IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Haftom Tsegay
March - 2022

Outline

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

Executive Summary

- The sources for the data analysis are collected using web scrapping and API techniques
- Different data wrangling techniques are applied to clean and standardize the data and falcon9 data is selected for further analysis.
- Visualizations like scatter plot ,bar chart and maps are used to analyze the relation ship between the dependent and independent variables and select the necessary features
- The identified features are converted to numeric using bag of words
- For predictive analysis models are KNN and Logistic regression models are selected according to their accuracies results.
- KSC LC-39A launch site is found with the highest success rate of landing with the range of payload mass 1000Kg to 5000Kg.
- Interactive dashboard is built to visualize the relation ship of each launch sites with payload mass and success rate.

Introduction

- Space Exploration Technologies Corp. (doing business as SpaceX) is an American aerospace manufacturer, a provider of space transportation services, and a communications corporation headquartered in Hawthorne, California
- 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.
- Falcon 9 is a partially reusable two-stage-to-orbit medium-lift launch vehicle designed and manufactured by SpaceX in the United States
- This data science capstone project is done predict if the Falcon 9 first stage will land successfully

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web scraping and parsing libraries like BeautifulSoup are used to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches.
- Perform data wrangling
  - Identifying missing data ,counting the landing outcomes to rewrite it as classes of 0 and 1 and success rate of Falcon 9 is calculated.
- Perform exploratory data analysis (EDA) using visualization and SQL

..Cont

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - correlated features selected from independent variables
  - data is standardized /scaled
  - train and test data split is done
  - best Hyperparameter for SVM, Classification Trees and Logistic Regression calculated by tuning parameters

Data Collection

- Data is collected using two processes

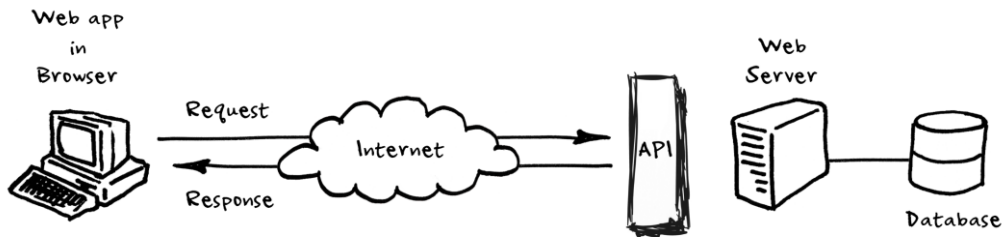
1. Web Scraping refers to the process of extracting data from a website or specific webpage

- Using request data is scrapped from "List of Falcon 9 and Falcon Heavy launches Wiki page updated on 9th June 2021" website

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

2. An API is typically defined as a set of specifications, such as Hypertext Transfer Protocol (HTTP) request messages, along with a definition of the structure of response messages, usually in an Extensible Markup Language (XML) or JavaScript Object Notation (JSON) format

- Data is requested from SpaceX API url = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN SkillsNetwork/datasets/API_call_spacex_api.json'

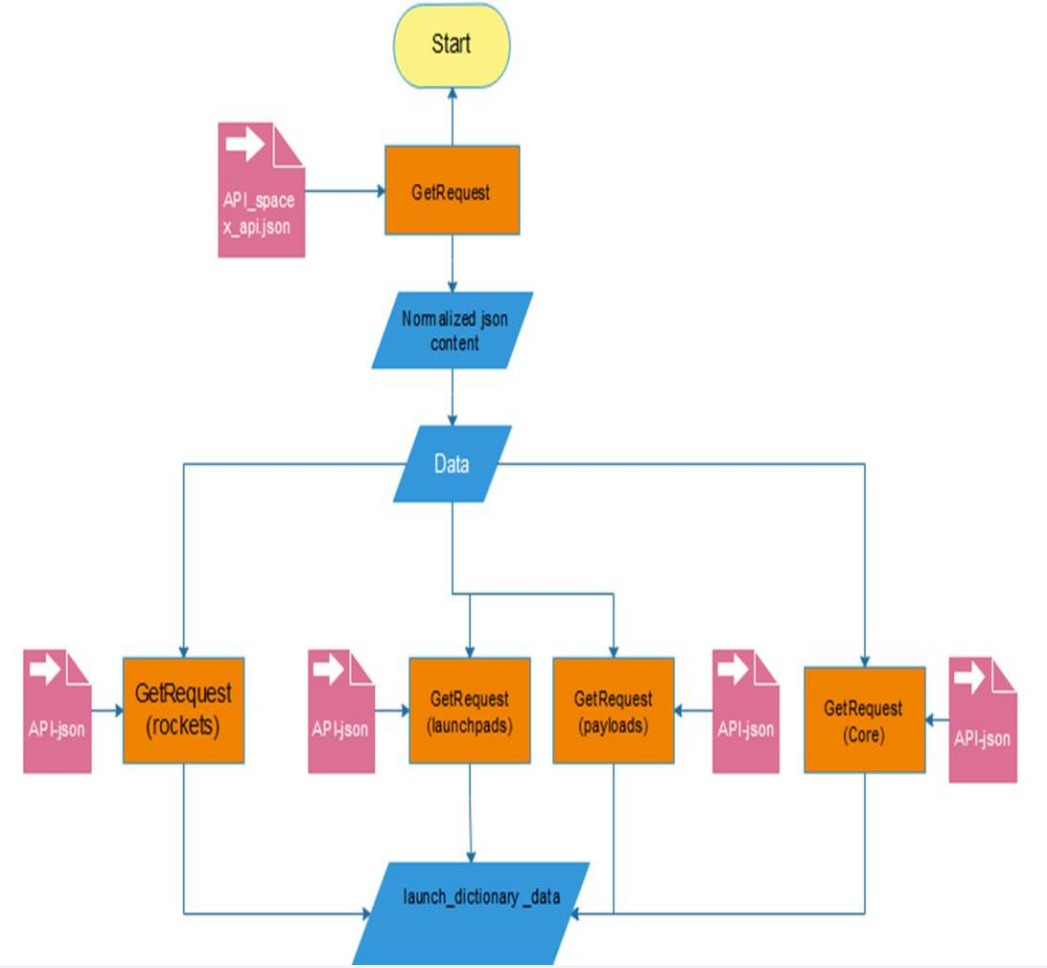


Data Collection – SpaceX API

- Request to the SpaceX API is send to get using the following commands
- response = requests.get(https://api.spacexdata.com/v4/launchpads/...).json()
- response = requests.get("https://api.spacexdata.com/v4/rockets/...).json()
- response = requests.get("https://api.spacexdata.com/v4/launchpads/...).json()
- response = requests.get("https://api.spacexdata.com/v4/payloads/"+load).json()
- response = requests.get("https://api.spacexdata.com/v4/cores/"+core['core']).json()
- response_js = requests.get(…..)

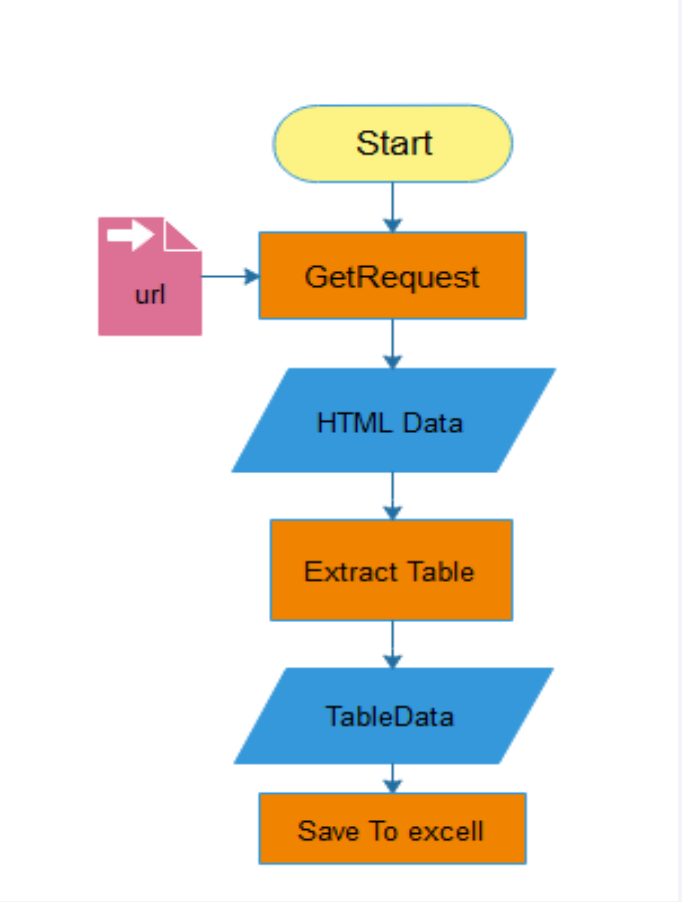
Accordingly

- BoosterVersion ,getLaunchSite, PayloadData and CoreData data are extracted using the given API
- Git Source : Applied-Data-Science-Capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · Haftom-sig/Applied-Data-Science-Capstone (github.com)



Data Collection - Scraping

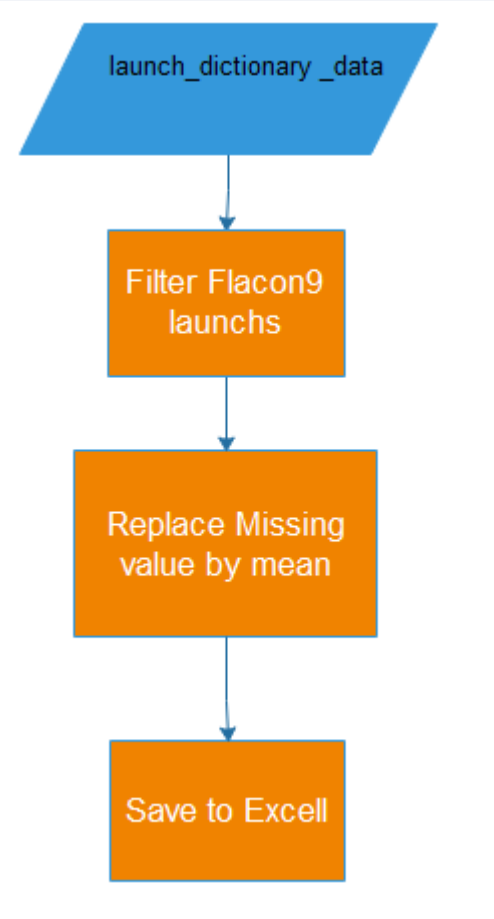
- Web scrapping is done on Falcon 9 launch records with BeautifulSoup python library
- Falcon 9 launch records HTML table from Wikipedia are extracted
- Tables from the extracted data searched and parsed into a Pandas data frame
- Git Source : [Applied-Data-Science-Capstone/jupyter-labs-webscraping.ipynb at main · Haftom-sig/Applied-Data-Science-Capstone (github.com)]



Data Wrangling

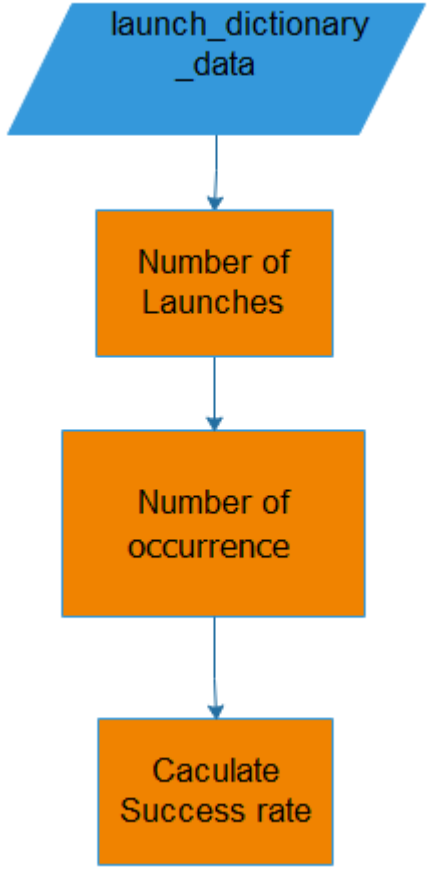
- Filter the data data frame using the BoosterVersion column to only keep the Falcon 9 launches
- Missing values in the dataset rows are identified (PayloadMass = 5 and LandingPad = 26) using *data_falcon9.isnull().sum()*
- For payload mass mean values are calculated over the entire payloadmass column and each missing are replaced by mean
- *data_falcon9_PM_mean = data_falcon9['PayloadMass'].mean()*
- *data_falcon9['PayloadMass']=data_falcon9['PayloadMass'].fillna(data_falcon9_PM_mean)*

- Git Source : Applied-Data-Science-Capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · Haftom-sig/Applied-Data-Science-Capstone (github.com)



…cont

- Number of launches on each site and
- the number occurrence of each orbit
- number and occurrence of mission outcome per orbit type are calculated using value_count() method
- A landing outcome label is created by assigning to new column landing_clas
- This new variable is used to calculate the success rate
- Git Source : [Applied-Data-Science-Capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · Haftom-sig/Applied-Data-Science-Capstone (github.com)]



EDA with Data Visualization

- To visualize the relation ship between the following parameters scatter plot are barchart are used ,because such plots easily help to visualize between parameters

Scatter Plot	Bar Chart
FlightNumber vs. PayloadMass	success rate vs orbit type
FlightNumber vs LaunchSite	Success vs year
Payload and Launch Site	
FlightNumber and Orbit type	
Payload and Orbit type	

The following are Gitsources

- Applied-Data-Science-Capstone/jupyter-labs-eda-dataviz.ipynb at main · Haftom-sig/Applied-Data-Science-Capstone (github.com)

EDA with SQL

- **The following are used to extract the SQL queries**
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass.
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 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
- The Sql quries statement are done and uploaded at the following github Applied-Data-Science-Capstone/jupyter-labs-eda-sql-coursera.ipynb at main · Haftom-sig/Applied-Data-Science-Capstone (github.com)

Build an Interactive Map with Folium

- The following objects with their importance are listed below

Object	Why are the objects used?
Circle	To clearly visualize the areas centered at the coordinates
Marker	To put icon and specify the areas name
Lines	To calculate the distance between the launch sites and its proximities like railway, city ,coastline etc

- Githubsouce : Applied-Data-Science-Capstone/lab_jupyter_launch_site_location.ipynb at main · Haftom-sig/Applied-Data-Science-Capstone (github.com)

Build a Dashboard with Plotly Dash

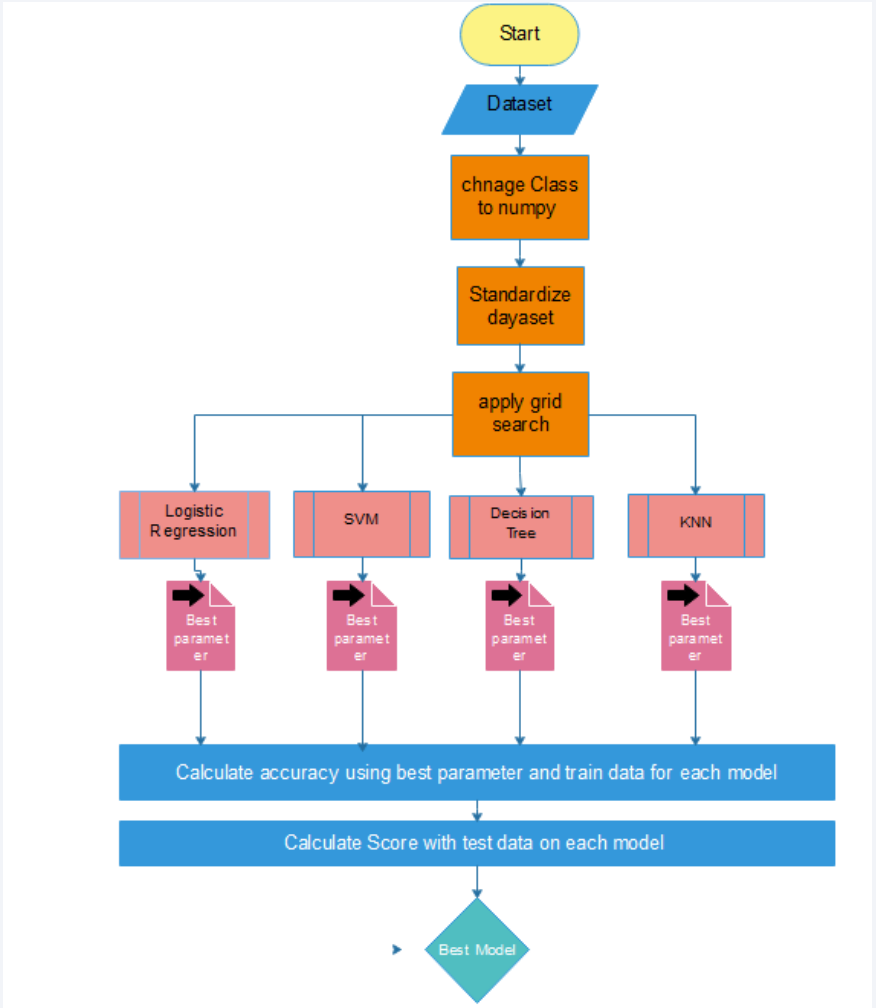
- Pichart is used to see the success rate of each launch sites because it the easiest to show the information.
- And scatter plot is applied in the dashboard to visualize the relation ship between payload mass and succuss for each sites
- Dropdown can easily help user to interact with the dashboard to sestet the launch sites as in put.
- To analyze the effect of payload mass Range Slider is used to see the success rate of each site along with different payload masses
- Github source Applied-Data-Science-Capstone/Applied_Ploty Dash.ipynb at main · Haftom-sig/Applied-Data-Science-Capstone (github.com)

Predictive Analysis (Classification)

- The data set has both features and target variables.
- After changing the target variable to numpy array the entire dataset is standardize to bring all the features data to the same scale
- Four models are selected to predict the success class

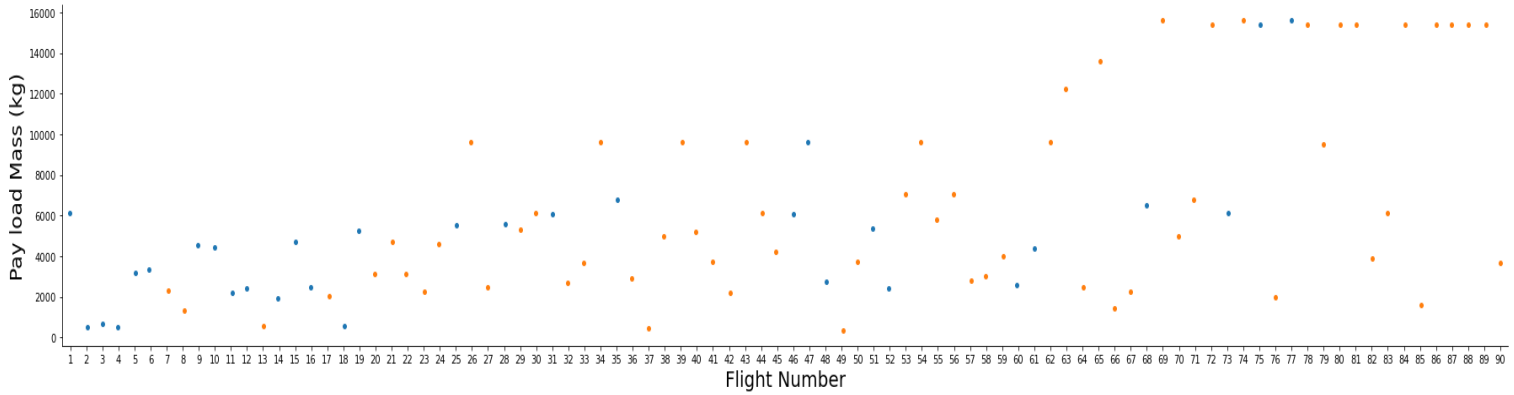
1. logistic regression
2. Support Vector motion
3. Decision tree
4. KNN

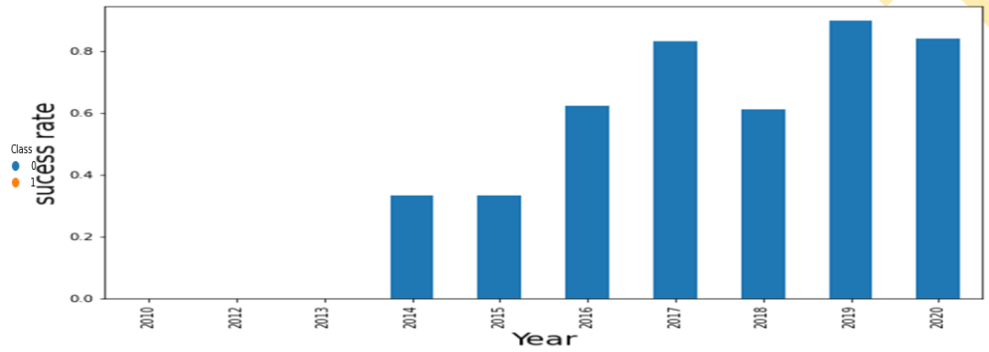
- Grid search is used to calculate the best parameters for each model
- By comparing the train test accuracies of each model best model for the data set is identified.
- During the training time KNN and logistic regression has best performance
- Github source : [Applied-Data-Science-Capstone/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · Haftom-sig/Applied-Data-Science-Capstone (github.com)]

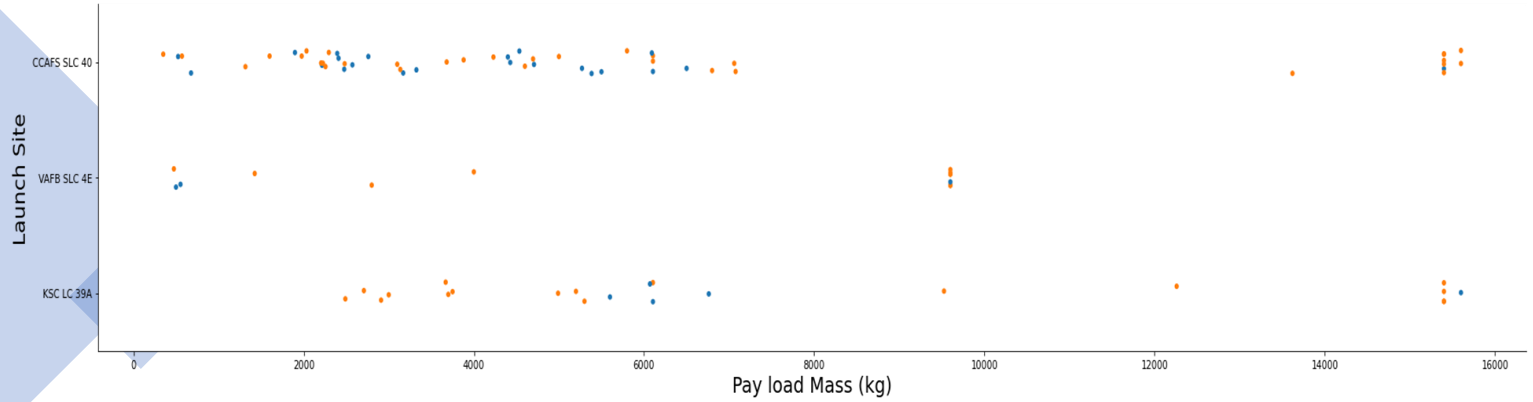


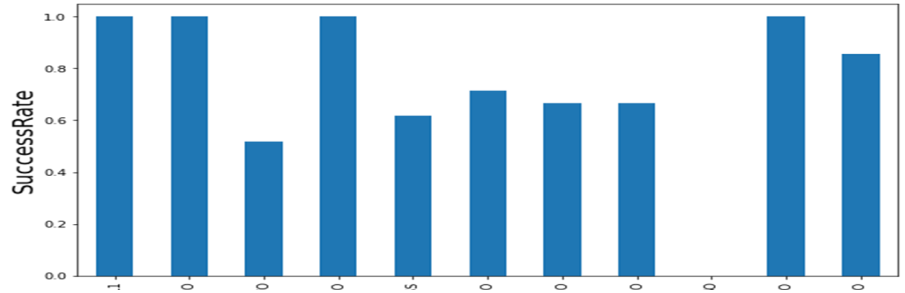
Results

- some exploratory data analysis results

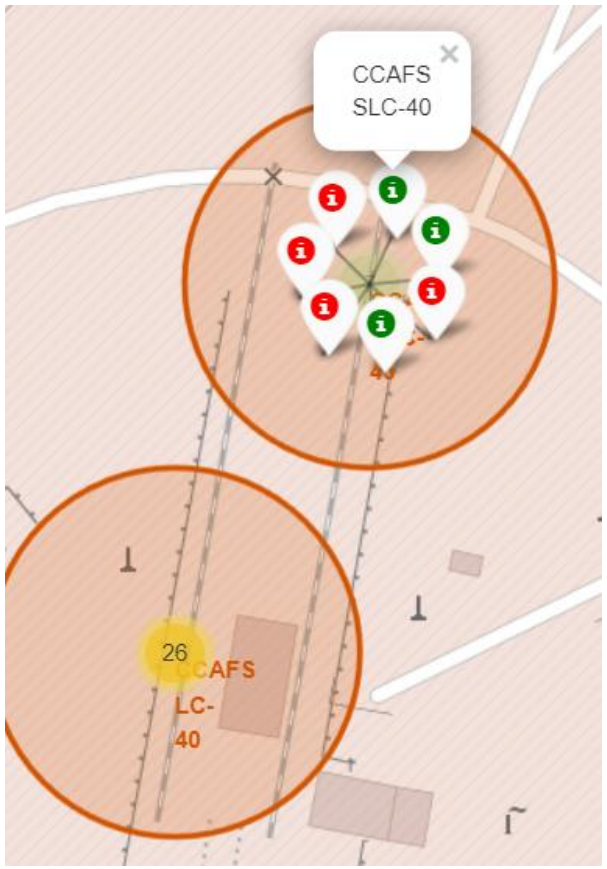


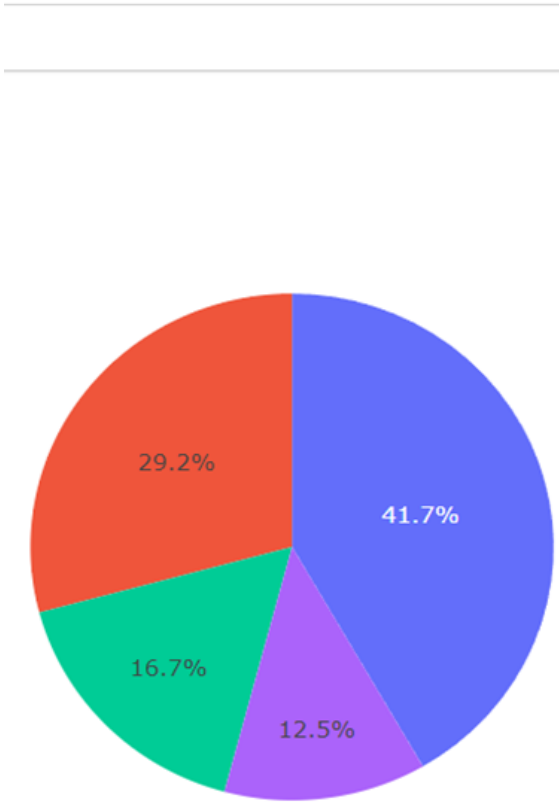


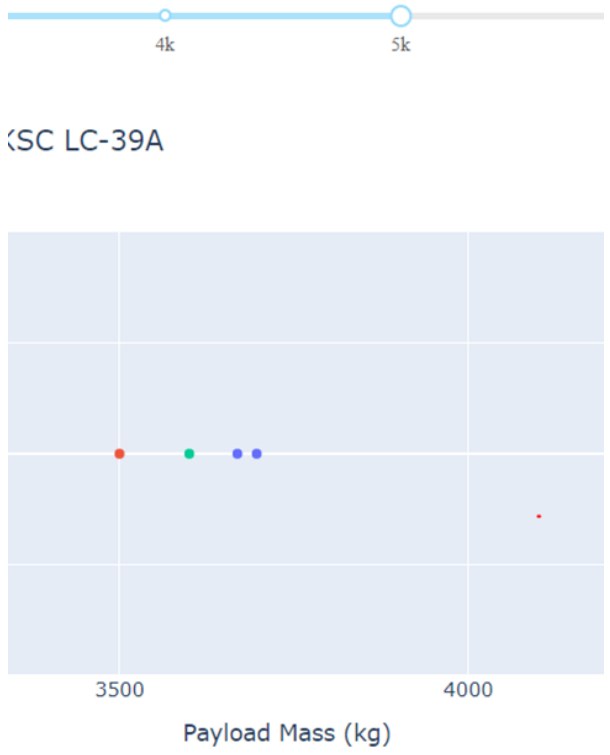




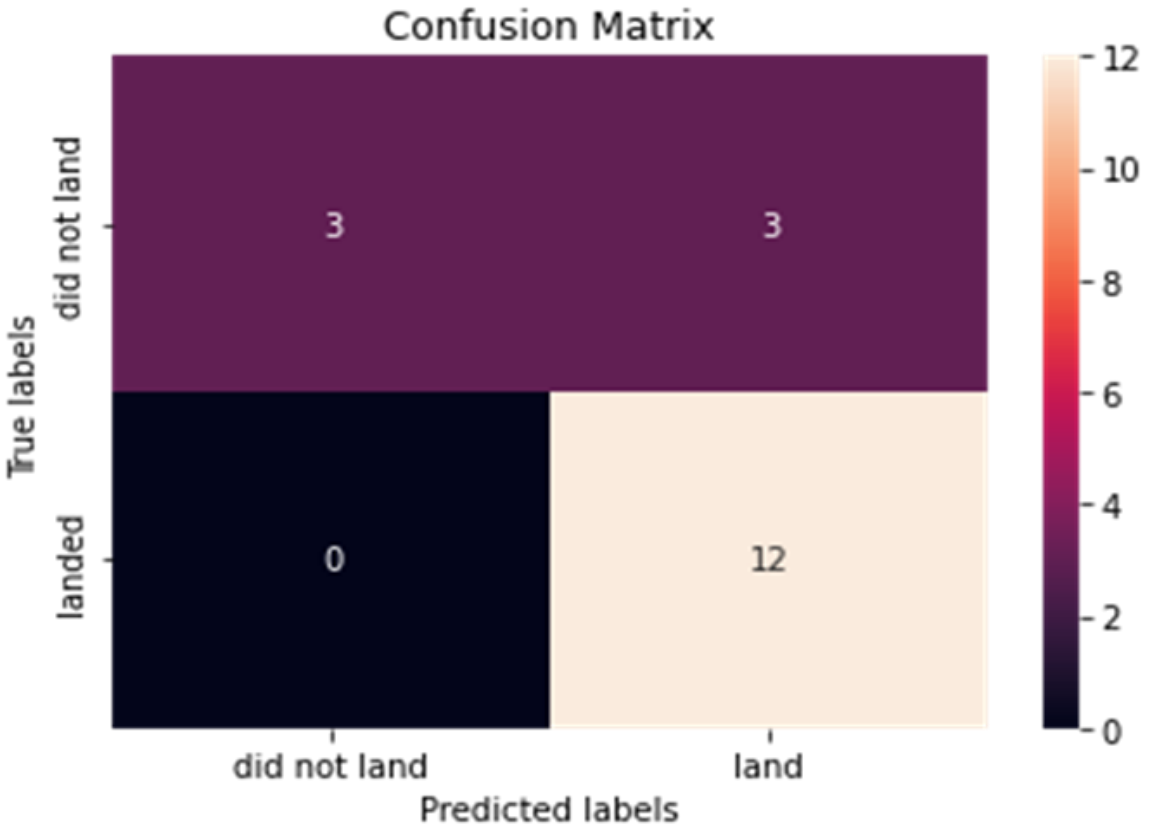
Interactive analytics demo in screenshots

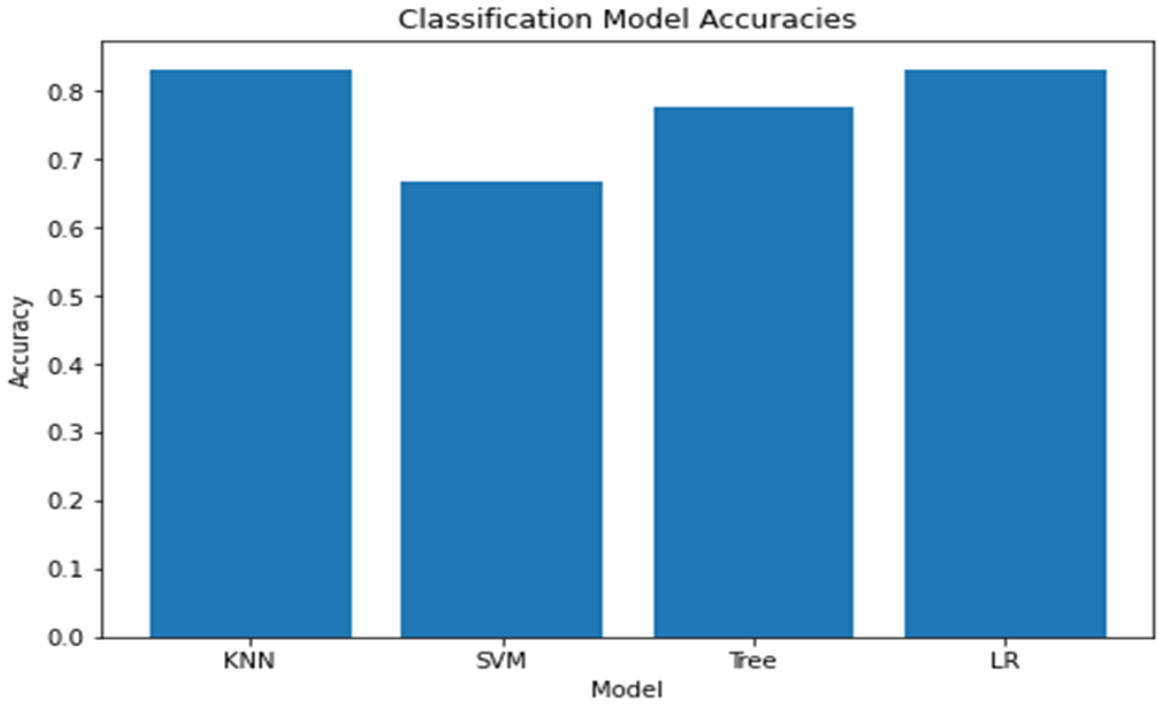


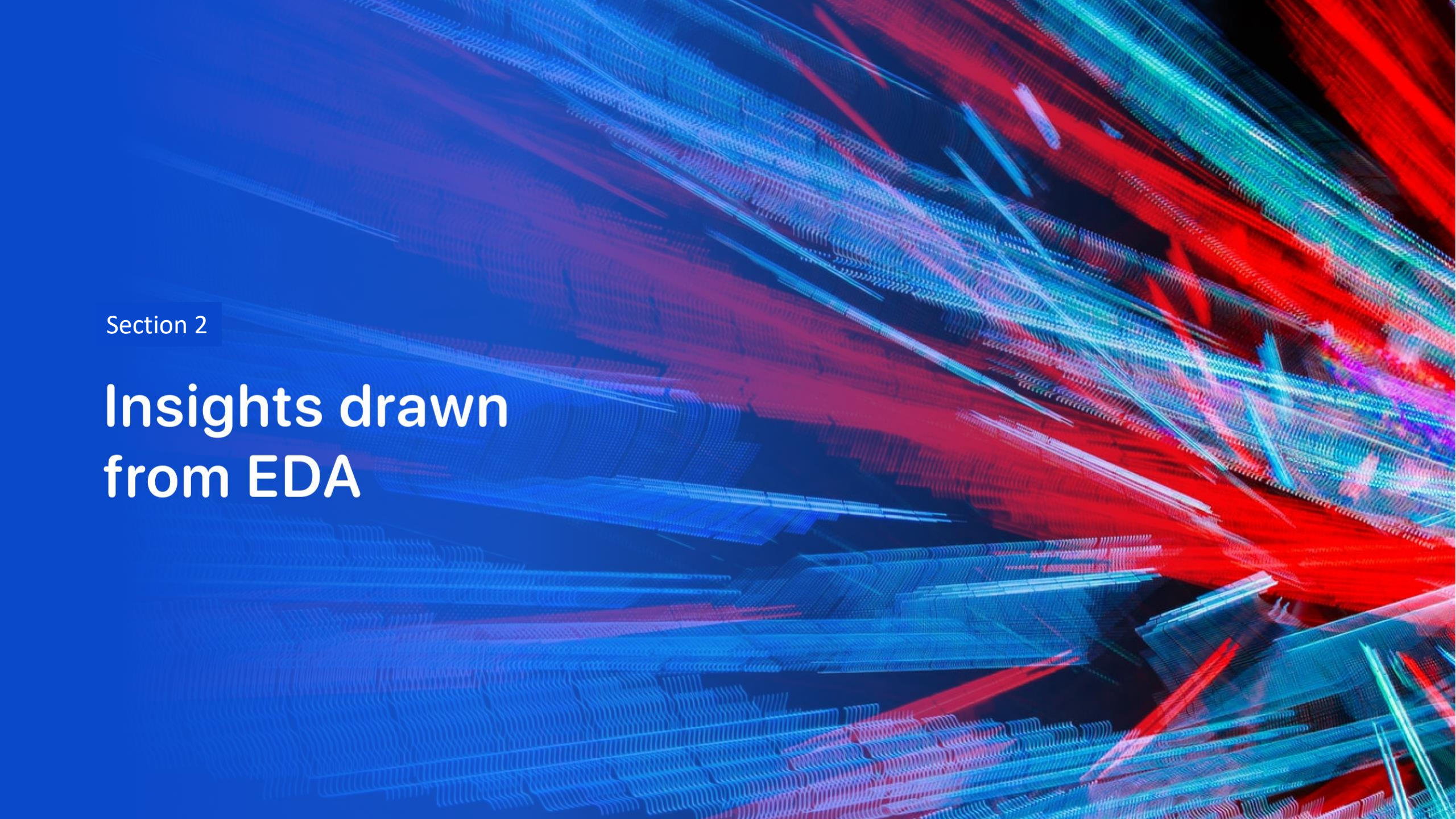




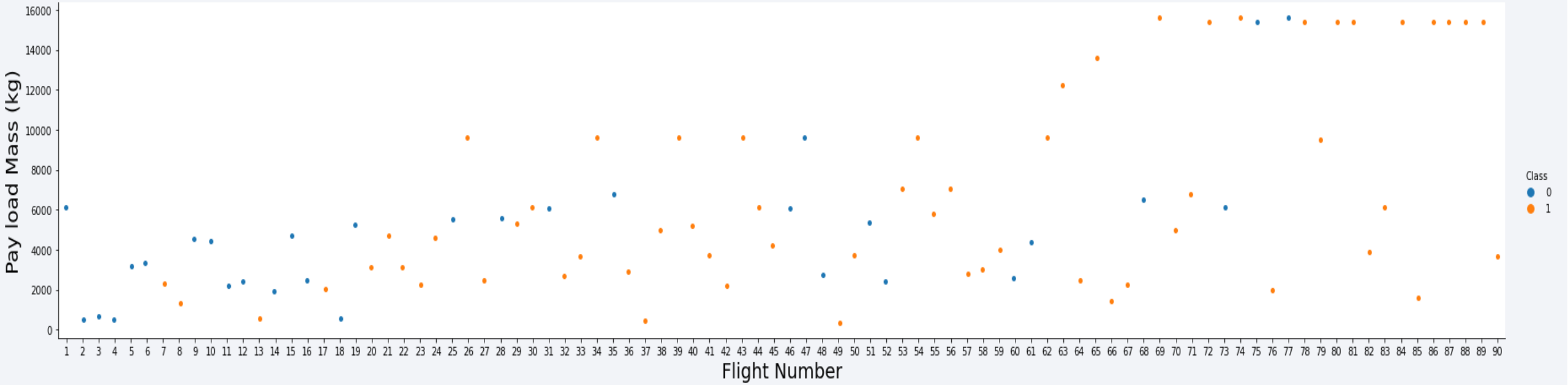
Predictive analysis results





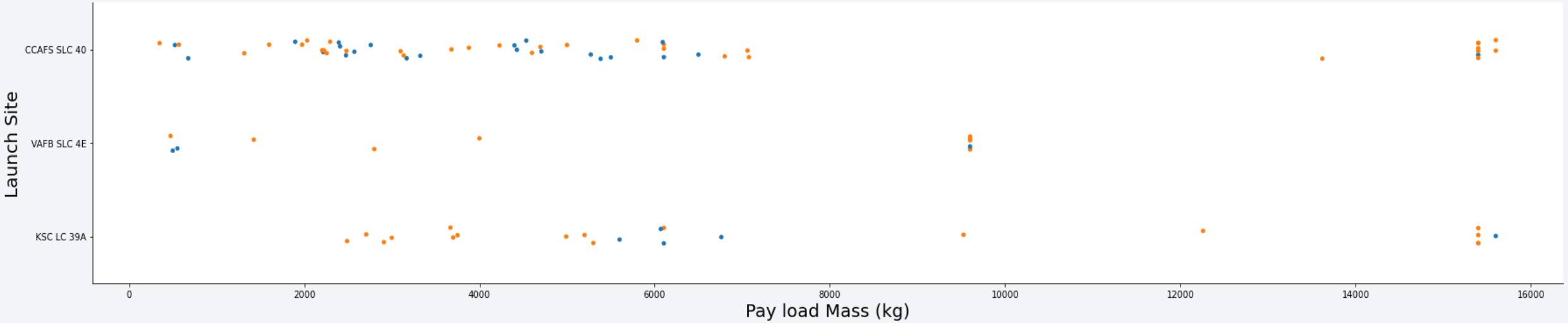

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



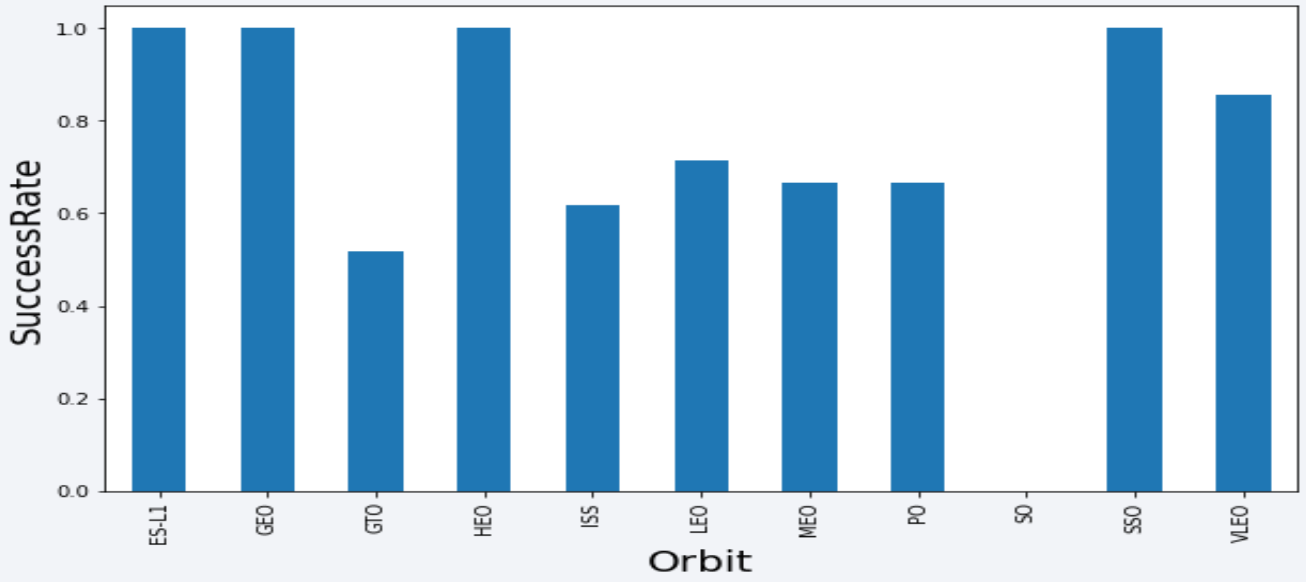
- As the flight number increases, the first stage is more likely to land successfully.
- The more massive the payload, the less likely the first stage will return.

Payload vs. Launch Site



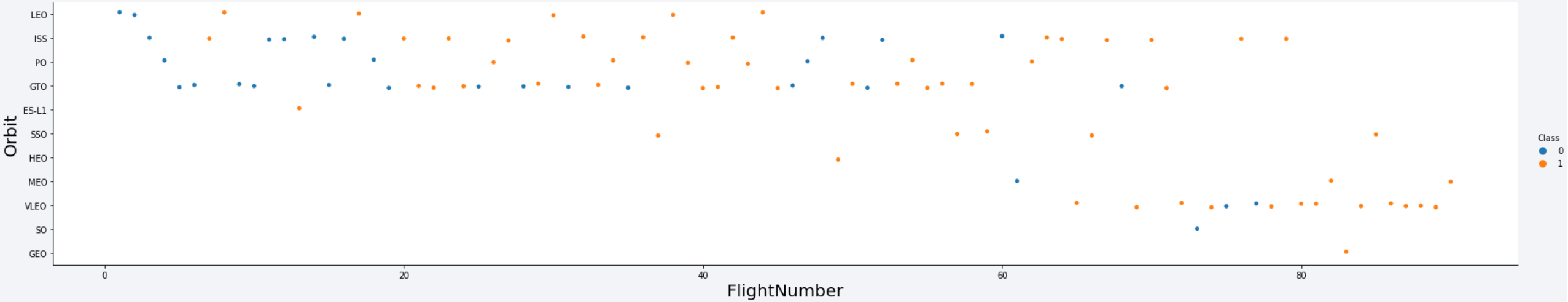
- For the VAFB-SLC laun chsite there are no rockets launched for heavy payload mass(greater than 10000).
- And CCAFS SLC40 and KSC LC 39A launch sites have much better success rate for rockets having heavier payload mass (greater than 10000)

Success Rate vs. Orbit Type



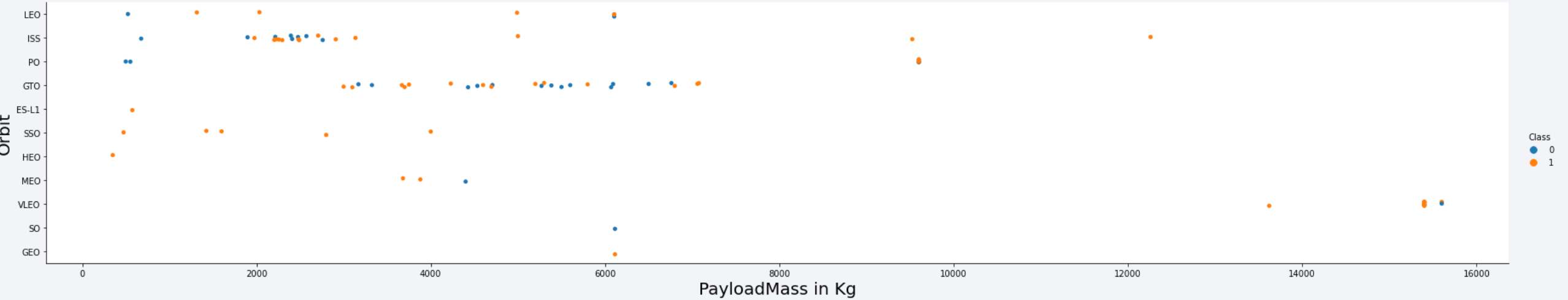
- ES-LQ ,GEO ,HEO and SSO orbits have higher success rate compared to other orbits

Flight Number vs. Orbit Type



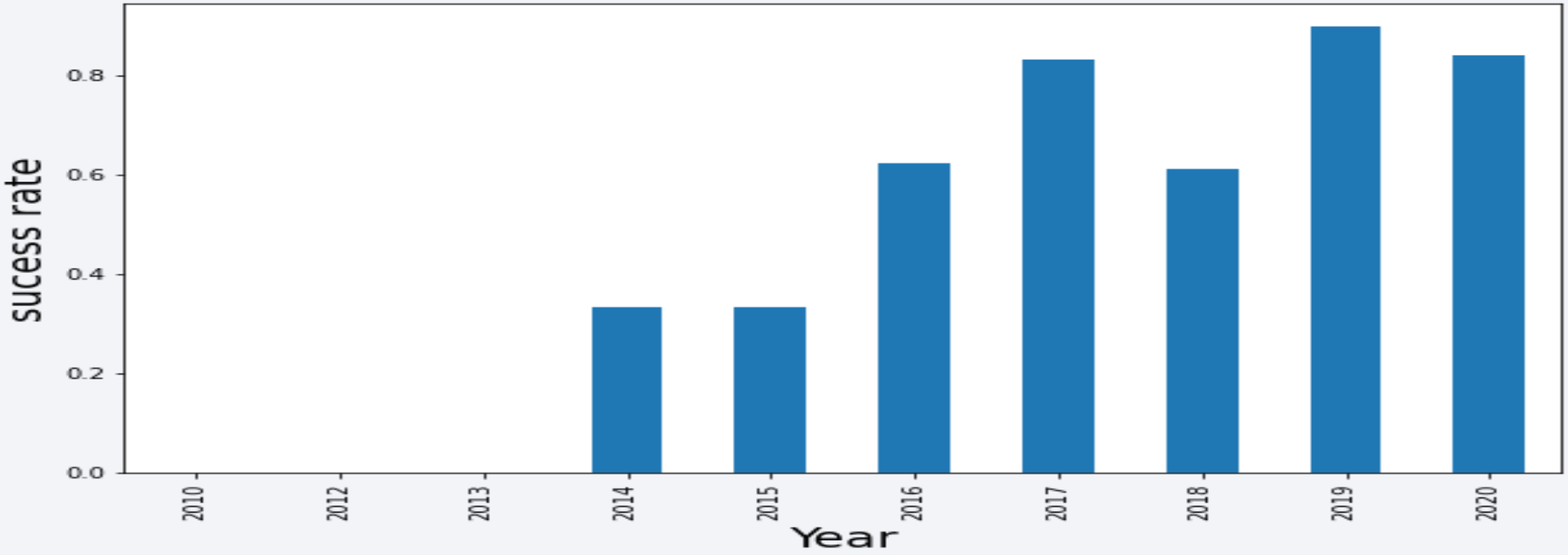
- For LEO and VLEO orbits the Success appears to be higher when the number of flights increase ; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO it can not be distinguished well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

All Launch Site Names

- The names of the unique launch sites are
- CCAFS LC-40
- KSC LC-39A
- VAFB SLC-4E
- CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- There are only two records where launch sites begin with the string 'CCA'
- CCAFS SLC-40
- CCAFS LC-40

Total Payload Mass

- The total payload mass carried by boosters launched by NASA (CRS) is 45596Kg
- It is calculated by adding the payload mass of the rockets for NASA(CRS) boosters .

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.40 Kg

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was in 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

➢ F9 FT B1022

➢ F9 FT B1031.2

➢ F9 FT B1026

➢ F9 FT B1021.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes are
- Successful mission = 100
- Failure = 1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass
- F9 B5 B1048.4 - F9 B5 B1060.2
- F9 B5 B1049.5 - F9 B5 B1058.3
- F9 B5 B1049.4 - F9 B5 B1056.4
- F9 B5 B1048.5 - F9 B5 B1051.6
- F9 B5 B1051.4 - F9 B5 B1060.3
- F9 B5 B1049.7 - F9 B5 B1051.3

2015 Launch Records

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are listed below

Year	booster versions	Site name
2015	F9 v1.1 B1012	CCAFS LC-40
2015	F9 v1.1 B1015	CCAFS LC-40

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

- The rank 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	Count	Rank_no
No attempt	10	8
Failure (drone ship)	5	5
Success (drone ship)	5	5
Success (ground pad)	5	5
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	3
Failure (parachute)	1	1
Precluded (drone ship)	1	1

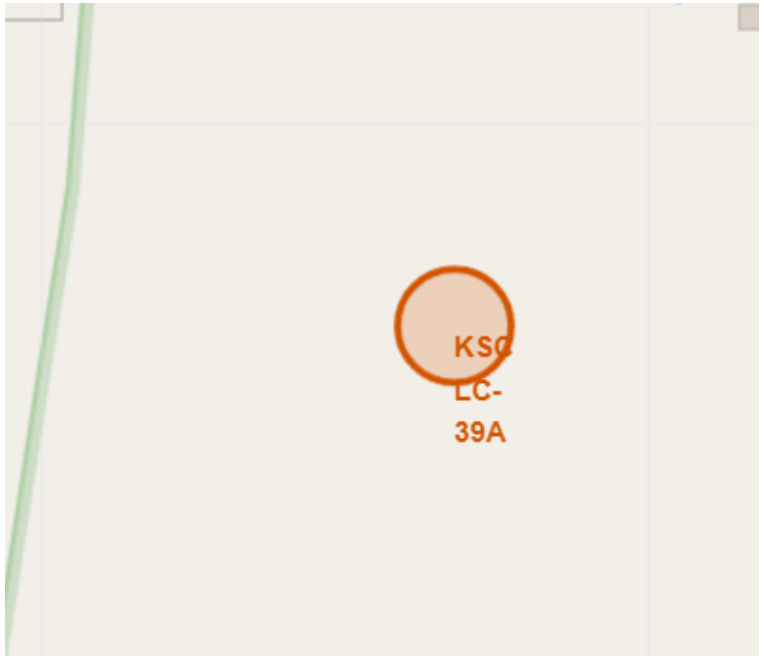
Section 3

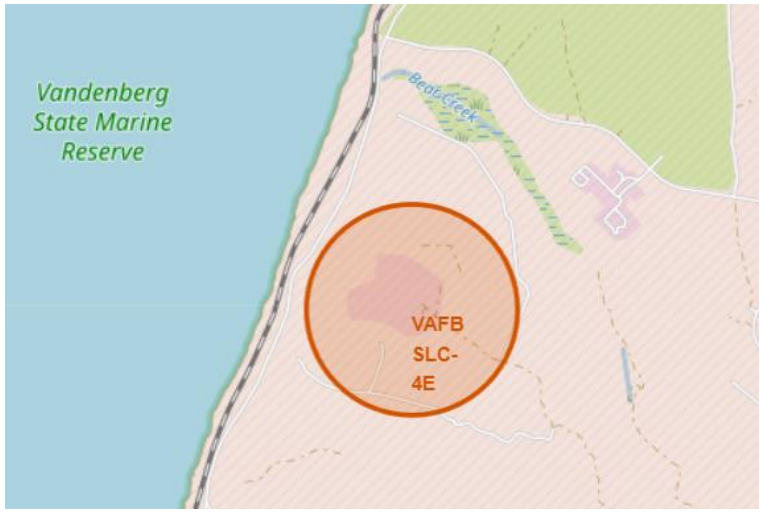
Launch Sites Proximities Analysis

Launch Sites Locations Analysis

From the map

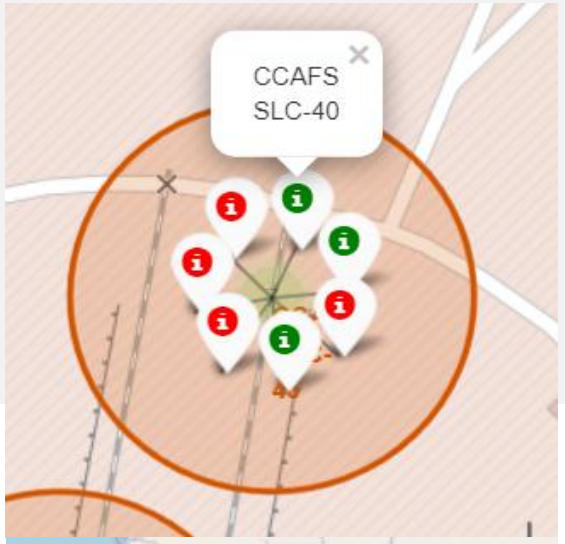
- The launch sites are very close to the coast
- Not far from the Equator

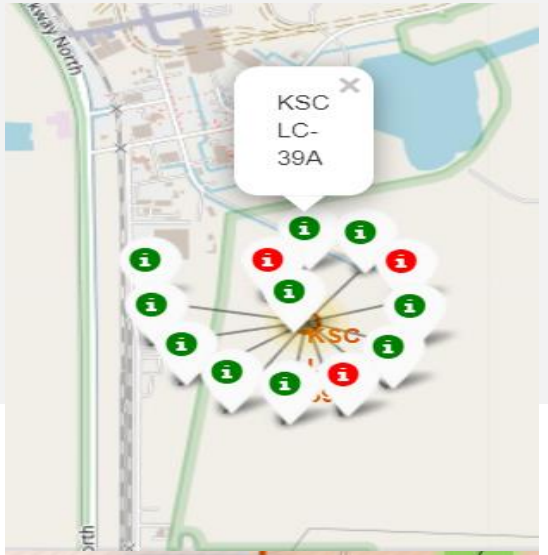


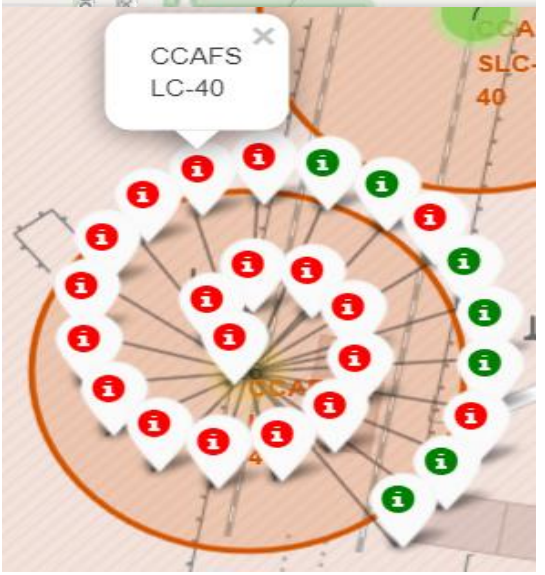


Launch Sites Locations and success rate

- KSC LC 39A Launch site has the highest success rate compared to all launch sites

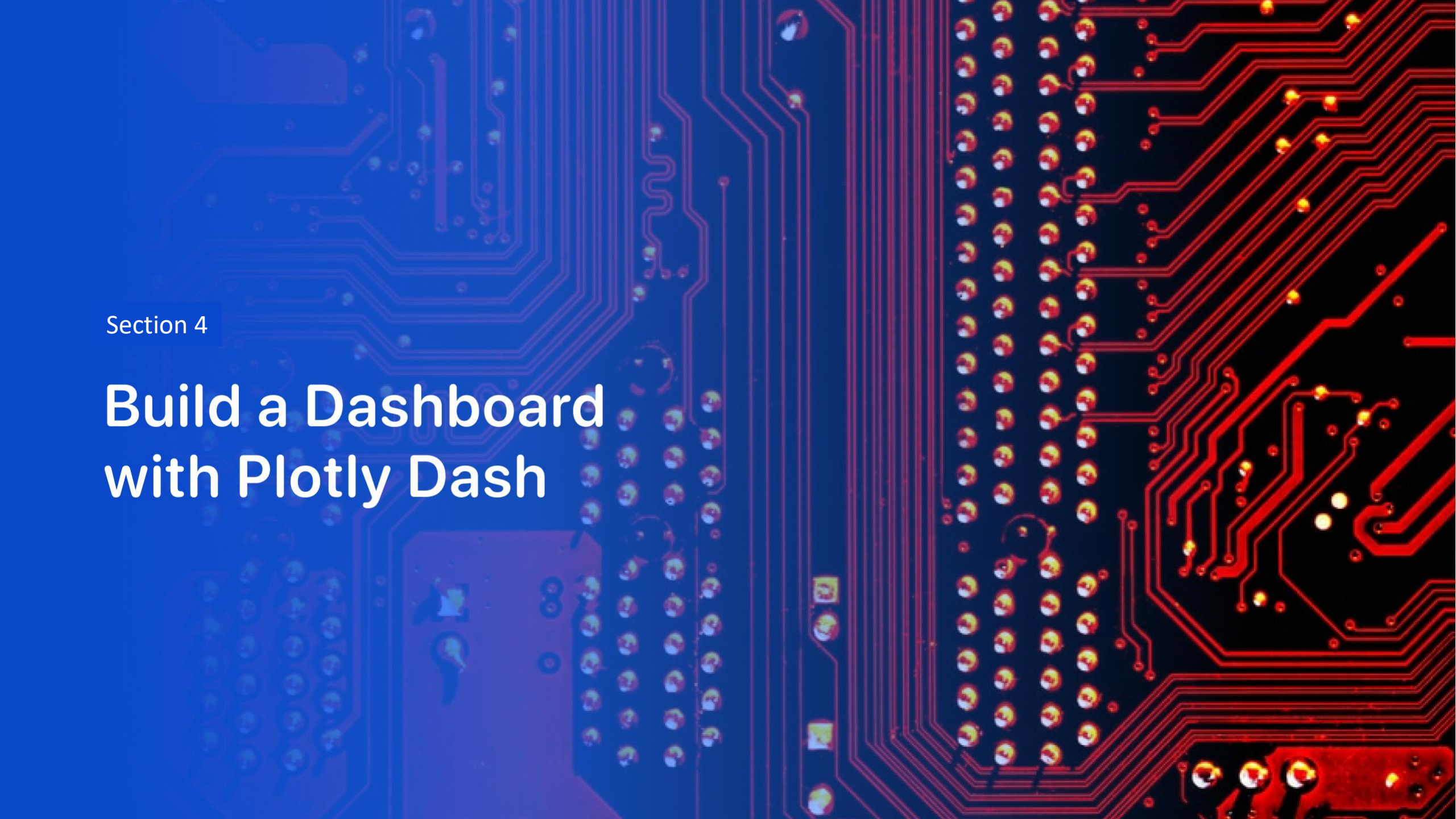






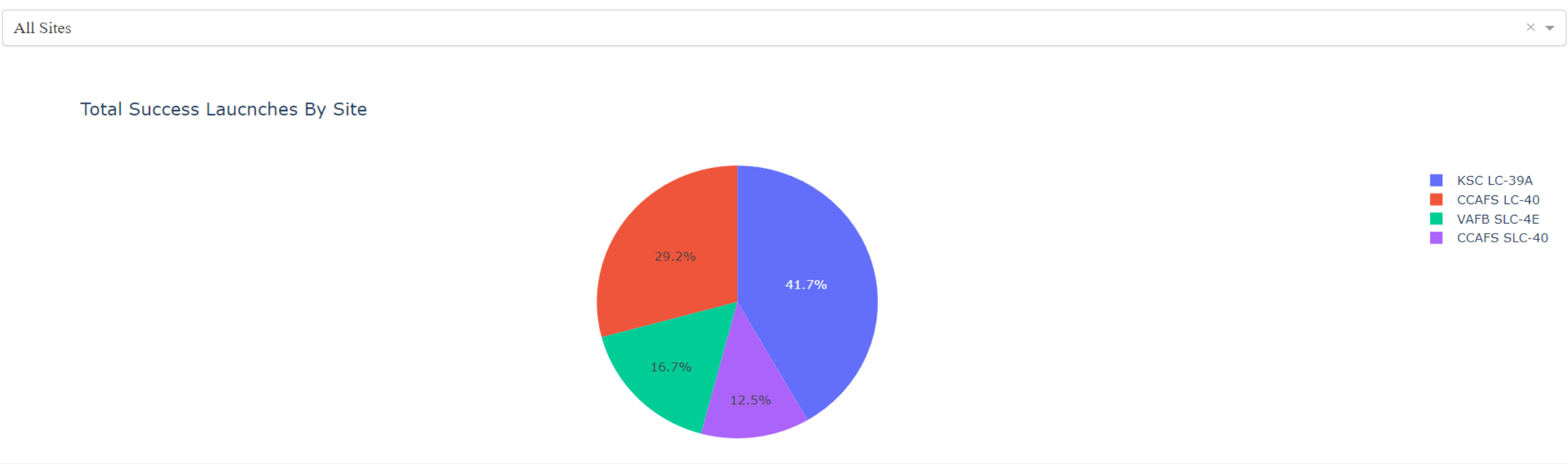
Distances between a launch site to its proximities

- The launch sites are close to the coastline
- They are far from the citifies but some of them example CAFS SLC-40 are close to railway


Section 4
Build a Dashboard with Plotly Dash

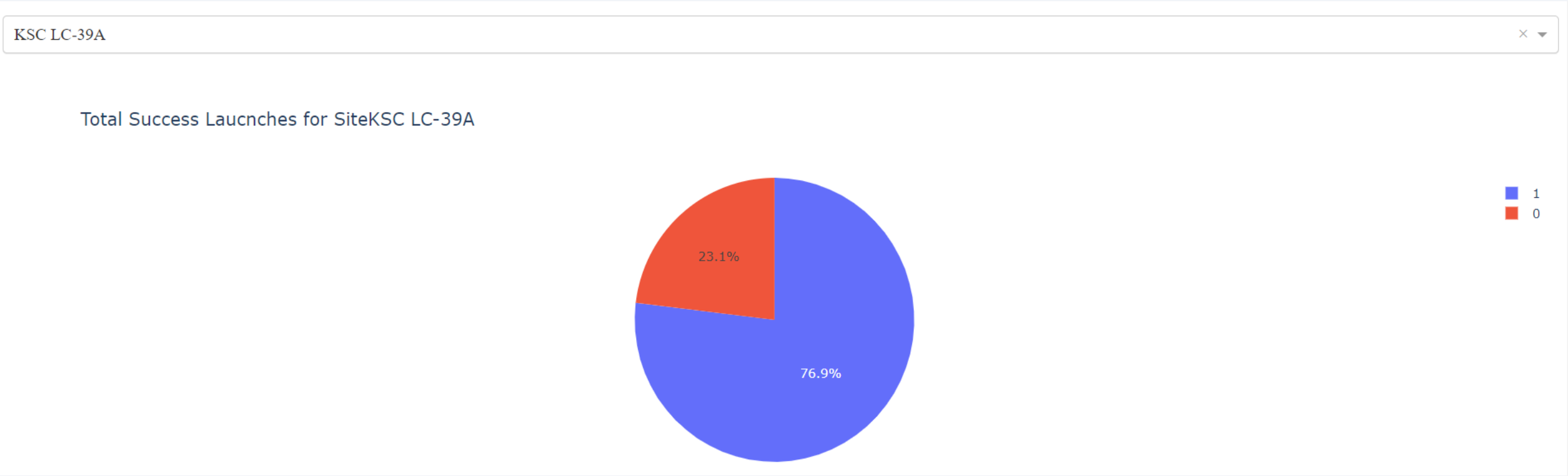
SpaceX Launch sites vs success rate Dashboard

- The Pichart indicates next to KSC LC-39A site CCAFS LC-40 launch site has the greater success rate of landing compared to CCAF SLC-40 and VAFB SLC-4E launch sites.



SpaceX highest success rate Launch site Dashboard

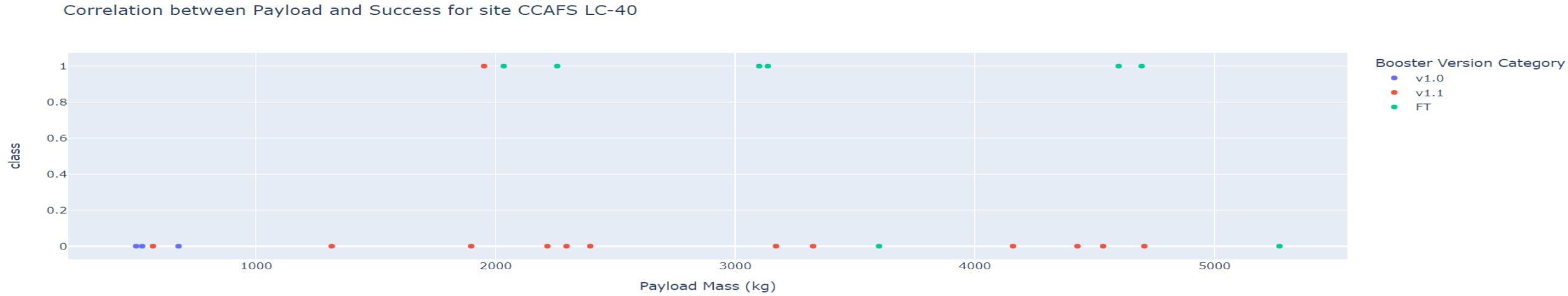
- The pichart in the dashboard shows that around 77% of KSC LC-39A landing tests were successful.



SpaceX Payload vs Launch Outcome dashboard

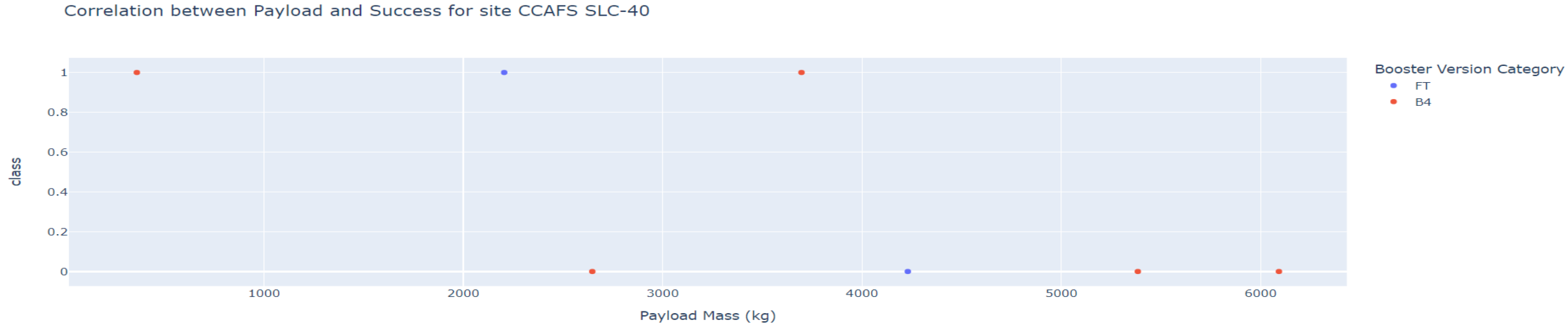
Payload range (Kg):

0 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k

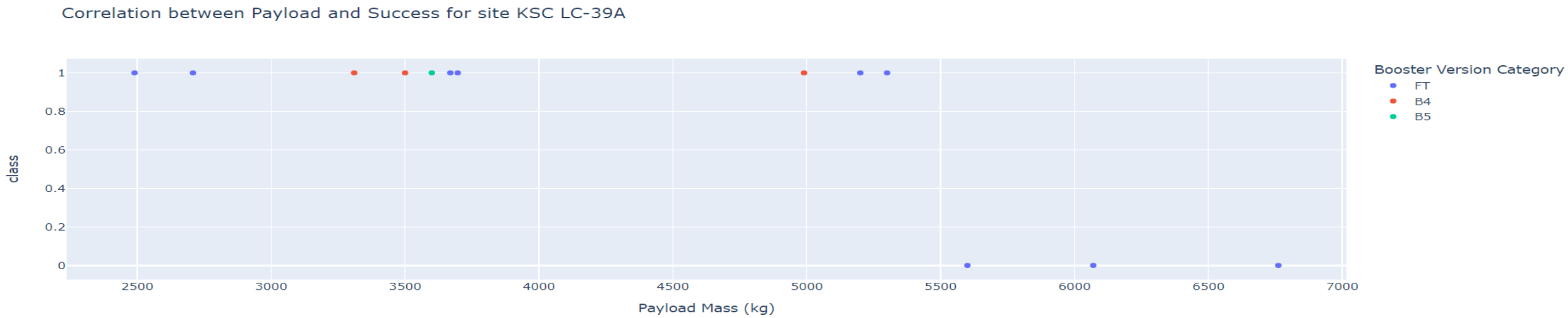


Payload range (Kg):

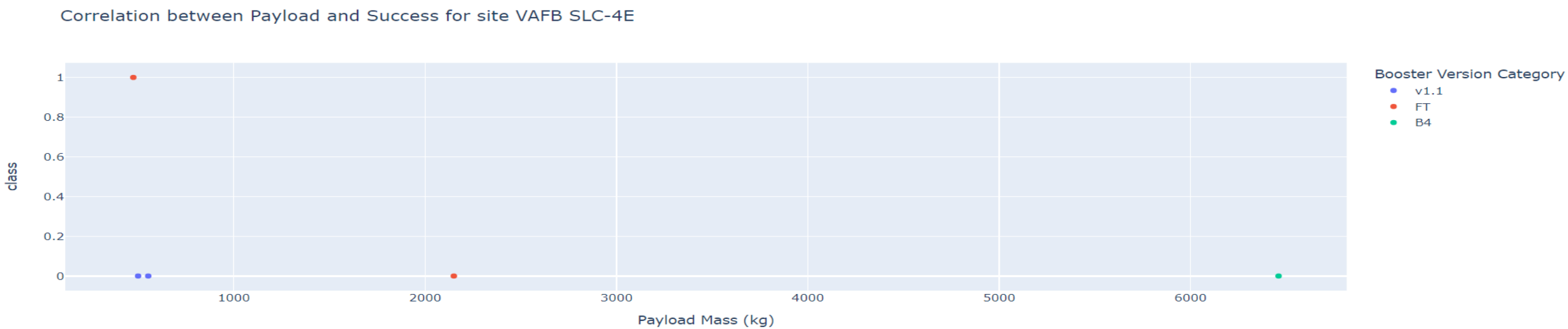
0 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k



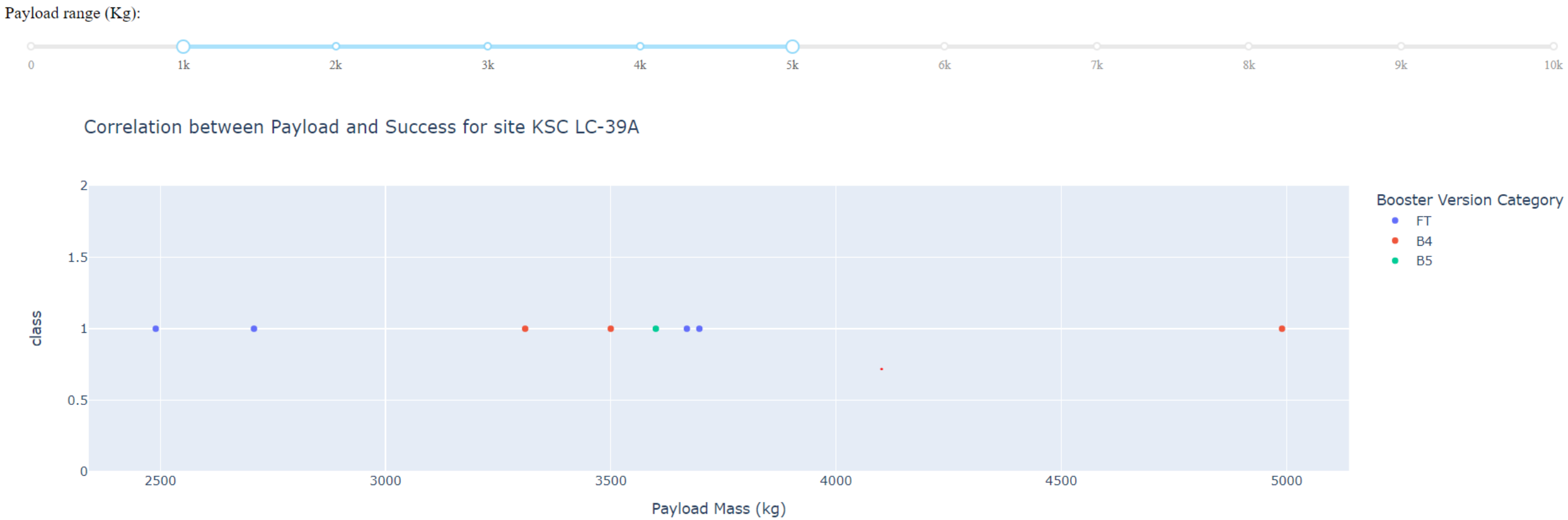
Payload range (Kg):



Payload range (Kg):



- From 1000kg to 5000Kg range payload mass the KSC LC-39A launch site has the highest success rate
- For less than 2000Kg payload the success rate is very low for almost all launch sites.

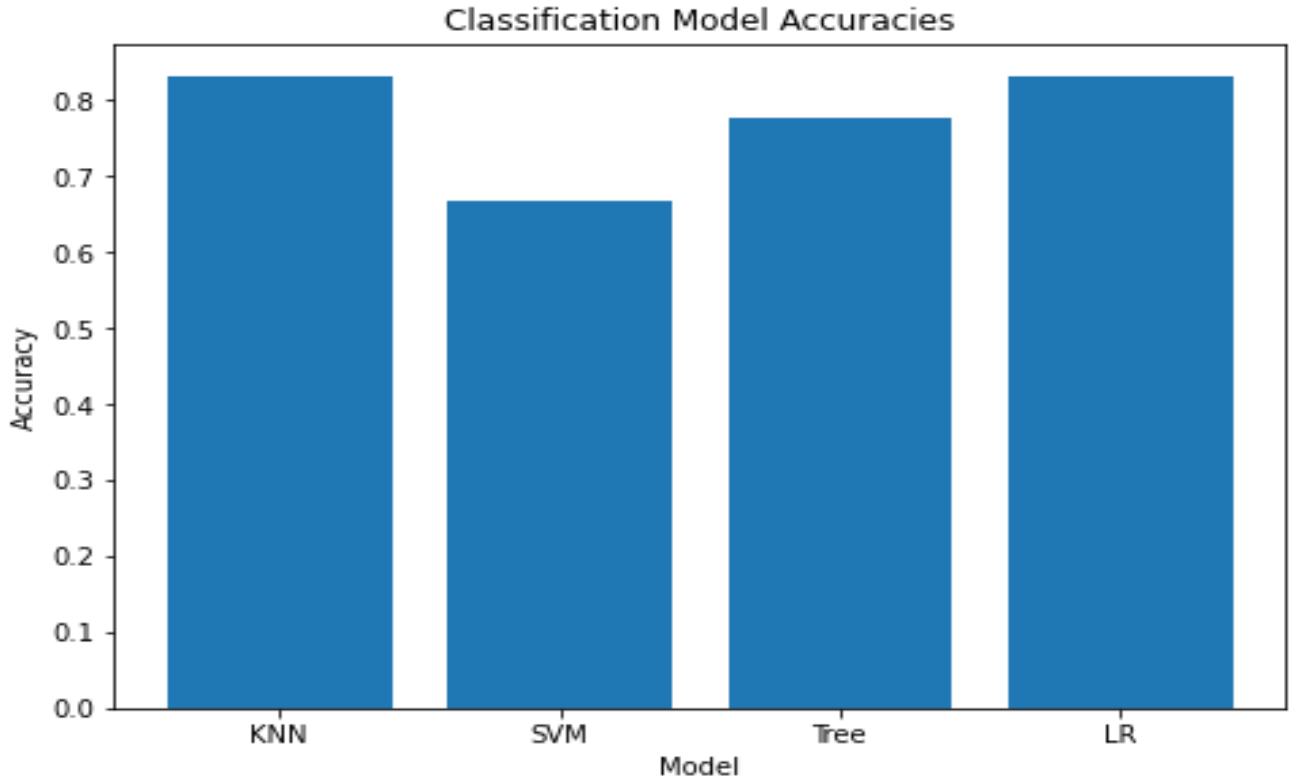


Section 5

Predictive Analysis (Classification)

Classification Accuracy

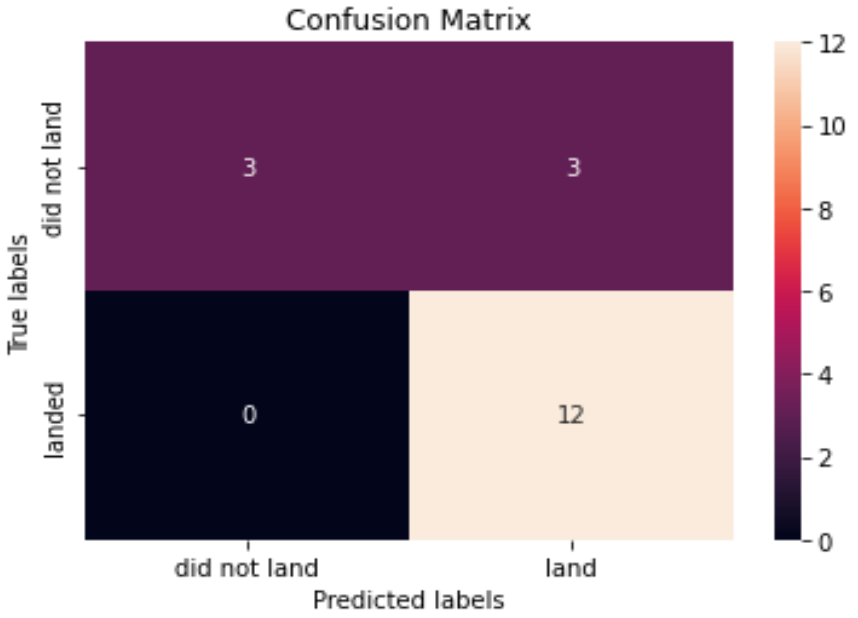
- Logistic regression and knn has similar and best train and test accuracy
- lr_train_score = 0.87
- lr_test_score = 0.83
- knn_train_score = 0.86
- knn_test_score = 0.83



Confusion Matrix

- According to the confusion matrix the errors are when it didn't land the model predicted incorrectly as landed which means it shows false positive errors .
- precision = TP/TP+FP = 12/(12+3) = 0.8
- Recall = TP/TP+FN = 12/(12+0) = 1
- F1 score = 2(P*R)/(P+R) = 2*(0.8*1)/(0.8+1) = 0.88
- Accuracy = TP + TN / TP + TN + FP + FN. = (12+3)/(12+3+3+0) = 0.833

The Model KNN and LR has an accuracy of 83.33% calculated from test data and from confusion matrix.



Conclusions

- The **Falcon 9 first-stage landing tests** were a series of controlled-descent flight tests conducted by SpaceX between 2013 and 2016. Since 2017, the first stage of Falcon 9 missions has been routinely landed.
- Flight number, payload mass, launch site, Orbit are highly correlated with the success rate of the tests.
- By learning from a number of flights the success rate of landing first stage increases yearly by adjusting the payload mass.
- CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Lower orbit, LEO tests has showed much better success compared to other orbits as flight number increases.
- By selecting the above parameters and using recent dataset KNN model can be used to predict the success of stage 1 landing of new launches .

Appendix

- All the project source codes are documented on github with the following link.
- Haftom-sig/Applied-Data-Science-Capstone: Final Presentation (github.com)
- External resources used are listed below
- Falcon 9 first-stage landing tests – Wikipedia
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Thank you!