN '04' THEN 'April'
    WHEN '05' THEN 'May'
    WHEN '06' THEN 'June'
    WHEN '07' THEN 'July'
    WHEN '08' THEN 'August'
    WHEN '09' THEN 'September'
    WHEN '10' THEN 'October'
    WHEN '11' THEN 'November'
    WHEN '12' THEN 'December'
    END AS Month_name,
  "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_name	Landing _Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

It lists 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 Landing Outcomes Between 2010-06-04 and 2017-03-20

```
%%sql select "Landing _Outcome", count("Landing _Outcome") as 'Count of Successful Landing' from SPACEXTBL
      where substr(Date,7)||substr(Date,4,2)||substr(Date,1,2) between '20100604' and '20170320'
      group by "Landing _Outcome"
      order by count("Landing _Outcome") DESC
```

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

Landing _Outcome	Count of Successful Landing
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

It lists a 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.

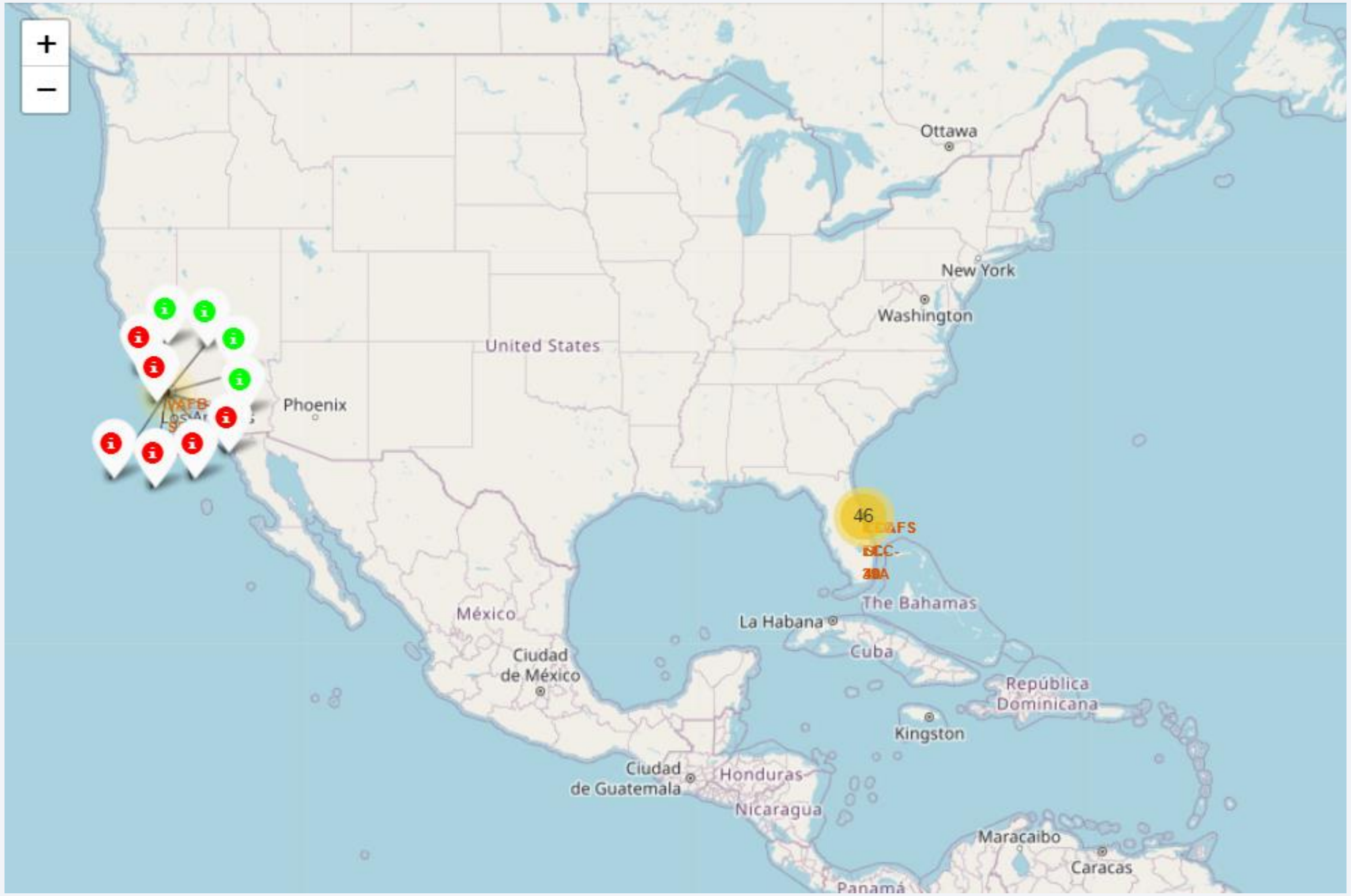
Section 3

Launch Sites Proximities Analysis

Launch Sites Locations

It shows the markers of the all Launch Site locations. There are very close between each other. And close to the coast, it is convenient for security problems.

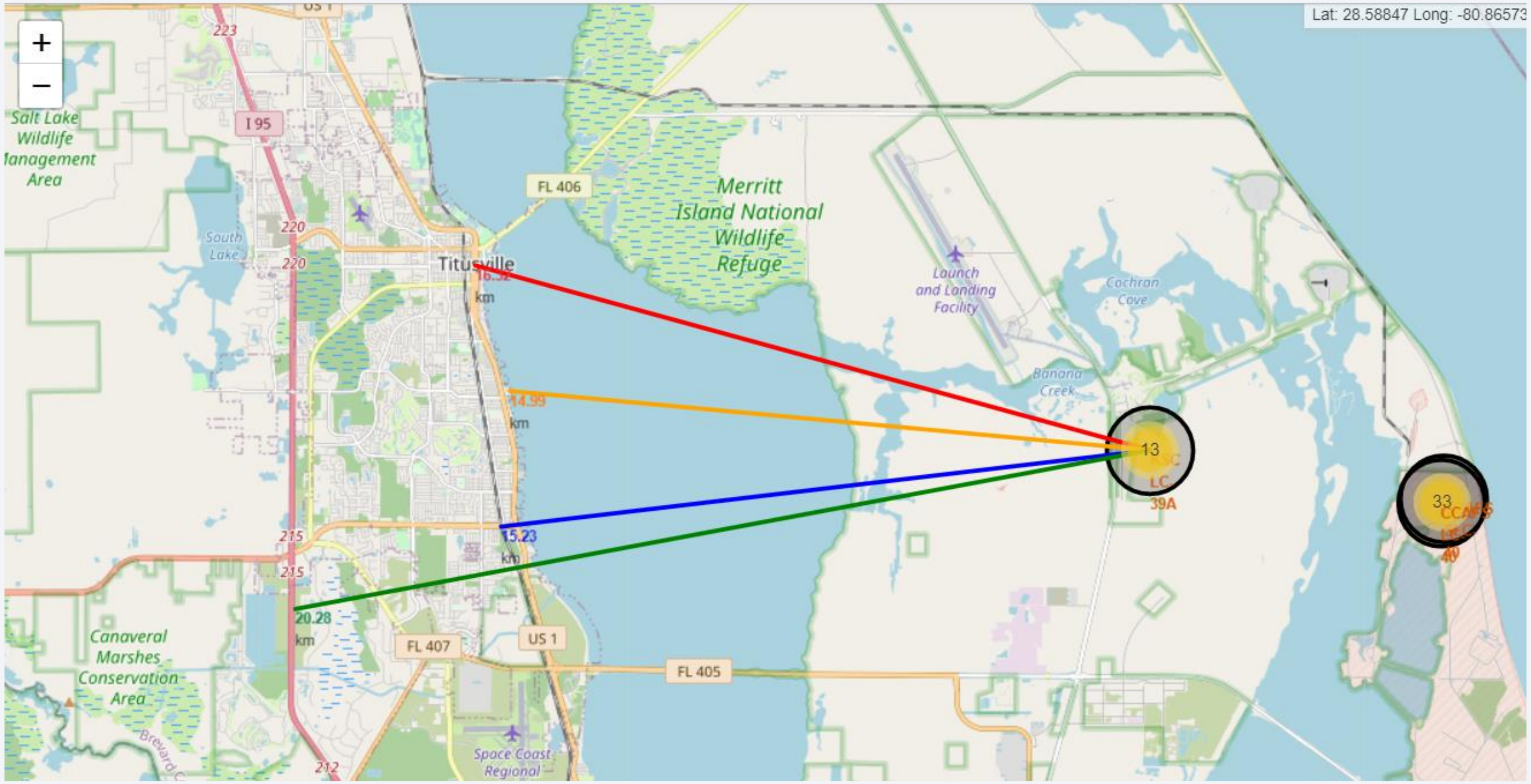
Launches markets with colors and grouped by Launch site



It shows the markers of the all Launches classify by colors with red will the not successful and the green with the successful.

From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

Distances between KSC LC-39A and its proximities

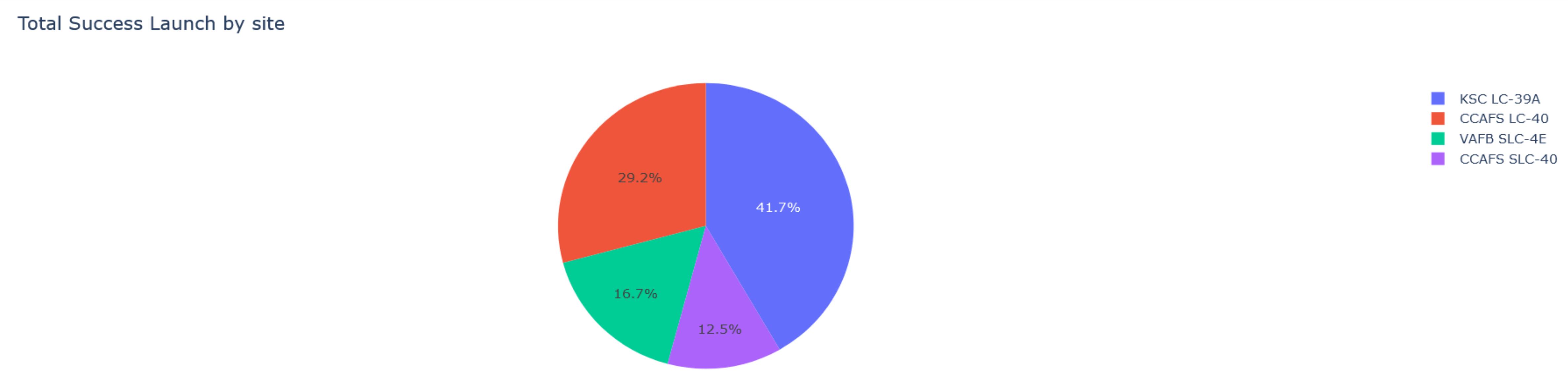


It shows the in close proximity to railways, coastline, highways, coastline and cities

Section 4

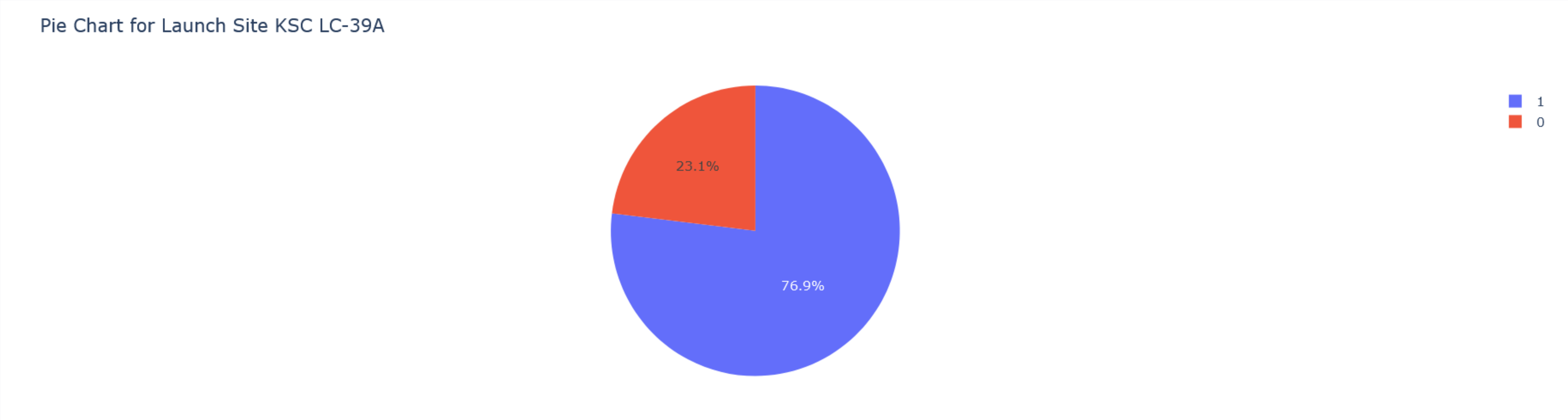
Build a Dashboard with Plotly Dash

Total Success Launch by site



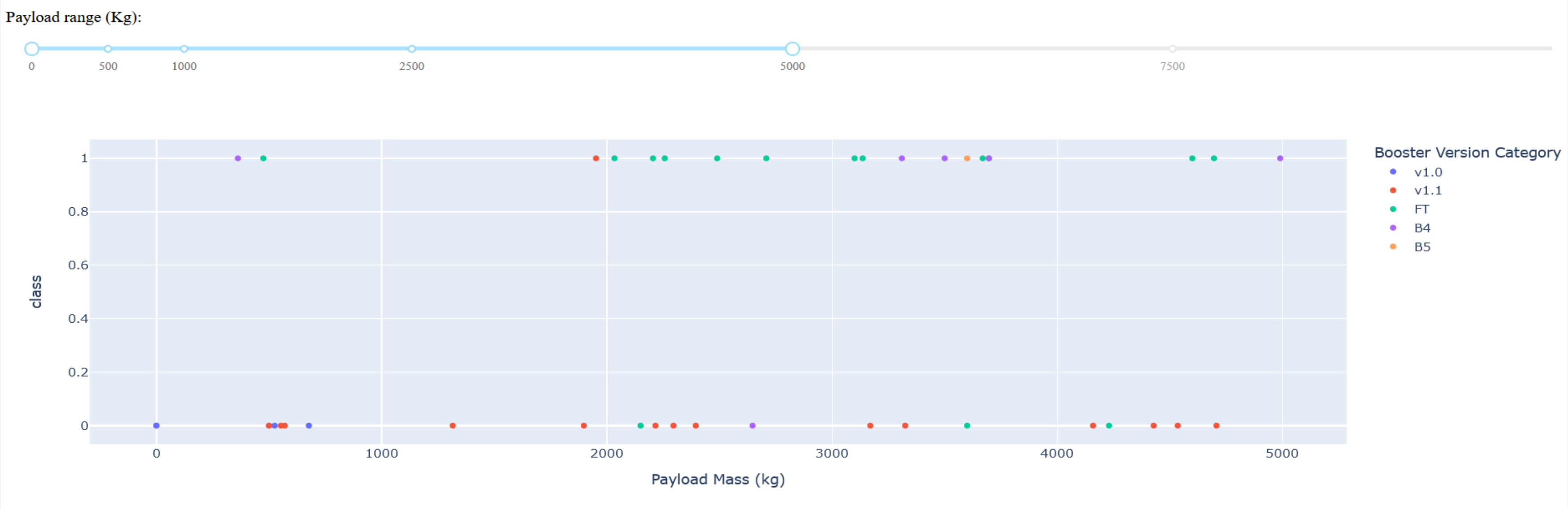
It shows a Pie Chart for the Total Success by Launch Site. The site KSC LC-39A has more successful launches.

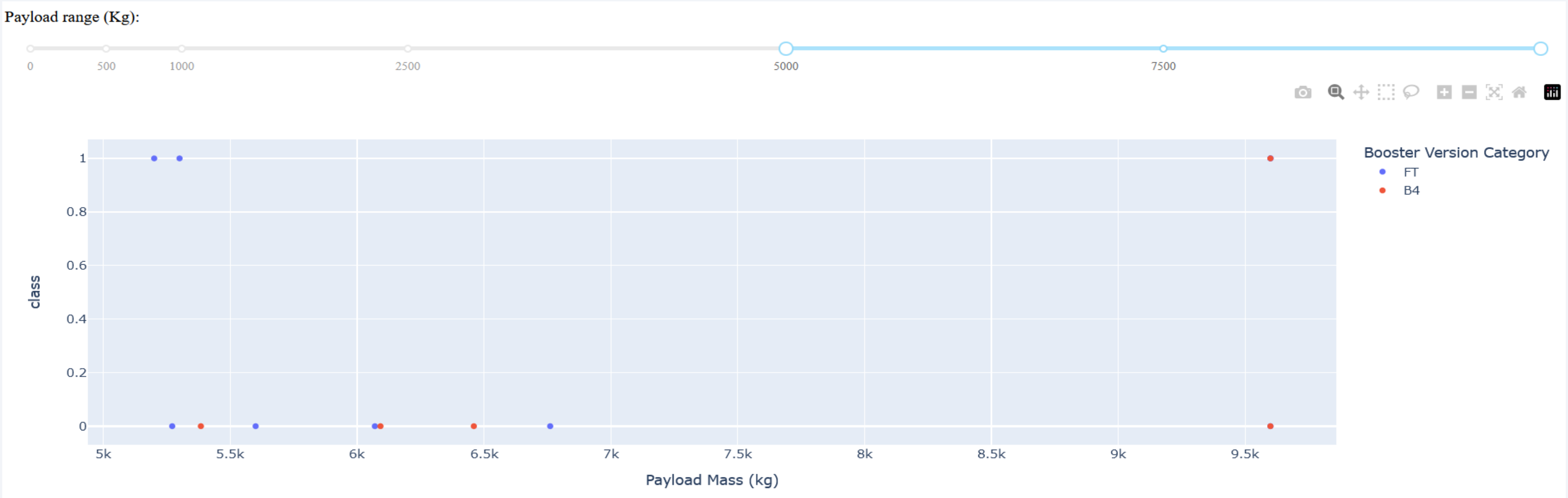
<Dashboard Screenshot 2>



It shows a Pie Chart for the Launch Site KSC LC-39A with more launch success

Payload Mass vs Launch Outcomes for all Launch Sites



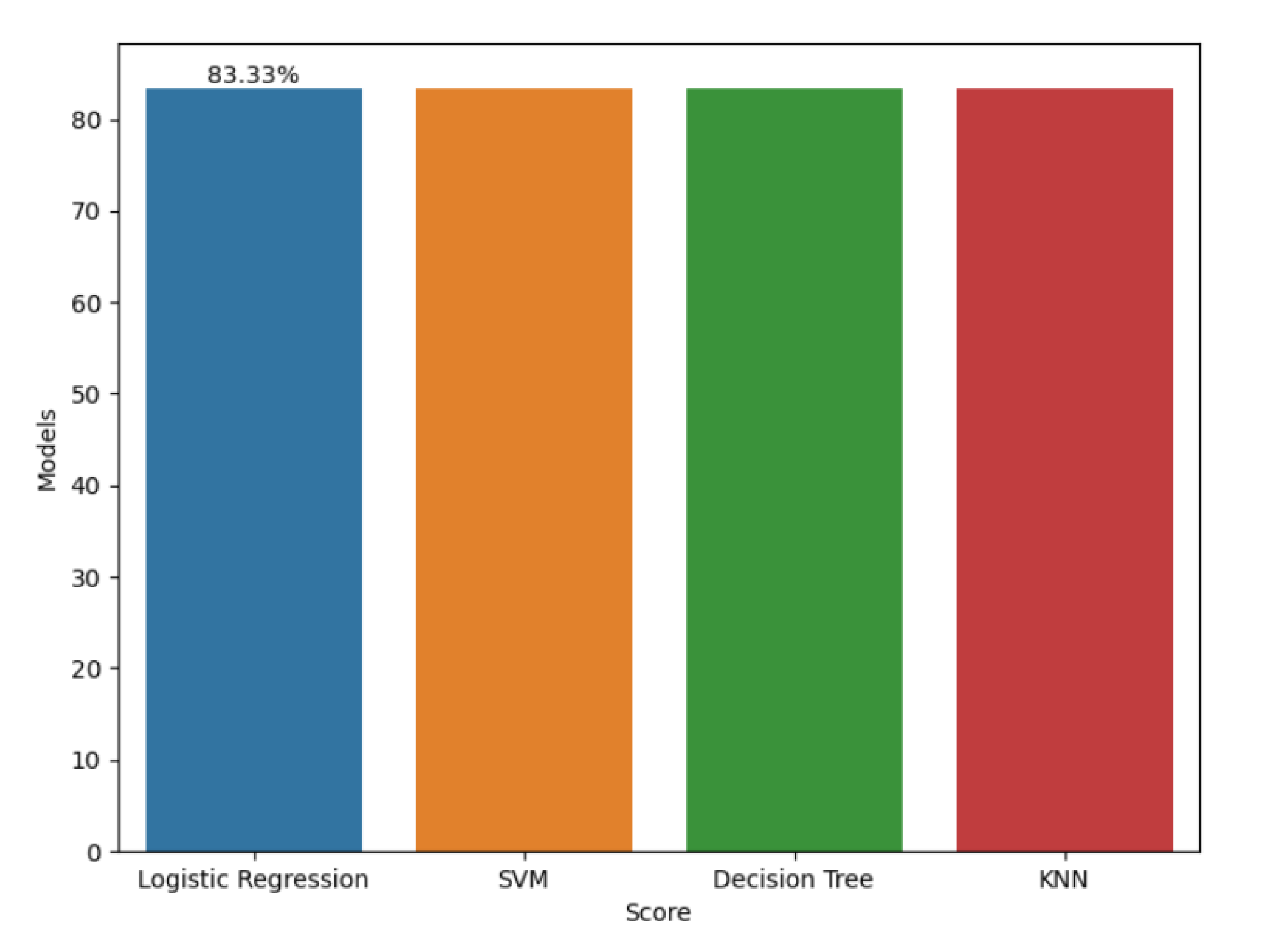


It shows a two Scatter Plot Payload Mass vs Class with different range of Payload. It displays for the figure 1 a more launch success between the range 0-5000 Kg of Payload Mass than the range 5000-9600 Kg.

Section 5

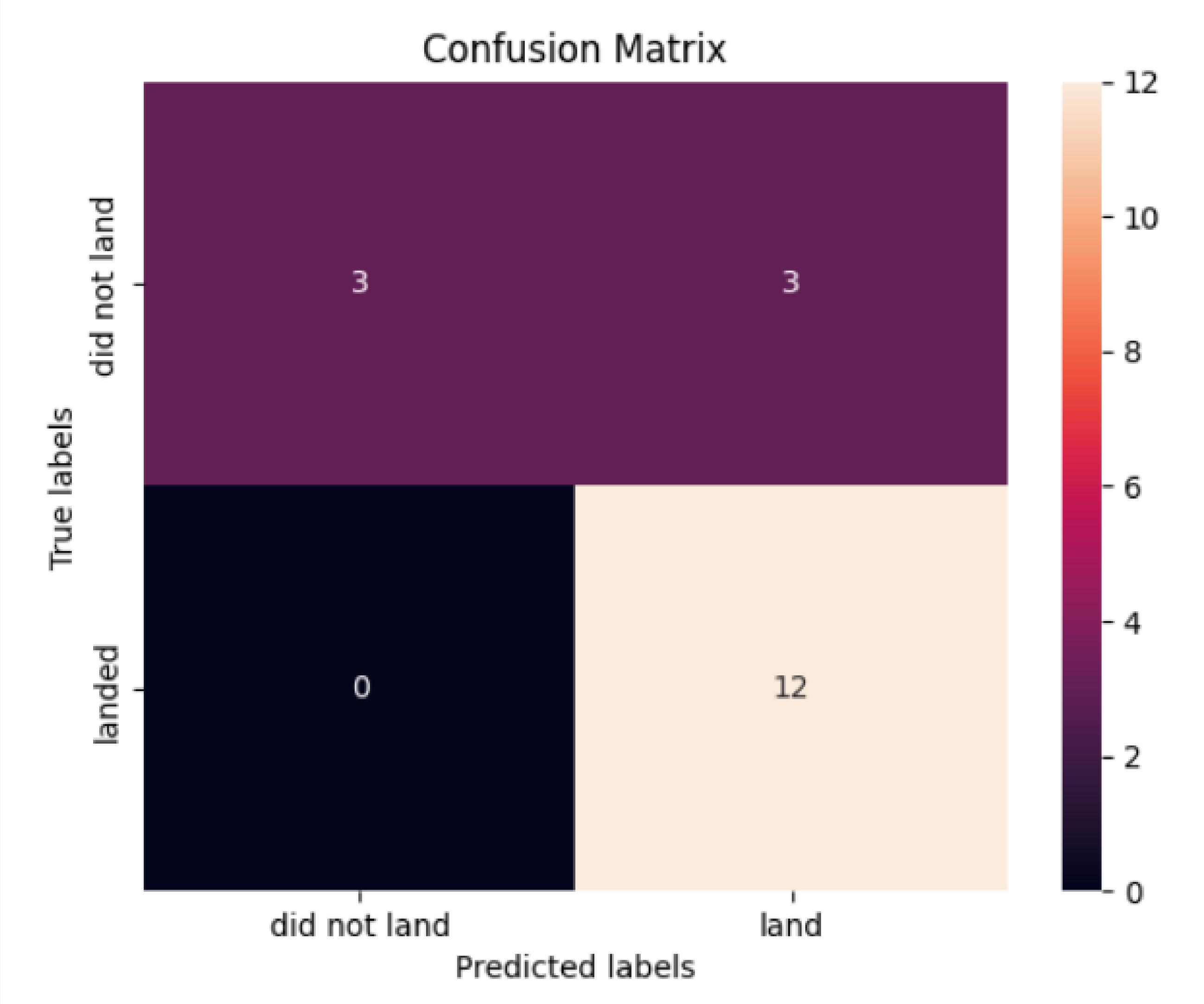
Predictive Analysis (Classification)

Classification Accuracy



It shows that all the models have the same value. With a score round of the 83.33%.

Confusion Matrix



The 4 Models have the same Confusion Matrix, because all have the same accuracy of 83.33%.

It reveals:

- The data is not balanced with 15 landed and only 3 did not landed.
- The data test is too small with only 18 samples.

Conclusions

- When the Number of Flights increase the success land increase.
- The success rate since 2013 kept increasing till 2020.
- The all Launch Site locations are very close between each other. And close to the coast.
- The site KSC LC-39A has more successful launches with a 76.9% of success.
- The Orbits ES-L1, GEO, HEO, SSO have 100% of success rate. However The Orbit SO has 0% of success rate.
- There are more launches success between the range 0-5000 Kg of Payload Mass than the range 5000-9600 Kg.
- The all machine learning models have a high accuracy with a 83.33%. However The data is not balanced with 15 landed and only 3 did not landed. The data test is too small with only 18 samples.

Recommendation

- Increase the number of samples for all models and run again for get the best model to predict a more reality result.

Appendix

- Source of the all code:

GitHub Applied Data Science Capstone

- IBM Data Science Professional Certificate

Thank you!

- Special Thanks:
  - Instructors
  - IBM
  - Coursera