IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Veronika Szabó
06/08/2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Methodology

- Data collection using web scraping (Wikipedia) and with the help of the Spacex API
- Data wrangling and cleaning with Pandas Numpy
- Exploratory data analysis (EDA) using visualization libraries Seaborn and Numpy and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis building and evaluating classification models (Support vector machine, Decision tree, K-nearest neighbour, Logistic Regression)

Results

- As it is clearly visible on the success rate yearly trend graph, the success rate is increasing with time.
- Based on the exploratory data analysis the success rate has a positive correlation with certain launch sites (KSC LC 39A and VAFB SLC 4E).
- We can see that for the launch site CCAFS SLC 40 the larger payload mass augmented the success rates.
- Based on the success rate graph of each orbit type, success rate of orbit type GEO, HEO, SSO and ES.L-1 is 1, wile other orbit types have lower rates of success.

Introduction

- The goal of the project is to determine if the Falcone 9 launch first stage could be recovered in the future. This is a crucial step for our company, SpaceY in gaining advantage on the rocket launch market
- The success probability will be studied with accounting for factors such as the type of orbit, payload mass or different booster versions, to check for the best candidates and other relevant variables
- After completing the analysis, recommendations can be made on the best combination of parameters to maximize the first stage recovery

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected using web scraping (Wikipedia) and with the help of the Spacex API
- Perform data wrangling
  - After selecting the relevant variables, the date was stored in a dataframe, normalized and the missing values were being dealt with
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different models were built and evaluated based on their classification accuracy

Methodology

Data collection

Requests to SpaceX API

Webscraping (Beautiful soup)

Data wrangling

Performing data wrangling and cleaning with Pandas Numpy

Determining training labels based on success of landing

Exploratory data analysis

SQL

Data Visualisation (Seaborn, matplotlib)

Interactive visual analytics

Interactve maps (Folium)

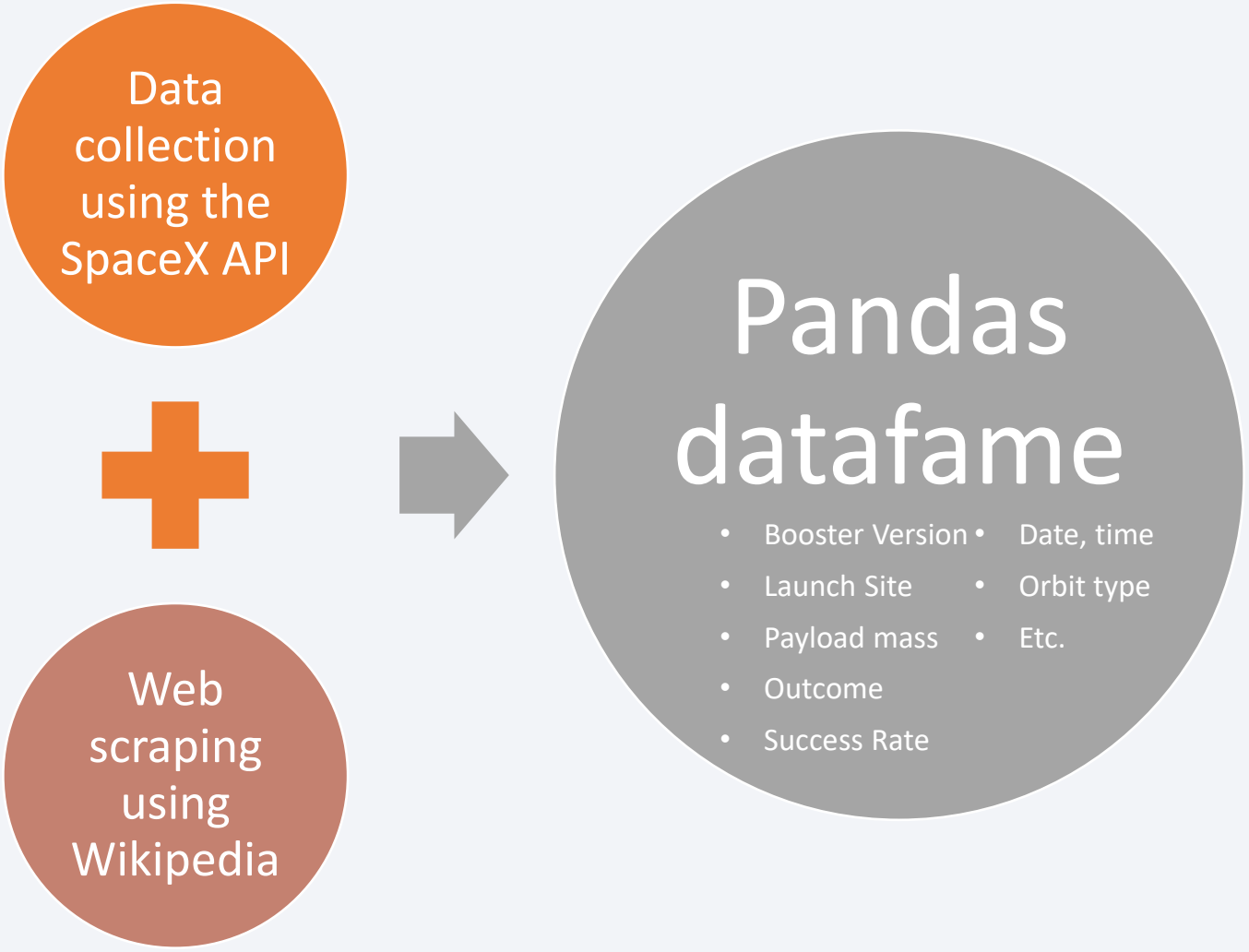
Interactive Dashboard (Plotly Dash)

Predictive analysis

Model Building

(Support vector machine, Decision tree, K-nearest neighbour, Logistic Regression) and evaluation based on classification accuracy

Data Collection



Data Collection – SpaceX API

Import and install libraries

- Requests
- Datetime
- Pandas
- Numpy

01

Define functions to call the API

- Booster Version
- Launch Sirte
- Payload mass
- Outcome etc.

02

Request data

- Request from the Spacex API and convert it to a pandas dataframe
- Select features to keep for the later analysis

03

Prepare and fill dataframe

- Create variables as empty lists to store data later on
- Use predefined functions to fill the dataframe, parsing the downloaded data

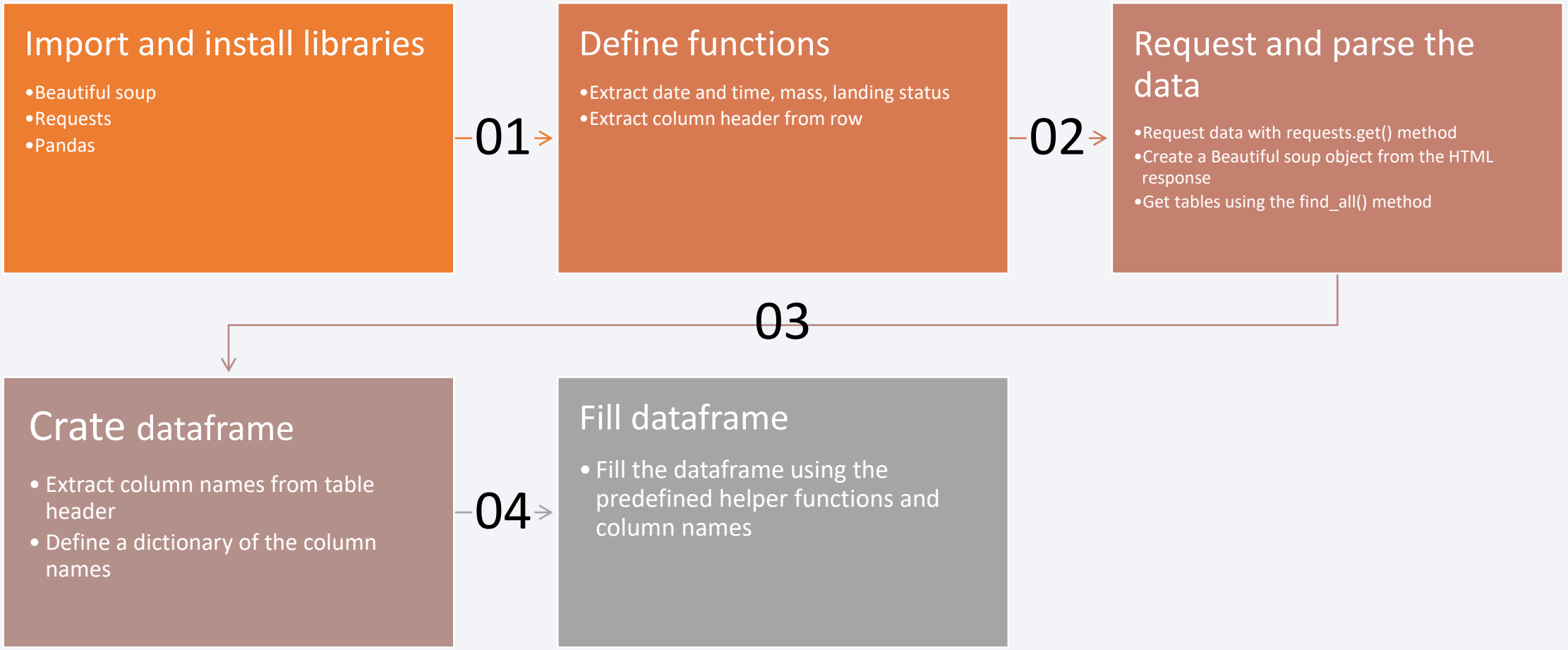
04

Dealing with missing values

Replace missing payload mass values with the mean of the column

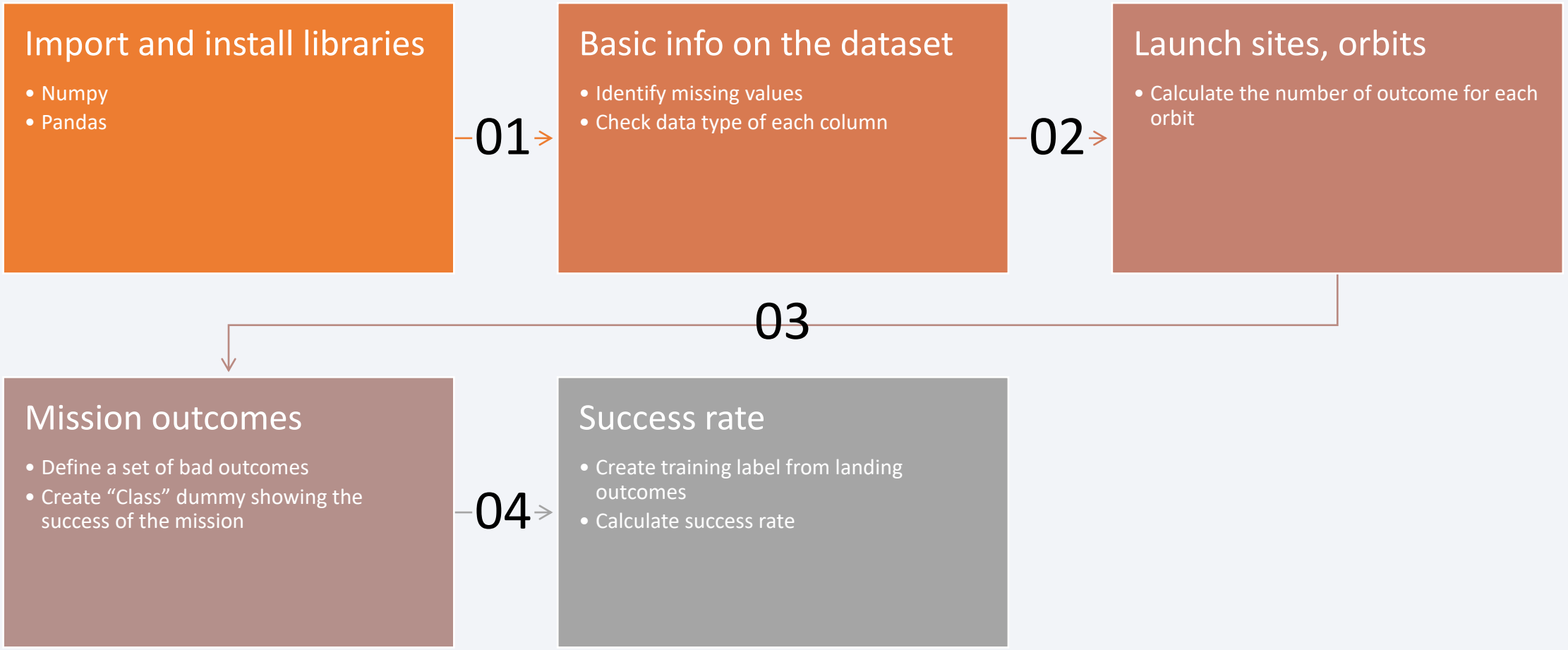
https://github.com/szbebo/capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



https://github.com/szbebo/capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling



https://github.com/szbebo/capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Visual of Flight number vs payload mass and Flight number vs launch site to visualize relationships
- Visual of Payload mass vs launch site to check if there is a correlation between the launch site and the payload mass carried
- Visual of Success rate, Payload mass and Flight number for each Orbit type to visualize patterns in the data related to the Orbit type
- Visual of Launch success yearly trend, to have a look at the general trend over time

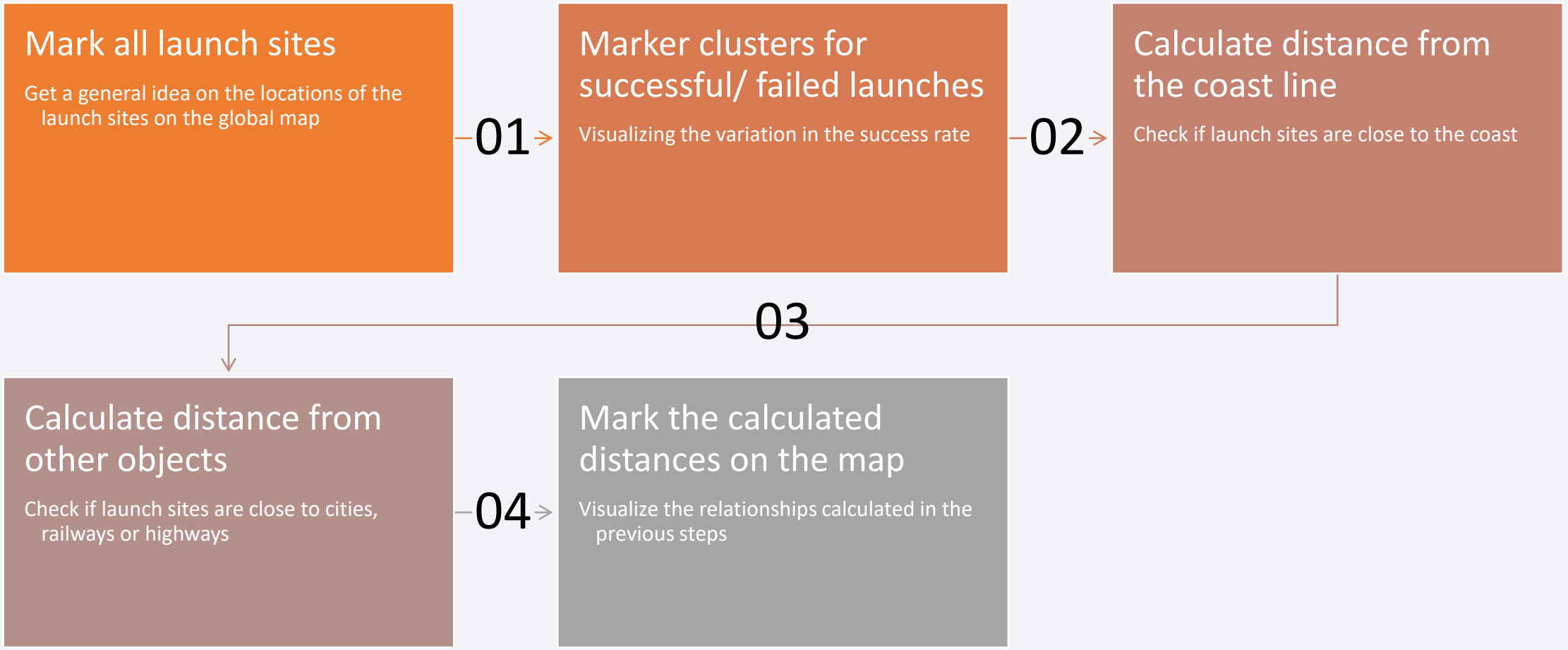
https://github.com/szbebo/capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Display the names of the unique launch sites in the space mission	SELECT Distinct LAUNCH_SITE FROM SPACEXTBL
5 records with launch sites starting with CCA	SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
The total payload mass carried by boosters launched by NASA (CRS)	SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer='NASA (CRS)'
The average payload mass carried by booster version F9 v1.1	SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%'
Date of the first successful ground pad landing	SELECT MIN(DATE) FROM SPACEXTBL WHERE [Landing _Outcome]='Success (ground pad)'
Booster versions with payload mass between 4000 and 6000 kg and successful landing (drone ship)	SELECT Distinct Booster_Version FROM SPACEXTBL WHERE [Landing _Outcome]='Success (drone ship)'AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
Total number of mission outcomes	SELECT DISTINCT [Mission_Outcome] FROM SPACEXTBL
Failed and successful mission outcome counts	SELECT COUNT([Mission_Outcome]) FROM SPACEXTBL WHERE [Mission_Outcome] LIKE 'Success%' SELECT COUNT([Mission_Outcome]) FROM SPACEXTBL WHERE [Mission_Outcome] LIKE 'Fail%'
List of the booster versions with maximum payload mass	SELECT Distinct Booster_Version FROM SPACEXTBL WHERE [PAYLOAD_MASS__KG_]=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
List the landing outcomes from 2015 which resulted in Failure (drone ship)	SELECT substr(Date,7,4) AS YEAR, (substr(Date, 4, 2)) AS MONTH, [Landing _Outcome] AS 'LANDING OUTCOME', [Booster_Version] AS 'BOOSTER VERSION', [Launch_Site] AS 'LAUNCH SITE' FROM SPACEXTBL WHERE [Landing _Outcome]='Failure (drone ship)' AND DATE LIKE '%2015%'
Rank successful landing outcomes in descending order	SELECT [Landing _Outcome], COUNT ([Landing _Outcome]) FROM SPACEXTBL WHERE DATE BETWEEN' 2010-06-04'AND '2017-03-20' AND [Landing _Outcome] LIKE 'Success%' GROUP BY [Landing _Outcome] ORDER BY COUNT([Landing _Outcome]) DESC

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

Build an Interactive Map with Folium



https://github.com/szbebo/capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Pie chart

to show the success read of the launch sites

Dropdown list

to enable Launch Site selection

Scatter chart

to show the correlation between payload and launch success

Slider

to select payload range

Callback function

translate input from the payload slider and the dropdown to the pie chart and the scatter plot

https://github.com/szbebo/capstone/blob/main/plotly.ipynb

Predictive Analysis (Classification)

Import and install libraries

Numpy, Pandas,
Matplotlib, Seaborn,
Sklearn

01

Define functions

Create the helper function for plotting the confusion matrix

02

Load data

Load the prepared Pandas dataframe

03

Mission outcomes

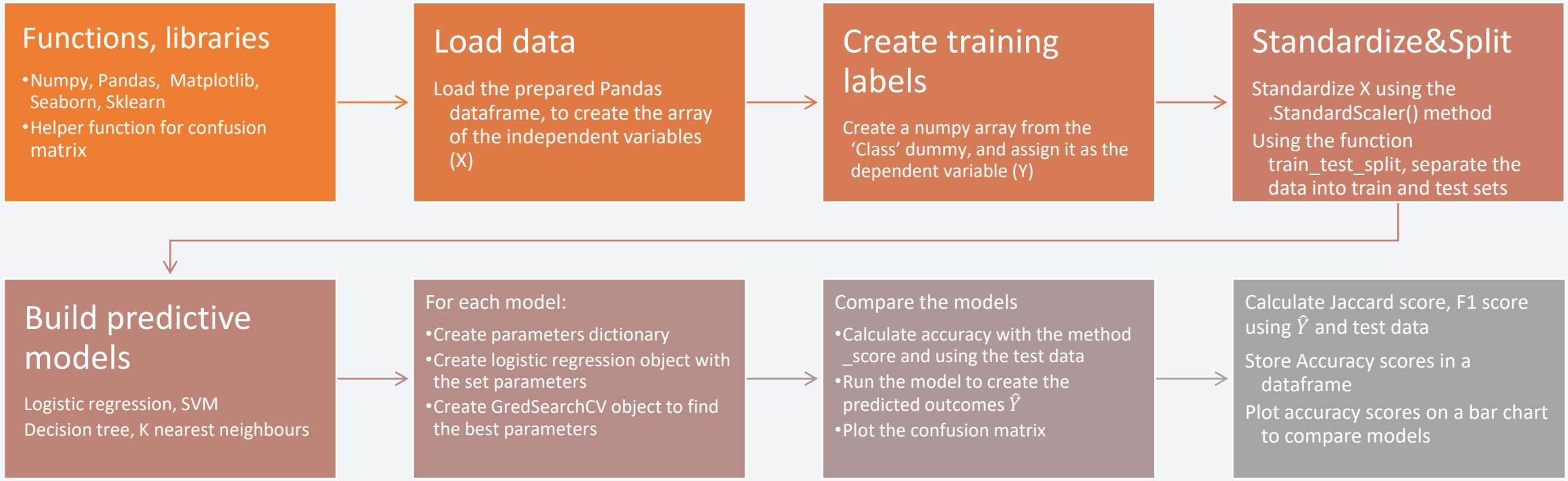
- Define a set of bad outcomes
- Create “Class” dummy showing the success of the mission

04

Success rate

- Create training label from landing outcomes
- Calculate success rate

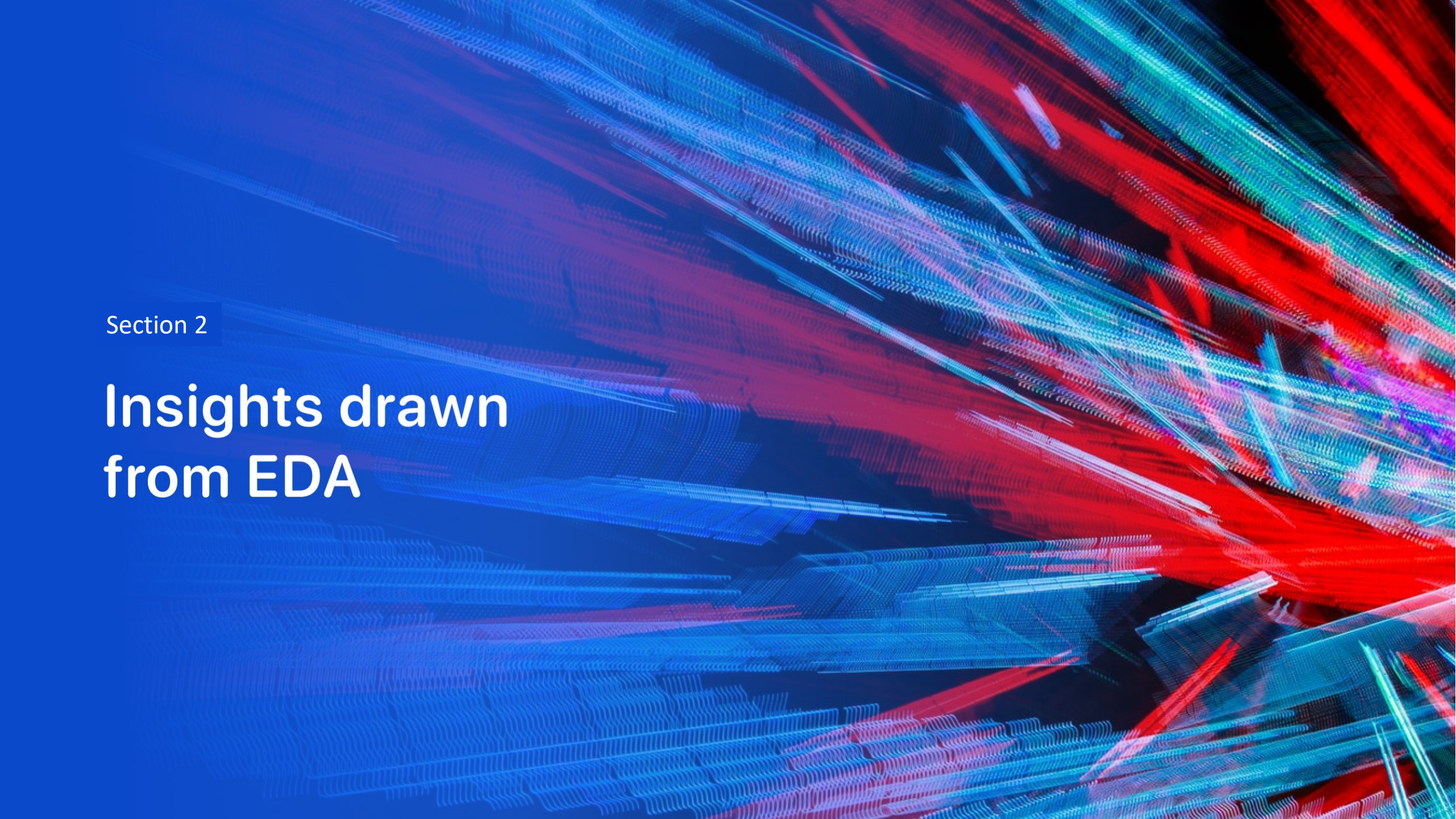
Predictive Analysis (Classification)



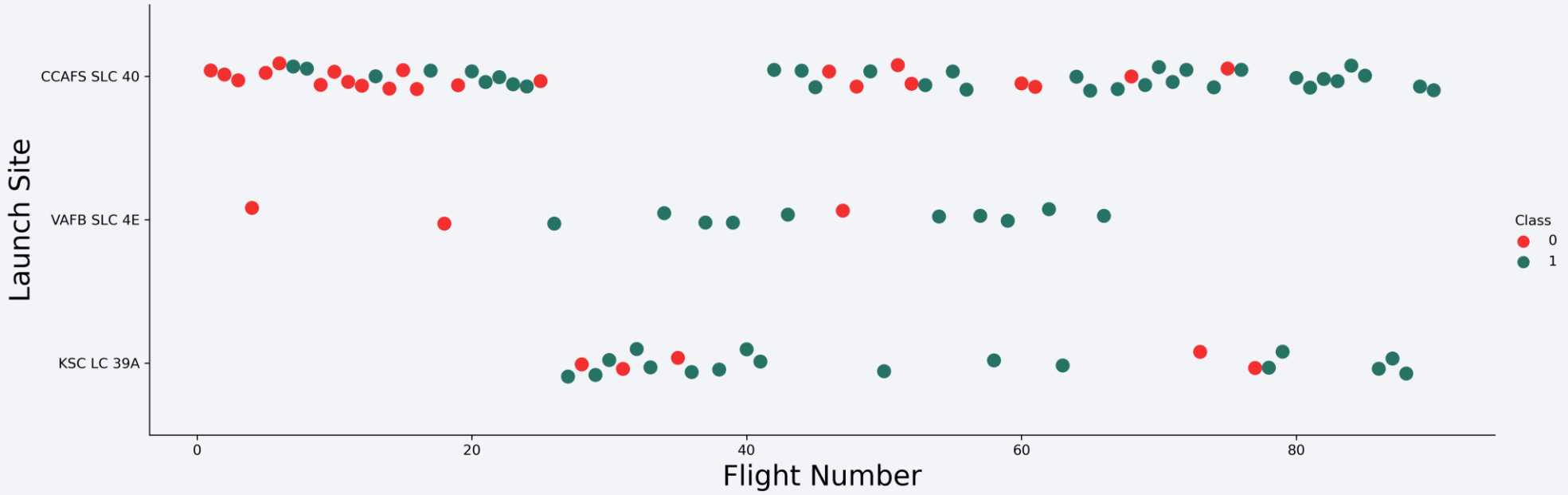
https://github.com/szbebo/capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



As we can see on the scatter plot, in the beginning more flights were launched from the site CCAFS SLC 40, followed by a period of more flights at KSC LC 39A, after which flights were launched from both launch sites.

The number of flights launched from the site VAFB SLC 4E remain low in volume compared to the other two. What we can see in general is that with the flight number the ratio of successful flights is increasing.

Payload vs. Launch Site

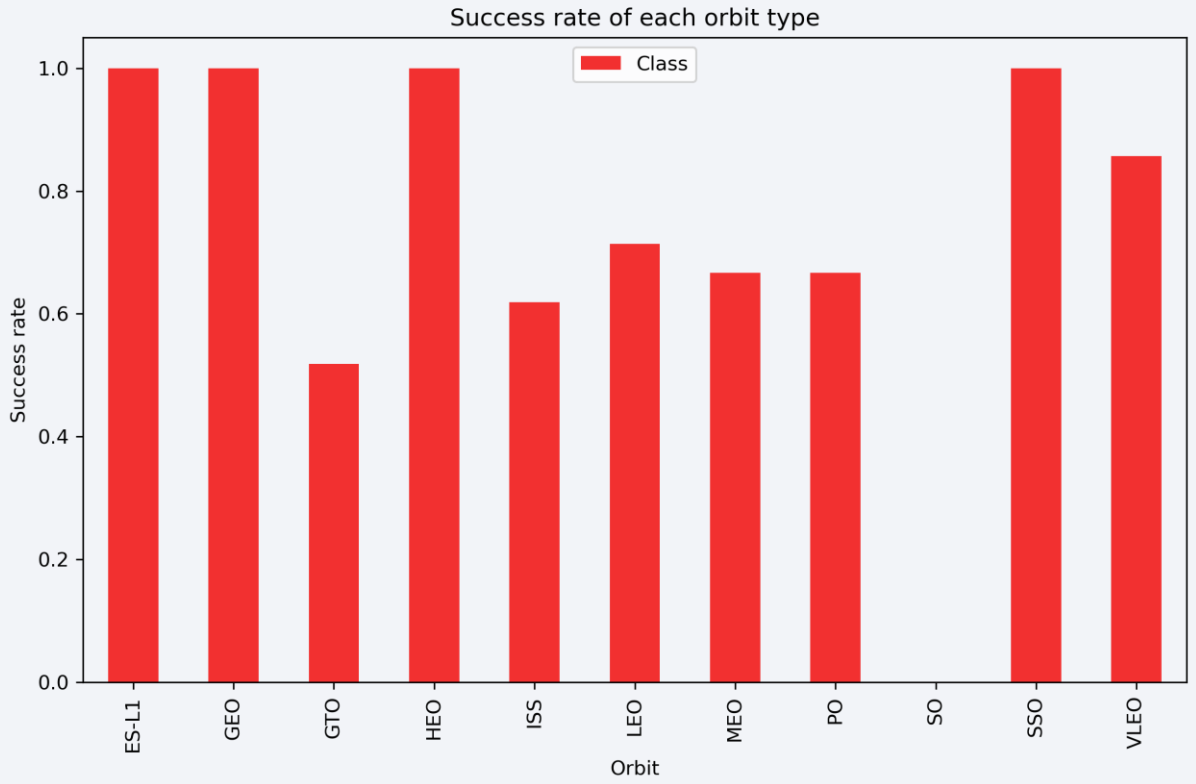
Flights with payload mass larger than 10000 kg were only launched from either KSC LC 39A or CCAFS SLC 40, but never from VAFB SLC 4E. All three sites were usen with lower payload mass.

Success Rate vs. Orbit Type

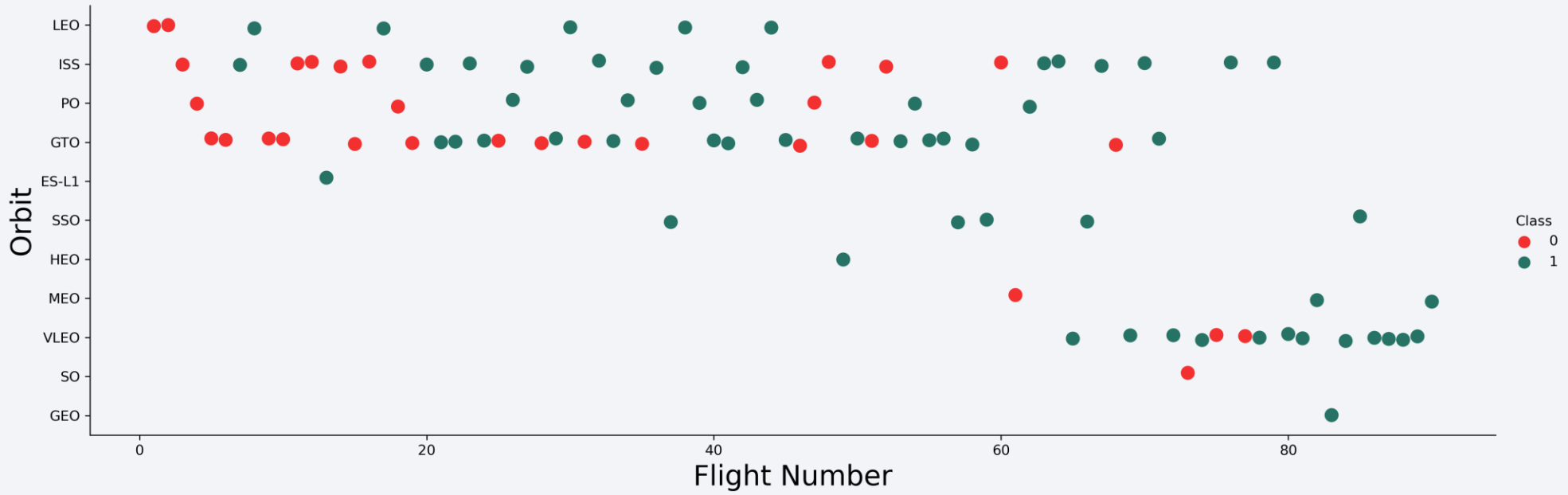
We can see that the success rate of orbit type GEO, HEO, SSO and ES.L-1 is 1, while below 0.8 for the others. For orbit type SO the success rate is 0.

These success rates must be interpreted with caution, as some orbit types (MEO, SO, GEO, ES.L-1), see the table below.

Orbit type	Number of records
GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3
SO	1
GEO	1
ES-L1	1
HEO	1

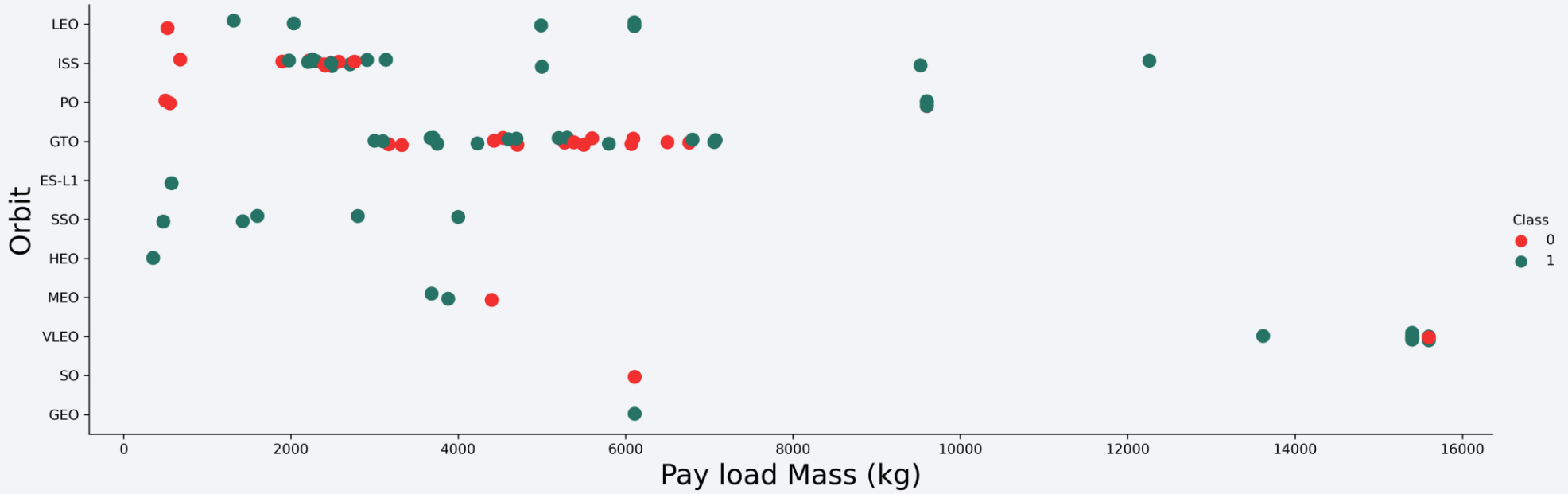


Flight Number vs. Orbit Type



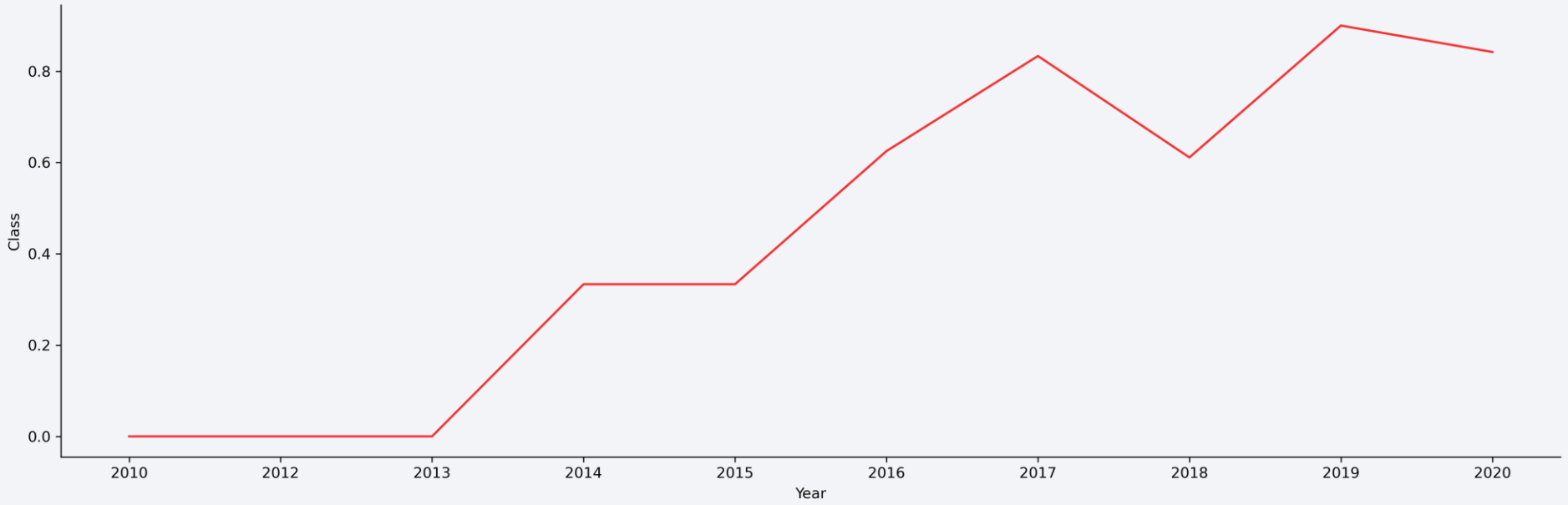
We can see that there is a change in the orbit types used. With higher flight numbers there are less flights in LEO orbit, but new types appear, for example VLEO and SSO, Also, there is a low amount of datapoints for SO, GEO MEO orbits.

Payload vs. Orbit Type



The variance of orbit types with lower payload mass is higher that with high payload mass. Flights with payload mass higher than 10000 kg were only in VLEO or ISS orbit.

Launch Success Yearly Trend



There is a clear positive trend when it comes to success rate over the years.

All Launch Site Names

- There are four launch site names in total:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

In [12]:
```
%sql SELECT Distinct LAUNCH_SITE FROM SPACEXTBL
```

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

Out[12]:

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

Launch Site Names Begin with 'CCA'

- The 5 records are all from the launch site CCAFS LC-40

```
In [13]:  1  %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

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

Out[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	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 is 45596 kg. This is the sum of the payloads where the customer is Nasa.

```
In [14]:  1 %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer='NASA (CRS)'
           * sqlite:///my_data1.db
          Done.
Out[14]: SUM(PAYLOAD_MASS__KG_)
                           45596
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534.7 kg

```
In [15]:  1 %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%'
 * sqlite:///my_data1.db
Done.
Out[15]: AVG(PAYLOAD_MASS__KG_)
         2534.6666666666665
```

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad is 01-05-2017.
- In order write the query, I need ed the possible landing outcomes, so that I could filter for the right type.

```
In [16]:
1 %sql SELECT Distinct [Landing _Outcome] FROM SPACEXTBL
```

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

Out[16]:

Landing _Outcome
Failure (parachute)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
Success (ground pad)
Success (drone ship)
Success
Failure
No attempt

```
In [17]:
1 %sql SELECT MIN(DATE) FROM SPACEXTBL WHERE [Landing _Outcome]='Success (ground pad)'
```

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

Out[17]:

MIN(DATE)
01-05-2017

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 are:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

```
In [13]:
1 %sql SELECT  Distinct Booster_Version FROM SPACEXTBL WHERE [Landing _Outcome]='Success (drone ship)'AND \
2 PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

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

Out[13]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes are 100:1, there is only one record where the mission outcome is failure, however the Success (payload status unclear) outcome exists as well

```
In [14]: 1 %sql SELECT DISTINCT [Mission_Outcome] FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

Out[14]:

Mission_Outcome
Success
Failure (in flight)
Success (payload status unclear)
Success

```
In [15]: 1 %sql SELECT COUNT([Mission_Outcome]) FROM SPACEXTBL WHERE [Mission_Outcome] LIKE 'Success%'
 * sqlite:///my_data1.db
Done.
```

Out[15]:

COUNT([Mission_Outcome])
100

```
In [16]: 1 %sql SELECT COUNT([Mission_Outcome]) FROM SPACEXTBL WHERE [Mission_Outcome] LIKE 'Fail%'
 * sqlite:///my_data1.db
Done.
```

Out[16]:

COUNT([Mission_Outcome])
1

Boosters Carried Maximum Payload

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

In [17]:
```
%sql SELECT  Distinct Booster_Version FROM SPACEXTBL WHERE [PAYLOAD_MASS__KG_]=(SELECT MAX(PAYLOAD_MASS__KG_) \
                                                                                FROM SPACEXTBL)
```

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

Out[17]:

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

- There are two records with failed landing_outcomes in drone ship. Their booster versions, and launch site names for in year 2015:

In [18]:
```
%sql SELECT substr(Date,7,4) AS YEAR,\
(substr(Date, 4, 2)) AS MONTH,\
[Landing _Outcome] AS 'LANDING OUTCOME',\
[Booster_Version] AS 'BOOSTER VERSION',\
[Launch_Site] AS 'LAUNCH SITE'\
FROM SPACEXTBL\
WHERE [Landing _Outcome]='Failure (drone ship)'\
AND DATE LIKE '%2015%'
```

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

Out[18]:

YEAR	MONTH	LANDING OUTCOME	BOOSTER VERSION	LAUNCH SITE
2015	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- 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 is visible on the screenshot. We can see that for the largest number of records there is no information on if the landing was successful on a ground pad or a drone ship.

In [19]:
```
%sql SELECT [Landing _Outcome], COUNT ([Landing _Outcome])  FROM SPACEXTBL\
WHERE DATE BETWEEN' 2010-06-04'AND '2017-03-20' AND [Landing _Outcome] LIKE 'Success%'\
GROUP BY [Landing _Outcome]\
ORDER BY COUNT([Landing _Outcome]) DESC
```

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

Out[19]:

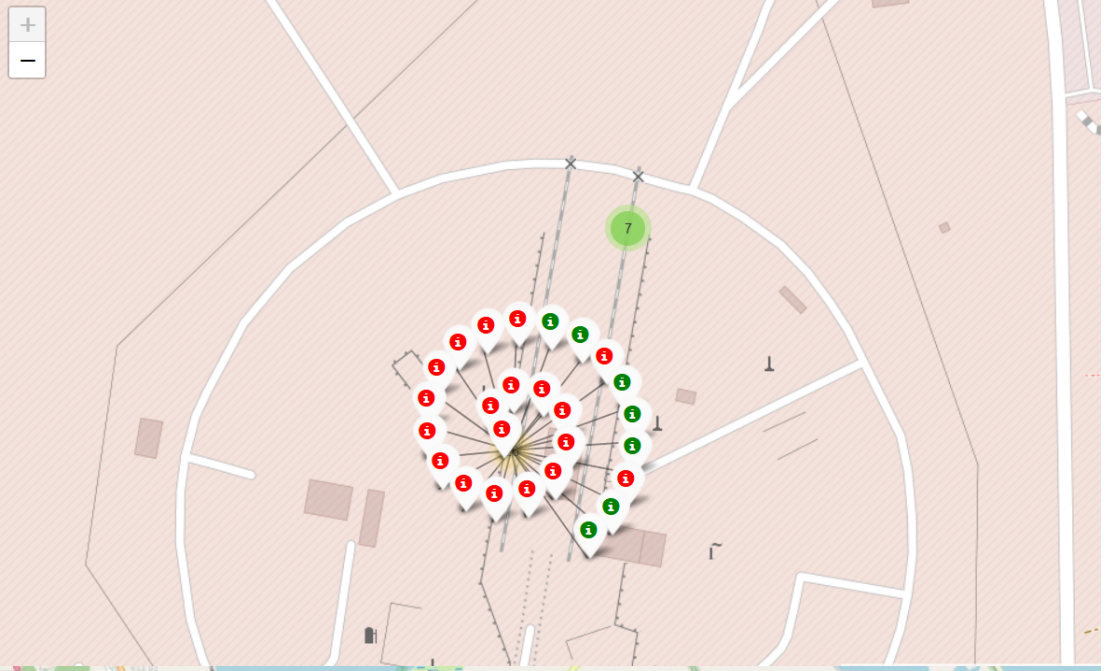
Landing _Outcome	COUNT (Landing _Outcome)
Success	25
Success (ground pad)	8
Success (drone ship)	8

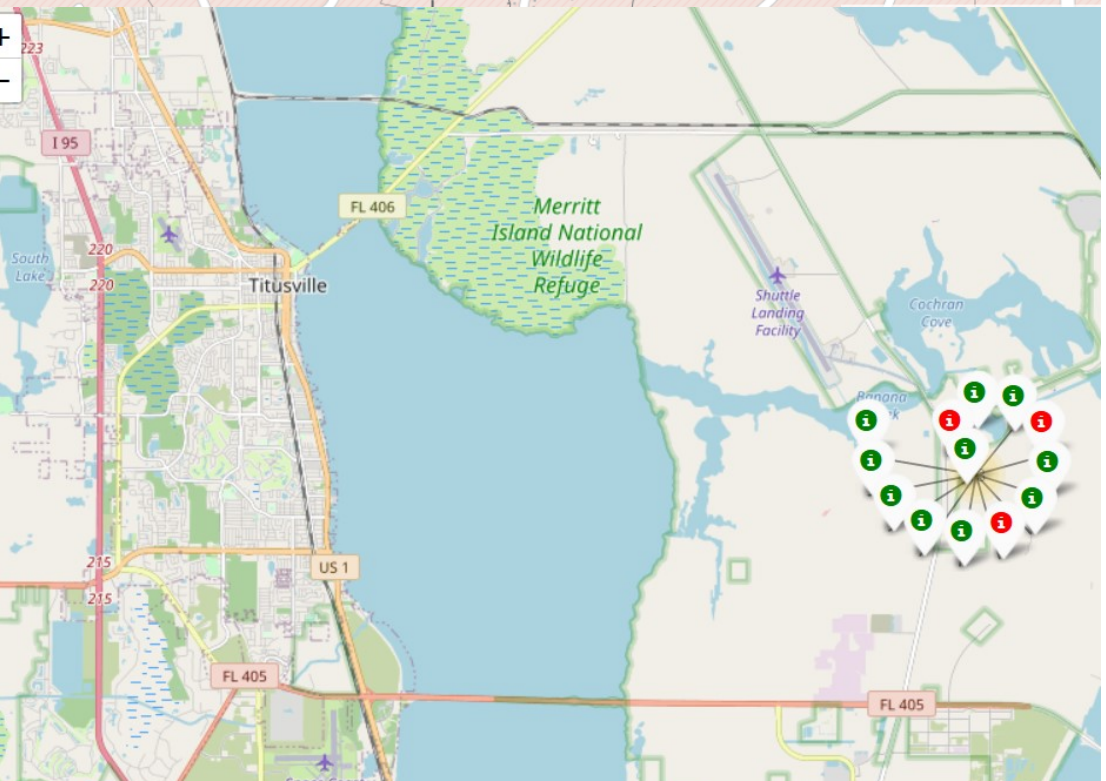
Section 3

Launch Sites Proximities Analysis

Location of the SpaceX Launch Sites

It is visible from the map that the launch sites are as close to the equator as possible, on the southern parts of the US. They are located close to the coastline.





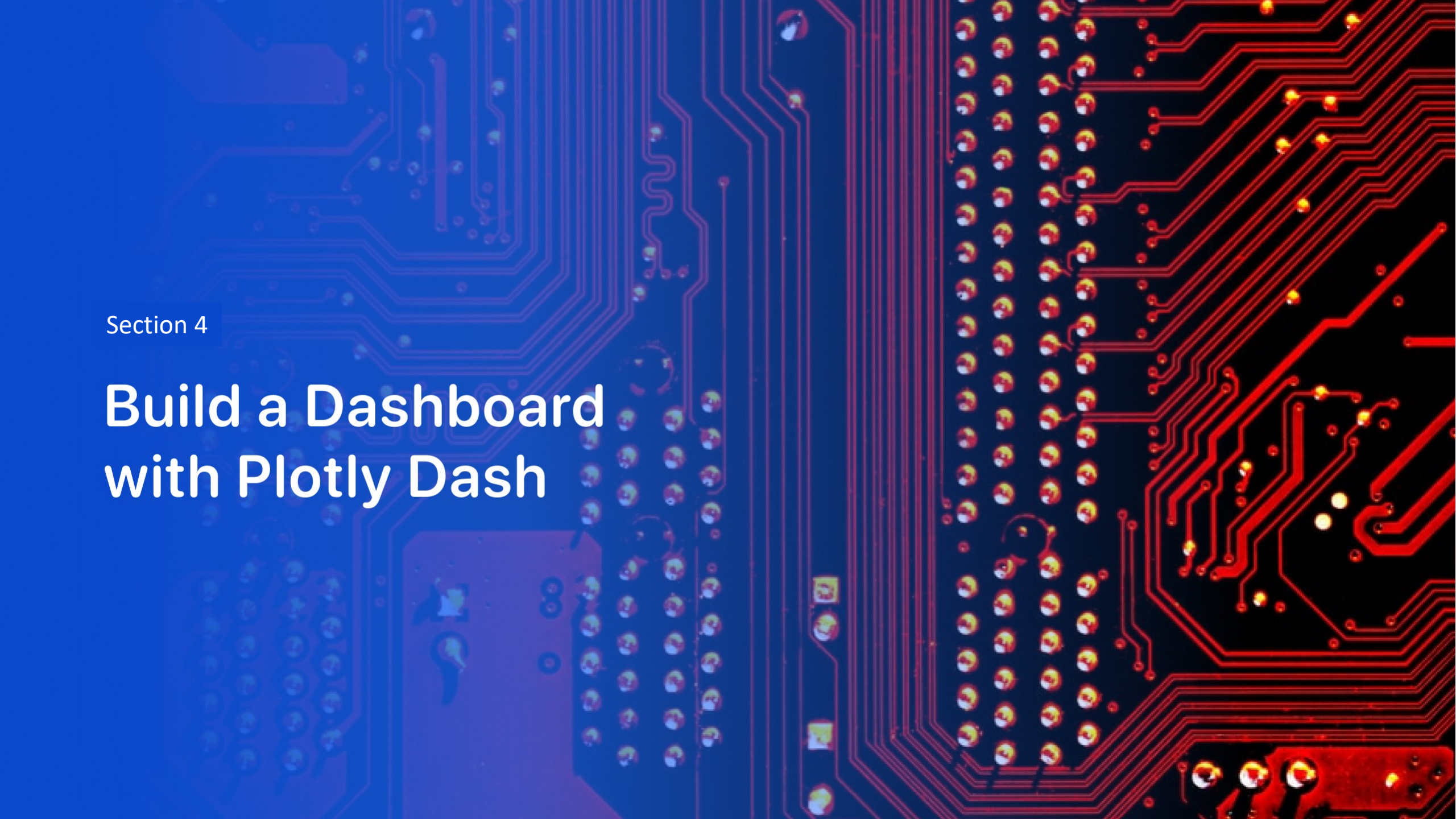
Successful and failed launches

- It is visible on the maps that there can be large differences in the success rate of missions between launch sites

Launch site and proximities

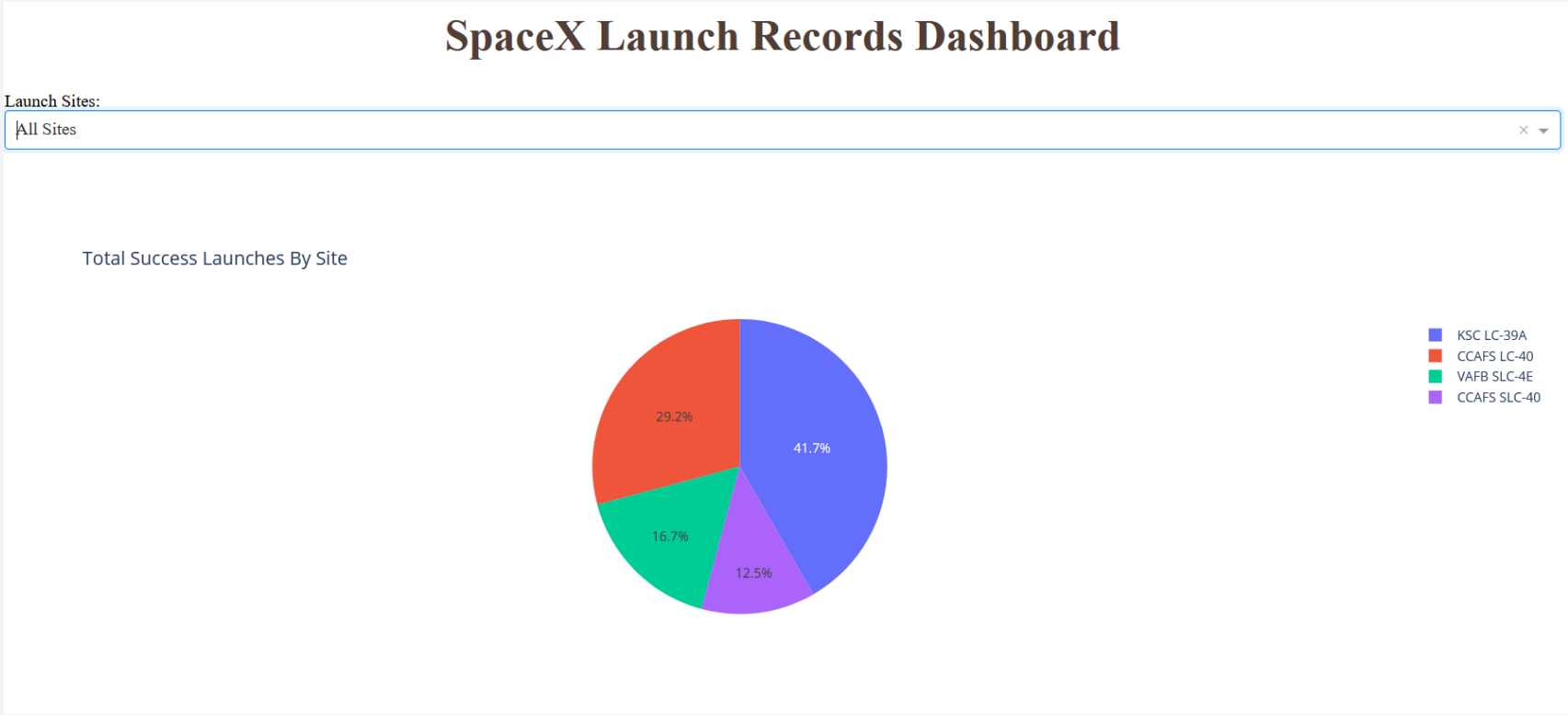
The map shows the launch site location and the distance from the nearest coastline, railway and city. As it is visible on the map, the launch site is only around 1 km away from the coast and the railway, but it is more than 10 km away from the nearest city.

To calculate and show the distance, the coordinates of the railway, coastline and city are extracted with the help or the Mouse Position method, the distance is calculated with the helper function and is shown with a folium.Marker object


Section 4
Build a Dashboard with Plotly Dash

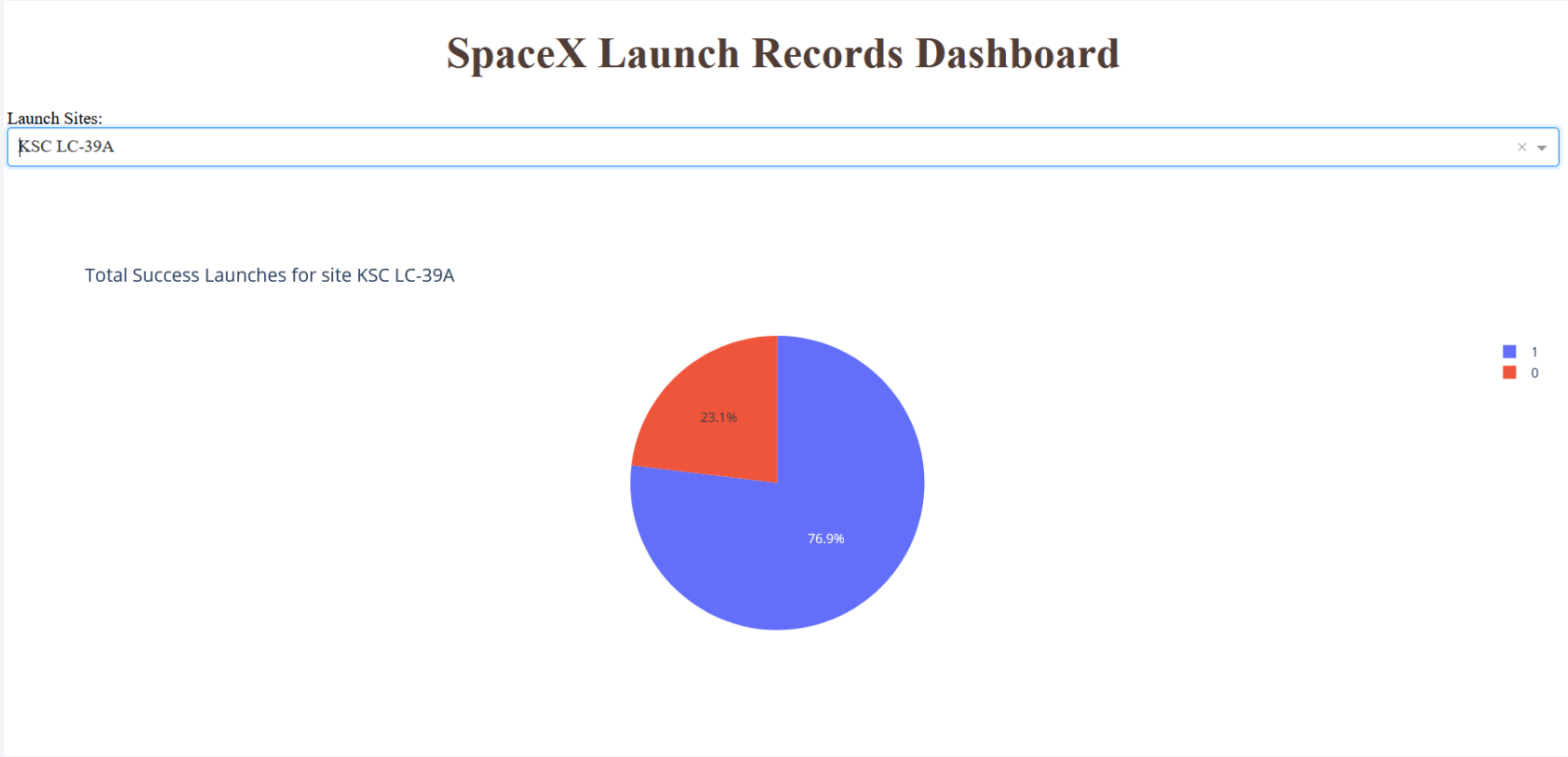
Dashboard – All sites

As visible on the screenshot, the site having the highest number of success launches is the CCAFS SLC 40



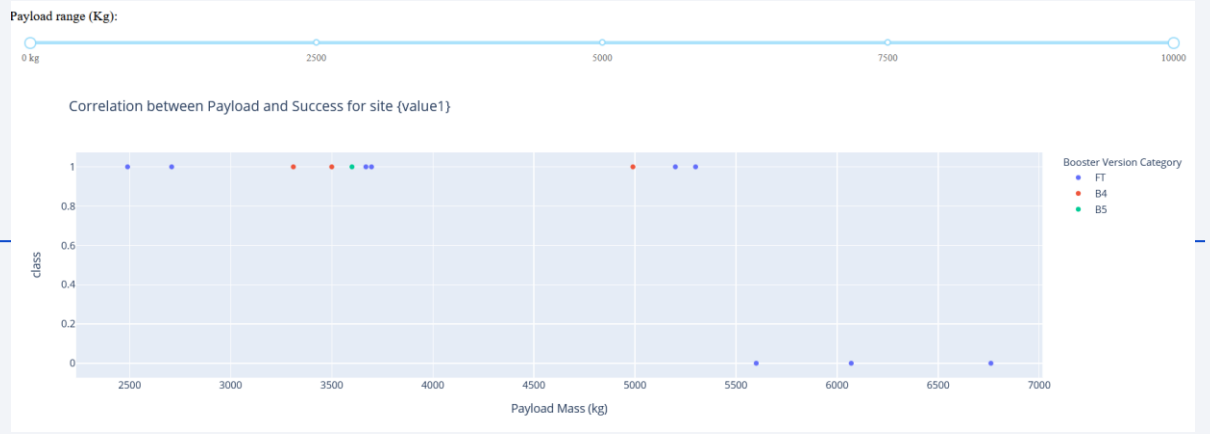
Dashboard – The highest success rate

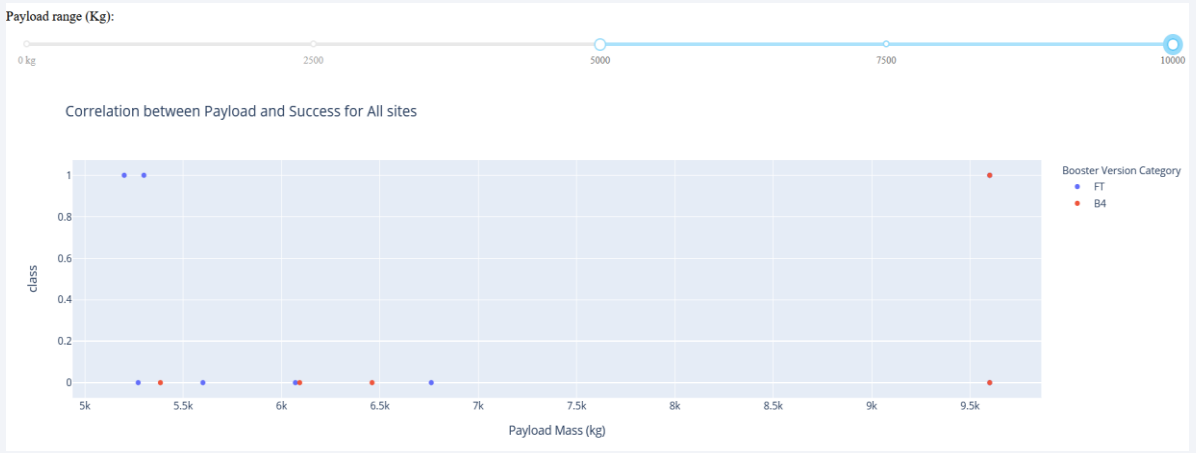
- The launch site with the highest success ration is the KSC LC-39A, with a success rate of 77%



Payload vs. Outcome

- It is visible on the screenshot that there are differences in the use of booster versions depending on the payload mass. With larger payload mass, the booster versions are B4 or FT, while at lower masses there is more variety.





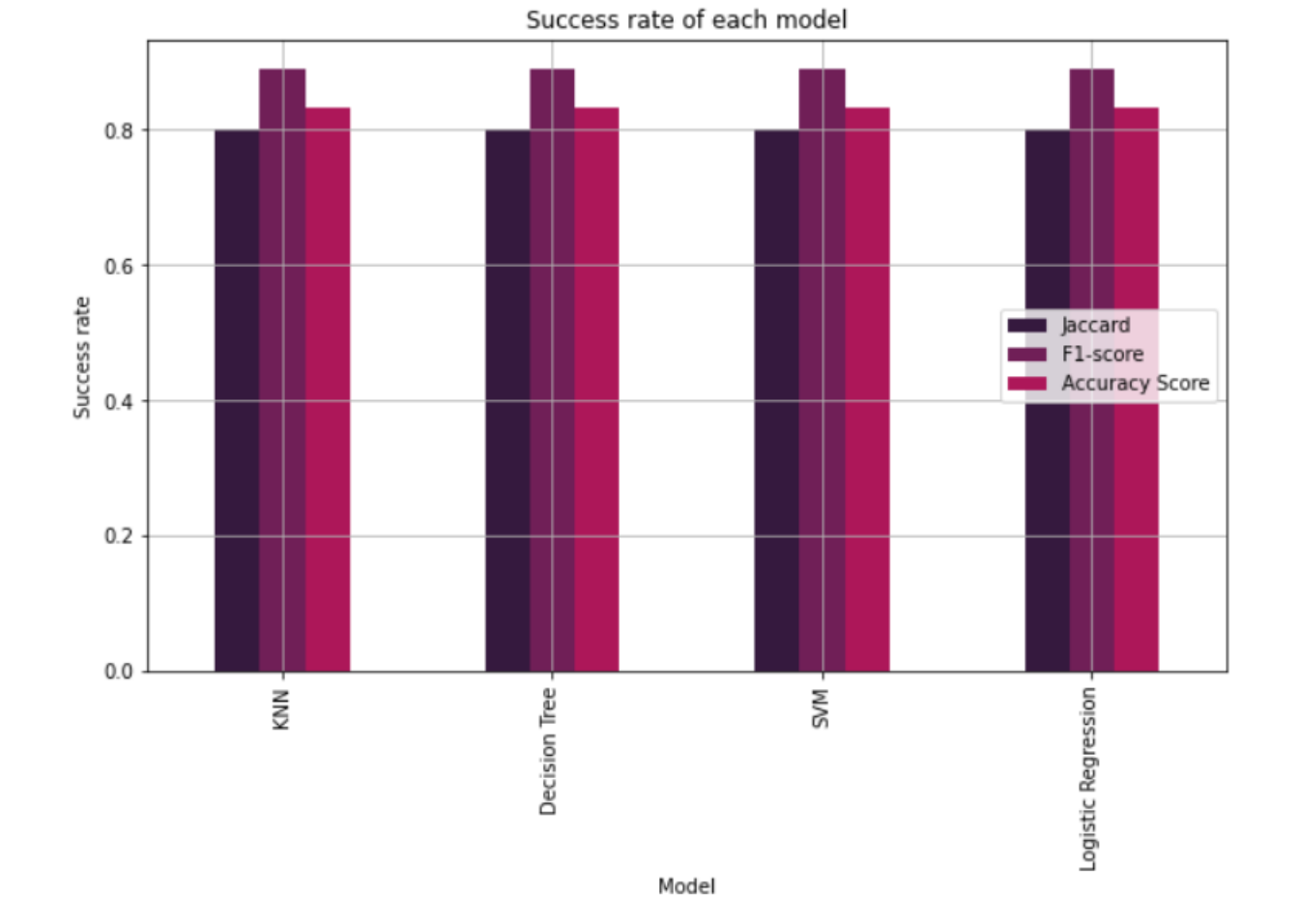
Section 5

Predictive Analysis (Classification)

Classification Accuracy

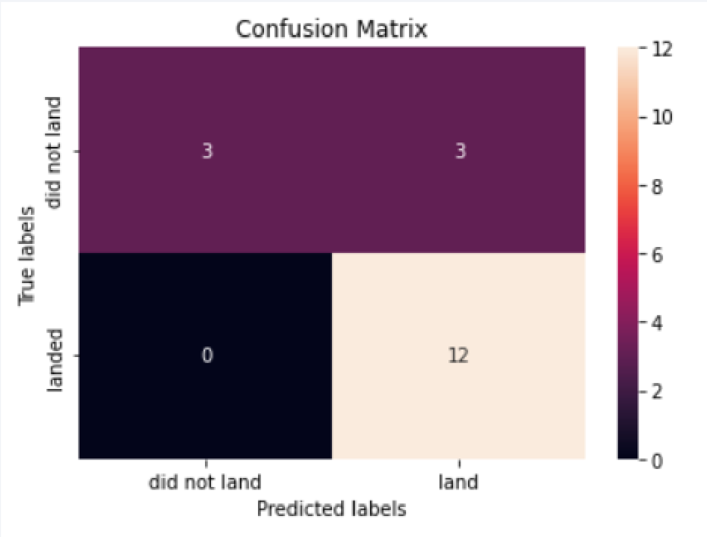
As it can be seen on the bar chart, all models have the same accuracy.

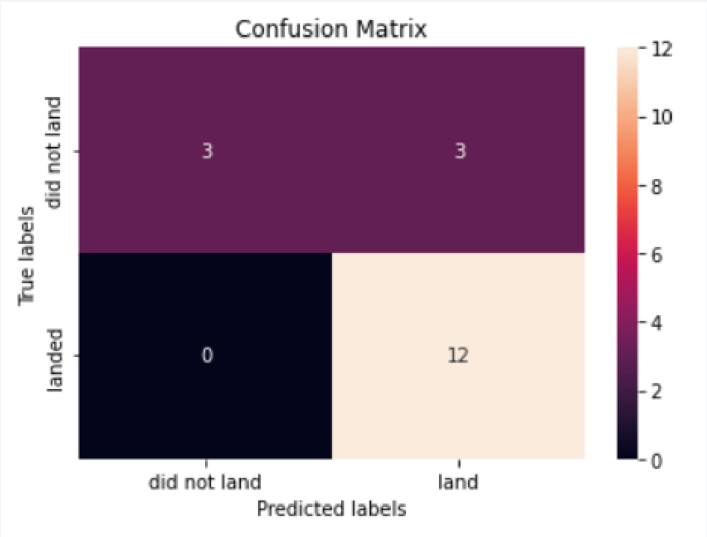
Model	Jaccard	F1-score	Accuracy Score
KNN	0.800000	0.888889	0.833333
Decision Tree	0.800000	0.888889	0.833333
SVM	0.800000	0.888889	0.833333
Logistic Regression	0.800000	0.888889	0.833333

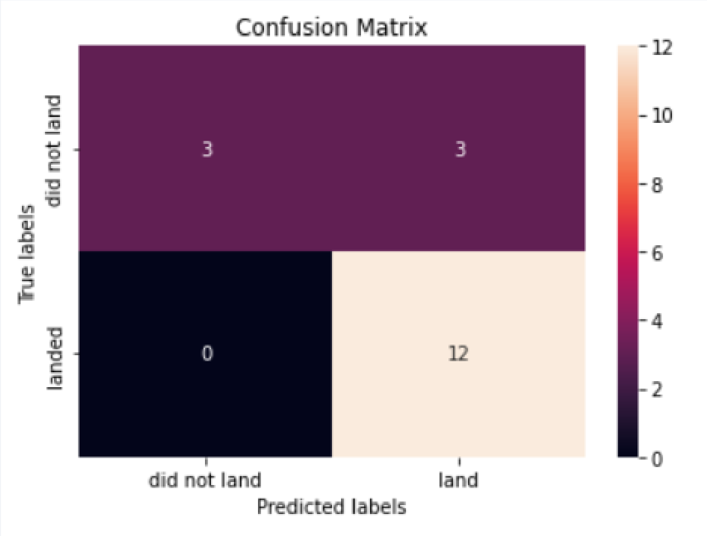


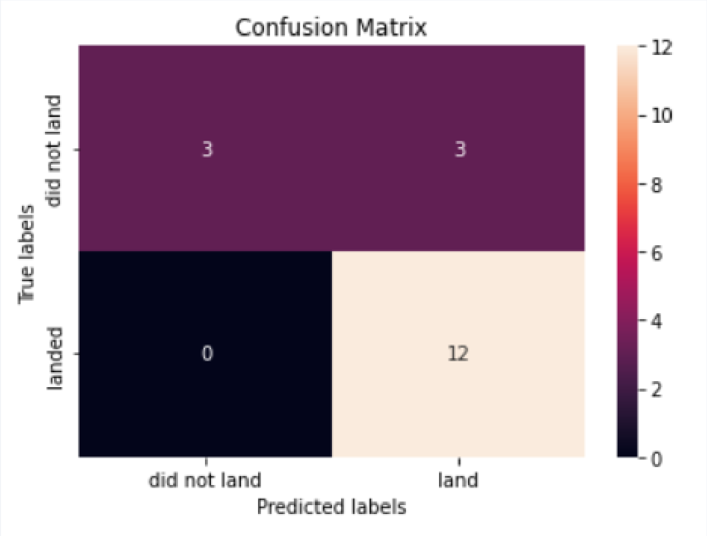
Confusion Matrix

- The confusion matrix is the same for all models, as they perform with the same accuracy level.









Conclusions

- Based on the exploratory data analysis the success rate has a positive correlation with certain launch sites (KSC LC 39A and VAFB SLC 4E). This relationship must be evaluated further with factoring for the effects of other variables, such as time, payload mass, etc.
- We can see that for the launch site CCAFS SLC 40 the larger payload mass augmented the success rates. This is not true though for the other launch sites.
- Based on the success rate graph of each orbit type, success rate of orbit type GEO, HEO, SSO and ES.L-1 is 1, wile other orbit types have lower rates of success. These success rates must be interpreted with caution, as some orbit types (MEO, SO, GEO, ES.L-1).
- As it is clearly visible on the success rate yearly trend graph, the success rate is increasing with time.

Thank you!