IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alex Ruvolo
1/7/2025

Outline

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

Executive Summary

Tasks Include:

- Data Collection with API and Web Scraping
- Data Wrangling
- Data Analysis with SQL
- Data Analysis for Data Visualization
- Interactive Visual Analytics
- Predictive Analysis

Introduction

- The goal is to predict outcomes for the new rocket company, Space X
- One goal is to predict the price of the rocket launches
- Another goal is to predict whether to reuse the first stage of each launch
- This is done by collecting information and using machine learning to predict future results

Section 1

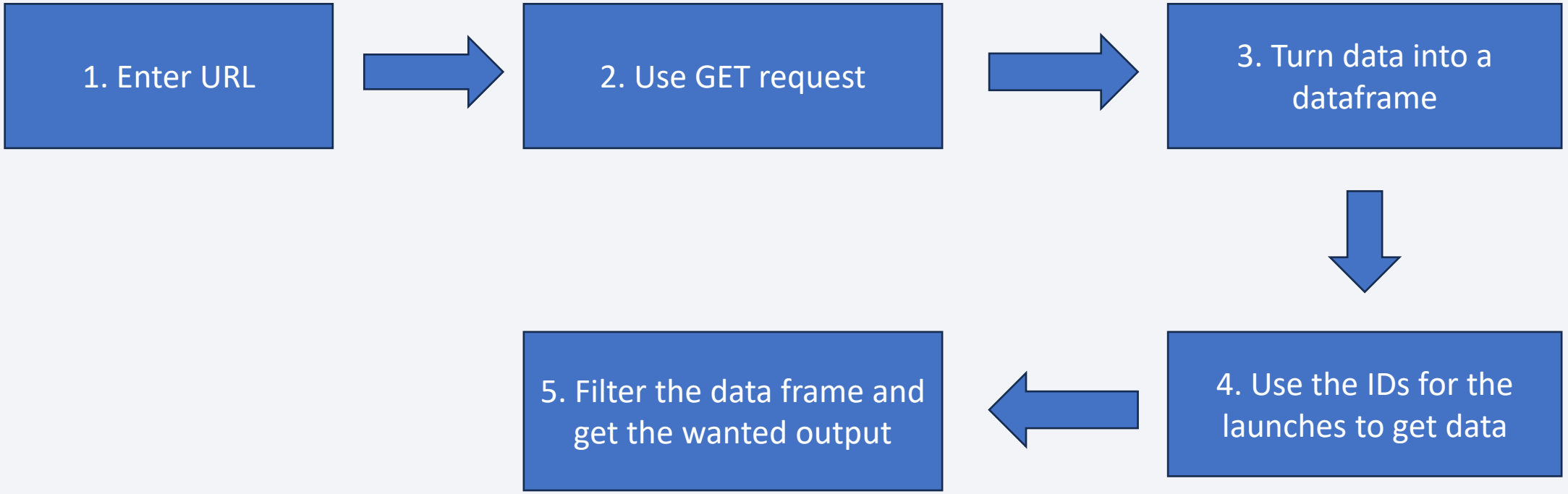
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Used the Space X API
- Perform data wrangling
  - Data was processed through cleaning the data of missing values, and changing units to fit the correct format
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Ran multiple models through machine learning to find the most accurate model for the data and task at hand

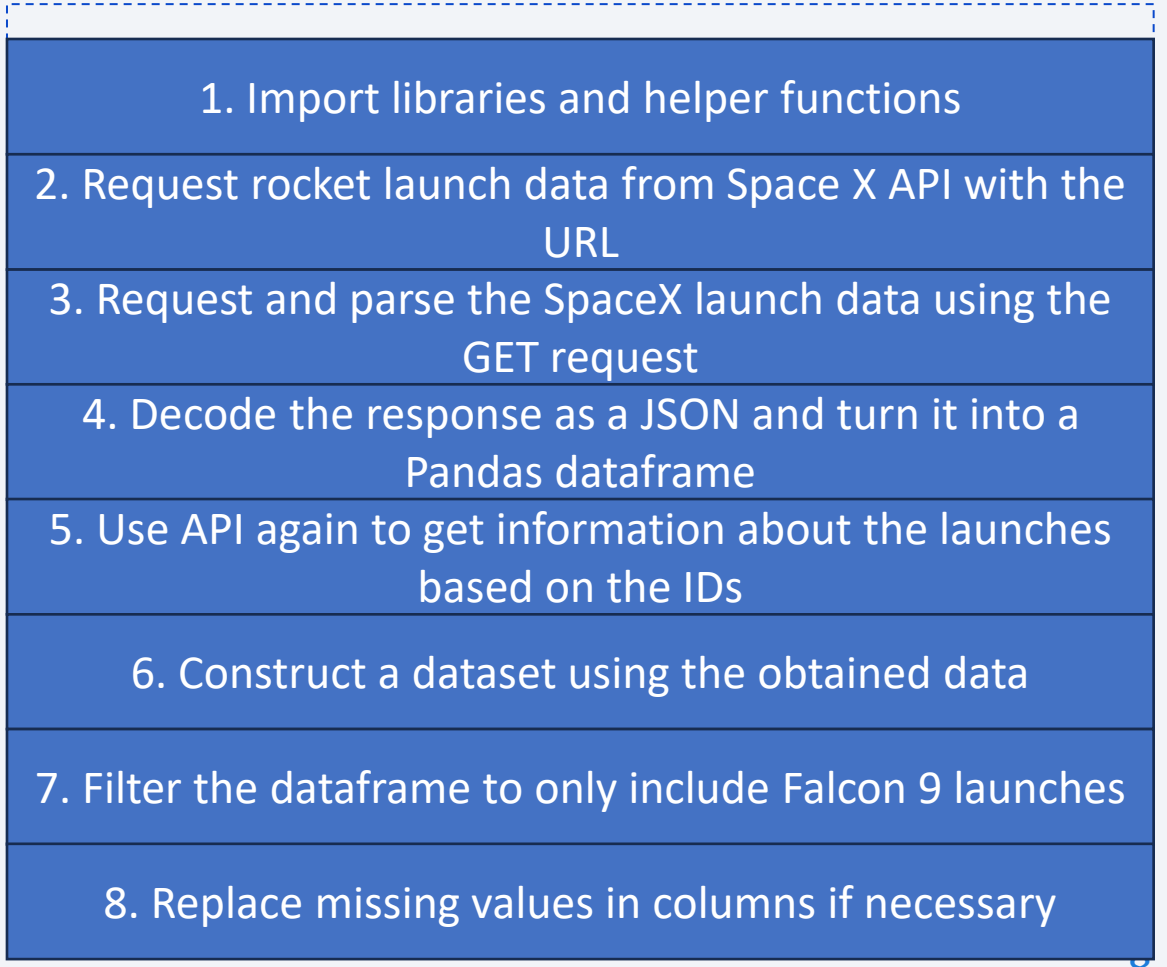
Data Collection



Data Collection – SpaceX API

- Data collection with SpaceX REST calls is shown in the flowchart on the right

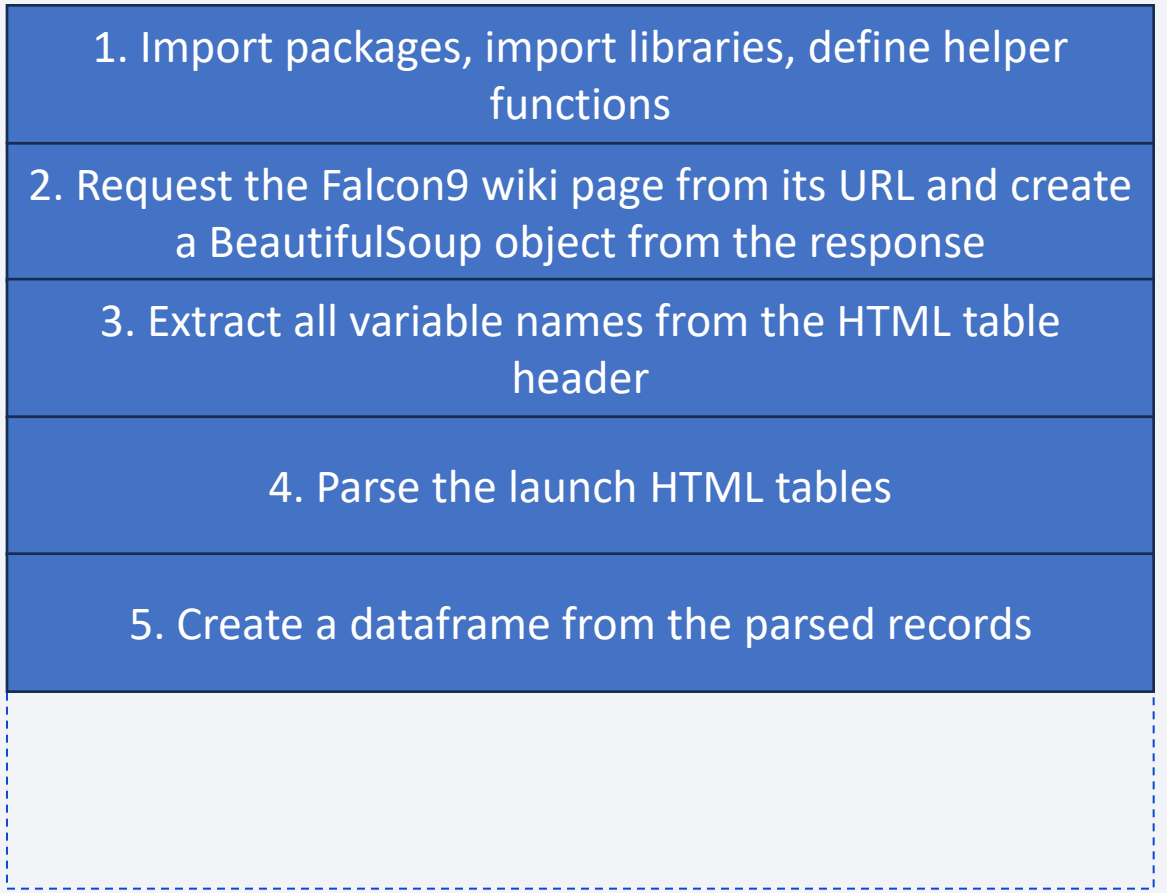
- GitHub URL: Capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · ruvoloa10/Capstone



Data Collection - Scraping

- My web scraping process using key phrases and flowcharts is to the right

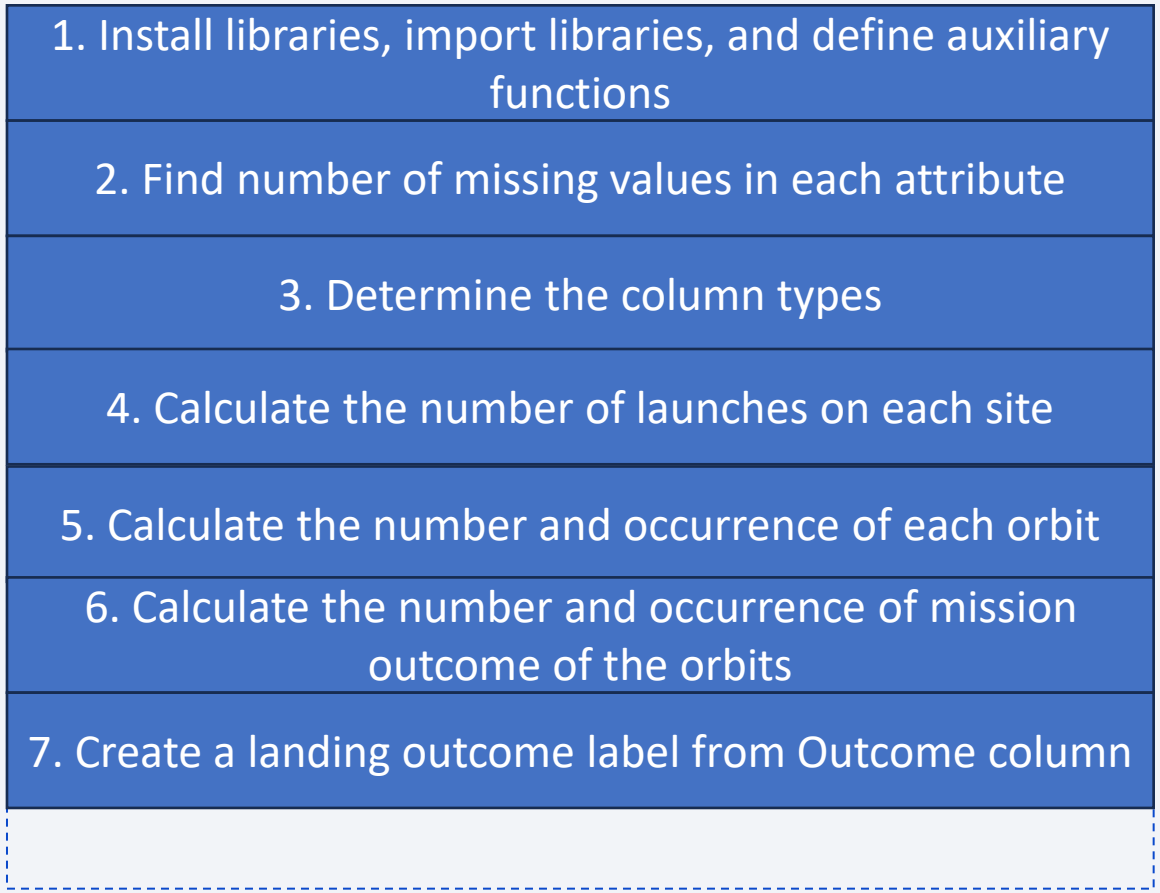
- GitHub URL: Capstone/jupyter-labs-webscraping.ipynb at main · ruvoloa10/Capstone



Data Wrangling

- Data Wrangling process is shown to the right

- GitHub URL: Capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · ruvoloa10/Capstone



EDA with Data Visualization

Plots:

- Scatterplot to view: FlightNumber vs PayloadMass, FlightNumber vs LaunchSite, and Payload Mass vs Launchsite, FlightNumber vs Orbit type, PayLoad Mass vs Orbit type
- Bar plot to view: Orbit and success rate
- Line Chart to view: Year vs Average Success Rate
- GitHub URL: Capstone/edadataviz.ipynb at main · ruvoloa10/Capstone

EDA with SQL

SQL Queries:

- **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 version F9 v1.1**
- **List the date when the first successful landing outcome in ground pad was achieved.**
- **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
- **List the total number of successful and failure mission outcomes**
- **List the names of the booster_versions which have carried the maximum payload mass. Use a subquery**
- **List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.**
- **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.**

GitHub URL: Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · ruvoloa10/Capstone

Build an Interactive Map with Folium

- Markers represent a launch site and launch records
- A green marker represents a successful and a red marker is used for a failed launch
- Marker clusters were used for multiple launches that shared the same coordinates with each other
- GitHub URL: Capstone/lab_jupyter_launch_site_location.ipynb at main · ruvoloa10/Capstone

Build a Dashboard with Plotly Dash

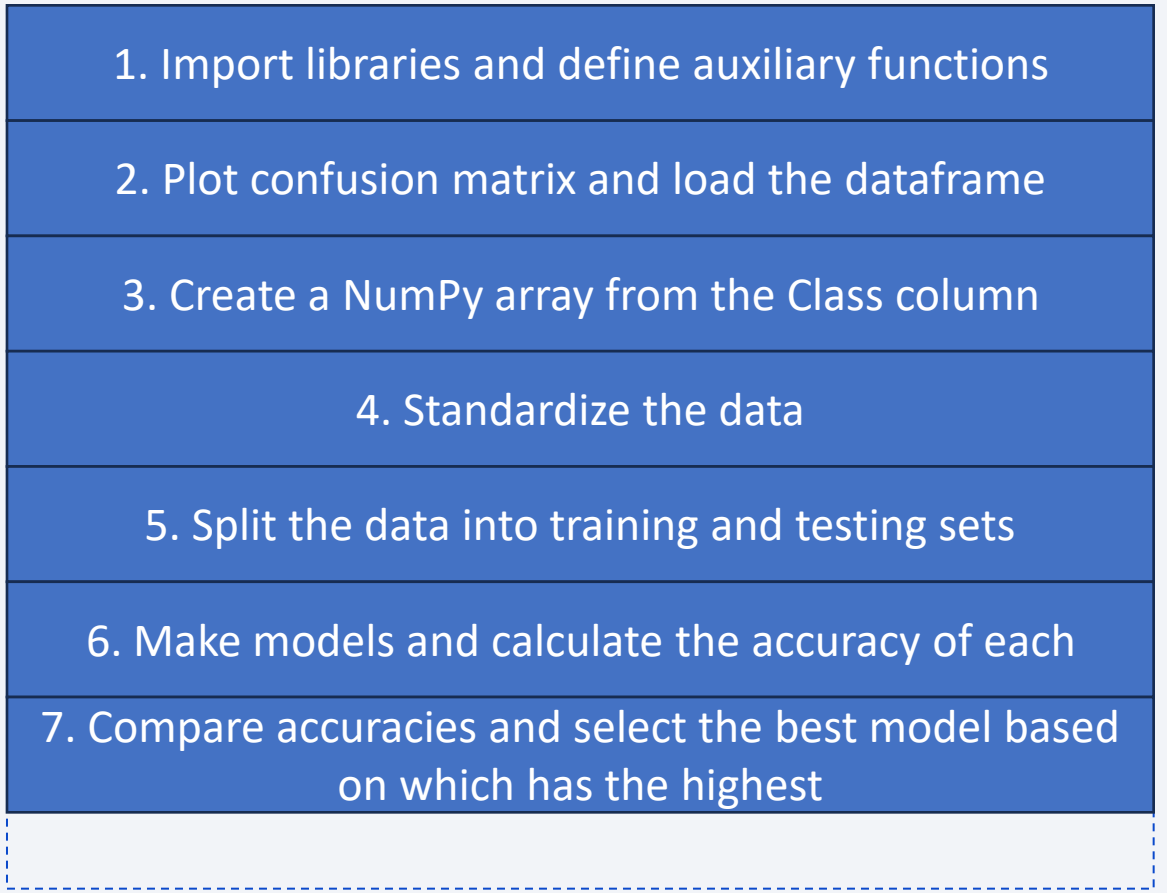
Added plots and graphs:

- Pie charts and scatter plots were added to visualize the launch records
- The charts showed the rocket launch success rate per site, and the possible influences on there launch rates
- Successes were labelled as a 1 and failures were labelled as a 0
- GitHub URL: Capstone/spacex_dash_app.py at main · ruvoloa10/Capstone

Predictive Analysis (Classification)

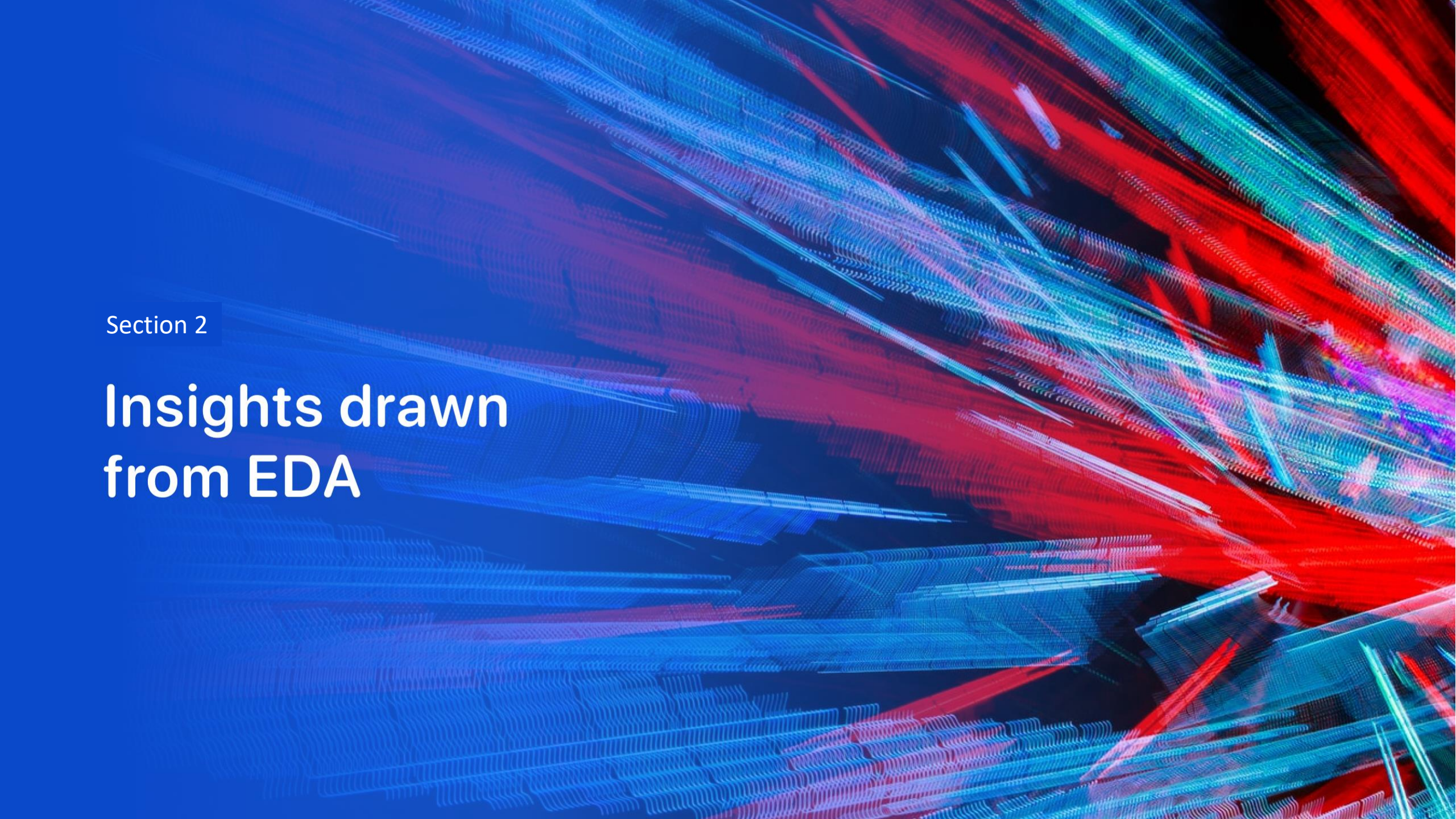
- Flowchart on process of finding best classification model on the right

- GitHub URL: Capstone/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · ruvoloa10/Capstone



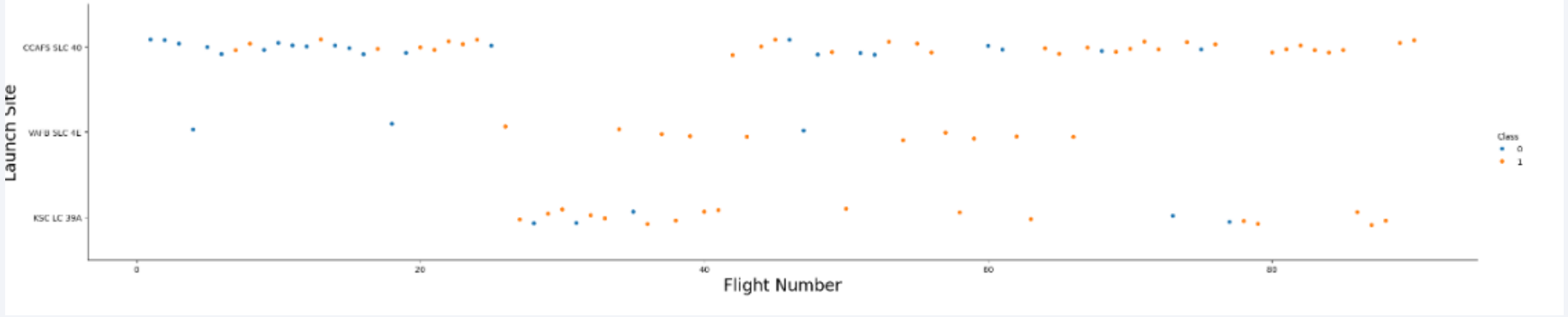
Results

- As flight number increases, the first stage is more likely to land successfully. The payload does not appear to be great factor because even with heavier payloads the first stage tends to return successfully
- Launch sites tend to be populous areas near the coast, which gives reason for having more resources, and being able to use the water for landing the rockets
- The different models used to predict the landing success of the rockets had an accuracy of 83.33%


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

```
# 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 va
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

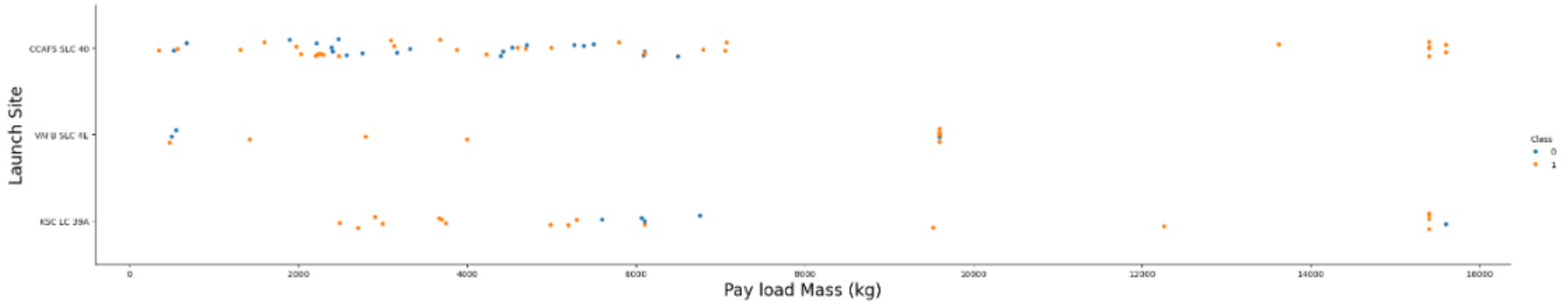


It appears there were more successful landings, for each launch site, as the number of flights increased.

Payload vs. Launch Site

In [7]:

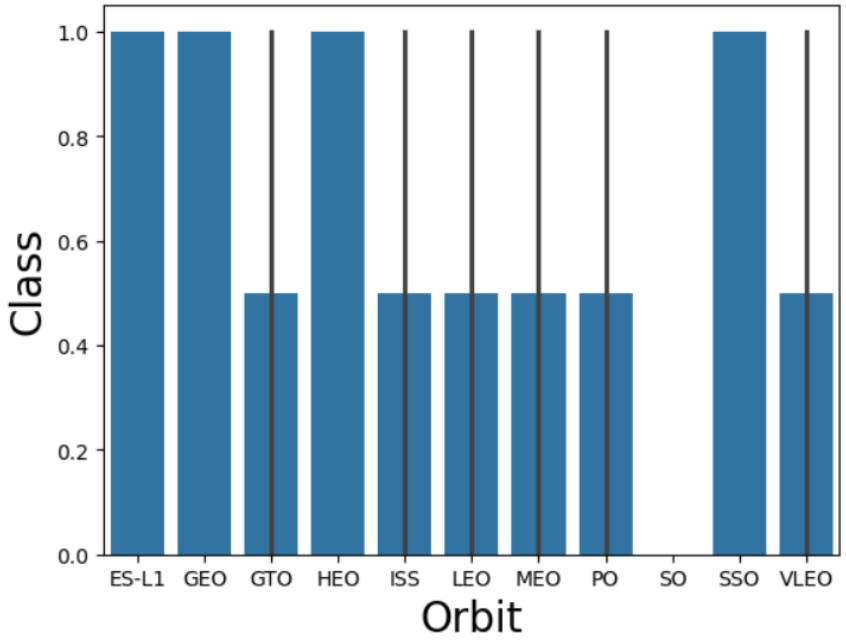
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the cla
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



There are no pay load masses that exceeded 10,000 kg. It appears that the pay load mass did not really impact whether the landing was successful, as even with the highest masses, there was only one failed landing.

Success Rate vs. Orbit Type

In [8]:
```
# HINT use groupby method on Orbit column and get the mean of Class column
t = df.groupby(['Orbit', 'Class'])['Class'].agg(['mean']).reset_index()
sns.barplot(y="Class", x="Orbit", data=t)

plt.xlabel("Orbit",fontsize=20)
plt.ylabel("Class",fontsize=20)
plt.show()
```

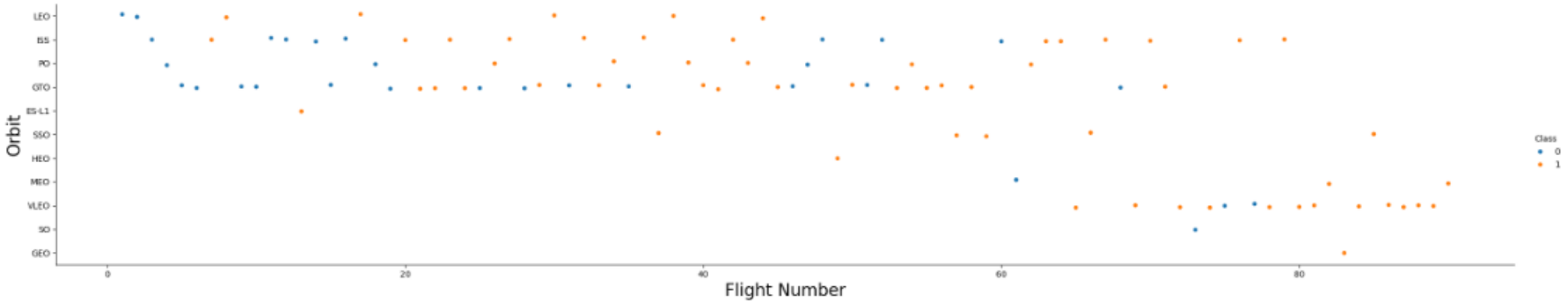


The orbit types with the highest success rates are ES-L1, GEO, HEO, and SSO. The remaining orbits, except SO, all have similar success rates around 0.5. The SO orbit has the lowest success rate with it being close to 0.

Flight Number vs. Orbit Type

In [9]:

```
# 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("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

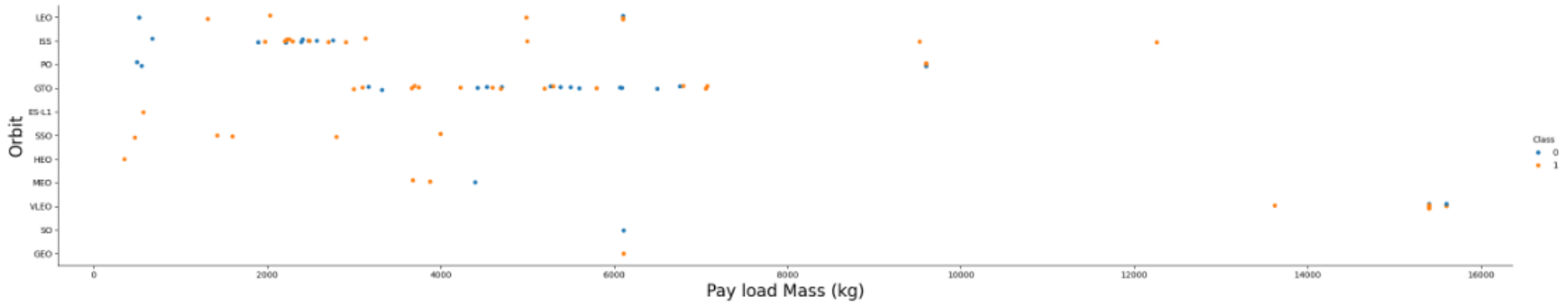


It appears the flight number might be related to the success for the LEO, PO, and GTO orbits. For these it appears that as the flight number increases, the number of successful landings also appears to increase. This does not appear to apply for the other orbits because there is either not enough information or not trend in the successes and failures.

Payload vs. Orbit Type

In [10]:

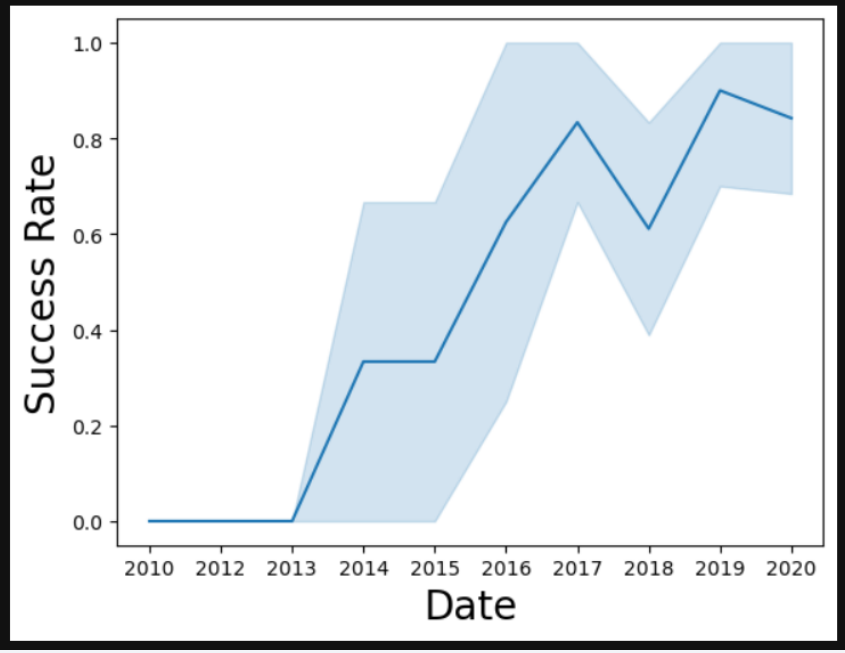
```
# Plot a scatter point chart with x axis to be Payload Mass and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



The pay load mass does not appear to affect the success rate of the landings for the different orbit types. For the heaviest pay loads, most of the landings were successful.

Launch Success Yearly Trend

```
[13]: # Plot a line chart with x axis to be the extracted year and y axis to be the success rate
      sns.lineplot(data=df, x="Date", y="Class")
      plt.xlabel("Date",fontsize=20)
      plt.ylabel("Success Rate",fontsize=20)
      plt.show()
```



It appears the yearly average success rate has been increasing from 2013 to 2020. There was a slight decrease from 2017 to 2018, but overall the rate has increased over time.

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
In [11]: %sql  SELECT DISTINCT LAUNCH_SITE  FROM SPACEXTBL
```

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

Out[11]:

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

The names of the launch sites are CCASFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Here, there are 5 records that have a launch site name that begins with ‘CCA’, where all of the launch-sites have the name “CCAFS LC-40”

Total Payload Mass

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

The total payload mass carries by boosters launched by NASA (CRS) is 45596 kg.

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

The average payload mass carried by booster version F9 v1.1 is 2928.4 kg.

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE Landing_Outcome = "Success (ground pad)"
```

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

MIN(DATE)
2015-12-22

The date when the first successful landing outcome in ground pad was achieved was 12-22-2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql SELECT BOOSTER_VERSION from SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' and  PAYLOAD_MASS__KG_  >4000 and
```

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

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

The names of the boosters which have success in drone ship and have a pay load mass between 4000 and 6000 kg are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight
```

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

count(MISSION_OUTCOME)
99

The total number of successful and Failure Mission Outcomes was 99 in total.

Boosters Carried Maximum Payload

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

```
%sql SELECT 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

The names of the booster versions that have carried the maximum payload mass are 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

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT substr(Date, 6,2),BOOSTER_VERSION,LAUNCH_SITE,LANDING_OUTCOME FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (
```

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

substr(Date, 6,2)	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)
01	F9 v1.1 B1017	VAFB SLC-4E	Failure (drone ship)
03	F9 FT B1020	CCAFS LC-40	Failure (drone ship)
06	F9 FT B1024	CCAFS LC-40	Failure (drone ship)

There were five records that had failed landing outcomes in the year 2015. These include the moth, booster version, and the launch site.

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(*) AS "Count"
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' and '2017-03-20'
GROUP BY landing_outcome
ORDER BY Count DESC
;
```

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

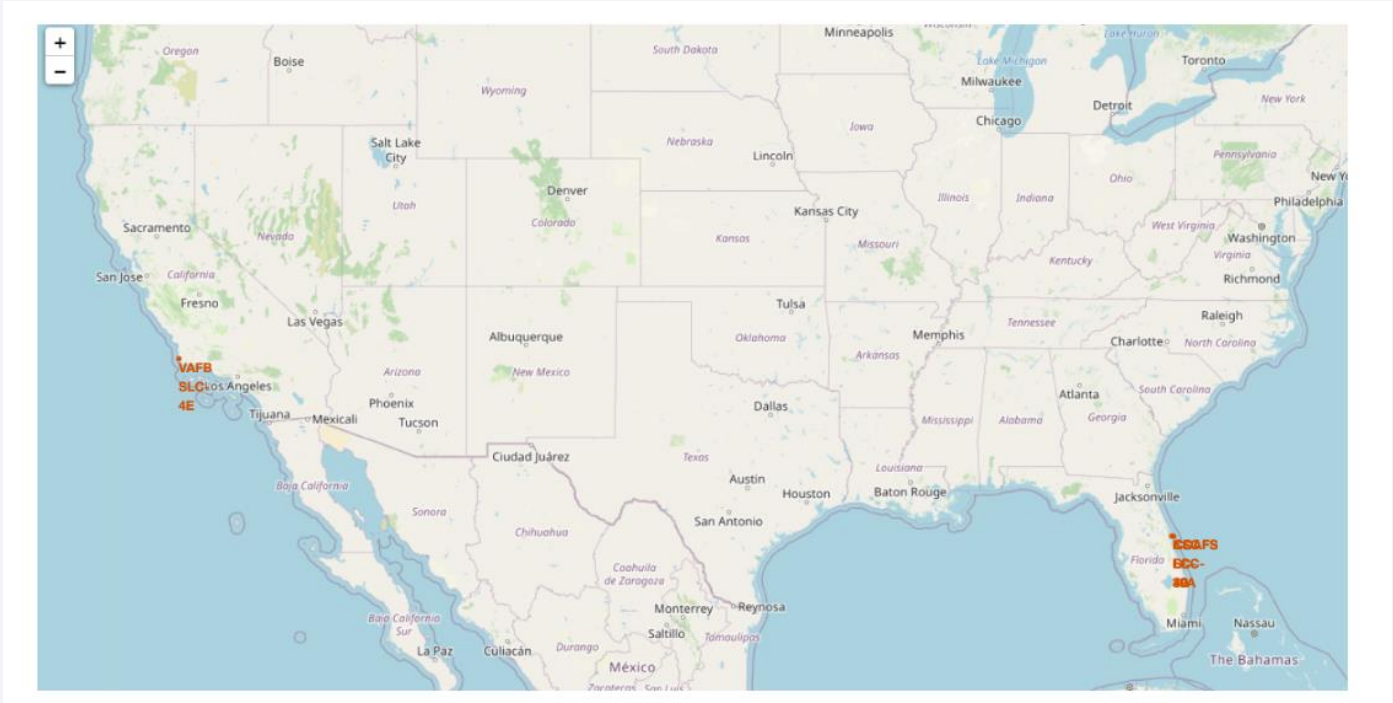
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

The landing outcomes in descending order between the specified dates are no attempt, success (drone ship) and failure (drone ship), success (ground pad) and controlled (ocean), uncontrolled (ocean) and failure (parachute), and precluded (drone ship).

Section 3

Launch Sites Proximities Analysis

All Launch Sites on US Map



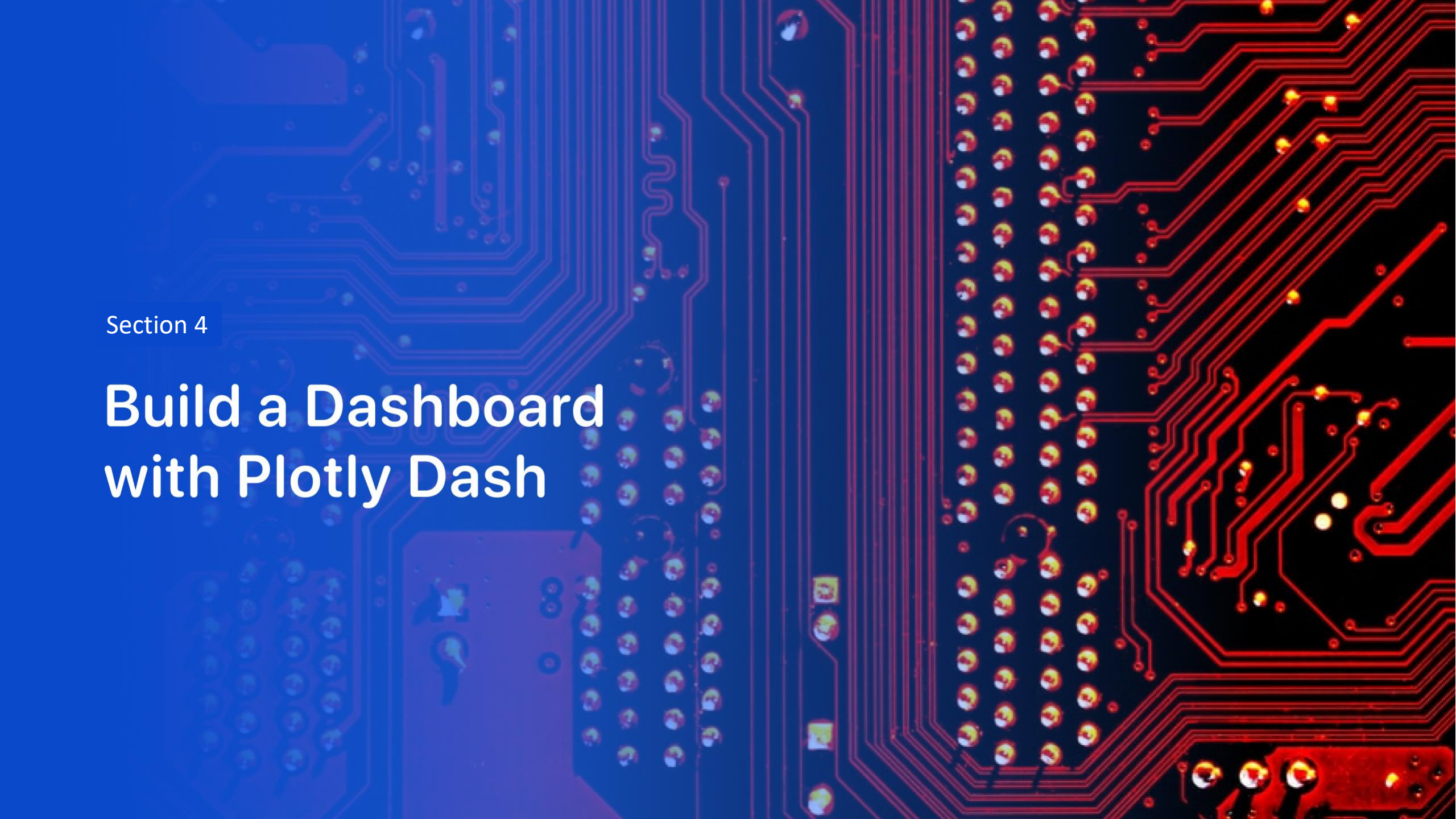
This map displays the launch sites across the United States, with some being in Florida and California

Success/Failure Launches at One Launch Location

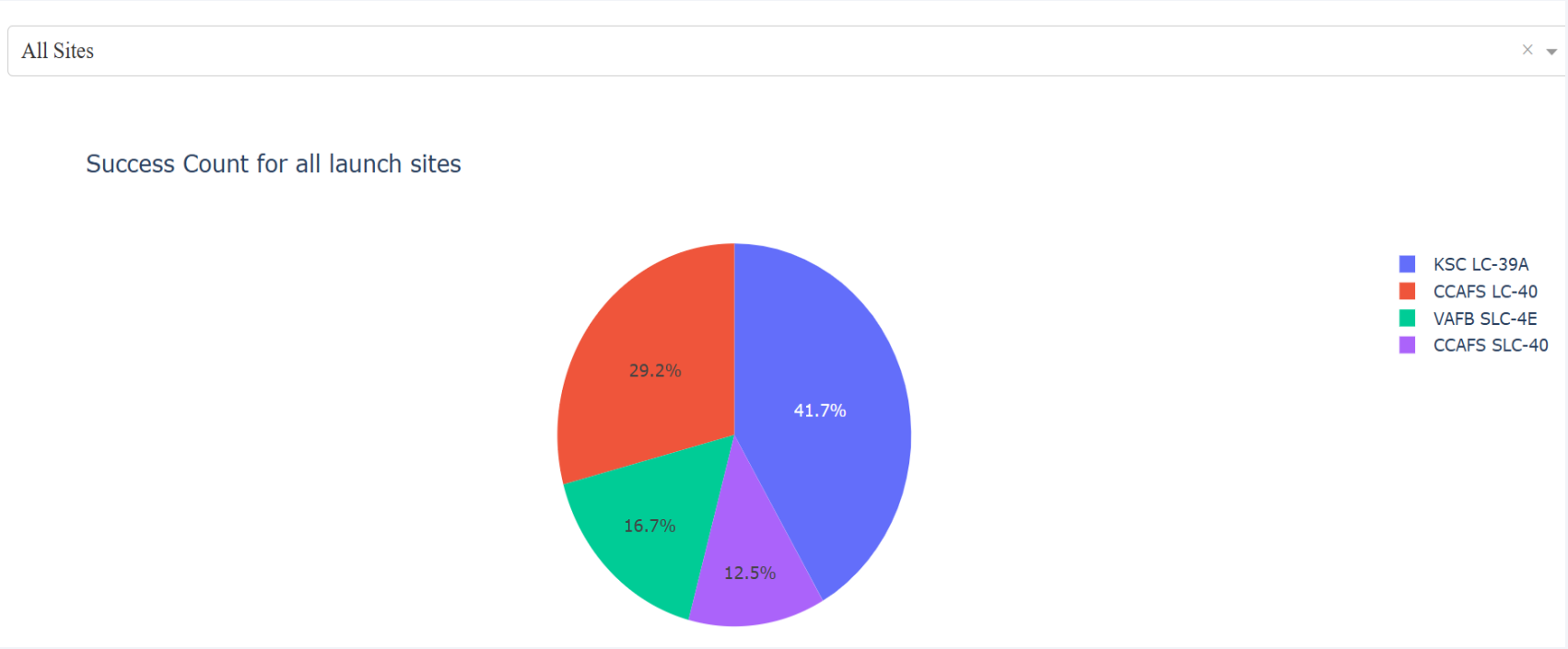
This displays the launches specifically at one of the Florida locations. The green markers are marks for successful landings and the red markers are marks for failed landings.

Proximity Distances

This displays the distance from the coastline from the 7th launch. The distance from the coast to this launch appears to be 0.9 km.

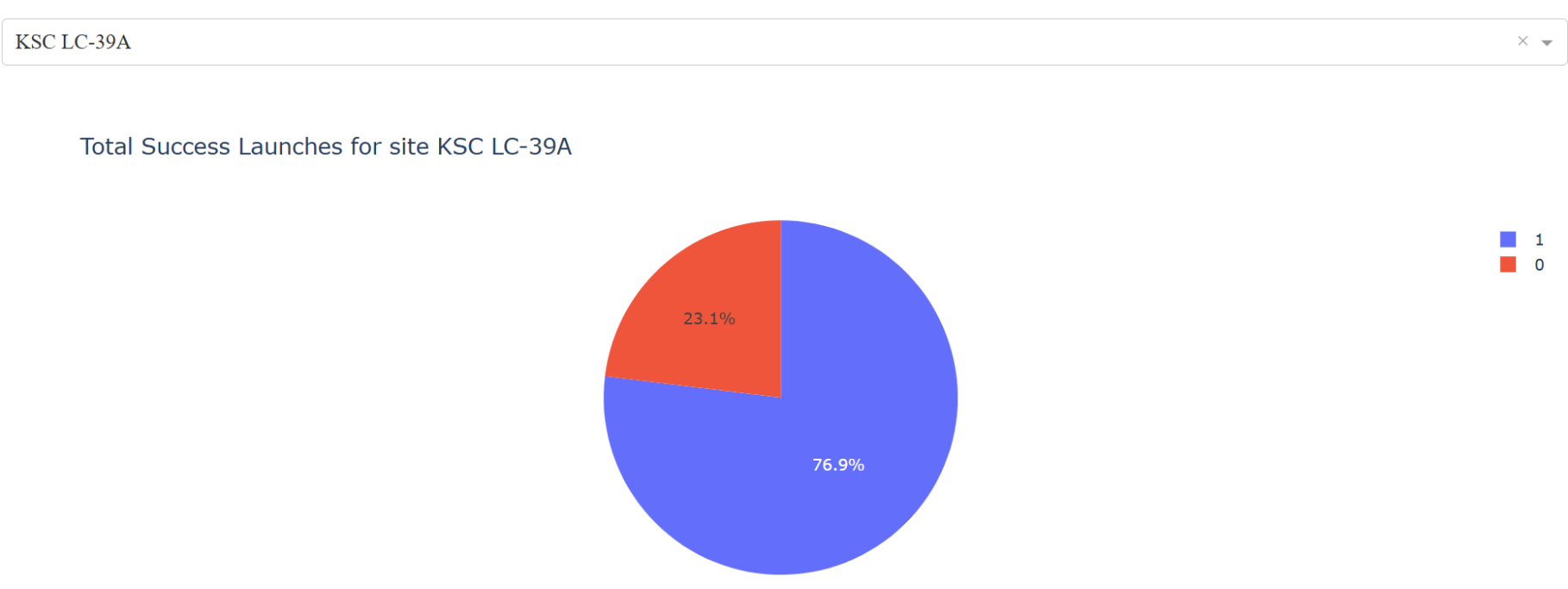

Section 4
Build a Dashboard with Plotly Dash

Dashboard All Sites Pie Chart



This is a pie chart displaying the proportion that each launch site had of the total successful landings. Here, the KSC LC-39A had the largest proportion of successful landings out of the total with a percentage of 41.7% and the CCAFS LC-40 had the lowest proportion of the total number of successes with a percentage of 12.5%.

Pie Chart of Highest Success Ratio



The launch site with the highest success ratio is KSC LC-39A. As shown by the chart, it had a success ratio of 76.9%, which is the highest success ratio of the four launch sites.

Payload vs Launch Outcomes for All Sites

As shown by the first scatterplot the FT booster had the highest number of successes out of the five different boosters. The second scatterplot shows that the highest number of successes appears in the payloads mass range from 3000 to 4000 kg with the next highest number of successes being in the range from 2000 to 3000 kg.

Section 5

Predictive Analysis (Classification)

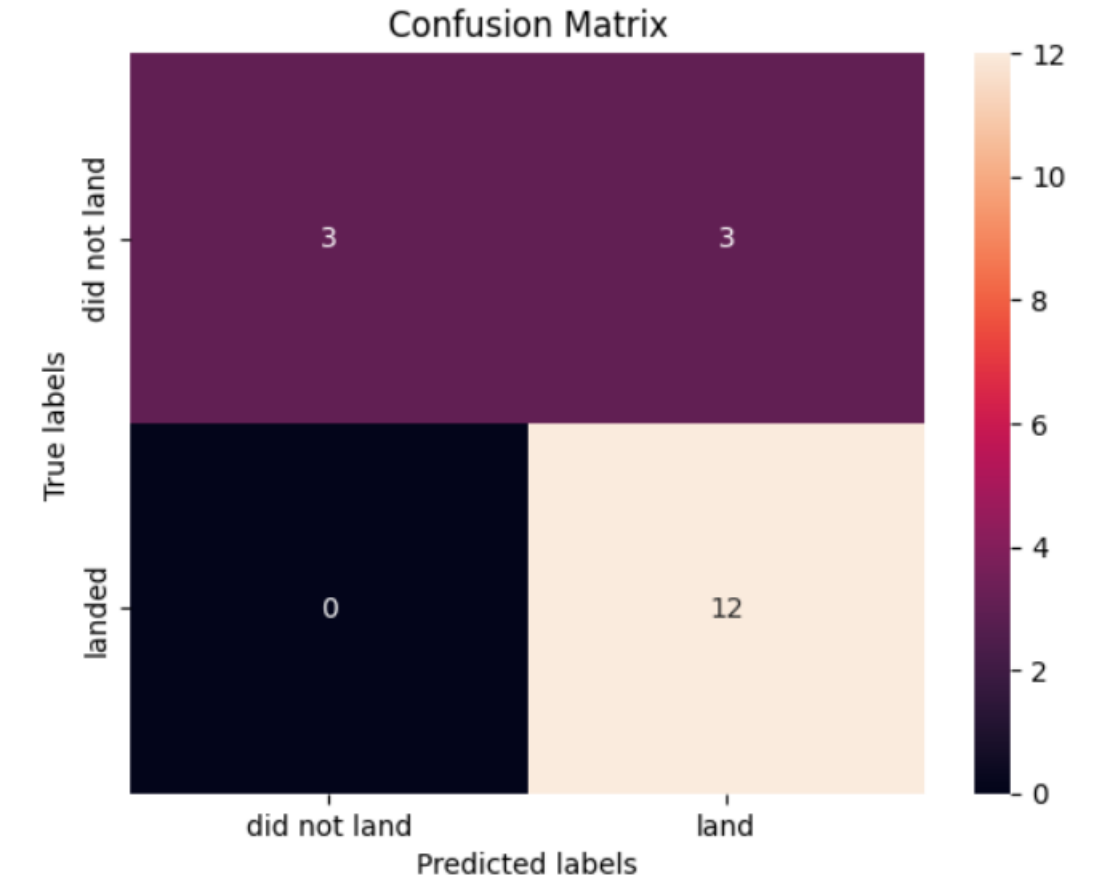
Classification Accuracy

```
accuracy = [accuracy, logreg_score, knnaccuracy, treeaccuracy]
accuracy = [i * 100 for i in accuracy]

method = ['Support Vector Machine', 'Logistic Regression', 'K Nearest Neighbour', 'Decision Tree']
models = {'ML Method':method, 'Accuracy Score (%)':accuracy}

ML_df = pd.DataFrame(models)
ML_df
```

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

All the models have the same average accuracy with 83.33%

Confusion Matrix



All the models performed the same and have the same confusion matrix. Here, it is shown that the models correctly predict that 12 of the landings will land safely. It also correctly predicted that 3 of the landings would not land safely. However, there are 3 landings which the models predicted would land and ended up not landing, which is considered a false positive.

Conclusions

- Multiple steps are involved in data science work, which could include data collection, data manipulation, query work, machine learning, and predictive analysis
- Space X has multiple factors that affect the landing outcomes of the launches. This includes the launch location, the type of landing, the thrusters used, the payload mass, the booster used, and even the timeframe
- These factors can be used to predict whether a launch in likely to safely succeed in a landing process
- Multiple machine learning methods can be used to find the best model for predicting, and this should be done by comparing their accuracies

Appendix

For reference all files, labs, and notes are to be found here:

ruvoloa10/Capstone

Thank you!