IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Muhammed Naseef
21/12/2021

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context

  Purpose of this project is to calculate the cost of launching rockets by predicting whether the first stage will land successfully or not so we can re use and save money. That way we, Space Y can create a good offer for our customers.

- Problems you want to find answers
  - What are the factors influencing the landing of launchers?

Section 1

Methodology

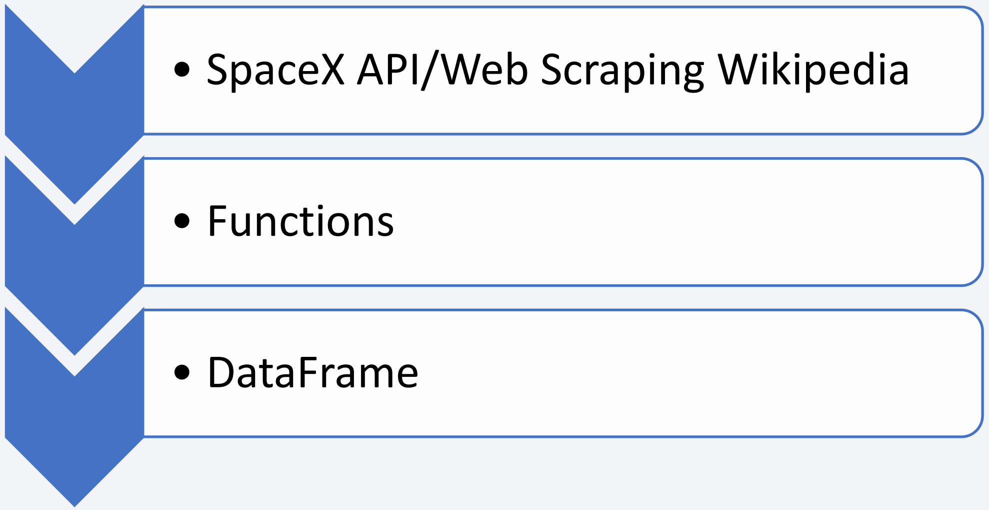
Methodology

Executive Summary

- Data collection methodology:
    - SpaceX API
    - Web Scraping Wikipedia
- Perform data wrangling
    - Removed Missing Values and Introduced class label to data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Best model and hyper parameters are found using GridSearchCV.

Data Collection

- Data were collected using SpaceX API / Web scraping Wikipedia and refined using functions to create a data frame.



Data Collection – SpaceX API

GitHub URL

1
- SpaceX API
- Taking data from url using request module

2
- Changing data to pandas dataframe using pandas json_normalize funtion

3
- Using functions to get data from the given data

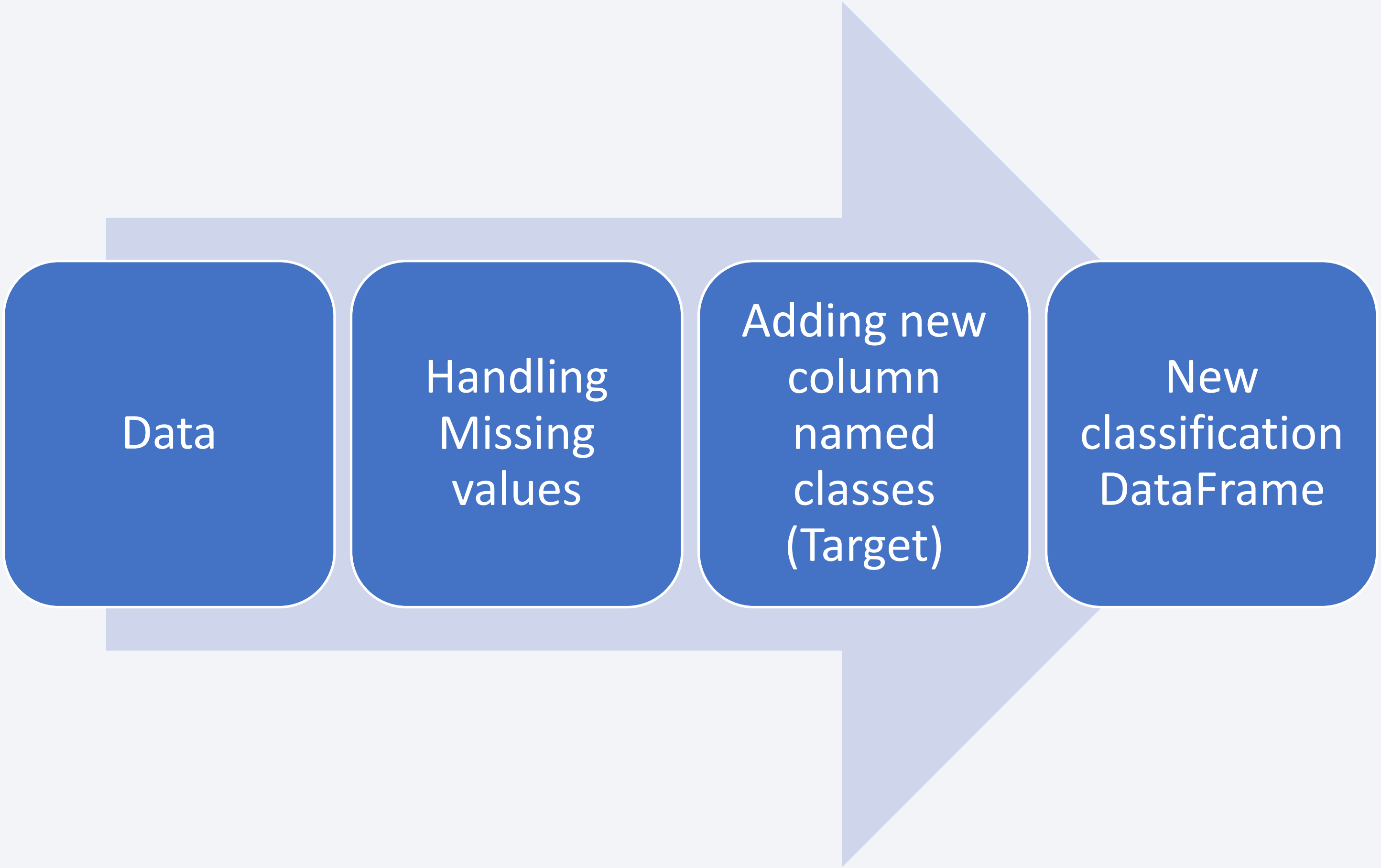
Data Collection - Scraping

[GitHub URL]

1. Wikipedia URL
2. Request module
3. Beautiful soup
4. Extracting tables
5. Changing into DataFrame
5. Adding new data
6. DataFrame

Data Wrangling

[GitHub URL to Notebook]



EDA with Data Visualization

GitHub URL to Notebook

Scatter Plots

- Flight Number Vs Payload
- Flight Number Vs Launch Site
- Flight Number Vs Orbit
- Payload Vs Launch Site
- Payload Vs Orbit

Reason: To find correlation between the values

- Bar chart of Orbits Avg. Success rate

Reason: To find Success rate of each orbit

- Line Plot of Avg. Success rate for Years.

Reason: To find correlation between year and success

EDA with SQL

GitHub URL to Notebook

Performed SQL queries to find information from data.

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

Build an Interactive Map with Folium [GitHub URL to Notebook]

- To Folium map object I have added:

- Circle – To show site locations.
- Marker – To show site location.
- Marker Cluster – To show details of launches for each site.
- Mouse Position – To get coordinates of certain objects.
- Polyline – To draw between site and it’s proximities

Build a Dashboard with Plotly Dash

[GitHub URL to Notebook]

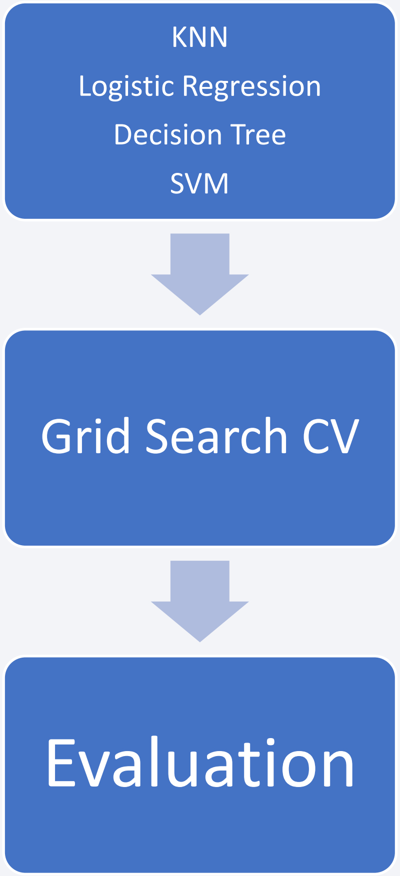
- Objects in Plotly Dash app:

- Dropdown: To change sites selected in graphs.
- Pie Chart: To Show Success ratio for sites.
- Slider: To change Payload range in Scatter plot.
- Scatter plot: To show correlation between success rate and payload

Predictive Analysis (Classification)

[GitHub URL to Notebook]

I have used Grid Search CV To find best hyper parameters for algorithms(i.e KNN, Logistic Regression, Decision Tree, SVM) and later evaluated using test set to find best algorithm. But The accuracy score was same for all four algorithms (83%). So any algorithm is ok for this problem. So I will be choosing Logistic Regression because SVM is bad for large data, Decision Tree tends to change highly, and KNN is not always optimal.



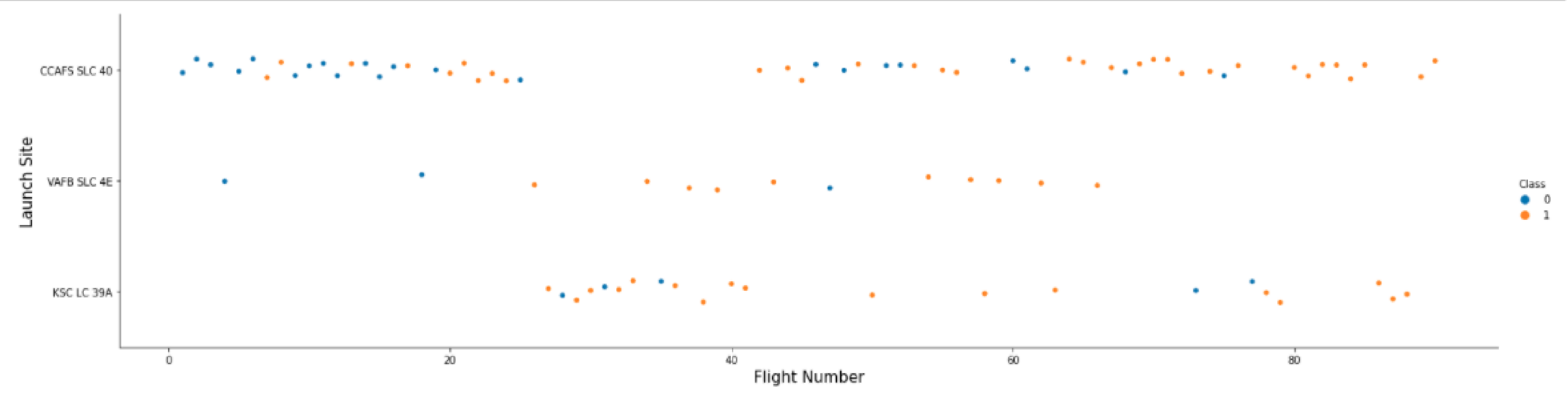
Results

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

Section 2

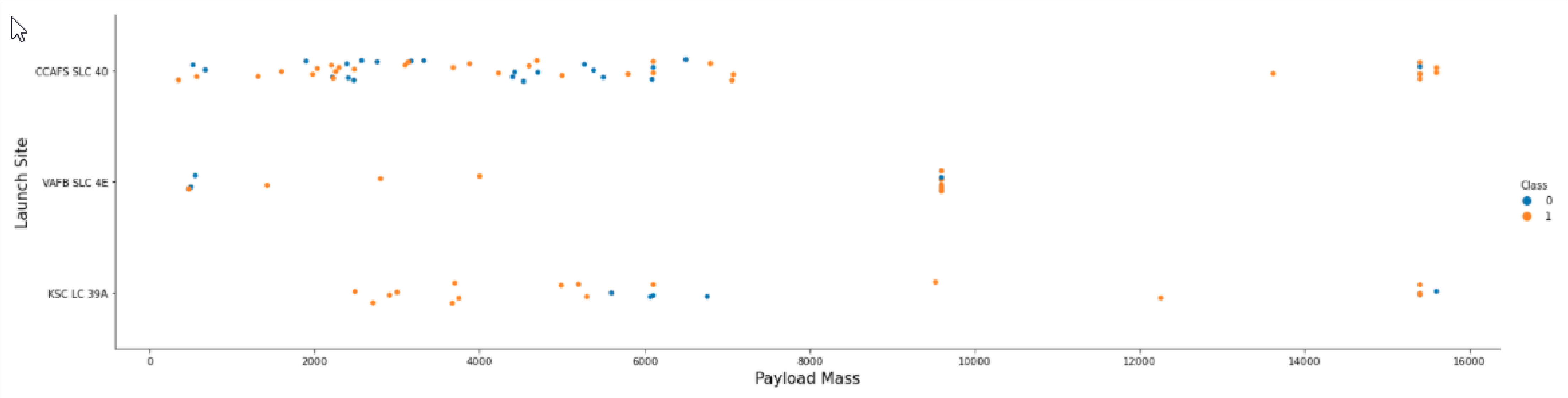
Insights drawn from EDA

Flight Number vs. Launch Site



We can say that increasing flight number also increases success rate for all launch site. And also CCAFS SLC-40 is a popular launch site.

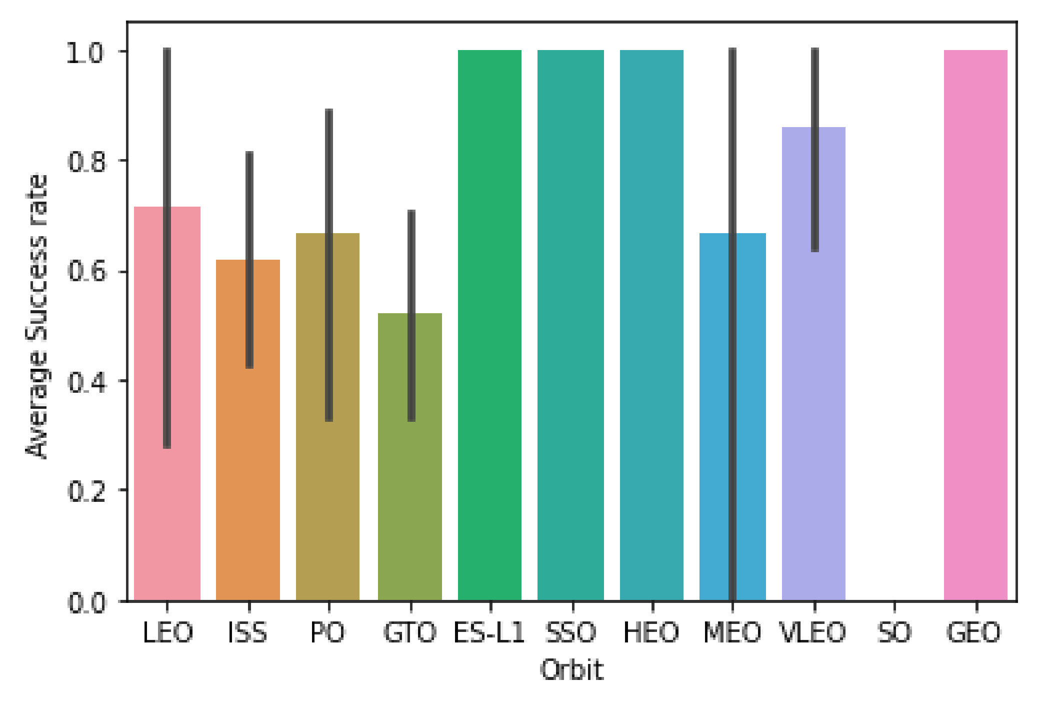
Payload vs. Launch Site



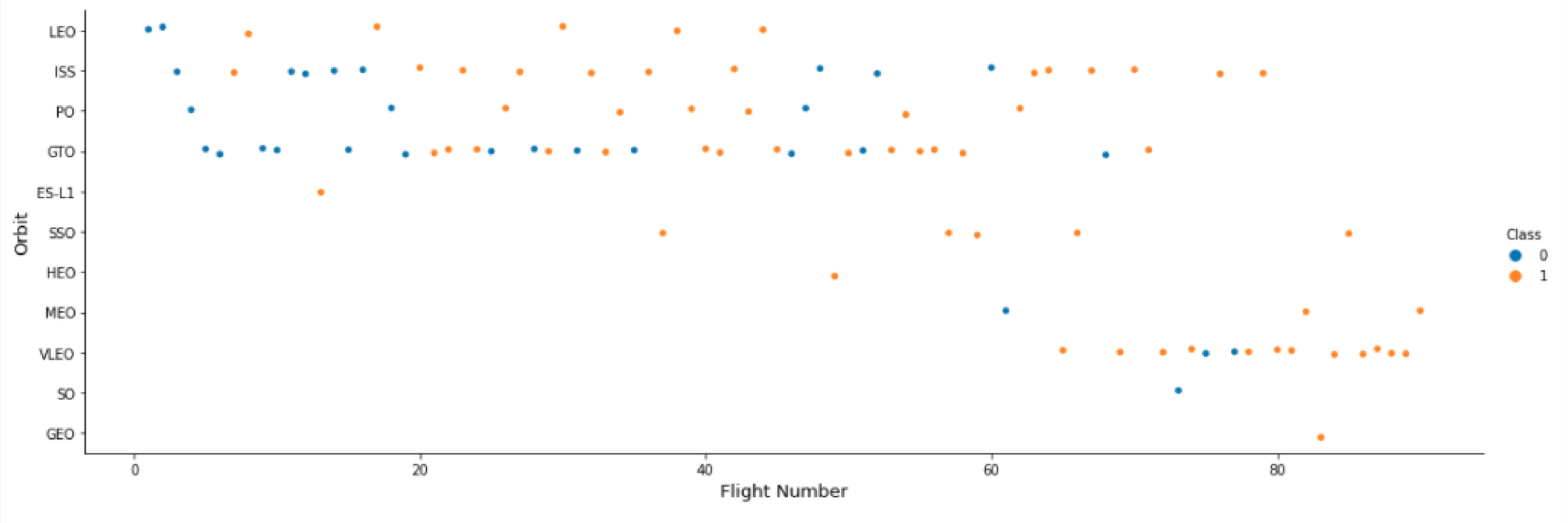
Site KSC LC 39A has a high correlation between success and payload It has good success rate at low payload. While site CCAFS SLC 40 seems to have no correlation between success and payload. VAFB SLC 4E has no launches with payload more than 1000kg.

Success Rate vs. Orbit Type

Orbit ES-L1, SSO, HEO and GEO have the highest average success rate. While SO has a very low success rate

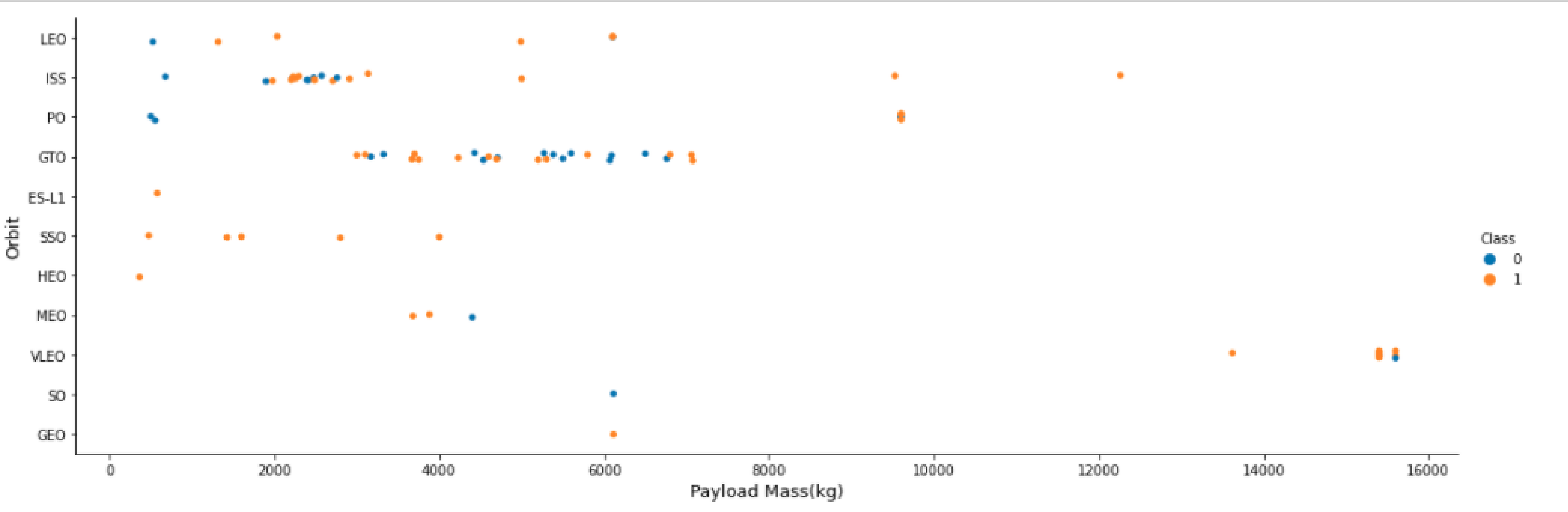


Flight Number vs. Orbit Type



LEO orbit's success seems to be highly related with flight number while GTO orbit's success has no relation.

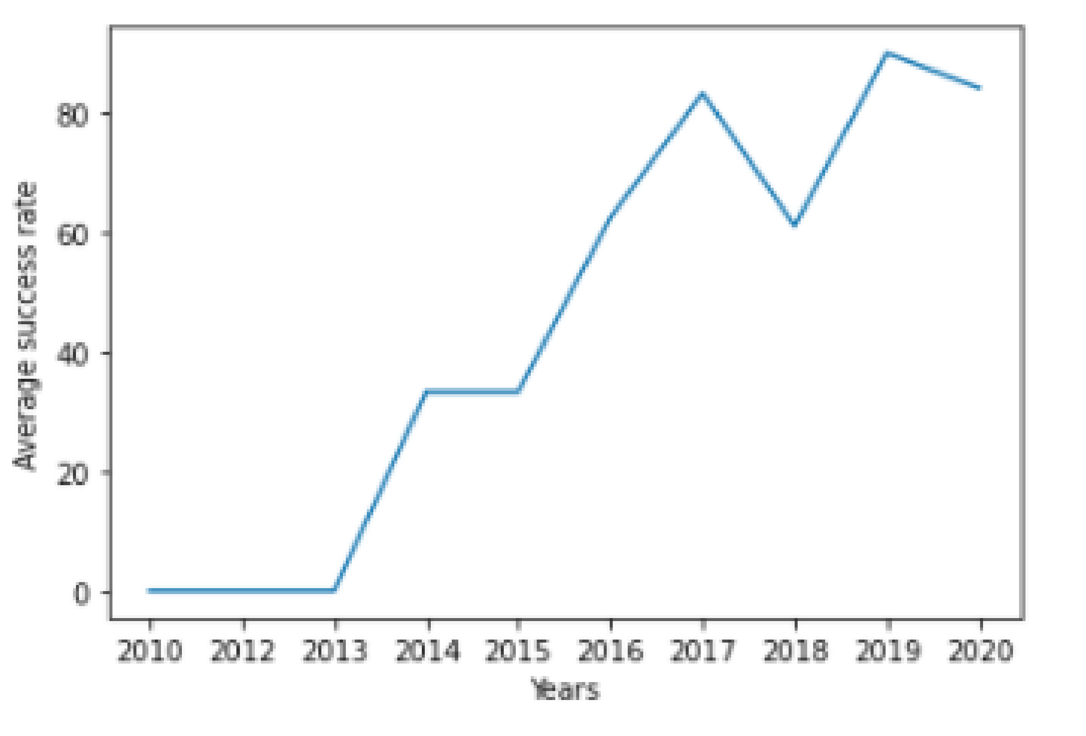
Payload vs. Orbit Type



For high payload Orbit LEO, ISS and PO have high success rate and MEO have low success rate and GTO seems not related.

Launch Success Yearly Trend

Mainly for Year and Average success rate have a positive correlation except 2018, 2020. For these two years there is small dip.



All Launch Site Names

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

There is four launch sites. And two of them(CCAFS LC-40 and CCAFS SLC-40) Seems to be highly related.

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

Main customers of this launch site seems to be NASA. And orbit is LEO

Total Payload Mass

Total Payload Mass for customer NASA is 45596 kg

45596

Average Payload Mass by F9 v1.1

Average Payload for F9 v1.1 booster is 2928 kg

avg_payload
2928

First Successful Ground Landing Date

First successful Ground Landing was on December 22nd in 2015

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters which landed successfully on Drone Ship with Payload between 4000 kg and 6000 kg are

F9 FT B1026
F9 FT B1021.2

Total Number of Successful and Failure Mission Outcomes

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

There are total of 100 Success and 1 failed Missions

Boosters Carried Maximum Payload

There are total of 12 Booster that carried Maximum Payload

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

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

Both of the Failed Outcome in Drone Ship in 2015 is launched from CCAFS LC-40

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

In Between July 4th 2010 and March 20th 2017 for 10 times there was no attempt to launch .

Section 4

Launch Sites Proximities Analysis

Launch Site Locations

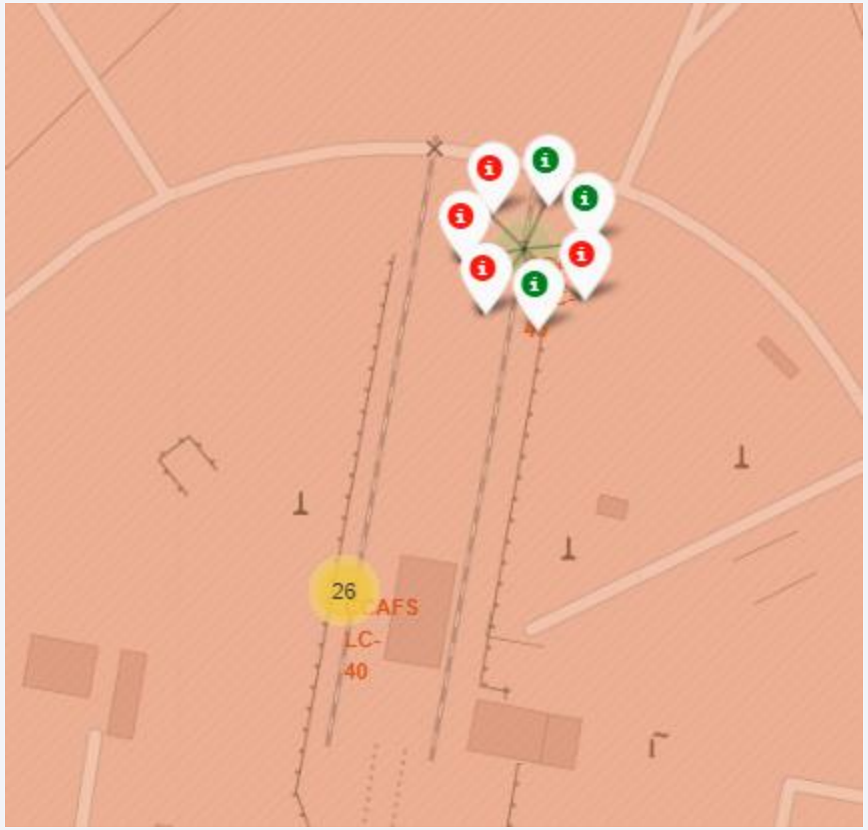
As you can see one of launch site is in California and other 3 is in Florida. A common property for all their location have is a small distance to coast.

Launches at CCAFS

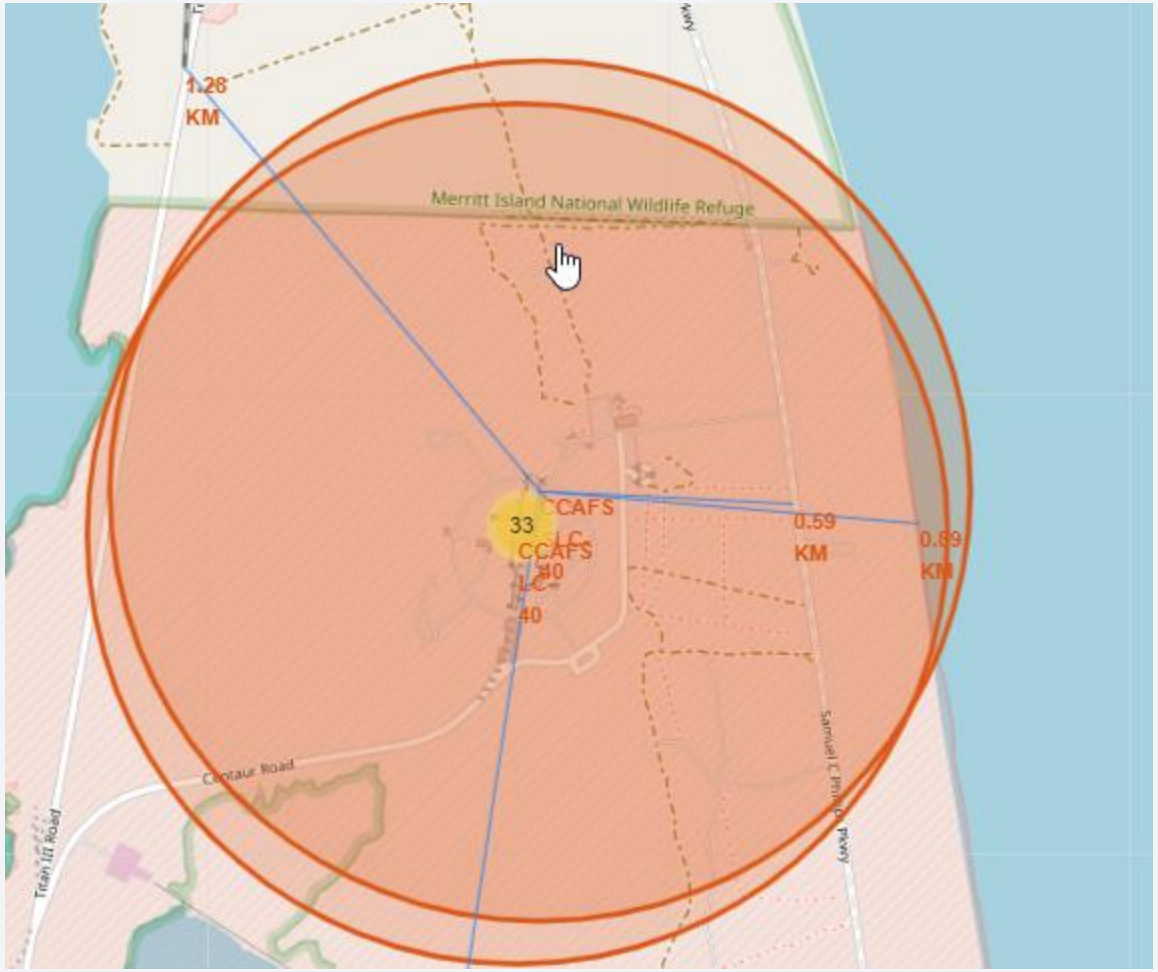
The sites in the given map are CCAFS SLC 40 (above) and CCAFS LC 40 (below).

The green marker represent success and red represent failure

The yellow one on LC 40 is a non zoom in marker we get red and green marker for it if we zoom in. The number represent total landings



CCAFS SLC 40 and it's proximities



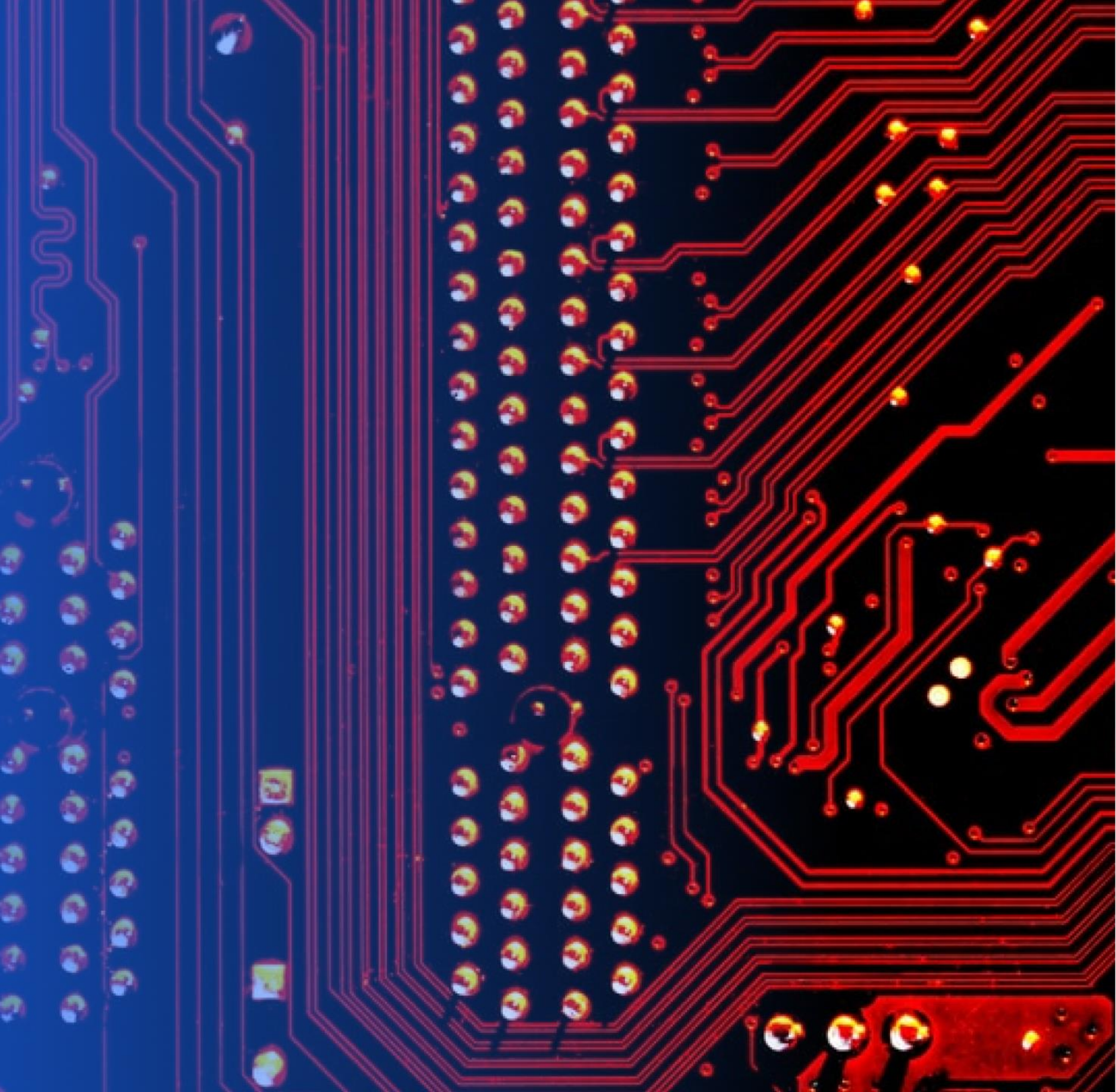
This show the distance between launch site CCAFS SLC 40 and it's proximities such as coast, highway, railroad.

The distances are:

- Coast ~ 890 m
- Highway ~ 590 m
- Railway ~ 1,280m

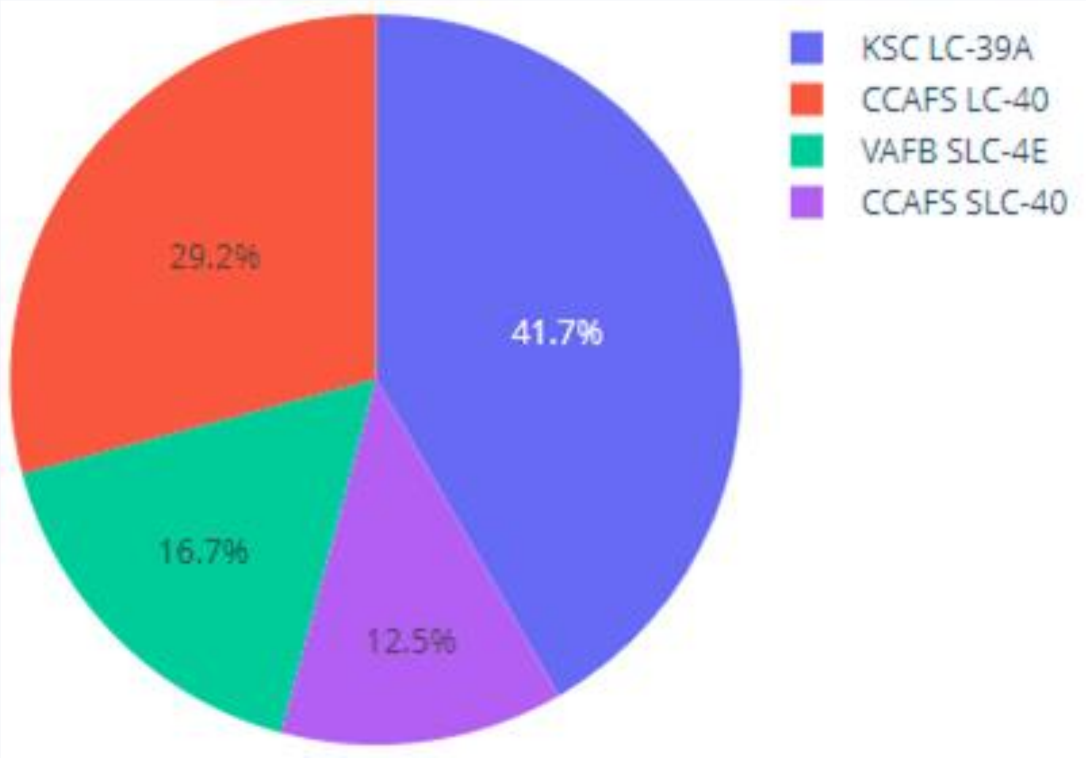
Section 5

Build a Dashboard with Plotly Dash



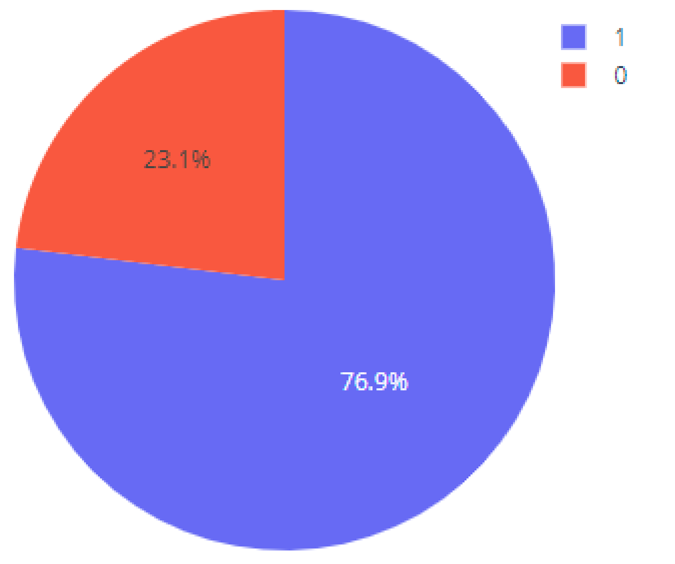
Success rate of All Sites

The pie chart shows success ratio of All Sites. We can see that KSC LC-39A High success ratio in Launch sites

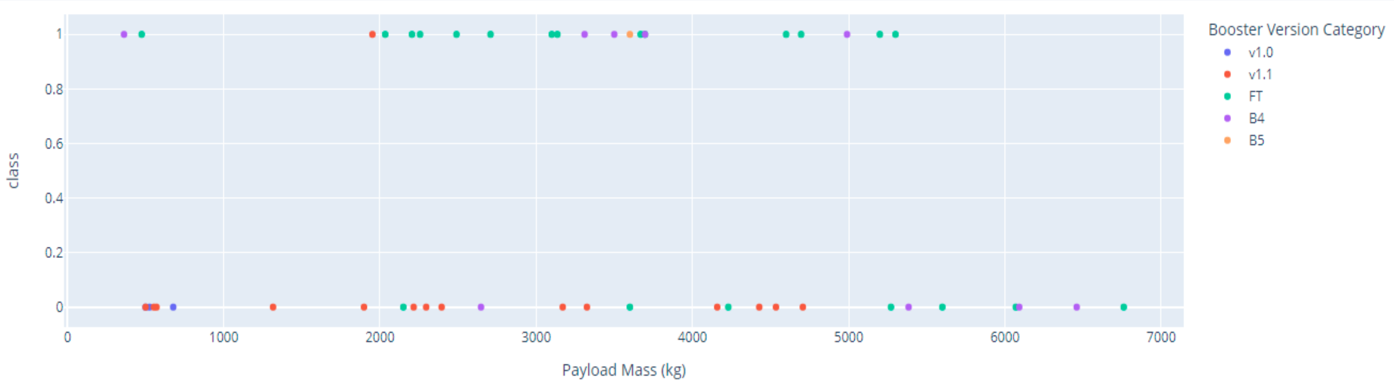


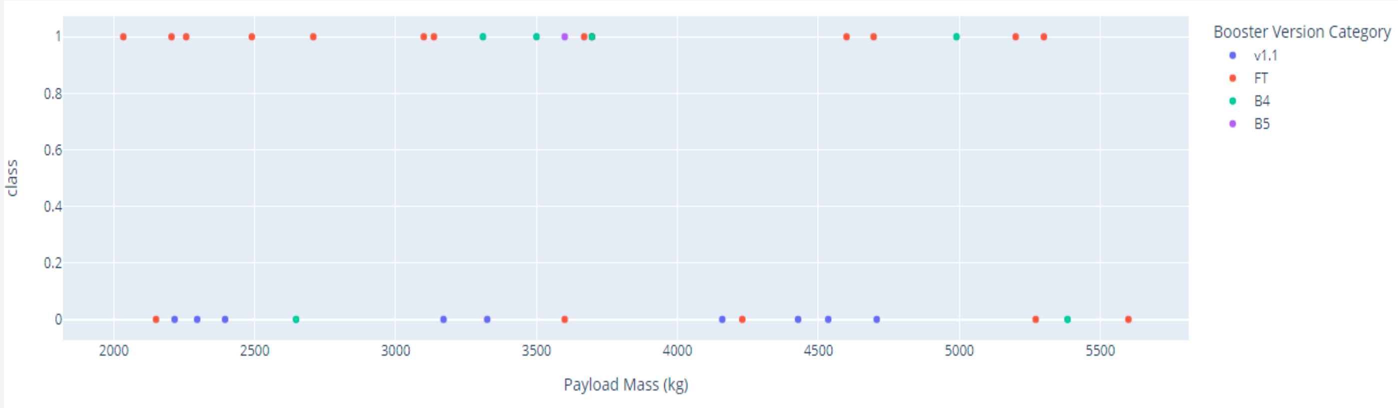
Success Percent of KSC LC-39A

The given pie chart is percent of success and failure of launches from KSC LC 39A. We can see that it have a success percent of 76.9% which is larger than any Launch site.



Payload Vs Success for All Sites



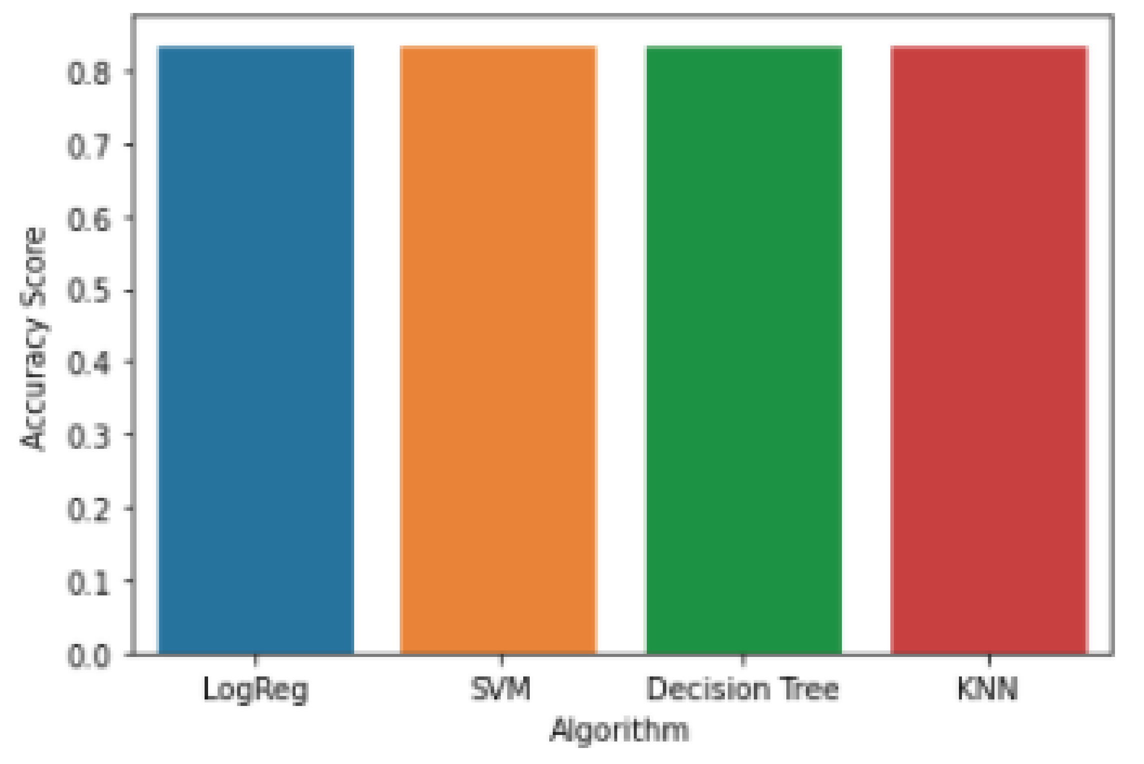


The scatter plot shows correlation between Payload and Success rate. The first one is a plot of all launches while second one is of payload between 2000kg and 6000kg. We can see that in that area the success rate is very high than any other. Also we can see that for higher payload there is low success rate.

Section 6

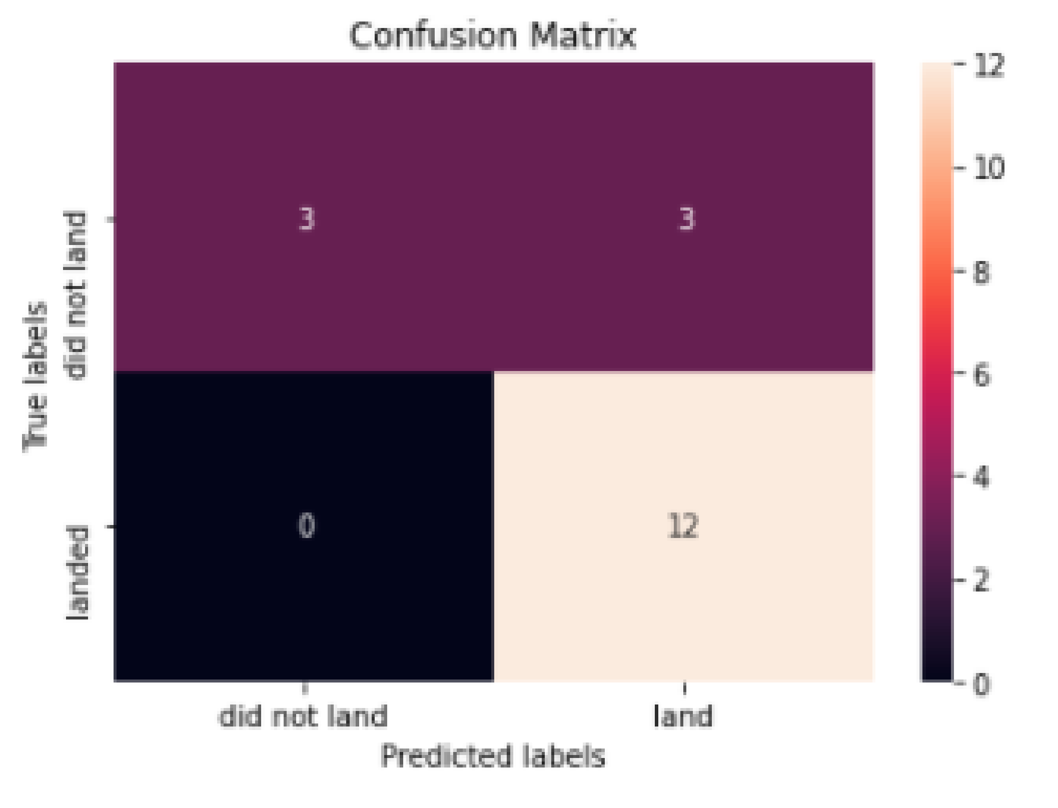
Predictive Analysis (Classification)

Classification Accuracy



All of the algorithm has a same accuracy value. So there is no best algorithm for this project.

Confusion Matrix



As I said before all of the algorithm have same accuracy So there is no best algorithm. All of them have same confusion Matrix too. Upper left 3 represent there are **3 not landed** values predicted **correctly**. Upper right 3 shows there are **3 wrongly predicted not landed** values. The lower left 0 shows that there is **there is no wrongly predicted landed values**. Lower right 12 shows there are **12 correctly predicted landed values**

Conclusions

- The best site to launch would be KSC LC-39A. Because it has higher success rate.
- The payload should be between 2000kg and 6000kg if it’s above or below the chance of failure is very high.
- The build model have a 83% accuracy.

Appendix

- All data, code and notebooks are in this GitHub repo:

https://github.com/Naseef03/IBM-Coursera

Thank you!