Data Science Final Project

SUJAL SHARMA
18-02-2024

Outline

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

Executive Summary

- Summary of methodologies
  - Collected data from public SpaceX API and SpaceX Wikipedia page.
  - Created labels column ‘class’ which classifies successful landings.
  - Explored data using SQL, visualization, folium maps, and dashboards. Gathered relevant columns to be used as features. Changed all categorical variables to binary using one hot encoding.
  - Standardized data and used GridSearchCV to find best parameters for machine learning models.
  - Visualize accuracy score of all models.
- Summary of all results

Four machine learning models were produced all produced similar results with accuracy rate of about 83.33%.

Introduction

Project background and context

- Human have already advanced into space but now it's time to commercialize it.
- Space X has best pricing ($62 million vs. $165 million USD).Largely due to ability to recover part of rocket (Stage 1).
- Space Y wants to compete with Space X.

Problems we want finding answers for

- Space Y wants us to train a machine learning model to predict successful Stage 1 recover

Section 1

Methodology

Methodology

- DATA COLLECTION :
  - It was taken from Space X API and Wikipedia page
- DATA WRANGLING:
  - Done to achieve classification between successful and unsuccessful landings.
- EDA :
  - Did EDA using SQL and Visualization
- DASHBOARD :
  - Created a dashboard using Plotly.
- PREDICTIVE ANALYSIS :
  - Used GridSearchCV

Data Collection

- Data collection process involves the combination of both the data taken from the API and the web scrapped data from the Wikipedia page of Space X.
- Space X API Data involved columns : Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude.
- Wikipedia Web scraped Data contains the following columns: Flight No., Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

Data Collection – SpaceX API

- Space X API was requested and in return we received a Json file the columns listed in the previous slides. The Json file was then normalized to a form of Data Frame, from which we were able to retrieve the dictionary based data which was casted to a data frame.
- The final data collected was of Falcon 9 launches and missing payload mass values which were filled with their mean value.
- Below is the GitHub link to the python notebook:

[Courcera-Capstone-Project/Data Science Capstone/1 Sujal Sharma DS Capstone (spacex-data-collection-api).ipynb at main · xxwizardxx117/Courcera-Capstone-Project (github.com)]

Data Collection - Scraping

- Space X Wikipedia page was requested and in return we received the complete html page. Using the help of the Beautiful Soup and html5lib parser we were able to extract the tables from the HTML file.
- The table data was extracted to a dictionary and then converted into Data frame.
- Below is the GitHub link to the python notebook:

Courcera-Capstone-Project/Data Science Capstone/2 Sujal Sharma DS Capstone (data-webscraping).ipynb at main · xxwizardxx117/Courcera-Capstone-Project (github.com)

Data Wrangling

- In Data Wrangling we create a training label with landing outcomes where successful = 1 & failure = 0. We pointed out Outcome column has two components: 'Mission Outcome' 'Landing Location'
- New training label column 'class' with a value of 1 if 'Mission Outcome' is True and 0 otherwise.
- Value Mapping:
  - True ASDS, True RTLS, & True Ocean were set to 1 and None None, False ASDS, None ASDS, False Ocean, False RTLS were set to 0

- Below is the GitHub link to the python notebook:-

  Courcera-Capstone-Project/Data Science Capstone/3 Sujal Sharma DS Capstone (spacex-Data wrangling).ipynb at main · xxwizardxx117/Courcera-Capstone-Project (github.com)

EDA with SQL

- Exploratory Data Analysis was done in the Jupyter Lite and SQL integration in python was used.
- The data was queried and information about launch site and their names, mission overall outcome, amount of payload of booster versions and outcome of various landing were extracted.
- Below is the GitHub link to the python notebook:- Courcera-Capstone-Project/Data Science Capstone/4 Sujal Sharma DS Capstone (eda-sql-coursera_sqllite).ipynb at main · xxwizardxx117/Courcera-Capstone-Project (github.com)

EDA with Data Visualization

- Exploratory Data Analysis was done in the Jupyter Lite and various plots were created using seaborn and Matplotlib.
- The various plots created are:-
    - Flight Number vs. Payload Mass
    - Flight Number vs. Launch Site
    - Payload Mass vs. Launch Site
    - Orbit vs. Success Rate
    - Flight Number vs. Orbit
    - Payload vs Orbit
    - Success Yearly Trend

Below is the GitHub link to the python notebook:-

Courcera-Capstone-Project/Data Science Capstone/5 Sujal Sharma DS Capstone (eda-dataviz).ipynb at main · xxwizardxx117/Courcera-Capstone-Project (github.com)

Build an Interactive Map with Folium

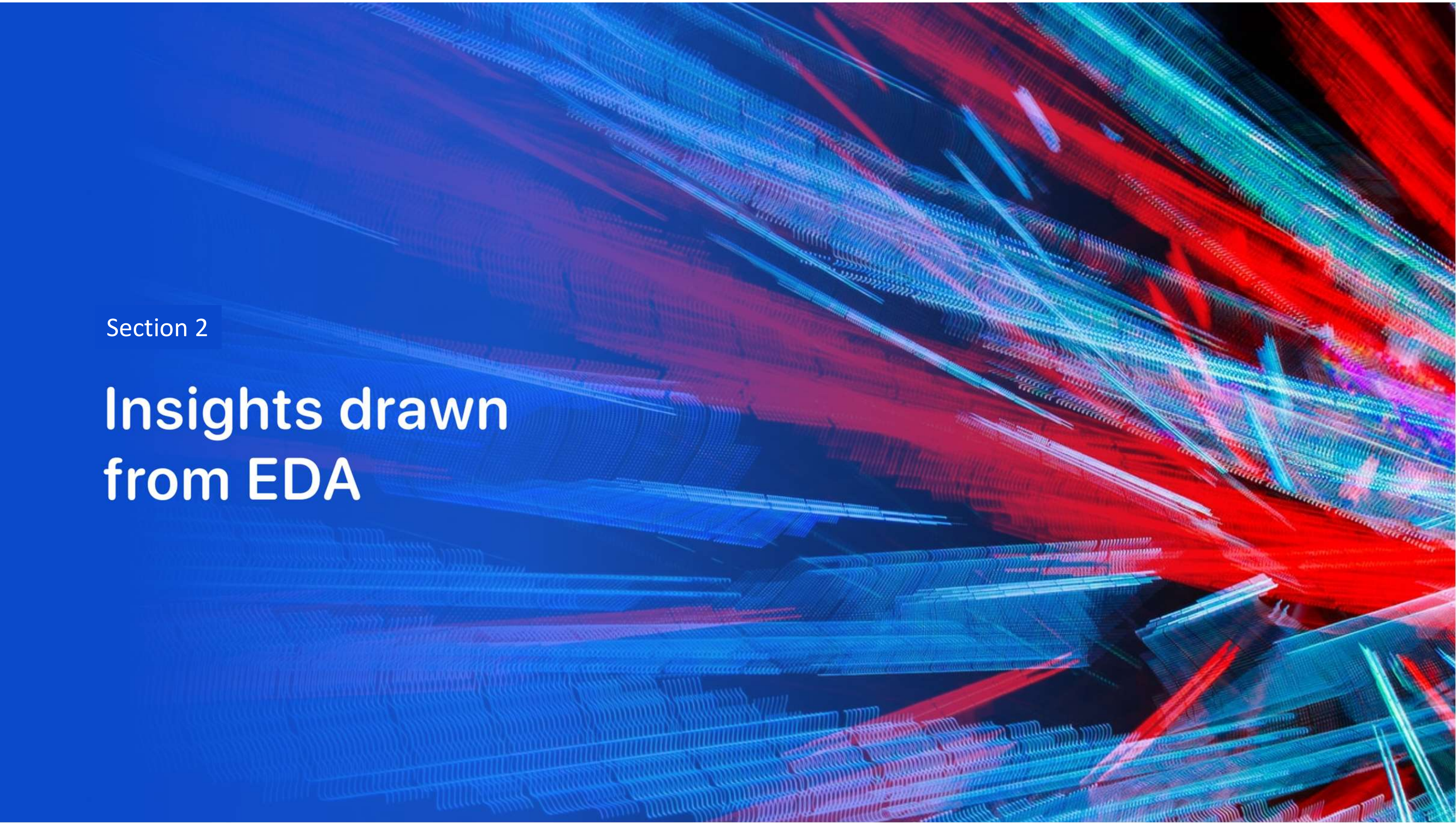
- We have created an interactive map which allows us to mark the Launch Sites and whether if the landing was successful or not .
- This map helps us to figure out what are the nearest cities, towns, highways, railway stations etc.
- Below is the GitHub link to the python notebook:- Courcera-Capstone-Project/Data Science Capstone/6 Sujal Sharma DS Capstopne (launch site map marking using Folium).ipynb at main · xxwizardxx117/Courcera-Capstone-Project (github.com)

Build a Dashboard with Plotly Dash

- We have created an dashboard which contains a pie chart and scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates. The pie chart is used to visualize launch site success rate.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg. The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.
- Below is the GitHub link to the python notebook

Predictive Analysis (Classification)

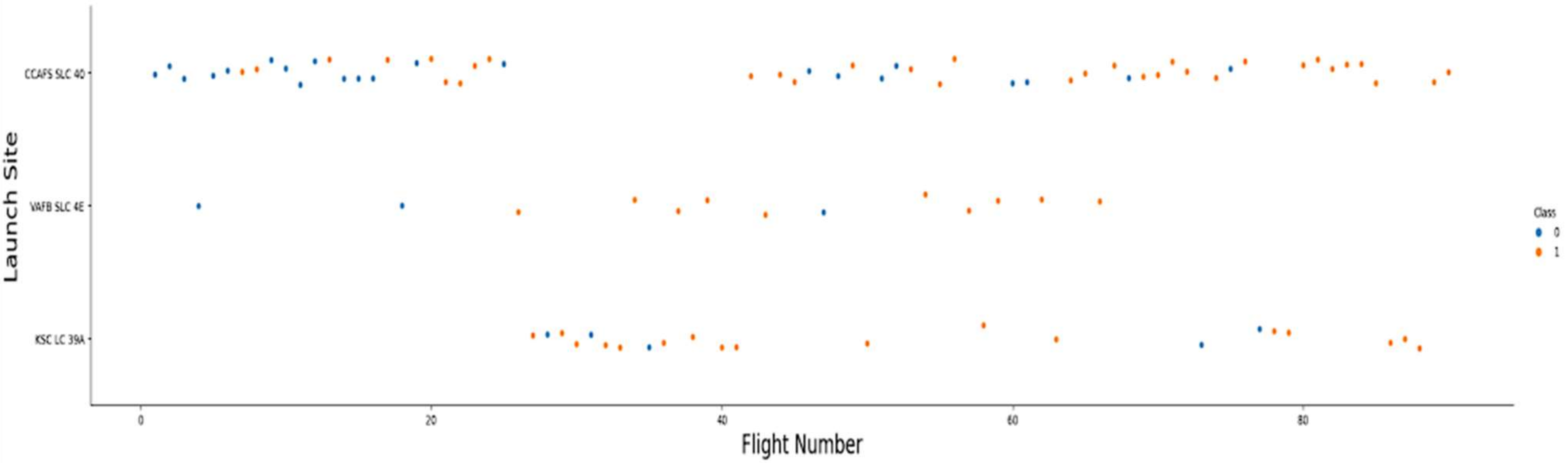
- We have used four Machine Learning Algorithms for this project.
- The 4 models used for are Logistic Regression, Support Vector Machine(SVM), Decision Tree, K Nearest Neighbors (KNN).
- We have used train test split and the size of the test data is 20% of the whole dataset. We have also used Standard Scaler to transform the data.
- We have use GridSearchCV with cv as IO to find the optimal parameters also created Confusion matrix and a Data Frame of the scores of each model.
- Below is the GitHub link to the python notebook:- Courcera-Capstone-Project/Data Science Capstone/8 Sujal Sharma DS Capstone (SpaceX_Machine_Learning_Prediction).ipynb at main · xxwizardxx117/Courcera-Capstone-Project (github.com)



Section 2

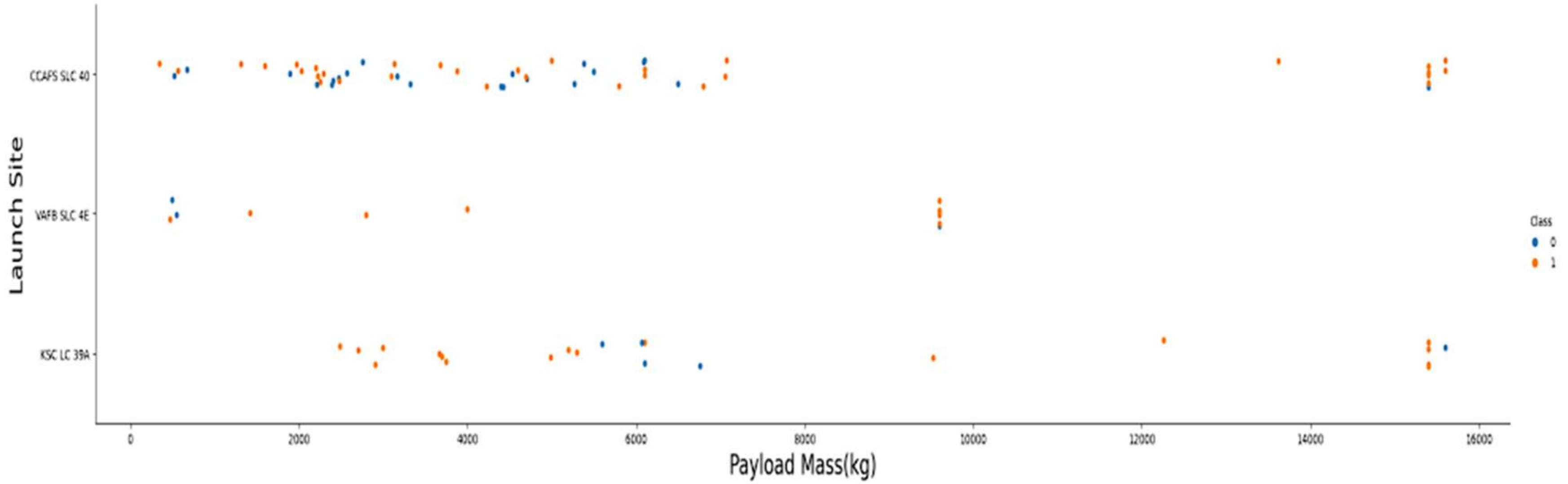
Insights drawn from EDA

Flight Number vs. Launch Site



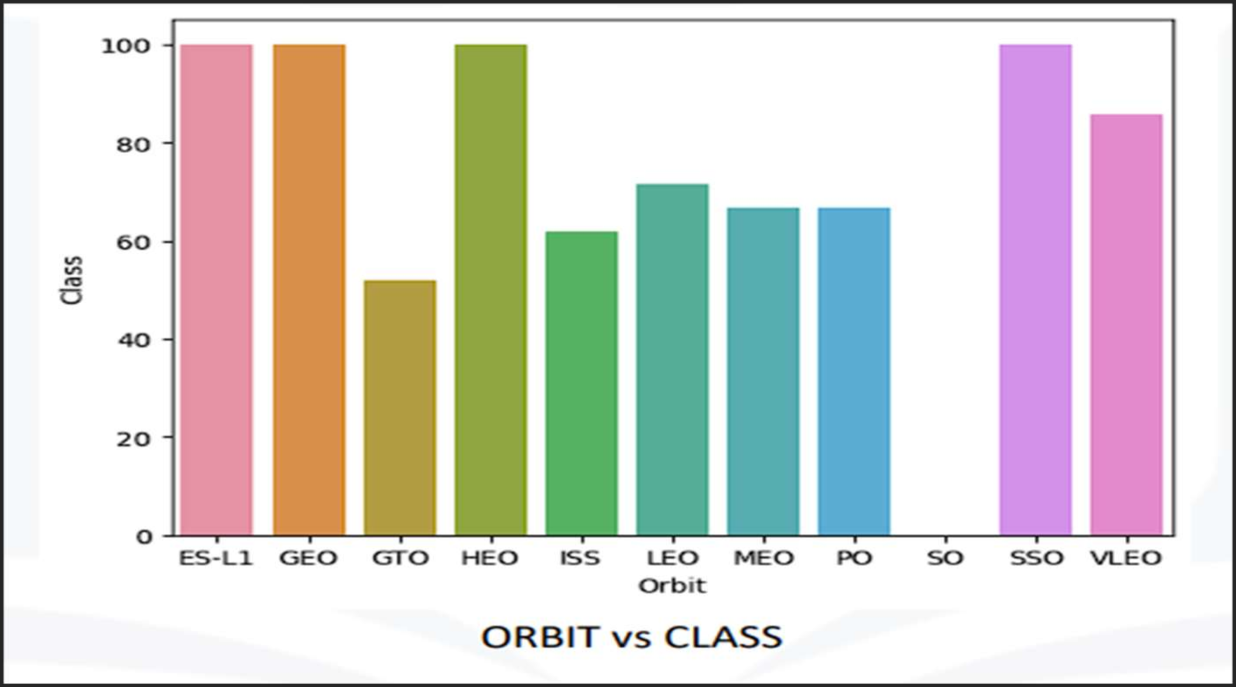
FLIGHT NUMBER vs LAUNCH SITE

Payload vs. Launch Site

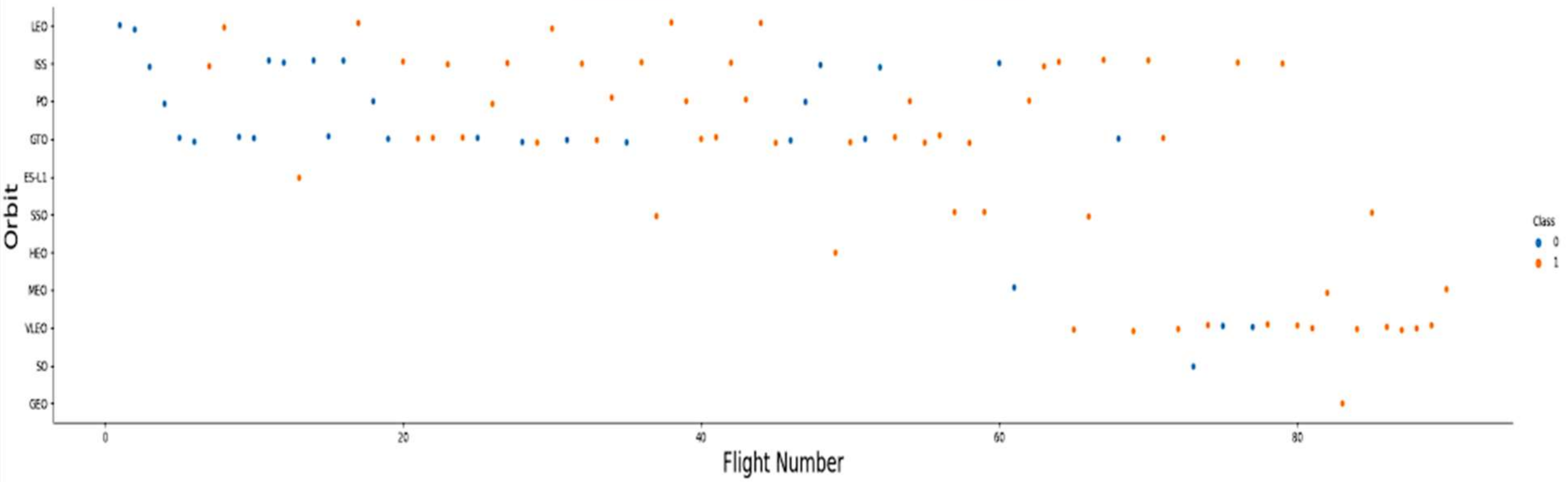


PAYLOAD MASS vs LAUNCH SITE

Success Rate vs. Orbit Type

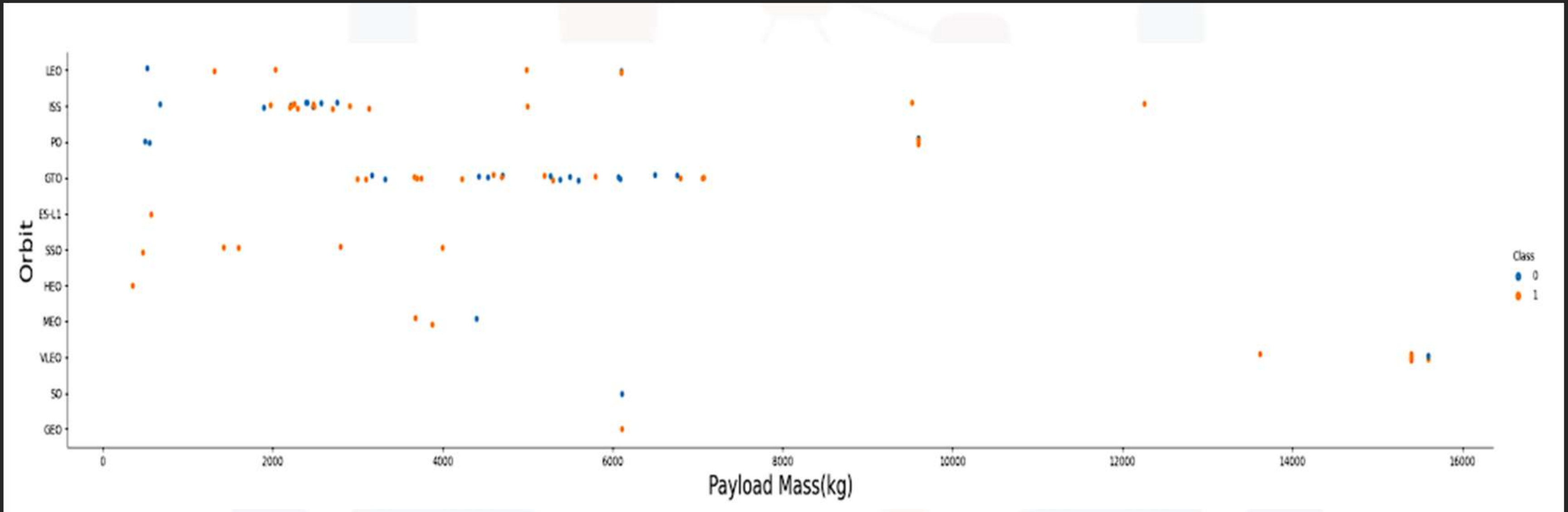


Flight Number vs. Orbit Type



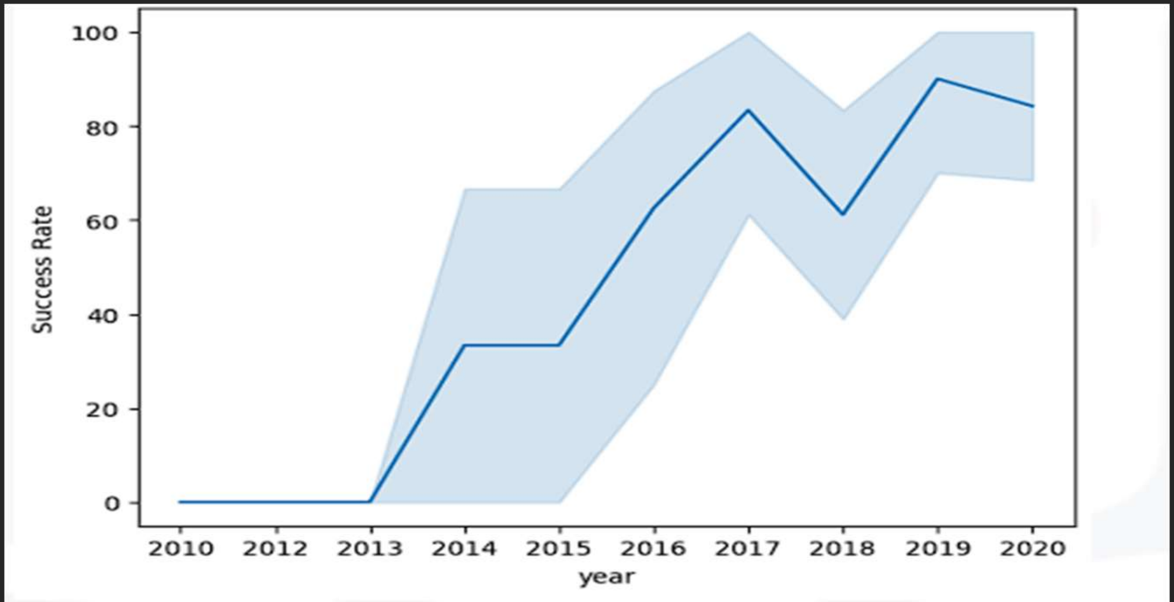
FLIGHT NUMBER vs ORBIT

Payload vs. Orbit Type



PAYLOAD MASS vs ORBIT

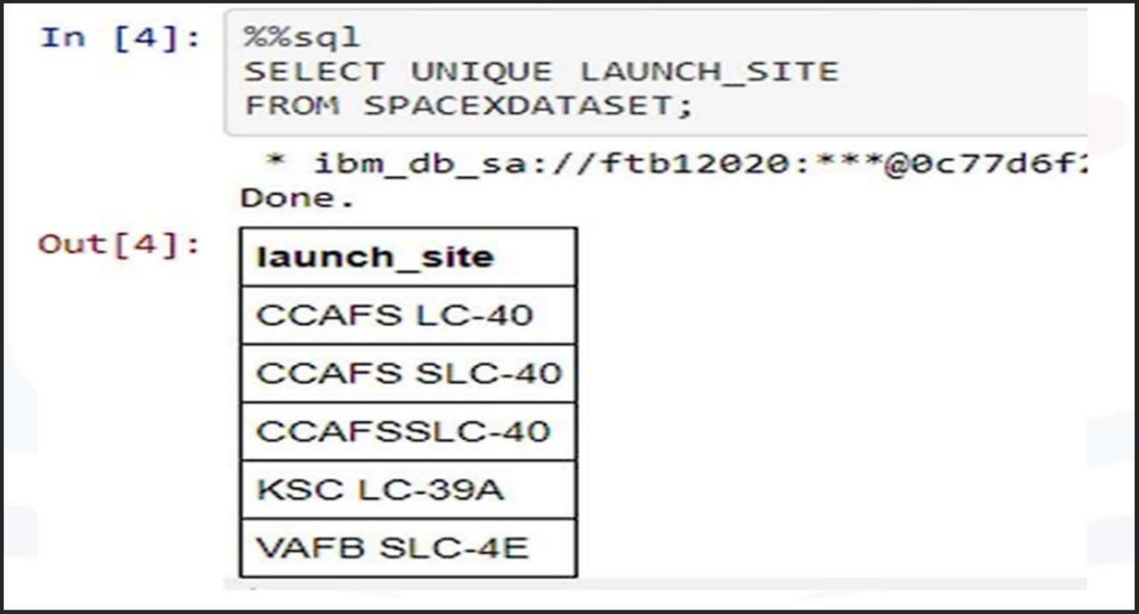
Launch Success Yearly Trend



YEAR vs SUCCESS RATE

All Launch Site Names

- Task 1: Display the names of the unique launch sites in the space mission



In [4]:
```
%%sql
SELECT UNIQUE LAUNCH_SITE
FROM SPACEXDATASET;
```

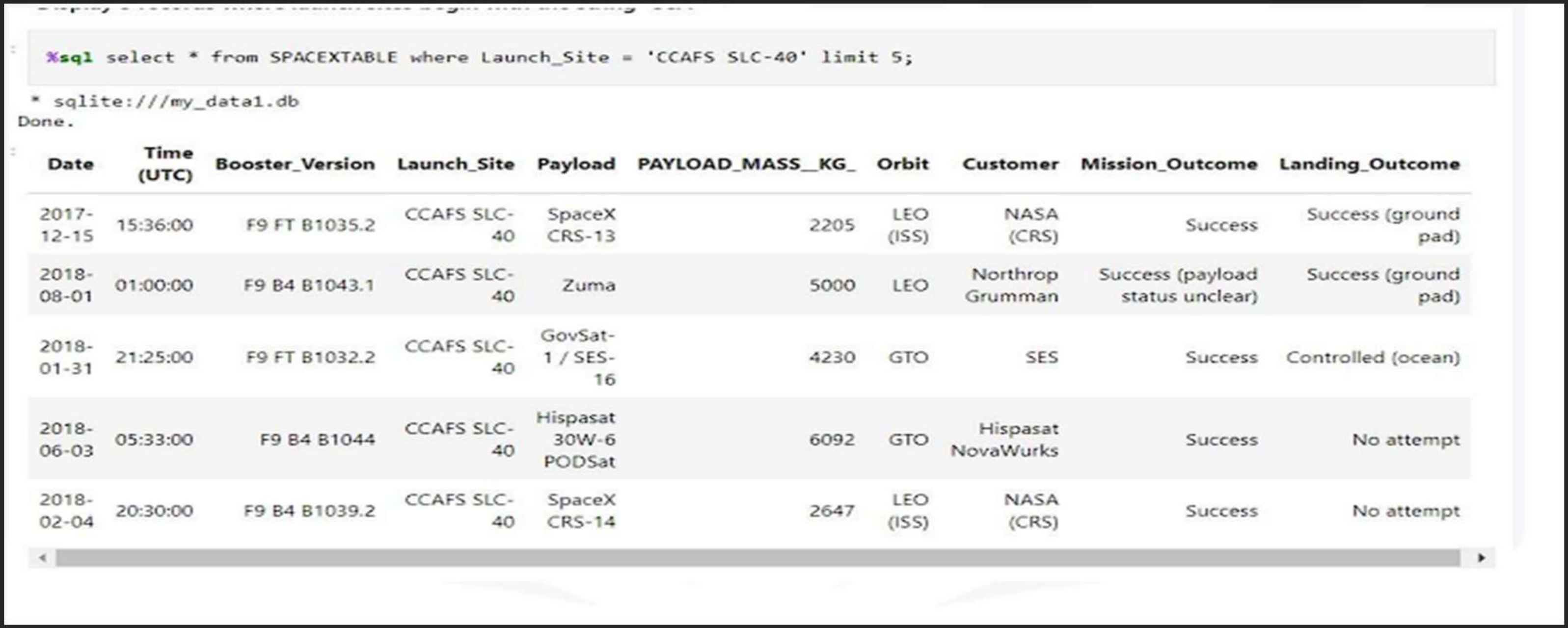
```
 * ibm_db_sa://ftb12020:***@0c77d6f
Done.
```

Out[4]:

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

Launch Site Names Begin with 'CCA'

- Task 2: Display 5 records where launch sites begin with the string 'CCN



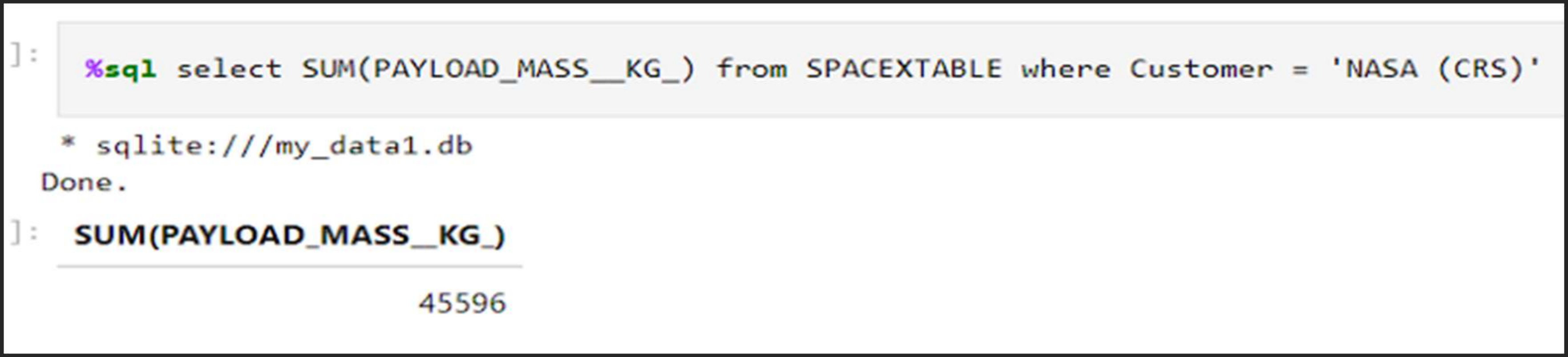
```
%sql select * from SPACEXTABLE where Launch_Site = 'CCAFS SLC-40' limit 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-12-15	15:36:00	F9 FT B1035.2	CCAFS SLC-40	SpaceX CRS-13	2205	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2018-08-01	01:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000	LEO	Northrop Grumman	Success (payload status unclear)	Success (ground pad)
2018-01-31	21:25:00	F9 FT B1032.2	CCAFS SLC-40	GovSat-1 / SES-16	4230	GTO	SES	Success	Controlled (ocean)
2018-06-03	05:33:00	F9 B4 B1044	CCAFS SLC-40	Hispasat 30W-6 PODSat	6092	GTO	Hispasat NovaWurks	Success	No attempt
2018-02-04	20:30:00	F9 B4 B1039.2	CCAFS SLC-40	SpaceX CRS-14	2647	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Task 3: Display the total payload mass carried by boosters launched by NASA (CRS)



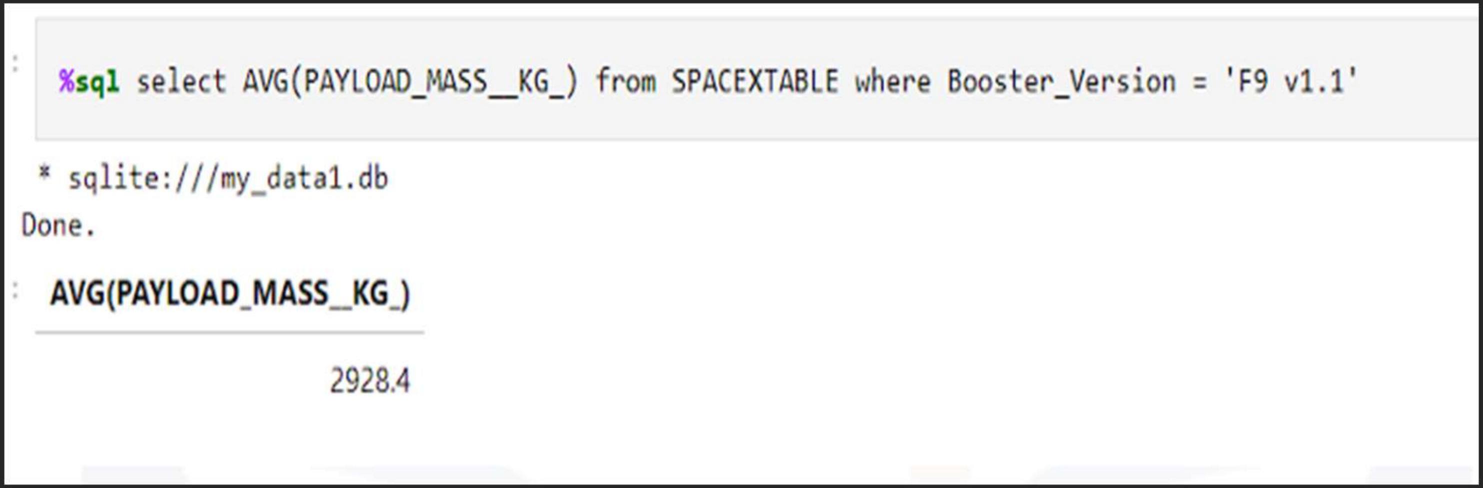
```
]:  %sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'

     * sqlite:///my_data1.db
    Done.
]:   SUM(PAYLOAD_MASS__KG_)

                      45596
```

Average Payload Mass by F9 v1.1

- Task 4: Display average payload mass carried by booster version F9 vl.l



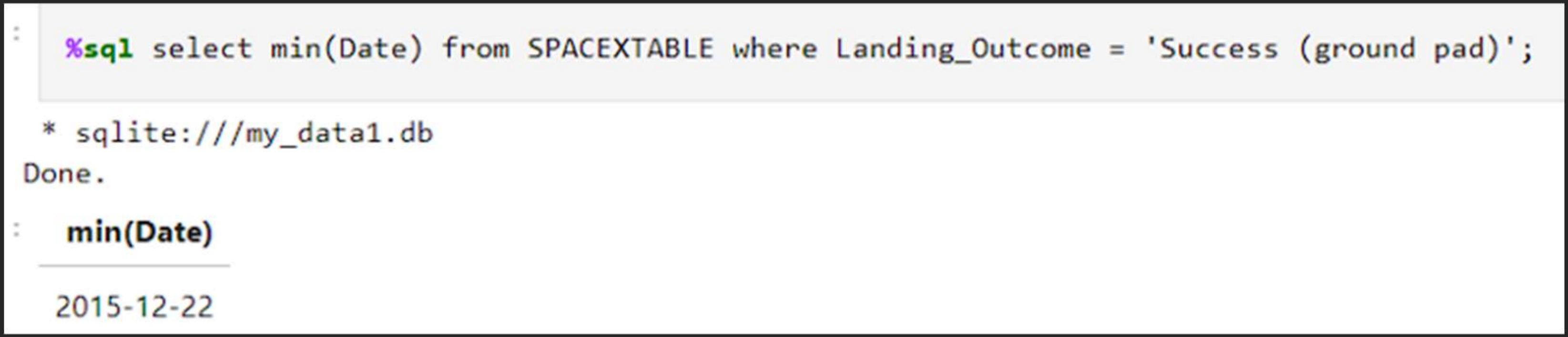
```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Task 5: List the date when the first successful landing outcome in ground pad was achieved.



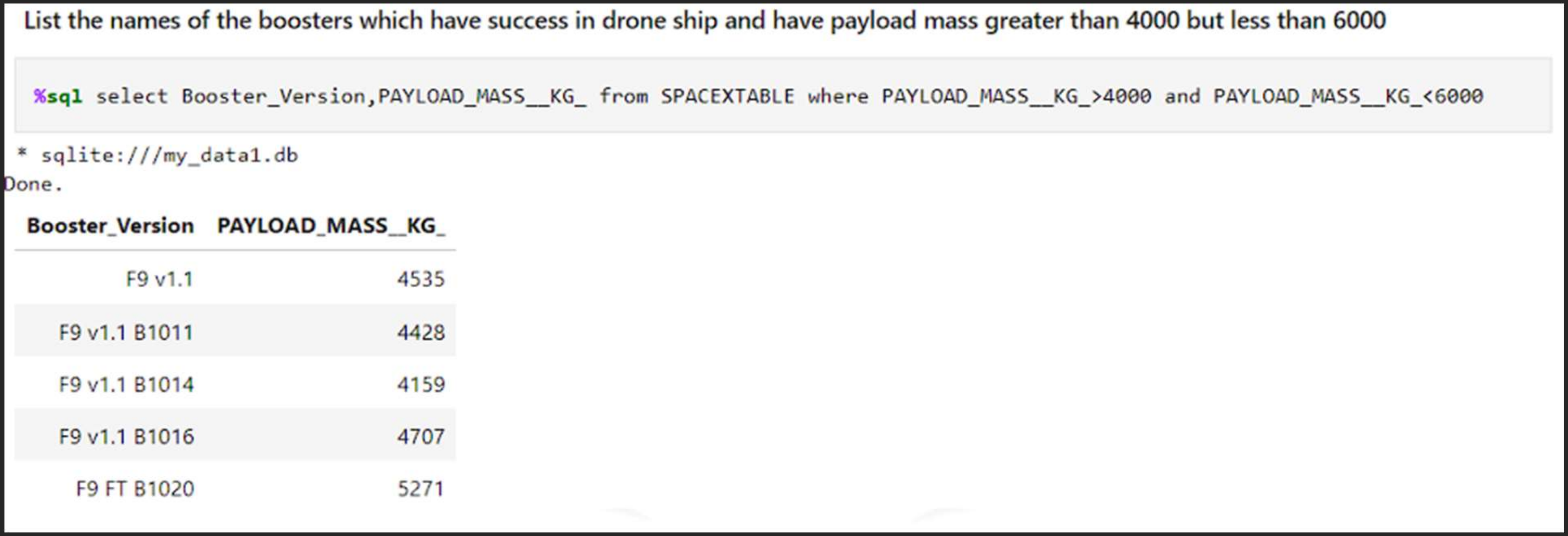
```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

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

min(Date)
2015-12-22

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



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

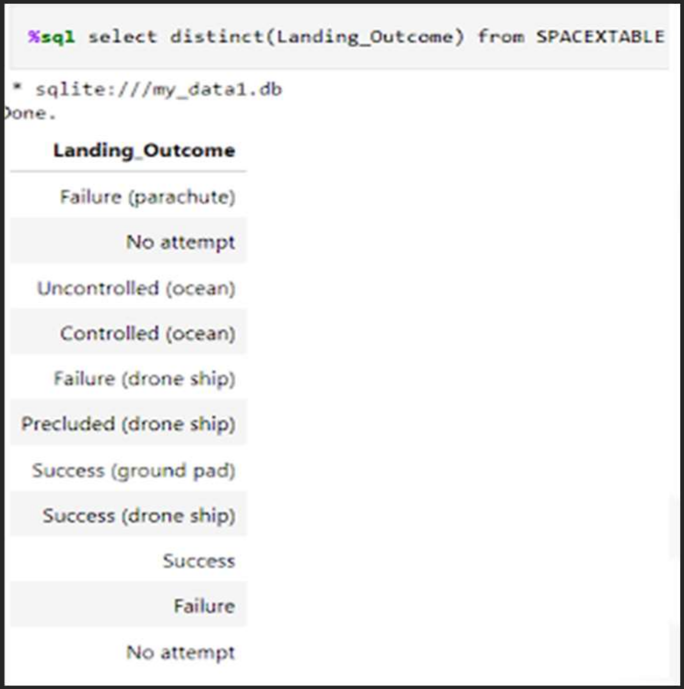
```
%sql select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000
```

```
 * 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

Total Number of Successful and Failure Mission Outcomes

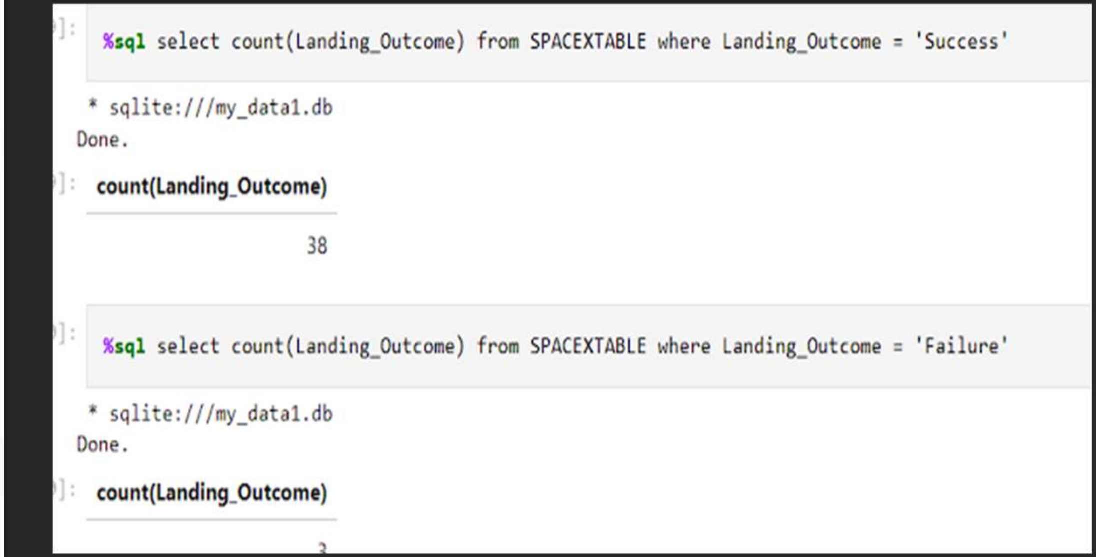
- Task 7: List the total number of successful and failure mission outcomes



```
%sql select distinct(Landing_Outcome) from SPACEXTABLE
```

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

Landing_Outcome
Failure (parachute)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
Success (ground pad)
Success (drone ship)
Success
Failure
No attempt



```
%sql select count(Landing_Outcome) from SPACEXTABLE where Landing_Outcome = 'Success'
```

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

count(Landing_Outcome)
38

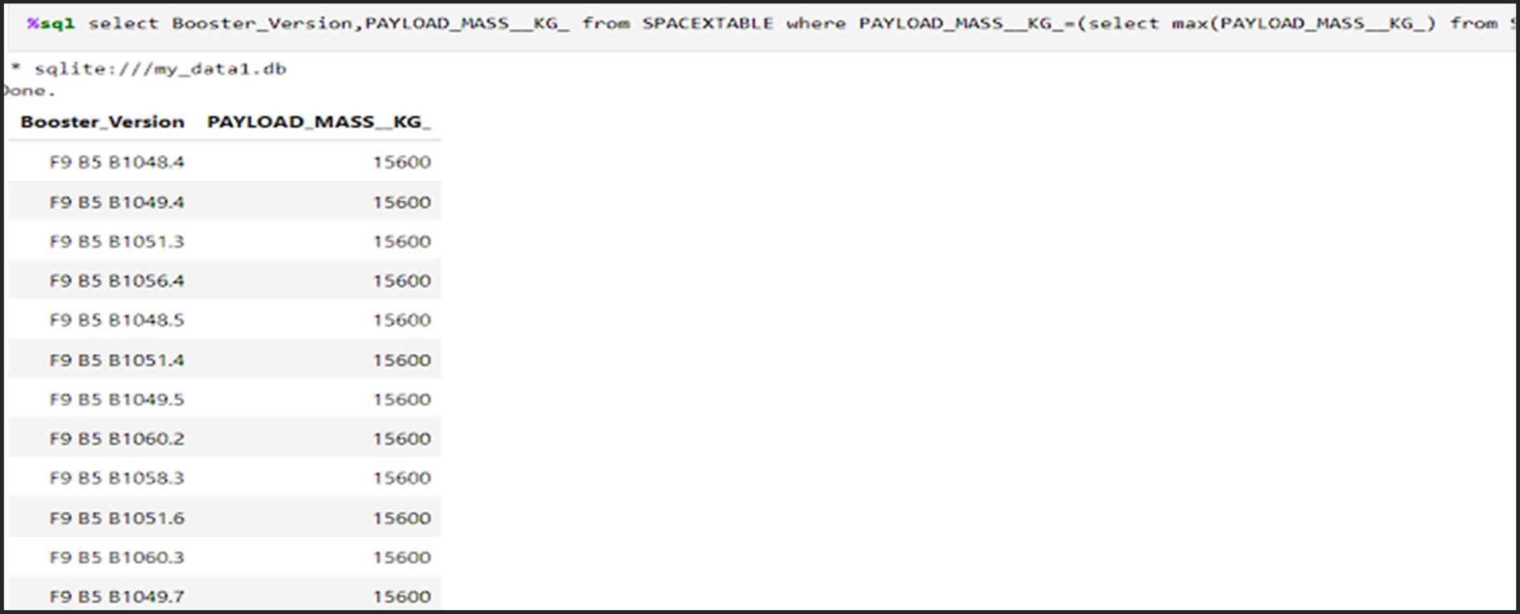
```
%sql select count(Landing_Outcome) from SPACEXTABLE where Landing_Outcome = 'Failure'
```

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

count(Landing_Outcome)
3

Boosters Carried Maximum Payload

- Task 8: List the names of the booster versions which have carried the maximum payload mass



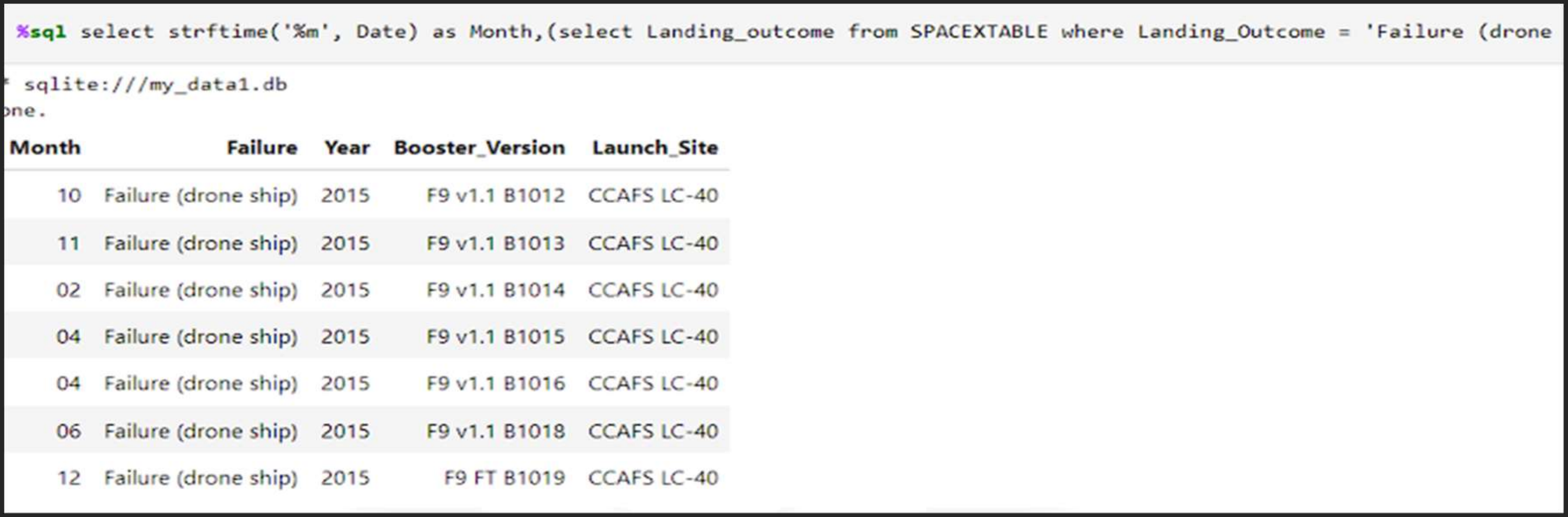
```
%sql select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from
```

```
* 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

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.



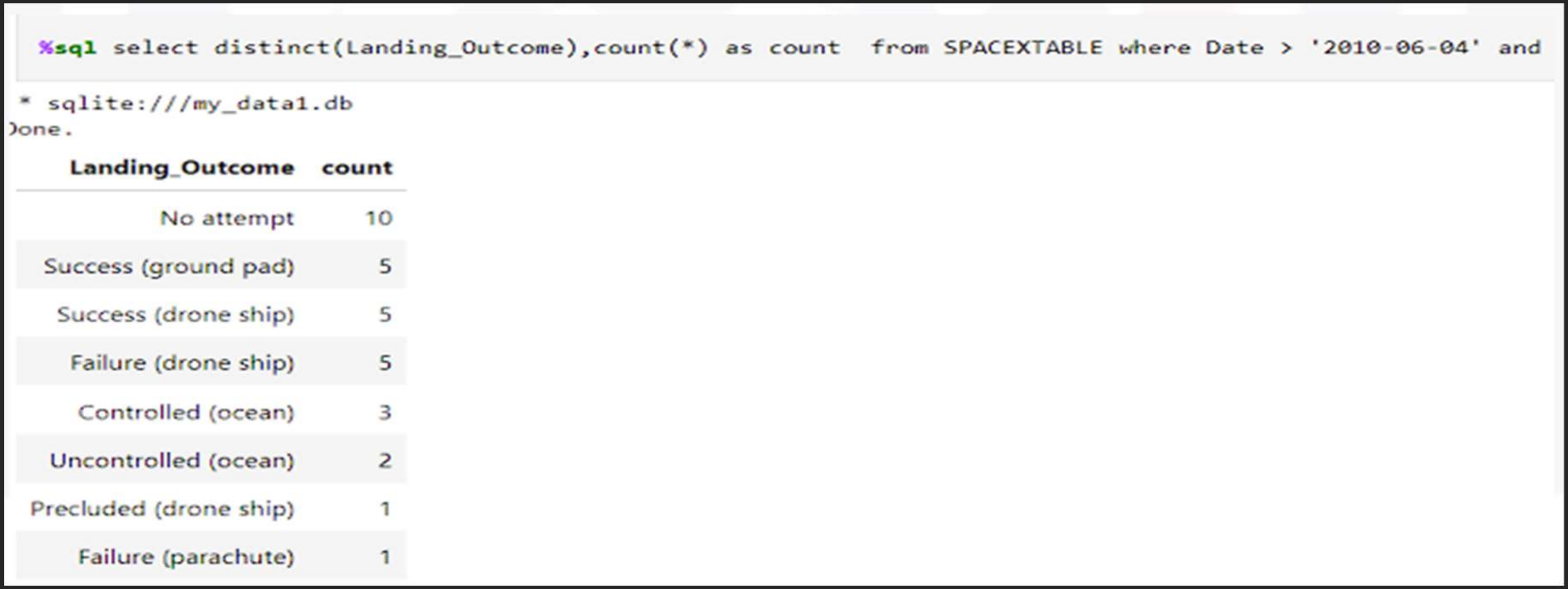
```
%sql select strftime('%m', Date) as Month,(select Landing_outcome from SPACEXTABLE where Landing_Outcome = 'Failure (drone
```

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

Month	Failure	Year	Booster_Version	Launch_Site
10	Failure (drone ship)	2015	F9 v1.1 B1012	CCAFS LC-40
11	Failure (drone ship)	2015	F9 v1.1 B1013	CCAFS LC-40
02	Failure (drone ship)	2015	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	2015	F9 v1.1 B1015	CCAFS LC-40
04	Failure (drone ship)	2015	F9 v1.1 B1016	CCAFS LC-40
06	Failure (drone ship)	2015	F9 v1.1 B1018	CCAFS LC-40
12	Failure (drone ship)	2015	F9 FT B1019	CCAFS LC-40

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.



```
%sql select distinct(Landing_Outcome),count(*) as count  from SPACEXTABLE where Date > '2010-06-04' and
```

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

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

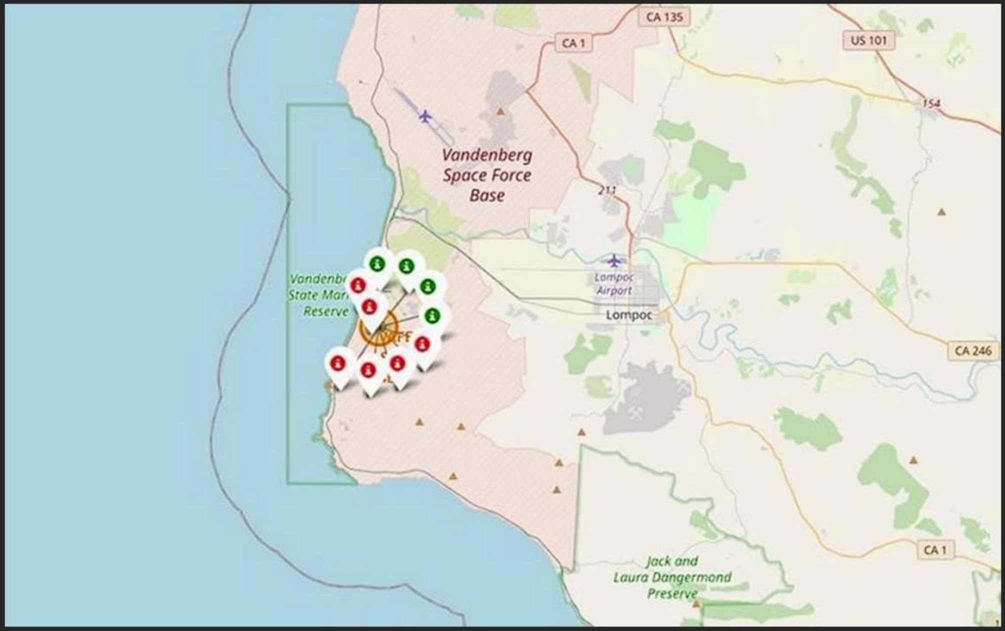
Section 3

Launch Sites Proximities Analysis

INTERACTIVE MAP USING FOLIUS

COLOR CODED MARKERS

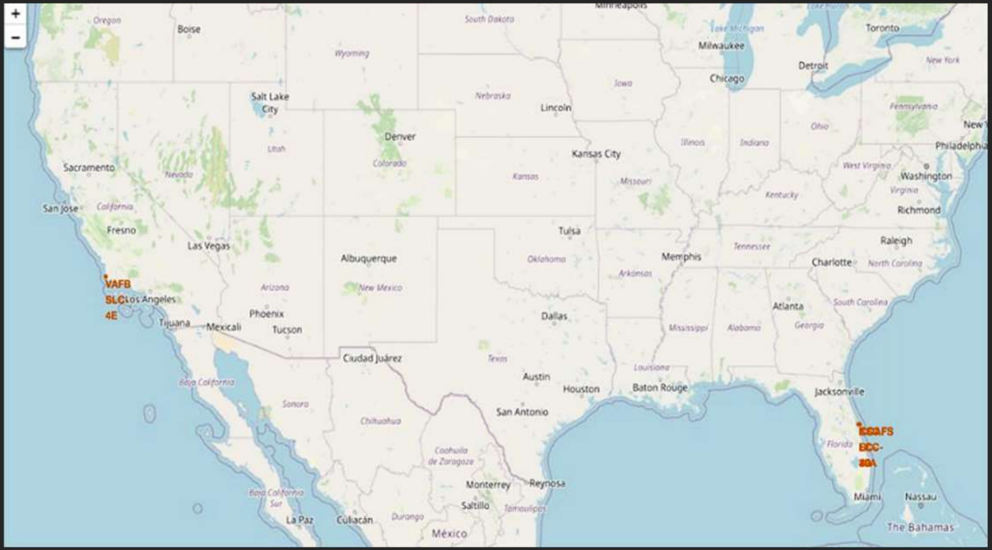
Color coded makers are used to represent whether the landing has been a failure or a success. The green markers represent the successful landing and red markers represent failure.

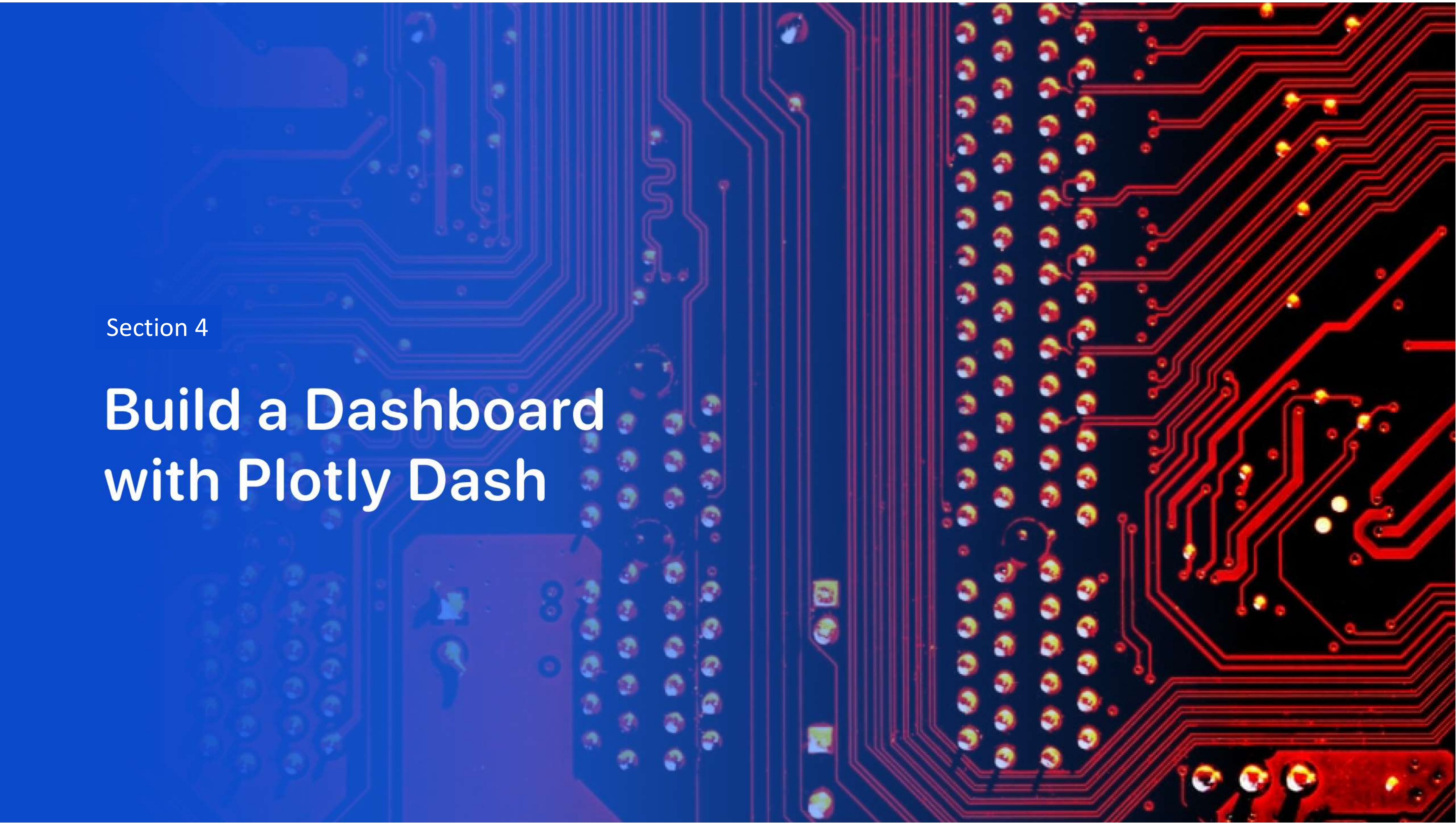


INTERACTIVE MAP USING FOLIUM

LAUNCH SITE LOCATIONS

The map shows all launch sites relative US map. The right side of the map shows the Florida launch sites on the east coast. The left side shows the launch site from the west coast / California. All launch sites are near the ocean.

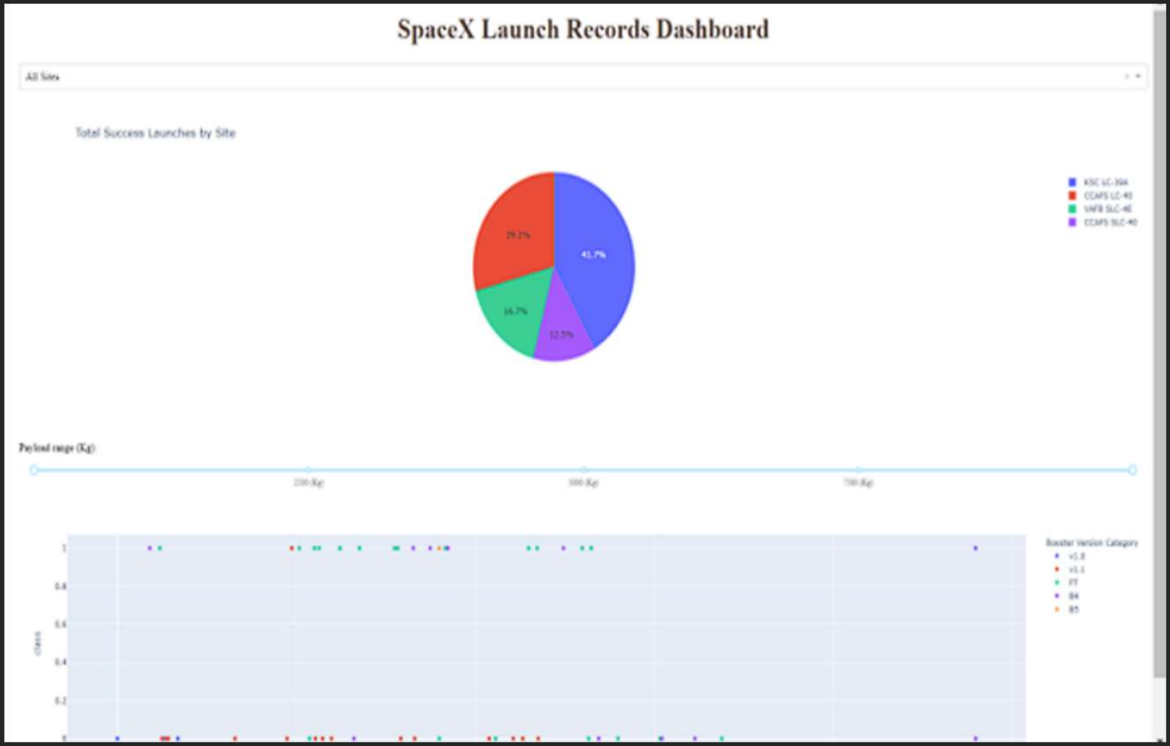




Section 4

Build a Dashboard with Plotly Dash

PLOTLY DASHBORAD



SpaceX Launch Records Dashboard

This the image of the Dashboard which consist of the result from EDA and Interactive Map with Folium

Section 5

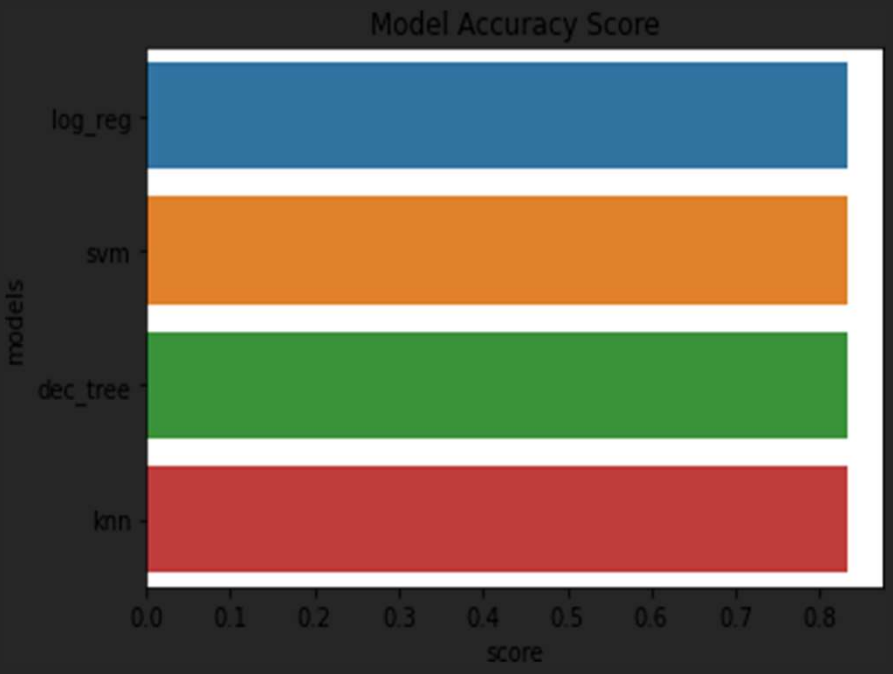
Predictive Analysis (Classification)

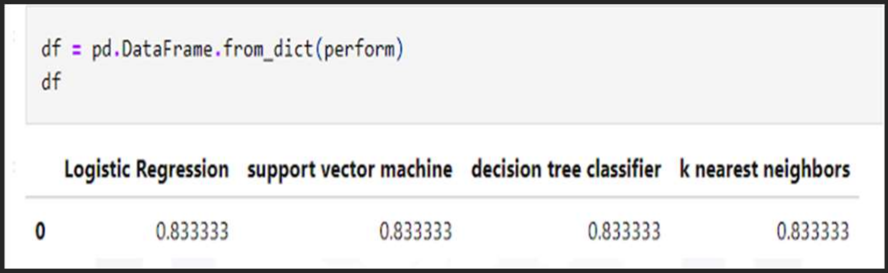
Classification Accuracy

- This the image gives the result of the score for each ML model we have used which comes out be 83.33% for all the model. Test size is small at only sample size of 18.

- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.

- We likely need more data to determine the best model.



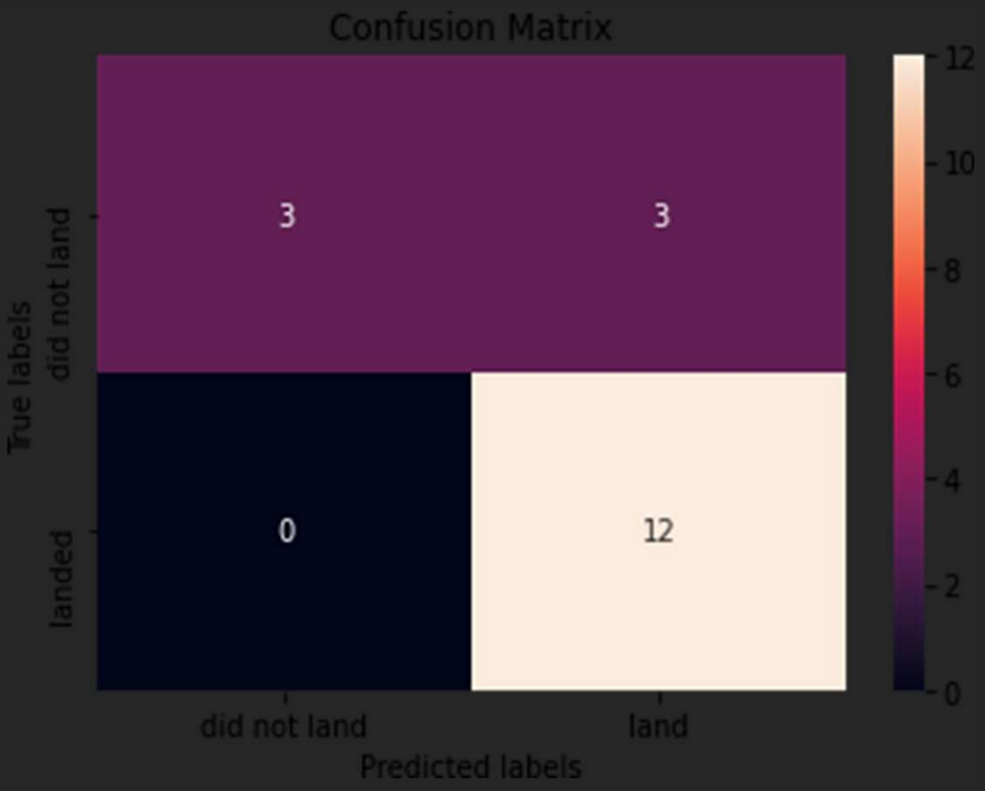


```
df = pd.DataFrame.from_dict(perform)
df
```

	Logistic Regression	support vector machine	decision tree classifier	k nearest neighbors
0	0.833333	0.833333	0.833333	0.833333

Confusion Matrix

- This the image depicts the confusion matrix for each ML model we have used in this project.
- The total number of correct predictions are 3+12 = 15 (diagonal)



Conclusions

- Our task: to develop a machine learning model for Space Y so that they can compete with Space X in the space race.
- The goal of model is to predict when Stage 1 will successfully land in-order to be financially improved than Space X. (around $100 million)
- We created a machine learning model with an accuracy of 83%.
- If possible more data should be collected to better determine the best machine learning model and improve accuracy.

Appendix

- **Instructors:**

  Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo

- **Special Thanks to All Instructors:**

  https://www.coursera.org/professional-certificates/ibm-datascience?#instructors

- **GITHUB Link:**

  xxwizardxx117/Courcera-Capstone-Project (github.com)

Thank you!