IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Elham heidari
7/23/2023

Outline

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

Executive Summary

Summary of methodologies

Data Collection: data using SpaceX REST API and web scraping techniques
Data Wrangling : data to create success/fail outcome variable
EDA with Data Visualization : data with data visualization techniques, considering the following factors
Analyze: the data with SQL, calculating the following statistics
Explore: launch site success rates and proximity to geographical markers
Visualize: the launch sites with the most success and successful payload ranges
Build Models :to predict landing outcomes using logistic regression, SVM, decision tree and KNN

Summary of all results

Exploratory Data Analysis
Visualization/Analytics
Predictive Analytics

Introduction

Project background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions to be answered

1. How payload mass, launch site, number of flights, and orbits affect first-stage landing success?
2. rate of successful landings increase over the years?
3. Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- data using SpaceX REST API and web scraping techniques
- Perform data wrangling
  - data to create success/fail outcome variable

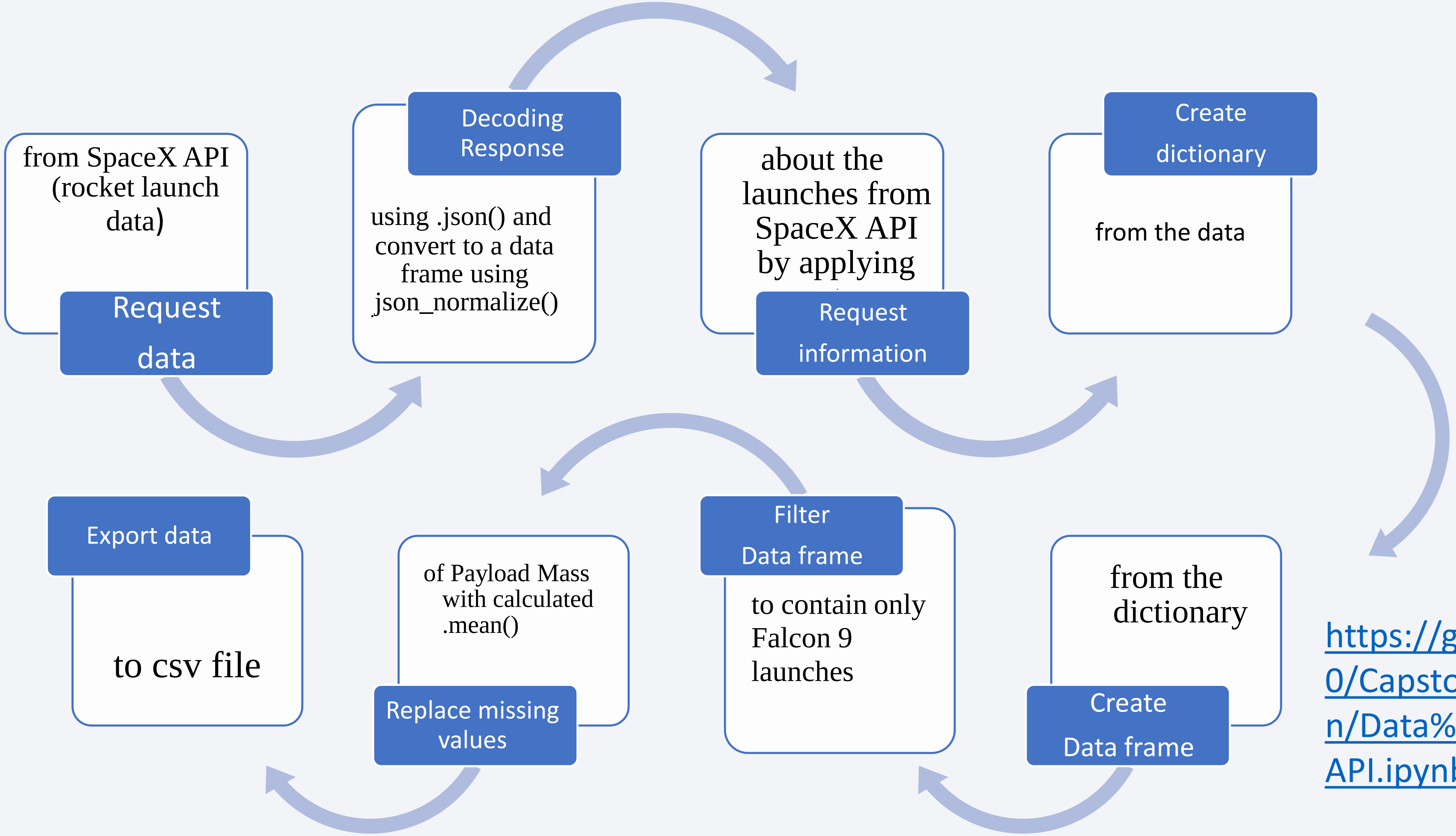
•Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- to predict landing outcomes using logistic regression, SVM, decision tree and KNN

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API



Request data
from SpaceX API (rocket launch data)

Decoding Response
using .json() and convert to a data frame using json_normalize()

Request information
about the launches from SpaceX API by applying

Create dictionary
from the data

Create Data frame
from the dictionary

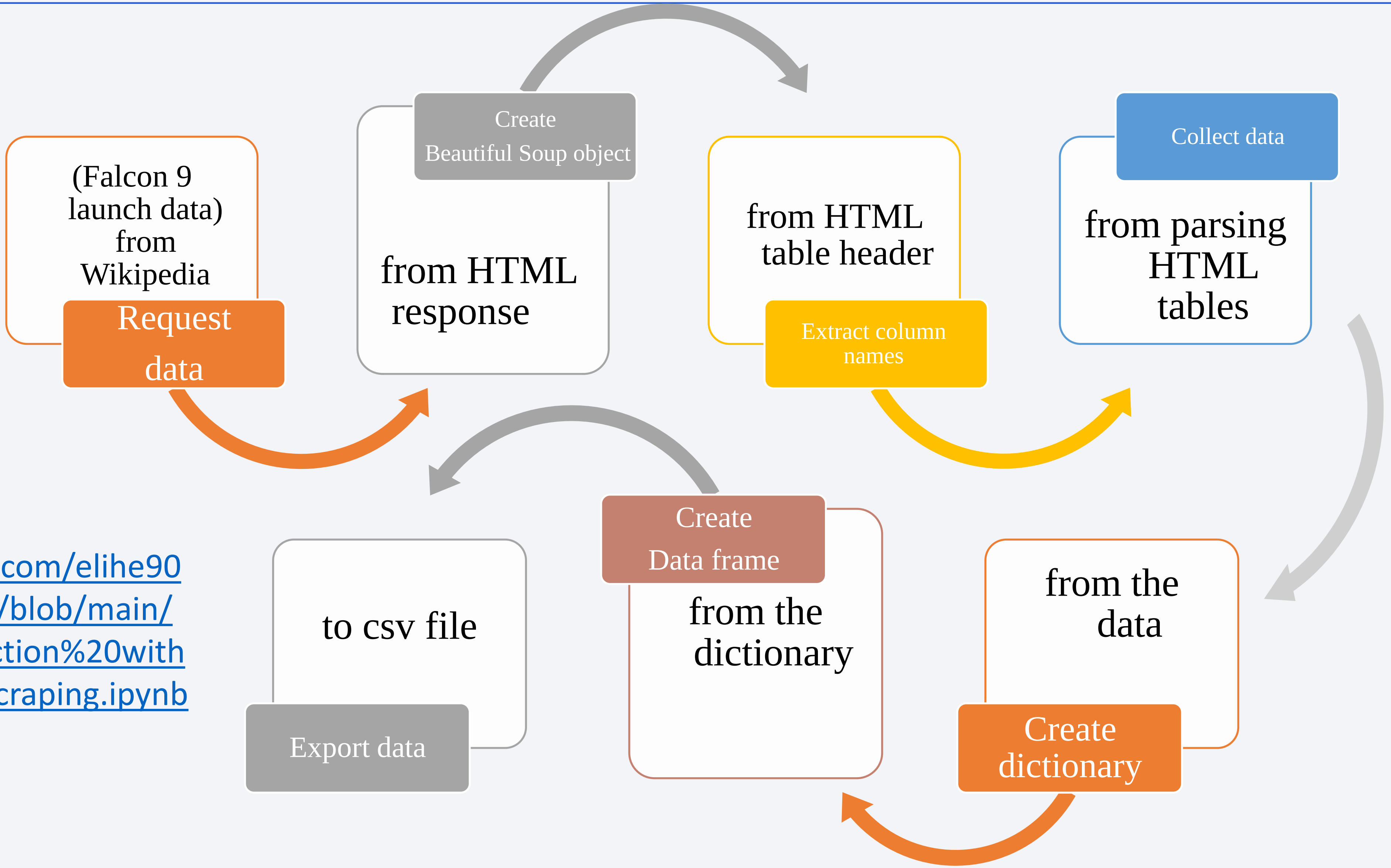
Filter Data frame
to contain only Falcon 9 launches

Replace missing values
of Payload Mass with calculated .mean()

Export data
to csv file

https://github.com/elihe90/Capston_IBM/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping



Request data

(Falcon 9 launch data) from Wikipedia

Create Beautiful Soup object

from HTML response

Extract column names

from HTML table header

Collect data

from parsing HTML tables

Create dictionary

from the data

Create Data frame

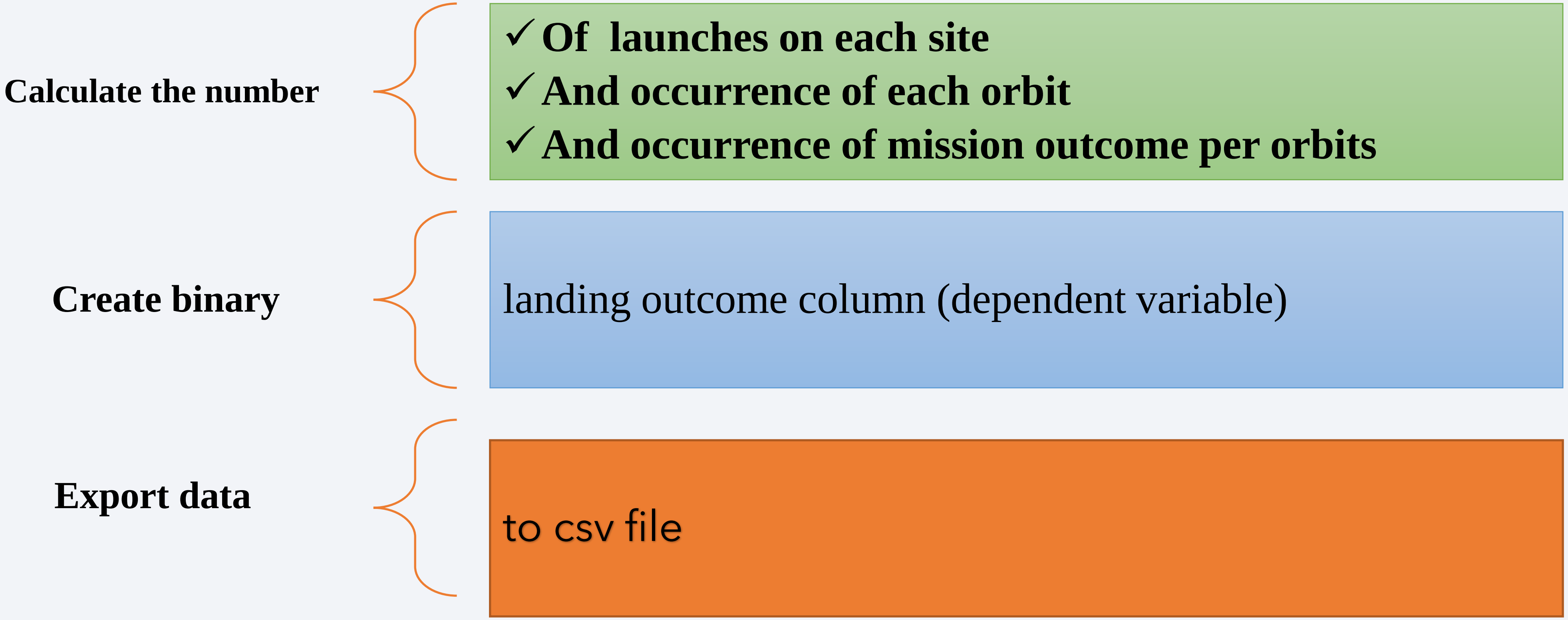
from the dictionary

Export data

to csv file

https://github.com/elihe90/Capston_IBM/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling



Calculate the number

- ✓ **Of launches on each site**
- ✓ **And occurrence of each orbit**
- ✓ **And occurrence of mission outcome per orbits**

Create binary

landing outcome column (dependent variable)

Export data

to csv file

https://github.com/elihe90/Capston_IBM/blob/main/EDA.ipynb

EDA with Data Visualization

Flight Number vs. Payload
Flight Number vs. Launch Site
Payload Mass (kg) vs. Launch Site
Payload Mass (kg) vs. Orbit type

https://github.com/elihe90/Capston_IBM/blob/main/jupyter-labs-eda-dataviz.ipynb

Scatter plots show the relationship between variables.
If a relationship exists, they could be used in machine learning model.
Bar charts show comparisons among discrete categories.
The goal is to show the relationship between the specific categories being compared and a measured value. Line charts show trends in data over time (time series).

EDA with SQL

Displaying

- the names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string ‘CCA’
- the total payload mass carried by boosters launched by NASA (CRS)
- average payload mass carried by booster version F9 v1.1

https://github.com/elihe90/Capston_IBM/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Listing

- date when the first successful landing outcome in ground pad was achieved
- names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- total number of successful and failure mission outcomes
- names of the booster versions which have carried the maximum payload mass
- failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

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.

Colored Markers of the launch outcomes for each Launch Site:

Added **colored markers** of **successful**(**green**) and **unsuccessful**(**red**) launches at each launch site to show which launch sites have high success rates

https://github.com/elihe90/Capston_IBM/blob/main/lab_jupyter_launch_site_location.ipynb

Distances between a Launch Site to its proximities
Added colored Lines to show distances between the Launch Site
CCAFS SLC- 40 and its proximity to the **nearest coastline, railway, highway, and city**

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

Allow a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches:

add user to see successful and unsuccessful launches as a percent of the total

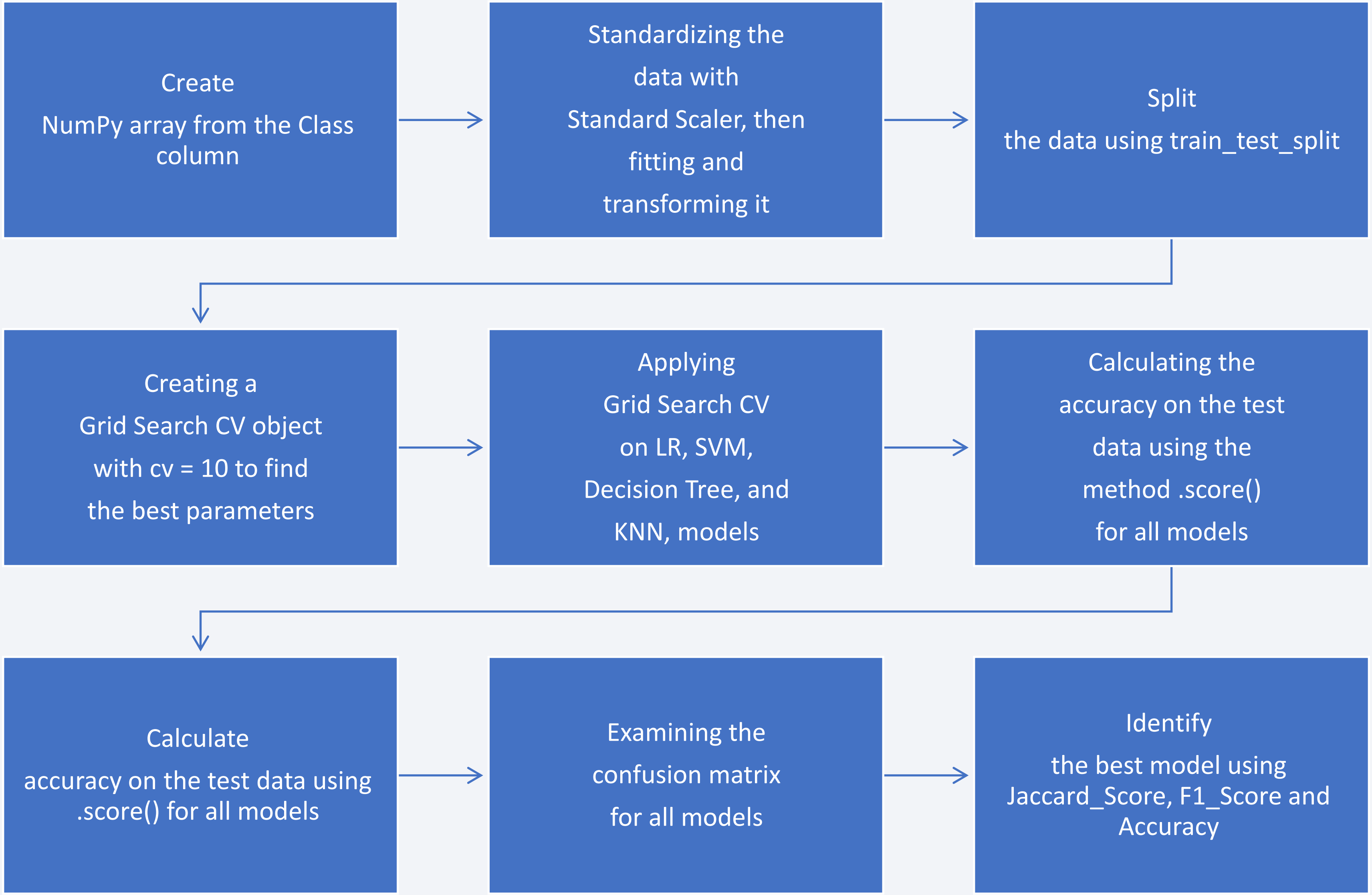
Slider of Payload Mass Range:

add a slider to select Payload range.

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

allow a scatter chart to see the correlation between Payload and Launch Success.

Predictive Analysis (Classification)



Create NumPy array from the Class column

Standardizing the data with Standard Scaler, then fitting and transforming it

Split the data using train_test_split

Creating a Grid Search CV object with cv = 10 to find the best parameters

Applying Grid Search CV on LR, SVM, Decision Tree, and KNN, models

Calculating the accuracy on the test data using the method .score() for all models

Calculate accuracy on the test data using .score() for all models

Examining the confusion matrix for all models

Identify the best model using Jaccard_Score, F1_Score and Accuracy

https://github.com/elihe90/Capston_IBM/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

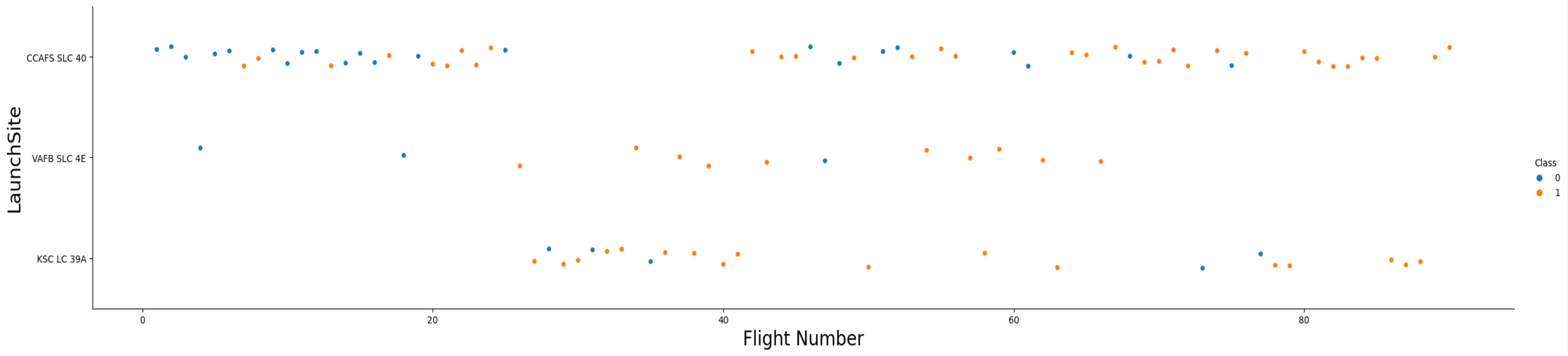
Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



EDA:

Zero class with blue color of failures and class 1 with orange color of successes

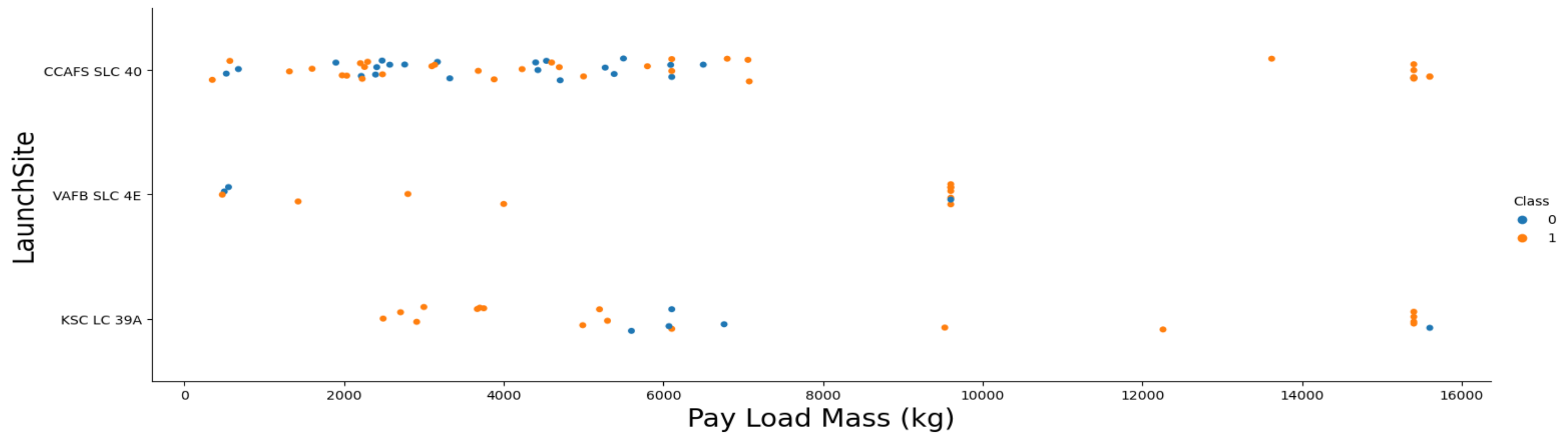
- Lower success first flights
- Future flights of higher success

Around half of launches were from CCAFS SLC 40 launch site

VAFB SLC 4E and KSC LC 39A have higher success rates

We can infer that new launches have a higher success rate

Payload vs. Launch Site



EDA:

For every launch site the **higher payload mass** (kg), the **higher** the **success rate**

very launches with payload mass over 7000 kg were successful.

KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type

❖ ES-L1, GEO, HEO and SSO

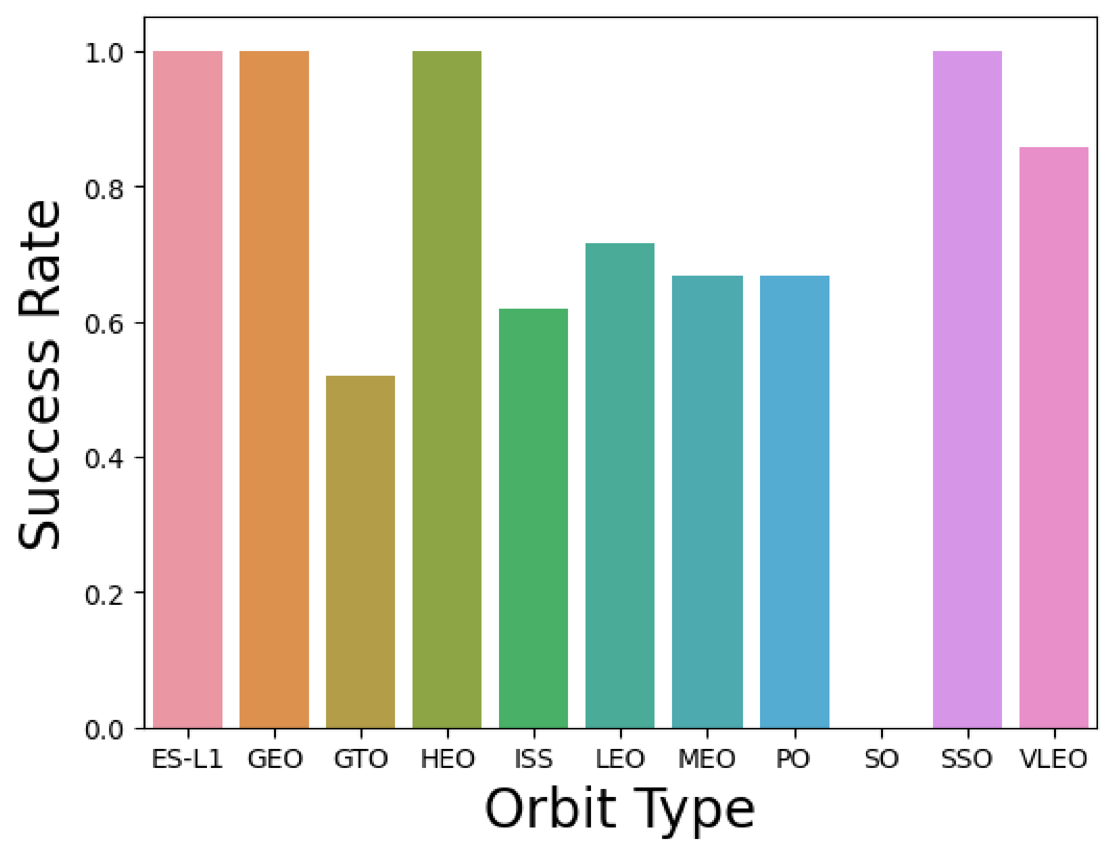
100% Success Rate

❖ GTO, ISS, LEO, MEO, PO

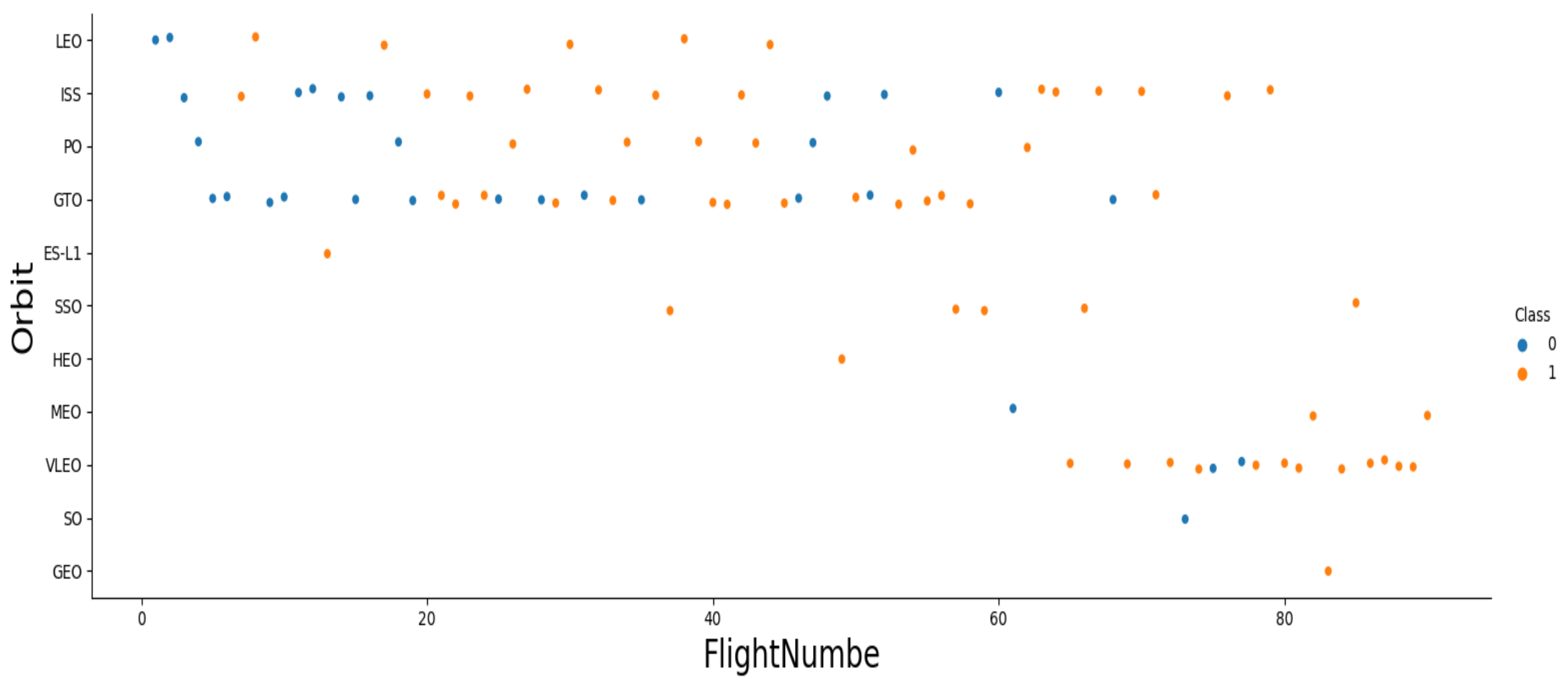
- **50%-80% Success Rate**

❖ SO

Orbits with 0% success rate



Flight Number vs. Orbit Type

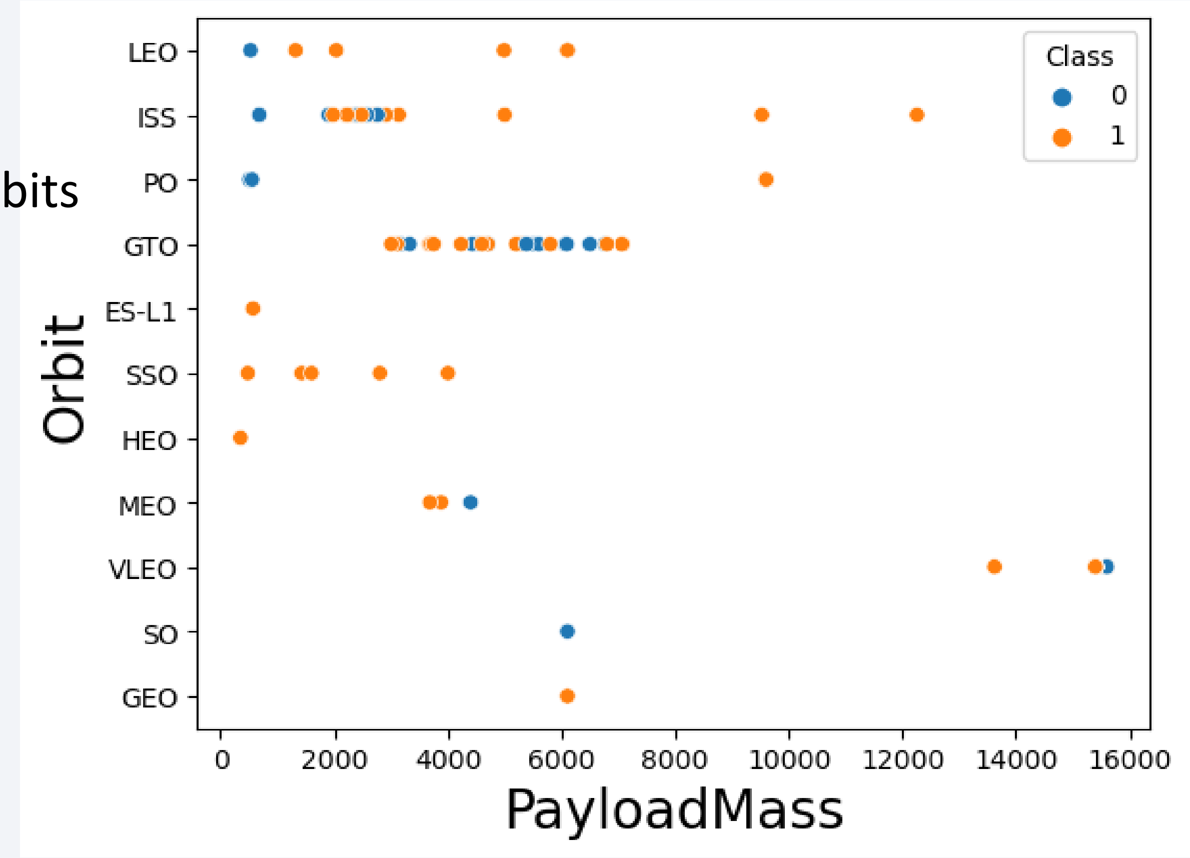


EDA:

The success rate seems to increase with the number of flights
In the LEO orbit the Success appears related to the number of flights
there seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type

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

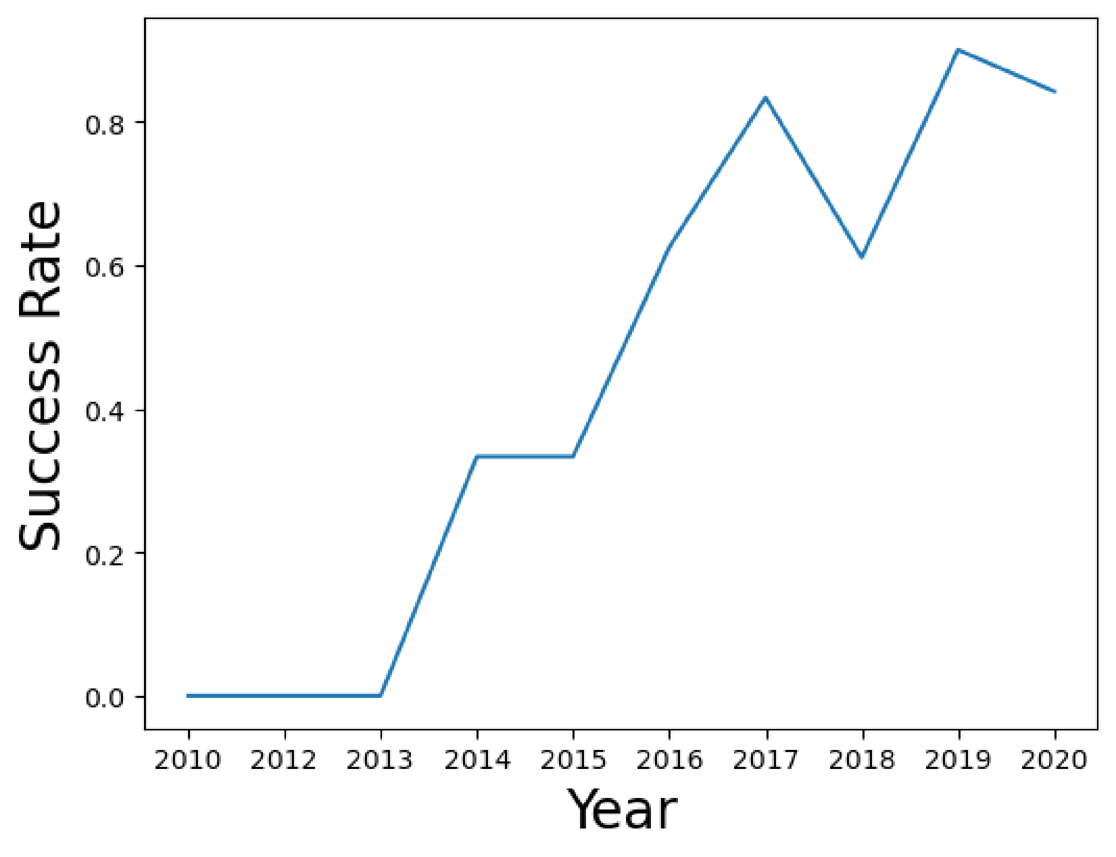


- Show the screenshot of the scatter plot with explanations

Launch Success Yearly Trend

The success rate since 2013 increasing till 2020.
The gate from 2014 to 2015 had a steady trend compared to the increase in 2013
From 2018 to 2019, there is a decrease compared to the previous period
but, the success rate has improved since 2013

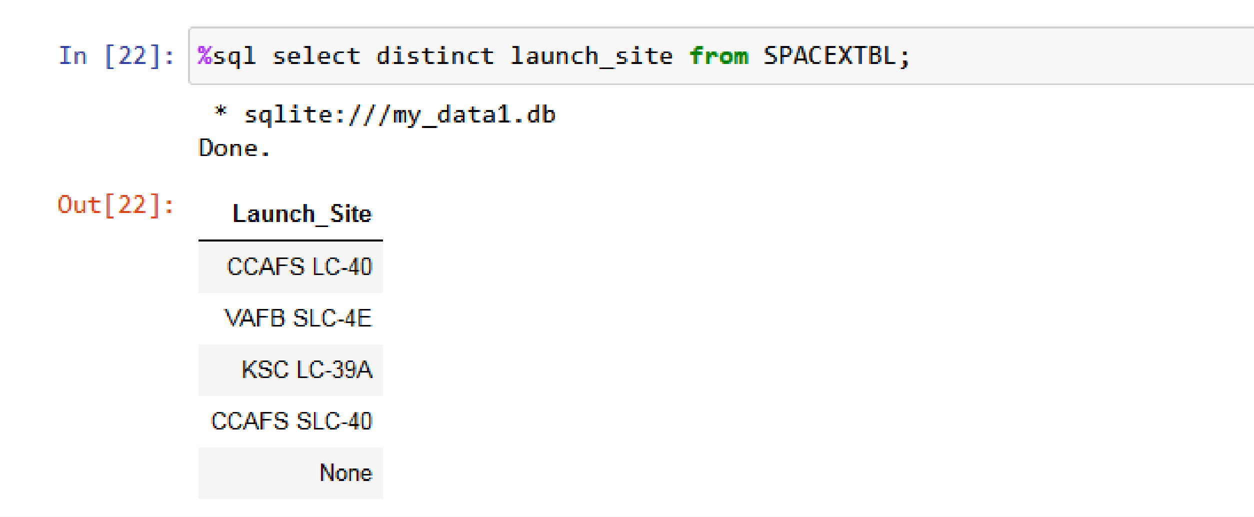
- Sho
  aver



- Show the screenshot of the scatter plot with explanations

All Launch Site Names

Launch Site Names
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E



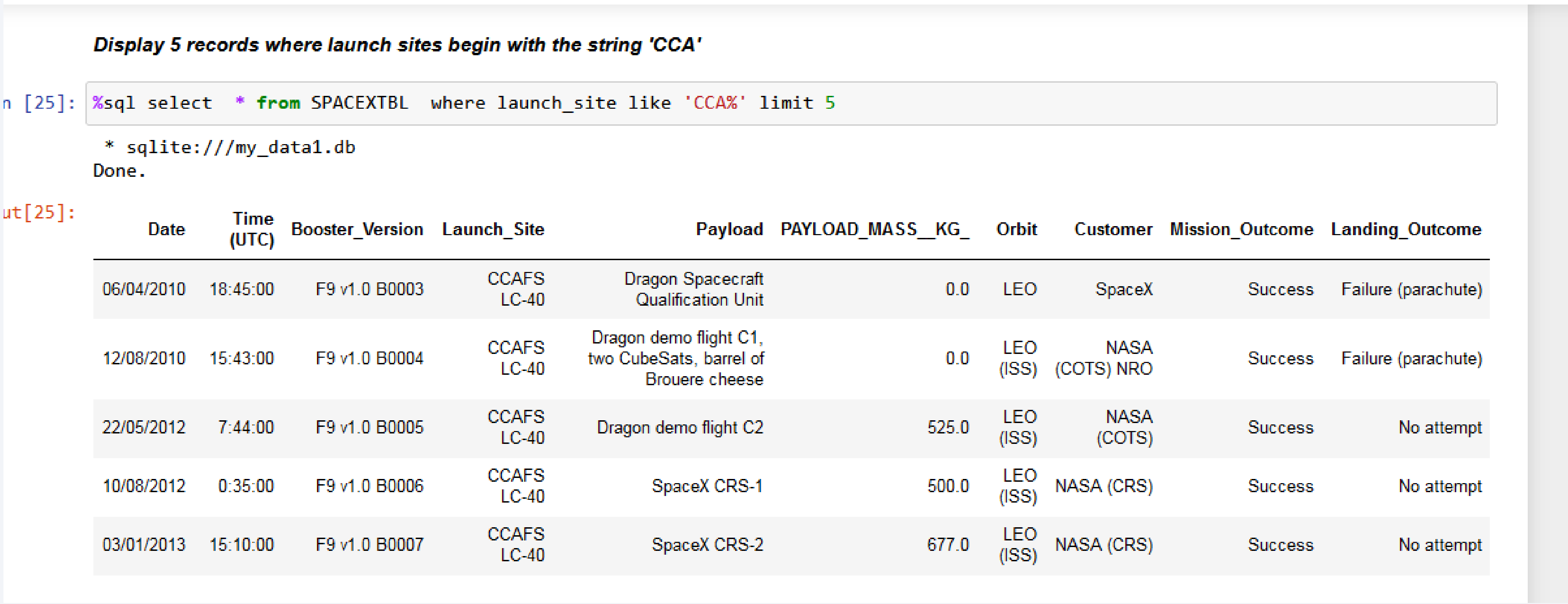
```
In [22]: %sql select distinct launch_site from SPACEXTBL;

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

Out[22]:

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

Launch Site Names Begin with 'CCA'

Records with Launch Site Starting with CCA



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

```
n [25]: %sql select  * from SPACEXTBL  where launch_site like 'CCA%' limit 5
```

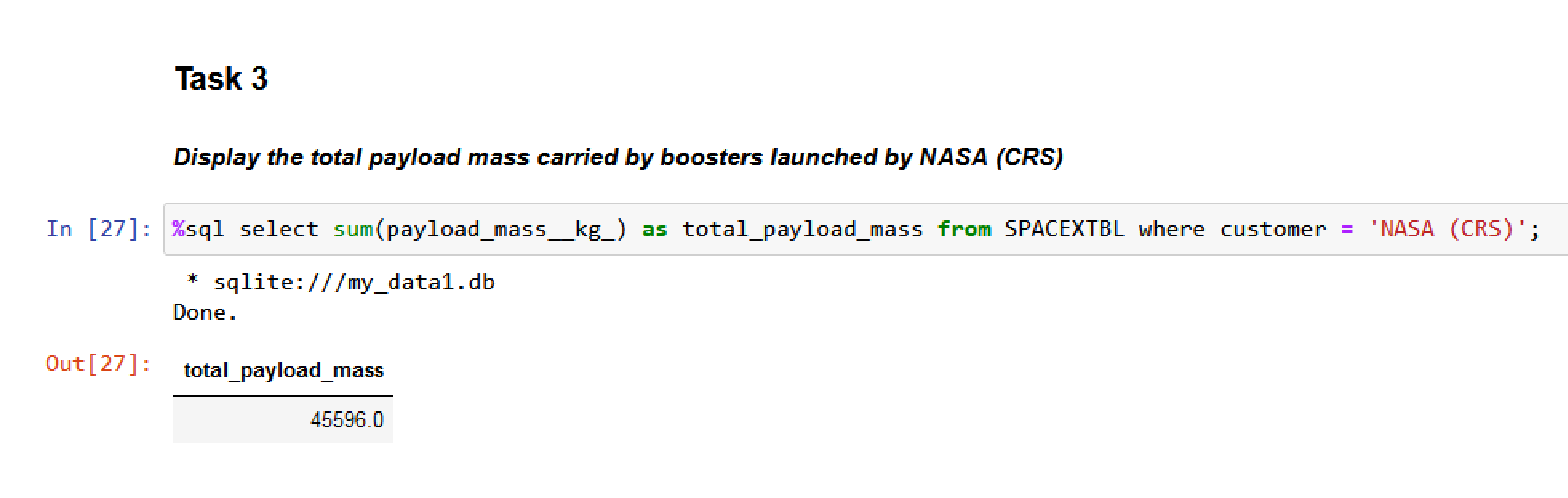
```
 * sqlite:///my_data1.db
Done.
```

ut[25]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total Payload Mass
45,596 kg (total) carried by boosters launched by NASA (CRS)



Task 3

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

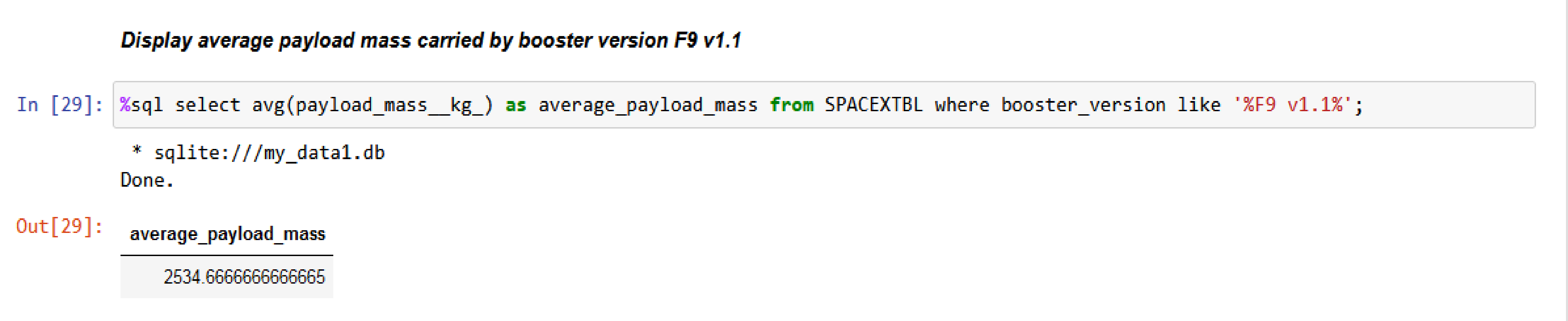
```
In [27]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';

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

Out[27]:

total_payload_mass
45596.0

Average Payload Mass by F9 v1.1

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



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

```
In [29]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL where booster_version like '%F9 v1.1%';
```

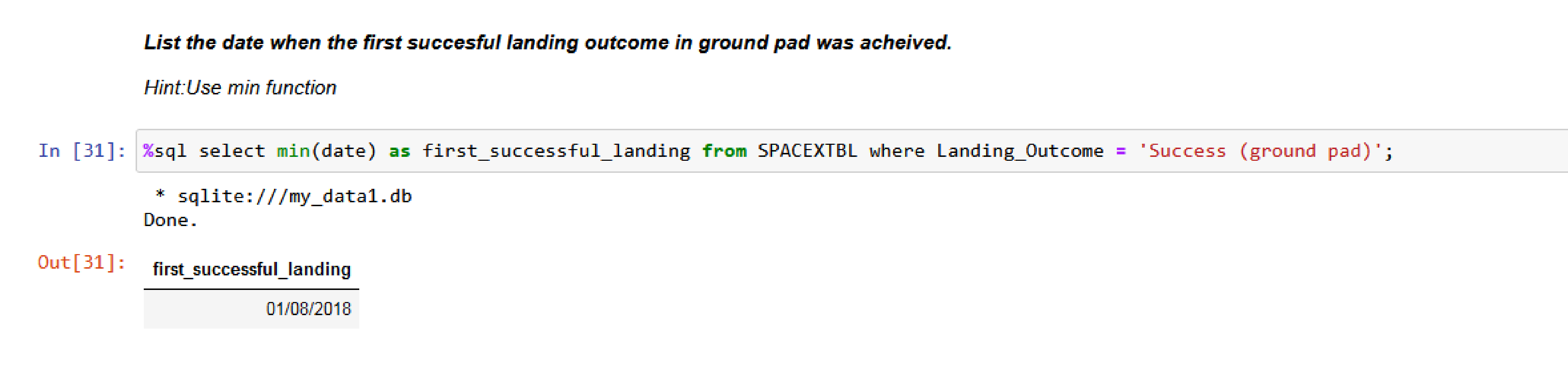
```
 * sqlite:///my_data1.db
Done.
```

Out[29]:

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

Listing the date when the first successful landing outcome in ground pad was achieved



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

Hint:Use min function

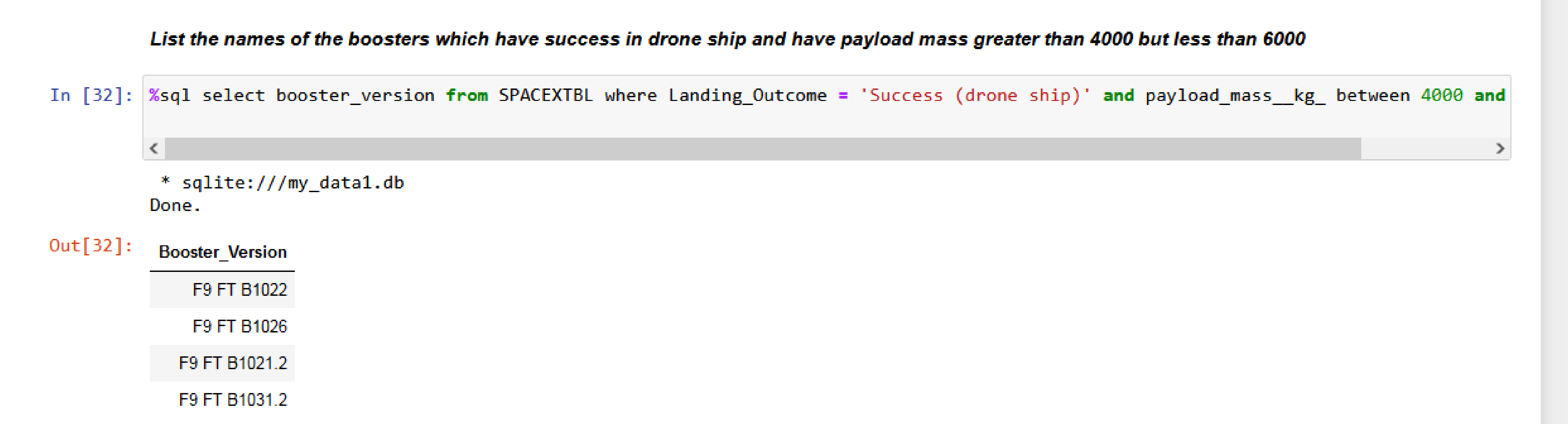
```
In [31]: %sql select min(date) as first_successful_landing from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';

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

Out[31]:

first_successful_landing
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

Listing 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

In [32]:
```
%sql select booster_version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and
```

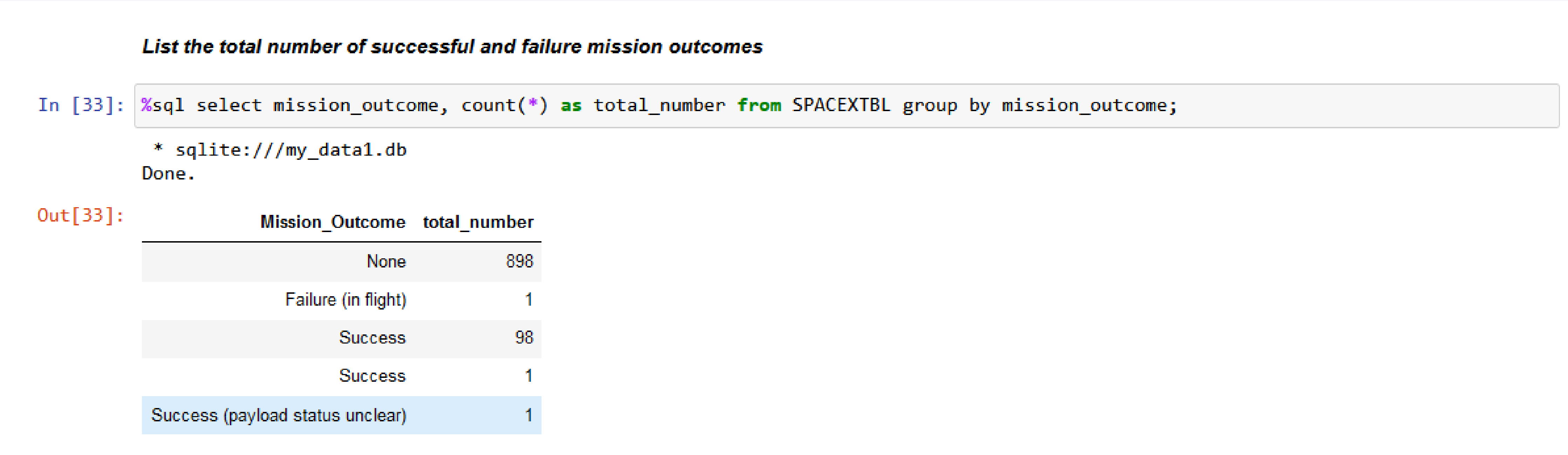
```
 * sqlite:///my_data1.db
Done.
```

Out[32]:

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

Total Number of Successful and Failure Mission Outcomes

Listing the total number of successful and failure mission outcomes.



List the total number of successful and failure mission outcomes

```
In [33]: %sql select mission_outcome, count(*) as total_number from SPACEXTBL group by mission_outcome;
```

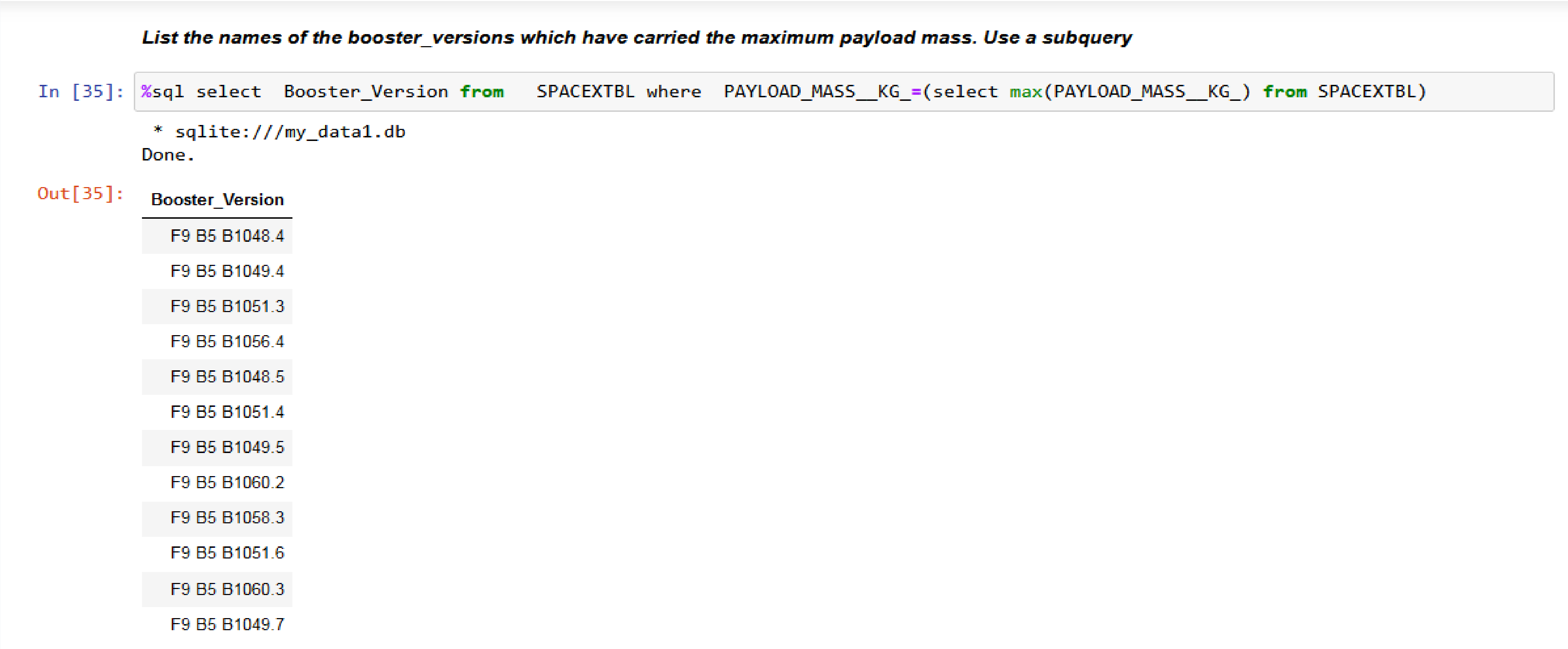
```
 * sqlite:///my_data1.db
Done.
```

Out[33]:

Mission_Outcome	total_number
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Listing the names of the booster versions which have carried the maximum payload mass.



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

```
In [35]: %sql select  Booster_Version from   SPACEXTBL where  PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

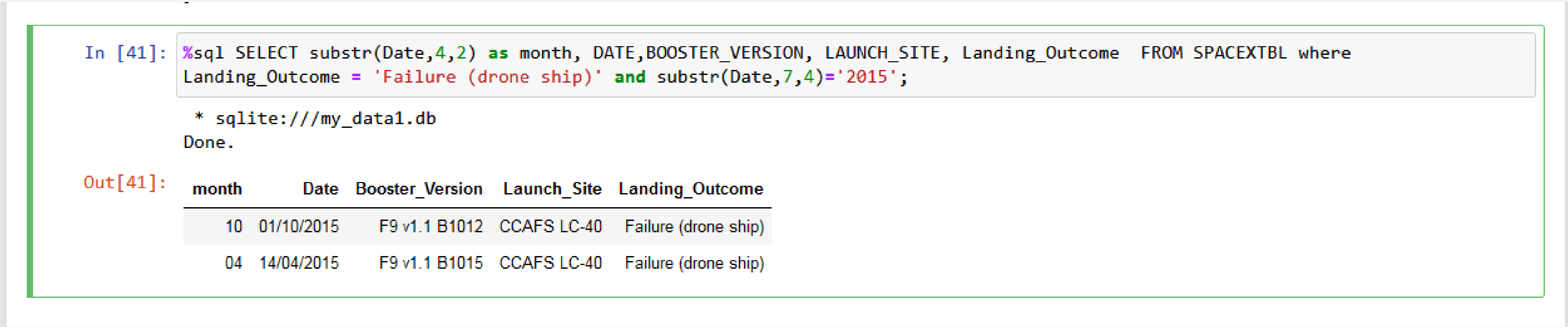
```
 * sqlite:///my_data1.db
Done.
```

Out[35]:

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

2015 Launch Records

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.



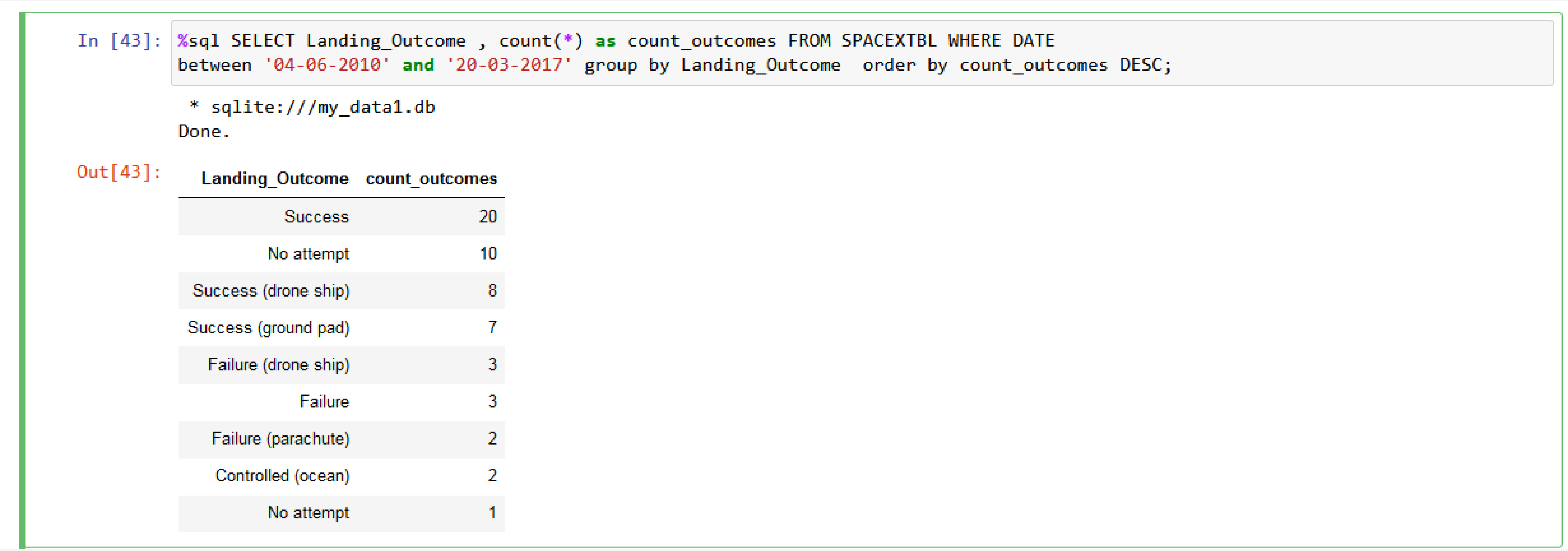
```
In [41]: %sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome  FROM SPACEXTBL where
         Landing_Outcome = 'Failure (drone ship)' and substr(Date,7,4)='2015';

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

Out[41]:

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

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

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



```
In [43]: %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[43]:
```

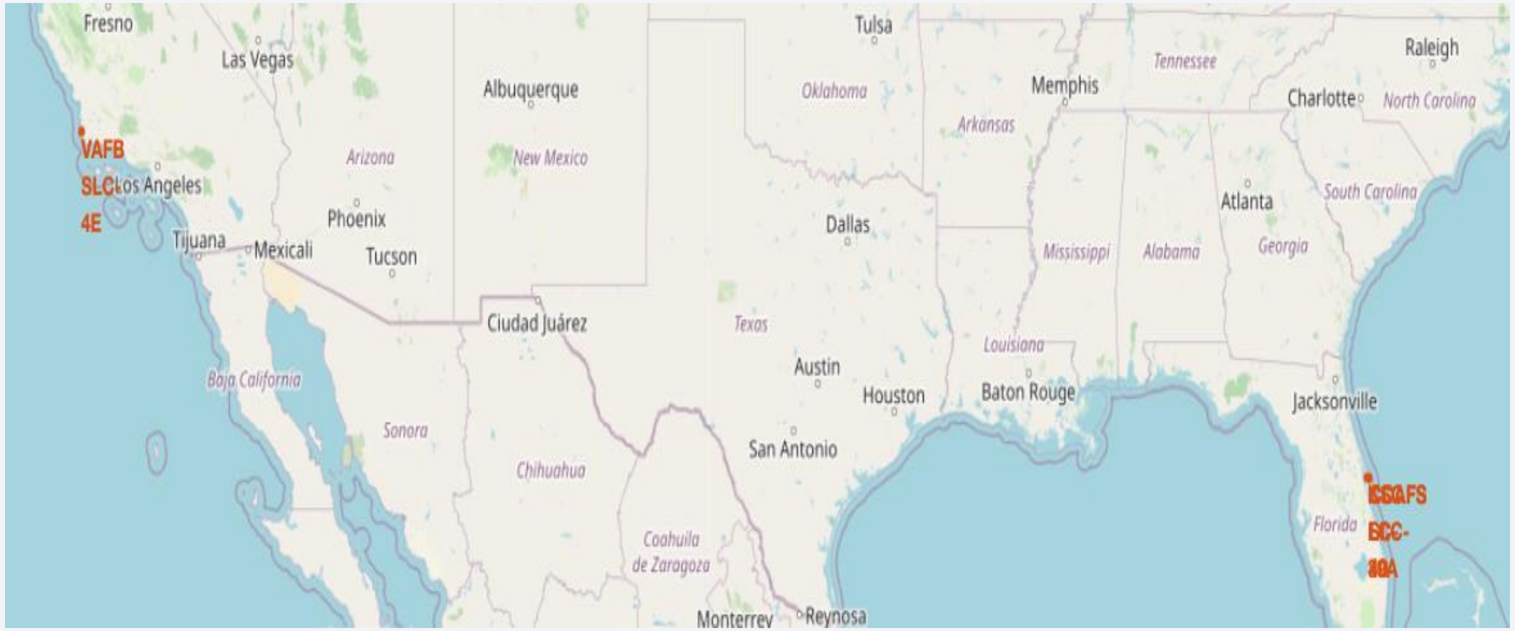
Landing_Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

All launch sites’ location global map

With Markers
Near Equator: the closer the launch site to the equator, the **easier**it is **to launch**to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an **additional natural boost**-due to the rotational speed of earth -that **helps save the cost**of putting in extra fuel and boosters.



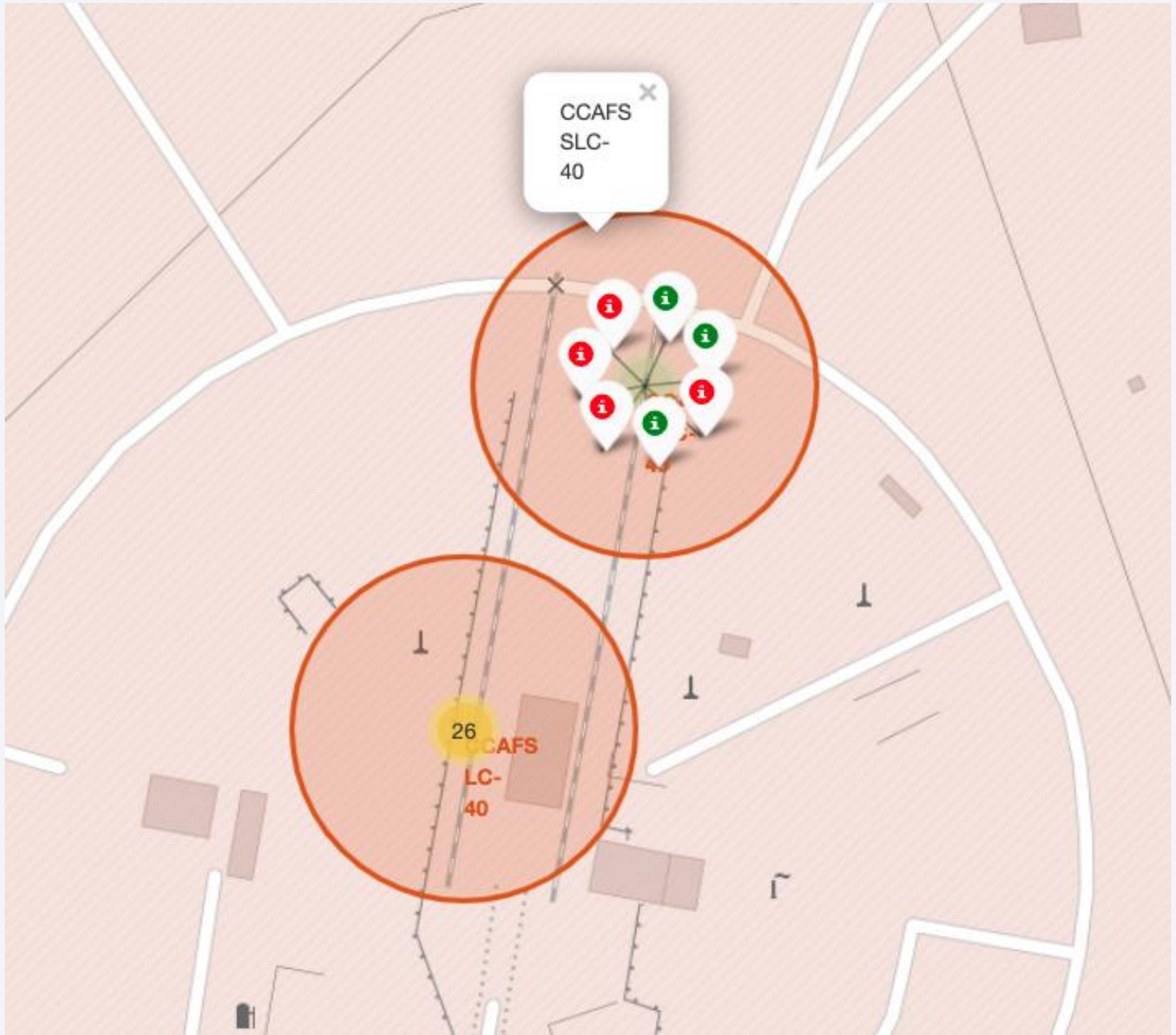
Color-labeled launch records on the map

At Each Launch Site Outcomes:

Green Marker = Successful Launch

Red Marker = Failed Launch

Launch site **CCAFS SLC-40** has a **3/7 success rate** (**42.9%)**

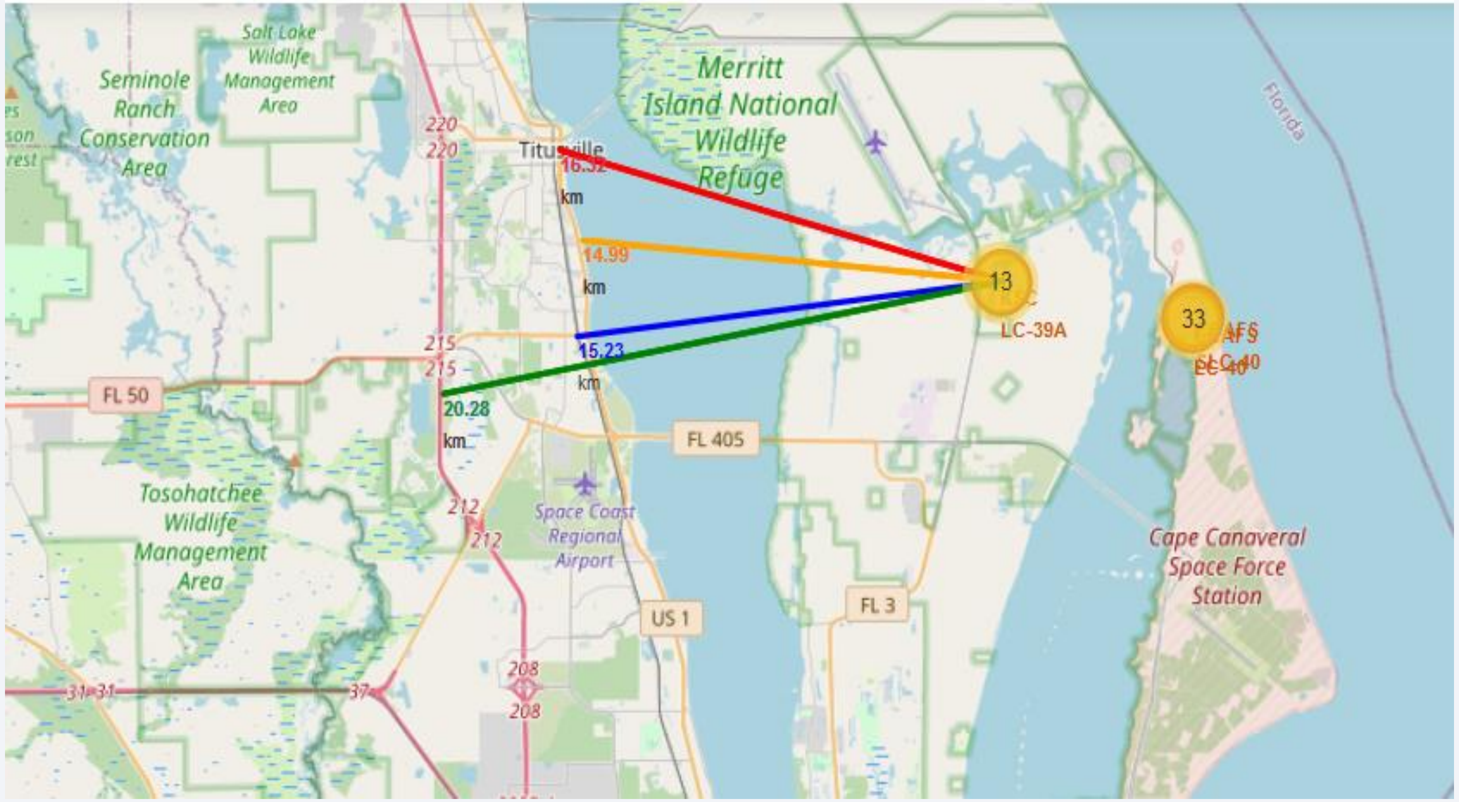


Distance to Proximities

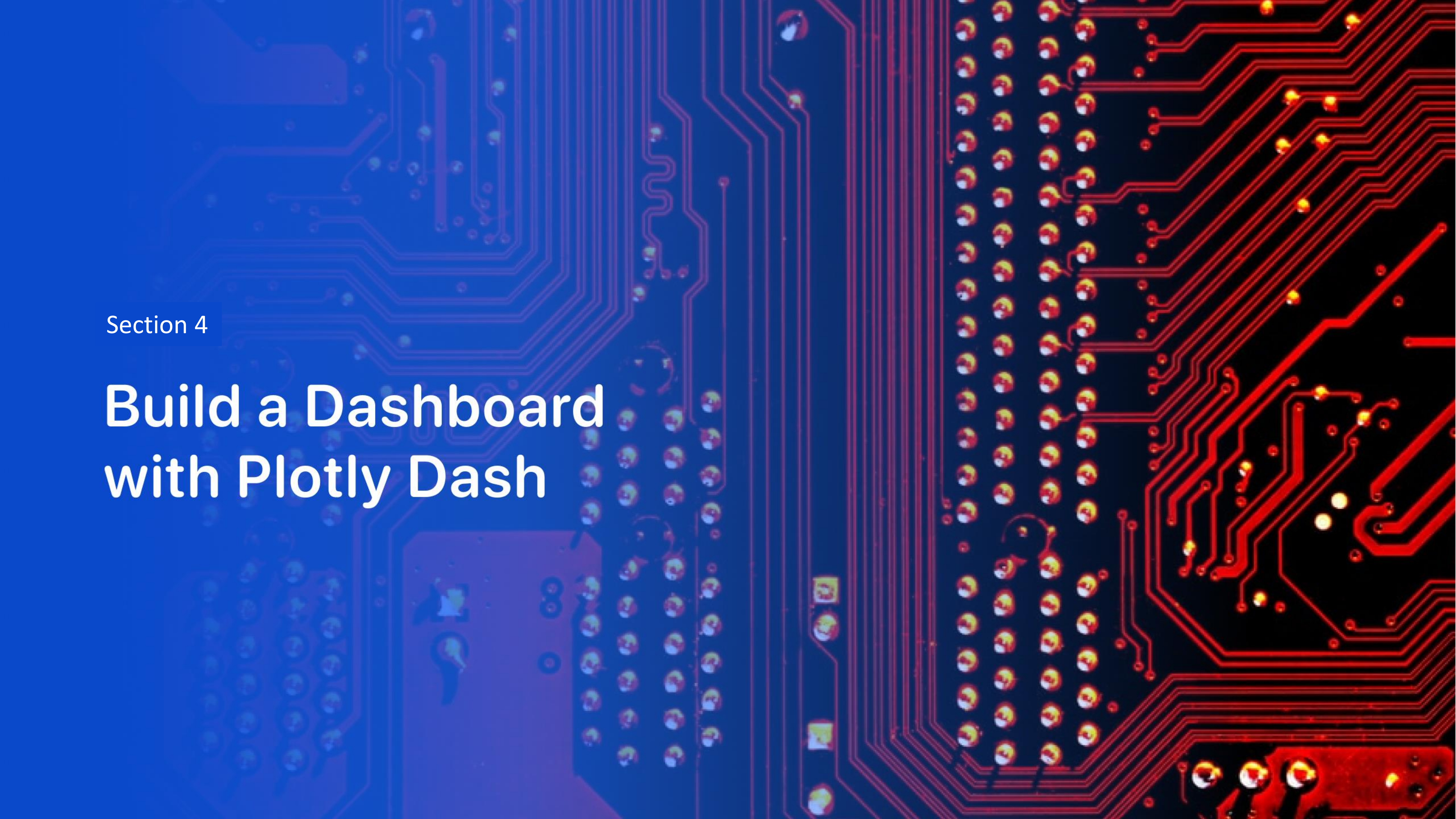
From the visual analysis of the launch site KSC LC-39A we can clearly see that it is

relative close to railway (15.23 km)
relative close to highway (20.28 km)
relative close to coastline (14.99 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.



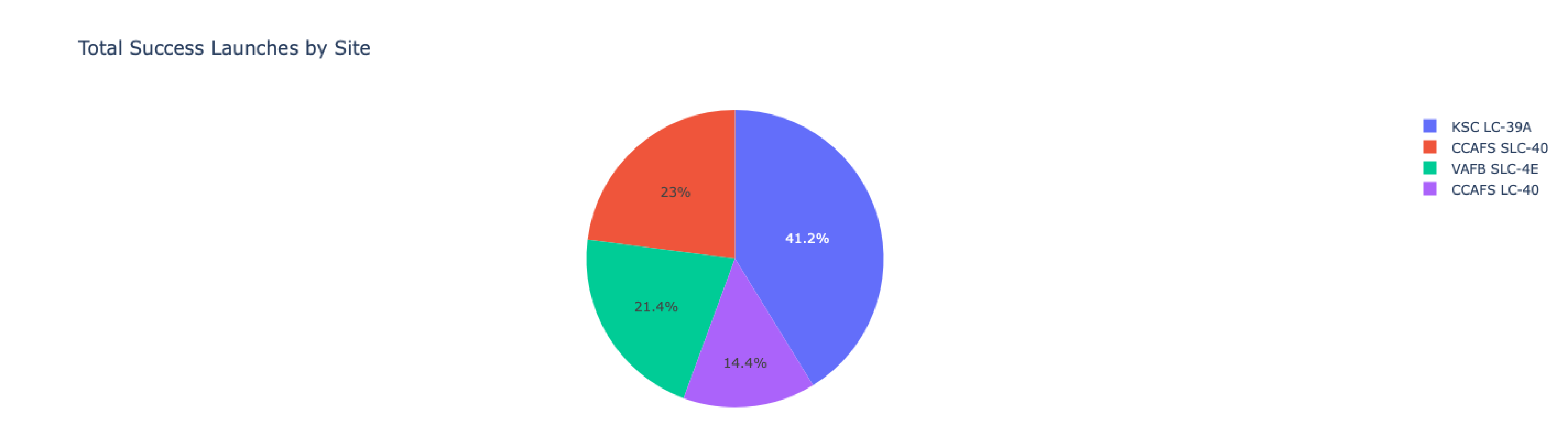
Seminole Ranch Conservation Area
Salt Lake Wildlife Management Area
Merritt Island National Wildlife Refuge
Florida
Titusville
16.32 km
14.99 km
15.23 km
20.28 km
13
LC-39A
33
AFS
SLC-40
LC-40
220
215
212
208
FL 50
FL 405
FL 3
US 1
31
37
Tosohatchee Wildlife Management Area
Space Coast Regional Airport
Cape Canaveral Space Force Station



Section 4

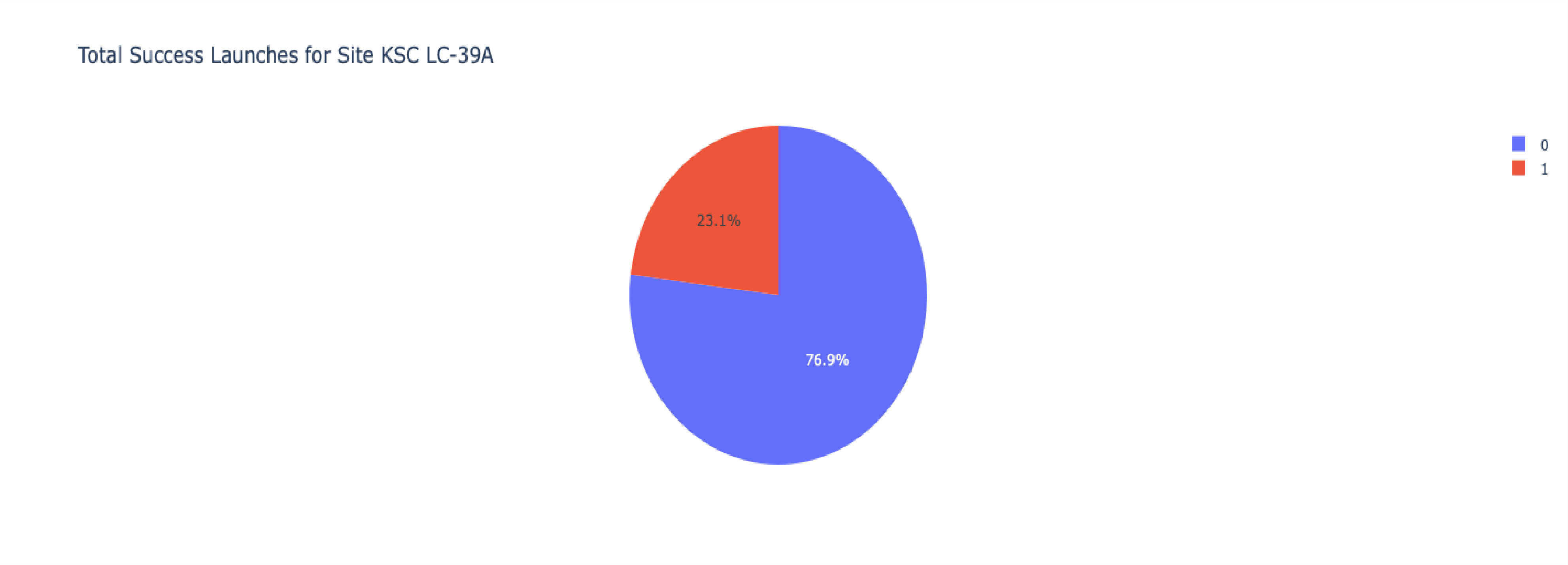
Build a Dashboard with Plotly Dash

Launch Success by Site



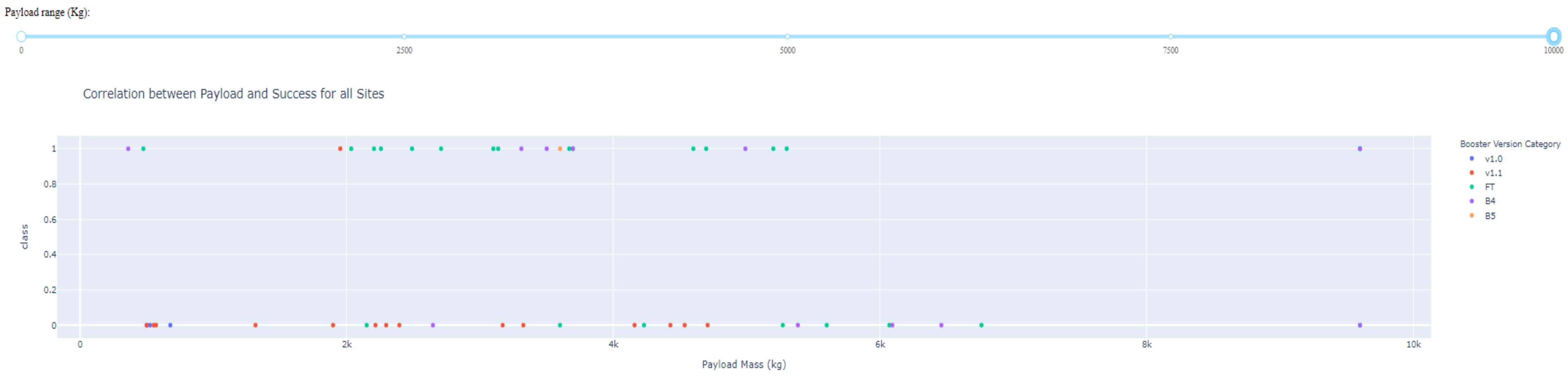
Success as Percent of Total
KSC LC-39A has the **most successful launches** amongst launch sites (**41.2%**)

Launch site with highest launch success ratio



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

Payload Mass and Success



By Booster Version

Payloads between2,000 kg and **5,000 kg** have the **highest success rate**

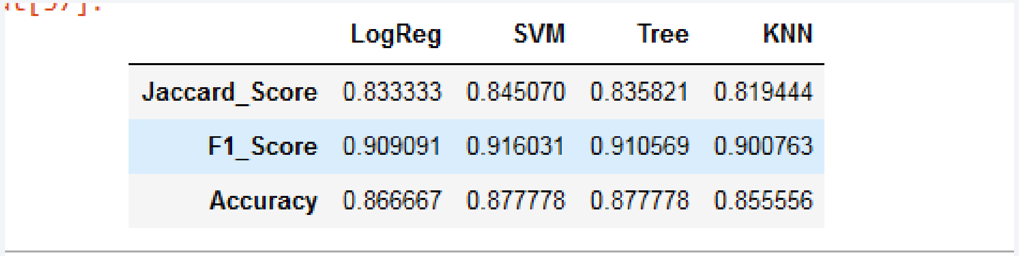
1 indicating successful outcome and 0 indicating an unsuccessful outcome

Section 5

Predictive Analysis (Classification)

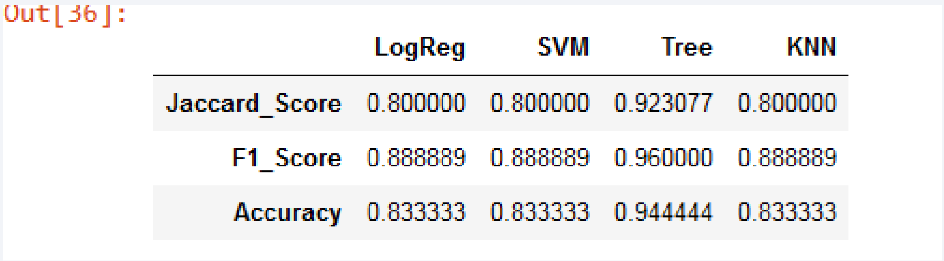
Classification Accuracy

Scores and Accuracy of the Entire Data Set



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.835821	0.819444
F1_Score	0.909091	0.916031	0.910569	0.900763
Accuracy	0.866667	0.877778	0.877778	0.855556

Scores and Accuracy of the Test Set



Out[36]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.923077	0.800000
F1_Score	0.888889	0.888889	0.960000	0.888889
Accuracy	0.833333	0.833333	0.944444	0.833333

Based on the scores of the Test Set, we can not confirm which method performs best.

Same Test Set scores may be due to the small test sample size (18 samples).

Therefore, we tested all methods based on the whole Dataset.

The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

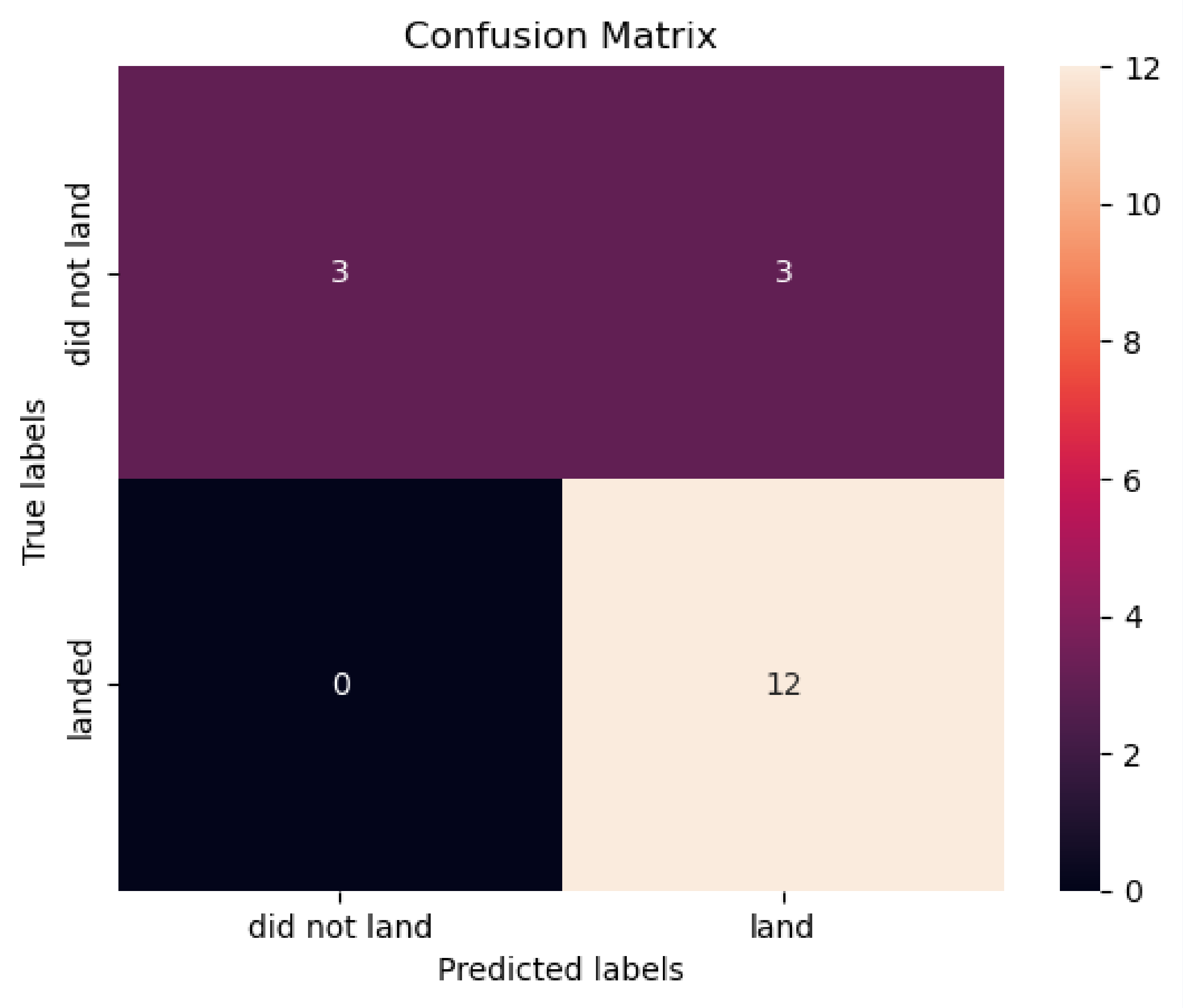
Confusion Matrix

A confusion matrix summarizes the performance of a classification algorithm

•All the confusion matrices were identical

•The fact that there are false positives (Type 1 error) is not good

•Confusion Matrix Outputs:



Conclusions

- **Model Performance**: The models performed similarly on the test set with the decision tree model slightly outperforming
- •**Equator**: Most of the launch sites are near the equator for an additional natural boost -due to the rotational speed of earth –which helps save the cost of putting in extra fuel and boosters
- •**Coast**: All the launch sites are close to the coast
- •**Launch Success**: Increases over time
- •**KSC LC-39A**: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- •**Orbits**: ES-L1, GEO, HEO, and SSO have a 100% success rate
- •**Payload Mass**: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!