IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Santhosh Kumar
23.11.2021

Outline

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

Executive Summary

Summary of methodologies

- Data collection (used rest API and web scrapping)
- Data wrangling
- Explorative data analysis with graphical visualisations
- Interactive visual representation including a dashboard
- Predictive modelling

Summary of all results

- The data was pre-processed keeping the underlying information
- The data was subsequently analysed with statistical and visual ways using open source tools (like plotly ans Dash)
- A classification model was developed
- Presenting the results to team members

Introduction

- Background

Falcon 9 developed by space X is one of its kind because it was developed for its reuse. With reuse of the falcon 9 vehicle, dramatically reduces the cost of space exploration. This has to be inline with safety. Falcon 9 should be a reliable and safe transportation method for people and payloads.

- Business problem definition

To come up with a predictive model for probability of successful landing of Falcon 9 using the given technical data regarding the rocket. This information will be used to estimate overall cost of the business if the success rate is identified.

Section 1

Methodology

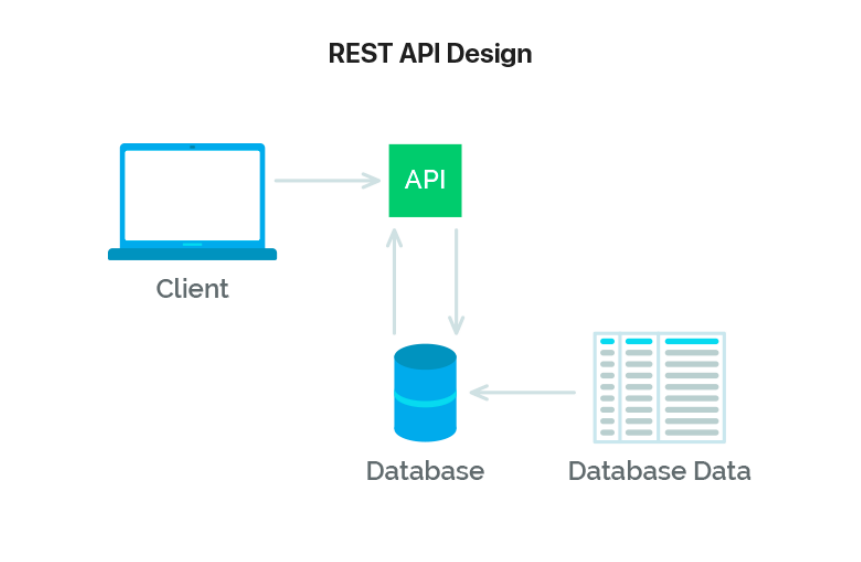
Methodology

Executive Summary

- Data collection methodology:
  - The data was collected with SpaceX rest API as well as web scrapping from Wikipedia
- Perform data wrangling
  - Transforming the data with python libraries (feature extraction)
  - Creation of more and useful data (as columns) which are correlated with the outcome
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - A classification model was build and tested with validation and test data

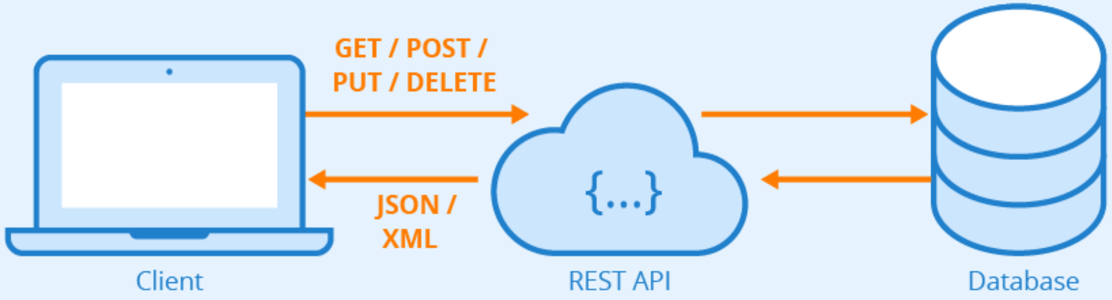
Data Collection

- SpaceX has its own RESTAPI which can be accessed without any credentials. The project used this tool.
- Web scraping is another useful tool which was used in this project.



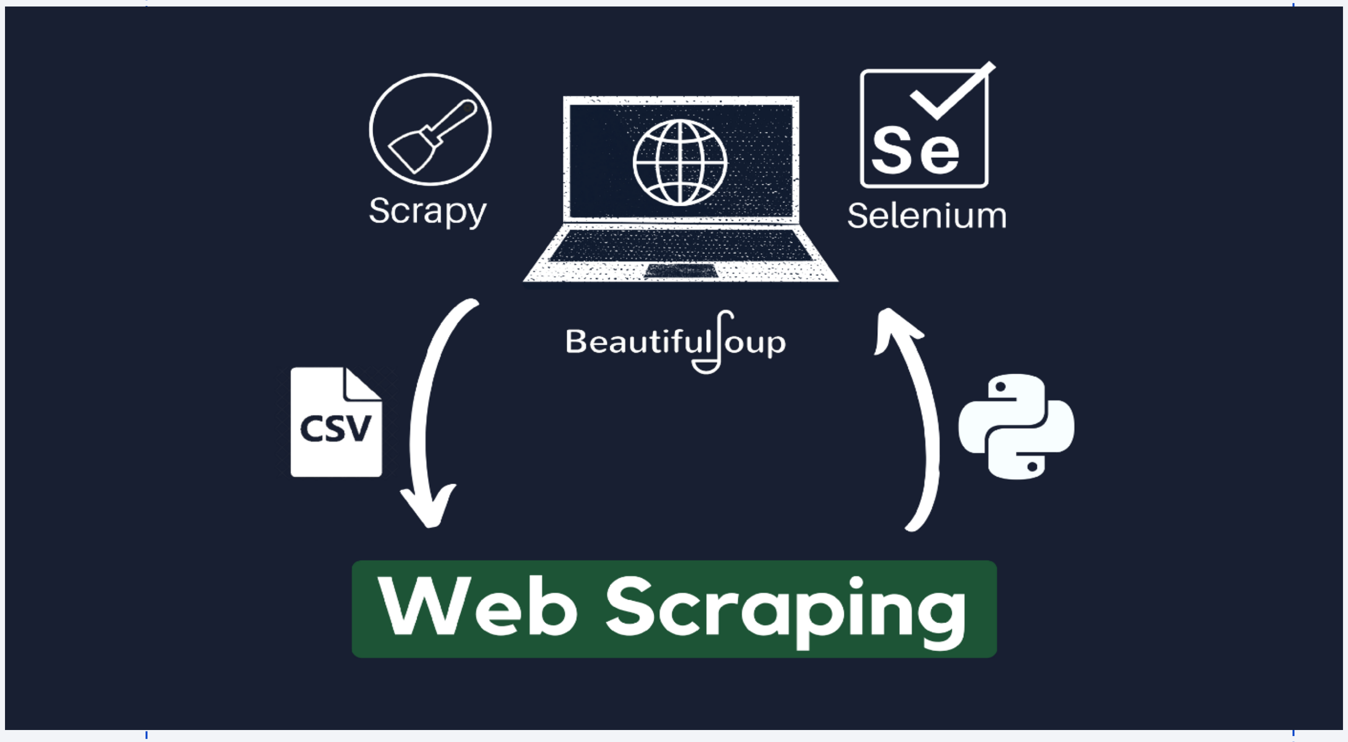
Data Collection – SpaceX API

- Rest API presents json data which was fed into a dataframe
- The API data obtained and the calls can be reviewed here
- https://github.com/gvsanthu10/IBM_datascience/blob/main/Data%20Collection%20API%20Lab.ipynb



Data Collection - Scraping

- With use of python libraries like requests, and beautifulsoup, additional data about falcon 9 was collected form Wikipedia
- The python code obtained and the data processing can be reviewed here
- https://github.com/gvsanthu10/IBM_datascience/blob/main/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb



Data Wrangling

- The collected data was preprocessed. The methods involved were replacing missing values, feature extraction and addition of new columns

- The code of the data wrangling can be viewed here:

https://github.com/gvsanthu10/IBM_datascience/blob/main/Data%20Wrangling%20Lab.ipynb

EDA with Data Visualization

- Several charts were plotted and the explained with inferences. The visual representation of the data was more initiative in explain few findings as well as finding important features
- The code for the EDA with visualization can be viewed here:
- https://github.com/gvsanthu10/IBM_datascience/blob/main/The%20EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- SQL was used to obtain important values. SQL was used because of simplistic approach to access the data stored in a database

- The source code is stored and accessed here:
- https://github.com/gvsanthu10/IBM_datascience/blob/main/EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

- An interactive map with Folium was developed to help understanding the data with graphs and maps. The map shows circular marks for each site with color coding depending on the outcome
- The code for the folium map is found here:
- https://github.com/gvsanthu10/IBM_datascience/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Another web based interactive dashboard was developed. This helped with sharing insights with the team members. Various types of the graphs shows relationship between different parametes(especially payload) with the outcome
- The source code of the dashboard can be viewed here:
- https://github.com/gvsanthu10/IBM_datascience/blob/main/spacex_dash_app.py

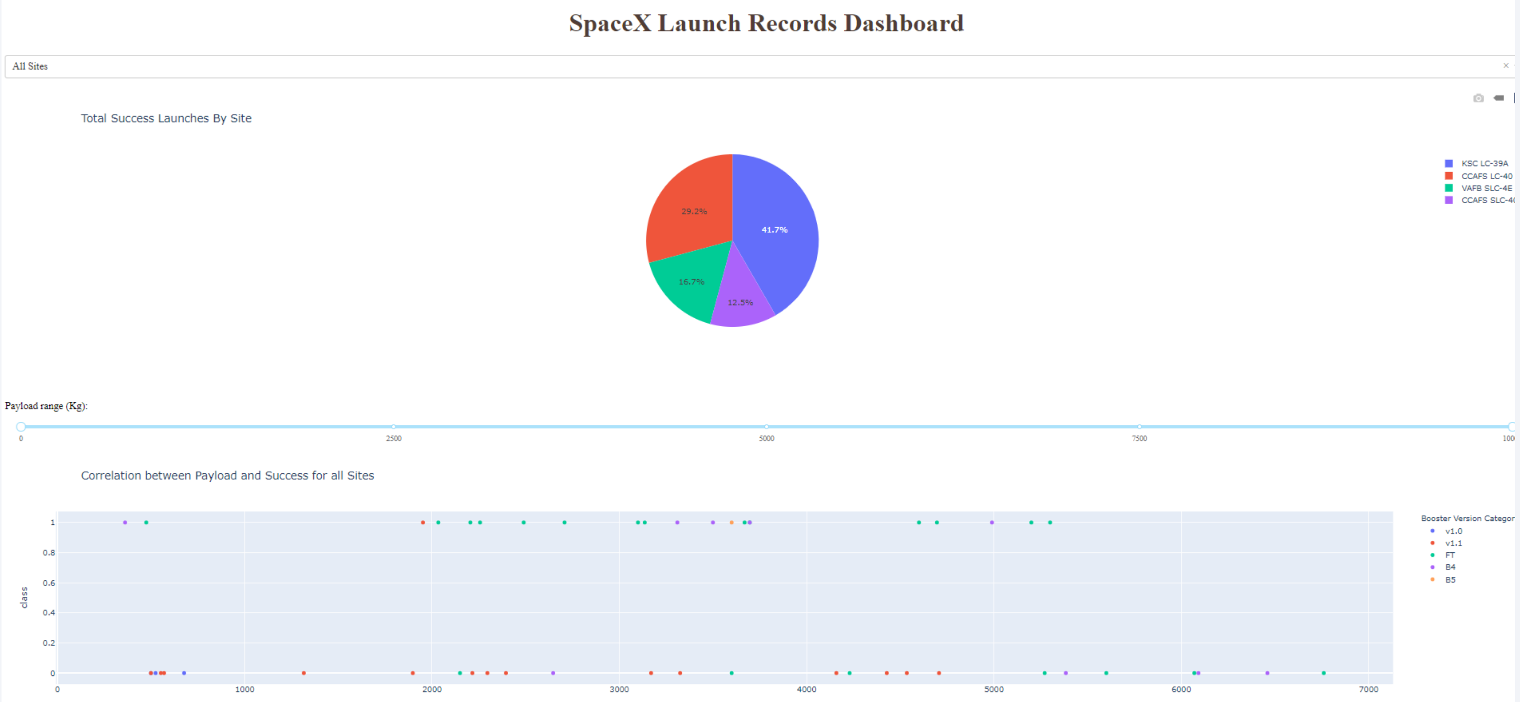
Predictive Analysis (Classification)

- Several classification model were developed and their performance on the validation data was compared. GridSearchCV was a handy tool for the quick selection of optimal hyperparameters.

- The source code can be viewed here:
- https://github.com/gvsanthu10/IBM_datascience/blob/main/Machine%20Learning%20Prediction%20lab.ipynb

Results

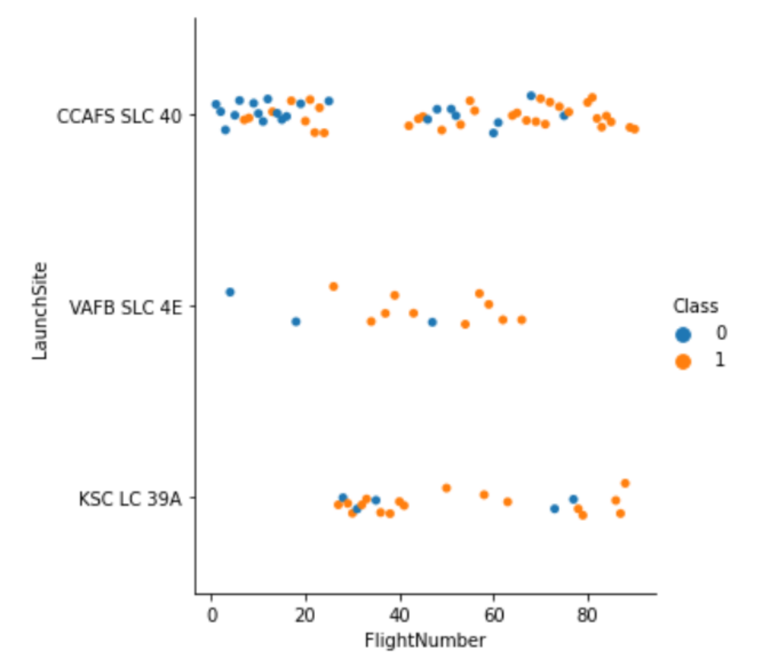
- EDA showed that orbit, launch site, payload and the number of flight were significantly correlated with the outcome
- The highest accuracy of the models were 83%



Section 2

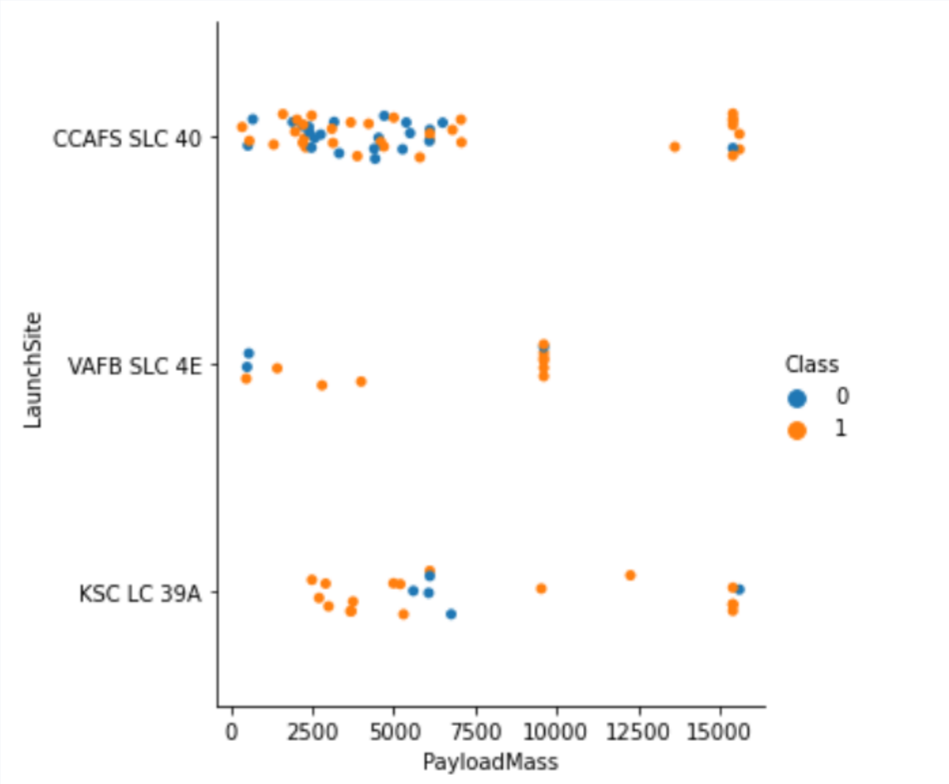
Insights drawn from EDA

Flight Number vs. Launch Site



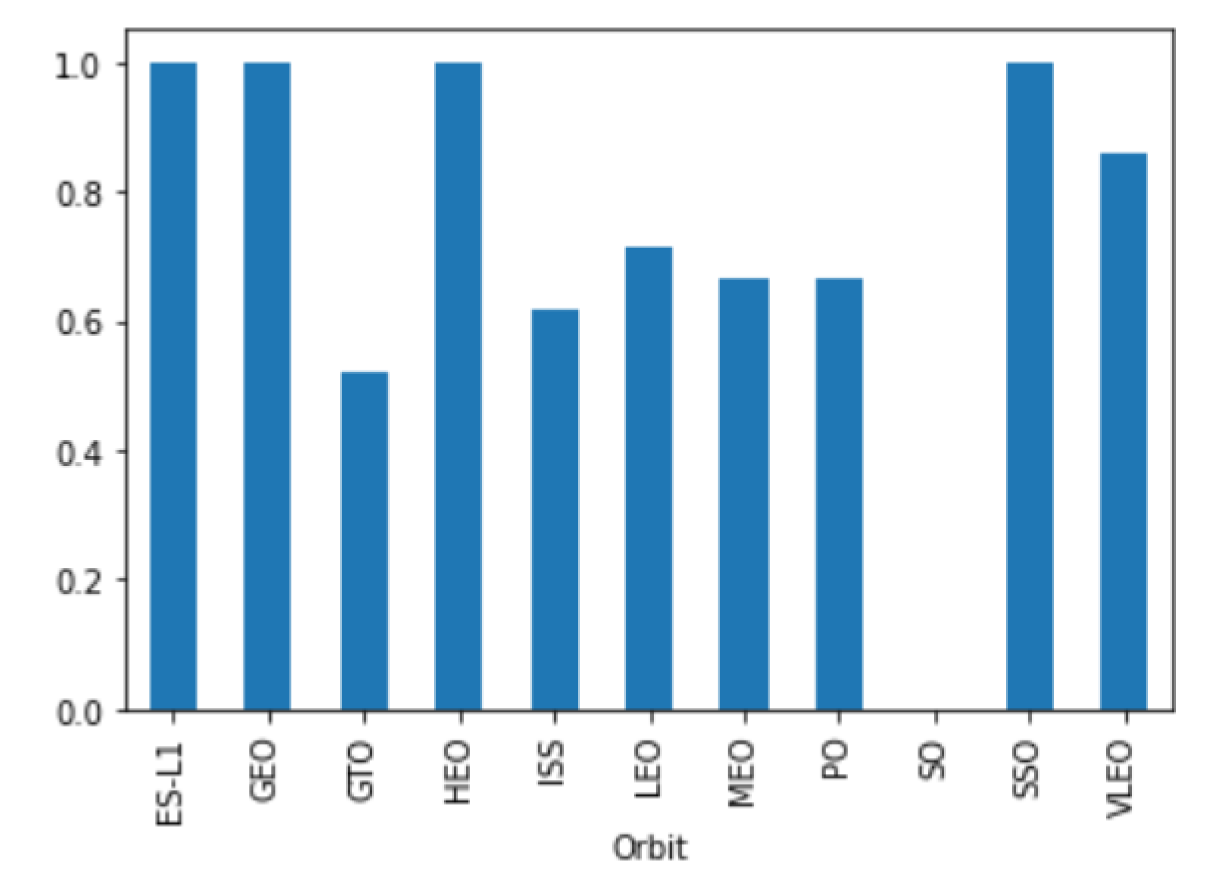
- The chart shows increased success rate with increased number of flights
- There is weak correlation with Launch site and outcome

Payload vs. Launch Site



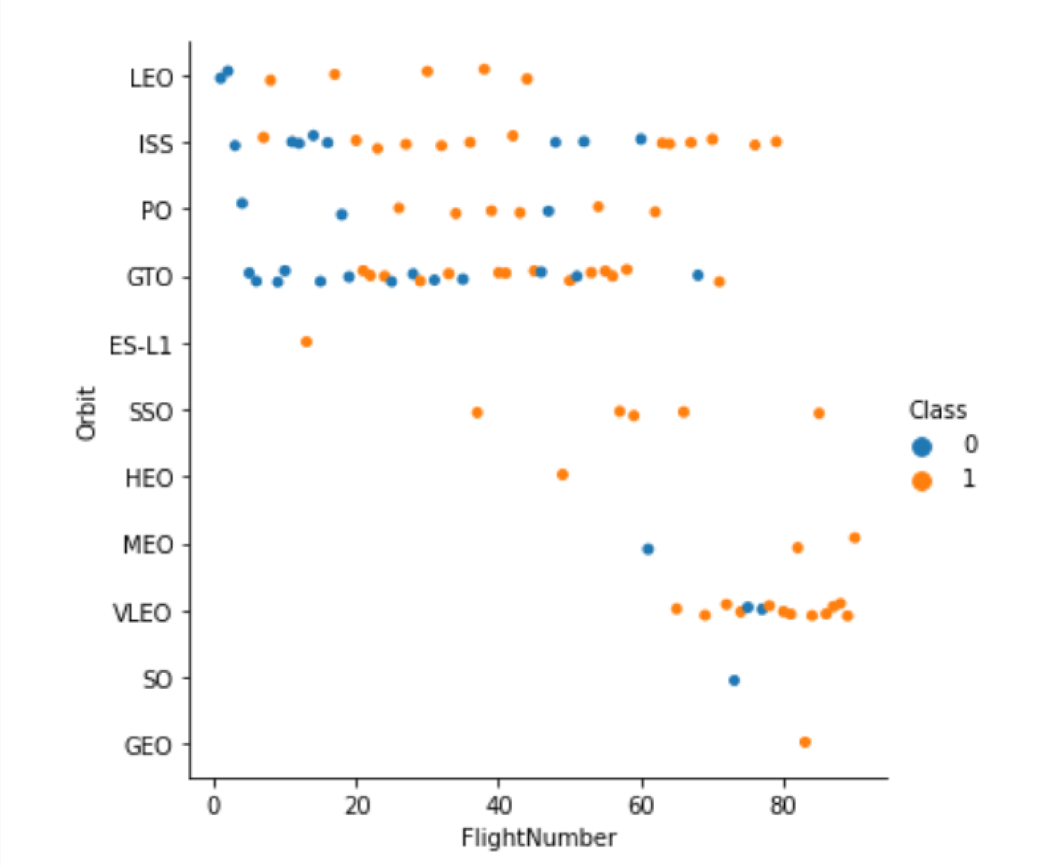
- The graph shows a slight correlation of payload with the outcome

Success Rate vs. Orbit Type



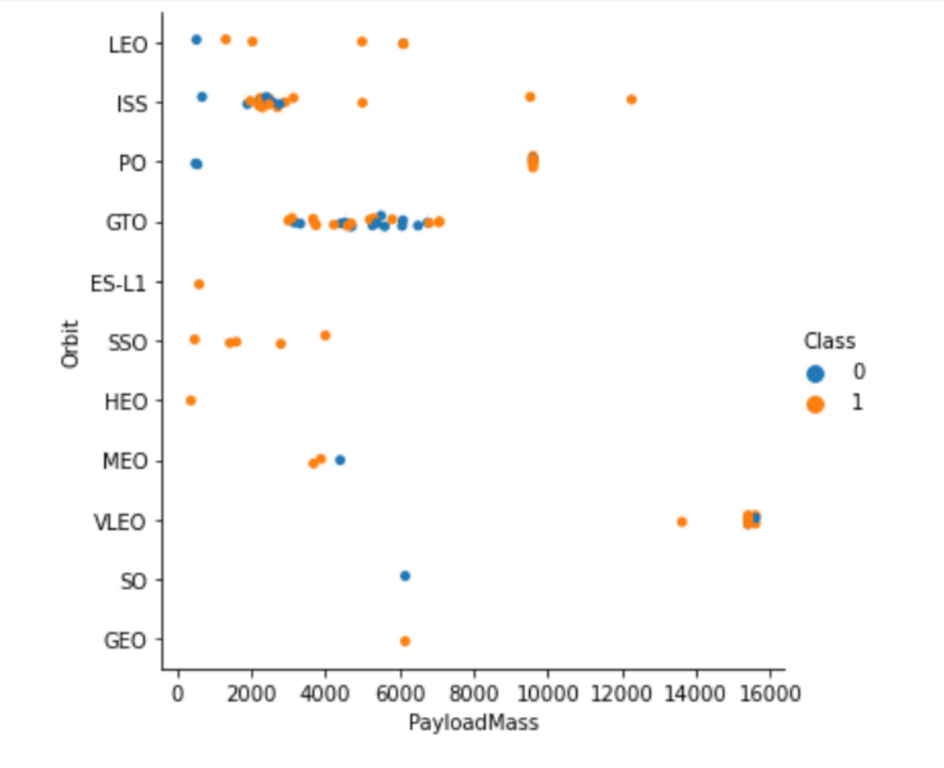
- Bar chart visualization of the Orbit type and the success rate. Few of the orbits are associated with less success rate

Flight Number vs. Orbit Type



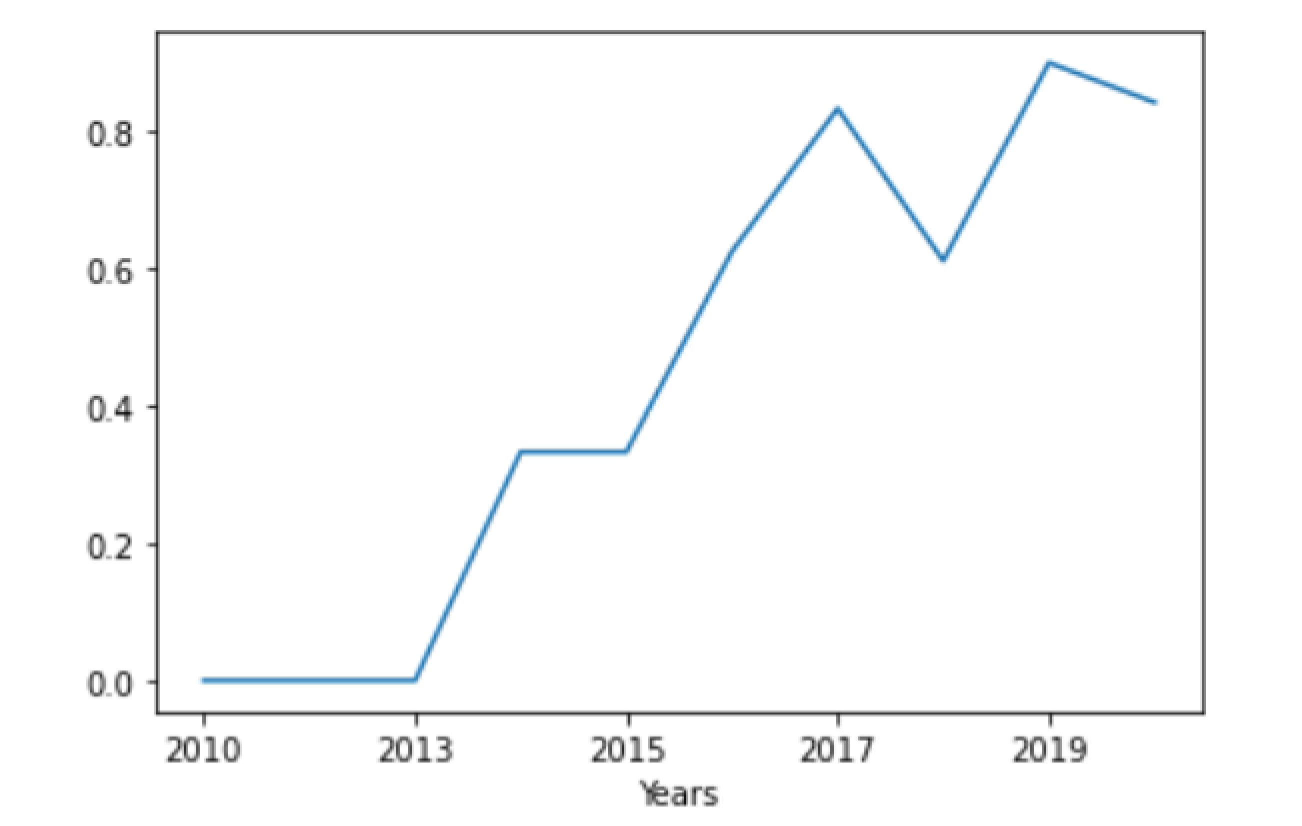
- This graph again shows increased success rate with increased flight number. The VLEO has good success rate partly due to increased flight number

Payload vs. Orbit Type



- Scatter plot with orbit and payload mass. There is a correlation of heavy payload with lower success rate

Launch Success Yearly Trend



- A line chart showing progressive increase in the success rate vs time (years)

All Launch Site Names

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

- These are our launch sites

Launch Site Names Begin with 'CCA'

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

- Result of query regarding launch site “CCA”
- First 5 rows are shown here

Total Payload Mass

- The total payload carried was 45596 kg. This might be an important figure which can be asked while estimating cost and profitability

Average Payload Mass by F9 v1.1

- The average payload mass carried by F9 v1.1: 2534 kg. The new version has improve payload carrying capacity

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad was 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- These are the list of booster versions with payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- The total count of success and failure of the mission show here

Boosters Carried Maximum Payload

booster_version
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

- These are the names of the boosters carried with maximum payload mass

2015 Launch Records

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- This table (SQL output) shows failed landing outcomes for booster versions and launch site for year 2015

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

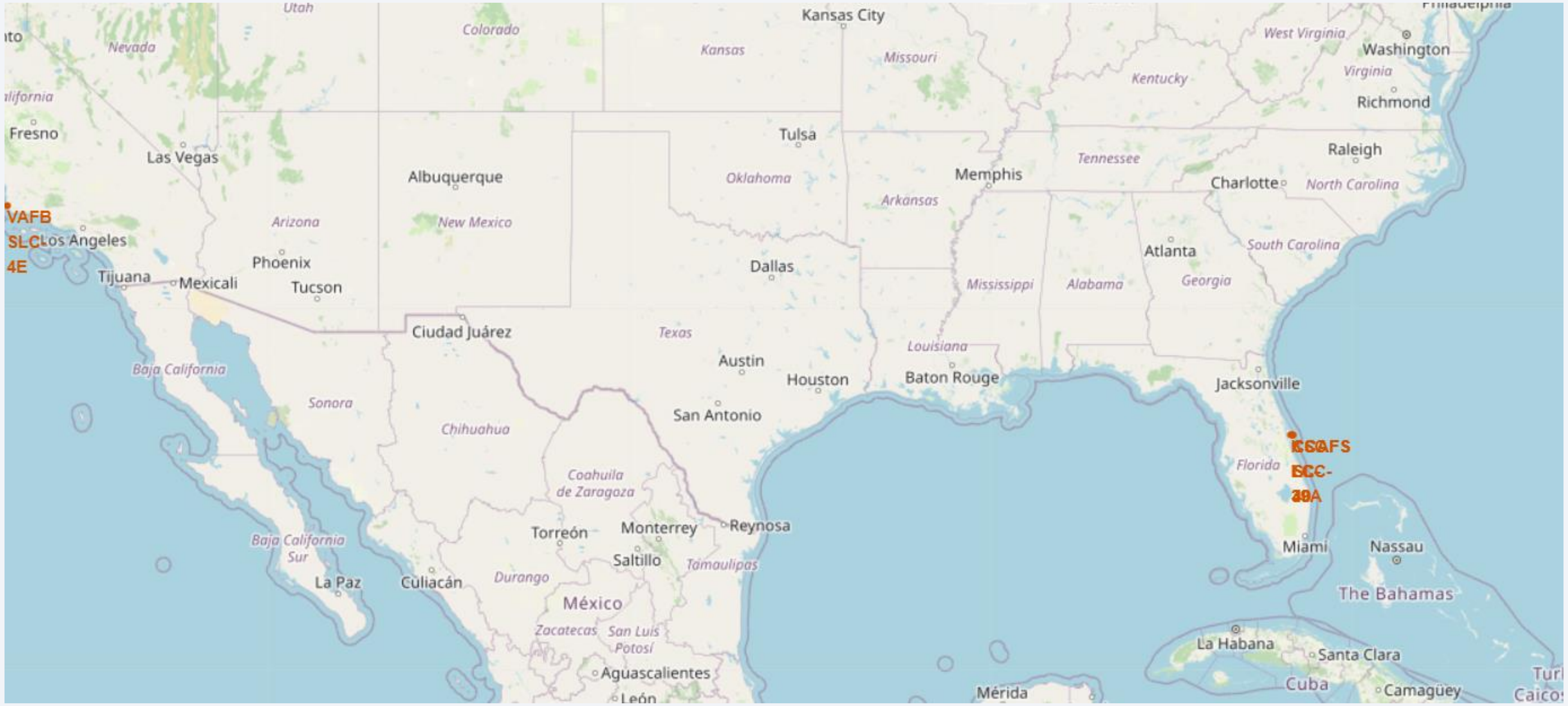
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

The total count of landing outcomes in for dates between 2010-06-04 and 2017-03-20

Section 4

Launch Sites Proximities Analysis

Lauch sites

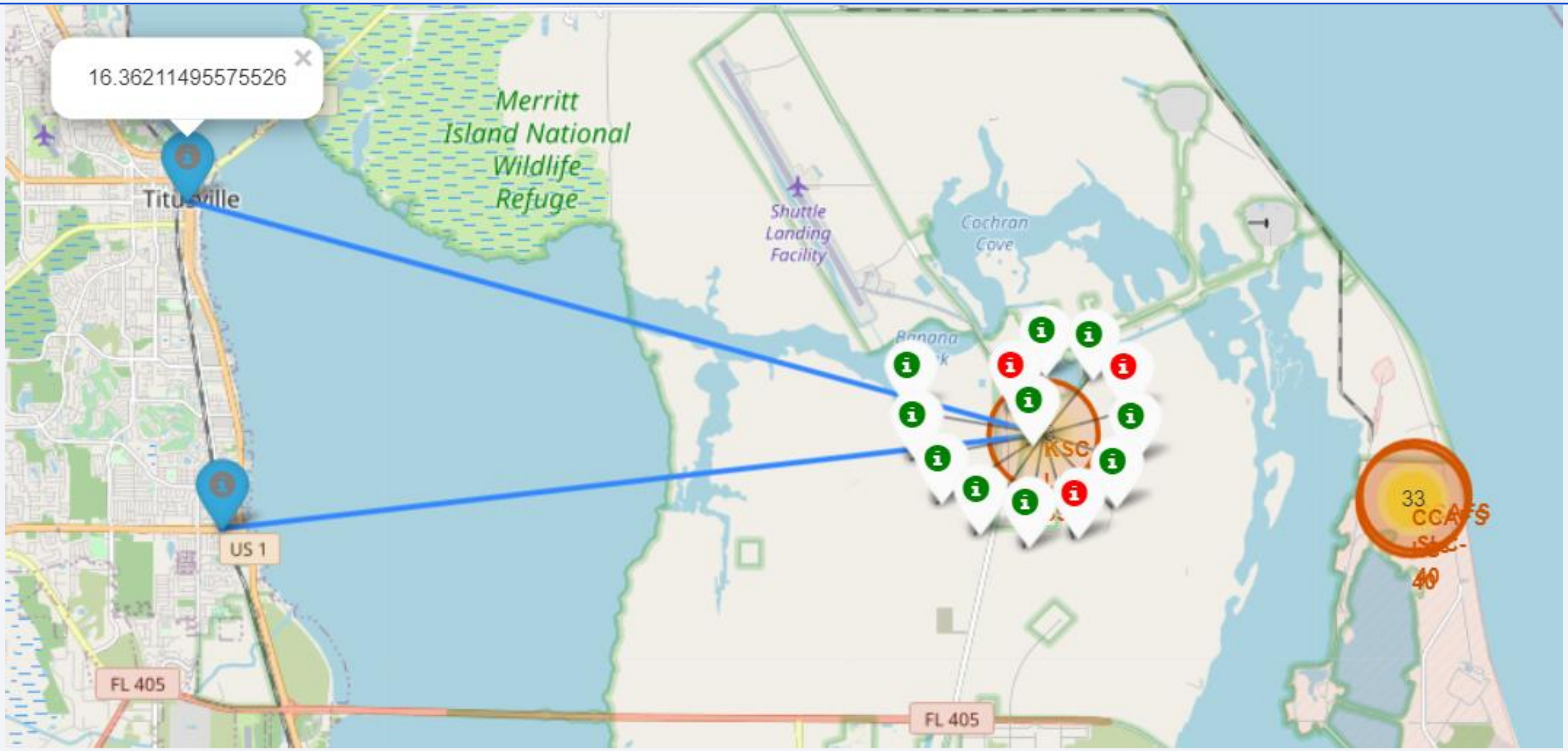


- This map with Folium showing all the lauch sites

Map Overled with number of lauchs

- 46 launches were made east cost and 10 in west coast

Map zoomed to east launch site



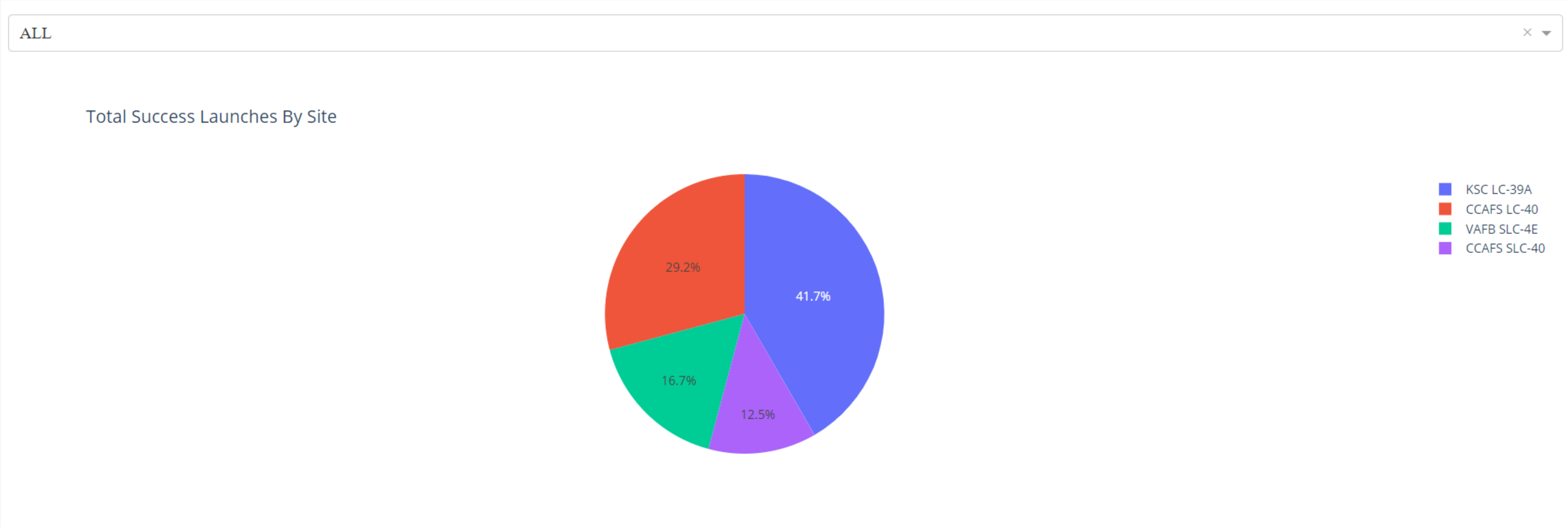
- This is the zoomed in map showing lauch sites with green as successful and red as unsuccessful outcomes

Section 5

Build a Dashboard with Plotly Dash

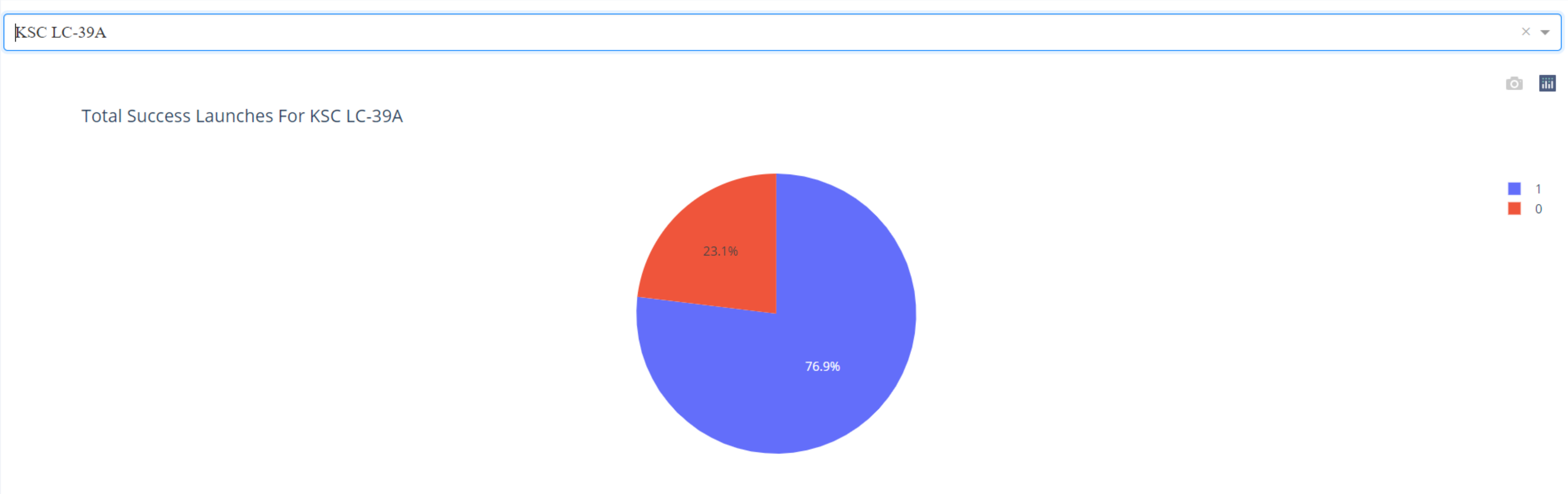
First graph in dashboard

- Pie chart showing successful launches distribution with lauch site



Pie chart KSC LC-39A

- Another screenshot showing KSC LC-39a success rate



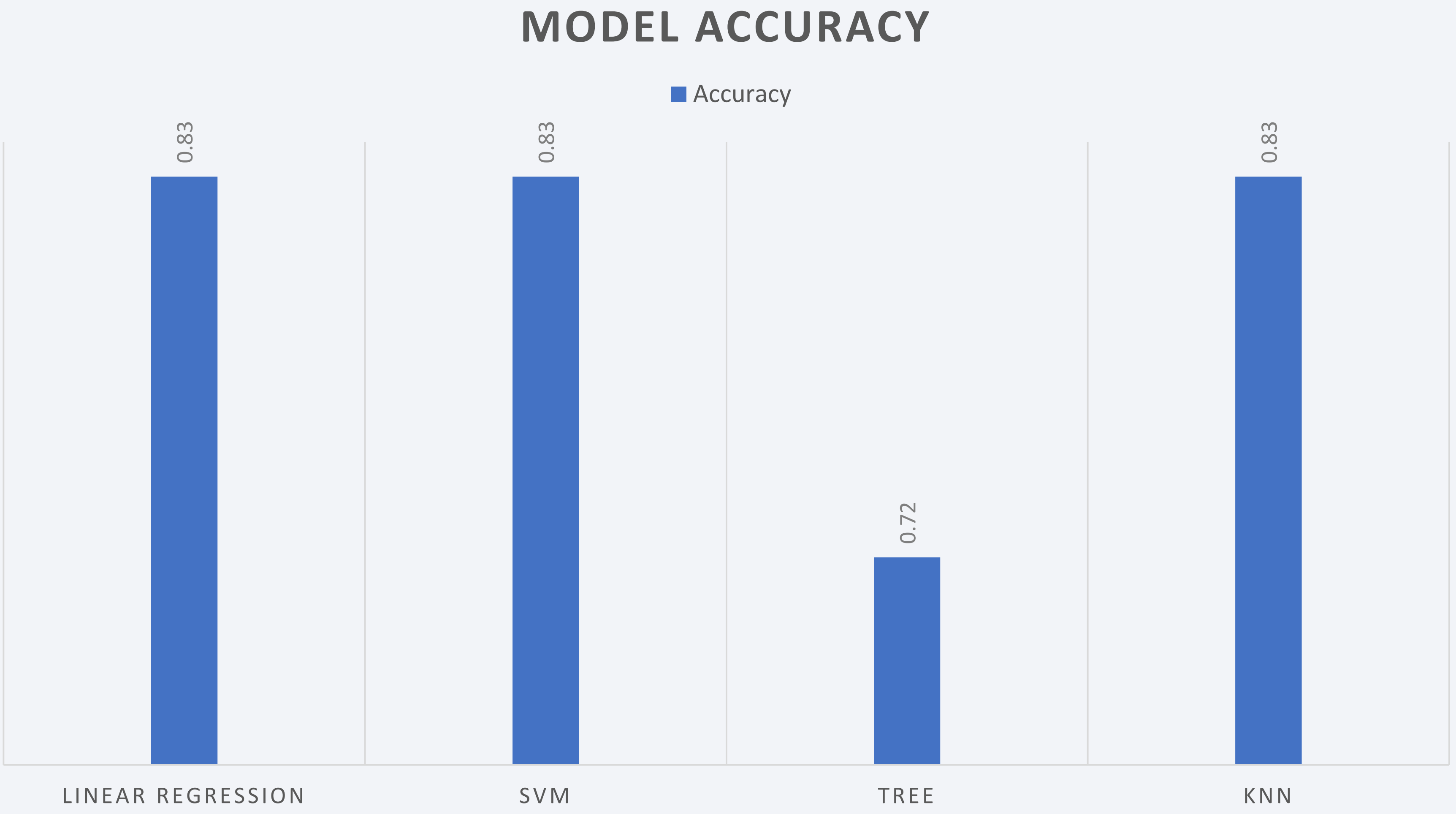
Payload vs class and outcome scatter plot

- Payload vs success scatter plot.

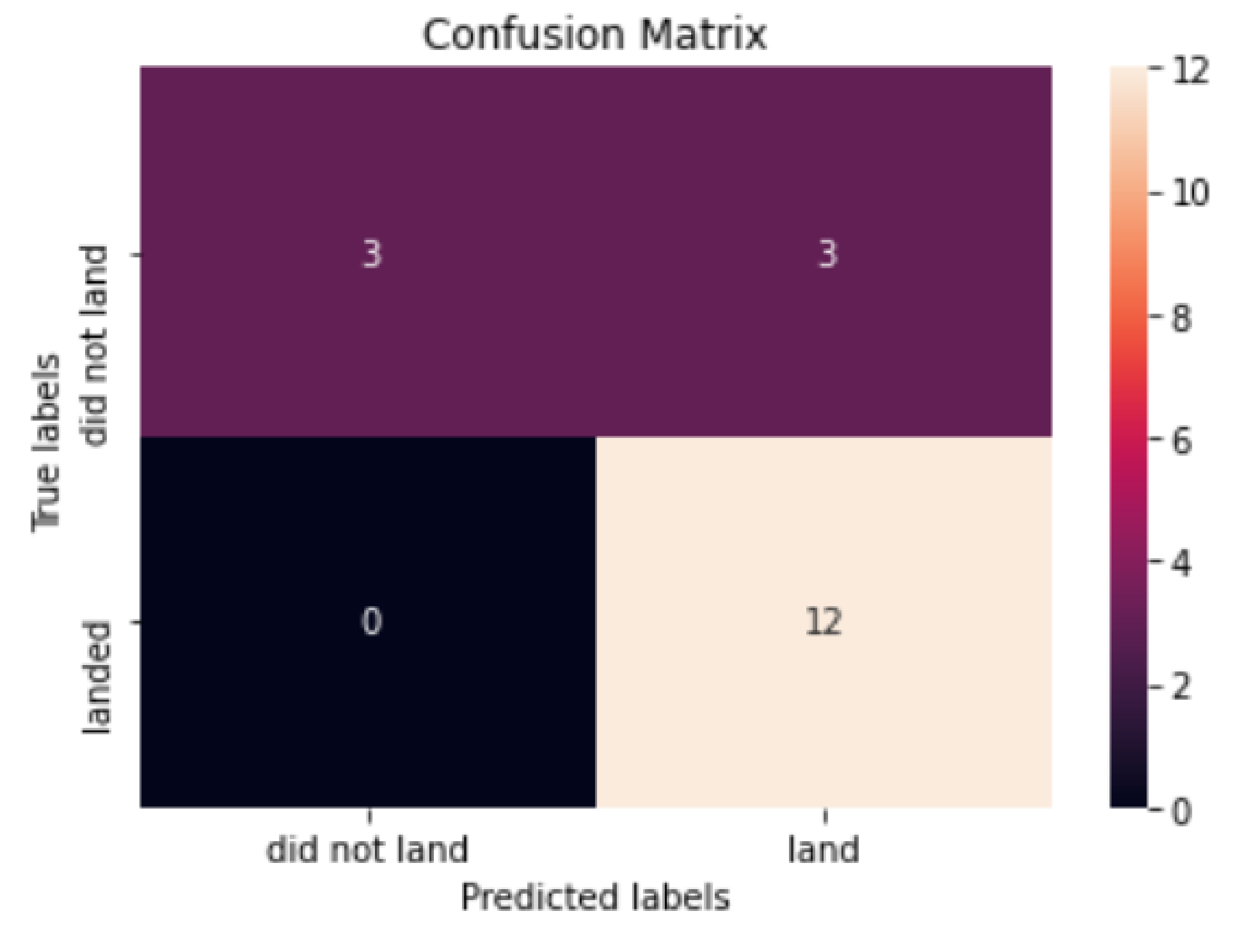
Section 6

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



- Confusion matrix showing true positive, false positive, true positive, true negative and false negative in numbers

Conclusions

- With predictive modelling technique, we predicted success rate of SpaceX launch outcome with 83% accuracy.
- The insights obtained were
- 1. lower payload increased the success rate. However, payload is also linked to profitability. We have to come up with a optimal payload for acceptable success rate (assuming that rest of the parameters are constant). Another way is to retest the hypothesis with new technology and invasion
- 2. effect of orbits on success rate
- 3. launch site related features. Are there any room for improvement?

Appendix

- All the source code and notebooks are in my github repository. I ran these notebooks in IBM Watson and/or google colab.
- Github repository:
- https://github.com/gvsanthu10/IBM_datascience

Future path

- Try to deploy the dashboard in heroku website
- The model should be trained with new data. If there significant drift or shift in the prediction, analyze the new parameters regarding change in the data
- Continuous monitoring of the model workflow (website) and its performance (feedback)

Thank you!