IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thomas Sinmegn
December 8/2023

Outline

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

Executive Summary

❖ Summary of Methodologies

- The research aims to predict if the Falcon 9 first stage will land successfully using machine learning algorithm. It also aims to determine which machine learning algorithms or models are more accurate in predicting the landing outcome. The data was downloaded from the Wikipedia page and it was re formatted, cleaned and transformed to make it suitable for further analysis. Exploratory data analysis was performed to understands fully the data set and finally predictive analysis was performed.
- The research used four machine learning models to predict the landing outcome and compared the accuracy of each model. The machine learning models used in this research are logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor(KNN).

❖ Summary of all results

- The landing success rate has been improved overtime
- KSC LC-39A has the highest success rate among landing sites
- All models performed similarly on the test set while the decision tree model has a higher accuracy rate.
- Orbits ES-L1, GEO,HEO, and SSO have a 100% success rate

Introduction

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. In this module, you will be provided with an overview of the problem and the tools you need to complete the course.

- Problem statement
- How payload mass, launch site, number flights, and orbits affect first stage landing success
- Rate of successful landings overtime
- Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from the SapceX Rest API (api.spacexdata.com/v4/launches/past)
- Perform data wrangling
  - Handling missing values
  - Dropping unnecessary columns
  - Creating outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The machine learning models used in this research are logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor(KNN).

Data Collection

Using API

Steps

- Request data from Spacex API(rocket launch data)
- Decode response using .json() and convert to a dataframe using.json_normalize.
- Request information about the launched from Spacex API using custom functions
- Create dictionary from the data
- Create dataframe from dictionary
- Filter dataframe to contain only Falcon 9 launches
- Replace missing values of Payload Mass

Web Scraping

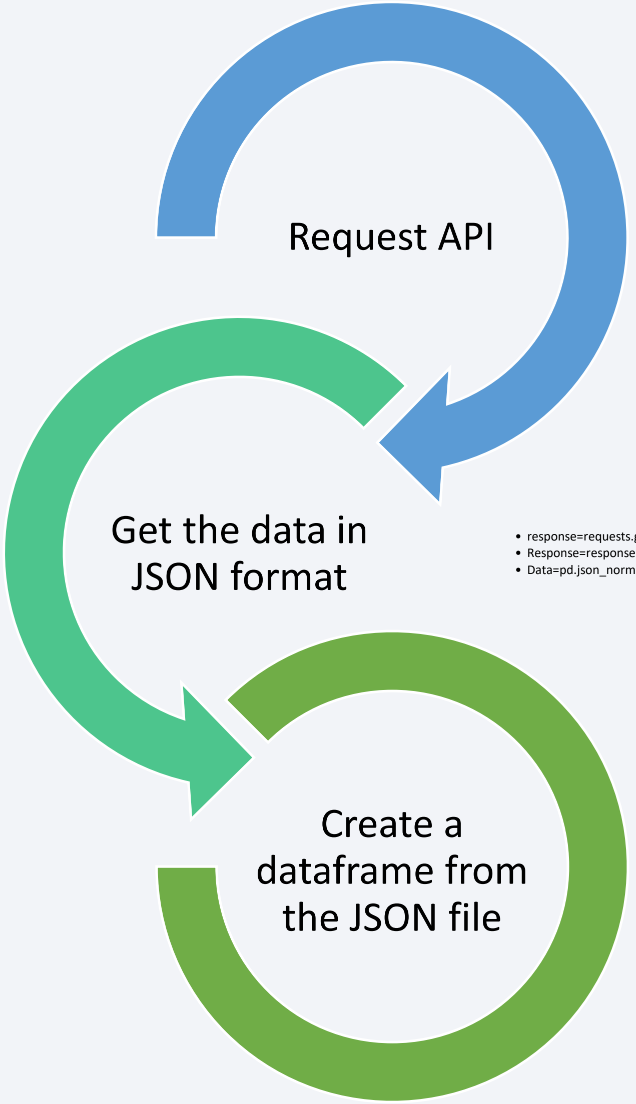
Steps

- Request data (Falcon 9 launch data) from Wikipedia
- Create BeautifulSoup object from HTML response
- Extract column names from HTML table header
- Collect data from parsing HTML tables
- Create dictionary from the data
- Create dataframe from the dictionary

Data Collection – SpaceX API

- Using the steps in the previous slide the procedures of data collection codes

- https://github.com/tsinmegn/Data-science-capstone-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Request API

- Request data from SpaceX API
- requests.get(Spacex url)

Get the data in JSON format

- response=requests.get (static_json_url)
- Response=response.json()
- Data=pd.json_normalize(response)

Create a dataframe from the JSON file

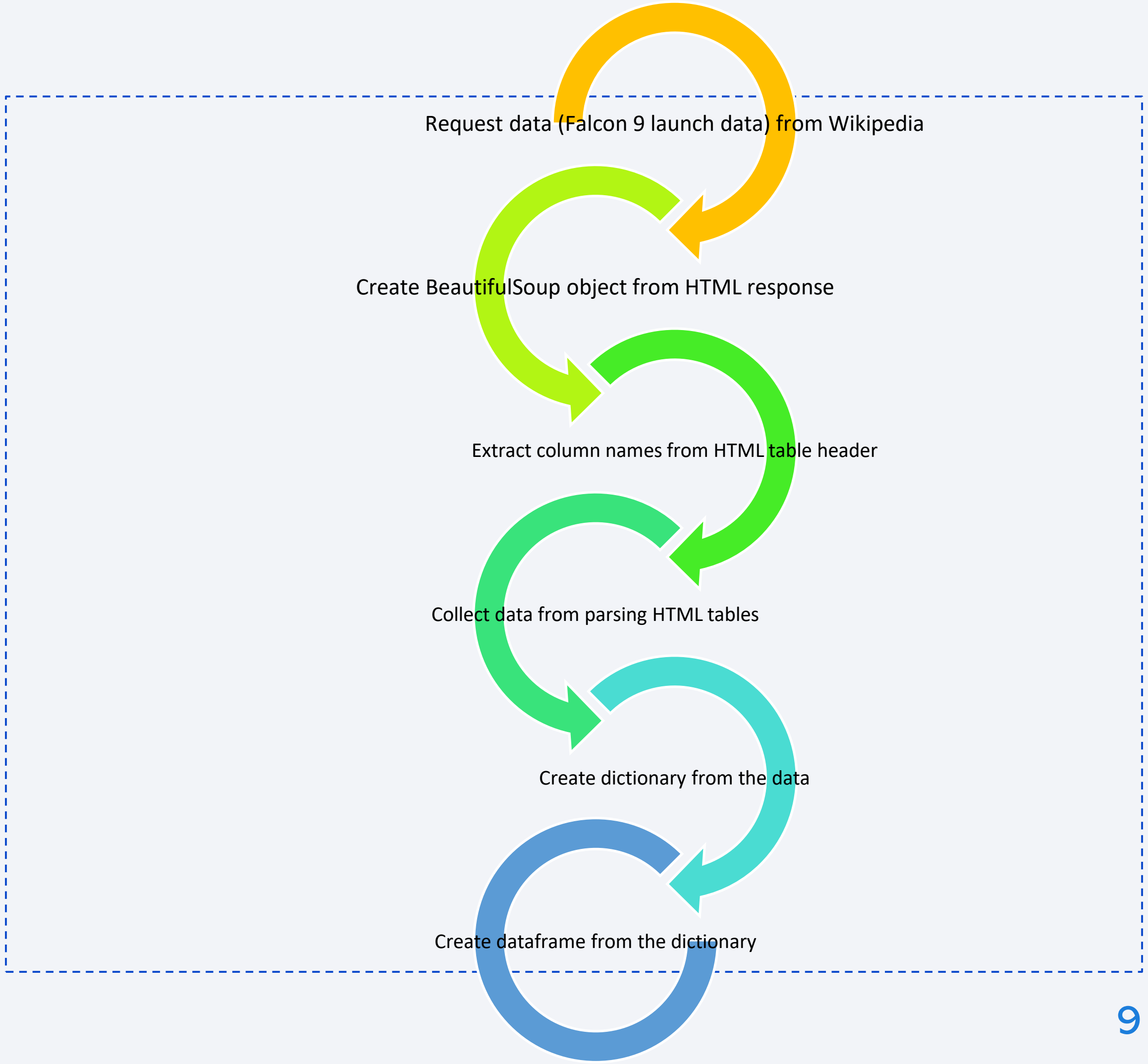
- Df=pd.DataFrame.from_dic(launch_dict)
- Df_falcon9=Df[Df['boosterVersion']!='Falcon_1']

Data Collection - Scraping

Steps

- Request data (Falcon 9 launch data) from Wikipedia
- Create BeautifulSoup object from HTML response
- Extract column names from HTML table header
- Collect data from parsing HTML tables
- Create dictionary from the data
- Create dataframe from the dictionary

- https://github.com/tsinmegn/Data-science-capstone-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Request data (Falcon 9 launch data) from Wikipedia

Create BeautifulSoup object from HTML response

Extract column names from HTML table header

Collect data from parsing HTML tables

Create dictionary from the data

Create dataframe from the dictionary

Data Wrangling

- The first step in any data analysis starts with understanding the data and checking the data quality for completeness, consistency and accuracy through exploratory Data Analysis(EDA).

- EDA involves checking for missing values, identifying the data type for each variable, creating new features, re-naming, deleting or adding columns and rows etc.

- https://github.com/tsinmegn/Data-science-capstone-project

EDA with Data Visualization

- Scatter plots, line charts and bar charts were plotted to explore the data.
- Scatter plots were used to explore the relation ship between two continuous variables. For example we used scatter plot to see how the FlightNumber and Payload variables would affect the launch outcome.
- Bar plots are used to explore the relationship between categorical variables. Bar plots are used to visualize the relationship between launch success rate and orbit type.
- Line charts are used to visualize the variables through time. For example success rates over the years is visualized using the line chart.
- https://github.com/tsinmegn/Data-science-capstone-project

EDA with SQL

- The queries used to solve tasks:
  - Display the names of unique launch sites in the space (%sql select DISTINCT Launch_Site FROM SPACEXTABLE;)
  - Display 5 records where launch sites begin with the string'CCA'(%sql select Launch_Site from SPACEXTABLE WHERE Launch_Site like 'CCA%' limit 5;)
  - Display the total payload mass carried by boosters launched by NASA(CRS) (%sql select sum (PAYLOAD_MASS__KG_) from SPACEXTABLE where CUSTOMER = 'NASA (CRS)')
  - Display Average payload mass carried by booster version F9 v1.1(%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where BOOSTER_VERSION = 'F9 v1.1')
  - List the date when the first succesful landing outcome in ground pad was achieved(%sql select Min(Date) from SPACEXTABLE where Landing_Outcome like 'Success%')
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000(%sql select BOOSTER_VERSION from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000)

Build an Interactive Map with Folium

Markers indicating launch sites

- Added blue circle at NASA Johnson Space Center’s coordinate with a popup label showing its name using its latitude and longitude coordinates
- Added red circles at all launch sites coordinates with a popup label showing its name

Colored markers of Launch Outcomes

- Added colored markers of successful and unsuccessful launches at each launch site to show which launch site have higher success rates.

Distance between a launch site to proximities

- Added colored lines to show distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline,railway, highway, and city.

Github link to the Notebook

- https://github.com/tsinmegn/Data-science-capstone-project

Build a Dashboard with Plotly Dash

Dropdown list with launch sites

- Allow selection of launch sites or a specific launch site

Pie chart showing successful launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range

- Allow user to select payload mass range

Scatter chart showing payload mass vs success rate by Booster version.

- Allow user to see the correlation between Payload and Launch Success

Predictive Analysis (Classification)

- Import the required libraries
- Acquire the data set
- Data preprocessing(EDA and visualization, One Hot Encoding to convert categorical variables to binary variables)
- Splitting the data into training data and test data
- Apply the machine learning model to the training data(Logistic Regression, KNN, Decision Trees, Support vector machines)
- Evaluate the accuracy of the models using (Accuracy Score, Jaccard Index, F1-Score ,LogLoss,R2-Score)
- Compare which models predict the success of landing. In our case Decision tree model has a slightly higher accuracy rate.

- 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

Results

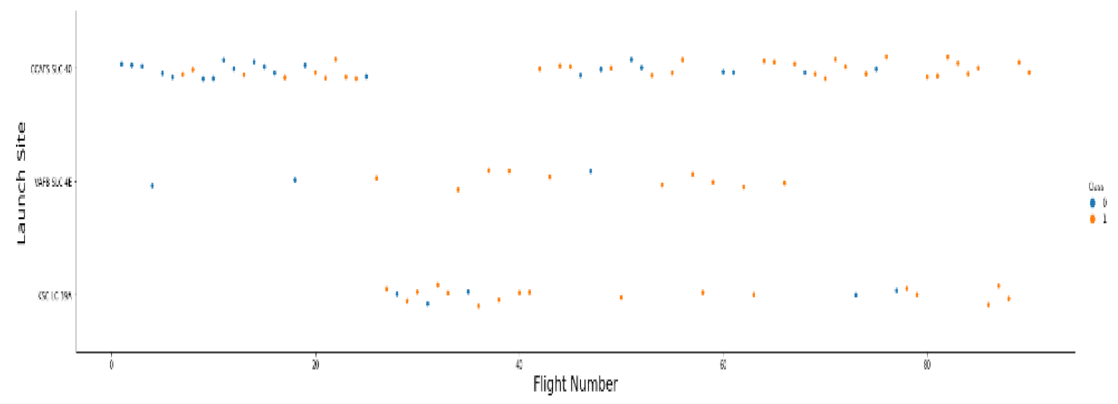
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

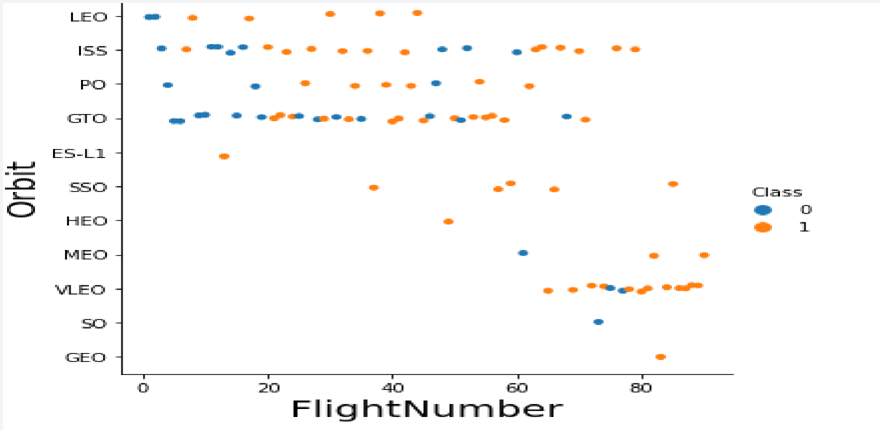
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

A scatter plot of Flight Number vs. Launch Site

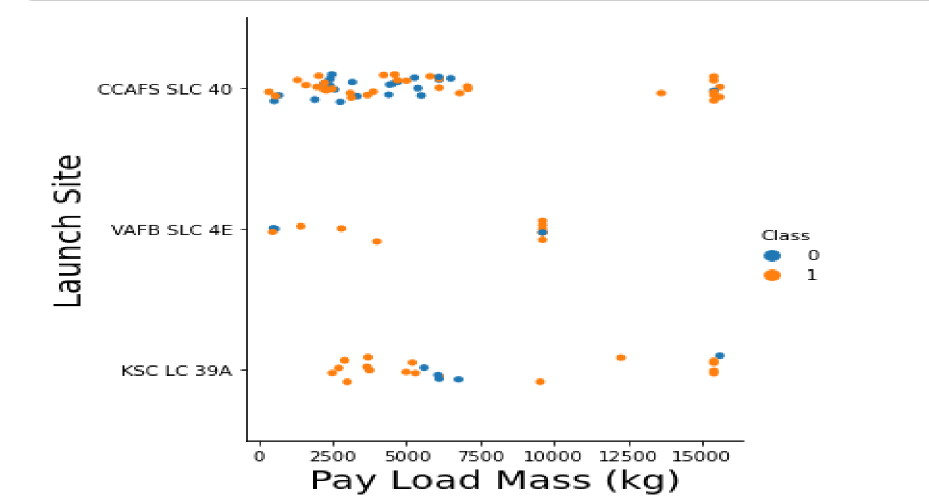


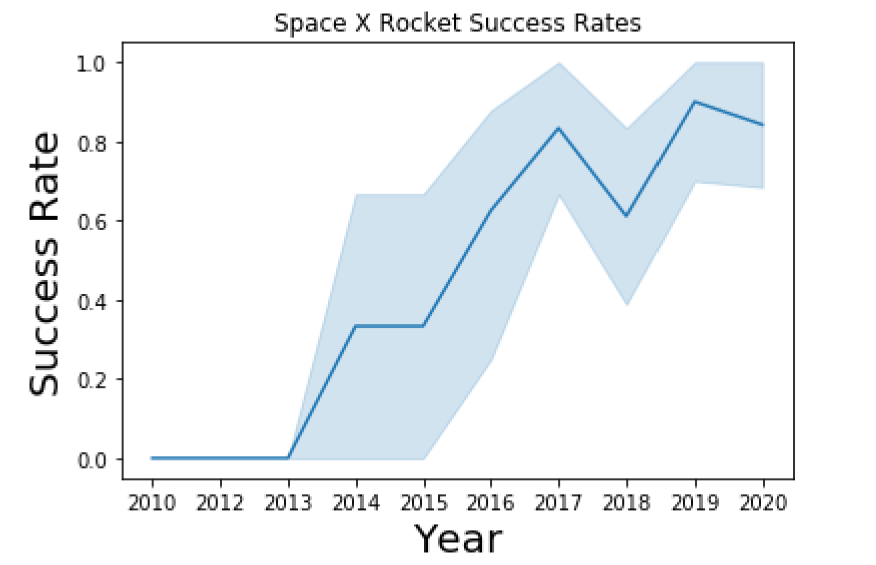


LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

A scatter plot of Orbit type with payload mass

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



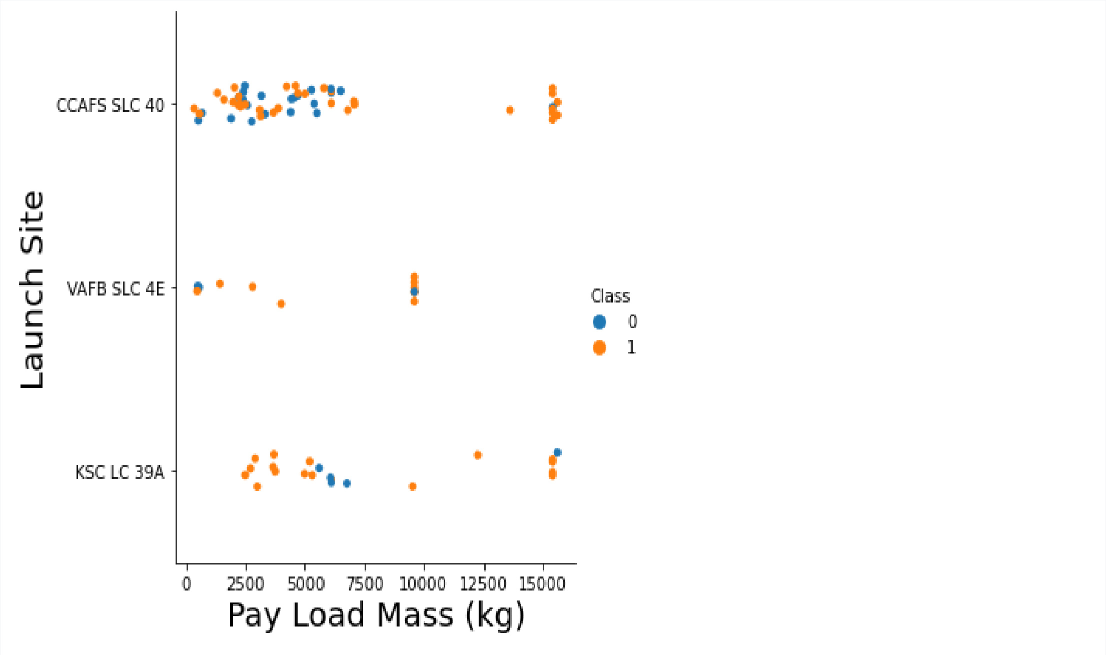


The success rate since 2013 kept increasing till 2020

Payload vs. Launch Site

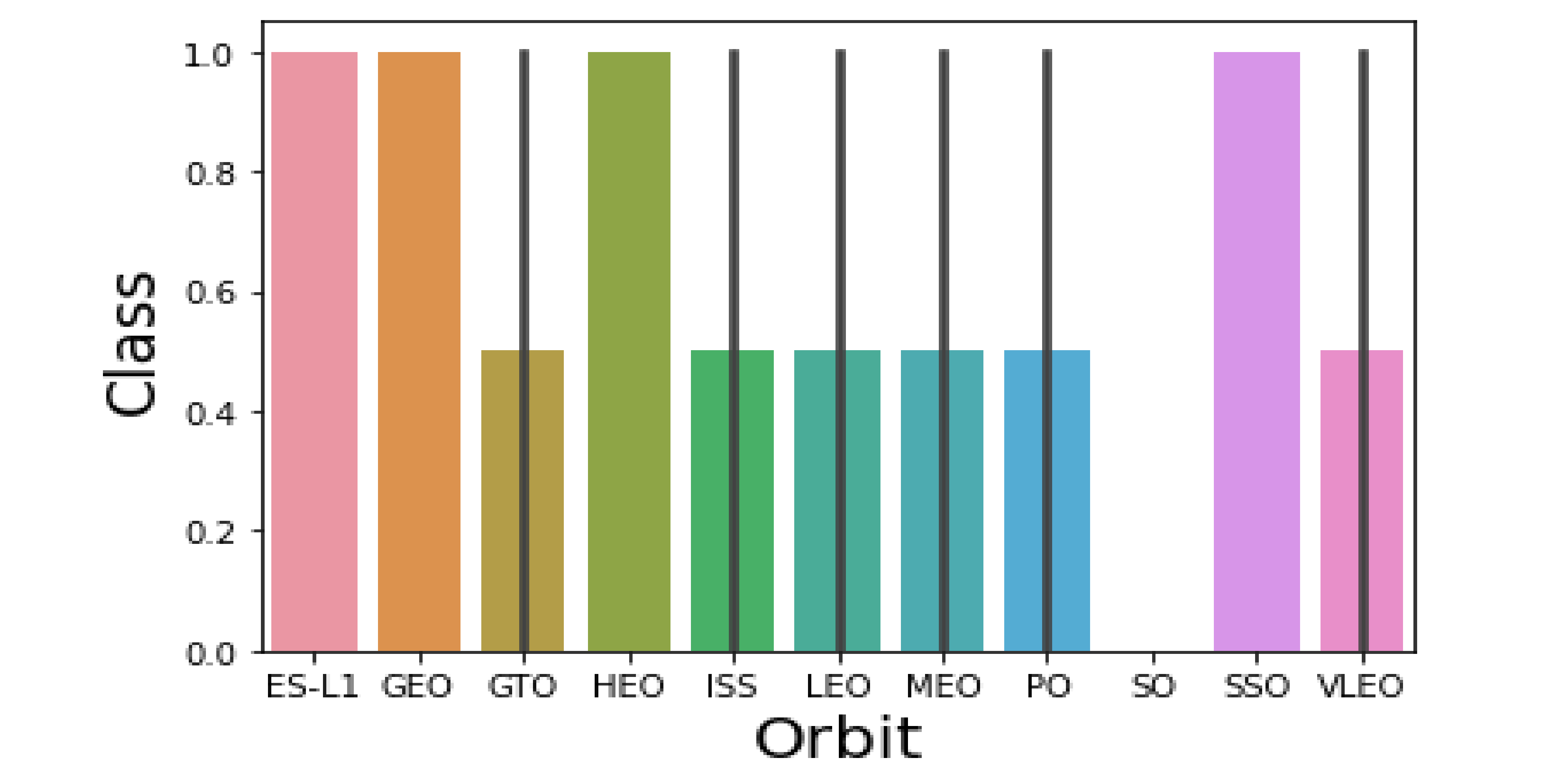
VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).

Payload Vs Launch Site scatter plot



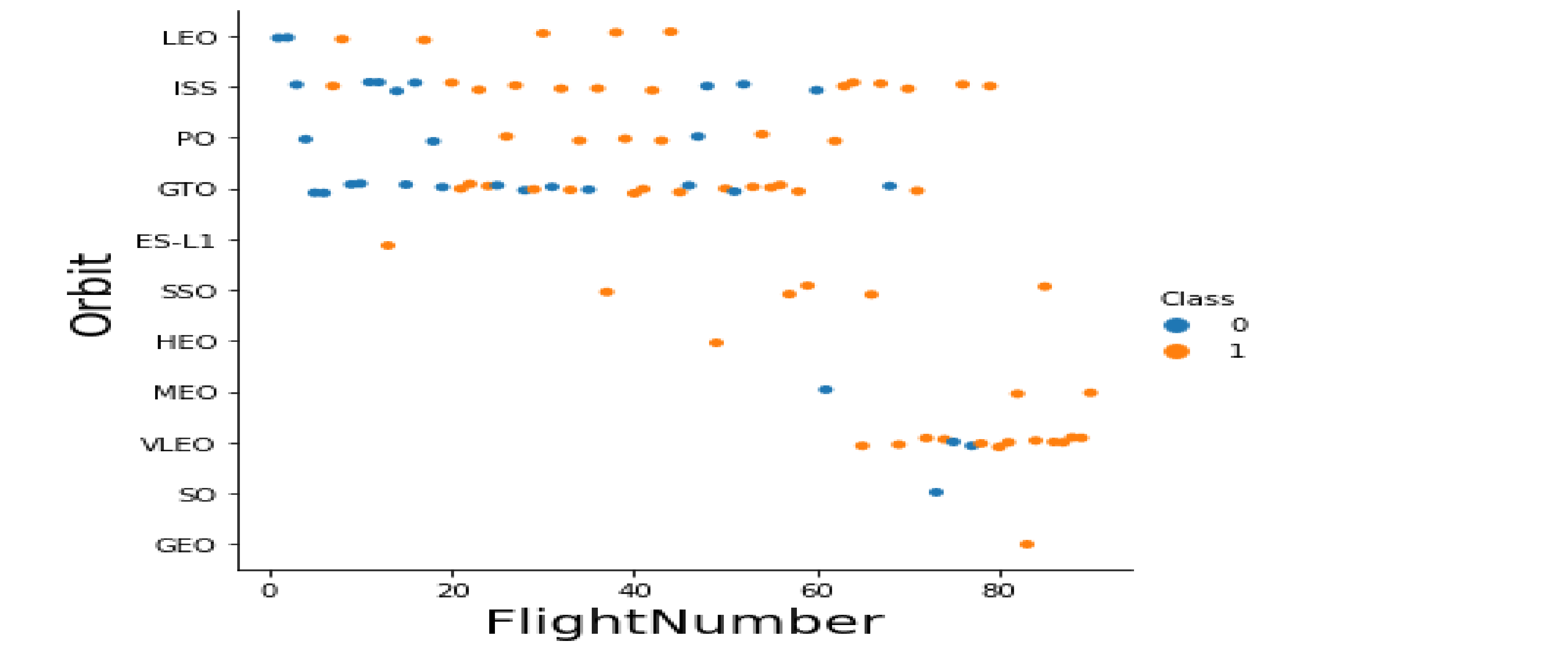
Success Rate vs. Orbit Type

- ES-L1, GEO,HEO and SSO have higher success rate



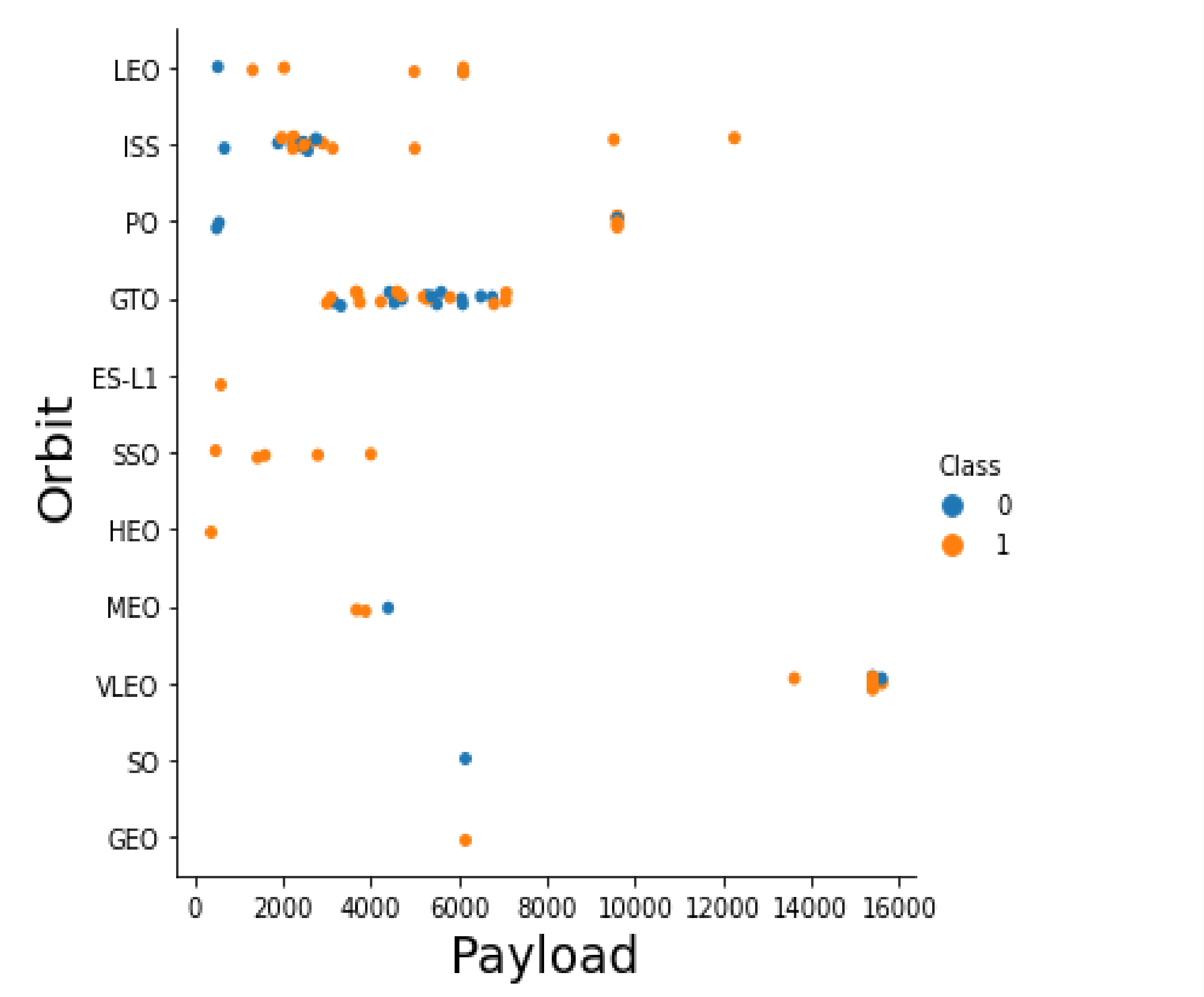
Flight Number vs. Orbit Type

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.



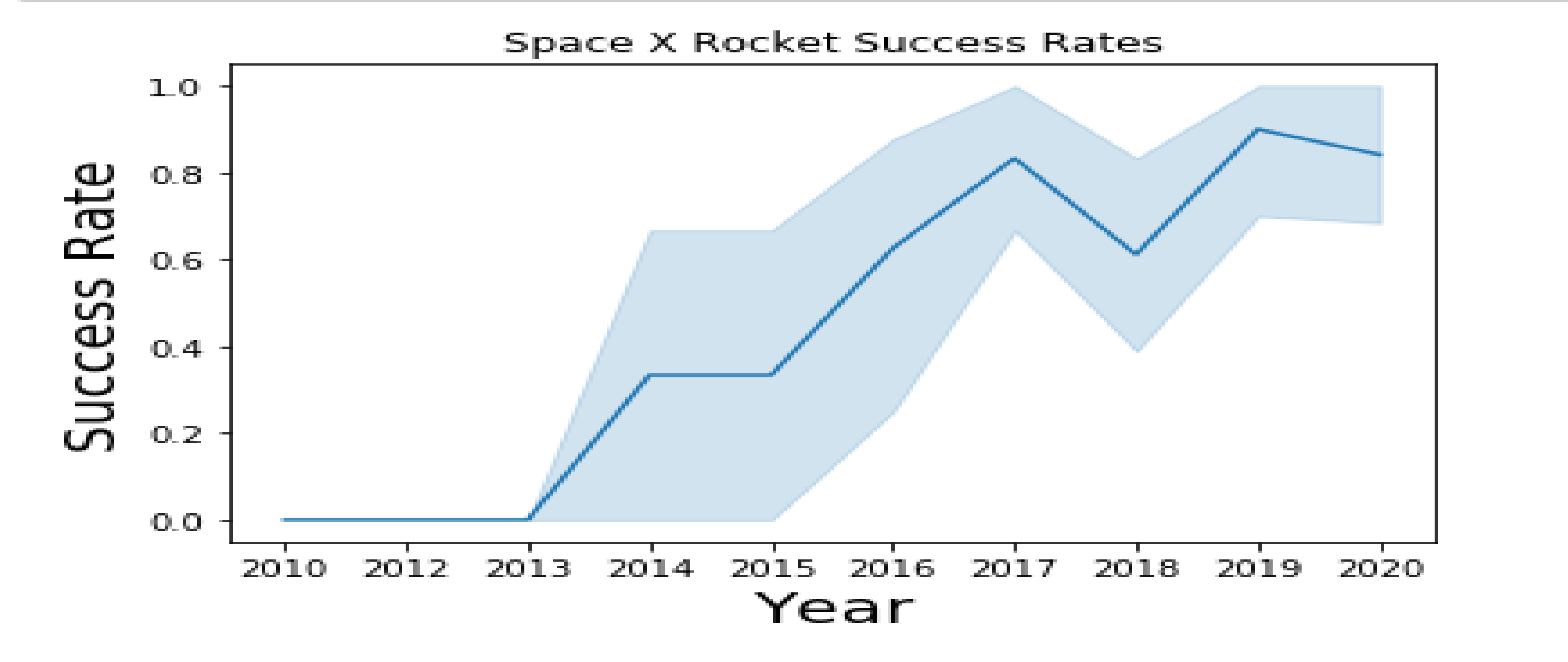
Payload vs. Orbit Type

- 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

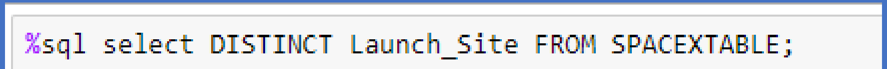
Success rate since 2013 kept increasing till 2020



All Launch Site Names

- The names of the unique launch sites

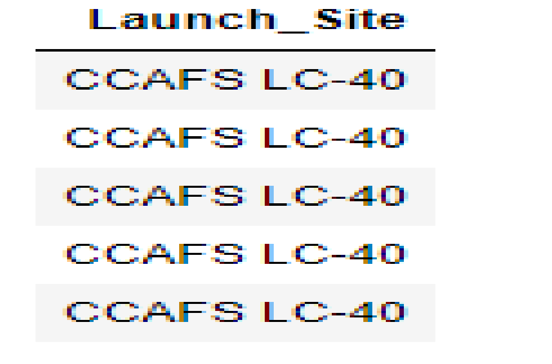
    1. CCAFS LC-40
    2. VAFB SLC-4E
    3. KSC LC-39A
    4. CCAFS SLC-40



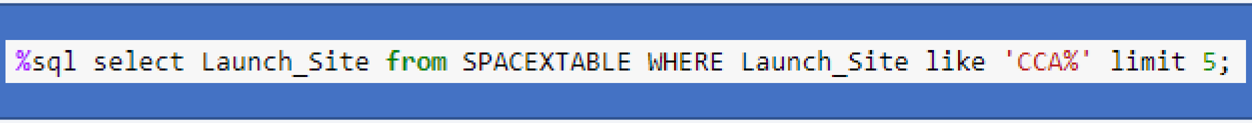
```
%sql select DISTINCT Launch_Site FROM SPACEXTABLE;
```

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`



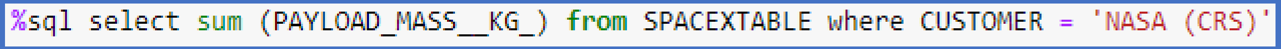
Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40



```
%sql select Launch_Site from SPACEXTABLE WHERE Launch_Site like 'CCA%' limit 5;
```

Total Payload Mass

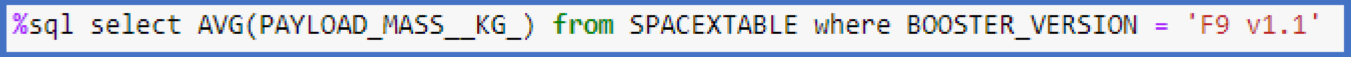
- The total payload carried by boosters from NASA is **45596** Kg.
- Present your query result with a short explanation here



```
%sql select sum (PAYLOAD_MASS__KG_) from SPACEXTABLE where CUSTOMER = 'NASA (CRS)'
```

Average Payload Mass by F9 v1.1

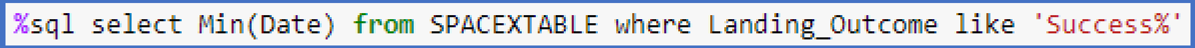
- The average payload mass carried by booster version F9 v1.1is **2928.4Kg**



```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where BOOSTER_VERSION = 'F9 v1.1'
```

First Successful Ground Landing Date

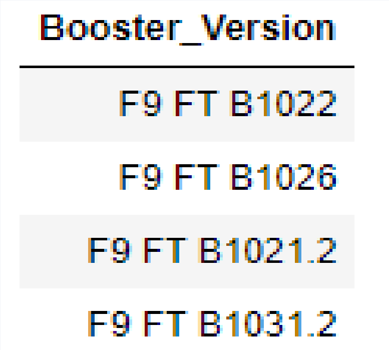
- The first successful landing outcome on ground pad had taken place on 2015-12-22



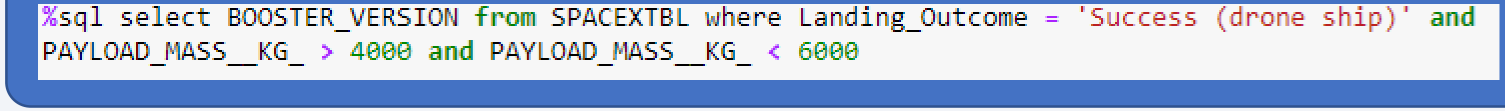
```
%sql select Min(Date) from SPACEXTABLE where Landing_Outcome like 'Success%'
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



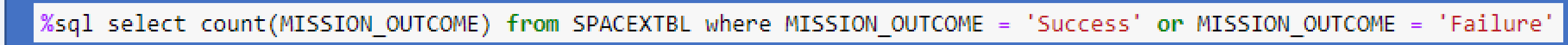
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2



```
%sql select BOOSTER_VERSION from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and
PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

Total Number of Successful and Failure Mission Outcomes

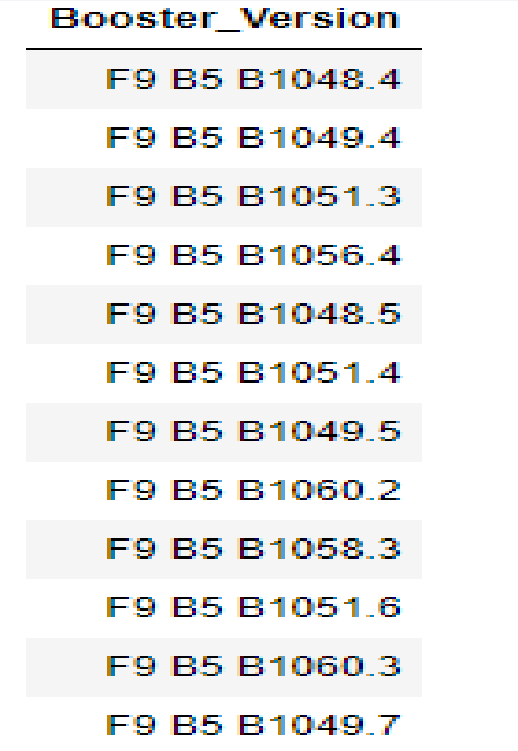
- The total number of successful and failure mission outcomes are 98



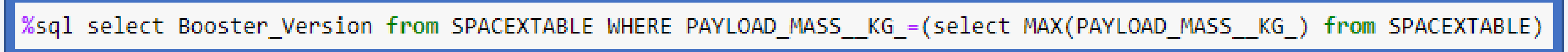
```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure'
```

Boosters Carried Maximum Payload

- Booster which have carried the maximum payload mass



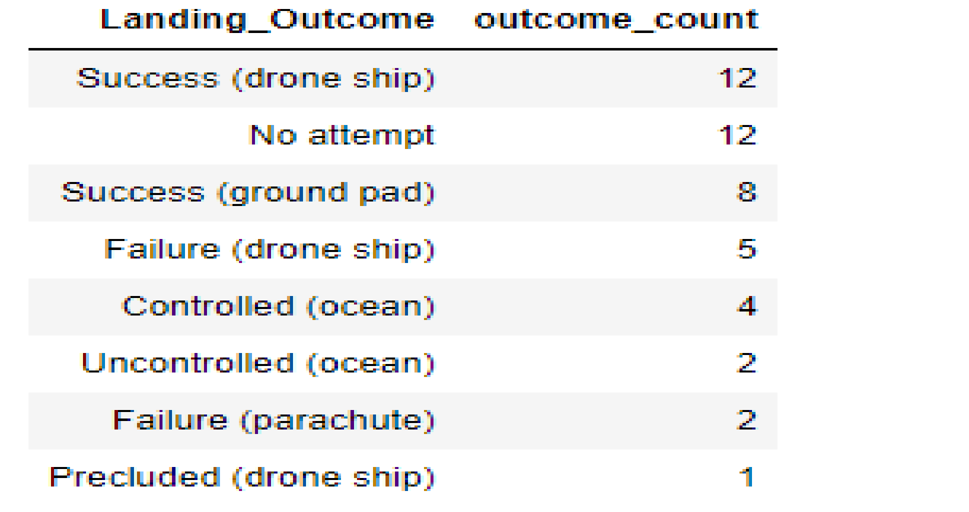
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7



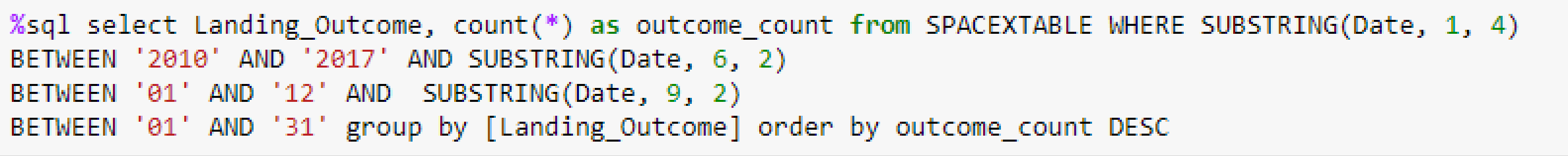
```
%sql select Booster_Version from SPACEXTABLE WHERE PAYLOAD_MASS__KG_=(select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

- 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



Landing_Outcome	outcome_count
Success (drone ship)	12
No attempt	12
Success (ground pad)	8
Failure (drone ship)	5
Controlled (ocean)	4
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1



```
%sql select Landing_Outcome, count(*) as outcome_count from SPACEXTABLE WHERE SUBSTRING(Date, 1, 4)
BETWEEN '2010' AND '2017' AND SUBSTRING(Date, 6, 2)
BETWEEN '01' AND '12' AND  SUBSTRING(Date, 9, 2)
BETWEEN '01' AND '31' group by [Landing_Outcome] order by outcome_count DESC
```

Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- Replace <Folium map screenshot 1> title with an appropriate title

- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map

- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 2>

- Replace <Folium map screenshot 2> title with an appropriate title

- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map

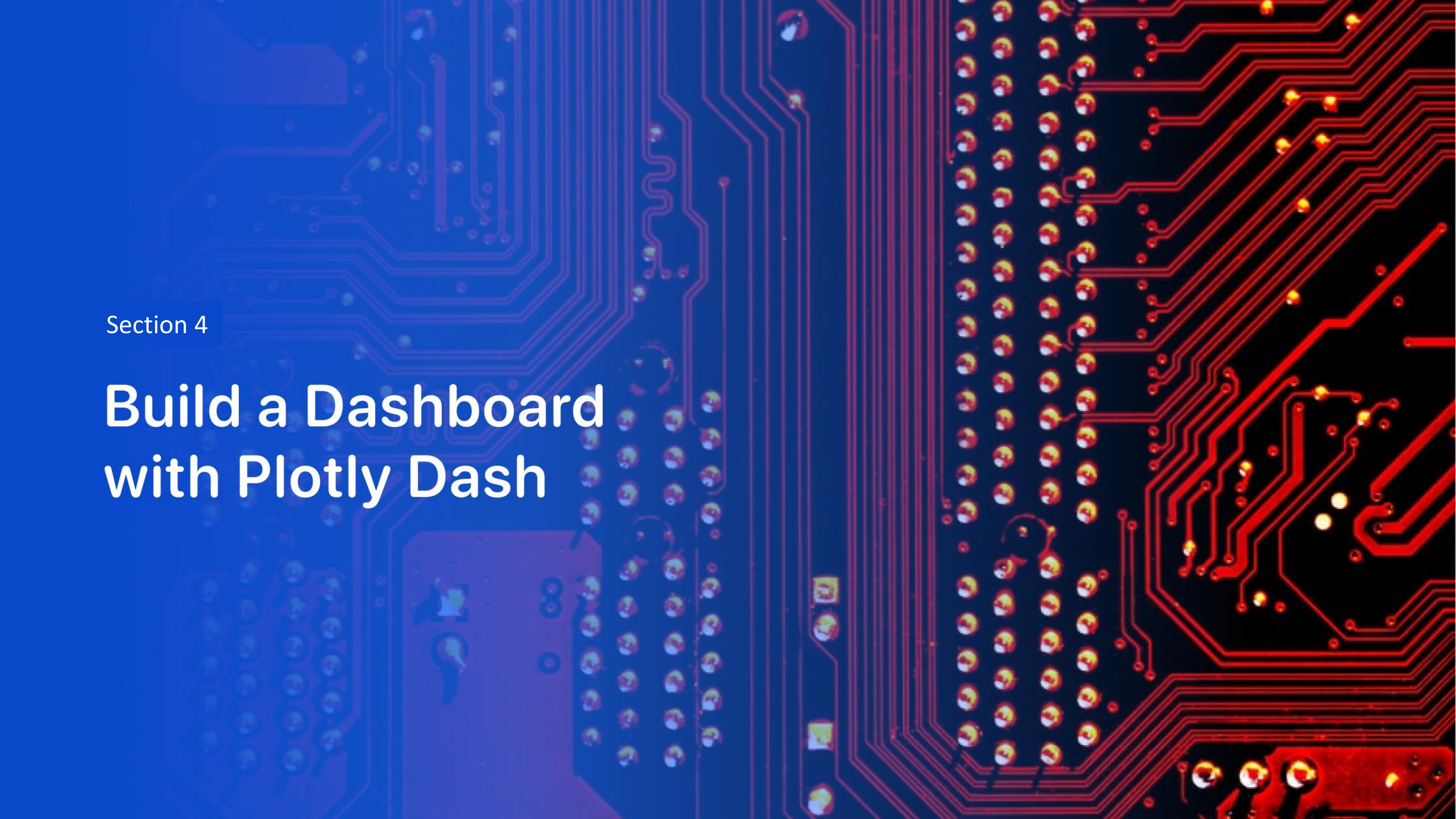
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title

- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

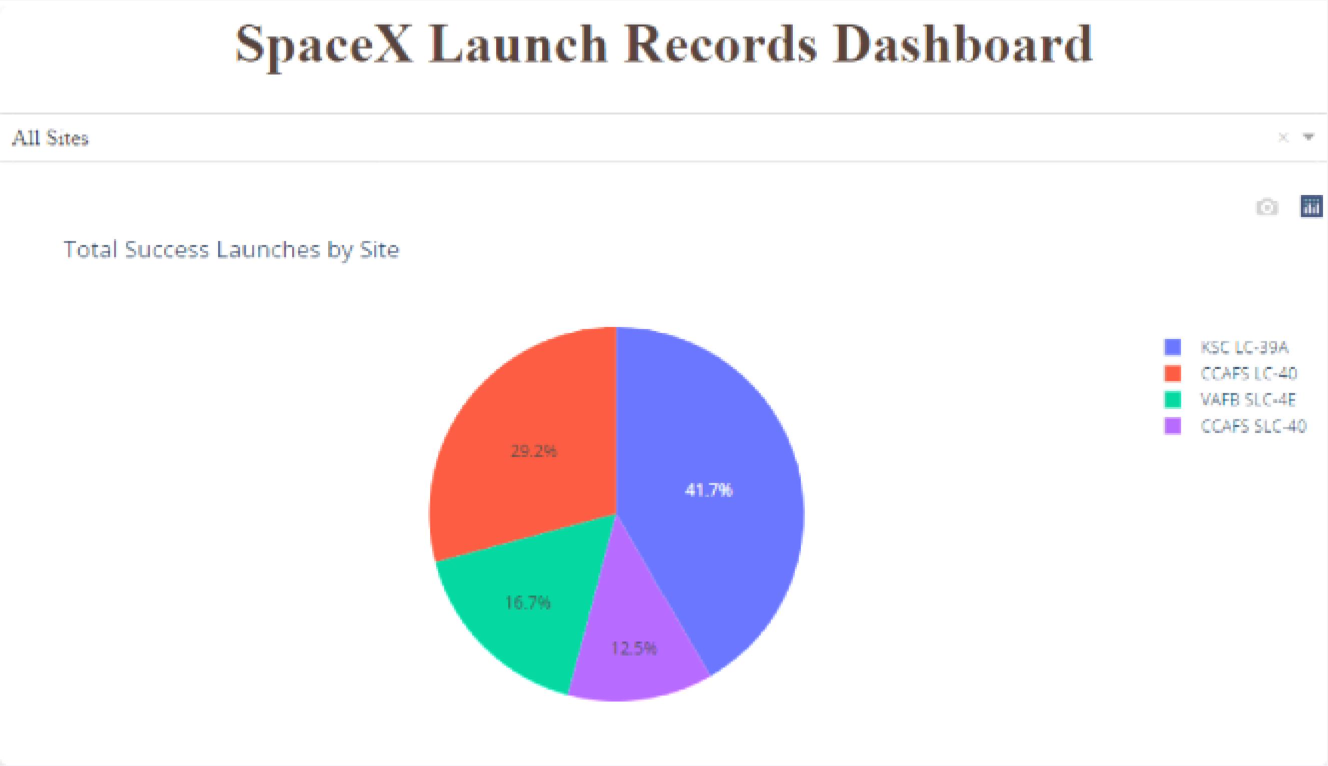
- Explain the important elements and findings on the screenshot



Section 4

Build a Dashboard with Plotly Dash

Launch Success Dashboard



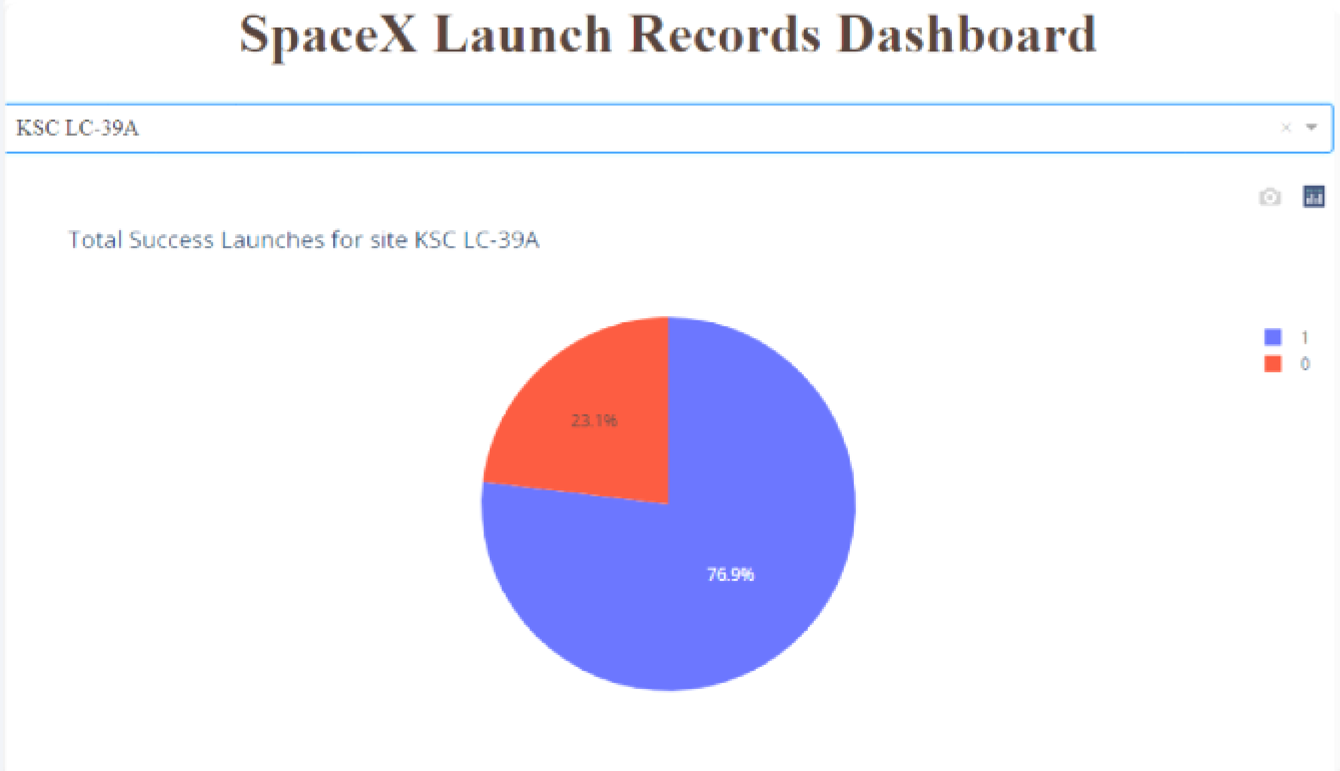
SpaceX Launch Records Dashboard

All Sites

Total Success Launches by Site

- The launch site KSC LC-39A had the most successful launches, with 41.7% of the total successful launches.

The launch site with the highest launch success rate

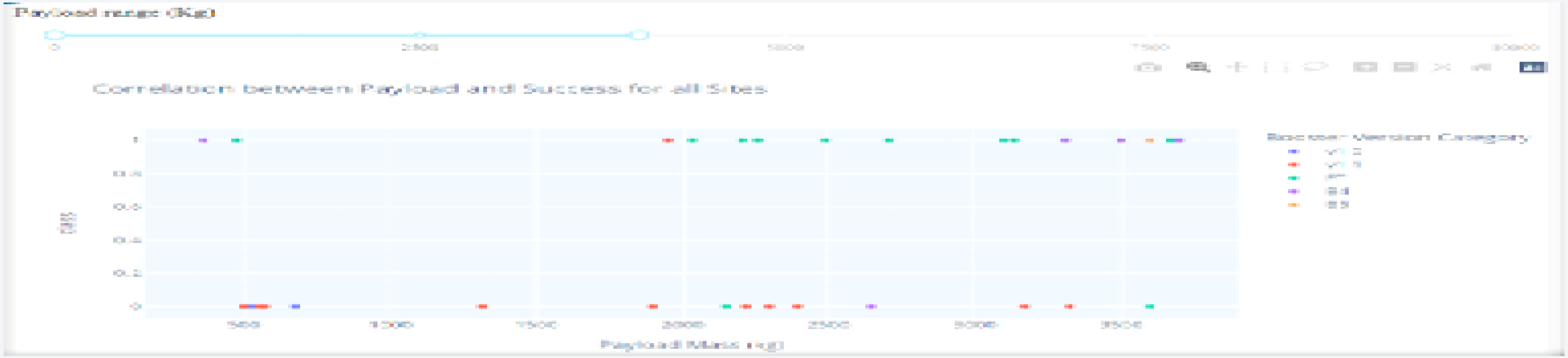


SpaceX Launch Records Dashboard

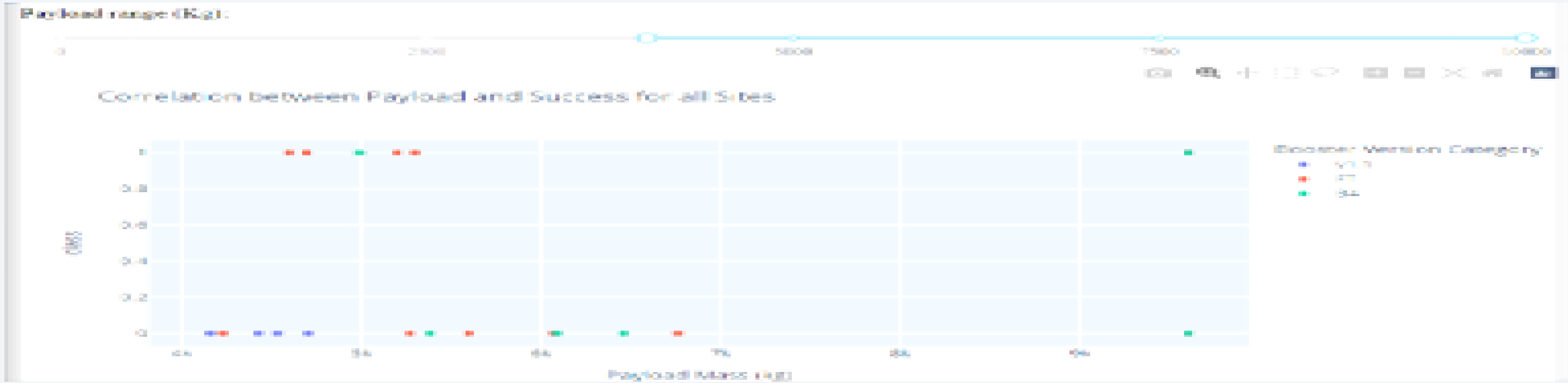
KSC LC-39A

The launch site KSC LC-39A also had the highest rate of successful launches, with a 76.9% success rate.

Launch Outcome vs. Payload Scatter Plot for all sites



Payload range (Kg):

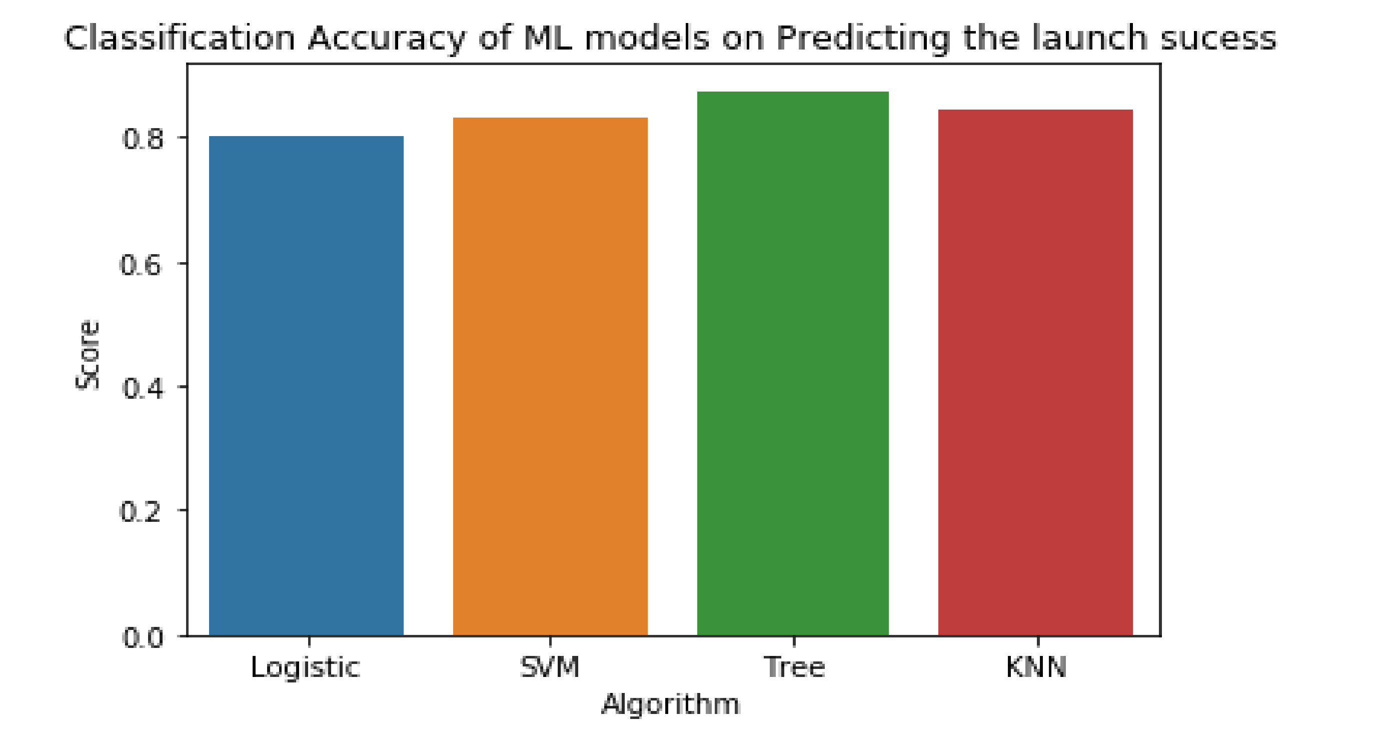


Payload range (Kg):

Section 5

Predictive Analysis (Classification)

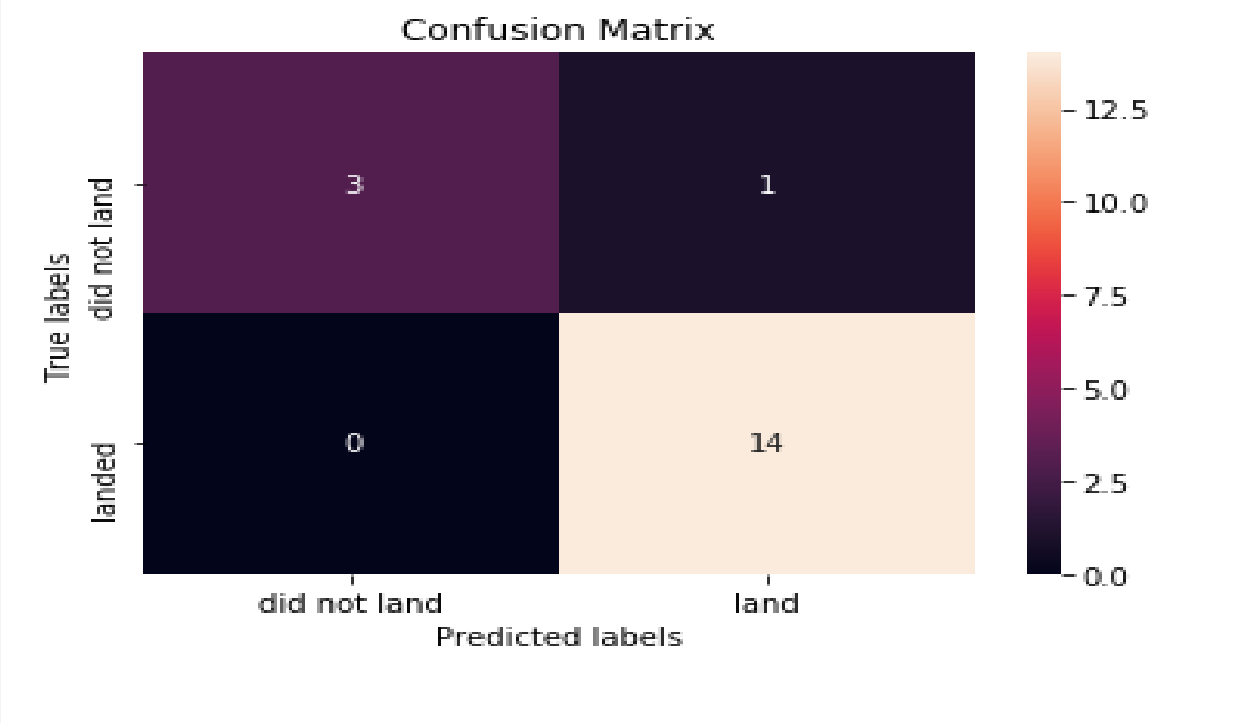
Classification Accuracy



- Decision tree classification model has the highest classification accuracy

Confusion Matrix

Decision tree confusion matrix on predicting launch success



- As can be seen from the confusion matrix only one observation is classified wrongly by the model.

Conclusions

- The best performing model for predicting the launch success in Decision tree with the highest accuracy prediction rate.

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!