IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Xiao Li>
<June 10, 2021>

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection: web scraping & API
  - Data analysis: EDA, data visualization, interactive visual analytics
  - Machine learning: log regression, SVM, decision tree, K nearest neighbor
- Summary of all results
  - Web scraping and API allows the data extraction from public resource
  - EDA can identify and plan success launching; visualization provides easy understanding results
  - Machine learning can provide the best suitable model to predict successfully launching with given data

Introduction

- Project background and context
  - This project is to help Space Y to determine if they can compete with Space X
  - All Space X data will be analyzed
- Problems you want to find answers
  - The best total cost for launches
  - The best location of lunches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Space X API
  - Web Scraping
- Perform data wrangling
  - Given data was enriched by creating landing outcome labels based on outcome data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was first normalized then split into train and test data following by different classification models. The accuracy of different models were discussed after

Data Collection

- Describe how data sets were collected.
  - SpaceX API:
  - Web Scraping:
- You need to present your data collection process use key phrases and flowcharts

- https://github.com/lixiaoaiden/Applied-data-science-capstone/blob/master/Data%20Collection%20.ipynb

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and used
- The API was used according to the flowchart beside and then data is persisted
- https://github.com/lixiaoaiden/Applied-data-science-capstone/blob/master/Data%20Collection%20.ipynb

Data Collection - Scraping

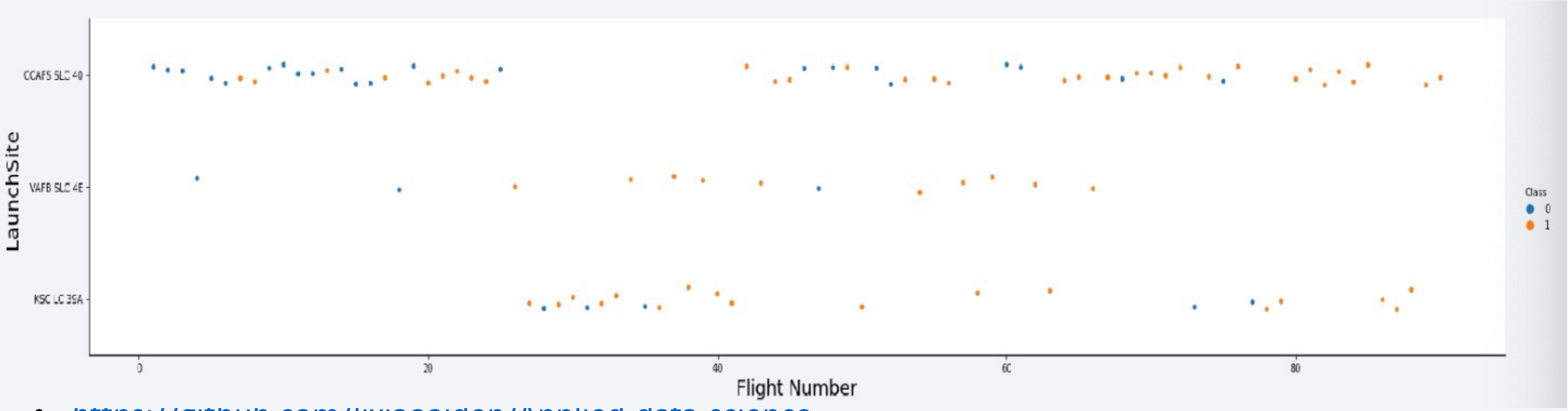
- Data from SpaceX launches can also obtained from Wikipedia
- https://github.com/lixiaoaiden/Applied-data-science-capstone/blob/master/Web%20scraping%20.ipynb

Data Wrangling

- Firstly apply some simple EDA to entire dataset
- Secondly summarize the launches per site, occurrences of each orbit and occurrence of mission outcome per orbit type
- Thirdly create outcome labels based on the outcome data
- https://github.com/lixiaoaiden/Applied-data-science-capstone/blob/master/Data%20wraggling.ipynb

EDA with Data Visualization

- Scatterplots and barplots were used to visualize the below features:
  - Payload Mass vs. Flight Number
  - Launch Site vs. Flight Number
  - Launch Site vs. Payload Mass
  - Orbit vs. Flight Number
  - Payload vs. Orbit



  - https://github.com/lixiaoaiden/Applied-data-science-capstone/blob/master/EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

- The following SQL queries were ran:
  - Names of the unique launch sites in the space missions
  - Top 5 launch sites whose name begin with "CCA"
  - Total payload mass carried by boosters launched by NASA(CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Total number of successful and failure mission outcomes
  - Name of the booster versions which have carried the max payload mass
  - Failed landing outcomes in drone ship with the booster versions, and launch sites name in year 2015
  - Rank of the count of landing outcomes between 2010-06-04 and 2017-03-20
- https://github.com/lixiaoaiden/Applied-data-science-capstone/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Fplium Maps
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates like NASA Johnson Space Center
  - Marker clusters indicate groups of events in each coordinate like launches in a launch site
  - Lines indicate distances between two coordinates
- https://github.com/lixiaoaiden/Applied-data-science-capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Percentage of launches by site and Payload range were used to visualize data in graphs and plots form
- This combination allows to quickly analyze the realtion between payloads and launch sites to identify where is the best place to launch according to the payload
- https://github.com/lixiaoaiden/Applied-data-science-capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Predictive Analysis (Classification)

- The following classification models were compared
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K-nearest neighbors

Data prep and standardization → Test of each model → Compare the results and accuracy

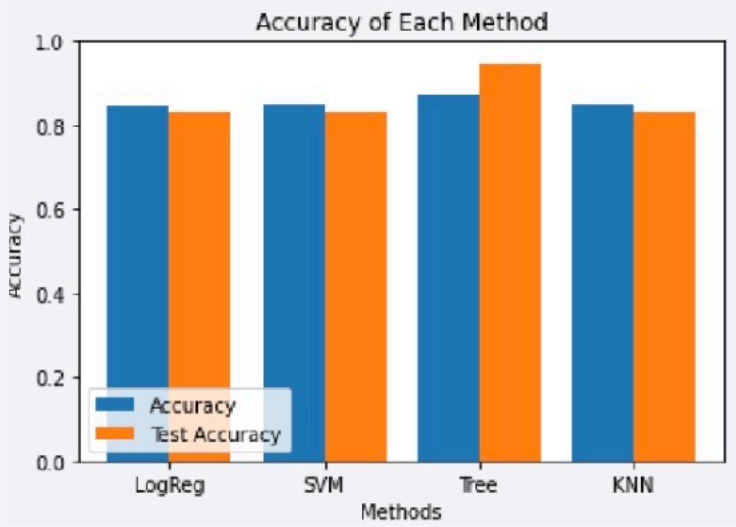
- https://github.com/lixiaoaiden/Applied-data-science-capstone/blob/master/Machine%20Learning%20Prediction%20lab.ipynb

Results

- Space X uses 4 different launch sites;
- The first launches were completed by Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2928kg;
- The first success landing outcome happened in 2015;
- Many F9 booster versions were successful at landing drone ships having payload above average;
- Almost 100% of mission outcomes were successful
- Two booster versions failed at landing in drone ships in 2015 are
  - F9 v1.1 B1012
  - F9 v1.1 B1015
- The number of landing outcomes were getting better as years passed;
- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around ;
- Most launches happens at east coast launch sites;

Results

- Predictive Analysis indicates that Decision Tree Classifier is the best model to predict successful landings which have the accuracy over 87% and accuracy for test data is over than 84%

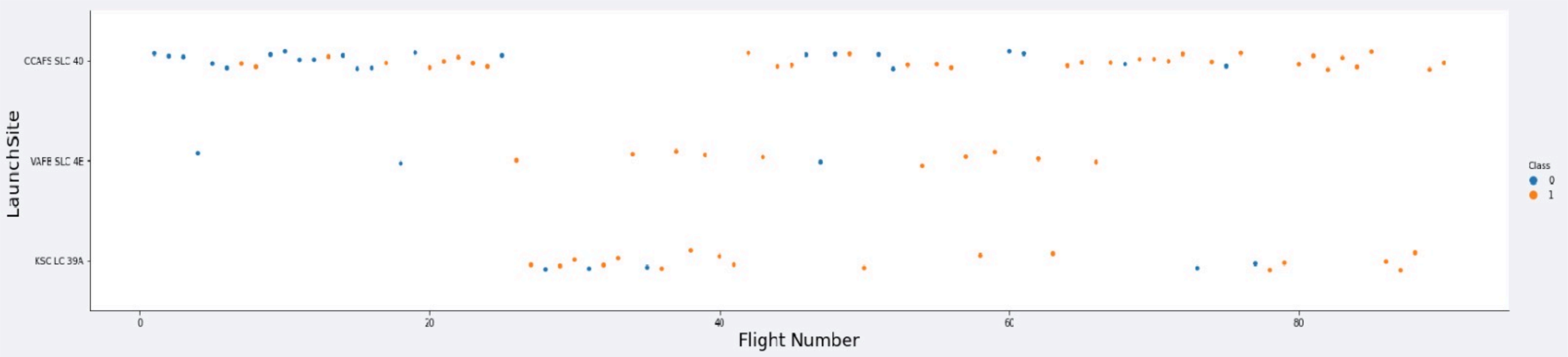


Section 2

Insights drawn from EDA

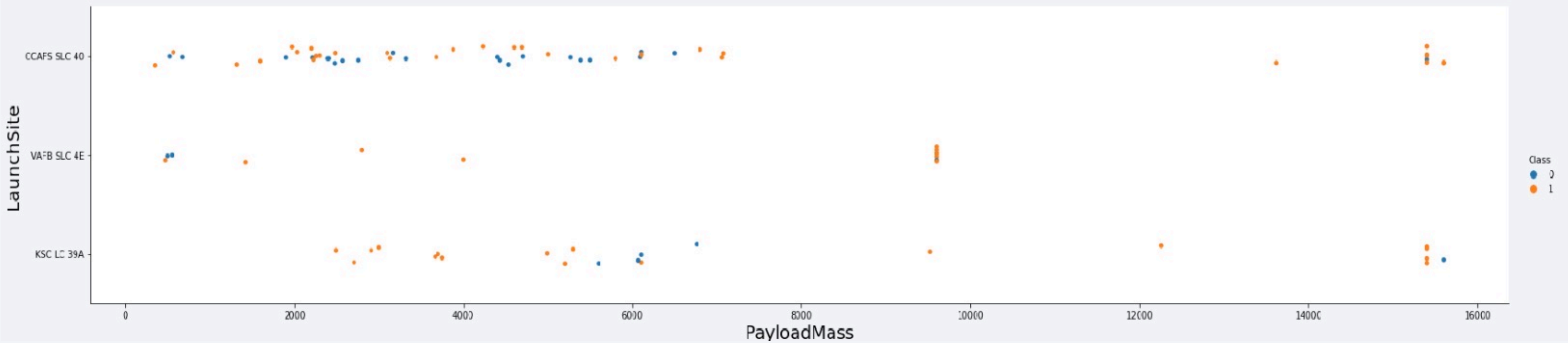
Flight Number vs. Launch Site

- The best launch site nowadays is CCAF5 SLC 40 where most of recent launches were successful;
- VAFB SLC 4E and KSC LC 39A were 2nd and 3rd best launch sites;
- The overall success rate was improved over the time



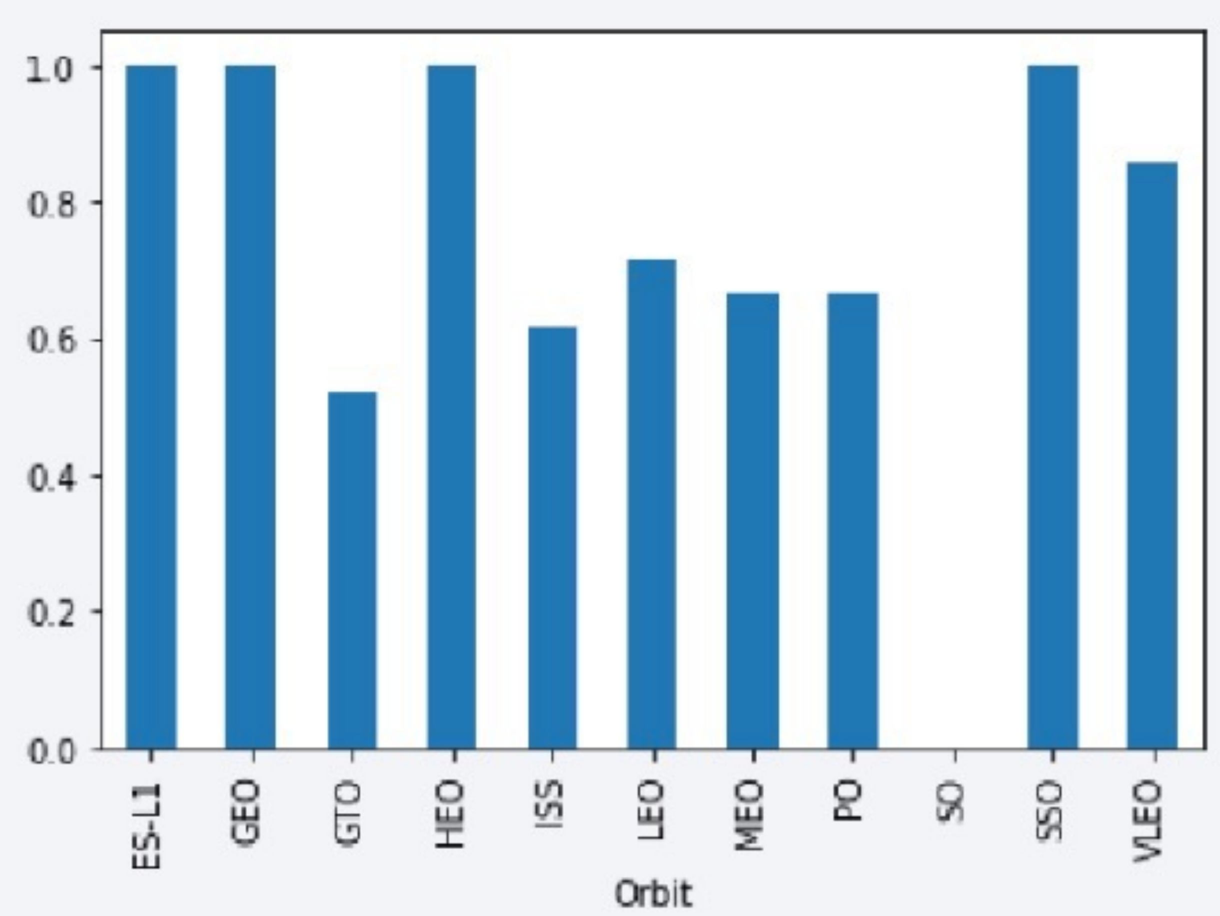
Payload vs. Launch Site

- Payloads over 9000kg have an excellent success rate;
- Payloads over 12000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites



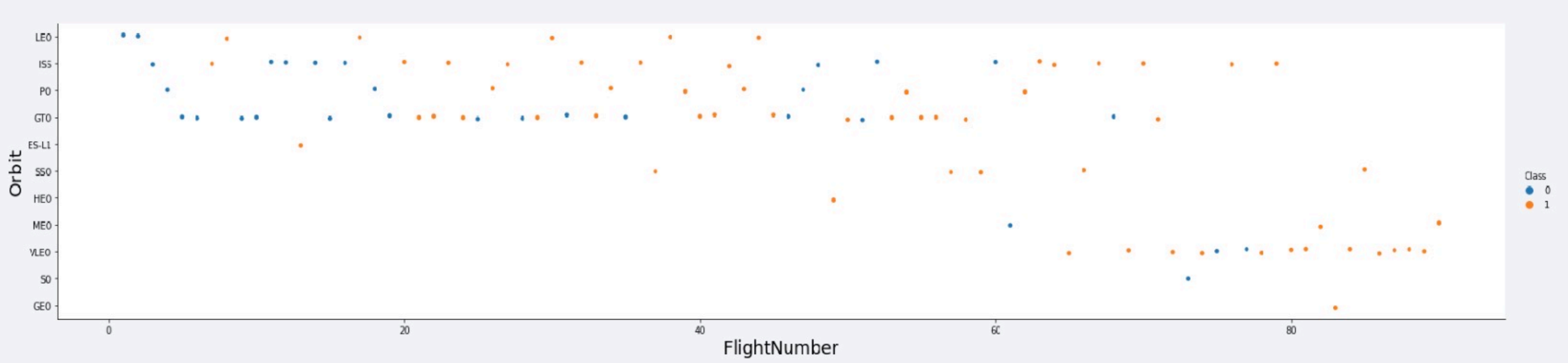
Success Rate vs. Orbit Type

- The highest success rates occurred at orbits
  - ES-L1;
  - GEO
  - HEO
  - SSO



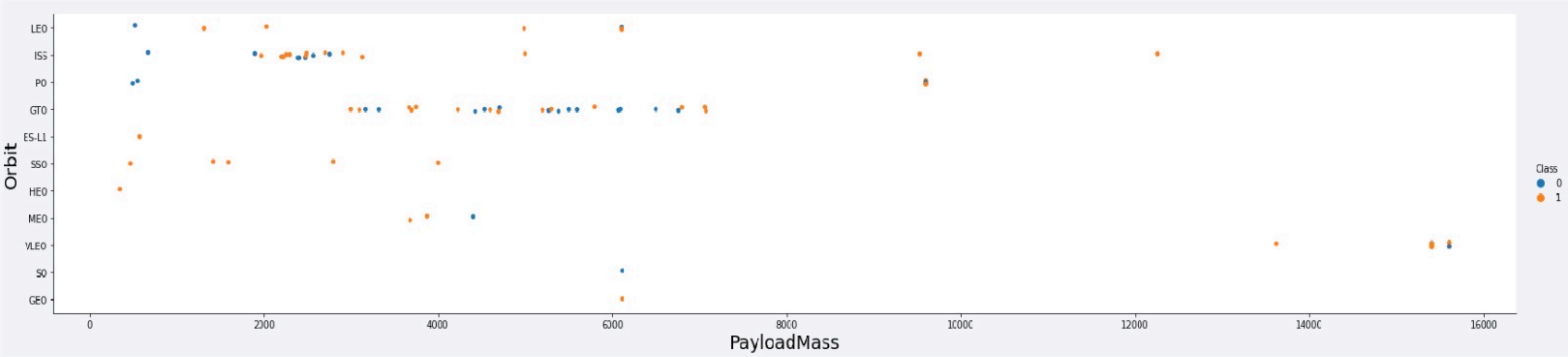
Flight Number vs. Orbit Type

- VELO orbit seems to be a new orbit as it's been used more recently
- Overall, all the orbits have improved in success rate over the time



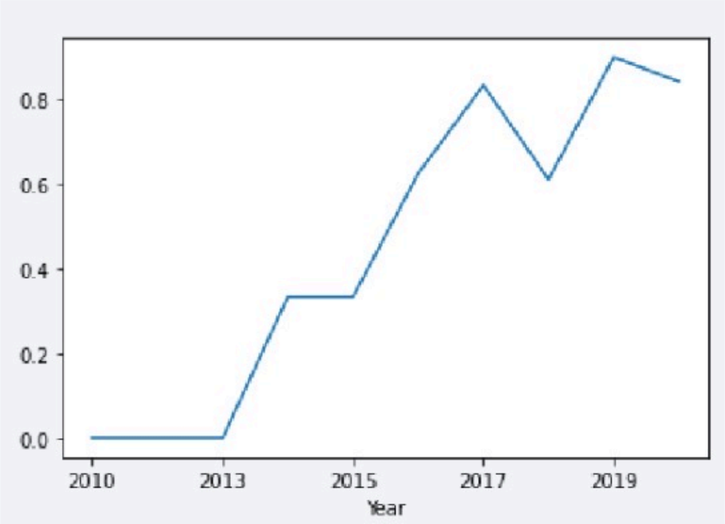
Payload vs. Orbit Type

- There is no relation between payload and success rate at orbit GTO
- ISS orbit has the widest range of payload with a good success rate
- There are few launches at orbits SO and GEO



Launch Success Yearly Trend

- Success rate started to increase in 2013 and till 2020
- It seems that the first three years were a period of adjusts and improvements of technology



All Launch Site Names

- There are four launch sites

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- Select unique occurrence of "launch_sites" values from the data

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites 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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Total Payload (kg)
111.268

- Total payload calculated by summing all payloads whose codes contain”CRS” which corresponds to NASA

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

Avg Payload (kg)
2.928

- Filtering data by the booster version above and calculating the average payload mass to be above vaule

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

Min Date
2015-12-22

- Filtering data by successful landing outcome on ground pad and getting the min value as of date. This will be your first successful date

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the 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 B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- Select the booster version as above and you will see four results

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- Grouping mission outcomes and counting the number for each group, you will see the result as above

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

Booster Version (...)	Booster Version
F9 B5 B1048.4	F9 B5 B1051.4
F9 B5 B1048.5	F9 B5 B1051.6
F9 B5 B1049.4	F9 B5 B1056.4
F9 B5 B1049.5	F9 B5 B1058.3
F9 B5 B1049.7	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1060.3

- Select the max payload mass for the booster versions, you will see result as above

2015 Launch Records

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

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- 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

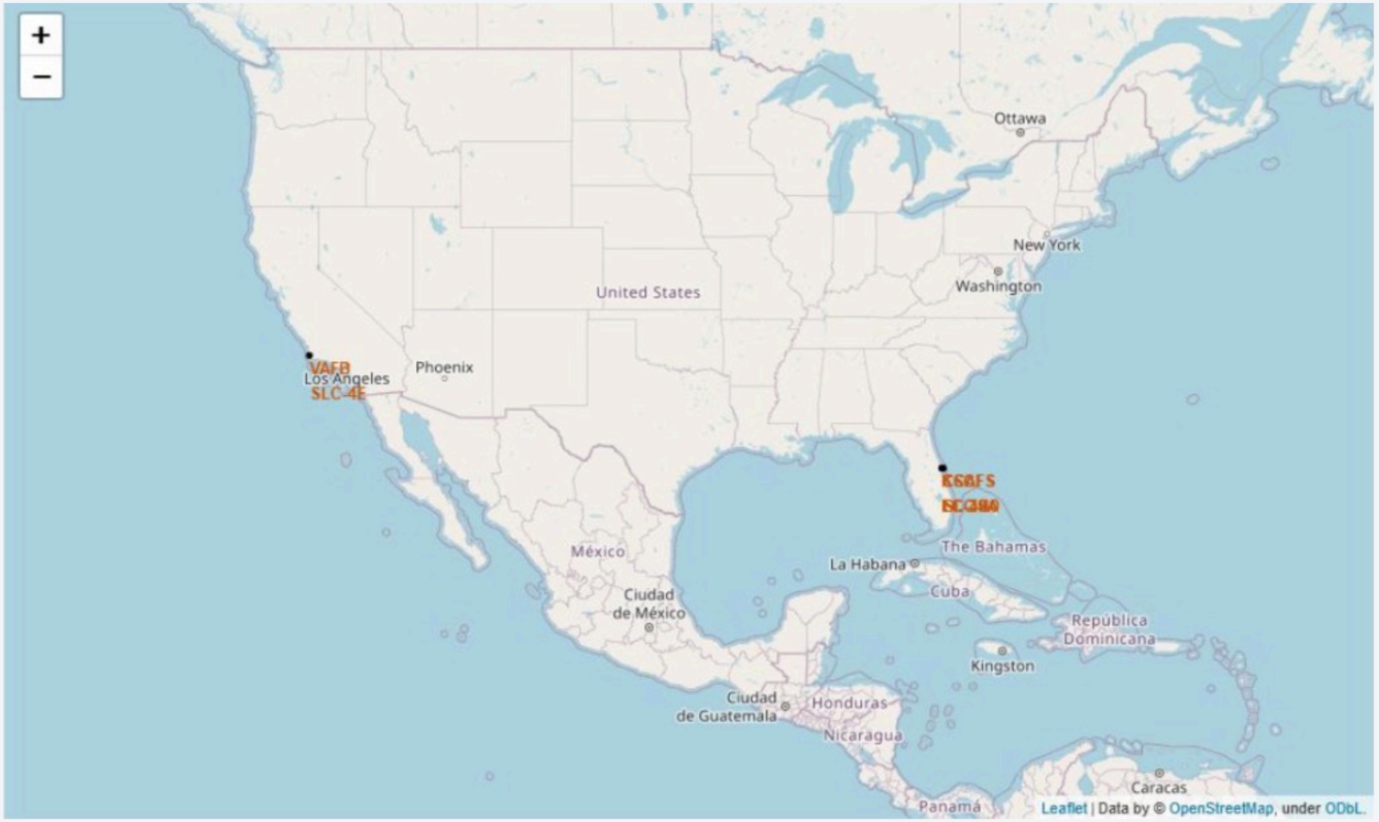
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

All Launch Sites

- All launch sites seem to be coastal area and also near to main roads.

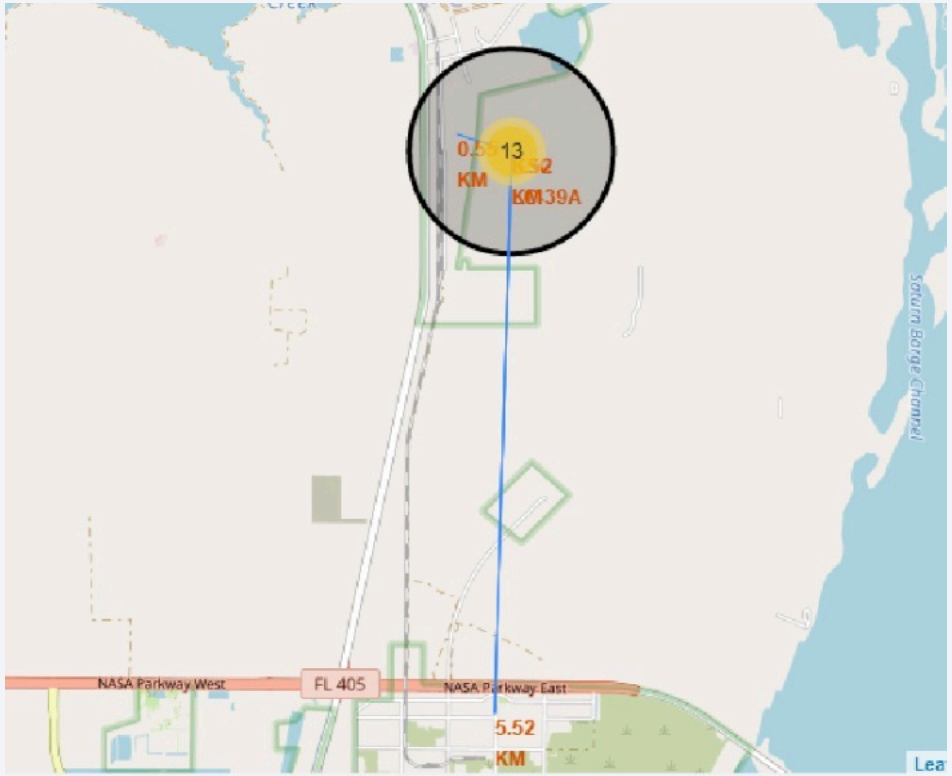


Launch Outcomes by Site

- Green markers indicate successful launch where read means failed one

Logistic and Safety

- Launch site KSC LC-39A has good logistics concerns

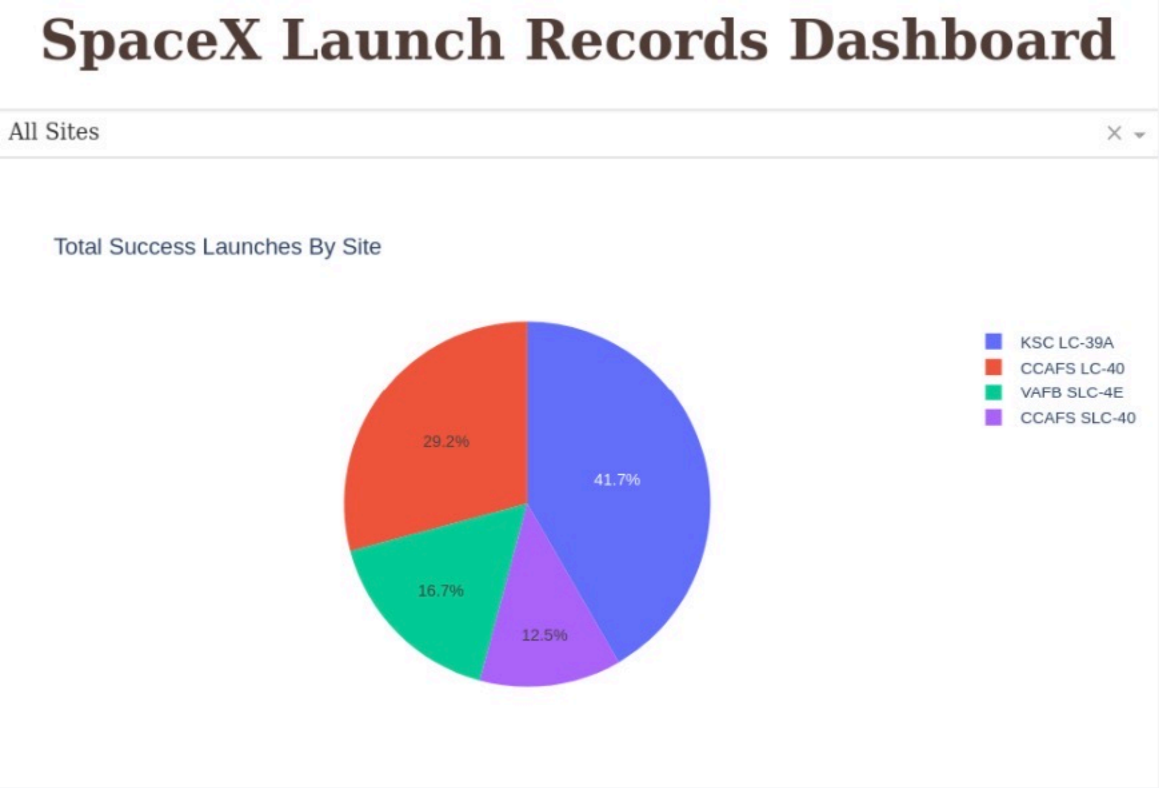


Section 4

Build a Dashboard with Plotly Dash

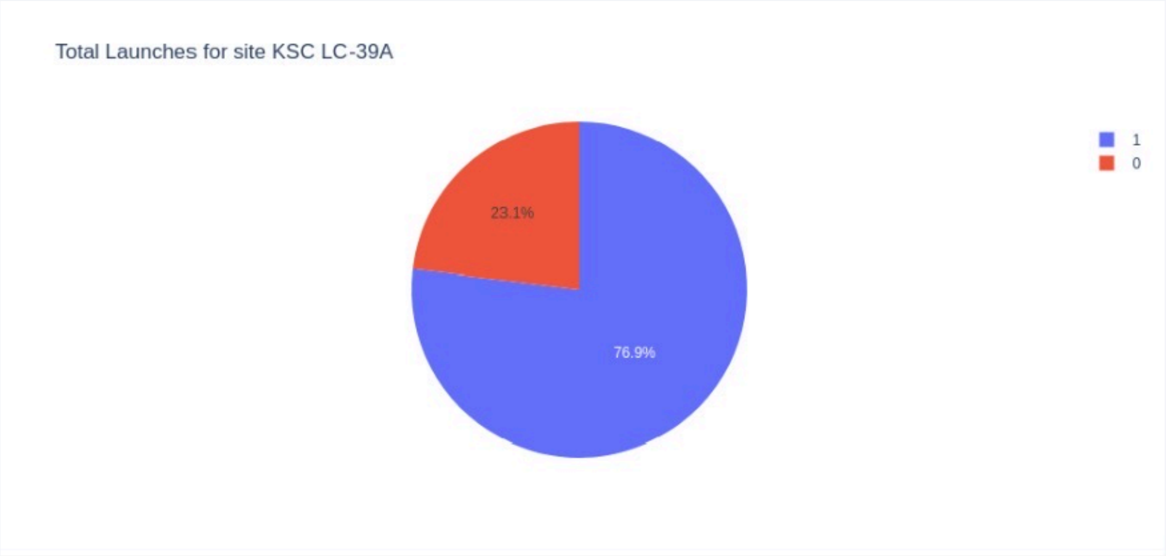
Successful Launches by Sites

- Launch location is a crucial factor to be considered for successful launches



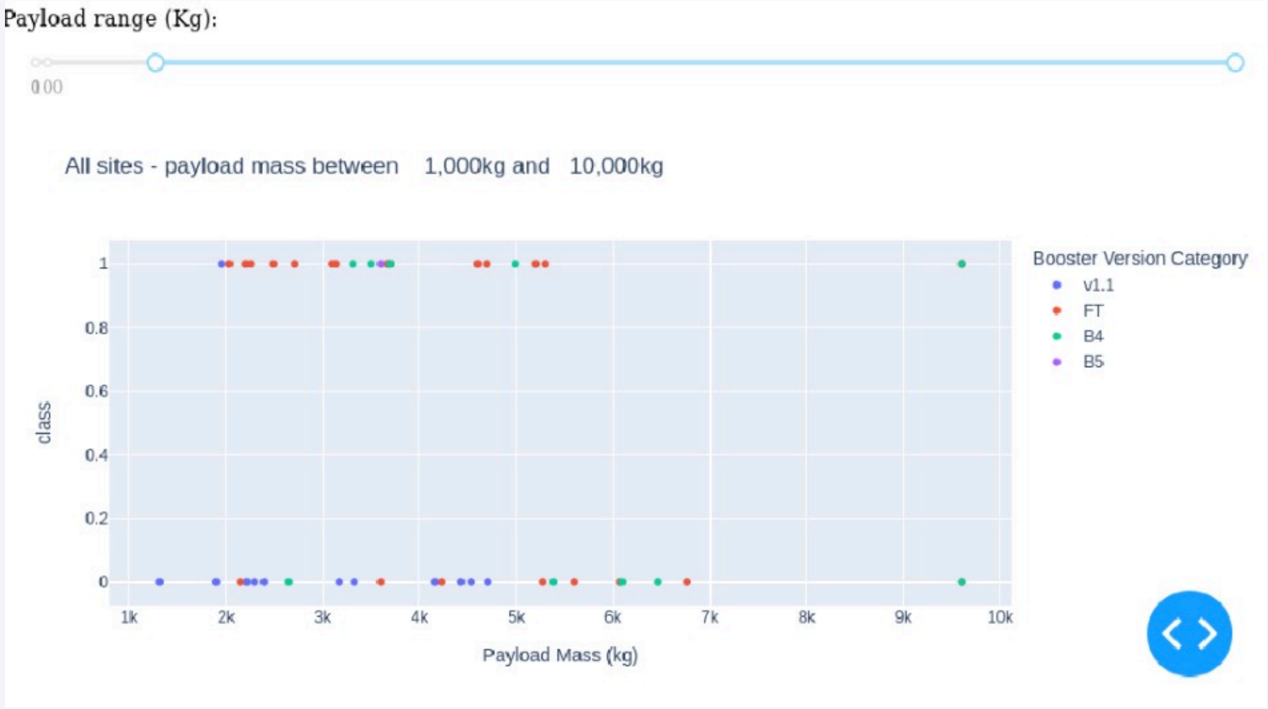
Launch Success Rate for KSC LC-39A

- This is a typical pie chart for launch location KSC LC-39A
- 1 means successful 0 means failed



Payload Range

- Payloads under 6000lg with FT booster are the most successful combination

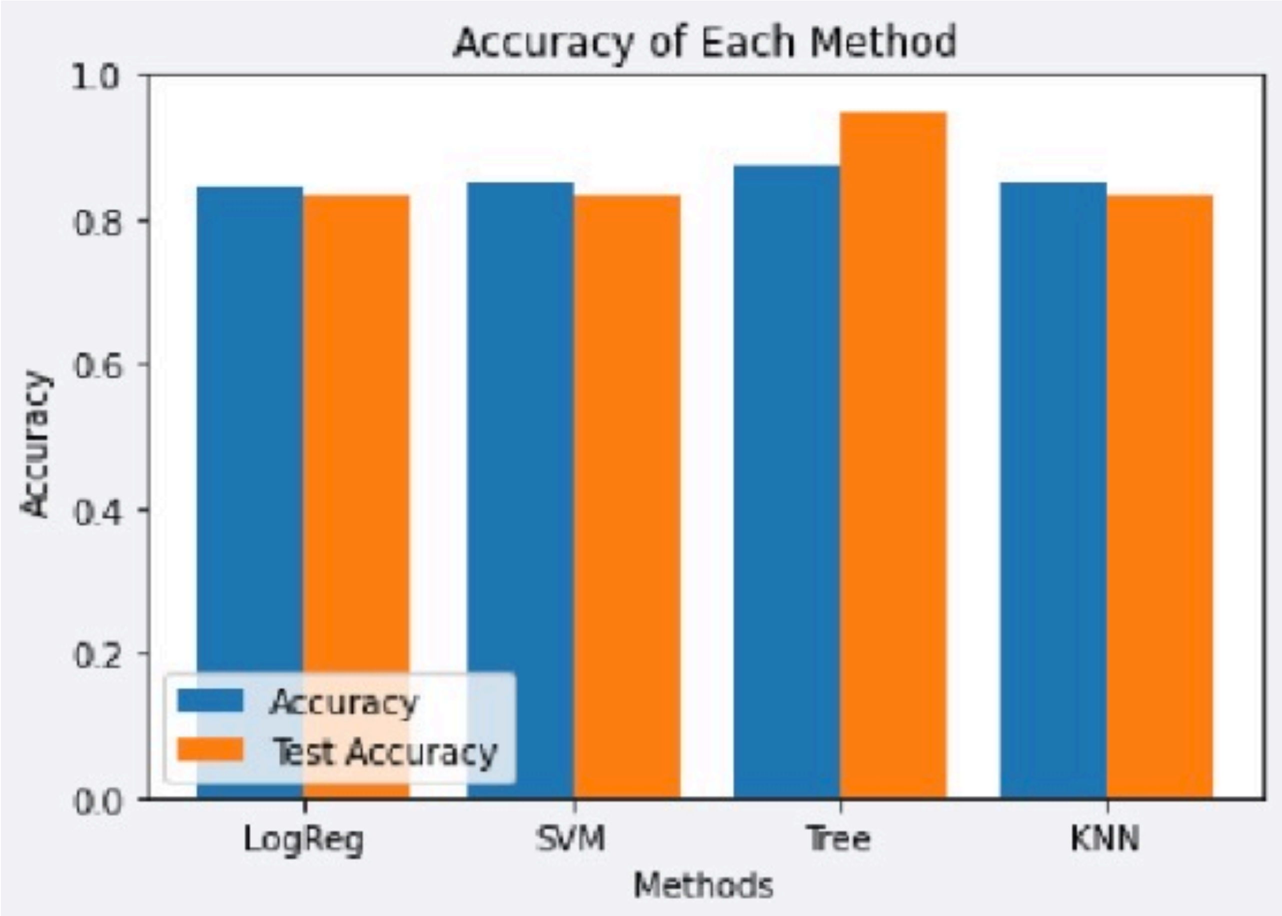


Section 5

Predictive Analysis (Classification)

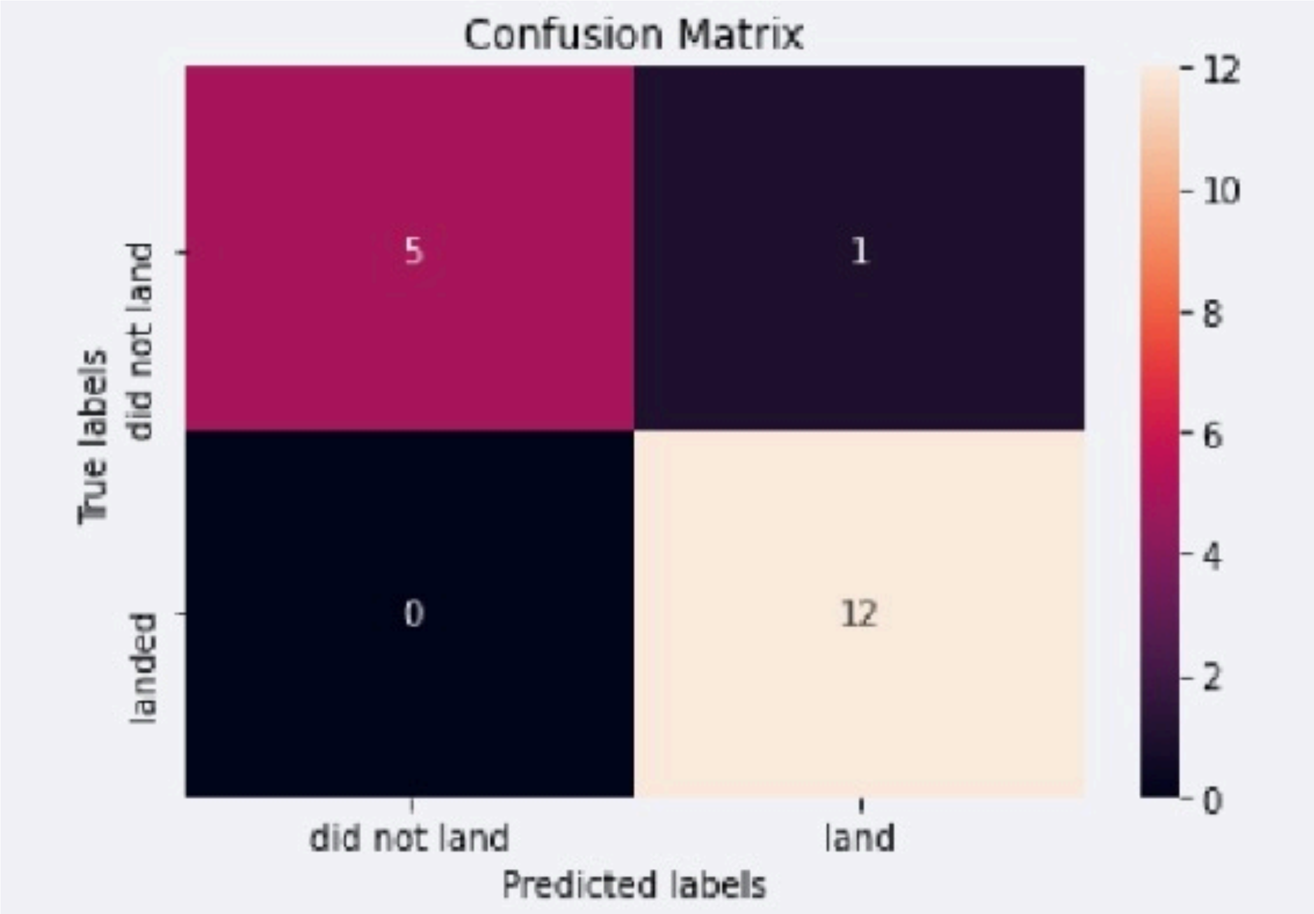
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- Decision tree has the highest accuracy which is more than 87%.



Confusion Matrix

- Decision tree confusion matrix proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones



Conclusions

- Best launch site is KSC LC-39A
- Launches with payload over than 7000kg could be dangerous
- Overall, successful landing outcomes seems to be improved over the time
- Decision tree classifier is the best model to predict successful landing outcomes

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!