IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vivian Chika-Ezeagu
12th February, 2023

Outline

Executive Summary

- The data for this project was collected from two sources: Space X API GET request and webscraping.
- The data was processed and then wrangled to gain insight; exploratory data analysis was carried out through visualizations.
- Interactive visual analysis was carried out using Folium and Plotly Dash.
- Predictive analysis was performed using classification models. The models were evaluated on train and test data to achieve high levels of accuracy.
- It was determined that heavy payload masses have a negative effect on the success of a launch while the success rate increases with higher flight numbers and time.
- All this information can be used to determine if Space X will reuse its first stage.

Introduction

- SpaceX's accomplishments include sending spacecrafts to the International Space Station.

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars while other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

- The aim of the project is to predict if the Falcon 9 first stage will land successfully.

- The models are trained and used for this purpose by determining the cost of each launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Rocket launch data was requested from SpaceX API using a URL.
  - Falcon launch 9 records were extracted from Wikipedia through web scraping.
  - Some columns of data were filtered out and data wrangled by the replacement of missing values with mean values.
- Performed exploratory data analysis (EDA) using visualization and SQL.
- Performed interactive visual analytics using Folium and Plotly Dash.
- Performed predictive analysis using classification models.

Data Collection

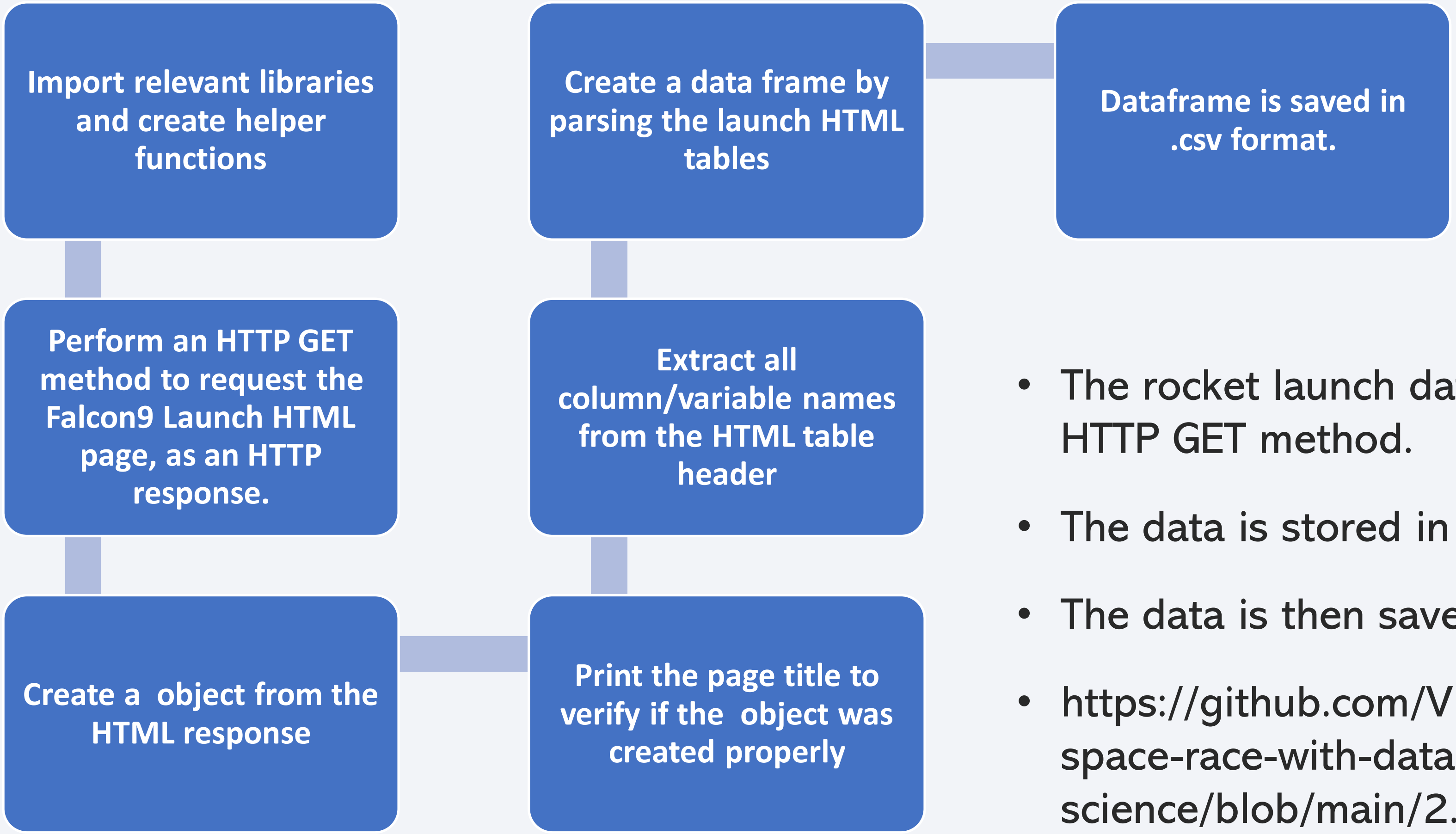
- Data was collected through requests from the Space X API.
- Data was also collected from wikipedia through a web scraping process.

Data Collection – SpaceX API

- Import Libraries and Define Auxiliary Functions
- Request rocket launch data from space X API using a URL
- Using the GET request , request and parse the data
- Data is stored in lists
- Update the lists using auxilliary functions
- Create a dictionary mapping the data with column names
- Create a dataframe from the dictionary
- Filter dataframe to include falcon 9 launches only
- Replace missing values with the mean.
- Dataframe is turned into .csv file.

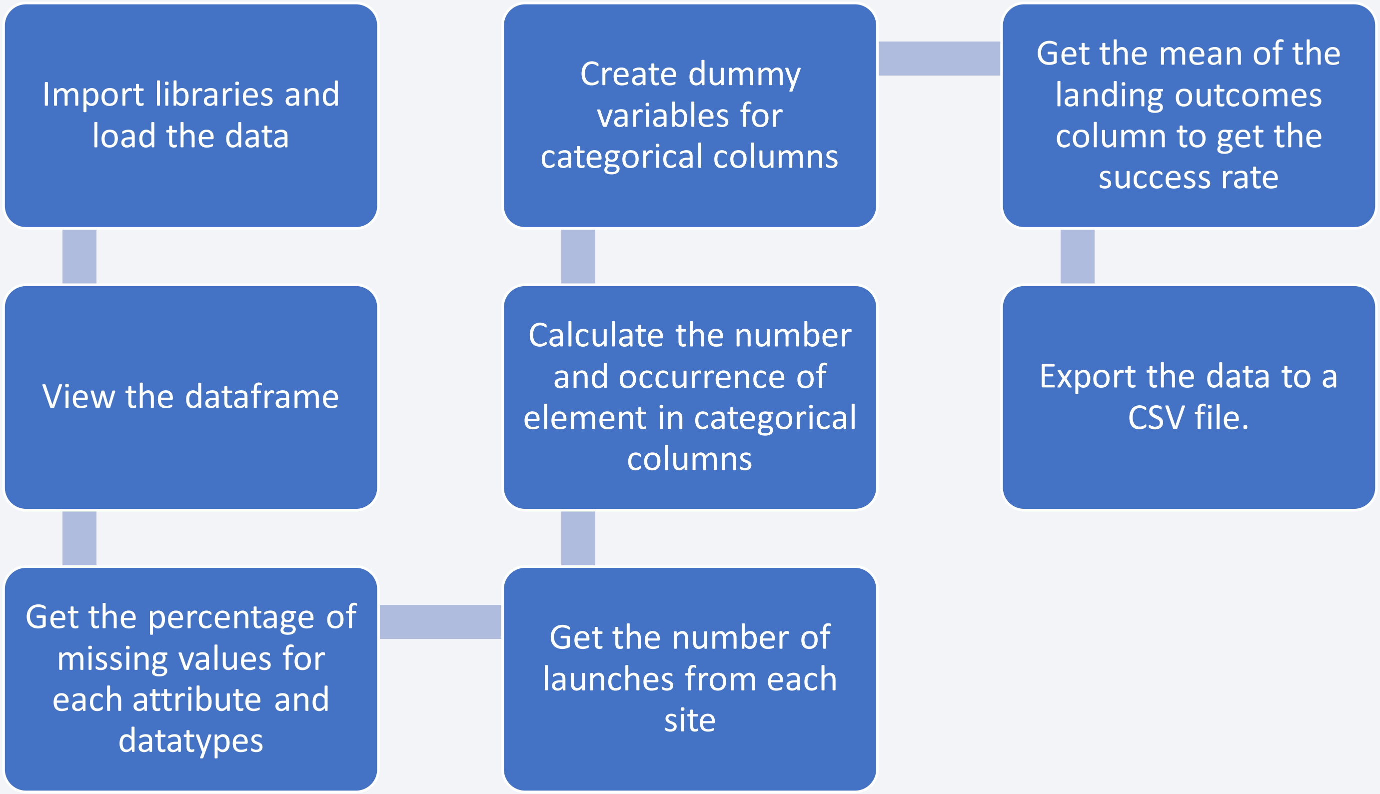
- Launch data is requested using a URL.
- The data is parsed and stored in a list.
- The data is stored in a dataframe, filtered, wrangled and stored in .csv format.
- https://github.com/VivianEzeagu/Winning-the-space-race-with-data-science/blob/main/1.%20Data-collection-API.ipynb

Data Collection – Web Scraping



- The rocket launch data is requested using the HTTP GET method.
- The data is stored in a data frame.
- The data is then saved in a .csv format.
- https://github.com/VivianEzeagu/Winning-the-space-race-with-data-science/blob/main/2.%20Data-collection-webscraping%20.ipynb

Data Wrangling



- The missing values were found and replaced by mean values.
- The categorical variables where presented and replaced with dummy variables.
- Determination of success rate of landing outcome.
- https://github.com/VivianEzeagu/Winning-the-space-race-with-data-science/blob/main/3.%20Data_wrangling.jupyterlite%20(1).ipynb

EDA with Data Visualization

- Numerous scatterplots were created to aid data visualization.
- The scatterplots were created in order to understand the relationships between variables and determine the extent of their effect on the outcome of the launch.
- The plot of the flight number vs payload mass shows that the first stage is more likely to succeed with a higher flight number and less likely to succeed with higher payload mass.
- A bar chart was created to visualize the success rate of each orbit.
- Based on the information gotten from the data visualization, a new table was created with the variables that can help predict the outcome of the first stage.
- https://github.com/VivianEzeagu/Winning-the-space-race-with-data-science/blob/main/4.%20EDA%20%26%20dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- The dataset was downloaded and loaded into the database table.
- A connection to the database was established.
- Exploratory analysis was carried out on the data to gain key insights.
- https://github.com/VivianEzeagu/ibm-projects/blob/main/login.html

Build an Interactive Map with Folium

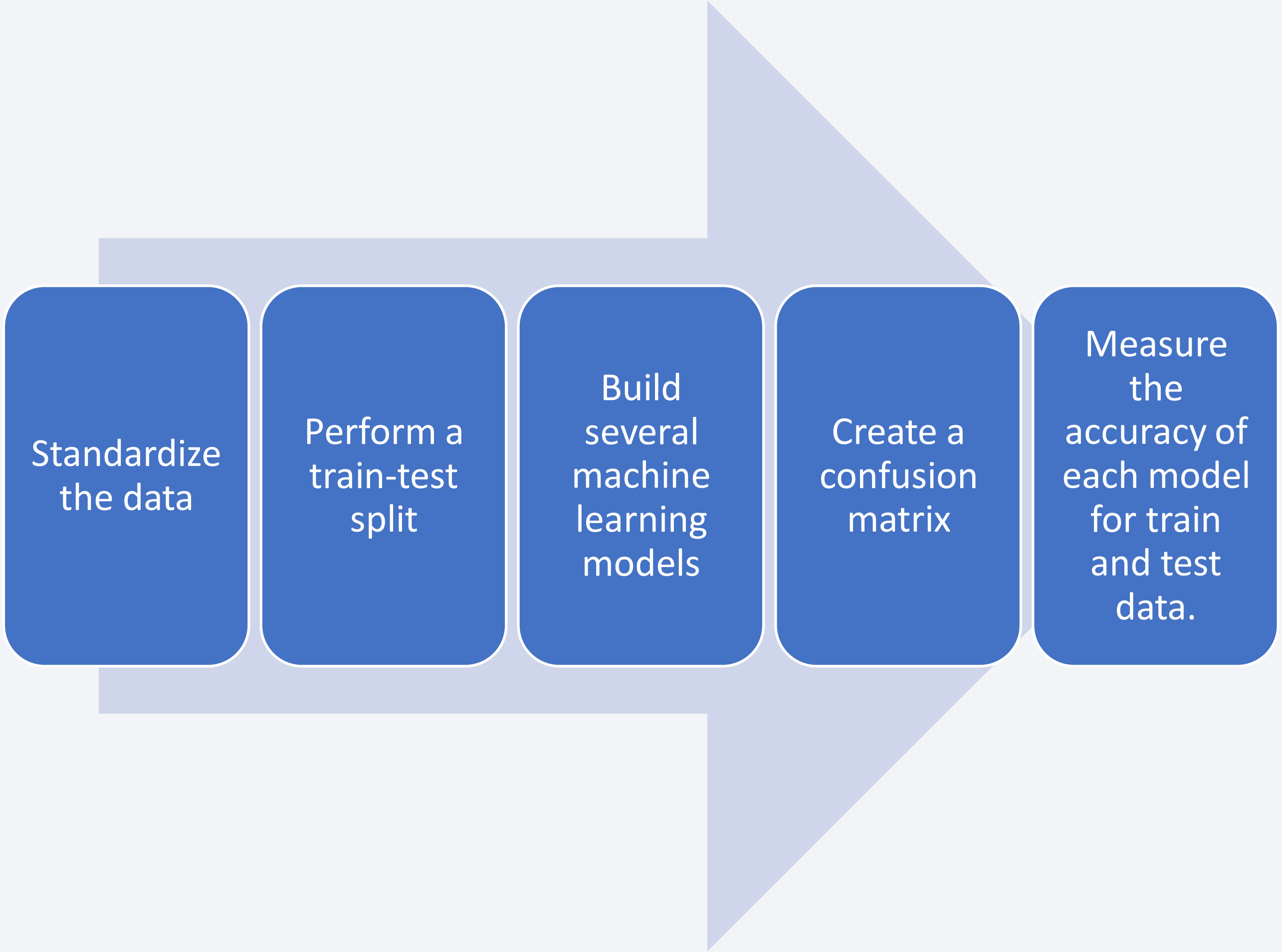
- Launch success may also depend on the location and proximity of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors, we discovered some of the factors by analyzing the existing launch site locations.
- Markers were created to visualize each of the launch sites on the map. The circle was created to highlight the site locations and aid visualization.
- A marker color object for each launch outcome with unique color identifiers was also added to enhance the map and see which sites have high success rates.
- A mouse pointer was added to enhance the ease of getting the longitude and latitudes of the launch sites.
- A line was added to visualize the distance between launch sites and significant structures like railways.
- https://github.com/VivianEzeagu/Winning-the-space-race-with-data-science/blob/main/Interactive%20visual%20analytics-folium.ipynb

Build a Dashboard with Plotly Dash

- A dropdown list was added to enable Launch Site selection.
- A pie chart was added to show the total successful launches count for all sites.
- Added a scatter chart to show the correlation between payload masses and launch success.
- https://github.com/VivianEzeagu/Winning-the-space-race-with-data-science/blob/main/Interactive%20visual%20analytics-plotly.ipynb

Predictive Analysis (Classification)

- An object was created for each model then a GridSearchCV object was also created.
- The object was fitted to find the best parameters from the dictionary parameters.
- An output of the GridSearchCV object for each model was produced.
- The best parameters were displayed and the accuracy on the validation data was calculated.
- https://github.com/VivianEzeagu/ibm-projects/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb



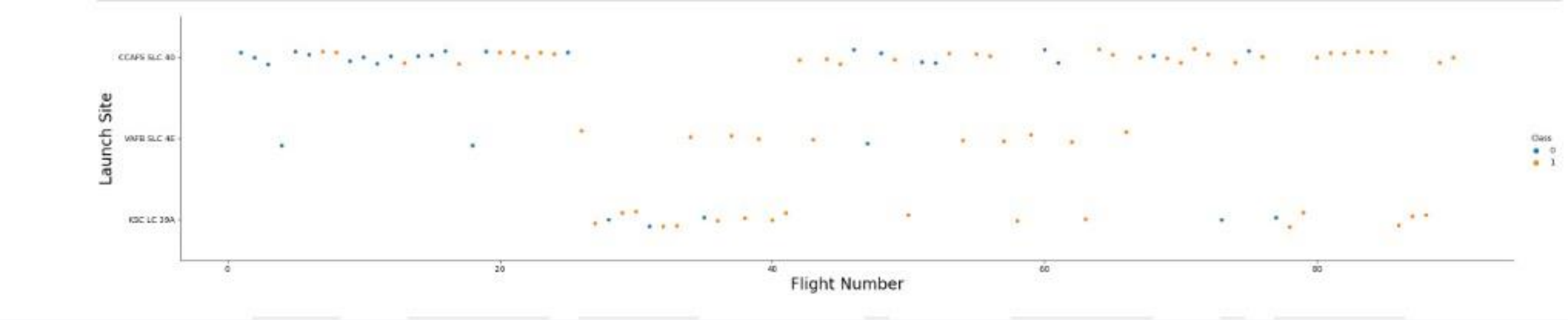
Results

- Insights drawn from Exploratory Data Analysis
- Launch sites proximities analysis
- Predictive analysis results

Section 2

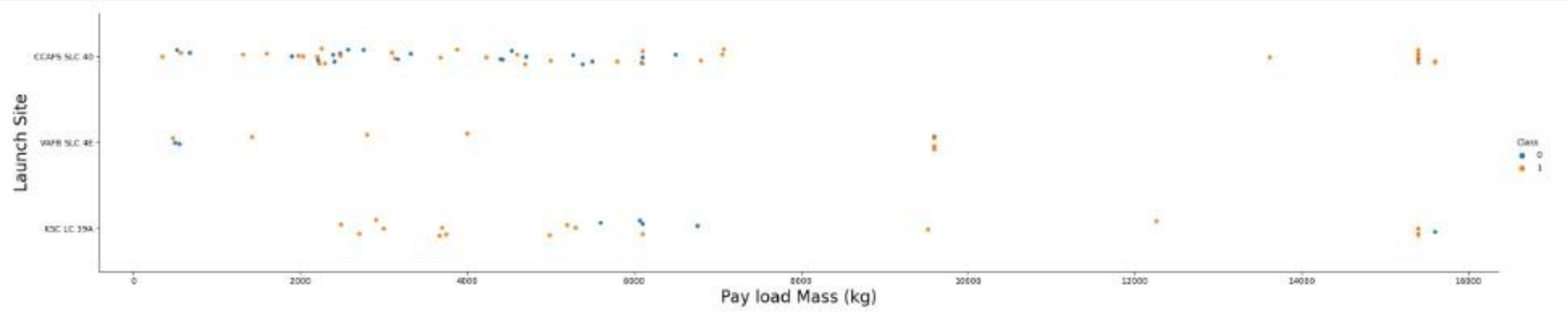
Insights drawn from EDA

Flight Number vs. Launch Site



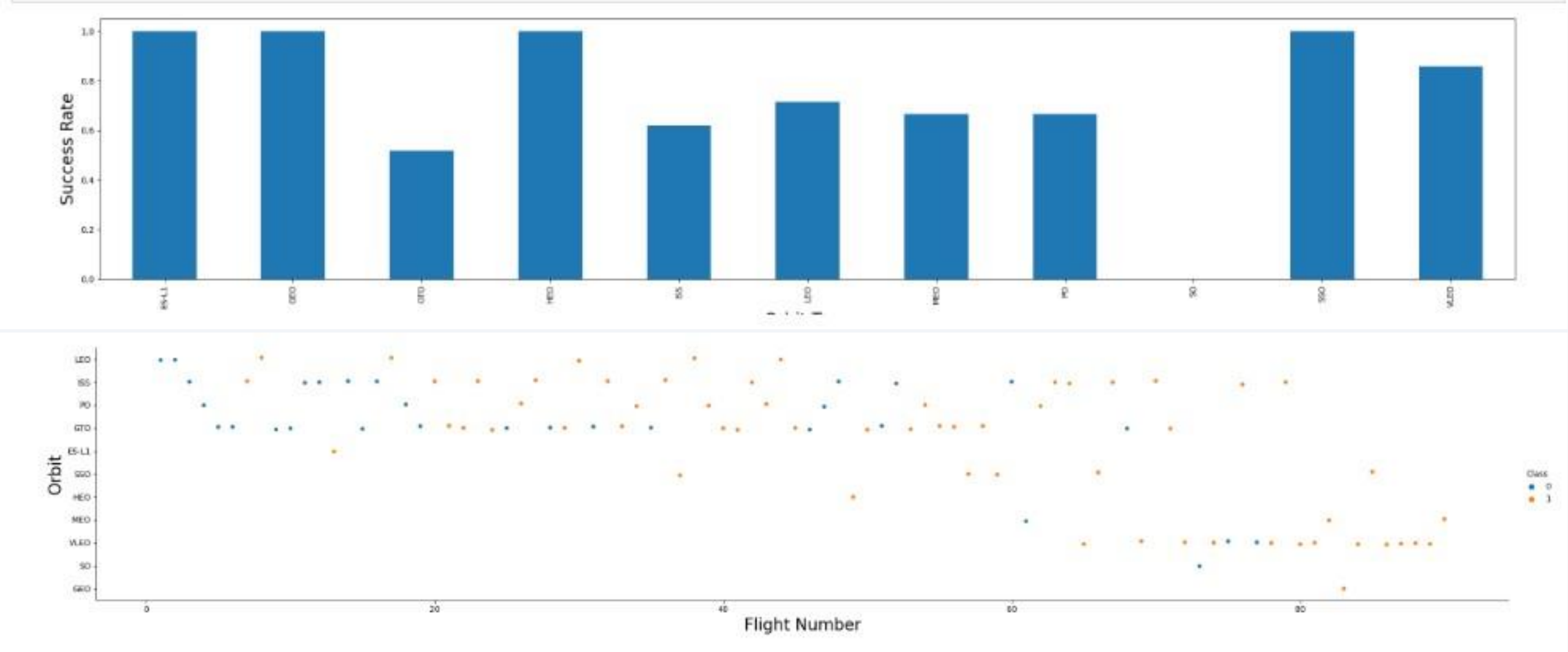
- This scatter plot shows that as the flight number increases the first stage is more likely to be successful.
- KSC LC 39A had more successful launches than unsuccessful.
- CCAFS SLC 40 had more unsuccessful launches at the initial flight attempts, the launches became more successful with increasing flight number.
- VAFB SLC 46 had more successful launches than unsuccessful.

Payload vs. Launch Site



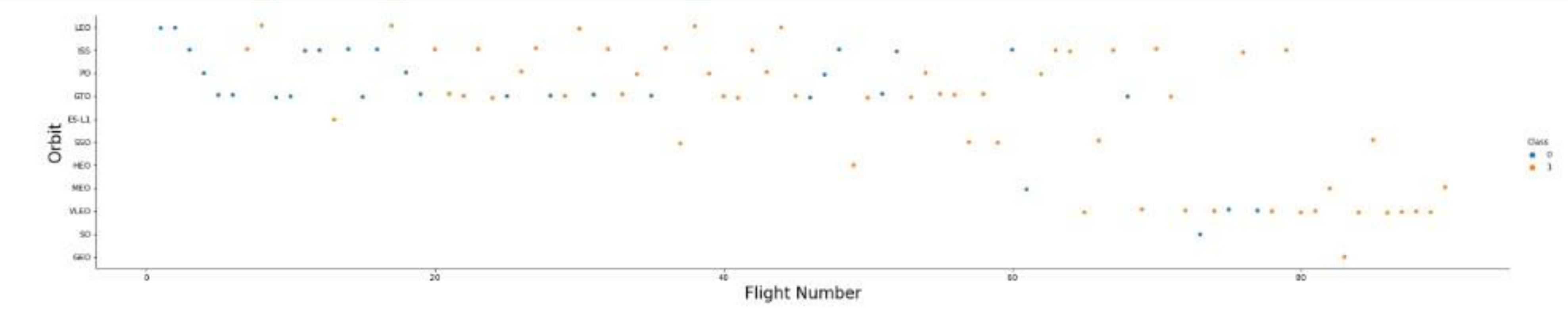
- VAFB SLC 46 had fewer launches than the other sites but had more successful launches.
- KSC LC 39A had more launches with a wide range of payload mass but had more successful launches.
- CCAFS SLC 40 had most of the launches with lower pay load mass but has a nearly fair number of successful and unsuccessful launches.
- More launches generally took place with lighter payload mass.

Success Rate vs. Orbit Type



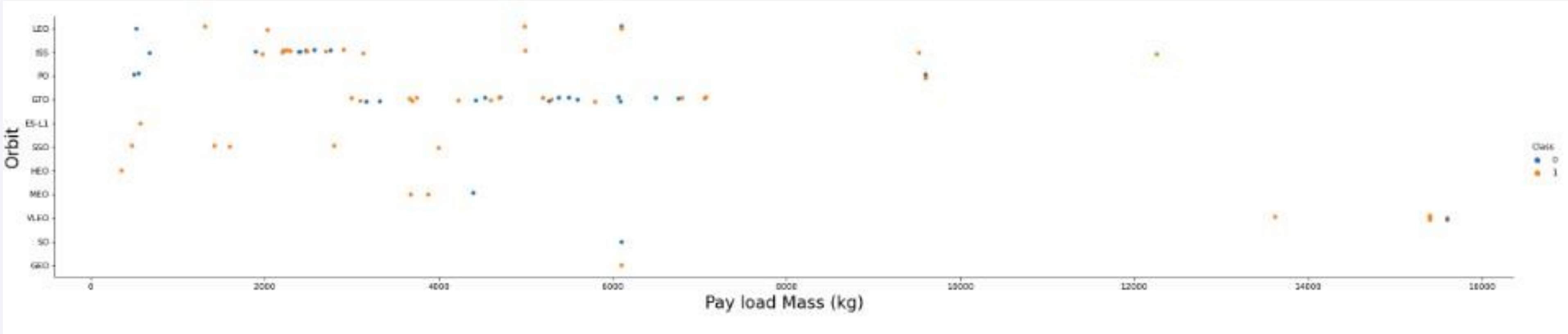
- The success rate of LEO, PO are moderate whereas VLEO success rate is a little higher.
- ES –L1, HEO, SSO and GEO had very high success rates.

Flight Number vs. Orbit Type



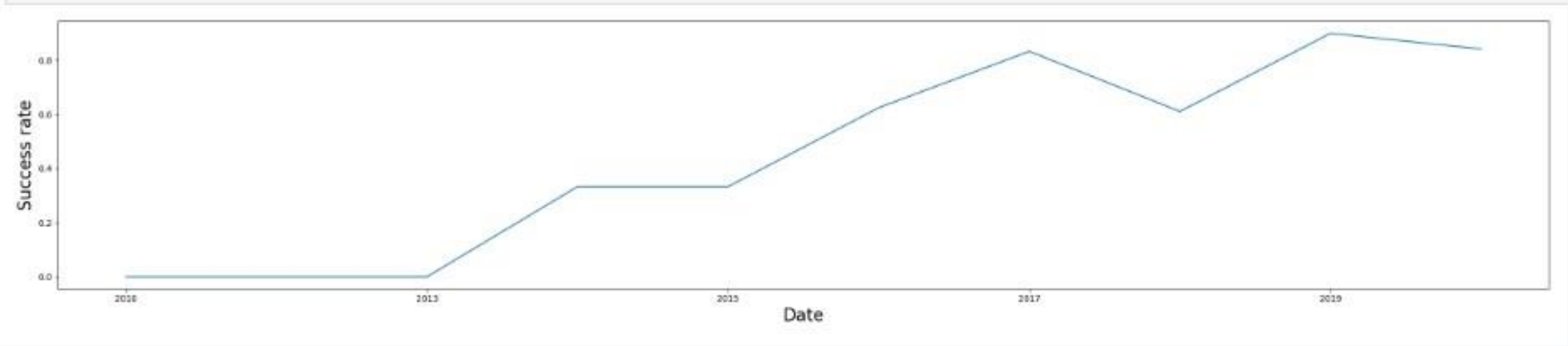
- The success rate of LEO and PO are mostly related to their flight numbers whereas GTO success rate has no correlation with it’s flight number.
- ES –L1, HEO, SSO and GEO had very few launch attempts with later flights which explains their success rate.

Payload vs. Orbit Type



- With heavy payloads, the rate of successful landings is higher for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.
- ES L1, SSO and HEO have launches with only light payload masses which were all successful.

Launch Success Yearly Trend



- The success rate was constant between 2010 and 2013.
- From 2013, the success rate had a continual increase till 2015, where there's a slight decrease, there's a continued increase after till 2017 ,where there's a decrease then another increase from 2018 till 2019 then a decrease past 2019.

All Launch Site Names

```
[12]: %%sql
select distinct LAUNCH_SITE from "SPACEXTBL"

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

[12]:

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

- There are four unique launch sites:
- CCAFS LC-40
- VAFB SLC -4E
- KSC LC -39A
- CCAFS SLC -40

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

19]:
```
%%sql
select launch_site
from "SPACEXTBL"
where launch_site like "CCA%"
limit 5
```

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

19]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- The above are five launch records that have launch site names beginning with “CCA”

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

[22]:
```
%%sql
select sum(payload_mass__kg_)
from SPACEXTBL
where customer="NASA (CRS)"
```

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

[22]:

sum(payload_mass__kg_)
45596

- The above shows the total payload mass carried by all the boosters launched by NASA(CRS).

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

```
[23]: %%sql
select avg(payload_mass__kg_)
from SPACEXTBL
where booster_version="F9 v1.1"

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

[23]:

avg(payload_mass__kg_)
2928.4

- The average payload mass carried by booster F9 v1.1 is 2928.4kg.

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

[61]:
```
%%sql
select min(DATE)
from SPACEXTBL
where "Landing _Outcome"  = 'Success (ground pad)'
```

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

[61]:

min(DATE)
01-05-2017

- The above query shows the first successful landing that took place on a ground pad, which was on the 1st of May, 2017.

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

[63]:
```
%%sql
select booster_version,payload_mass__kg_
from SPACEXTBL
where "Landing _Outcome" = "Success (drone ship)" and payload_mass__kg_ between 4000 and 6000
```

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

[63]:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- The above shows boosters which has landing success in drone ship with a payload mass between 4000kg and 6000kg.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

[72]:
```
%%sql
select count(mission_outcome)
from SPACEXTBL
where mission_outcome like "Success%"
```

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

[72]:

count(mission_outcome)
100

- The above shows the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
[75]: %%sql
select booster_version,payload_mass__kg_
from SPACEXTBL
where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)

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

[75]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- The boosters that have carried maximum payload mass are shown in the table above.

2015 Launch Records

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

[78]:
```
%%sql
select substr(Date, 4, 2),booster_version,"Landing _Outcome",launch_site from SPACEXTBL
where substr(Date,7,4)='2015'
and "Landing _Outcome" = "Failure (drone ship)"
```

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

[78]:

substr(Date, 4, 2)	Booster_Version	Landing _Outcome	Launch_Site
01	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

- The result shows the months and other records in 2015 that landing failed in the drone ship.

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

Task 10

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

[100]:
```
%%sql  SELECT Date, "Landing _Outcome",count("Landing _Outcome")as LANDING_OUTCOME_COUNT
from SPACEXTBL
where   substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) between '20100604'and '20170320'
and  "Landing _Outcome" like "%Success%"
group by "Landing _Outcome"
order by count("Landing _Outcome")
desc
```

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

[100]:

Date	Landing _Outcome	LANDING_OUTCOME_COUNT
08-04-2016	Success (drone ship)	5
22-12-2015	Success (ground pad)	3

- The number of successful landing outcomes between 4th of June,2010 and 20th of March, 2017 are shown in descending order.

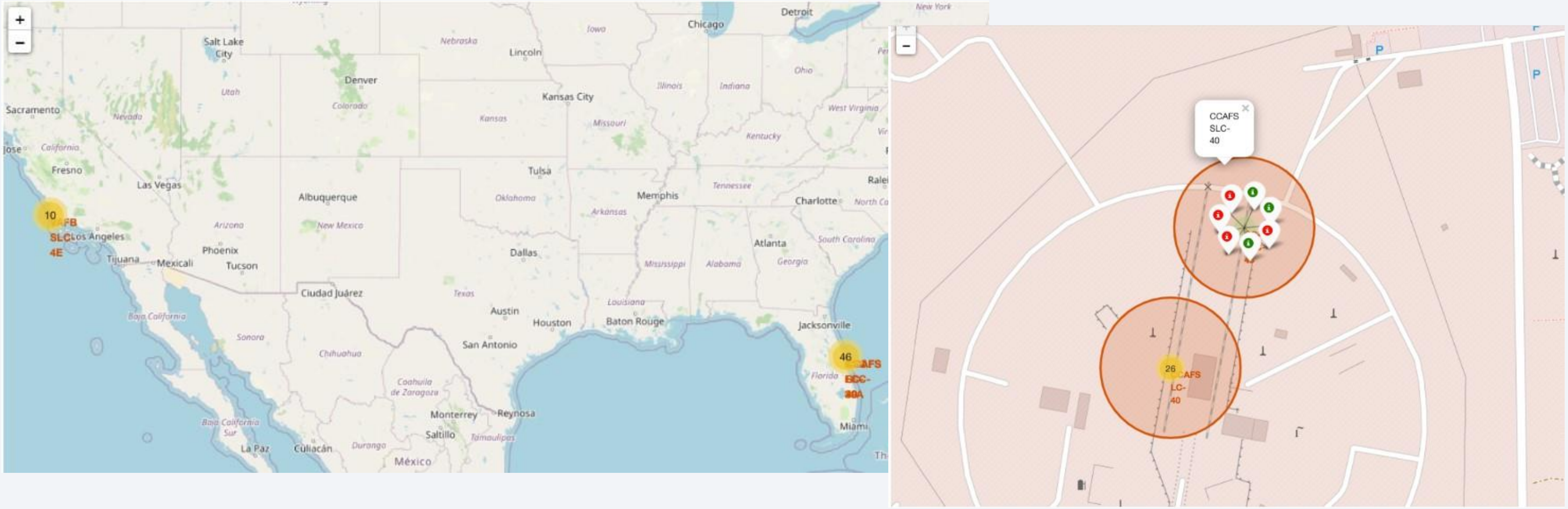
Section 3

Launch Sites Proximities Analysis

Global map displaying launch site locations

- This map shows the locations of all the launch sites on a global map.
- The launch sites are all at close proximity to the coastline which is safe area because it is far away from cities, highroads and other areas where it could be harmful to humans.

Folium map displaying the launch outcomes of the launch sites



- The map shows the color-labeled launch outcomes for each site, where green represents successful launch outcomes and red represents the unsuccessful outcomes.

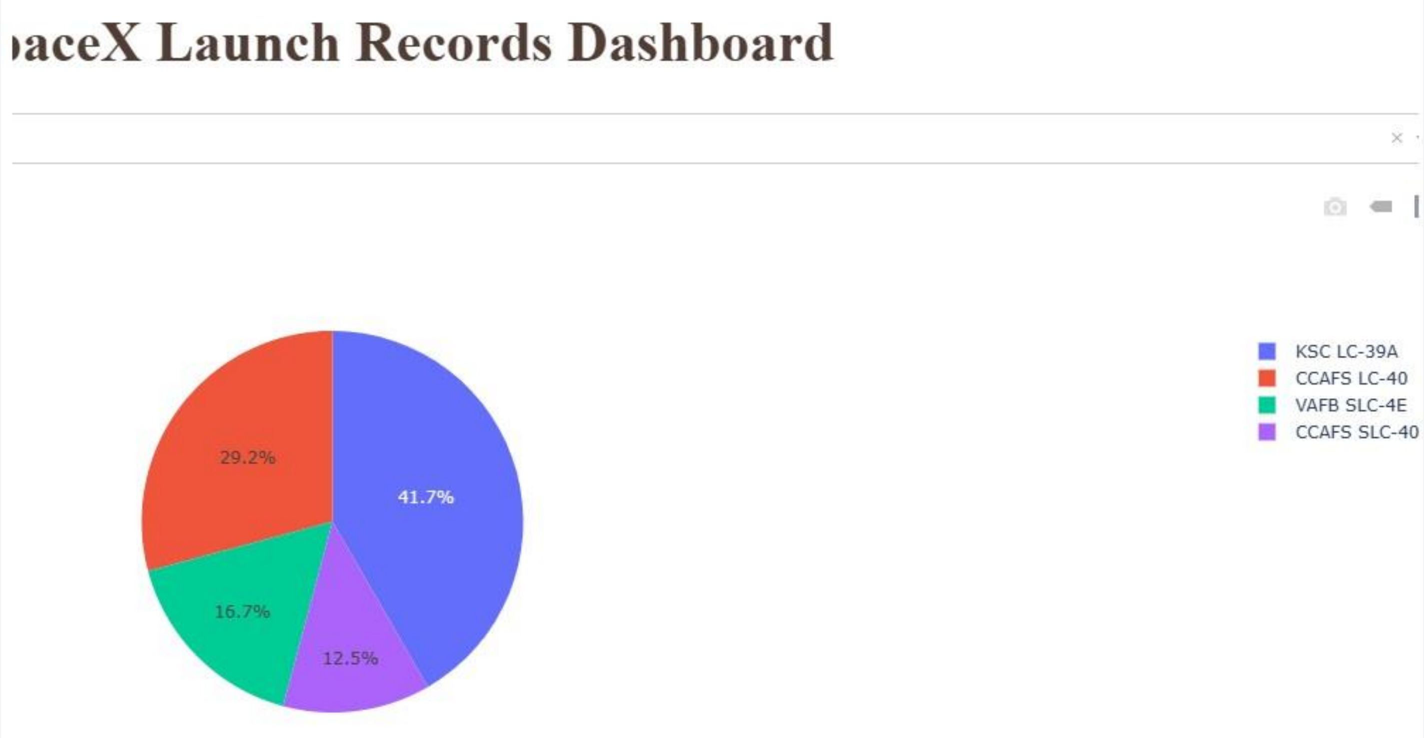
Folium map displaying the distance of VFAB SLC-4E from the coastline

- This shows the distance of VFAB SLC-4E from the coastline.

Section 4

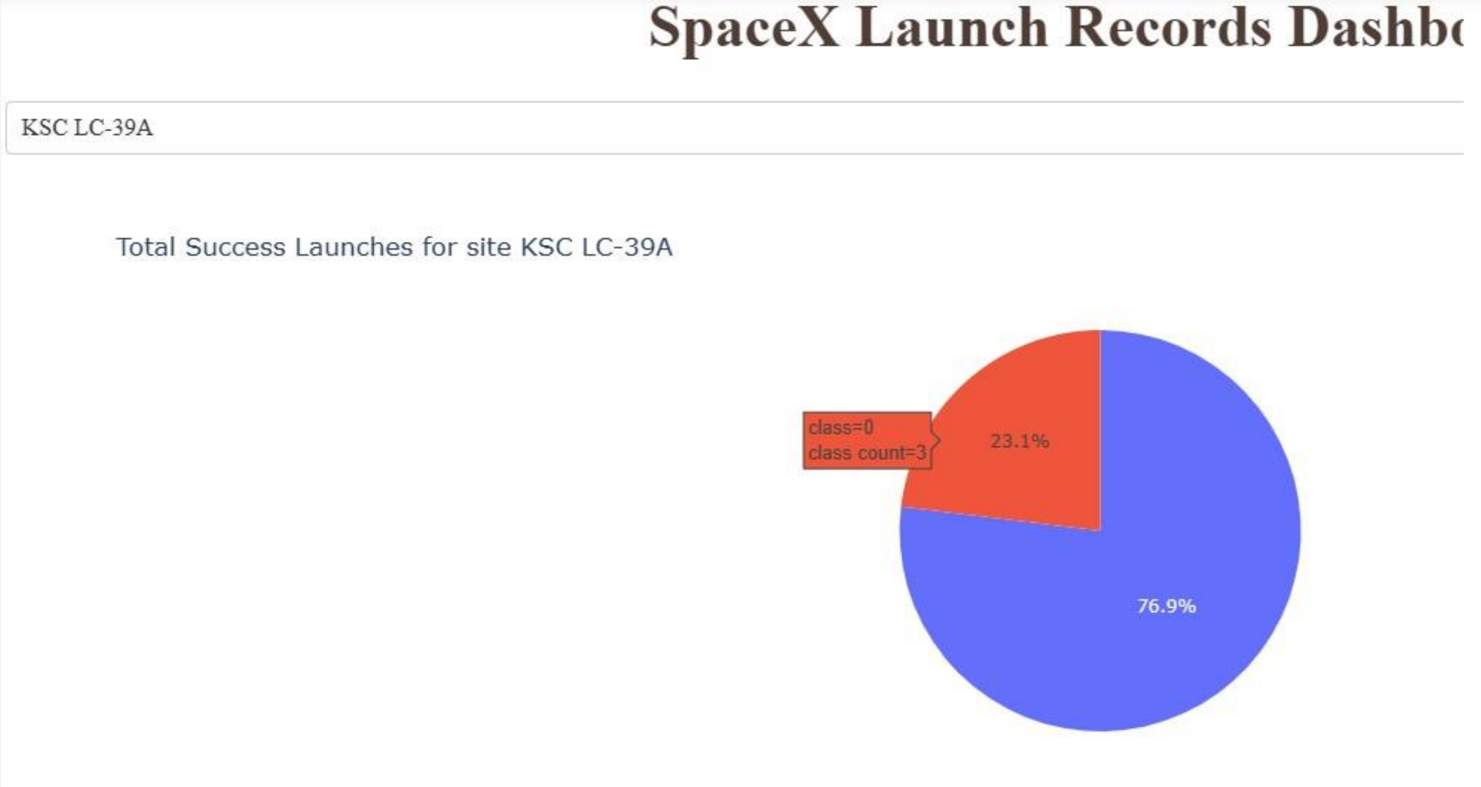
Build a Dashboard with Plotly Dash

Pie chart of success count for all launch sites



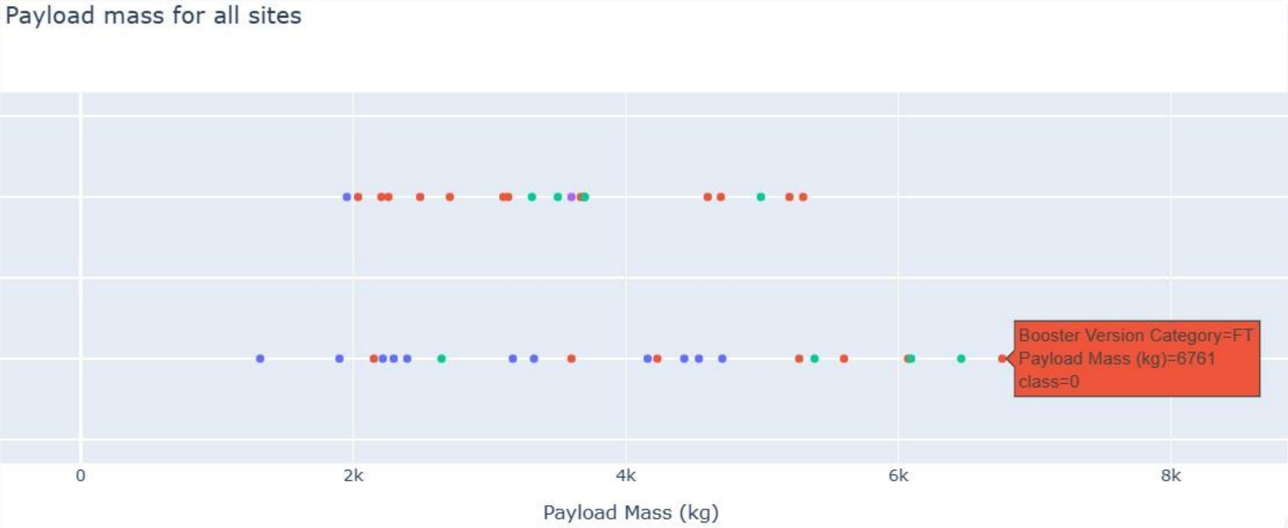
- The KSC LC 39A has the highest percentage of successful landings.
- The CCAFS SLC -40 has the least successful launches.

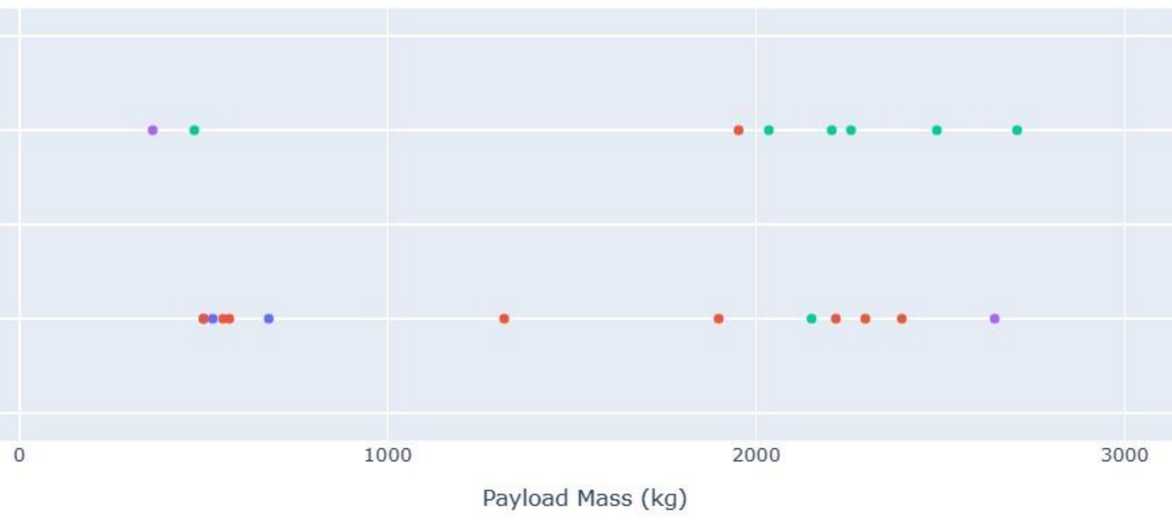
Pie chart of the KSC LC -39A

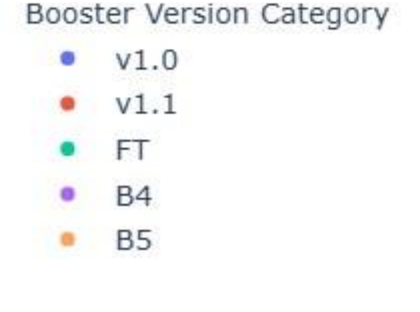


- The KSC LC -39A launch site has the most successful launches overall with 76.9% of the launches being successful.

Scatter plots of Payload vs. Launch Outcome







- The scatter plots show the success rate of the booster versions per payload mass.
- Booster v1.1 has high success rate with payload mass between 2000kg and 6000kg while v1.0 experiences the opposite.
- Booster FT has more successes within 500kg-3000kg payload range.

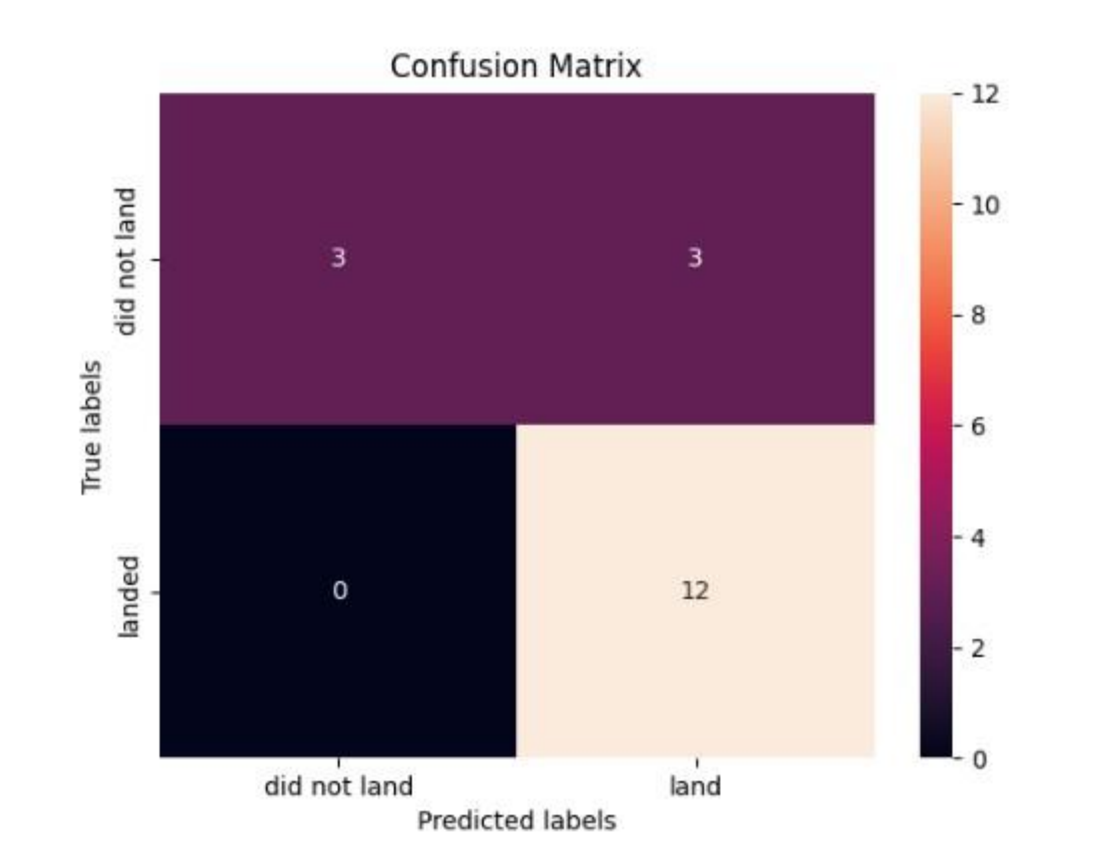
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The decision tree model has the highest accuracy with an accuracy of 0.94.

Confusion Matrix



- The best performing model is the decision tree model.
- The confusion matrix shows that 12 successful launches were correctly predicted and 3 incorrect, 3 unsuccessful launches were correctly predicted.

Conclusions

- This data set revealed interesting insights; the payload can determine the landing outcome of the launch depending on the booster version that carries it.
- As the flight number increases, the success rate increases.
- The booster v1.1 has a lot of successful outcomes across wide payload range.
- The KSC LC -39A has the most successful launches, which could be influenced by orbit, boosters and payload mass.
- The optimal machine learning model to be used is the decision tree model because of the high accuracy level of 0.94.
- A higher flight number, booster v1.1 and launch site KSC LC -39A are great ingredients for the successful launch of the first stage of Falcon9. The orbit and payload mass have varied effects on the success of launch depending on some other factors.

Appendix

[5]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	Reused(
78	79	2020-05-30	Falcon 9	9525.000000	ISS	KSC LC 39A	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	
79	80	2020-06-04	Falcon 9	15400.000000	VLEO	CCAFS SLC 40	True ASDS	5	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	
80	81	2020-06-13	Falcon 9	15400.000000	VLEO	CCAFS SLC 40	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
81	82	2020-06-30	Falcon 9	3880.000000	MEO	CCAFS SLC 40	True ASDS	1	True	False	True	5e9e3033383ecbb9e534e7cc	5.0	
82	83	2020-07-20	Falcon 9	6104.959412	GEO	CCAFS SLC 40	True ASDS	2	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	
83	84	2020-08-18	Falcon 9	15400.000000	VLEO	CCAFS SLC 40	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
84	85	2020-08-30	Falcon 9	1600.000000	SSO	CCAFS SLC 40	True RTLS	4	True	True	True	5e9e3032383ecb267a34e7c7	5.0	
85	86	2020-09-03	Falcon 9	15400.000000	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
86	87	2020-10-06	Falcon 9	15400.000000	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
87	88	2020-10-18	Falcon 9	15400.000000	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	

- Dataset worked on is shown above.

Thank you!