IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tshepo Teffo

18-11-2023

Outline

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

Executive Summary

- Data collection methodology:
  - ➢ Data was collected through SpaceX API and
  - ➢ Web-scraping of SpaceX wiki pages
- Perform data wrangling
  - ➢ Check for percentage of null values in each attribute
  - ➢ Null values of a payload mass column were replaced with the mean of values in that column
  - ➢ Perform one-hot-encoding to categorical columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Logistic regression, SVM, Decision tree, and Kneighbours models were develop using Sklearn Library. Model's evaluation was done using score() method and confusion matrix. Best hyperparameters were found using GridSearchCV object.

Summary of all results

- Exploratory data analysis with sql revealed the following: (1) 2015-12-22 is launch date of the first successful landing, (2) There has been 99 success mission outcomes, 1 failed, and 1 success with unclear payload status
- Further analysis and visualization in python indicated that the landing outcome is dependent on Flight number, orbit type, payload mass, and launch site variables – success rate appeared to be greater where flight numbers and payload mass were higher. It also varied between orbits and launch sites.
- There has been zero success rate from 2010 to 2013 while an increasing trend was notable from 2013 to 2020. However, success rate has been the same in 2014 and 2015.
- CCAFS SLC-40 launch site had the highest success rate of them all. It appears there has been more failed launches than successful ones for all combined sites.
- All classification models built have accuracy score of no less than 83.33% with decision tree classifier having the highest accuracy of them all. The accuracy and misclassification rates are sitting at 83.33% and 16.67%, respectively.

Introduction

- Project background and context:

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.

- Problem:

The cost to launch a rocket is more for other service providers than it is for Space-X, and hence this service providers cannot successfully compete with Space-X in the market.

Section 1

Methodology

"

Methodology

Executive Summary

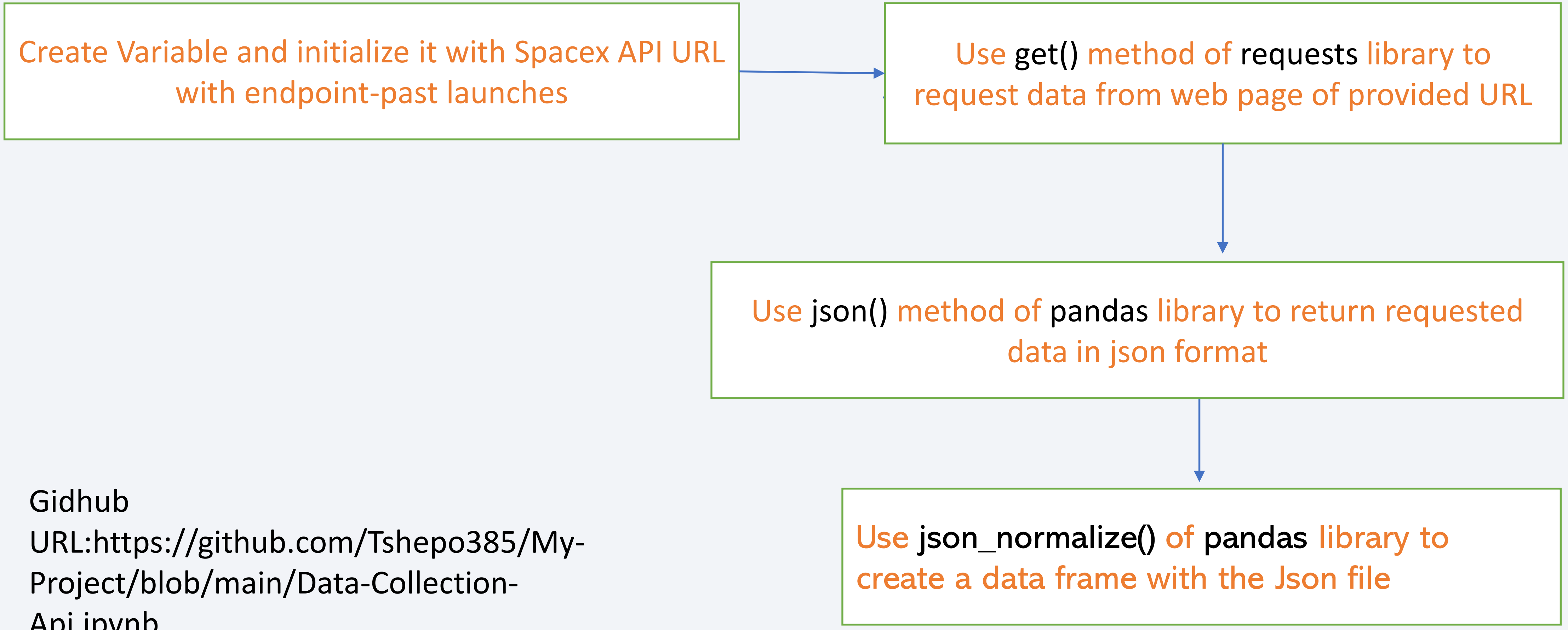
- Data collection methodology:
  - ➢ Data was collected through SpaceX API and
  - ➢ Web-scraping of SpaceX wiki pages
- Perform data wrangling
  - ➢ Checked for percentage of null values in each attribute
  - ➢ Null values of a payload mass column were replaced with the mean of values in that column
  - ➢ Perform one-hot-encoding to categorical columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

Data sets were collected using two methods:

- ➢ SpaceX API – by targeting Space-X url with endpoint, “past launches” and using get() method of requests library to extract the data.
- ➢ Web scrapping – a soup object of the BeautifulSoup library was created. The soup object was then used to extract data from SpaceX wiki page. Tables and table header titles were extracted using find_all() method.

Data Collection – SpaceX API



Create Variable and initialize it with Spacex API URL with endpoint-past launches

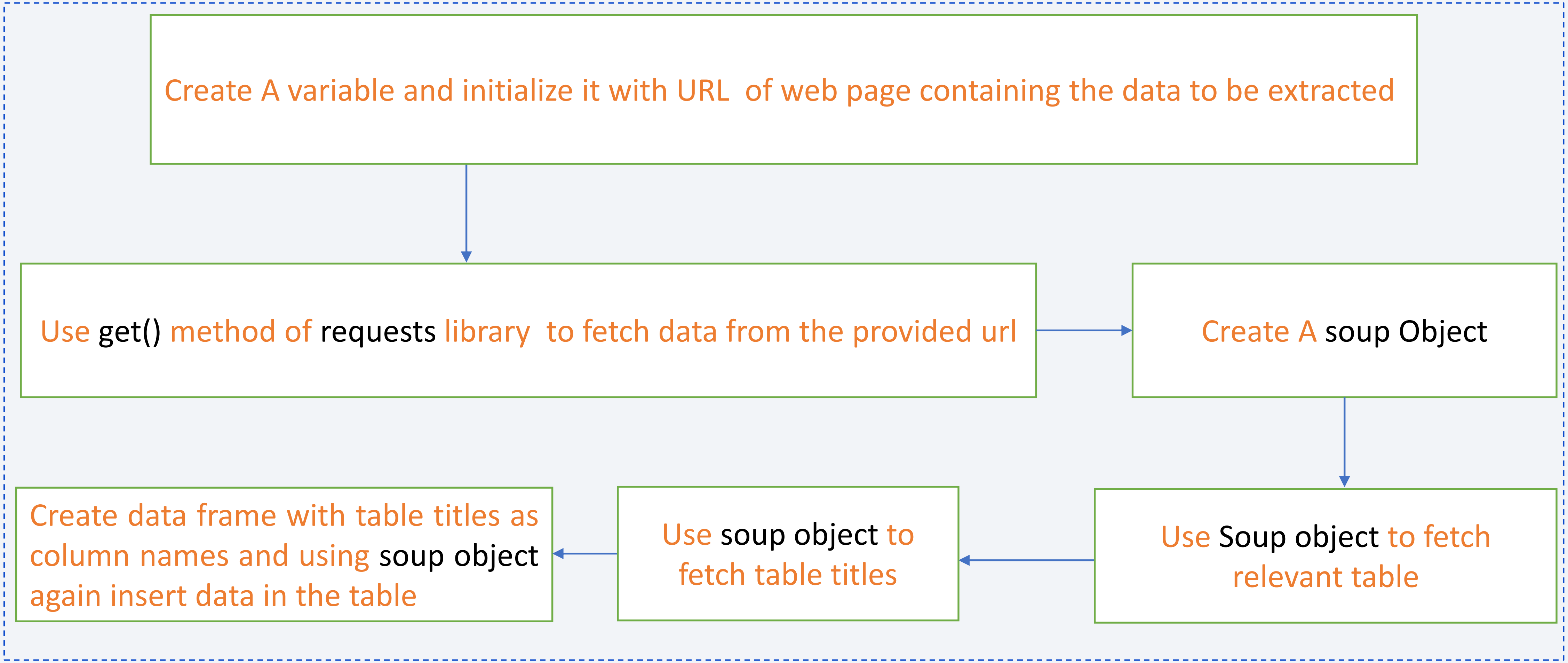
Use get() method of requests library to request data from web page of provided URL

Use json() method of pandas library to return requested data in json format

Use json_normalize() of pandas library to create a data frame with the Json file

Gidhub URL:https://github.com/Tshepo385/My-Project/blob/main/Data-Collection-Api.ipynb

Data Collection - Scraping



Create A variable and initialize it with URL of web page containing the data to be extracted

Use get() method of requests library to fetch data from the provided url

Create A soup Object

Use Soup object to fetch relevant table

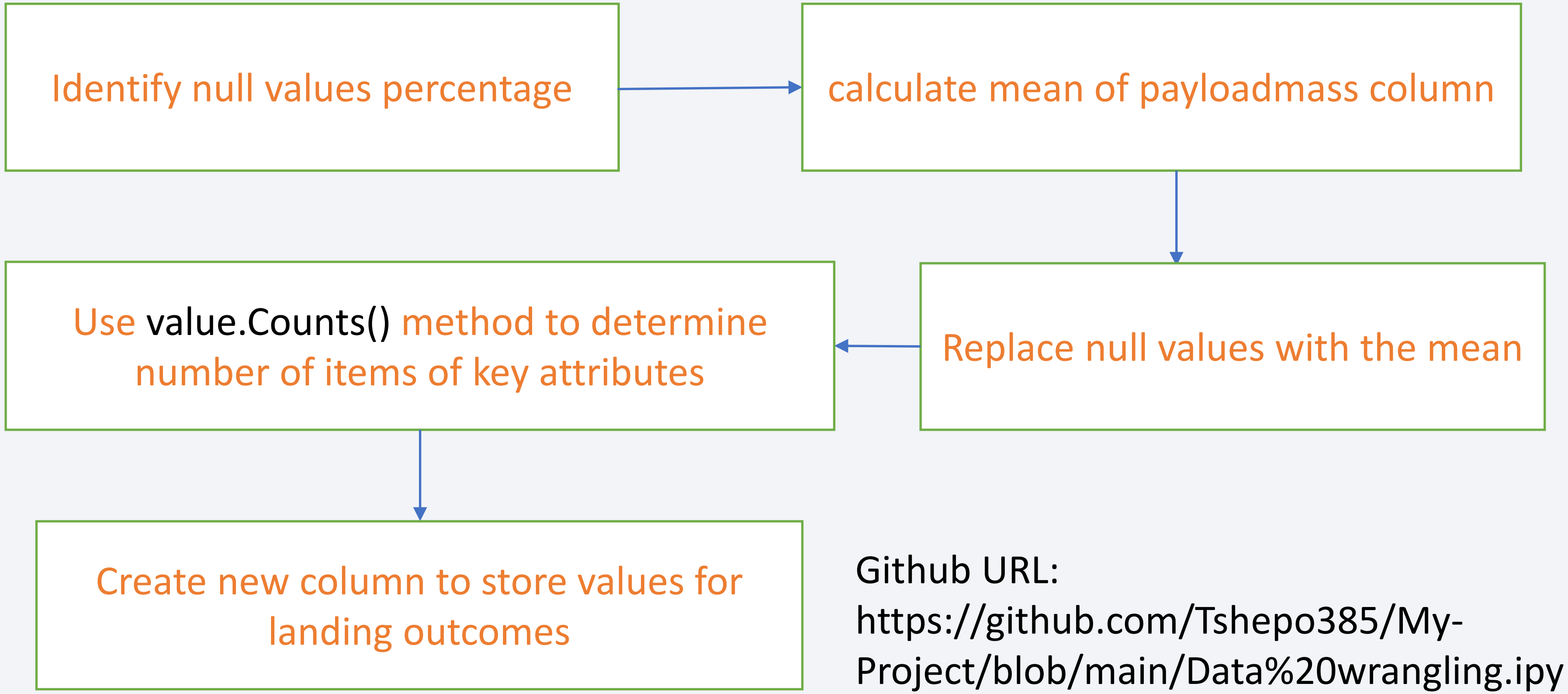
Use soup object to fetch table titles

Create data frame with table titles as column names and using soup object again insert data in the table

Data Collection – Scraping Notebook

Github URL:https://github.com/Tshepo385/My-Project/blob/main/webscraping.ipynb

Data Wrangling



Identify null values percentage

calculate mean of payloadmass column

Replace null values with the mean

Use value.Counts() method to determine number of items of key attributes

Create new column to store values for landing outcomes

Github URL:
https://github.com/Tshepo385/My-Project/blob/main/Data%20wrangling.ipynb

EDA with Data Visualization

The following Charts were plotted:

- Flight Number vs Launch site catplot and scatter plots
- Payload mass vs Launch site catplot and scatter plots
- Flight number vs Orbit catplot and scatter plots
- Yearly success rate line chart
- Success rate vs Orbit bar chart

The charts were used to assess the relationship between the variables and also the outcome.

GitHub URL: https://github.com/Tshepo385/My-Project/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

The executed sql queries were to:

- ➢ Display the names of the unique launch sites in the space mission
- ➢ Display 5 records where launch sites begin with the string 'CCA
- ➢ List 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 booster_versions which have carried the maximum payload mass
- ➢ List the total number of successful and failure mission outcomes
- ➢ List the date when the first succesful landing outcome in ground pad was achieved

Github URL: https://github.com/Tshepo385/My-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build a Dashboard with Plotly Dash

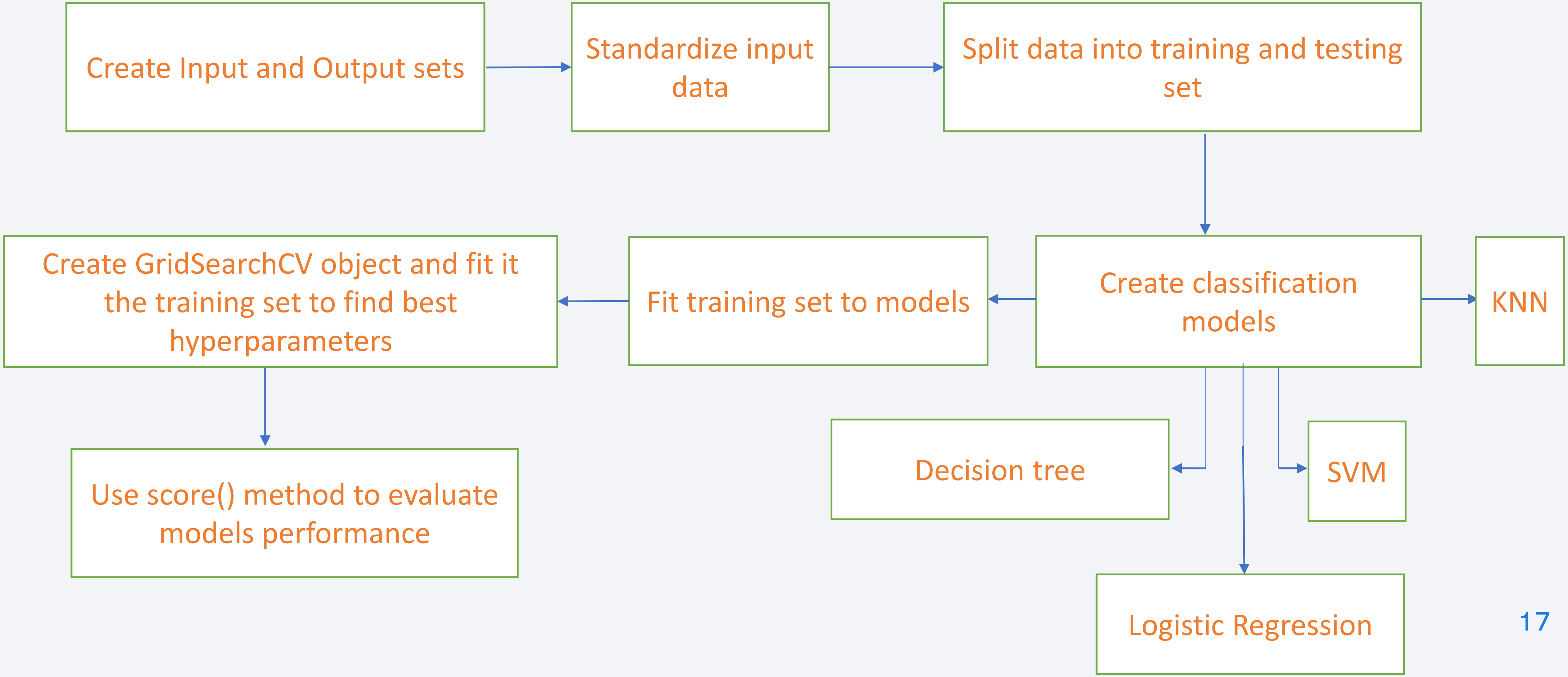
- Pie chart and scatter plots were added to the dashboard to determine the success rate of each site and that of all sites combined.
- A range slider was added to afford the chart the interaction at different payload masses by movement of the slider

Predictive Analysis (Classification)

- Created variables X and Y whose values are input set and output set, respectively.
- Use standardscalar() method to Standardize the input set.
- Divide data frame into training set and testing set using train_test_split() method
- Created model objects namely: logistic regression, SVM, decision tree, and Kneighbours.
- Use fit() method to fit training set to created models
- Create GridSearchCV object and fit it the training set to find best hyperparameters.
- Evaluate model predictions with score() method .

GitHub URL :https://github.com/Tshepo385/My-Project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Predictive Analysis (Process)



Create Input and Output sets

Standardize input data

Split data into training and testing set

Create classification models

KNN

Decision tree

SVM

Logistic Regression

Fit training set to models

Create GridSearchCV object and fit it the training set to find best hyperparameters

Use score() method to evaluate models performance

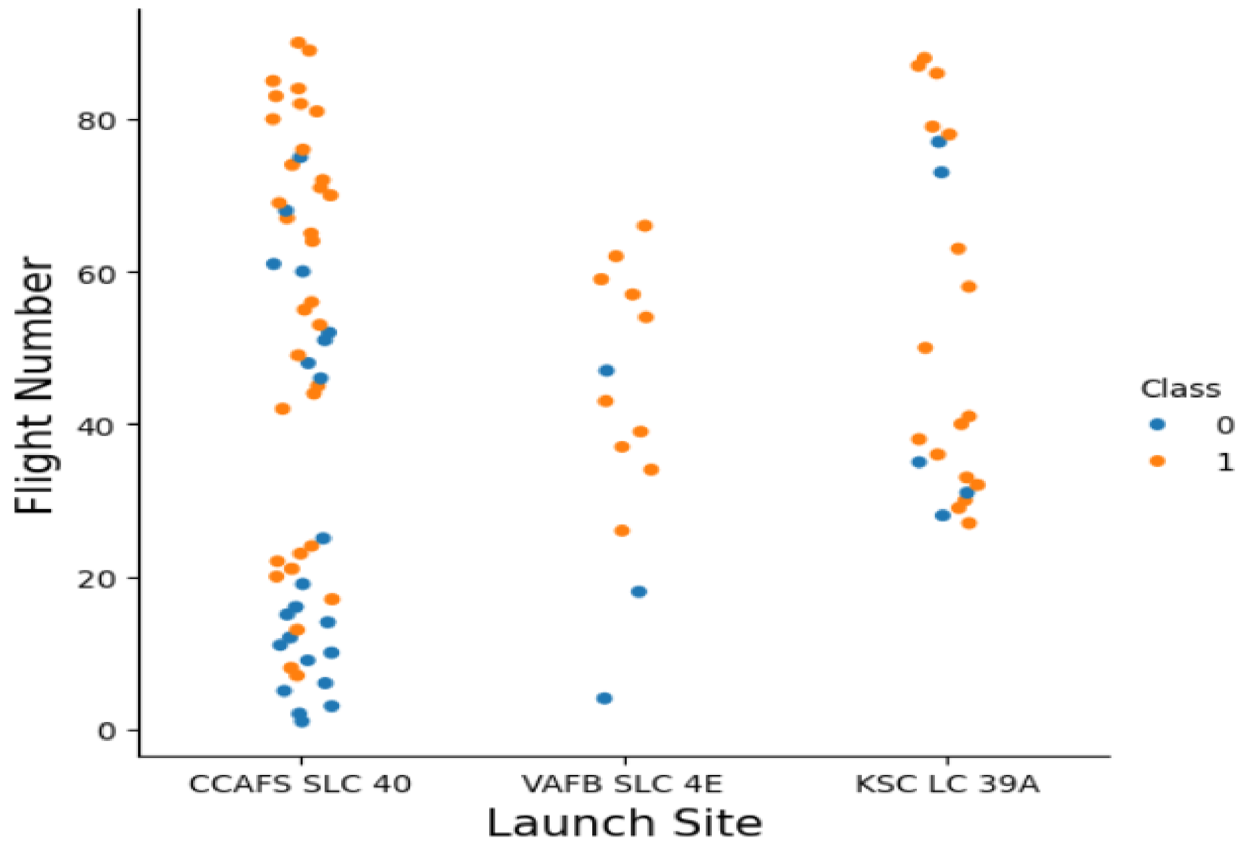
Results

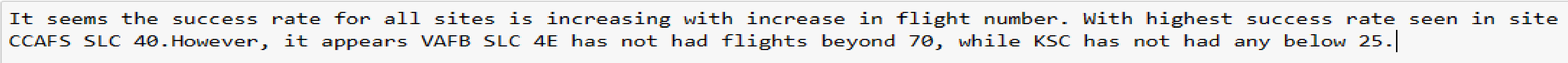
- ➢ Exploratory data analysis results
- ➢ Interactive analytics demo in screenshots
- ➢ Predictive analysis results

Section 2

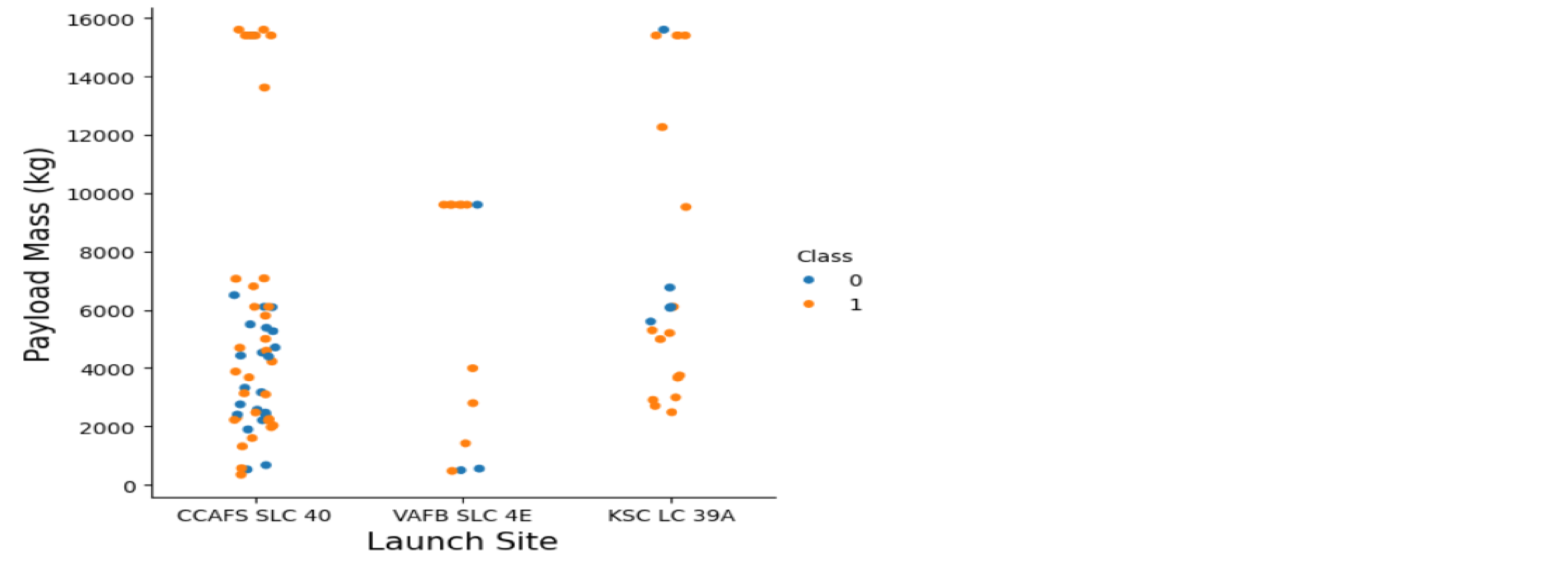
Insights drawn from EDA

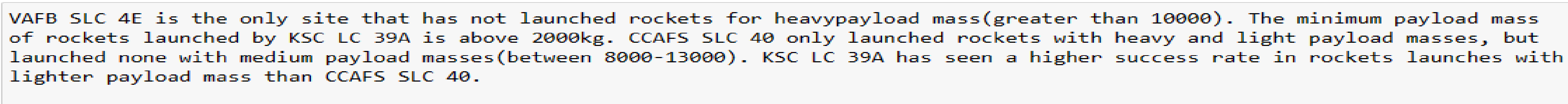
Flight Number vs. Launch Site




It seems the success rate for all sites is increasing with increase in flight number. With highest success rate seen in site CCAFS SLC 40.However, it appears VAFB SLC 4E has not had flights beyond 70, while KSC has not had any below 25.

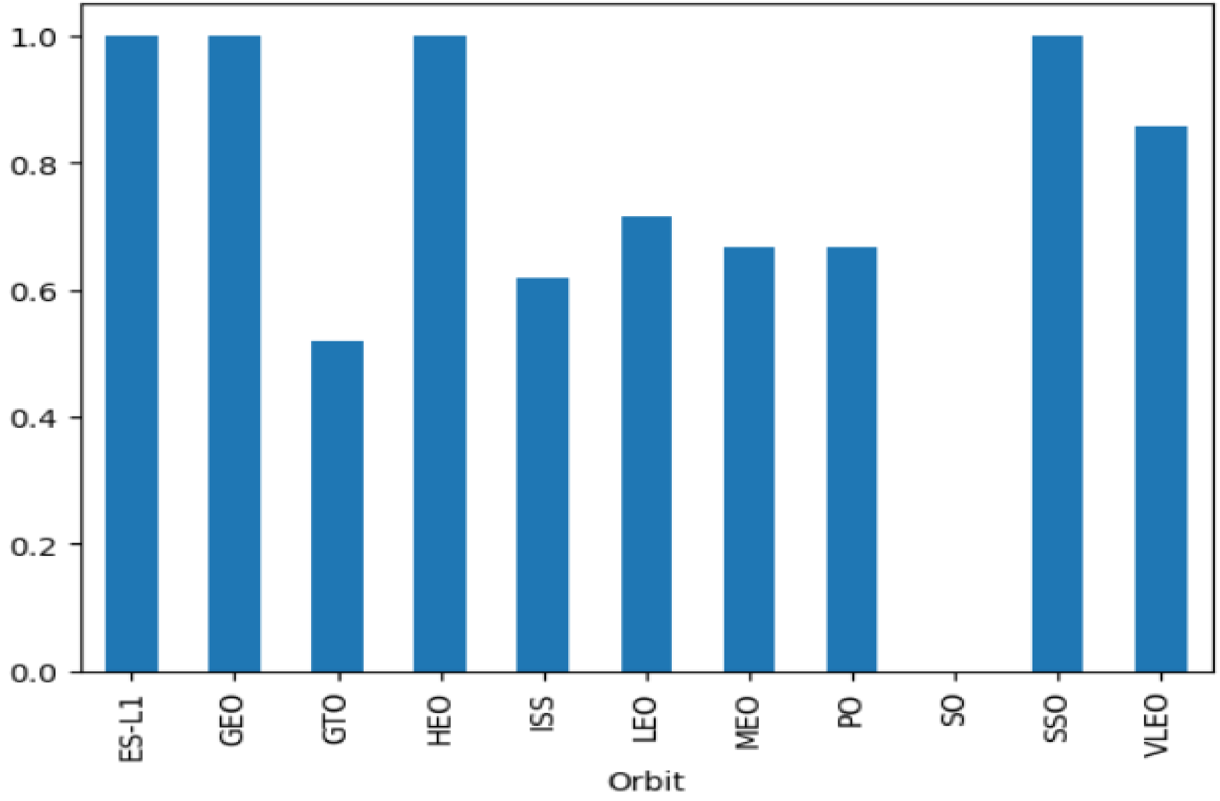
Payload vs. Launch Site





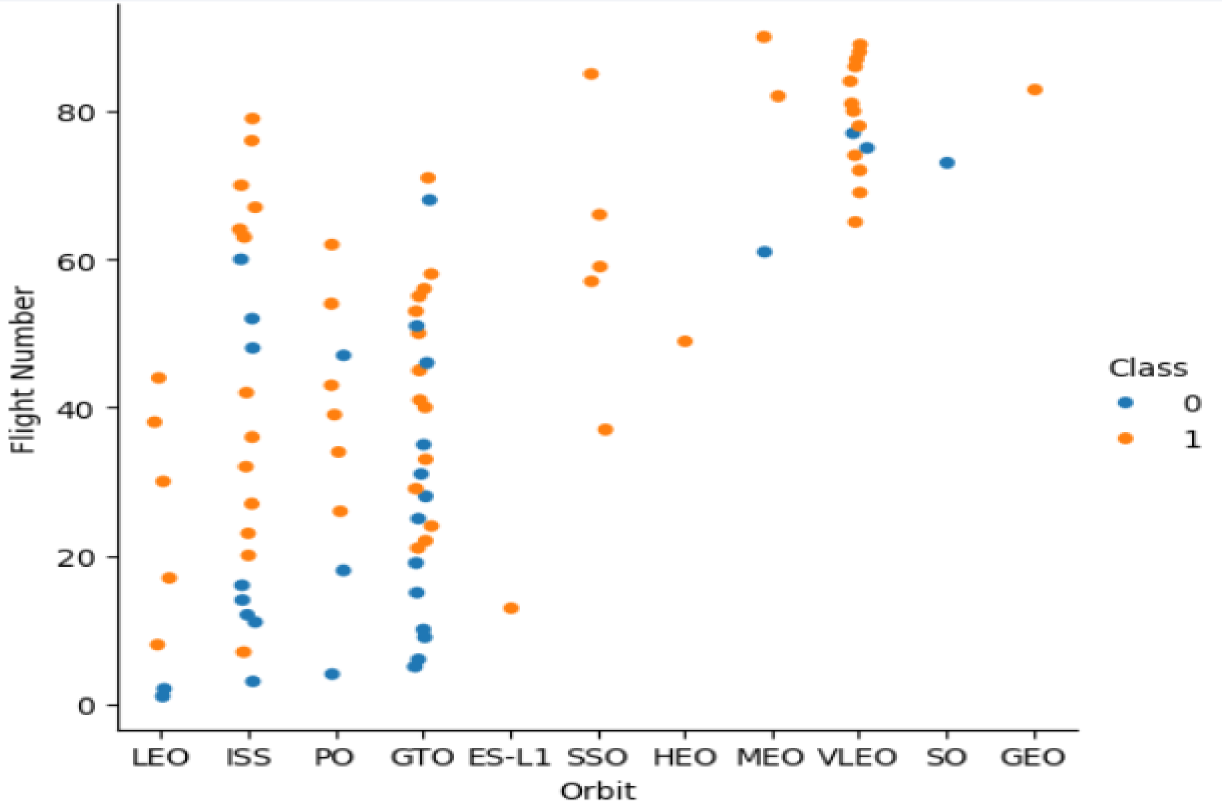
```
VAFB SLC 4E is the only site that has not launched rockets for heavypayload mass(greater than 10000). The minimum payload mass
of rockets launched by KSC LC 39A is above 2000kg. CCAFS SLC 40 only launched rockets with heavy and light payload masses, but
launched none with medium payload masses(between 8000-13000). KSC LC 39A has seen a higher success rate in rockets launches with
lighter payload mass than CCAFS SLC 40.
```

Success Rate vs. Orbit Type



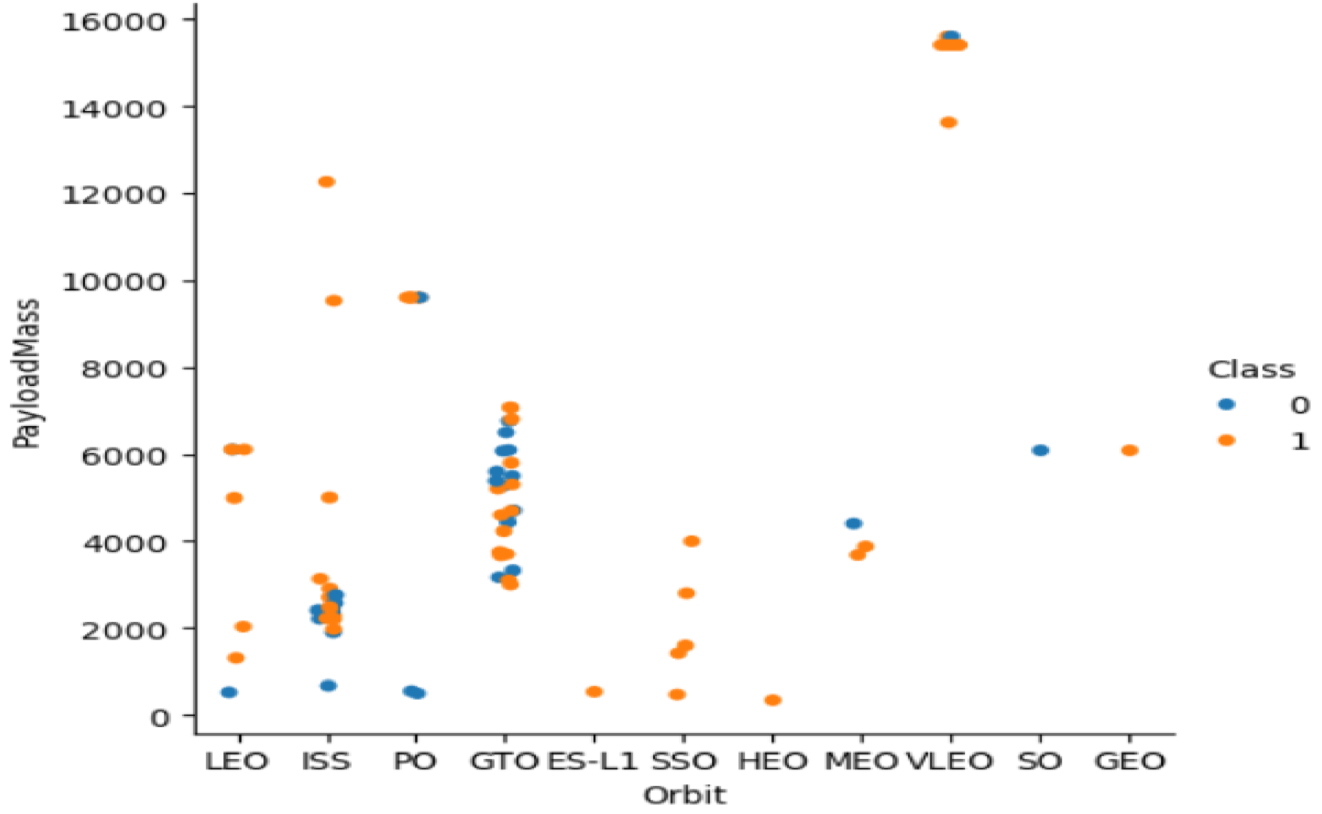
ES-L1, GEO, HEO, and SSO are orbits with the highest sucess rate. VLEO is the second highest. GTO is the only orbit with the lowest success rate.

Flight Number vs. Orbit Type



You should 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.

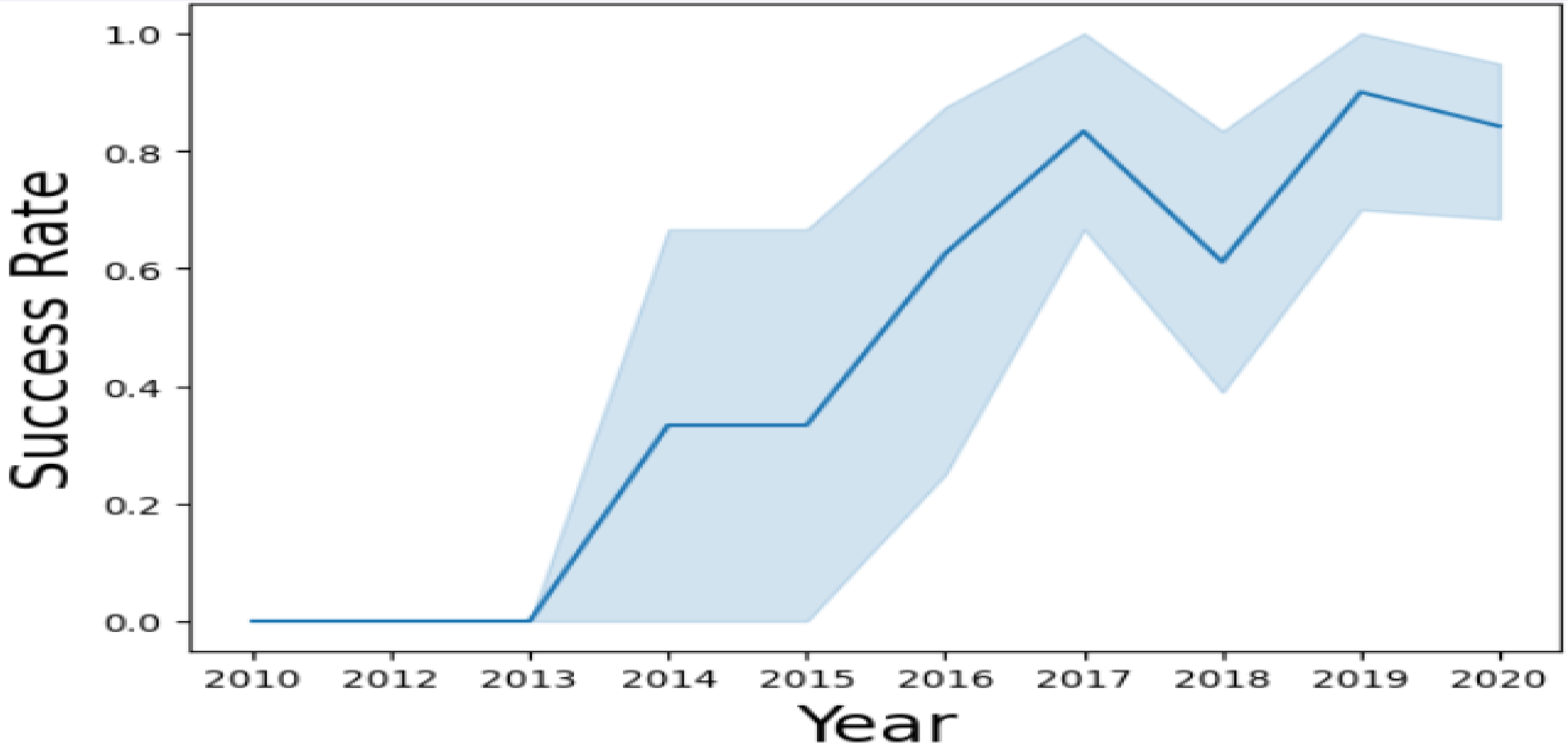
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

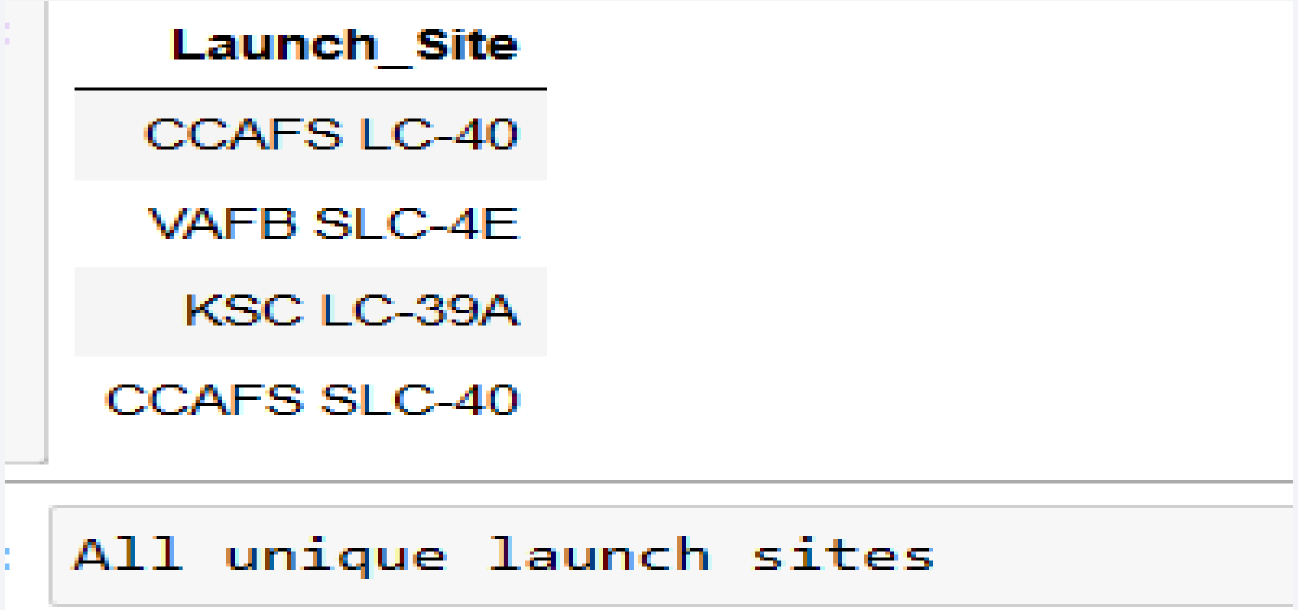
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

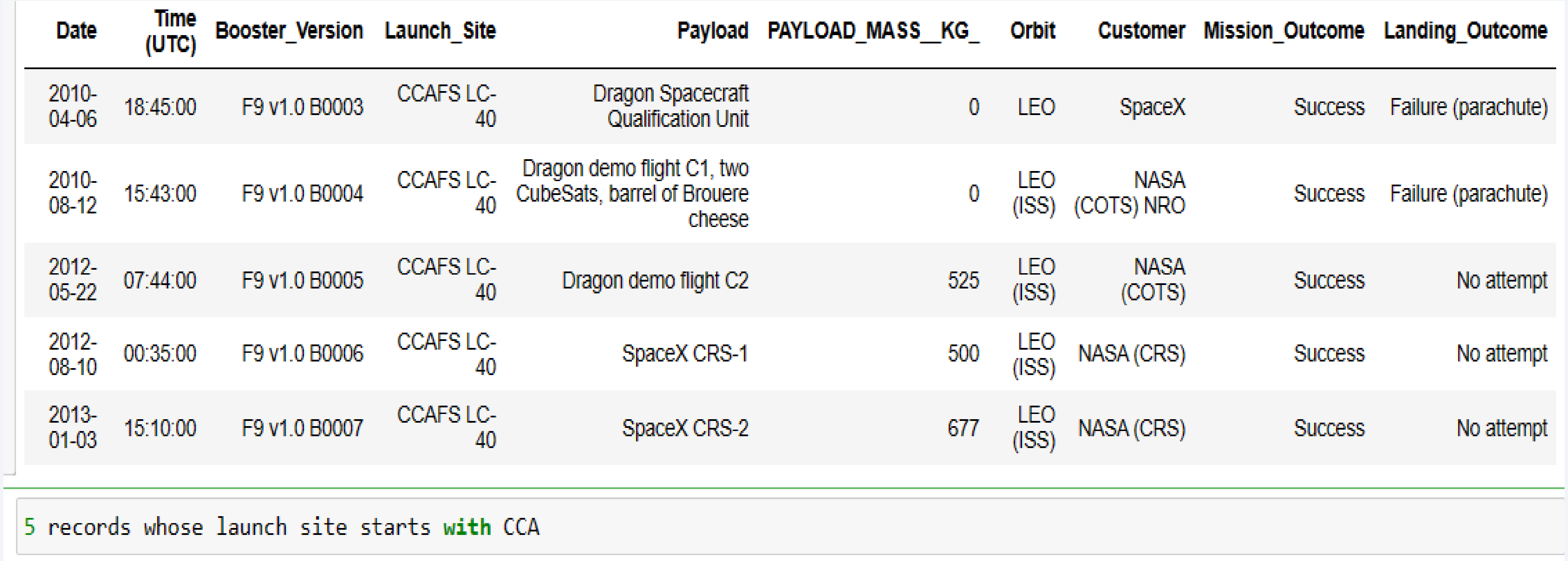
All Launch Site Names



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

```
All unique launch sites
```

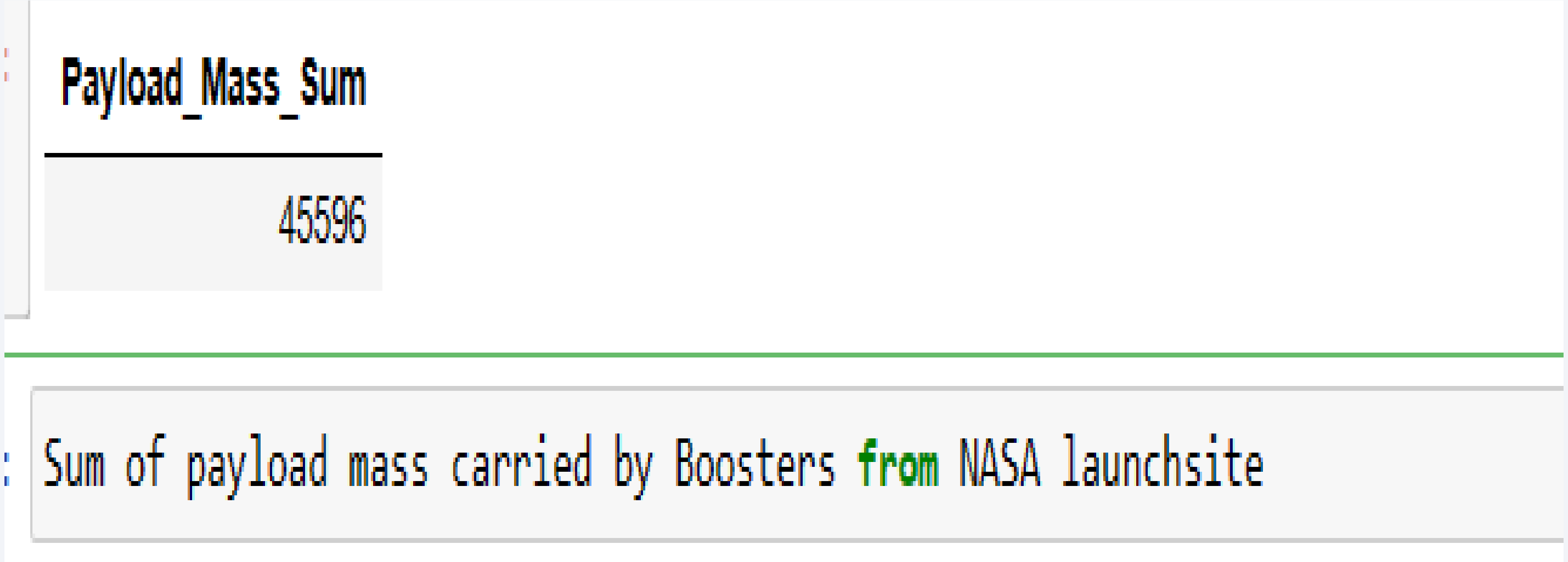
Launch Site Names Begin with 'CCA'



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

```
5 records whose launch site starts with CCA
```

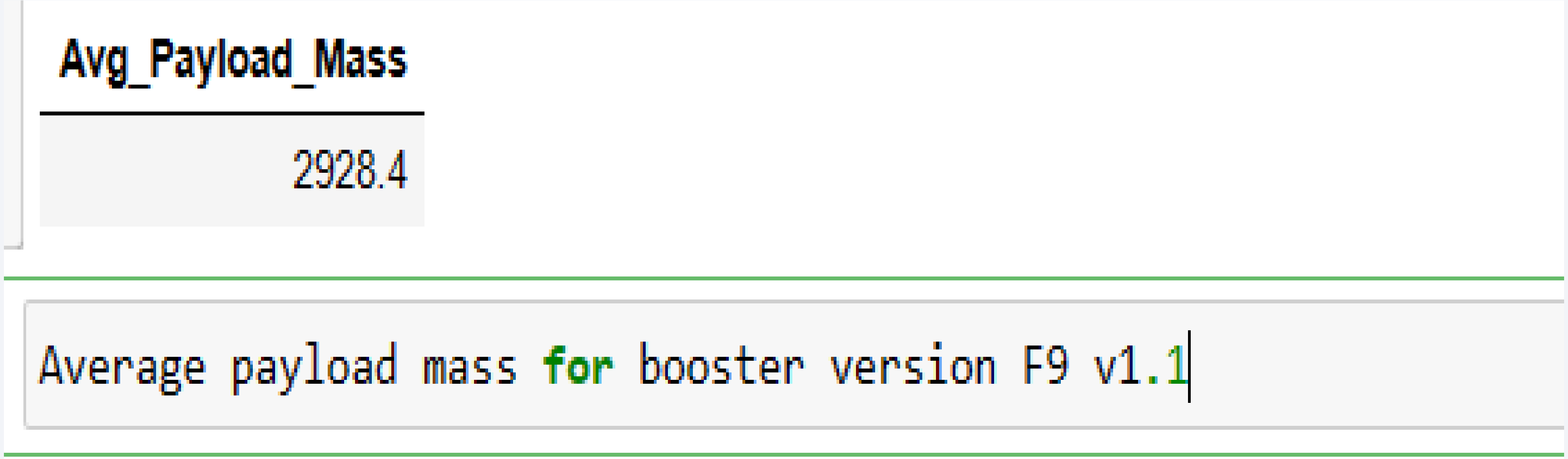
Total Payload Mass



Payload_Mass_Sum
45596

```
Sum of payload mass carried by Boosters from NASA launchsite
```

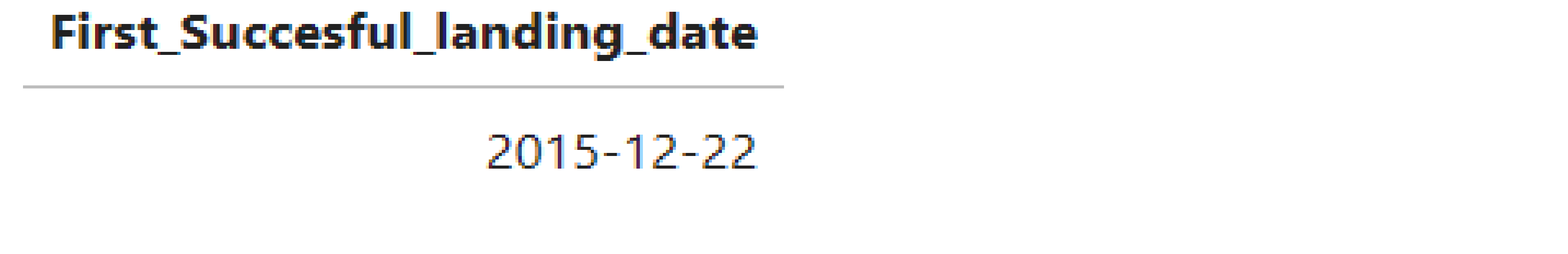
Average Payload Mass by F9 v1.1



Avg_Payload_Mass
2928.4

```
Average payload mass for booster version F9 v1.1
```

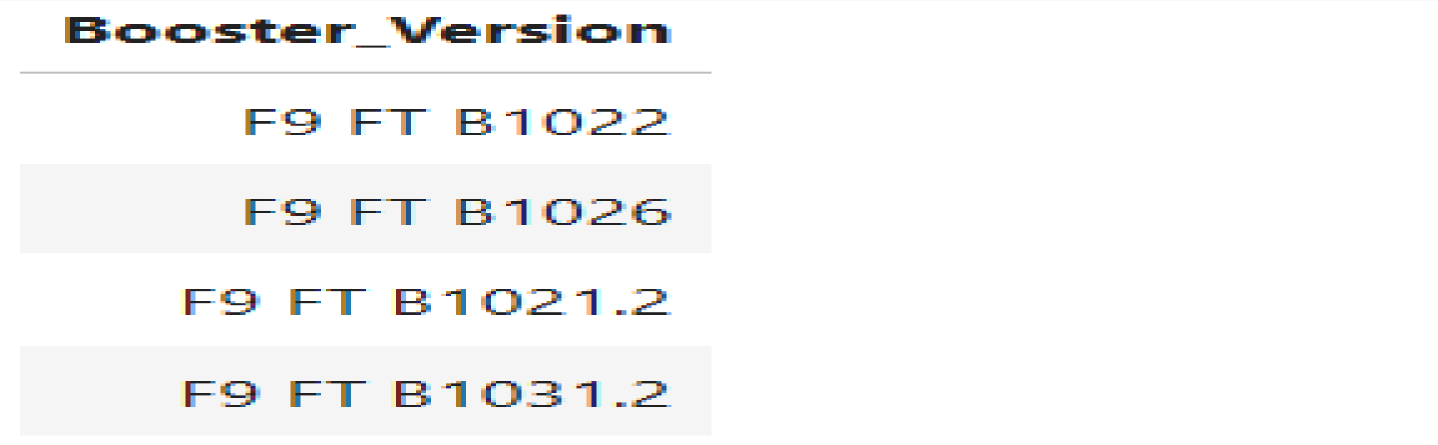
First Successful Ground Landing Date



First_Succesful_landing_date
2015-12-22

- Launch date of the first successful landing

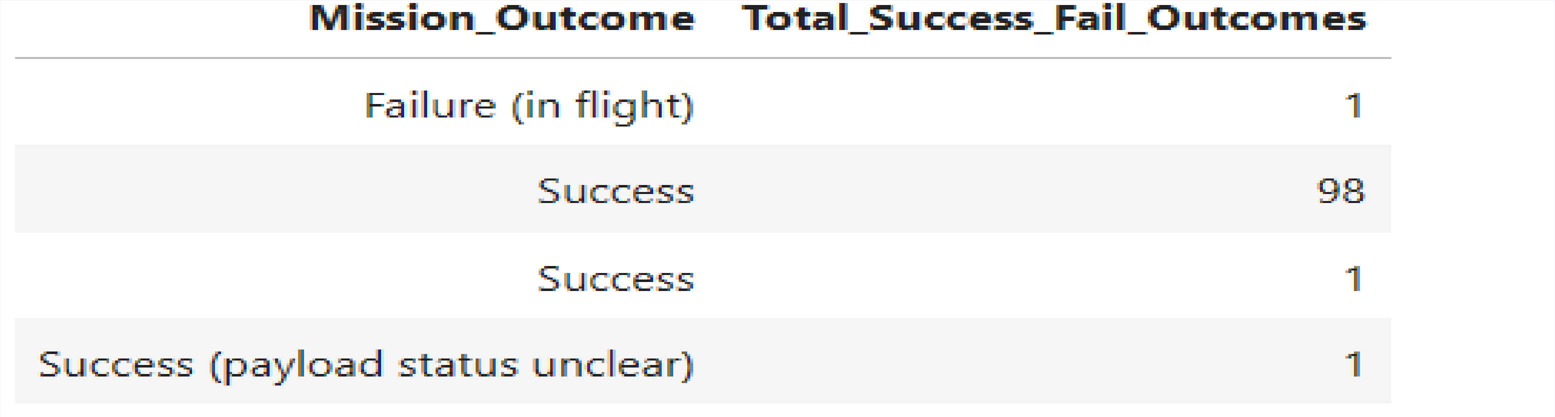
Successful Drone Ship Landing with Payload between 4000 and 6000



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

- The above booster versions have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

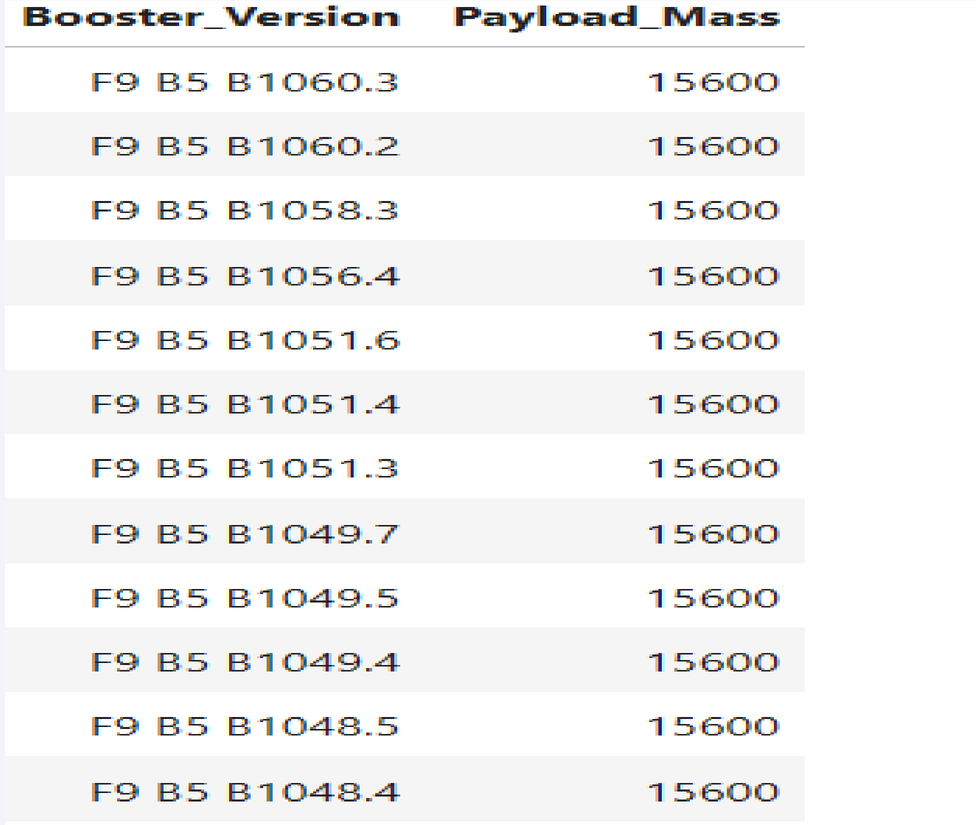
Total Number of Successful and Failure Mission Outcomes



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

- ➢ It appears there have been 99 success mission outcomes, 1 failed, and 1 success with unclear payload status

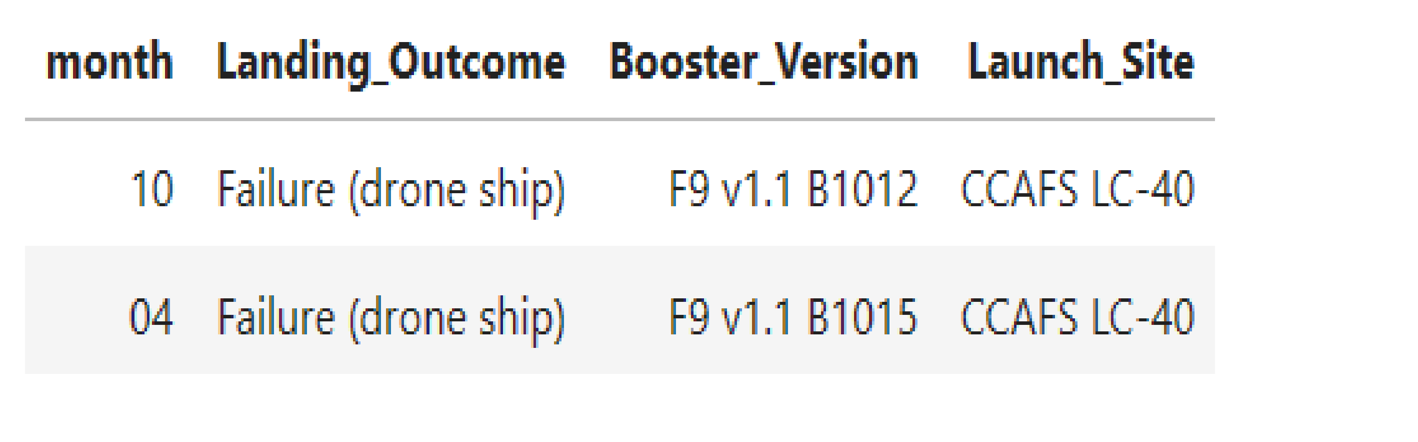
Boosters Carried Maximum Payload



Booster_Version	Payload_Mass
F9 B5 B1060.3	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1056.4	15600
F9 B5 B1051.6	15600
F9 B5 B1051.4	15600
F9 B5 B1051.3	15600
F9 B5 B1049.7	15600
F9 B5 B1049.5	15600
F9 B5 B1049.4	15600
F9 B5 B1048.5	15600
F9 B5 B1048.4	15600

- ➢ The above booster versions carried the maximum payload mass than other versions

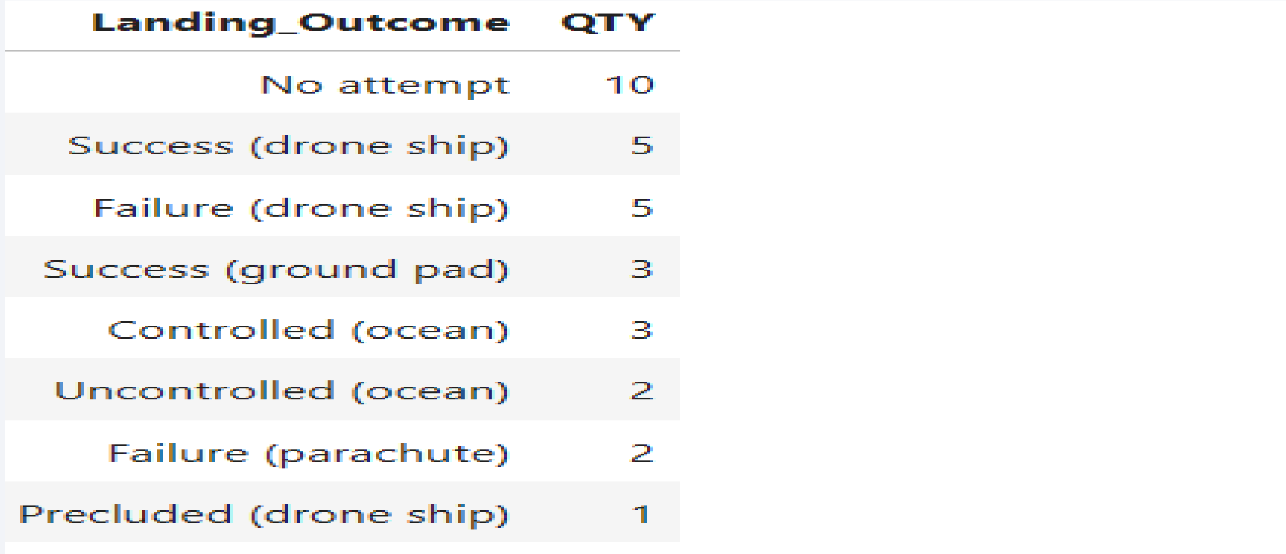
2015 Launch Records



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

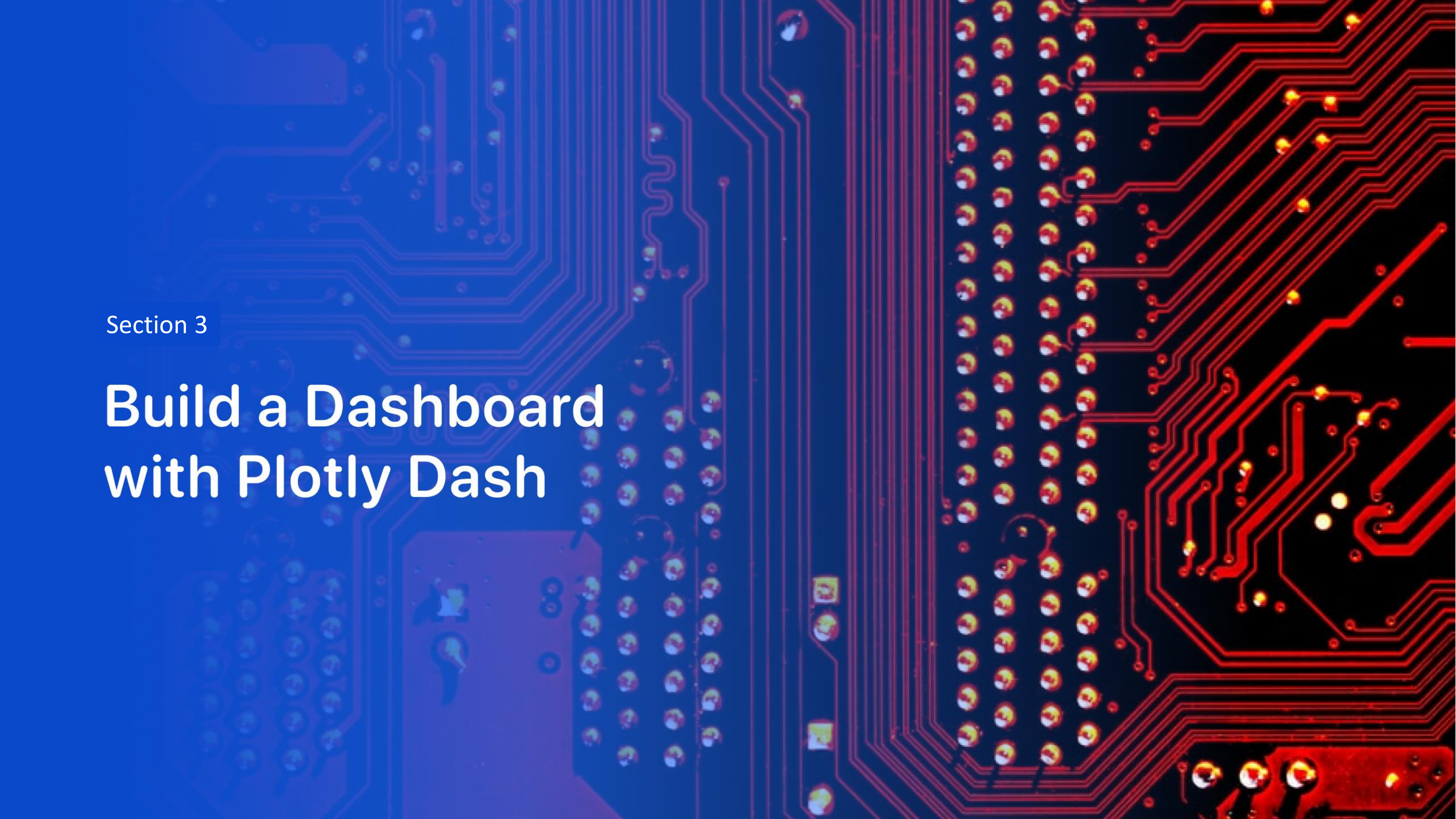
- CCAFS LC-40 appears to be the only launch site with failure (drone ship) landing outcome in months 10 and 04 of 2015

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



Landing_Outcome	QTY
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

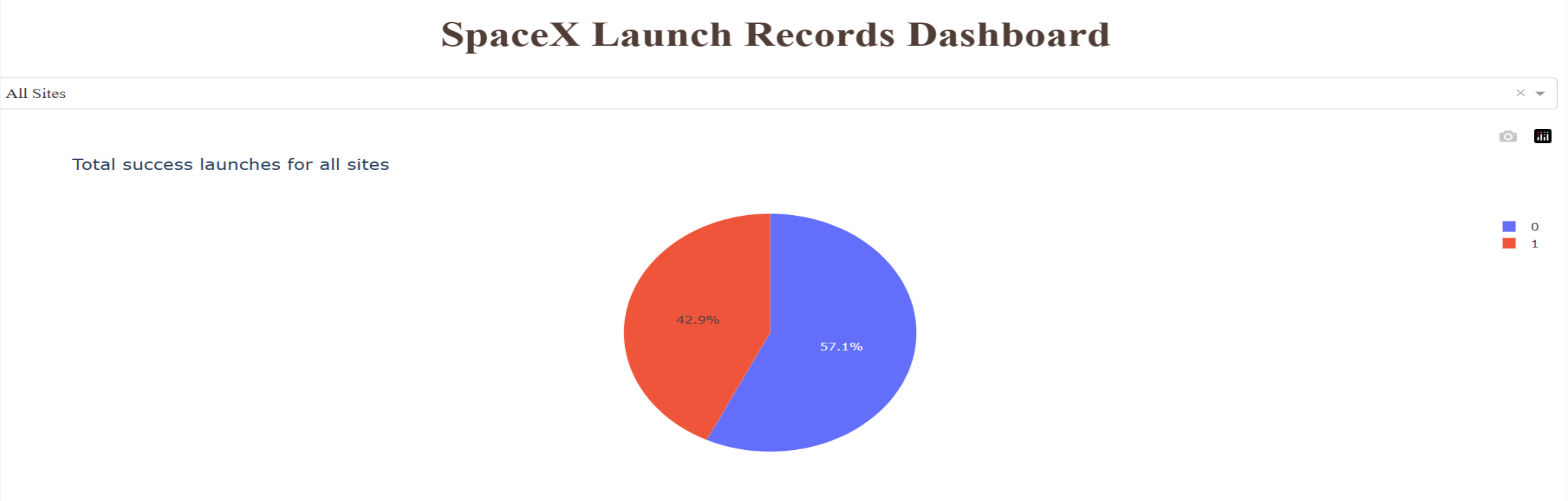
- There has been the least number of Precluded (drone ship) landing outcome than any other landing outcome. Number of No attempts is greatest at ten. While the number of success (drone ship) is 8.



Section 3

Build a Dashboard with Plotly Dash

Total Success Launches For All Sites



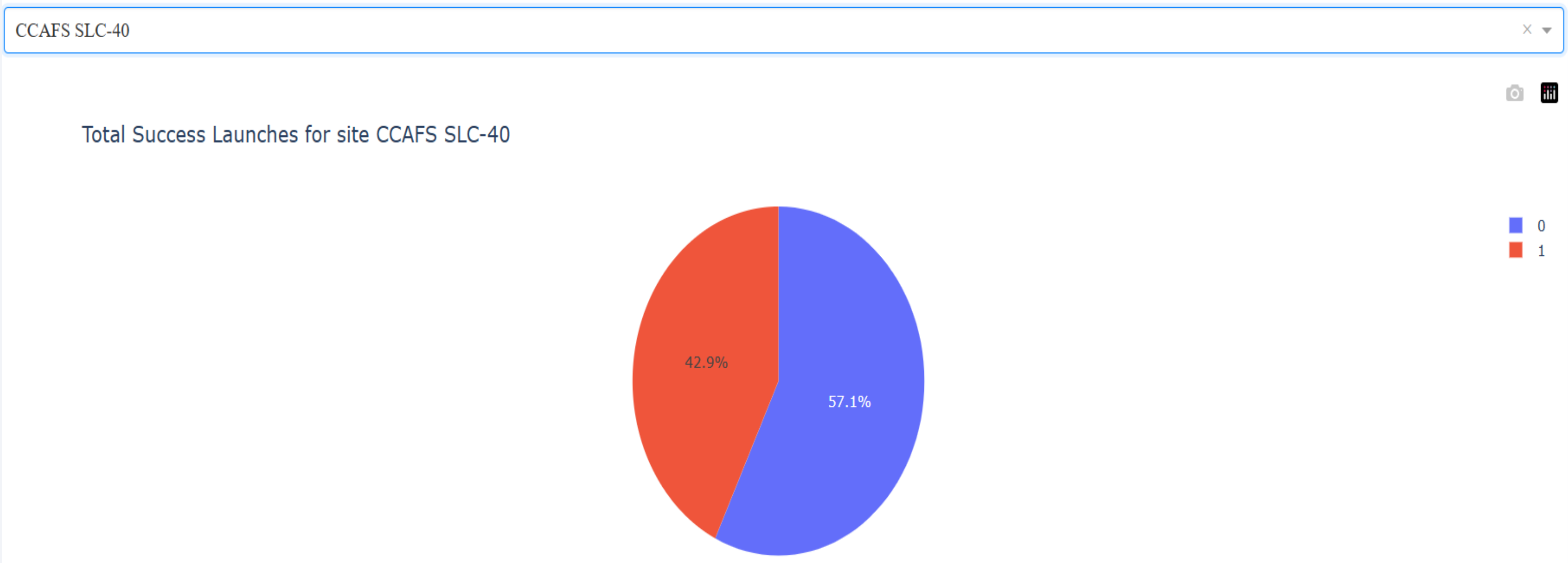
SpaceX Launch Records Dashboard

All Sites

Total success launches for all sites

- ➢ It appears there has been more failed launches than successful ones for all sites. As can be seen, failed launches are sitting at 57.1% while successful ones are 42.9%.

Launch Site With The Highest Success Ratio



CCAFS SLC-40

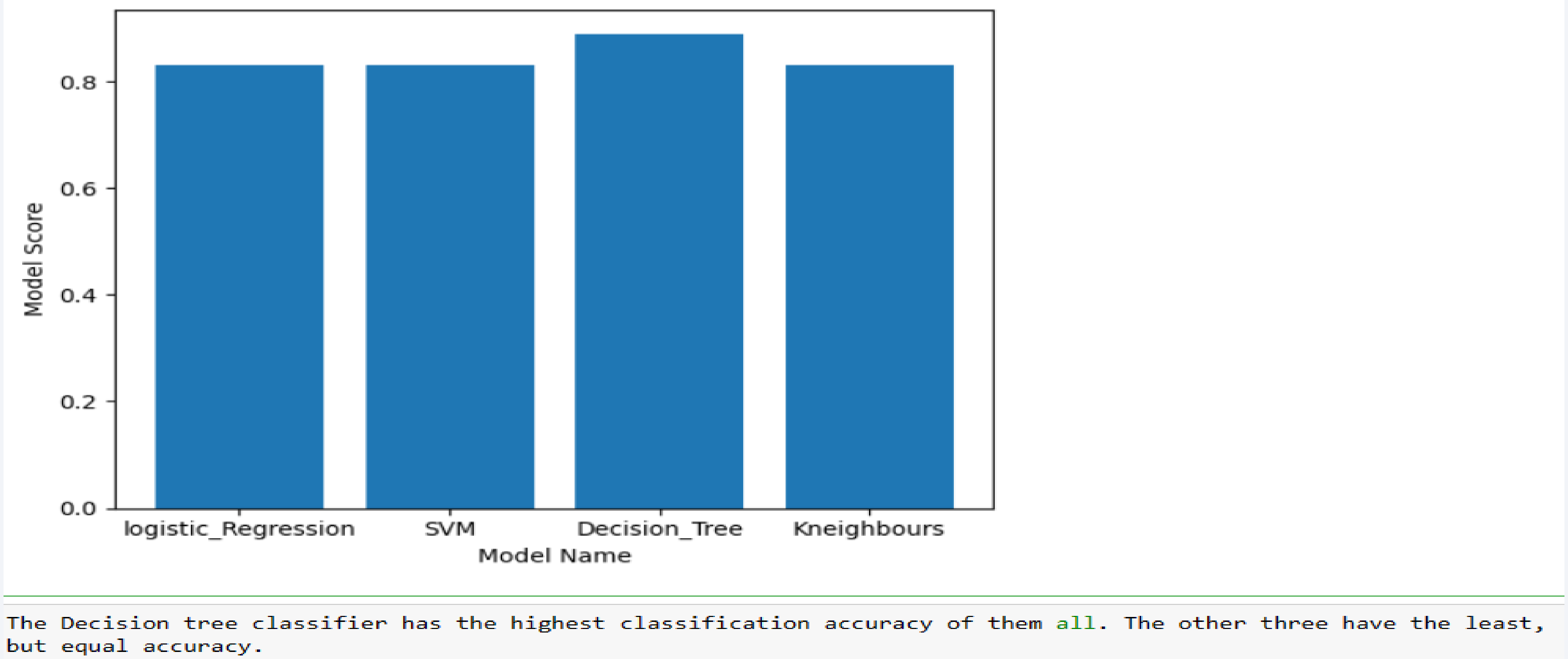
Total Success Launches for site CCAFS SLC-40

- ➢ CCAFS SLC-40 has seen the highest success ratio of launches than any other launch site.

Section 4

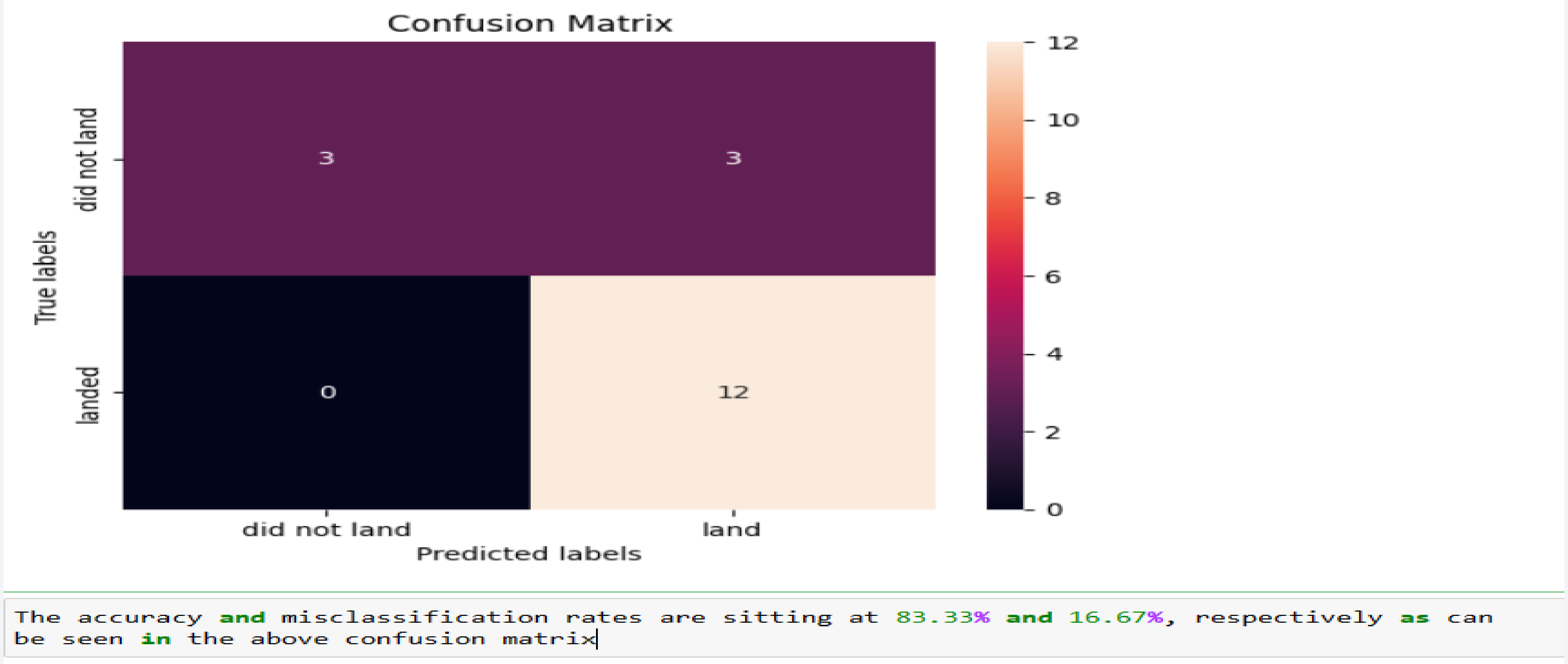
Predictive Analysis (Classification)

Classification Accuracy



```
The Decision tree classifier has the highest classification accuracy of them all. The other three have the least,
but equal accuracy.
```

Confusion Matrix



The accuracy and misclassification rates are sitting at 83.33% and 16.67%, respectively as can be seen in the above confusion matrix

Conclusions

- Objective: To reduce the cost to launch a rocket by using Space-X dataset to predict if the first stage will successfully land and advocate for re-use of such stage.
- Datasets were successfully collected using Space-X API and Web scrapping Space-X wiki pages.
- Exploratory data analysis with sql revealed the following:1) 2015-12-22 is launch date of the first successful landing, 2) There have been 99 success mission outcomes, 1 failed, and 1 success with unclear payload status
- Further analysis and visualization in python indicated that the landing outcome is dependent on Flight number, orbit type, payload mass, and launch site.
- There has been zero success rate from 2010 to 2013 while an increasing trend was notable from 2013 to 2020. However, success rate has been the same in 2014 and 2015.
- CCAFS SLC-40 launch site had the highest success rate of them all. It appears there has been more failed launches than successful ones for all combined sites.
- All classification models built have accuracy score of no less than 83.33% with decision tree classifier having the highest accuracy of them all. The accuracy and misclassification rates are sitting at 83.33% and 16.67%, respectively. This coupled with the above points indicate that there is 83.33% probability that the first stage will indeed land successfully.

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!