IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shibasis Bhowmik
04/28/2024

Outline

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

Executive Summary

- Falcon 9 Launch Prediction methodologies
  - Exploratory data analysis(EDA) with SQL
  - Interactive visual analytics using Folium and Plotly Dash
  - Predictive analysis using classification model

- Summary of all Prediction results
  - accuracy_logreg : 0.8464285714285713
  - accuracy_svm : 0.8482142857142856
  - accuracy_tree : 0.8642857142857142
  - accuracy_knn : 0.8482142857142858

Introduction

- Our objective is to predict whether SpaceX will attempt to land a rocket or not
- We will create a machine learning pipeline to predict if the first stage will land given the which can estimate the Falcon 9 launch Cost.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Space X Rest API
  - Web Scrapping
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for trai ning supervised models.

Data Collection

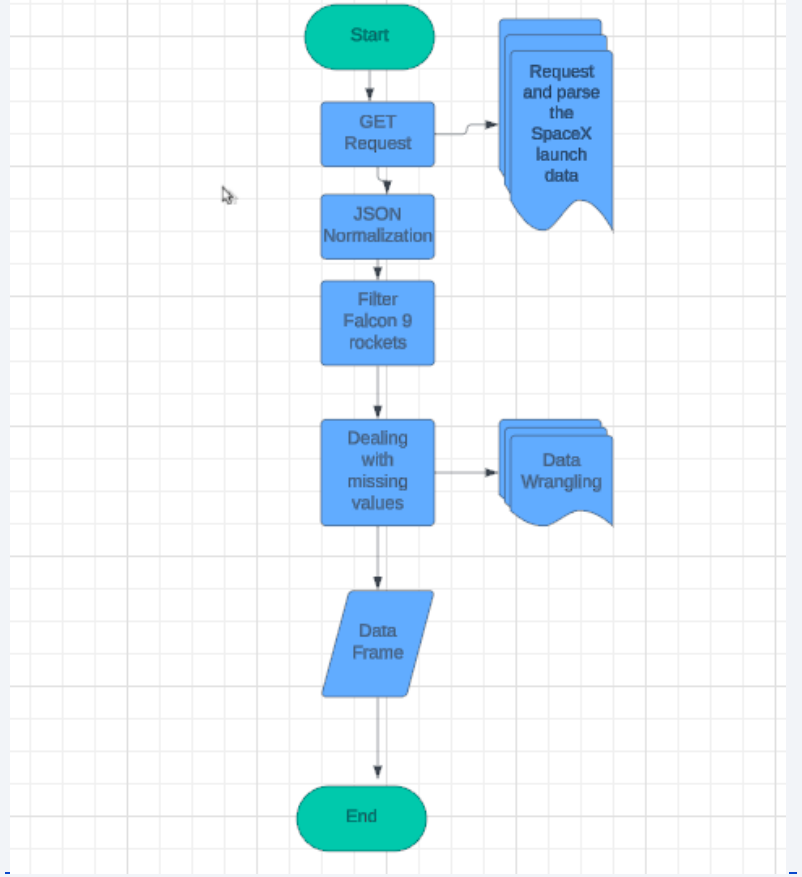
- Request to the SpaceX API and Url for all past rocket launces
- Clean the requested data and construct data set
- Filter the data frame with only Falcon 9 launches
- Dealing with missing values

Data Collection – SpaceX API

- GET request
- JSON normalization
- Filter data
- Handle Missing values
- Final Data Frame

- Github Link

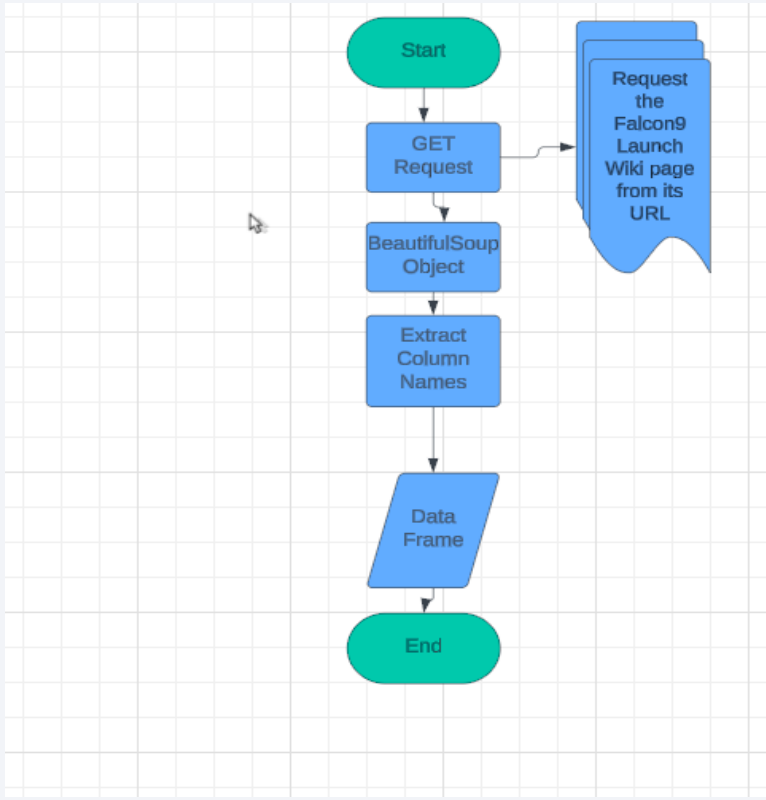
https://github.com/shibho05/Data-Science-Assigmnets/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb



Data Collection - Scraping

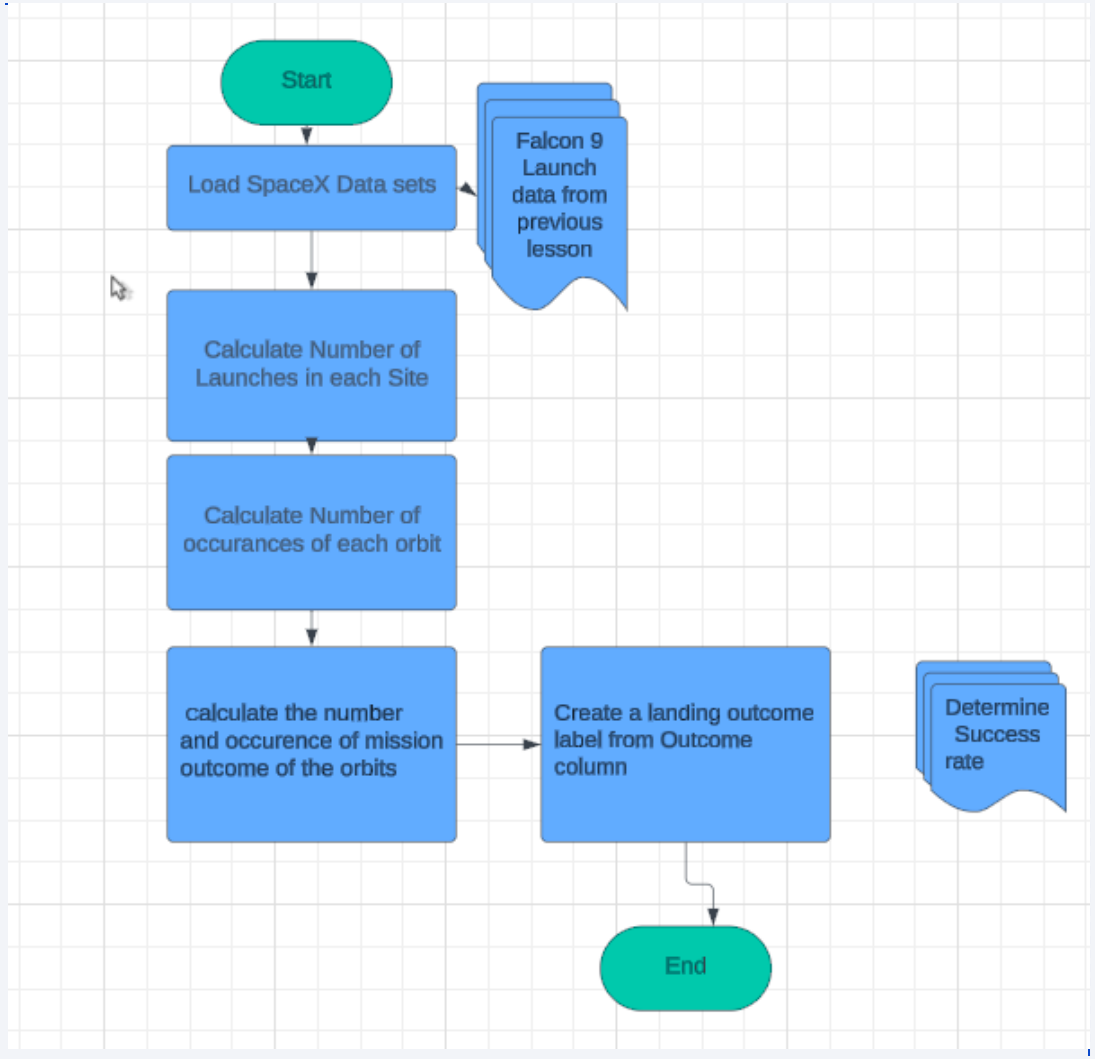
- Get response from URL
- BeautifulSoup obejct
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables
- Github Link

https://github.com/shibho05/Data-Science-Assigmnets/blob/main/jupyter-labs-webscraping%20(1).ipynb



Data Wrangling

- Load Falcon 9 Data set from the last session.
- Calculate Launches in Each Site
- Calculate Occurrences in each Orbit
- Calculate Number and occurance of mission outcome in the Orbit
- Create a landing outcome label
- Github Link
- https://github.com/shibho05/Data-Science-Assigmnets/blob/main/labs-jupyter-spacex-Data%20wrangling-Copy1.ipynb



EDA with Data Visualization

- Visualize the relationship between Flight Number and Launch Site with Scattered point chart to find out success rate
- Visualize the relationship between success rate of each orbit type WITH Bar Chart. This finds which orbits have high success rate.
- Visualize the relationship between Payload and Orbit type. With scattered point chart . This shows heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- Visualize the launch success yearly trend. With a Line chart. to get the average launch success trend.
-
- GitHub: https://github.com/shibho05/Data-Science-Assigmnets/blob/main/edadataviz%20(2).ipynb

EDA with SQL

- Display Names of Unique Launch Sites.
- 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 achieved.
- 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

EDA with SQL

- 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

- 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.

- GitHub:

https://github.com/shibho05/Data-Science-Assigmnets/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

Map Objects used:

- folium.Circle -->To add a highlighted circle area with a text label on a specific coordinate
- Folium.map.Marker--> to add markers into Folium maps
- MarkerCluster-->to group nearby markers on a map into clusters,
- MousePosition-->o display the geographic coordinates (latitude and longitude) of the mouse cursor's current position
- PolyLine-->to draw lines on a map between coordinates
- the map
- GitHub

https://nbviewer.org/github/shibho05/Data-Science-Assigmnets/blob/main/lab_jupyter_launch_site_location.ipynb#L1C0

Build a Dashboard with Plotly Dash

Below are plots/graphs and interactions added to a dashboard:

- Dropdown list and a range with launch sites. Default value shows "All Sites".
- Added an interactive pie chart to show the total successful launches count for all sites vrs. each site.
- Added a slider to interact with a pie chart and a scatter point chart.

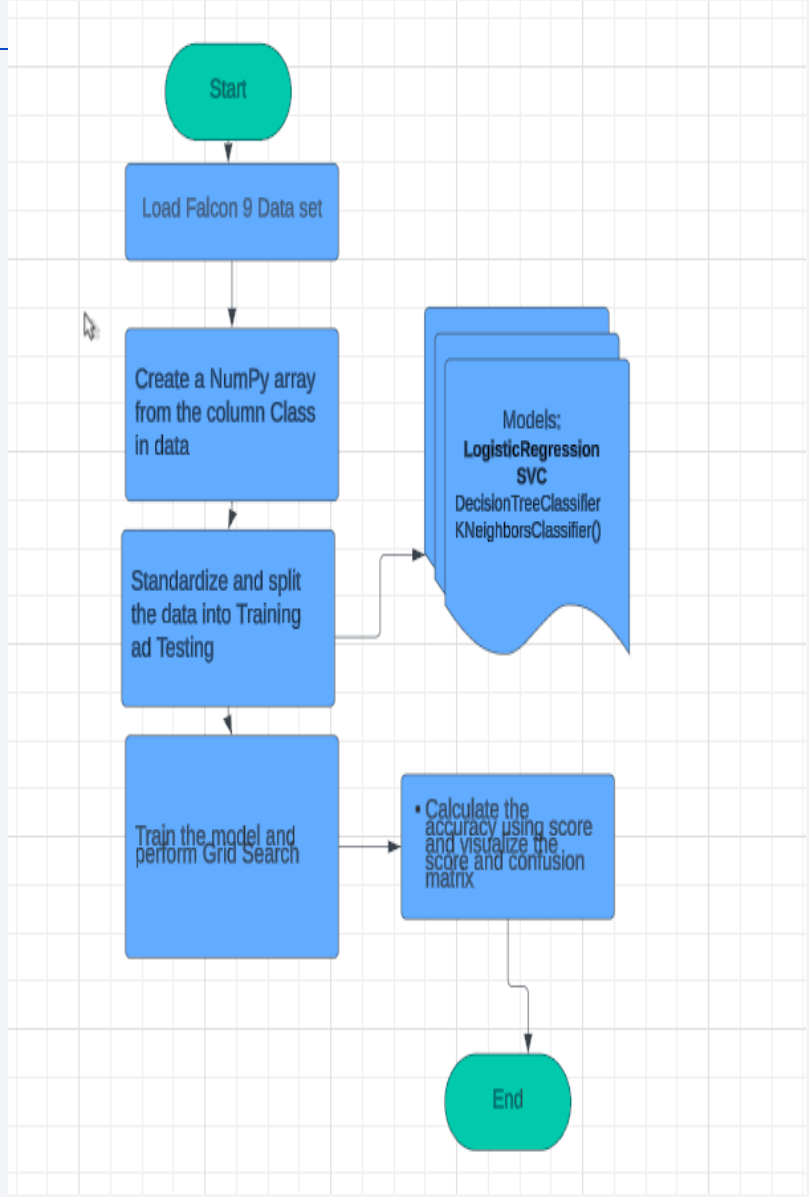
- GitHub

https://github.com/shibho05/Data-Science-Assigmnets/blob/main/spacex_dash_app%20(2).py

Predictive Analysis (Classification)

- Loaded Falcon 9 dataset
- Create a NumPy array from the column Class in data
- Split the data into training and testing
- Trained the model and perform Grid Search, allowing us to find the hyperparameters that allow a given algorithm to perform best.
- Calculate the accuracy using score and visualize the score and confusion matrix

- GitHub
- https://nbviewer.org/github/shibho05/Data-Science-Assigmnets/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20%281%29.ipynb

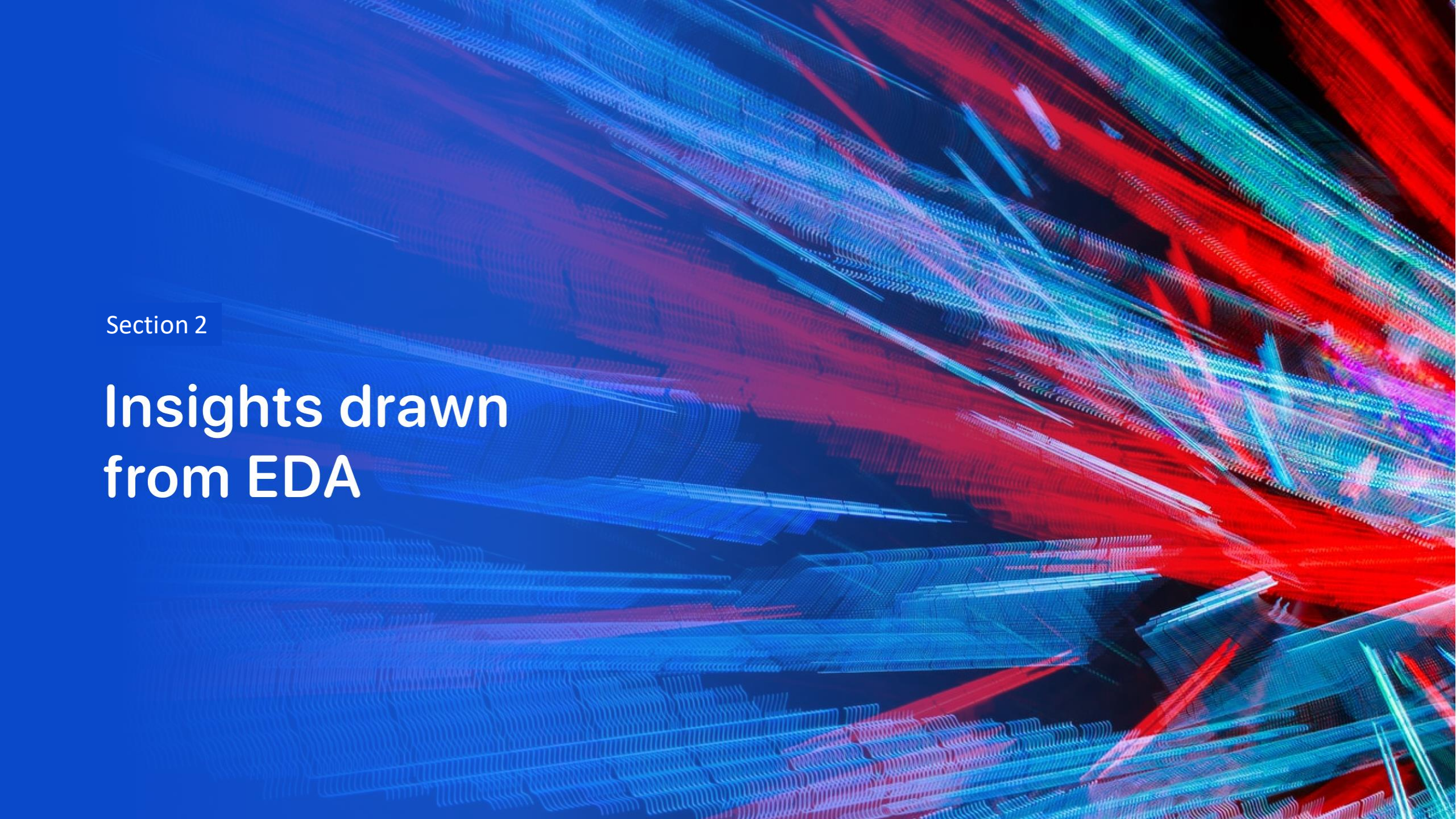


Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results:
  - print("accuracy :",logreg_cv.best_score_)
  - print("accuracy :",svm_cv.best_score_)
  - print("accuracy :",tree_cv.best_score_)
  - print("accuracy :",knn_cv.best_score_)
  -
    - accuracy : 0.8464285714285713
    - accuracy : 0.8482142857142856
    - accuracy : 0.875
    - accuracy : 0.9

Section 2

Insights drawn from EDA



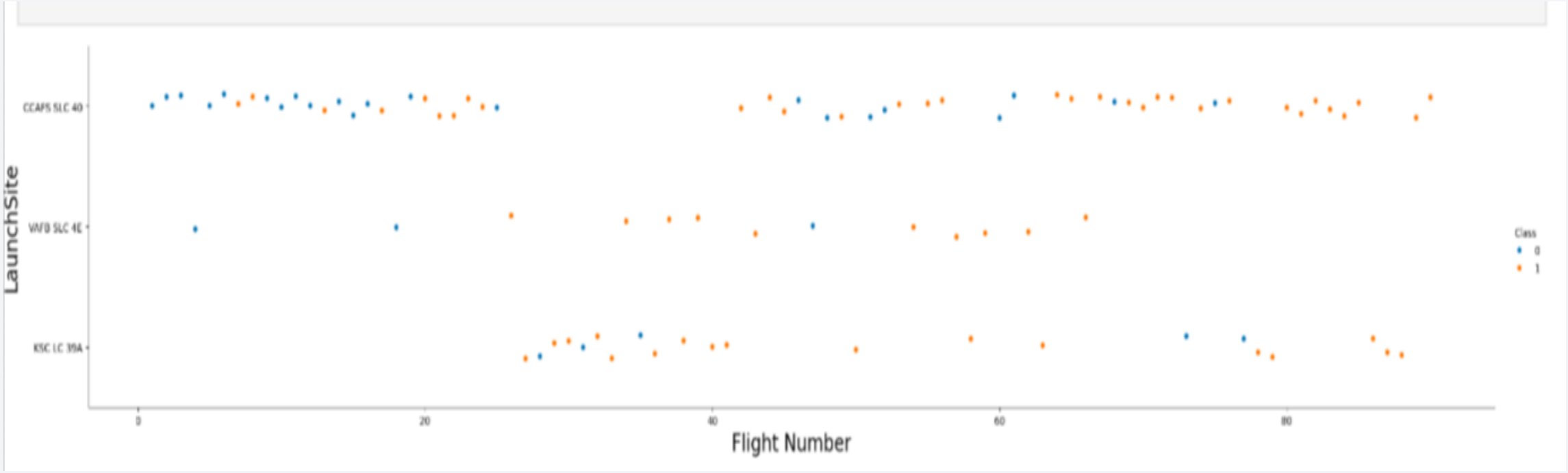
Flight Number vs. Launch Site

Plotting query:

```
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```

Flight Number vs. Launch Site

We can see from the plot that CCAFS SLC40 has more attempts between Flight 0 and 20 than other sites and beyond 80 they have 100% success rate, whereas VAFB SLC 4E site has lowest number of attempts from 0 to 20 however they have more success between 20 and 70.KSC LC 39A has lowest failure over attempts.



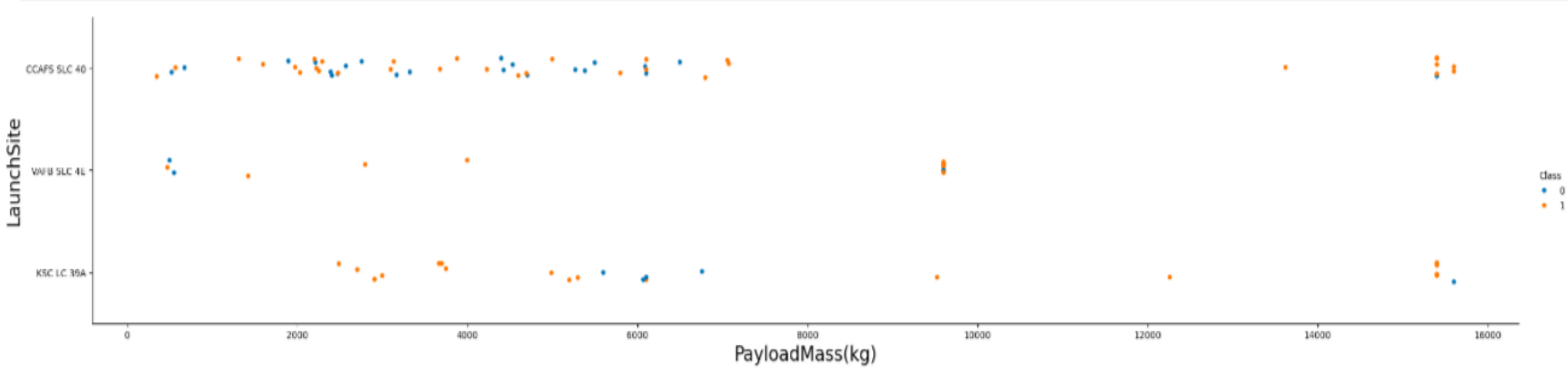
Payload vs. Launch Site

Plotting query

```
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass(kg)",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```

Payload vs. Launch Site

- CCAFS SLC-40 has launched more lightweight to moderate weight(less than 800kg) rockets and between 8000 and 14000kg there is only one attempt which is also successful whereas VAFB SLC did not launch any heavy rockets(beyond 10000kg)
- KSC 39 has more attempt rockets weight between 2000 and 10000kg.



Success Rate vs. Orbit Type

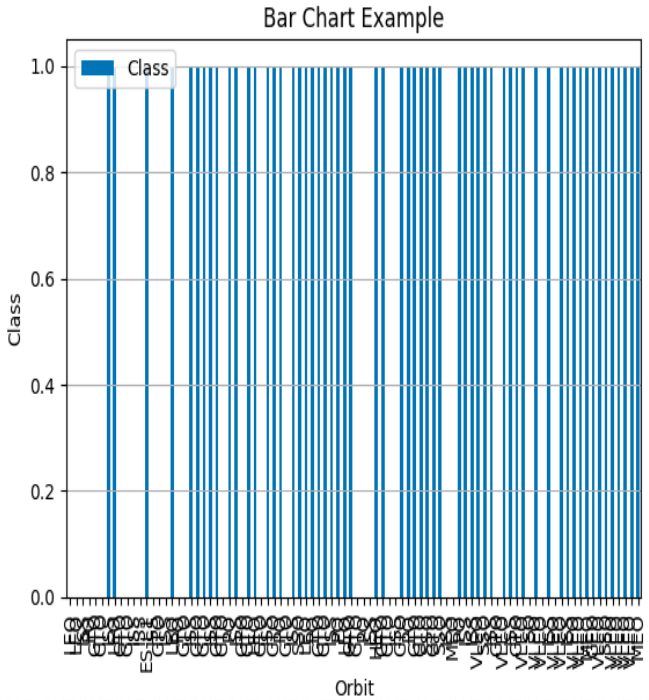
Plotting Query:

```
groupedby = df.groupby("Orbit")["Class"].mean()
print(groupedby)

# Plotting a bar chart
df.plot(x='Orbit', y='Class', kind='bar')

# Customizing the plot
plt.title('Bar Chart Example')
plt.xlabel('Orbit')
plt.ylabel('Class')
plt.grid(axis='y')  # Show gridlines only on the y-axis
plt.show()
```

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- Orbit SO has zero success rate whereas ES-L1, GEO, HEO, SSO have maximum success rates

```
Orbit
ES-L1    1.000000
GEO      1.000000
GTO      0.518519
HEO      1.000000
ISS      0.619048
LEO      0.714286
MEO      0.666667
PO       0.666667
SO       0.000000
SSO      1.000000
VLEO     0.857143
Name: Class, dtype: float64
```



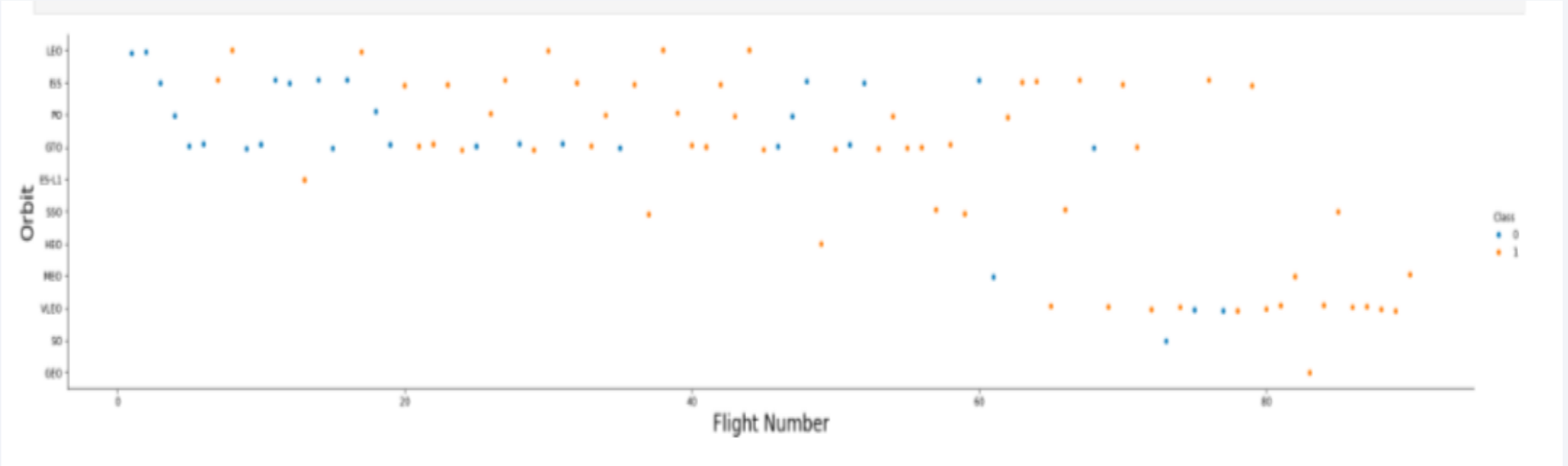
Flight Number vs. Orbit Type

Plotting Query:

```
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)

plt.xlabel("Flight Number",fontsize=20)

plt.ylabel("Orbit",fontsize=20)

plt.show()
```

Flight Number vs. Orbit Type

- LEO orbit has more success rate in co-relation with number of flights
- On the other hand, there seems to be no relationship between flight number when in GTO orbit.



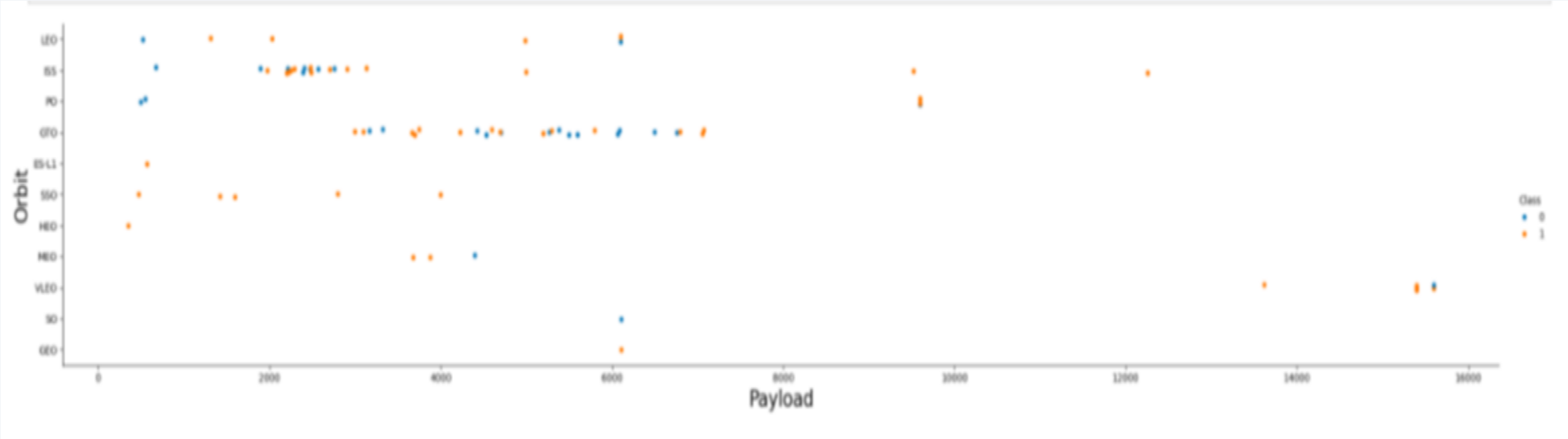
Payload vs. Orbit Type

Plotting Query:

```
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

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(unsuccessful mission) are both there here.

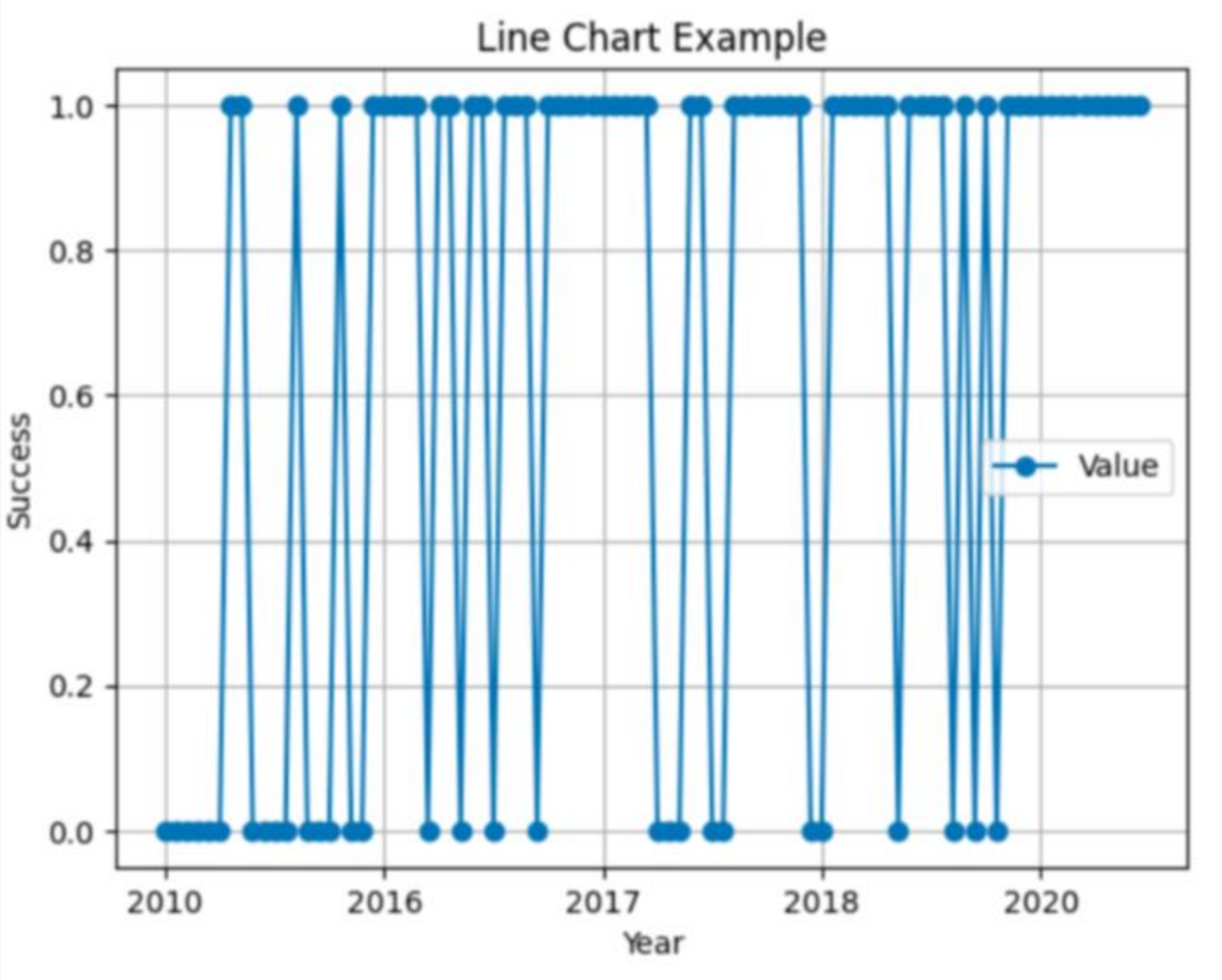


Launch Success Yearly Trend

- Plotting Query:
- *# A function to Extract years from the date*
- year=[]
- **def** Extract_year():
- **for** i **in** df["Date"]:
- year.append(i.split("/")[0])
- **return** year
- Extract_year()
- df['Date'] = year
- df.head()

Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites
  - %sql select distinct "Launch_Site" from SPACEXTABLE
- Result:
  - "Distinct" keyword fetches unique values from a column

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

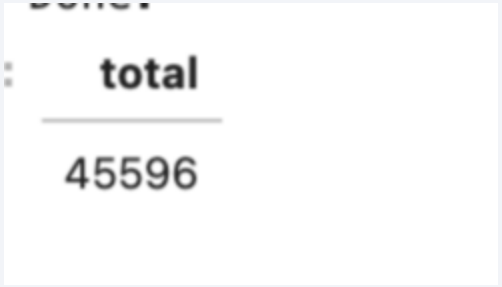
Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
  - %sql select * from "SPACEXTABLE" where "Launch_Site" like "CCA%" limit 5
- Result: Like key word does a string match

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - %sql select sum("PAYLOAD_MASS__KG_") as total from "SPACEXTABLE" where "Customer" = "NASA (CRS)"
- Result:
  - Sum function in SQL does a total of values of all records for a column



total
45596

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_") as total from "SPACEXTABLE" where "Booster_Version" = "F9 v1.1"
  -
- Result:
  - AVG function does the average of values of a column in a table

]:

total
2928.4

First Successful Ground Landing Date

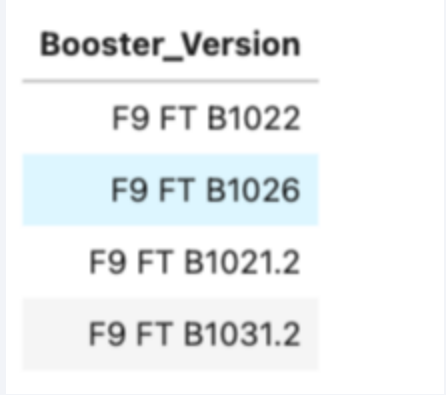
- Find the dates of the first successful landing outcome on ground pad
- **%sql** select min(Date) from "SPACEXTABLE" where "Landing_Outcome" = "Success (ground pad)"
- Result:
  - Min function does calculate minimum value of a column in a table

min(Date)
2015-12-22

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 Booster_Version from "SPACEXTABLE" where "Landing_Outcome" = "Success (drone ship)" and PAYLOAD_MASS__KG_ between "4000" and "6000"
-

- Result:
  - Simpe SQL query with where complex where clause



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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
  - **%sql** select "Mission_Outcome" , count(*) from "SPACEXTABLE" group by "Mission_Outcome"
  -
- Result:
  - Group by clause count is used here

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
  - **%sql** select Booster_Version, "PAYLOAD_MASS__KG_" from (select * from "SPACEXTABLE" order by "PAYLOAD_MASS__KG_" Desc limit 1)
  -
- Result:
  - Use of a Nested query

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - **%sql** select substr(Date, 6, 2) as month, "Landing_Outcome", "Booster_Version", "Launch_Site" from "SPACEXTABLE" where "Landing_Outcome" = "Failure (drone ship)" and substr(Date,0,5)='2015'
  -
- Result:
  - Substr() function is used to offset the year as month

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

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
  - **%sql** select "Landing_Outcome" , count(*) as ranking from "SPACEXTABLE" where "Date" between "2010-06-04" and "2017-03-20" group by "Landing_Outcome" order by ranking desc
  -
- Result:
  - Order by clause is used the sorting ascending or descending.
  - Between clause is used for determine a range

Landing_Outcome	ranking
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

All Launch Sites in map

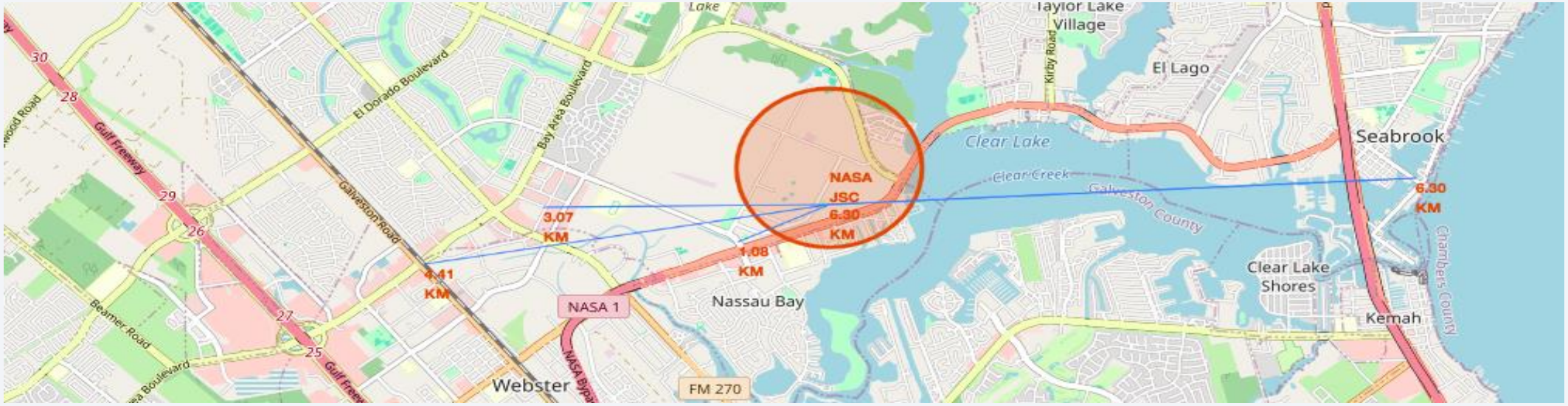
- Three sites are very adjacent to each other except VAFB SLC-4E.
- Are all launch sites in proximity to the Equator line?
- Are all launch sites in very close proximity to the coast?

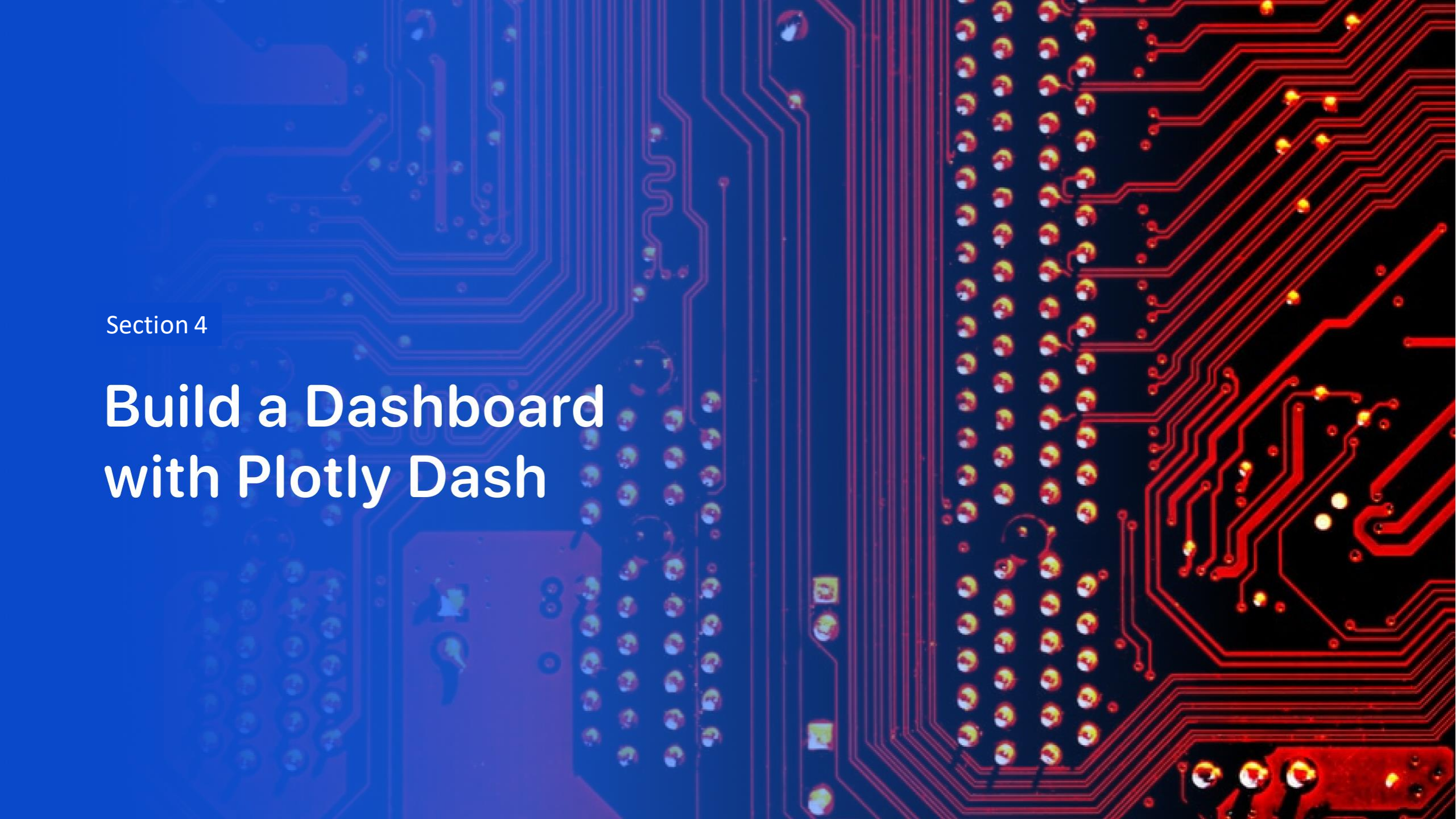
Launch Outcome for Sites

- We can see Site VAFB SLC-4E has only 4 success out of total of 10 launces.
- Others ae sites ae showing a in clusters upon zooming of same we can

Distance between Launch site to its proximities

- NASA JSC Site is 4.41km away from Railway
- 6.30km away from Coastline
- 3.07km away from Cities
- 1.08km away from Highway

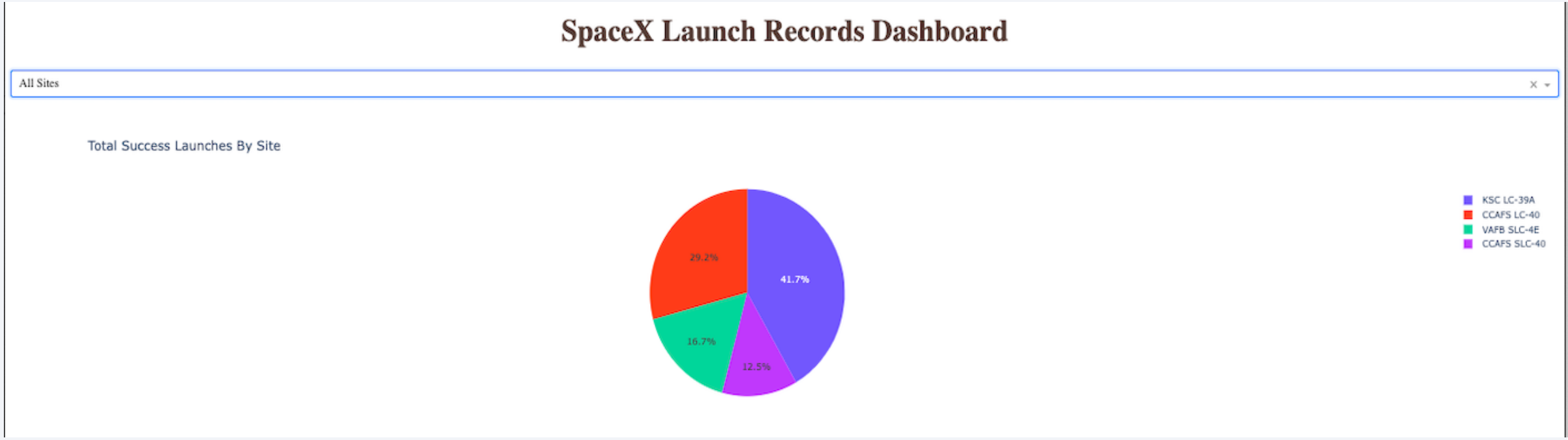



Section 4
Build a Dashboard with Plotly Dash

All Launch Sites Dash board

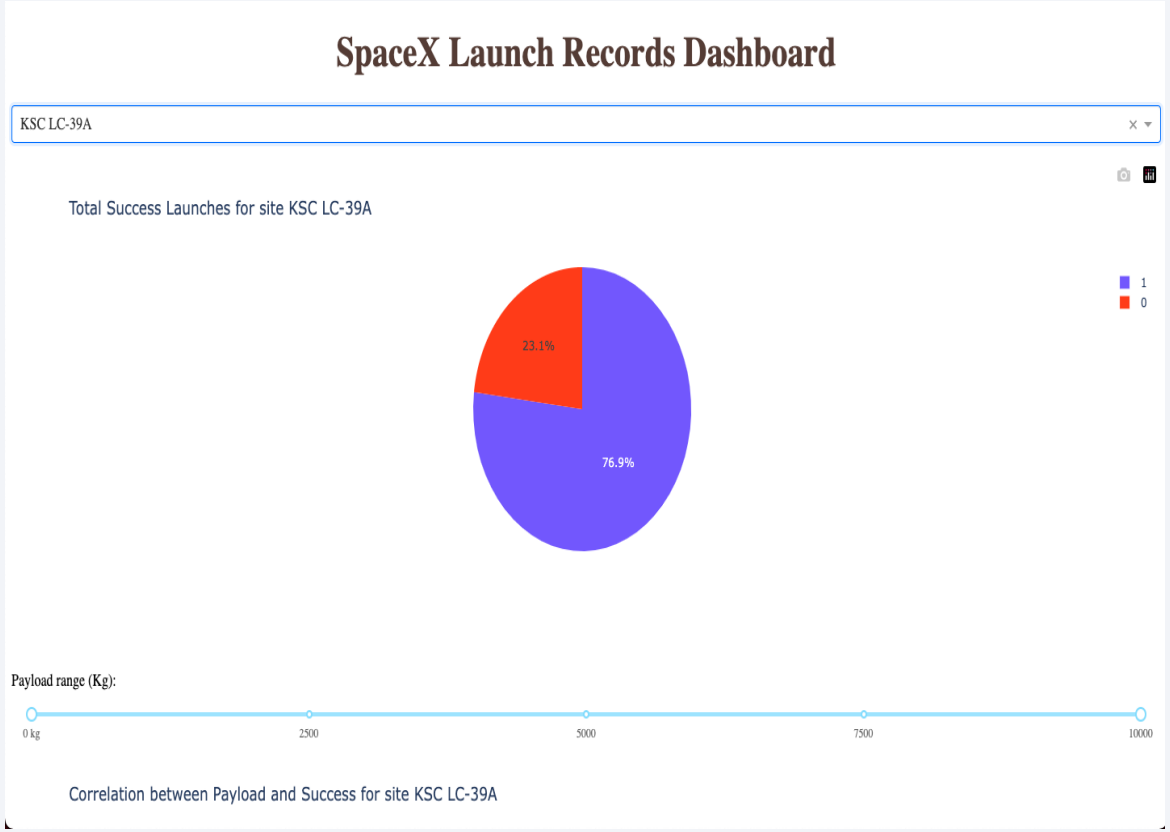
Explain the important elements and findings on the screenshot

- KSC LC-39 launch site has highest (41%) success rate and others site also we can see their respective success percentages



Highest Success Launch Site Dash Board

- KSC LC 39-A site has most success rate among all other sites.
- We see it has 76.9% successful



Correlation Dash board

- This das board shows for every payload range, what is the success rate for each or all sites and what are the booster version were contributed to the same.

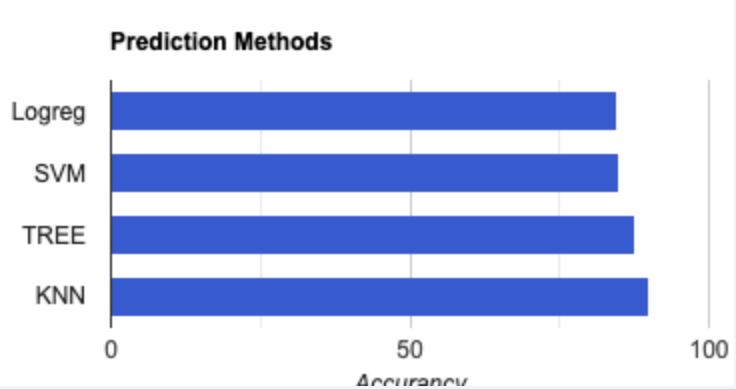
Section 5

Predictive Analysis (Classification)

Classification Accuracy

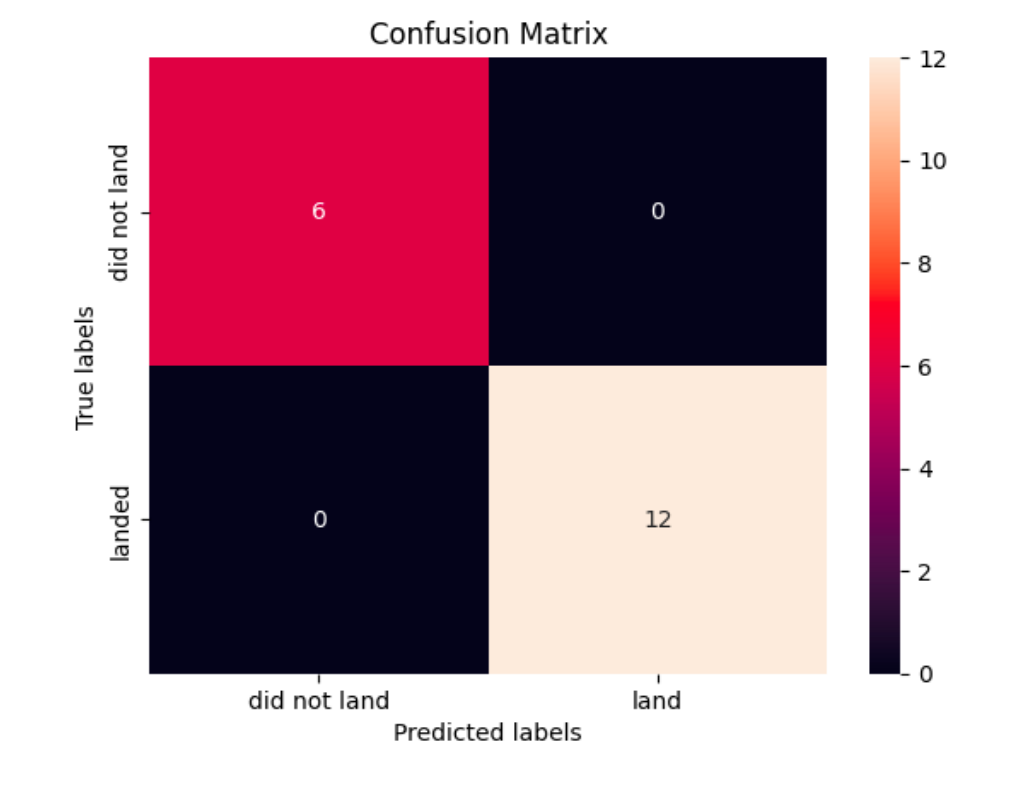
- We can visualize the accuracy comparison in bar chart from all methods.

- KNN model has highest accuracy



Confusion Matrix

- KNN or KNeighborsClassifier has the highest score accuracy



Conclusions

- As we review Predictive Analysis for Falcon 9 landings
- Trained the model and perform Grid Search using the best hyperparameter values,
- Determining the model with the best accuracy using the training data.
- Decision Tree Classifier, and K-nearest neighbors and finally, we will output the confusion matrix.

Appendix

- SPACE X Api used
  - https://api.spacexdata.com/v4/rockets/
  - https://api.spacexdata.com/v4/launchpads/
  - https://api.spacexdata.com/v4/payloads/
  - https://api.spacexdata.com/v4/cores/

- Datasets used for Data Wrangling
  - https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csvhttps://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv

Thank you!