Winning Space Race with Data Science

<Zara Zhao>
<Jun 26th 2022>

IBM Developer SKILLS NETWORK

Outline

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

Executive Summary

- Summary of methodologies
  - ➢ Data collection
  - ➢ Data wrangling
  - ➢ EDA with SQL
  - ➢ EDA with visualization
  - ➢ Build an Interactive Map with Folium
  - ➢ Build dashboard using plotly dash
  - ➢ Predictive analysis (Classification)
- Summary of all results
  - ➢ EDW with SQL results
  - ➢ Visualization results
  - ➢ Predictive analysis results

Introduction

- Project background and context

  The commercial space age is here, Companies are making space travel affordable for everyone. As a new company in the market, Space Y is trying to compete with Space X. Executive team need to determine the price of each launch by analyzing data for Space X and create dashboard to show executive team. We will train machine learning model and use public information to predict if Space X will reuse the first stage.

- Problems you want to find answers

  The project will use ML algorithm to predict whether the first stage of the Space X Falcon 9 rocket launch will successfully land

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - ➢ Collecting data with SpaceX REST API or
    - ➢ Using Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records
- Perform data wrangling
    - ➢ Wrangling data using an API
    - ➢ Sampling data
    - ➢ Dealing with Nulls
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - ➢ The project will train logistic regression, KNN, SVM and decision tree models to make prediction and evaluate models performance and choose the best fitting model.

Data Collection

- Describe how data sets were collected.

  - ➢ we will be working with SpaceX launch data that is gathered from **SpaceX REST API**. This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

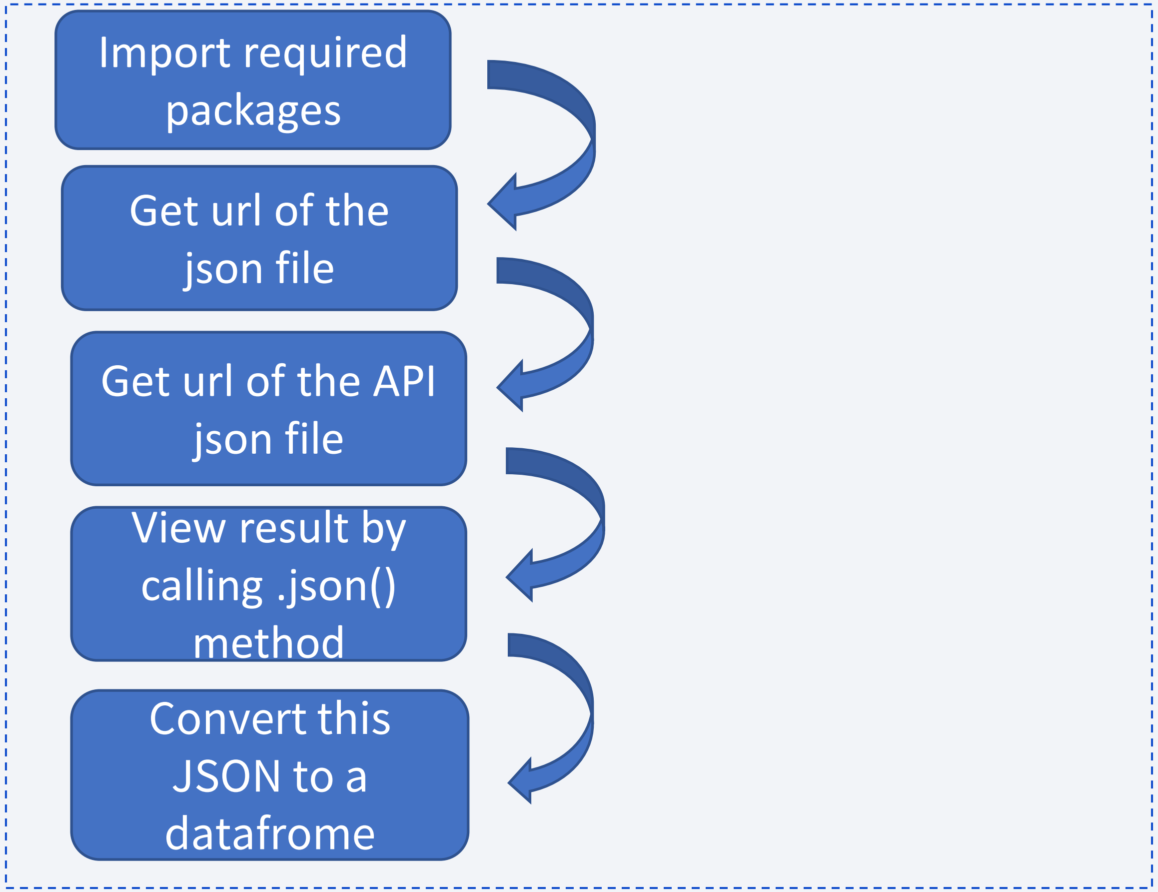
  - ➢ We will perform a get request using the requests library to obtain the launch data, which we will use to get the data from the API. This result can be viewed by calling the **.json()** method. Then we can use the **json_normalize** function to convert this JSON to a datafrome. This function will allow us to “normalize” the structured json data into a flat table.

  - ➢ Another popular data source for obtaining Falcon 9 Launch data is web scraping related Wiki pages. In this lesson, you will be using the Python **BeautifulSoup** package to web scrape some HTML tables that contain valuable Falcon 9 launch records. Then you need to parse the data from those tables and convert them into a Pandas data frame for further visualization and analysis. We want to transform this raw data into a clean dataset which provides meaningful data on the situation we are trying to address: Wrangling Data using an API, Sampling Data, and Dealing with Nulls.

Data Collection – SpaceX API

GitHub URL of the completed SpaceX API calls notebook:

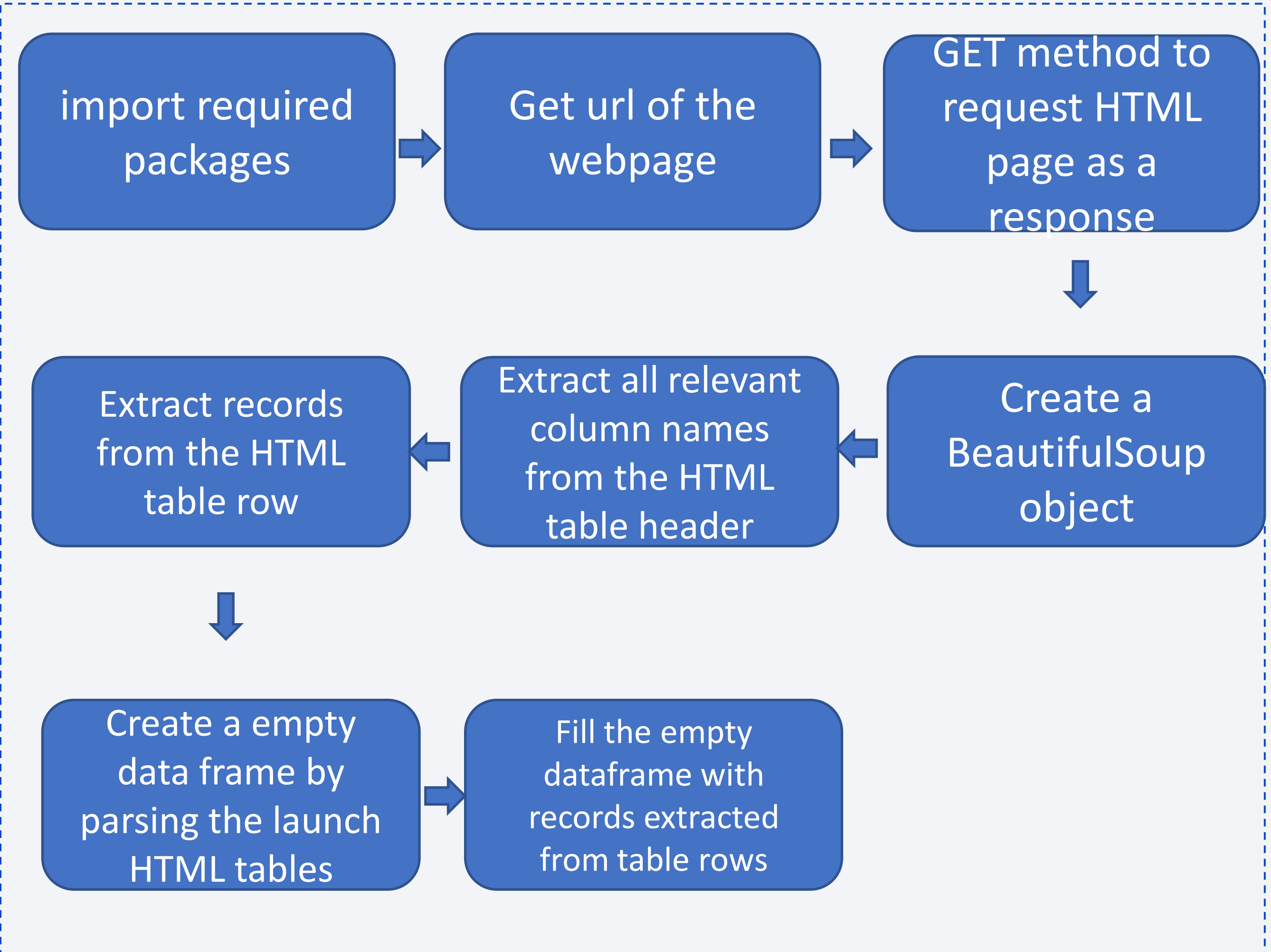
https://github.com/zara01991/DataScience/blob/master/Capstone%20-%20SpaceX%20-%20Data%20Collection%20API.ipynb



Data Collection - Scraping

GitHub URL of the completed SpaceX API calls notebook:

https://github.com/zara01991/DataScience/blob/master/spacex-data-collection-webscraping-soup.ipynb



Data Wrangling

- Describe how data were processed

➢ Perform Exploratory Data Analysis (EDA) to find some patterns such as Identify and calculate the percentage of the missing values in each attribute, Identify which columns are numerical and categorical, etc.

➢ Determine Outcomes would be the label for training supervised models,

➢ Convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

- GitHub URL:

https://github.com/zara01991/DataScience/blob/master/spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

➢ Scatter plot the FlighNumber vs. PayloadMass and overlay the outcome of the launch to see how the FlightNumber and Payload would affect the launch outcome

➢ Scatter plot FlightNumber vs. LauchSite as we see different launch sites have different success rates

➢ Scatter plot Payload vs. LauchSiates to observe if there is relationship between launch sites and their payload mass

➢ Bar chat to visually check if there are relationship between success rate and orbit type

➢ Scatter plot FlightNumber vs. Orbit type to see if there is relationship between FlightNumber and Orbit type

➢ Scatter plot Payload vs. Orbit type to reveal the relationship between Payload and Orbit type

➢ Plot line chart Year vs average success rate to get the average launch success trend

- GitHub URL:

https://github.com/zara01991/DataScience/blob/master/spacex-eda-dataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

➢ Install sqlalchemy and load SQL extension

➢ Perform EDA to analyzed the data to see if the attributes can be used as features to determine if the Falcon 9's second stage will land and correlated with a successful land, such as:

  - ➢ Launch sites
  - ➢ Different success rate for different launch sites
  - ➢ Total number of successful and failure mission outcomes
  - ➢ Success rate over years
  - ➢ Average, minimum, maximum payload mass
  - ➢ Booster_versions which have carried the maximum payload mass
  - ➢ ……

- GitHub:

https://github.com/zara01991/DataScience/blob/master/spacex-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

➢ Map showing all launch sites' location markers on a global map

➢ Map showing the color-labeled launch outcomes on the map

➢ Map showing a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

- Explain why you added those objects

➢ The maps helps provide visualization of the allocation of the sites and the landing outcome;

➢ The maps helps implicitly explain the reason why they are located this way;

➢ Gives the vision if there is relationship between site location and outcome;

- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

https://github.com/zara01991/DataScience/blob/master/spacex-launch_site_location%20Folium.ipynb

Build a Dashboard with Plotly Dash

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

➢ Pie chart for all launch sites landing counts and success ratio

➢ Scatter plot for all launch sties landing outcome vs payload mass (kg)

- Explain why you added those plots and interactions

➢ Pie chart provides visualization how each launch site performs in terms of land success, also tells which site has highest success rate;

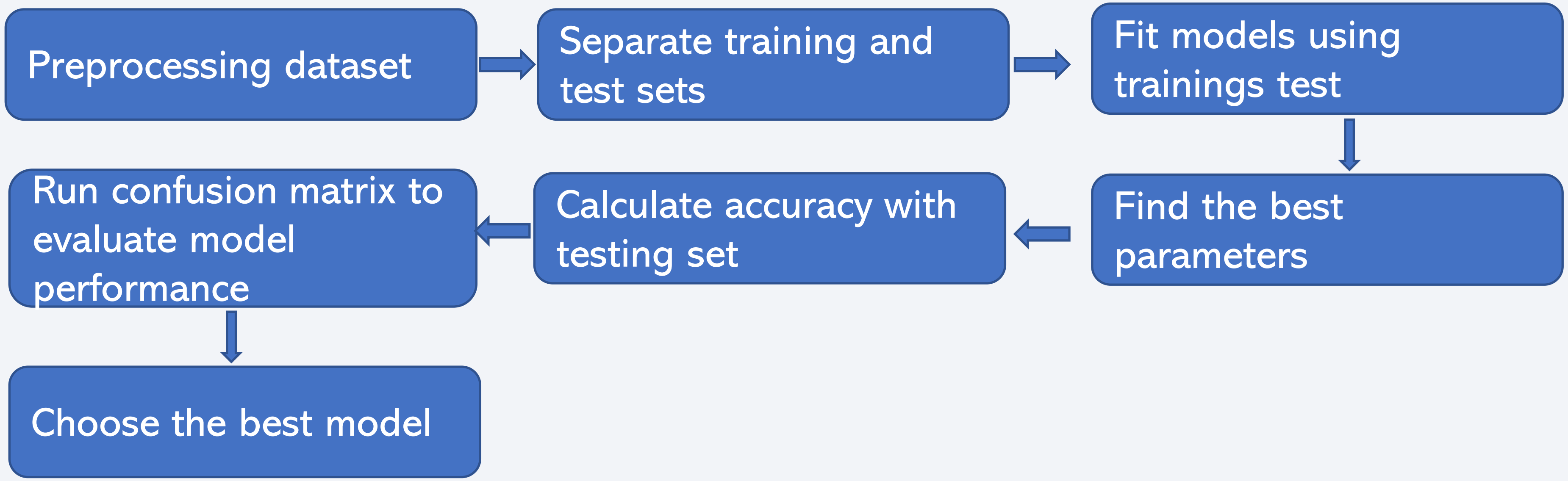
➢ Scatter plot helps visualize the relationship between payload mass and outcome;

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

https://github.com/zara01991/DataScience/blob/master/spacex_dash_app2.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model



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

https://github.com/zara01991/DataScience/blob/master/Spacex_Machine%20Learning%20Prediction_KNN_LR_SVM_DT.ipynb

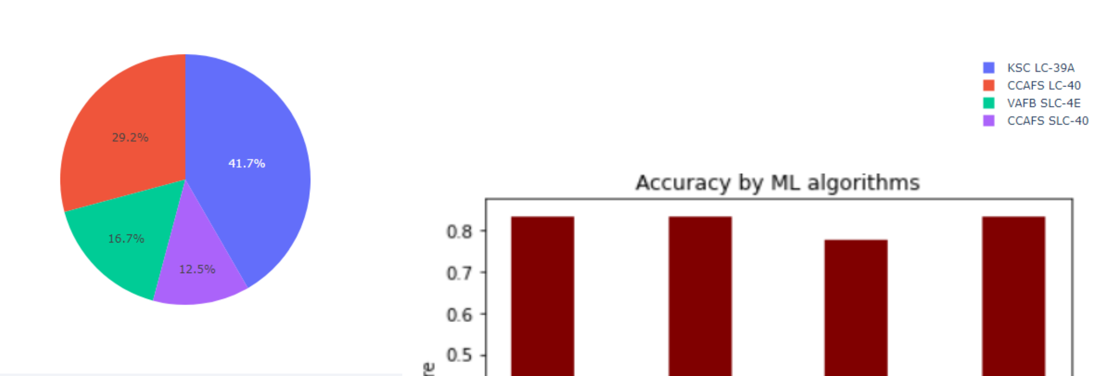
Results

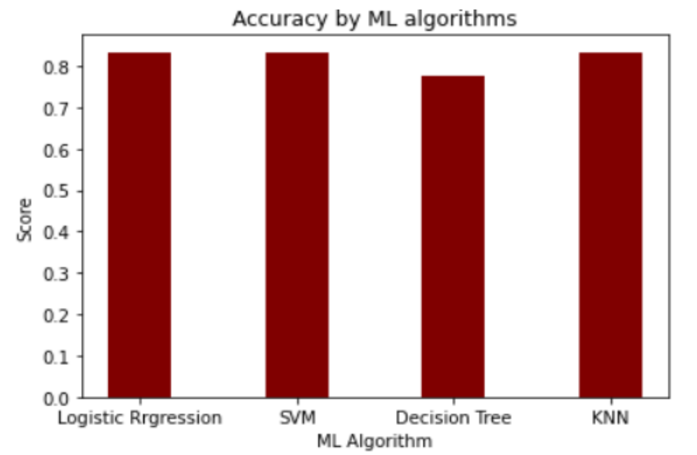
- Exploratory data analysis results

➢ KSC LC 39A has the most successful landing

➢ ES-L1, GEO, HEO, SSO have the highest success rate

➢ Overall, the success rate keeps increasing since 2013





- Interactive analytics demo in screenshots

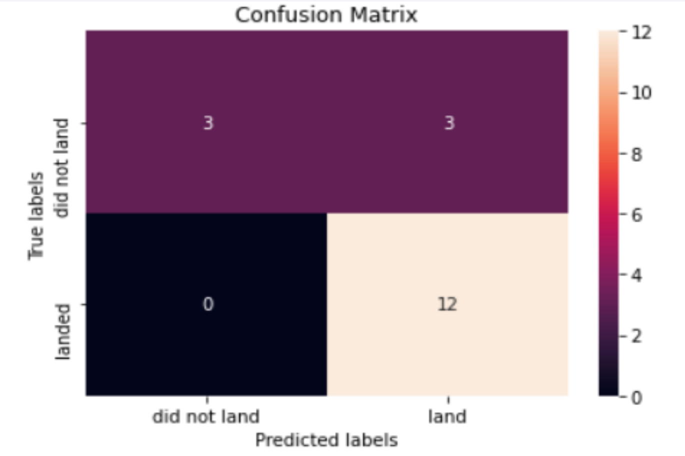
➢ KSC LC-39A has the most success launches;

➢ CCAFS SLC-40 has the least success launches;

➢ As payload is increasing, the success rate seems getting lower;

- Predictive analysis results

➢ KNN, LR and SVM methods have the highest classification of **83.3%**



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site

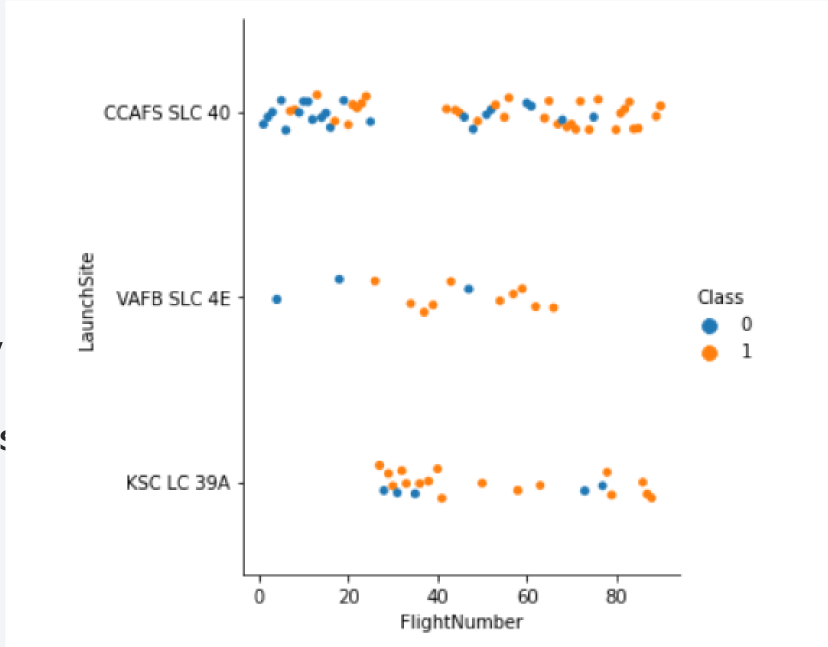
- Observations:

➢ 1) VAFB SLC 4E and KSC LC 39A are more likely to land successfully

➢ 2) For CCAFS SLC 40, as the flight number increases, the first flight is more likely to land successfully

➢ 3) CCAFS SLC 40 has the most launches while VAFB SLC 4E has the least

➢ 4) KSC LC 39A doesn't have launches for flight number between 0 – 20, VAFB SLC 4E doesn’t have launches for flight number over 60



Payload vs. Launch Site

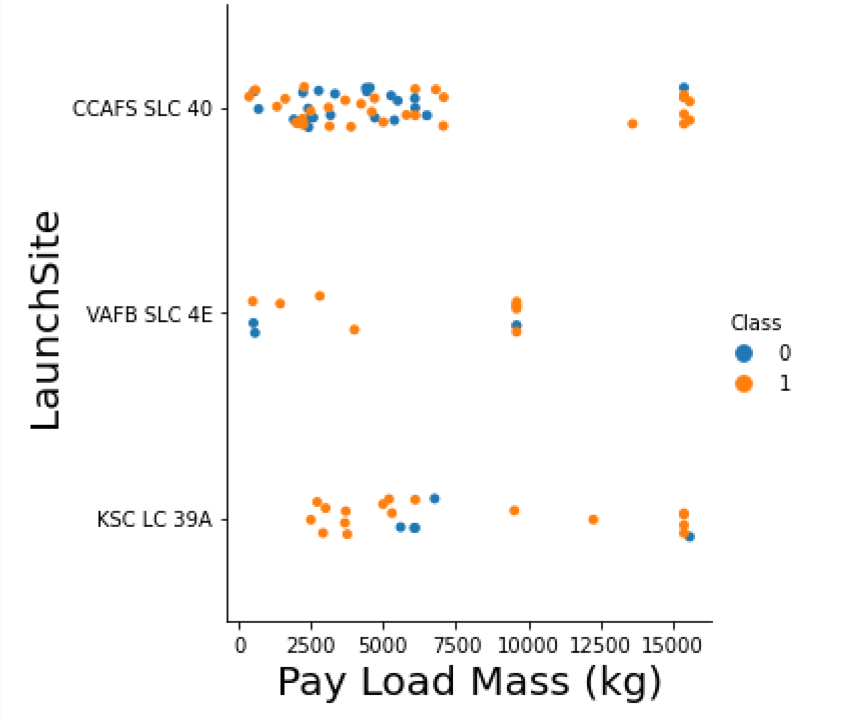
- Show a scatter plot of Payload vs. Launch Site
- Observations:

➢ 1) VAFB SLC launch site has the least launches and has no launch for heavy payload mass greater than 10000

➢ 2) It seem like after reaching a threshold, the more massive pay load is, the more likely the launch will be successful

➢ 3) KSC LC 39A has the most successful landing

➢ 4) there seems to be a relationship between outcome and payload mass



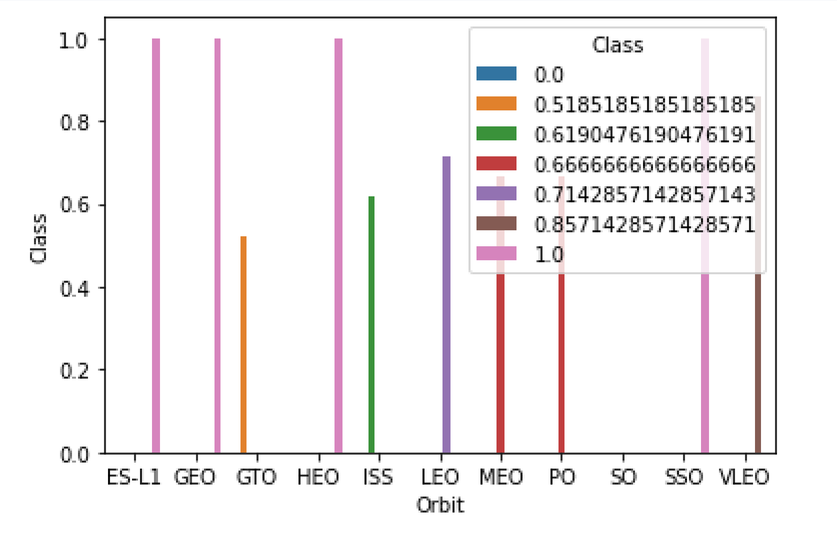
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- Observations:

➢ 1) ES-L1, GEO, HEO, SSO have the highest success rate

➢ 2) GTO has the least success rate

➢ 3) ISS, LEO, MEO, PO have the average success rate

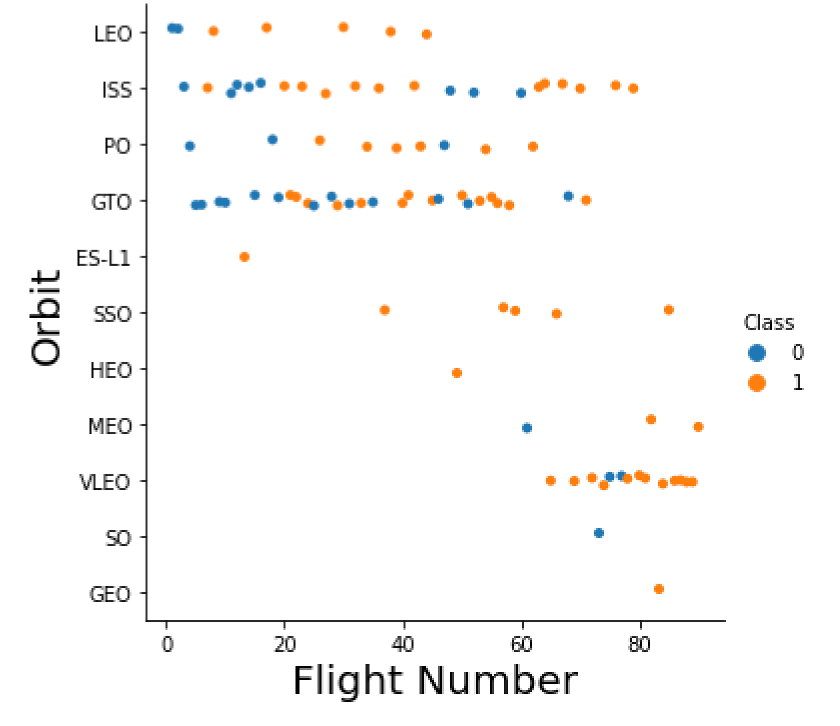


Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type
- Observations:

➢ 1) For LEO the success appears related to the number of flights while for GTO there seems to be no relationship between flight number and success.

➢ 2) SSO has relatively low launches but higher success rate



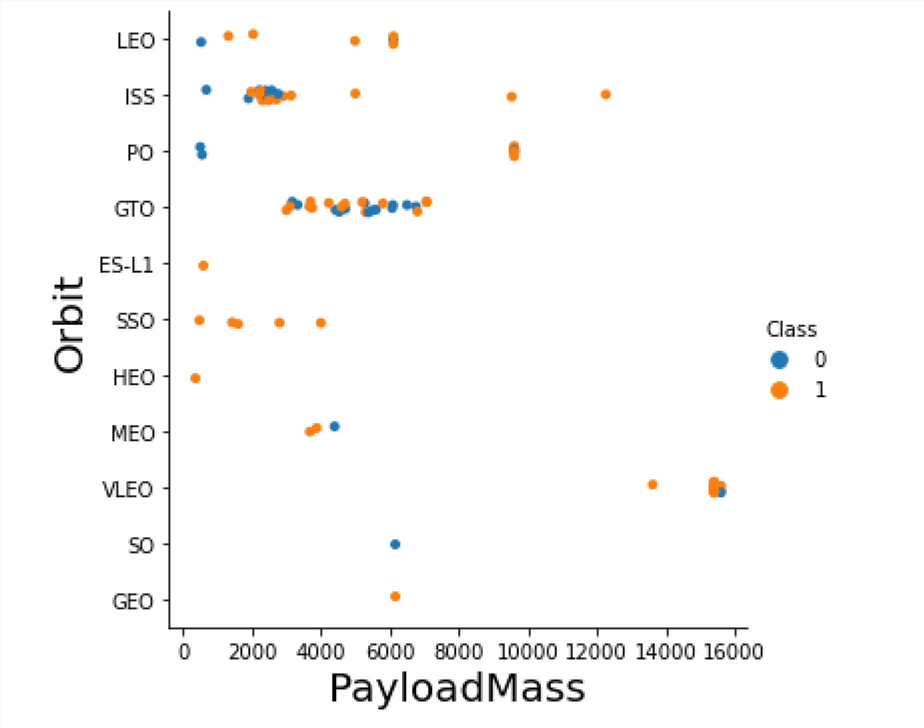
Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type
- Observations:

➢ 1) With heavy payloads the successful landing rate are more for LEO, PO, ISS

➢ 2) There seems no relationship between payload mass and success for GTO

➢ 3) SSO has high success rate for small payload mass

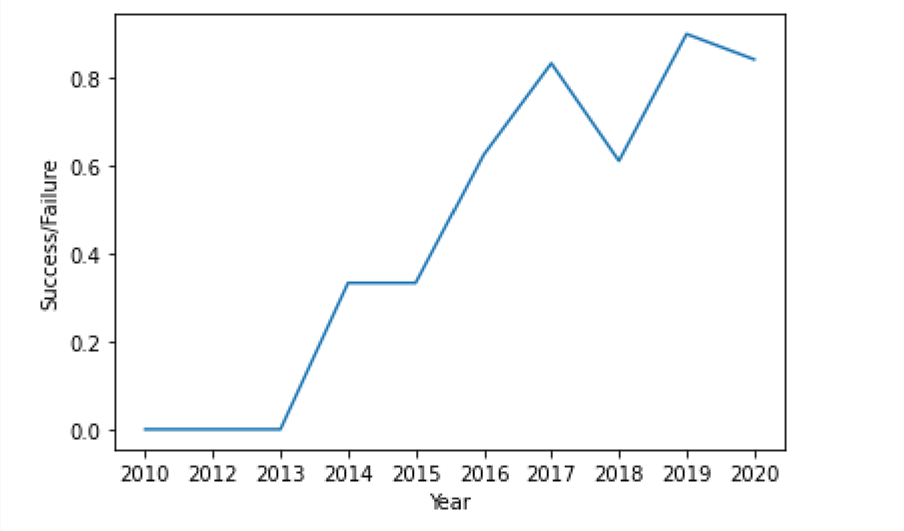


Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Observations:

➢ 1) Overall, the success rate keeps increasing since 2013

➢ 2) There was a big jump in 2018 and increased to highest rate in 2019



All Launch Site Names

- Find the names of the unique launch sites
- Observation:

➢ 1) There are 4 launch sites

```
%%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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites 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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%%sql
select  sum(PAYLOAD_MASS__KG_) from SPACEXTBL
where Customer  = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

- Observation:

➢ 1) NASA boosters carried massive payload in total

Average Payload Mass by F9 v1.1

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

```
%%sql
select  avg(PAYLOAD_MASS__KG_) from SPACEXTBL
where Booster_Version  like 'F9 v1.1%'
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- Observations:

➢ 1) F9 V1.1 carried small payload on average

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%%sql
select  min(Date) from SPACEXTBL
where [Landing _Outcome] ='Success (ground pad)'
```

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

min(Date)
01-05-2017

- Observations:

➢ 1) The first ground pad successful landing happened in 2017-1-5

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%%sql
select distinct payload from SPACEXTBL
where  [Landing _Outcome] ='Success (drone ship)'
and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

- Observations:
- 1) There are 4 boosters have successfully landed on drone ship and had meadieum payload mass

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%%sql
select case when [Landing _Outcome] like 'Success%' then 'Success' else 'Failure' end as Outcome, [Landing _Outcome] ,count(*) from SPACEXTBL
where [Landing _Outcome] like 'Success%' or [Landing _Outcome] like 'Failure%'
group by case when [Landing _Outcome] like 'Success%' then 'Success' else 'Failure' end, [Landing _Outcome]
```

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

Outcome	Landing _Outcome	count(*)
Failure	Failure	3
Failure	Failure (drone ship)	5
Failure	Failure (parachute)	2
Success	Success	38
Success	Success (drone ship)	14
Success	Success (ground pad)	9

- Observations:

➢ 1) Success rate is much higher than failure rate

Boosters Carried Maximum Payload

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

➢ 1) F9 B5 booster version has the greatest payload mass

```
%%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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Observations:

➢ 1) Two booster versions had failure drone ship landing in 2015

➢ 2) These failure landing happened in CCAFS LC-40

```
%%sql

select substr(Date, 4, 2) as month, Launch_Site, Booster_Version,[Landing _Outcome]
from SPACEXTBL
where [Landing _Outcome] = 'Failure (drone ship)'
and substr(Date,7,4)='2015'
```

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

month	Launch_Site	Booster_Version	Landing _Outcome
01	CCAFS LC-40	F9 v1.1 B1012	Failure (drone ship)
04	CCAFS LC-40	F9 v1.1 B1015	Failure (drone ship)

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

- 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

- Observations:

➢ 1) Successful landing rate is higher than failure during this time range, around 60%

➢ 2) Successful landing rate in all means are higher than failure

```
%%sql
select [Landing _Outcome], count(*) as sucesslanding from SPACEXTBL
where Date between '04-06-2010' and '20-03-2017'  --and [Landing _Outcome] = 'Success'
group by [Landing _Outcome]
order by count(*) desc
```

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

Landing _Outcome	sucesslanding
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

Mark Each Launch Site

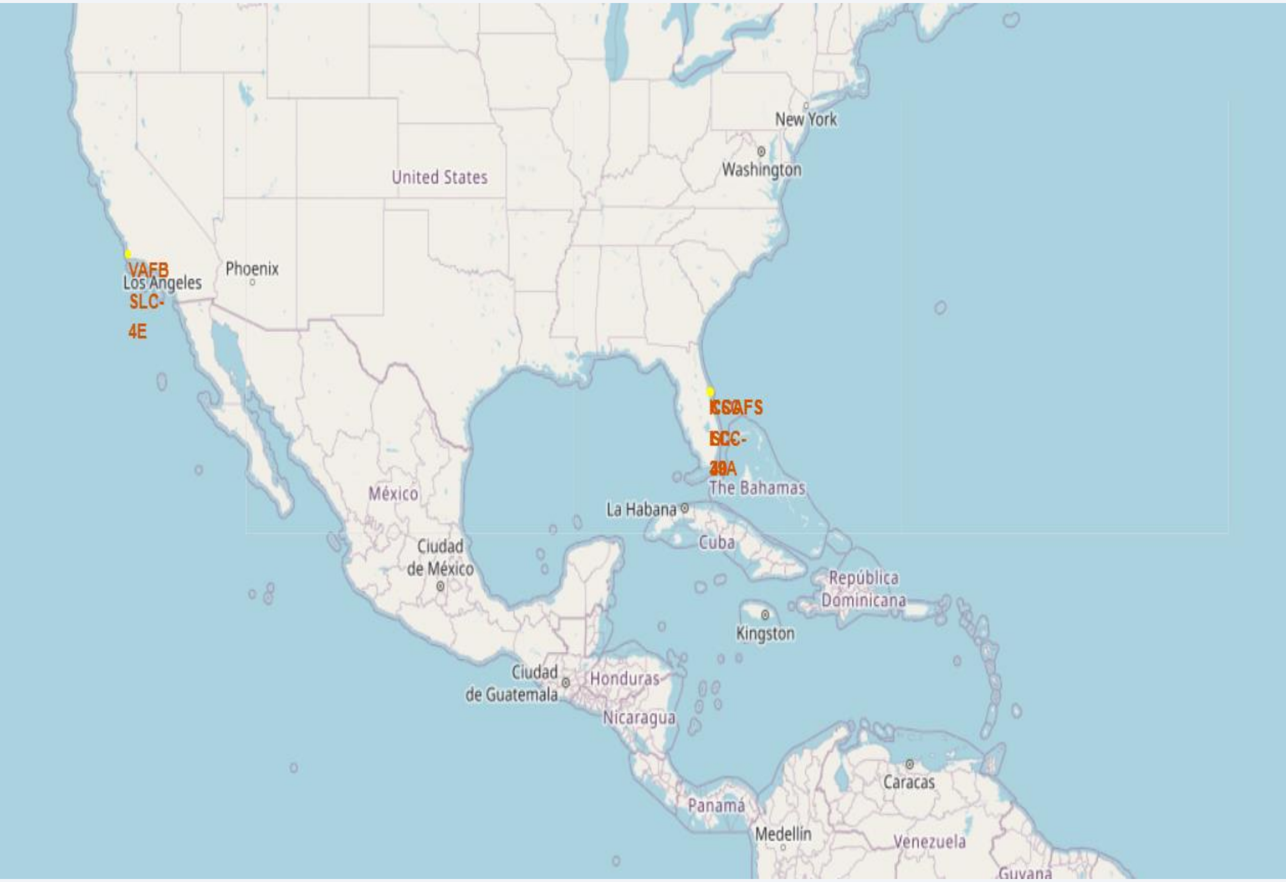
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot

➢ There're 4 launch sites;

➢ One at west coast and other three are all together in east coast and close by;

➢ All are in very close proximity to the coast;

➢ All sites are close to Equator line;



Mark the success/failed launches for each site

- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map

- Explain the important elements and findings on the screenshot

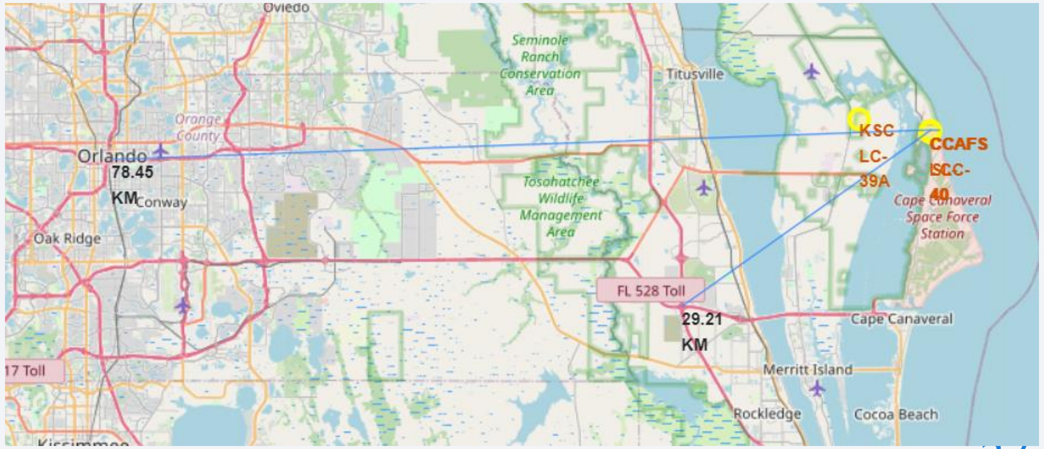
Distances between a launch site to its proximities

- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

- Explain the important elements and findings on the screenshot

➢ Launch sites are more likely to close to coastline;

➢ Launch sites are not likely to close to highway or cities with higher population;



Section 4

Build a Dashboard with Plotly Dash

<Total Success Launches>

- Show the screenshot of launch success count for all sites, in a piechart

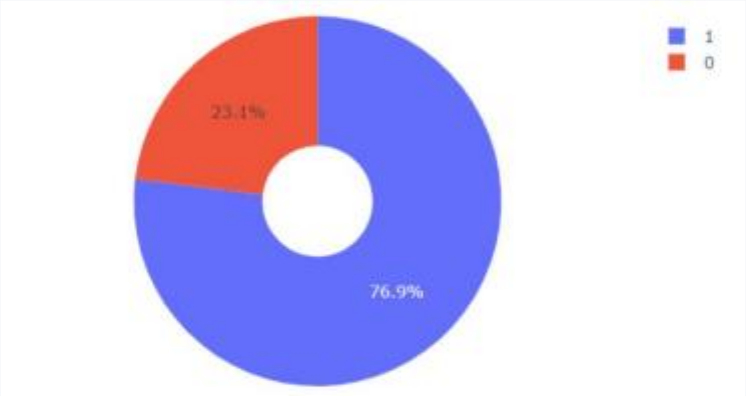
- Explain the important elements and findings on the screenshot

➢ KSC LC-39A has the most success launches 41.7% across all sites;

➢ CCAFS SLC-40 has the least success launches 12.5% across all sites;

< Launch site with highest launch success ratio >

- Show the screenshot of the piechart for the launch site with highest launch success ratio



- Explain the important elements and findings on the screenshot

➢ CCAFS SLC-40 has the highest launch success ratio at 76.9%;

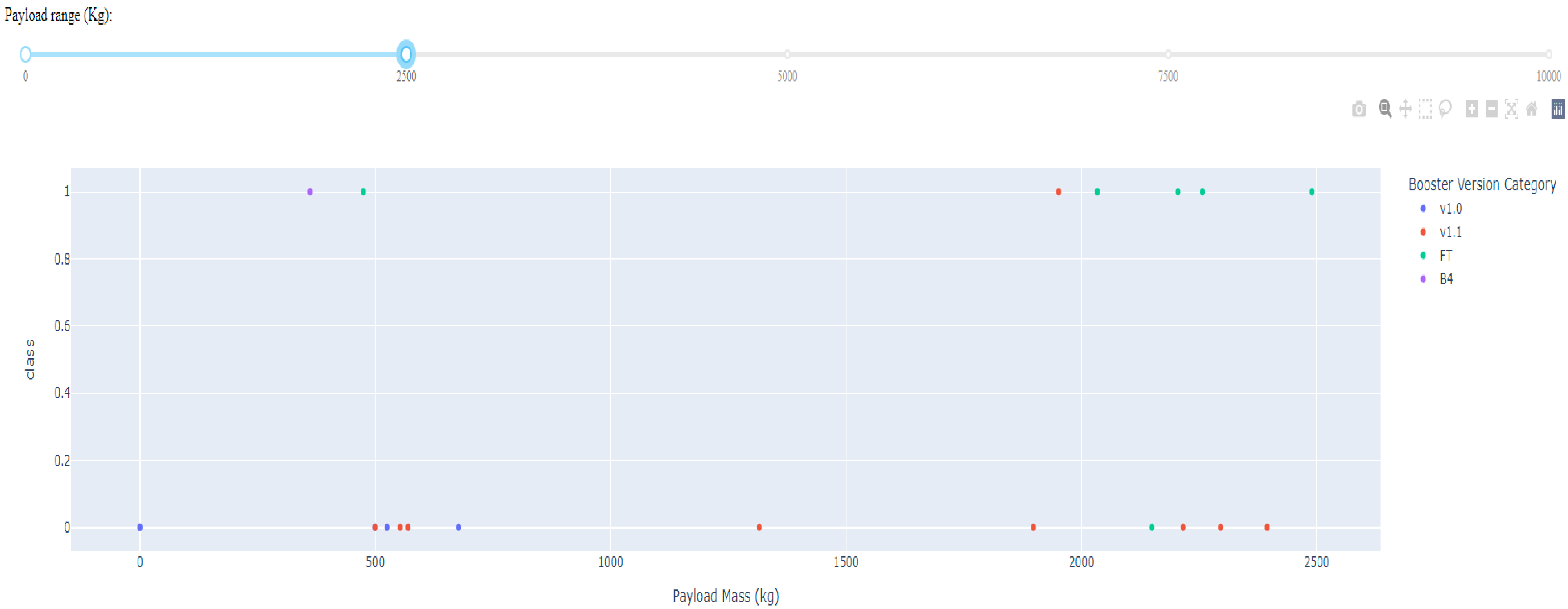
< Payload vs. Launch Outcome >

- 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, such as which payload range or booster version have the largest success rate, etc.

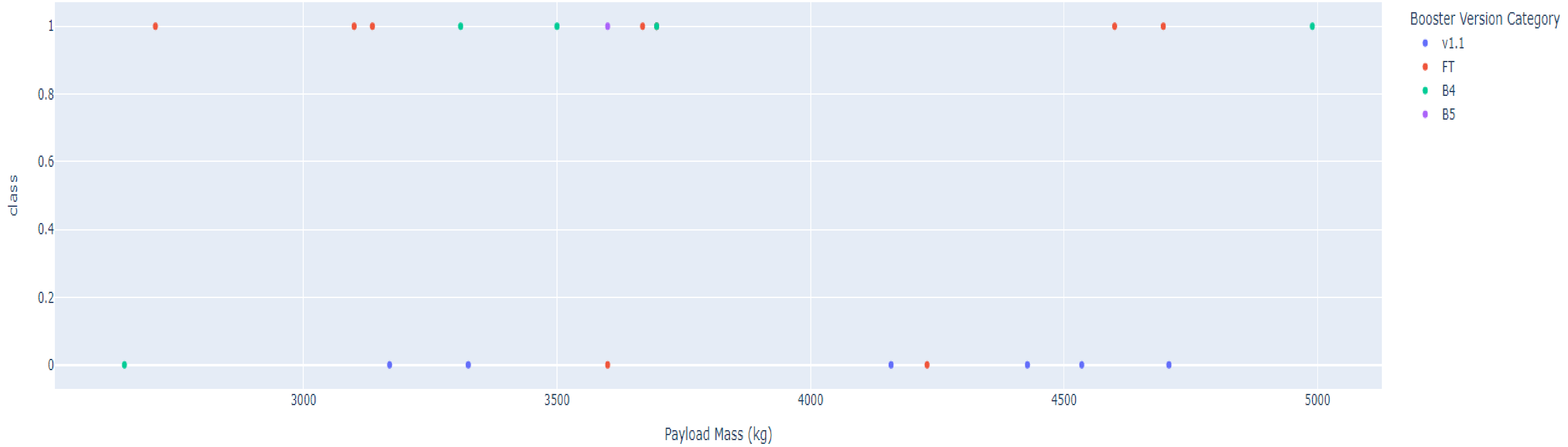
➢ Payload ranging from 2500 to 5000 seems to have the highest success rate;

➢ As payload is increasing, the success rate seems getting lower;

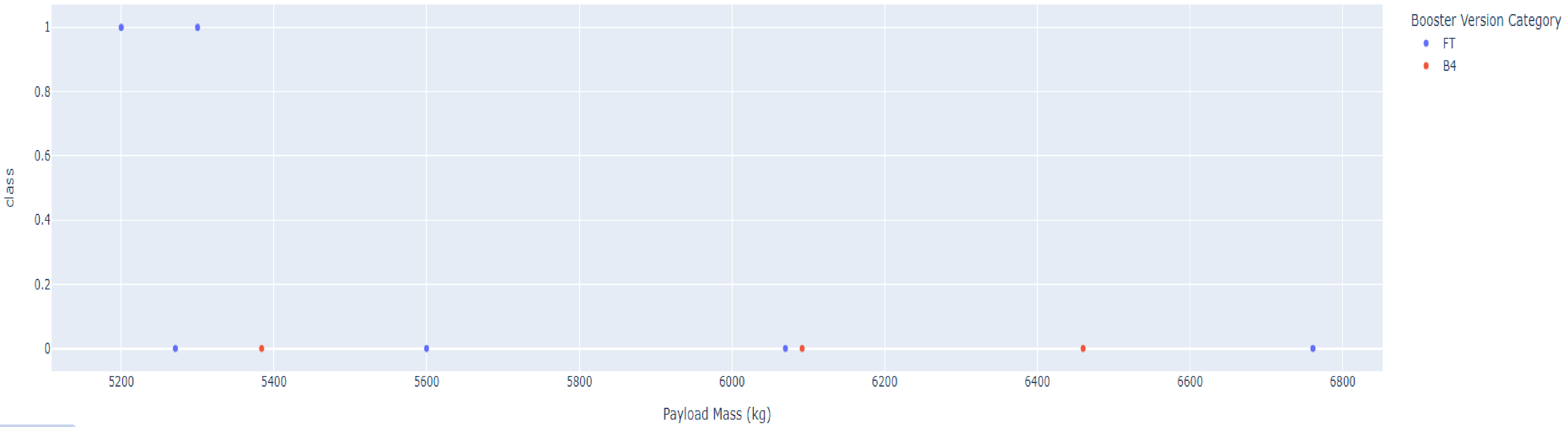

Payload range (Kg):
0
2500
5000
7500
10000
class
1
0.8
0.6
0.4
0.2
0
0
500
1000
1500
2000
2500
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4

Payload range (Kg):

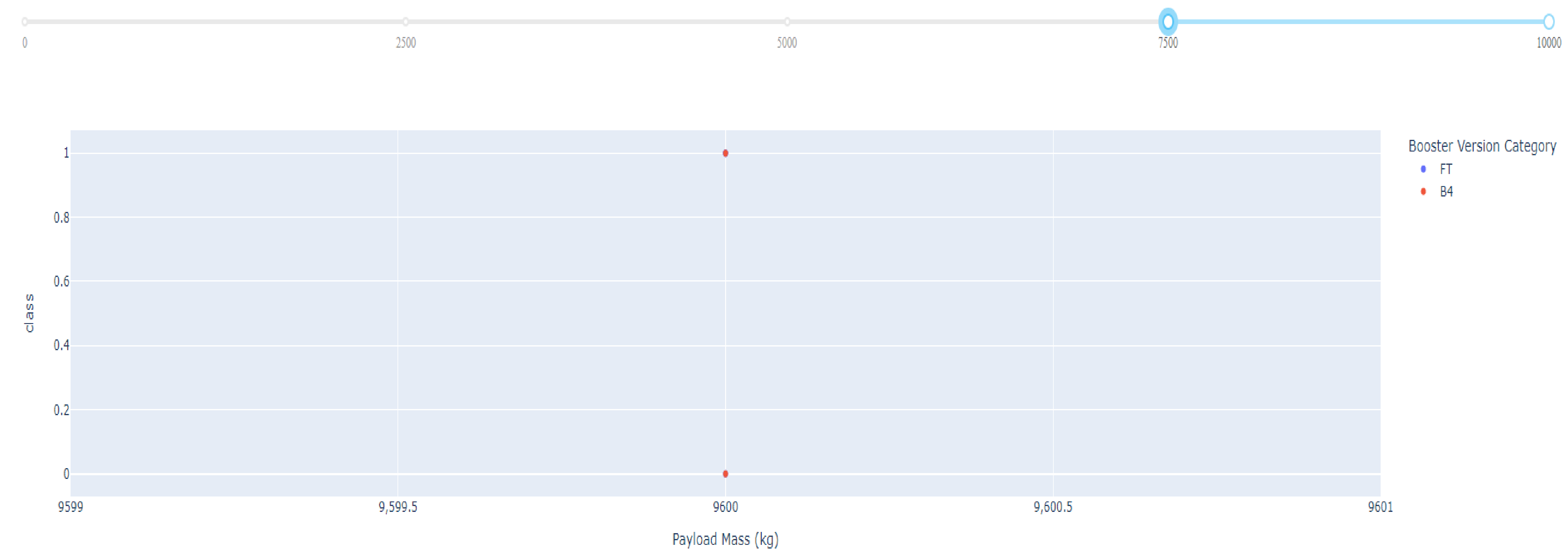
0 2500 5000 7500 10000



Payload range (Kg):



Payload range (Kg):



Section 5

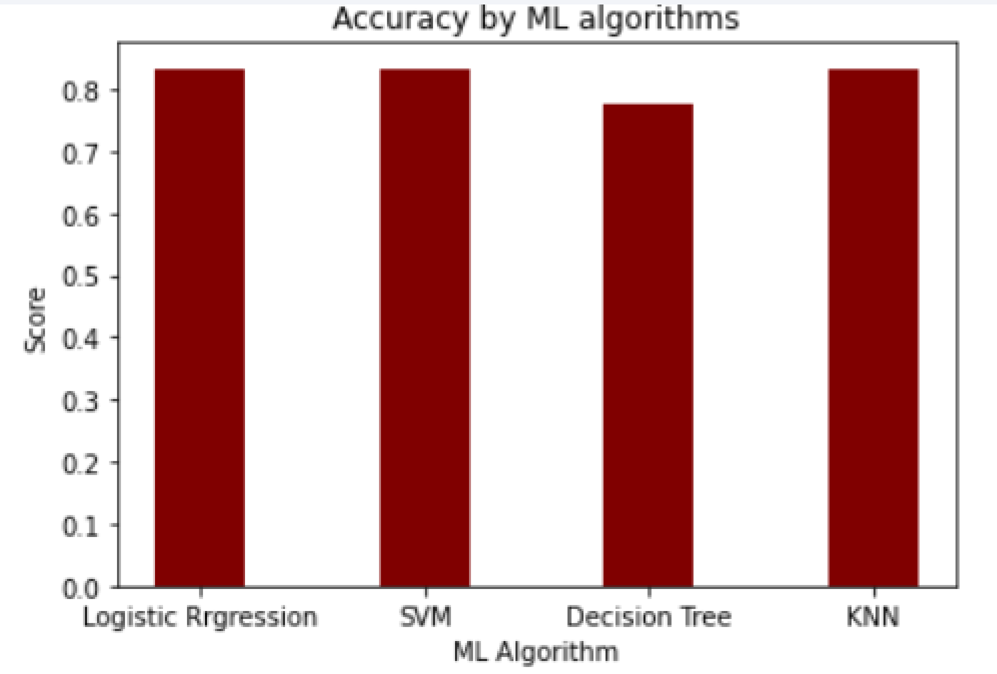
Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

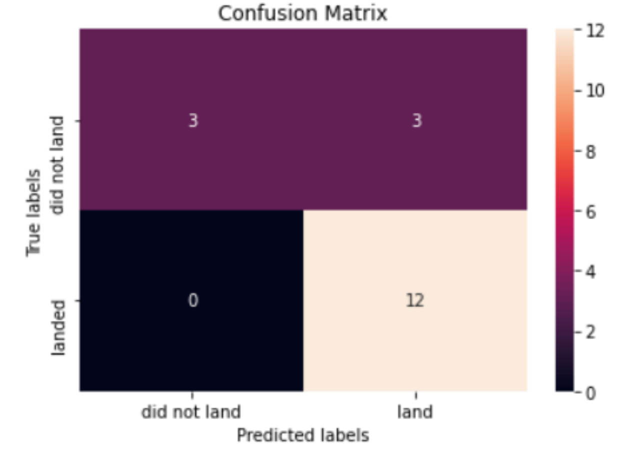
- Find which model has the highest classification accuracy

➢ LR, SVM and KNN methods have the similar classification performance of 83.3%

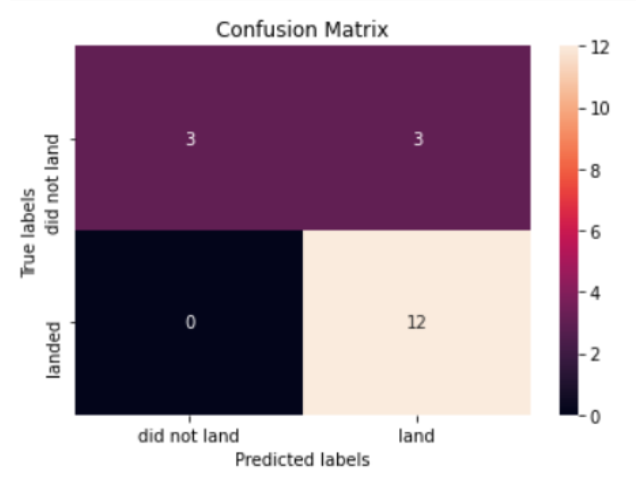



```
Logistic Rrgression: 0.8333333333333334
SVM: 0.8333333333333334
Decision Tree: 0.7222222222222222
K nerest Neighbor: 0.8333333333333334
```

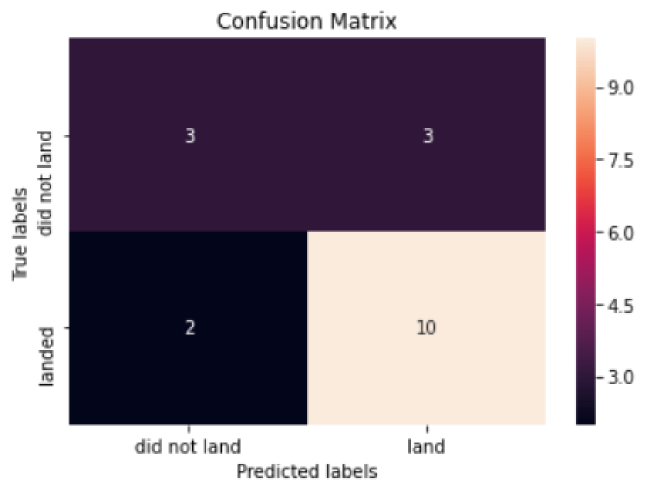
Classification Accuracy



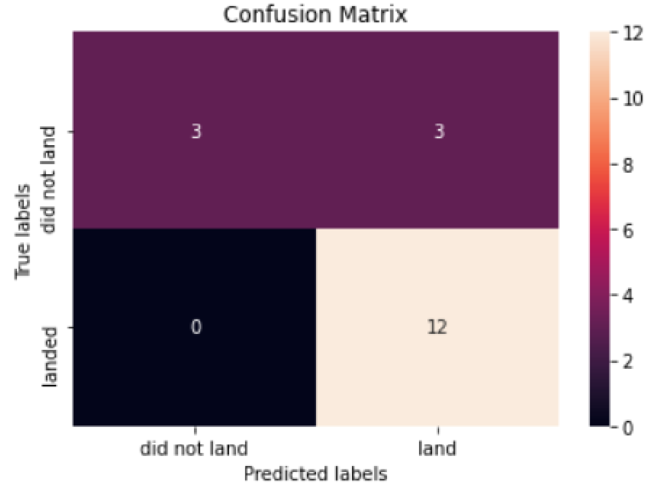
Logistic Regression



SVM



Decision Tree



KNN

Confusion Matrix

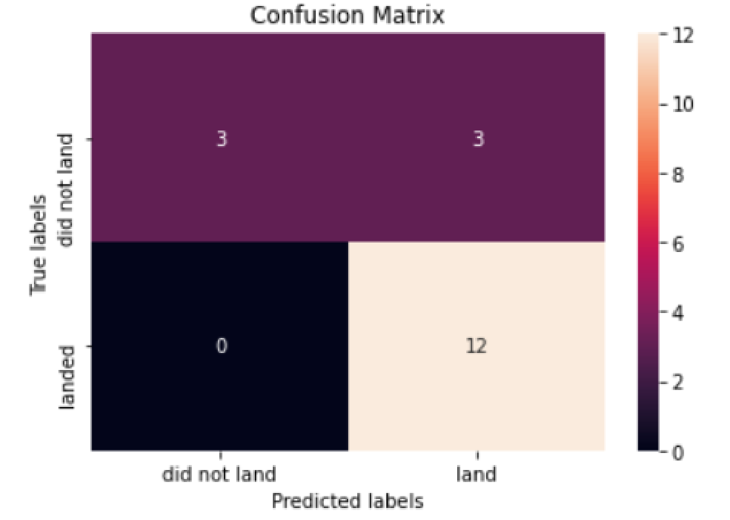
- Show the confusion matrix of the best performing model with an explanation

➢ TP: 12

➢ TN: 3

➢ FP: 3

➢ FN: 0



➢ **Accuracy** of an algorithm is represented as the ratio of correctly classified land (TP+TN) to the total number of launches (TP+TN+FP+FN) = **83.3%**

➢ **Precision** of an algorithm is represented as the ratio of correctly classified land (*TP*) to the total launches predicted to land (*TP+FP*) = **80%**

➢ **Recall** metric is defined as the ratio of correctly classified land (*TP*) divided by total number of launches than are landing *(TP+FN)* = **100%**

➢ **F1** score is also known as the F Measure. The F1 score states the equilibrium between the precision and the recall *(2*precision*recall)/(precision + recall)* = **88.9%**

Conclusions

- LR, SVM and KNN models have similar classification performance and have highest classification of 83.3% for this training and testing datasets
- Decision tree has the lowest accuracy rate at 72%
- KSC LC-39A site has the highest launch success rate
- Launch sites are more likely to be close to coastline and away from cities and highway
- There is relationship between site location and success land
- As payload is increasing, the success rate seems getting lower
- ES-L1, GEO, HEO, SSO have the highest success rate
- Overall, the success rate keeps increasing since 2013

Thank you!