IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Christian Reyes>
<July 1 2023>

Outline

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

Executive Summary

Summary of methodologies

- ➔ Perform exploratory data analysis (EDA) using visualization and SQL
- ➔ Perform interactive visual analytics using Folium and Plotly Dash
- ➔ Perform predictive analysis using classification models
- ➔ How to build, tune, evaluate classification models

Summary of all results

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

Introduction

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

Desirable answers:

The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;

Where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

Using SpaceX Rest API

- Using Web Scrapping from Wikipedia

Perform data wrangling

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data to a binary classification

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

-Building, tuning and evaluation of classification models to ensure the best results.

Data Collection

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

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

Data Columns are obtained by using SpaceX REST API:

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

Data Columns are obtained by using Wikipedia Web Scraping:

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

Data Collection – SpaceX API

Requesting rocket launch data from SpaceX API

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

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

Constructing data we have obtained into a dictionary

Creating a dataframe from the dictionary

Filtering the dataframe to only include Falcon 9 launches

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

Exporting the data to CSV

https://github.com/Christactic/Data-Science-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Web Scraping

Requesting Falcon 9 launch data from Wikipedia → Creating a BeautifulSoup object from the HTML response → Extracting all column names from the HTML table header → Collecting the data by parsing HTML tables → Constructing data we have obtained into a dictionary → Creating a dataframe from the dictionary → Exporting the data to CSV

https://github.com/Christactic/Data-Science-Capstone-Project/blob/main/jupyter-labs-webscraping%20(1).ipynb

Data wrangling

Perform exploratory Data Analysis and determine Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting the data to CSV

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; an example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean mean the mission outcome was unsuccessfully landed to a specific region of the ocean.

https://github.com/Christactic/Data-Science-Capstone-Project/blob/main/module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

Charts were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

Scatter plots show the relationship between variables. If a relationship exits, they could be used in machine learning model. Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value

Line charts show trends in data over time (time series).

https://github.com/Christactic/Data-Science-Capstone-Project/blob/main/module_2_EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

Performed SQL queries:

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

https://github.com/Christactic/Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

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

Distances between a Launch Site to its proximities:

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

https://github.com/Christactic/Data-Science-Capstone-Project/blob/main/module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected

Slider of Payload Mass Range:

- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.

https://github.com/Christactic/Data-Science-Capstone-Project/blob/main/Hands-on%20Lab:%20Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash

Predictive Analysis (Classification)

Creating a NumPy array from the column “Class” in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy on the test data using the method.score() for all models

Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

https://github.com/Christactic/Data-Science-Capstone-Project/blob/main/module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

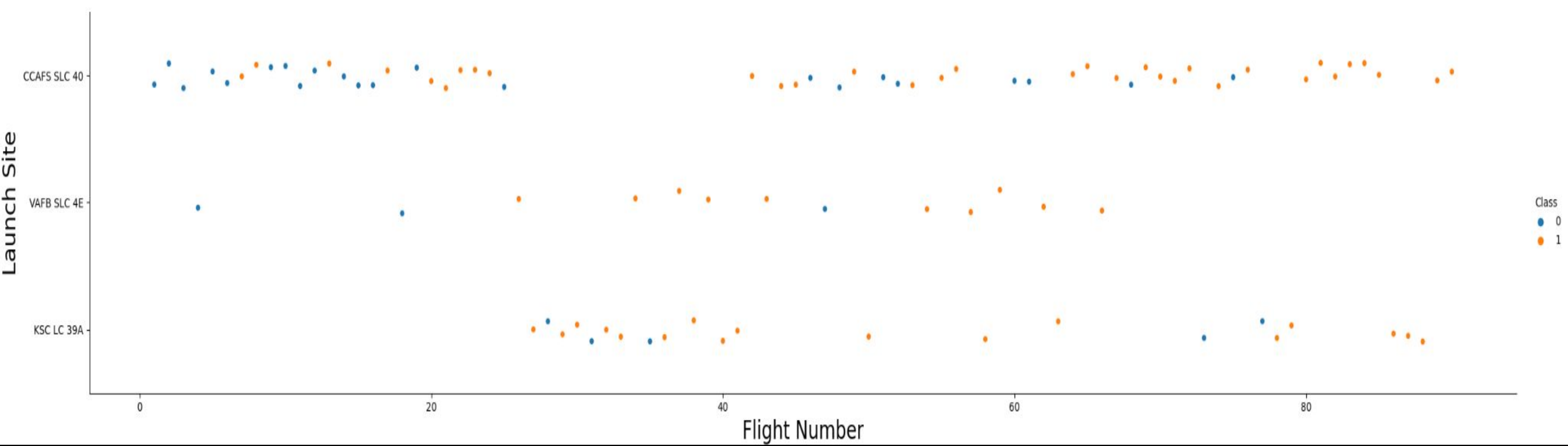
Results

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

Section 2

Insights drawn from EDA

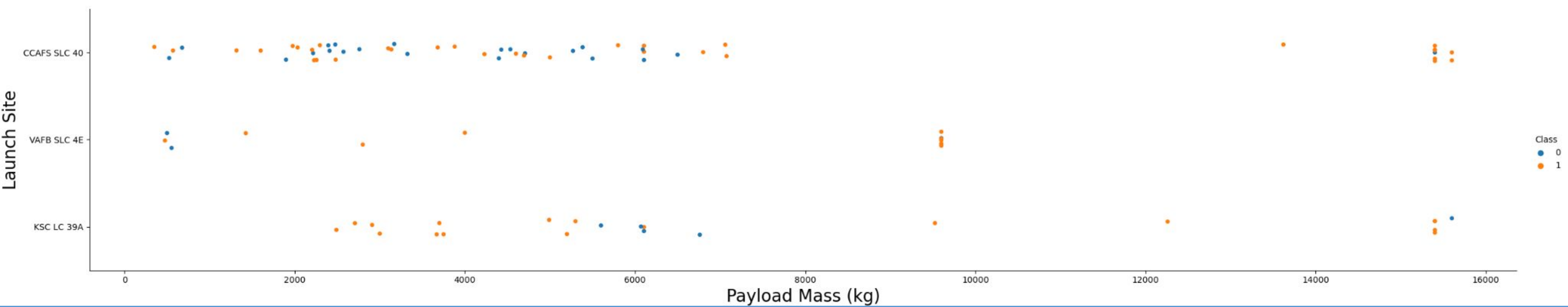
Flight Number vs. Launch Site



Explanation:

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

Payload vs. Launch Site



Most of the launches with payload mass over 7000 kg were successful.

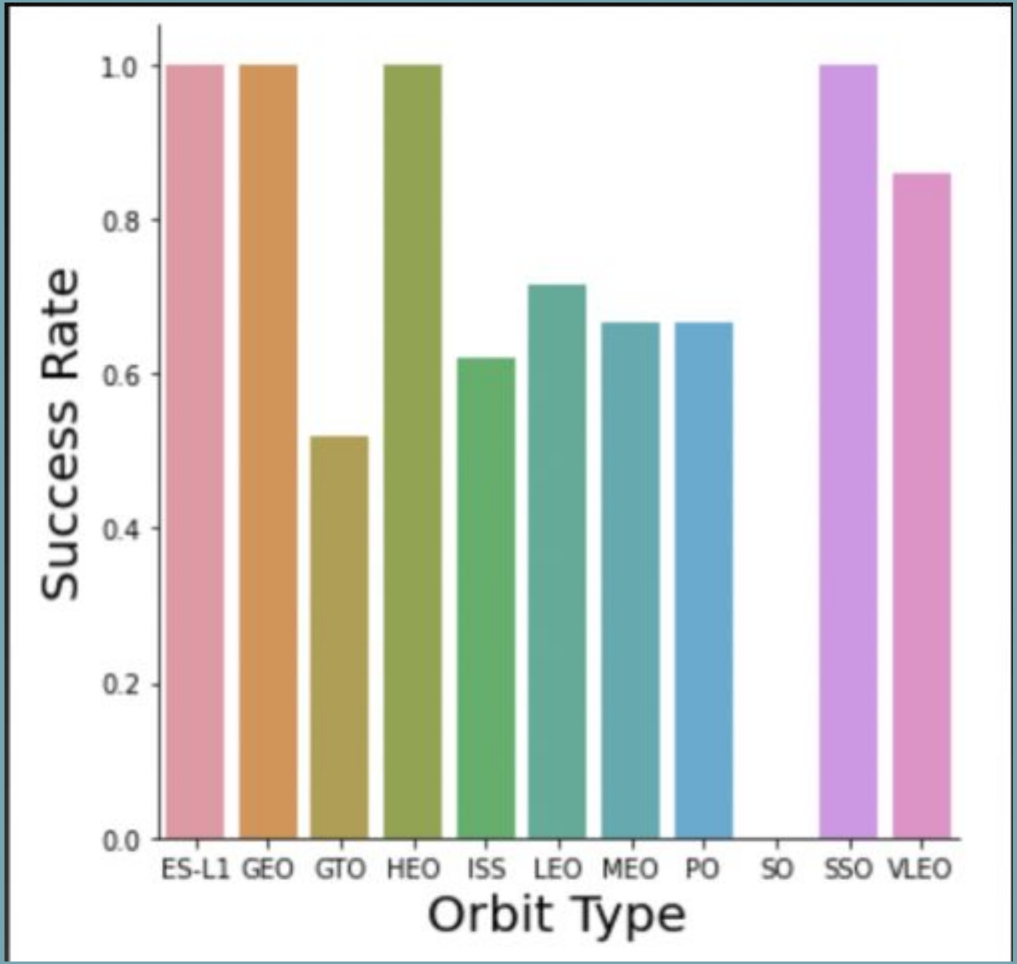
KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

For every launch site the higher the payload mass, the higher the success rate.

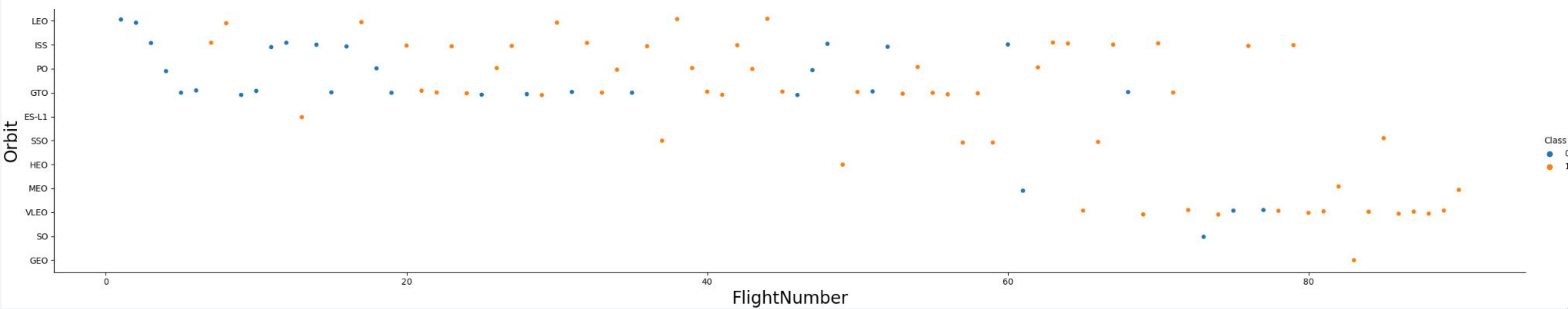
Success Rate vs. Orbit Type

Explanation:

- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO
- Orbits with 0% success rate:
  - SO



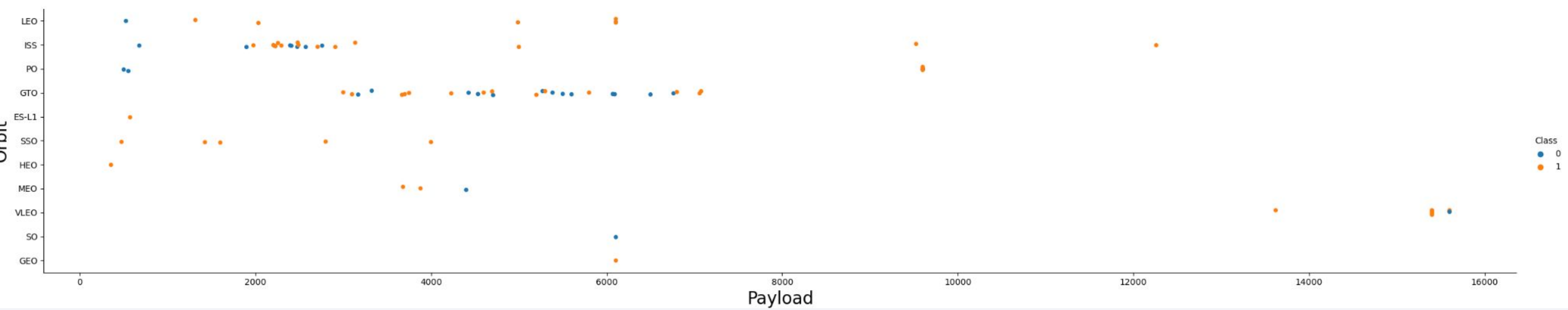
Flight Number vs. Orbit Type



Explanation:

The LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



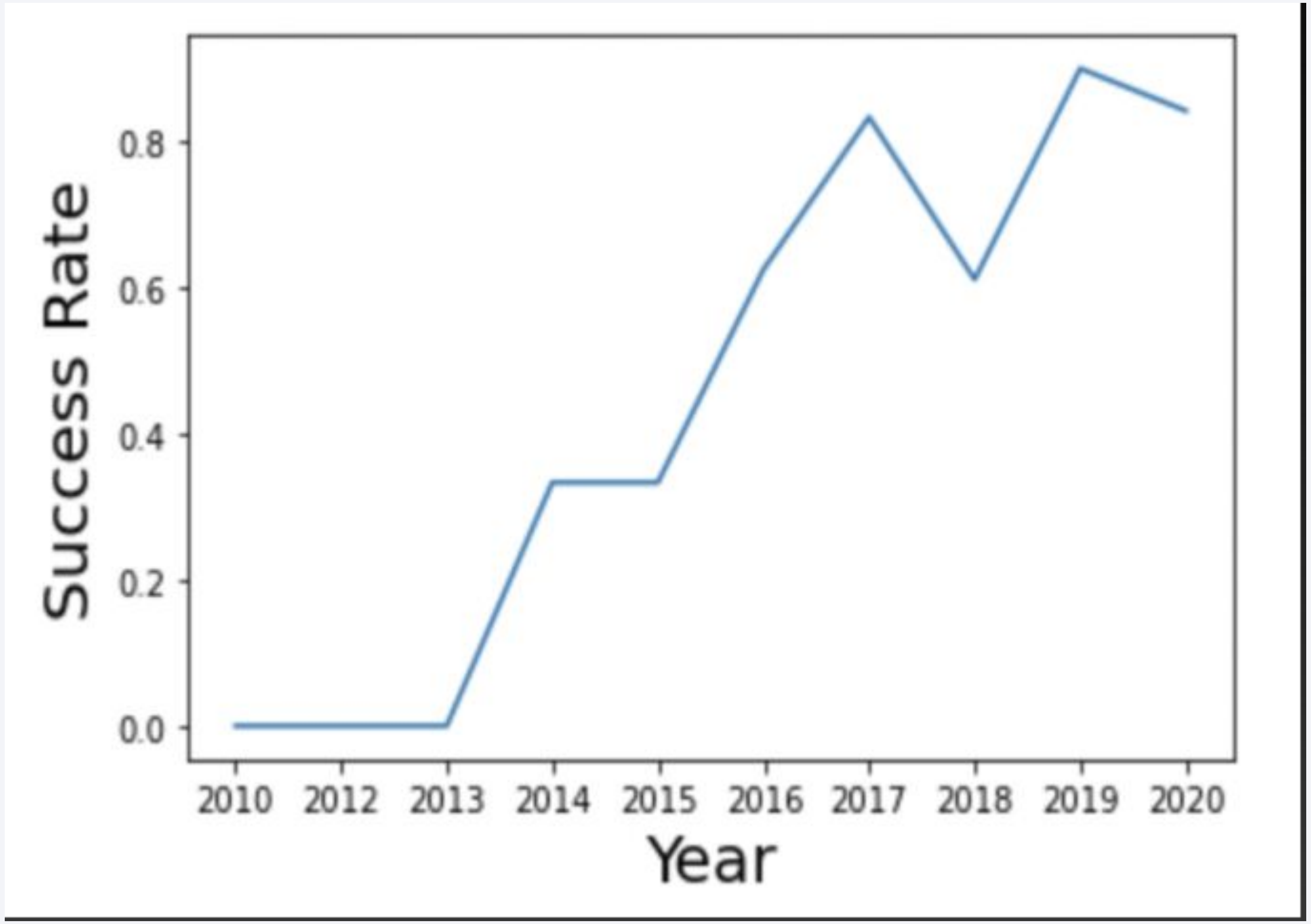
Explanations

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

Launch Success Yearly Trend

Explanation:

The success rate since 2013 kept increasing till 2020.



All Launch Site Names

Out[4]:

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

Explanation:

- Displaying the names of the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

Out[5]:

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:

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

Total Payload Mass

Out[6]:

total_payload_mass
45596

Explanation

- Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

Out[7]:

average_payload_mass
2534

Explanation

- Displaying average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

Out[8]:

first_successful_landing
2015-12-22

Explanation:

- Listing the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
Done.
Out[9]:
```

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

Explanation:

- Listing 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

Out[10]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Explanation:

- **Listing the total number of successful and failure mission outcomes**

Boosters Carried Maximum Payload

Out[34]:

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:

- **Listing the names of the booster versions which have carried the maximum payload mass.**

2015 Launch Records

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation:

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

landing__outcome	count_outcomes
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:

- **Ranking 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.**

Section 3

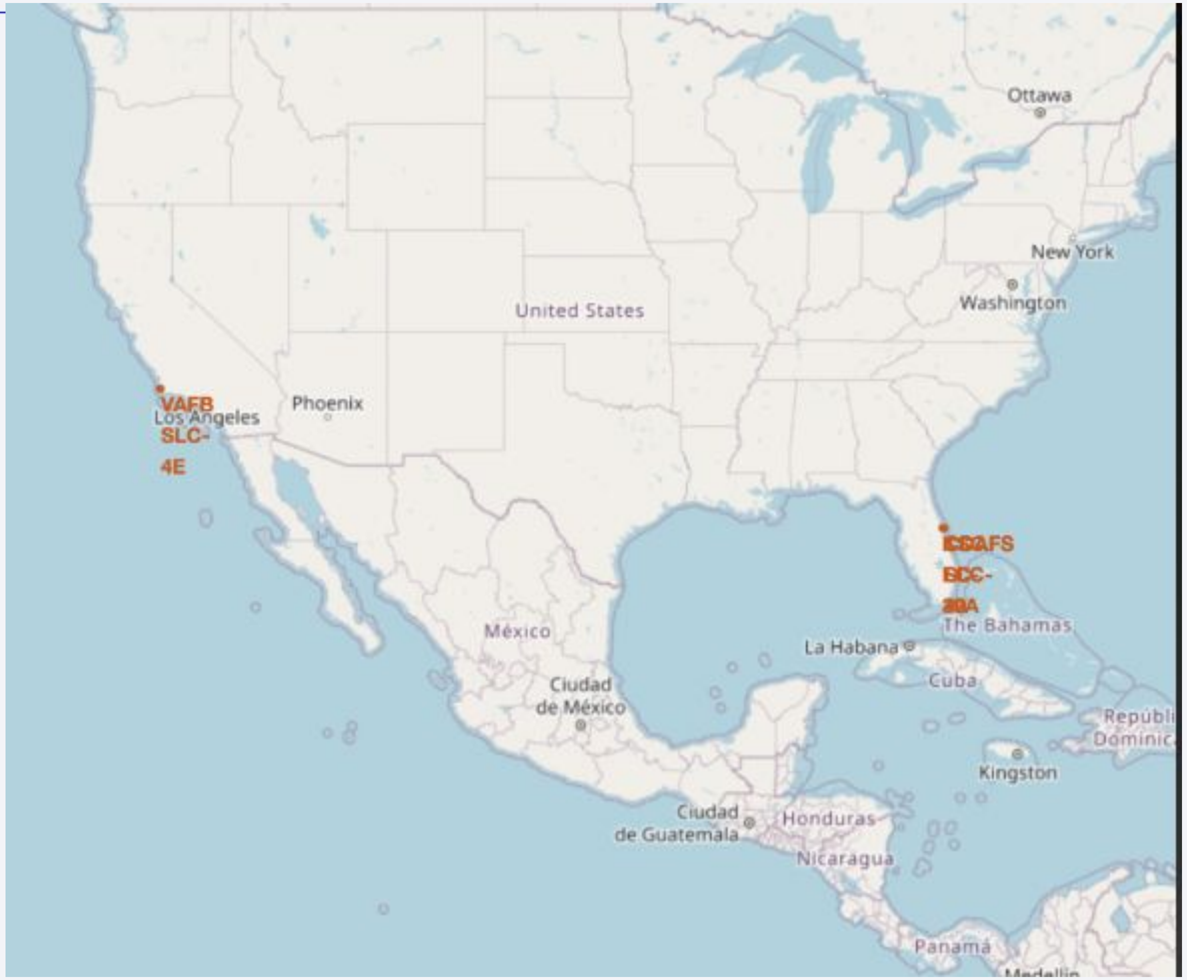
Launch Sites Proximities Analysis

Location markers on a global map

Explanation:

- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.

- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth.

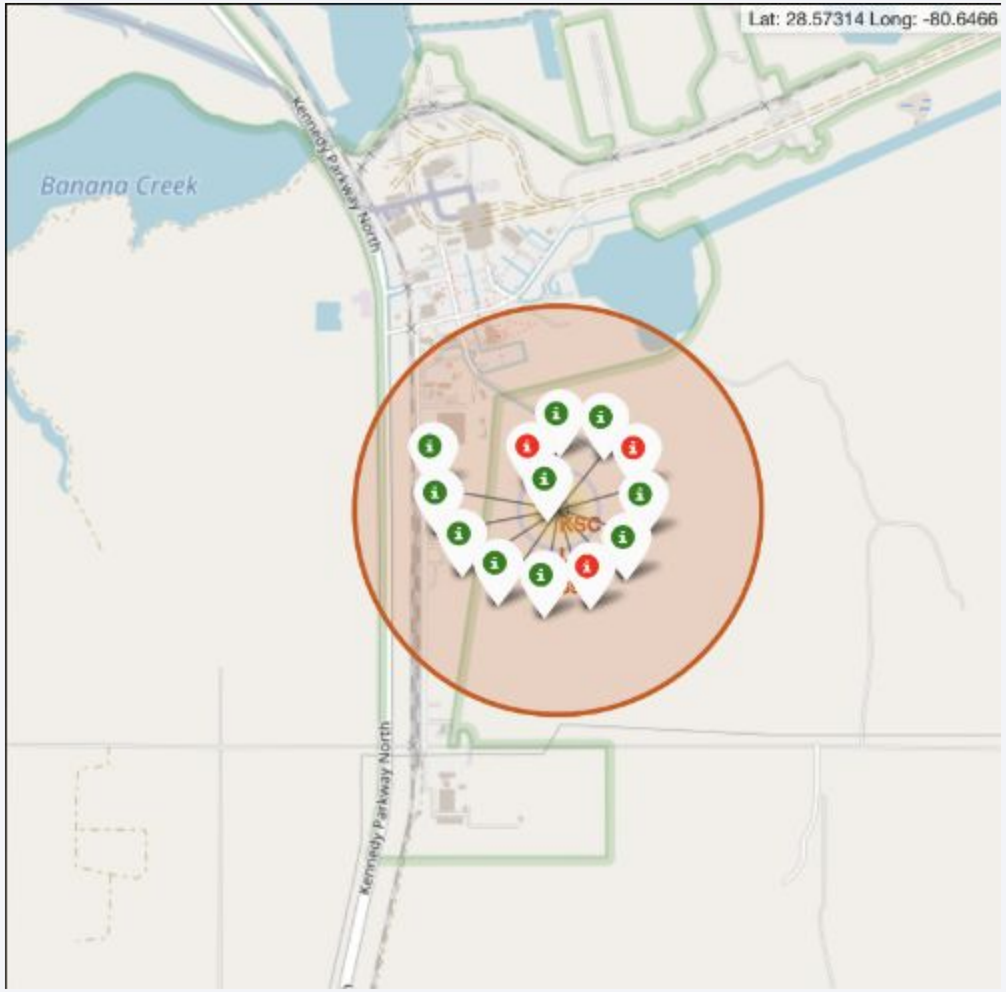


Records on the map of Colour-labeled launch

Explanation:

- From the colour-labeled markers we should be easily identify which launch sites have relatively high success rates.
- Green Marker = Successful Launch
- Red Marker = Failed Launch

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

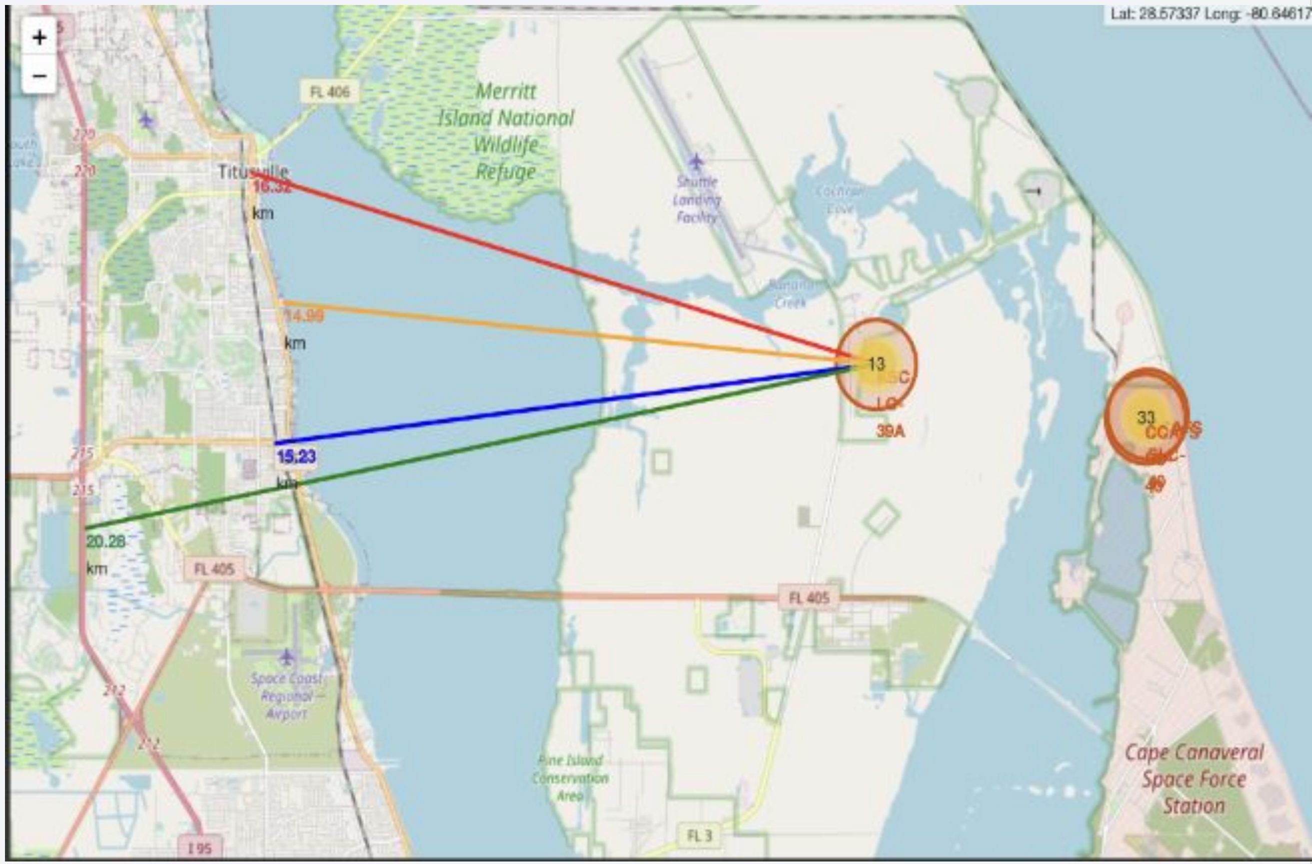


Distance from the launch site KSC LC-39A to its Proximities

Explanation:

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relatively close to railway (15.23km)
  - relatively close to highway (20.28km)
  - relatively close to coastline (14.99km)

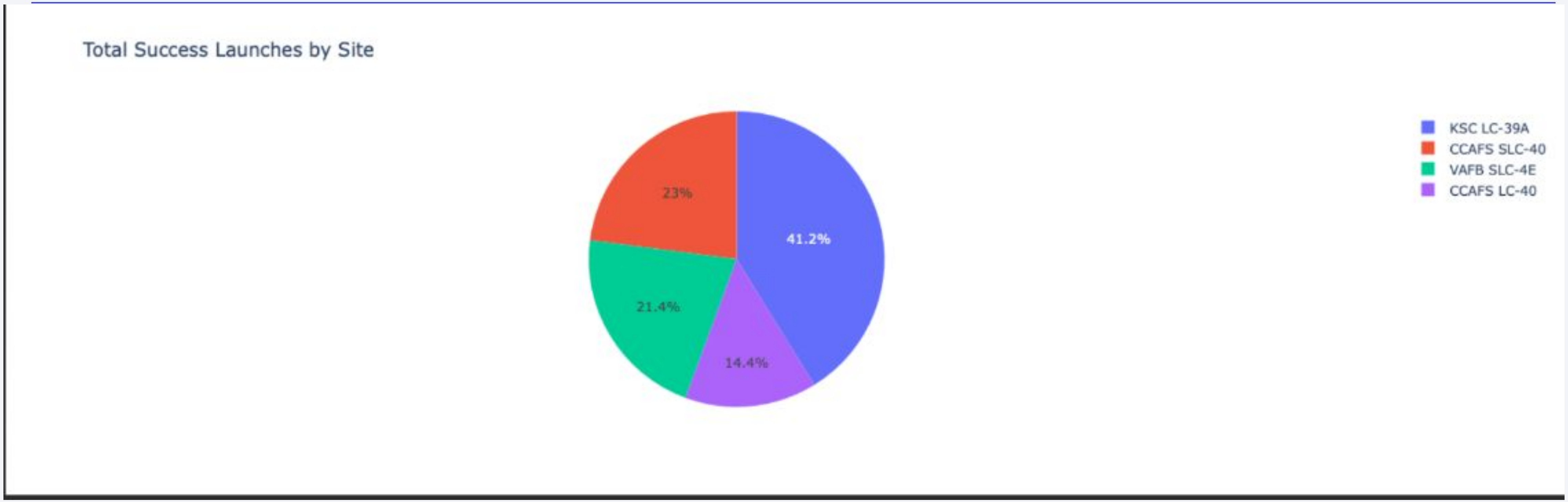
Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



Section 4

Build a Dashboard with Plotly Dash

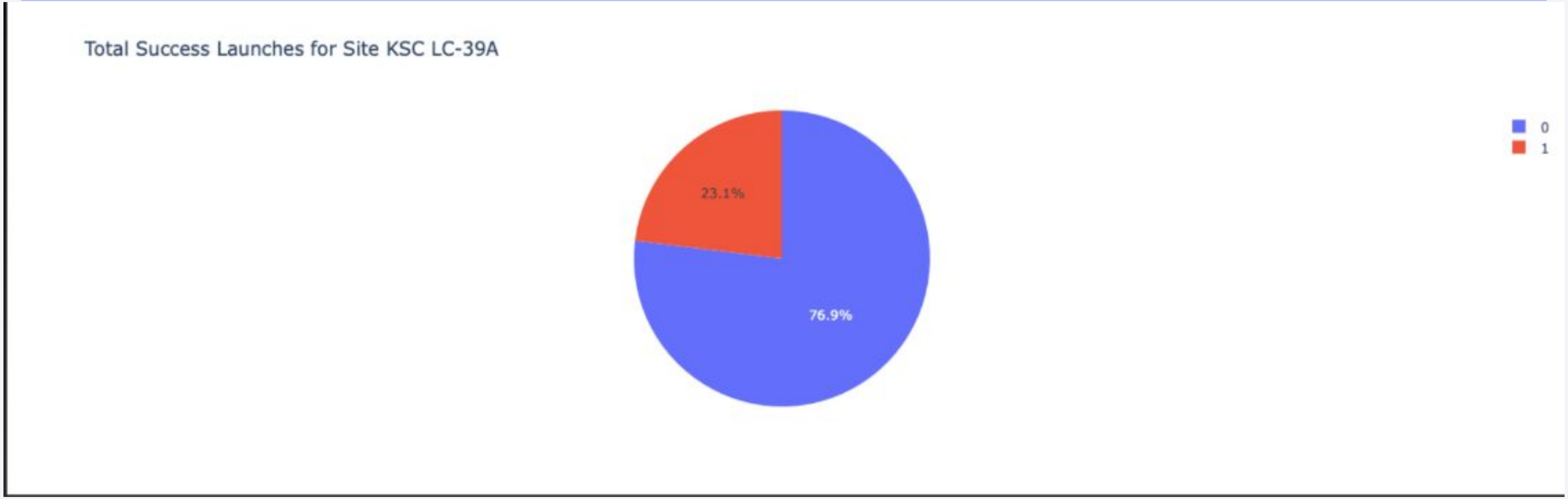
Launch success count for all sites



Explanation:

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

Launch site with highest launch success ratio



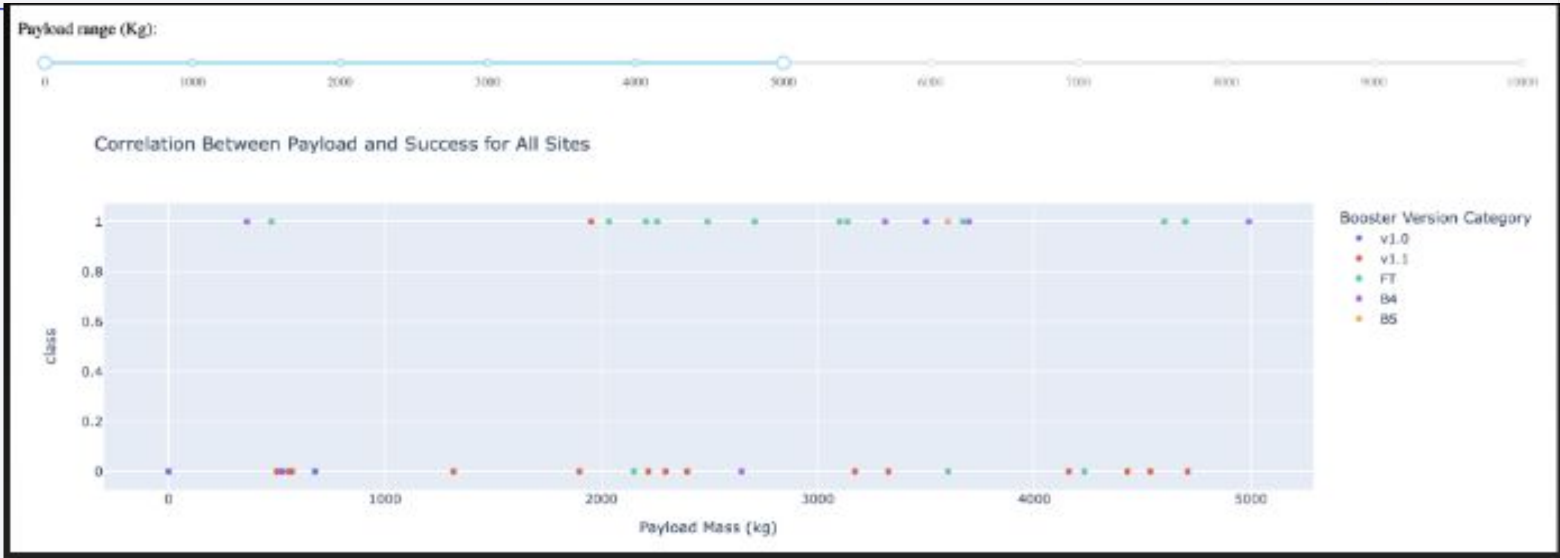
Explanation:

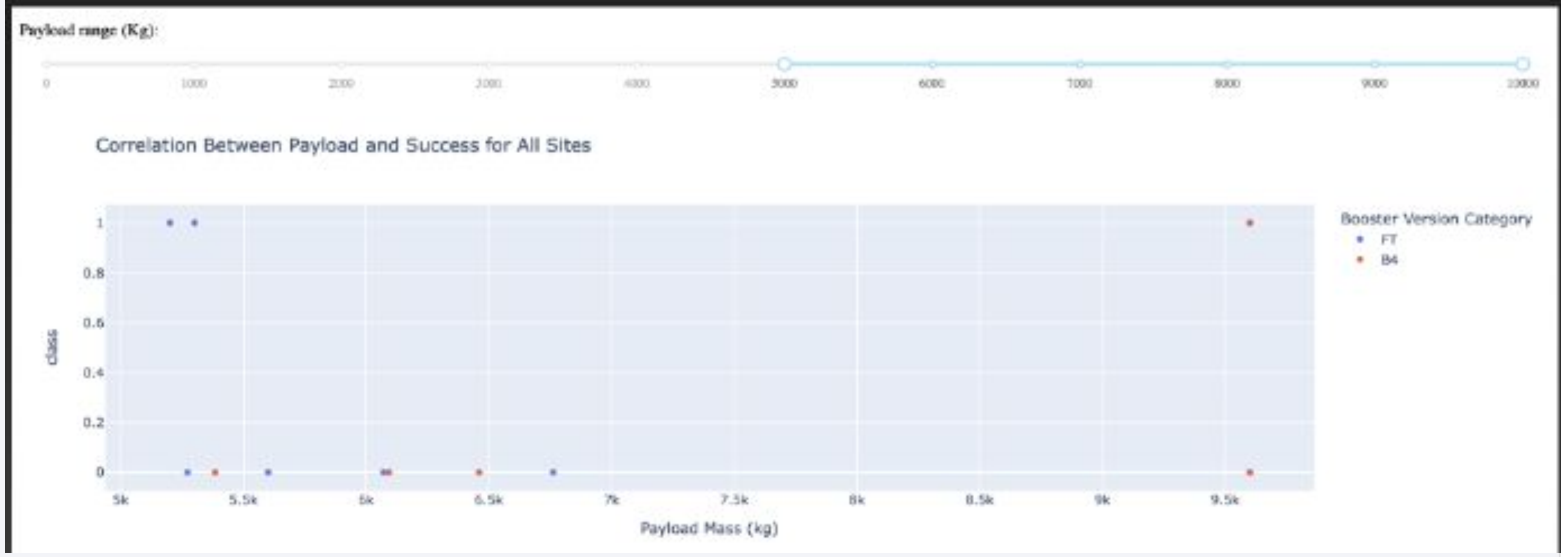
- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload Mass vs. Launch Outcome for all sites

Explanation:

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





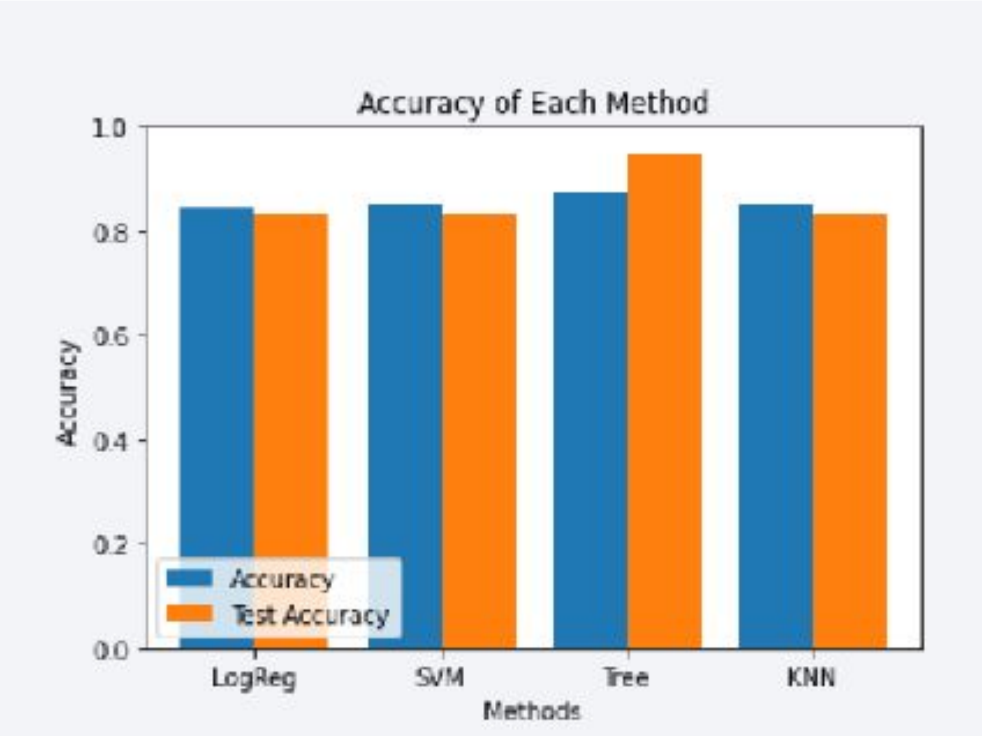
Section 5

Predictive Analysis (Classification)

Classification Accuracy

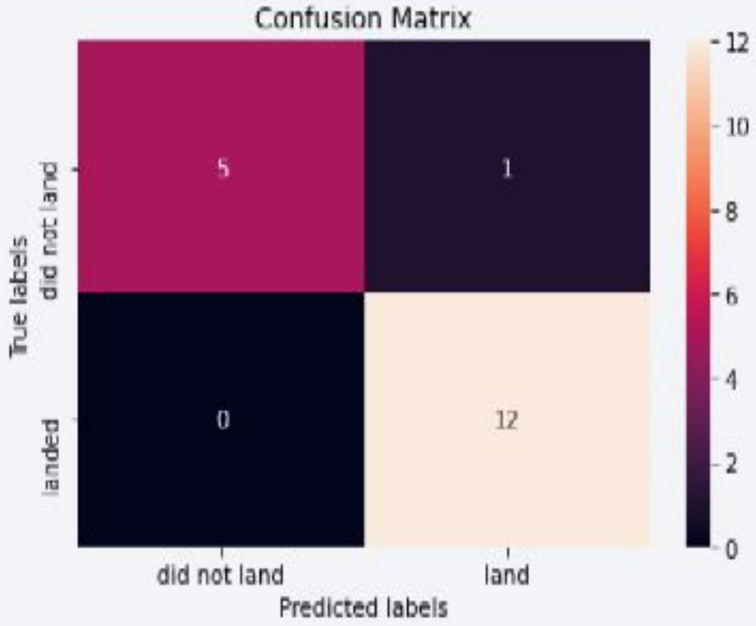
- Four classification models were tested, and their accuracies are plotted beside;

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones



Conclusions

- The best launch site is KSC LC-39A
- Launches above 7,000 kg are less risky;
- Decision Tree Classifier can be used to predict successful landings and increase profits
- Different data sources were analyzed, refining conclusions along the process;
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.

Appendix

As an improvement for model tests, it’s important to set a value to np.random.seed variable;

Thank you!