Winning Space Race with Data Science

Lucas De Cunto Costanzo
08/08/2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- The main methodologies used for the analysis were:
  - Data Collection
  - EDA with Data Visualization
  - EDA with SQL
  - Interactive maps with Folium
  - Dashboard with Plotly
  - Predictive analysis
- Summary of all results
  - Preliminary analysis based EDA
  - Interactive maps and dashboards
  - Predictive results

Introduction

- Currently several companies are trying to make space travel more affordable. One of the most successful companies to achieve that is SpaceX, in which launches are announced for 62 million dollars, while other providers can cost up to 165 million dollars. This reduction in price is due to the fact that SpaceX can retrieve and reuse the rockets first stage.

- With this analysis we will try to determine if the landing of a rockets first stage will be successful, using SpaceX historical data, so we can estimate better the flight cost for a new company and reduce the final price as a consequence.

Section 1

Methodology

Methodology

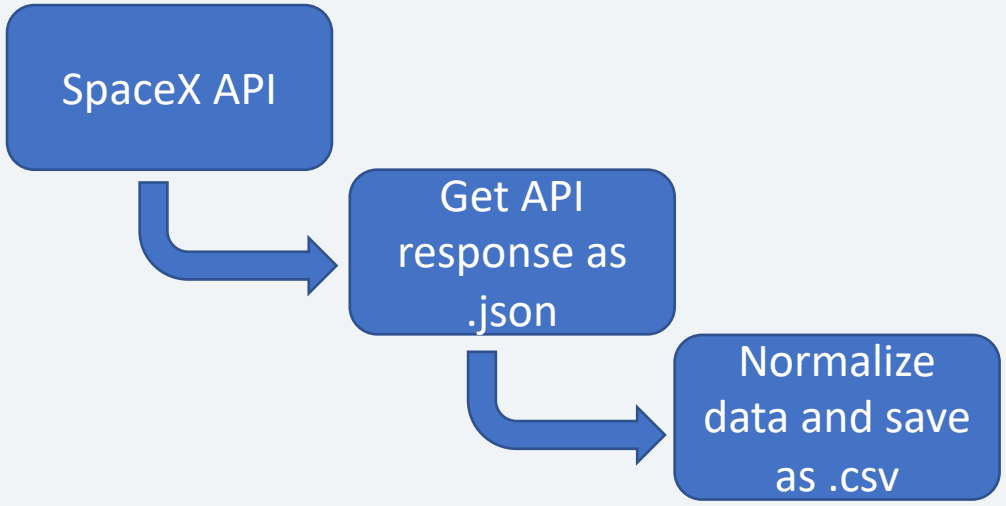
Executive Summary

- Data collection methodology:
  - The data was collected using two main approaches, SpaceX REST API and web scraping
- Perform data wrangling
  - One hot encoding to prepare data for machine learning algorithms
- 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 preparation and model building, tuning e evaluation

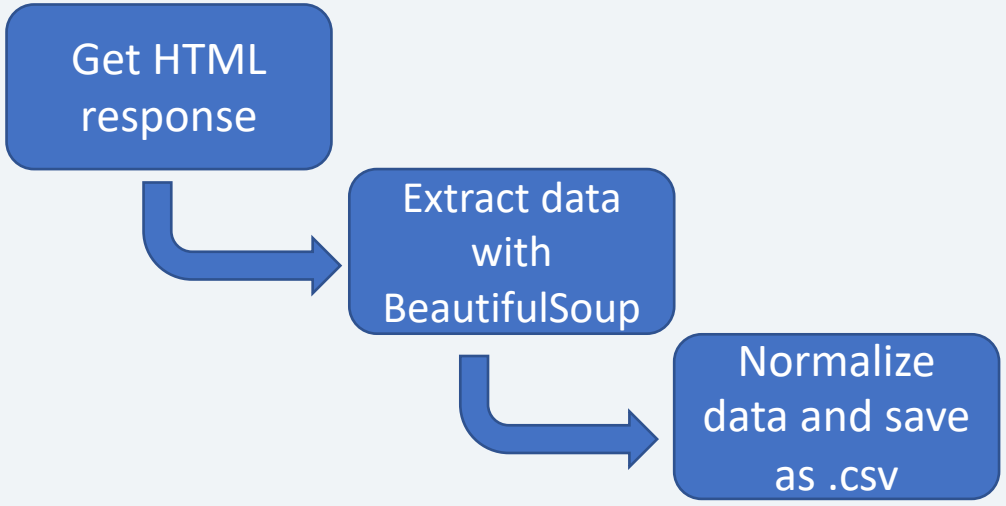
Data Collection

- The Datasets used in this analysis were collected using two main methodologies
  - SpaceX API
  - Web Scrapping in Wikipedia

- **SpaceX API**



- **Web Scrapping**



Data Collection – SpaceX API

Getting data from API

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

Converting to .json file

```
data = pd.json_normalize(response.json())
```

Cleaning data (Predefined functions)

```
# Call getBoosterVersion
getBoosterVersion(data)
# Call getLaunchSite
getLaunchSite(data)
# Call getPayloadData
getPayloadData(data)
# Call getCoreData
getCoreData(data)
```

Constructing Dataset

```
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}
data = pd.DataFrame(launch_dict)
```

Filtering Dataset

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

- SpaceX API Lab Link

Data Collection - Scraping

Getting data from html

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

Creating Object

```
soup = BeautifulSoup(response.text, "html.parser")
```

Finding Tables

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

Getting Columns

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

Creating Dictionary

```
launch_dict= dict.fromkeys(column_names)
del launch_dict['Date and time ( )']
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'] = []
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

Appending data

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

Creating Dataframe

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

- Webscraping Lab Link

Data Wrangling

EDA on Dataset

```
df.isnull().sum()/df.count()*100
```

```
df.dtypes
```

Number of launches at each site

```
df['LaunchSite'].value_counts()
```

Occurrence of each orbit

```
df['Orbit'].value_counts()
```

Occurrence of landing outcome

```
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

Landing Outcome Label from Outcome Column

```
landing_class = []
for i,outcome in enumerate(df['Outcome']):
    if outcome in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)

landing_class
```

Success Rate

```
df["Class"].mean()
```


- Data Wrangling Lab Link

EDA with Data Visualization

- Scatter Plots
  - Flight Number VS. Payload Mass
  - Launch Site VS. Flight Number
  - Payload Mass VS. Launch Site
  - Flight Number VS. Orbit
  - Payload Mass VS. Orbit
- Bar Chart
  - Orbit Type VS. Success Rate
- Line Graph
  - Year VS. Success Rate

Scatter plots will be used to show the relationship between two variables and how they affect each other, which is also called **correlation**.

Bar Chart are good to compare sets of data between two different sets, where the X axis will represent a categorical variable and the Y axis a discrete value.

Line graphs are very useful to visualize trends in the relation between two variables. Because of that, they can help make predictions about future data.

- Data Visualization Lab Link

EDA with SQL

- Performed SQL query's
  - 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 F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the name of the boosters which have success in drone ship and have payload mass grater 4000 but less than 6000
  - List the total number of success and failure mission outcomes
  - List the number of the booster_versions which have carried the maximum payload mass
  - List records which will display the month names, failure landing_outcomes in drone ship, booster versions and launch_site for the months in year 2015
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order
- SQL Lab Link

Build an Interactive Map with Folium

- The map objects added to the map were
  - Markers
  - Circles
  - Lines
- This map objects were added to give an idea of the places launches were performed, how many were carry out in each place and their distance to areas like railways, highways and coastlines.

- Folium Lab Link

Build a Dashboard with Plotly Dash

- In the created dashboard, the following objects were added:
  - Dropdown menu: Select all or one of the launch sites to visualize data.
  - Pie chart: Shows the proportion of successful and failure landing outcomes for the selected landing site.
  - Range slider: Selects the payload range of data to be displayed in the scatter plot.
  - Scatter plot: Shows correlation between payload mass and success rate, distinguishing by color booster versions.
- Those objects were added to interactively show the selected data graphically.

- Plotly Dash Lab Link

Predictive Analysis (Classification)

- Model Building
  - We first imported the data to NumPy and Pandas
  - Transformed data to work with machine learning algorithms
  - Split the data in training and testing sets
  - Select the machine learning algorithm
  - Set parameter to be tested with GridSearchCV
  - Train our model with data training set and the best parameters found by GridSearchCv
- Evaluating Model
  - Check accuracy for each model using data test set
  - Plot confusion matrix to visually analyze the performance
- Improving model
  - Feature engineering
  - Algorithm tuning

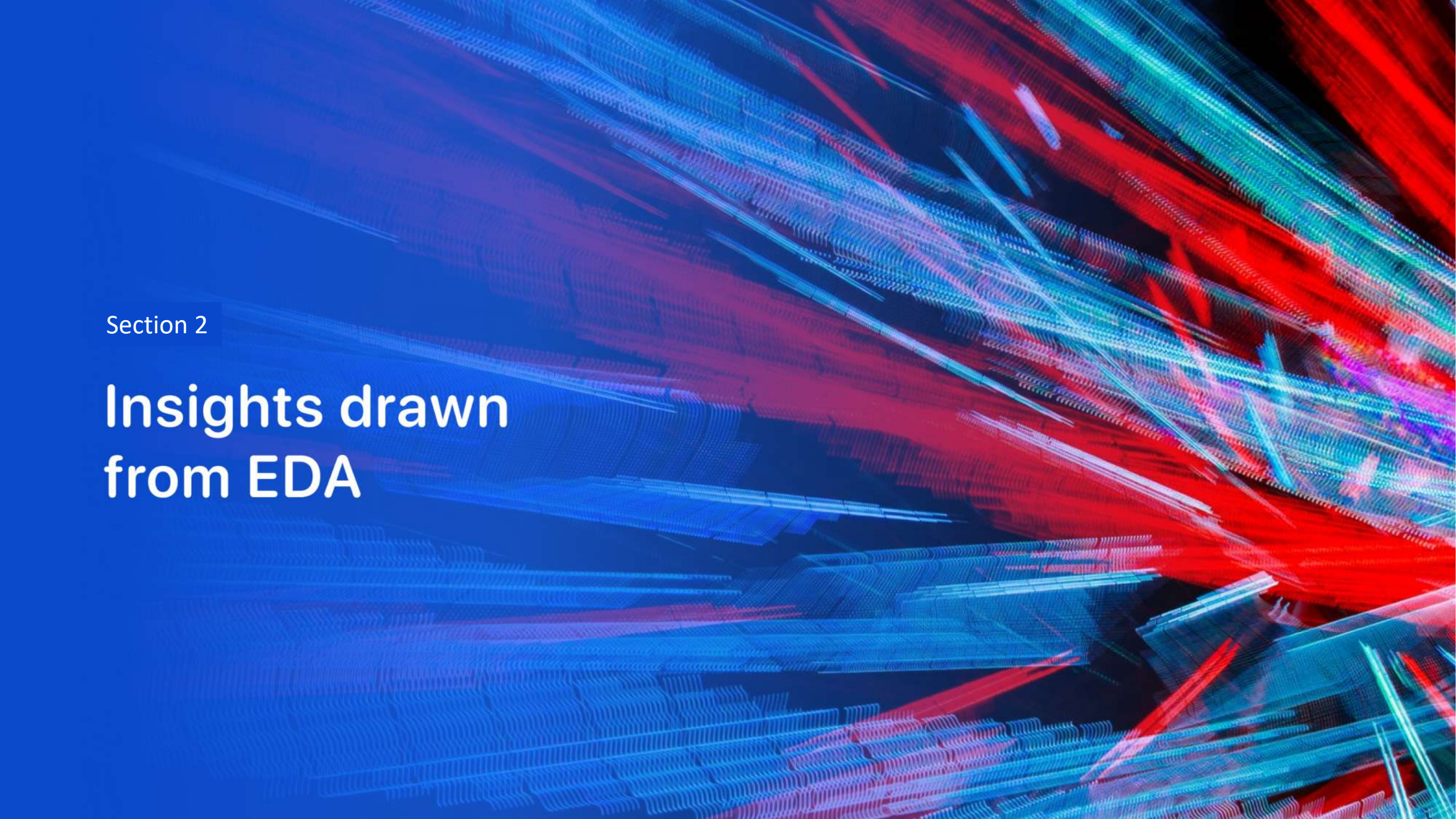
- Predictive Analysis Lab Link

Results

Exploratory data analysis results

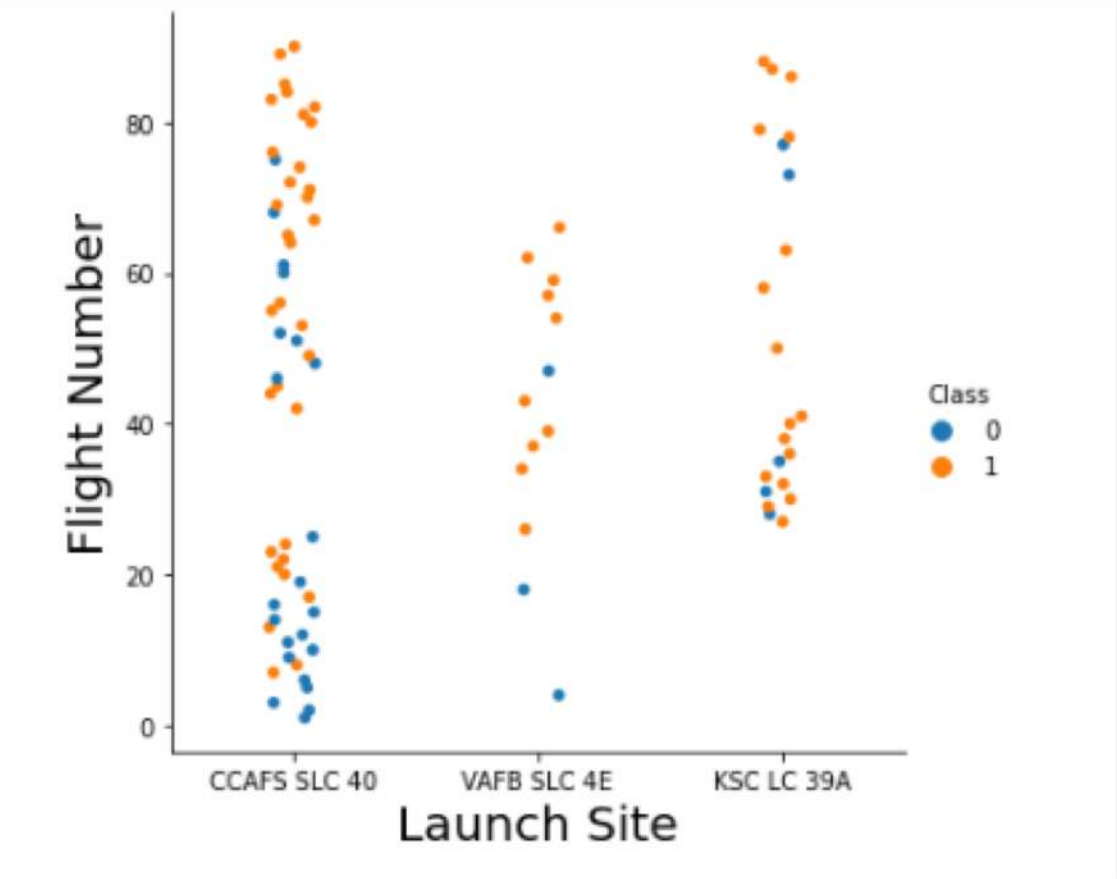
Interactive analytics demo in screenshots

Predictive analysis results


Section 2
Insights drawn
from EDA

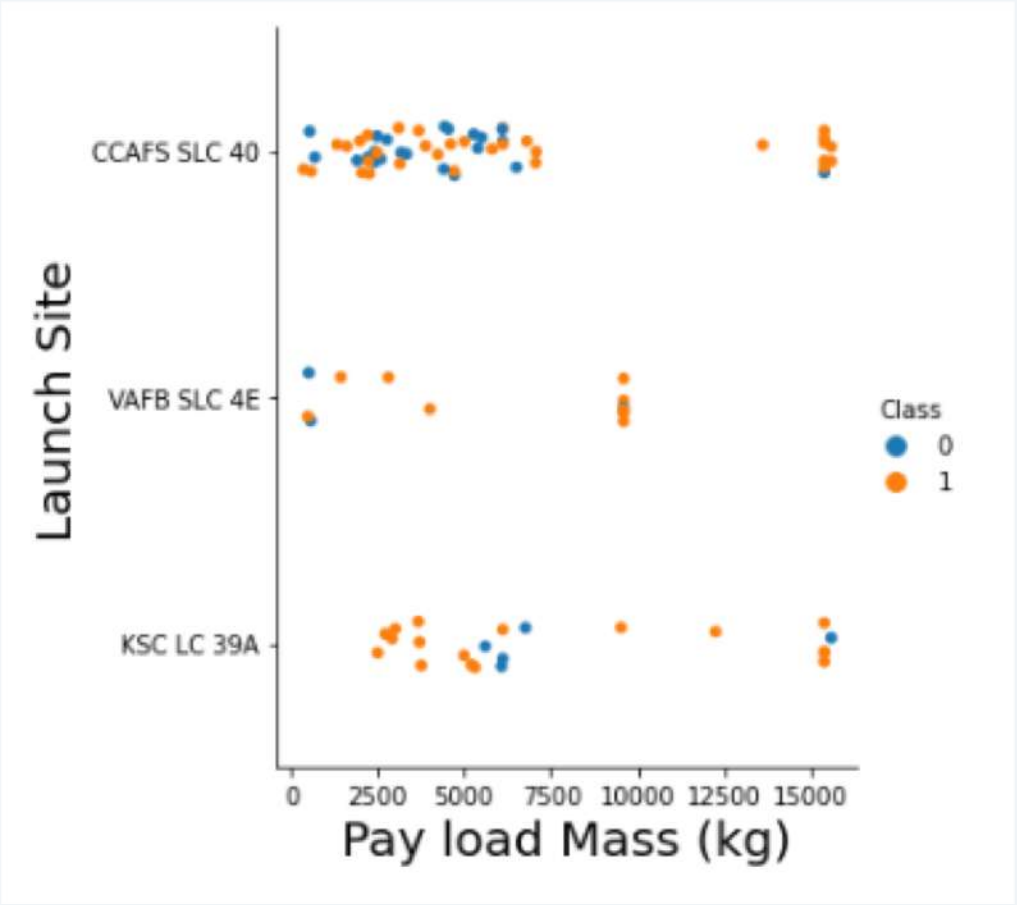
Flight Number vs. Launch Site

- From this scatter plot we can observe that the successful landings rate increases with the flight number.
- Furthermore, the landings tends to be more successful in the launch sites “VAFB SLC 4E” and “KSC LC 39A”.



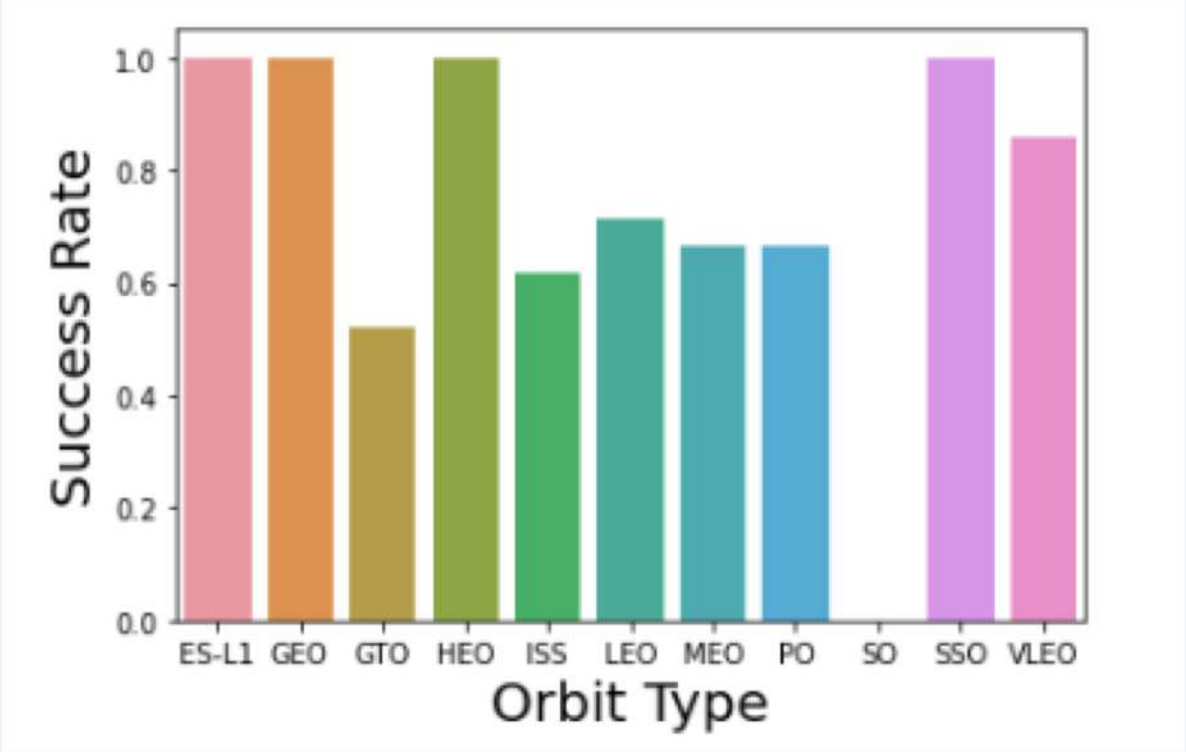
Payload vs. Launch Site

- This graph shows that the success rate tends to increase with the payload mass.
- Besides, it supports the idea that the launch site has influence in the success.



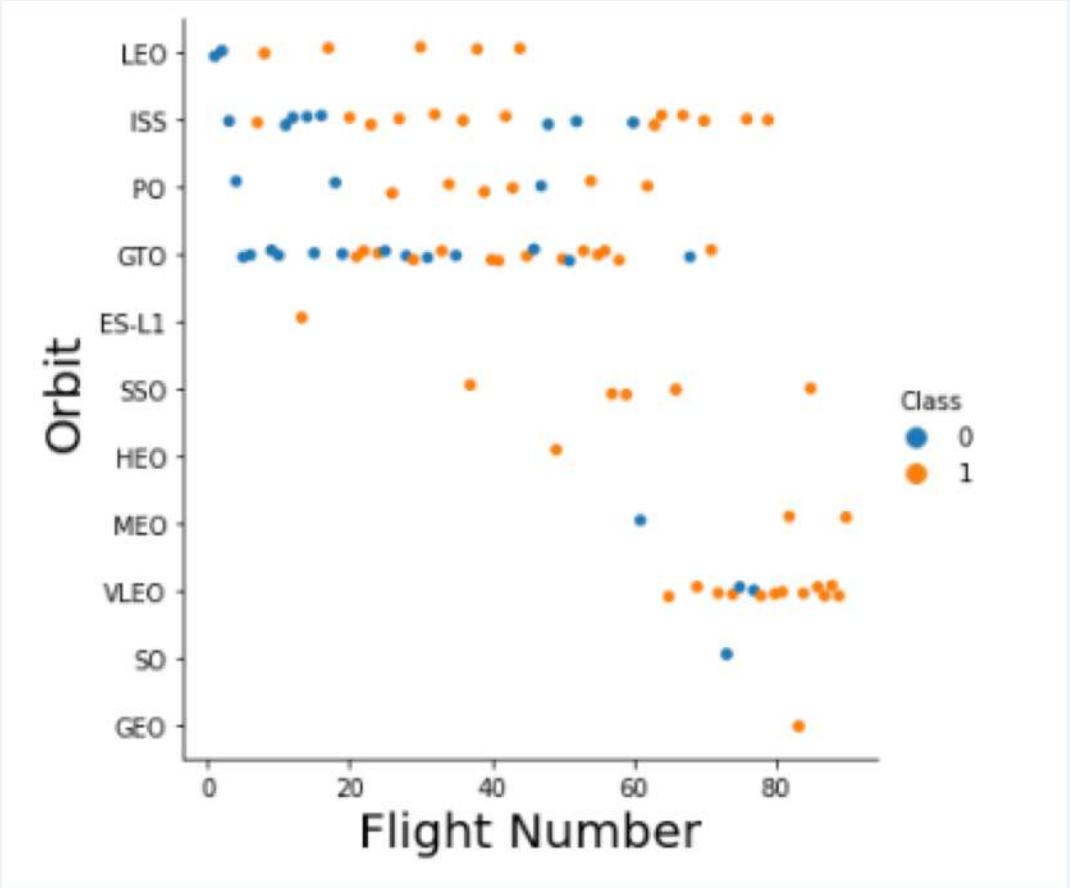
Success Rate vs. Orbit Type

- The missions performed in orbits “ES-L1”, “GEO”, “HEO” and “SSO” present the highest chance of being successful, with a success rate of 100%. Conversely, missions in the “SO” orbit present a low chance of success with a success rate of 0%.



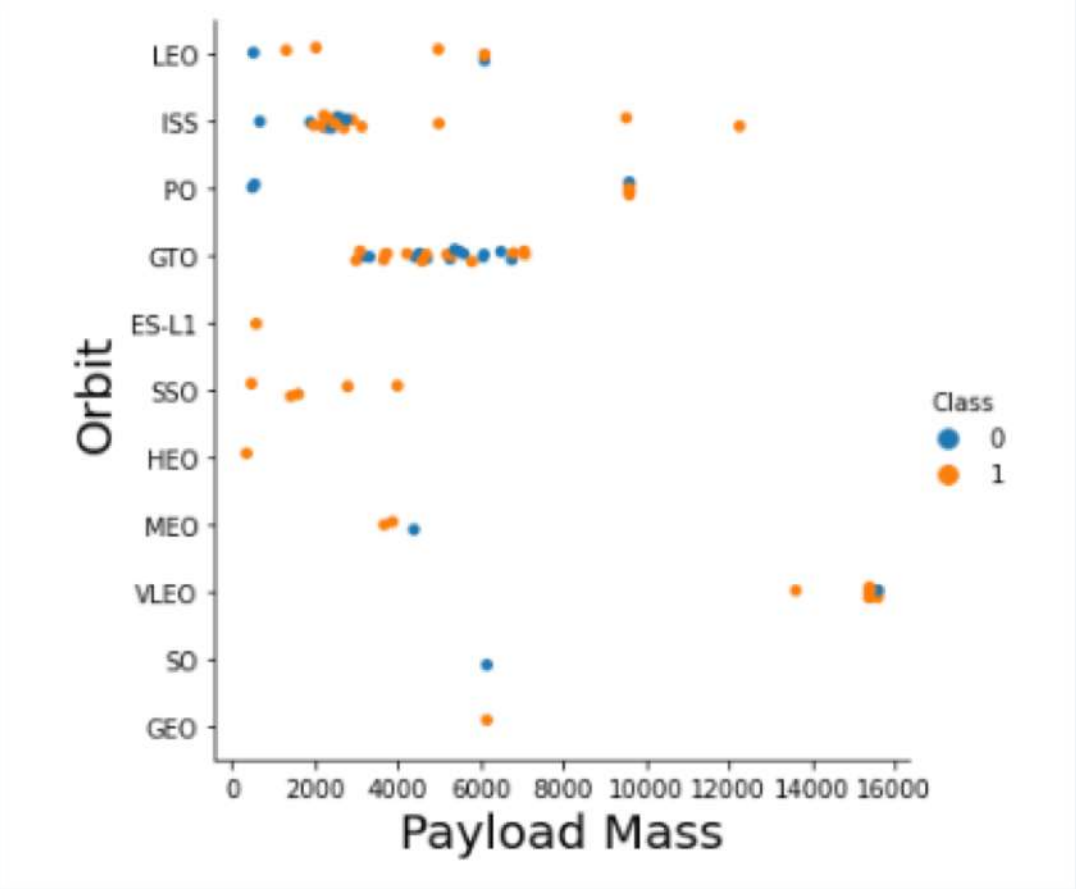
Flight Number vs. Orbit Type

- In this scatter plot we can see that the success rate increases with the flight number in the ”LEO” orbit, while it seems to have no relation in the “GTO” orbit.



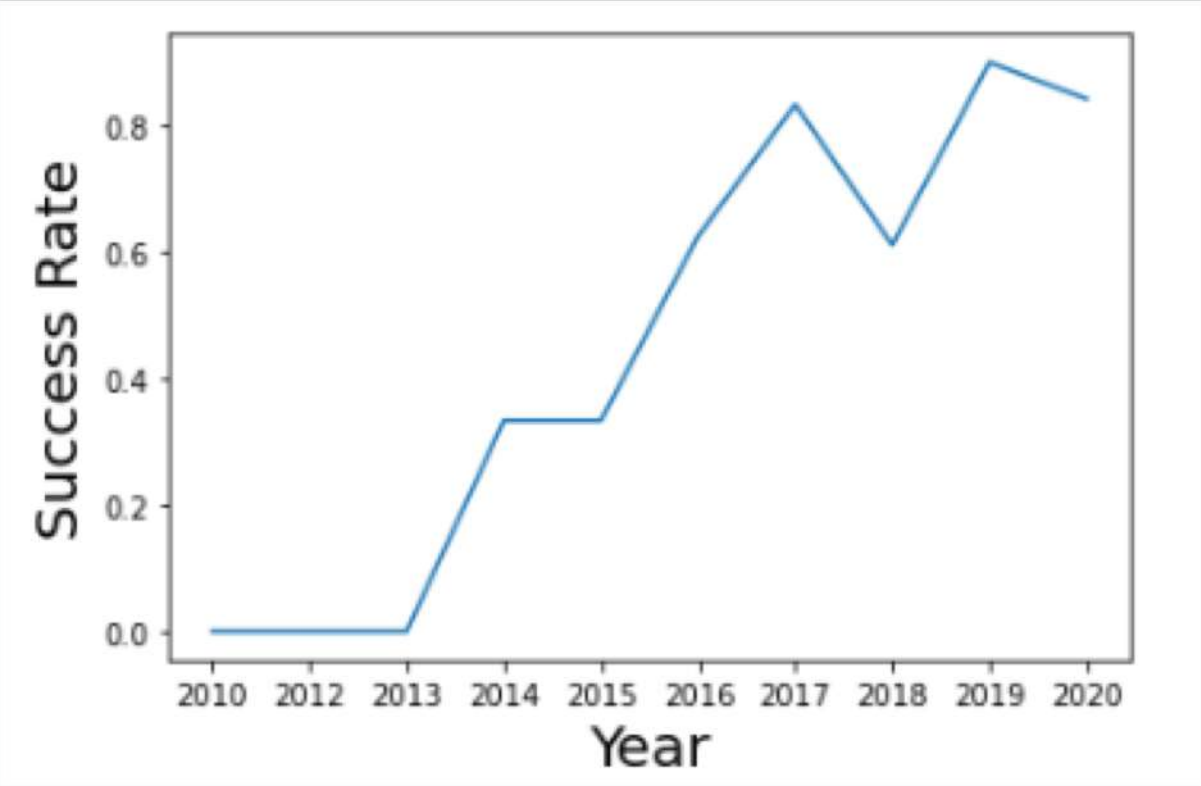
Payload vs. Orbit Type

- From this plot we can conclude that the payload mass has a positive effect in the success rate in “LEO”, “ISS” and “Polar” orbits as it increases. However, we can’t conclude about it’s influence in “GTO” orbit.



Launch Success Yearly Trend

- From this line chart we can observe that, even with the descend trend in specific years, the success rate tends to increase over the years.



All Launch Site Names

- The query used to find the names of the unique launch sites was

```
select distinct "Launch_Site" from SPACEXTBL
```

The distinct statement selects the unique values in column “Launch_Site” in table SPACEXTBL, and it result is presented in the figure aside.

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

Launch Site Names Begin with 'CCA'

- The query was used to find the first 5 records in the table in which the launch site begins with “CCA”. Its result is presented in the figure below

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 query used to calculate the total payload carried by NASA was

```
%sql select sum("PAYLOAD_MASS__KG_") from SPACEXTBL where "Customer" = "NASA (CRS)”
```

where the sum() statement sums all values in the selected column while the where statement filter only the rows with the desired “Customer”.

- Its result is presented in the figure below

Total Payload Mass
45596

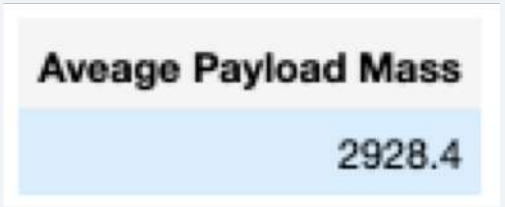
Average Payload Mass by F9 v1.1

- The query used to calculate the average payload carried by booster version “F9 v1.1”was

```
%sql select avg("PAYLOAD_MASS__KG_") from SPACEXTBL where "Booster_Version" = "F9 v1.1”
```

where the avg() statement averages all values in the selected column while the where statement filter only the rows with the desired booster version.

- Its result is presented in the figure below



Aveage Payload Mass
2928.4

First Successful Ground Landing Date

- The query used to calculate the first successful ground landing date was

```
%sql select min("Date") from SPACEXTBL where "Landing _Outcome" = "Success (ground pad)”
```

where the min() statement selects the minimum all values in the “Date” column while the where statement filter only the rows with the desired landing outcome.

- Its result is presented in the figure below

min("Date")
2015-12-22 00:00:00

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query used to calculate all successful drone ship landings with payload in the desired interval

```
%sql select "Booster_Version", "PAYLOAD_MASS__KG_", "Landing _Outcome" from SPACEXTBL \
where "Landing _Outcome" = "Success (drone ship)" and ("PAYLOAD_MASS__KG_" between 4000 and 6000)
```

In which the where clause will have two filters, the landing outcome and the payload mass.

- Its result is presented in the figure below

Booster_Version	PAYLOAD_MASS__KG_	Landing _Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- The query used to calculate the total number of successful and failure mission outcomes drone ship landings with payload in the desired interval

```
%sql select "Mission_Outcome", count("Mission_Outcome") from SPACEXTBL \
group by "Mission_Outcome"
```

In which the count statement counts the number of occurrences of each outcome and the group by statement groups the unique mission outcomes.

- Its result is presented in the figure below

Mission_Outcome	count("Mission_Outcome")
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The query used to list the names of all boosters that have carried the maximum payload was

```
%sql select distinct "Booster_Version", "PAYLOAD_MASS__KG_" from SPACEXTBL \
where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTBL)
```

where the distinct statement avoid repetition of booster versions and the subquery selects the maximum payload carried by the boosters.

- Its result is presented in the figure below

Booster_Version	PAYLOAD_MASS__KG_		
F9 B5 B1048.4	15600	F9 B5 B1049.5	15600
F9 B5 B1049.4	15600	F9 B5 B1060.2	15600
F9 B5 B1051.3	15600	F9 B5 B1058.3	15600
F9 B5 B1056.4	15600	F9 B5 B1051.6	15600
F9 B5 B1048.5	15600	F9 B5 B1060.3	15600
F9 B5 B1051.4	15600	F9 B5 B1049.7	15600

2015 Launch Records

- The query used to list the failure landing outcomes in 2015 was

```
%sql select substr("Date",6,2) as Month_Name, "Landing _Outcome", "Booster_Version", "Launch_Site"  \
from SPACEXTBL \
where ("Landing _Outcome" = "Failure (drone ship)") and (substr("Date", 1, 4) = "2015")
```

In which the substr() statement selects part of the string values in a column.

- Its result is presented in the figure below

Month_Name	Landing _Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The query used to rank the successful landing outcomes in the selected period of time, ordered by descending date order, was

```
%sql select "Date", "Landing _Outcome" from SPACEXTBL \
where ("Landing _Outcome" like "%Success%") and ("Date" between "2010-06-04" and "2017-03-20") \
order by "Date" desc
```

where the desc statement order by descent order of the selected column.

- Its result is presented in the figure below

Date	Landing _Outcome
2017-03-06 00:00:00	Success (ground pad)
2017-02-19 00:00:00	Success (ground pad)
2017-01-14 00:00:00	Success (drone ship)
2017-01-05 00:00:00	Success (ground pad)
2016-08-14 00:00:00	Success (drone ship)
2016-08-04 00:00:00	Success (drone ship)
2016-07-18 00:00:00	Success (ground pad)
2016-06-05 00:00:00	Success (drone ship)
2016-05-27 00:00:00	Success (drone ship)
2015-12-22 00:00:00	Success (ground pad)

Section 3

Launch Sites Proximities Analysis

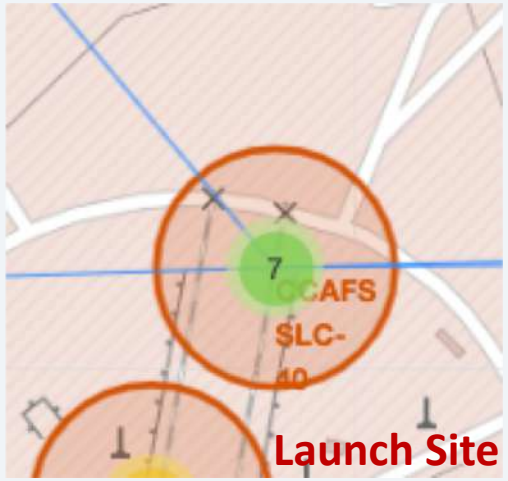
Launch Sites Map

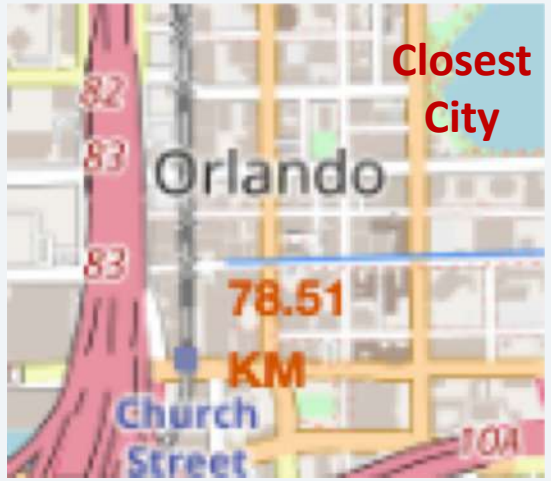
- We can see that the launch sites are located in two regions of the United States, three of them in the east coast and one in the west coast.

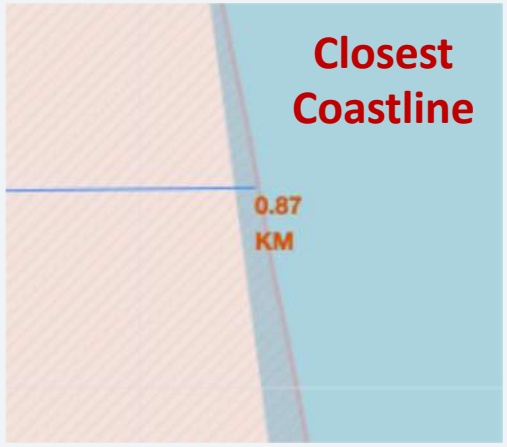
Success/Failure in Launch Sites

- We can see in the figures the number of launches at each site and if the landing was successful (in green) or failure (in red).

Launch Sites Distance to Landmarks Using CCAFS SLC-40 as Reference



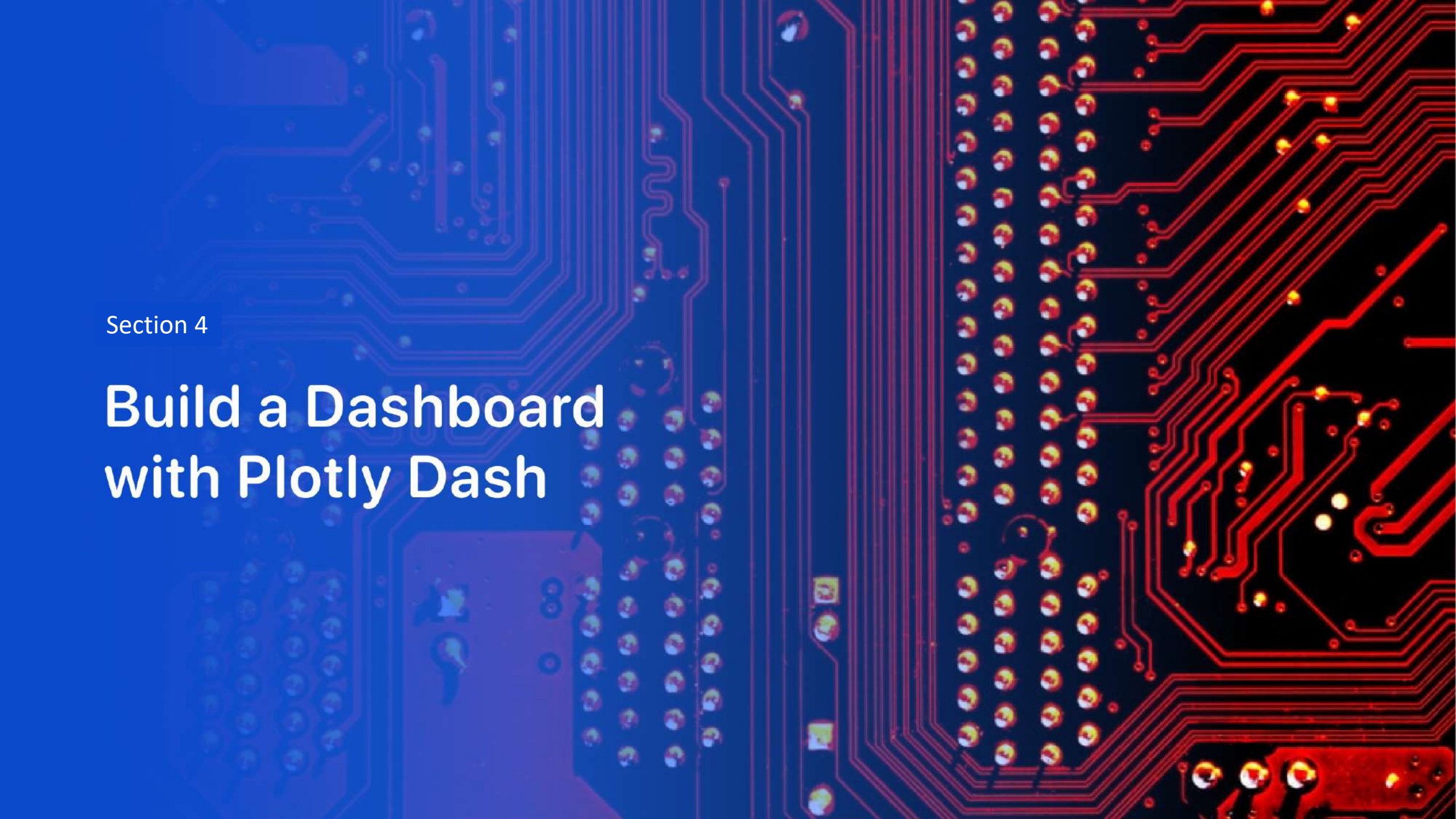




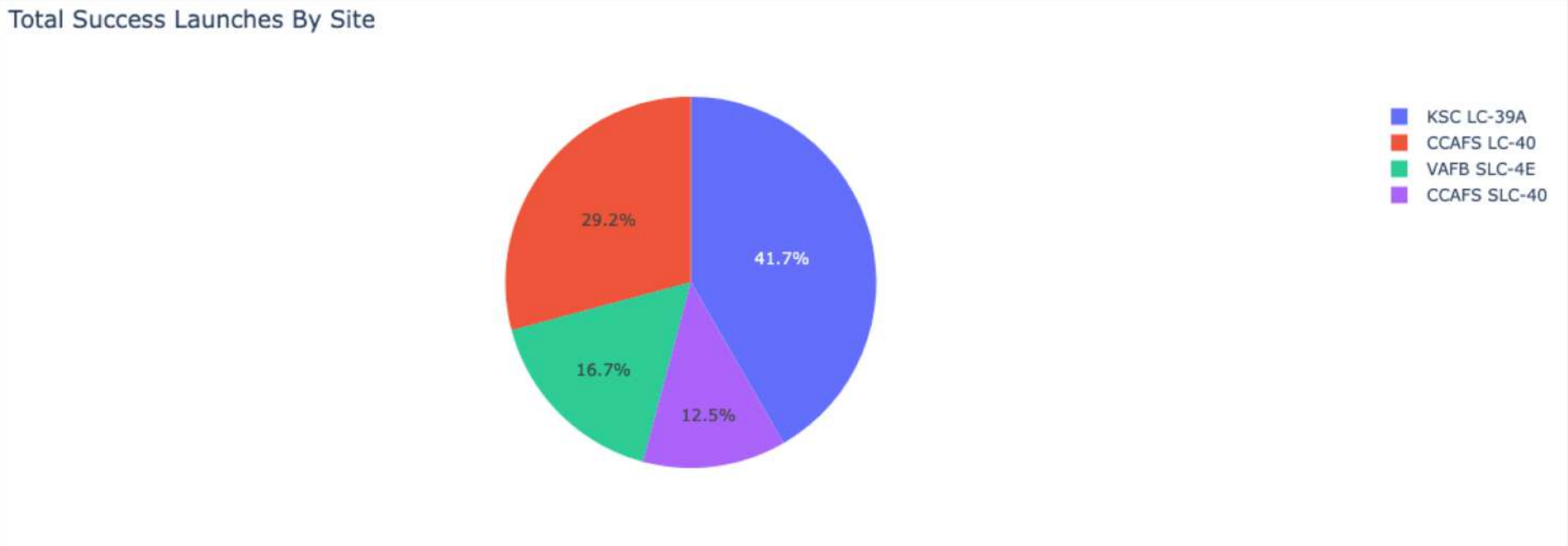
- Are launch sites in close proximity to railways? Yes
- Are launch sites in close proximity to highways? Yes
- Are launch sites in close proximity to coastlines? Yes
- Do launch sites keep certain distance away to cities? Yes

Section 4

Build a Dashboard with Plotly Dash

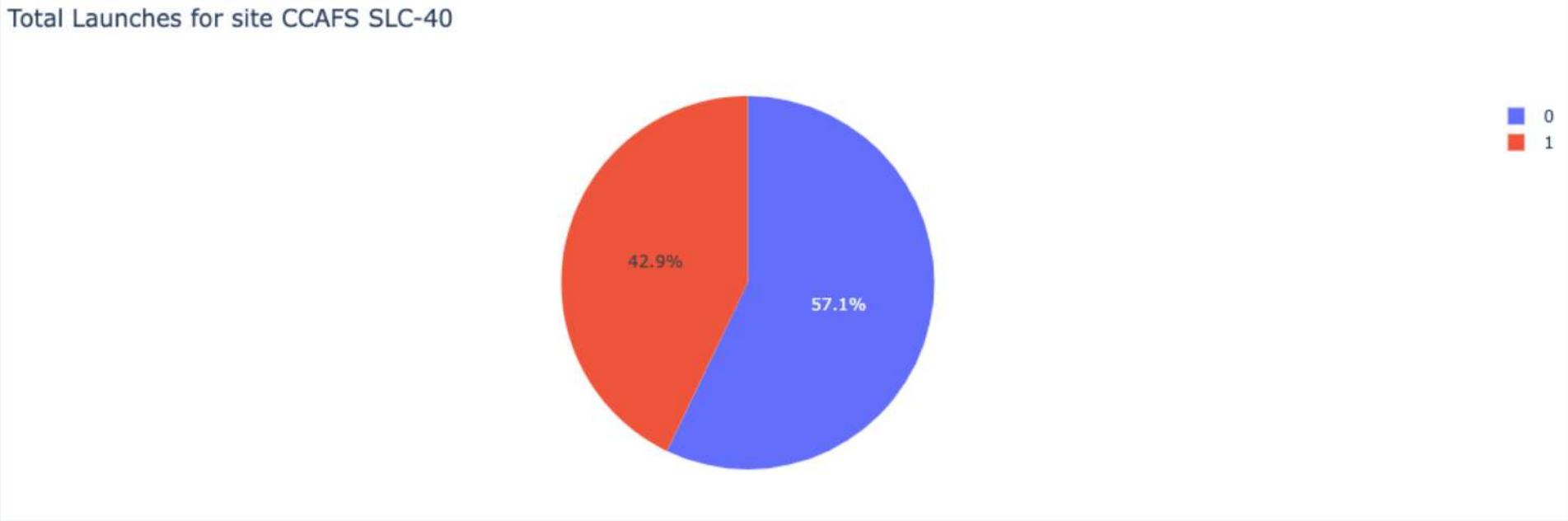


Total Success Launches by Site



- Shows the total success launches proportion for all sites.

Highest Success Rate

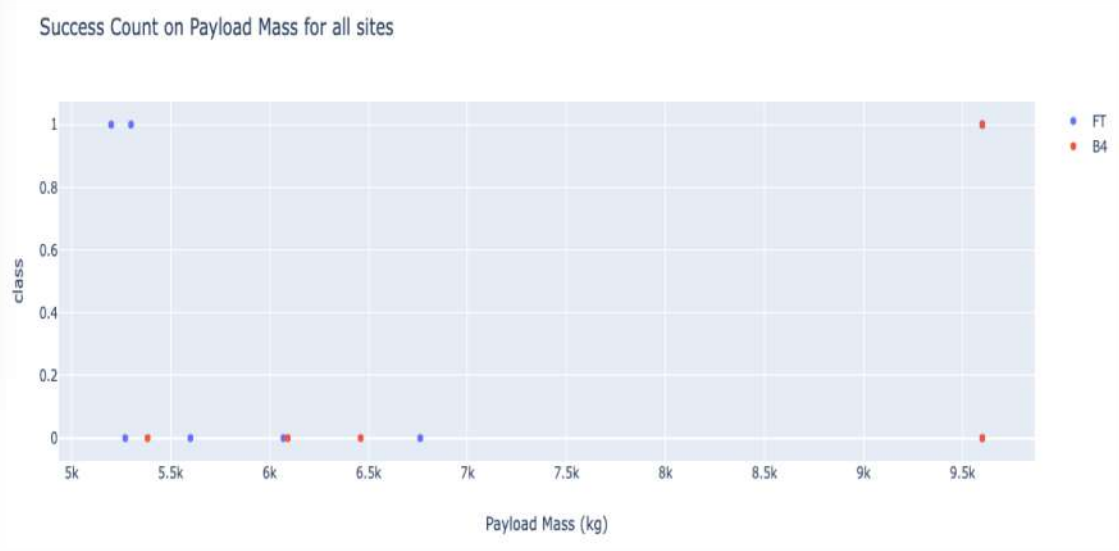


- Shows proportion of success/failure for the site with highest success rate

Success Count on Payload Mass for All Sites

Payload mass 0-5000 kg

Payload mass 5000-10000 kg

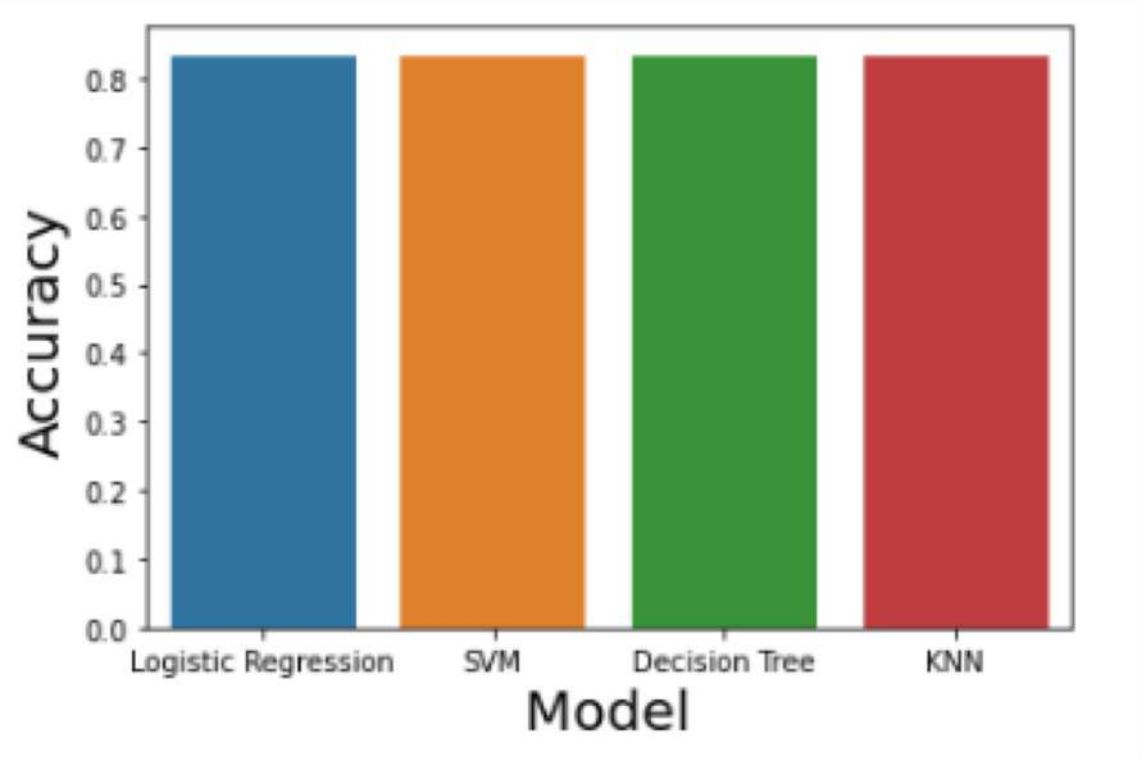


- Shows the relation of success/failure and the payload mass for all sites. The booster version is shown by the different colors.

Section 5

Predictive Analysis (Classification)

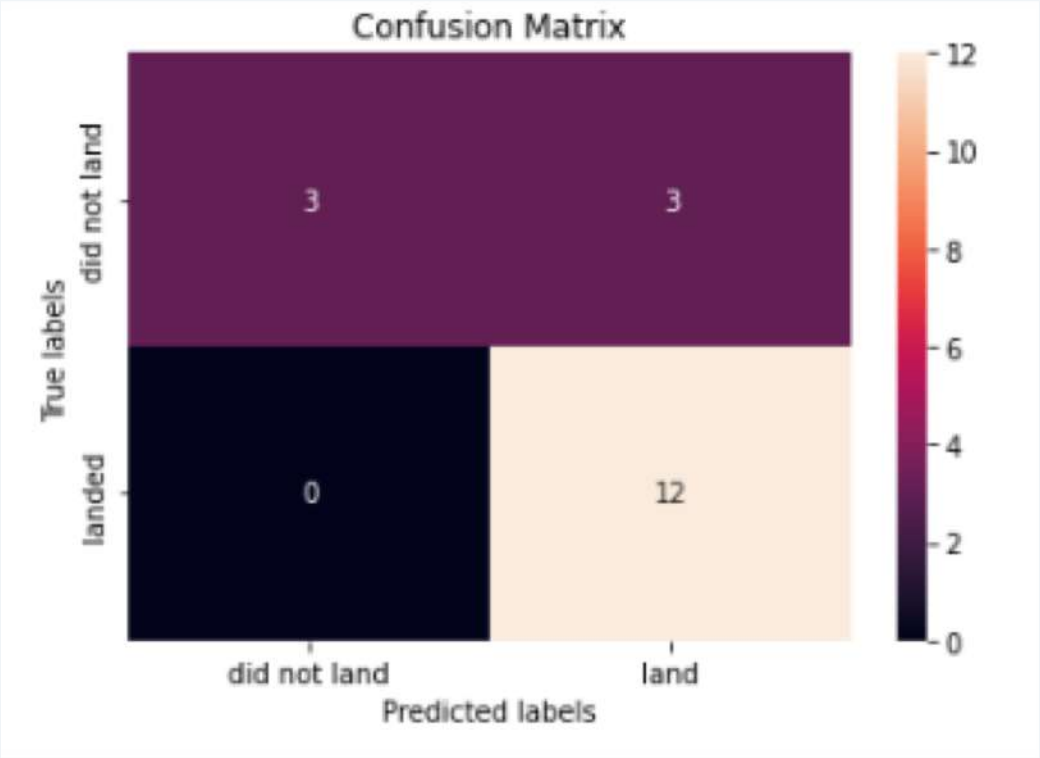
Classification Accuracy



- All classification models presented the same accuracy

Confusion Matrix

- All tested models presented the same accuracy and the same confusion matrix, shown below



Conclusions

- All classifiers presented the same accuracy to predict the outcome of future missions.
- The success rate for SpaceX missions increase with the number of launches over time.
- Launch site CCAFS SLC-40 had the highest rate of successful landing outcomes.
- Orbits ES-L1, GEO, HEO, SSO presented the highest chance of successful landings.

Thank you!