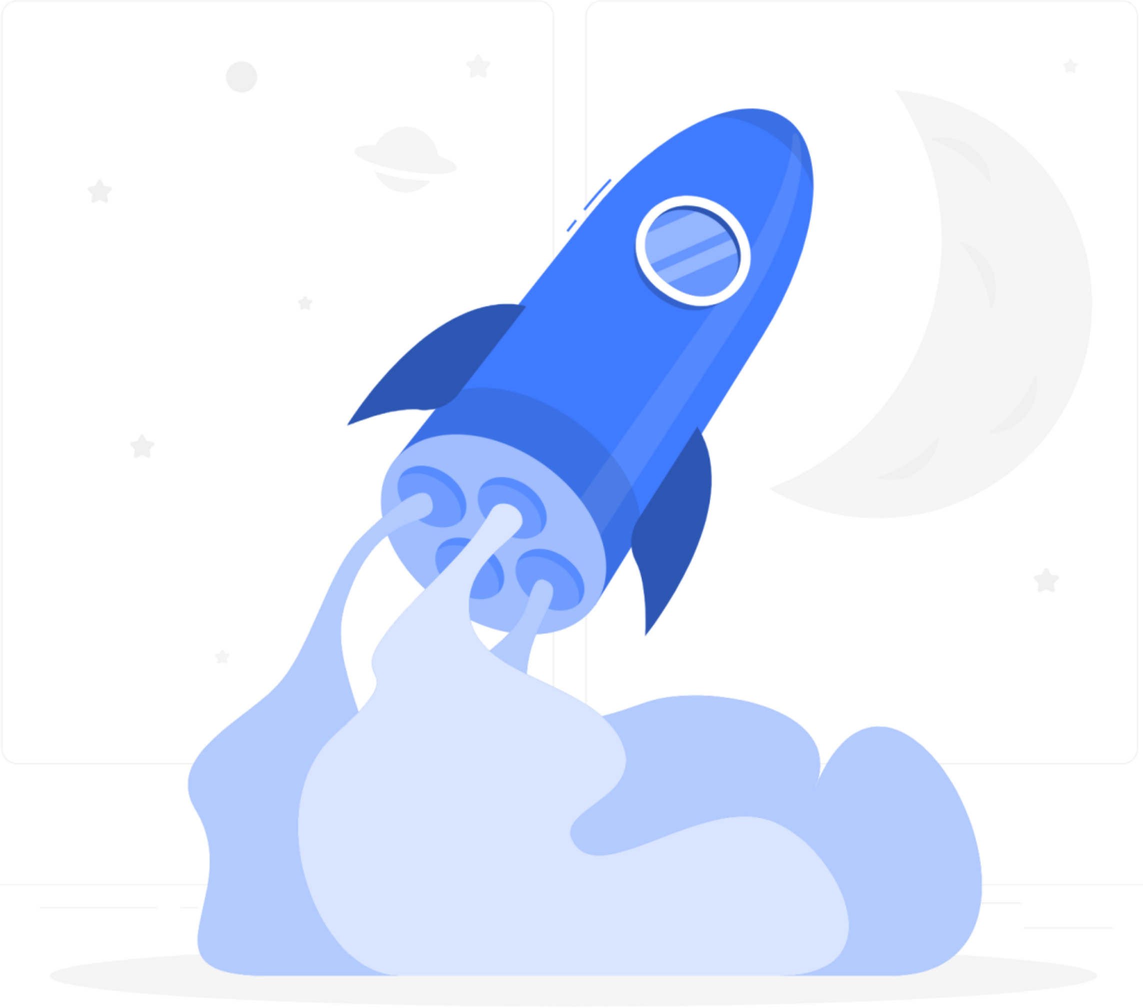
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

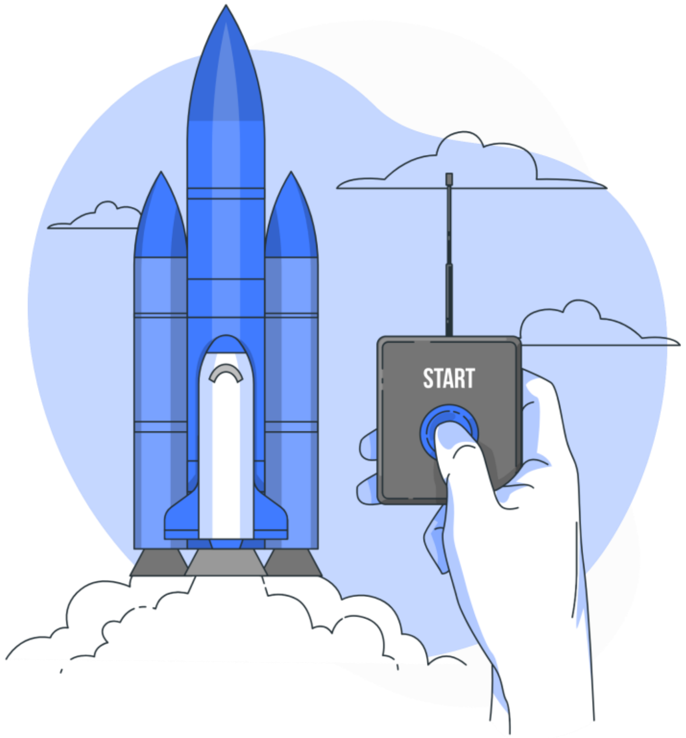
Jaime Solis
June 28^{th}, 2022

Table of Contents



IBM Developer SKILLS NETWORK

Executive Summary

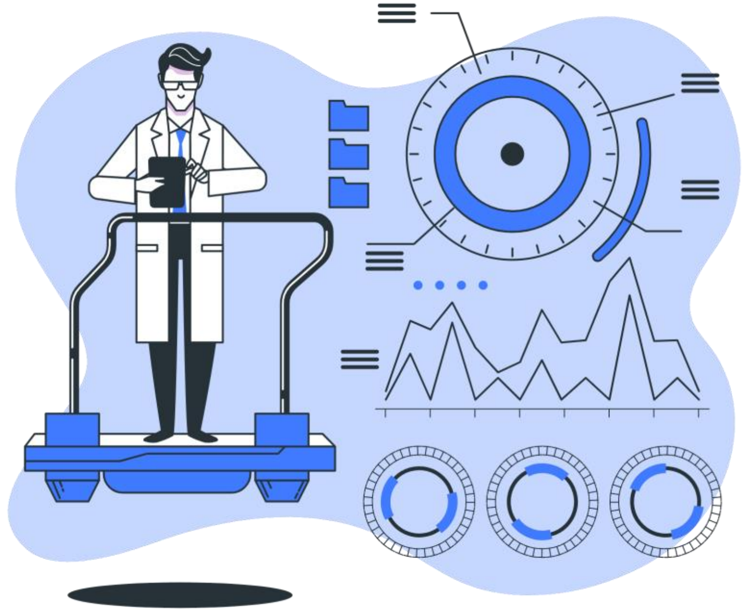


- **SpaceY**, a new rocket launch company, wants to compete with **SpaceX**.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of **62 million dollars**; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- If it can be determined if the first stage will land, it can be determined the cost of a launch.
- After collecting and analyzing data some models were trained to predict if the first-stage rocket booster will land successfully after a launch with an **accuracy level of 83.33%**, given some parameters like payload mass and desired orbit.
- As a result, SpaceY will be able to make more informed bids against SpaceX by using 1st stage landing prediction as a proxy for launch costs.

IBM Developer SKILLS NETWORK

Introduction — Backgound

- This report has been prepared as the final deliverable for the Applied Data Science Capstone course, part of IBM Data Science Professional Certification.
- In this project, the Data Science methodology has been followed, involving data collection, data wrangling, exploratory data analysis, data visualization, model development, model evaluation, and reporting results.
- Using data science findings and models, the data science team of the fictional company SpaceY, will be able to make more informed bids against SpaceX for a rocket launch.



IBM Developer
SKILLS NETWORK

Introduction — Business Problem

- 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, this report aims to accurately **predict the likelihood of the first stage rocket landing successfully** as a proxy to determine the cost of a launch, using parameters like payload mass, desired orbit, and launch site, among others.

Section 1

Methodology

Methodology

For this report, the data science methodology followed can be outlined as such:

1st step
Data Collection

2nd step
Data Wrangling

3rd step
Perform exploratory data analysis (EDA)

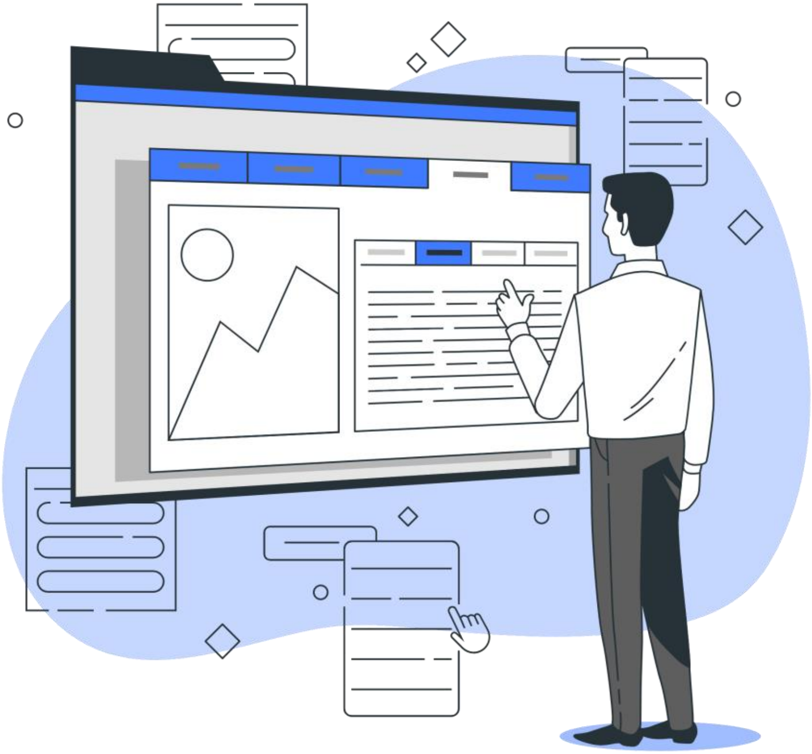
4th step
Data visualization

5th step
Model Development

Reporting Results

IBM Developer SKILLS NETWORK

Data Collection



Falcon 9 launch dataset was limited to launches before December 2020 per instructions

IBM Developer SKILLS NETWORK

1) API

Historical launch data from [SpaceX REST API]

Filtered data to include only Falcon 9 launches

Request and parse the SpaceX launch data using the GET request

Replaced missing payload mass values with mean

2) Web Scraping

Historical launch data from a Wikipedia page called "List of Falcon 9 and Falcin Heavy Launches[

Extract all column/variable names from the HTML table header

Request the Falcon9 Launch Wiki page from its URL with BeatifulSoup

Created a data frame by parsing the launch HTML tables

Data Wrangling

- Explored data to determine the label for training supervised models
  - Calculated the number of launches on each site
  - Calculated the number and occurrence of each orbit
  - Calculated the number and occurrence of mission outcome per orbit type
- Created a landing outcome label ‘Class’ from Outcome column

True ASDS	41
None None	19
True RTLS	14
False ASDS	6
True Ocean	5
False Ocean	2
None ASDS	2
False RTLS	1

First stage booster did not land successfully
Class = 0

None None: not attempted
None ASDS: unable to be attempted due to launch failure
False ASDS: drone ship landing failed
False Ocean: ocean landing failed
False RTLS: ground pad landing failed

First stage booster landed successfully
Class = 1

True ASDS: drone ship landing succeeded
True RTLS: ground pad landing succeeded
True Ocean: ocean landing succeeded

IBM Developer
SKILLS NETWORK

Exploratory Data Analysis

a) with SQL

Ran SQL queries to display information about

- Launch sites
- Payload masses
- Booster versions
- Mission outcomes
- Booster landings

b) with visualization

Used Matplotlib and Seaborn libraries to plot

- Flight Number x Payload Mass
- Flight Number x LaunchSite
- Payload x LaunchSite
- Orbit type x Success rate
- Flight Number x Orbit type
- Payload x Orbit type
- Year x Success rate

Data Visualization

Launch Sites Location Analysis

- Used Folium, an interactive mapping library for Python
- Marked all launch sites on map
- Marked the successful/failed launches for each site
- Calculated distances between a launch site and its proximities

Launch Records Dashboards

- Used Plotly Dash, an interactive dashboarding library for Python
- Added pie chart showing success rate by site
- Added scatter chart showing payload mass vs landing outcome
- Added drop-down menu to choose between all sites and individual launch site
- Added range slider for limiting payload amount

IBM Developer SKILLS NETWORK

Predictive Analysis (Classification)

Imported libraries and load data frame

Pandas
Numpy
Matplotlib
Seaborn
Scikit-learn

Created a column for training label 'Class'

Created during data wrangling

Standardized the data

Split data into train and test sets

Fit training data to models

Logistic Regression
SVM
Decision Tree
KNN

Used cross-validated grid-search with hyperparameter to select the best ones

GridSearchCV

Evaluated accuracy of each model using test data to select the best model

IBM Developer SKILLS NETWORK

Results

- Exploratory data analysis results
- Interactive analytics demo
- Predictive analysis results

IBM Developer SKILLS NETWORK

Section 2

Insights drawn from EDA

EDA with SQL

- Names of the launch sites
  CCAFS LC-40
  CCAFS SLC-40
  KSC LC-39A
  VAFB SLC-4E

- Records where launch sites begin with `CCA`
  Last launch from CCAFS LC-40 was 2016-08-14
  First launch from CCAFS SLC-40 was 2017-12-15

- Total payload carried by boosters from NASA
  45,596 KG.

- Average payload mass carried by booster version F9 v1.1
  2,534 KG

- The dates of the first successful landing outcome on ground pad
  01-05-2017

EDA with SQL

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

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

- The total number of successful and failed mission outcomes

Failure (in flight)	1
Success	99
Success (payload status unclear)	1

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

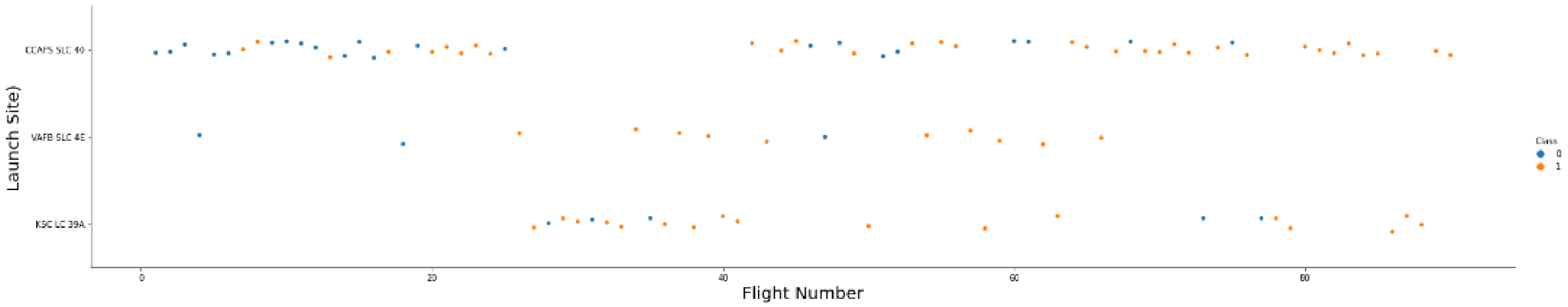
  B1048.4, B1049.4, B1051.3, B1056.4, B1048.5, B1051.4, B1049.5, B1060.2, B1058.3, B1051.6, B1060.3, B1049.7

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

January	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

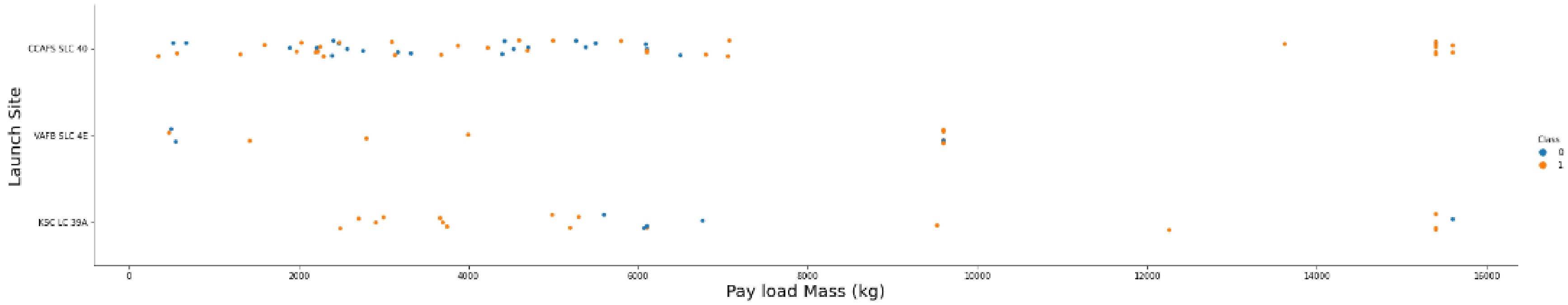
EDA with Visualization

- Launch Site vs Flight Number



CCAFS SLC 40 has more flights than the other launch sites combined and the first third of them were mostly failures but as the flight number increases, the first stage is more likely to land successfully.

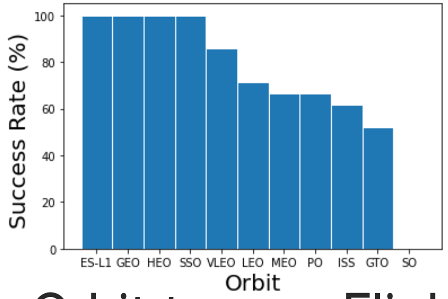
- Launch Site vs Payload Mass



VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

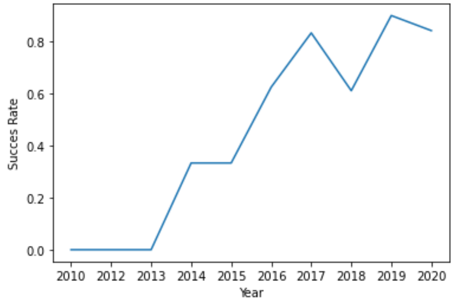
EDA with Visualization

- Success rate of each orbit type



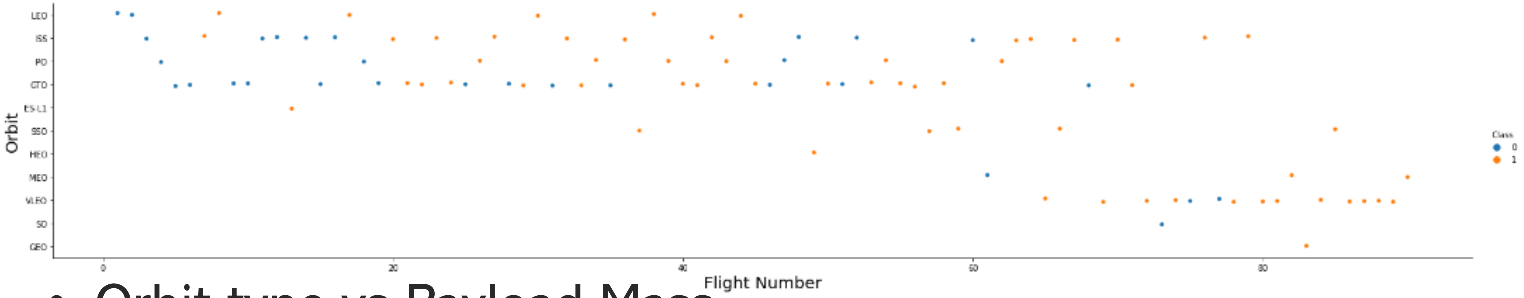
All orbit types have had successful landings except "SO"

- Yearly average success rate



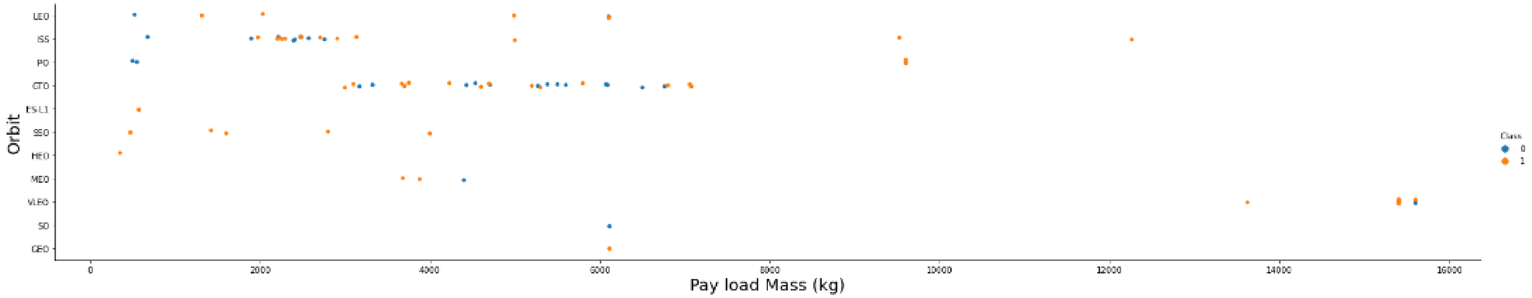
The success rate since 2013 kept increasing till 2020

- Orbit type vs Flight number



In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

- Orbit type vs Payload Mass



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

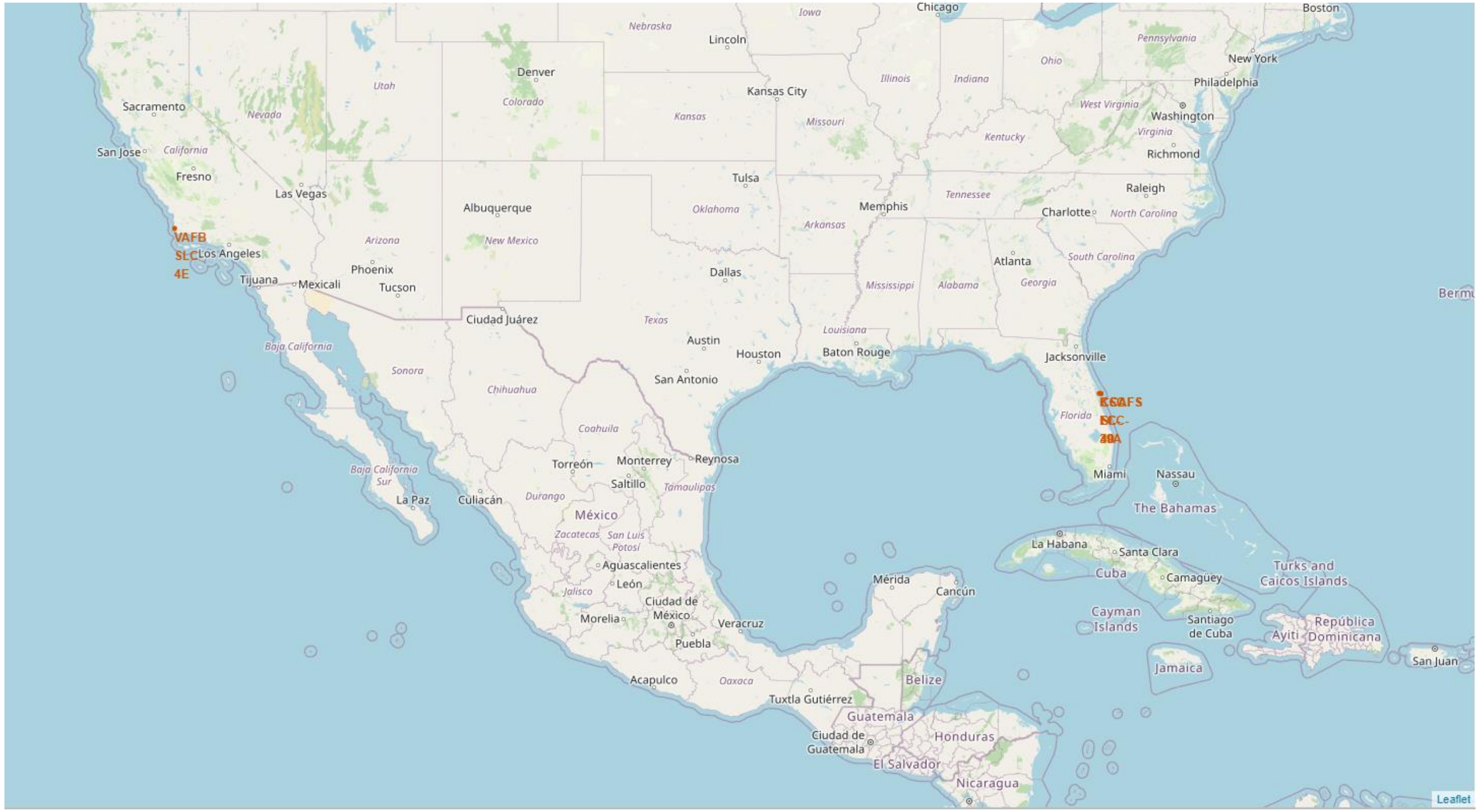
IBM Developer SKILLS NETWORK

Section 3

Launch Sites Proximities Analysis

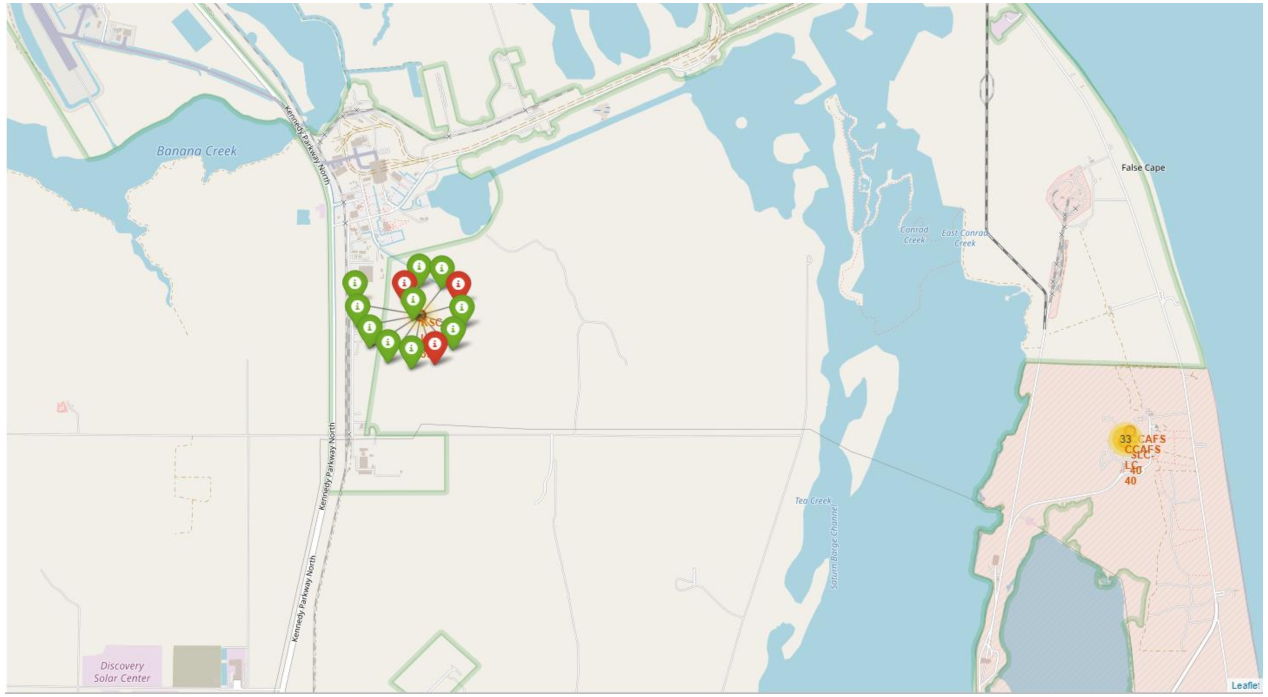
All Launch Sites

Visualizing all the launch sites on a map help to highlight the importance of launch site proximity to coasts and the equator.



Success/failed launches for each site

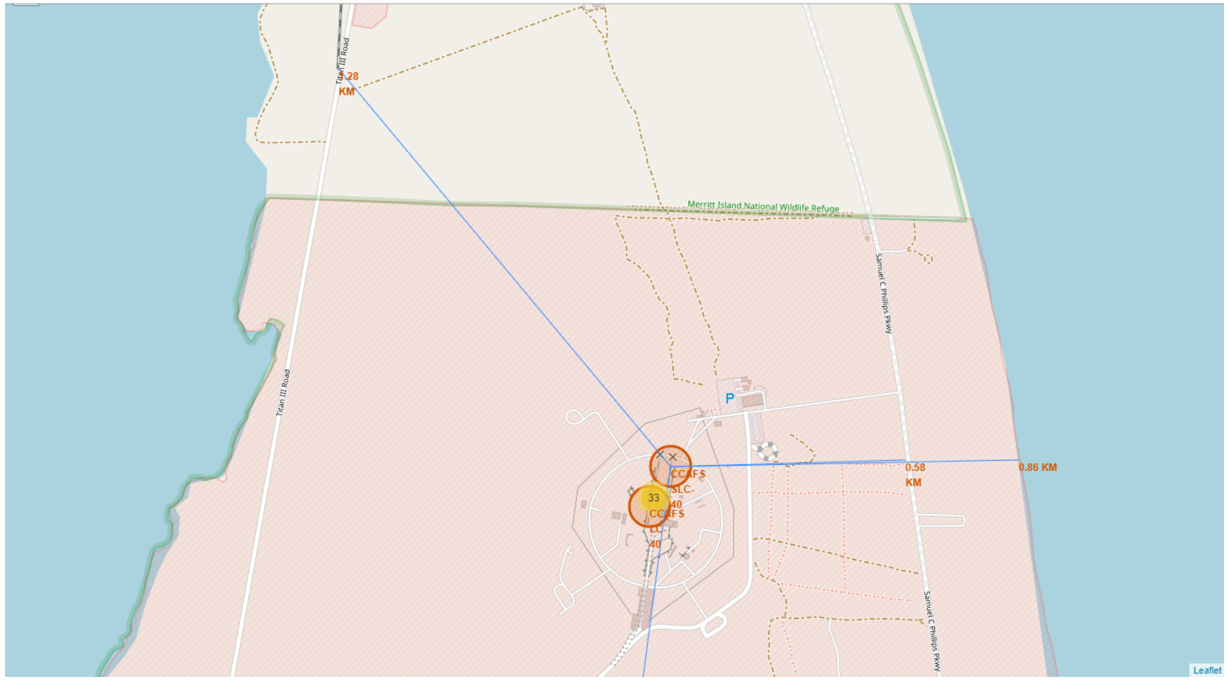
Visualizing landing outcomes for each launch site highlights which launch sites have relatively high success rates.



Distances between a launch site to its proximities

Visualizing proximities with railways, highways, coastlines and cities allows to conclude that launch sites are **near railways** (the transport for heavy cargo), **highways**(for easily transport required people and equipment), and **coastlines** (so they can fly over the ocean during launch if a launch abort is required and attempt a water landing to minimize risk for people and property from falling debris).

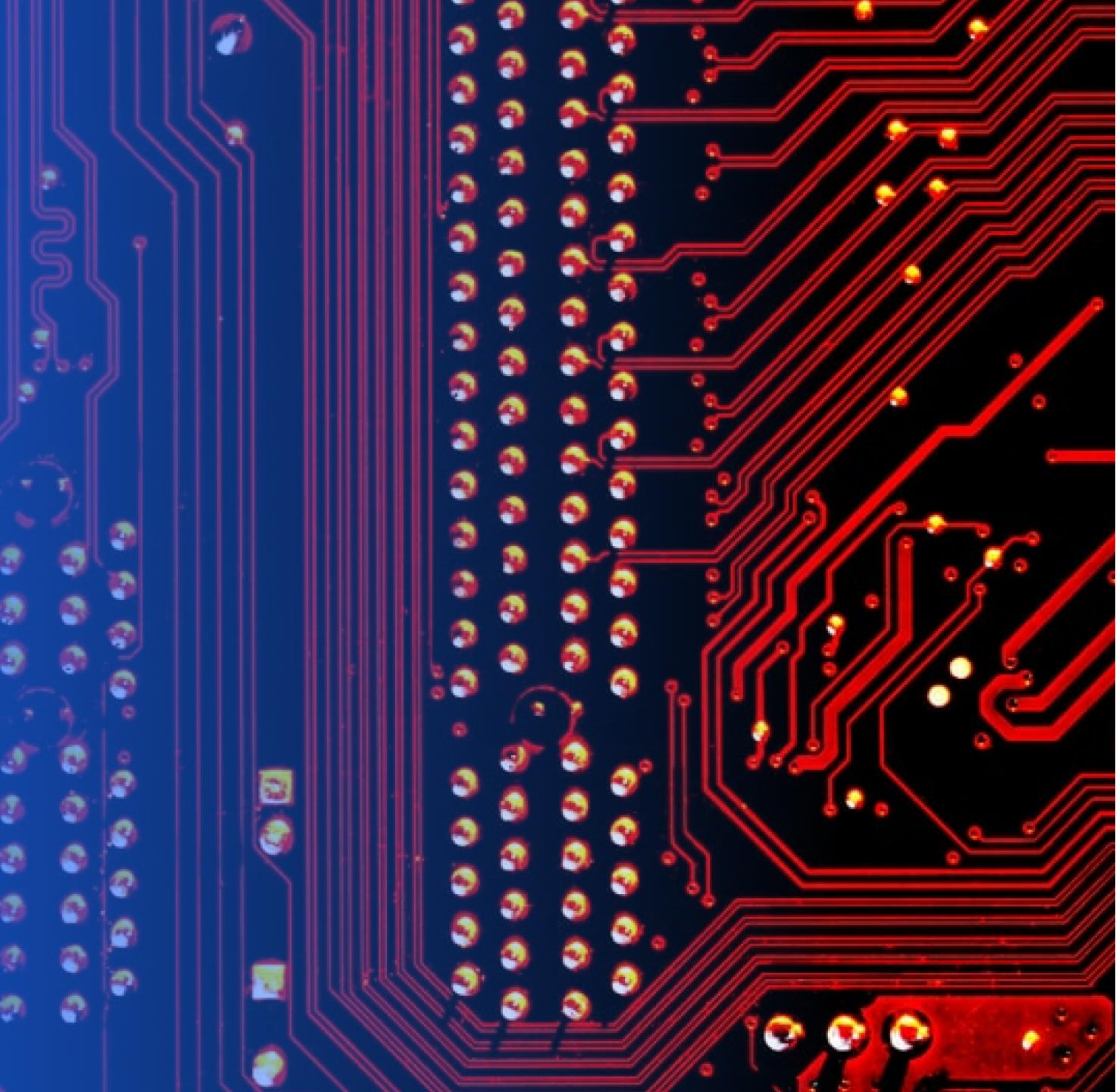
Launch sites **are not in close proximity to cities**, which minimizes danger to population-dense areas.



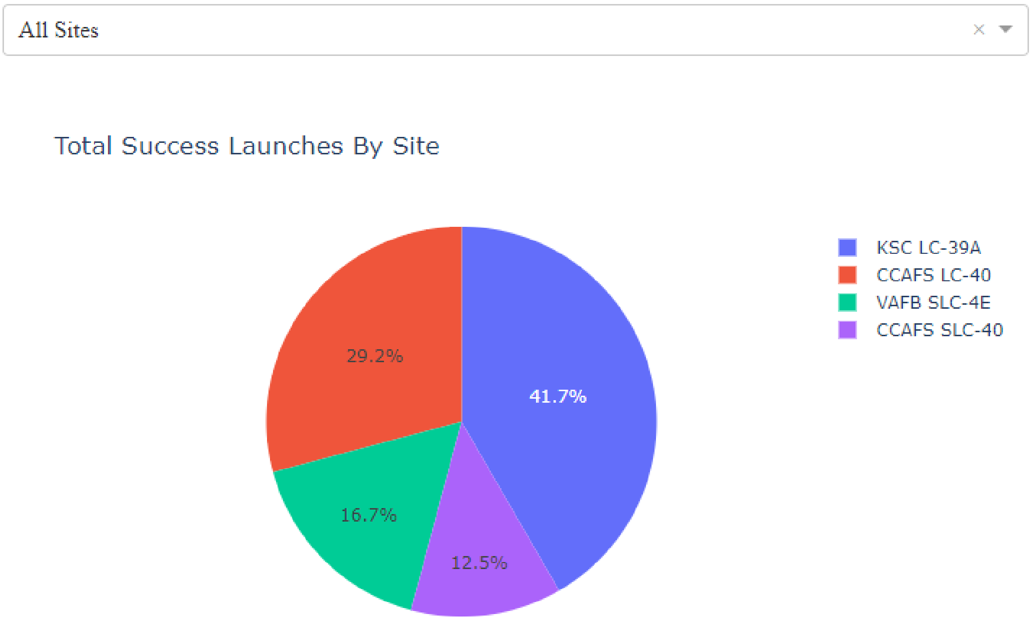
IBM Developer SKILLS NETWORK

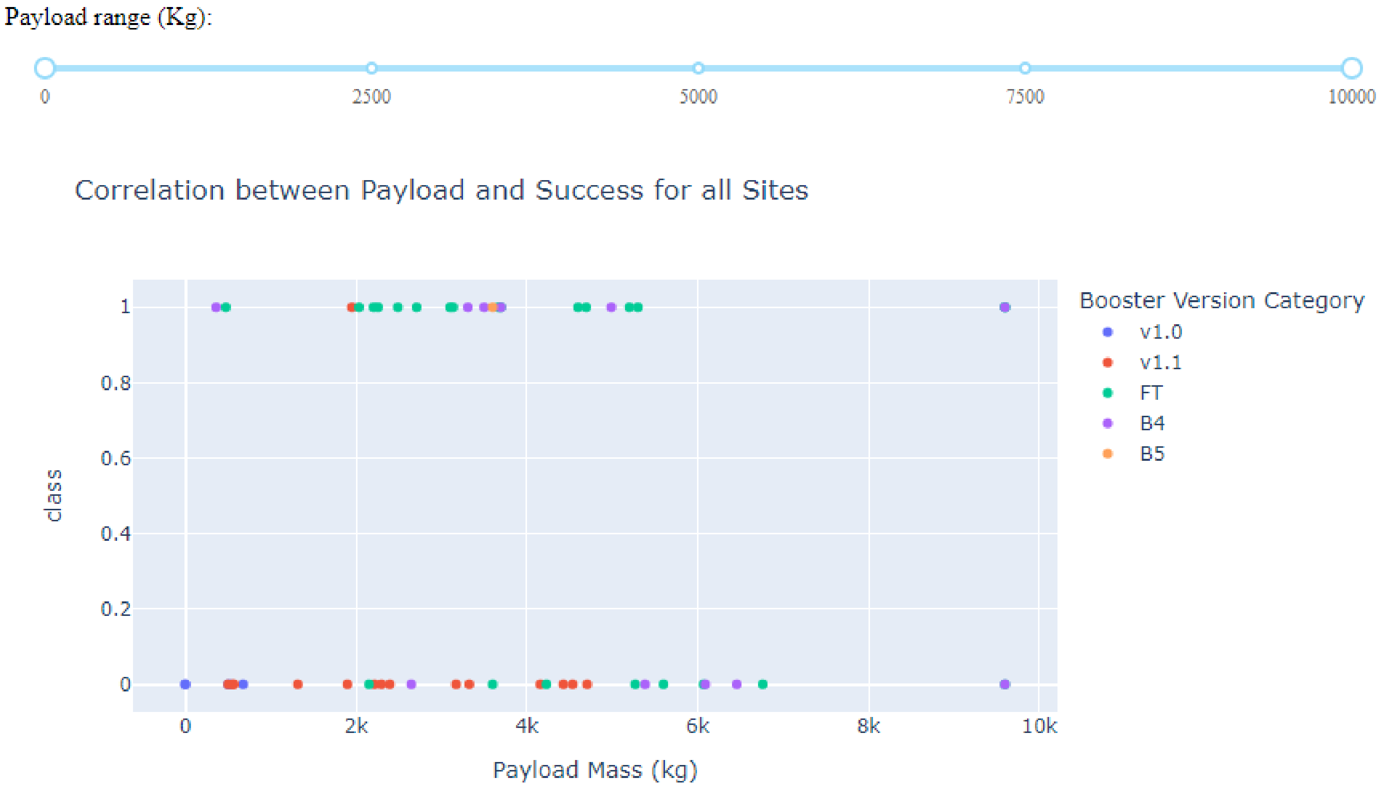
Section 4

Build a Dashboard with Plotly Dash



SpaceX Launch Records Dashboard



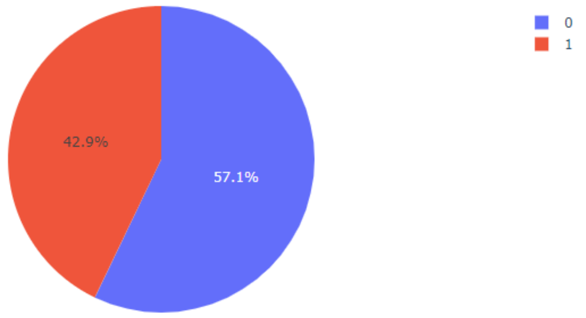


- KSC LC-39A is the launch site with the highest landing success rate considering all launches
- The heavier the payload is, the more likely failure landing is
- VAFB SLC-4E has the heaviest succesful landing.

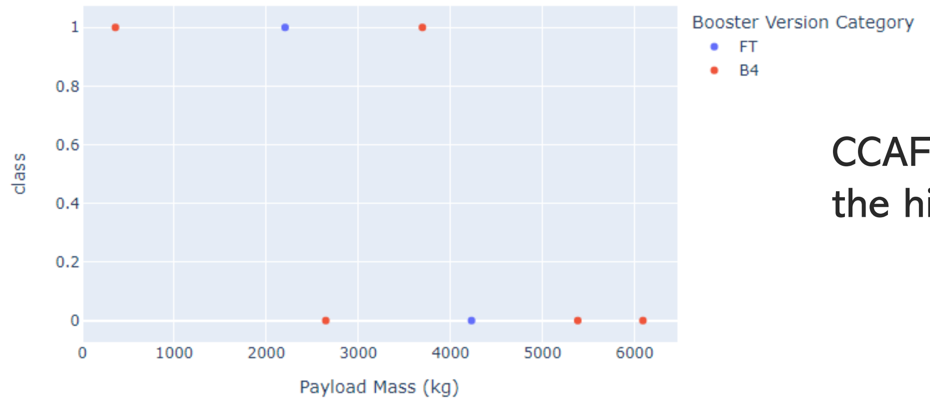
IBM Developer SKILLS NETWORK

Observations

Total Success Launches for site CCAFS SLC-40



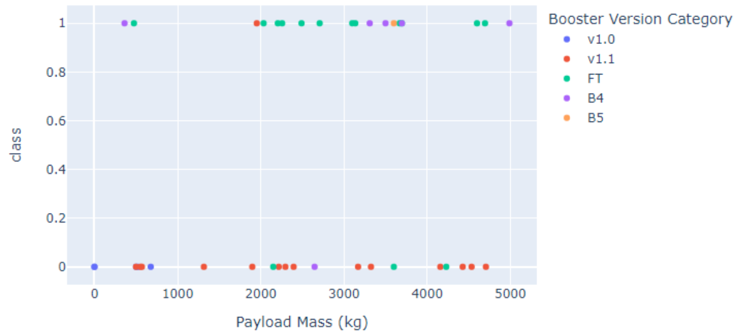
Correlation between Payload and Success for site CCAFS SLC-40



CCAFS SLC-40 is the launch site with the highest success rate of 42.9%.

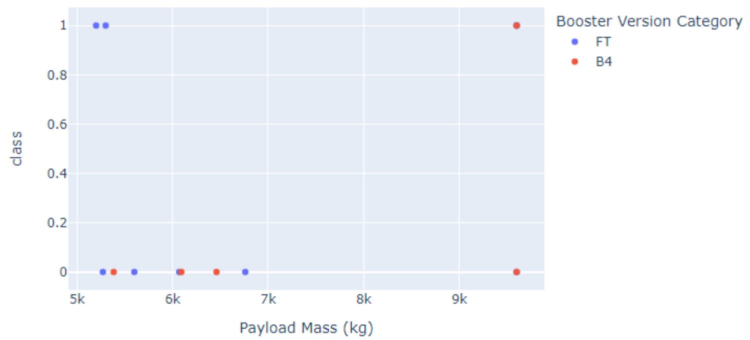
Payload range (Kg):

Correlation between Payload and Success for all Sites



Payload range (Kg):

Correlation between Payload and Success for all Sites



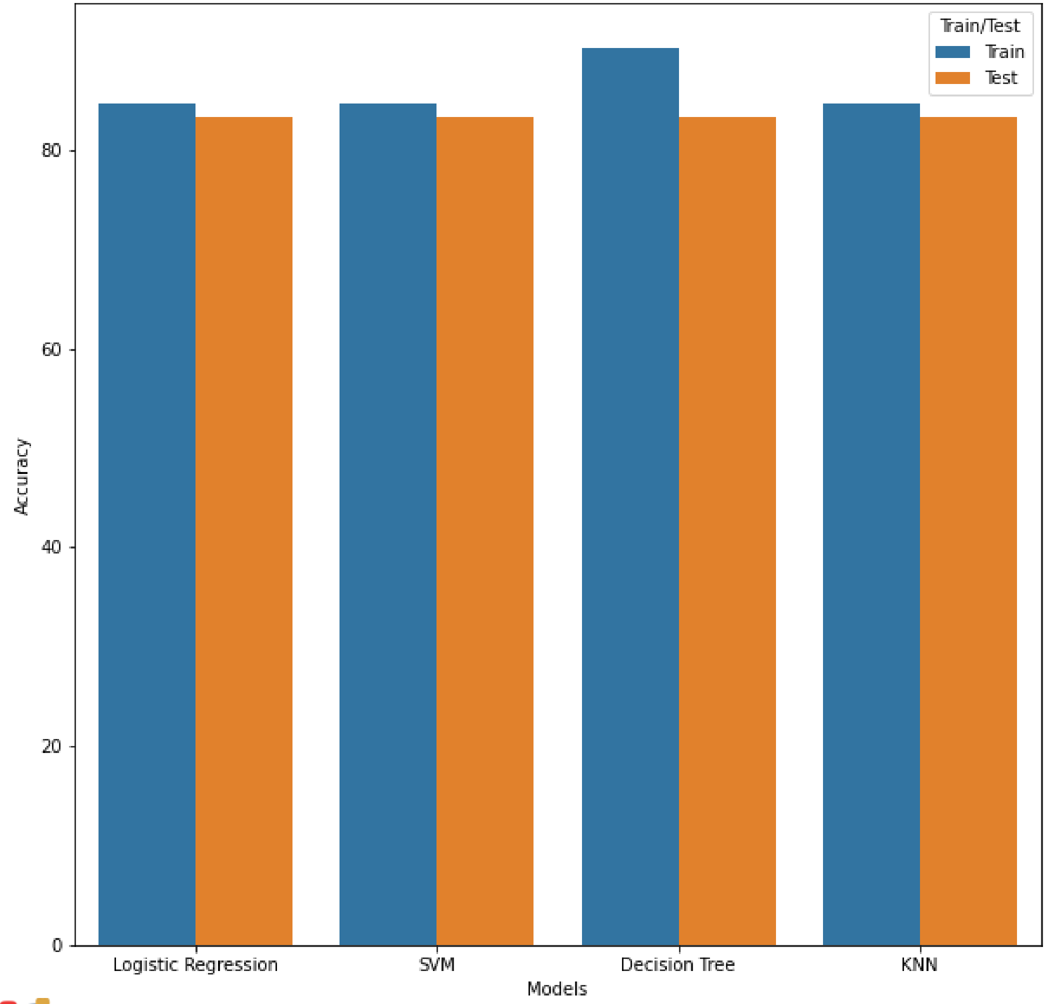
Payloads < 5000 kg have higher booster landing success.

IBM Developer SKILLS NETWORK

Section 5

Predictive Analysis (Classification)

Classification Accuracy

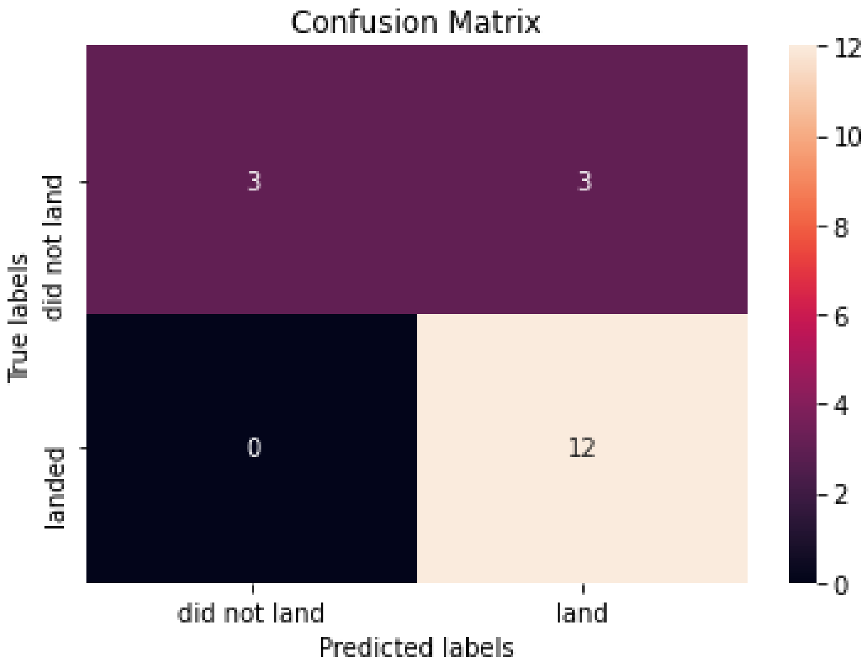


- All models performed practically the same, except for the Decision Tree Model which fitted better on train data but worse on test data.
- Each of the four models retrieved the same accuracy score of 83.33%

IBM Developer SKILLS NETWORK

Confusion Matrix

- The confusion matrices for all four models are the same.
- The main issue with the models is false positives as they incorrectly predicted 3 failure landings as successful ones out of 18 samples in the test set.



Conclusions

- Using the models from this report SpaceY can predict when SpaceX will successfully land the 1st stage booster with **83.3% accuracy**
- This will enable SpaceY to make more informed bids against SpaceX, since they will have a good idea of when to expect the SpaceX bid to include the cost of sacrificed 1st stage booster
- Biggest opportunities going forward to make even more informed bids:
  - Retrain the models with the whole dataset using the same hyperparameters previously defined
  - Incorporate additional launch data to the dataset and model as it becomes available

Appendix

- Notebook files to recreate this same project can be found on this github repository: https://github.com/JaimeSolisS/IBM-Applied-Data-Science-Capstone
  - Collect data:
    https://github.com/JaimeSolisS/IBM-Applied-Data-Science-Capstone/blob/main/01-Collecting_the_Data/01-spacex-data-collection-api.ipynb
    https://github.com/JaimeSolisS/IBM-Applied-Data-Science-Capstone/blob/main/01-Collecting_the_Data/02-webscraping.ipynb
  - Data Wrangling
    https://github.com/JaimeSolisS/IBM-Applied-Data-Science-Capstone/blob/main/02-Data_Wrangling/01-spacex-Data%20wrangling.ipynb
  - Exploratory Analysis
    https://github.com/JaimeSolisS/IBM-Applied-Data-Science-Capstone/blob/main/03-Exploratory_Analysis/01-eda-sql_sqllite.ipynb
    https://github.com/JaimeSolisS/IBM-Applied-Data-Science-Capstone/blob/main/03-Exploratory_Analysis/02-eda-dataviz.ipynb
  - Data Visualization
    https://github.com/JaimeSolisS/IBM-Applied-Data-Science-Capstone/blob/main/04-Interactive_Visual_Analytics_and_Dashboard/01-launch_visual_analytics_folium.ipynb
    https://github.com/JaimeSolisS/IBM-Applied-Data-Science-Capstone/blob/main/04-Interactive_Visual_Analytics_and_Dashboard/02-spacex_dash_app.py
  - Predictive Analysis
    https://github.com/JaimeSolisS/IBM-Applied-Data-Science-Capstone/blob/main/05-Predictive_Analysis_Classification/01-spacex_machine%20_learning_prediction.ipynb
- Illustrations from storyset.com
- Icons from slidesgo.com

Thank you!