IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Csilla Cecília Takács>
<17.03.2023.>

Outline

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

Executive Summary

For performing this scientific work, we used

- data collection
- data wrangling
- exploratory data analysis with data visualization and SQL
- building an interactive map with Folium
- building a dashboard with Plotly Dash
- predictive analysis

Summary of all results

- we opted for the optimal model and visualization to help decision making

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, while other providers cost upward to 165 million dollars each
- Falcon 9 is a reusable, two-stage rocket designed and manufactured by SpaceX for the reliable and safe transport of people and payloads into Earth orbit and beyond. It is the world's first orbital class reusable rocket.
- Reusability allows SpaceX to re-fly the most expensive parts of the rocket, which in turn drives down the cost of space access
- We have used data science methodology to predict
  1. if the Falcon 9 first stage will land successfully
  2. what factors determine the success rate of the landing
  3. the interaction amongst various features that determine the success rate
  4. the operating conditions needed to ensure a successful landing program

Section 1

Methodology

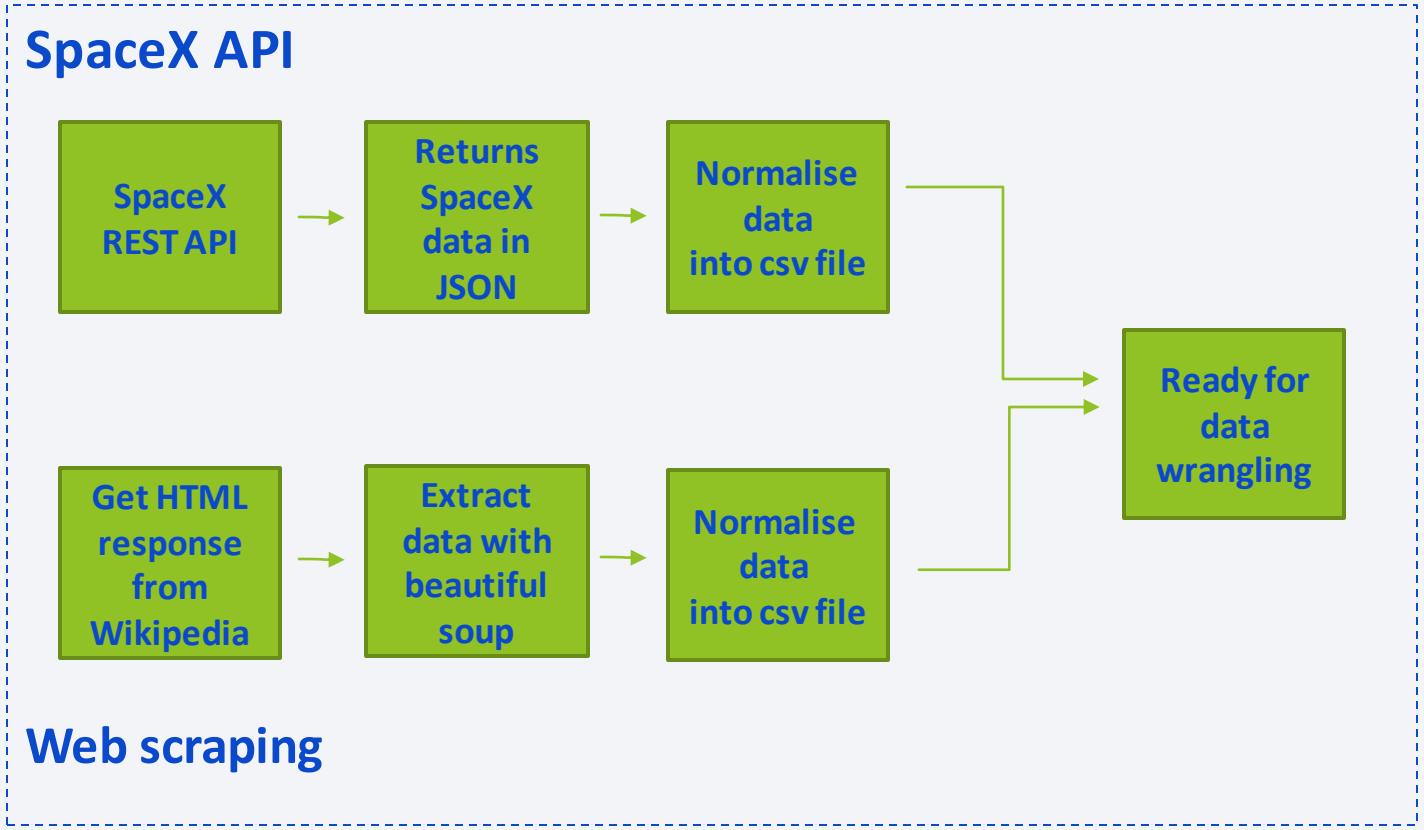
Methodology

Executive Summary

- Data collection methodology
  - SpaceX REST API and web scraping from Wikipedia
- Data wrangling
  - One-hot encoding to the categorical features
  - Data cleaning of null values and irrelevant columns
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - K nearest number, logistic regression, support vector machine and decision tree

Data Collection

SpaceX launch data was gathered from the SpaceX REST API and from Wikipedia



Data Collection – SpaceX API

Getting response from the API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url).json()
```

Converting the response to a JSON file

```
response = requests.get(static_json_url).json()
data = pd.json_normalize(response)
```

Apply custom functions to clean data

```
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

```
getBoosterVersion(data)
```

Assign list to dictionary then to dataframe

```
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}
```

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

Filter dataframe then export to csv file

```
data_falcon9 = df.loc[df['BoosterVersion']!="Falcon 1"]
```

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

https://github.com/TCsillaC/IBM-Data-Science-Professional-Certificate---Capstone-Project/blob/e348c9fd439649c35c0fa4e473a5b6ef15e6d571/jupyter-labs-spacex-data-collection-api%20(3).ipynb

Data Collection - Scraping

1. Getting response from HTML

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

2. Creating Beautiful Soup object

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

3. Finding tables

```
html_tables = soup.find_all("table")
print(html_tables)
```

4. Getting column names

```
column_names = []

# Apply find_all() function with `th` element on
# Iterate each th element and apply the provided
# Append the Non-empty column name (`if name is n

temp = soup.find_all('th')
for x in range(len(temp)):
    try:
     name = extract_column_from_header(temp[x])
     if (name is not None and len(name) > 0):
        column_names.append(name)
    except:
     pass
```

5. Creation of a dictionary

```
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
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']=[]
```

6. Converting dictionary to dataframe

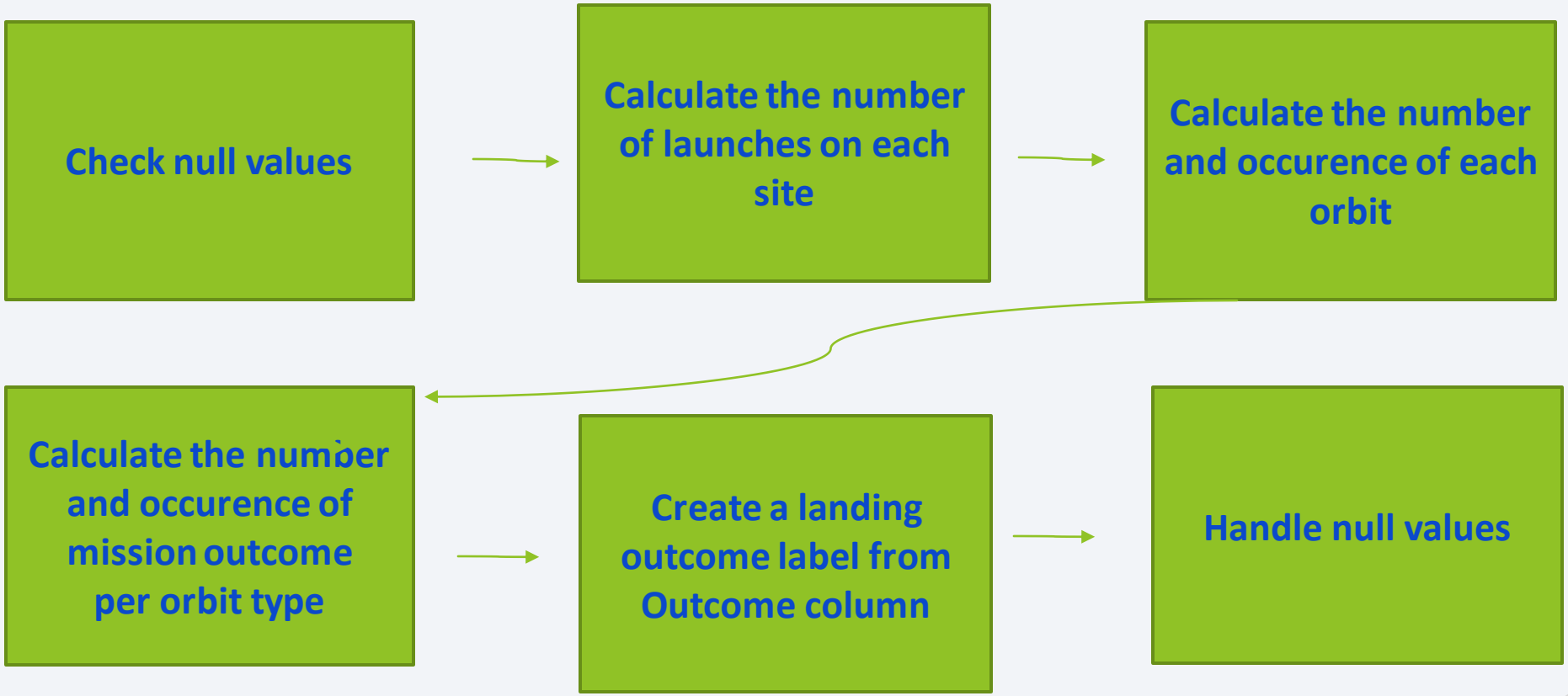
```
df=pd.DataFrame(launch_dict)
```

7. Saving dataframe to csv file

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

https://github.com/TCsillaC/IBM-Data-Science-Professional-Certificate---Capstone-Project/blob/e348c9fd439649c35c0fa4e473a5b6ef15e6d571/jupyter-labs-webscraping%20(1).ipynb

Data Wrangling

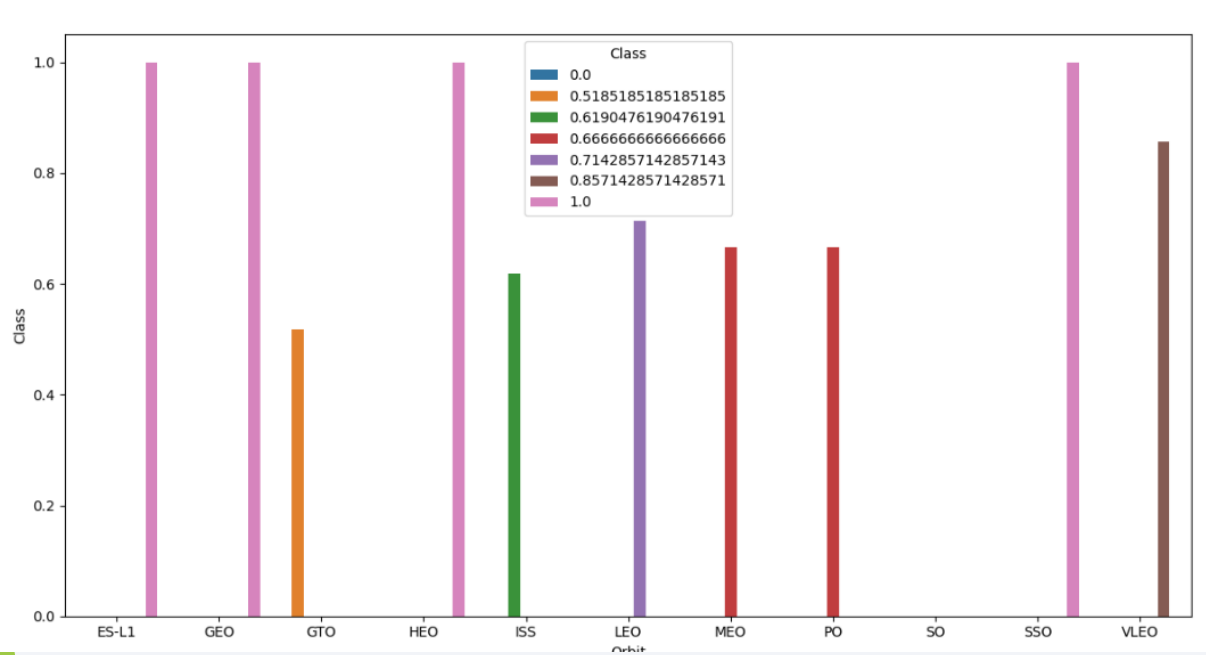


https://github.com/TCsillaC/IBM-Data-Science-Professional-Certificate---Capstone-Project/blob/e348c9fd439649c35c0fa4e473a5b6ef15e6d571/jupyter-labs-eda-sql-coursera_sqllite.ipynb

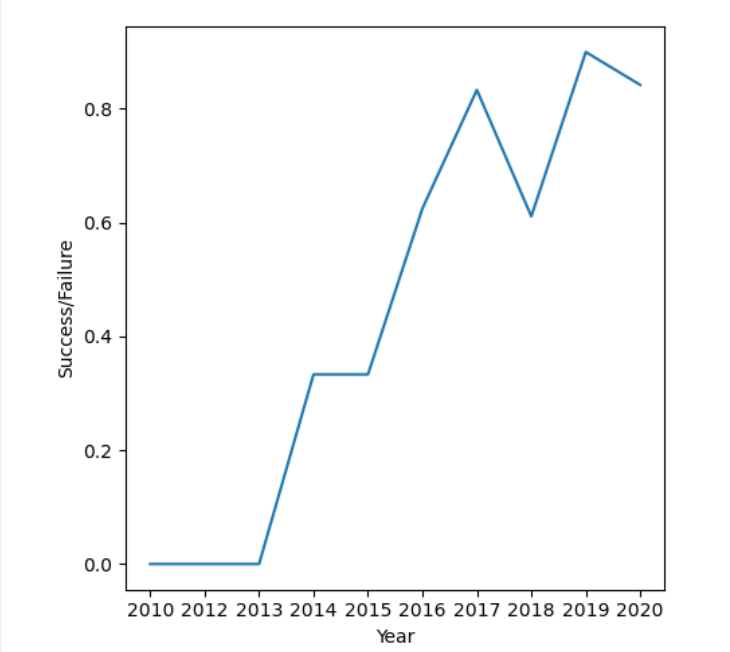
EDA with Data Visualization

We plotted bar charts, scatter plots, trend visualizations, interactive maps.

For example, we generated a bar chart to find out which orbits have the highest success rates:



We analyzed the yearly trend of the launch succes rate:



https://github.com/TCsillaC/IBM-Data-Science-Professional-Certificate-–Capstone-Project/blob/e348c9fd439649c35c0fa4e473a5b6ef15e6d571/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

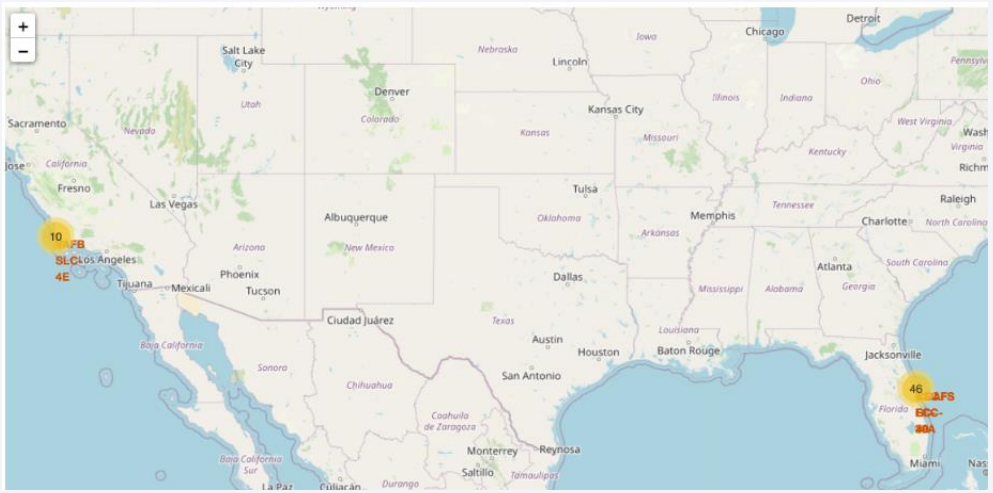
Exploratory Data Analysis with SQL

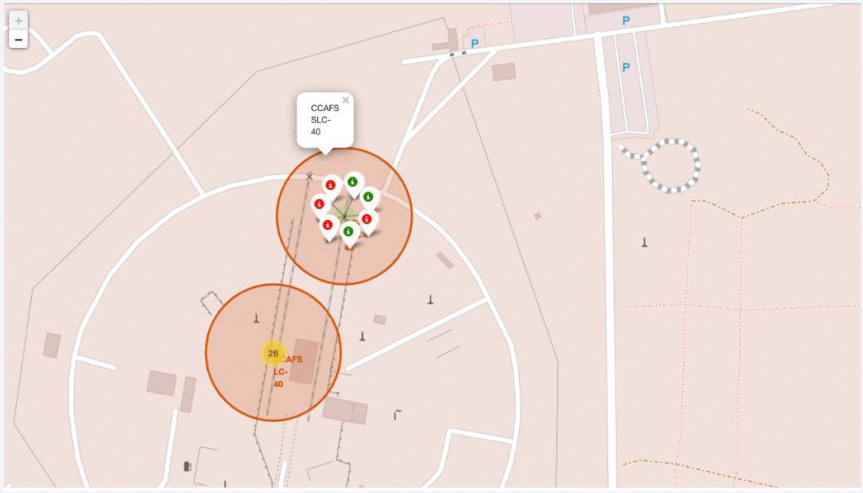
- We connected to the SpaceX launch database
- We displayed the names of the unique launch sites in the space mission
- We displayed the total payload mass carried by boosters launched by NASA
- We displayed the average payload mass carried by booster version F9 v1.1
- We listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- We listed the total number of successful and failure mission outcomes
- With the use of a subquery, we listed the names of the booster versions which have carried the maximum payload mass.

https://github.com/TCsillaC/IBM-Data-Science-Professional-Certificate---Capstone-Project/blob/e348c9fd439649c35c0fa4e473a5b6ef15e6d571/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.

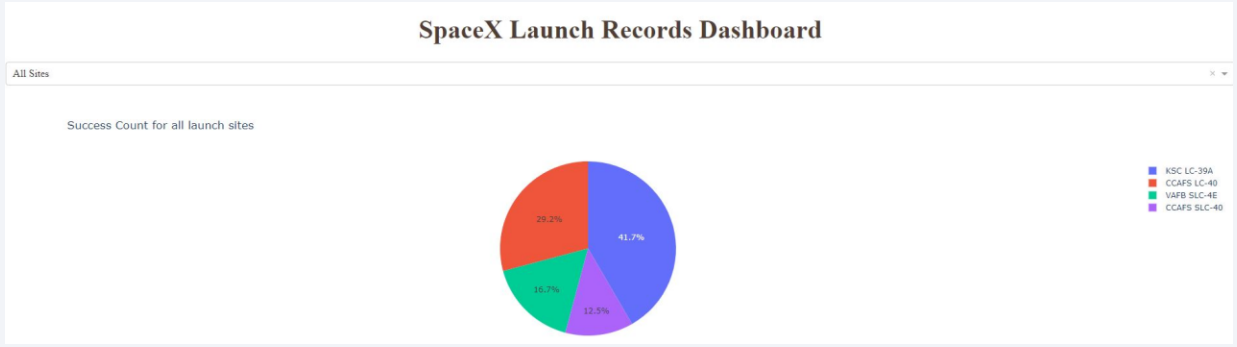




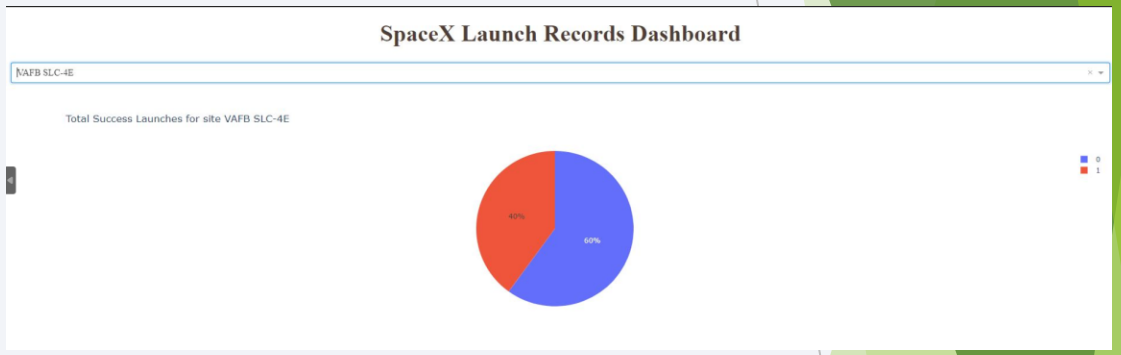
https://github.com/TCsillaC/IBM-Data-Science-Professional-Certificate---Capstone-Project/blob/e348c9fd439649c35c0fa4e473a5b6ef15e6d571/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

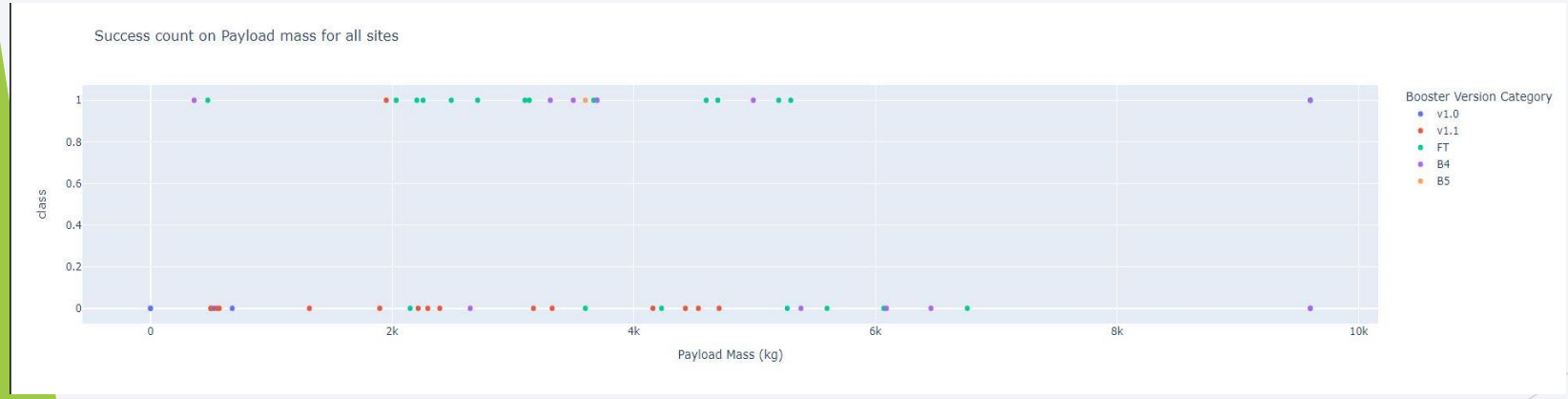
We built an interactive dashboard with Plotly dash



We plotted pie charts showing the launches by site



We plotted scatter graphs showing the relationship with outcome and payload mass for the different booster versions



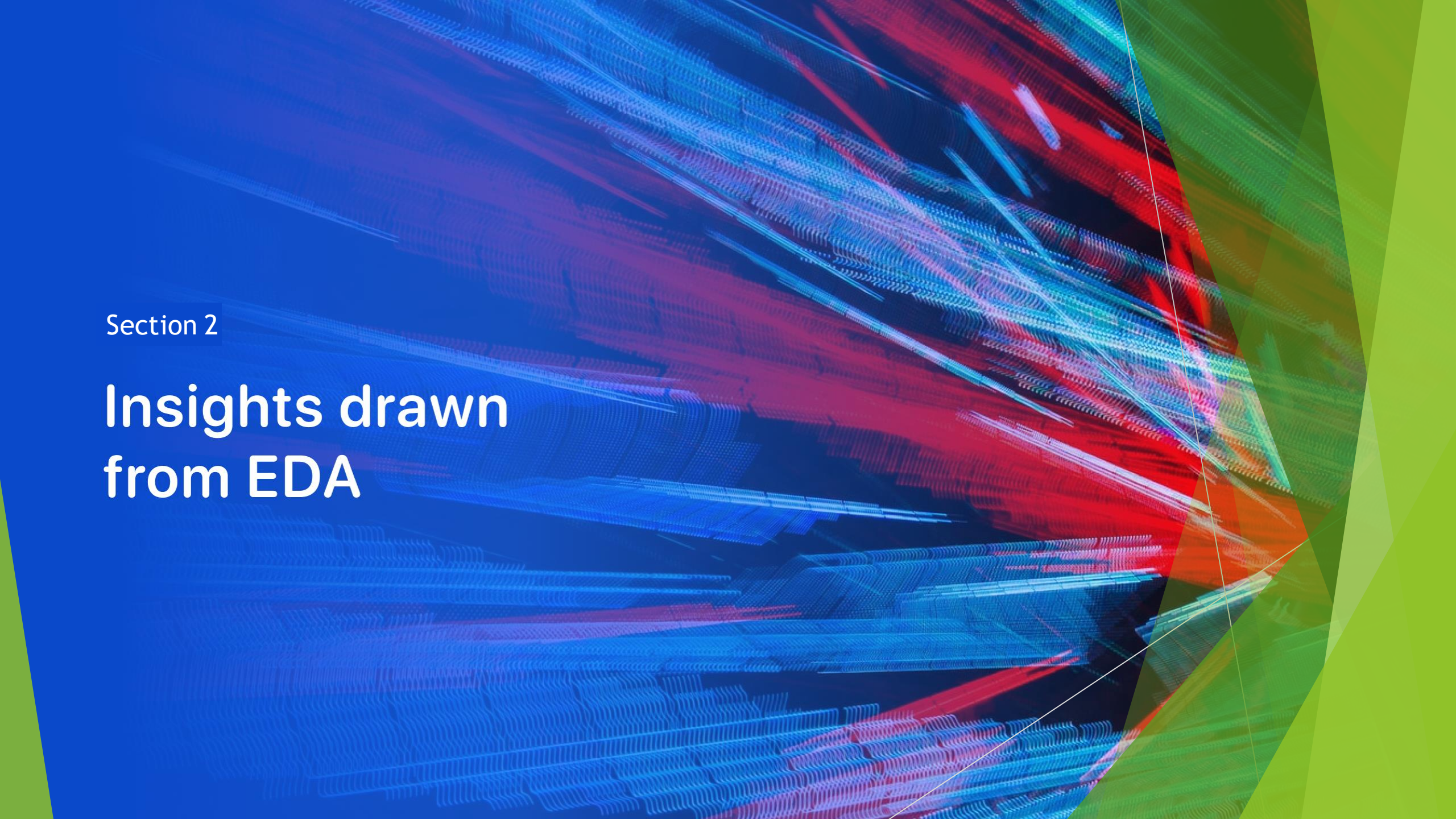
Predictive Analysis (Classification)

- We loaded the data using numpy and pandas
- We transformed and split the data into training and test sets
- We built different machine learning models and tuned different hyperparameters using grid search cv
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning
- We found, that the support vector machine, K nearest number and the logistic regression model achieved the highest accuracy at 83.3%.
- The support vector machine model performed best in terms of area under the curve at 0.958.

https://github.com/TCsillaC/IBM-Data-Science-Professional-Certificate---Capstone-Project/blob/e348c9fd439649c35c0fa4e473a5b6ef15e6d571/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork_labs_Module%206_ML0101EN_SkillUp_FinalAssignment.jupyterlite.ipynb

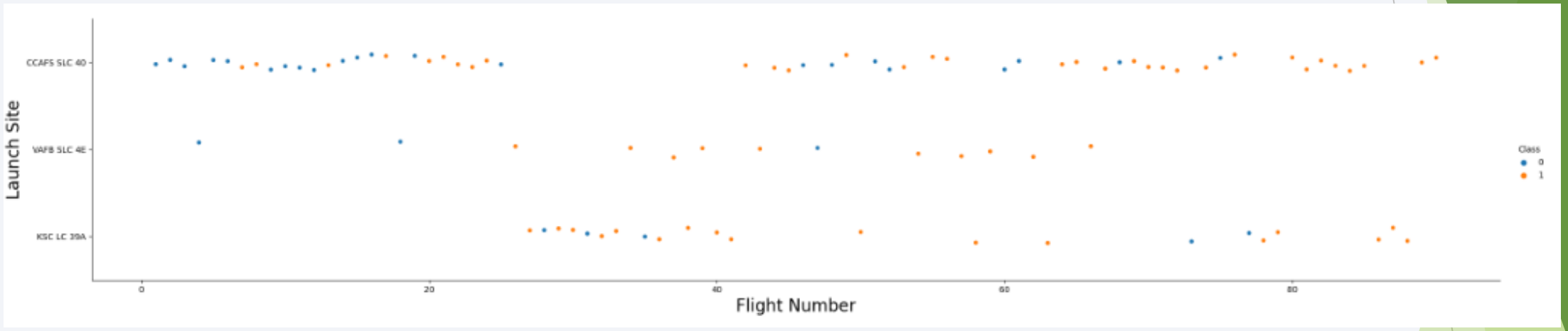
Results

- The support vector machine, K nearest number and the logistic regression model performed best in terms of prediction
- The success rate of SpaceX launches increased with time
- Rockets with lower payloads performed better
- KSC LC 39A site had the most successful launches
- Orbits GEO, HEO, SSO, ES L1 had the best success rate


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

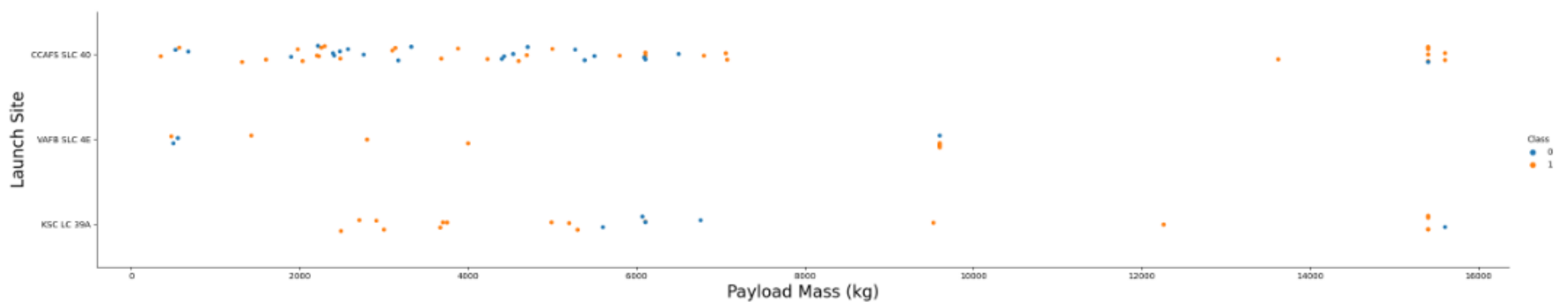
Scatter plot of flight number versus launch site



- There are significantly more launches from CCAFS SLC 40, then from the other sites
- The probability of a good outcome of a launch increases with flight number.

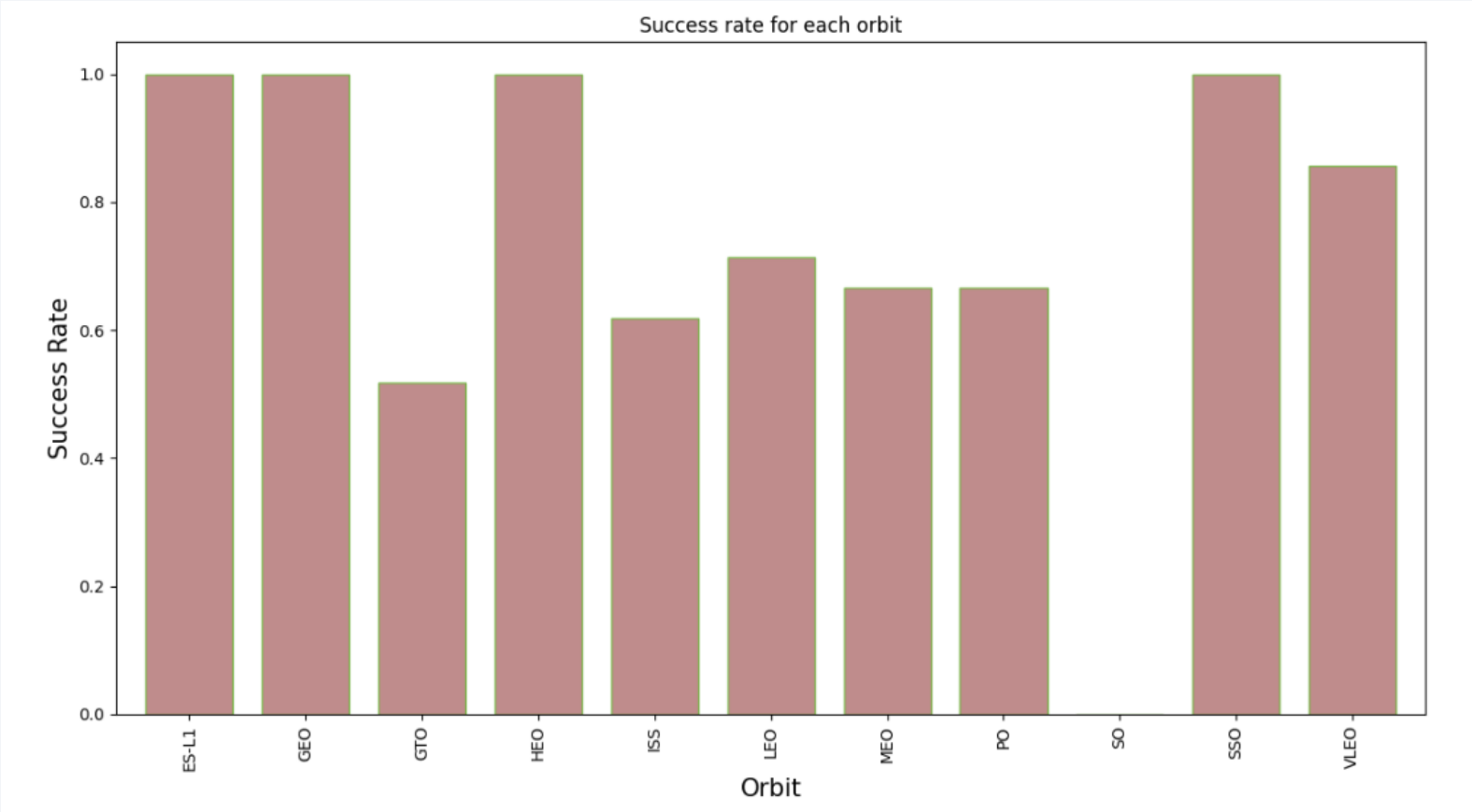
Payload vs. Launch Site

Scatter plot of payload versus launch site



We can observe that for the VAFB-SLC launch site there are no rockets with heavy payload mass (greater than 10000 kg).

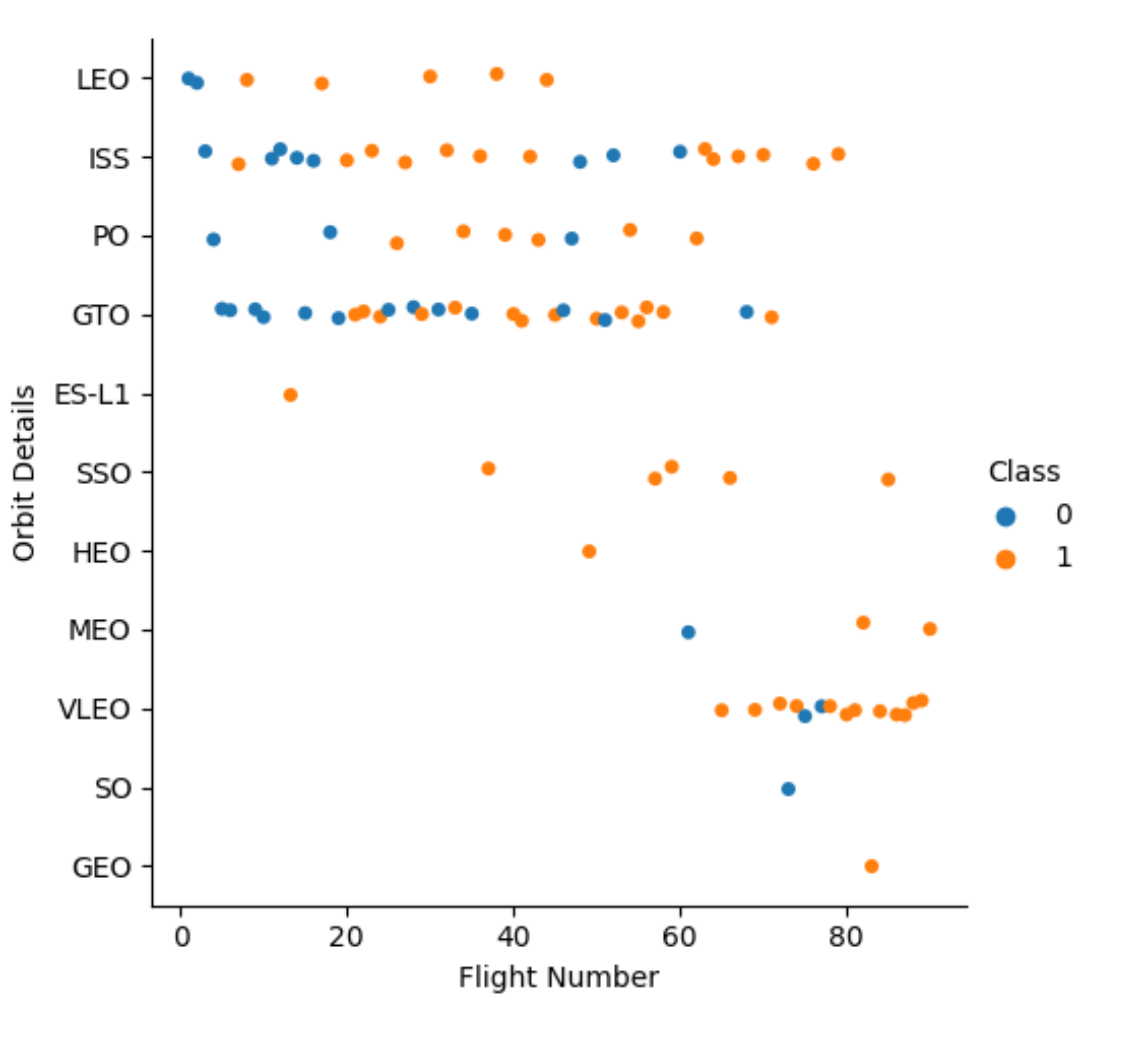
Success Rate vs. Orbit Type



ES-L1, GEO, HEO and SSO orbits have the highest success rate.

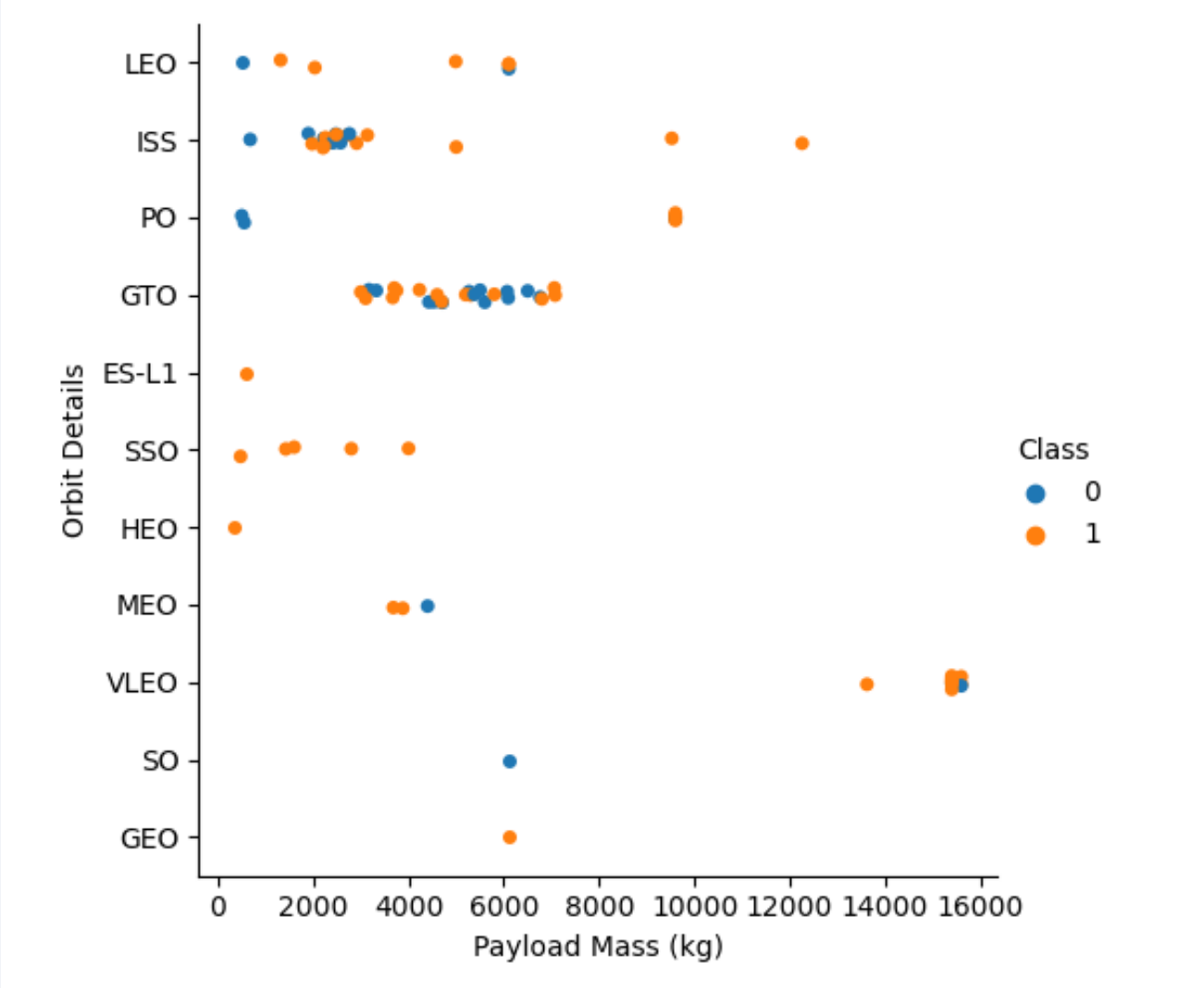
Flight Number vs. Orbit Type

- We can observe that in the LEO orbit the successful landings are more frequent with the increase of the number of flights
- There seems to be no relationship with flight number when in GTO orbit.

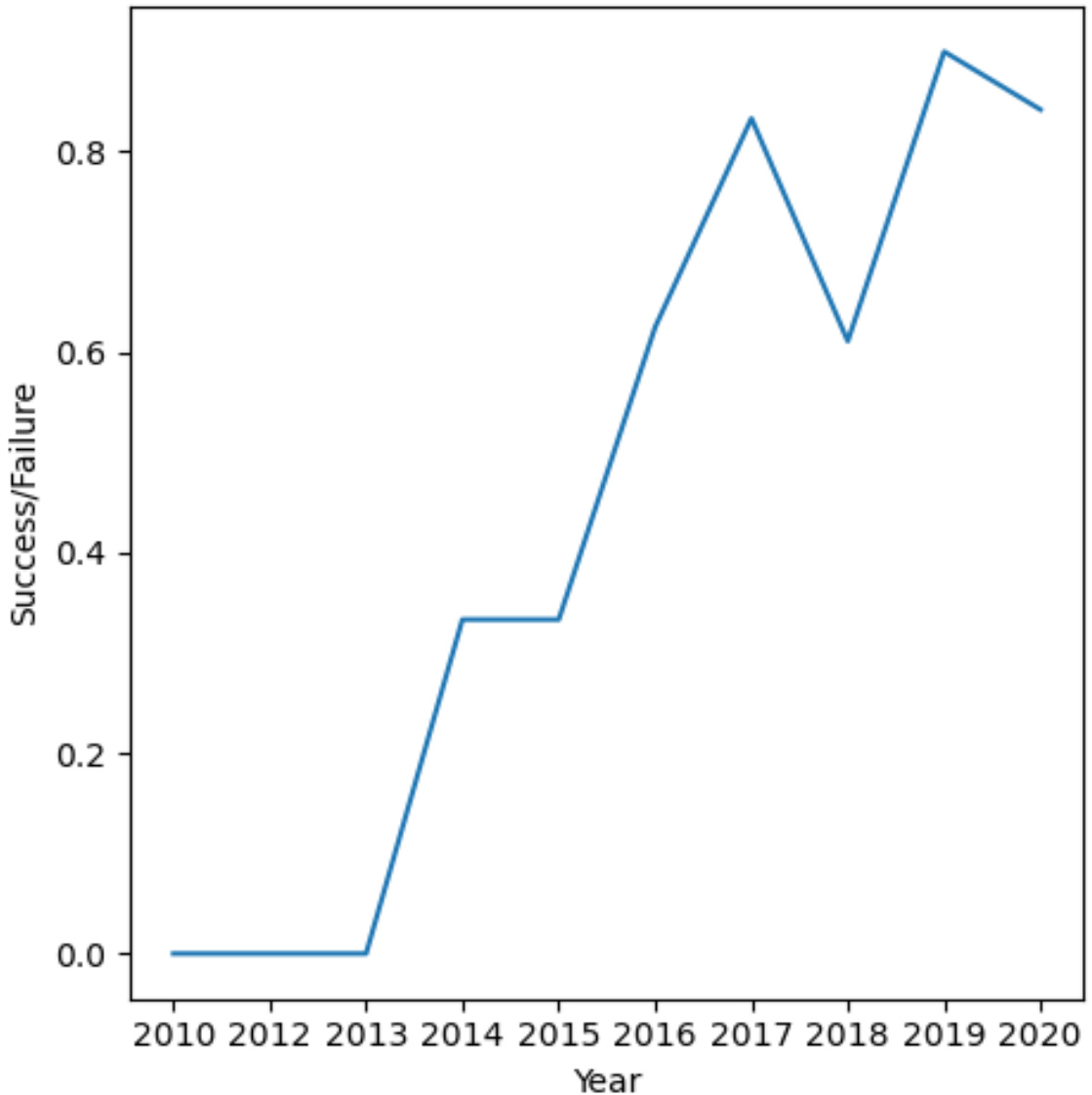


Payload vs. Orbit Type

We can observe that the positive landing rate with heavy payloads is more characteristic to the PO, LEO and ISS orbits.



Launch Success Yearly Trend



The success rate of SpaceX launches increased significantly with time, which could be an effect of learning and improvement of the technology

All Launch Site Names

Find the names of the unique launch sites:

```
%sql select Distinct(LAUNCH_SITE) from SPACEXTBL;
```

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

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

- Using the word "Distinct" in the SQL statement assures that it will show only the unique values in the LAUNCH_SITE column from the SpaceX table.

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`

```
%sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Using the "%" wildcard assures that we select only the rows containing 'CCA'

Using the "LIMIT 5" in the SQL statement assures that it will show the first 5 rows from the SpaceX table.

Total Payload Mass

Calculate the total payload carried by boosters from NASA

```
%sql select sum(PAYLOAD_MASS__KG_) as payloadmax from SPACEXTBL where CUSTOMER = 'NASA (CRS)';
```

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

payloadmax
45596

The function "sum" summates the values in the "PAYLOAD_MASS__KG_" column.

The "where" clause filters the result to the rows pertaining to "NASA (CRS)"

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.

```
%sql select avg(PAYLOAD_MASS__KG_) as payloadmax from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1';
```

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

payloadmax
2928.4

The function "avg" calculates the average of the values in the "Payload_Mass__kg_" column

The "where" clause filters the result to the rows pertaining to the 'F9 v1.1' values in the "BOOSTER_VERSION" column

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

```
%sql select min(DATE) from SPACEXTBL where LANDING_OUTCOME = "Success (ground pad)";
```

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

1
2015-12-22

The "min" function finds the first value in the DATE column

The "where" clause filters the result to the rows pertaining to the 'Success (ground pad)' values in the "LANDING_OUTCOME" column

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME = "Succes (drone ship)" and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000;
```

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

The "where" clause filters the result to the rows pertaining to the 'Success (drone ship)' values in the "LANDING_OUTCOME" column

The "and" clause specifies additional filtering conditions

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

```
%sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

missionoutcomes
1
98
1
1

We selected and counted all the values from the "MISSION_OUTCOME" column

Boosters Carried Maximum Payload

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

```
%sql select BOOSTER_VERSION as booster_version from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

We have used the "max" clause to select the maximal values of "PAYLOAD_MASS__KG_" from the "BOOSTER_VERSION" column

2015 Launch Records

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

```
%sql SELECT substr(Date,4,2) as month,MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where substr(Date,7,4)='2015';
```

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

month	Mission_Outcome	Booster_Version	Launch_Site
01	Success	F9 v1.1 B1012	CCAFS LC-40
02	Success	F9 v1.1 B1013	CCAFS LC-40
03	Success	F9 v1.1 B1014	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40
04	Success	F9 v1.1 B1016	CCAFS LC-40
06	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

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

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

```
%sql SELECT count(MISSION_OUTCOME) FROM SPACEXTBL WHERE DATE BETWEEN '04-06.2010' AND '20-03-2017' ORDER BY DATE DESC;
```

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

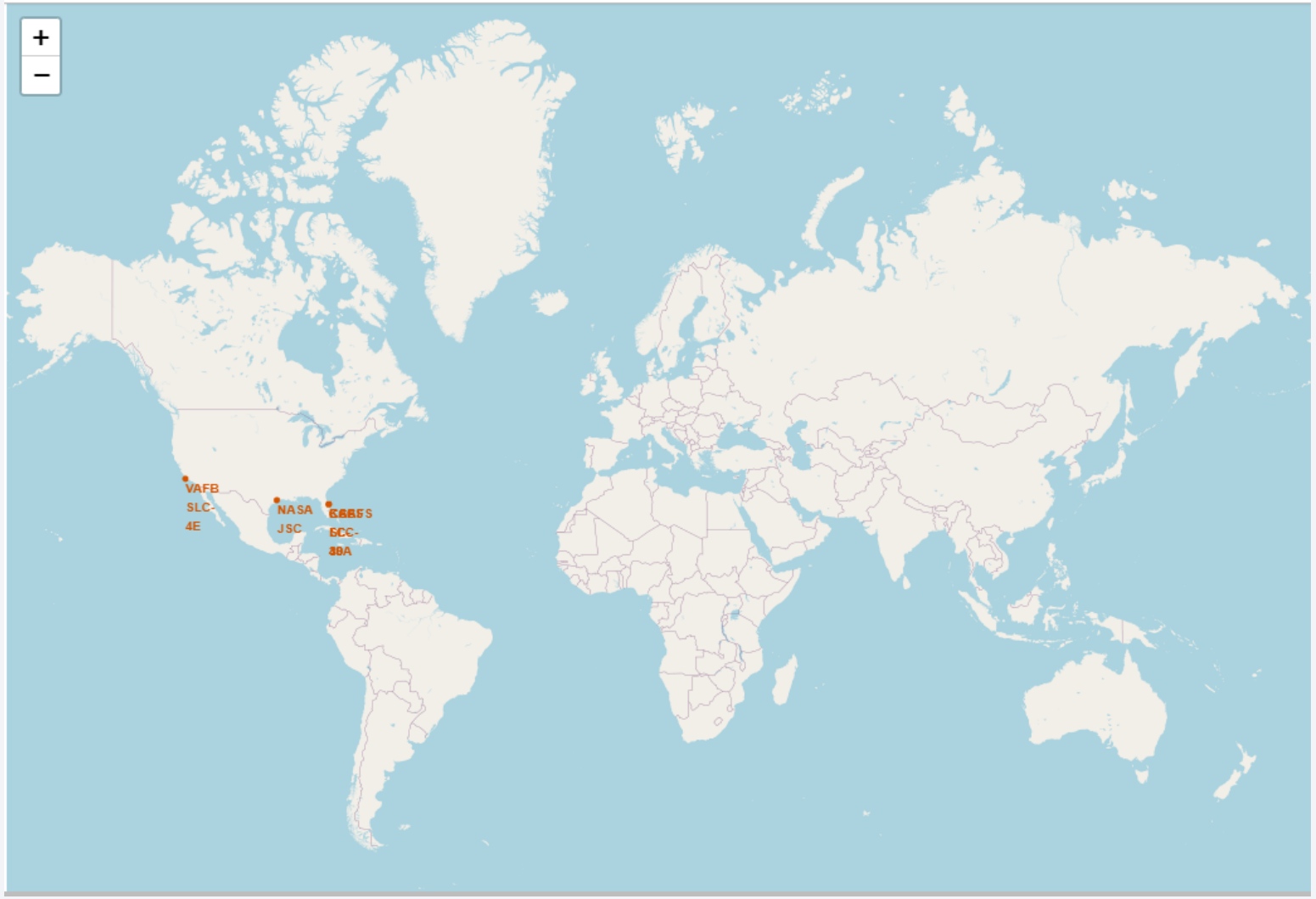
count(MISSION_OUTCOME)
54

We used a combination of the "where", "like", "and", "between" clauses to filter for all landing outcomes.

Section 3

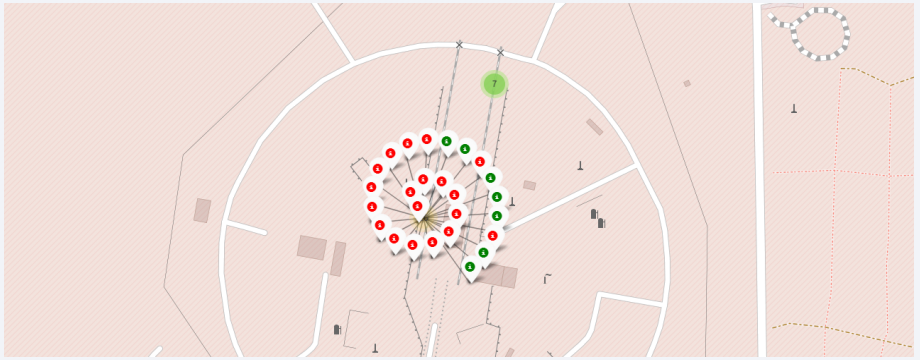
Launch Sites Proximities Analysis

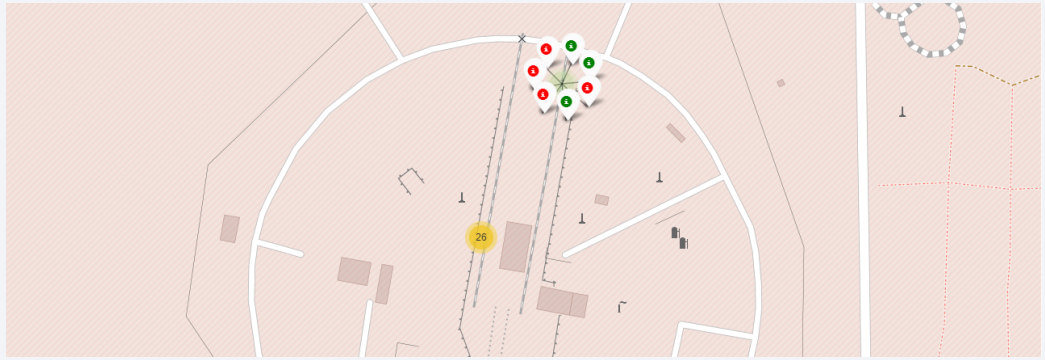
Rocket Launch Sites on a Global Map



We can observe, that all sites can be found on coastal areas of the USA

Launch Outcomes of the Different Sites



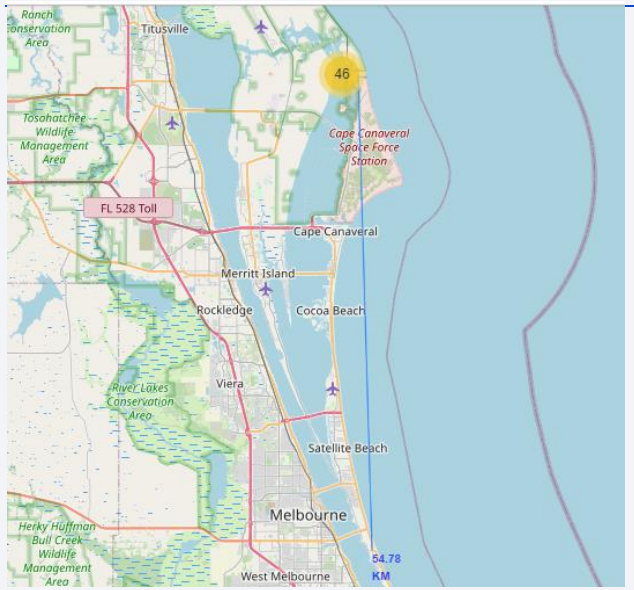


Cape Canaveral Space Launch Complex 40

Vandenberg Space Launch Complex 6 in California

Kennedy Space Center Launch Complex 39A in Florida

Position of a Launch Site



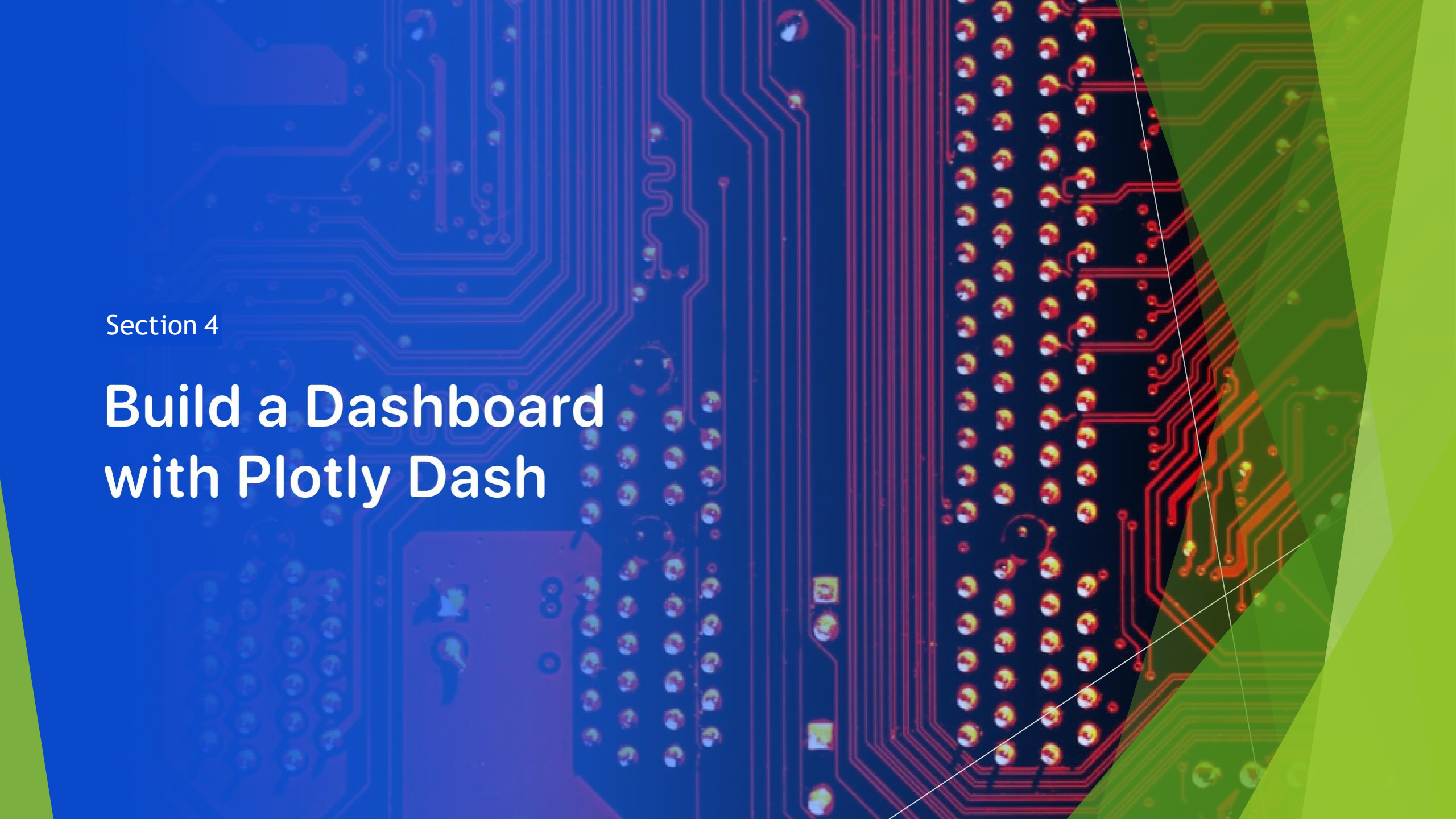

```
lat = 28.56221
long = -80.56777
distance_coastline = calculate_distance(launch_sites_df.iloc[0]["Lat"], launch_sites_df.iloc[0]["Long"], lat, long)
distance_coastline
```

```
0.9365874197351569
```

Distance of Cape Canaveral Space Launch Complex 40 from Melbourne

Distance of Cape Canaveral Space Launch Complex 40 fr om the coastline

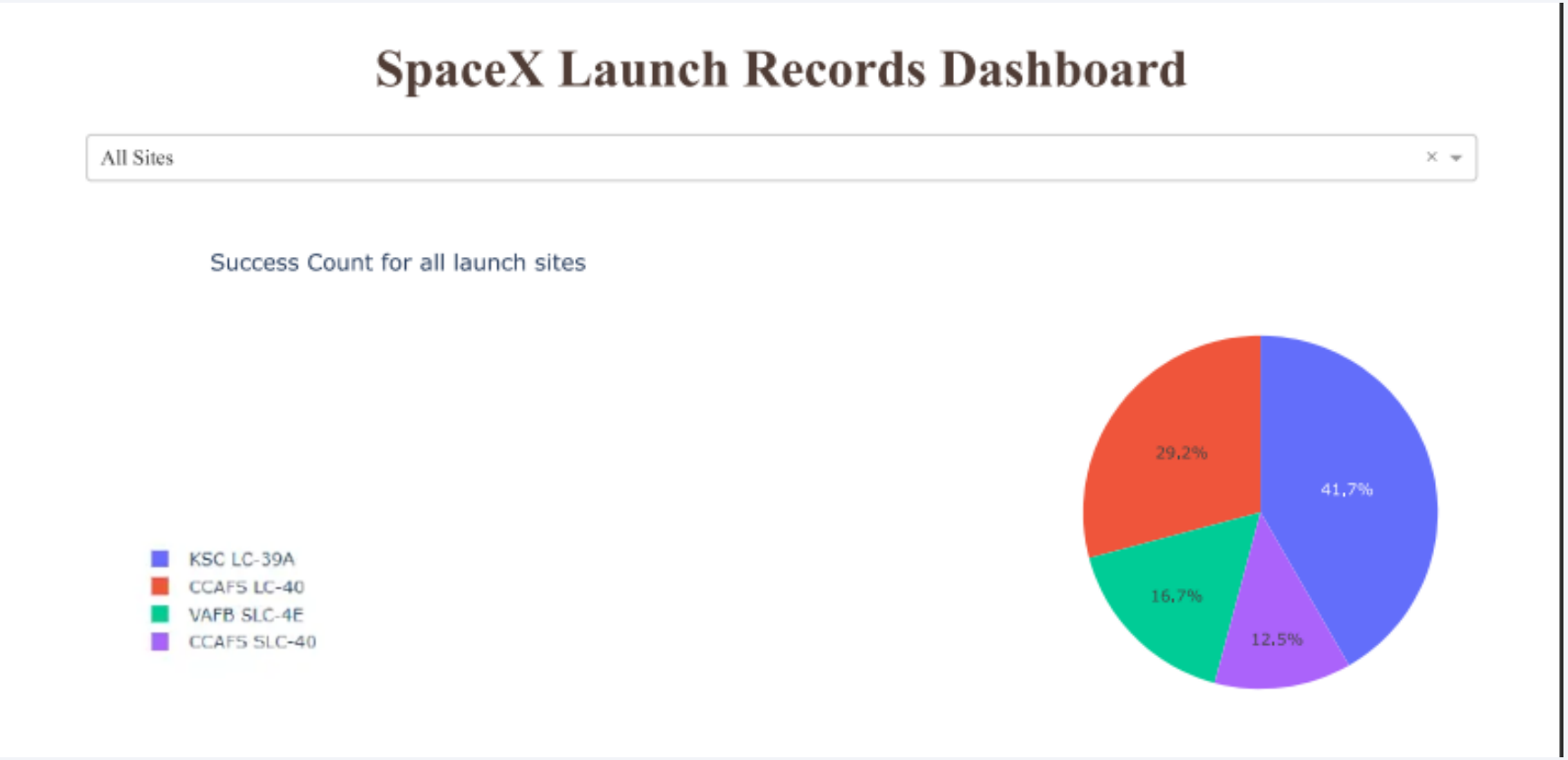
We observed that launch sites are located near to the coast line and relatively far from cities, railways and highways



Section 4

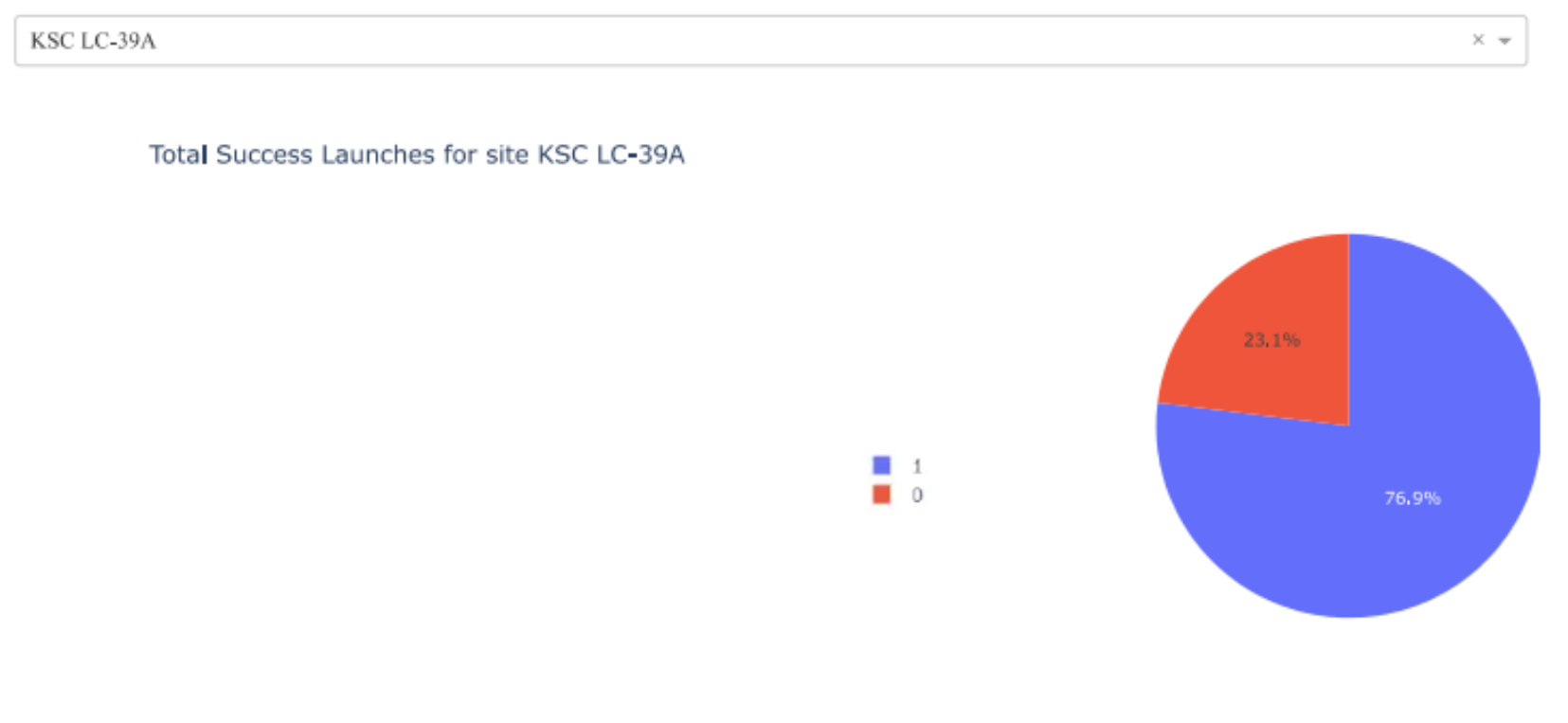
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



We can observe that the Kennedy Space Center Launch Complex 39A had the most successful launches.

The Launch Site with the Highest Success Ratio



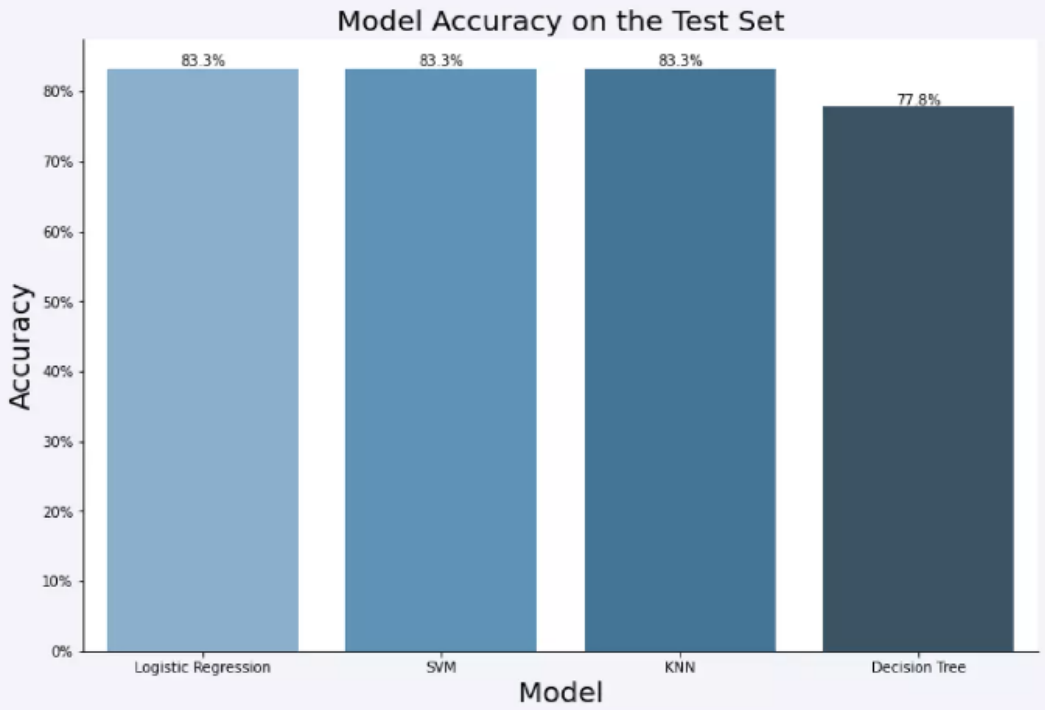
We can observe that the Kennedy Space Center Launch Complex 39A had a success ratio of 76.9%.

Payload vs. Launch Outcome for All Sites

Section 5

Predictive Analysis (Classification)

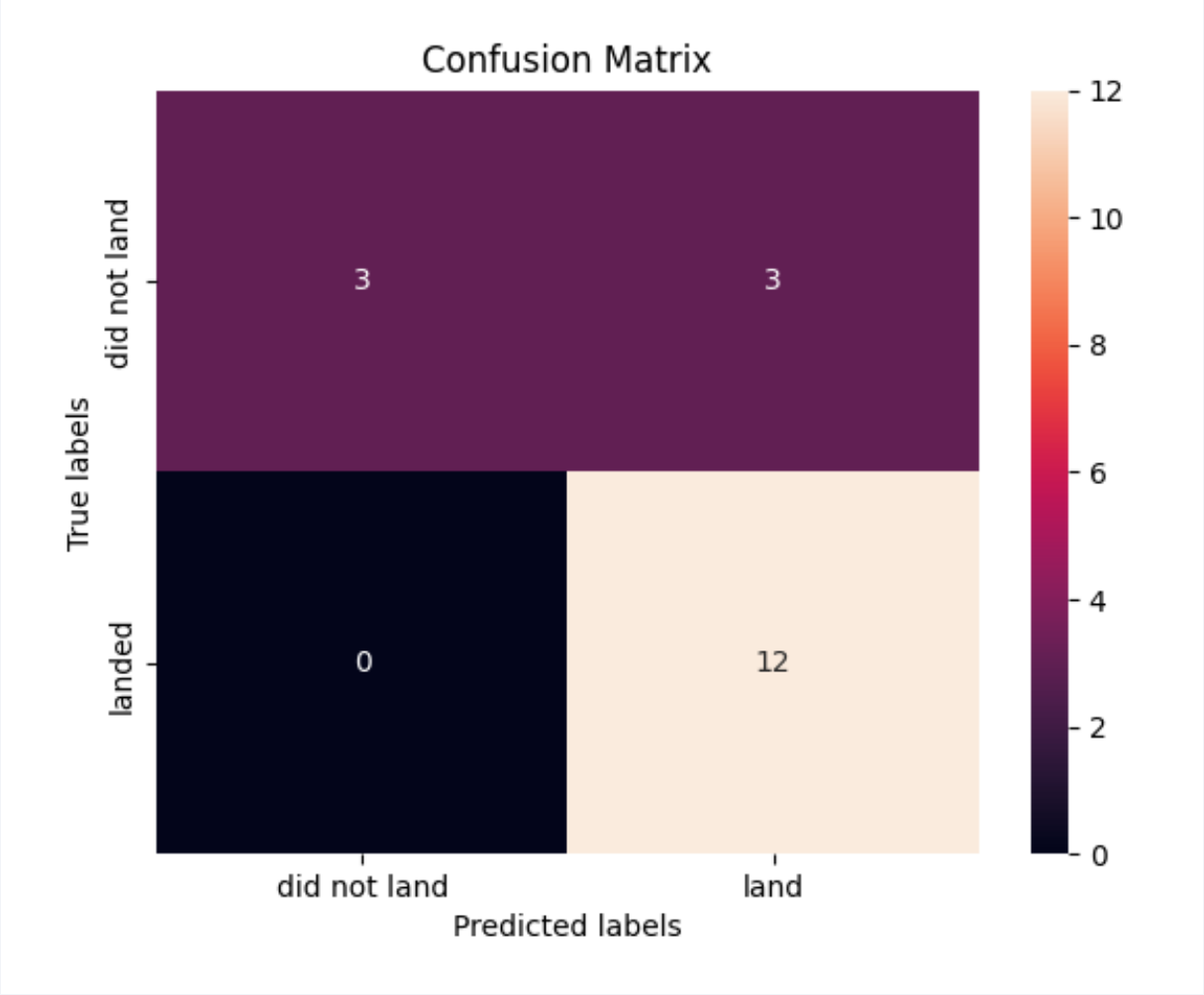
Classification Accuracy



We found, that the support vector machine, K nearest number and the logistic regression model achieved the highest accuracy.

Confusion Matrix

The confusion matrix of the best performing model:



Conclusions

- The support vector machine, K nearest number and the logistic regression model performed best in terms of prediction
- Launch of rockets with lower payloads is more frequently successful
- The success rate of the launches increased significantly with time, which could be an effect of learning and improvement of the technology
- The Kennedy Space Center Launch Complex 39A had the most successful launches.
- HEO, GEO, SSO, ES L1 orbits had the best success rate
- We can conclude that there is a high probability of a successful landing of the Falcon 9 rocket.

Thank you!