IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Deny Rahmalianto
September 4, 2022

Outline

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

Executive Summary

- The methodologies that are used in the presentation:
  - Data collection using API and Webscrapping
  - Preprocessing
  - Exploratory data analysis through data visualization
  - Machine learning prediction
- Summary of all results
  - Taking time to preprocess and learn about the data is extremely crucial in order to make a good model.

Introduction

- The project is a part of coursera course for Applied Data Science Capstone. In this project we learn the end to end of data scientist steps, from collecting data, preprocessing the data, exploring the data, creating models, as well as making final presentation.
- This project would like to answers what are the steps that are needed in making data science analysis.

Section 1

Methodology

Methodology

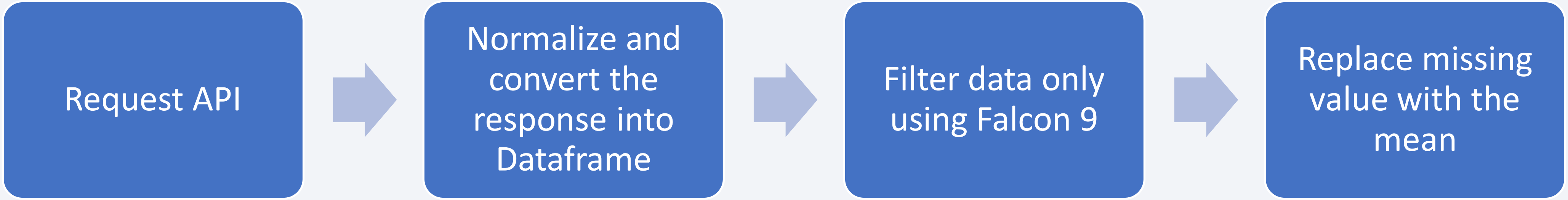
Executive Summary

- Data collection methodology:
  - Data is collected from Space X using API as well as web scrapping
- Perform data wrangling
  - Convert categorical data into numeric
  - Replacing null data with the mean
- 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 were split into train and test to avoid bias
  - Several machine learning method were compared in order to find the best model
  - Several parameters were compared at hyperparameter tuning in order to find the best parameter

Data Collection

- Data were collected using 2 ways:
  - API from: https://api.spacexdata.com/v4/payloads/
  - Web Scrapping from: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API



Request API

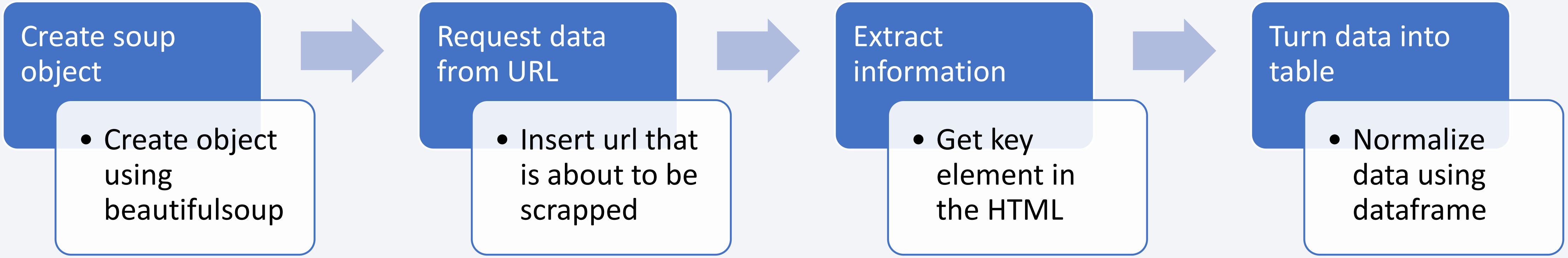
Normalize and convert the response into Dataframe

Filter data only using Falcon 9

Replace missing value with the mean

Source code: https://github.com/rahmalianto/Coursera-Applied-Data-Science-Capstone/blob/master/Data%20Collection%20API.ipynb

Data Collection - Scraping



Create soup object

- Create object using beautifulsoup

Request data from URL

- Insert url that is about to be scrapped

Extract information

- Get key element in the HTML

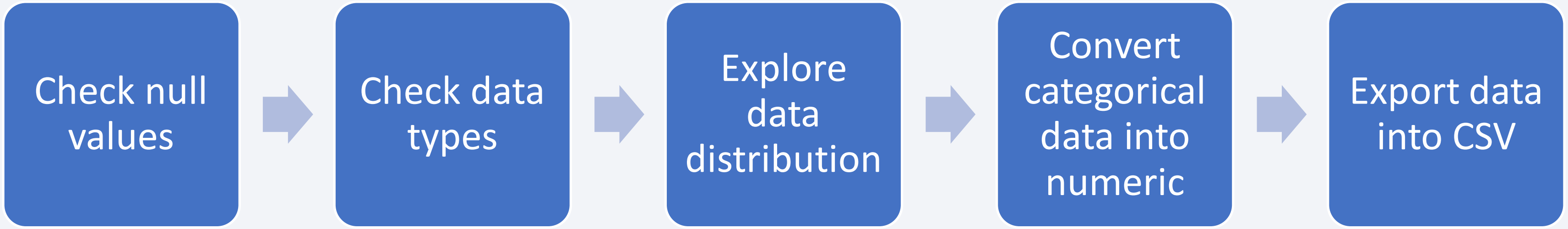
Turn data into table

- Normalize data using dataframe

Github Link: https://github.com/rahmalianto/Coursera-Applied-Data-Science-Capstone/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- First, check null values in every columns
- Second check whether data types is already as expected
- Third learn the data distribution
- Convert data categorical into numeric
- Export the data


Check null values

Check data types

Explore data distribution

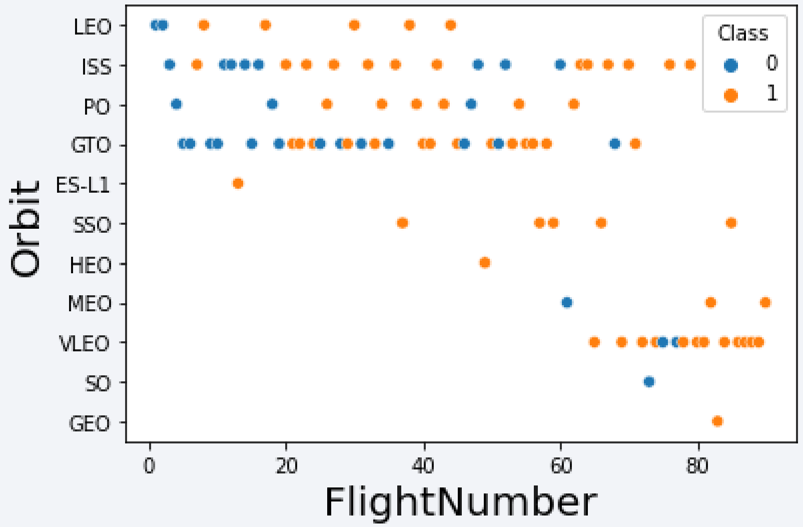
Convert categorical data into numeric

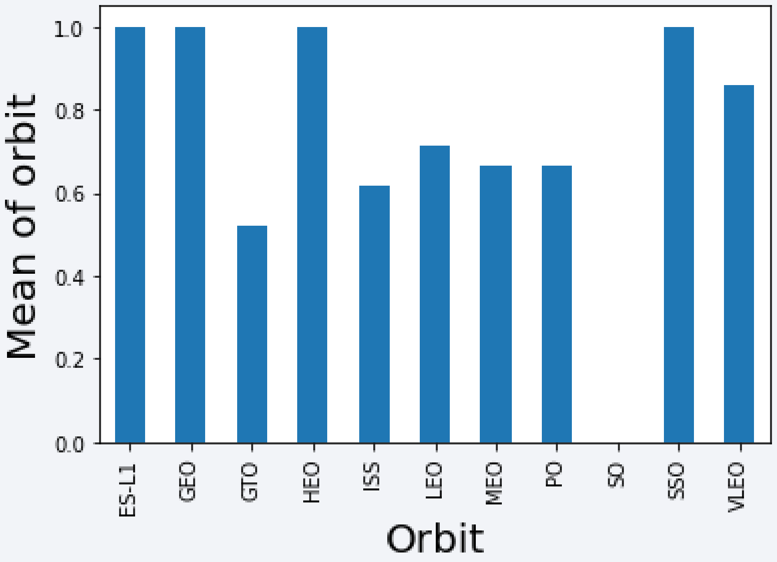
Export data into CSV

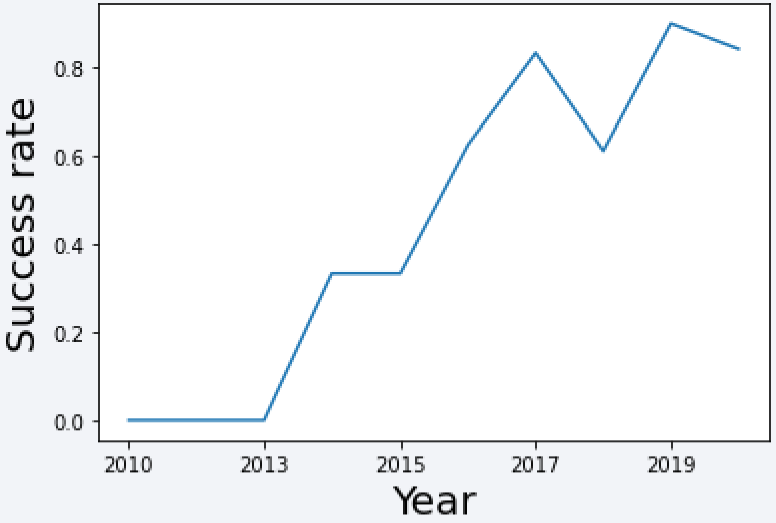
Source code: https://github.com/rahmalianto/Coursera-Applied-Data-Science-Capstone/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

- Visualize the relationship between variables using scatter plot
- Check the success rate of each orbit type
- Check the success rate trend annually







Source code: https://github.com/rahmalianto/Coursera-Applied-Data-Science-Capstone/blob/master/EDA%20with%20Visualization.ipynb

EDA with SQL

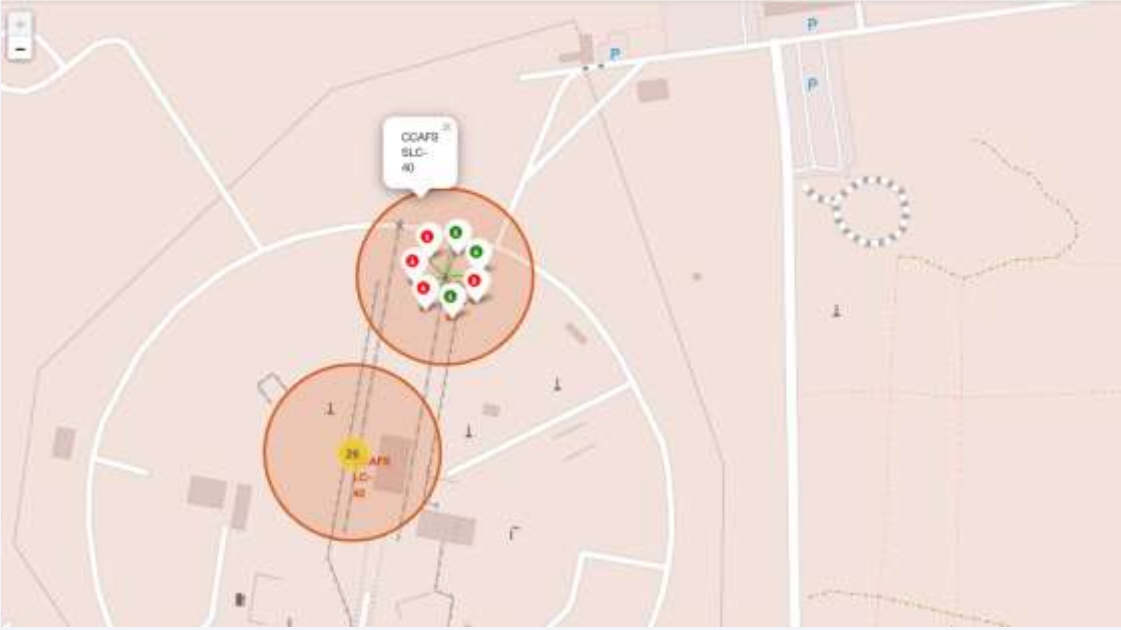
Here are the list of tasks perform during the course:

- 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 acheived.
- 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. Use a subquery
- 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

Source code: https://github.com/rahmalianto/Coursera-Applied-Data-Science-Capstone/blob/master/EDA%20with%20SQL%20Lab.ipynb

Build an Interactive Map with Folium

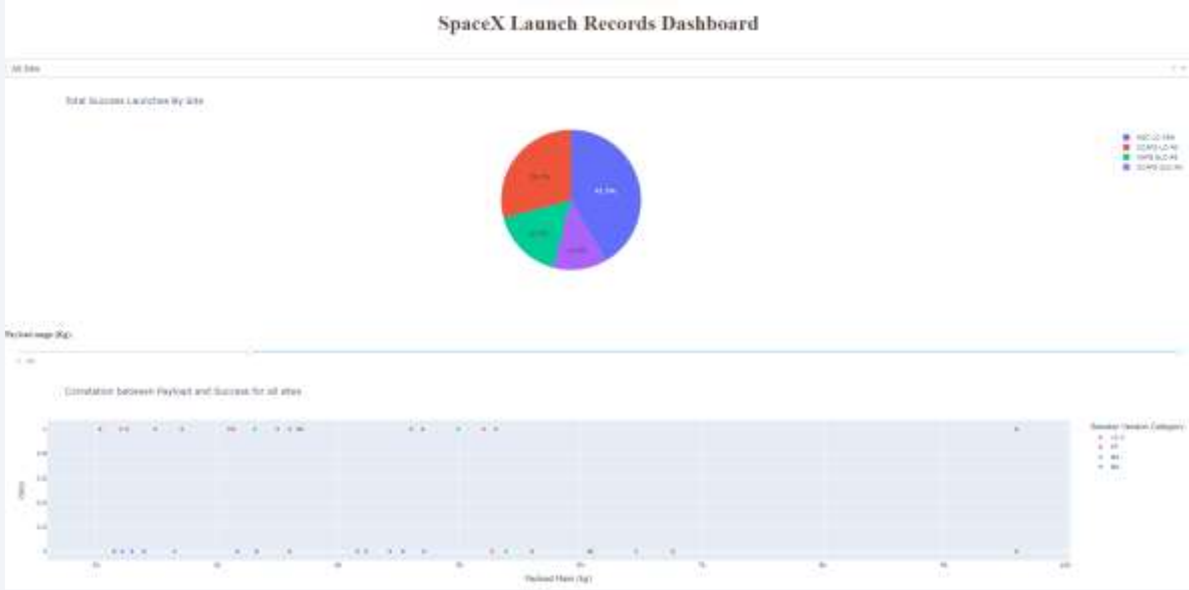
- Markers are used to mark launch sites
- Circles are used to highlight areas;
- Marker clusters are used to groups several markers
- Lines are used to annotate distances between two points.



Source code: https://github.com/rahmalianto/Coursera-Applied-Data-Science-Capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Pie chart is used to show proportion
- Scatter plot is used to show the relationship between variables

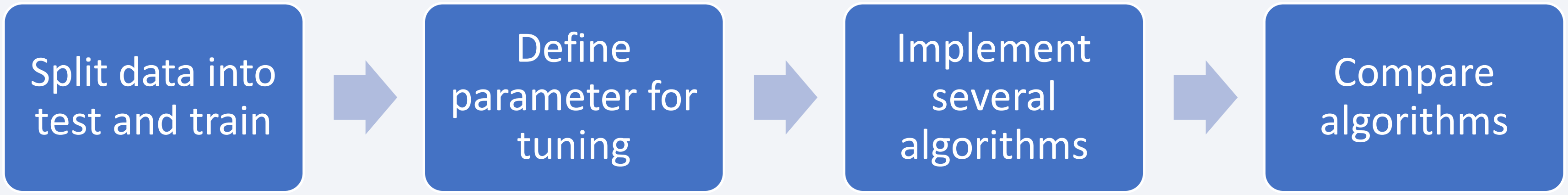


SpaceX Launch Records Dashboard

Source code: https://github.com/rahmalianto/Coursera-Applied-Data-Science-Capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- First we split dataset into train and test
- Define parameters for hyperparameter tuning
- Apply to several algorithms
- Compare which algorithms and parameters perform best



Split data into test and train

Define parameter for tuning

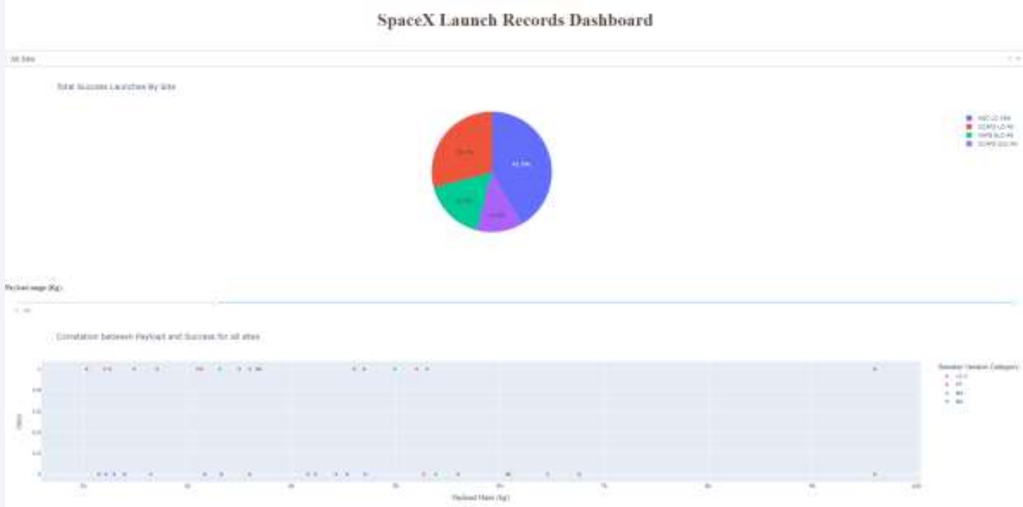
Implement several algorithms

Compare algorithms

Source code: https://github.com/rahmalianto/Coursera-Applied-Data-Science-Capstone/blob/master/Predictive%20Analysis.ipynb

Results

- Payload mass affects the probability of launch success
- Certain orbit type has a better success rate
- Decision tree perform better in predicting success/failure of a rocket launch

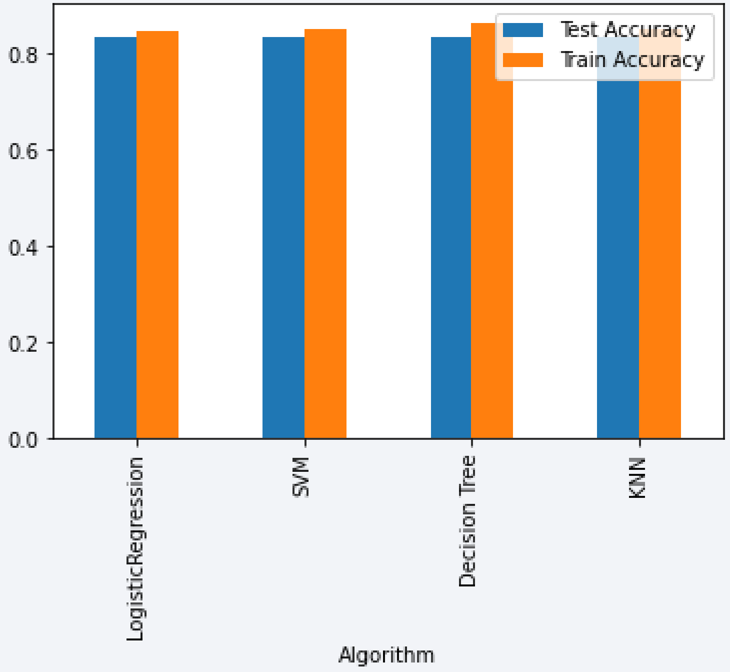

SpaceX Launch Records Dashboard
All Sites
Total Success Launches By Site
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
Correlation between Payload and Success for all sites
Booster Version Category
Payload Mass (kg)

Section 2

Predictive Analysis (Classification)

Classification Accuracy

- For test accuracy, almost all model perform the same, however for train accuracy, **decision tree perform way better** than the other.



Conclusions

- As the flight number increases, the first stage is more likely to land successfully
- 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
- Certain orbit type has a better success rate
- Decision tree perform better in predicting success/failure of a rocket launch

Appendix

- Github repository: https://github.com/rahmalianto/Coursera-Applied-Data-Science-Capstone/tree/master

Thank you!