IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Murat An
08.11.2022

Outline

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

Executive Summary

- **Summary of methodologies**
- Data Collection
- Data Wrangling
- EDA with Data Visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis

- **Summary of all results**
- Exploratory Data Analysis Results
- Interactive analytics Demo in Screenshots
- Predictive Analysis Results

Introduction

- Project background and content
  - The future lies on the commercial space
  - SpaceX is one of the companies to take advantage of space race and make profit
  - SpaceX has the purpose of establishing a vast connecting of internet by satellites, sending spacecraft to the International Space Station, and sending manned mission into space.

- Problems you want to find answers
  - We want to find out what is the success rate of SpaceX's rocket launch mission.
  - Predict the future success and find out under what condition the missions are the most succesful

Section 1

Methodology

Methodology

Executive Summary

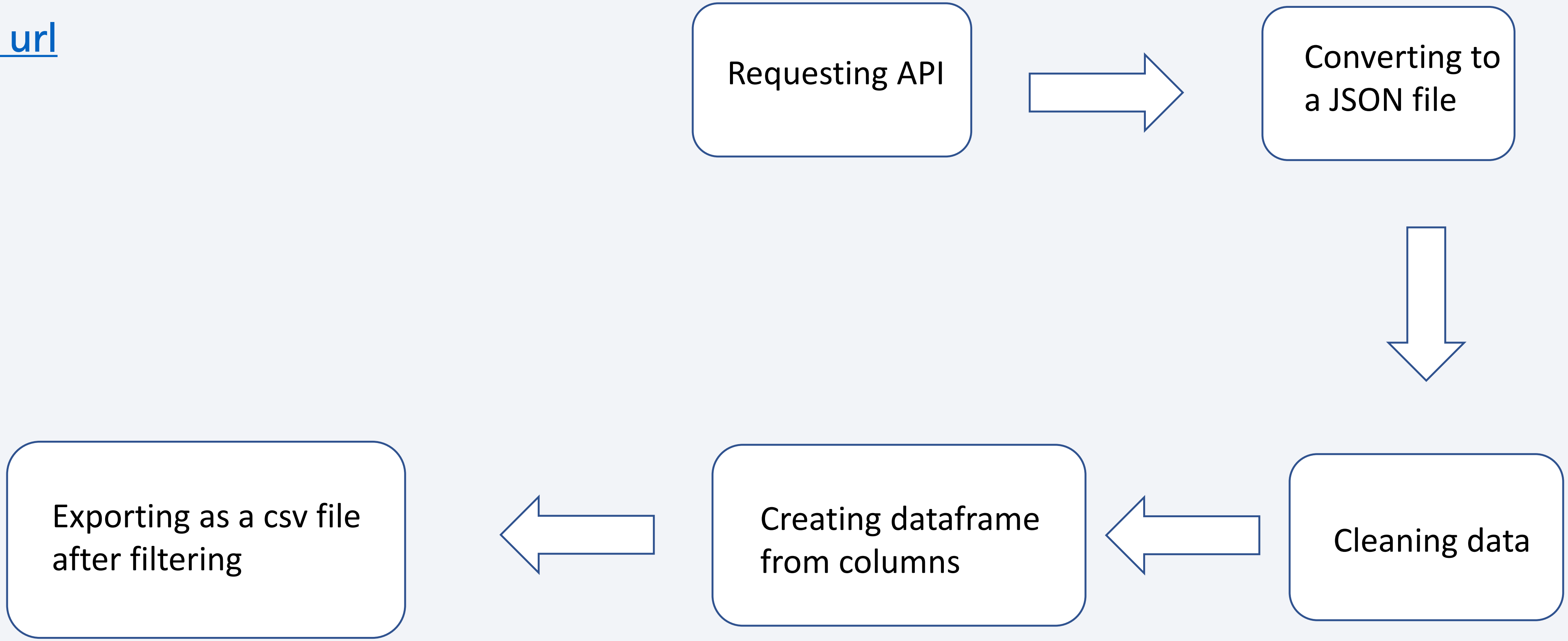
- Data collection methodology:
  - Web Scraping from SpaceX Wikipedia page and requesting Space API
- Perform data wrangling
  - Filtration and replacing missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - SVM, Classification Trees and Logistic Regression models are used by finding best hyperparameters

Data Collection

- First, we collect data from Wikipedia page through Web Scraping and SpaceX API through API requests processes
- Then we get the data columns from API : FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- From Wikipedia: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch Outcome, Version Booster, Booster landing, Date, Time

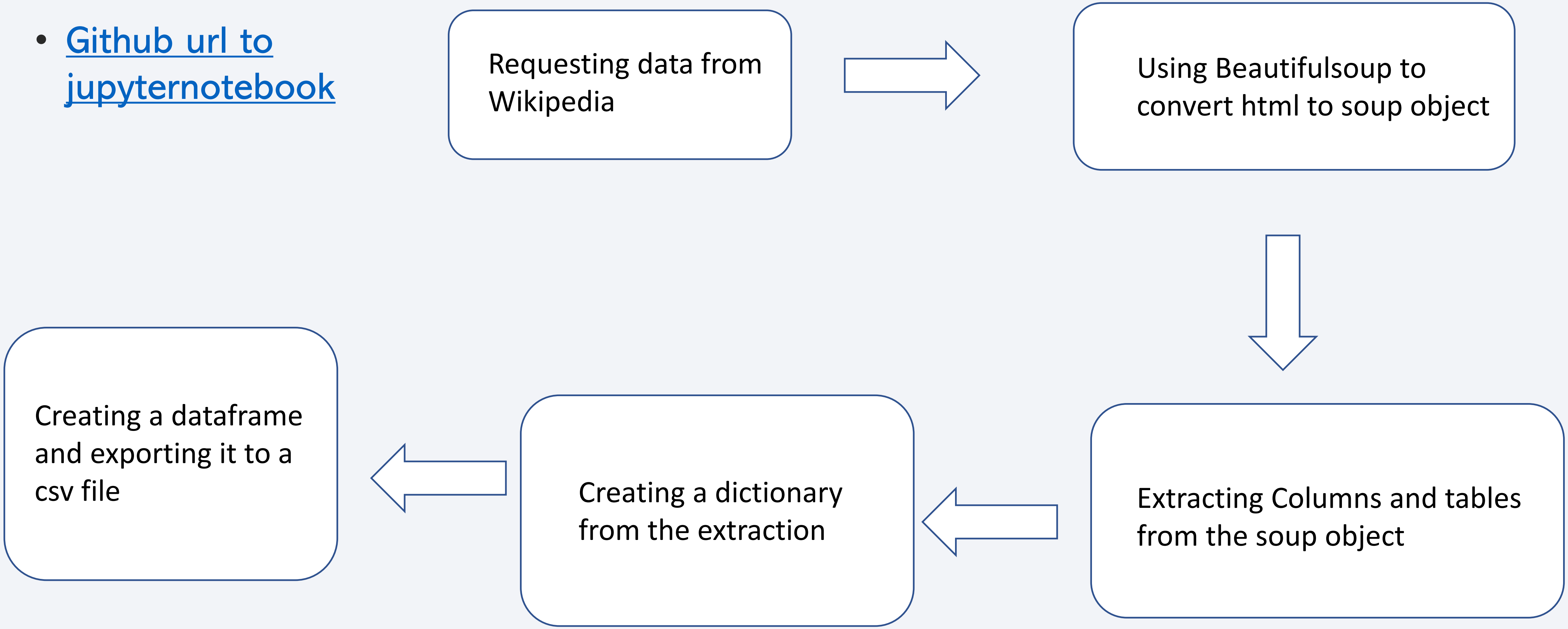
Data Collection – SpaceX API

- [Github url]()



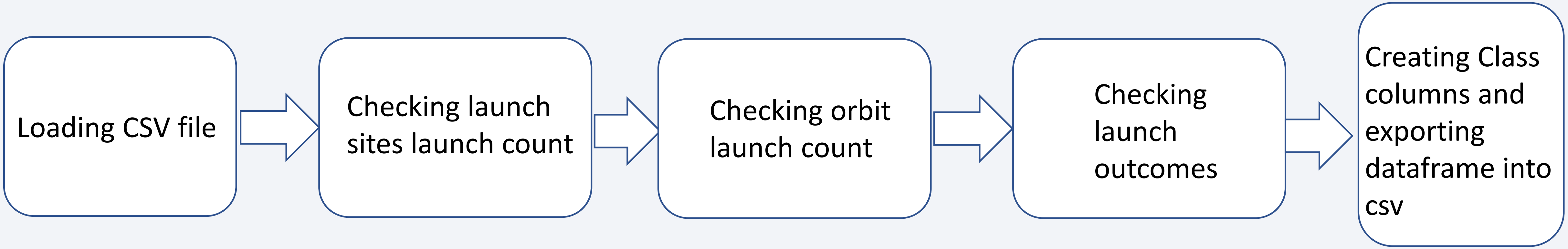
Data Collection - Scraping

- [Github url to jupyternotebook]



Data Wrangling

- Data wrangling is based on success landing and failing landing on the dataset
- Covert the results into binary integers as successful =1 and failure=0 into new column ‘Class’



[github url jupyter notebook]

EDA with Data Visualization

- Scatter Charts:
  - Flight Number vs Launch Site, Payload vs Launch Site, Flight Number vs Orbit, Payload vs Orbit
  - A scatter chart is a good way to see correlation between variables
- Bar Chart:
  - Orbit vs Success Rate
  - A bar chart is ideal for classification inside a column
- Line Chat:
  - Year vs Success Rate
  - A line chart shows the changes with dynamic variable like time

github url jupyter notebook

EDA with SQL

- SQL queries:
    - Names of the unique launch sites
    - 5 records of launch sites begin with the string ‘CCA’
    - Total payload mass by NASA(CRS)
    - Average payload mass by booster F9 v1.1
    - List the date when ground pad landing become first successful
    - List the names of successful boosters in drone ship with payload mass between 4000 and 6000
    - List the total number of success and failures in mission outcome
    - List the names of the booster version with maximum payload mass
    - List the failed outcomes in drone ship with the months, the booster version and launch site names in year 2015
    - Rank the count of landing outcomes or Success between 2010-06-04 and 2017-03-20 in descending way

[Github url jupyter notebook]()

Build an Interactive Map with Folium

- Circle
  - For NASA headquarters and multiple launch sites
- Marker
  - For the coordinate of each rocket launch
- Marker Cluster
  - For nearby markers
- Icon
  - For coloring launch outcome of each rocket
- Mouse Position
  - For coordinates of the mouse pointer in the map
- Line
  - For showing the distance to the nearest public places like Railway, Highway,etc.

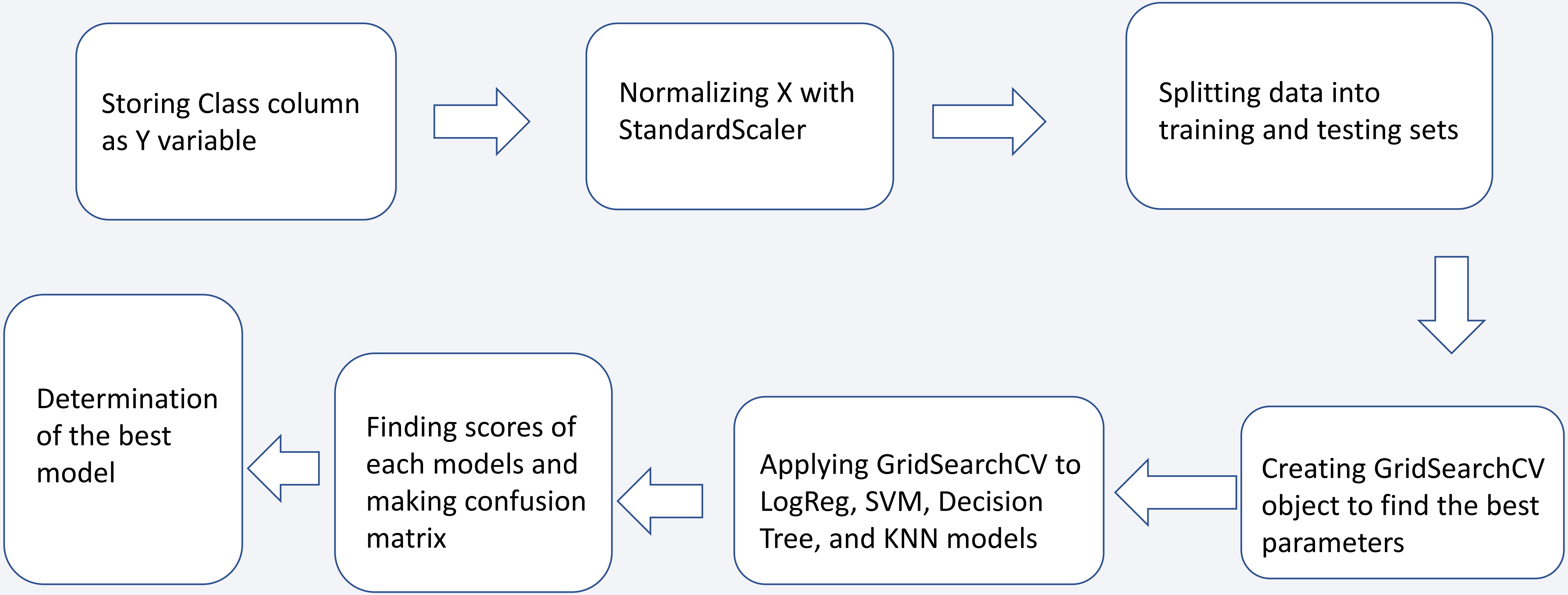
[Github url jupyter notebook]()

Build a Dashboard with Plotly Dash

- Interactions
  - Droplist : To select the launch site
  - Range Slider Bar: To select the range of Payload mass
- Charts
  - Pie Chart: To illustrate the success rate in different launch sites
  - Scatter Plot: To illustrate the payload mass, success rate, and booster version

Github url python

Predictive Analysis (Classification)



[Github url jupyter notebook]

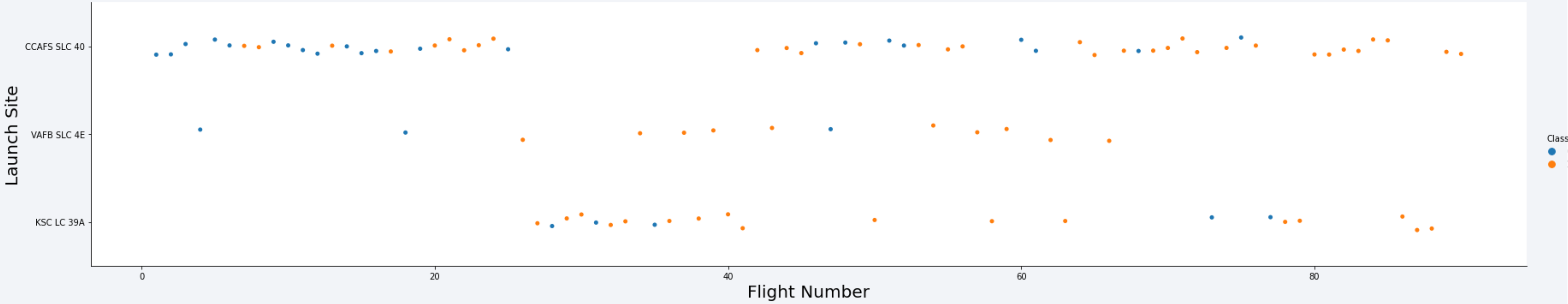
Results

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

Section 2

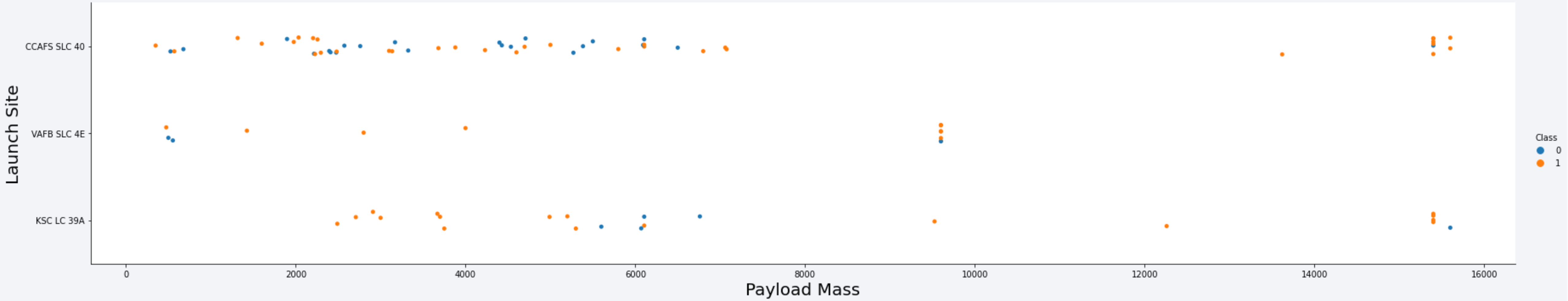
Insights drawn from EDA

Flight Number vs. Launch Site



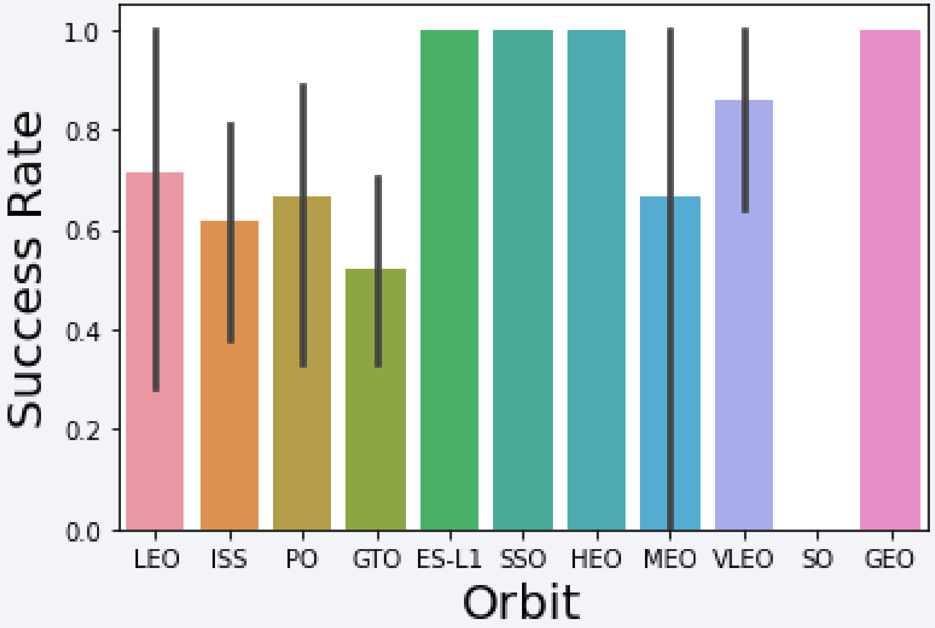
- There is an increasing success rate over time. CCAFS looks like the main launch site for launches

Payload vs. Launch Site



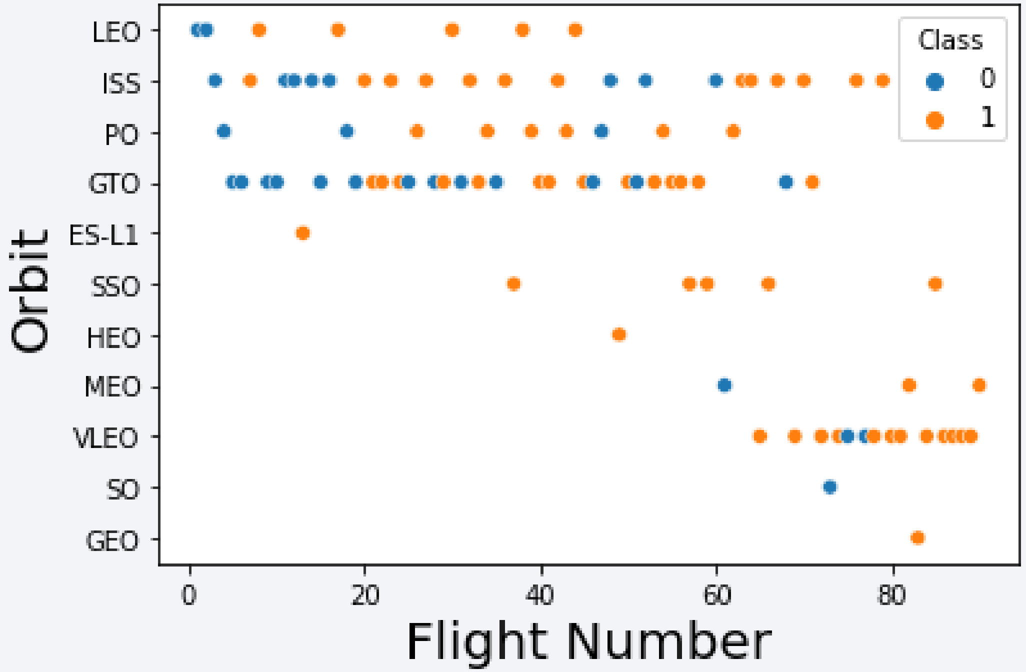
- The success rate is increasing with payload mass for each launching site. However KSC LC 39A launch site has also a very high success rate under 6000 kg

Success Rate vs. Orbit Type



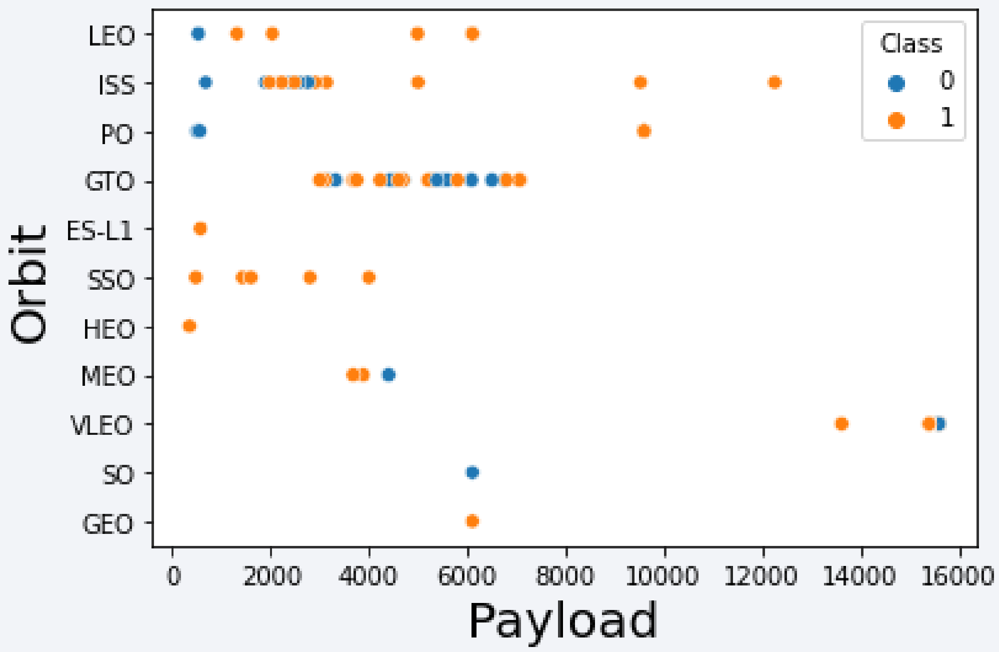
- ES-L1, SSO, HEO, and GEO orbits have %100 success rate
- GTO has success rate of %50 which is the lowest after SO which has %0 success rate with one launch

Flight Number vs. Orbit Type



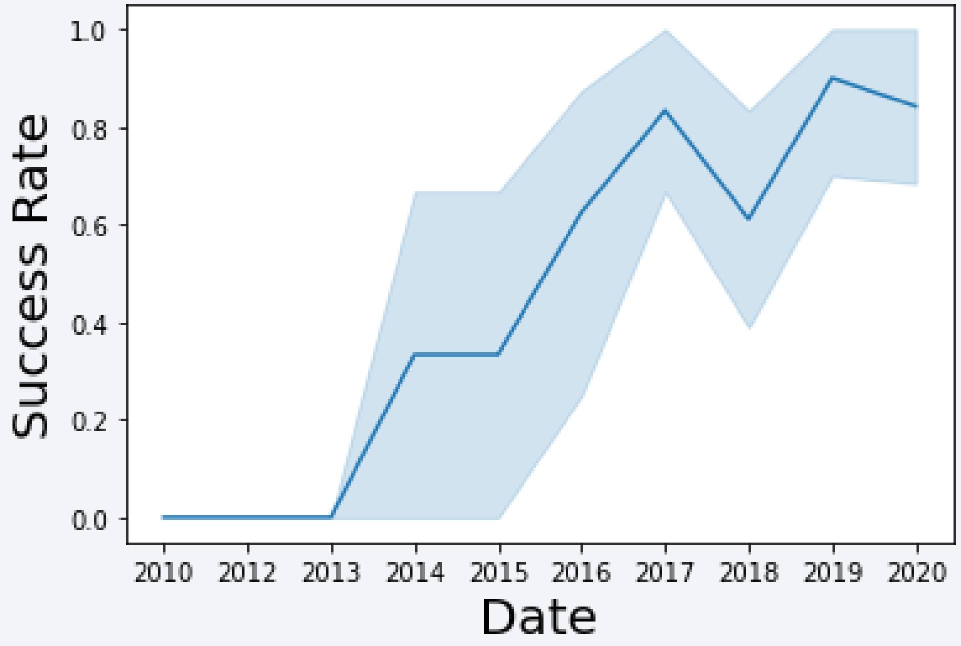
- Orbit types have different launch numbers according to preferences
- LEO had moderate success then VLEO is used for recent launches
- Launches seems to be more successful on lower orbits

Payload vs. Orbit Type



- There is some correlation between payload mass and orbit
- Lighter payload masses prefers LEO and heavier one prefers VLEO, GTO is close to moderate payload mass
- HEO, SSO, ES-L1 has light payload mass

Launch Success Yearly Trend



- The success rate is in increasing trend from 2013 and recent success rate is around %80
- There is only a slight decrease in 2018

All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Selecting launch sites from SPACEX Table(SPACEXTBL)
- Distinct function is used to prevent repetition of the same names

Launch Site Names Begin with 'CCA'

```
[ ]  %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Selecting items from table where launch site begin with CCA like and …% means place with % can be anything
- Limit 5 for giving us only 5 results

Total Payload Mass

```
[ ] %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';

 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

- SUM takes summation of payload masses and WHERE limits only launches by NASA (CRS)

Average Payload Mass by F9 v1.1

```
[ ] %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1';

 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

- AVG takes of average payload mass and WHERE limits only F9 v1.1 by equation

First Successful Ground Landing Date

```
[ ]  %sql SELECT MIN(DATE) AS FIRST_SUCCESS FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)';

 * sqlite:///my_data1.db
Done.
```

FIRST_SUCCESS
01-05-2017

- MIN function finds the earliest date WHERE landing is successful by equation of landing_outcome

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[ ] %%sql
    SELECT DISTINCT(Booster_Version) FROM SPACEXTBL
     WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
     AND "Landing _Outcome" = 'Success (drone ship)' ;

     * sqlite:///my_data1.db
    Done.
```

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

- Finding booster version DISTINCT prevent repetition and WHERE put limits on payload mass between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

```
[ ] %sql SELECT Mission_Outcome, COUNT(*) AS TOTAL FROM SPACEXTBL GROUP BY Mission_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- COUNT counting mission outcomes and GROUP BY groups for different mission outcomes

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT(Booster_Version) FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * sqlite:///my_data1.db
Done.
```

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

- There are two queries, the last one finds the maximum payload mass and first one finds booster version of this payload mass

2015 Launch Records

```
%%sql
SELECT SUBSTR(Date,4,2) as MONTH, Booster_Version, Launch_Site   FROM SPACEXTBL
WHERE SUBSTR(Date,7,4) = '2015'
AND "Landing _Outcome" = 'Failure (drone ship)';
```

```
 * sqlite:///my_data1.db
Done.
```

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

SUBSTR(Date,4,2) finds the month and we are finding month, booster version, launch site where year (SUBSTR (Date,7,4)) is 2015

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

```
%%sql
--Date format is incorrect
SELECT "LANDING _OUTCOME", COUNT(*) AS Success
FROM SPACEXTBL
WHERE "LANDING _OUTCOME" like 'Success%'
GROUP BY "LANDING _OUTCOME"
ORDER BY Success DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

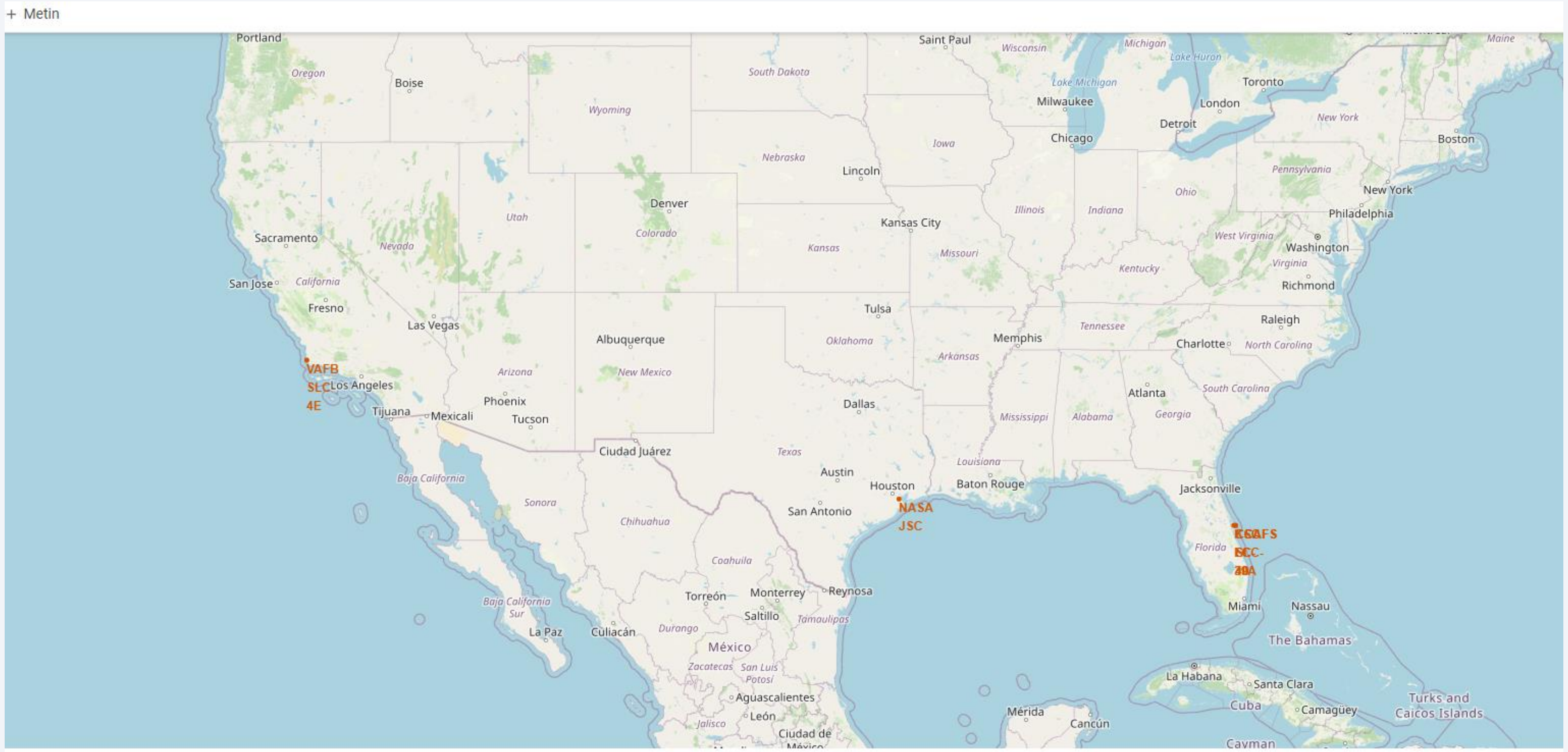
Landing _Outcome	Success
Success	38
Success (drone ship)	14
Success (ground pad)	9

- Finding landing outcomes from 2010-06-04
- There is a mistake in date format so we find until the last date of table
- Group by groups in terms of landing outcome and order by success count orders from high to low (DESC)

Section 3

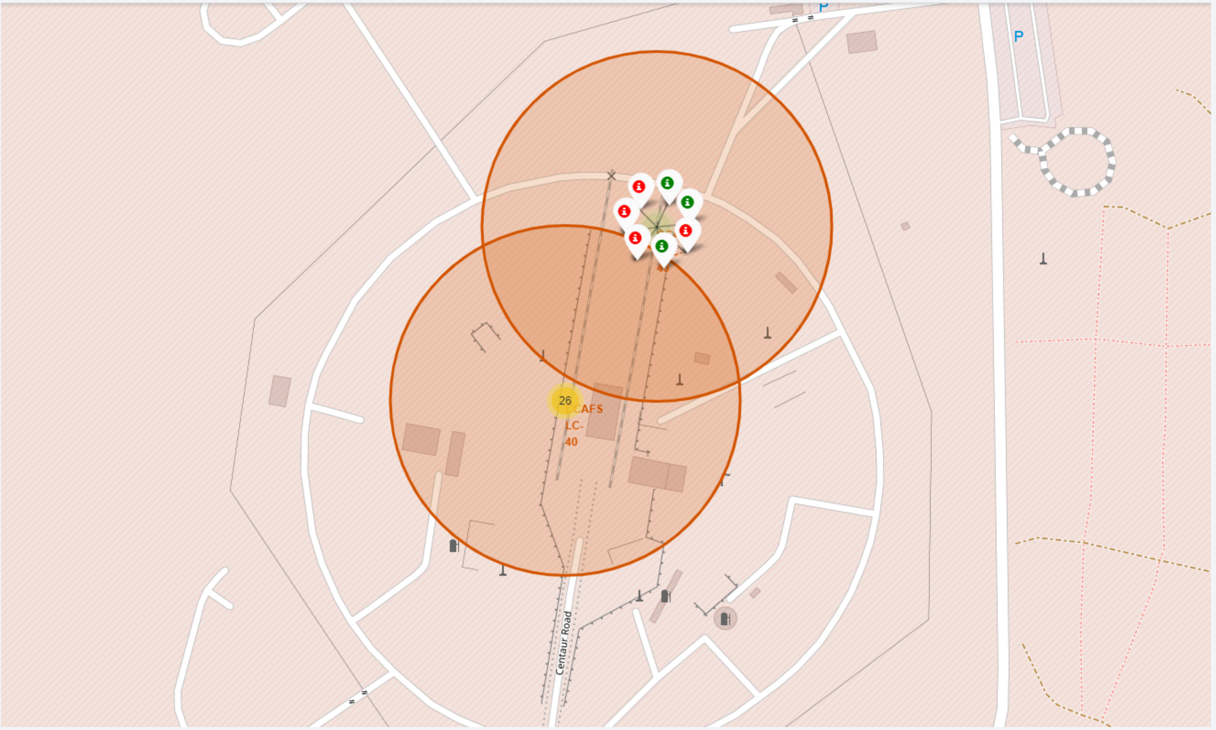
Launch Sites Proximities Analysis

Launch Sites Locations



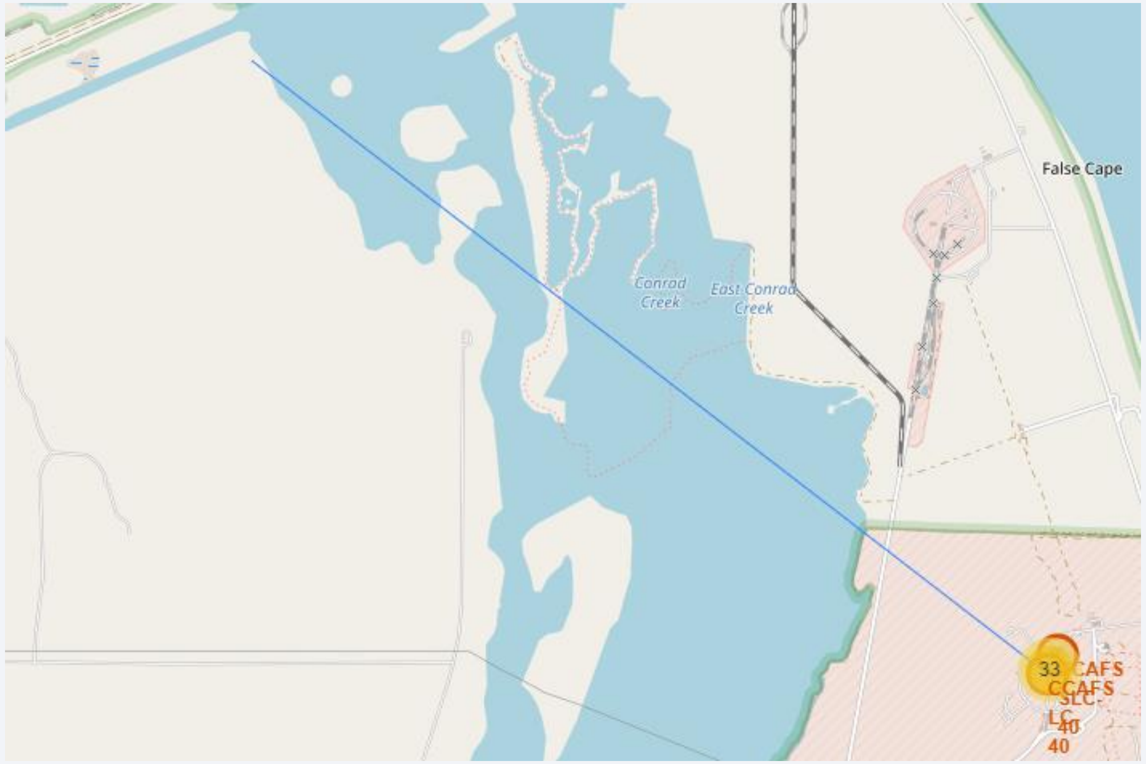
- Showing all launch site with NASA headquarters
- All launch sites are located near coastline

Records Labeled with Color on the Map



- Failed (red) and successful (green) landings can be seen by clicking on Clusters in Folium map

Distances to Public Places

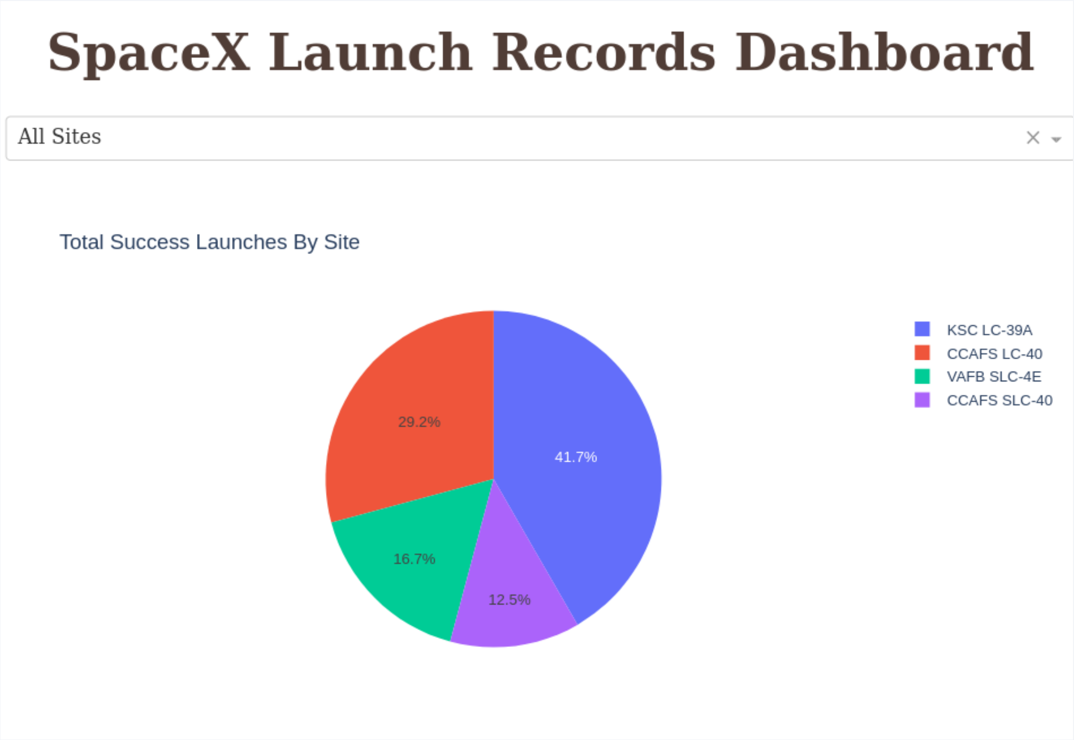


- We can spot the distances to public places on Folium map
- Here is the distance to Train station

Section 4

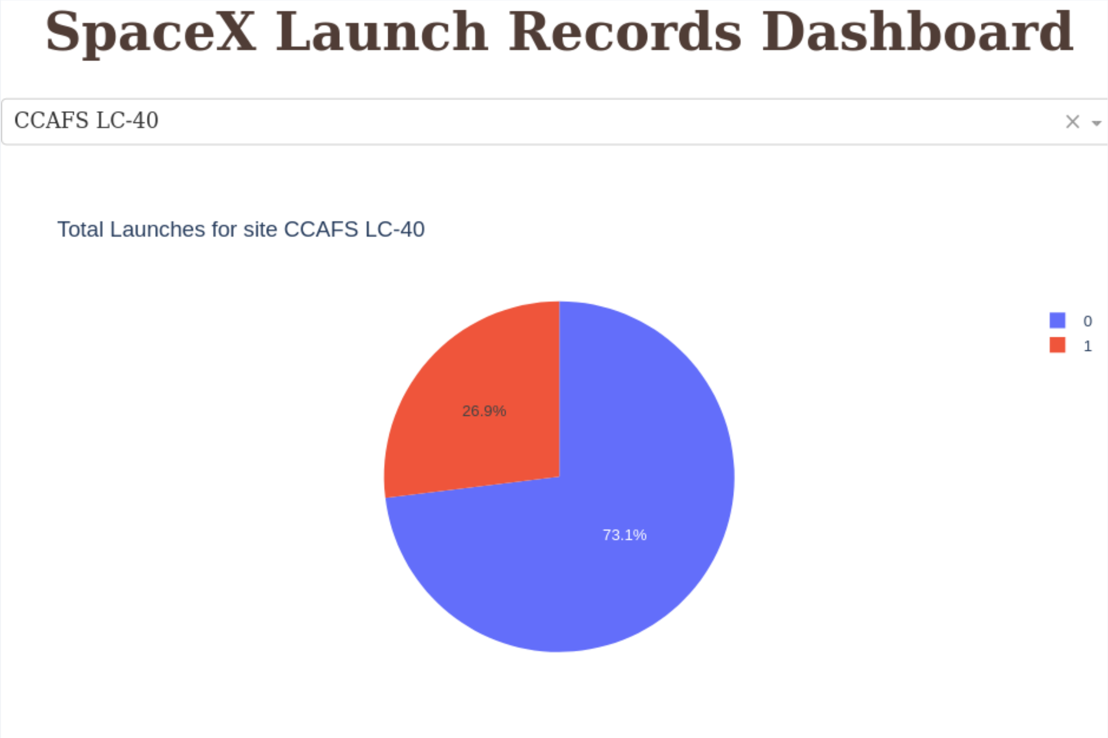
Build a Dashboard with Plotly Dash

Total Success by Launch Sites



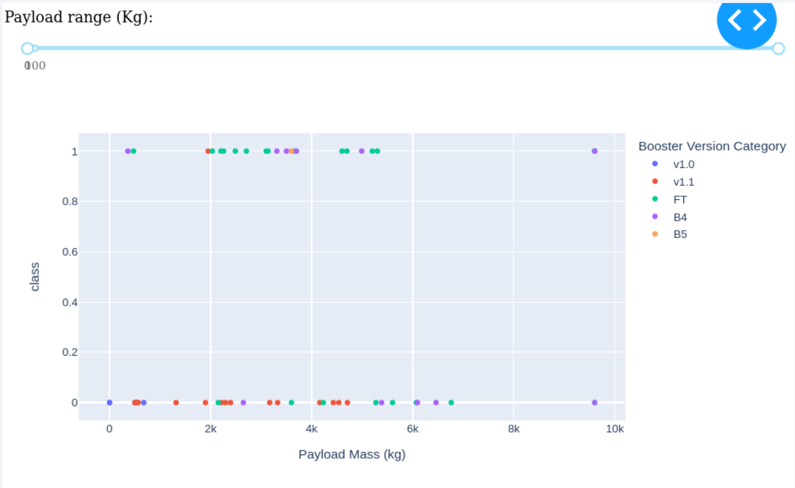
- It is obvious that KSC LC-39A is the launch site with highest success rate

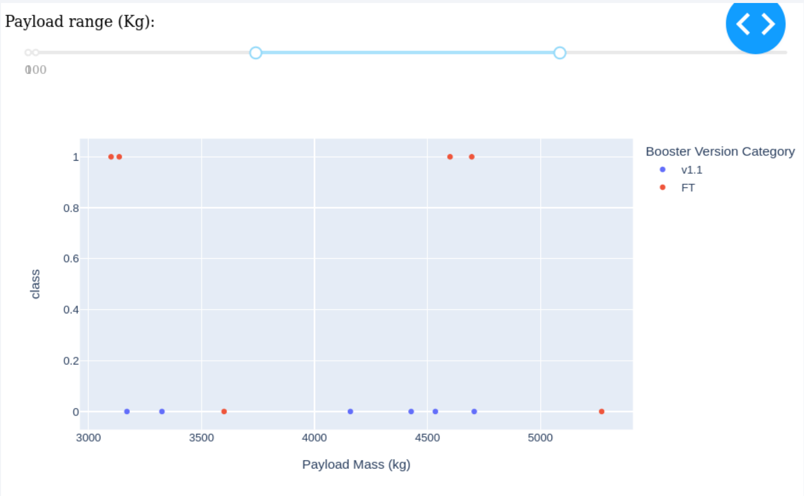
Highest Successful Launch Site



- CCAFS LC-39A is the highest successful launch site
- I put the second highest success launch site CCAFS LC-40 here because the seconds need to be appreciated also

Payload mass and success rate



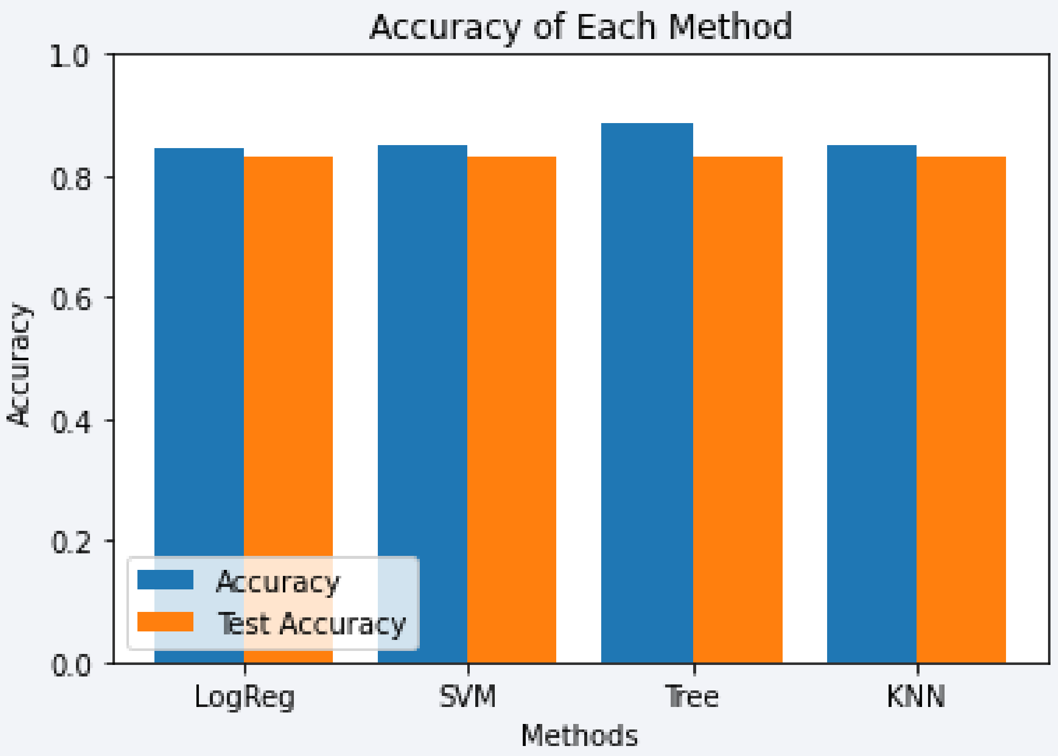


- Payload between 2000 and 5500 has highest success rate
- FT booster is the most successful among other boosters

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- All models accuracy are the same %83.33 rate
- However in test accuracy Decision tree model is a little better

Confusion Matrix

- The confusion matrix is also the same for all models like accuracy
- We have only 3 True Negative so only 3 predictions was wrong
- Models are predicting successful landing

Conclusions

- Over time, successful landing rating is increased up to %80
- SSO, HEO, GEO, and ES L-1 are the orbit types with the highest success rate
- Light payload mass has a higher success rate than heavy ones
- The launch sites are close to coastline
- There is high success rate in model with %83.33 accuracy but there need to be more data for more precise prediction

Appendix

- Github URL

Thank you!