IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Author: Luis Gonzalez
October 23, 2023

Outline

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

Executive Summary

The goal of the present Project was to analyze historical SpaceX rocket launch data, in order to develop machine learning models capable of predicting the ocurrence of a successful or unsuccessful launch.

- **Methodology**: The project was done entirely using the python language and SQL, and followed the following process:

Data Collection **->** Data wrangling **->** Exploratory Data Analysis **->** Interactive Visual Analytics and Dashboarding **->** Machine Learning Prediction Modelling

- **Results:** Using independent, explanatory variables like booster model, payload mass, launch site among others, an out-of-sample accuracy of 83.33% was achieved for predicting the occurrence of a successful or unsuccessful launch of s SpaceX rocket.

Introduction

- SpaceX is a very prominent Company in the sector of space exploration, and their success comes largely from their ability to offer rocket missions that can cost up to 100M US dollars less tan their competitors.
- The background for this project is the existence of a hypothetical competing company to SpaceX, that wants to identify the factors leading to them having successful rocket missions, so they may gain insights and apply them to their own processes.
- The most important objective, is to develop a machine learning model capable of predicting with enough accuracy if a rocket misión will be successful, given a variety of explanatory variables, like payload mass and booster version.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from a pre-prepared online csv file
- Perform data wrangling
  - Data was wrangled using Python libraries *numpy* and *pandas*
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression, SVM, Decision Trees and KNN were considered as classification models.

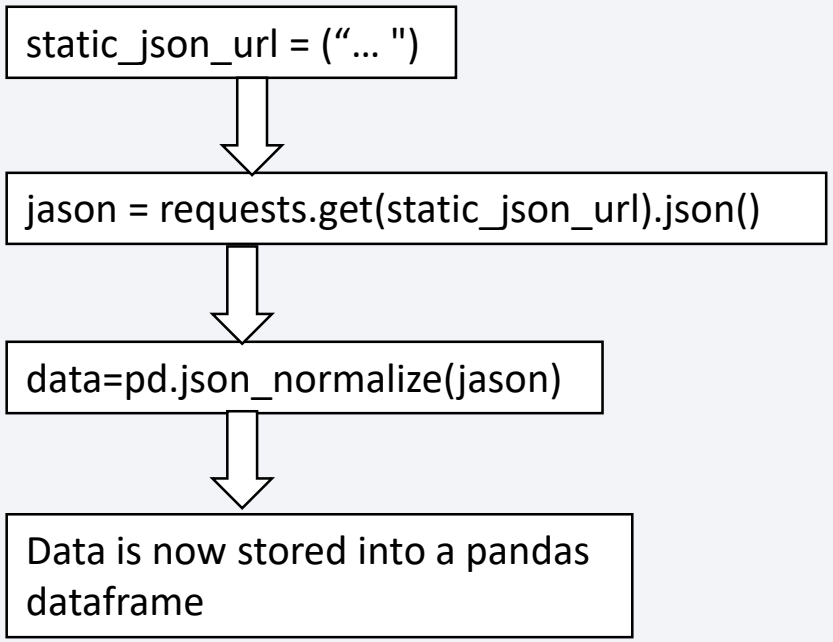
Data Collection

- The original dataset is available at a spacex API:

https://api.spacexdata.com.

- In some posterior instances, pre-prepared records of the data were used, provided by IBM instructors.
- The data was collected from the web using the *requests* library and stored into a Python *pandas* dataframe object.

Data Collection – SpaceX API



Jupyter notebook with the detailed process is available at:

https://github.com/luisfg10/IBM-Data-Science-Capstone-Project/blob/main/1-Data_Collection.ipynb

Data Wrangling

- Data was processed using python libraries *pandas* and *numpy.*
- For exploratory analysis, methods and attributes from pandas were used:

.columns, .dtypes, .value_counts(), etc.

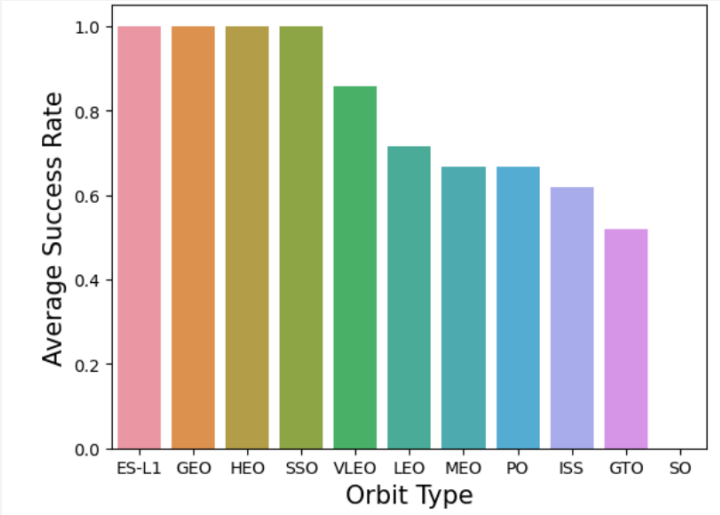
- The number of launches per type of orbit was calculated.
- Column “Outcome”, containing categorical data about the mission result, was converted to a numerical, binary column named “class”. “1” was adopted for successful missions, and “0” for failed ones.
- **It was found that 2/3 of all missions ended successfully.**

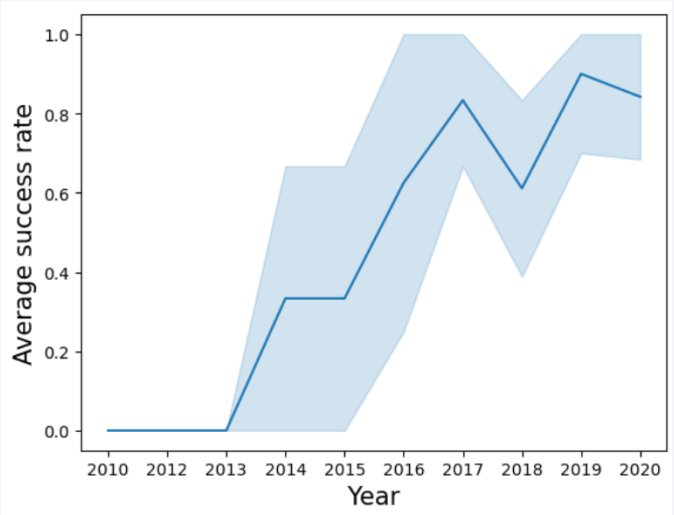
Jupyter notebook with the detailed process is available at:

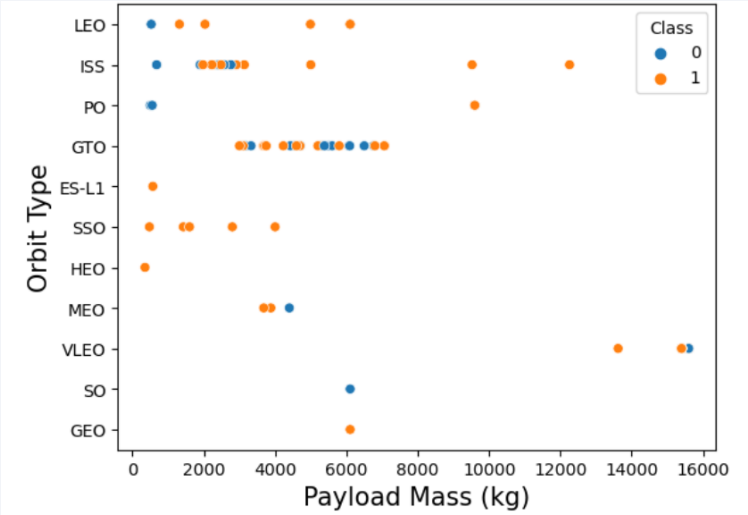
https://github.com/luisfg10/IBM-Data-Science-Capstone-Project/blob/main/2-Data_wrangling.ipynb

EDA with Data Visualization

- Plotting libraries *matplotlib* and *seaborn* were used.
- Some notable findings of the EDA follow:







- It´s interesting to observe the success rate is highly dependant of the orbit type and the previous experience (success improves with the years).

Jupyter notebook with the detailed process is available at:

https://github.com/luisfg10/IBM-Data-Science-Capstone-Project/blob/main/4-EDA_Viz.ipynb

EDA with SQL

- Library *sqlite3* was used, along with the sql magic extension.
- After the dataframe was loaded into an sqlite3 database, several queries were run. An example follows:

```
SELECT DISTINCT(Launch_Site)

 FROM SPACEXTABLE;
```

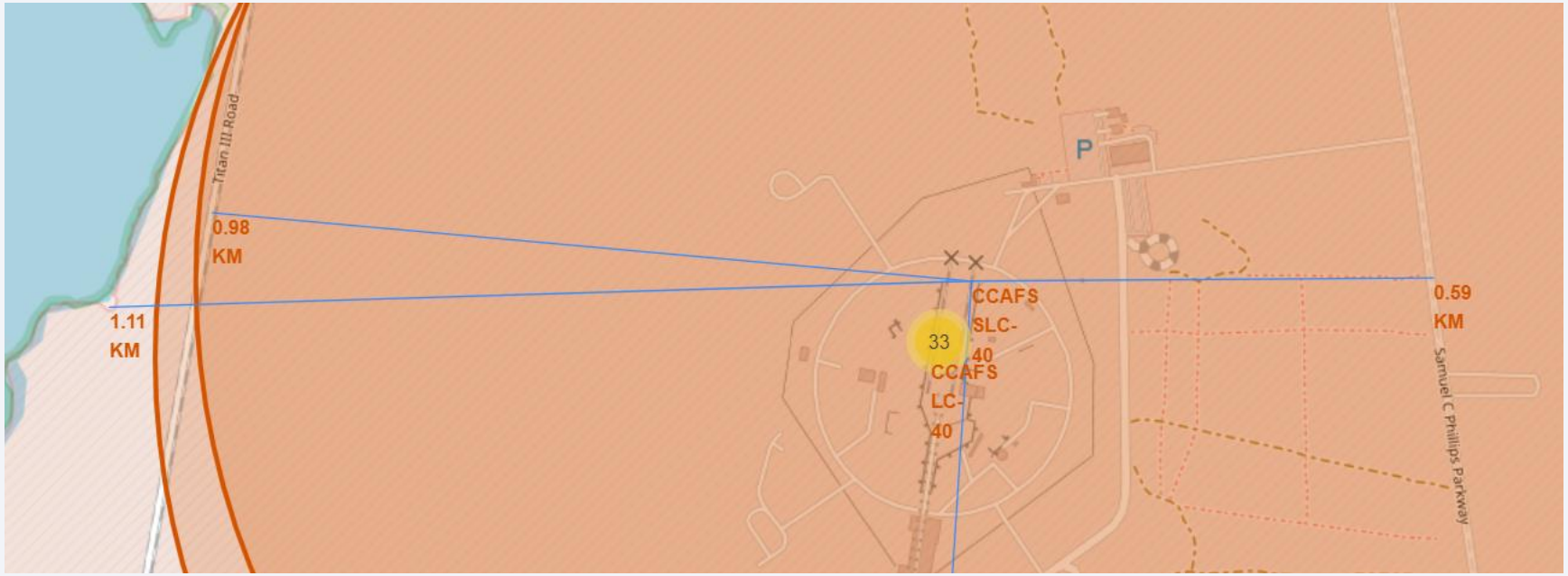
- It´s important to remark, given the low number of observations (less than 200), and the small overall size of the dataset, the use of SQL and relational databases is actually unnecessary and doesn´t give any benefits additional to pandas.

Jupyter notebook with the detailed process is available at:

https://github.com/luisfg10/IBM-Data-Science-Capstone-Project/blob/main/3-Exploratory-DA-SQL.ipynb

Build an Interactive Map with Folium

- To facilitate the visualization of geospatial data(launch sites), *folium* library was used.
- Additional tolos like mouse position, marker clusters and polylines were used in order to enhance the overall utility of the map.

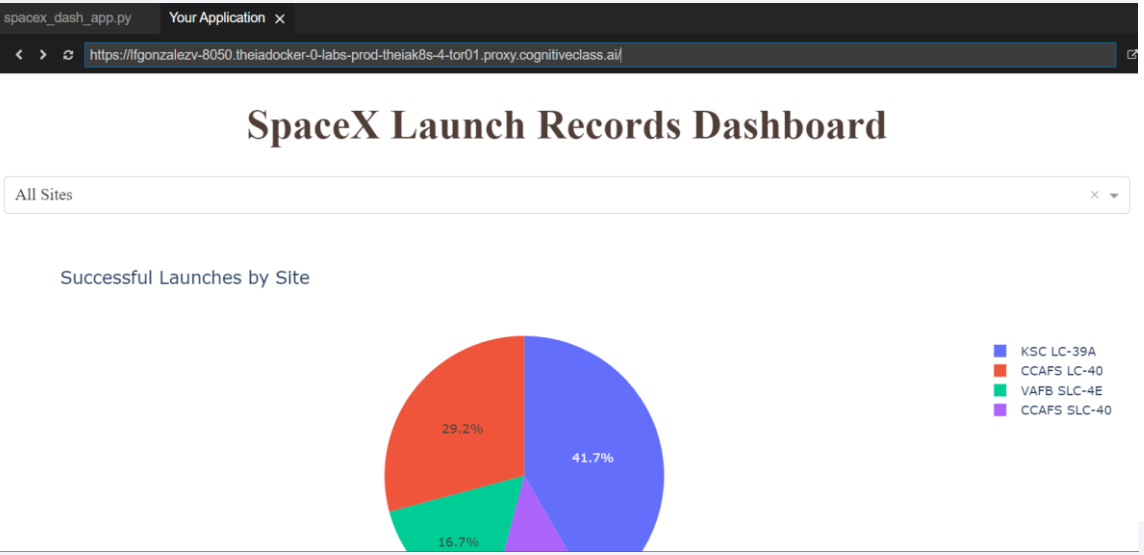


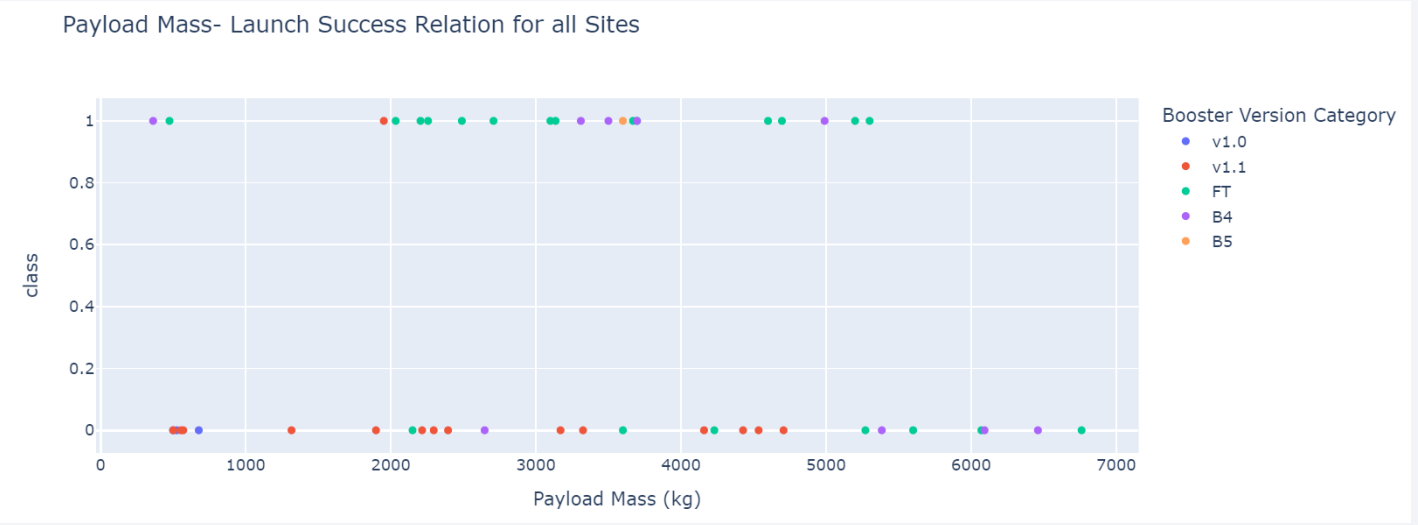
Jupyter notebook with the detailed process is available at:

https://github.com/luisfg10/IBM-Data-Science-Capstone-Project/blob/main/5-Folium.ipynb

Build a Dashboard with Plotly Dash

- The web app for interactive dashboards is capable of showing pie charts and scatter plots for a selected launch site, and range of payload mass.





Python script with the detailed process is available at:

https://github.com/luisfg10/IBM-Data-Science-Capstone-Project/blob/main/6-spacex_dash_app_complete.py

Predictive Analysis (Classification)

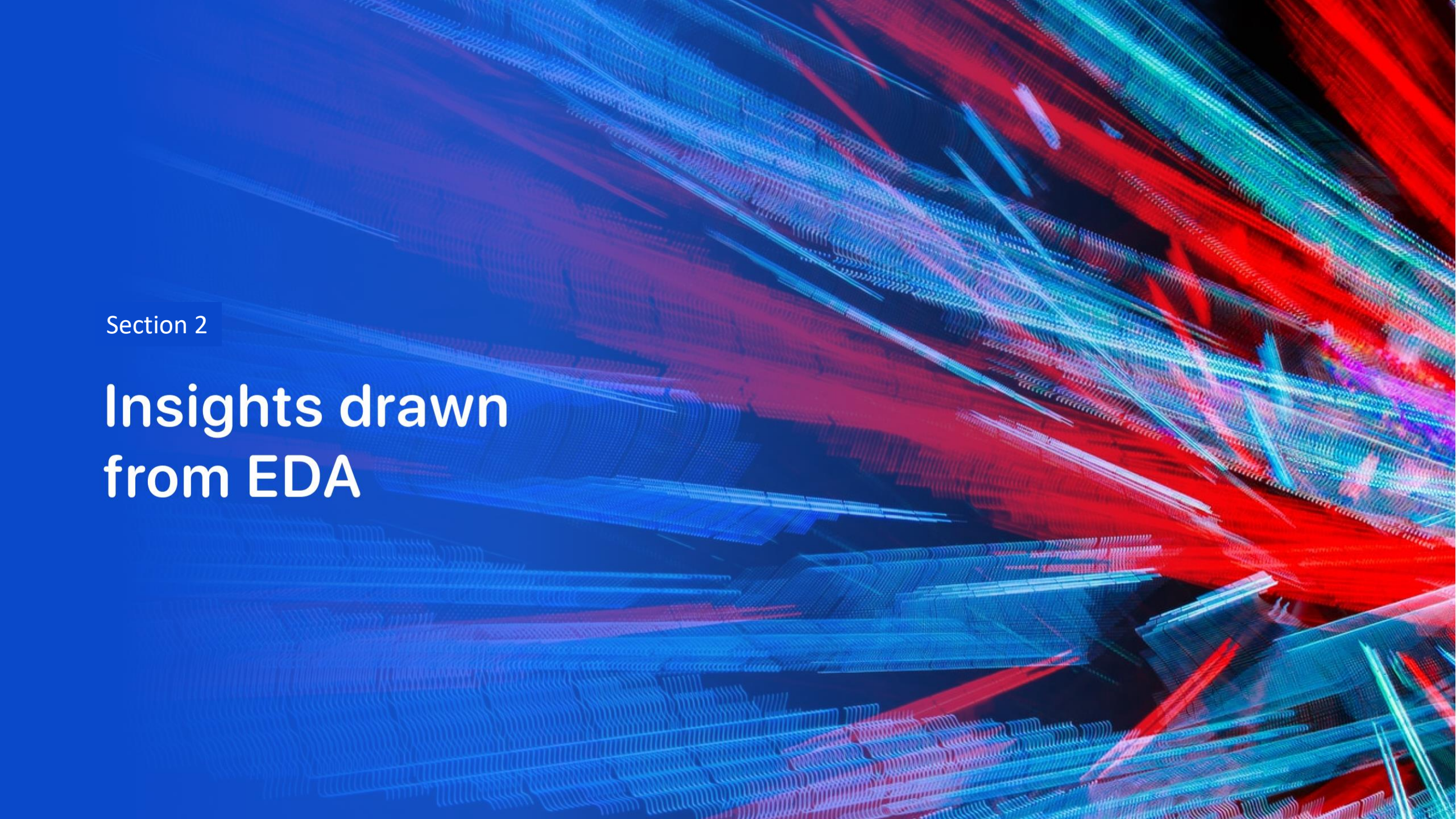
- 4 classification models were considered: Logistic Regressiom, K-Nearest Neighbors, Decision Trees and Support Vector Machine.
- A train-test split of the data was created, with a 20% allocation for testing.
- Using GridSearch, a tool for cross validation, the ideal parameters for each type of algorithm were found.
- All 4 models had the same out-of-simple accuracy of 83.33%. However, **logistic regression** was chosen as the best model, because of its easy interpretability and higher model simplicity.

Python script with the detailed process is available at:

https://github.com/luisfg10/IBM-Data-Science-Capstone-Project/blob/main/7-ML_Prediction.ipynb

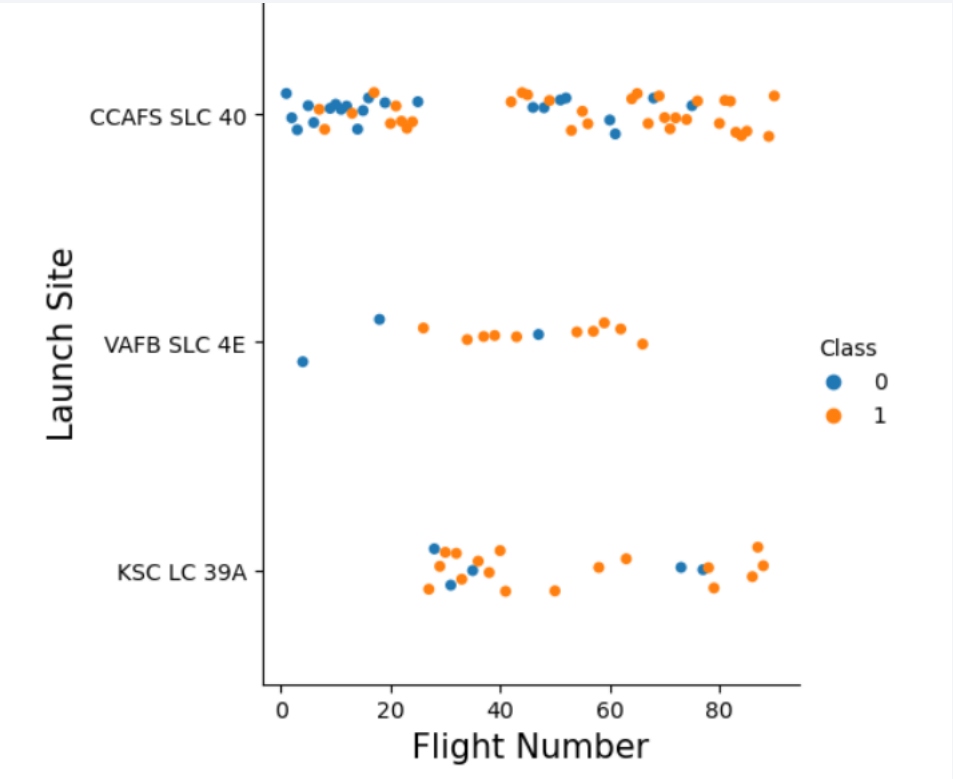
Results

- The developed logistic regression model is capable of predicting with an accuracy of 83.33% if a rocket misión will successfully land or not, given its payload mass, launch site, booster versión, among other variables.


Section 2
Insights drawn from EDA

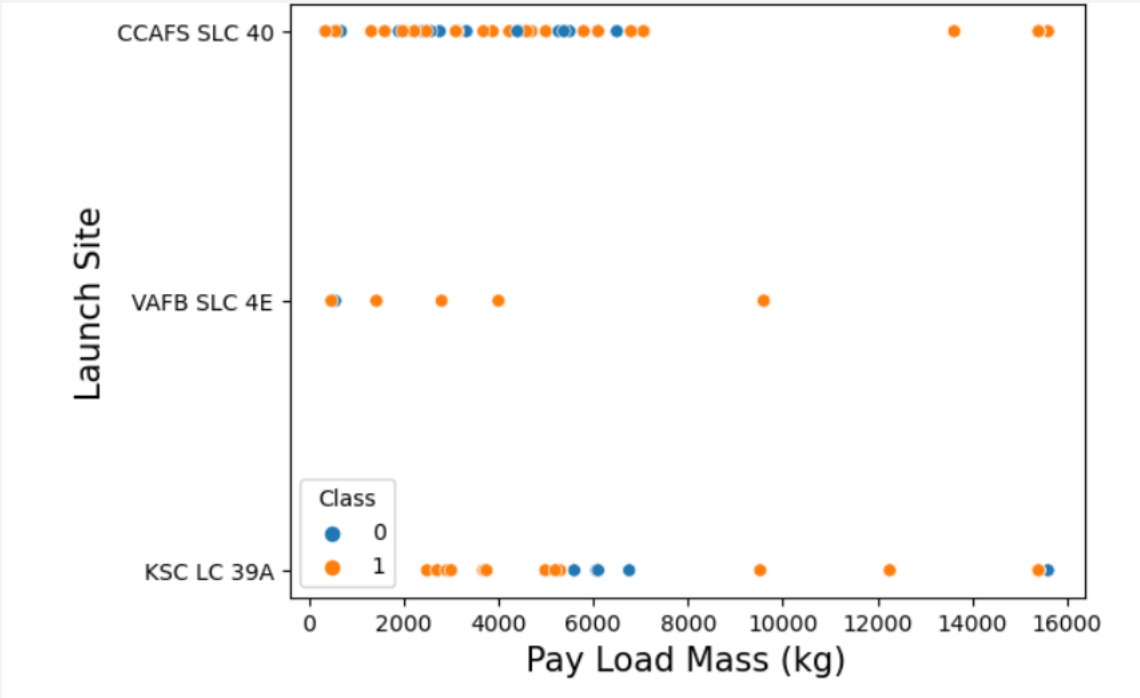
Flight Number vs. Launch Site

- The first launch site appears to have the most failures and the last the most successes.



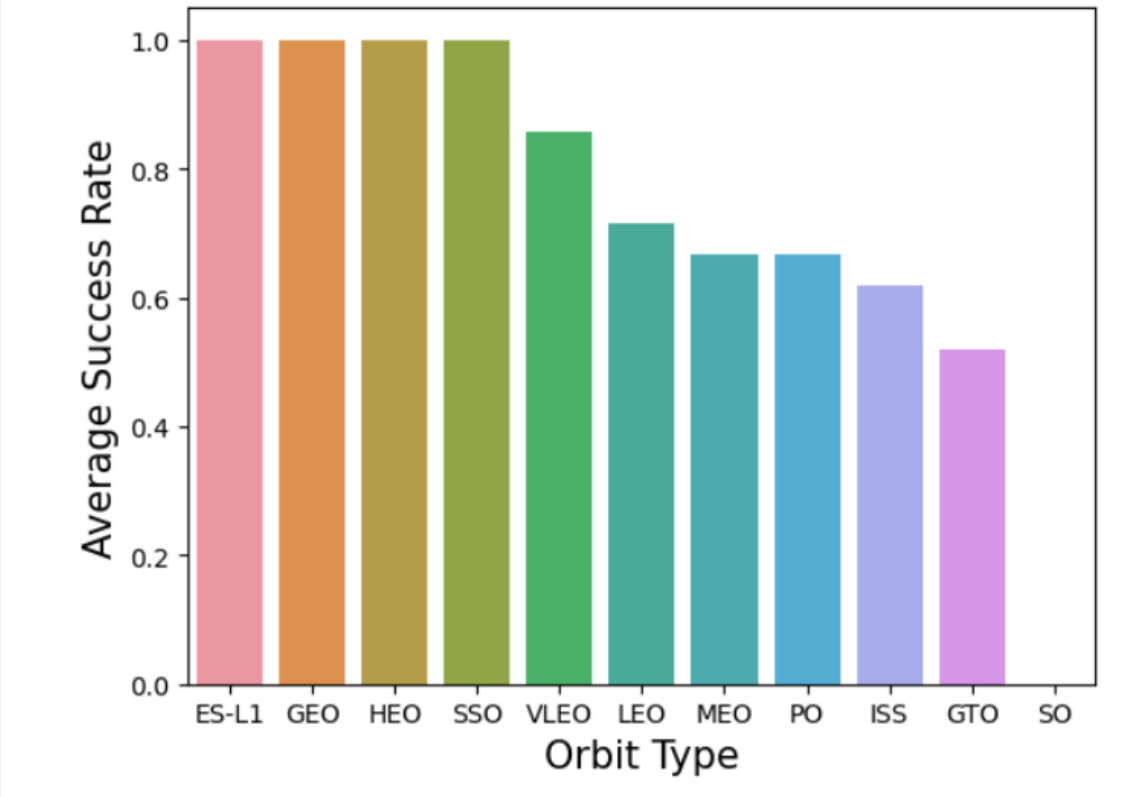
Payload vs. Launch Site

- The VAFB-SLC launchsite has no rockets launched for heavy payload mass.



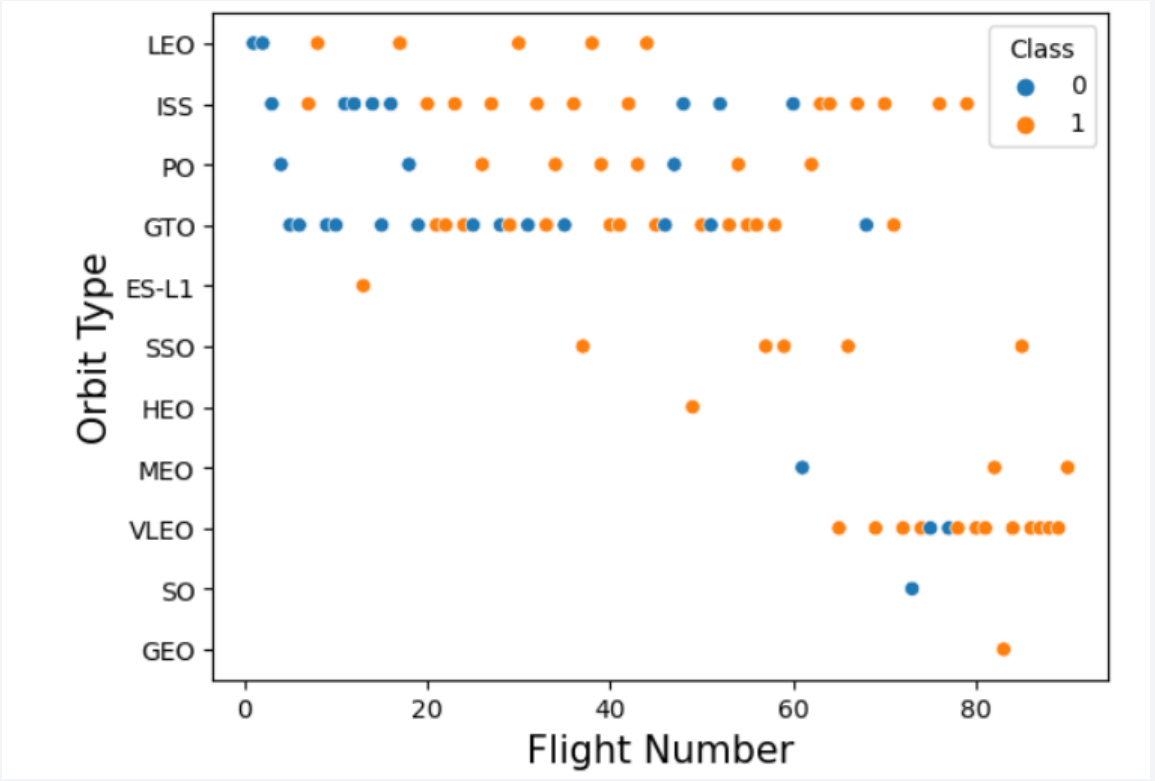
Success Rate vs. Orbit Type

- ES-L1 orbit has the highest success rate, and SO the lowest.



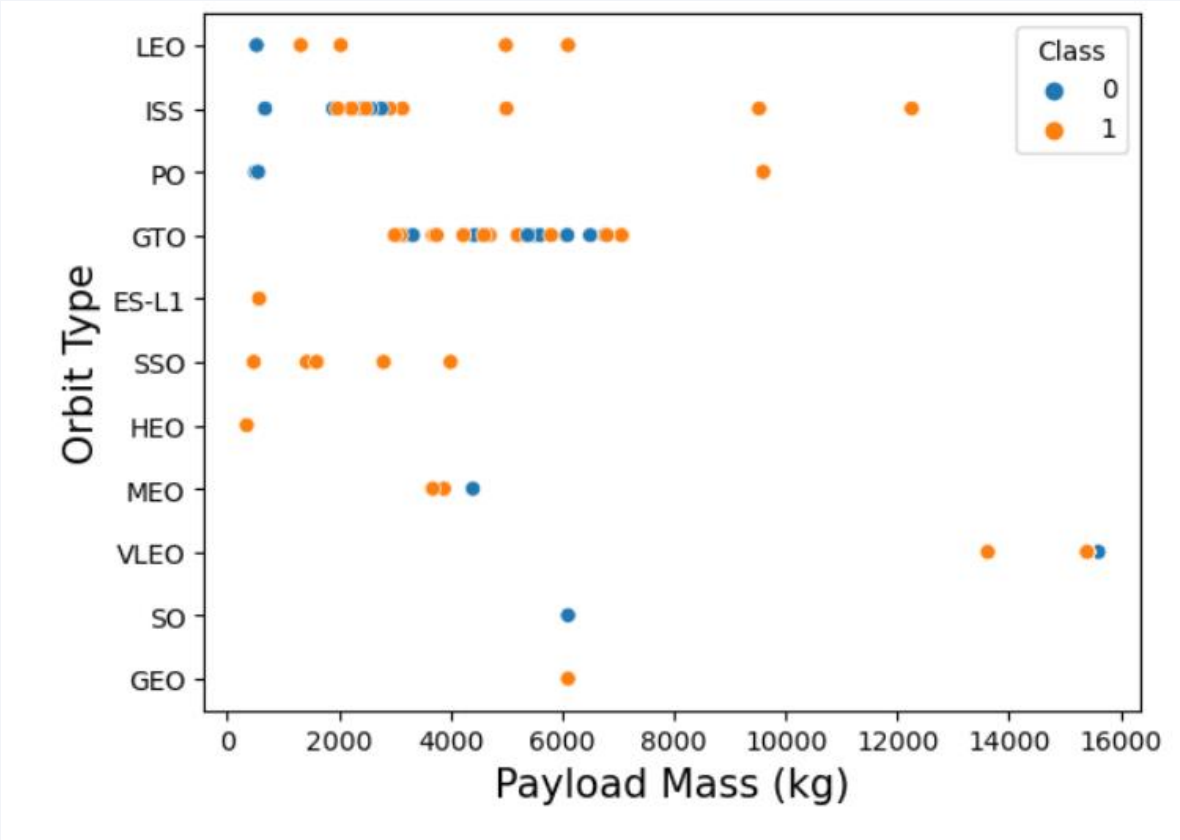
Flight Number vs. Orbit Type

- It appears orbit types have moved towards VLEO, from flight 60 onwards.



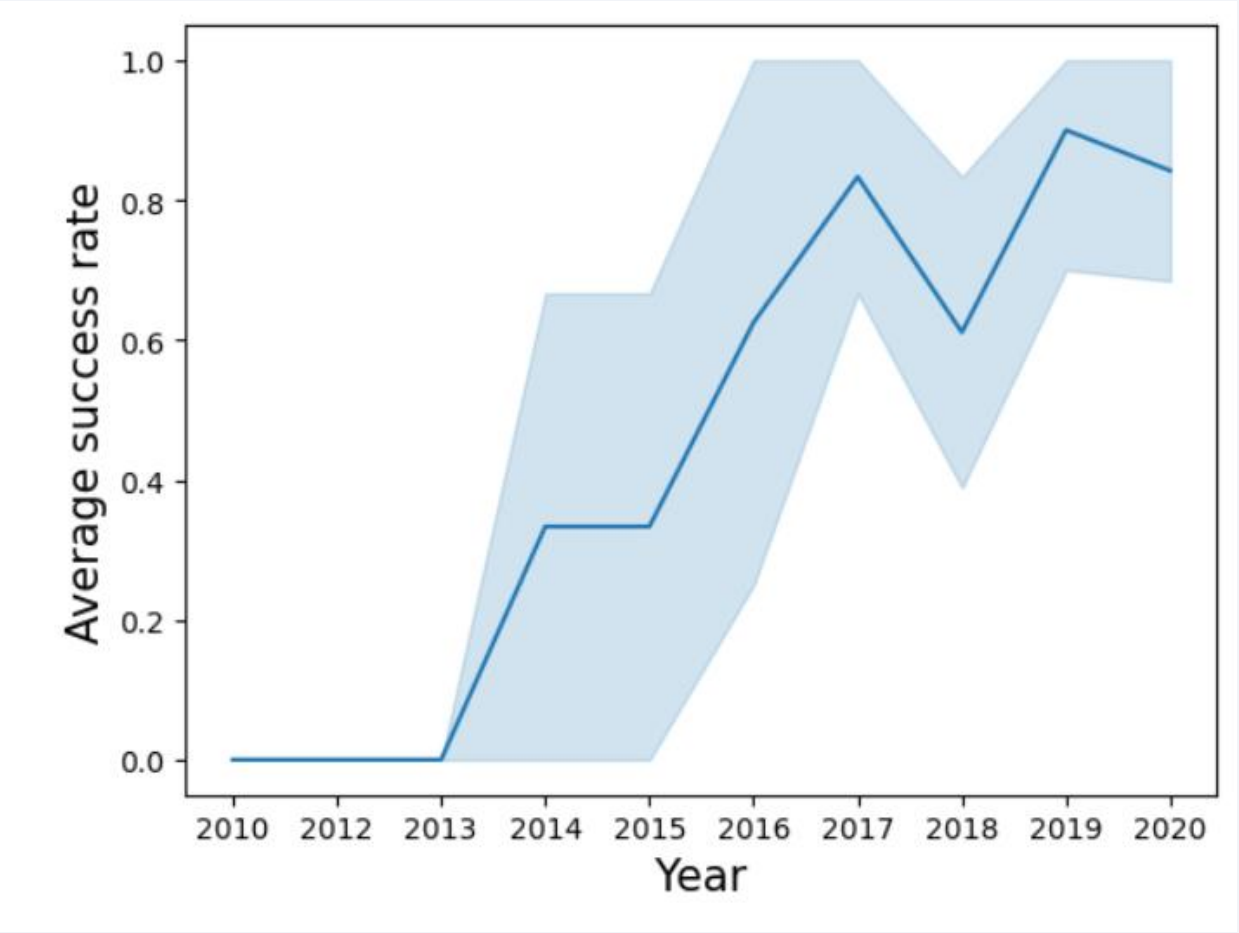
Payload vs. Orbit Type

- There appears to be no obvious relation between payload mass and orbit type.



Launch Success Yearly Trend

- We can see the success rate has consistently improved for SpaceX along the years.



All Launch Site Names

- There´s four different launch site names.

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

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass from NASA

- The total carried mass on NASA missions is 45596 kg.

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

- On average, F9 v1.1 boosters carry 2534.7 kg of load.

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- This is the first registered record of a successful ground pad landing.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

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

Boosters Carried Maximum Payload

- These are the booster names that have carried the maximum registered load.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- Failed drone ship missions

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-10-01	09:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

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

- These outcomes happened during the selected timeframe:

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

Section 3

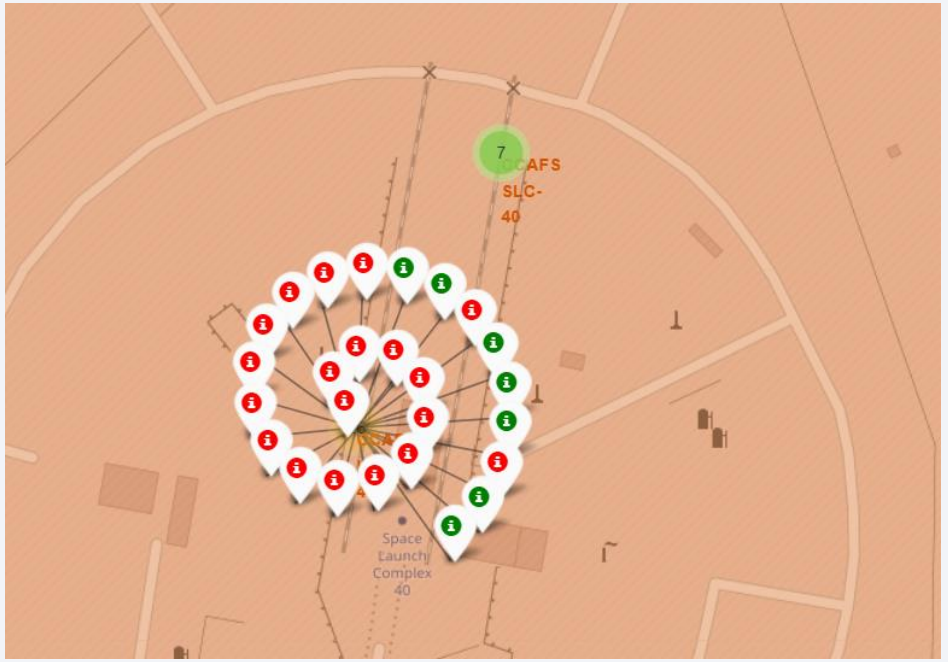
Launch Sites Proximities Analysis

Launch Site Locations

- Here´s the location of the different launch sites for SpaceX missions. One is in LA, the other 3 in Florida.

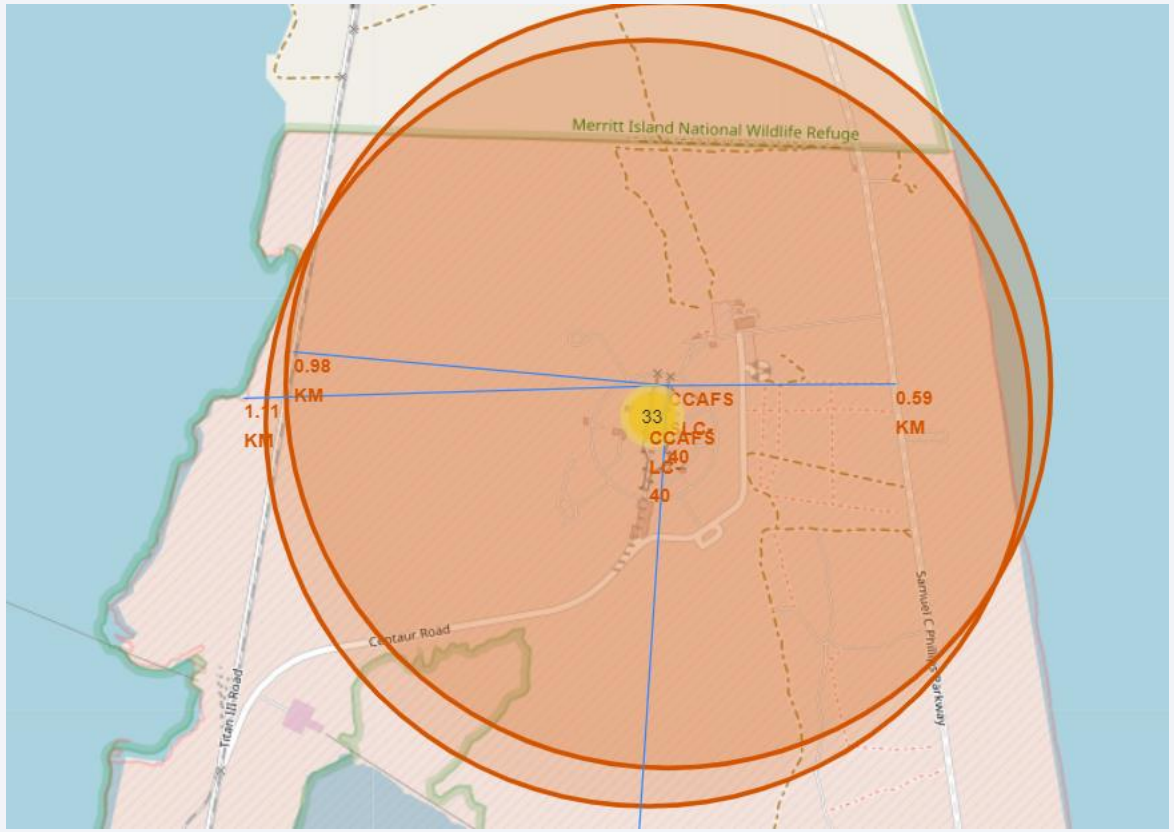
Launches by Site

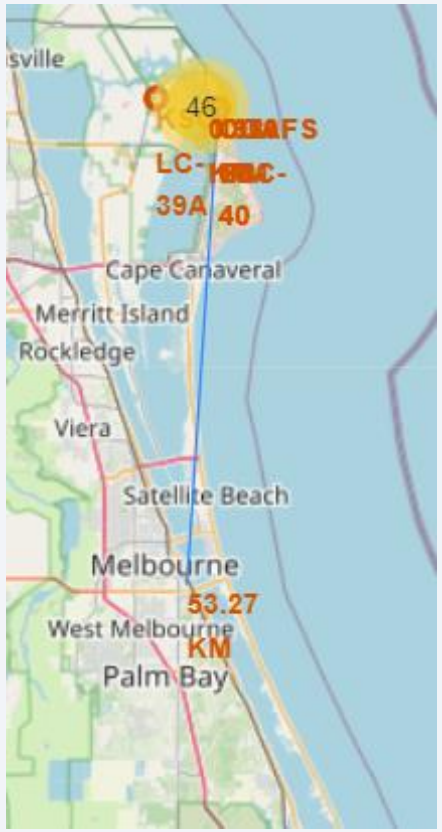
- Here, we can see all the launches, or ocurrences of launches, per site. For image simplicity, only those in Florida are shown.

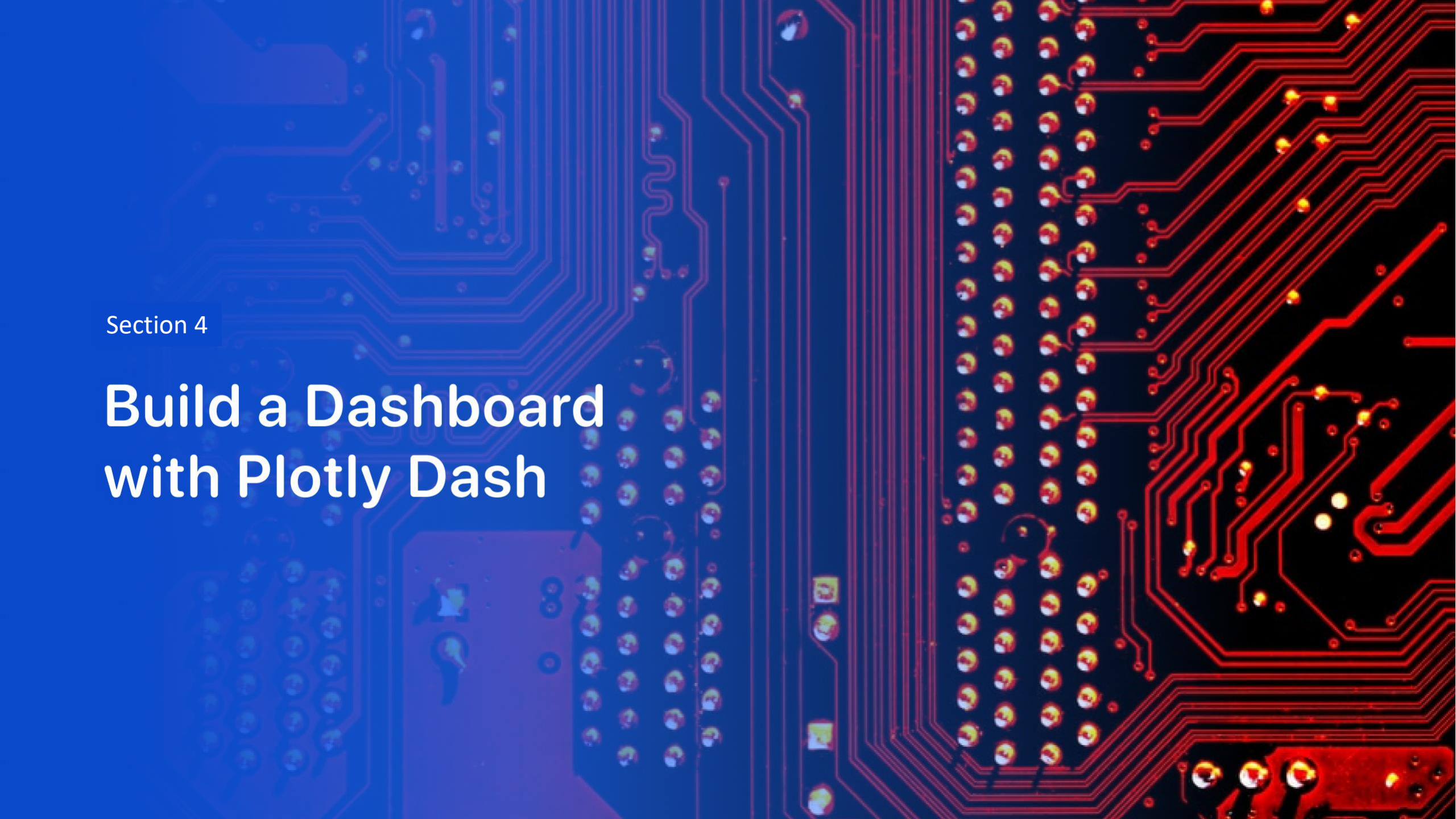


Launch Site Proximities

- In this image, we can see launch site proximities to coastlines, railways and cities.

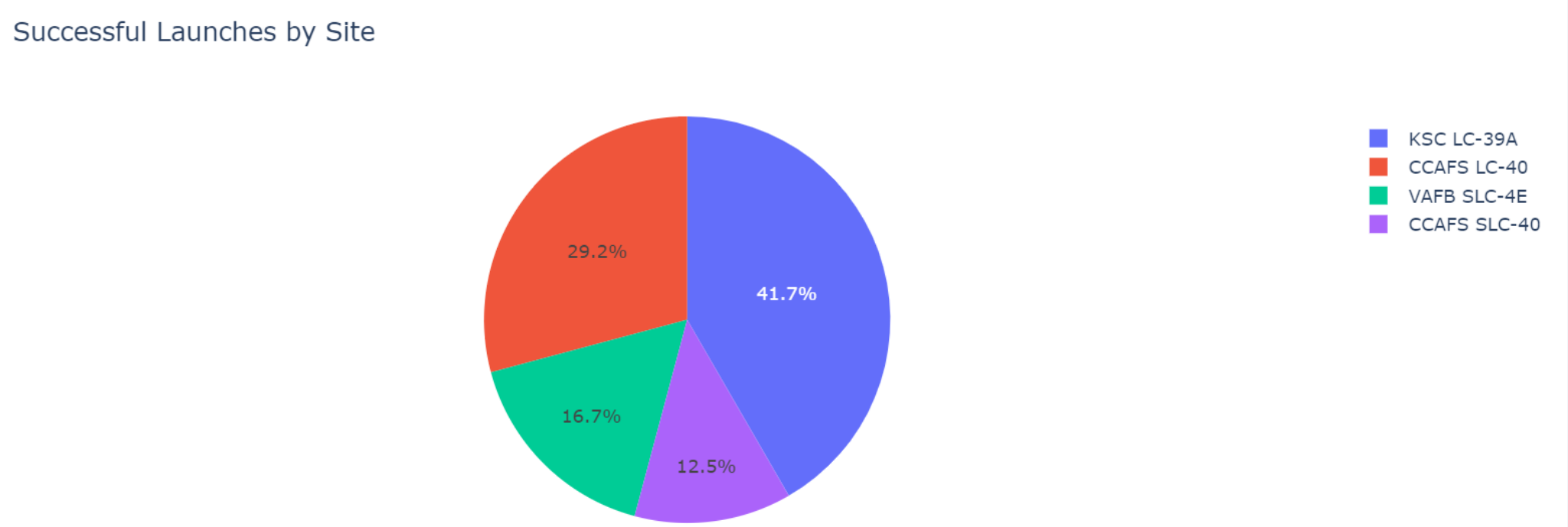





Section 4
Build a Dashboard with Plotly Dash

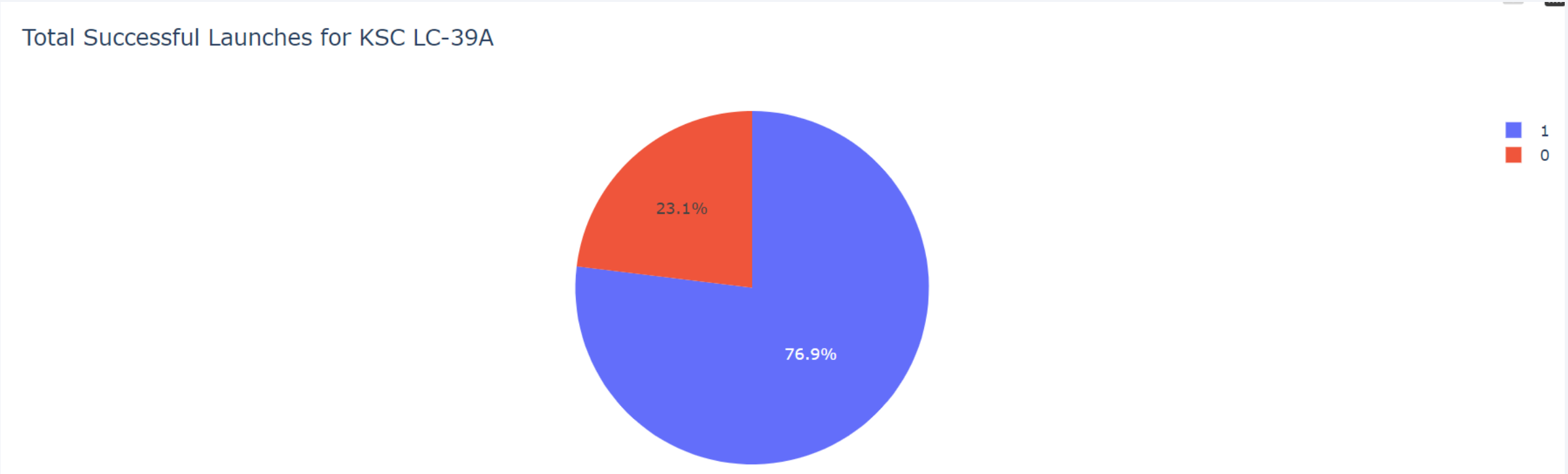
Successful launches by site

The most successful launches come from site KSC LC-39A.



Site with highest success ratio

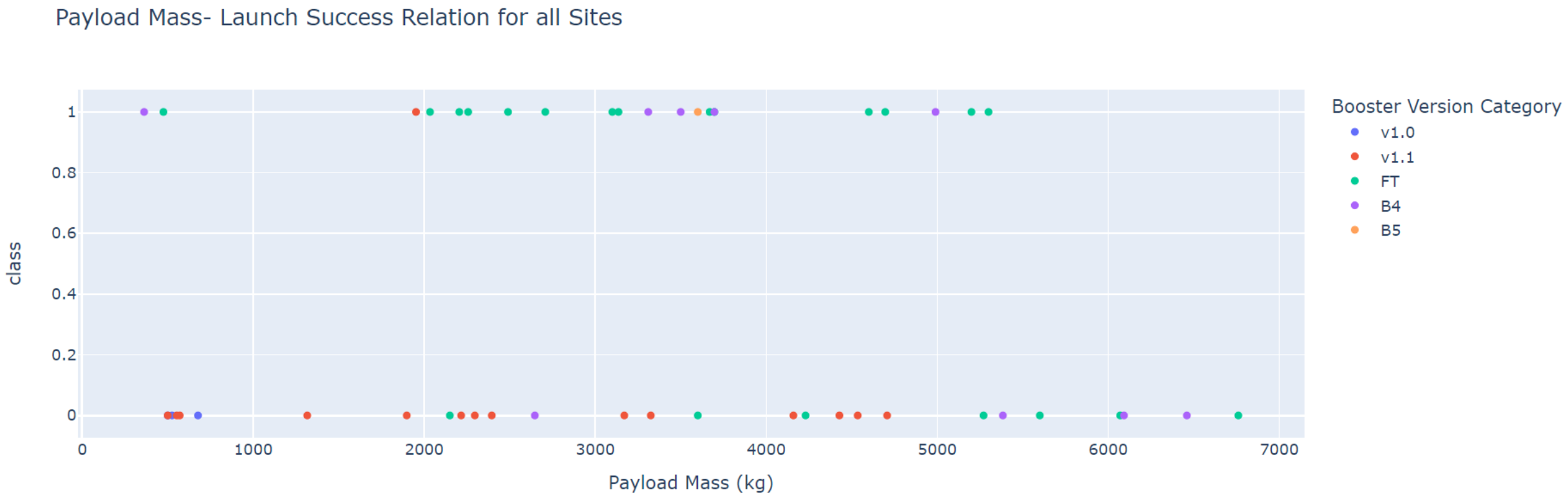
- Site KSC LC-39A has the highest success ratio.



Payload vs outcome for all sites

Payload range (Kg):

0 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k

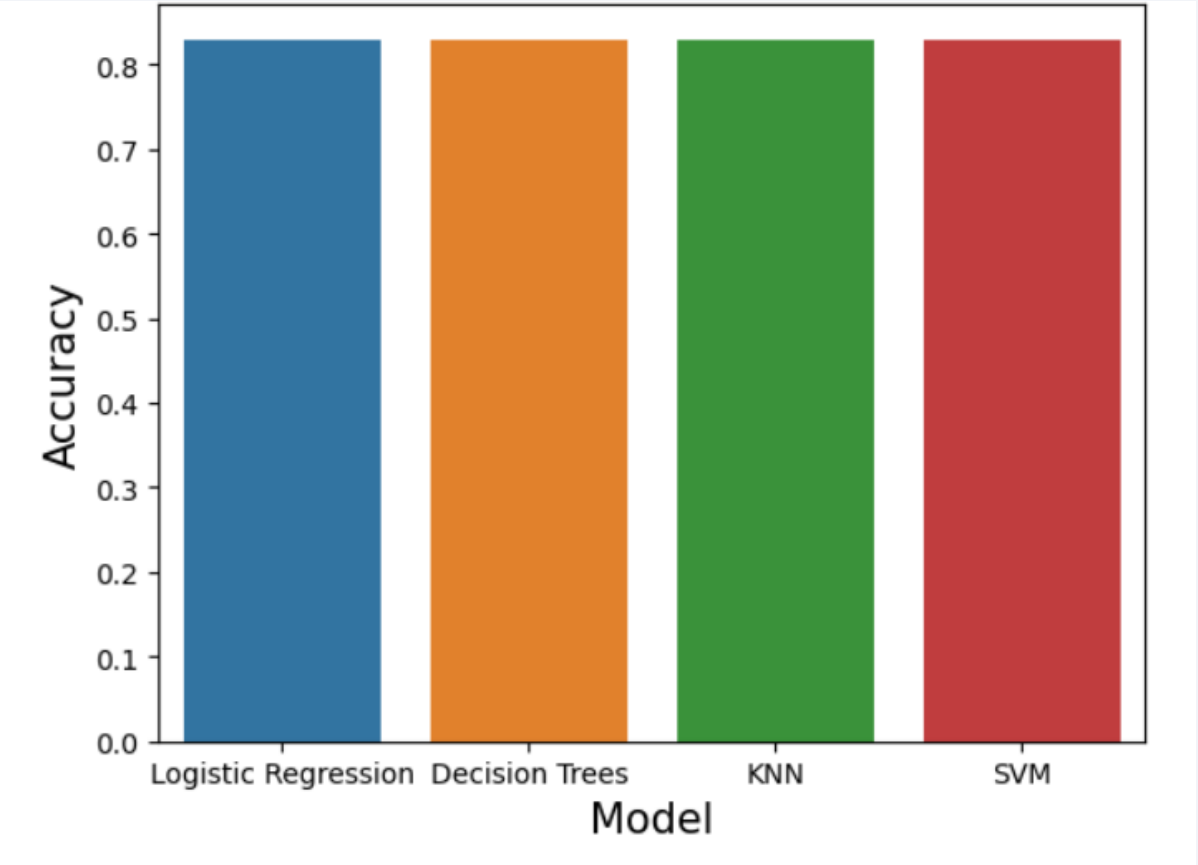


Section 5

Predictive Analysis (Classification)

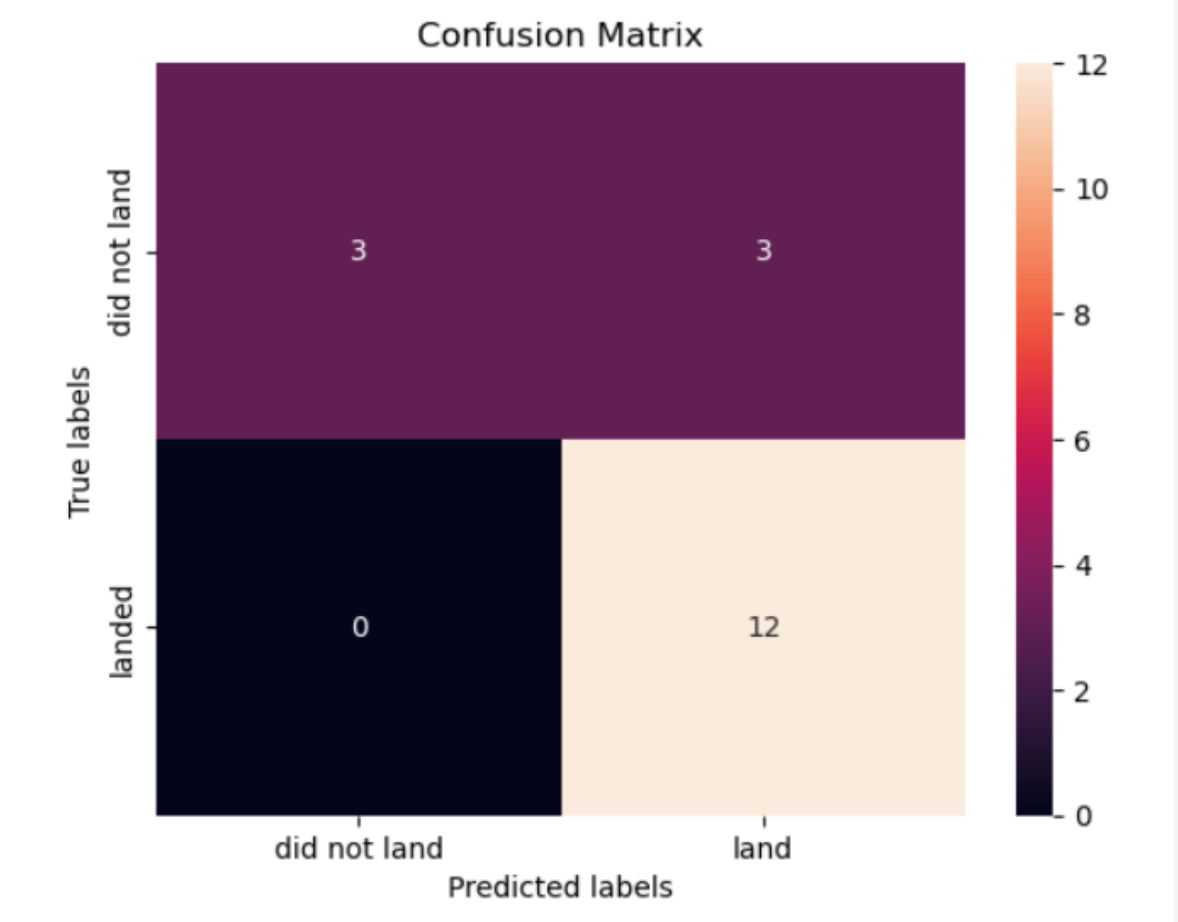
Classification Accuracy

Even though all models show the same out-of-simple accuracy, **logistic regression** is chosen because of its easy interpretability and simplicity.



Confusion Matrix

- We can see that 12 results are true positives, 3 are true negatives, 0 are false negatives and 3 are false positives.



Conclusions

- Logistic regression is the most convenient ML model for predicting the occurrence of success or failure of a rocket mission.
- Part of SpaceX´s success lies in improving year after year, as they have been able to improve their success ratios consistently for years.
- To improve the accuracy of the existing models, it´s important to explore more parameter tuning, as well as to have more data available, if necessary from other companies.

Thank you!