Data Science Capstone project

Zubair Kaif

16 August 2021

Outline

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

Executive Summary

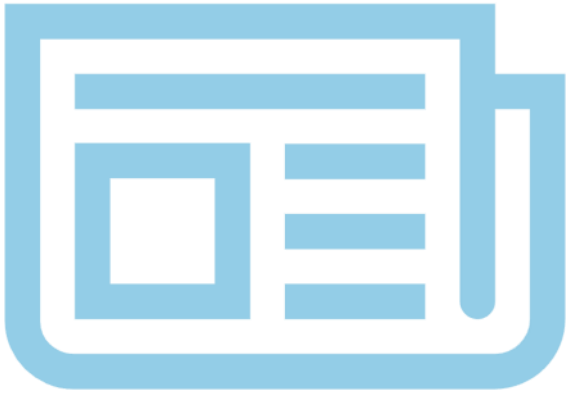
- Summary of methodologies:
- **Data Collection/Data Wrangling**
- **Exploratory Data Analysis (EDA)**
- **Data visualization**
- **Predictive Analysis (Classification)**

- Summary of all results:
- **In this project we use data to find out whether the first stage of falcon 9 rocket land successfully or not using machine learning algorithm like SVM, Classification Trees, and Logistic Regression.**

Introduction

- Project background and context:
- **Falcon 9 rocket of SpaceX is most cost effective rocket (cost around 65 million) SpaceX reuse its first stage. So if we can predict the success rate of landing of first stage then we can use these results to rival SpaceX and can help companies build cheaper rockets.**
- Problems you want to find answers:
- **Success rate of Falcon 9 rocket first stage landing.**

Methodology



- Data collection methodology:
  - Describe how data were collected
- Perform data wrangling
  - Describe how data were processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Methodology

Data collection

- Data sets were collected from SpaceX api.

1. url = “https://api.spacexdata.com/v4/launches/past”

2. Perform a get request.

3. View result by calling Json() method on response.

Data collection – SpaceX API

Added a flowchart of SpaceX API calls here

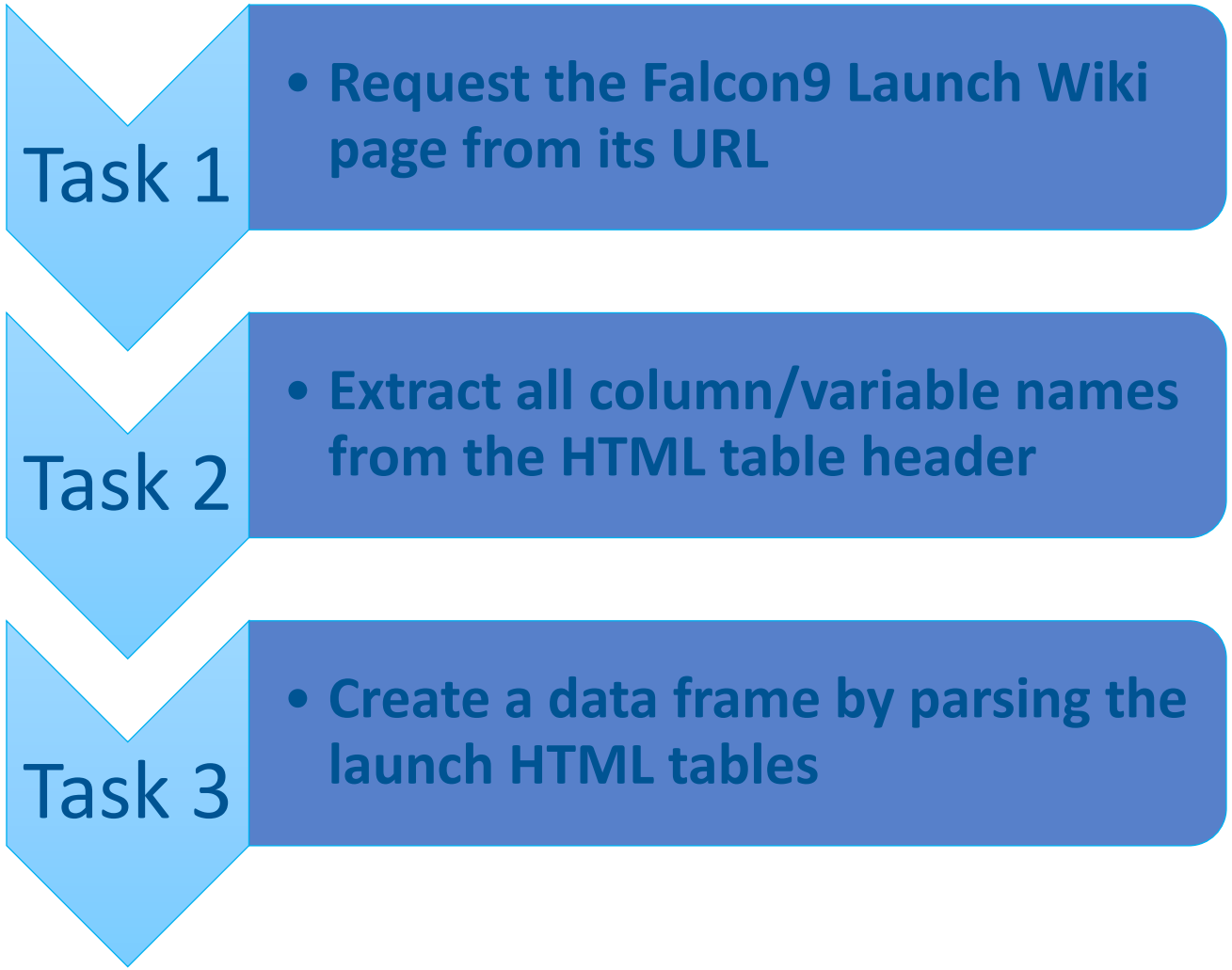
Github link:
https://github.com/zubairKaif/Coursera-assignment/blob/8bd87dad71ee4e1e21b863029bbf263e038976a1/Data%20Collection%20API%20Lab.ipynb

Data collection – Web scraping

Objectives

Web scrap Falcon 9 launch records with BeautifulSoup:

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame



GitHub URL:https://github.com/zubairKaif/Coursera-assignment/blob/0aad3fd81411d18f943765f606ff633e3b7fcab8/Data%20wrangling.ipynb

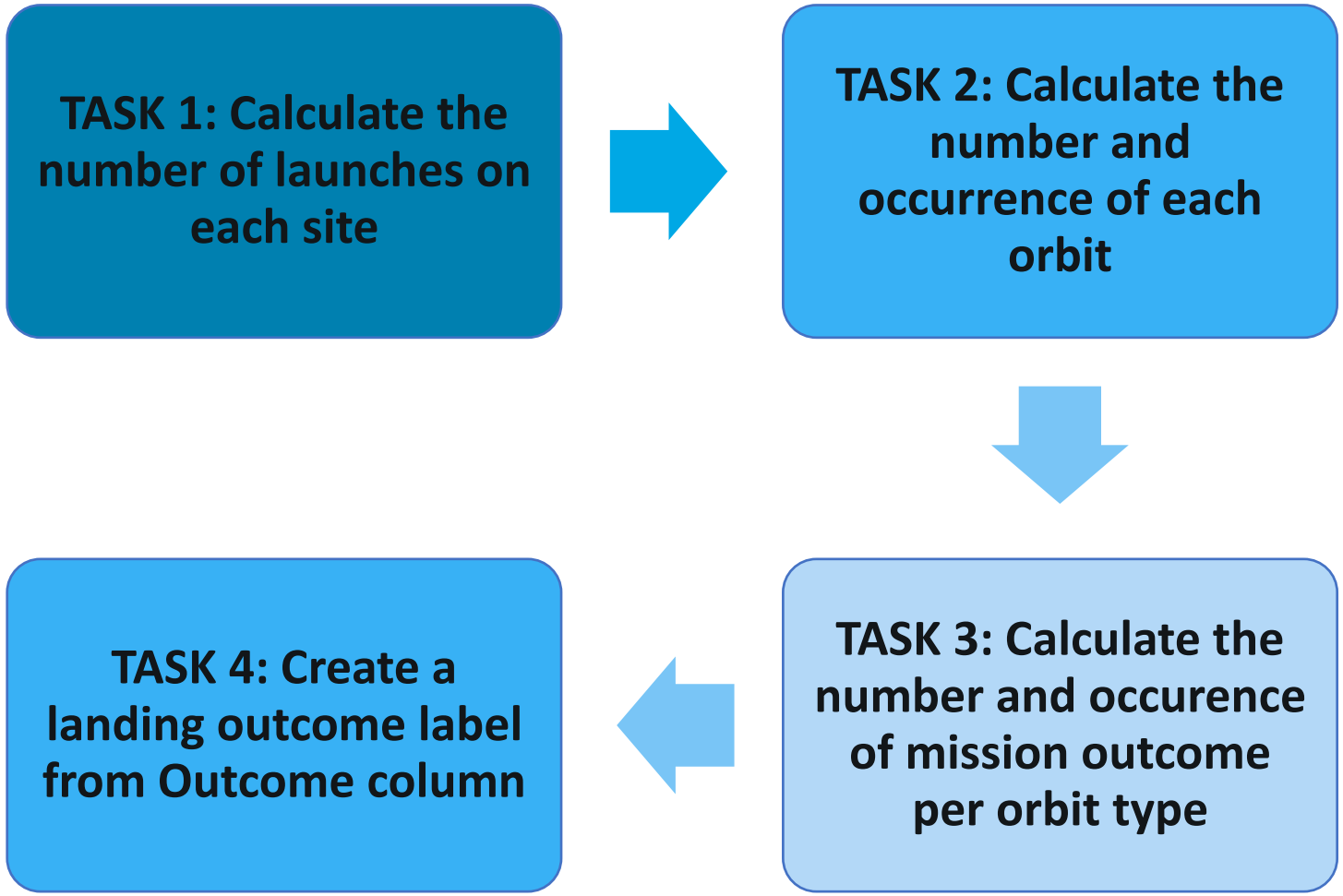
Data wrangling

In this lab, we had perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models

Objectives

Perform exploratory Data Analysis and determine Training Labels

- Exploratory Data Analysis
- Determine Training Labels
- Describe how data were processed



https://github.com/zubairKaif/Coursera-assignment/blob/0fcf48535175e6ad4f6fcc469885035d9c1faa5c/Data%20wrangling.ipynb

EDA with data visualization

- In this lab, I had performed Exploratory Data Analysis and Feature Engineering.

- Objectives

Perform exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib

Exploratory Data Analysis

Preparing Data Feature Engineering

GitHub Url:https://github.com/zubairKaif/Coursera-assignment/blob/dd81ff244b8a34047bb58401d5fd1af602e75157/EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- Summarize performed SQL queries using bullet points

- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Build an interactive map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard

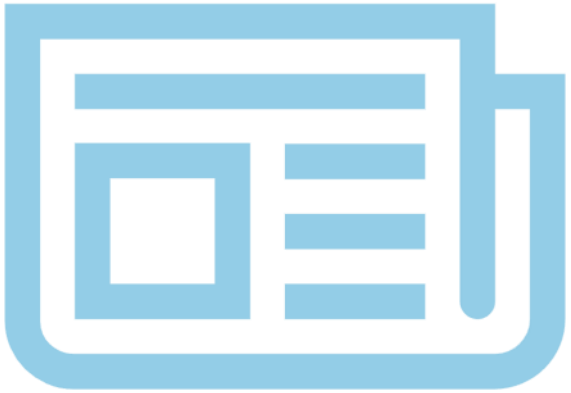
- Explain why you added those plots and interactions

- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive analysis (Classification)

- 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

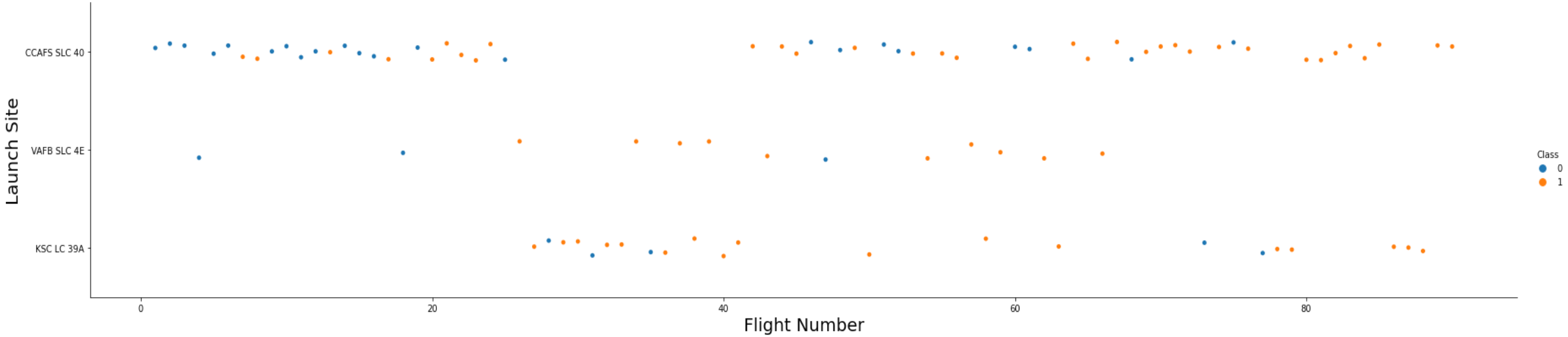
Results



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

EDA with Visualization

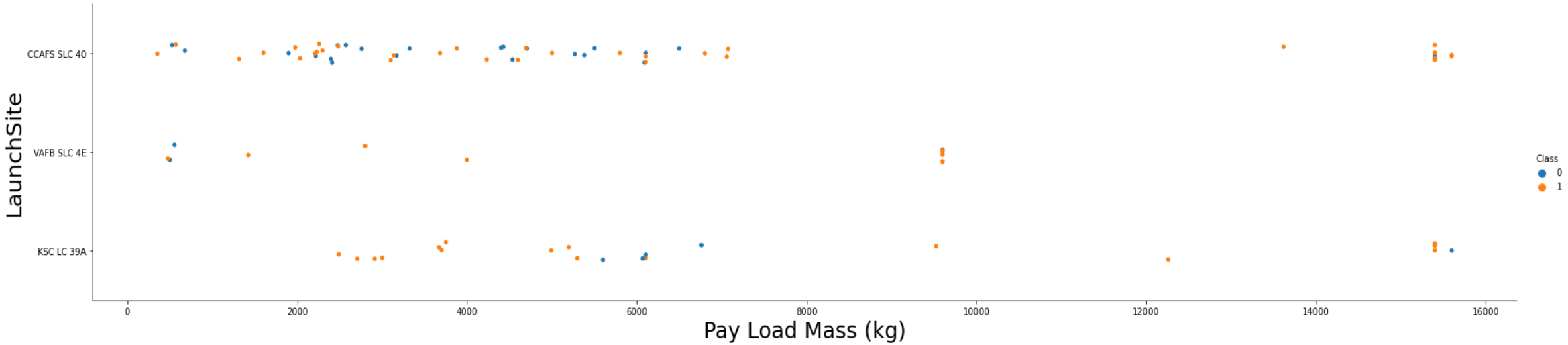
Flight Number vs. Launch Site



scatter plot of Flight Number vs. Launch Site

Explanation: Most of the flights and successful flights are launced from lauch site ‘CCAPS SLC 40’

Payload vs. Launch Site



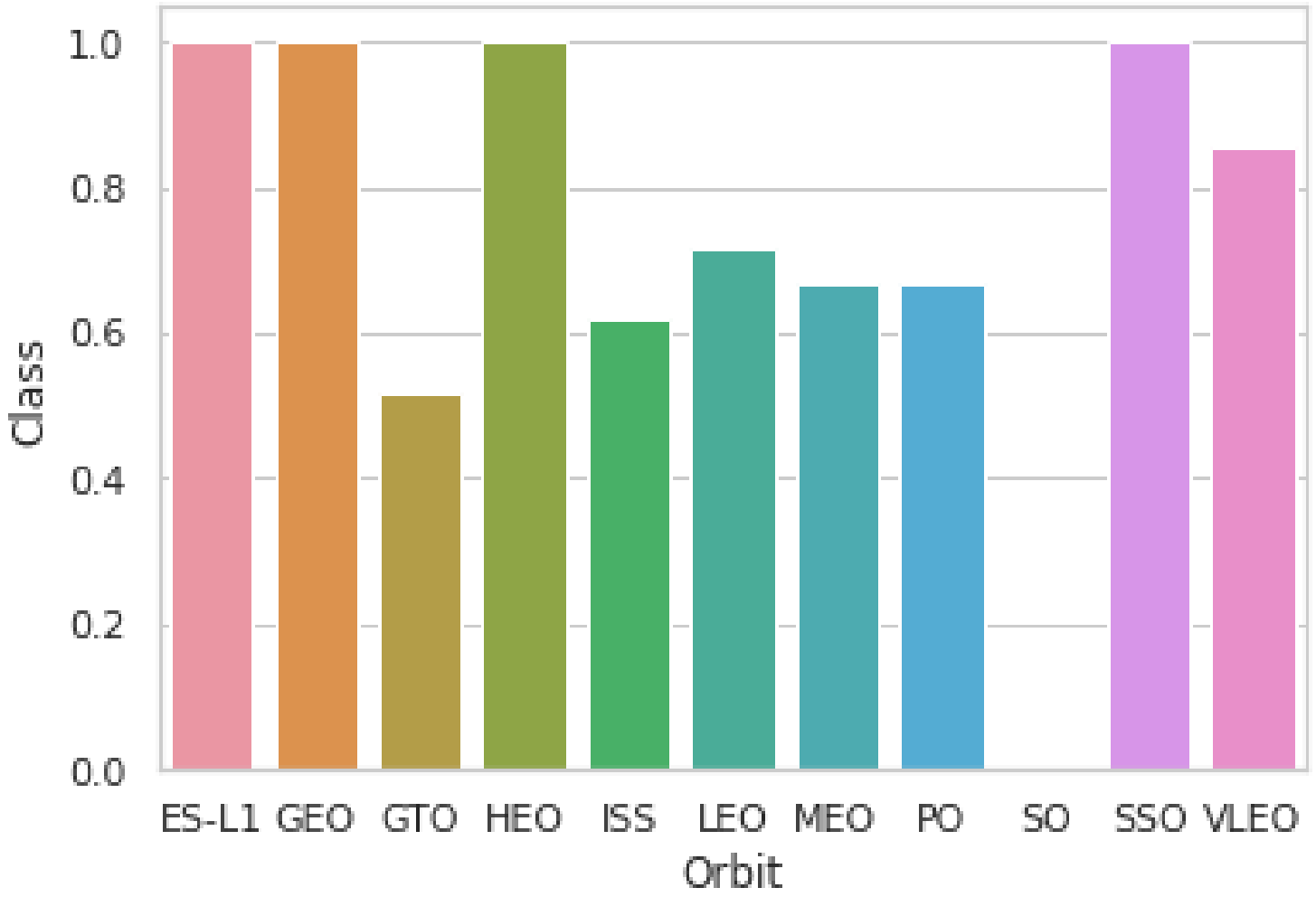
Scatter plot of Payload vs. Launch Site

Explanations: Highest payload mass launched from only two launch sites ‘CCAPS SLC 40’, ‘KSC LC 39A’

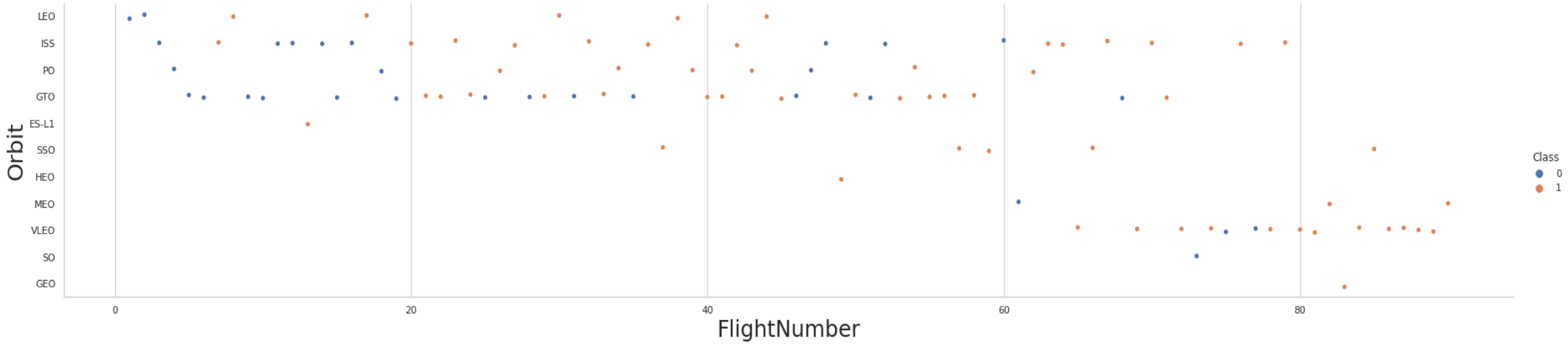
Success rate vs. Orbit type

A bar chart for the success rate of each orbit type

Explanations: minimum success rate is from orbit ‘GTO’ and maximum are ‘ES-L1’, ‘GEO’, ‘HEO’ and ‘SSO’



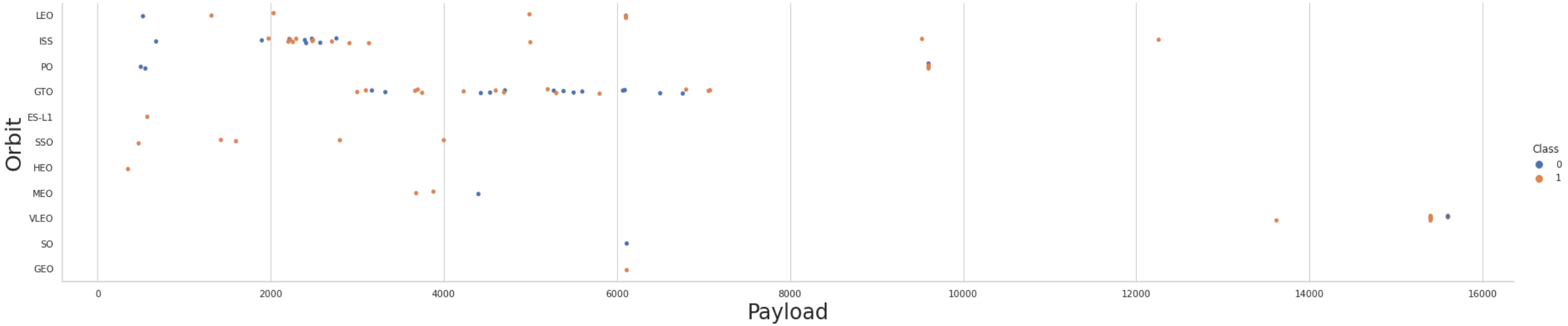
Flight Number vs. Orbit type



A scatter point of Flight number vs. Orbit type

Explanations: first almost 80 launches are to orbits LEO, ISS, PO, GTO and later most launches are to VLEO

Payload vs. Orbit type



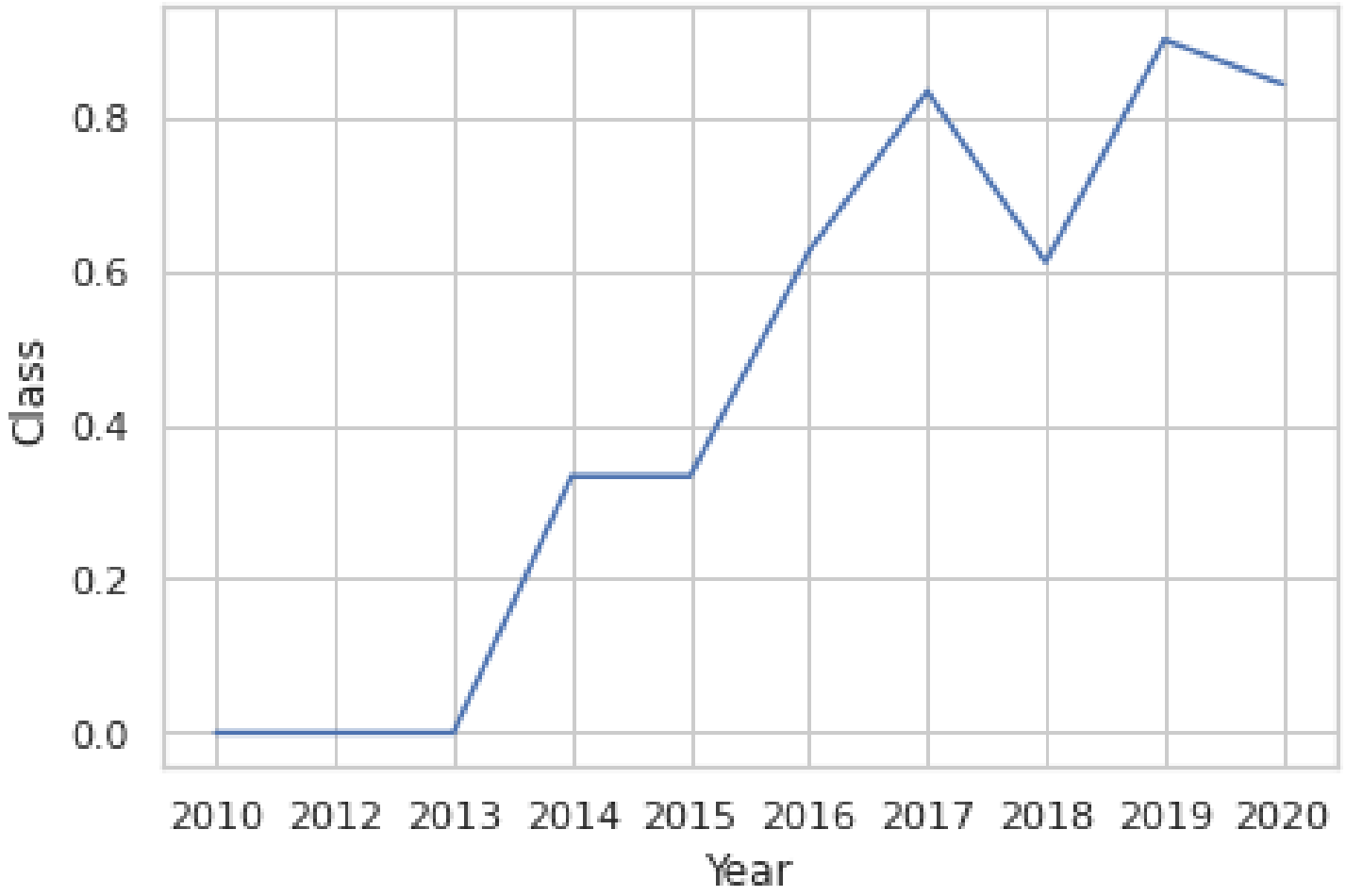
A scatter point of payload vs. orbit type

Explanations: Heaviest payload is launched only to orbit ‘VLEO’

Launch success yearly trend

A line chart of yearly average success rate

Explanations: as the year increases success rate also improved



EDA with SQL

All launch site names

- There are 5 different launch sites

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

```
%%sql
SELECT DISTINCT LAUNCH_SITE FROM SPACEX ;
```

```
 * ibm_db_sa://zxz53985:***@dashdb-txn-sbo
Done.
```

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

Launch site names begin with `CCA`

```
%%sql
SELECT Distinct(Launch_Site) FROM SPACEX WHERE Launch_Site LIKE 'CCA%'
```

```
 * ibm_db_sa://zxz53985:***@dashdb-txn-sbox-yp-lon02-06.services.eu-gb
Done.
```

5]:

launch_site
CCAFS LC-40
CCAFS SLC-40
CCAFSSLC-40

- There are three launch sites starting with CCA

Total payload mass

```
: %%sql
SELECT SUM(payload_mass__kg_) FROM SPACEX WHERE Customer = 'NASA (CRS)' ;
```

```
 * ibm_db_sa://zxz53985:***@dashdb-txn-sbox-yp-lon02-06.services.eu-gb.blue
Done.
```

26]:

1
45596

- The total payload mass is 45596 kg

Average payload mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- RESULT

- average payload mass = 2928.40 KG

```
:  %%sql
   SELECT AVG(payload_mass__kg_) FROM SPACEX WHERE booster_version ='F9 v1.1'

    * ibm_db_sa://zxz53985:***@dashdb-txn-sbox-yp-lon02-06.services.eu-gb.b
   Done.

27]:
```

1
2928.400000

First successful ground landing date

- Find the date when the first successful landing outcome in ground pad
- RESULT
- Date = 2015-12-22

```
%%sql
SELECT MIN(DATE) FROM SPACEX WHERE landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://zxz53985:***@dashdb-txn-sbox-yp-lon02-06.services.eu-gb.blu
Done.
```

3]:

1
2015-12-22

Successful drone ship landing with payload between 4000 and 6000

- List the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- RESULT

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

```
%%sql
SELECT booster_version FROM SPACEX WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 and 6000 ;
```

```
 * ibm_db_sa://zxz53985:***@dashdb-txn-sbox-yp-lon02-06.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

9]:

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

Total number of successful and failure mission outcomes

- Calculate the total number of successful and failure mission outcomes
- RESULT

- Successful outcomes = 100
- Failure outcome = 1

Boosters carried maximum **payload**

- List the names of the booster which have carried the maximum payload mass
- RESULT

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

2015 launch records

- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- RESULT

1	booster_version	launch_site
January	F9 v1.1 B1012	CCAFS LC-40
April	F9 v1.1 B1015	CCAFS LC-40
January	F9 v1.1 B1017	VAFB SLC-4E
March	F9 FT B1020	CCAFS LC-40
June	F9 FT B1024	CCAFS LC-40

Rank success count between 2010-06-04 and 2017-03-20

- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.
- RESULT

landing__outcome	2
Success (drone ship)	5
Success (ground pad)	3

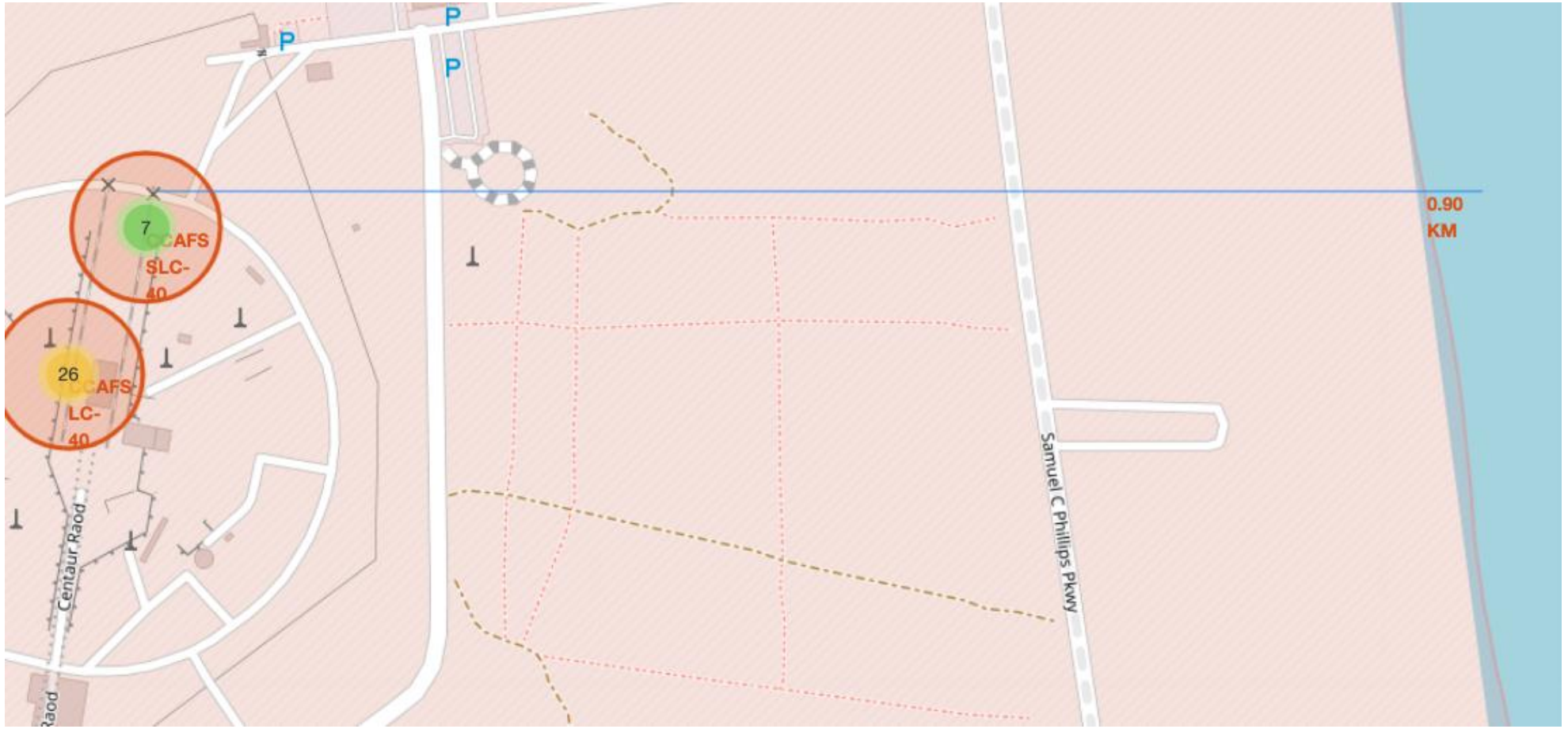
Interactive map with Folium

Map with marked launch sites

All the launch sites are near coast line

Marker Cluster on Maps

Distance between launch sites



Build a Dashboard with Plotly Dash

<Dashboard screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot

Predictive analysis (Classification)

Classification Accuracy

Visualize all the built model accuracy for all built models, in a barchart

Find which model has the highest classification accuracy

Confusion Matrix

Show the confusion matrix of the best performing model with explanation

CONCLUSION

- We use api and web scrapping of for data collection
- SQL is very helpful in EDA
- We see from charts as the Time is increasing the success rate is also improving
- Launch sites are near coastal line area
- Dash Application is a very interactive and informative way to represent data

APPENDIX

- Pandas Library is very good for data frame building it makes calculation faster and code easier to implement