IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alin-Constantin Paraschiv
2023-10-10

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected using the SpaceX API and Web Scrapping
  - Exploratory Data Analysis (Data Preprocessing and Data Visualization, including Interactive Visual Analytics)
  - Machine Learning modeling
- Summary of all results
  - Data was collected from public sources
  - Using EDA, I identified the best features to use in the modeling
  - Using ML algorithms I managed to predict the target class

Introduction

- Project background and context
  - The objective is to evaluate the viability of SpaceY, a new competitor to SpaceX
- Problems you want to find answers
  - Where is the best place for launching
  - Estimate the total cost of launches and how to reduce it, that means we need to predict successful landing of the first stage

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX API and from Wikipedia using WebScrapping
- Perform data wrangling
  - A new label called landing_outcome was created based on a previous feature
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - After the collection step was finished, data was normalized, splitted into train/test and fitted to 4 different modes using different grids of parameters for each

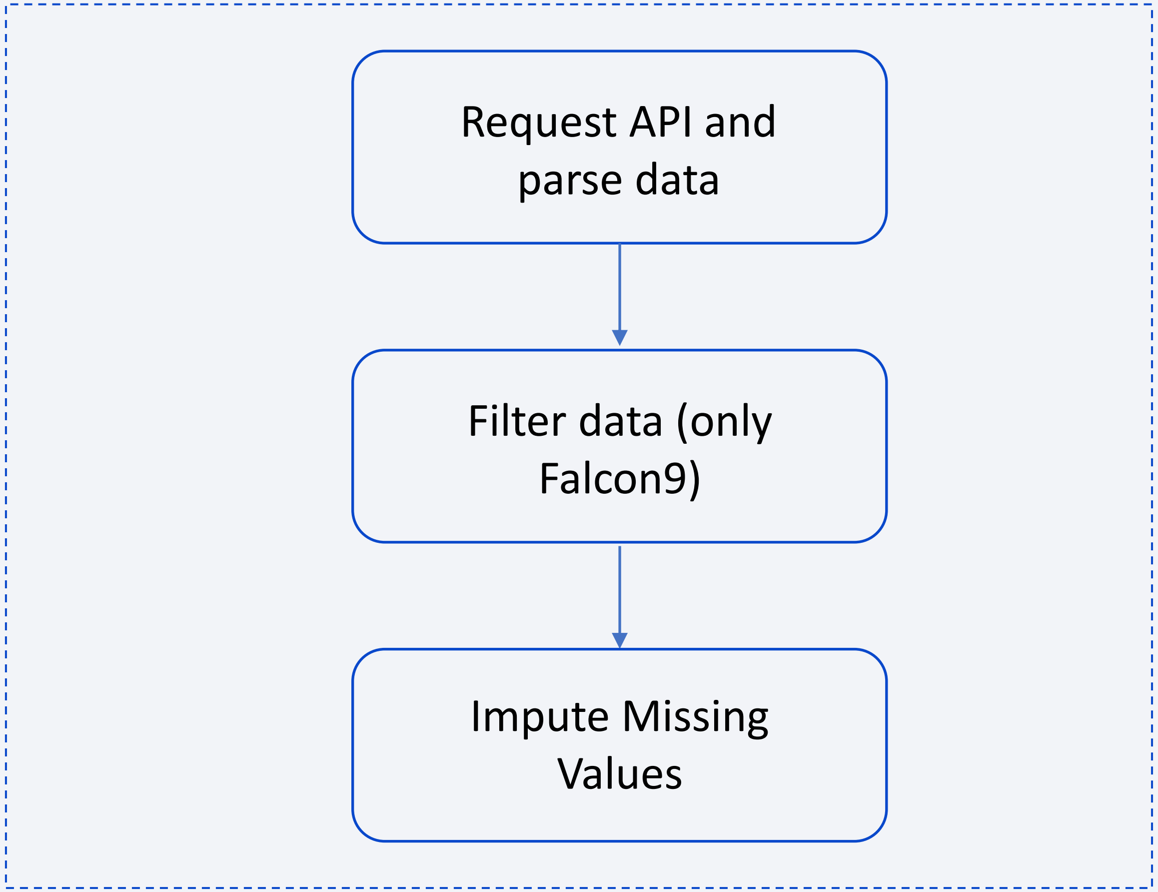
Data Collection

- Data was collected from:
  - SpaceX API -> https://api.spacexdata.com/v4/rockets/
  - Wikipedia -> https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used
- This API was used according to the flowchart beside and then data is persisted.

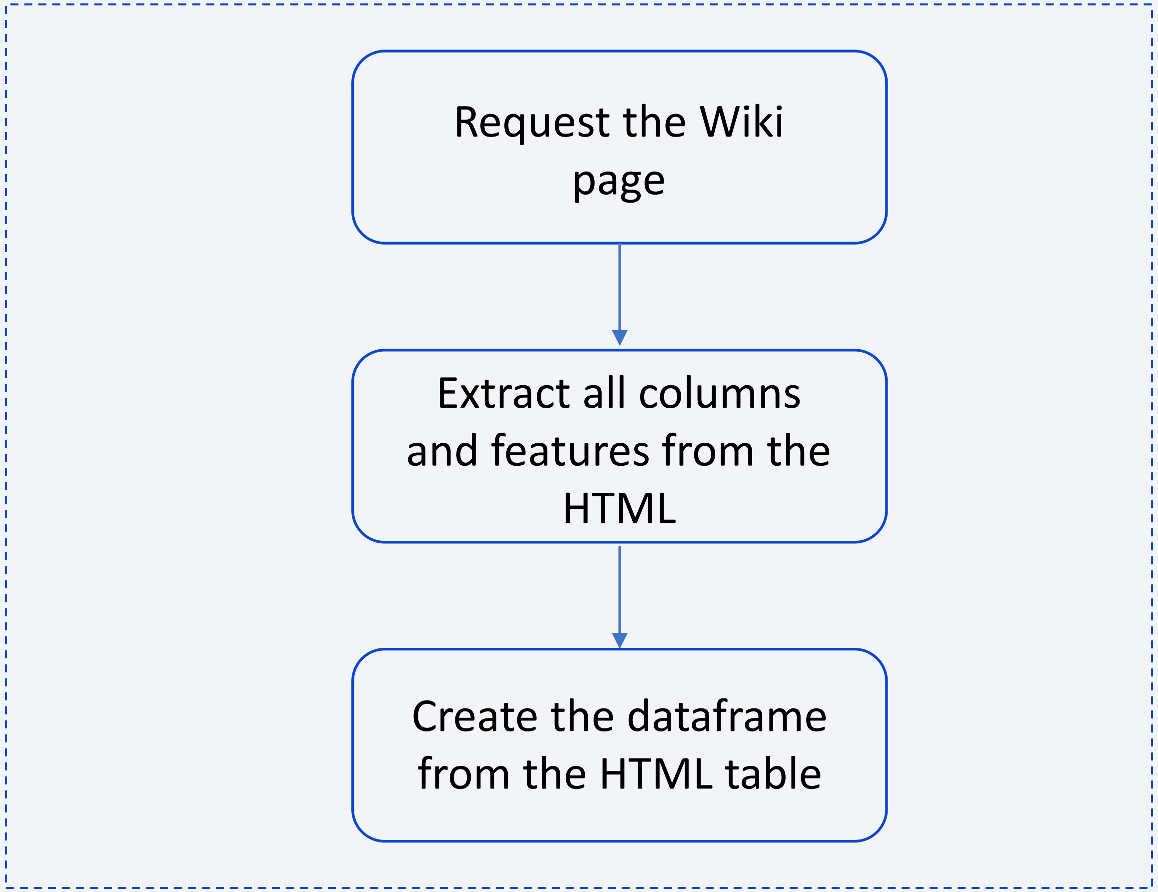
https://github.com/alinparaschiv0107/capstone_Coursera/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Data from SpaceX launches was also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the graph.

https://github.com/alinparaschiv0107/capstone_Coursera/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- I performed Exploratory Data Analysis (EDA) on the dataset to see if anything is not in order.
- Summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- The target, landing outcome label was created from Outcome column.

EDA + Summarizations = Creation of the target

https://github.com/alinparaschiv0107/capstone_Coursera/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- To explore the data, scatterplots and barplots were used to visualize the relationship between pair of features, including
    - Payload Mass vs Flight Number
    - Launch Site vs Payload Mass
    - Launch Site vs Flight Number
    - Orbit vs Flight Number
    - Payload vs Orbit

https://github.com/alinparaschiv0107/capstone_Coursera/blob/main/EDA%20with%20Data%20Viz.ipynb

EDA with SQL

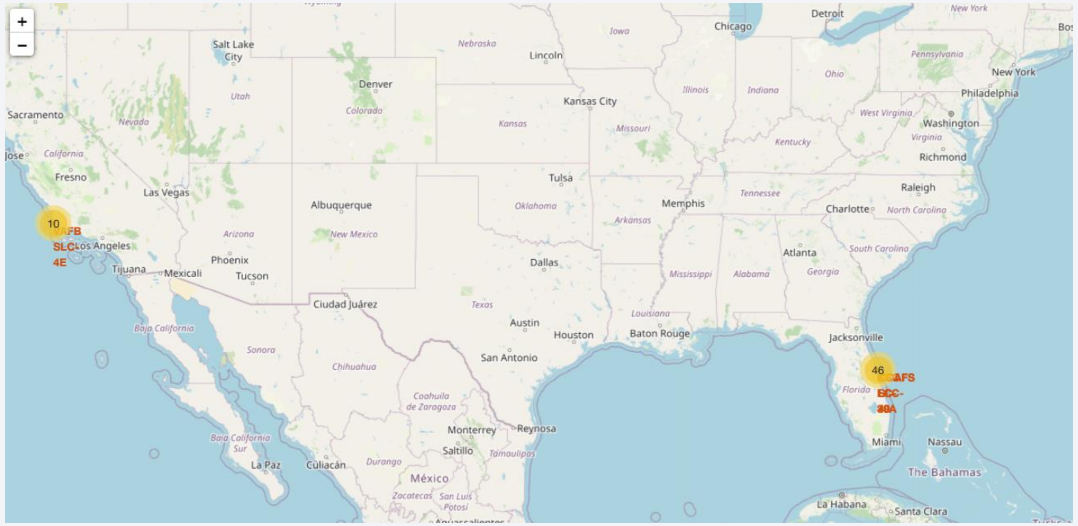
- The following SQL queries were done:
  - Names of the unique launch sites in the space mission
  - Top 5 launch sites whose name begin with the string 'CCA’
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

https://github.com/alinparaschiv0107/capstone_Coursera/blob/main/jupyter-labs-eda-sql-coursera_sqllite-Copy1.ipynb

Build an Interactive Map with Folium

Markers, circles, lines and marker clusters were used with Folium Maps

- Markers indicate points like launch sites
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site
- Lines are used to indicate distances between two coordinates



https://github.com/alinparaschiv0107/capstone_Coursera/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

The following graphs and plots were used to visualize data

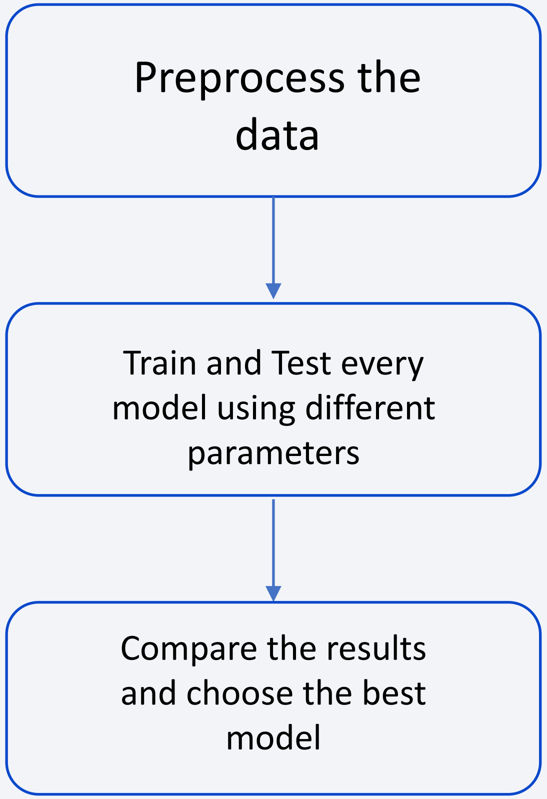
- Percentage of launches by site
- Payload range

I used those 2 plot to underline the relationship between payloads and launch sites, helping to identify best launching place according to payloads.

https://github.com/alinparaschiv0107/capstone_Coursera/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- I tried four different classification models using a different grid of parameters used as parameter in the GridSearchCV method:
    1. Logistic Regression
    2. Support Vector Machines
    3. Decision Trees
    4. KNN



https://github.com/alinparaschiv0107/capstone_Coursera/blob/main/SpaceX%20Machine%20Learning%20Prediction.ipynb

Results

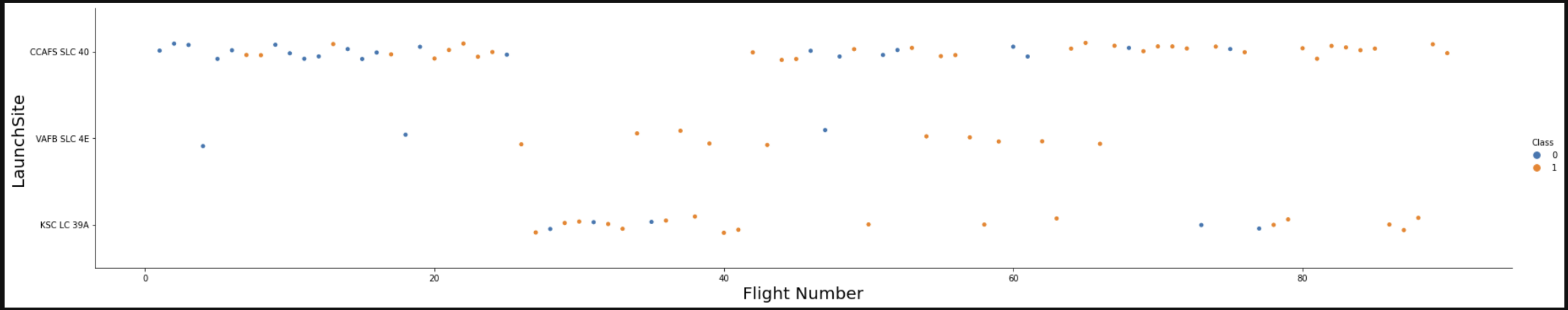
- Space X uses 4 different launch sites
- The first launches were done to Space X itself and NASA
- The first success landing outcome happened in 2015 fiver year after the first launch
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
- Almost 100% of mission outcomes were successful
- The number of landing outcomes increased as years passed
- Most launches are from sites near water
- The best model used to model the data is the decision tree which achieved over 94% accuracy on test data

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

```
In [7]: # Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
        sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
        plt.xlabel("Flight Number",fontsize=20)
        plt.ylabel("LaunchSite",fontsize=20)
        plt.show()
```

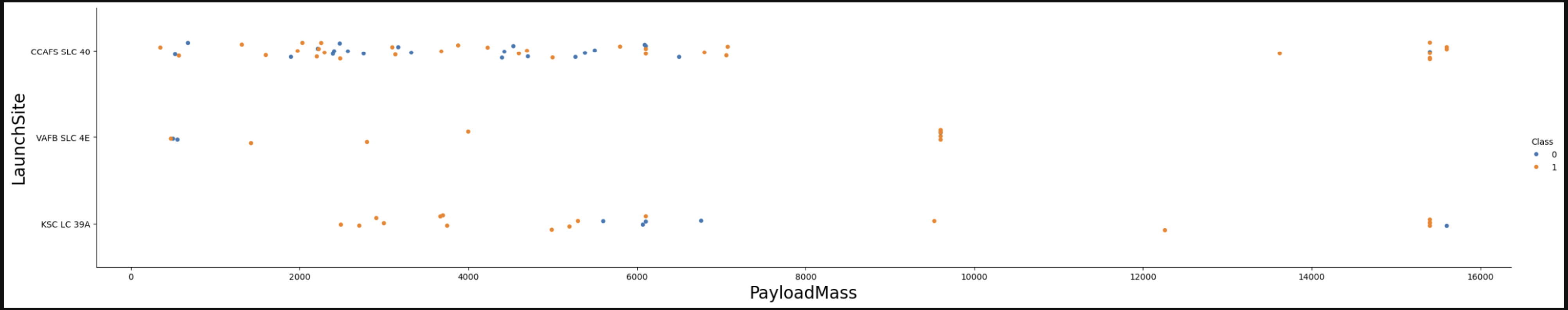


Success rate improved over time

Payload vs. Launch Site

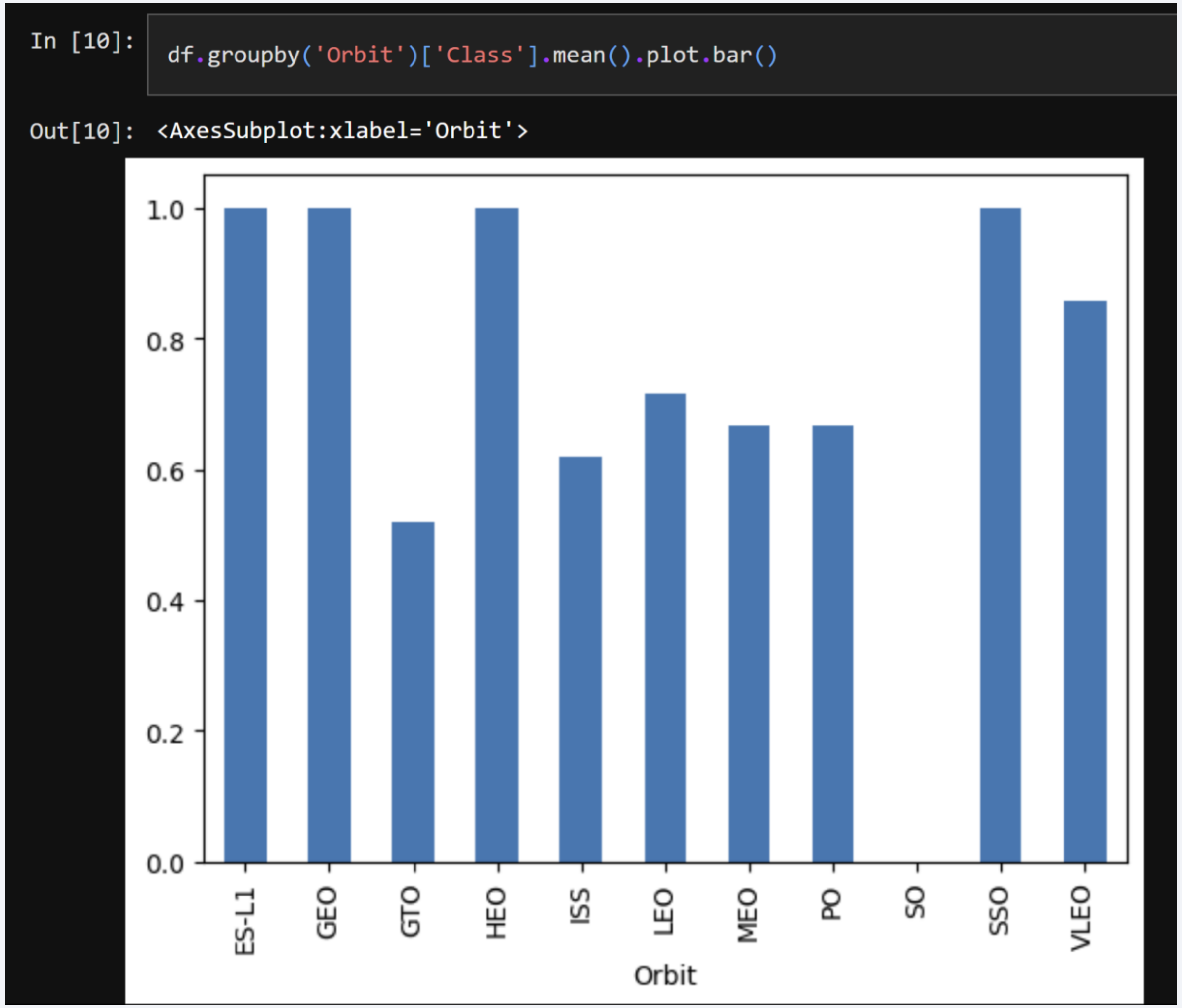
In [8]:

```
### TASK 2: Visualize the relationship between Payload and Launch Site
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites

Success Rate vs. Orbit Type



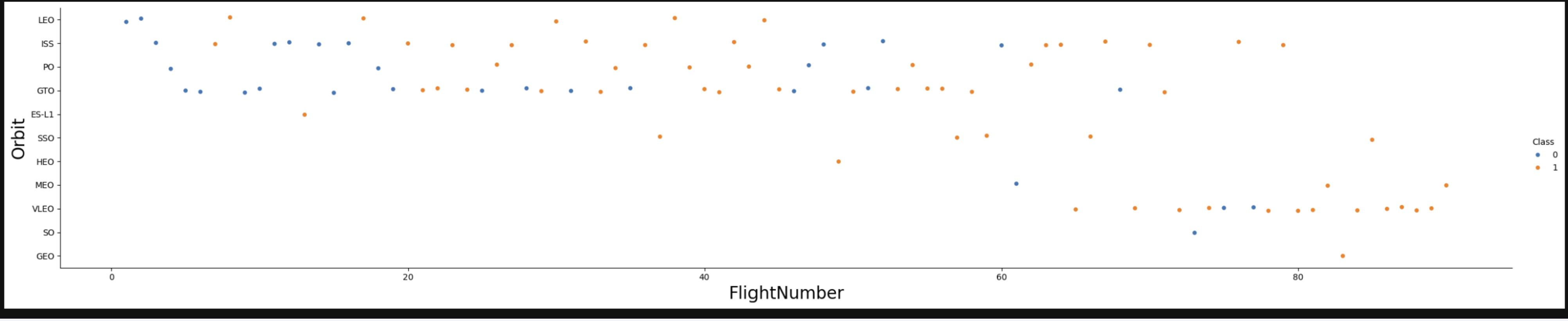
The biggest success rates happens to orbits:

- ES-L1
- GEO
- HEO
- SSO

Flight Number vs. Orbit Type

In [11]:

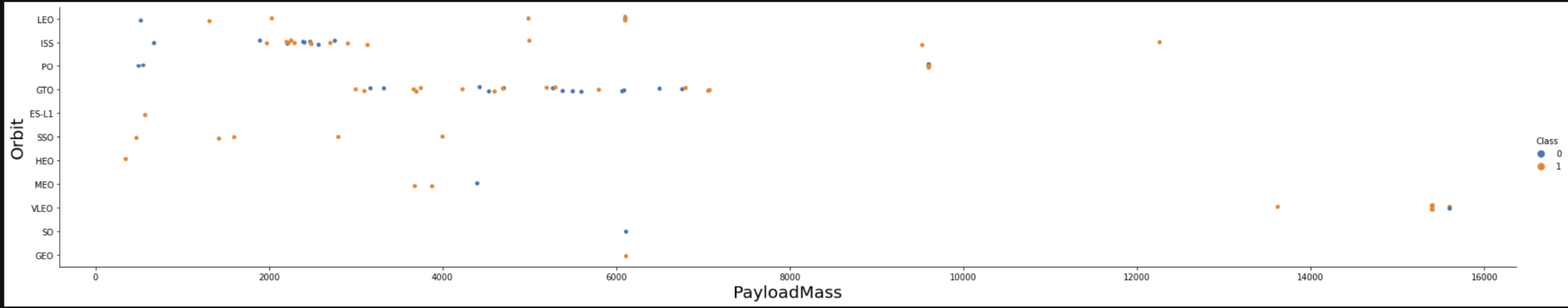
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



Success rate improved over time

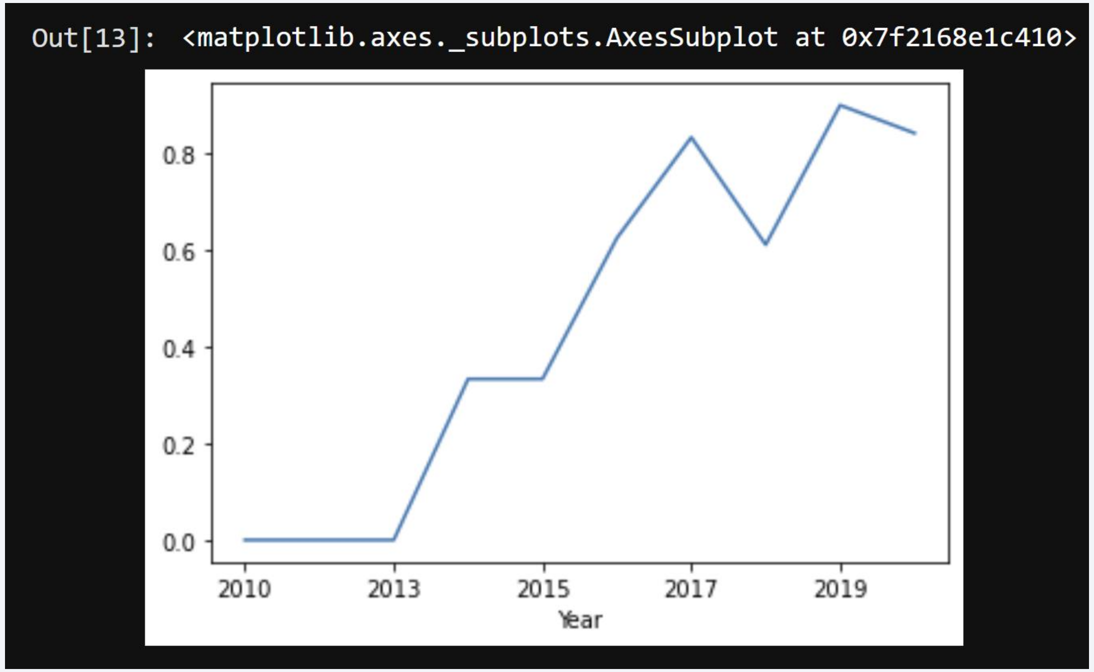
Payload vs. Orbit Type

In [11]:
```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



There are less launches to the orbits SO and GEO.

Launch Success Yearly Trend



Success rate started increasing in 2013 and kept until 2019/2020

All Launch Site Names

```
In [11]: %sql SELECT DISTINCT Launch_Site from SPACEXTABLE

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

Out[11]:

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

Launch Site Names Begin with 'CCA'

```
In [12]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site like 'CCA%' LIMIT 5

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

Out[12]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

In [13]:
```
%sql SELECT sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where customer = 'NASA (CRS)'
```

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

Out[13]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

In [14]:
```
%sql SELECT avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1 %'
```

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

Out[14]:

avg(PAYLOAD_MASS__KG_)
2337.8

First Successful Ground Landing Date

In [16]:
```
%sql SELECT min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

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

Out[16]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [19]: %sql SELECT Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 40

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

Out[19]:

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

Total Number of Successful and Failure Mission Outcomes

```
In [20]: %sql SELECT Mission_Outcome, COUNT(*) from SPACEXTABLE group by Mission_Outcome

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

Out[20]:

Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

In [23]:
```
%sql SELECT distinct Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacexta
```

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

Out[23]:

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

```
In [27]: %sql SELECT substr(Date, 6,2) as month, landing_outcome, booster_version, launch_site from spacextable where landing_outcome

 * sqlite:///my_data1.db
Done.
Out[27]:
```

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

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

In [35]:
```
%sql Select landing_outcome, count(*) from spacextable where date between '2010-06-04'and '2017-03-20' group by landing_outc
```

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

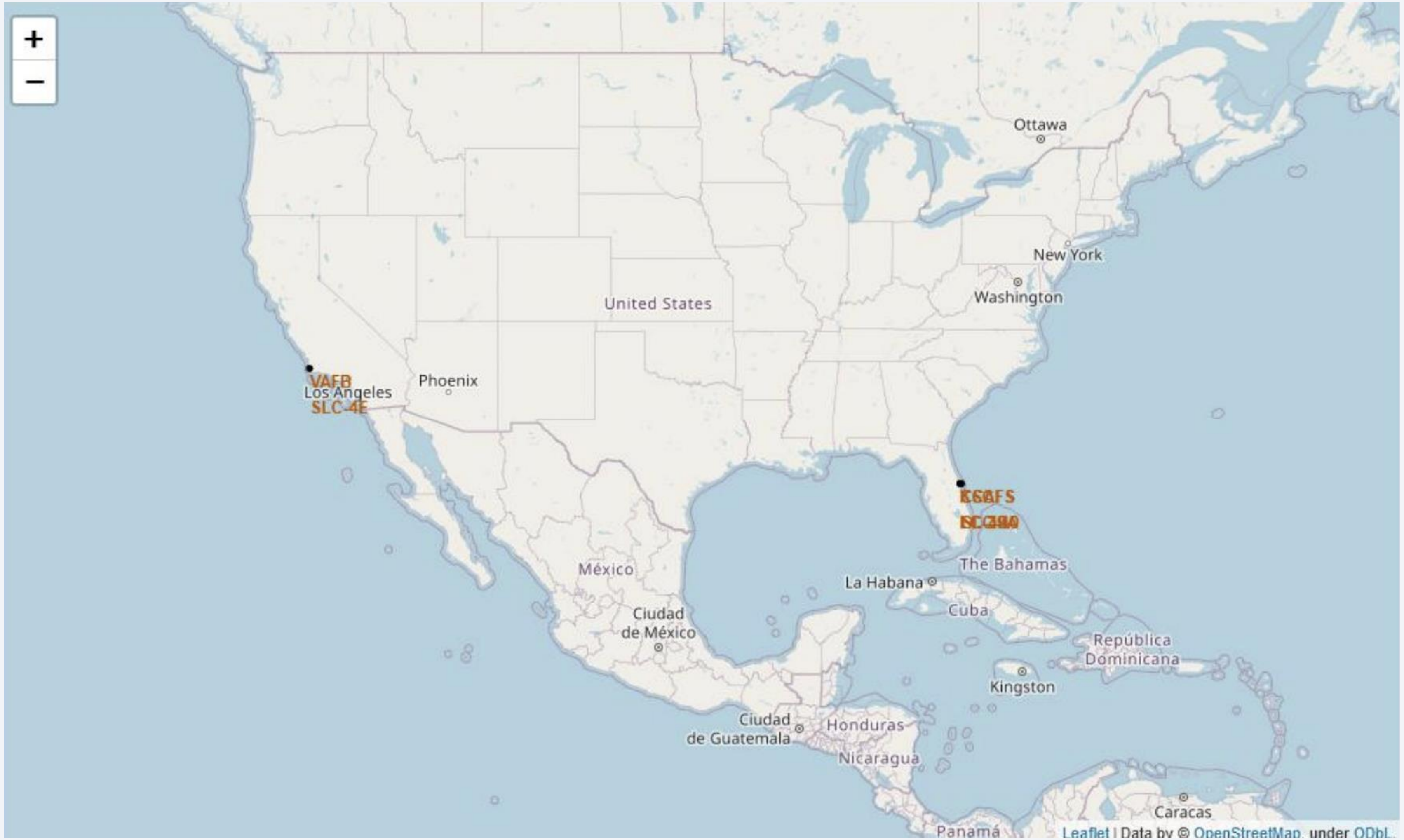
Out[35]:

Landing_Outcome	count(*)
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

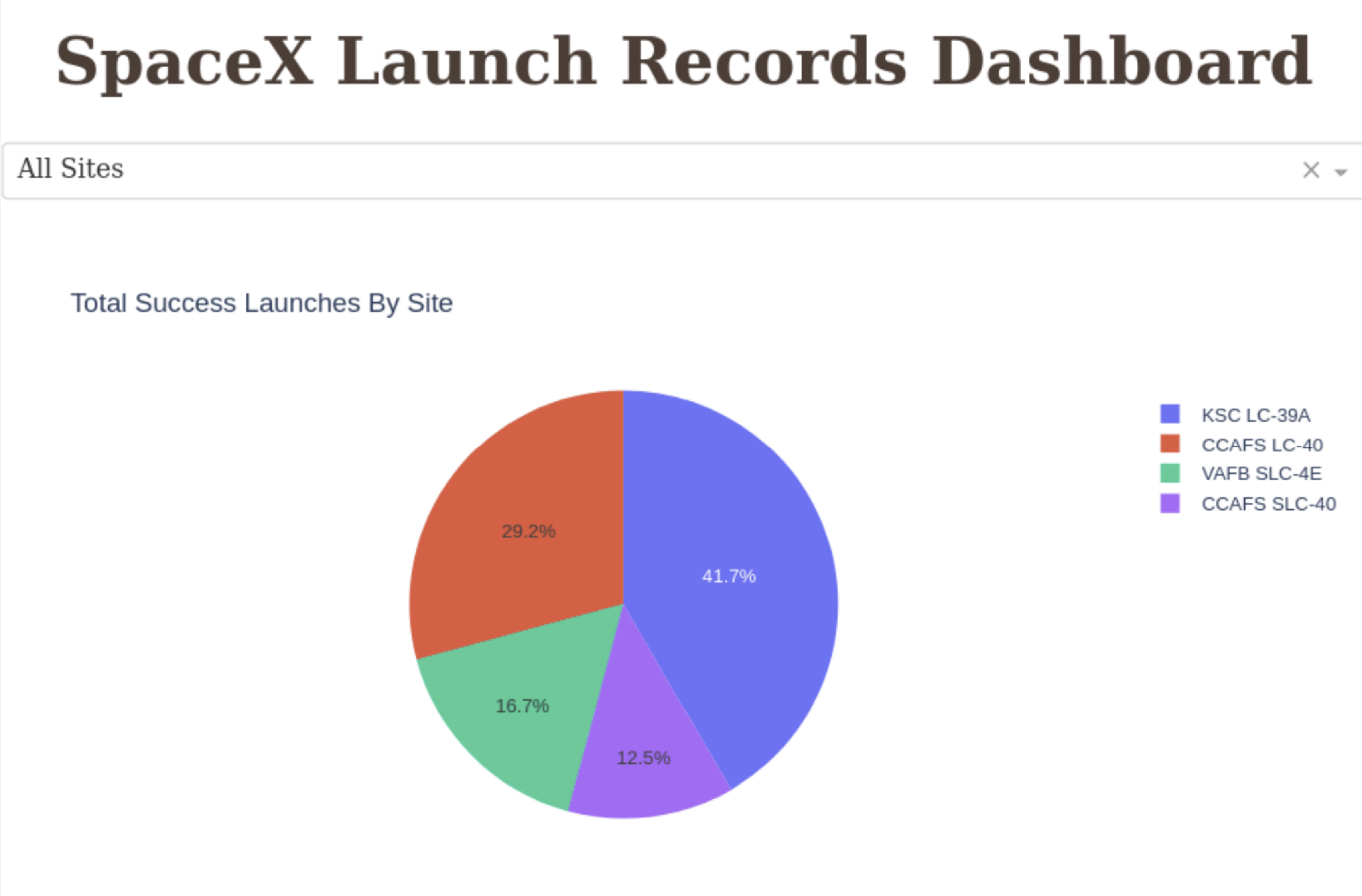
All launch sites



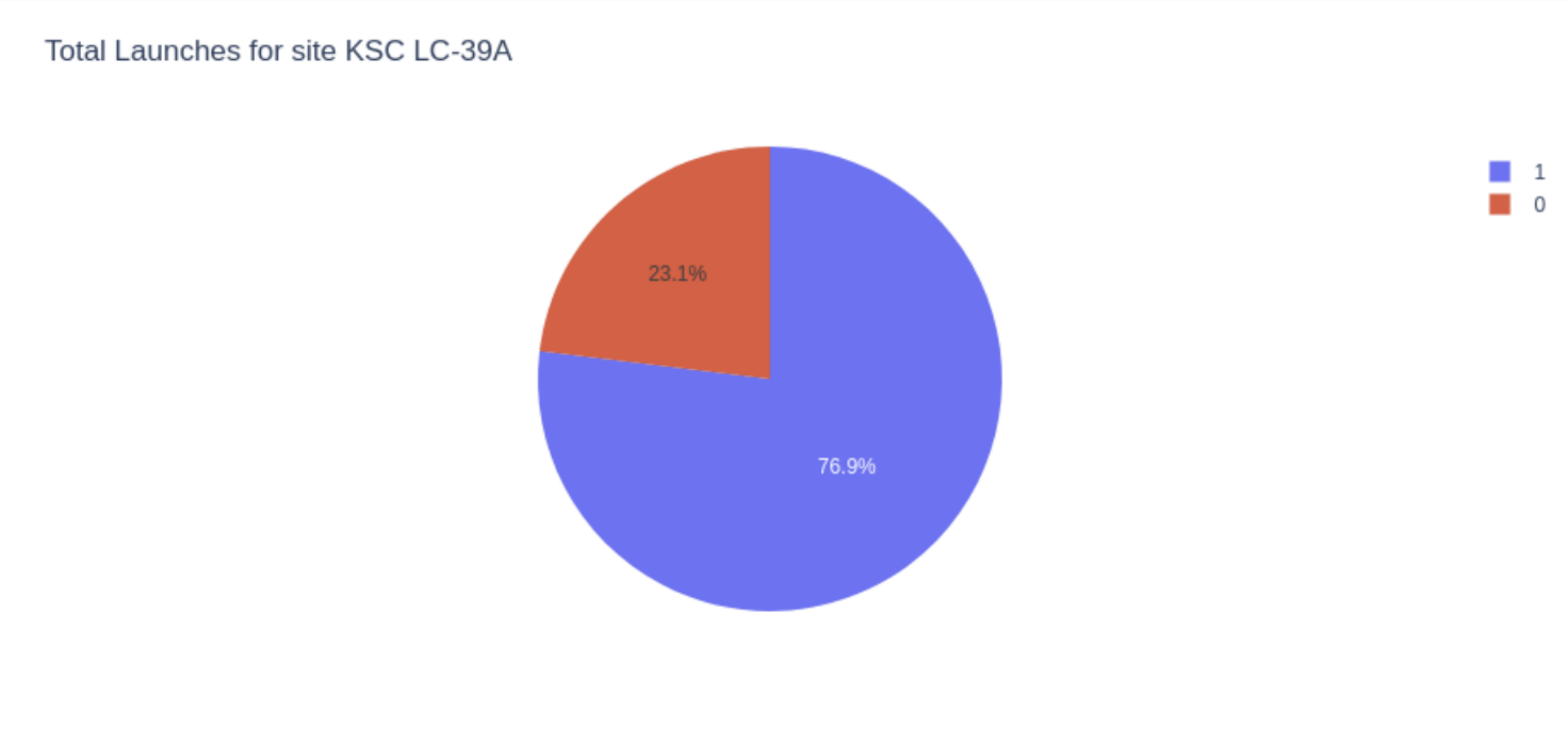
Section 4

Build a Dashboard with Plotly Dash

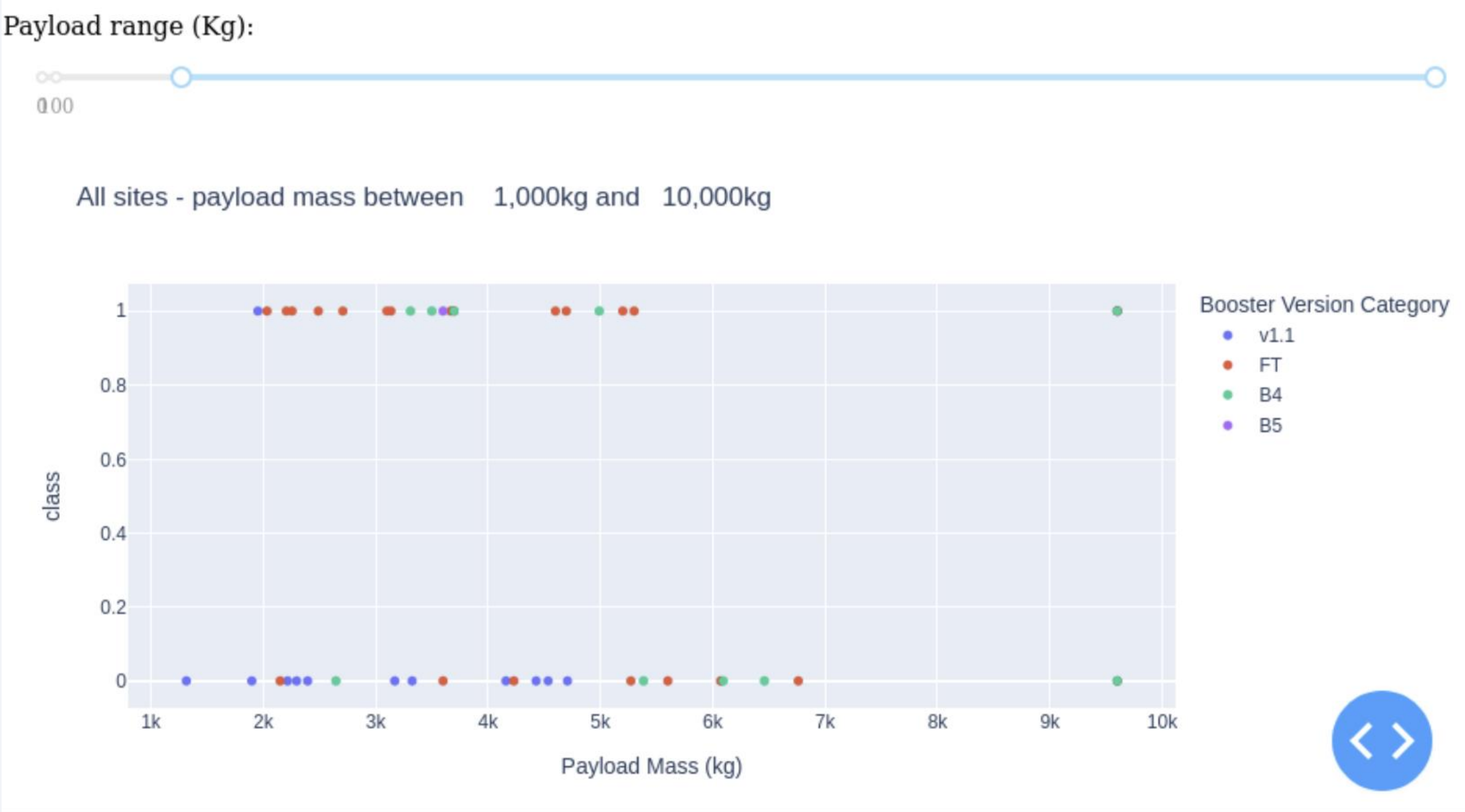
Successful launches / Site



77% Success Rate for KSC LC-39A



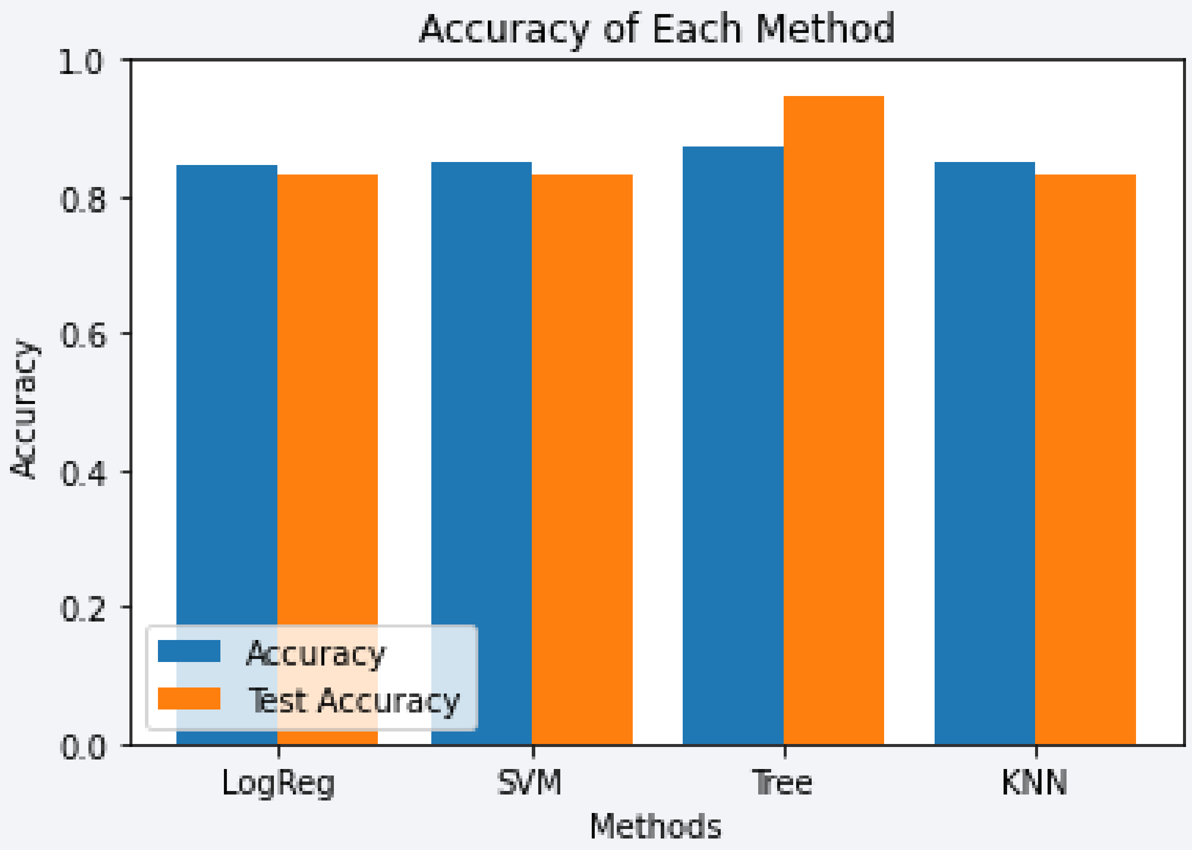
Payload x Launch - <6k kg + FT



Section 5

Predictive Analysis (Classification)

Classification Accuracy



Decision Tree performed the best with more than 94% accuracy

Confusion Matrix

Conclusions

- The best launch site is KSC LC-39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- Every query and code written is available on github at:
- https://github.com/alinparaschiv0107/capstone_Coursera/tree/main

Thank you!