IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Becerra Flores Jose Eduardo
24 November 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Summary of Methodologies:

Data from SpaceX launches **was collected and preprocessed,** including payload mass, launch site, booster version, orbit type, and landing outcomes. Exploratory data analysis, **SQL queries**, and **visualizations (scatter, follium and pie charts)** were used to identify **patterns**. **Machine learning models** (Logistic Regression, SVM, Decision Tree, KNN) were trained and tuned with GridSearchCV to predict successful first-stage landings.

Summary of most important Results:

- Launch site KSC-LC-39A shows the highest success rate (76.9%).
- Most successful launches occur with payloads between 1,000–6,000 kg and flight numbers above 40.
- Logistic Regression proved the most reliable model for predicting landing success.

Introduction

Context

Today, companies are making **space travel more affordable. SpaceX** has been the most successful in this regard, primarily by **reusing the first stage of their rockets** when they land safely. This approach generates significant cost savings.

Objective

In this study, we aim to predict whether SpaceX will **successfully reuse the first stage** of a rocket. Instead of relying on physical experiments, we approach this problem using data science and machine learning models.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and Web Scraping in Wikipedia.
- Perform data wrangling
  - Transforming categorical features to numerical and converting data to the right type.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Each model was build using functions from sklearn library, tune with them own parameters and using GridSearch to obtain the best parameters and evaluate with the function score().

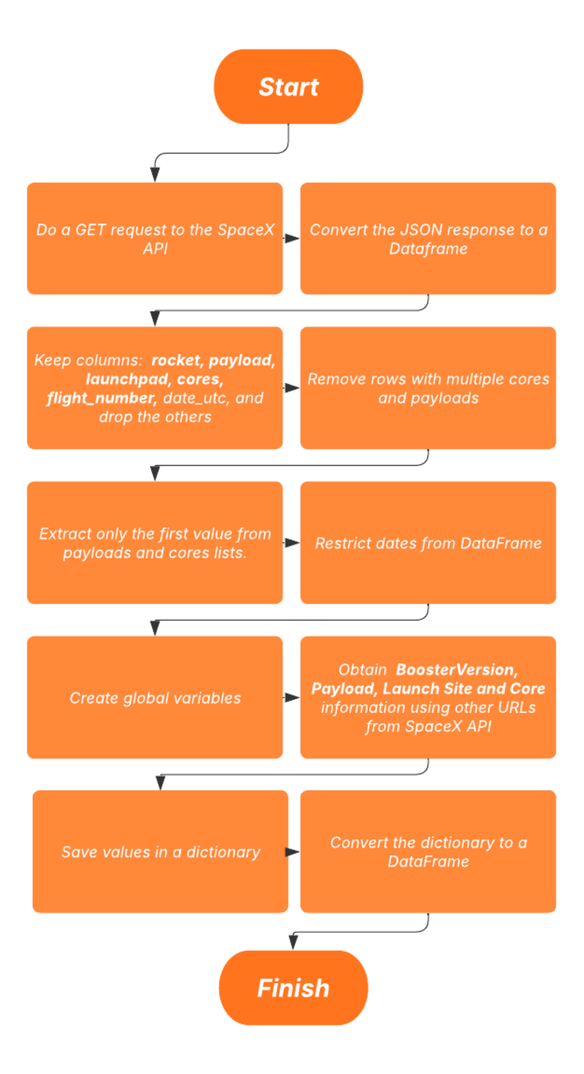
Methodology

- Parameters for classification models
  - Logistic Regression. C, Penalty, Solver.
  - Support Vector Machine. Kernel, C, Gamma.
  - Decision Tree Classifier. Criterion, Splitter, Max depth, Max features, Minimum samples leaf, Minimum samples split.
  - K- Nearest neighbors. N-Neighbors, Algorithm, P.

Data Collection – SpaceX API

We made a **GET request** to the **SpaceX API** to obtain the **JSON data** and then converted it into a **DataFrame**. Finally, using additional URLs and our extracted data, we obtained the required information for this exercise.

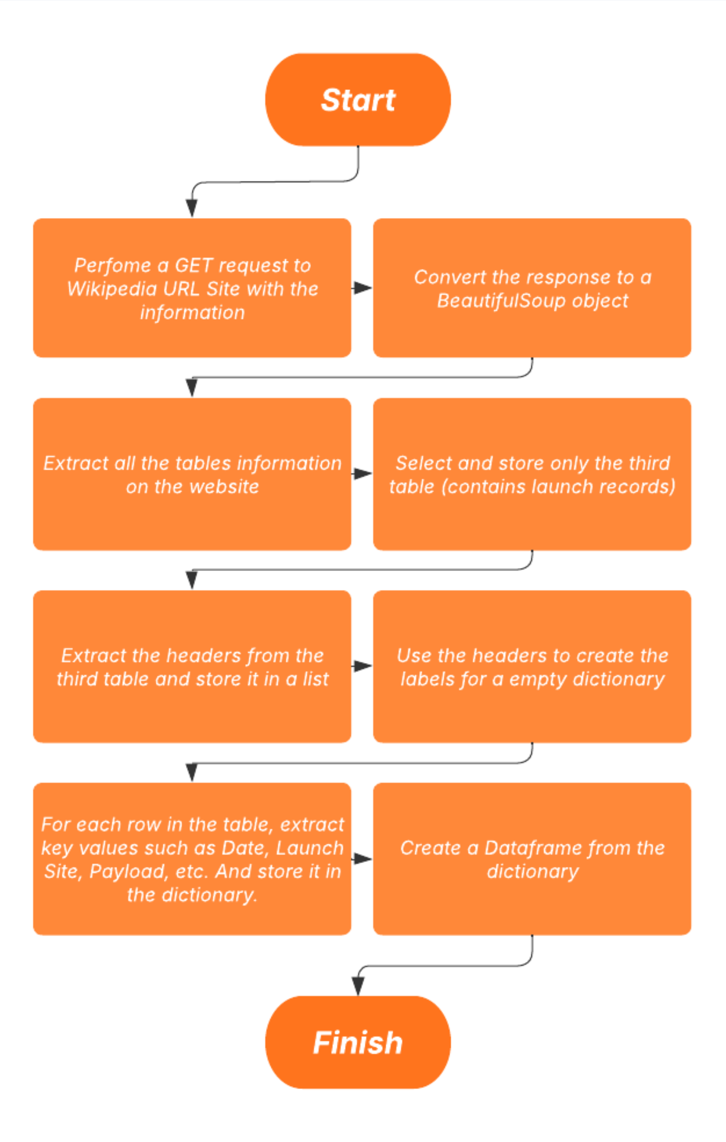
- GitHub URL: https://github.com/Edu4rd02/IBM-SpaceX/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

We send a **GET request** to the Wikipedia page and parse the HTML using **BeautifulSoup**. Then we extract **all tables**, select the **third one**, and obtain its headers to build a dictionary. Next, we loop through each row to collect key launch information and store it in the dictionary. Finally, we convert **the dictionary into a DataFrame**.

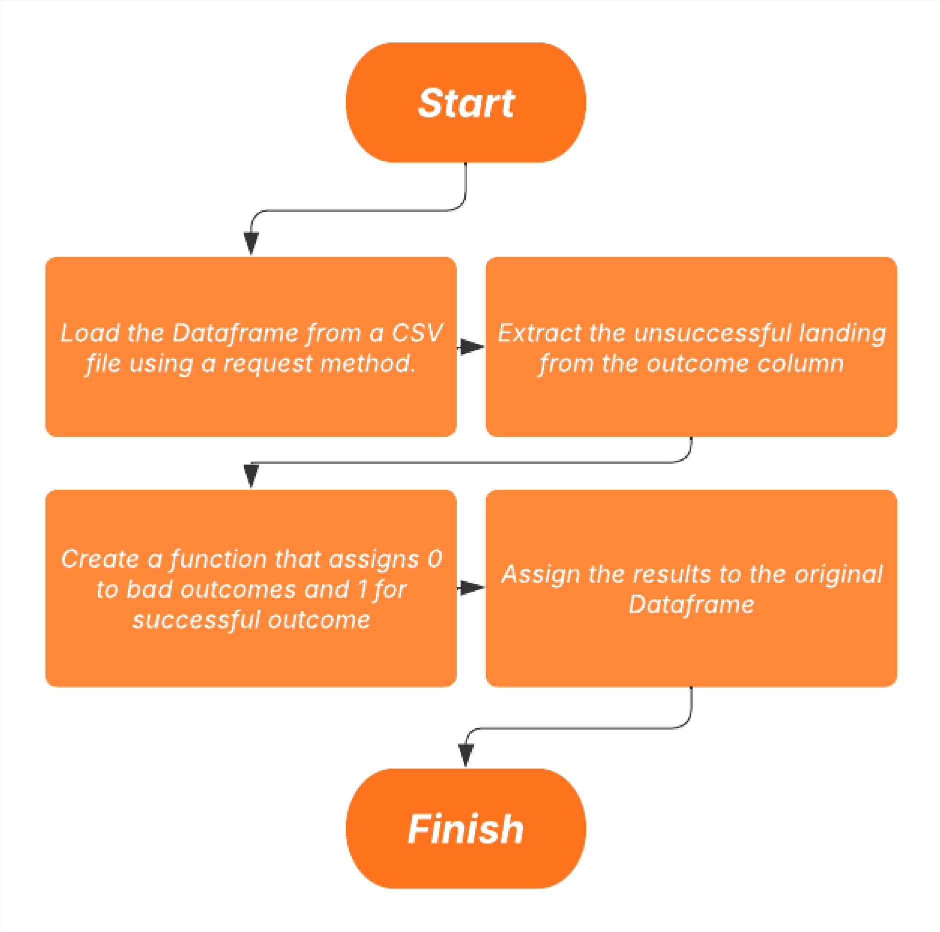
- GitHub URL: https://github.com/Edu4rd02/IBM-SpaceX/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

Using a **Dataframe from a previous section**, we extracted the successful and bad returns from the outcome column. **Then we assigned a value** from each categorie, **0 to unsuccessful** and **1 to great returns** from first stage.

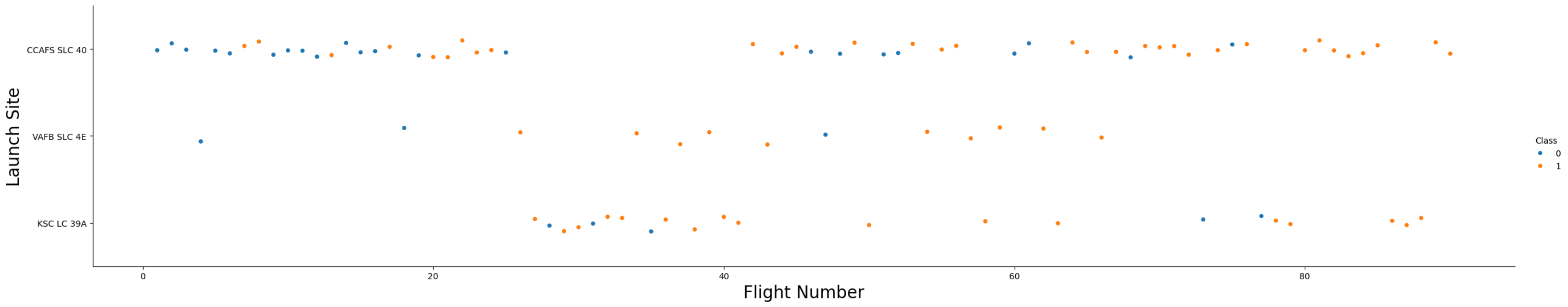
- GitHub URL: https://github.com/Edu4rd02/IBM-SpaceX/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



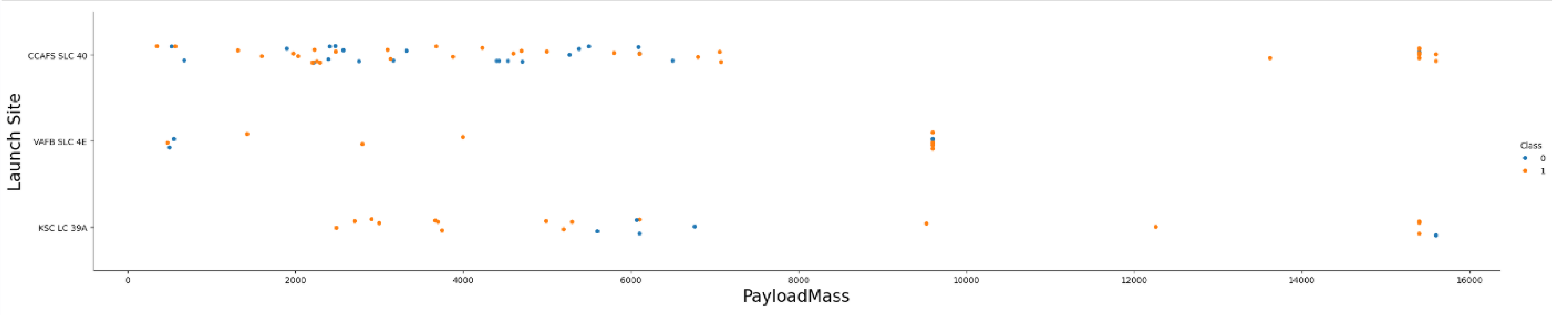
EDA with Data Visualization

- **Relation between Flight Number and Launch Site (Scatter plot).**

Scatter plots are useful exploring relations between two variable, makes so easy view trends, patterns and correlations.



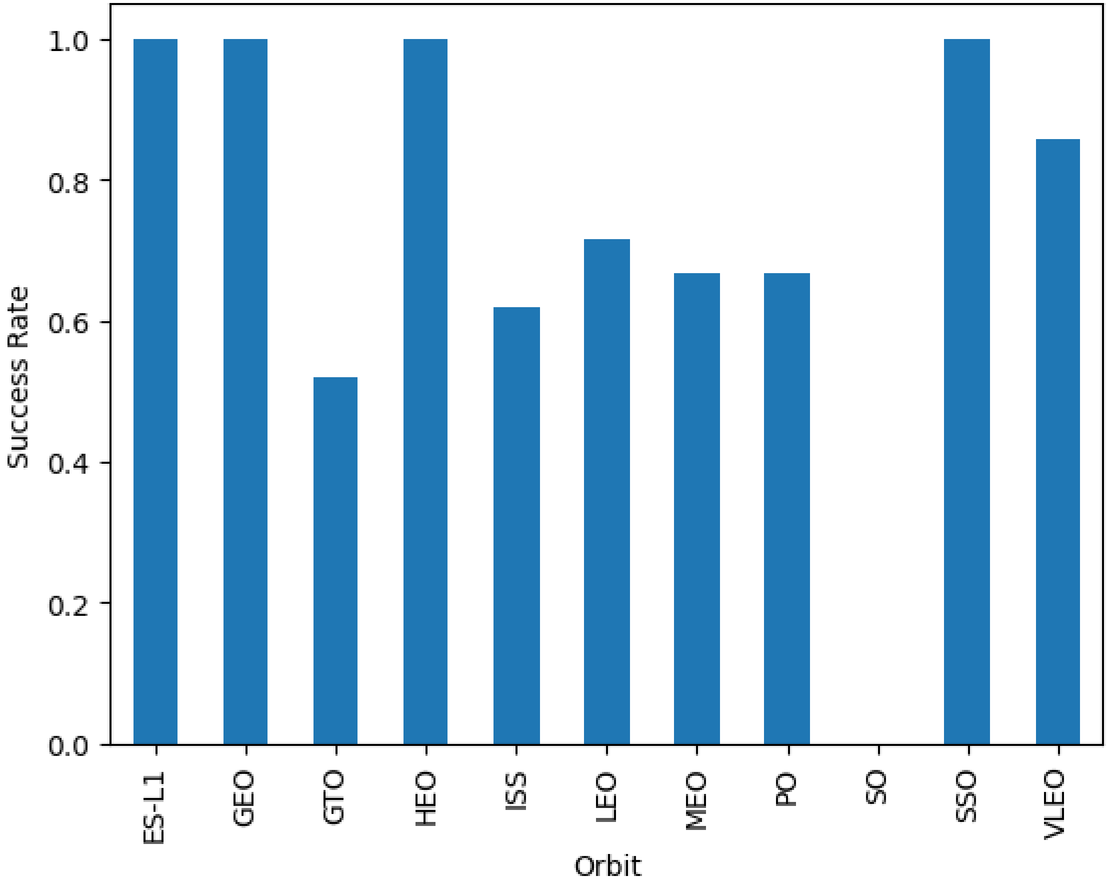
- **Relation between Payload Mass and Launch Site (Scatter plot).**



EDA with Data Visualization

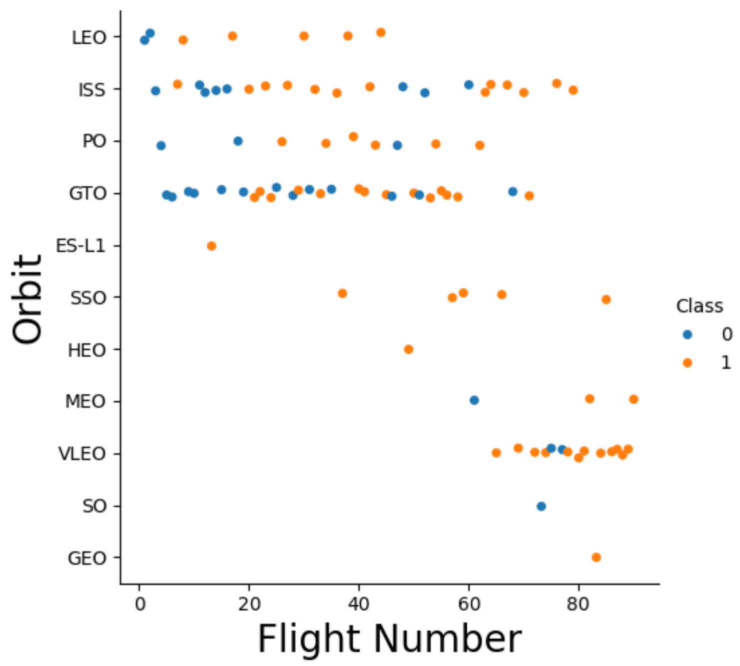
- **Relation between Success Rate and Orbit type (Bar chart).**

Bar charts helps to compare categories, and his values or means.

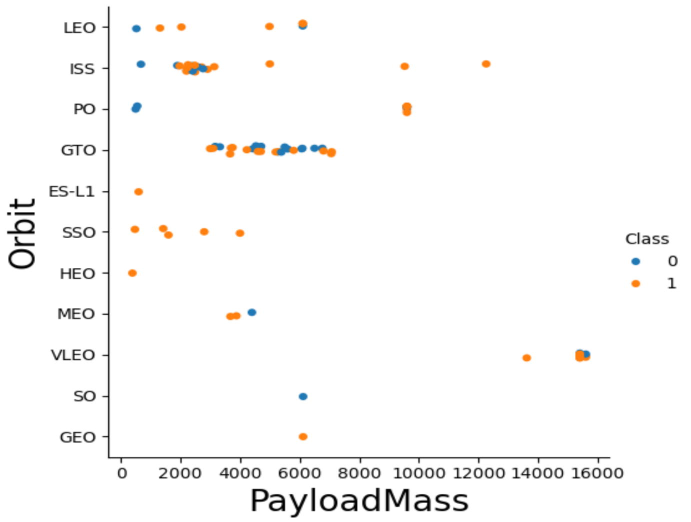


EDA with Data Visualization

- **Relation between Flight Number and Orbit type (Scatter plot).**

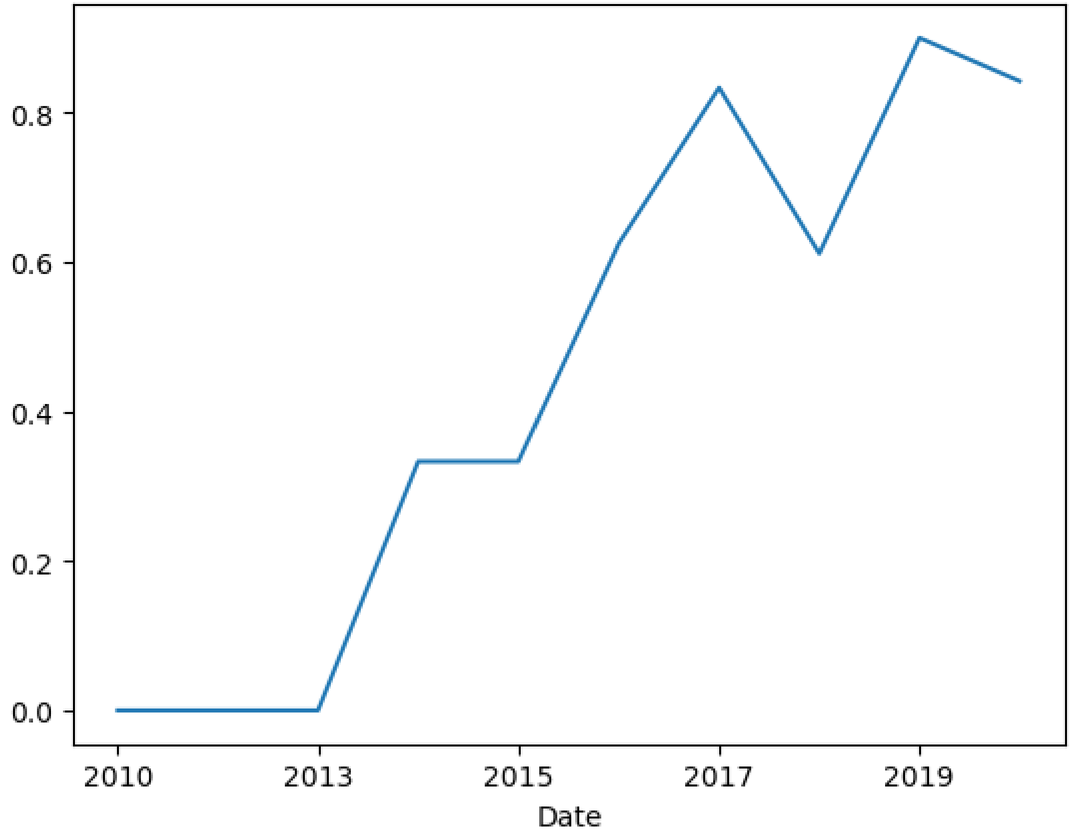


- **Relation between Orbit type and PayLoad Mass**



EDA with Data Visualization

- **Launch Success yearly trend**



- Github link: https://github.com/Edu4rd02/IBM-SpaceX/blob/main/edadataviz.ipynb

EDA with SQL

Display the names of the unique launch sites in the space mission

Query: *SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE*

- Select all the distinct values or names from Launch Site feature.
- Return a list of all the distinct names from Launch Site.

Display 5 records where launch sites begin with the string ‘CCA’

Query: *SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5*

- Select all the rows on the table.
- Filters only the Launch sites names that start with ‘CCA’.
- Return a list of all table rows where Lauch Sites which start with ‘CCA’.

EDA with SQL

Display the total payload mass carried by boosters launched by NASA (CRS)

Query: *SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer == 'NASA (CRS)’*

- Calculate the sum of payload mass values.
- Filters only the rockets launched by ‘NASA (CRS)’.
- Return a single number that is the total payload mass.

List the date when the first successful landing outcome in ground pad was achieved

Query: *SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome == 'Success (ground pad)’*

- Select the lowest value in the row Date.
- Filters only the Landing Outcome in ground pad that was success.
- Return a value that is the date of the first successful landing in ground pad.

EDA with SQL

Display the total payload mass carried by boosters launched by NASA (CRS)

Query: *SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer == 'NASA (CRS)’*

- Calculate the sum of payload mass values.
- Filters only the rockets launched by ‘NASA (CRS)’.
- Return a single number that is the total payload mass.

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Query: *SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome == 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_<6000)*

Selects the Booster_Version from the table.

- Filters rows where the Landing_Outcome is ‘Success (drone ship)’.
- Includes only launches with a payload mass between 4000 and 6000 kg.
- Returns the booster versions that meet both conditions (successful drone-ship landing and payload in that range).

EDA with SQL

Display the total payload mass carried by boosters launched by NASA (CRS)

Query: *SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer == 'NASA (CRS)’*

- Counts how many missions were labeled as successful using Mission_Outcome LIKE 'Success%’.
- Counts how many missions were labeled as failures using Mission_Outcome LIKE 'Failure%’.
- Uses CASE WHEN inside SUM() to convert each row into a 1 or 0 depending on the condition.
- Returns two numeric values: Success_Count and Failure_Count.

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

Query: *SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)*

- Selects the Booster_Version from the table.
- Uses a subquery to find the maximum payload mass in the entire dataset.
- Retrieves the booster version(s) associated with that highest payload mass.
- Returns the booster(s) that carried the heaviest payload ever recorded in the table.

EDA with SQL

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

Query: *SELECT substr(Date,6,2) as Month_Num, substr(Date,0,5) as Year, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE (Date >= '2010-01-01' AND Date<='2015-12-31') AND Landing_Outcome = 'Failure (drone ship)')*

- Extracts the month number from the Date column using substr(Date, 6, 2).
- Extracts the year from the Date column using substr(Date, 0, 5).
- Filters records where the date is between 2010 and 2015 (inclusive).
- Selects only rows where the Landing_Outcome is ‘Failure (drone ship)’.
- Returns the month number, year, landing outcome, booster version, and launch site for those failed drone-ship landings.
- Returns the booster(s) that carried the heaviest payload ever recorded in the table.

EDA with SQL

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.

Query: *SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTABLE WHERE (Date >= '2010-06-04' AND Date<='2017-03-20') GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC*

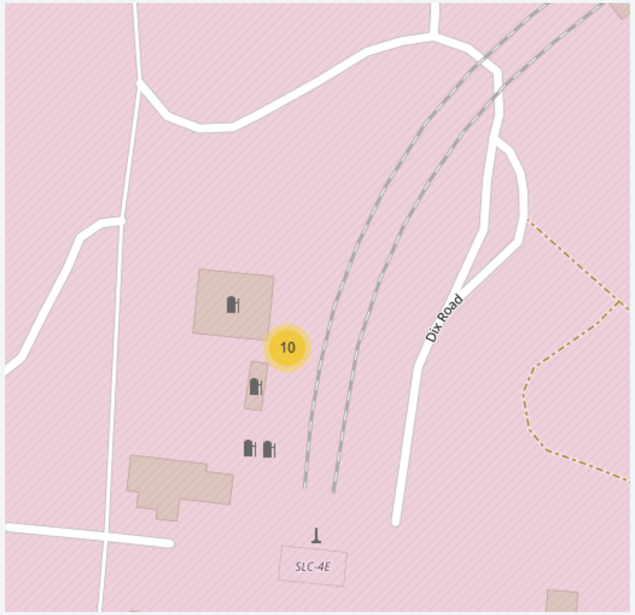
- Counts how many records fall under each Landing_Outcome category.
- Filters rows with dates between June 4, 2010 and March 20, 2017.
- Groups the results by Landing_Outcome to produce one count per category.
- Orders the results in descending order, showing the most frequent outcomes first.

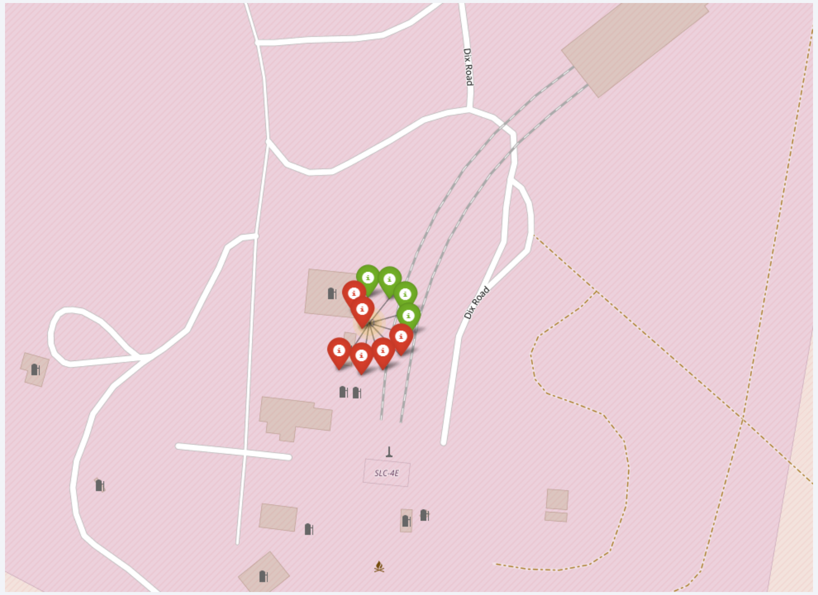
Github link: https://github.com/Edu4rd02/IBM-SpaceX/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Follium is a Python library for **geographic data visualization.** In these maps, you can add elements such as markers, circles, lines, text, etc.

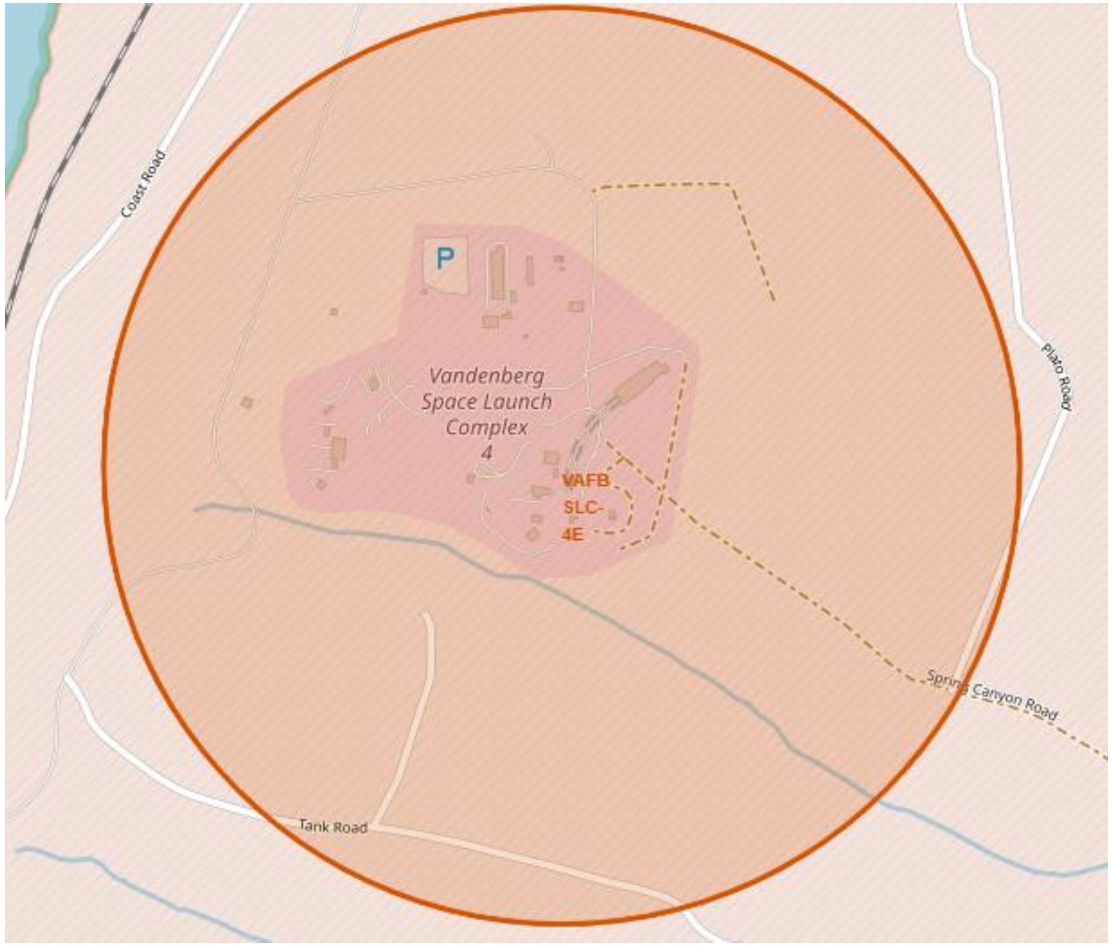
Markers. They are used to show whether each launch in the Dataframe was success (green) or failure (red).





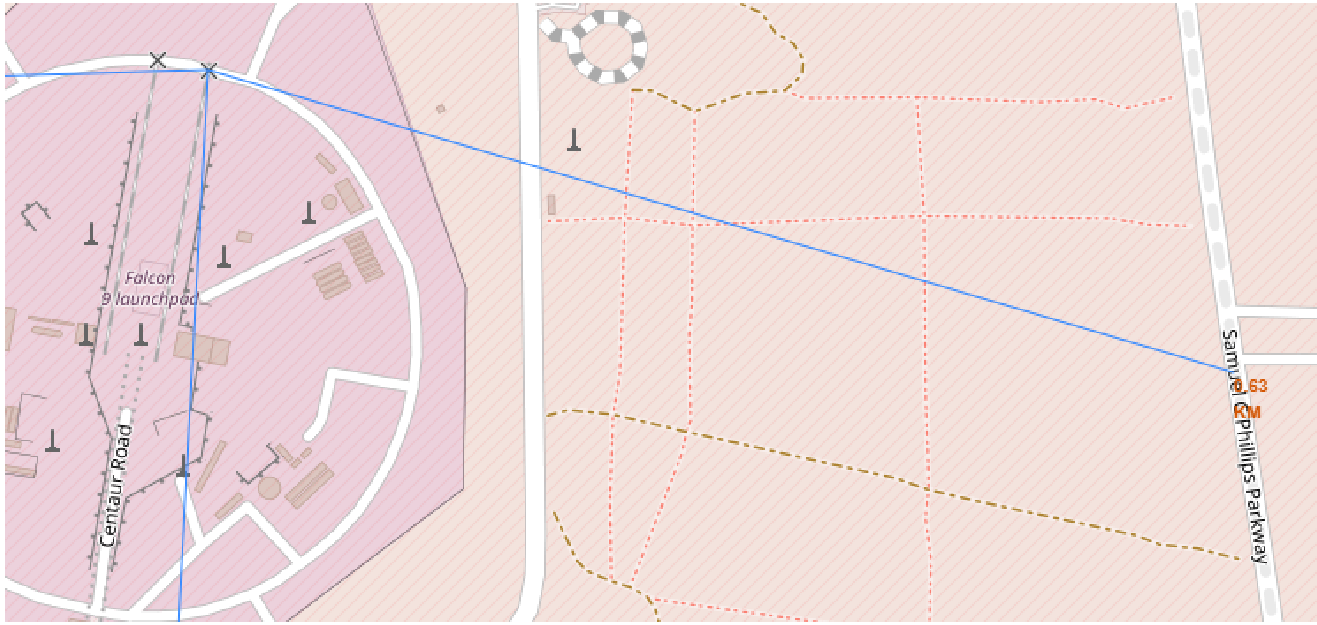
Build an Interactive Map with Folium

Circles. They are used to show the area for each Launch sites.



Build an Interactive Map with Folium

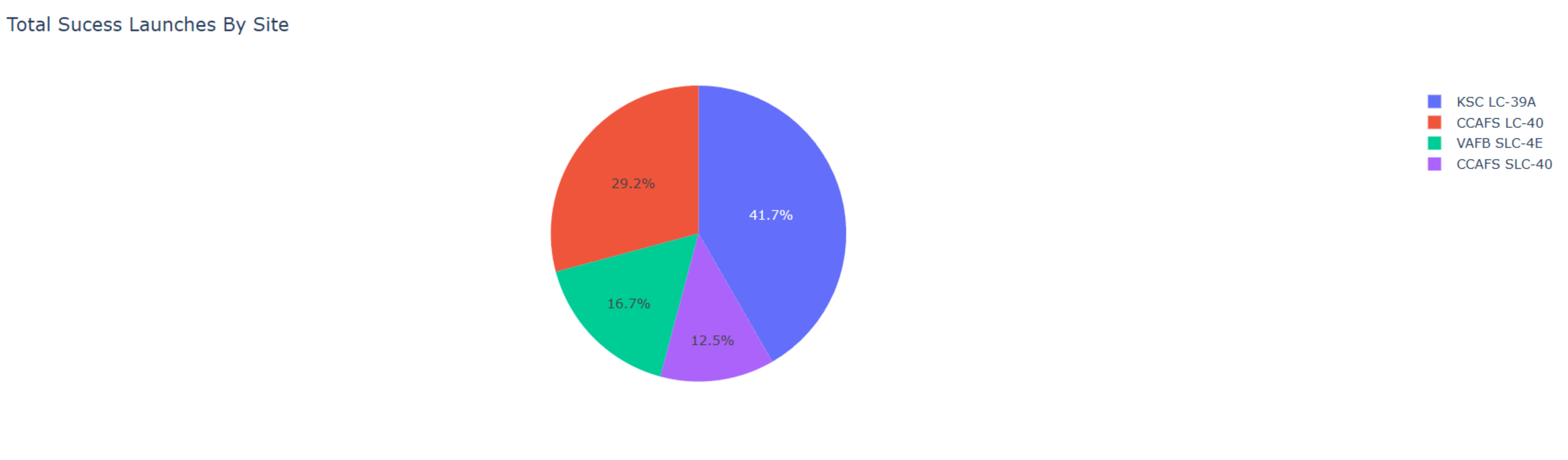
Lines and texts. Were used to show distance between launch sites and other elements like coasts, cities or highways and display the number of kilometers.



- GitHub URL: https://github.com/Edu4rd02/IBM-SpaceX/blob/main/DV0101EN-Exercise-Generating-Maps-in-Python.ipynb

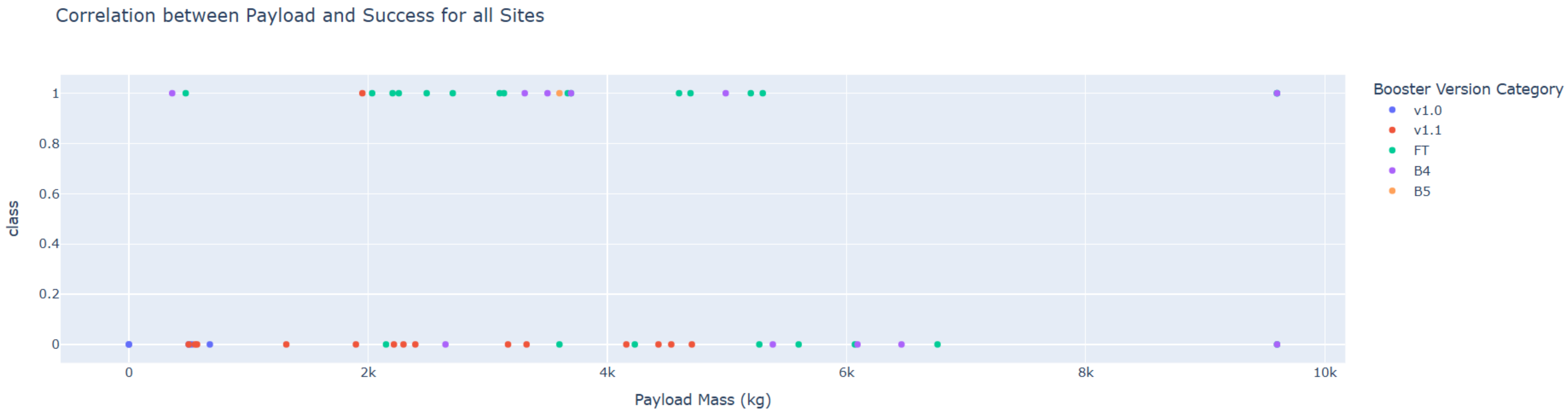
Build a Dashboard with Plotly Dash

We added a pie chart to more easily visualize which launch sites have a better success rate and their corresponding percentages.



Build a Dashboard with Plotly Dash

We also added a scatter plot to provide a different perspective and to help analyze and detect trends in the data.



Github link: https://github.com/Edu4rd02/IBM-SpaceX/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

Each model was built using functions from the **scikit-learn** library in Python. They were tuned with **GridSearchCV** to find the best hyperparameters and evaluated using the *.score()* method on the test dataset.

Although Logistic Regression, SVM, and KNN achieved similar accuracy scores on the test data, Logistic Regression showed the highest accuracy on the training data.

Therefore, Logistic Regression is considered the best-performing model.

```
parameters = {'criterion': ['gini', 'entropy'],
     'splitter': ['best', 'random'],
     'max_depth': [2*n for n in range(1,10)],
     'max_features': ['auto', 'sqrt'],
     'min_samples_leaf': [1, 2, 4],
     'min_samples_split': [2, 5, 10]}

tree = DecisionTreeClassifier()
```

```
tree_cv = GridSearchCV( estimator=tree,
                        param_grid = parameters,
                         cv=10,
                         scoring='accuracy'
                        )

tree_cv.fit(X_train,Y_train)
```

Calculate the accuracy of tree_cv on the test data using the method `score`:

```
[24]: tree_accuracy = (tree_cv.score(X_test,Y_test))
      tree_accuracy
```

```
[24]: 0.7222222222222222
```

Predictive Analysis (Classification)

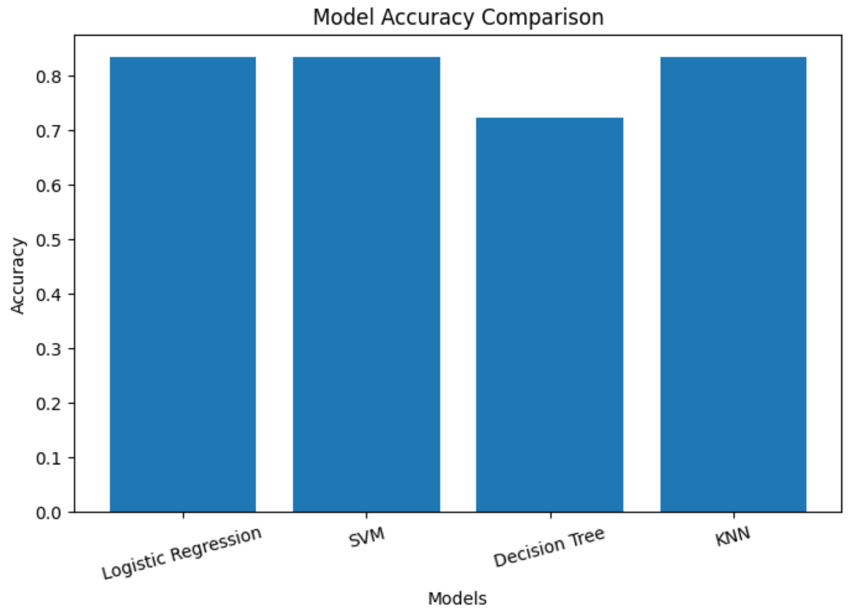
To find the best model, each best accuracy result for model was store in a variable and compare with the others to see who was the best.

- GitHub URL: https://github.com/Edu4rd02/IBM-SpaceX/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Find the method performs best:

```
[31]: best_model = max([
          ("Logistic Regression", logreg_accuracy),
          ("SVM", svm_accuracy),
          ("Decision Tree", tree_accuracy),
          ("KNN", knn_accuracy)
      ], key=lambda x: x[1])

      print(f"\nBest Model: {best_model[0]} with accuracy: {best_model[1]:.4f}")
```



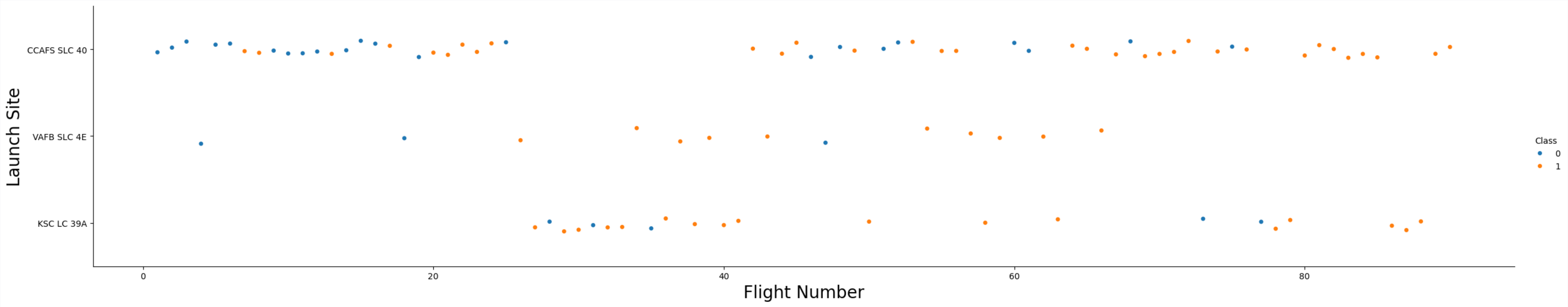
Results

- **Exploratory data analysis results**
  - Launch sites with more payload mass of flights tend to have higher success rates.
  - ES-L1, GEO, HEO and SSO were the orbits with the most successful rate.
  - The launch success rate increase in 2013 till 2020.
- **Interactive analytics demo in screenshots**
  - The best launch site was KSC LC-39A with a 41.7% in the total success.
  - The most successful launches occur when the payload mass is in the range of 1,000 to 6,000 kilograms.
- **Predictive analysis results**
  - The best model was Logistic Regression.

Section 2

Insights drawn from EDA

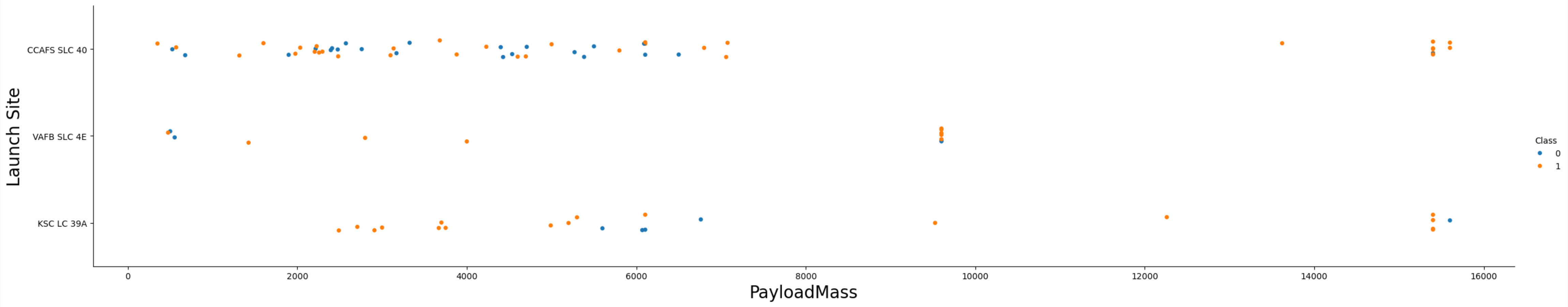
Flight Number vs. Launch Site



This is a scatter plot that shows relationship between Flight Numbers and Launch Sites. The insights are:

- Launch sites with a larger number of flights tend to have higher success rates.
- VABF SLC 4E is the launch site with less flight numbers.

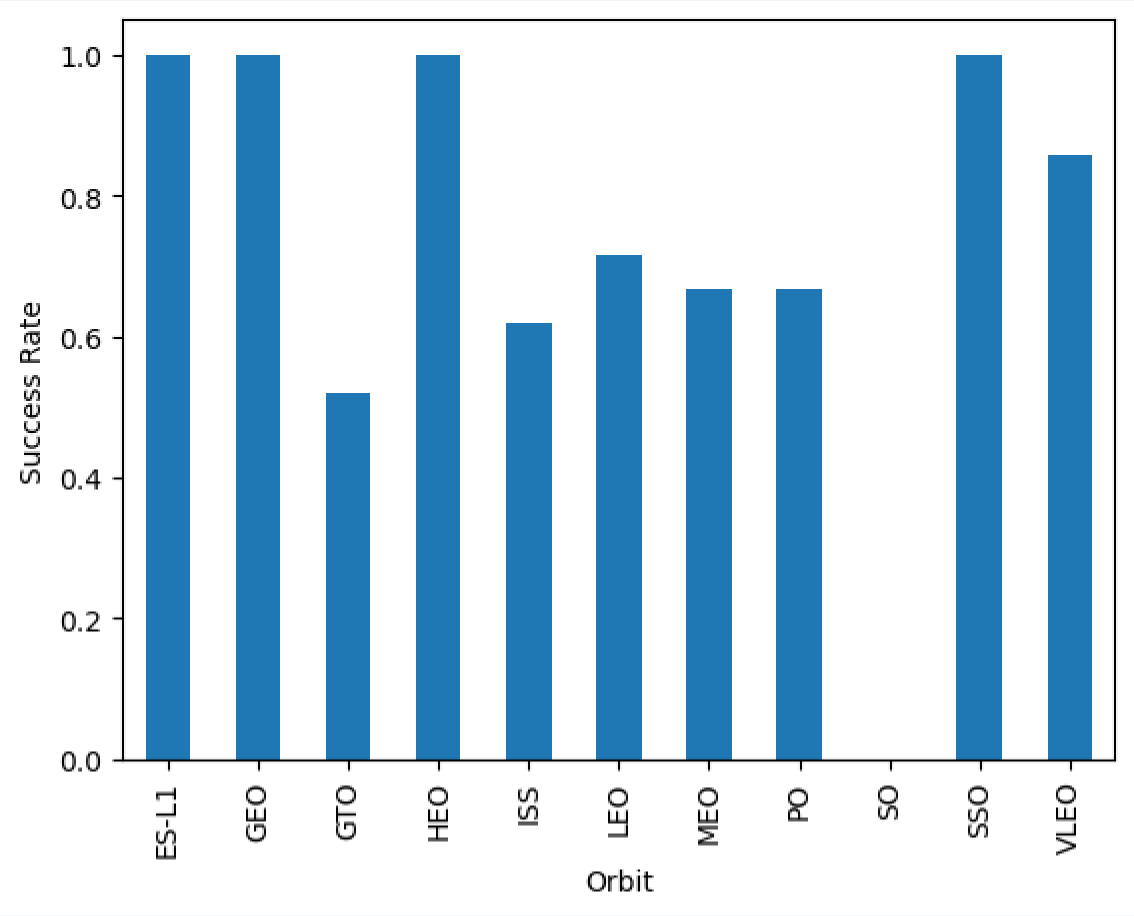
Payload vs. Launch Site



This is a scatter plot that shows relationship between Payload and Launch Sites. The insights are:

- Launch sites with more payload mass of flights tend to have higher success rates.
- VAFB-SLC 4E launch site doesn't have rockets launched for a payload mass greater than 10000.

Success Rate vs. Orbit Type

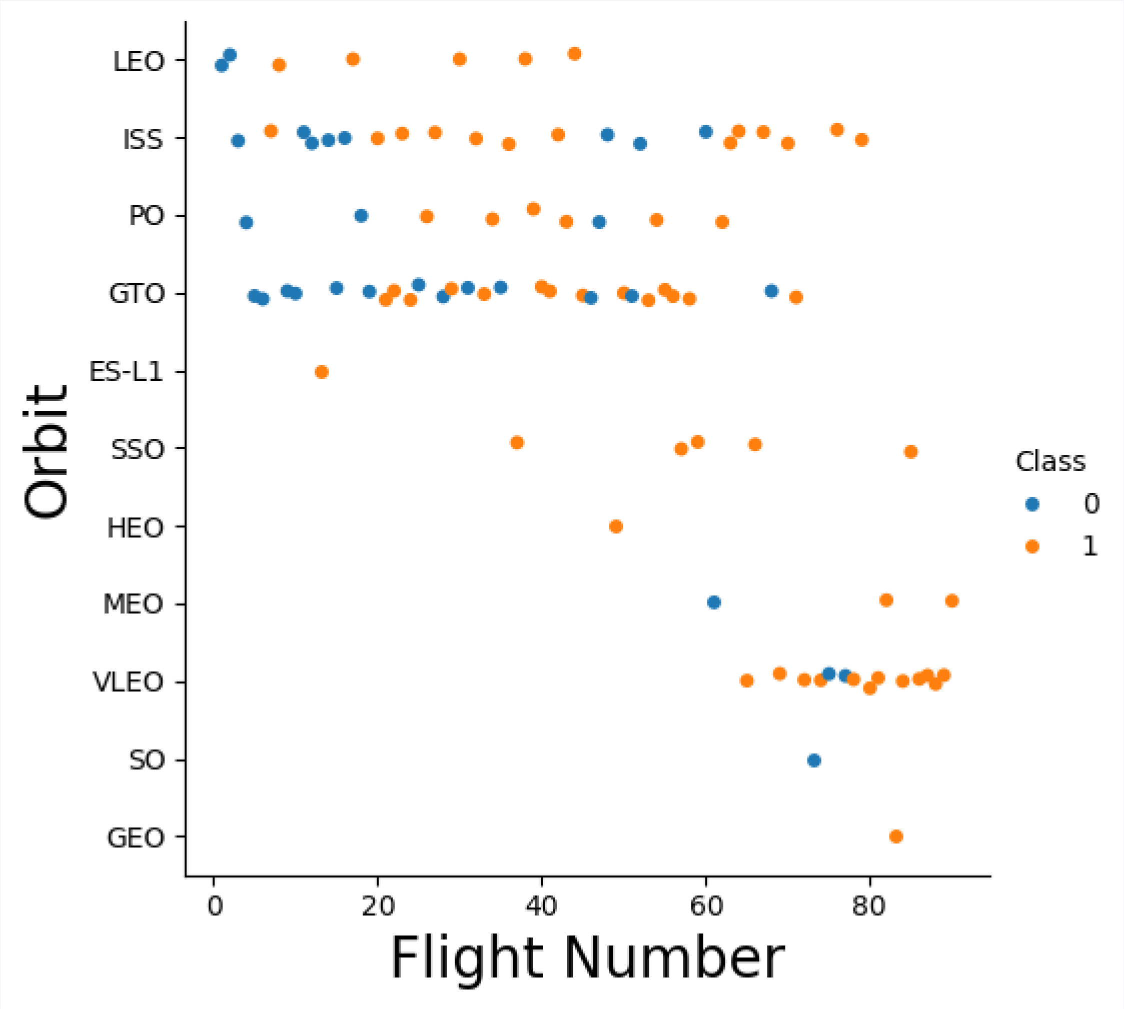


- This is a bar chart that shows relationship between Orbit type and Success Rate. The insights are:
- ES-L1, GEO, HEO and SSO are the orbit type with most success rate in all the dataset.
- GTO has the worst success rate.

Flight Number vs. Orbit Type

This is a scatter plot that shows relationship between Orbit Type and Flight Numbers. The insights are:

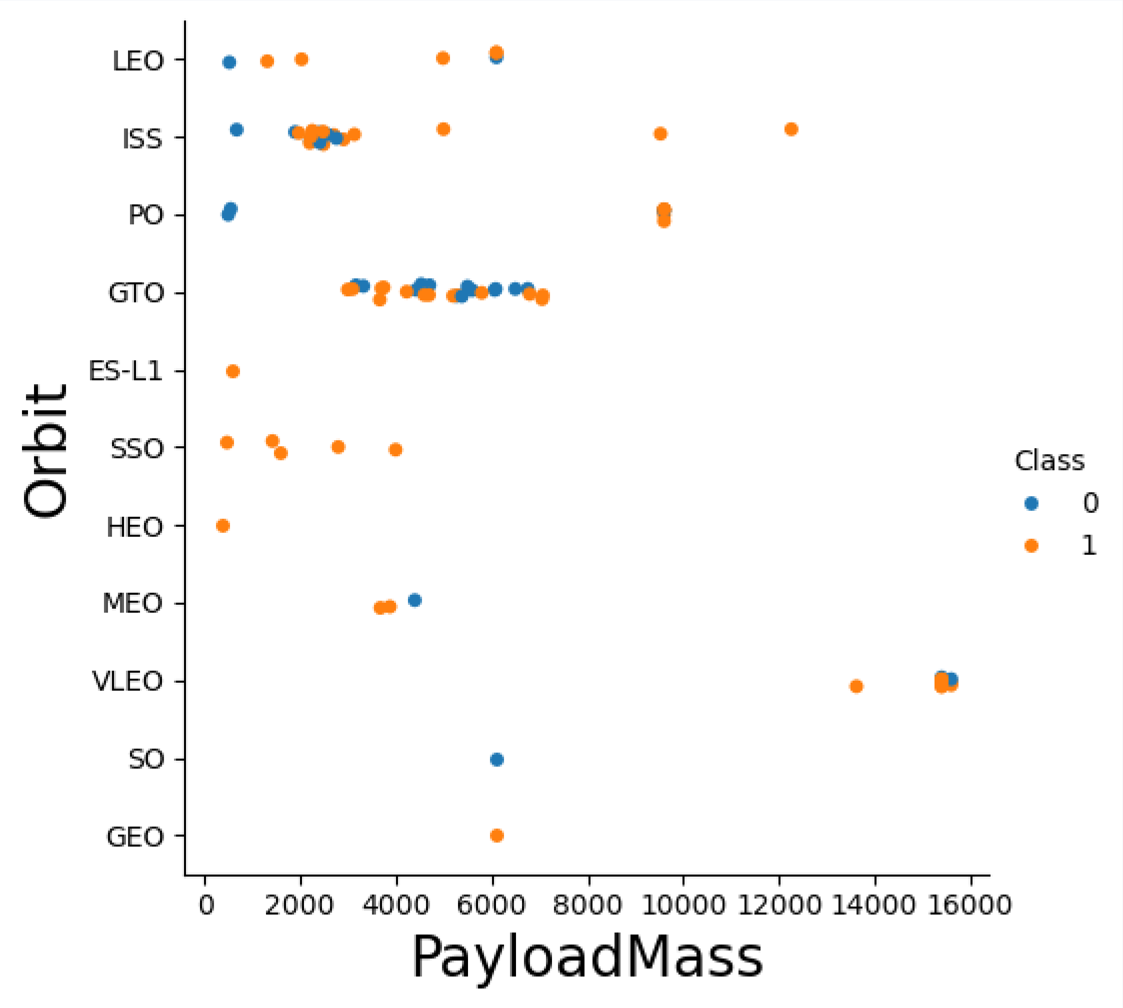
- ES-L1 and SO have only one point to display in the relation.
- SSO, HEO, MEO, VLEO, SO and GEO have only flight numbers grater than 40.
- For flight numbers greater than 40, we see there are much successful returns.



Payload vs. Orbit Type

This is a scatter plot that shows relationship between Orbit Type and Payload mass. The insights are:

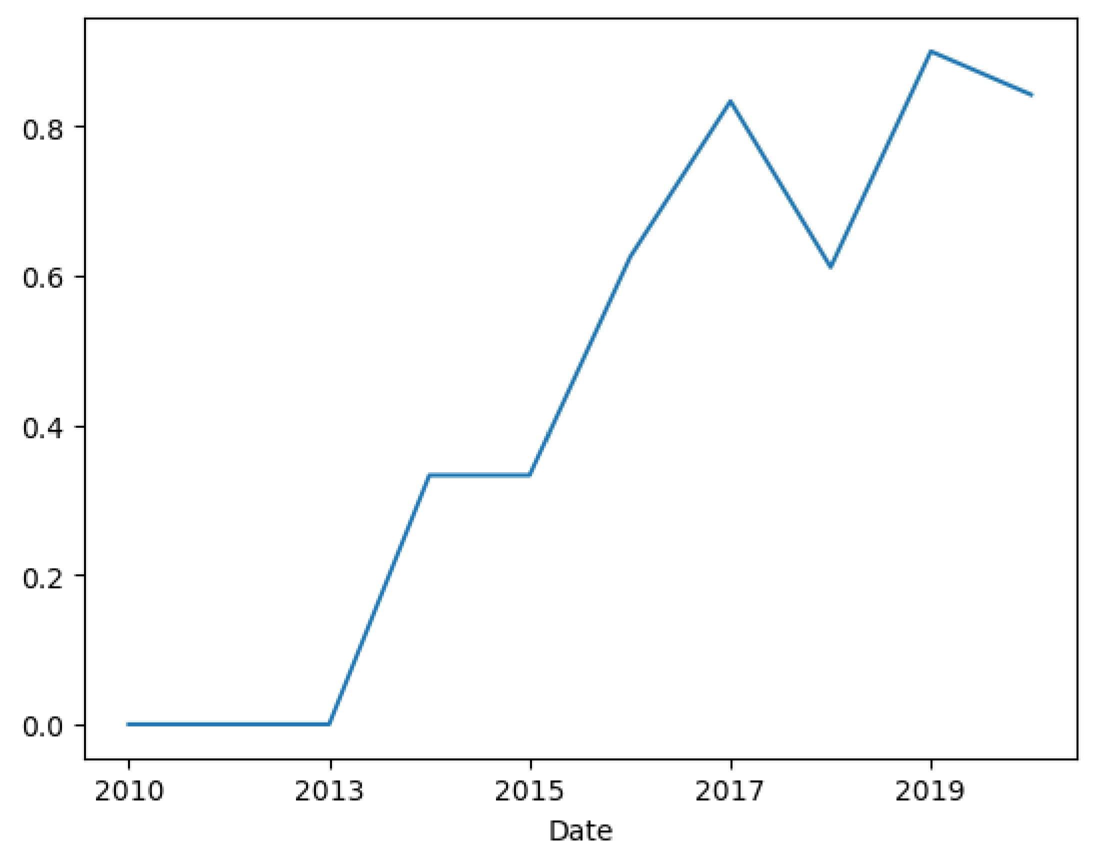
- The orbits ES-L1, SSO and HEO have only successful returns of first stage.
- ES-L1, SO and GEO have only one launch of 6000 payload mass each.



Launch Success Yearly Trend

This is a line chart that shows relationship between Launch success and Yearly trend. The insights are:

- This is a line chart that show the success rate across all the years.
- Since 2013 the success rate still increasing until the year 2020.



All Launch Site Names

```
%%sql
SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
```

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

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

With this query, we selected the unique names for the launch sites and eliminate duplicate values.

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT * FROM SPACEXTABLE 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
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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

We selected 5 results where the launch site name starts with the letters "CCA". The % operator is used to match zero or more characters after those letters.

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer == 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

We selected only the payload mass for the customer ‘NASA (CRS)’, and by using the SUM function, we obtained the total payload mass associated with that customer.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version == 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

We selected only the payload mass for the booster version ‘F9 v1.1’, and by using the AVG function, we obtained the average payload mass associated with that booster version.

First Successful Ground Landing Date

```
%%sql
SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome == 'Success (ground pad)'
```

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

MIN(Date)
2015-12-22

We selected the earliest date on which a “Success (ground pad)” occurred. To do this, we used the MIN function to retrieve the smallest (earliest) value in the Date column that matches this outcome.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome == 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_<6000)
```

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

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

We selected the booster versions that achieved a “Success (drone ship)” and carried a payload mass greater than 4000 kg and less than 6000 kg.

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT
    SUM(CASE WHEN Mission_Outcome LIKE 'Success%' THEN 1 ELSE 0 END) AS Success_Count,
    SUM(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 ELSE 0 END) AS Failure_Count
FROM SPACEXTABLE;
```

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

Success_Count	Failure_Count
100	1

We counted the number of successful and failed missions by checking the Mission_Outcome column. Using CASE expressions inside the SUM function, we counted entries that start with “Success” and “Failure” separately.

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

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

We selected the booster version that carried the maximum payload mass. To do this, we compared each record's payload mass with the highest payload mass in the table using a subquery.

2015 Launch Records

```
%sql SELECT substr(Date,6,2) as Month_Num, substr(Date,0,5) as Year, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE (Date >= '2010-01-01' AND Date<='2015-12-31') AND Landing
```

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

Month_Num	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

We selected the month, year, landing outcome, booster version, and launch site for all launches that occurred between 2010 and 2015 and ended in “Failure (drone ship)”. The substr function is used to extract the month and year from the Date column.

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

```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTABLE WHERE (Date >= '2010-06-04' AND Date<='2017-03-20') GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC
```

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

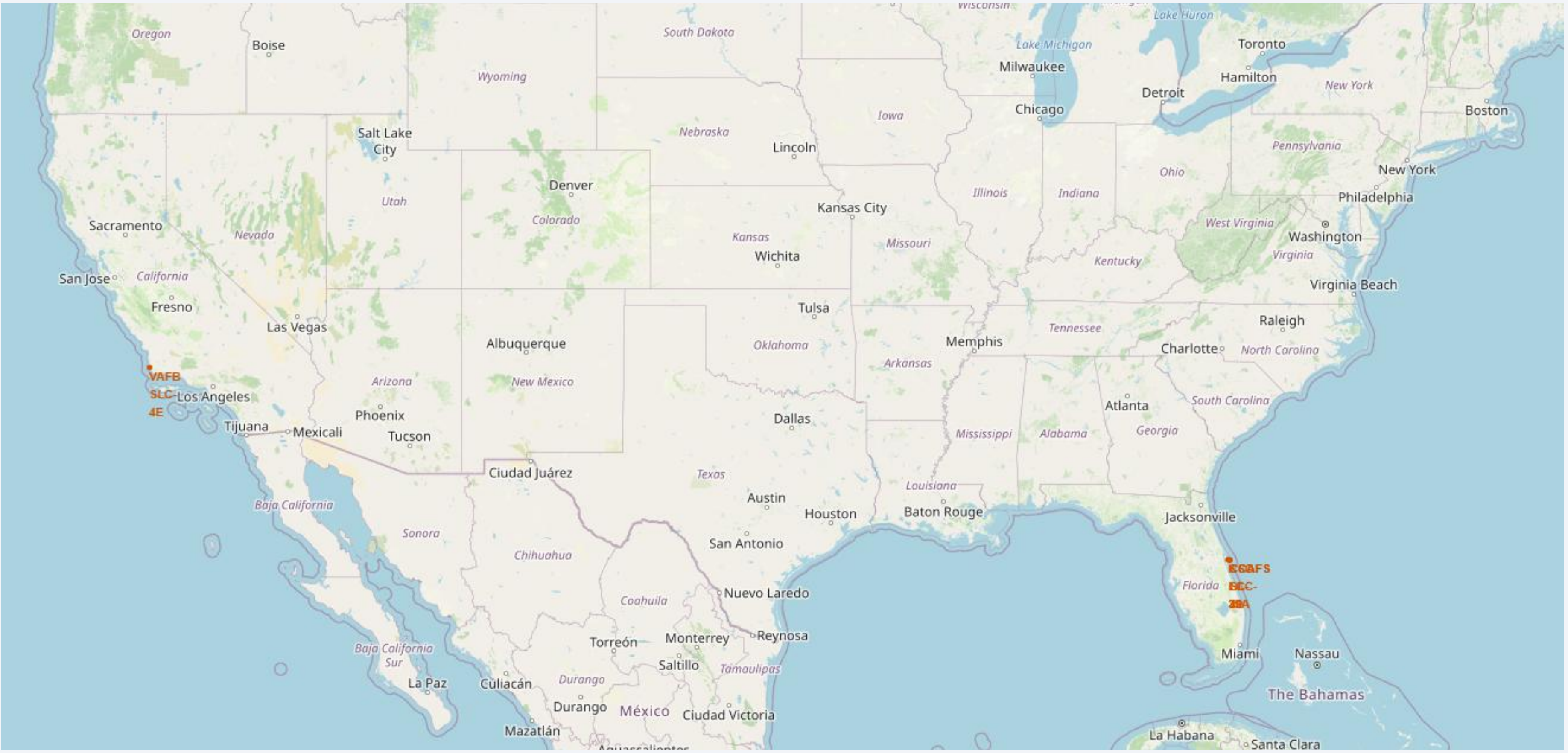
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

We counted how many times each landing outcome occurred between June 4, 2010 and March 20, 2017. The results are grouped by Landing_Outcome and ordered from the most frequent to the least frequent.

Section 3

Launch Sites Proximities Analysis

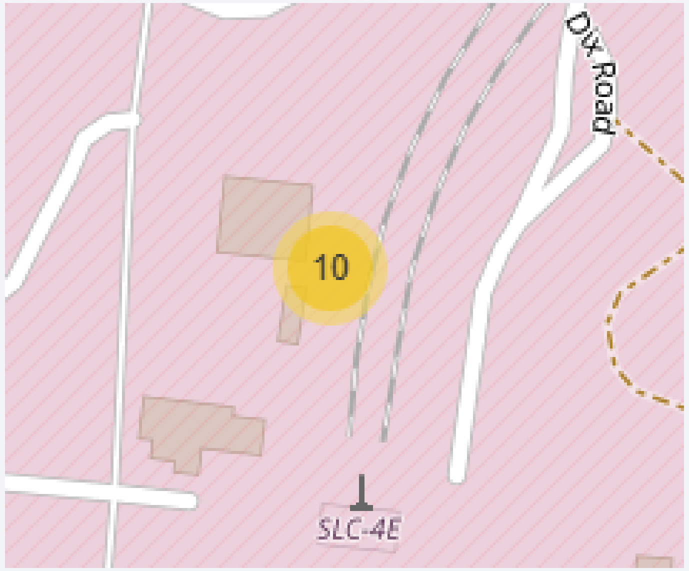
Launch sites map



We identified and mapped the different launch sites using circles and observed that all of them are located near the coast.

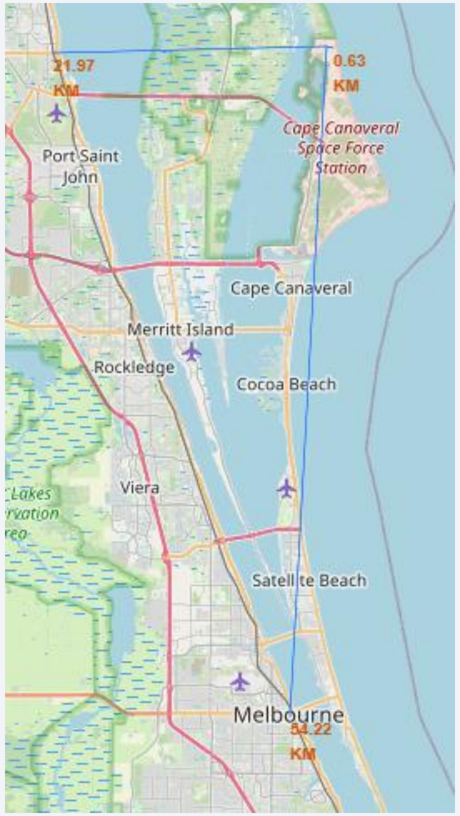
For each launch

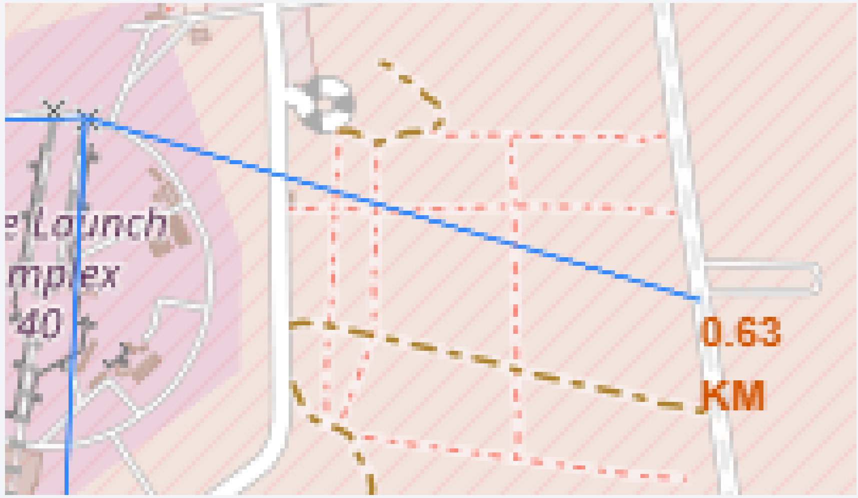
At each launch site, successful launches are marked in green and unsuccessful ones in red. Launches are also grouped to display the total number of launches per site.



Distance between points

We can observe the differences between sites, with some located near the coastline and others farther inland, closer to the city.

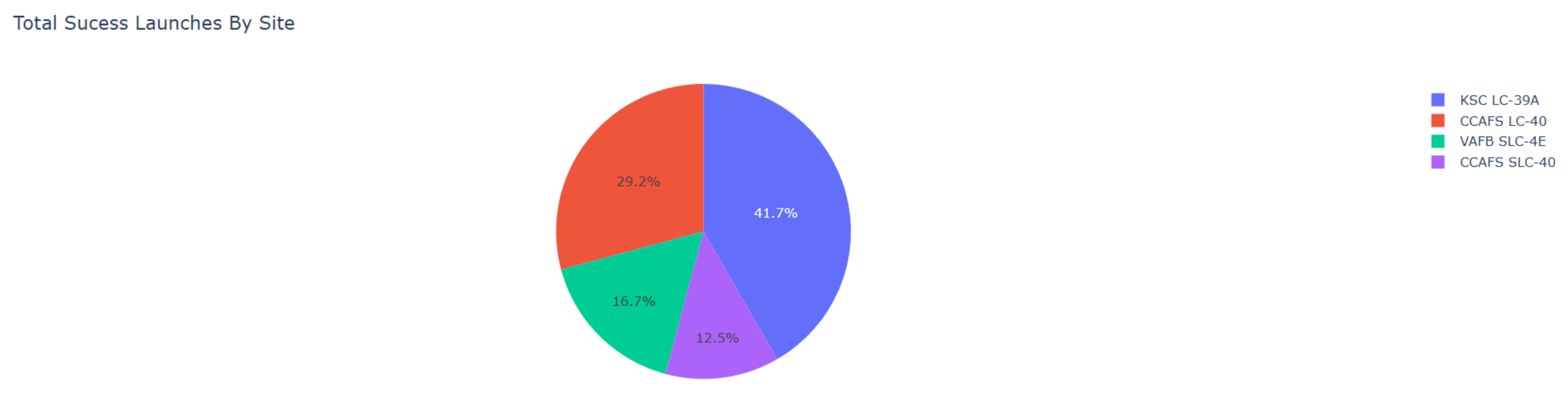




Section 4

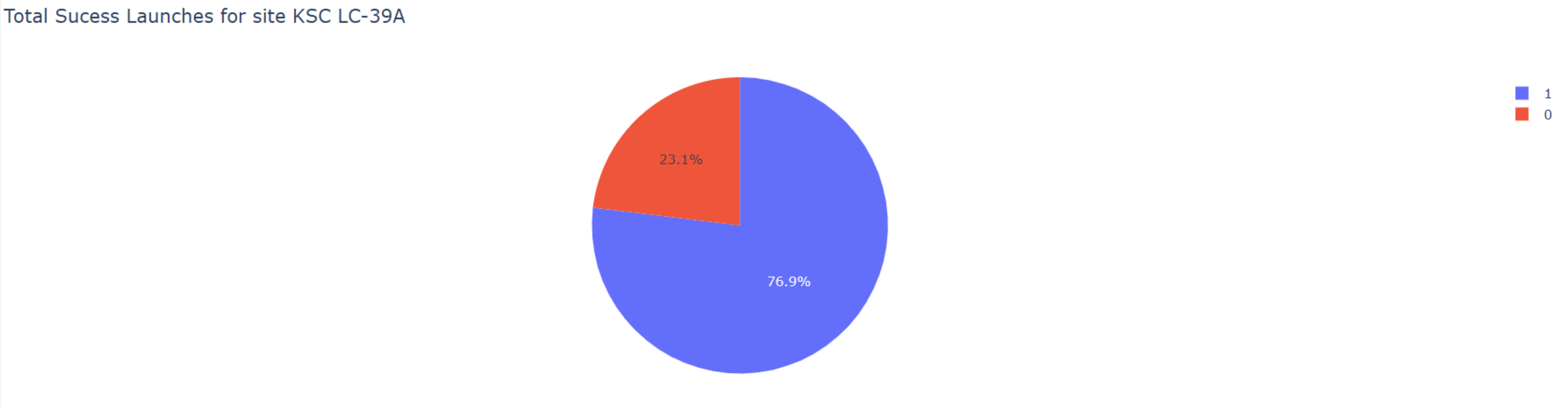
Build a Dashboard with Plotly Dash

Pie chart success



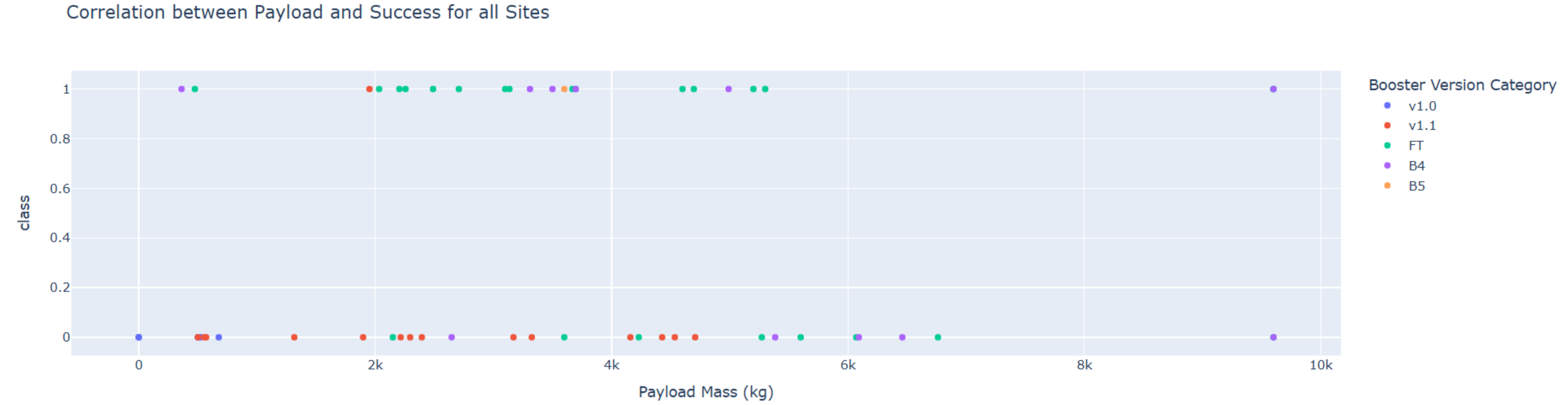
We can see that the launch site KSC LC-39A has the highest success rate, while CCAFS SLC-49 has the lowest.

Total success for KSC LC-39A



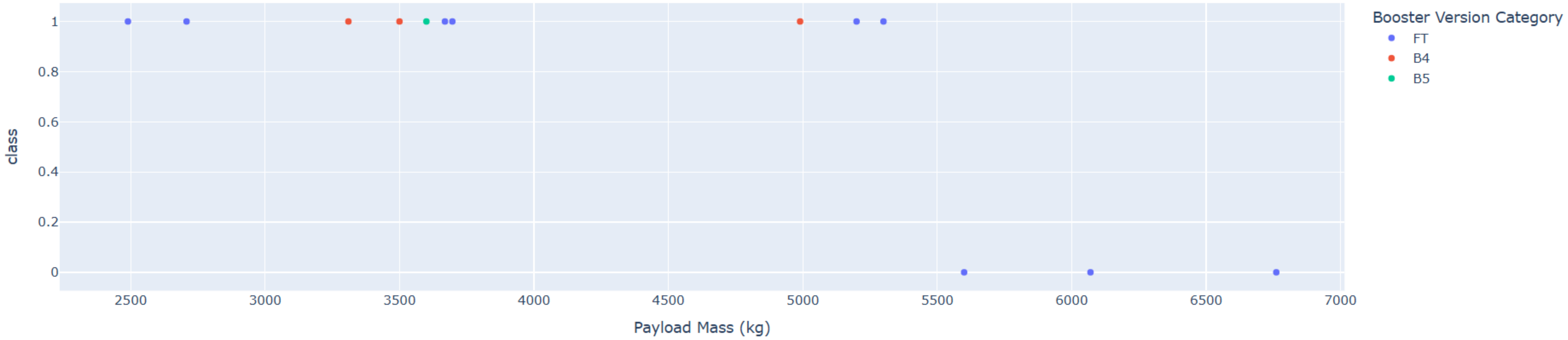
In this chart, we can see the total number of successful (76.9%) and unsuccessful (23.1%) launches for the best site, which is KSC LC-39A.

Scatter plot best site



We can see that most successful launches occur when the payload mass is in the range of 1,000 to 6,000 kilograms.

Scatter plot best site



We can also see that, within the preferred range for successful returns, the KSC-LC-39A landing site has a high percentage of achievements.

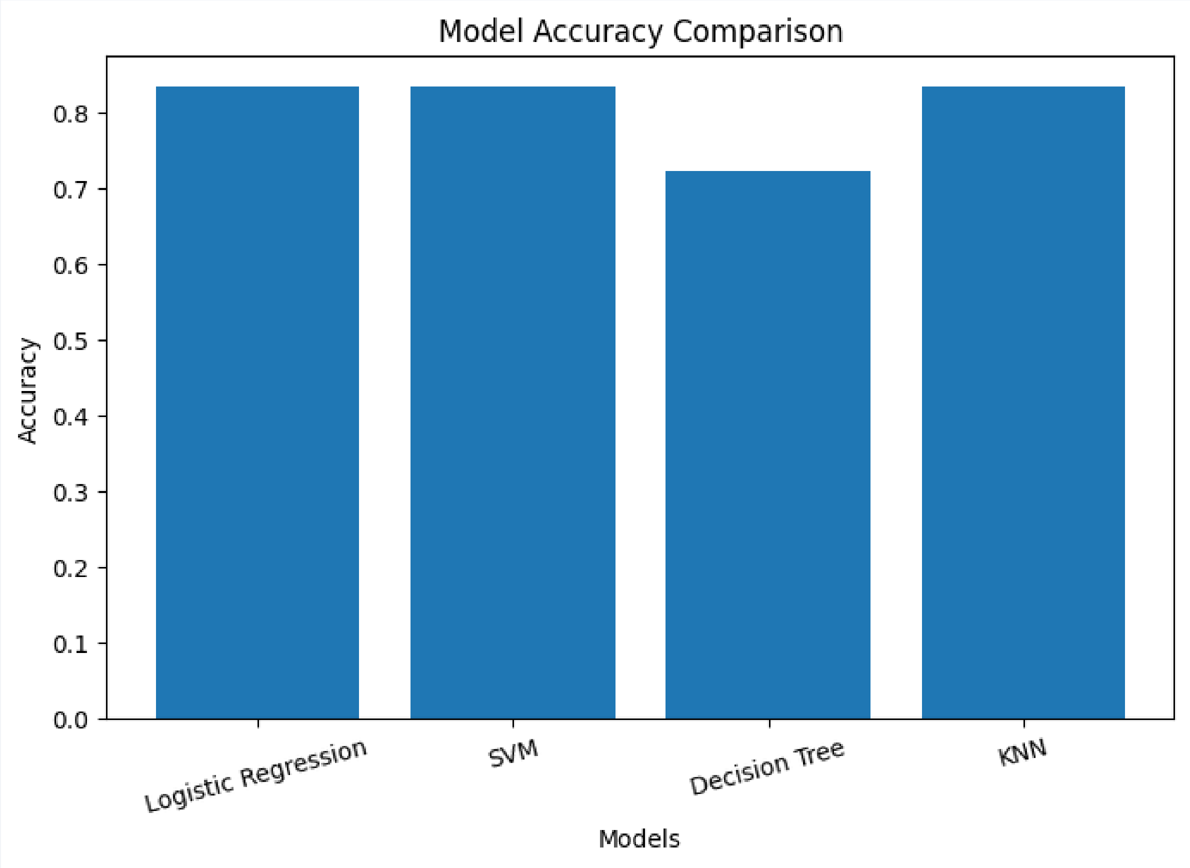
Section 5

Predictive Analysis (Classification)

Classification Accuracy

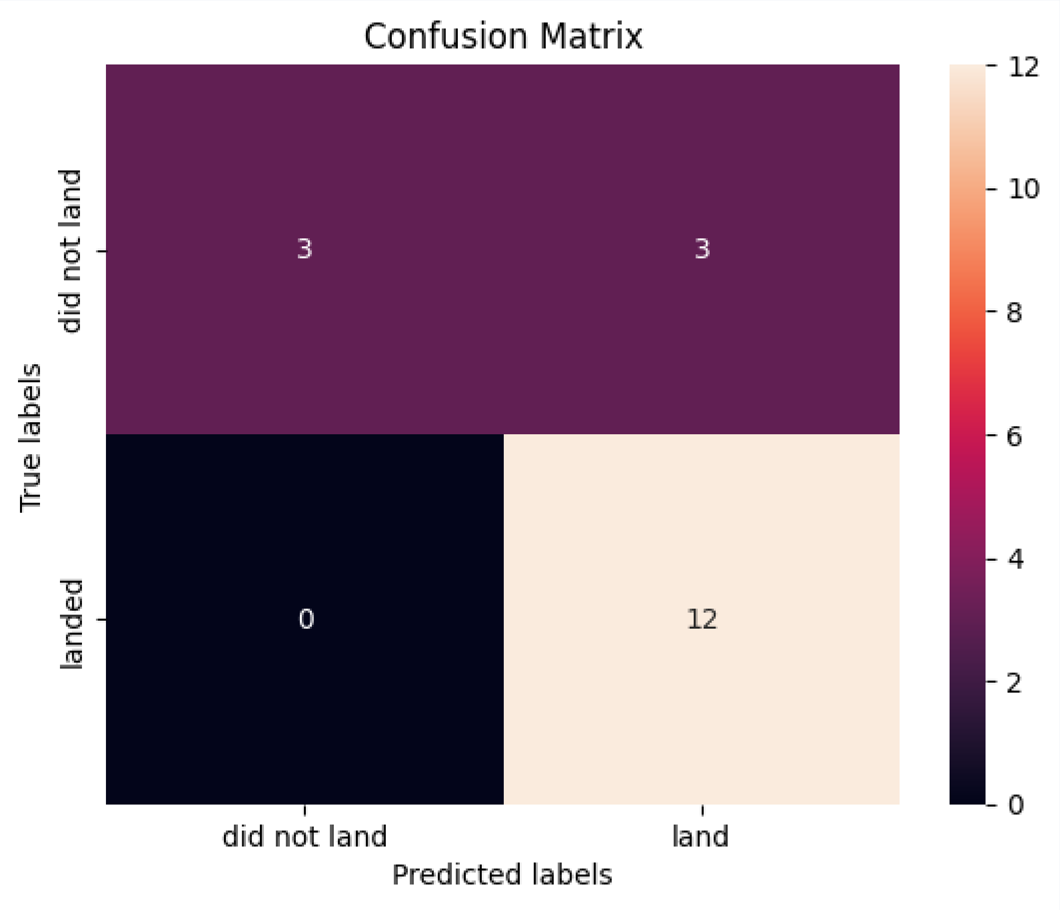
The Logistic Regression, SVM, and KNN models have the same accuracy when using the ***score*** method, but Logistic Regression shows the highest value when evaluated with the training data.

Therefore, the best model is **Logistic Regression**.



Confusion Matrix

The model has 3 False Positive errors (it predicted 'landing' when there was none) and 0 False Negative errors. The correct predictions consist of 12 True Positives and 3 True Negatives.



Conclusions

- Launch sites with more payload mass of flights tend to have higher success rates.
- ES-L1, GEO, HEO and SSO were the orbits with the most successful rate.
- The launch success rate increase in 2013 till 2020.
- The best launch site was KSC LC-39A with a 41.7% in the total success.
- The most successful launches occur when the payload mass is in the range of 1,000 to 6,000 kilograms.
- The best model was Logistic Regression.
- This was an incredible project that help me to applicate all the knowledge from topics on this course

Thank you!