IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Edvaldo Junior
August / 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected via API calls and web scrapping. The process of data wrangling cleaned the data and granted only the essential features were left in the dataset. Some useful insights were obtained with the EDA using SQL and some data visualization. Then, the predictive analysis were performed to select the best model for the task.
- Summary of all results
  - KNN, SVM and Logistic Regression got the same score in the test set: 83.333%
  - Best classifier found was Decision Tree, with an accuracy of 94.444%

Introduction

- SpaceX 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- In this capstone, we will predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

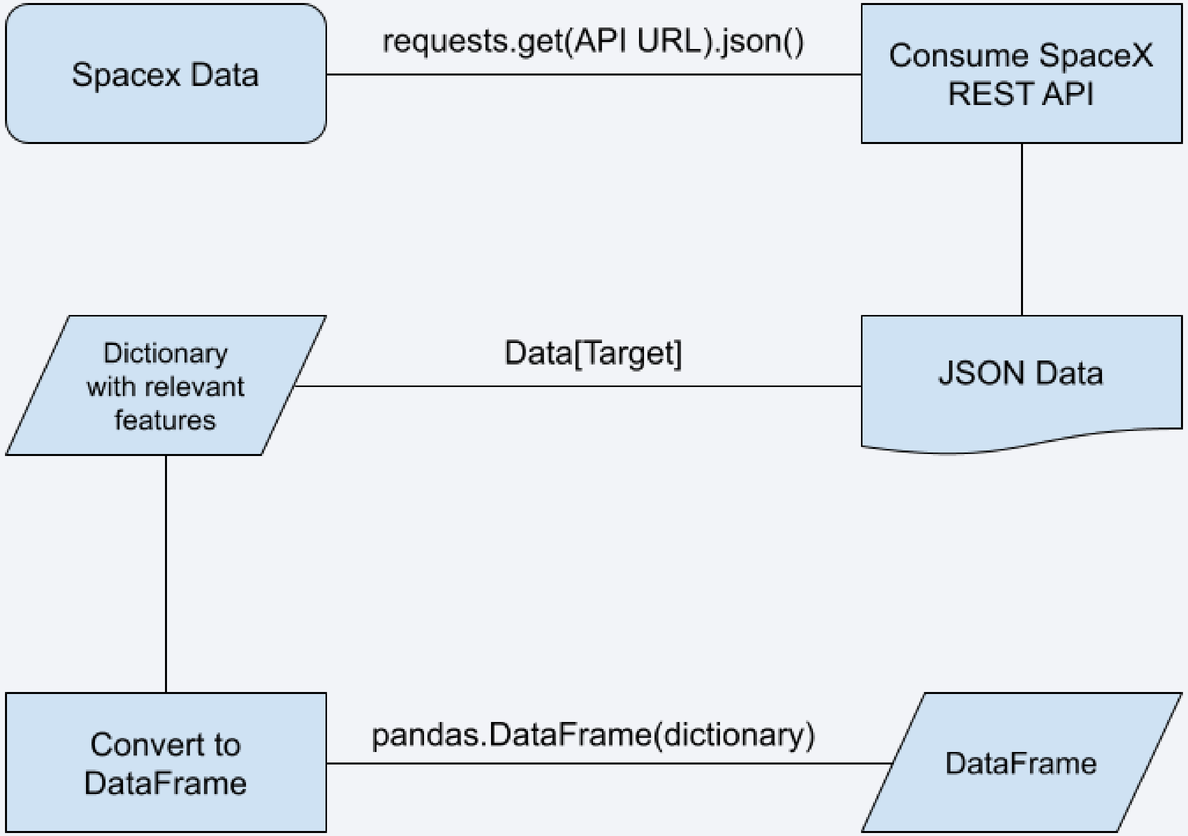
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using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

Data Collection

- One way to get data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome, consists of consuming the SpaceX REST API
- Another popular data source for obtaining launch data is web scraping related Wiki pages
- Those two methods were used and are explained in the next slides

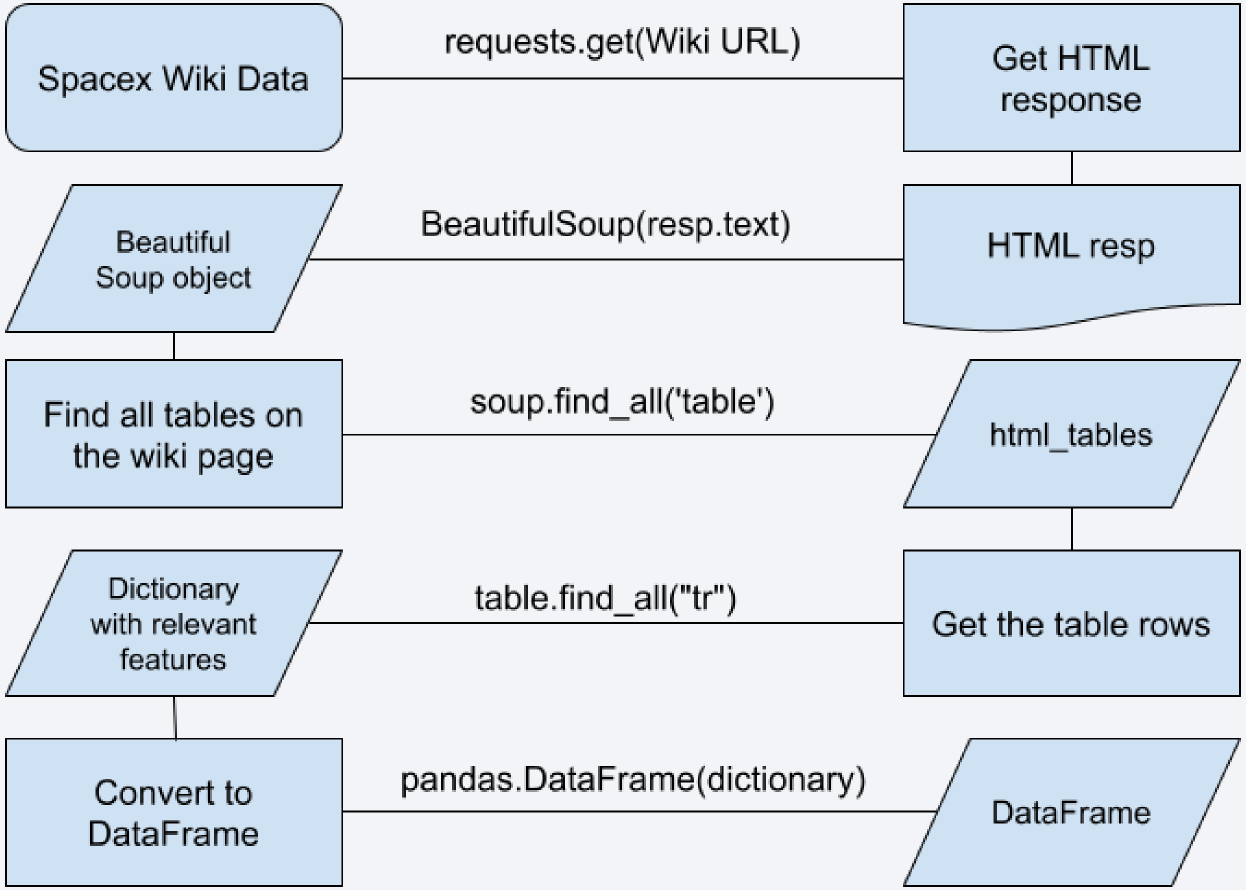
Data Collection – SpaceX API

- Flowchart of SpaceX REST API calls
- https://github.com/edvdjr/spacex-launch-analysis/blob/main/1-spacex-data-collection-api.ipynb


Spacex Data
requests.get(API URL).json()
Consume SpaceX REST API
Dictionary with relevant features
Data[Target]
JSON Data
Convert to DataFrame
pandas.DataFrame(dictionary)
DataFrame

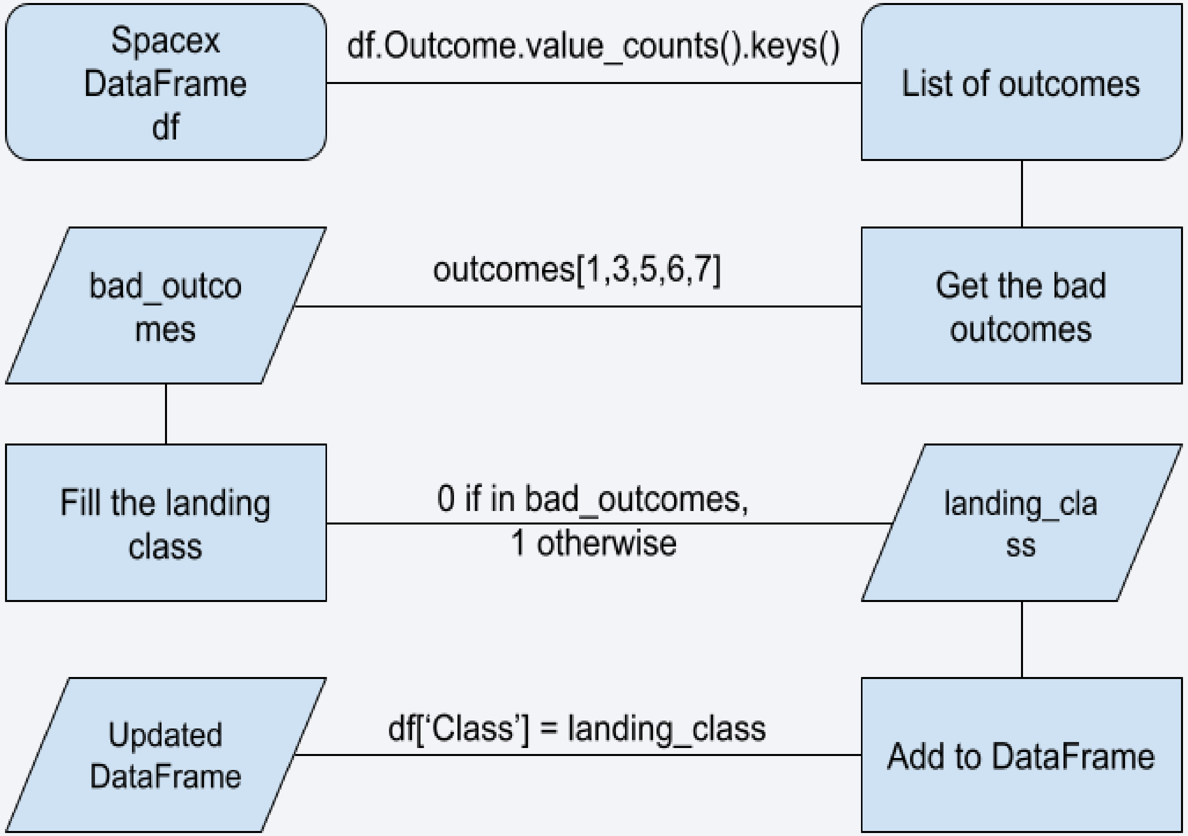
Data Collection – Scraping

- Flowchart of web scraping process
- https://github.com/edvdjr/spacex-launch-analysis/blob/main/2-webscraping.ipynb



Data Wrangling

- The dataset contains several different cases where the booster did not land successfully
- The outcomes of the landings were converted into Training Labels with 1 meaning the booster successfully landed and 0 meaning it was unsuccessful.
- https://github.com/edvdjr/spacex-launch-analysis/blob/main/3-spacex-data-wrangling.ipynb



EDA with Data Visualization

- 4 Scatter Plots to show the relationship between:
  - the Flight Number and the Launch Site
  - the Payload Mass and the Launch Site
  - the Flight Number and the Orbit
  - the Payload Mass and the Orbit
- Bar chart to check success rate vs. orbit type
- Line chart with x axis as the year and y axis as the average success rate, to get the average launch success trend
- https://github.com/edvdjr/spacex-launch-analysis/blob/main/5-eda-dataviz.ipynb

EDA with SQL

- All Launch Site Names
- Launch Site Names Begin with 'CCA'
- Total Payload Mass
- Average Payload Mass by F9 v1.1
- First Successful Ground Landing Date
- Successful Drone Ship Landing with Payload between 4000 and 6000
- Total Number of Successful and Failure Mission Outcomes
- Boosters Carried Maximum Payload
- 2015 Launch Records
- Rank Landing Outcomes Between 2010-06-04 and 2017-03-20
- https://github.com/edvdjr/spacex-launch-analysis/blob/main/4-eda-sqllite.ipynb

Build an Interactive Map with Folium

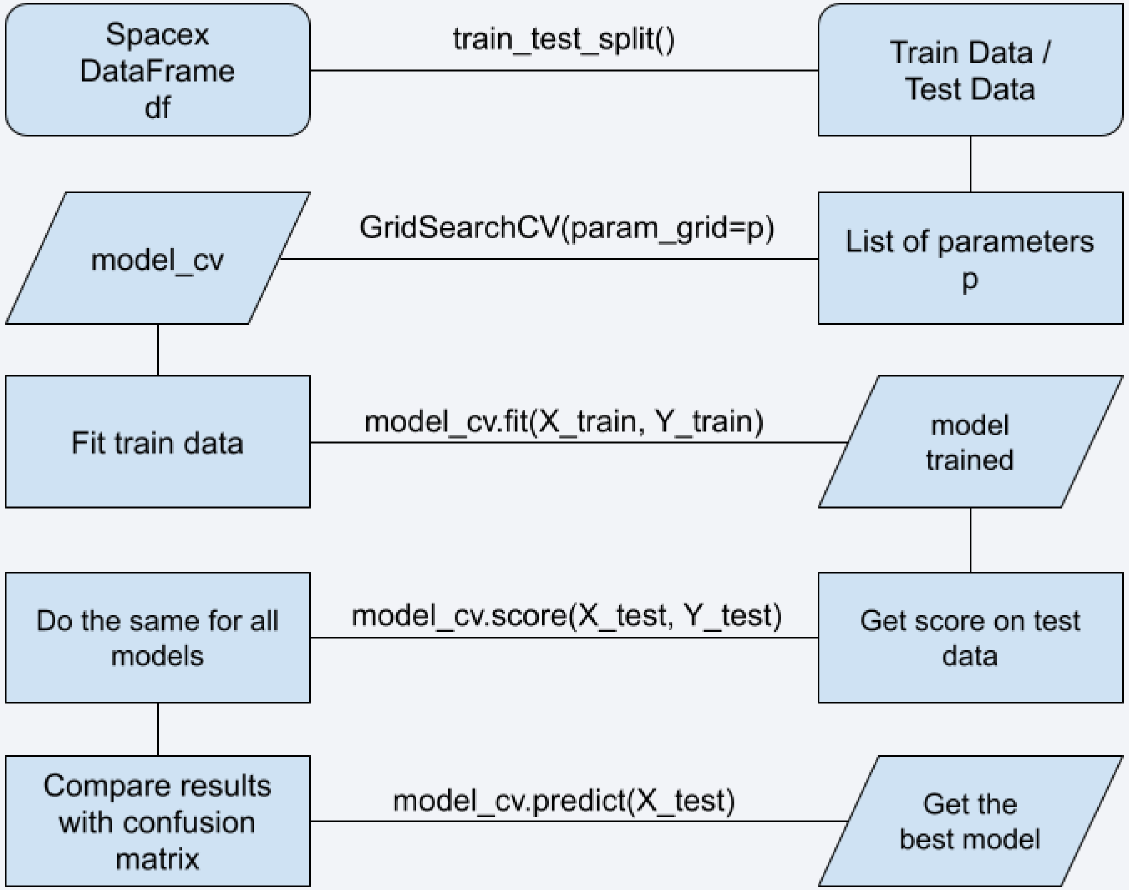
- Many map objects were added to a folium map to make it easier to see important geographic features of launch sites.
- Markers and Circles are used to easily locate the sites on the map;
- Colored icons inside a Marker Cluster indicate the success/failed launches for each site;
- A Polyline draws a line between a launch site to its closest city, railway and highway, so we can see how close the sites are to those relevant points;
- https://github.com/edvdjr/spacex-launch-analysis/blob/main/6-launch-site-location.ipynb

Build a Dashboard with Plotly Dash

- A Launch Site Dropdown Input Component;
- A Pie chart for success rate based on the selected site dropdown;
- A Range Slider to Select Payload;
- A scatter plot Payload x Launch Outcome, so we can visually observe how payload may be correlated with mission outcomes for selected site(s);
- https://github.com/edvdjr/spacex-launch-analysis/blob/main/7-spacex-dash-app.py

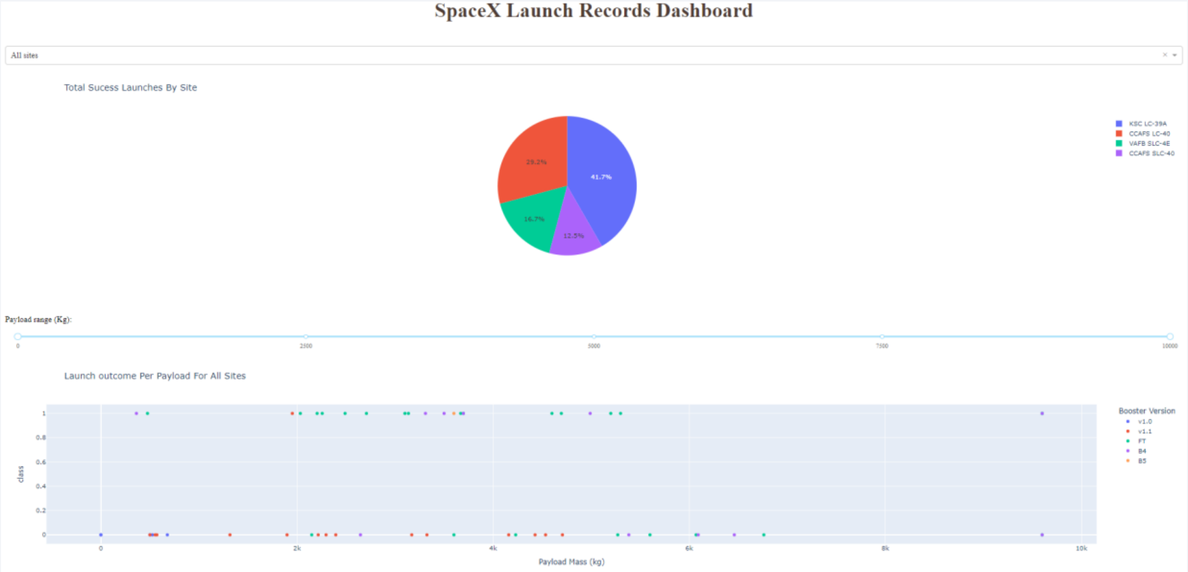
Predictive Analysis (Classification)

- Split the data into train and test
- Use data to train, get the best parameters for each model and compare their results to get the best classifier for this task
- https://github.com/edvdjr/spacex-launch-analysis/blob/main/8-Prediction.ipynb



Results

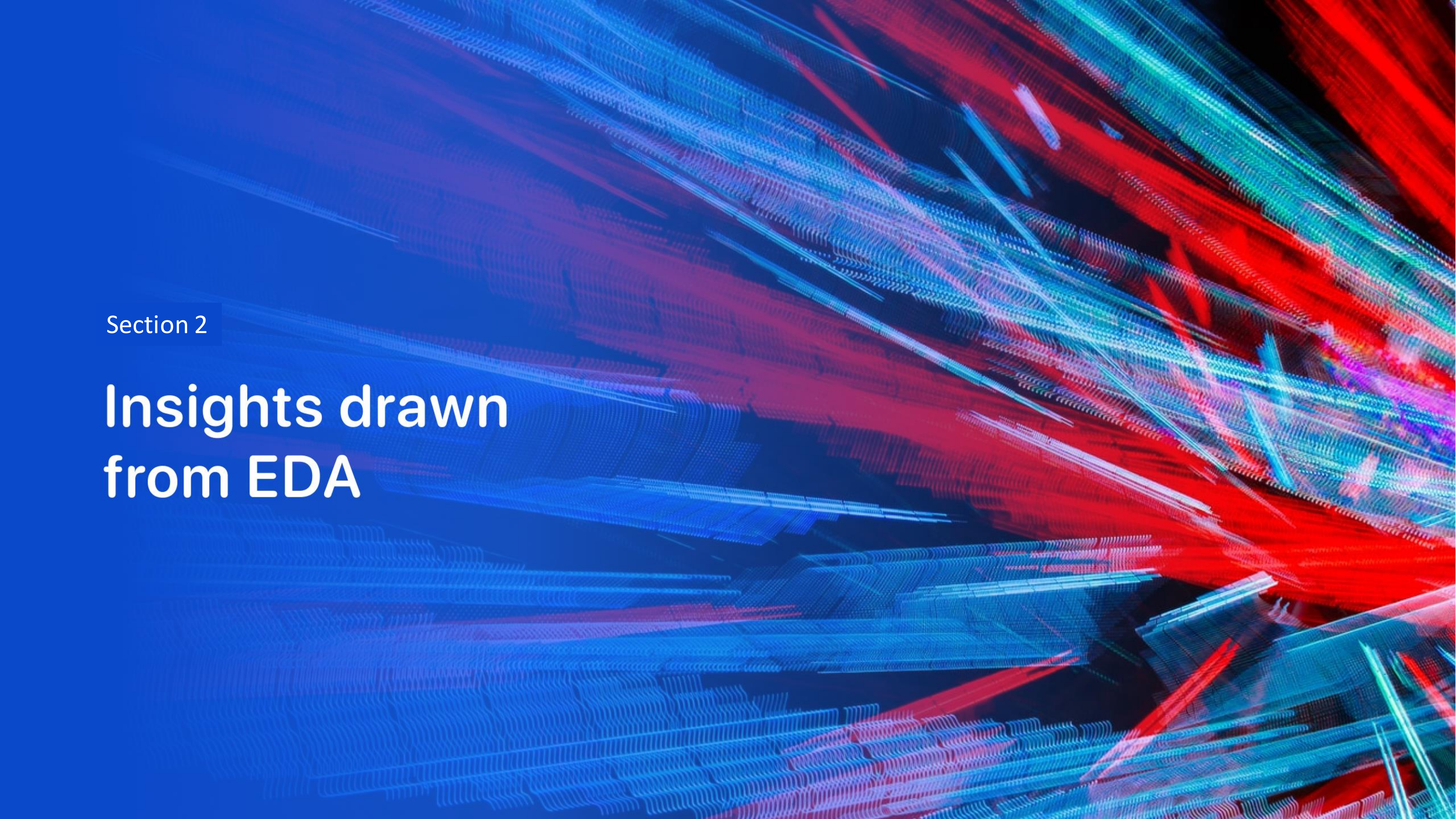
- Exploratory data analysis revealed some relationship among some features, like between Payload mass and Launch site;
- The interactive analytics allows easy verification of specific data in the dataset;
- The predictive analysis found the best Hyperparameters for SVM, Classification Tree, Logistic Regression and KNN, then found the method that performs best using test data.



SpaceX Launch Records Dashboard

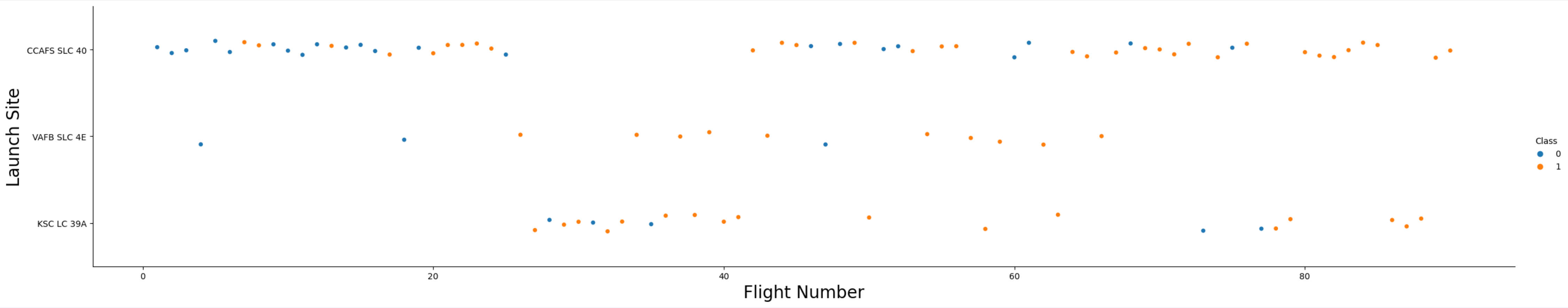
Section 2

Insights drawn from EDA



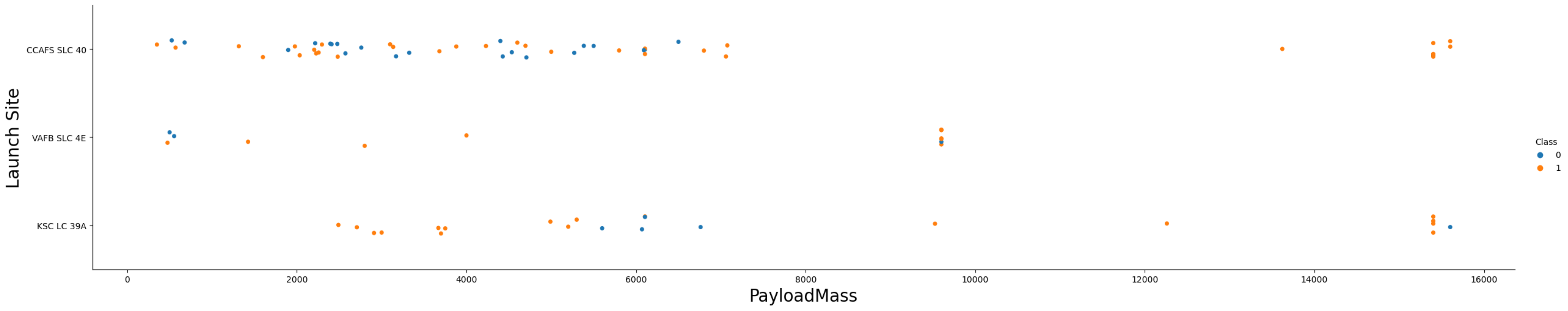
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site
- For the site KSC LC 39A, a flight number greater than or equal to 40 means a success rate of about 85%; For the site VAFB SLC 4E, a Flight number greater than 20 means 90% of success landing; For CCAFS SLC 40, all the landings succeeded for flight numbers greater than 80



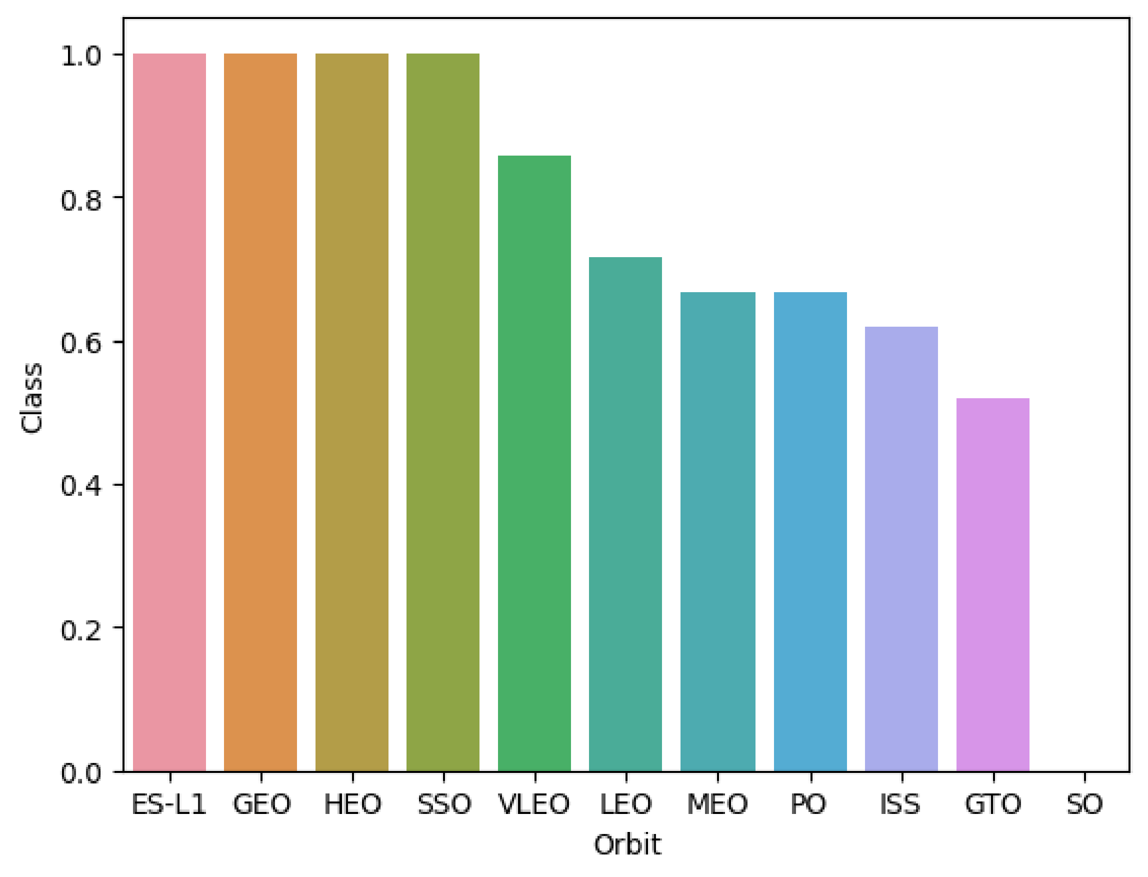
Payload vs. Launch Site

- Scatter Plot to show the relationship between the Launch Site and the Payload Mass;
- Almost all the landings with a Payload Mass greater than 7000 Kg were successful
- We can find that for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).



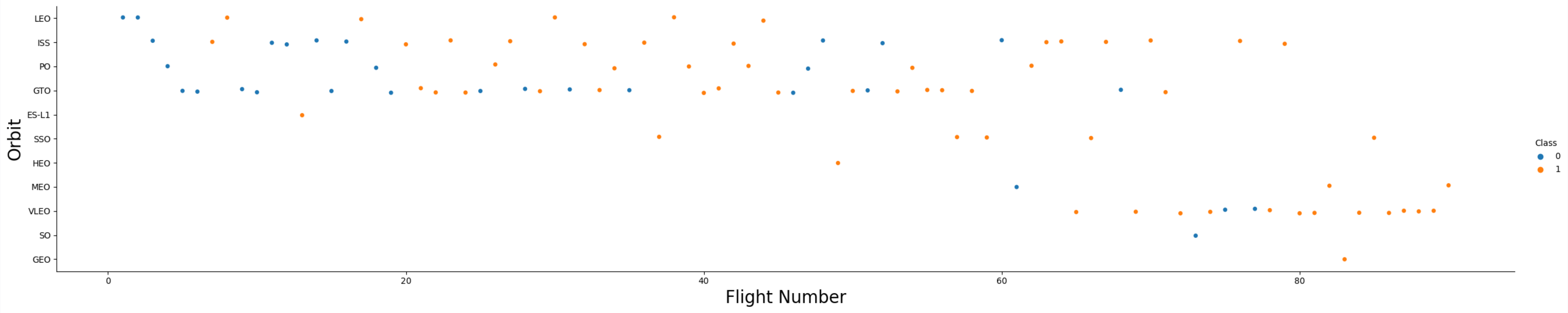
Success Rate vs. Orbit Type

- Bar chart to check if there are any relationship between success rate and orbit type
- ES-L1, GEO, HEO and SSO had 100% success rate, indicating that they are very promising orbits for launches



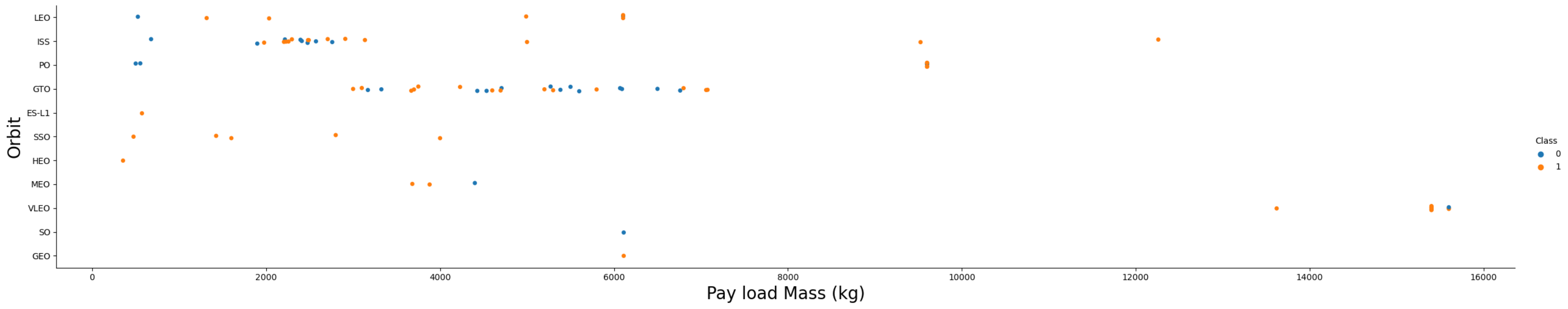
Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type
- We can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

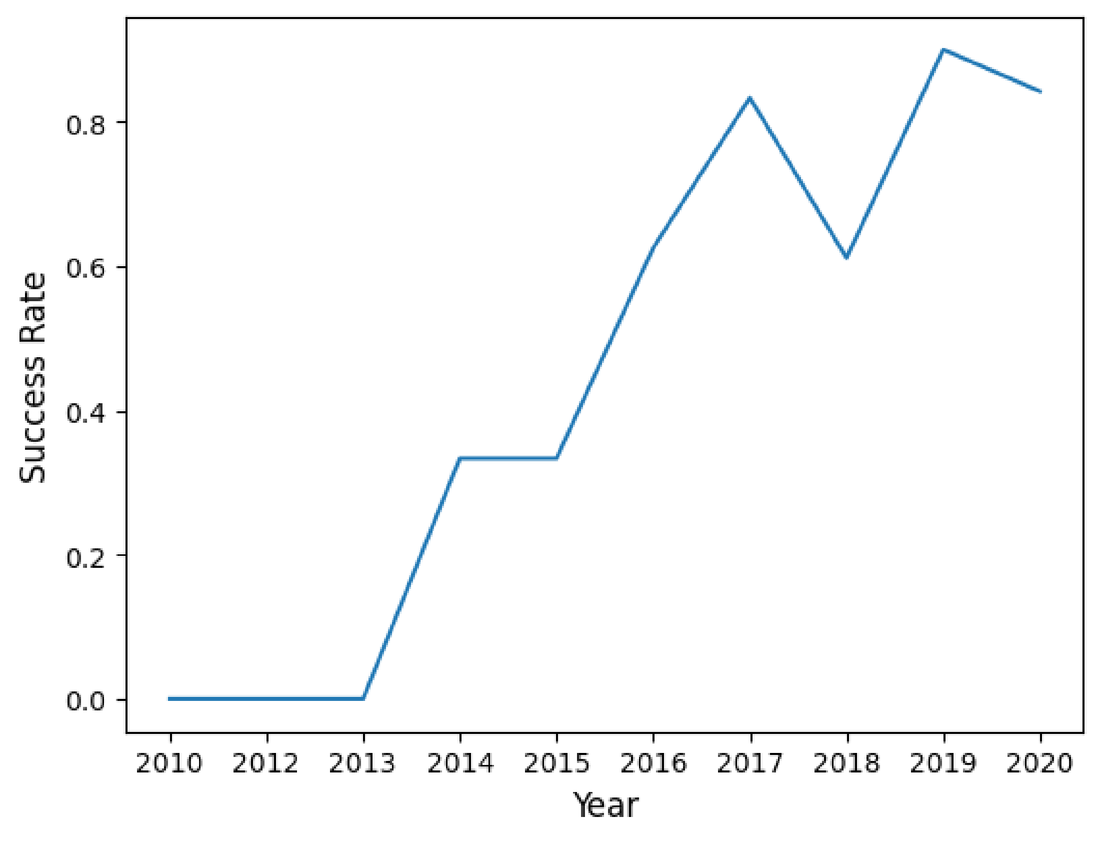
- Show a scatter point of payload vs. orbit type
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there.



Launch Success Yearly Trend

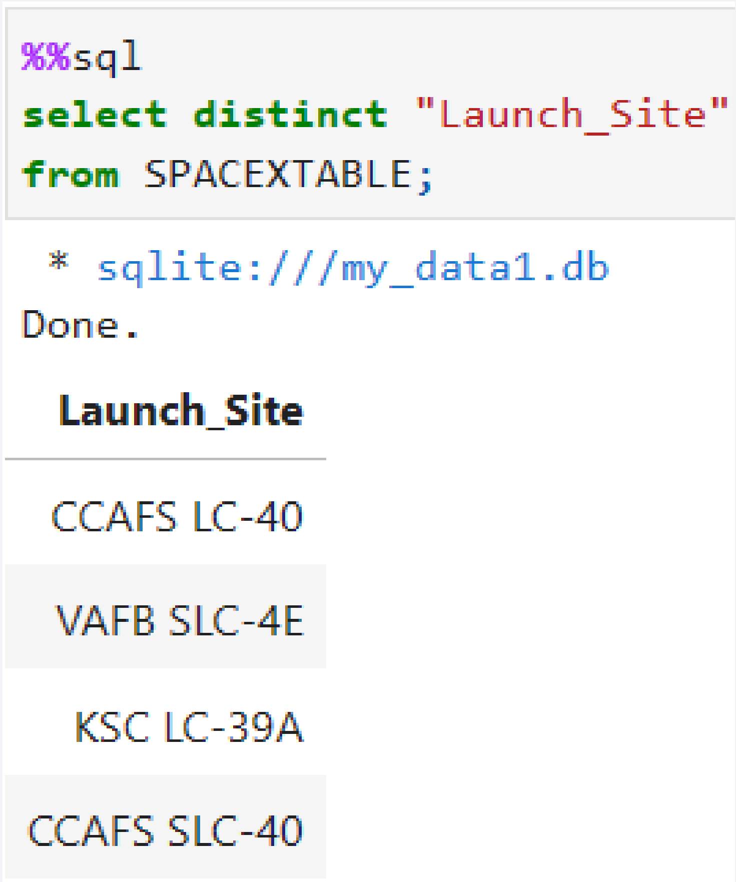
- Line chart with x axis as the year and y axis as the average success rate, to get the average launch success trend
- We can observe that the success rate greatly improved since 2013 till 2020
- GitHub URL of the completed EDA with data visualization notebook:

[spacex-launch-analysis/5-eda-dataviz.ipynb at main · edvdjr/spacex-launch-analysis (github.com)]()



All Launch Site Names

- Find the names of the unique launch sites
- The SELECT DISTINCT statement is used to return only different values



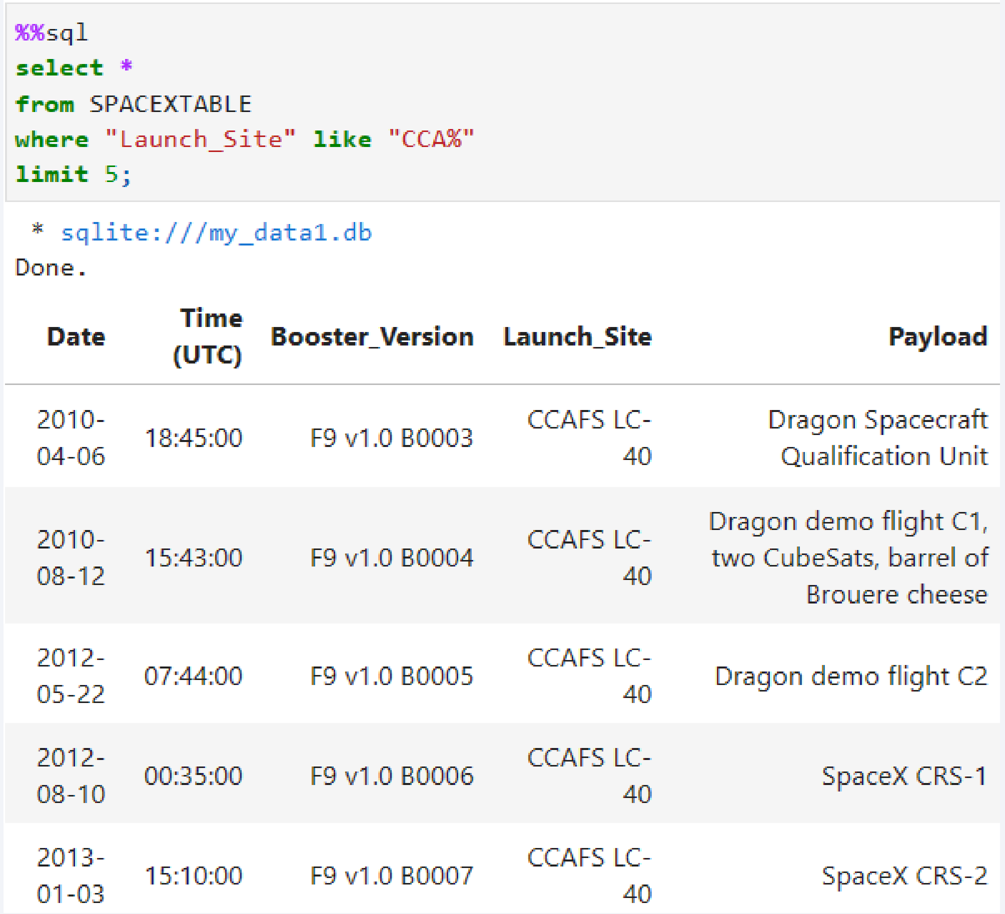
```
%%sql
select distinct "Launch_Site"
from SPACEXTABLE;
```

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

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Wildcard characters are used with the LIKE operator. The LIKE operator is used in a WHERE clause to search for a specified pattern in a column. The % wildcard represents any number of characters.



```
%%sql
select *
from SPACEXTABLE
where "Launch_Site" like "CCA%"
limit 5;
```

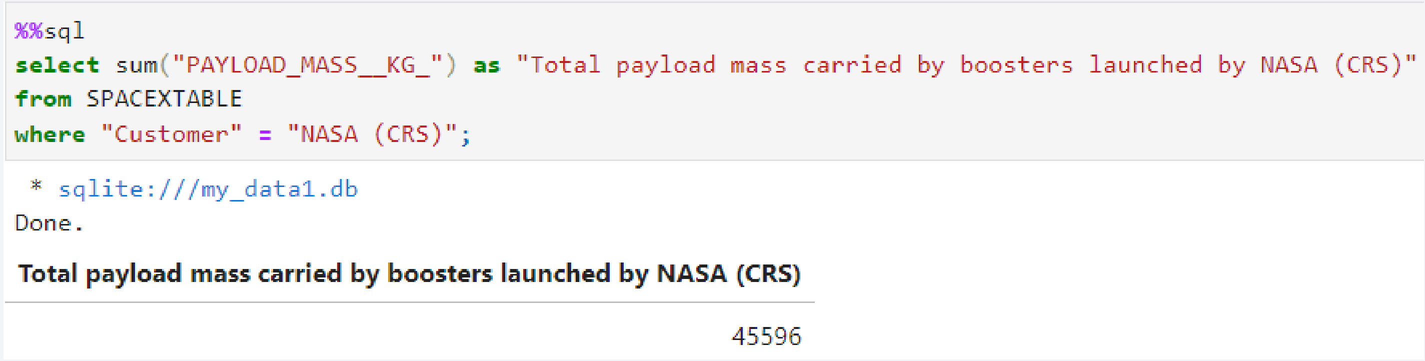
```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

- The SUM() function returns the total sum of a numeric column.



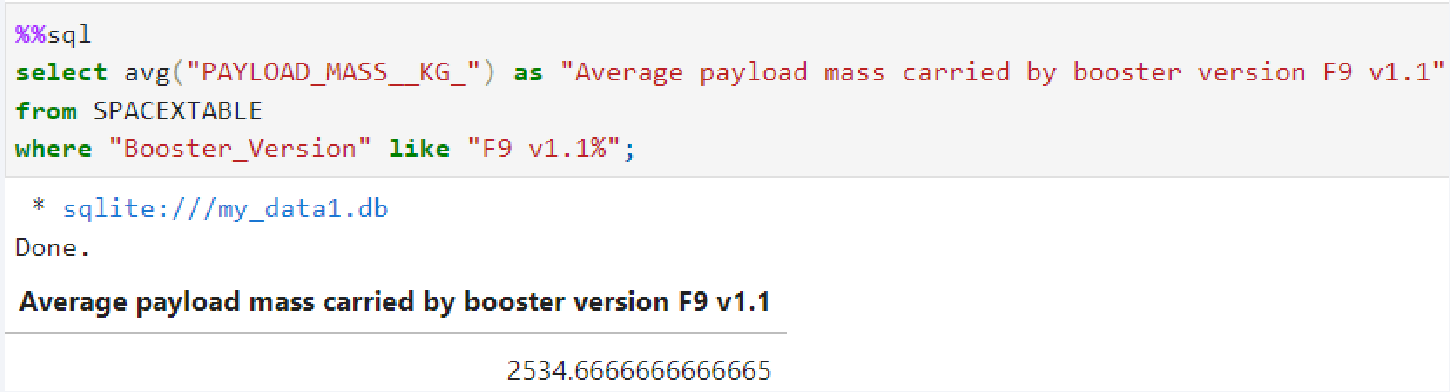
```
%%sql
select sum("PAYLOAD_MASS__KG_") as "Total payload mass carried by boosters launched by NASA (CRS)"
from SPACEXTABLE
where "Customer" = "NASA (CRS)";
```

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

Total payload mass carried by boosters launched by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- The AVG() function returns the average value of a numeric column.



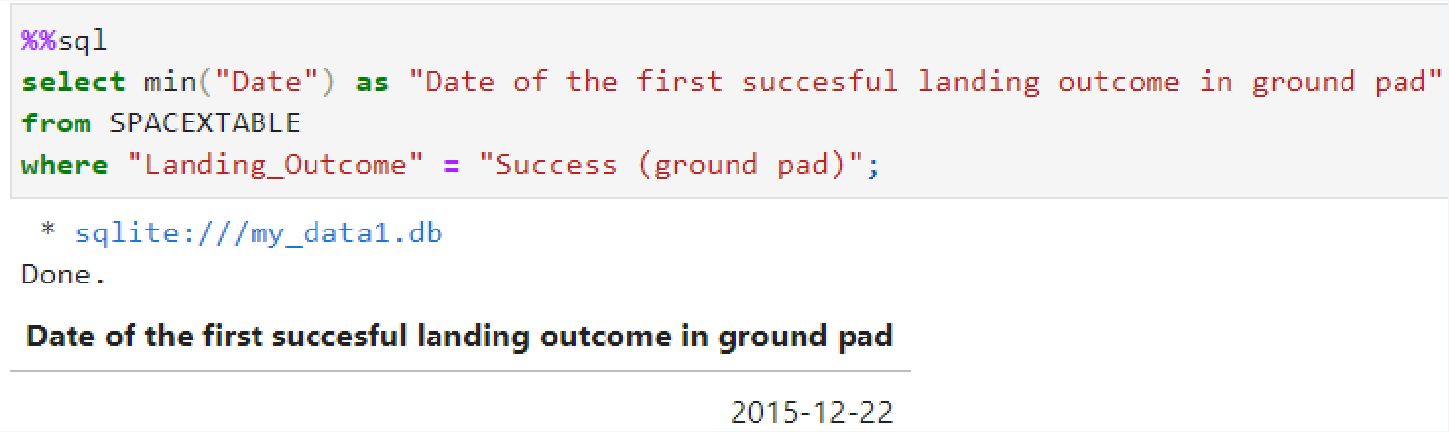
```
%%sql
select avg("PAYLOAD_MASS__KG_") as "Average payload mass carried by booster version F9 v1.1"
from SPACEXTABLE
where "Booster_Version" like "F9 v1.1%";
```

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

Average payload mass carried by booster version F9 v1.1
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The MIN() function returns the smallest value of the selected column.



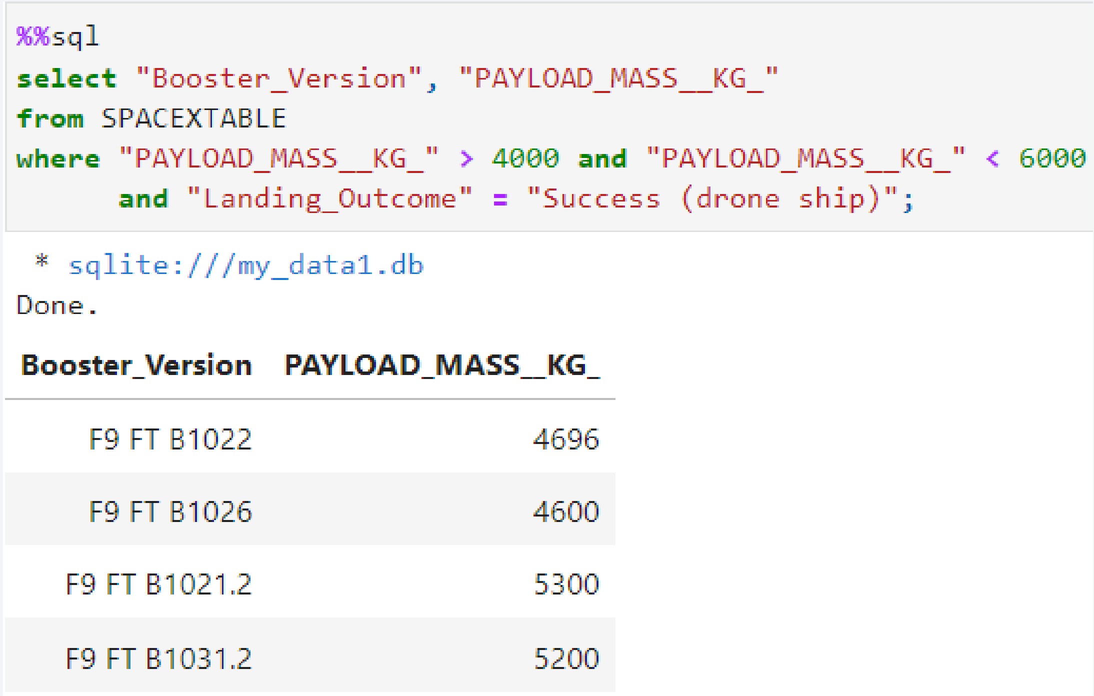
```
%%sql
select min("Date") as "Date of the first succesful landing outcome in ground pad"
from SPACEXTABLE
where "Landing_Outcome" = "Success (ground pad)";
```

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

Date of the first succesful landing outcome in ground pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- The AND operator displays a record if all the conditions are TRUE.



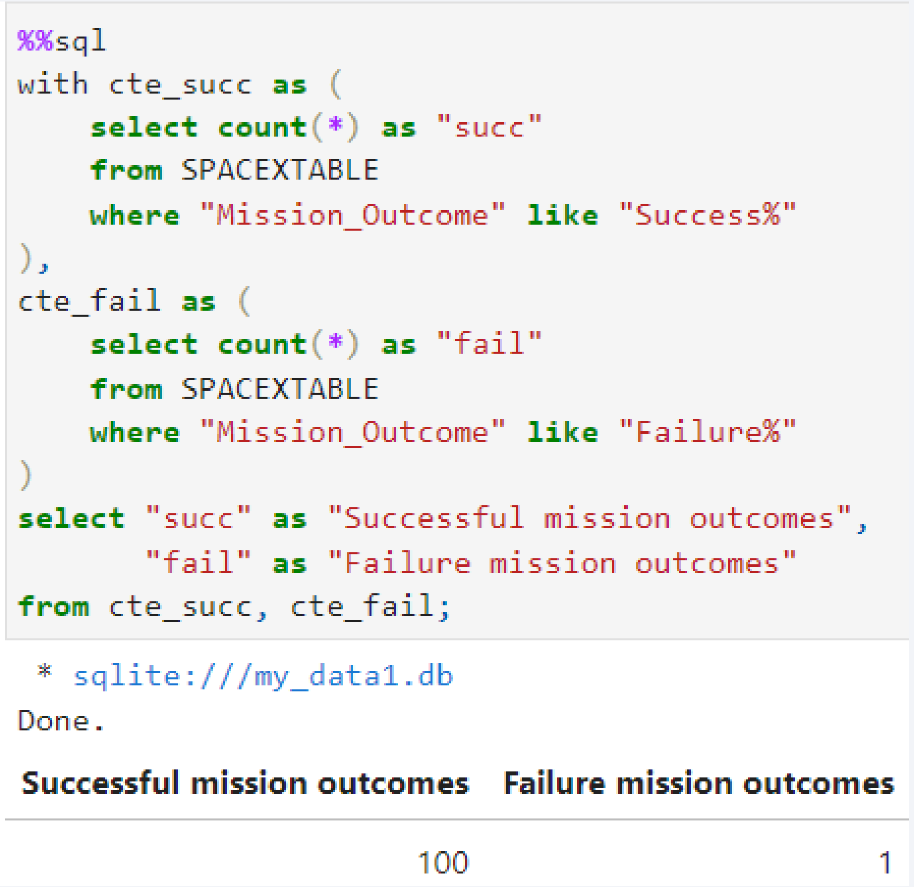
```
%%sql
select "Booster_Version", "PAYLOAD_MASS__KG_"
from SPACEXTABLE
where "PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000
      and "Landing_Outcome" = "Success (drone ship)";
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- A common table expression, or CTE, is a temporary named result set created from a simple SELECT statement that can be used in a subsequent SELECT statement.



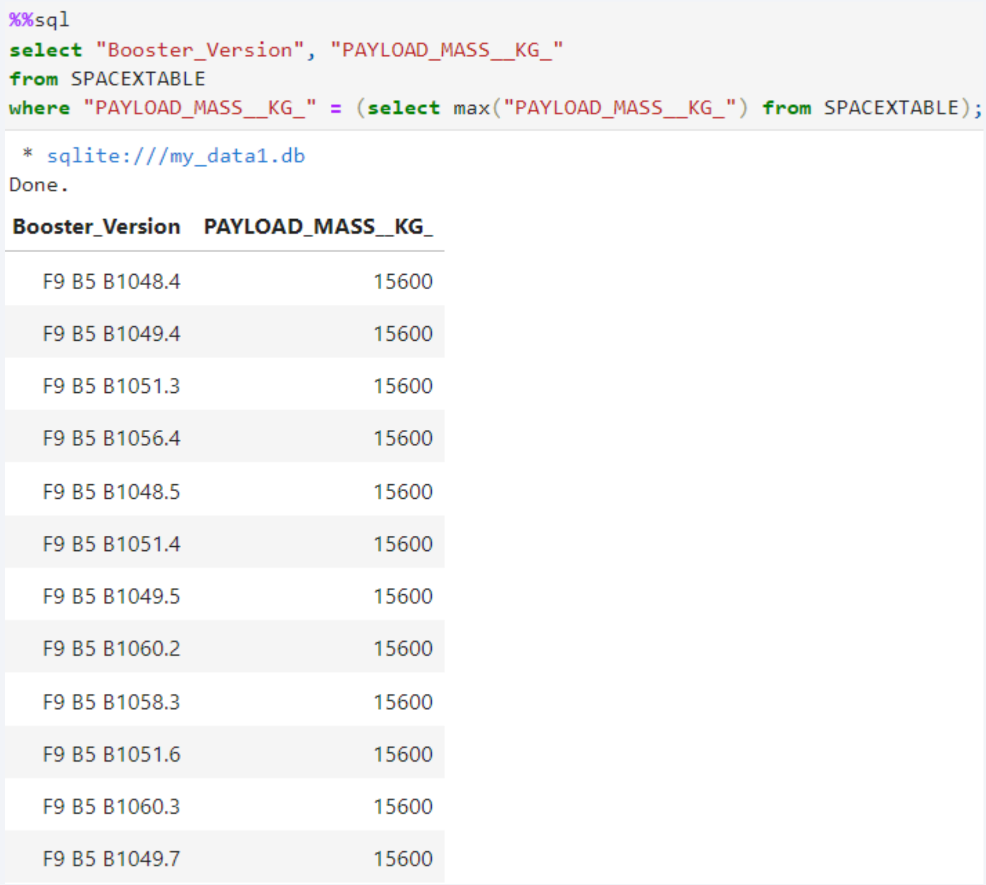
```
%%sql
with cte_succ as (
    select count(*) as "succ"
    from SPACEXTABLE
    where "Mission_Outcome" like "Success%"
),
cte_fail as (
    select count(*) as "fail"
    from SPACEXTABLE
    where "Mission_Outcome" like "Failure%"
)
select "succ" as "Successful mission outcomes",
       "fail" as "Failure mission outcomes"
from cte_succ, cte_fail;
```

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

Successful mission outcomes	Failure mission outcomes
100	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- The MAX() function returns the largest value of the selected column. A subquery is a SQL query nested inside a larger query.



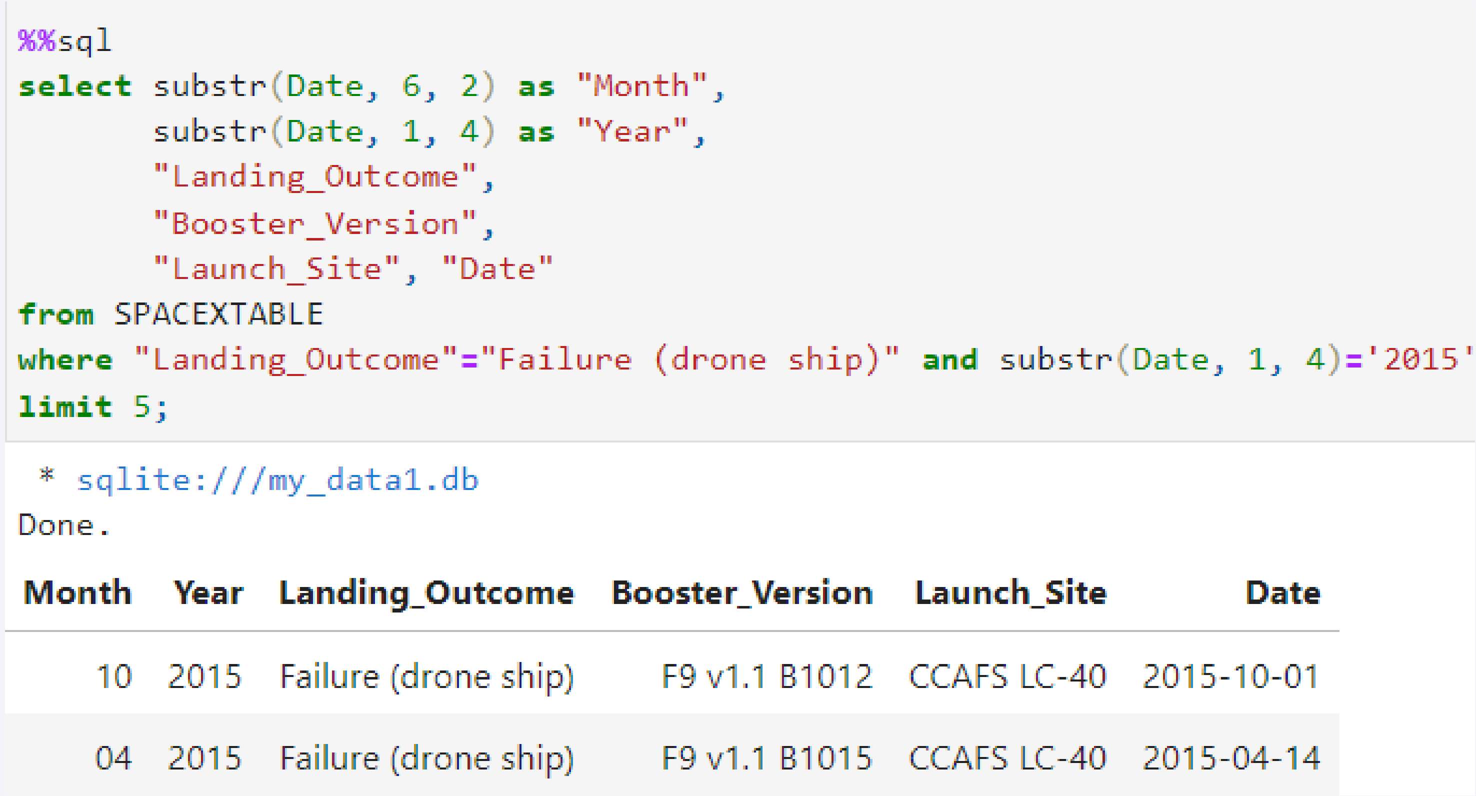
```
%%sql
select "Booster_Version", "PAYLOAD_MASS__KG_"
from SPACEXTABLE
where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE);
```

```
 * 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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- The SUBSTR() function extracts a substring from a string (starting at any position).



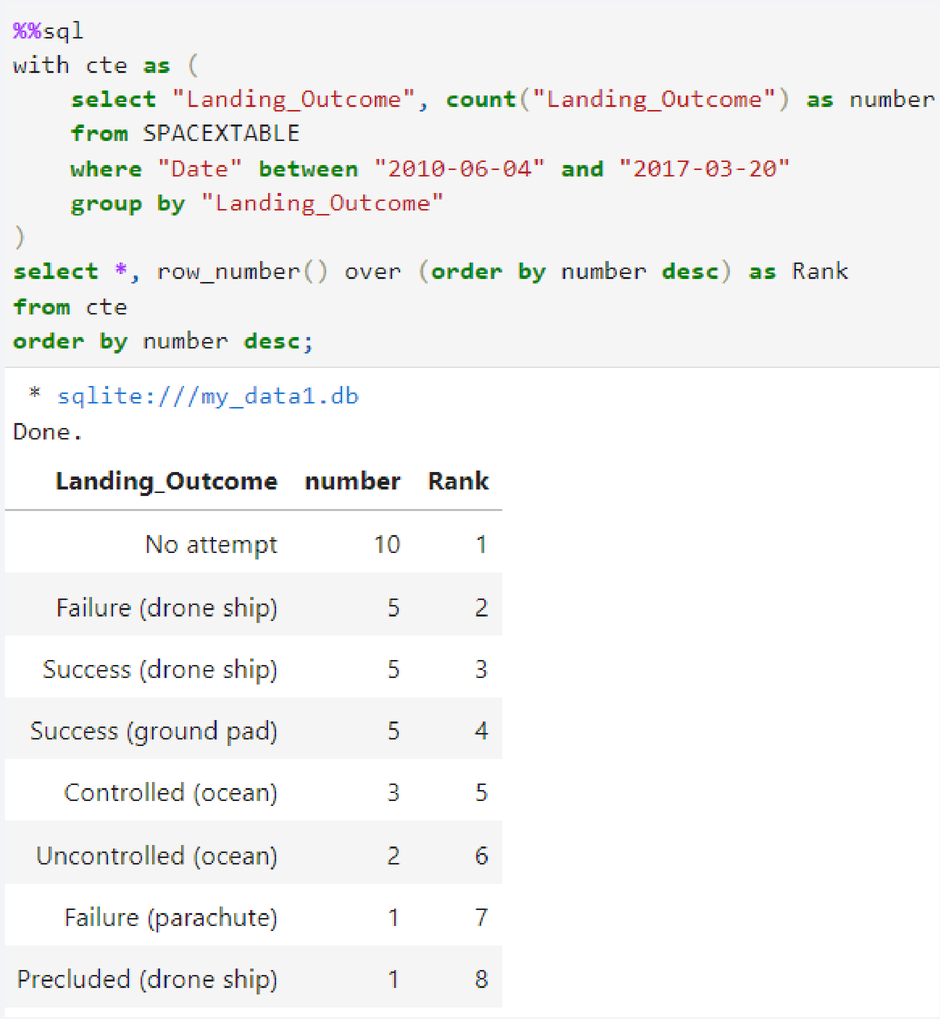
```
%%sql
select substr(Date, 6, 2) as "Month",
       substr(Date, 1, 4) as "Year",
       "Landing_Outcome",
       "Booster_Version",
       "Launch_Site", "Date"
from SPACEXTABLE
where "Landing_Outcome"="Failure (drone ship)" and substr(Date, 1, 4)='2015'
limit 5;
```

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

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

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
- The ROW_NUMBER() function numbers the output of a result set.



```
%%sql
with cte as (
    select "Landing_Outcome", count("Landing_Outcome") as number
    from SPACEXTABLE
    where "Date" between "2010-06-04" and "2017-03-20"
    group by "Landing_Outcome"
)
select *, row_number() over (order by number desc) as Rank
from cte
order by number desc;
```

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

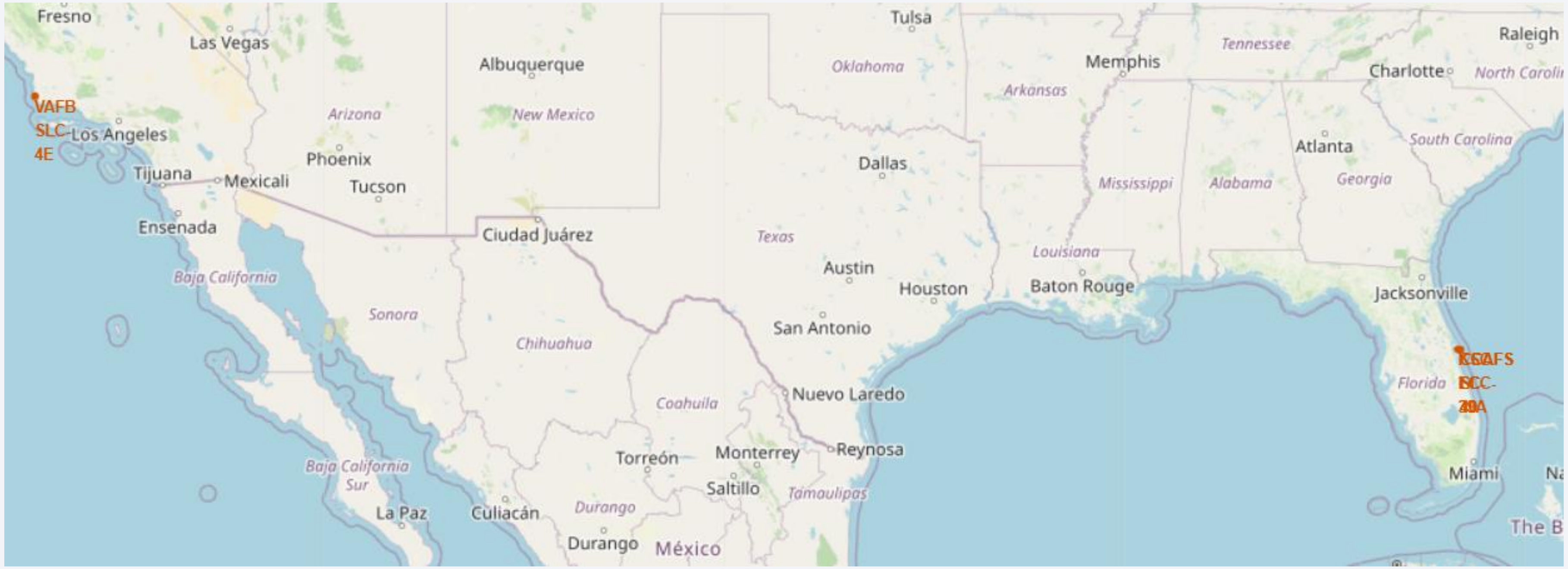
Landing_Outcome	number	Rank
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	3
Success (ground pad)	5	4
Controlled (ocean)	3	5
Uncontrolled (ocean)	2	6
Failure (parachute)	1	7
Precluded (drone ship)	1	8

Section 3

Launch Sites Proximities Analysis

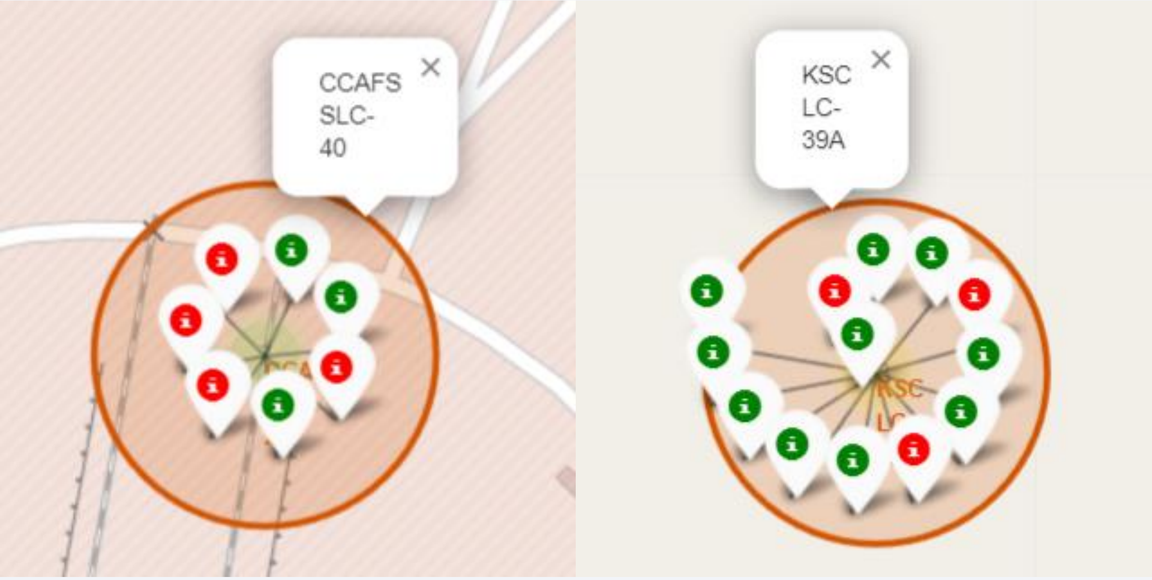
Launch Sites

- Most of the launch sites are around 28 degrees of latitude (one of them is at 34 degrees), far from the Equator Line. But all of them are in very close proximity to the coast.



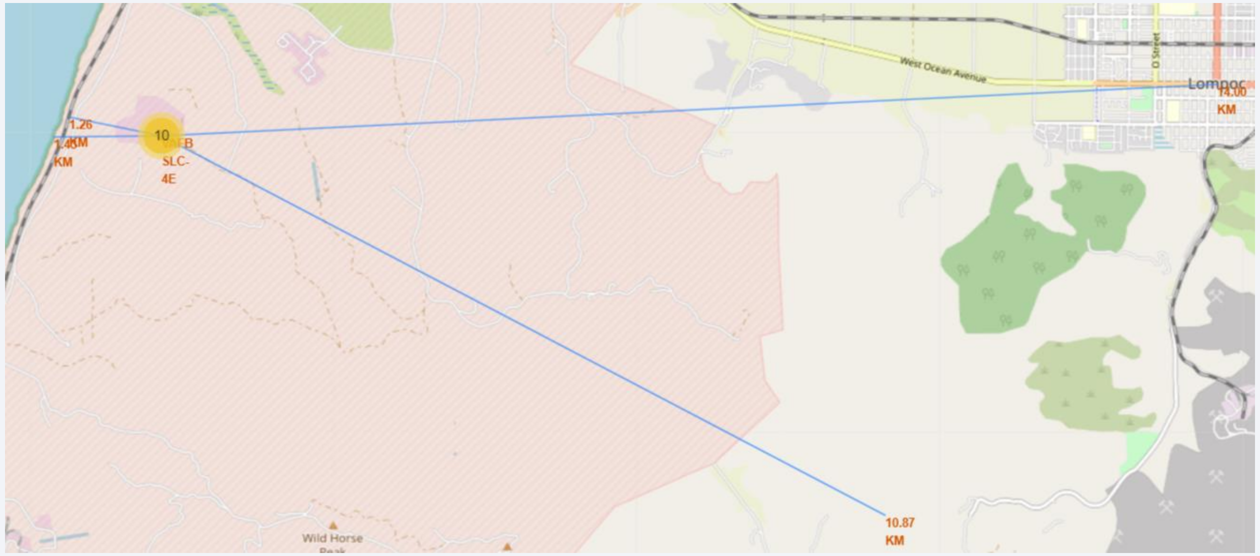
Success/Failure of the Launch Outcomes

- The marker's icon color indicates if the launch was succeeded (green) or failed (red)
- From the color-labeled markers in marker clusters, it is easy to identify which launch sites have relatively high success rates, like KSC LC-39A, and which ones have low success rate, like CCAFS SLC-40.



Launch Sites And Its Proximities

- Are launch sites in close proximity to railways?
  - Yes, less than 2 Km from the closest railway



•Are launch sites in close proximity to highways?

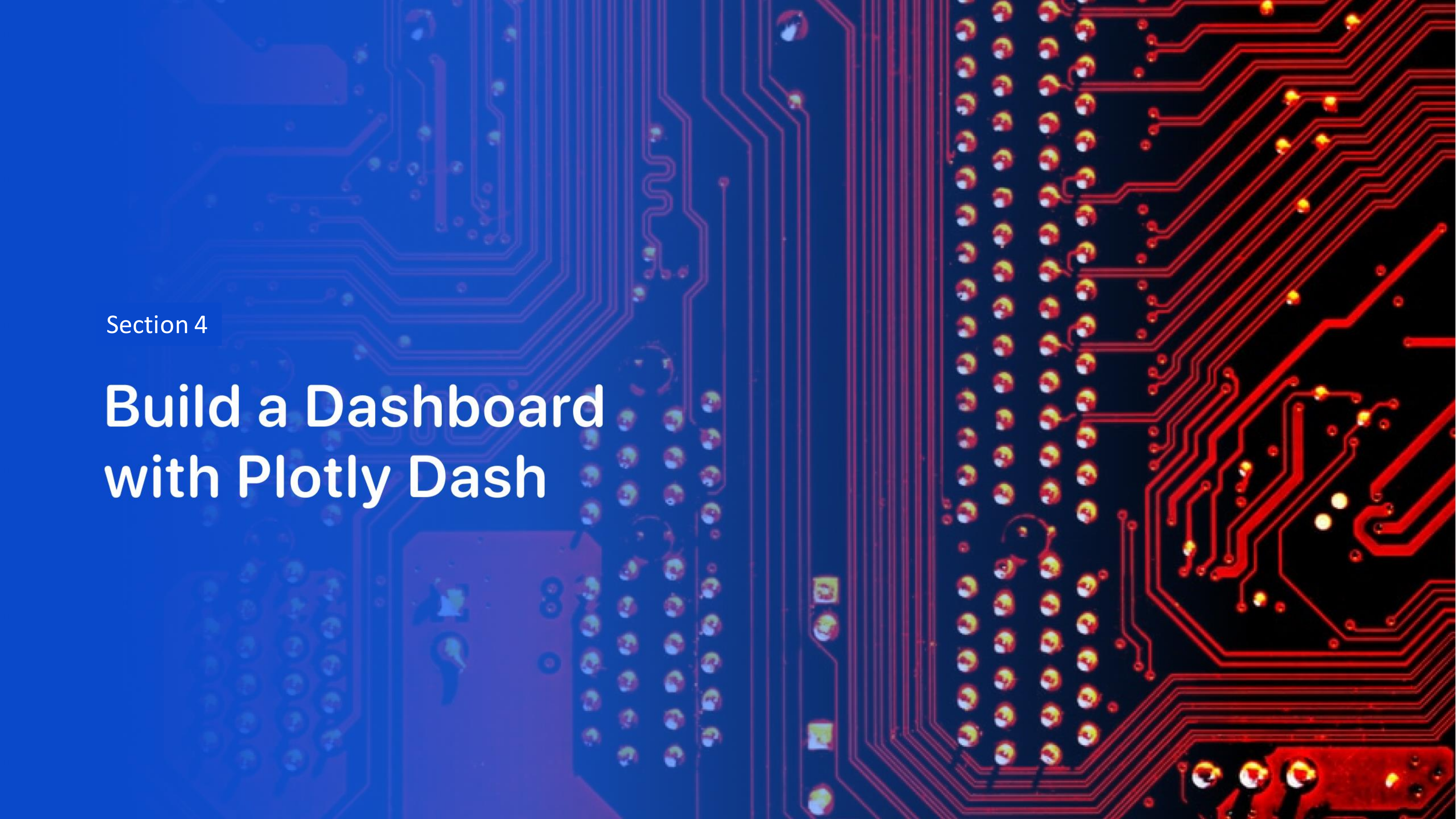
•Yes, less than 2 Km from the coastline

•Are launch sites in close proximity to coastline?

•Yes, less than 2 Km from the coastline

•Do launch sites keep certain distance away from cities?

•Yes, More than 10 Km from the closest city

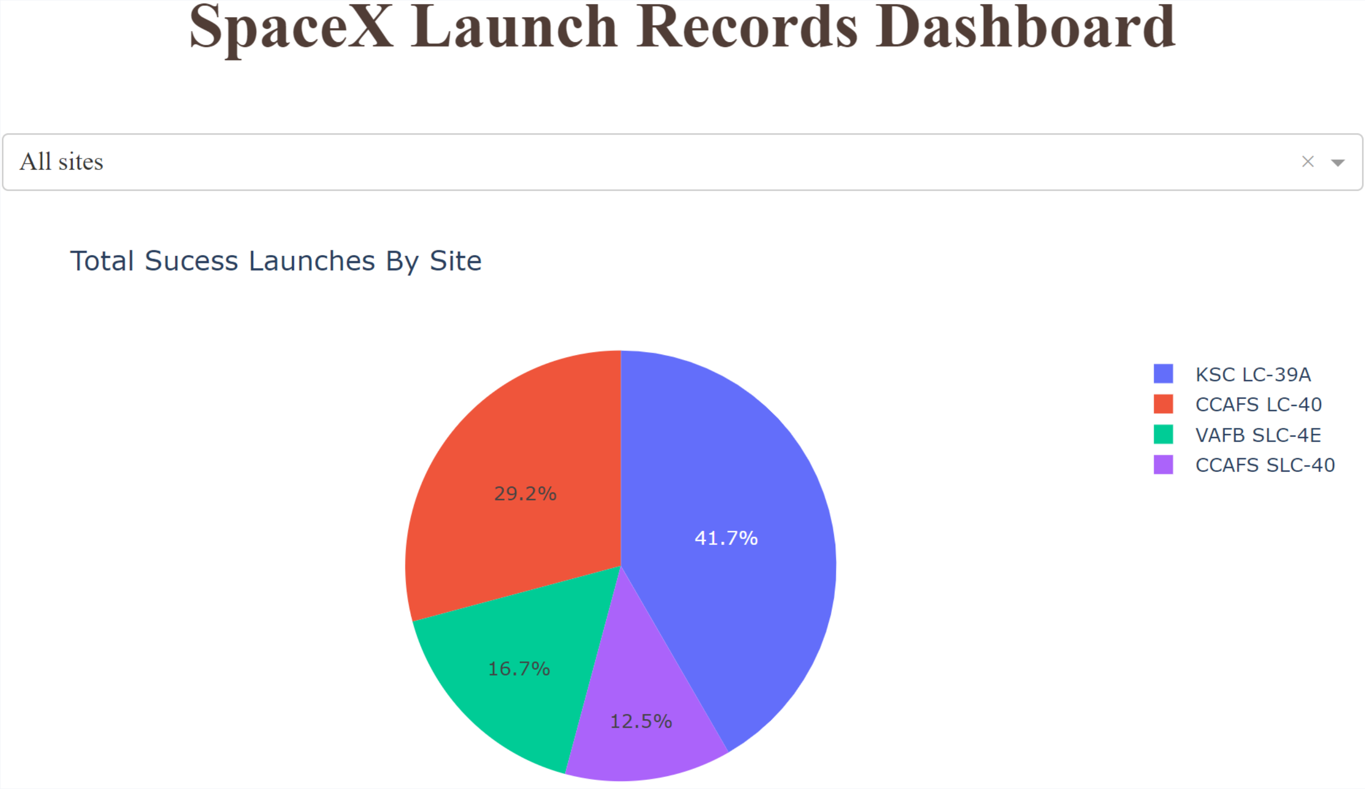


Section 4

Build a Dashboard with Plotly Dash

Success Launches By Site

- KSC LC-39A is responsible for 41.7% of the succeeded launches and CCAFS SLC-40 launched only 12.5% of the well succeeded outcomes



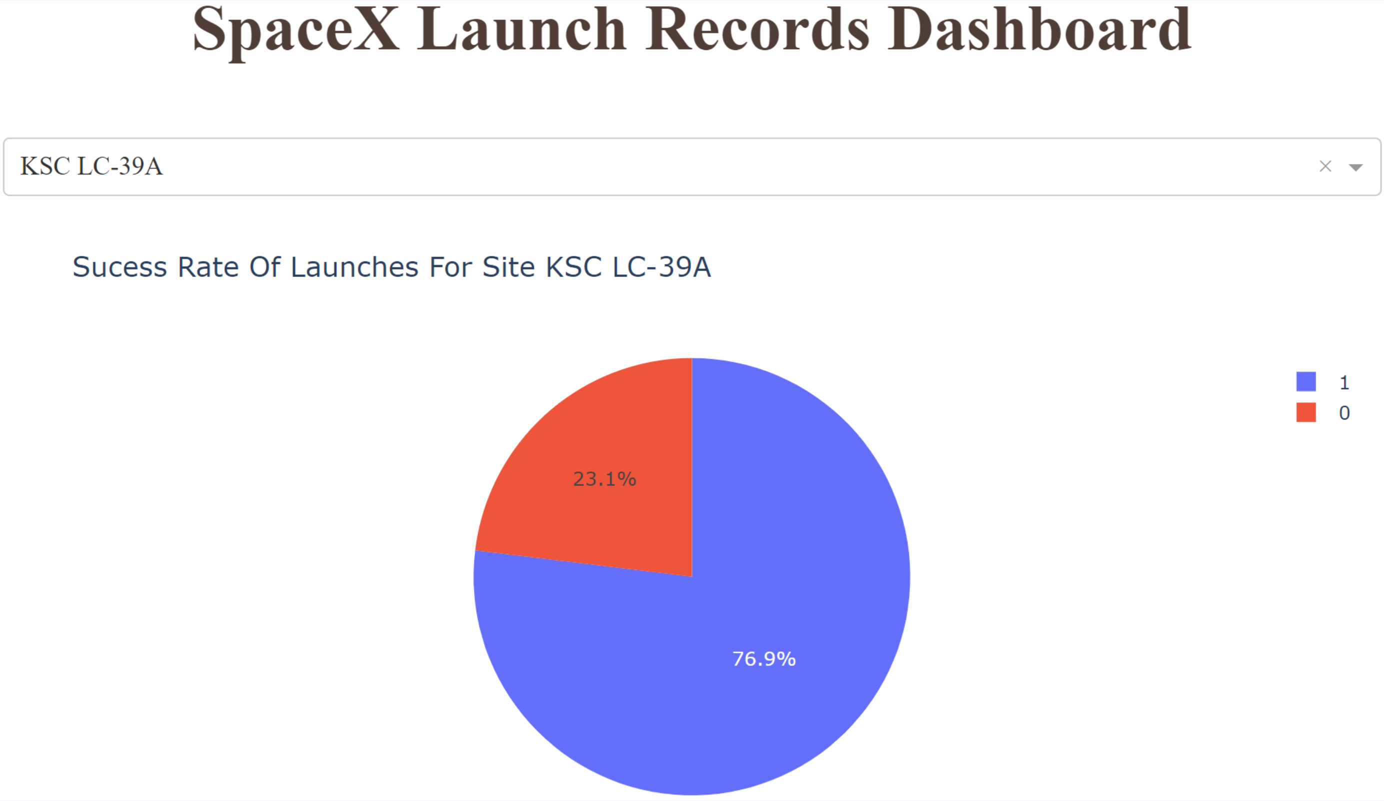
SpaceX Launch Records Dashboard

All sites

Total Sucess Launches By Site

Success Rate For Site KSC LC-39A

- KSC LC-39A was the launch site with the highest launch success ratio of 76.9%



SpaceX Launch Records Dashboard

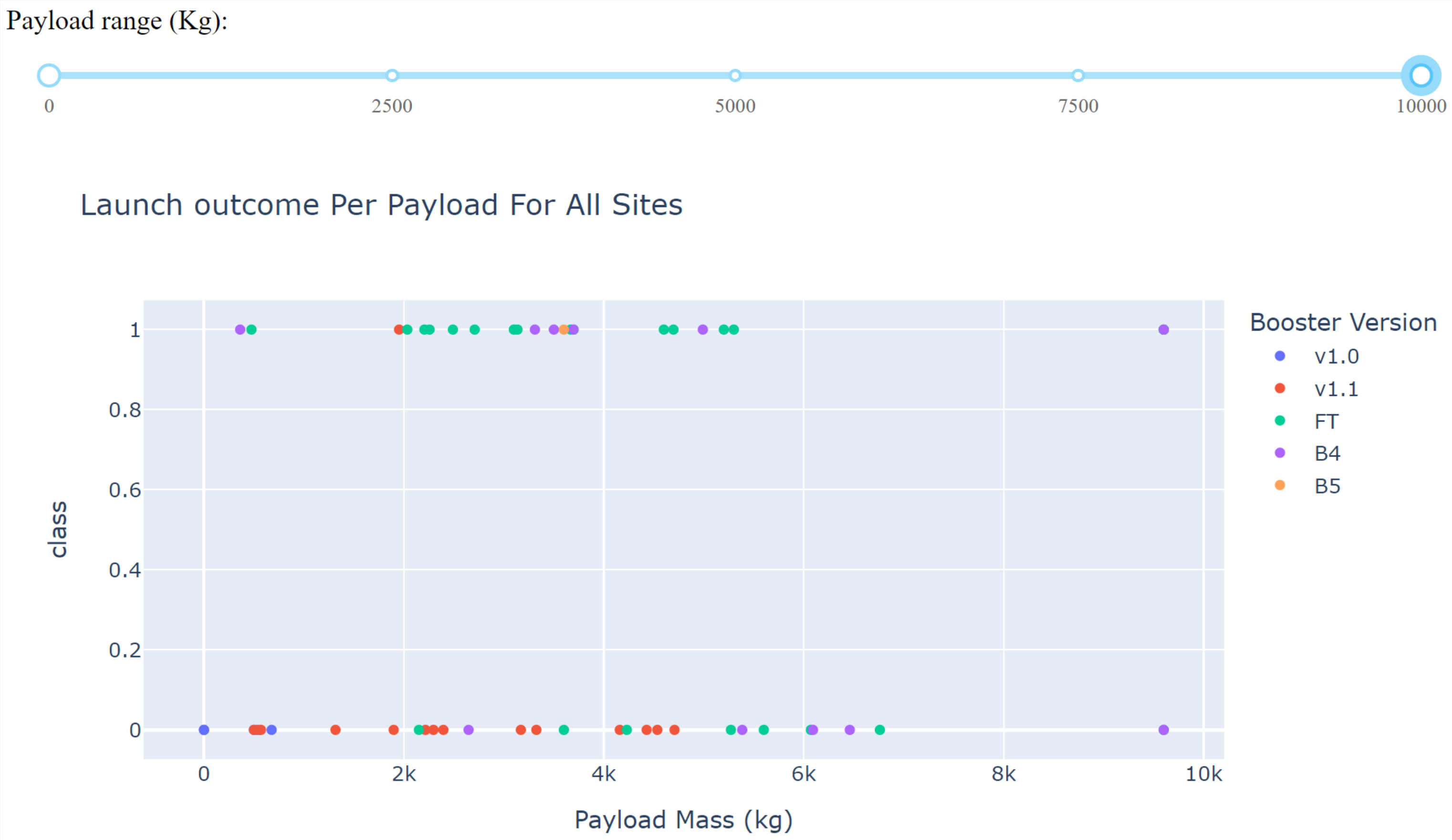
KSC LC-39A

Sucess Rate Of Launches For Site KSC LC-39A

Payload vs. Launch Outcome For All Sites

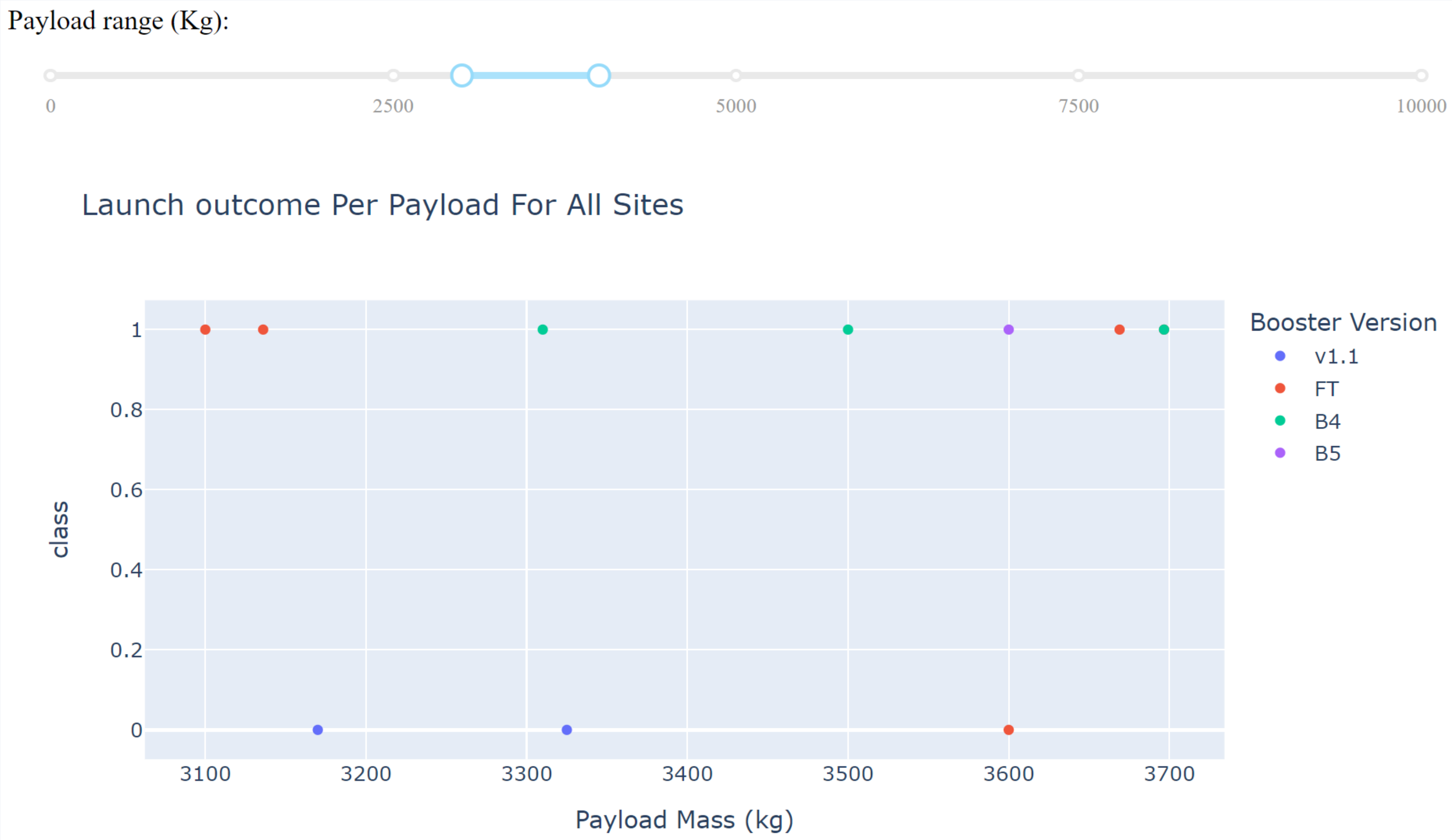
- Payload vs. Launch Outcome scatter plot for all sites, with all the payload mass registered

- F9 Booster version v1.0 was never succeeded. The FT version got a success rate of 65% and the version B5 was launched once and was succeeded



Payload vs. Launch Outcome For All Sites

- Payload vs. Launch Outcome scatter plot for all sites, showing the payload range (about 3000Kg to 4000Kg) with the largest success rate (70%)

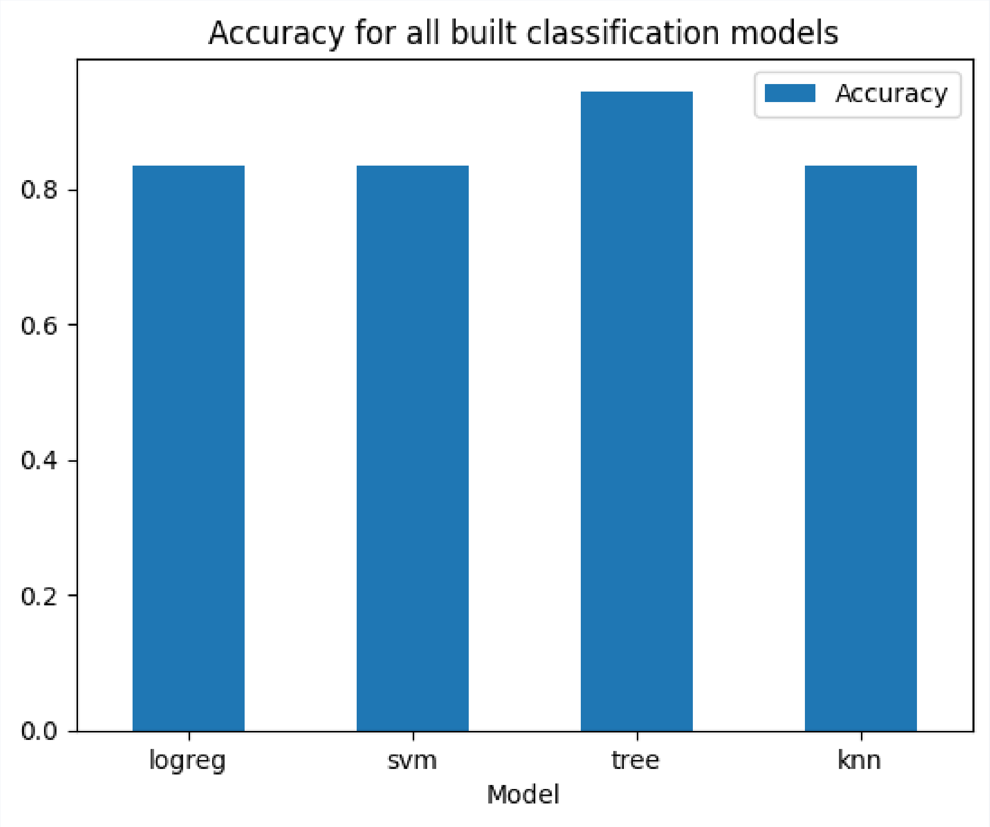


Section 5

Predictive Analysis (Classification)

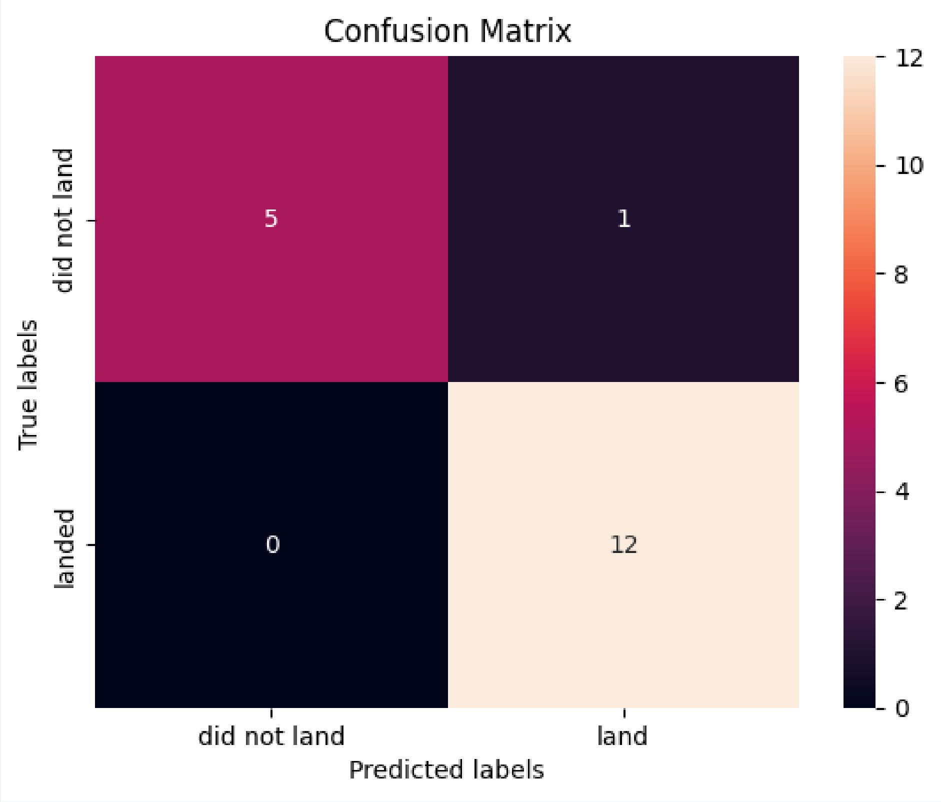
Classification Accuracy

- KNN, SVM and Logistic Regression got the same score in the test set: 83.333%
- The Decision Tree Classifier got the best accuracy, with: 94.444%
- Tuned hyperparameters (best parameters): {'criterion': 'entropy', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'best'}



Confusion Matrix

- The Decision Tree Model predicted almost all the landing outcomes of the test set. Only one outcome labeled as 0 (did not land) was wrongly predicted as 1 (landed), as can be seen in the confusion matrix alongside.

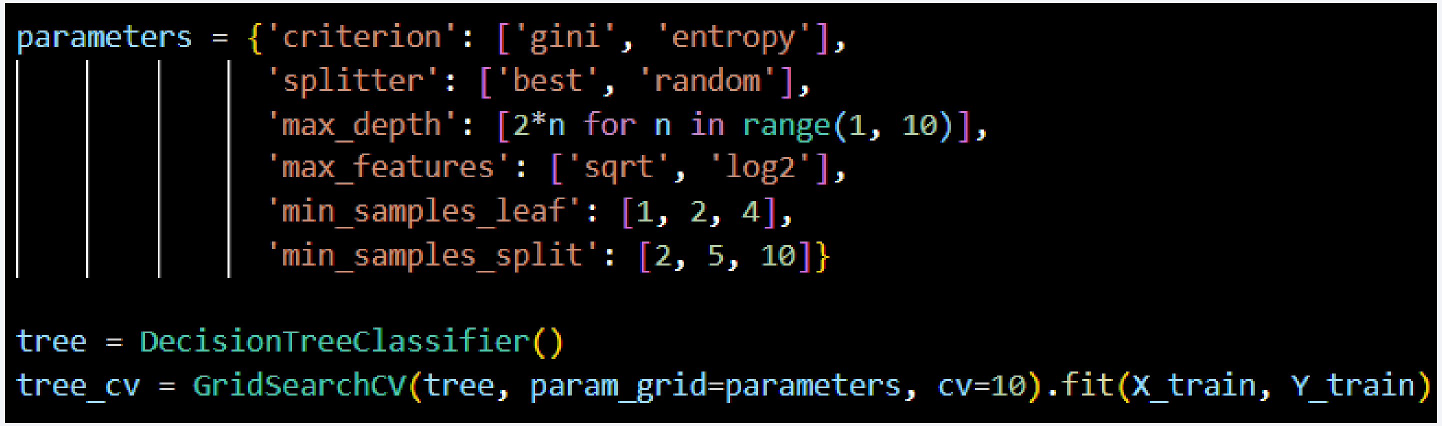


Conclusions

- ES-L1, GEO, HEO and SSO are the orbits with the highest success rate;
- From the EDA with SQL we know that the first successful landing outcome in ground pad was achieved in 2015/12/22;
- From interactive visual analytics: KSC LC-39A was the launch site with the highest launch success ratio;
- The best classifier to predict the landing outcome of the Falcon 9 Rocket was a Decision Tree Model

Appendix

- Construction of the best classification model:



```
parameters = {'criterion': ['gini', 'entropy'],
              'splitter': ['best', 'random'],
              'max_depth': [2*n for n in range(1, 10)],
              'max_features': ['sqrt', 'log2'],
              'min_samples_leaf': [1, 2, 4],
              'min_samples_split': [2, 5, 10]}

tree = DecisionTreeClassifier()
tree_cv = GridSearchCV(tree, param_grid=parameters, cv=10).fit(X_train, Y_train)
```

- Github url with all the relevant files of this project:
- https://github.com/edvdjr/spacex-launch-analysis

Thank you!