Data Science Capstone project

NHAN NGUYEN CAO

28 August 2021

Outline

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

Executive Summary

For this Capstone project, the methodologies of **Data Collection**, **Data Preparation** (using Data Wrangling), **Data Understanding** (using some Exploratory Data Analysis (EDA) techniques) as well as **Modeling** and **Evaluation** were used to achieve the goals.

Generally speaking, the project was able to draw a complete and general overviewing of the dataset to point out some insights as well as successfully build a predictive model to predict the set problem.

Introduction

Taking the scenario of a newly-created rocket company – **Space Y** – that is aiming to compete with Space X, the project's main goal is to draw some **beneficial insights** from public dataset of Space X about the launches of Falcon on the first stage (focus on Falcon 9) and used those insights to reach the **answer to the question**.

For this Capstone project, the question to answer is whether the provided dataset can be used to build a good predictive model that predict *if the Falcon 9 first stage will land successfully*.

Methodology

- Data collection methodology:
  - Using REST API and Web Scraping
- Perform data wrangling
  - Dealing with Null value and build predict labels.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Evaluate on 4 types of models, using GridSearch for hyperparameters tuning.

Methodology

Data collection

Dataset for the Capstone project was gathered using 2 methods from 2 sources:

1. Gathering data from SpaceX REST API, an open-source API storing data relating to SpaceX. The project focus on the dataset about past launches of rockets of SpaceX.
2. Gathering data about Falcon 9 and Falcon Heavy launches by using Web Scraping technique from Wikipedia page about **"List of Falcon 9 and Falcon Heavy launches"**

Data collection – SpaceX API

1. The data is collected by using **GET method** on the provided static URL using the requests library.

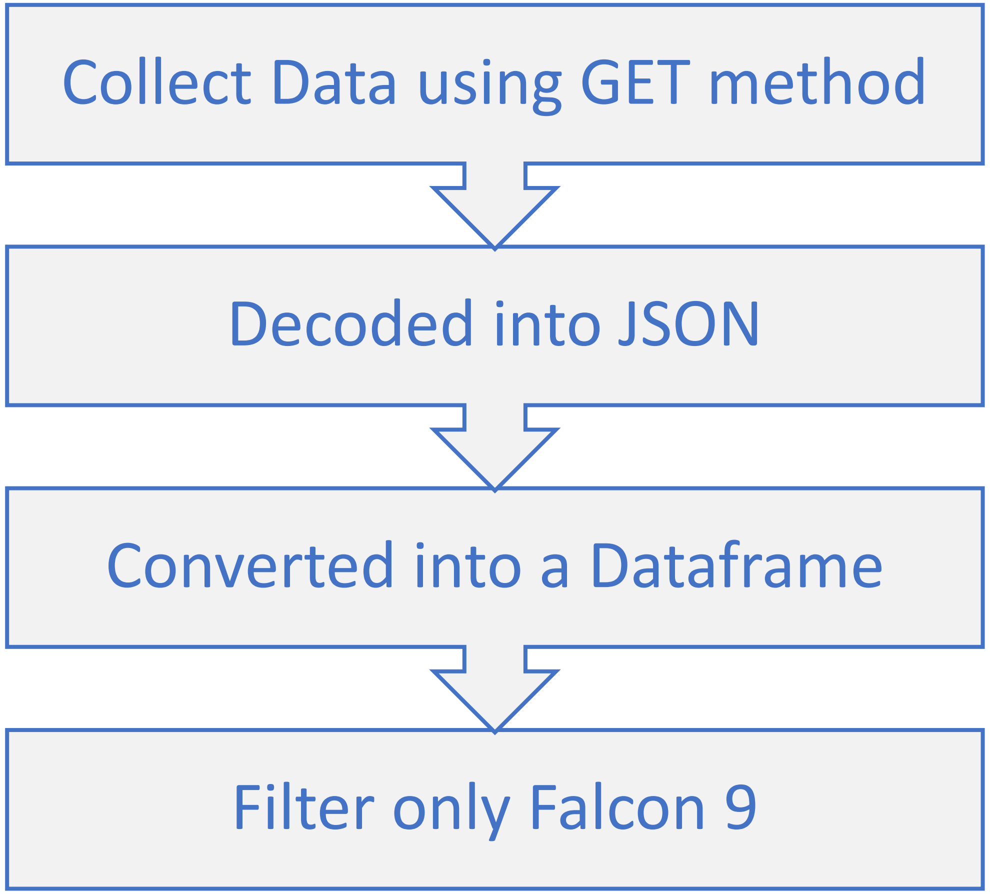
2. The response is then **decoded into JSON** format using .json() function.

3. Then, the decoded response is **converted into a DataFrame** using .json_normalize().

4. Finally, we filter the dataframe to **only contain rows about Falcon 9** launches.

Notebook: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/Data%20Collection.ipynb

Flowchart



Data collection – SpaceX API
Final Data

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	7	B1060	-80.603956	28.608058
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	7	B1058	-80.603956	28.608058
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	9	B1051	-80.603956	28.608058
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	7	B1060	-80.577366	28.561857
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	1	B1062	-80.577366	28.561857

90 rows × 17 columns

Data collection – Web scraping

1. The data is collected by using **GET method** on the provided static URL using the requests library.

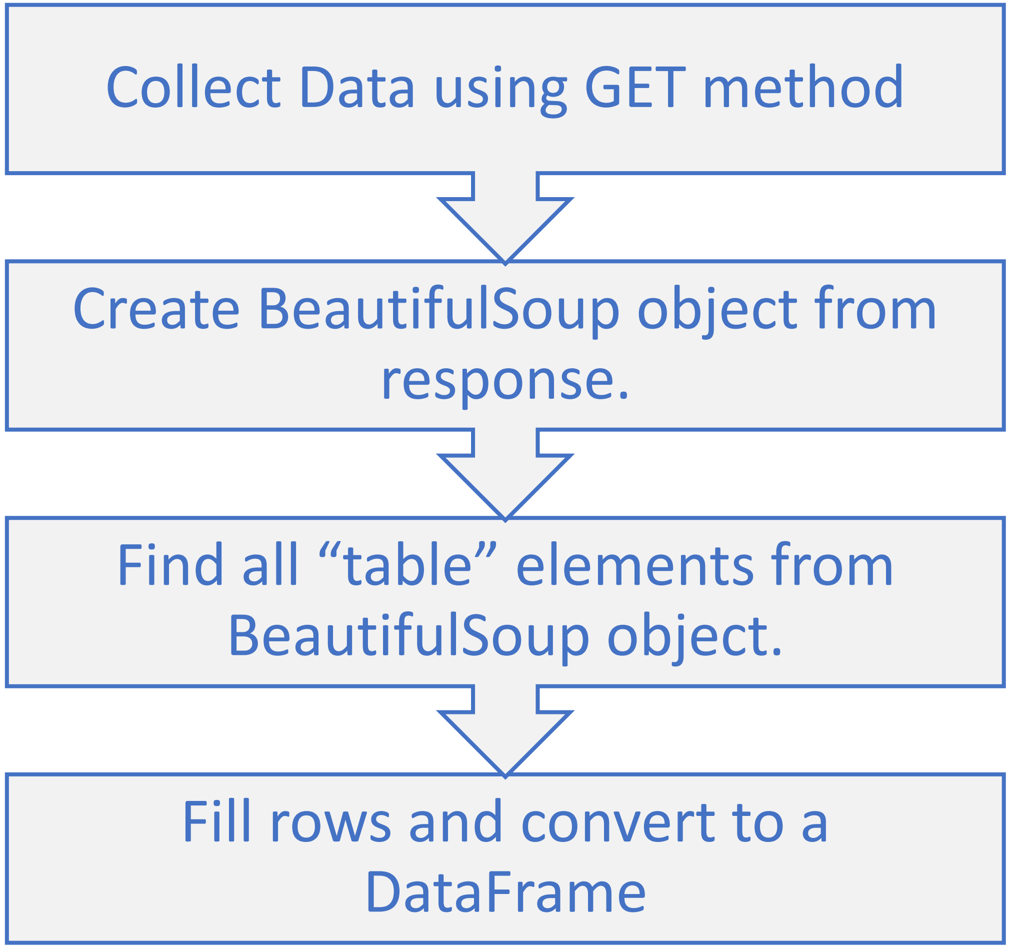
2. Create a **BeautifulSoup** object from the received response.

3. Find **“table” elements** from the BeautifulSoup object then extract column names.

4. Fill a dictionary with rows from the table, then convert to a **DataFrame** for usage.

Notebook: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Flowchart



Data collection – Web scraping

Final Data

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	SpaceX	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2	Success	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA	Success\n	F9 B5B1067.1	Success	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success	6 June 2021	04:26

121 rows × 11 columns

Data wrangling

- After Data Collection, the Data wrangling is performed. The main goals of this process includes dealing with null data and identifying target labels, creating appropriate training labels from that for the model.

Examine null data and data types for each columns → Calculate number of launches on each site and occurrence of each orbit → Determine available labels for Outcome, group into 2 groups (success and failure) → Create Outcome label (0 for failure, 1 for success) as target label for model

Notebook: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/Data%20Wrangling.ipynb

Data wrangling – Dealing with Null data

- Overviewing the description of data, the only column containing Null data is “LandingPad”, which is not a numerical attribute. Therefore, no further action was performed.

Data wrangling – Creating the target labels

- Since the goal of the project is creating a predictive model capable of predicting if each Falcon 9 launch will be successful or not. Therefore, the target labels need to be able to demonstrate this.
- Examining the “Outcome” column points out that there are 8 available values for this column, each of which has a clear outcome of either success or failure. Therefore, classifying the 8 labels into 2 groups and encoding each group with an appropriate label (0 for failure and 1 for successful) was the method chose.

True ASDS	41
None None	19
True RTLS	14
False ASDS	6
True Ocean	5
None ASDS	2
False Ocean	2
False RTLS	1

0	0
1	0
2	0
3	0
4	0
5	0
6	1

Data wrangling - Result

- After Data Wrangling process, the final dataset looks like below. The "Class" column will be used as the target label for the model.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

EDA with data visualization

- The first method used for Exploratory Data Analysis (EDA) was using Data Visualization.
- By using different types of charts and graphs, insights were pointed out from the data as well as the relationships between attributes inside data.
- Notebook: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/EDA%20with%20Visualization.ipynb

EDA with data visualization

- 6 charts were created to draw insights from data, including
  - **Flight Number vs. Launch Site**: use ***scatter plot*** to find relationship between the two variables.
  - **Payload vs. Launch Site**: use ***scatter plot*** to find relationship between the two variables.
  - **Success rate vs. Orbit type**: use a ***bar chart*** to easily compare the success rate between orbit types.
  - **Flight Number vs. Orbit type**: use ***scatter plot*** to find relationship between the two variables.
  - **Payload vs. Orbit type**: use ***scatter plot*** to find relationship between the two variables.
  - **Launch success yearly trend**: use ***line chart*** to better illustrate the trend and changes in success rates through yeras

EDA with SQL

- Further into Exploratory Data Analysis, some SQL queries were also used to gather more specific details and insights.
- The performed SQL queries include:
  - 1. Display the names of the unique launch sites
  - 2. Display 5 records where launch sites begin with the string 'CCA’
  - 3. Display the total payload mass carried by boosters launched by NASA (CRS)
  - 4. Display average payload mass carried by booster version F9 v1.1
  - 5. List the date when the first succesful landing outcome in ground pad was acheived.
  - 6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - 7. List the total number of successful and failure mission outcomes
  - 8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - 9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015
  - 10. 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.
- Notebook: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/EDA%20with%20SQL.ipynb

Build an interactive map with Folium

- An interactive map was added using Folium library, showing the launch sites for different launches.
  - Each launch site is displayed using a **Circle** and a **Marker**.
  - Each launch is illustrated using a **Marker** containing an Icon showing the outcome of the launch (using color of the Icon), and is included in a **MarkerCluster** of the launch site.
  - Distance from launch sites to some locations, cities were illustrated using **Markers** and **PolyLines** for better showing the estimated Euclidean distance**.**
- Notebook: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

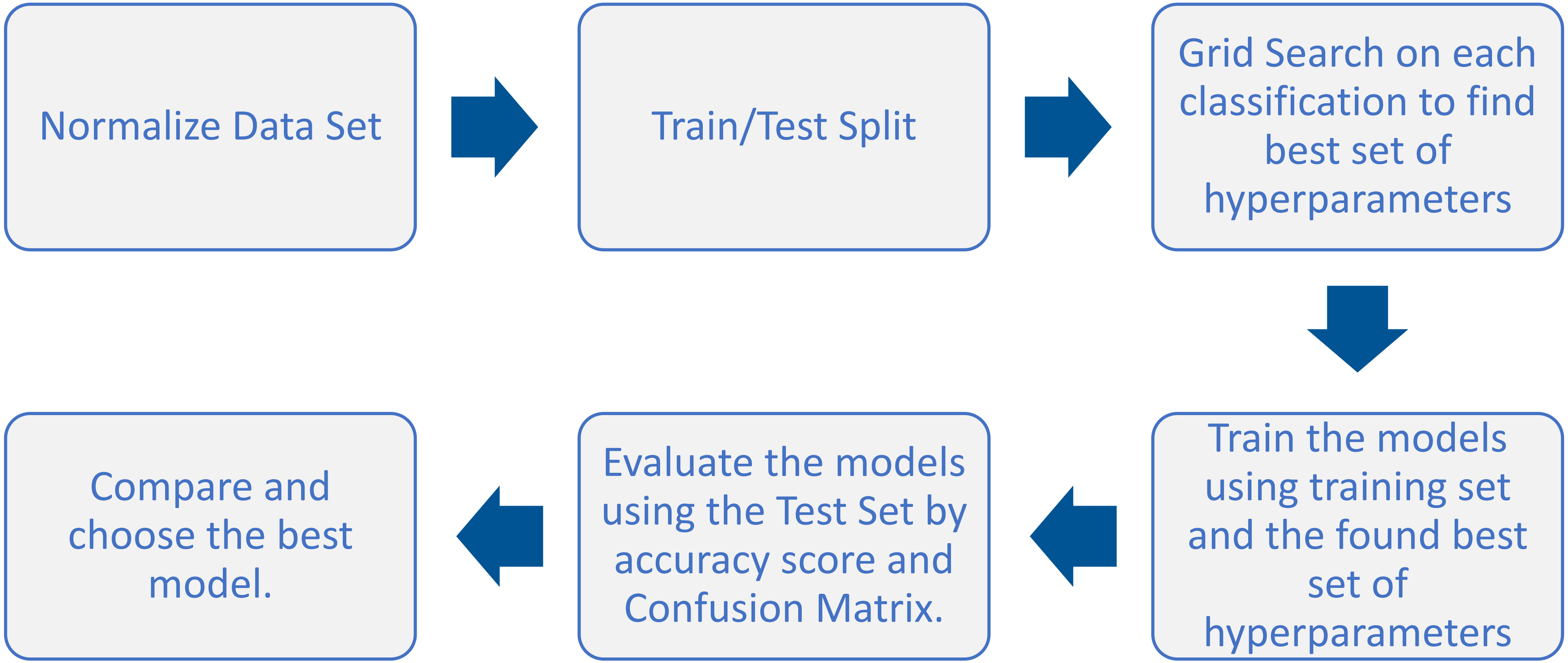
- A Dashboard was built for interactively visualization from the dataset.
- A total of 2 plots/graphs were used to build the Dashboard, include:
  - A pie chart showing either the proportion of success rate of different launch sites compared to the overall success rate or the success/failure proportion for each launch site.
  - A scatter plot showing the correlation between Payload and Success rate for either all sites or of a site, categorized by booster version. This plot is linked with a slider for interactive input range of payload mass.
- Dash file: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/spacex_dash_app.py

Predictive analysis (Classification)

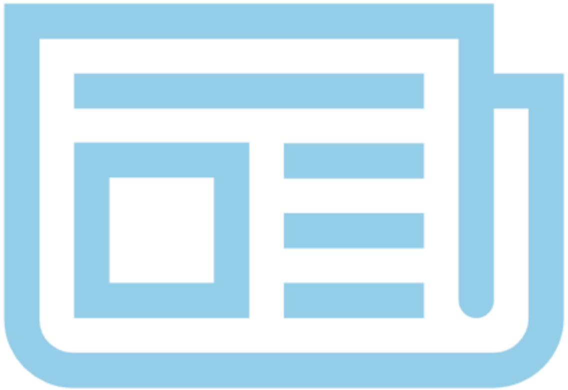
- Building a model to predict if a Falcon 9 launch will land successfully, 4 different classification models were chosen to be evaluate to choose the best one including **Logistic Regression**, **SVM**, **Decision Tree** and **K-Nearest Neighbors**.
- The dataset is split into 8/2 train/test ratio and **normalized** using Standard Scaler.
- After that, **Grid Search** was performed for each model to find the best set of hyperparameters for each one. Then **evaluate the accuracy** for each model to choose the best one.
- Notebook: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/Machine%20Learning%20Prediction.ipynb

Predictive analysis (Classification)

Model Development Process Workflow



Results

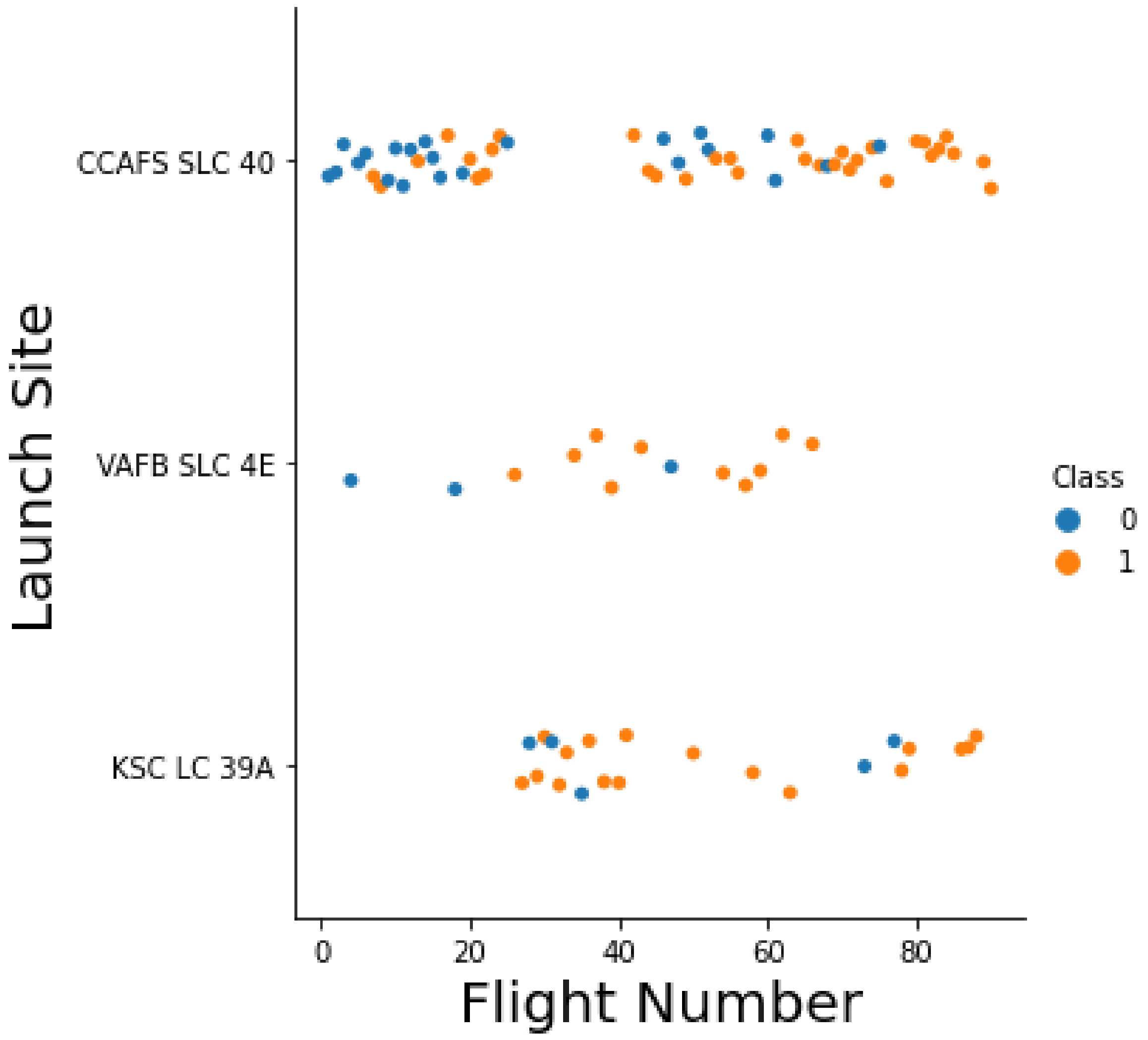


- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

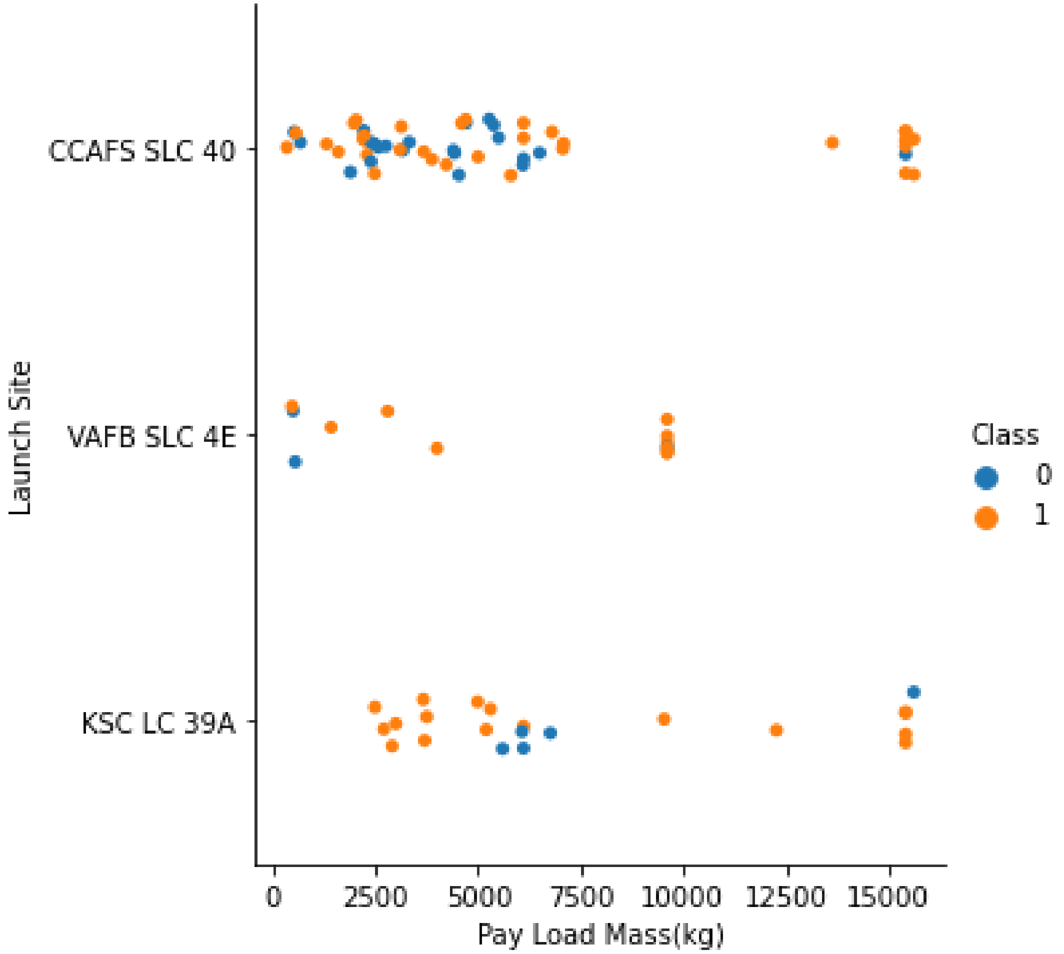
Flight Number vs. Launch Site

- There is no really clear relationships can be found between Flight Number and Launch Site with the outcome of the launch.
- Generally speaking, VAFB SLC 4E tends to be more likely to be successful for higher flight number.
- However, for the remaining launch sites, the relationships are not clearly available at all.



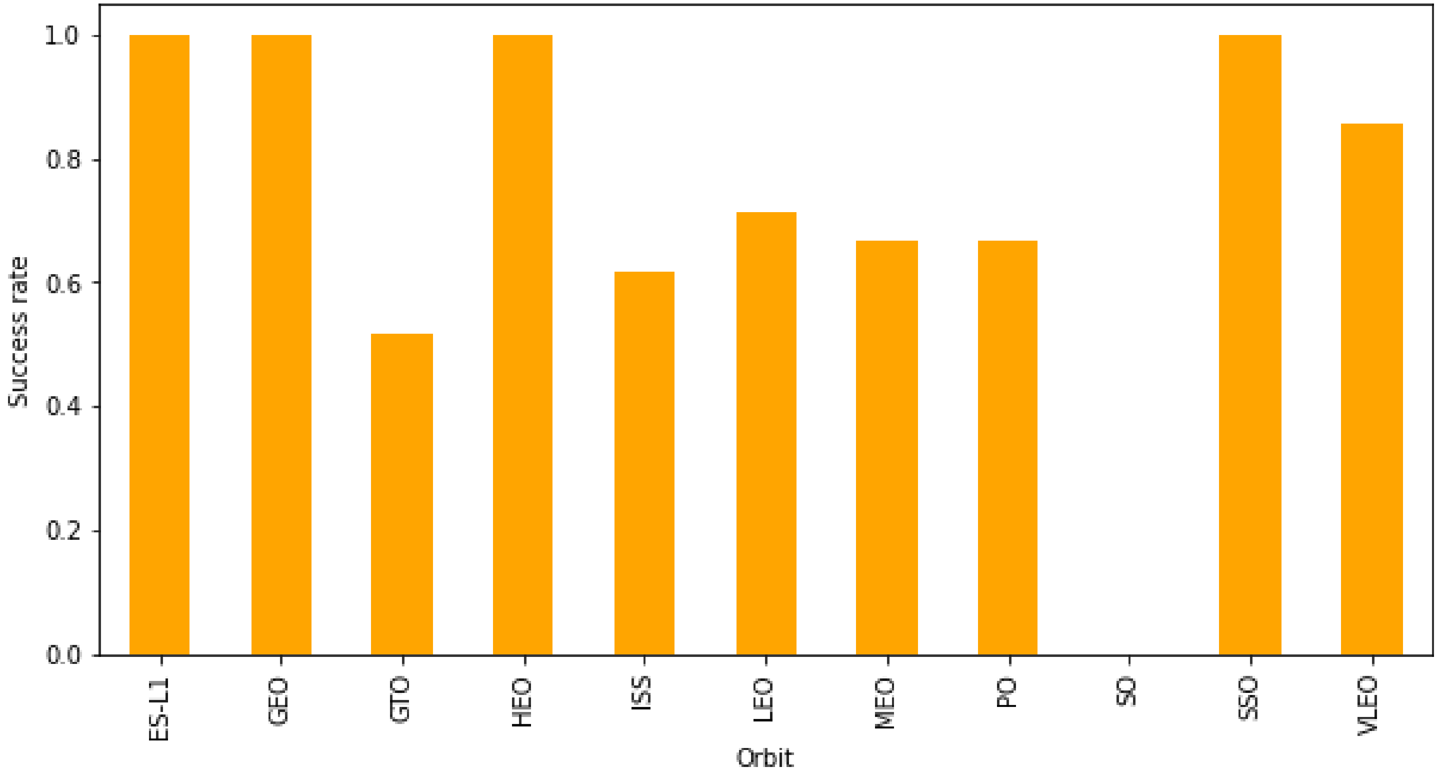
Payload vs. Launch Site

- The relationship between payload mass and launch sites in contrast is more clear than flight number.
- We can see from the scatter plot that for all three launch sites, the heavier the payload, the more likely it is to be successful.
- For CCAFS SLC 40, the result for the ones lighter than 10000 kg is harder to be determined while for the others, there exists a range in which the results are likely to be unsuccessful (for example: close to 0 for VAFB SLC 4E and around 5000 to 7000 for KSC LC 39A).



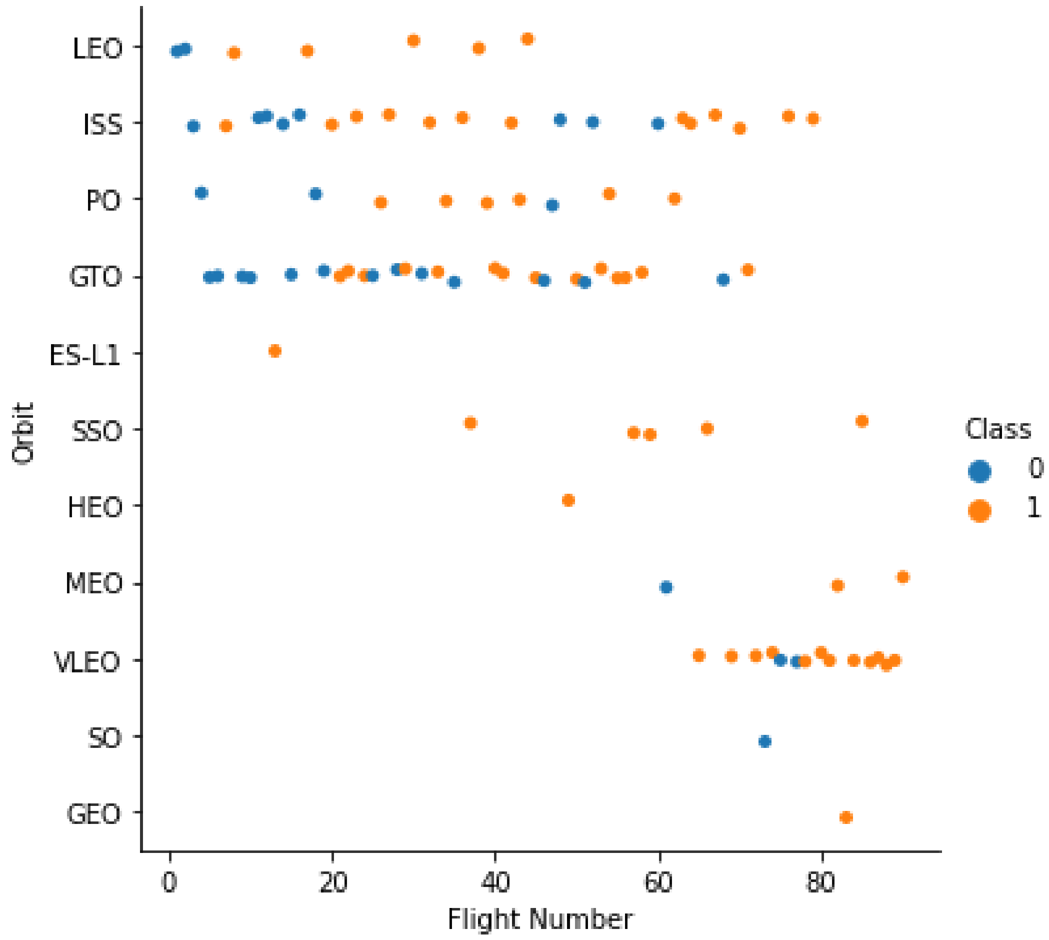
Success rate vs. Orbit type

- The bar chart pointed out that there are some types of Orbit that are much more likely to be successful compared to each other.
- ES-L1, GEO, HEO, SSO are orbit types with success rates of 100%, while for SO, it is a surprisingly 0% success rate. The remaining varies but in general the success rate is higher than a half.



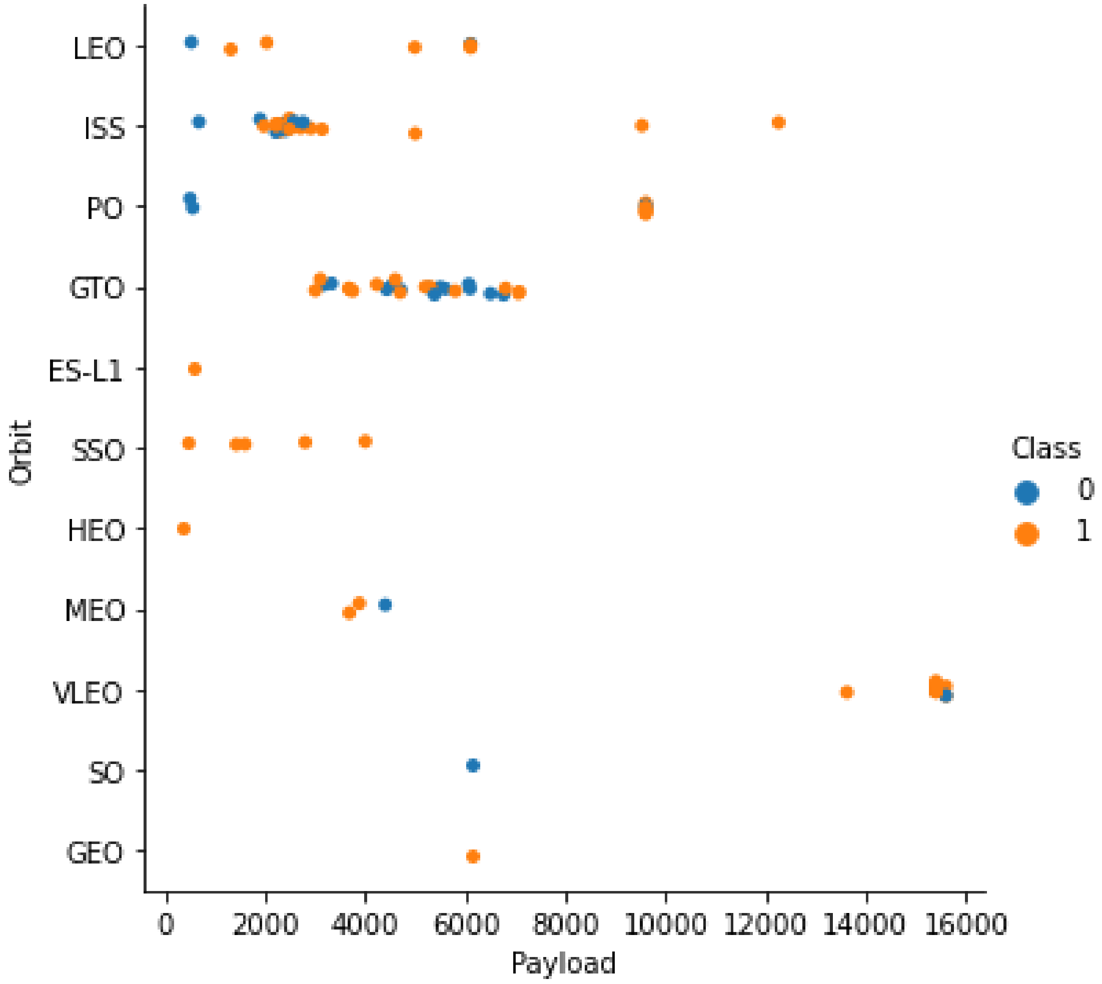
Flight Number vs. Orbit type

Not all types of Orbit have a clear relationship between success rate and flight number except for LEO or MEO where the flights with higher flight number seems to be more likely to be successful.



Payload vs. Orbit type

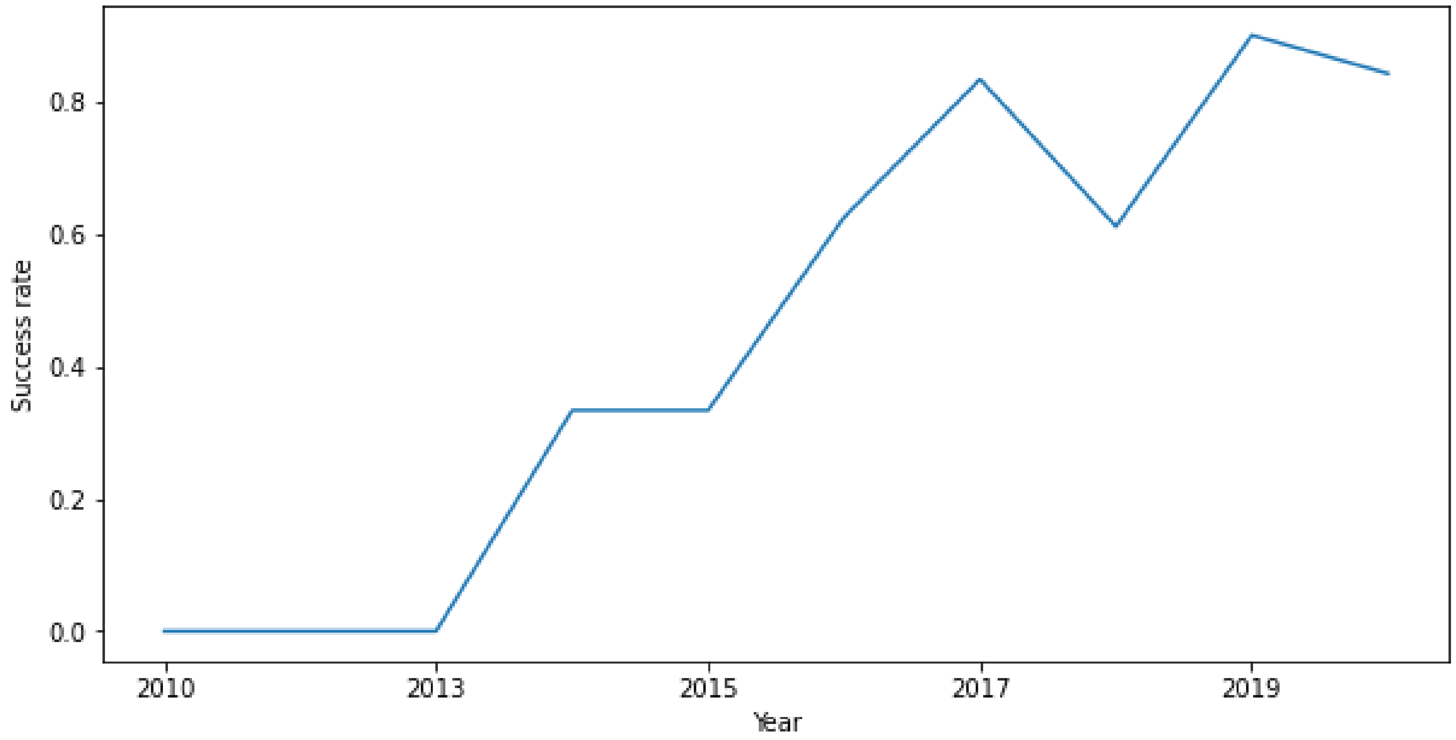
The relationship between Orbit types and Payload seems to be more clear. We can point out that for some orbits, the heavier are more likely to be success, for example LEO, ISS, PO.



Launch success yearly trend

We can point out that as the project grows further into the time, the more likely it is to be success in launching. Despite the slight fall of success rate to 60% in 2018, the success rate of 2019 reaches the peak of nearly 90%, a noticeable rate.

Generally speaking, the success rates are surely more than a half since 2016 until later on.



EDA with SQL

All launch site names

Query:

```
%%sql
SELECT UNIQUE(LAUNCH_SITE)
FROM SPACEXDATASET
```

Result:

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

Explanation: Simply select all unique values of “LAUNCH_SITE” columns using the SELECT query.

Launch site names begin with `CCA`

Query

```
%%sql
SELECT * FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

Result

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	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
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation: SELECT all available columns from table, with additional condition of site name begins with ‘CCA’ (using Wildcard character) inside WHERE clauseand limit the result with 5 first record using LIMIT statement

Total payload mass

Query

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)'
```

Result

1
45596

Explanation: Simply select all payload_mass of the launches of customer ‘NASA (CRS)’ (using WHERE clause to filter) and then using SUM aggregate function to add up the total payload mass.

Average payload mass by F9 v1.1

Query

```
%%sql
SELECT AVG(CAST(PAYLOAD_MASS__KG_ AS FLOAT))
FROM SPACEXDATASET
WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'
```

Result

1
2534.6666666666665

Explanation: SELECT all payload_mass filtered with booster_version ‘F9 v1.1’ only (using Wildcard character in WHERE statement) then using AVG aggregate function to calculate average. Casting the payload_masses into float in order for the average to be displayed in decimal.

First successful ground landing date

Query

```
%%sql
SELECT MIN(DATE)
FROM SPACEXDATASET
WHERE LANDING__OUTCOME LIKE '%Success%'
```

Result

1
2015-12-22

Explanation: Since DATE datatype is comparable in SQL, simple use MIN aggregate function to select the first date. The examining records are filter with "Success" landing outcome only using Wildcard Character inside WHERE statement.

Successful drone ship landing with payload between 4000 and 6000

Query

```
%%sql
SELECT PAYLOAD
FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

Result

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Explanation: SELECT the name and filter out appropriate Landing_Outcome and Payload_Mass inside WHERE statement.

Total number of successful and failure mission outcomes

Query

```
%%sql
SELECT * FROM
(SELECT COUNT(*) AS Success_Mission_Count FROM SPACEXDATASET WHERE MISSION_OUTCOME LIKE 'Success%') T1,
(SELECT COUNT(*) AS Failure_Mission_Count FROM SPACEXDATASET WHERE MISSION_OUTCOME LIKE 'Failure%') T2
```

Result

success_mission_count	failure_mission_count
100	1

Explanation: Execute the COUNT function for successful and failure missions in different tables, then INNER JOIN the tables to get the appropriate result.

Boosters carried maximum **payload**

Query

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
                           FROM SPACEXDATASET)
```

Result

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

Explanation: Using the sub-query to find out the maximum payload, then add the sub-query into the WHERE statement to filter out the records with maximum payloads and SELECT out the requirement attribute.

2015 launch records

Query

```
%%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Failure (drone ship)'
AND YEAR(DATE) = 2015
```

Result

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Explanation: Filter the year of launch by applying YEAR function on Date column and compare with 2015, then filter the appropriate Landing_Outcome also inside WHERE statement before SELECT the required attributes to display.

Rank landing outcome count between 2010-06-04 and 2017-03-20

Query

```
%%sql
SELECT LANDING__OUTCOME, COUNT(*) AS Number_Of_Landing
FROM SPACEXDATASET
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY Number_Of_Landing DESC
```

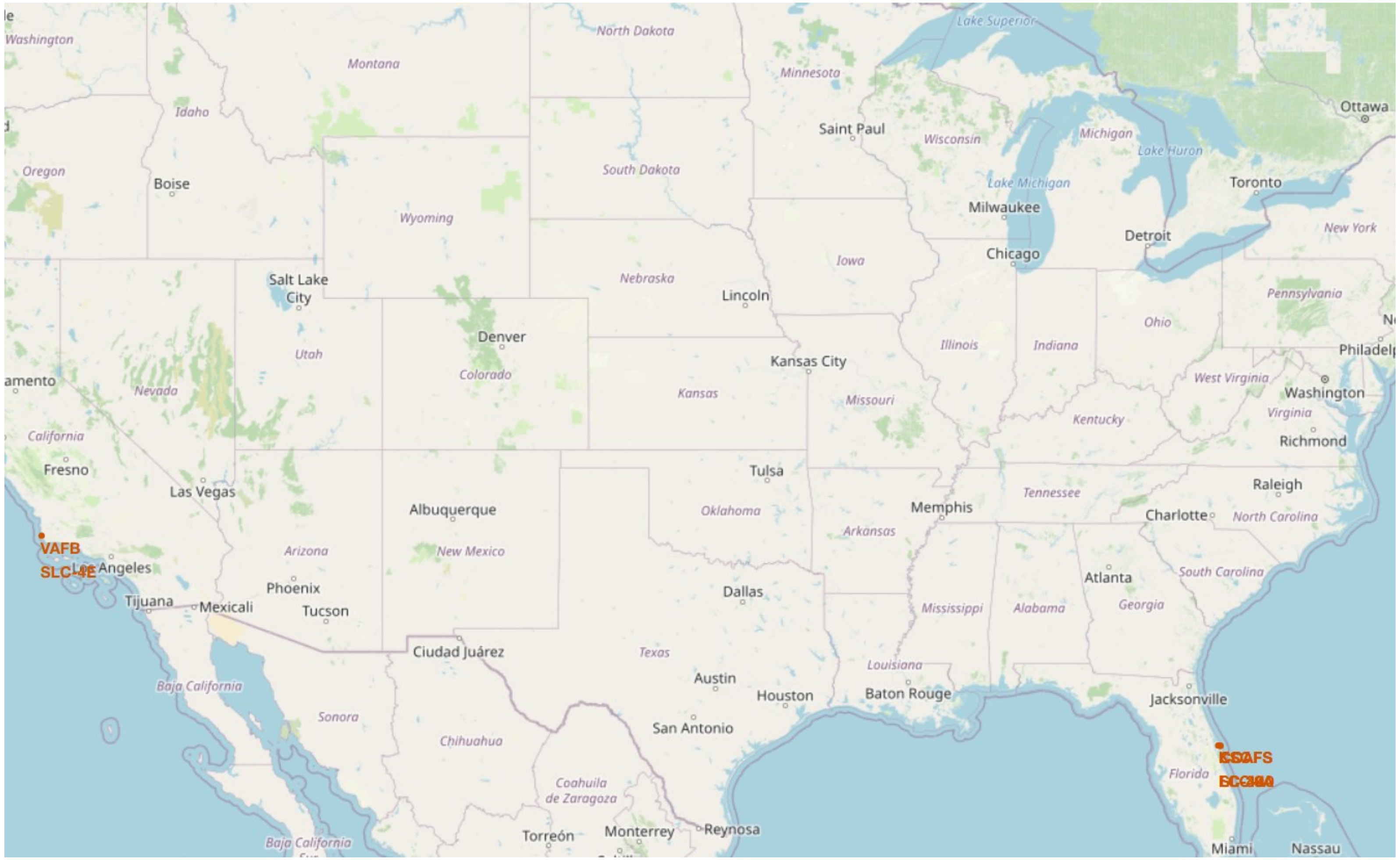
Result

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

Explanation: Select the landing outcomes, then use GROUP BY clause to group by that attribute and use COUNT aggregate function to retrieve the count of each outcome. Rank the displayed result by using ORDER BY clause with DESC option.

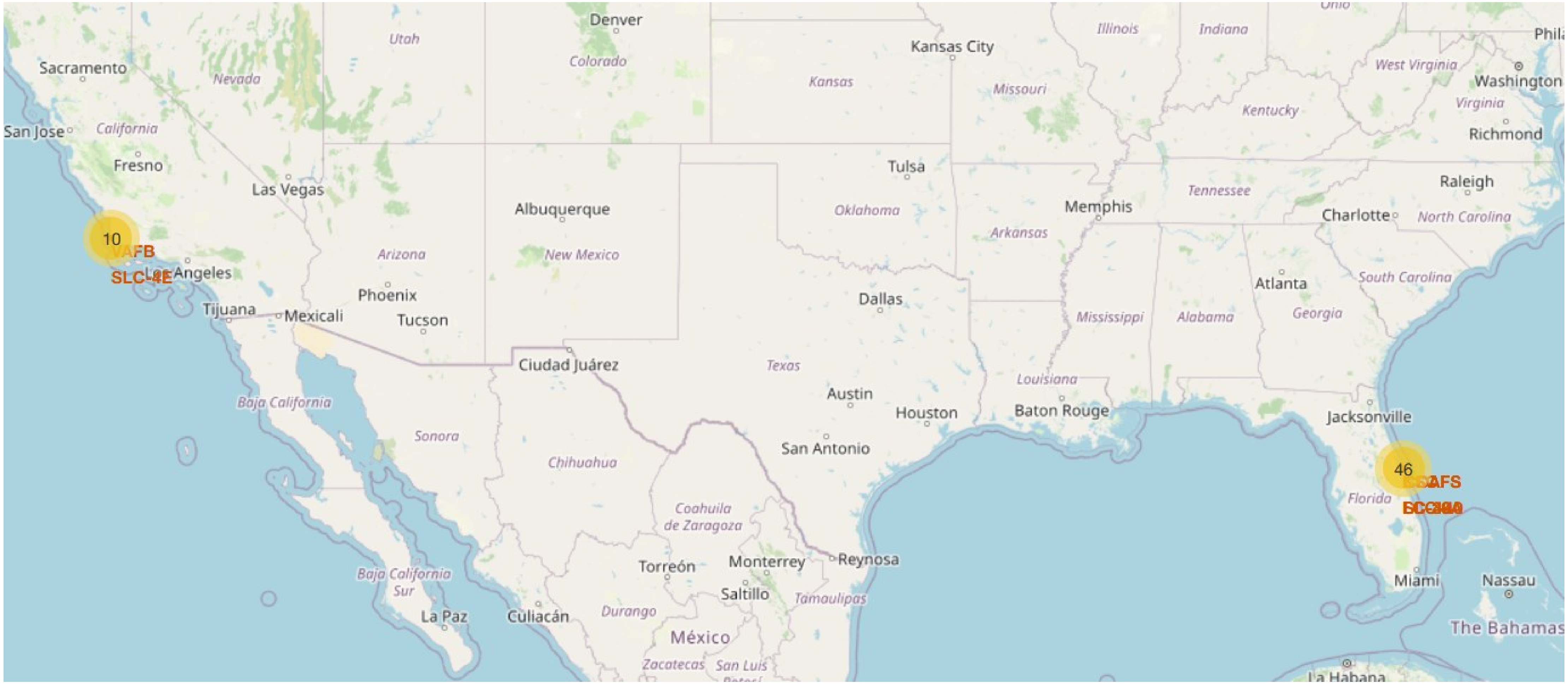
Interactive map with Folium

SpaceX Falcon 9 Launch Sites



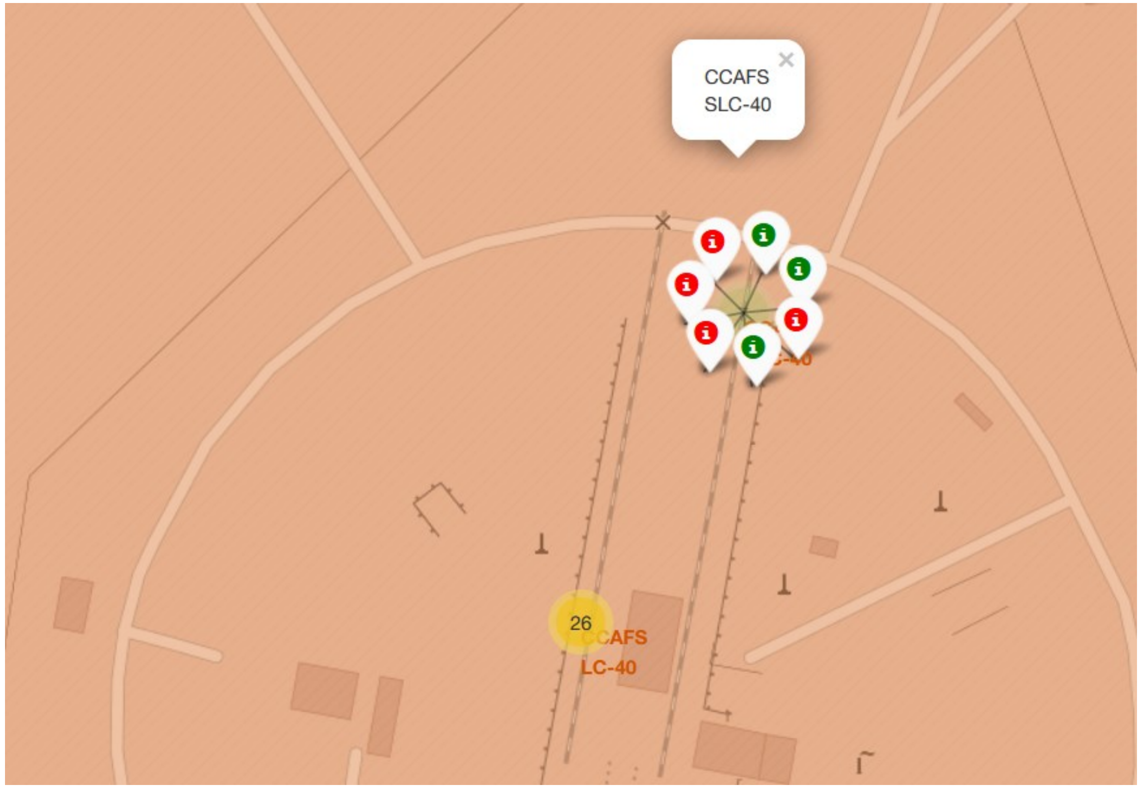
Although there are 4 different launch sites, they are all cluster around 2 positions. From this map, we can point out that Florida and California are the states where SpaceX chose to launch their rockets, mostly at locations very close to the coastline of the states.

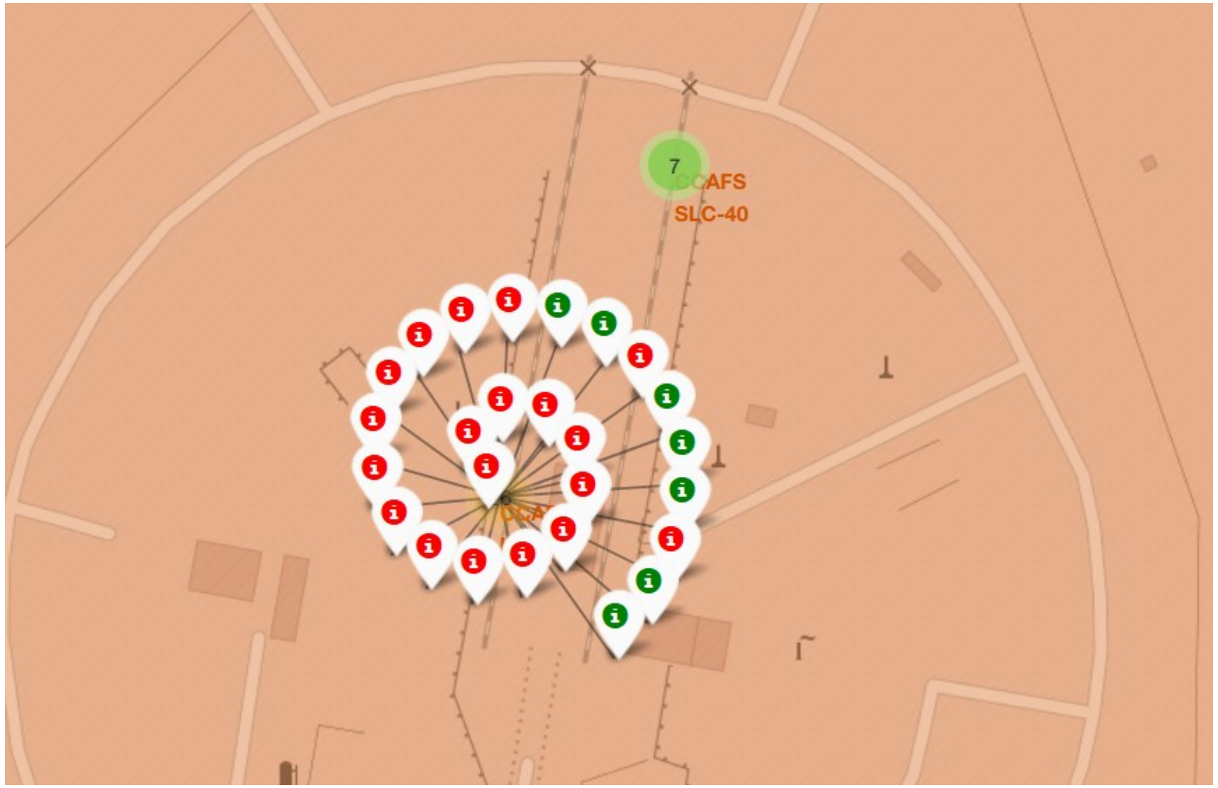
SpaceX Falcon 9 Launches Locations and Outcomes



From the map, it seems like the Eastern launch sites are preferrable by SpaceX for testing its rockets, which 46 launches were made their compared to only 10 launches in the Western launch site.

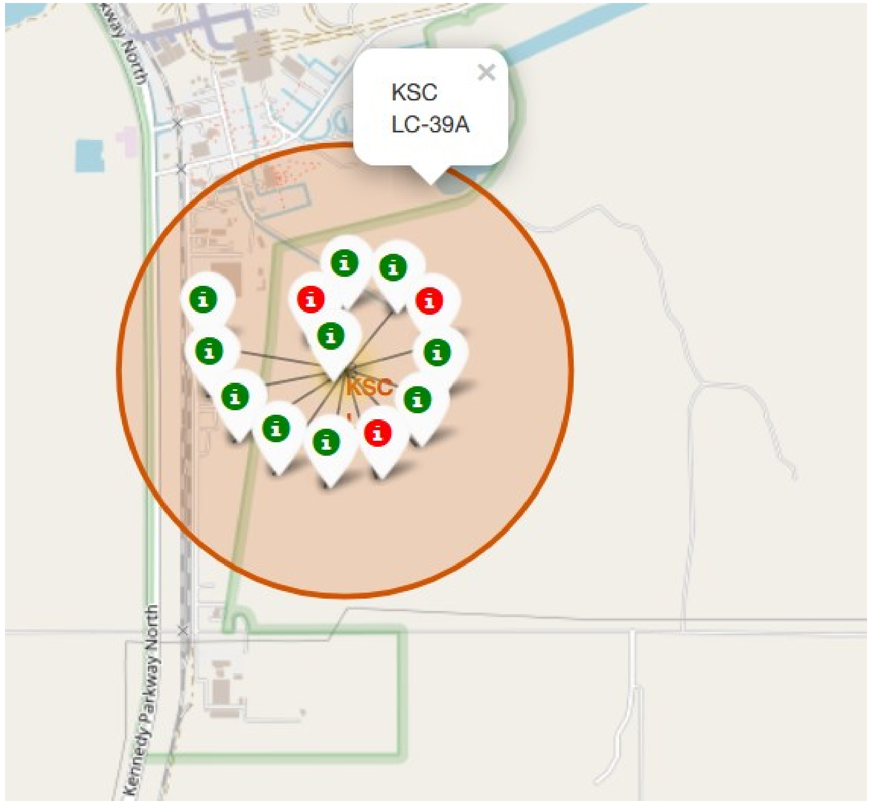
SpaceX Falcon 9 Launches Locations and Outcomes

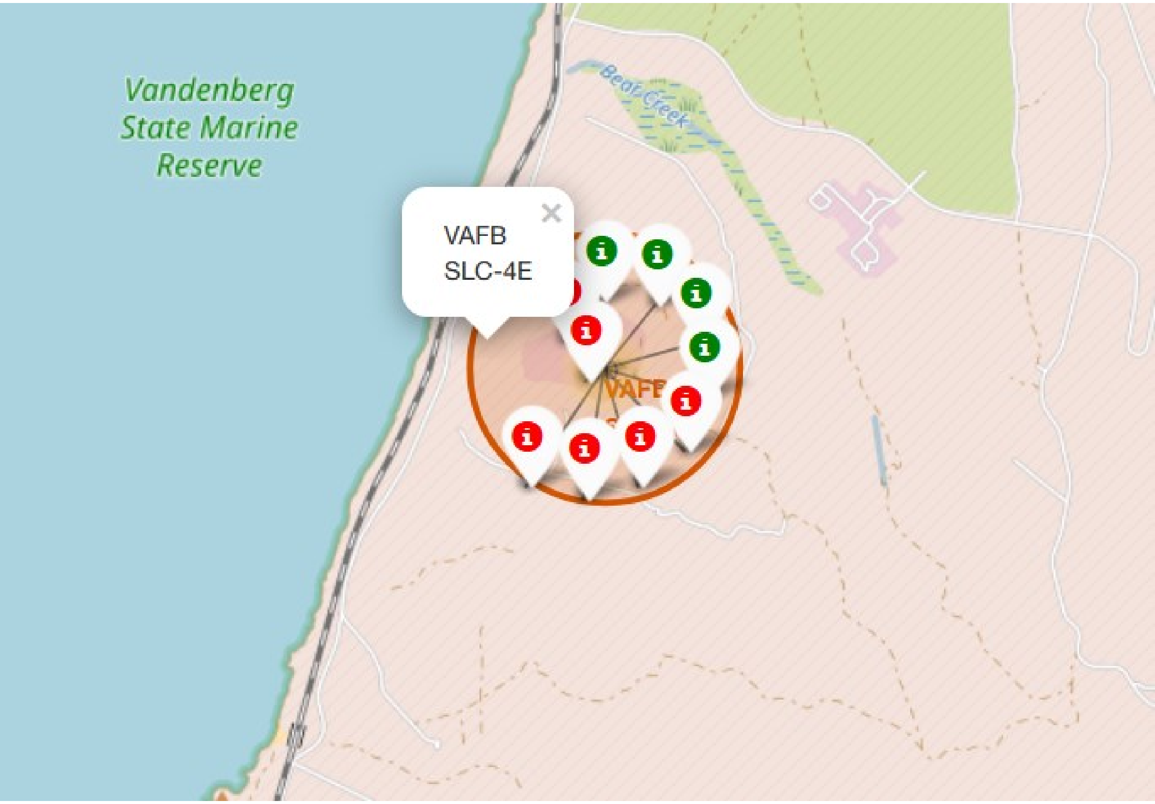




The group of nearby launch sites CCAFS LC-40 and CCAFS SLC-40 are the common one, fill up to 33 out of 56 launches. Although the success rate for those launch sites are considered quite low.

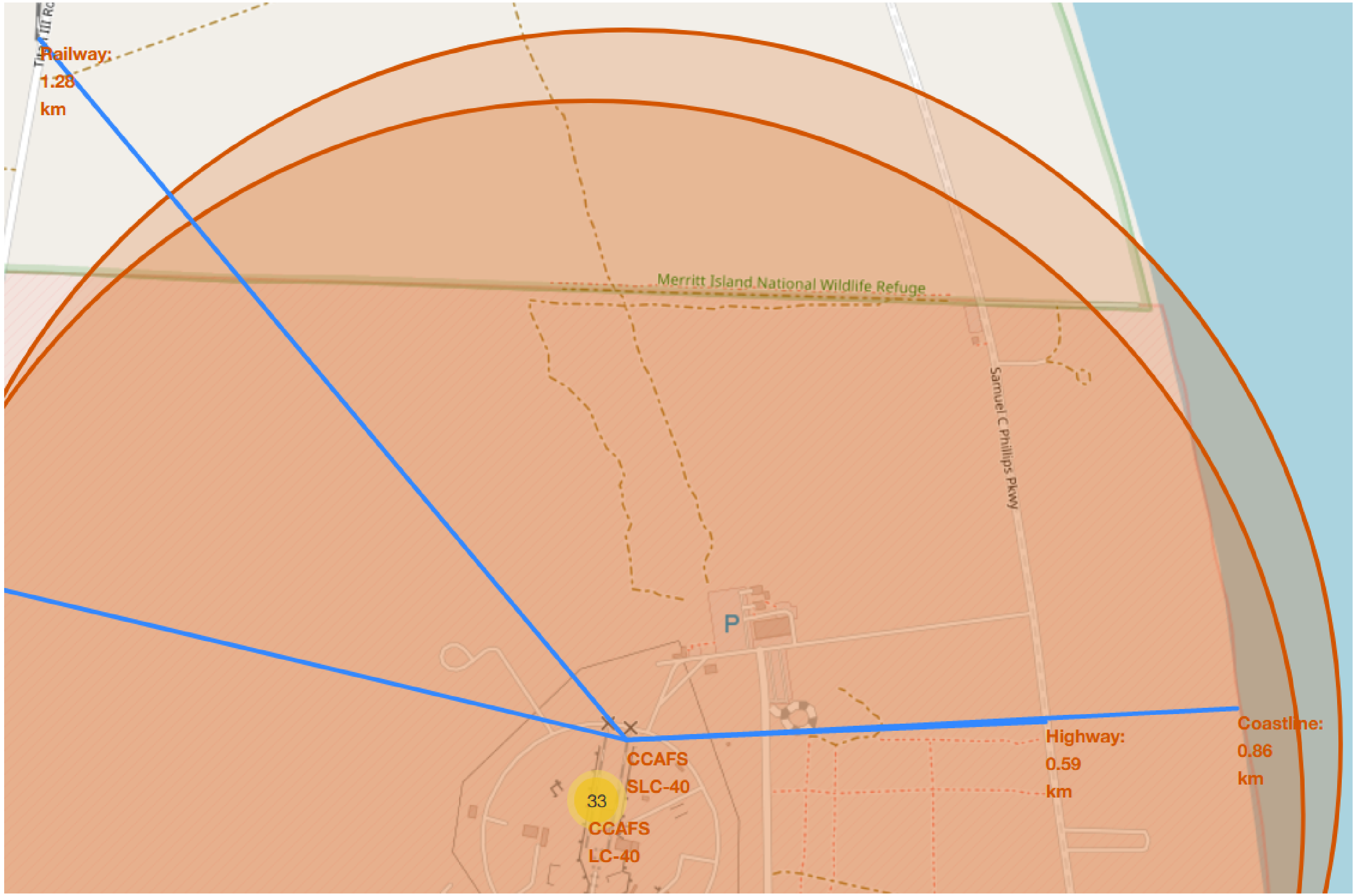
SpaceX Falcon 9 Launches Locations and Outcomes





The Eastern launch site also has a quite low success rate of less than 50% while for the remaining launch site in the Western cluster, the success rate is much higher. It can be concluded that *there could be a possibility that the launch sites closer to the land are likely to be more successful than the ones near the beaches and coastlines*.

Location of Launch Sites to Public Points



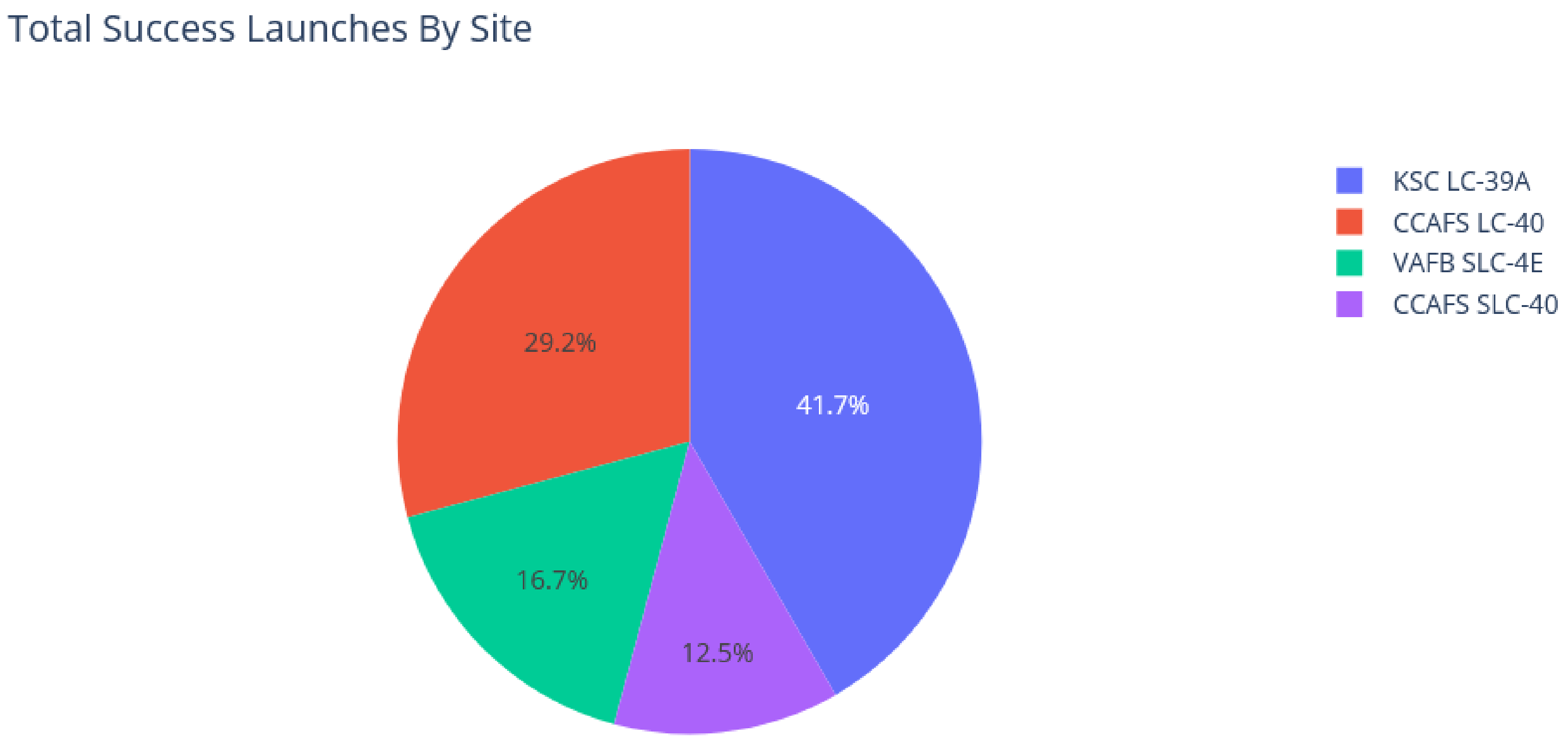
Examine the CCAFS SLC-40 launch site, we can see that the launch site is not very far from some nearby points like Highway, Coastline or Railway. However, almost all of those points are the private ones, not accessible by the public.

Location of Launch Sites to Public Points

The nearest city point of Titusville is located approximately 23 km away from the launch sites. We can point out that *the launch sites tend to be chosen at points that are far from the public*. This could be for the purpose of safety in case the launches are unsuccessful.

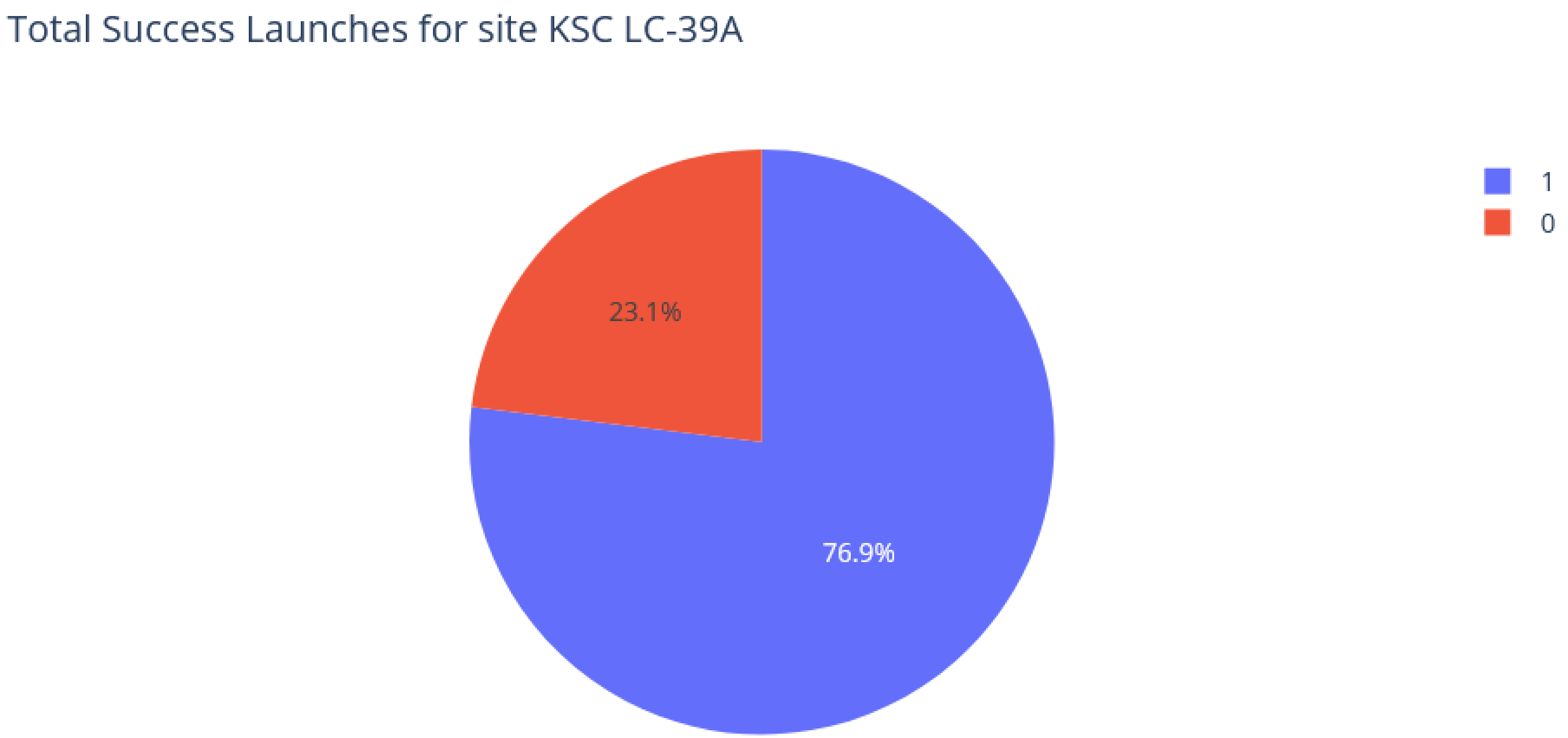
Build a Dashboard with Plotly Dash

Total Success Launches By All Sites



3 out of 4 “near-coastline” launch sites has an overall proportion rate of less than 30% (with CCAFS SLC-40 has a limited proportion of only one eighth of all success launches). However, KSC LC-39A in contrast has a noticeably proportion that takes up to more than 40% of all the successful launches.

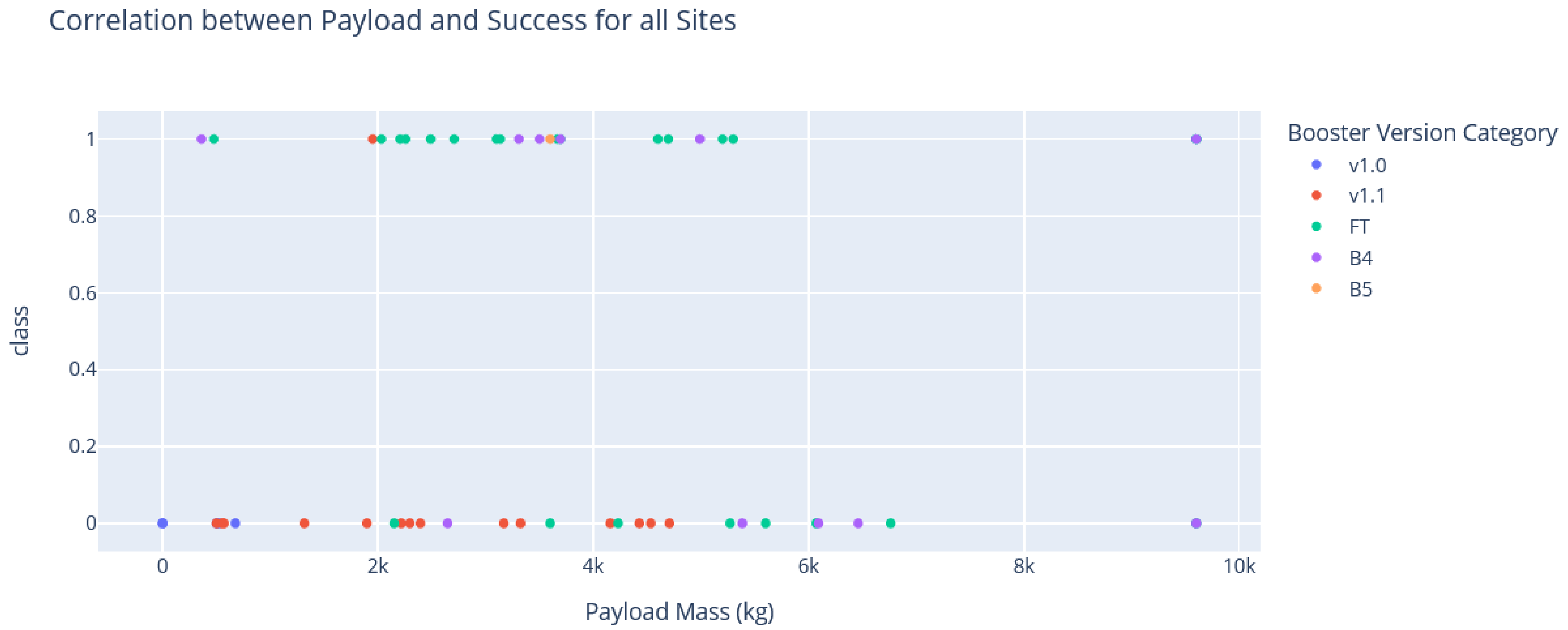
Total Success Launches for site KSC LC-39A



The success rate of site KSC LC-39A is the highest of all four, reaching up to *more than three fourth* of the total launches at this site.

Payload vs Launch Outcome for all Sites

Payload range (Kg):



Generally in the whole payload range, FT seems to be the booster version that are more likely to be successful while v1.1 in contrast has quite a lot of failed launches.

Payload vs Launch Outcome for all Sites

The payload range 1000 – 6000 kg seems to be the common for all launches. The distribution of outcome in this range is similar to the overall range, with FT being the most successful one.

Payload vs Launch Outcome for all Sites

However for the high range from 6000 – 10000, the success rates decrease rapidly as the number of launches in this range is very limited. From 6000 to 9000 only 2 booster versions are available and all of the launches were failed (even FT). Not until a little higher than 9500 that there is a successful launch of booster version B4.

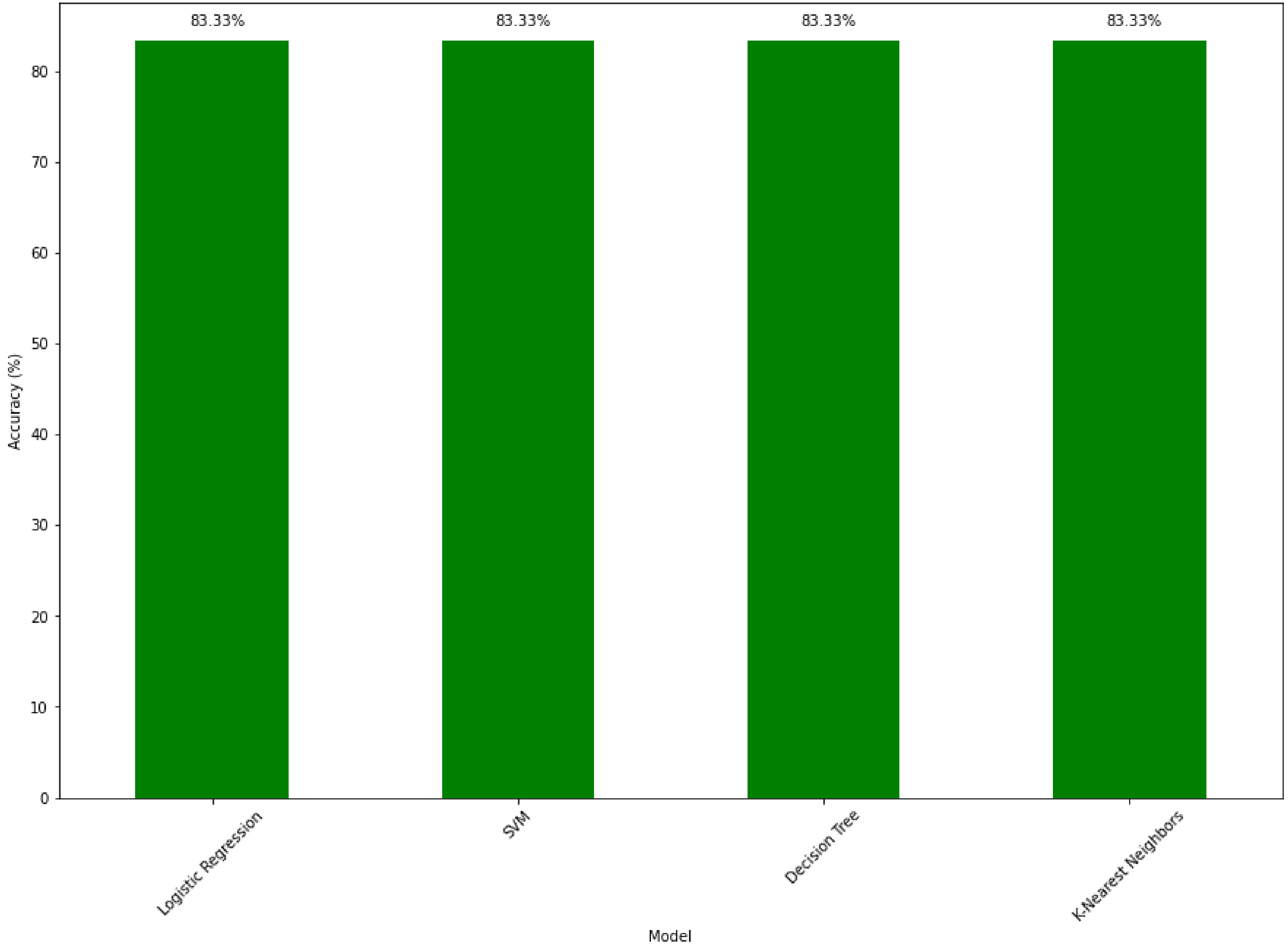
Predictive analysis (Classification)

Classification Accuracy

Surprisingly, the performance of all four used models are equal to the others, reaching **83.33% accuracy** on test set.

This can be because the size of data set is quite small, so after splitting only 18 samples were included in the test set.

A larger data set can be useful for showing the difference in accuracy between the 4 models.

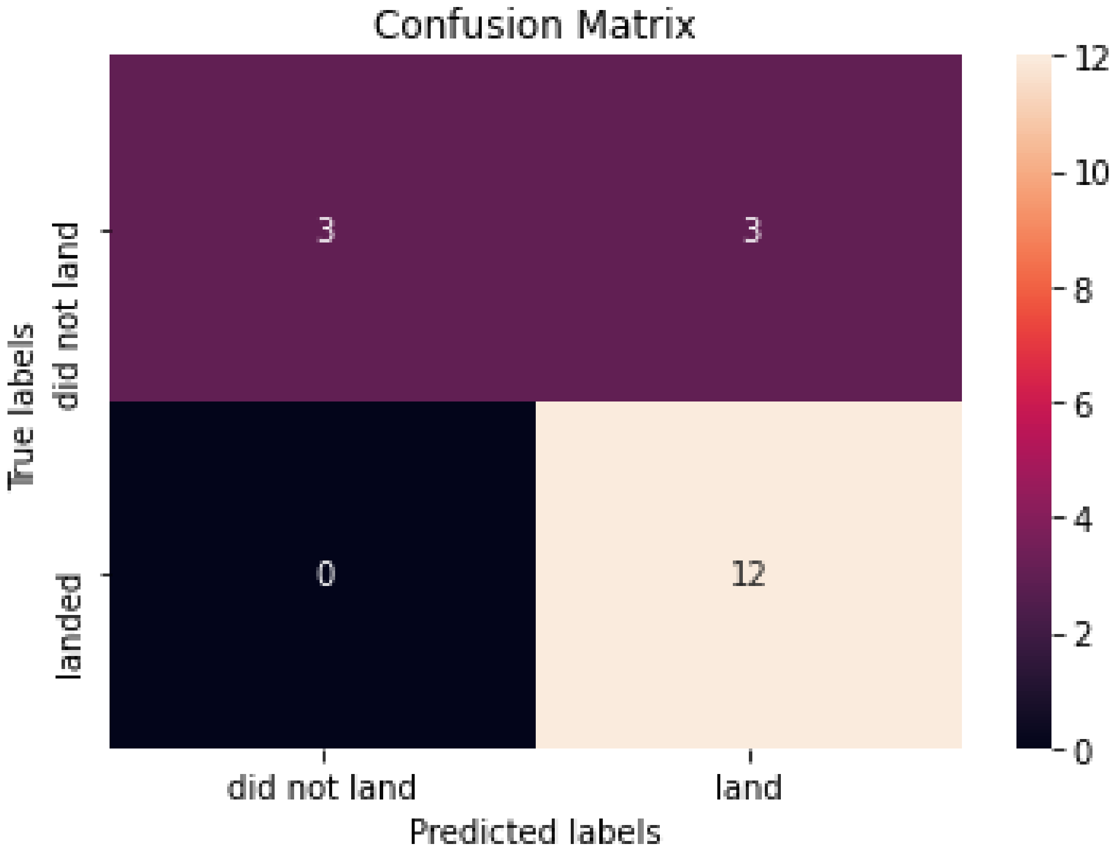


Confusion Matrix

Similarly, the Confusion Matrices of the 4 models are identical to the others.

In general, all the models can predicted quite accurately the test samples, where **all landed (successful) launches were predicted correctly**.

However, optimization can be made for all the models as among the 6 "unsuccessful" models, only 50% were correctly predicted. This lead to the case that the models tend to *predict "landed" labels many more times then "did not land" labels* (although the distribution of labels in training set are quite equals), which cause some serious drawbacks and **decrease the result of the model on the different evaluation methods**.



CONCLUSION

- In conclusion, we can point out some points that can better predict whether a launch of Falcon 9 will be successful or not, in detailed:
  - The heavier the payload, the more likely it is to be a successful launch. This is not true for the payload range of above 6000 kg.
  - ES-L1, GEO, HEO and SSO are the orbits that are more likely to be successful.
  - If the launch site is further from the coastline, then there is a higher chance that the launch will be successful.
  - With lower payload range (common range), FT seems to be the booster version with highest success rate.

APPENDIX

- Github Repository: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project.git
- Notebooks:
  - Data Collection with API: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/Data%20Collection.ipynb
  - Data Collection with Web Scraping: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

APPENDIX

- Data Wrangling: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/Data%20Wrangling.ipynb
- EDA with Visualization: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/EDA%20with%20Visualization.ipynb
- EDA with SQL: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/EDA%20with%20SQL.ipynb

APPENDIX

- Visual Analytics with Folium: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/Visual%20Analytics%20with%20Folium.ipynb
- Plotly Dashboard: https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/spacex_dash_app.py
- Machine Learning Prediction

  https://github.com/nguyencaonhan271201/IBM-Data-Science-Certificate-Capstone-Project/blob/main/Machine%20Learning%20Prediction.ipynb