IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

HuyCQ
2022/01/30

# Outline

- Project Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

## Project Summary

- This project will go through from data collection to data prediction
- This project aim to generalize and summarize all knowleadge of what and how a Data Scienctist applied both math, programming and data-driven decision making to solve real-world prblem
- The problem we will cover today is SpaceX Falcon 9 first stage Landing Prediction

# Introduction

- This project is a final project for IBM Data Science
- The problem is about SpaceX, we will use its past landing data to predict if the future launch's landing successfull rate.
- SpaceX, or Space Exploration Technologies Corp., is an American aerospace manufacturer space transportation services, and communications corporation headquartered in Hawthorne, California. SpaceX was founded in 2002 by Elon Musk with the goal of reducing space transportation costs to enable the colonization of Mars. SpaceX manufactures the Falcon 9 and Falcon Heavy launch vehicles, several rocket engines, Cargo Dragon, crew spacecraft, and Starlink communications satellites.

Section 1

# Methodology

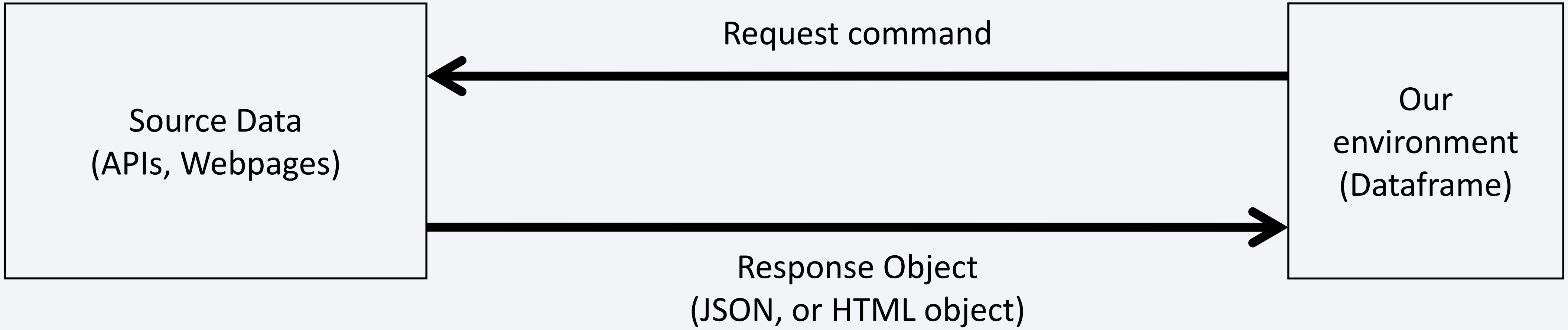
# Methodology

## Executive Summary

- Data collection methodology:
  - The data is collected from 2 sources: APIs and Web-scraiping
- Perform data wrangling
  - Raw data is preprocessed with Null replacement and Standarization
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - This project will evaluate many models base on their performance on GridSeasrch, then the best model is chosen and test with unknown data.

# Data Collection

- The raw data in this project will be collected in two methods: API and Web-scraping



Source Data
(APIs, Webpages)

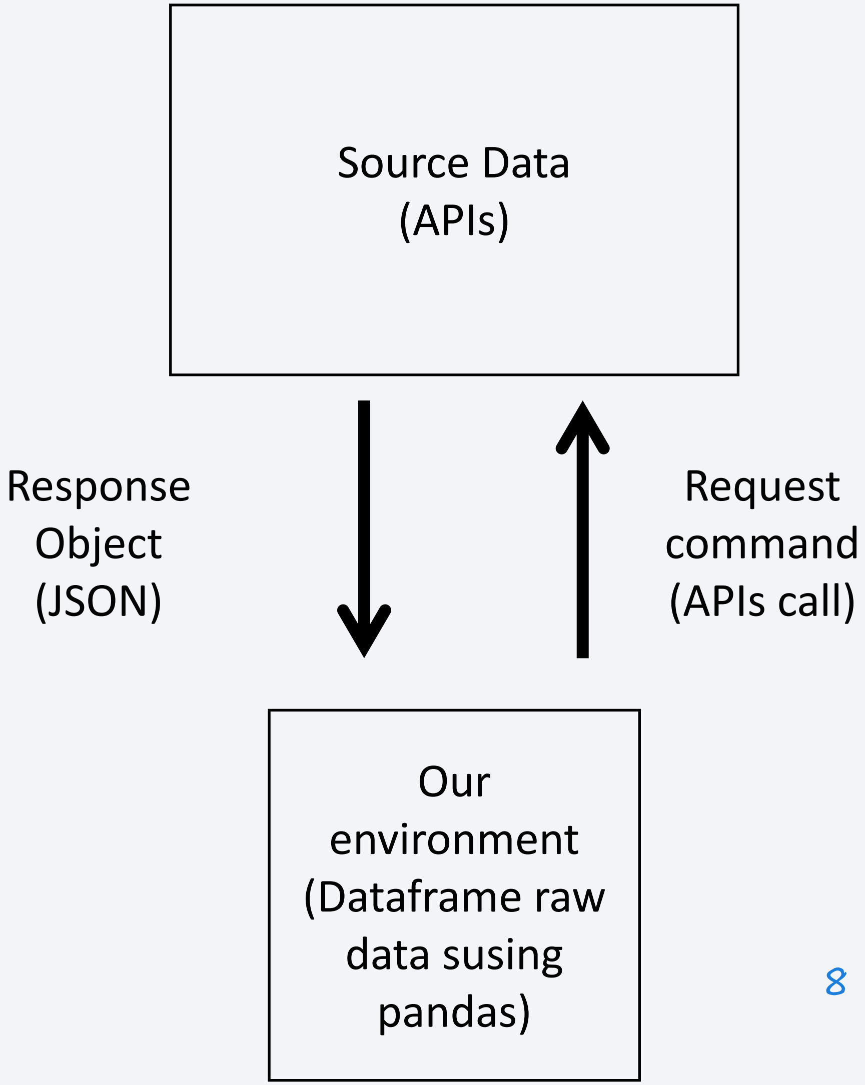
Request command

Response Object
(JSON, or HTML object)

Our
environment
(Dataframe)

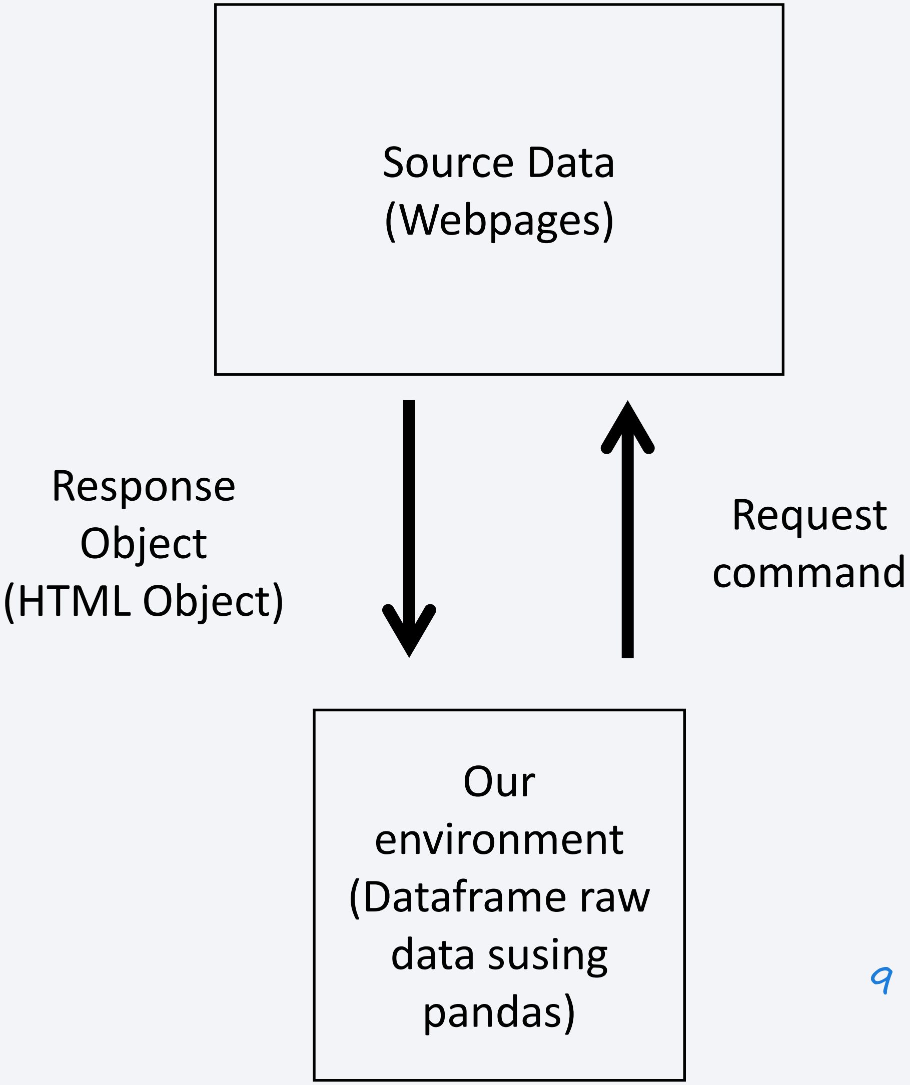
# Data Collection – SpaceX API

- APIs source url is here
- Response object type is JSON.
- Target raw data format is DataFrame (converted by pandas)
- For more information click here to see the notebook



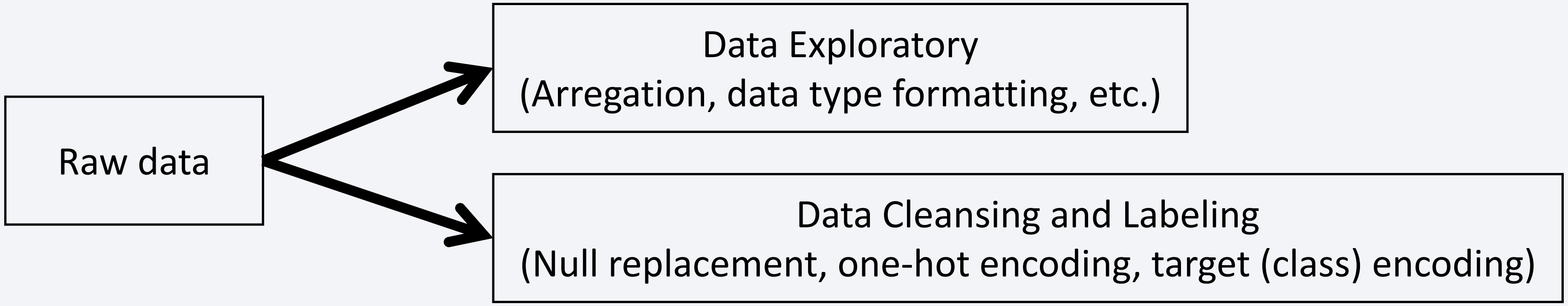
# Data Collection – Scraping

- Source url is here
- Response object type is HTML Object.
- Target raw data format is DataFrame (converted by pandas)
- For more information click here to see the notebook


Source Data (Webpages)
Response Object (HTML Object)
Request command
Our environment (Dataframe raw data susing pandas)

# Data Wrangling

- The data after collected will be used for EDA, or Exploratory Data Analysis, which in this project consist of these following step:


Raw data
Data Exploratory
(Arregation, data type formatting, etc.)
Data Cleansing and Labeling
(Null replacement, one-hot encoding, target (class) encoding)

- For more information click here to see the notebook

(Note: the one-hot encoding is from this notebook, in the Feature Engineering step)

# EDA with Data Visualization

- In this stage, we will continue the EDA process by using visualization to help us have more comprehensive understanding about the dataset.
- In this step, these following chart was used to graph the visualization of the dataset:
  - Scatter plot
  - Bar chart
  - Line chart
- For more information click here to see the notebook

# EDA with SQL

- We keep on EDA-ing but with different approach. In this step, we will use SQL command to query data from the datasource.
- SQL command covered in this step is:
  - Basic DQL commands such as SELECT, FROM, WHERE, GROUP BY,..
  - Some arregation fucntion such as SUM(), COUNT(), MAX(), MIN(),..
- For more information click here to see the notebook

# Build an Interactive Map with Folium

- This next step, we will also visualize our dataset, but in the different way: plot data on a MAP using package Folium.
- This step give us an insight about the distance about lauch field base, also organize the cluster the failed/success lauch coordinate around each lauch field base, make us understand more about if the successful rate depend on the launch field?
- For more information click here to see the notebook

# Build a Dashboard with Plotly Dash

- To maximize the experience for user, also in this project, we crate a easy-friendly web-based dashboard that users can interact with the dataset and plot some interesting visualization.
- In this webpage, two main visualization we want to emphasize is Pie chart to visualize the successfull rate of every launch sites, and a scatter plot chart to visualize the correlation between payload body mass with the successfull rate. Hence, we can see what feauter of the dataset is most effect to the target-successfull landing rate.
- For more information click here to see the server source code

(Note: you need a python IDE to run this source code)

# Predictive Analysis (Classification)

- Finnally, we conduct a study to find which following method:
  - Logistic Regression method
  - Support Vector Machine (SVM)
  - Decision Tree
  - K- nearest neigbors (kNN)

return the best performance on this dataset:

- After spliting the dataset into training and test set, we conduct a GridSearch with 10-fold to derive our best hyper-parameters for each model, then these best of each model will be finally test with the real unknown test data to find the best performance one.
- For more information click here to see the notebook

# Results

- Exploratory data analysis results
  - After EDA, we can conclude that our dataset is cleaned, with null preplacement, standarization and one-hot encoding for string data.
  - We have labeled the dataset with the new column ‘class’, which we will use as the target for prediction.
- Interactive analytics demo in screenshots

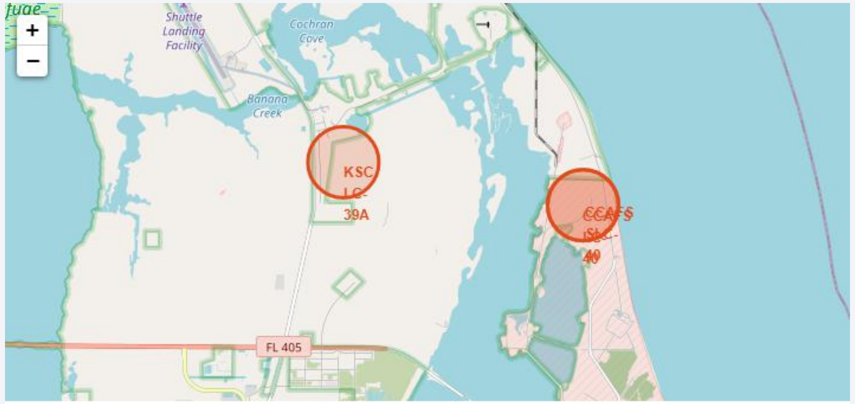


Fig1. Mark the launch site on Map

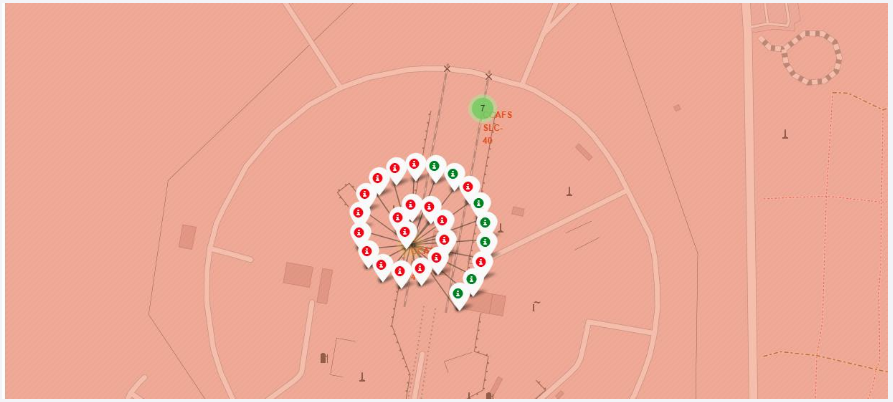
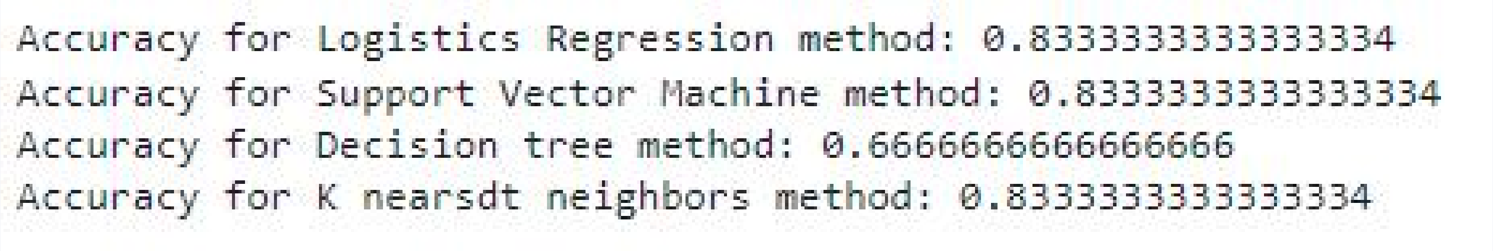


Fig2. Clustering launch outcome on each launch site

# Results

## Predictive analysis results

- We found that the kNN, SVM, and Logistic Regression method will give performance on unknown test data with accuracy of 83.33%.
- We also noticed that although the Decision Tree give the good accuracy on the train set (87.5%), it gave the poor performance on the test set with only 66.67%



```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.6666666666666666
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

Section 2

# Insights drawn from EDA

# Flight Number vs. Launch Site

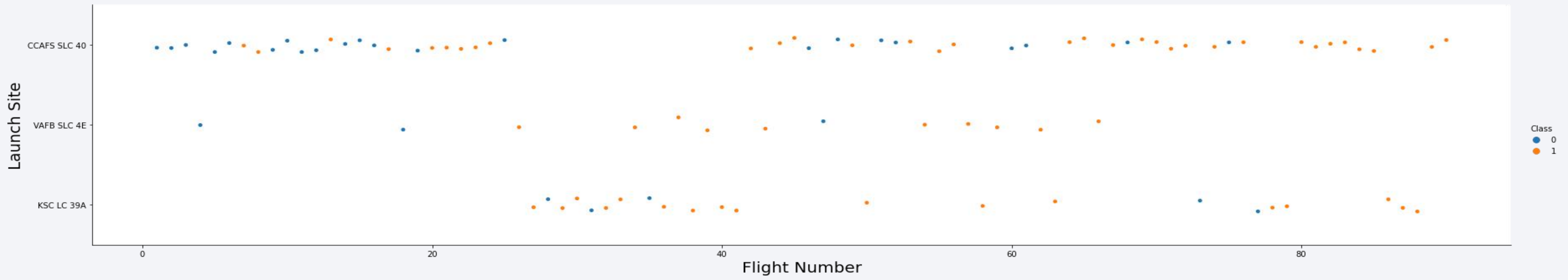


Fig3. Scatter Plot between Flight Number with Launch Site

- According to Fig3., we can concolude that KSC LC-39A have most flight number between 25 to 40 flight numbers, also the chance that the landing is success when launced at this site is considered to be high.

- With launch site VAFB SLC 4E, the flight number is varies from 5 to around 65 flight numbers, with more flight number, the chance of success landing also increase.

- the CCAFS SLC 40 landing site are the site that is varies on most flight number, but got a break at 25 to 40 flight numbers. we saw that the chance of success landing for this site is not high when compare to other two sites.

# Payload vs. Launch Site

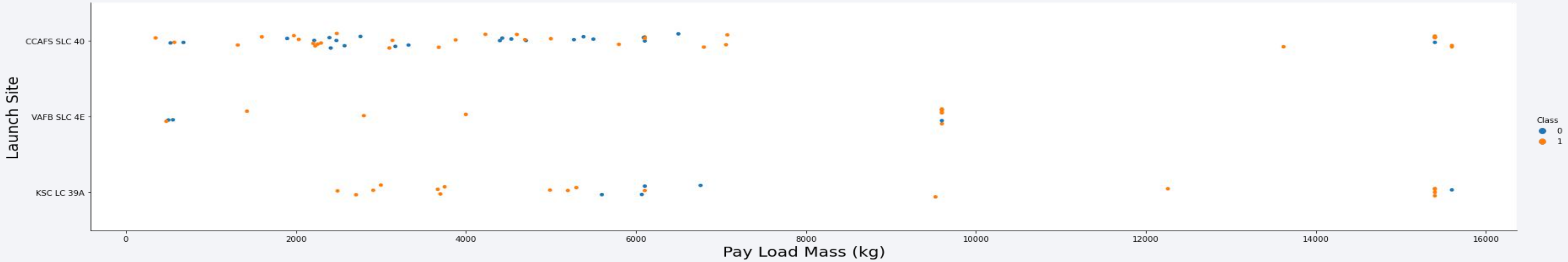


Fig4. Scatter Plot between Payload with Launch Site

- With this figure, we can derived that for all three site, the payload is mostly gathered below 7000 (kg). We also noticed that for all payloads that are over 7000 (kg) lauched by three site, the chance it had a successful landing rate is considerable high (only 3 cases failed in total 14 cases)
- With KSC LC-29A site, the more light the payload is, the higher rate of successful landing we got. It is opposite to VAFB SLC 4E, which had a higher chance of successful landing rate when the mass of payload is increased.
- The  CCAFS SLC 40 site give us an ambiguos insight since the chance of fail and succedd to land for every payload mass is varied randomly

# Success Rate vs. Orbit Type

- From this chart, we saw that with these orbit, the success rate is 100%:
    - ES-L1
    - GEO
    - HEO
    - SSO
- The GTO orbit returned the lowest success rate with the rate approximate 50%
- The VLEO orbit can be considered returned high rate of success since its result is over 80%

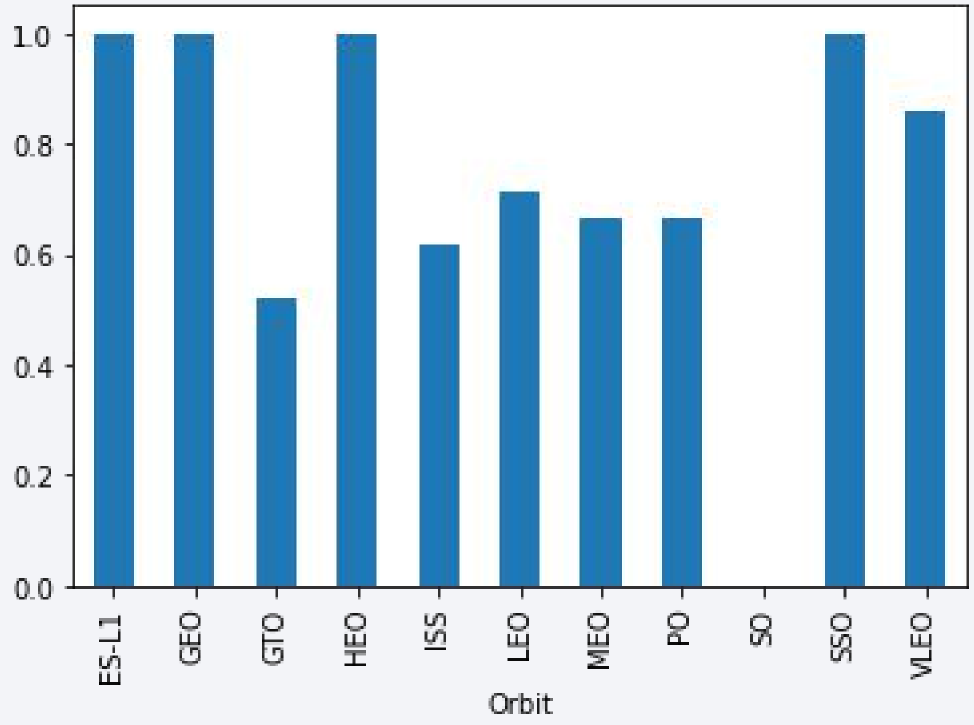


Fig5. Bar chart for Success rate with Orbit

# Flight Number vs. Orbit Type

- With this chart, we concluded that the group of LEO, ISS, PO, GTO orbits is where the flight number vary widely from 1 to over 80 flight numbers.
- Meanwhile, in the ES-L1, HEO and GEO orbit, the filght number only gather in specifical values such as around 17 (for ES-L1), around 50 (for HEO) and around 83 (for GEO)
- We also noticed that the success rate to land to ES-L1, SSO, HEO, GEO oribits is 100%. This rate is also hight for SO, ISS orbits.

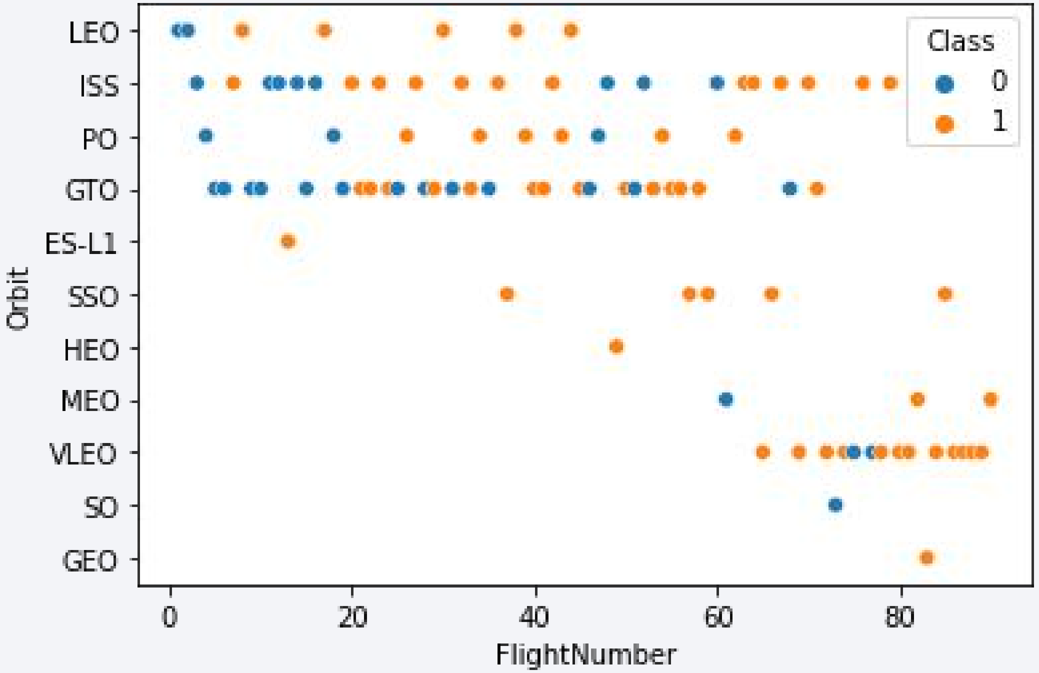


Fig6. Scatter plot for Flight number with Orbit

# Payload vs. Orbit Type

- According to this Fig7., we concluded that for a specific type of orbit, the payload mass is limited, for example, the payload mass for GEO and SO orbit is just 6000 (kg).
- We also see that the dispersal of data points for GTO is the lowest since it gathers only from 3000 to 8000 (kg), compare to others
- The group of ES-L1, SSO, HEO orbits gave us 100% chance to land successfully, this rate drop slightly for ISS.

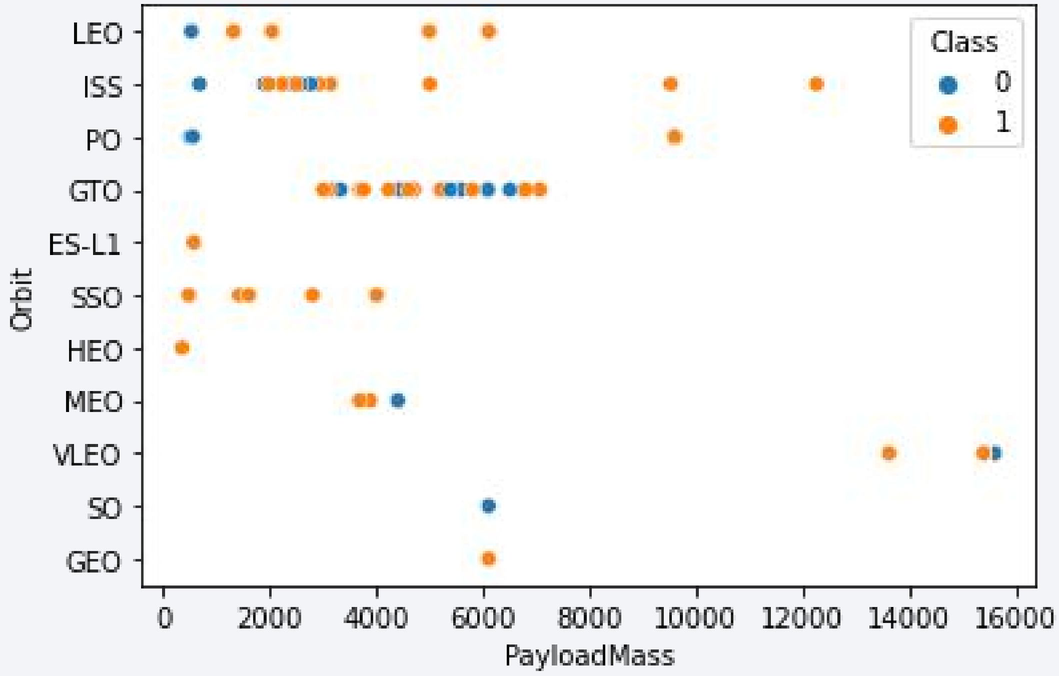


Fig7. Scatter plot for Payload with Orbit

# Successful Land Yearly Trend

- As this figure shown, as the year went by, the success rate reach higher.
- This rate rise significantly from 2015 to 2017 but drop dramatically (~20%) in the next year. But gain back its performance at the rate over 80% in 2019
- In the 2020, the succesful rate is slightly decrese but still at the rate of 80%

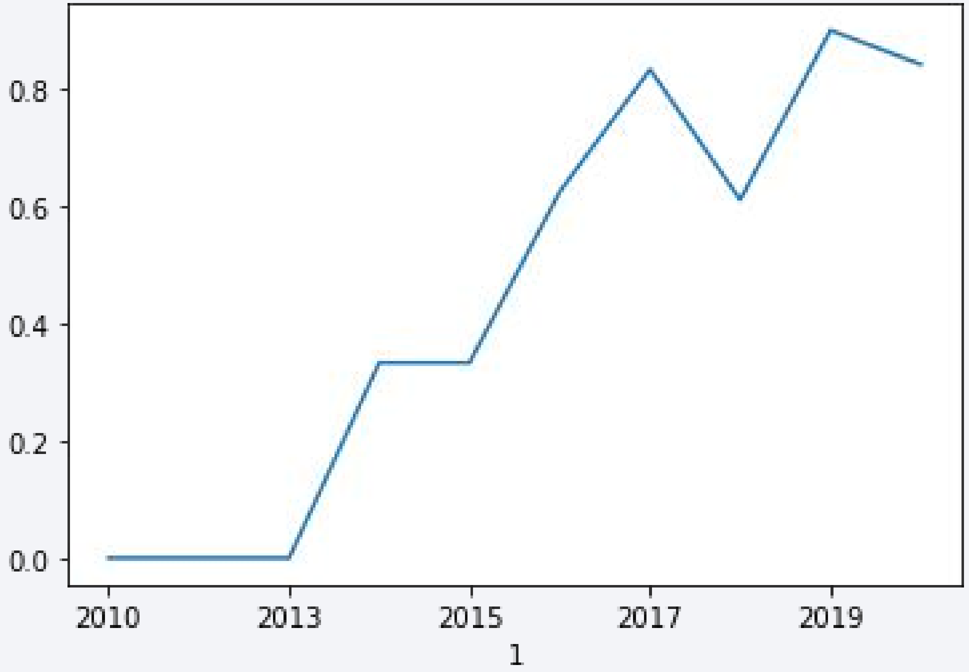


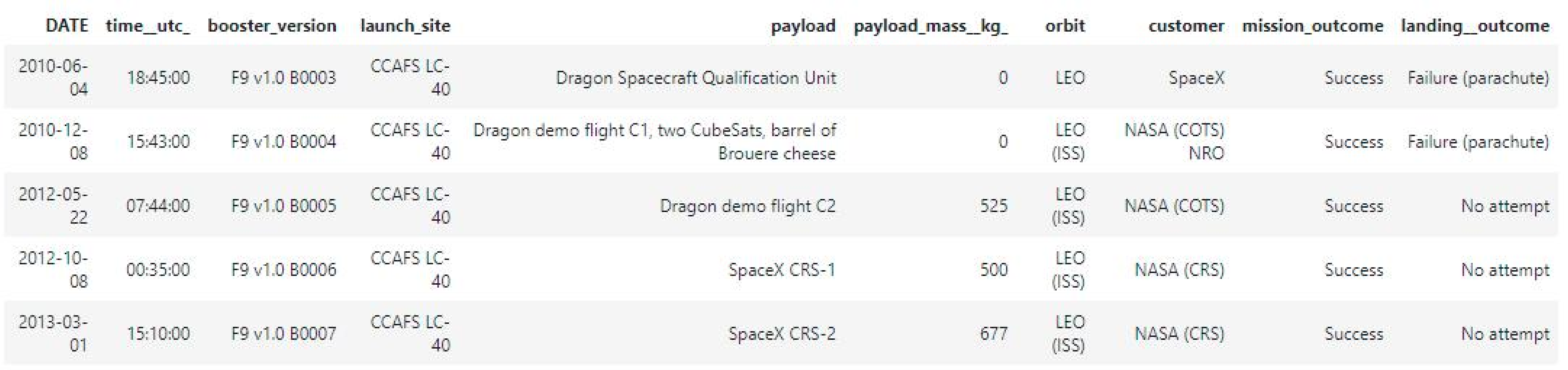
Fig8. Line graph show the trend of success rate over years

# All Launch Site Names

- In this dataset, we will study the data of this following launch site:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- To get the above data, we just simply put this code to query the data from the dataset:
  - %sql select distinct launch_site from SPACEXTBL

# Launch Site Names Begin with 'CCA'

- In order to get the data with the lauch site name begin with ‘CCA’, we use the following query:
    - %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
- The result is shown here:



| 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 |

# Total Payload Mass

- We apply this query to get the total payload mass:
    - %sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)'
- The query return the following result: 45596 (kg) is our total payload mass

# Average Payload Mass by F9 v1.1

- To get this, we use this query:
  - %sql select avg(payload_mass__kg_) from SPACEXTBL where booster_version = 'F9 v1.1'
- The above returned 2928(kg), which is the average payload mass by booster version of F9 v1.1

# First Successful Ground Landing Date

- We conclude that the first successful ground landing date was at 2015-12-22, to know this, we aplly this query:
  - %sql select min(date) from SPACEXTBL where landing__outcome = 'Success (ground pad)'

# Successful Drone Ship Landing with Payload between 4000 and 6000

- We can list the name of booster that satisfy the condition above is:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2
- The query command to get this result is:

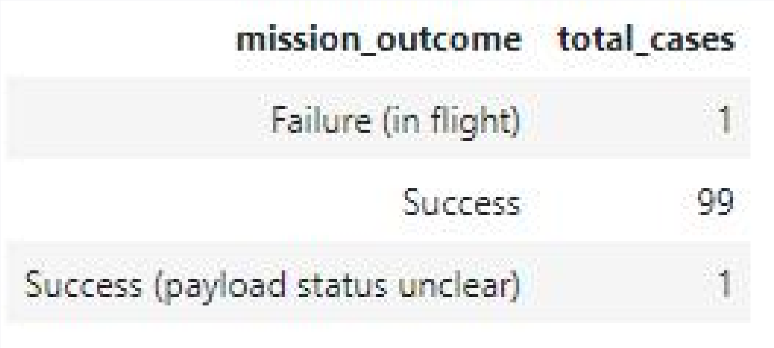
```
%sql select booster_version from SPACEXTBL where landing__outcome = 'Success (drone ship)' and payload_mass__kg_
between 4000 and 6000;
```

# Total Number of Successful and Failure Mission Outcomes

- To get the total number of successful and failure mission outcome, we use the following query:

```
%sql select mission_outcome, count(mission_outcome) as total_cases from SPACEXTBL group by mission_outcome
```

- The result is shown here:



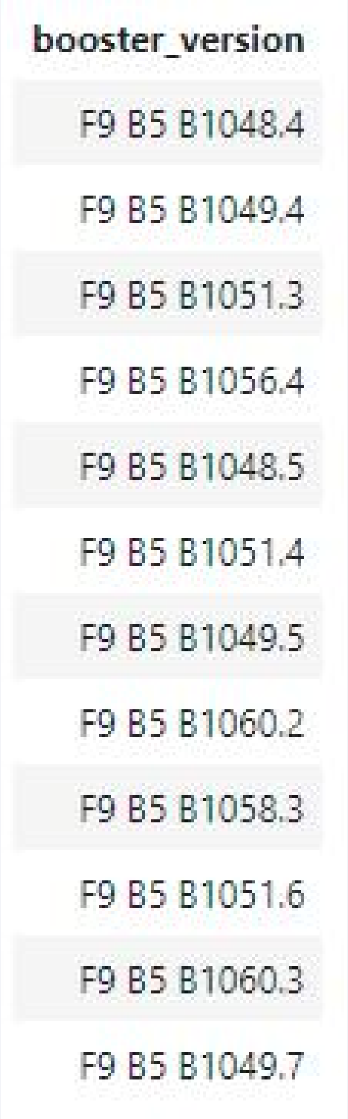
| mission_outcome | total_cases |
|---:|---:|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

# Boosters Carried Maximum Payload

- Using

```
%sql select booster_version from SPACEXTBL where
payload_mass__kg_ = (select max(payload_mass__kg_) from
SPACEXTBL)
```

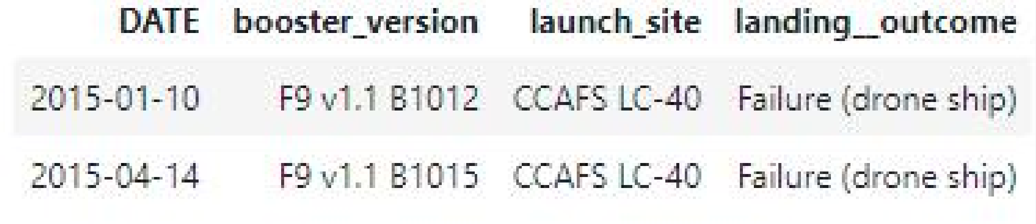
we can derive the right result, which satisfied the condition of get boosters carried the maximum payload.



| 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 |

# 2015 Launch Records

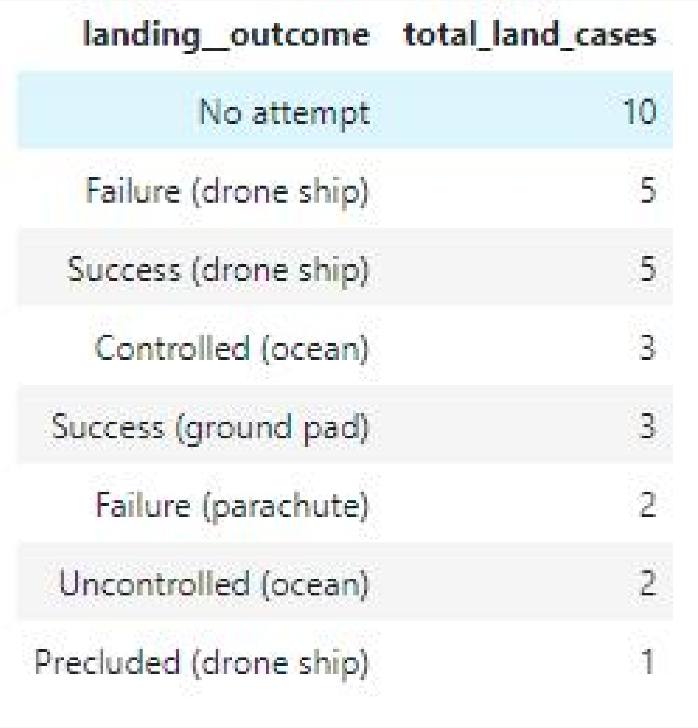
- To know the year 2015 Launch records, we simply apply this query:
    - %sql select date, booster_version, launch_site, landing__outcome from SPACEXTBL where landing__outcome = 'Failure (drone ship)' and year(date) = 2015
- And the result we get is the 2015 launch recorded derived from our dataset:



| DATE | booster_version | launch_site | landing__outcome |
|---|---|---|---|
| 2015-01-10 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 2015-04-14 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

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

- 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, we used the following query:
    - %sql select landing_outcome, count(landing_outcome) as total_land_cases from SPACEXTBL where date between '2010-06-04' and '2017-03-20' group by landing_outcome order by total_land_cases desc
- The result is shown in the right table



| landing_outcome | total_land_cases |
|---:|---:|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

Section 3

# Launch Sites Proximities Analysis

# Mark all launch sites on a map

- Arcording to the map, we saw that 3 of 4 launch sites lcoated in Florida, remain launch site, VBA SLC-4E locate in California.
- Also, we acknowledged that two site CCAFS SLC-40 and CCAFS LC-40 located very near together.

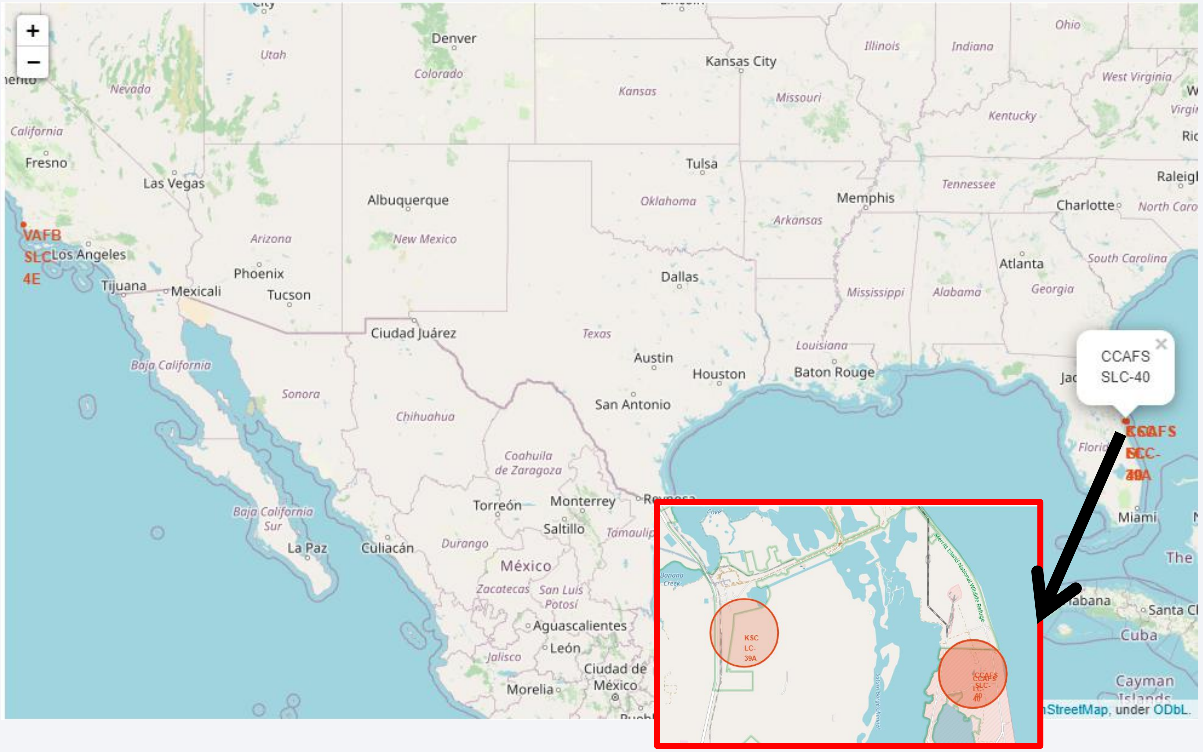


Fig9. All SpaceX launch site on the US map

# Mark the success/failed launches for each site on the mapp

- With this map, we can calculate the number of success/failure rate of launching booster at each site. Beside, the color at each site indicate that the success rate, the more red the color, the lower success rate get.

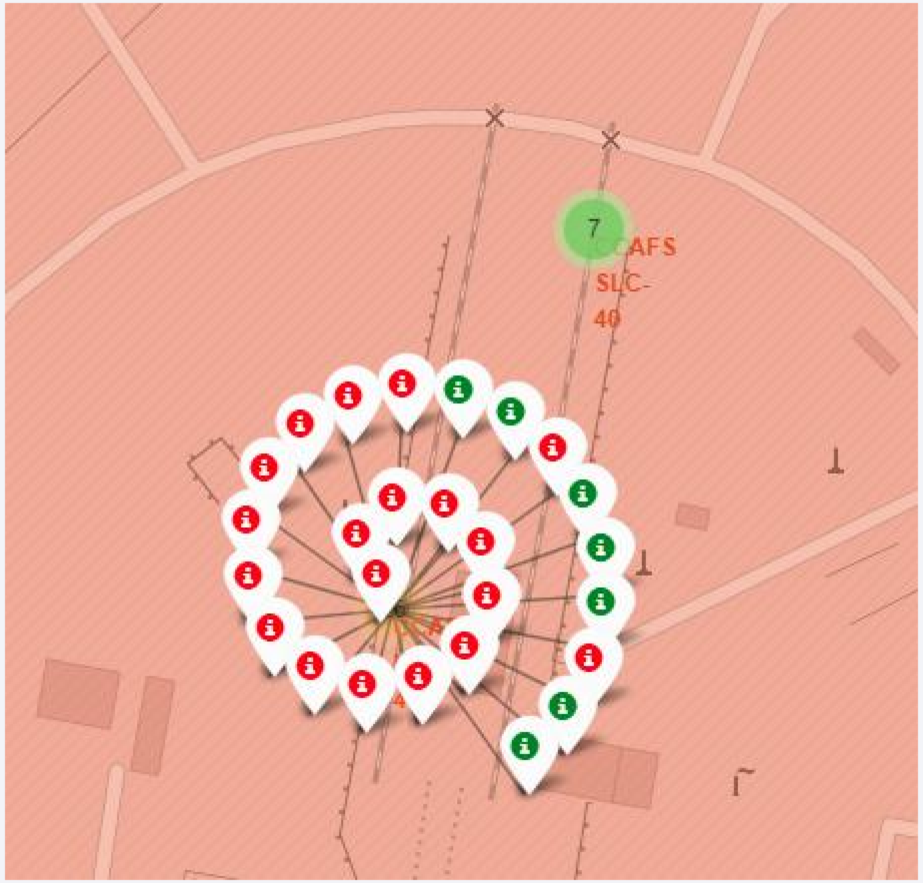


Fig10. CCAFS LC-40 launch site with labeled outcome

# Calculate the distances between a launch site to its proximities

- We can also calculate at some specific location, how far does it take to reach our lauch site.
- For example, the distance from the CCAFS LC-40 launch site to Melbource Beach is approsimately 54.78 (km)

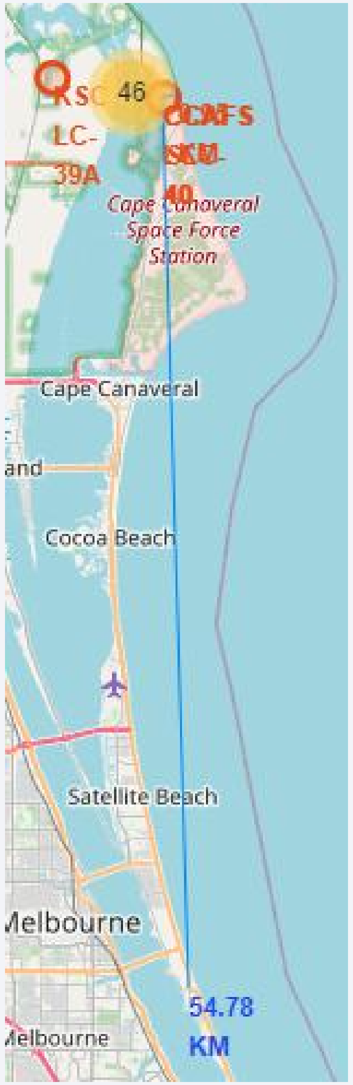
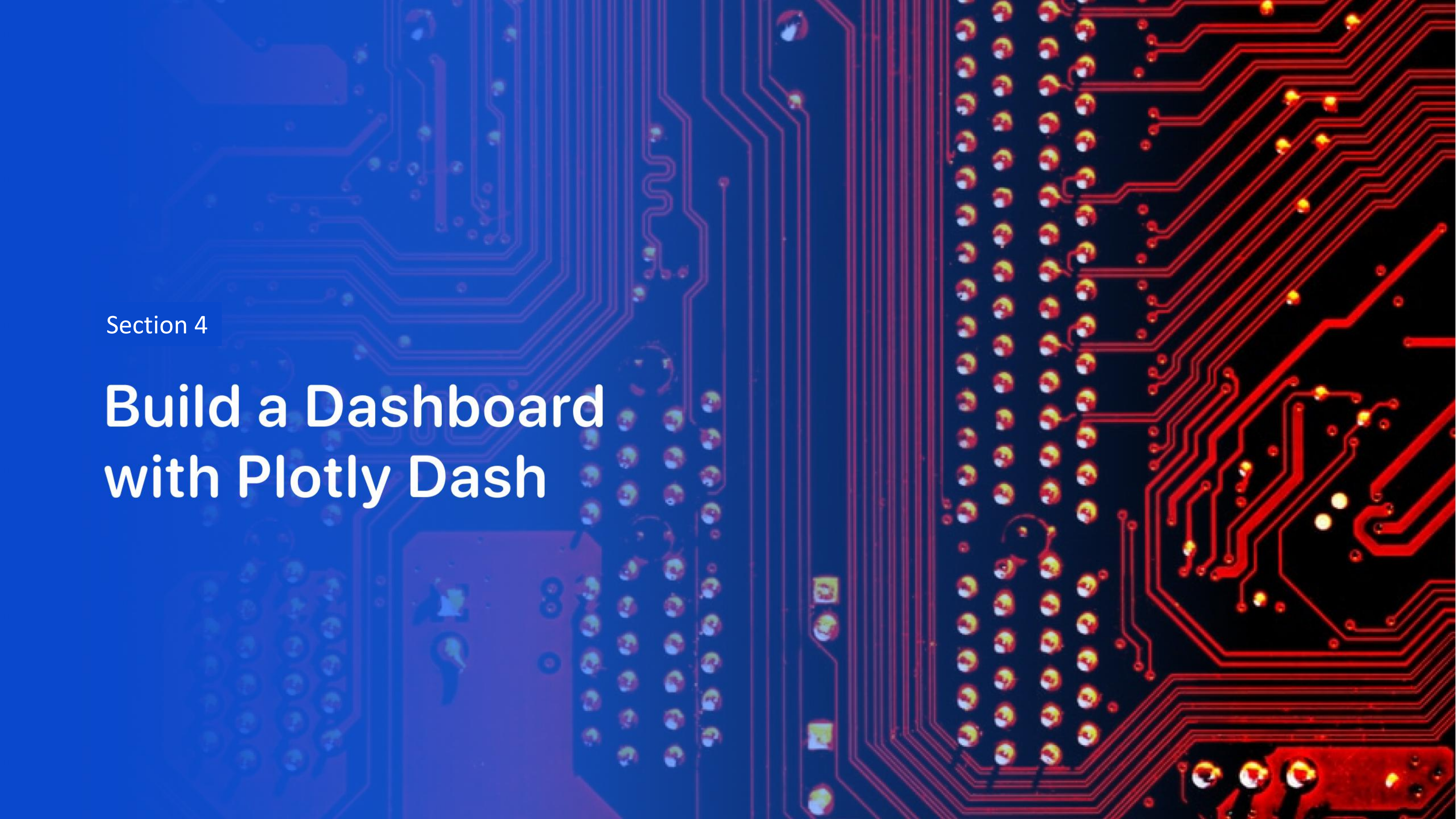


Fig11. Distance from CCAFS LC-40 launch site to  Melbource Beach

Section 4

# Build a Dashboard with Plotly Dash



# Success rate based for all site

- As shown in the Pie chart, the KSC LC-39A site had the most success rate ratio compare to other three site, the success rate ratio is as follow:
  - KSC LC-39A : 41.7%
  - CCAFS LC-40: 29.2%
  - VAFB SLC-4E: 16.7%
  - CCAFS SLC-40: 12.5%

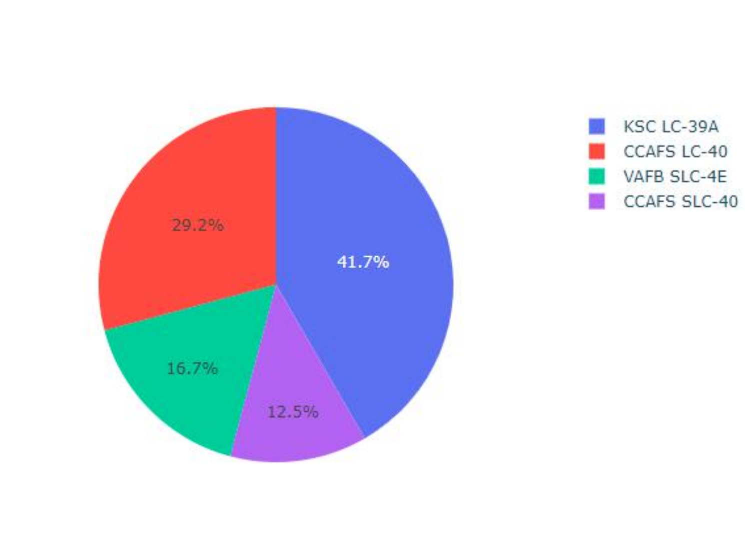


Fig12. Success rate by launch site

# Success launch rate at KSC LC-39A site

- For this graph, the success rate of the KSC LC-39A launch site is further detaily analyze. We can see that in this site 76.9% launched booster landed successfully, otherwise 23.1% failed to land.

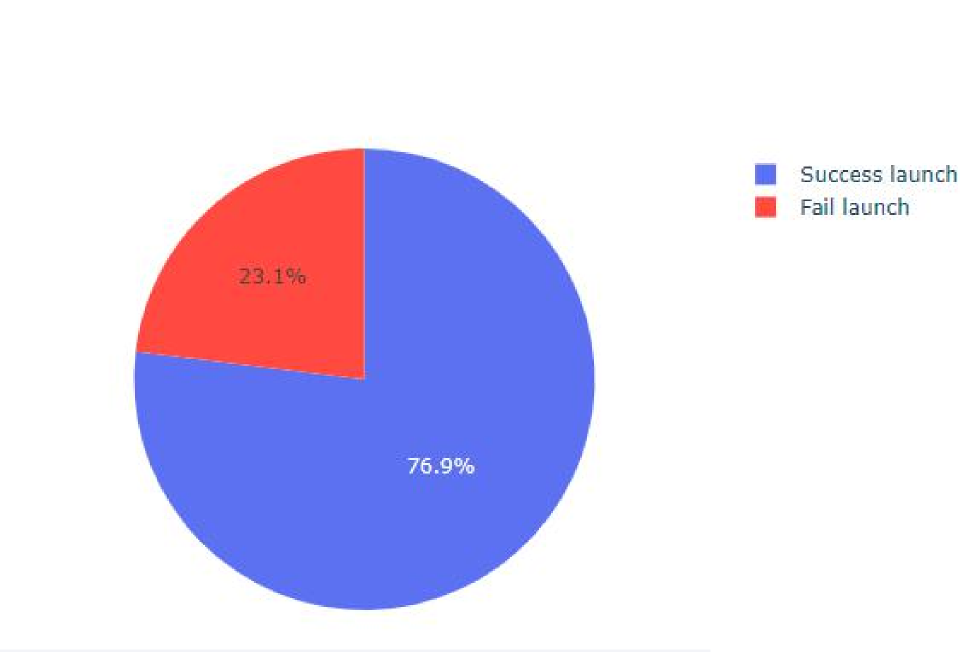


Fig13. Success rate at KSC LC-39A site

# Payload Mass and Success rate correlation

- Fig13. shows the correllation between payload mass and success rate. We can see that the payload mass point is gather mostly from 2000 (kg) to 6000 (kg), thus we used slider to zoom out the payload mass range. We derived Fig.14
- Fig14. shows that the version FT gave more successful landing others version.

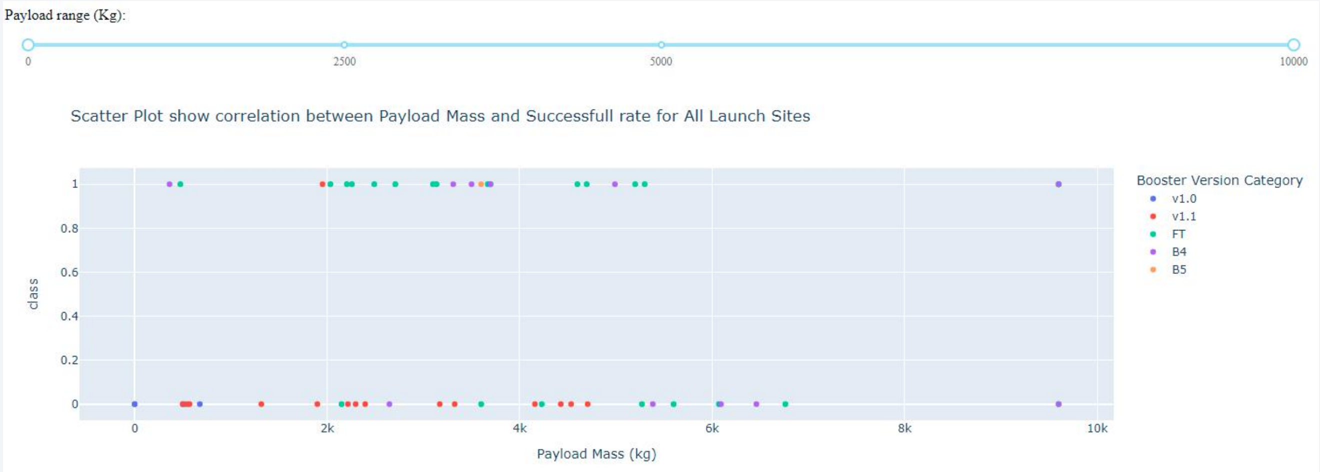


Fig13. Correlation between Payload mass and success rate (all mass)

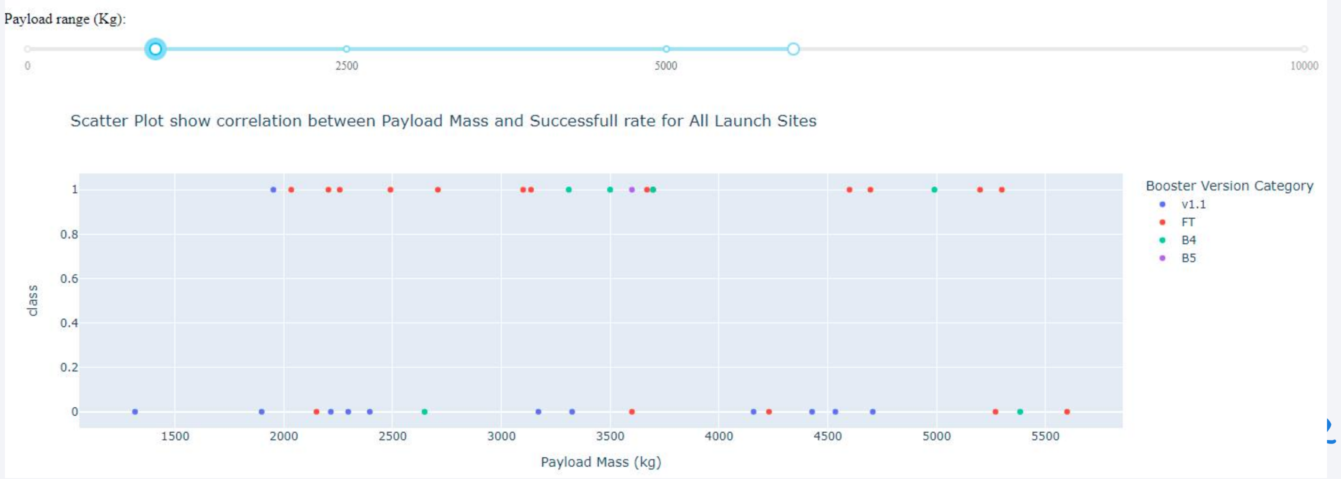


Fig14. Correlation between Payload mass and success rate (2000kg to 6000kg)

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

- According to this result graph, we conclude that Logistic Refression, Support Vector Machine (SVM), and k-Nearest Neighbor (kNN) models all had 83.33% accuracy on the test set with an average 84.5% on training accuracy.
- Decision Tree model, in the other hand, gave 88.93% training accuracy, but recieved poor performance on the test set (66.67%)

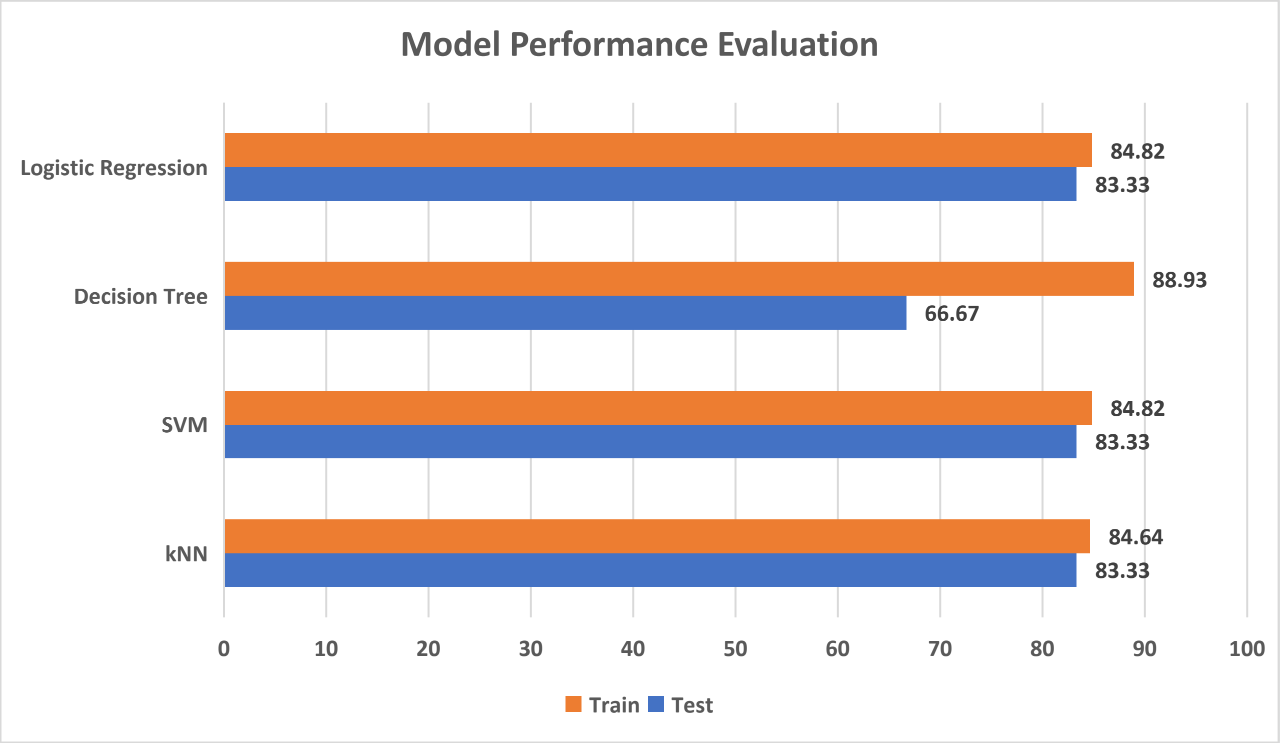
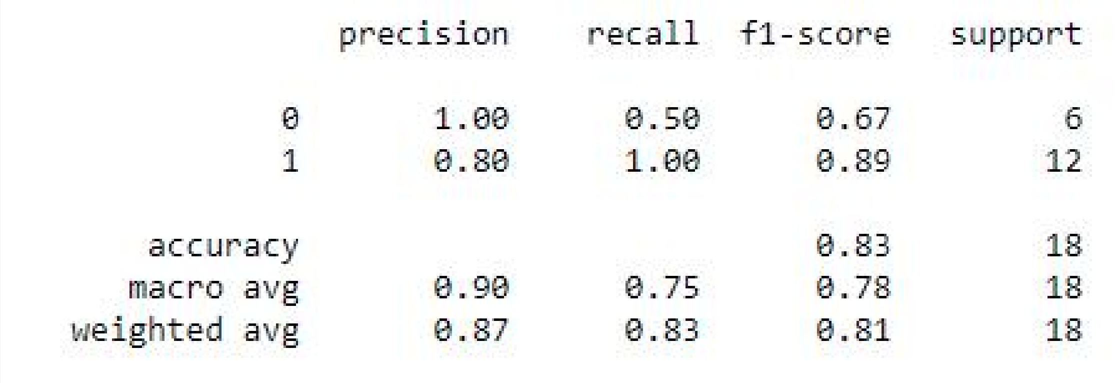


Fig15. Classification accuracy

# Confusion Matrix

- The confusion matrix of all three best model (logistic regression, SVM, kNN) are the same and shown as the right figure.
- From this confusion matrix, we make a classification report to see how our best model performed on each classes.



```
              precision    recall  f1-score   support

           0       1.00      0.50      0.67         6
           1       0.80      1.00      0.89        12

    accuracy                           0.83        18
   macro avg       0.90      0.75      0.78        18
weighted avg       0.87      0.83      0.81        18
```

Fig17. Classification Report

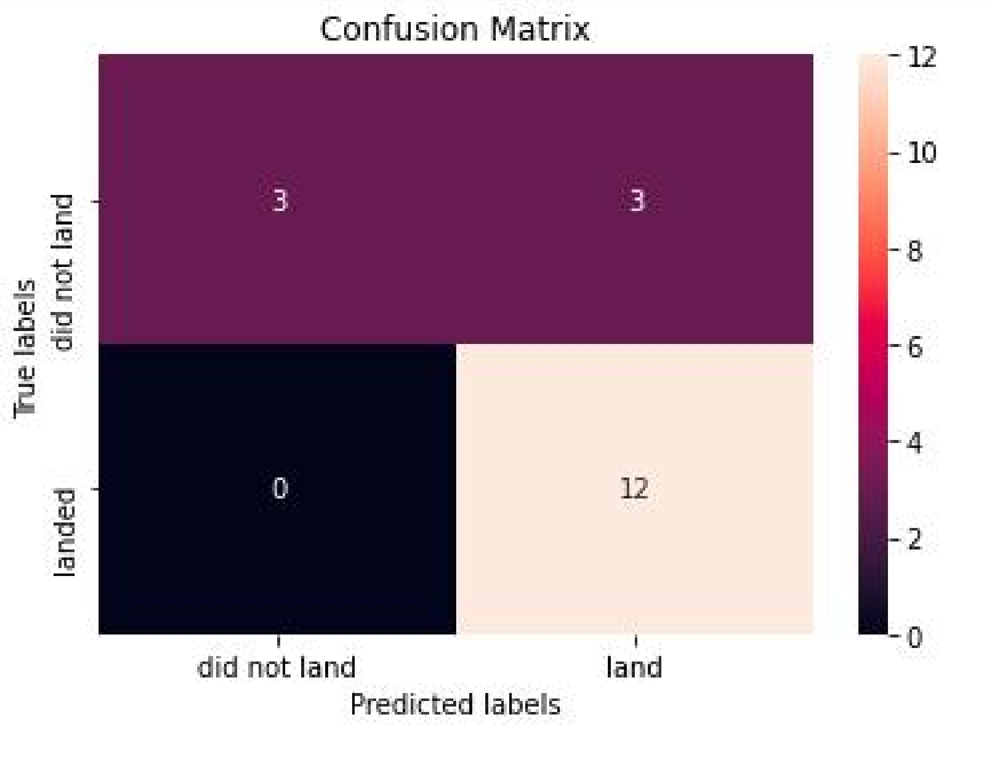


Fig16. Confusion Matrix

# Conclusions

- In this project, we had:
  - Retriving data from API request, or Web-scraping
  - Perform Preprocessing data and EDA through many methods (visulization, SQL). From that, we derieved insight from each visualization from scatter plot, bar graph and line graph. Proved that the success rate increase dramatically through years.
  - Built a web-based interactive dashboard, enhanced user experience. From this step, we conclude that KSC LC-39A had best success rate ratio with 76.9% success rate.
  - Finally, we had built a machine model that can predict wheather in the future launch, does the booster has how many probability to land successfully, this project provide that we can use Logistic Regression, Support Vector Machine, and k Nearest Neighbors with appropriate hyper-paramters can 83.33% predict correctly the chance of successful land.

# Appendix

- For all source code and notebook, refer my personal Github respository https://github.com/CQHofsns/IBM-FinalCapStone

Thank you!