Winning Space Race with Data Science

Nam Hyuk KIM

April 15th, 2024

Outline

Executive Summary

- Space X advertised that Falcon 9's launch cost much less than its competitors by reusing the first stage. In this project, I tried to determine if we can predict the landing of the first stage of Falcon 9 to assess the cost of a launch and the validity of Space X's advertisement.
- To execute this project, I created the machine learning pipeline as follows: I collected relevant data from Space X API and web-scraping and performed data wrangling on the data. Then, I performed exploratory data analysis (EDA) using data visualization, SQL queries, Folium maps and Plotly dash to better understand the data. Finally, I trained and tested the classification models with the data and suggested the best performing models.
- As a result, I found that Falcon 9's mission outcome recorded excellent success rate, while the first-stage landing success was not as impressive as mission itself. I also found that the success or failure outcome of the Falcon-9 first-stage landing can be predicted with 0.833 % of accuracy by using the Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) classification models.
- Also, I attained certain insights that launch sites, payload mass, and orbit types may influence the Falcon 9's first stage landing success rate. Future study may delve into this influence and correlation in detail.

Introduction

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, while, for other providers, rocket launches cost upward of 165 million dollars each. Much of the savings comes from Space X's ability to reuse the first stage.
- Therefore, if we can determine if the first stage will land, we can assess the cost of a launch and the validity of Space X's advertisement. This information can be used if an alternate company wants to bid against space X for rocket launch projects.
- In this capstone project, I created a machine learning pipeline to predict the landing of the first stage based on the data collected from Space X API or web-scraping, so as to find the best machine learning models for prediction.
- First, I will introduce the methodology for the machine learning pipeline. Then, I will share the major results and my insights from them for each pipeline stage. Finally, I will try to answer the question "Can we predict if the first stage will land?" and suggest the best model.

Section 1

Methodology

Methodology

Executive Summary

- Data Collection: Relevant data were collected from Space X API using the GET request and turned into a dataframe and a CSV file after filtering irrelevant information. Also, another CSV file was created by web scraping based on the BeautifulSoup library.
- Data Wrangling : Missing data were detected and replaced by the mean value of their column. The new binary Class column including target variable (landing success or failure) was created from the first-stage landing outcome column.
- EDA using data visualization and SQL : Scatter charts, bar plots, and a line chart were created to identify correlations and trends among variables A set of information was called by SQL queries.
- Interactive visual analytics using Folium and Plotly Dash : Interactive maps made by Folium denoted the location of launch sites and landing outcomes. The dashboard created by Plotly Dash exhibited pie charts and scatter charts to show important ratios and correlations.
- Predictive analysis using classification models : Logistic regression, support vector machine, decision tree classification models were trained and tested to find out their best parameters and the best performing models for prediction.

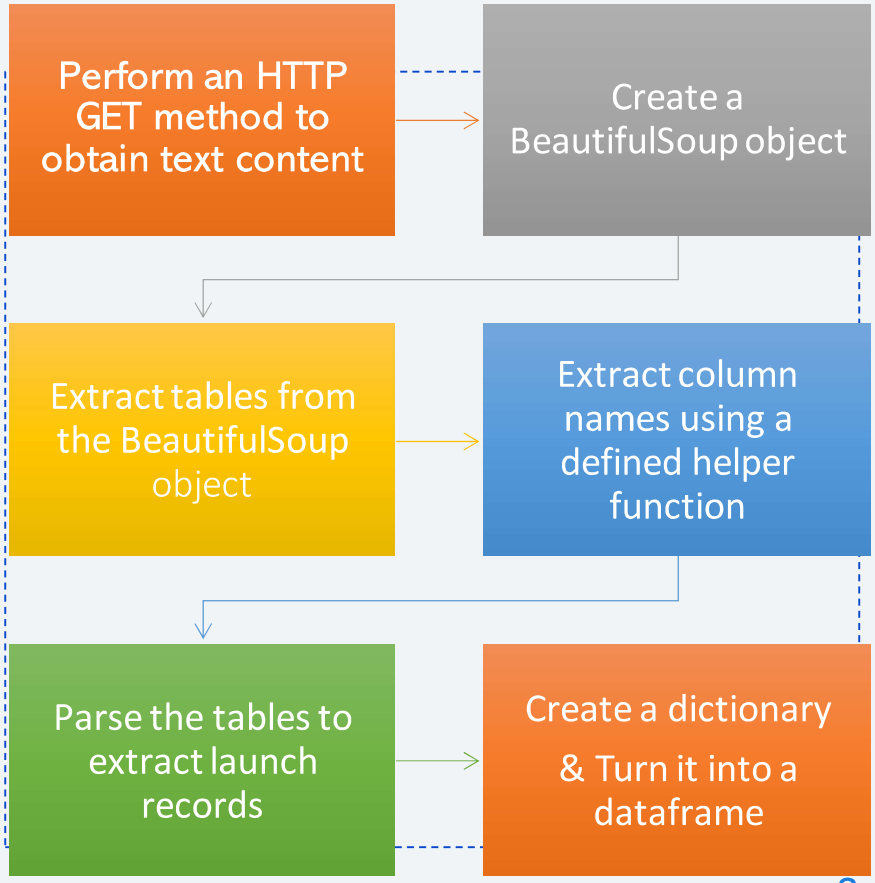
Data Collection – SpaceX API

- First of all, I envisioned relevant columns for the SpaceX launch dataframe.
- Second, I identified necessary endpoints in SpaceX API to extract those columns.
- Third, I performed GET request using the requests library to obtain contents from each SpaceX API endpoints.
- Fourth, I decoded the obtained contents as Json files using .jason() method and turn them into a list format using .append() method.
- Fifth, I combined these lists in the dictionary format with necessary adjustment, so as to turn them into a dataframe with relevant columns.
- Finally, I removed the Falcon 1 launches to keep only the Falcon 9 launches, using a filtering method then resetting flight numbers.
- The outcome is the file "dataset_part_1.csv"

GitHub URL : https://github.com/knh444/Capstone-Presentation-for-Data-Science/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

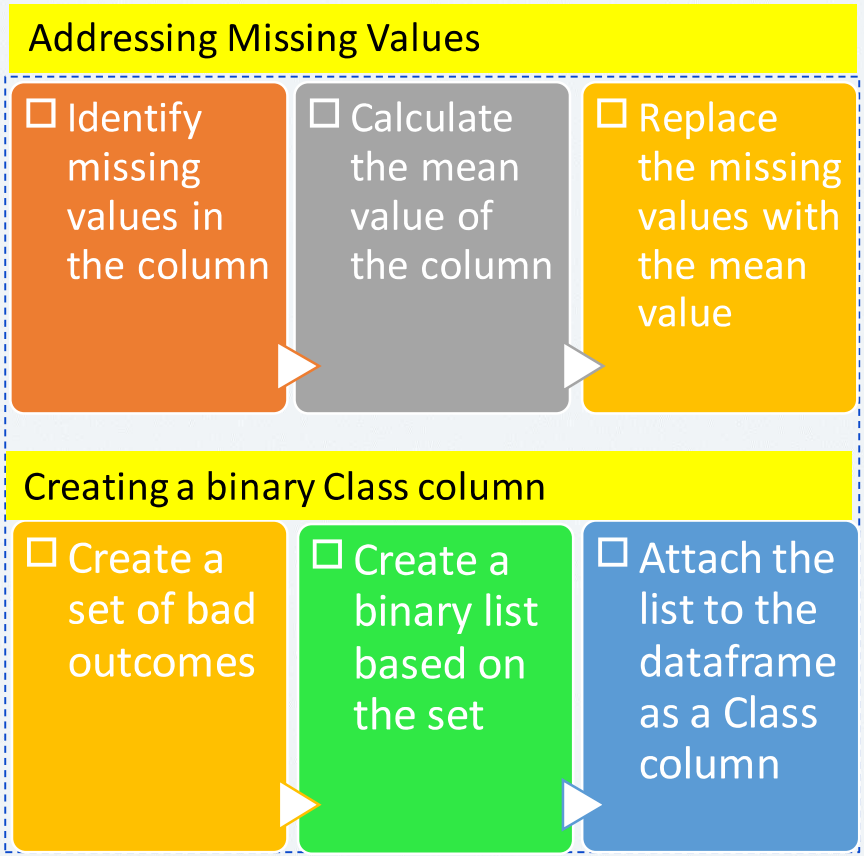
- First, I performed an HTTP GET method to obtain text content from the Falcon 9 Launch HTML page.
- Second, I used BeautifulSoup() to create a BeautifulSoup object from the text content.
- Third, I extracted all the tables from the BeautifulSoup object using the "find_all" function,
- Fourth, I extracted column names one by one through <th> elements, using the helper function.
- Fifth, I parsed all the tables to extract launch records under the column names, using for loop and defined helper functions.
- Finally, I created a dictionary with the column names and their records; and turn it into a dataframe.
- The outcome is the file "spacex_web_scraped.csv"



GitHub URL : https://github.com/knh444/Capstone-Presentation-for-Data-Science/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- To address missing values in the “PayloadMass” column of the dataframe (file : dataset_part_1.csv), I identified them, using .isnull().sum().
- Then, I replaced them by the mean value of the column, using .mean() and .replace().
- To transform the strings in the “Outcome” column of the dataframe to binary variables, I created a set of bad outcomes that the first stage did not land successfully.
- Then, I created a list where the element is zero if the corresponding row in Outcome is in the set of bad outcome; otherwise, it's one, using for loop.
- Finally, I attached the list to the dataframe as a new column ‘Class’ and save the data as “dataset_part_2.csv”.



GitHub URL : https://github.com/knh444/Capstone-Presentation-for-Data-Science/blob/main/jupyter-labs-webscraping.ipynb

EDA with Data Visualization

No.	I created the plots below from "dataset_part_2.csv"	Why?
1	a scatter chart with x axis = FlightNumber and y axis = PayloadMass, and hue to the Class value.	To see relationship between the flight number and the payload mass, and the landing outcome
2	a bar plot with x axis = LaunchSite and y axis = Success Rate.	To see the success rate for each launch site.
3	a scatter chart with x axis = FlightNumber and y axis = LaunchSite, and hue to be the Class value.	To see the relationship between the flight number and the launch site, and the landing outcome
4	a scatter chart with x axis = PayloadMass and y axis = LaunchSite, and hue to be the Class value.	To see the relationship between the payload mass and the launch site, and the landing outcome
5	a bar plot with x axis = Orbit and y axis = Success Rate.	To see the success rate for each orbit type.
6	a scatter chart with x axis = FlightNumber and y axis = Orbit, and hue to be the Class value.	To see the relationship between the flight number and the orbit type, and the landing outcome.
7	a scatter chart with x axis = PayloadMass and y axis = Orbit, and hue to be the Class value.	To see the relationship between the payload mass and the orbit type, and the landing outcome
8	a line chart with x axis = Year and y axis = the average success rate.	to get the average launch success trend by year

GitHub URL : https://github.com/knh444/Capstone-Presentation-for-Data-Science/blob/main/edadataviz.ipynb

EDA with SQL (1)

- First, I loaded the ipython-sql extension, connected to a SQLite database, created a cursor object, and connected the Jupyter session to a SQLite database. Then I loaded "Spacex.csv" and stored it in SQLite database.
- For Task 1, I displayed the names of the unique launch sites, using SELECT DISTTINCT query
- For Task 2, I displayed 5 records where launch sites begin with the string 'CCA', using SELECT, WHERE, LIKE, CCA%, LIMIT
- For Task 3, I displayed the total payload mass carried by boosters launched by NASA (CRS), using SELECT SUM, WHERE, LIKE, NASA%
- For Task 4, I displayed average payload mass carried by booster version F9 v1.1, using SELECT AVG, WHERE, LIKE, %v1.1%
- For Task 5, I listed the date when the first successful landing outcome in ground pad was achieved, using SELECT MIN, WHERE, LIKE, %Success%

GitHub URL : https://github.com/knh444/Capstone-Presentation-for-Data-Science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with SQL (2)

- For Task 6, I listed the names of boosters which have success in drone ship and have payload mass greater than 4000 kg but less than 6000 kg, using SELECT DISTINT, WHERE, AND, BETWEEN
- For Task 7, I listed the total number of successful and failure mission outcomes, using SELECT, COUNT, GROUP BY
- For Task 8, I listed the names of the booster version which have carried the maximum payload mass, using SELECT DISTINCT, subquery with WHERE, SELECT MAX
- For Task 9. I listed the records displaying the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015, using SELECT SUBSTR, WHERE, AND, SUBSTR
- For Task10, I ranked 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, using SELECT, COUNT, WHERE, BETWEEN, GROUP BY, ORDER BY COUNT, DESC

GitHub URL : https://github.com/knh444/Capstone-Presentation-for-Data-Science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

I created map objects below from "spacex_launch_geo.csv"	Why?	Functions of the objects
Circles	To identify the area where the launch sites are located.	Drawing a circle around a set distance from a set coordinate.
Markers	To identify the name of the launch sites	Showing the name of the location at the circled coordinate
Marker Clusters	To identify relative landing success rates of the launch sites	Showing colored features of data points related to the circled coordinates
Distance Markers	To calculate and show distance between the launch site and landmarks	Showing distance between two coordinates
LInes	To mark the shortest distance between the launch site and landmarks	Drawing a line between two coordinates

GitHub URL : https://github.com/knh444/Capstone-Presentation-for-Data-Science/blob/main/lab_jupyter_launch_site_location.ipynb

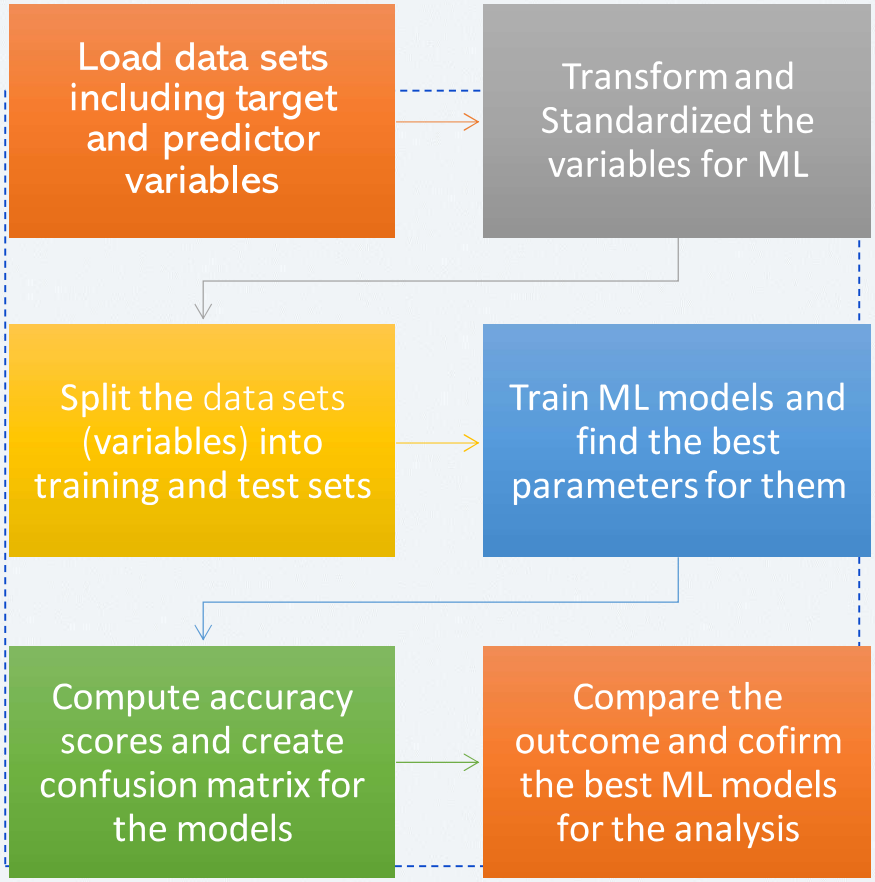
Build a Dashboard with Plotly Dash

I added Plot, Graphs, & Interactions below from "spacex_launch_dash.csv"	Why?	Functions
Dropdown list	The selection activates callback function to show relevant charts.	Enabling the selection of the launch site in the list
Pie chart	The pie chart shows successful landing ratio by the launch site and the percentage of success and failure for each site.	Presenting the composition (%) of each category
Callback function for the pie chart	This function enables the selection at the dropdown list to show a relevant pie chart for each site.	Linking the dropdown list and the chart
Scatter chart	The scatter chart shows correlation between payload mass and success for all sites and each site.	Showing correlation between two elements
Callback function for the scatter chart	This function enables the selection at the dropdown list to show a relevant scatter chart for each site.	Linking the dropdown list and the chart
Slider	The selection of the range of payload mass in the slider enables the scatter chart to show relevant plot.	Enabling the selection of a range

GitHub URL : https://github.com/knh444/Capstone-Presentation-for-Data-Science/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- First, I created "dataset_part_3.csv" by applying "features engineering" on "dataset_part_2.csv".
- Second, I loaded these two sets of data as dataframe, which respectively include the Class variable (a target) and the features related the launches (predictors).
- Third, I transformed the target variable into an array using Numpy, and standardized the predictor variables as arrays using StandardScaler.
- Fourth, I split the data sets (variables) into the training set (80%) and the test set (20%) using train_test_split.
- Fifth, I created the object of ML models - Logistic Regression, SVM, Decision Tree & KNN - and trained them using GridSearchCV, finding a model with the best parameters.
- Sixth, I computed the accuracy score of each model using the test set and created the confusion matrix.
- Finally, I compared the outcomes and concluded that Logistic Regression, SVM, KNN are the best models with the same and highest accuracy score.



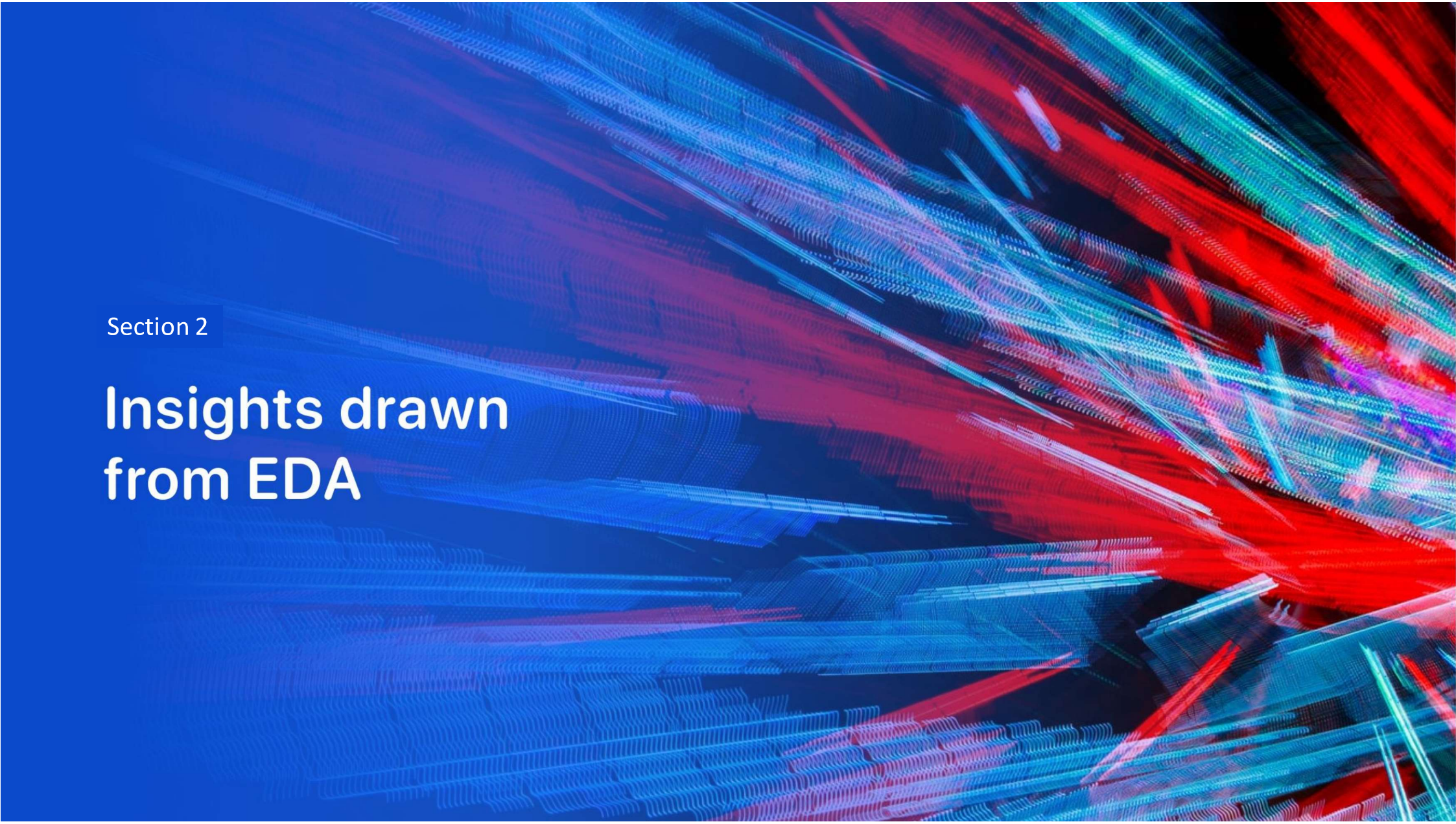
GitHub URL : https://github.com/knh444/Capstone-Presentation-for-Data-Science/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results (EDA & Dashboard Analysis)

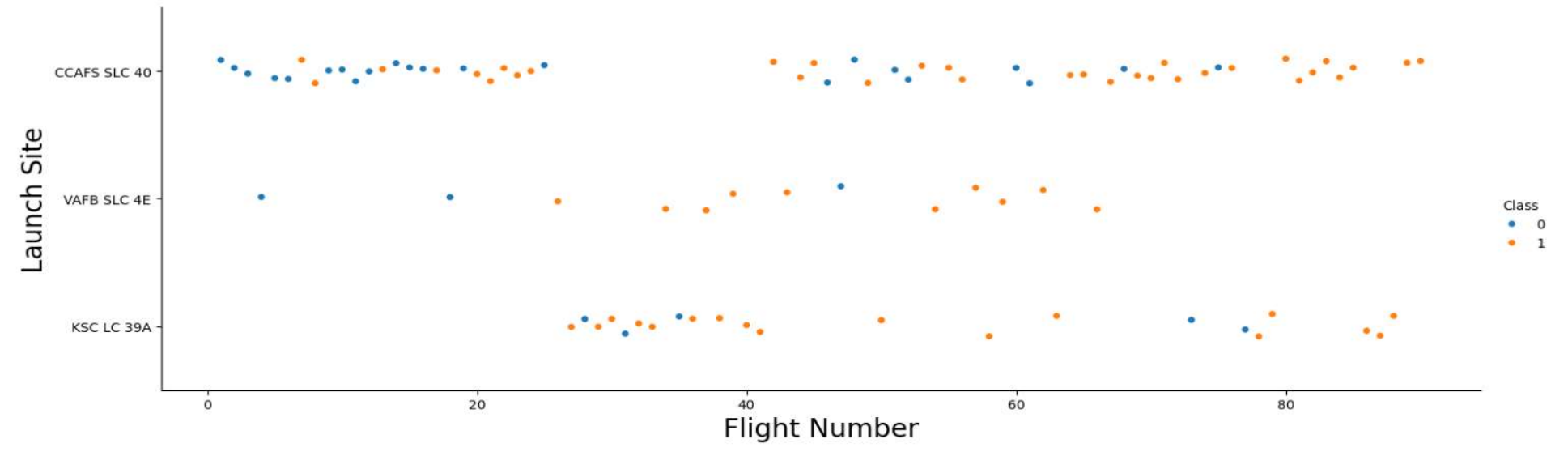
- Falcon 9's mission outcome recorded excellent success rate, while the first stage landing success was not as impressive as mission itself.
- The first stage landing had started to show success from 2013 and the launch success was in upward trend generally since then.
- CCAFS SLC-40 site was most used for Falcon-9 rocket launch and KSC LC-39A site seemed to be used as an alternative due to the high failure rate of the first stage landing in CCAFS SLC-40 in earlier flights.
- CCAFS LC-40 recorded the largest number of landing success (highest ratio), but its success rate is only 26.9%.
- LEO, PO, ISS and GTO orbits were used for early and mid-period launches, while SSO and VLEO orbits were used for later launches and recorded successful first stage landing for heavier payload mass.
- The correlation between payload and landing success seemed somewhat negative.

Results (ML Classification Analysis)

- The analysis found that **the success or failure outcome of the Falcon 9's first stage landing can be predicted with 0.833 % of accuracy** by using **the Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN)** classification models.
- The Decision Tree classification model showed only 0.611% accuracy in prediction.
- Logistic Regression, SVM, and KNN showed the same accuracy score and the identical confusion matrix. They all perform best for the first stage landing prediction.

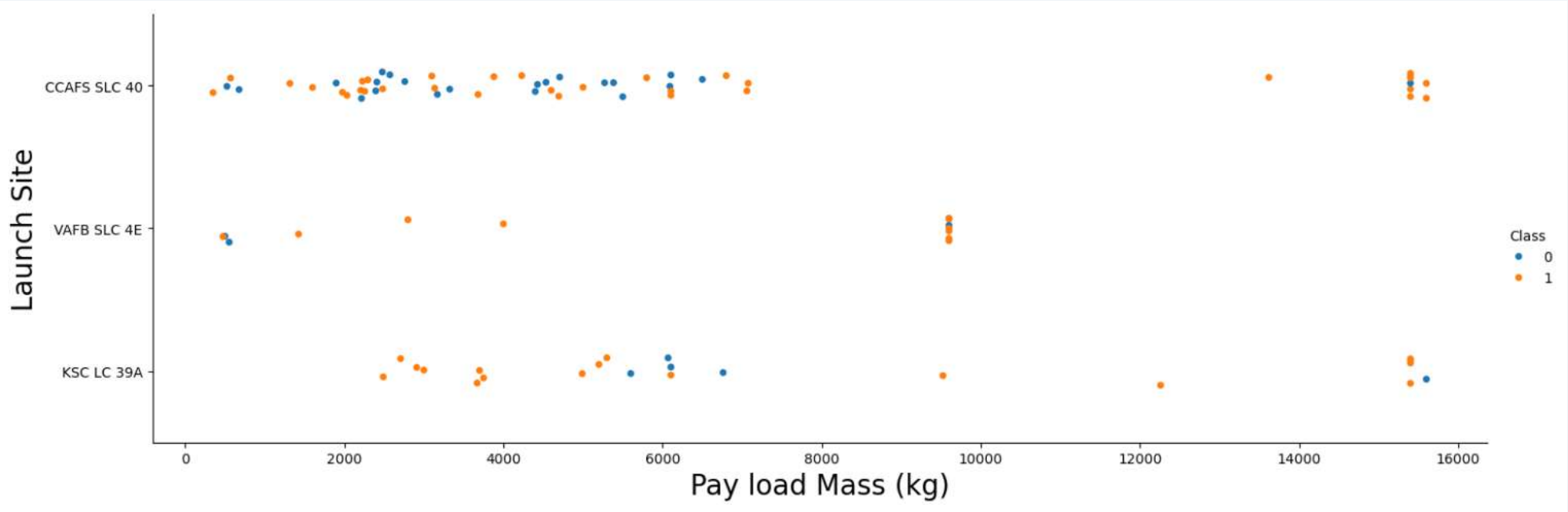

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



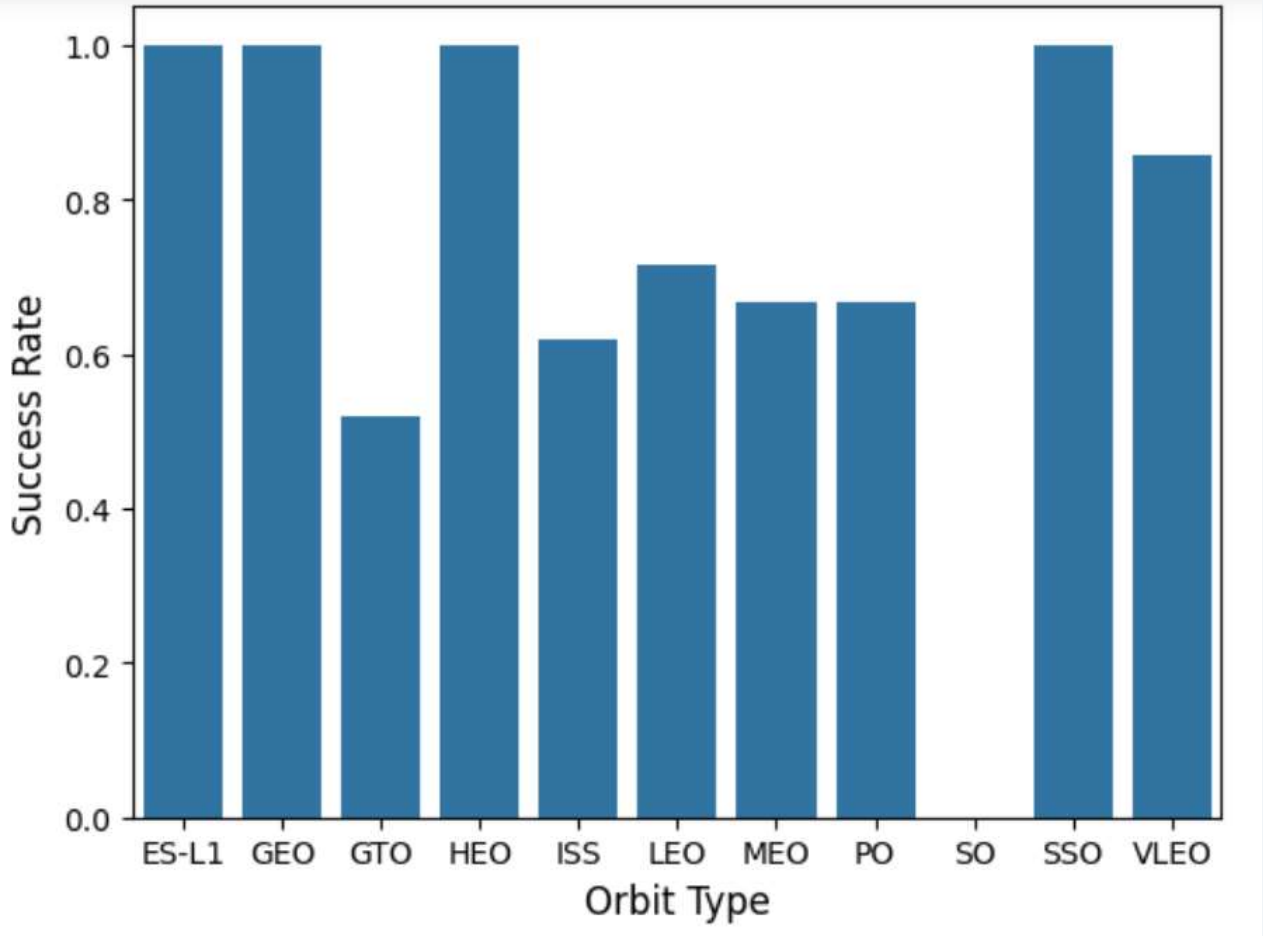
- The scatter chart shows CCAFS was most used among three sites. It was used for earlier flights before flight number 30 and reused after flight number 40. Before flight number 30, CCAFS showes high failure ratio.
- While CCAFS was not used, KSC started to be used; however it was used less frequently after flight number 40 when CCAFS was used again.
- VAFB was used from earlier flights but less frequently than the other sites. It was not used after flight number 70.

Payload vs. Launch Site



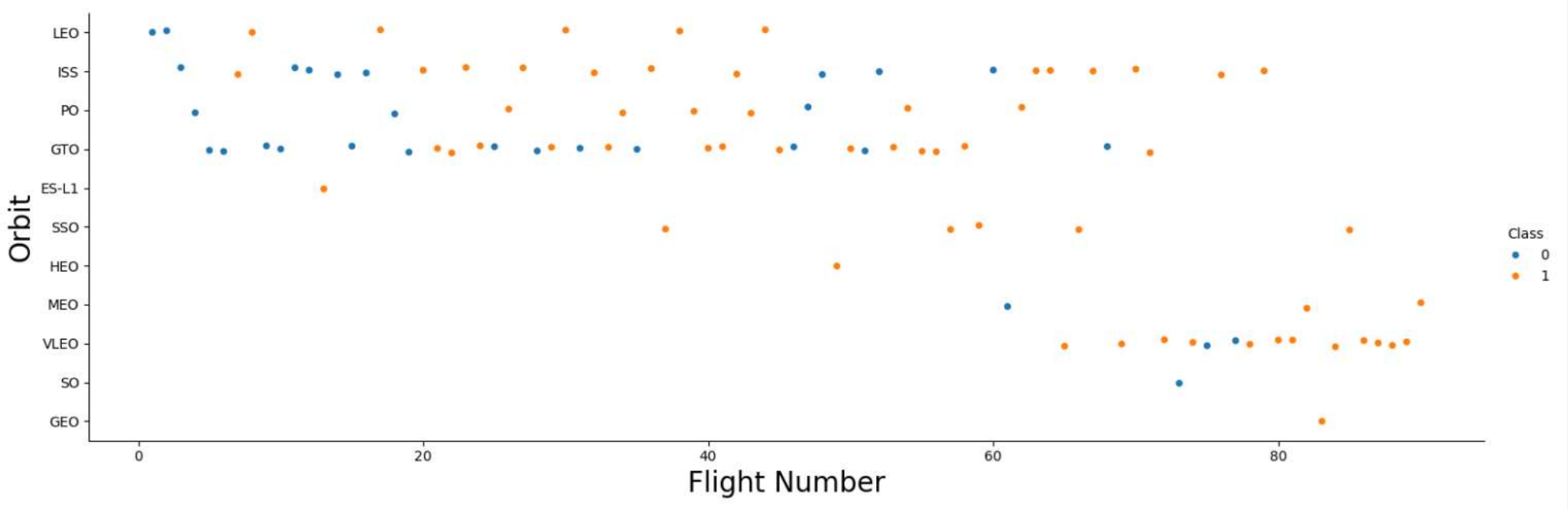
- CCAFS shows higher success rate of the first stage landing for pay load mass heavier than 13000 kg, while VABF hasn't recorded any launch for pay load mass heavier than10000 kg.
- KSC shows relatively high success rate of the first stage landing across the range of pay load mass.

Success Rate vs. Orbit Type



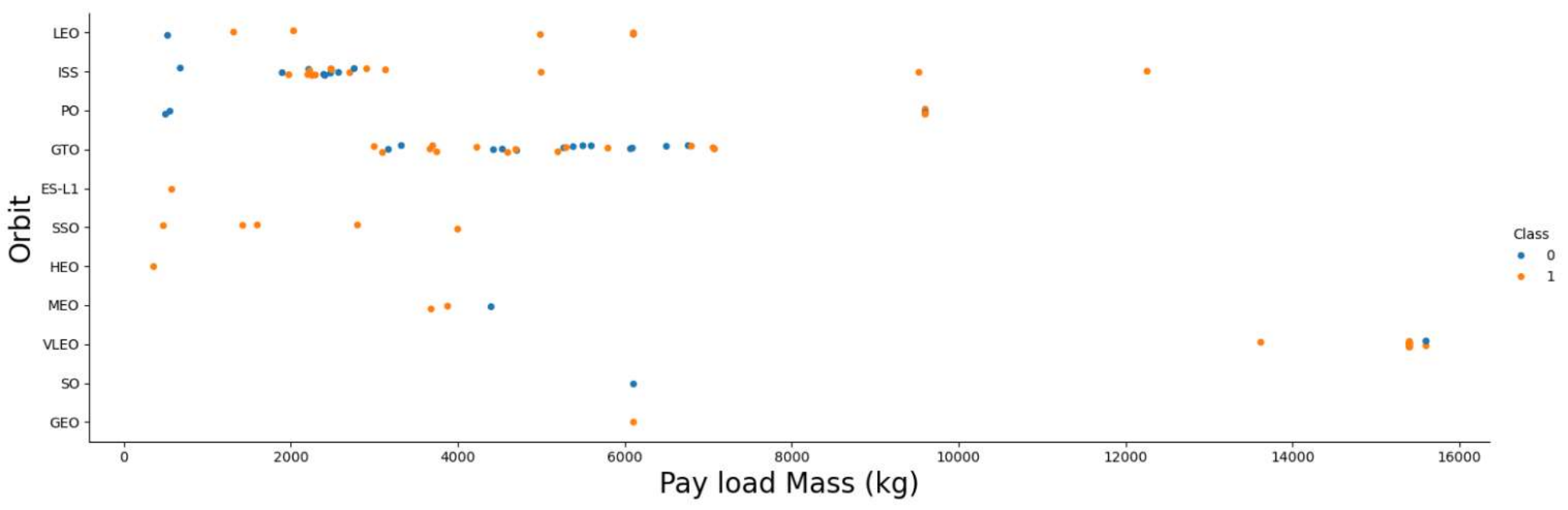
- According the bar chart, ES-L1, GEO, HEO, SSO and VLEO orbits have high success rates.
- However, ES-L1, GEO, HEO orbits have only one launch record respectively; therefore, their first stage landing success rates cannot give us meaningful insights.
- SSO and VLEO orbits have 5 and 10 records respectively; therefore, we can assess that they have somewhat meaningfully high success rate.

Flight Number vs. Orbit Type



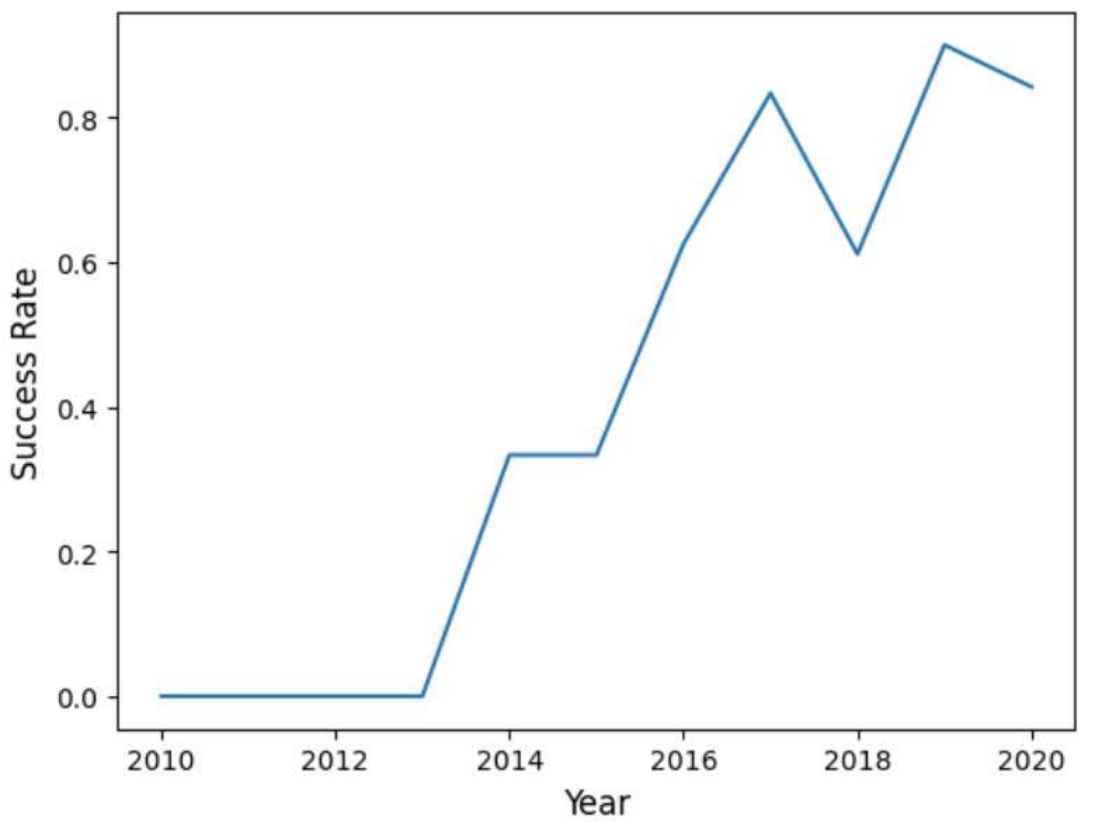
- LEO and PO orbits were used for earlier launches, while VLEO and SSO orbits were used for later launches.
- ISS and GTO orbits were used for a wider range of flight numbers, but not used after flight number 80.

Payload vs. Orbit Type



- VLEO, ISS, and PO orbits recorded successful first stage landing for heavy payload mass beyond 8000kg. Other orbits were used for lighter heavy payload.

Launch Success Yearly Trend



- First stage landing had started to succeed after 2013. Since then, the landing success was in upward trend except setbacks in 2018 and 2020.
- The highest success rate was recorded in 2019.

All Launch Site Names

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

- The left side shows the 4 unique launch sites from the SpaceX data set. They were found using the SQL query.

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 above shows 5 records where launch sites begin with `CCA`. They were found using the SQL query.

Total Payload Mass

Total Payload Mass from NASA
99980

- The total payload carried by boosters from NASA was calculated as 99980 kg as shown above, using the SQL query

Average Payload Mass by F9 v1.1

Average Payload Mass carried by F9 v1.1
2534.6666666666665

- The average payload mass carried by booster version F9 v1.1 was calculated as 2534.67 kg as shown above, using the SQL query.

First Successful Ground Landing Date

First Success Date
2015-12-22

- The above indicates the dates of the first successful landing outcome on ground pad as December 22nd 2015. It was found using the SQL query.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- The left side shows the list of 4 boosters which have successfully landed on drone ship and had payload mass greater than 4000 kg but less than 6000 kg. They were found using the SQL query.

Total Number of Successful and Failure Mission Outcomes

Mission Outcome	Total
Successful mission	100
Failure mission	1

- The left side shows the total number of successful and failure mission outcomes out of 101 outcomes. They were found using the SQL query.

Boosters Carried Maximum Payload

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

- The left side shows the list of the boosters which have carried the maximum payload mass. They were found using the SQL query.

2015 Launch Records

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

- The left side shows the list of the failed landing outcomes in drone ship, their booster versions, and launch site names in 2015. They were found using the SQL query.

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

Landing_Outcome	Total
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

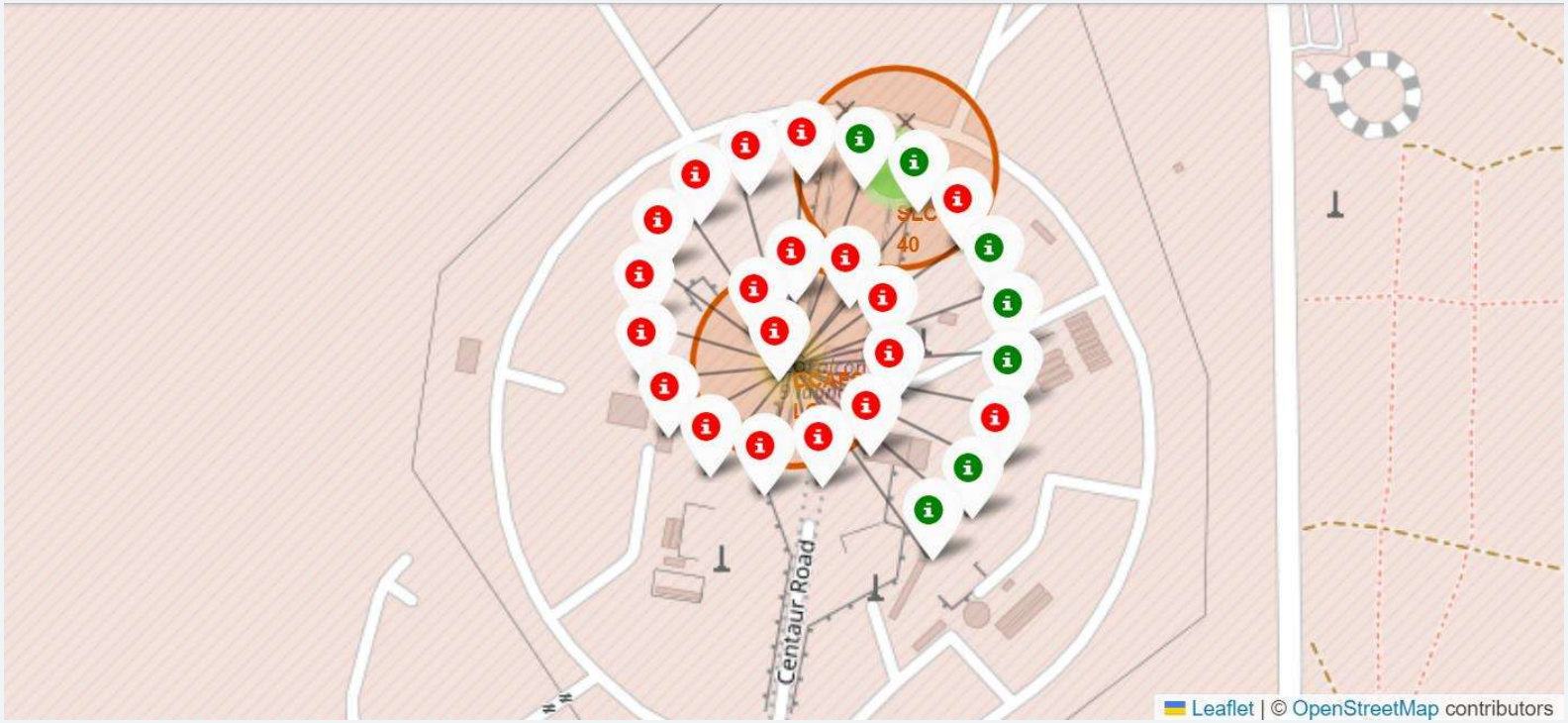
- The left side table ranks the count of landing outcomes, such as Failure (drone ship) or Success (ground pad), between the date 2010-06-04 and 2017-03-20, in descending order.
- The table was called using the SQL query.


Section 3
Launch Sites
Proximities Analysis

All launch sites' location markers

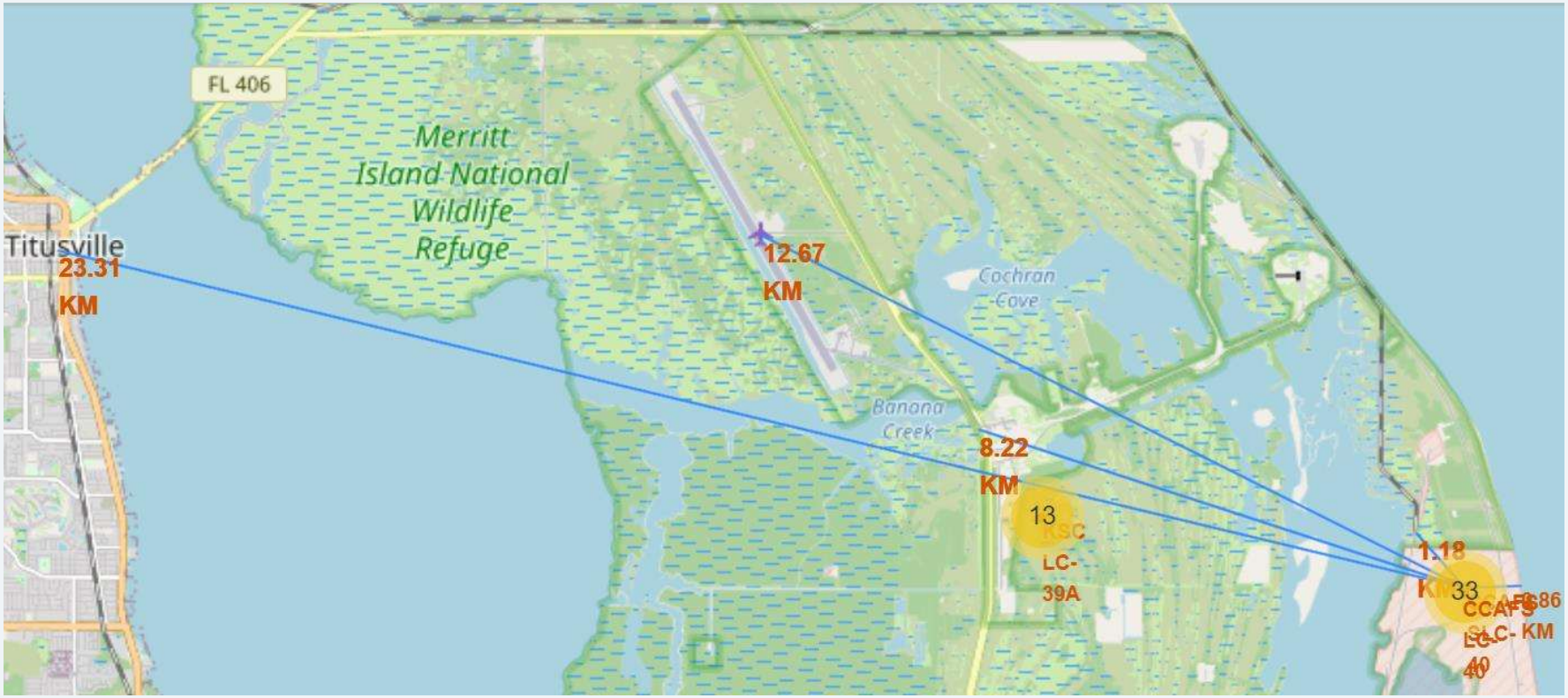
- The red dots (small circles) on the above US map shows 4 launch sites of the Falcon 9 rocket, one in California and three in Florida. They were found in areas between latitudes 28 North and 34 North.

Color-labeled Landing Outcomes at CCAFS LC-40

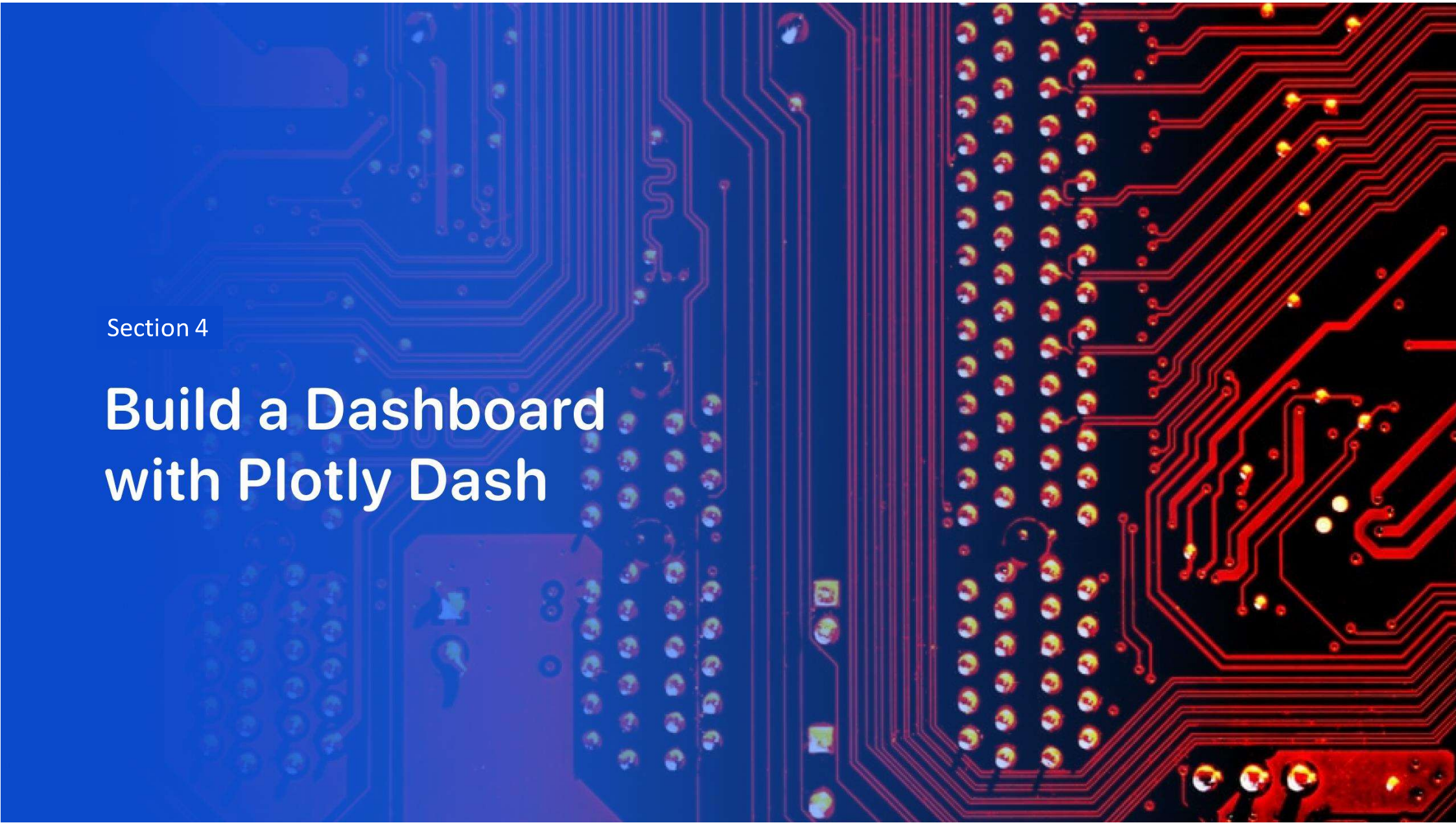


- The above captured map shows the first stage landing outcomes for the launches at the site CCAFS LC-40 in Florida. Out of 26 launches at the site, 7 succeeded (green) and 19 failed (red) in the first stage landing based on the data from 'spacex_launch_geo.csv'.

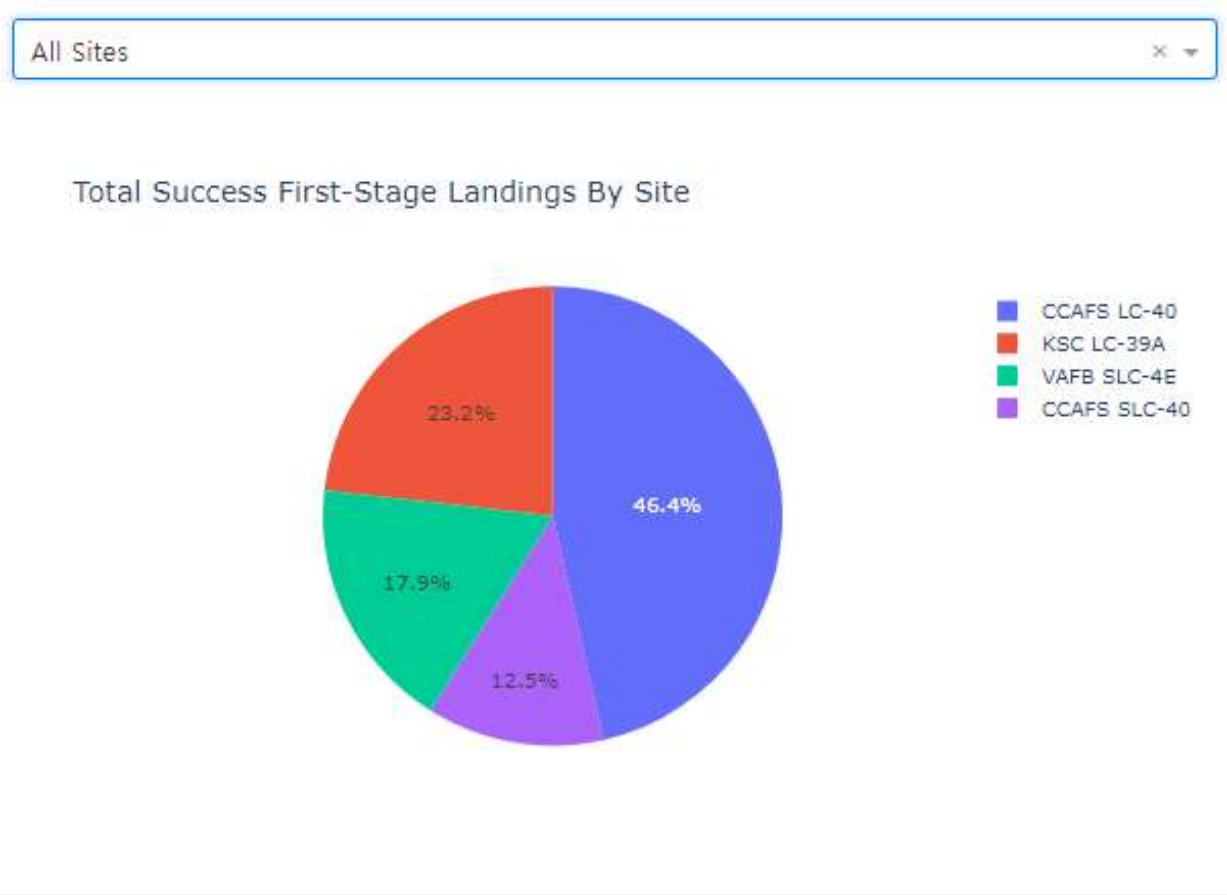
Distance between Landmarks and CCAFS SLC-40



- The blue lines indicate the shortest distances between the launch site CCAFS SLC-40 in Florida and the nearest landmarks (a city, an airport, a highway, and a railway). The site is 23.31 km away from Titusville, 12.67km away from the airport, 8.22km away from the highway, and 1.18km away from the railway.

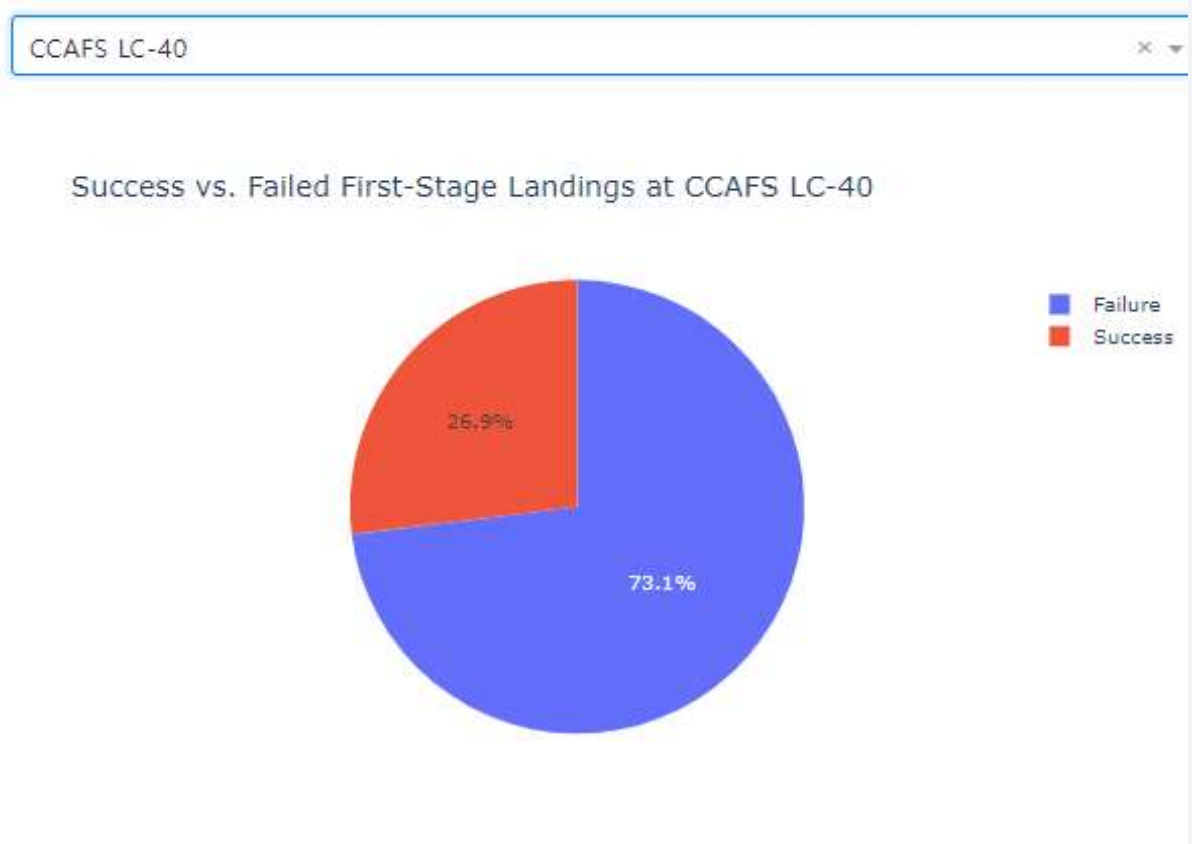

Section 4
Build a Dashboard with Plotly Dash

Total Success Landing Ratio by Site



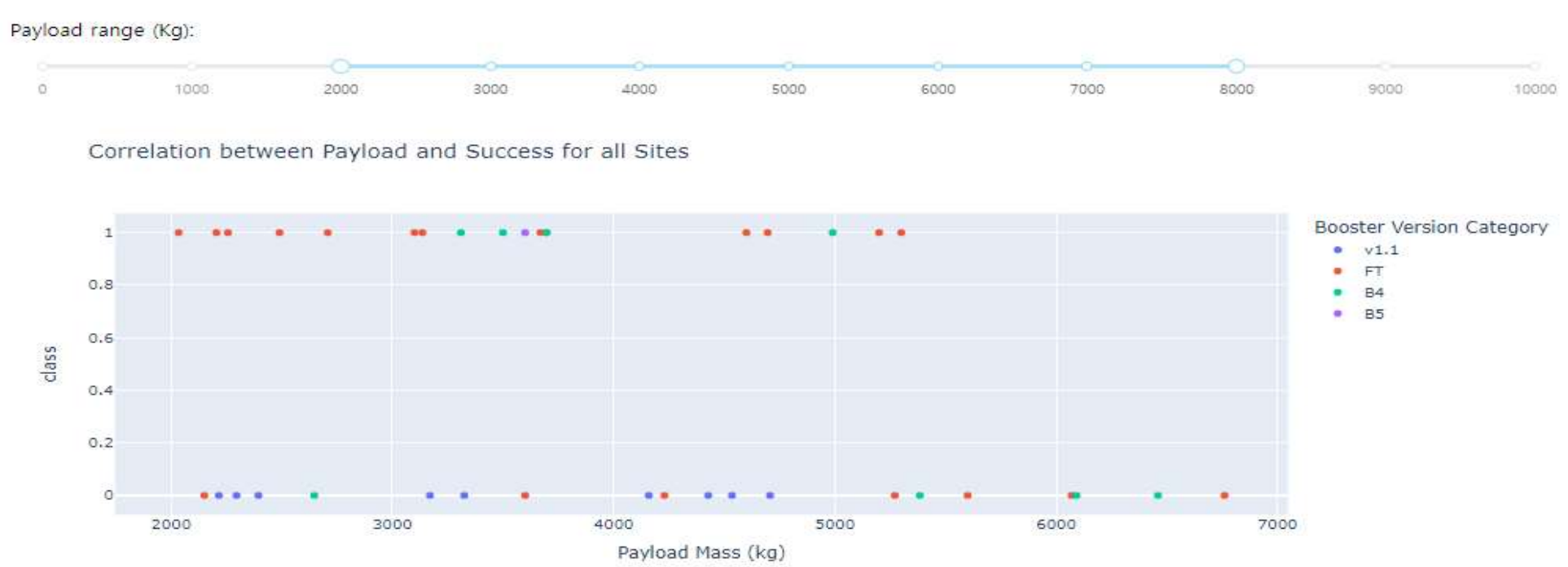
- The pie chart shows the ratio of the first stage landing success count of each site against the launch success count of all sites based on the data from ‘spacex_launch_dash.csv’.
- CCAFS LC-40 recorded the largest number of landing success (highest ratio), while CCAFS SLC-40 the smallest number (lowest ratio).

Launch Site with Highest Success Ratio



- The pie chart shows the success-failure landing ratio for CCAFS LC-40, the launch site with the highest landing ratio found in the previous page.
- Successful first stage landings account for 26.9% of the missions, and failure landings 73.1% at the site.

Payload vs. Landing Outcome Scatter Plot for All Sites

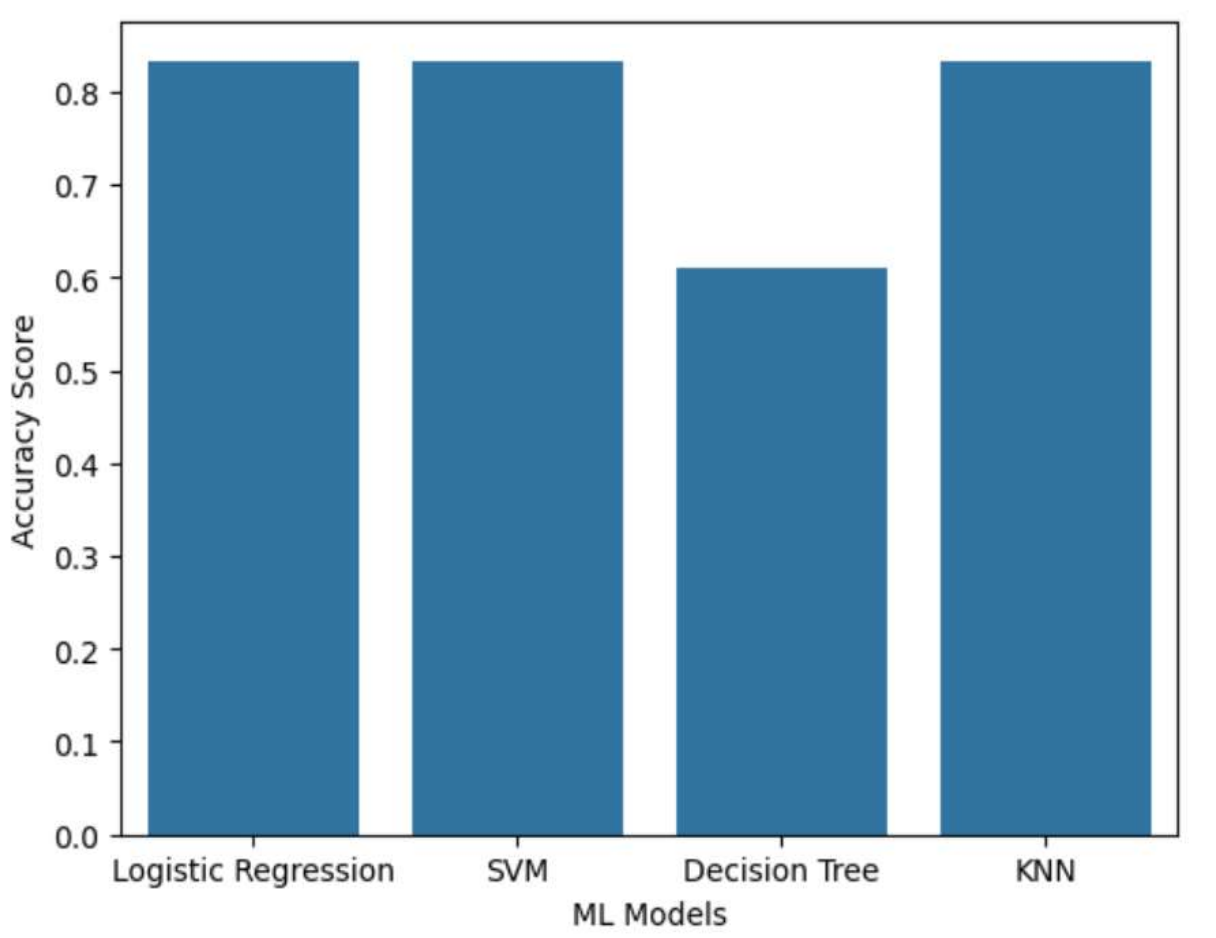


- This scatter chart shows correlation between payload mass and landing success for all sites in the range of payload mass between 2000 kg and 8000 kg, hue to the booster version category.
- The correction is somewhat negative. When payload is between 5500 kg and 8000 kg, no successful landing is recorded based on the data from 'spacex_launch_dash.csv'.

Section 5

Predictive Analysis (Classification)

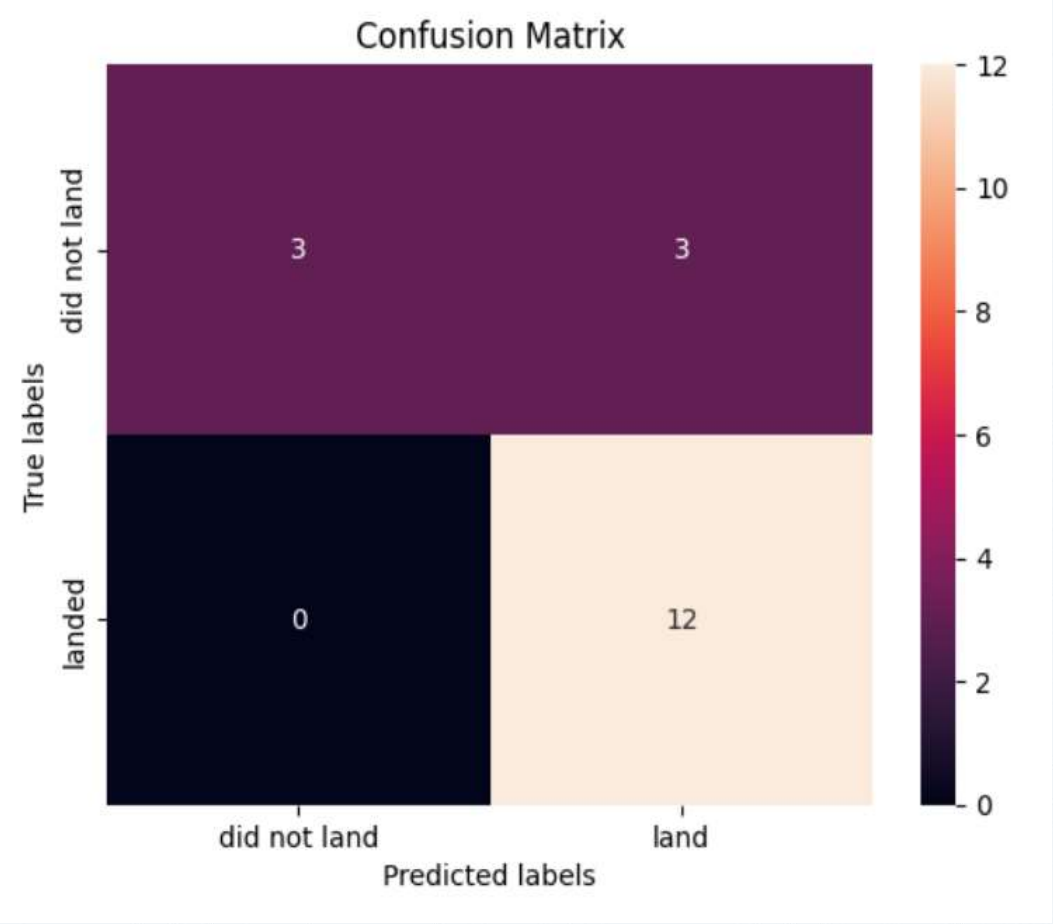
Classification Accuracy



- Logistic Regression, SVM, KNN have the same and highest accuracy score(0.833) on test samples.
- They all perform best.

- Decision Tree's accuracy score was only 0.611.

Confusion Matrix



- This is the confusion matrix of KNN, one of the best performing models.
- The KNN model predicted all correctly out of its 3 'failure' predictions, while it predicted 12 correctly out of its 15 ‘success’ predictions.
- Accuracy score is 15/18 = 0.833

Conclusions

- Point 1 : The Falcon 9's first stage landing outcome may be predicted by using the machine learning technique based on SpaceX launch data with accuracy of more than 80%.
- Point 2 : The best machine learning model for this analysis is Logistic Regression, Support Vector Machine, and K-Nearest Neighbor classification models, not Decision tree.
- Point 3 : The first-stage landing success rate seems not so high as that of the Space-X mission itself, even though it seems increasing with the passage of time.
- Point 4 : Launch sites, payload mass, or orbit types may have influence on the landing success rate. Researchers can perform further study on such relationships in the future to assess the merit of SpaceX more precisely.

Appendix

- Visit the following Github URL to find notebooks, python files and data files I created or used during this project.
- URL : https://github.com/knh444/Capstone-Presentation-for-Data-Science

Thank you!