IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Celdrick Ndze Kuta
January 22, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory data analysis with data visualization
  - Exploratory data analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive Analytics
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- SpaceX (space exploration technology Corp), an American aerospace manufacturer space transportation services and communications corporation.
- SpaceX Falcon 9 rocket launches are relatively inexpensive ($62million) compared to other competitors (who spend around $165million) 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 project seeks to answer the following:
  - Will SpaceX reuse the first stage i.e Will the first stage successfully land?
  - What will be the price of a launch ?
  - What are the determinants for a successful launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Launch data requested from SpaceX website
  - Web scraping Falcon 9 Launch data from related Wiki pages
- Perform data wrangling
  - Landing outcomes converted to two classes 0 (unsuccessful landing) or 1(successful landing)
  - Filtering only Falcon 9 data from all launches that also contained Falcon 1 launches
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - New target column named ‘Class’ , split data into train ,test sets , validation set for choosing best parameters to train four different models and then evaluate models on test set.

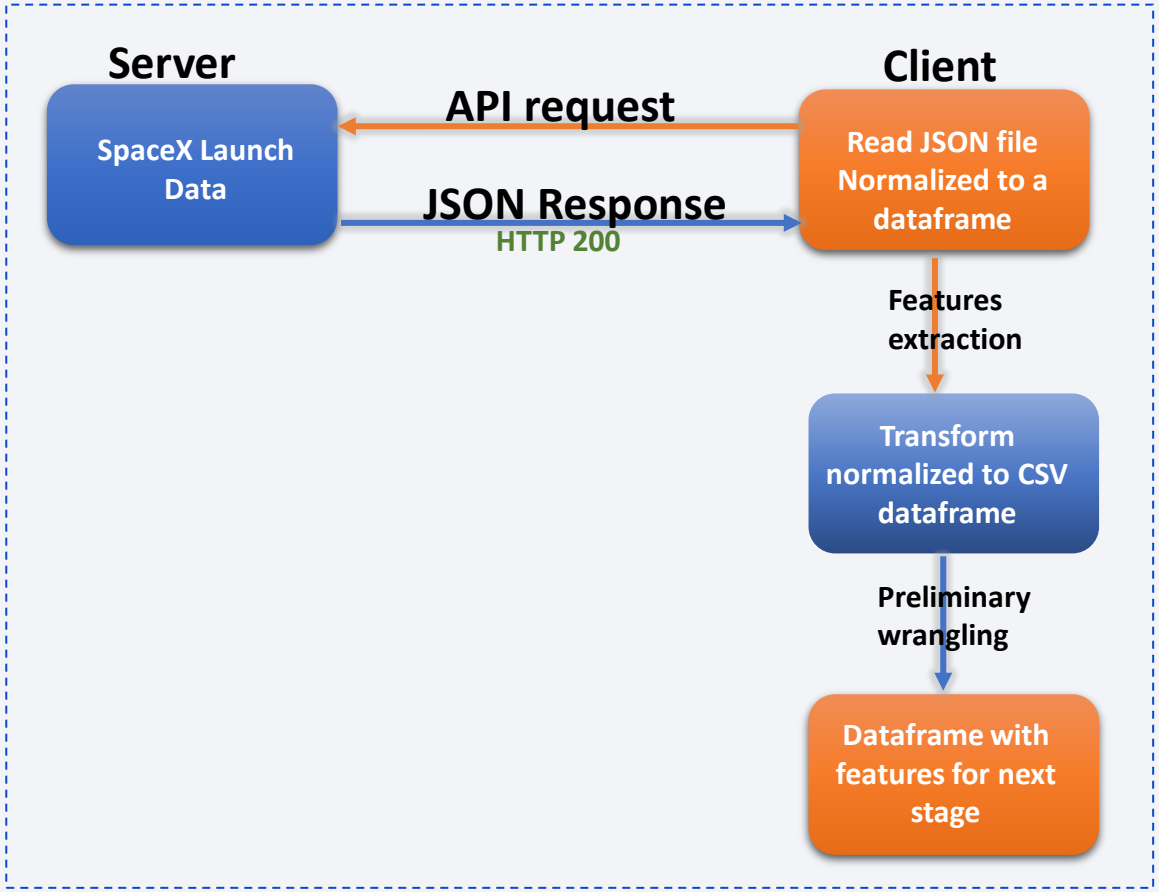
Data Collection

- Data collection is a process of collecting data from multiple sources. Rocket launch data was obtained from the spaceX website and from multiple wiki pages.
- Hypertext Transfer Protocol (HTTP) is the protocol in which information is transmitted (fetched) via web browsers over internet usually in html format.
- Rest API is a set of functions that helps two application to communicated over the internet using HTTP framework. In essence it delivers users response to the system and sends system response to the user usually in a JSON format.
- JavaScript Object Notation (JSON) is an open standard file format that uses human-readable text to store and transmit data objects consisting of attributes, value pairs and arrays.

Data Collection – SpaceX API

- Request past rocket launch data from SpaceX API
- Extract attributes from the API using their identification numbers and append them in a list
- Normalize the JSON files to CSV files
- Construct the data set by combining all attributes into a table
- Filter table only to include Falcon 9 data

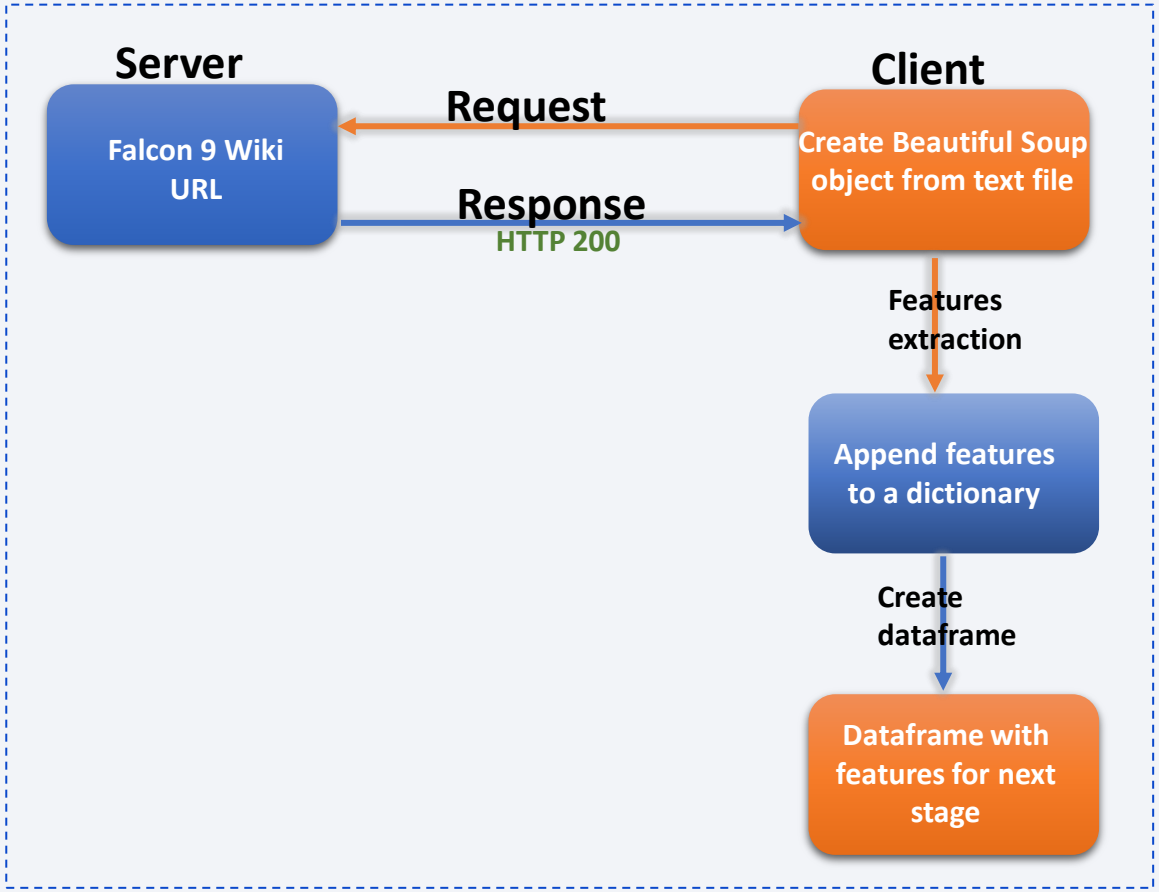
GitHub



Data Collection - Scraping

- Request Falcon 9 launch wiki page as HTML file
- Create Beautiful Soup object from file
- Extract all columns and variable names from the file
- Create data frame by parsing launch HTML tables

GitHub



Data Wrangling

- Data Analysis of features;
  - Check for missing values and data types
  - Calculate the number of launches on each site(each launch aim to an orbit)
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome per orbit type and label them bad_outcomes if mission outcome was unsuccessfully landed.
- Create new feature in dataframe called ‘Class’ from outcome feature with labels 0 for bad_outcomes and 1 if mission successfully landed in a specific region.

GitHub

EDA with Data Visualization

- Exploratory Data Analysis (EDA) help give more inside on the relationship between the features. The following charts where plotted with outcome overlay
  1. **Scatter plot**: Explains the correlation between two numerical features making patterns easy to observe
     - The following variables where plotted to visualize their relationships.
       - Flight Number and Payload Mass with outcome overlay
       - Flight Number and Launch Site with outcome overlay
       - Payload Mass and Launch site with outcome overlay
       - Flight Number and Orbit type with outcome overlay
       - Payload Mass and Orbit type with outcome overlay
  2. **Bar chart:** Use to explain categorical variables , comparing group importance by visualizing length of bars. The length of the bars are proportional to values represented by the variable.
     - The following bar chart was plotted to visualize variable relationship
       - Relationship between success rate and each orbit type
  3. **Line chart:** Explains how a variable changes with time. Depicts trends with time.
     - Following line plot was plotted to visualize trends with over time
       - Success rate and year to get average launch success trend.

GitHub

EDA with SQL

- The following queries where performed on an IBM DB2 instance in cloud to further understand the data.
  - Names of the unique launch sites in the space mission
  - Five records where launch site begins with the string ‘CCA’
  - The total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - The date where successful landing outcome in ground pad was achieved
  - Names of boosters with success in drone ship and payload mass greater than 4000 and less than 6000
  - Total number of successful and failure mission outcomes
  - Name of booster version with highest payload mass
  - Records which display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015
  - Rank the count of successful landing outcomes between the date of 04-06-2010 and 20-03-2017 in descending order

GitHub

Build an Interactive Map with Folium

- Folium map is a tool to visualize geospatial data. For instance, good understanding of location and launch site proximities, finding an optimal position for launch sites.
- The following map objects where added on the map:
  - A folium circle and marker to mark all launch site on map this helps to visualize locate launch sites on the map
  - A folium marker cluster to mark the success/failed launches for each site on the map to help in visually enhancing the outcome for each site.
  - A mouse position and a polyline to visualize the distances between a launch site to its proximities
- The Reason for the folium object was to answer the followings
  - Are launch sites in proximity to the Equator line? YES
  - Are all launch sites in very close proximity to the coastline? YES
  - Are launch sites in close proximity to railways, highway? YES
  - Do launch sites keep certain distance away from cities? YES

GitHub

Build a Dashboard with Plotly Dash

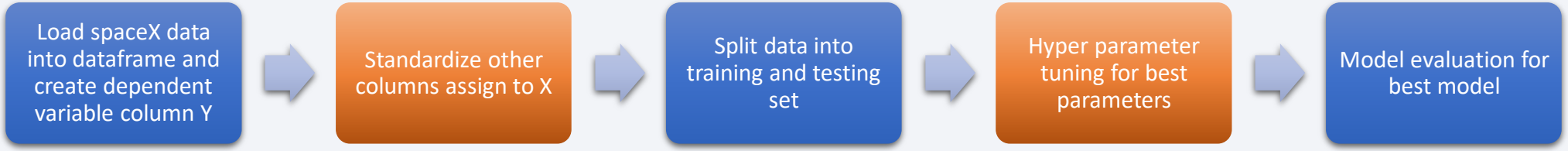
- Plotly Dash helps to provide interactive visual analytics on spaceX launch data in real time.
- Following plots and interactions where added.
  - A launch site drop down input component to help filter dashboard visuals for all four or a particular launch site
  - A callback function to render success-pie-chart based on selected launch site dropdown to help render a pie chart visualization of launch success counts
  - A range slider to select a payload range to help identify visual patterns/ correlation between payload to mission outcome
  - A callback function to render the success-payload-scatter-chart scatter plot to visualize how payload is correlated with mission outcomes for selected site(s)
- The plots and interactions helps in answering the following question
  - Which site has the largest successful launches? KSC LC-39A with 10
  - Which site has the highest launch success rate? KSC LC-39A with 76.9%
  - Which payload range(s) has the highest launch success rate? 2000-5000kg
  - Which payload range(s) has the lowest launch success rate? 0-2000 and 5500-7000
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc) has the highest launch success rate? FT

GitHub Link spaceX_dash_app hosted

Predictive Analysis (Classification)

- Predictive Analysis is a way to help identify future outcomes based on historical data. In order to achieve this, the following Classification machine learning pipeline was optimal for 4 algorithms used(SVM, Logistic regression, KNN, Decision Tree).
  - Load the spaceX data into a data frame and create a column for the dependent variable Y
  - Standardize the remaining columns and then assign them to independent variable X
  - Split the data into training (for model building) and testing (for model testing) sets.
  - Improve model by finding best hyperparameter using Grid Search cross validation
  - Model evaluation by comparing the model accuracy to a threshold on training and test data set.
  - The best classification model obtain by comparing all model accuracy on training and testing set with the best performing algorithm been the Decision tree with model accuracy of 0.88%

GitHub Link

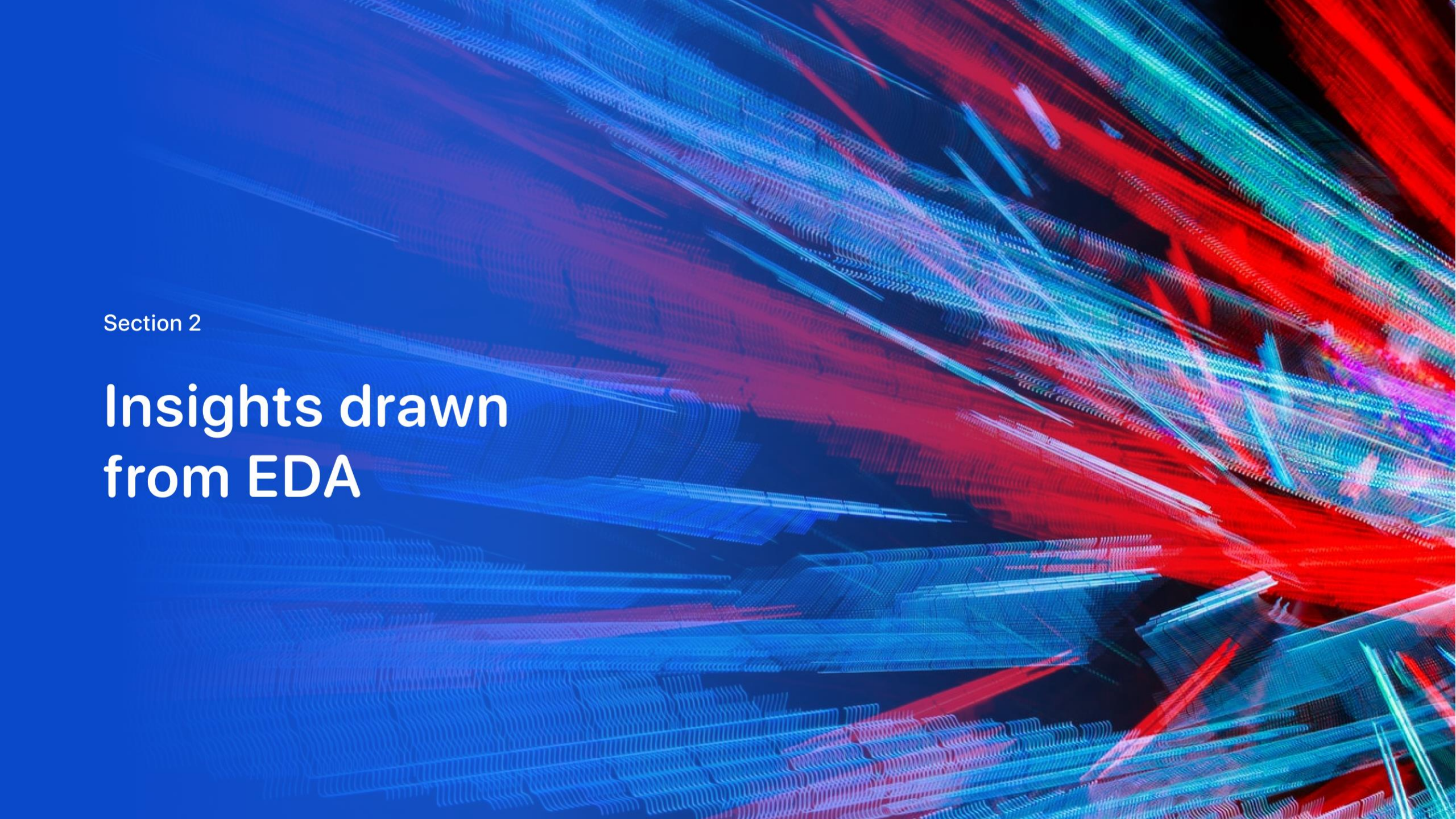


Results

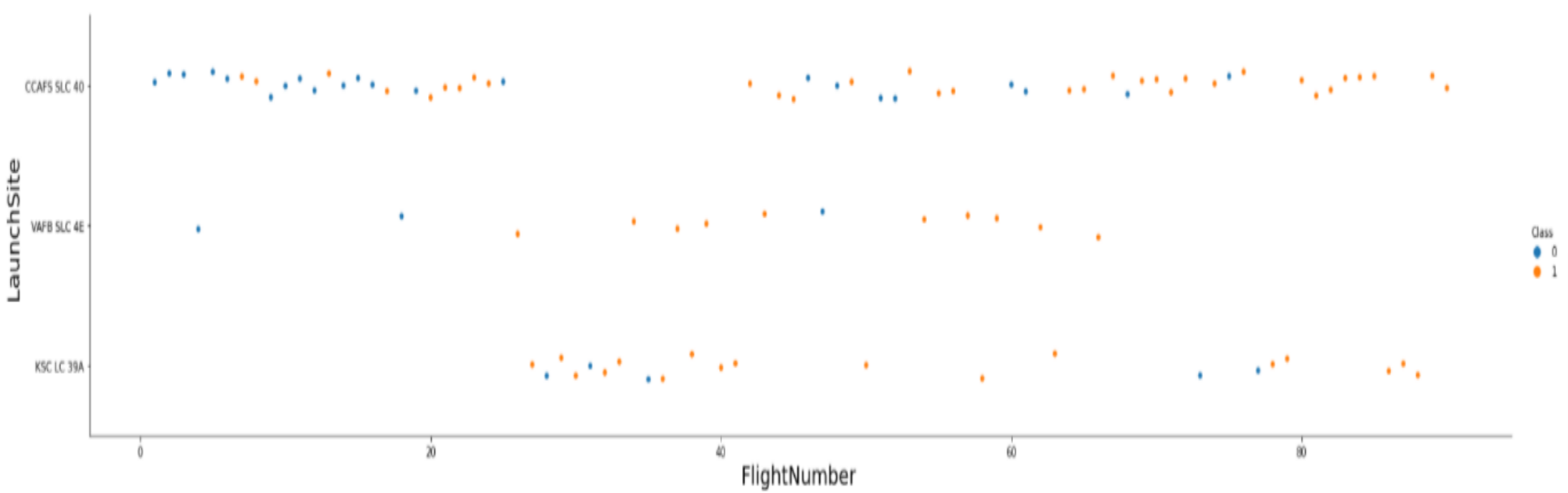
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

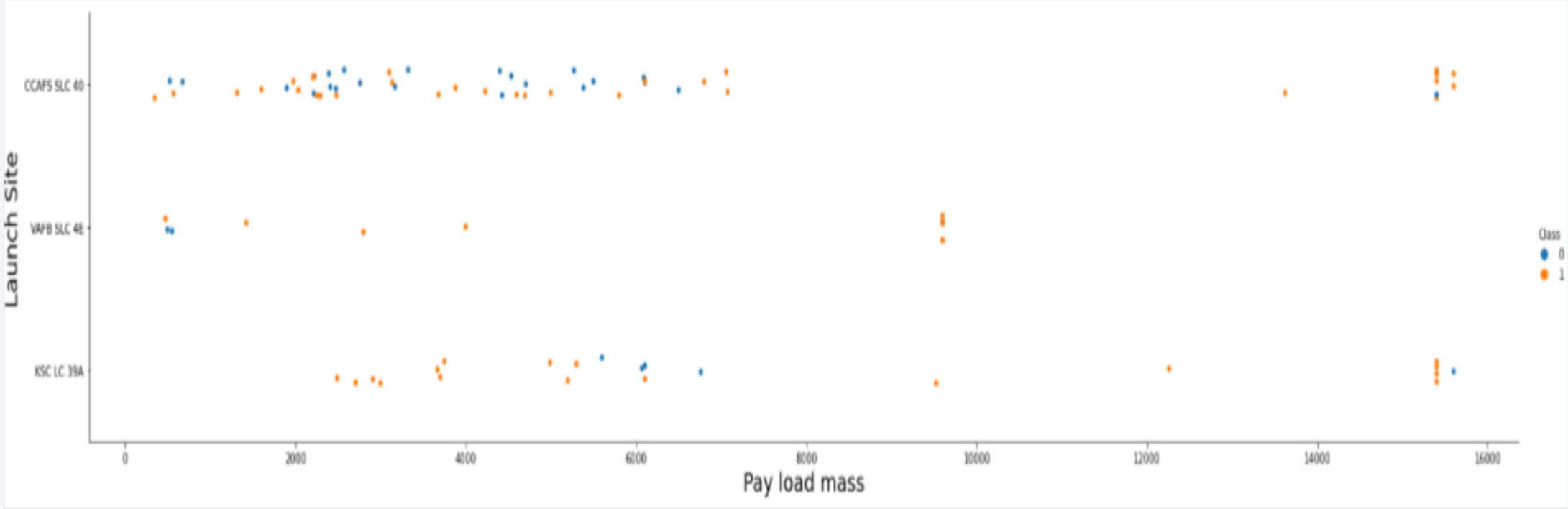


Flight Number vs. Launch Site



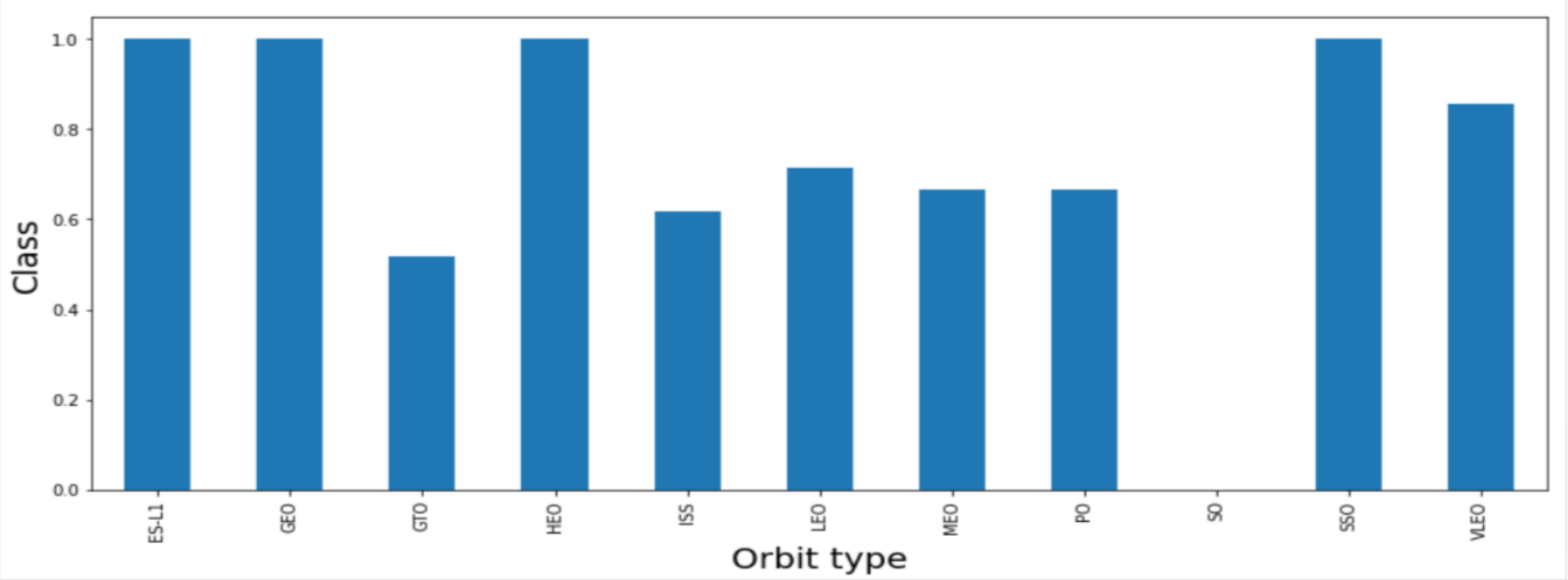
- The more the number of lunches on a site, the greater the chance for a successful landing (class =1) in first launch
- For launch site ‘KSC LC 39A’ it takes at least 25 launches before the first successful launch

Payload vs. Launch Site



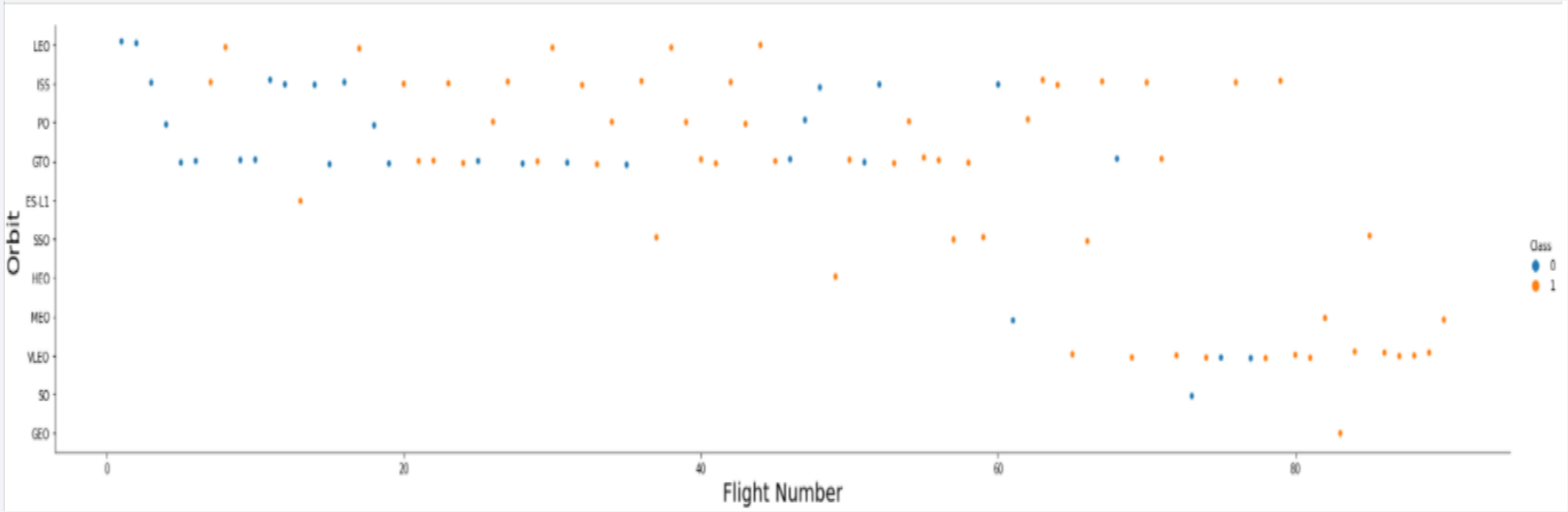
- For launch site ‘VAFB SLC 4E’ there are no launches for loads heavier than 10000kg
- For launch site ‘CCAFS SLC 40’, success rate increases for greater payload mass
- There seem to be no clear correlation or pattern between Launch sites and pay load mass

Success Rate vs. Orbit Type



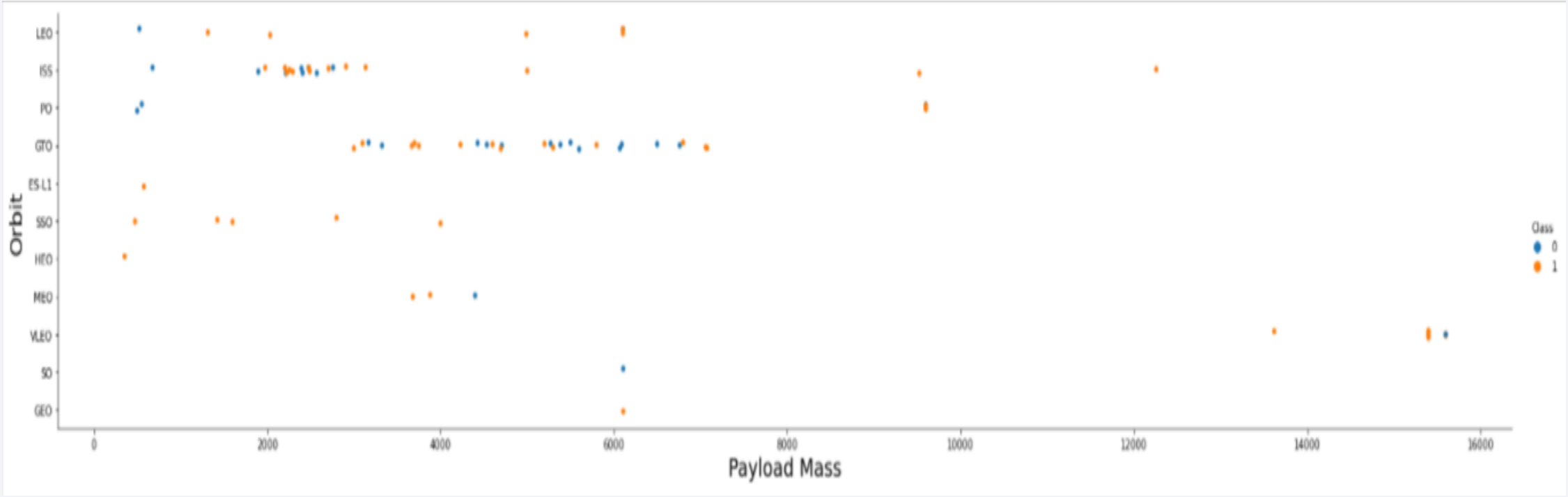
- Orbits ES-L1, GEO, HEO, SSO has highest success rate. First Launches in these orbits will be successful
- GTO has the lowest success rate
- SO no information given

Flight Number vs. Orbit Type



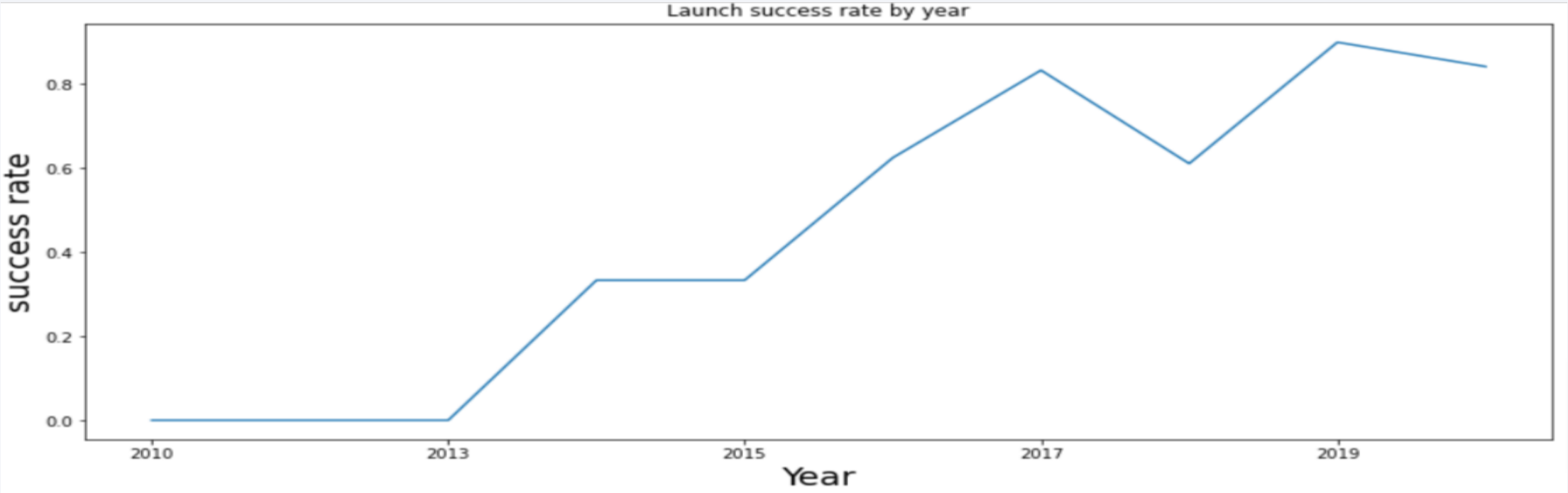
- GEO has one and only one landing only after 80+ launches and is successful.
- SO has no successful landing for all launches
- For the most part LEO, ISS, PO, SSO, MEO, VLO has increase successful landing rate as flight number increases
- No correlation between Flight number and GTO orbit

Payload vs. Orbit Type



- With heavy payloads mass, successful landing rate are more for LEO, ISS, and PO
- GTO, MEO does not really have a clear correlation with payload mass
- High success rate for SSO for masses less that 4000kg

Launch Success Yearly Trend



- Success rate gets better with time since 2013 to 2020
- Very high success rate for year 2019

All Launch Site Names

- **SQL Query unique launch sites names**

```
select DISTINCT(Launch_Site) from SPACEXTBL
```

- **Results**

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

- **Explanation**
  - The DISTINCT clause helps return only unique launch sites from the spacextbl table in the database.

Launch Site Names Begin with 'CCA'

- **SQL Query 5 records**

```
%sql  select * from SPACEXTBL \
      where Launch_Site LIKE 'CCA%' LIMIT 5
```

- **Result**

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

- **Explanation**
  - %sql a sequel magic command helps write the code in multiple lines, specify each line with backslash \
  - The LIKE limits the search criteria to only strings that begins with CCA and LIMIT only the first five search

Total Payload Mass

- **SQL Query by boosters from NASA**

```
%sql select Customer, SUM(PAYLOAD_MASS_KG_) AS 'Payload_Mass' \
        from SPACEXTBL where Customer = 'NASA (CRS)'
```

- **Result**

Customer	Payload_Mass
NASA (CRS)	45596

- **Explanation**
  - SUM function returns the sum in payload mass column with Payload_Mass using AS for alias
  - WHERE clause specifies the criteria for NASA (CRS) customers only

Average Payload Mass by F9 v1.1

- **SQL Query booster version F9 v1.1**

```
%%sql select Booster_Version, AVG(PAYLOAD_MASS_KG) as 'Payload Mass'
    from SPACEXTBL
        where Booster_Version = 'F9 v1.1'
```

- **Result**

Booster_Version	Payload Mass
F9 v1.1	2928.4

- **Explanation**
  - AVG functions calculates the average in the PAYLOAD_MASS_KG column
  - WHERE clause specifies the filtering criteria to Booster_Version F9 v1.1

First Successful Ground Landing Date

- **SQL Query ground pad**

```
%%sql select Date from SPACEXTBL
    where “Landing _Outcome”=‘Success (round pad)’ LIMIT 1
```

- **Result**

Date
22-12-2015

- **Explanation**
  - %%sql just like %sql but no backslash \
  - WHERE clause specifies the filtering criteria to Landing _Outcome of Success (round pad)
  - LIMIT 1 returns only the first result

Successful Drone Ship Landing with Payload between 4000 and 6000

- **SQL Query**

```
%%sql select Booster_Version from SPACEXTBL
 where “Landing _Outcome” =‘Success (drone ship)’
 AND (PAYLOAD_MASS_KG_ >4000 AND PAYLOAD_MASS_KG_ <6000)
```

- **Result**

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

- **Explanation**
  - AND clause specifies all PAYLOAD_MASS_KG whose masses lies between 4000 and 6000 and whose landing outcome are successful drone ships.

Total Number of Successful and Failure Mission Outcomes

- **SQL Query**

```
%%sql select (select count(Mission_Outcome) from SPACEXTBL where
Mission_Outcome LIKE '%Success%') as 'Successful_Outcomes',
(select count(Mission_Outcome) from SPACEXTBL where
Mission_Outcome LIKE  '%Failure%' as 'Failure_Outcomes'
```

- **Result**

Successful_Outcomes	Failure_Outcomes
100	1

- **Explanation**
  - Count function counts each row in the spacex table with Mission outcome by the LIKE criteria
  - The wildcard '**%abc%**' specifies any character before and after the string in **abc**

Boosters Carried Maximum Payload

- **SQL Query**

```
%%sql select Booster_Version from SPACEXTBL
    where PAYLOAD_MASS_KG_ =(select MAX(PAYLOAD_MASS_KG_) from
    SPACEXTBL)
```

- **Results**

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

- **Explanation**
  - MAX function returns the maximum Payload mass from the PAYLOAD_MASS_KG column

2015 Launch Records

- **SQL Query**

  %%sql select substr(Date,4,2) as Month, Booster_Version, Launch_Site,

  "Landing _Outcome" from SPACEXTBL where "Landing _Outcome" ='Failure

  (drone ship)' and substr(Date, 7,4) ='2015'

- **Result**

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

- **Explanation**
  - Substr is a substring function the first parameter 'Date' is the specified column, the second parameter or first number '4 or 7' is the position to start from considering special characters and the second number '2 or 4' is the number of strings to filter including special characters.

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

- **SQL Query**

%%sql select “Landing _Outcome”, count(“Landing _Outcome”) as “Successful landing outcomes cnt” from SPACEXTBL where (Date between ’04-06-2010’ and ‘20-03-2017’) and “Landing _Outcome” LIKE ‘Success%’ group by “Landing _Outcome” Order by count(“Landing _Outcome”) DESC

- **Result**

Landing _Outcome	Successful landing outcomes cnt
Success	20
Success(drone ship)	8
Success(ground pad)	6

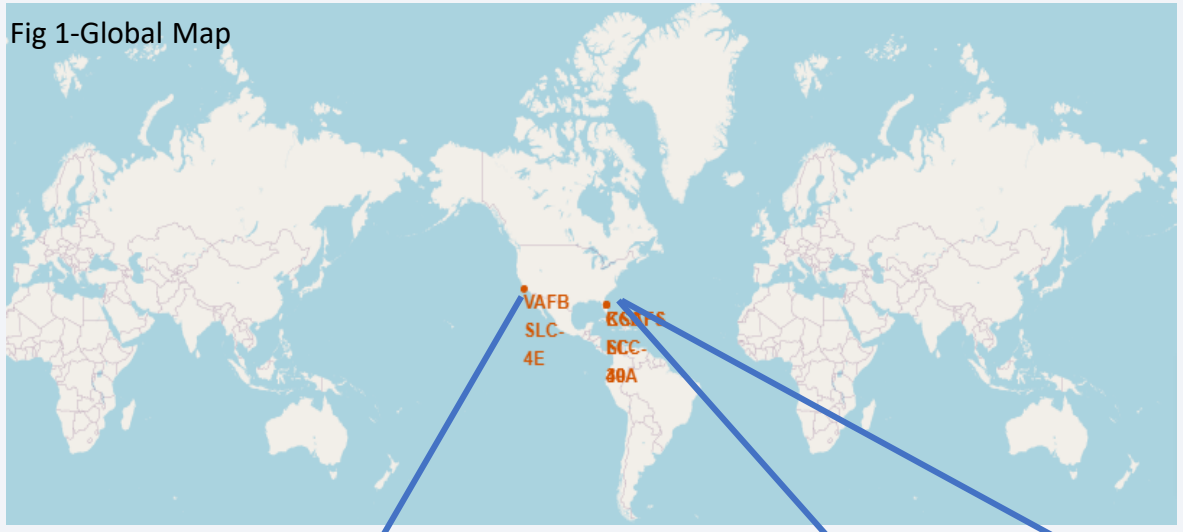
- **Explanation**
  - Between clause use with and clause specifies a range in a period of time inclusively
  - Group by is the grouping criteria while the Order is the arrangement in descending order by DESC

Section 4

Launch Sites Proximities Analysis

Global Map For Launch Sites

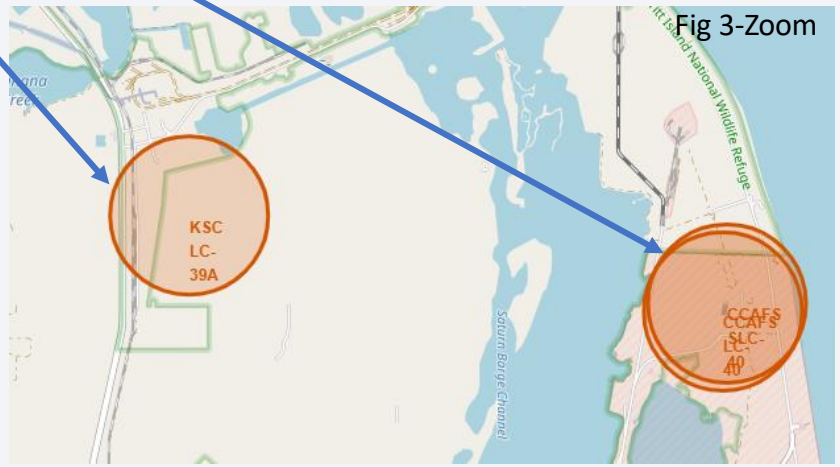
Fig 1-Global Map



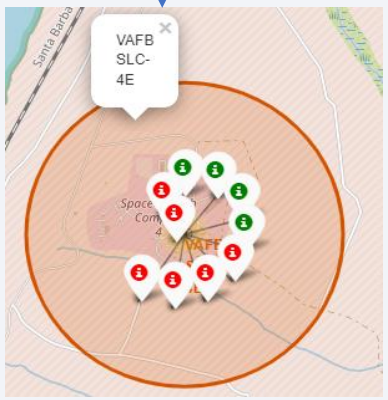
- SpaceX Falcon 9 Launch sites are located in the USA off the coasts of California and Florida
- Fig 2 and Fig3 are zoom from the glabal map Launch sites to see the markers (VAFB SLC-4E in California while KSC LC-39A, CCAFS SLC-40 , CCAFS LC-40 in Florida)

Fig 2-Zoom

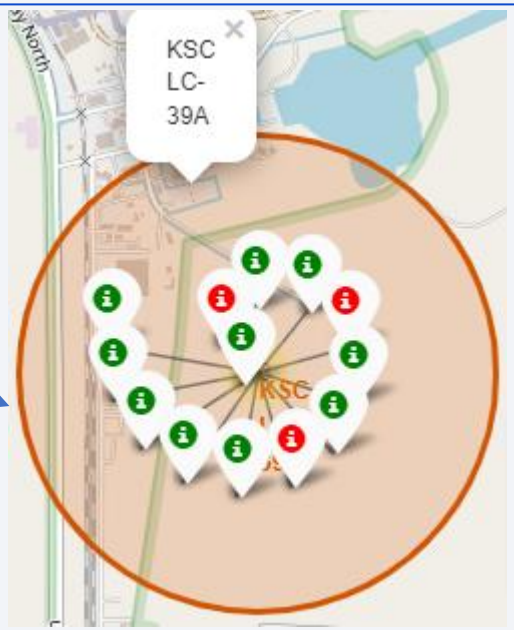
Fig 3-Zoom

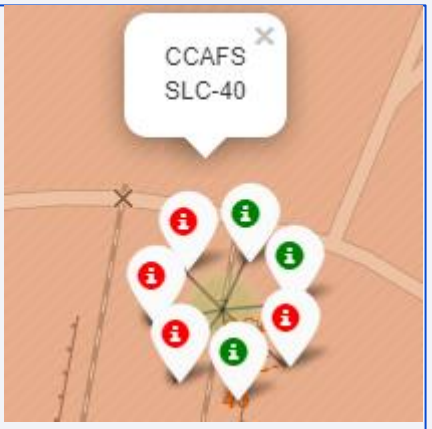


Success/Failed Launches



California Launch site success and Failure rate. The Green markers are For success rate and the Red markers For failure rate on each sites

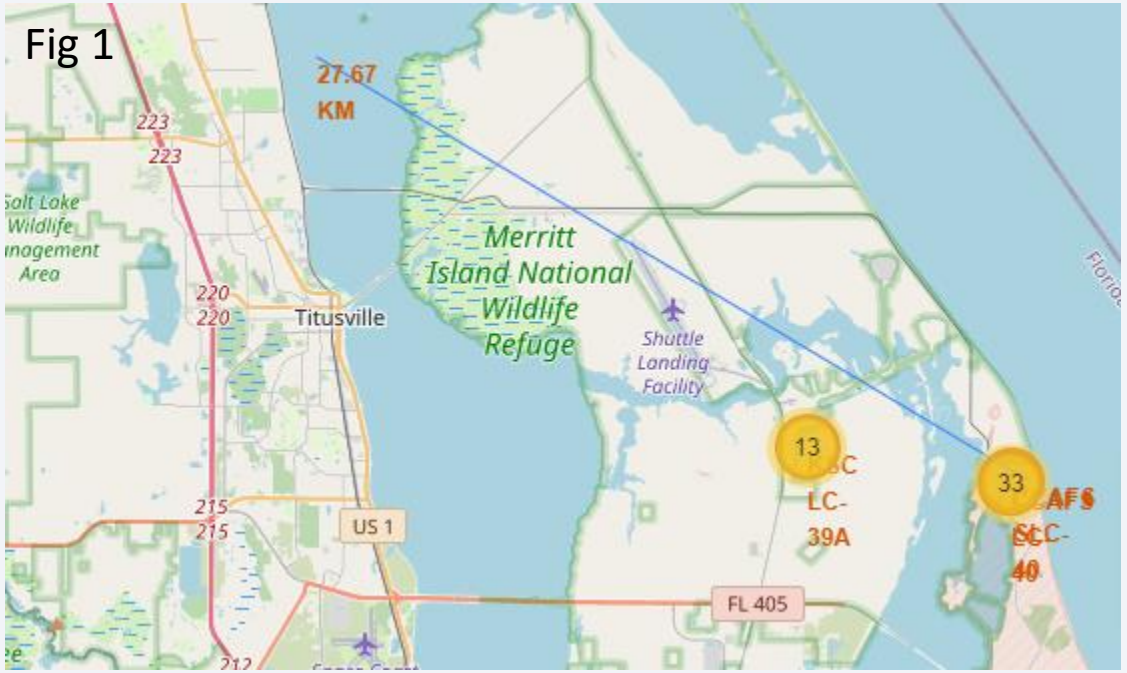




Florida launch sites Success/failure rate.

Distance between sites and proximities

Fig 1



Zoom

The Zoom figure shows the launch site in consideration From the global fig 1.

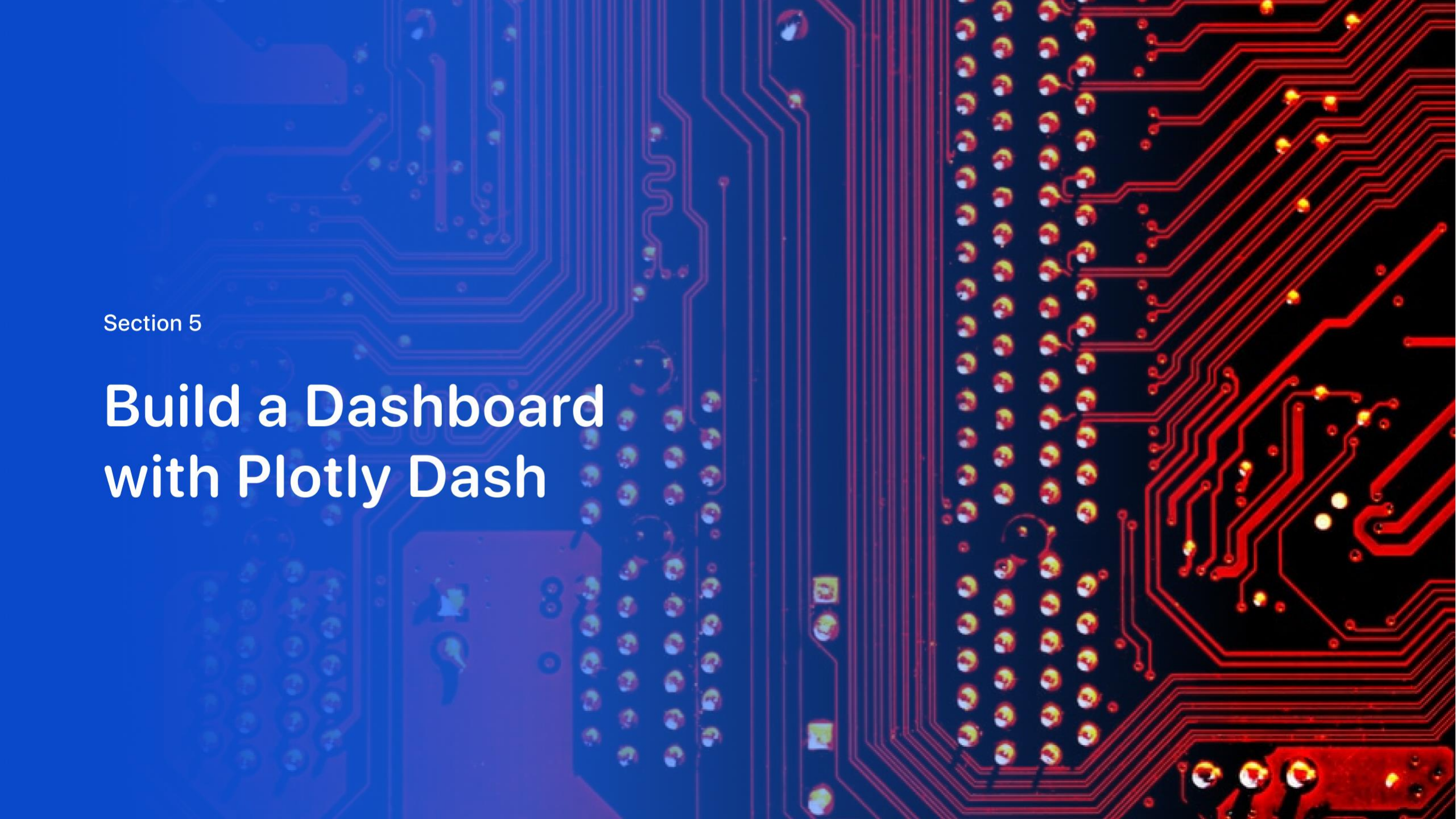
Distance proximity to CCAFS SLC-40

Launch Sites are located closer to coastlines, Shuttle Landing facilities, railroad and highways.

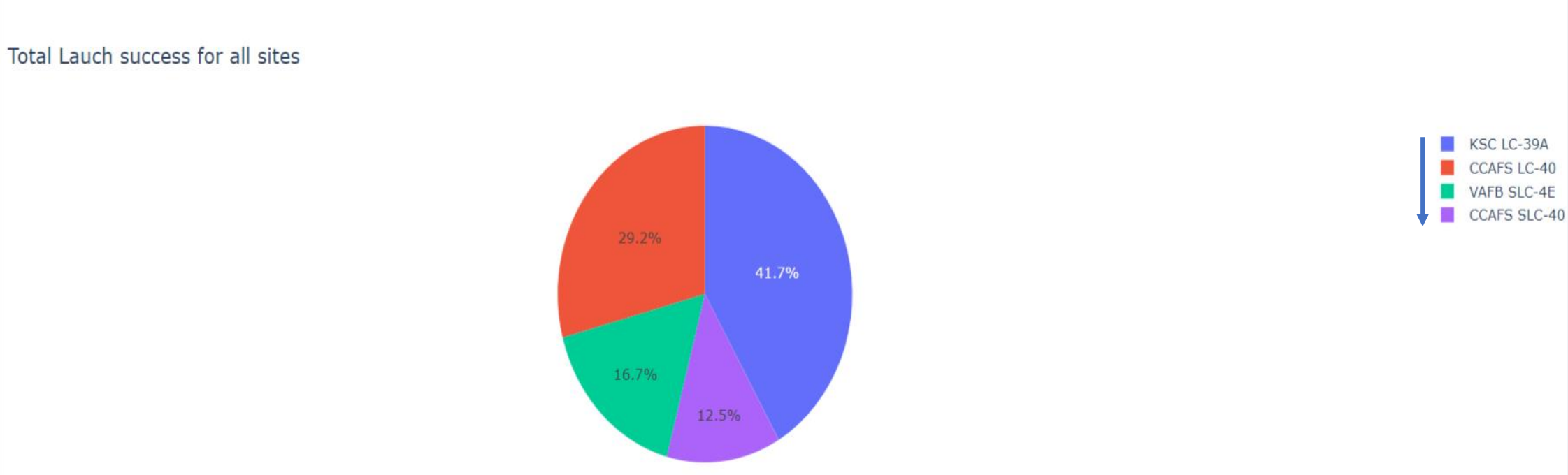
Also noticed that the city is located further from Launch sites.

Section 5

Build a Dashboard with Plotly Dash



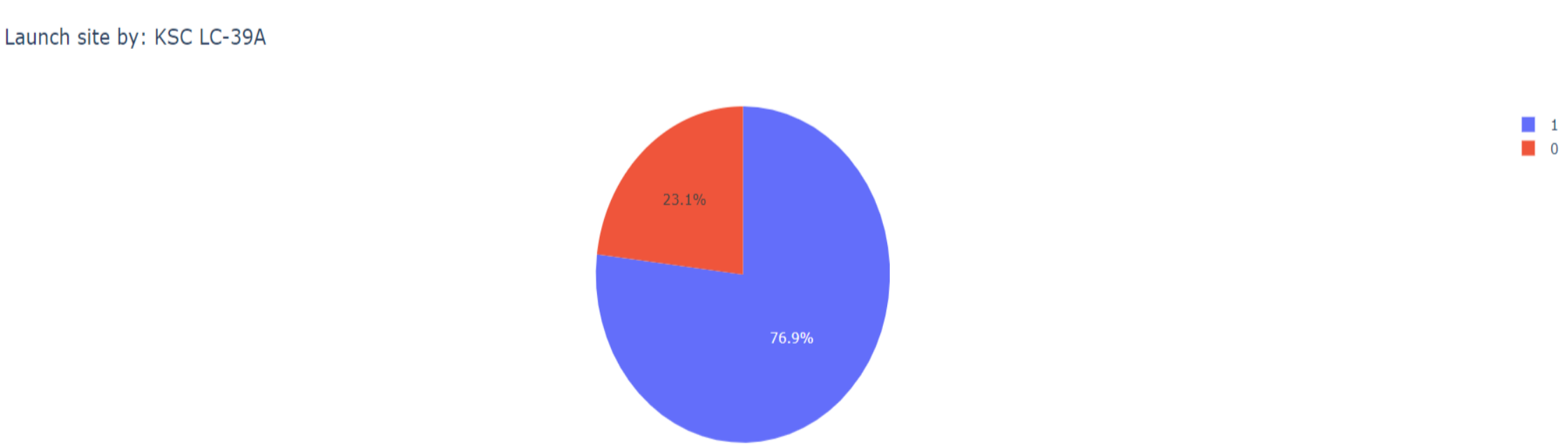
Launch Site Success Counts



The arrow in the legend shows a decrease in success counts for various launch sites

- KSC LC-39A has highest launch site success count and CCAFS SLC-40 the lowest success counts

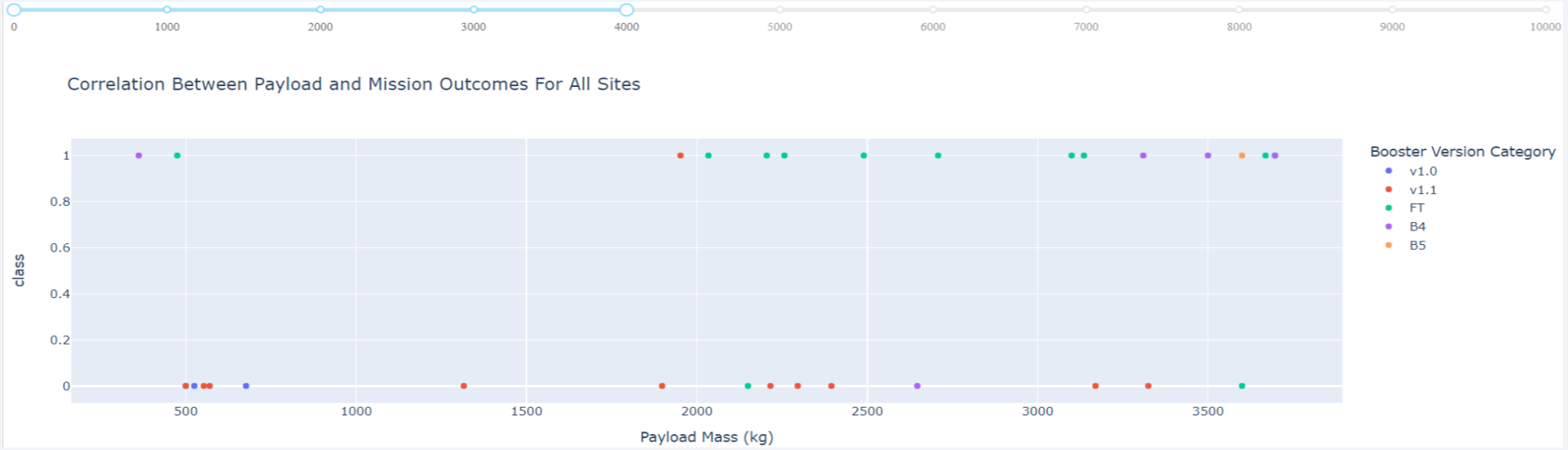
Site with highest launch success



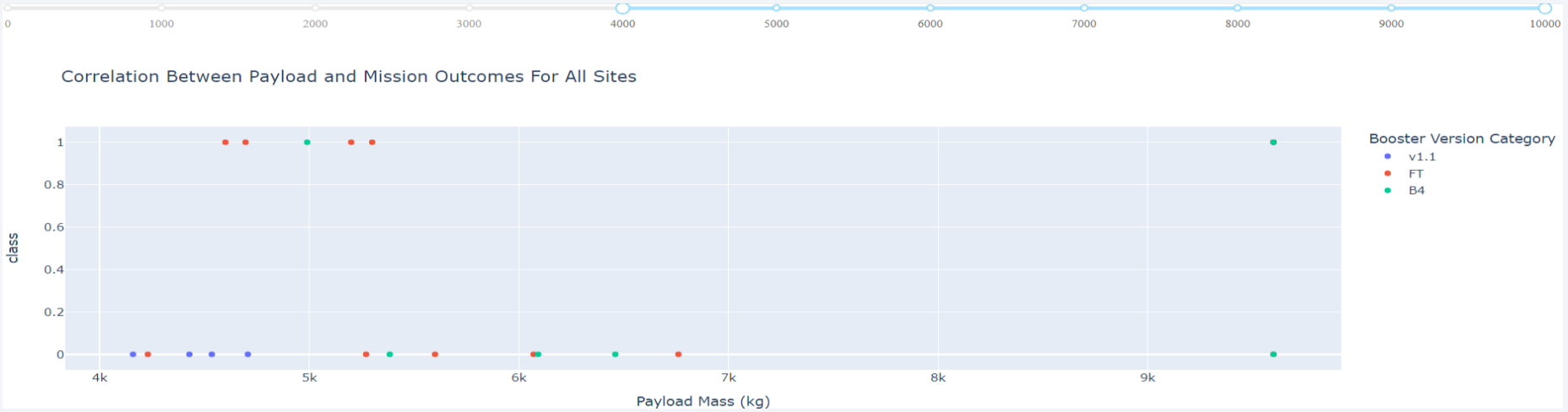
KSL LC-39A achieved a 76.9% success rate and a 23.1% failure rate

NB: Red fill indicate failure rate and Blue indicate success rate.

Payload vs. Launch Outcome



- For payloads in Range 0-4000kg, The success rate is More than payloads In the range 4000-10000kg

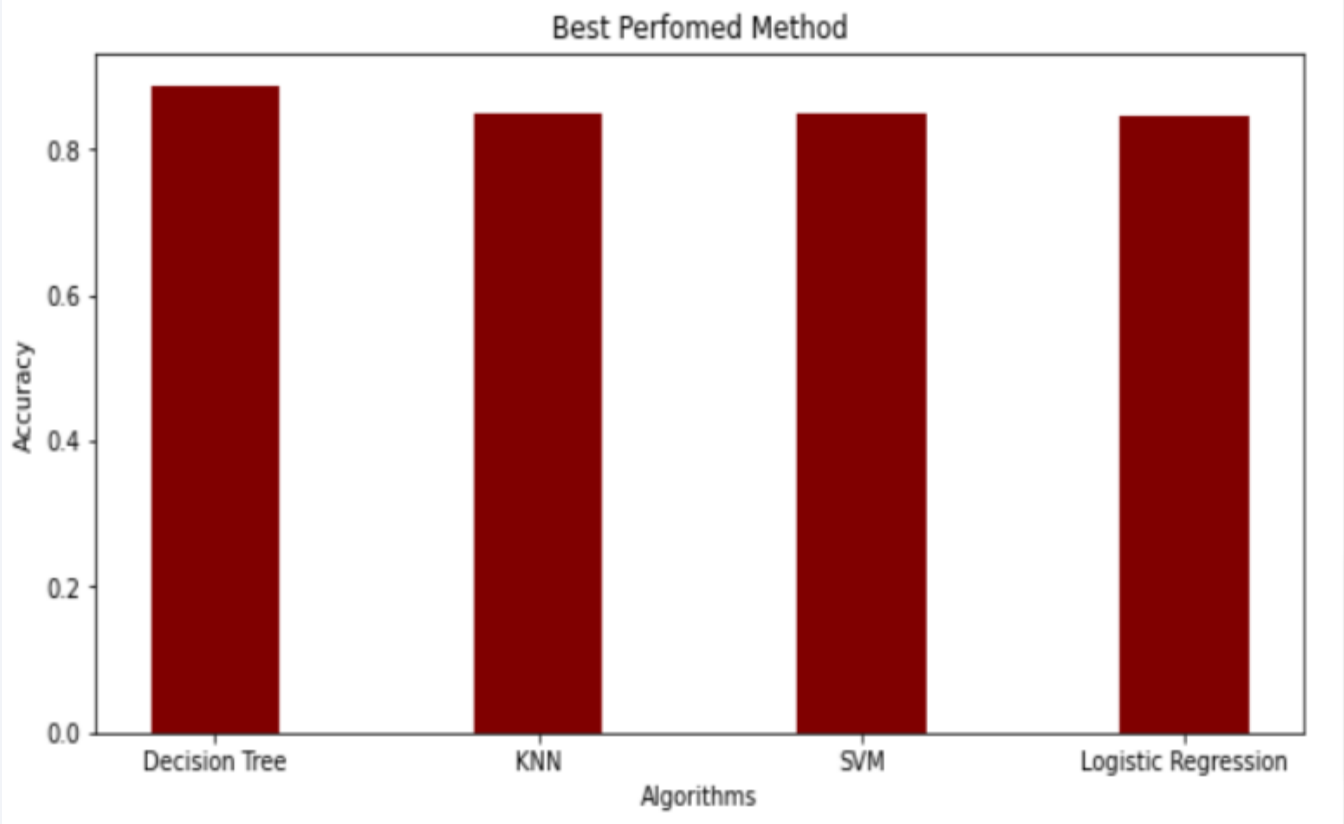


- In conclusion , Higher payload Mass will have Less success rates And Low payload, High success rate

Section 6

Predictive Analysis (Classification)

Classification Accuracy

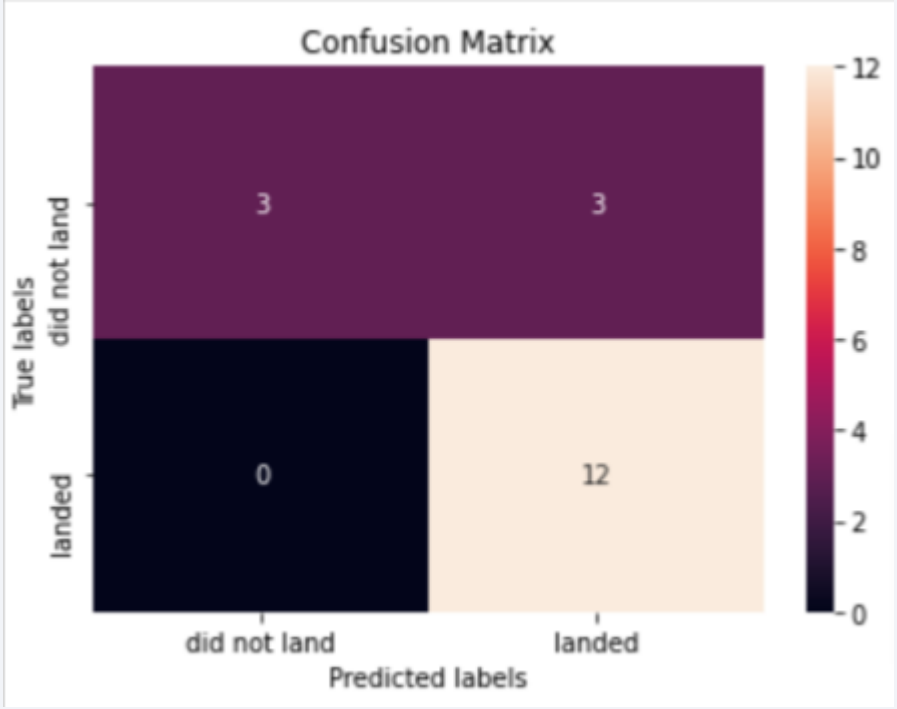


Algorithm	Accuracy Score	Test Data Accuracy Score
Decision Tree	0.875000	0.833333
KNN	0.848214	0.833333
SVM	0.848214	0.833333
Logistic Regression	0.846429	0.833333

- From Bar charts, the Decision Tree Algorithm outperforms the KNN(K-nearest neighbors), SVM(support Vector machine) and the Logistic Regression Algorithms On the training data Eventhough the test scores are the same.

- The Decision Tree model has an accuracy of 87.5% on the training data set.

Confusion Matrix



- The Decision Tree classifier made 18 total predictions
  Predicted 3 correct unsuccessful outcome(TN-True Negatives),
  12 correct successful Outcomes (TP- True positives)
- 3 incorrect predictions: the algorithm predicted that 3 launches Where successful when actually there where not successful. These Are called False positives (FP)
- Classifier did not predict any False Negatives (FN)
- On the Test set the accuracy is ((TP+TN)/Total = 0.83) and the Misclassification rate is 0.16
- Precision = (TP)/(TP + FP), Recall = (TP)/(TP+FN)

Conclusions

- The Decision Tree Classifier is the best Machine Learning algorithm for this data set
- Orbits ES-L1, GEO, HEO, SSO has highest success rate.
- LEO, ISS, PO, SSO, MEO, VLO has increase success rate as flight number increases
- Success rate of Falcon 9 SpaceX launches gets better with time as there keep improving on the launches learning from previous
- Highest success rate from Florida launch site KSC LC -39A
- Launches with low payload masses will have higher success rate than higher payload mass

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!