Data Science Capstone Project

Minh-Hoang DANG

Summary

1. Executive Summary
2. Introduction
3. Methodology
4. Results
5. Conclusion
6. Appendix

Executive Summary

This capstone project encapsulates notions seen throughout the courses, namely:

- Summary of methodologies:
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)

- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

Background and Motivation

SpaceX stands as the leading company in the commercial space era, making space travel more affordable. They promote Falcon 9 rocket launches on their website at a price of $62 million, significantly lower than other providers, who charge over $165 million. A major factor in this cost reduction is SpaceX's ability to reuse the rocket's first stage. Therefore, by predicting whether the first stage can be recovered, we can estimate the launch cost. Using publicly available data and machine learning models, we aim to forecast if SpaceX will reuse the first stage.

Interesting questions

- How do factors like payload mass, launch site, number of flights, and orbit type influence the success of the first stage landing?
- Has the success rate of first stage landings improved over time?
- Which algorithm is most effective for binary classification in this scenario?

General methodology

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Performed data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

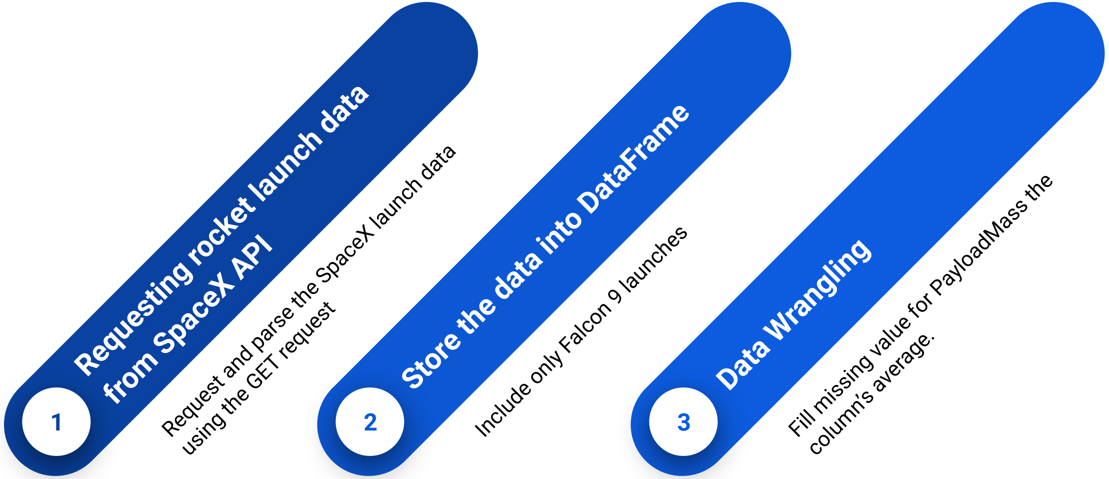
Methodology

Data collection

- The data collection process involved using a combination of API requests from the SpaceX REST API and web scraping from a table on SpaceX's Wikipedia page. Both methods were necessary to gather comprehensive information on the launches, enabling a more thorough analysis.

REST API	Web Scraping
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude	Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data collection – SpaceX API



1 **Requesting rocket launch data from SpaceX API**

Request and parse the SpaceX launch data using the GET request

2 **Store the data into DataFrame**

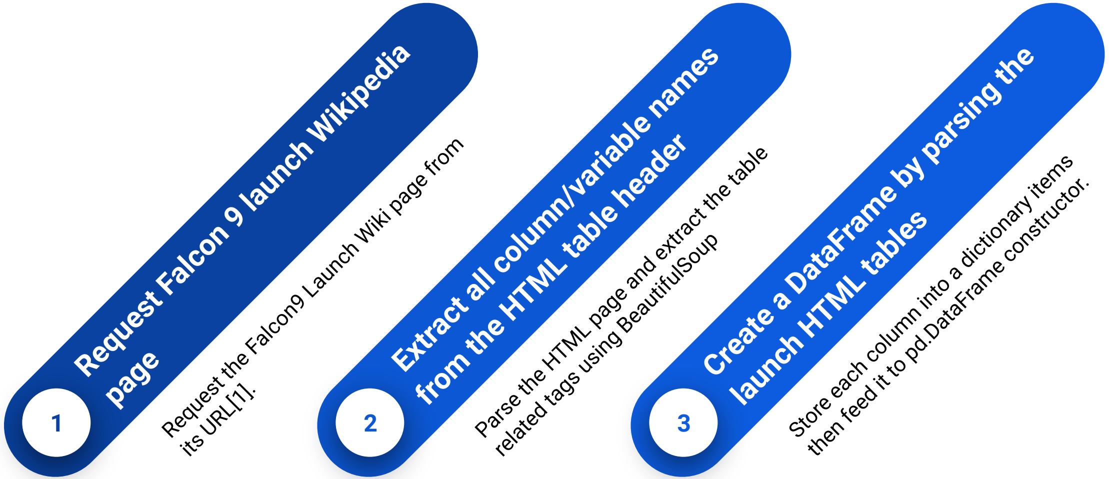
Include only Falcon 9 launches

3 **Data Wrangling**

Fill missing value for PayloadMass the column's average.

Data Collection API - Notebook -GitHub

Data Collection - Web Scraping



1 **Request Falcon 9 launch Wikipedia page**

Request the Falcon9 Launch Wiki page from its URL[1].

2 **Extract all column/variable names from the HTML table header**

Parse the HTML page and extract the table related tags using BeautifulSoup

3 **Create a DataFrame by parsing the launch HTML tables**

Store each column into a dictionary items then feed it to pd.DataFrame constructor.

Data Collection Web Scraping - Notebook -GitHub

[1] https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Wrangling

Objective: find patterns in the data and determine what would be the label for training supervised models.

1 Calculate the number of launches on each site

2 Calculate the number and occurrence of each orbit

3 Calculate the number and occurence of mission outcome of the orbits

4 Create a landing outcome label from Outcome column

Data Wrangling - Notebook - GitHub

EDA with data visualization

The aim is to visually assert the association between variables and select appropriate feature for Machine Learning model:

- Flight Number vs. Payload Mass,
- Flight Number vs. Launch Site,
- Payload Mass vs. Launch Site,
- Orbit Type vs. Success Rate,
- Flight Number vs. Orbit Type,
- Payload Mass vs Orbit Type
- Success Rate Yearly Trend

EDA Dataviz - Notebook -GitHub

EDA with SQL

The aim is to explore the data by asking relevant questions then obtain the answer from the database:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

EDA SQL - Notebook - GitHub

Build an interactive map with Folium

Markers of all Launch Sites

Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site

Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities

Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

[Folium - Notebook - GitHub]()

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List

Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches

Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range

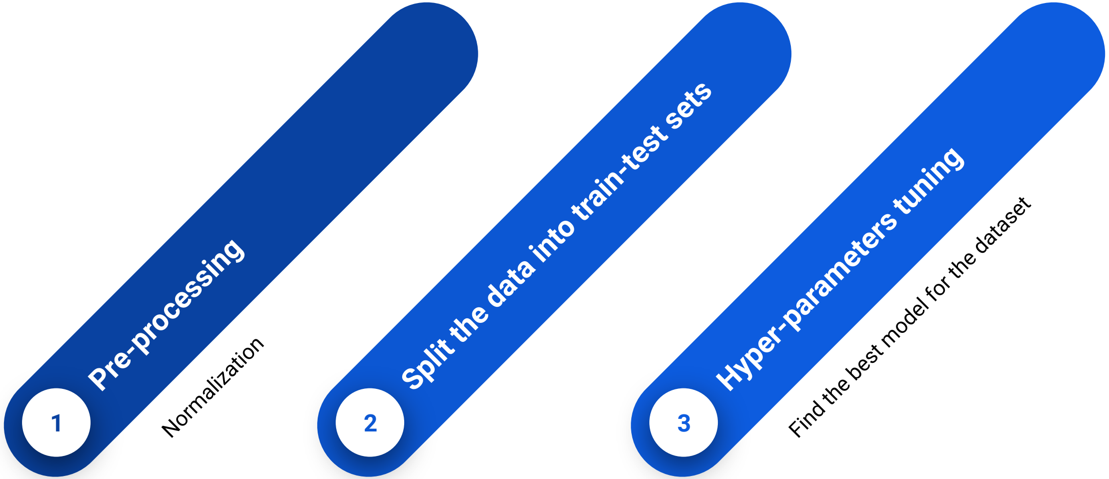
Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate

Added a scatter chart to show the correlation between Payload and Launch Success.

[Plotly Dash - Python - GitHub]

Predictive analysis



1 Pre-processing

Normalization

2 Split the data into train-test sets

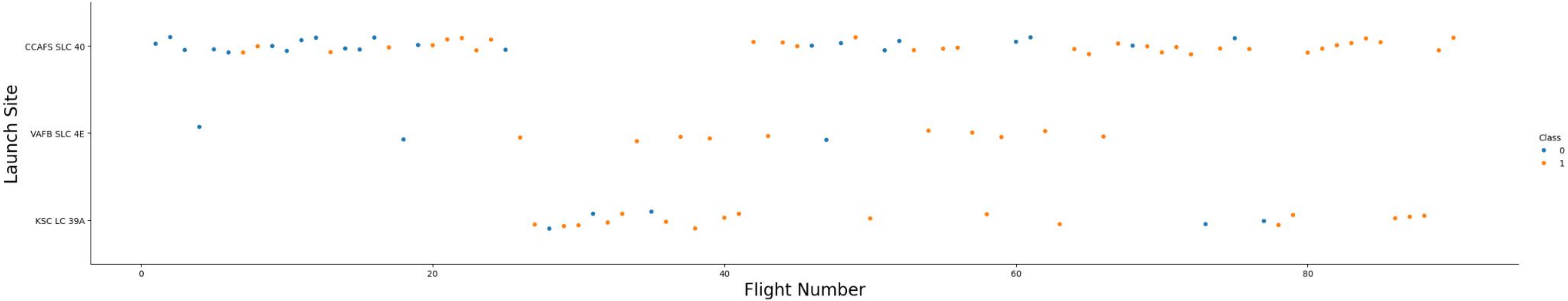
3 Hyper-parameters tuning

Find the best model for the dataset

[Predictive Analysis - Notebook - GitHub]()

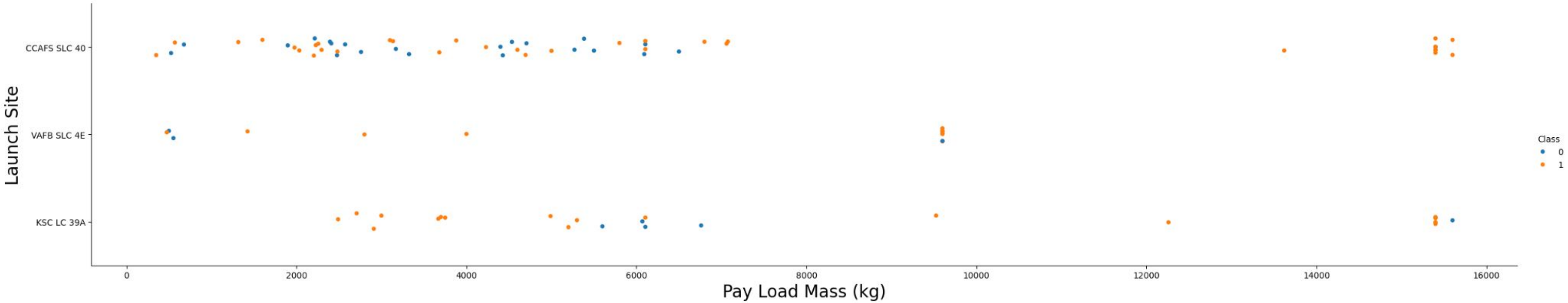
Results

EDA with vizualization - Flight Number vs. Launch Site



- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- Each new launch has a higher rate of success.

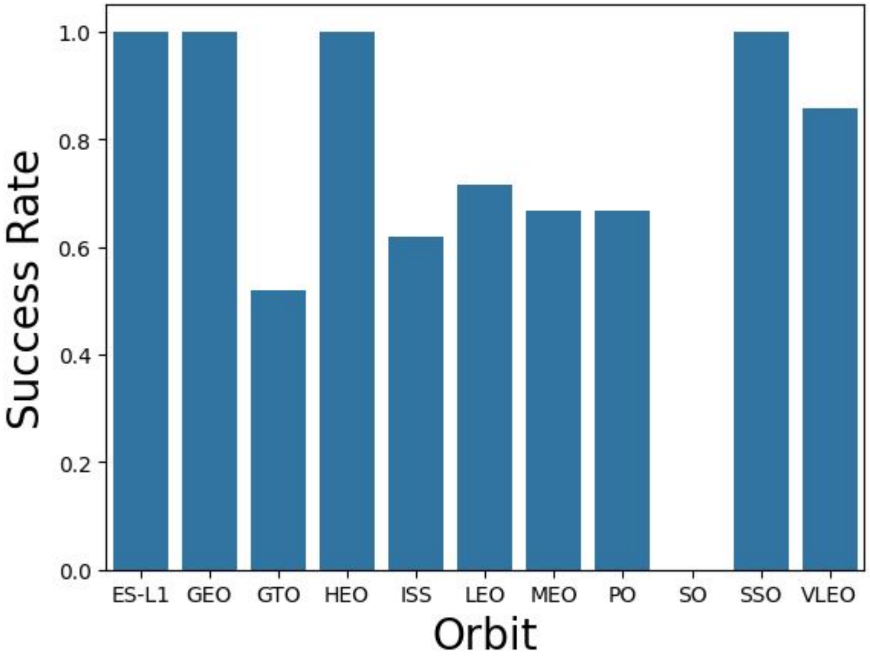
EDA with vizualization - Payload vs. Launch Site



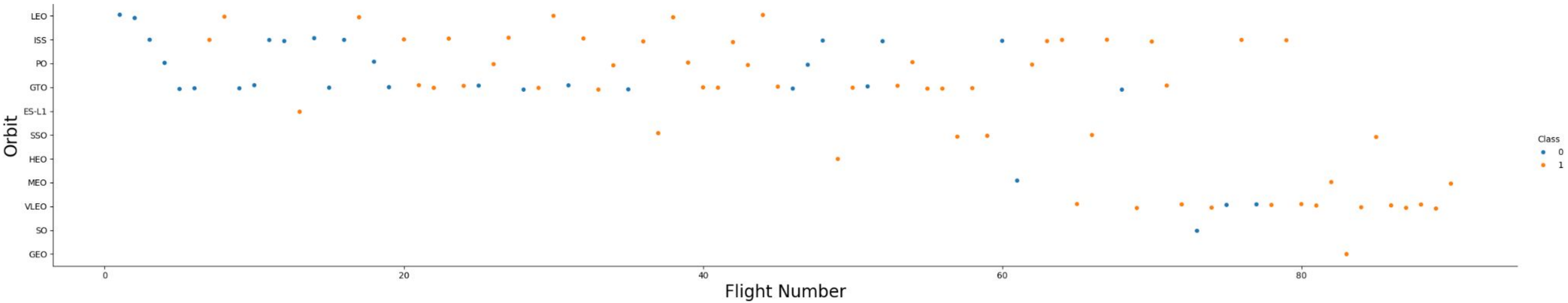
- For every launch site: the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg as well.

EDA with vizualization - Success rate vs. Orbit type

- Orbits with 100% success rate:
    - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
    - SO
- Orbits with success rate between 50% and 85%:
    - GTO, ISS, LEO, MEO, PO

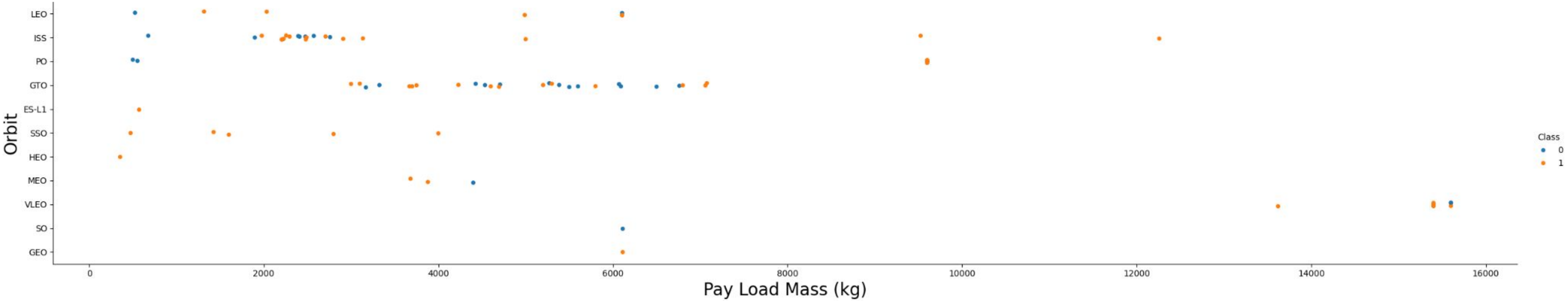


EDA with vizualization - Flight Number vs. Orbit type



- In the LEO orbit the Success appears related to the number of flights.
- There is no relationship between flight number when in GTO orbit.

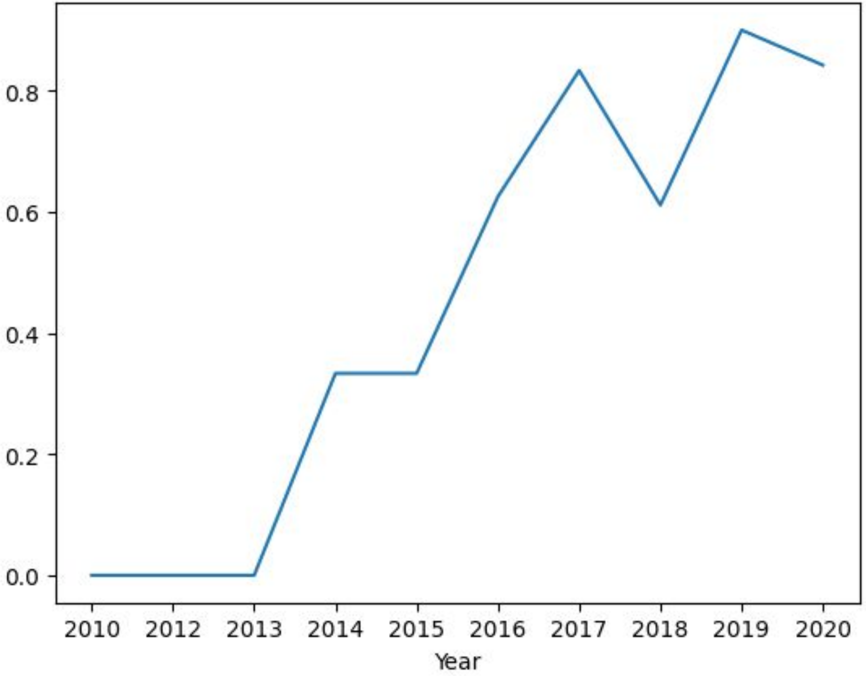
EDA with vizualization - Payload Mass vs. Orbit type



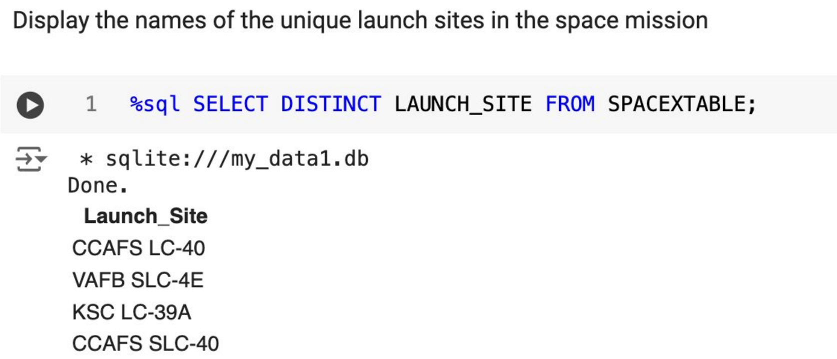
- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

EDA with vizualization - Success rate vs. Orbit type

- The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



EDA with SQL - Q1

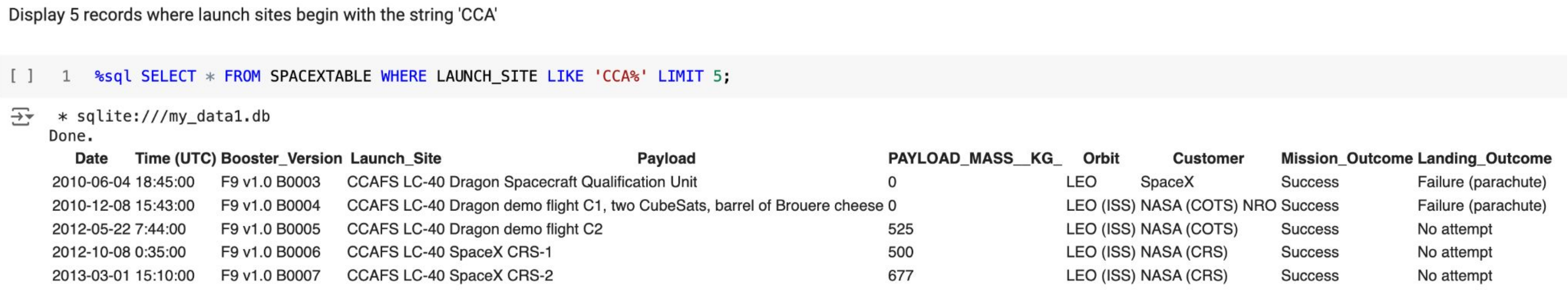


Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
  Launch_Site
 CCAFS LC-40
 VAFB SLC-4E
 KSC LC-39A
 CCAFS SLC-40
```

EDA with SQL - Q2



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

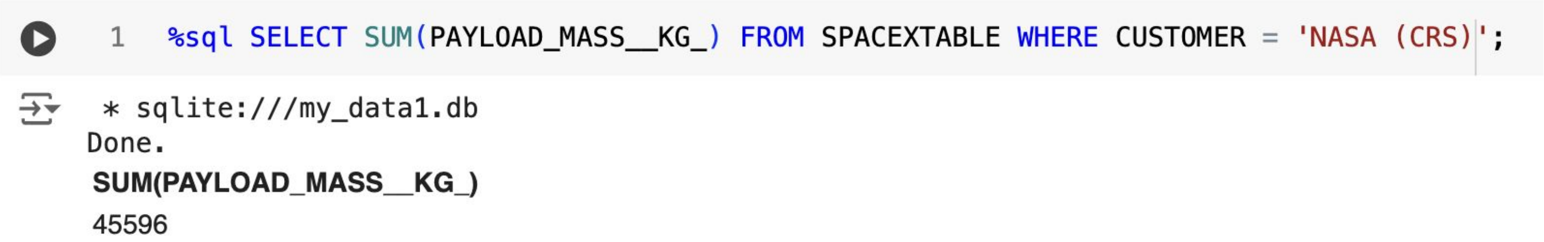
```
%sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

EDA with SQL - Q3

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



```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER = 'NASA (CRS)';
```

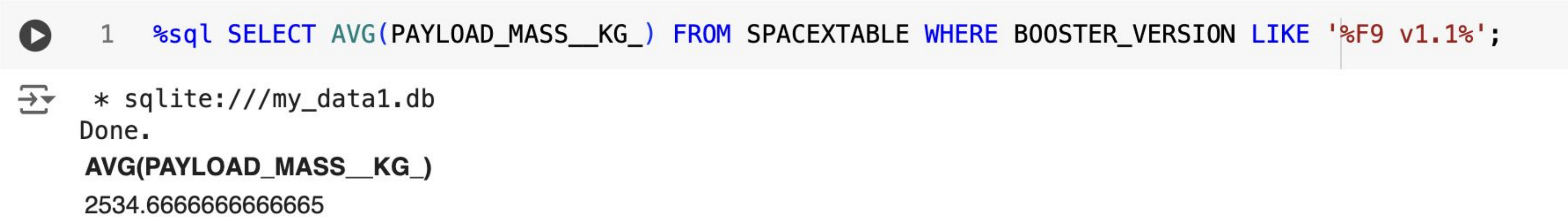
```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)

45596

EDA with SQL - Q4

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



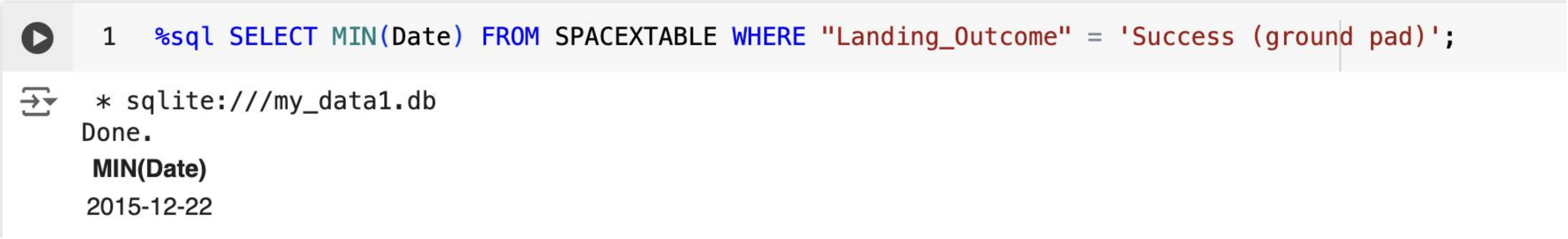
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE BOOSTER_VERSION LIKE '%F9 v1.1%';
```

```
 * sqlite:///my_data1.db
Done.
 AVG(PAYLOAD_MASS__KG_)
 2534.6666666666665
```

EDA with SQL - Q5

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

Hint:Use min function

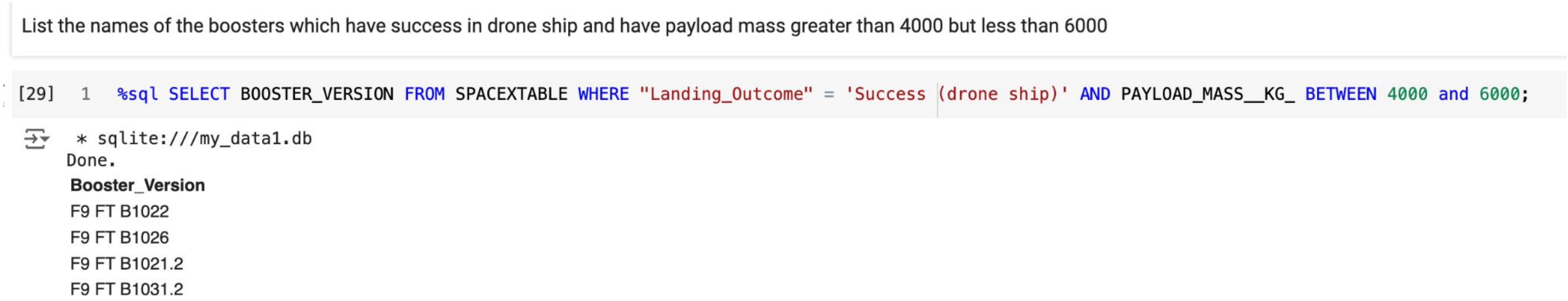


```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

MIN(Date)
2015-12-22

EDA with SQL - Q6



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

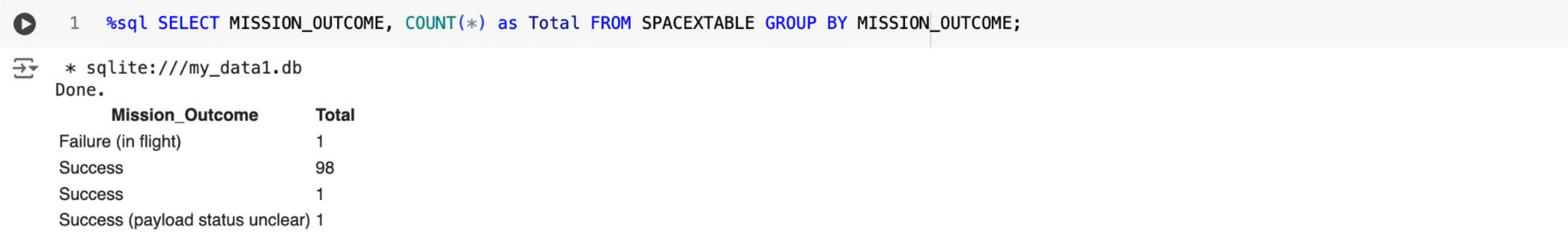
```
[29]  %sql SELECT BOOSTER_VERSION FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

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

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

EDA with SQL - Q7

List the total number of successful and failure mission outcomes

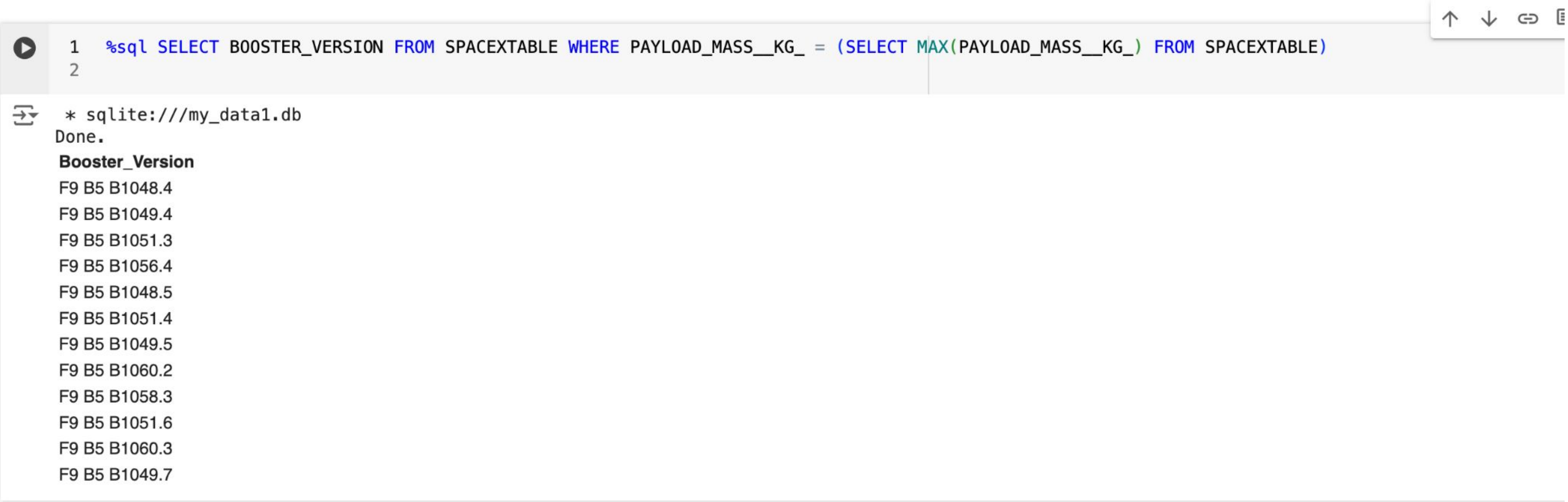


```
%sql SELECT MISSION_OUTCOME, COUNT(*) as Total FROM SPACEXTABLE GROUP BY MISSION_OUTCOME;
```

```
 * sqlite:///my_data1.db
Done.
        Mission_Outcome          Total
Failure (in flight)              1
Success                          98
Success                          1
Success (payload status unclear) 1
```

EDA with SQL - Q8

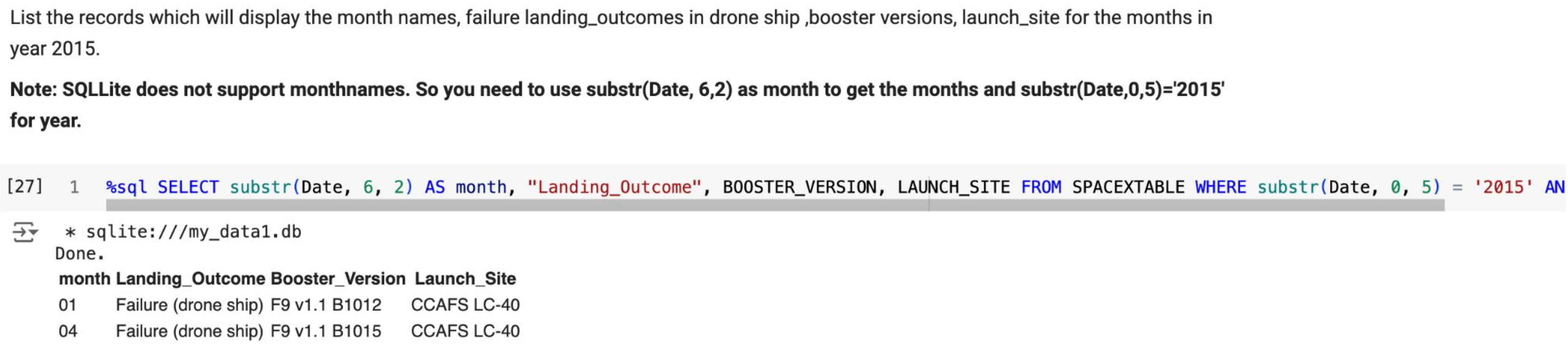
List the names of the booster_versions which have carried the maximum payload mass. Use a subquery



```
%sql SELECT BOOSTER_VERSION FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * 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
 F9 B5 B1058.3
 F9 B5 B1051.6
 F9 B5 B1060.3
 F9 B5 B1049.7
```

EDA with SQL - Q9



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, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

[27]

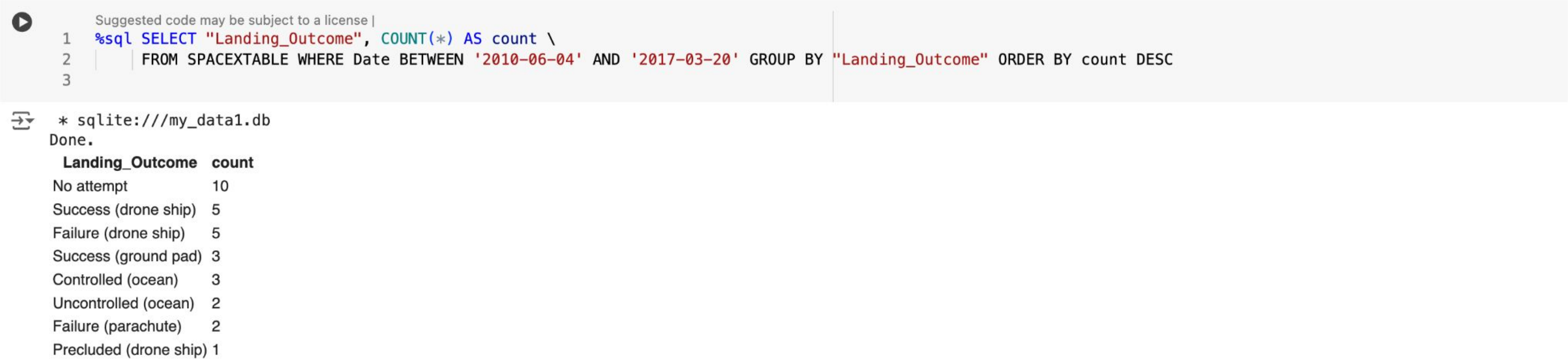
```
%sql SELECT substr(Date, 6, 2) AS month, "Landing_Outcome", BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTABLE WHERE substr(Date, 0, 5) = '2015' AN
```

```
 * 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

EDA with SQL - Q10

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.



Suggested code may be subject to a license |

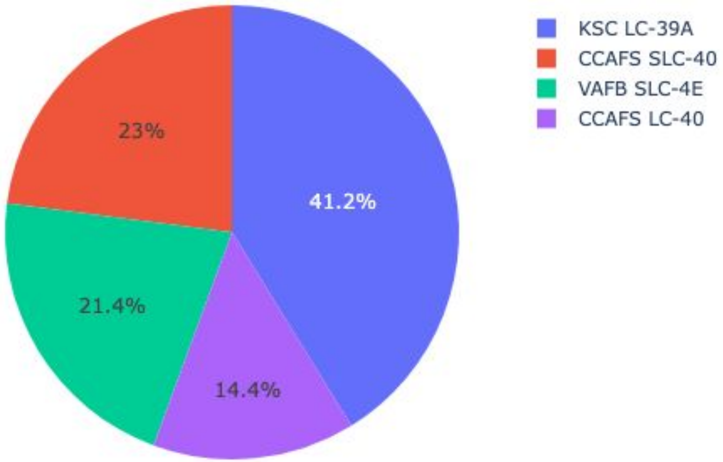
```
%sql SELECT "Landing_Outcome", COUNT(*) AS count \
    FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY count DESC

```

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

Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

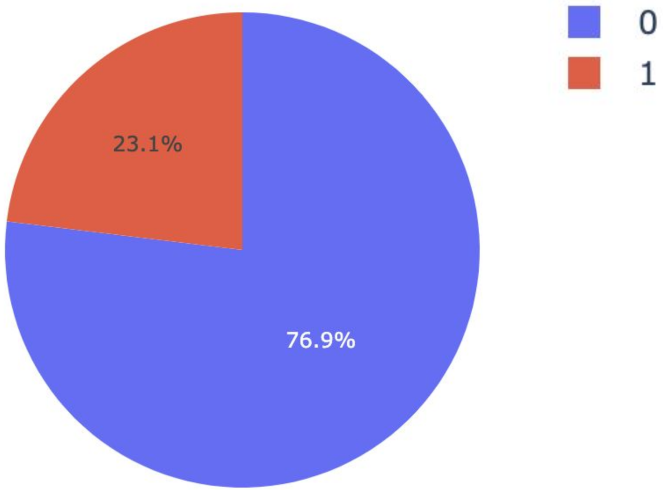
Dashboard - Launch success count for all sites

Of all the launch sites, KSC LC-39A has the most successful launches.

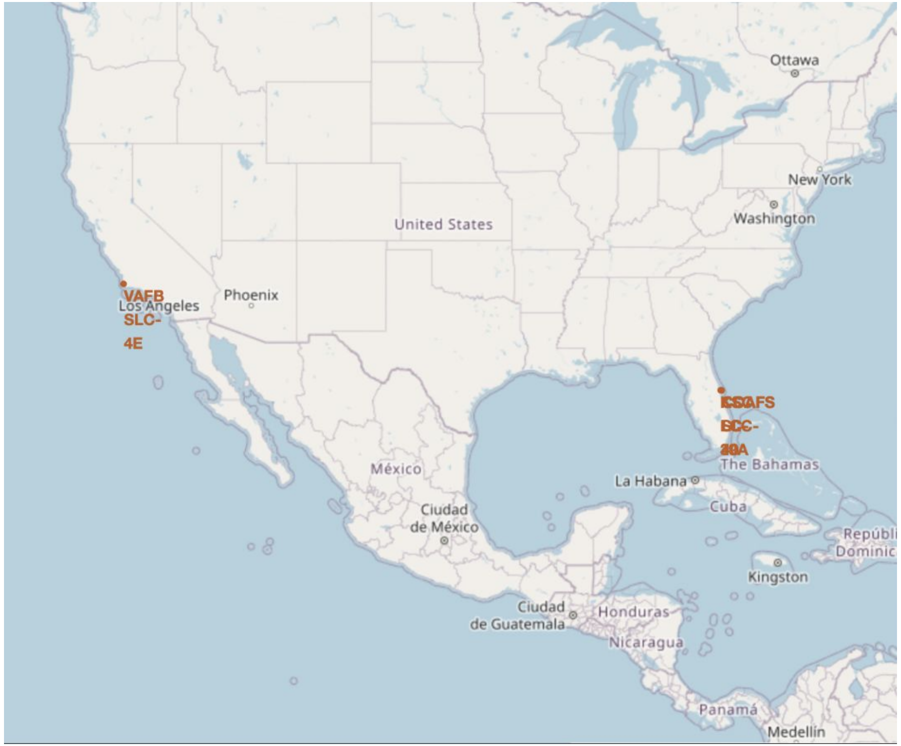


Dashboard - Launch success count for all sites

KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

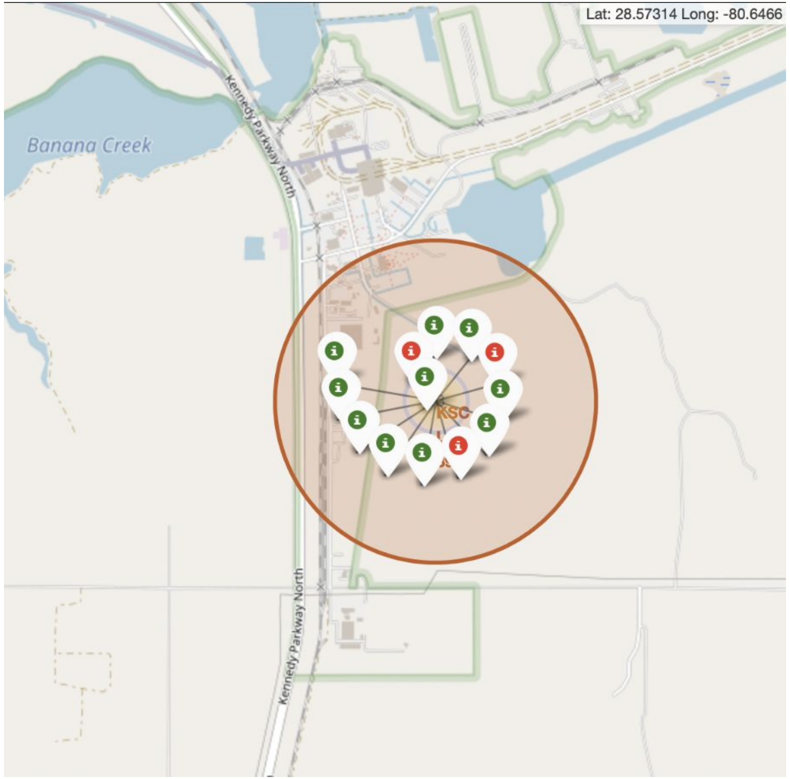


Interactive map with Folium



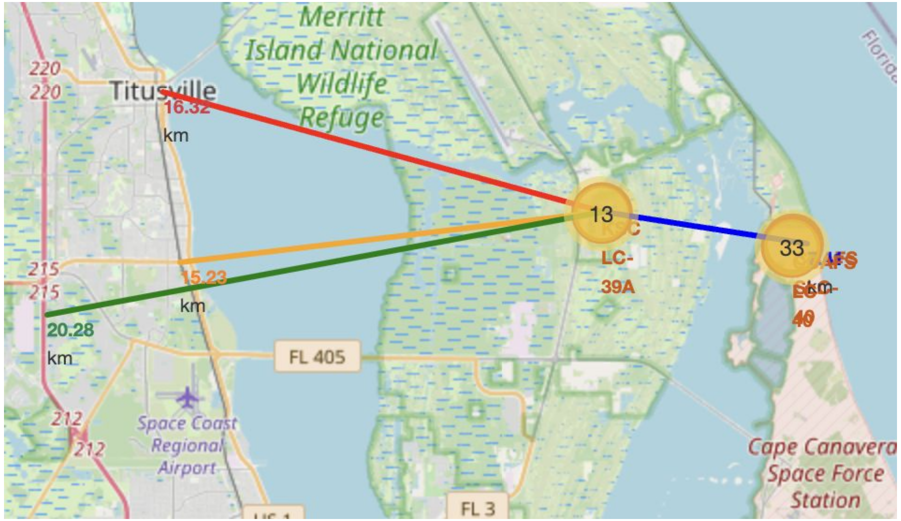
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.

Interactive map with Folium



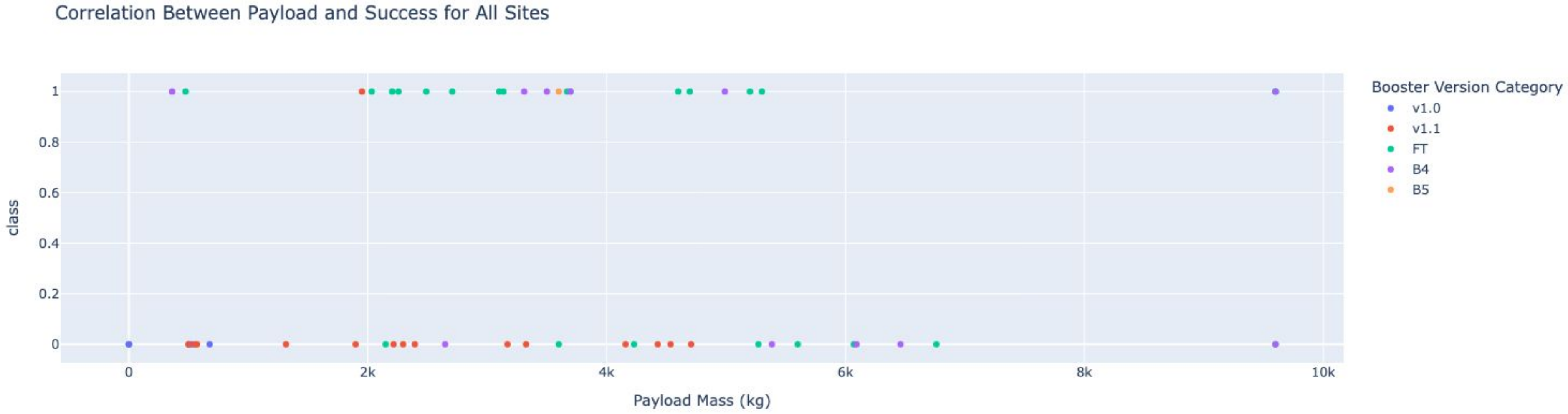
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.

Interactive map with Folium


Merritt Island National Wildlife Refuge
Titusville
16.32 km
15.23 km
20.28 km
13
33
KSC LC-39A
CCAFS SLC-40
FL 405
FL 3
Space Coast Regional Airport
Cape Canaveral Space Force Station
220
215
212

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relative close to city (16.32 km)
  - relative close to railway (15.23 km)
  - relative close to highway (22.74 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

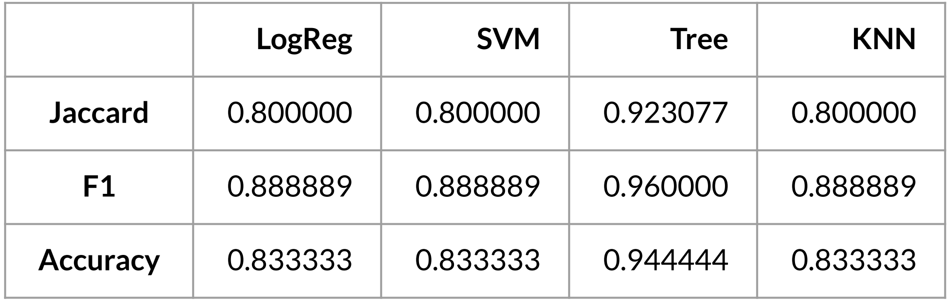
Dashboard - Payload Mass vs. Launch Outcome



The charts show that payloads between 2000 and 5500 kg have the highest success rate.

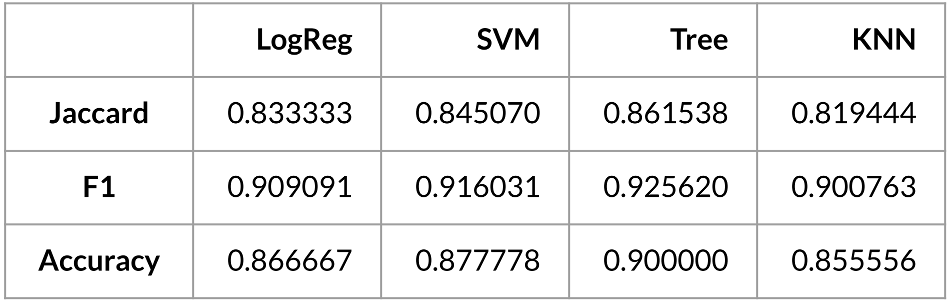
Predictive Analysis - Classification Performance

- Based on the scores of the Test Set, we can not confirm which method performs best.
- The best model is the Decision Tree Model, having the highest Jaccard, F1 and Accuracy.



	LogReg	SVM	Tree	KNN
Jaccard	0.800000	0.800000	0.923077	0.800000
F1	0.888889	0.888889	0.960000	0.888889
Accuracy	0.833333	0.833333	0.944444	0.833333

Figure A: Scores and Accuracy of the test set

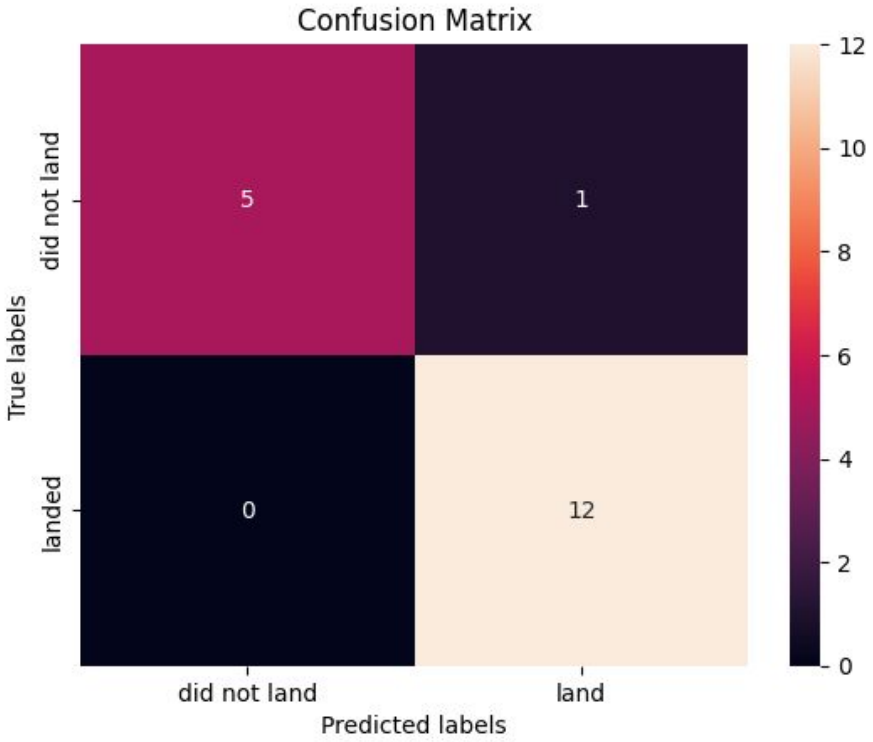


	LogReg	SVM	Tree	KNN
Jaccard	0.833333	0.845070	0.861538	0.819444
F1	0.909091	0.916031	0.925620	0.900763
Accuracy	0.866667	0.877778	0.900000	0.855556

Figure B: Scores and Accuracy of the whole dataset

Predictive Analysis - Confusion Matrix

- The Decision tree method perform very well with very few instances of mis-classification (only 1 false positive).



Conclusion

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.