Data Science Capstone Project - SpaceX

Nalli Ajay Kumar 07 July 2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Conclusion
- Appendix

EXECUTIVE SUMMARY

➢ **Summary of methodologies**

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

➢ **Summary of all results**

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

INTRODUCTION

background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon9 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage

Problem Statement : The objective is to evaluate the viability of the new company Space Y to compete with Space X.

Questions to be answered

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

METHODOLOGY

- **Data collection methodology:**
  - Using Data from SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- **Performed data wrangling**
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- **Performed exploratory data analysis (EDA) using visualization and SQL**
- **Performed interactive visual analytics using Folium and Plotly Dash**
- **Performed predictive analysis using classification models**
  - Building, tuning the model using GridSearchCV and evaluation of classification models to ensure the best results

Data collection

Data collection process involved a combination of API requests from SpaceX RESTAPI and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

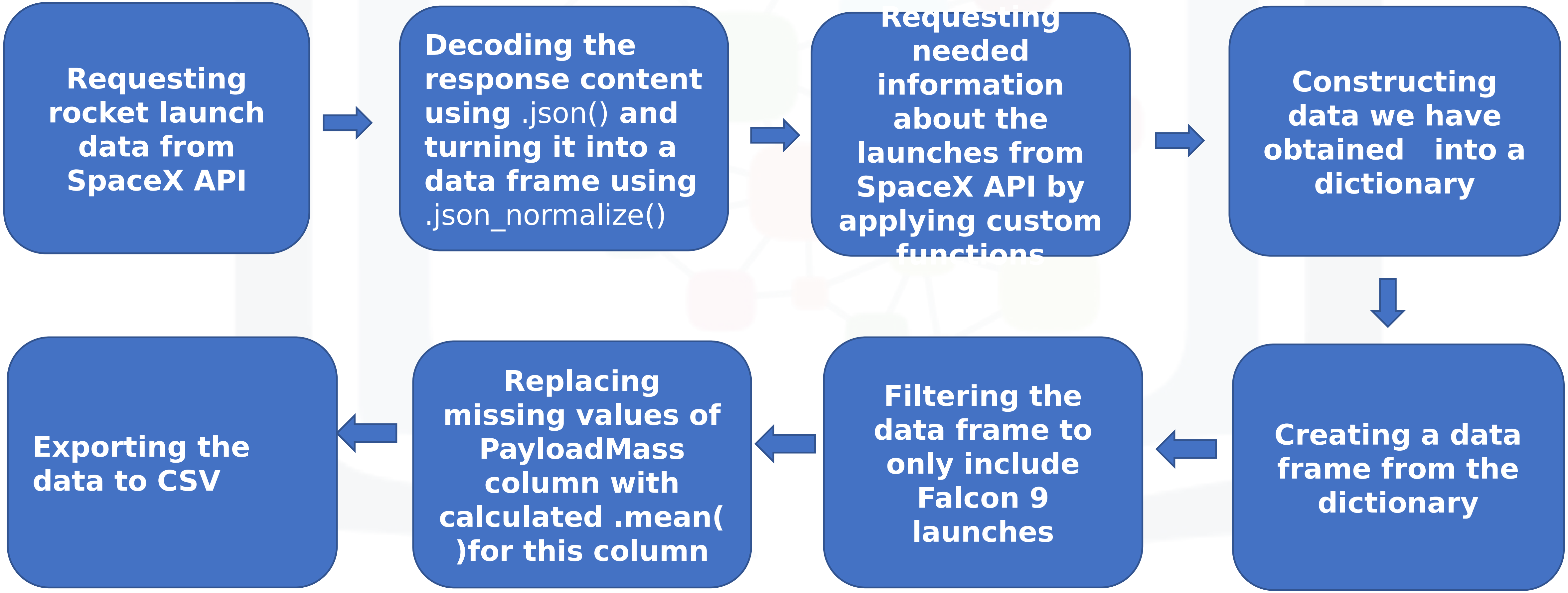
Data Columns are obtained by using Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, LaunchOutcome, Version Booster, Booster landing, Date, Time

Data collection – SpaceX API

Flow chart of data collection from API

Github url : click here



Requesting rocket launch data from SpaceX API

Decoding the response content using .json() and turning it into a data frame using .json_normalize()

Requesting needed information about the launches from SpaceX API by applying custom functions

Constructing data we have obtained into a dictionary

Creating a data frame from the dictionary

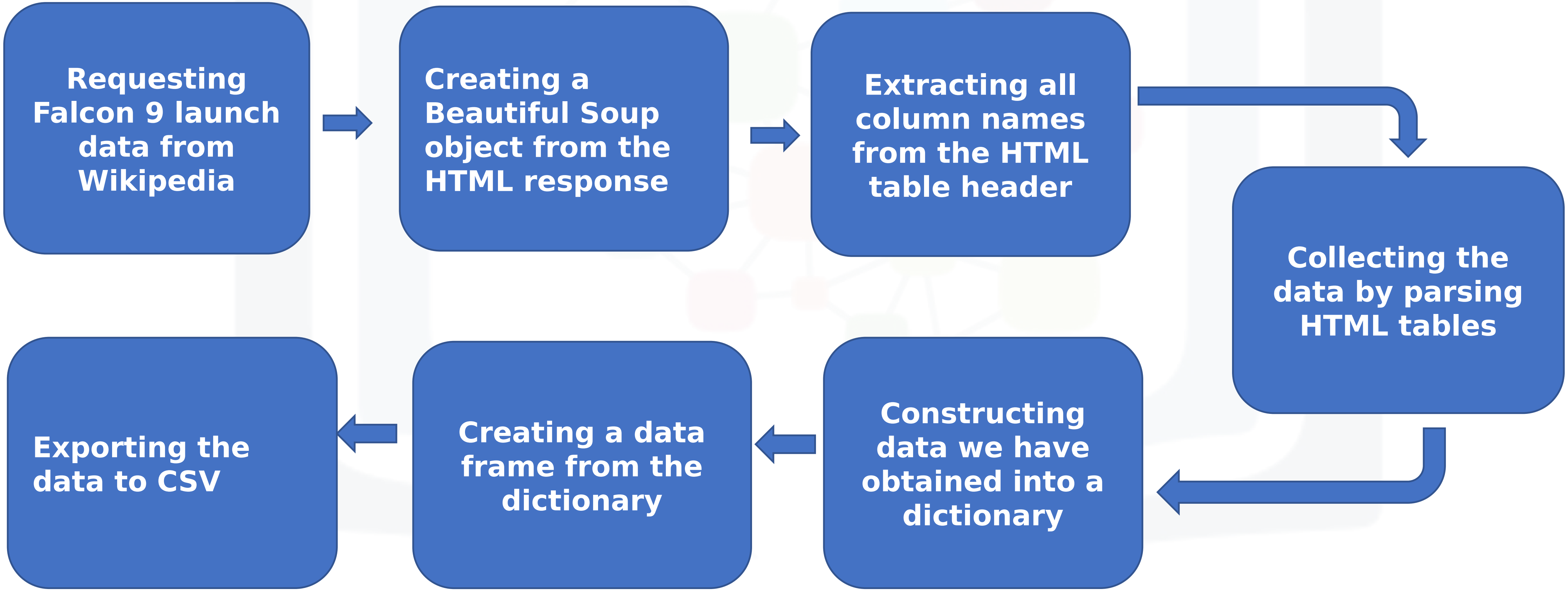
Filtering the data frame to only include Falcon 9 launches

Replacing missing values of PayloadMass column with calculated .mean( )for this column

Exporting the data to CSV

Data collection – Web scraping

Flow chart of data collection from web scraping   Github url : click here



Requesting Falcon 9 launch data from Wikipedia

Creating a Beautiful Soup object from the HTML response

Extracting all column names from the HTML table header

Collecting the data by parsing HTML tables

Constructing data we have obtained into a dictionary

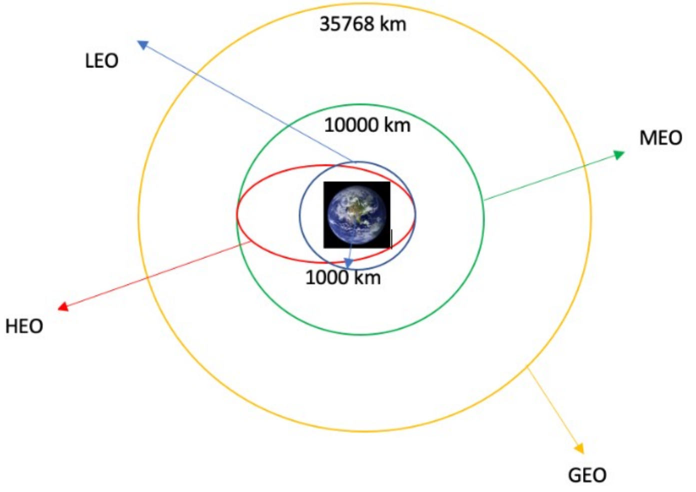
Creating a data frame from the dictionary

Exporting the data to CSV

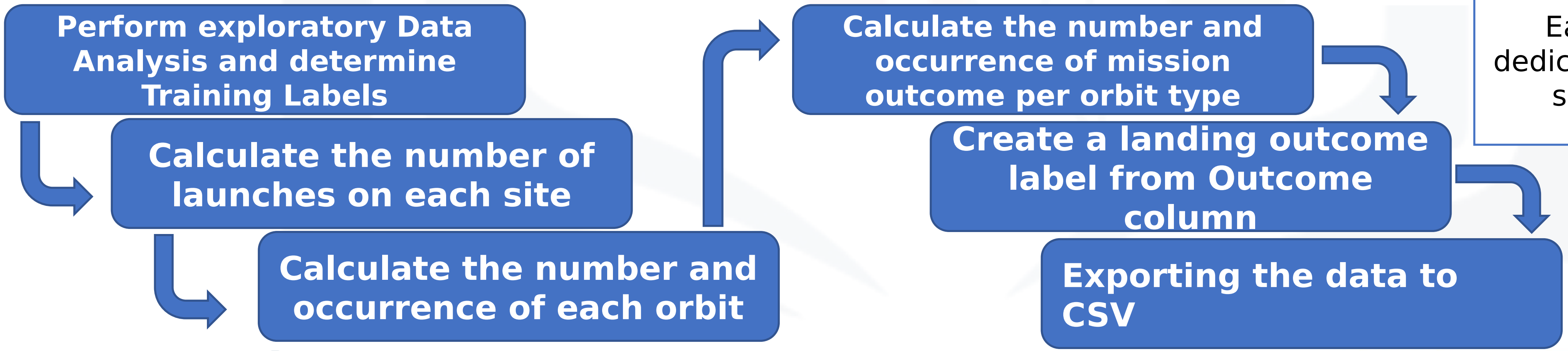
Data wrangling

jupyter notebook: click here

- 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.
- We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful



Each launch aims to a dedicated orbit, and here are some common orbits



Perform exploratory Data Analysis and determine Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting the data to CSV

EDA with data visualization

Scatter Graphs being drawn :

Flight Number VS Payload Mass

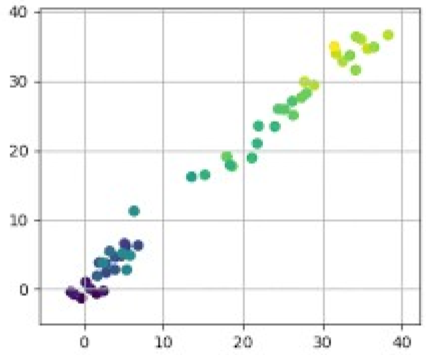
Flight Number VS launch Site

Payload VS Launch Site

Orbit VS Flight Number

Orbit VS Orbit type

Orbit VS Payload Mass



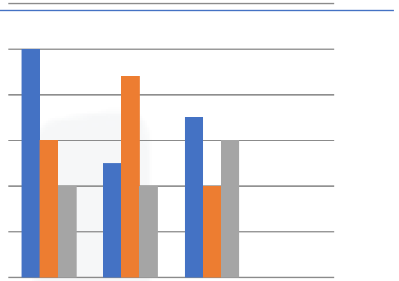
Scatter plot shows how much one variable is effected another variable.

The relation between two variables is called their correlation.

EDA with data visualization note book : click here

Bar Graph being drawn :

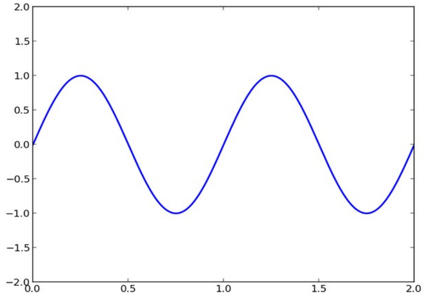
Mean VS Orbit



A Bar graph makes it easy to compare sets of data between different groups at a glance. The graph represents categories on axis and discrete values in another

Line Graph being drawn :

Success Rate VS Year



A line graphs are useful in that they shows data variables and trends very clearly and can help to make predictions about the results of the data not yet recorded

EDA with SQL

jupyter Notebook : click here

Performed SQL queries :

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA‘
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date2010-06-04 and 2017-03-20 in descending order

Build an interactive map with Folium

- **Markers of all Launch Sites:**
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

- **Coloured Markers of the launch outcomes for each Launch Site:**

  -Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

- **Distances between a Launch Site to its proximities:**
  - Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

  - jupyter NoteBook : click here

Build a Dashboard with Plotly Dash

- **Launch Sites Dropdown List:**
  - - Added a dropdown list to enable Launch Site selection.

- **Pie Chart showing Success Launches (All Sites/Certain Site):**
  - - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

- **Slider of Payload Mass Range:**
  - - Added a slider to select Payload range.

- **Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:**
  - - Added a scatter chart to show the correlation between Payload and Launch Success.

Plotly Dash : click here

Predictive analysis (Classification)

Building Model :

- Load our data in to numpy and pandas
- Transform data
- Split our data in training and test dataset
- Check how many test samples we have
- Decide which type of machine learning algorithm we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our data into GridSearchCV objects and train our data

Evaluating The Model :

- Check accuracy for each model
- Get tuned hyperparameters for each type of algorithm
- Plot confusion matrix

jupyter NoteBook : click here

Improving Model :

- Feature engineering
- Algorithm tuning

Finding The Best Performing Classification Model :

- The model with best Accuracy score wins the best performing model
- In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook

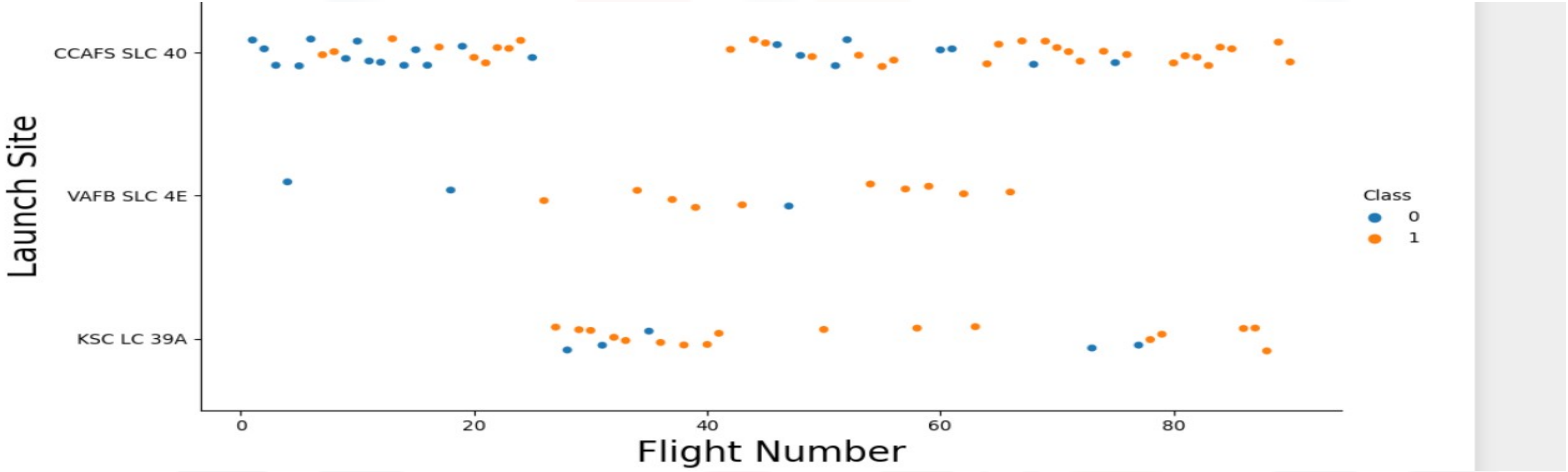
Results

✓ **Exploratory data analysis results:**

- Space X uses 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

EDA with Visualization

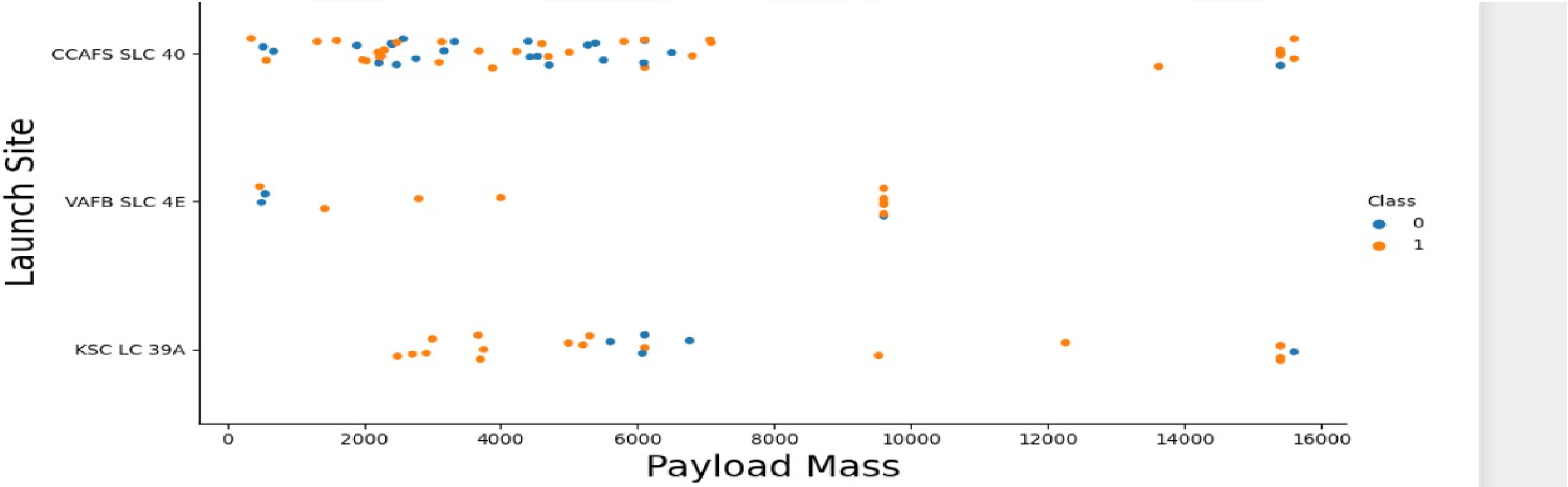
Flight Number vs. Launch Site



❑ **Explanation:**

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Payload vs. Launch Site



❑ **Explanation:**

- For every launch site the higher the payload mass, the higher the successrate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

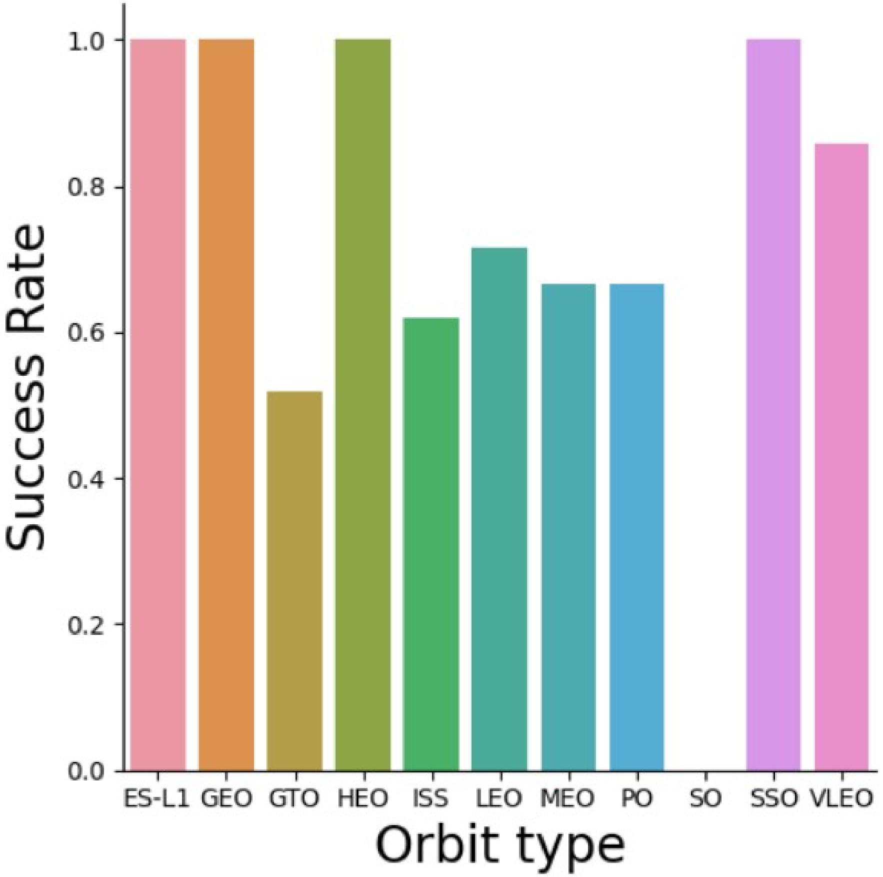
Success rate vs. Orbit type

Explanation:

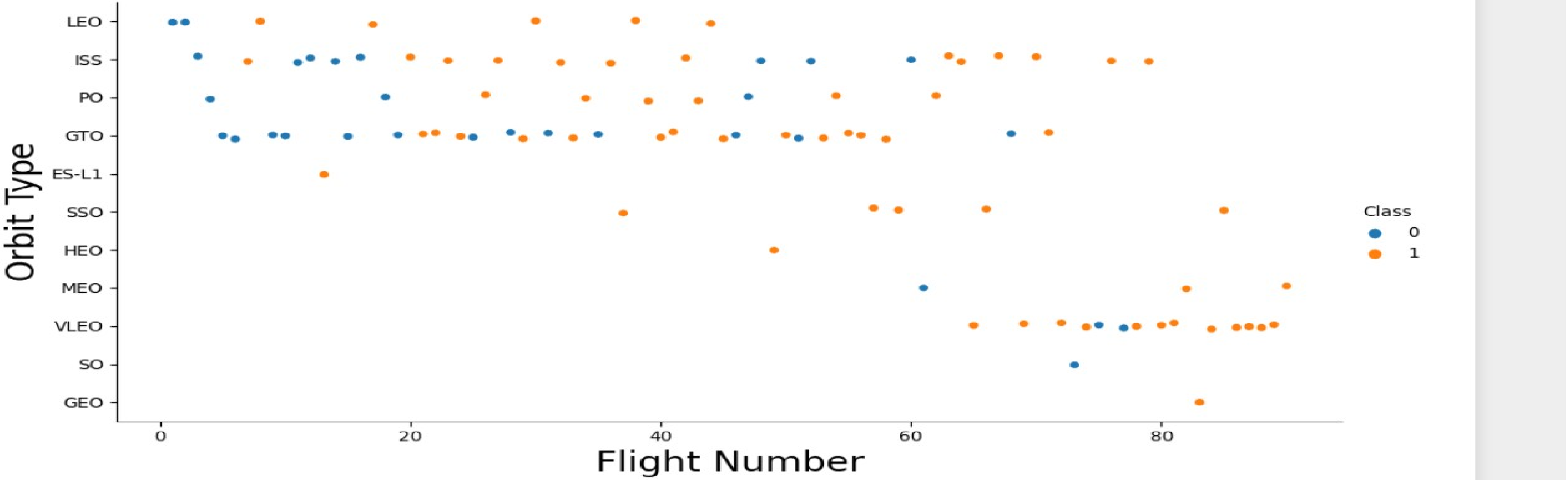
- ❑ Orbits with 100% success rate:-
  - ▪ ES-L1, GEO, HEO, SSO

- ❑ Orbits with 0% success rate:
  - ▪ -SO

- ❑ Orbits with success ratebetween 50% and 85%:
  - ▪ -GTO, ISS, LEO, MEO, PO

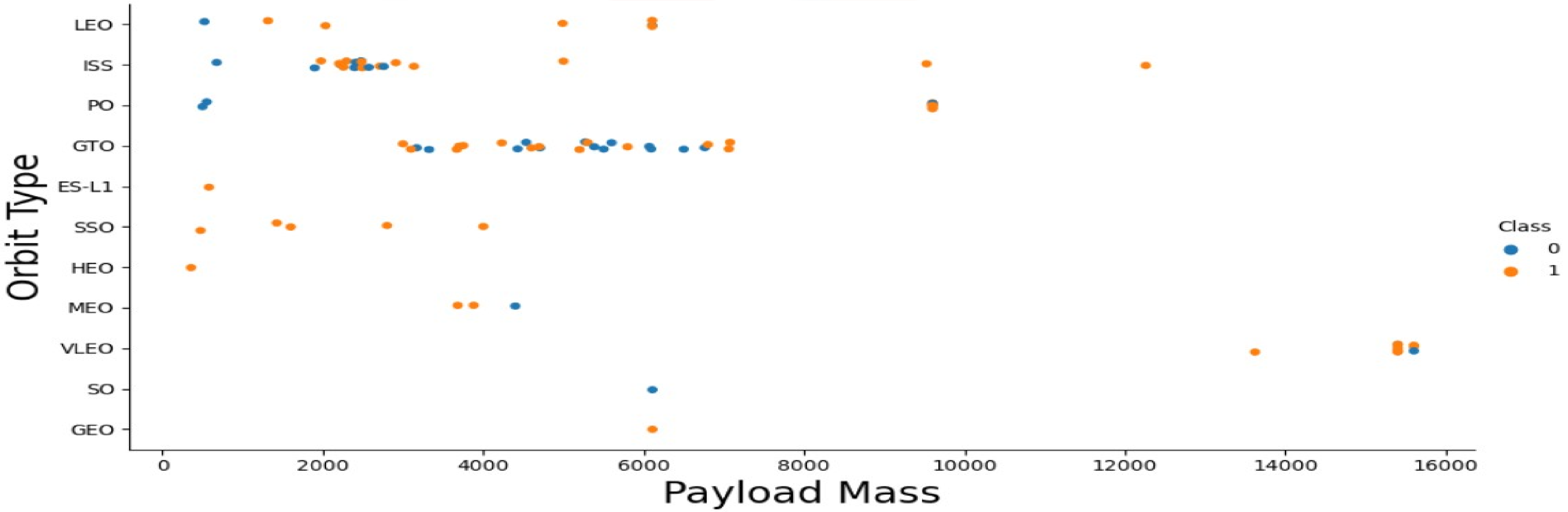


Flight Number vs. Orbit type



- **Explanation:**
  - 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.

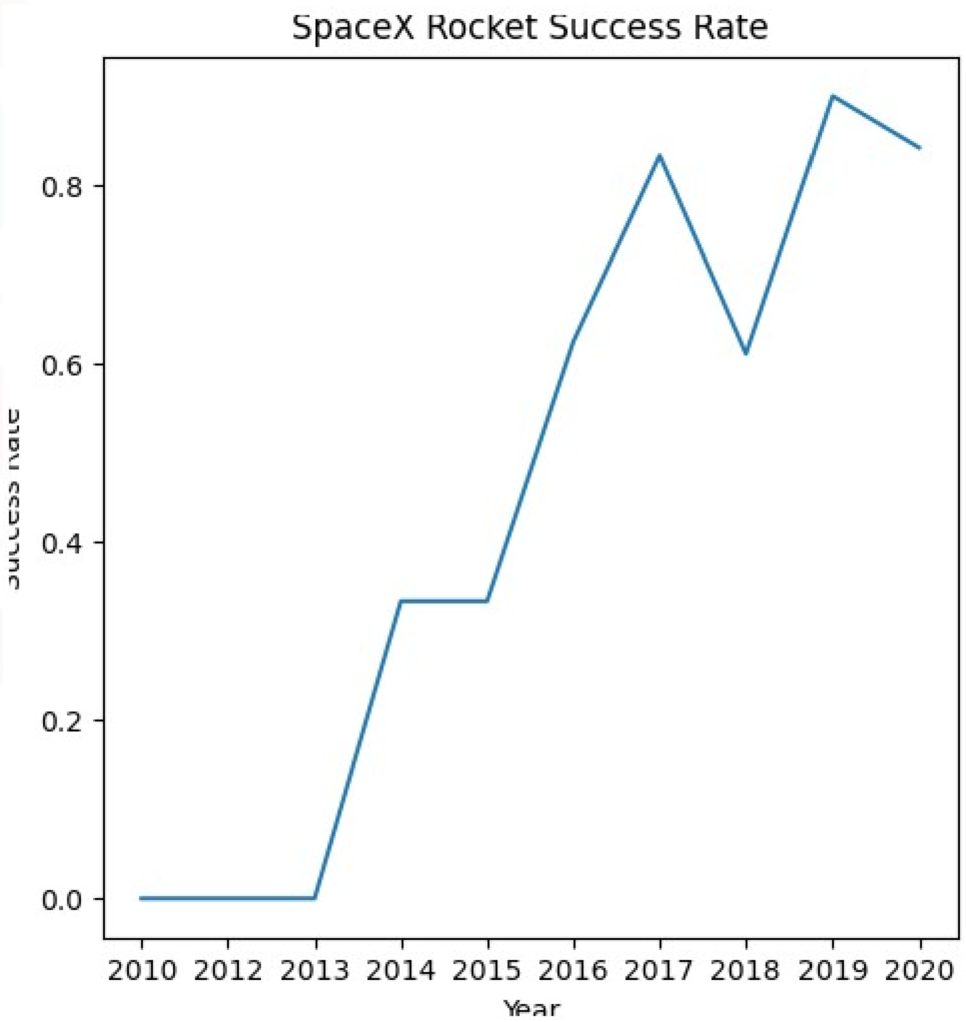
Payload Mass vs. Orbit type



❑ **Explanation:**

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch success yearly trend



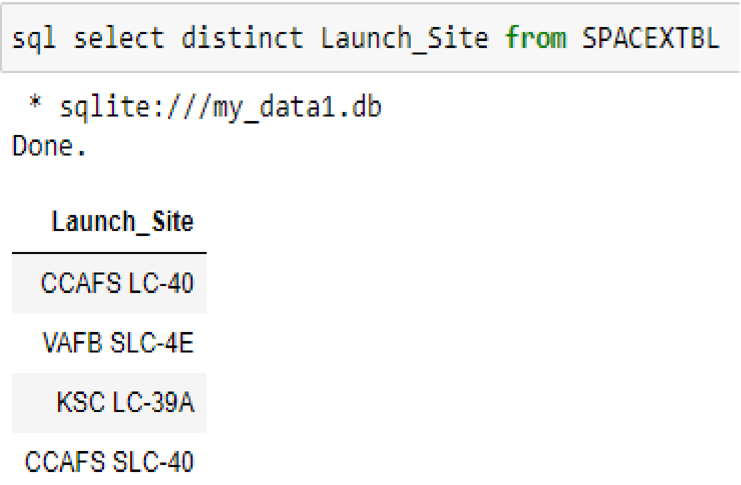
- **Explanation:**
  - The success rate since 2013 keep increasing till 2020.

EDA with SQL

Unique launch site names

SQL QUERY

```
sql select distinct Launch_Site
from SPACEXTBL
```



```
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

QUERY EXPLANATION:

using the keyword DISTINCT in the query means it will only shows the unique values in the Launch_site column from tblSpaceX

Launch site names begin with `CCA`

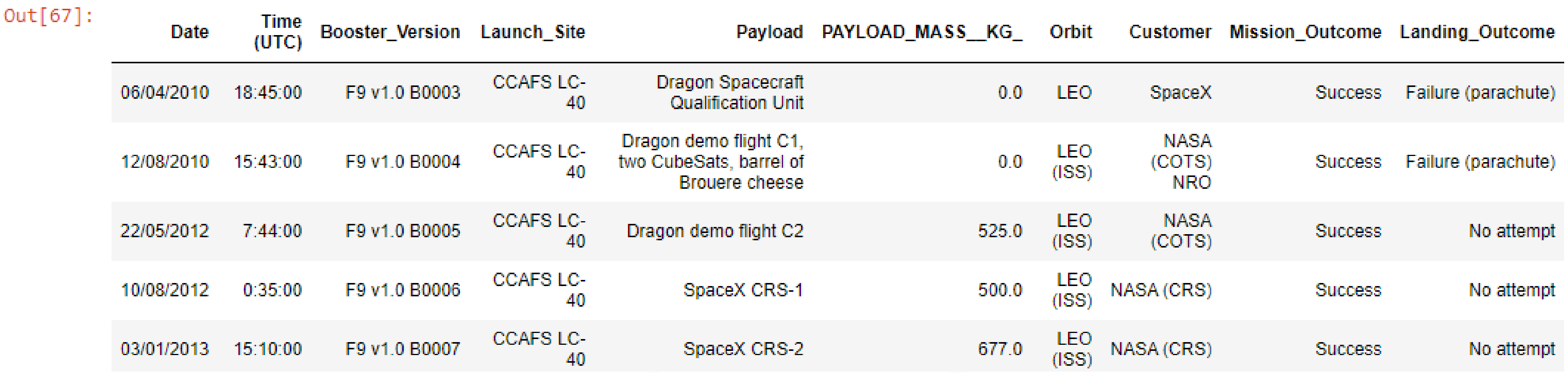
❑ **QUERY EXPLANATION :**

using the key word means ***TOP5*** in the query means that it will shows 5 records from the spacextbl

and ***like*** key word is a wild card with the words ***‘CCA%’*** the percentage symbol in the end suggests that the launch site name must starts with ***‘CCA’***

SQL QUERY :

```
    sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```



Out[67]:

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
	06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
	12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
	22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
	10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
	03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

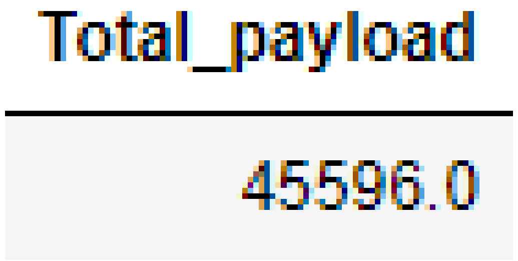
IBM Developer

SKILLS NETWORK

Total payload mass

SQL QUERY :

```
sql select sum(PAYLOAD_MASS__KG_) as
Total_payload from SPACEXTBL
where Customer = 'NASA (CRS)'
```



Total_payload
45596.0

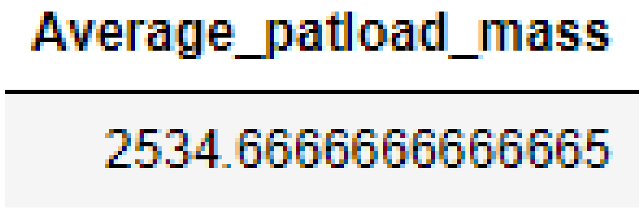
❑ **QUERY EXPLANATION :**

- Using the SUM summates the total in the column PAYLOAD_MASS__KG_
- THE WHARE clause filters the data set to only perform calculations ***Customer_Nasa(crs)***

Average payload mass by F9 v1.1

SQL QUERY :

```
sql select avg(PAYLOAD_MASS__KG_)
as Average_patload_mass from SPACEXTBL
where Booster_Version like 'F9 v1.1%'
```



Average_patload_mass
2534.6666666666665

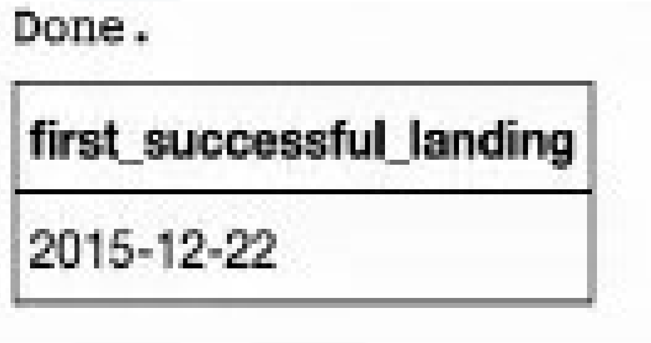
❑ **QUERY EXPLANATION:**

- Using the function ***AVG*** calculate the average in the column ***PAYLOAD_MASS__KG_***
- THE WHARE clause filters the data set to only perform calculations **Booster_Version 'F9 v1.1**

First successful ground landing date

SQL QUERY :

```
sql select  min(Date) as first_success_land
 from SPACEXTBL where Landing_Outcome
 like 'Success (ground pad)';
```



Done.

first_successful_landing
2015-12-22

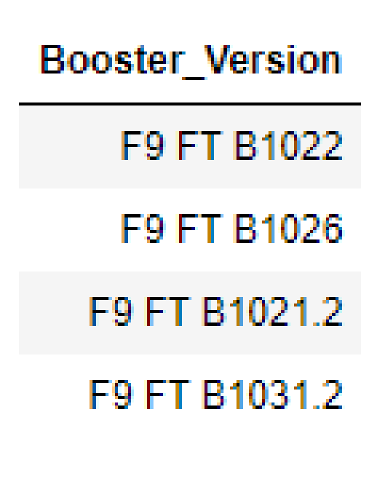
❑ **Query Explanation:**

- The MIN key word gives the minimum of that column
- Like is a wild card key word

Successful drone ship landing with payload between 4000 and 6000

SQL QUERY :

```
sql select Booster_Version from
 SPACEXTBL where Landing_Outcome
 like 'Success (drone ship)' and
 PAYLOAD_MASS__KG_ between 4000 and 6000
```



Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

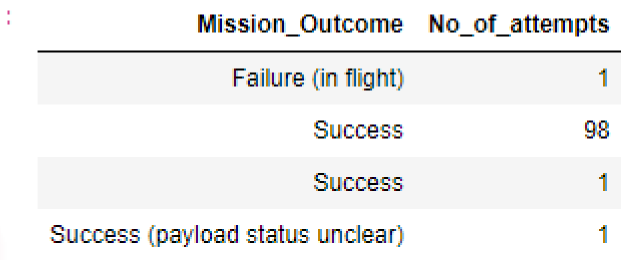
❑ QUERY EXPLANATION:

- Selecting only selects the booster version
- THE ***WHARE*** clause filters the data set to only perform calculations based on two statements
- the ***AND*** clause specifies(apply filter) the true when both conditions are true

Total number of successful and failure mission outcomes

SQL QUERY :

```
sql select Mission_Outcome,
count(Mission_Outcome) as No_of_attempts
from SPACEXTBL group by Mission_Outcome;
```



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

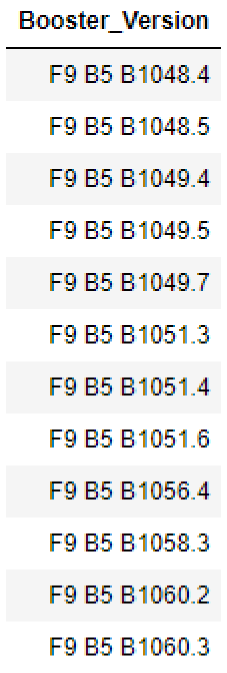
❑ QUERY EXPLANATION:

- Count Key word count the specific variable based on the condition and it give how many times it is repeated
- As key word acts as a alias

Boosters carried maximum payload

SQL QUERY :

```
sql select Booster_Version from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_)
from SPACEXTBL) order by Booster_Version
```



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

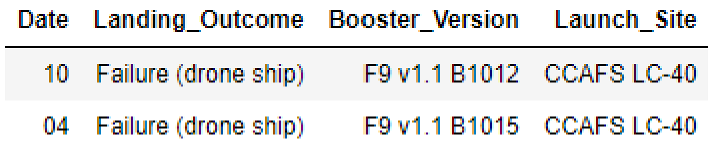
QUERY EXPLANATION:

- Here we using sub query to compare max payload mas from PAYLOAD_MASS__KG_
- Based on the max payload mass
- Select booster versions which have a max payload mass

2015 launch records

SQL QUERY :

```
 sql select substr(Date,4,2) as Date,
Landing_Outcome, Booster_Version,Launch_Site
from SPACEXTBL where Landing_Outcome
like 'Failure (drone ship)' and
(substr(Date,7,4) = '2015')
```



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

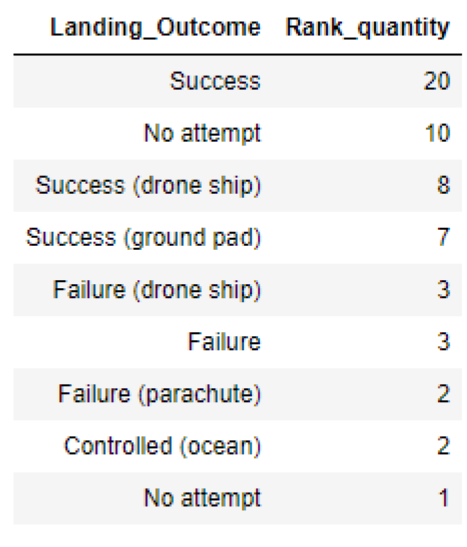
QUERY EXPLANATION:

- SUBSTR key word takes the some part of string from a string
- Using substr we take year from the total date
- Here WHERE clause takes two statements

Rank success count between 2010-06-04 and 2017-03-20

SQL QUERY :

```
        sql select Landing_Outcome,
        count(*)as Rank_quantity from SPACEXTBL
         where Date between '04-06-2010' and '20-03-2017‘
         group by Landing_Outcome
         order by Rank_quantity desc
```



Landing_Outcome	Rank_quantity
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

QUERY EXPLANATION:

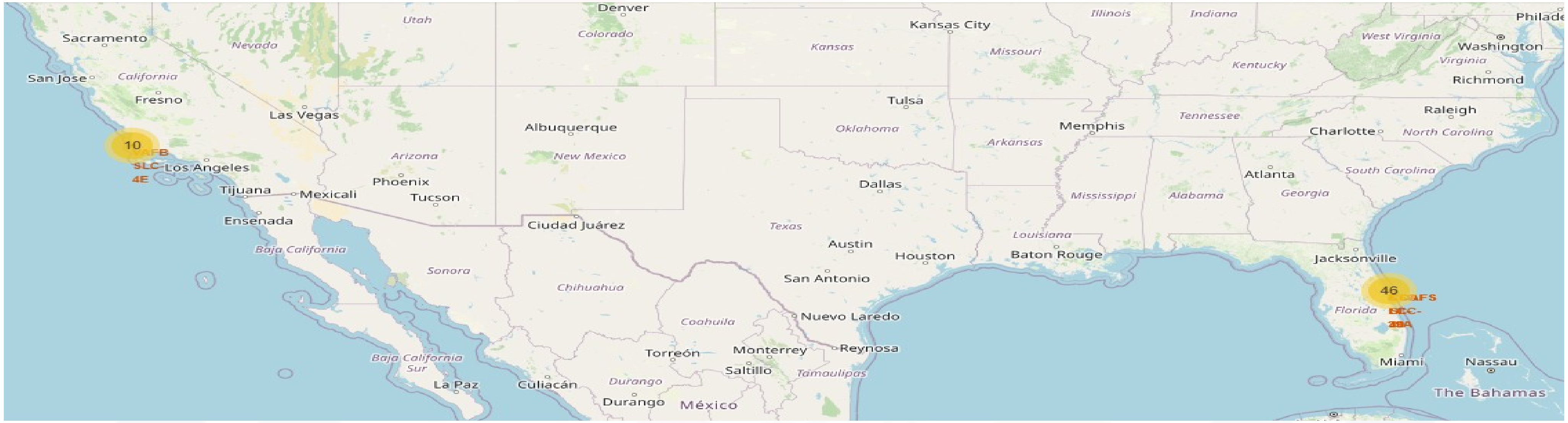
- Using ***GROUP BY*** we combine same type of data in to one group
- An ***ORDER BY*** Gives the data in a ascending order by default
- ***COUNT*** counts the variable how many no of times its repeated

Interactive map with Folium

IBM Developer

SKILLS NETWORK

All launch sites’ location markers on a global map



Explanation:

- •Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- •All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.

Colour-labeled launch records on the map

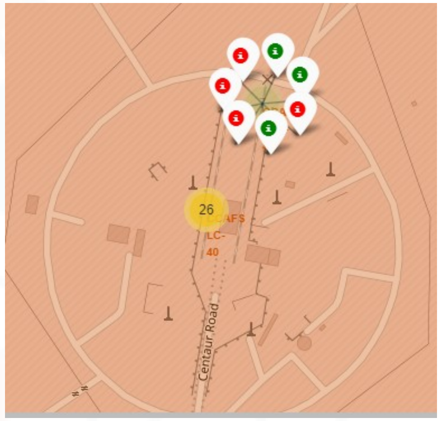
❑ **Explanation:**

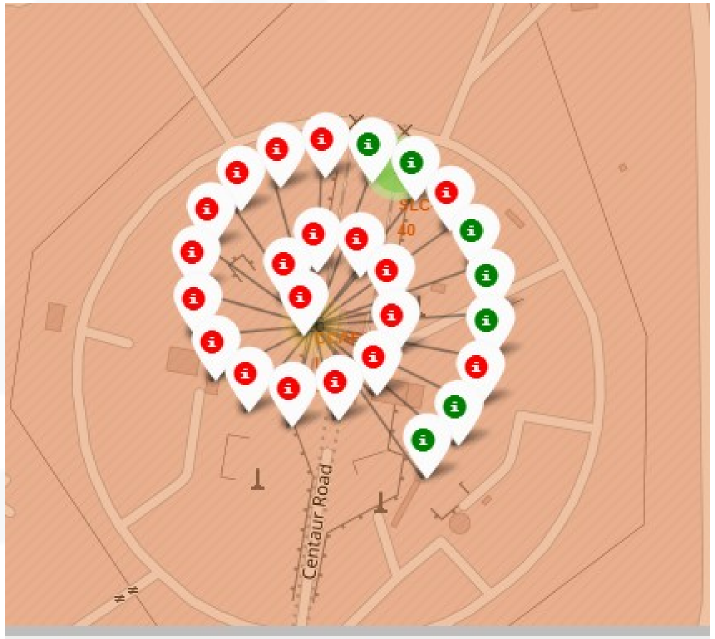
•From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

-Green Marker = Successful Launch

-Red Marker = Failed Launch

•Launch Site KSC LC-39A has a very high Success Rate.



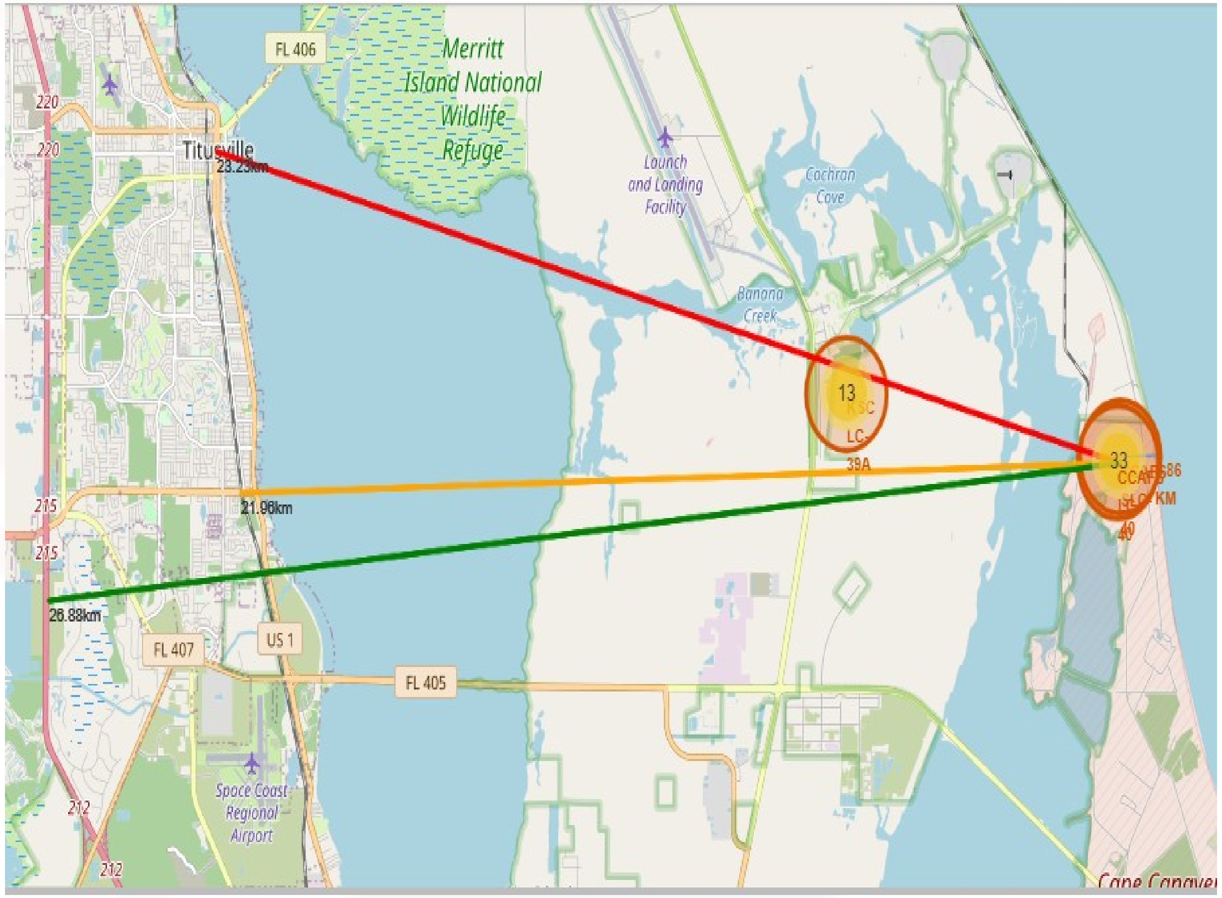


IBM Developer

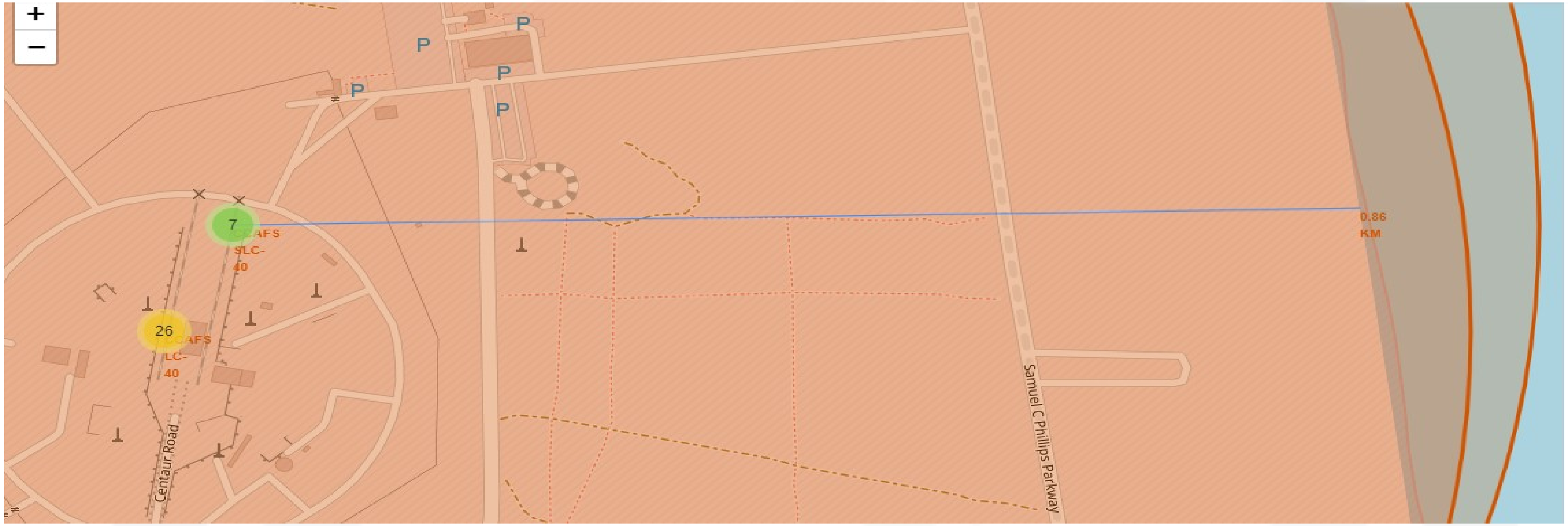
SKILLS NETWORK

Distance from the launch site CCAFS SCL 40 to its proximities

- **Explanation:**
- •From the visual analysis of the launch site CCAFS SCL 40 we can clearly see that it is:
- -relative close to railway (21.96 km)
- -relative close to highway (26.88 km)
- -relative close to coastline (0.86 km)
- •Also the launch site CCAFS SCL 40 is relative close to its closest city Titusville (23.2 km).
- •Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



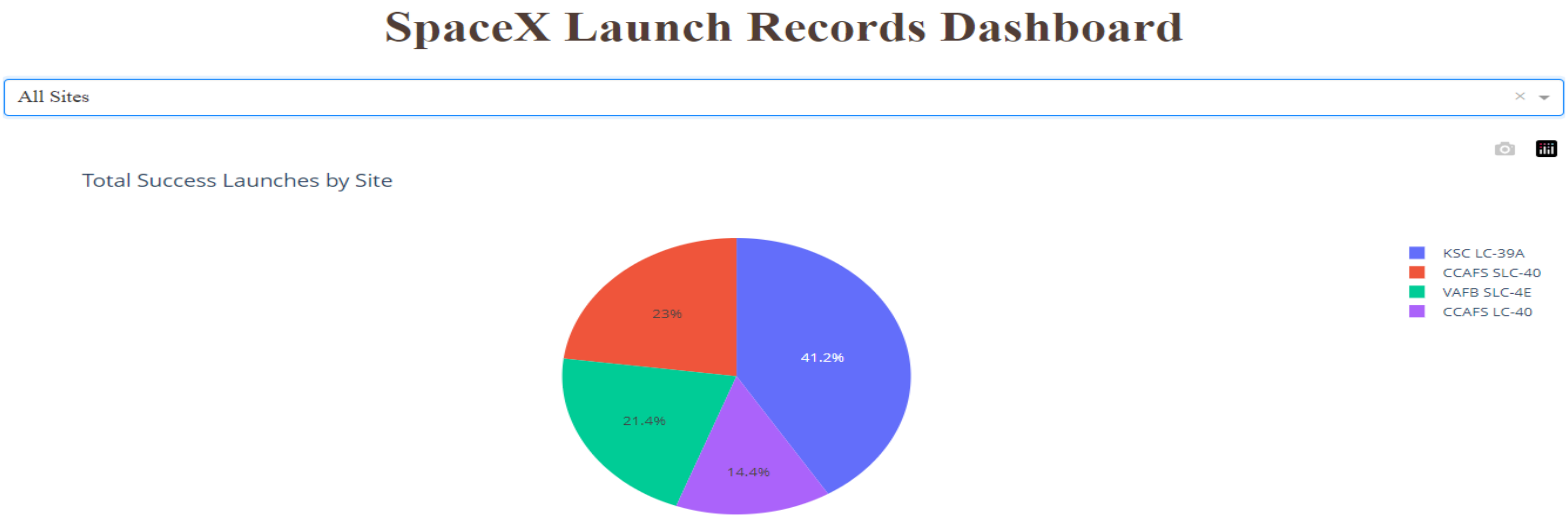
a PolyLine between a launch site to the selected coastline point



- the costal line distance from CCAFS SCL 40 is 0.86 is nearly ~ 0.9

Build a Dashboard with Plotly Dash

Launch success count for all sites



SpaceX Launch Records Dashboard

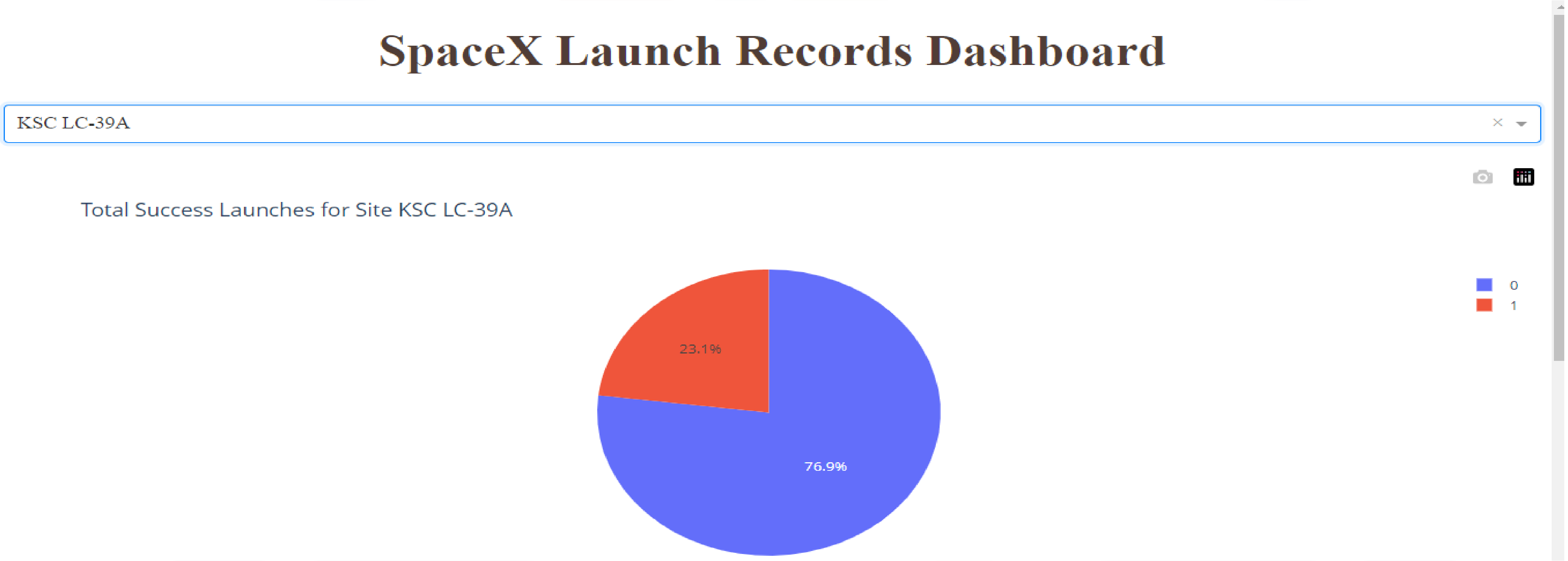
All Sites

Total Success Launches by Site

❑ **EXPLANATION:**

- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch site with highest launch success ratio



SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for Site KSC LC-39A

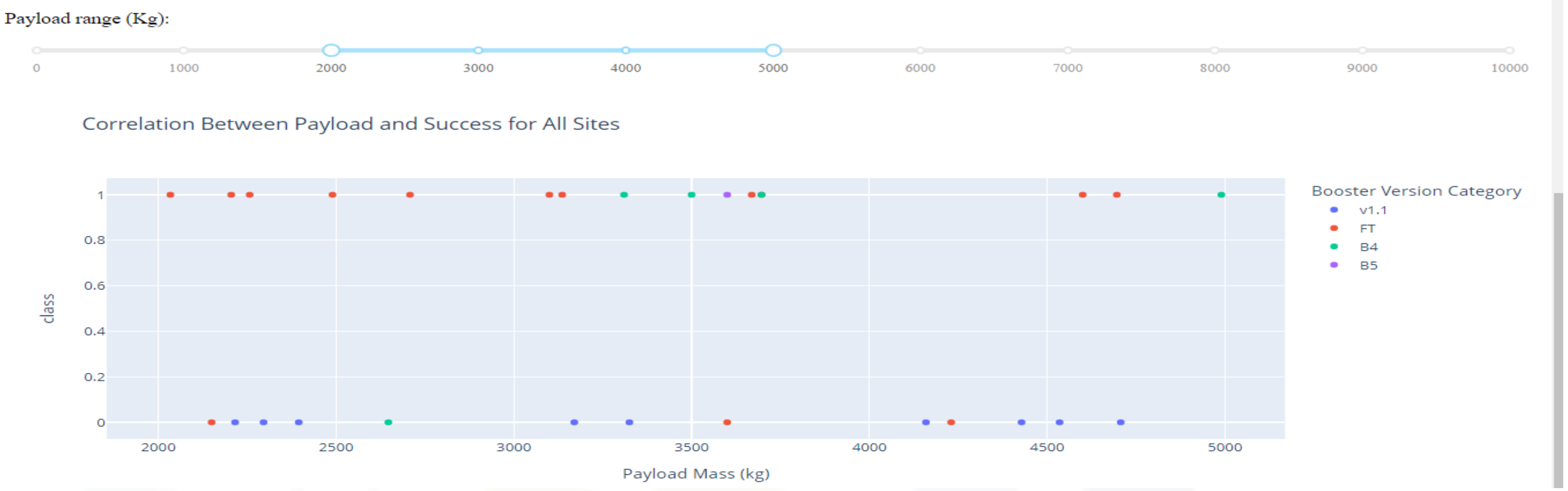
❑ EXPLANATION:

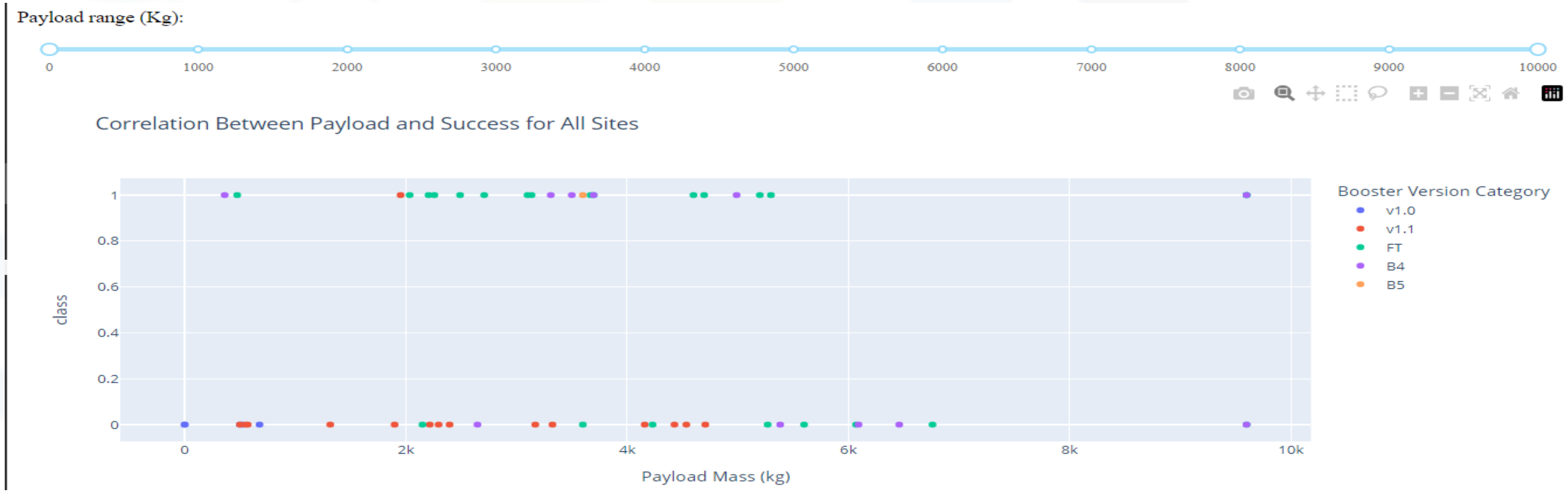
- using the keyword DISTINCT in the query means it will only shows the unique values in KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload Mass vs. Launch Outcome for all sites

❑ **Explanation:**

The charts show that payloads between 2000 and 5500 kg have the highest success rate.



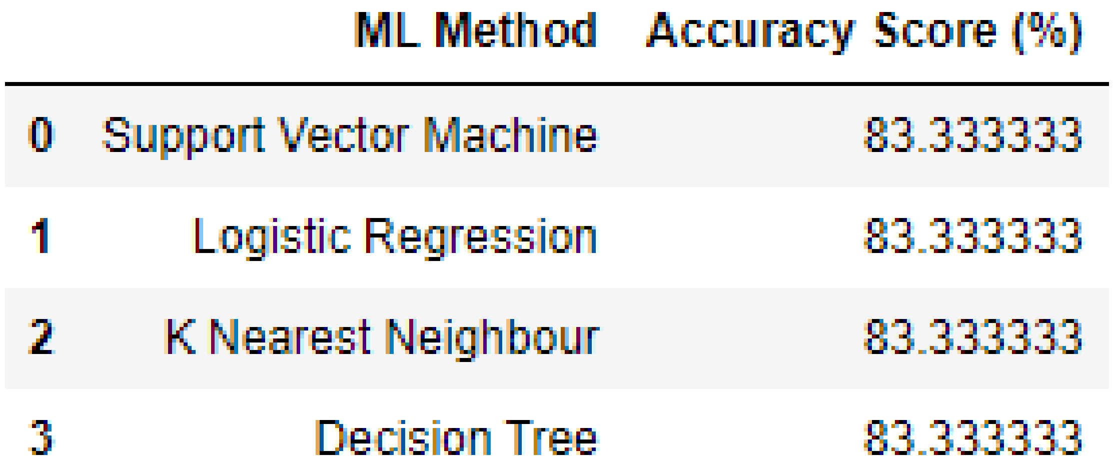


Predictive analysis (Classification)

Classification Accuracy

- ❑ **Accuracy Classification Using Training Data:**

  As we can see our Accuracy is
  extremely close all the models have
  a same accuracy on test data set is 83.33%
  it should be noted the size of the sample
  is small at only sample size of 18

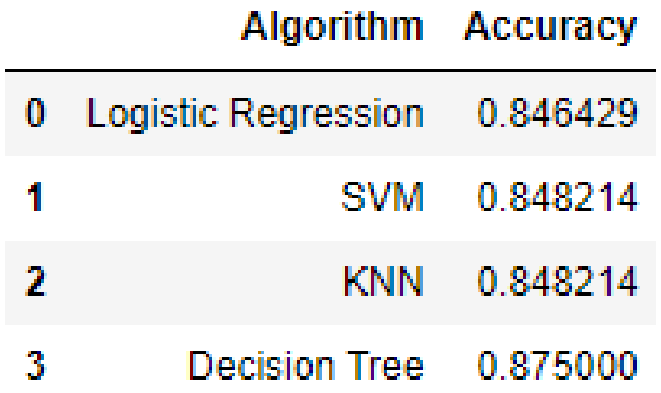


	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

Classification Accuracy

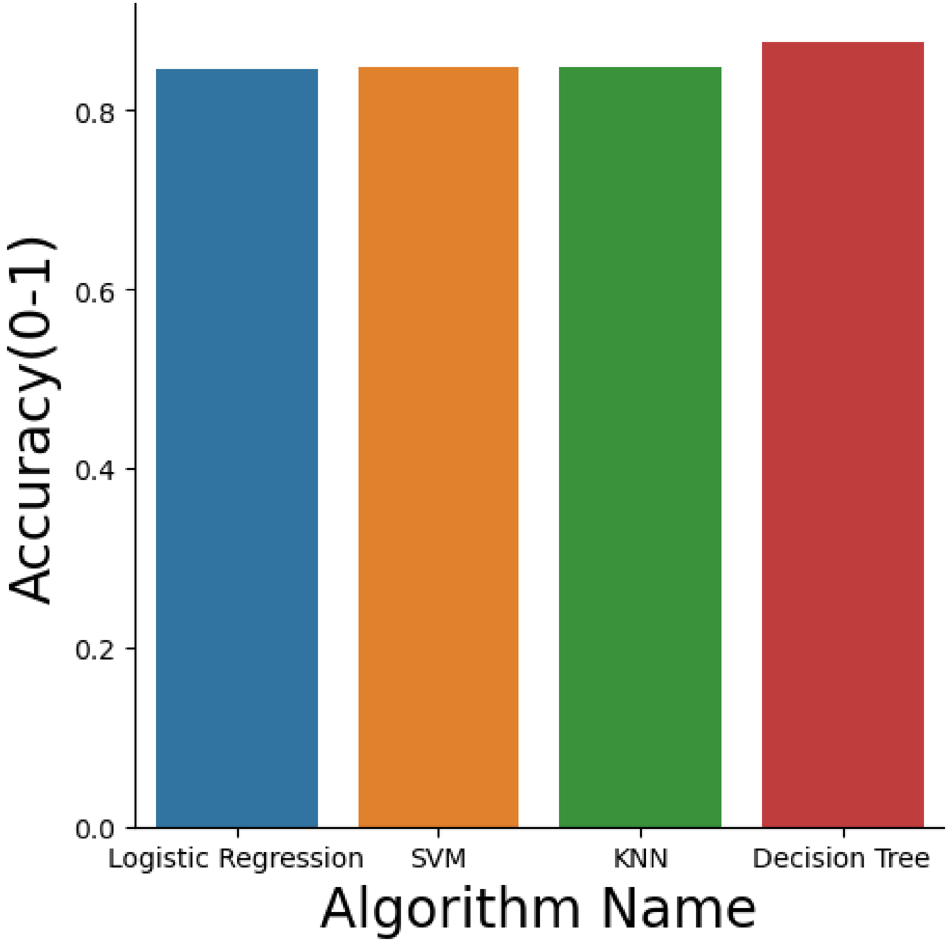
❑ **Accuracy Classification Using Training Data:**

here we see that we show the data based on the best_score we see the small difference and got the solution



	Algorithm	Accuracy
0	Logistic Regression	0.846429
1	SVM	0.848214
2	KNN	0.848214
3	Decision Tree	0.875000

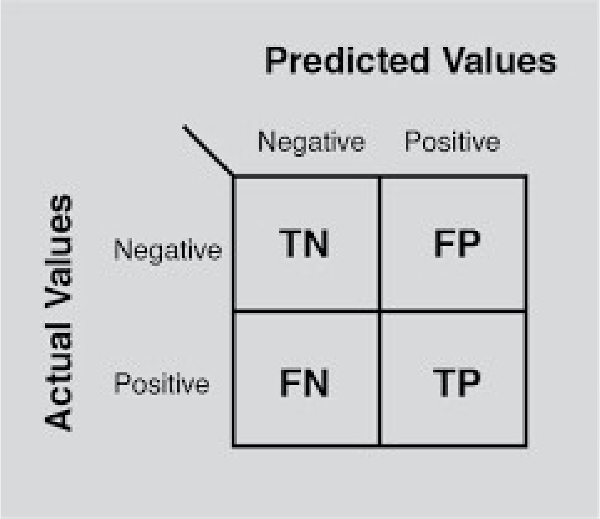
Here we see that decision tree do the best performance



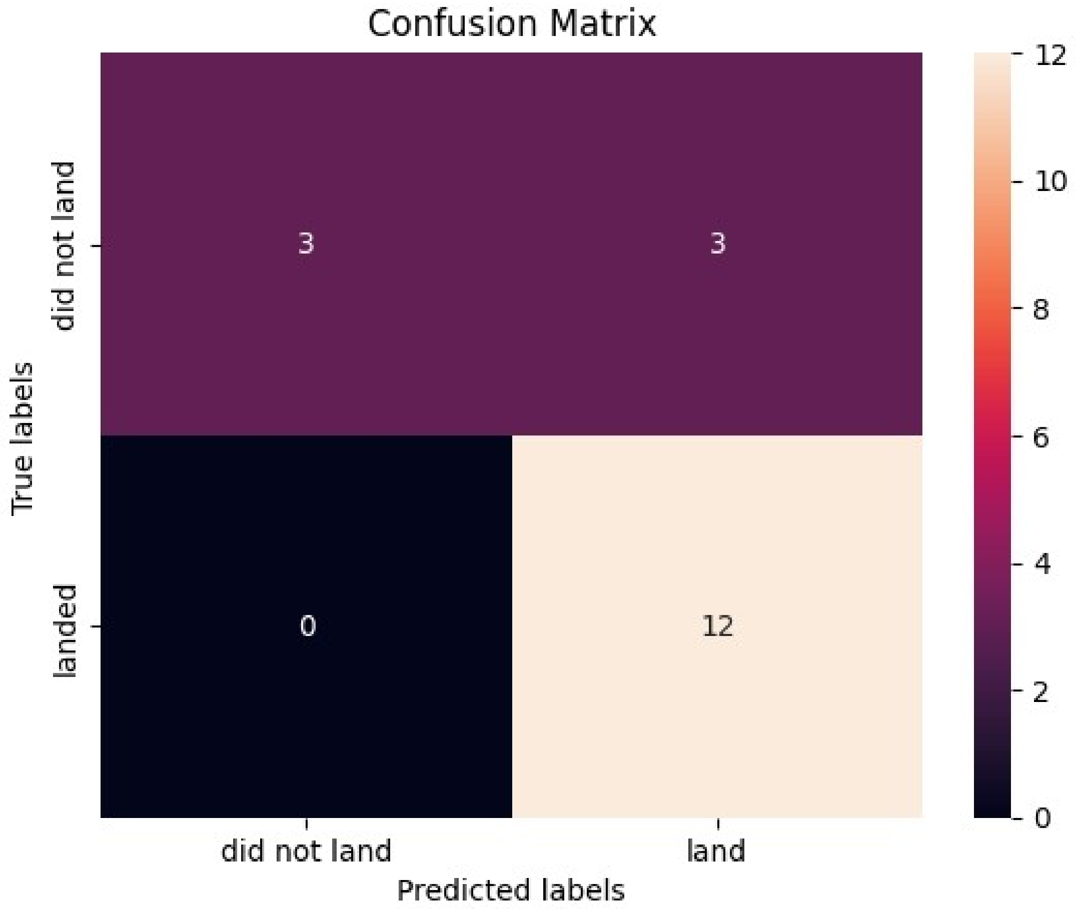
Confusion Matrix

❑**EXPLANATION:**

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN	FP
	Positive	FN	TP



CONCLUSION

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

APPENDIX

- Special thanks to
- Instructors:
  - Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo
  - Coursera
  - IBM