Winning Space Race with Data Science

<Sarut Sunpawatr>
<20/07/2022>

Outline

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

Executive Summary

- Summary of methodologies
  - Collecting the Data
  - Data Wrangling
  - Exploratory Data Analysis
  - Visual Analytics
  - Classification Analysis
- Summary of all results
  - For 4 machine learning algorithms for predice the landing outcome the best percentage was only 83.3% . It showed that the data was not enough. Therefore, we need to find more attributes that effect on landing outcome to improve our accuracy

Introduction

- Project background and context
  - As the Space Y company, we would like to predict the outcome of a rocket landing. Which characteristic has the most effect on rocket landing? Consequently, we decide to get the information from the Wikipedia website that SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.

- Problems you want to find answers
  - Find correlation between each attribute
  - Perform the best model to predict landing outcome correctly

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Rest SpaceX API
  - Web scrapping
- Perform data wrangling
  - Analyzed attributed, Classified between successful landing and fail to land
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Select the model according to the problem
  - Find the best parameters for each model
  - Compare accuracy with each other

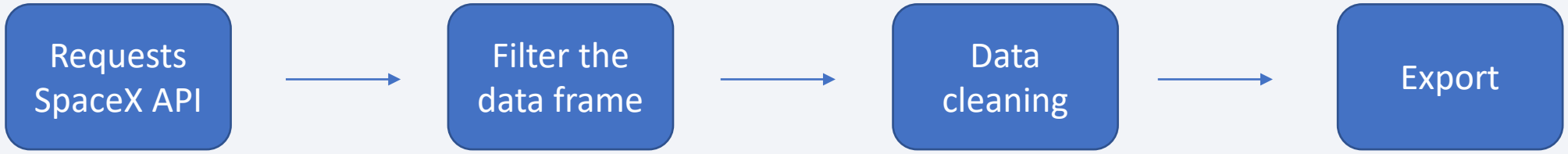
Data Collection

Data sets were collected in 2 different ways as follows:

1. Rest API
    - Collected from SpaceX API
        - https://api.spacexdata.com/v4/rockets/
        - https://api.spacexdata.com/v4/launchpads/
2. Web scrapping
    - Collected from Wikipedia
        - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

1. Collected the data from SpaceX API that shows on the previous page
2. Process the data following with flowchart showed below



Github : https://github.com/parklhbd45/Data-project/blob/7dbfd7352bf352d82881408bdd30d1a73bdc8318/jupyter-labs-spacex-data-collection-api.ipynb

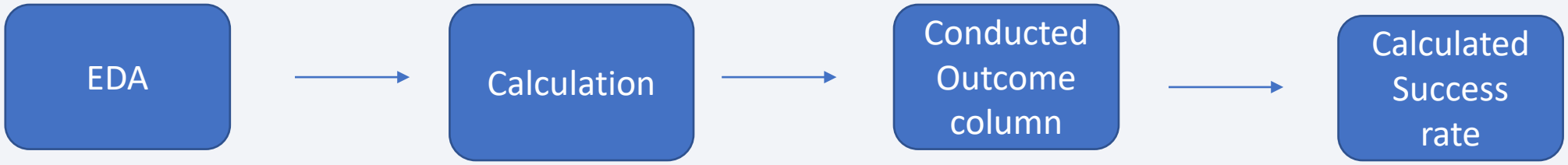
Data Collection - Scraping

1. Collected the data from Wikipedia that shows on the previous page
2. Process the data following with flowchart shown below

Github : https://github.com/parklhbd45/Data-project/blob/7dbfd7352bf352d82881408bdd30d1a73bdc8318/jupyter-labs-webscraping.ipynb

Data Wrangling

1. Perform Exploratory Data Analysis (EDA) to find suitable data for training supervised model
2. Calculate the number of launches on each site, the occurrence of each orbit, and mission outcome per orbit type
3. After that, Conducted the landing outcome label from Outcome column. Therefore, Calculated the success rate



Github : https://github.com/parklhbd45/Data-project/blob/7dbfd7352bf352d82881408bdd30d1a73bdc8318/jupyter-labs-spacex-data-collection-api.ipynb

EDA with Data Visualization

In this project, There are three types of chart

1. Scatterplot (Easy to inspect spreading of data)
   - Analyze Class depending on PayloadMass and FlightNumber
   - Analyze Class depending on LaunchSite and FlightNumber
   - Analyze Class depending on PayloadMass and LaunchSite
   - Analyze Class depending on Orbit and FlightNumber
   - Analyze Class depending on Orbit and PayloadMass
2. Barplot (Easy to understand and analysis)
   - Analyze between Class and Orbit
3. Lineplot (Easy to inspect the trend)
   - Analyze between Year and Success rate

Github : https://github.com/parklhbd45/Data-project/blob/a1adc634764ab56ea3b25fe0142d7c03c9330b5b/jupyter-labs-eda-dataviz%20(1).ipynb

EDA with SQL

executed SQL queries to get the following information

- Names of the unique launch sites in the space mission
- Top five launch sites whose names contain ‘CCA’
- Total payload mass carried by NASA boosters launched
- Average payload mass carried by F9 v1.1
- First date that achieved landing in a ground pad
- Name of the boosters which have success in drone ship and payload mass greater than 4000 but less than 6000
- Total number of each mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Display the month names, failure landing outcomes in drone ship, booster versions, and launch sites for the months in the year 2015.
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Github : https://github.com/parklhbd45/Data-project/blob/9523117eddd4129f2e9c2578e65e4920a567b8fc/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- There are three kinds of map objects:

    1. Markers : Showed the landing outcomes. For example, if the landing was successful the marker would be green, on the other hand, if the landing was a failure the marker would be red
    2. Circles : Showing the location of each launch site
    3. Lines: showed a distance Launch site to the nearest railway, highway, coastline, and city

Github : https://github.com/parklhbd45/Data-project/blob/ce57eb052358215de46ffd518000a87835724325/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- There are two types of graphs that help to identify the best location according to payload mass and booster version

    1. Pie chart :

        1.1 Compared the successful percentage of all sites

        1.2 Compared the successful and failure percentage of each site

    2. Scatter plot :

        2.1 Showing the correlation between outcome and payload mass of each booster version

Github : https://github.com/parklhbd45/Data-project/blob/ce57eb052358215de46ffd518000a87835724325/Plotly%20dash%20visualization.ipynb

Predictive Analysis (Classification)

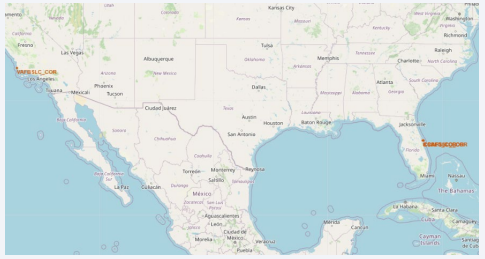
Procedure of Analysis

1. Analyse type of problem: In this section is Classification
2. Choose model : there are four model for classification analysis : K nearest neighbors, decision tree, support vector machine, logistic regression
3. Find the best parameter for each model : Using GridSearchCV
4. Compared accuracy

Github : https://github.com/parklhbd45/Data-project/blob/ab83d764f9d4765164b56f691071d5db74510f32/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

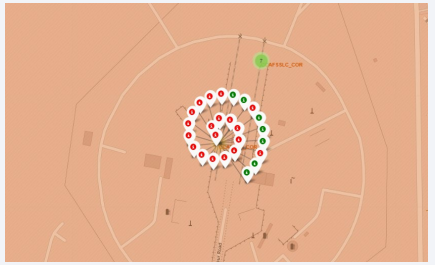
Results

- **Exploratory data analysis results**
  - There were four launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
  - the first successful landing outcome in the ground pad in the year 2015
  - Most F9 could carry high payload mass
  - Only 1 time of failure mission outcomes
  - By 2015, Two booster versions fail to land in drone ship
  - The number of success landings were better by time
- **Interactive analytics demo in screenshots**



Location of Launch sites

The Number of Launching



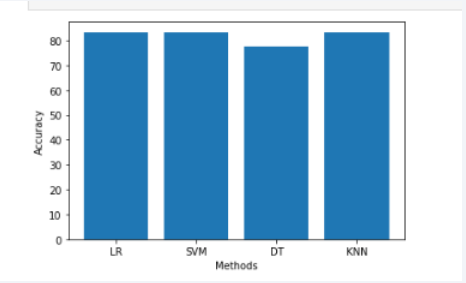
Successful and failure landing markers

Distance between two coordination

- **Predictive analysis results fail to land in drone ship**

  According to the graph, Logistis Regression, support vector machine, and decision tree have same accuracy over83%.

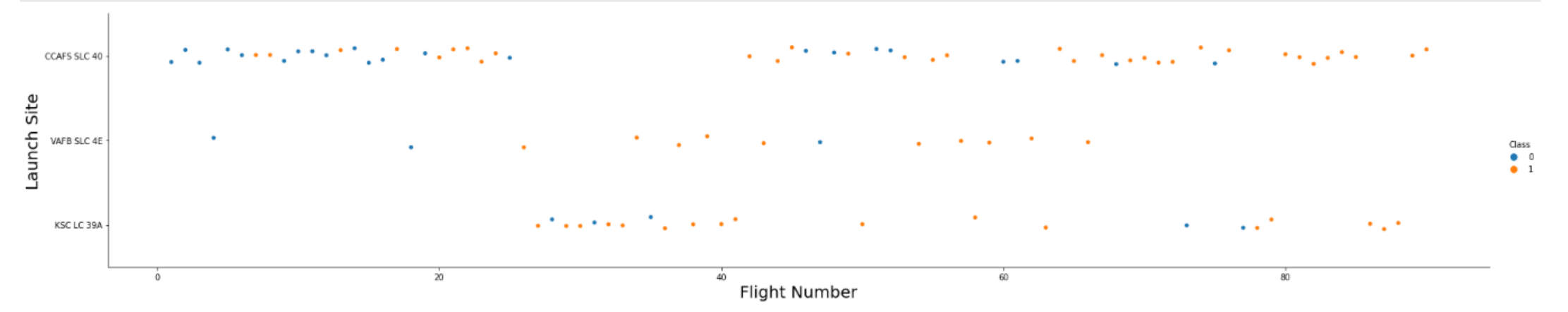
  However, K nearest neighbors is the worst model having accuracy only 78%



Section 2

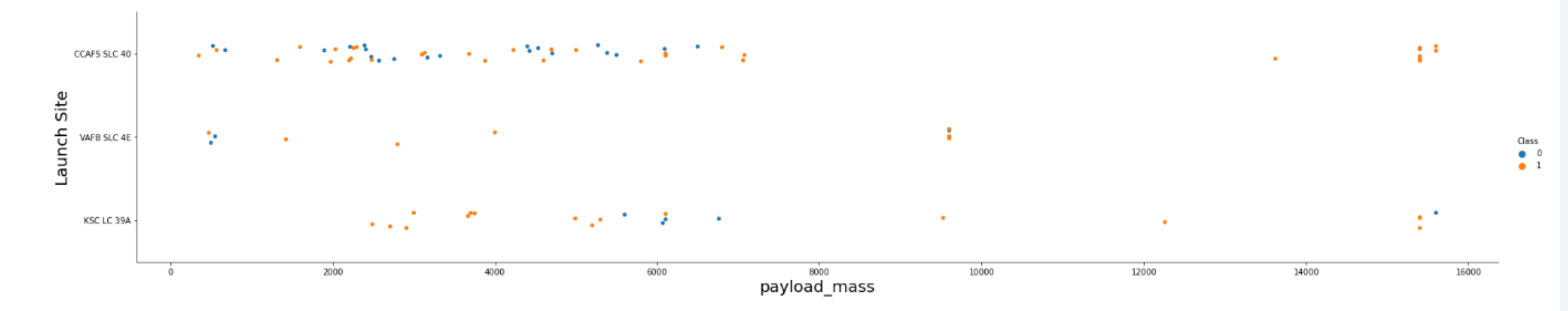
Insights drawn from EDA

Flight Number vs. Launch Site



- According to the graph, the Location with a high success percentage was VAFB SLC4E. Followed by KSC LC 39A and CCAF5 SLC40, respectively.
- The highest number of successful landings appears at CCAF5 SLC40
- The success rate is increased over the period.

Payload vs. Launch Site



- The graph showed that CCAFS SLC 40 site has a better successful percentage as payload mass
- Almost 100% success rate appears at the payloads over 8000 kg

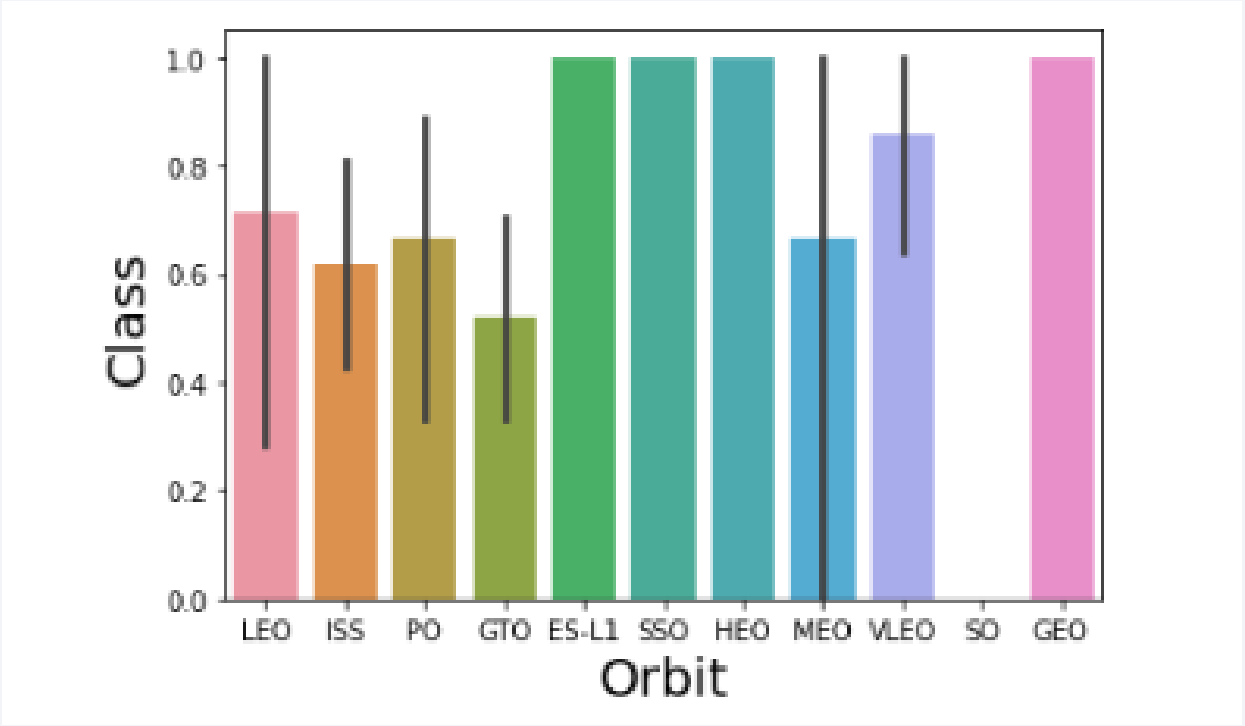
Success Rate vs. Orbit Type

- There are four highest success percentage

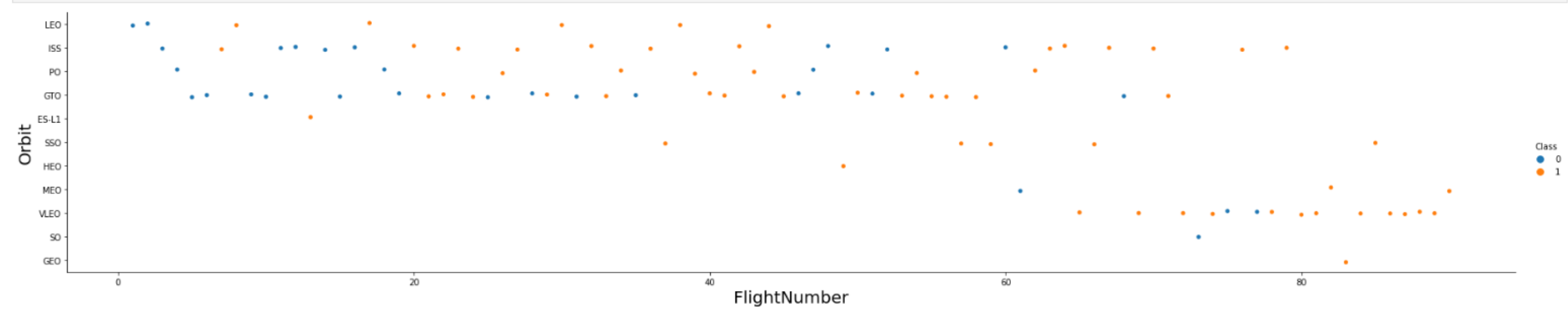
1. ES-L1
2. SSO
3. HEO
4. GEO

- Followed by

1. VLEO
2. LEO

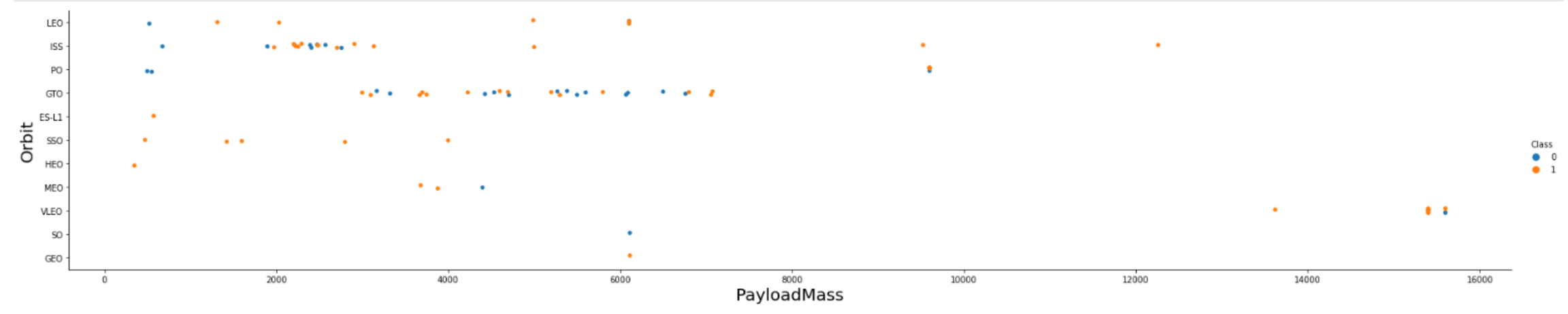


Flight Number vs. Orbit Type



- The success rate is increased over the period.
- VLEO orbit has an excellent success percentage in the short past years

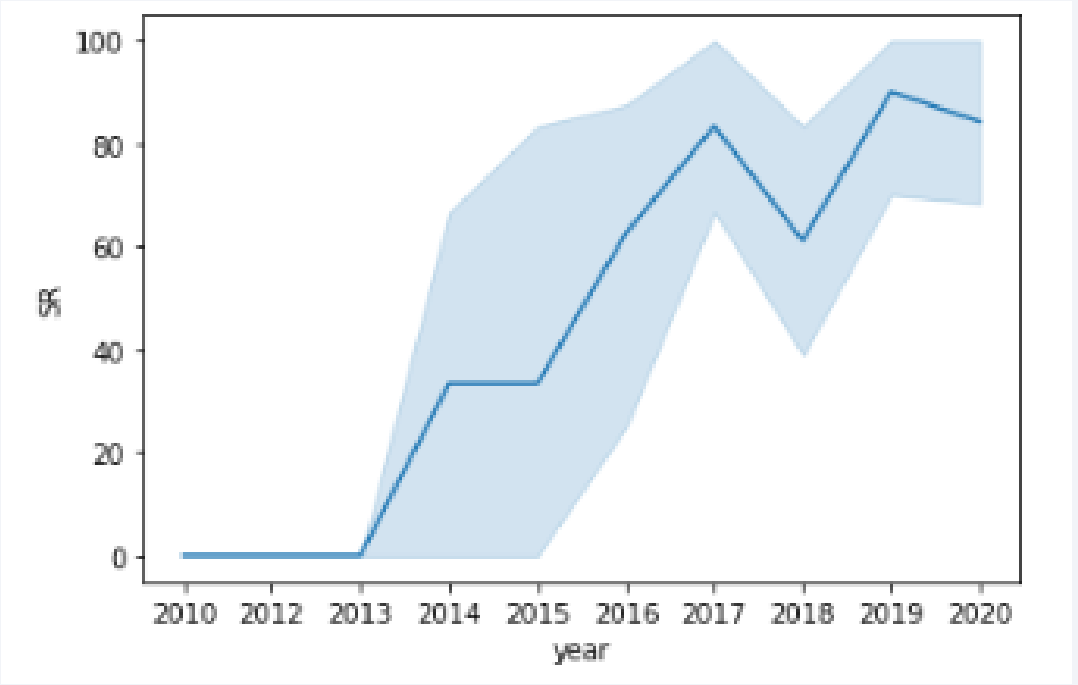
Payload vs. Orbit Type



- With low payloadmass, SSO orbit has 100% of a successful landing
- No correlation between payloadmass and GTO orbit
- Leo orbit has a great job as higher payloadmass

Launch Success Yearly Trend

- Apparently, from the years 2010 to 2013 were improved state
- Success rate dramatical increased between the years 2013 and 2017
- In 2018-2019, had a significant drop in success rate



All Launch Site Names

- Names of the unique launch sites

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

- Obtained by Distinct method which duplicates the same data name

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

index	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	08-12-2010	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)
2	22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Selected the column 'Launch_site' in dataframe, then limit with 5 numbers

Total Payload Mass

- total payload carried by boosters from NASA

sum(PAYLOAD_MASS__KG_)
48213

- Use the sum function of customer names that contain the word ‘NASA (CRS)’

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- Use the average function of the booster version that contains the word ‘F9 v1.1’

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad

Date
22-12-2015

- Selected the data whose successful landing outcomes on ground pad. After that, ordered the date and limit by 1

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

- Filtered by three attributes, Mission outcome, payloadmass,and landingout come according to the problem

Total Number of Successful and Failure Mission Outcomes

- **Total number of successful and failed mission outcomes**

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

- **Count the number that group by mission outcome**

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

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

- Use a subquery to select the boosterversion that contain max payload mass

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

- Select data from failed landing outcome occurring in the year 2015

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

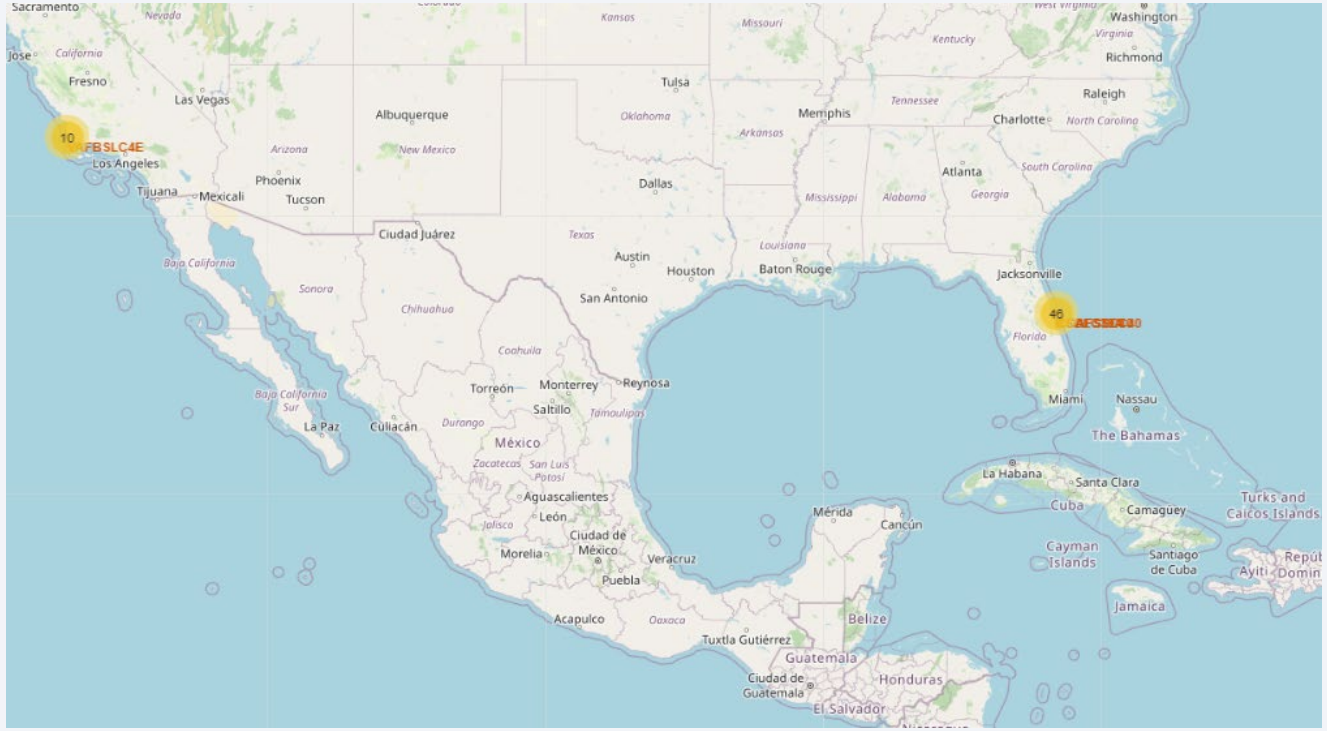
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Landing_Outcome	number
Success	20
Success (drone ship)	8
Success (ground pad)	6

- Filtered to get only successful landing outcome and order by count


Section 3
Launch Sites
Proximities Analysis

Overall



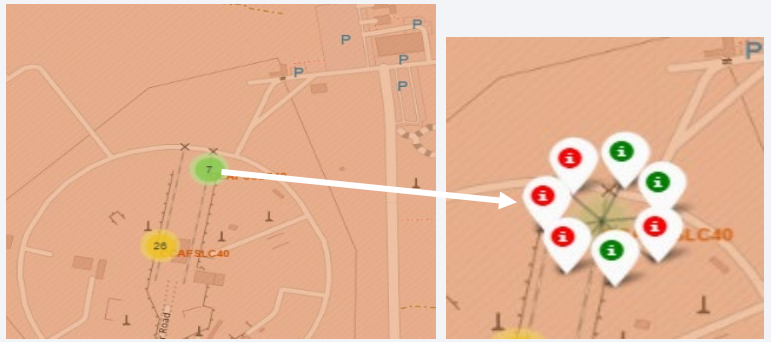
- Launch Sites are far from living place and city. However, they are close to the sea
- Apparantly, There were 56 times rocket testing

Number of Testing on each site

CCAFS LC-40

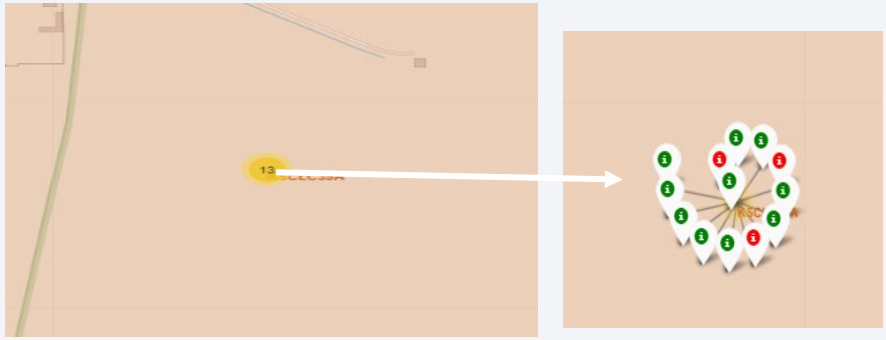
Number of Testing = 26 times
Success = 7 times
Fail = 19 times

CCAFS SLC-40



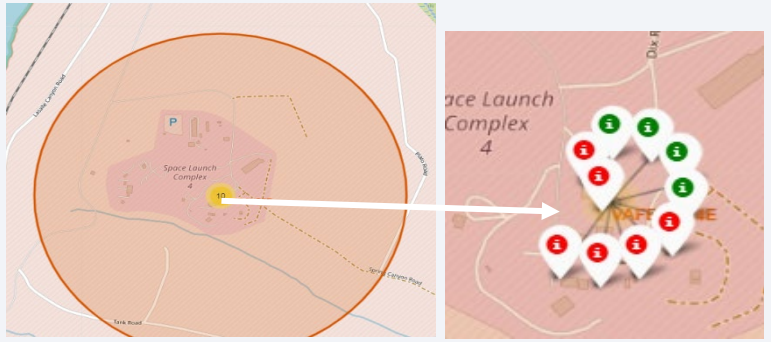
Number of Testing = 7 times
Success = 3 times
Fail = 4 times

VAFB SLC-4E



Number of Testing = 13 times
Success = 10 times
Fail = 3 times

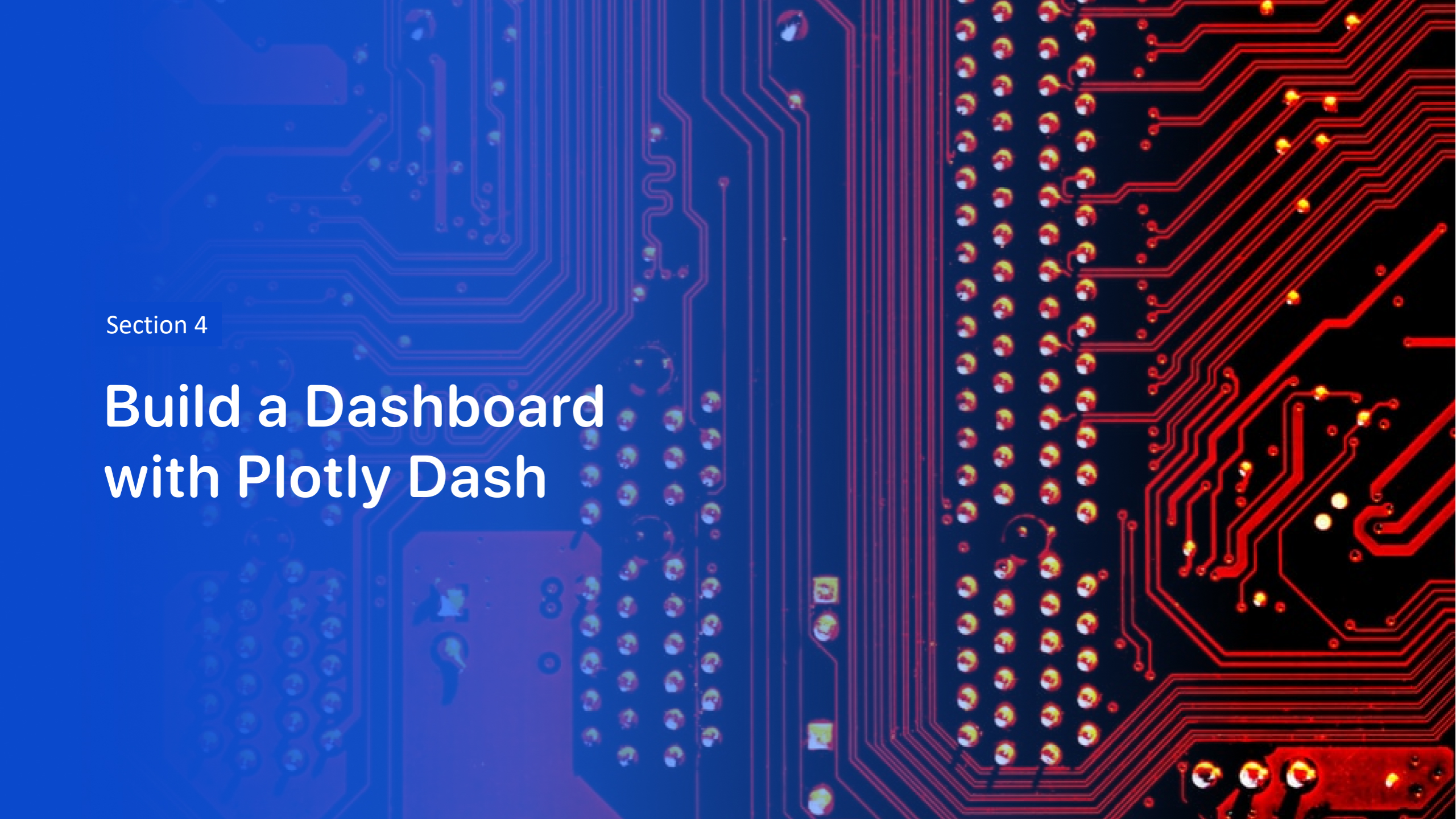
KSC LC-39A



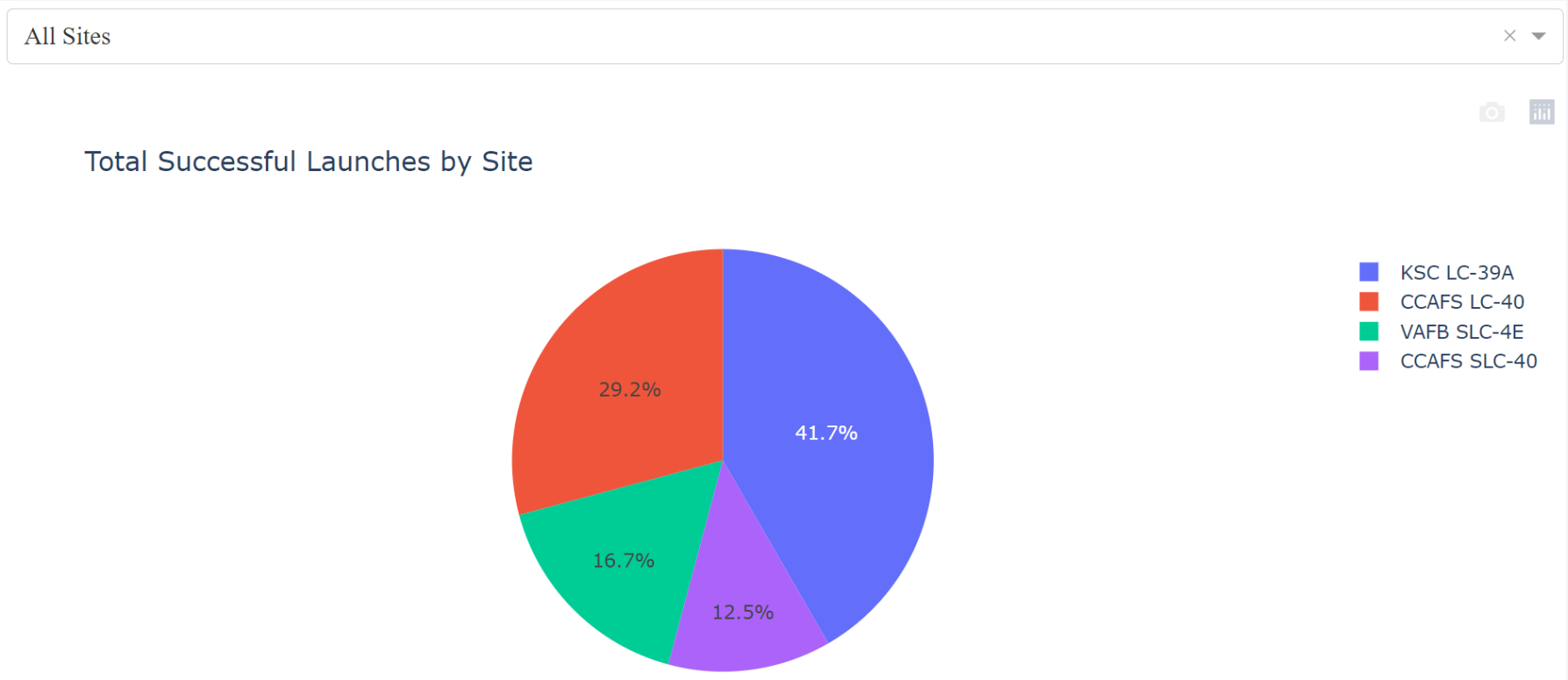
Number of Testing = 10 times
Success = 4 times
Fail = 6 times

Location

- KSC LC-39A has the highest success percentage.
- The location is suitable for testing due to the long distance from the living place and the coast


Section 4
Build a Dashboard with Plotly Dash

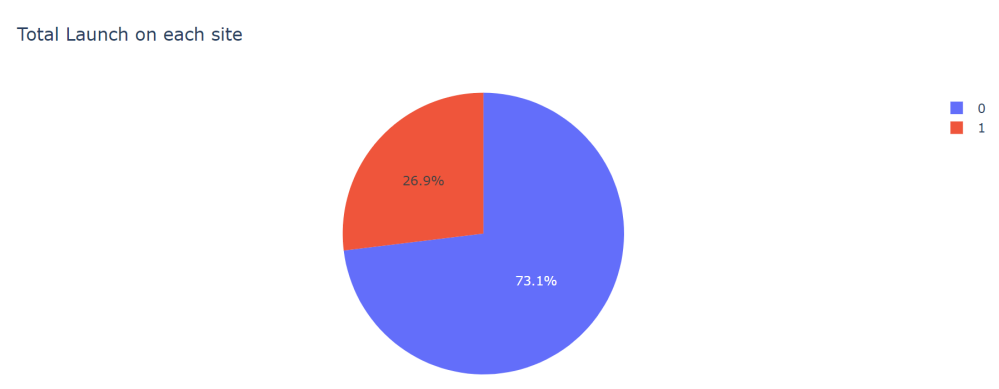
Overall pie chart



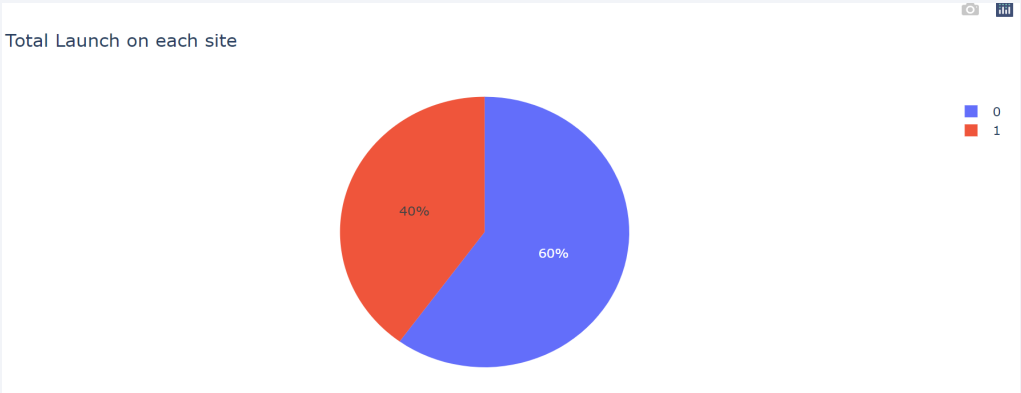
- KSC LC-39A has the highest success percentage. This was followed by CCADS LC-40 and VAFB SLC-4E, which were 29.2% and 16.7% respectively
- The worst success percentage shows on CCAFS SLC-40

Success rate

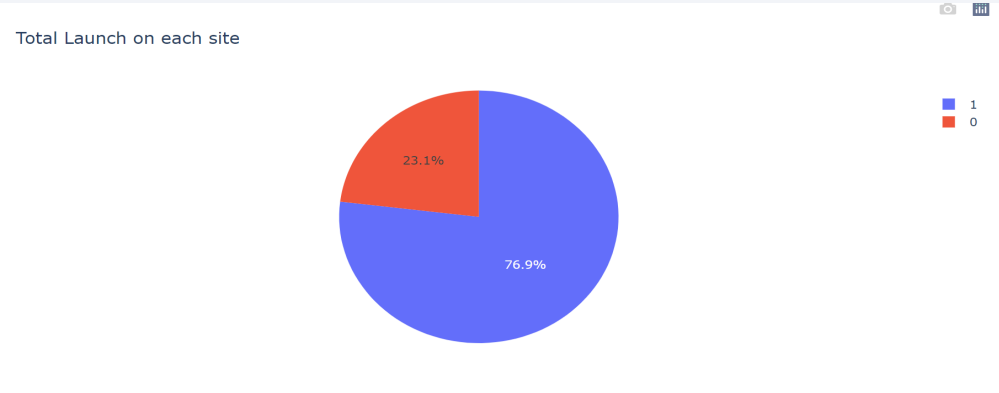
CCAFS LC-40



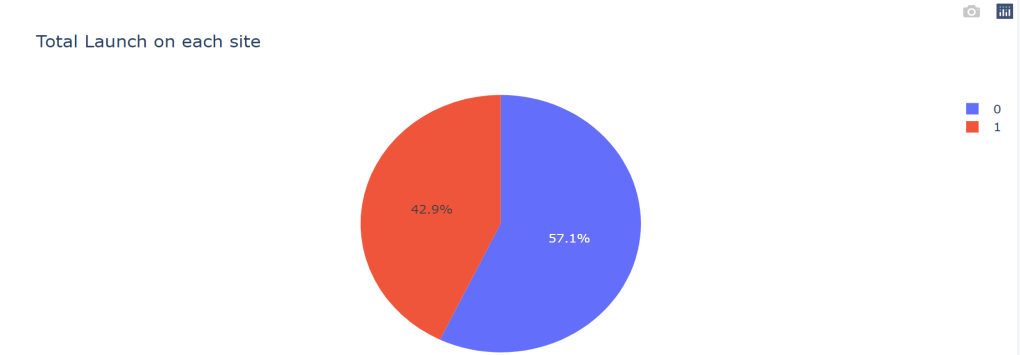
VAFB SLC-4E



KSC LC-39A



CCAFS SLC-40



Launch outcome and payload mass

- FT Booster Version has the highest number of landing successes. On the other hand, the highest number of landing failure appear in v 1.1 Booster Version
- Not enough evidence for analyzing the payload mass which weight more than 6500 kg

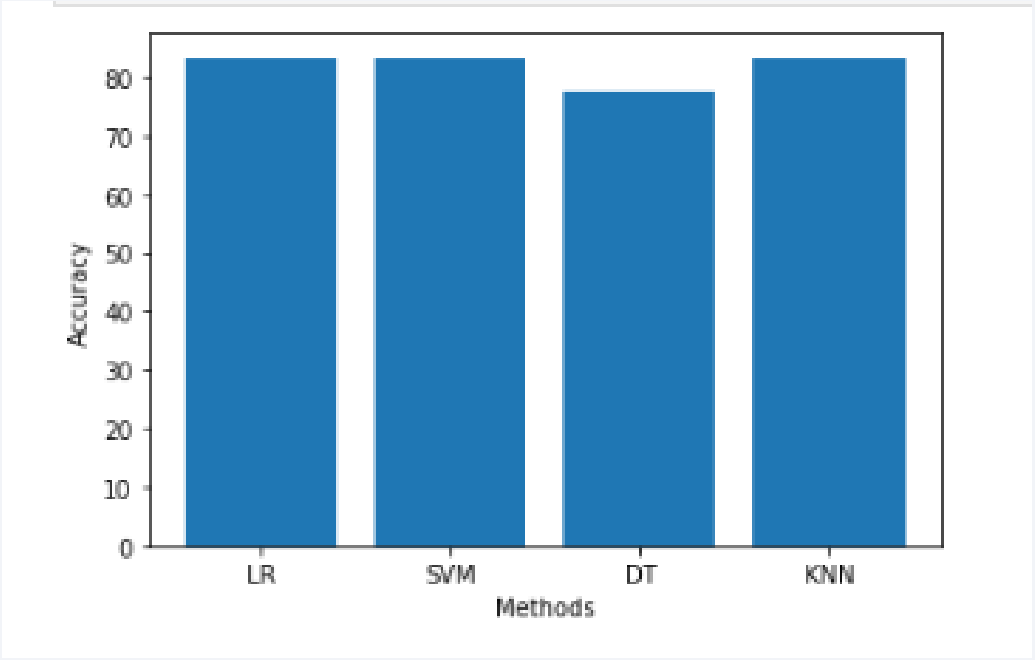
Section 5

Predictive Analysis (Classification)

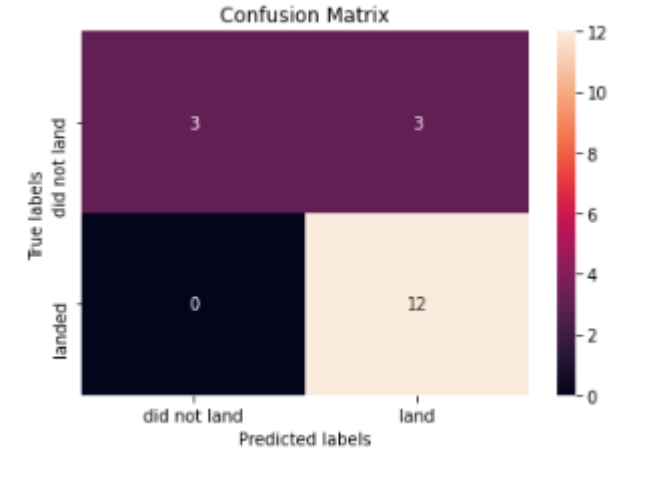
Classification Accuracy

- Four classification methods that show in the right graph contain Logistic regression, support vector machine, decision tree, and k nearest neighbor.

- There is the same accuracy for 3 models at 83.3%



Confusion Matrix



- Confusion matrix shows that the prediction for successful landing was completely collected. However, Forecasting for failure was only 50% correct.

Conclusions

- Our work is to find the best model to predict the landing outcome of the rocket on different attributes
- Data obtained in 2 ways: Rest API and Web scrapping
- Perform EDA to analyze the relationship between each feature
- The best model accuracy is 83.3%
- The best launch site is KSC LC-39A
- Need more information to improve model

Appendix

My GitHub repository = https://github.com/parklhbd45/Data-project

Instructors

- Rav Ahuja
- Alex Aklson
- Aije Egwaikhide
- Svetlana Levitan
- Romeo Kienzler
- Polong Lin
- Joseph Santarcangelo
- Azim Hirjani
- Hima Vasudevan
- Saishruthi Swaminathan
- SAEED AGHABOZORGI
- Yan Luo

Thank you!