IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ifeanyi Anthony Okpala
31st December, 2022.

Outline

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

Executive Summary

Summary of methodologies:

- Data collection
- Data wrangling
- EDA with SQL
- EDA with data visualization
- Using Folium to build interactive maps
- Building a Dashboard with Plotly Dash
- Predictive analysis with Machine Learning

Summary of all results:

- EDA results
- Interactive analytics
- Predictive analysis

Introduction

Project background and context

- Our goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.
- You will be collecting data from various sources. After your raw data has been collected, you will need to improve the quality by performing data wrangling.
- We can then use these features with machine learning to automatically predict if the first
- stage can land successfully

Problems you want to find answers

- The project task is to predicting if the first stage of the SpaceX Falcon 9 rocket will land successfully

Section 1

Methodology

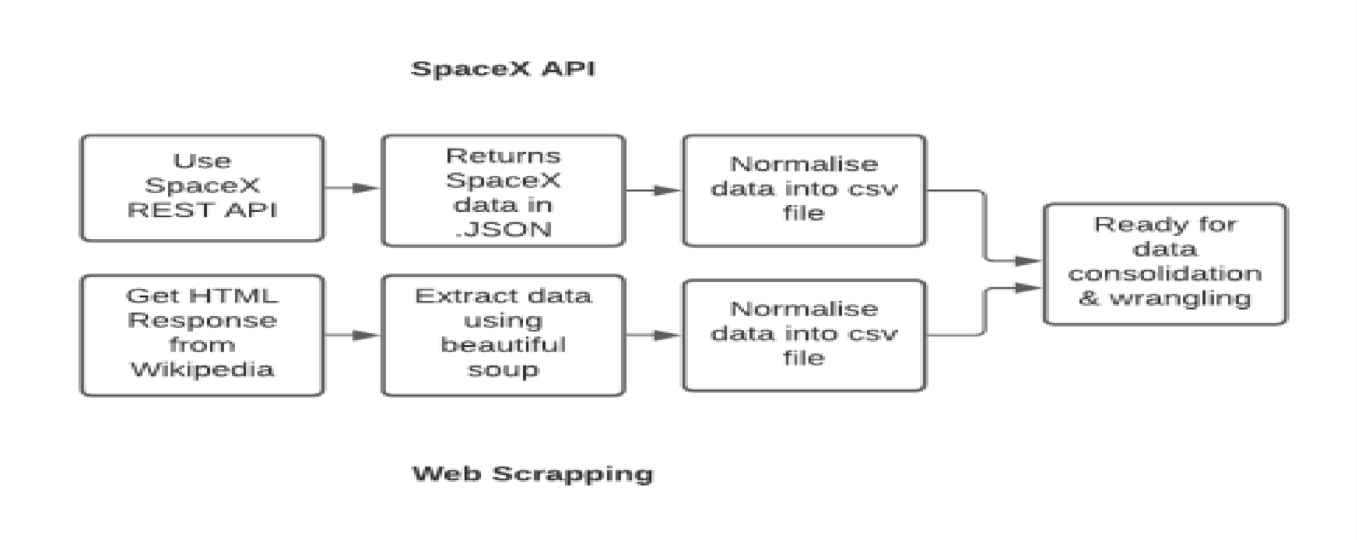
Methodology

Executive Summary

- Data collection methodology:
  - With Rest API and Web Scrapping
- Perform data wrangling
  - Both data transformation and one hot encoder were used before later applying machine learning models
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Seaborn scatter plots visualization techniques were used for the EDA
- Perform interactive visual analytics using Folium and Plotly Dash
  - Dash and Folium were used to achieve this goal
- Perform predictive analysis using classification models
  - Machine learning models such as LR, KNN, SVM and DTree were built to achieve these goals

Data Collection

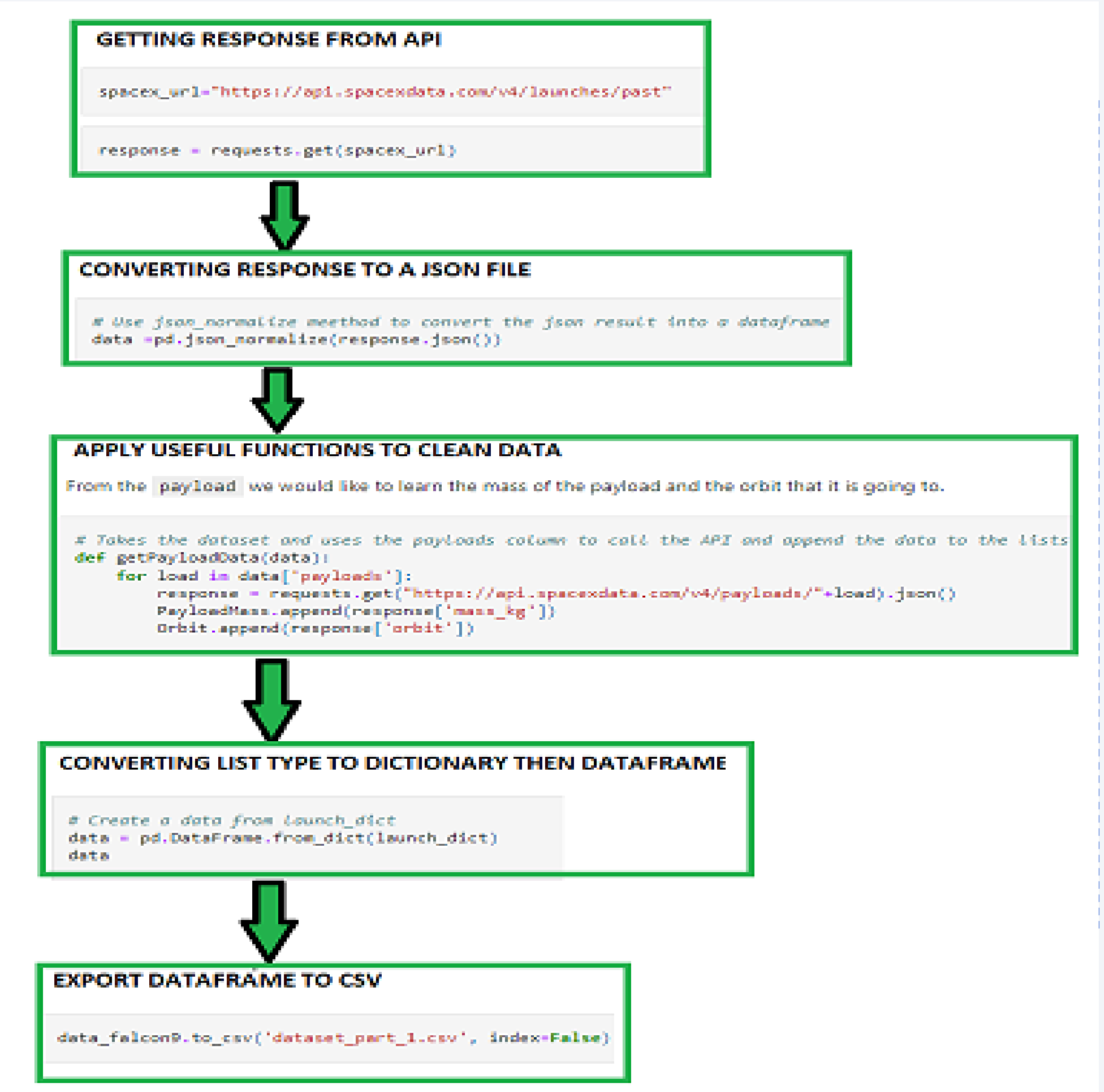
- Describe how data sets were collected.
- Data sets were collected using the API call from several websites, I collected rocket, launchpad, payloads, and cores data from https://api.spacexdata.com/v4 website.
- You need to present your data collection process use key phrases and flowcharts



Data Collection – SpaceX API

GitHub URL:

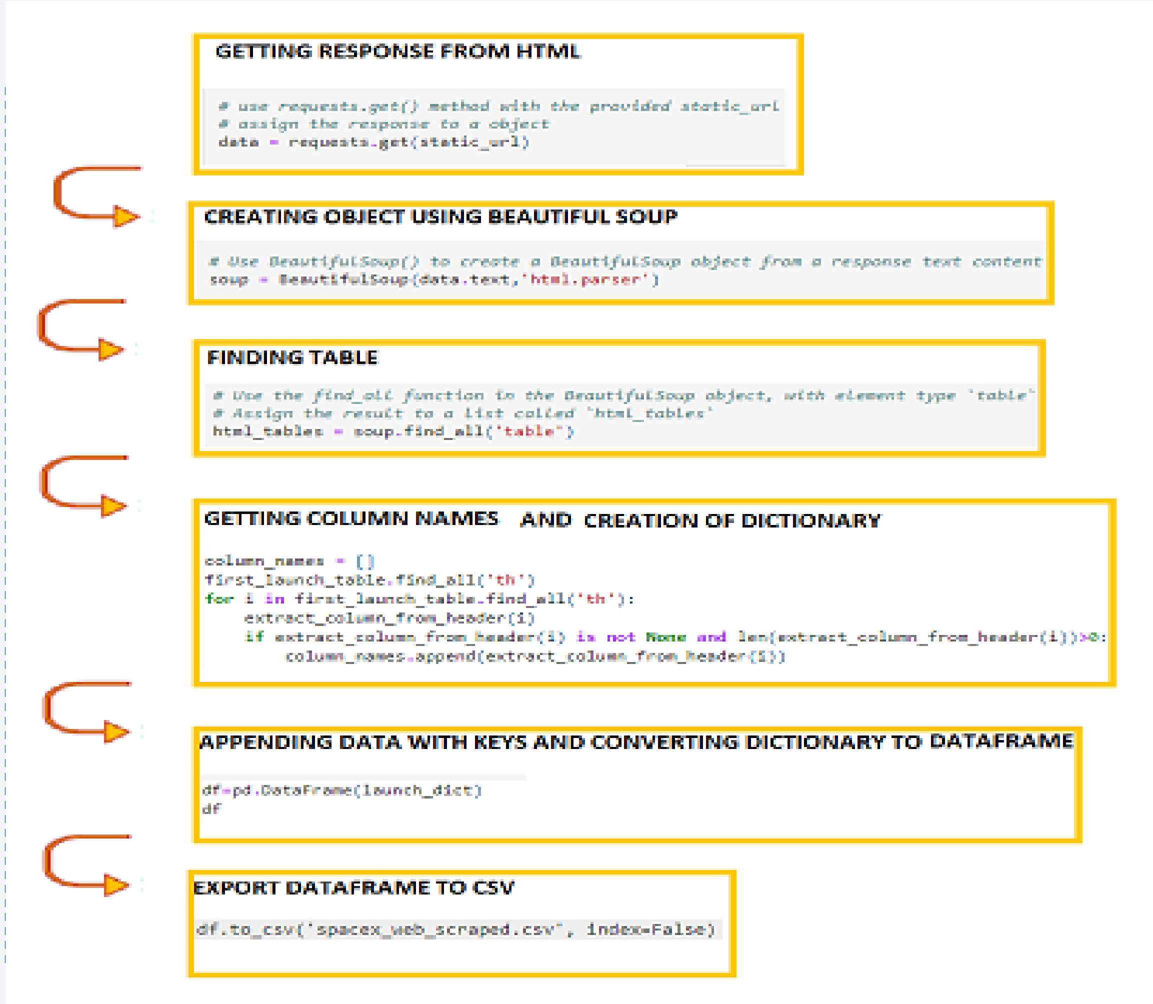
https://github.com/ifeanyiokpala/Applied-Data-Science-Capstone/blob/main/Space%20Y%20Capstone%20Project%20-%20WEEK%201a.ipynb



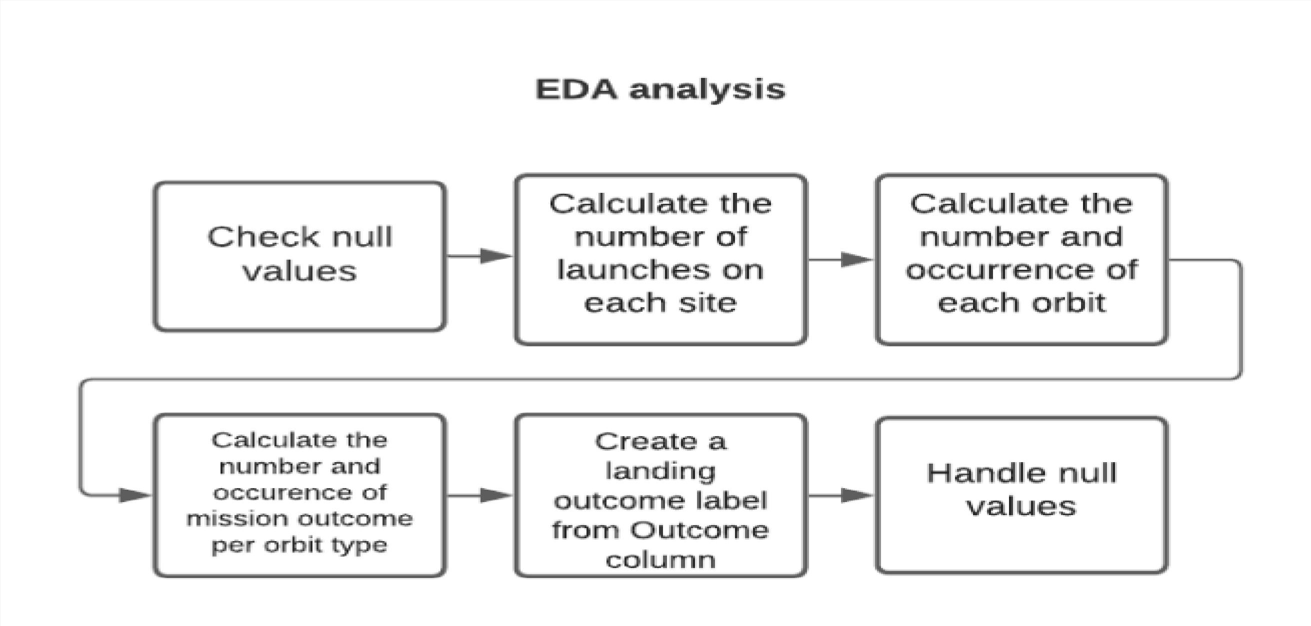
Data Collection - Scraping

GitHub URL:

https://github.com/ifeanyiokpala/Applied-Data-Science-Capstone/blob/main/Space%20Y%20Capstone%20Project%20-%20WEEK%201b.ipynb

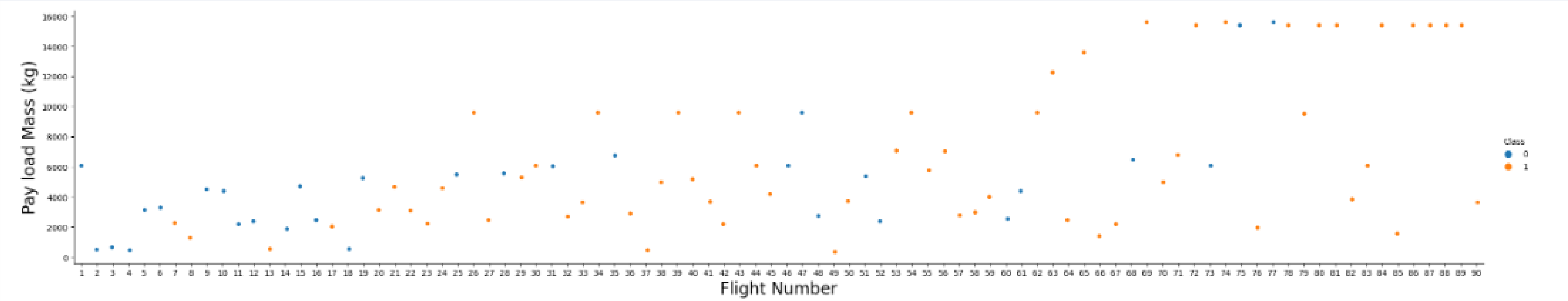


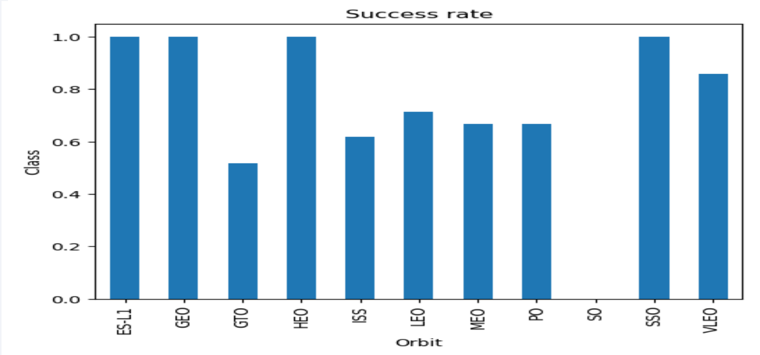
Data Wrangling

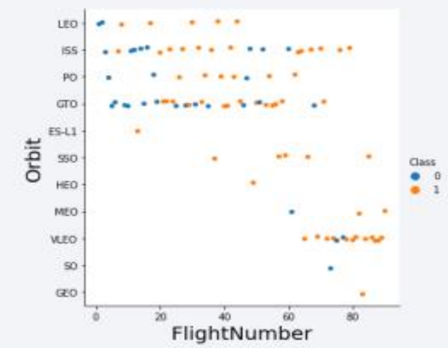


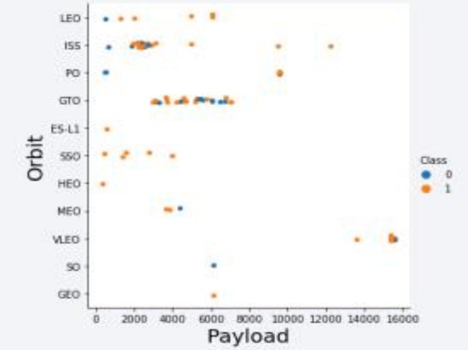
GitHub URL: https://github.com/ifeanyiokpala/Applied-Data-Science-Capstone/blob/main/Space%20Y%20Capstone%20Project%20-%20WEEK%201c.ipynb

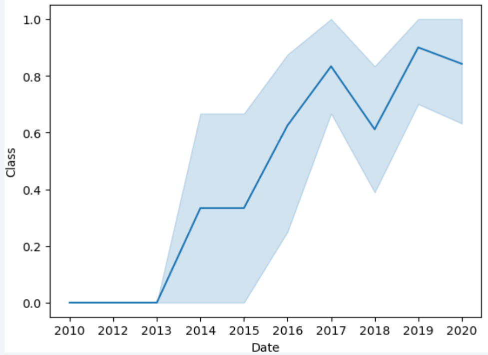
EDA with Data Visualization











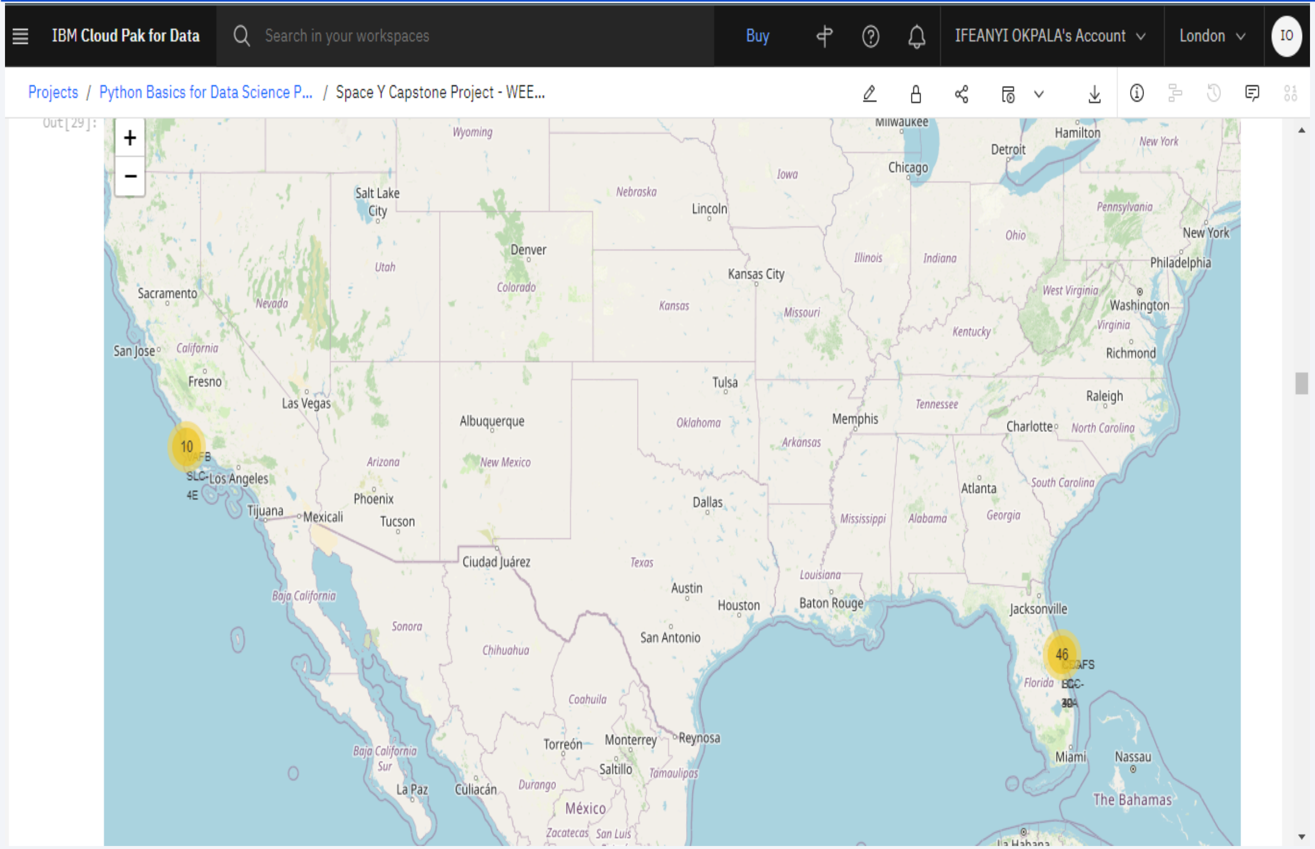
GitHub URL:
https://github.com/ifeanyiokpala/Applied-Data-Science-Capstone/blob/main/Space%20Y%20Capstone%20Project%20-%20WEEK%202b.ipynb

EDA with SQL

- 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for the 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

GitHub URL:
https://github.com/ifeanyiokpala/Applied-Data-Science-Capstone/blob/main/Space%20Y%20Capstone%20Project%20-%20WEEK%202a.ipynb

Build an Interactive Map with Folium

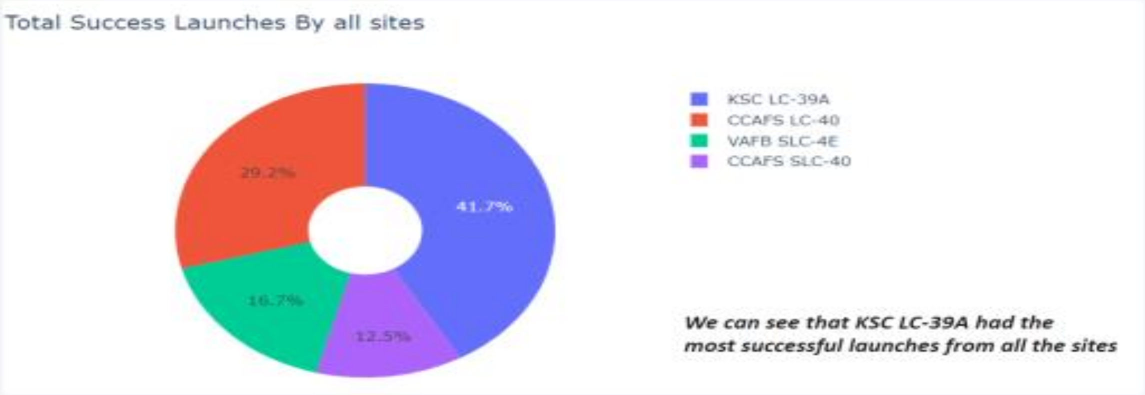


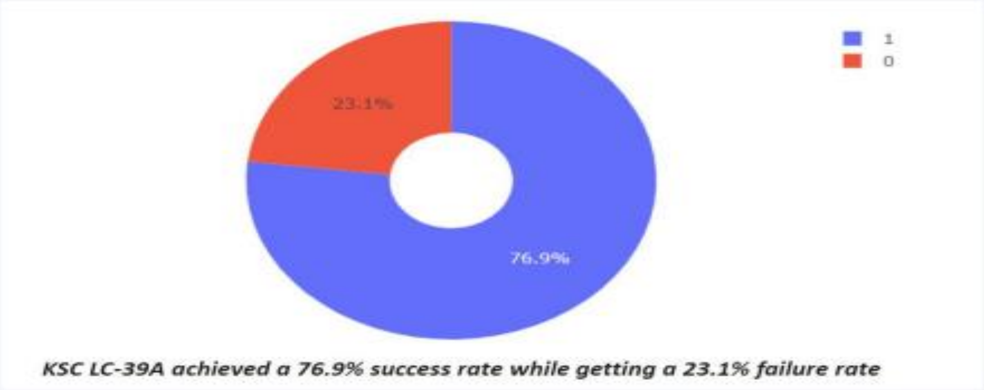
- folium.Marker() was used to create marks on the maps.
- folium.Circle() was used to create a circles above markers on the map.
- folium.Icon() was used to create an icon on the map.
- folium.PolyLine() was used to create polynomial line between the points.
- folium.plugins.AntPath() was used to create animated line between the points.
- markerCluster() was used to simplify the maps which contain several markers with identical coordination.

GitHub URL:
https://github.com/ifeanyiokpala/Applied-Data-Science-Capstone/blob/main/Space%20Y%20Capstone%20Project%20-%20WEEK%203.ipynb

Build a Dashboard with Plotly Dash

- Dash and html components were used as they are the most important thing and almost everything depends on them, such as graphs, tables, dropdowns, etc.
- Pandas was used to simplifying the work by creating dataframe.
- Plotly was used to plot the graphs.
- Pie chart and scatter chart were used to for plotting purposes.
- Rangeslider was used for payload mass range selection.
- Dropdown was used for launch sites.

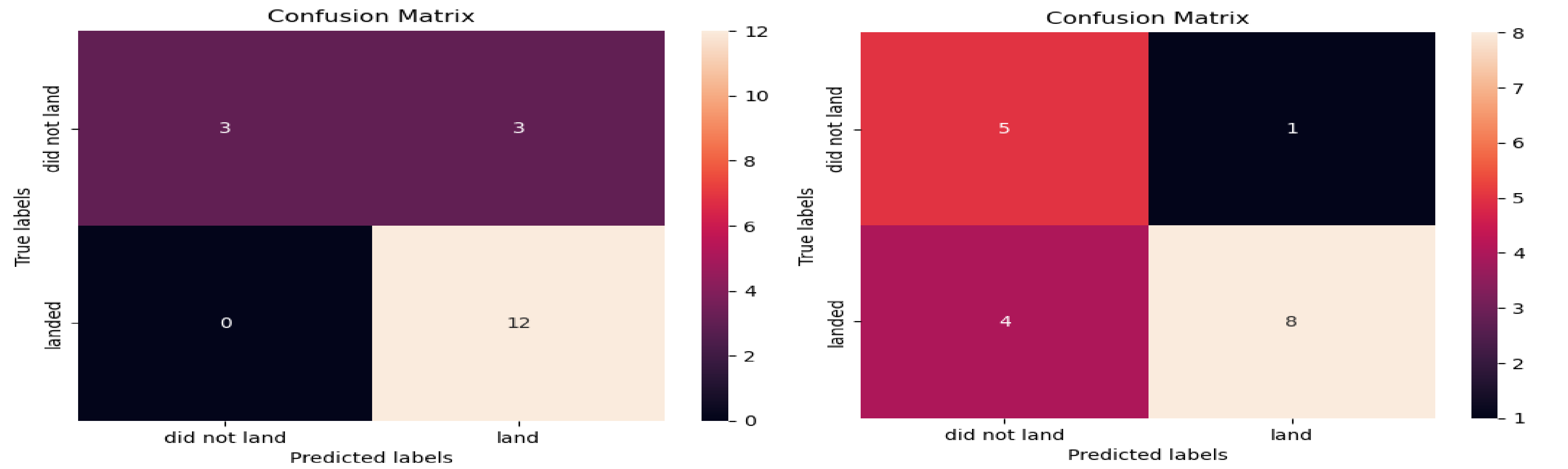




GitHub URL:
https://github.com/ifeanyiokpala/Applied-Data-Science-Capstone/blob/main/Space%20Y%20Capstone%20Project%20dashboard.ipynb

Predictive Analysis (Classification)

KNN, SVM and Logistic Regression tend to perform best with a score accuracy of 83.33%, while Classification Trees had an accuracy score of 72.22%



GitHub URL:
https://github.com/ifeanyiokpala/Applied-Data-Science-Capstone/blob/main/Space%20Y%20Capstone%20Project%20-%20WEEK%204.ipynb

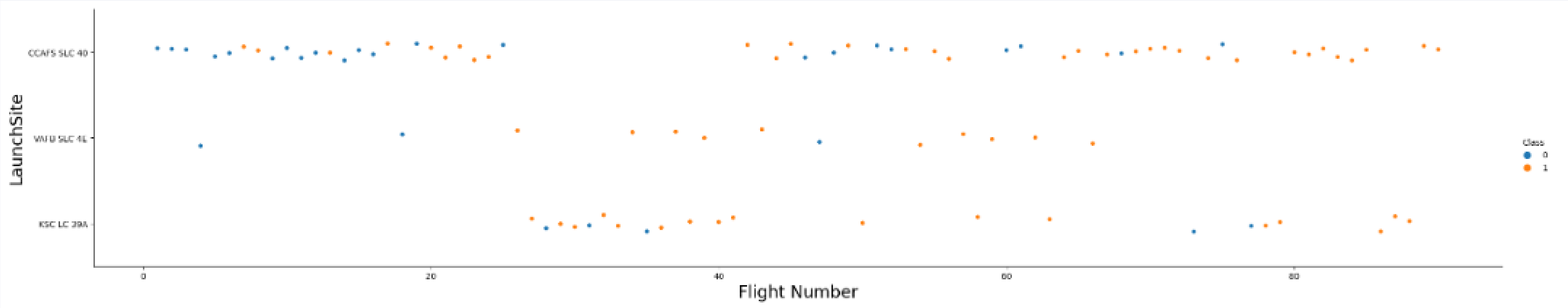
Results

- **The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy.**
- **Low weighted payloads perform better than the heavier payloads.**
- **Orbit GEO,HEO,SSO,ES L1 has the best Success Rate.**
- **KSC LC 39A had the most successful launches from all the sites.**

Section 2

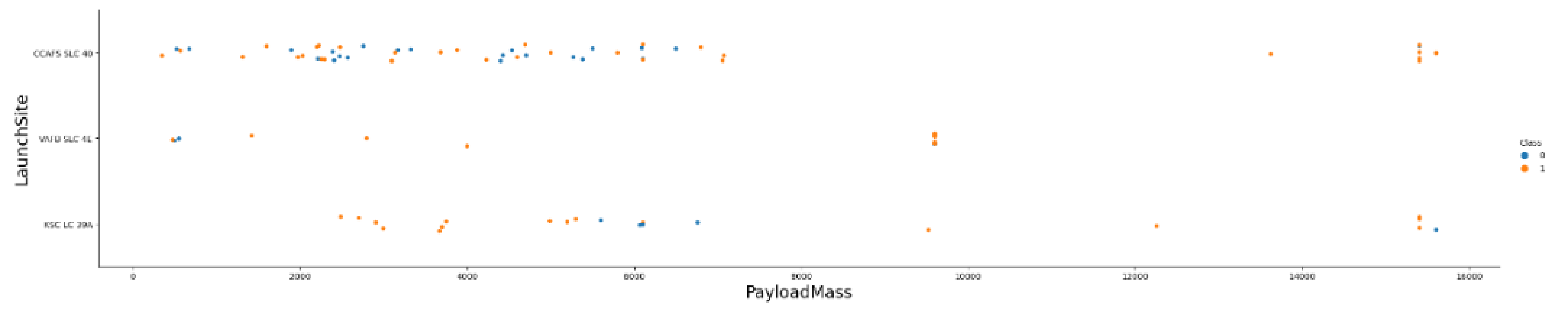
Insights drawn from EDA

Flight Number vs. Launch Site



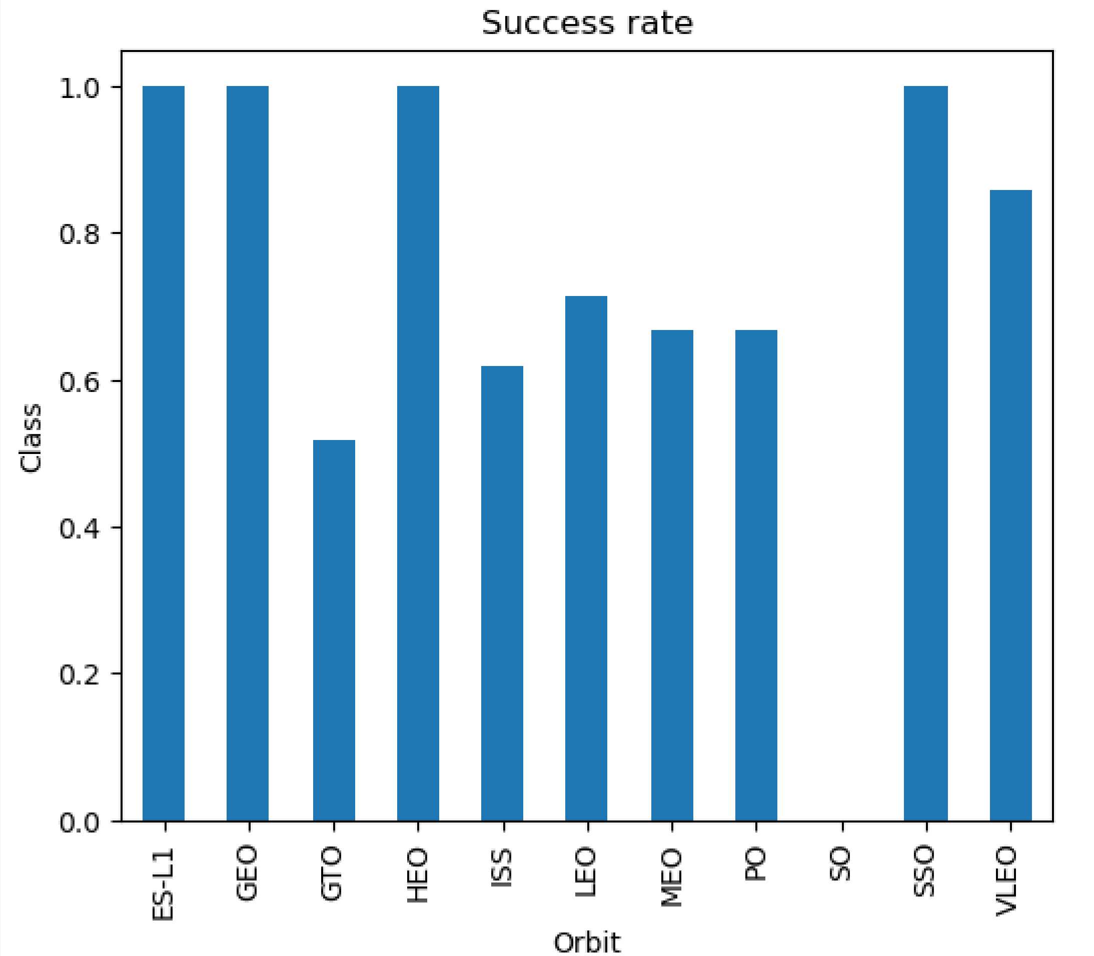
The success percentage at the launch sites is rising along with the number of flights.

Payload vs. Launch Site



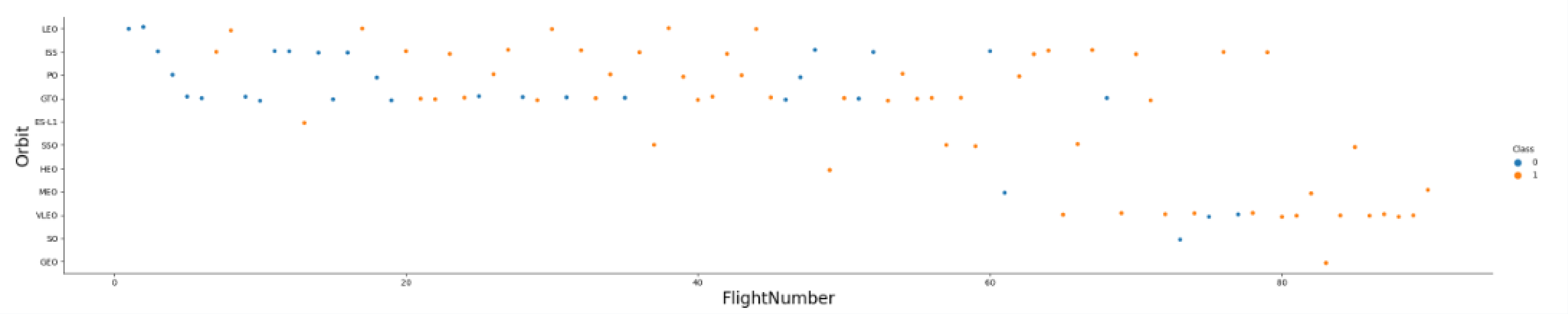
The success rate at launch locations is rising along with the increase in payload mass.

Success Rate vs. Orbit Type



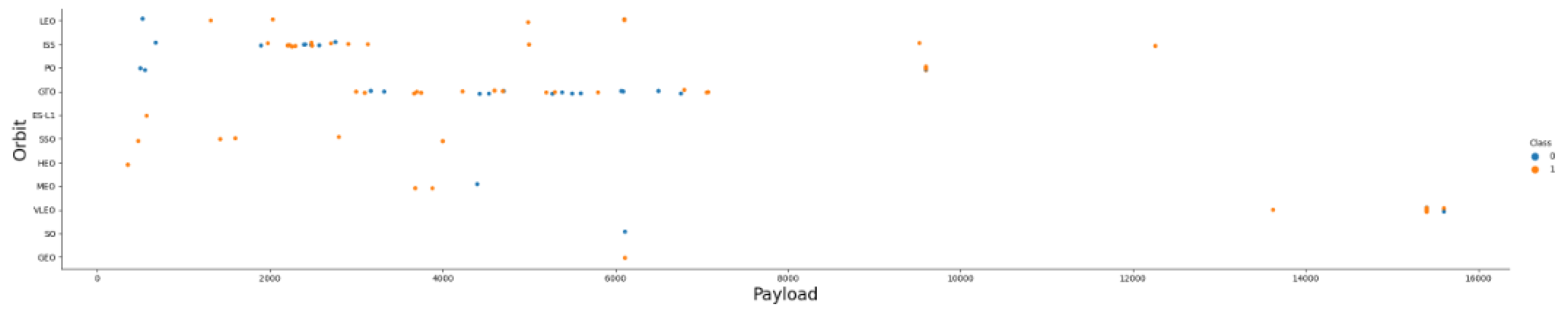
ES-L1, GEO, HEO, and SSO have a success rate of 100% while SO has a success rate of 0%

Flight Number vs. Orbit Type



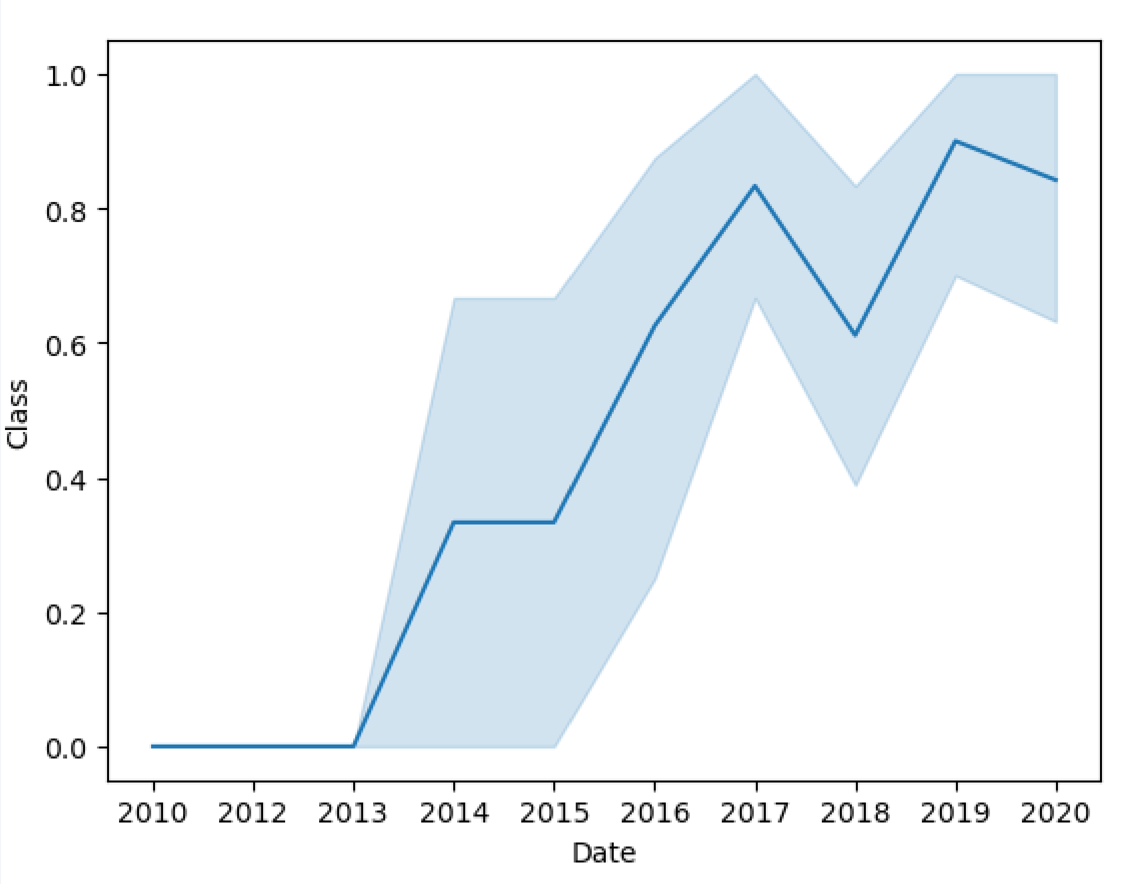
Although it is difficult to tell anything here, we can state that there is no real connection between flight number and GTO.

Payload vs. Orbit Type



The first item to check is how the ISS is affected by payload masses between 2000 and 3000.The GTO is also impacted by Pay load Mass between 3000 and 7000.

Launch Success Yearly Trend



The success rate has significantly increased since 2013. It did, however, experience a slight decline in 2018, but later recovered.

All Launch Site Names

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

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

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

We can get the unique values by selecting DISTINCT values

Launch Site Names Begin with '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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Launch site that begins with CCA can be limited to 5 by using limit 5

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

We can get the sum of all values by using SUM()

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE Booster_Version like 'F9 v1.1%'
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

We can get the average of all values by using AVG()

First Successful Ground Landing Date

```
%sql select min(Date) from SPACEXTBL where Mission_Outcome like 'Success%'
```

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

min(Date)
01-03-2013

We can get the minimum of all values by using MIN()

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTBL where Mission_outcome = 'Success' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 v1.1	4535
F9 v1.1 B1011	4428
F9 v1.1 B1014	4159
F9 v1.1 B1016	4707
F9 FT B1020	5271
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1030	5600
F9 FT B1021.2	5300
F9 FT B1032.1	5300
F9 B4 B1040.1	4990
F9 FT B1031.2	5200
F9 FT B1032.2	4230
F9 B4 B1040.2	5384
F9 B5 B1046.2	5800
F9 B5 B1047.2	5300
F9 B5 B1046.3	4000
F9 B5 B1048.3	4850
F9 B5 B1051.2	4200
F9 B5B1060.1	4311
F9 B5 B1058.2	5500
F9 B5B1062.1	4311

The payload mass data was taken between 4000 and 6000 only, and the landing outcome was determined to be successful.

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(Mission_Outcome) as 'Successful Mission'from SPACEXTBL where Mission_Outcome like 'Success%'
```

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

Successful Mission
100

```
%sql select count(Mission_Outcome) as 'Failed Mission'from SPACEXTBL where Mission_Outcome like 'Fail%'
```

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

Failed Mission
1

We can get the number of all the successful and failed mission by using COUNT and LIKE

Boosters Carried Maximum Payload

```
%sql select booster_version,PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ =(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

We can get the maximum payload masses by using MAX()

2015 Launch Records

```
%sql select date, substr(Date,4,2) as Month,Booster_Version,Launch_Site,Mission_Outcome from SPACEXTBL where Mission_Outcome like 'succ%' and substr(D
```

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

Date	Month	Booster_Version	Launch_Site	Mission_Outcome
10-01-2015	01	F9 v1.1 B1012	CCAFS LC-40	Success
11-02-2015	02	F9 v1.1 B1013	CCAFS LC-40	Success
02-03-2015	03	F9 v1.1 B1014	CCAFS LC-40	Success
14-04-2015	04	F9 v1.1 B1015	CCAFS LC-40	Success
27-04-2015	04	F9 v1.1 B1016	CCAFS LC-40	Success
22-12-2015	12	F9 FT B1019	CCAFS LC-40	Success

We can get the months by using substr() and in the WHERE function we assigned the year value is 2015 and mission outcome is success

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

```
%sql select date,Mission_Outcome from SPACEXTBL where Mission_Outcome like 'succ%' and date between '04-06-2010' and '20-03-2017' order by date desc
```

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

Date	Mission_Outcome
19-02-2017	Success
19-01-2020	Success
18-10-2020	Success
18-08-2020	Success
18-07-2016	Success
18-04-2018	Success
18-04-2014	Success
18-03-2020	Success
17-12-2019	Success
17-02-2020	Success
17-01-2016	Success
16-11-2020	Success
16-03-2017	Success
15-12-2017	Success
15-11-2018	Success
15-06-2016	Success
15-05-2017	Success
14-08-2017	Success
14-08-2016	Success
14-07-2014	Success
14-04-2015	Success
14-01-2017	Success
13-06-2020	Success
12-06-2019	Success
11-11-2019	Success
11-10-2017	Success
11-05-2018	Success
11-02-2015	Success
11-01-2019	Success
10-09-2018	Success
10-01-2015	Success
09-10-2017	Success
08-12-2010	Success
08-10-2018	Success
08-10-2012	Success
08-04-2016	Success
08-01-2018	Success (payload status unclear)
07-09-2017	Success
07-09-2014	Success
07-08-2020	Success
07-08-2018	Success
07-03-2020	Success
07-01-2020	Success
06-12-2020	Success
06-10-2020	Success
06-08-2019	Success
06-05-2016	Success
06-03-2018	Success
06-01-2014	Success
05-12-2019	Success
05-12-2018	Success
05-11-2020	Success
05-08-2014	Success
05-07-2017	Success
04-06-2020	Success
04-06-2018	Success
04-06-2010	Success

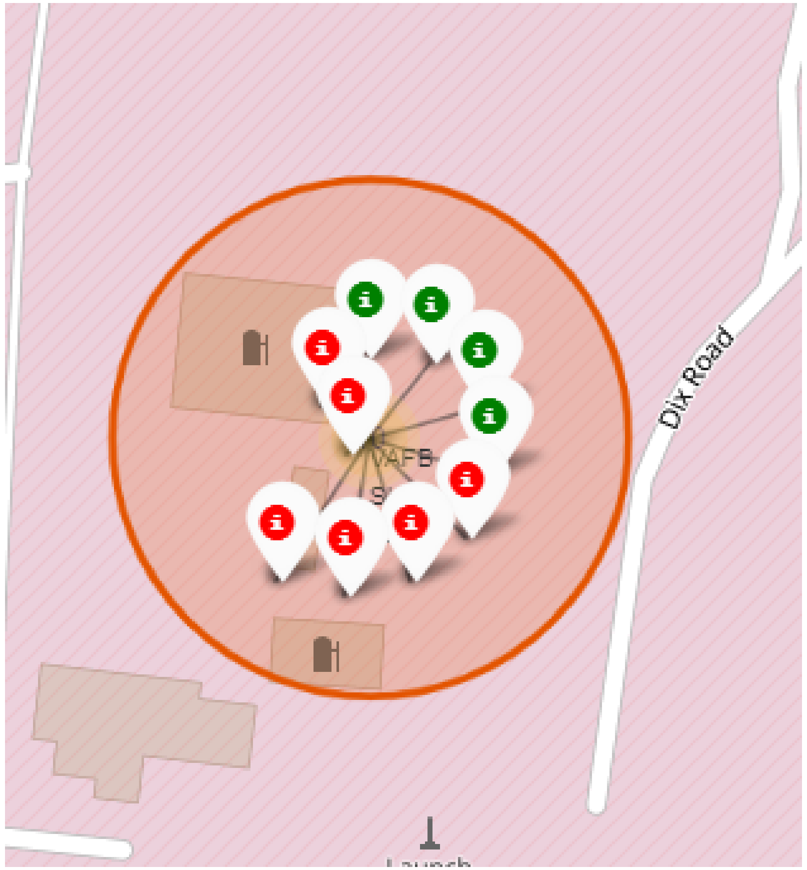
We can rank landing outcome within a range by using the ORDER BY

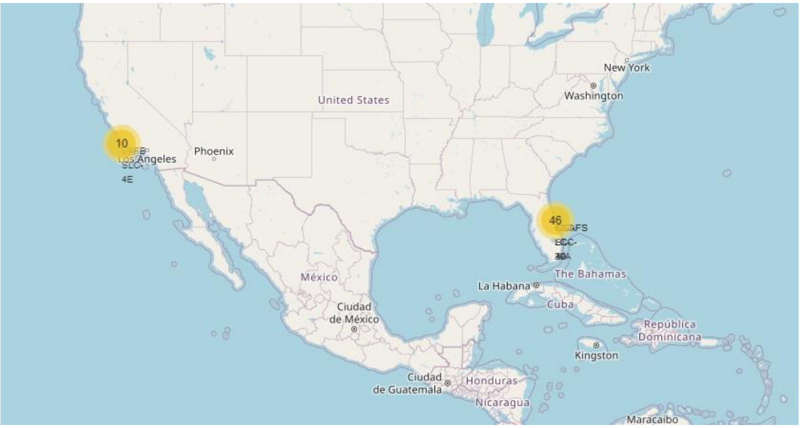
Section 3

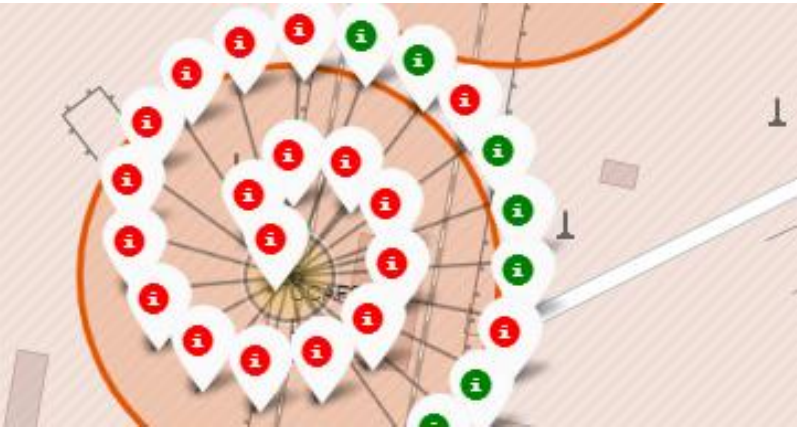
Launch Sites Proximities Analysis

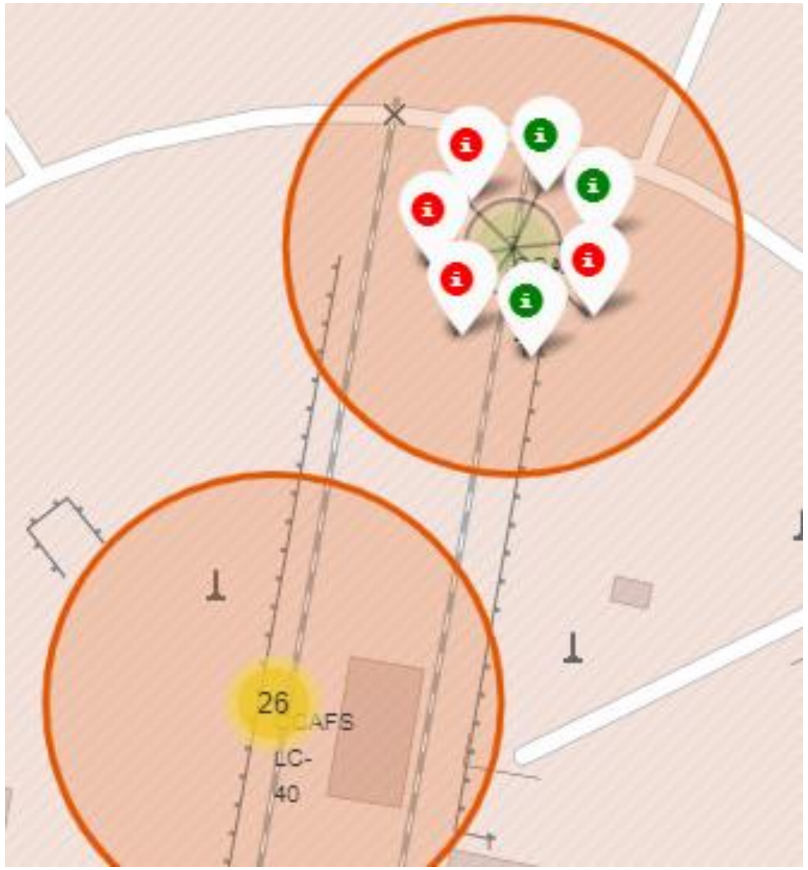
All Launch Sites’ Location Markers

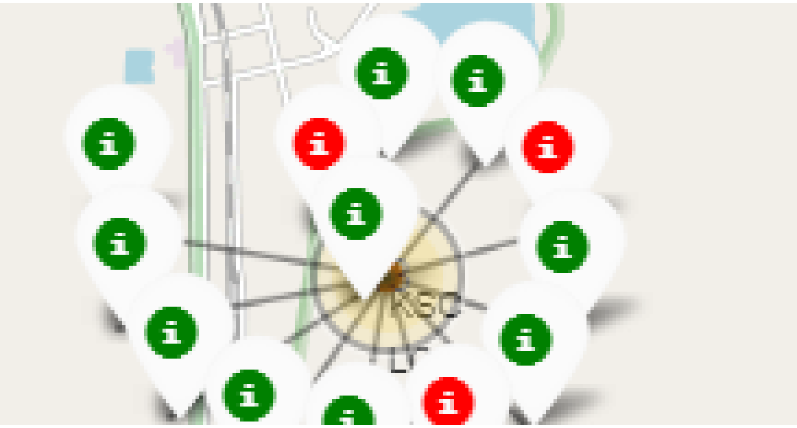
All the launches are near USA, Florida, and California









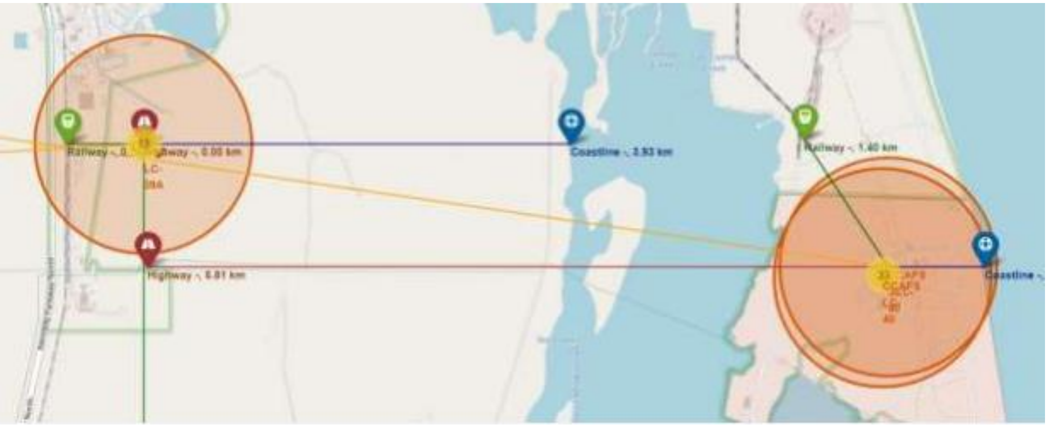


Color-labeled Launch Outcomes

(**Green means successful Red means Failure**)

Launch Sites to its Proximities

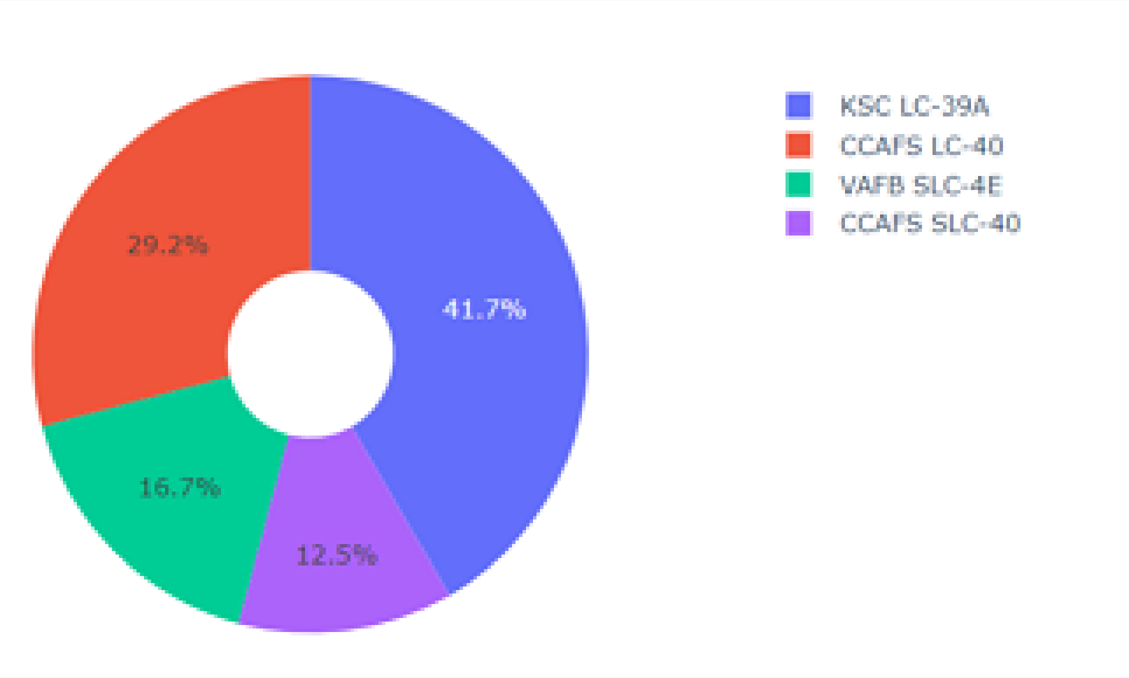
The distances from all the launch sites to its proximities were not far from railway tracks.



Section 4

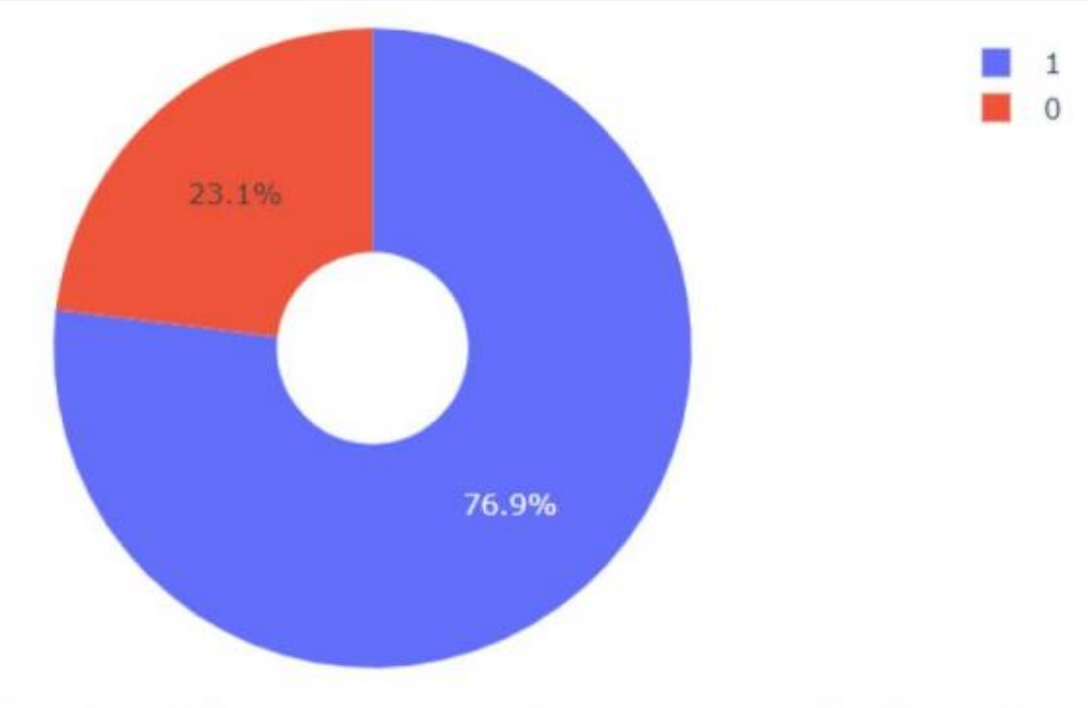
Build a Dashboard with Plotly Dash

Launch Success Count



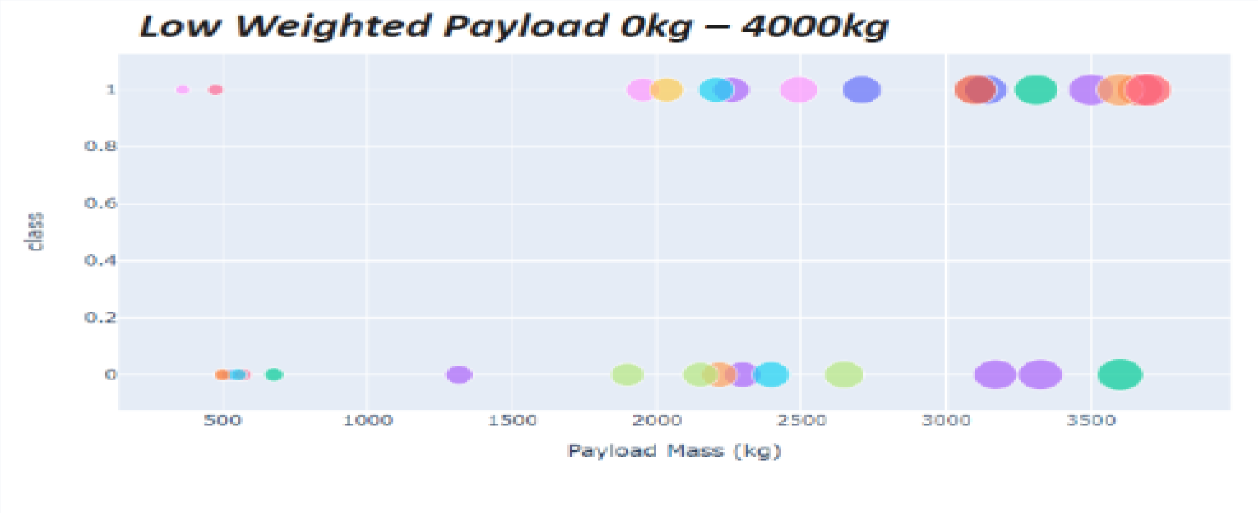
From the plot we can see that the most successful lunch is KSC LC-39A with 41.7%

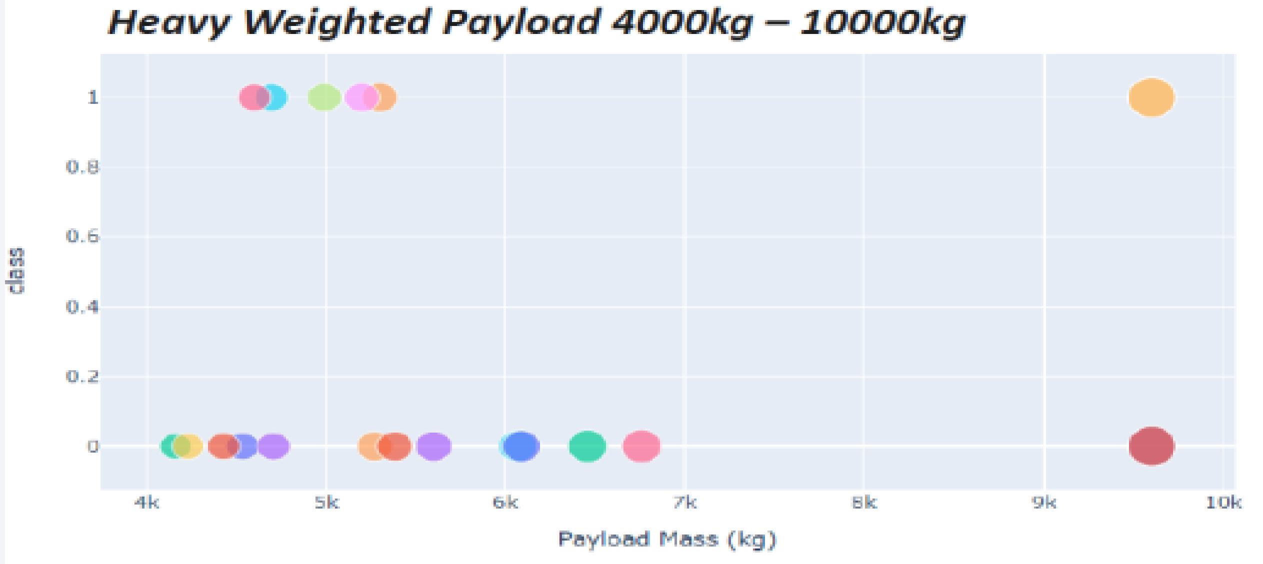
Launch Site with Highest Score



KSC LC-39A has the highest success score of 76.9% while unsuccessful score is 23.1%

Payload vs. Launch Outcome



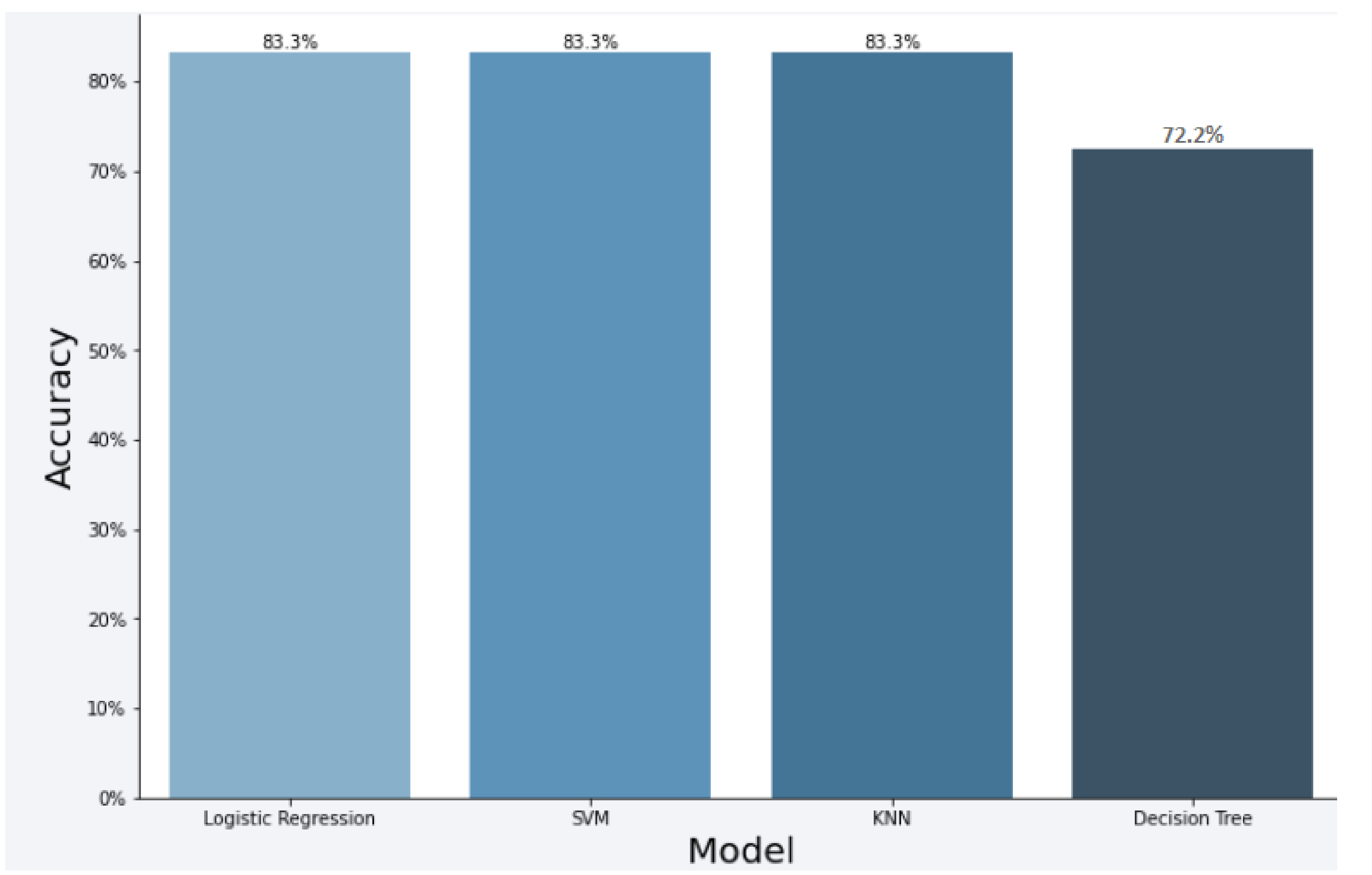


There is a higher success rate for the low weighted payload of 0kg – 4000kg than the higher payload.

Section 5

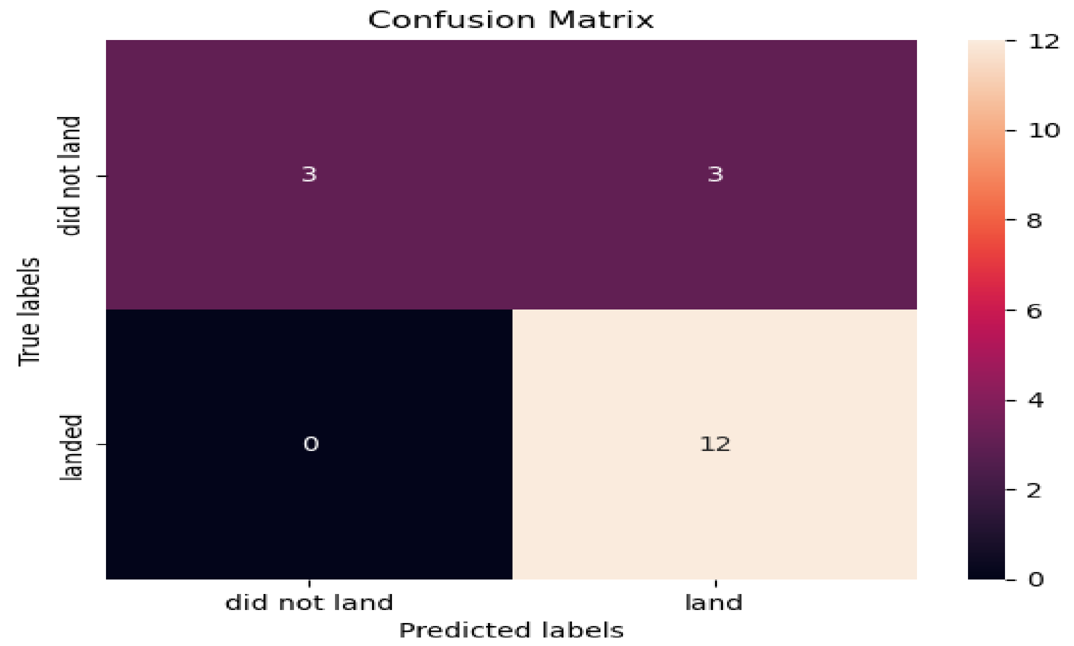
Predictive Analysis (Classification)

Classification Accuracy

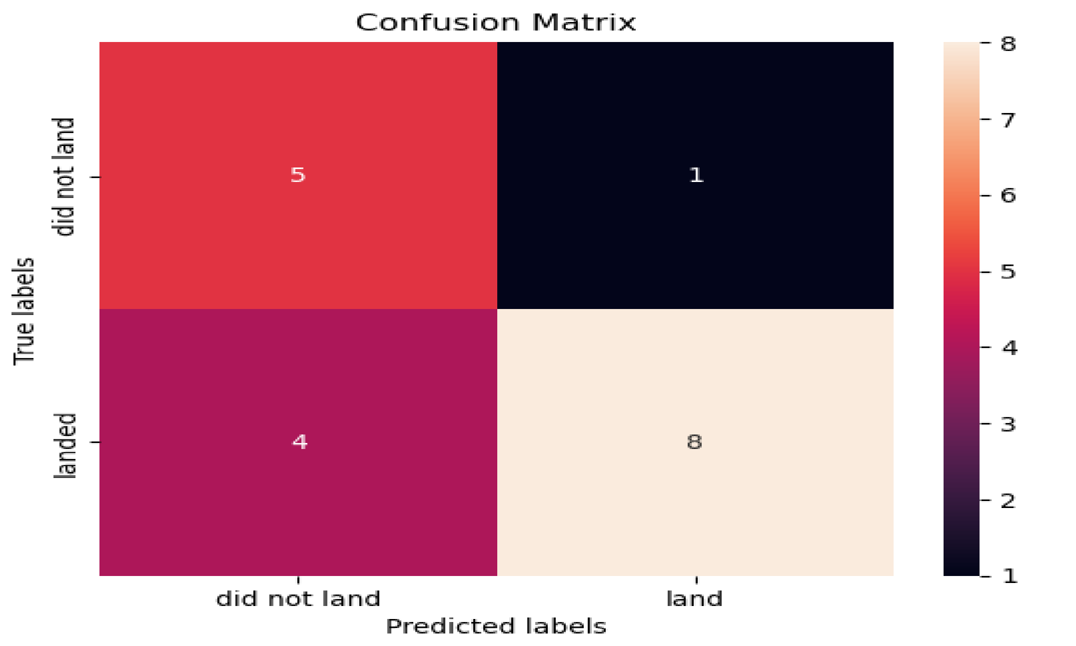


Decision Tree has the lowest score accuracy of 72.2%, with the remaining models with the same accuracy of 83.3%

Confusion Matrix



KNN, SVM and Logistic Regression



Decision Tree

From the Confusion Matrix we can see a better accuracy in prediction for the KNN, SVM and Logistic Regression with just 3 wrong prediction False Positive (FP), although for the Decision Tree we can see 1 FP and 4 FN which is less accurate.

Conclusions

- SVM, KNN, and Logistic Regression had the best score accuracy.
- The payload of 0 kg to 5000 kg was more diverse than 6000 kg to 10000 kg
- KSC LC 39A had the most successful launches from all the sites.
- We calculated the launch sites distance to its proximities

Appendix

GitHub URL:
https://github.com/ifeanyiokpala/Applied-Data-Science-Capstone

Thank you!