IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dario Sammaruga
31/12/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using SpaceX API and web scraping from Wikipedia
  - Exploratory Data Analysis (EDA) with data visualization and Dashboard
  - Machine Learning
- Summary of all results
  - Data collection has been possible using free sources
  - EDA allowed to detect best feature to predict successful launches
  - Good predictions reached, so we can discuss about insights

Introduction

- SpaceX has been the first company capable to reuse first stage of rockets and this means less costs for every launch
- The target of the project is to investigate the possibility for SpaceY to enter in the space company market in competition with SpaceX
- To do it we need to:
  - Estimate total costs for launches after predicting successful landings for the first stage of SpaceX rockets
  - Understand if there are some launch sites that has better performance than others

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected with SpaceX API
  - Web Scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Data has been enriched after One Hot Encoding process and after creating a binary outcome label based on outcome categories provided.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Data has been queried and displayed in charts to select which feature should be used for predictive analysis and to understand first basic insights.

Methodology

Executive Summary

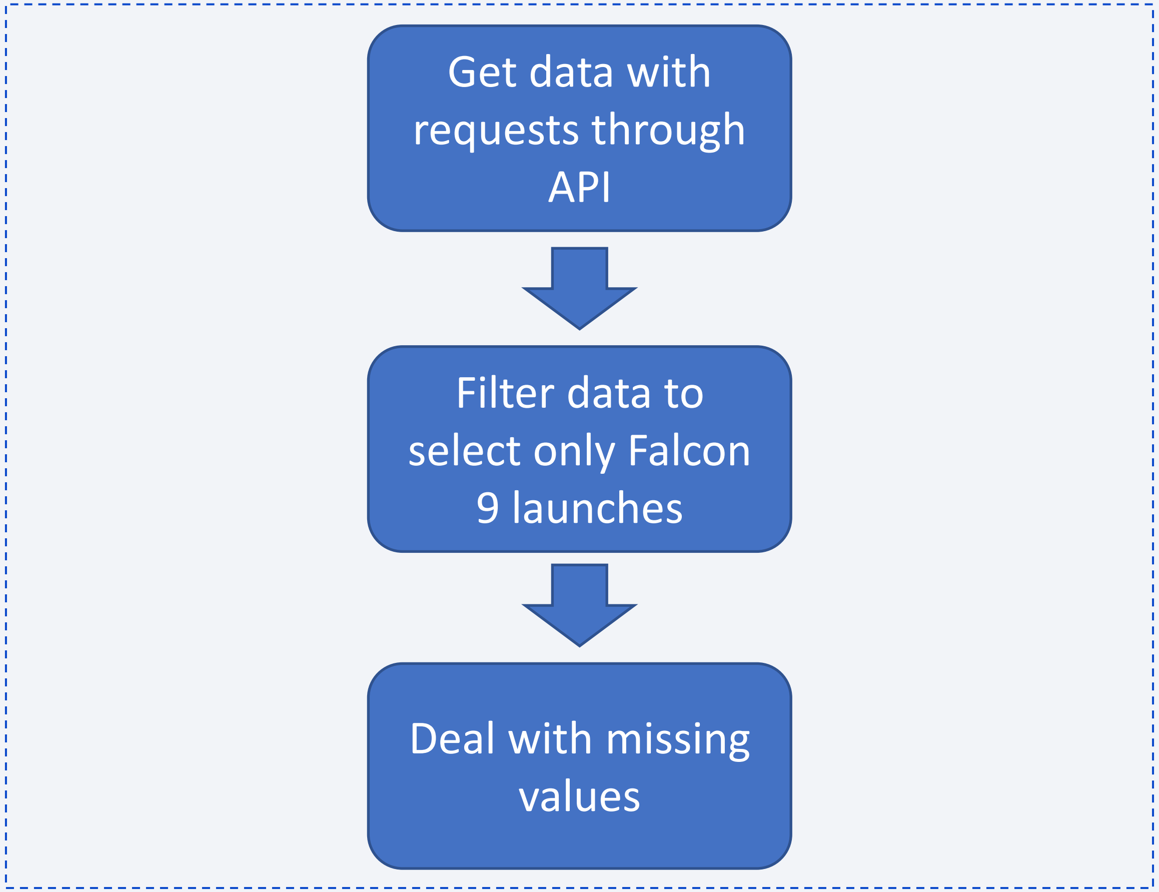
- Perform interactive visual analytics using Folium and Plotly Dash
  - Displaying Launch sites and distance from infrastructures like railways, highway, cities and coastline on a map has been possible to understand why some sites had less launches than others.
  - An interactive dashboard has been proposed to drill-down on data.
- Perform predictive analysis using classification models
  - Data has been normalized, then splitted in training and test datasets.
  - Four classification models has been evaluated on datasets.
  - Every model has been optimized with different combinations of parameters and finally I compared accuracy for each optimized model.

Data Collection

- Data has been collected with two methods:
    - SpaceX API (https://api.spacexdata.com/v4/rockets)
    - Web Scraping from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- We get data from SpaceX REST API using four endpoints:
  - https://api.spacexdata.com/v4/rockets/
  - https://api.spacexdata.com/v4/launchpads/
  - https://api.spacexdata.com/v4/payloads/
  - https://api.spacexdata.com/v4/cores/
- GITHUB URL:
  - https://github.com/BeanRepo/IBM-Data-Science-Capstone-Project/blob/main/week%201/jupyter-labs-spacex-data-collection-api.ipynb



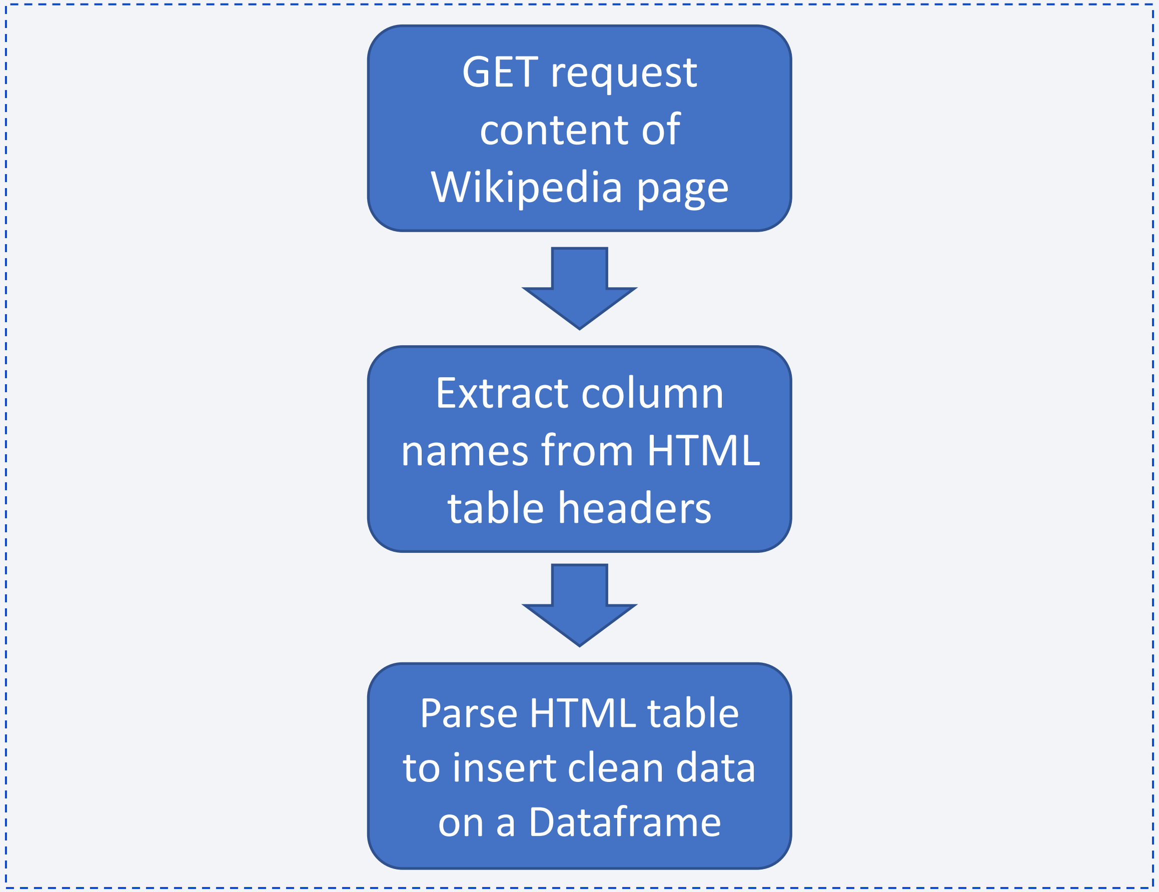
Get data with requests through API

Filter data to select only Falcon 9 launches

Deal with missing values

Data Collection - Scraping

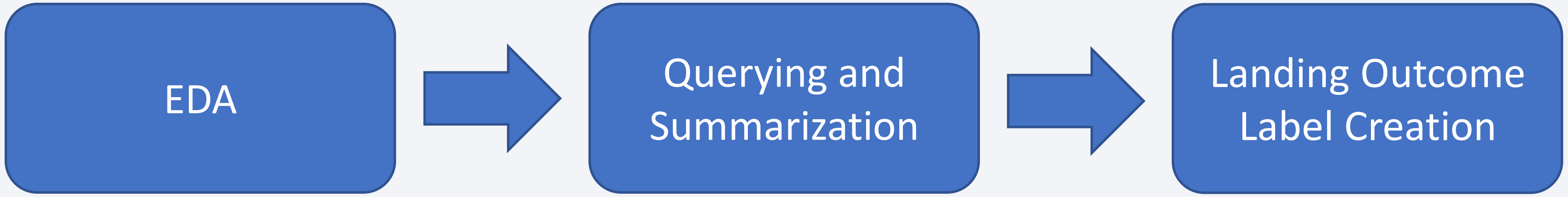
- Source code of Wikipedia website has been extracted to collect data in a table.
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- GITHUB URL:
  - https://github.com/BeanRepo/IBM-Data-Science-Capstone-Project/blob/main/week%201/jupyter-labs-webscraping.ipynb



GET request content of Wikipedia page
Extract column names from HTML table headers
Parse HTML table to insert clean data on a Dataframe

Data Wrangling

- Exploratory Data Analysis has been performed to understand:
  - KPI about launches, landing sites, success rate, payload wheight
  - Relationships between features
- Querying, Summarization, one-hot encoding processes has been done
- Landing Outcome binary label for classification has been created



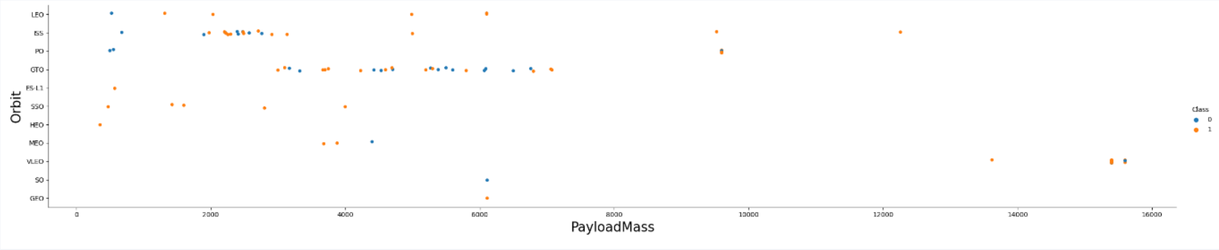
EDA

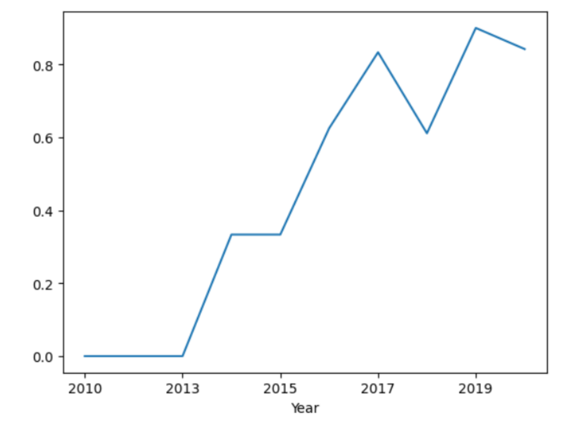
Querying and Summarization

Landing Outcome Label Creation

EDA with Data Visualization

- To better understand data, some scatter plots and bar plots has been created to understand relationship with couples of features:
  - Payload Mass and Flight Number
  - Launch Site and Flight Number
  - Launch Site and Payload Mass
  - Orbit and Flight Number
  - Orbit and success rate
  - Payload and Orbit
  - Success Rate trend over time





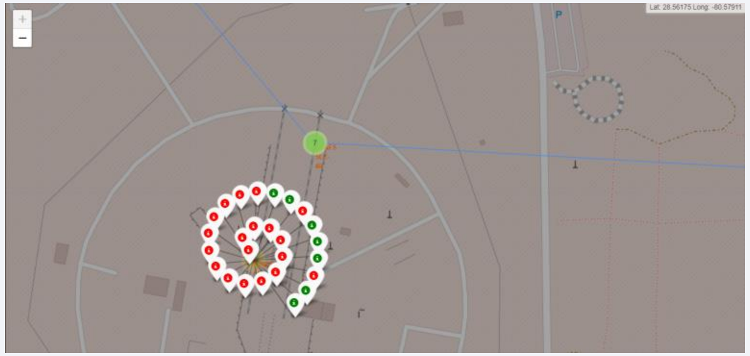
- GITHUB URL:
  - https://github.com/BeanRepo/IBM-Data-Science-Capstone-Project/blob/main/week%202/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries performed
  - how many launches came from launch sites,
  - average payload for every launch,
  - first successful landing date,
  - successful landings on ship,
  - Total number of successful and failure missions
  - Booster version with max payload carried
  - Resume of missions in 2015 by month
  - Landing outcomes between 04-06-2010 and 20-03-2017
- GITHUB URL:
  - https://github.com/BeanRepo/IBM-Data-Science-Capstone-Project/blob/main/week%202/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

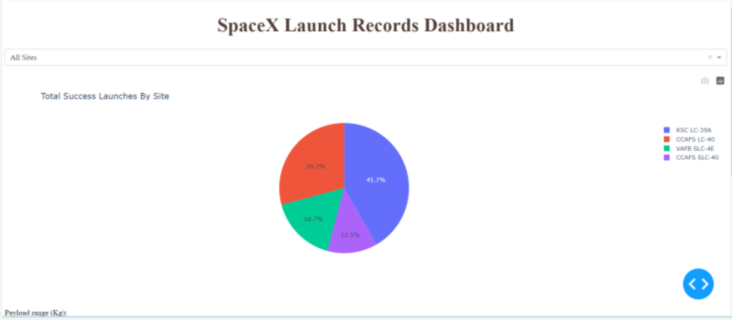
- markers, circles, lines has been added to a folium map, where:
  - Markers are displayed as launch sites
  - Circles are displayed as areas around launch sites
  - Marker clusters display binary landing outcomes for each launch site
  - Lines has been plotted between a launch site and 2 coordinates: closest seashore and closest railway



- GITHUB URL:
  - https://github.com/BeanRepo/IBM-Data-Science-Capstone-Project/blob/main/week%203/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Two graphs are displayed in order to visualize:
  - Percentage of launches by site
  - Payload range



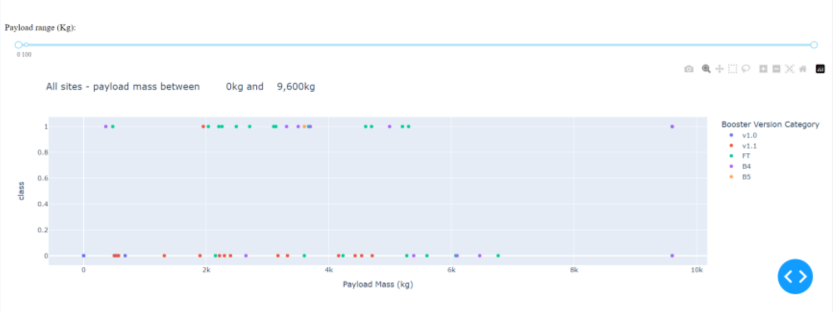
SpaceX Launch Records Dashboard

All Sites

Total Success Launches By Site

Payload range (Kg):

- Investigating relationships between payloads and launch sites has been fundamental to understand the best launch site based on payloads.



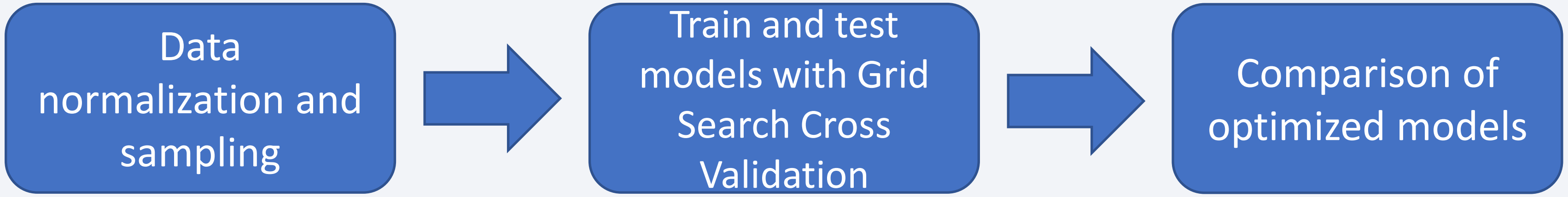
Payload range (Kg):

All sites - payload mass between 0kg and 9,600kg

- GITHUB URL:
  - https://github.com/BeanRepo/IBM-Data-Science-Capstone-Project/blob/main/week%203/lab_dash_app.py

Predictive Analysis (Classification)

- Has been performed a Data normalization and the dataset has been splitted in Train and Test sets.
- 4 Classification Methods has been deployed:
  - Logistic regression
  - Support Vector machines (SVM)
  - Decision Tree Classifier
  - K-Nearest-neighbour (KNN)
- For every model a different dictionary of parameters has been used to find the best ones using Grid-Search Cross validation and accuracy scores has been calculated on train and test sets
- GITHUB URL: https://github.com/BeanRepo/IBM-Data-Science-Capstone-Project/blob/main/week%204/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Data normalization and sampling

Train and test models with Grid Search Cross Validation

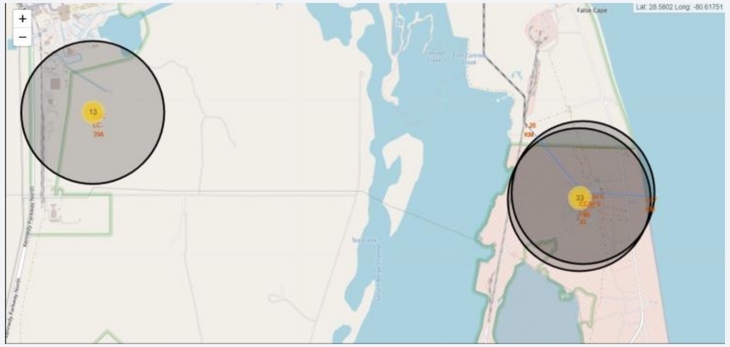
Comparison of optimized models

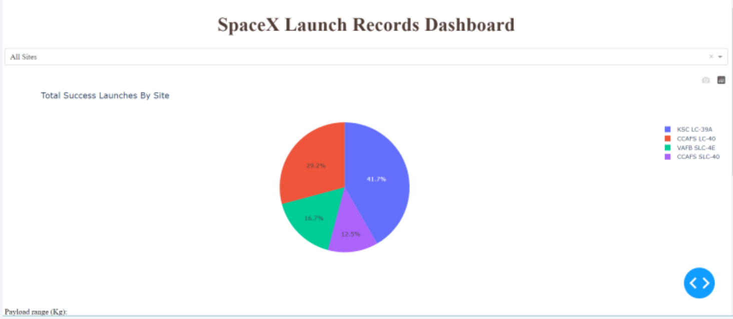
Results

- Exploratory data analysis results
    - Space x uses four launch sites
    - The first launch has been done by SpaceX and NASA
    - Average payload of F9 x1.1 booster is 2928 kg
    - The first successful landing has been in 2015 after five years of launches
    - Almost every mission reached his target. This is a different concept compared to landing outcome
    - Many booster version has successful landing outcome on drone ship with an above average payload
    - Only two booster versions failed at landing on drone ship in 2015
    - The success rate on landing increases after time

Results

- Interactive analytics

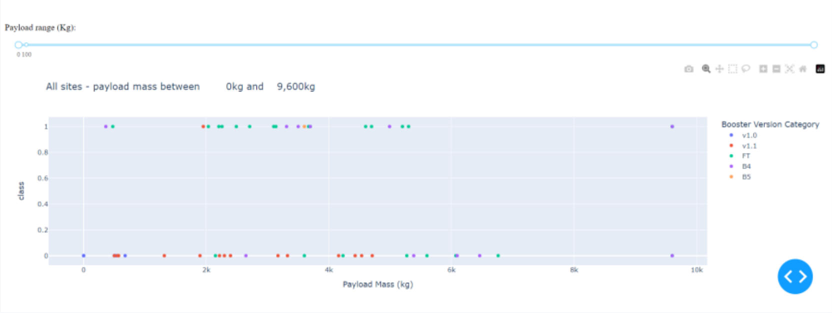



SpaceX Launch Records Dashboard

All Sites

Total Success Launches By Site

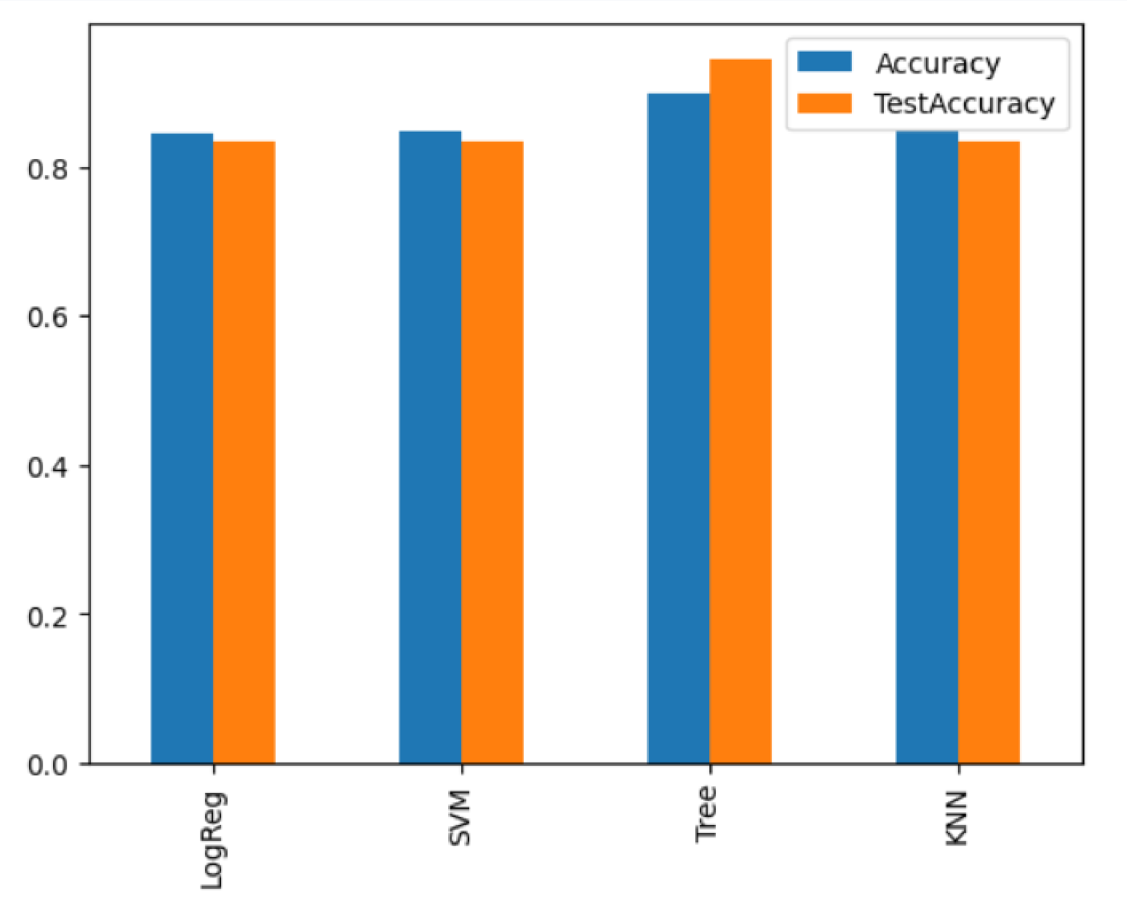
Payload range (Kg):


Payload range (Kg):

All sites - payload mass between 0kg and 9,600kg

Results

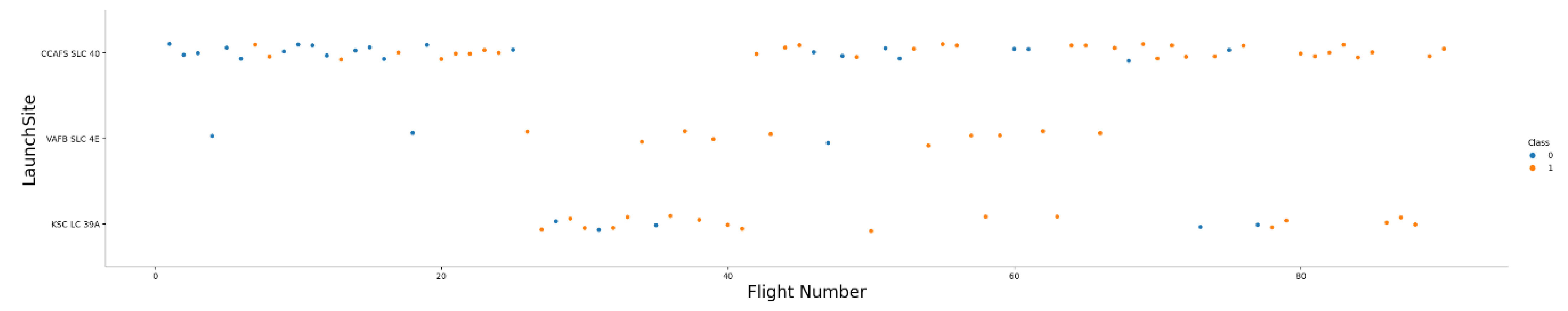
- Predictive analysis results
  - Decision tree classifier is the best model deployed
  - Train accuracy over 87%
  - Test accuracy over 94%



Section 2

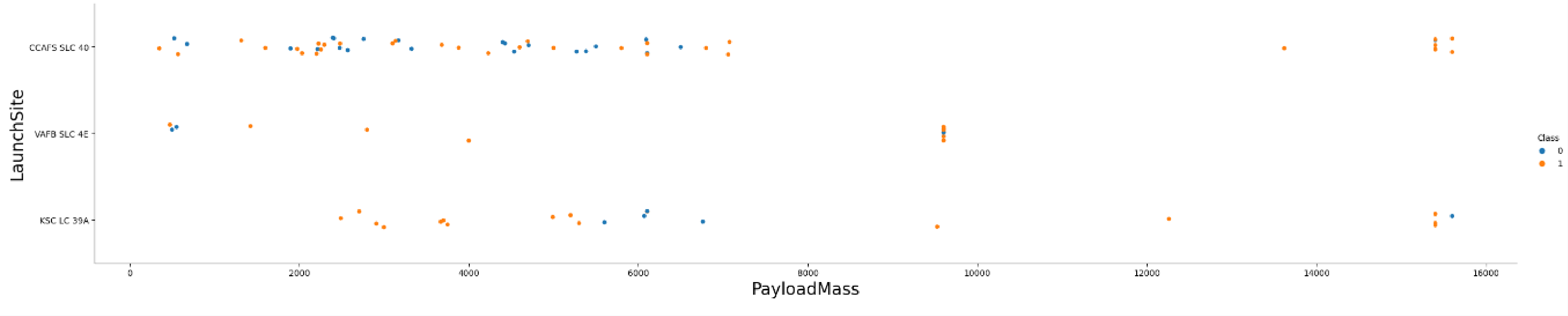
Insights drawn from EDA

Flight Number vs. Launch Site



- The best Launch site is CCAF5 SLC 40
- The second one is VAFB SLC 4E
- Orange dots represents successful landings and is it clear that success rate increases over time

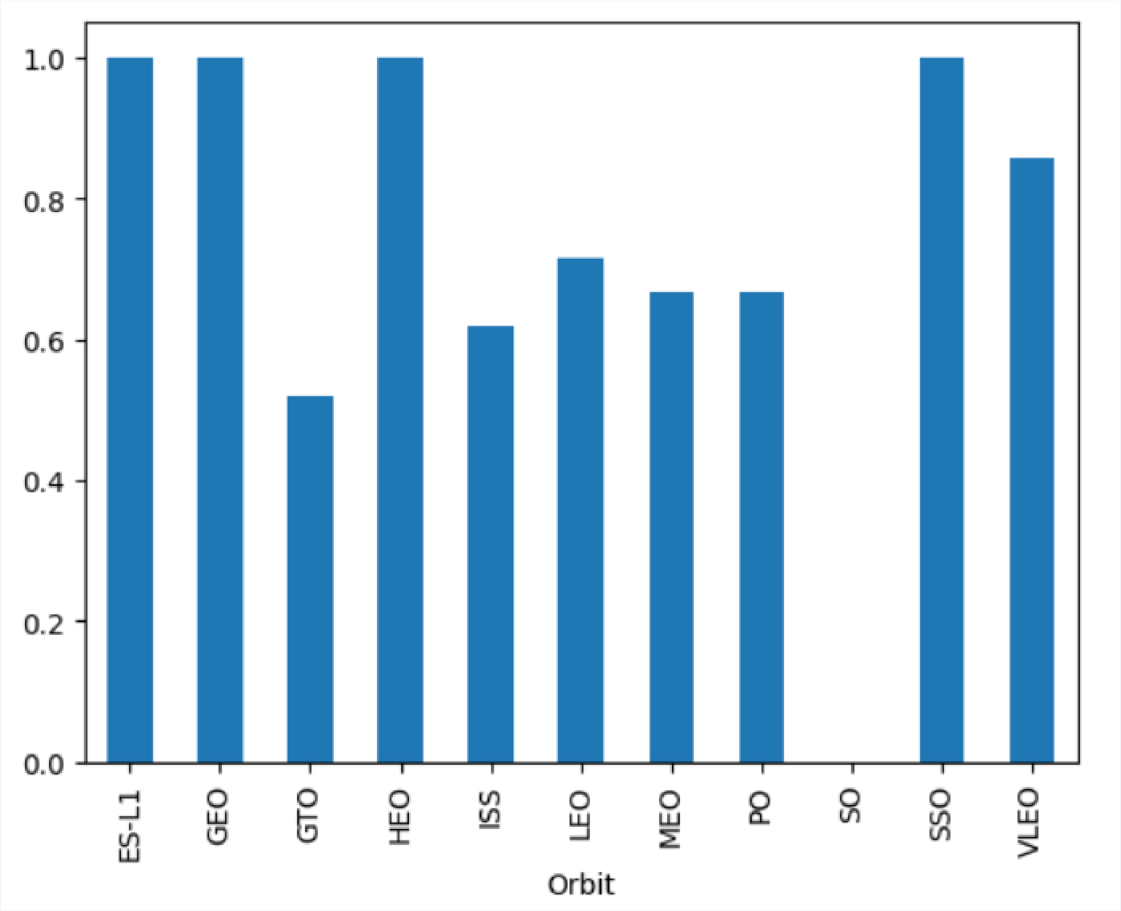
Payload vs. Launch Site



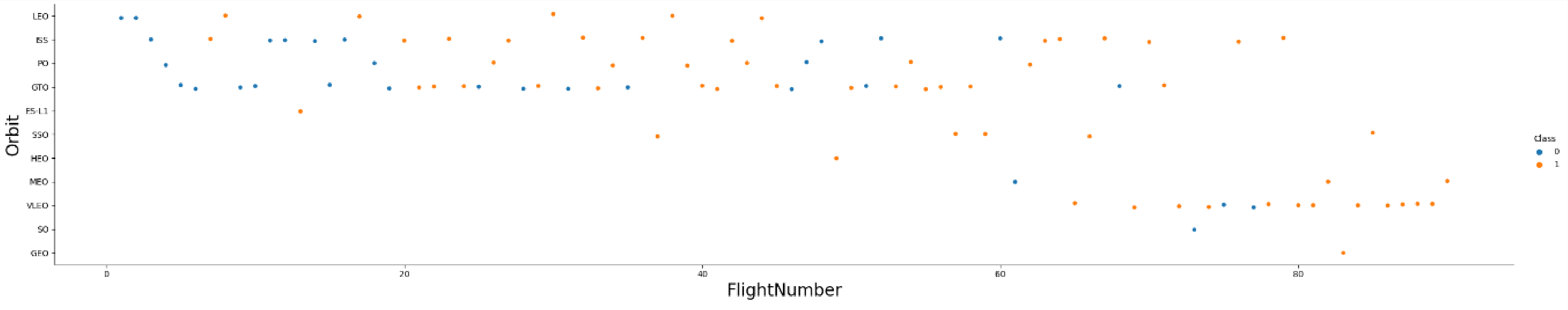
- Payload over 9000kg has success rate better than lower than 9000kg of payload launches but are less in number
- Payloads over 12000kg has been launched only on KSC LC 39A and on CCAFS SLC 40

Success Rate vs. Orbit Type

- Best success rates has been reached on:
  - ESL-1
  - GEO
  - HEO
  - SSO
- Also VLEO and LEO has high success rate, over 70%

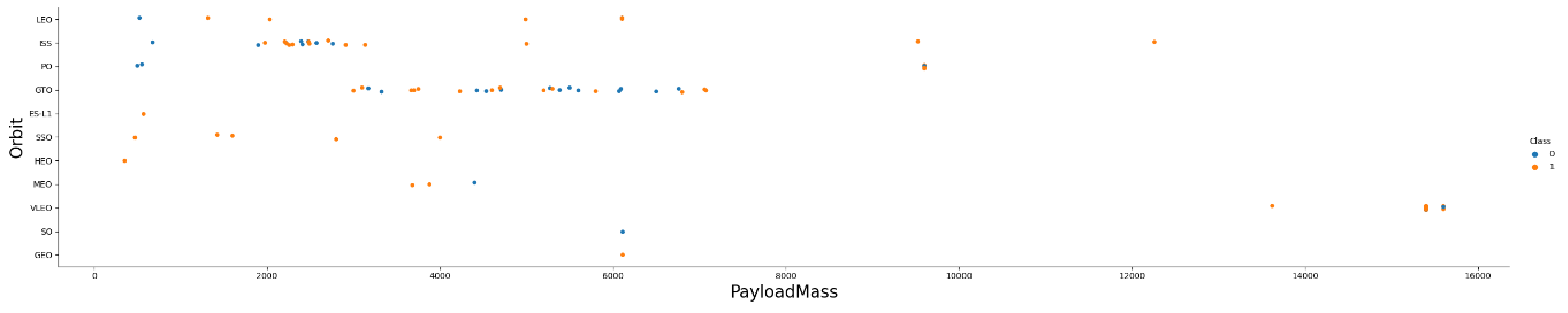


Flight Number vs. Orbit Type



- Success rate increases over time on all orbits
- VLEO orbit attempts are more recent. Is this a new opportunity or a more difficult orbit to reach that has been tested later due to its difficulty?

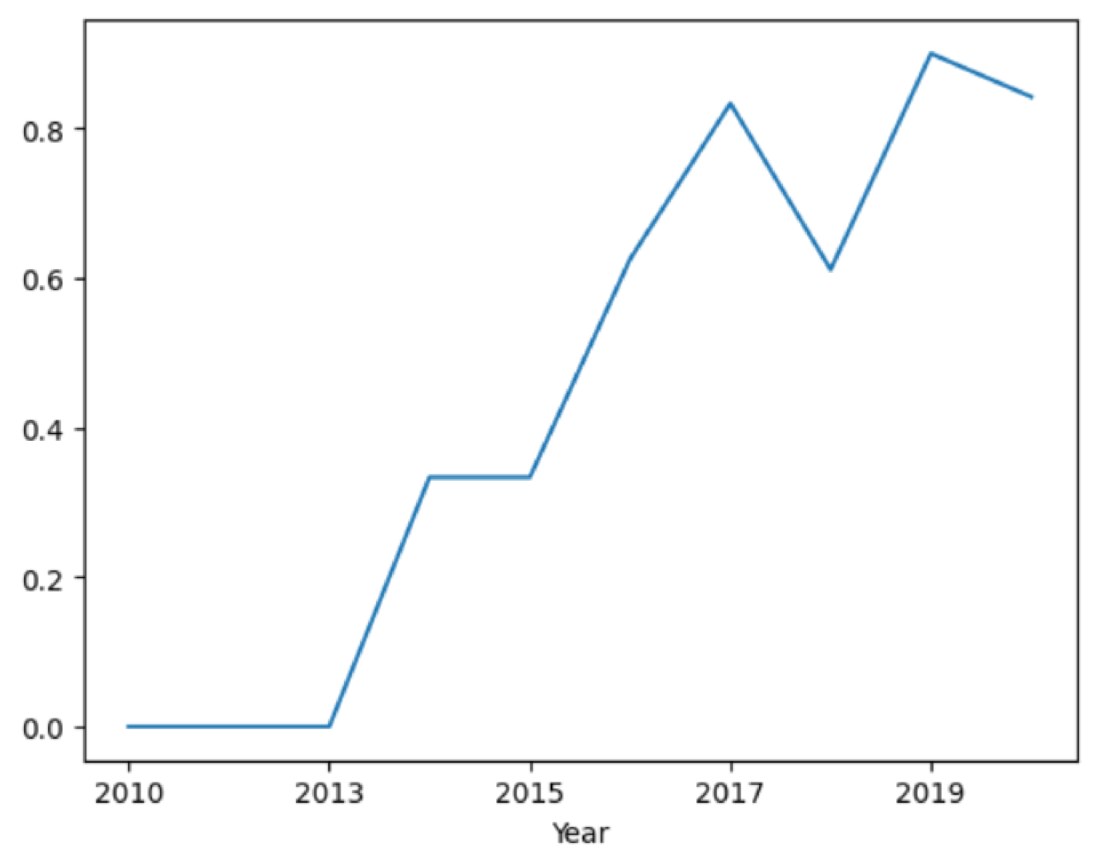
Payload vs. Orbit Type



- ISS obit has a wide range of payload launched and an overall good success rate
- No relation between payload and success rate on GTO orbit
- Too Few launches to orbits SO and GEO, is it not a business opportunity or a new one?

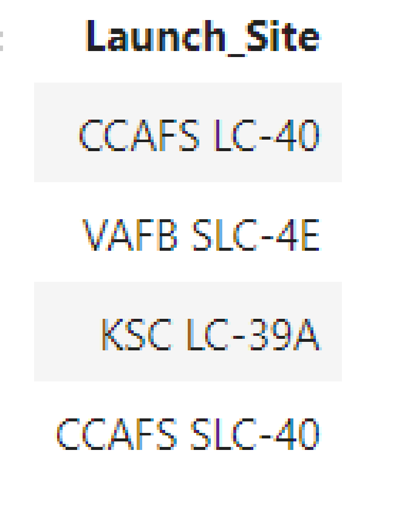
Launch Success Yearly Trend

- Success rate increases over time
- First successes comes after 2013
- Three years needed before first successful attempt



All Launch Site Names

- Here the only four Launch sites used by SpaceX

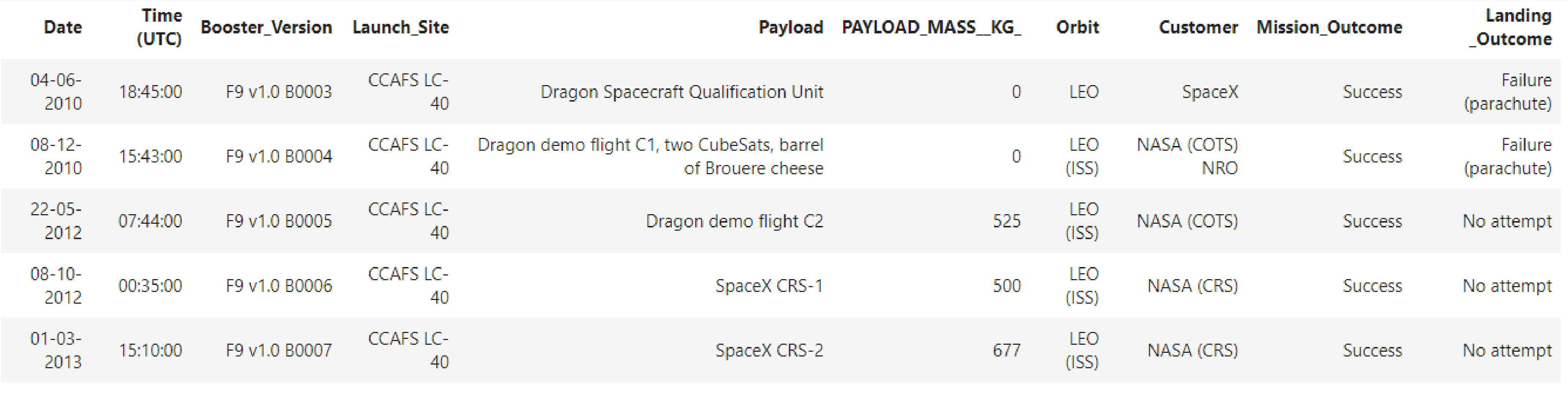


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- Calculated as distinct values of Launch Site field on the table

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

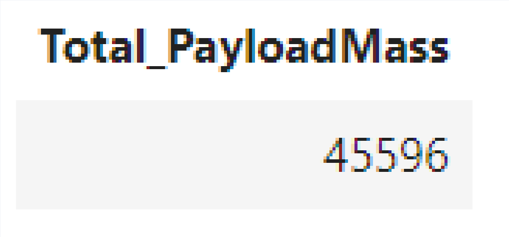


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
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
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Selecting top 5 records on the table where Launch Sites are located at Cape Canaveral

Total Payload Mass

- total payload carried by boosters from NASA

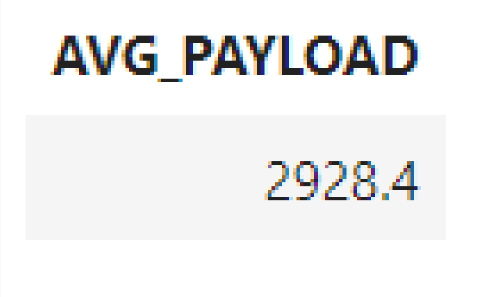


Total_PayloadMass
45596

- Sum of Payload Mass Kg where Customer is like NASA (CRS)

Average Payload Mass by F9 v1.1

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

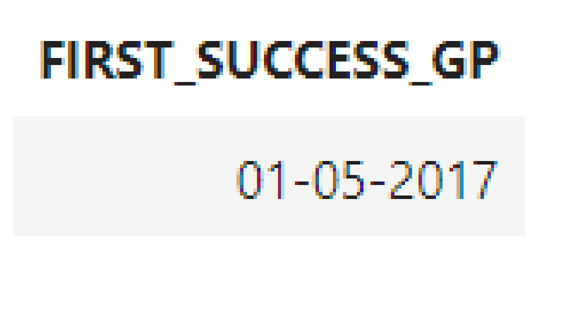


AVG_PAYLOAD
2928.4

- Calculating average of Payload Mass Kg field only where booster version is F9 v1.1

First Successful Ground Landing Date

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

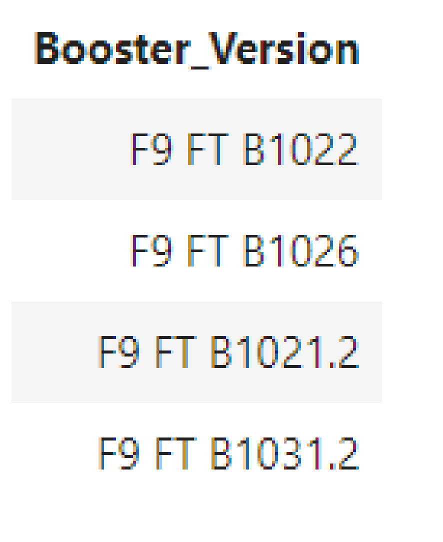


FIRST_SUCCESS_GP
01-05-2017

- Selecting minimum value of date from table where Landing Outcome is Success on Ground Pad

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

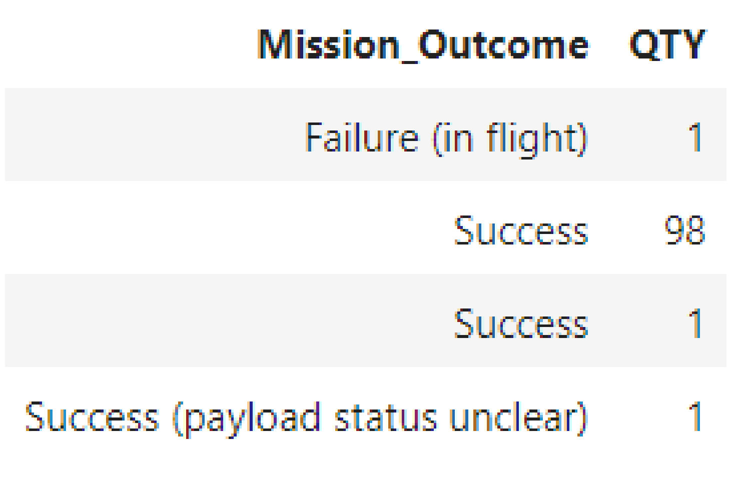


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Selecting distinct values of Booster version field where payload mass is between 4000 and 6000 and where Landing Outcome is Successful on drone ship

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

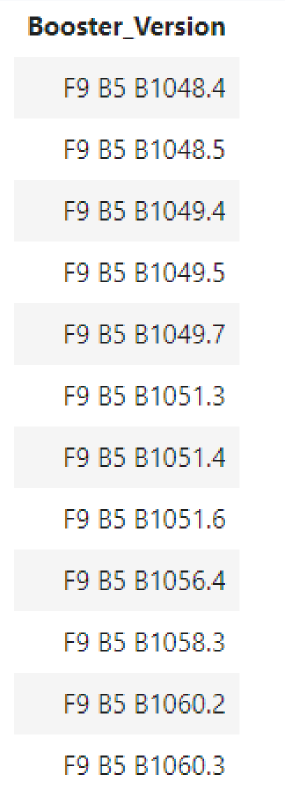


Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Count of Mission Outcomes grouped by distinct Mission Outcome

Boosters Carried Maximum Payload

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

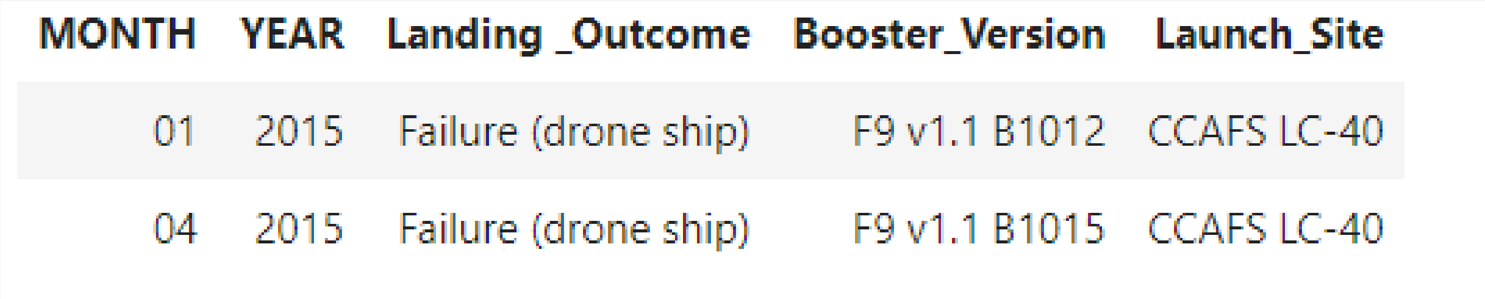


Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- Selecting distinct Booster versions with maximum payload mass on the table

2015 Launch Records

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

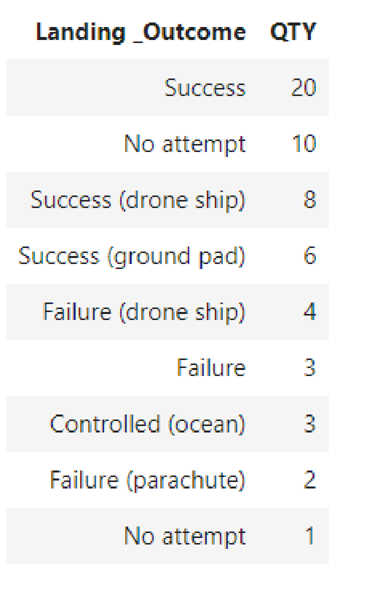


MONTH	YEAR	Landing _Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Selecting landing outcomes, booster versions and launch site for year 2015 by month only where Landing Outcome is Failure on Drone Ship

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



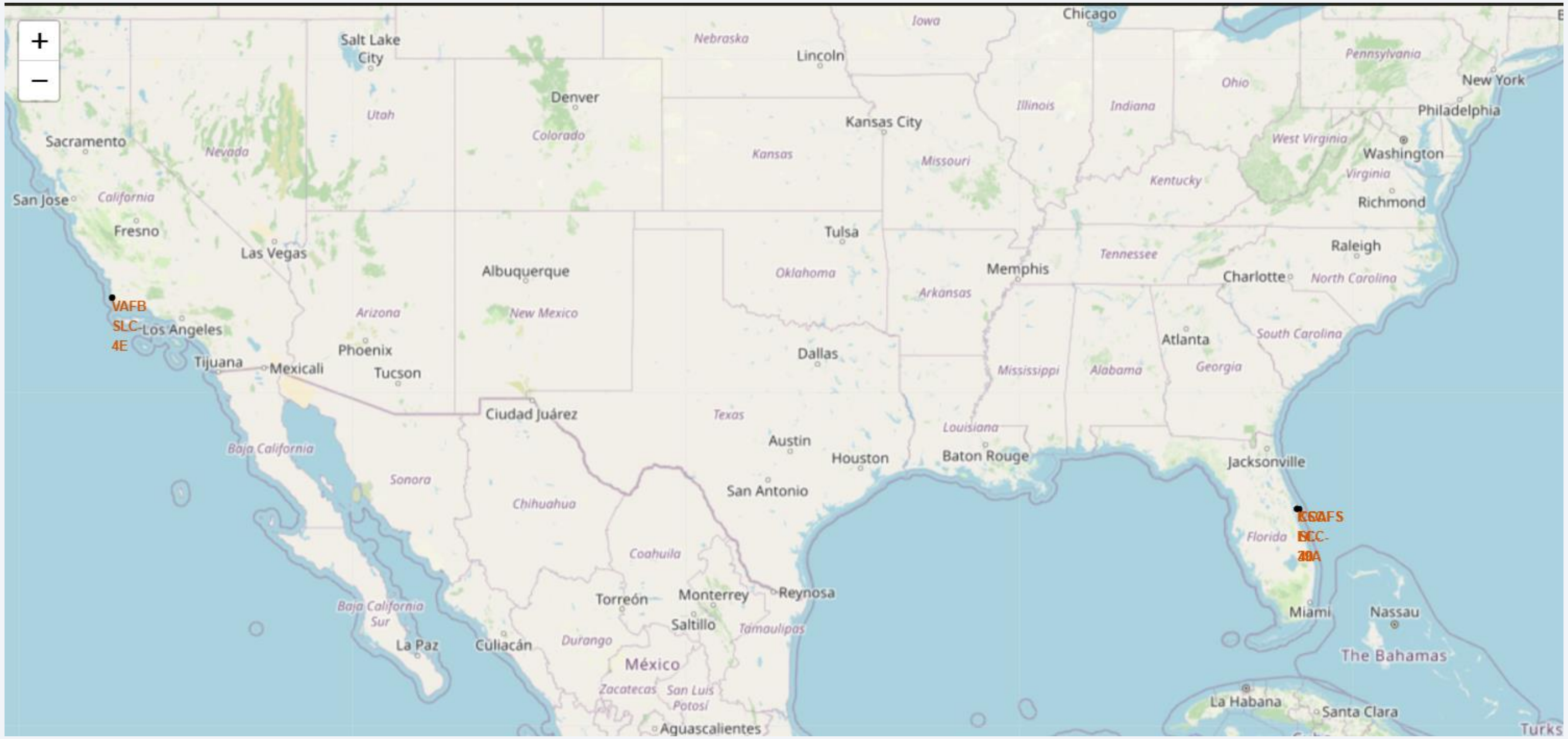
Landing _Outcome	QTY
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

- Count of distinct Landing outcomes where date is between 2010-06-04 and 2017-03-20 and ordered by quantity in descending order

Section 3

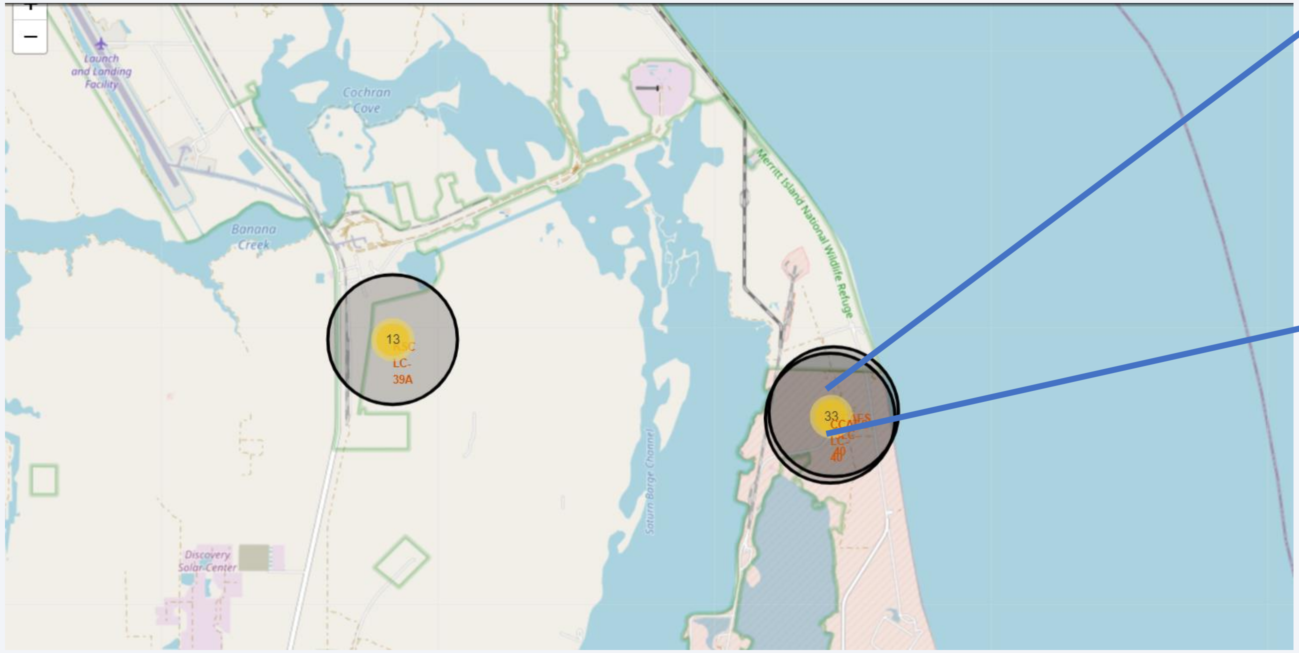
Launch Sites Proximities Analysis

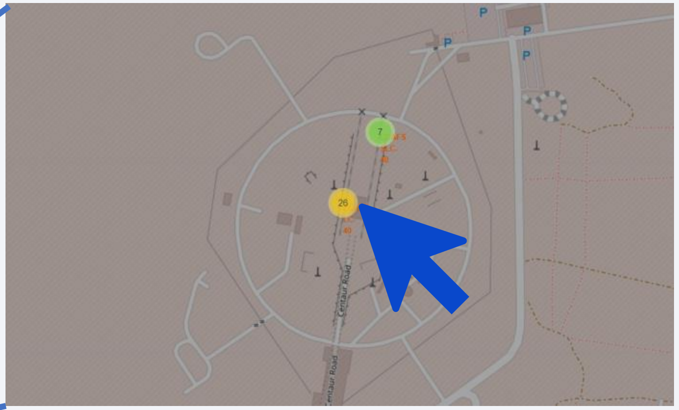
SpaceX Launch Sites

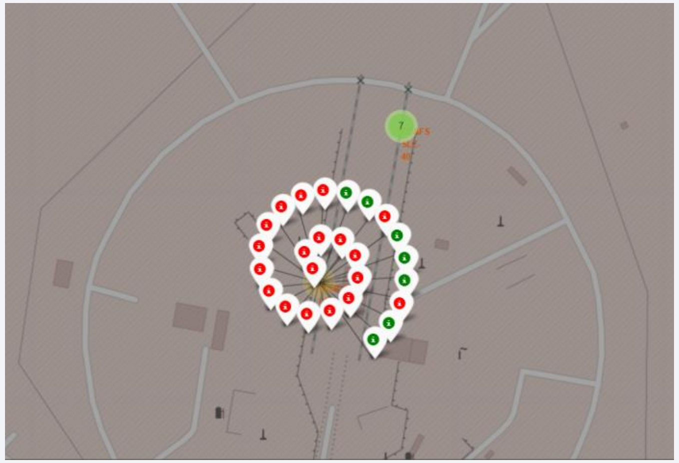


- Launch sites are on seashores for safety reason and they are also close to infrastructures like railways and highways to allow supply chain close connections

Launch Outcomes by Launch Site





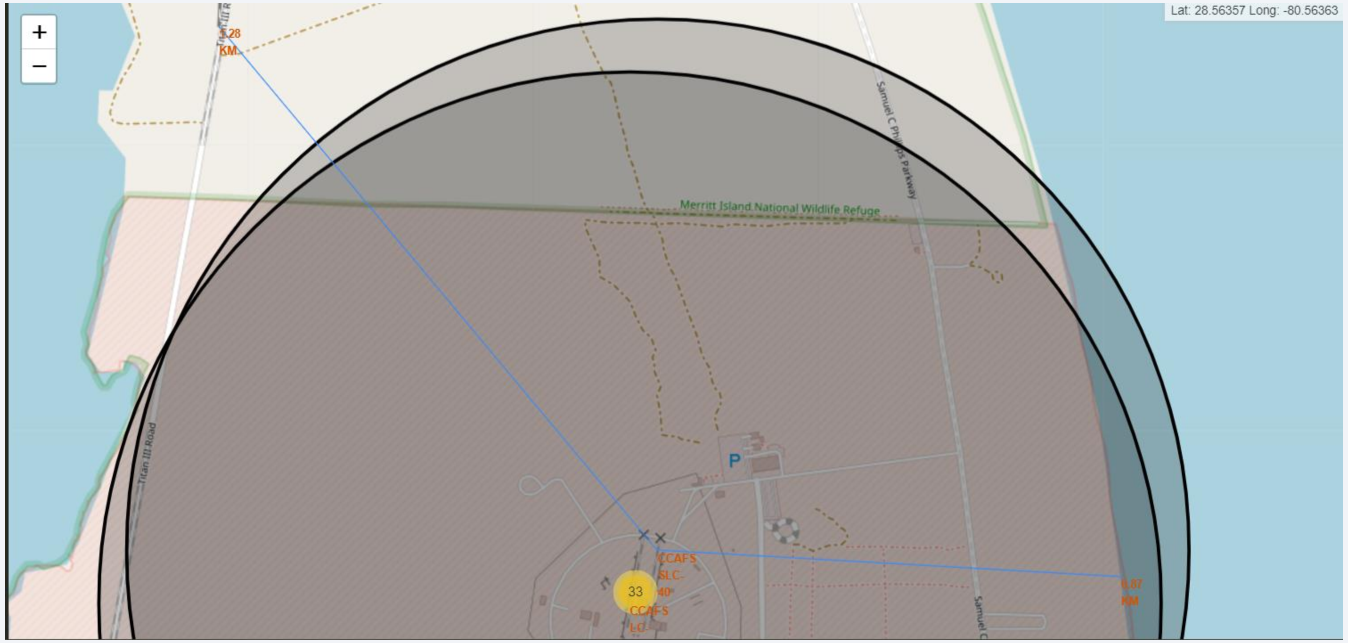


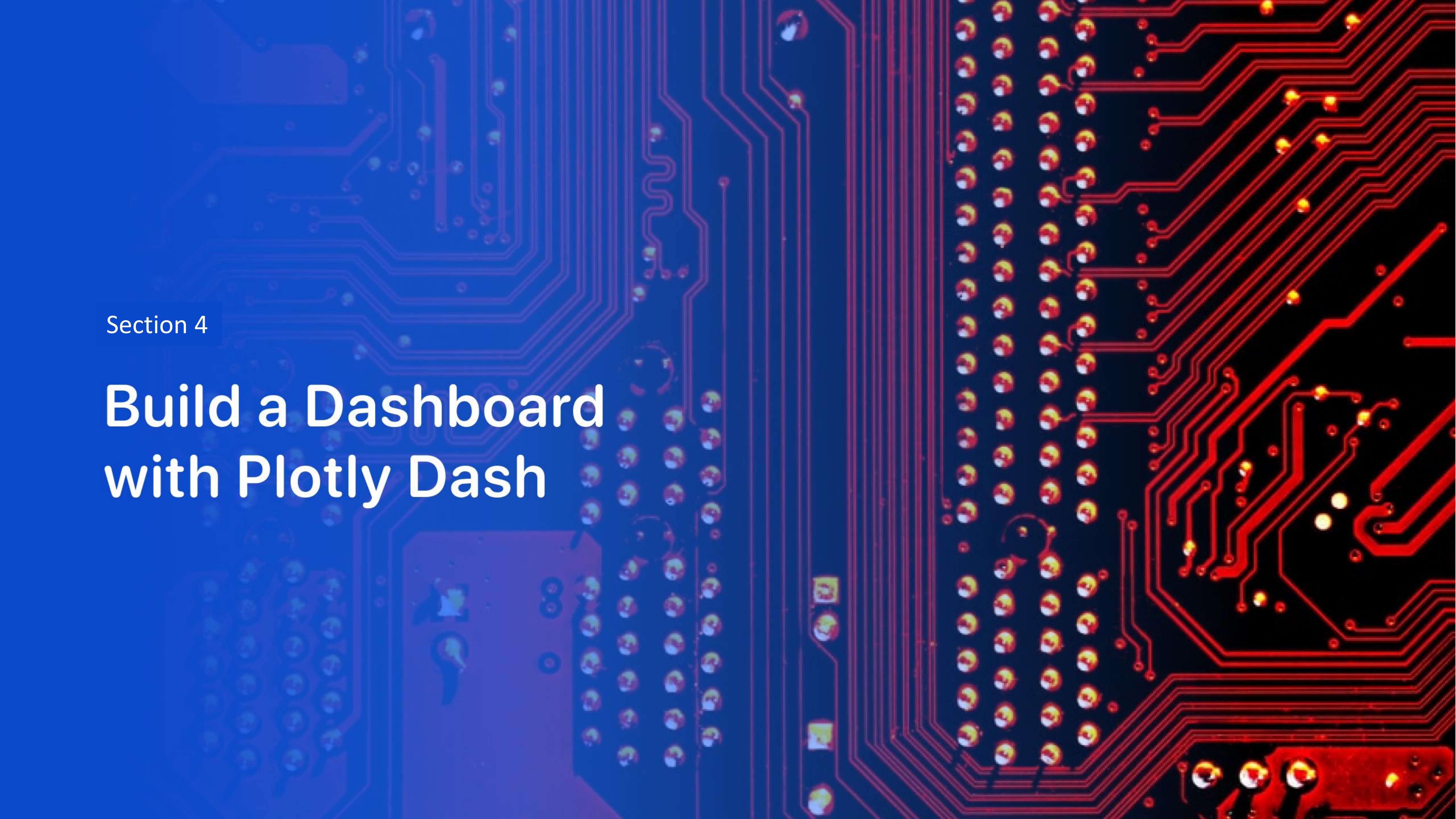
- Zooming on a launch site we can click on it
- Then we visualize successful attempts in green and failures in red

Safety and infrastructures

CCAFS SLC 40 launch site:

- is 0.87Km far from the seashore
- is only 1.28Km far from railways
- is far from the first high density city


Lat: 28.56357 Long: -80.56363
+
−
1.28 KM
Titan III Road
Merritt Island National Wildlife Refuge
Samuel C Phillips Parkway
P
CCAFS SLC-40
33
CCAFS LC-
0.87 KM
Samuel C

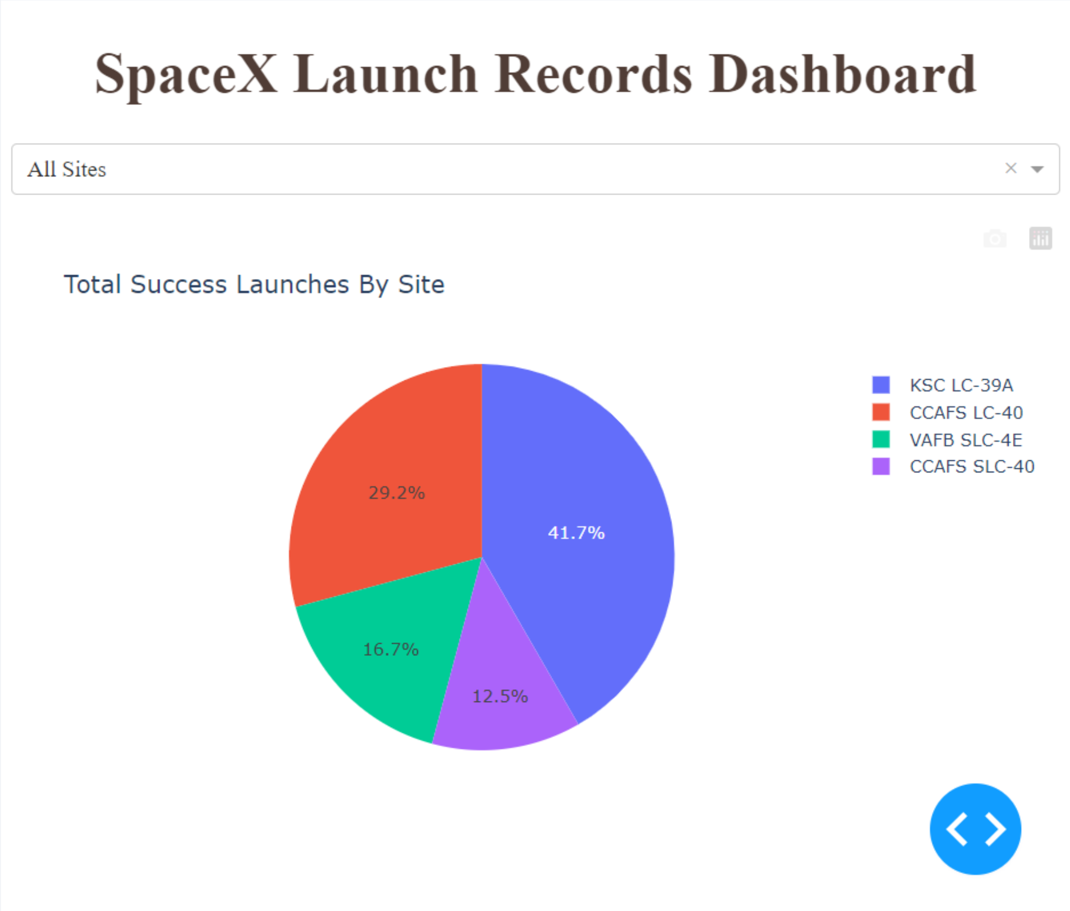


Section 4

Build a Dashboard with Plotly Dash

Successful Launches by sites

- Most successful launches comes from Cape Canaveral

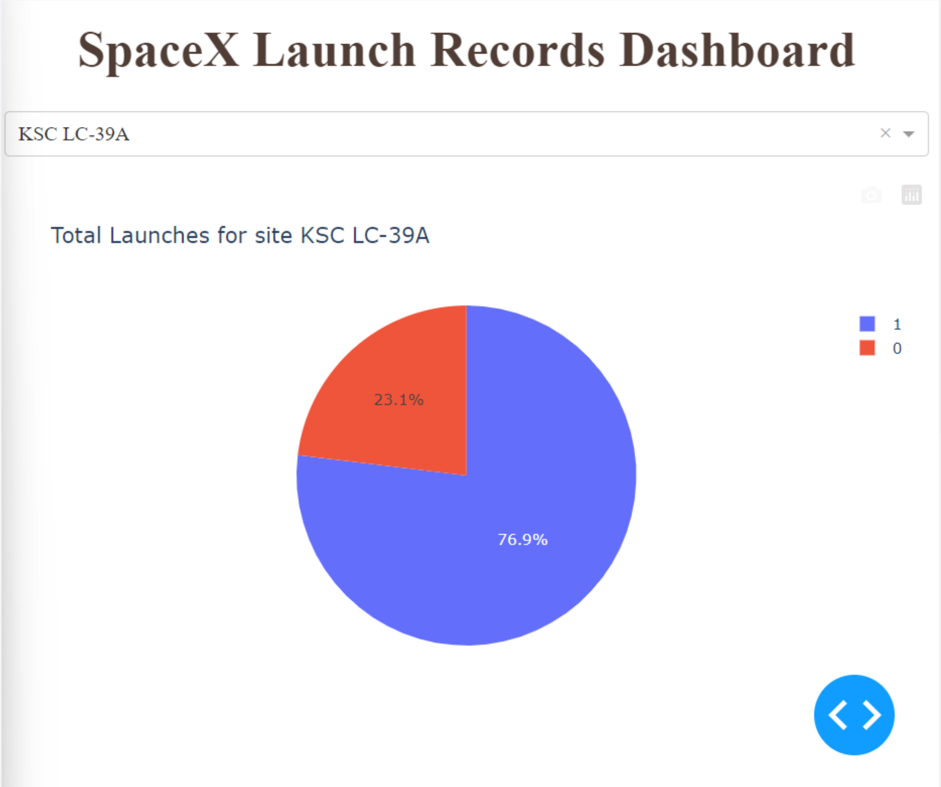


SpaceX Launch Records Dashboard

All Sites

Total Success Launches By Site

Success Rate for a Launch site



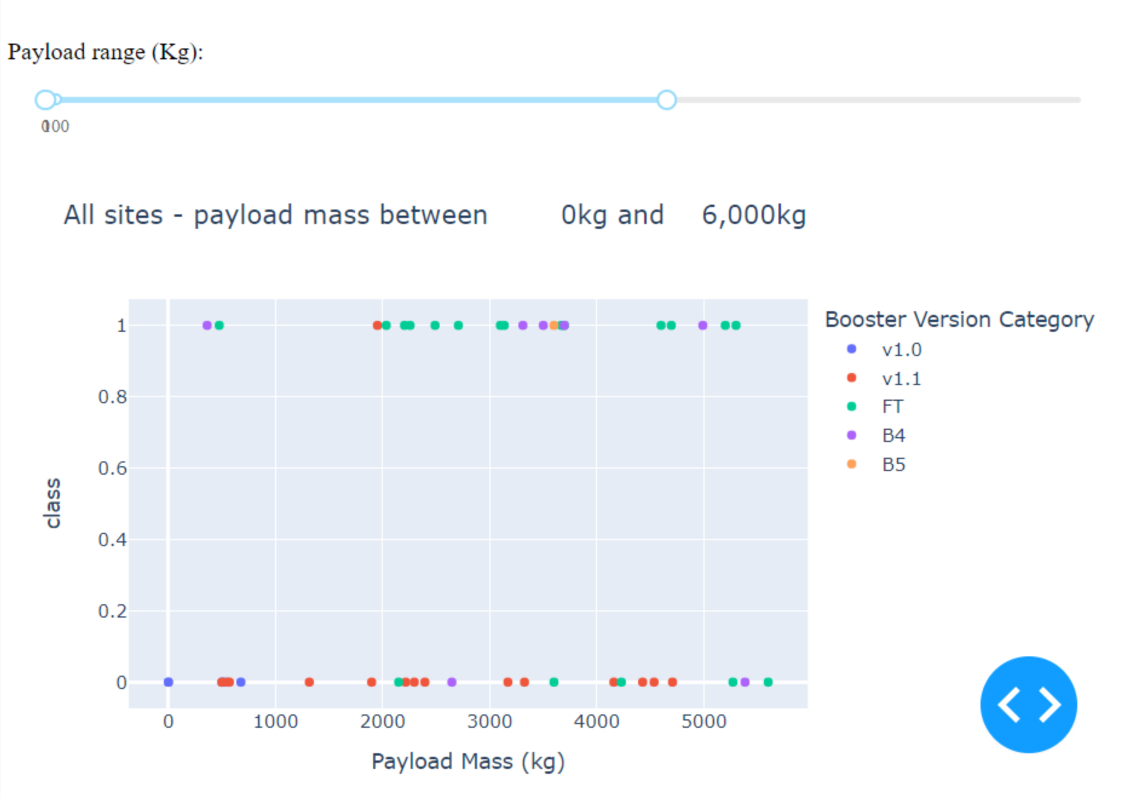
SpaceX Launch Records Dashboard

KSC LC-39A

Total Launches for site KSC LC-39A

- Drilling down to KSC LC-39A we discover that it has the best Success Rate

Payload and Launch Outcome



Payload range (Kg):

All sites - payload mass between 0kg and 6,000kg

- After some filtering and drill downs it is clear that the combination of payload under 6000Kg and FT boosters delivers the best combination for successful attempts.

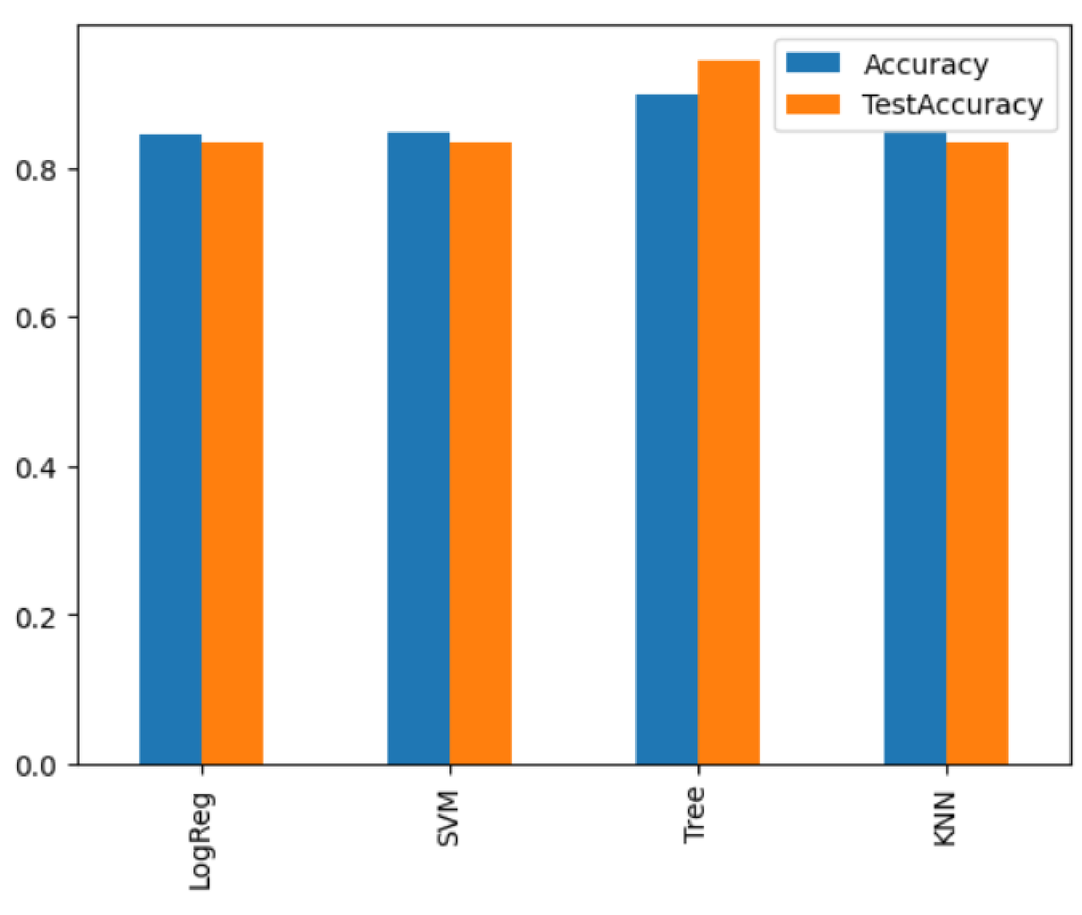
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- On the right we can see accuracies for the classification models deployed

- Decision Tree Classifier has performed better than other models with accuracies over 87%



Confusion Matrix

Confusion matrix for the best model

- TRUE POSITIVE

  The model predicted correctly 11 successful attempts on 12 total successful attempts

- TRUE NEGATIVE

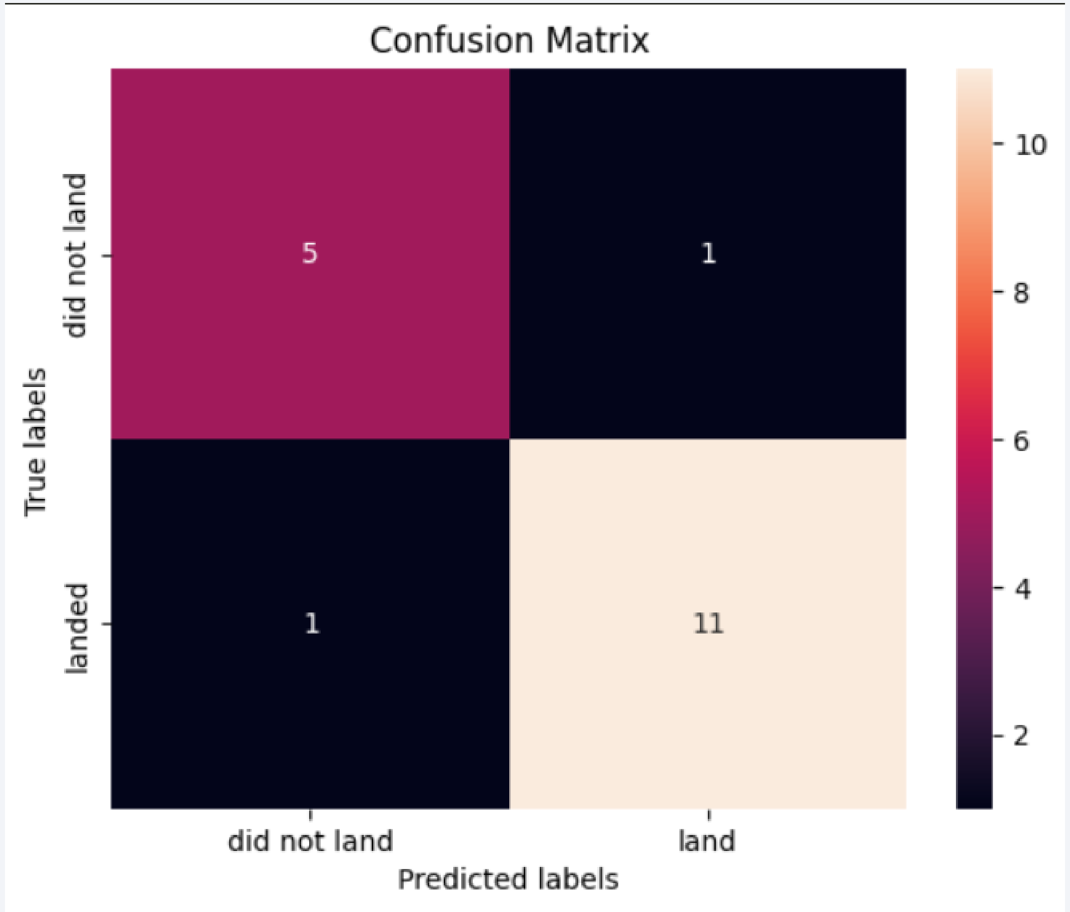
  The model predicted correctly 5 failures on 6 total failures

- FALSE POSITIVE

  1 failure has been incorrectly classified as successful attempt

- FALSE NEGATIVE

  1 successful attempt has been incorrectly classified as failure



Conclusions

- The best launch site is KSC LC-39A, followed by CCAFS SLC 40
- Launches with payloads over 7000Kg has higher success rate
- To reach successful mission it takes approximately 3 years of tests and to reach first successful landing it takes approximately 5 years from start. After this period the success rate increases.
- VLEO orbit launches seems to be a new market opportunity with improvement margins
- The classification model can be useful to predict successful landings in order to estimate costs.

Appendix

- Folium maps are not available when opening notebook on GitHub, please consider to download them.

Thank you!