



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Mohamad Farabi Mohd  
Nasir  
28<sup>th</sup> March 2023



# Outline

---

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

# Executive Summary

---

- Summary of methodologies
  - Two ways of data collection.
  - Data wrangling for easier data analysis.
  - Exploratory Data Analysis (EDA) by using SQL.
  - Exploratory Data Analysis (EDA) by using data visualization on Python.
  - Create a dashboard using Plotly Dash.
  - Using 4 models for machine learning prediction.
- Summary of all results
  - By using SQL and data visualization, data can be easily understood. This can be proved by which variables affect the launch outcome.
  - The launch outcomes showed an upward trend throughout the year.
  - The best model in machine learning prediction is Decision Tree model.

# Introduction

---

- Project background
  - The project background is that the commercial space age is here and companies like Virgin Galactic, Rocket Lab, Blue Origin, and SpaceX are making space travel affordable and accessible to everyone. SpaceX has been particularly successful, with a track record of sending spacecraft to the International Space Station, launching Starlink satellite internet constellation, and conducting manned missions to space. One of the reasons for SpaceX's success is its ability to reuse the first stage of its Falcon 9 rocket, which significantly reduces launch costs.
- Problems statement
  - The problem that the project aims to solve is to determine the cost of a launch by predicting if the first stage of the Falcon 9 rocket will be successfully recovered and reused. This information will help a new rocket company, Space Y, founded by Billionaire industrialist Allon Musk, compete with SpaceX. The project will use data science techniques and machine learning models to analyze public information about SpaceX's launches and predict if the first stage will be reusable. The project will also create dashboards for the Space Y team to visualize the data and make informed decisions about launch costs.



Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection methodology:
  - SpaceX launch data can be collected in two ways:
    - SpaceX Rest API endpoint or URL ([api.spacexdata.com/v4/](https://api.spacexdata.com/v4/))
    - Web scraping related Wiki pages ([https://en.wikipedia.org/wiki/List\\_of\\_Falcon\\_9\\_and\\_Falcon\\_Heavy\\_launches](https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches))
- Perform data wrangling
  - The number of launches on each site, the number and occurrence of each orbit, and the number and occurrence of mission outcomes per orbit type were calculated. Lastly, in order to produce a classification variable Y, landing outcomes are converted to binary classes, with 0 denoting a bad outcome and 1 denoting a good one.
- Perform exploratory data analysis (EDA) using visualization and SQL

# Methodology

---

## Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data was split into training data and testing data to determine the training labels and find the best hyperparameter for all the classification models: Support Vector Machine, Classification Trees, k Nearest Neighbours and Logistic Regression. Comparison of the accuracy will be done to find the best model.

# Data Collection

---

- SpaceX launch data can be collected in two ways, SpaceX Rest API endpoint or URL ([api.spacexdata.com/v4/](https://api.spacexdata.com/v4/)) and web scraping related Wiki pages ([https://en.wikipedia.org/wiki/List\\_of\\_Falcon\\_9\\_and\\_Falcon\\_Heavy\\_launches](https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)).



# Data Collection – SpaceX API

---

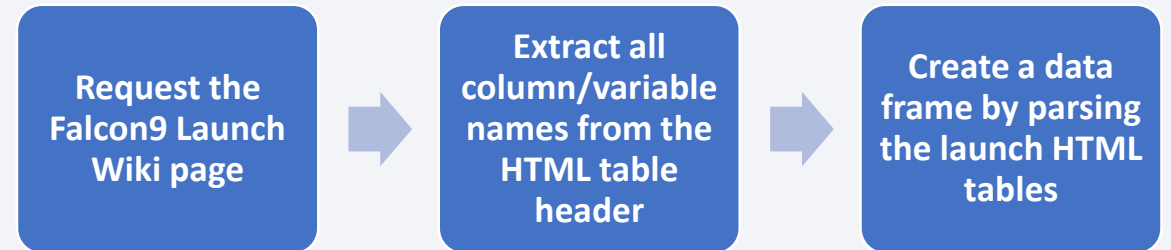
- SpaceX launch data that is gathered from an API, specifically the SpaceX REST API.
- This API will give us data about launches including other information about the launches.
- Source code:  
[https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/1\\_jupyter-labs-spacex-data-collection-api.ipynb](https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/1_jupyter-labs-spacex-data-collection-api.ipynb)



# Data Collection - Scraping

---

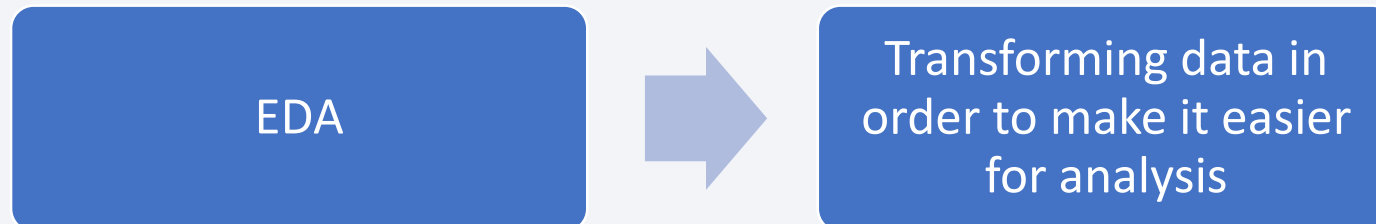
- Another way to collect the is Web scraping
- Data was obtained from Wikipedia
- Source code:  
[https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/2\\_jupyter-labs-web scraping.ipynb](https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/2_jupyter-labs-web scraping.ipynb)



# Data Wrangling

---

- Exploratory Data Analysis (EDA) was done on the dataset.
- The number of launches on each site, the number and occurrence of each orbit, and the number and occurrence of mission outcomes per orbit type were calculated.
- In order to produce a classification variable that represents the outcome of each launch, landing outcomes are converted to binary classes, with 0 denoting a bad outcome and 1 denoting a good one.



- Source code: [https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/3\\_labs-jupyter-spacex-Data%20wrangling.ipynb](https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/3_labs-jupyter-spacex-Data%20wrangling.ipynb)

# EDA with Data Visualization

---

- To help in understanding of the dataset, a few charts were plotted to see the relationship between the variables:
- Between Launch Site and Flight Number, between Payload Mass and Launch Site, between Success Rate and Orbit, between Flight Number and Orbit, between Payload Mass and Orbit.
- Lastly, linear chart of success yearly trend was plotted.
- Source code: [https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/5\\_jupyter-labs-eda-dataviz.ipynb](https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/5_jupyter-labs-eda-dataviz.ipynb)

# EDA with SQL

---

- A few SQL queries were executed to understand and explore the dataset. The SQL queries as follows:
  - Names of the unique launch site in the space mission.
  - 5 records where launch site begin with the 'CCA'.
  - Total payload mass carried by boosters launched by NASA (CRS).
  - Average payload mass carried by booster version F9 v1.1.
  - First successful landing outcome in ground pad was achieved.
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg.
  - Total number of successful and failure mission outcomes.



# EDA with SQL

---

- Names of the booster versions which have carried the maximum payload mass.
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
- 
- Source code: [https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/4\\_jupyter-labs-eda-sql-coursera\\_sqlite.ipynb](https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/4_jupyter-labs-eda-sql-coursera_sqlite.ipynb)

# Build an Interactive Map with Folium

---

- By using Folium Maps:
  - All launch sites were marked on a map.
  - Success or failed launches for each site were marked on a map.
  - By using the distances between a launch site to its proximities, a line was drawn and the distance between can be seen.
- Source code: [https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/6 lab jupyter launch site location.ipynb](https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/6%20lab%20jupyter%20launch%20site%20location.ipynb)

# Build a Dashboard with Plotly Dash

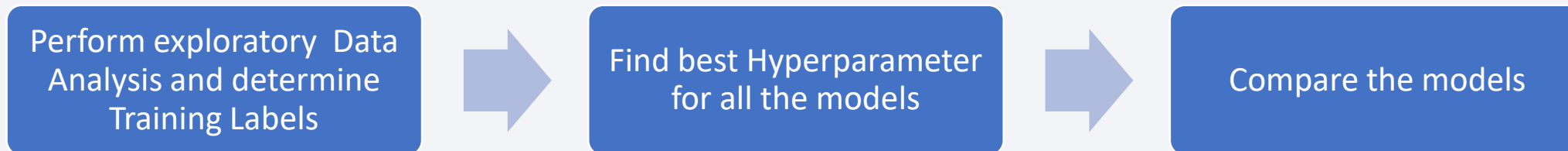
---

- Interactive dashboard to visualize data using charts.
  - Selectable launch site to display pie chart.
  - Range slider to select payload and display scatter plot based on the selected payload.
- Source code (python file): [https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/7.2\\_dashboard\\_with\\_plotly\\_dash.py](https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/7.2_dashboard_with_plotly_dash.py)
- Source code (jupyter notebook): [https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/7\\_dashboard\\_with\\_plotly\\_dash.ipynb](https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/7_dashboard_with_plotly_dash.ipynb)

# Predictive Analysis (Classification)

---

- The data was split into training data and testing data to determine the training labels
- Find the best hyperparameter for all the classification models: Support Vector Machine, Classification Trees, k Nearest Neighbors and Logistic Regression.
- Comparison of the accuracy will be done to find the best model.



- Source code: [https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/8\\_SpaceX\\_Machine%20Learning%20Prediction\\_Part\\_5.ipynb](https://github.com/mofarabi/IBM-Applied-Data-Science/blob/main/8_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb)

# Results

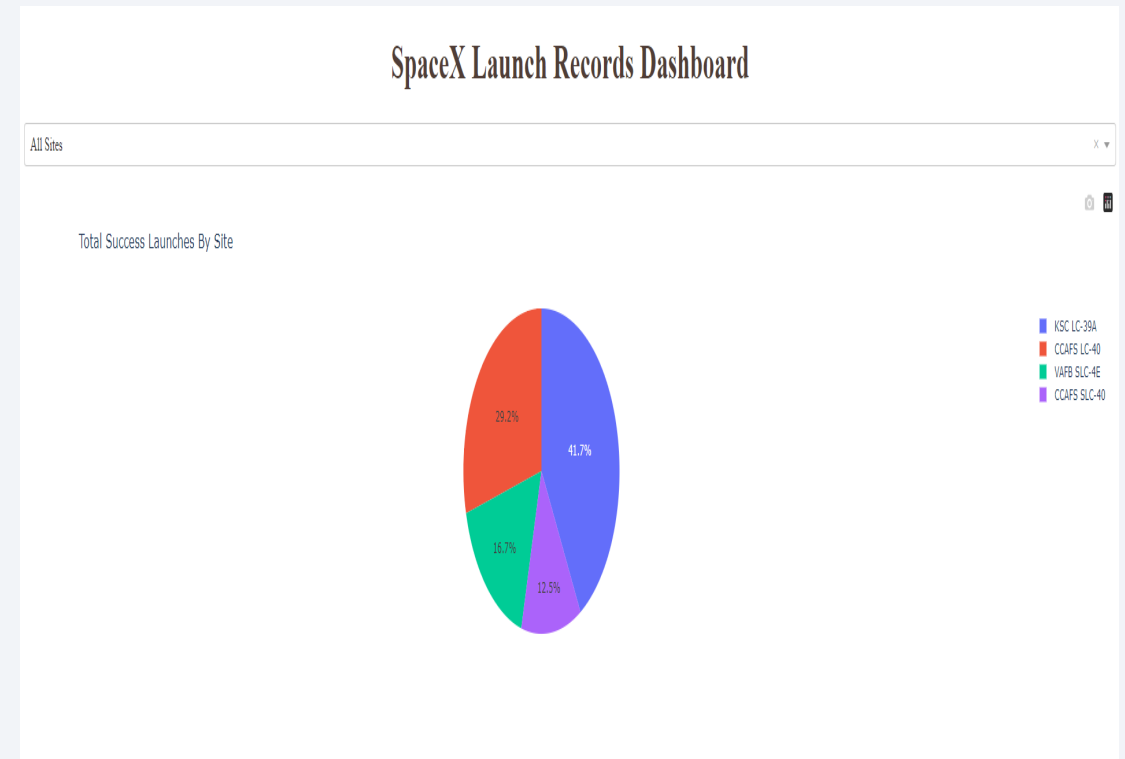
---

- Exploratory data analysis results:
  - CCAF5 SLC 40 is the most used launch site.
  - ES-L1, GEO, HEO and SSO have the highest success rate.
  - The success rate has been increasing over the years.
  - The total payload mass by boosters from NASA (CRS) is 45596 kg and the average payload mass by F9 v1.1 is 2534 kg.
  - The first successful ground landing was on 22<sup>th</sup> December 2015 and there were only two occurrences of failed landings in 2015, and they happened in January and April.



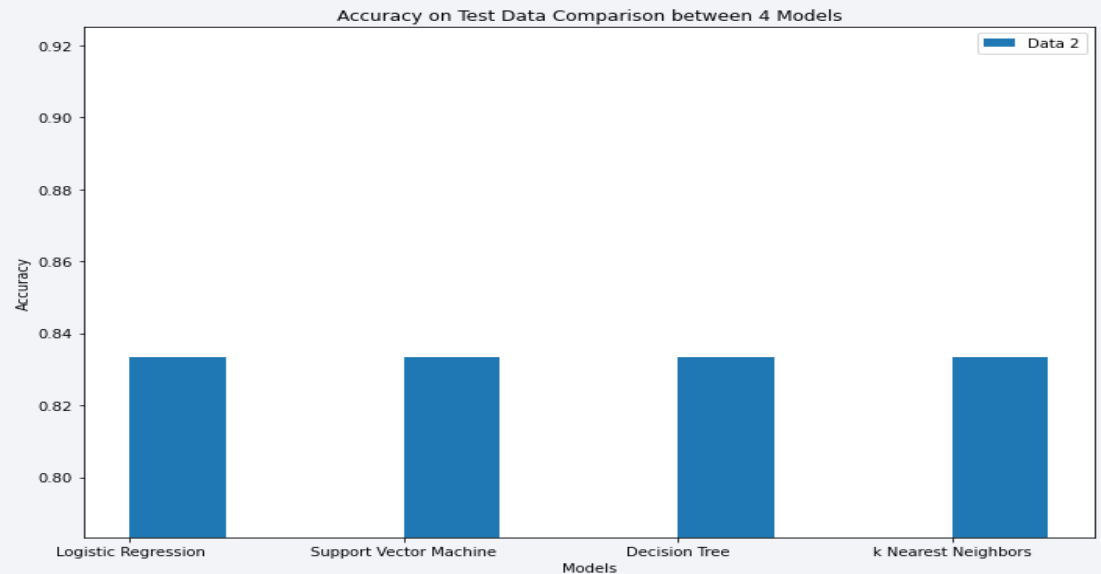
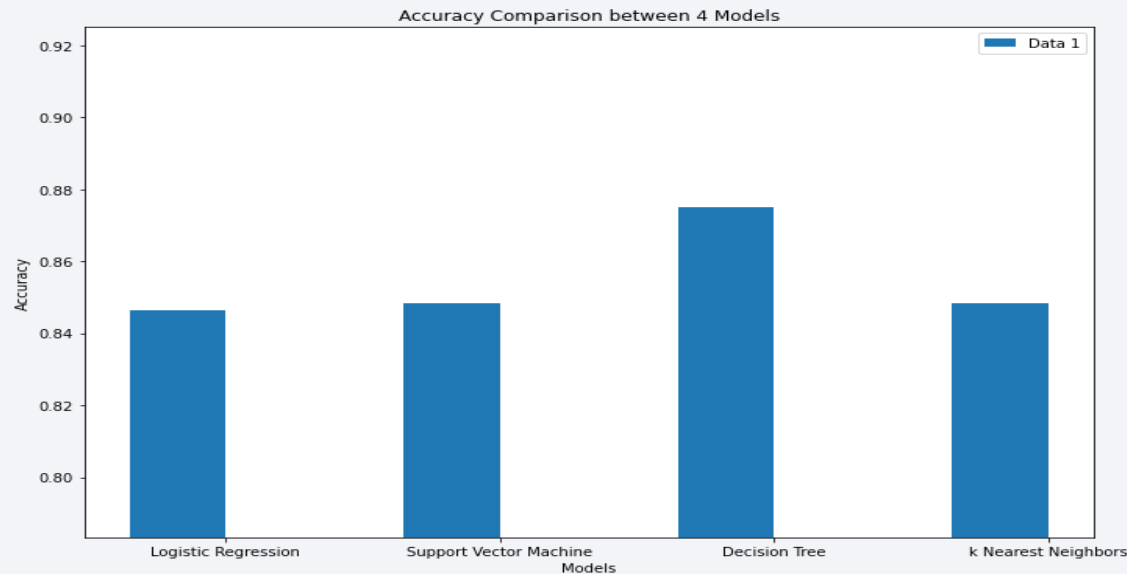
# Results

- Interactive analytics demo in screenshots:
  - KSC LC-39A is has the highest successful launches with 41.7%
  - With 29.2% of all launches, CCAFS LC-40 comes in second, followed by VAFB SLC-4E at 16.7% and CCAFS SLC-40 , which accounts for 12.5% of all launches.



# Results

- Predictive analysis results:



- Decision Tree model is concluded as the best model as it has the highest accuracy when all the models have been optimized using the grid search technique with their best hyperparameters and the same dataset.



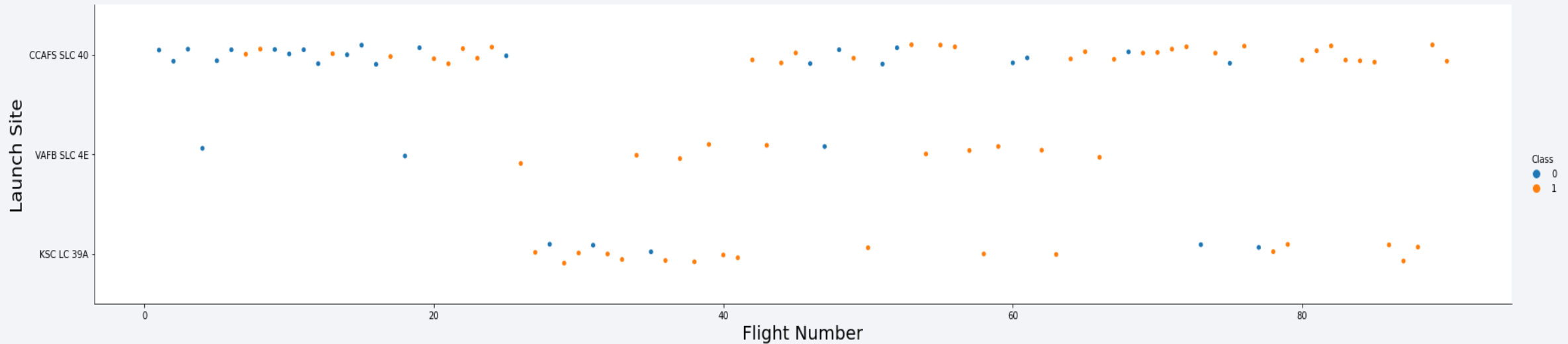
The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

# Insights drawn from EDA

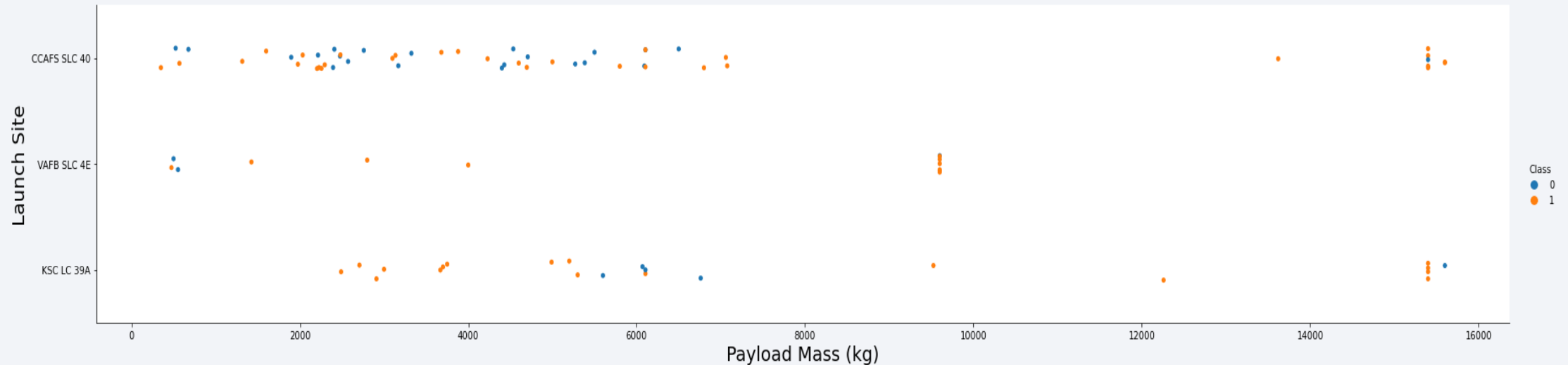


# Flight Number vs. Launch Site



- Based on the chart, CCAFS SLC 40 had the most flight numbers and had the most successful launches.
- Followed by KSC LC 39A in second place and VAFB SLC 4E in last place.

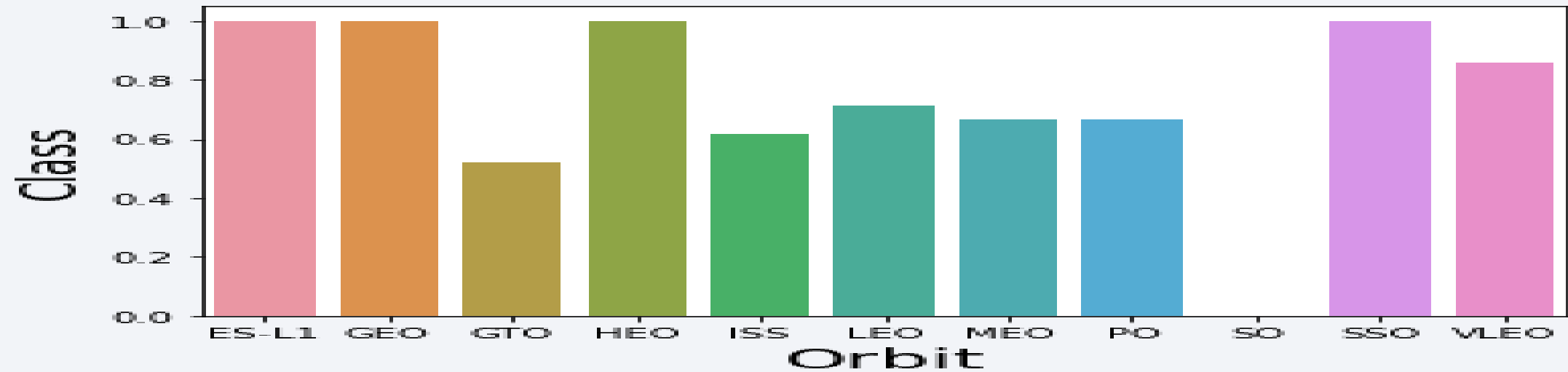
# Payload vs. Launch Site



- Most of the launches that exceeded the payload mass of 8000 kg were success.
- However, for VAFB SLC 4E, there were no rockets launched for payload mass over 10,000 kg.

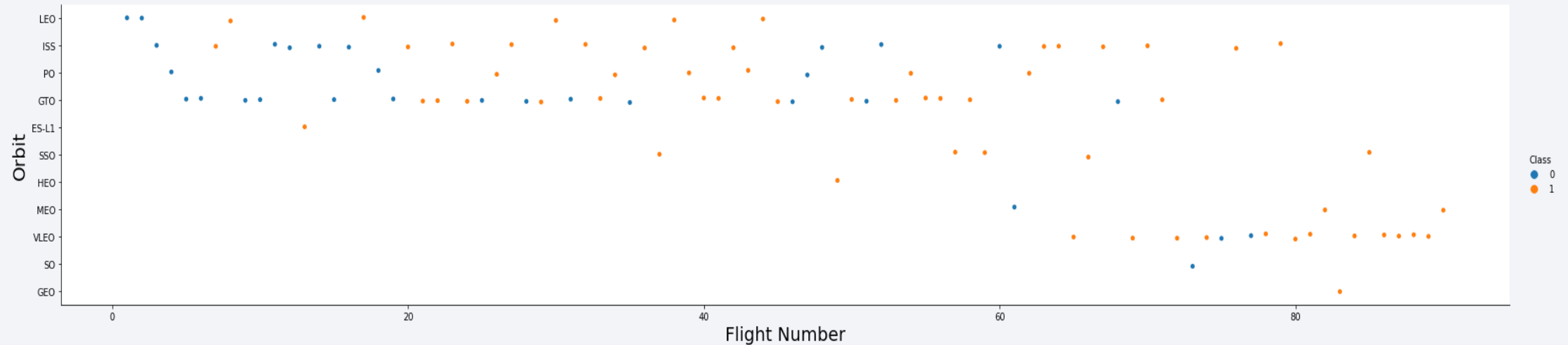


# Success Rate vs. Orbit Type



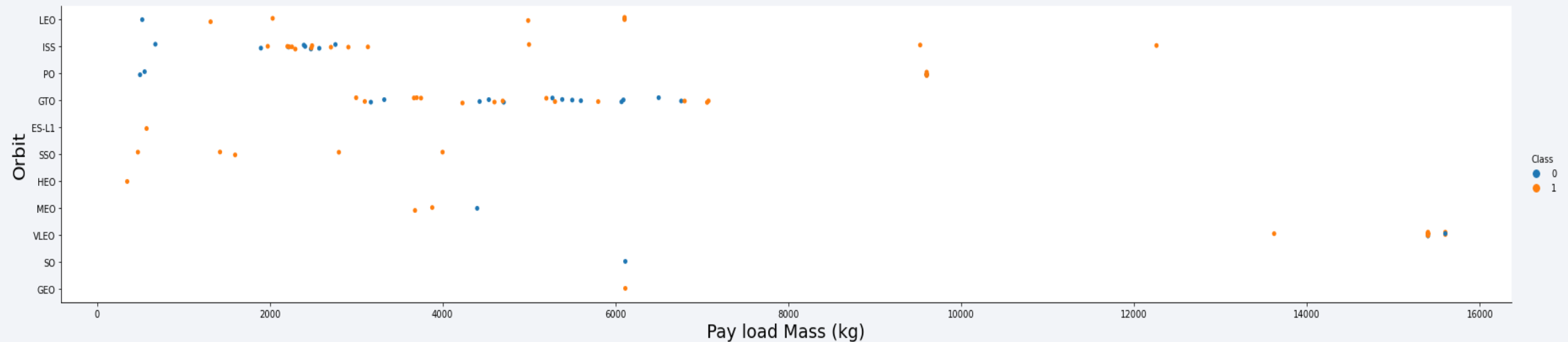
- The orbits that have the highest success rate are ES-L1, GEO, HEO and SSO.

# Flight Number vs. Orbit Type



- Based on this chart, GTO and ISS have the most flight number, but a low success rate.
- However, VLEO has a good success rate with a lot of occurrences, based on the flight number.

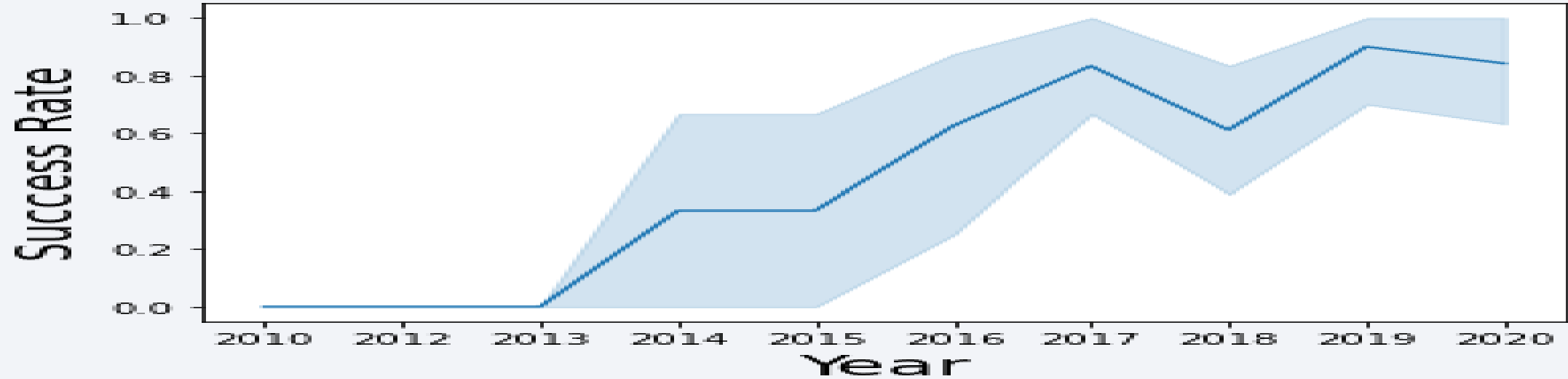
# Payload vs. Orbit Type



- There was no relationship between payload and orbit for GTO.
- While for PO, LEO and ISS, with heavy payload the successful landing are more.

# Launch Success Yearly Trend

---



- Success rate has been increasing from 2013.
- Improvement in technology and experienced helped in the increasing of the success rate.

# All Launch Site Names

---

- There are four launch sites in the dataset.
- This result can be gotten by using distinct in the SQL queries.

**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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- 5 records where launch sites begin with 'CCA'.
- The results can be gotten by using like 'CCA%'.

# Total Payload Mass

---

- The total payload carried by boosters from NASA (CRS) is 45596 kg.
- The value is from summing all payload where the customer is NASA (CRS).

**Total\_Payload\_Mass**

**45596**

# Average Payload Mass by F9 v1.1

---

- The average payload mass carried by booster version F9 v1.1 is 2534 kg.
- By filtering the data to booster version F9 v1.1, average payload mass for the booster can be calculated.

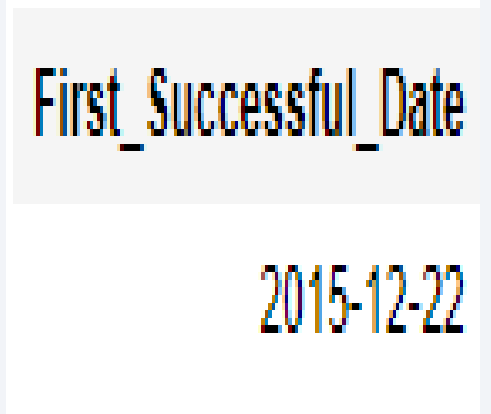
Average\_Payload\_Mass

2534.6666666666665

# First Successful Ground Landing Date

---

- The dates of the first successful landing outcome on ground pad is as shown in the figure.
- The date was on December 22, 2015. By using min on the Date and filter it by successful landing outcome on ground pad, the result will be gotten.



First\_Successful\_Date

2015-12-22

## Successful Drone Ship Landing with Payload between 4000 and 6000

---

- Based on the figure, the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- Filtering out the booster version by their payload mass greater than 4000 but less than 6000 will result in these 4 booster versions.

**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 number of successful and failure mission outcomes.
- The result can be gotten by grouping up mission outcomes and counting records for each group.

Mission_Outcome	Total_Number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

# Boosters Carried Maximum Payload

---

- Names of the booster which have carried the maximum payload mass.
- By using subquery, maximum payload mass can be gotten to filter out the name of the booster version.

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 of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

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

- There were only two occurrences of failed landings in 2015, and they happened in January and April.



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

- 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 is shown in the figure
- 'No attempt' must be taken into consideration in farther analysis.

Landing_Outcome	Success_Count
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

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

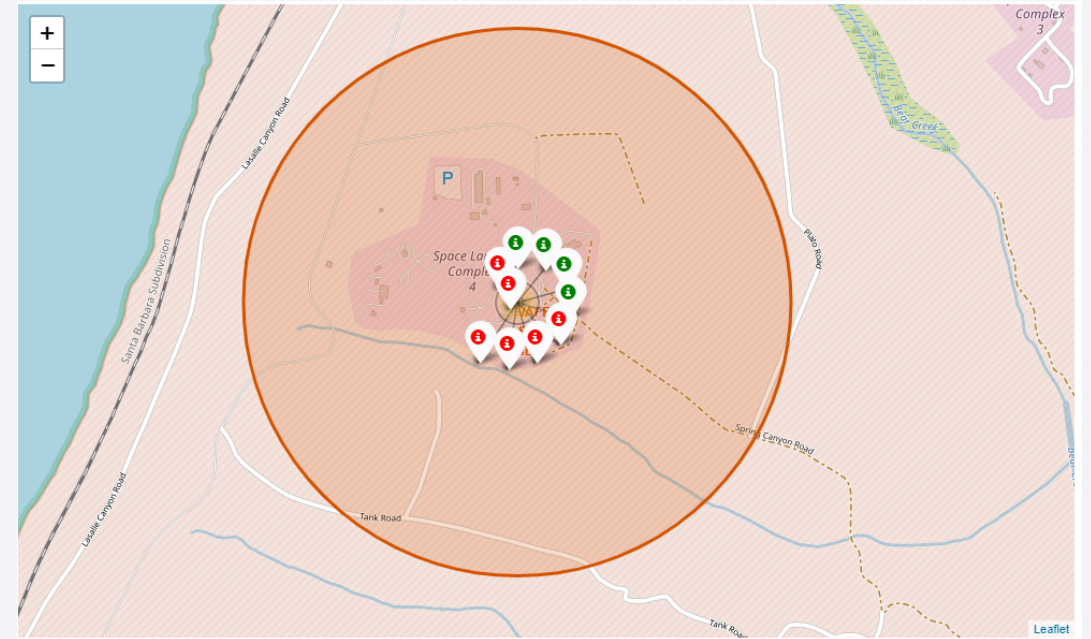
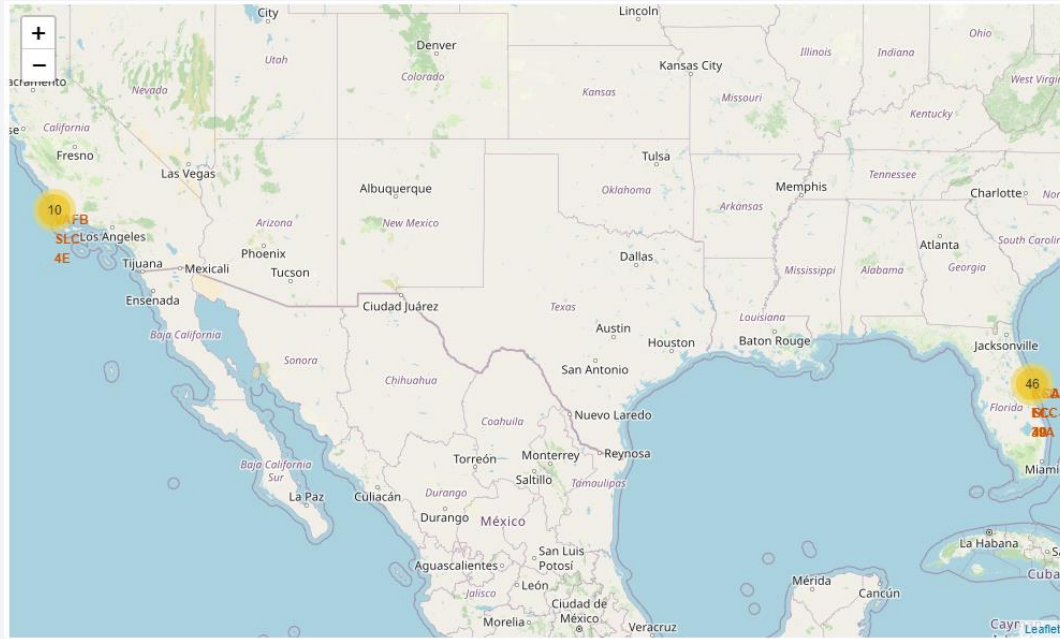
# Launch Sites Proximities Analysis

# All Launch Sites on a Map



- Based on the figure, the launch sites were located near the sea to lower the risk of launching over populous regions, improving third parties' safety.
- Other than that, out of 4 launch sites, only one is farther from the other.

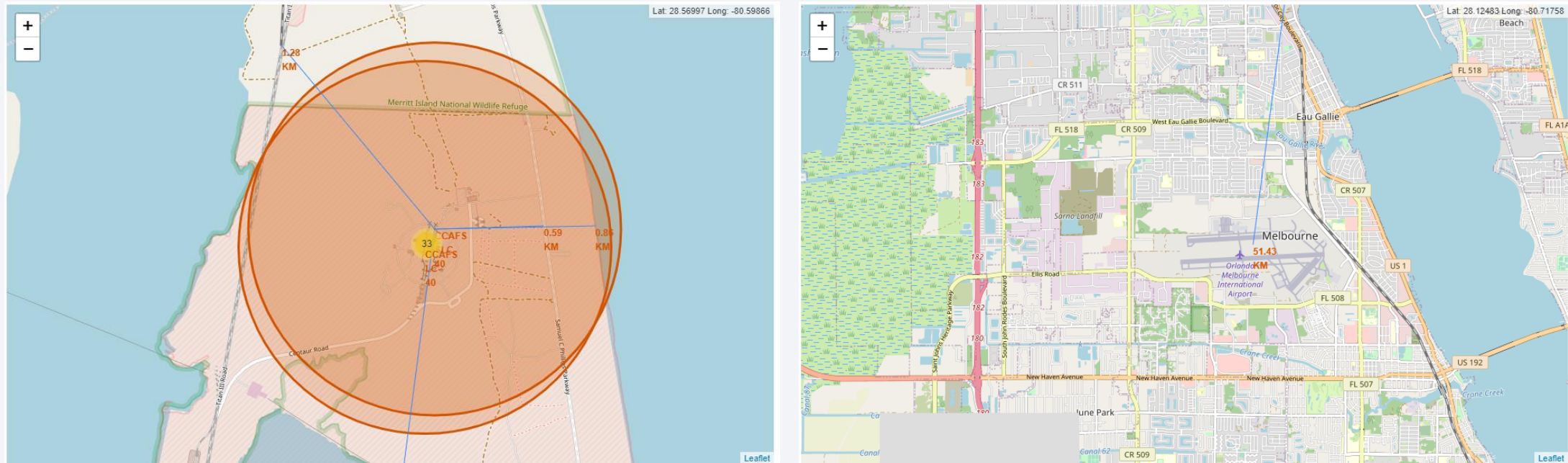
# Launch Outcomes for Each Site



- One of the launch site, VAFB SLC 4E is used as an example.
- From the color-labeled markers in the marker clusters, launch sites that have relatively high success rates can be easily identified.

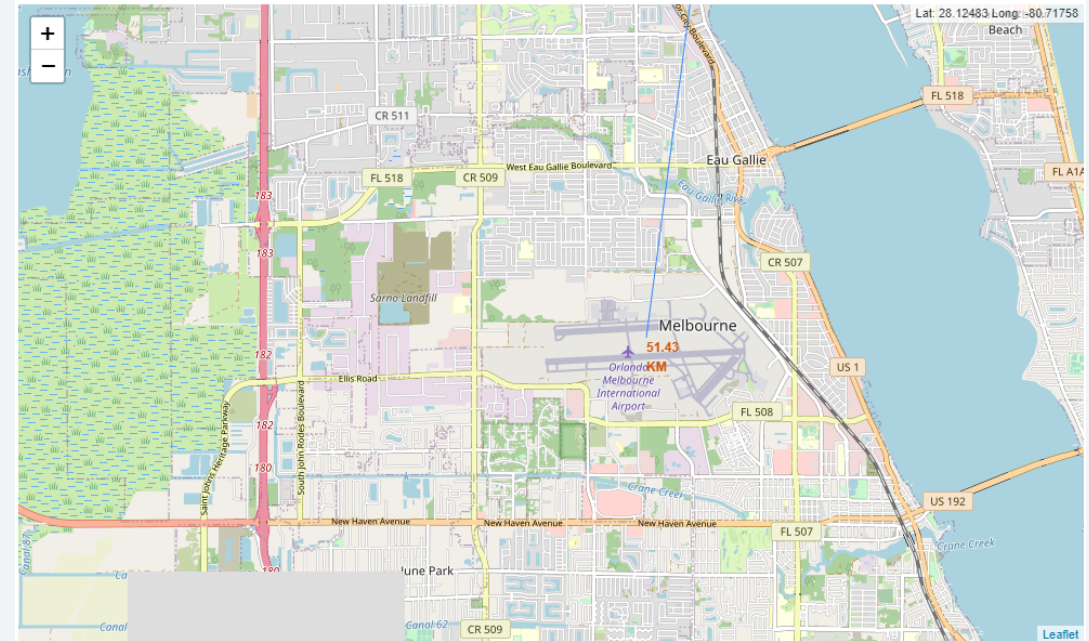
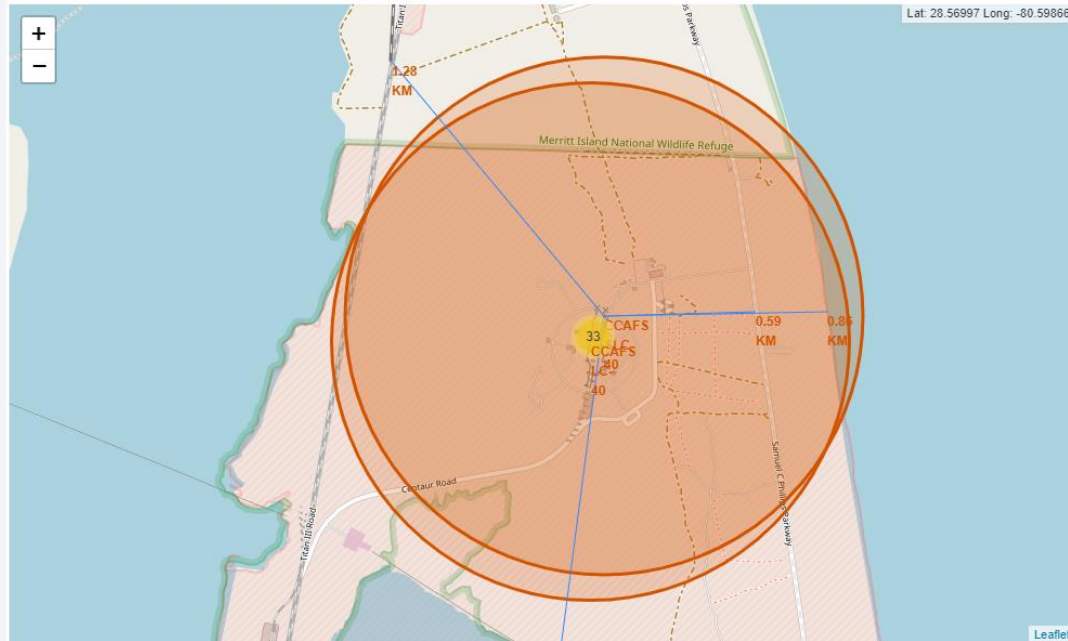


# Launch Site to its Proximities



- Launch site CAFS SLC 40 is used in this example, and the distance between the launch site and its proximities were calculated and shown in the map.
- The coastline, highways, and railways are respectively 0.8 km, 0.59 km, and 1.28 km away from the launch site, while the city is 51.43 km away.

# Launch Site to its Proximities



- As mentioned before, the launch site was located near the sea, which reduced the risk of launching over densely populated areas and improved safety for third parties. Being far from the city, the local community will not be affected by the rocket launch.
- Furthermore, having railways and highways close to the launch site will help in transporting materials needed for launching the rocket and developing the launch site.





Section 4

# Build a Dashboard with Plotly Dash

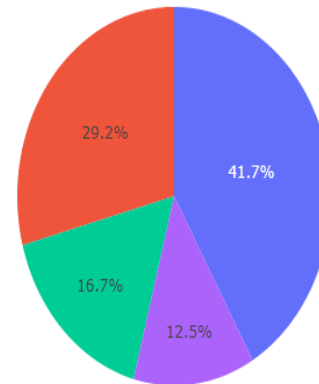
# Successful Launches All Sites

## SpaceX Launch Records Dashboard

All Sites



Total Success Launches By Site



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

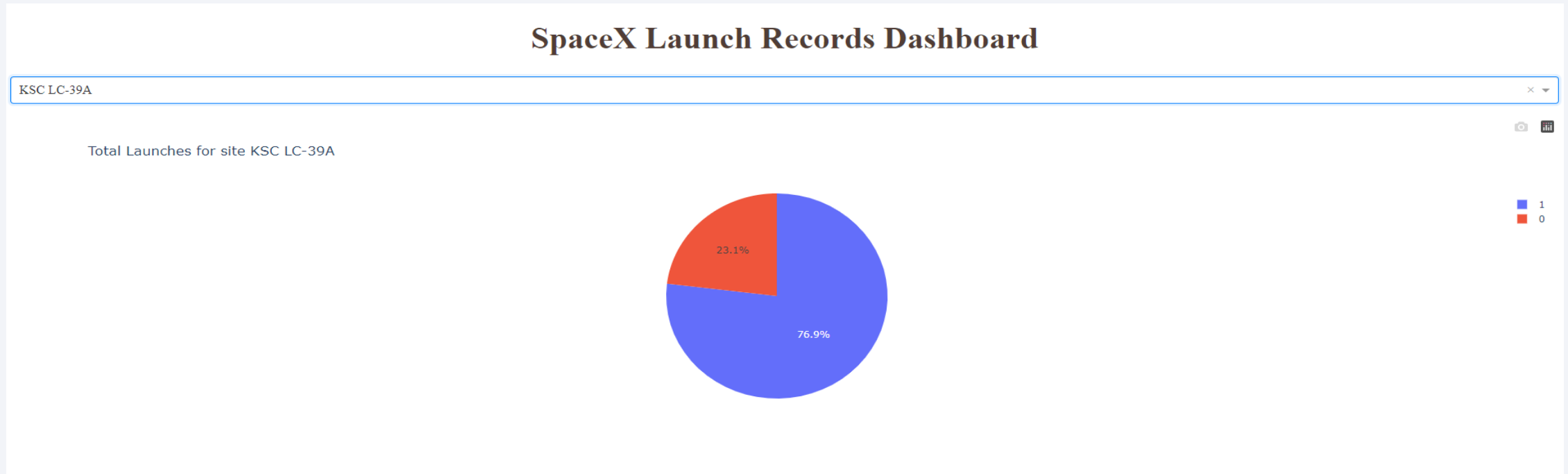


# Successful Launches All Sites

---

- The chart shows that KSC LC-39A is the majority for successful launches (41.7%).
- With 29.2% of all launches, CCAFS LC-40 comes in second, followed by VAFB SLC-4E at 16.7% and CCAFS SLC-40 , which accounts for 12.5% of all launches.
- The pie chart useful for quickly grasp how launches are distributed among the various launch sites. It demonstrates that the majority of launches take place at KSC LC-39A and CCAFS LC-40 , the two main launch sites, whereas fewer launches occur at VAFB SLC-4E and CCAFS SLC-40 .

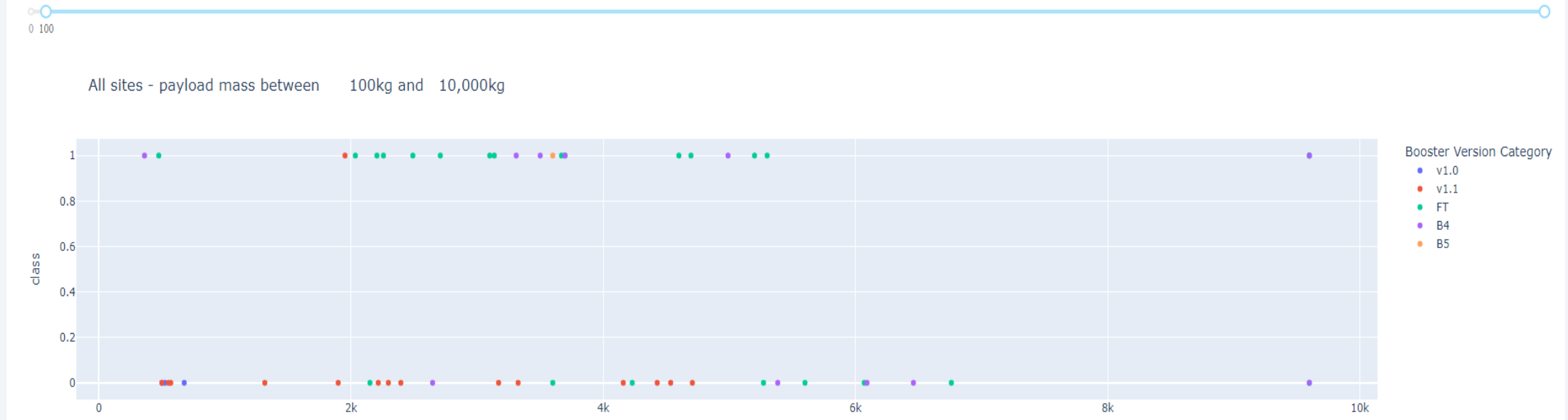
# Successful Launch by Sites



- The figure shows that KSC LC-39A as it has the highest success rate of any launch site at 76.9% for all of its launches.
- It demonstrates that KSC LC-39A has a comparatively high percentage of success, which can be attributable to a number of elements including its location, infrastructure, and launch procedures.

# Payload vs. Launch Outcome All Sites

Payload range (Kg):

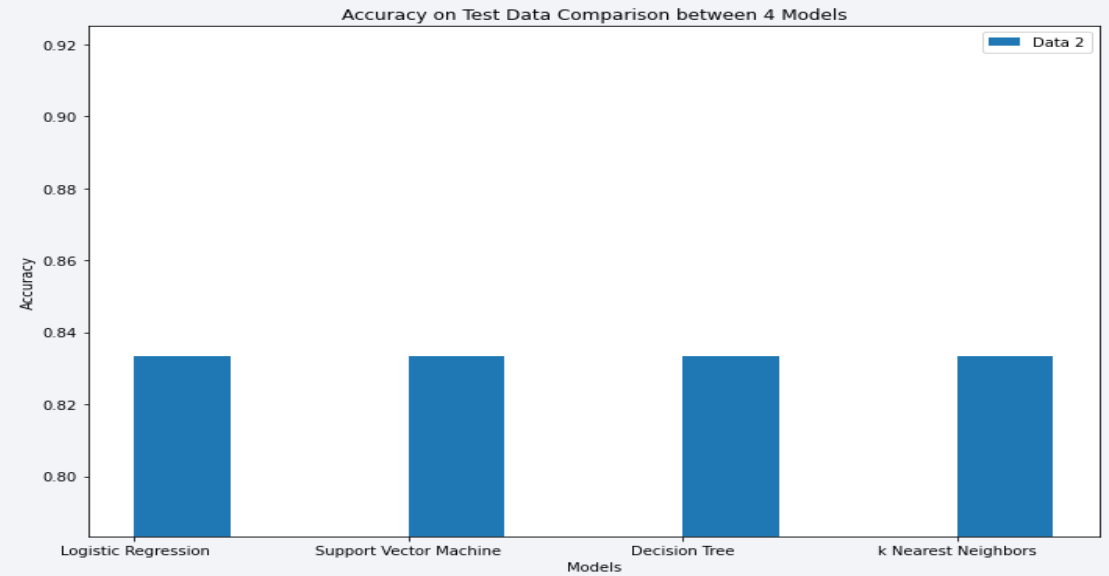
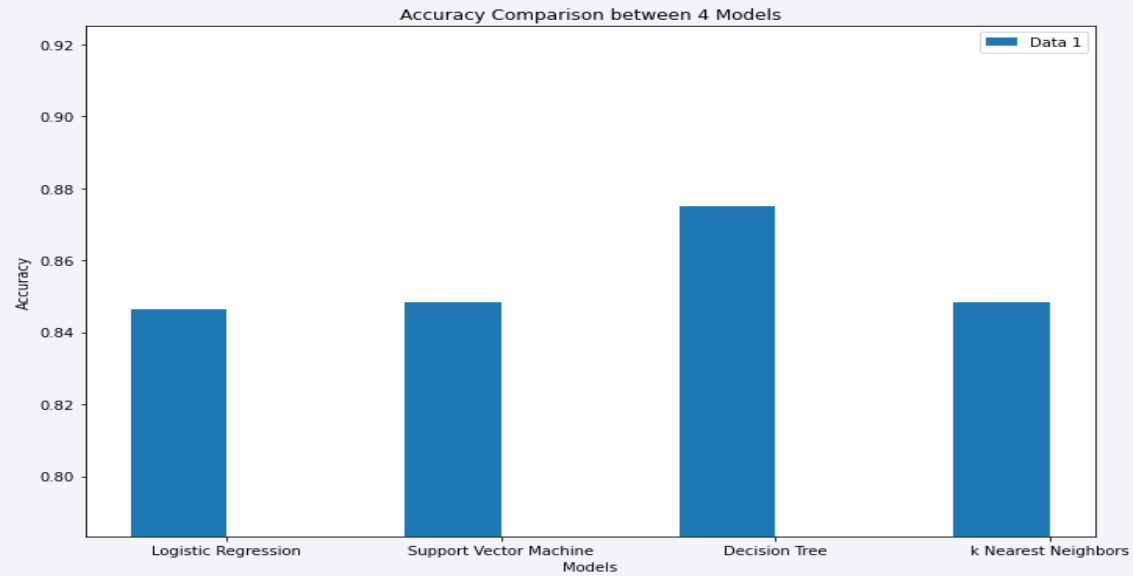


- Most of the successful launch outcomes were under 6000 kg of payload.
- Other than that, booster version B4 and FT were used the most.

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy



Model	Accuracy	Accuracy on Test Data
Logistic Regression	0.846429	0.833333
Support Vector Machine	0.848214	0.833333
Decision Tree	0.875000	0.833333
k Nearest Neighbors	0.848214	0.833333

# Classification Accuracy

---

- Based on the second chart, the value for accuracy on test data are the same for all four models. Hence, accuracy between four model will be used to determine the best model.
- The first chart shows highest performance score achieved by a model during the grid search process. This score is typically determined by some evaluation metric, such as accuracy. It is to evaluate the model during the hyperparameter tuning process.
- Hence, based on the highest value of accuracy, which is 87.5%, it can be concluded that Decision Tree is the best model.

# Confusion Matrix

- Based on the confusion matrix for Decision Tree, evaluation metrics for the model can be calculated. The accuracy for the model is 83%. That means the model can correctly predicted 83% of the cases in the dataset.



	Predicted value		
Actual value		Negati ve	Positiv e
	Negati ve	TN	FP
	Positiv e	FN	TP

# Conclusions

---

- Data could be obtained in two ways.
- Exploratory Data Analysis (EDA) is important to understand the dataset.
- Orbit type affect the success rate of the rocket launch.
- The success rate has been increasing over the year.
- All the launch site are close to coastline or the sea.
- KSC LC-39A has the most successful launches.
- Payload above 6000kg shows a low success rate.
- Decision Tree model is the best machine learning prediction model for this dataset.



# Appendix

---

[Github](#)

Thank you!

