IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

T.Can Özay
15.10.2021

Outline

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

Executive Summary

In this study, target is to find possible solution by using python and machine learning toolkits. Web fetching by RestAPI and Beautifulsoup takes great role. Pandas module put all together dataframe in order to proceed EDA steps.

The resource indicates that it is possible to bring ~80% of accuracy solution by implementing Supervised Classification methods. However, considering the cost of such project, accuracy should be targeted to be improved in future.

Introduction

- SpaceX is targeting unique ambitions for human race and trying to reach the goal of making dream to affordable reality and it is import step and challenge to manage.
- As SpaceX attempts are highly cost oriented and we would like to measure the risk of failure and success before launching device. With such target, tried to cover possible scientific approaches which model can offer the best possible outcome of success.

Section 1

Methodology

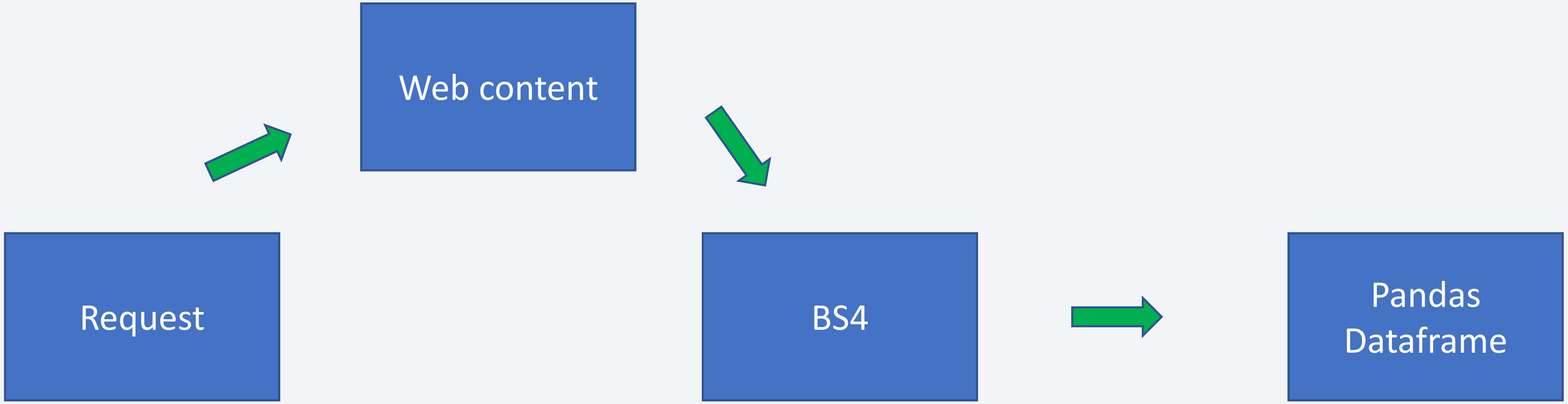
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected and covered by IBM training programme where historical data of SpaceX project were presented. Data is collected by beautiful soup & RestAPI where we can gather the data by web fetching methods.
- Perform data wrangling
  - Filtered the data for Falcon1. Beautifulsoup module helped to find all relative html content and shaped into pandas DataFrame structure.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Supervised classification problem is attempted to be solved by implementing by Logistic Regression, Decision Tree, KNN and Support Vector Machines models.

Data Collection

- Dataset is collected from several web resources by applying to gather html content in python environment.
- Beautifulsoup module helped to gather html content and find all relative tables where we put together them into dataframe format of pandas module.

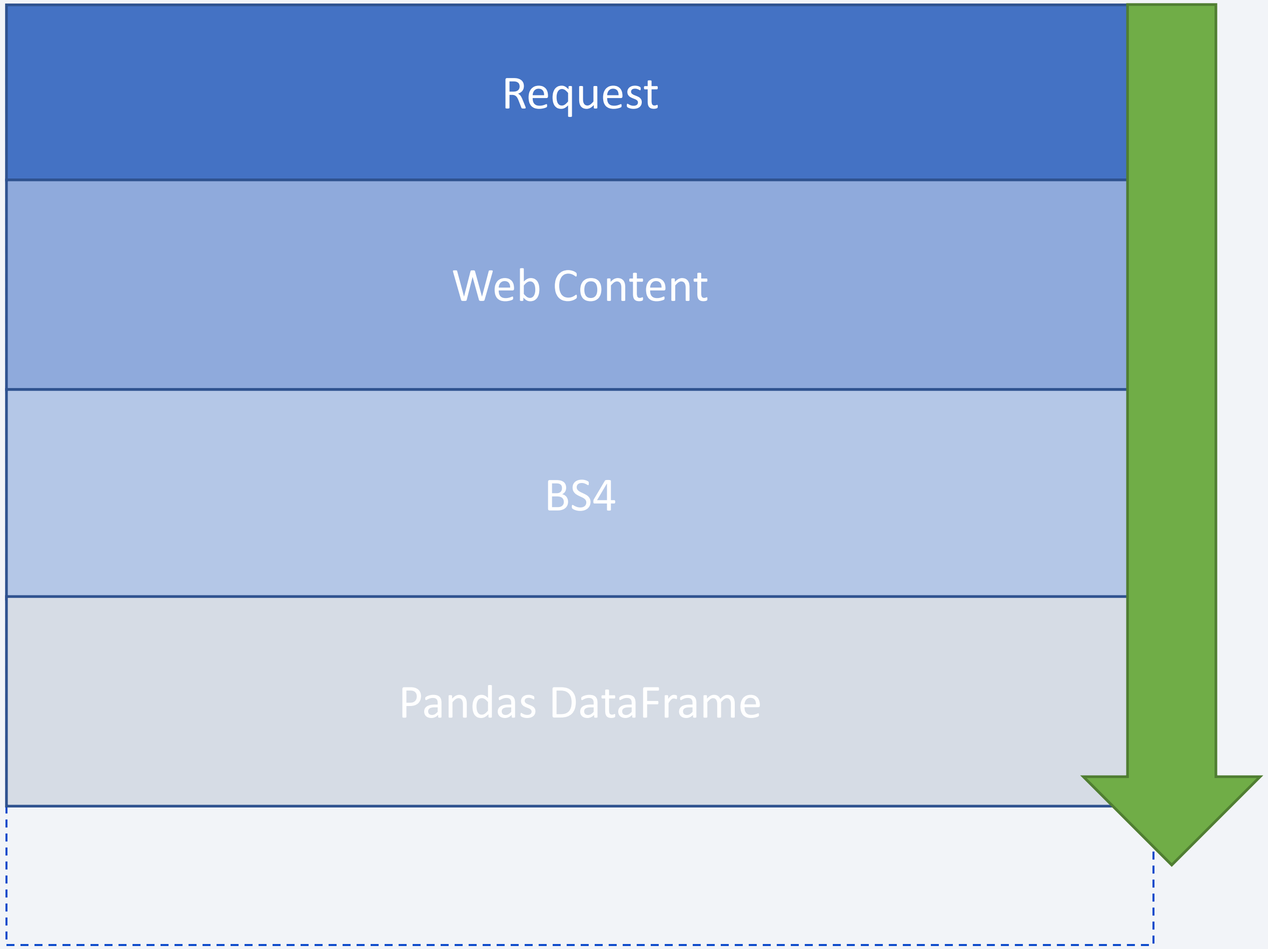


Data Collection – SpaceX API

- Please see data collection with SpaceX REST calls and flowcharts

```
def getBoosterVersion(data):
    for x in data['rocket']:
        response = requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()
        BoosterVersion.append(response['name'])
```

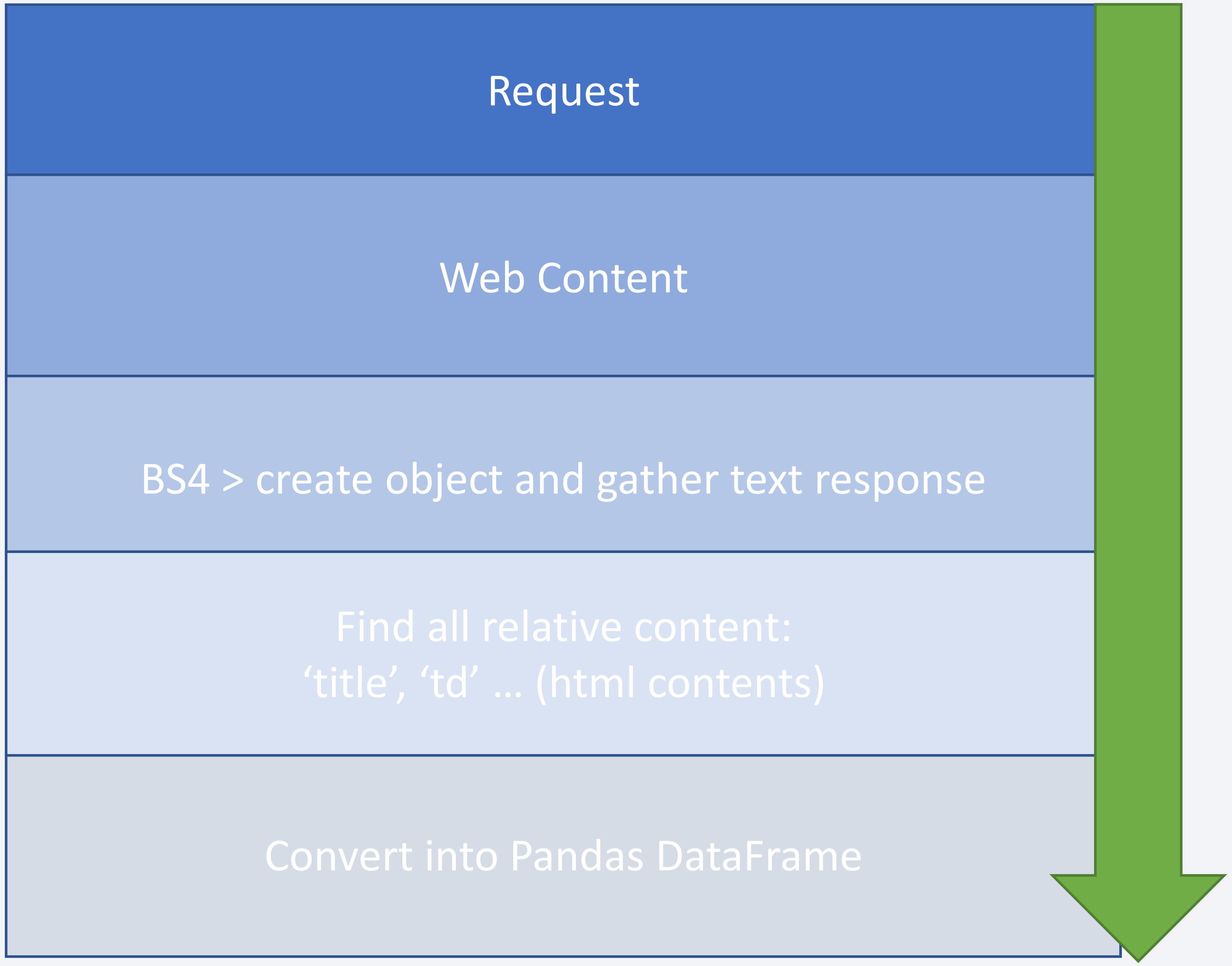
- Github:https://github.com/qosacan21/Spacex/blob/main/project1%20websc.ipynb



Data Collection - Scraping

- Please see the web scraping process using key steps and flowcharts

- GitHUB: https://github.com/qosacan21/Spacex/blob/main/project%20ws2.ipynb



Data Wrangling

- Used the method .value_counts() to determine the number and occurrence of each orbit in the column Orbit!
- Created a list where the element is zero if the corresponding row in Outcome is in the set bad_out!
- comeGithub: https://github.com/qosacan21/Spacex/blob/main/EDA1.ipynb

EDA with Data Visualization

- Seaborn and Matplotlib plotting techniques are used for data visualization.
- Applied catplot chart to visualize observation distributions of PayloadMass and FlightNumber in order to understand whether there is impact on success ratio (class variable) or not!. Also same method applied between Launch Sites and FlightNumber to understand whether there is impact on success ratio.
- Scatter point and bar charts applied to understand if there is significant or meaningful difference between studied groups.
- Github: https://github.com/qosacan21/Spacex/blob/main/EDA2%20viz.ipynb

EDA with SQL

- Applied simple SQL queries to understand relationship between variables.
- To observe unique values!
- To observe count amounts!
- To observe max, mean and sum values!
- Applied conditional filters by where command.
- Applied order by command for ascending or descending orders.

Build an Interactive Map with Folium

- Added Launch Sites coordinates to map and labelled them.
- In case of detailed view, possible to zoom in the map and see success/fail attempts of each launches per Launch Site locations.
- Objects are added to map in order to understand success ratio of attempts per location and understand the location of importance how such big projects can be supported logistically (so you can observe airports, railways … and coast line distances) And finally estimated the distances between a launch site to its proximities.
- https://github.com/qosacan21/Spacex/blob/main/Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- There are two main plot techniques which applied into the project. First one is pie chart and second one is scatter plot.
- Interactions are added to filter out different condition and compare the visual results. Dropdown component filters out the entered Launch site and range components collects min and max filtered values of Payload KGs and then filters conditions.

- Github:
- PDF> https://github.com/qosacan21/Spacex/blob/main/mydash.pdf
- PY> https://github.com/qosacan21/Spacex/blob/main/spacex_dash_app%20(1).py

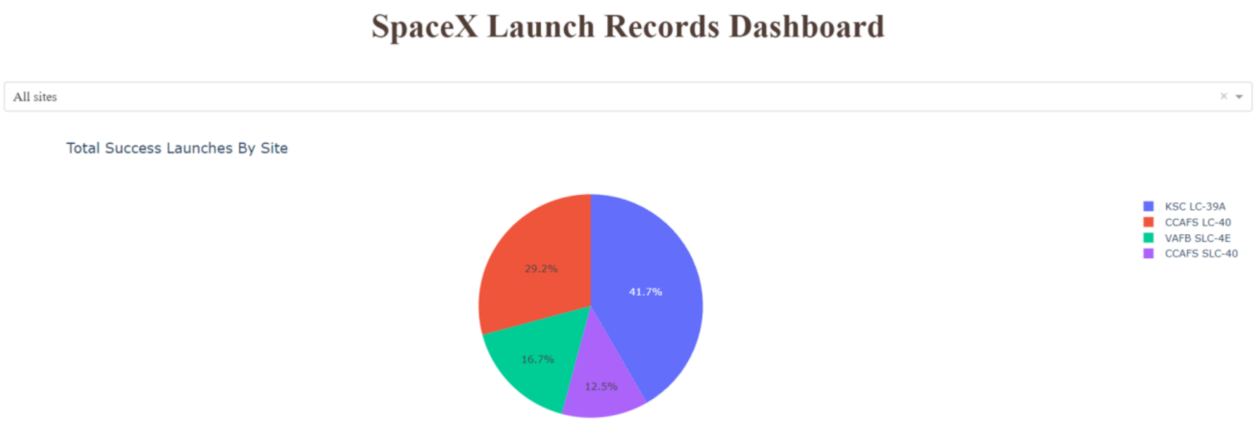
Predictive Analysis (Classification)

- Supervised classification methods applied in order to solve the problem.
- Firstly Gridseach CV method implemented of training data in order to understand the best tuning parameters for Logistic Regression, SVM, KNN and Decision Trees method. Then it is followed by testing the developed models in test data. The reason we use the testing data is observing whether model performs similar accuracy on different data and avoiding overfitting.

- https://github.com/qosacan21/Spacex/blob/main/ml.ipynb

Results

- Exploratory data analysis results shows that PayloadMass decisive indicator on success ratio
- Interactive analytics dashboard:



- Predictive analysis results can be seen below the classification methods have approximately same accuracy on Test data!

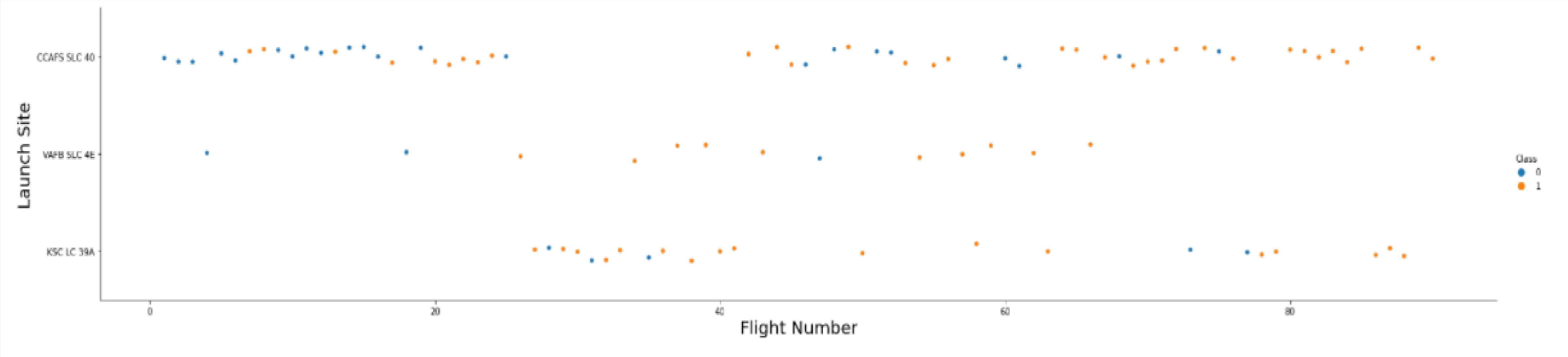
	Test Score
Logistic Regression	0.833333
SVM	0.833333
Tree	0.833333
KNN	0.833333

Section 2

Insights drawn from EDA

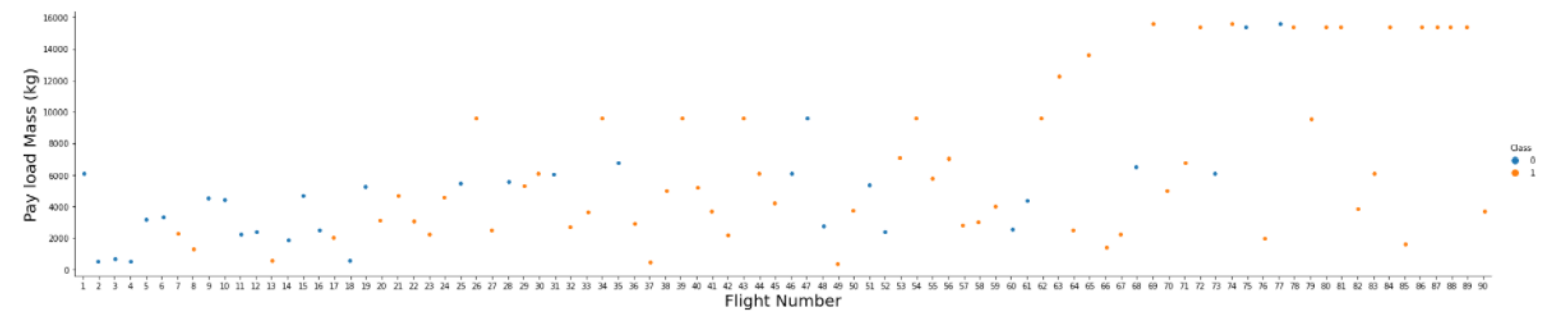
Flight Number vs. Launch Site

- Below graph shows that first flight numbers are more likely to fail than latest flight numbers.

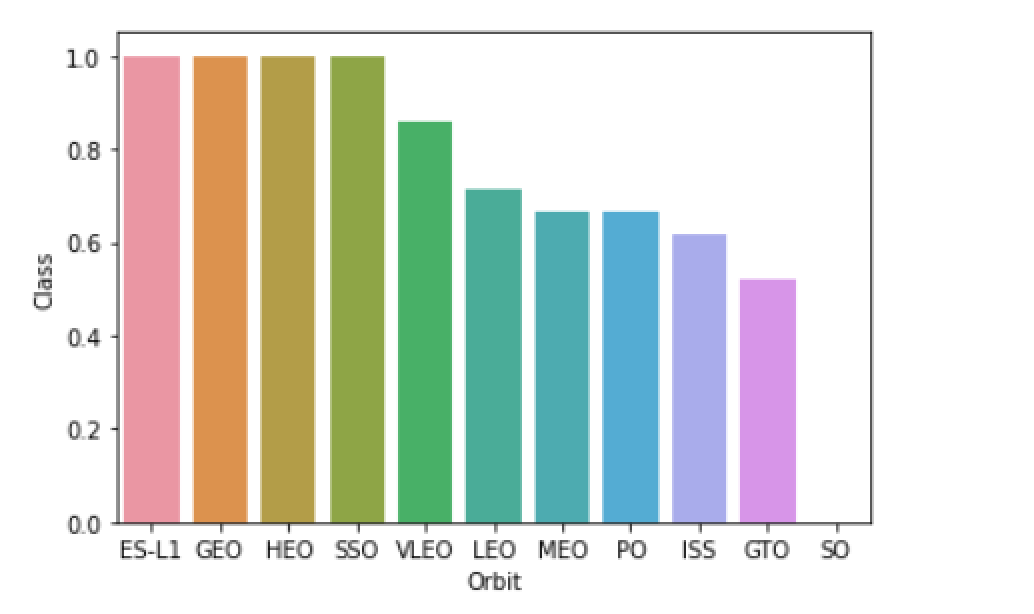


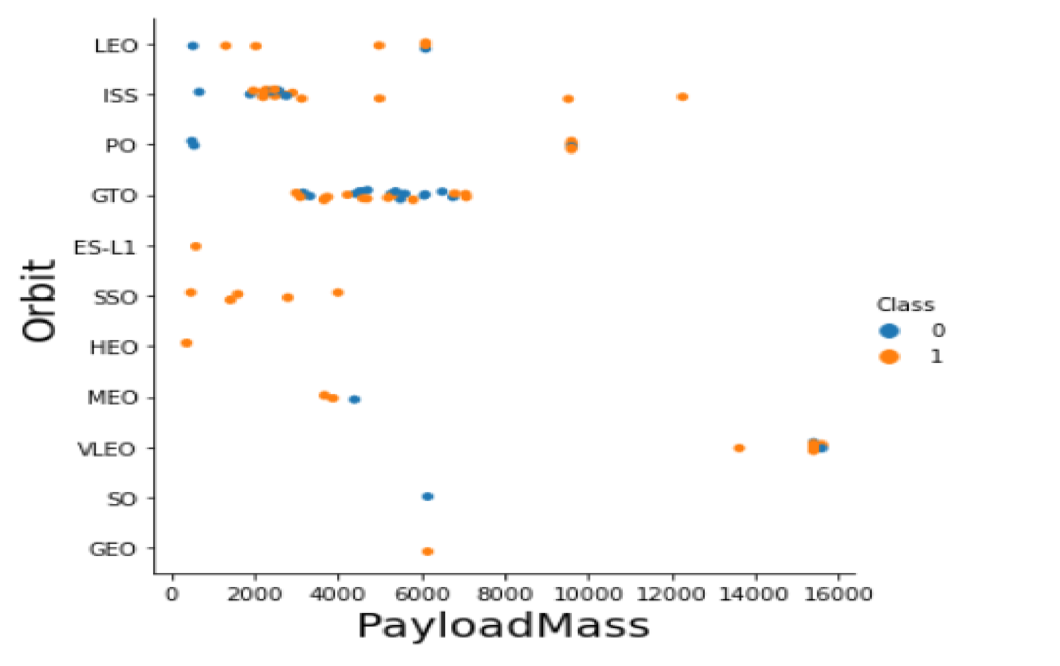
Payload vs. Launch Site

- Payload versus Launch Site scatter plot graphs is not able to provide the clear picture that we can conclude certain output but heavier Playload mass indicates the higher success ratio.



Success Rate vs. Orbit Type

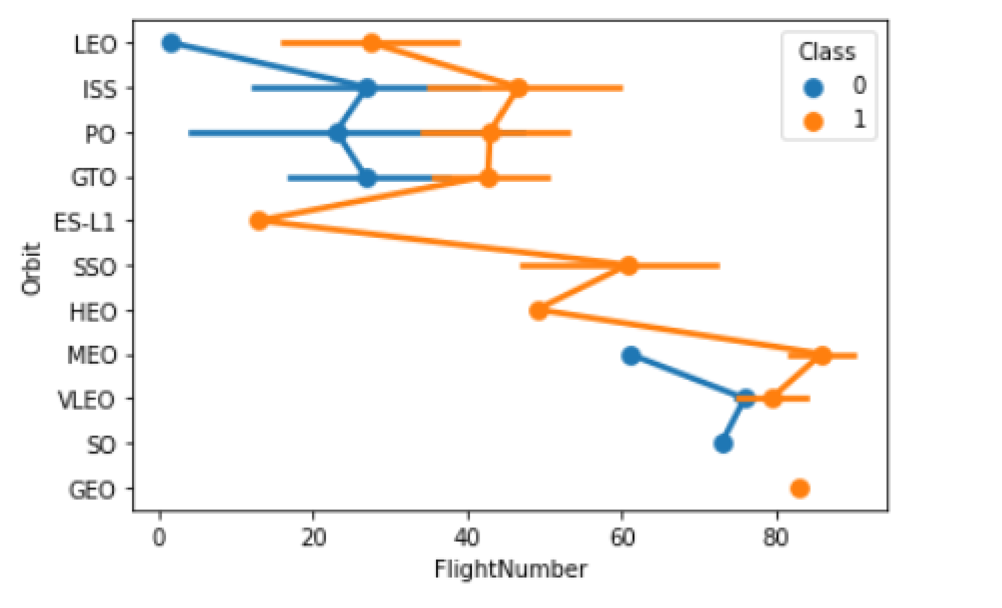




- Bar chart indicates that ES-L1, GEO, HEO and SSO orbits more likely to be successful. Meanwhile ISS, GTO and SO have lowest success ratio!
- Sactter plot contributes to understand that Payloadmass has impact on success ratio!

Flight Number vs. Orbit Type

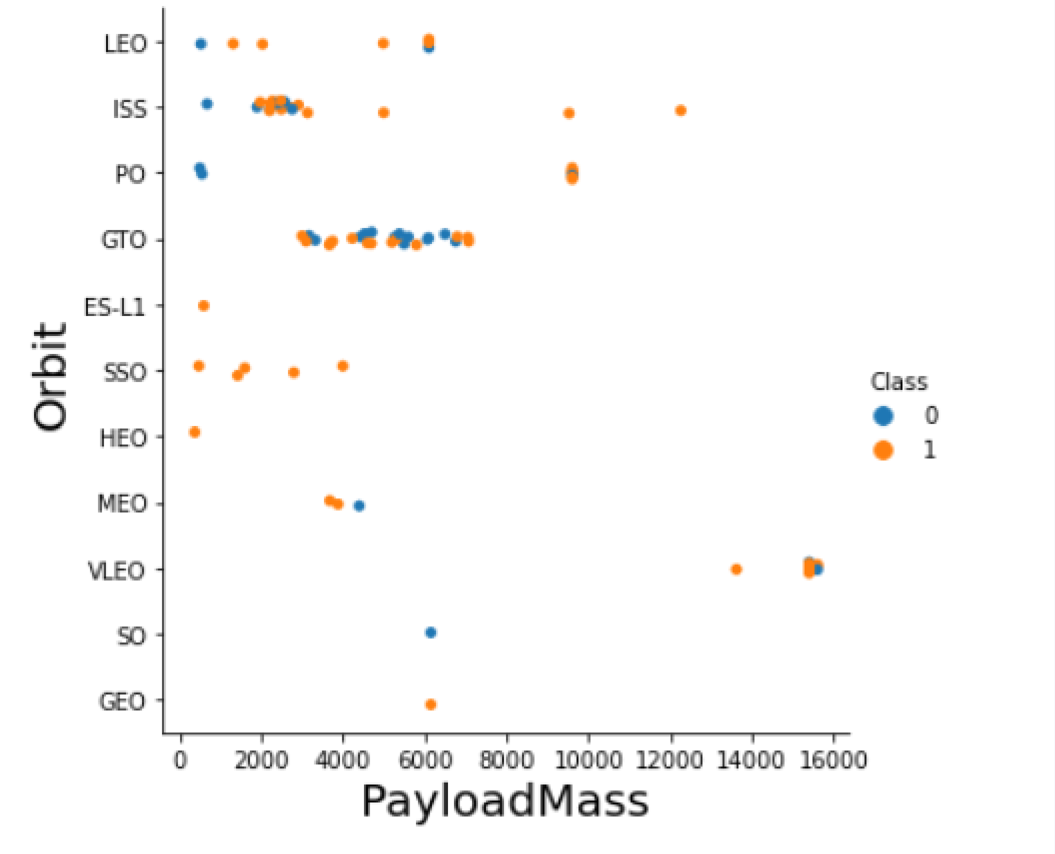
Orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

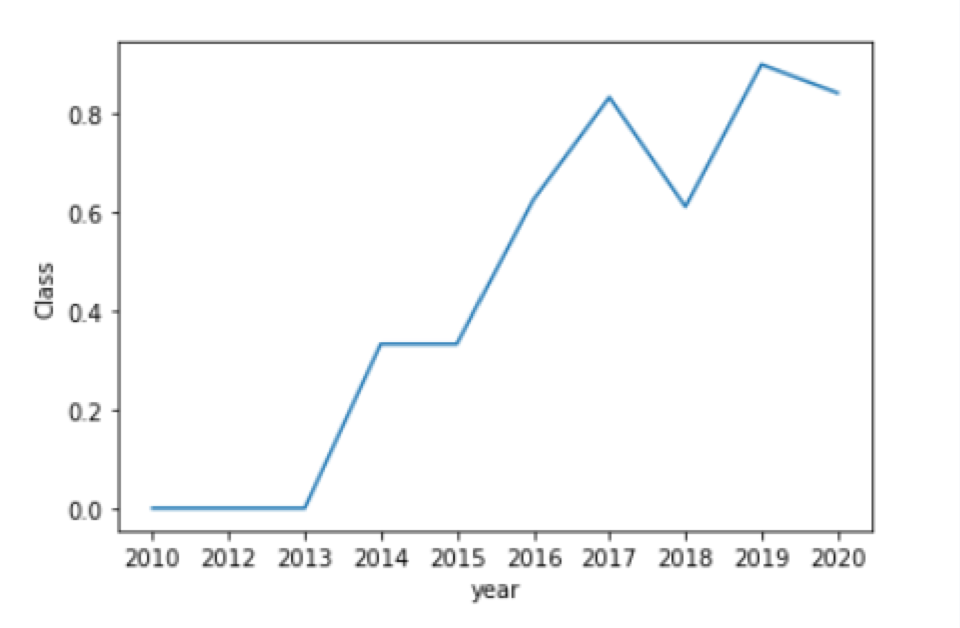
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

- However for GTO we cannot distinguish this well as both positive landing rate and negative landing.



Launch Success Yearly Trend

Sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- Unique Launch sites from the result of query.
- There are 4 different Launch sites.

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
0	4/6/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	8/12/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	8/10/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	1/3/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculated the total payload carried by boosters from NASA: 45596
- SELECT SUM(PAYLOAD_MASS__KG_) FROM DF WHERE Launch_Site = ‘NASA (CRS)’;

45596

Average Payload Mass by F9 v1.1

- Calculated the average payload mass carried by booster version F9 v1.1: 2928.4
- SELECT SUM(PAYLOAD_MASS__KG_) FROM DF WHERE Booster_Version = ‘F9 v1.1’;

2928.4

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad:

8/4/2016

- SELECT Date FROM DF ORDER BY Date LIMIT 1;

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

```
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1
```

- SELECT Booster_Version FROM DF WHERE (Landing _Outcome = “Success (ground pad)”) AND (PAYLOAD_MASS__KG_ > 4000) AND (PAYLOAD_MASS__KG_ < 6000);

Total Number of Successful and Failure Mission Outcomes

- Calculated the total number of successful and failure mission outcomes

Success	98
Success	1
Failure (in flight)	1
Success (payload status unclear)	1

- SELECT COUNT(Mission_Outcome) FROM DF;

Boosters Carried Maximum Payload

- Listed the names of the booster which have carried the maximum payload mass

F9 B5 B1060.2
F9 B5 B1051.6
F9 B5 B1051.4
F9 B5 B1048.5
F9 B5 B1056.4
F9 B5 B1049.5
F9 B5 B1051.3
F9 B5 B1058.3
F9 B5 B1048.4
F9 B5 B1060.3

- SELECT Booster_Version FROM DF ORDER BY PAYLOAD_MASS__KG_ DESC;

2015 Launch Records

- Listed the records which displays the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

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

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

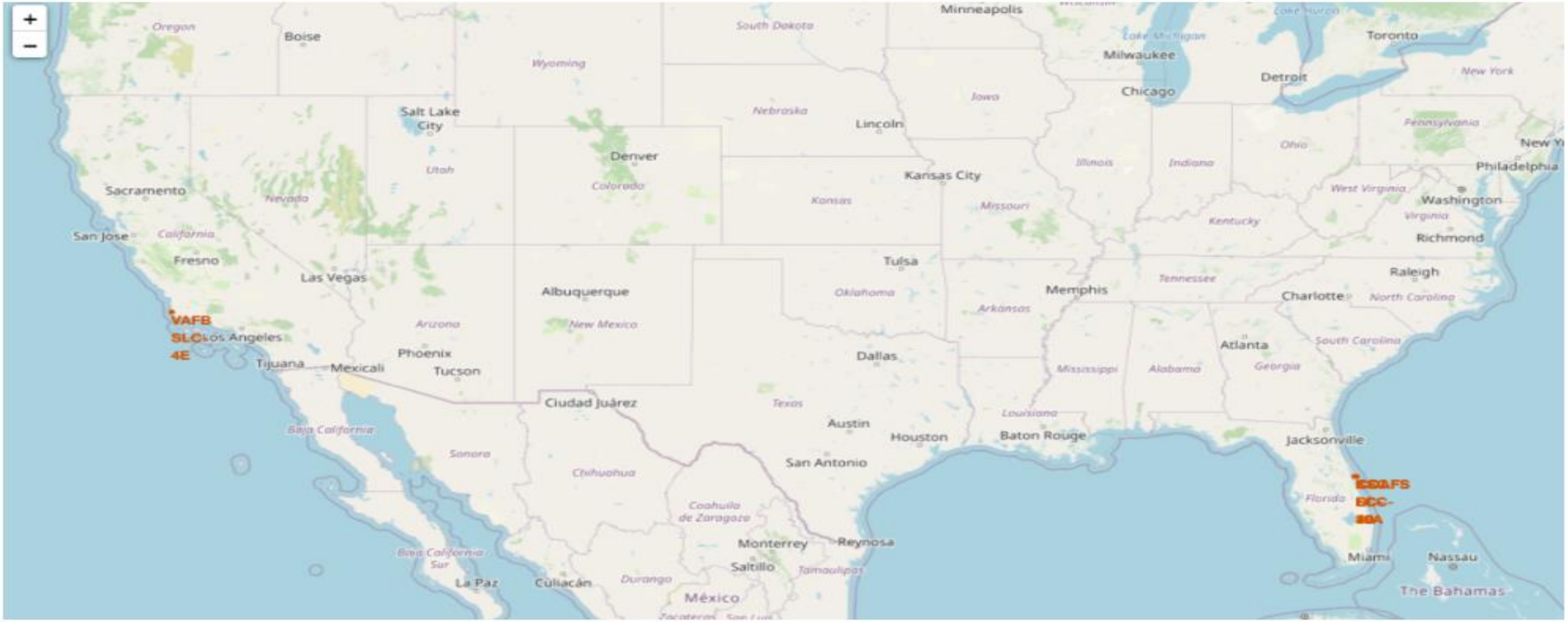
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-06	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-01-14	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
2017-01-05	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-04	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-06-15	14:29:00	F9 FT B1024	CCAFS LC-40	ABS-2A Eutelsat 117 West B	3600	GTO	ABS Eutelsat	Success	Failure (drone ship)
2016-06-05	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)

Section 4

Launch Sites Proximities Analysis

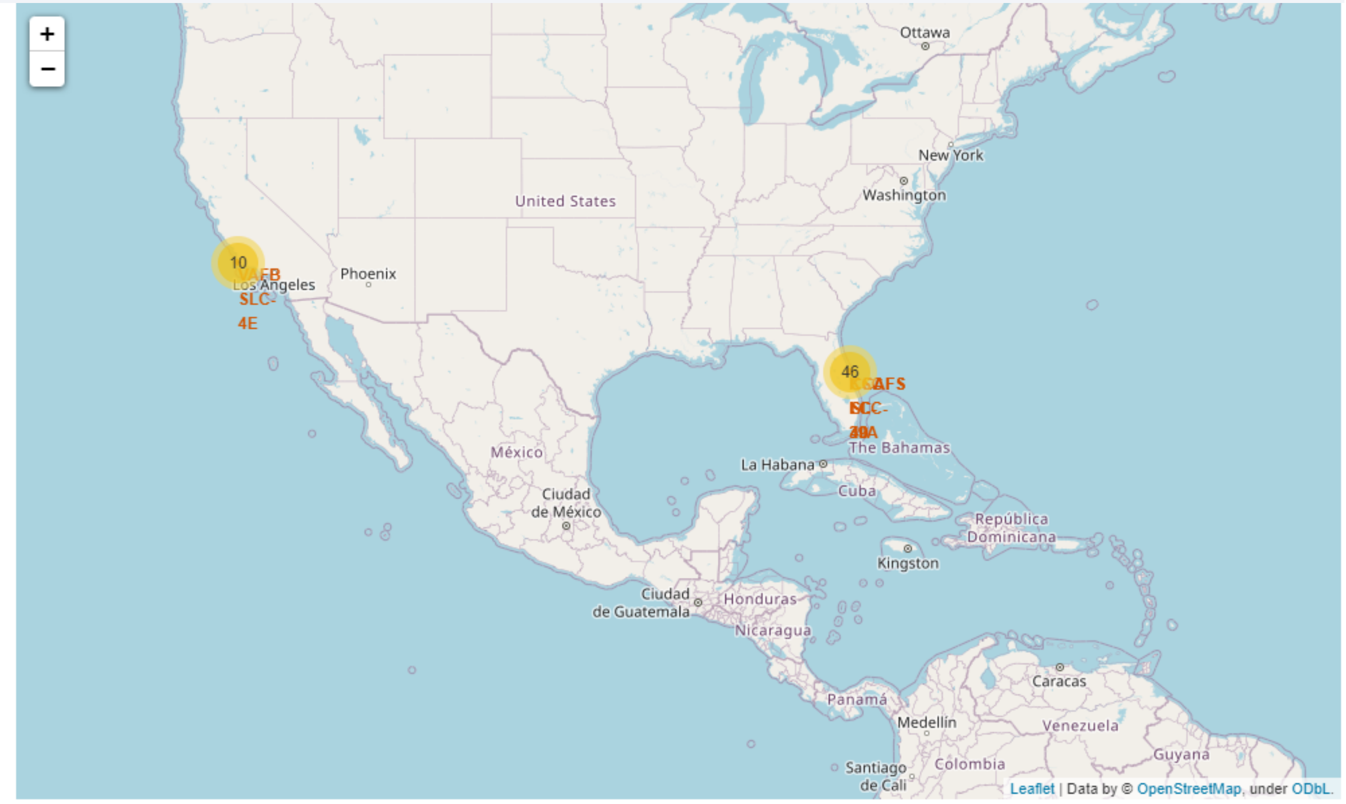
Launch Site Location – Folium 1

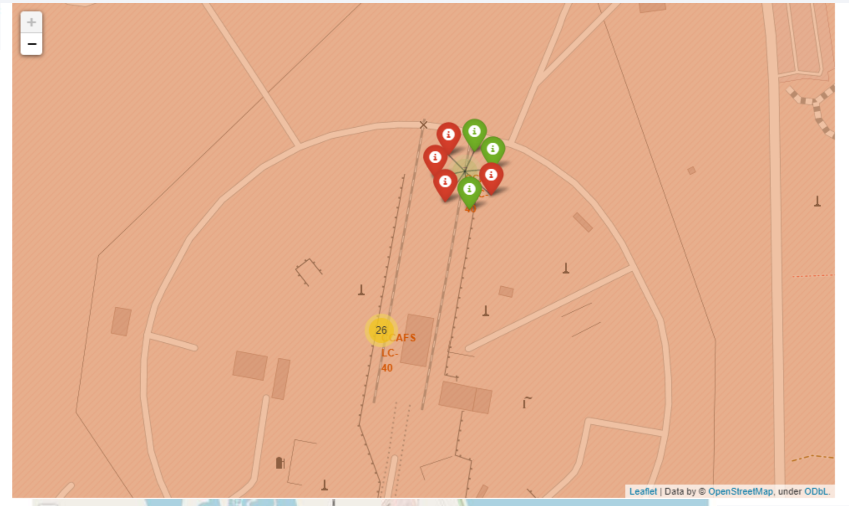
- All launch site locations can be seen below. Common points of each launch sites are being close to costal area and near the transportation points such as railway, airport …

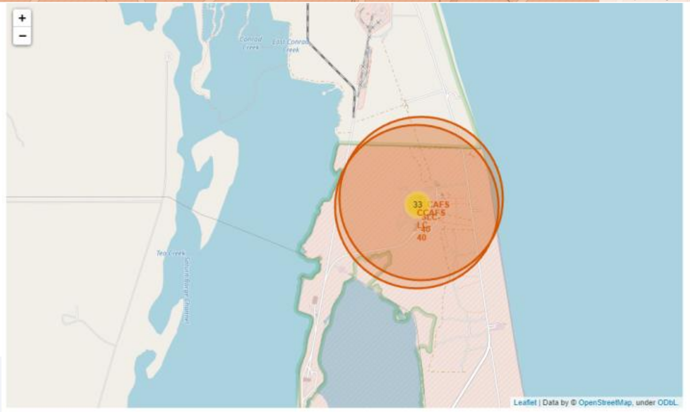


Launch Site Location – Folium2

- From the color-labeled markers in marker clusters, we can easily identify which launch sites have relatively high success rates by zooming in and out.

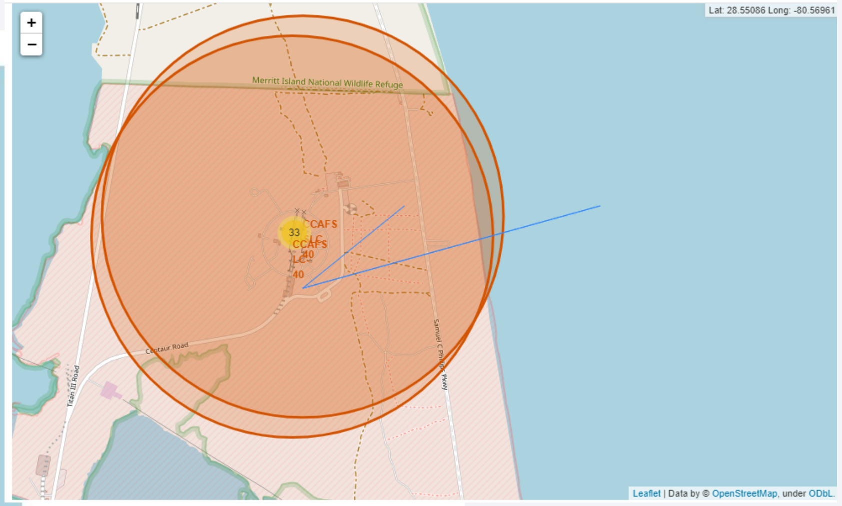


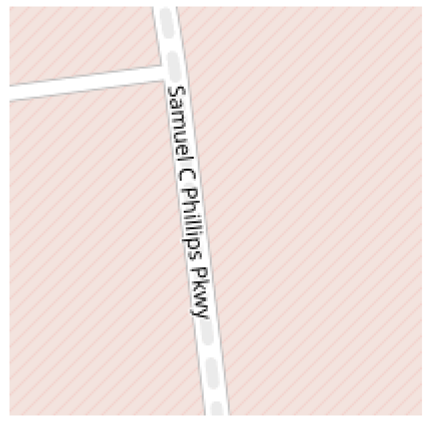




Launch Site Location – Folium3

- We can see that Launch site is very close the coast. And there are highway road and railway around.



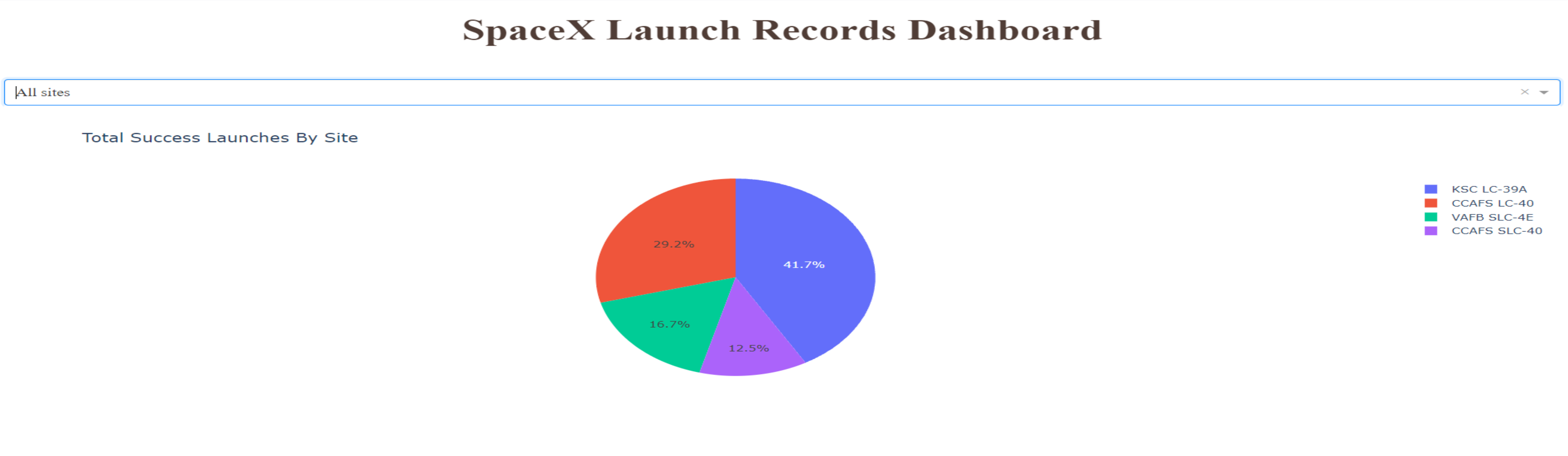


Section 5

Build a Dashboard with Plotly Dash

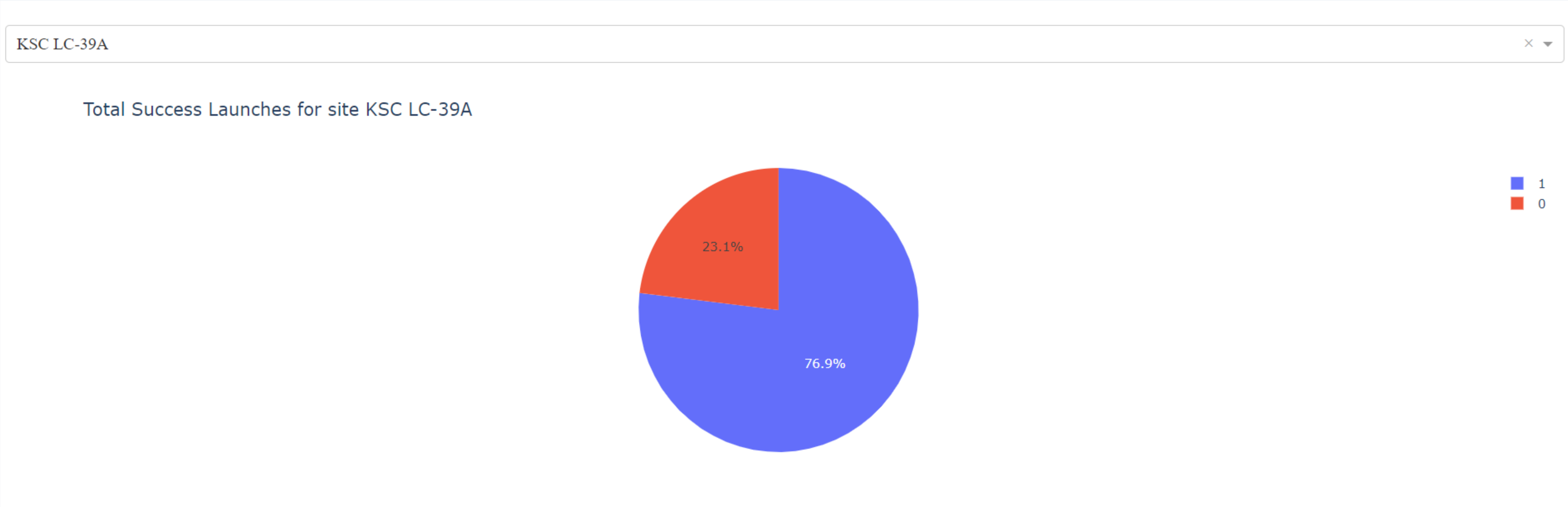
Total Success Launches By Site (DASH)

- We can observe from below pie chart that KSC-LC-39A Launch Site location have highest total success figures!



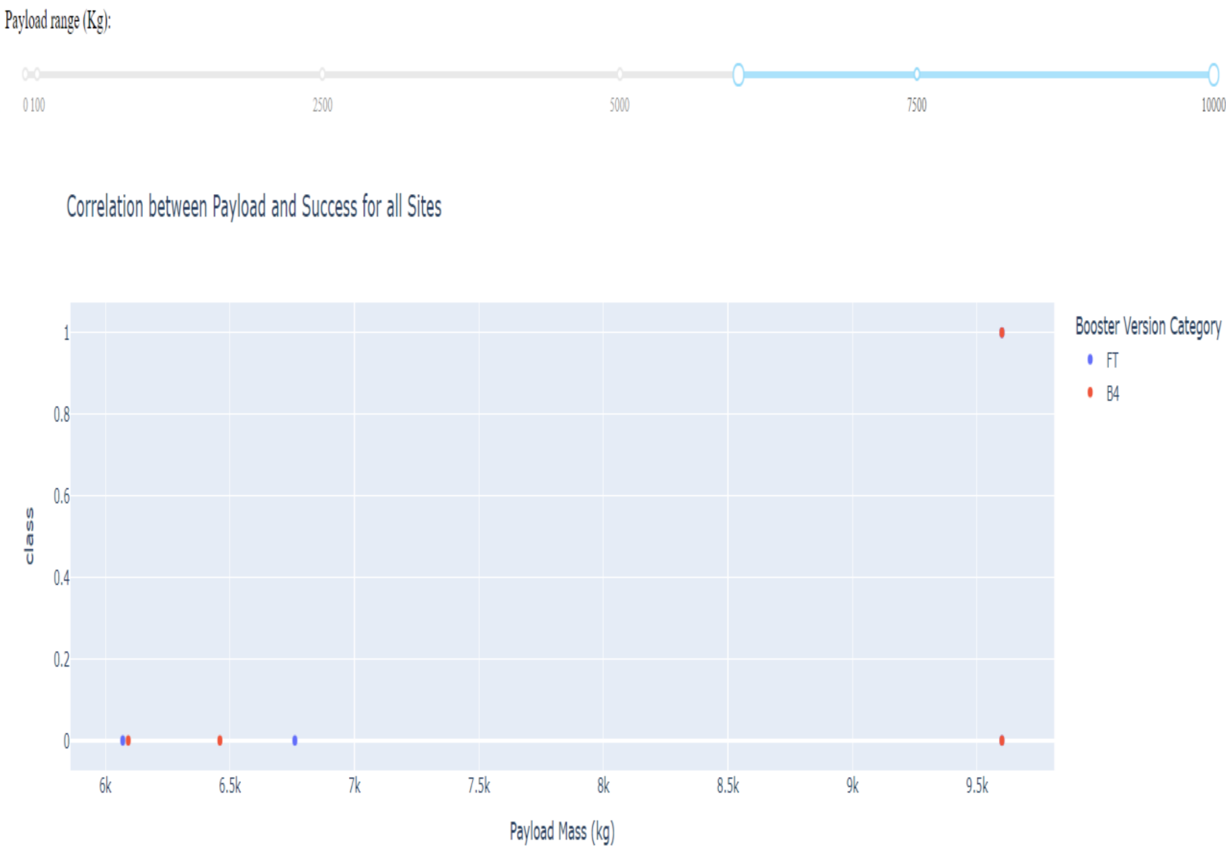
Highest Launch Success Ratio(DASH)

- Pie chart shows that for the launch site with highest launch success ratio:



Role of Payload Amount on Success Ratio(DASH)

- We can observe that lighter payload had higher success ratio than heavy ones!

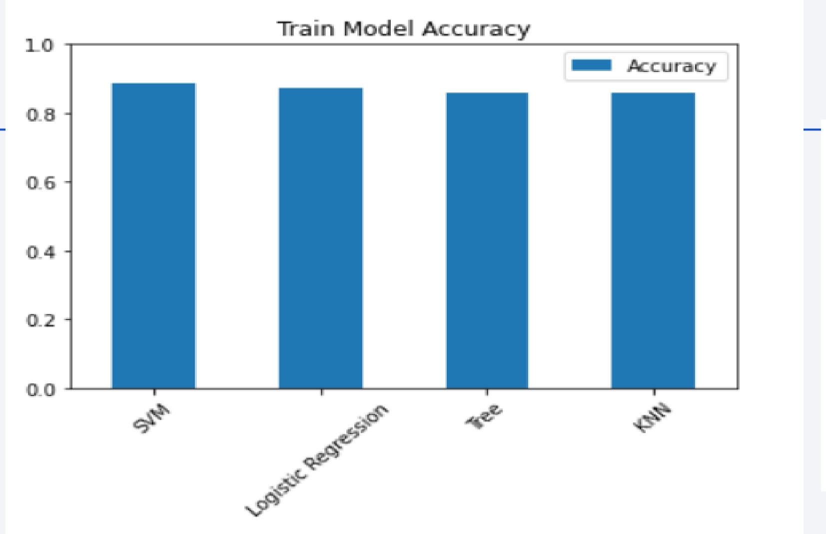


Section 6

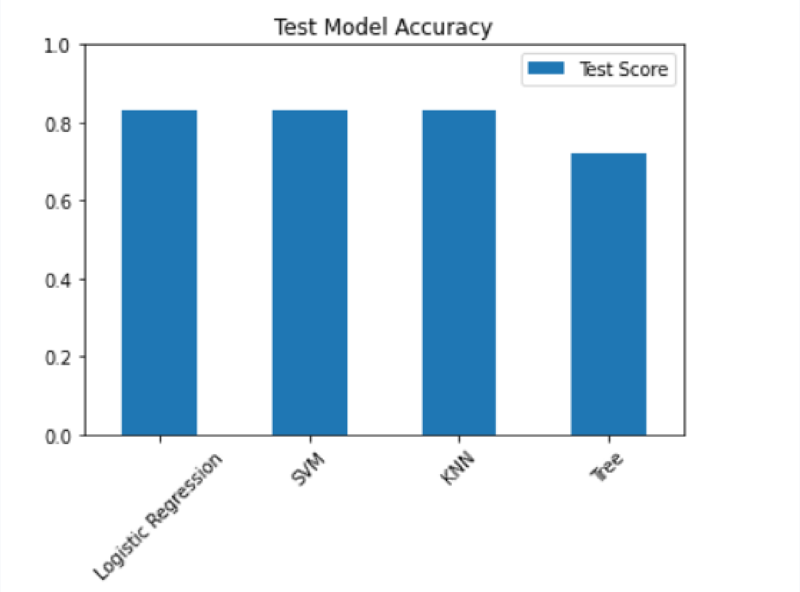
Predictive Analysis (Classification)

Classification Accuracy

Once we only focus the result of Train Model Accuracy, then SVM accuracy is slightly higher than others. However, Test model accuracy is more reliable and it indicates very same score for several models which involves Logistic Regression, SVM and KNN.



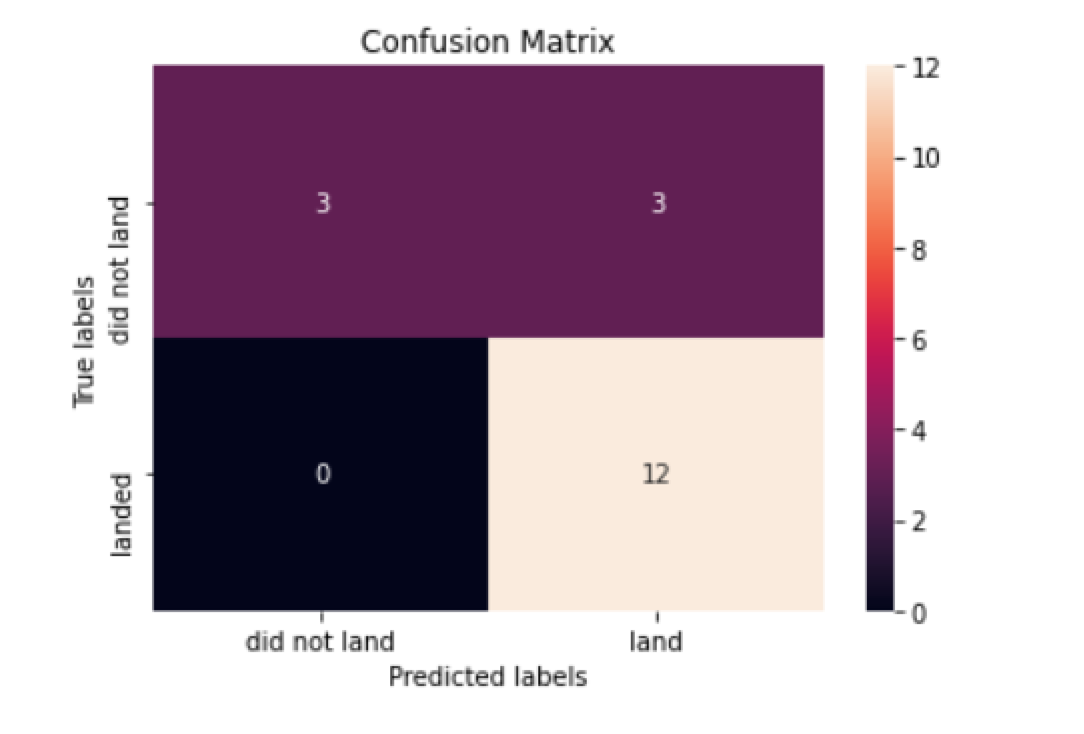
	Accuracy
SVM	0.888889
Logistic Regression	0.875000
Tree	0.861111
KNN	0.861111



	Test Score
Logistic Regression	0.833333
SVM	0.833333
KNN	0.833333
Tree	0.722222

Confusion Matrix

- As we can observe from confusion matrix. Only 3 observations are mislabeled by model. Even though 3 observations actually did not land successfully, we predicted the probability of them that they can land.

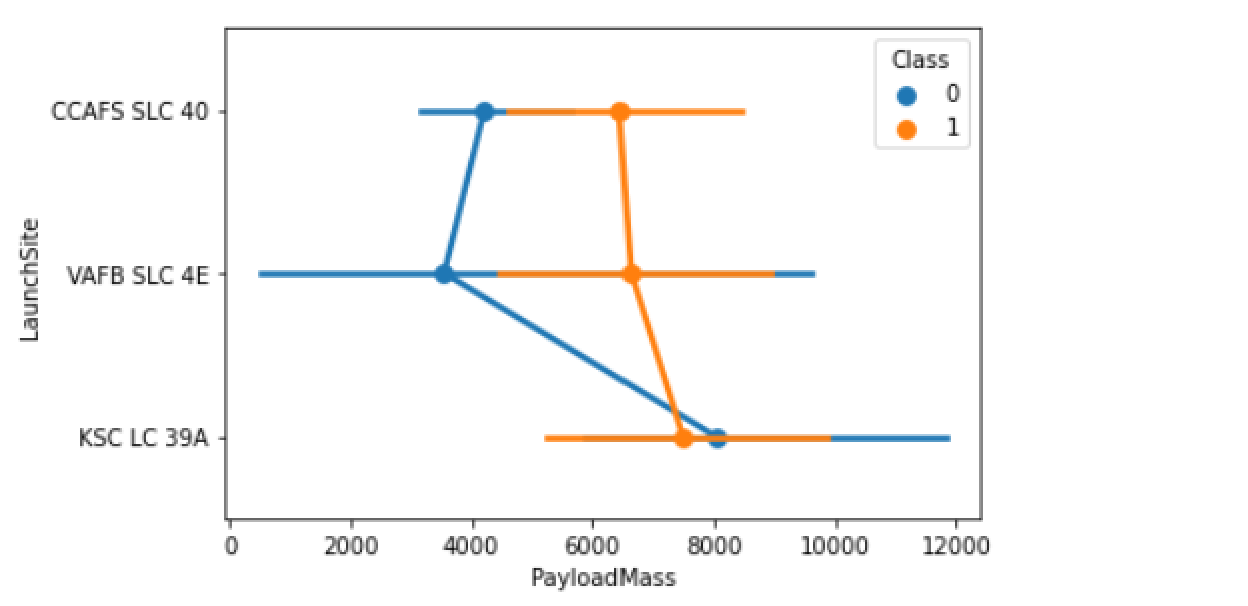


Conclusions

- It is highly possible to calculate outcome of models by Supervised Classification methods like Logistic Regression, SVM or KNN methods.
- Possible to calculate over 83% accuracy with limited amount of data.
- Payload mass important indicator.
- Success ratio of launches have a better results in some certain launch sites.
- Launch site location should be close coast and transportation point.

Appendix

- https://github.com/qosacan21/Spacex
- Payload vs Launch Site



Thank you!