Rocket Science

with Data Science

IBM Developer
SKILLS NETWORK

Flávio Dias - March 19, 2023

OUTLINE

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

EXECUTIVE SUMMARY

Given the complexity of predicting the successful landing of the rocket's first stage with mathematical models, a more approachable way is to analyze data from the past launches, taking insights and predicting the outcome based on Machine Learning Models.

Achieving 83.3% accuracy on the prediction model's results, we are able to determine a set of conditions that can improve the likelyhood of a good first stage landing.

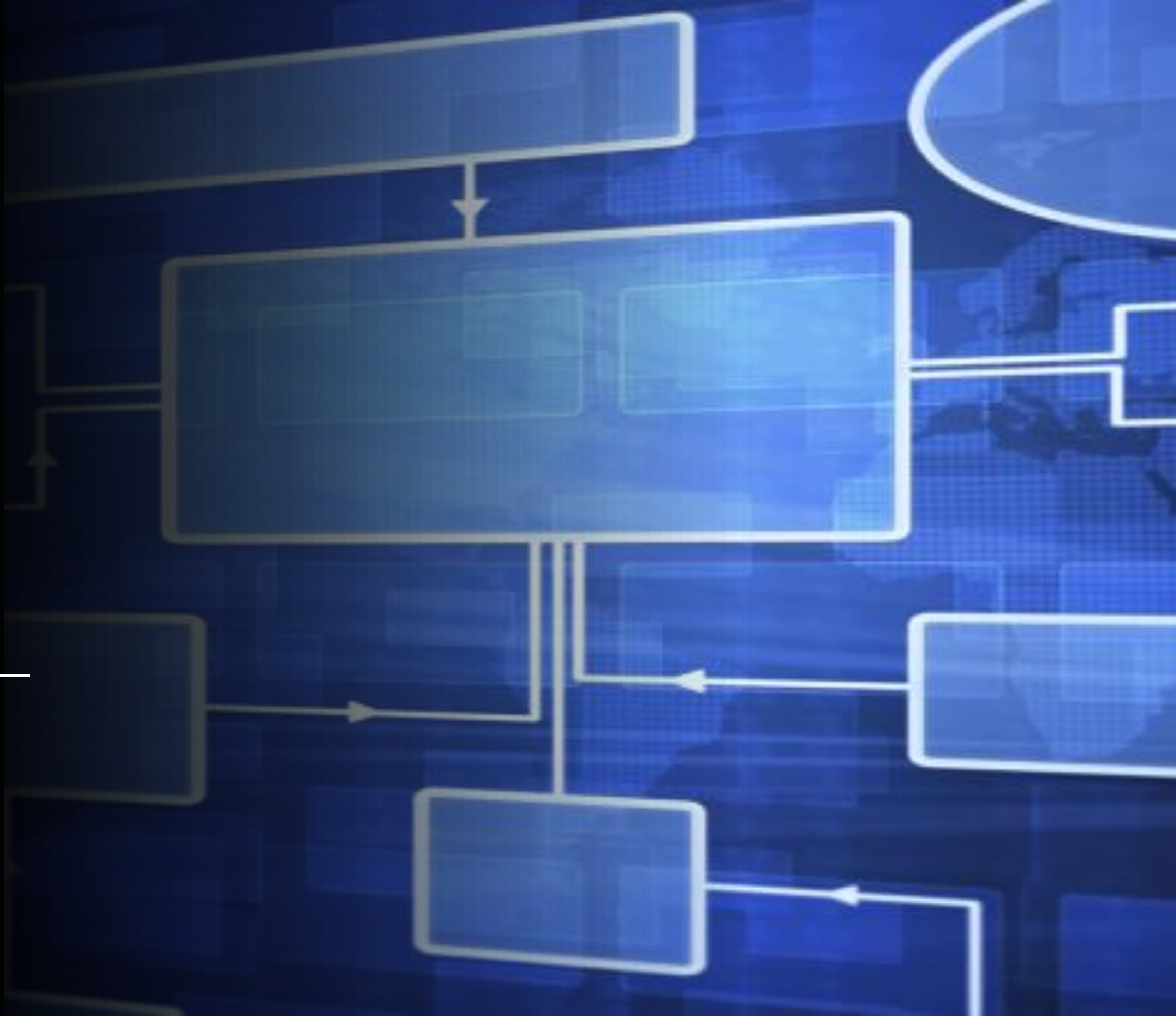
INTRODUCTION

Space Y company, competing directly with SpaceX, is planning to service space transportation in the near future. SpaceX managed to save about 100 million dollars by reutilizing the first stage of their rockets and, in order to be competitive, Space Y needs to follow up this challenge.

Hence, we are set to predict wheather the first stage will land succssesfully or not, by using the SpaceX public date to feed machine learning models.

In adition to the models, Exploratory Data Analysis is performed to create valuable insights, aiming to answer questions like the following : What Orbit has a higher success rate? Which Lounch Site have better success rate? What Booster Version do better carrying specific payloads?

METHODOLOGY



METHODOLOGY MAIN SCOPE

DATA

Data Collection of the SpaceX public Data via API.

Perform Data Wrangling to clean and preprocess the data.

ANALYSIS

Perform EDA using visualization and SQL.

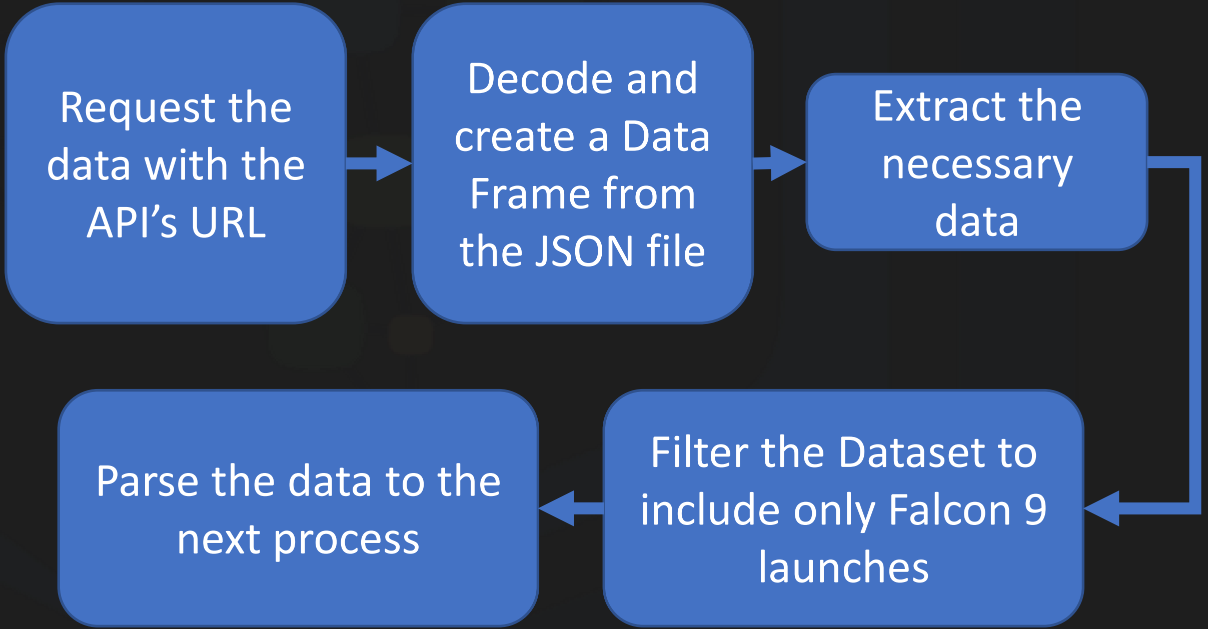
Perform Visual Analytics using Folium and Plotly Dash.

Predictions

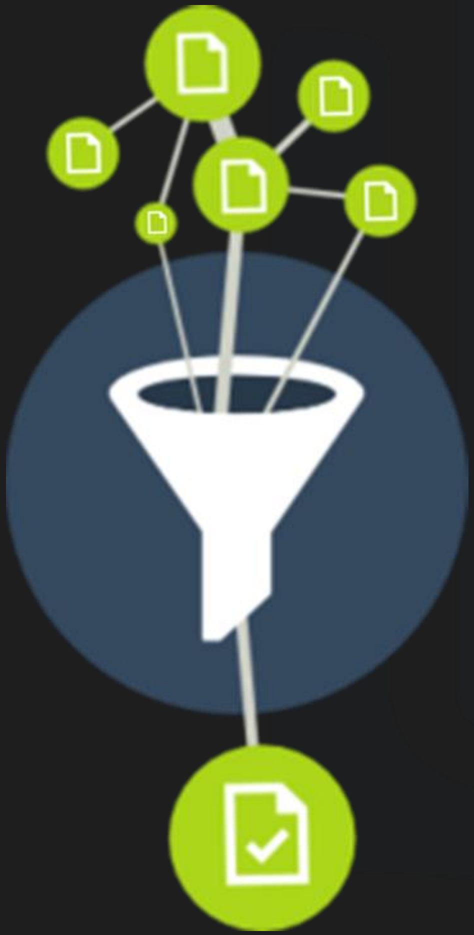
Perform predictive analysis using classification models.

DATA COLLECTION

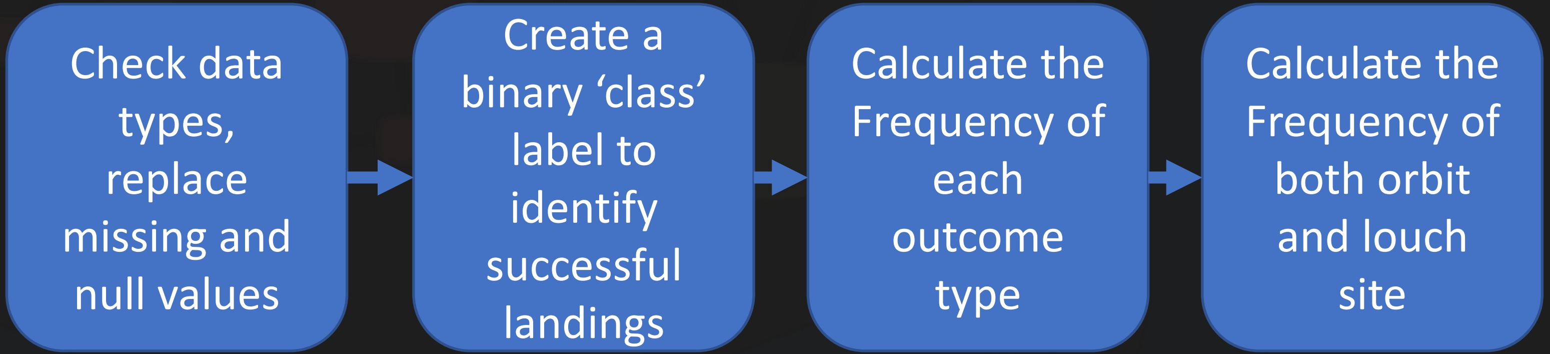
The flowchart explains the steps of the entire process of the data collection:


Request the data with the API's URL
Decode and create a Data Frame from the JSON file
Extract the necessary data
Filter the Dataset to include only Falcon 9 launches
Parse the data to the next process

DATA WRANGLING



The flowchart explains the steps of the entire process of the data wrangling:



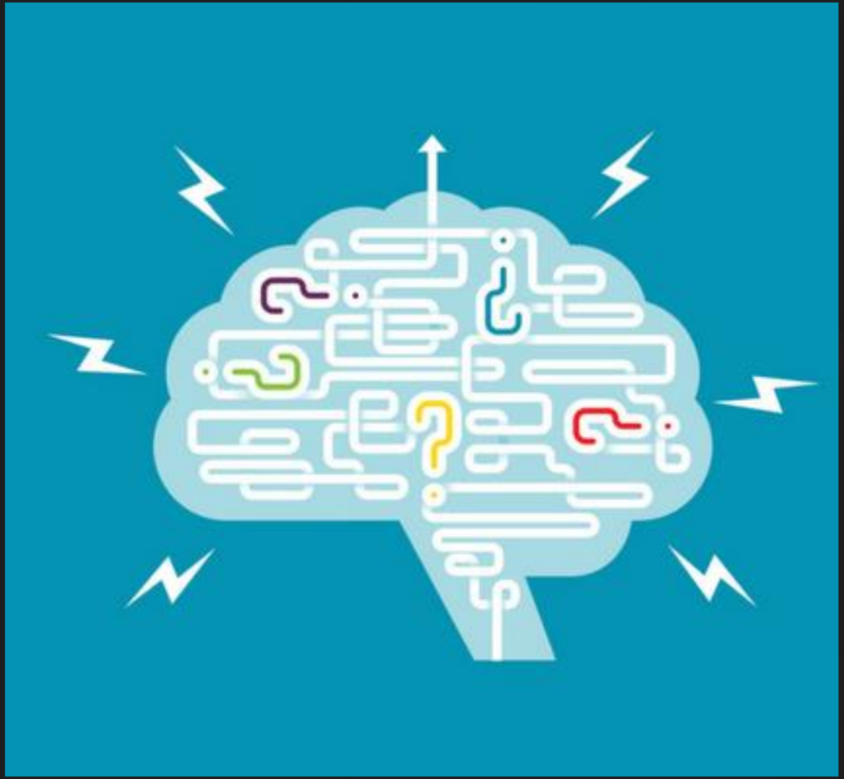
Check data types, replace missing and null values → Create a binary ‘class’ label to identify successful landings → Calculate the Frequency of each outcome type → Calculate the Frequency of both orbit and louch site

DATA VISUALIZATION

EDA were performed through charts like scatter, bar and line plots we are able to identify the relationship between variables and its landing success via a color indicator, providing a better understanding not only of the data but also the factors that influence most on the landing outcome.

In addition, an interactive map and dashboard were created, providing a more flexible analysis.

PREDICTIVE ANALYSIS



After preprocessing the data to train the models and normalizing the predictor variables and separaring the target values, the data spliting were performed to separate the train and test sets.

Then different models were trained, such as logistic regression, SVM, decision tree and KNN.

And an evaluation was made to check the performance of each of the models.

RESULTS

EDA – EXPLORATORY DATA ANALYSIS

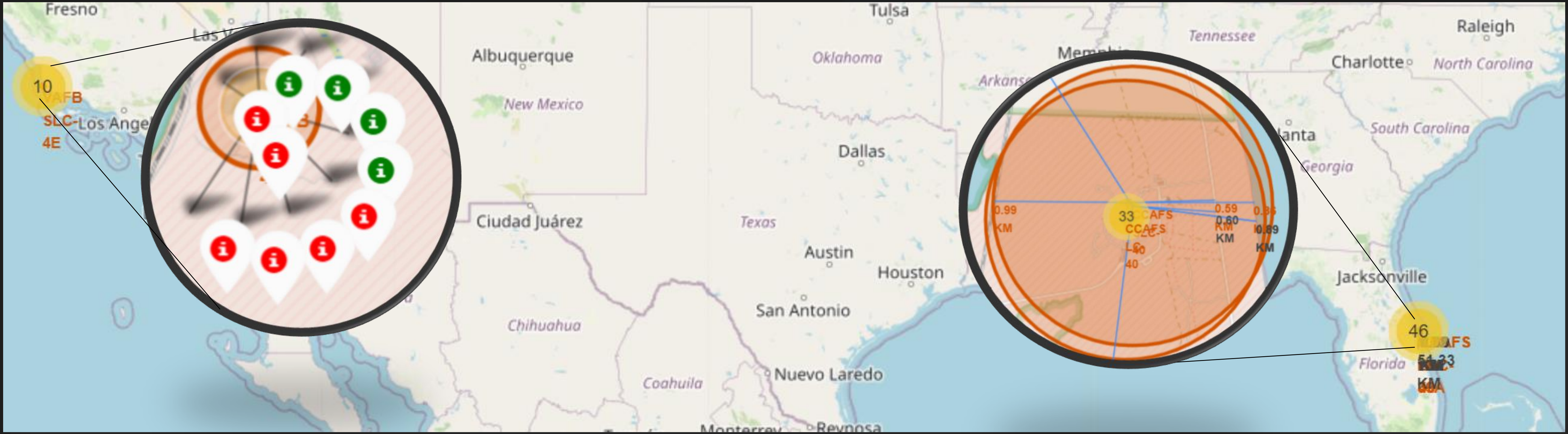
Many valuable insights were found by performing the EDA. Some are the success rate over the years, the boosters that are more likely to succeed and the success rate related with the orbits.

The main results are:

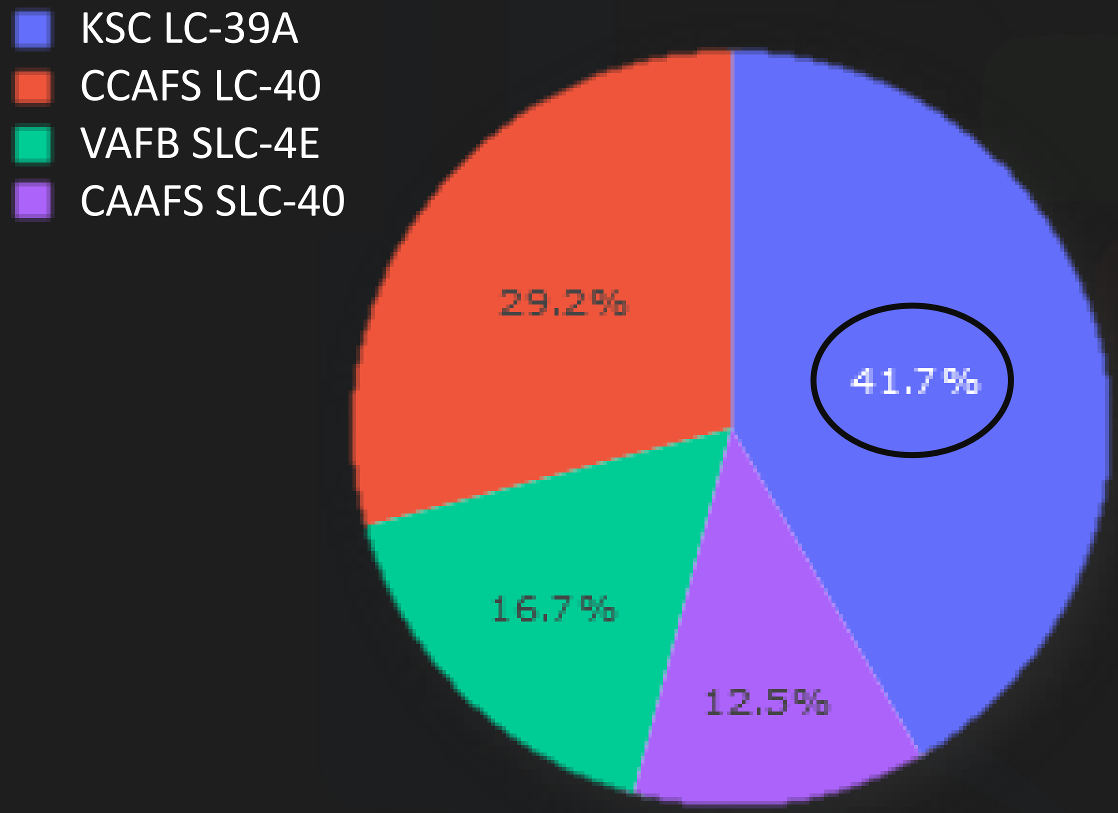
- Space X uses 4 launch sites in the US
- The average payload mass carried by booster F9 v1.1 is 2534 kg
- The first successful landing outcome in ground pad was achieved in 2017
- Almost 100% of the mission outcomes were successful

MAP WITH FOLIUM

Markers, clusters, circles and lines where drown on top of an interactive map, using the coordinates of each sites. The lines marks the distances between reference points, such as railways, highways and the coastside distance, allowing us to see that the sites are located near the sea with good logistic infrastructure around.



PLOTLY DASHBOARD

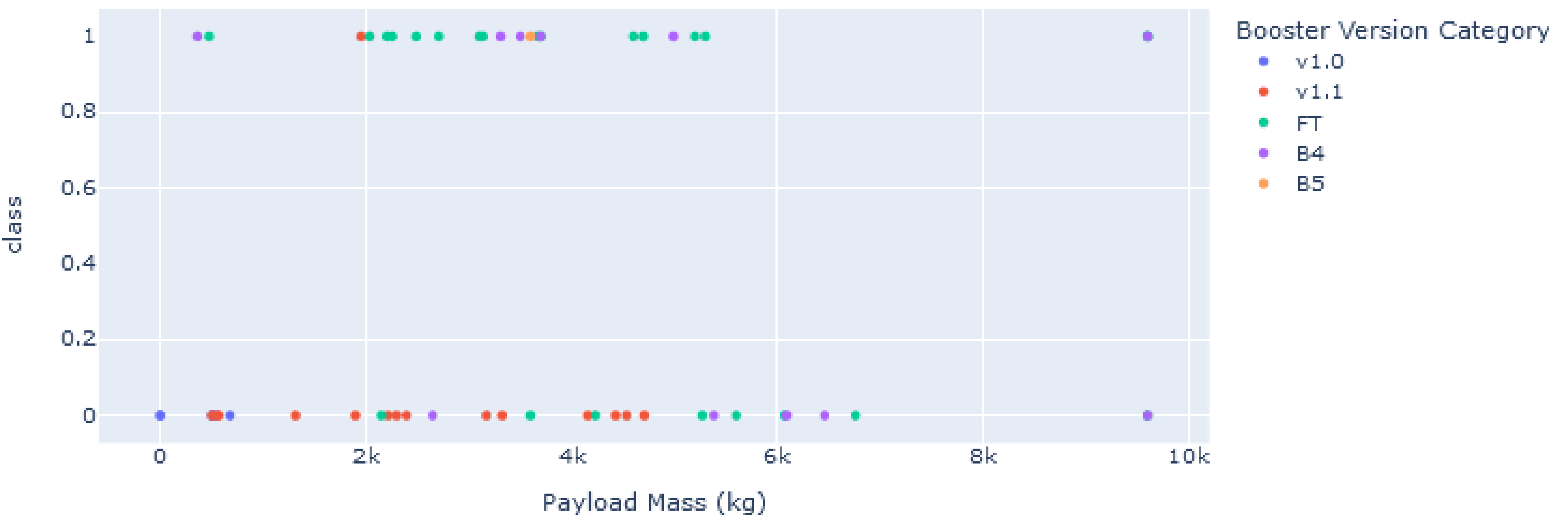


From the interactive dashboard, while performing the EDA, we found that 41.7% of the landings at KSC LC-39A were successful. That represents 42% more likely to have a successful landing at CCAFS LC-40 site.

But this can be deceiving, because the latest flights had a higher success rate than the earliest, so, deeper investigation is needed to securely conclude that.

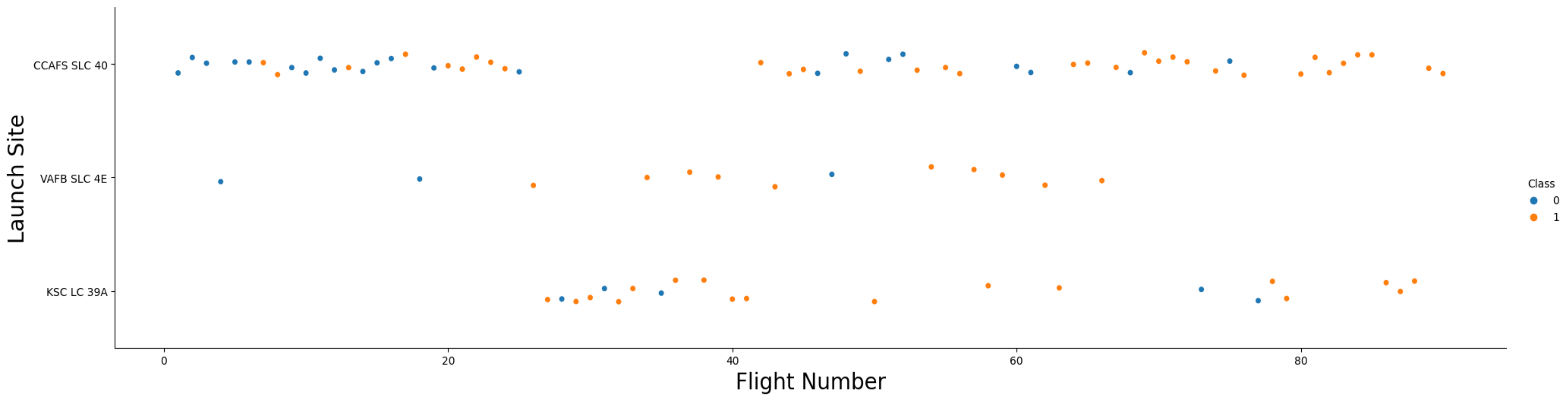
PLOTLY DASHBOARD

Its also clear the success rate of the FT booster while in the range of 2-6k kg payload mass. Showing that it can be a good candidate in those situations for our Space Y company. B4 booster has around 50% success when operating with more than 9k kg.



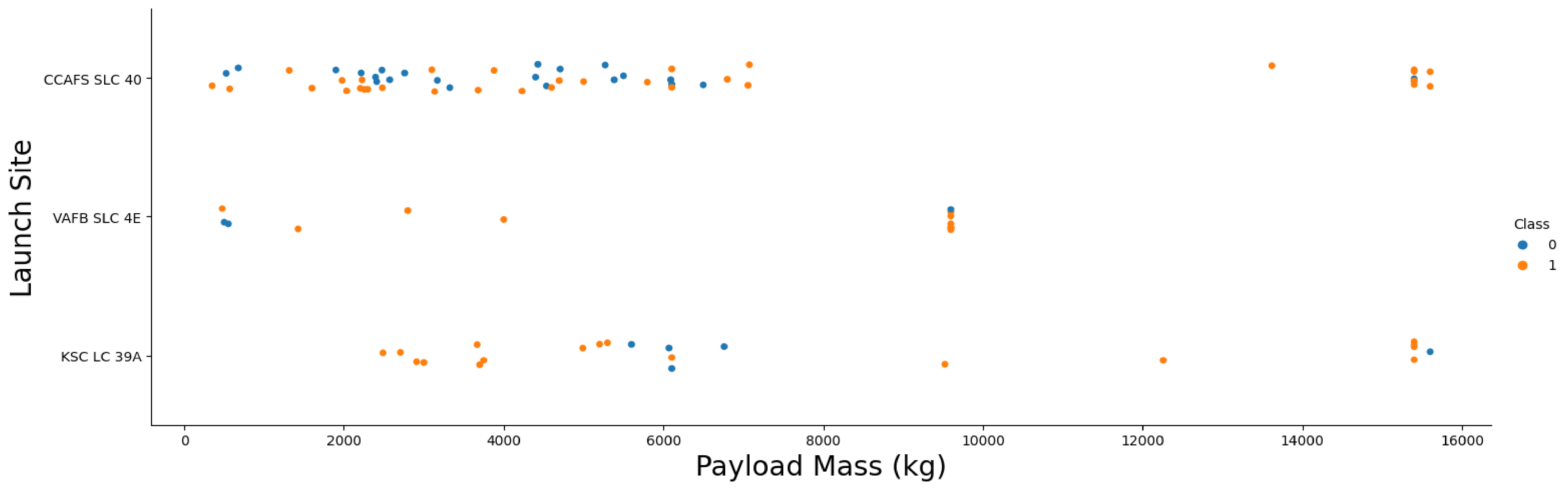
EDA – FLIGHT NUMBER vs LAUNCH SITE

The graph shows that the main launch site is CCAFS SLC-40 and since flight number 20 the success rate has been increasing on all launch sites, explaining the pie chart observations.



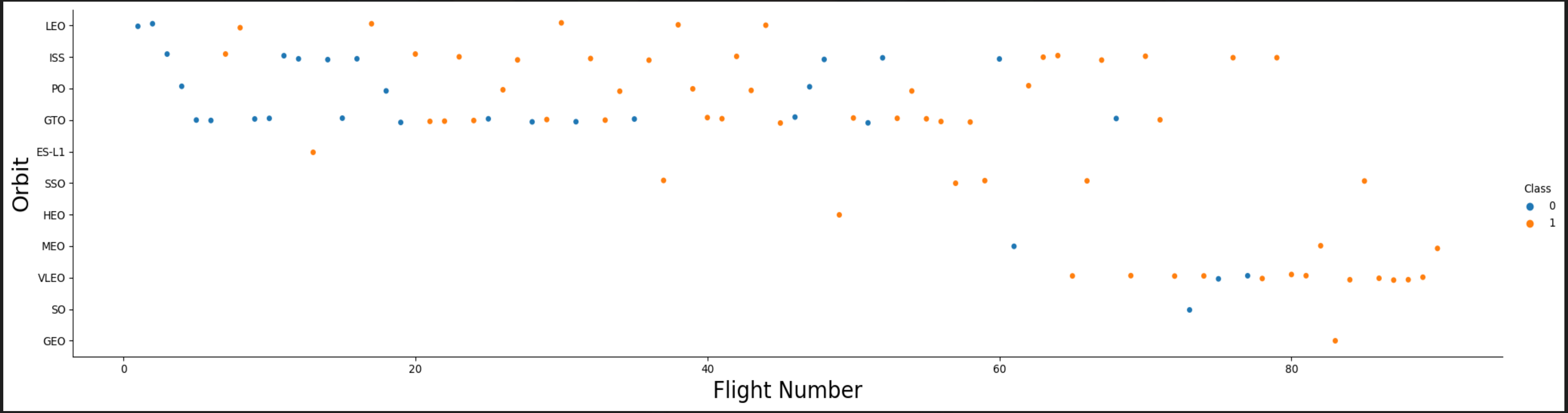
EDA – PAYLOAD vs LAUNCH SITE

The graph shows that the higher the payload mass the higher the success rate and some launch sites have a better success rate for specific payloads. KSC LC 39A has high success rate when carrying between 2-6k kg of payload, while for CCAFS SLC 40 has its high around 15-16k kg.

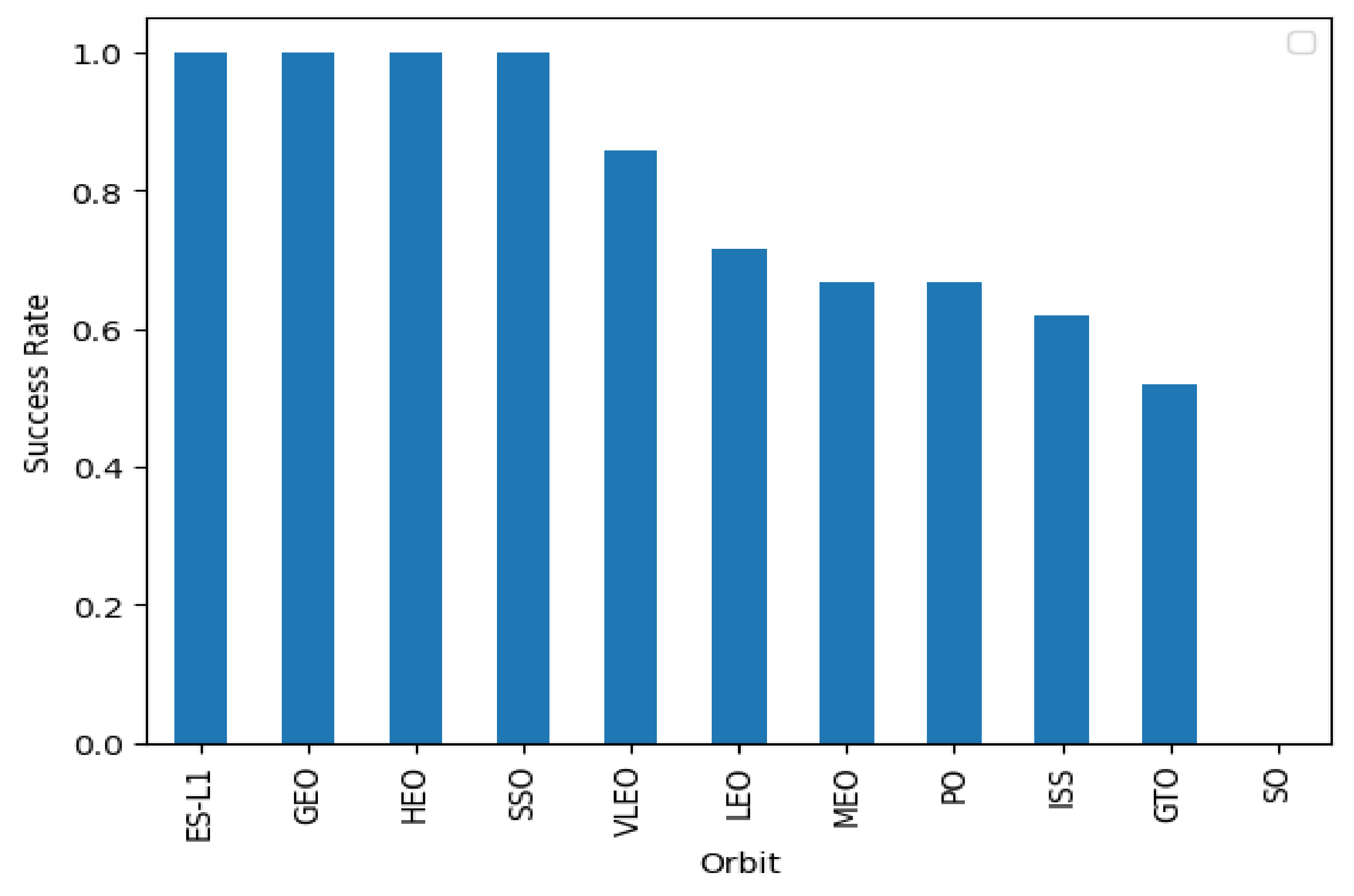


EDA – FLIGHT NUMBER vs ORBIT

Here we can see that there may be a relationship between LEO and flight number, as success increases with the flight number. It’s valid to mention that orbits with highest success rates have fewer missions than the other orbits, like VLEO and HEO.



EDA – SUCCESS RATE PER ORBIT

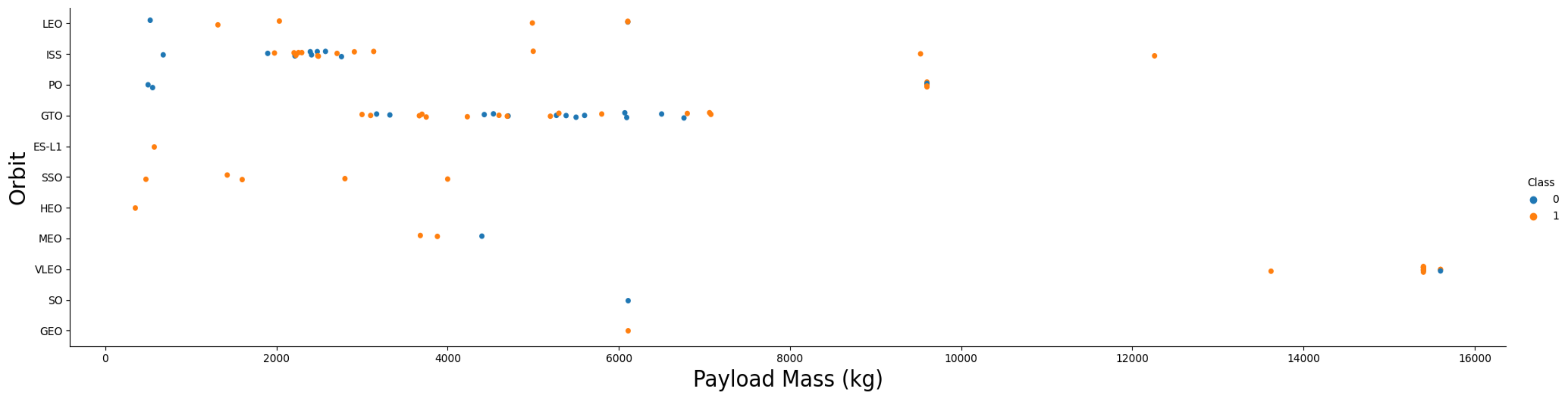


The graph shows the boosters on missions to orbits such as ES-L1, GEO, HEO and SSO are guaranteed to land back successfully. Meanwhile GTO has success rate at about 50% and SO orbit had no previous success.

Though, from the last graph, it was shown that those “guaranteed success” sites have fewer launches.

EDA – PAYLOAD vs ORBIT

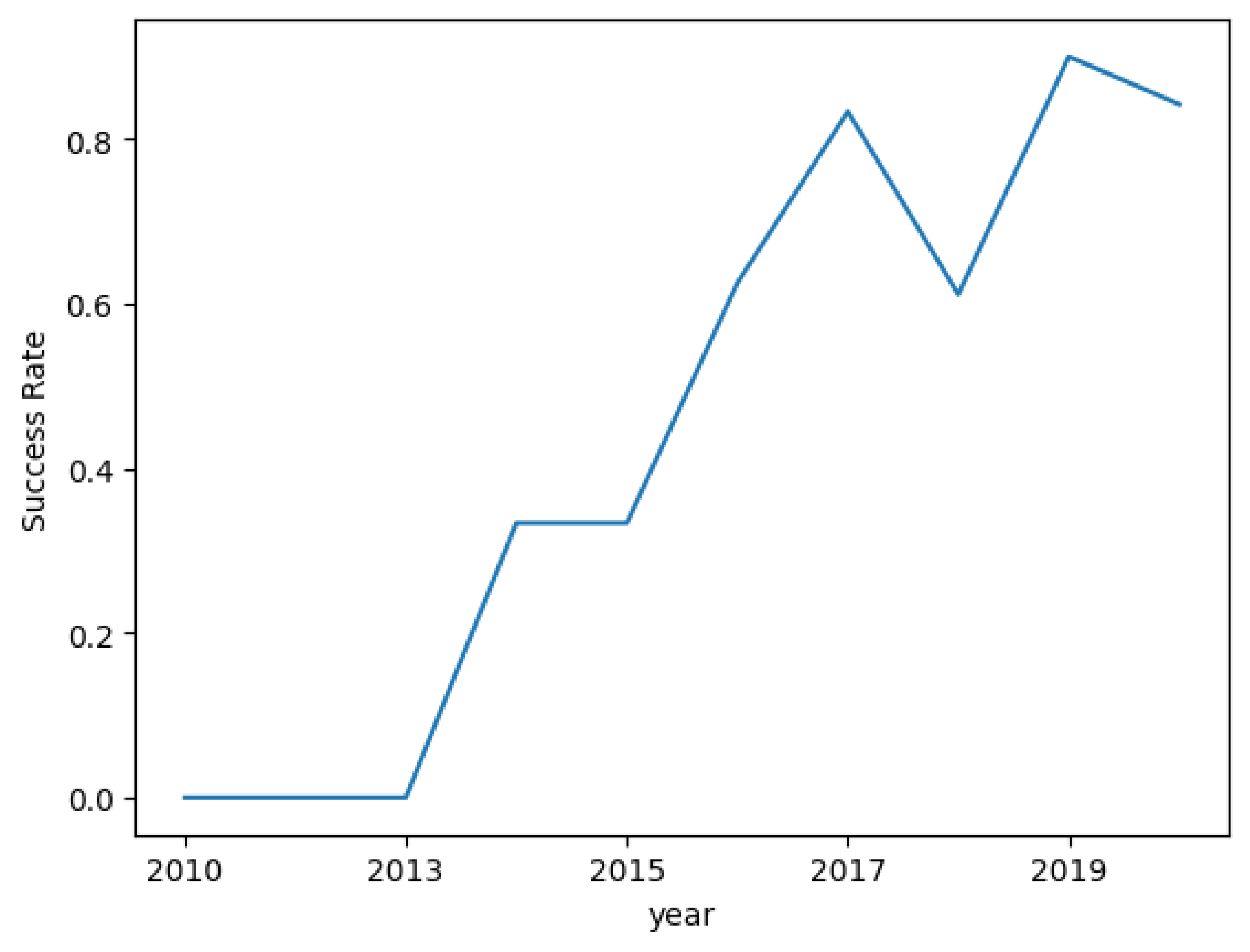
Data shows that heavier payloads seems to achieve better outcomes for orbits like VLEO, LEO, ISS and PO. On the other hand, orbits such as ES-L1, SSO, GEO and HEO seem to have a higher success rate for lighter payloads, since those don’t have launches with payloads past 8000 kg.



IBM Developer

SKILLS NETWORK

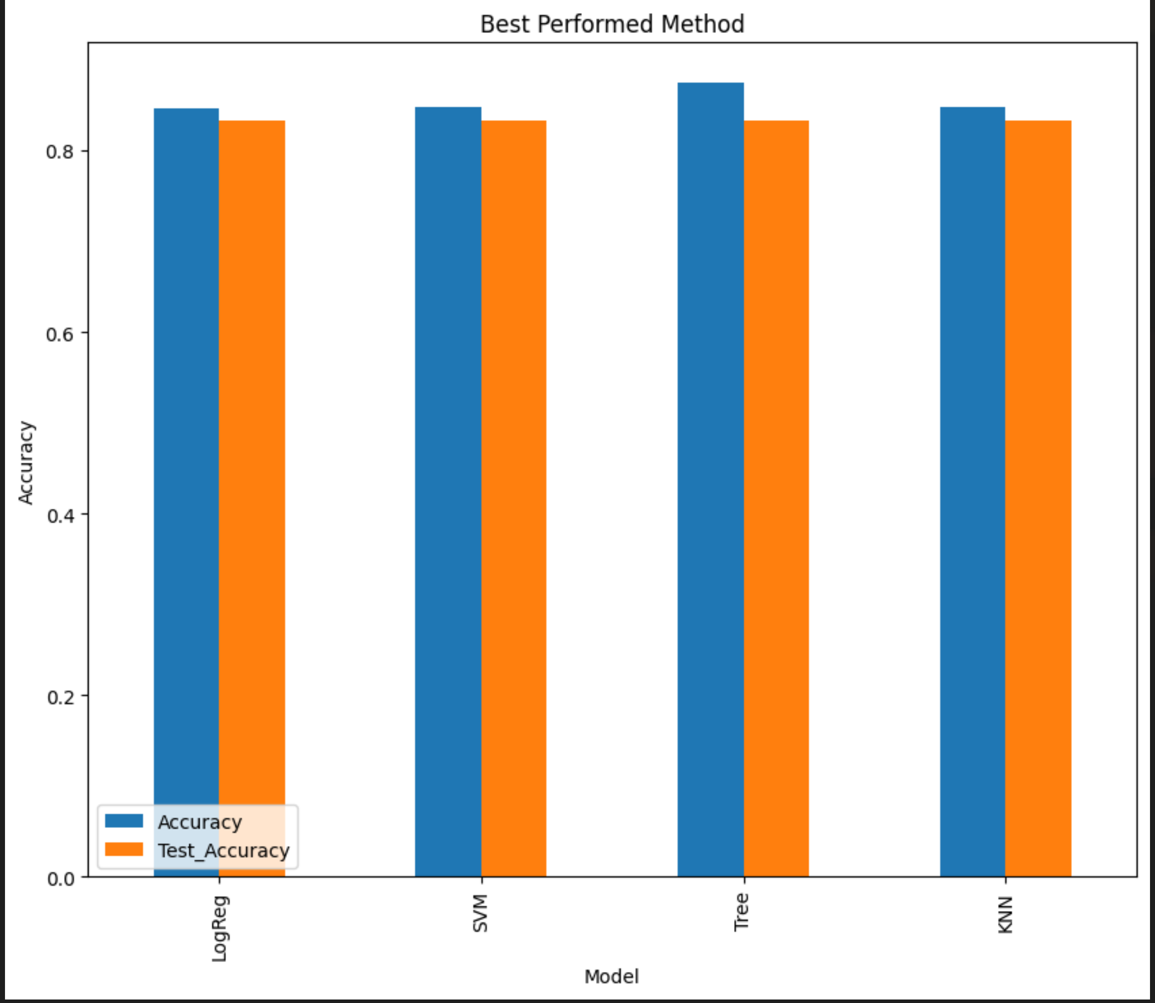
EDA - EVOLUTION OF SUCCESS RATE



Success rate has been increasing over the years, with its peak at about 90% in 2019.

Although, from 2014 to 2015 did not increase and from 2017 to 2018 and from 2019 to 2020 the success rate did decreased.

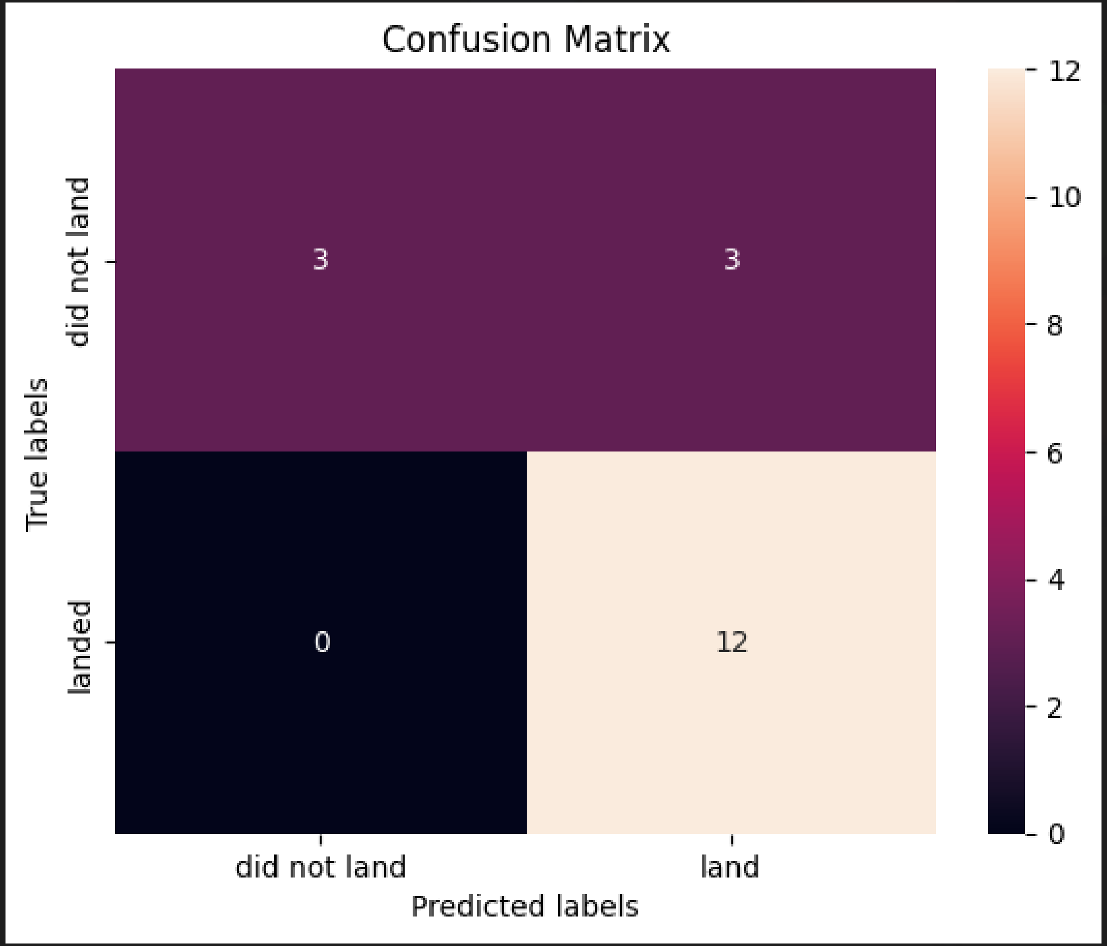
PREDICTIVE ANALYSIS - CLASSIFICATION



Based on the training accuracy score, the decision tree performed slightly better than the other models, with 87.5% accuracy.

All classification models were found to perform equally (83.3% accuracy over the test set), which can be explained by the size of the dataset being small (understandable due the nature of the events). And to choose the best model, other benefits were considered, such as the estimated risk and the score on the training set.

PREDICTIVE ANALYSIS – CONFUSION MATRIX



For all four models, the same confusion matrix were obtained, which is expected given the known equal accuracy of the models.

It can be seen that the models have good chance to predict the successful landing. Where we've got 12 true positives, 3 true negatives, 3 false positives and 0 false negative predictions.

CONCLUSION

KSC LC-39A have the most successful launch rate.

Payloads over 9000 kg have higher success rate.

Successful landing outcome improved over time.

Machine learning models have a prediction accuracy of 83.3%.

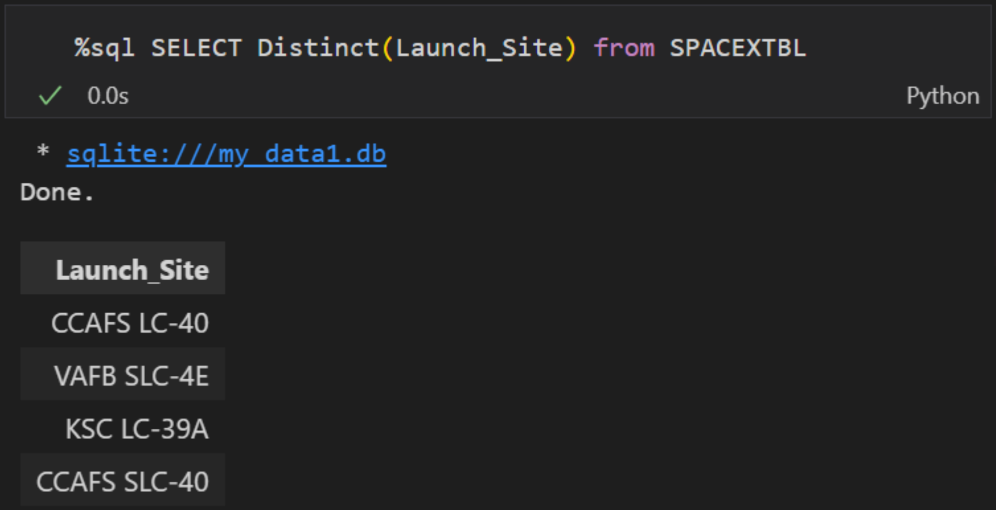
The models can be used to predict with relative high accuracy whether the first stage of the rocket will land successfully.

More data is needed to improve the model’s accuracy, enabling to select one model that stands out from the others.

APPENDIX – EDA CODES

Task 1

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



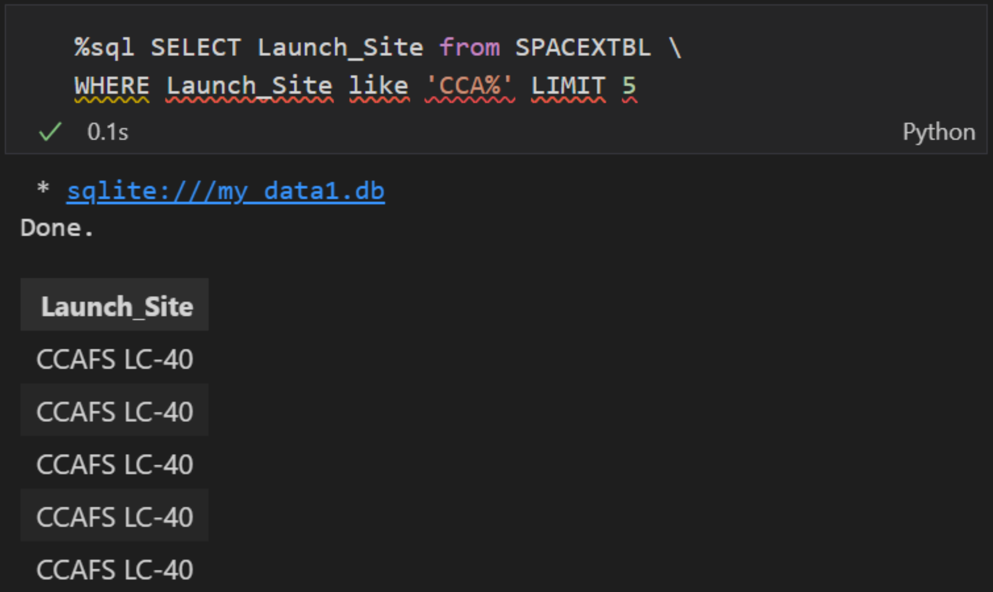
```
%sql SELECT Distinct(Launch_Site) from SPACEXTBL
```

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

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

Task 2

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

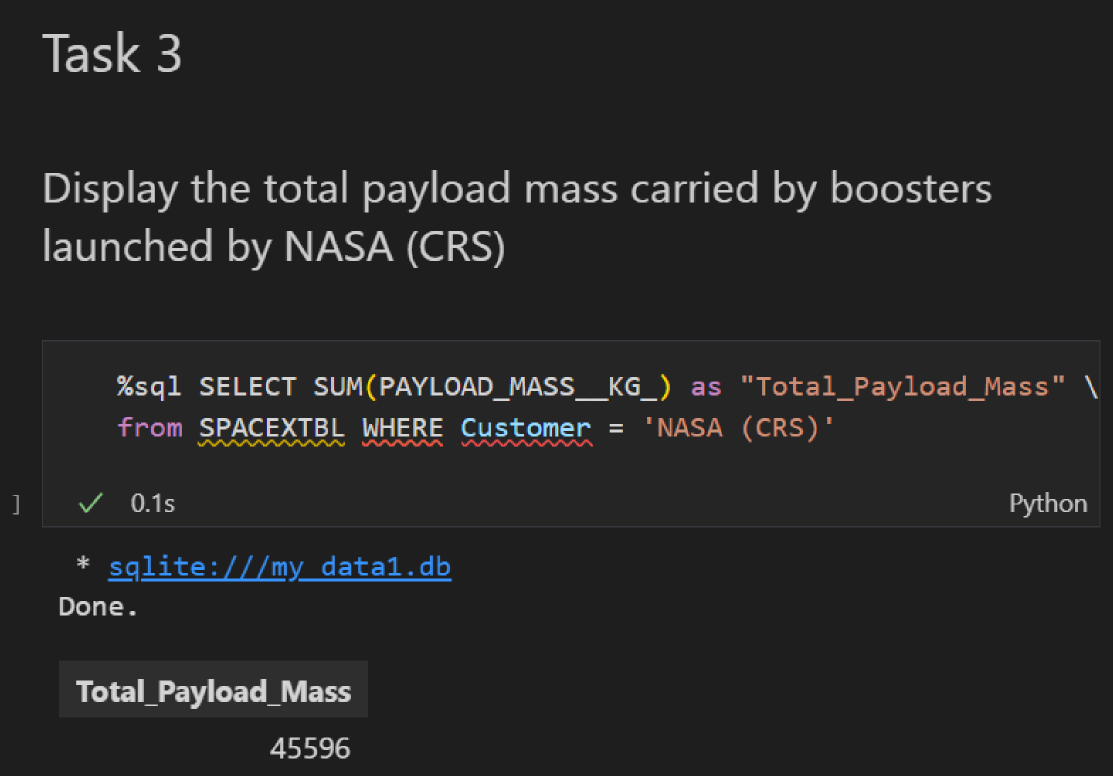


```
%sql SELECT Launch_Site from SPACEXTBL \
WHERE Launch_Site like 'CCA%' LIMIT 5
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

APPENDIX – EDA CODES



Task 3

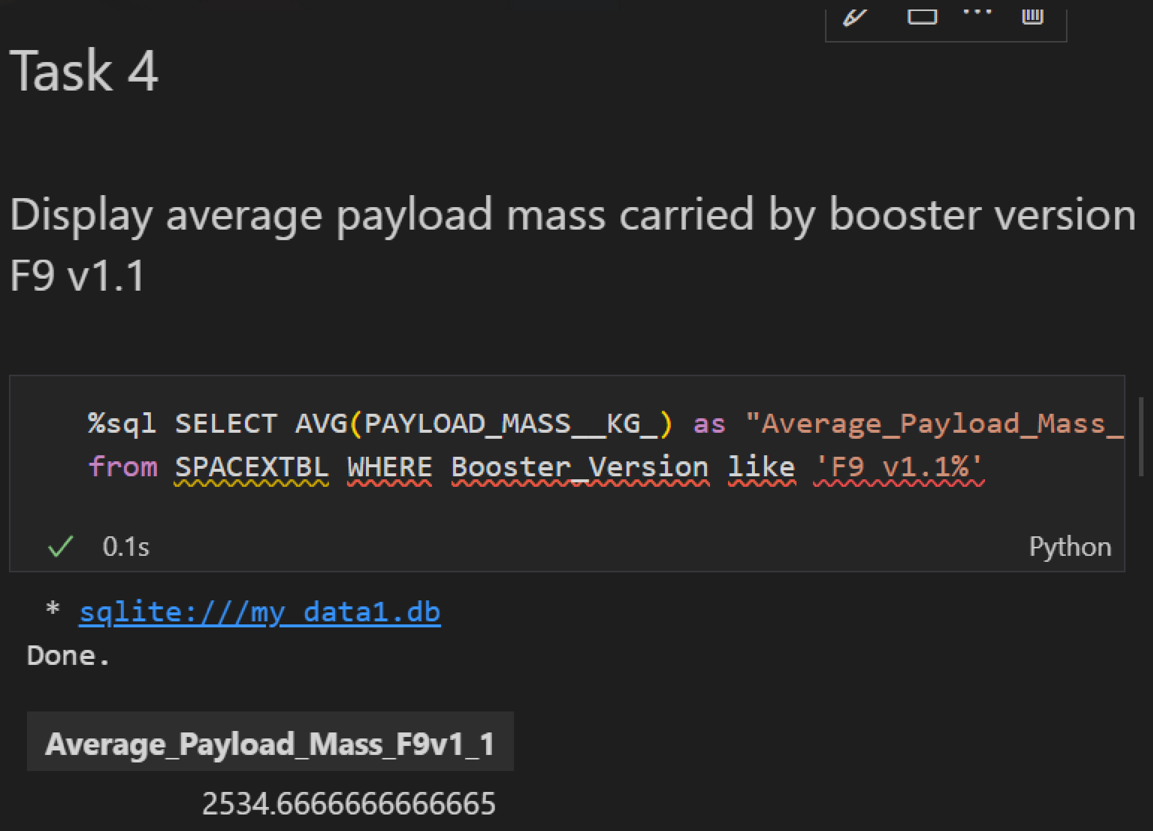
Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total_Payload_Mass" \
from SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

✓ 0.1s Python

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

Total_Payload_Mass
45596



Task 4

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as "Average_Payload_Mass_
from SPACEXTBL WHERE Booster_Version like 'F9 v1.1%'
```

✓ 0.1s Python

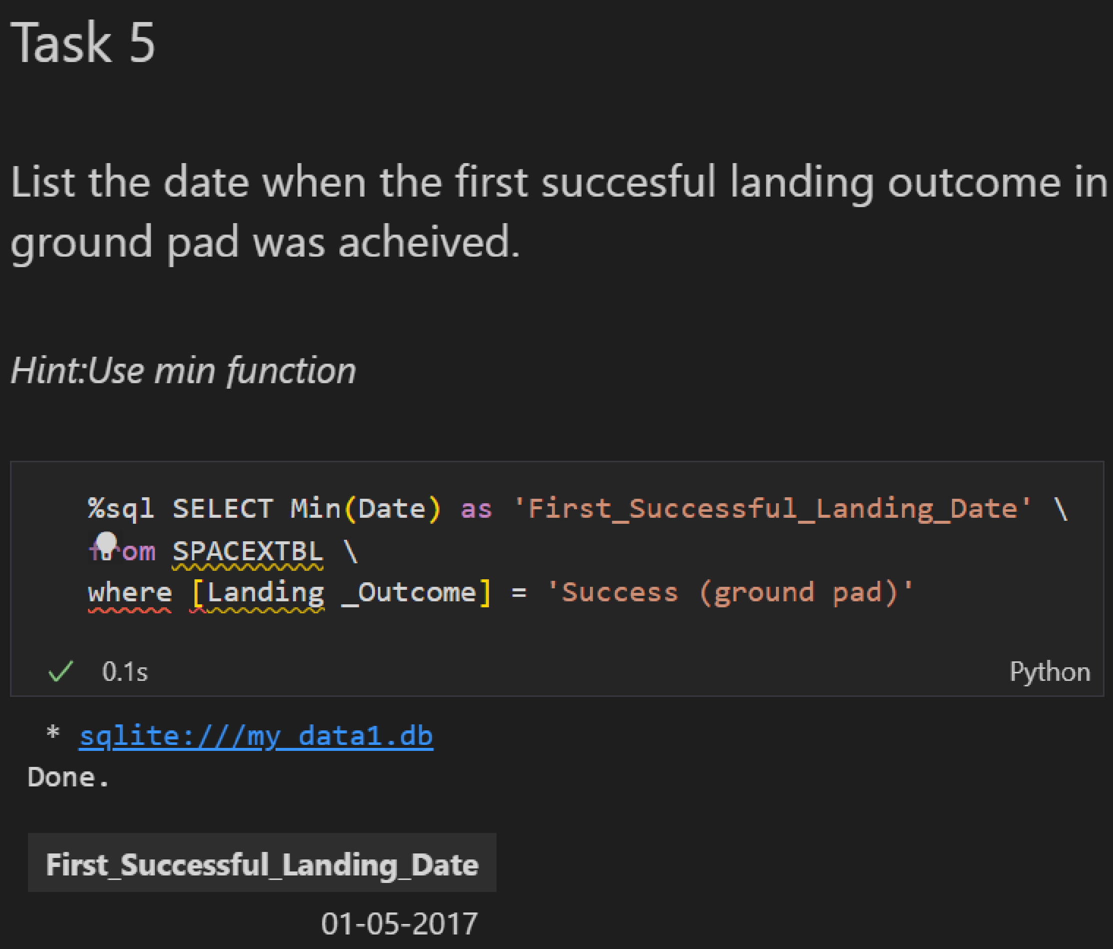
```
 * sqlite:///my_data1.db
Done.
```

Average_Payload_Mass_F9v1_1
2534.6666666666665

IBM Developer

SKILLS NETWORK

APPENDIX – EDA CODES



Task 5

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

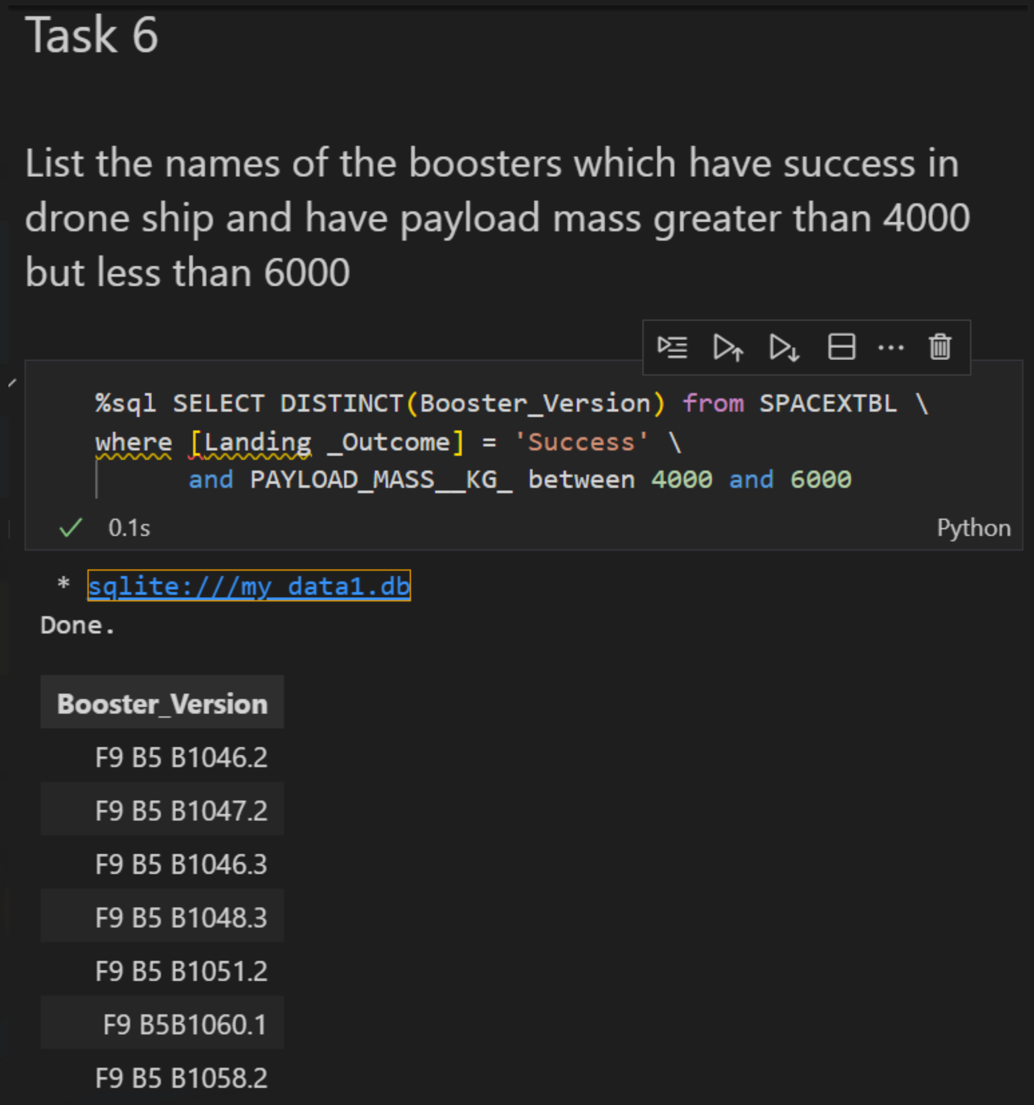
Hint:Use min function

```
%sql SELECT Min(Date) as 'First_Successful_Landing_Date' \
from SPACEXTBL \
where [Landing _Outcome] = 'Success (ground pad)'
```

✓ 0.1s Python

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

First_Successful_Landing_Date
01-05-2017



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

```
%sql SELECT DISTINCT(Booster_Version) from SPACEXTBL \
where [Landing _Outcome] = 'Success' \
    and PAYLOAD_MASS__KG_ between 4000 and 6000
```

✓ 0.1s Python

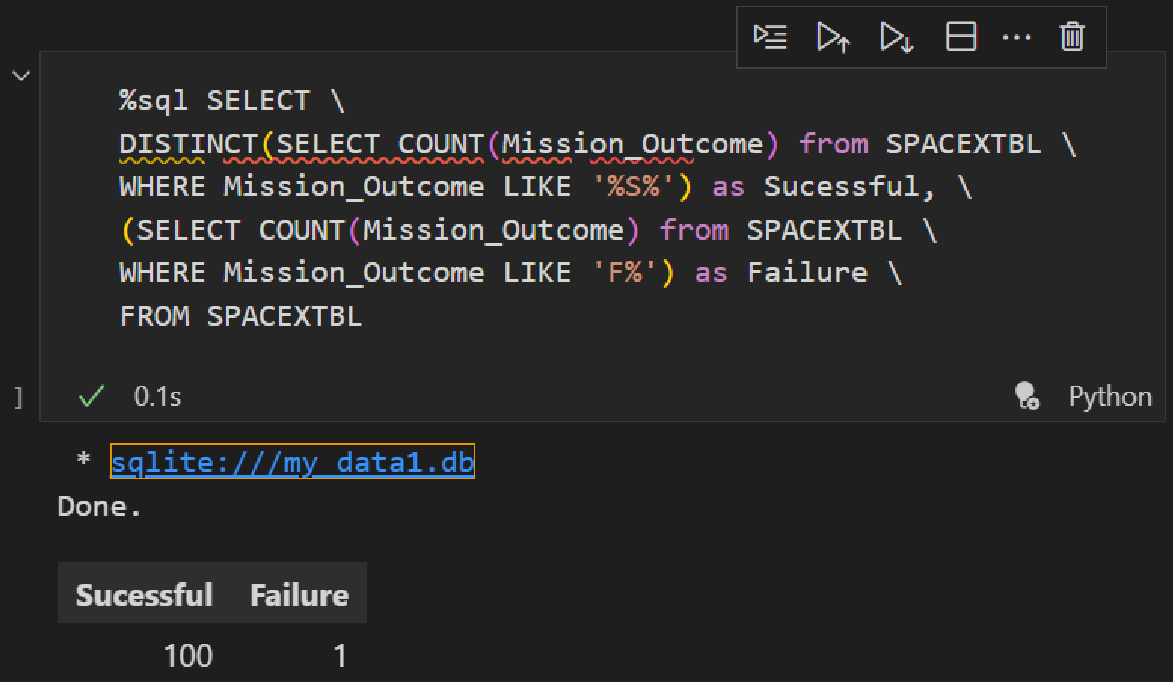
```
* sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2

APPENDIX – EDA CODES

Task 7

List the total number of successful and failure mission outcomes



```
%sql SELECT \
DISTINCT(SELECT COUNT(Mission_Outcome) from SPACEXTBL \
WHERE Mission_Outcome LIKE '%S%') as Sucessful, \
(SELECT COUNT(Mission_Outcome) from SPACEXTBL \
WHERE Mission_Outcome LIKE 'F%') as Failure \
FROM SPACEXTBL
```

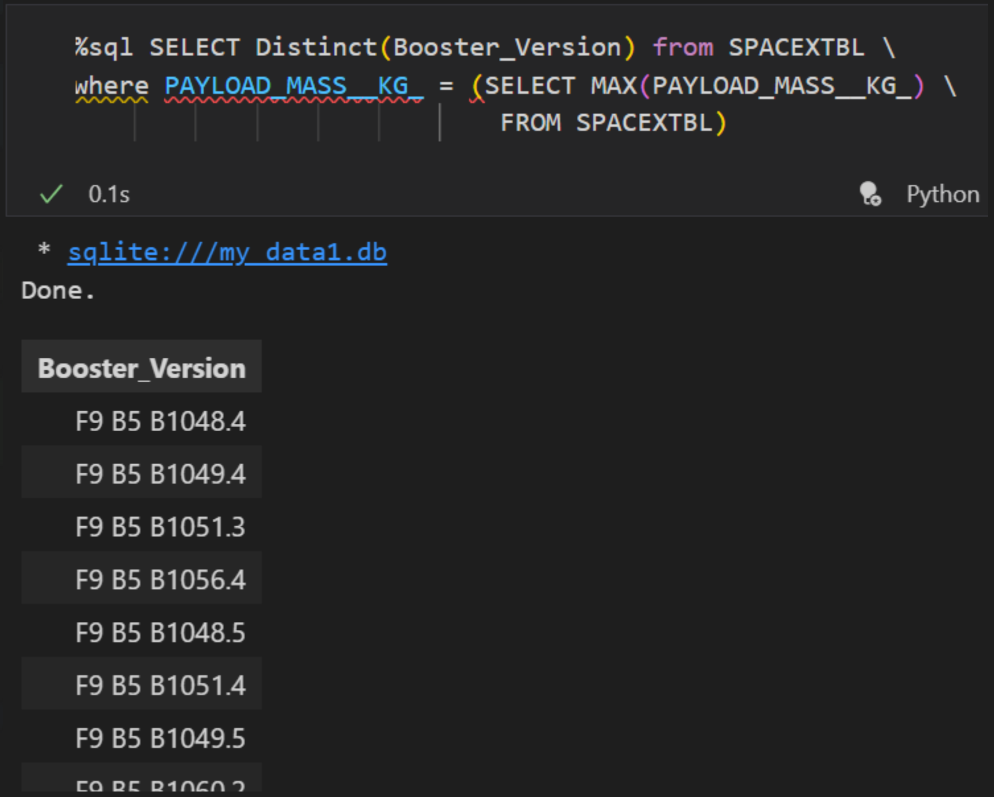
✓ 0.1s Python

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

Sucessful	Failure
100	1

Task 8

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



```
%sql SELECT Distinct(Booster_Version) from SPACEXTBL \
where PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) \
                           FROM SPACEXTBL)
```

✓ 0.1s Python

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

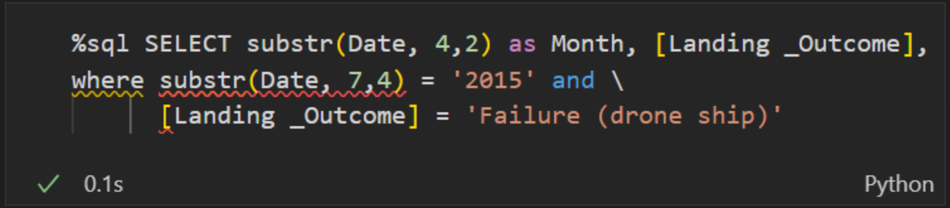
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2

APPENDIX – EDA CODES

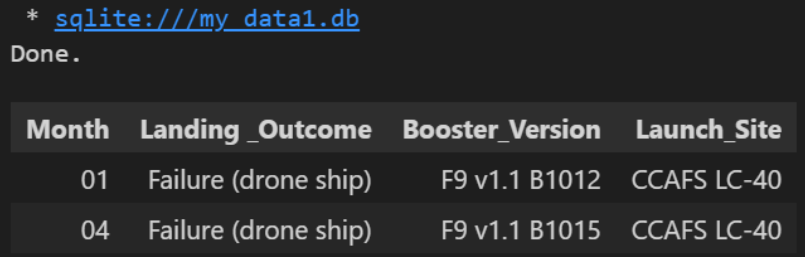
Task 9

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, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.



```
%sql SELECT substr(Date, 4,2) as Month, [Landing _Outcome],
where substr(Date, 7,4) = '2015' and \
      [Landing _Outcome] = 'Failure (drone ship)'
```

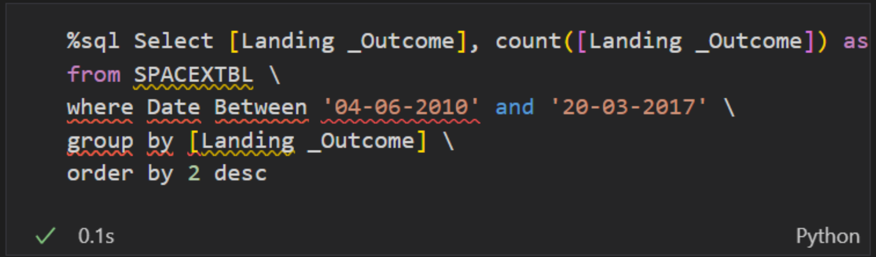


```
 * sqlite:///my_data1.db
Done.
```

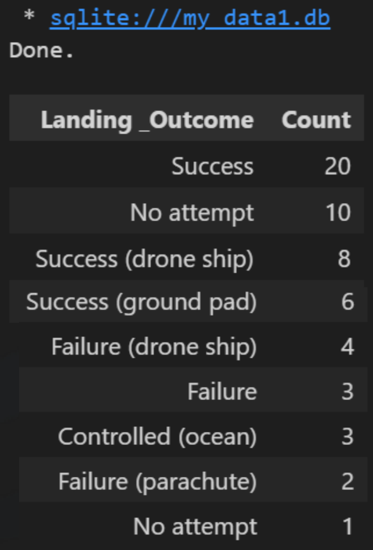
Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Task 10

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



```
%sql Select [Landing _Outcome], count([Landing _Outcome]) as
from SPACEXTBL \
where Date Between '04-06-2010' and '20-03-2017' \
group by [Landing _Outcome] \
order by 2 desc
```



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

Landing _Outcome	Count
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