IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sarat Smitinont

21 July 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection: retrieved data by web scrapping
  - Data Wrangling
  - Exploratory Analysis
  - Data Visualization: built an interactive dashboard
  - Predictive Analysis: Classification model
- Summary of all results
  - After applied each classification model to predict the landing outcome, I found that SVM and K nearest method had the most accurate outcome which was 77.78%

Introduction

- Project background and context
  - Launching rocket can cost the company a lot. However, SpaceX is able to reuse the first stage of the rocket that has been launched. As a result, they can reduce the cost of launching from 165 millions to 62 millions. So, it would be beneficial to the company if they know the factor that affect launching outcome. This mean they can predict the cost of launching as well.

- Problems you want to find answers
  - Which factor can be used to predict the launching out come of the first stage

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Used API to retrieve data from Wikipedia by web scrapping method
- Perform data wrangling
    - Classified landing outcome with the new label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - I separated the data into 2 sets which are train set and test set. I used train set of data to build the classification model. I also used GridSearchCV method to help tuning the model to find out the best parameters. Then, I test the model with test set of data to find out the most accurate model.

Data Collection

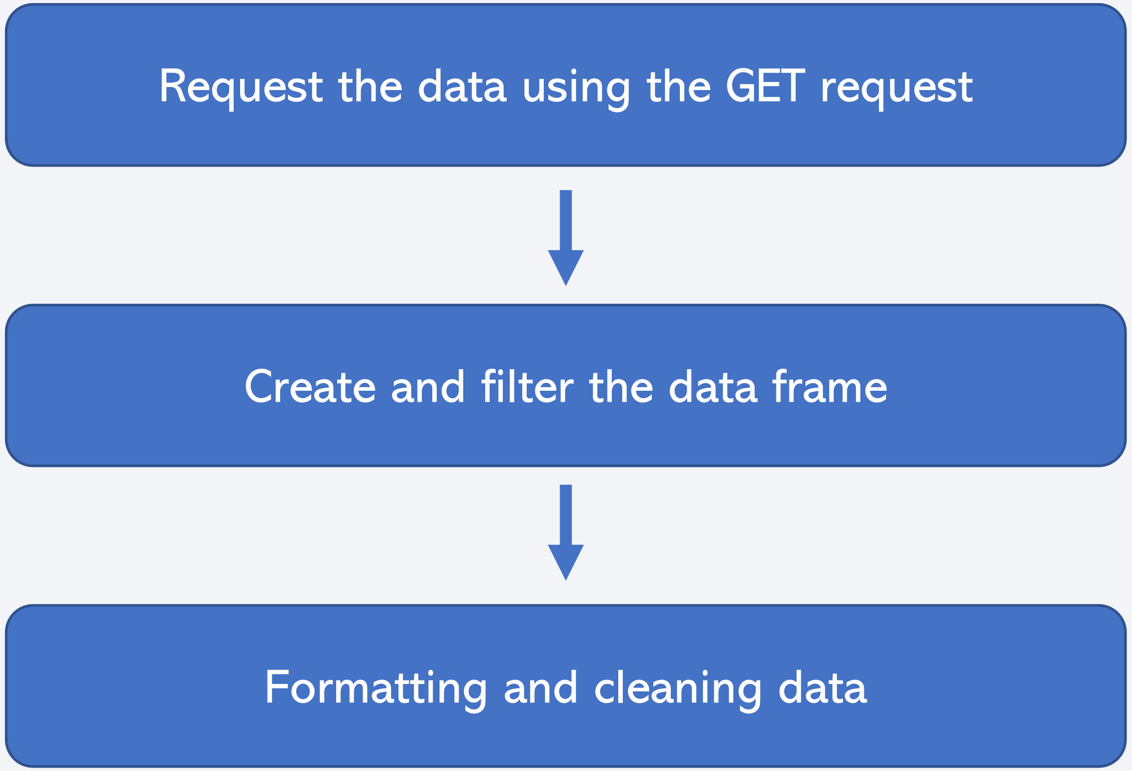
1. Used API to retrieve data

    - https://api.spacexdata.com/v4/rockets
    - https://api.spacexdata.com/v4/launchpads
    - https://api.spacexdata.com/v4/payloads
    - https://api.spacexdata.com/v4/cores
    - https://api.spacexdata.com/v4/launches/past

2. Scrapping from Wikipedia

    - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
    - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API



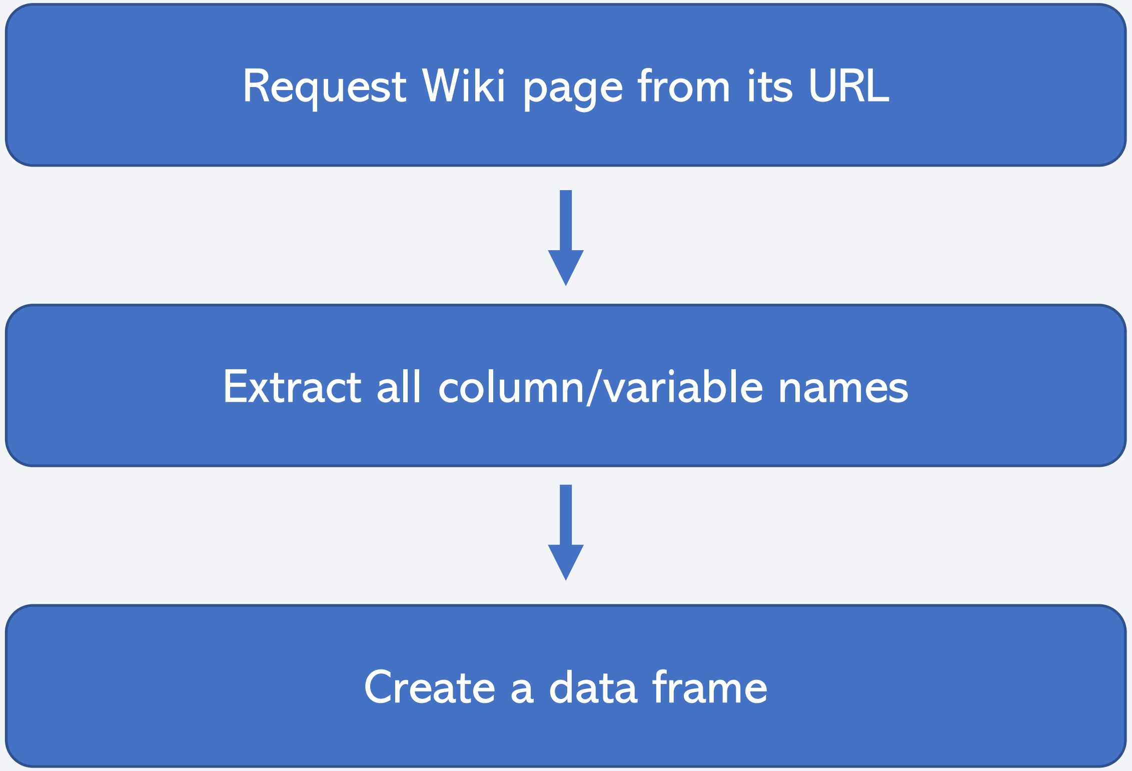
Request the data using the GET request

Create and filter the data frame

Formatting and cleaning data

GitHub: https://github.com/ninesmit/Data-Science-first-project/blob/8e58d955588bff5484bd45c081b83e13dafcfc83/Collecting%20the%20Data.ipynb

Data Collection - Scraping



Request Wiki page from its URL

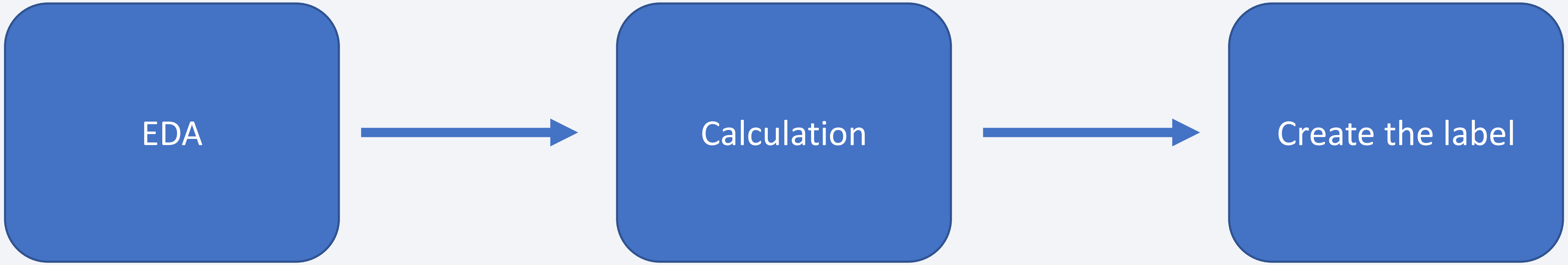
Extract all column/variable names

Create a data frame

GitHub: https://github.com/ninesmit/Data-Science-first-project/blob/e2d80b021a8c73159dbf84556c9d92bc6d80a798/Web%20scraping.ipynb

Data Wrangling

- Perform Exploratory Data Analysis to select the data which will be used to train the model
- Perform calculation by each factors
- Create the label for new dataframe



EDA

Calculation

Create the label

GitHub: https://github.com/ninesmit/Data-Science-first-project/blob/7173ae5f1cdb02fcab8e577dcfa89ebe3acce75c/Data%20Wrangling.ipynb

EDA with Data Visualization

Scatter plot (Easy to see the distribution of data)

- Relationship between payload mass and flight number
- Relationship between launch site and flight number
- Relationship between launch site and pay load mass
- Relationship between orbit type and flight number
- Relationship between orbit type and payload mass

Bar Chart (Easy to compare and understand)

- Relationship between success rate and orbit type
- Relationship between launch site and flight number
- Relationship between launch site and pay load mass

Line chart (Easy to understand)

- Relationship between success rate and years

GitHub: https://github.com/ninesmit/Data-Science-first-project/blob/1f260bd936dcee235754e218062fab69a685110b/EDA%20Visualization.ipynb

EDA with SQL

Querying the information as listed below

- The names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- the total payload mass carried by boosters launched by NASA (CRS)
- average payload mass carried by booster version F9 v1.1
- the date when the first succesful landing outcome in ground pad was achieved
- the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- the total number of successful and failure mission outcomes
- the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in 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/ninesmit/Data-Science-first-project/blob/668191024de66fd19974ccb50fa23d7768b090f2/EDA%20SQL.ipynb

Build an Interactive Map with Folium

Three types of object can be seen on the map

1. Markers
   - Normal Mark: use to mark the launch site
   - Cluster Mark: use to mark the total success and failure of each launch site
2. Circles
   - Use to show the area of the launch site
3. Lines
   - Use to show the distance and direction from launch site to nearest coastline, highway, railway, and city

GitHub: https://github.com/ninesmit/Data-Science-first-project/blob/f0d82703584b6c0a2433836724b30a9156a13ed2/Foilium%20Maps.ipynb

Build a Dashboard with Plotly Dash

Two types of graph are shown on the dashboard

1. Pie chart
   - Showed the success of launching by each site
   - Showed the proportion between success and failure of each site
2. Scatter plot
   - Showed the relationship between launching outcome and payload mass of each booster version

GitHub: https://github.com/ninesmit/Data-Science-first-project/blob/e4e38c324a878d358b3bdda7cdba0bdbb72f57c2/Plotly%20Dashboard.ipynb

Predictive Analysis (Classification)

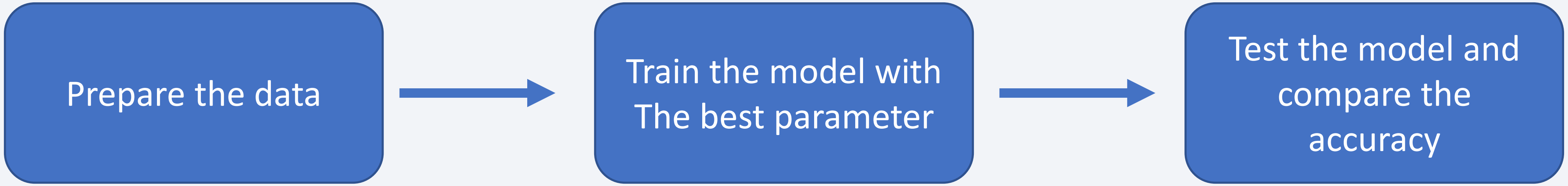
Analysis process

1. Prepare the data to be used for training model
   - Building the dataframe
   - Normalizing the data
   - Split the data into 2 set: train set and test set

2. Train all 4 models and find the best parameter using GridsearchCV

   - Use train set to fit the model consist of k nearest model, decision tree, SVM, and logistic reression

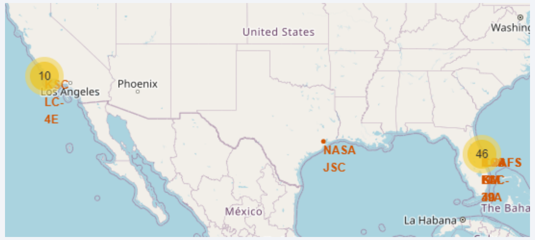
3. Test and compare the accuracy

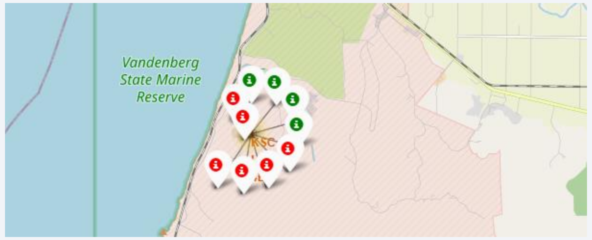

Prepare the data
Train the model with The best parameter
Test the model and compare the accuracy

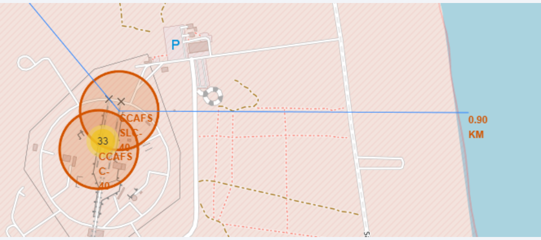
GitHub: https://github.com/ninesmit/Data-Science-first-project/blob/9b2739d1dbcf47c422faf0c54888b48c5b418bbd/Predictive%20Analysis.ipynb

Results

- Exploratory data analysis results
  - Most of Falcon 9 can carry high payload mass
  - Failure mission outcome happened once
  - The success rate of landing has increased over time
  - Two booster version failed landing in drone ship
- Interactive analytics demo in screenshots

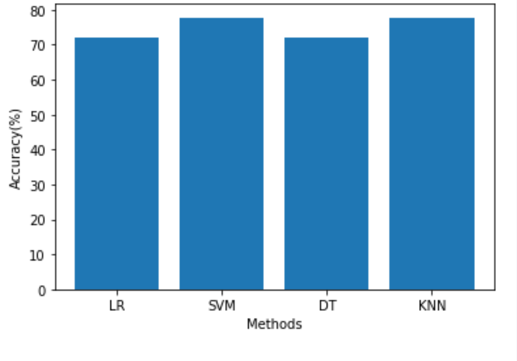






- Predictive analysis results
  - According to the graph, support vector machine and K nearest have the highest accuracy

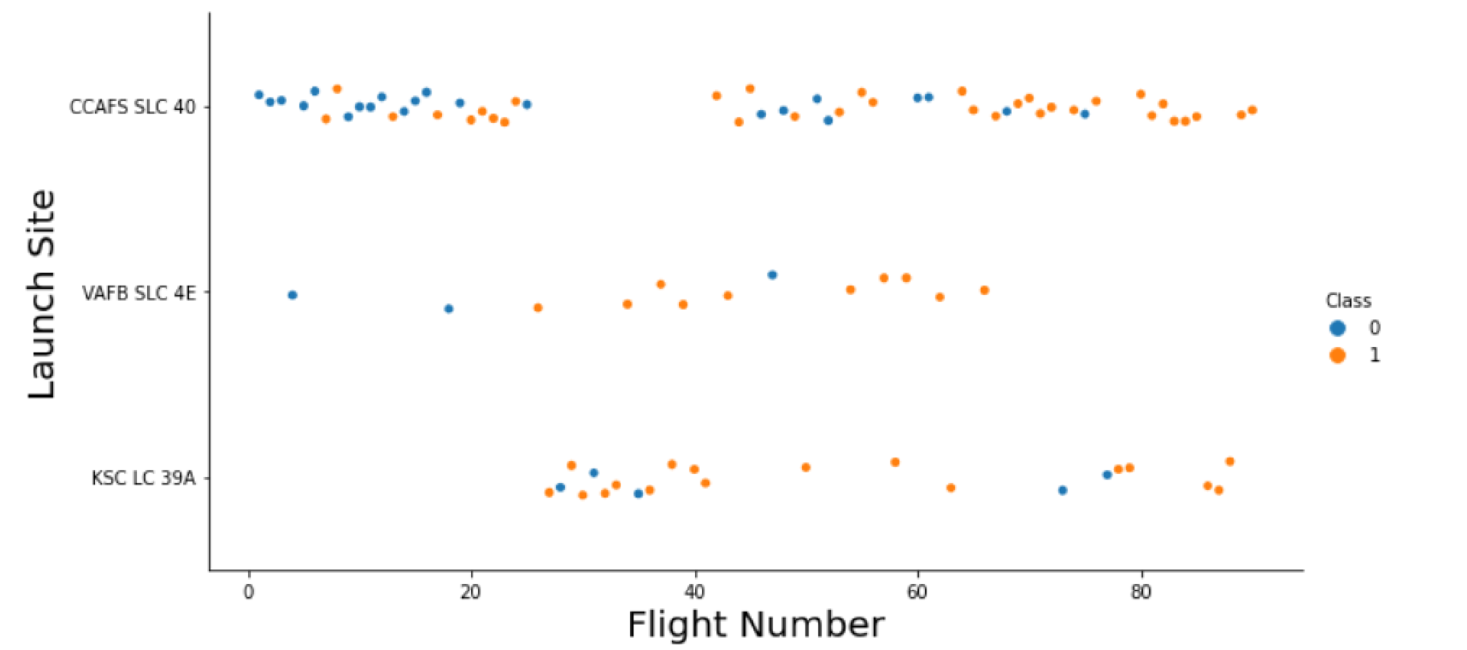
Which is 78 percent, followed by logistic regression and decision tree.



Section 2

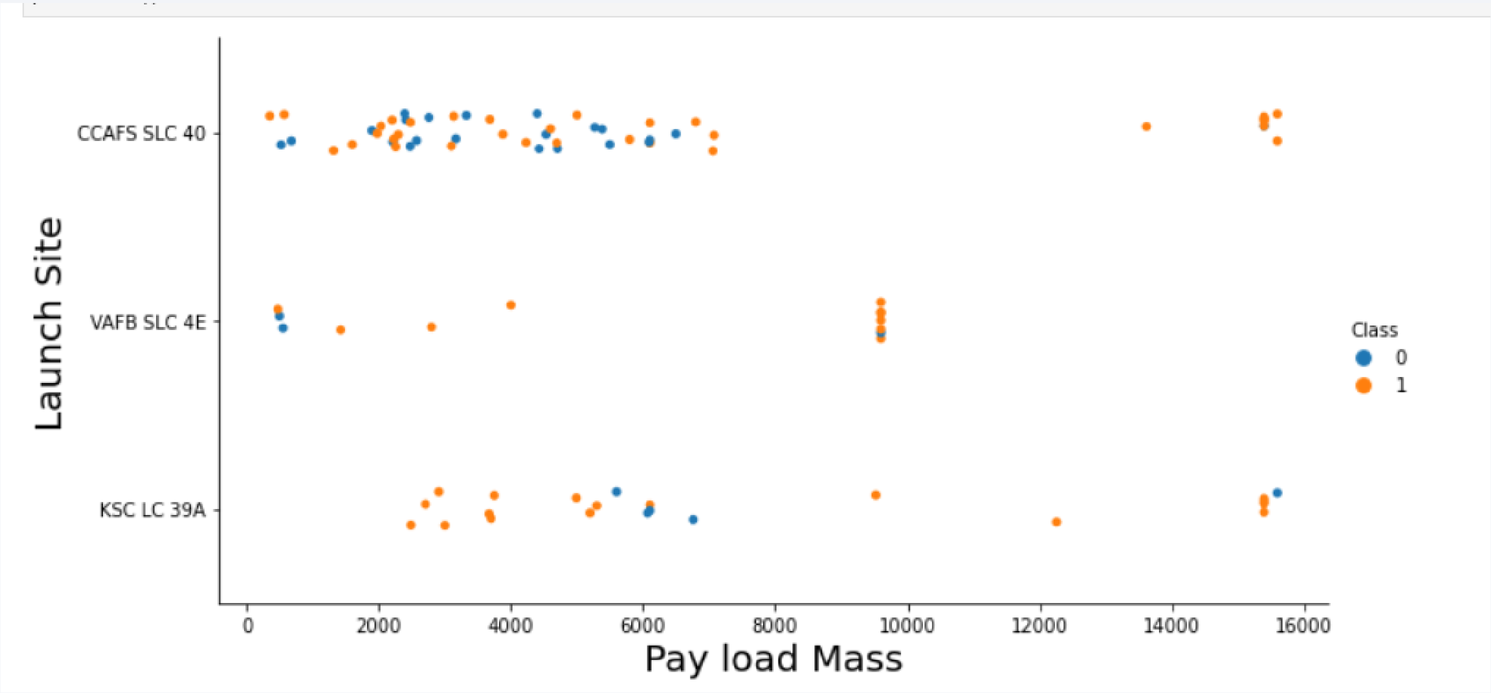
Insights drawn from EDA

Flight Number vs. Launch Site



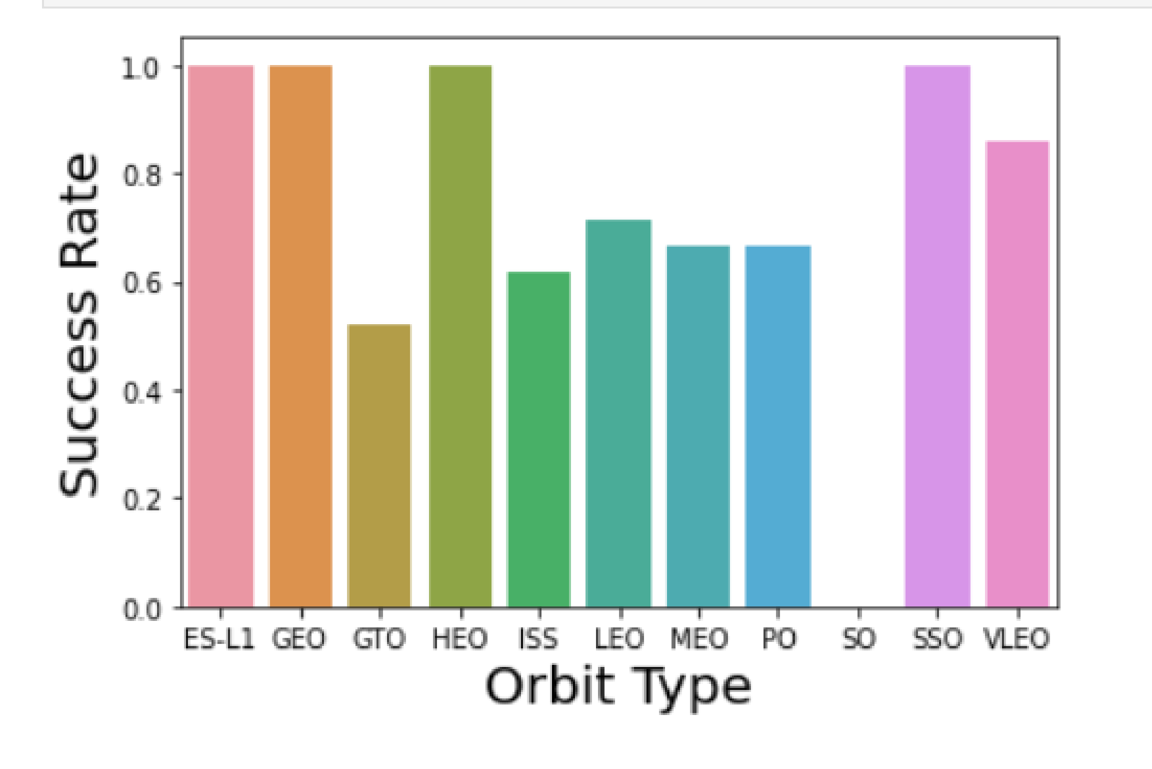
- Site VAFB SLC 4E has the most success rate over the others
- The success rate at site CCAFS SLC 40 has improved over time

Payload vs. Launch Site



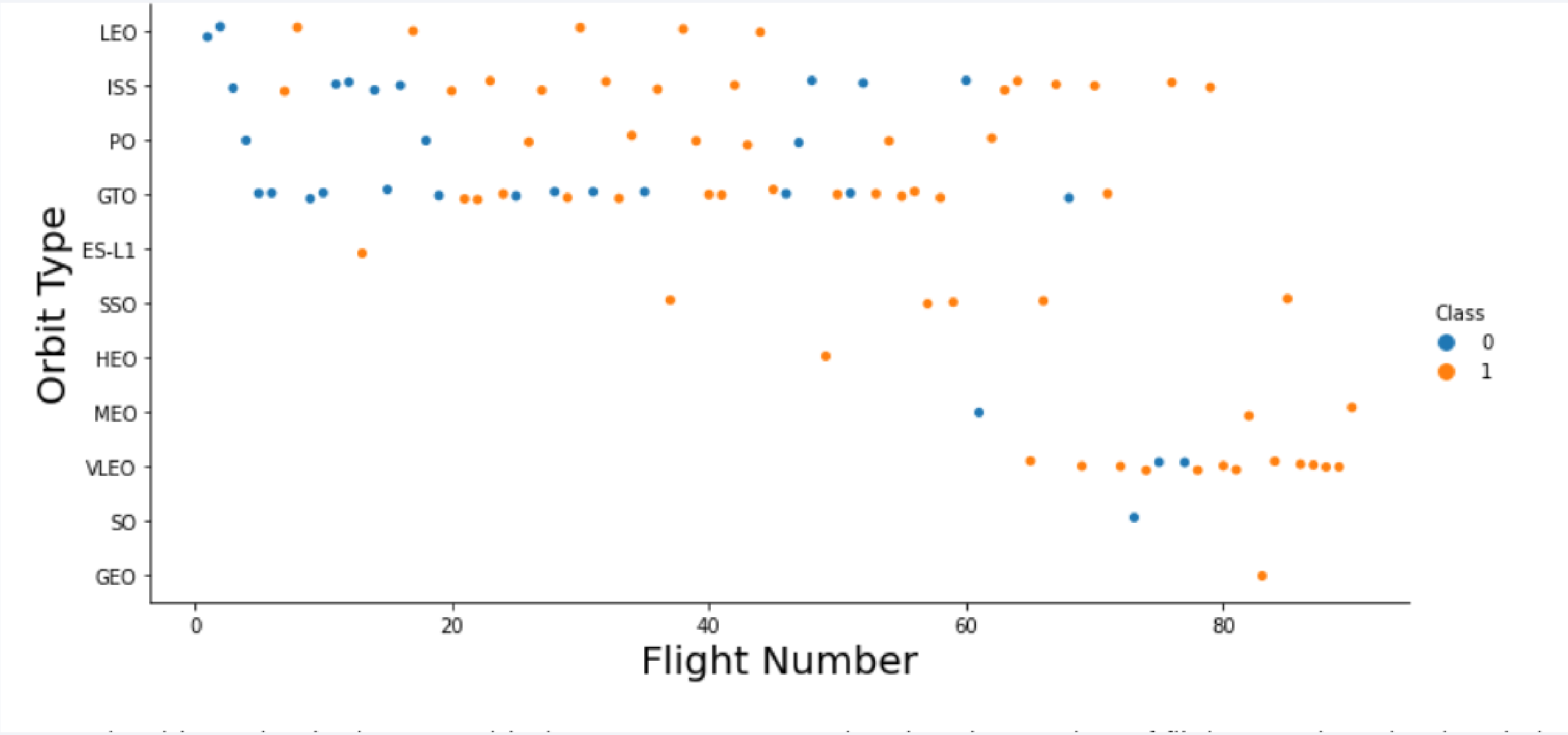
- Site CCAFS SLC 40 has 100% success rate at high payload mass
- Site KSC LC 39A has 100% success rate between 2000 and 4000 payload mass
- Site VAFB SLC 4E has almost 100% success rate at around 10000 pay load mass

Success Rate vs. Orbit Type



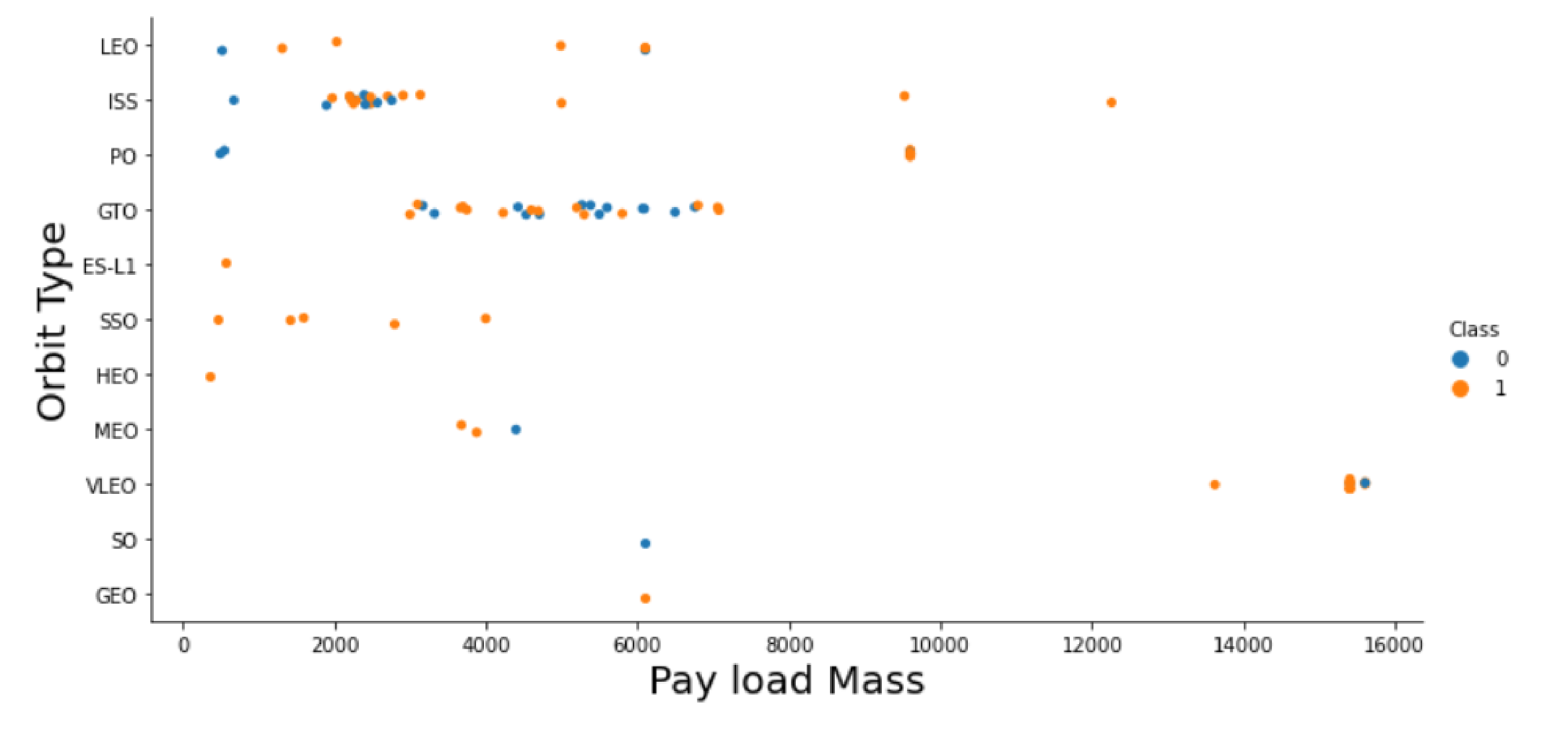
- Orbit ES-L1, GRO, HEO, and SSO have 100% success rate
- The success rate has dropped significantly at orbit ISS, LEO, MEO, and PO
- The lowest success rate was at orbit GTO

Flight Number vs. Orbit Type



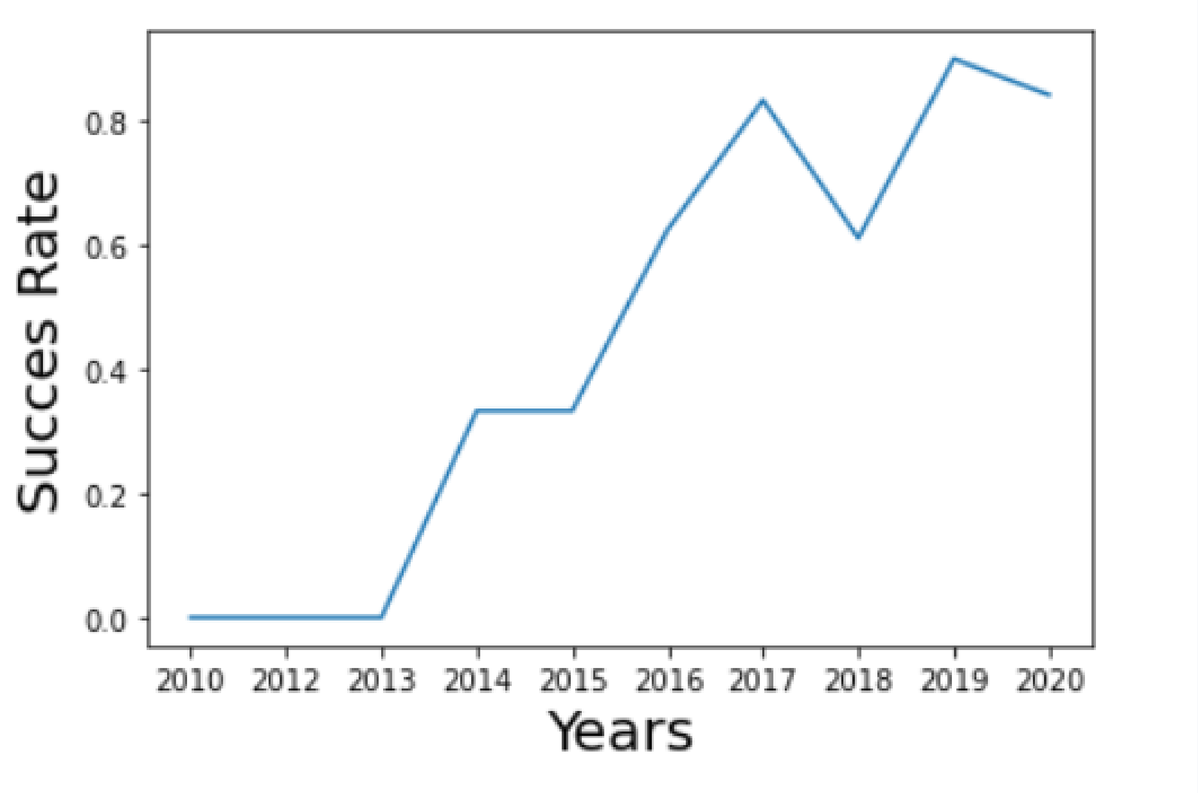
- The success rate has increased over the number of flight
- Orbit SSO has 100% success rate

Payload vs. Orbit Type



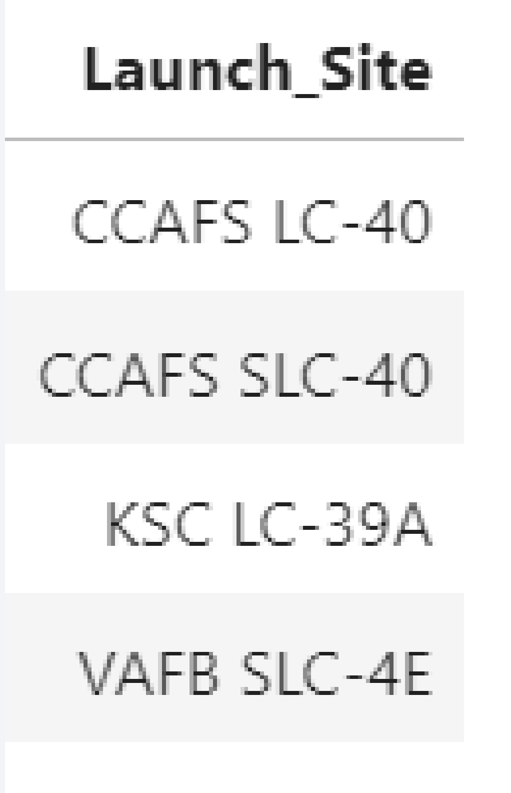
- Orbit SSO has 100% success rate at below 4000 payload mass
- Orbit LEO and ISS have 100% success rate at over 4000 payload mass

Launch Success Yearly Trend



- The success rate has increased obviously over time
- The success rate dopped notably from 2017 to 2018

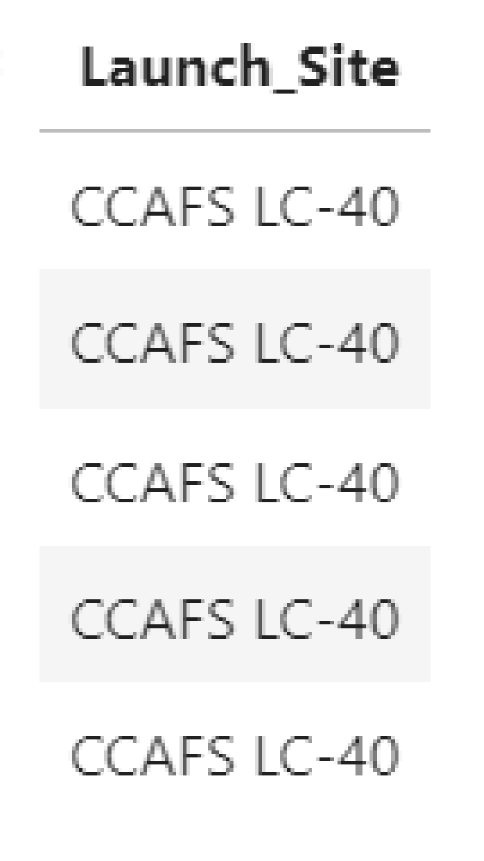
All Launch Site Names



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

- Query the “ Launch Site” column from table SPACEXTBL

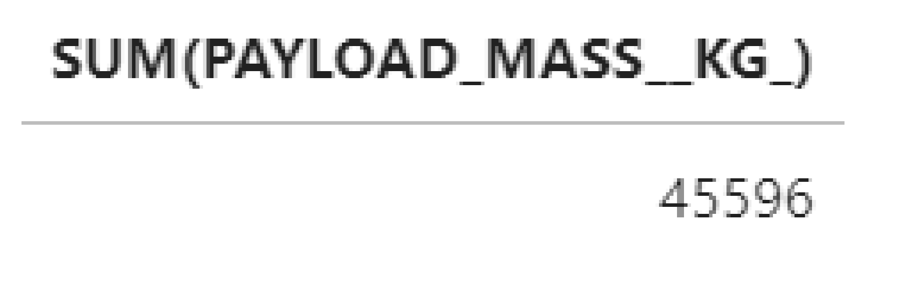
Launch Site Names Begin with 'CCA'



Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

- Query the “ Launch Site” column from table SPACEXTBL with the condition that it has to start with CCA

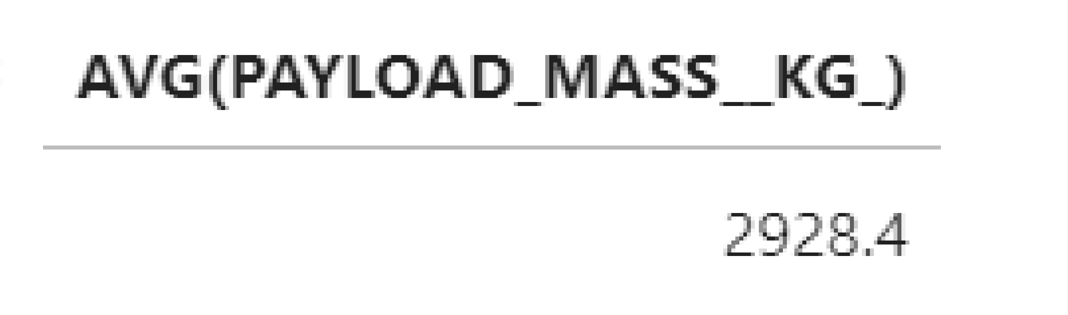
Total Payload Mass



SUM(PAYLOAD_MASS__KG_)
45596

- Query the “Payload mass” column with summation function from table SPACEXTBL with the condition as NASA(CRS) is customer

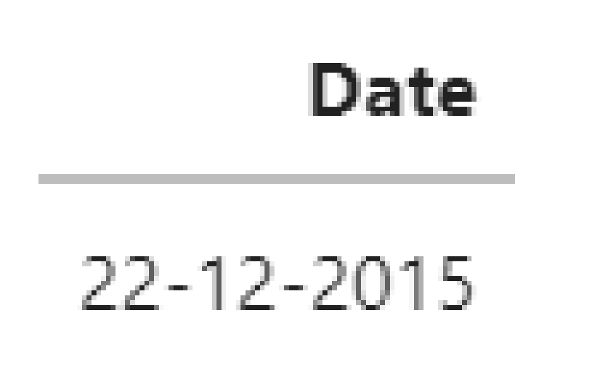
Average Payload Mass by F9 v1.1



AVG(PAYLOAD_MASS__KG_)
2928.4

- Query the “Payload mass” column with average function from table SPACEXTBL with the condition as Booster version is F9 V1.1

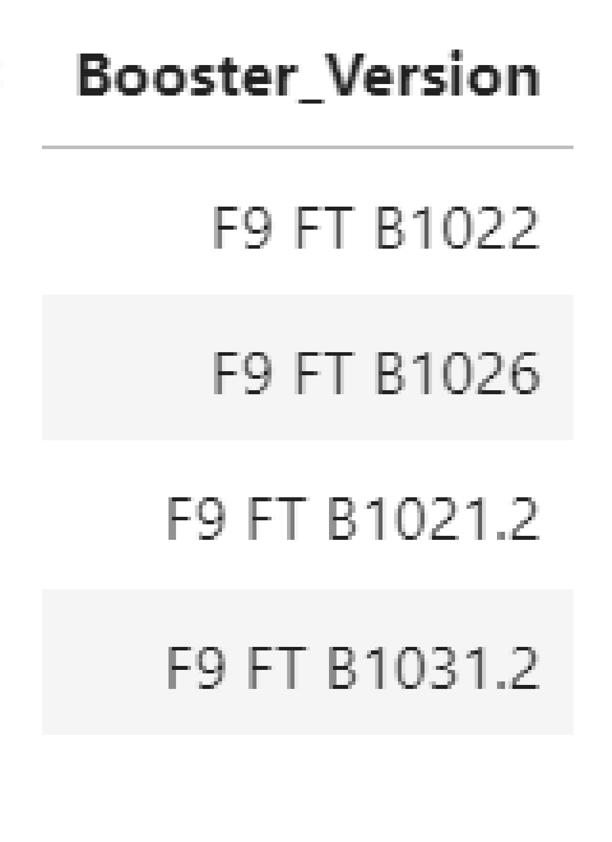
First Successful Ground Landing Date



Date
22-12-2015

- Query the “DATE” column from table SPACEXTBL with the condition as landing outcome is success (ground pad)
- Order the Queried data by DATE and limit the number of row to 1

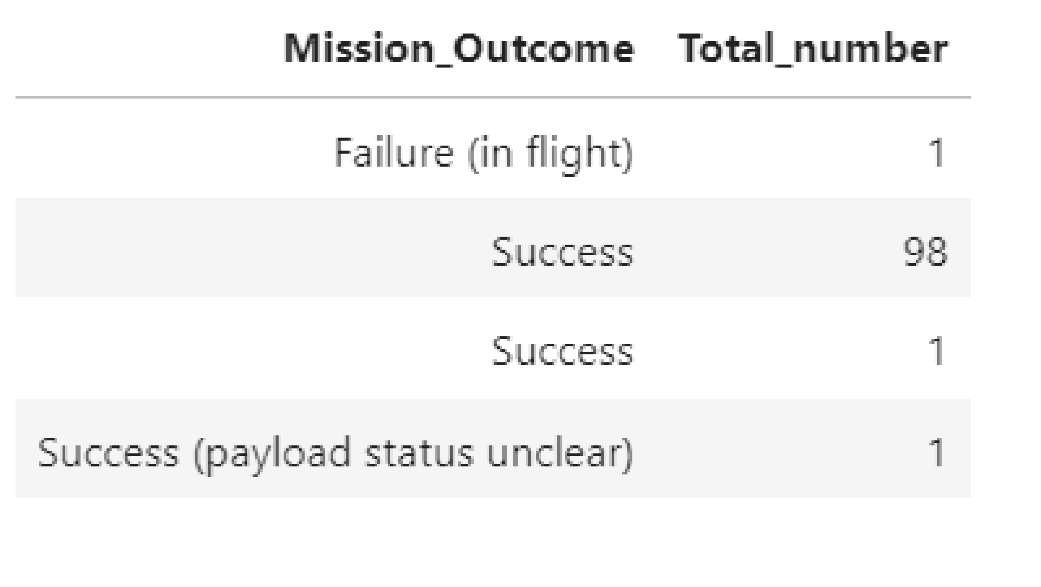
Successful Drone Ship Landing with Payload between 4000 and 6000



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

- Query the “Booster version” column from table SPACEXTBL with the condition as landing outcome is success (drone ship)
- And another condition is payload need to be in between 4000 and 6000

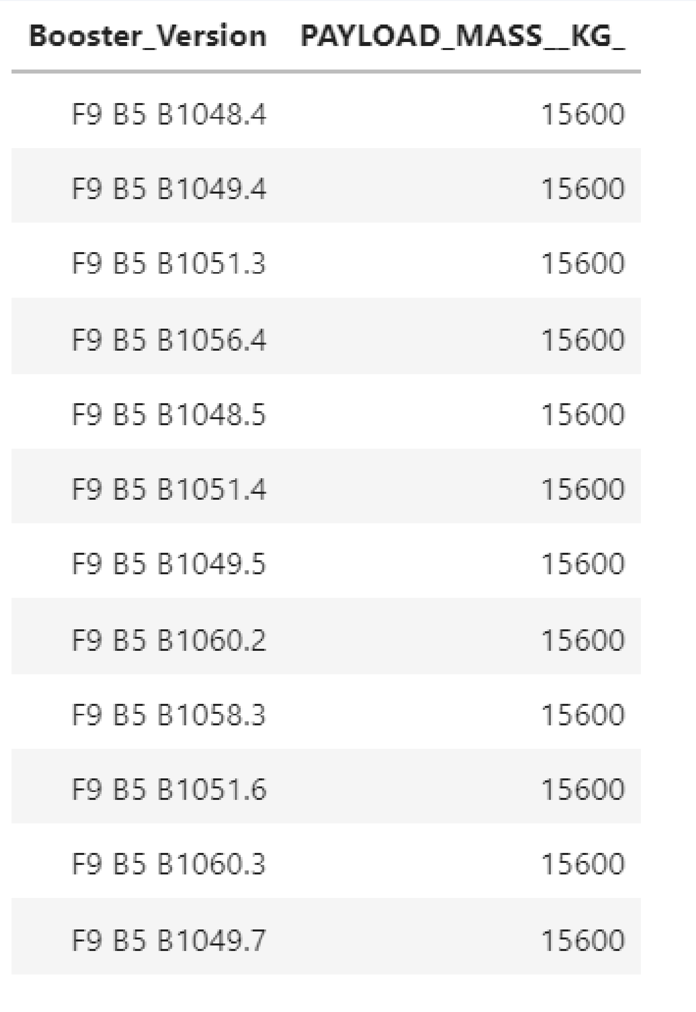
Total Number of Successful and Failure Mission Outcomes



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

- Query the “Mission outcome” column with and without count function from table SPACEXTBL and group the data with mission outcome

Boosters Carried Maximum Payload

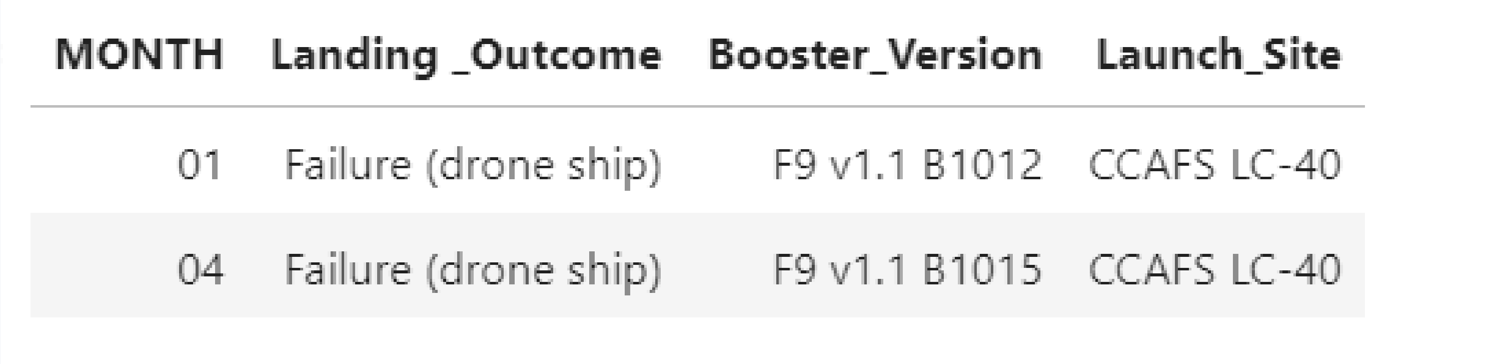


Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- Query the “Booster version” and “ Payload mass” column from table SPACEXTBL

- Query the “Payload mass” with maximum function and use it as a sub-query

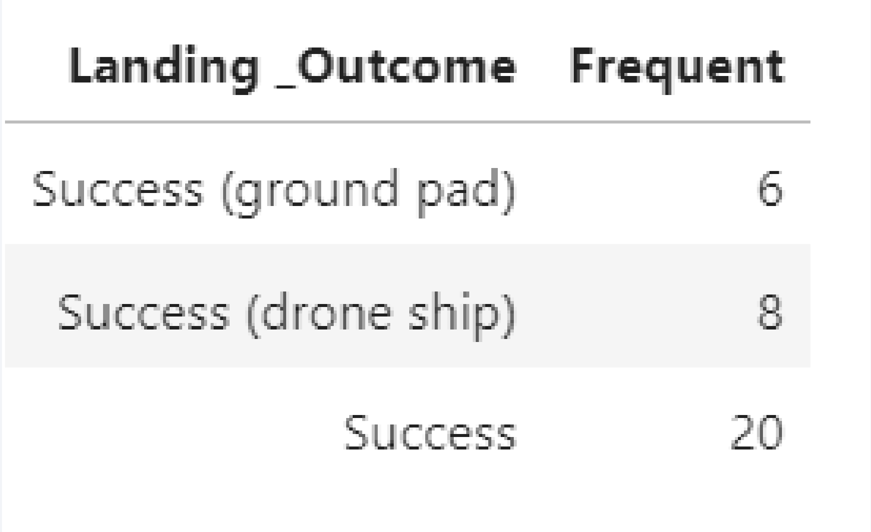
2015 Launch Records



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

- Query the “DATE” column with condition as in year 2015 from table SPACXTBL and set condition as landing outcome is Failure (drone ship)

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



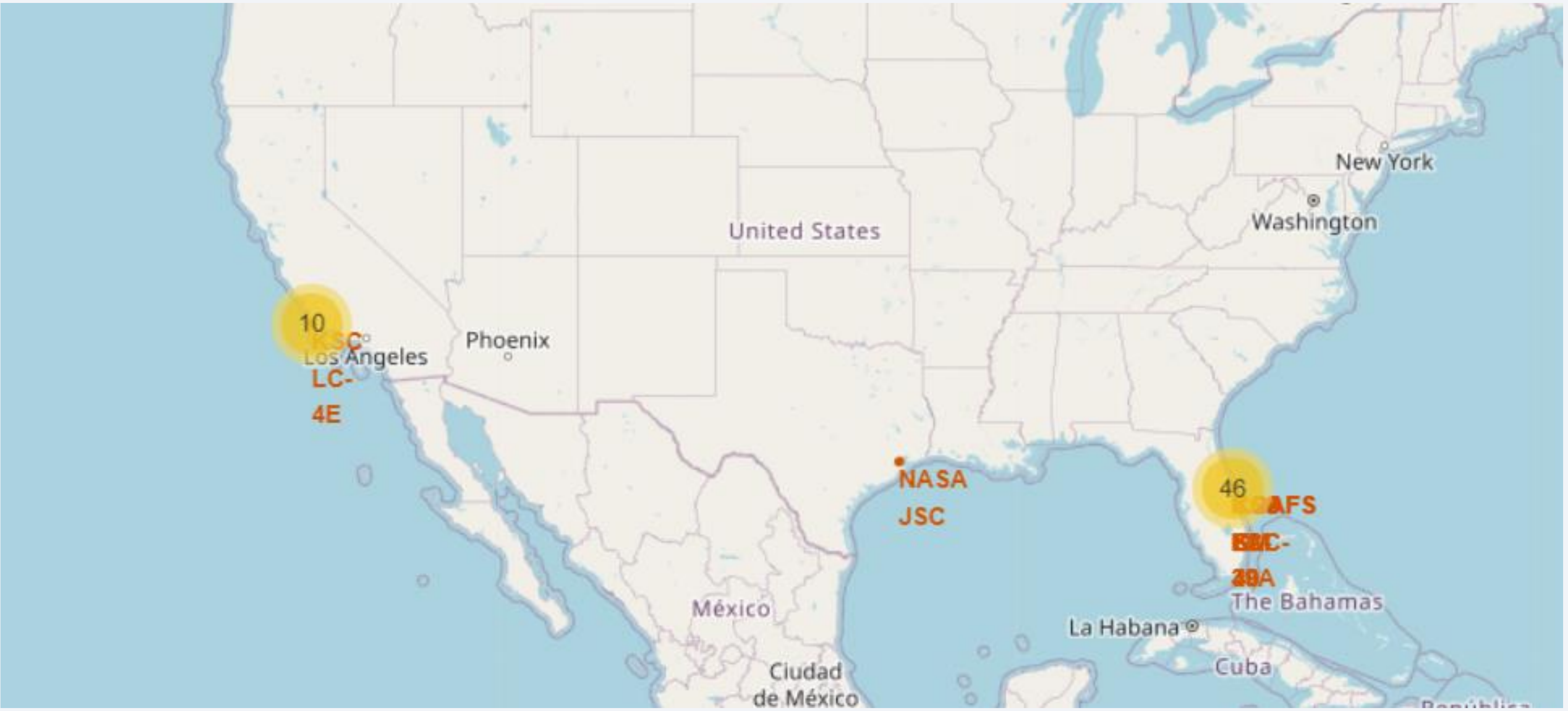
Landing _Outcome	Frequent
Success (ground pad)	6
Success (drone ship)	8
Success	20

- Query the “Landing outcome” column with and without count function from table SPACXTBL
- Group data by landing outcome

Section 3

Launch Sites Proximities Analysis

Overview of the Launch site



- There are 2 main area of launch site located near the coastline on both east and west side of USA
- There are 56 launch times

Success and Failure rate of each site

KSC
LC-
4E
Launch
Tower

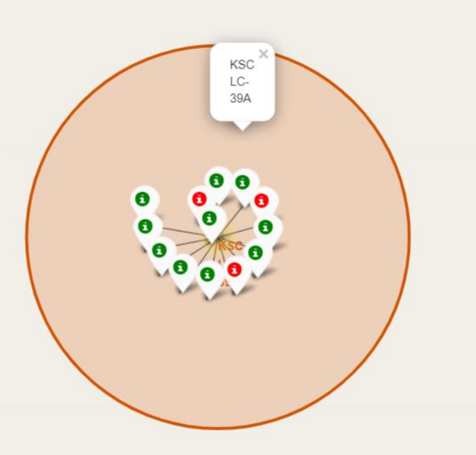
Site: KSC LC-4E

- Launch times: 10
- Success times: 4
- Failed times: 6
- Success rate: 40%

CCAFS
SLC-40
26

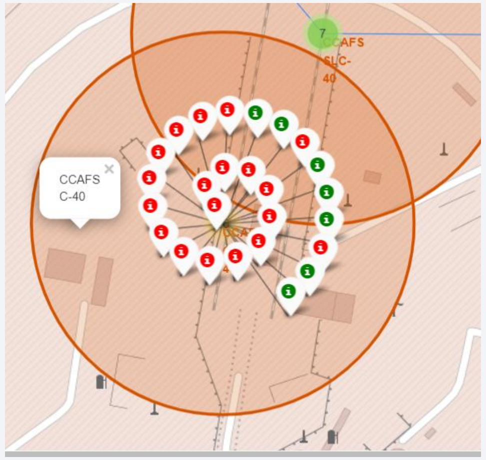
Site: CCAFS SLC-40

- Launch times: 7
- Success times: 3
- Failed times: 4
- Success rate: 57%



Site: KSC LC-39A

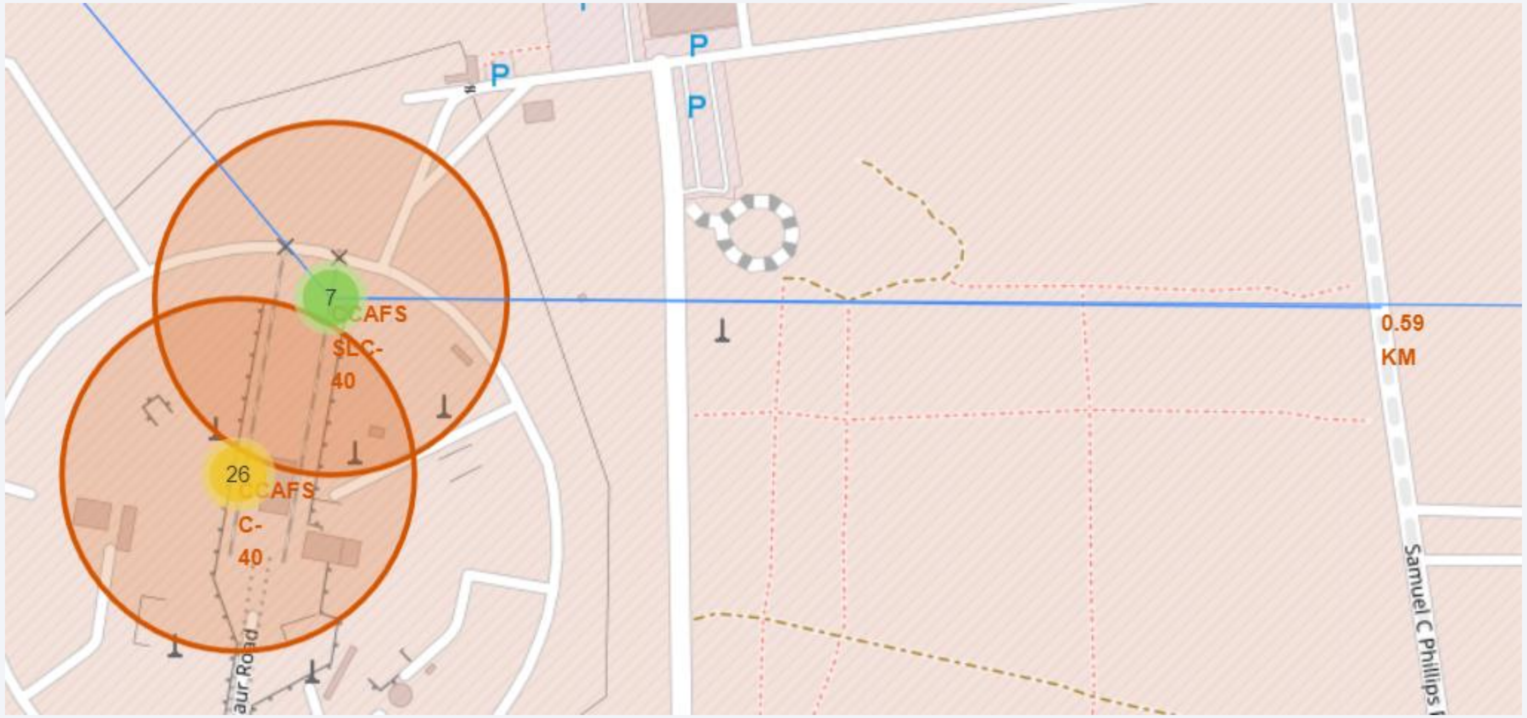
- Launch times: 13
- Success times: 10
- Failed times: 3
- Success rate: 77%



Site: CCAFS C-40

- Launch times: 26
- Success times: 7
- Failed times: 19
- Success rate: 27%

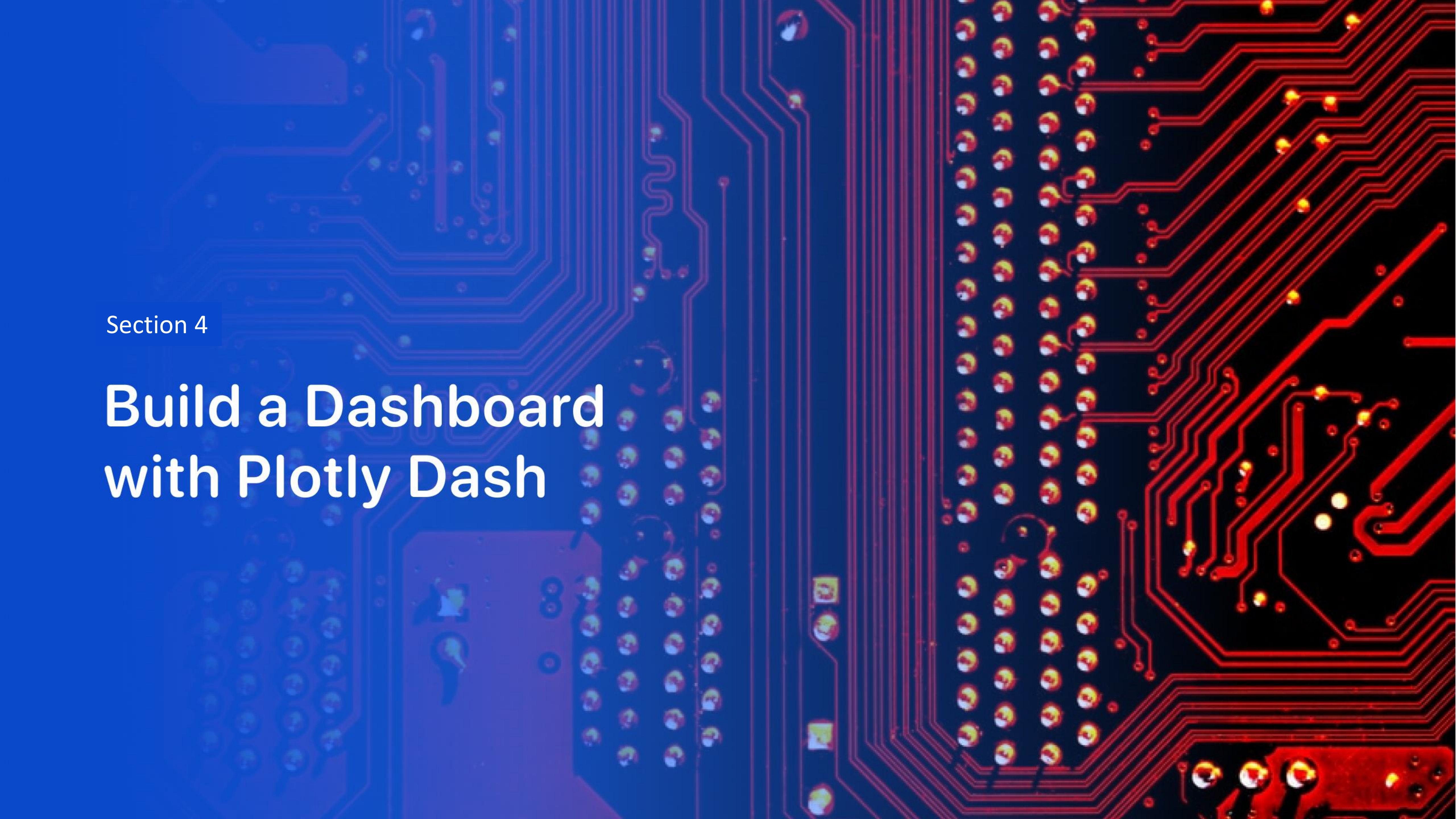
Distance between Launch site and highway



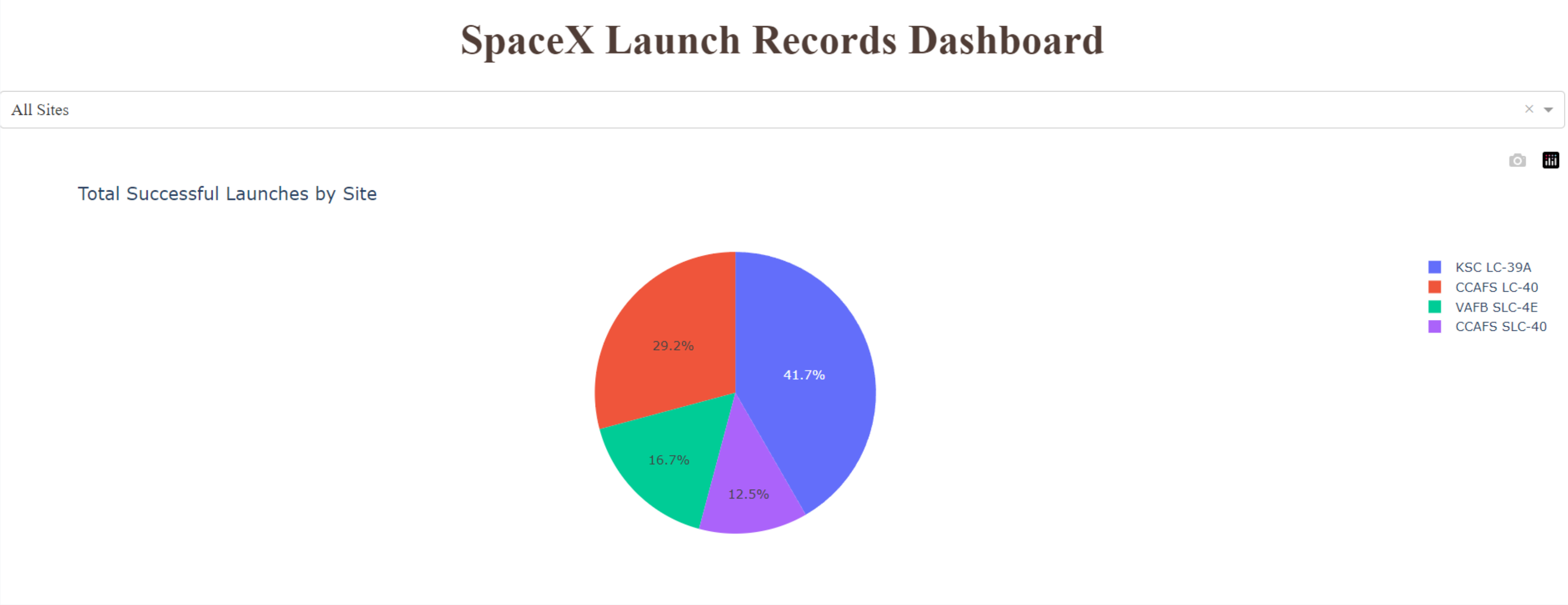
- Selected launch site: CCAFS SLC-40
- Distance from highway: 0.59 KM.

Section 4

Build a Dashboard with Plotly Dash



Overview of success rate

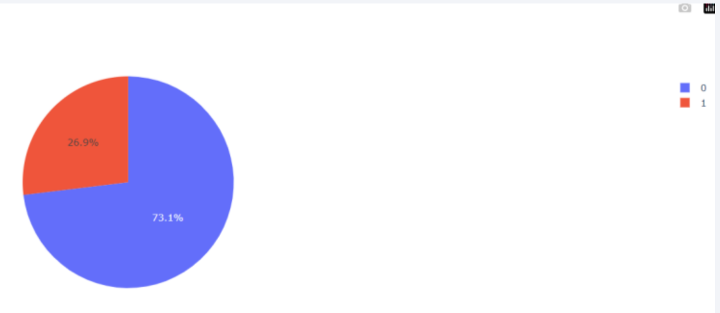


SpaceX Launch Records Dashboard

All Sites

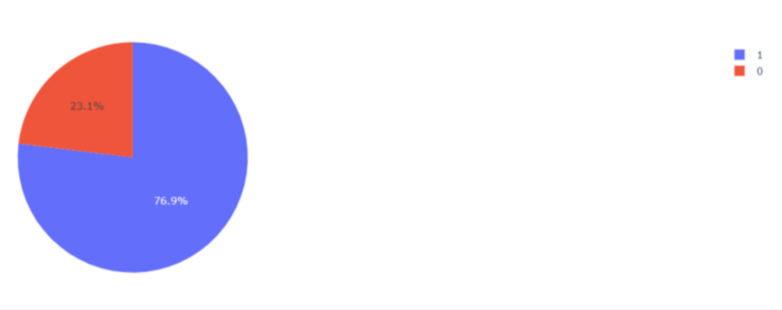
- Site KSC LC-39A has the most success times over other site.
- Site CCAFS SLC-40 has the least success times.

Success rate of each site



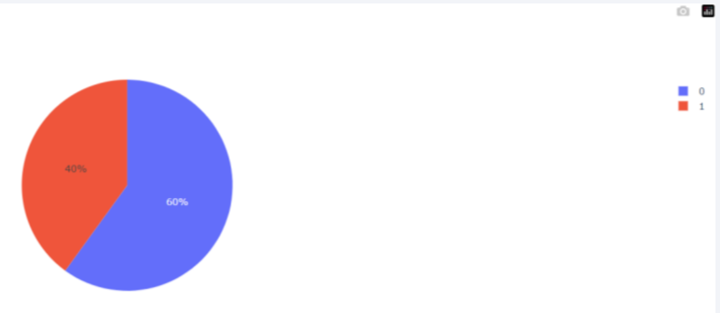
Site: CCAFS LC-40

- Success rate: 26.9%
- Failure rate: 73.1%



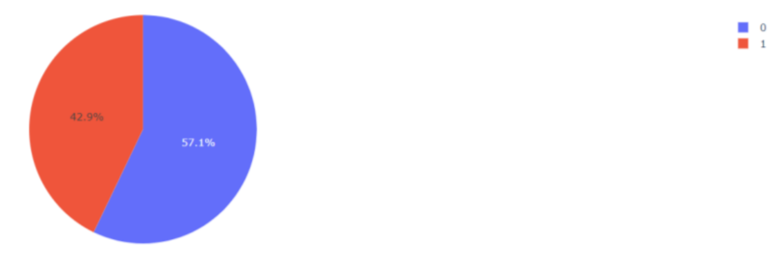
Site: KSC LC-39A

- Success rate: 76.9%
- Failure rate: 23.1%



Site: VAFB SLC-4E

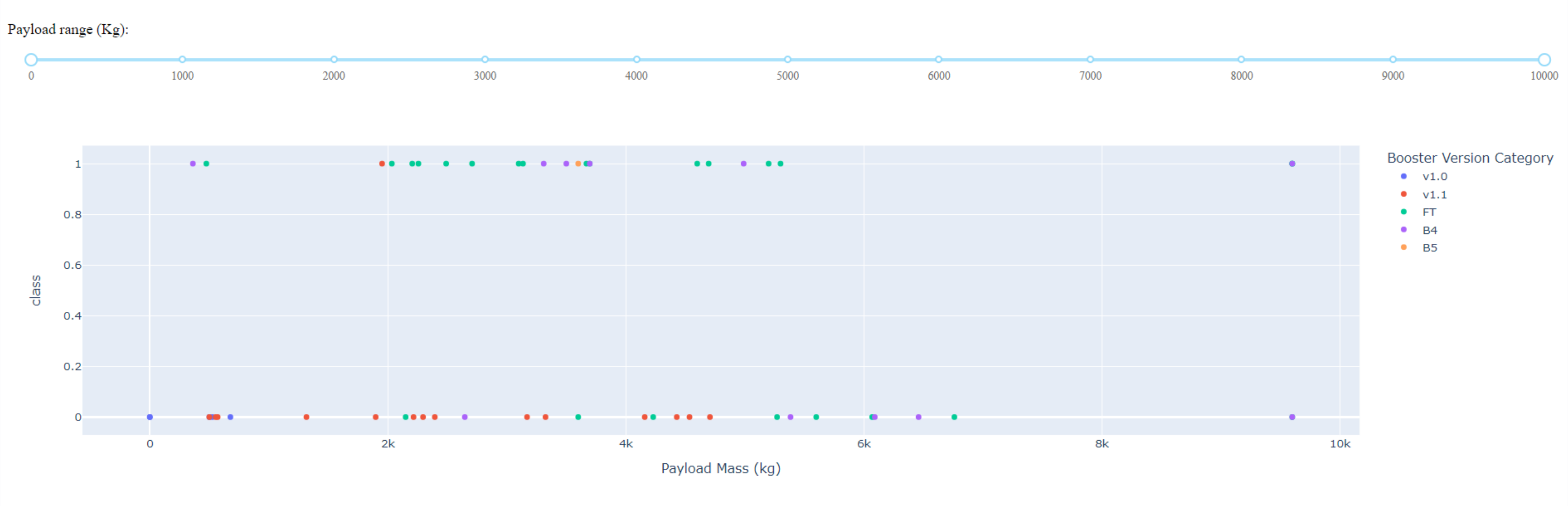
- Success rate: 40%
- Failure rate: 60%



Site: CCAFS SLC-40

- Success rate: 42.9%
- Failure rate: 57.1%

The best booster version

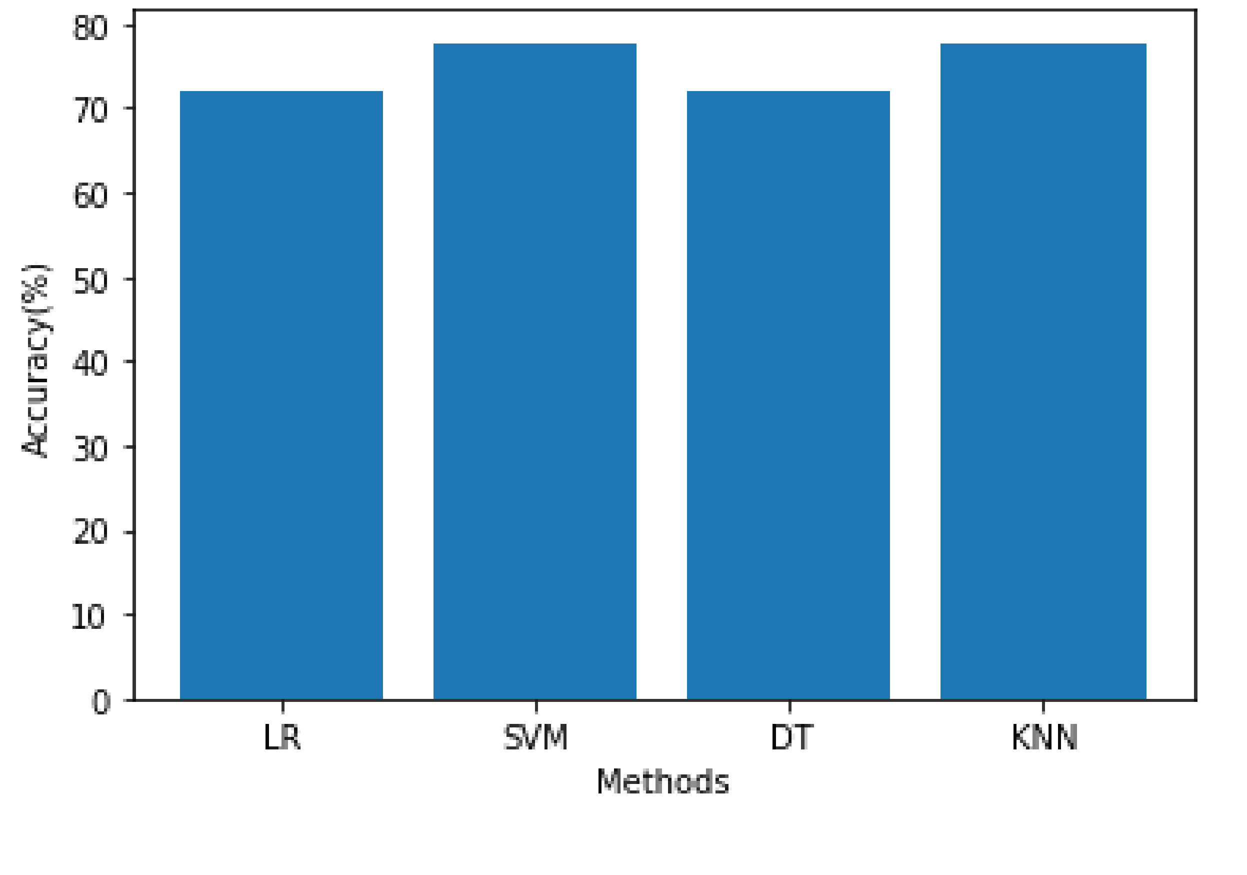


- FT booster achieve the highest success rate from range 3000 kg to 10000 kg

Section 5

Predictive Analysis (Classification)

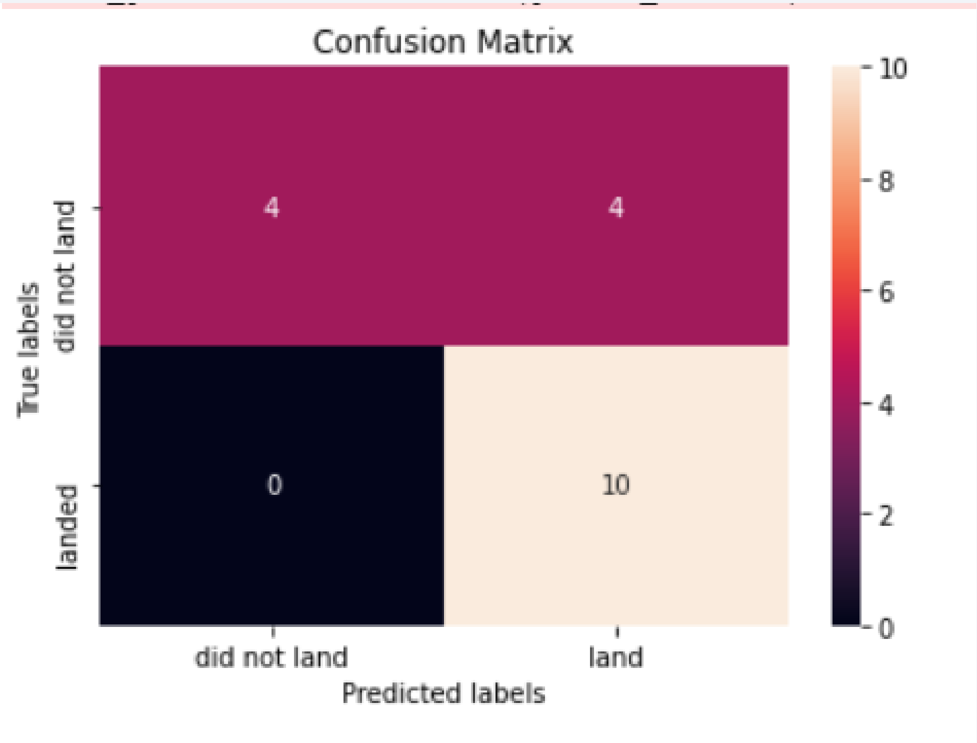
Classification Accuracy



- According to the bar chart, the highest accuracy is 78% from SVM model and K nearest model.

- Logistic regression model and decision tree have accuracy only 72%

Confusion Matrix



- Confusion matrix from K nearest model
- The model predict correctly at did not land cases
- The model still need to be improved because there are mistakes of predicting land cases

Conclusions

- The best model in this case are K nearest model and support vector machine
- The most success launch site is KSC LC-39A
- The best booster is FT version
- Orbit ES-L1, GRO, HEO, and SSO have 100% success rate

Appendix

- GitHub: https://github.com/ninesmit/Data-Science-first-project.git

Thank you!