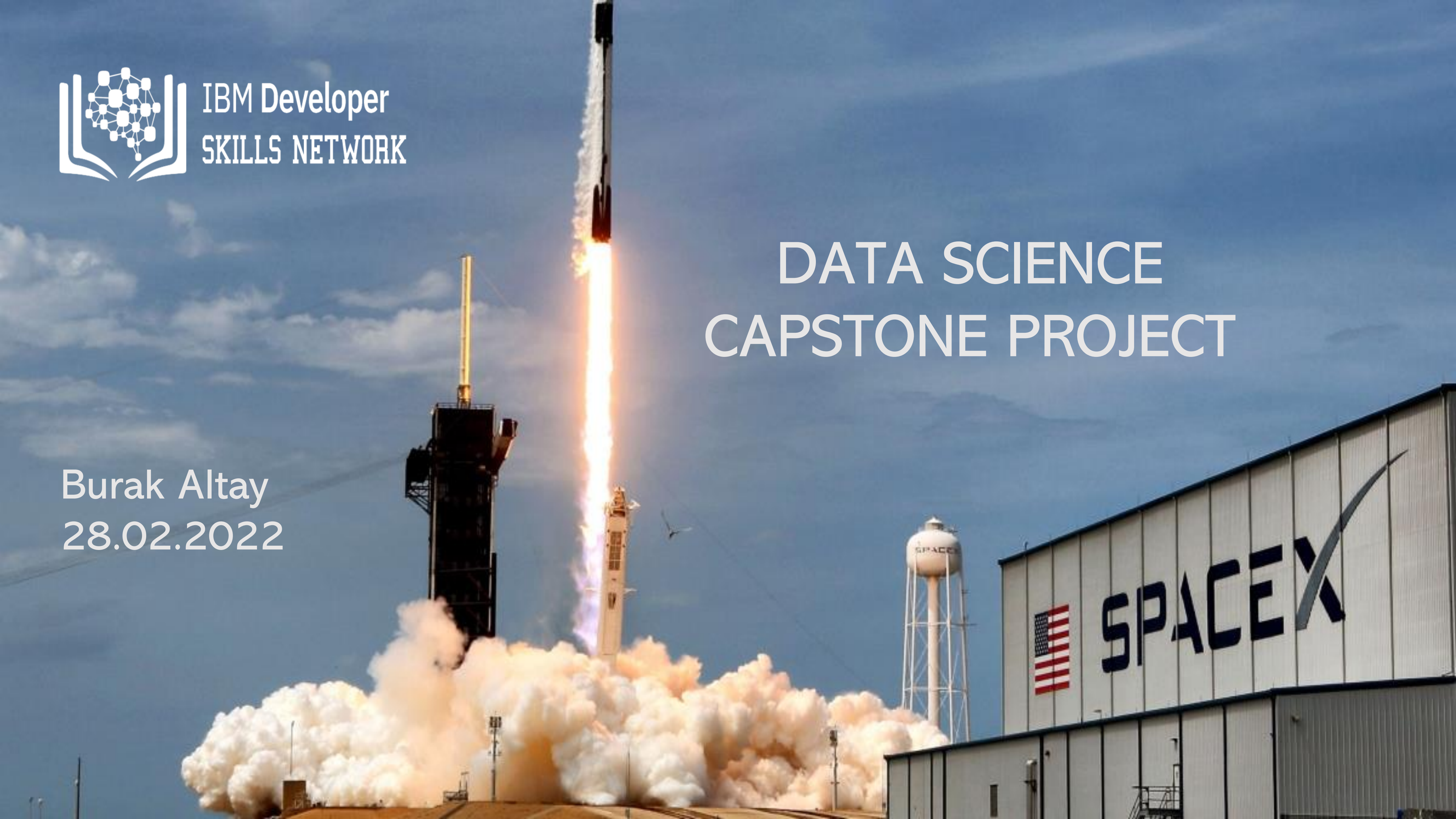
IBM Developer
SKILLS NETWORK

DATA SCIENCE CAPSTONE PROJECT

Burak Altay
28.02.2022



Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection
  - Data Wrangling
  - Exploratory Analysis Using SQL
  - Exploratory Analysis Using Pandas and Matplotlib
  - Interactive Visual Analytics and Dashboards
  - Predictive Analysis (Classification)
- **Summary of all results**
  - Data Collection and Wrangling Results
  - Exploratory Analysis Results
  - Data Visualization Results
  - Predictive Analysis Results

Introduction

Project background and context

- SpaceX is an American space transportation company headquartered in Hawthorne, California, USA and advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. In this lab, you will collect and make sure the data is in the correct format from an API. The following is an example of a successful and launch.

Problems you want to find answers

- What are the factors that determine whether a rocket will land successfully or not?
- Where are the best places to launch rockets?
- What is the best algorithm to predict successful launches?

Methodology

Section 1

Methodology

Executive Summary

Data collection methodology:

- **SpaceX Rest API**
- **Web Scrapping from Wikipedia**

Perform data wrangling

- **Data summary and visualization**
- **Dealing with missing data**
- **Using one hot coding and feature selection**

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

Data Collection

- Data was collected through API requests made from the SpaceX REST API and Web Scraping in a table in SpaceX's Wikipedia entry.

SpaceX API:

1. Getting response from API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

2. Decoding the response content as a Json and turning it into a dataframe

```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

3. Cleaning data with custom functions

```
# Call getBoosterVersion
getBoosterVersion(data)
```

```
# Call getLaunchSite
getLaunchSite(data)
```

```
# Call getPayloadData
getPayloadData(data)
```

```
# Call getCoreData
getCoreData(data)
```

4. Combining data in a dictionary

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

5. Creating a dataframe from the dictionary

```
# Create a data from launch_dict
data = pd.DataFrame(launch_dict)
```

6. Filtering dataframe

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

7. Exporting dataframe to .CSV

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

GitHub: Data Collection API

Web scraping:

1. Requesting data from Wikipedia

```
response = requests.get(static_url)
```

2. Creating BeautifulSoup Object from response

```
soup = BeautifulSoup(response.content,'html.parser')
```

3. Extracting column/variable names

```
findall = soup.find_all('table')
html_tables = list(findall)
html_tables
```

```
column_names = []
for i in first_launch_table.find_all('th'):
    name = extract_column_from_header(i)
    if name is not None and len(name) > 0:
        column_names.append(name)
```

4. Creating a dictionary with keys from the extracted column names

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

5. Appending data into keys

```
extracted_row = 0
#Extract each table
for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders collapsible")):
   # get table row
    for rows in table.find_all("tr"):
        #check to see if first table heading is as number corresponding to launch a number
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.string.strip()
                flag=flight_number.isdigit()
```

6. Creating dataframe from dictionary

```
df=pd.DataFrame(launch_dict)
```

7. Exporting dataframe to .CSV

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

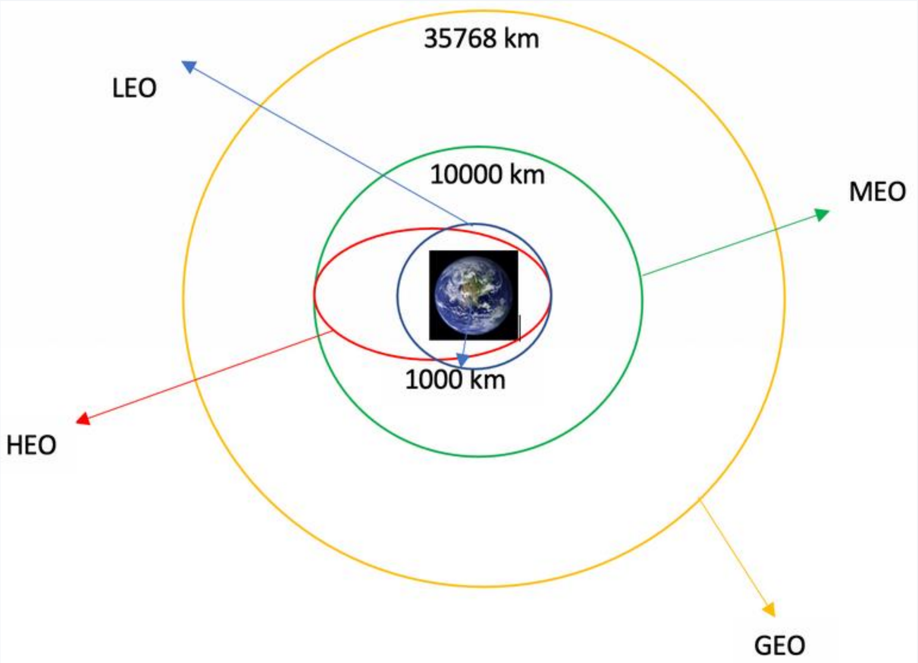
[GitHub: Web Scraping]

Data Wrangling

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; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

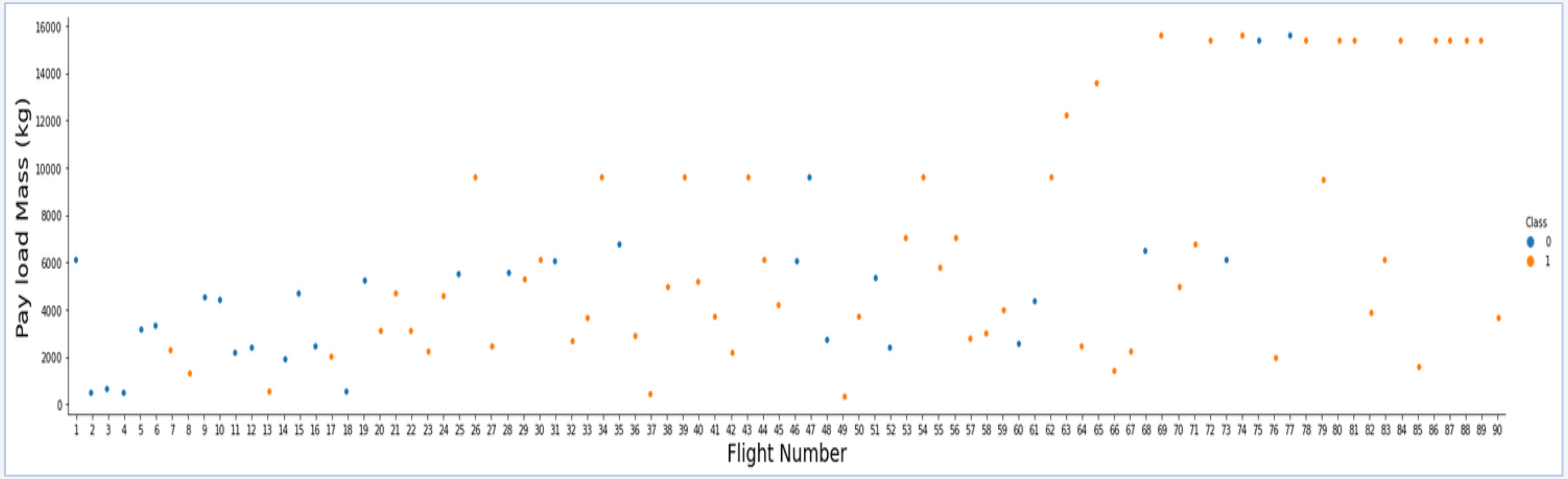
- Perform EDA on dataset
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Work out success rate for every landing in dataset
- Export data to .CSV



[GitHub: Data Wrangling]

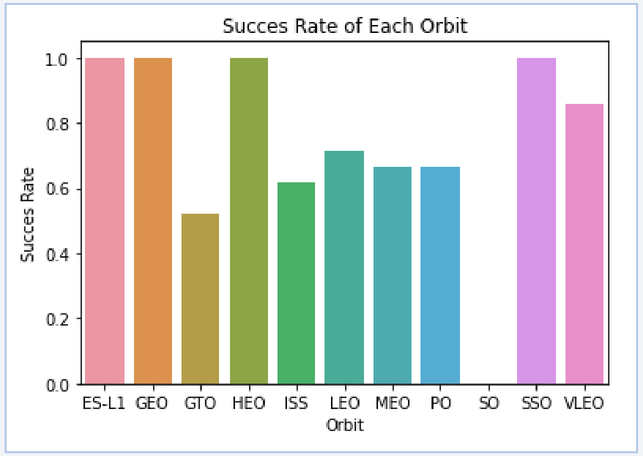
EDA with Data Visualization

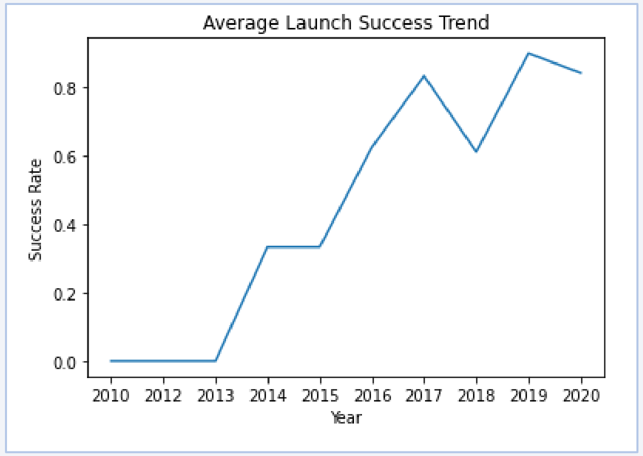
Scatterplots were used to visualize the relationship between features, a bar plot was used to visualize the relationship between the success rate of each orbit type. and a line chart was used to visualize the yearly trend of launch success. .



Scatter plots for;

- ✓ Flight Number VS. Payload Mass
- ✓ Flight Number VS. Launch Site
- ✓ Payload VS. Launch Site
- ✓ Orbit VS. Flight Number
- ✓ Payload VS. Orbit Type
- ✓ Orbit VS. Payload Mass





[GitHub: EDA with Data Visualization]

EDA with SQL

The following SQL queries are executed to answer the assignment questions;

- Displaying 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 names 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 in descending order,

```
%sql select UNIQUE launch_site from SPACEXTBL;
```

```
 * ibm_db_sa://trl47770:***@b1bc1829-6f45-4cd4-b
Done.
```

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

GitHub: EDA with SQL

Build an Interactive Map with Folium

We performed more interactive visual analyzes of the launch data using Folium.

For visual analysis, markers, circles, lines, and marker clusters were used with folium maps .

- ✓ First, we added each site's location to a map using the site's latitude and longitude coordinates,
- ✓ Next, we tried to improve the map by adding the launch results for each site and looked at which sites had high success rates.
- ✓ Next, we explored and analyzed the proximity of the launch sites.
- ✓ For this, we first add a MousePosition to the map to get the coordinate of the mouse over a point on the map. So we were able to easily find the coordinates of any point of interest (such as a railway) while exploring the map.
- ✓ After drawing distance lines to proximity, we were able to more easily answer questions such as "Are the launch sites close to the coastline or city centers?"

[GitHub:Build an Interactive Map with Folium]

Build a Dashboard with Plotly Dash

We created a Plotly Dash application for them to perform interactive visual analytics on SpaceX launch data in real time.

This built-in dashboard applet includes input components such as a dropdown list and a range slider to interact with the pie chart and scatter point chart.

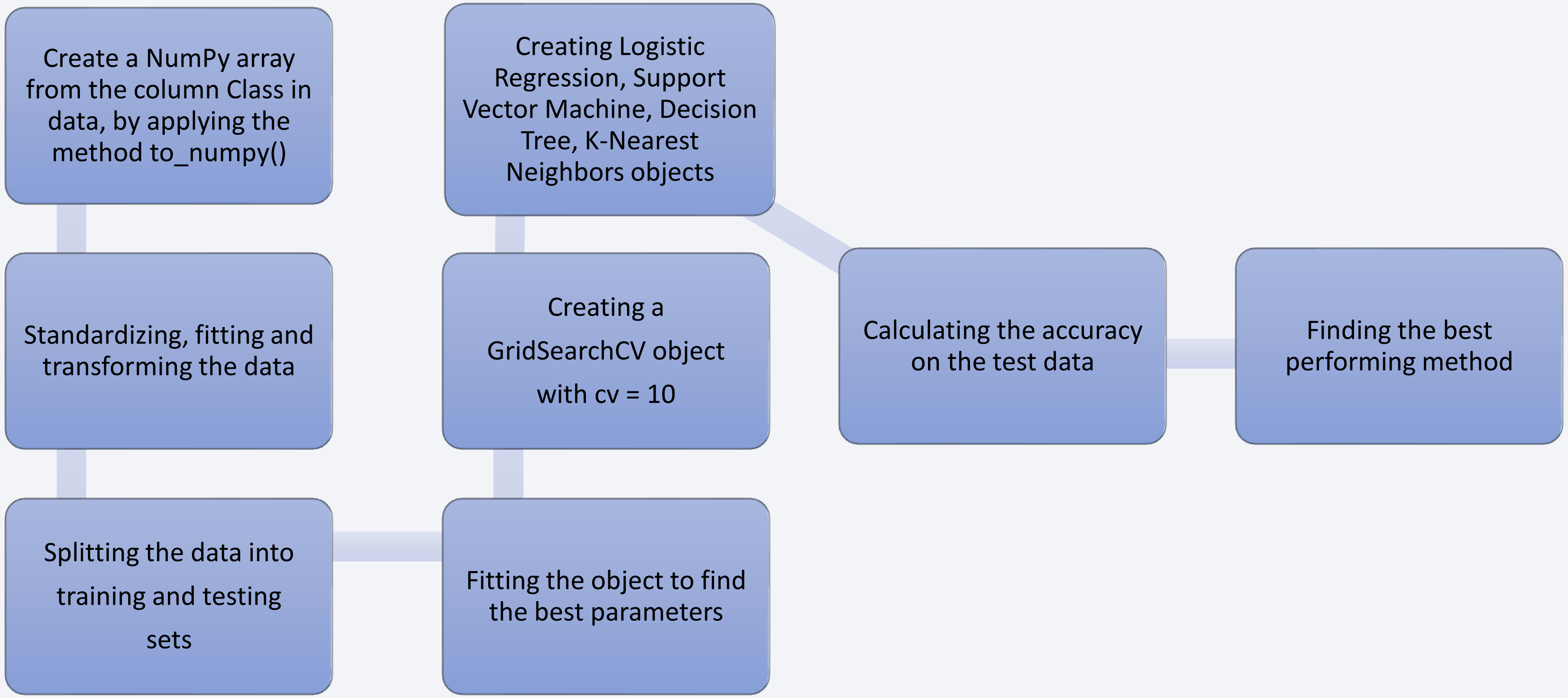
After visual analysis using the dashboard, necessary information was obtained, and answers were sought to the following questions.

- Which site has the biggest successful launches?
- Which site has the highest launch success rate?
- Which load range(s) have the highest startup success rate?
- Which load range(s) have the lowest startup success rate?
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest startup success rate?

GitHub: Build a Dashboard with Plotly Dash

Predictive Analysis (Classification)

We used machine learning to determine if Falcon 9's first stage would land successfully. The data was split into training data and test data to find the best Hyperparameter for SVM, Classification Trees and Logistic Regression. Then find the best performing method using the test data



GitHub: Predictive Analysis

Results

Exploratory data analysis results

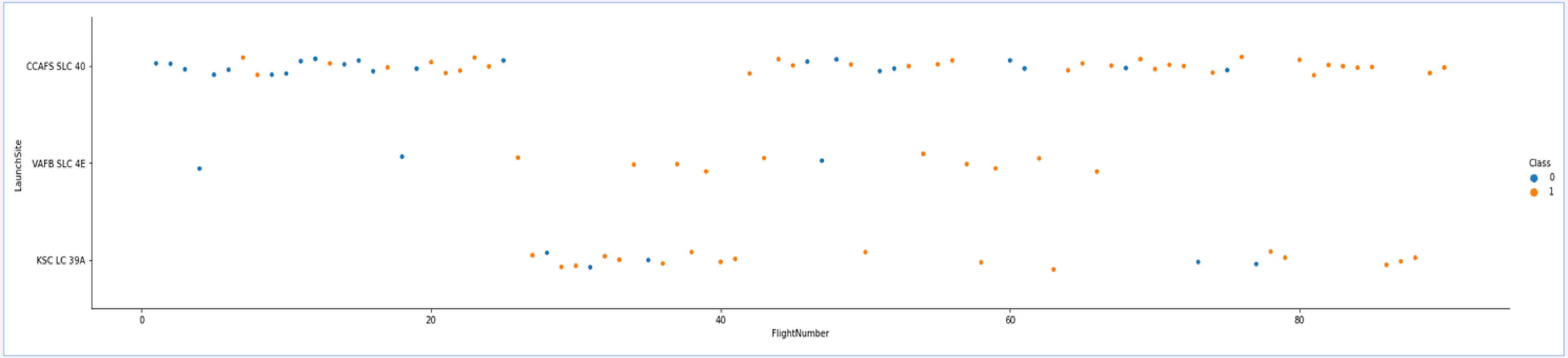
Interactive analytics demo in screenshots

Predictive analysis results

EDA with Visualization

Section 2

Flight Number vs. Launch Site

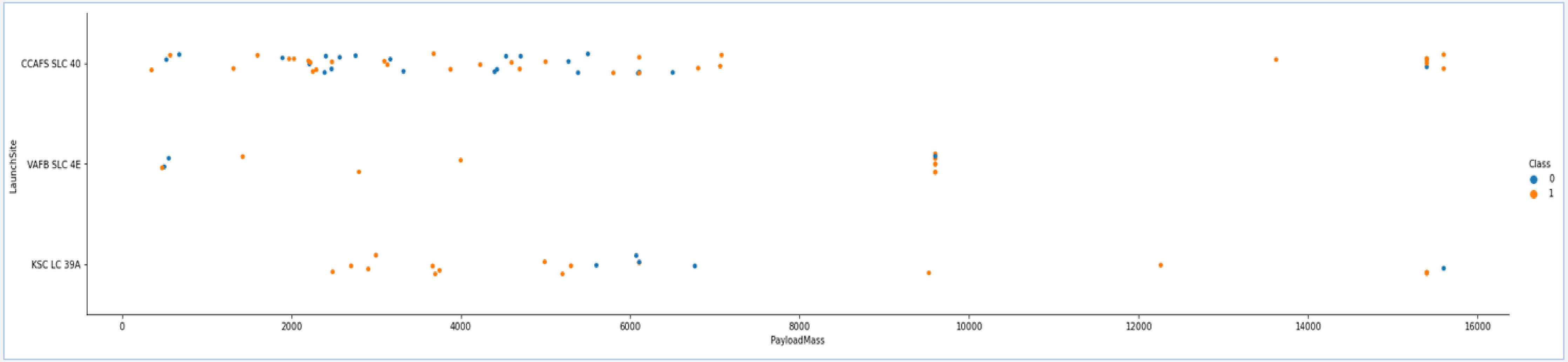


All last 5 launches were successful at all launch sites.

The overall success rate has increased over time.

The last launches are mostly successful in "CCAF5 SLC 40", which is the most used launch area.

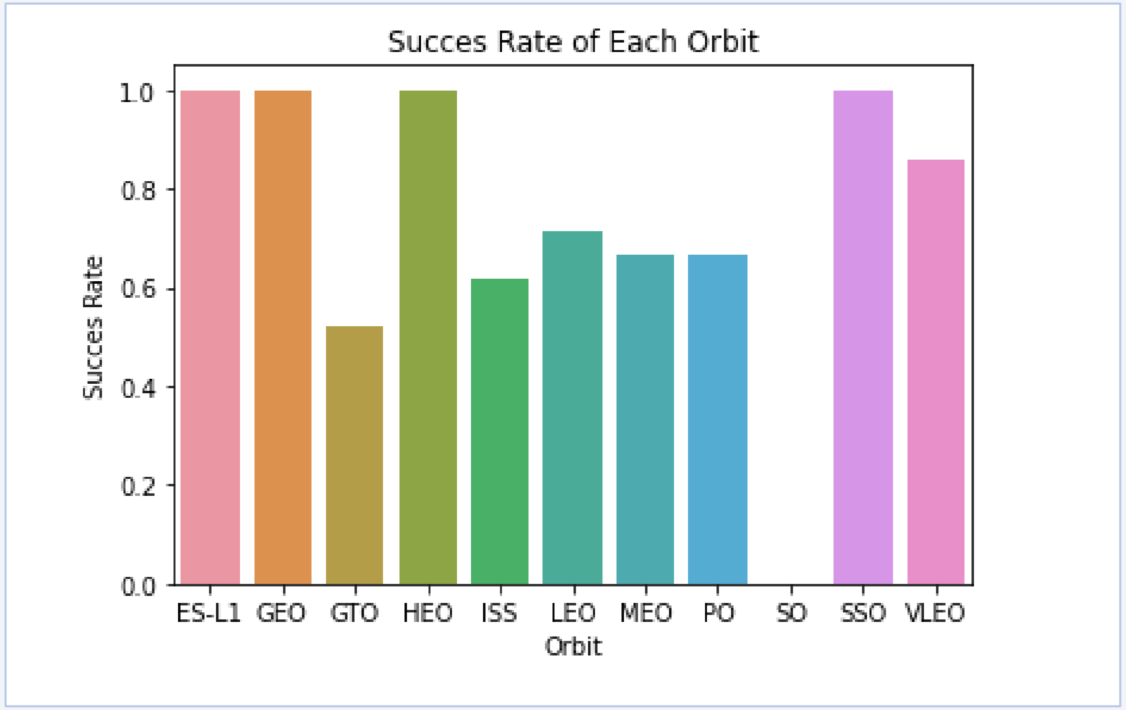
Payload vs. Launch Site



Most launches were made with a payload mass of up to 7000 kg.

Launches with a payload mass of more than 7000 kg are mostly successful.

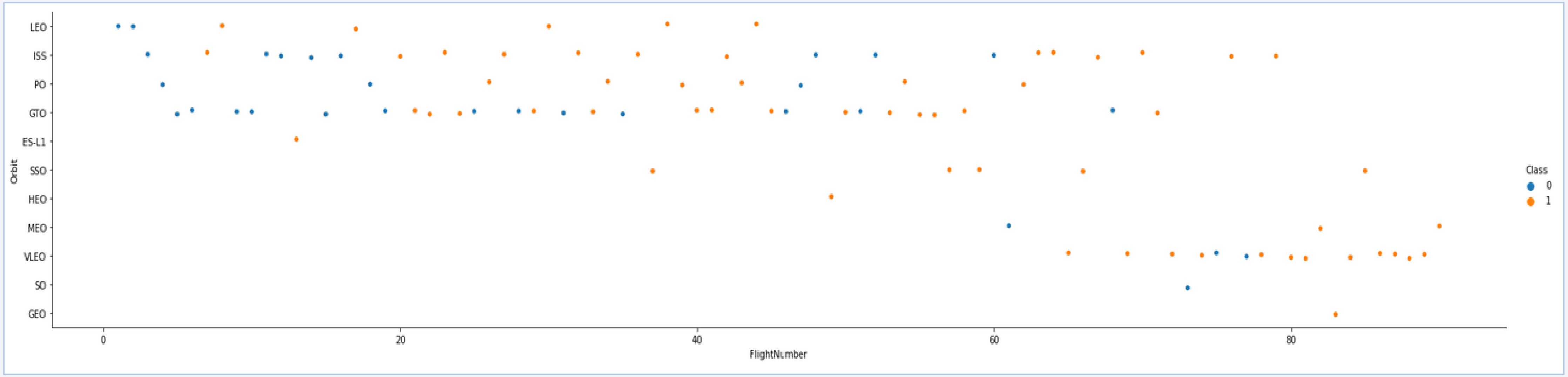
Success Rate vs. Orbit Type



While the success rate is 0 in SO orbital, the success rate is 100% in ES-L1, GEO, HEO, SSO orbits.

In other orbits, the success rate is between 50% and 85%.

Flight Number vs. Orbit Type

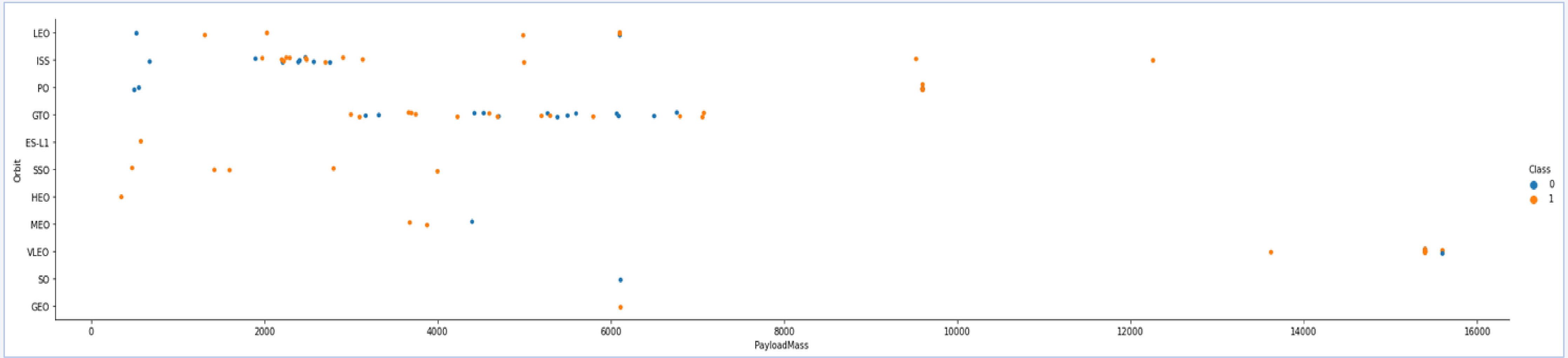


It is observed that the success rate increases over time in all trajectories.

Recently, the VLEO orbit has been preferred more than other orbits.

While the first two flights were unsuccessful in LEO orbit, all other flights were successful, while in ISS and GTO orbits, unsuccessful flights were encountered from time to time.

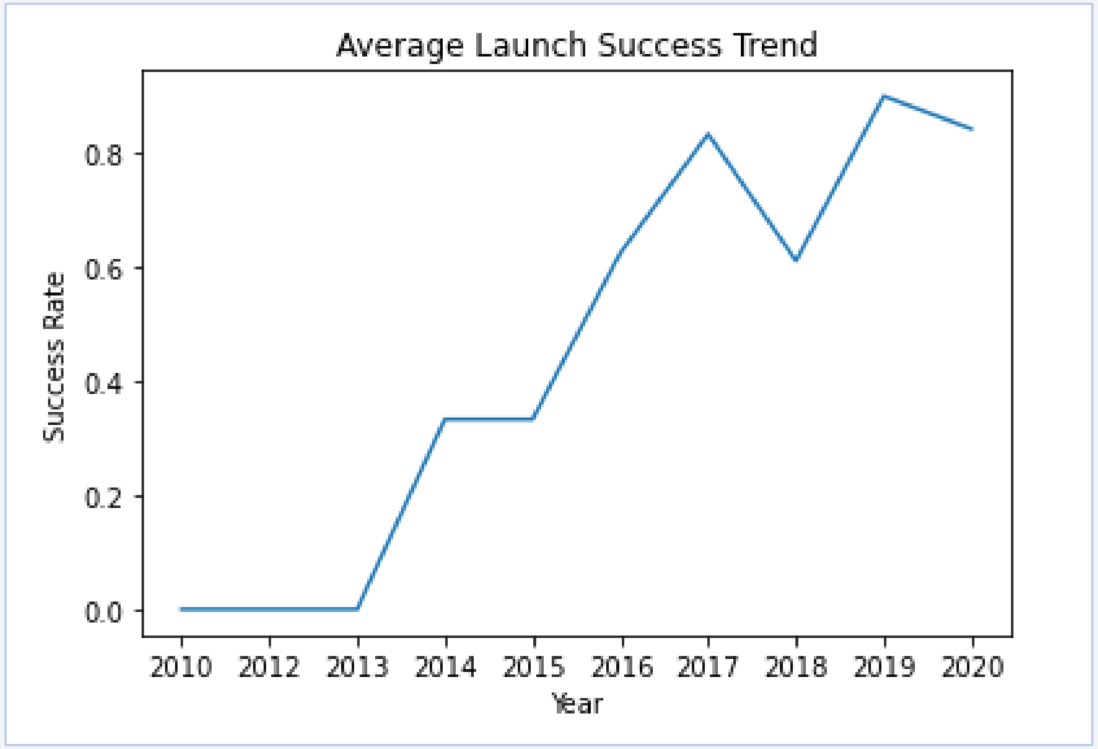
Payload vs. Orbit Type



While there is no relationship between the payload to the GTO orbit and the success rate, all payloads heavier than 3000 KG on the ISS are successful.

SSO orbit is preferred only for payloads of 4000 kg and below.

Launch Success Yearly Trend



At the end of the first three-year period, it started to increase as of 2013 and the successful flight rate continued to increase.

EDA

with

SQL

Section 3

All Launch Site Names

```
%sql select UNIQUE launch_site from SPACEXTBL;
```

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

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

✓ According to data, there are four launch sites.

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 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

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

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as Total_PM from SPACEXTBL where customer = 'NASA (CRS)';
```

total_pm
45596

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

✓ The total payload mass carried by boosters is **45596** kg.

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as AVG_PM from SPACEXTBL where booster_version like 'F9 v1.1%';
```

avg_pm
2534

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

✓ The average payload mass carried by the booster version F9 v1.1 is **2534** kg.

First Successful Ground Landing Date

```
%sql select min(DATE) as min_date from SPACEXTBL where mission_outcome = 'Success' and landing__outcome like '%pad%';
```

min_date
2015-12-22

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

✓ The first successful landing result was obtained on 22/12/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql select booster_version,payload_mass__kg_ as pmkg,mission_outcome,landing__outcome
from SPACEXTBL where landing__outcome = 'Success (drone ship)' and payload_mass__kg_  between 4000 and 6000;
```

booster_version	pmkg	mission_outcome	landing__outcome
F9 FT B1022	4696	Success	Success (drone ship)
F9 FT B1026	4600	Success	Success (drone ship)
F9 FT B1021.2	5300	Success	Success (drone ship)
F9 FT B1031.2	5200	Success	Success (drone ship)

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

```
%sql select mission_outcome,count(*) from SPACEXTBL group by mission_outcome;
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);
```

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

Listing the names of the booster_versions which have carried the maximum payload mass. Use a subquery

2015 Launch Records

```
%sql select landing__outcome, booster_version, launch_site,DATE from SPACEXTBL where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

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

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

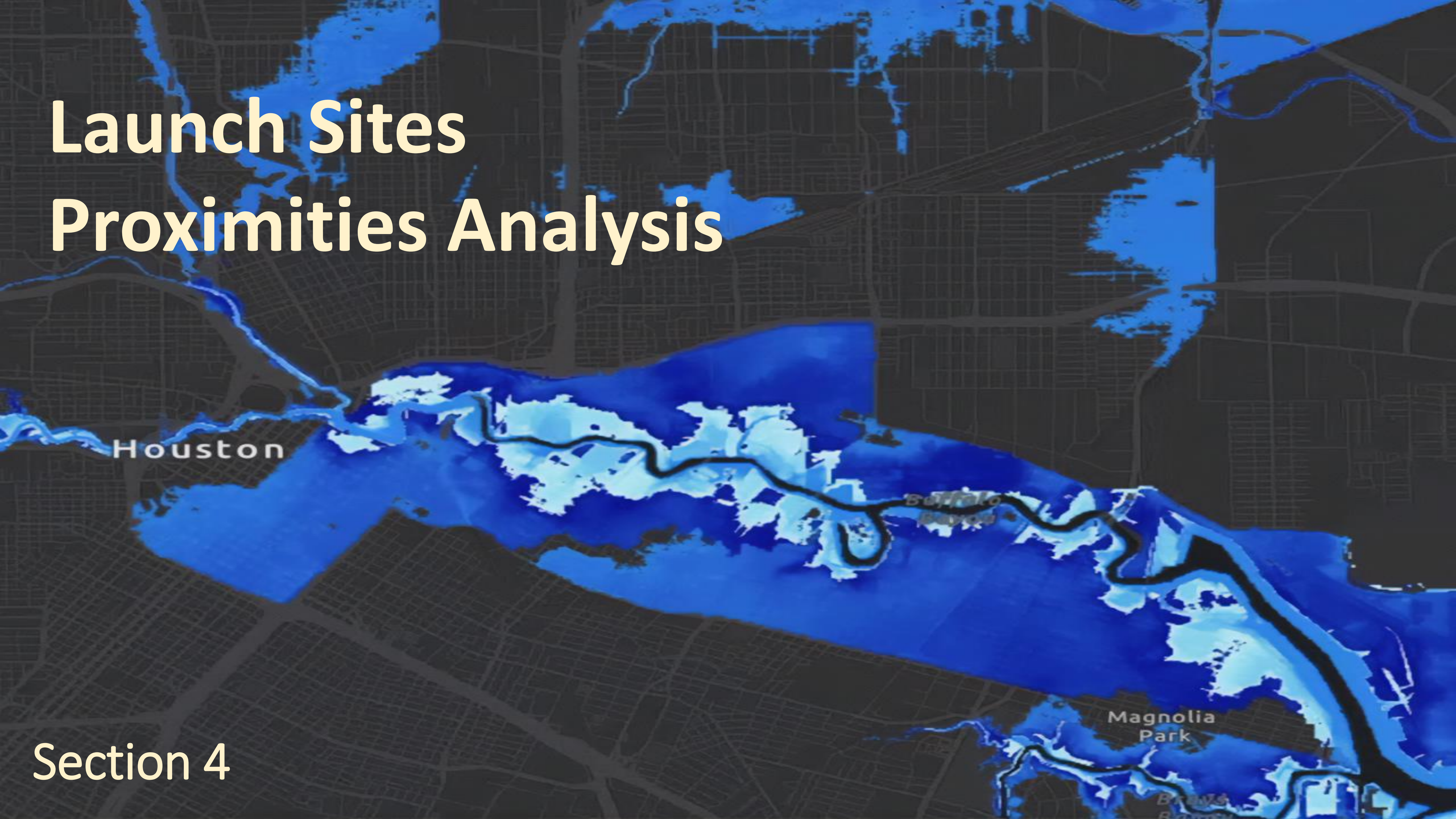
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

```
%%sql select landing__outcome, count(*) as count_oc
from SPACEXTBL
where date between '2010-06-04' and '2017-03-20'
group by landing__outcome
order by count_oc desc;
```

landing__outcome	count_oc
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

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.

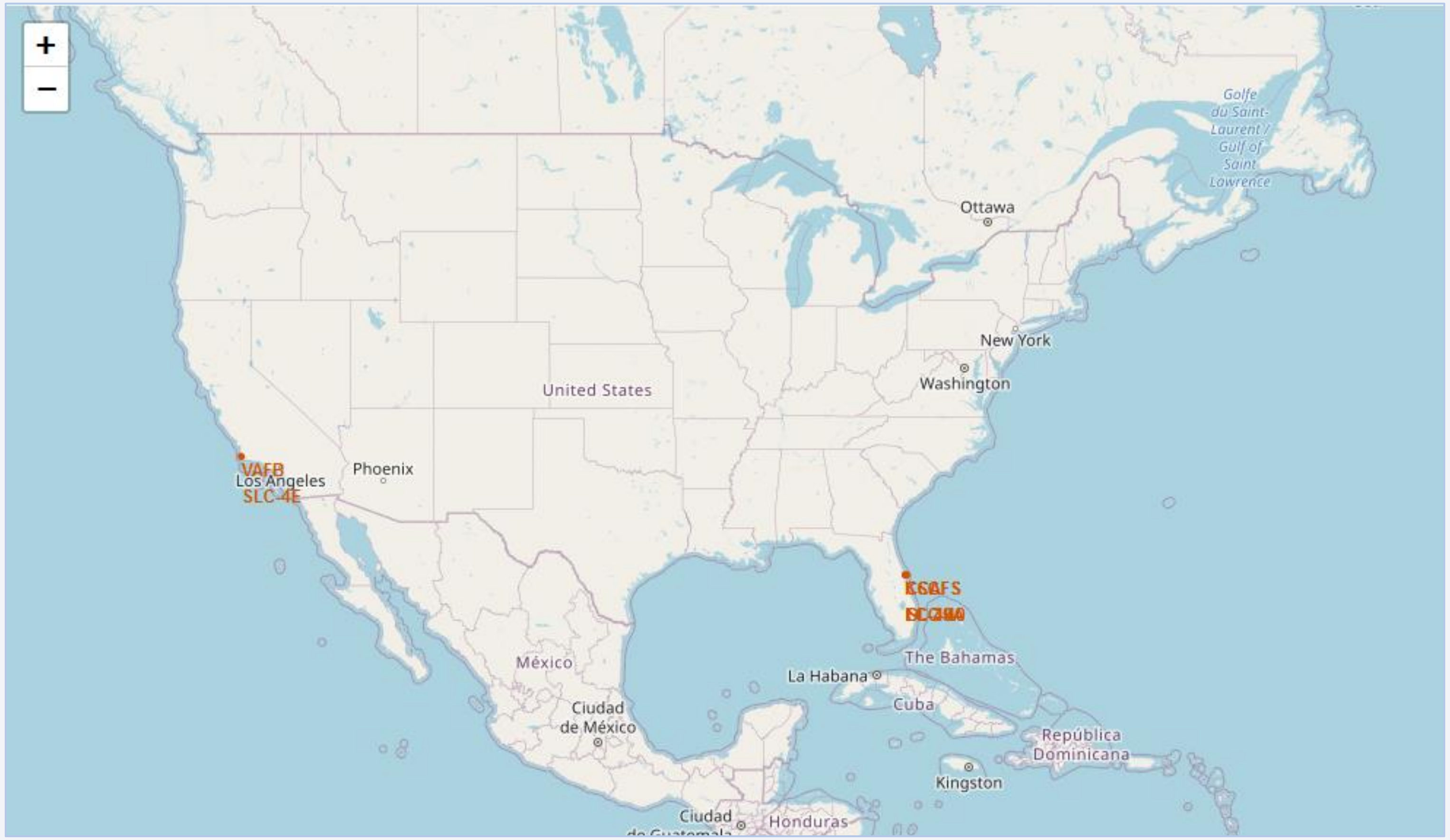
Launch Sites Proximities Analysis



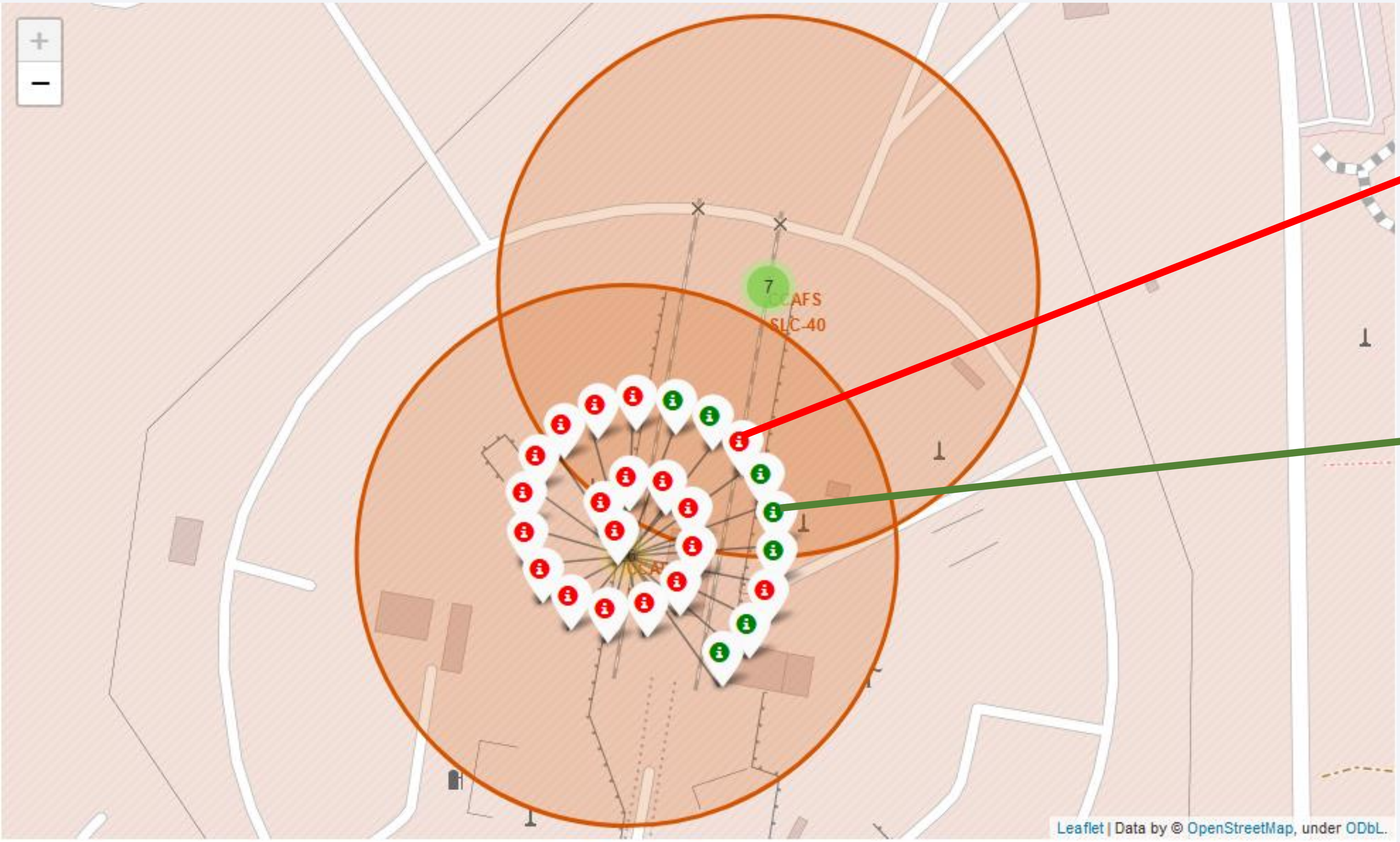
Section 4

Marking all launch sites on a map

SpaceX launch sites are very close to the equator line and the coast.



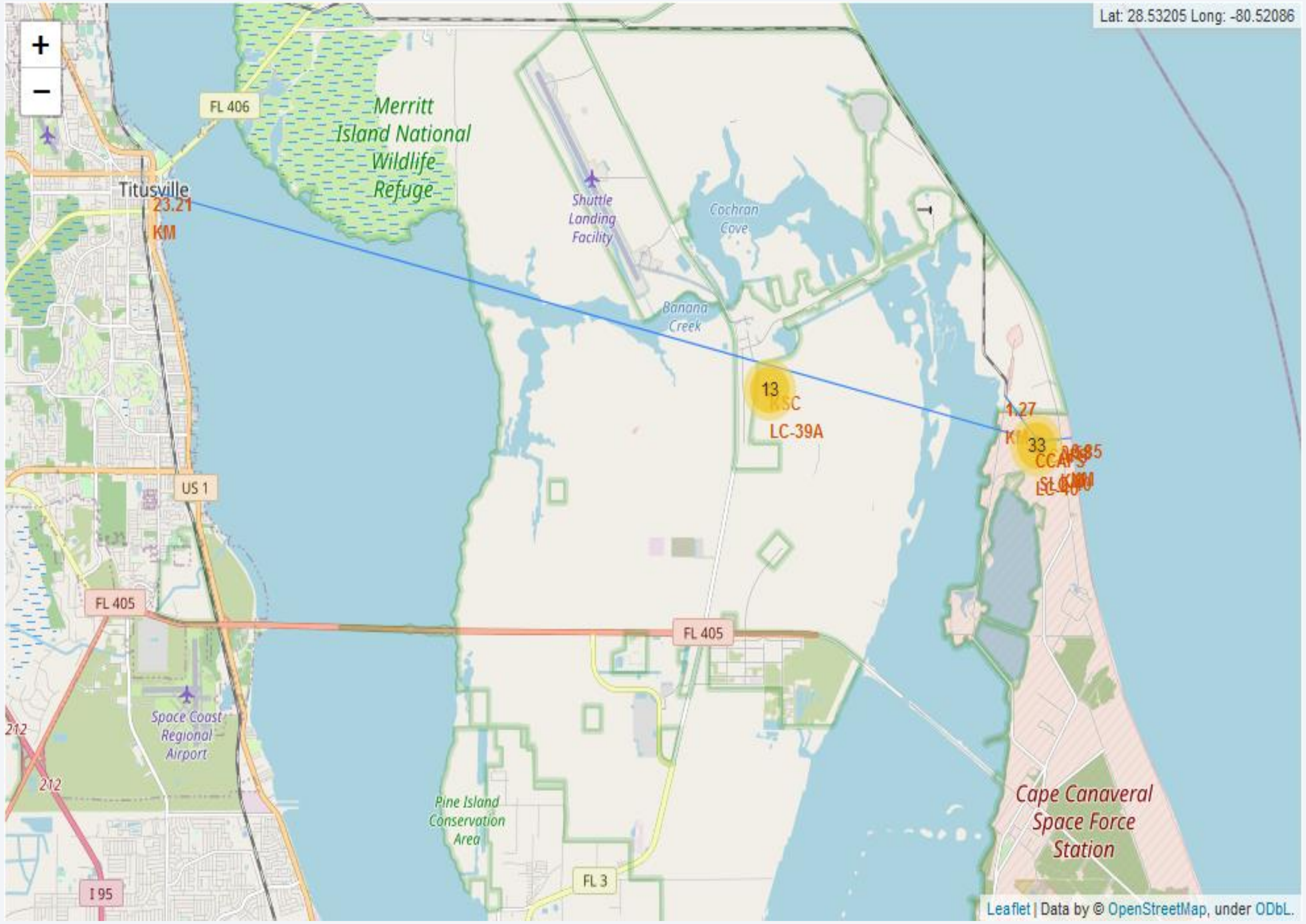
Marking the success/failed launches for each site on the map



Red markers shows failed launchs

Green markers shows successful launchs

Calculate the distances between a launch site (CCAFS SLC-40) to its proximities



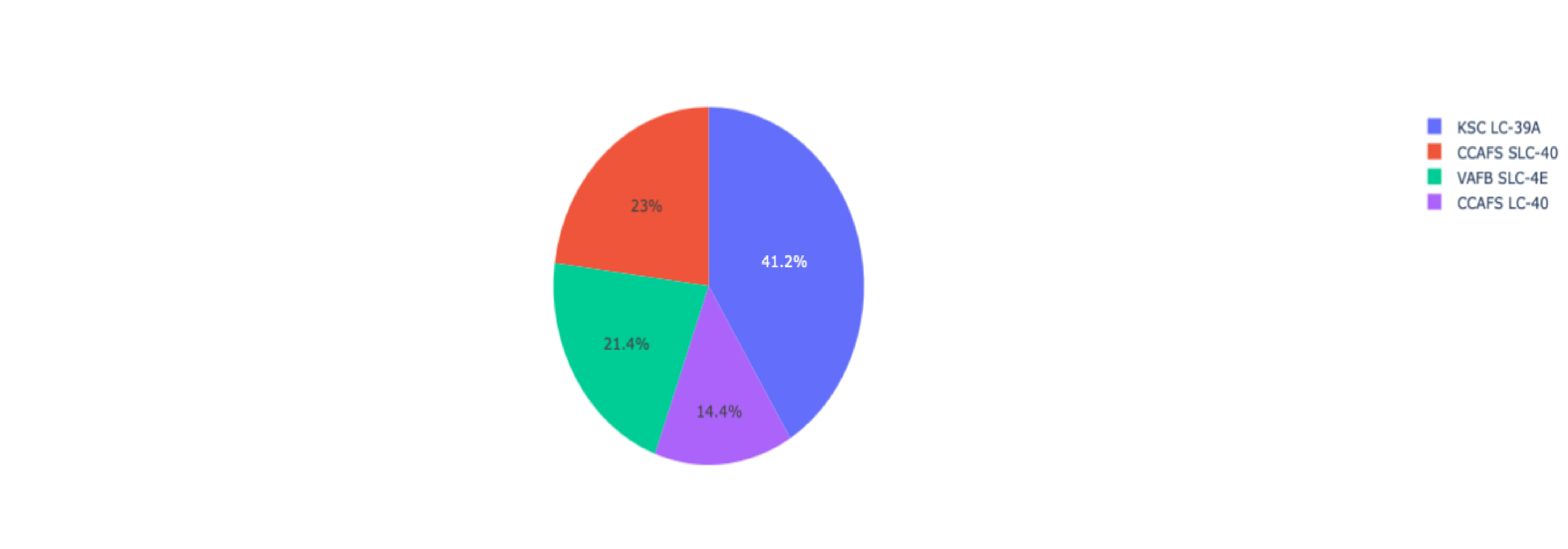
- distance to railway: 1.27 km
- distance from the highway: 0.58 km
- distance from the coastline: 0.85 km
- distance to nearest city center: 23.21 km

While the launch area is very close to the highway, railway and the beach, the nearest city center is outside of the danger zone.

Build a Dashboard with Plotly Dash

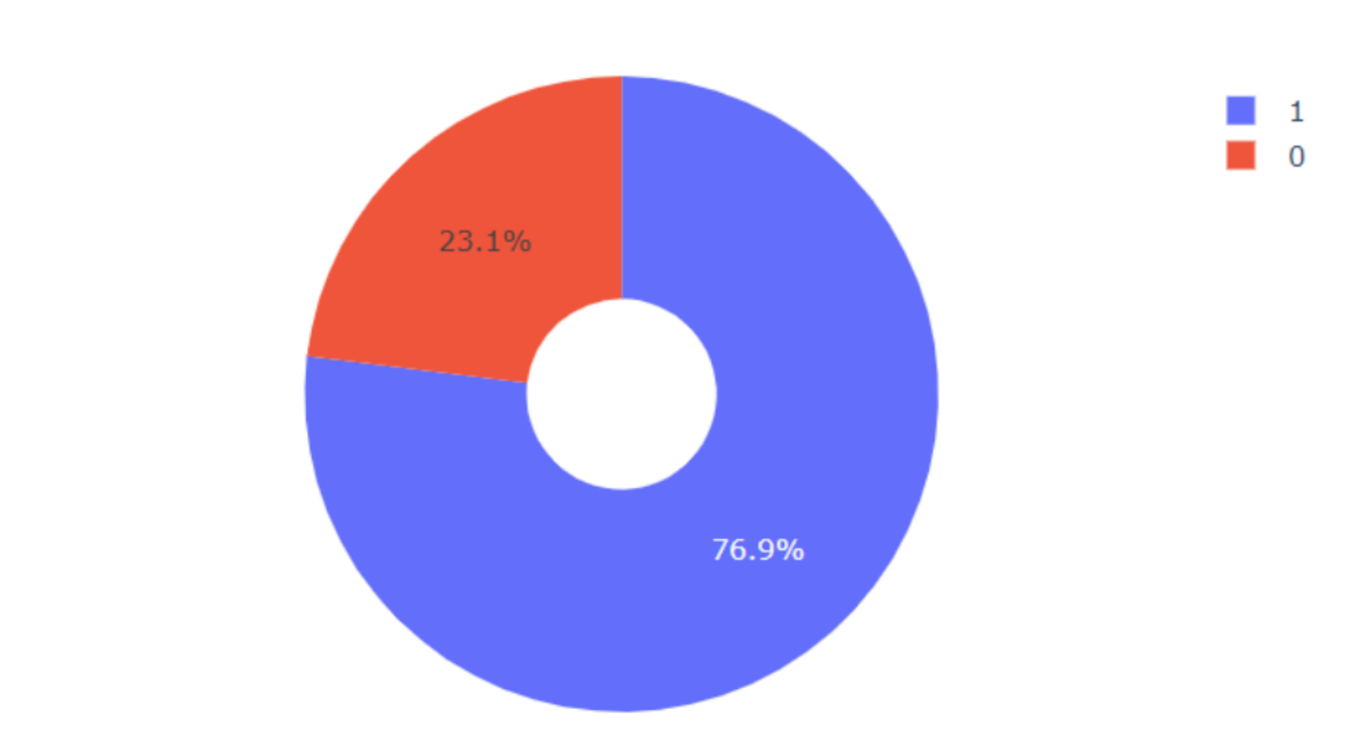
Section 5

Total Success Launches for All Sites



In terms of the place where the launches are made, the "KSC LC-39A" launch area is more successful than other areas.

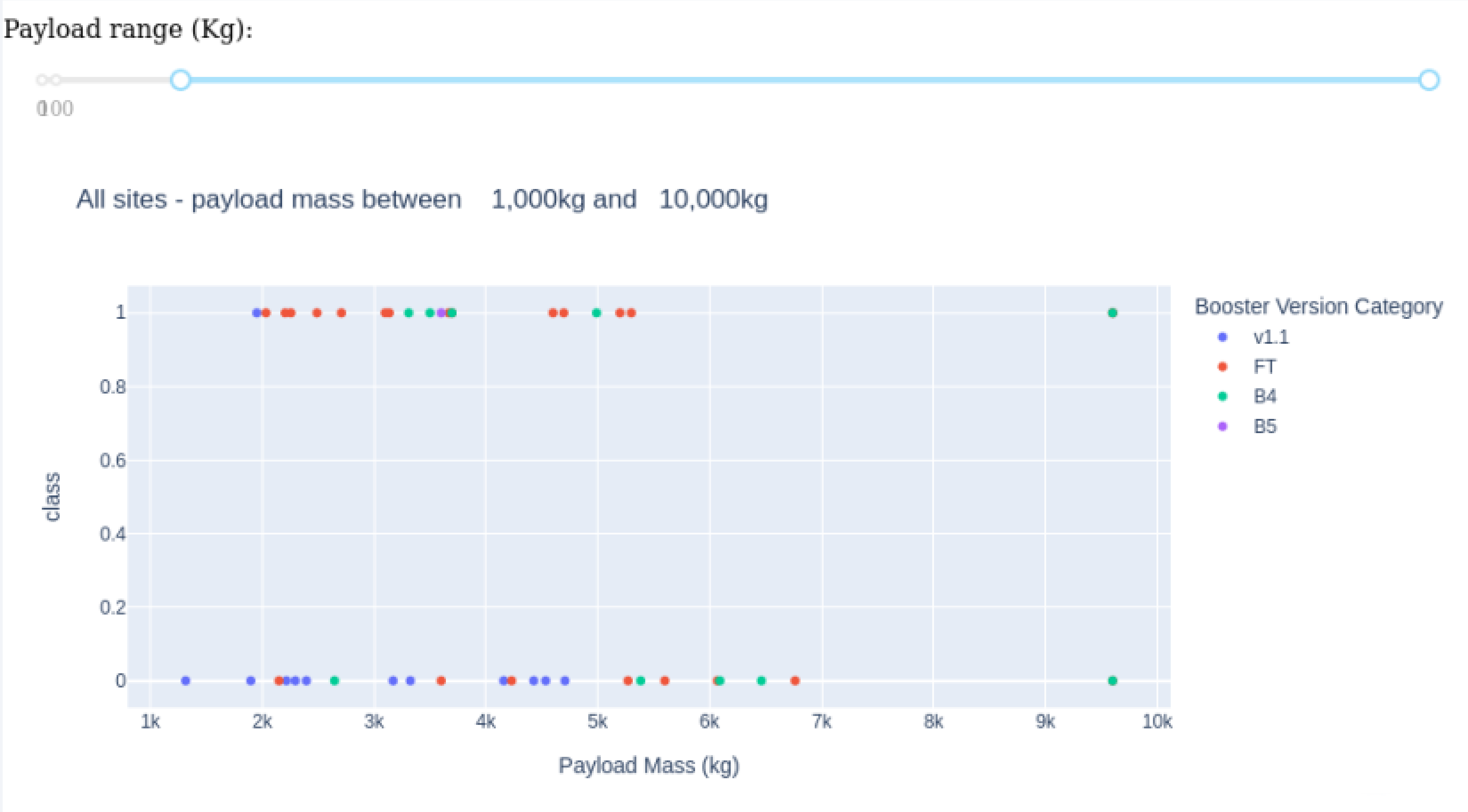
Total Success Launches for KSC LC-39A



In the "KSC LC-39A" launch site, the total failure rate is 23.1%, while the total success rate is 76.9%.

Payload Mass vs Success for All Sites

The highest success rate is between 2500 and 5500 kg payload and FT boosters.



Predictive Analysis

(Classification)

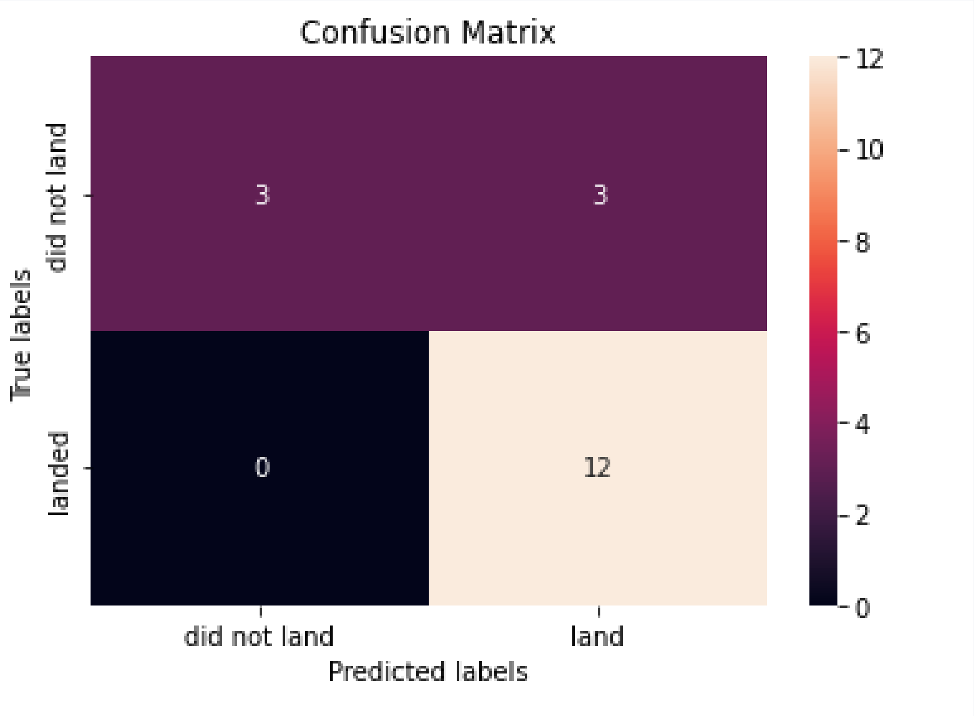
Section 6

Classification Accuracy

Four classification models were tested, and their accuracy calculated.

As a result, the best score was obtained with 87.67% Decision Tree classifier.

Confusion Matrix



By examining the confusion matrix, the ability of logistic regression to distinguish between different classes can be observed.

Conclusions

- Most of the launch sites are located close to the Equatorial line, the coast, the railroad and the highway, while being within a safe distance from the city center,
- The KSC LC-39A launch site has the highest success rate,
- It has been observed that launches with low payload mass are more successful than launches with larger payload mass,
- While the success rate is 0 in SO orbital, the success rate is 100% in ES-L1, GEO, HEO, SSO orbits,
- The rate of successful launches increases over the years and as the number of flights increases,
- Decision Tree Classifier was determined as the best algorithm for the data set.

Thank You!