IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Koji Sera
Apr 18th, 2025

Outline

This presentation is composed of the following chapters:

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

Executive Summary

- Summary of methodologies
  - Data collection through API and web scraping (Falcon 9 launch data)
  - Data wrangling
  - Exploratory Data Analysis with visualization and SQL
  - Interactive visual analytics with Folium and Plotly Dash
  - Predictive data analysis with classification models
- Summary of all results
  - EDA result with visualization and SQL
  - Interactive visualization results with Folium Map and Plotly Dash
  - Predictive data analysis results with SVM, KNN, Decision Tree and Logistic Regression

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches with 62 million dollars while others can do with 165 million dollars. SpaceX can save more than 100 million for each launch only if Falcon 9 can reuse the first stage.
  - As one step before building a model to predict the first stage landing, we focus on prediction of successful launch.
- Problems to find answers
  - We want to build a model to predict if Falcon9 can launch successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web scraping and API were used for Falcon 9 launch records
- Perform data wrangling
  - 8 categorical values of landing outcome were reassigned to 0 (bad) or 1 (good) landing as 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
  - 4 categorical supervised ML tools (SVM, K-nearest neighbors, Decision Tree, and Logistic Regression) were used with parameter tuning. Confusion matrix was used for model accuracy evaluation.

Data Collection

- Data collection methods

    1. WEB SCRAPING (Beautiful Soup)

        - Extract a HTML table of Falcon 9 launch data from Wikipedia
        - Parse the table for the items (Flight No, Launch Site, Payload, Payload mass, Orbit, Customer, Launch outcome, Booster Version, Booster landing, Date, and Time)
        - Convert the table to a data frame

    2. API (GET REQUEST)

        - Send a get request to Space X for the item ID information
        - Convert it to pandas data frame with json_normalize( )
        - Send another set of get request to extract the needed items based on the item ID
        - Convert it to pandas data frame

Data Collection – SpaceX API

- A get request was sent to SpaceX REST for ID numbers for each item. Response was converted to a data frame. Then based on the data frame, another set of get request was sent to SpaceX REST for the exact name/value of each column.
- GitHub URL: https://github.com/G-flatminor/Capstone-project-Space-X-/blob/main/jupyter-labs-webscraping.ipynb

Send API to get ID numbers for the required items
(format example: 5e9d0d95eda69955f7…)

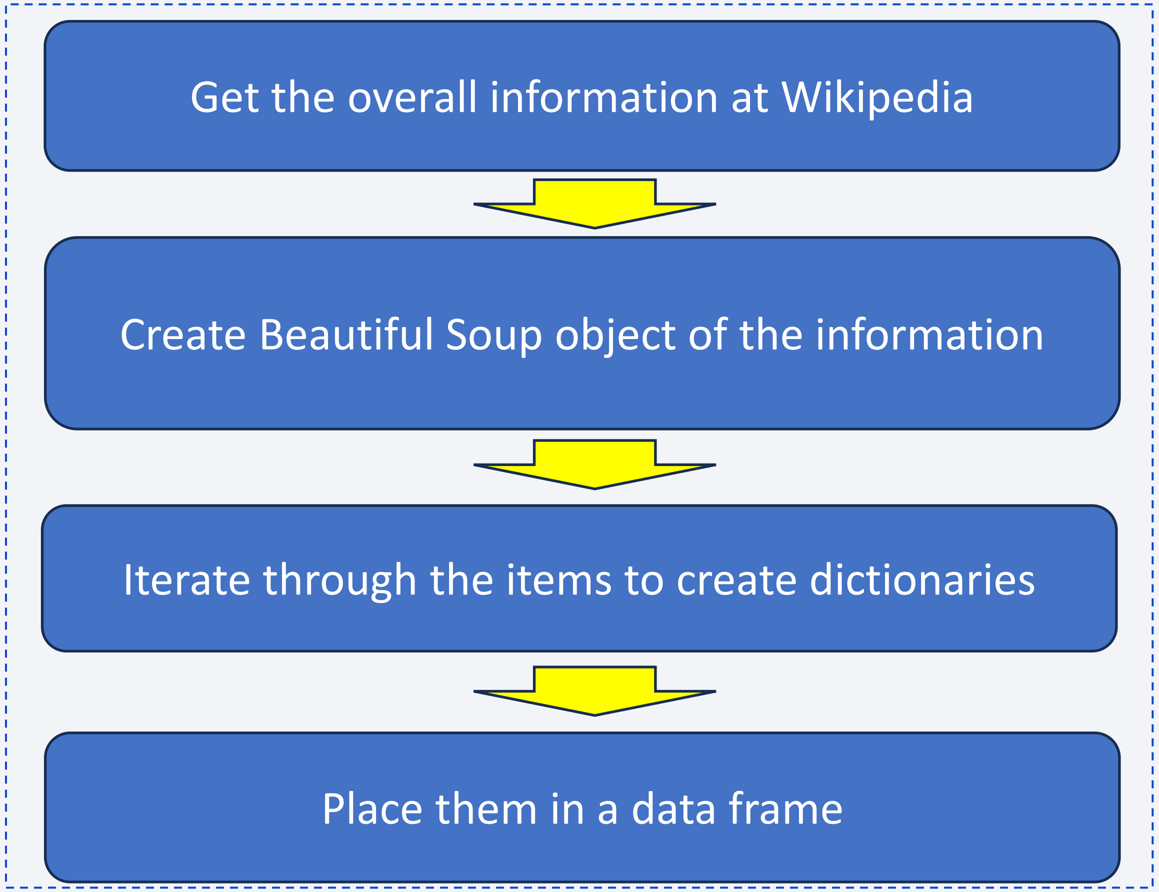
Send another set of APIs with ID to get exact
name/values based on item ID
(name example: booster version: Falcon 9)

Place them in a pandas data frame

Filter the data frame with Falcon 9

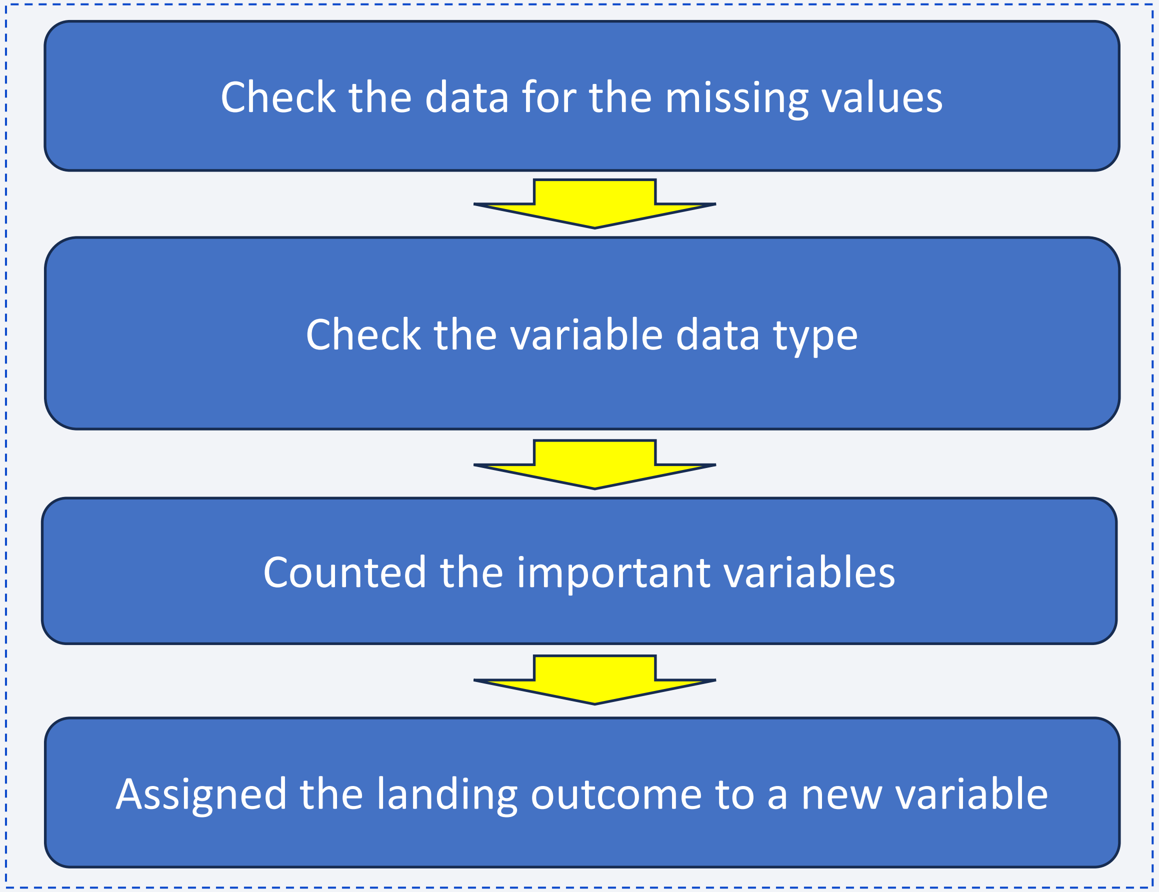
Data Collection - Scraping

- A HTML table of Falcon 9 launch data was extracted from Wikipedia. The extracted table was parsed for the items (Flight No, Launch Site,Payload mass, Orbit, etc.). The data was converted to a data frame.
- GitHub URL: https://github.com/G-flatminor/Capstone-project-Space-X-/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- The data was checked for the missing values and data type, first. Then important variables were counted for each category with value_counts( ): Launch Site, Orbit, and Outcome.
- The outcome has 8 categories, and each one was assigned to a new variable, “Class” (1: success or 0: failure).
- GitHub URL: https://github.com/G-flatminor/Capstone-project-Space-X-/blob/main/labs-jupyter-spacex-Data%20wrangling-v2.ipynb



EDA with Data Visualization

- Seaborn scatter plot is used for a relationship between:
  - Flight Number vs Payload mass (kg)
  - Flight Number vs Launch site
  - Launch site vs Payload mass
  - Flight Number vs Orbit type
  - Payload Mass vs Orbit type
- Seaborn bar chart is used to show success rate by orbit type.
- Seaborn line chart shows success rate transition on a yearly basis.
- GitHub URL: https://github.com/G-flatminor/Capstone-project-Space-X-/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- SQL database can be created or imported with a command from jupyter notebook:
  - %load_ext sql (# loading external database)
  - %sql sqlite:///my_data1.db (# create a database named my_data1.db)
  - df = pd.read_csv(“example.csv")df.to_sql("SPACEXTBL", con, if_exists='replace', index=False, method="multi") (# import “example.csv” as “SPACEXTBL”)
- Once imported, you can handle SQL database from jupyter notebook:
  - %sql select * from SPACEXTBL (# call all variables from database)
- GitHub URL: https://github.com/G-flatminor/Capstone-project-Space-X-/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium.Marker is used for Space X launch site in the map.
- Folium.Circle is located in the map for the launch site area.
- Marker clusters are used to show all launches with its success/failure in each launch site.
- Lines are drawn with distance between one launch site and;
    1. Railway
    2. Highway
    3. Coast
    4. City
- Folium map enables users to zoom in and out of the map and to see the geographical conditions of the launch sites.
- GitHub URL: URL:https://github.com/G-flatminor/Capstone-project-Space-X-/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

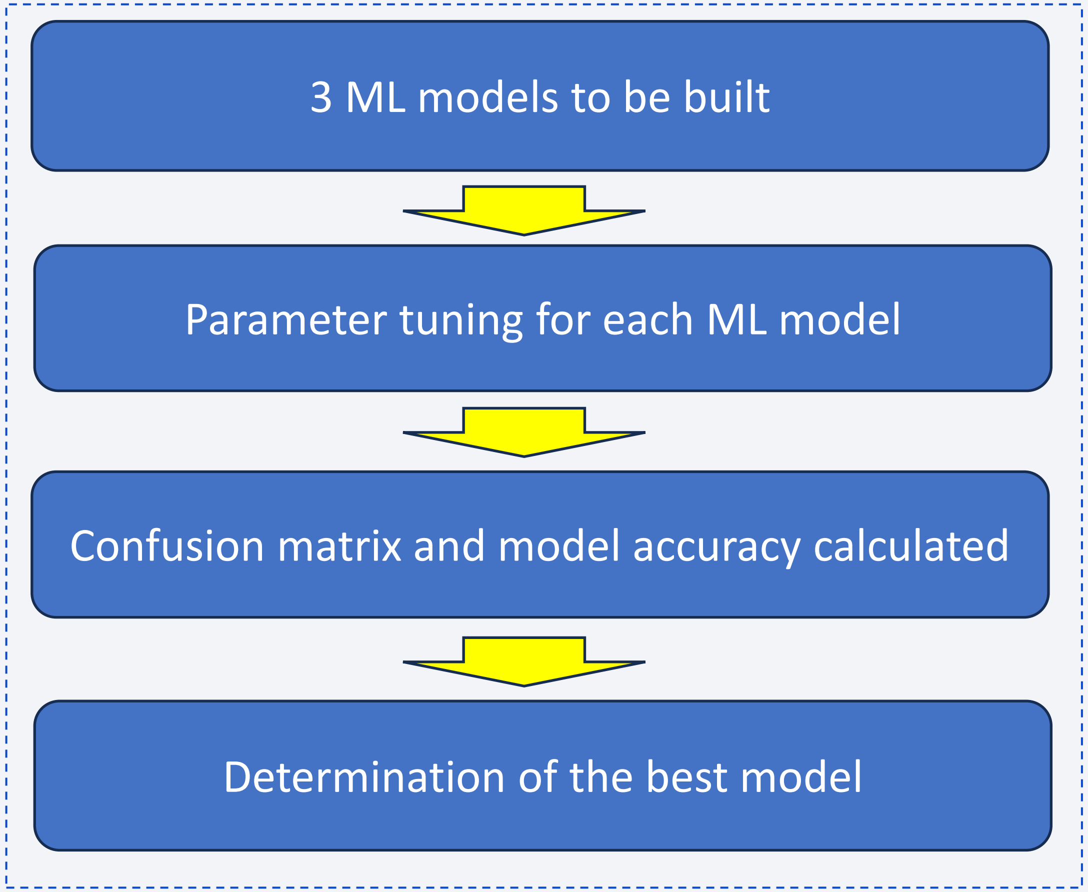
- Plotly Dash dashboard is composed of ;

    1. A dropdown menu for launch sites
    2. A payload slider
    3. Success/Failure rate for all sites (aggregated all sites or independent site) in a pie chart
    4. Payload mass vs Landing outcome for all sites (aggregated all sites or independent site) in a scatter plot

- The dashboard provides an easy understanding of relationships among payload * success/failure rate * launch sites.
- GitHub URL: https://github.com/G-flatminor/Capstone-project-Space-X-/blob/main/dash_spacex.py

Predictive Analysis (Classification)

- We try to build a model to predict success/failure of Falcon 9 first stage landing through three different machine learning tools. The model will be modified with parameter tuning.
- Each of those three models will complete confusion matrix for the success/failure classification accuracy to determine the best model.
- GitHub URL: https://github.com/G-flatminor/Capstone-project-Space-X-/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb



Results

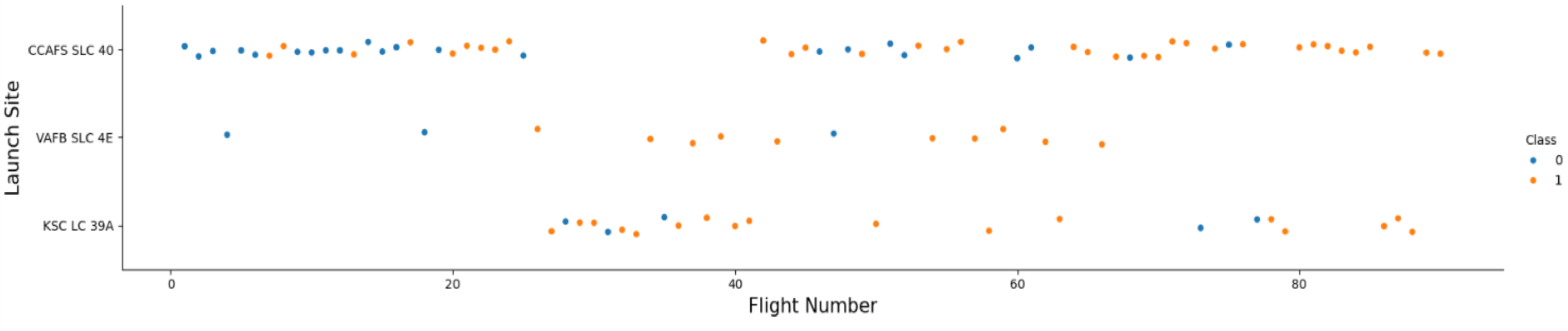
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

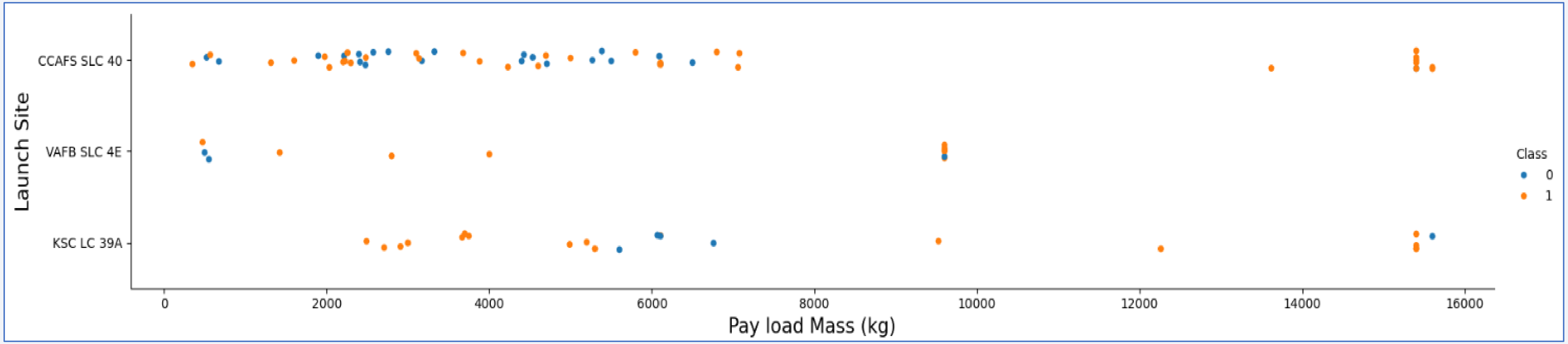
Flight Number vs. Launch Site

- CCAFS SLC 40 has most launches and longest history of Falcon 9 launches.
- Most of the first 20 launches are failed.
- As they launched more, their success rate of the first stage landing has become higher. The launch No. 80 and above are all success.



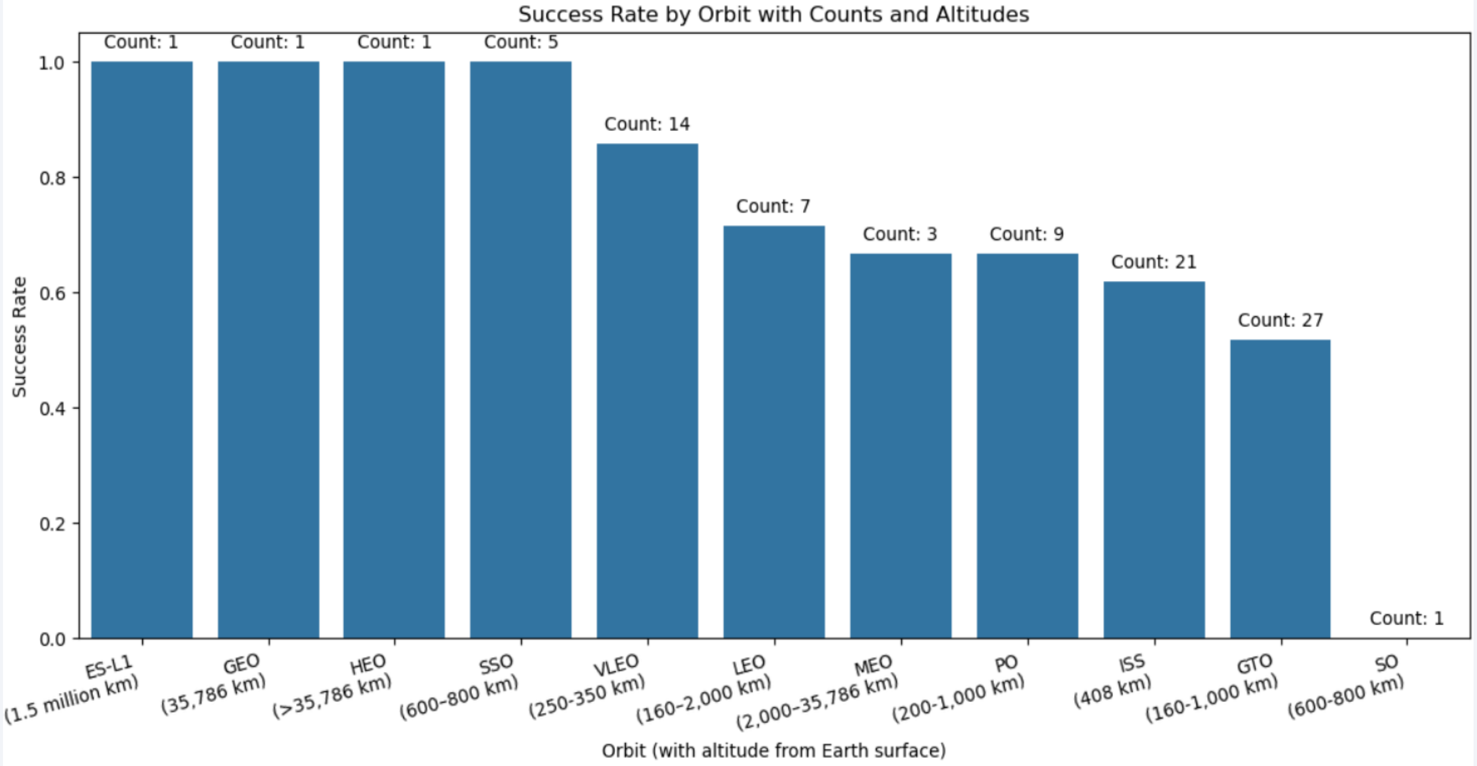
Payload vs. Launch Site

- Most common payload range is 0 through 7,000 kg. Most failures are seen in the same range.
- It seems Space X has tried less payload in the early stage of trials and has increased the payload mass in the later stage.



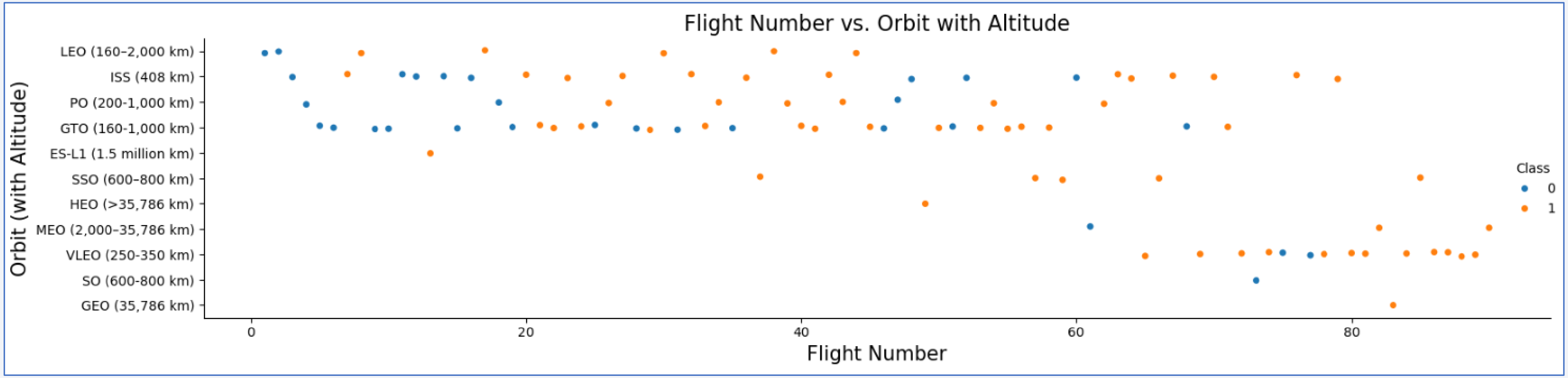
Success Rate vs. Orbit Type

- A bar chart for the success rate of each orbit type is shown on the right along with the number of launches and orbit altitude (km).
- Success rate is influenced by the orbit altitude and the number of launches.
- Space X launched more to LEO/VLEO/ISS/GTO for Starlink products or other satelites.



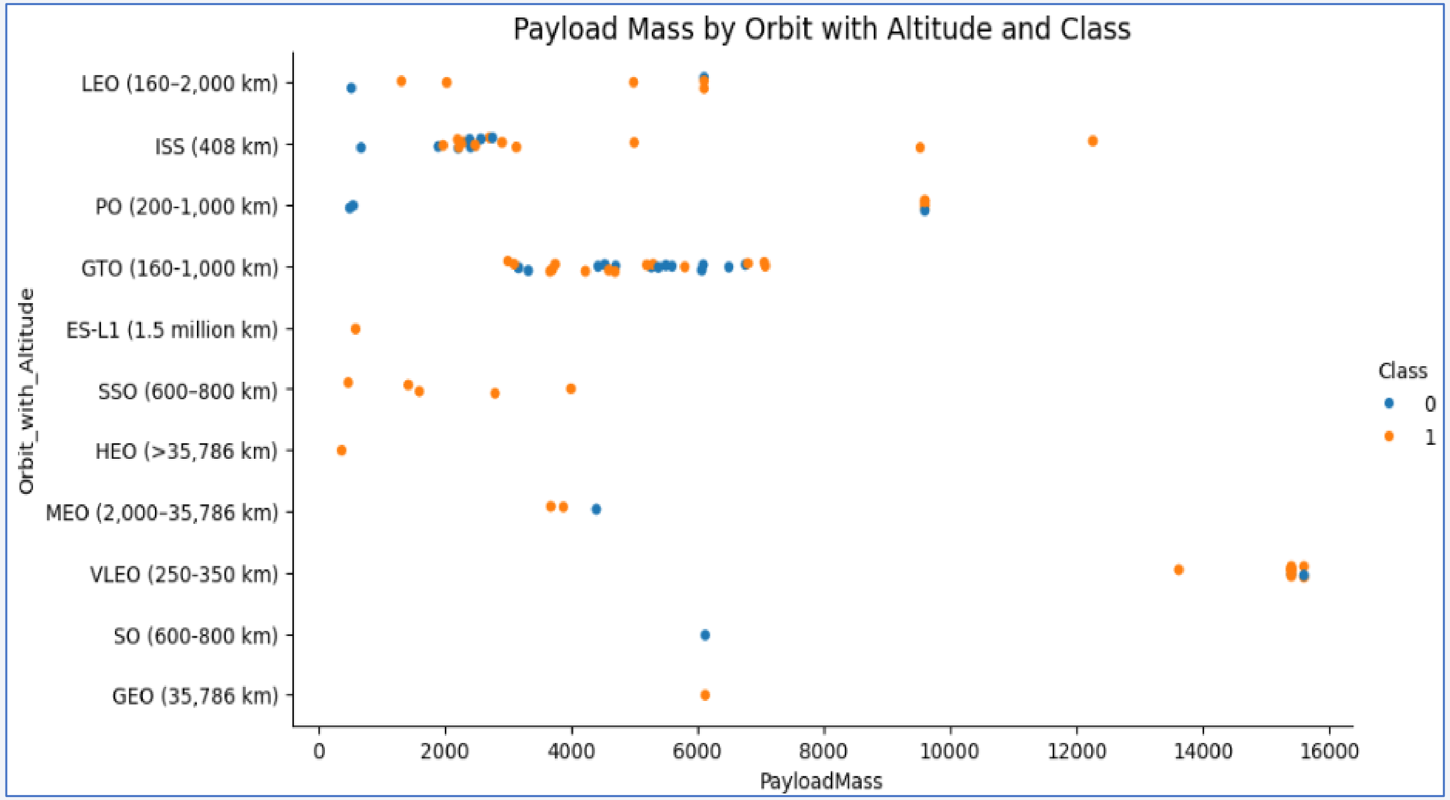
Flight Number vs. Orbit Type

- Most launches were conducted less than altitude 450 km, which are used for satellites and Starlink.
- Falcon 9 first stage was failed to land in the earlier flight (flight No. 20 or less).
- As they launched more, their success rate got higher.



Payload vs. Orbit Type

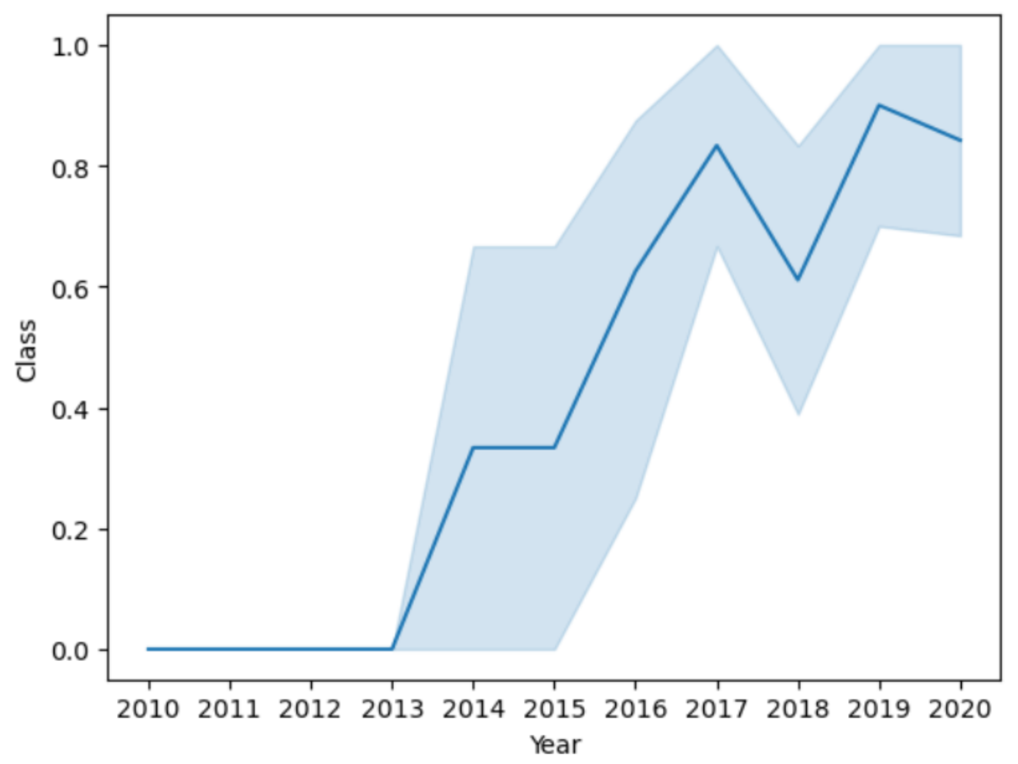
- Overall number of launches are mostly concentrated in their practical purposes which are ISS, Starlink, and satellite.
- While the payload for ISS is rather lighter (2,000 – 3,000 kg), the ones for GTO are heavier (3,000 – 7,000 kg). The ones for VLEO are extremely heavy (15,000 – 16,000 kg).
- Many failures occurred either ISS or GTO.



Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- Show the screenshot of the scatter plot with explanations



All Launch Site Names

- “select distinct” command found four launch sites;
    1. CCAFS LC-40
    2. VAFB SLC-4E
    3. KSC LC-39A
    4. CCAFS SLC-40

```
%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'

- The following command was sent to SQLITE database to display the first 5 records where launch sites begin with ‘CCA’:

```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5

 * sqlite:///my_data1.db
```

- The following 5 records were returned. The launch was successful, but the landing wasn’t.

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

Total Payload Mass

- The following command was sent to calculate the total payload carried by boosters from NASA.

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)';
```

- The total payload carried is 45,596 kg.

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The following command was run to calculate the average payload mass carried by booster version F9 v1.1.

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
```

- The average payload mass is 2928.4 kg.

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The following command was run to find the dates of the first successful landing outcome on ground pad.

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome = 'Success';
 * sqlite:///my_data1.db
```

- The query found out the first successful landing date is July 22, 2018.

min(Date)
2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The following command was run to 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 SPACEXTBL where Landing_Outcome = 'Success (drone ship)' AND Payload_Mass__KG_ < 6000 AND Payload_Mass__KG_ > 4000;
```

- 4 boosters were found to meet the criteria as seen below.

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 original Mission Outcome column showed four different categories. So the outcome needs to be re-categorized to ‘Success’ or ‘Failure’.
- The query on the right was run to create a new column and count the total number of successful and failure mission outcomes.

```
%%sql
Select
    Case
        When Mission_Outcome Like '%Success%' Then 'Success'
        When Mission_Outcome Like '%Failure%' Then 'Failure'
        Else 'Other'
    End As Outcome_Category,
    Count(*) As Total
From SPACEXTBL
Group by Outcome_Category;
```

- The query result is shown on the right.

Outcome_Category	Total
Failure	1
Success	100

Boosters Carried Maximum Payload

- The following query is to list the names of the booster which have carried the maximum payload mass. A subquery was used to do this task.

- The query result is shown on the right.

```
%%sql Select
    Booster_Version, Payload_Mass__KG_
    from SPACEXTBL
    Where Payload_Mass__KG_ = (select max(Payload_Mass__KG_) from SPACEXTBL)
    order by Booster_Version;
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

- The following query is to list the failed landing_outcomes in drone ship, the month names, their booster versions, and launch site names for in year 2015.

```
%%sql select substr(Date, 6,2) as Month, Booster_Version, landing_outcome, Launch_site
    from SPACEXTBL where substr(Date, 0,5)='2015' and Landing_Outcome = 'Failure (drone ship)';
```

- The query found two records that meet the criteria, one for Jan, 2015 and the other for Apr, 2015.

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

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

- The query on the right was run to 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(Landing_Outcome) as Outcome_Count
    from SPACEXTBL where Date between '2010-06-04' and '2017-06-04'
    group by Landing_outcome order by Outcome_Count DESC;
```

Landing_Outcome	Outcome_Count
No attempt	11
Success (drone ship)	6
Success (ground pad)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- The query result is shown on the right.

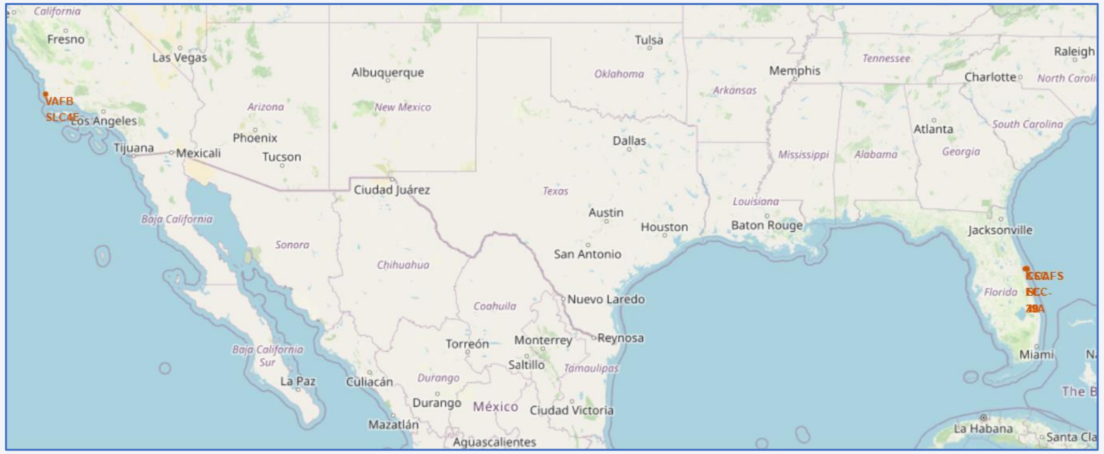
Section 3

Launch Sites Proximities Analysis

Launch Sites in Folium Map

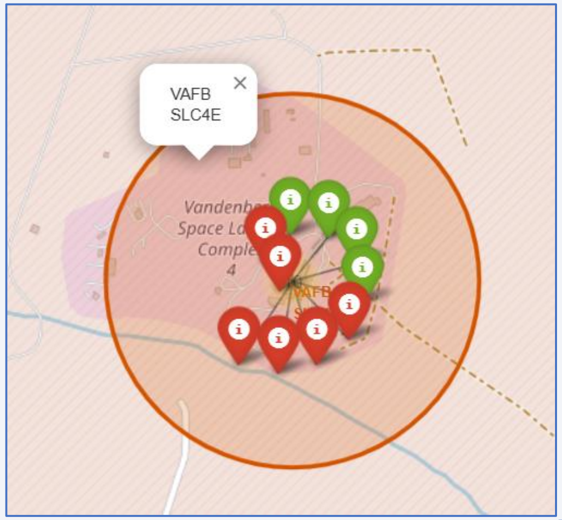
- Four launch sites are located in Folium Map with markers and circles. All sites are close to the ocean.

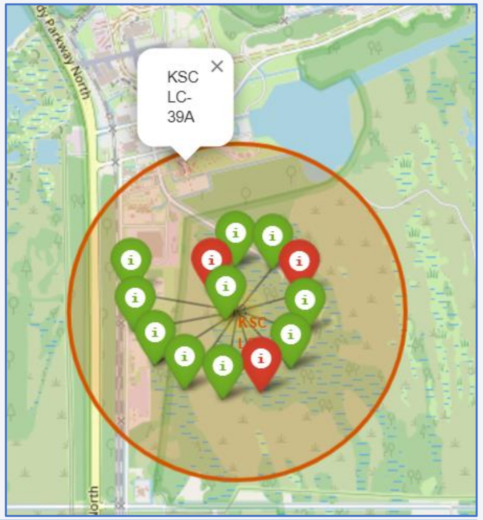
- A closer look at three sites in Florida provides a better view of circles and markers of each site.
- CCAFS LC-40 and CCAFS SLC-40 are very close to each other, so smaller size of circles were applied.

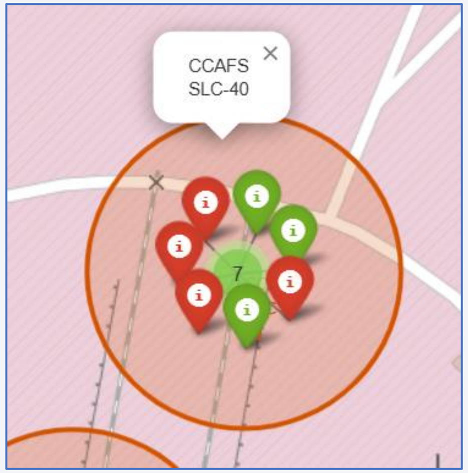


Launch Outcome in Folium Map

- Four launch sites are shown respectively with marker cluster to show launch outcome, success or failure.
- Green markers indicate success, and red markers mean failure.







Proximity to major items in Folium Map

- A launch site was connected with major items (closest highway, closest coastline, closest railroad, and closest city) with a line along with its distance.
- Distance from CCAFS LC-40:
    1. Highway: 0.65km
    2. Coastal Line: 0.92km
    3. Railroad: 0.95km
    4. City: 17.46km
- Top 3 items are close to the site while the last one, city, is away from the site.

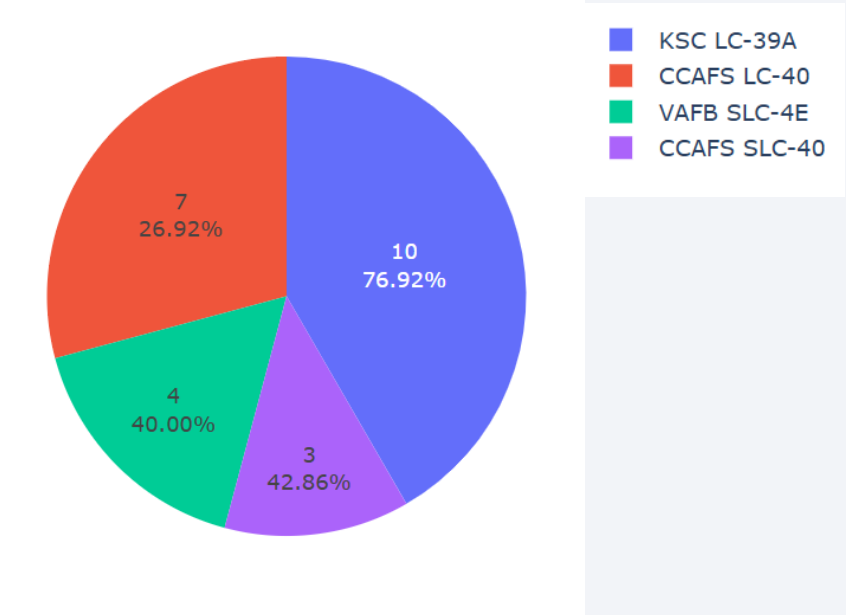
Section 4

Build a Dashboard with Plotly Dash

Successful Launch Count by Site in Dashboard

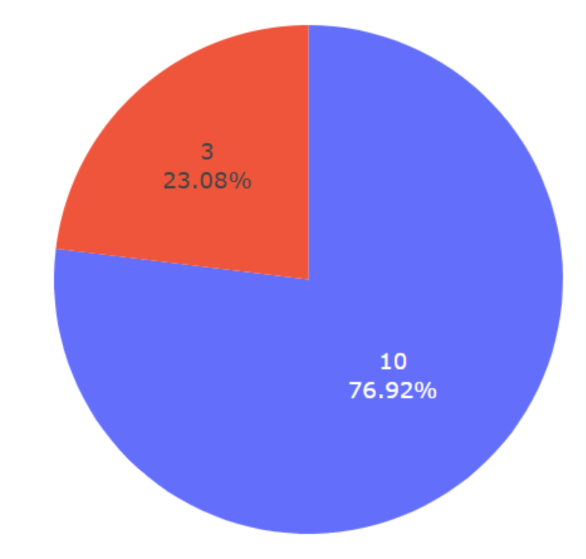
- The screenshot of launch success count and success ratio for all sites is shown in a pie chart.

- KSC LC-39A has succeeded the most launches (10 launches) of four sites, and their success rate is also the highest (76.92%)

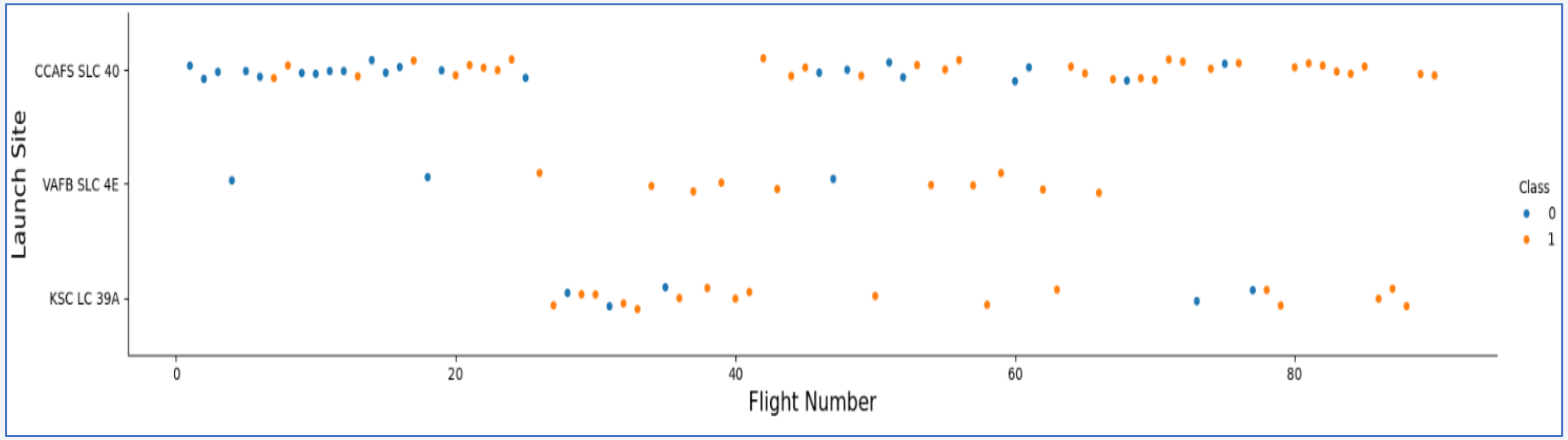


Highest Success Rate Site

- KSC LC-39A shows the highest launch success ratio (76.92%).
- This may be because the site started their launch after the first 20 launches which experienced the highest failure rate.



Success vs Failure for KSC LC-39A



Success Launches with Payload Mass

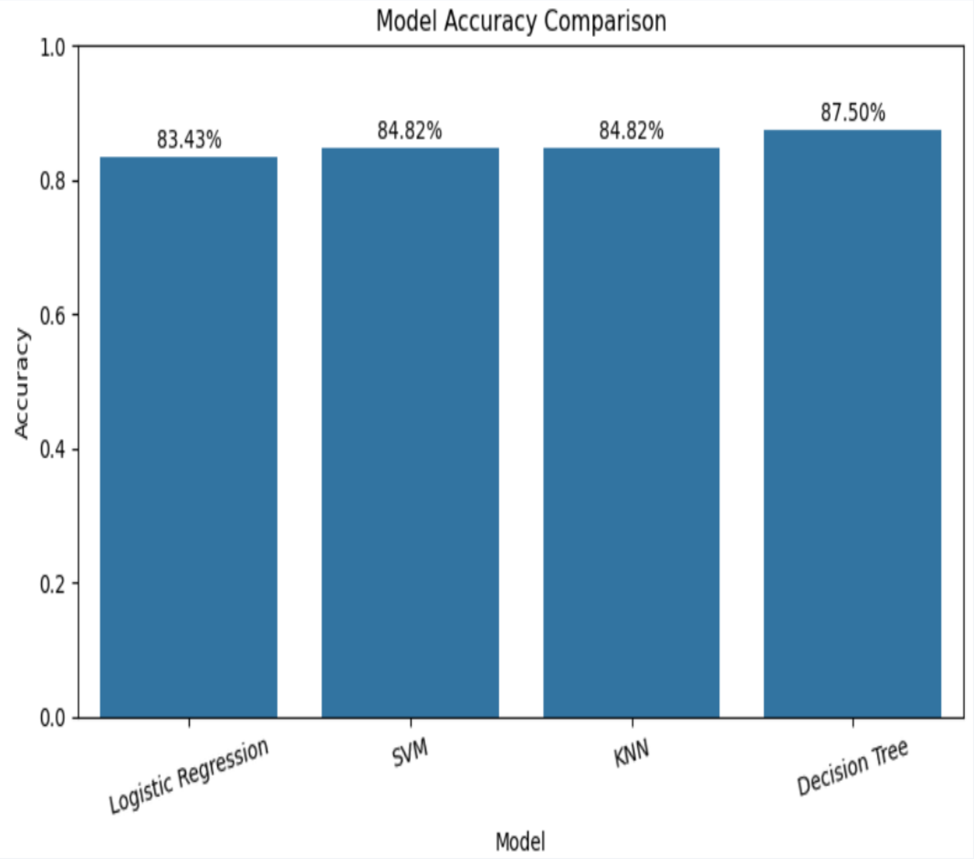
- Space X started their launch trials with rather small payload mass, less than 2,000 kg. They experienced more failure than success.
- As they increased payload mass and launched more, success rate has got higher.

Section 5

Predictive Analysis (Classification)

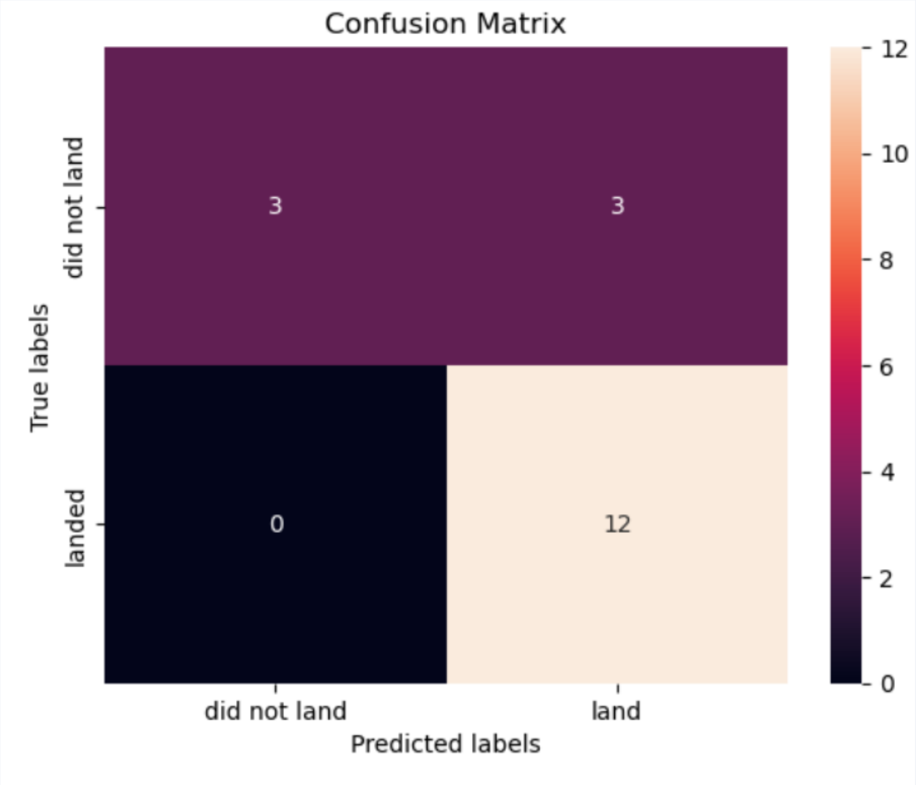
Classification Accuracy

- The predictive classification analysis applied;
  - Logistic Regression
  - Support Vector Machine
  - K-nearest neighbors
  - Decision Tree
- Each model used the grid search technique for the best parameter setting. The accuracy score was calculated with the best parameter.
- The highest accuracy was found with Decision Tree.



Confusion Matrix

- The confusion matrix of the decision tree model is shown on the right.
- Type I error rate is 50%. There is a room to improve the model to make it better.
- Type II error rate is 0%.

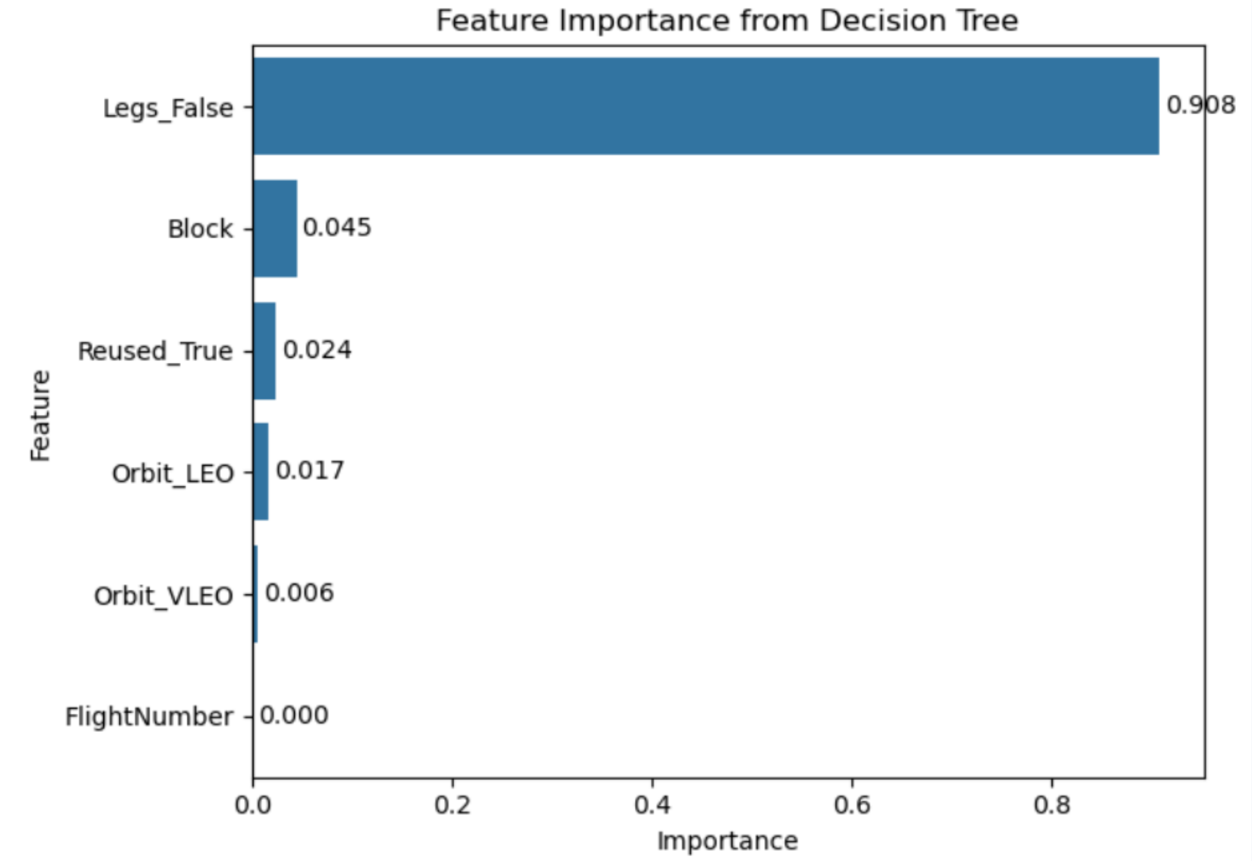


Conclusions

- Flight Number is an important factor to predict the launch success/failure.
  - First 20 launches experienced many failure.
  - The success rate got better after those 20 launches.
- Launch site has an impact on the success rate.
  - The first 30 launches were mostly done at CCAFS SLC-40. So the site experienced many failures.
  - Conversely, the other sites have experienced more successful launches.
- LEO/ISS/PO/GTO (~400km) were the most orbit targets.
  - They are for satellite, StarLink, and ISS.
- Launch Year is also important for successful launch rate.
- Machine Learning algorithm successfully built a prediction mode at 87% accuracy.

Appendix

- After instructed procedure of Machine Learning prediction, the feature importance from decision tree was calculated.
- The feature importance figured out the most important factor for successful launch. Launch without legs (Legs_False) is the most important (0.908).
- Other factors including Block, Orbit, or Payload Mass have almost no impact on success/failure of launch.



Thank you!