IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jose Pablo Mora Villalobos
23/4/2023

Outline

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

Executive Summary

- This project was developed to analyze data from SpaceX rocket landings and determine whether a future landing would be successful or not based on past record.
- To do this, data was collected from Wikipedia. Then it was cleaned and stored on a database. Preliminary analysis was made to determine the nature of the data. Afterwards, different Machine Learning models were train and to be able to classify landings as successful or not. The models perform well, being able to correctly predict most of the validation cases.

Introduction

- SpaceX is a company that is trying to revolutionize space traveling. They have being testing rocket landings since 2010. The data from this tests is available online through the SpaceX API and Wikipedia.
- The objective of this project is to determine if future landings are going to be successful.
- The project focused on the landing of the SpaceX Falcon 9 model series.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Some data was collected using "GET request" to the SpaceX API. Additionally, some extra data was collected using web scrapping techniques on Wikipedia.
- Perform data wrangling
  - Exploratory data analysis was conducted to check the nature of the data.
  - Null values were filled with the media of the corresponding data or deleted.
  - The number of categories of some categorical data was calculated.
  - The target variable (class), that determines the success of the landing, was added.
- 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
  - Four ML models were created to train on the data: Logistic regression, Support vector machine, Decision tree classifier and k-nn classifier.
  - Data was separated in test and train partitions.
  - Models were trained using cross validation to determine the best performing parameters.
  - To evaluate the accuracy of the models, R squared score and confusion matrices were used.

Data Collection

- SpaceX API
  - Data was collected using "Get request" to the SpaceX API.
  - Additionally, extra data from rocket, payloads, launchpad, and cores were collected.
  - A data set was created using only data for the Falcon 9 Booster Version.
- Web Scrapping
  - Data from the Wikipedia page was collected using a "Get request".
  - The Falcon 9 data was collected from a specific table.
  - A dataframe was created using the content of this table.
- Some minimal data cleansing was performed at this stage on both data collection techniques.

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- https://github.com/JosePabloMV/Applied-Data-Science-Capstone/blob/master/Data%20Collection%20API%20Lab.ipynb

Task 1: Request and parse the SpaceX launch data using the GET request

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsN
```

Task 2: Filter the dataframe to only include `Falcon 9` launches

Finally we will remove the Falcon 1 launches keeping only the Falcon 9 launches. Filter the data dataframe using the `BoosterVersion` column to only keep the Falcon 9 launches. Save the filtered data to a new dataframe called `data_falcon9`.

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = data[data['BoosterVersion']!='Falcon 1']
data_falcon9
```

Task 3: Dealing with Missing Values

Calculate below the mean for the `PayloadMass` using the `.mean()`. Then use the mean and the `.replace()` function to replace `np.nan` values in the data with the mean you calculated.

```
# Calculate the mean value of PayloadMass column
mean = data_falcon9['PayloadMass'].mean()
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan, mean,inplace = True)
data_falcon9
```

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- https://github.com/JosePabloMV/Applied-Data-Science-Capstone/blob/master/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

TASK 1: Request the Falcon9 Launch Wiki page from its URL

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
```

TASK 2: Extract all column/variable names from the HTML table header

Next, we want to collect all relevant column names from the HTML table header

Let's try to find all tables on the wiki page first. If you need to refresh your memory about `BeautifulSoup`, please check the external reference link towards the end of this lab

```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables = soup.find_all("table")
```

TASK 3: Create a data frame by parsing the launch HTML tables

We will create an empty dictionary with keys from the extracted column names in the previous task. Later, this dictionary will be converted into a Pandas dataframe

```
launch_dict= dict.fromkeys(column_names)
```

Data Wrangling

- Data wrangling was performed to find patterns on the data set and create the target variable.
- First, the number of different launching sites, orbits and outcomes were calculated.
- A new variable 'class' was created based on the outcome. This variable had the value of 1 if the landing was successful and 0 if not.
- https://github.com/JosePabloMV/Applied-Data-Science-Capstone/blob/master/EDA%20Lab.ipynb

EDA with Data Visualization

- Charts were used to analyzed the behavior of the data and the relationship with in the features.
- A scatterplot was used to see the relationship Flight Number and Launch Site, Payload and Launch Site and between Flight Number and Orbit type.
- A bar plot showed the average success rate of every orbit type and every year.
- https://github.com/JosePabloMV/Applied-Data-Science-Capstone/blob/master/EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

SQL queries were used to get additional insights on the data. For example:

- Display the names of the unique launch sites
- Display records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- List the date when the first successful landing outcome in ground pad was achieved.
- List the failed landing_outcomes in drone ship.
- https://github.com/JosePabloMV/Applied-Data-Science-Capstone/blob/master/EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

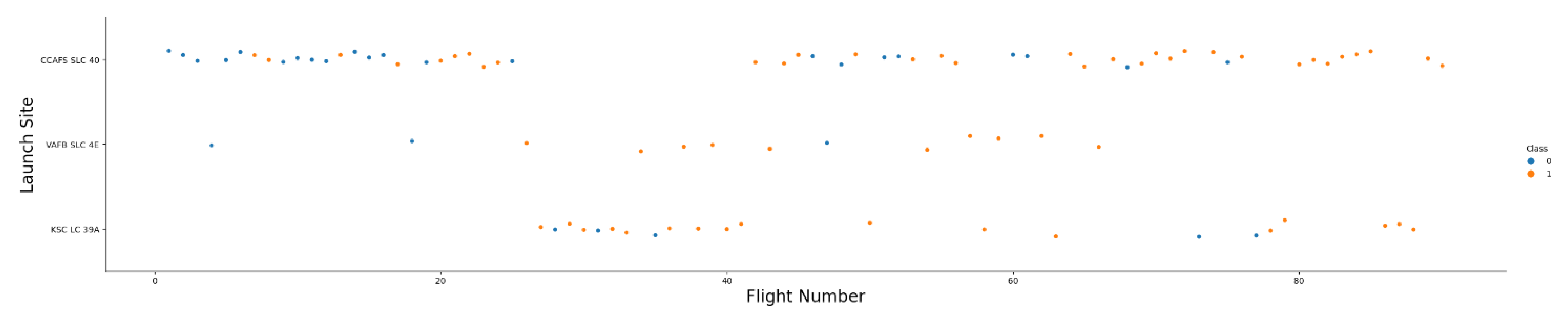
Results

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

Section 2

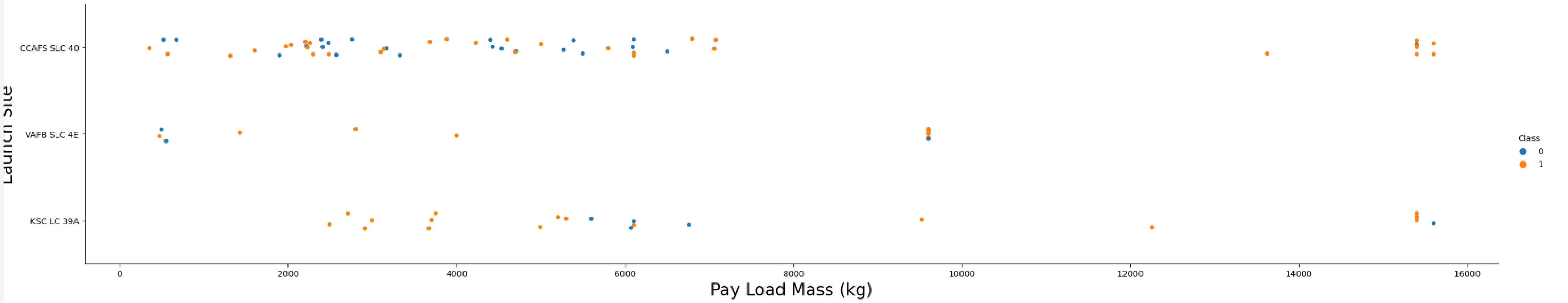
Insights drawn from EDA

Flight Number vs. Launch Site



- Most launches were made on CCAFS SLC 40
- Most of the failed lauches happend at the beggining.

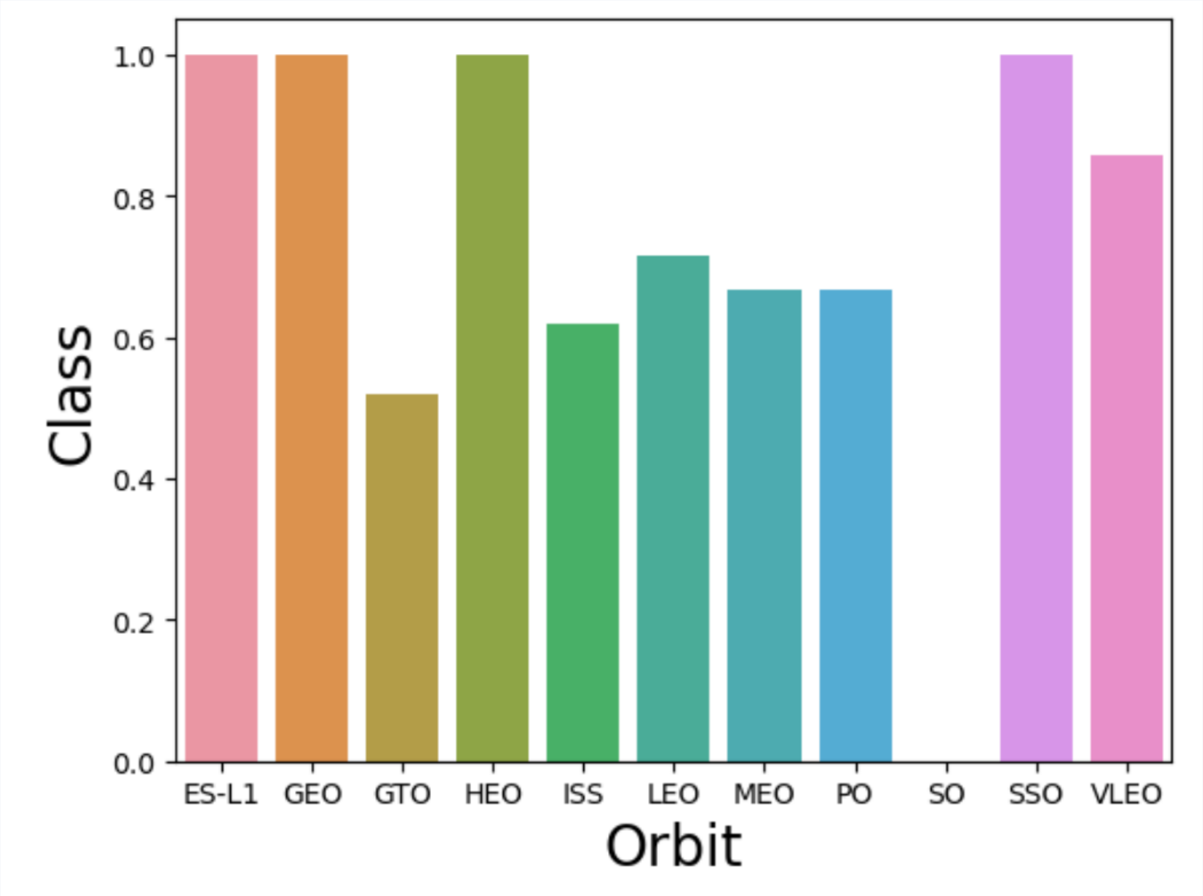
Payload vs. Launch Site



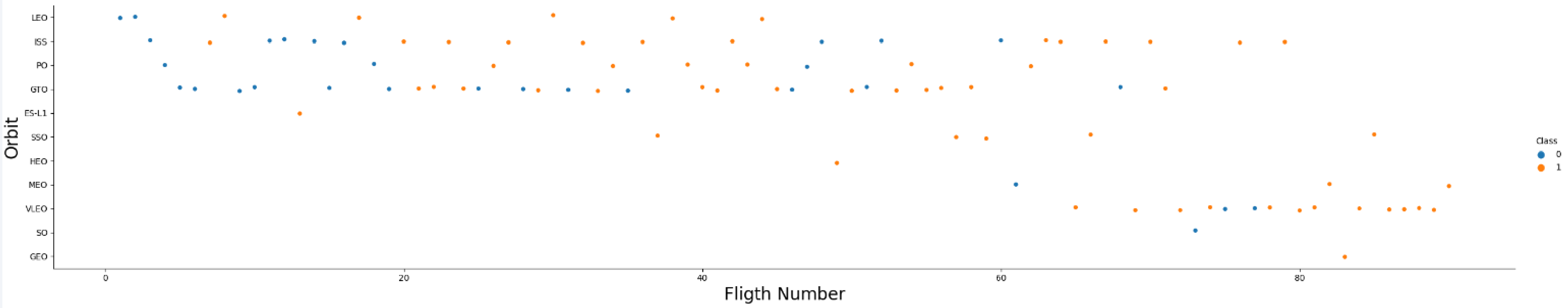
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- As we can see, several orbit types have a 100% success rate, 4 in total.
- Also, most orbit types have a success rate over 50%, except SO which has 0.

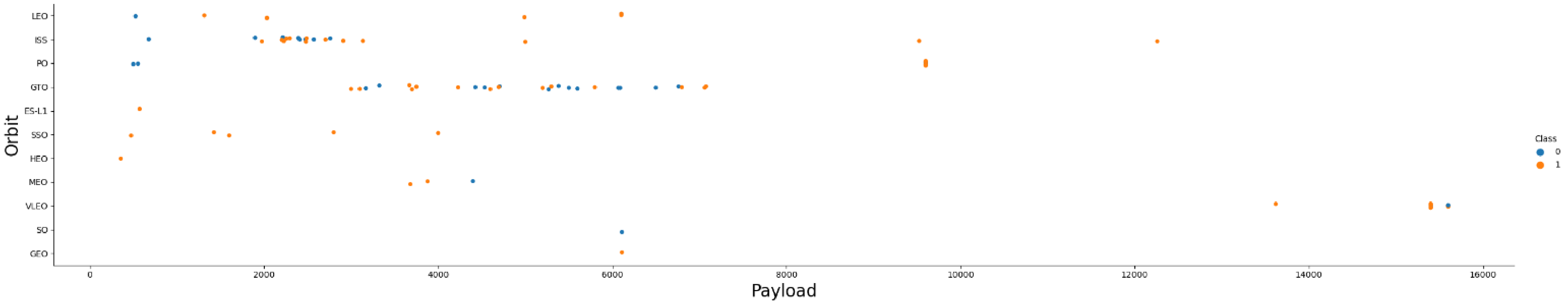


Flight Number vs. Orbit Type



You should see that in 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

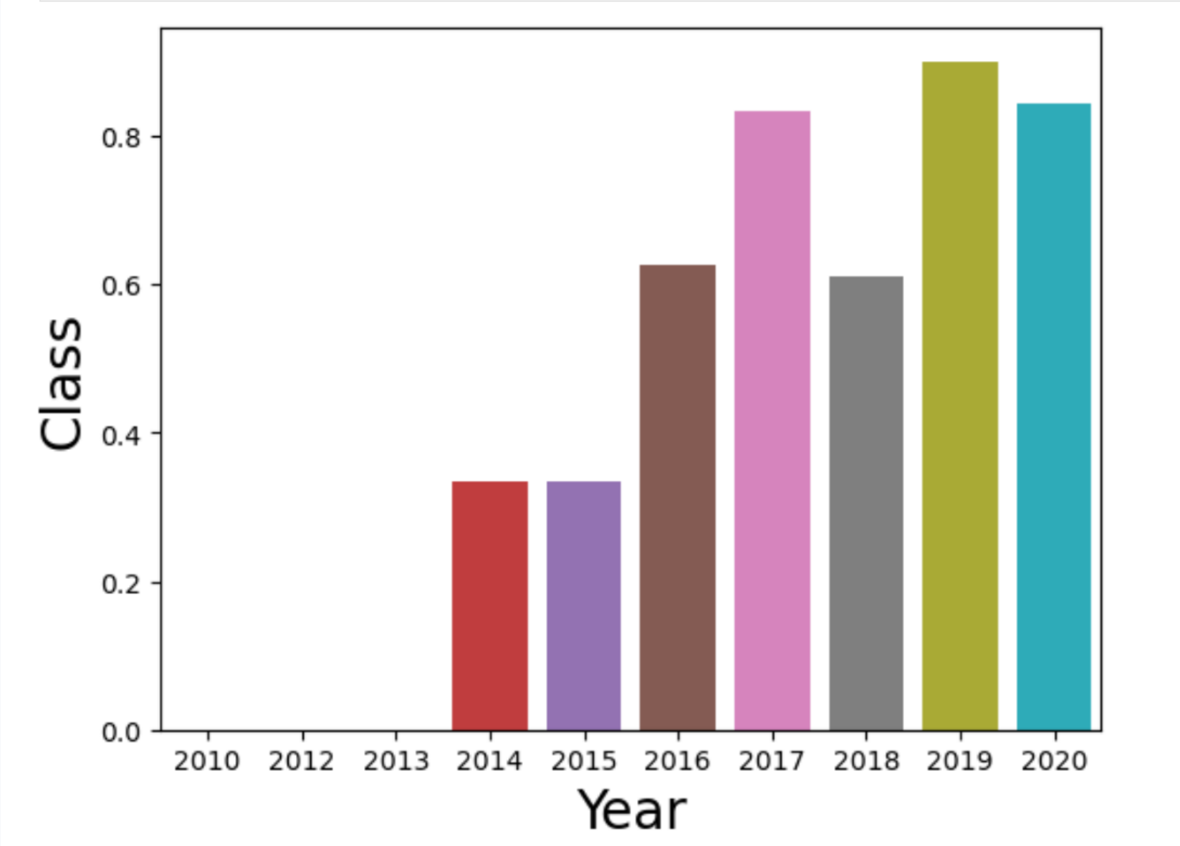


With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Success rates are increasing as years pass.



All Launch Site Names

- There are a total of 4 launch sites.

```
%%sql
SELECT UNIQUE(launch_site) FROM SPACEX
```

```
 * ibm_db_sa://yts96916:***@b1bc1829-6f45-4c
DB
Done.
```

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

Launch Site Names Begin with 'CCA'

- Ther first 5 records for the launch site starting with 'CAA' were not successful or not attempted at all.

```
%%sql
SELECT * FROM SPACEX
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

```
* ibm_db_sa://yts96916:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/BLUDB
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	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

Total Payload Mass

- Nasa launched a total of 45596 kg across all launches.

```
%%sql
SELECT SUM(payload_mass__kg_) AS "TOTAL PAYLOAD MASS" FROM SPACEX
WHERE CUSTOMER = 'NASA (CRS)'
```

```
* ibm_db_sa://yts96916:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tg
one.
```

TOTAL PAYLOAD MASS
45596

Average Payload Mass by F9 v1.1

- Booster version F9 v1.1 moved on average 2928 kg per launch.

```
%%sql
SELECT AVG(payload_mass__kg_) AS "AVG PAYLOAD MASS" FROM SPACEX
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://yts96916:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1o
Done.
```

AVG PAYLOAD MASS
2928

First Successful Ground Landing Date

- The first launch happened in 2010

```
%%sql
SELECT MIN(DATE) FROM SPACEX
```

```
* ibm_db_sa://yts96916:***@b1bc
Done.
```

1
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- 2 Different boosters carried between 4000 and 6000 kg

```
%%sql
SELECT booster_version FROM SPACEX
WHERE landing__outcome = 'Success (drone ship)'
AND payload_mass__kg_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://yts96916:***@b1bc1829-6f45-4cd4-bef4
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- A total of 41 launches were attempted

```
%%sql
SELECT COUNT(*) FROM SPACEX
WHERE landing__outcome LIKE 'Success'
OR landing__outcome LIKE 'Failure'
```

```
 * ibm_db_sa://yts96916:***@b1bc1829-6f45-4
Done.
```

1
41

Boosters Carried Maximum Payload

- This boosters have carried the maximum amount of payload

```
%%sql
SELECT booster_version FROM SPACEX
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEX)
```

```
 * ibm_db_sa://yts96916:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0t
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

2015 Launch Records

- Only 2 attempts were a failure in 2015

```
%%sql
SELECT landing__outcome, booster_version, launch_site FROM SPACEX
WHERE landing__outcome = 'Failure (drone ship)'
AND YEAR(DATE) = 2015
```

```
 * ibm_db_sa://yts96916:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1og
Done.
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- There was a total of 38 successes

```
%%sql
SELECT landing__outcome, COUNT(*) FROM SPACEX
GROUP BY landing__outcome
ORDER BY COUNT(*) DESC
LIMIT 5
```

```
* ibm_db_sa://yts96916:***@b1bc1829-6f45-4cd4-bef
Done.
```

landing__outcome	2
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5

Section 3

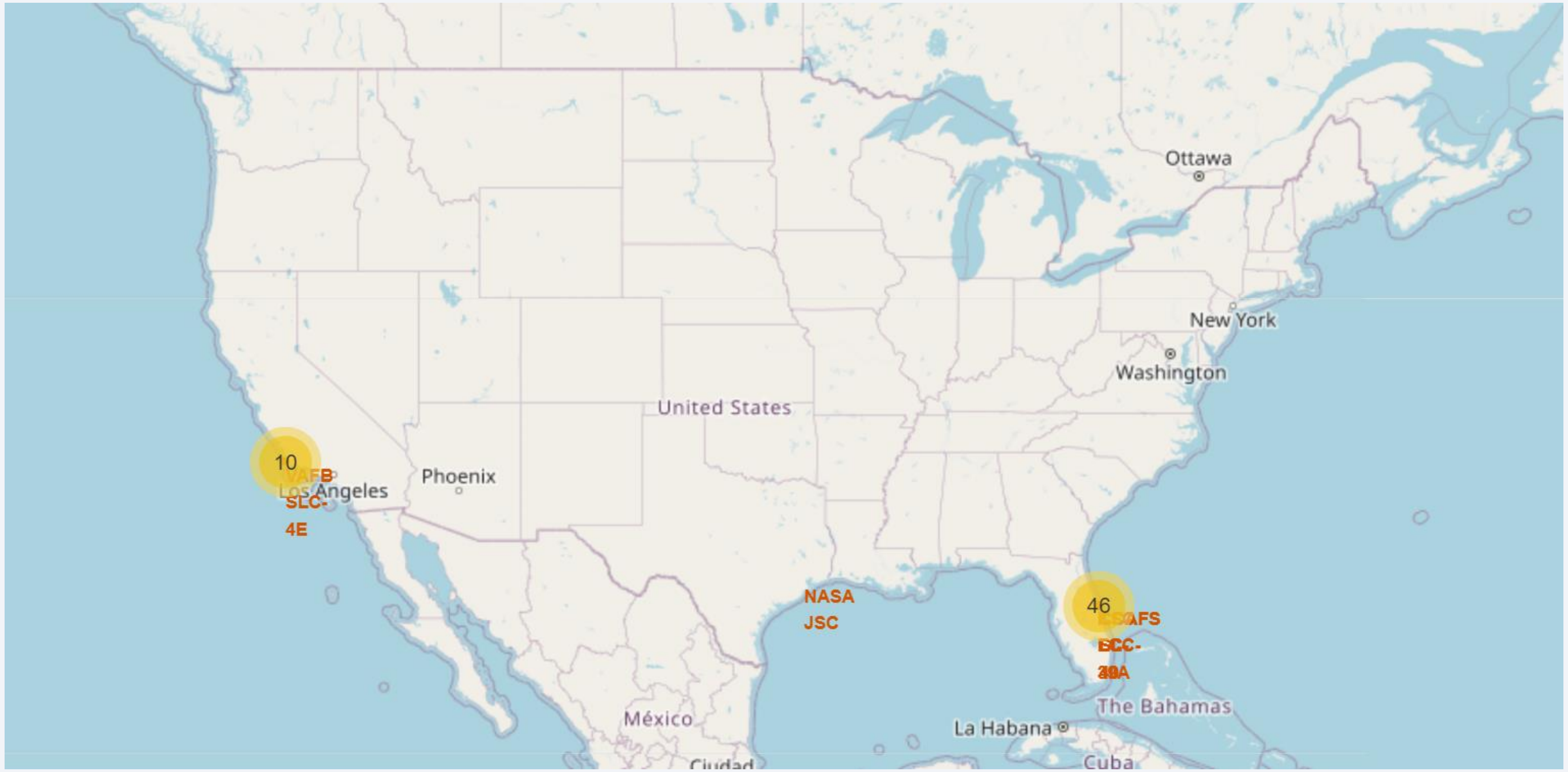
Launch Sites Proximities Analysis

Launch sites locations

- Most of the launch sites are near the coasts of the United States

Number of launches per launch site

- There were a total of 10 launches on the west side and 46 on the east.

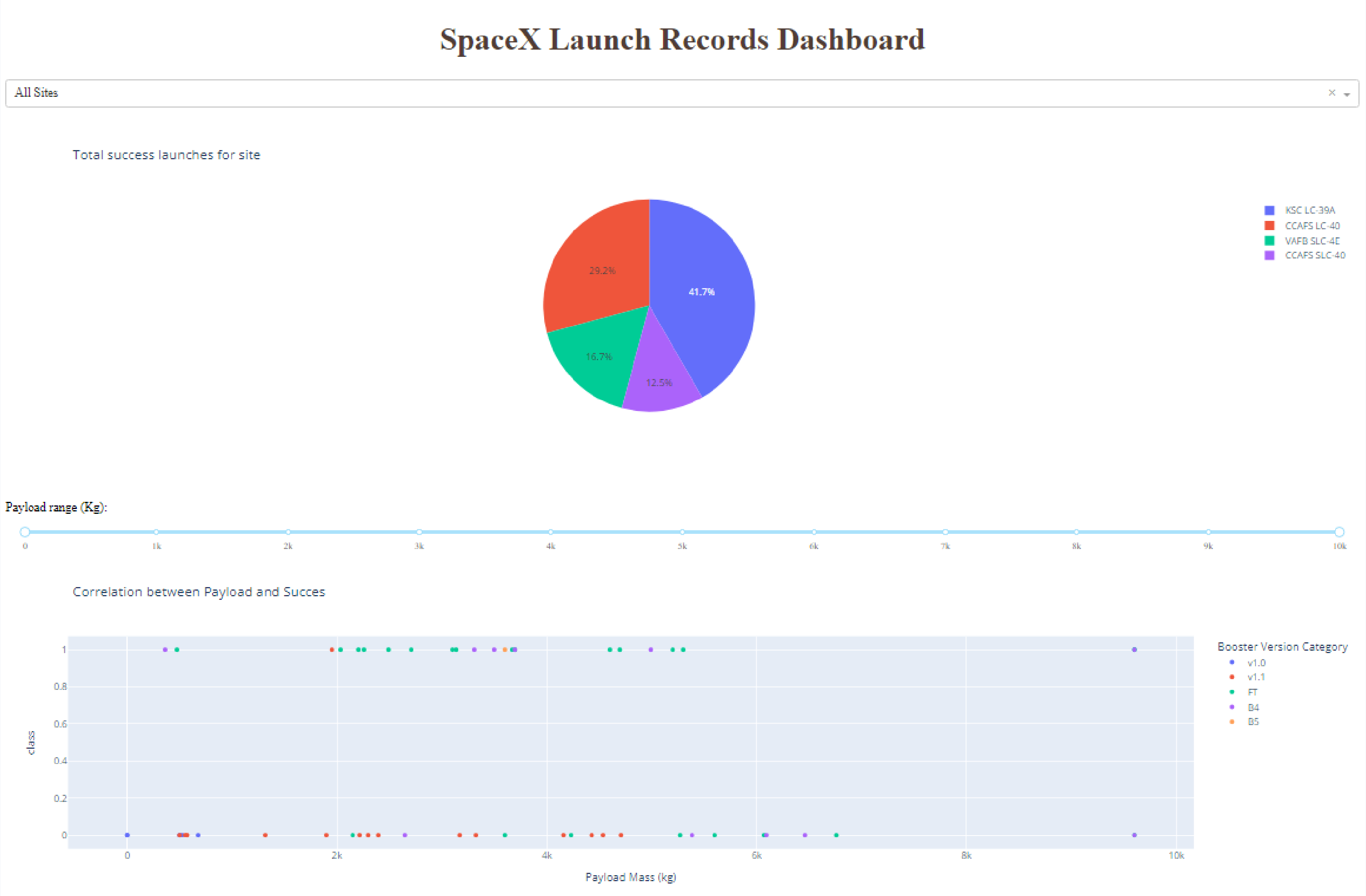


Section 4

Build a Dashboard with Plotly Dash

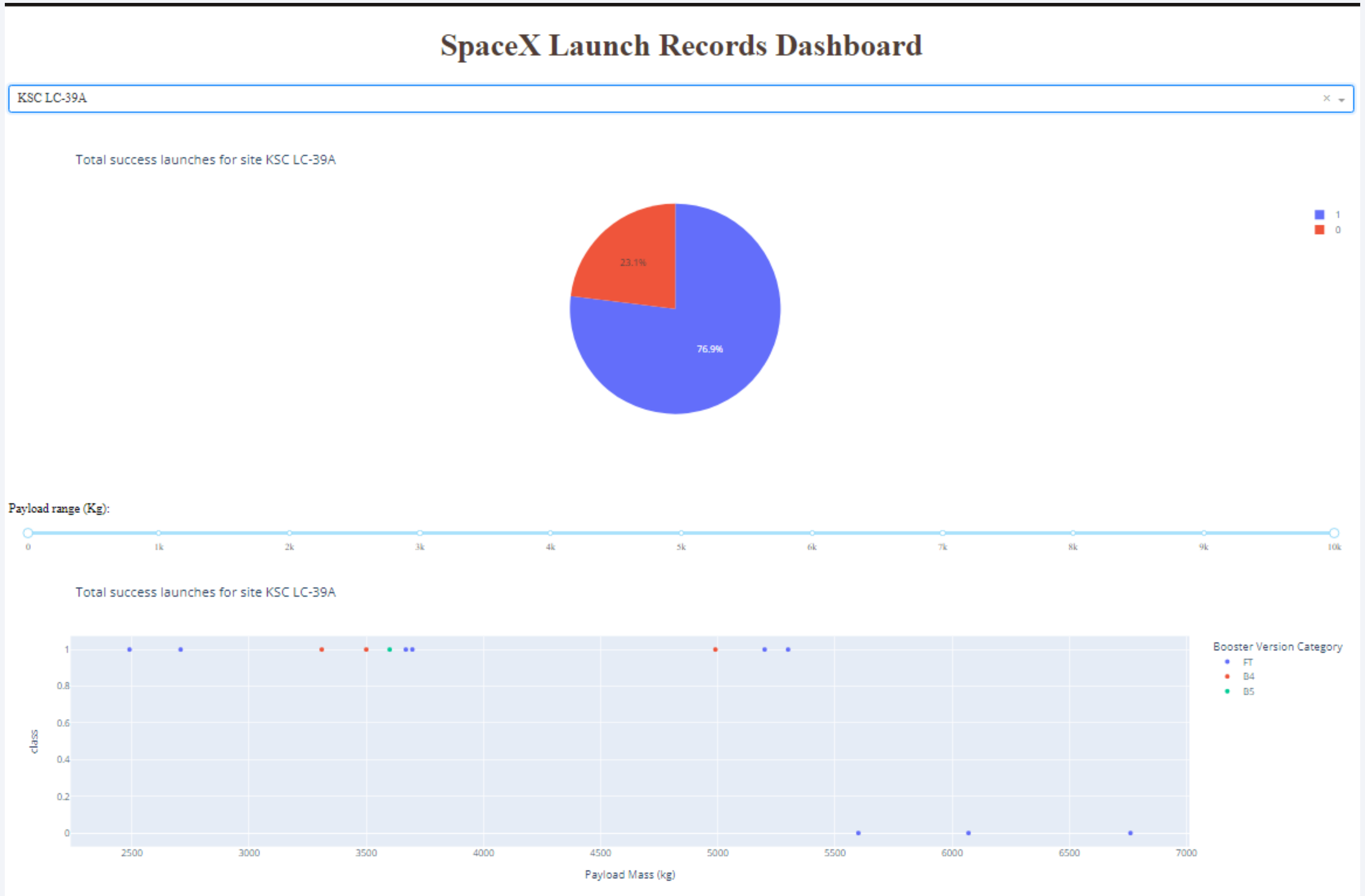
Total launches per launch site

- Most launches were made on KSC LC-39A, whereas CCAFS SLC-40 had the least number of launches.



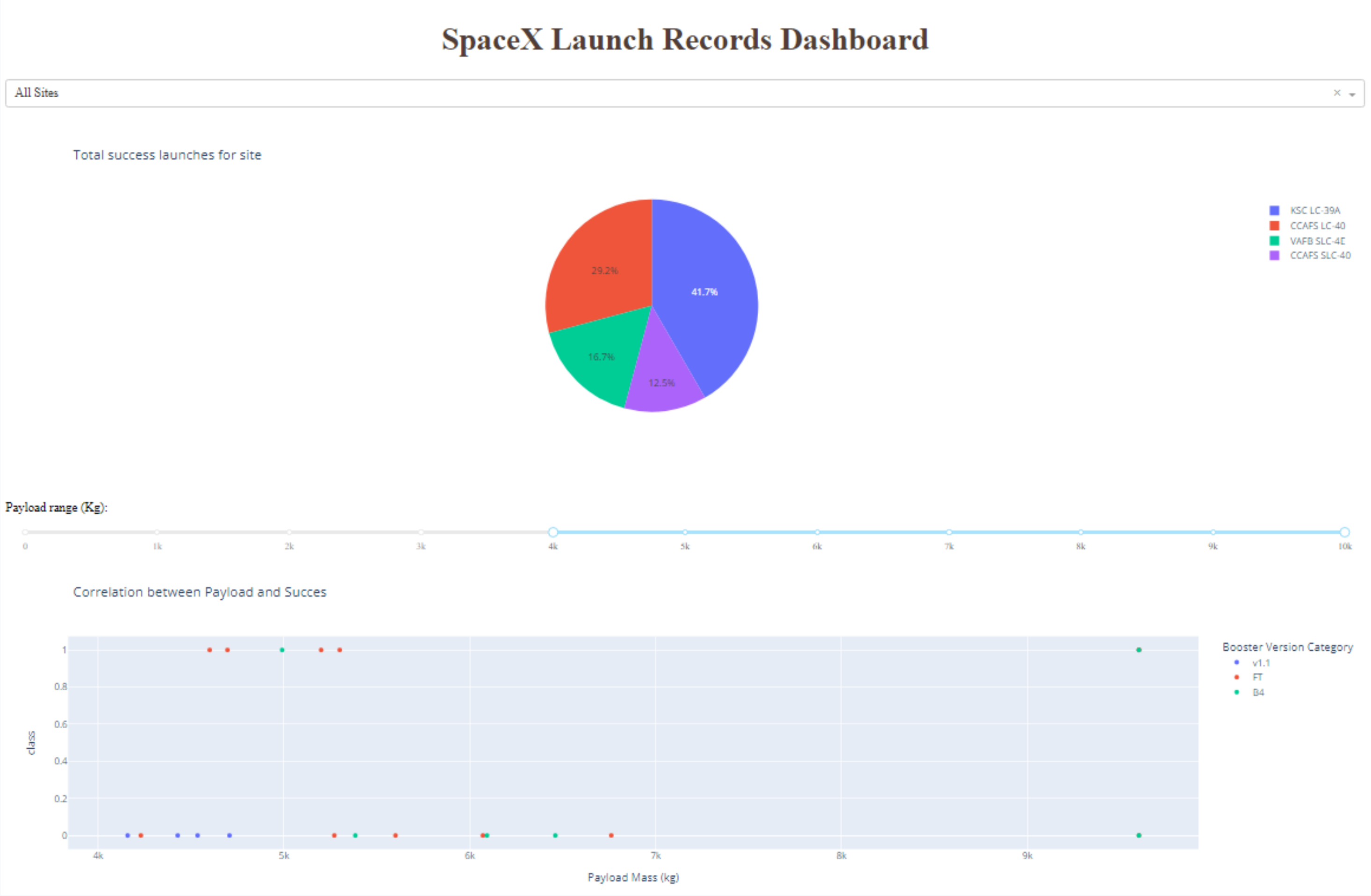
Most successful launch site

- The launch site with the most success ratio was KSC LC-39A.



Payload mass per Booster version

- Out of all 5 different booster versions, only 3 attempted to carry more than 4000 kg of payload.
- The B4 (green) booster version was the only one to carry more than 9000 kg of payload.



Section 5

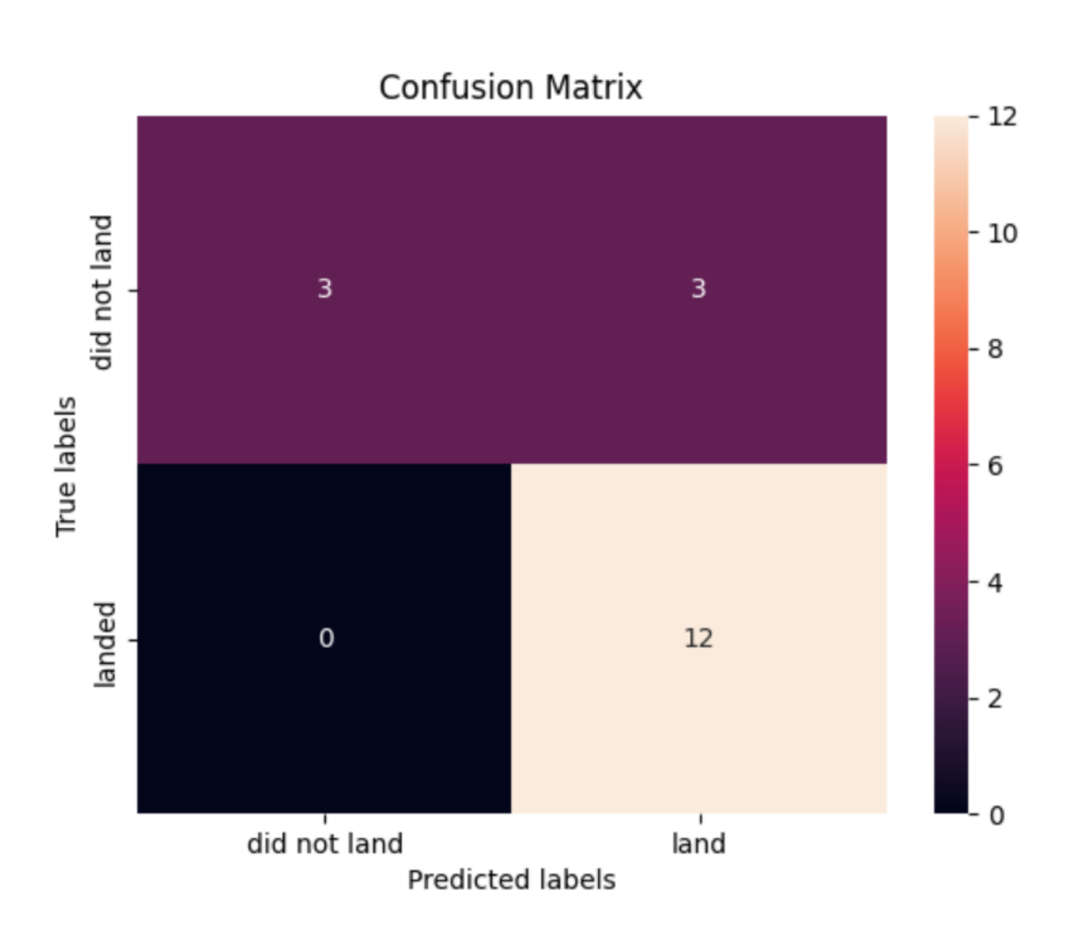
Predictive Analysis (Classification)

Classification Accuracy

- Four models were trained in total: Logistic regression, Support vector machine, Decision tree classifier and k-nn classifier.
- All model got an accuracy of 83.33% when tested with R squared on the test dataset.
- The model that had the highest accuracy with the training model was the decision tree with 88.75%.

Confusion Matrix

- The model did not have problem misclassifying false negatives, the problem was classifying false positives.
- It predicted 3 successful landings on launches that failed.



Conclusions

- Most data is not available in a structured form, but it can be transformed to be of value to the investigation.
- EDA can help identify problematic values on the dataset and understand the nature of the data.
- Visualization tools also help identifying patterns in the data an relationships with its attributes that are worth exploring.
- A ML model can be trained with the data to predict future landings success.
- The model that performed the best was the Decision Tree, but it had problems misclassifying false positives.

Thank you!