Winning Space Race with Data Science

Mallikarjun Naik
01-Jan-2022

Author : Mallikarjun Naik

Outline

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

Executive Summary

Project background and context

- Data collection using API, Web scraping and SQL.
- Data wrangling, Loading and Analysis.
- EDA and interactive visual analytics.
- Predictive analysis methodology.

Summary of results

- Exploratory Data Analysis Results
- Interactive map with Folium results
- Plotly Dash dashboard results
- Predictive analysis (classification) results

Author : Mallikarjun Naik

Introduction

Project background and context

SpaceX advertises Falcon 9 rocket launches on its websites with a cost of 62 million dollars whereas other providers are charging 165 million dollars each. SpaceX can reuse the first stage. Due to this SpaceX can save considerable cost in rocket launches. Therefore, if we can determine whether first stage will land successfully. The gathered information can be used to determine if other companies can bid against SpaceX for a rocket launch.

Problems you want to find answers

- Use public information on SpaceX rocket launched to determine factors contributing to successfully landing of the rocket.
- Apply Data Science methodology with findings what data and information will lead to successful reuse of first stage of SpaceX rocket launches.

Author : Mallikarjun Naik

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- SpaceX Rest API
- Web Scraping from Wikipedia List of Falcon 9 and Falcon Heavy launches

Perform data wrangling

- One hot encoding data fields for machine learning and removing non applicable fields.

Perform exploratory data analysis (EDA) using visualization and SQL

- Scatter and Bar charts to visualize patterns between data

Perform interactive visual analytics

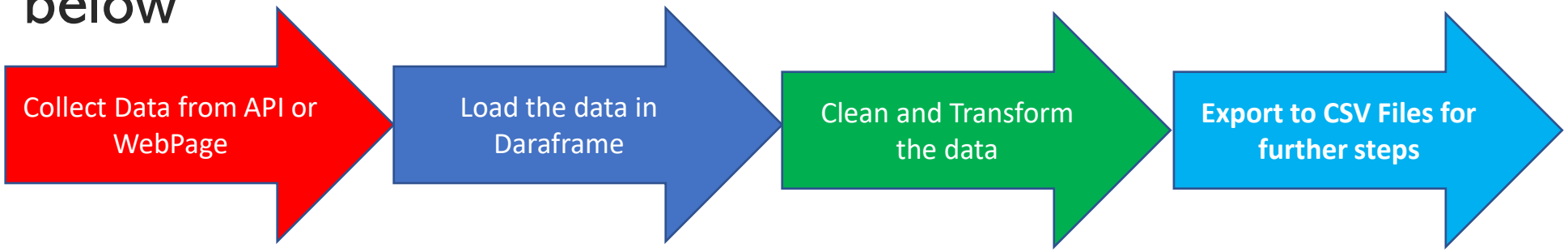
- Using Folium and Plotly Dash

Perform predictive analysis using classification models

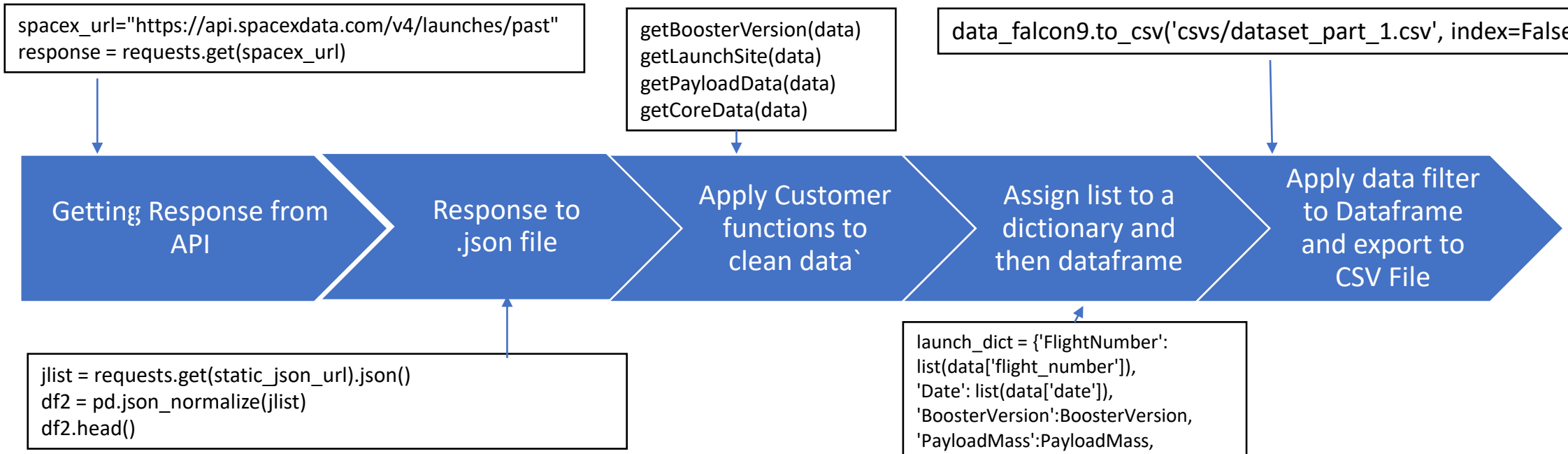
- How to build, tune, evaluate classification models

Data Collection

- Data was collected from SpaceX Rest API which provided information on launches, rocket used, Payload, launch specification and landing outcome.
- The Data can be collected via Web scraping from Wikipedia using BeautifulSoup
- The steps involved in data collection and Transformation goes as below



Data Collection – SpaceX API GituHu



GITHub URL

Author : Mallikarjun Naik

Data Collection – Web Scraping

Getting response from HTML

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

```
data = requests.get(static_url).text
```

Creating BeautifulSoup Object

```
soup = BeautifulSoup(data, 'html5lib')
```

Retrieving Tables

```
html_tables=soup.find_all("table")
 html_tables
```

Getting Column Names

```
column_names = []
ths = first_launch_table.find_all('th')
  for th in ths: name = extract_column_from_header(th)
        if name is not None and (name) > 0:
                column_names.append(name)
```

Creation of dictionary and appending data to keys

```
launch_dict= dict.fromkeys(column_names)
```

Converting and saving dictionary to dataframe

```
df = df.replace('\n','', regex=True)
df.head()
```

Saving the Dataframe in CSV format

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

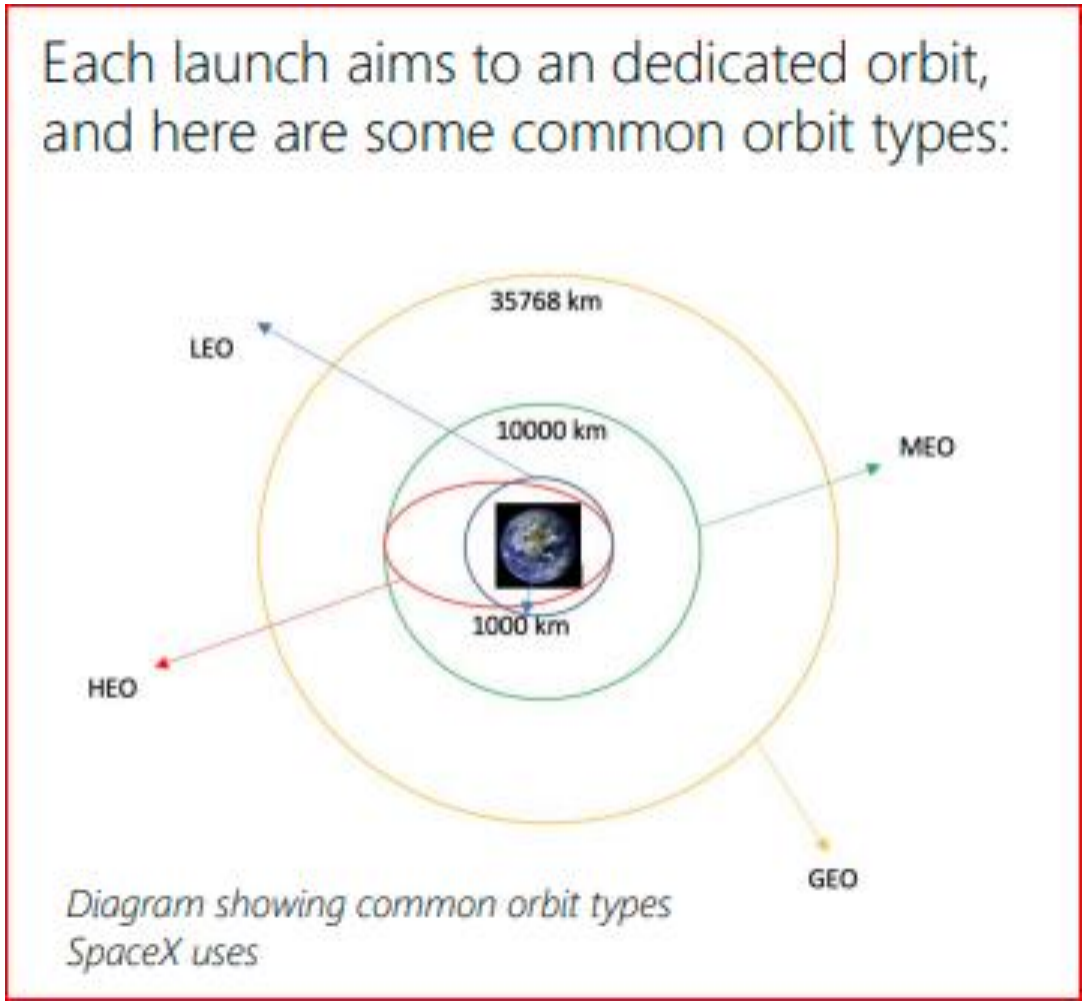
[GITHub Notebook URL]

Author : Mallikarjun Naik

Data wrangling

Data wrangling is the process of cleaning, structuring and enriching raw data into a desired format for better decision making in less time.

- Perform Exploratory Data Analysis EDA on dataset
- Calculate the number and launches at each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit site
- Export Dataset as CSV Type
- Create landing outcome label from Outcome column
- Workout success rate for every landing in dataset



EDA with Data Visualization

- Scatter Graphs being draw
  - Flight Number VS. Payload Mass
  - Flight Number VS. Launch Site
  - Payload VS. Launch Site
  - Orbit VS. Flight Number
  - Payload VS. Orbit Type
  - Orbit VS. Payload Mass

Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation . Scatter plots usually consist of a large body of data.

- Bar Graph being drawn:

Mean VS. Orbit

A bar diagram makes it easy to compare sets of data between different groups at a glance. The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes. Bar charts can also show big changes in data over time

- Line Graph being drawn:

Success Rate VS. Year

Line graphs are useful in that they show data variables and trends very clearly and can help to make predictions about the results of data not yet recorded

EDA with SQL

Below SQL queries are executed in the dataset :

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000
- but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_ versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_ outcomes in ground pad ,booster
- versions, launch_site for the months in year 2017
- Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

03-Jan-2022

Author : Mallikarjun Naik

Build an Interactive Map with Folium

To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.

We assigned the dataframe launch outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()

Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. Lines are drawn on the map to measure distance to landmarks

- Example of some trends in which the Launch Site is situated in.
- •Are launch sites in close proximity to railways? No
- •Are launch sites in close proximity to highways? No
- •Are launch sites in close proximity to coastline? Yes
- •Do launch sites keep certain distance away from cities? Yes

Build a Dashboard with Plotly Dash

- The dashboard is built with Flask and Dash web framework.
- Graphs
  - Pie Chart showing the total launches by a certain site/all sites
  - display relative proportions of multiple classes of data.
  - size of the circle can be made proportional to the total quantity it represents.

Scatter Graph showing 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 you a non-linear pattern.
  - The range of data flow, i.e. maximum and minimum value, can be determined.
  - Observation and reading are straightforward.

Predictive analysis (Classification)

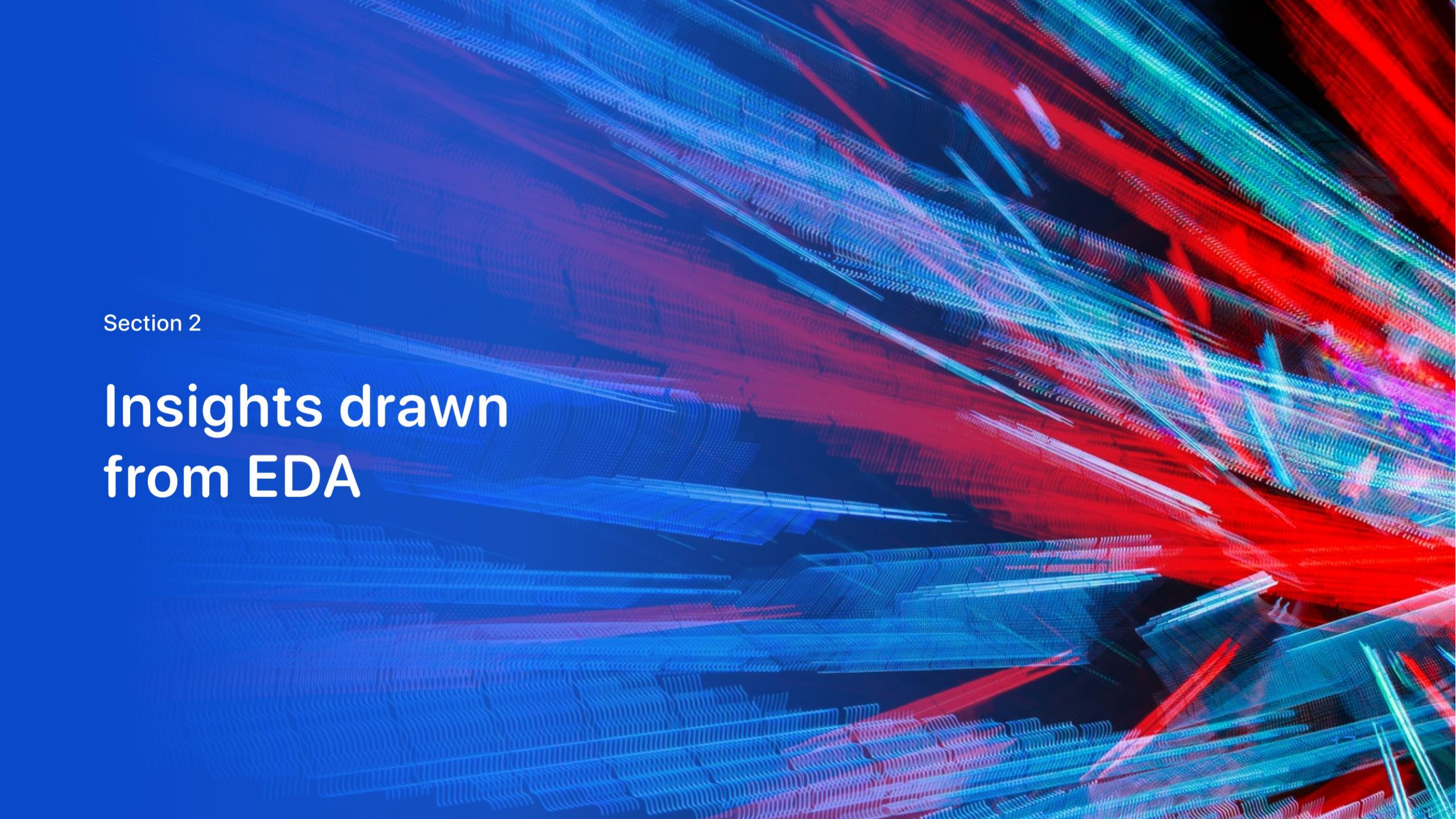
- BUILDING MODEL
  - Load our dataset into NumPy and Pandas
  - Transform Data
  - Split our data into training and test data sets
  - Check how many test samples we have
  - Decide which type of machine learning algorithms we want to use
  - Set our parameters and algorithms to GridSearchCV
  - Fit 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
- IMPROVING MODEL
  - Feature Engineering
  - Algorithm Tuning
- FINDING THE BEST PERFORMING CLASSIFICATION MODEL
  - The model with the best accuracy score wins the best performing model
  - In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook.

Author : Mallikarjun Naik

Results

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

Author : Mallikarjun Naik



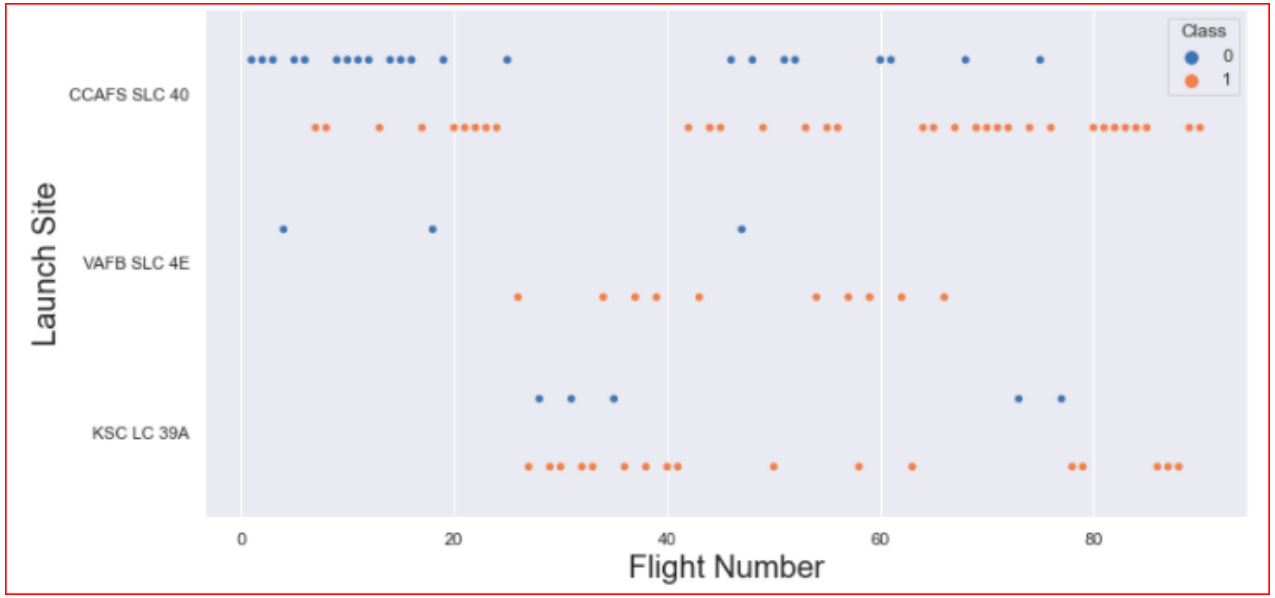
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- **With more flight numbers (after 40) higher the success rate for the Rocket is increasing**.
- *But theres no clear pattern to make a decision if the Flight Number is dependant on Launch Site for a success launch.*

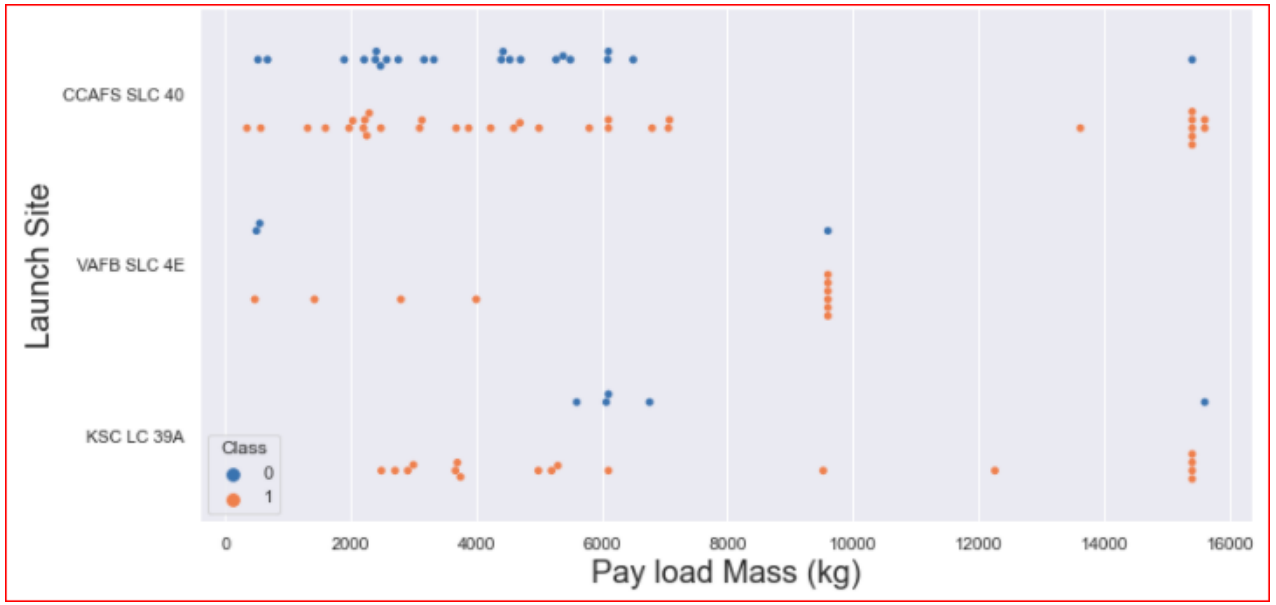
GITHub Notebook URL



Author : Mallikarjun Naik

Payload vs. Launch Site

- The **greater the payload mass (greater than 8000) higher the success rate for the Rocket**. *But theres no clear pattern to make a decision if the Launch Site is dependant on Pay Load Mass for a success launch.*

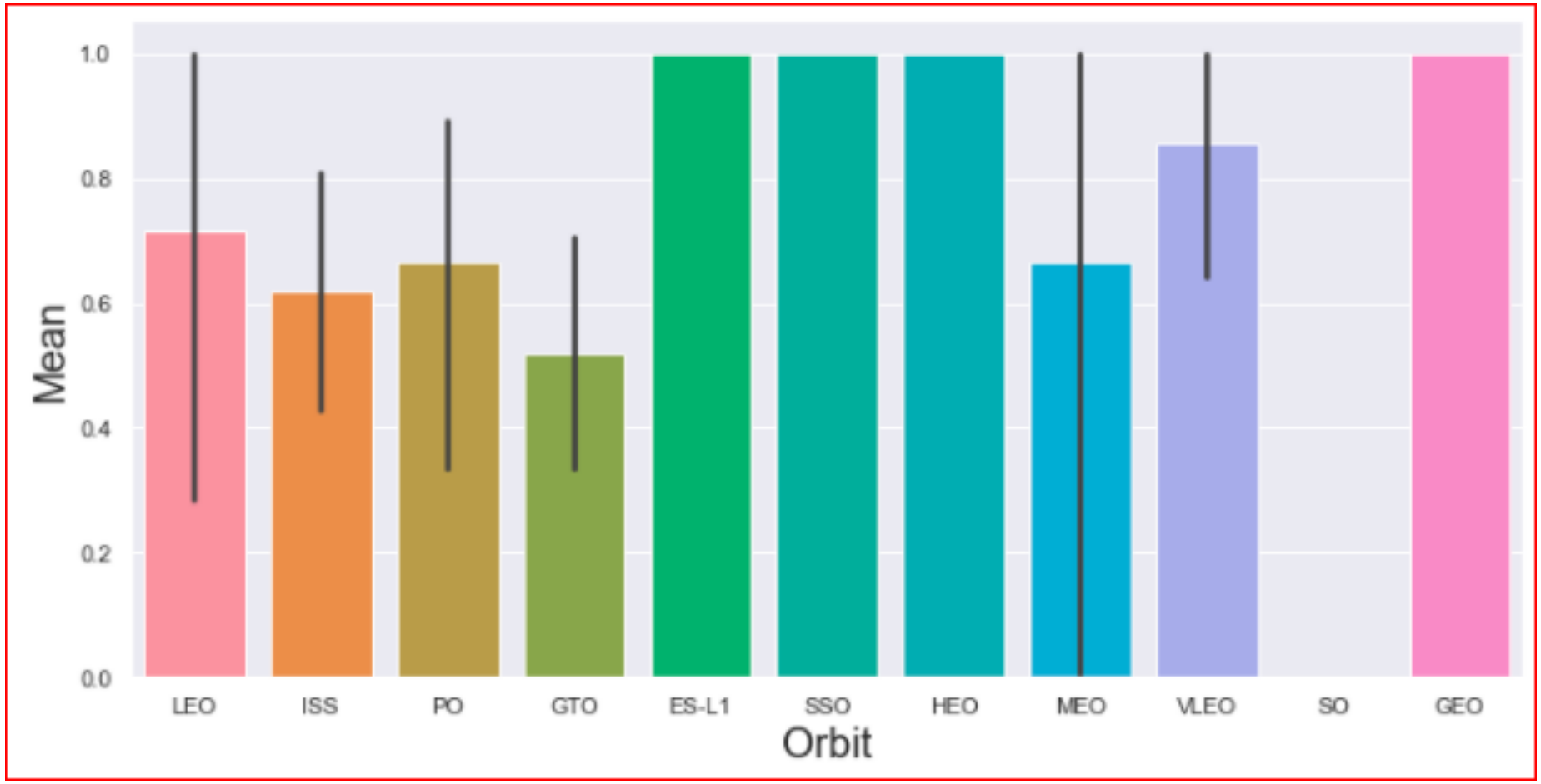


[GITHub Notebook URL]

Author : Mallikarjun Naik

Success Rate vs. Orbit Type

ES-L1, GEO, HEO, SSO has highest Sucess rates. *SO has poorest.*

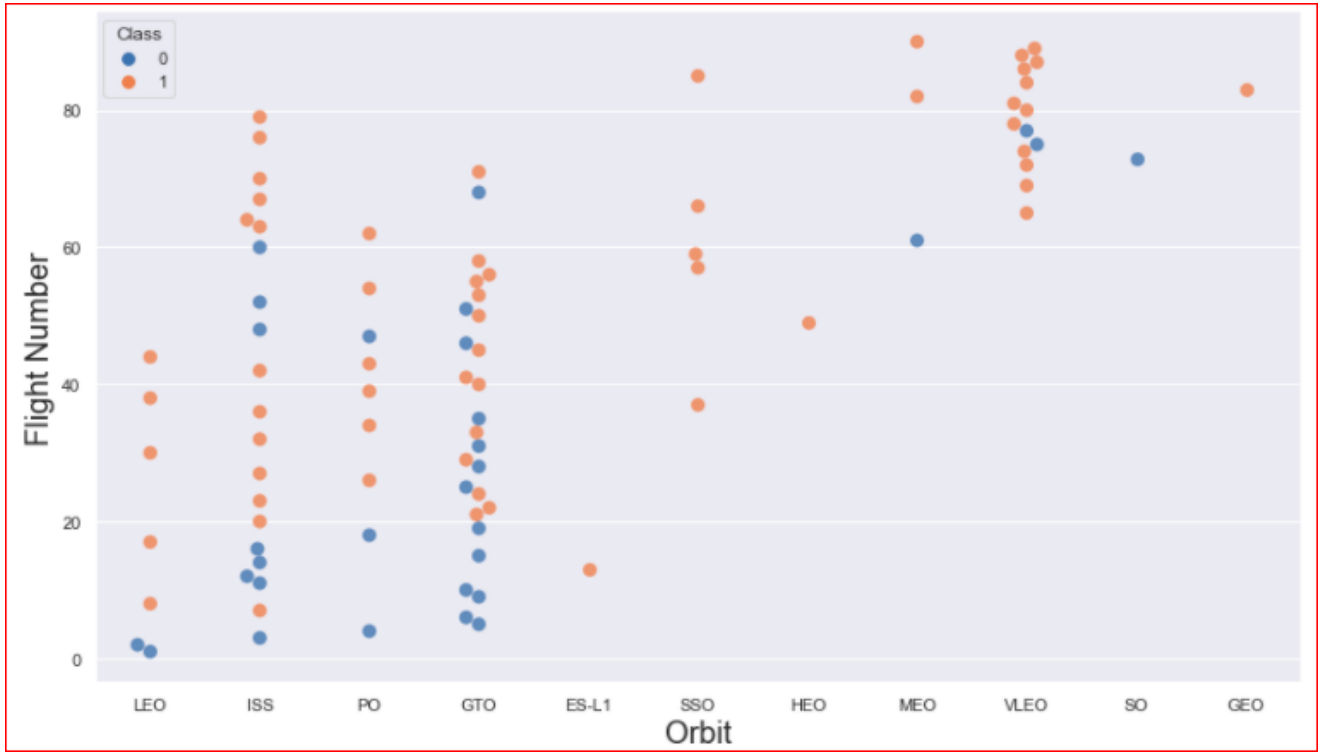


Author : Mallikarjun Naik

GITHub Notebook URL

Flight Number vs. Orbit Type

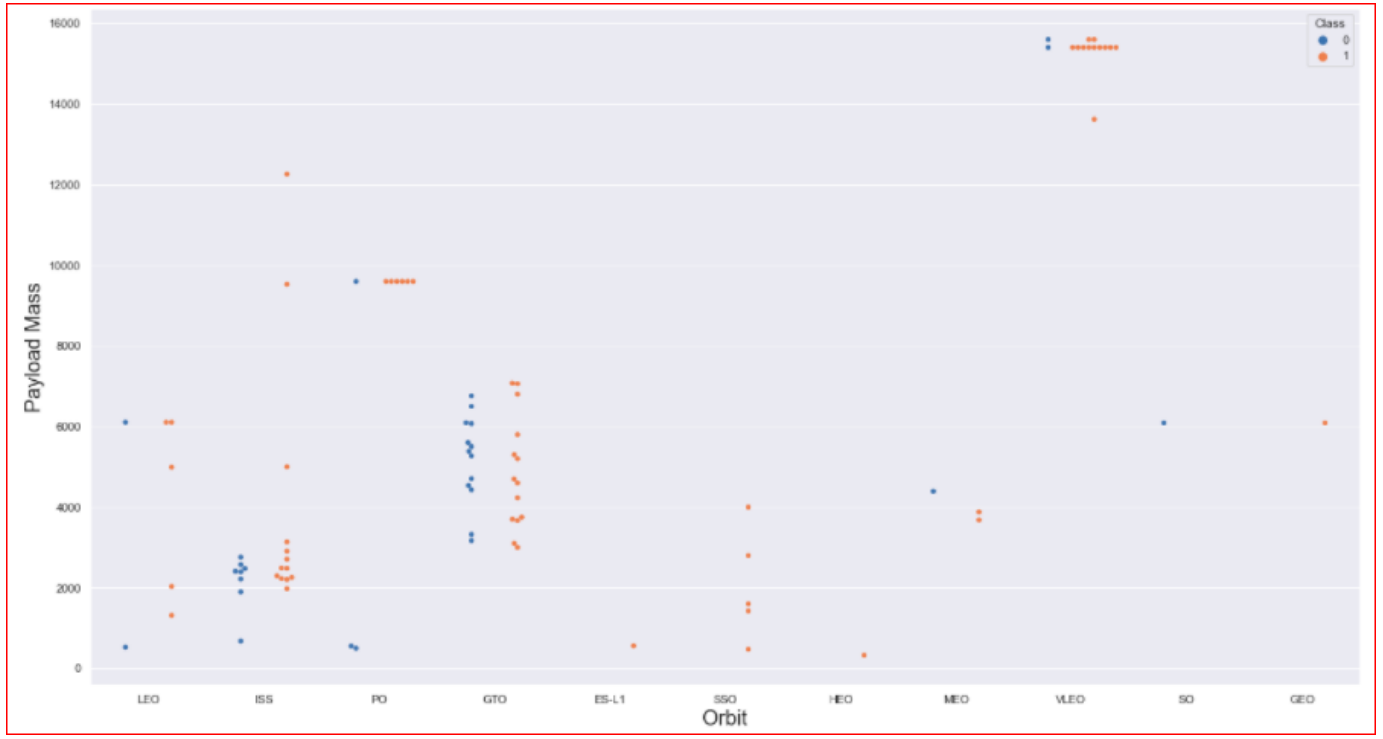
- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



GITHub Notebook URL

Payload vs. Orbit Type

- You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



GITHub Notebook URL

You can observe that the success rate since 2013 kept increasing till 2020

Launch Success Yearly Trend

GITHub Notebook URL

Author : Mallikarjun Naik

03-Jan-2022

Unique Launch Sites

- SQL Query
  - %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEX;
- Description

  Using the word DISTINCT in the query means

  that it will only show Unique values

  in the Launch_Site column from tblSpaceX

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

GITHub Notebook URL

Launch Site Names Begin with 'CCA'

- SQL Query
  - %sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
- Query Description
  - Using the word TOP 5 in the query means that it will only show 5 records from tblSpaceX and LIKE keyword has a wild card with the words ‘KSC%’ the percentage in the end suggests that the Launch_Site name must start with KSC

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

GITHub Notebook URL

Total Payload Mass

- SQL
  - %sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload Mass by NASA (CRS)" FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)';
- Description
  - QUERY EXPLAINATION Using the function SUM summates the total in the column PAYLOAD_MASS_KG_ The WHERE clause filters the dataset to only perform calculations on Customer NASA (CRS)

Total Payload Mass by NASA (CRS)
45596

GITHub Notebook URL

Average Payload Mass by F9 v1.1

- SQL Query

  %sql SELECT MIN(DATE) AS "First Succesful Landing Outcome in Ground Pad" FROM SPACEX \

  WHERE LANDING__OUTCOME = 'Success (ground pad)';

- Description
  - Using the function AVG works out the average in the column PAYLOAD_MASS_KG_ The WHERE clause filters the dataset to only perform calculations on Booster_version F9 v1.1

Average Payload Mass by Booster Version F9 v1.1

2928

GITHub Notebook URL

First Successful Ground Landing Date

- SQL Query
  - %sql SELECT MIN(DATE) AS "First Successful Landing Outcome in Ground Pad" FROM SPACEX \
  - WHERE LANDING__OUTCOME = 'Success (ground pad)';
- Description
  - QUERY EXPLAINATION Using the function MIN works out the minimum date in the column Date The WHERE clause filters the dataset to only perform calculations on Landing_Outcome Success

First Successful Landing Outcome in Ground Pad

2015-12-22

GITHub Notebook URL

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL Query

```
%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

- Description

Selecting only Booster_Version The WHERE clause filters the dataset to Landing_Outcome = Success (drone ship) The AND clause specifies additional filter conditions Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000

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

GITHub Notebook URL

Total Number of Successful and Failure Mission Outcomes

- SQL Queries
  - %sql SELECT COUNT(MISSION_OUTCOME) AS "Successful Mission" FROM SPACEX WHERE MISSION_OUTCOME LIKE 'Success%’;
  - %sql SELECT COUNT(MISSION_OUTCOME) AS "Failure Mission" FROM SPACEX WHERE MISSION_OUTCOME LIKE 'Failure%';
- Description

  There are 2 queries being retrieved to get the below results

Successful Mission	Failure Mission
100	1

GITHub Notebook URL

Boosters Carried Maximum Payload

- SQL Query

  %sql SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" FROM SPACEX \

  WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX);

- Description
  - Using the word DISTINCT in the query means that it will only show Unique values in the Booster_Version column from tblSpaceX GROUP BY puts the list in order set to a certain condition. DESC means its arranging the dataset into descending order

GITHub Notebook URL

Booster Versions which carried the Maximum Payload Mass
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- SQL Query

  %sql SELECT {fn MONTHNAME(DATE)} as "Month", BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE year(DATE) = '2015' AND \

  LANDING__OUTCOME = 'Failure (drone ship)';

- Description

  a much more complex query as I had my Date fields in SQL Server stored as NVARCHAR the MONTH function returns name month.

Month	booster_version	launch_site
January	F9 v1.1 B1012	CCAFS LC-40
April	F9 v1.1 B1015	CCAFS LC-40

GITHub Notebook URL

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

- SQL Query

```
%sql SELECT LANDING__OUTCOME as "Landing Outcome", COUNT(LANDING__OUTCOME) AS "Total Count" FROM SPACEX \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY  LANDING__OUTCOME \
ORDER BY COUNT(LANDING__OUTCOME) DESC ;
```

- Description

  Lists failure and Success landing outcomes

Landing Outcome	Total Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

GITHub Notebook URL

Section 4

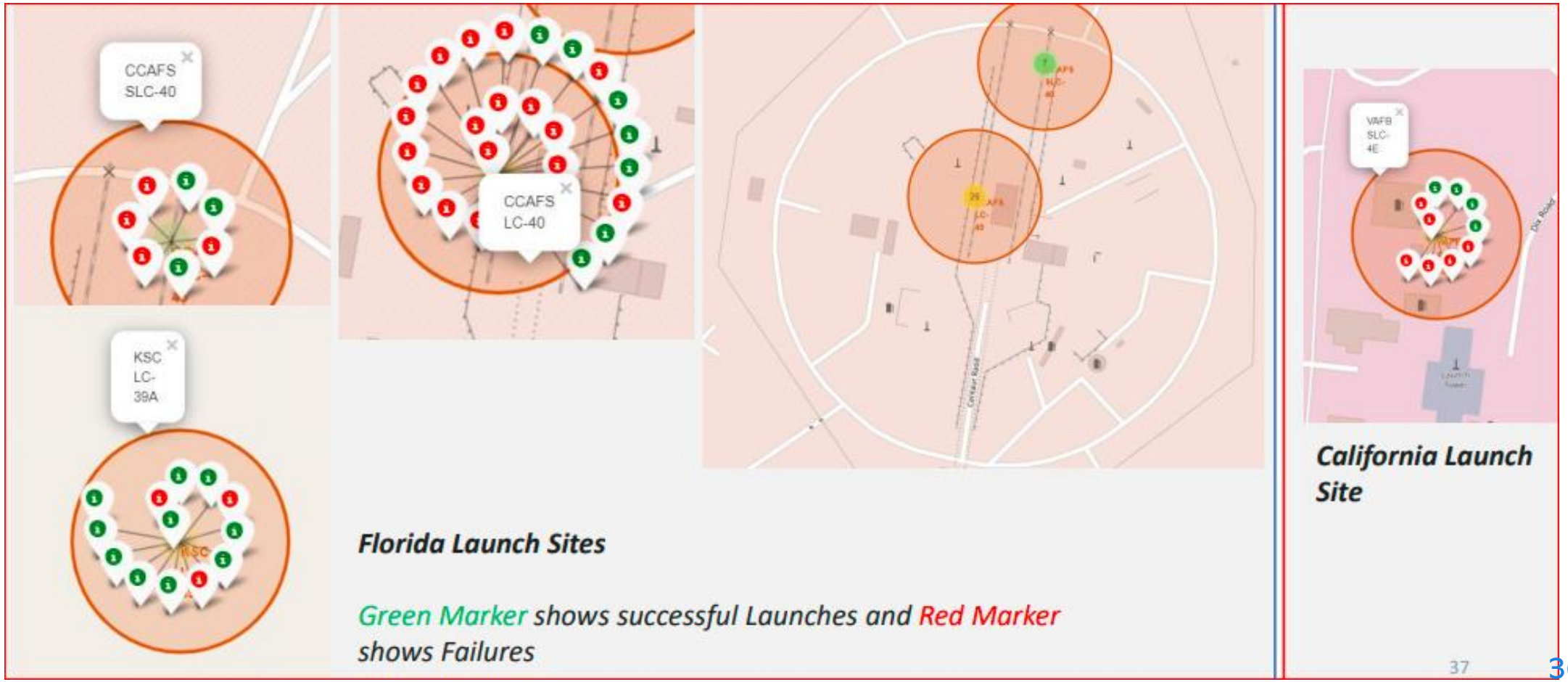
Launch Sites Proximities Analysis

All launch sites global map markers

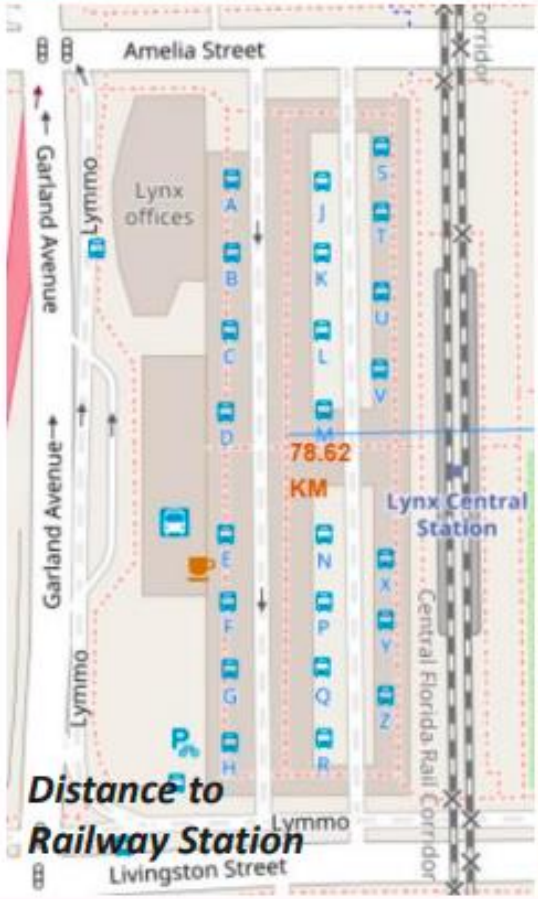
GITHub Notebook URL

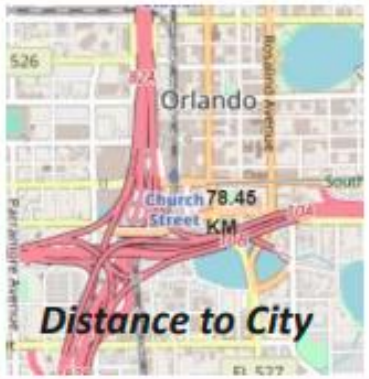
We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

Colour Labelled Markers



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

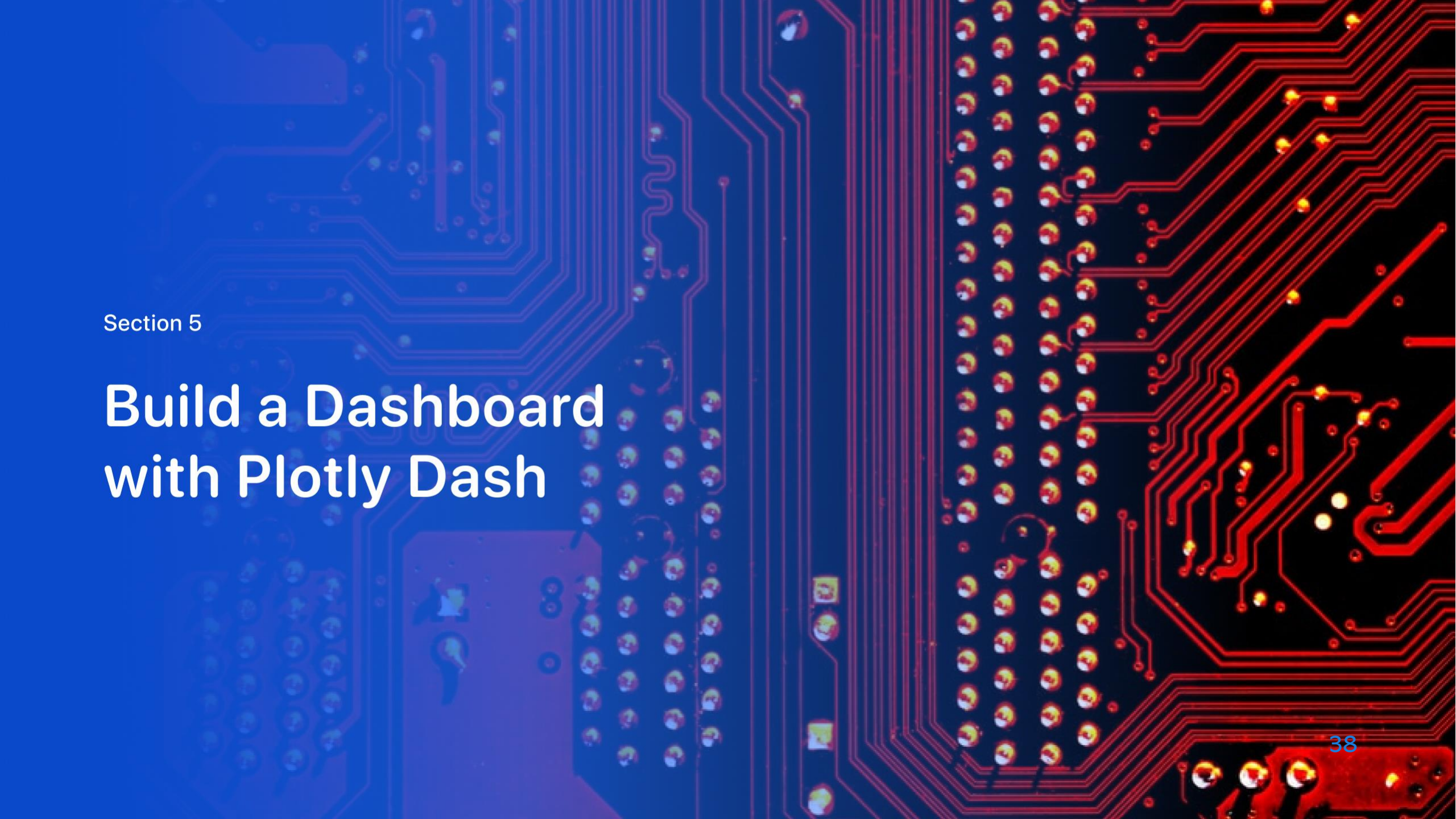




- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Author : Mallikarjun Naik

GITHub Notebook URL

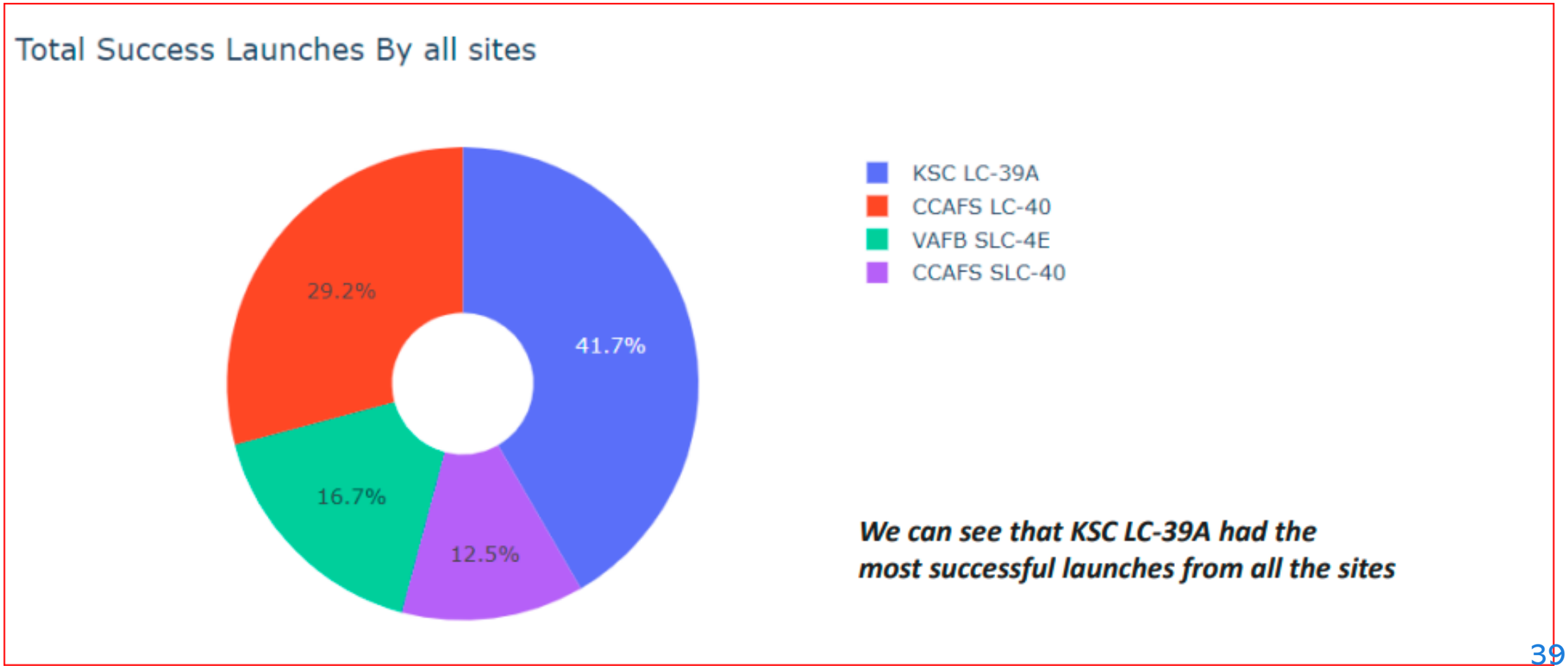


Section 5

Build a Dashboard with Plotly Dash

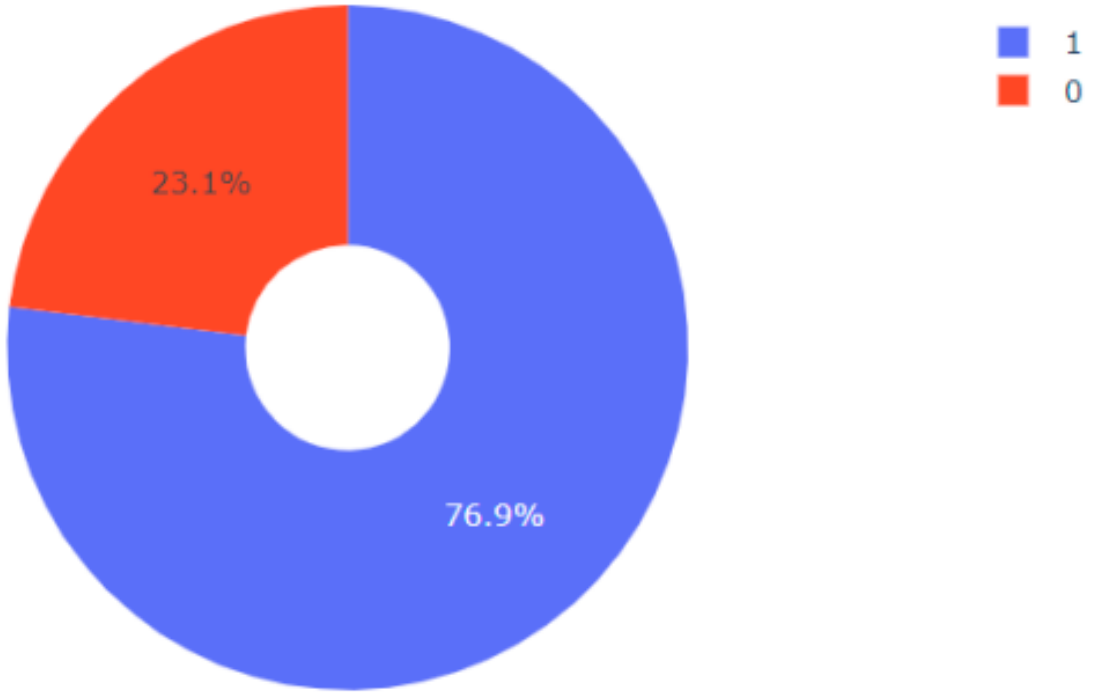
DASHBOARD – Pie chart showing the success percentage achieved by each launch site

GITHub Notebook URL



DASHBOARD – Pie chart for the launch site with highest launch success ratio

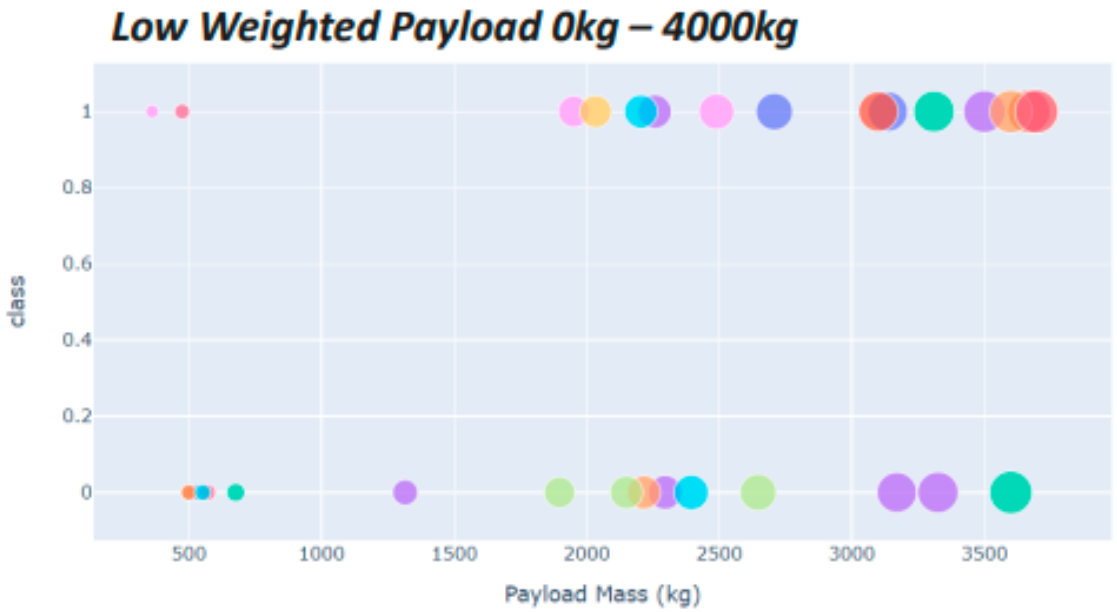
GITHub Notebook URL

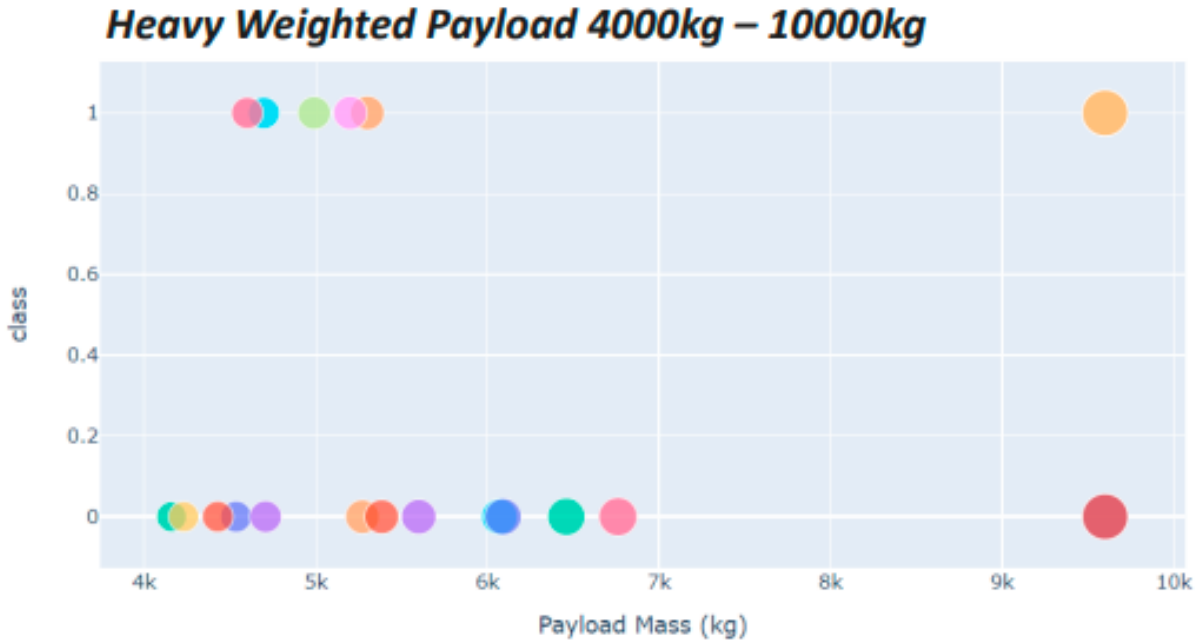


KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

DASHBOARD – Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

GITHub Notebook URL

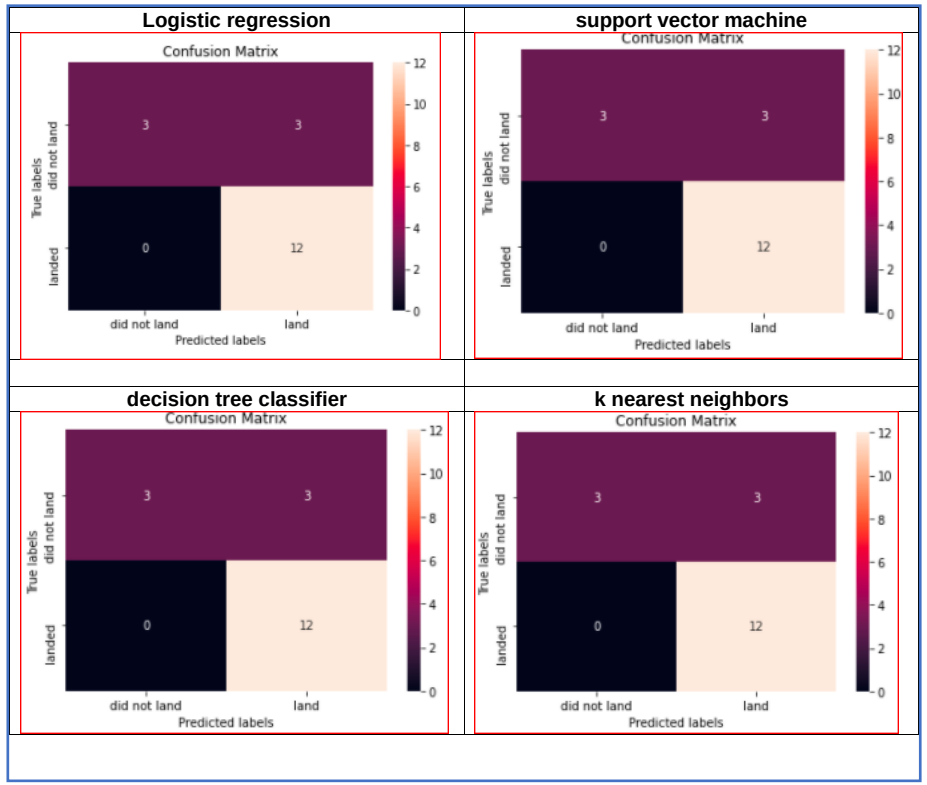




We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

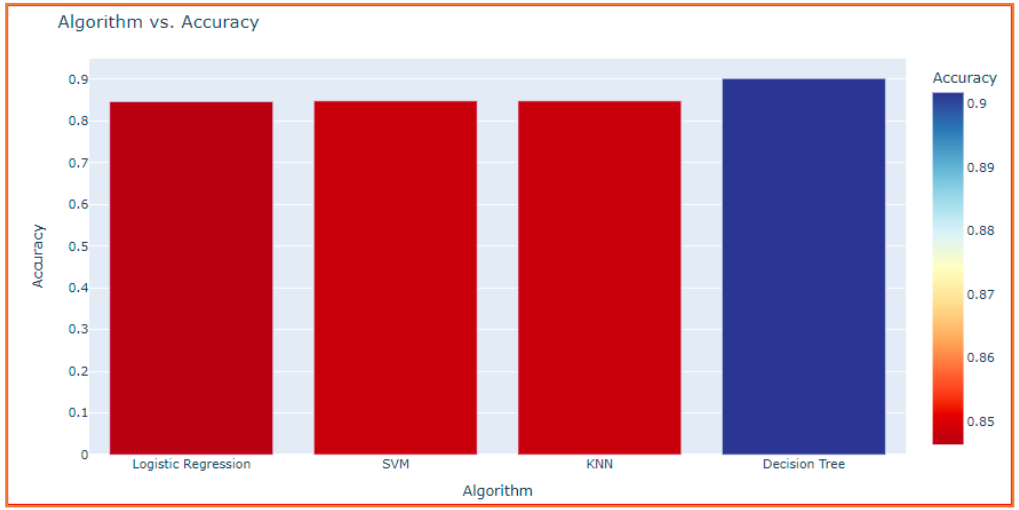
Predictive Analysis (Classification)

Classification Accuracy



- As you can observe accuracy of all the algorithms are vey close but the most accurate down to decimal point is Decision Tree Algorithm

Sl No	Algorithm	Accuracy
0	Logistic Regression	0.846429
1	SVM	0.848214
2	KNN	0.848214
3	Decision Tree	0.901786



Conclusion

THE TREE CLASSIFIER ALGORITHM IS THE BEST FOR MACHINE LEARNING FOR THIS DATASET

LOW WEIGHTED PAYLOADS PERFORM BETTER THAN THE HEAVIER PAYLOADS

THE SUCCESS RATES FOR SPACEX LAUNCHES IS DIRECTLY PROPORTIONAL TIME IN YEARS THEY WILL EVENTUALLY PERFECT THE LAUNCHES

WE CAN SEE THAT KSC LC-39A HAD THE MOST SUCCESSFUL LAUNCHES FROM ALL THE SITES

ORBIT GEO,HEO,SSO,ES-L1 HAS THE BEST SUCCESS RATE

Author : Mallikarjun Naik

Appendix

Python Anywhere

Introduction

PythonAnywhere makes it easy to create and run Python programs in the cloud. You can write your programs in a web-based editor or just run a console session from any modern web browser. There's storage space on our servers, and you can preserve your session state and access it from anywhere, with no need to pay for, or configure, your own server. Start work on your work desktop, then later pick up from where you left off by accessing exactly the same session from your laptop.

Your website

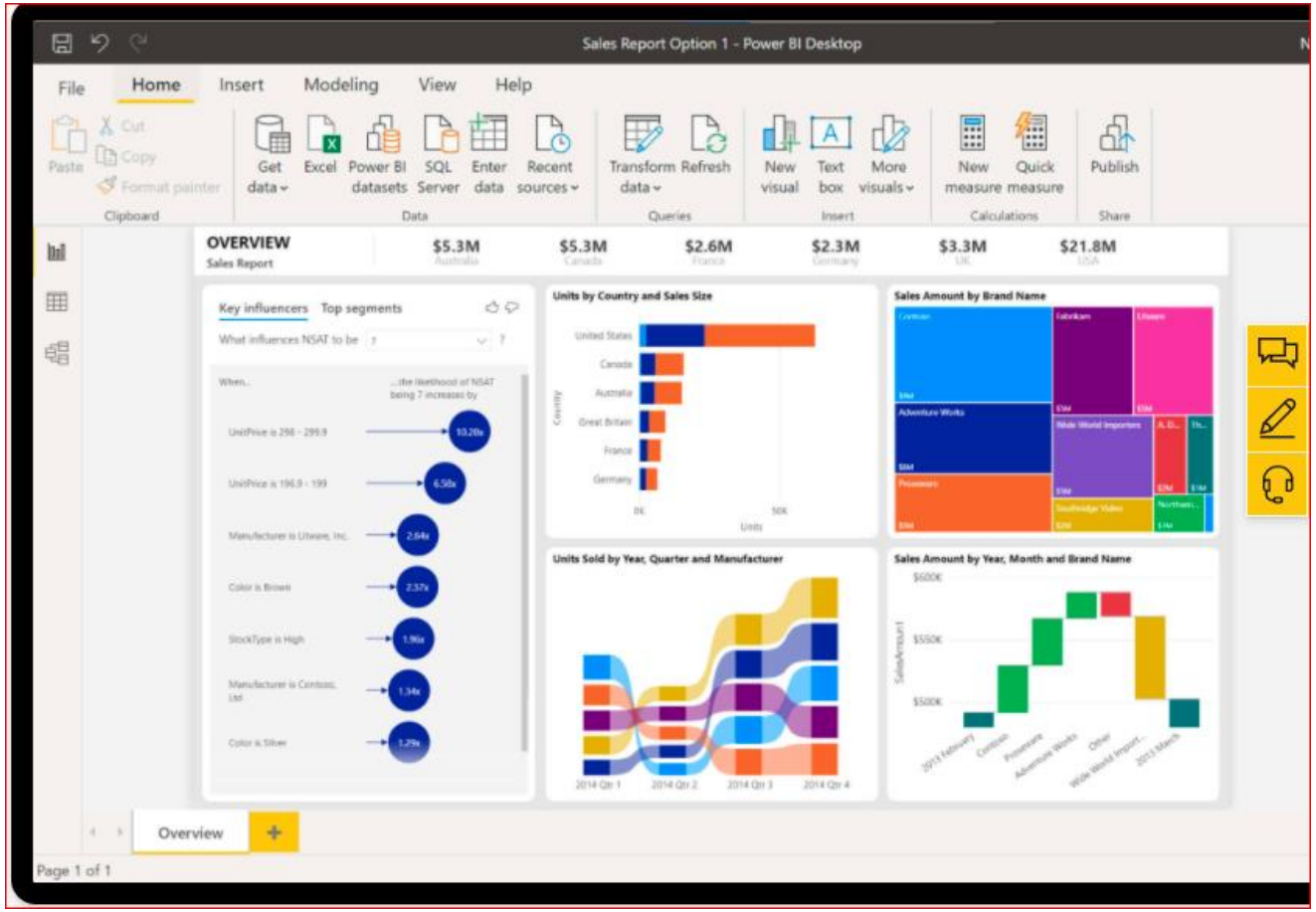
Want to host your own domain at PythonAnywhere? Our paid accounts do that for you. And free users don't get left out — http://yourusername.pythonanywhere.com/ works for everyone.

Easy scaling

A $5 Hacker account at PythonAnywhere can easily support a 10,000 hit/day website. But when your site grows and you need to support 100 times that traffic, we're still there — you just need to upgrade your account.

Microsoft Power BI- A great Business Intelligent Tool…

- **Get self-service analytics at enterprise scale**
- Reduce the added cost, complexity, and security risks of multiple solutions with an analytics platform that scales from individuals to the organization as a whole.
- **Use smart tools for strong results**
- Find and share meaningful insights with hundreds of data visualizations, built-in AI capabilities, tight Excel integration, and pre-built and custom data connectors.
- **Help protect your analytics data**
- Keep your data secure with industry-leading data security capabilities including sensitivity labeling, end-to-end encryption, and real-time access monitoring.



Thank you!

Author : Mallikarjun Naik