IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Elizabeth Ifedayo Owolabi
1st January, 2022

Outline

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

Executive Summary

- To carry out the data analysis on SpaceX Falcon 9 launch data, the data was gathered from SPACEX REST API and Wikipedia page on Falcon 9 launching outcomes (using Web scraping). After gathering the data, the data was wrangled to deal with missing values and properly label landing outcomes. The data was explored using data visualization, SQL queries, interactive map with Folium and Plotly Dash which highlighted a lot of information about the data such as
    - how payload is surprisingly not a good determinant of landing outcomes,
    - how flight number, launch sites and orbit are correlated,
    - KSC LC-39A the most successful launch site,
    - there have been an upward trend of successful launch rates since 2013
- Finally, predictive Analysis was done to determine if a landing outcome will be successful on the first stage using various classification models. This resulted in our model being able to significantly predict successful and failed landing outcomes

Introduction

- Companies are making space travel available for everyone. A key company in this field is SpaceX who has been able to accomplish a lot. SpaceX is able achieve a lot because their rocket launches are relatively inexpensive as they can reuse the first stage
- The first stage is where most of the work is done and it is much larger than the second stage. At the first stage, sometimes
  - the first stage does not land,
  - the first stage will crash
  - SpaceX will sacrifice the first stage due to mission parameters like payload, orbit and customer
- This stage is quite large and expensive. However, unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage.
- SpaceY would like to bid against SpaceX. For successful bidding, we are going to analyse SpaceX data to determine the price of each launch based on if the first stage lands and determine if SpaceX will reuse the first stage

Section 1

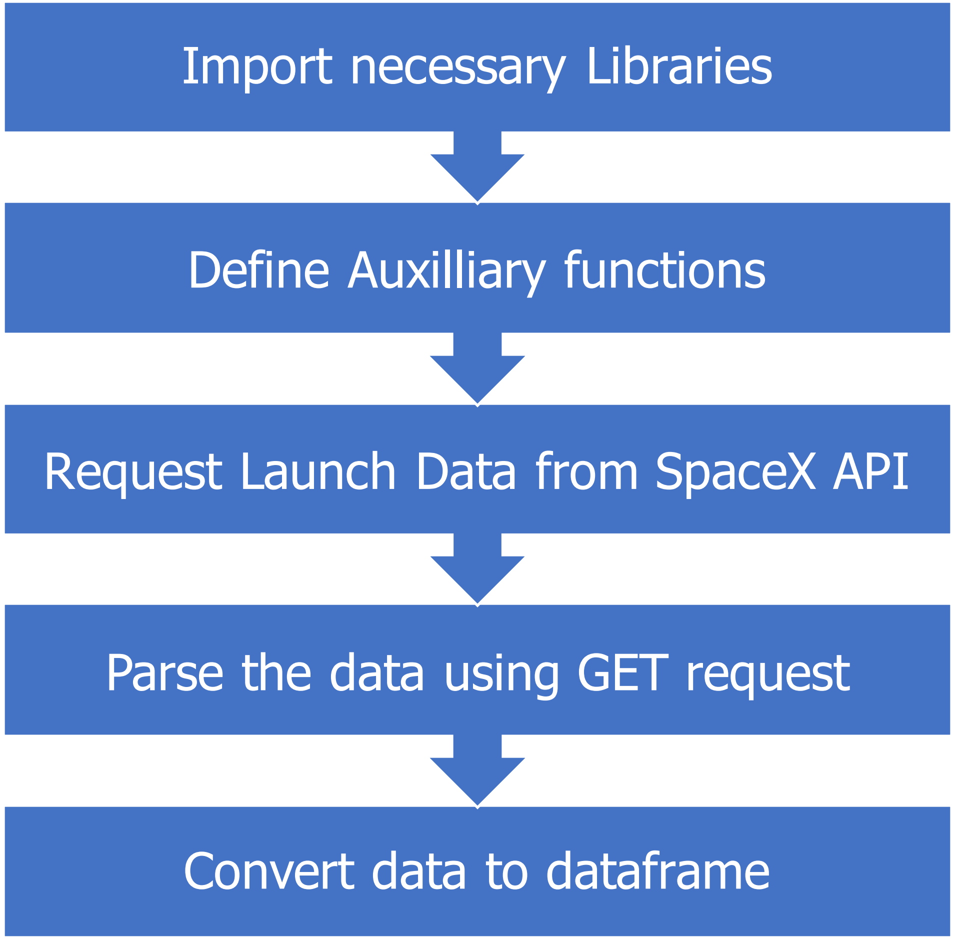
Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - Data was gathered via api from the SpaceX REST API and Web scraping Wikipedia pages for launch outcome results using Beautiful Soup
- Perform data wrangling
  - Missing Values were treated and a label column to easily identify launching outcomes was created
- 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 – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts
- Data was gathered from SpaceX REST API which provide information on Rocket used, Payload delivered, Launch specification, Landing specification and Landing outcome
- To get data from the api, we targeted a specific endpoint of the API to get past launch data by using a GET request to obtain the launch data
- The result returned as a json object which was then parsed and converted to a dataframe
- Data collection via SpaceX API notebook



Import necessary Libraries

Define Auxilliary functions

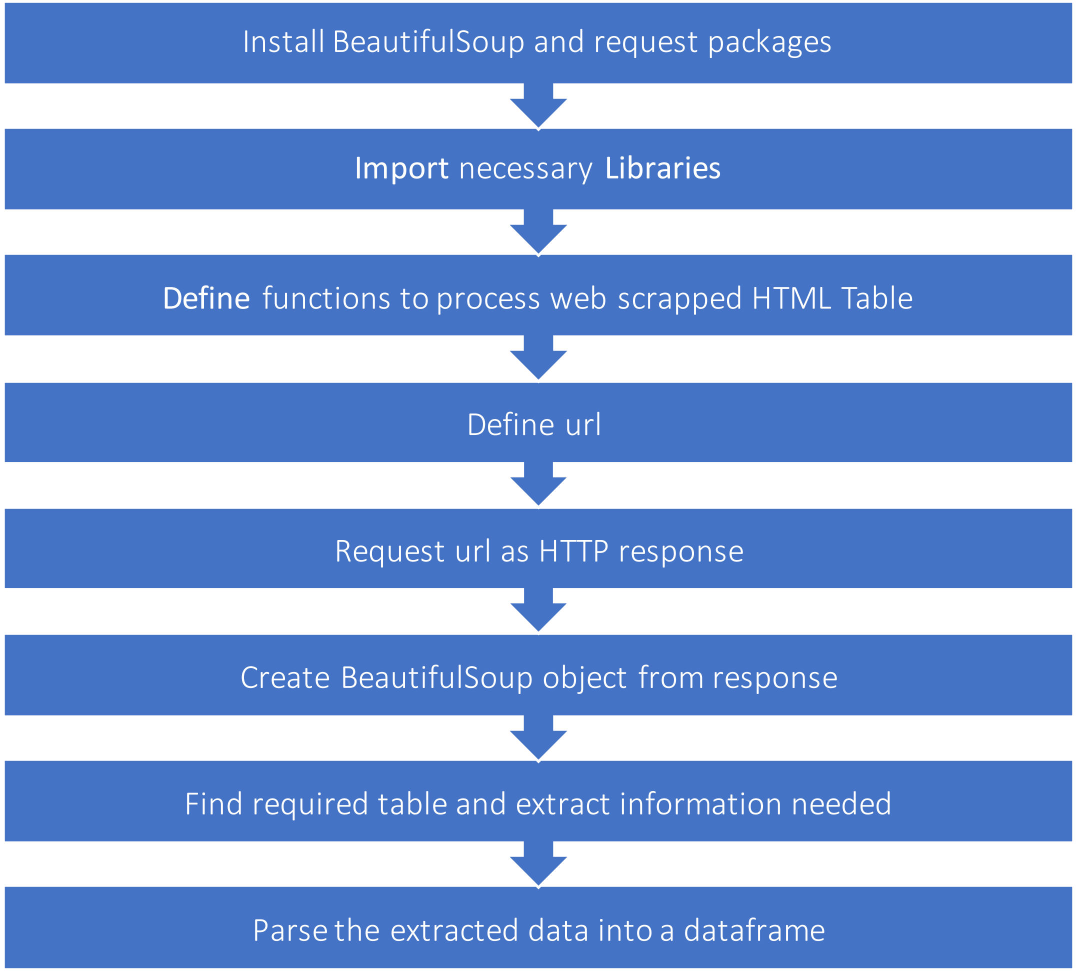
Request Launch Data from SpaceX API

Parse the data using GET request

Convert data to dataframe

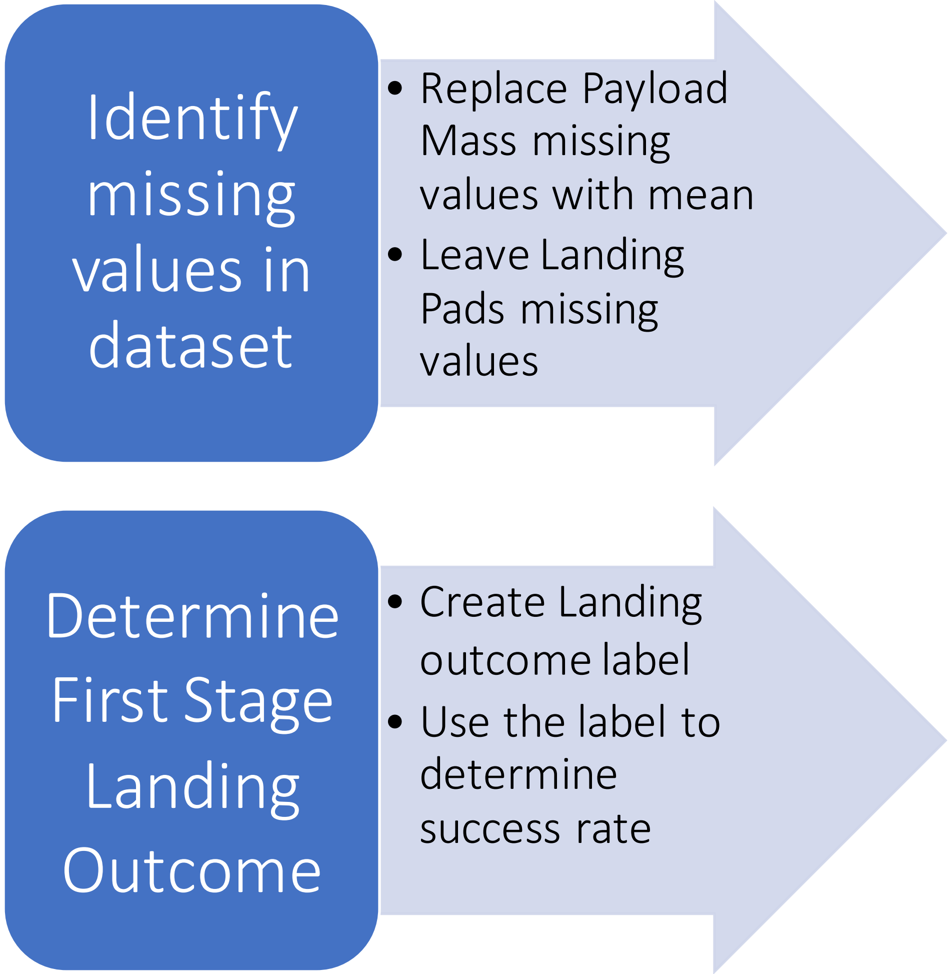
Data Collection - Scraping

- Addition data was gathered using web scrapping via Beautiful Soup package
- This data was collected from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches which provided Falcon 9 historical launch records
- The url where the data was stored was requested as HTTP response. A BeautifulSoup object was created on the response.
- Elements from the required table data were extracted and parsed into a dataframe
- Data collection via web scrapping


Install BeautifulSoup and request packages
Import necessary Libraries
Define functions to process web scrapped HTML Table
Define url
Request url as HTTP response
Create BeautifulSoup object from response
Find required table and extract information needed
Parse the extracted data into a dataframe

Data Wrangling

- Two key data wrangling tasks were done
- Identifying Missing Values
  - Missing values were identified on the Payload Mass and Landing Pads columns
  - For Payload Mass, the missing values were replaced with the mean while Landing Pad's were left to represent when landing pads were not used
- Determining Success rate of First stage
  - Firstly, a landing outcome label was created by identifying if an outcome was bad. 0 is used to represent failed first stage while 1 is used to represent successful first stage
  - Based on the labels above, the success rate for the first stage can be identified
- Data Wrangling (Missing Values)
- Data Wrangling (Landing Outcome)



Identify missing values in dataset

- Replace Payload Mass missing values with mean
- Leave Landing Pads missing values

Determine First Stage Landing Outcome

- Create Landing outcome label
- Use the label to determine success rate

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- To explore the data the following charts were plotted
    - Scatter Plot: This was used to understand how two variables affect Launch Outcomes such as
        - Flight Number and Launch Site
        - Payload and Launch Site
        - Flight Number and Orbit Type
        - PayLoad and Orbit Type
    - Bar Chart: This was used to show the relation between success rate and orbit type
    - Line Chart: This was used to show the yearly trend of average success rate
- EDA with Visualization

EDA with SQL

- Using SQL, a lot of information were derived some of the derived answers are shown below
- Unique names of launch sites in the space mission
- The total payload mass carried by boosters launched by NASA (CRS) was 45,596kg
- Average payload mass carried by booster version F9 v1.1 was 2928.4kg
- The first successful landing outcome in ground pad was achieved on 22nd Dec, 2015
- There were 101 successful mission outcomes and 1 failure mission outcome
- 12 booster_versions have carried the maximum payload mass
- The full SQL queries and results are in this notebook EDA with SQL

Build an Interactive Map with Folium

- Launch Site Location Analysis was done using Folium. For proper analysis, the following objects were used
- Map: used to create a map and center NASA Johnson Space Center at Hoston, Texas
- Circle: add a highlighted circle area with a text label on a specific coordinate
- Markers: mark the location of the coordinates
- Marker clusters: used to simplify a map containing many markers having the same coordinate
- Mouse Position: used to get the coordinate (Lat, Long) for a mouse hover on the map
- PolyLine: use a line to show the distance between a selected Launch Site and specific coastal, railway, highway and city
- The full analysis with Folium are in this notebook Interactive Map with Folium

Build a Dashboard with Plotly Dash

- A Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time
- This app has two interactive items for the user
  - a *dropdown* for the user to select specific Launch site or all Launch sites
  - a *PayLoad slider* for the user to choose a range for the PayLoad Mass
- The map has two visuals
  - A *pie chart* showing the total success launches for the various Launch sites and the success & failed count for selected sites
  - A *Scatter Plot* to observe how payload may be correlated with mission outcomes for selected site(s) color coded by Booster Version
- The completed plotly Dash Lab can be found in this notebook Plotly Dash Lab

Predictive Analysis (Classification)

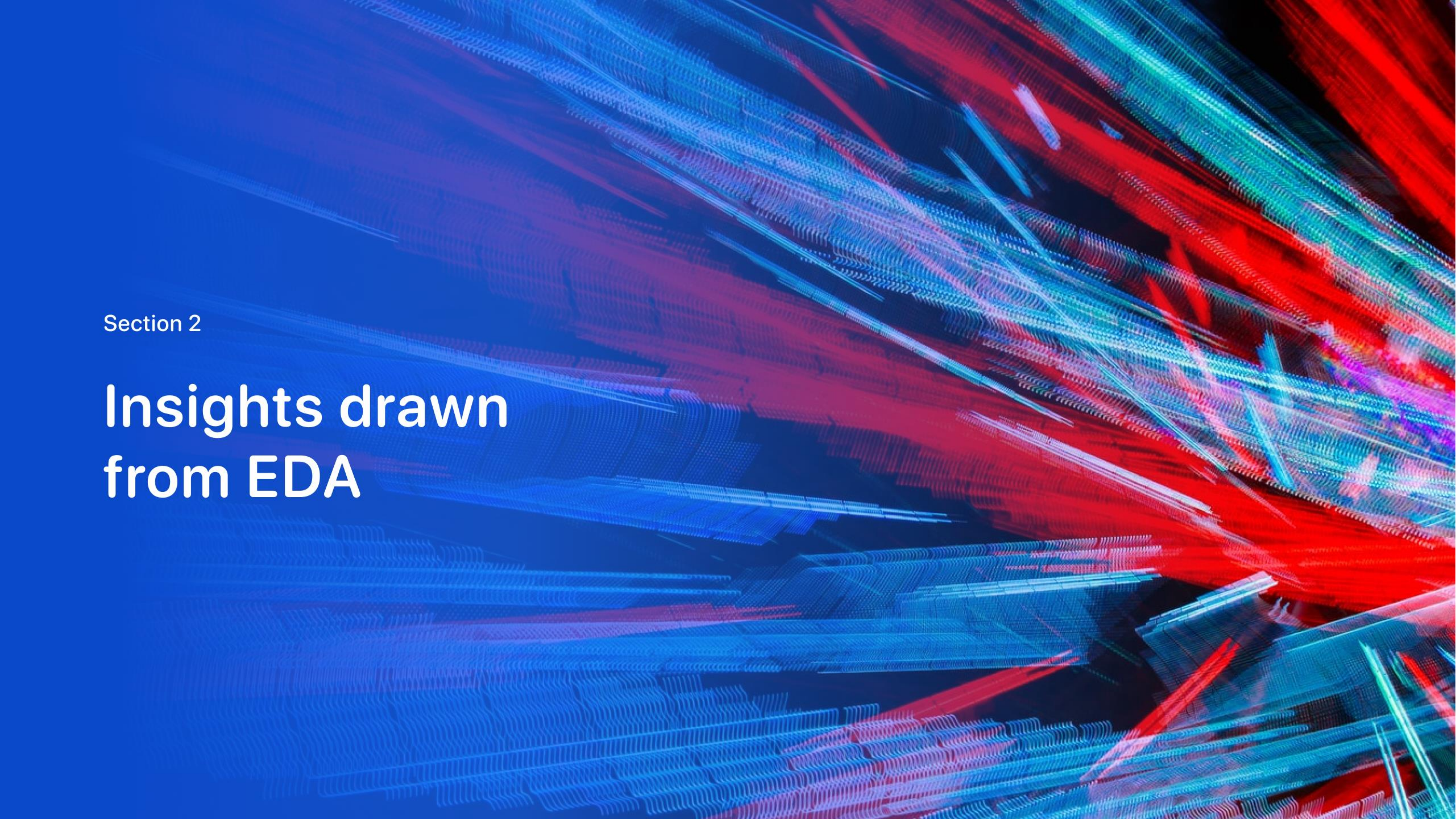
- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
- Four classification models were used to determine the best model to predict landing outcomes. The models are Logistic Regression, SVM, Decision Tree and KNN
- By standardizing the data, splitting the data into train & test datasets and fitting the model using GridSearch CV to find the best parameters for the model, the following was discovered
- Using the training dataset, Decision Tree performed best with an accuracy score of 87.5%
- Using the test data however, all four models had the same accuracy score of 83.33%
- The predictive analysis can be found in this notebook Predictive Analysis

Results

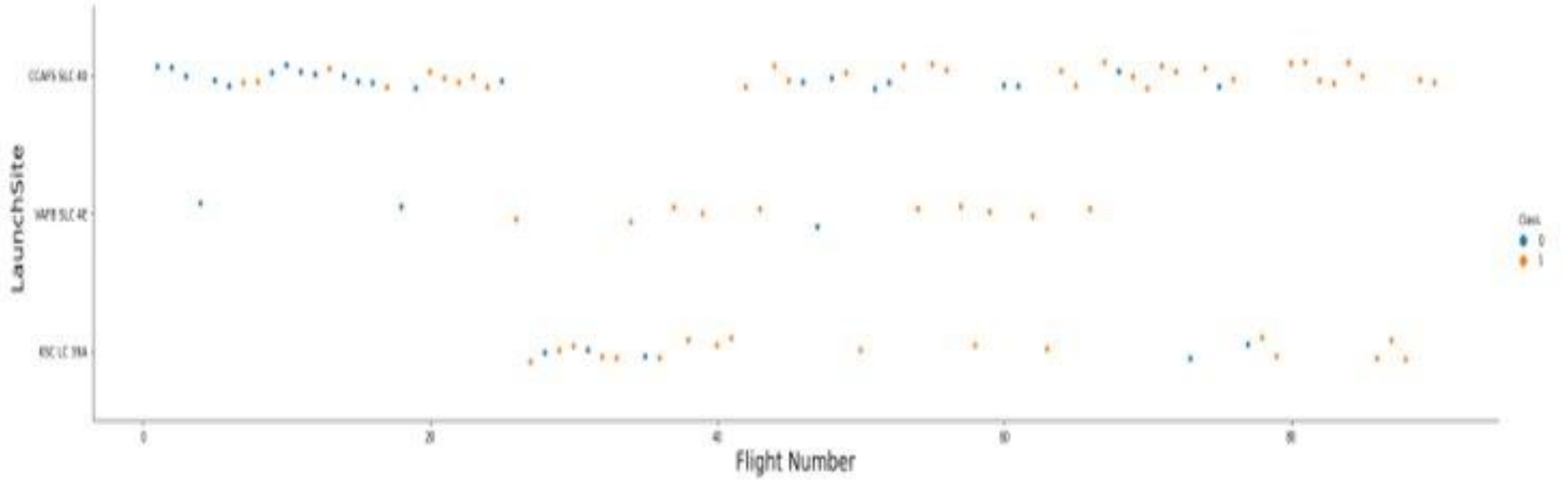
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

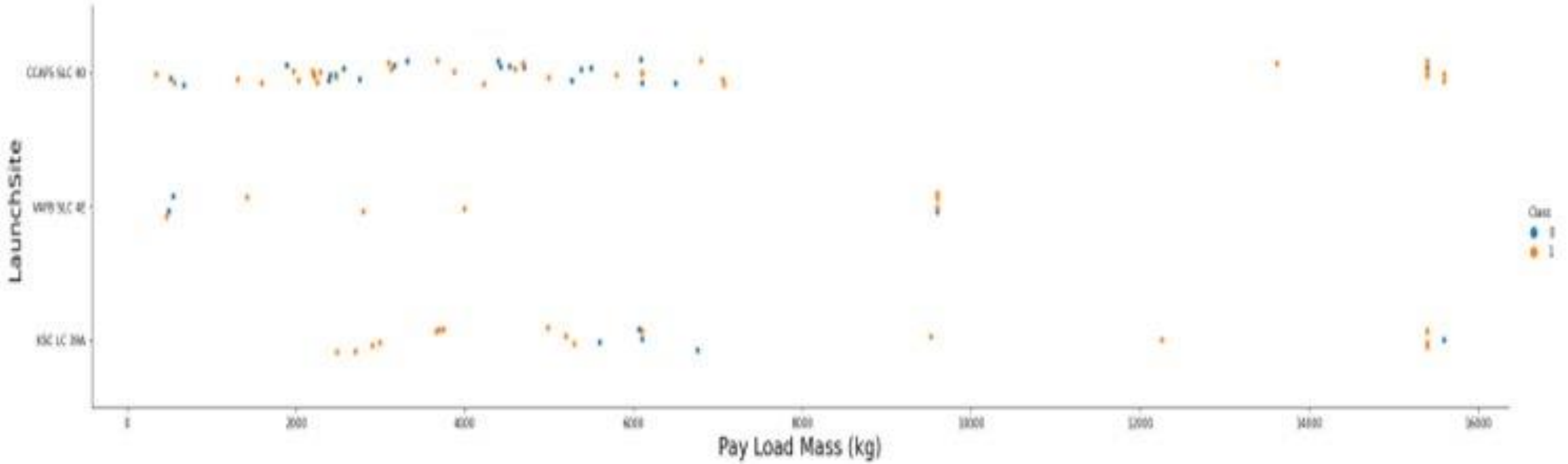


Flight Number vs. Launch Site



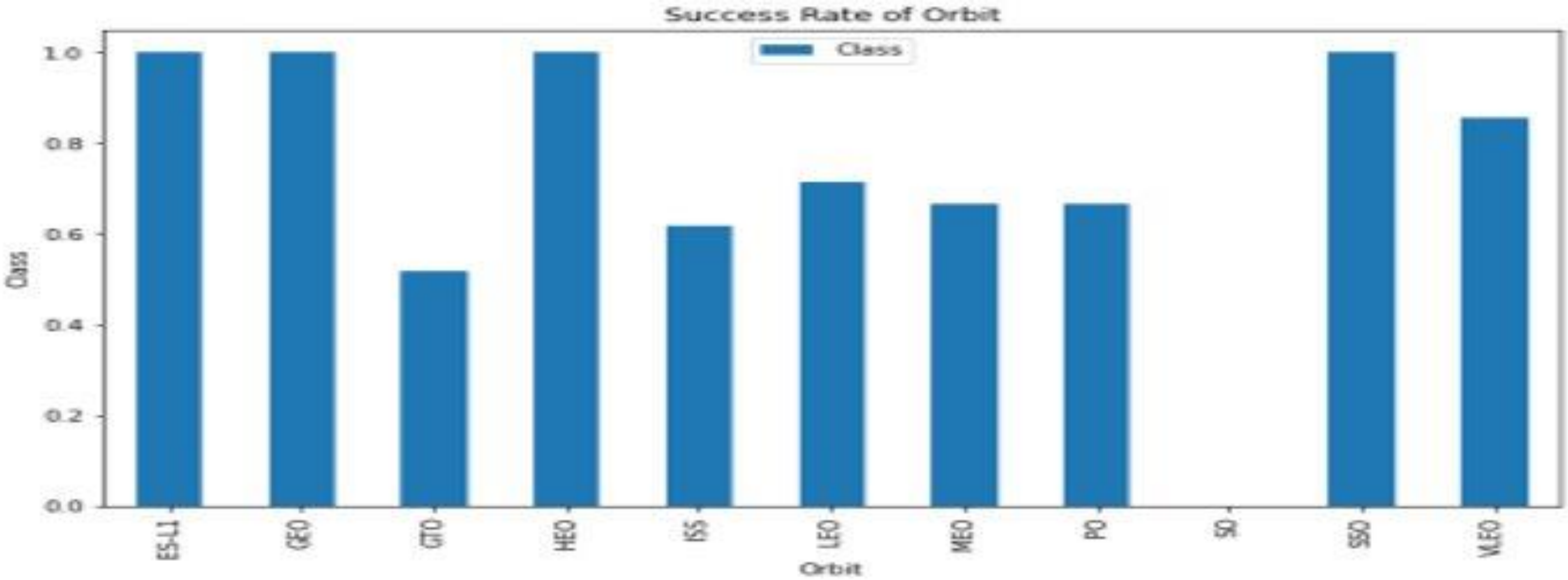
- as the flight number increases, the first stage is more likely to land successfully for each launch site
- VAFB SLC 4E has lower launches compared to others but also a higher successful launch rate

Payload vs. Launch Site



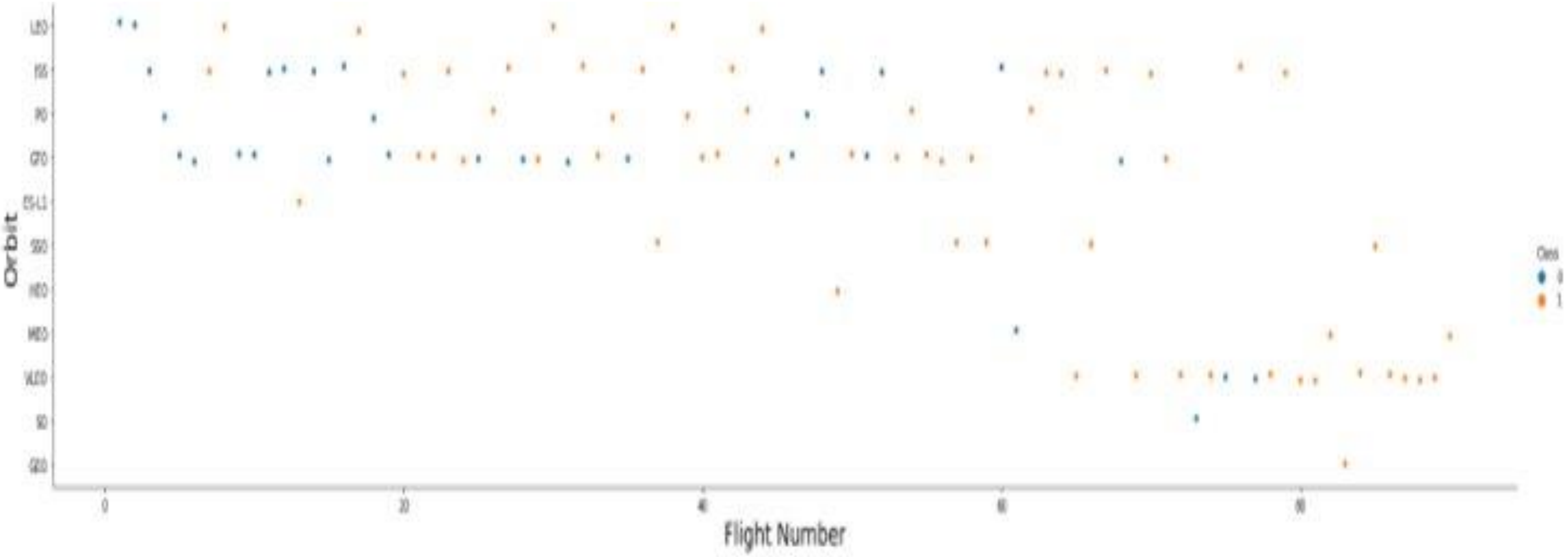
- as the payload mass increases, the first stage is more likely to land successfully for each launch site
- for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000)

Success Rate vs. Orbit Type



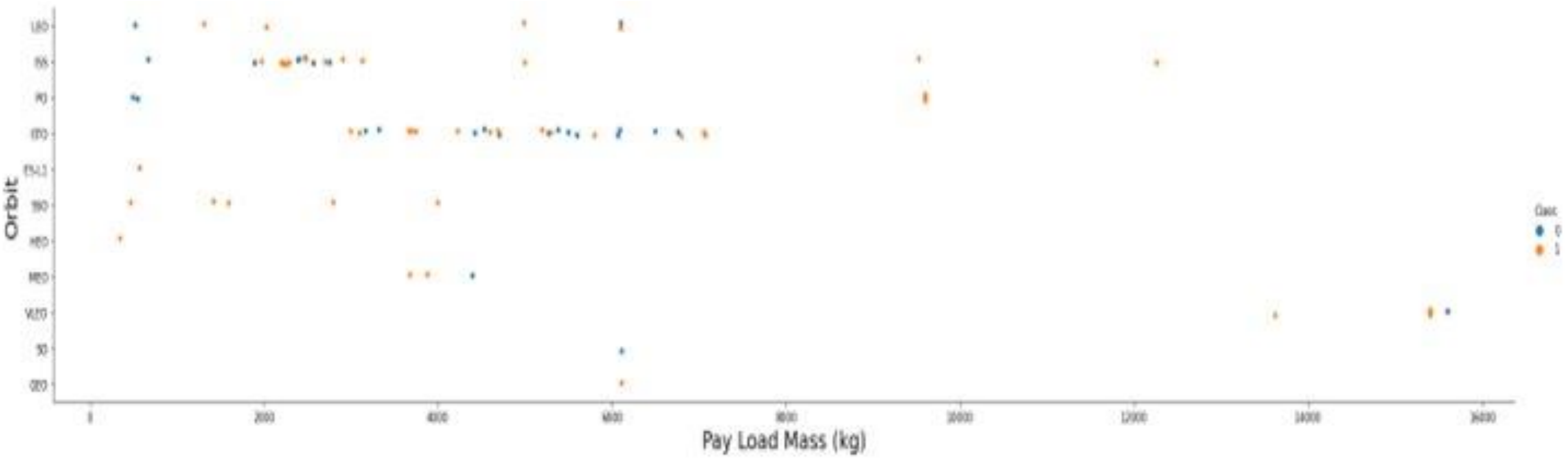
- ESL1, GEO, HEO, SEO all has 100% Success rate
- SO have had no successful launch

Flight Number vs. Orbit Type



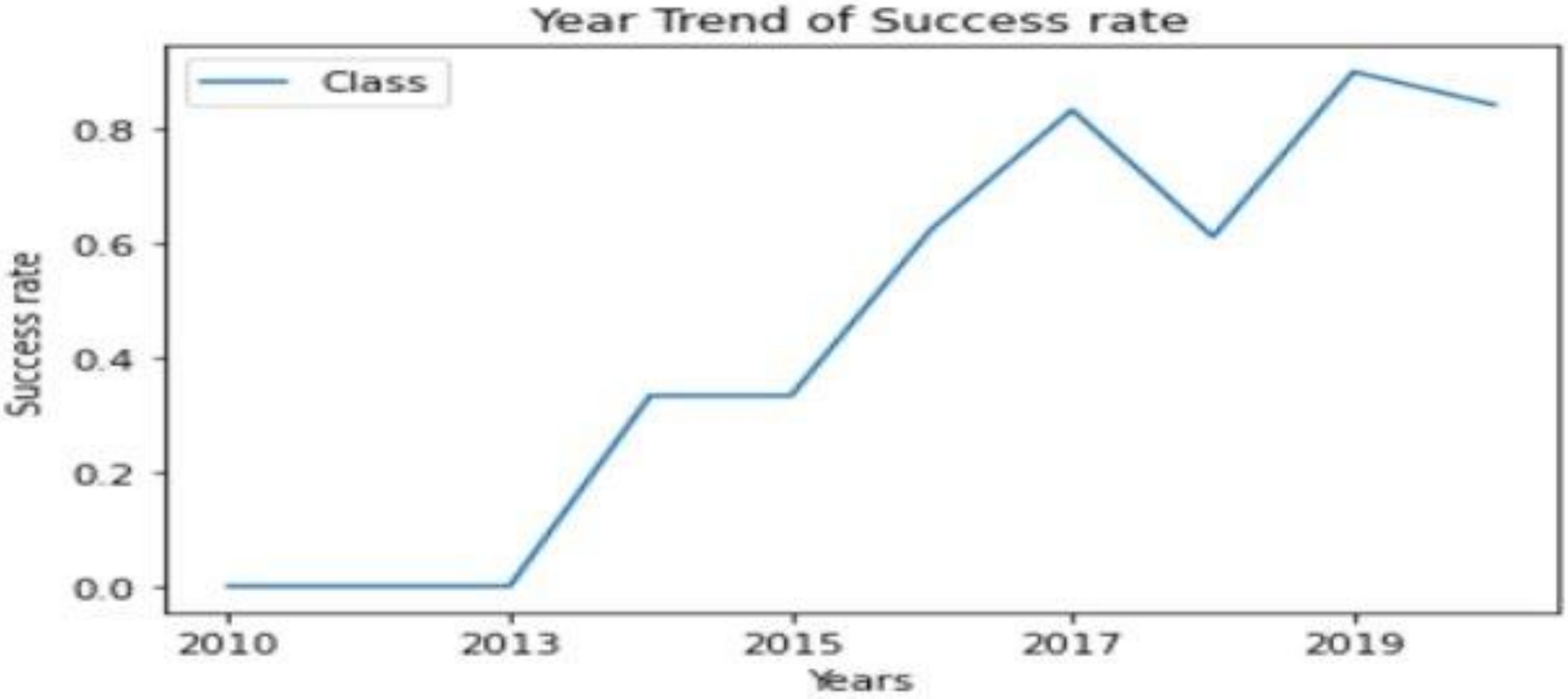
- LEO and VLEO orbit Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

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.

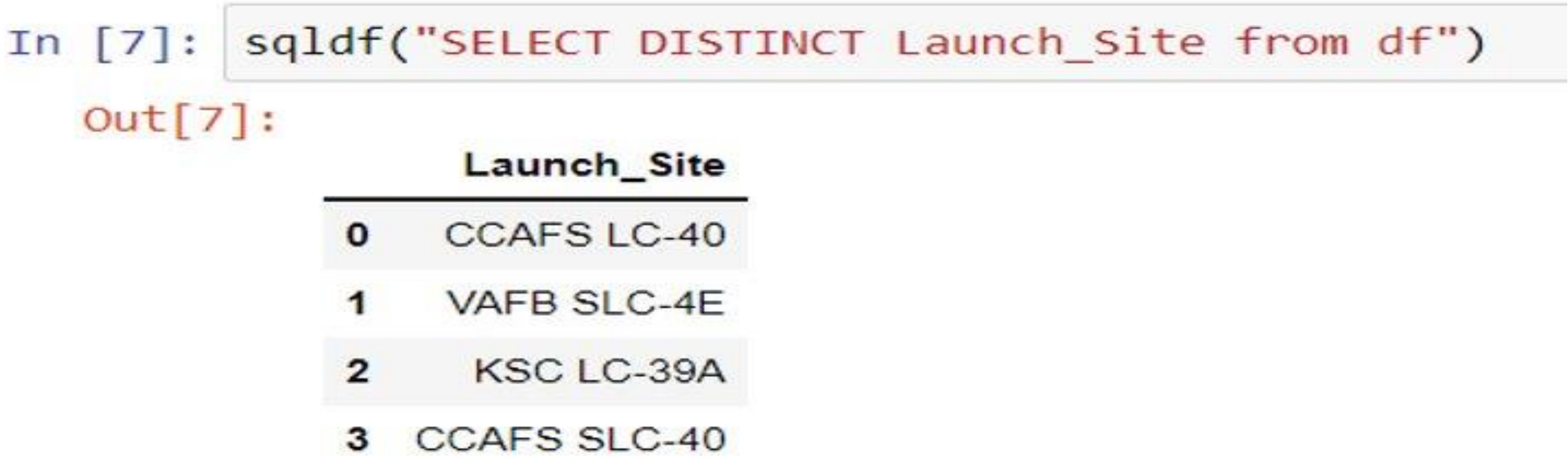
Launch Success Yearly Trend



- the sucess rate since 2013 kept increasing till 2020. There a was decline in 2018.

All Launch Site Names

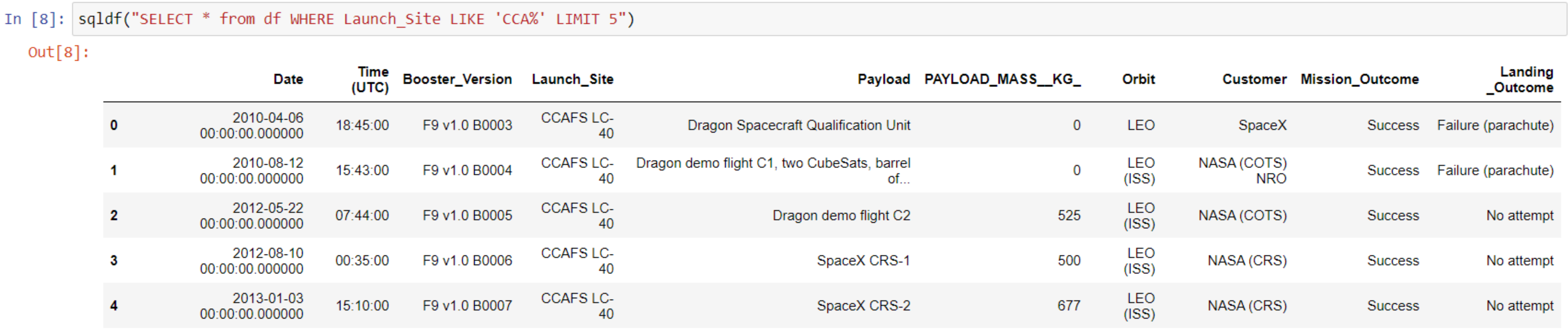
- There are four unique launch sites



```
In [7]: sqldf("SELECT DISTINCT Launch_Site from df")
   Out[7]:
                 Launch_Site
            0   CCAFS LC-40
            1   VAFB SLC-4E
            2    KSC LC-39A
            3  CCAFS SLC-40
```

Launch Site Names Begin with 'CCA'

- launch sites beginning with `CCA` are the most commonly used launch sites. We can see a sample of 5 record of launches from these sites



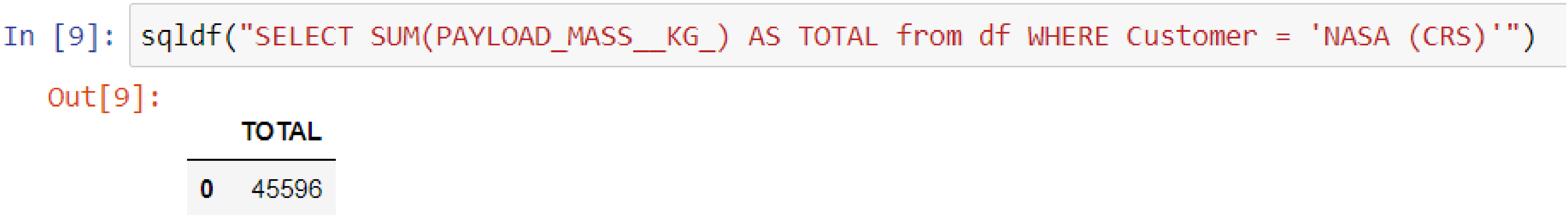
```
In [8]: sqldf("SELECT * from df WHERE Launch_Site LIKE 'CCA%' LIMIT 5")
```

Out[8]:

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
0	2010-04-06 00:00:00.000000	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12 00:00:00.000000	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	2012-05-22 00:00:00.000000	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	2012-08-10 00:00:00.000000	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03 00:00:00.000000	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload carried by boosters from NASA was 45,596 kg

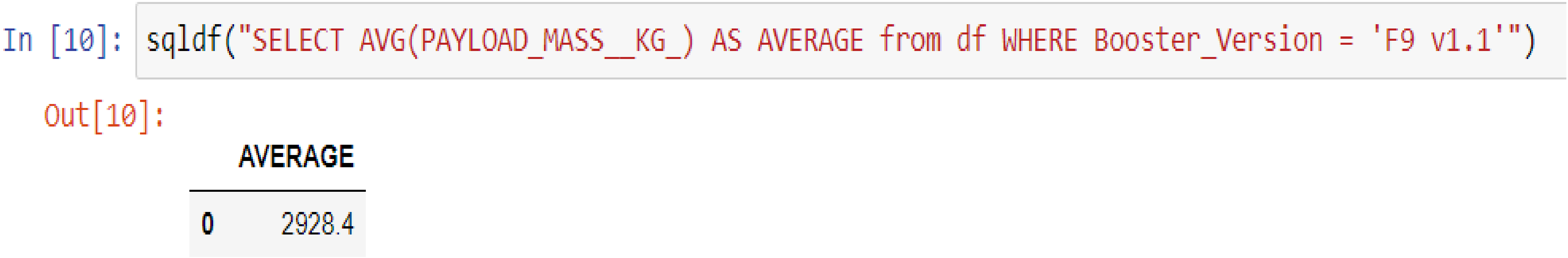


```
In [9]: sqldf("SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL from df WHERE Customer = 'NASA (CRS)'")

Out[9]:
           TOTAL
        0  45596
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2,928.4 kg

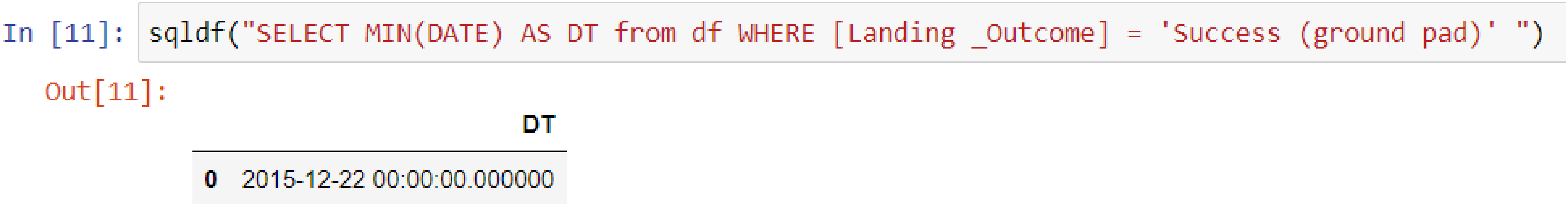


```
In [10]: sqldf("SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE from df WHERE Booster_Version = 'F9 v1.1'")
```

```
Out[10]:
     AVERAGE
0     2928.4
```

First Successful Ground Landing Date

- The first successful landing outcome in ground pad was achieved on 22nd Dec, 2015

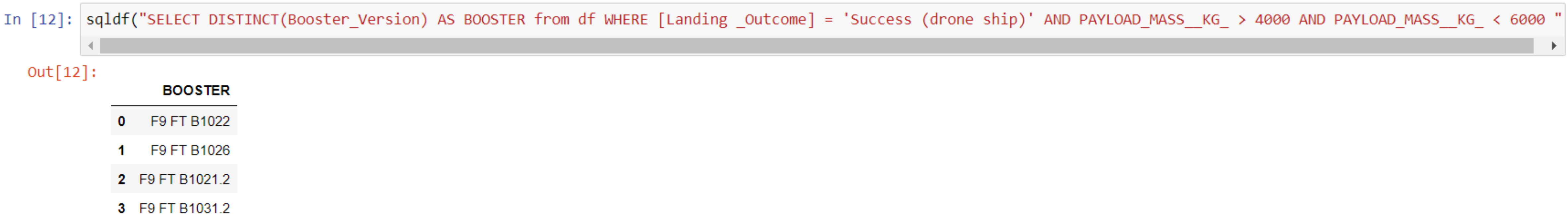


```
In [11]: sqldf("SELECT MIN(DATE) AS DT from df WHERE [Landing _Outcome] = 'Success (ground pad)' ")
```

```
Out[11]:
                          DT
0   2015-12-22 00:00:00.000000
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2



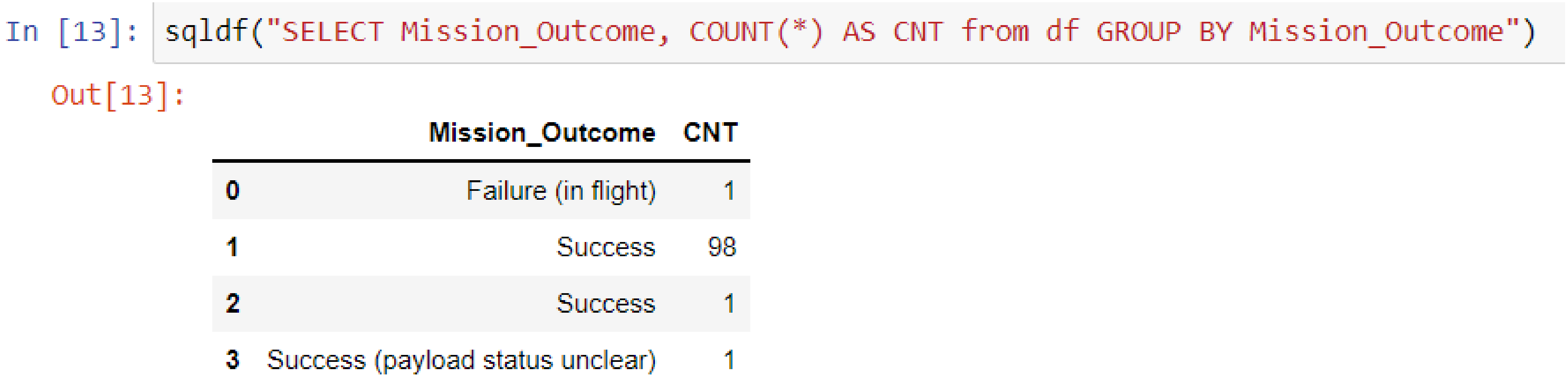
```
In [12]: sqldf("SELECT DISTINCT(Booster_Version) AS BOOSTER from df WHERE [Landing _Outcome] = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000 "
```

Out[12]:

	BOOSTER
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- There were 101 successful mission outcomes and 1 failure mission outcome



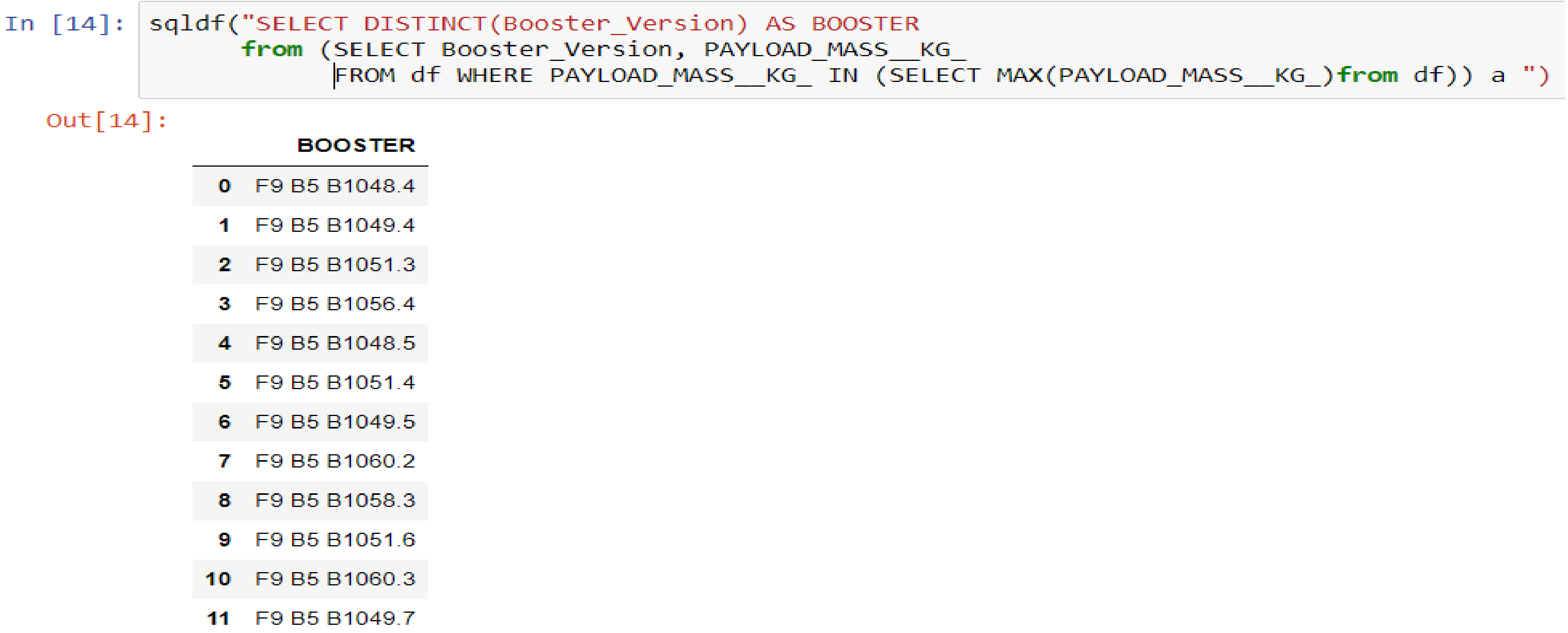
```
In [13]: sqldf("SELECT Mission_Outcome, COUNT(*) AS CNT from df GROUP BY Mission_Outcome")
```

Out[13]:

	Mission_Outcome	CNT
0	Failure (in flight)	1
1	Success	98
2	Success	1
3	Success (payload status unclear)	1

Boosters Carried Maximum Payload

- 12 booster_versions have carried the maximum payload mass. These are shown in the screenshot below

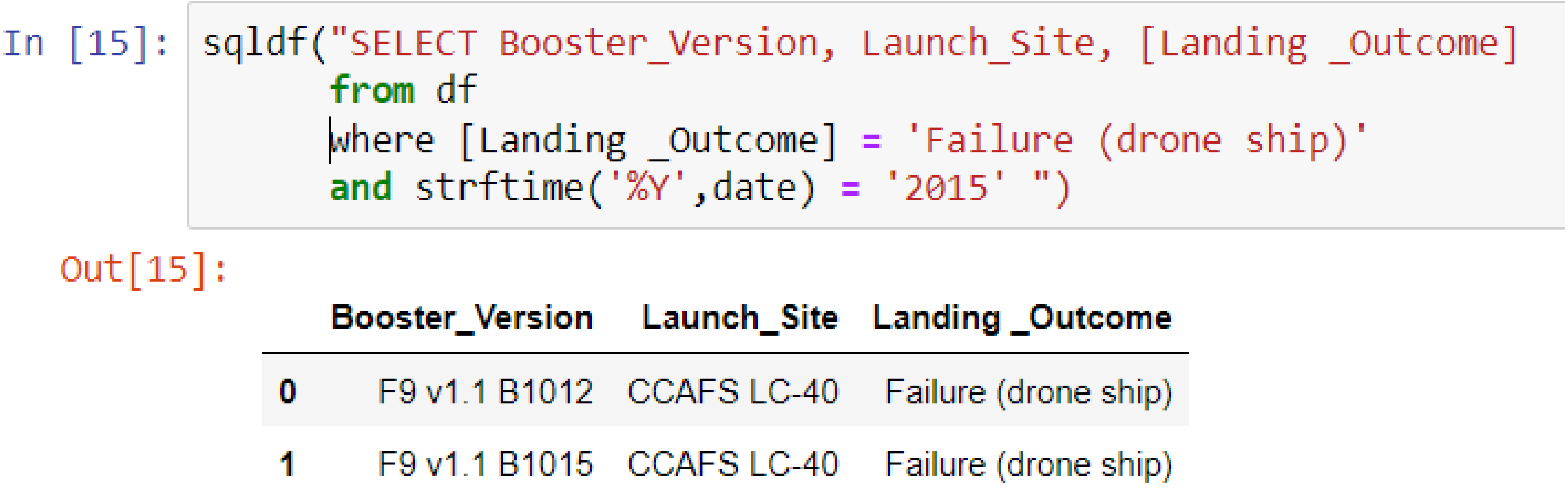


```
In [14]: sqldf("SELECT DISTINCT(Booster_Version) AS BOOSTER
               from (SELECT Booster_Version, PAYLOAD_MASS__KG_
                     FROM df WHERE PAYLOAD_MASS__KG_ IN (SELECT MAX(PAYLOAD_MASS__KG_)from df)) a ")
```

```
Out[14]:
               BOOSTER
     0    F9 B5 B1048.4
     1    F9 B5 B1049.4
     2    F9 B5 B1051.3
     3    F9 B5 B1056.4
     4    F9 B5 B1048.5
     5    F9 B5 B1051.4
     6    F9 B5 B1049.5
     7    F9 B5 B1060.2
     8    F9 B5 B1058.3
     9    F9 B5 B1051.6
    10    F9 B5 B1060.3
    11    F9 B5 B1049.7
```

2015 Launch Records

- There were two failed landing_outcomes in drone ship in 2015.
- Their booster versions, and launch site names are shown in the screenshot below



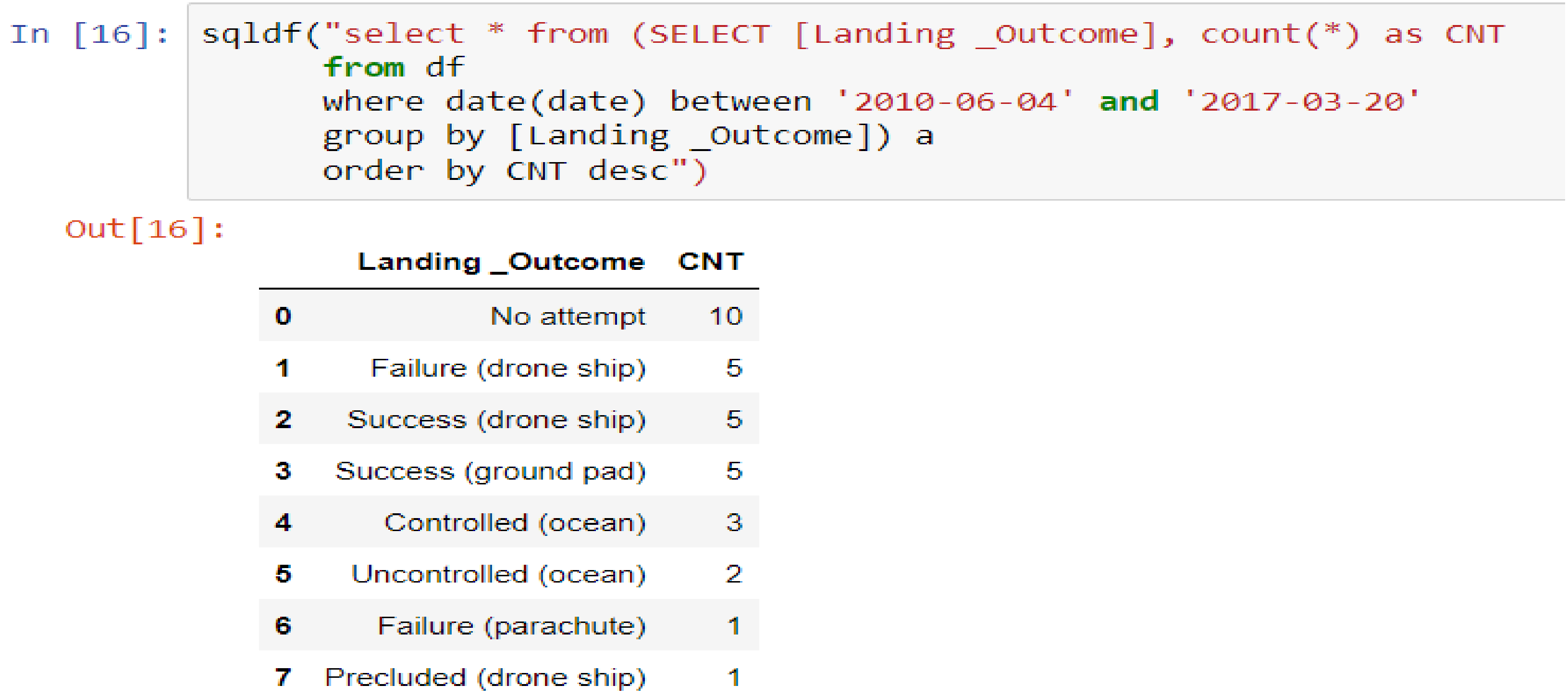
```
In [15]: sqldf("SELECT Booster_Version, Launch_Site, [Landing _Outcome]
               from df
               where [Landing _Outcome] = 'Failure (drone ship)'
               and strftime('%Y',date) = '2015' ")
```

Out[15]:

	Booster_Version	Launch_Site	Landing _Outcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Below is the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20
- No attempt has the most count of 10 while Failure by parachute and Precluded by droneship have the lowest count of 14 each



```
In [16]: sqldf("select * from (SELECT [Landing _Outcome], count(*) as CNT
                from df
                where date(date) between '2010-06-04' and '2017-03-20'
                group by [Landing _Outcome]) a
                order by CNT desc")
```

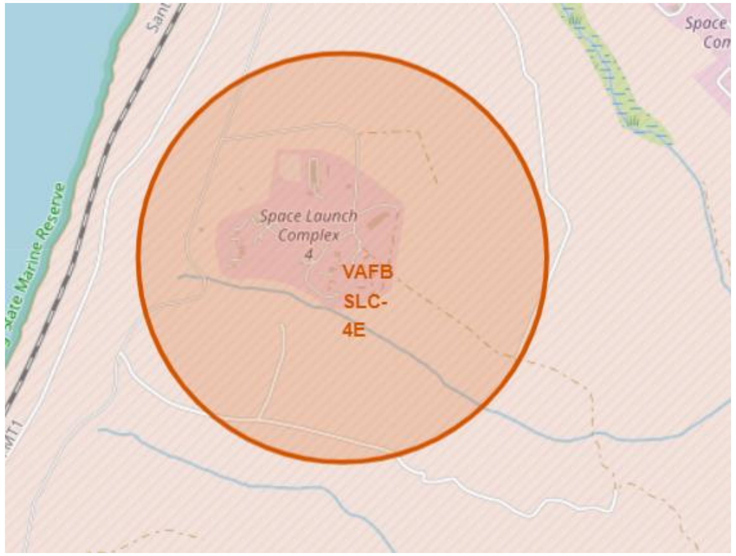
Out[16]:

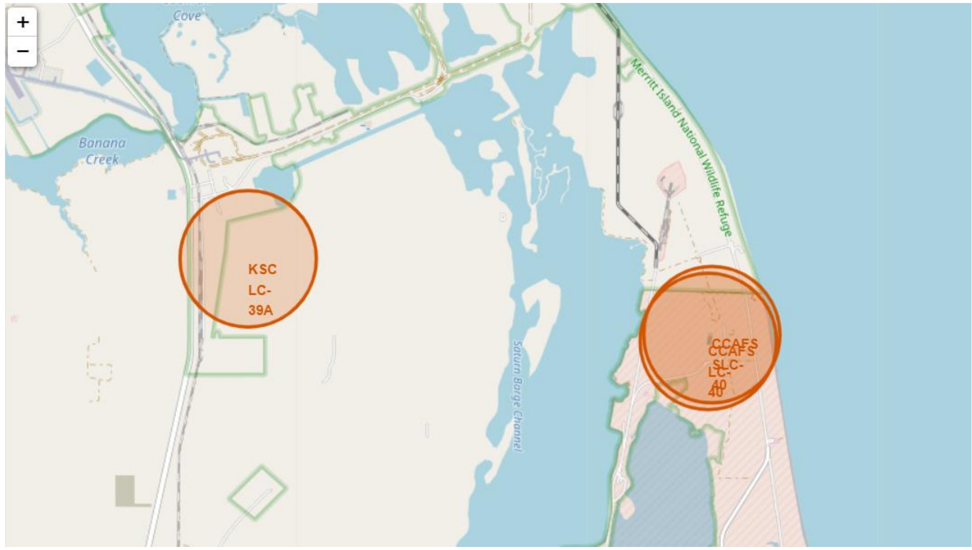
	Landing _Outcome	CNT
0	No attempt	10
1	Failure (drone ship)	5
2	Success (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Failure (parachute)	1
7	Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

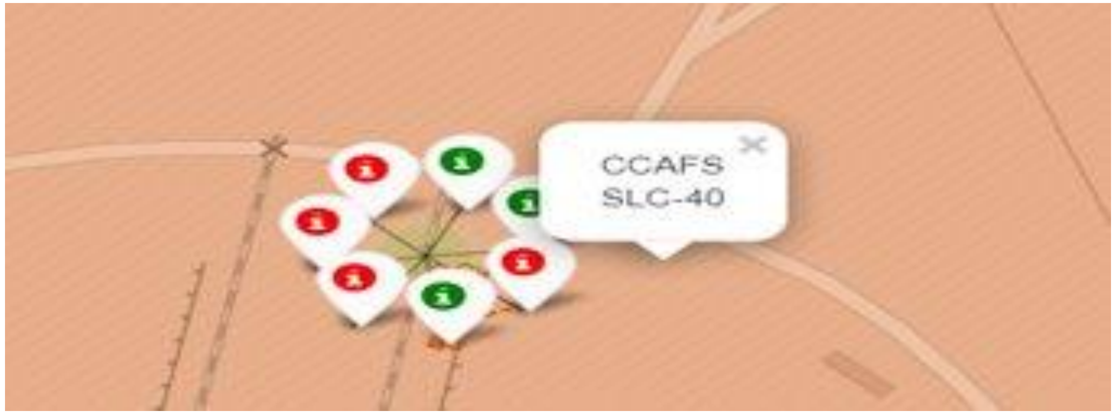
Interactive Map with Folium (Launch Site Locations)

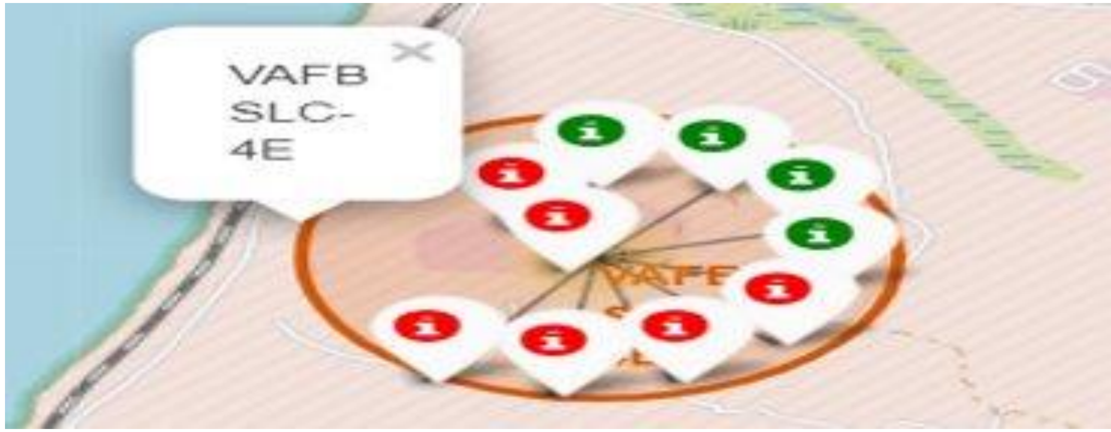


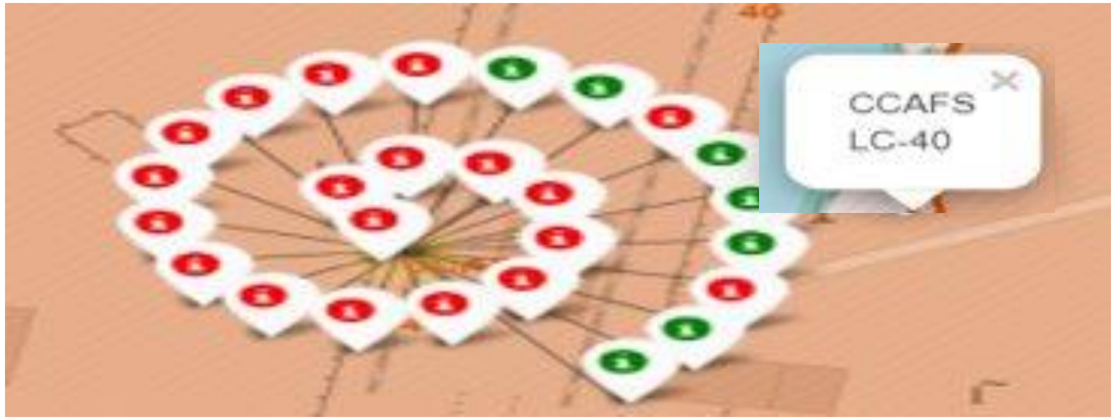


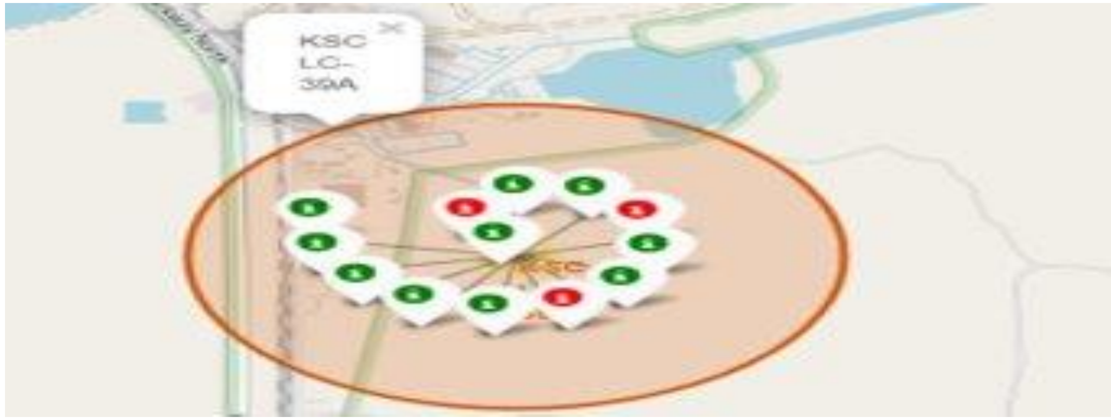
- VAFB SLC-4E is further from the other launch sites

Interactive Map with Folium (Color Labeled Outcomes)



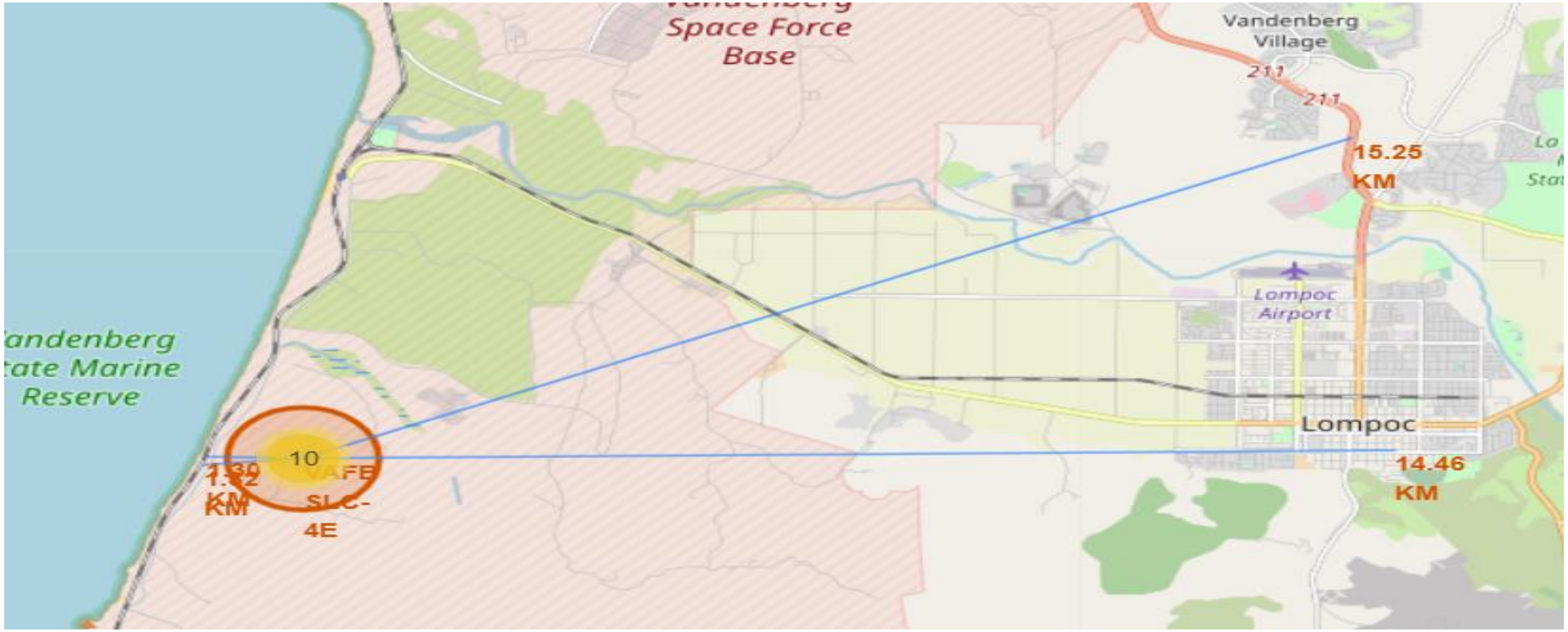




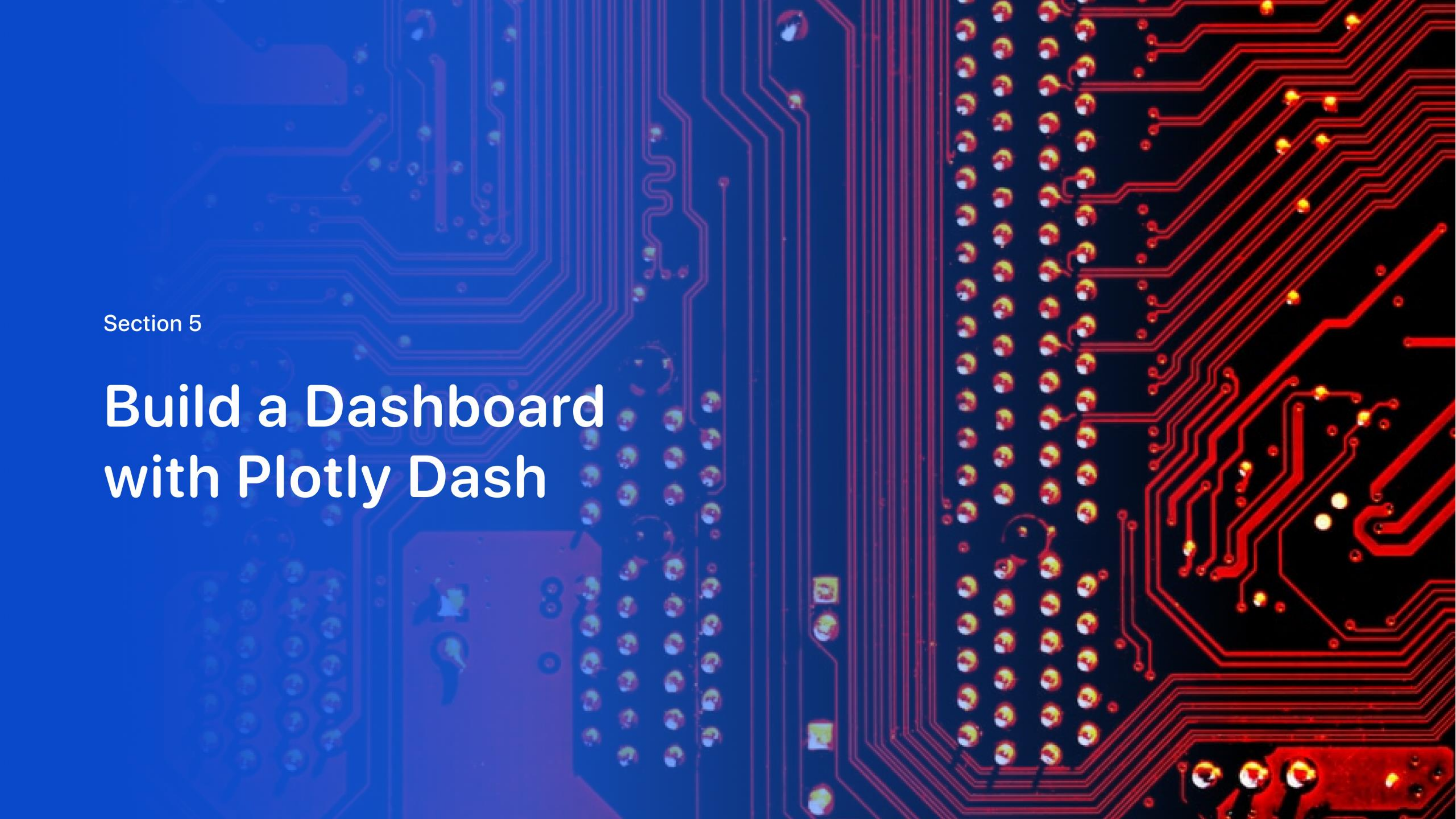


- KSC LC-39A has had the most successful launches. CCAFS LC-40 , VAFB SLC-4E and CCAFS SLC-40 have more failed launch than successful launch

Interactive Map with Folium (Distance)



- VAFB SLC-4E is much closer to coastal line (distance of 1.30km) followed by highway (distance of 1.32km) then Lompoc city (distance of 14.46km) and finally railway (distance of 15.25km)

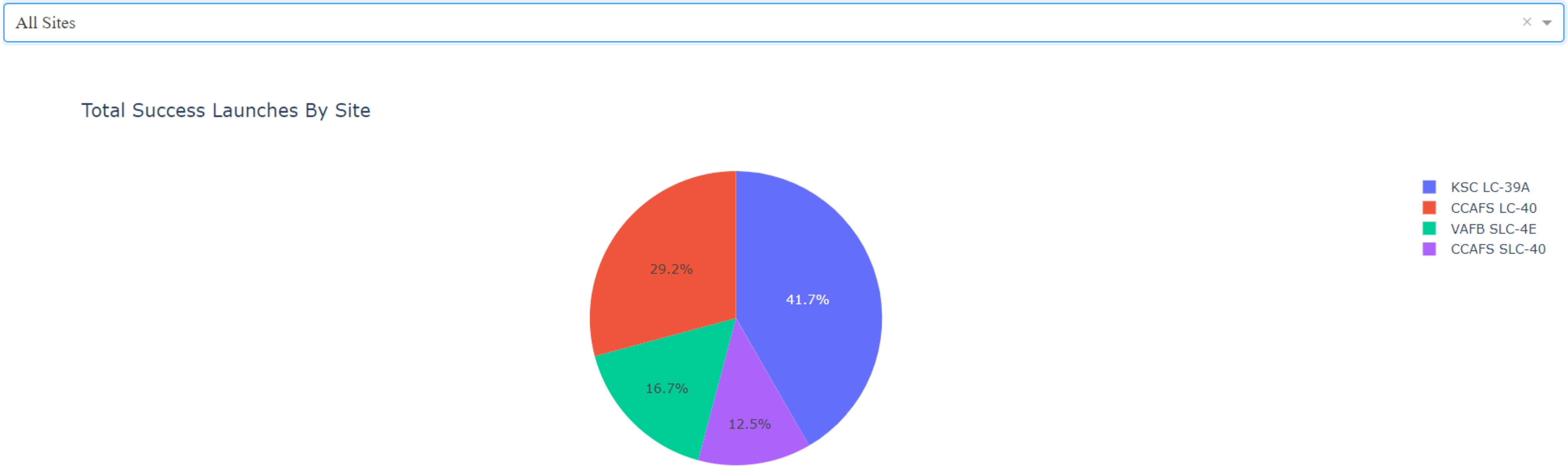


Section 5

Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard (All sites Pie chart)

- KSC LC-39A has the most successful launches (41.7% of all launches) while CCAFS SLC-40 has the least with 12.5%

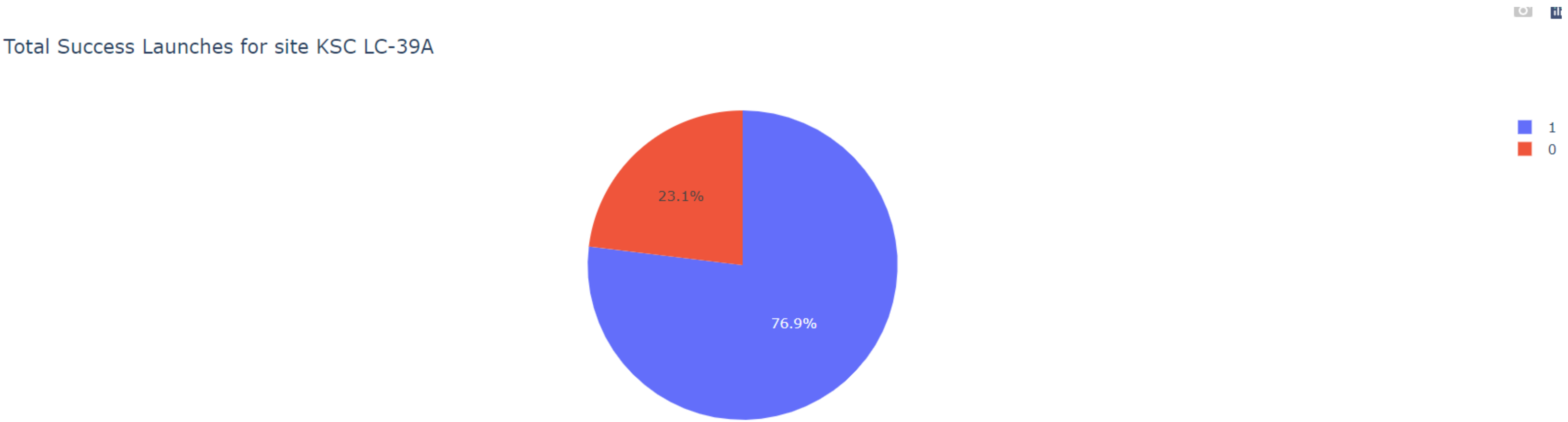


All Sites × ▾

Total Success Launches By Site

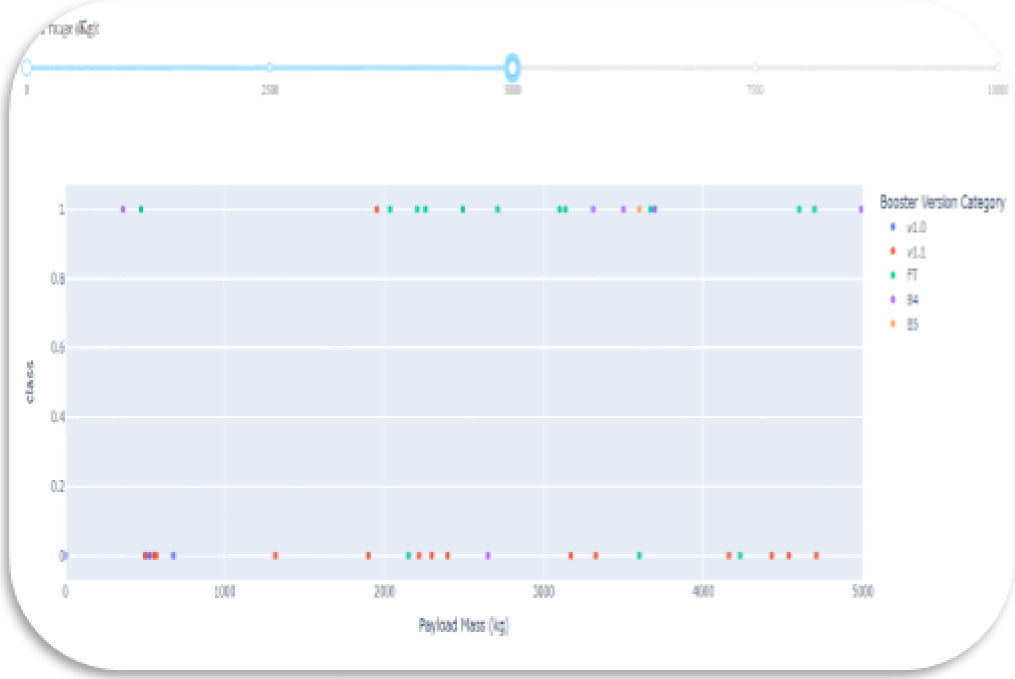
SpaceX Launch Records Dashboard (KSC LC-39A Pie chart)

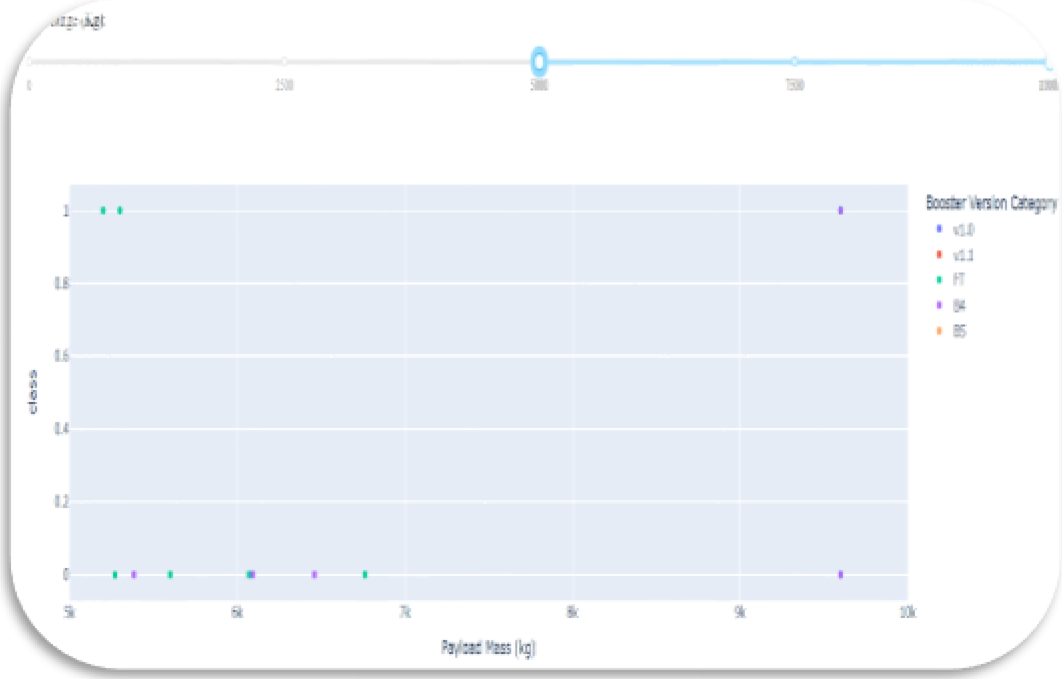
- KSC LC-39A has the highest launch success ratio of 76.9% (I.e. 10 of their 13 launches were successful)



SpaceX Launch Records Dashboard (All sites Scatter Plot)

- Within payload range of 0 and 5,000kg, overall there were a little bit more failed launches than successful launches with Booster version FT having the most successful launch rate
- Within payload range of 5,000 and 10,000kg, overall there were a lot more failed launches than successful launches. Booster Version FT and B4 were only ones launching at this range
- Overall there seems to be no correlation between Payload Mass and Success Rate



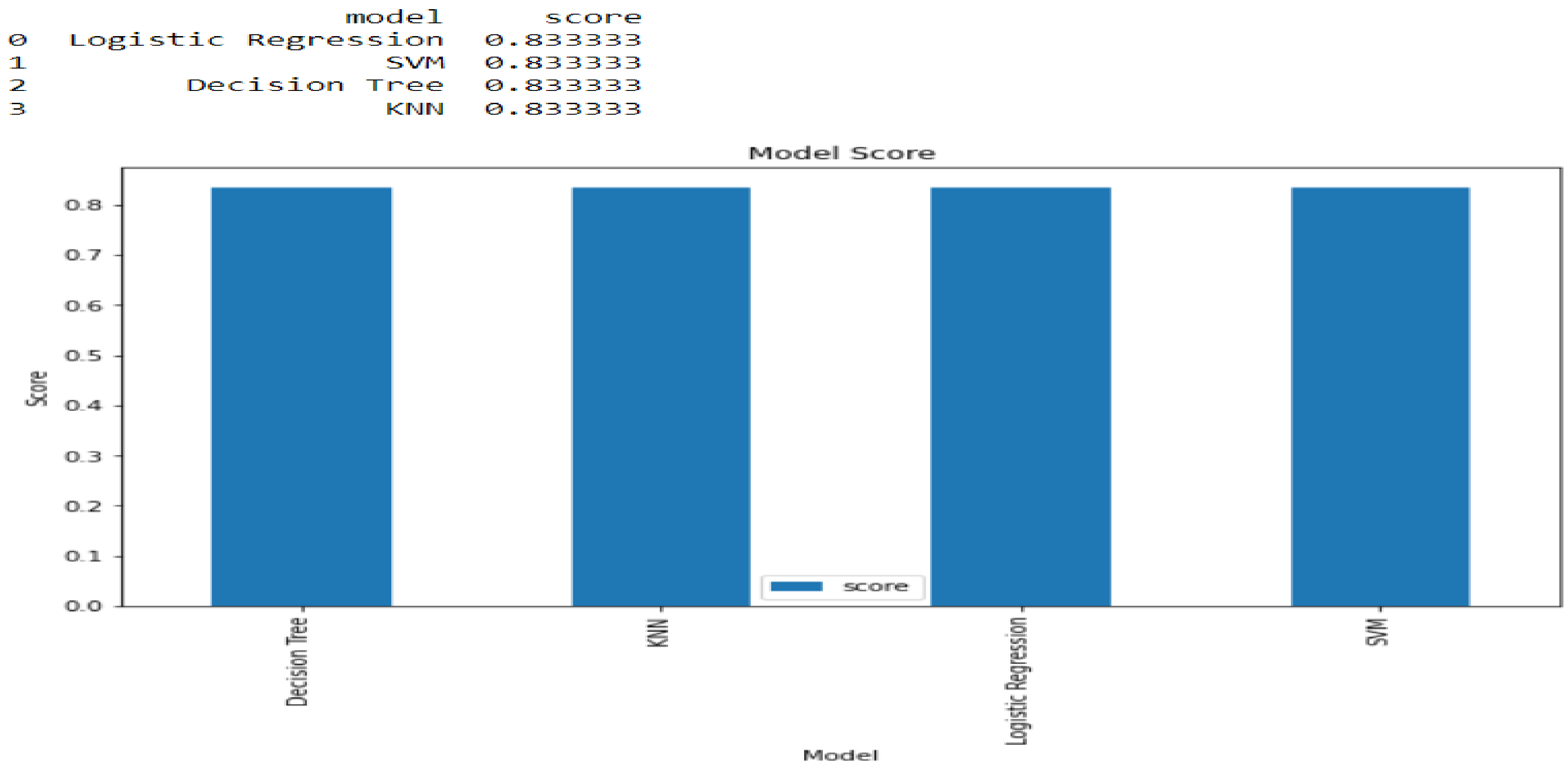


Section 6

Predictive Analysis (Classification)

Classification Accuracy

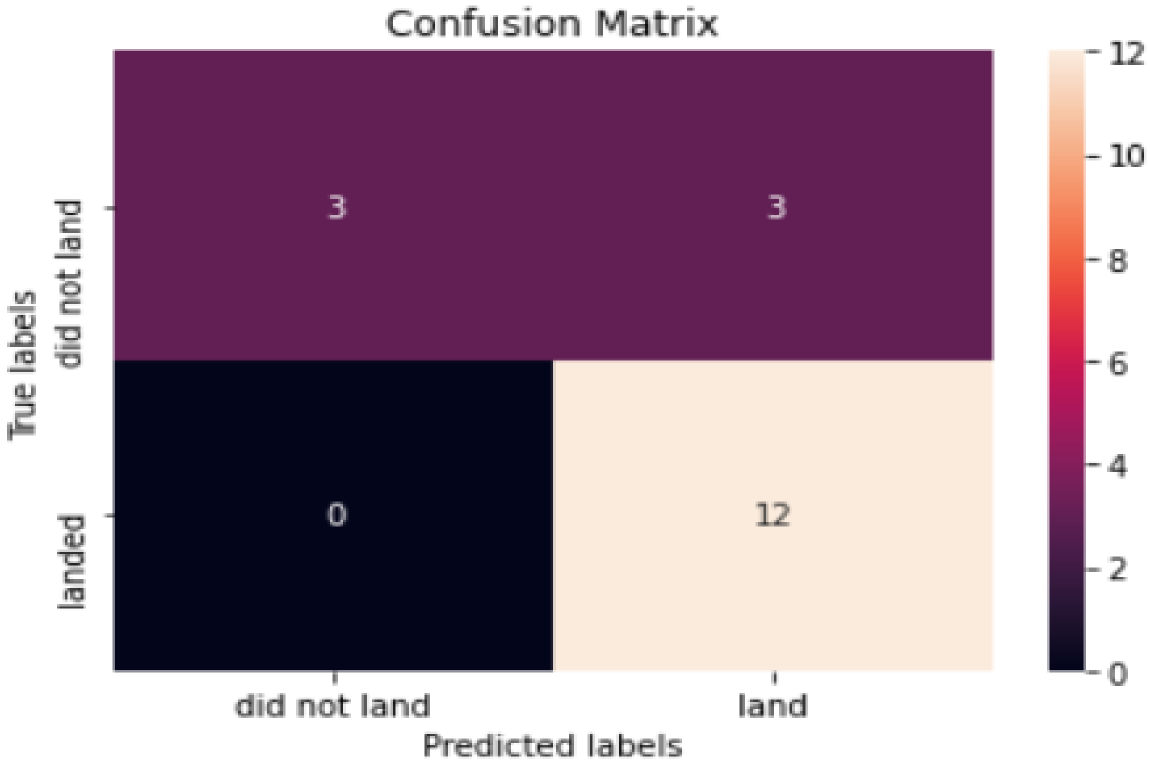
- All four models have the same score accuracy



```
                 model     score
0  Logistic Regression  0.833333
1                  SVM  0.833333
2        Decision Tree  0.833333
3                  KNN  0.833333
```

Confusion Matrix

- The models can to a large extent correctly predict successful and failed landing but at times can incorrectly classify unsuccessful landing as successful



Conclusions

Based on the above analysis, the following conclusions can be made

- Flight Number and Lauch sites significantly affect Launch outcomes while Payload surprisingly does not seem to affect the launch outcomes
- Since 2013, there have been a positive trend of successful launch rates
- KSC LC-39A is the best site to carry out a launch to have a better chance for a successful launch outcome and rate
- The Launch Sites a closer to coastal areas for safe launches
- Our model can to a large extent correctly predict successful and failed landing but at times can incorrectly classify unsuccessful landing as successful

Appendix

Notebook links

- Data collection via SpaceX API notebook
- Data collection via web scrapping
- Data Wrangling (Missing Values)
- Data Wrangling (Landing Outcome)
- EDA with Visualization
- EDA with SQL
- Interactive Map with Folium
- Plotly Dash Lab
- Predictive Analysis

Thank you!