IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Visali Devi Nerla
14 April 2025

Outline

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

Executive Summary

Methodologies Used:

- Collecting the Data
- Data Wrangling
- Exploratory Data Analysis (EDA)
  - Using SQL
  - Using Pandas and Matplotlib
- Interactive Visual Analysis
  - Using Folium
  - Using Plotly Dash
- Predictive Analysis

Results Provided:

- Exploratory Data Analysis
- Geospatial Analysis
- Interactive Dashboard
- Predictive Analysis

Introduction

- Spaces X's Falcon 9 launch regular rockets. This launch has two stages.
    - The first stage is responsible for providing the initial thrust to lift the rocket off the ground and propel it into space.
    - Once the first stage has burned through its fuel, it separates from the second stage and begins its descent back to Earth for landing.
- Space Y is a new player in commercial rocket launch market and is trying to mimic Space X's formula
- Space Y wants to use machine learning techniques to predict success rate of Falcon 9 in different stages and leverage the findings to make informed decision for their own projects
- Data Scientists at Space Y wants to develop a model to accurately predict the success rate of first stage landing of Space X's Flacon 9

Section 1

Methodology

Methodology

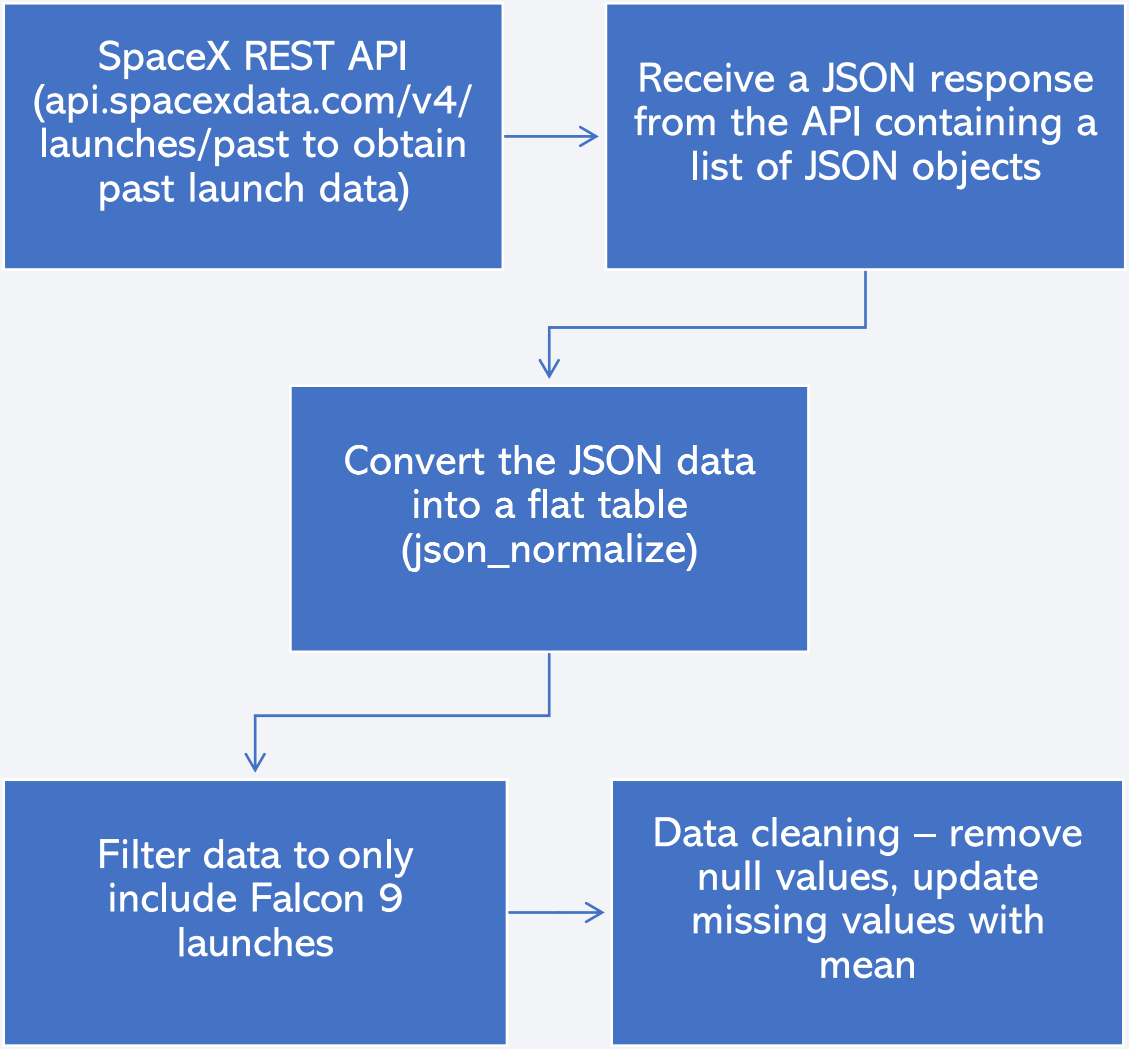
Executive Summary

- Data collection methodology:
  - Using SpaceX's REST APIs and with Web Scraping
- Perform data wrangling
  - Raw data needs to be cleaned, and transformed to suit the analysis requirements. This involves removing duplicates, updating missing values, and standardizing formats
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Access the accuracy of each classification model to determine best one

Data Collection – SpaceX API

- **Data Source:**
    - SpaceX REST API (api.spacexdata.com/v4/launches/past to obtain past launch data)
- **GET Request and Response:**
    - Make GET request to the API and convert the response to .json that contains a list of JSON objects
- **Data Transformation:**
    - Convert .json file to a pandas DataFrame
    - Define lists to store data and store this as dictionary
    - Filter data to only include Falcon 9 lauches
    - Replace missing values of PayLoadMass with the mean PayLoadMass value

[GithubLink_to_DataCollectionAPI]



SpaceX REST API (api.spacexdata.com/v4/launches/past to obtain past launch data)

Receive a JSON response from the API containing a list of JSON objects

Convert the JSON data into a flat table (json_normalize)

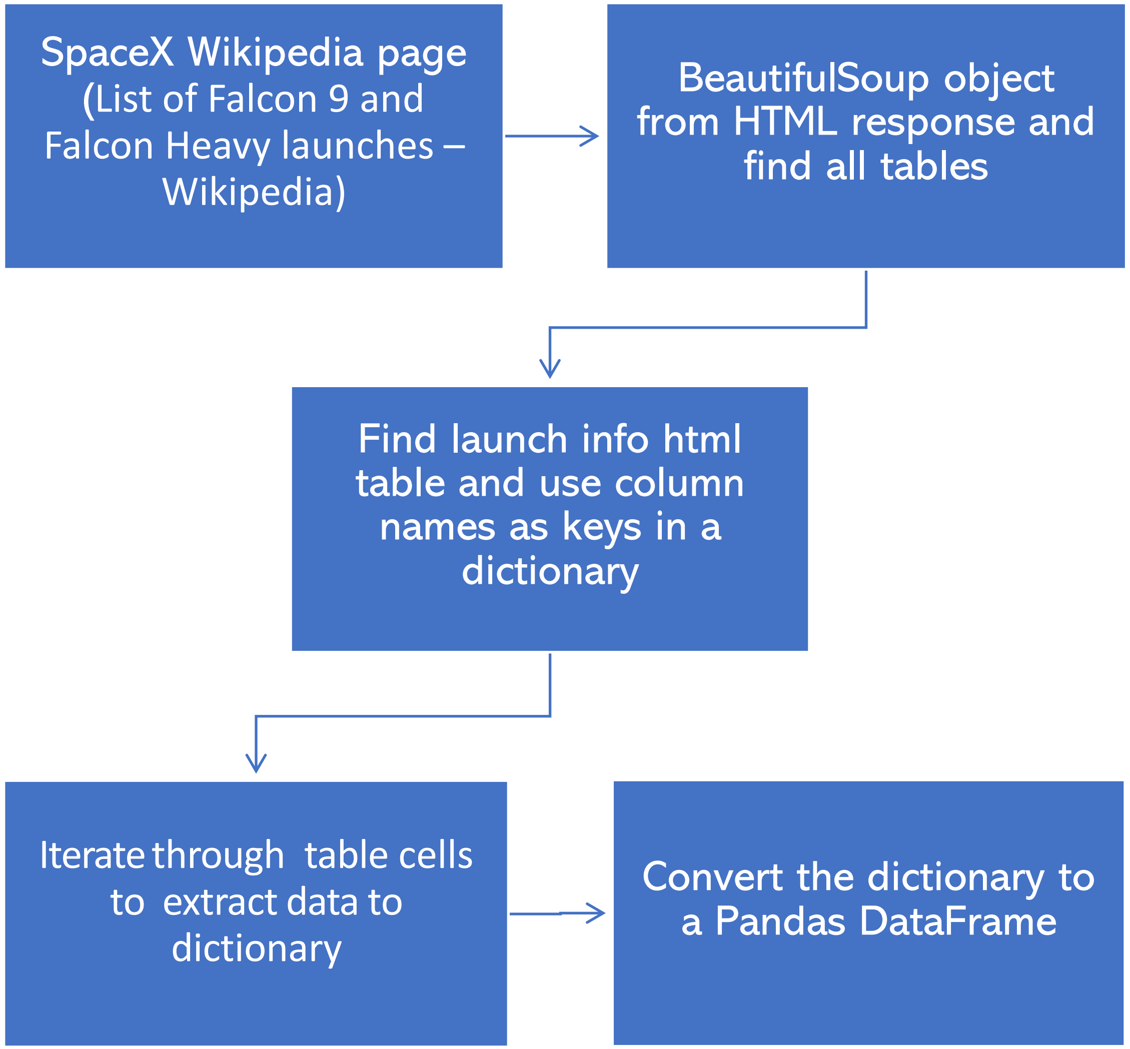
Filter data to only include Falcon 9 launches

Data cleaning – remove null values, update missing values with mean

Data Collection – Web Scraping

- **Data Source:**
  - SpaceX Wikipedia page (List of Falcon 9 and Falcon Heavy launches – Wikipedia)
- **Web Scraping:**
  - Utilizing the Python BeautifulSoup package to extract data from HTML tables
- **Data Parsing and Conversion:**
  - Collect all column names from the tables found within the HTML page
  - Use column names as keys in a dictionary
  - Use custom functions and logic to parse all launch tables to fill the dictionary values
  - Convert the dictionary to a Pandas DataFrame

GitHubLink_to_WebScraping


SpaceX Wikipedia page (List of Falcon 9 and Falcon Heavy launches – Wikipedia)
BeautifulSoup object from HTML response and find all tables
Find launch info html table and use column names as keys in a dictionary
Iterate through table cells to extract data to dictionary
Convert the dictionary to a Pandas DataFrame

Data Wrangling

- Main outcome of the data wrangling is to convert the outcomes into Training Labels - with 1 means the booster successfully landed, 0 means it was unsuccessful
  - True ASDS, True RTLS, & True Ocean are set to 1
  - None None, False ASDS, None ASDS, False Ocean, False RTLS are set to 0
- Process involved:
  - Load Space X dataset (data collected from API and web scraping)
  - Calculate the percentage of the missing values in each attribute
  - Identify which columns are numerical and categorical
  - Create a landing outcome label from Outcome column
- Calculate various metrics to access the performance of Falcon 9 at each orbit - number of launches on each site, number and occurrence of mission outcome of the orbits, number and occurrence of each orbit

  GitHubLink_to_DataWrangling

EDA with Data Visualization

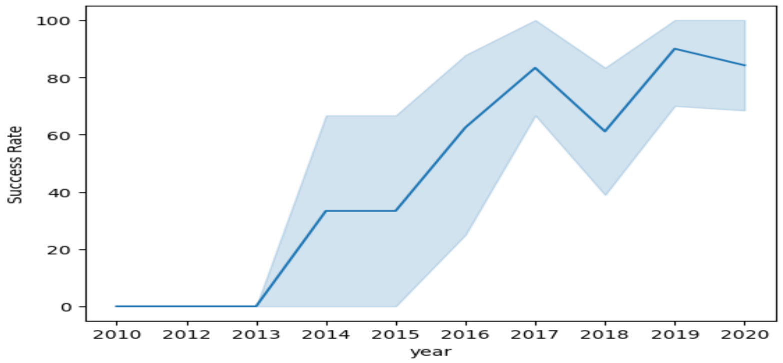
- In this phase, the data is further analyzed to predict if the Falcon 9 first stage will land successfully
  - This is important as the cost savings (compared to other competitors come from the fact that SpaceX can reuse the first stage
- Process involved:
  - Read the SpaceX dataset into a Pandas dataframe
  - Plot various scatter, line, and bar charts to understand the relationship between various variables
  - Plots include
    - Scatter plots - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Flight Number vs. Orbit, Payload vs Orbit
    - Bar chart - Orbit vs. Success Rate
    - Line chart - Success Yearly Trend

GitHubLink_to_EDAwithVisualization

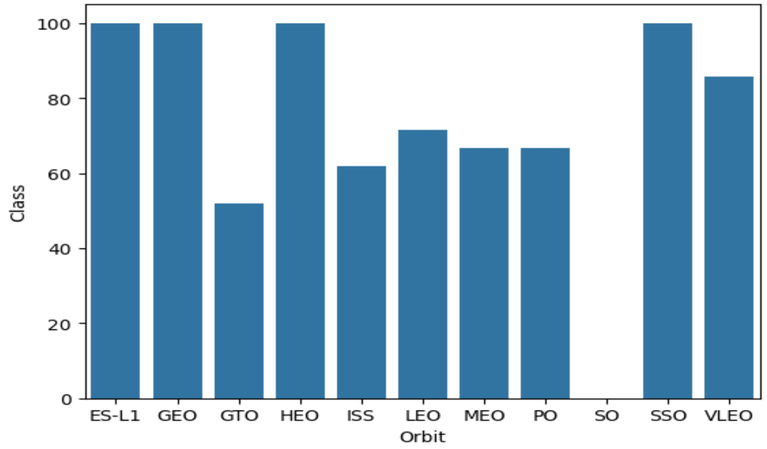
EDA with Data Visualization – contd.

- Important outcomes:
  - VAFB-SLC launchsite has no rockets launched for heavy payload mass (greater than 10000)
  - ES-L1, GEO, HEO, SSO orbits have highest success rates
  - LEO orbit success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success
  - With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. For GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present
  - The success rate of the launches constantly increasing since 2013 till 2020

Launch success yearly trend



Success rate of each orbit type



EDA with SQL

- This is to further understand the Spacex DataSet by loading the dataset into the corresponding table in a Db2 database and querying the data to get answers to the questions related to dataset
- Process involved:
    - Load the dataset into the corresponding table in a Db2 database
    - Query the db tables to perform below tasks
        - Remove blank rows from table
        - Display the names of the unique launch sites in the space mission
        - Display average payload mass carried by booster version F9 v1.1
        - Display the total payload mass carried by boosters launched by NASA (CRS)
        - List the date when the first successful landing outcome in ground pad was achieved
        - List the total number of successful and failure mission outcomes

GitHubLink to EDAwithSQL

Build an Interactive Map with Folium

- Purpose of this step is to do the following tasks:
  - TASK 1: Mark all launch sites on a map
  - TASK 2: Mark the success/failed launches for each site on the map
  - TASK 3: Calculate the distances between a launch site to its proximities
- Folium Map object has to be initialized and folium.Circle and folium.Marker has been added for each site
- Marker color to be assigned to success (1) as green and failure (0) as red
- Create an icon as text label, assigning the icon_color as the marker_color
- Using lat and long values calculate the distance between points
- folium.polyline can be used to display the distance line between two points

GitHubLink to InteractiveDataAnalyticsUsingFolium

Build a Dashboard with Plotly Dash

- Main purpose of a Dashboard with Plotly Dash includes to answer the following questions:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?
- Process involved:
  - Building a pie chart to show all sites and individual sites upon filtering vs. their success rates – use px.pie(), dcc.dropdown() object
  - Building a scatter graph to show the correlation between outcome (success vs. unsuccess) and payload mass (kg) – use px.scatter(), RangeSlider() object

  GitHubLink_to_InteractiveDashboardwithPlotly

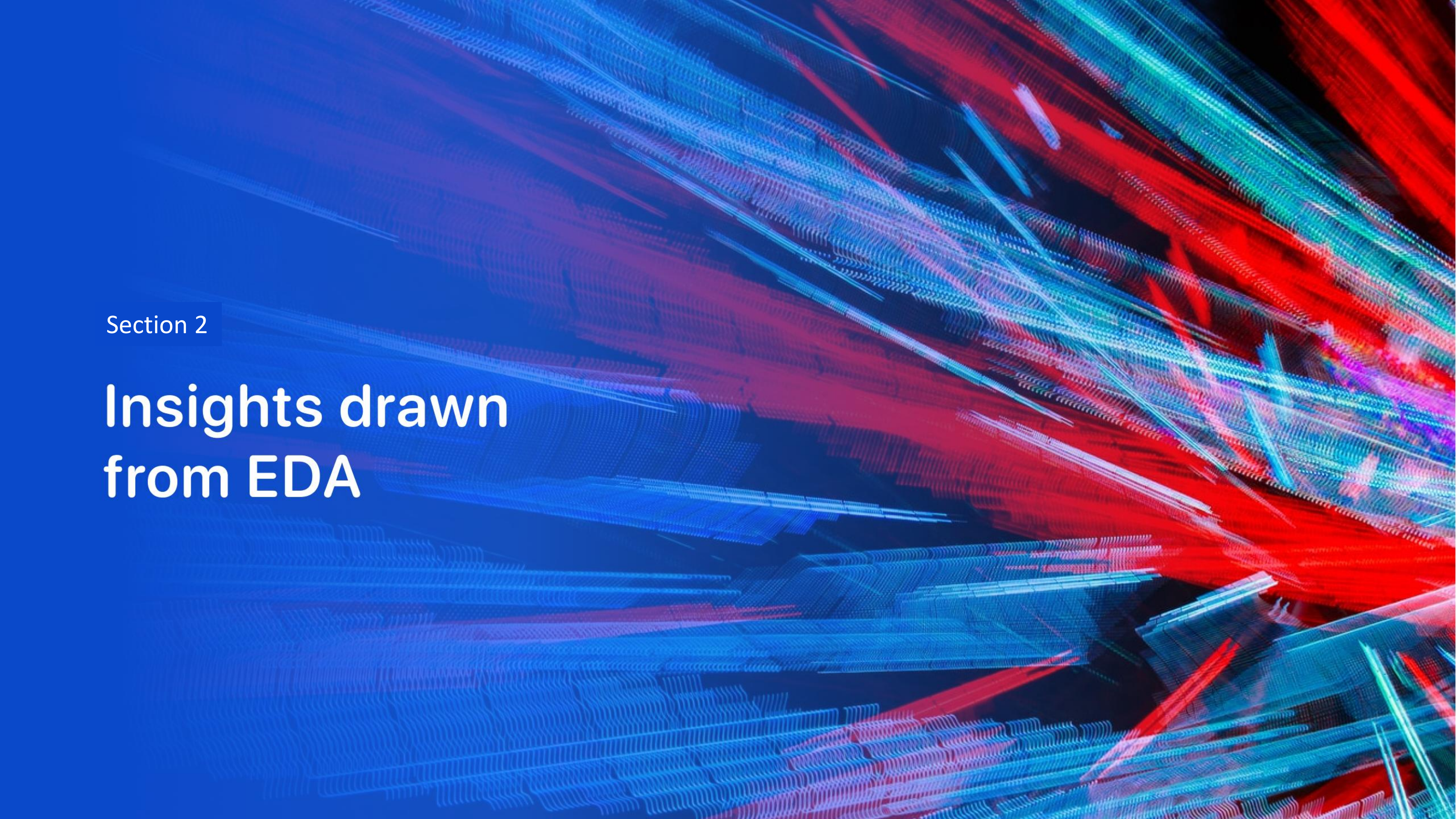
Predictive Analysis (Classification)

- Objective of this analysis is to find the best Hyperparameter for SVM, Classification Trees and Logistic Regression and find the method performs best using test data
- Process involved:
  - Create a column for the class: Create a NumPy array from the column Class in data
  - Standardize the data: Standardize the data using StandardScalar() and fit_transform() methods and reassign it to the dataframe
  - Split into training data and test data: Use a 20% of the data for testing and remaining for training the model. Create a logistic regression, support vector machine, decision tree classifier, and KNN objects and fit these objects to find best parameters.
  - Identify the best method with high accuracy: Calculate the accuracy and scores of each of these methods to identify best methods.

  GitHubLink_to_MachineLearningPredictionAnalysis

Results

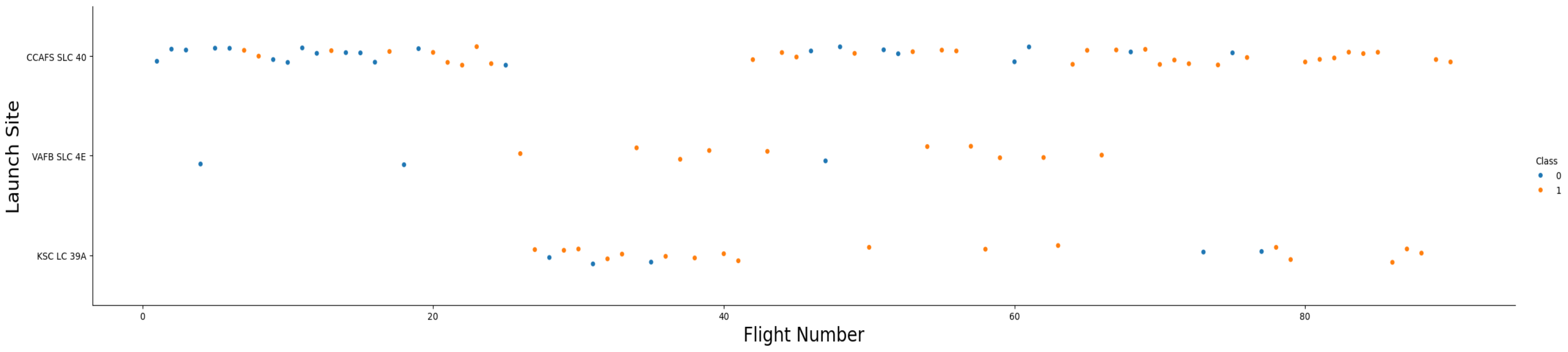
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

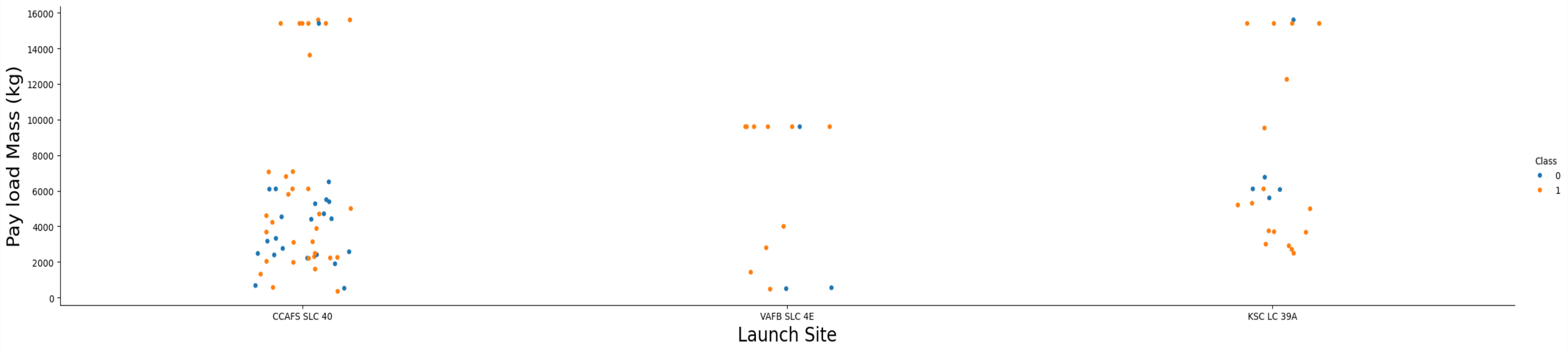
Insights drawn from EDA

Flight Number vs. Launch Site



- Orange dots indicate successful launch, and blue dots indicate unsuccessful launch
- As the number of flights increase, the successful launches increase as well
- Above 40, the successful launches increased significantly

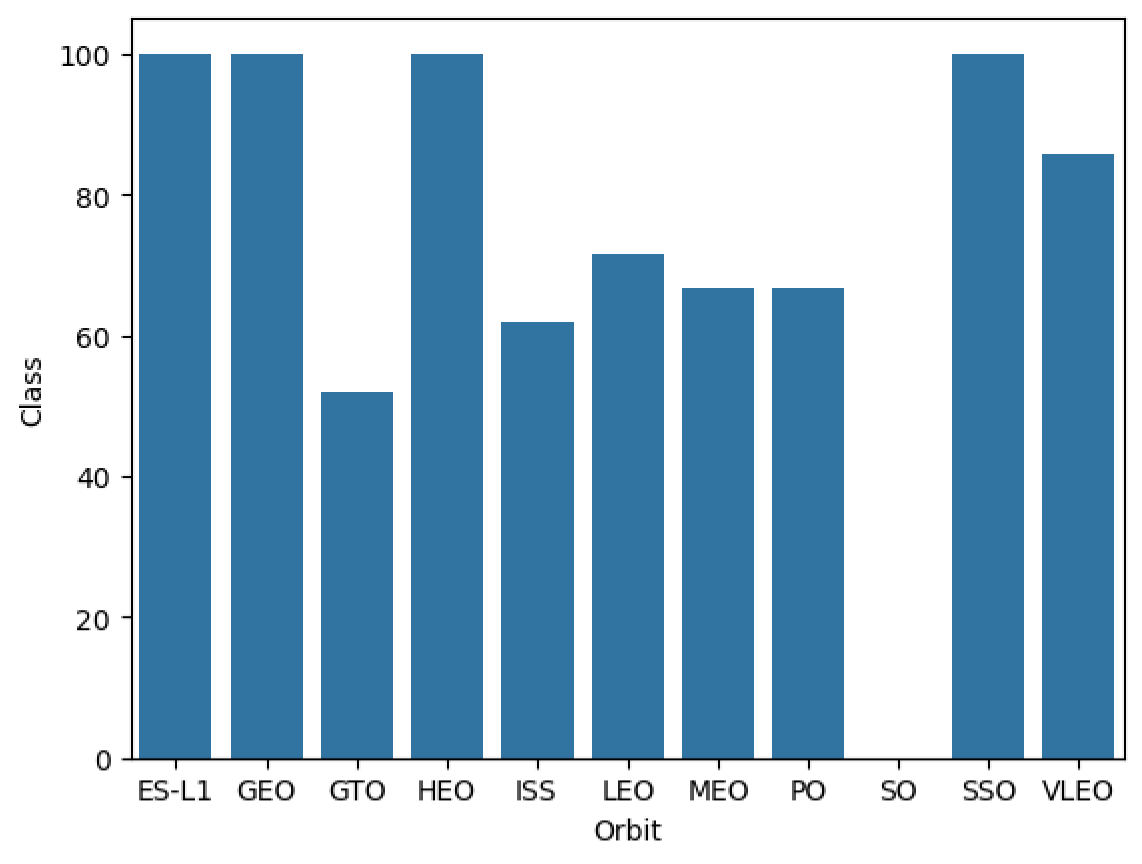
Payload vs. Launch Site



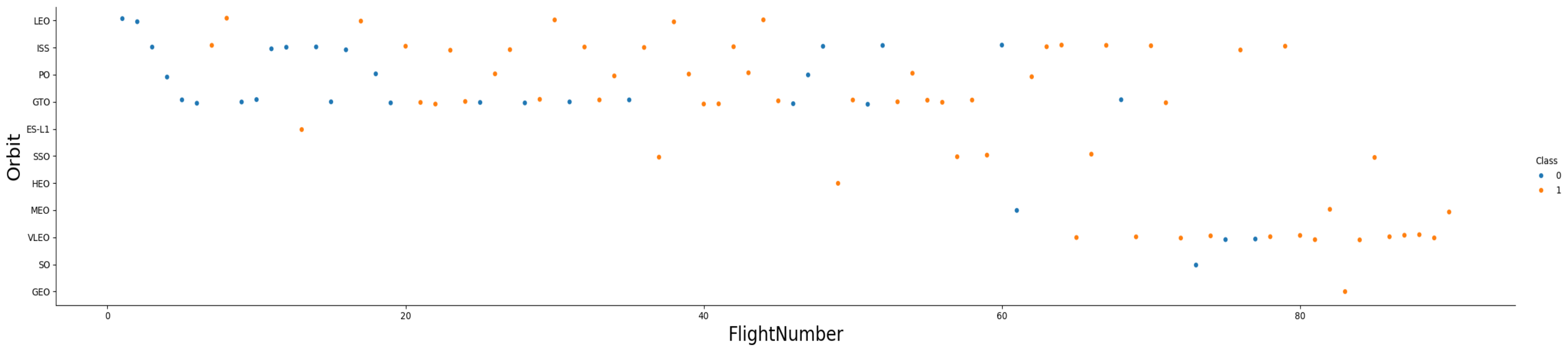
- Orange dots indicate successful launch, and blue dots indicate unsuccessful launch
- Unsuccessful events are relatively less above 6,000 kg payload
- All the sites launched high and low payloads. CCAPS SLC 40 launched relatively smaller payloads

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO orbits have 100% success rate
- While SO orbit type has the least success rate i.e., 0%

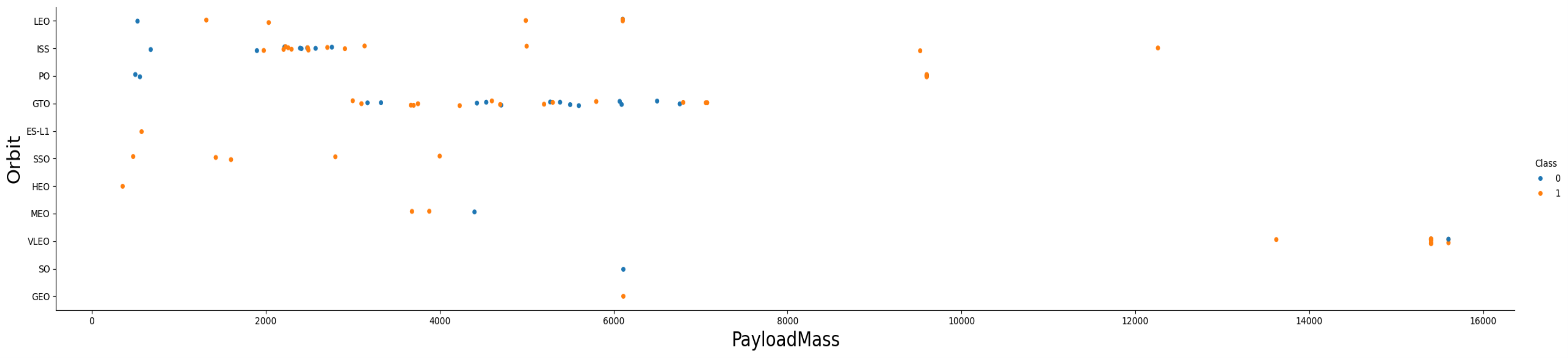


Flight Number vs. Orbit Type



- As observed in the previous bar chart, ES-L1, GEO, HEO, and SSO has 100% success rate. Here, we can see these orbits have only one flight except SSO which has 5 flights.
- Other than this, there is little relationship between successful flights and orbits

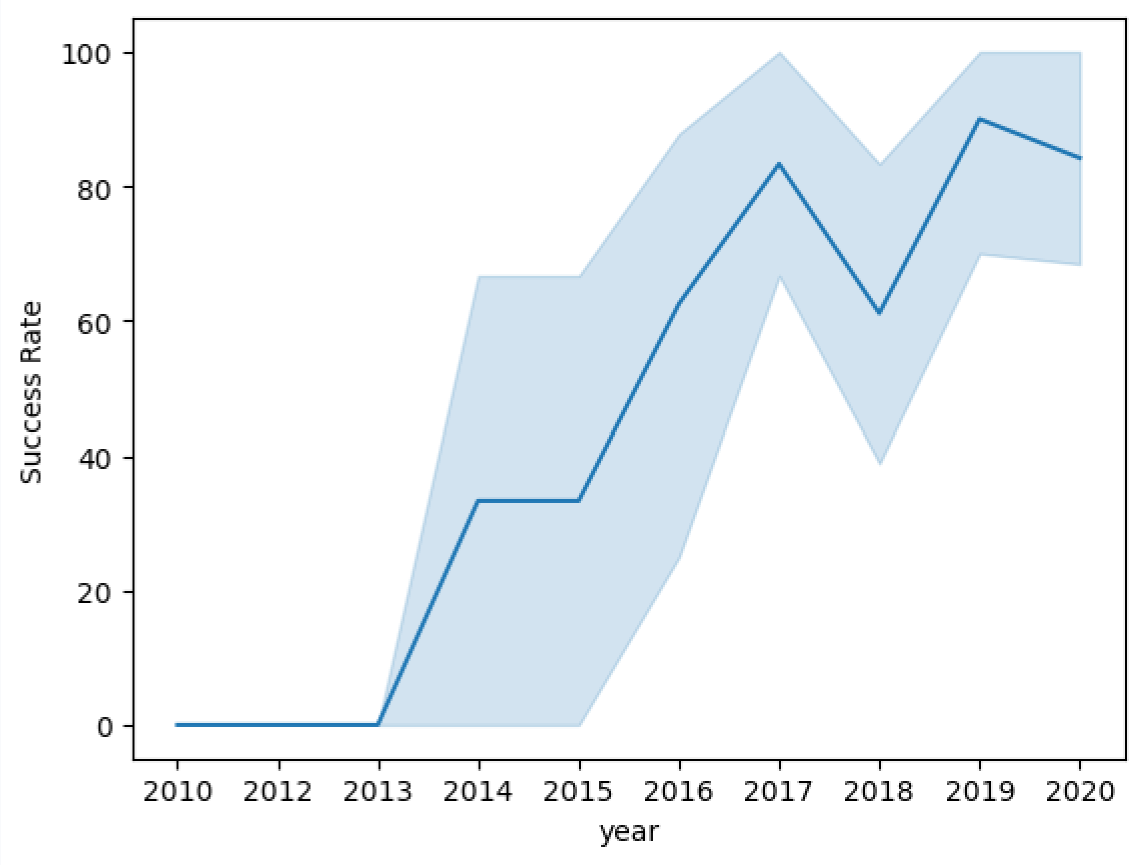
Payload vs. Orbit Type



- Success rate of the launch is higher for higher payloads irrespective of orbit
- Some orbits have high rate of success for lower payload mass
  - ES-L1, SSO, HEO, LEO

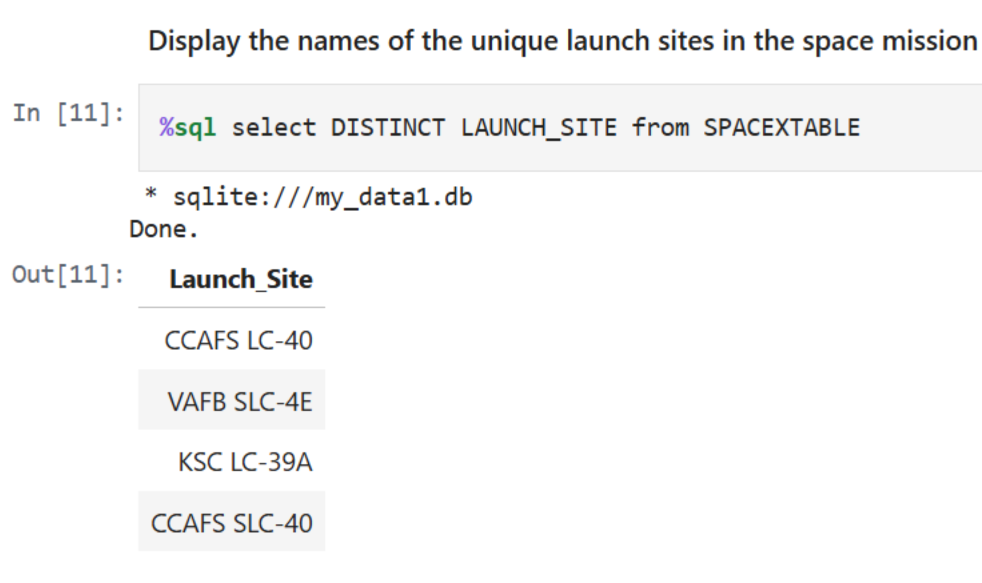
Launch Success Yearly Trend

- Success rate has been constantly increasing since 2013 until 2017
- Year 2018 saw a dip in success rate and it has been recovered in subsequent years
- It is good idea to further analyze the external factors involved in low success rate in the year 2018



All Launch Site Names

- Below screenshot shows the list of all launch sites



Display the names of the unique launch sites in the space mission

```
In [11]:  %sql select DISTINCT LAUNCH_SITE from SPACEXTABLE
```

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

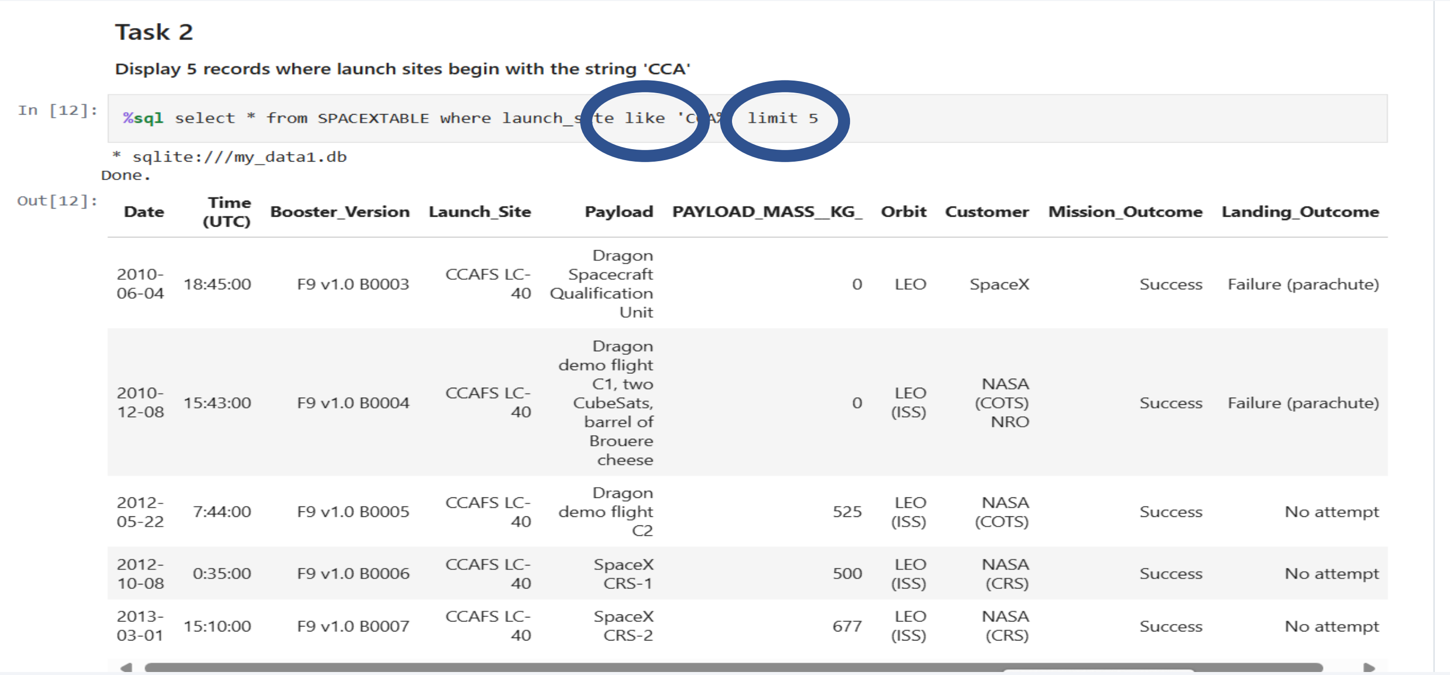
Out[11]:

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

- CCAFS LC-40 and CCAFS SLC-40 could be the same sites and need some additional information to validate this assumption and accurately represent all metrics

Launch Site Names Begin with 'CCA'

- ‘like’ and ‘limit 5’ helps in retrieving 5 records where launch sites begin with the string ‘CCA’



Task 2

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

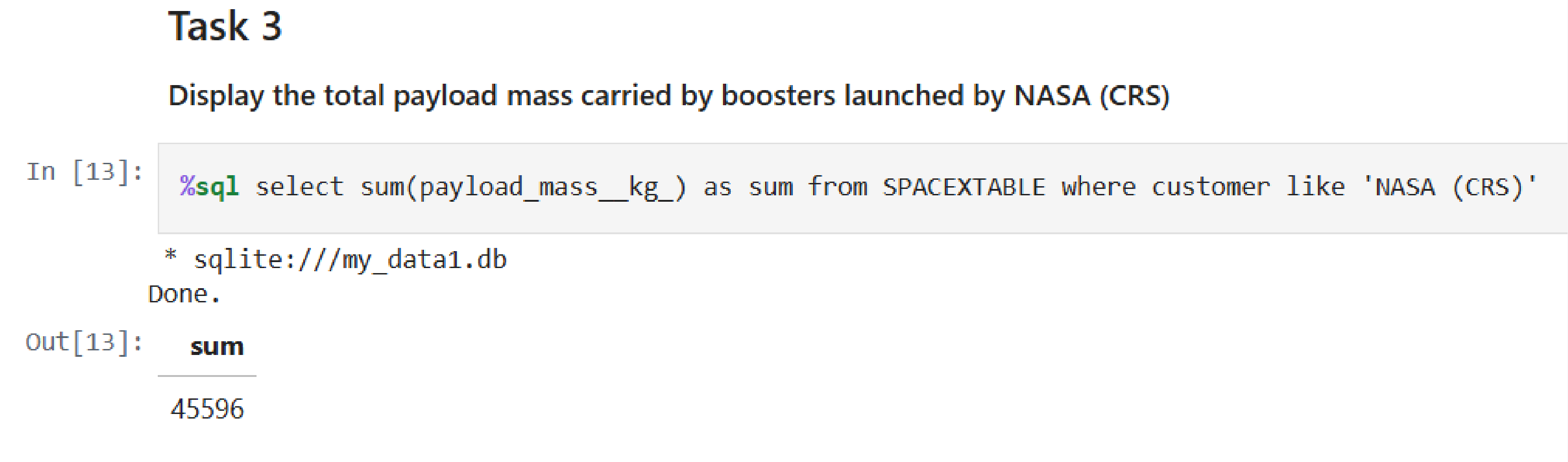
```
In [12]:  %sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5

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

Out[12]:

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

Total Payload Mass

- Total payload carried by boosters from NASA is 45,596 kgs
- ‘SUM’ is the keyword used here to calculate total payload



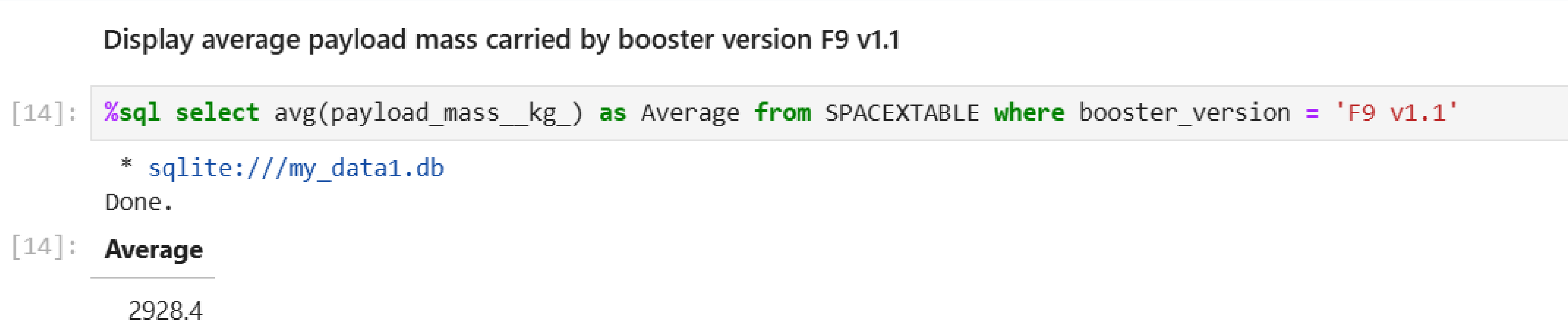
Task 3

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

```
In [13]: %sql select sum(payload_mass__kg_) as sum from SPACEXTABLE where customer like 'NASA (CRS)'

 * sqlite:///my_data1.db
Done.

Out[13]:  sum
         45596
```

Average Payload Mass by F9 v1.1

- On an average 2,928 kg payload mass carried by booster version F9 v1.1
- ‘avg’ is used to calculate the average payload; ‘like’ or ‘=‘ can be used to specify the condition based on the specific requirements



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

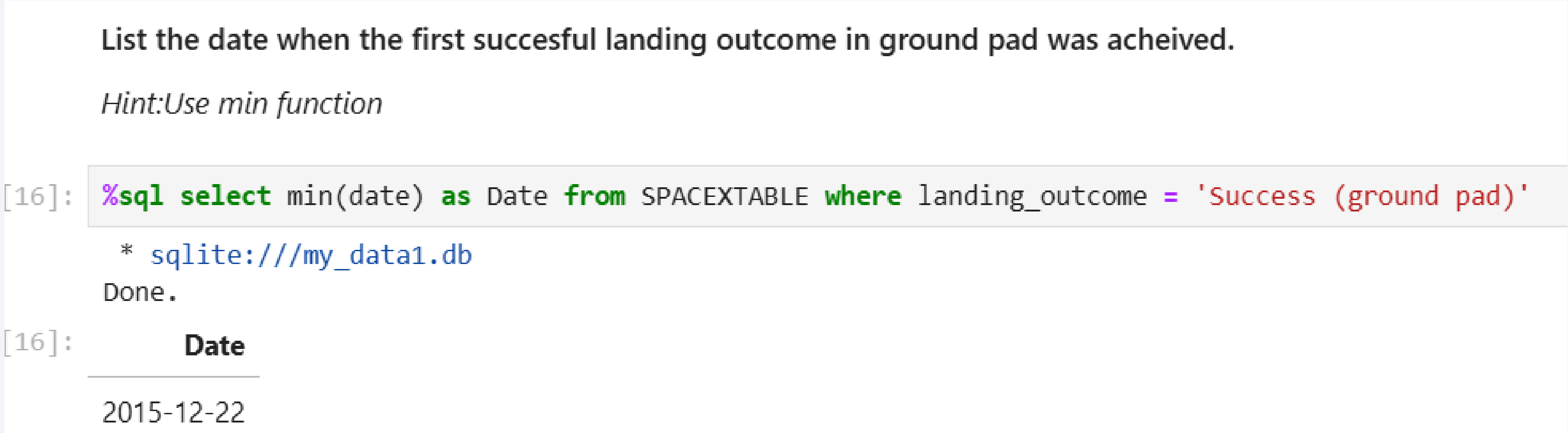
```
[14]: %sql select avg(payload_mass__kg_) as Average from SPACEXTABLE where booster_version = 'F9 v1.1'

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

[14]:

Average
2928.4

First Successful Ground Landing Date

- The first successful landing outcome on ground pad happened on December 22, 2015
- ‘min’ is used to find the earliest date of success



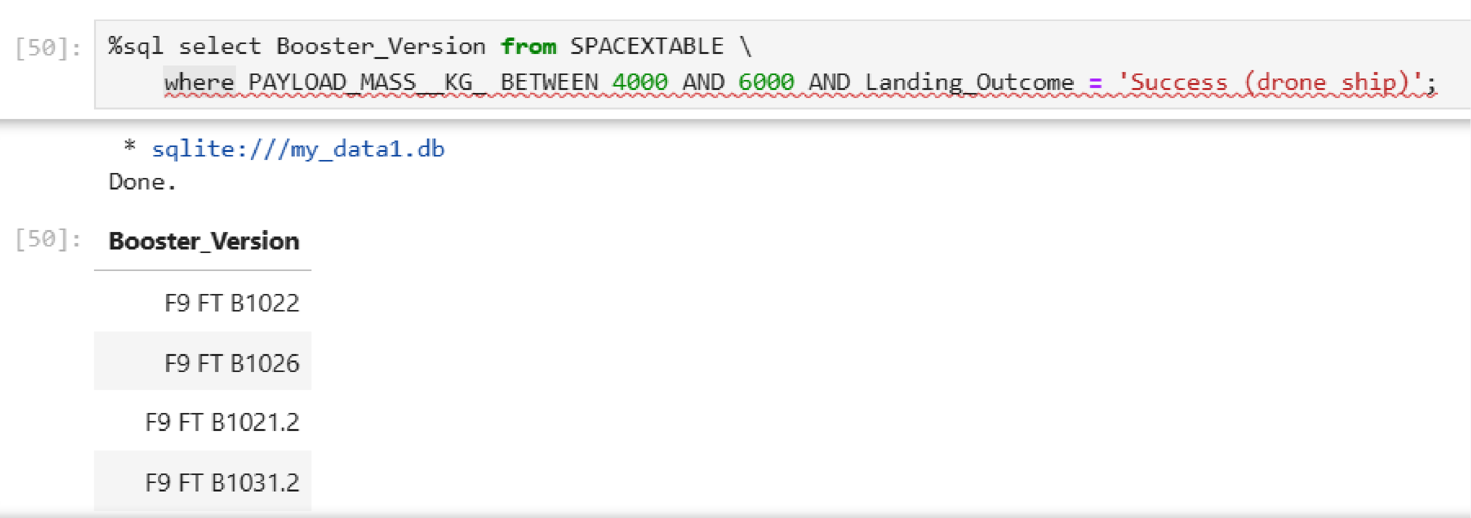
List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
[16]: %sql select min(date) as Date from SPACEXTABLE where landing_outcome = 'Success (ground pad)'

 * sqlite:///my_data1.db
Done.

[16]:       Date
      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 is determined using ‘between’ and ‘and’ conditions
- There are total 4 of them



```
[50]: %sql select Booster_Version from SPACEXTABLE \
          where PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Success (drone ship)';
```

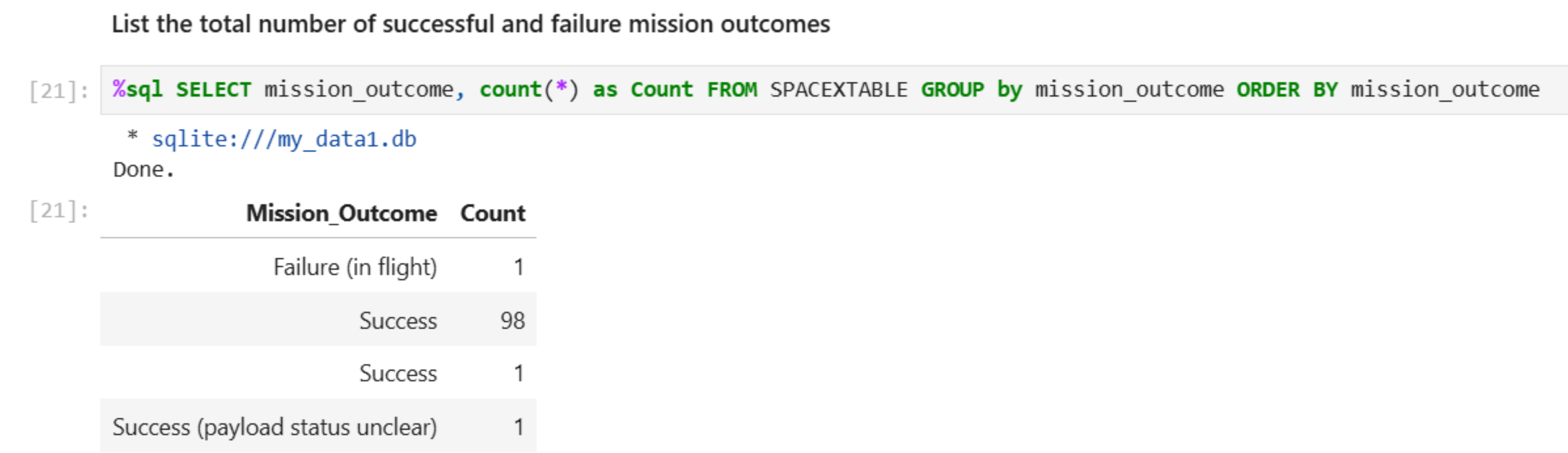
```
 * sqlite:///my_data1.db
Done.
```

[50]:

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

Total Number of Successful and Failure Mission Outcomes

- As per the results below, more than 99% of the mission outcomes are success
- Out of 101 mission outcomes, only 1 is a failure. This shows the immense planning and tight execution of missions



List the total number of successful and failure mission outcomes

```
[21]: %sql SELECT mission_outcome, count(*) as Count FROM SPACEXTABLE GROUP by mission_outcome ORDER BY mission_outcome
```

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

[21]:

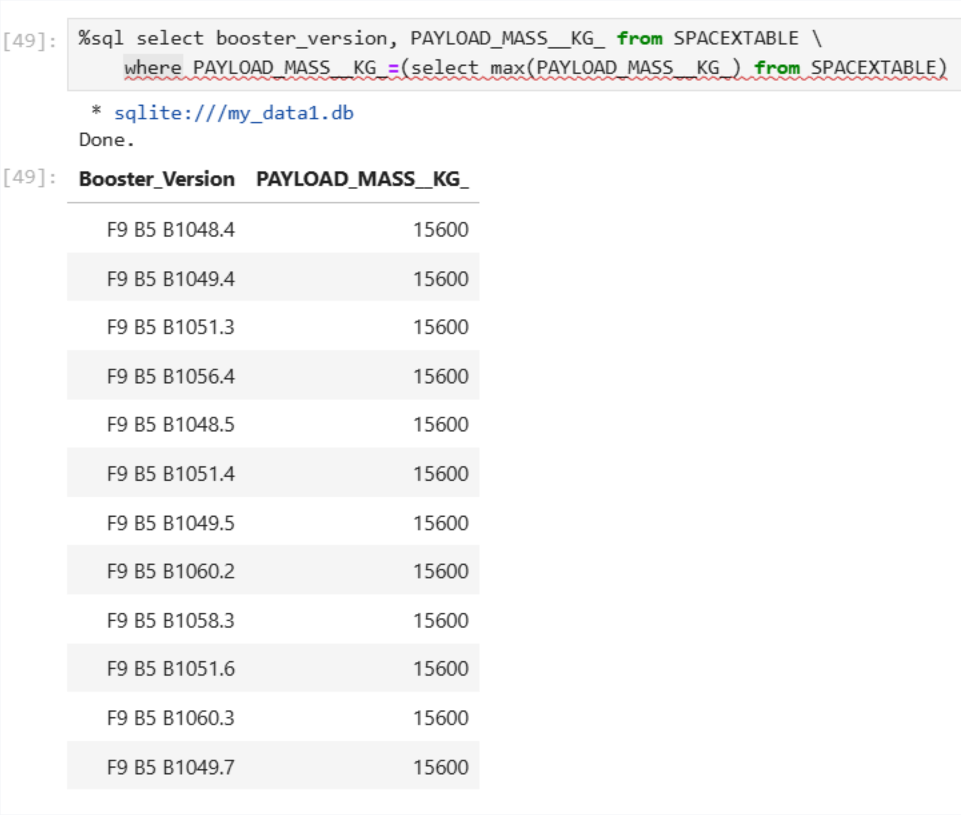
Mission_Outcome	Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

I have two rows of mission outcome as “Success”. This may be due to space at the beginning or ending.

This can be avoided by further data cleansing.

Boosters Carried Maximum Payload

- This query used ‘max’ function to get the list of booster versions that carried maximum payload
- ‘F9 B5 B10XX.X’ are the versions carried maximum payloads
- Further analysis needed:
    - Are there any ‘F9 B5 B10XX.X’ boosters which doesn’t carry maximum payloads?
    - What are the technical limitations or other reasons for any other versions not carrying maximum?



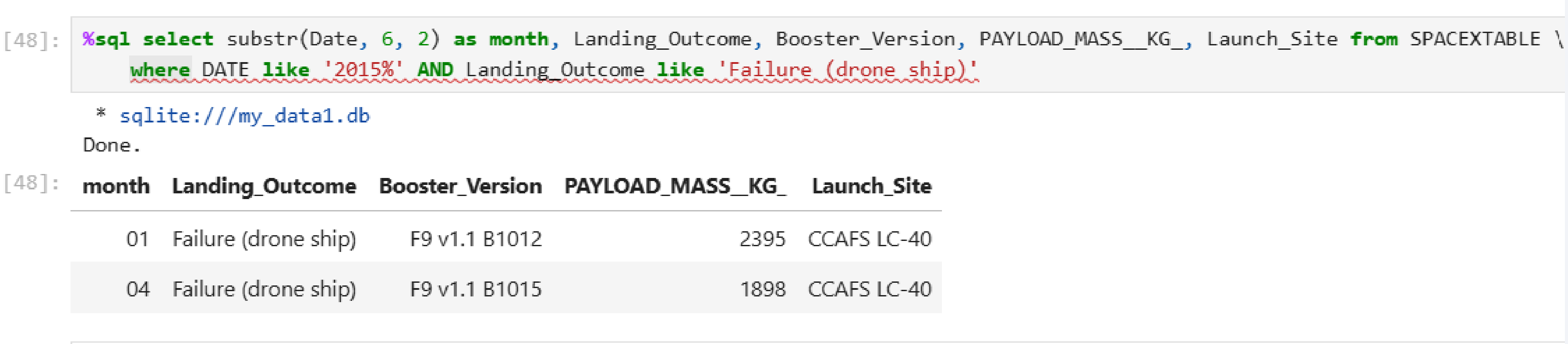
```
[49]: %sql select booster_version, PAYLOAD_MASS__KG_ from SPACEXTABLE \
          where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)

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

[49]:

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

- There are two failures in 2015 in drone ship. These two happened in the months of January and April
- Payloads on these both failed missions are on lower side



```
[48]: %sql select substr(Date, 6, 2) as month, Landing_Outcome, Booster_Version, PAYLOAD_MASS__KG_, Launch_Site from SPACEXTABLE \
          where DATE like '2015%' AND Landing_Outcome like 'Failure (drone ship)'
```

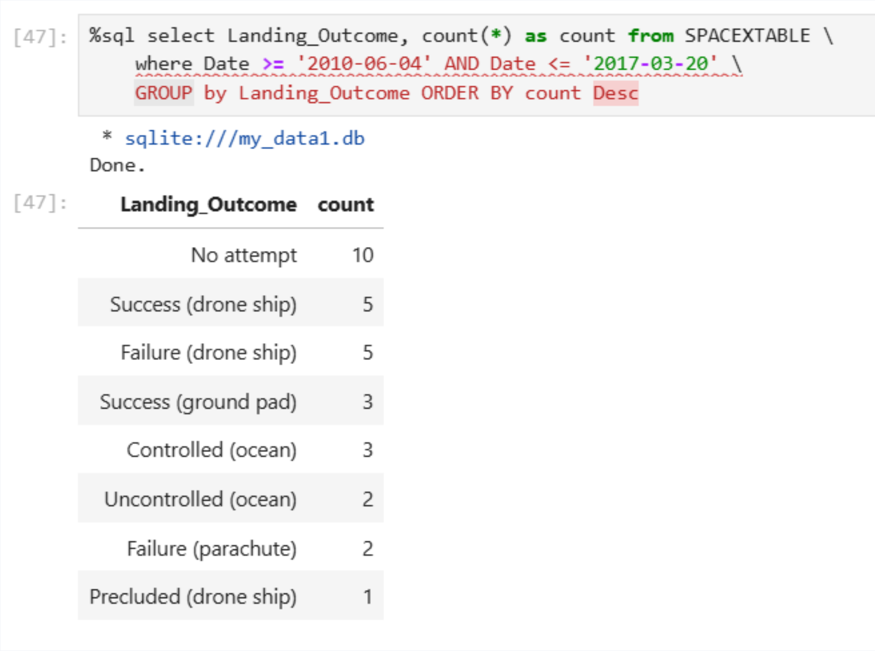
```
 * sqlite:///my_data1.db
Done.
```

[48]:

month	Landing_Outcome	Booster_Version	PAYLOAD_MASS__KG_	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

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

- Between the date 2010-06-04 and 2017-03-20, highest category of the landing outcome is labelled as “No Attempt”. This need to be further investigated
- ‘where’ clause is used along with ‘>=, <=‘ operators to get the desired outcome



```
[47]: %sql select Landing_Outcome, count(*) as count from SPACEXTABLE \
          where Date >= '2010-06-04' AND Date <= '2017-03-20' \
          GROUP by Landing_Outcome ORDER BY count Desc

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

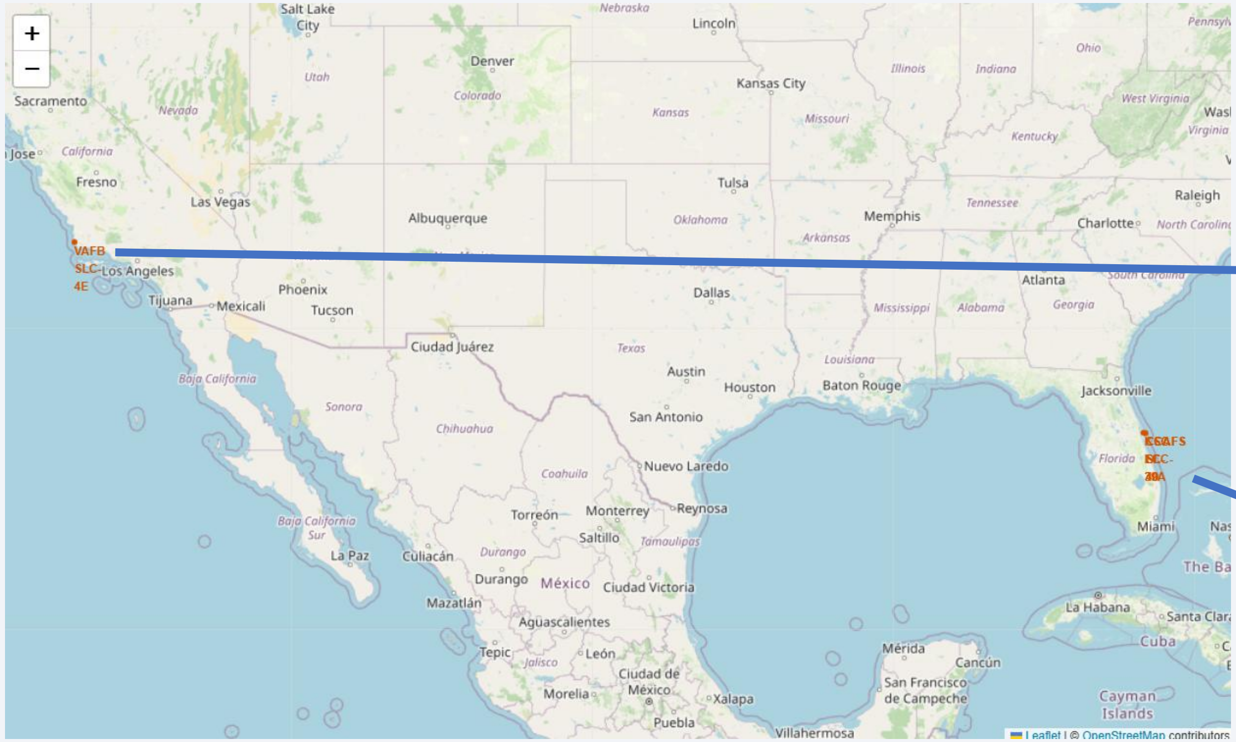
[47]:

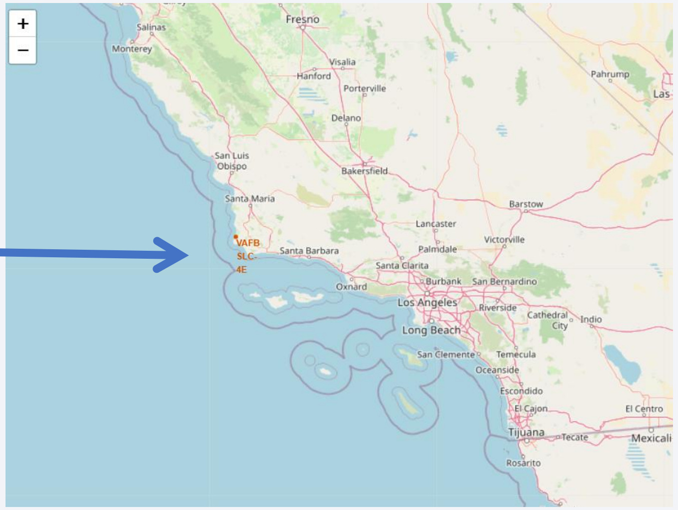
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

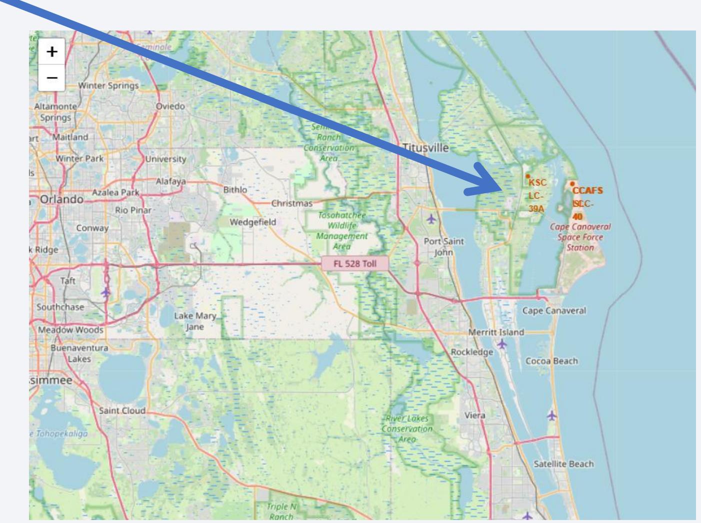
Section 3

Launch Sites Proximities Analysis

Launch Sites Folium Map

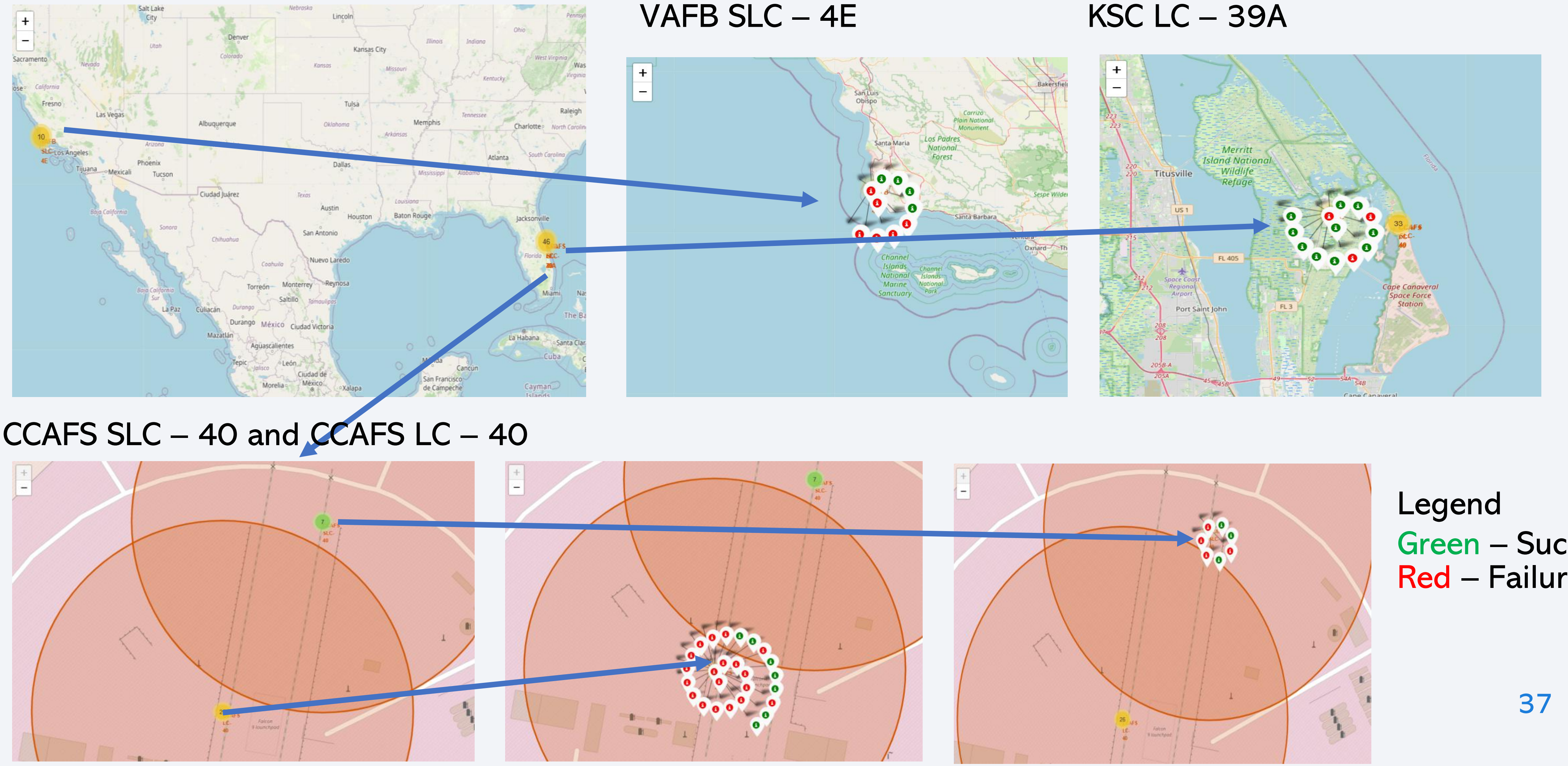






- These folium maps showcase the launch sites. Zoomed-in versions are provided for California and Florida launch sites.

Success and Failed Launches for Each Site



VAFB SLC – 4E

KSC LC – 39A

CCAFS SLC – 40 and CCAFS LC – 40

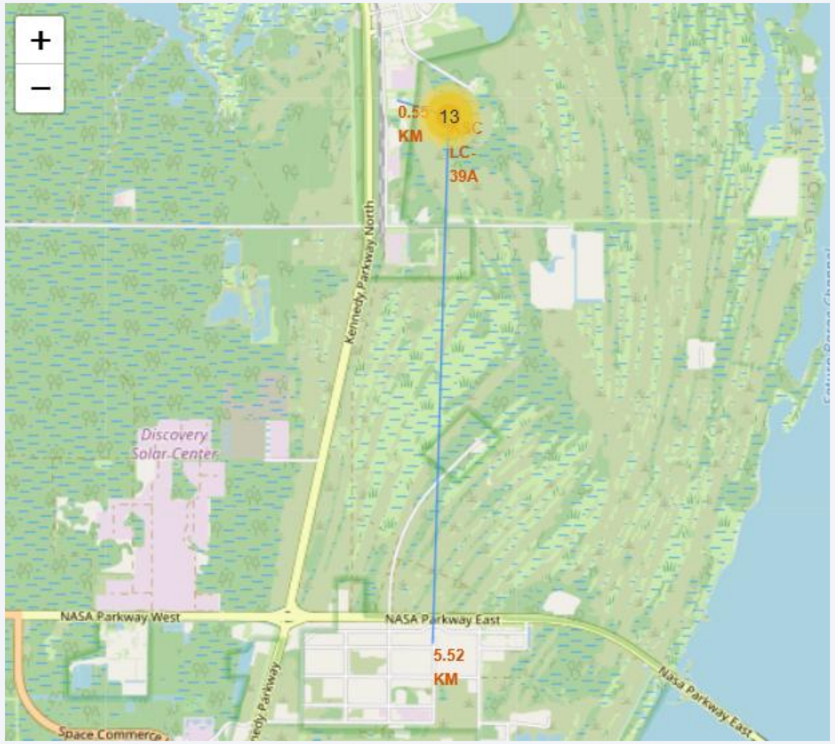
Legend

Green – Success

Red – Failure

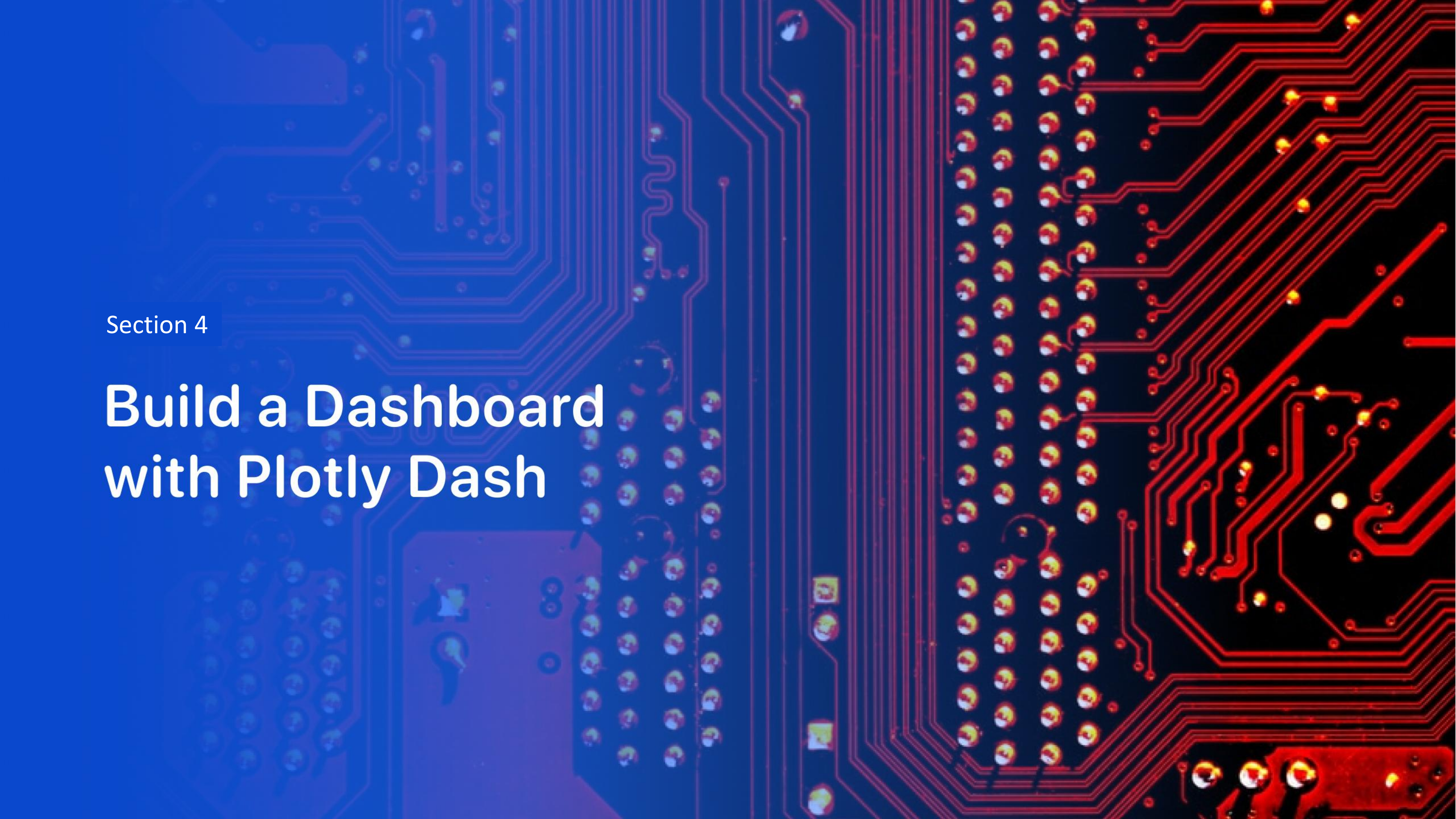
Proximity Map of the Launch Sites

- Given sample (KSC LC-39A) shows that the launch sites are near to the railways, highways, and coasts for convenience in terms of supply chain and transportation activities

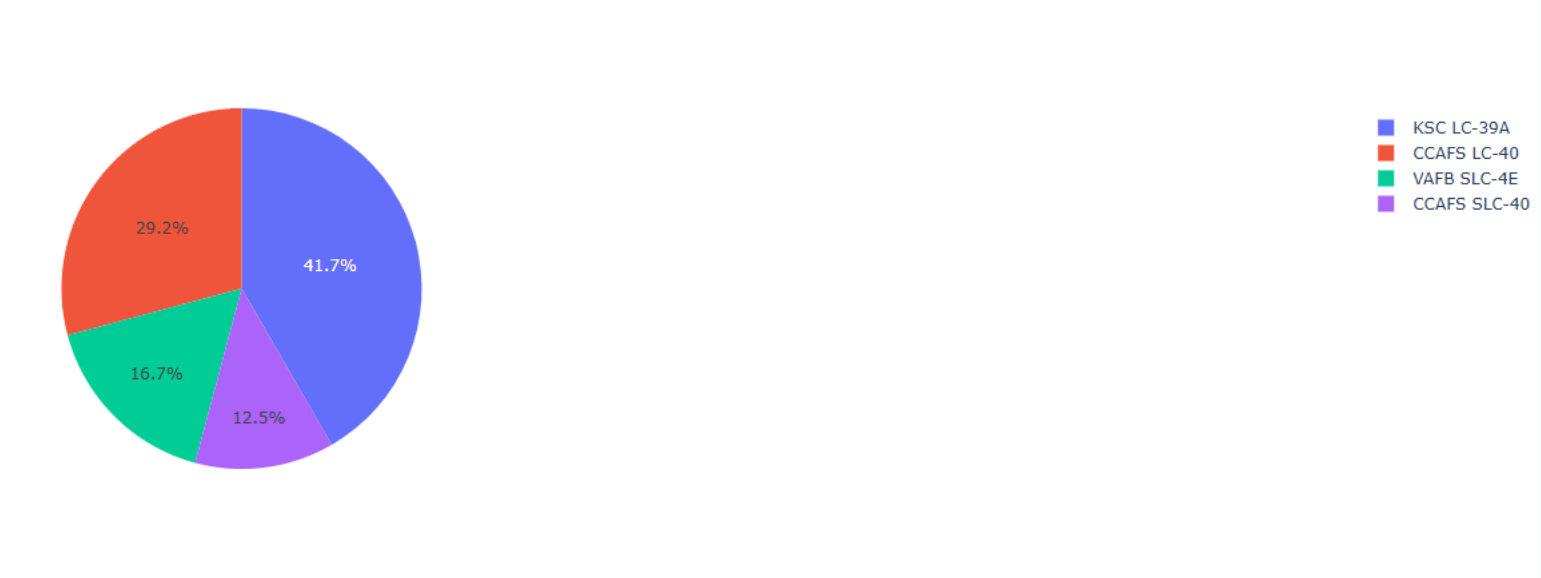


Section 4

Build a Dashboard with Plotly Dash



Total Success Launches by Site

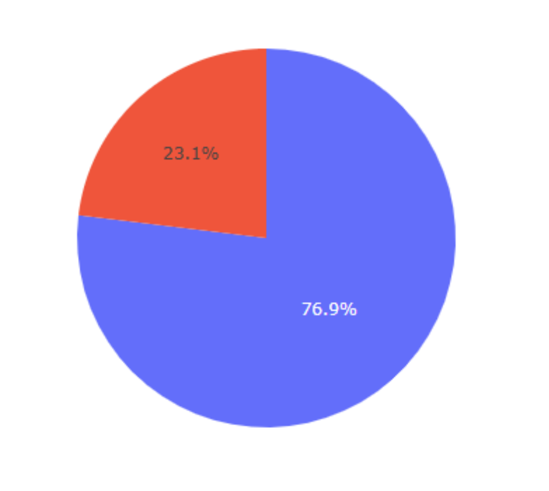


- Out of all KSC LC – 39A has highest percentage of successful launches
- Based on map, CCAFS LC – 40 and CCAFS SLC – 40 seems to be the same site
  - So, this site contributed same amount of successful launches as KSC LC – 39A site

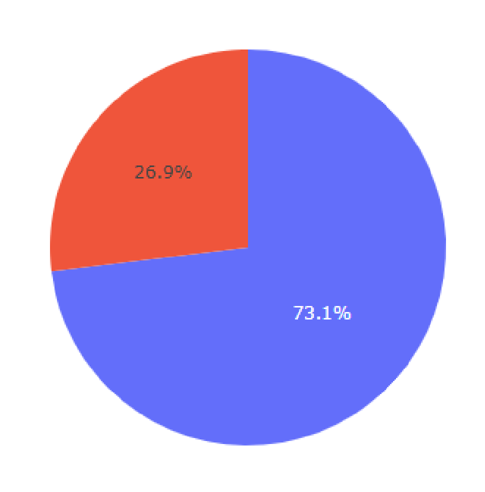
Success Rate by Launch Site

- KSC LC – 39A site has almost 77% of success rate out of total launches.
- CCAFS SLC – 40 site has lowest success rate of all site at 57.1%

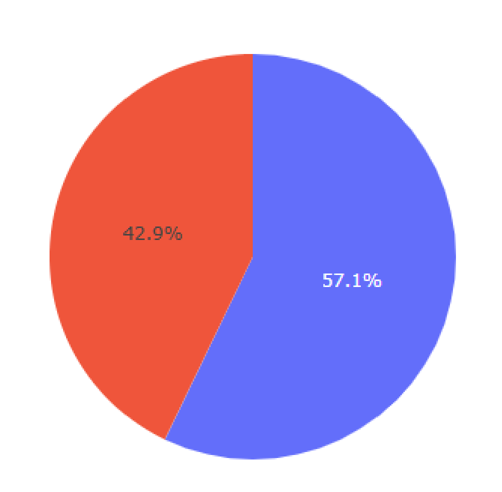
KSC LC – 39A



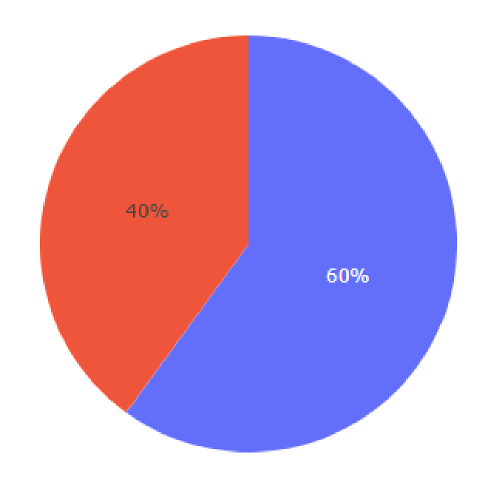
CCAFS LC – 40



CCAFS SLC – 40

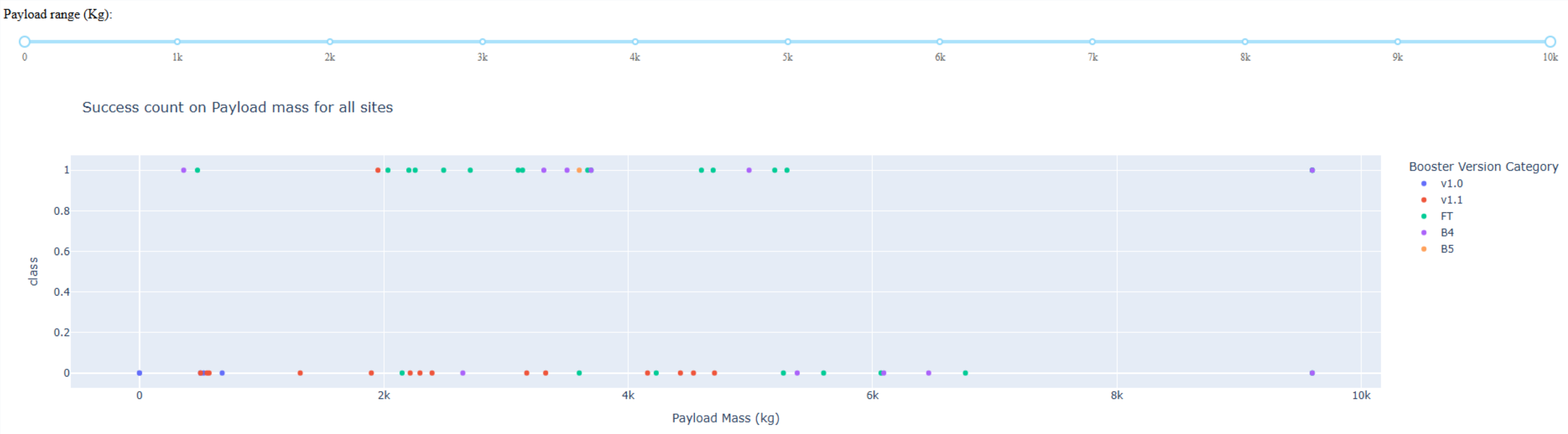


VAFB SLC – 4E



Launch Outcome Vs. PayLoad for All Sites

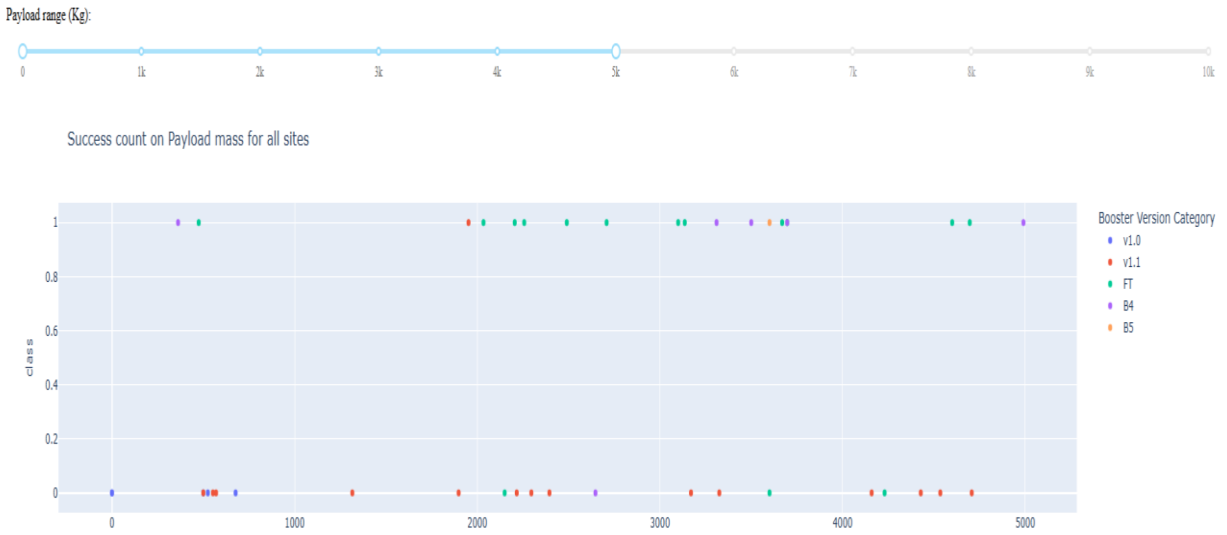
- Booster Version Category v1.1 has lot of unsuccessful launches comparatively
- This can be further analyzed by dividing it into small (0 to 5K kg) vs big payloads (5K to 10K kg)

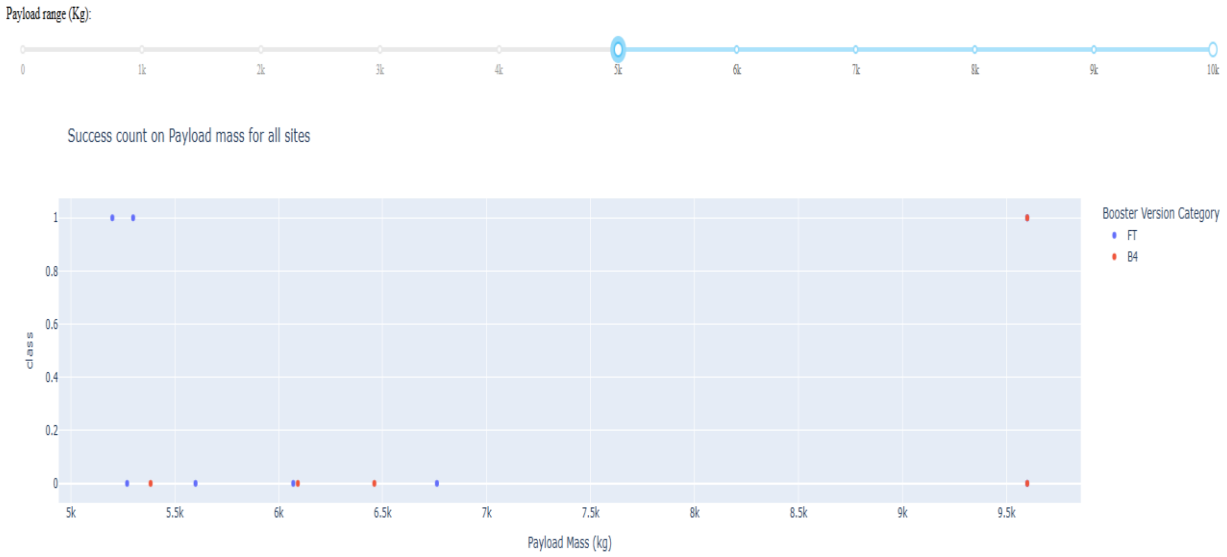


Launch Outcome Vs. PayLoad for All Sites – Contd

- Bigger payloads have less success rate compared to smaller payloads

- v1.0 and v1.1 booster version categories are specifically carrying smaller payloads and their success rate is much lower



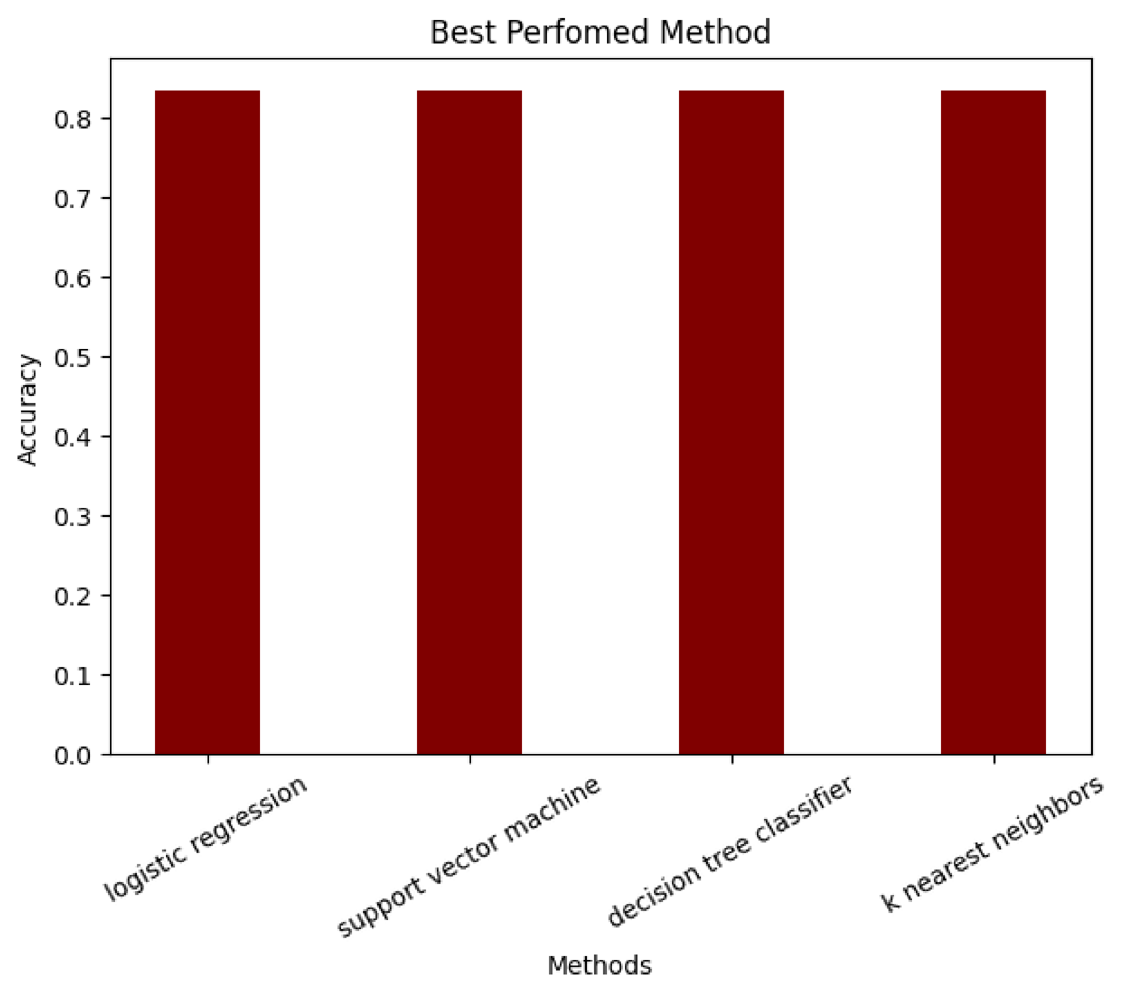


Section 5

Predictive Analysis (Classification)

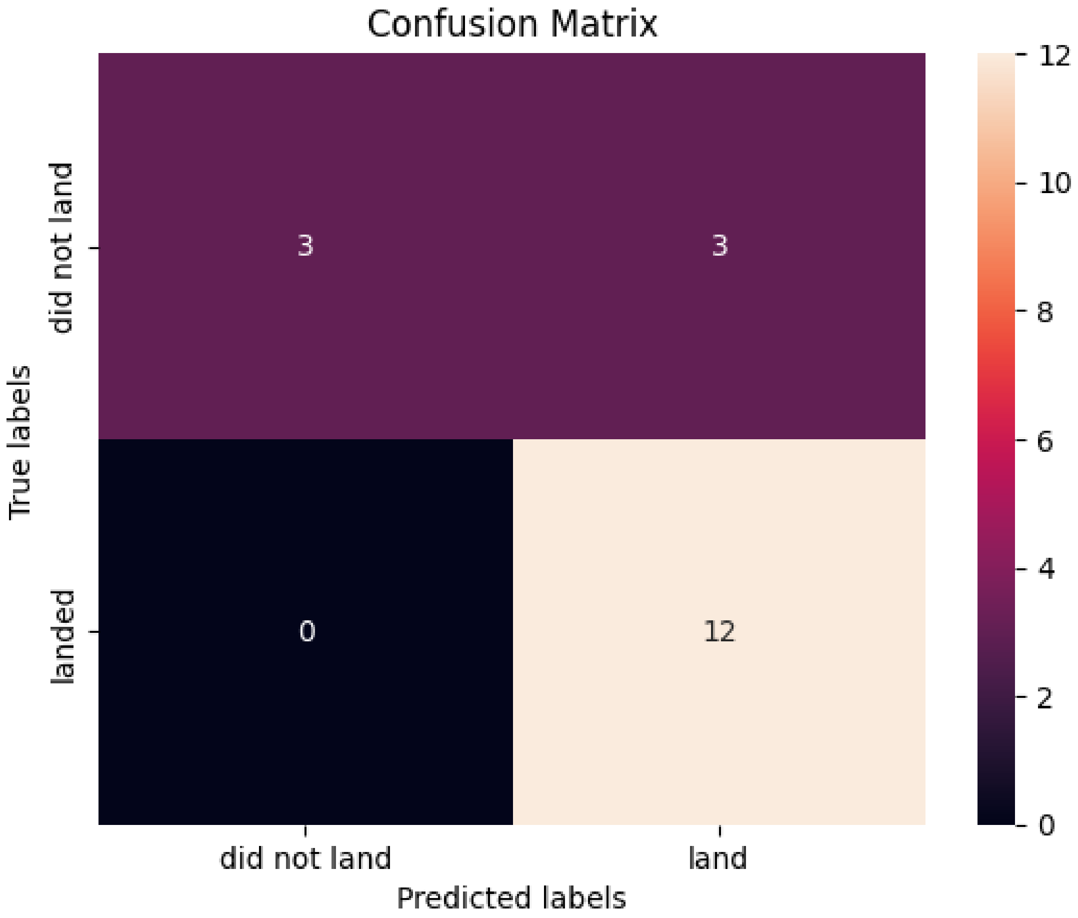
Classification Accuracy

- Based on the accuracy scores all the methods are providing same amount of accuracy
- However, decision tree classifier score is slightly higher than other methods.
- Small test sample size could be the reason for this same accuracy levels and this analysis could be done with bigger testing sample size



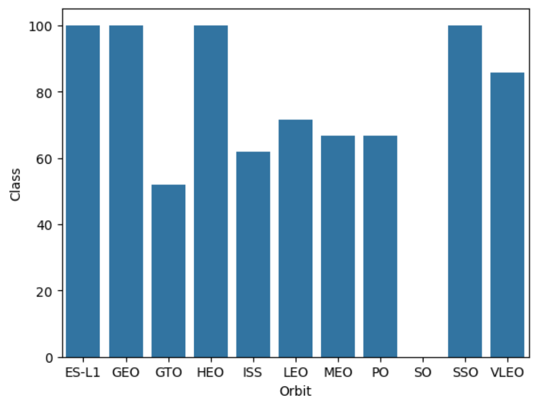
Confusion Matrix

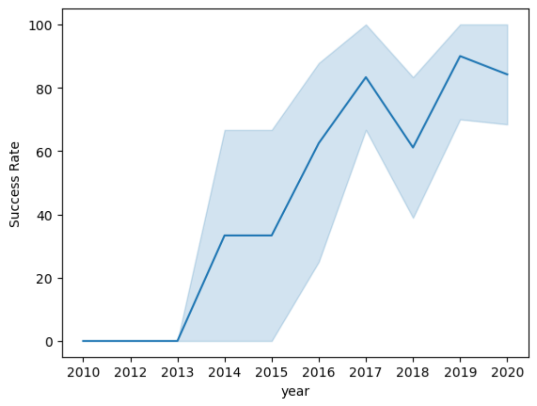
- Confusion matrix shows 3 of the occurrences which are actually not landed are labeled as landed (false positives)
- The model did not show any false negative error
- Increasing test sample size might give us different set of results and this is necessary to identify best performing model.

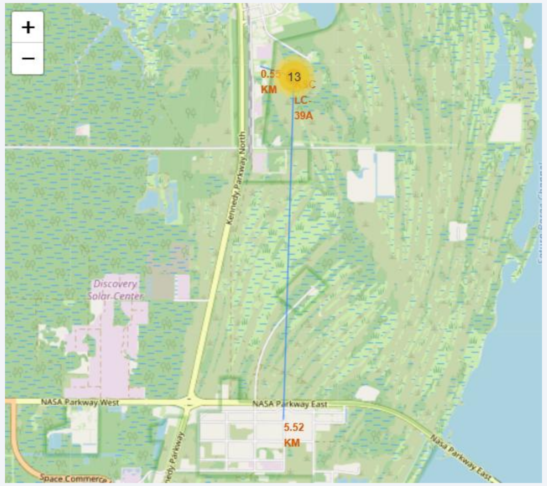


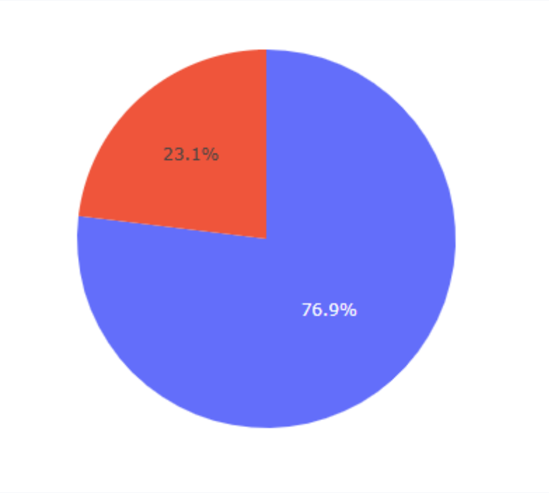
Conclusions

- ES-L1, GEO, HEO, and SSO orbits have 100% success rate
  - While SO orbit type has the least success rate i.e., 0%
- Year 2018 saw a sudden dip in successful launches compared to constant increase of successful launches from 2013 to 2017
- Launch sites are located near by railway stations, airports, and coasts for easy logistics access
- KSC LC – 39A is the site with highest success % around 77%
- All the predictive models are giving similar accuracy % - This can further be validated by changing the test and train sample sizes









Appendix

- All the jupyter lab files, and python files have been uploaded to GitHub repository. Use the below link to access those files

  VisaliNerla_DataScienceCapstone_Lab_Files

Thank you!