IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Talgat Azhibekov
25/09/2023

# OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

# EXECUTIVE SUMMARY

In this data science capstone project an outcome of Space X ship Falcon 9 first stage launch was predicted. First off all, data was collected using two methods requesting API and Web scraping. Then, data was processed by using pandas library followed by EDA techniques. EDA was performed with data visualization applying seaborn library and with SQL. During these stages data was cleaned and transformed to necessary formats, interesting feature relationships and dependencies were identified, landings success rates were calculated for each orbits and launch sites. To visually observe identified insights Plotly dashboards were created. Afterwards, each launch sites with landing outcomes were depicted on the world map by means of Folium functionality. Finally, predictive analysis was performed by using four classification models with accuracy 83.33%.

# INTRODUCTION

In 1961 first human successfully made one orbit around Earth. Since then active space exploration developed. The main drawback of such missions is extremely high budget. Nowadays, due to the usable first stage technology a Space X company can launch Falcon 9 rocket with a cost of 62 million dollars. While other providers offers start from 162 million dollars.

In this capstone, based on historical data, the Falcon 9 first stage successful landing will be predicted. Therefore, the cost of a launch can be determined. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

# Methodology

# METHODOLOGY

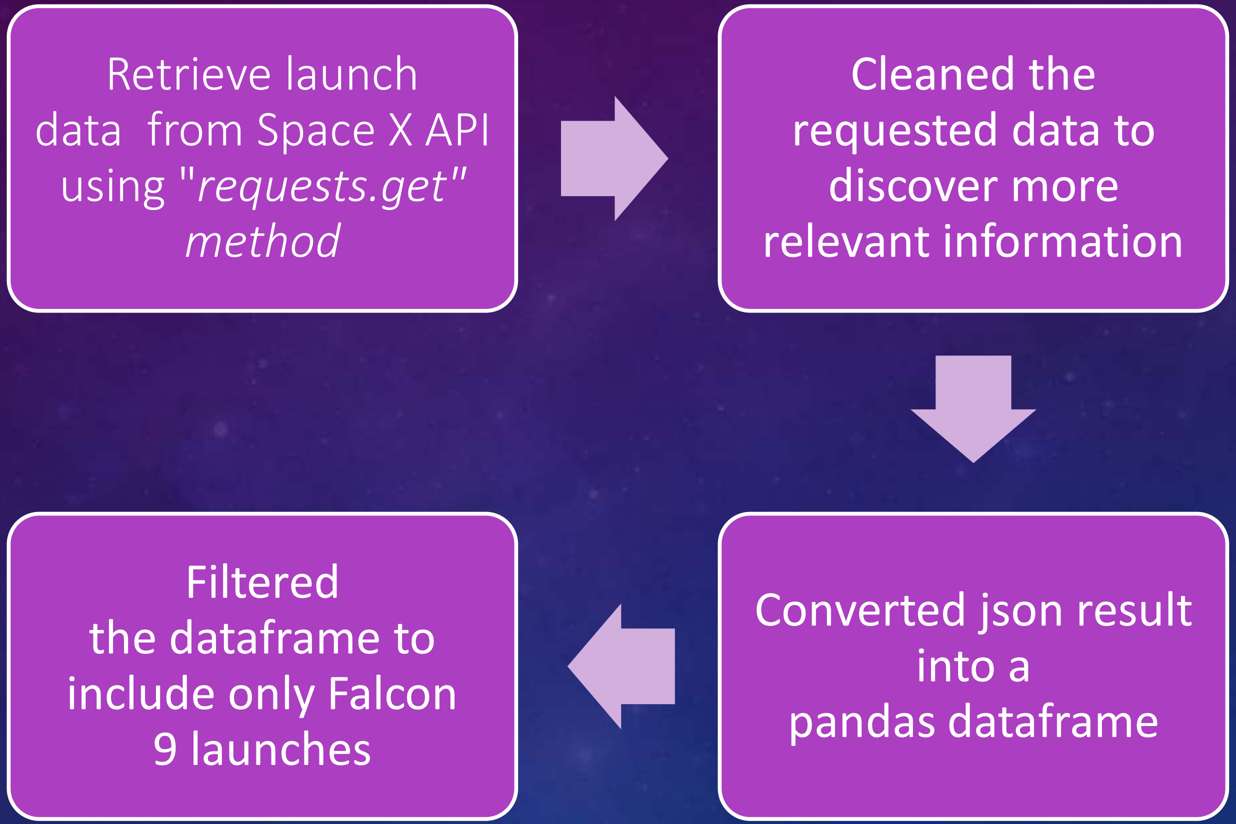
- Executive Summary
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

# DATA COLLECTION

Data sets were collected by 2 ways:

- Made a get request to the Space X API.
- Web scraping from the Wikipedia page

# DATA COLLECTION – SPACE X API



Retrieve launch data  from Space X API using "*requests.get*" *method*

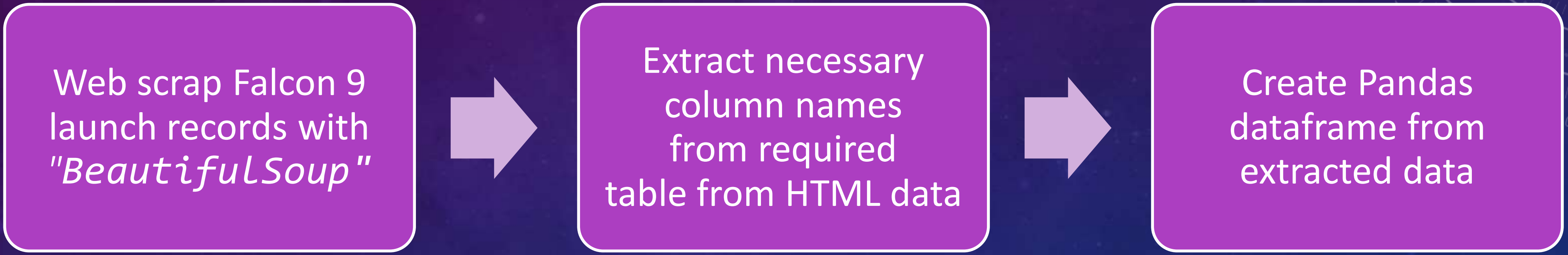
**Cleaned the requested data to discover more relevant information**

Converted json result into a pandas dataframe

Filtered the dataframe to include only Falcon 9 launches

GitHub URL:
[https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git](https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git)

# DATA COLLECTION – WEB SCRAPING



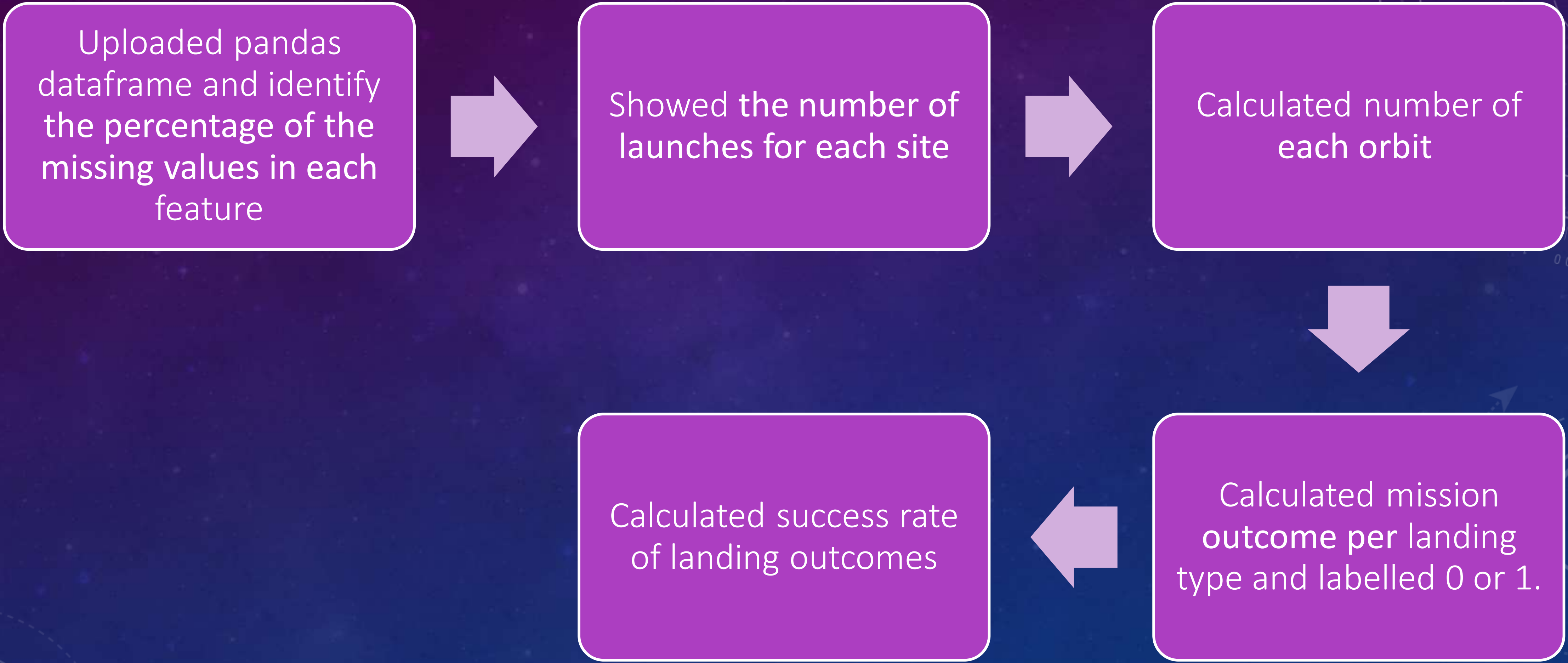
Web scrap Falcon 9 launch records with *"BeautifulSoup"*

Extract necessary column names from required table from HTML data

Create Pandas dataframe from extracted data

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# DATA WRANGLING



Uploaded pandas dataframe and identify **the percentage of the missing values in each** feature

Showed **the number of launches for each site**

Calculated number of **each orbit**

Calculated mission **outcome per** landing type and labelled 0 or 1.

Calculated success rate of landing outcomes

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# EDA WITH DATA VISUALIZATION

By using the seaborn Python data visualization library the following  main charts were plotted to identify useful relationships between data features:

- **FlightNumber vs LaunchSite** – By increasing the flight numbers the number of successful landings increased from all launch sites.
- **FlightNumber vs PayloadMass** – More massive the payload, the less likely the first stage will return.
- **FlightNumber vs Orbit** – LEO orbit's Success is related to the number of flights. On the other hand, there is no relationship between flight number in GTO orbit.
- **Success rates of the orbits** –  ES-L1, GEO, HEO and SSO have the highest success rates.
- **Success rate vs Years** –  Success rate since 2013 kept increasing till 2020 with slight decrease in 2018.

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# EDA WITH SQL

After loading the SQL extension and establishing a connection with the database the following main SQL queries were performed:

- Displayed the unique launch sites and the total payload mass carried by boosters launched by NASA (CRS).
- Identified the date when the first successful landing outcome in ground pad was achieved and the total number of successful and failure mission outcomes.
- Listed the names of the boosters which have success in drone ship and have payload mass between 4000kg and 6000kg.
- By using subquery found the names of the booster_versions which have carried the maximum payload mass.
- By using complicated queries extracted necessary launch information occurred in specific date frame.

GitHub URL: https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# BUILD AN INTERACTIVE MAP WITH FOLIUM

To perform more interactive visual analytics using Folium the following main map objects were created:

- Created circles for each launch sites with a popup labels showing their names.
- Created markers at each launch sites with icons showing their names.
- Created marker clusters at each launch sites with colour codes. Green – successful, Red- unsuccessful.
- Calculated distances from launch sites to coastlines, railways, highways, cities and drew lines to show closest routes.

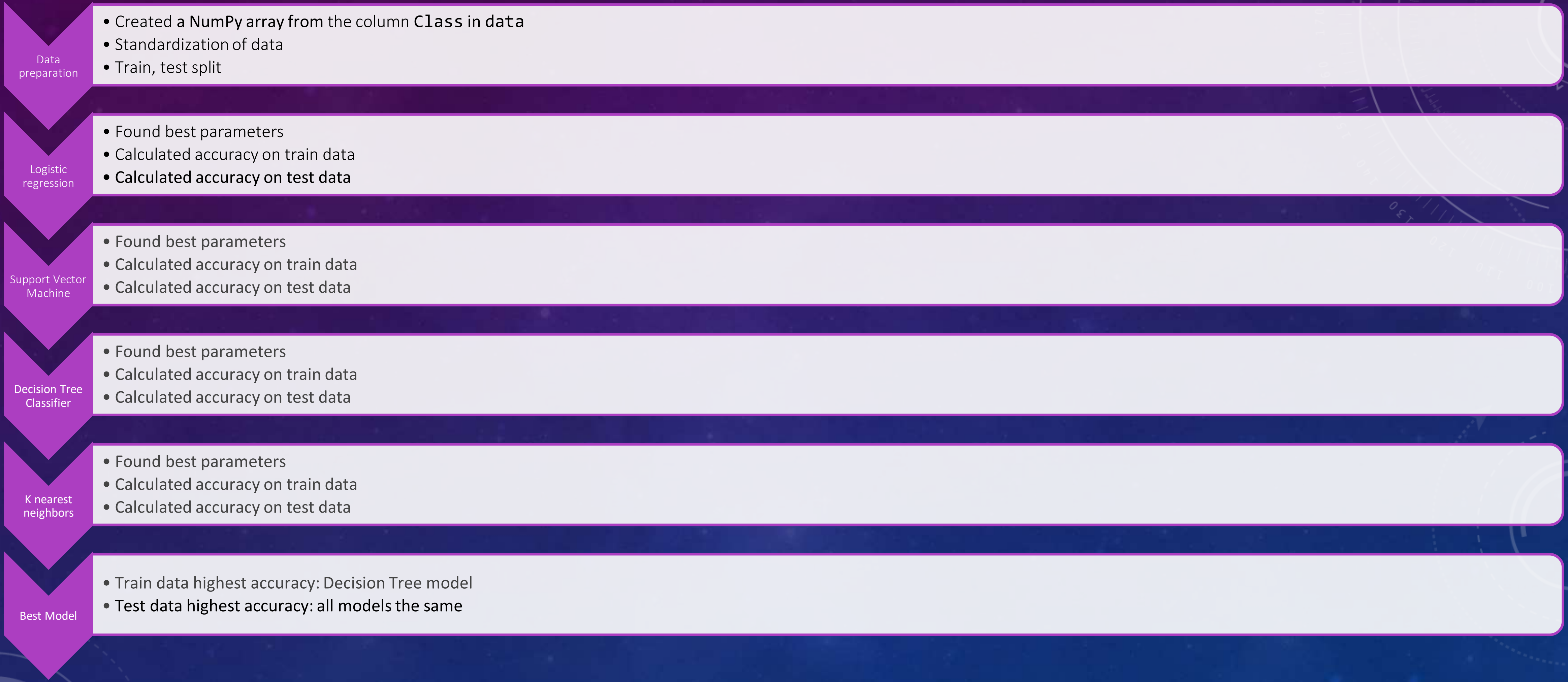
GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# BUILD A DASHBOARD WITH PLOTLY DASH

- Added a Launch Site Drop-down Input Component – to see which one has the largest success count.
- Added a callback function and displayed pie-chart – to show success rates of launch sites on pie-chart based on selected site dropdown.
- Added a Range Slider to Select Payload – to easily select different payload range and see if variable payload is correlated with mission outcome.
- Added a callback function and displayed scatter chart – to observe how payload may be correlated with mission outcomes for selected launch sites.

GitHub URL:
[https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git](https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git)

# PREDICTIVE ANALYSIS (CLASSIFICATION)



Data preparation
- Created a NumPy array from the column `Class` in `data`
- Standardization of data
- Train, test split

Logistic regression
- Found best parameters
- Calculated accuracy on train data
- Calculated accuracy on test data

Support Vector Machine
- Found best parameters
- Calculated accuracy on train data
- Calculated accuracy on test data

Decision Tree Classifier
- Found best parameters
- Calculated accuracy on train data
- Calculated accuracy on test data

K nearest neighbors
- Found best parameters
- Calculated accuracy on train data
- Calculated accuracy on test data

Best Model
- Train data highest accuracy: Decision Tree model
- Test data highest accuracy: all models the same

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# RESULTS

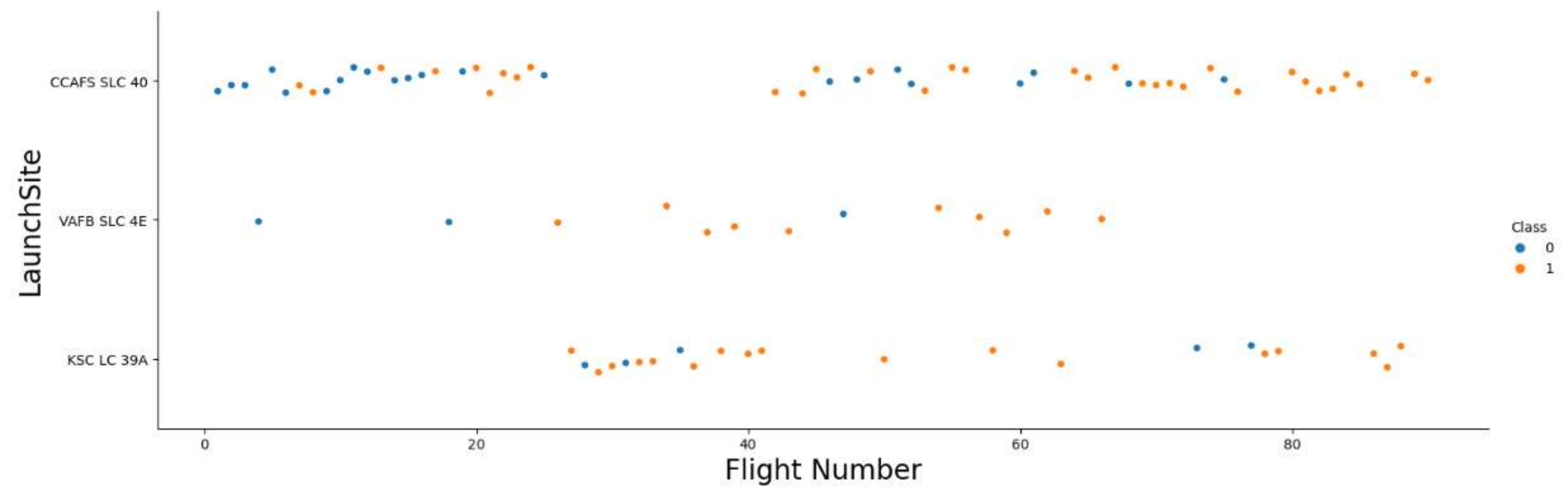
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

Section 2

# Insights drawn from EDA

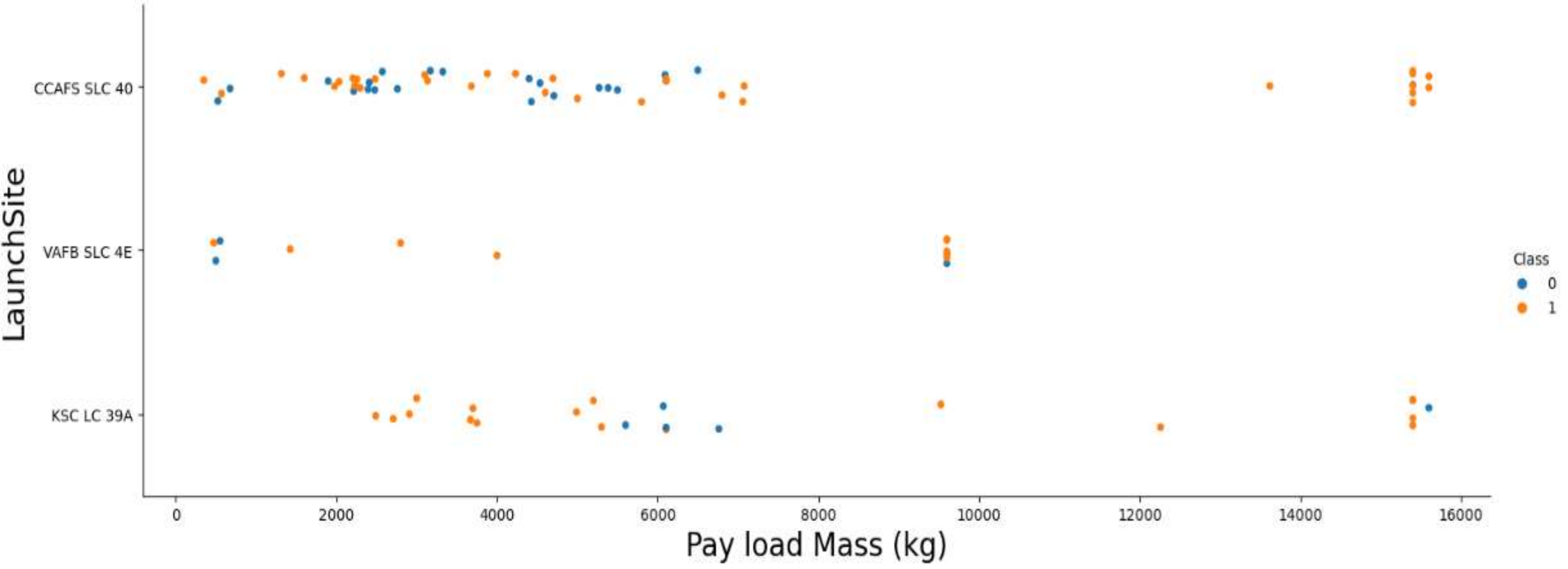
# FLIGHT NUMBER VS. LAUNCH SITE



By increasing the flight numbers the number of successful landings increased from all launch sites

GitHub URL:
[https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git](https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git)

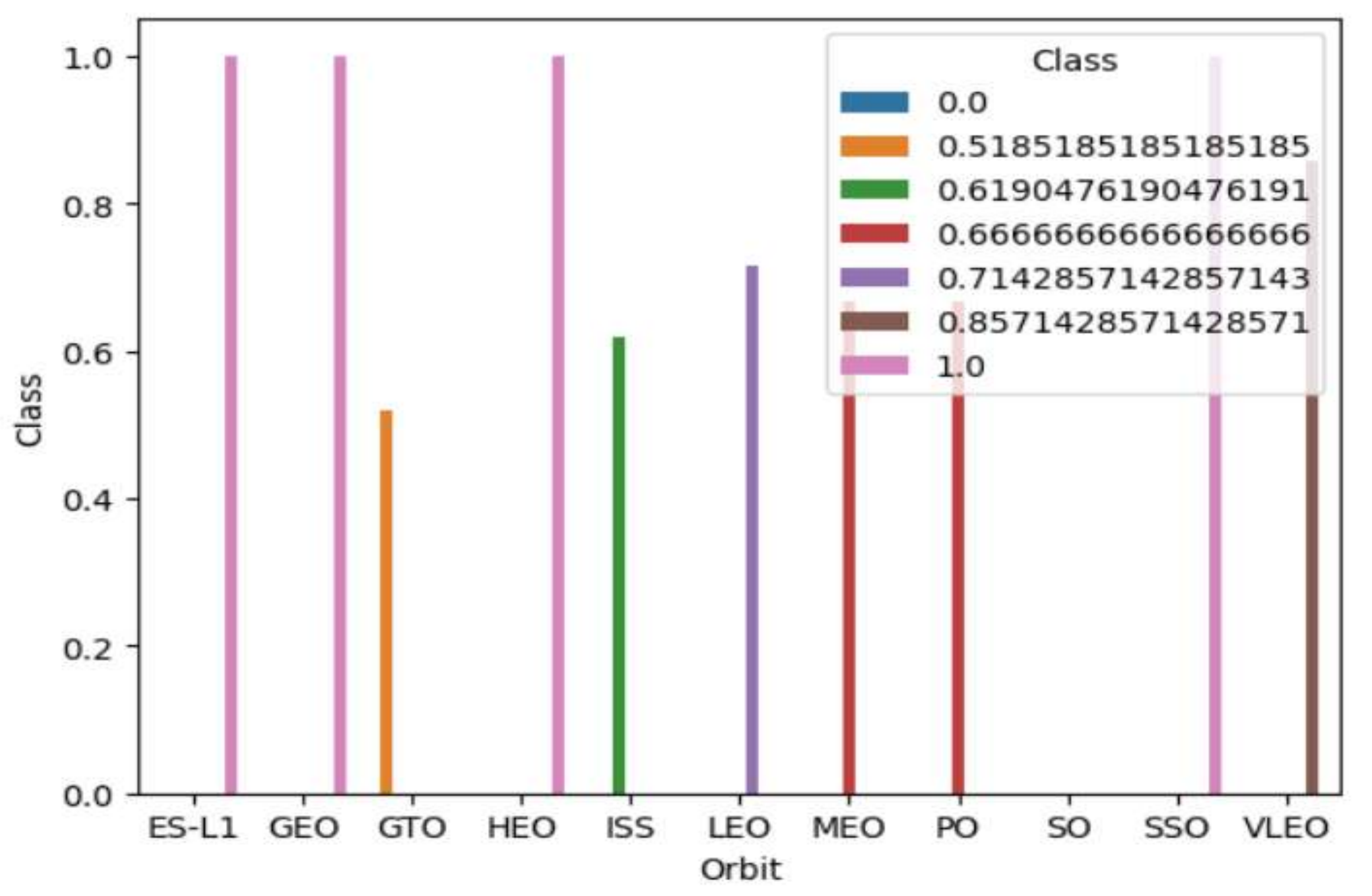
# PAYLOAD VS. LAUNCH SITE



It is shown on this scatter point chart that the VAFB-SLC 4E launchsite has no rockets launched for heavy payload mass(greater than 10000kg). Also for CCAFS SLC 40 launchsite success rate increases above pay load mass 8000kg.

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

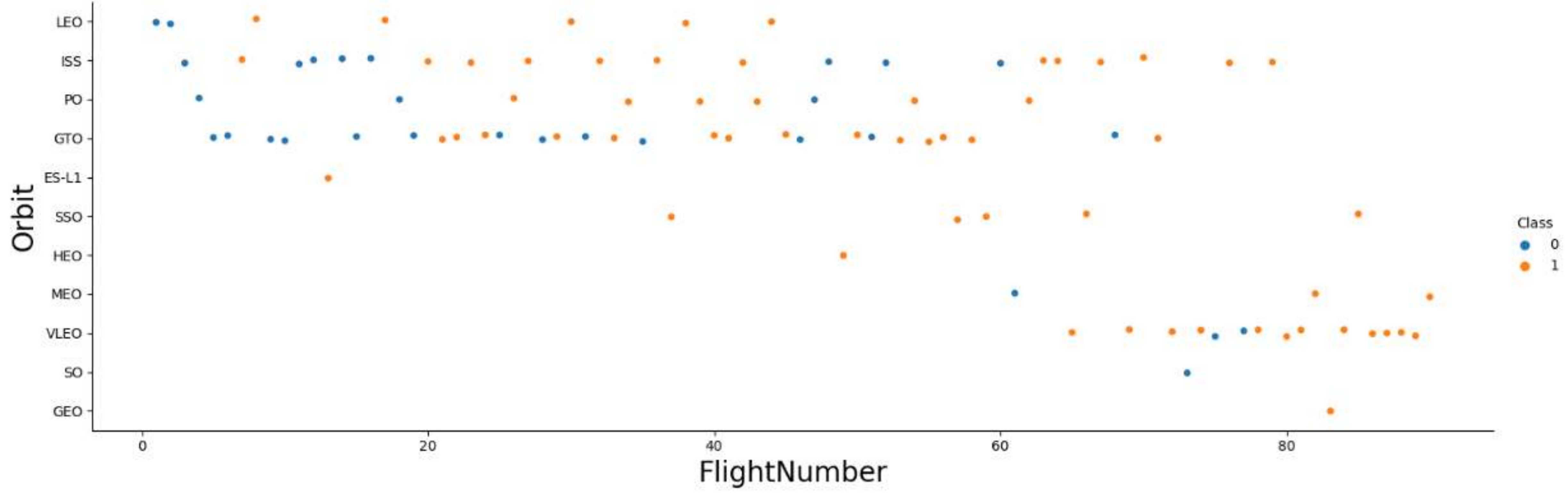
# SUCCESS RATE VS. ORBIT TYPE



It is shown on this bar chart that the orbits ES-L1, GEO, HEO and SSO have the highest success rates.

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

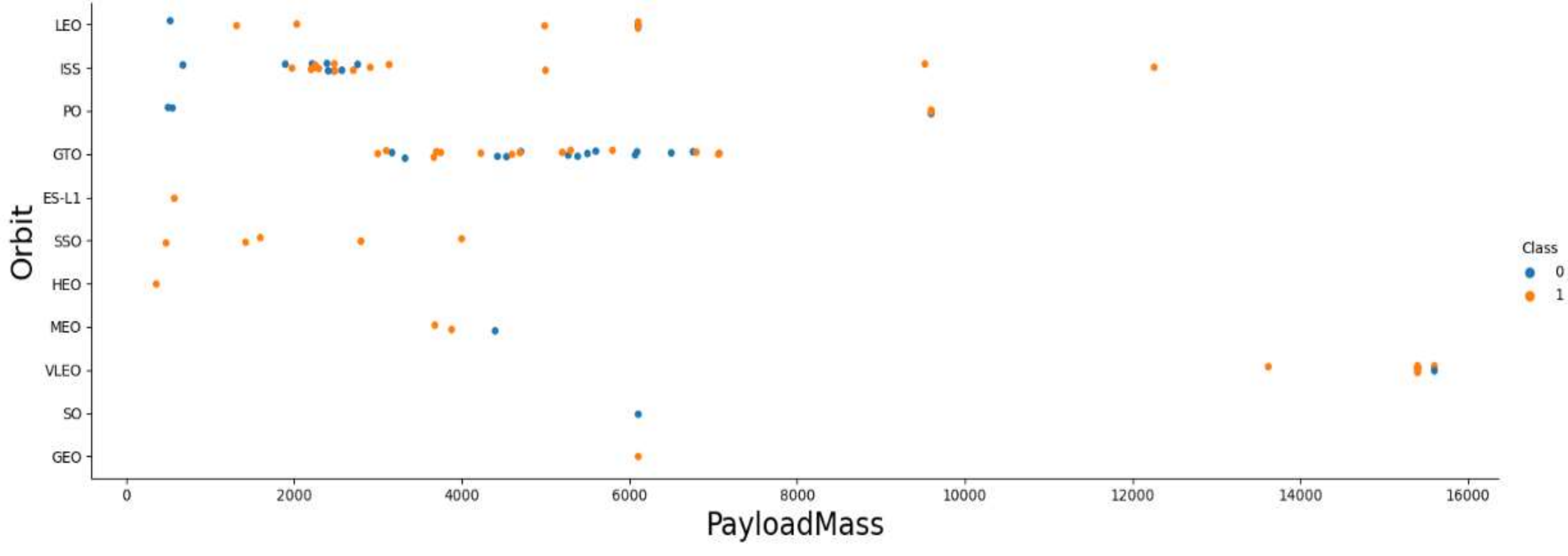
# FLIGHT NUMBER VS. ORBIT TYPE



LEO orbit's Success is related to the number of flights. On the other hand, there is no relationship between flight number in GTO orbit.

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

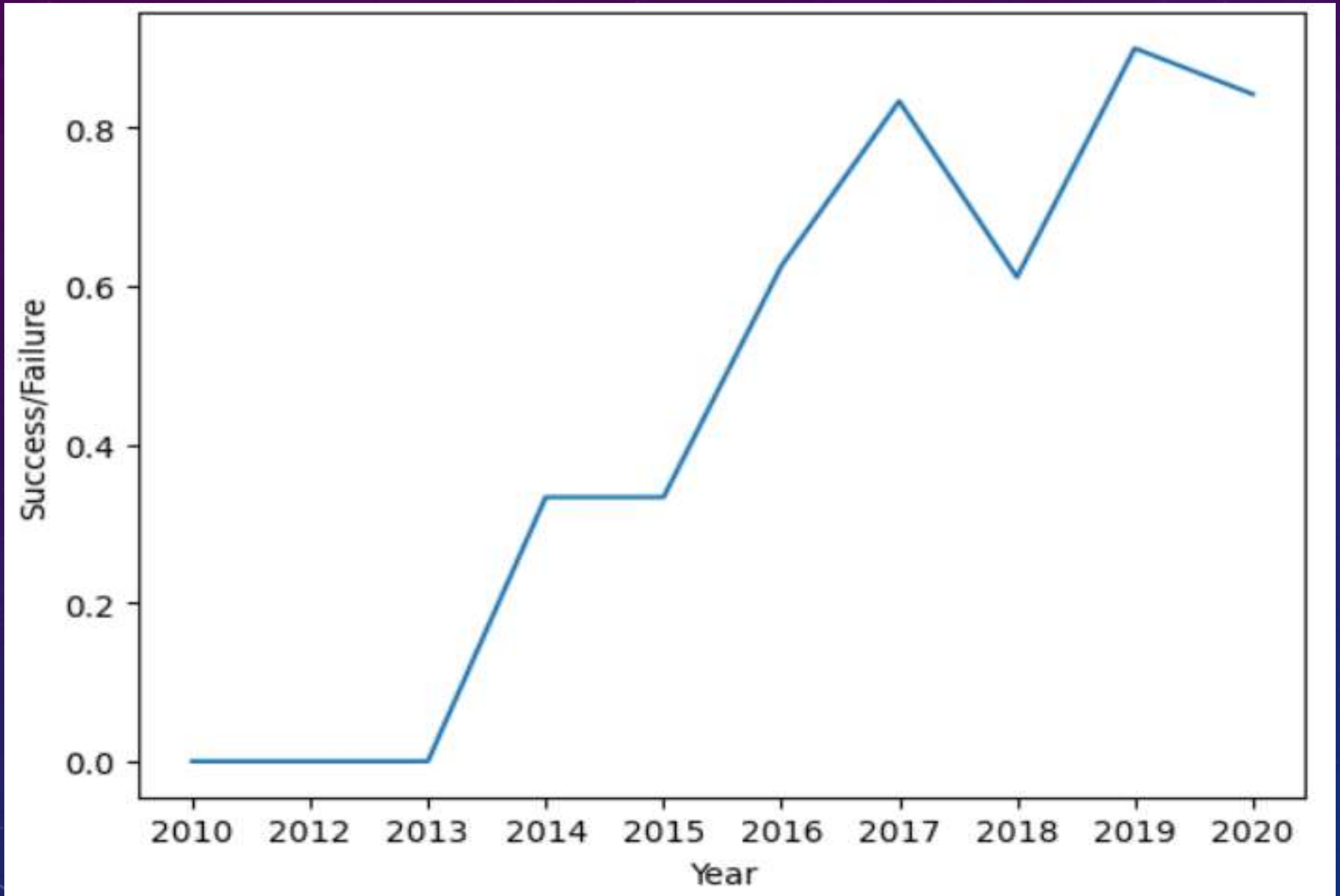
# PAYLOAD VS. ORBIT TYPE



With heavy payloads the successful landing rate are higher for PO, LEO and ISS.
However for GTO it cannot be distinguished well as both positive and negative landings are both there.

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

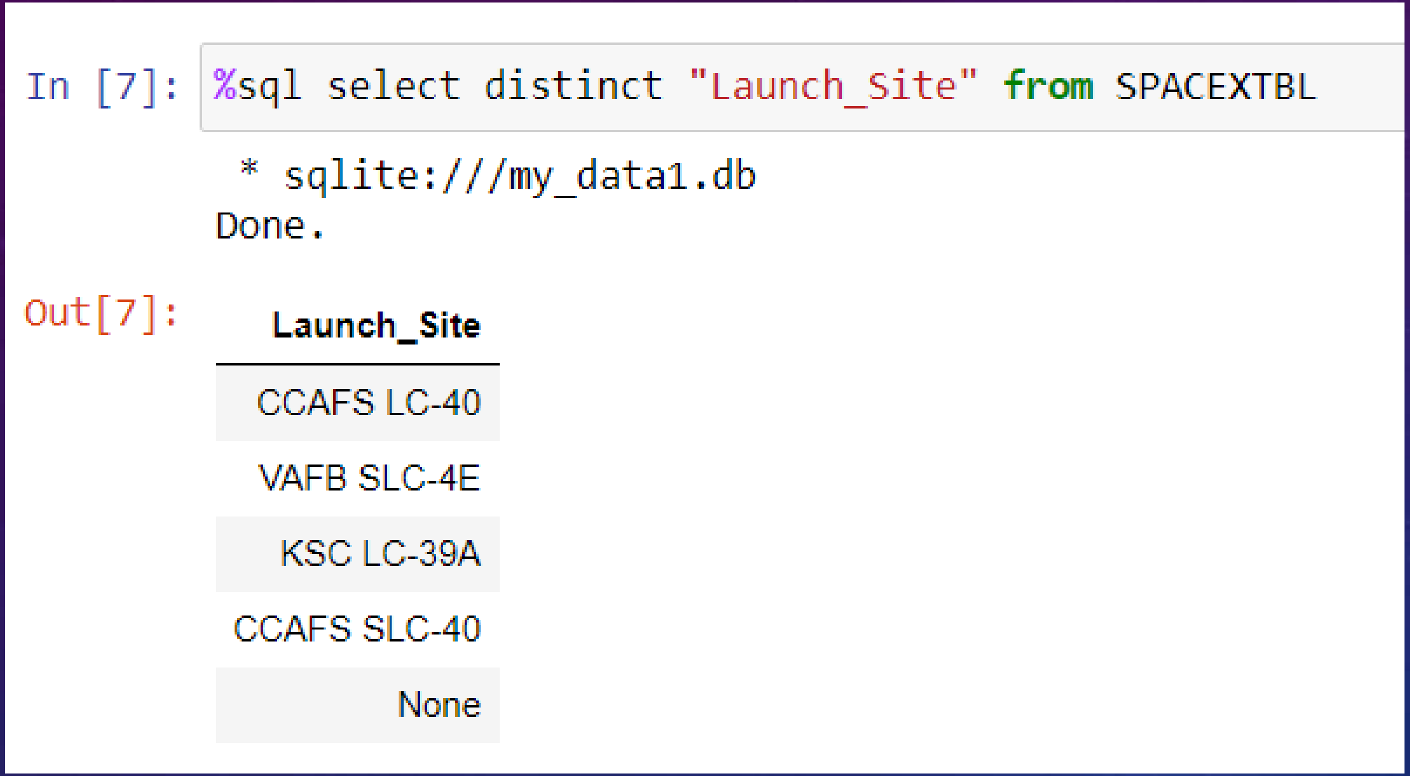
# LAUNCH SUCCESS YEARLY TREND



Success rate since 2013 kept increasing till 2020 with slight decrease in 2018

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# ALL LAUNCH SITE NAMES



```
In [7]: %sql select distinct "Launch_Site" from SPACEXTBL

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

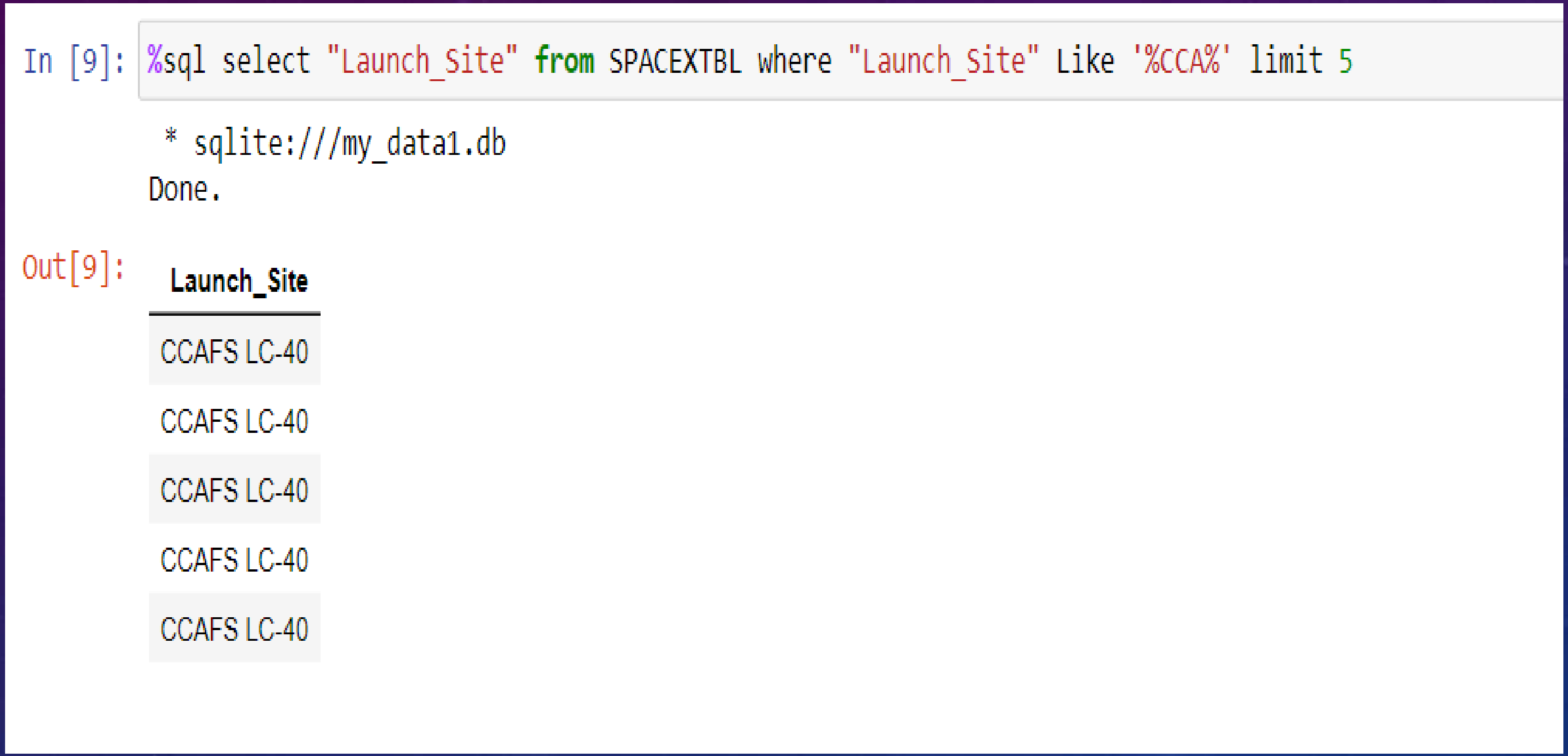
Out[7]:

| Launch_Site |
|---|
| CCAFS LC-40 |
| VAFB SLC-4E |
| KSC LC-39A |
| CCAFS SLC-40 |
| None |

By implementing SQL request the 4 unique launch sites were identified.

GitHub URL: https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# LAUNCH SITE NAMES BEGIN WITH 'CCA'

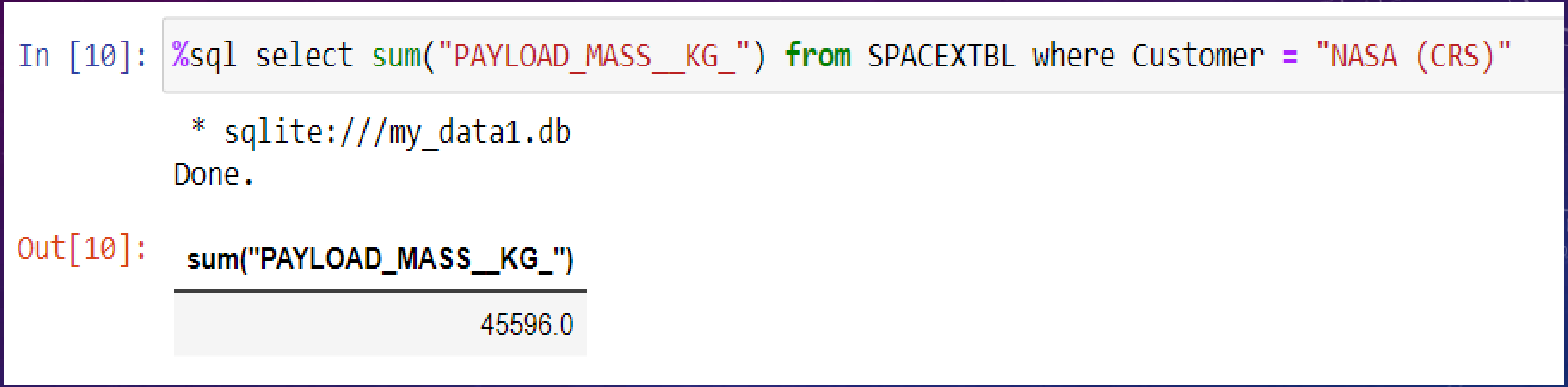


```
In [9]: %sql select "Launch_Site" from SPACEXTBL where "Launch_Site" Like '%CCA%' limit 5

         * sqlite:///my_data1.db
        Done.

Out[9]:  Launch_Site
        CCAFS LC-40
        CCAFS LC-40
        CCAFS LC-40
        CCAFS LC-40
        CCAFS LC-40
```

By implementing SQL request with limitation the 5 launch sites beginning from CCA were identified.

GitHub URL: [https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git](https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git)

# TOTAL PAYLOAD MASS



```
In [10]: %sql select sum("PAYLOAD_MASS__KG_") from SPACEXTBL where Customer = "NASA (CRS)"

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

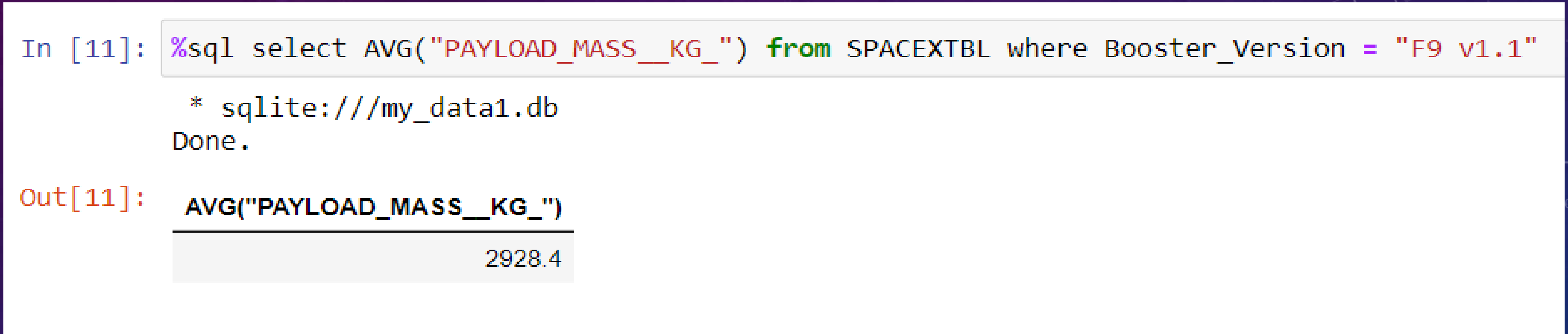
Out[10]:

| sum("PAYLOAD_MASS__KG_") |
|---:|
| 45596.0 |

By using sum function the total payload mass was calculated for "NASA (CRS)".

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# AVERAGE PAYLOAD MASS BY F9 V1.1

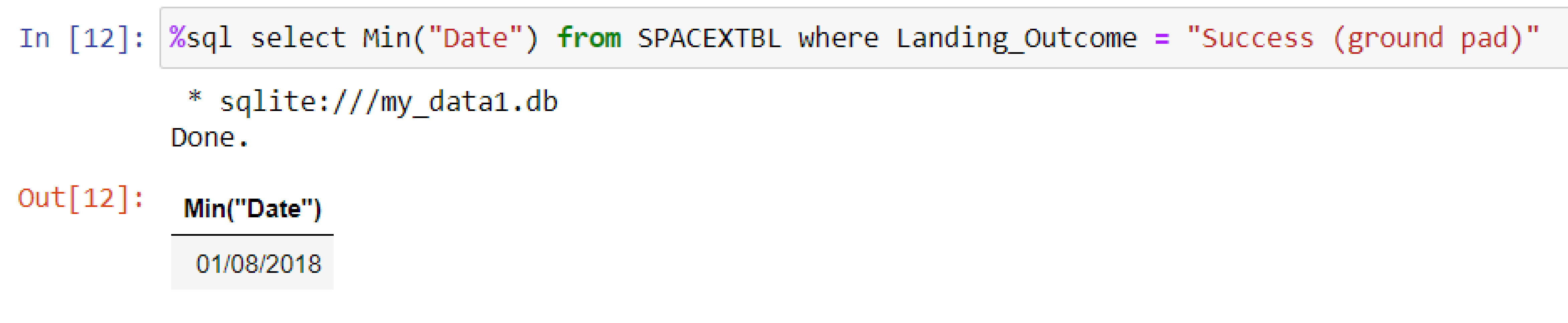


```
In [11]: %sql select AVG("PAYLOAD_MASS__KG_") from SPACEXTBL where Booster_Version = "F9 v1.1"

          * sqlite:///my_data1.db
         Done.

Out[11]: AVG("PAYLOAD_MASS__KG_")
                           2928.4
```

By using AVG function the average payload mass was calculated for booster version "F9 v1.1".

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# FIRST SUCCESSFUL GROUND LANDING DATE



```
In [12]: %sql select Min("Date") from SPACEXTBL where Landing_Outcome = "Success (ground pad)"
```

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

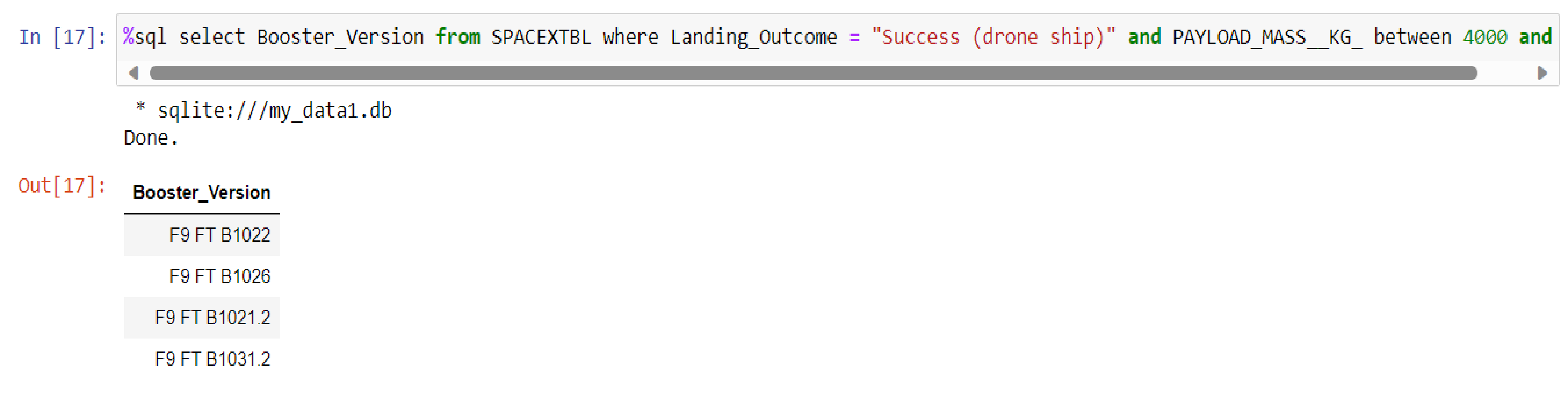
Out[12]:

| Min("Date") |
|---|
| 01/08/2018 |

By using Min  function the first successful ground landing date was identified.

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000



```
In [17]: %sql select Booster_Version from SPACEXTBL where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and
```

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

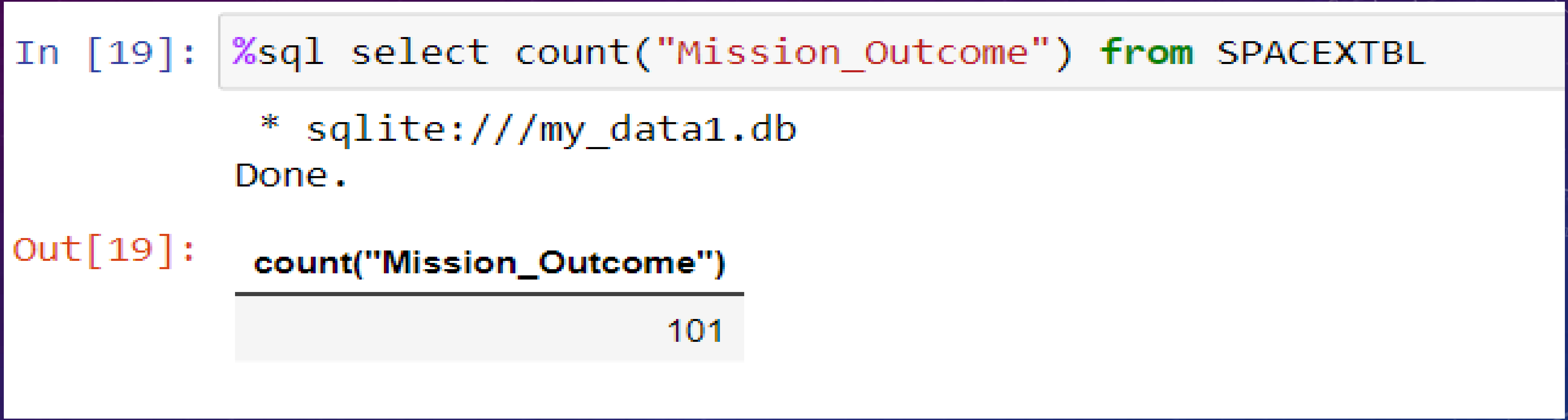
Out[17]:

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

By using between function the  successful drone ship landing with payload more than 4000kg and less than 6000kg were identified.

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

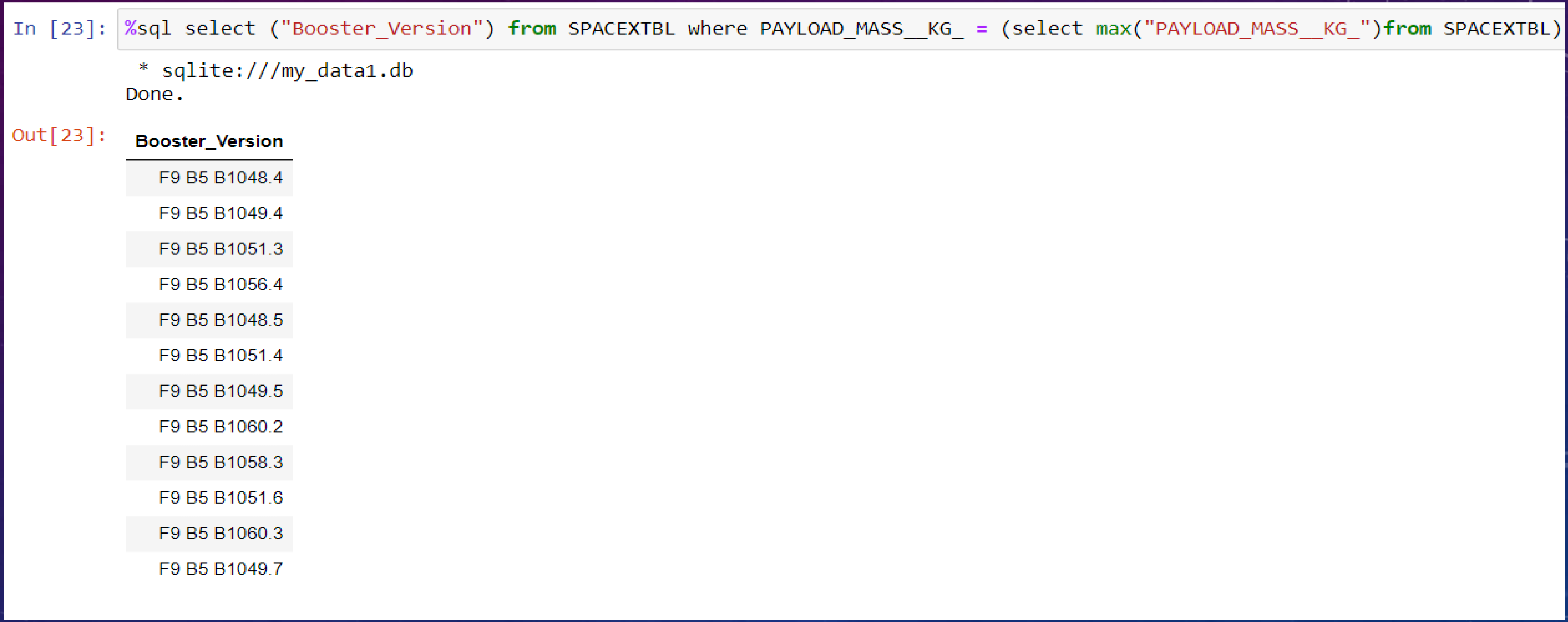


```
In [19]: %sql select count("Mission_Outcome") from SPACEXTBL

          * sqlite:///my_data1.db
         Done.

Out[19]:  count("Mission_Outcome")
         ---------------------------
                               101
```

By using count function the total number of successful and failure mission outcomes were found.

GitHub URL:
[https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git](https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git)

# BOOSTERS CARRIED MAXIMUM PAYLOAD



```
In [23]: %sql select ("Booster_Version") from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max("PAYLOAD_MASS__KG_")from SPACEXTBL)

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

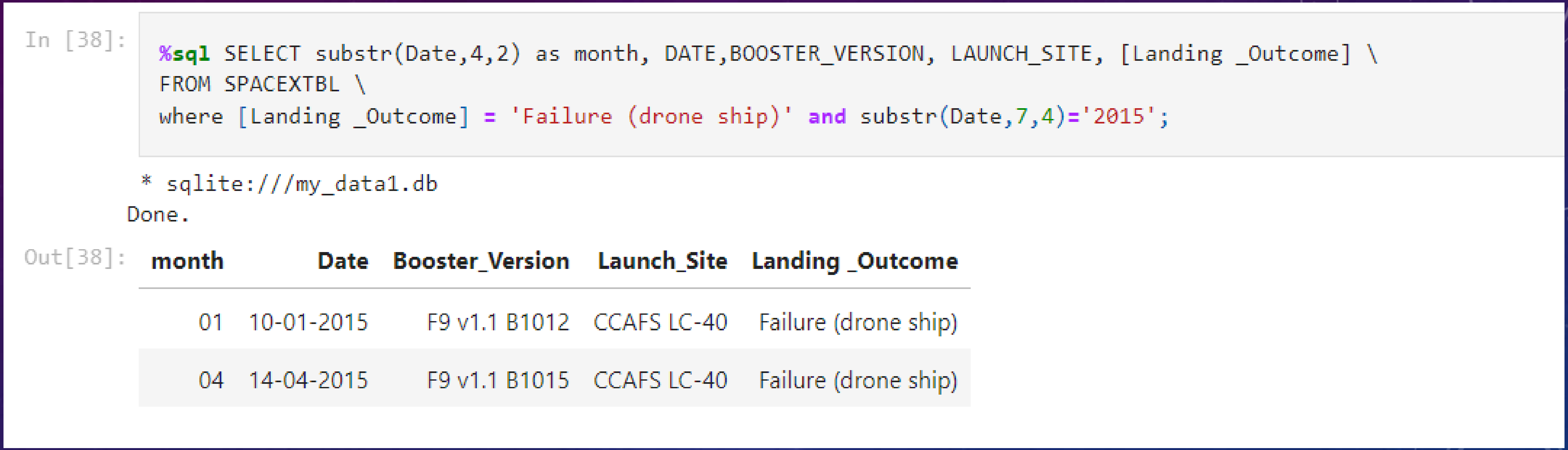
Out[23]:

| 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 |

By using subquery and max function boosters with maximum payloads were listed.

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# 2015 LAUNCH RECORDS



```
In [38]: %sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] \
         FROM SPACEXTBL \
         where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';

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

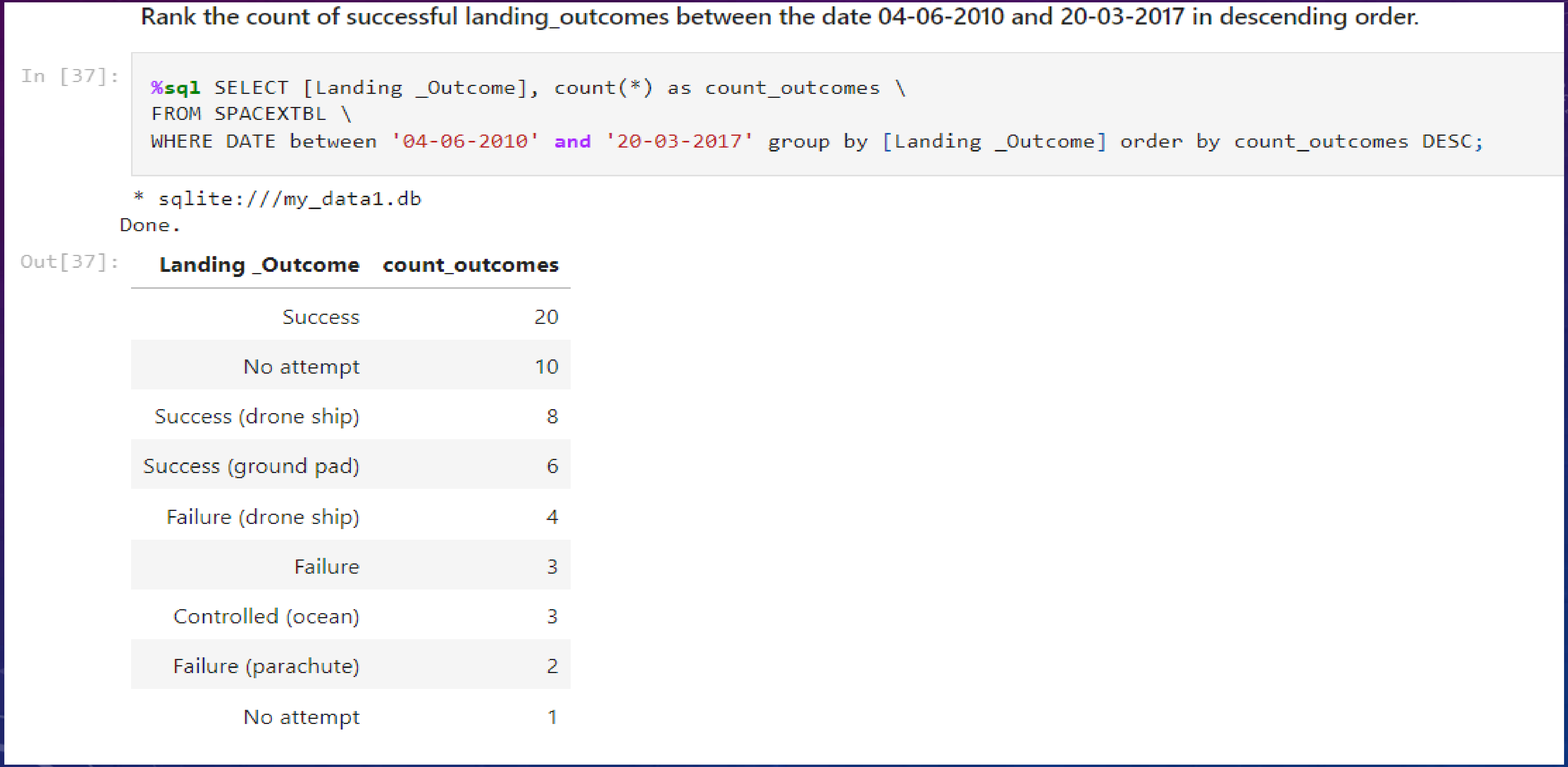
Out[38]:

| month | Date | Booster_Version | Launch_Site | Landing _Outcome |
|---|---|---|---|---|
| 01 | 10-01-2015 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 04 | 14-04-2015 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

By using special date format displayed main information regarding failure drone ship in 2015

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20



Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

In [37]:

```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

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

Out[37]:

| Landing _Outcome | count_outcomes |
|---|---|
| Success | 20 |
| No attempt | 10 |
| Success (drone ship) | 8 |
| Success (ground pad) | 6 |
| Failure (drone ship) | 4 |
| Failure | 3 |
| Controlled (ocean) | 3 |
| Failure (parachute) | 2 |
| No attempt | 1 |

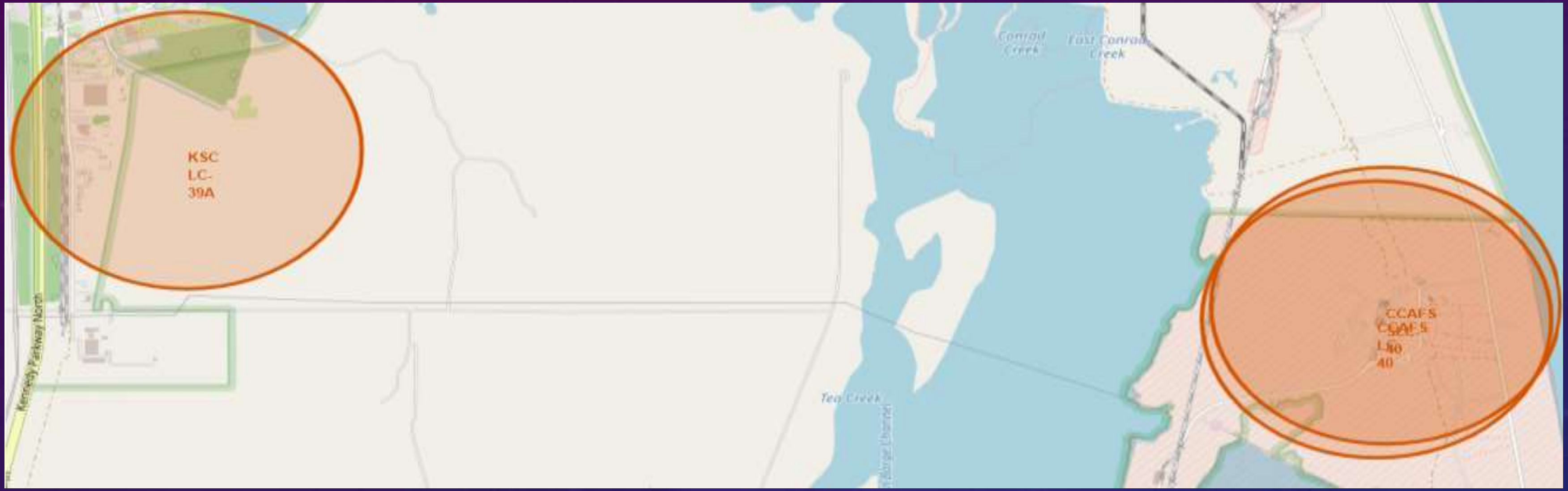
By using count function listed landing outcomes between 04/06/2010 and 20/03/2017 in descending order.

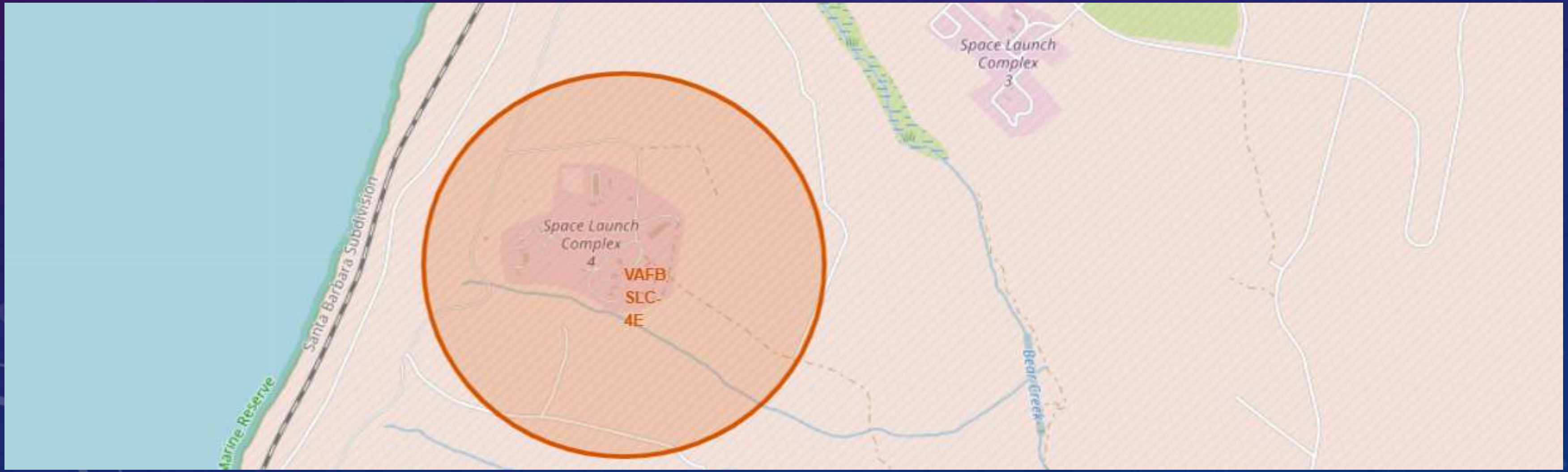
GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

Section 3

# Launch Sites Proximities Analysis

# LAUNCH SITES LOCATIONS ON THE MAP

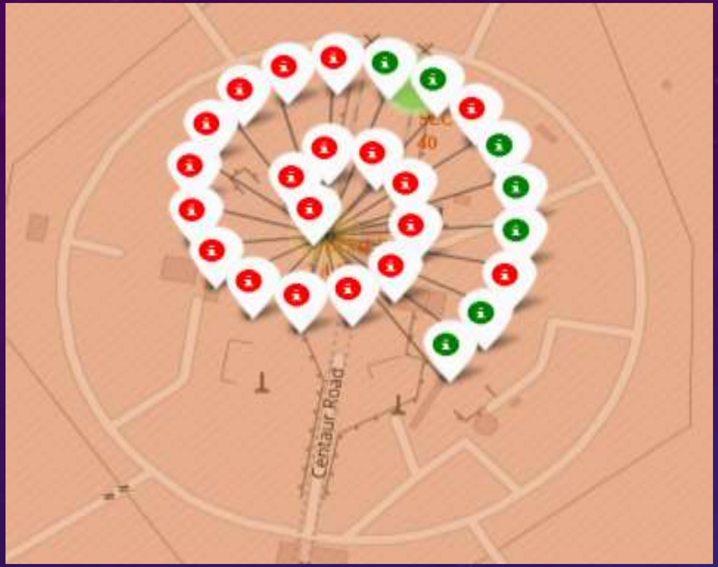


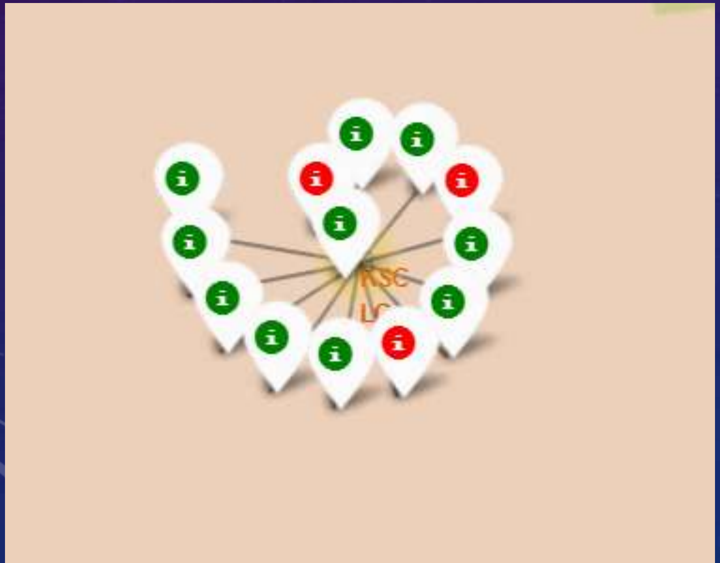


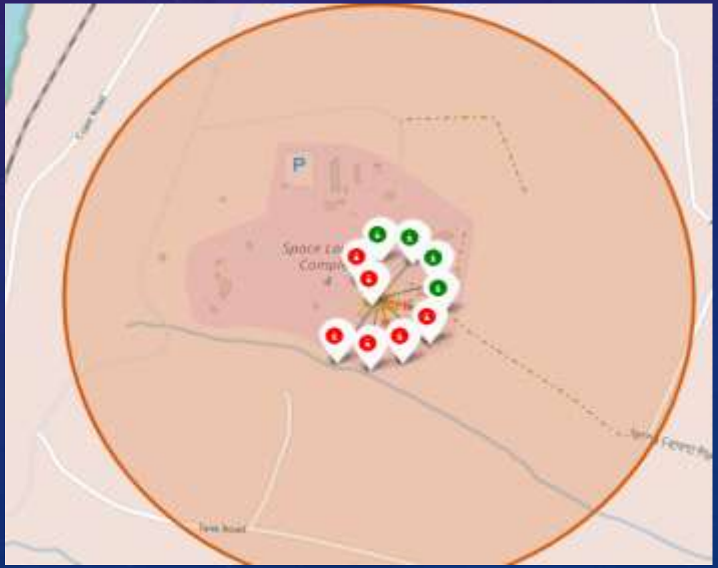
By creating  Folium map markers and circles displayed 4 launch site locations on the map. It can be observed that 3 launch sites are located very close to each other and 2 of them overlapping.

GitHub URL: [https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git](https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git)

# LAUNCH OUTCOMES







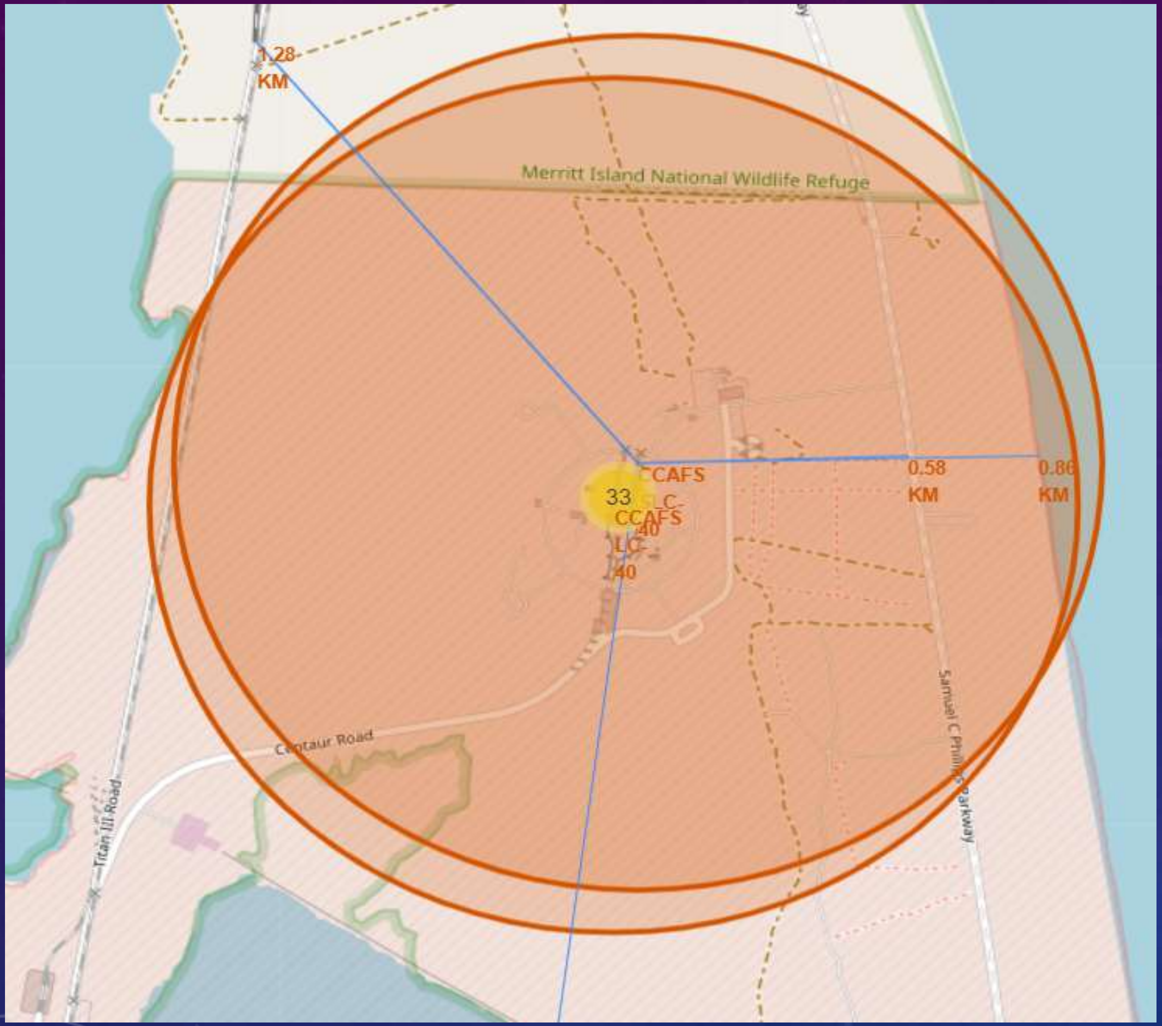
Green – Successful landing
Red – Unsuccessful landing

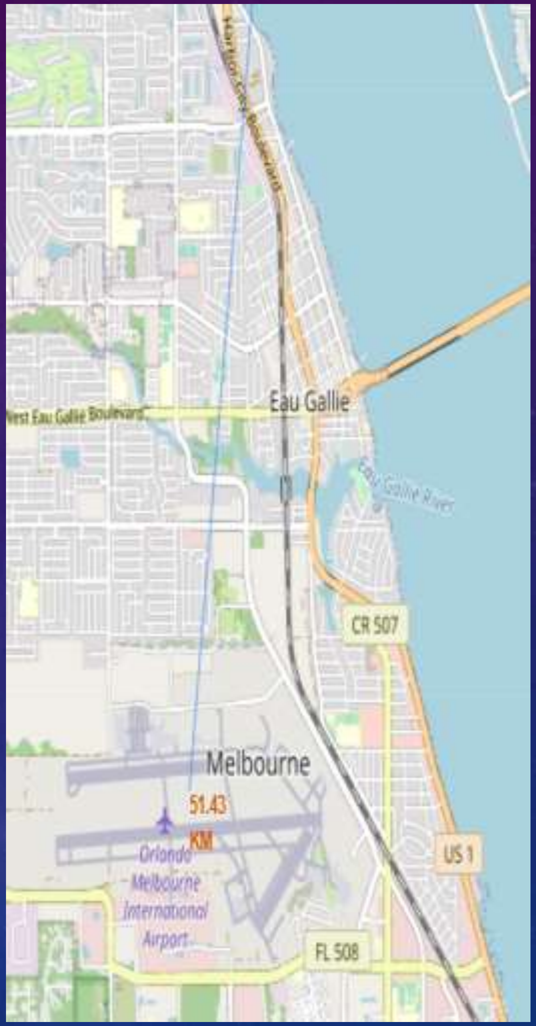
KSC-LC-39A: highest success rate

CCAFS-LC-40: largest number of launches, but most of them were unsuccessful

GitHub URL: https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

# DISTANCES FROM CCAFS-SLC-40 LAUNCH SITE TO ITS PROXIMITIES


1.28 KM
Merritt Island National Wildlife Refuge
CCAFS SLC-40
33
0.58 KM
0.86 KM
Centaur Road
Titan III Road
Samuel C Phillips Parkway


West Eau Gallie Boulevard
Eau Gallie
Eau Gallie River
CR 507
Melbourne
51.43 KM
Orlando Melbourne International Airport
US 1
FL 508

Distances:

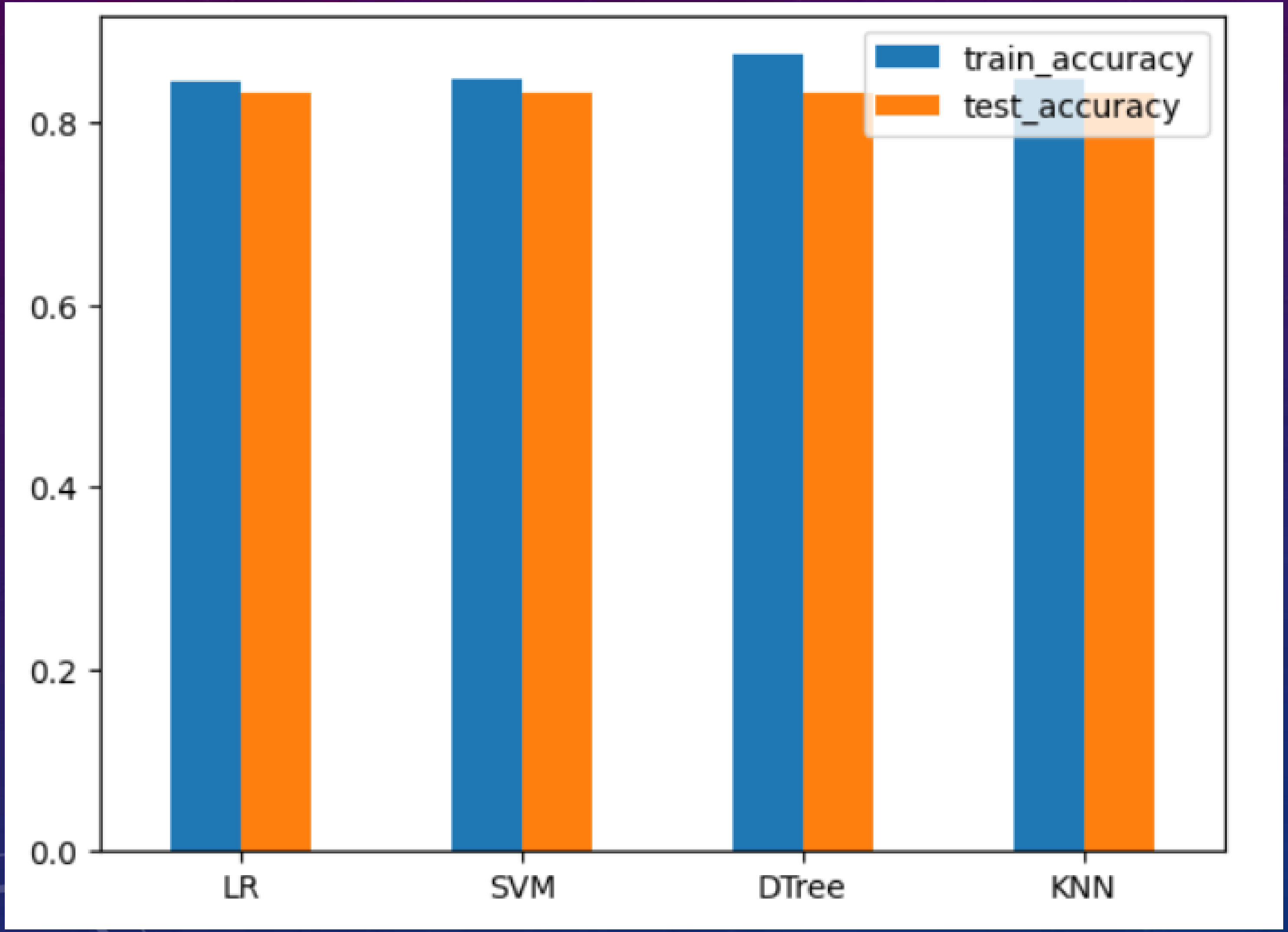
To the coastline: 0.86 km
To the highway: 0.58 km
To the railroad: 1.28 km
To the city: 51.43 km

GitHub URL:
https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git

Section 5

# Predictive Analysis (Classification)

# CLASSIFICATION ACCURACY



Test_data accuracy for all models is the same: 0.8333

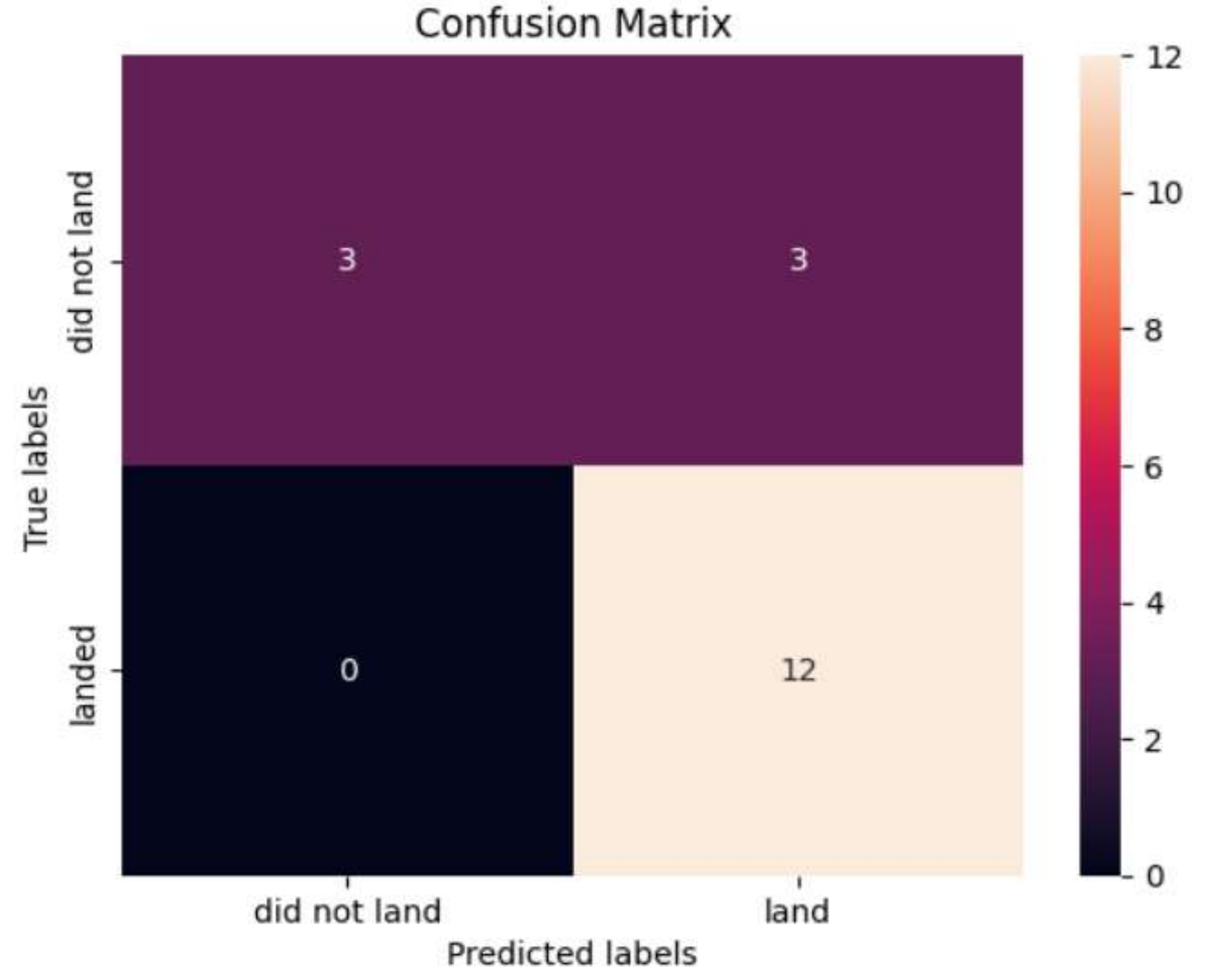
Train_data accuracy:

1) Decision Tree: 0.875
2) Support Vector Machine: 0.8482
3) K nearest neighbour: 0.8482
4) Linear Regression: 0.8464

As can be seen train data accuracies almost the same within the models. Only Decision Tree model has a slight higher score.

GitHub URL: [https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git](https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git)

# CONFUSION MATRIX



Confusion matrixes of all models revealed the same, as test_data accuracy was the same.

It is seen that models can distinguish between the different classes quite good. The major problem is false positives, where models predicted 3 launches as successful landings, in fact they did not land.

GitHub URL: [https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git](https://github.com/TalgatAzhibekov/IBM_DS_Capstone.git)

# CONCLUSIONS

It was a quite long journey which included all main stages of a data science project.

From EDA with visualization it was revealed that landing success increased with the number of flights. Overall success rate kept increasing from 2013 till 2020 with slight decrease in 2018.

From Folium maps activities the launching site locations hidden peculiarities discovered. So that they are located close to the equator line, coastlines and railroads. On the other hand, far away from big cities.

It is interesting to note that all four machine learning models (LR, SVM, DTree, KNN) showed exactly the same accuracy on test data 83.33%. While on train data the decision tree model calculated the best score 87.5% with minor difference from other models.

In conclusion, being able to predict the landing outcome of the Falcon 9's first stage with 83.33% accuracy gives enough confidence to be more prepared for future launches.

Thank you!