DAV 6150 Module 3 Assignment

Cleaning "Messy" Data

***** You may work in small groups of no more than three (3) people for this Assignment *****

This assignment requires the application of many of the data preparation concepts covered in the required readings for Module 3. Specifically, you will be exploring and "cleaning" a data set that suffers from a variety of data integrity + usability issues. The data set you will be using is comprised of information on more than 12,700 wines, with many of the attributes being related to the chemical composition of the wines. An overview of the data attributes is provided below:

Data Set Attribute	**Description**
INDEX	Unique ID
TARGET	Response Variable (indicates # of cases of wine sold)
AcidIndex	Measures total acidity of wine via a weighted average
Alcohol	Alcohol Content
Chlorides	Chloride content of the wine
CitricAcid	Citric Acid content of the wine
Density	Density of the wine
FixedAcidity	FixedAcidity of the wine
FreeSulfurDioxide	Sulfur Dioxide content of the wine
LabelAppeal	Subjective marketing score that indicates the appeal of the design of the label on the bottle
ResidualSugar	Residual sugar content of the wine
STARS	Wine rating as determined by experts (4 = excellent; 1 = Poor
Sulphates	Sulfate content of the wine
TotalSulfurDioxide	Total sulfur dioxide content of the wine
VolatileAcidity	Volatile acid content of the wine
pH	pH of the wine

*** If you are not fully familiar with the meanings and typical data values for any of these metrics, you might want to consider developing some **domain knowledge** prior to undertaking the required tasks for this assignment. Data science practitioners are expected to be adept at acquiring and incorporating domain knowledge whenever needed. ***

Once you are comfortable in your understanding of the various data attributes, get started on the assignment as follows:

1) Load the provided M3_Data.csv file to your DAV 6150 Github Repository.
2) Then, using a Jupyter Notebook, read the data set from your Github repository and load it into a Pandas dataframe.
3) Using your Python skills, perform exploratory data analysis (EDA) on all of the provided data attributes and document your findings. Your EDA writeup should include any insights you are able to derive from your statistical analysis of the attributes and the accompanying exploratory graphics you create (e.g., bar plots, box plots, histograms, line plots, etc.). It is up to you as the data science practitioner to decide how you go about your EDA, including selecting appropriate statistical metrics to be calculated

+ which types of exploratory graphics to make use of. You should also identify any and all potential data integrity issues (e.g., missing data values; invalid data values; etc.) and, based on your analysis, determine whether any data attributes might need to be transformed prior to being used within a machine learning model. Your goal should be to provide an EDA that is thorough and succinct without it being so detailed that a reader will lose interest in it.

4) Using your Python skills, perform appropriate data preparation tasks relative to the data integrity + usability issues you identified via your EDA work. Describe the ways in which you have transformed / prepared the data for use within a machine learning algorithm, e.g., have you deleted any observations? Used imputation to fill missing data values? Created any new variables? Transformed data via mathematical transforms (e.g., Box-Cox, logarithms, etc.) or binning? etc. Be sure to explain your justification for each adjustment you have made to the data.
5) Using Python, re-run your EDA analysis on any variables you have adjusted during Data Preparation and compare / contrast your results to those you saw prior to performing your Data Preparation adjustments. Describe how each of your Data Preparation adjustments have improved the data set for purposes of using it within a machine learning algorithm.

Your deliverable for this assignment is your Jupyter Notebook. It should contain a combination of Python code cells and explanatory narratives contained within properly formatted Markdown cells. The Notebook should contain (at a minimum) the following sections (including the relevant Python code for each section):

1) **Introduction (5 Points)**: Summarize the problem + explain the steps you plan to take to address the problem
2) **Exploratory Data Analysis (35 Points)**: Explain + present your EDA work including any conclusions you draw from your analysis regarding the integrity + usability of the data in its raw state. This section should include any Python code used for the EDA
3) **Data Preparation (45 Points)**: Describe + show the steps you have taken to address the data integrity + usability issues you identified in your EDA. This section should include any Python code used for Data Preparation
4) **Prepped Data Review (10 Points)**: Explain + present your post-Data Prep EDA analysis. This section should include any Python code used for re-running your EDA on the variables adjusted during your Data Preparation work.
5) **Conclusions (5 Points)**

Your deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Upload your Jupyter Notebook to your online DAV6150 GitHub directory. Be sure to save your Notebook using the nomenclature we‘ve been using, i.e., **first initial_last name_M3_assn**" (e.g., J_Smith_M3_assn_). Then submit the resulting web link via Canvas within the Module 3 Assignment page. ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***