IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vasyl Kulyk
October 17, 2024

Outline

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

Executive Summary

The search for an answer to the main question involved data collection, data preparation, exploratory data analysis (EDA) and model development. Data was collected by using API queries and web scraping. Data standardization was also performed. EDA enabled the identification of patterns in the data. Logistic regression, SVM, decision tree classifier and KNN were used to predict the success of the launches.

As a result, the importance of several factors to the success of launches has been identified. These include number of flights, orbit type, payload and launch site locations. On test data, all the classification models used in the project show the same accuracy. It is equal to 0.833. This is quite a high value.

Introduction

- We are a young startup that is entering the space launch market. There are several companies competing in this market. Among them are SpaceX, NASA, Blue Origin, Virgin Galactic and others.
- The main problem is the high cost of space launches.
- We want to reduce the cost of space launches (objective) to win the space race (goal).
- One way to do this is to use the first stage of a rocket repeatedly.
- To plan and budget space launches, we need to predict their success.
- Therefore, the main question is: Will the first stage of the rocket land successfully?

Section 1

Methodology

Methodology

Executive Summary

This section describes how the following tasks were carried out:

- Perform data collection
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification model

Data Collection

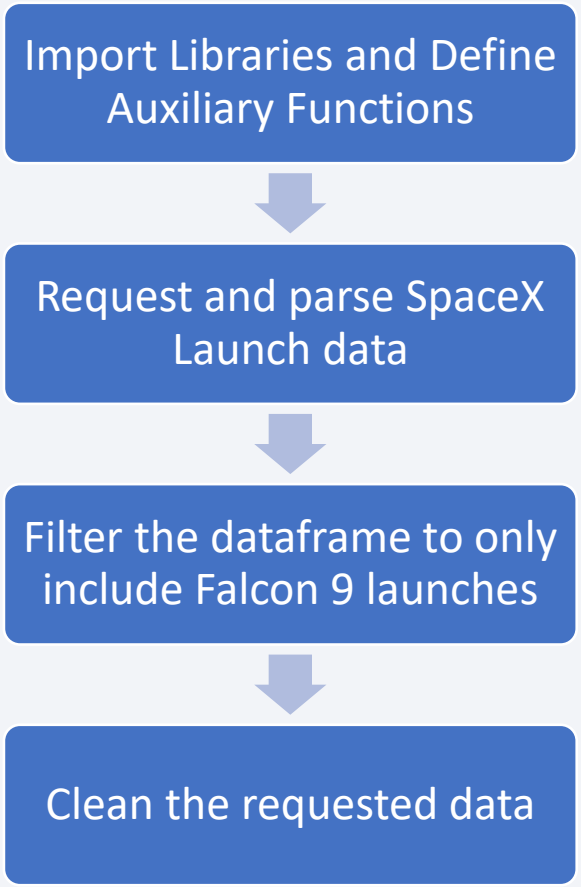
Data were collected from two sources.

The first source is SpaceX API.

The second source is the Wiki page about Falcon 9.

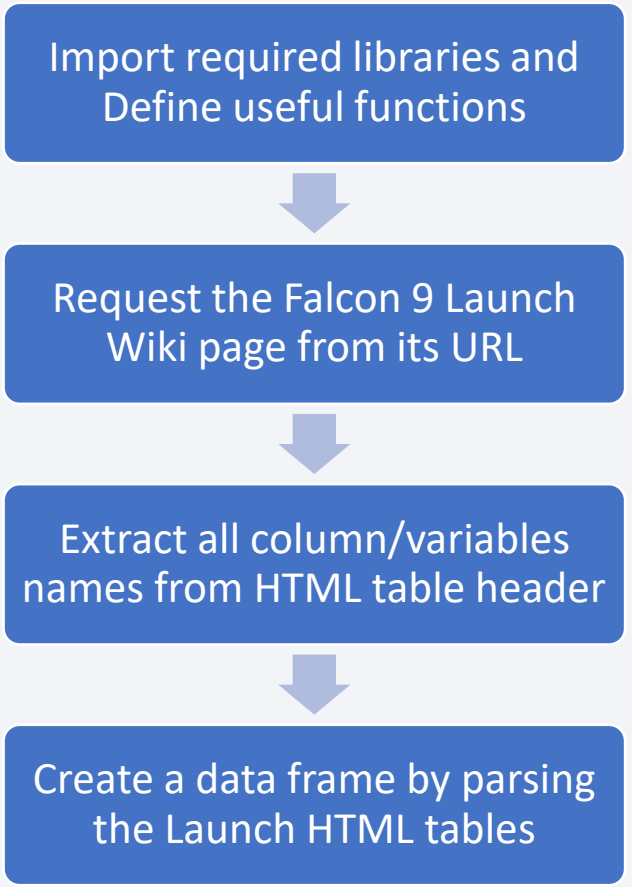
Data Collection – SpaceX API

- We collected the data from SpaceX API
- The data collection process included the following steps
- The results of the data collection process can be viewed in the following notebook: https://github.com/vhkulyk/capstone-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



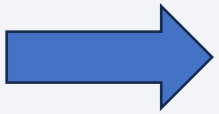
Data Collection - Scraping

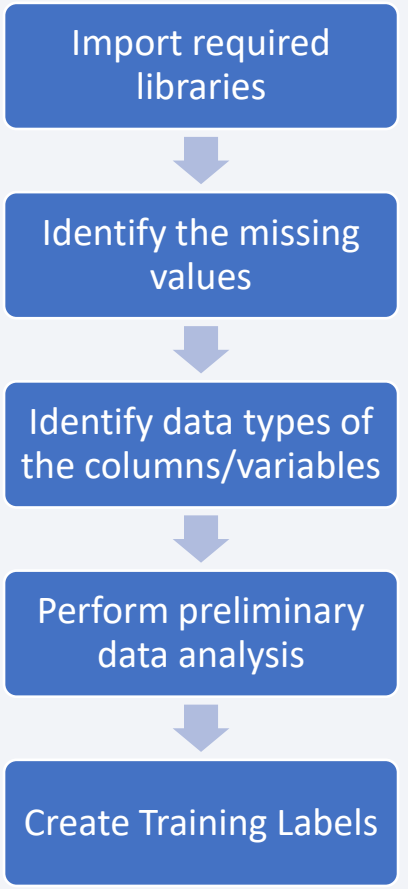
- We collected data about Falcon 9 and Falcon Heavy launches from Wikipedia
- The data collection process included the following steps
- The results of the data collection process can be viewed in the following notebook: https://github.com/vhkulyk/capstone-project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- We performed the necessary data preparation to build the predictive model
- The data wrangling process included the following steps
- The results of the data wrangling process can be viewed in the following notebook: https://github.com/vhkulyk/capstone-project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb





EDA with Data Visualization

- We used three types of graphs.
- The first of them is scatter plot. It was used to visualize the relationships between three variables (all continuous).
- The second one is a bar chart. It was needed to visualize the relationships between two variables (categorical and continuous).
- The third one is a line chart. It was used to visualize the trend in the change of a continuous variable over time.
- The results of EDA with Data Visualization can be viewed in the following notebook: https://github.com/vhkulyk/capstone-project/blob/main/edadataviz.ipynb

EDA with SQL (I)

The SQL queries were used to:

- ✓ Display the names of the unique launch sites in the space mission,
- ✓ Display 5 records where launch sites begin with the string 'CCA',
- ✓ Display the total payload mass carried by boosters launched by NASA (CRS),
- ✓ Display average payload mass carried by booster version F9 v1.1,
- ✓ List the date when the first successful landing outcome in ground pad was achieved,

EDA with SQL (II)

The SQL queries were used to:

- ✓ List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000,
- ✓ List the total number of successful and failure mission outcomes,
- ✓ List the names of the booster versions which have carried the maximum payload mass,
- ✓ 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,
- ✓ 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.

The results of the EDA with SQL can be viewed in the following notebook:

https://github.com/vhkulyk/capstone-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Objects such as circles, markers and lines were created and added to the map.

- Circles were used to mark NASA and launch sites on the map.
- Markers were used to mark launch outcomes on the map.
- Lines were used to mark on the map the distances between the launch site and its proximities.

Interactive Map can be viewed in the following notebook:

https://github.com/vhkulyk/capstone-project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

The pie chart and scatter plot are used in the dashboard.

The first of them shows the results of launches. The second shows the correlation between payload and launch success.

Besides them, dropdown list and slider are used.

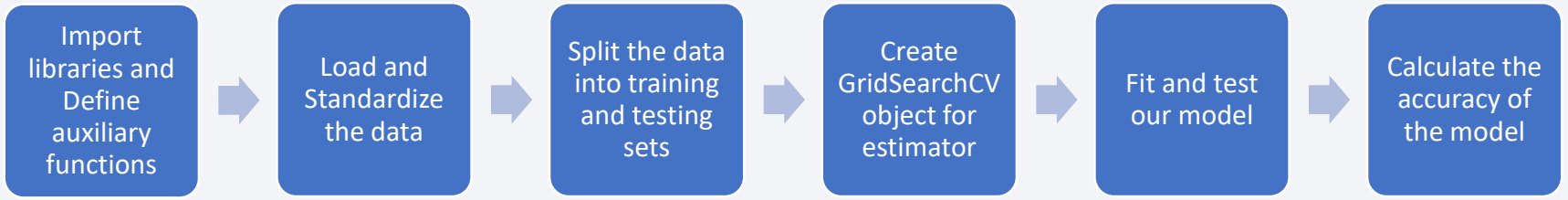
The first one helps to select a launch site, the second one - payload range.

The code to run the dashboard can be viewed at the following link:

https://github.com/vhkulyk/capstone-project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- GridSearchCV from scikit-learn library was used to find the best performing classification model. Logistic regression, SVM, decision tree classifier and KNN were used as estimators.
- The model development process included the following steps:



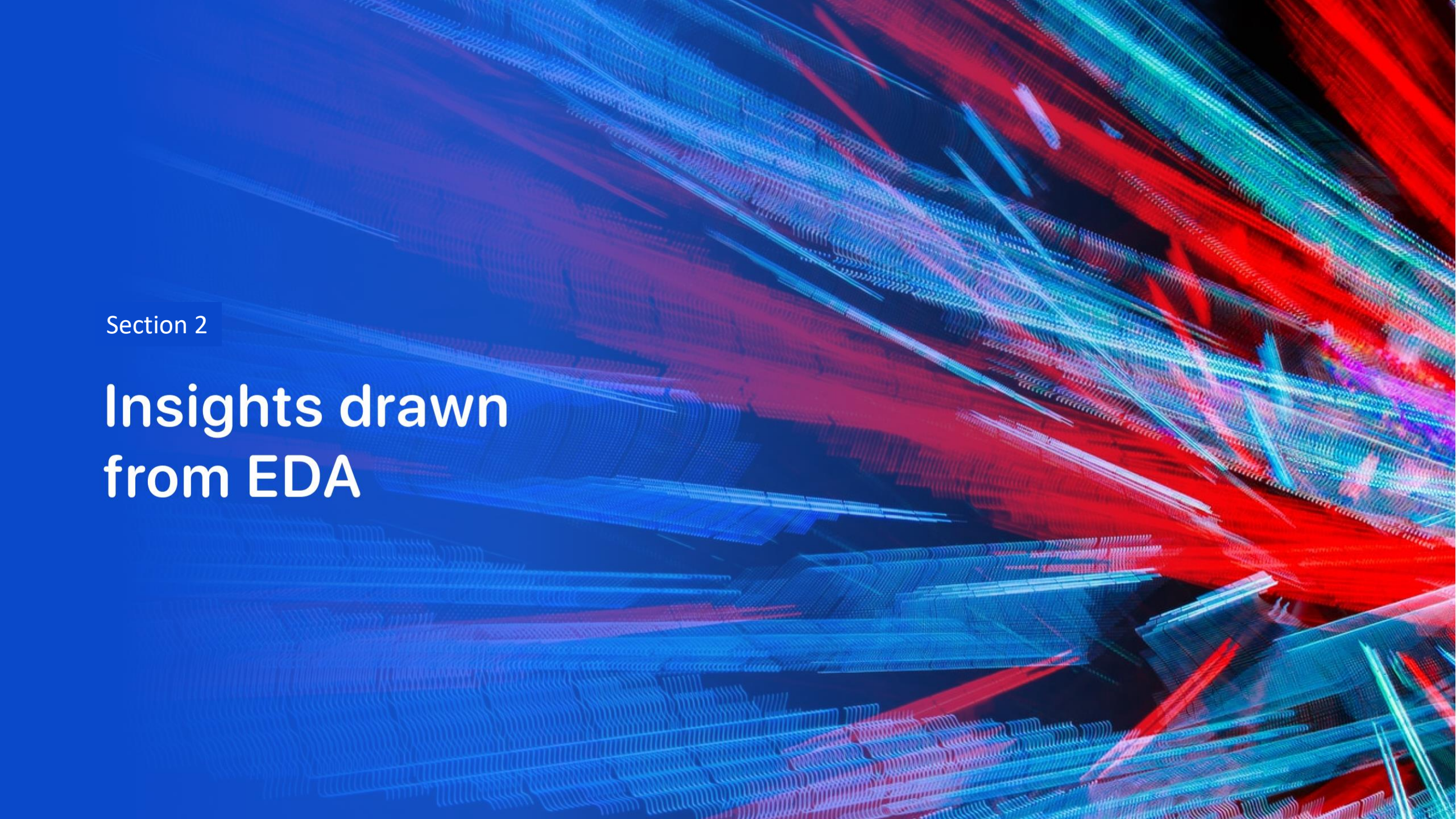
- Results of predictive analysis can be viewed at the following link:

https://github.com/vhkulyk/capstone-project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

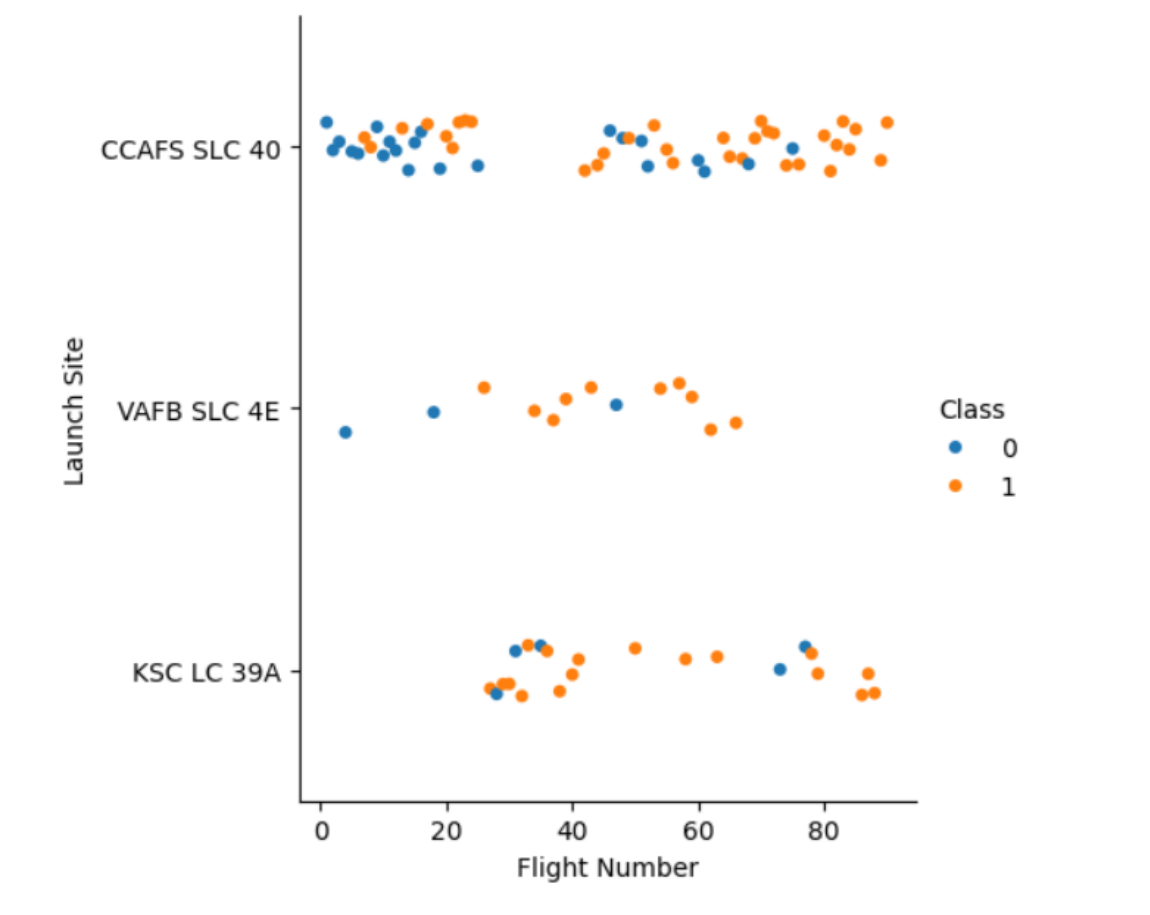
Results

The results are presented in the following sections:

- Insights drawn from EDA
- Launch Sites Proximities Analysis
- Build a Dashboard with Dash Plotly
- Predictive Analysis (Classification)


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

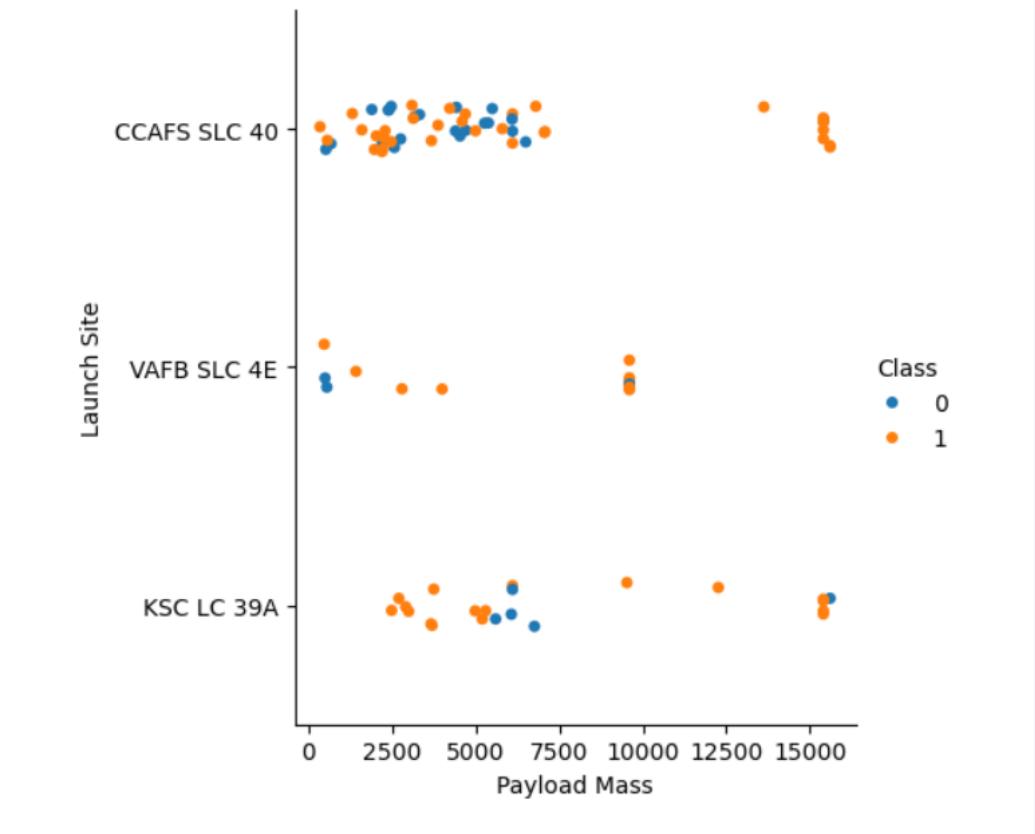


Interpretation:

As the number of flights increases, the success rate increases.

This relationship is seen for all launch sites.

Payload vs. Launch Site

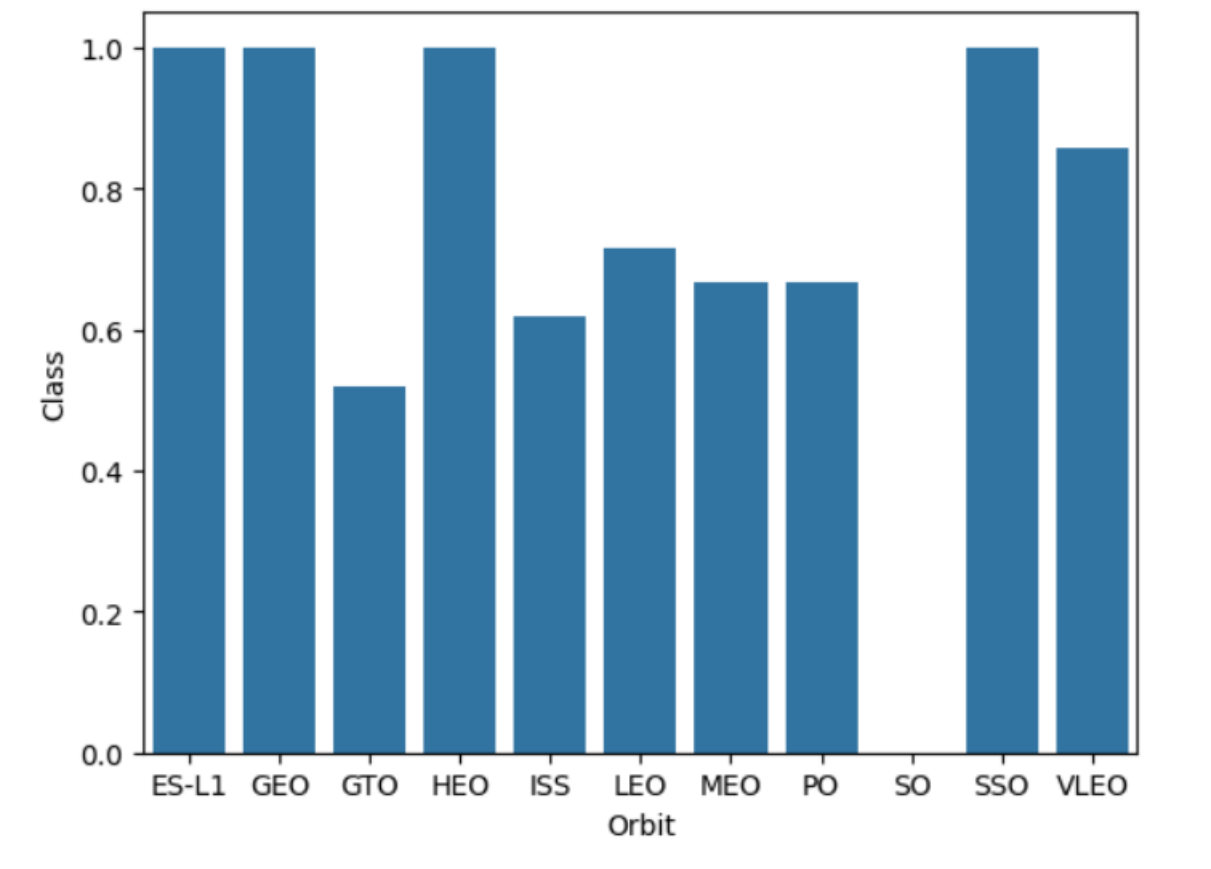


Interpretation:

In general, the value of payload mass has no clearly visible effect on the success rate.

Failures are observed in different payload mass ranges.

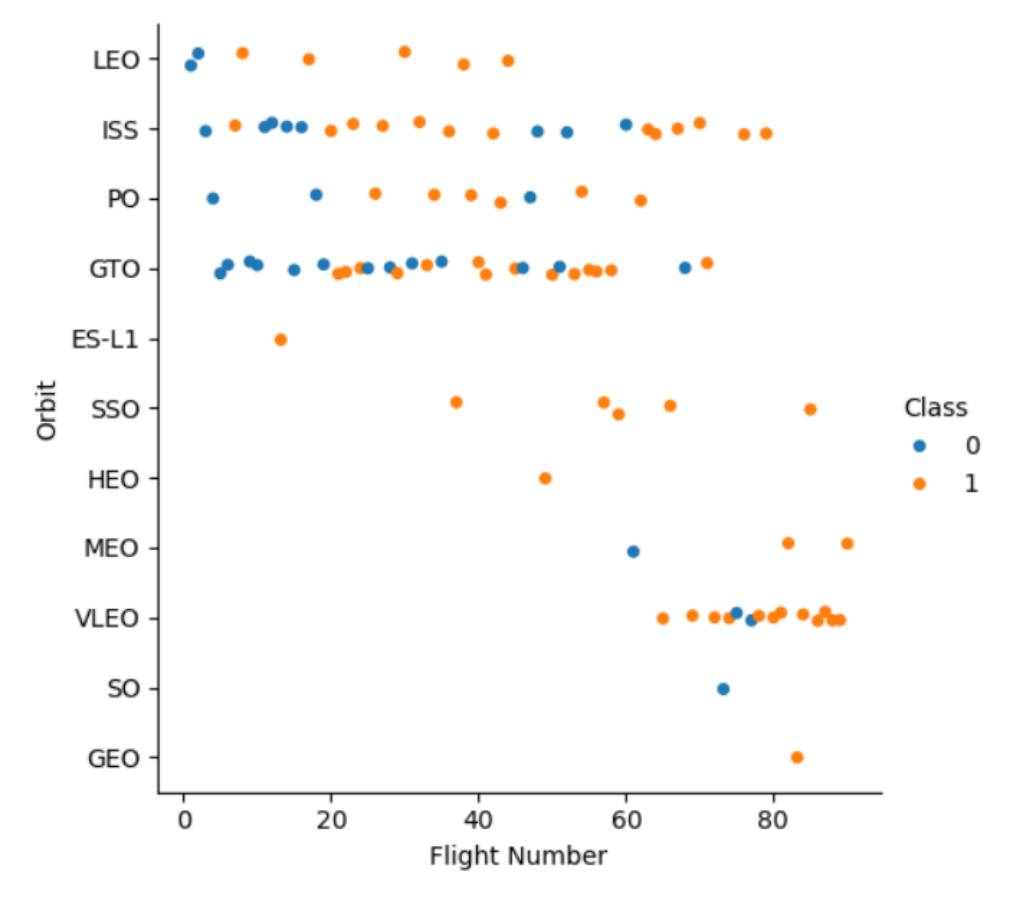
Success Rate vs. Orbit Type



Interpretation:

The highest success rate has orbit types such as ES-L1, GEO, HEO, SSO.

Flight Number vs. Orbit Type

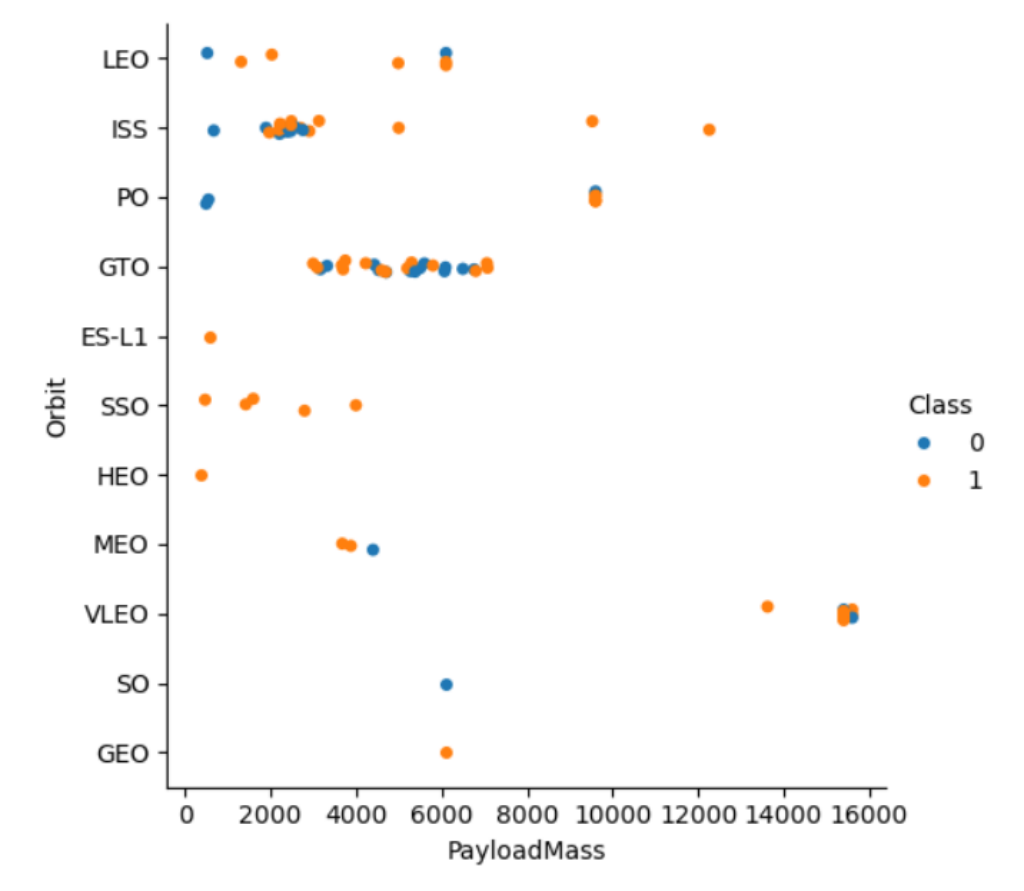


Interpretation:

In the case of LEO, success has a relationship with the number of flights.

No such relationship is observed for other orbit types.

Payload vs. Orbit Type

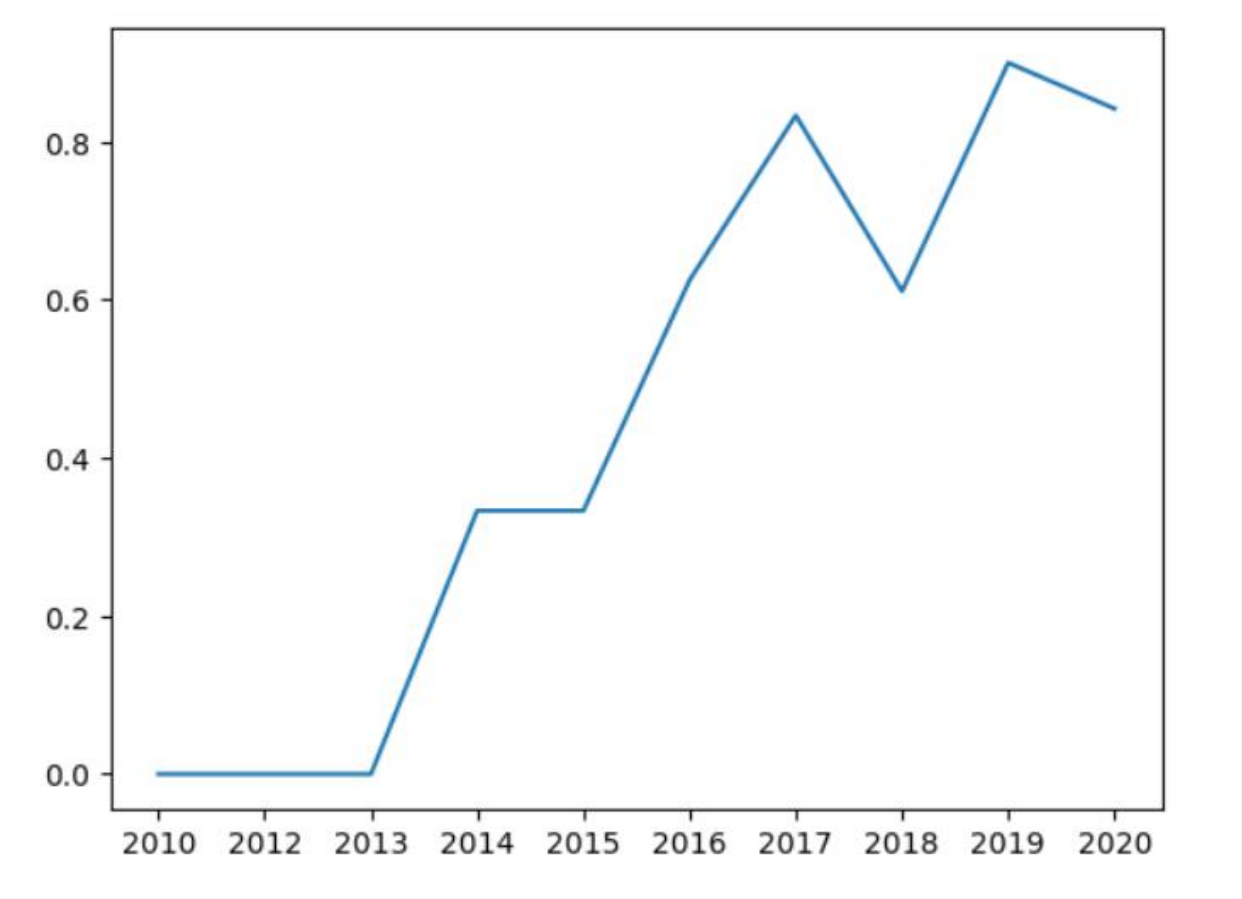


Interpretation:

The relationship between payload mass and success rate is seen in the case of ISS.

For other orbit types no relationship is observed.

Launch Success Yearly Trend



Interpretation:

Success rate increases from 2013 to 2020.

All Launch Site Names

The names of the unique launch sites

Launch sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

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

Total Payload Mass

The total payload carried by boosters from NASA (kg)

Customer	Total
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

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

Booster Version	Average mass (kg)
F9 v1.1	2928.4

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad

Date
2015-12-12

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

Mission Outcome	Count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass

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

The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Date	Month	Landing Outcome	Booster version	Launch Site
2015-01-10	01	Failure (drone ship)	F9 V1.1 B1012	CCAFS LC-40
2015-04-14	04	Failure (drone ship)	F9 V1.1 B1015	CCAFS LC-40

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

Count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

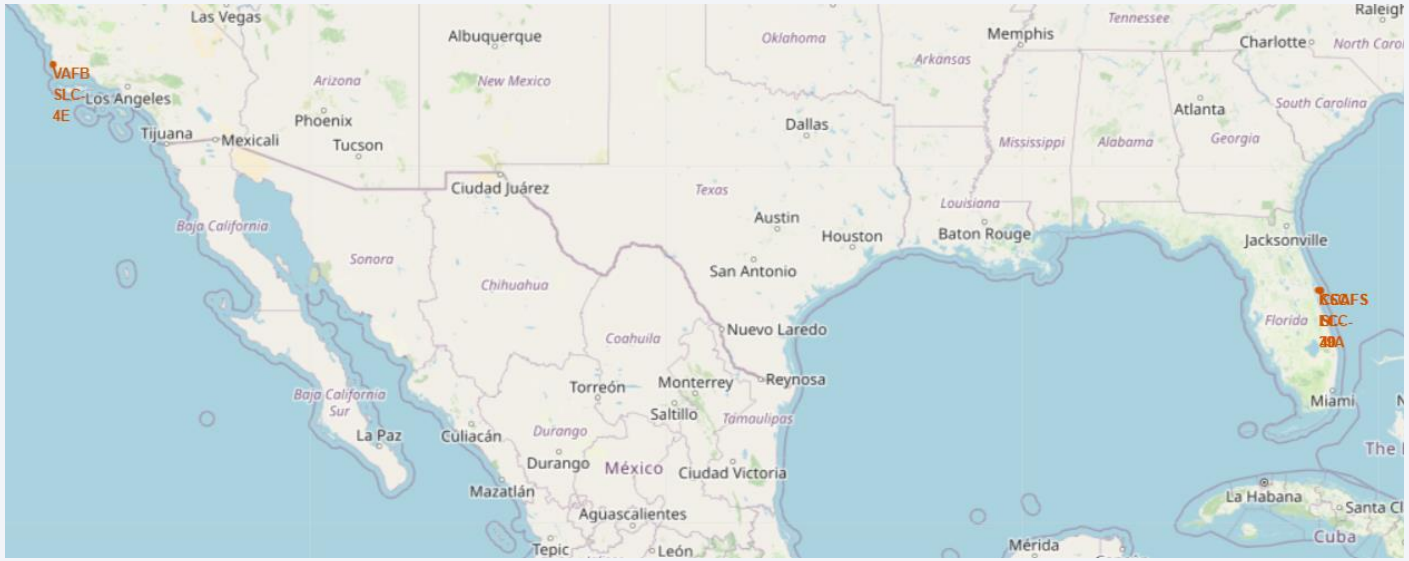
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Sites Locations

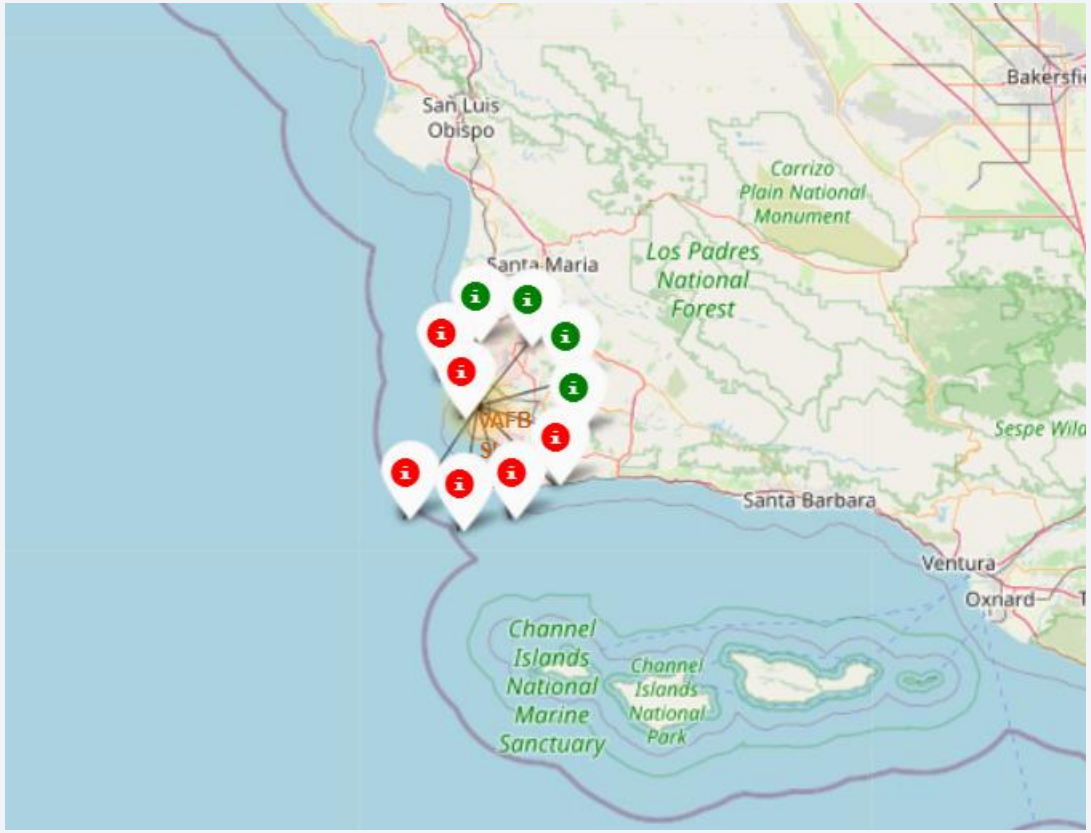
Launch sites' location markers on a global map



Interpretation:
All launch sites are very close to the coast.
VAFB SLC-4E is further from the equator than the other launch sites.

Launch Outcomes

The color-labeled launch outcomes on the map

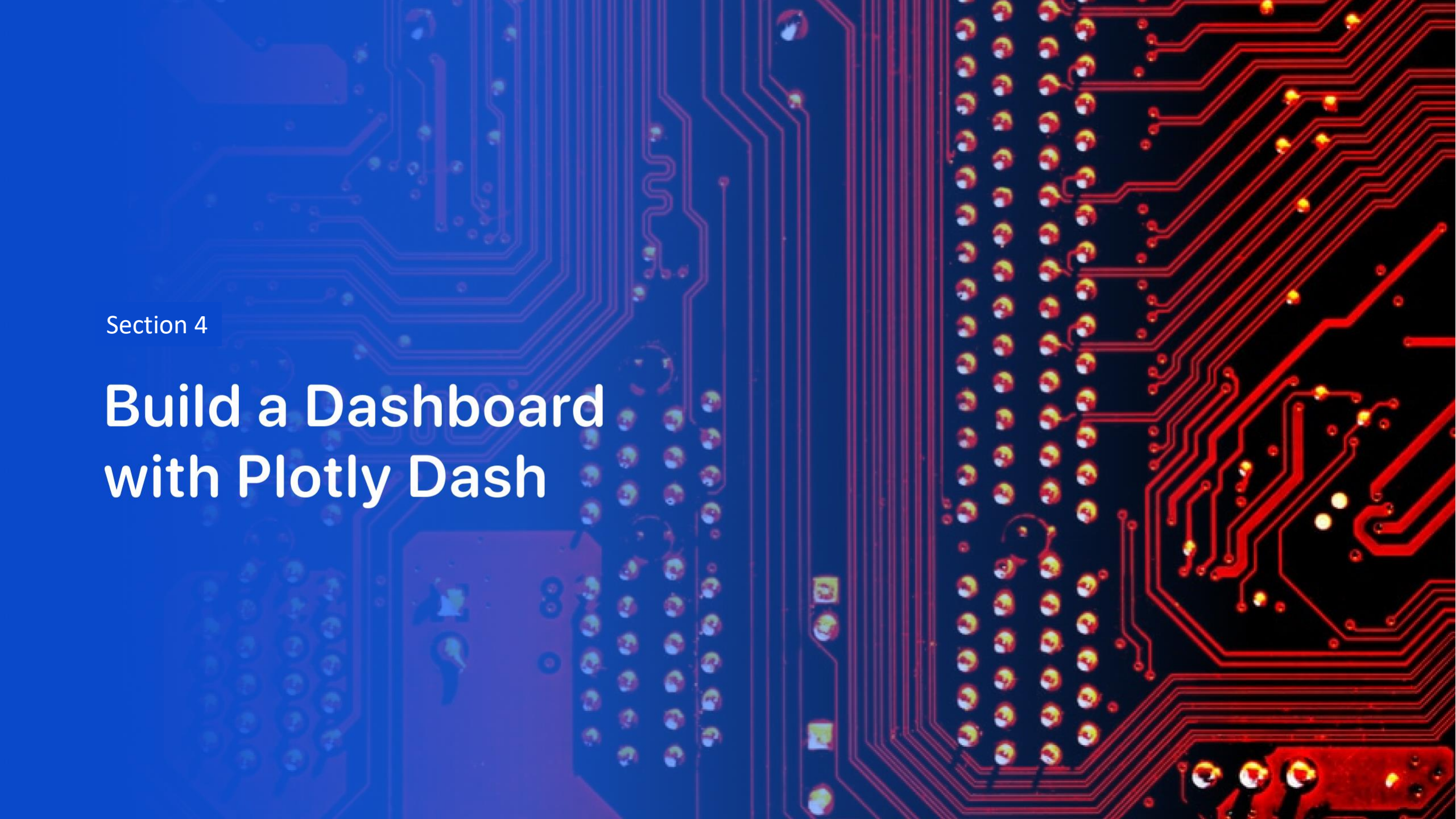


Interpretation:
Red color is used for unsuccessful launches and green for successful launches.
For example, the VAFB SLC-4E has four successful launches and six failed launches.

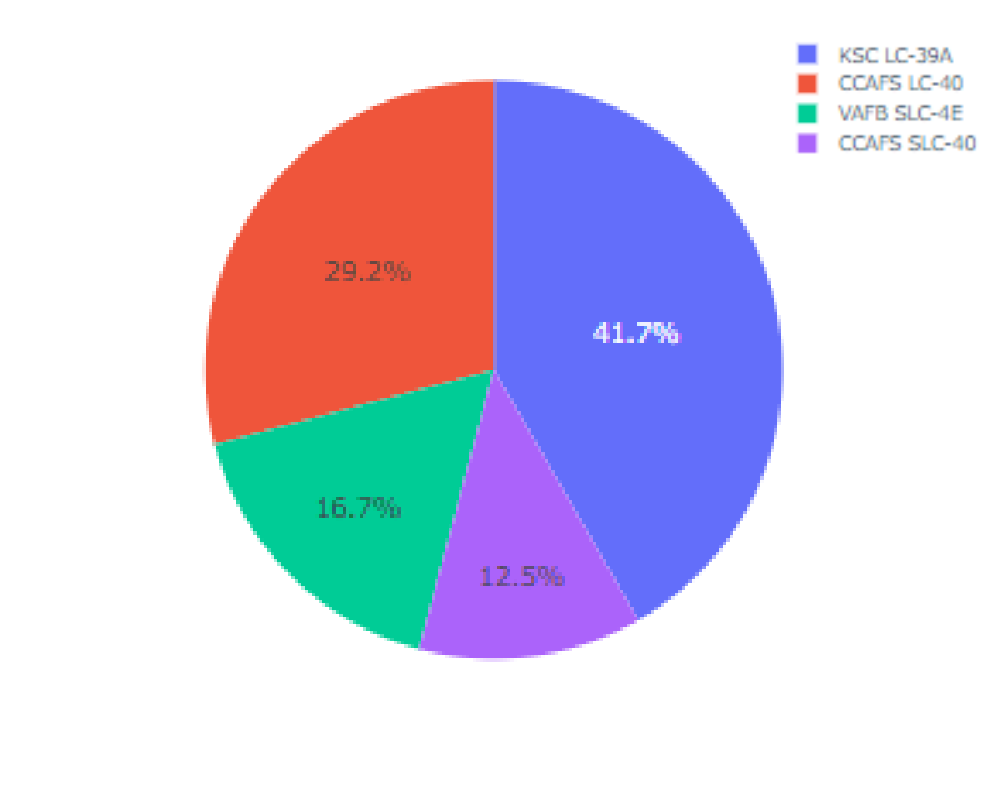
Proximities of Launch Sites

Launch site and its proximities such as railway, coastline,

with distance calculated and displayed

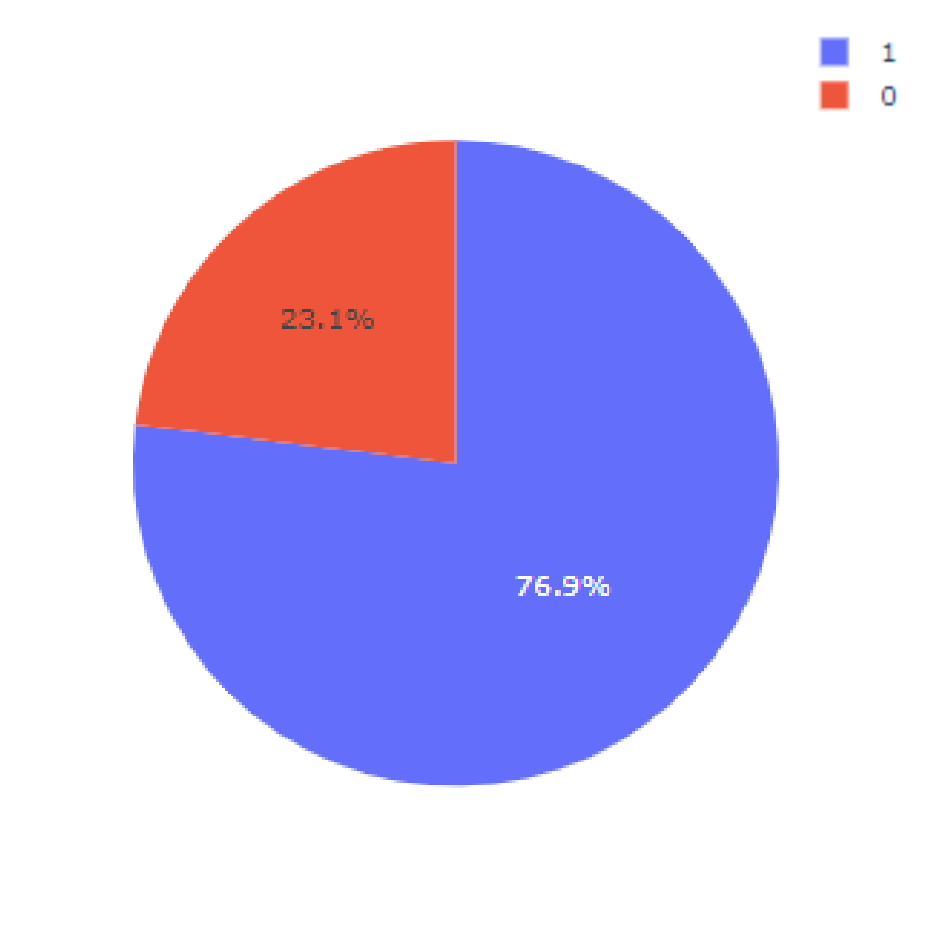

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



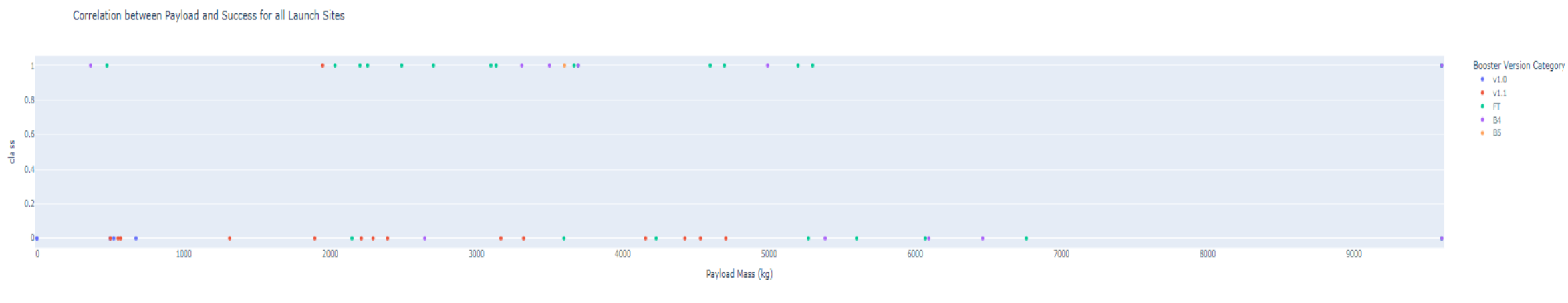
Interpretation:
KSC LC-39A holds a 41.7% share of successful launches.
It's the highest.

Highest Launch Success Ratio



Interpretation:
The KSC LC-39A has the highest successful launch rate.
It equals 76.9%.

Payload vs. Launch Outcome



Interpretation:

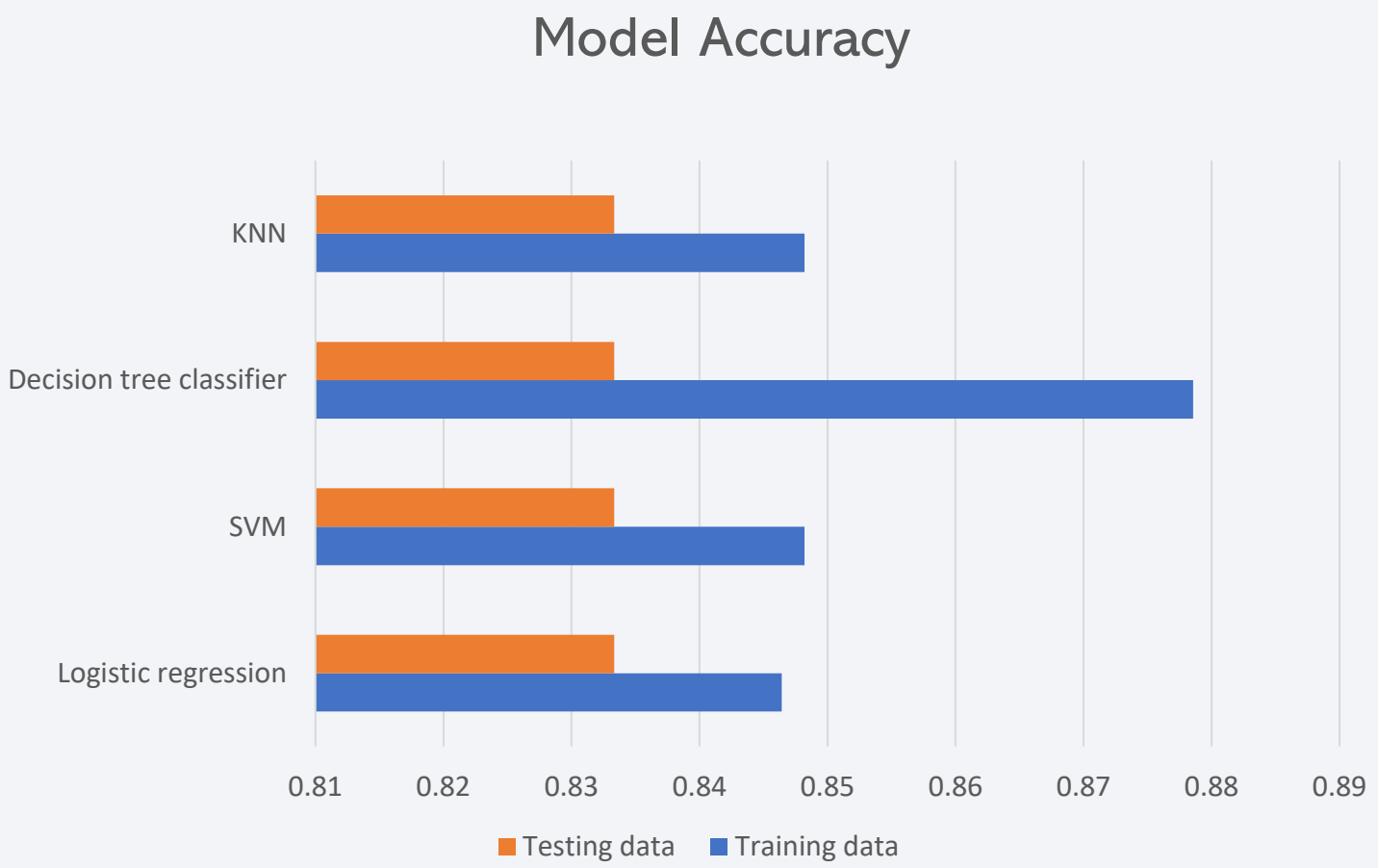
Payload range from 3000 to 4000 kg has the highest launch success rate.

Booster Version “FT” has the highest launch success rate.

Section 5

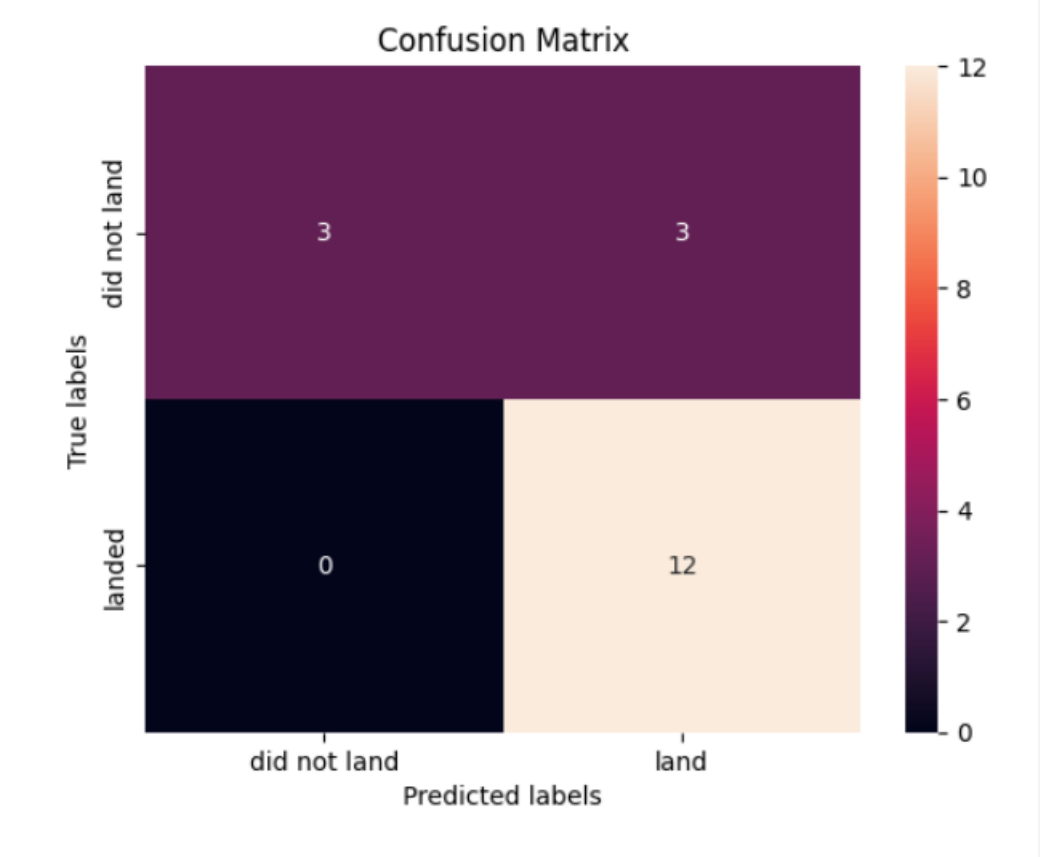
Predictive Analysis (Classification)

Classification Accuracy



Interpretation:
On training data, all models show almost the same accuracy. The difference is mainly measured in thousandths. The bar chart emphasizes this visually. On test data all models show the same accuracy.

Confusion Matrix



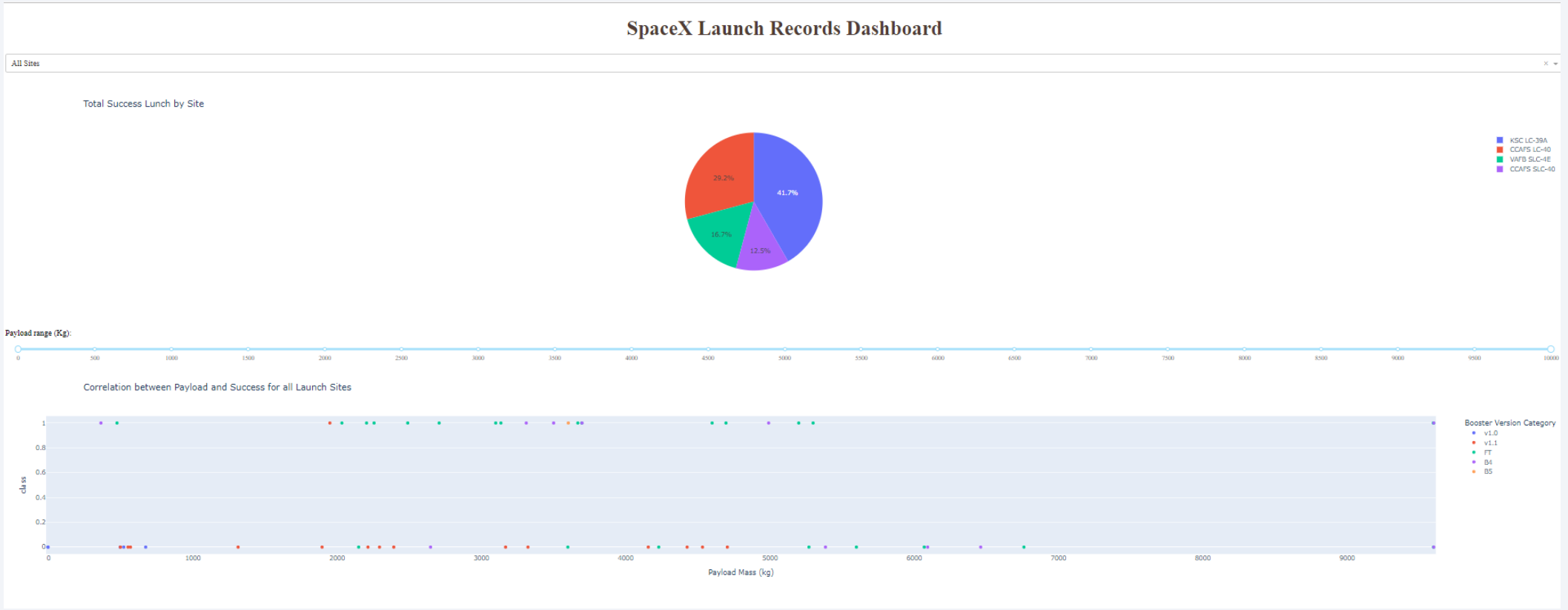
Interpretation:
All classification models have the same distribution of cases in the confusion matrix.
The main problem is False Positives (3 cases). All models predicted that the 3 first stages are landed but they are not.

Conclusions

- The number of flights, type of orbit, launch site locations matter for the success of launches.
- If the number of flights increases, the success rate increases.
- Orbit types such as ES-L1, GEO, HEO, SSO have the highest success rate.
- KSC LC-39A is the best launch site.
- Using payload from 3000 to 4000 kg and Booster Version ‘FT’ has the highest success rate.
- Classification models show high prediction accuracy. It is equal to 0.833. However, all models have a problem, namely False Positives.

Appendix

1) Dashboard



Thank you!