IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aqil Helmi Wan Nordin
9th December 2025

Outline

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

Executive Summary

- Data was collected through SpaceX REST API and web scraping through a Wikipedia page titled as “List of Falcon 9 and Falcon Heavy Launches”.
- All the classification model has the same classification accuracy on the test data but Classification Tree has the highest accuracy on the train data.
- Launch site KSC LC-39A has the highest launch success ratio.
- KSC LC-39A has the highest number of successful launch compared to other launch sites.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Goal is to determine whether the first stage will land or not, in order to determine the cost of the launch.
- Also to understand which factor and variables plays the major role in making the first stage land.

Section 1

Methodology

Methodology

Executive Summary

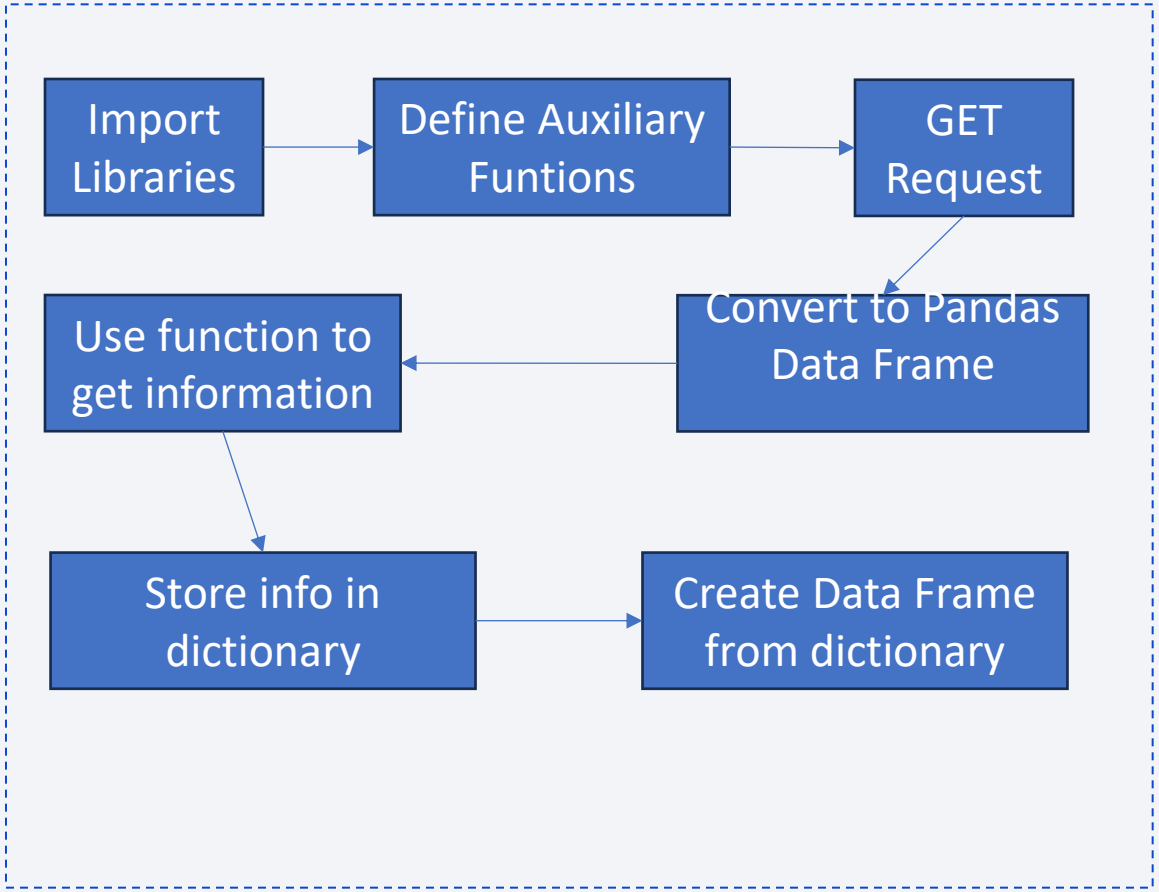
- Data collection methodology:
  - SpaceX REST API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Identify missing values, labelling the predictor and target variable
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data into training and testing set, create GridSearchCV object, calculate test accuracy

Data Collection

- Data collected through SpaceX REST Application Programming Interface (API) and Web Scraping from Wikipedia page titled “List of Falcon 9 and Falcon Heavy Launches.
- The link of the Wikipedia web page is as follow: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

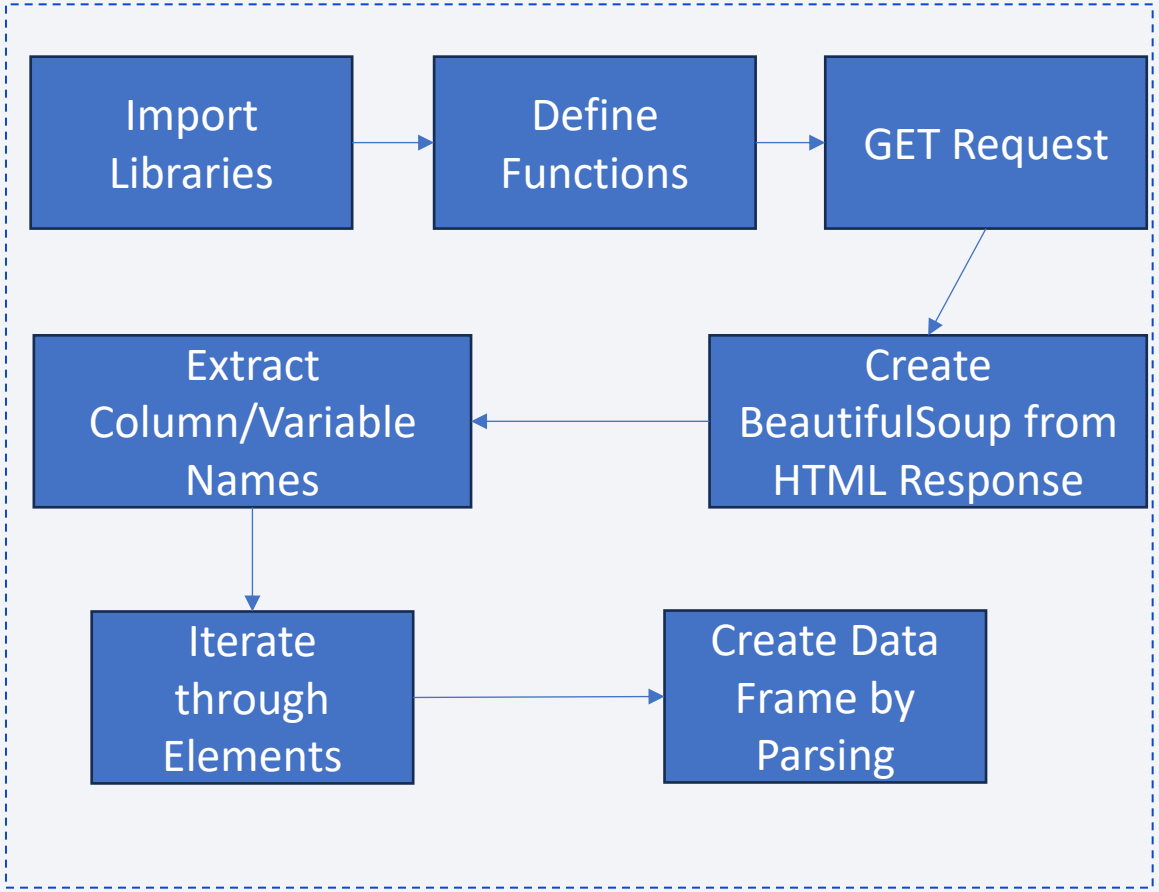
Data Collection – SpaceX API

- GitHub URL: https://github.com/Qiloteh/IbmDataScienceCapstoneProject-SpaceX/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- GitHub URL: https://github.com/Qiloteh/IbmDataScienceCapstoneProject-SpaceX/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Identify missing values -> Identify categorical and numerical columns -> Calculate number of launches on each site -> Calculate number and occurrence of each orbit -> Calculate number and occurrence of mission outcome of the orbits -> Labelling the Outcome 0 as did not land and 1 as landed -> Calculate the success rate of landing
- GitHub URL: https://github.com/Qiloteh/IbmDataScienceCapstoneProject-SpaceX/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Most of the charts are scatter plots. This is because scatter plots are suitable in comparing two numerical variables. Also, I used hue="Class" so that it can compare 3 variables at once. This is crucial to see the underlying patterns between landed successfully and did not land.
- There is also a bar chart I used to plot Success Rate by Orbit. This is because orbit is a categorical variable. Hence, bar chart is the most suitable visualization tool.
- GitHub URL: https://github.com/Qiloteh/IbmDataScienceCapstoneProject-SpaceX/blob/main/edadataviz.ipynb

EDA with SQL

- Queried the name of the launch sites.
- Queried 5 records where the launch sites starts with “CCA”.
- Displayed average payload mass carried by booster version F9 v1.1.
- Listed the date when the first succesful landing outcome in ground pad was achieved.
- Listed the total number of successful and failure mission outcomes.
- GitHub URL: https://github.com/Qiloteh/IbmDataScienceCapstoneProject-SpaceX/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

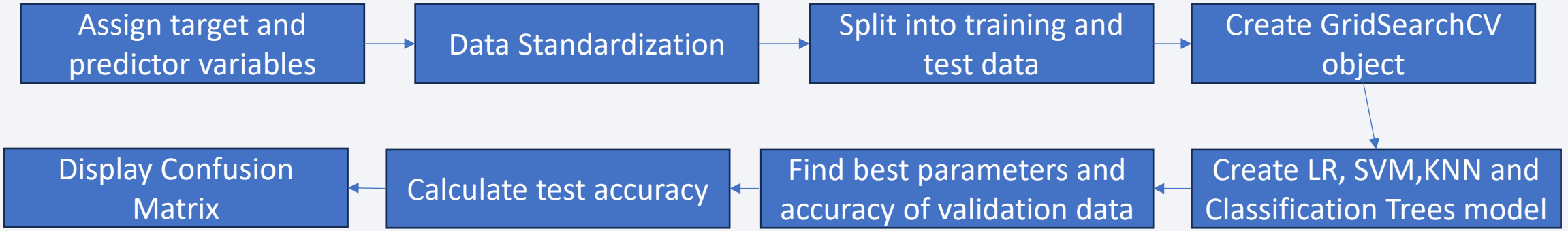
- Markers, circles and polyline were added to the Folium map.
- Markers were placed at specific coordinate points to highlight exact locations of interest such as cities, landmarks, or data points.
- Circles were added to represent areas around a point, useful when showing coverage zones, buffer regions, or when visualizing quantities using radius size.
- Polyline were drawn between coordinates to represent paths, routes, boundaries, or connections between locations.
- GitHub URL: https://github.com/Qiloteh/IbmDataScienceCapstoneProject-SpaceX/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie Chart – Gives a quick high-level overview of SpaceX launch performance, effective in showing proportions such as success rates.
- Scatter Plot – Helps understand the correlation between payload and success, shows how different booster versions perform under different payload weights.
- Dropdown Menu – Enables interactive filtering of the charts to select either All or a specific launch site.
- Payload Range Slider – Lets user a range of payload weights.
- GitHub URL: https://github.com/Qiloteh/IbmDataScienceCapstoneProject-SpaceX/blob/main/dashboard.py

Predictive Analysis (Classification)

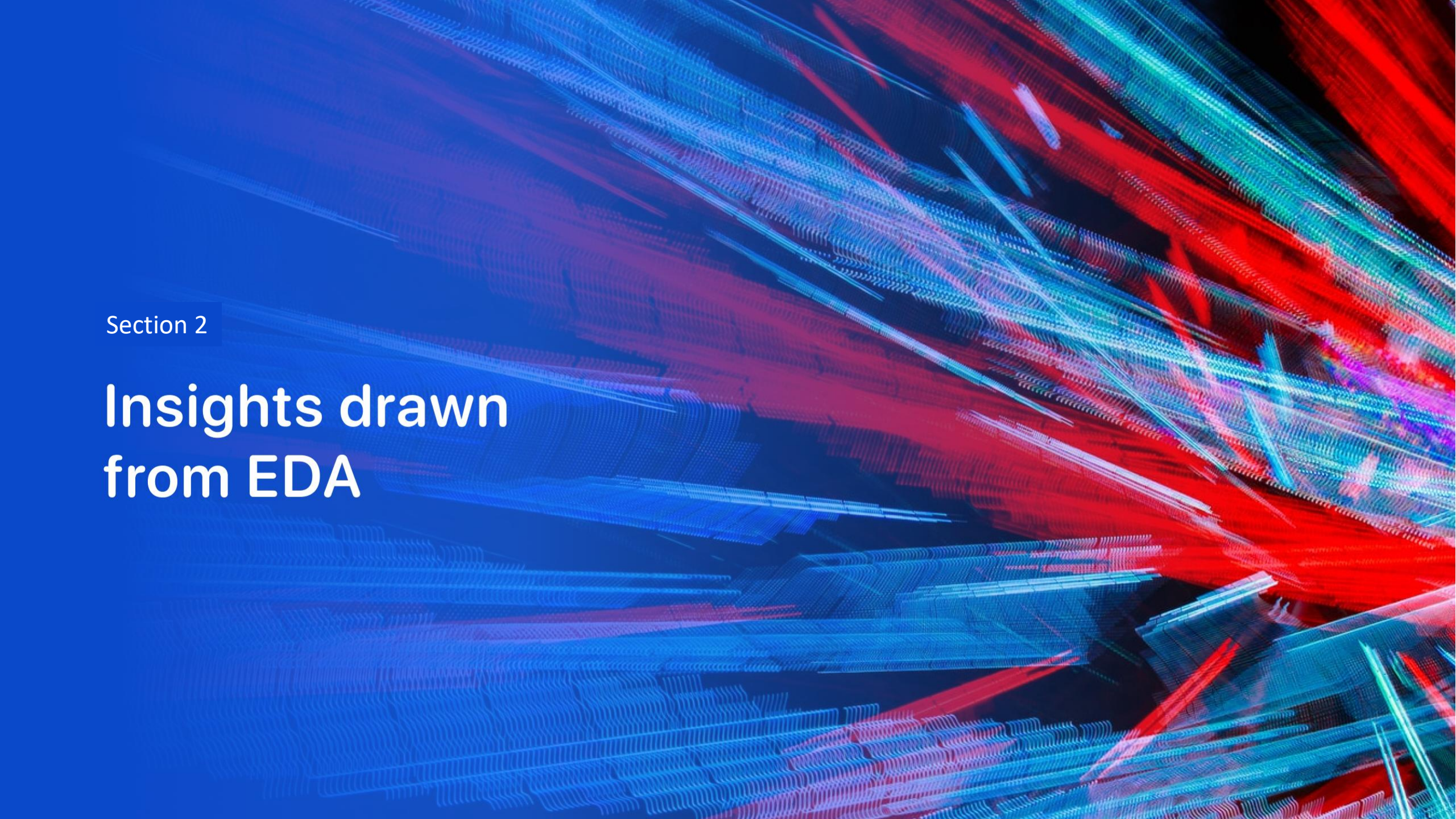
- Linear Regression, SVM, KNN and Classification Trees were built to perform predictive analysis.
- Flowchart:



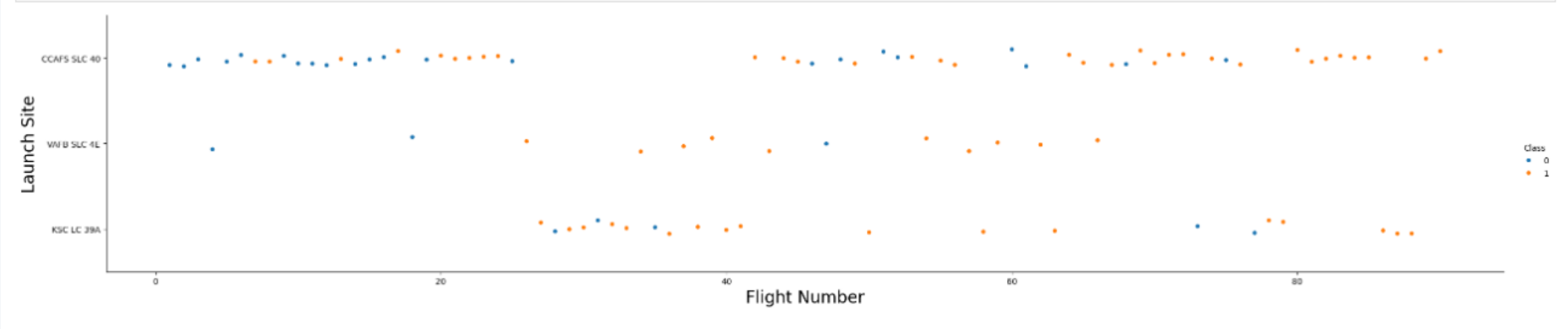
- GitHub URL: https://github.com/Qiloteh/IbmDataScienceCapstoneProject-SpaceX/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Orbit ES-L1, GEO, HEO and SSO has the highest success rate which is 100% while orbit SO has the lowest which is 0%.
- Success rate since 2013 kept increasing till 2020.
- Average payload mass carried by booster version F9 v1.1 is 2928.4KG.
- Maximum payload mass carried by any booster is 15600KG.
- All the classification model has the same classification accuracy on the test data but Classification Tree has the highest accuracy on the train data.

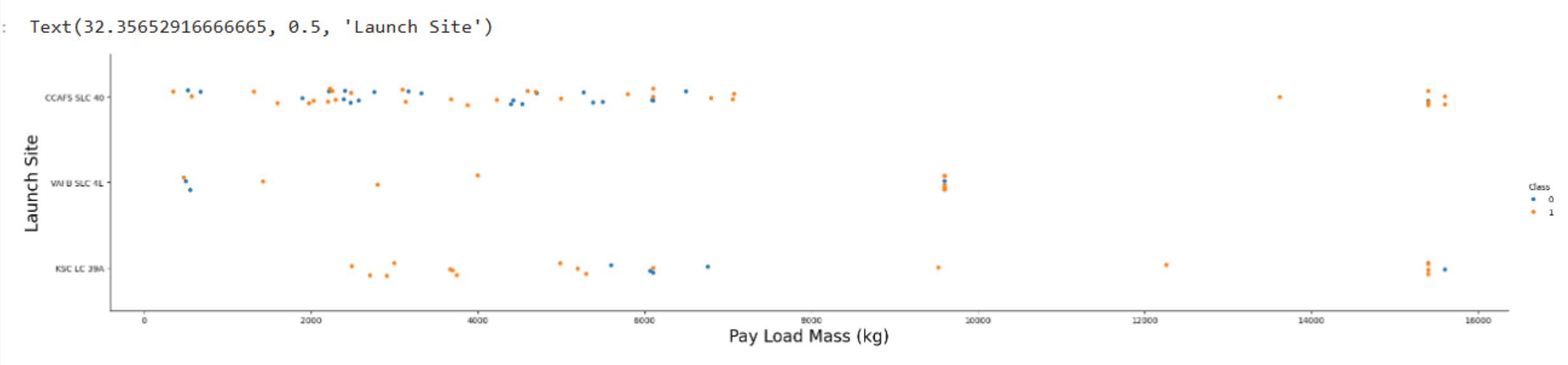

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Higher flight number in CCAFS SLC 40 has the highest rate of successful landing.

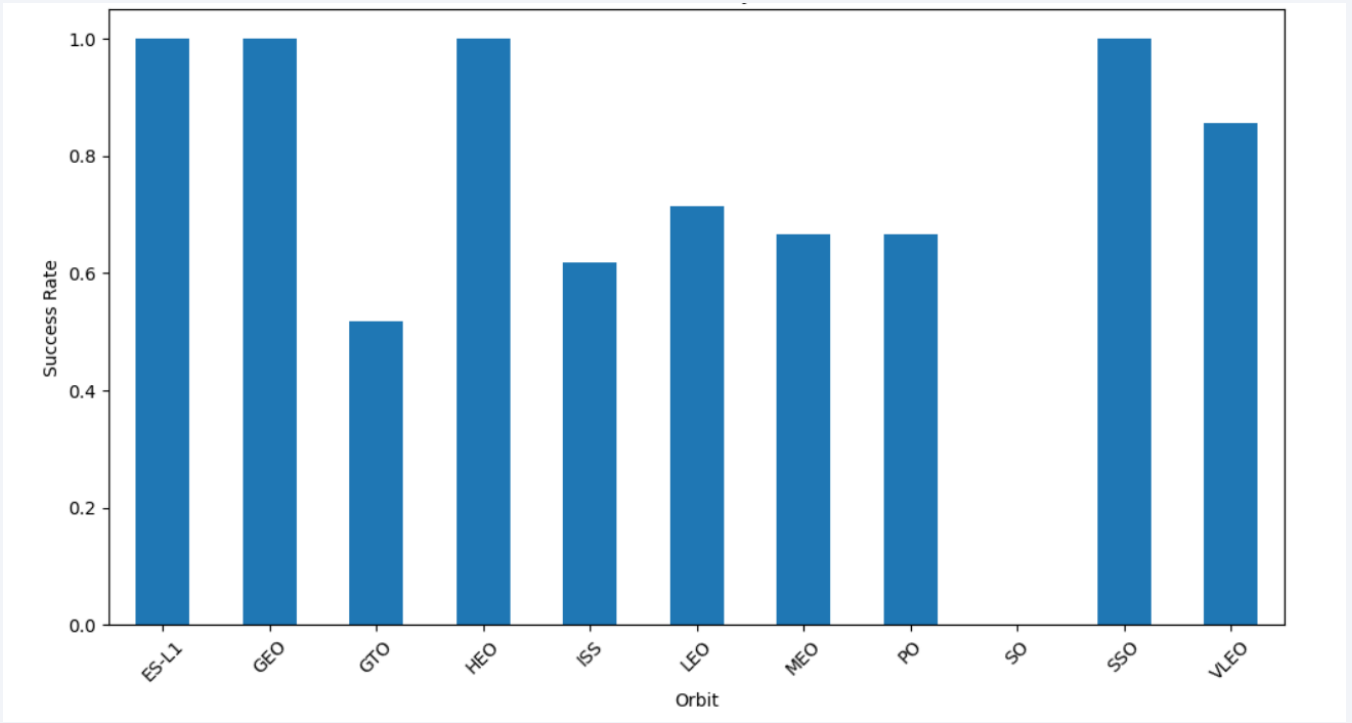
Payload vs. Launch Site



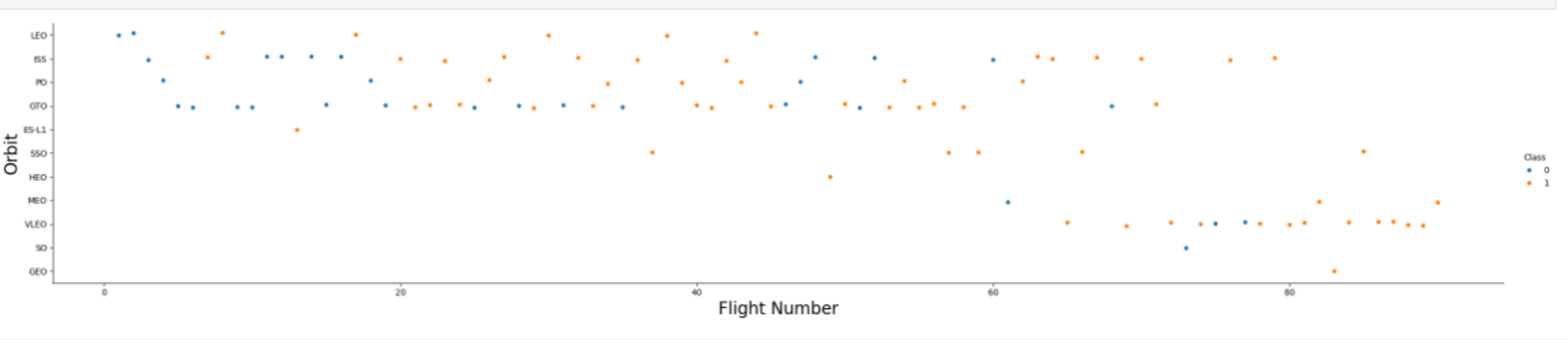
- Higher successful landing rate in KSC LC 39A with a pay load mass ranging from 2000kg to 6000kg

Success Rate vs. Orbit Type

- Orbit ES-L1, GEO, HEO and SSO has the highest success rate which is 100% while orbit SO has the lowest which is 0%.

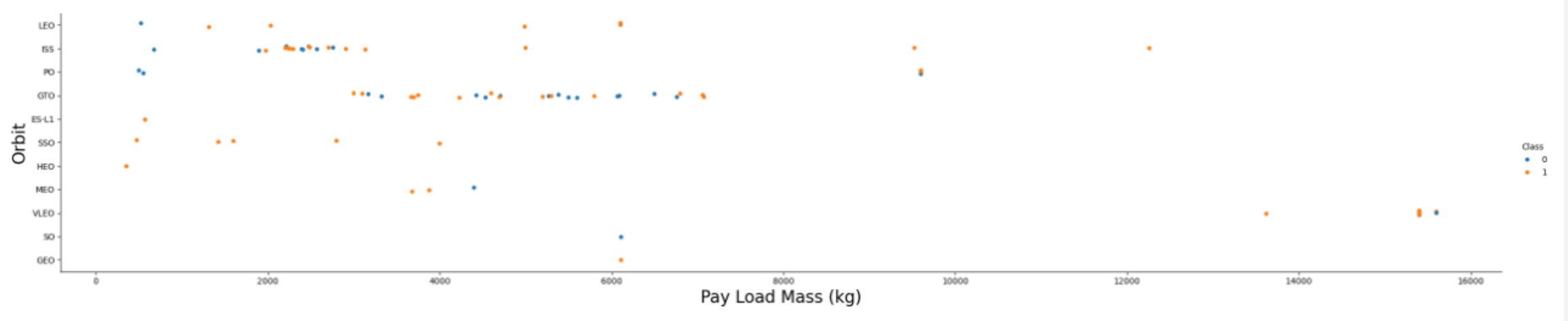


Flight Number vs. Orbit Type



- In the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type

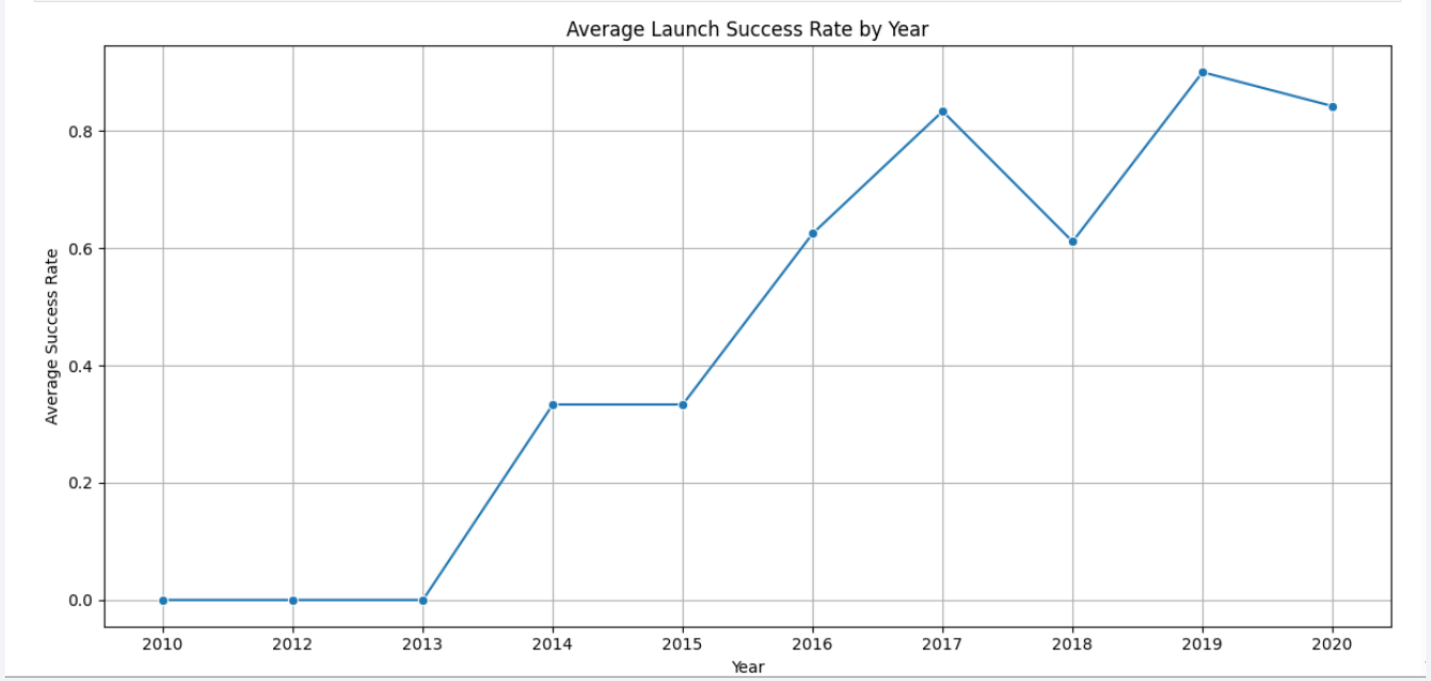


- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- Success rate since 2013 kept increasing till 2020



All Launch Site Names

- The name of the launch sites are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40

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

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The records shown are from the 5 Launch Site names that begin with ‘CCA’

Total Payload Mass

- The Total Payload Mass is 45,596KG

Done.

TotalPayloadMass
45596

Average Payload Mass by F9 v1.1

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

AveragePayloadMass
2928.4

First Successful Ground Landing Date

- The first successful ground landing date is on 22nd December 2015

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2 are boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

- The total successful mission is 100 missions.

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

Boosters Carried Maximum Payload

- Attached are the screenshot of booster version carrying maximum payload amounting to 15600KG

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

2015 Launch Records

- These are record of the failed landing outcomes in year 2015 together with their booster versions and launch site names

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

- These are the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order.
- No attempt landing outcome has the highest count which is 10

Landing_Outcome	OutcomeCount
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

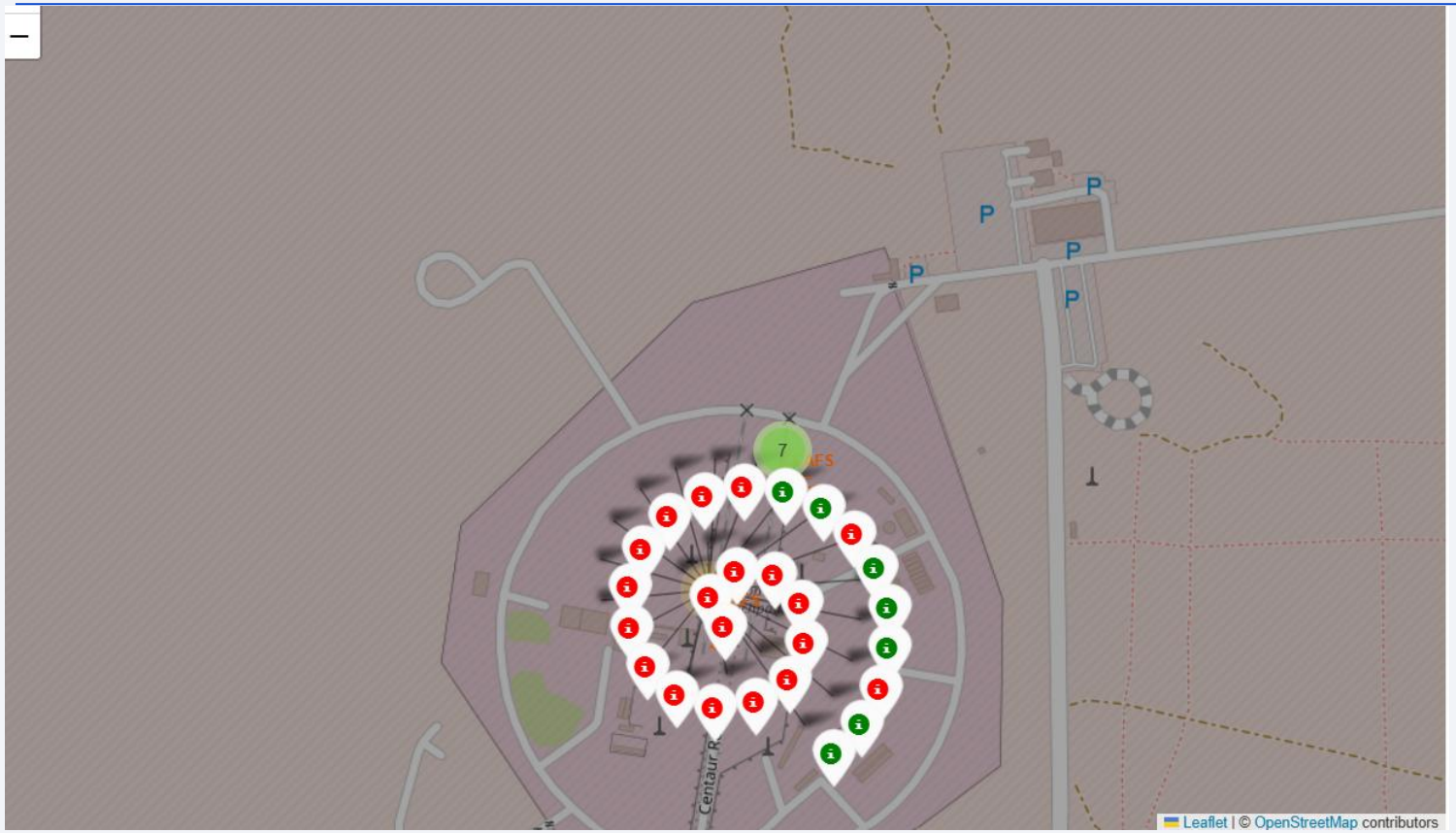
Section 3

Launch Sites Proximities Analysis

Launch Site Locations Markers

- The map shows markers with each of the launch site locations.

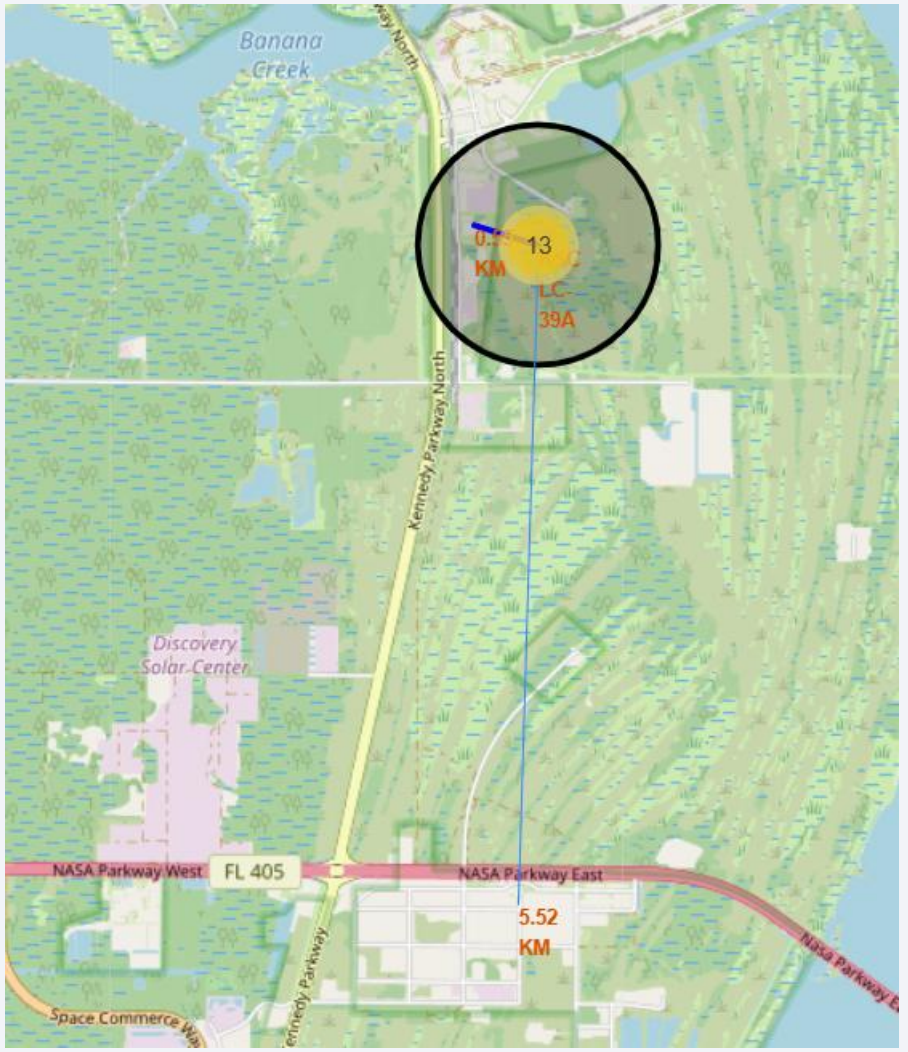
Color Labeled Launch Site Markers

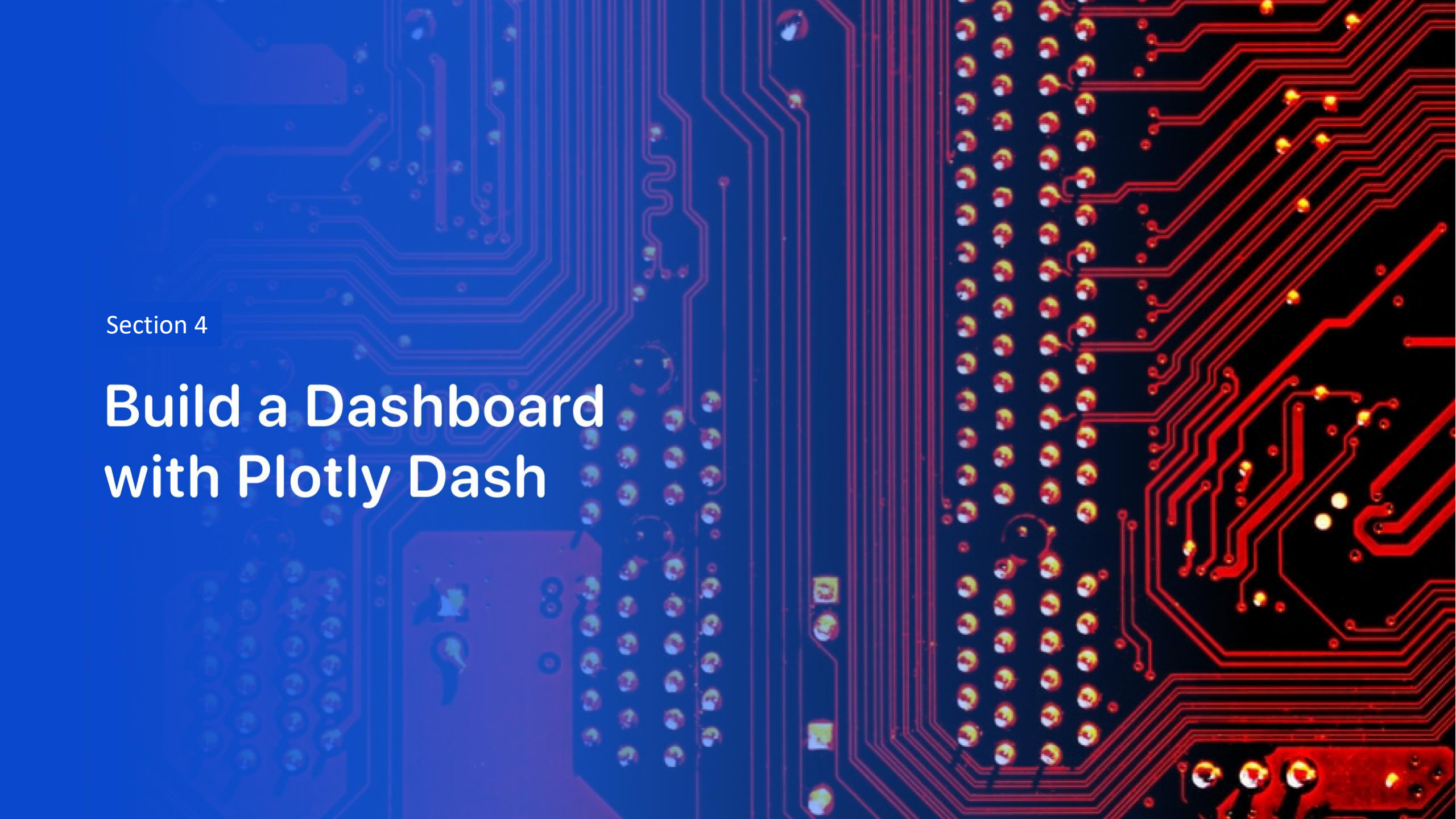


- The color label marker indicates the successful and failed launch outcome. Red means fail, green means successful.

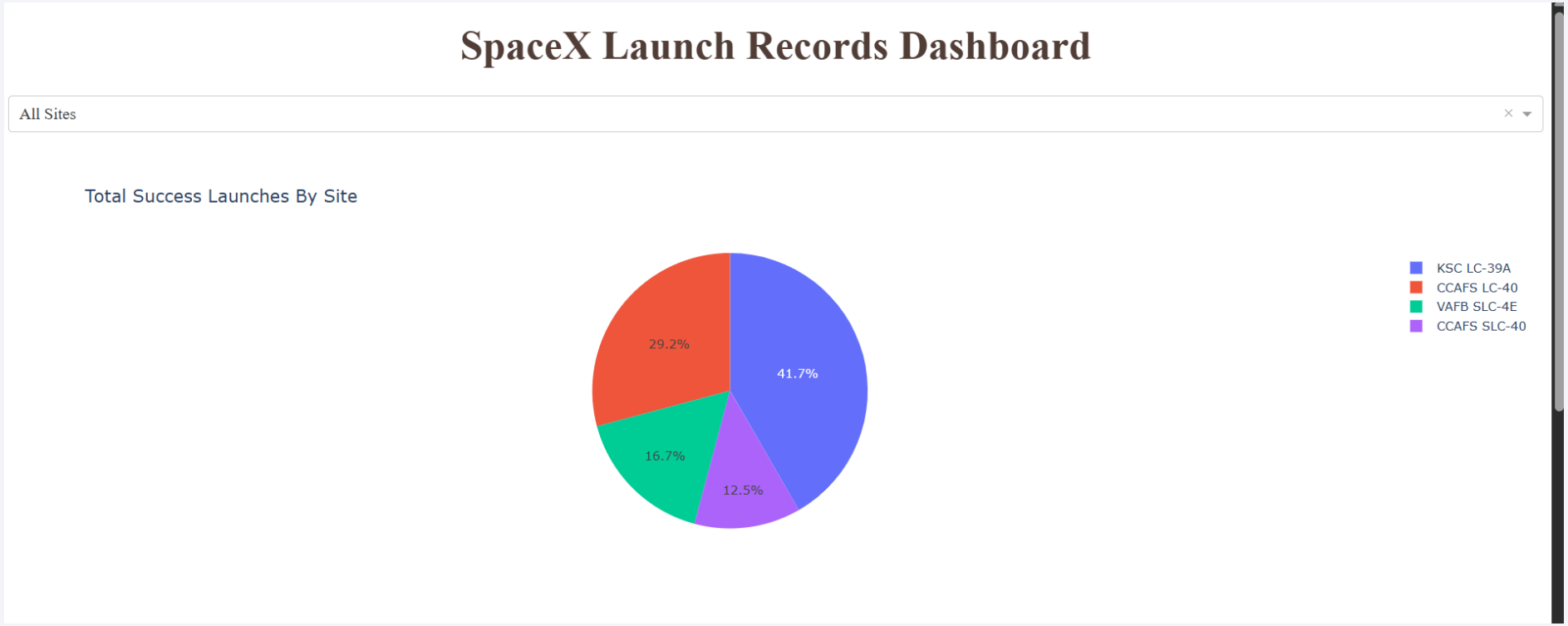
Launch Site Close Proximities

- The polyline indicates the distance of the launch site towards nearest facilities.
- This is helpful to measure any consequences towards the nearest facility due to SpaceX rocket launches.



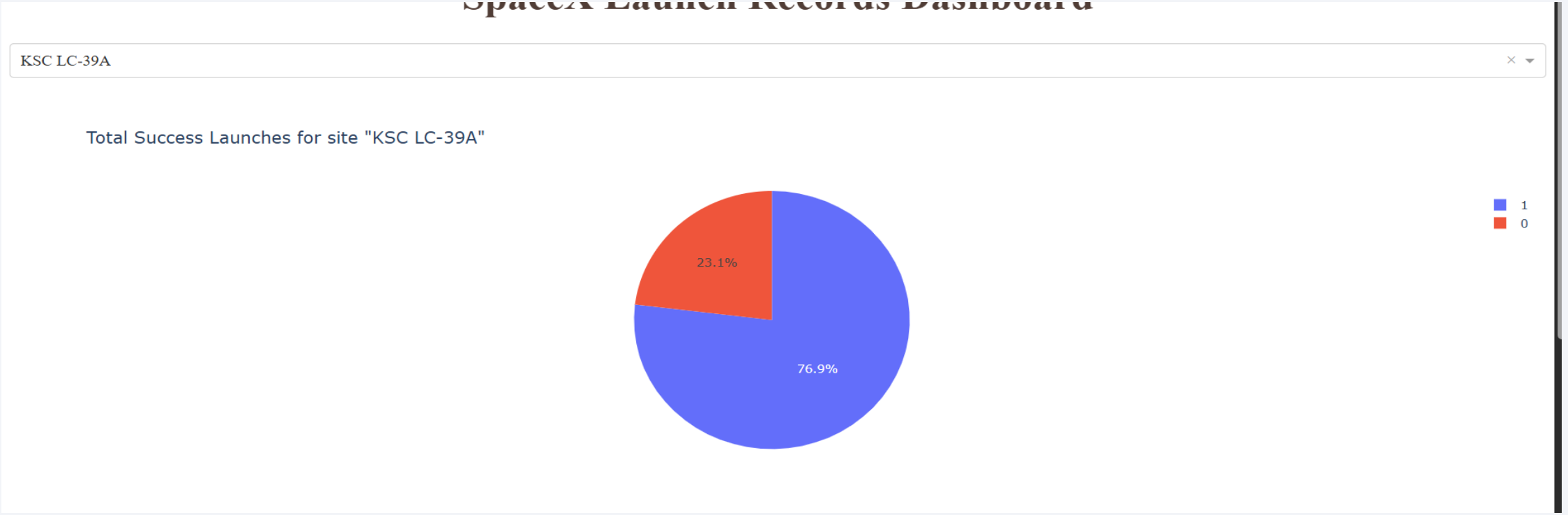

Section 4
Build a Dashboard with Plotly Dash

Successful Launch Count for All Site



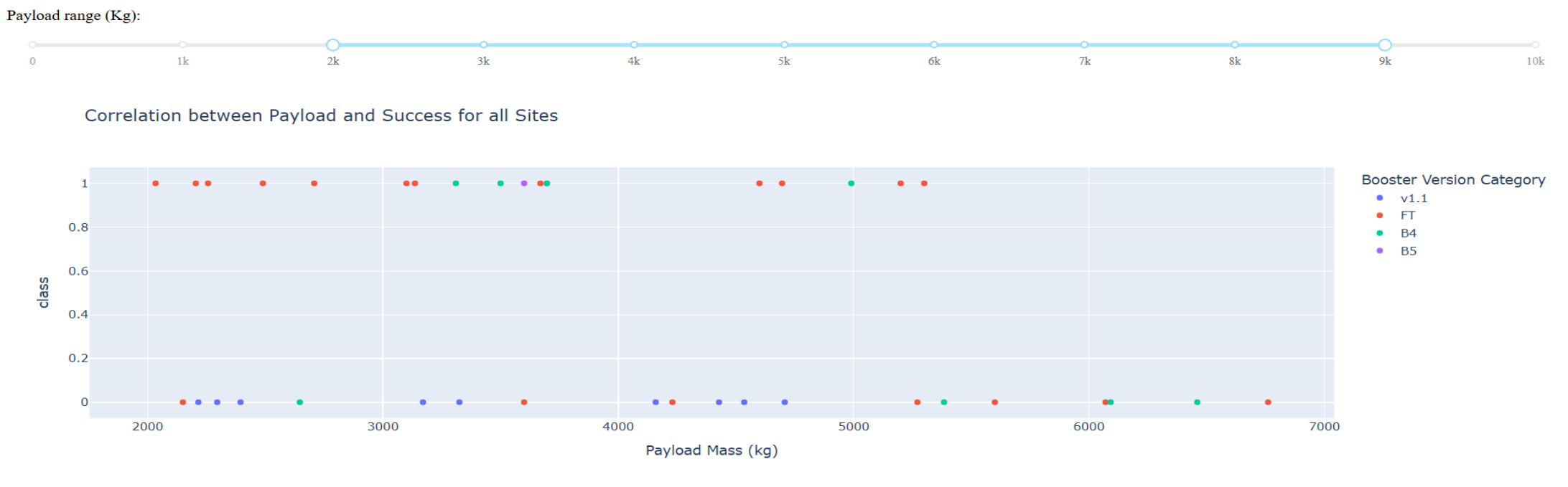
- The pie chart shows proportion of successful launches in each site.

Highest Launch Success Ratio



- Using the dropdown feature, we can choose each site and see that KSC LC-39A has the highest launch success ratio.

Payload vs Launch Outcome



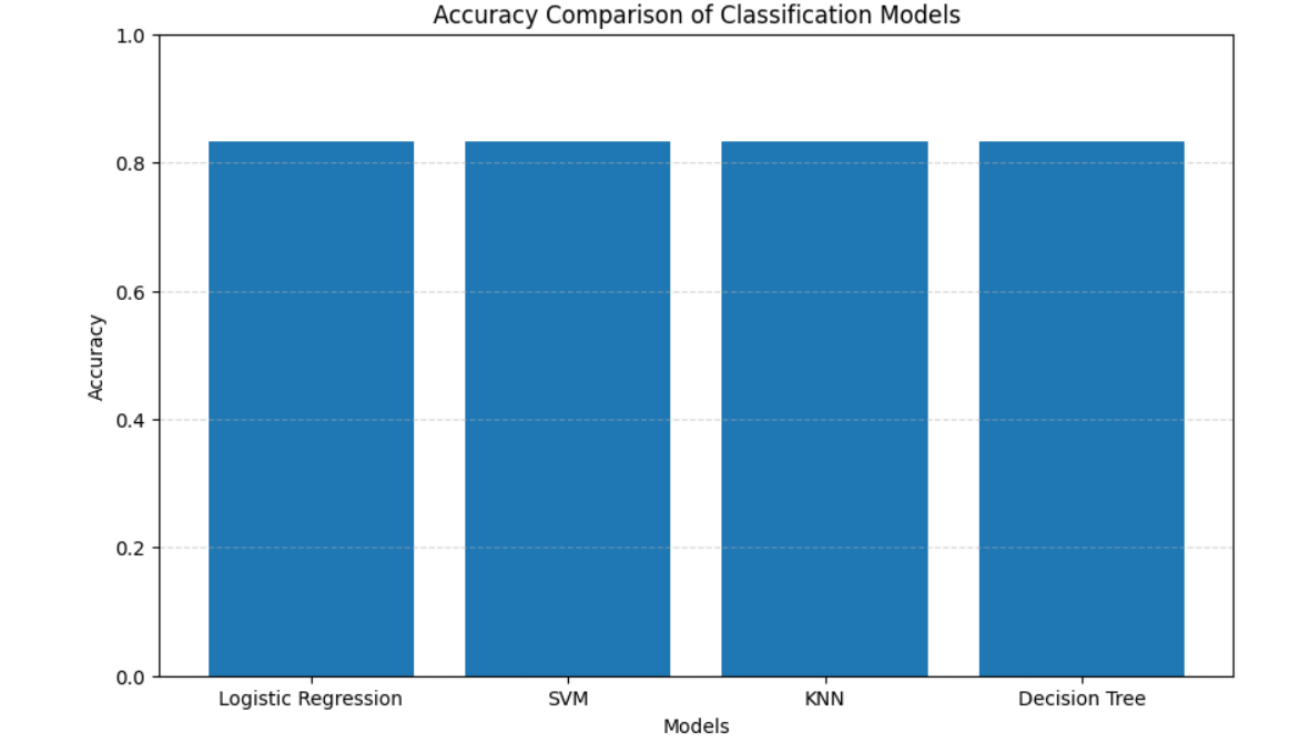
- The range slider enables us to select a range of payload and see the interactivity on how payload affects success through the scatter plot.

Section 5

Predictive Analysis (Classification)

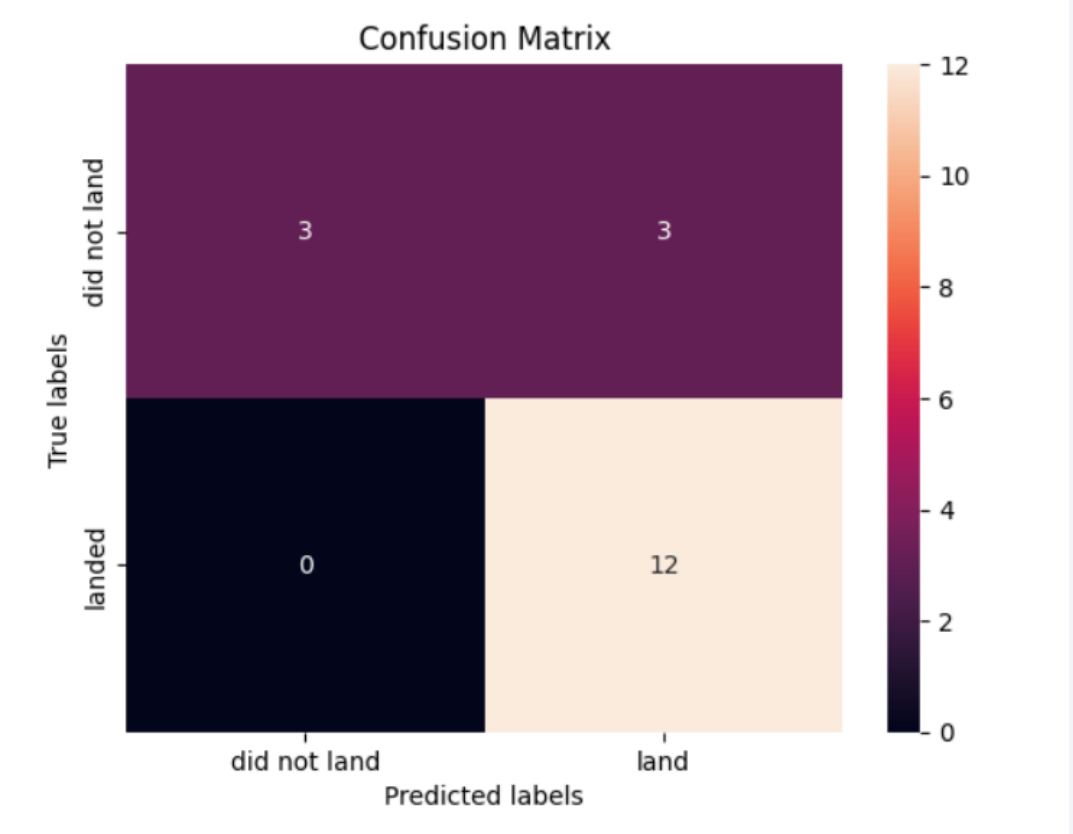
Classification Accuracy

- All the model has the same classification accuracy on the test data but Classification Tree has the highest accuracy on the train data.



Confusion Matrix

- True Postive - 12 (True label is landed, Predicted label is also landed)
- False Postive - 3 (True label is not landed, Predicted label is landed)



Conclusions

- All the model has the same classification accuracy on the test data but Classification Tree has the highest accuracy on the train data.
- Launch site KSC LC-39A has the highest launch success ratio.
- KSC LC-39A has the highest number of successful launch compared to other launch sites.
- Payload range between 3000-5000kg has the highest launch success rate.
- FT F9 booster version has the highest success launch rate.

Appendix

Create a logistic regression object then create a GridSearchCV object `logreg_cv` with cv = 10. Fit the object to find the best parameters from the dictionary `parameters`.

In [12]:
```
parameters ={'C':[0.01,0.1,1],
             'penalty':['l2'],
             'solver':['lbfgs']}
```

In [13]:
```
parameters ={"C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs']}# l1 lasso l2 ridge
lr=LogisticRegression()
logreg_cv=GridSearchCV(lr,parameters,cv=10)
logreg_cv.fit(X_train,Y_train)
```

Out[13]:
```
GridSearchCV(cv=10, estimator=LogisticRegression(),
             param_grid={'C': [0.01, 0.1, 1], 'penalty': ['l2'],
                         'solver': ['lbfgs']})
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

We output the `GridSearchCV` object for logistic regression. We display the best parameters using the data attribute `best_params_` and the accuracy on the validation data using the data attribute `best_score_`.

In [14]:
```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

- Linear Regression model

Thank you!