IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Robert Polony

11.11.2024

Outline

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

Executive Summary

- Summary of methodologies

Various techniques have been used to analyze SpaceX data, such as downloading data from the SpaceX API, web scraping, data processing and cleaning (data wrangling), exploratory data analysis (EDA) using SQL and visualization using Matplotlib, Seaborn and Folium, and predictive analysis using Machine Learning classification models.

- Summary of all results

Launch success rate increased since 2013 till 2020. The launch site KSC LC-39A achieved the largest number of successful launches. Orbits: ES-L1, GEO, HEO, SSO has the 100% success rates. All Machine Learning models performed similarly, but the Decision Tree model achieved the highest accuracy (~0.87).

Introduction

- Project background and context

SpaceX is a leader in the space industry, known for groundbreaking achievements such as the first successful recovery of a rocket from low Earth orbit by a private company in December 2010. One of the key elements of SpaceX's strategy is to reuse rockets through safe first stage landings, significantly reducing launch costs.

The standard launch cost of a Falcon 9 rocket is $62 million, which is significantly cheaper compared to offerings from other companies (whose costs can exceed $165 million). The big savings come from rocket reusability, but this depends on the success of the first stage landing.

Therefore, the goal of the project is to analyze the factors affecting landing success and develop predictive models. Successful prediction of landing results can be valuable both to SpaceX and to competing companies considering entering the space market. The analysis includes aspects such as payload mass, launch site location, orbit type and the use of advanced analytical methods such as classification models.

Introduction

- Problems you want to find answers

- What factors affect the success rate of SpaceX launches?

We want to study which variables, such as payload mass, orbit type or launch site location, have a significant impact on mission outcome. Understanding these relationships will help identify the optimal conditions for successful missions.

- How does launch site location affect mission success?

We analyze whether and how differences in launch site locations and their proximity to specific infrastructure elements (e.g., coastline, highways, railroads) can affect success rates.

- Which classification models best predict mission outcome?

The goal is to develop predictive models that can effectively predict whether the first stage of a rocket will land successfully. Highly accurate predictions are crucial for assessing mission risk and reducing operational costs.

Introduction

- Problems you want to find answers

- Does payload mass affect mission outcome?

By examining the relationship between payload mass and success rate, we can determine how the load on the rocket affects the chances of a successful landing.

- How can SpaceX optimize launch site selection for future missions?

Analysis will identify optimal locations and conditions that can increase the probability of mission success.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data was collected by using request to the SpaceX API and by using web scraping from Wikipedia
- Perform data wrangling
    - Data was filtered to only include Falcon 9 launches. Missing values was replaced by mean. One-hot encoding was used for categorical data.
- Perform exploratory data analysis (EDA) using visualization and SQL

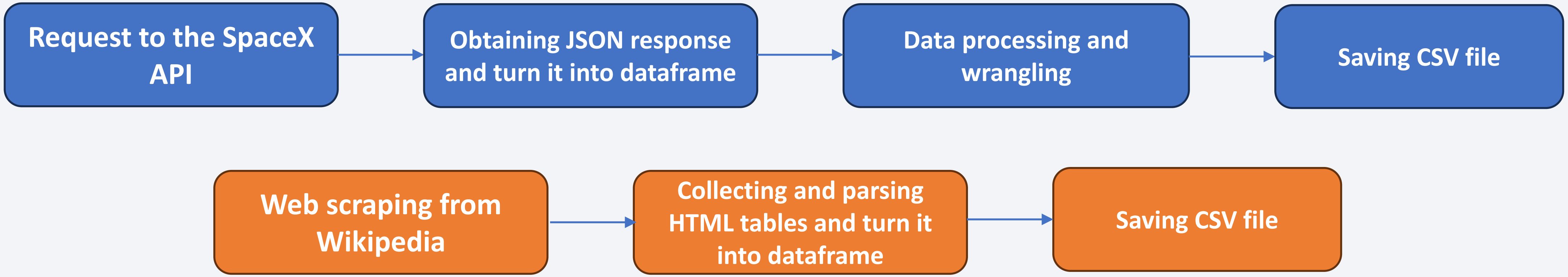
Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Developed classification models, performed hyperparameter tuning using Grid Search, and evaluated their performance using metrics such as accuracy.

Data Collection

- The data was collected using the SpaceX API and web scraping. The process involved obtaining a JSON response from the API, processing the data, and saving it to a CSV file for further analysis. Additional information about historical launch records was then gathered using web scraping form Wikipedia website.



Data Collection – SpaceX API

- Defined a series of functions that helped to use the API to extract information using IDs in the launch data
- Requested rocket launch data from SpaceX API
- Obtained response object
- Decoded the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()
- Used API again to get information about the launches by using created functions and the IDs given for each launch
- Constructed dataset by using obtained data and combined it into a dictionary
- Created dataframe from the dictionary
- Filtered the dataframe to only include Falcon 9 launches and replacing missing values with its mean value
- Saved CSV file: dataset_part_1.csv

- The link for SpaceX API calls notebook:
https://github.com/Robson2k7/IBM-SpaceX-Applied-Data-Science-Capstone/blob/main/1-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Created a series of functions to process web scraped HTML table
- Requested Falcon 9 launch data from Wikipedia
- Created BeautifulSoup object from the HTML response
- Extracted all column/variable names from the HTML table header
- Collected data by parsing the launch HTML tables
- Constructed dataset by using obtained data and combined it into a dictionary
- Created dataframe from the dictionary
- Saved CSV file: spacex_web_scraped.csv

- The link for web scraping notebook: https://github.com/Robson2k7/IBM-SpaceX-Applied-Data-Science-Capstone/blob/main/2-jupyter-labs-spacex-webscraping.ipynb

Data Wrangling

Loaded data from CSV file: dataset_part_1.csv

Performed Exploratory Data Analysis

Calculated the number of launches on each site

Calculated the number and occurrence of each orbit

Calculated the number and occurrence of mission outcome of the orbits

Created a landing outcome label from Outcome column (0- unsuccessful, 1 – successful)

Saved CSV file: dataset_part_2.csv

- The link for data wrangling notebook:
https://github.com/Robson2k7/IBM-SpaceX-Applied-Data-Science-Capstone/blob/main/3-jupyter-labs-spacex-data-wrangling.ipynb

EDA with Data Visualization

- Charts:

Flight Number vs Payload Mass

The chart analyzes how consecutive launch attempts (FlightNumber) and payload mass (PayloadMass) affect the mission outcome. The goal is to examine whether SpaceX's increasing experience and heavier payloads improve the likelihood of a successful first-stage landing.

Flight Number vs Launch Site

The chart analyzes how SpaceX's experience (Flight Number) impacts launch outcomes at different launch sites. The goal is to identify which locations are the most reliable and contribute to mission success.

Payload Mass vs Launch Site

The chart analyzes how payload mass was distributed across different launch sites. The goal is to identify which locations handle heavier or lighter payloads and how it affects mission success.

- The link for EDA with data visualization notebook: https://github.com/Robson2k7/IBM-SpaceX-Applied-Data-Science-Capstone/blob/main/5-jupyter-labs-spacex-eda-data-vizualization.ipynb

EDA with Data Visualization

- Charts:

Success Rate (Mean of Class) vs Orbit Type

The bar chart analyzes which orbits have the highest success rates.

Flight Number vs Orbit Type

The chart analyzes the relationship between flight number and the type of orbit to which payloads were launched. The goal is to examine whether mission success is influenced by SpaceX's experience across different orbits.

Payload Mass vs Orbit Type

The chart analyzes how payload mass impacts landing outcomes across different orbits. The goal is to examine which orbits favor successful landings based on weight of the payloads.

The Launch Success Yearly Trend

The chart analyzes the yearly trend of SpaceX's launch success. The goal is to examine how the success rate has changed over time and whether SpaceX has achieved greater reliability in recent years.

- The link for EDA with data visualization notebook:
https://github.com/Robson2k7/IBM-SpaceX-Applied-Data-Science-Capstone/blob/main/5-jupyter-labs-spacex-eda-data-vizualization.ipynb

EDA with SQL

SQL queries for:

- The names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- The date when the first succesful landing outcome in ground pad was achieved
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- The names of the booster_versions which have carried the maximum payload mass
- The list of failure landing outcomes on drone ship, their booster version and launch site for the months in 2015
- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

- The link for EDA with SQL notebook: https://github.com/Robson2k7/IBM-SpaceX-Applied-Data-Science-Capstone/blob/main/4-jupyter-labs-spacex-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Marked all launch sites on a map

- Added orange circle at NASA Johnson Space Center's coordinate with a popup label showing its name
- Added red circles at all launch sites coordinates with a popup labels showing ist names

Marked the success/failed launches for each site on the map

- Added markers: green, for successful launches, and red for unsuccessful launches to all launch sites coordinates

Calculated the distances between a launch site to its proximities

- Added blue lines to show distance between launch site CCAFS SLC-40 and the nearest coastline and railway

- The link for notebook with interactive map build with Folium:

https://github.com/Robson2k7/IBM-SpaceX-Applied-Data-Science-Capstone/blob/main/6-jupyter-lab-spacex-launch-site-location-updated.ipynb

Build a Dashboard with Plotly Dash

Created Dropdown List with launch sites allowing to select all of them or a certain launch site

Created Pie Chart showing successful and unsuccesfull launches as a percent of the total

Created Slider of Payload Mass Range

Created Scatter Chart showing the correlation between Payload Mass and Launch Success basing on Booster Version

- The link for Plotly Dash lab:
https://github.com/Robson2k7/IBM-SpaceX-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Created a NumPy array from the column Class in data

Standardized the data in X then reassigned it to the variable X using the StandardScaler

Splited the data by using train_test_split function

Created a GridSearchCV object with cv=10 for parameter optimalization

Applied GridSearchCV on: Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor algorithms

Calculated accuracy on the test data for all models

Created the confusion matrix for all models

- The link for Predictive Analysis lab:

https://github.com/Robson2k7/IBM-SpaceX-Applied-Data-Science-Capstone/blob/main/7-jupyter-lab-spacex-machine-learning-prediction.ipynb

Results

- Exploratory data analysis results

Launch success rate increased since 2013 till 2020. The launch site KSC LC-39A achieved the largest number of successful launches. Orbits: ES-L1, GEO, HEO, SSO has the 100% success rates.

- Predictive analysis results

All Machine Learning models performed similarly, but the Decision Tree model achieved the highest accuracy (~0.87).

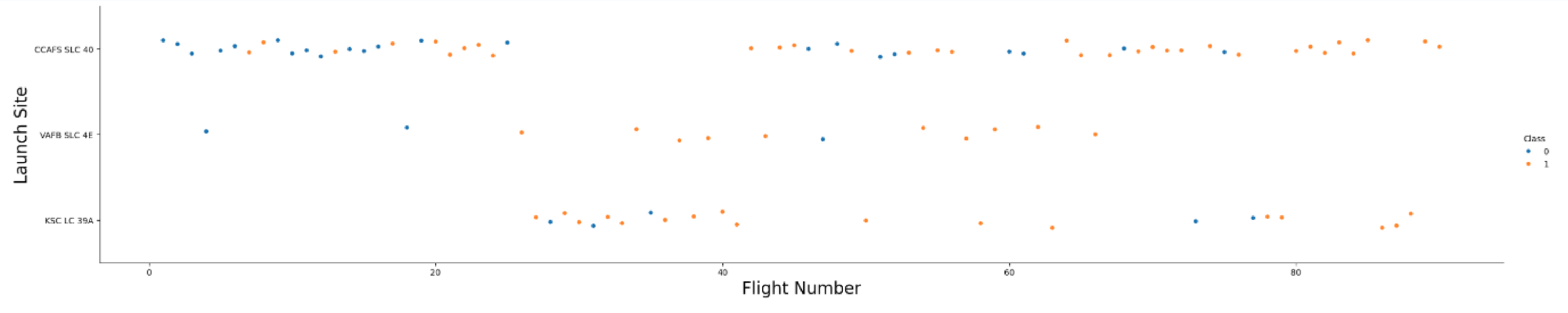
- Interactive analytics demo in screenshots

Section 2

Insights drawn from EDA

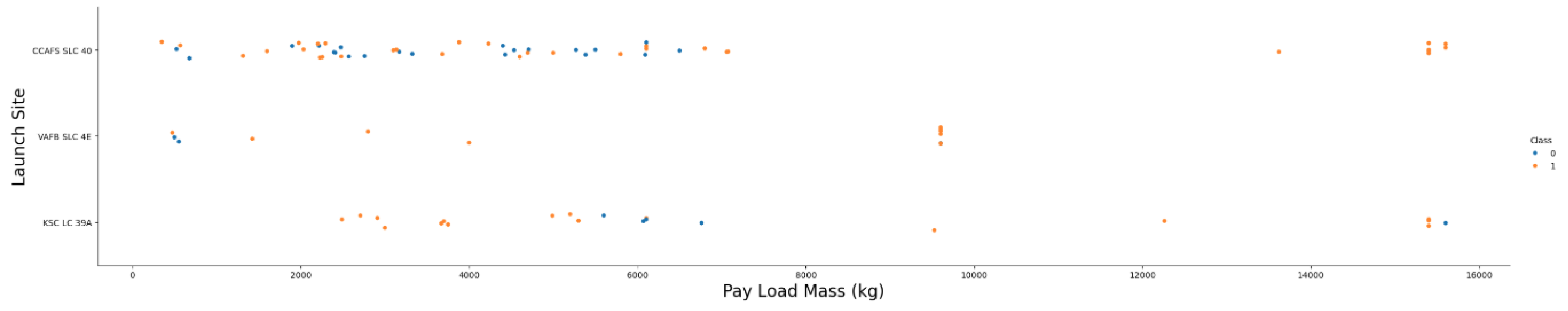
Flight Number vs. Launch Site

In summary, the chart suggests that with experience, SpaceX is achieving increasingly better results in its launches, especially at the launch sites KSC LC 39A and CCAFS SLC 40.

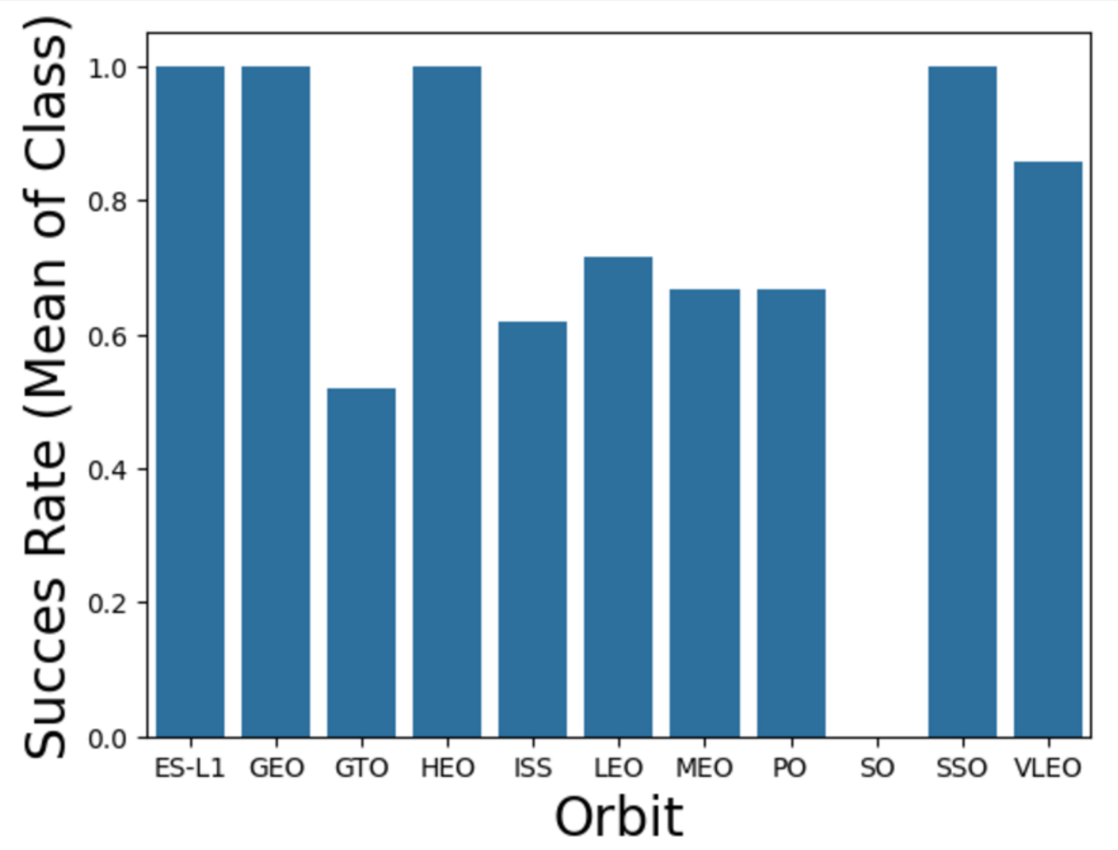


Payload vs. Launch Site

The chart shows that the higher payload mass, the higher the success rate. Also VAFB SKC 4E has not launched anything greater than 10000 kg.



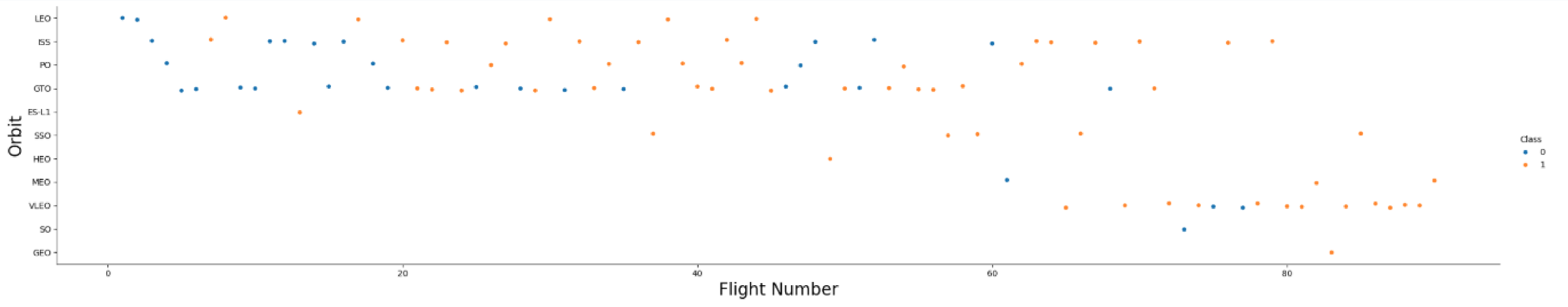
Success Rate vs. Orbit Type



The bar chart shows that orbits ES-L1, GEO, HEO, SSO have the highest (100%) success rates.

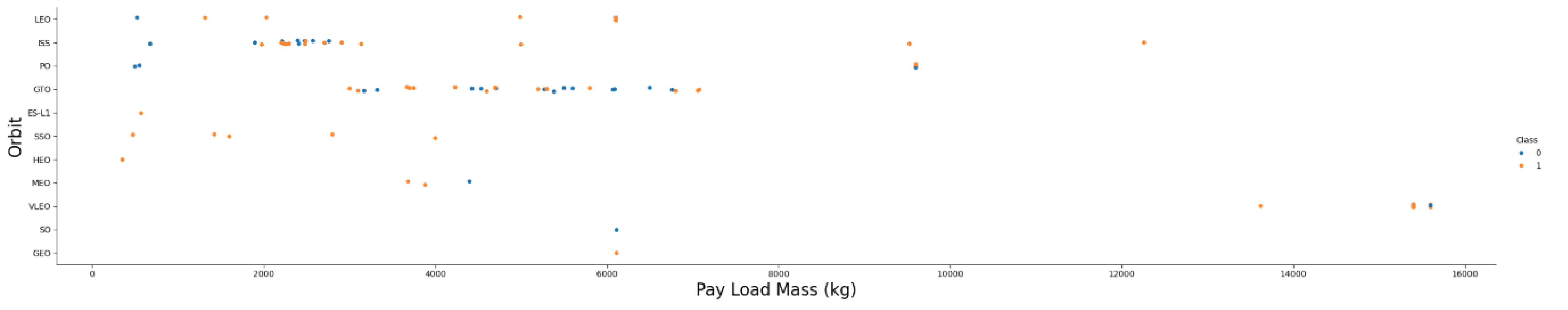
Flight Number vs. Orbit Type

The LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

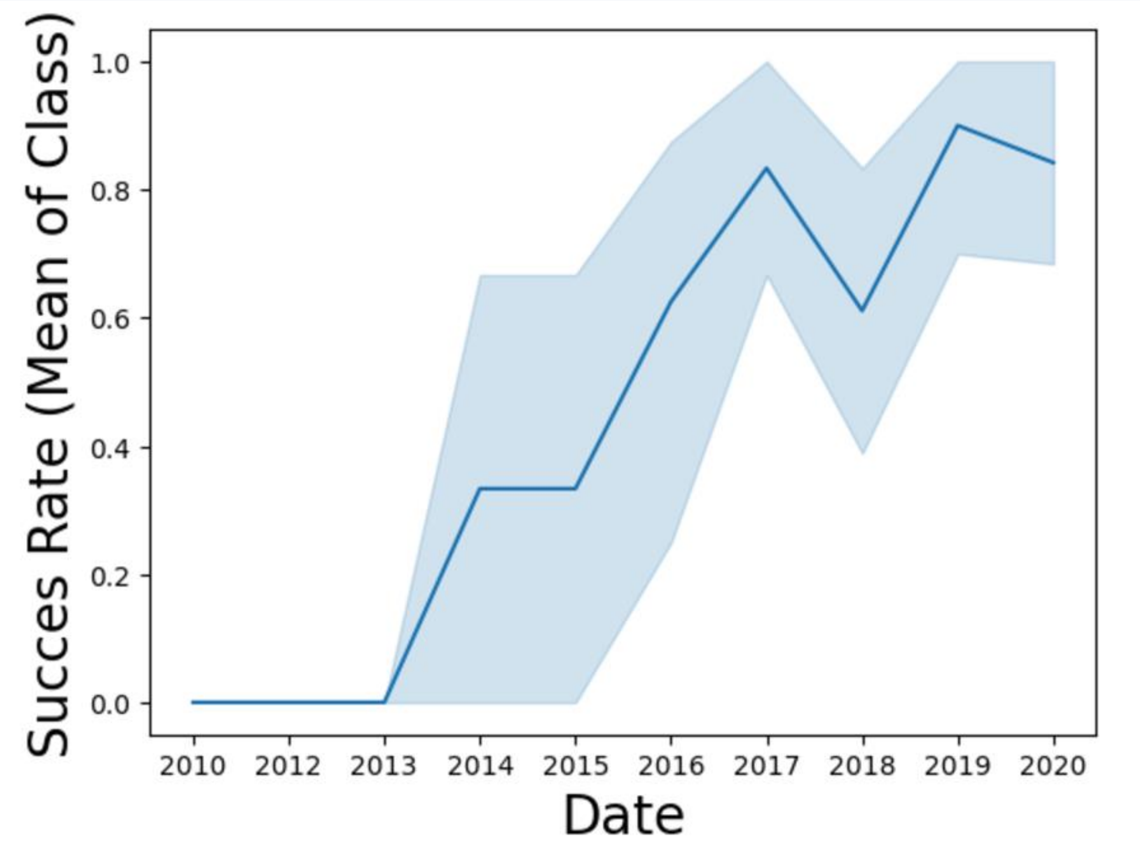


Payload vs. Orbit Type

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend



Launch success rate increased since 2013 till 2020.

All Launch Site Names

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

In [13]:
```
%sql SELECT DISTINCT "Launch_Site" from SPACEXTABLE;
```

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

Out[13]:

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

Used DISTINCT to show unique launch sites from the data.

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

In [15]:
```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE '%CCA%' LIMIT 5;
```

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

Out[15]:

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

Used LIKE to display records where launch sites name Begin with „CCA".

LIMIT to show only 5 rows.

Total Payload Mass

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

In [19]:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer = "NASA (CRS)";
```

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

Out[19]:

Total_Payload_Mass
45596

Calculated the total payload mass carried by boosters launched by NASA (CRS) with SUM.

The Total Payload Mass is 45596 kg.

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [20]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1";
```

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

Out[20]:

Average_Payload_Mass
2928.4

Calculated the average payload mass carried by booster version F9 v1.1 with AVG.

The Average Payload Mass is 2928,4 kg.

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [22]:

```
%sql SELECT MIN(Date) AS First_Succesful_Landing FROM SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)";
```

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

Out[22]:

First_Succesful_Landing
2015-12-22

Used MIN for Date column and WHERE on Landing_Outcome to show when was the first succesful landing in ground pad.

It is 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [24]:

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 40
```

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

Out[24]:

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

Used WHERE to filter boosters which have succesful landing outcome on drone ship and also used AND to determine succesful landing with payload mass 4000 and 6000.

FULL CODE: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [28]:
```
%sql SELECT Mission_Outcome, COUNT(*) AS Total_Count FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

Out[28]:

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

Used COUNT to count all succesful and unsuccesful missions.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [31]:
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

Out[31]:

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

Searched for the boosters that have carried maximum payload mass by using a subquery with MAX function in WHERE function.

2015 Launch Records

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [37]:

```
%sql SELECT CASE WHEN SUBSTR(Date, 6, 2) = '01' THEN 'January' WHEN SUBSTR(Date, 6, 2) = '02' THEN 'February' WHEN SUBSTR(Da
```

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

Out[37]:

Month	Booster_Version	Launch_Site	Landing_Outcome
January	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Replaced months numbers for its names. Example: 01 to January, etc.

FULL CODE: %sql SELECT CASE WHEN SUBSTR(Date, 6, 2) = '01' THEN 'January' WHEN SUBSTR(Date, 6, 2) = '02' THEN 'February' WHEN SUBSTR(Date, 6, 2) = '03' THEN 'March' WHEN SUBSTR(Date, 6, 2) = '04' THEN 'April' WHEN SUBSTR(Date, 6, 2) = '05' THEN 'May' WHEN SUBSTR(Date, 6, 2) = '06' THEN 'June' WHEN SUBSTR(Date, 6, 2) = '07' THEN 'July' WHEN SUBSTR(Date, 6, 2) = '08' THEN 'August' WHEN SUBSTR(Date, 6, 2) = '09' THEN 'September' WHEN SUBSTR(Date, 6, 2) = '10' THEN 'October' WHEN SUBSTR(Date, 6, 2) = '11' THEN 'November' WHEN SUBSTR(Date, 6, 2) = '12' THEN 'December' END AS Month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date, 0, 5) = '2015';

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

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.

```
In [39]: %sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GRO
```

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

Out[39]:

Landing_Outcome	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

COUNT of landing outcomes and used WHERE function to filter them from dates BETWEEN 2010-06-04 to 2017-03-20. GROUPED and ordered output data by descending ORDER.

FULL CODE: %sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;

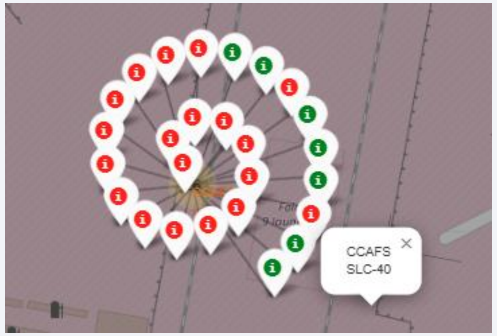
Section 3

Launch Sites Proximities Analysis

All Launch Sites

All launch sites are in very close proximity to the coast: VAFB SLC-4E in California, KSC LC-39A and CCAFS SLC-40 and CCAFS LC-40 in Florida. They all are in proximity to the Equator line.

Launch Outcomes at each Launch Site

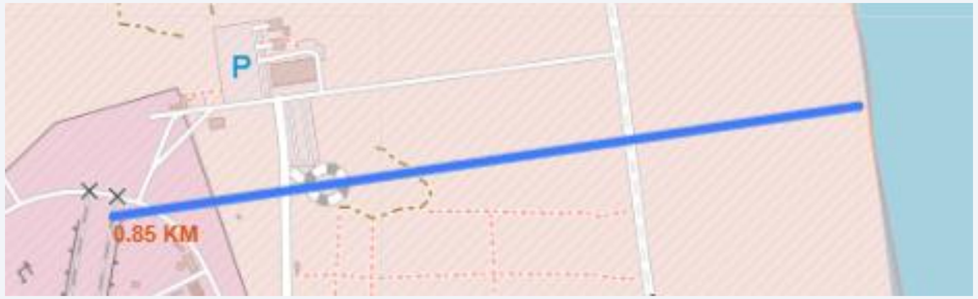


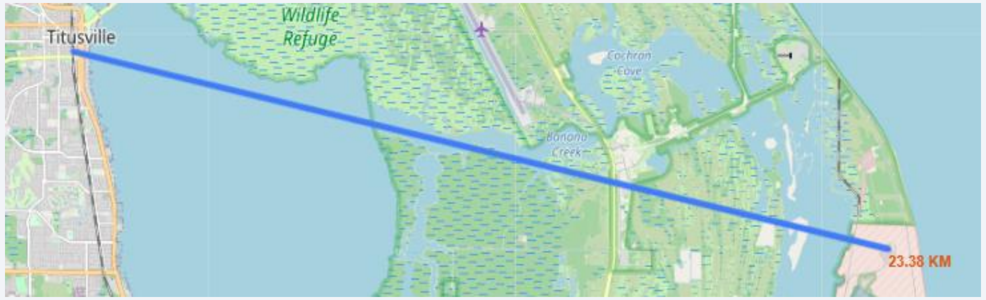
Green markers for succesful launches

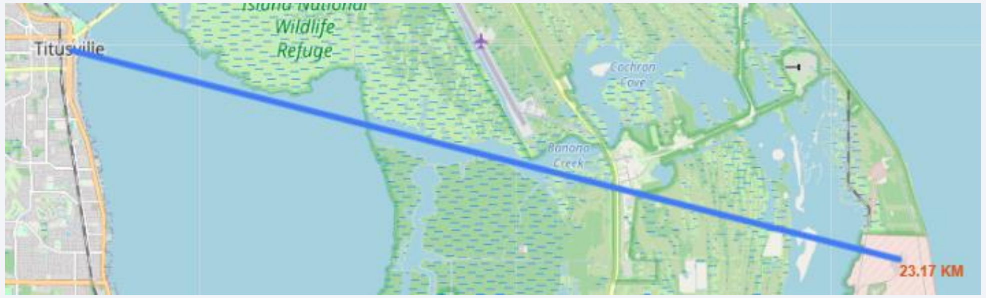
Red markers for unsuccesful launches

The launch site KSC LC-39A achieved the largest number of successful launches.

Distance to Proximities





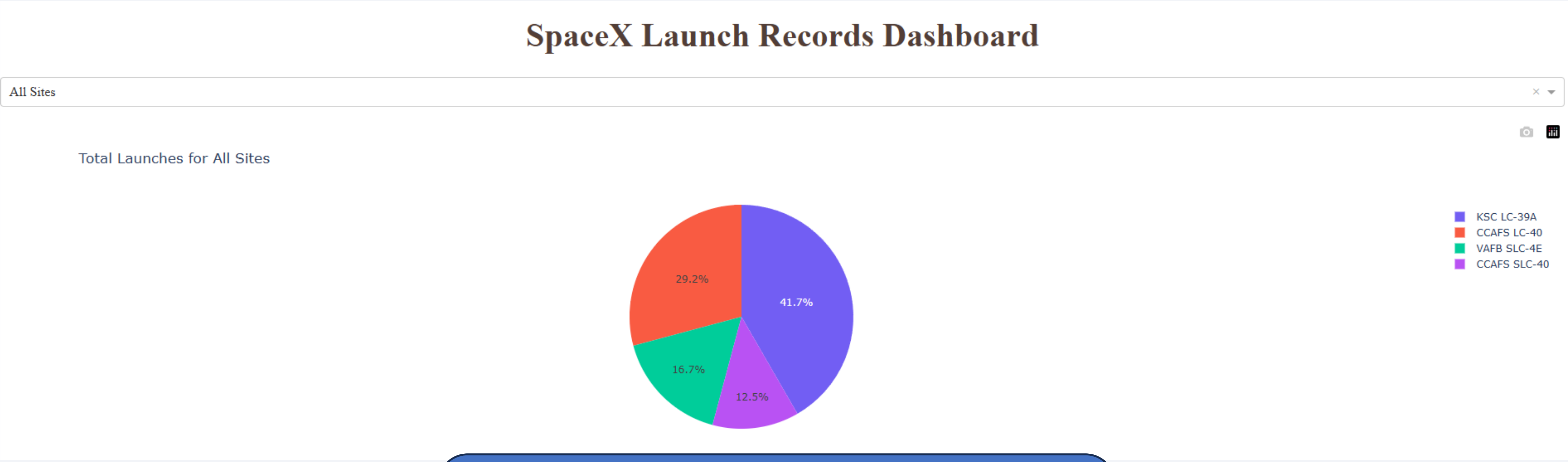


Example: CCAFS SLC-40 keeps certain distance away from cities, railways and highways. But, as said before, its in close proximity to coastline.

Section 4

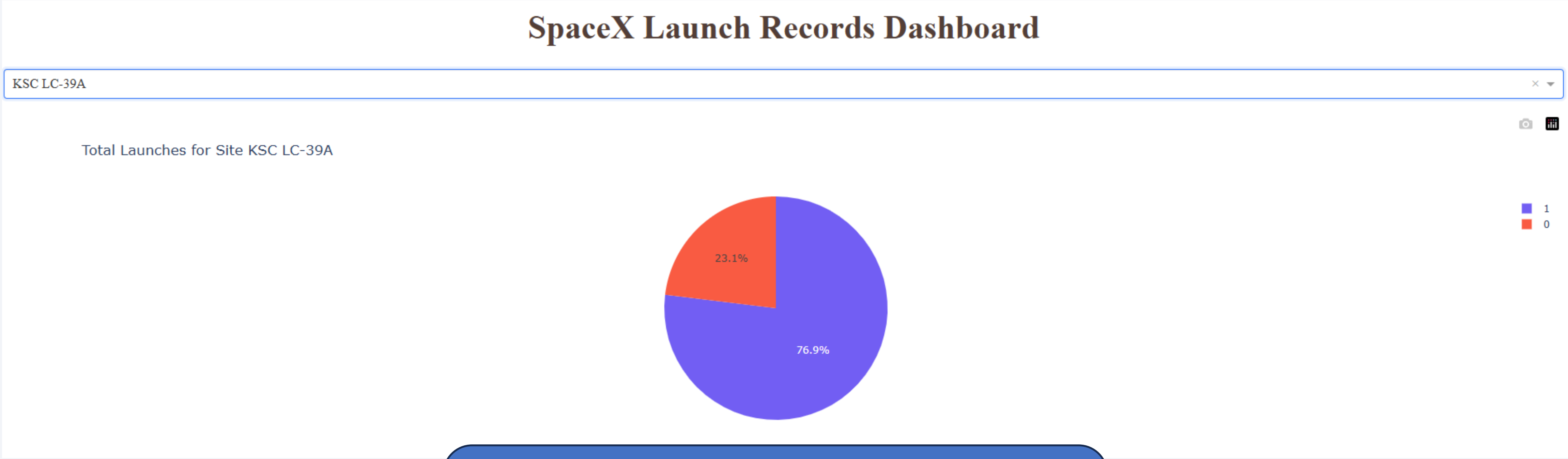
Build a Dashboard with Plotly Dash

Launch Success by site



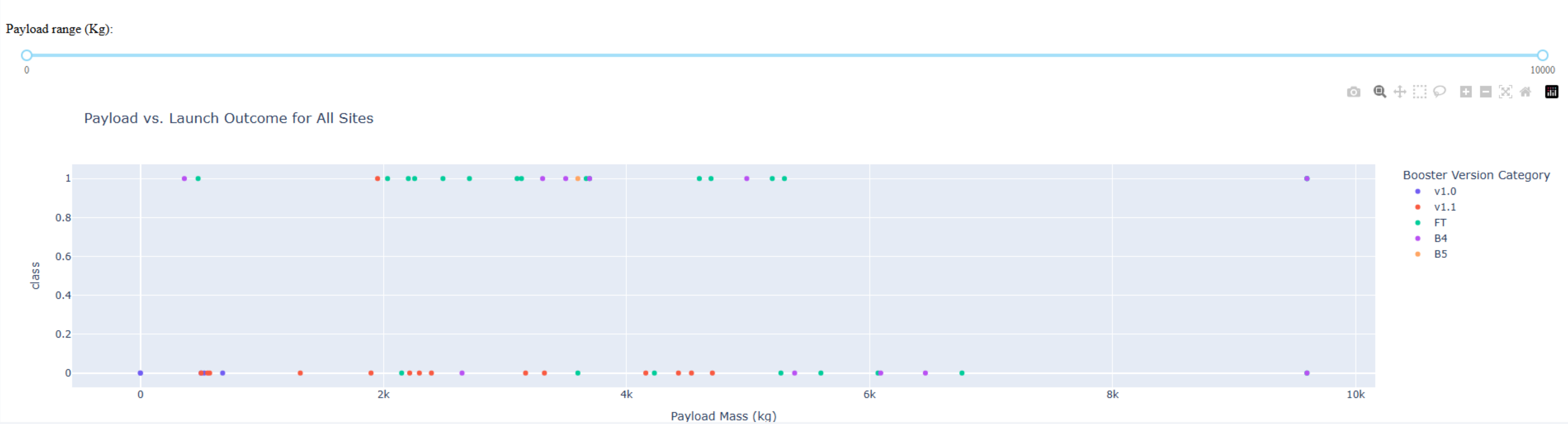
The launch site KSC LC-39A achieved the largest number of successful launches (41.7%).

Launch Site with highest success rate



The launch site KSC LC-39A has the highest success rate amongst launch sites (76.9%).

Payload Mass and Success



Payloads between circa 2000kg and 5000kg has the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Logistic Regression

```
In [14]: print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
         print("accuracy :",logreg_cv.best_score_)

tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

Support Vector Machine

```
In [21]: print("tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
         print("accuracy :",svm_cv.best_score_)

tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856
```

Decision Tree

```
In [27]: print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
         print("accuracy :",tree_cv.best_score_)

tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 14, 'max_features': 'sqrt', 'min_samples_leaf':
4, 'min_samples_split': 10, 'splitter': 'best'}
accuracy : 0.8714285714285713
```

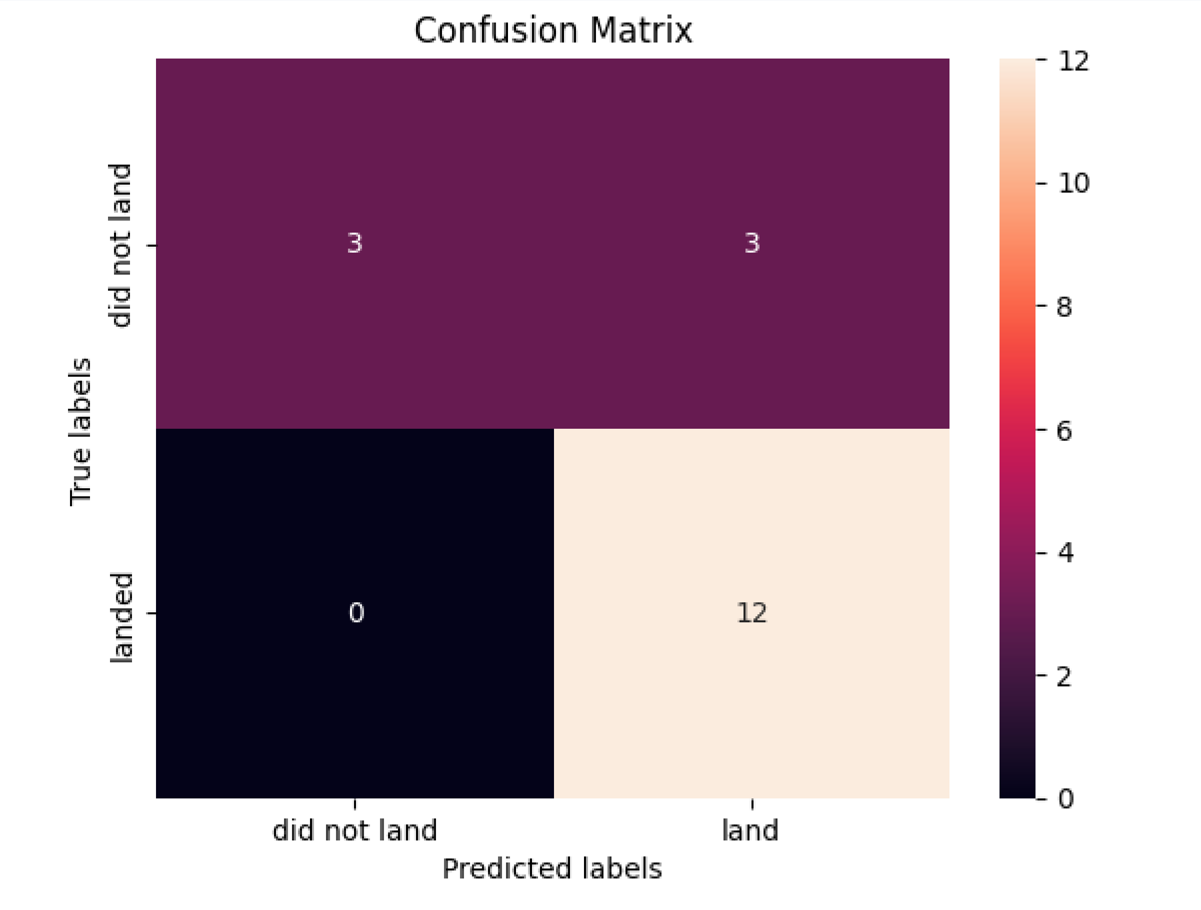
K-Nearest Neighbor

```
In [32]: print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
         print("accuracy :",knn_cv.best_score_)

tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858
```

Decision Tree is the model with highest clasification accuracy.

Confusion Matrix



Overview:
True Postive - 12 (True label is landed, Predicted label is also landed)
False Postive - 3 (True label is not landed, Predicted label is landed)

Confusin Matrix for all models were practically the same. The situation that there are 3 false postitives suggest that it is not good. Because they are marked by the model as succesful landings.

Conclusions

- With experience, SpaceX is achieving increasingly better results in its launches.
- Launch success rate increased since 2013 till 2020.
- The launch site KSC LC-39A achieved the largest number of successful launches.
- Orbits: ES-L1, GEO, HEO, SSO has the 100% success rates.
- The higher payload mass, the higher the success rate.
- All launch sites keeps certain distance away from cities, railways and highways. But they are in close proximity to coastline and also to the Equator line.
- All Machine Learning models performed similarly, but the Decision Tree model achieved the highest accuracy (~0.87).

Thank you!