IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

EL BARHICHI Mohammed
29th July, 2023

Outline

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

Executive Summary

- Using data from multiple sources, cleaning it, and processing it using visualizations, SQL, python dash and machine learning, we were able to analyze the publicly available data regarding SpaceX launch sites in order to come to a conclusion regarding whether taking on SpaceX is a wise decision or not.
- The main finding of this report is that given the data, we should not bet against SpaceX as it is bound to succeed.

Introduction

- This project aims to find out whether SpaceX's next launch will be a success or not.
- We have used the key concepts learned in our Data Science professional certification in order to collect, process and interpret the data using various tools such as python, sql, plotly dash, machine learning etc.
- Learning whether the next launch will be a success or a failure will help us decide if competing with spaceX is a good idea or not.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API and webscraping
- Perform data wrangling
  - Data was processed using Jupyterlite in python. Pandas library was used to visualize the data and help in wrangling.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We built, tuned, evaluated classification models using scikit learn on python.

Data Collection

- We collected data through information available at spaceX API and cross-referenced it with data taken through web-scrapping the relevant Wikipedia page.
- Past launch data was obtained through get request and the resulting json file was converted to pandas dataframe
- It was also obtained through web scraping Wikipedia using Beautiful Soup.

Data Collection – SpaceX API

- GitHub URL: https://github.com/qzemad/DSCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Requested data from SpaceX API using GET request

API was used for information about each launch such as rocket type, payload etc

Filtered the data to include information about falcon 9 only

Dealt with missing values of payload by replacing them with mean

Data Collection - Scraping

- GitHub URL: https://github.com/qzemad/DSCapstone/blob/main/jupyter-labs-webscraping.ipynb

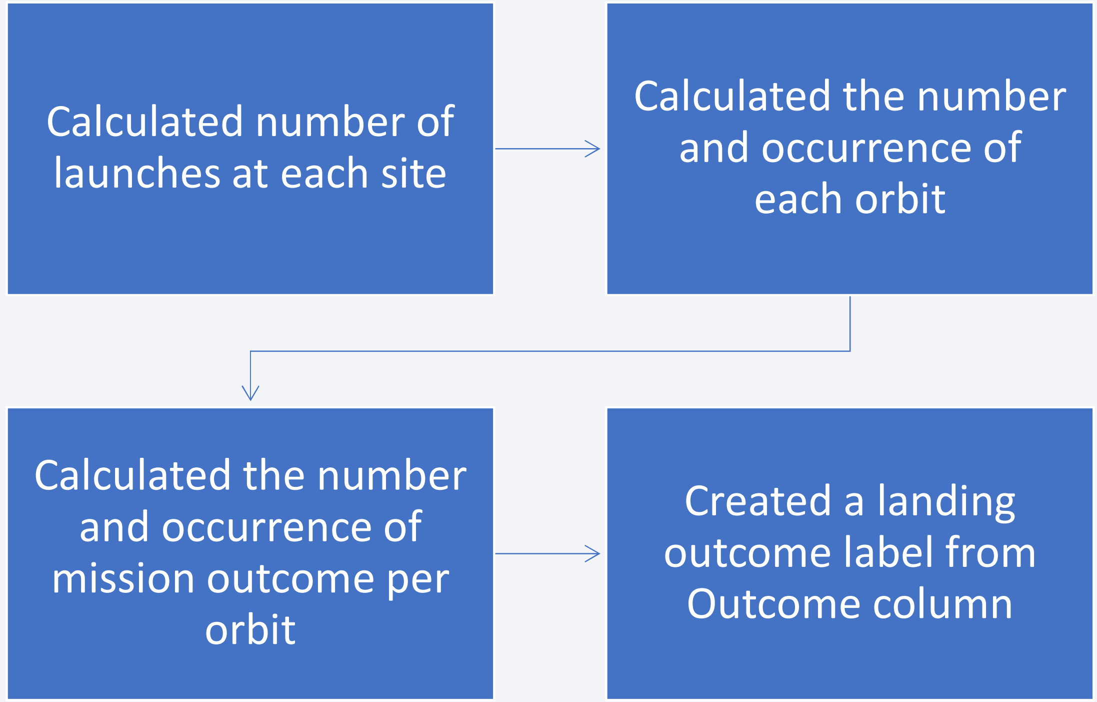
Requested Falcon 9 launch from Wikipedia using GET

Extracted column names and table headers using beautiful soup

Created a dataframe by parsing the launch HTML table.

Data Wrangling

- GitHub URL: https://github.com/qzemad/DSCapstone/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization (1)

- Charts plotted:
  - Payload Mass vs Flight Number (Scatter)
    - This showed that the more massive payloads were only attempted later on, and they had a better success rate
  - Launch Site vs Flight Number(Scatter)
    - Showed how distribution of flight number is across launch sites
  - Launch Site vs Payload Mass (Scatter)
    - Showed how distribution of flight number is across launch sites
  - Orbit vs class (bar)
    - Showed success rates for different orbits

EDA with Data Visualization (2)

- Charts plotted:
  - Orbit vs Flight Number (Scatter)
    - Showed which orbits were attempted when.
  - Orbit vs Payload Mass (Scatter)
    - Showed distribution of payload mass across orbits
  - Class vs Date (line graph)
    - Showed the average success rate and how it changed over time.
- GitHub URL: https://github.com/qzemad/DSCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- SQL queries performed:
  - Unique launch sites
  - Launch records beginning with ‘CCA’
  - Total payload mass carried by boosters launched by NASA
  - Average payload carried by Booster version f9 v1.1
  - Date of first successful launch on a ground pad
  - Names of boosters with success in drone ship with payload mass between 4000 and 6000
  - Total number of successful and failed outcomes
  - Names of boosters that have carried maximum payload
  - Record of success and failure in 2015
  - Rank landing outcomes in descending order
- GitHub URL: https://github.com/qzemad/DSCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

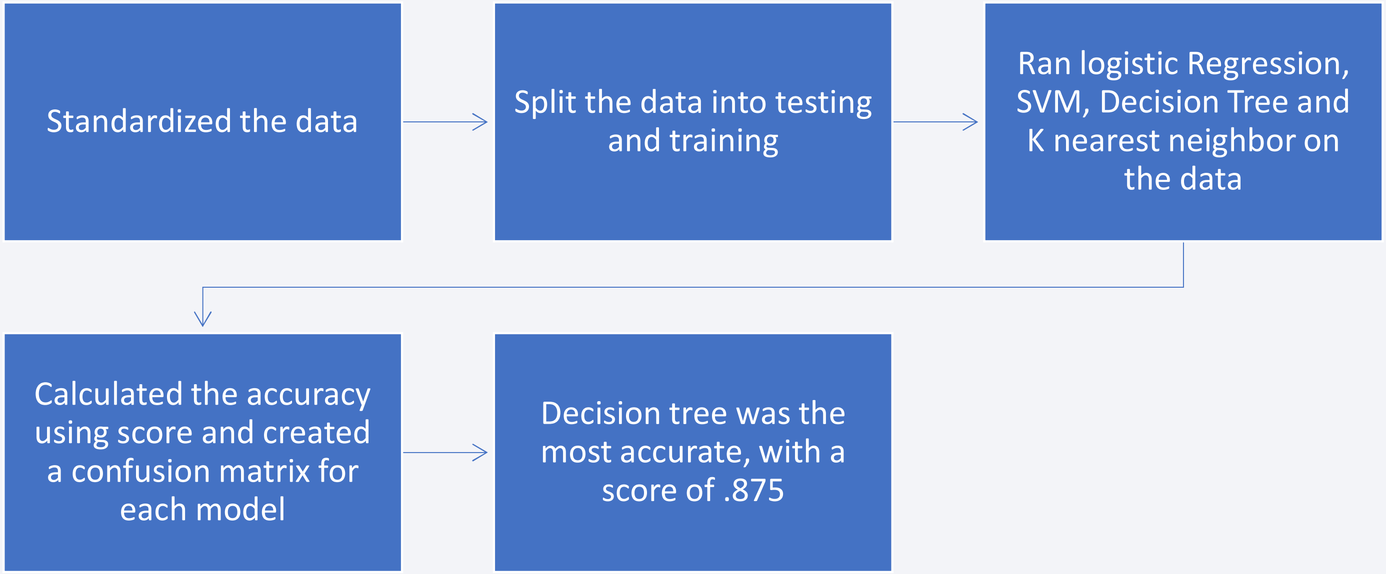
- Created a map that showed launch site locations
- Tried to add clusters but failed
- GitHub URL: https://github.com/qzemad/DSCapstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- A pie chart which showed successful attempts vs failed attempts at each site.
  - Users can explore and compare success rates at different sites.
- A scatter graph for successful attempts for various payload masses.
  - Users can explore different ranges of payload masses and their success or failures
- GitHub URL: https://github.com/qzemad/DSCapstone/blob/main/plotly%20dash%20script%20final.txt

Predictive Analysis (Classification)

- GitHub URL: https://github.com/qzemad/DSCapstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

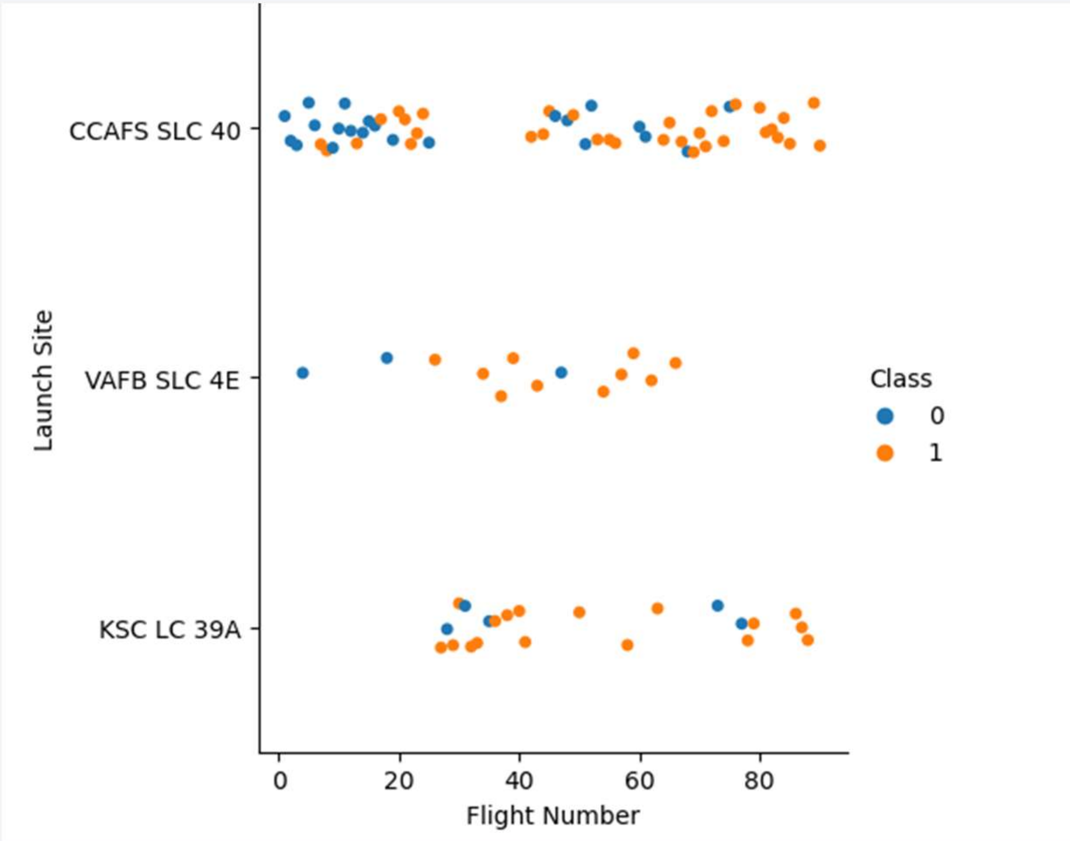
- The next slides will show:
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Section 2

Insights drawn from EDA

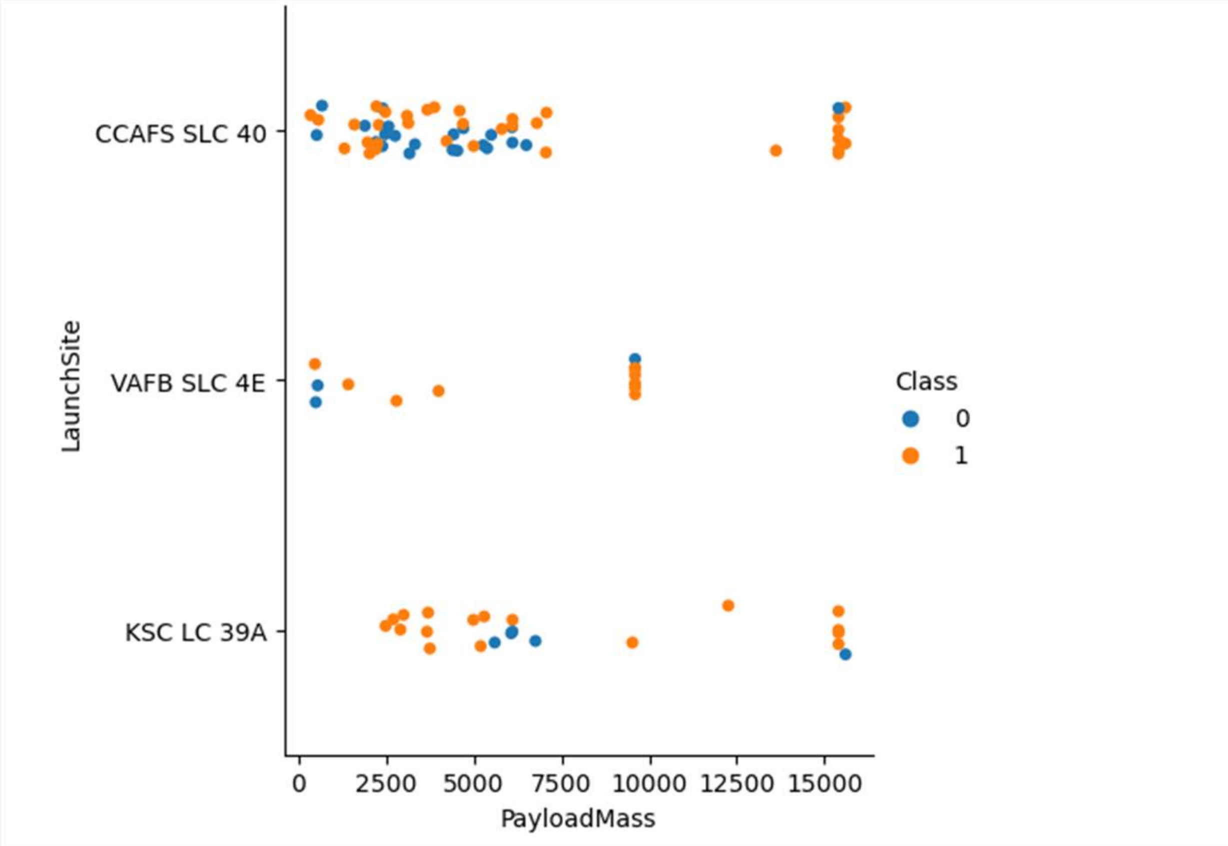
Flight Number vs. Launch Site

- This shows when each of the launch sites was used and their outcomes.
- We see that CCAFS SLC 40 was where most of the initial testing was performed, and we see a concentration of failed outcomes at the beginning, with more successes as more flights were tested.
- The other sites have a comparatively cleaner record.



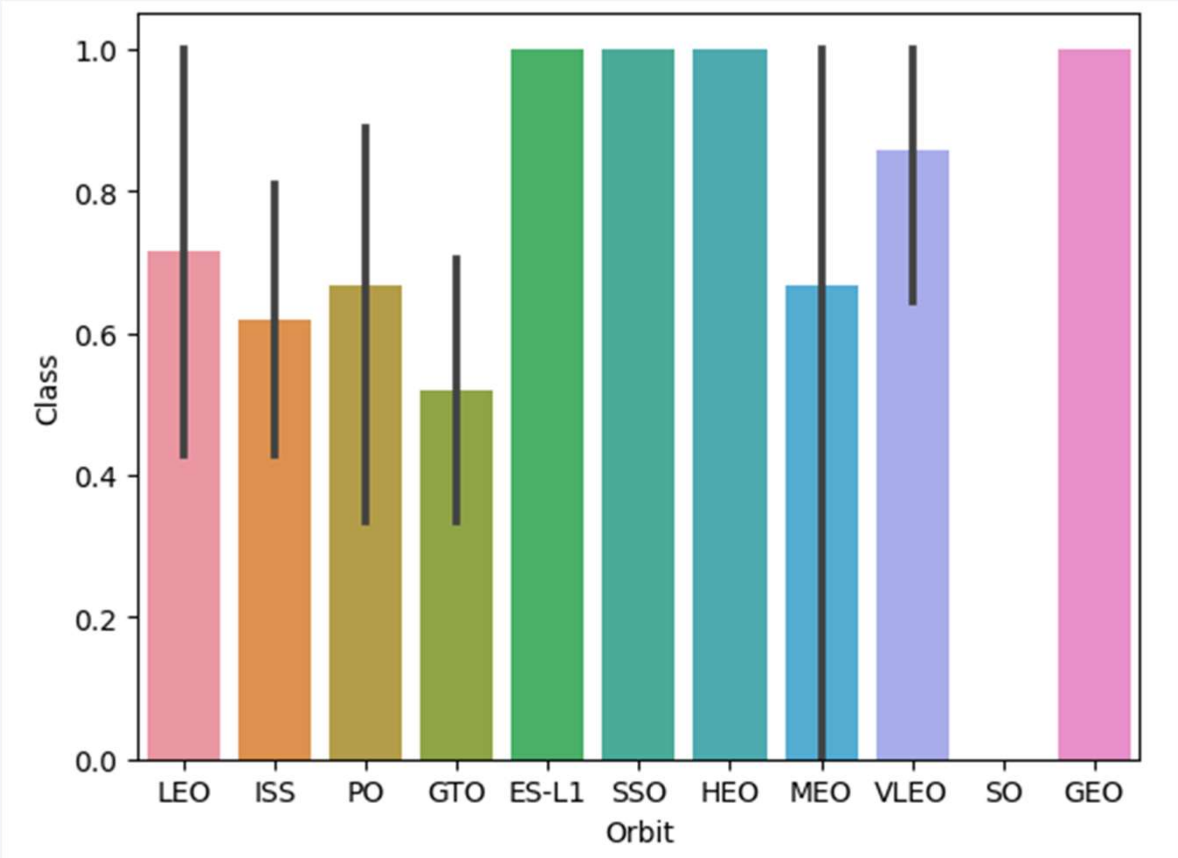
Payload vs. Launch Site

- This plot shows how much payload was sent through each launch site and its outcome.
- We see that massive payloads were either sent from CCAFS SLC 40 or from KSC LC 39A. With VAFB SLC 4E only sending payloads of 10000 kgs
- We also see how most of the failures were for low mass payloads



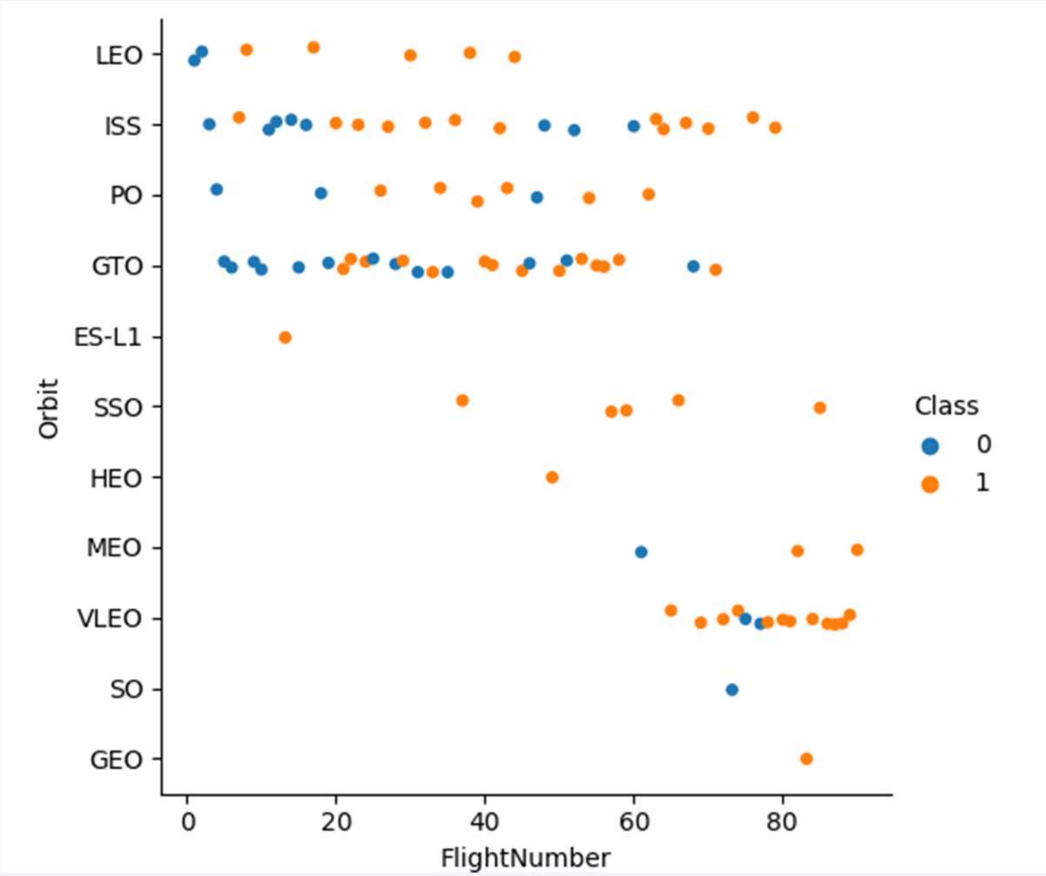
Success Rate vs. Orbit Type

- This bar chart shows the success rates for different orbit launches
- ES-L1, SSO, HEO and GEO orbits have the maximum success rate of 1
- There have been no successful launches at SO orbit.



Flight Number vs. Orbit Type

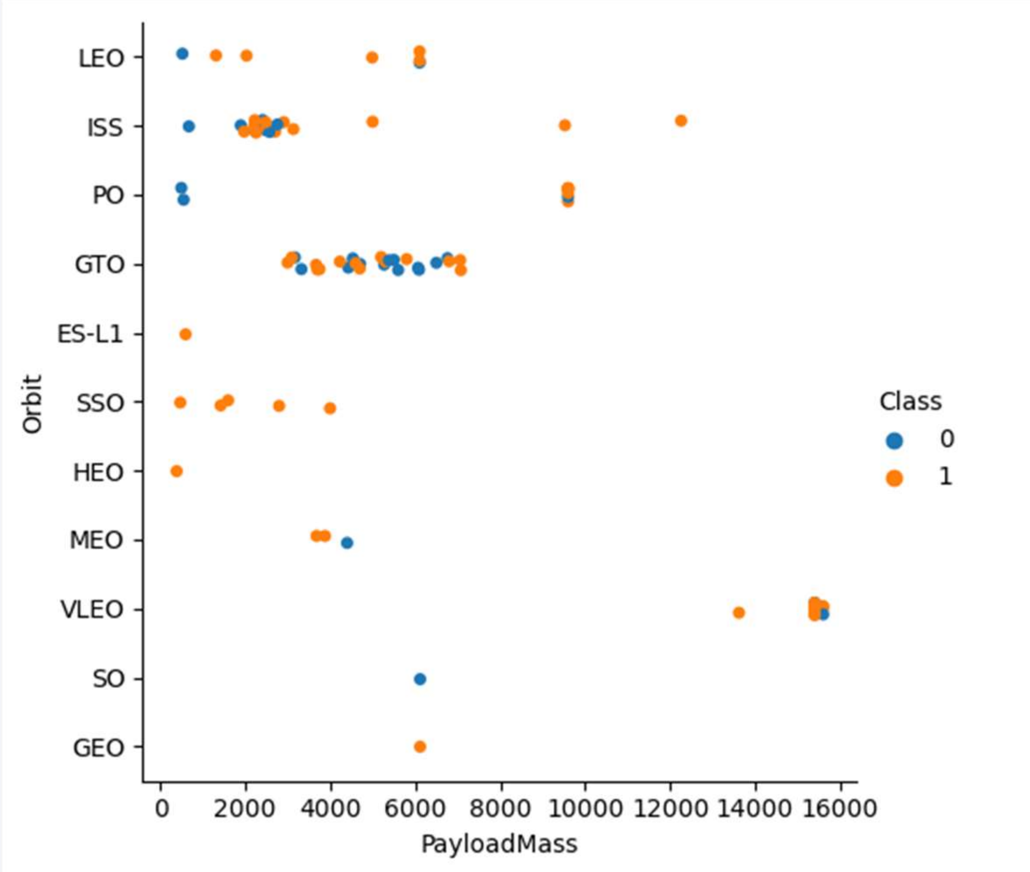
- Different orbit launches and their flight numbers.
- Clearly shows how intial flights were sent to LEO, ISS, PO and GTO whereas other orbits, such as SSO and VLEO were only attempted later on.



Payload vs. Orbit Type

- Shows a scatter point of payload vs. orbit type

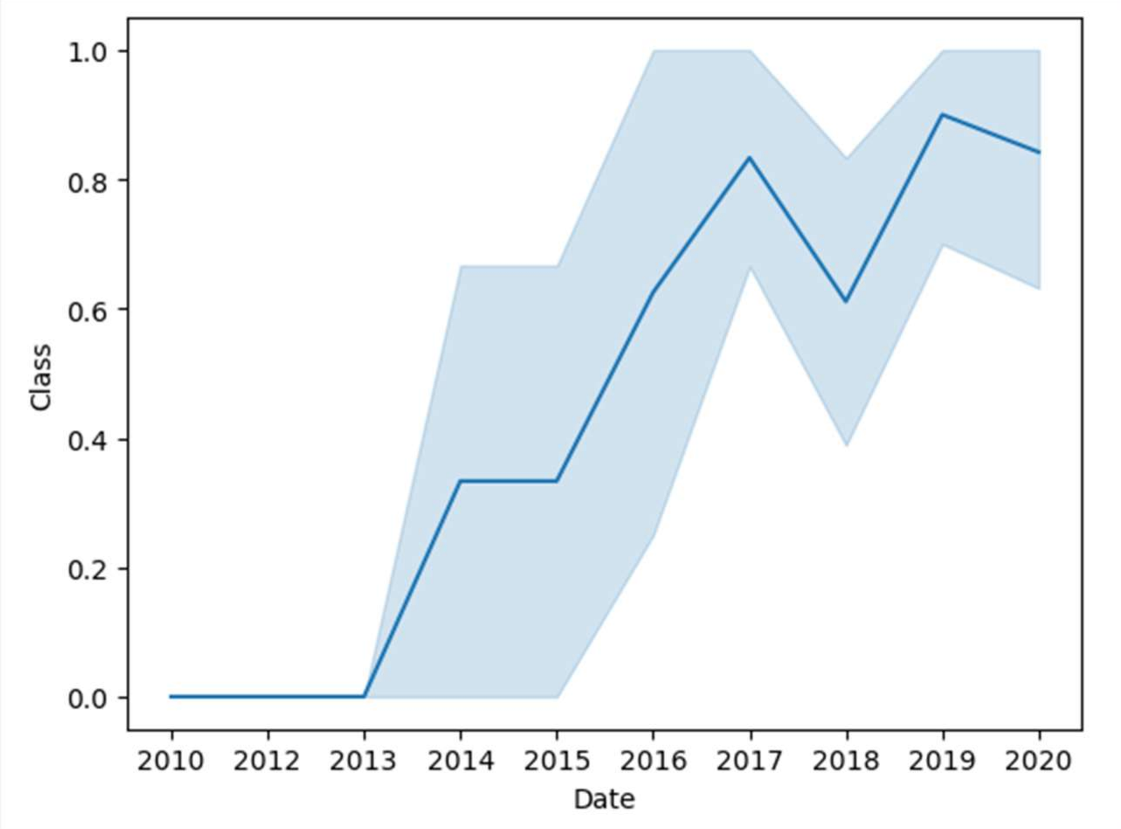
- We can see that more massive payloads were generally sent to VLEO orbit whereas other orbits tend to have less massive payloads sent to them



Launch Success Yearly Trend

- Shows a line chart of yearly average success rate

- Success rate has been climbing steadily and approaching 100 percent over time.



All Launch Site Names

- The names of the unique launch sites in the dataset:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No att
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No att
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No att

Total Payload Mass

- 45596 KGs were carried by boosters from NASA

```
[22]: %sql select sum(PAYLOAD_MASS__KG_) as total_mass from spacextbl where customer = "NASA (CRS)"

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

[22]:

total_mass
45596.0

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

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

```
]:  %sql select avg(payload_mass__kg_) as average_mass from spacextbl where booster_version like "F9 v1.1%"
```

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

]:

average_mass
2534.6666666666665

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

[40]:
```
%sql select min(date) from spacextbl where landing_outcome = "Success (ground pad)"
```

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

[40]:

min(date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [32]:
```
%sql select distinct booster_version from spacextbl where payload_mass__kg_>4000 and payload_mass__kg_ < 6000
```

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

Out[32]:

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(mission_outcome) from spacextbl group by mission_outcome
```

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

Mission_Outcome	count(mission_outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

```
%sql select booster_version, payload_mass__kg_ from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql select substr(date, 4, 2) as month, landing_outcome, booster_version, launch_site from spacextbl
where substr(date,7,4)="2015" and landing_outcome = 'Failure (drone ship)'
```

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

month	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

Task 10

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 landing_outcome, count(landing_outcome) as total from spacextbl where date between '04/06/2010' and '20/03/2017
```

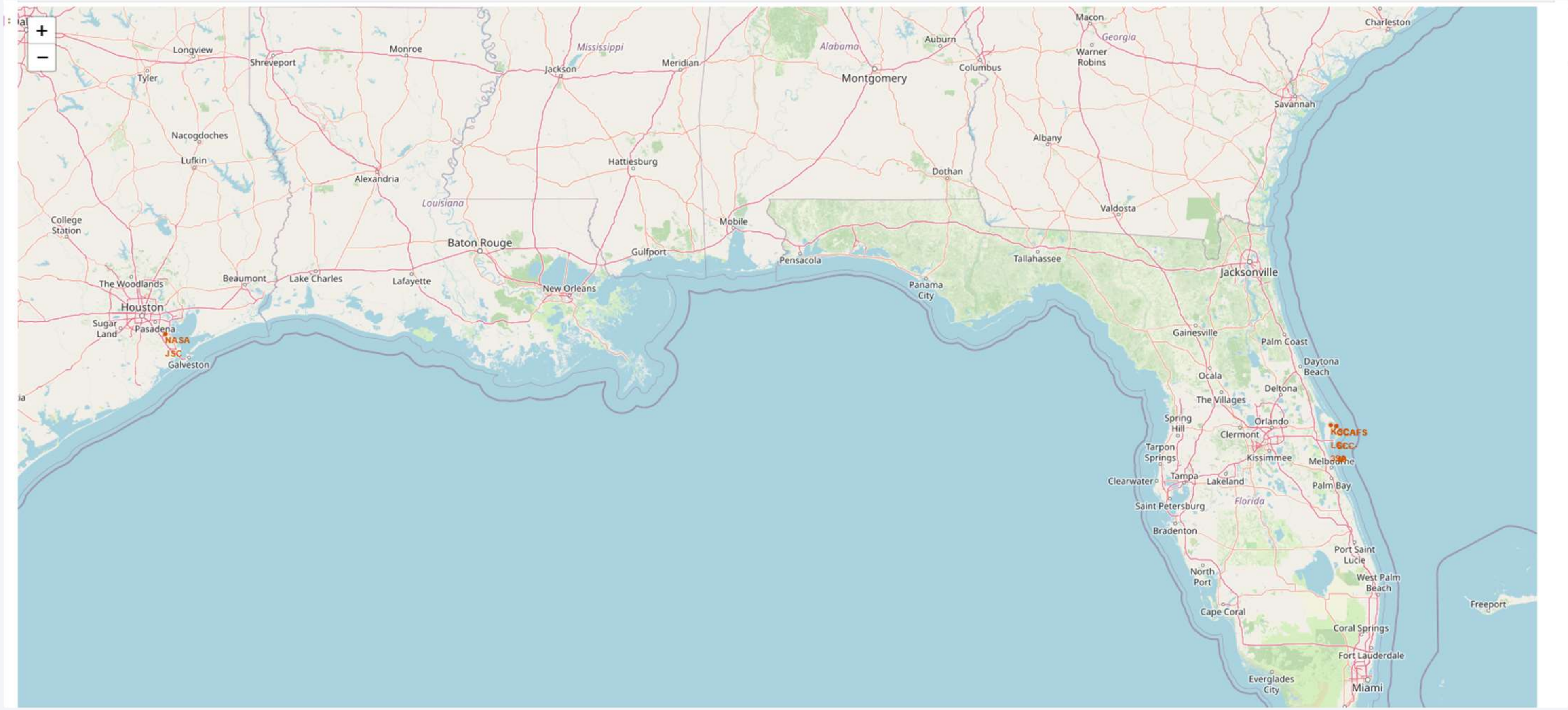
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	total
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

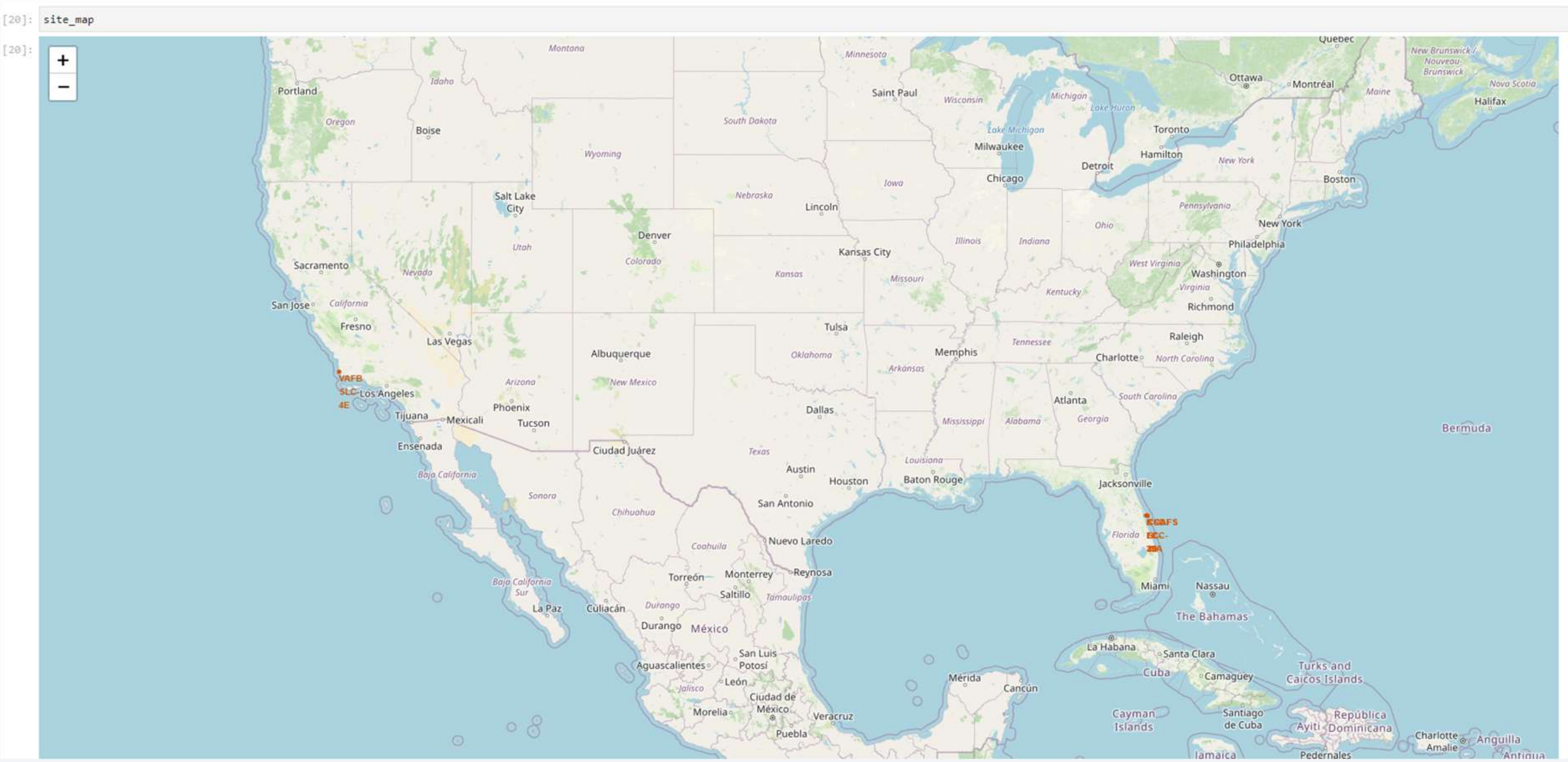
Launch Sites Proximities Analysis

First Folium Map



Second Folium Map showing launch sites

- All launch sites are situated near the equator, however, they are located on both east and west coast of USA



Folium Map Showing clusters of successful and failed attempts

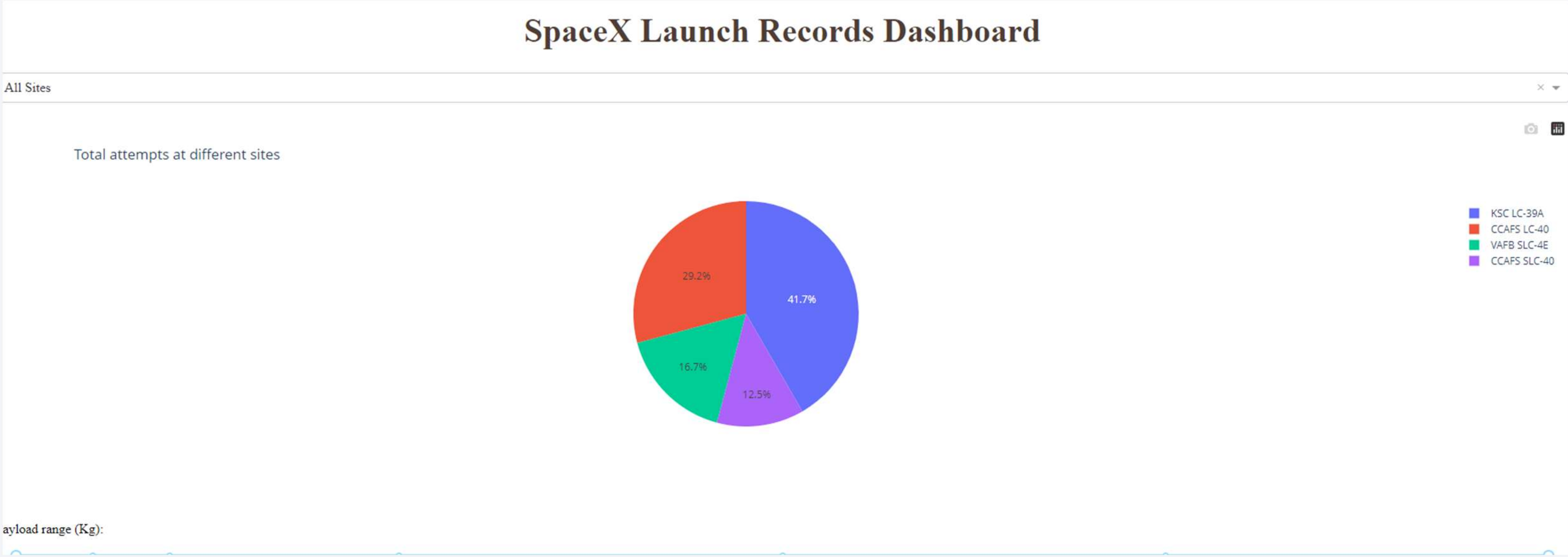
- I failed to produce this map.

Section 4

Build a Dashboard with Plotly Dash

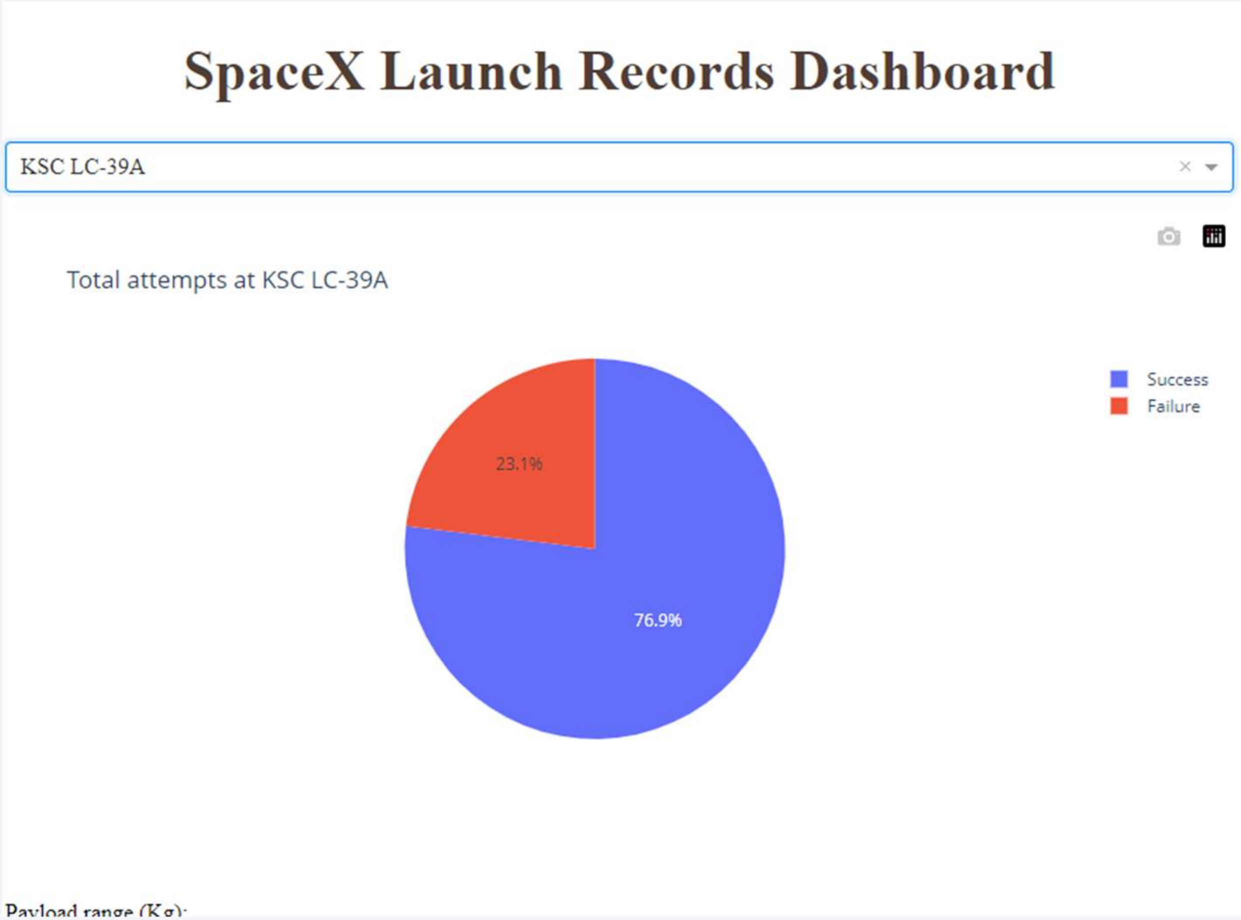
Dashboard Pie chart 1

- Pie chart showing attempts at all the sites is presented here
- KSC LC 39A was the busiest launch site



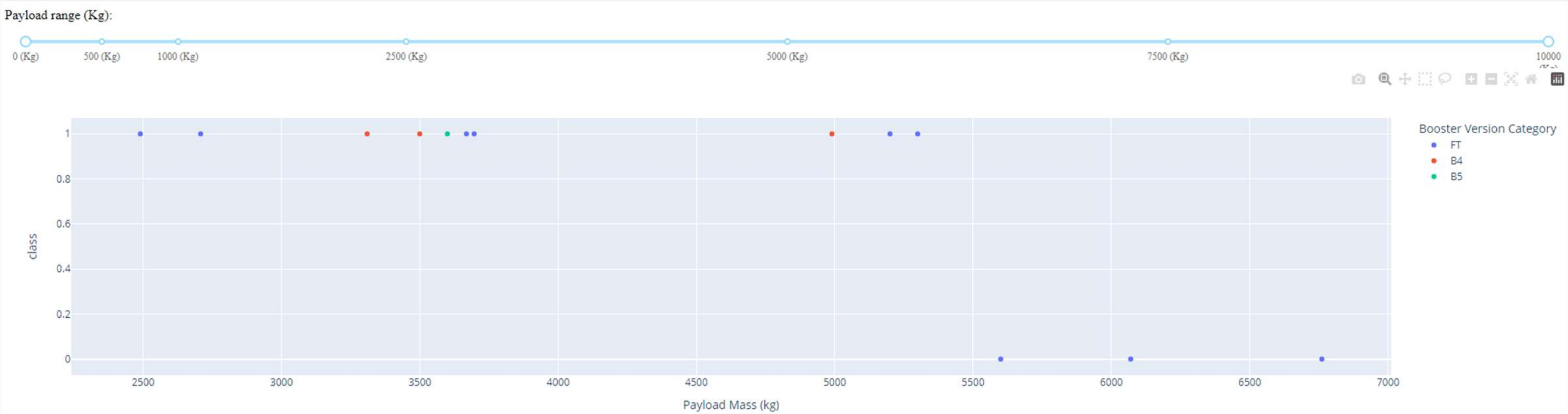
Dashboard Pie chart 2

- piechart for the launch site with highest launch success ratio
- This was
- Explain the important elements and findings on the screenshot



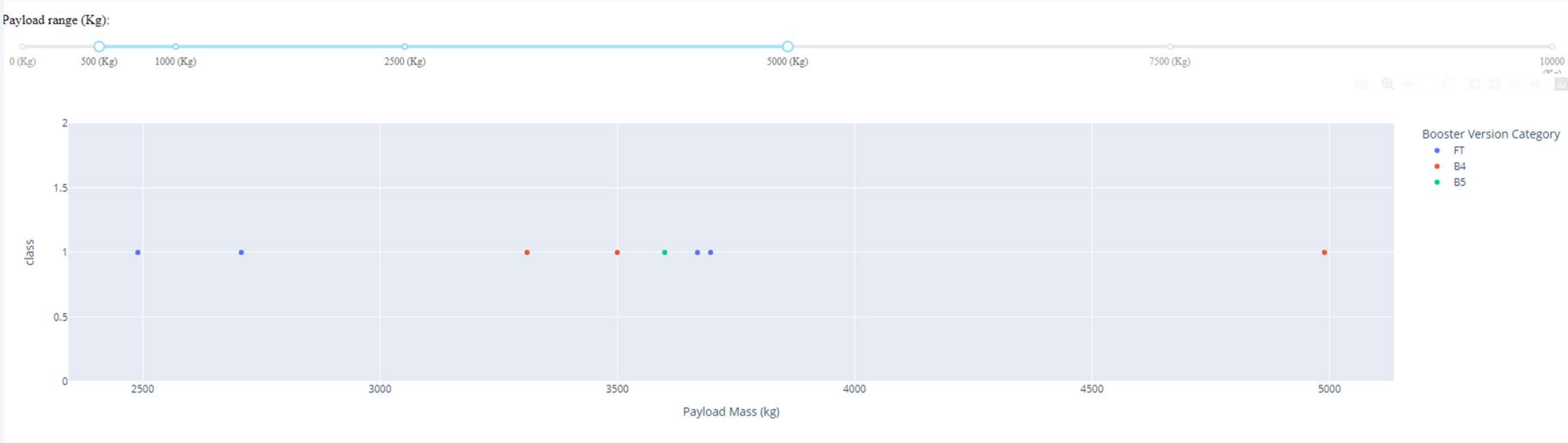
Dashboard scatterplot (1)

- Scatter plot of outcome vs payload
- This shows that booster category FT has the poorest record



<Dashboard scatter plot (2)

- Different range of payload selected:



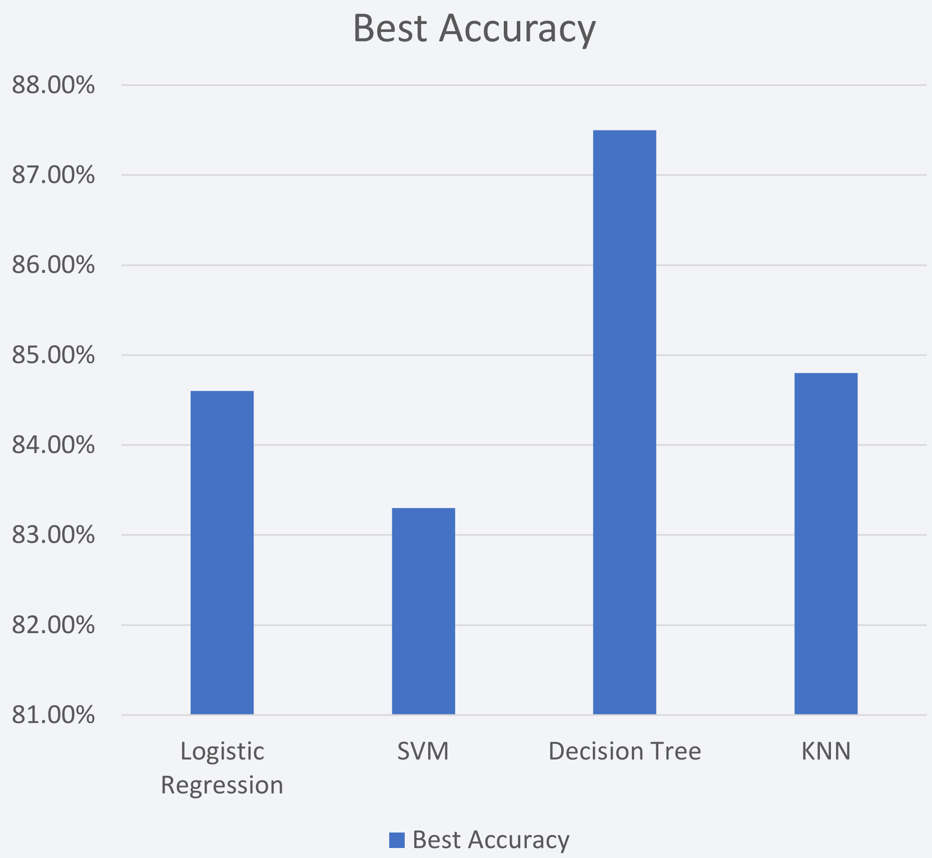
Section 5

Predictive Analysis (Classification)

Classification Accuracy

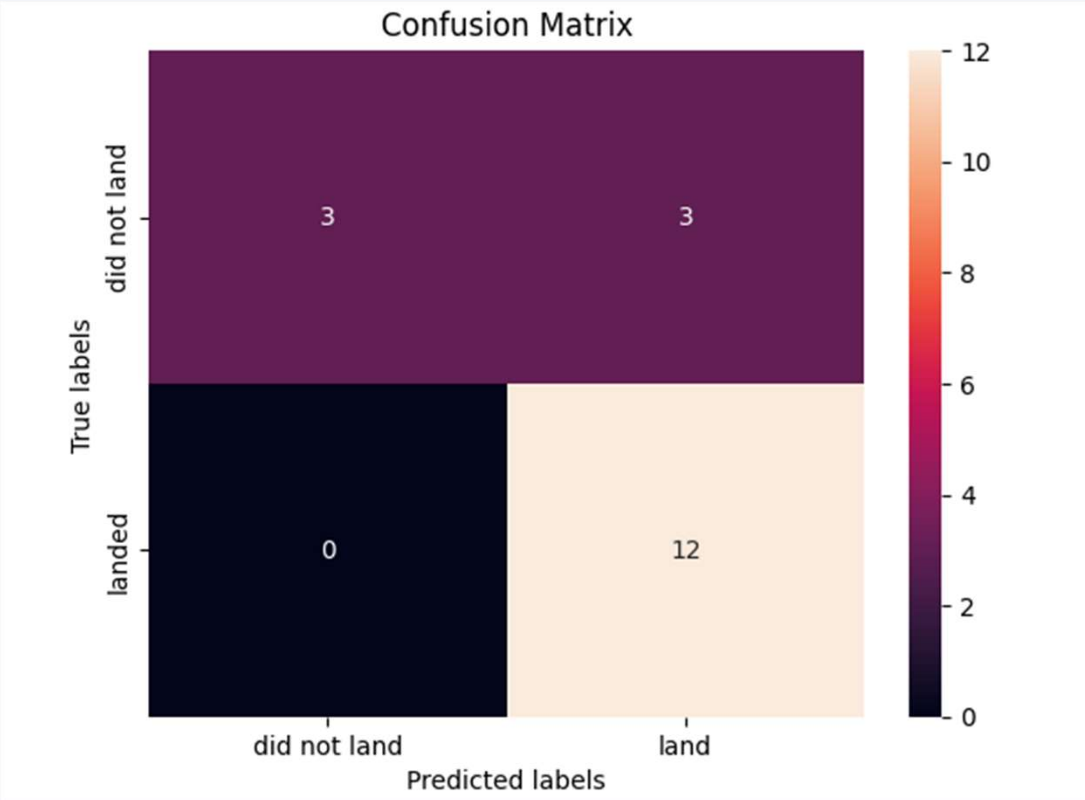
- Built models' accuracy for all built classification models, in a bar chart

- Decision tree showed the highest accuracy.



Confusion Matrix

- Confusion matrix of decision tree



Conclusions

- SpaceX has been continually improving its success rates.
- Specially for the more massive payload in the VLEO orbit, the success rate is extremely impressive.
- Data shows a clear trend towards improvement, so I wouldn't bet against SpaceX.
- Better to stick to some other business!

Appendix

- Text file for Plotly dashboard is provided in the github repository.

```
# Import required libraries
import pandas as pd
import dash
import dash_html_components as html
import dash_core_components as dcc
from dash.dependencies import Input, Output
import plotly.express as px

# Read the airline data into pandas dataframe
spacex_df = pd.read_csv("spacex_launch_dash.csv")
max_payload = spacex_df['Payload Mass (kg)'].max()
min_payload = spacex_df['Payload Mass (kg)'].min()

# Create a dash application
app = dash.Dash(__name__)

# Create an app layout
app.layout = html.Div(children=[html.H1('SpaceX Launch Records Dashboard',
                                        style={'textAlign': 'center', 'color': '#503D36',
                                               'font-size': 40}),
                                # TASK 1: Add a dropdown list to enable Launch Site selection
                                # The default select value is for ALL sites
                                dcc.Dropdown(id='site-dropdown', options=[{'label': 'All Sites', 'value': 'ALL'},
                                                                          {'label': 'CCAFS LC-40', 'value': 'CCAFS LC-40'},
                                                                          {'label': 'VAFB SLC-4E', 'value': 'VAFB SLC-4E'},
                                                                          {'label': 'KSC LC-39A', 'value': 'KSC LC-39A'},
                                                                          {'label': 'CCAFS SLC-40', 'value': 'CCAFS SLC-40'},
                                                                          ],
                                                                           value='ALL',
                                                                           placeholder="Please select the launch site",
                                                                           searchable=True),
                                html.Br(),

                                # TASK 2: Add a pie chart to show the total successful launches count for all sites
                                # If a specific launch site was selected, show the Success vs. Failed counts for the site
                                html.Div(dcc.Graph(id='success-pie-chart')),
                                html.Br(),

                                html.P("Payload range (Kg):"),
                                # TASK 3: Add a slider to select payload range
```


Thank you!