IBM Data Science Capstone Project – Space X

Anna Semenova

Nijmegen, The Netherlands

14 June 2022

Outline

❑ Executive Summary

❑ Introduction

❑ Methodology

❑ Results

❑ Conclusion

Executive Summary

SpaceX dominates the market because of the reusability of its rockets. To outcompete them, we need to find a way to predict the landing outcome of our rockets after they are launched.

Solution:

We collected data on SpaceX's Falcon 9 rocket launches and used it to train machine learning models that will predict whether the first stage of the rocket will land successfully. This was done by:

- ☐ Data Collection.
- ☐ Data Wrangling
- ☐ Data Analysis
- ☐ Predictive Modeling using Machine Learning Algorithms

Outcome:

Powerful data science analytics and machine learning tools not only increase the competitiveness, but they also increase the customer confidence in the product offering. Predictive models enabled our company to determine the success of a rocket landing. This was especially important, because a failed landing could result in a loss of tens of millions of dollars. In addition, it allowed our company to determined costs and made more competitive offer than SpaceX.

The full project repository: https://github.com/Annatje1503/Capstone-Project/tree/master

Introduction

❑ Project background and context:

The commercial space age is here, companies are making space travel affordable for everyone. Perhaps the most successful company is SpaceX. 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. The primary objective of this Data Science project is to help our company to compete with SpaceX. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. As result, our company can make more informed bids against SpaceX for a rocket launch.

❑ Business problem question:

“Will the Falcon 9 first stage will land successfully and consequently be reused?”

❑ Questions to be answered:

“What is the success rate of landing the 1st stage of Falcon 9 and what is the price of each Falcon 9 rocket launch?”

Methodology

□ **Perform Data Collection:**

- Data collection of SpaceX launch data with an API - the SpaceX REST API;
- Web scraping related Wiki pages.

□ **Perform Data Wrangling** to transform the raw data into a clean dataset for Machine Learning:

- Wrangling data using an API;
- Sampling Data;
- Dealing with Nulls.

□ **Perform Exploratory Data Analysis (EDA)** using SQL queries and Panda and Matplotlib to visualize and determine what attributes are correlated with successful landings

□ **Perform Interactive Visual Analytics** using Folium and Plotly Dash

□ **Perform Predictive Analysis** using classification models Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors and evaluate the best classifier

METHODOLOGY: Data Collection

Methodology: Data Collection

Data Collection

The data collection stage is the most crucial stage in the project because it is used to train machine learning models to make precise predictions. There are different ways to collect data. There are two ways were used:

- ❑ Data collection by SpaceX API request.
- ❑ Data collection by Web Scraping

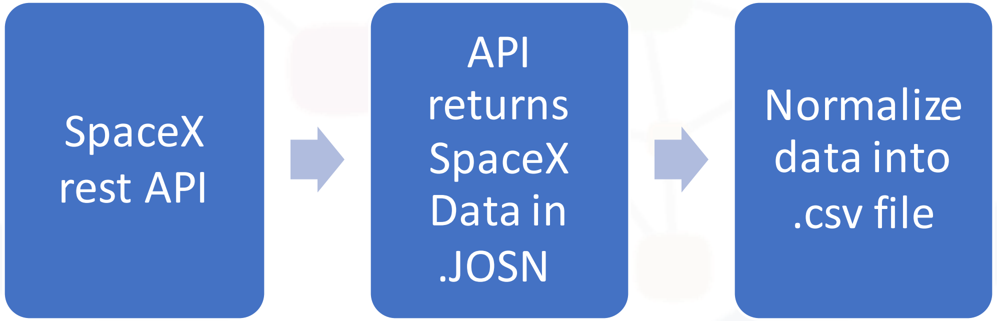
These methods require only a well-functioning internet connection.

Methodology: Data Collection

Data Collection – SpaceX API:

- □ SpaceX launch data was gathered from the SpaceX REST API. This API gave us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- □ The data was checked to make sure that it is in the correct format and some basic data wrangling and formatting was performed in order to clean the requested data.
- □ Our data frame was converted into a CSV dataset.

SpaceX API



SpaceX rest API

API returns SpaceX Data in .JOSN

Normalize data into .csv file

GitHub URL to notebook: https://github.com/Annatje1503/Capstone-Project/blob/master/Final_notebook_1.ipynb

Methodology: Data Collection

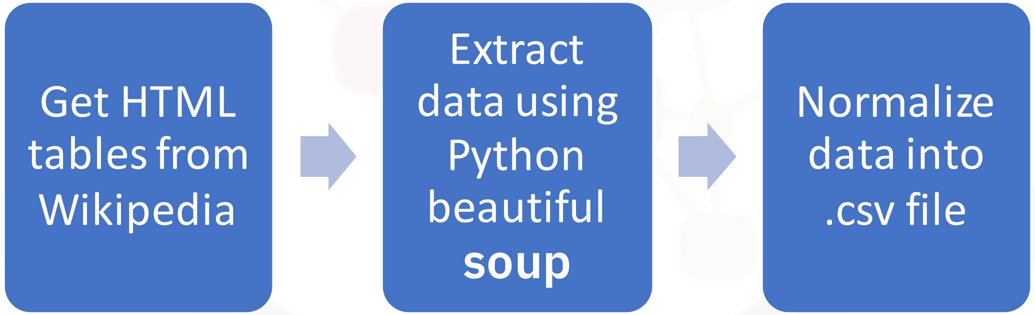
Data Collection - Web Scraping:

□Another popular data source for obtaining Falcon 9 launch data was used by web scraping related Wiki pages. For this the Python BeautifulSoup package was used.

□Launch records were stored in an HTML table.

□The table was parsed and convered it into a CSV dataset.

Web scraping



Get HTML tables from Wikipedia

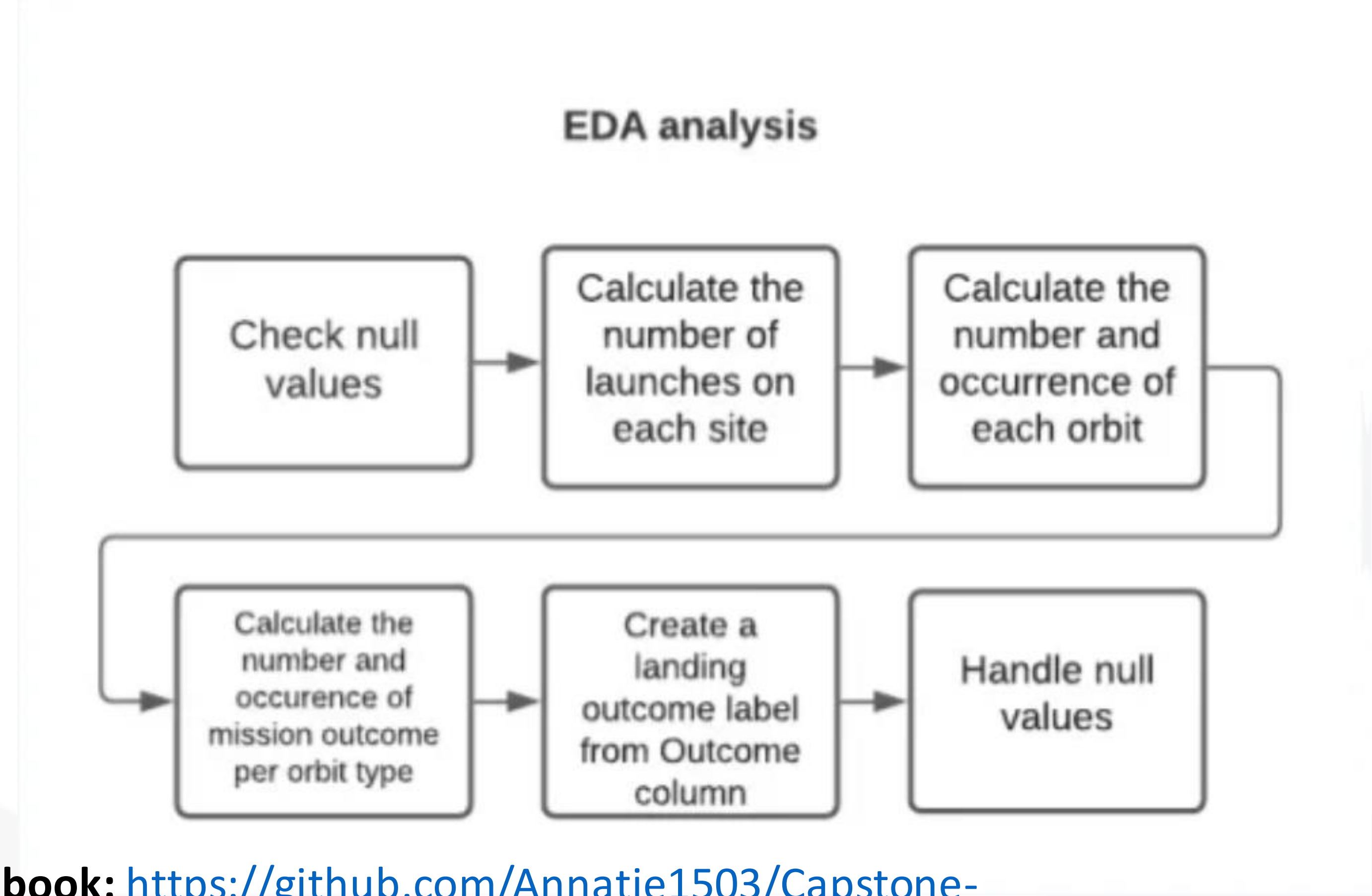
Extract data using Python beautiful **soup**

Normalize data into .csv file

GitHub URL to notebook: https://gist.github.com/Annatje1503/7ae253910f64a7baf0a839824f806134

METHODOLOGY: Data Wrangling

Methodology: Data Wrangling


EDA analysis
Check null values
Calculate the number of launches on each site
Calculate the number and occurrence of each orbit
Calculate the number and occurence of mission outcome per orbit type
Create a landing outcome label from Outcome column
Handle null values

GitHub URL to notebook: https://github.com/Annatje1503/Capstone-Project/blob/master/Final_assignment_2.ipynb

METHODOLOGY: EDA with Visualization

Methodology: EDA with Visualization (using Pandas and Matplotlib)

Scatter Graphs were drawn:

- Flight Number versus Payload Mass
- Flight Number versus Launch Site
- Payload Mass versus Launch Site
- Flight Number versus Orbit Type
- Payload Mass versus Orbit Type

Scatter plots are useful data visualization tools for illustrating a trend. In particularly scatter plots are the graphs that present the relationship between two variables in a data-set. The relationship is called their correlation.

Bar Graph were drawn:

- Orbit Type versus Success Rate of Each Orbit

A bar graph is a visual tool that uses bars to compare data among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value. The longer the bar, the greater its value.

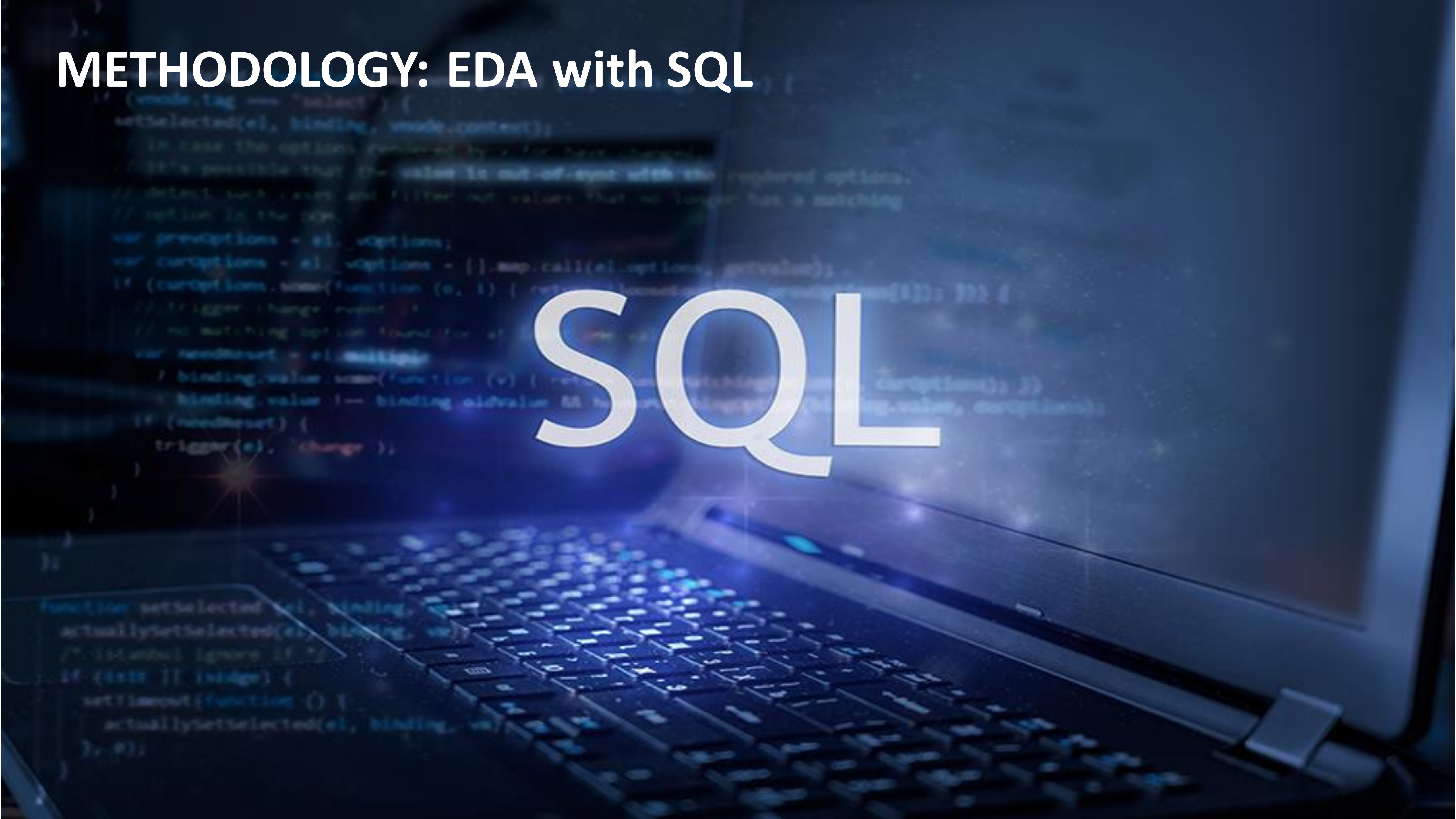
Line Graph were drawn:

- Year versus Average Success Rate

A line chart is a type of chart used to show information that changes over time. Line charts are used to track changes over short and long periods of time.

GitHub URL to notebook: https://github.com/Annatje1503/Capstone-Project/blob/master/Final_assignment_4.ipynb

METHODOLOGY: EDA with SQL



Methodology: EDA with SQL

The Spacex DataSet was loaded into corresponding table in Db2 dataset and SQL queries were written and executed to solve the assignment tasks:

1. Display the names of the unique launch sites in the space mission.
2. Display 5 records where launch sites begin with the string 'CCA'.
3. Display the total payload mass carried by boosters launched by NASA (CRS).
4. Display average payload mass carried by booster version F9 v1.1.
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
7. List the total number of successful and failure mission outcomes.
8. List the names of the booster_versions which have carried the maximum payload mass.
9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for the year 2015.
10. Rank 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.

GitHub URL to notebook: https://github.com/Annatje1503/Capstone-Project/blob/master/EDA%20with%20SQL.ipynb

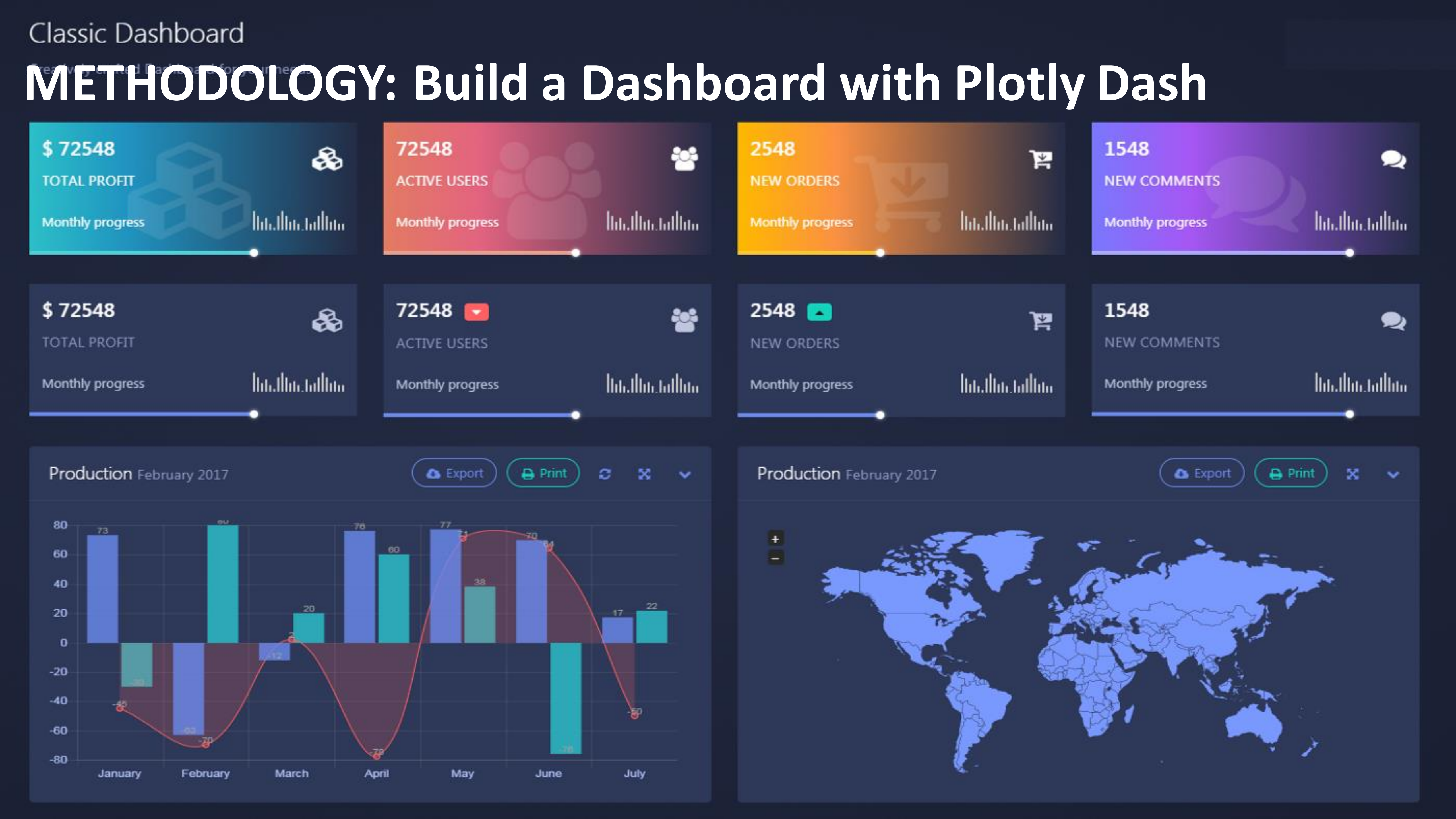
METHODOLOGY: Interactive Map with Folium

Methodology: Building an Interactive Map with Folium

- ❑ Folium Markers were used to show the SpaceX launch sites and their nearest important landmarks like railways, highways, cities and coastlines.
- ❑ Polylines were used to connect the launch sites to their nearest land marks.
- ❑ Folium Circles were used to highlight circle area of launch sites.
- ❑ To mark the success/failed launches for each site, marker clusters were used on the map. Red color represents the rocket launch failures while Green color represents the successes.

GitHub URL to notebook: https://github.com/Annatje1503/Capstone-Project/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

METHODOLOGY: Build a Dashboard with Plotly Dash



Methodology: Build a Dashboard with Plotly Dash

The dashboard is built with Plotly Dash web framework.

Graphs

❑ **Pie Chart** shows the total launches by a certain site/all sites

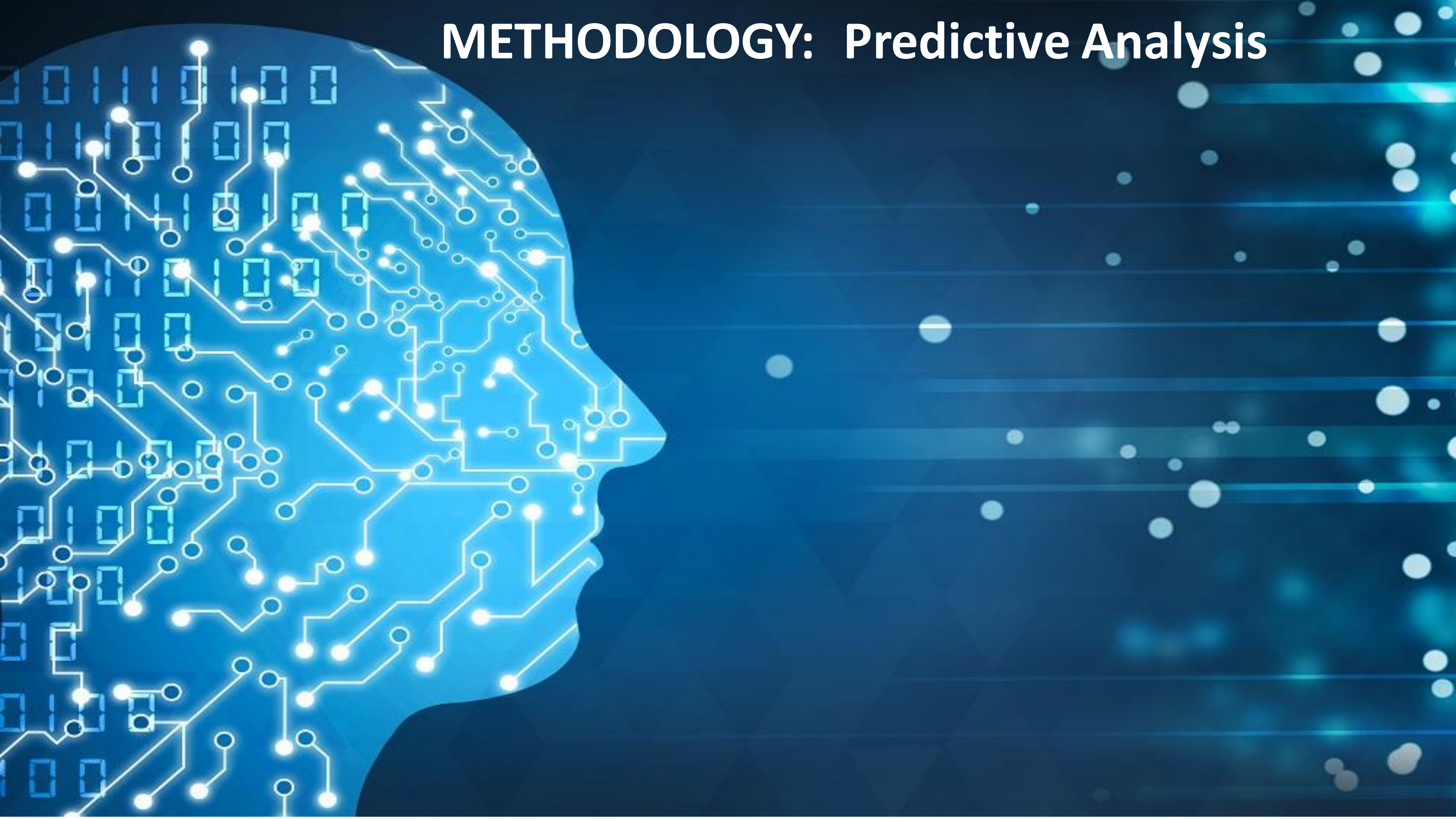
- displays relative proportions of multiple classes of data
- Displays the proportion of a particular data class to the total quantity

❑ **Scatter Graph** shows the relationship with Outcome and Payload Mass (Kg) for the different Booster Versions

- It shows the relationship between two variables.
- It is the best method to show a non-linear pattern.
- The range of data flow, i.e. maximum and minimum value, can be determined.
- Observation and reading are straightforward.

GitHub URL: https://github.com/Annatje1503/Capstone-Project/blob/master/Dashboard%20Application%20with%20Plotly%20Dash.txt

METHODOLOGY: Predictive Analysis



Methodology: Predictive Analysis (Classification)

BUILDING MODEL

- ❑ Load the dataset into NumPy and Pandas
- ❑ Transform Data
- ❑ Split the data into training and test data sets using the function train_test_split
- ❑ Check how many test samples we have
- ❑ Decide which type of machine learning algorithms we want to use
- ❑ Set our parameters and algorithms to GridSearchCV
- ❑ Fit our datasets into the GridSearchCV objects and train our dataset.

EVALUATING MODEL

- ❑ Check accuracy for each model
- ❑ Get tuned hyperparameters for each type of algorithms
- ❑ Plot Confusion Matrix for each type of algorithms

IMPROVING MODEL

- ❑ Feature Engineering
- ❑ Algorithm Tuning

FINDING THE BEST PERFORMING CLASSIFICATION MODEL

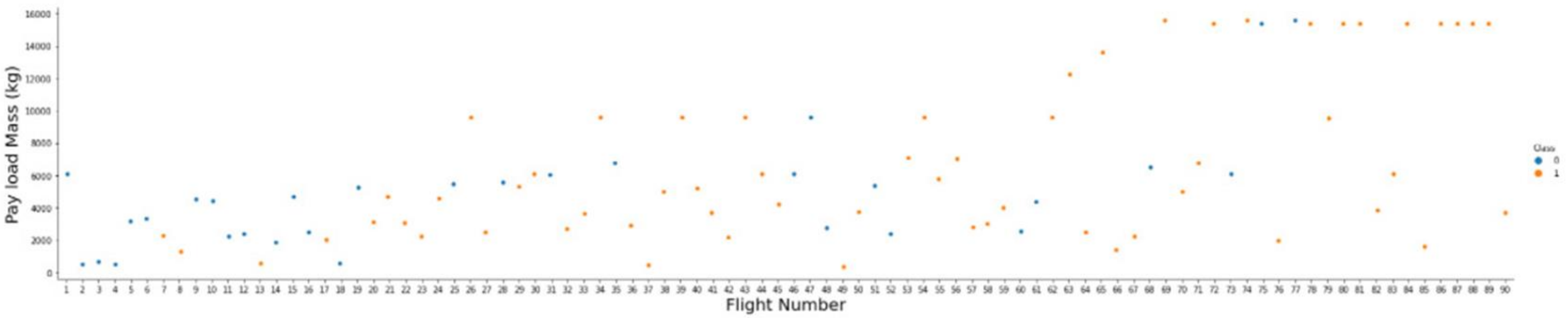
- ❑ The model with the best accuracy score wins
- ❑ In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook

GitHub URL to notebook: https://github.com/Annatje1503/Capstone-Project/blob/master/Machine%20Learning%20Predictive%20Analysis.ipynb

RESULTS: EDA with Visualization

RESULTS: EDA with visualization (using Pandas and Matplotlib)

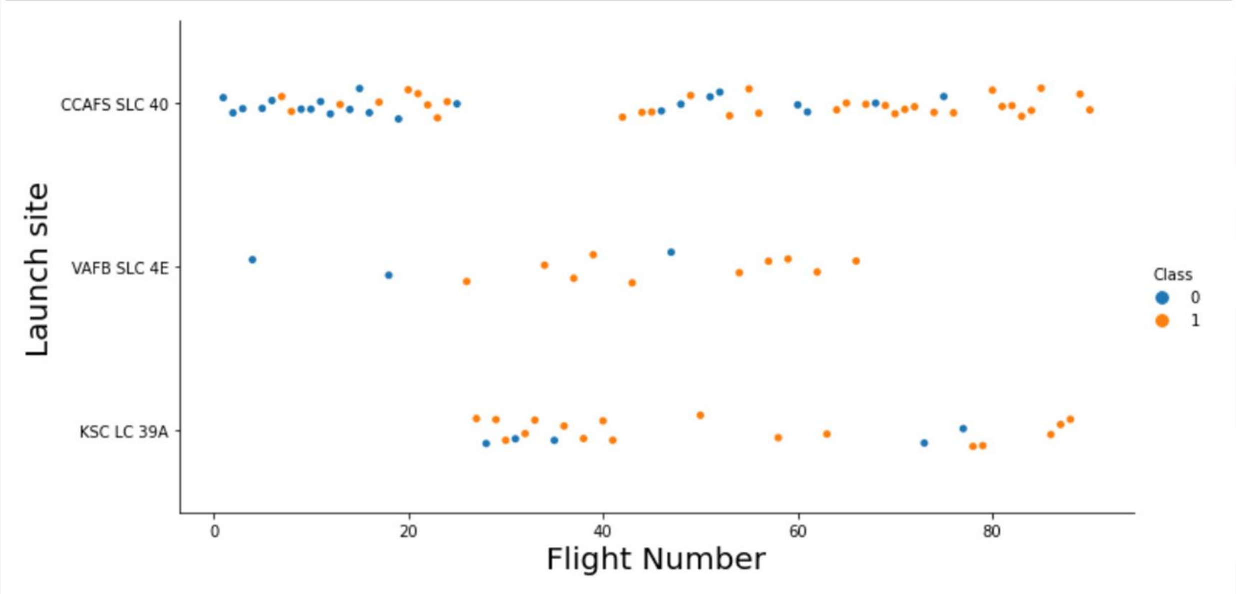
Scatter Graph: Flight Number versus Payload Mass



The flight number increases, the first stage is more likely to land successfully. The payload mass is also important. The more massive the payload, the less likely the first stage will return.

RESULTS: EDA with visualization (using Pandas and Matplotlib)

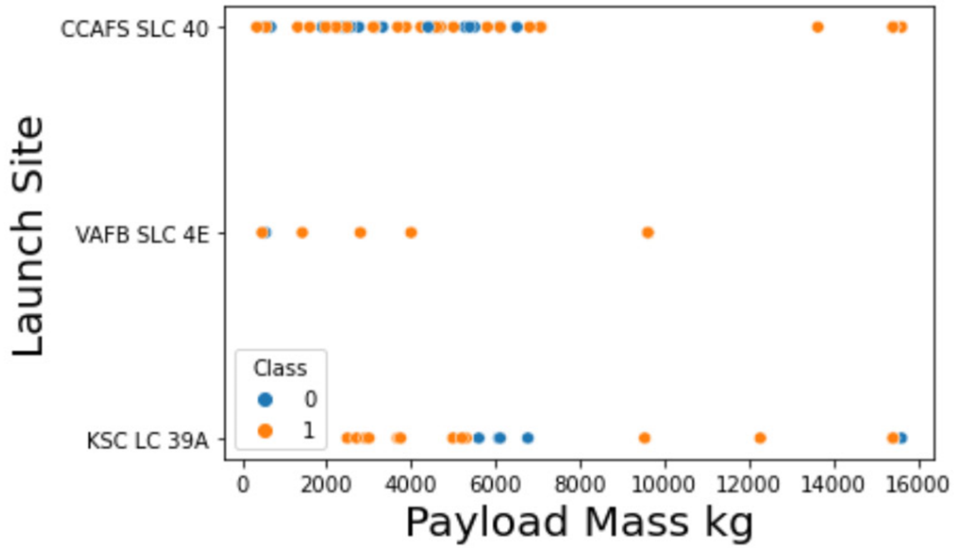
Scatter Graph: Flight Number versus Launch Site



Launches from the site of CCAFS SLC40 are significantly higher than launches from other sites. The more amount of flights at a launch site the greater the success rate at a launch site.

RESULTS: EDA with visualization (using Pandas and Matplotlib)

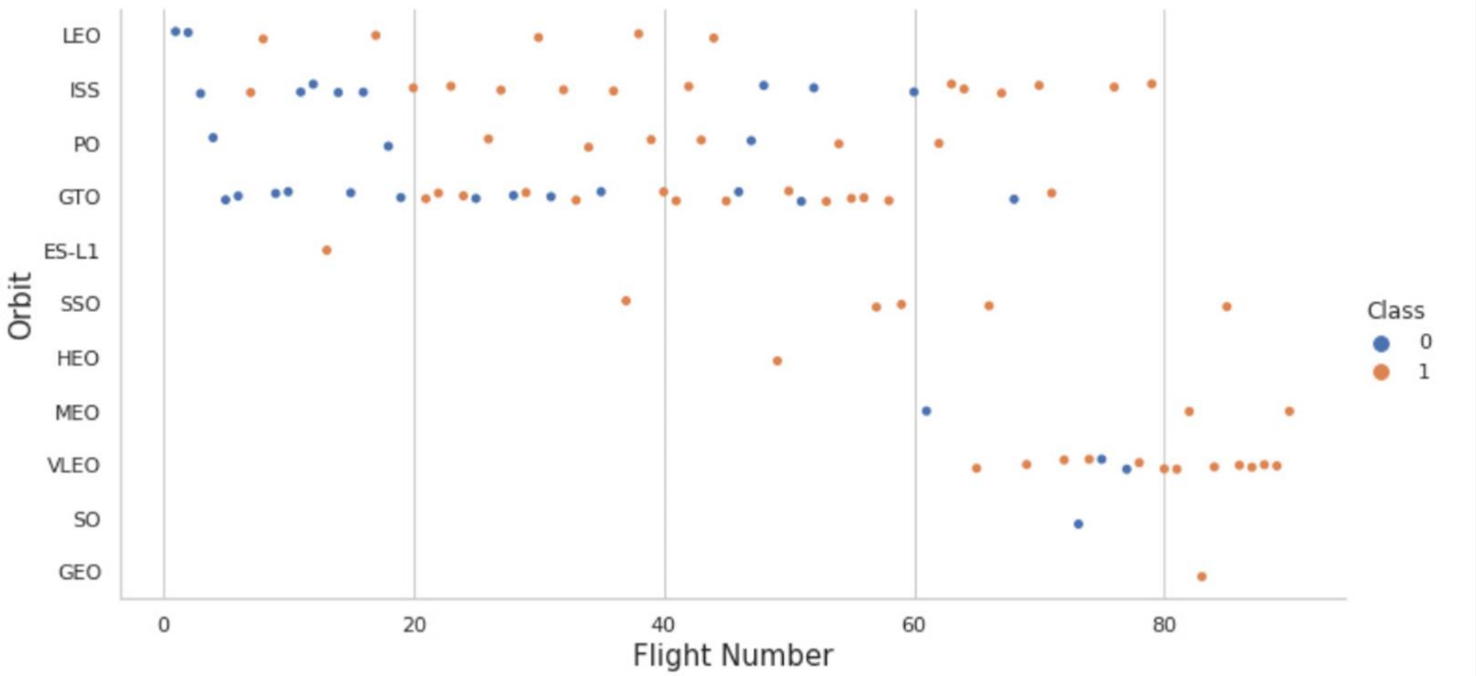
Scatter Graph: Payload Mass versus Launch Site



For the VAFB-SLC Launch Site there are no rockets launched for heavy payload mass(greater than 10000). The majority of payloads with lower mass have been launched from CCAFS SLC 40. The greater the payload mass for Launch Site CCAFS SLC 40, the higher the success rate for the Rocket. There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dependent on Pay Load Mass for a success launch.

RESULTS: EDA with visualization (using Pandas and Matplotlib)

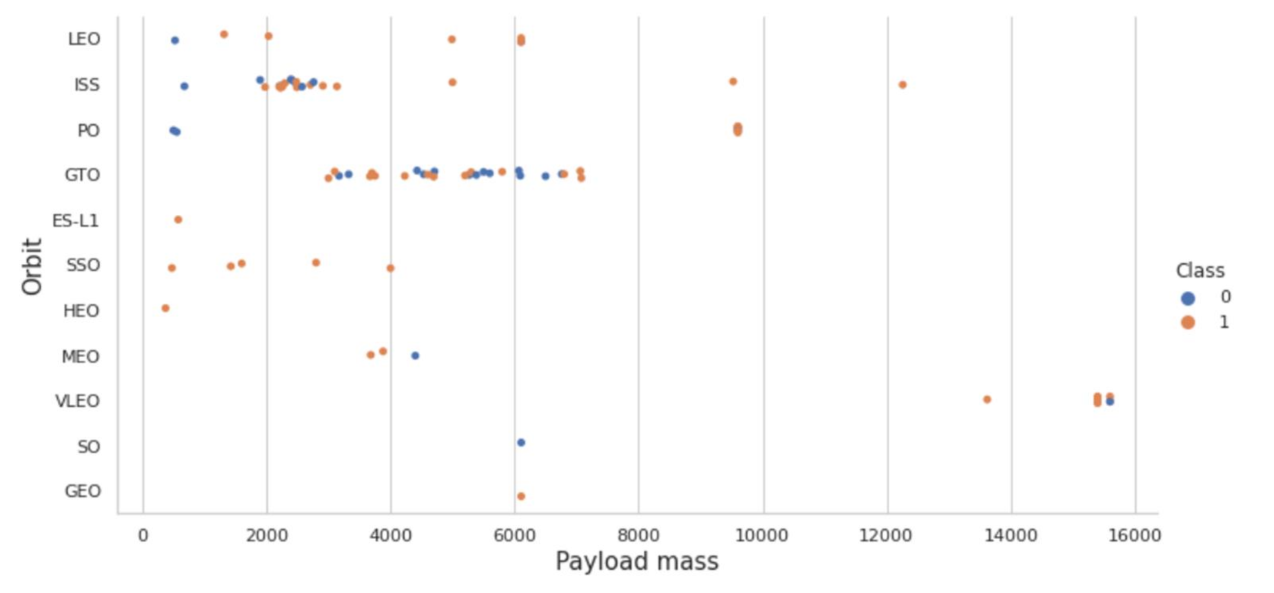
Scatter Graph: Flight Number versus Orbit Type



A trend can be observed of shifting to VLEO launches in recent years. In the LEO orbit the success is related to the number of flights; There's no relationship between flight number in GTO orbit.

RESULTS: EDA with visualization (using Pandas and Matplotlib)

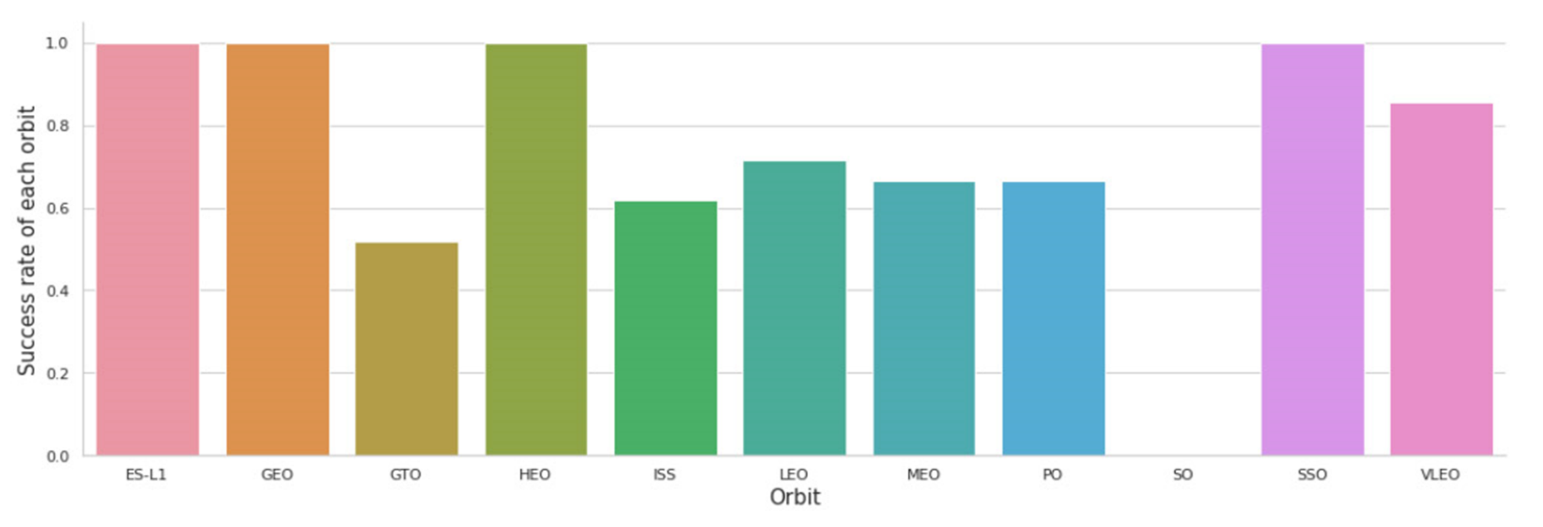
Scatter Graph: Payload Mass versus Orbit Type



There are strong correlation between ISS and Payload at the range around 2000, as well as between GTO and the range of 4000-8000

RESULTS: EDA with visualization (using Pandas and Matplotlib)

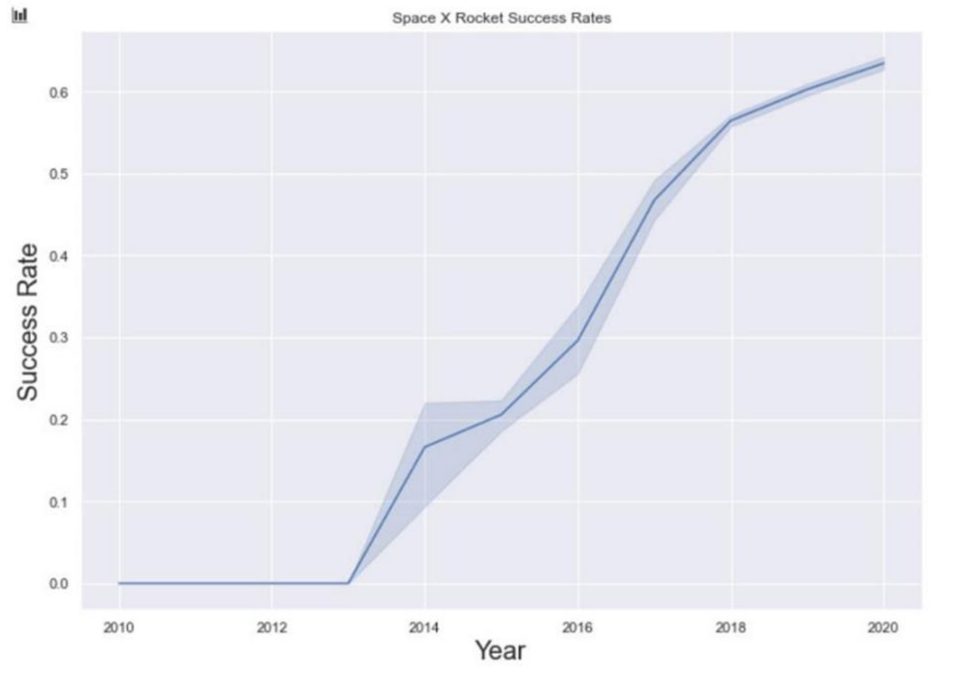
Bar Graph: Flight Number versus Orbit Type



Orbits ES-L1, GEO, HEO and SSO have the best success rate

RESULTS: EDA with visualization (using Pandas and Matplotlib)

Line Graph: Year versus Average Success Rate



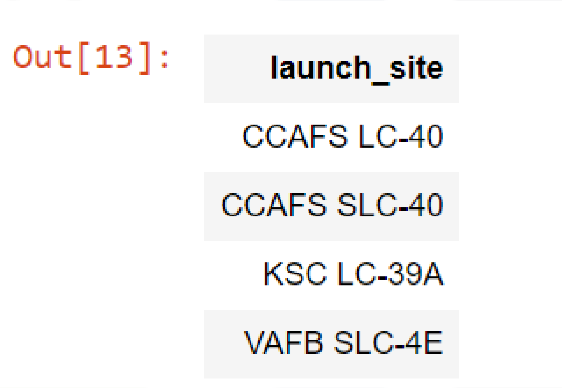
The success rate since 2013 kept increasing till 2020 possibly because of lessons learnt and technological progress

RESULTS: EDA with SQL

RESULTS: EDA using SQL queries

1. Unique Launch Sites

select Unique(LAUNCH_SITE) from SPACEXTBL4



Out[13]:

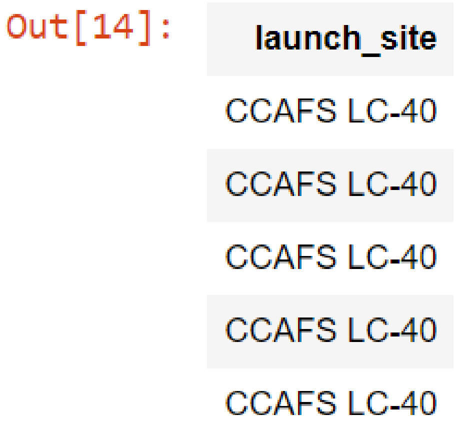
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

These are the names of the launch sites where different rocketlandings where attempted: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

RESULTS: EDA using SQL queries

2. Launch site names begin with ‘CCA’

```
%sql SELECT LAUNCH_SITE from SPACEXTBL4
where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5
```



Out[14]:

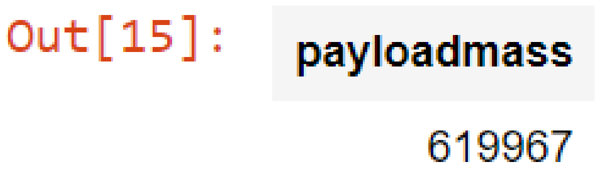
launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

These are 5 records where launch sites begin with the letters 'CCA'.

RESULTS: EDA using SQL queries

3. Total payload mass by boosters launched by NASA (CRS)

```
%sql select sum(PAYLOAD_MASS__KG_) as
payloadmass from SPACEXTBL4
```



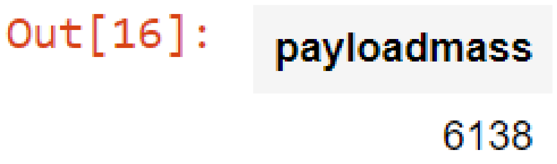
```
Out[15]: payloadmass
              619967
```

Total payload mass by boosters launched by NASA (CRS) is 619,967 kg.

RESULTS: EDA using SQL queries

4. Average payload mass by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) as
payloadmass from SPACEXTBL4
```



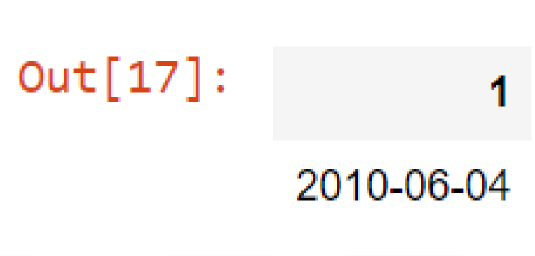
```
Out[16]:  payloadmass
                 6138
```

The average payload mass carried by F9 v1.1 was 6138 kg.

RESULTS: EDA using SQL queries

5. Date of the first successful landing outcome

```
%sql select min(DATE) from SPACEXTBL4
```



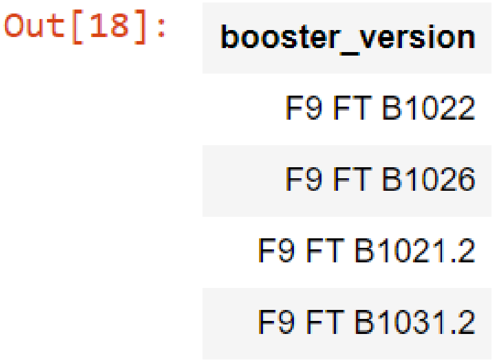
```
Out[17]:          1
         2010-06-04
```

The first successful ground pad landing took place in June 2010.

RESULTS: EDA using SQL queries

6. Boosters’ names which are successful in drone ship (4000< payload mass <6000)

```
%sql select BOOSTER_VERSION from SPACEXTBL4 where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
```



Out[18]:

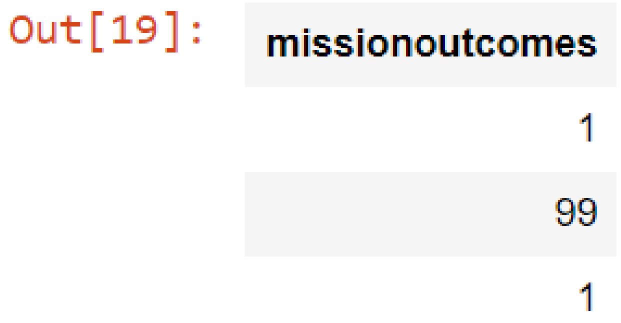
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

There were only 4 Boosters with a payload mass between 4000 and 6000 and all of them had successful landing outcome.

RESULTS: EDA using SQL queries

7. Total number of successful and failure mission outcomes

%sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL4 GROUP BY MISSION_OUTCOME



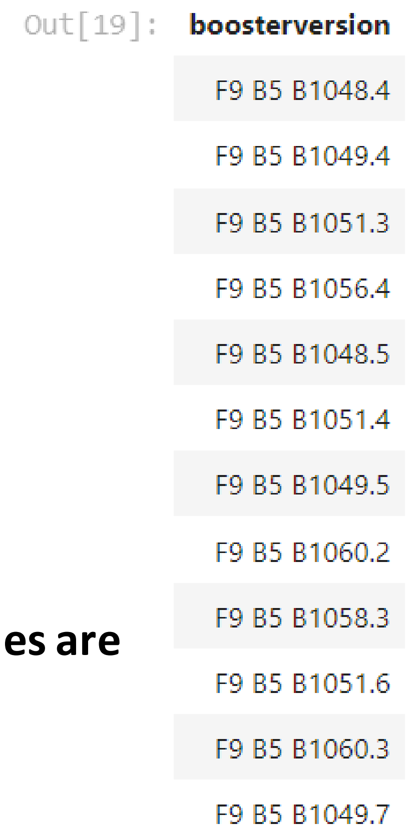
```
Out[19]:  missionoutcomes
                        1
                       99
                        1
```

Missions generally tend to be successful with the exception of one failure.

RESULTS: EDA using SQL queries

8. Booster_versions with maximum payload mass

%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL4 where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL4)



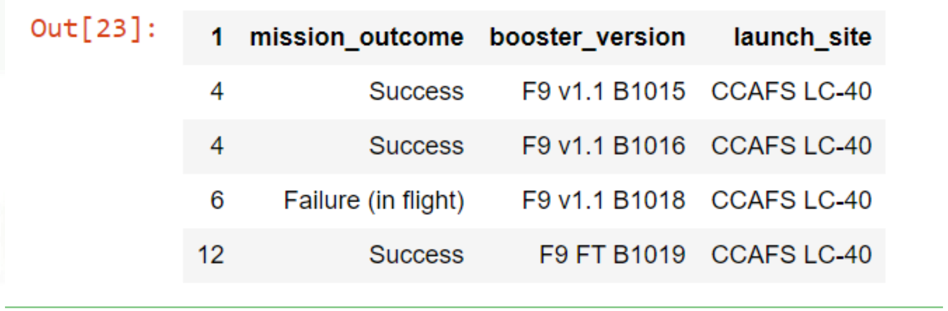
```
Out[19]:  boosterversion
          F9 B5 B1048.4
          F9 B5 B1049.4
          F9 B5 B1051.3
          F9 B5 B1056.4
          F9 B5 B1048.5
          F9 B5 B1051.4
          F9 B5 B1049.5
          F9 B5 B1060.2
          F9 B5 B1058.3
          F9 B5 B1051.6
          F9 B5 B1060.3
          F9 B5 B1049.7
```

12 boosters have carried the maximum payload mass. Since the version names are similar, they might be from the same manufactures.

RESULTS: EDA using SQL queries

9. Failed landing_outcomes in drone ship in 2015

```
%sql SELECT
MONTH(DATE),MISSION_OUTCOME,BOOSTE
R_VERSION,LAUNCH_SITE FROM SPACEXTBL
where EXTRACT(YEAR FROM DATE)='2015'
```



Out[23]:

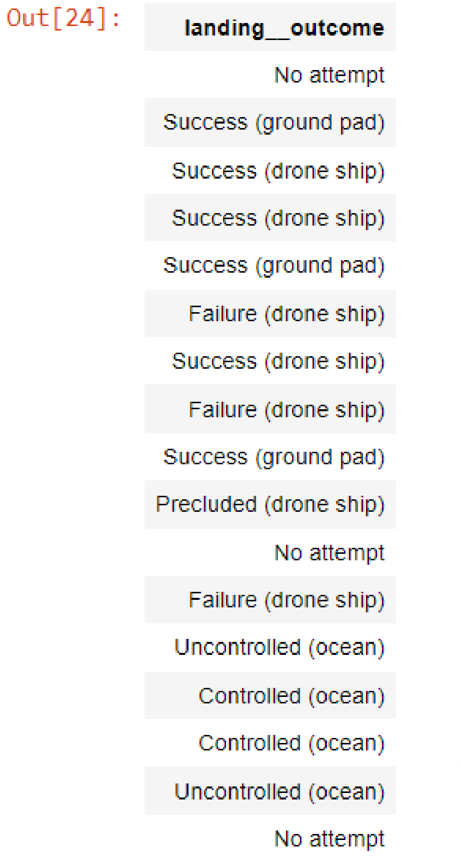
1	mission_outcome	booster_version	launch_site
4	Success	F9 v1.1 B1015	CCAFS LC-40
4	Success	F9 v1.1 B1016	CCAFS LC-40
6	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

Booster version F9v1.1 B1018 failed to land in 2015.

RESULTS: EDA using SQL queries

10. Landing outcomes rank (descending) between 2010-06-04 and 2017-03-20

```
%sql SELECT LANDING__OUTCOME FROM
SPACEXTBL WHERE DATE BETWEEN '2010-06-
04' AND '2017-03-20' ORDER BY DATE DESC
```



Out[24]:

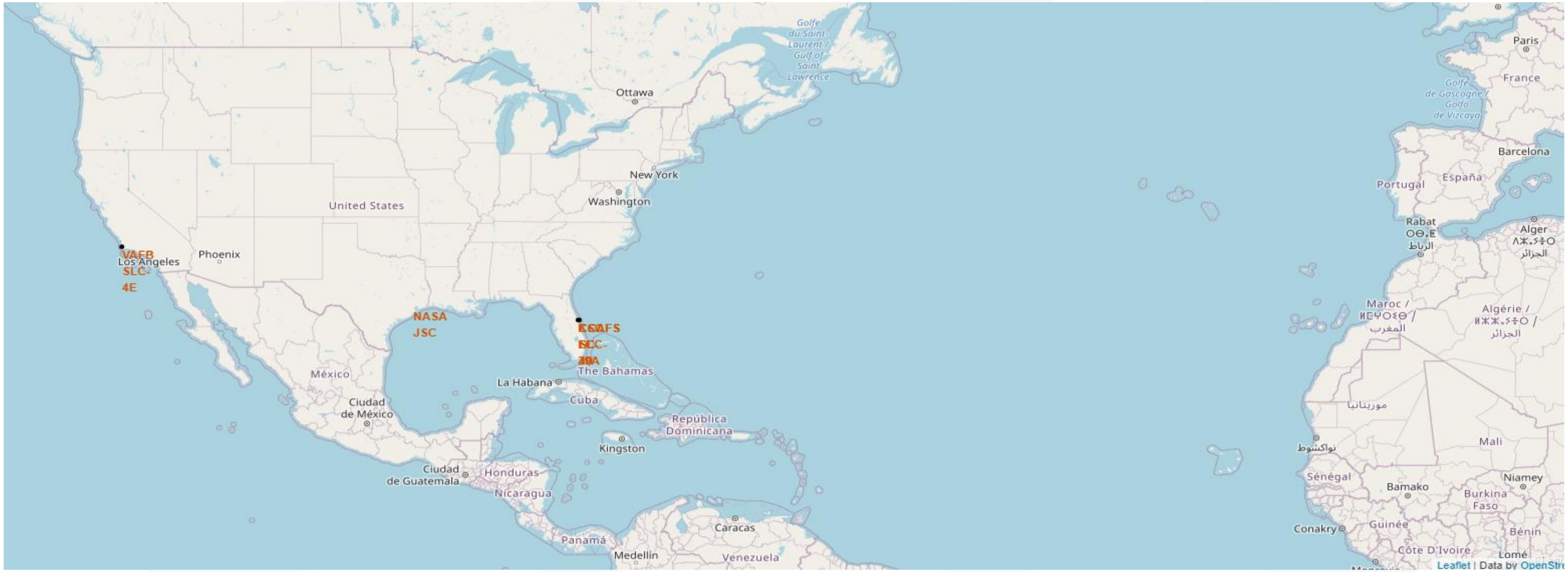
landing__outcome
No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Failure (drone ship)
Success (ground pad)
Precluded (drone ship)
No attempt
Failure (drone ship)
Uncontrolled (ocean)
Controlled (ocean)
Controlled (ocean)
Uncontrolled (ocean)
No attempt

The number of successful landings have increased since 2013. Before 2013, it seems that there were no attempts to land the boosters.

RESULTS: Interactive Map with Folium

RESULTS: Interactive Map with Folium

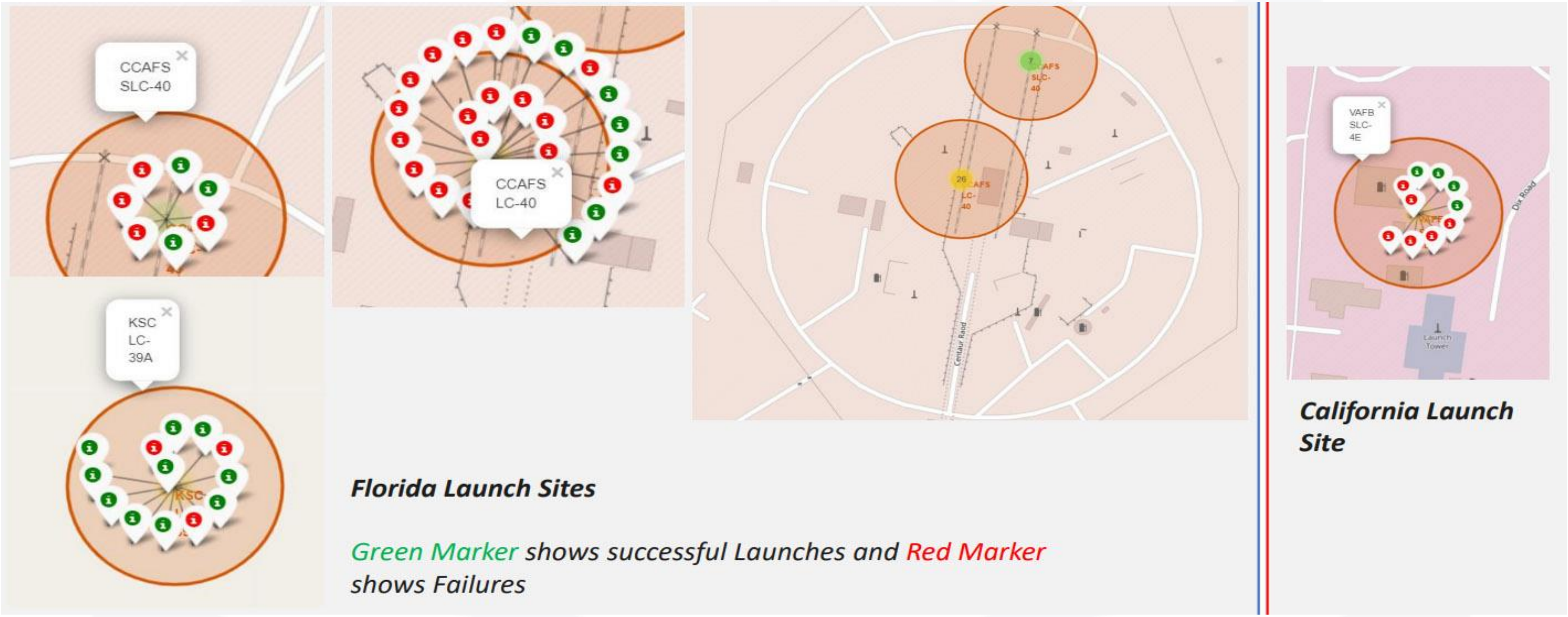
All launch sites are marked on a map



All launch sites are in very close proximity to the coast and they are also a couple thousand kilometers away from the equator line. Most launch sites are concentrated near Florida and California.

RESULTS: Interactive Map with Folium

Color Labelled Markers



Florida Launch Sites

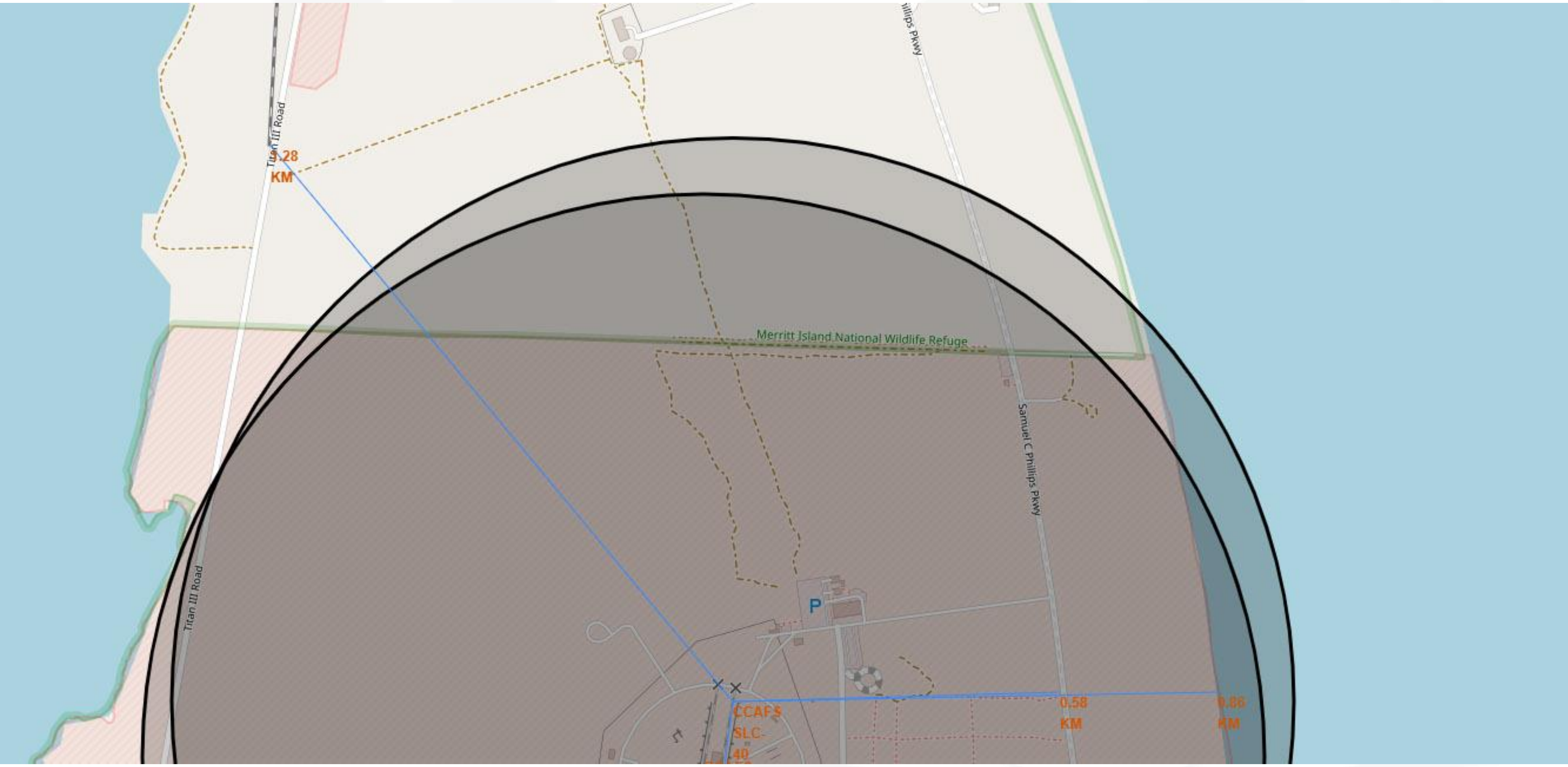
Green Marker shows successful Launches and Red Marker shows Failures

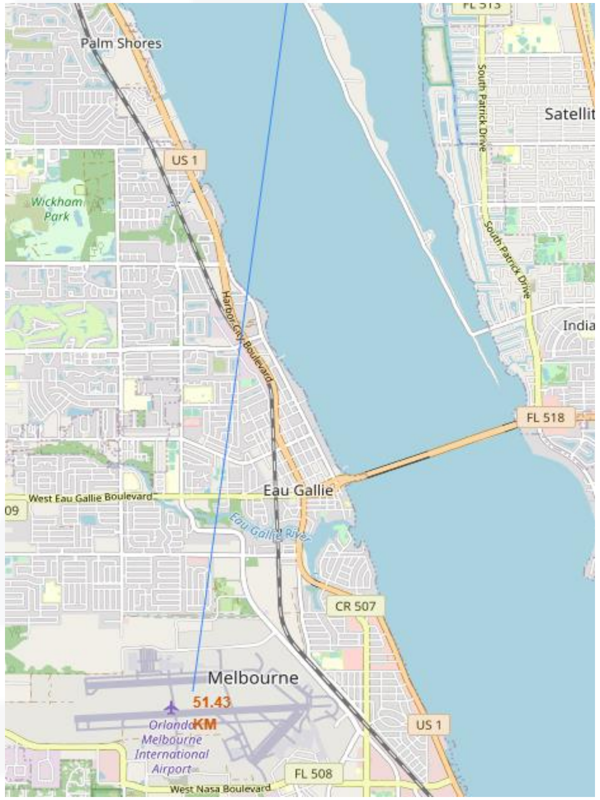
California Launch Site

KSC LC- 39A had the highest success rate of rocket launches compared to other launch sites.

RESULTS: Interactive Map with Folium

Working out Launch Sites distance to landmarks to find trends with Haversine formula using CCAFS-SLC-40 as a reference





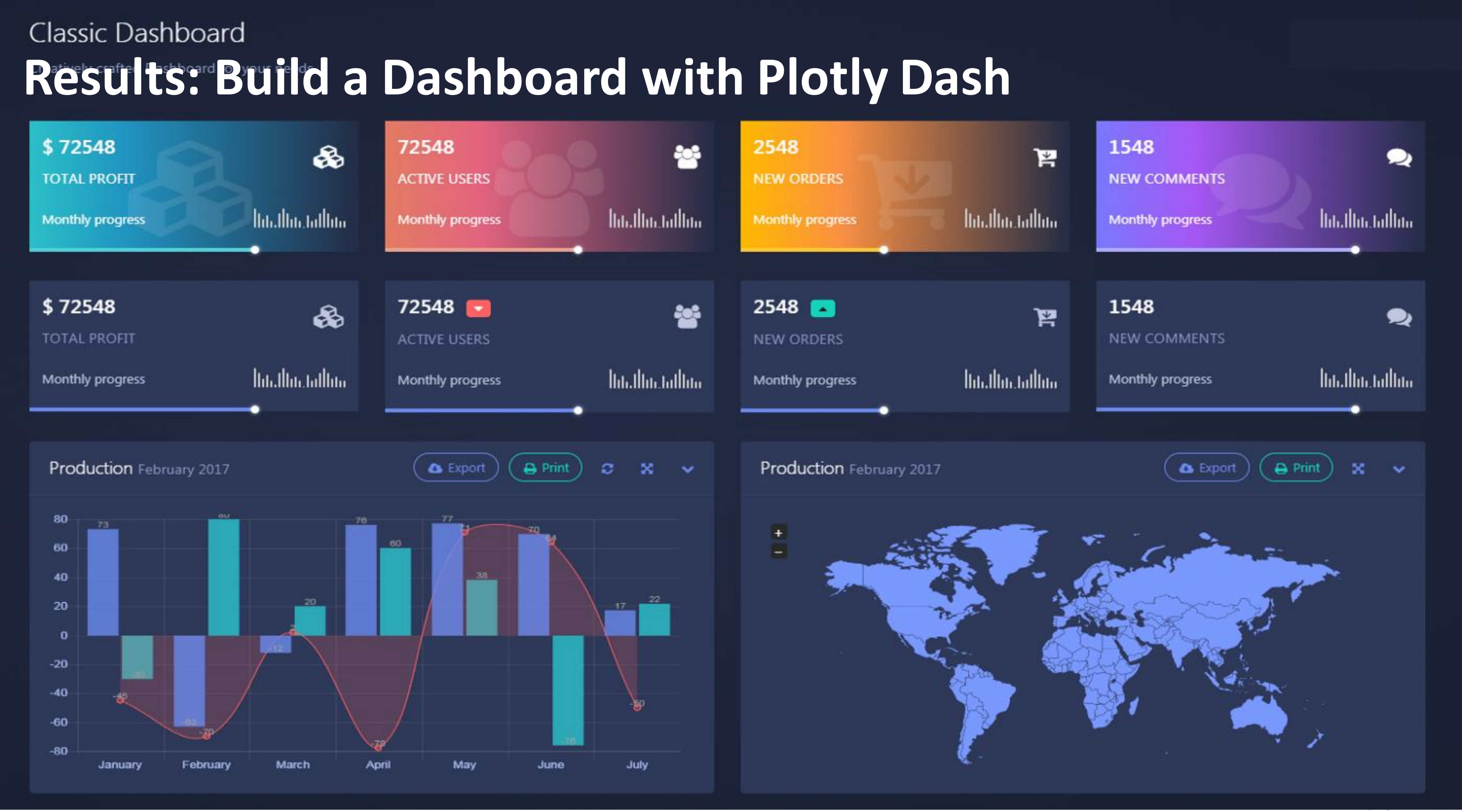
RESULTS: Interactive Map with Folium

Working out Launch Sites distance to landmarks to find trends with Haversine formula using CCAFS-SLC-40 as a reference

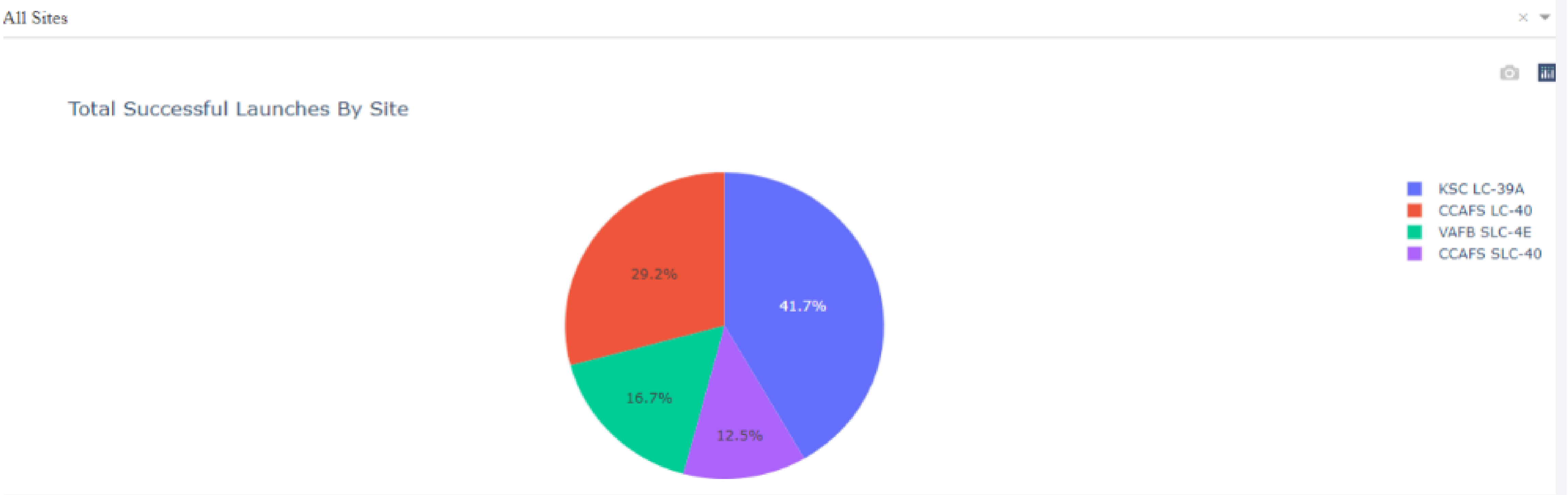
FINDINGS:

- Launch sites are in close proximity to equator to minimize fuel consumption by using Earth's ~ 30km/sec eastward spin to help spaceships get into orbit.
- Launch sites are in close proximity to coastline so they can fly over the ocean during launch, for at least two safety reasons:
  (1) crew has option to abort launch and attempt water landing
  (2) minimize people and property at risk from falling debris.
- Launch sites are in close proximity to highways, which allows for easily transport required people and property.
- Launch sites are in close proximity to railways, which allows transport for heavy cargo.
- Launch sites are not in close proximity to cities, which minimizes danger to population dense areas.

Results: Build a Dashboard with Plotly Dash



RESULTS: A Dashboard with Plotly Dash

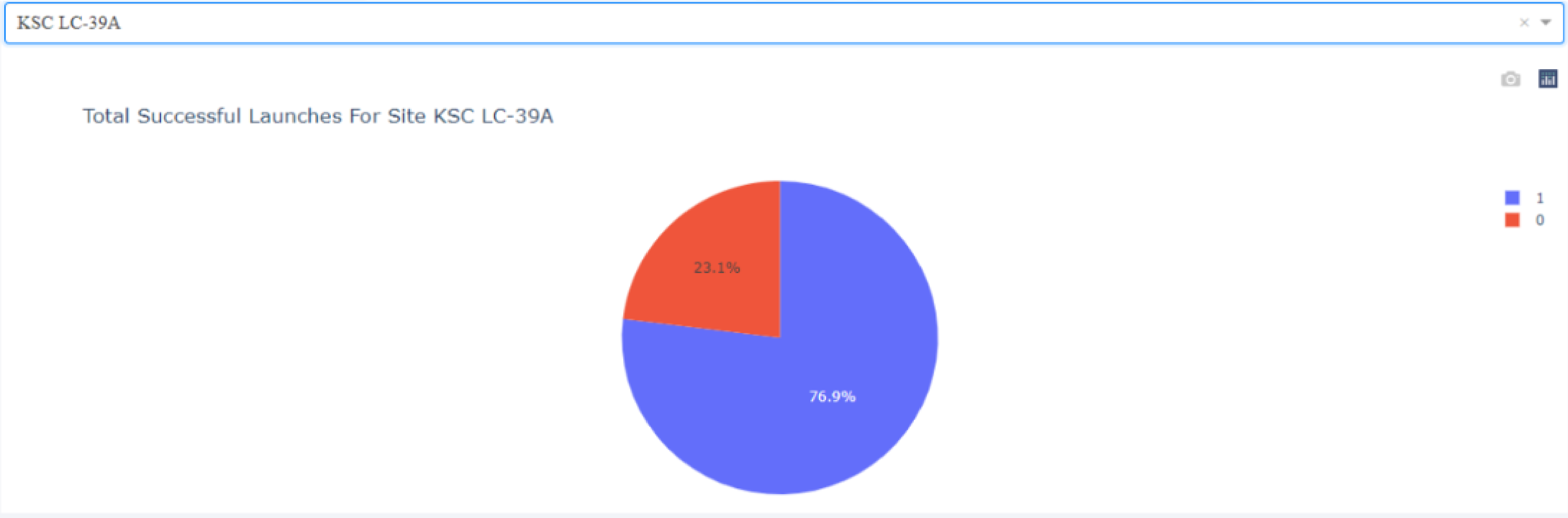


All Sites

Total Successful Launches By Site

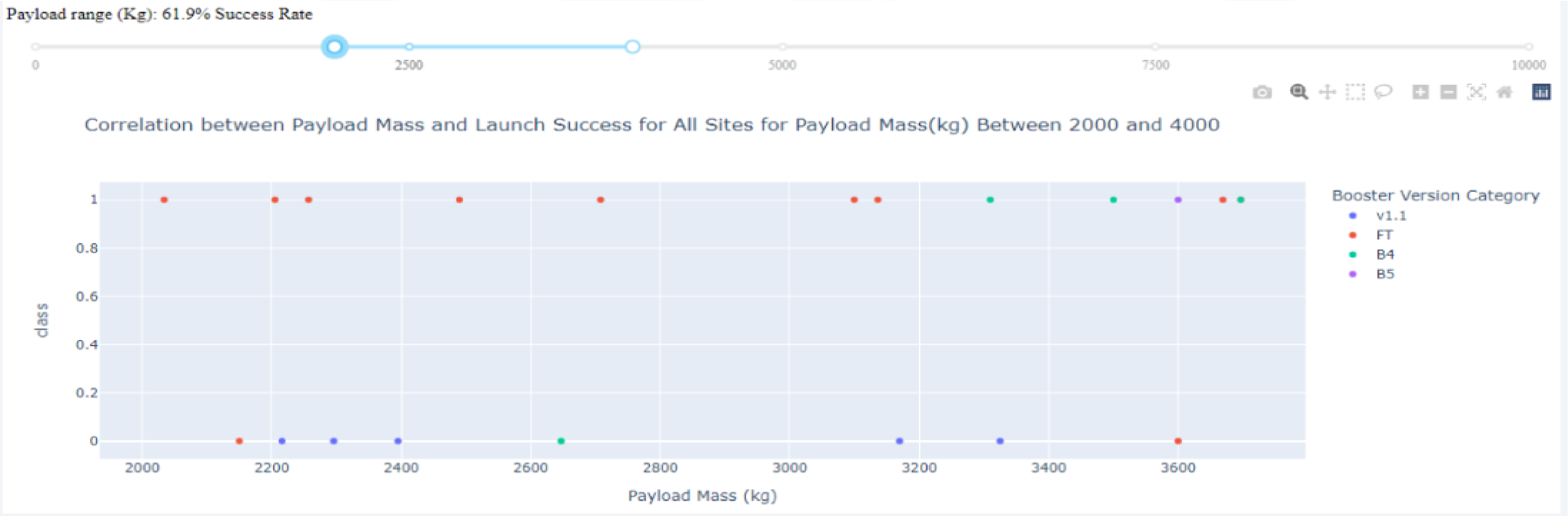
KSC LC-39A has the largest successful launches as well the highest launch success rate.

RESULTS: A Dashboard with Plotly Dash



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate This is the highest success rate of all the different launch sites. However, this success rate was only around 3% higher than the runner up; site CCAFS LC-40.

RESULTS: A Dashboard with Plotly Dash

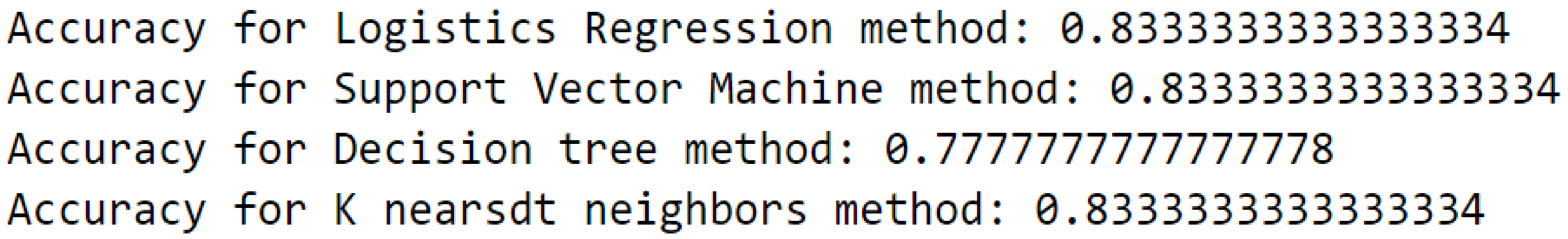

Payload range (Kg): 61.9% Success Rate

- The payload range between 2000 kg and 4000 kg has the highest success rate.
- The launch success rate was also dramatically low between the payload range of 0kg and 2500kg. Perhaps very low masses decrease launch success.
- The booster version FT has the highest launch success rate than other booster versions.

RESULTS: Predictive Analysis

RESULTS: Predictive Analysis (Classification)

Classification accuracy using training data

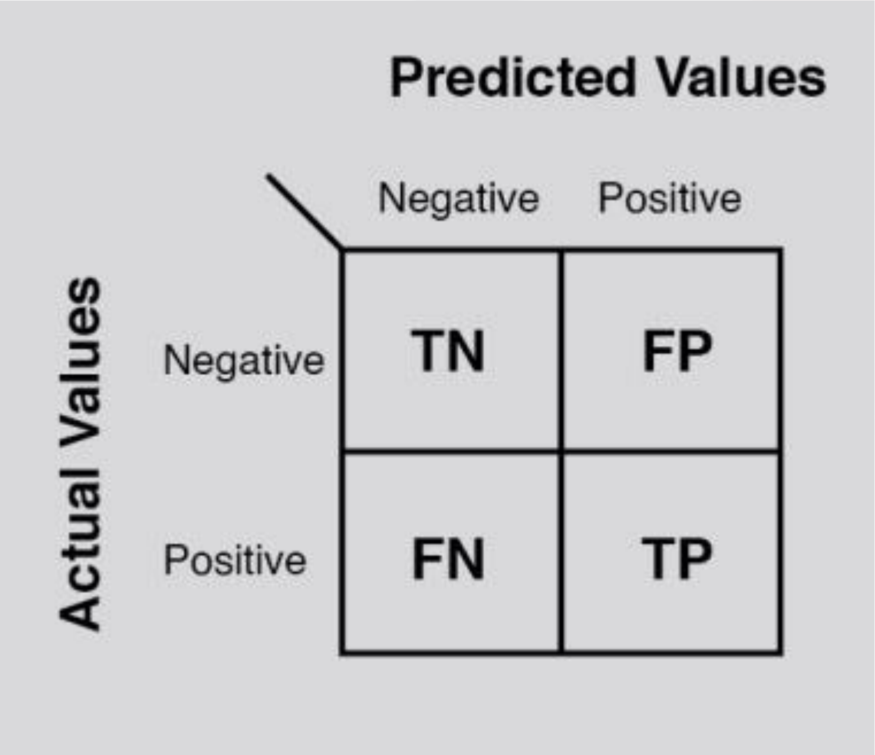


```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.7777777777777778
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

SVM, KNN and Logistic Regression models are the best in terms of prediction accuracy for this data set and all have an identical accuracy score of 83.33%.

RESULTS: Predictive Analysis (Classification)

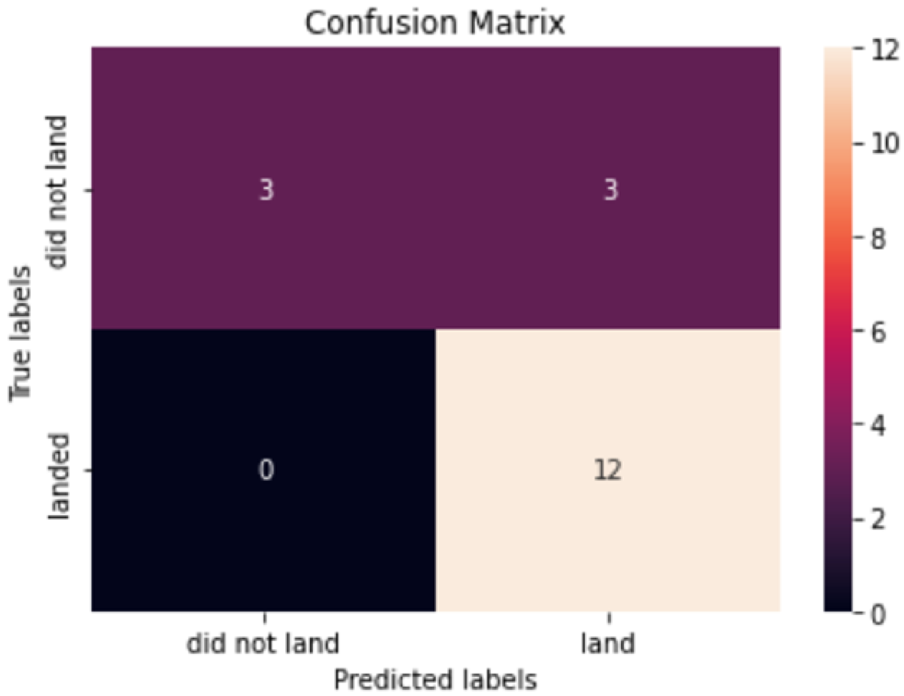
Confusion Matrix



Predicted Values

Actual Values

	Negative	Positive
Negative	TN	FP
Positive	FN	TP



Examining the confusion matrix for SVM, KNN and Logistic Regression models, we see that the major problem is false positives. The models only failed to accurately predict 3 labels.

CONCLUSION

In order to compete with SpaceX, it is crucial to analyze their data. Through the data analysis process, an overall picture of their success was created.

❑All launch sites are in close proximity to equator to minimize fuel consumption

❑All launch sites are located near the coast in order the crew has option to abort launch and attempt water landing, and minimize people and property at risk from falling debris.

❑All launch sites are located away from nearby cities which minimizes danger to population dense areas, but in close proximity to highways and railways for easy logistic.

❑Starting from 2013, the success rate of SpaceX landing has significantly increased possibly because of lessons learnt and technological progress. It is also obvious that the launch success has increased with the increase of the number of flights.

❑KSC LC-39A had the most successful launches from all sites.

❑The payload range between 2000 kg and 4000 kg had the highest success rate. Very low masses decrease the aunch success.

❑The booster version FT has the highest launch success rate than other booster versions.

❑Orbit GEO,HEO,SSO,ES-L1 had the best success rate.

❑All this data was used to train a machine learning models that are able to predict the landing outcome of rocket launches with 83.33% accuracy. The SVM, KNN and Logistic Regression models are the best in terms of prediction accuracy for this data set.

These outcomes will help our company to make more attractive offer than SpaceX.

Thank you for your attention!