SpaceX Falcon 9 first stage Landing Prediction using Data Science.

Ramadhani Maulid

15th February 2024

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings and Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Data science has been employed to determine if the SpaceX Falcon 9 first stage will land successfully. First stage does most of work, larger than second stage and highly determines the cost of lunch. Cost of the first stage lunch influences decision of opting alternative bids from other space agencies.
- Summary of Methodologies
  - Data Collection through SpaceX REST API and Web Scraping.
  - Data Transformation through Data Wrangling.
  - Exploratory data analysis (EDA) using SQL , and Visualization by Pandas and Matplotlib.
  - Interactive Visual Analytics using Folium (Map) and Plotly Dash (Dashboard)
  - Predictive analysis using machine learning techniques (support vector machine (SVM), logistic regression, decision trees and K-nearest neighbor (KNN))
- Summary of all Results
  - EDA Results.
  - Interactive Visual Analytics Results.
  - Predictive Analysis Results.

INTRODUCTION

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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. Instead of using rocket science to determine if the first stage will land successfully, machine learning model was trained and used public information to predict if SpaceX will reuse the first stage.
- Questions to be answered
  - What are factors that determine successfully landing of a first stage.
  - What is rate of change of success rate of landing with time.
  - What are required operating conditions for successfully first stage landing.
  - What method correctly predict success of landing of first stage.

METHODOLOGY

- Reference is made summaries of the methodologies in the executive summary.
- Data Collection through SpaceX REST API
  - Get request performed to obtain launch data from API.
  - Launch data from get request decoded using .json() method and then converted to dataframe using .json_normalize function.
  - Defined a series of helper functions that helped using the API to extract information using identification numbers in the launch data.
  - Create dictionary from data and create a dataframe from it.
  - Retain only data from launch 9 falcon on the dataframe.
  - Replace missing values of Payload Mass with calculated .mean()
  - Export data in the comma separated value (CSV) format.

METHODOLOGY

- **Data Collection through Web Scraping**
  - Request the Falcon9 Launch Wiki page from its URL and hence Create a BeautifulSoup object from the HTML response.
  - Extract all column/variable names from the HTML table header.
  - Defined a series of helper functions that helped using the API to extract information using identification numbers in the launch data.
  - Create a data frame by parsing the launch HTML tables which is attained by first creating dictionary from data then convert to data frame.
  - Export data in the comma separated value (CSV) format.

METHODOLOGY

- **Data Transformation through Data Wrangling.**
  - Exploratory data analysis has been through:
  - Identifying and calculating the percentage of the missing values in each attribute.
  - Identifying which columns are numerical and categorical.
  - Calculating the number of launches on each site.
  - Calculating the number and occurrence of each orbit.
  - Calculating the number and occurrence of mission outcome of the orbits.
  - Creating a landing outcome label from Outcome column.
  - Data obtained exported in CSV format.

METHODOLOGY

- **EDA Through Visualization**
  - Performed exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib.
  - Created plots (Scatter, Bar) used to Visualize the relationship between:

(Flight Number vs Payload, Flight Number vs Launch Site, Payload Mass (kg) vs Launch Site, Payload Mass (kg) vs Orbit type, Success Rate vs Year).

  - Preliminary insights obtained about how each important variable would affect the success rate, used to select the features that will be used in success prediction.

METHODOLOGY

- **EDA Through SQL**
  - Data from CSV loaded to database.
  - Performed SQL queries to get insights from data such as:
  - 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.

METHODOLOGY

- **EDA Through SQL (Continued)**
  - 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.
  - 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.
  - 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

METHODOLOGY

- **Interactive Visual Analytics using Folium**
  - Marked all launch sites on a map and observed their proximity to equator line and the coast (created and inserted map objects such as markers with popup labels, circles, lines to mark the success or failure of launches for each site on the folium map).
  - Marked the success/failed launches for each site on the map by enhancing map by adding the launch outcomes for each site, and see which sites have high success rates by using the color-labeled marker clusters. Data frame used has detailed launch records, and the class column indicates if this launch was successful or not (0 which means failure mapped to red marker,1 which means success marked to green marker for the map).

METHODOLOGY

- **Interactive Visual Analytics using Plotly and Dash**
  - Built a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real time.
  - The dashboard application contains input components dropdown list (You can select all sites or specific site) and a range slider to interact with a pie chart and a scatter point chart.
  - After visual analysis using the dashboard (pie and scatter charts), some insights obtained to answer the questions such as Which site has the largest successful launches?, Which site has the highest launch success rate?

METHODOLOGY

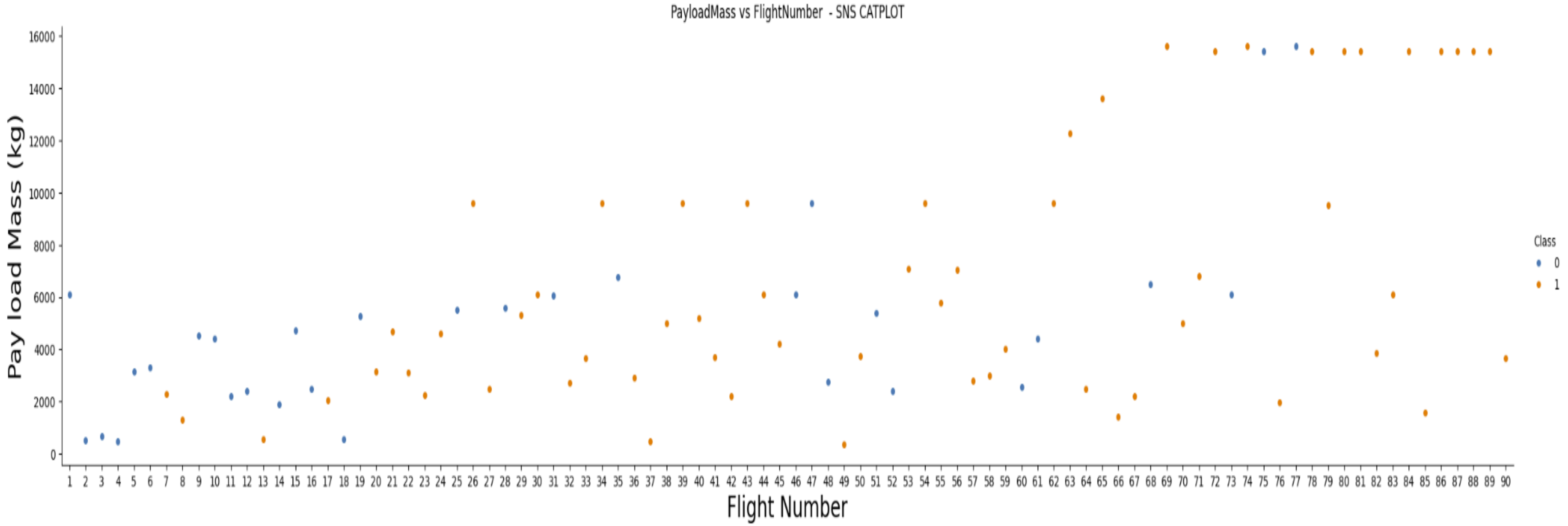
- **Predictive Analysis was achieved through:**
  - Performing exploratory Data Analysis and determine Training Labels
  - *Creating a column for the class*
  - *Standardizing the data*
  - *Split into training data and test data*
  - Finding best Hyperparameter for Support Vector Machines (SVM), Classification Trees, Logistic Regression and K-Nearest Neighbor(KNN)
  - Find the method performs best using test data.

RESULTS

- **EDA Through Visualization Results**
  - To achieve EDA number of plots have been created to visualize relationships between features.
  - The results for relationship are covered in the next slides(TITLE shows the features/columns involved in visualization of their relationship)
  - Scatter(catplot) and Bar plots were used due to nature of the nature .

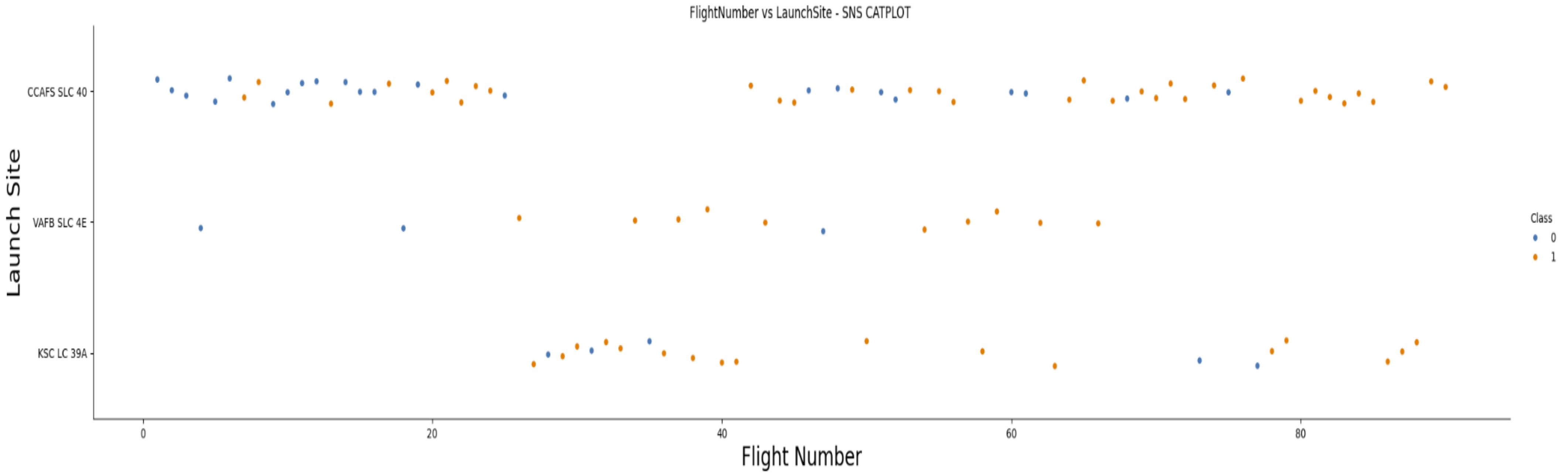
RESULTS

- **EDA Through Visualization Results**



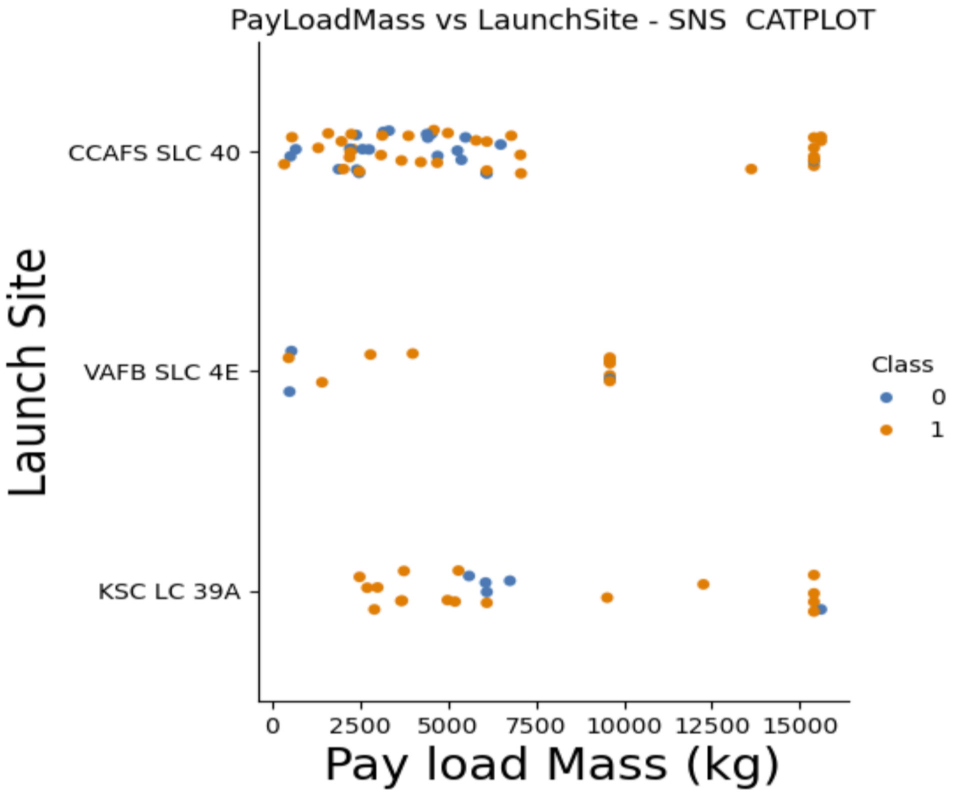
RESULTS

- **EDA Through Visualization Results**



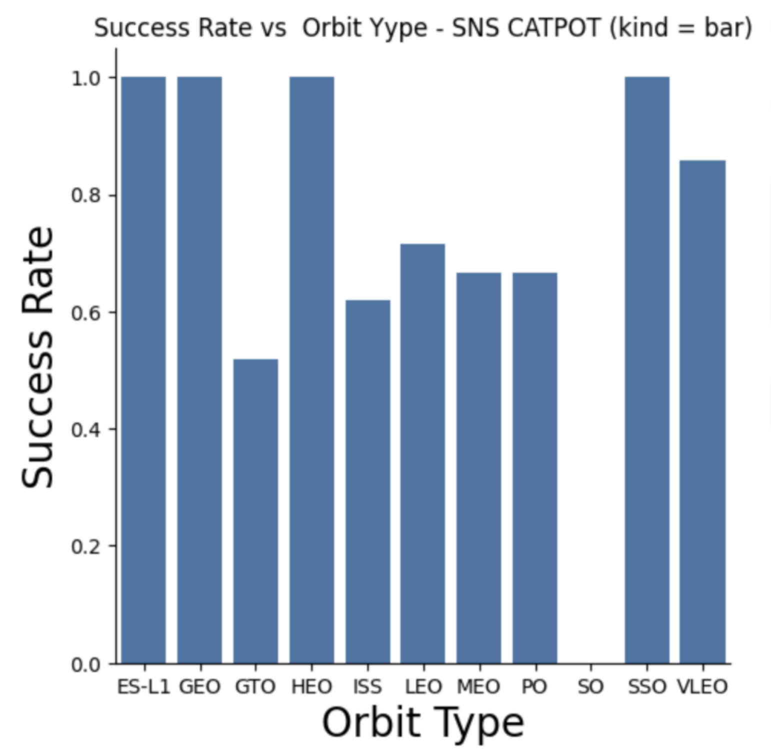
RESULTS

- **EDA Through Visualization Results**



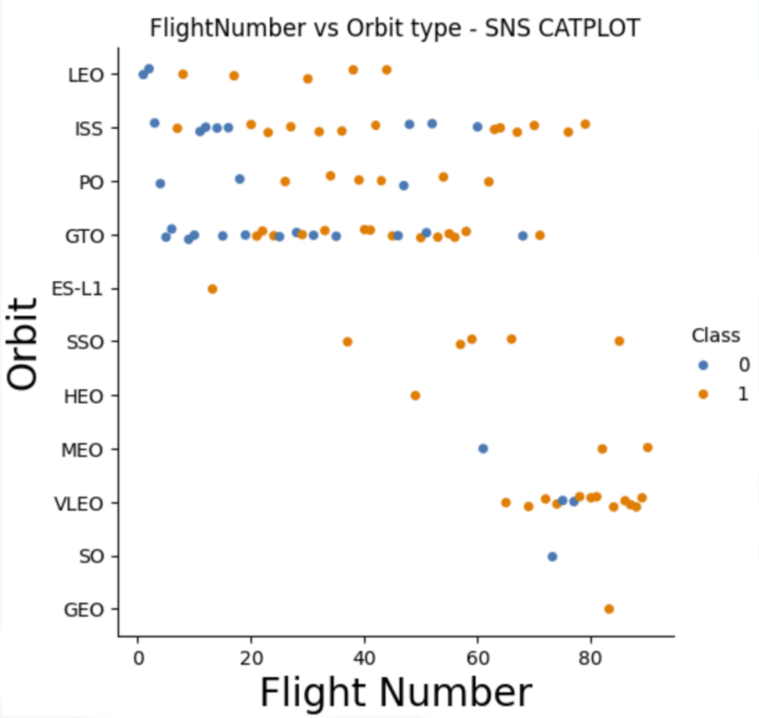
RESULTS

- **EDA Through Visualization Results**



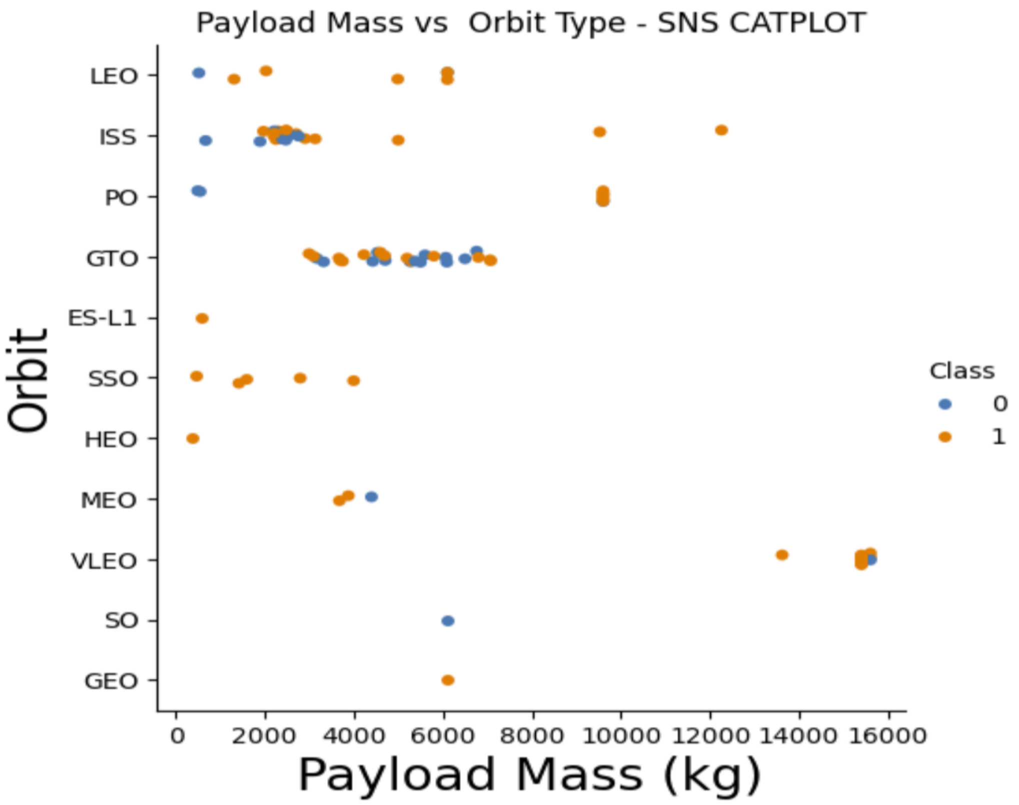
RESULTS

- **EDA Through Visualization Results**



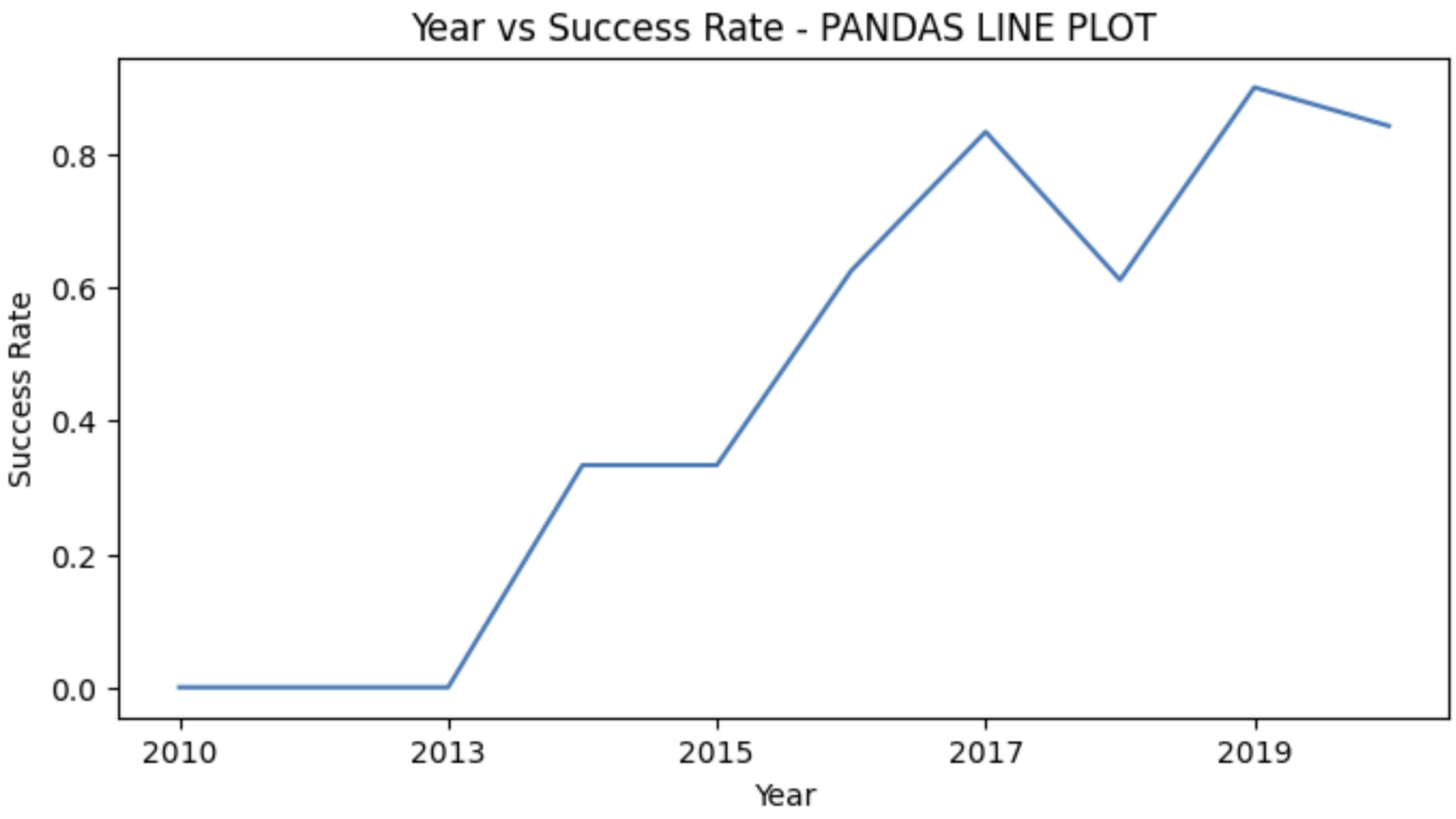
RESULTS

- EDA Through Visualization Results



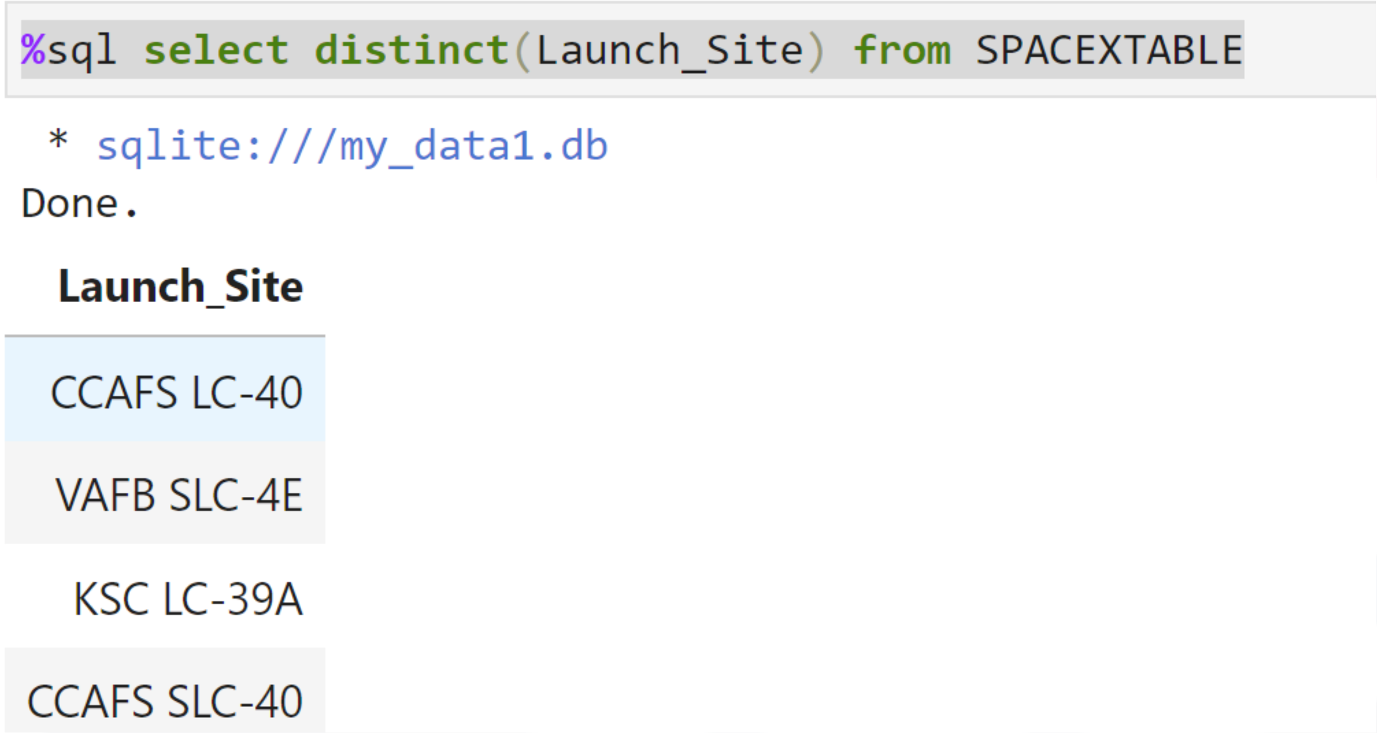
RESULTS

- EDA through Visualization Results



RESULTS

- EDA Through SQL Results (names of the unique launch sites)



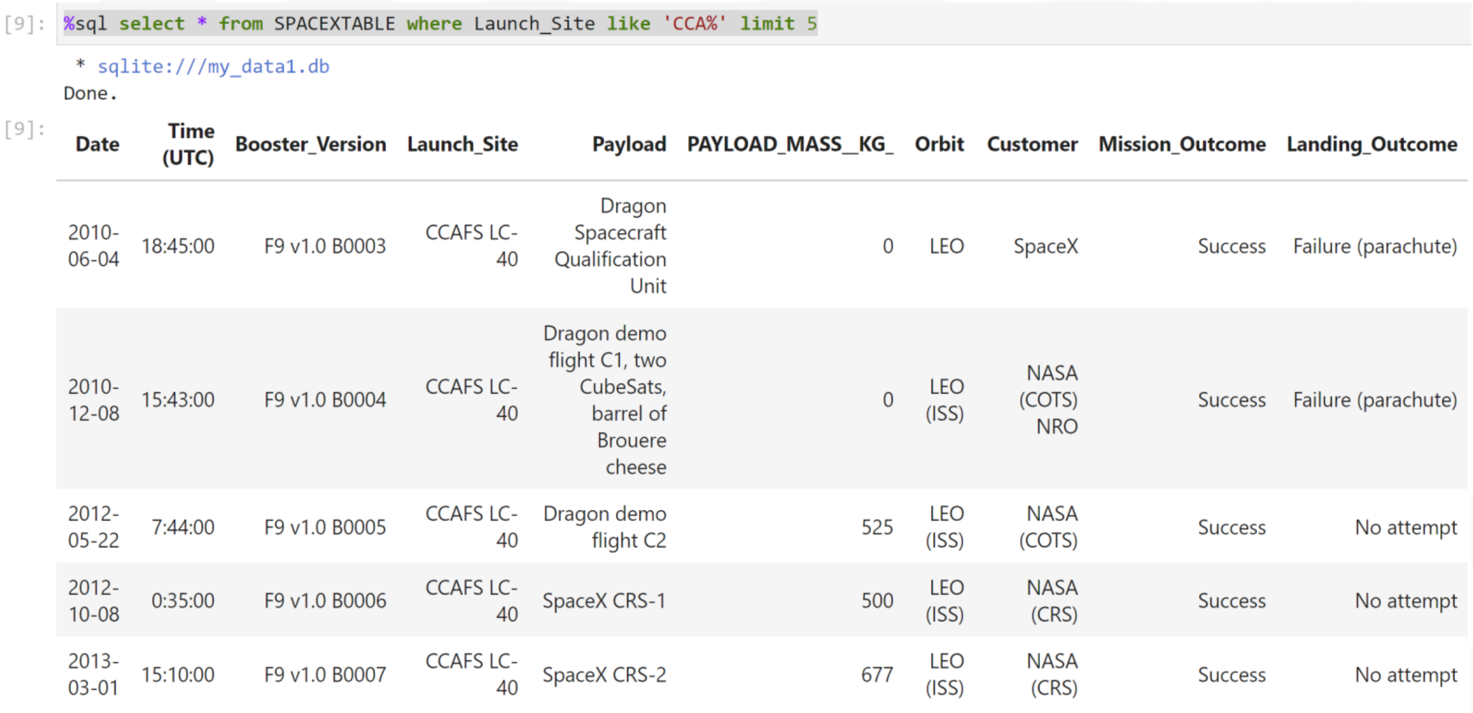
```
%sql select distinct(Launch_Site) from SPACEXTABLE
```

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

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

RESULTS

- **EDA Through SQL Results (5 records that launch sites begin with 'CCA')**



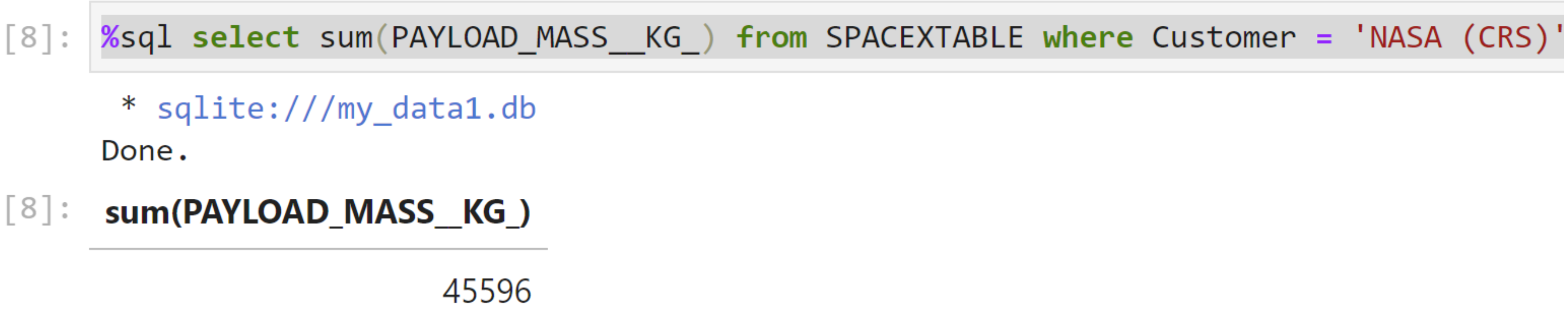
```
[9]: %sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5

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

[9]:

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

RESULTS

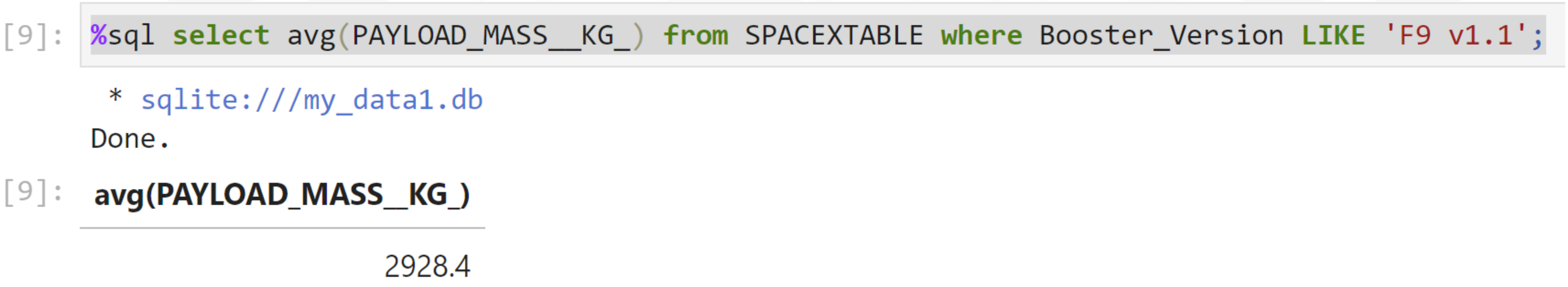
- EDA Through SQL Results (total payload mass by NASA(CRS) boosters)



```
[8]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'

      * sqlite:///my_data1.db
     Done.

[8]: sum(PAYLOAD_MASS__KG_)
     ----------------------
                      45596
```

RESULTS

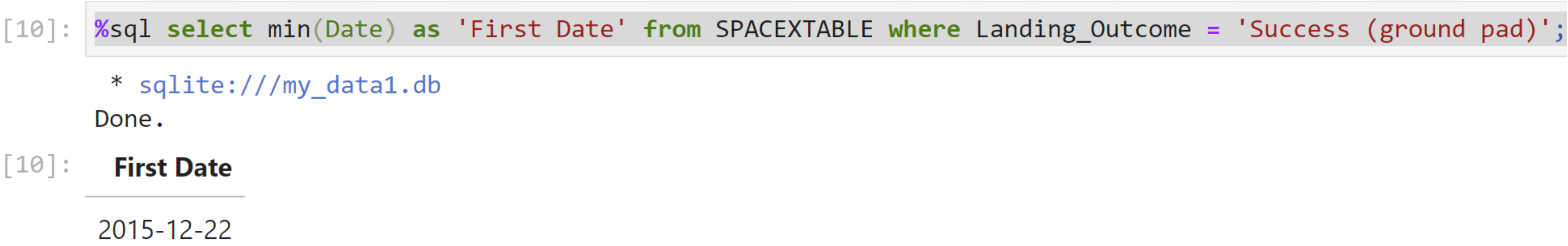
- EDA Through SQL Results (average payload mass carried by booster version F9 v1.1)



```
[9]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version LIKE 'F9 v1.1';

 * sqlite:///my_data1.db
Done.

[9]: avg(PAYLOAD_MASS__KG_)
                    2928.4
```

RESULTS

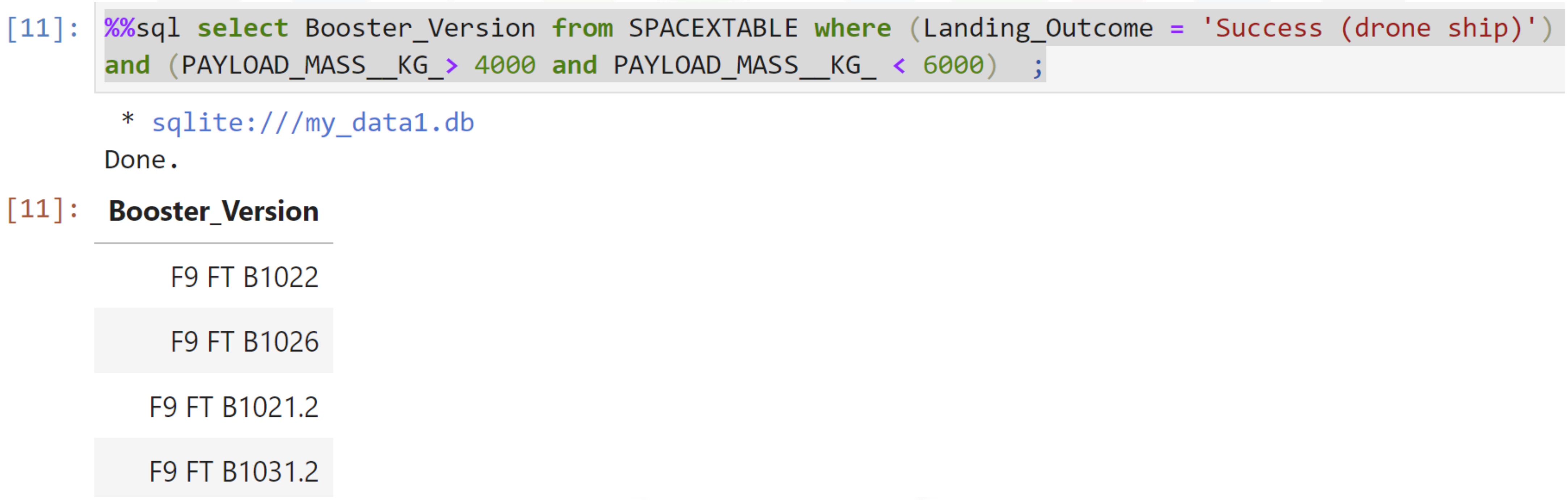
- **EDA Through SQL Results (date when the first successfully landing outcome in ground pad was achieved)**



```
[10]: %sql select min(Date) as 'First Date' from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';

 * sqlite:///my_data1.db
Done.

[10]:  First Date
       2015-12-22
```

RESULTS

- EDA Through SQL Results (names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000)



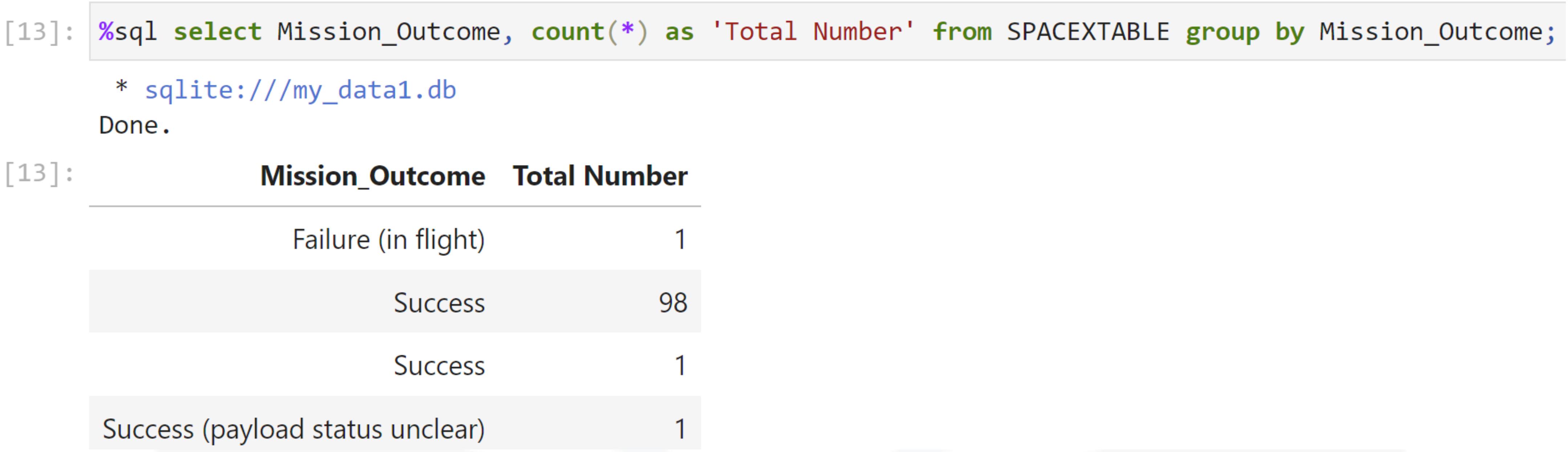
```
[11]: %%sql select Booster_Version from SPACEXTABLE where (Landing_Outcome = 'Success (drone ship)')
      and (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000)  ;

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

[11]:

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

RESULTS

- **EDA Through SQL Results (total number of successful and failure mission outcomes)**



```
[13]: %sql select Mission_Outcome, count(*) as 'Total Number' from SPACEXTABLE group by Mission_Outcome;
```

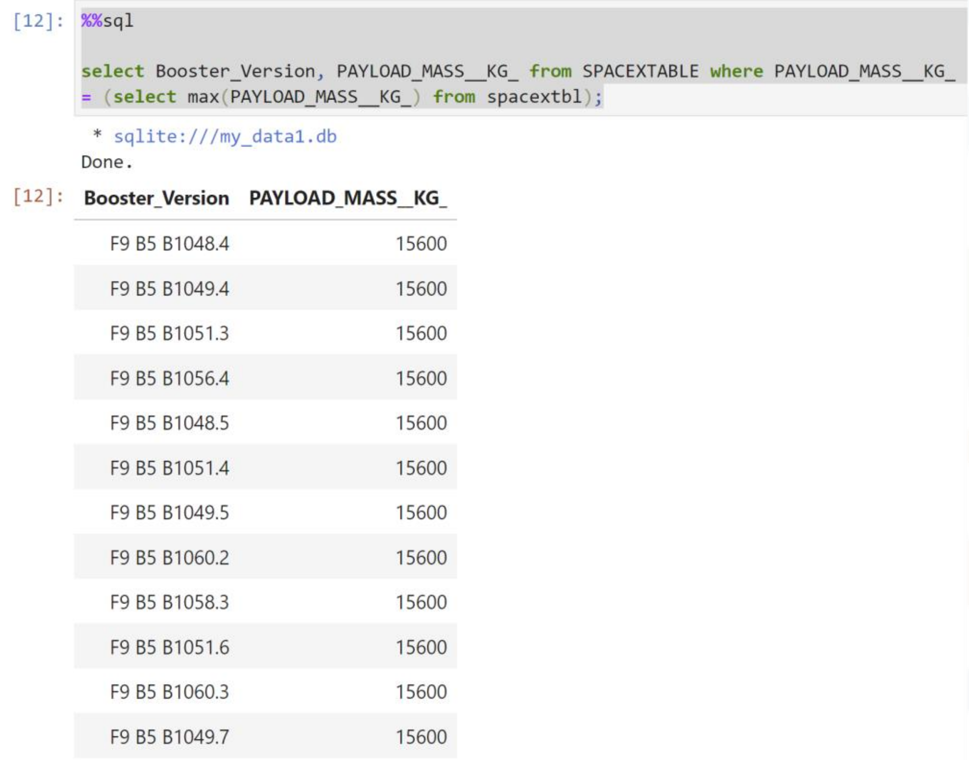
```
 * sqlite:///my_data1.db
Done.
```

[13]:

Mission_Outcome	Total Number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

RESULTS

- EDA Through SQL Results (names of the booster_versions which have carried the maximum payload mass use a subquery)



[12]:
```
%%sql

select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_
= (select max(PAYLOAD_MASS__KG_) from spacextbl);
```

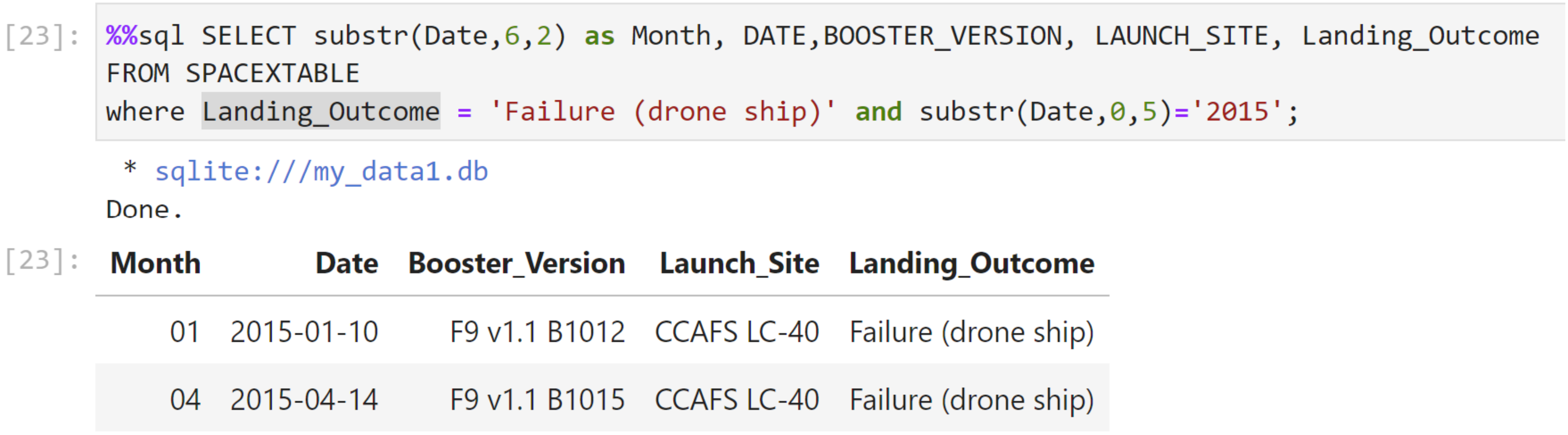
```
 * sqlite:///my_data1.db
Done.
```

[12]:

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

RESULTS

- EDA Through SQL Results (he month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015).



```
[23]: %%sql SELECT substr(Date,6,2) as Month, DATE,BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome
      FROM SPACEXTABLE
      where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

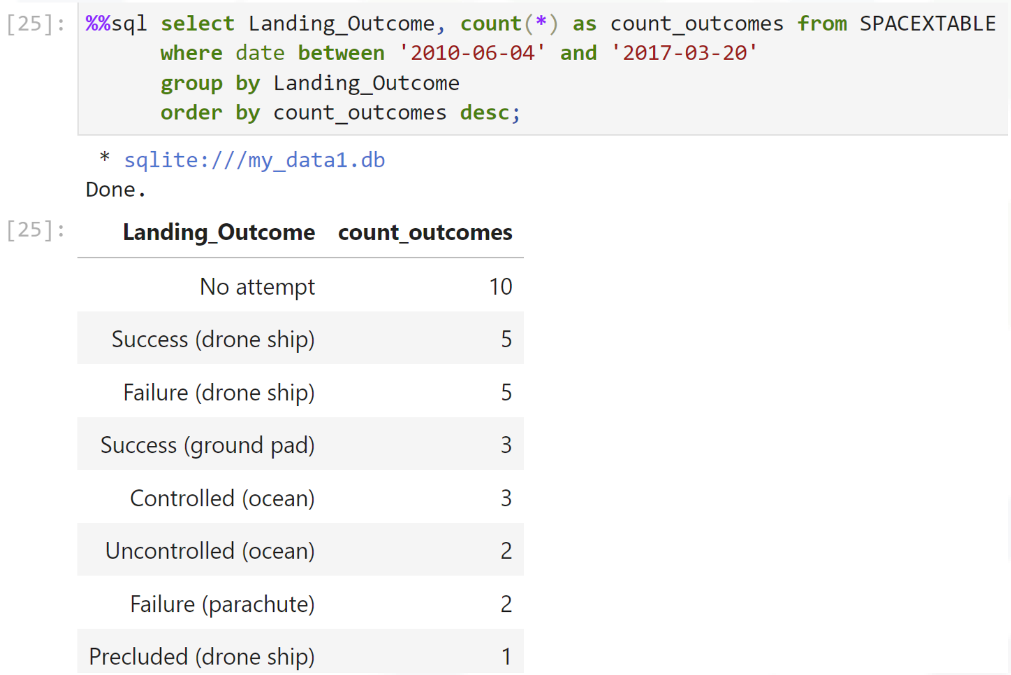
```
 * sqlite:///my_data1.db
Done.
```

[23]:

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

RESULTS

- **EDA Through SQL Results (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)**



```
[25]: %%sql select Landing_Outcome, count(*) as count_outcomes from SPACEXTABLE
          where date between '2010-06-04' and '2017-03-20'
          group by Landing_Outcome
          order by count_outcomes desc;
```

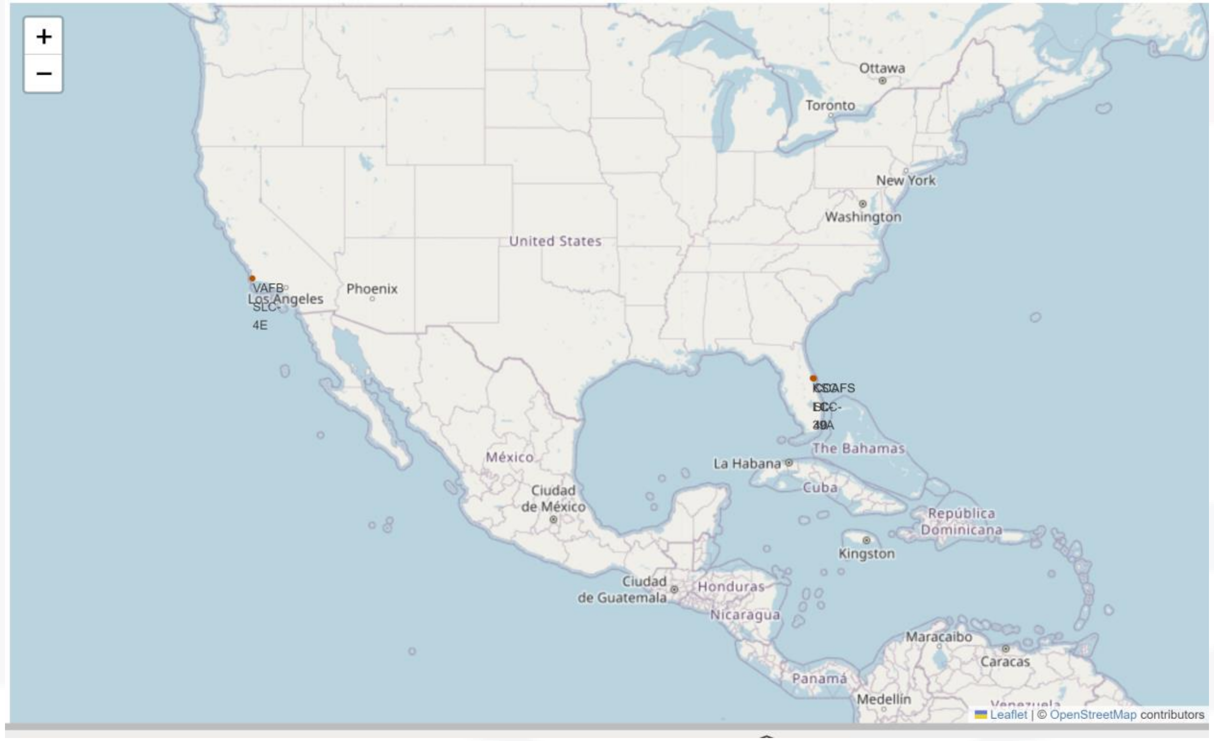
```
 * sqlite:///my_data1.db
Done.
```

[25]:

Landing_Outcome	count_outcomes
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

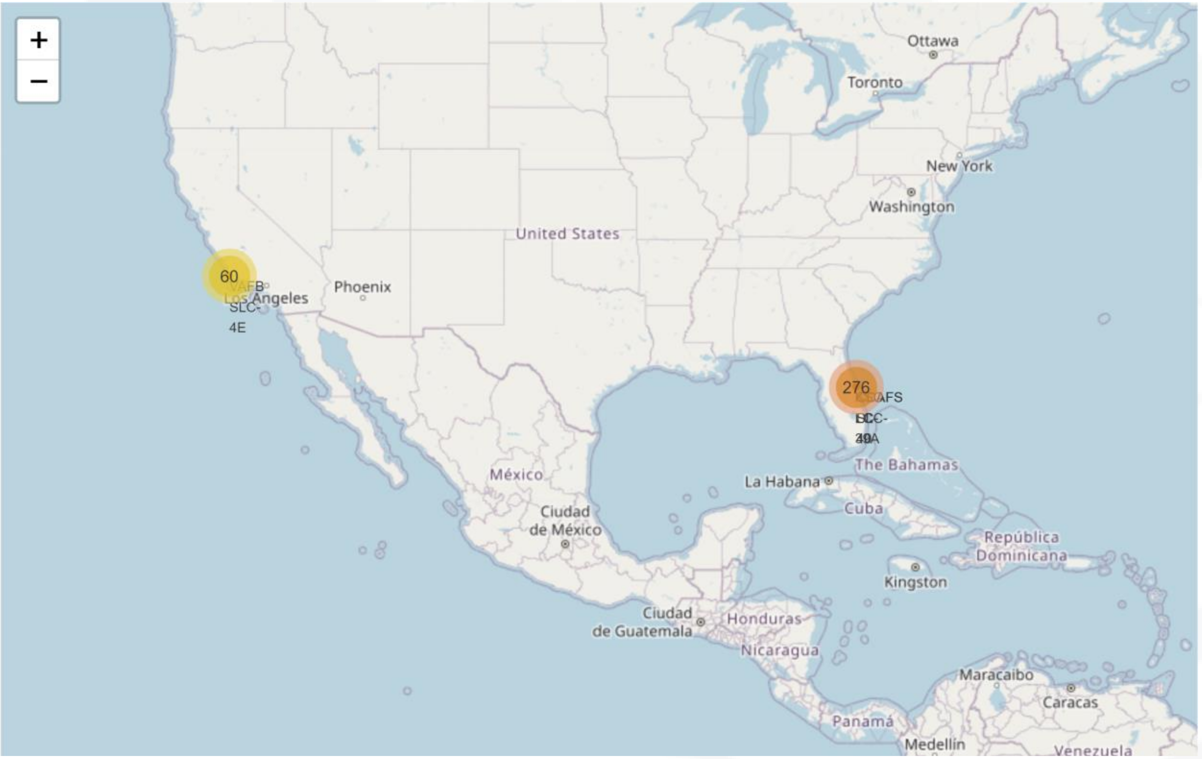
RESULTS

- **Interactive Visual Analytics with Folium**
- Mark all launch sites on a map (All sites are close proximity to equator and coast)



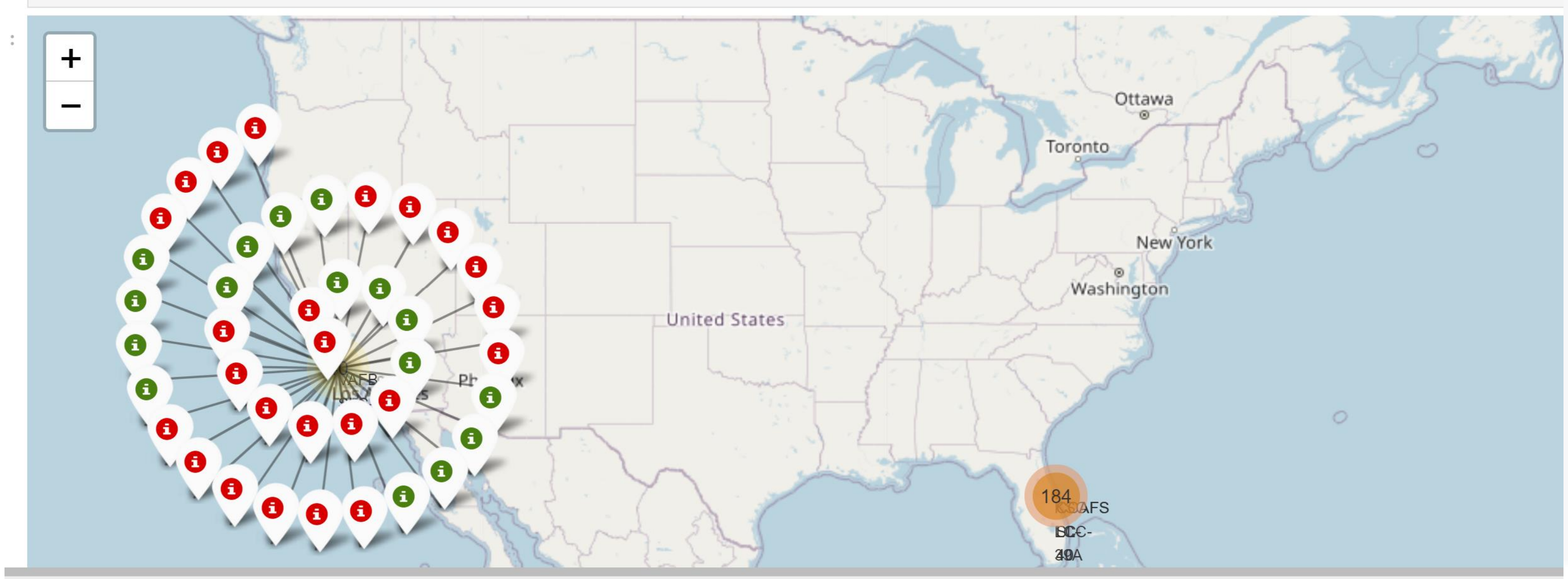
RESULTS

- **Interactive Visual Analytics with Folium**
- Mark the success/failed launches for each site on the map



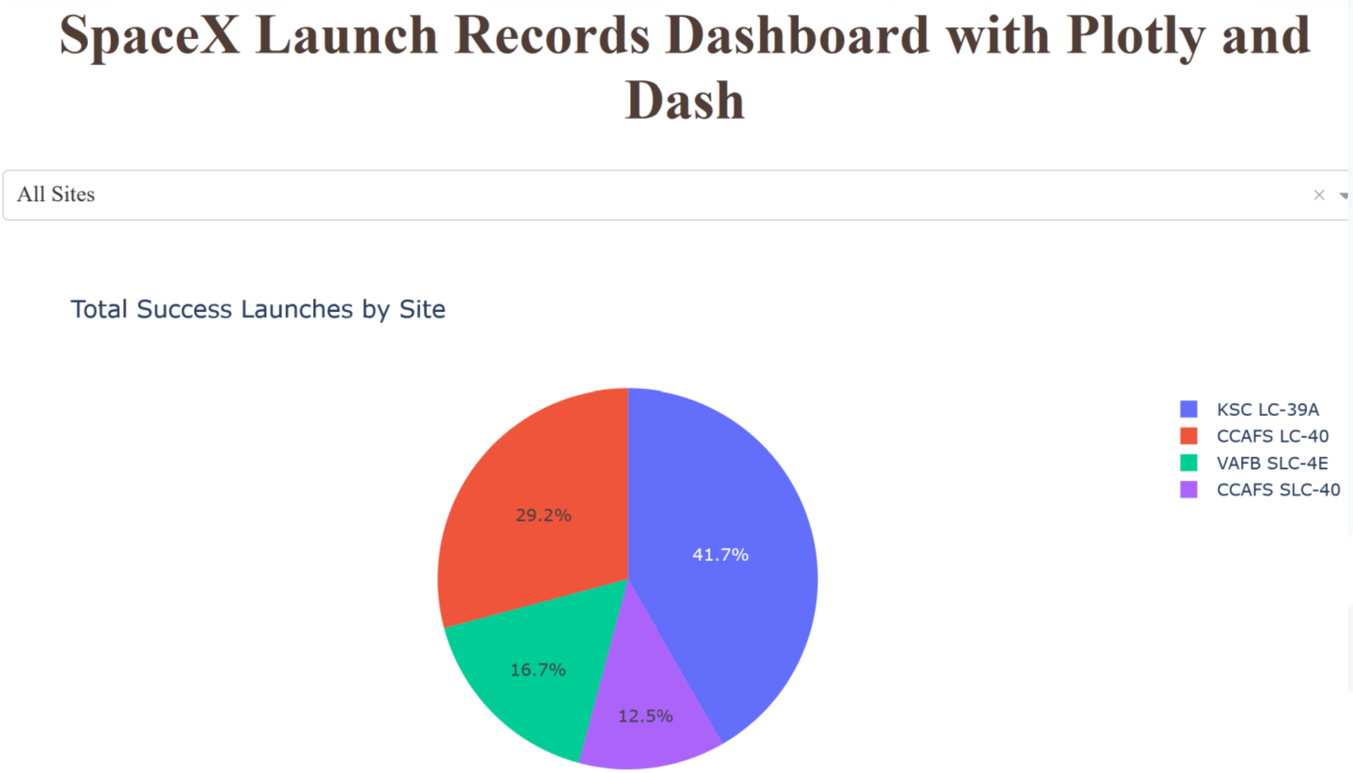
RESULTS

- **Interactive Visual Analytics with Folium**
- Mark the success/failed launches for each site on the map (Green for Success, Red for Failure, this is obtained upon double clicking of the circles in previous plot)



RESULTS

- **Interactive Visual Analytics with Plotly and Dash**
- Total Success Launches by Site (KSC LC-39A has the largest successful launches).



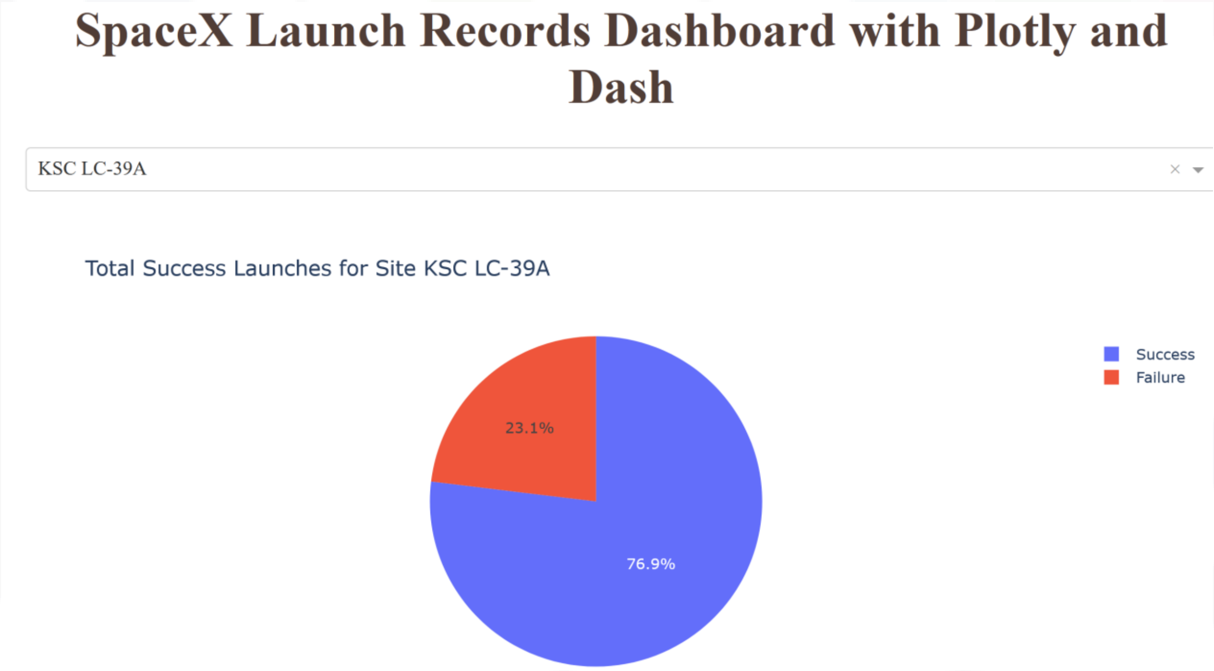
SpaceX Launch Records Dashboard with Plotly and Dash

All Sites

Total Success Launches by Site

RESULTS

- **Interactive Visual Analytics with Plotly and Dash**
- Success Launches Rate by Selecting Specific Site (KSC LC-39A has the highest launch success rate (76.9)).



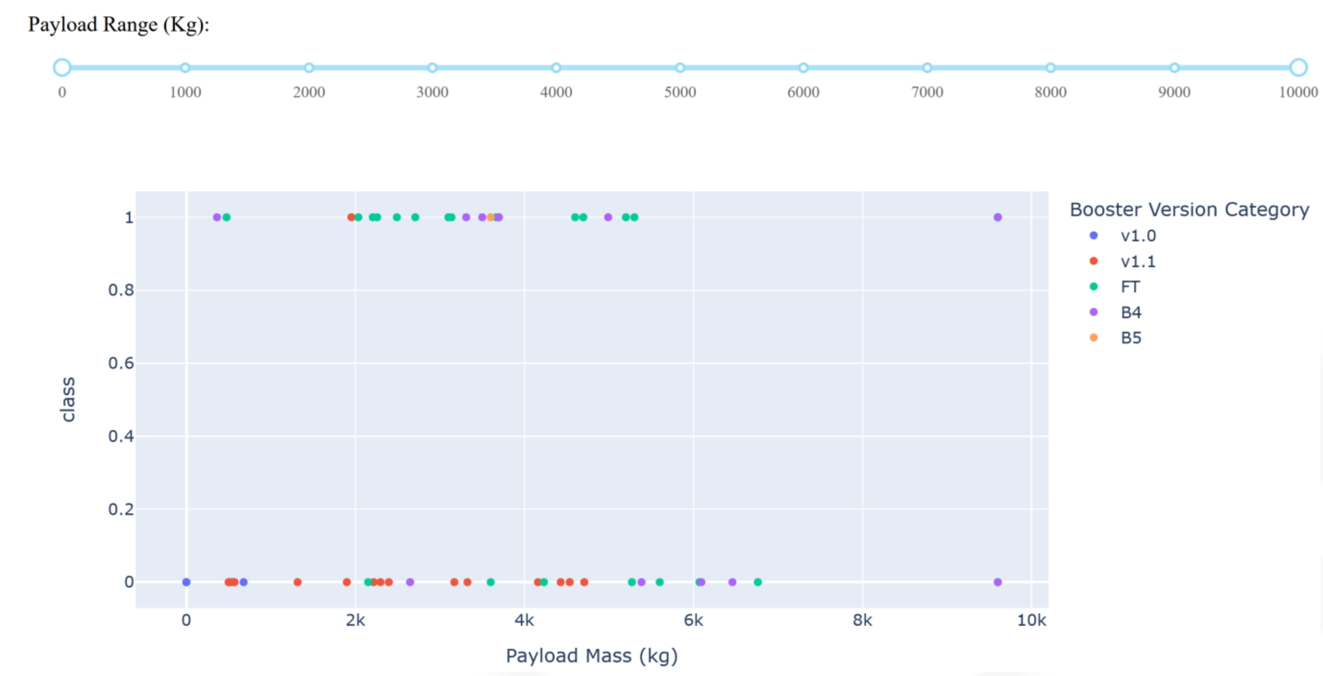
SpaceX Launch Records Dashboard with Plotly and Dash

KSC LC-39A

Total Success Launches for Site KSC LC-39A

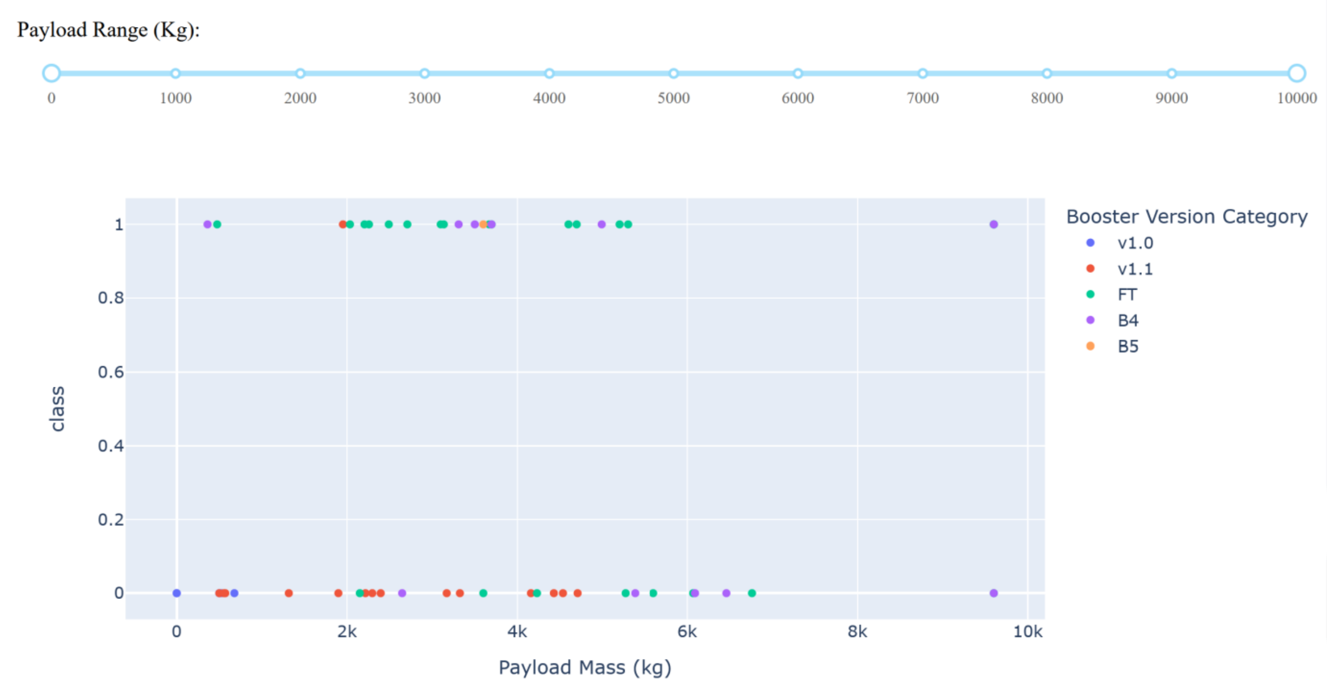
RESULTS

- **Interactive Visual Analytics with Plotly and Dash**
- Considering Payload Mass and Success by Booster Version (Payloads between 2,000 Kg and 5,000 Kg have the highest success rate)



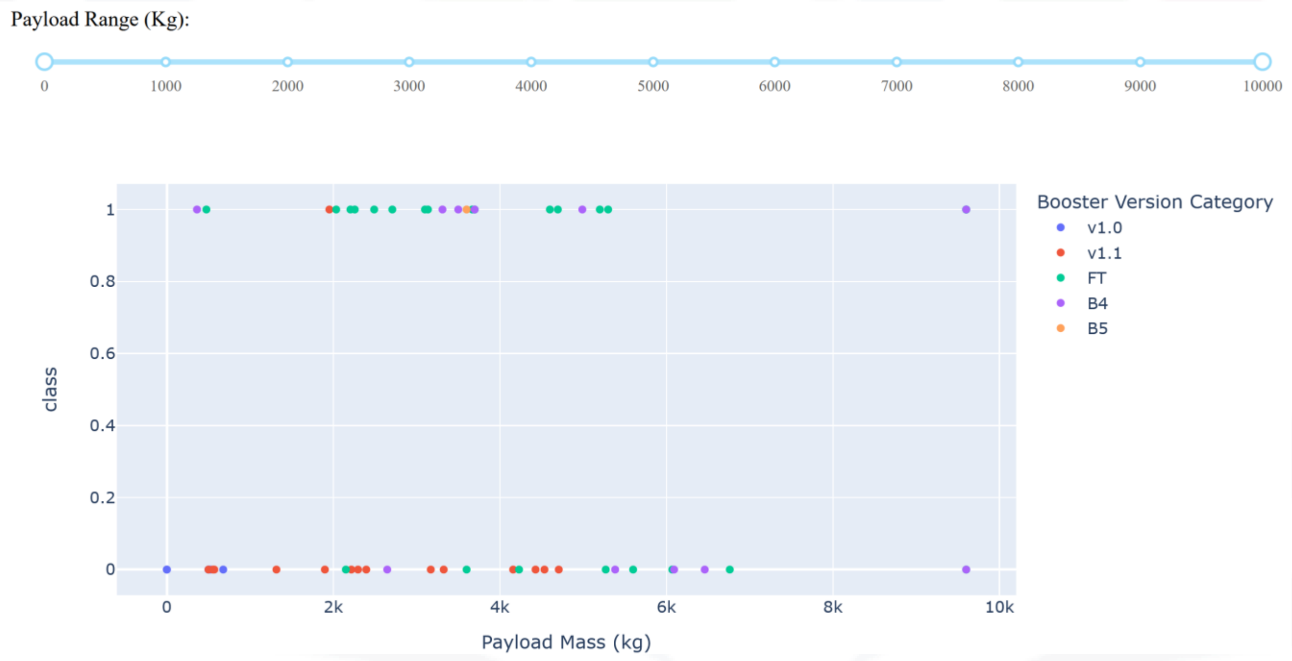
RESULTS

- **Interactive Visual Analytics with Plotly and Dash**
- Considering Payload Mass and Success by Booster Version (Payloads above 6,000Kg have the lowest success rate).



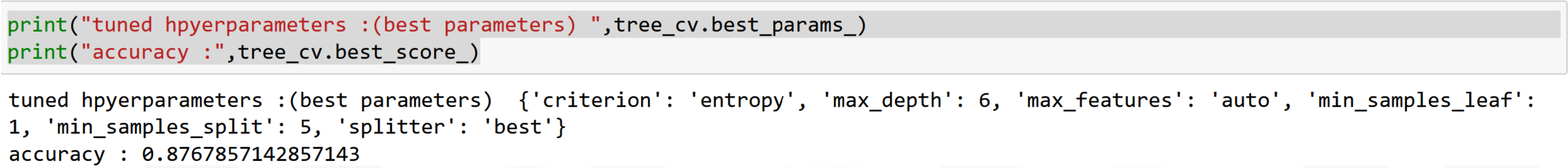
RESULTS

- **Interactive Visual Analytics with Plotly and Dash**
- Considering Payload Mass and Success by Booster Version (F9 Booster version FT has the highest launch success rate).



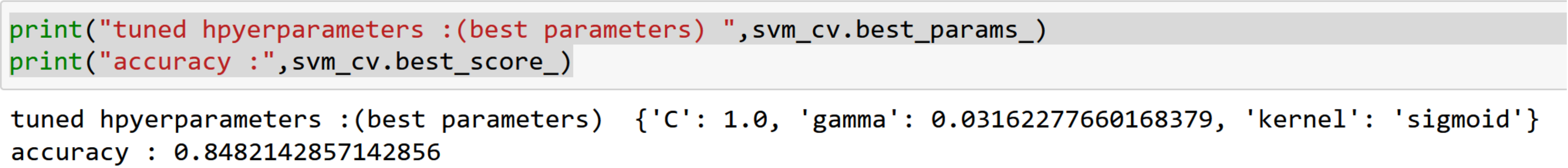
RESULTS

- **Predictive Analysis Results**
- Compare Accuracy of all models (Decision Tree is the best with Accuracy of 0.8768)



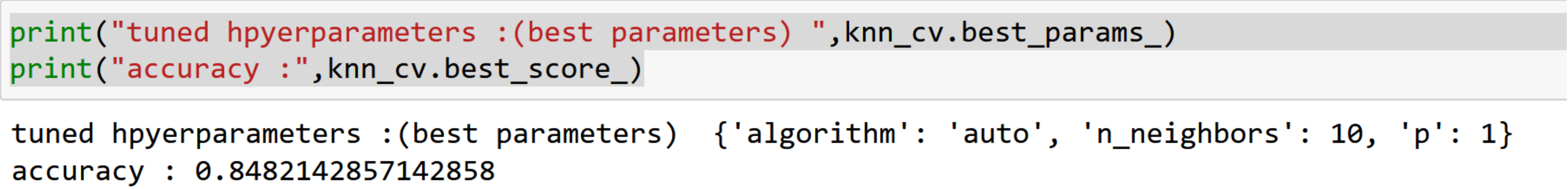
```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf':
1, 'min_samples_split': 5, 'splitter': 'best'}
accuracy : 0.8767857142857143
```



```
print("tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
print("accuracy :",svm_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856
```



```
print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
print("accuracy :",knn_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858
```

OVERALL FINDINGS & IMPLICATIONS

Findings

- All site are near equator line
- All sites are near the coast
- No much difference in accuracy among the used methods
- With time launch success increased

Implications

- Low fuel consumption
- High safety to surroundings
- Selection of any methods mentioned for predict is okay
- Models should be trained with more data for better results.

OVERALL FINDINGS & IMPLICATIONS

Findings

- Some orbits(S-L1, GEO, HEO, and SSO) have success rate of 100%.
- There is a huge difference in success rate among the sites (KSC LC-39A has the highest success rate)
- There is a range of payload mass that has high success rate(2000kg to 5000kG)

Implications

- Selection of orbit is important for high success rate.
- Selection of launching site is important for high success rate.
- Payload mass per booster version should be within the proper range for high success rate.

CONCLUSION

- Project showed that data science can well predict if first launch can land successfully.
- More dataset is needed to increase accuracy in prediction.
- For generalization of results, needed to include data far away from equator line.
- Data preprocessing with methods(Multiscale Principal Component Analysis(MSPCA)) that reduce dimension should be used for accuracy prediction
- Methods that can compare results from multiple methods (Bayser thorem) should be used to decide the final answer for reliable classfication results.

APPENDIX

- Source Codes are found at:
- ramadhani28/AppliedDataScienceCapstone (github.com)