



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

KHIN ZAR NI HTUN
04 DEC 2023



TABLE OF CONTENTS

SPACE Y



- Executive Summary
- Introduction
- Methodology
- Results
 - EDA with Visualization
 - EDA with SQL
 - Interactive Maps with Folium
 - Plotly Dash Dashboard
- Discussion
 - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY



EXECUTIVE SUMMARY

- New rocket launch company SpaceY have to compete with SpaceX rocket launch provider.
- SpaceX announced that Falcon9 rocket launch mission is starting at \$62 million cost (while other provider cost upwards of \$165 million) and cost saving is because of SpaceX can reuse first stage.
- Mission parameter like payload mass, orbit, customer, models produced in this report will predict the successful accuracy of 83.33% of the first stage rocket booster landing.
- Train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

INTRODUCTION

- **Background and Business Problem**

- SpaceX's accomplishments include are sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars when other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.
- If we can determine if the first stage will land, we can determine the cost of a launch.
- I will take the role of a data scientist working for a new rocket company, Space Y that would like to compete with SpaceX.



Section 1

Methodology

METHODOLOGY



METHODOLOGY

Outline of using data science methodology in this report are:

- **Data Collection**
 - API
 - Web Scraping
- **Data Wrangling**
- **Exploratory data analysis (EDA)**
 - EDA with SQL
 - EDA with Visualization
- **Data Visualization**
 - Launch Sites Locations Analysis with Folium
 - SpaceX Launch Records Dashboard
- **Model development**
- **Report result for the stakeholder**

METHODOLOGY

- **Data Collection – API**
 - Historical rocket launch data can request from SpaceX API
 - Request and parse the SpaceX launch data using the GET request
 - Filter the data frame to only include Falcon 9 launches
 - Missing values in payload mass column were replace with mean value
- **Data Collection – Web Scraping**
 - Acquired Falcon 9 and Falcon Heavy Launches Records from Wikipedia page
 - Request the Falcon9 Launch Wiki page from its URL
 - Extract all column/variable names from the HTML table header
 - Create a data frame by parsing the launch HTML tables

METHODOLOGY

- **Data Wrangling**

- Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
 - Calculate the number of launches on each site
 - Calculate the number and occurrence of each orbit
 - Calculate the number and occurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome column
 - Training label named 'Class'
 - Value of Class '0' is the first stage did not land successfully
 - Value of Class '1' is the first stage land successfully
 - Landing outcome contents
 - True ASDS : drone ship landing succeeded
 - True RTLS : ground pad landing succeeded
 - True Ocean : ocean landing succeeded
 - None None : not attempted
 - False ASDS : drone ship landing failed
 - False Ocean : ocean landing failed
 - None ASDS : due to launch failure unable to attempted
 - False RTLS : ground pad landing failed

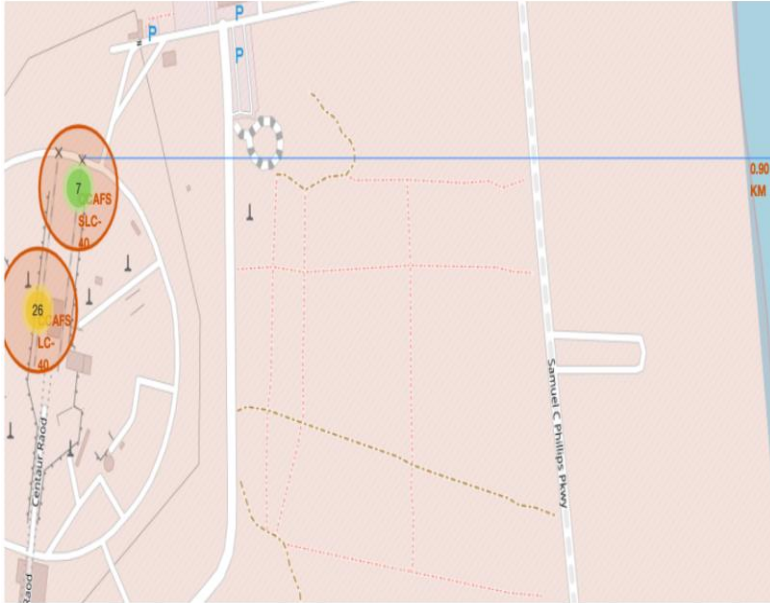
METHODOLOGY

- **Exploratory Data Analysis**
 - EDA with SQL
 - Load the dataset into the corresponding table in a Db2 database
 - SQL queries to display and list on
 - Launch sites
 - Payload mass
 - Booster landing
 - Mission outcome
 - Booster Version

METHODOLOGY

- **Exploratory Data Analysis**
 - EDA with Visualization
 - read the SpaceX dataset into a Pandas dataframe
 - Using panda and seaborn libraries to visualize the data between
 - Flight number and payload mass
 - Flight number and Launch site
 - Payload mass and Launch site
 - Orbit type and success rate
 - Flight number and orbit
 - Payload mass and orbit
 - Launch success yearly trend
 - Year and success rate

METHODOLOGY

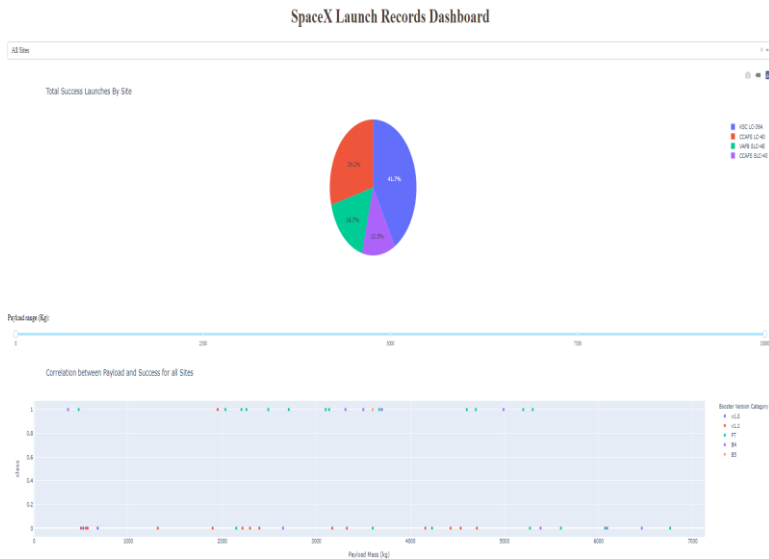


- **Data Visualization**

- Launch Sites Locations Analysis with Folium

- Performing more interactive visual analytics using Folium
 - Mark all launch sites on a map
 - Mark the success/failed launches for each site on the map
 - Calculate the distances between a launch site to its proximities
 - Cities
 - Railway
 - Highway
 - Coastline

METHODOLOGY

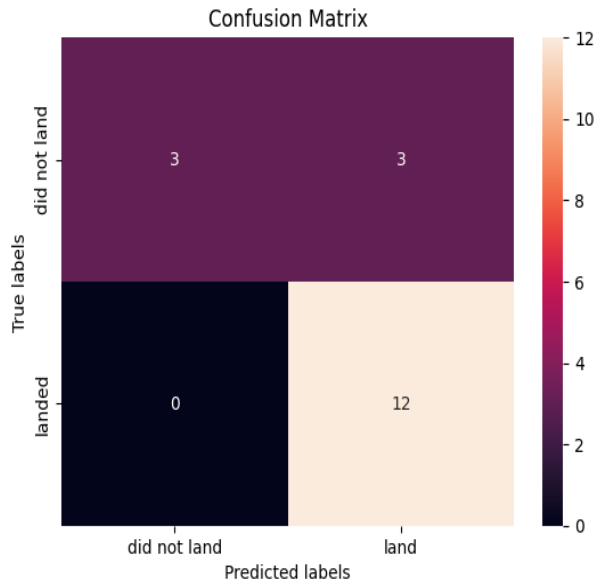


- **Data Visualization**

- SpaceX Launch Records Dashboard

- Pie Chart Showing Successful Launches
 - Slider of Payload Mass Range
 - Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version
 - Dropdown List with Launch Sites

METHODOLOGY



Showing 15 corrected predictions and 3 false positive of confusion matrix by the use of logistic regression model.

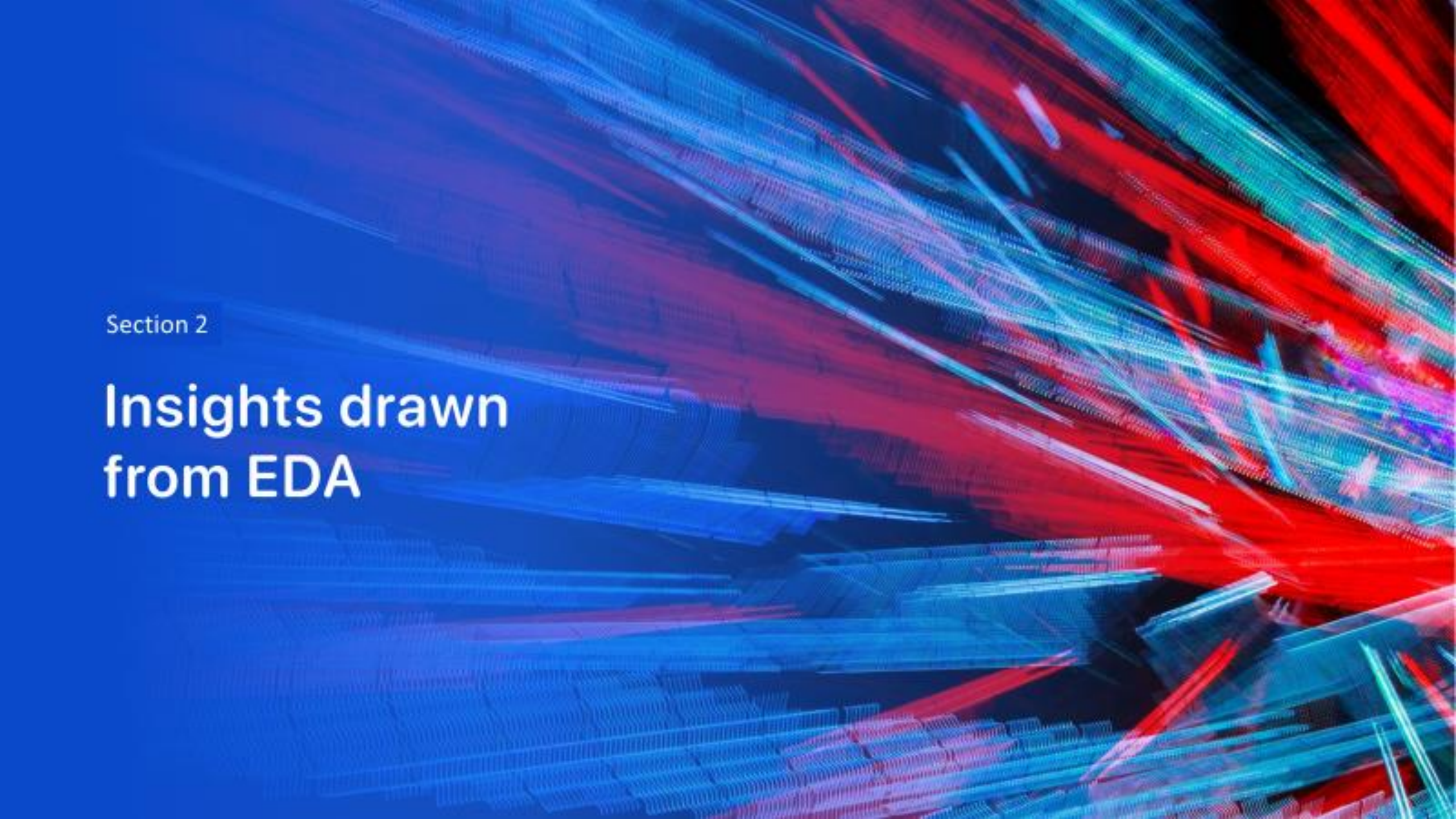
- **Model Development**

- Create a column for training label 'Class'
- Standardize the data with StandardScaler. Fit and transform the data.
- Split the data into training and testing using `train_test_split`
- Create a GridSearchCV object with `cv=10` for parameter optimization
- Fit the training data on different models such as logistic regression , support vector machine, decision tree, K-Nearest Neighbor
- Using test data calculated accuracy of each models to choose for the best
- Assess the confusion matrix for all models

RESULTS SUMMARY



- Exploratory Data Analysis
- Visual Analytics
- Predictive Analytics

The background of the slide is an abstract composition. It features a solid blue area on the left side, which transitions into a dark, almost black, central region. Overlaid on this are numerous bright, diagonal streaks in shades of red and cyan. These streaks vary in thickness and intensity, creating a sense of motion and depth. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant, adding a technical or data-oriented feel to the design.

Section 2

Insights drawn from EDA

RESULTS – EDA WITH SQL

- Launch sites used by SpaceX
 - CCAFS LC-40
 - CCAFS SLC-40
 - KSC LC-39A
 - VAFB SLC-4E
- Total payload mass carried by boosters launched by NASA (CRS)
 - 45596 KG
- Average payload mass carried by booster version F9 v1.1
 - 2928.4 KG

RESULTS – EDA WITH SQL

- List the total number of successful and failure mission outcomes

Out[25]:

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

- List the names of the booster_versions which have carried the maximum payload mass

Out[30]:

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

RESULTS – 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

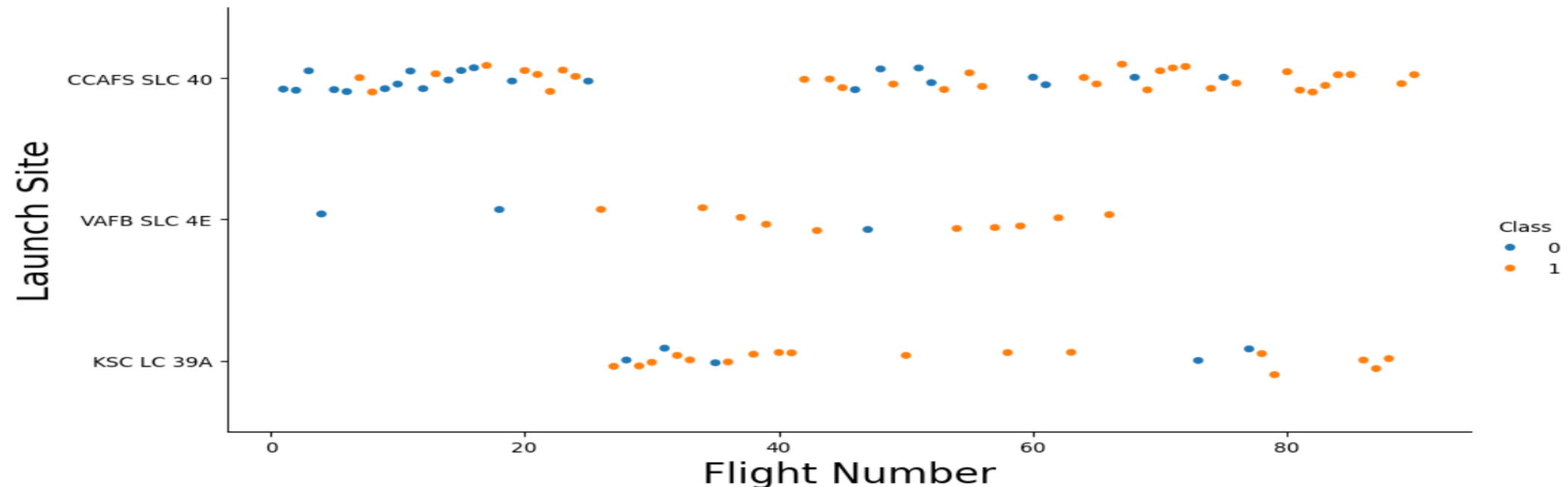
month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

RESULTS – EDA WITH VISUALIZATION

- **Flight Number & Launch Site**

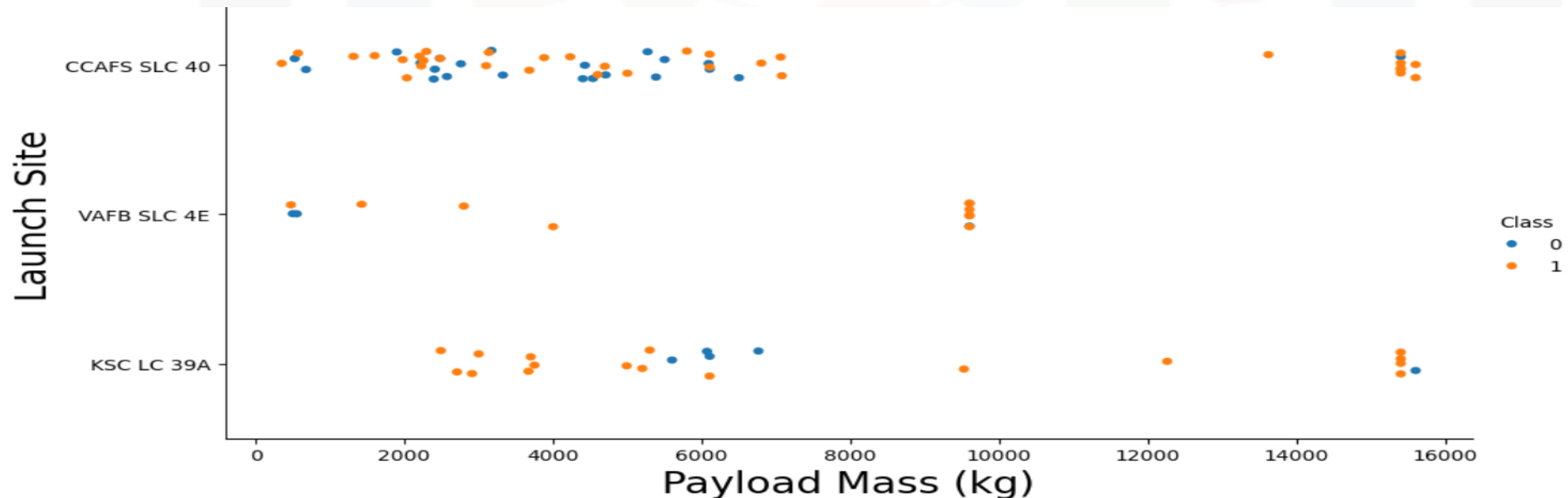
- Lower success rate are found in earlier flights(Class 0 = blue)
- Higher success rate are found in later flights(Class 1 = orange)
- Launch sites VAFB SLC 4E and KSC LC 39A have higher success rate than CCAFS SLC 40 launch site



RESULTS – EDA WITH VISUALIZATION

- **Payload Mass(kg) & Launch Site**

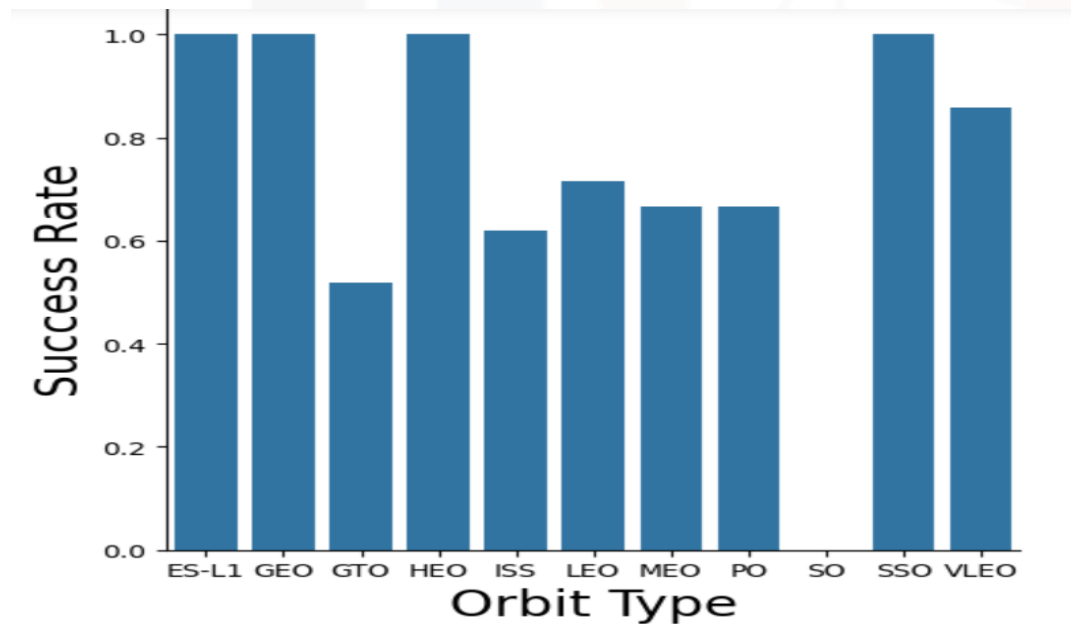
- Less than of payload mass 5000kg for KSC LC 39A are found 100% success rate
- Higher success rate are found in higher payload mass (Class 1= orange)
- Greater than 7000kg payload mass of launch sites are with success rate



RESULTS – EDA WITH VISUALIZATION

- **Success Rate of Each Orbit**

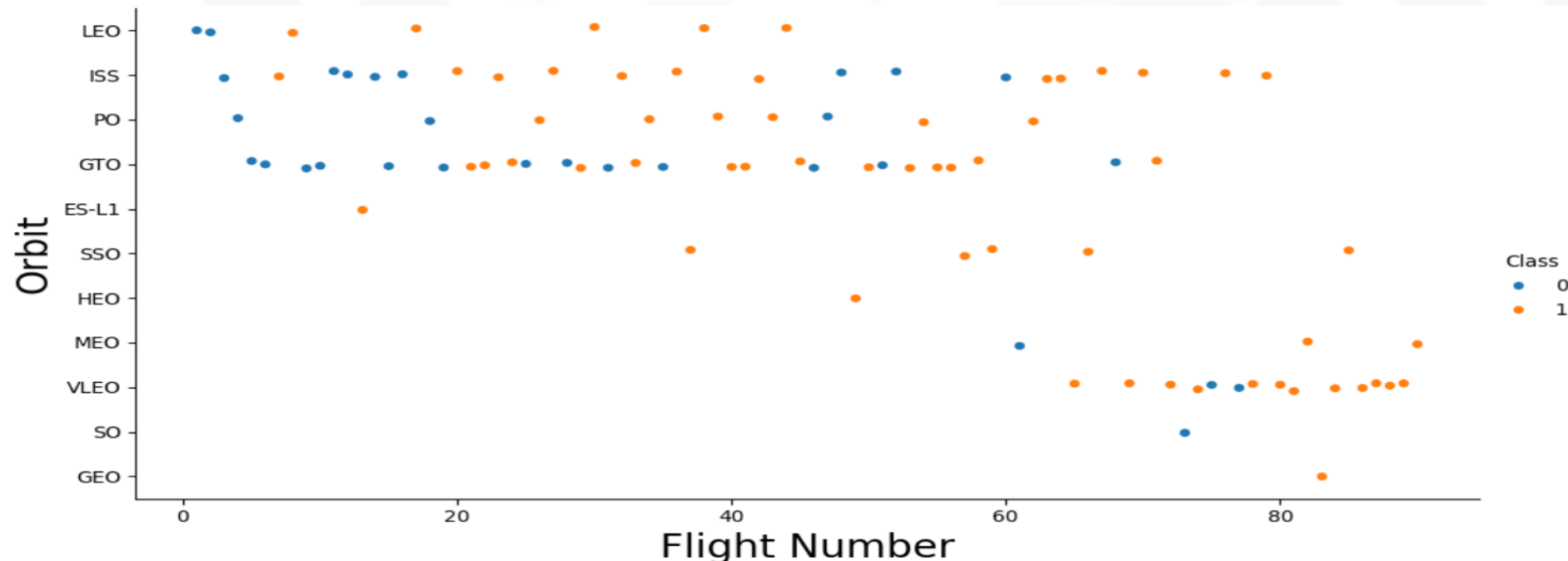
- Orbit SO have 0% success rate
- Orbit GTO, ISS, LEO, MEO, PO, and VLEO had 50% - 80% success rate
- Orbit ES-L1,GEO,HEO,AND SSO had 100% success rate



RESULTS – EDA WITH VISUALIZATION

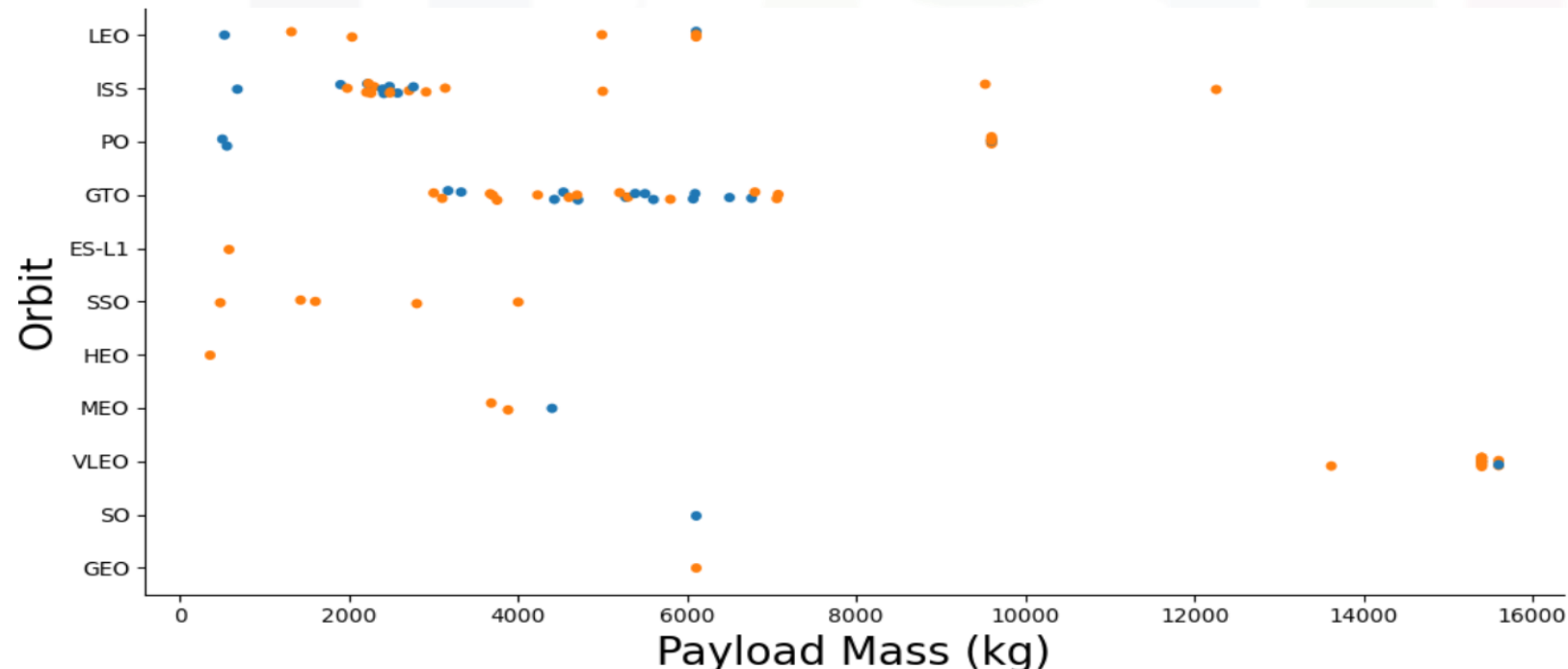
- **Flight Number & Orbit Type**

- All orbit type with higher flight number have success rate and positively correlated
- Orbit GTO is not following in this trend.



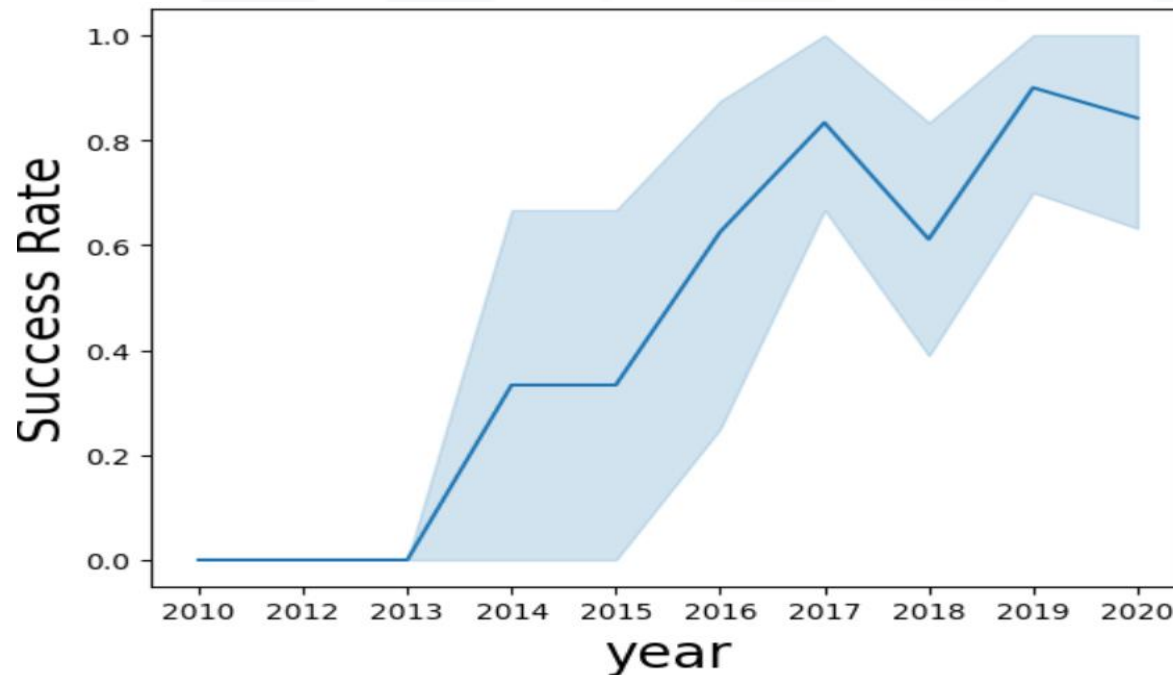
RESULTS – EDA WITH VISUALIZATION

- **Payload Mass (kg) & Orbit Type**
 - On GTO orbit, payload mass are negatively correlated with success rate
 - Orbit ISS, payload mass are positively correlated with success rate



RESULTS – EDA WITH VISUALIZATION

- **Year & Success Rate**
 - Since 2013 success rate is positively correlated with yearly



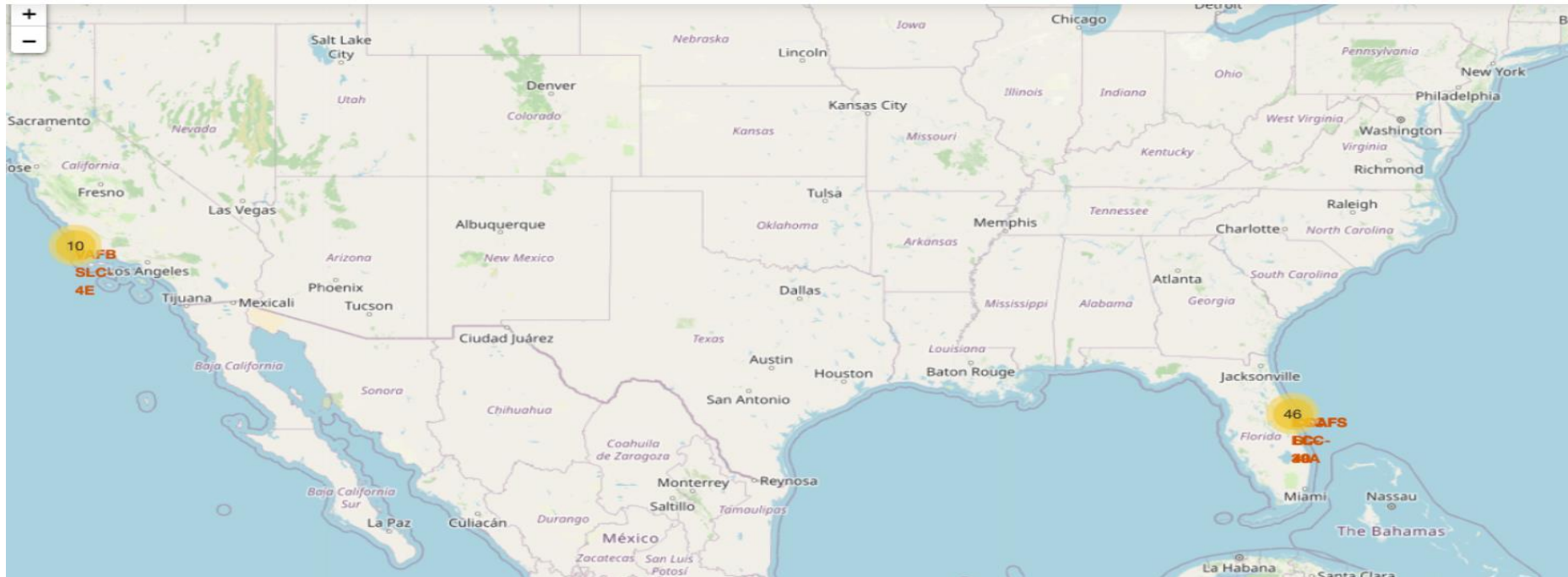
A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a dark blue sky with stars and a view of the Earth's surface from space. The Earth's surface is mostly dark, with a dense network of yellow and orange lights representing city lights at night. The lights are concentrated in the lower right portion of the image, following the curve of the Earth. The upper portion of the image shows the dark blue sky with a few stars.

Section 4

Launch Sites Proximities Analysis

RESULTS – LAUNCH SITES LOCATION ANALYSIS

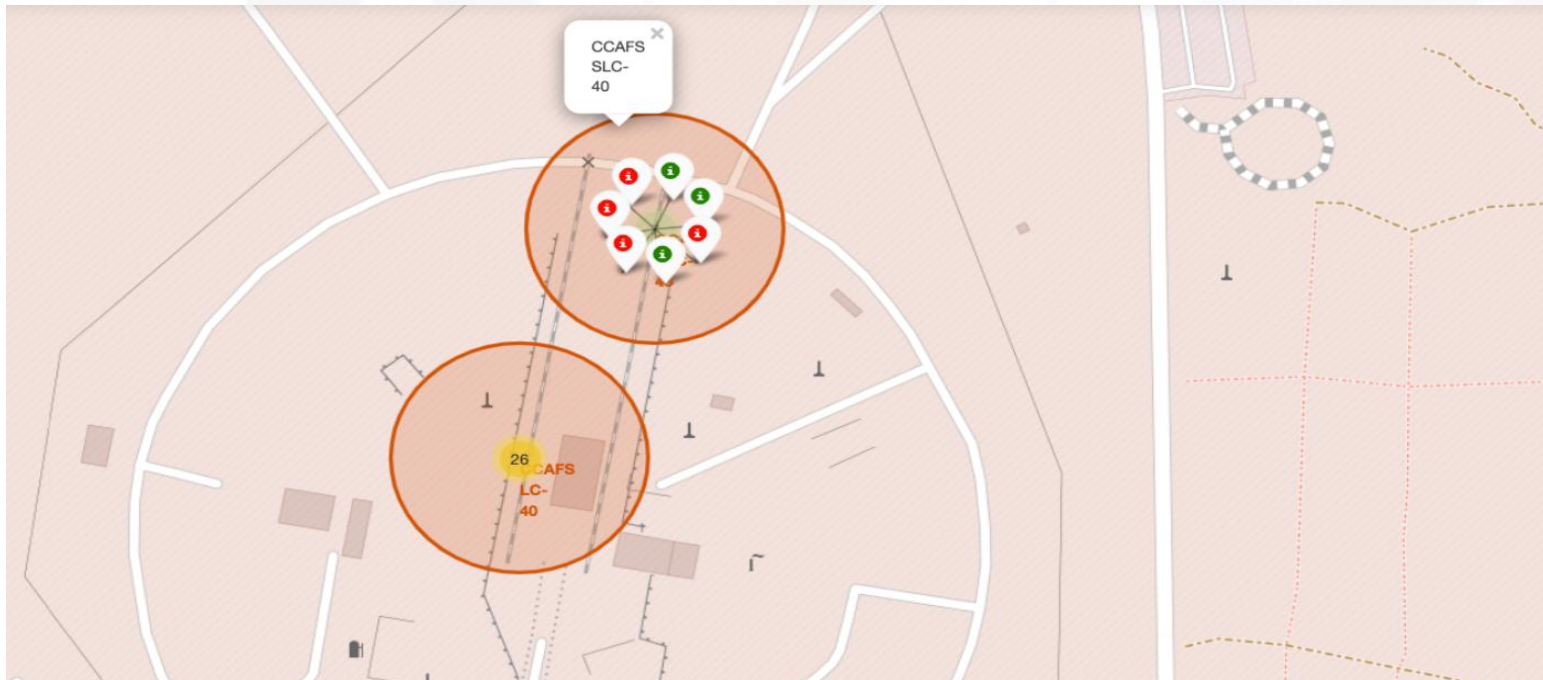
- Launch Sites
 - The biggest advantage from launching near the equator is that earth's rotation gives the payload a boost reaching orbit.



RESULTS – LAUNCH SITES LOCATION ANALYSIS

- **Launch Outcomes**

- Successful launch display with green markers and unsuccessful launch display with red markers.
- Visualizing the success rate of launch site CCAFS SLC-40 is 42.9%.



RESULTS – LAUNCH SITES LOCATION ANALYSIS

- Distance to Proximities
 - Visualizing the launch site CCASF SLC-40 is
 - 26.88 km distance from nearest highway
 - 23.33 km distance from nearest city
 - 21.96 km distance from nearest railway
 - .86 km distance from nearest coastline





Section 5

Build a Dashboard with Plotly Dash

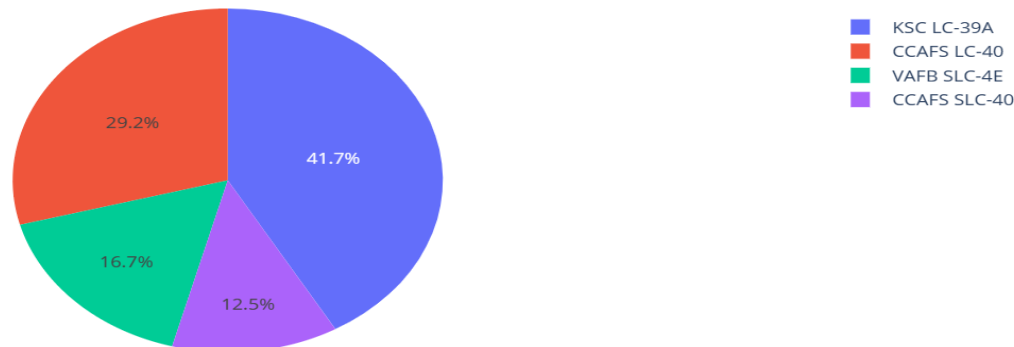
RESULTS - DASHBOARD WITH PLOTLY

- **Success Launches by Site**
 - KSC LC-39A launch site has 41.2%, it is the highest rate of successful launch

SpaceX Launch Records Dashboard

All Sites

Success Count for all launch sites



RESULTS - DASHBOARD WITH PLOTLY

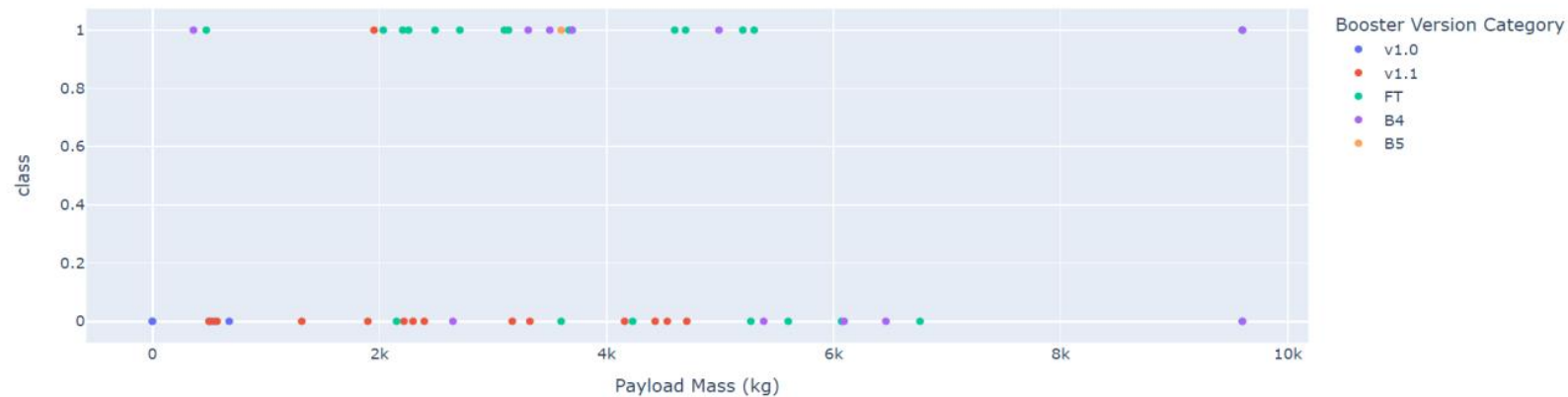
- **Payload Mass (KG) & Success**

- Less than of payload mass 5,000(kg) are increasing booster version success rate.
- Greater than of payload mass 5,000(kg) are decreasing booster version success rate.

Payload range (Kg):



Correlation Between Payload and Success for All Sites





Section 6

Predictive Analysis (Classification)

RESULTS – CLASSIFICATION ACCURACY

- Accuracy
 - All of the ML models are at same accuracy scores of 83.33%
 - Best model is DecisionTree with the score of 0.889

Out[56]:

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

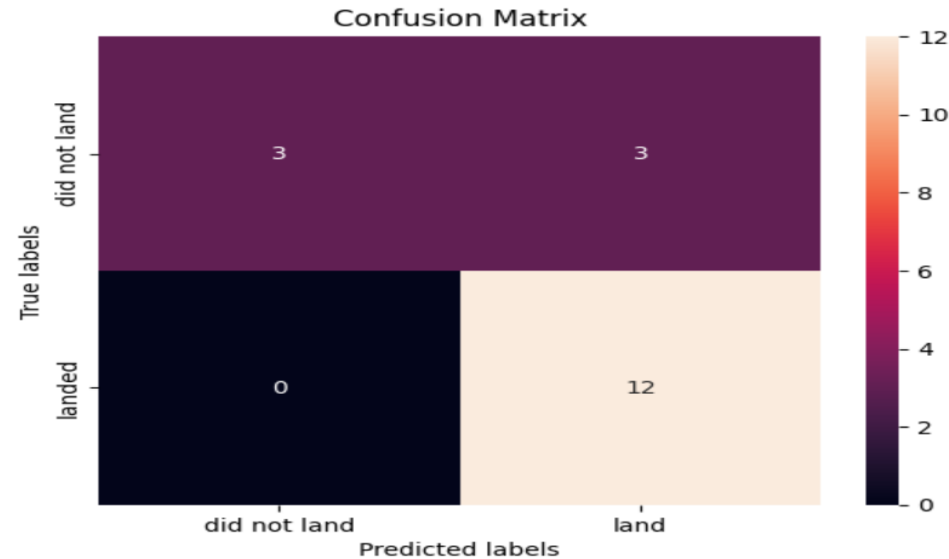
```
In [57]: models = {'KNeighbors': knn_cv.best_score_,
                  'DecisionTree': tree_cv.best_score_,
                  'LogisticRegression': logreg_cv.best_score_,
                  'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm, 'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)

Best model is DecisionTree with a score of 0.8892857142857145
Best params is : {'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'best'}
```

RESULTS – CONFUSION MATRIX

- **Performance Summary**
 - Confusion matrix of all models are identical.
 - Confusion matrix outcomes are 3 True negative, 3 False positive, 12 True positive, and 0 False negative.
 - 3 False positive is Type 1 error which mean model predicting incorrectly.



CONCLUSION

- Model Performance: All ML models are performing identical with the decision tree model is better than other.
- Equator and Coast: Most launch sites are near the equator. The extra boost given to rockets launched near the equator helps save money by not having to put in more fuel. Launching a rocket from the east coast gives an additional boost to the rocket, due to the rotational speed of Earth.
- Successful launch outcome: Becoming higher at all time
- KSC LC-39A: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- Orbits: ES-L1, GEO, HEO, and SSO have a 100% success rate
- Payload Mass: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Appendix

- **Github**
 - <https://github.com/zarnikhinkyi/Coursera-AppliedDataScienceCapstoneProject>
- **References**
 - <https://www.spacex.com/vehicles/falcon-9/>
 - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
 - <https://science.nasa.gov/learn/basics-of-space-flight/chapter14-1/>
- **Acknowledgements**
 - Thanks you to instructors, teaching assistants, forum moderators, other contributors, and staffs at IBM for letting me to learn the courses and materials.

Thank you!

