Winning Space Race with Data Science

Zoubir Omar
30/11/2021

Outline

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

Executive Summary

- Summary of methodologies
  - **Collecting the Data**
  - **Data Wrangling**
  - **Exploratory Analysis Using SQL**
  - **Exploratory Analysis Using Pandas and Matplotlib**
  - **Interactive Visual Analytics and Dashboard**
  - **Predictive Analysis (Classification)**
- Summary of all results
  - **Plots and charts using matplotlib and seaborn**
  - **SQL queries**
  - **Folium maps**
  - **Plotly dash dashboard**

Introduction

- My name is Zoubir Omar. In this project, I will apply your data science skills as a Data scientist for a private space launch company in this project.
- The purpose of this project is to explore and analyze various data related to SpaceX rocket launch in order to find the best parameters that can predict a successful landing for the first stage of rocket launchers.

Section 1

Methodology

Methodology

Executive Summary

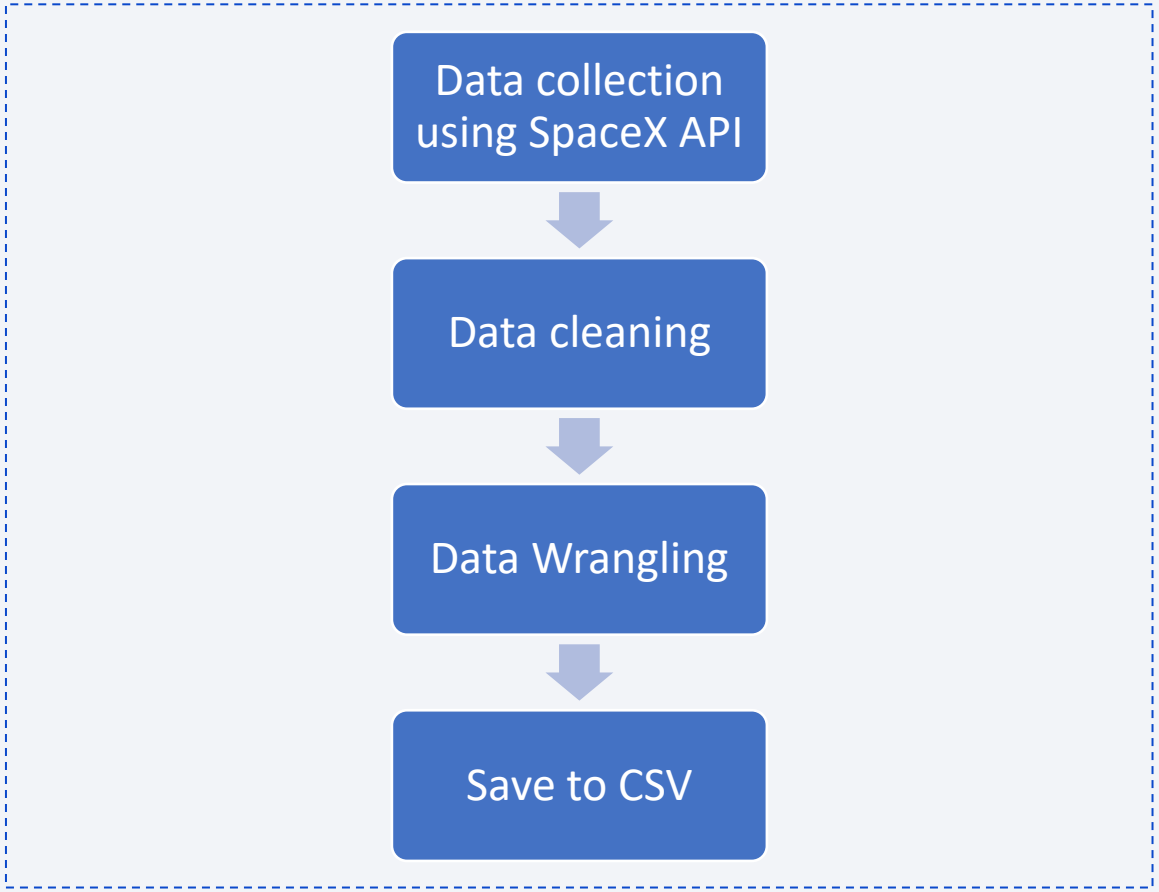
- Data collection methodology:
  - Requesting and cleaning data using the SpaceX API
  - Web scraping to collect Falcon 9 historical launch records
- Perform data wrangling
  - We get to explore the data using some SQL queries
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Data was collected both from SpaceX API and from historical launch records of the Falcon 9 From Wikipedia
- In both steps, the data was converted to pandas DataFrames and cleaned later on
- At the end we saved the data to csv files

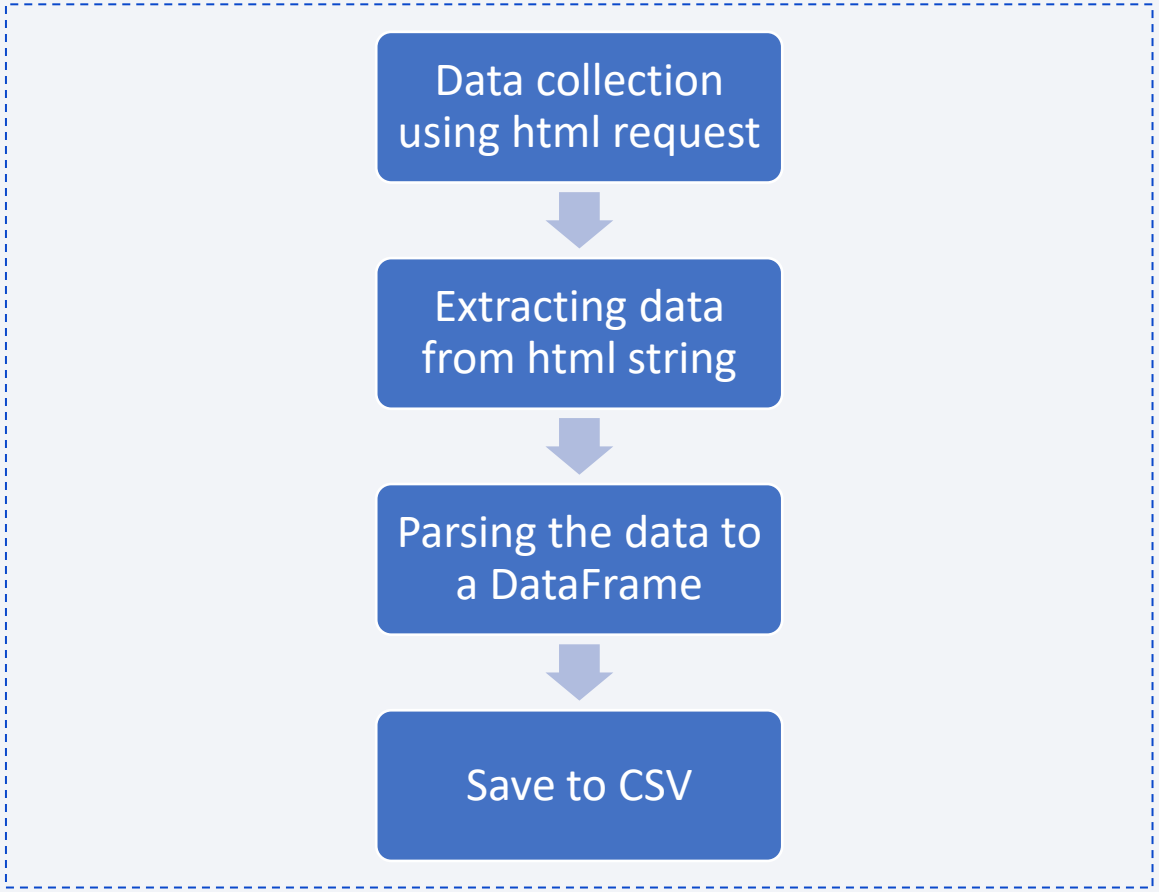
Data Collection – SpaceX API

- Data was collected first as a JSON file from the SpaceX API then parsed into a pandas DataFrame
- The collected data was normalized and wrangled by changing some column types and replacing null values with the mean
- **Notebook:** https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/5b02c80d-7557-46c3-b69d-2ea10f26e339/view?access_token=6becc8aac30954d8c35c38d96f0baae74ceeb2094e41e45484b37b8e2aa7f96a



Data Collection – Web Scraping

- Raw Falcon 9 launch history data was collected using an html request
- Relevant Data was extracted from the html string and converted to a DataFrame
- **Notebook:** https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/23e79c04-a45b-44c1-9b74-4a74cfb0fcea/view?access_token=2d3c47452edf233ac8480d1ae04bbdd2fe8e3d28fafc27990b4860dee6f4af2c



Data Wrangling

- Understanding the data and Exploring the categories of each of the categorical columns
- Formatting the data related to landing success to a Boolean value for ease of manipulation on later stages
- **Notebook:** https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/b0f3ecb3-3a87-4c63-8a6f-36f9bb886068/view?access_token=7e02c0620e20acec611eb2f07b5650cc4cd726b0778929c0a0b08d7181976977

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Using Seaborn Catplot we visualized the relationships between some parameters and their success rates.
- Catplot allows us to visualize relationships numerical and categorical variables. In addition, the hue parameter was useful to differentiate the class of each data point (aka. the landing success of each data point)
- **Notebook:** https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/9601bd7f-42a8-4d78-ac60-1481038e4fb8/view?access_token=d0495b241df64050274717d3fc5c7c31d501d73087f05960a33bfdd20b91a421

EDA with SQL

The following are the SQL data exploration tasks used on the dataset imported to ibm cloud DB2

i. *Display the names of the unique launch sites in the space mission*

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

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

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

v. *List the date when the first successful landing outcome in ground pad was acheived*

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

vii. *List the total number of successful and failure mission outcomes*

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

ix. *List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015*

x. *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*

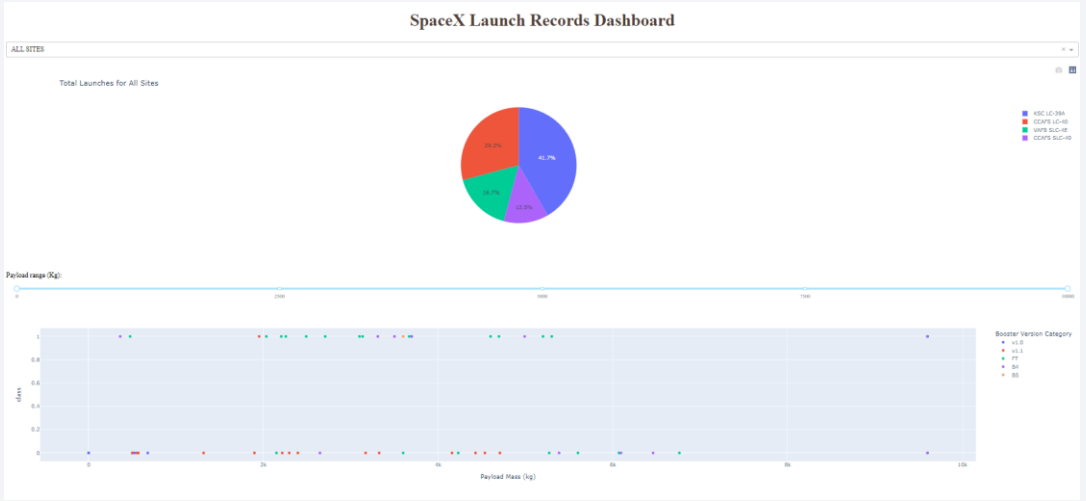
- **Notebook:** https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/de97b51c-e01a-4640-b666-b9898cc22454/view?access_token=dea219ea55baa05119ca995b309711a240411349b13cf051411eecc7f44521b0

Build an Interactive Map with Folium

- We used folium.circle and folium.marker to highlight the launch sites
- The marker color was set to red and green depending on the class of the landing
- We also calculated the distances between a launch site to its proximities
- **Notebook:** https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/aa6c2e9f-bc4b-4118-a700-3304ddbbfced/view?access_token=6eac5c30131044a0493cea3638bf25e37714af1cb87ad2c244c65c001aa44c5d

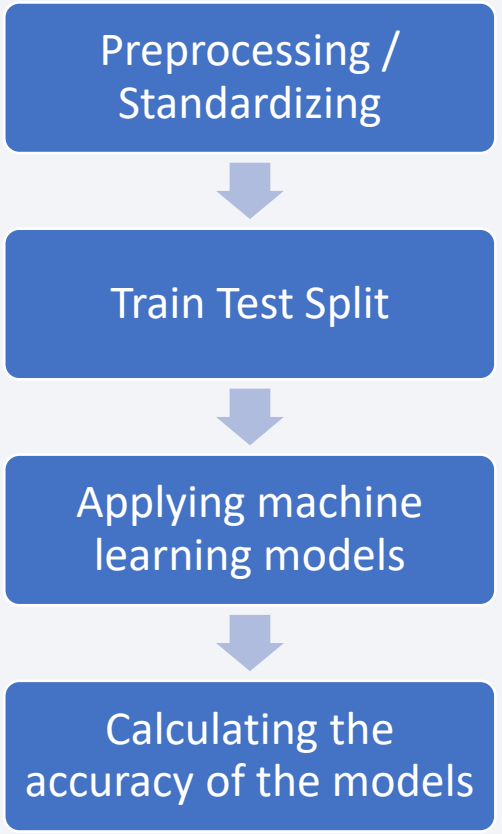
Build a Dashboard with Plotly Dash

- We used plotly dash to build the following plots/graphs
  - A pie chart showing the proportion of successful landings by landing site
  - A scatter plot visualizing the relationship between the payload mass, the booster version category and the class
- We added the input for the payload range as a slider
- **Github link:** https://github.com/MagicaOmar/Data-Science-Capstone-Project/blob/96ccb7b66b489bc96f210cf6dc3320da4b88dd83/spacex_dash_app.py



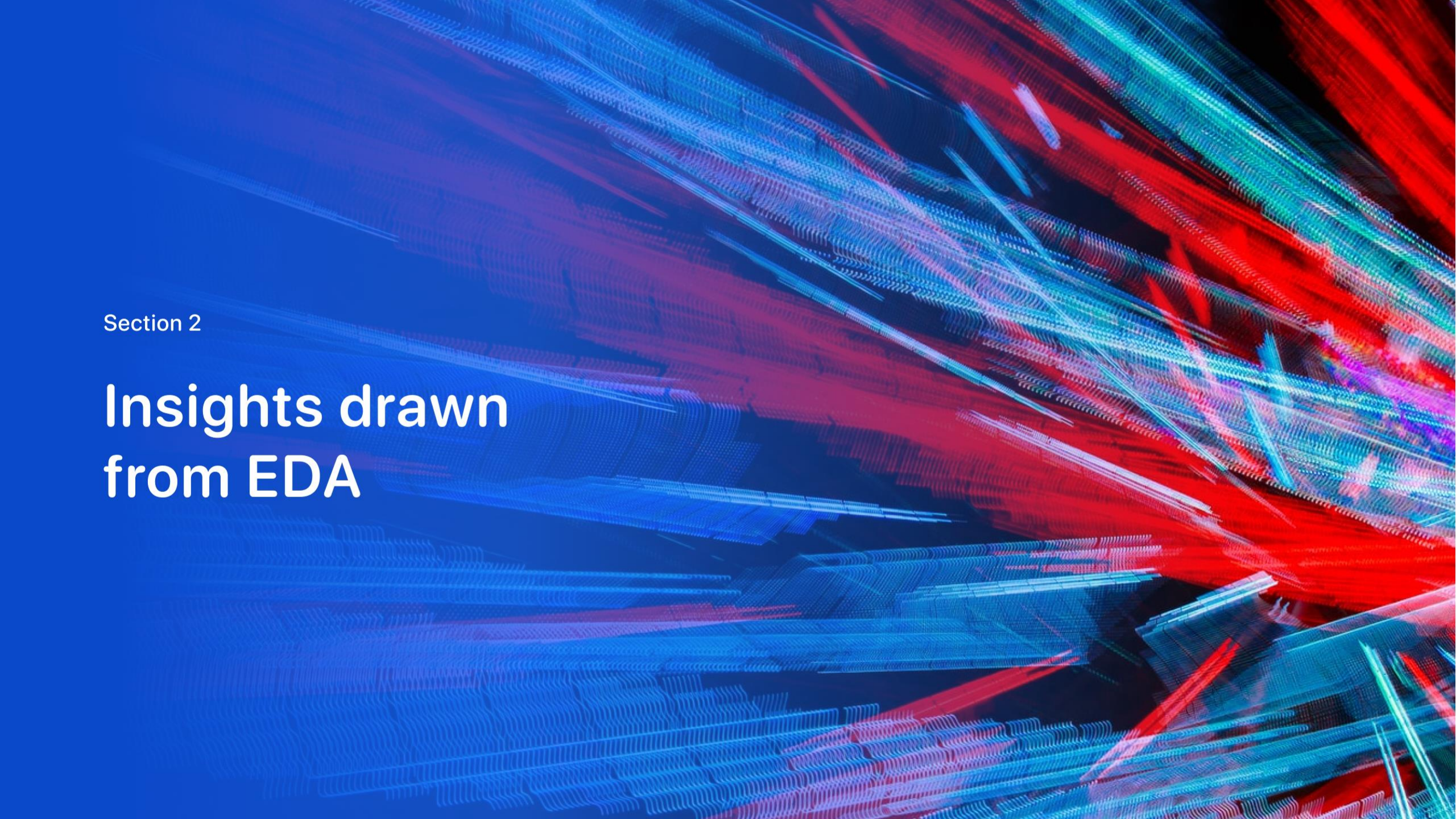
Predictive Analysis (Classification)

- We started by standardizing the data and preparing it for modeling
- We split the data to a training set and a testing set
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose



Section 2

Insights drawn from EDA



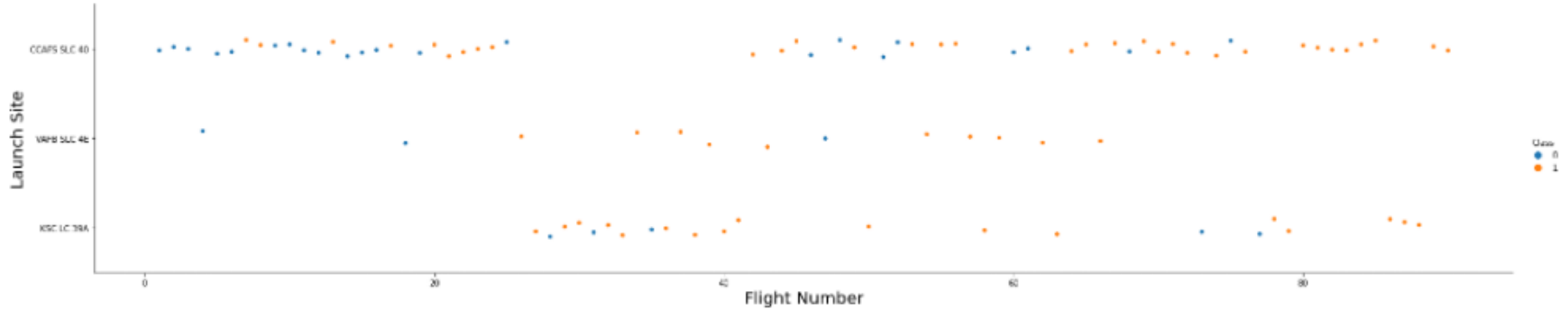
Flight Number vs. Launch Site

- Scatter plot showing the relationship between Flight numbers and Launch sites and the success for each data point

TASK 1: Visualize the relationship between Flight Number and Launch Site

Use the function `catplot` to plot `FlightNumber` vs `LaunchSite`, set the parameter `x` parameter to `FlightNumber`,set the `y` to `Launch Site` and set the parameter `hue` to `'class'`

In [4]:
```
# 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("Launch Site",fontsize=20)
plt.show()
```



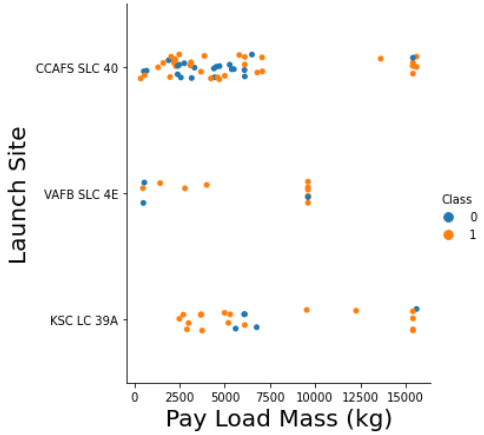
Payload vs. Launch Site

- Scatter plot showing the relationship between Payload mass and Launch sites and the success for each data point

TASK 2: Visualize the relationship between Payload and Launch Site

We also want to observe if there is any relationship between launch sites and their payload mass.

In [5]:
```
# 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 class valu
e
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df)
plt.xlabel("Pay Load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



Success Rate vs. Orbit Type

- Bar chart showing the success rate of each orbit type

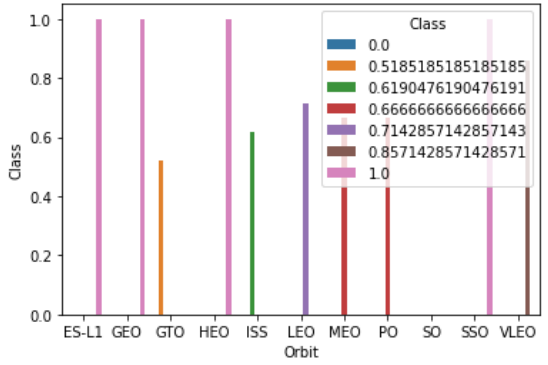
TASK 3: Visualize the relationship between success rate of each orbit type

Next, we want to visually check if there are any relationship between success rate and orbit type.

Let's create a `bar chart` for the sucess rate of each orbit

In [8]:
```
# HINT use groupby method on Orbit column and get the mean of Class column
orbit_success = df.groupby('Orbit').mean()
orbit_success.reset_index(inplace=True)
sns.barplot(x="Orbit",y="Class",data=orbit_success,hue='Class')
```

Out[8]: `<AxesSubplot:xlabel='Orbit', ylabel='Class'>`



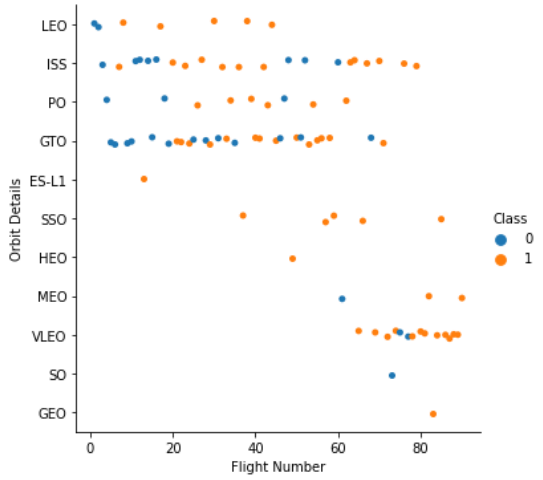
Flight Number vs. Orbit Type

- Scatter plot showing the relationship between flight number and orbit type and the success for each data point

TASK 4: Visualize the relationship between FlightNumber and Orbit type

For each orbit, we want to see if there is any relationship between FlightNumber and Orbit type.

In [10]:
```
# 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(x='FlightNumber',y='Orbit',data=df,hue='Class')
plt.xlabel('Flight Number')
plt.ylabel('Orbit Details')
plt.show()
```



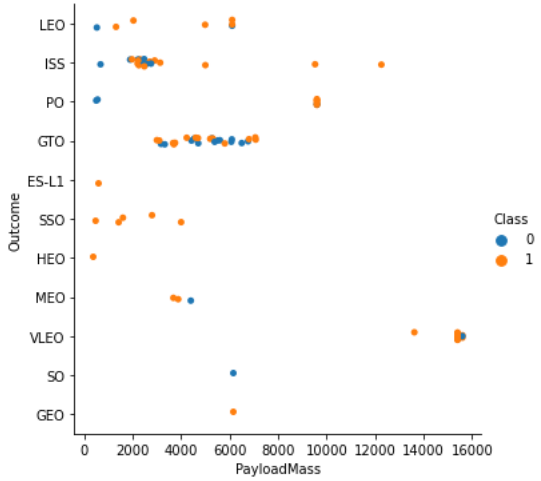
Payload vs. Orbit Type

- Scatter plot showing the relationship between payload and orbit type and the success for each data point

TASK 5: Visualize the relationship between Payload and Orbit type

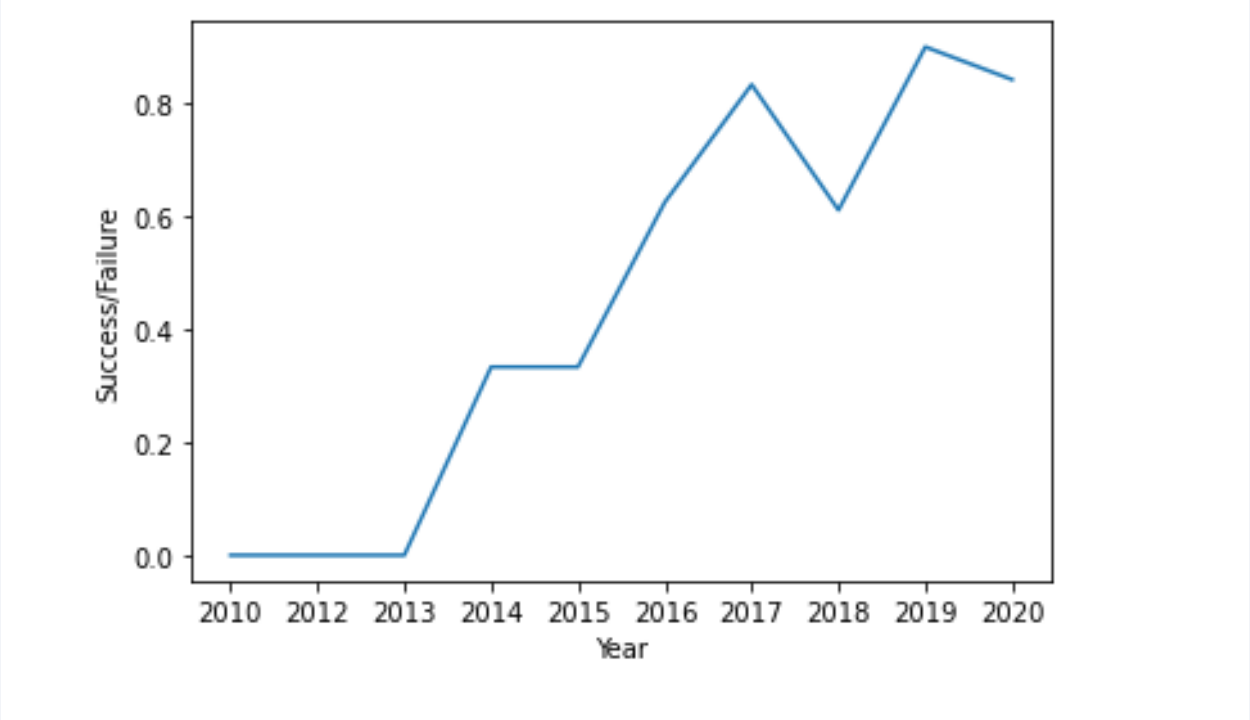
Similarly, we can plot the Payload vs. Orbit scatter point charts to reveal the relationship between Payload and Orbit type

In [9]:
```
# 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(x='PayloadMass',y='Orbit',data=df,hue='Class')
plt.xlabel('PayloadMass')
plt.ylabel('Outcome')
plt.show()
```



Launch Success Yearly Trend

- A line chart of yearly average success rate
- It shows an increase in success rate with the increase of years



All Launch Site Names

- Unique launch sites

Task 1

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

In [5]:
```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

```
 * ibm_db_sa://vqk72733:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[5]:

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

Launch Site Names Begin with 'CCA'

- Records where launch sites begin with `CCA`

Task 2

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

In [6]:
```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

```
 * ibm_db_sa://vqk72733:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[6]:

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)
2012-08-10	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
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Total Payload Mass

- The total payload carried by boosters from NASA

Task 3

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

In [7]:
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://vqk72733:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[7]:

1
22007

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

Task 4

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

In [8]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://vqk72733:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[8]:

1
3676

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad

Task 5

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

Hint:Use min function

In [9]:
```
%sql select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://vqk72733:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[9]:

1
2017-01-05

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [10]:
```
%sql select BOOSTER_VERSION from SPACEXTBL where Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

```
 * ibm_db_sa://vqk72733:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[10]:

booster_version
F9 FT B1022
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

In [11]:
```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
```

```
 * ibm_db_sa://vqk72733:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[11]:

1
44

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass

Task 8

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

In [12]:
```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * ibm_db_sa://vqk72733:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[12]:

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

2015 Launch Records

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

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [15]:
```
%sql SELECT MONTH(DATE),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where EXTRACT(YEAR FROM DATE)='2015';
```

```
 * ibm_db_sa://vqk72733:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[15]:

1	mission_outcome	booster_version	launch_site
10	Success	F9 v1.1 B1012	CCAFS LC-40
11	Success	F9 v1.1 B1013	CCAFS LC-40
2	Success	F9 v1.1 B1014	CCAFS LC-40

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

- 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

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

In [14]:
```
%sql select * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by da
te desc
```

```
 * ibm_db_sa://vqk72733:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

Out[14]:

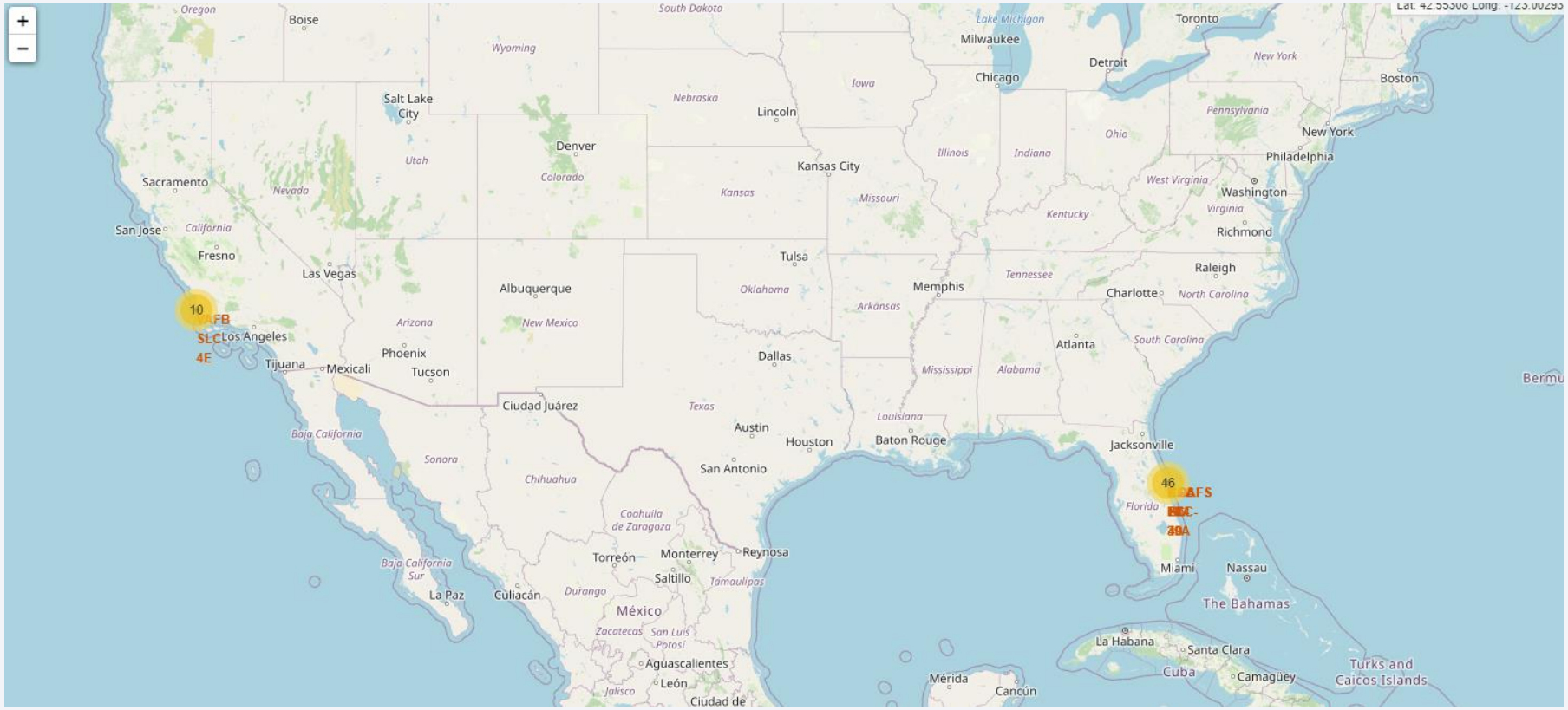
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-03-06	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-01-05	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2016-08-04	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2016-06-05	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)

Section 4

Launch Sites Proximities Analysis

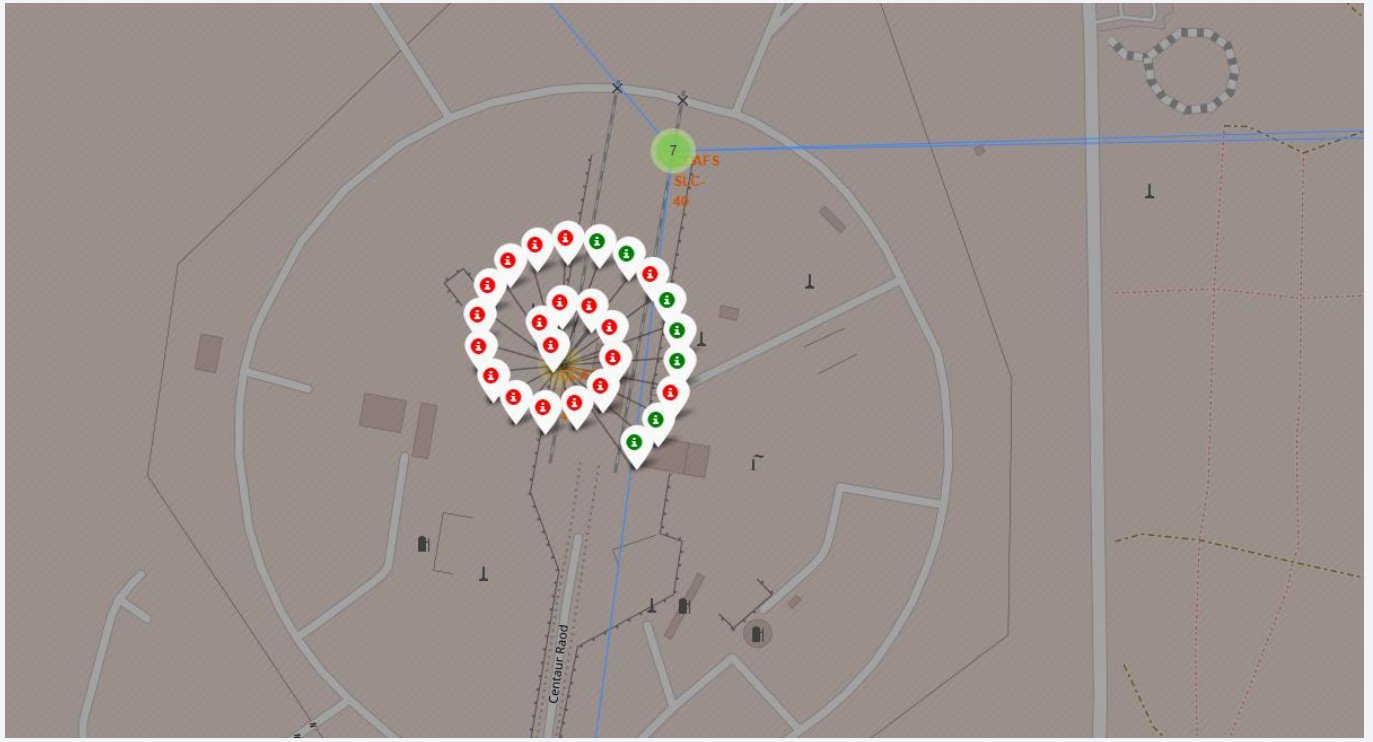
Location markers

- Placing circle markers on the map for each launch site



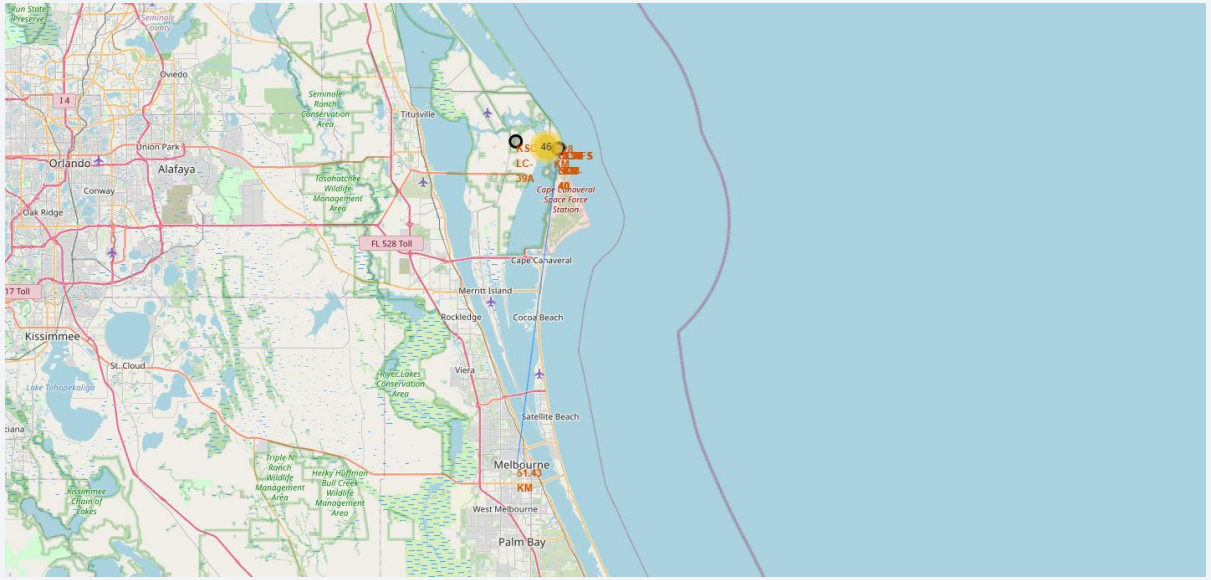
Launch outcome

- Color coding launch outcome using colored markers (Red for failure and green for success)



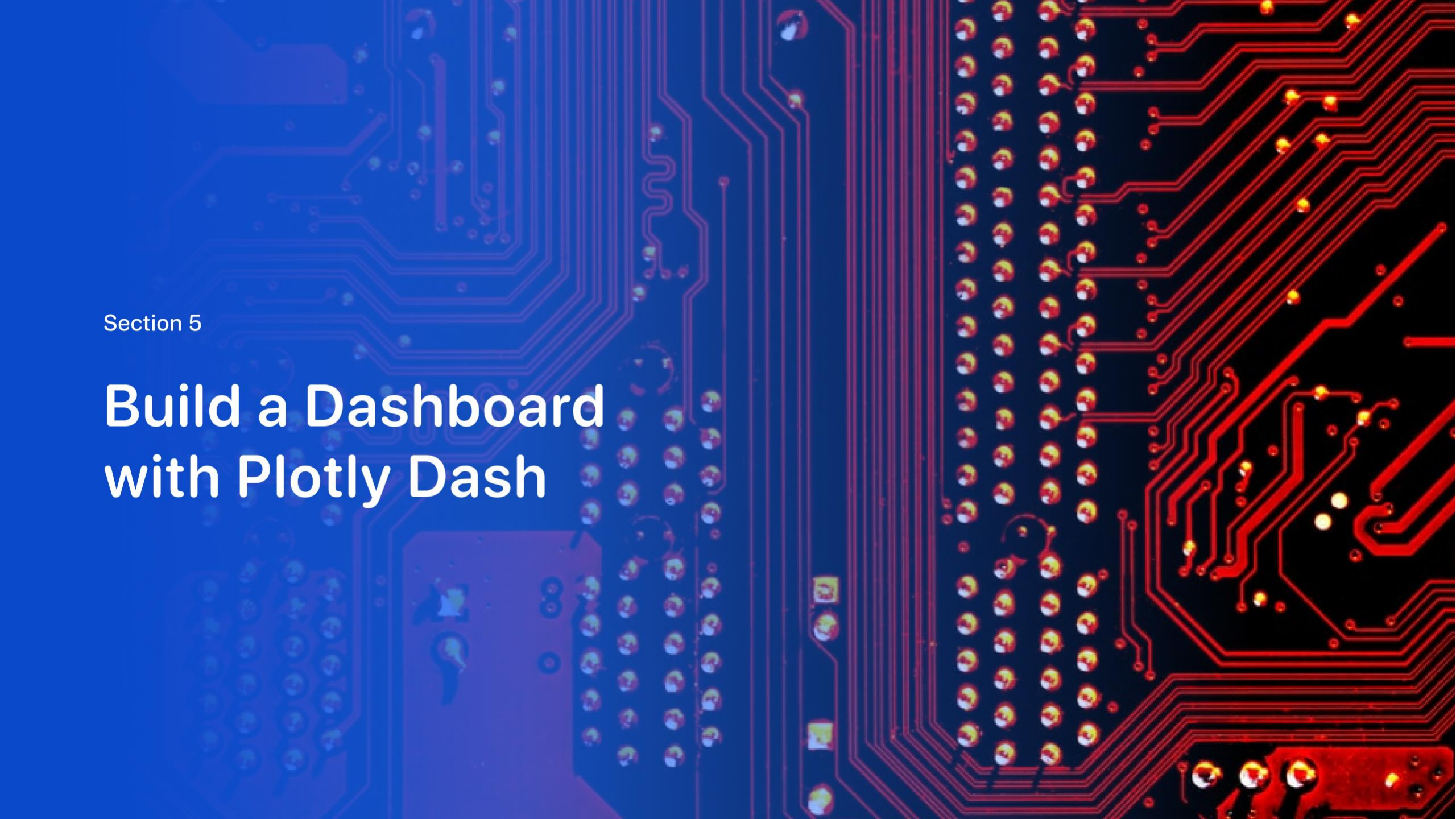
Proximity of launch sites

- We updated the map to show the proximity to railways, highways and coastlines
- Showcasing the distance as well as the successful and unsuccessful landings allows us to chose the best landing site



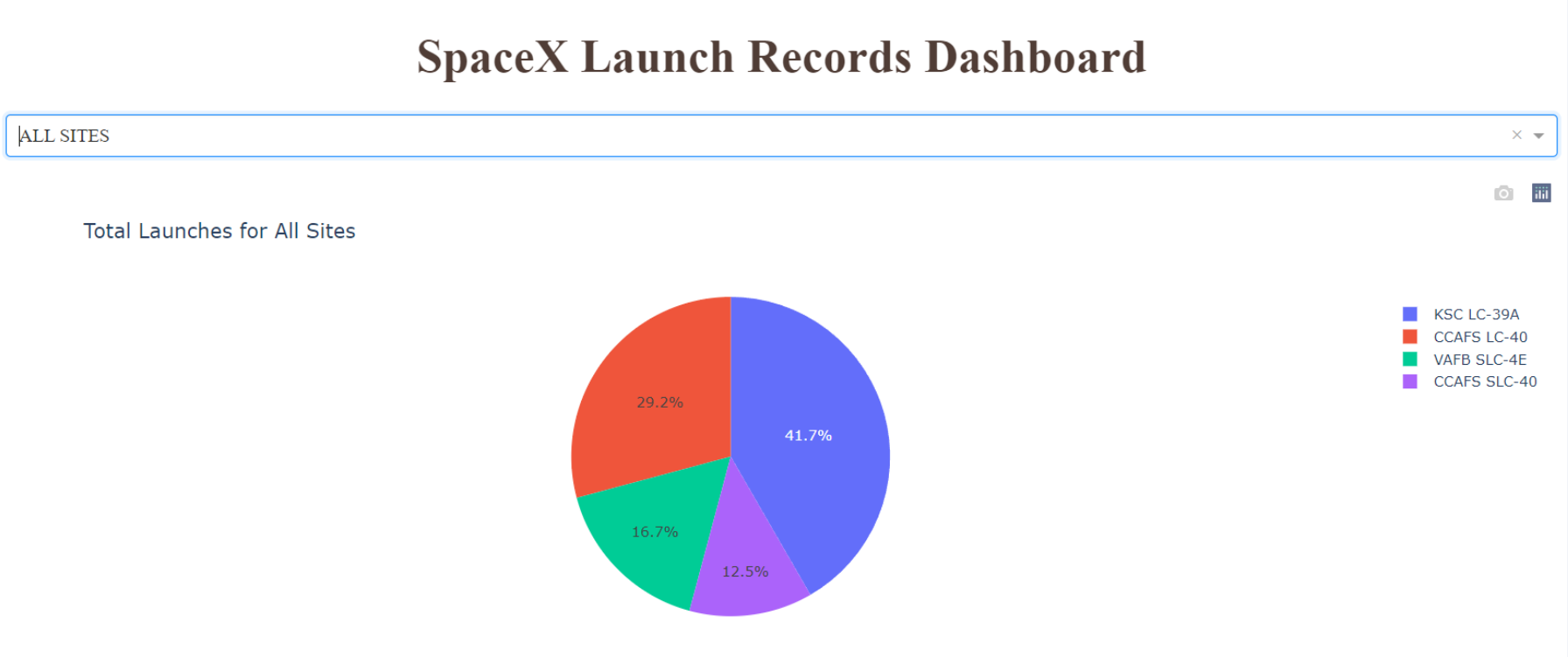
Section 5

Build a Dashboard with Plotly Dash



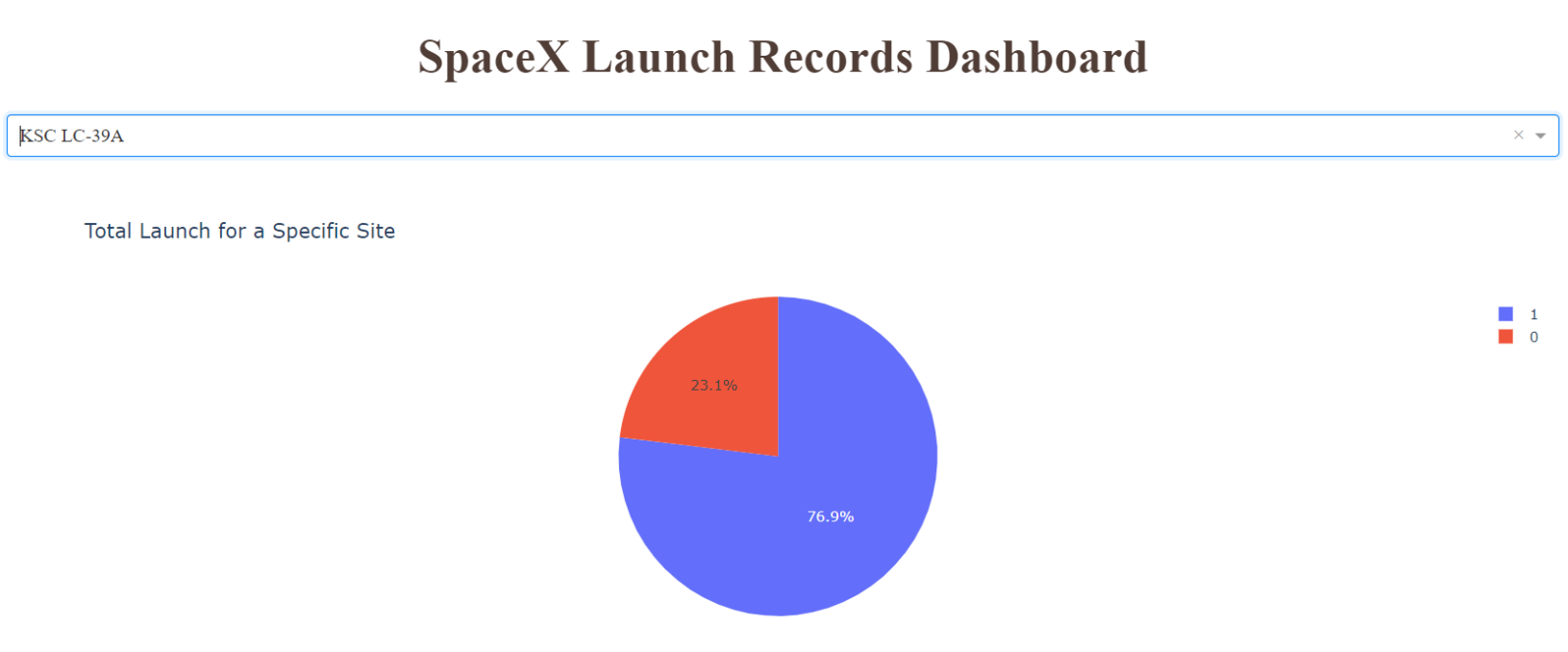
Proportion of successful launches

- A pie chart showing the proportion of successful launches by site



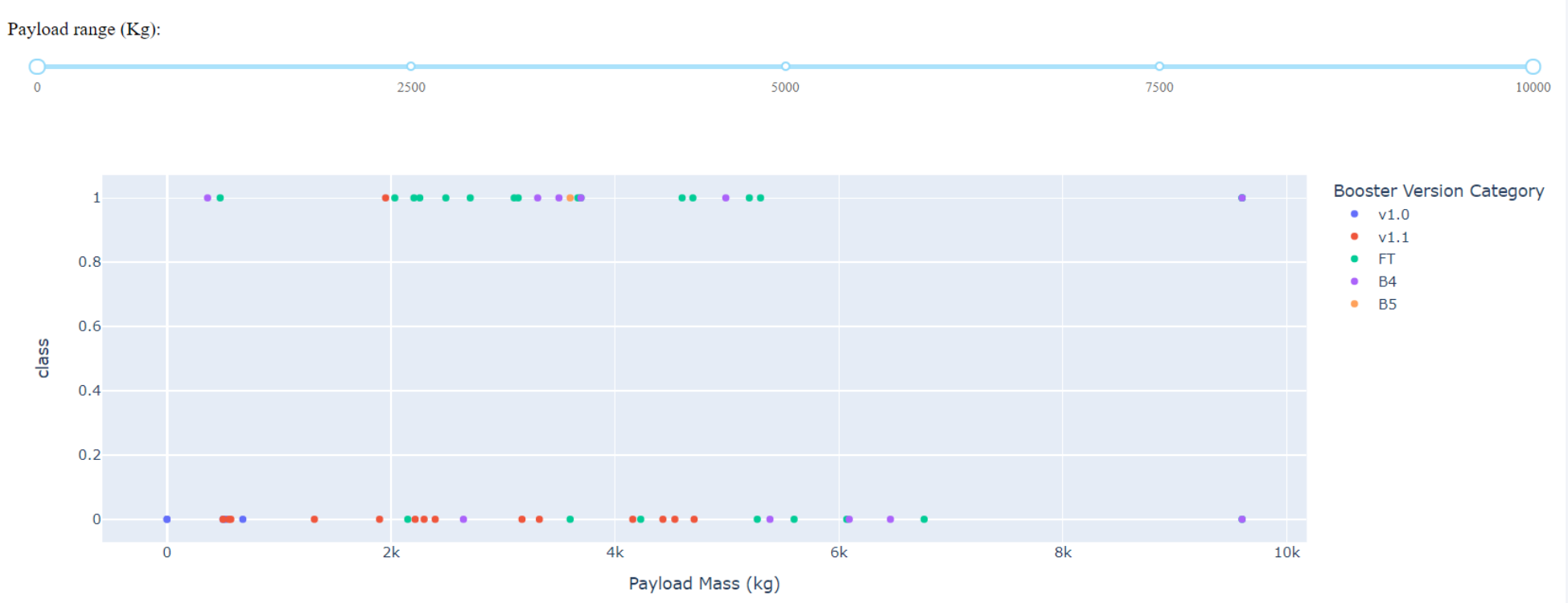
Success ration

- Pie chart showing the success ratio in a selected launch site



Booster version & Payload mass

- An interactive scatter plot showing the relationship between the payload mass, the class and the booster version category

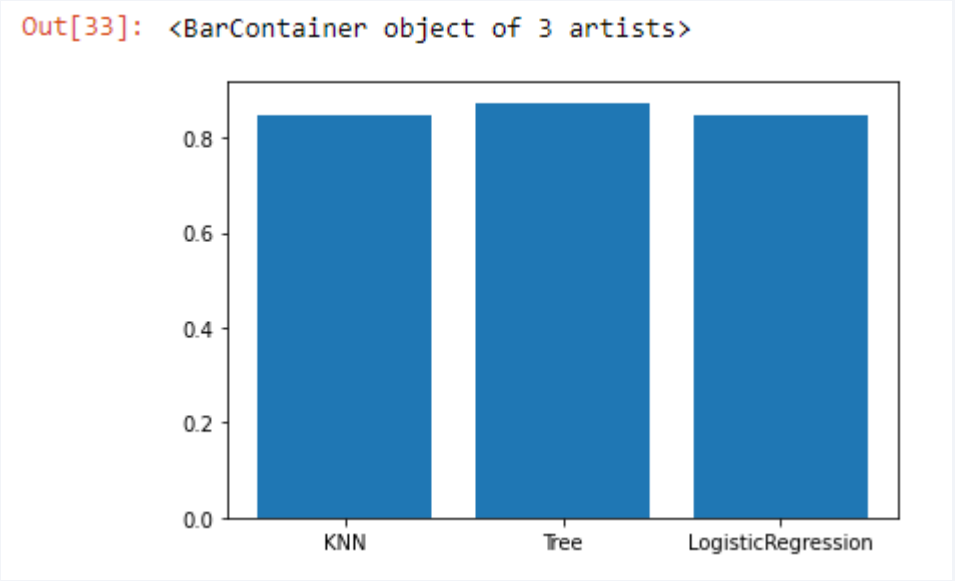


Section 6

Predictive Analysis (Classification)

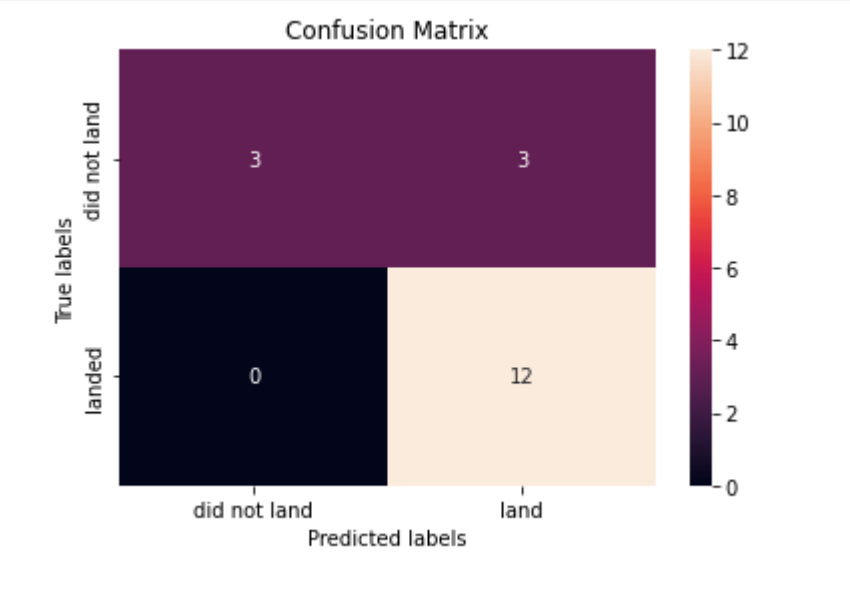
Classification Accuracy

- A bar chart visualizing the built model accuracy for all built classification models,
- The decision tree is the most accurate model



Confusion Matrix

- The confusion matrix of the best performing model (decision tree) with a correctly predicted negative landings and a weak prediction for positive ones.



Conclusions

- Launch site with the most success rate is KSC LC 39A
- High success rate for the booster version FT
- The decision tree is the most accurate model, although it’s not very precise.
- The landing sites are in a safe area away from cities and roadways

Appendix

- Github link: https://github.com/MagicaOmar/Data-Science-Capstone-Project

Thank you!