IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Olivia Njinjoh
28 March 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Summary of methodologies

- Data Collection using API
- Data Collection with web scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Introduction

- In the capstone project, I have the role of a data scientist working for a new rocket company to determine the price of each launch of SPACEX.
- It is done by gathering information about Space X and creating dashboards for the team.
- It also determines if SpaceX will reuse the first stage
- Instead of using rocket science to determine if the first stage will land successfully, a machine learning model will be trained using public information to predict if SpaceX will reuse its first stage.

Section 1

Methodology

Methodology

Executive Summary

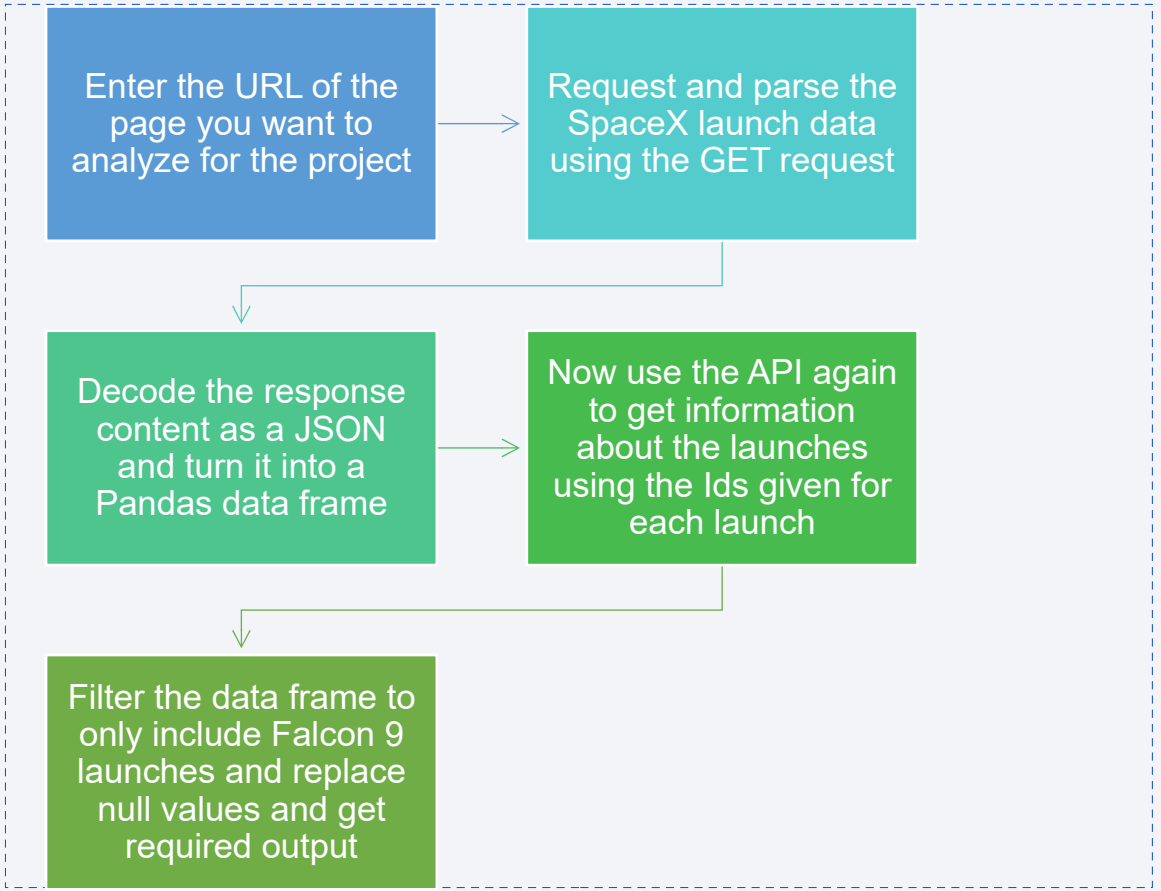
- Data collection methodology:
  - Get requests to the SpaceX API and web scraping from wikipedia
- Perform data wrangling
  - Clean the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create the best Machine Learning Model

Data Collection

The data sets are collected by

- SpaceX API request
- Web Scraping

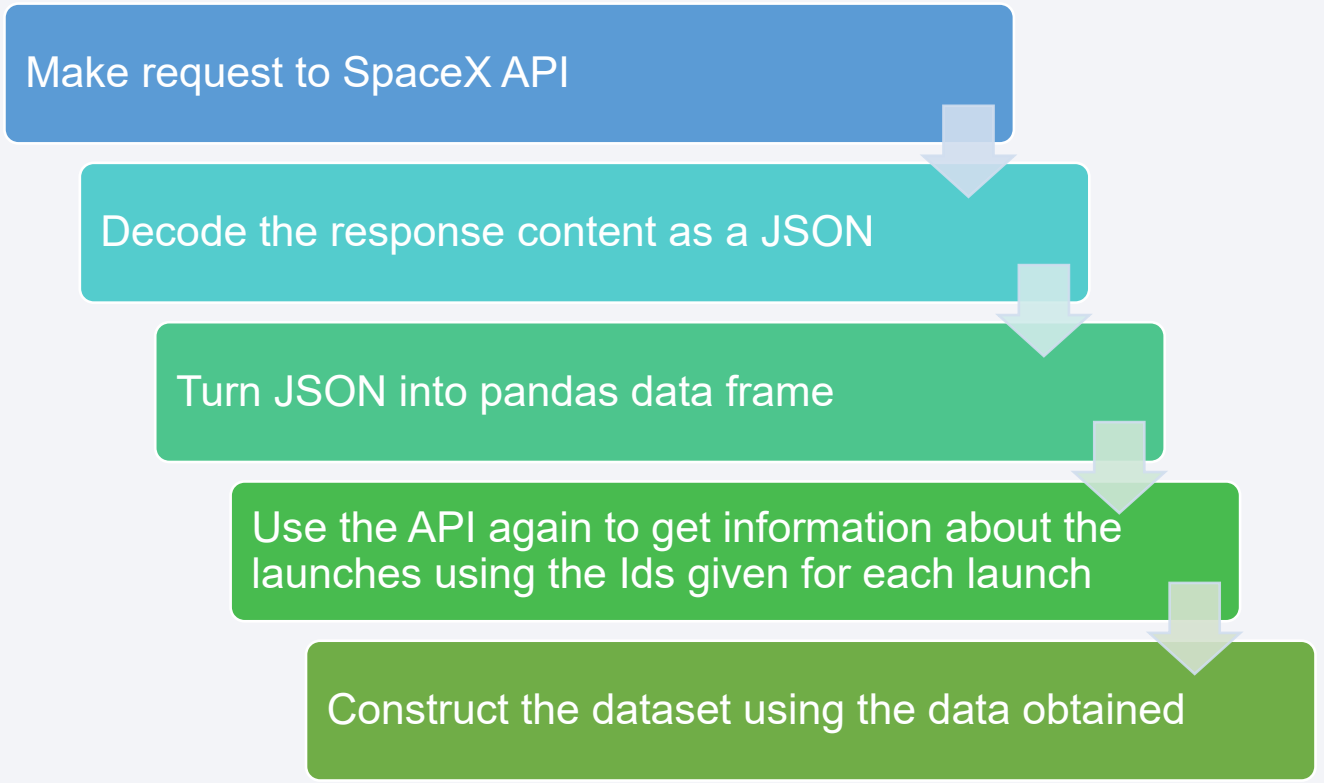
GitHub URL link: https://github.com/OliviaNNgit/SpaceX-Falcon9-DataScience-Capstone/blob/main/Data%20Collection%20API.ipynb



Data Collection – SpaceX API

This is how Data Collection is done as shown in the flow chart. The complete notebook is found in the link below.

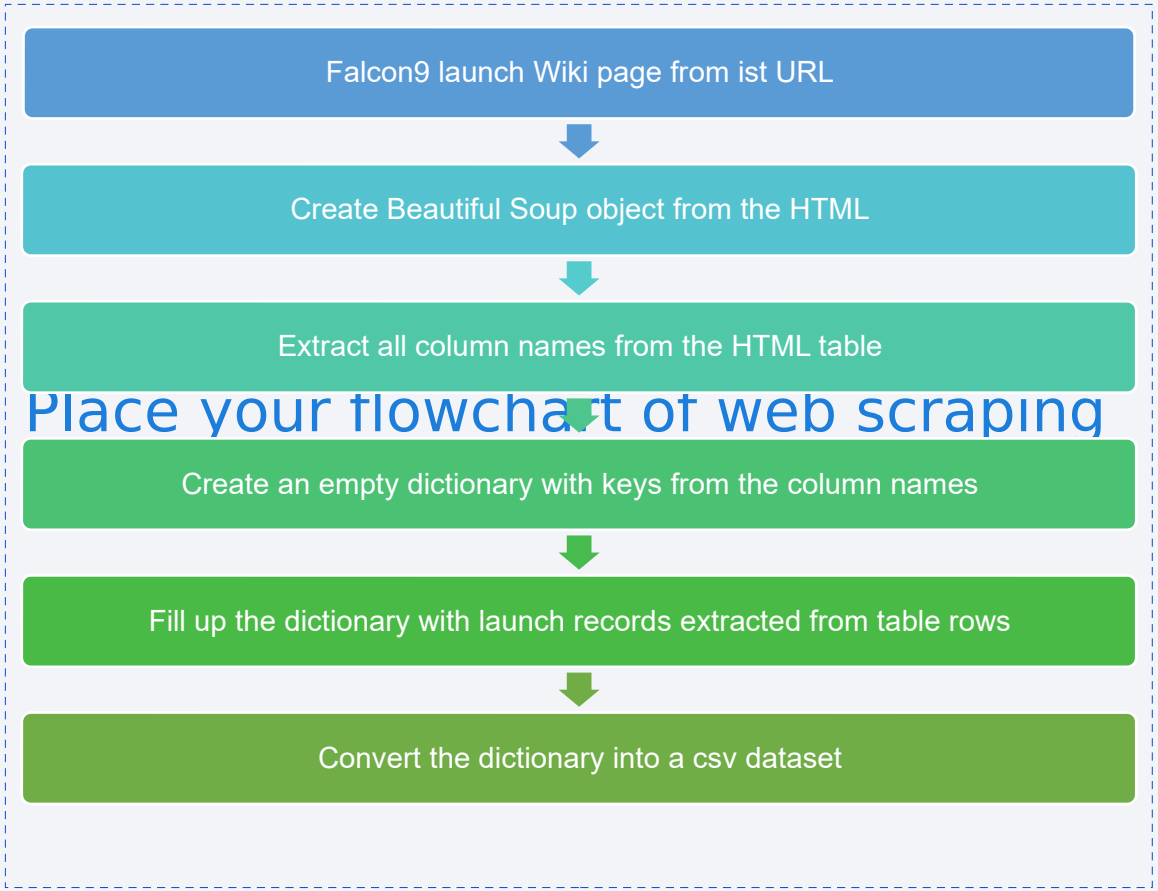
Github URL link: https://github.com/OliviaNNgit/SpaceX-Falcon9-DataScience-Capstone/blob/main/Data%20Collection%20API.ipynb



Data Collection – Web Scraping

- Data Collection by web scraping process is shown in the flow chart for an overview. The complete notebook is found in the link below.

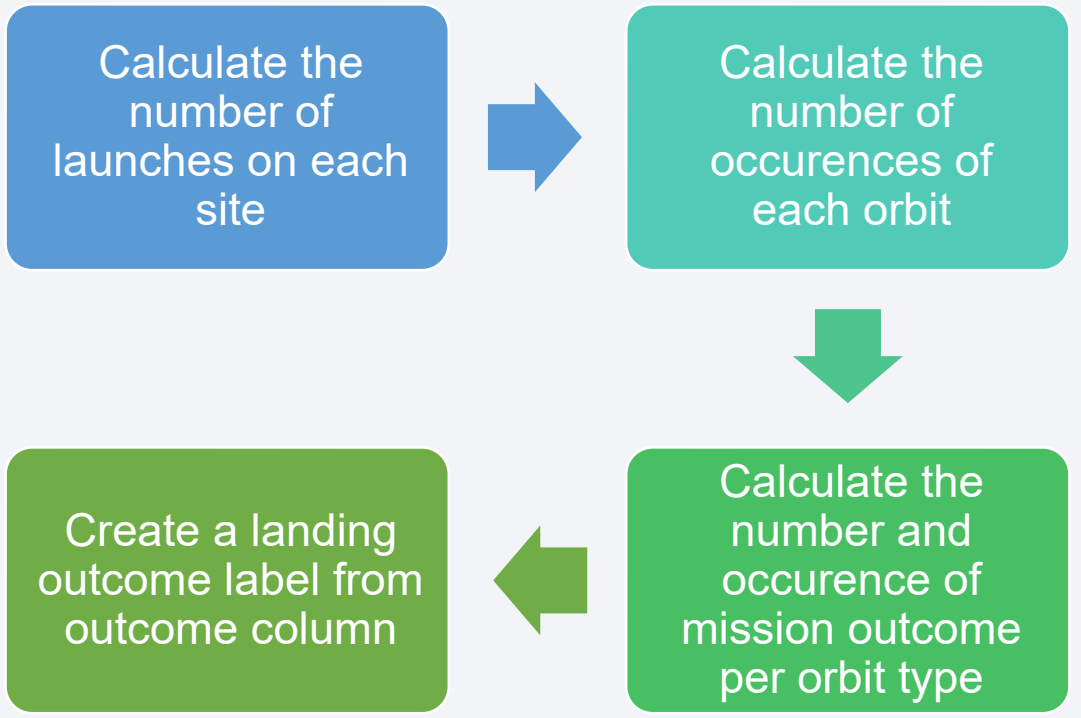
Github URL link: https://github.com/OliviaNN git/SpaceX-Falcon9-DataScience-Capstone/blob/main/Data%20Collection%20with%20webscraping.ipynb



Data Wrangling

Data Wrangling process is shown in this flow chart overview. The complete notebook can be found in the link below.

GitHub URL link: https://github.com/OliviaNNgit/SpaceX-Falcon9-DataScience-Capstone/blob/main/Data%20wrangling.ipynb



EDA with Data Visualization

Types of charts used:

Scatter plot – Flight Number vs Payload Mass, Flight Number vs Launch Sites, Payload and Launch Sites, Flight Number and Orbit Type, Payload and Orbit Type

Bar chart – Success rate of each orbit

Line plot – Success rate and Data

GitHub URL link: https://github.com/OliviaNNgit/SpaceX-Falcon9-DataScience-Capstone/blob/main/EDA%20Visualization.ipynb

EDA with SQL

Summary of SQL queries that were used:

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 that 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 booster versions which have carried the maximum payload mass. Use a subquery

List the failed landing outcomes in drone ship, their booster versions and launch site names 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 link: https://github.com/OliviaNNgit/SpaceX-Falcon9-DataScience-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Folium Markers are used to show the Space X launch sites and their nearest important landmarks like railways, highways, cities and coastlines.

Polylines were used to connect the launch sites to their nearest landmark.

Red represents rocket launch

Green represents the success

GitHub URL link: https://github.com/OliviaNNgit/SpaceX-Falcon9-DataScience-Capstone/blob/main/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Pie charts and scatter charts are used to visualize the launch records of Space X

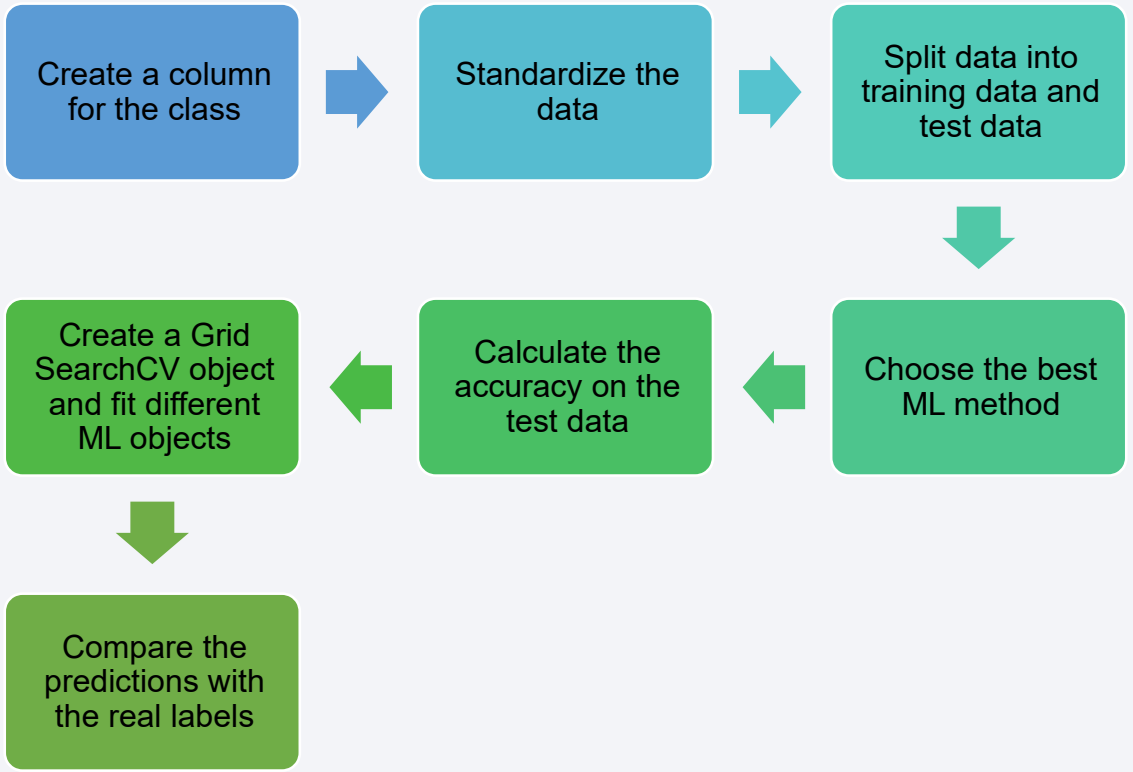
This charts displayed the rocket launch success rate per launch site. I was able to get an understanding of the factors that may have been influencing the success rate at each site. Such as the payload mass and booster versions.

Successful launches were represented by 1 while the failures were represented by 0.

Predictive Analysis (Classification)

Scikit-learn is a Machine Learning library that was used for predictive analysis. The following took place:

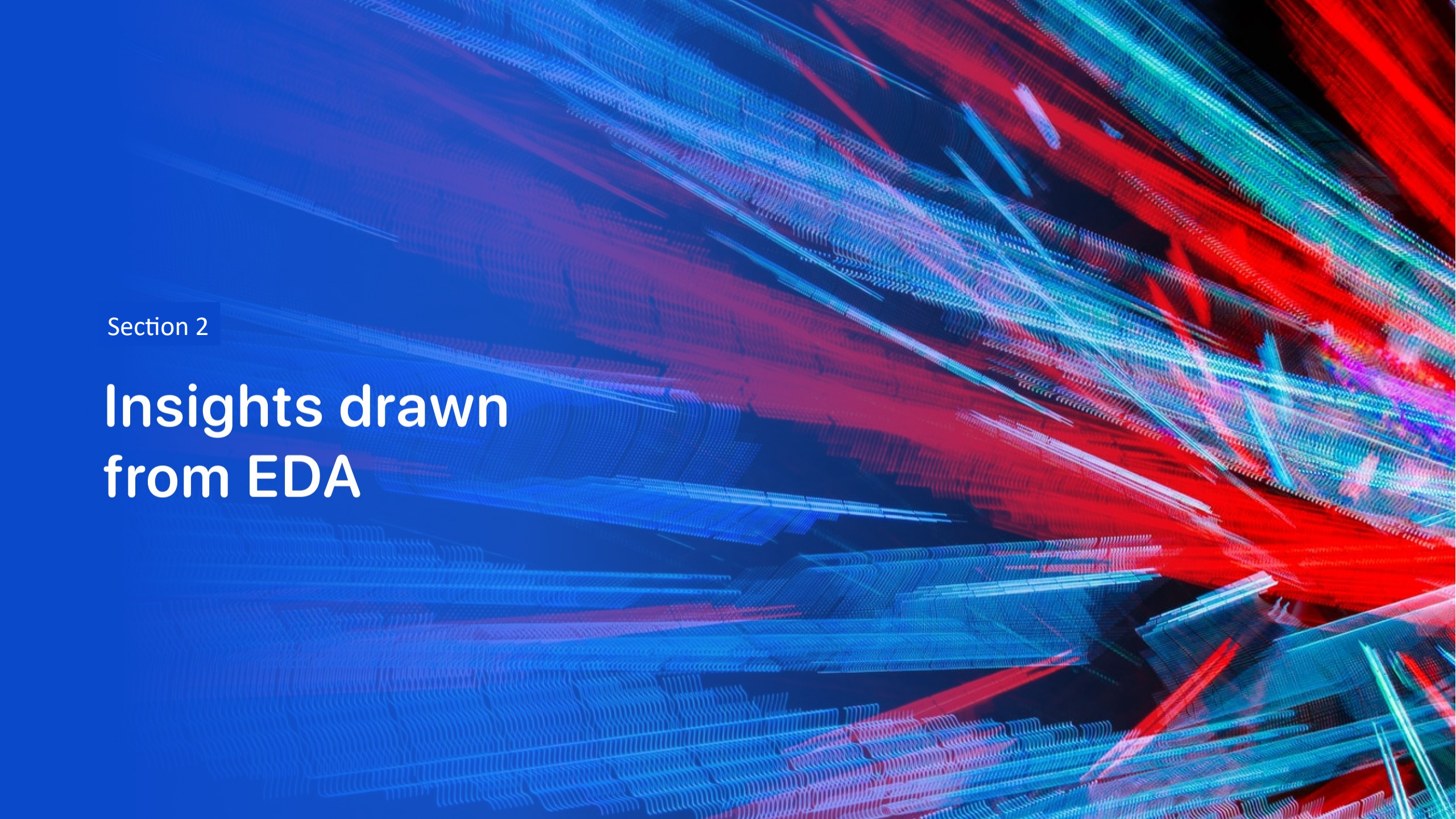
A machine learning pipeline was created to predict if the first stage will land given the data.



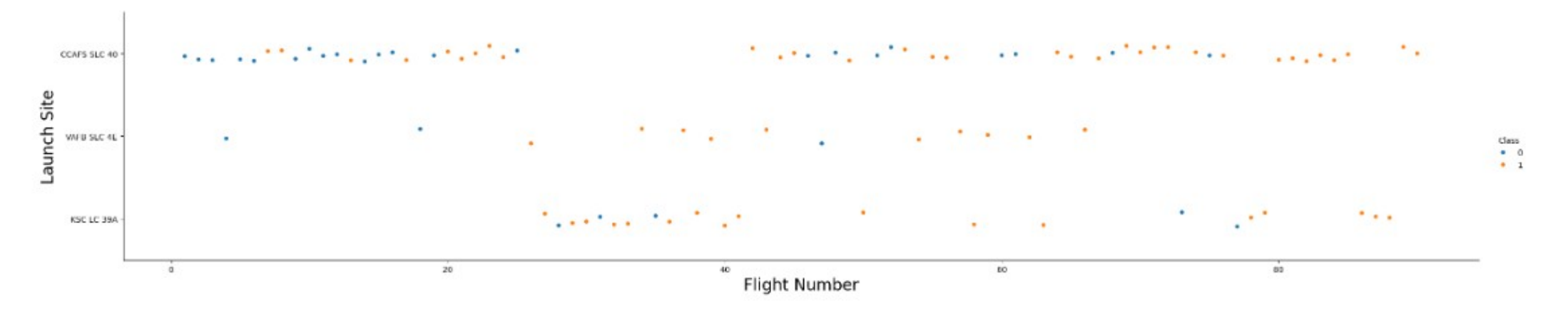
GitHub URL link:
https://github.com/OliviaNNgit/SpaceX-Falcon9-DataScience-Capstone/blob/main/SpaceX%20Machine%20Learning%20Prediction.ipynb

Results

- The exploratory data analysis has shown us that successful landing outcomes are some what correlated with flight number. It was also apparent that the successful landing outcomes have had a significant increase since the year 2015.
- All launch sites are located near the coast line. Perhaps, this makes it easier to test the rocket landings in the water.
- Sites are also located near highways and railways. This may facilitate transportation of equipment and research material.
- The machine learning model was able to predict the landing success of rockets with an accuracy of 83.33%

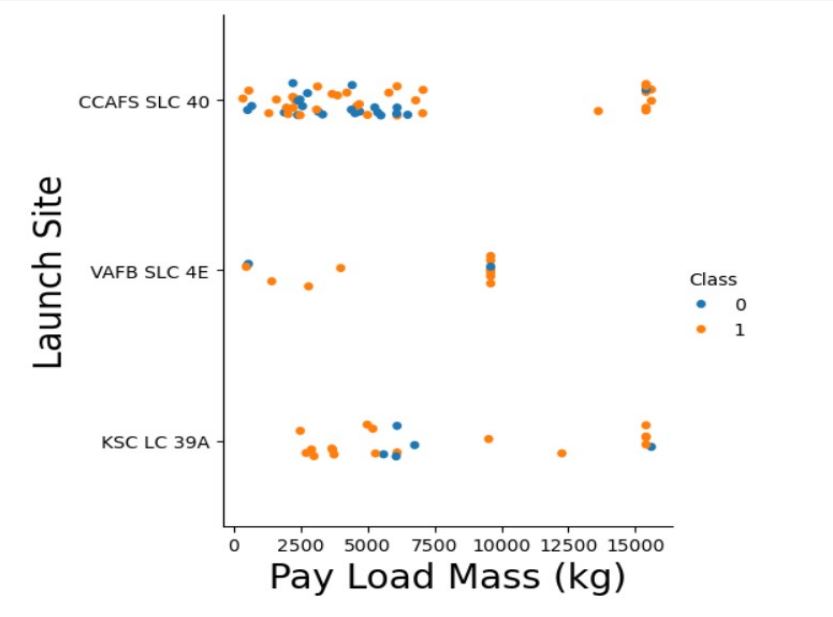

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



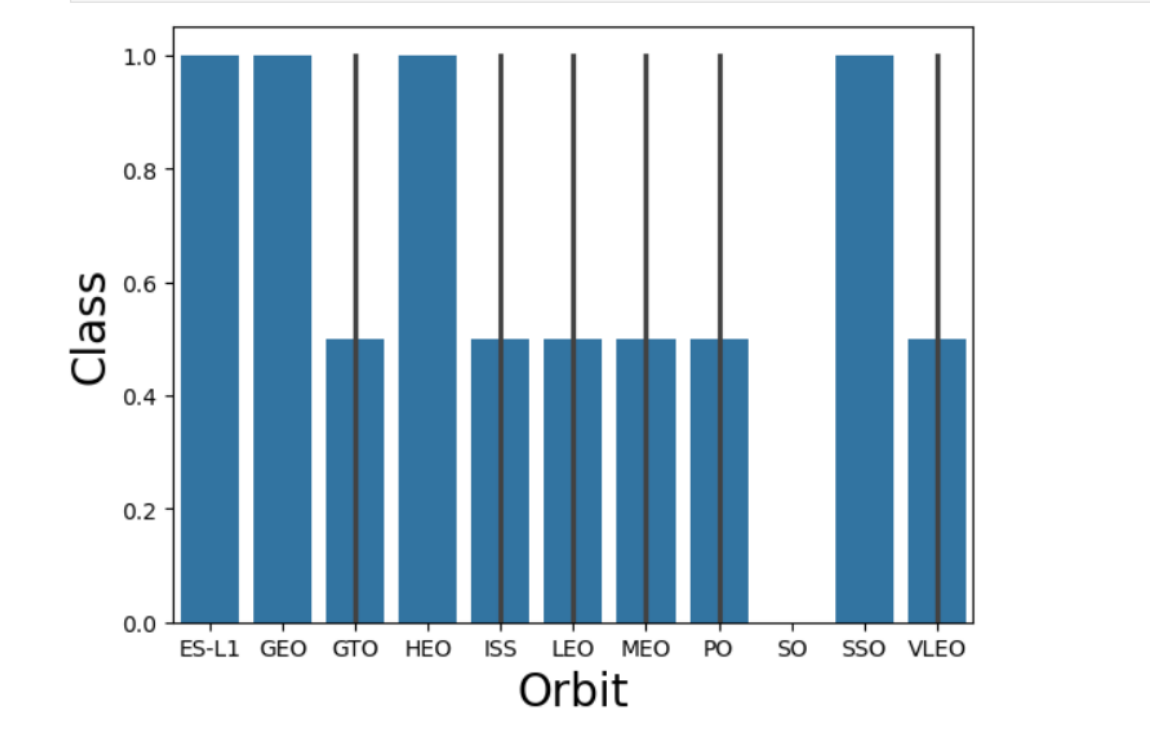
- It appears that there were more successful landings as the flight numbers increased. Laun´ch site CCAFS SLC 40 had the most number of landing.

Payload vs. Launch Site



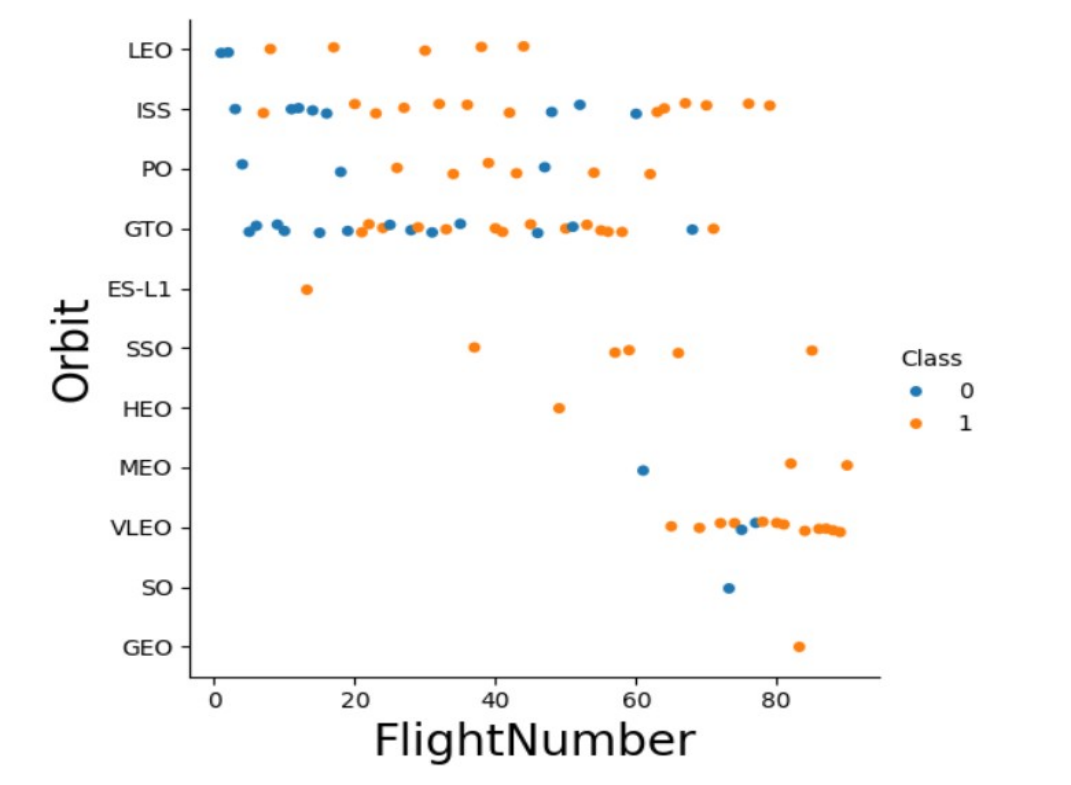
- The scatter plot chart shows that for the VAFB SLC 4E launch site, there are no rockets launched for heavy payload mass (greater than 10000)

Success Rate vs. Orbit Type



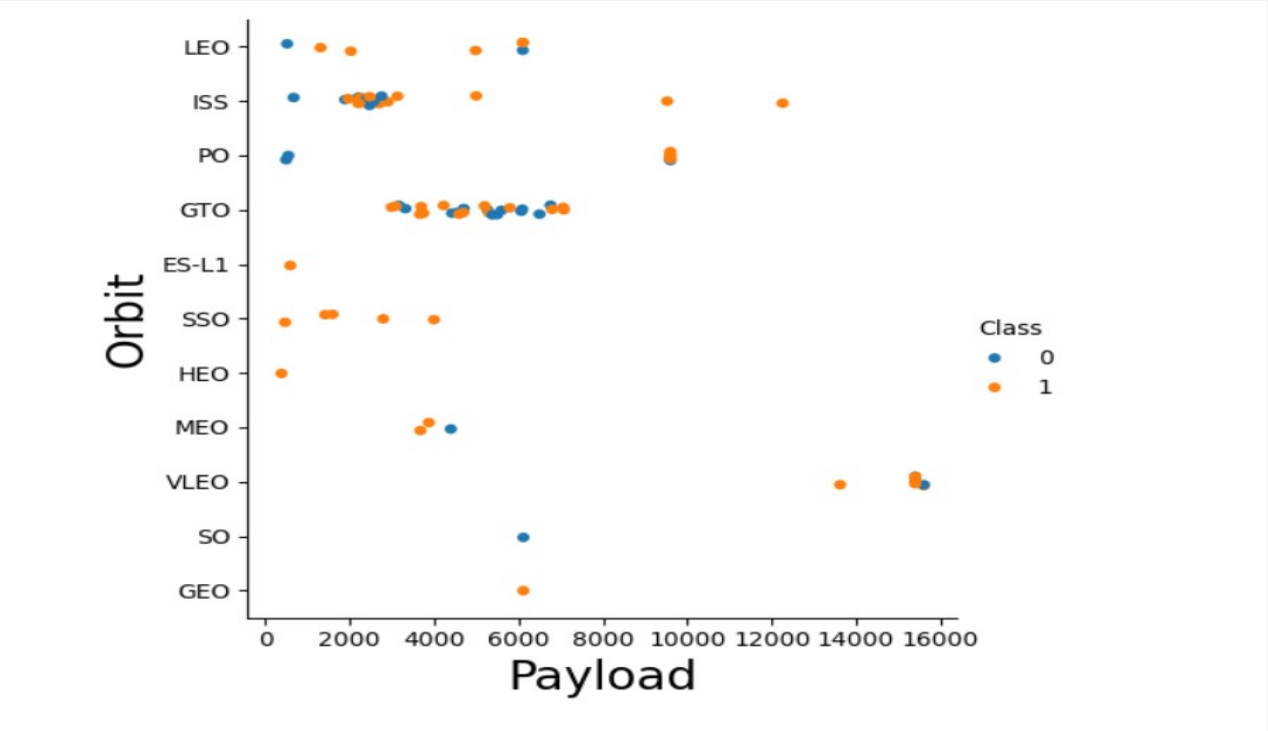
- The highest success rate Orbits are ES-L1, GEO, SSO and HEO

Flight Number vs. Orbit Type



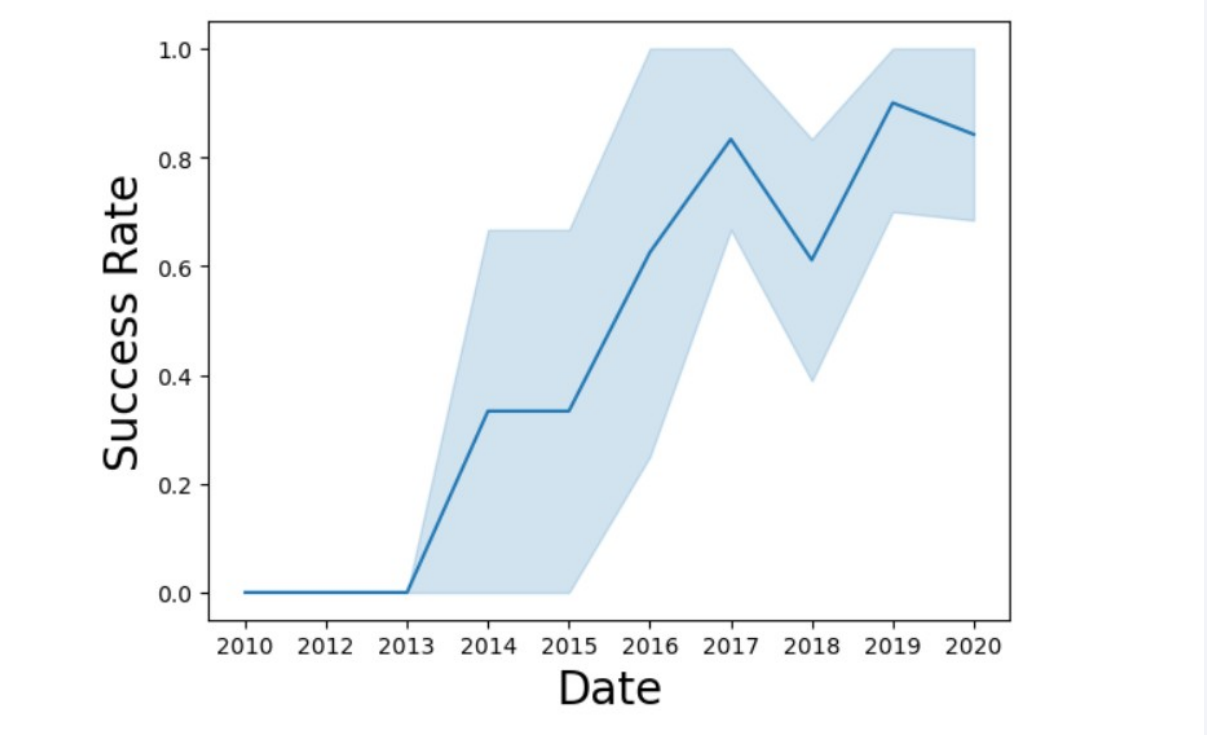
It can be seen that in the LEO orbit, the success appears related to the number of flights. On the other hand, there seems to be no relationship between the flight number and success in the GTO orbit.

Payload vs. Orbit Type



- With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISS
- However for GTO, it cannot be well distinguished since both positive landings and negative landings are there.

Launch Success Yearly Trend



It is clear that the success rate has significantly increased from 2013 to 2020.

All Launch Site Names

According to the data provided, these are the names of the launch sites where the different rocket landings were attempted:

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
In [9]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[9]:

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

These are 5 records where launch sites begin with the letters ‘CCA’. As we can see, there are other organizations beside Space X that are testing their rockets.

Total Payload Mass

Task 3

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

In [10]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA(CRS)';
```

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

Out[10]:

SUM(PAYLOAD_MASS__KG_)
None

The total payload carried by boosters from NASA is zero KG.

Average Payload Mass by F9 v1.1

Task 4

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

In [11]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
```

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

Out[11]:

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

Task 5

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

Hint:Use min function

In [12]:
```
%sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';
```

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

Out[12]:

min(DATE)
2015-12-22

The date of the first successful landing outcome on ground pad was 22nd of December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

```
[14]: ELECT BOOSTER_VERSION from SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' and  PAYLOAD_MASS__KG_  >4000 and PAYLOAD_MASS__KG_ <6000;
```

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

[14]:

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

These are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
[15]: %sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
 * sqlite:///my_data1.db
Done.
```

[15]:

count(MISSION_OUTCOME)
99

The total number of successful and failure mission outcomes is 99

Boosters Carried Maximum Payload

Task 8

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

In [15]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[15]:

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 which have carried the maximum payload mass are listed above

2015 Launch Records

Task 9

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.

[17]:
```
%sql SELECT BOOSTER_VERSION,LAUNCH_SITE,LANDING_OUTCOME FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' and DATE('2015')
```

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

[17]:

Booster_Version	Launch_Site	Landing_Outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1017	VAFB SLC-4E	Failure (drone ship)
F9 FT B1020	CCAFS LC-40	Failure (drone ship)
F9 FT B1024	CCAFS LC-40	Failure (drone ship)

Above is the list of failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

Task 10

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.

```
[22]: %sql select * from SPACEXTBL where Landing_Outcome = 'Success (ground pad)' and (DATE between '2010-06-04' and '2017-03-20') order by date desc
```

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

[22]:

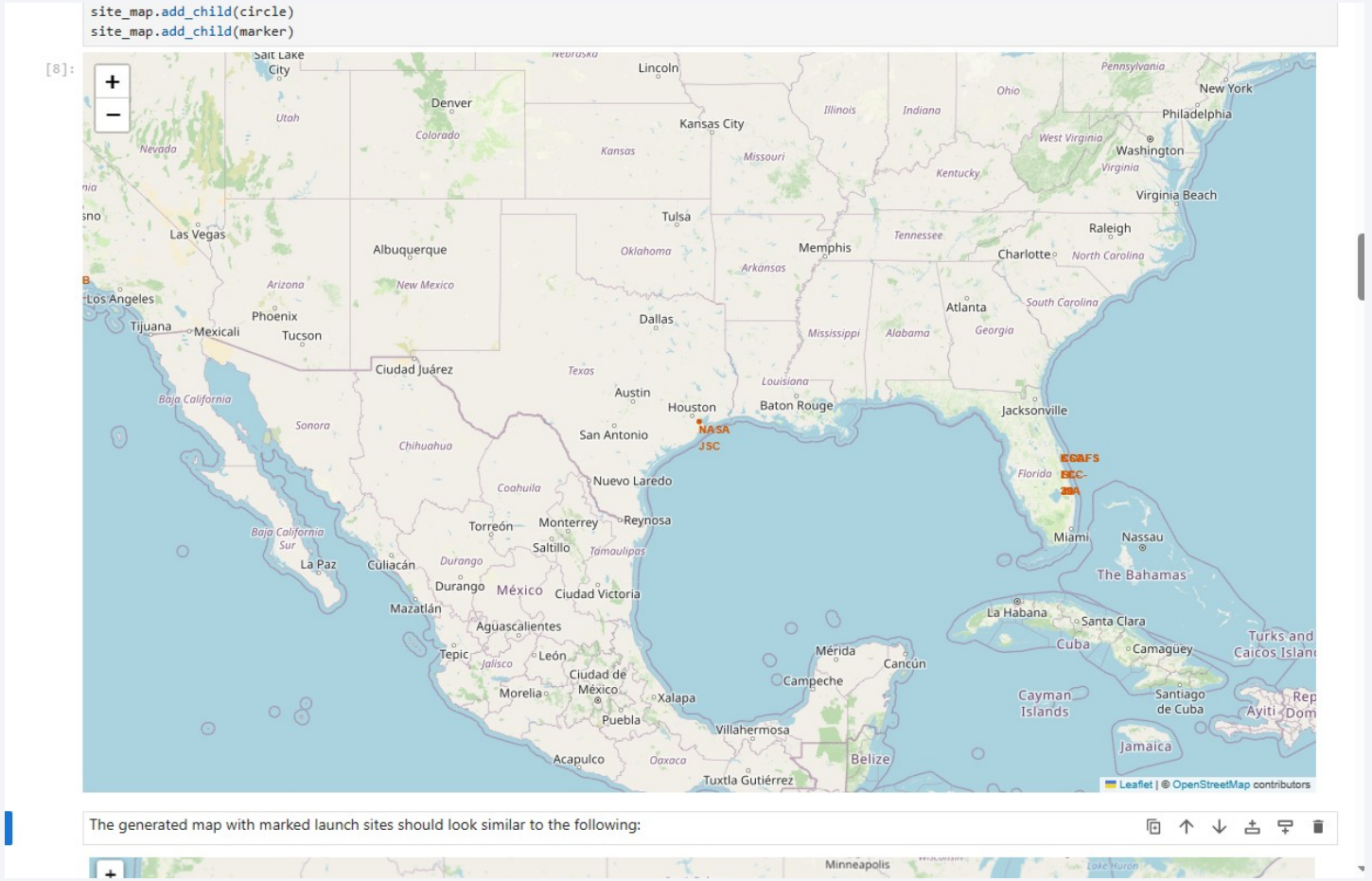
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Above is the count ranking 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

Section 3

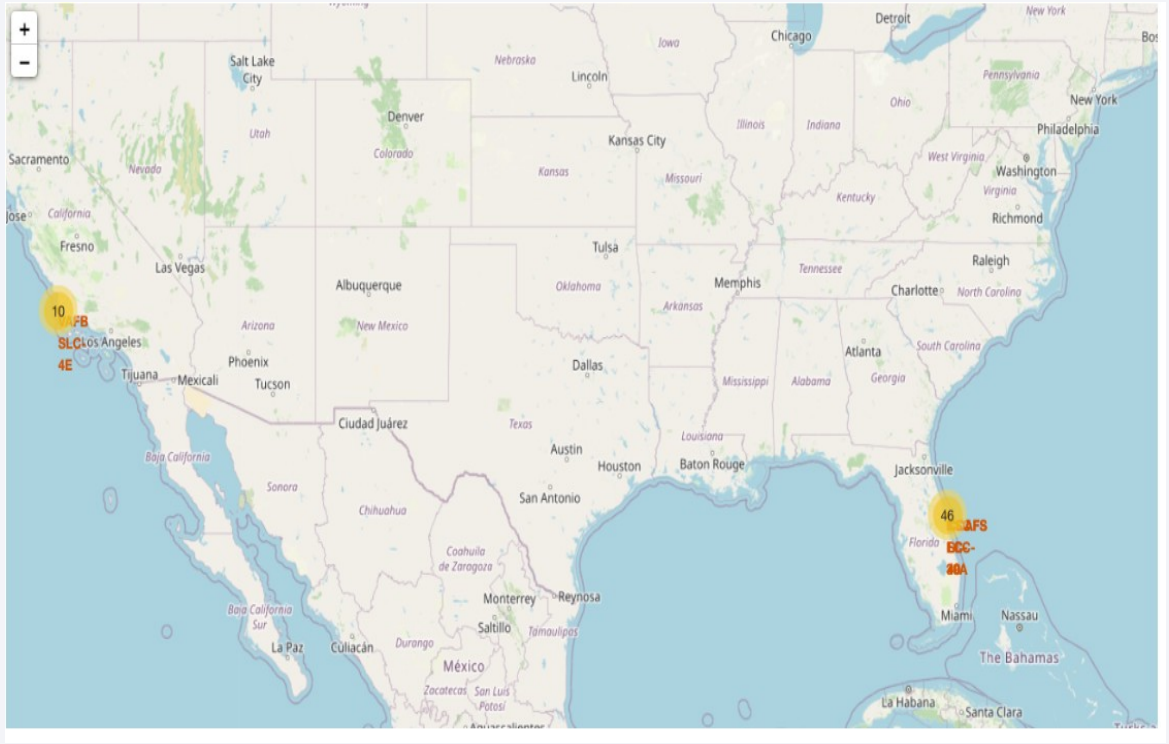
Launch Sites Proximities Analysis

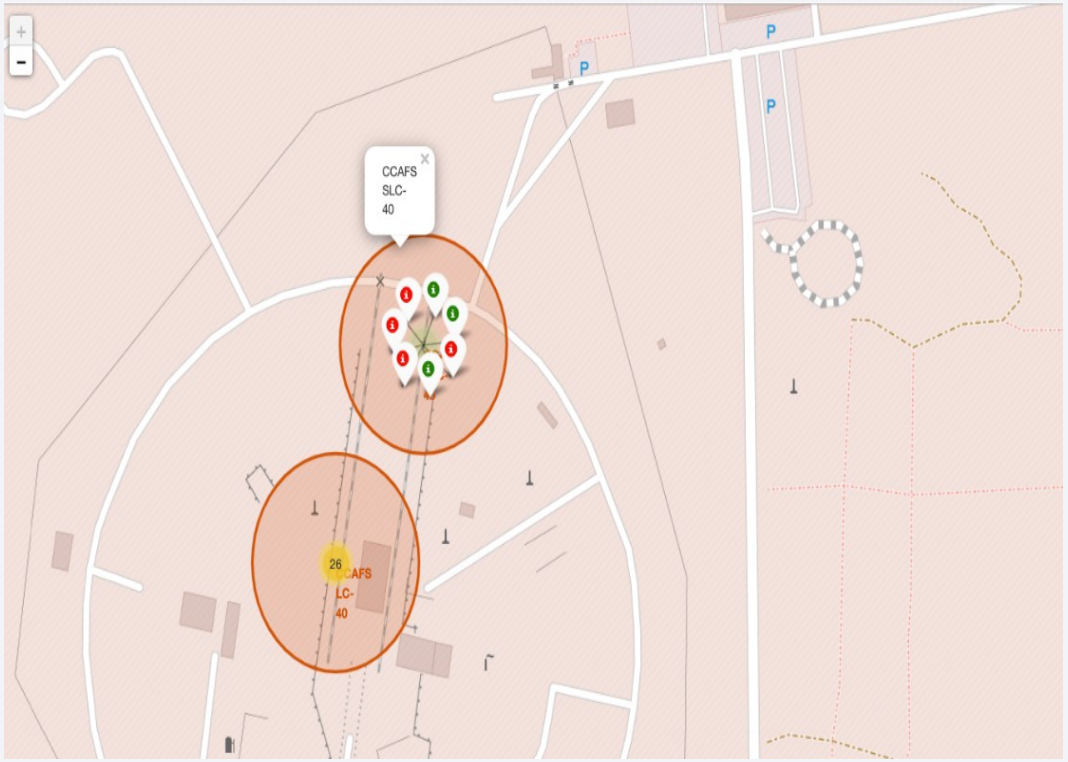
All launch sites on a map



All the 4 launch site are shown on the map in red

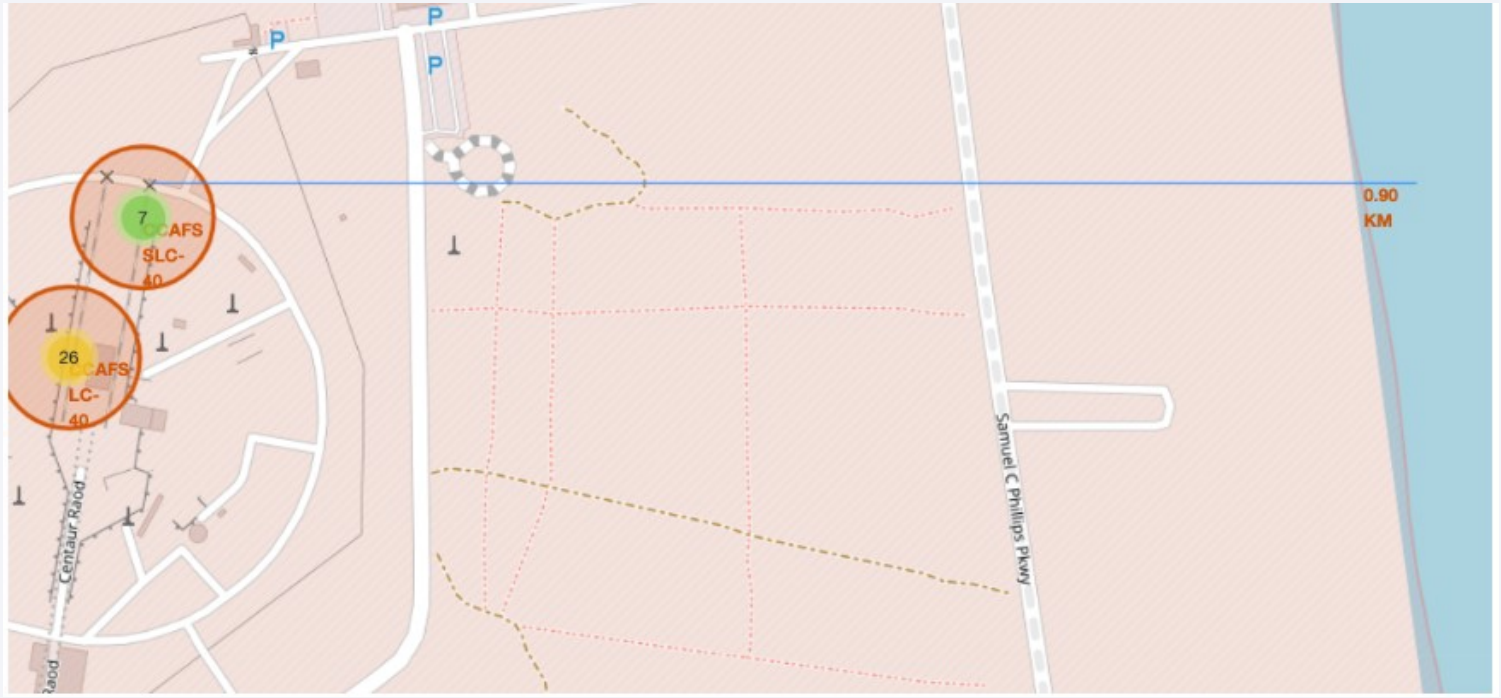
Success or failed launches for each site



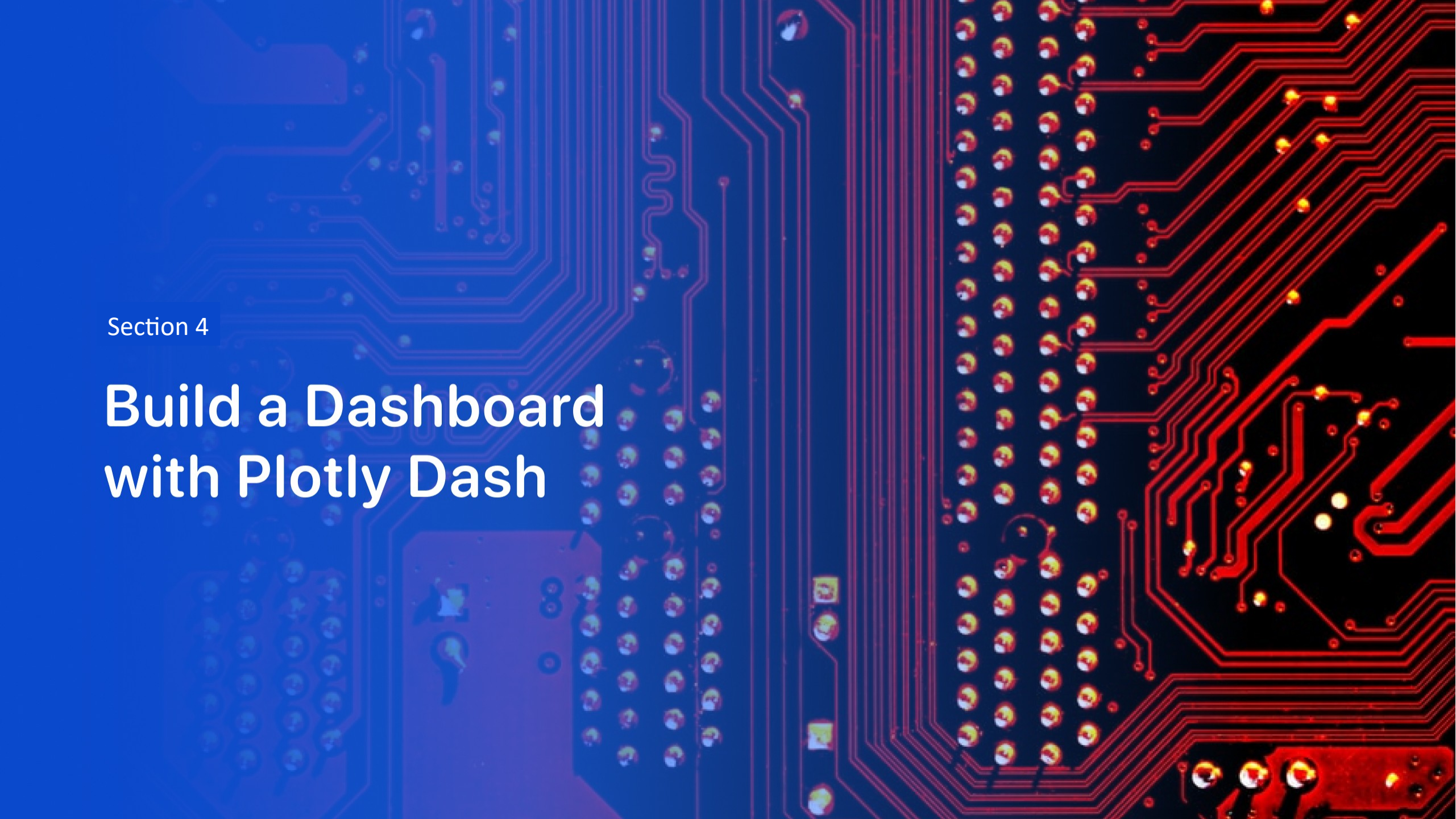


The green color-labeled marker in the marker cluster shows a successful launch while the red shows a failed launch.

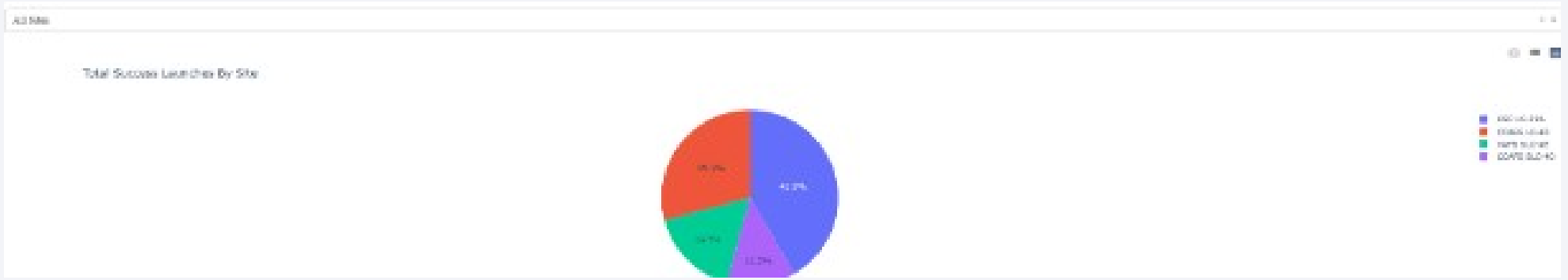
Distance between a launch site and is proximities



A line can also be drawn between a launch site to its closest city, railway or highway. A MousePosition shold be used to find the their coordinates on the map first


Section 4
Build a Dashboard with Plotly Dash

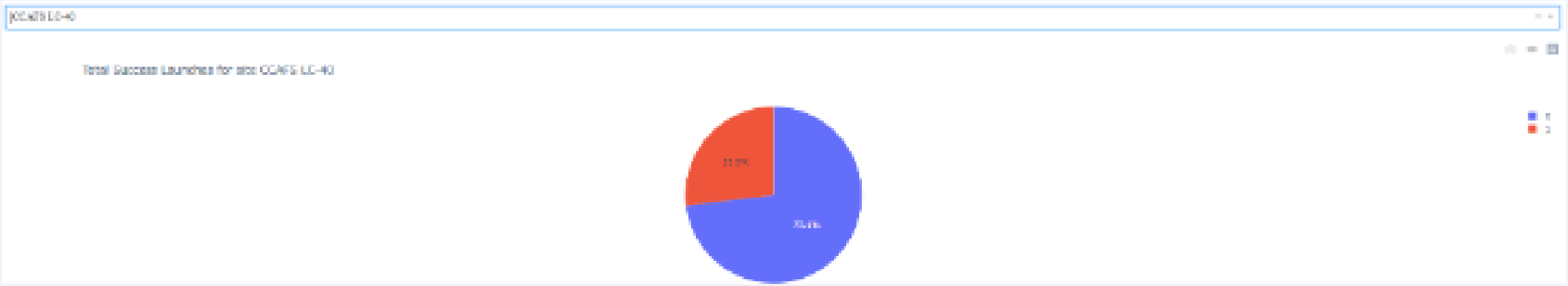
Launch success count for all sites in a piechart



The piechart shows the success count for all sites.

Github URL link: https://github.com/OliviaNNgit/SpaceX-Falcon9-DataScience-Capstone/blob/main/spacex_dash_app.py

Launch site with highest launch success ratio



The piechart for the launch site with highest launch success ratio

Githun URL Link: https://github.com/OliviaNNgit/SpaceX-Falcon9-DataScience-Capstone/blob/main/spacex_dash_app.py

Payload vs. Launch Outcome for all sites

Githun URL Link: https://github.com/OliviaNNgit/SpaceX-Falcon9-DataScience-Capstone/blob/main/spacex_dash_app.py

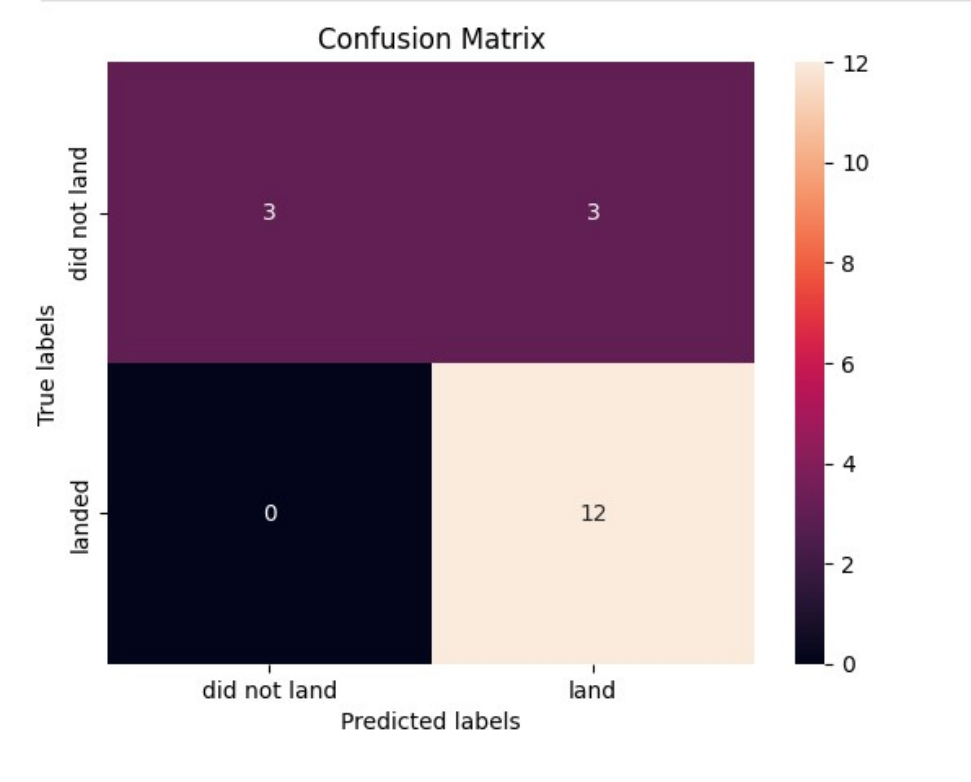
Section 5

Predictive Analysis (Classification)

Classification Accuracy

All 4 methods Support Vector Machine, Logistic Regression, K Nearest Neighbour and Decision Tree have the same accuracy

Confusion Matrix



12 launches that were predicted to land by the model landed and 3 of the 6 launches that were predicted not to land were wrongly classified.

Conclusions

In this capstone:

1. Data was successfully collected from SpaceX API and Wikipedia page titled List of Falcon 9 and Falcon Heavy launches

2. Exploratory Data Analysis was performed and Training Labels were determined

3. Trends in the data were visualized

4. Launch sites were also examined

5. Four machine learning models were used to predict if the Falcon 9 first stage will land successfully.

6. The machine learning models all have an accuracy of 83.33% to predict this outcome

Thank you!