DOUGLAS COLLEGE – Winter 2025

CSIS 3290 – Lab 01

- **Create a folder and rename it according to the folder structure and naming convention stated below**
- **All the files you are required to submit for the assignment should be placed inside this folder.**
- **You will lose points if you just cut and paste materials from close exercises (e.g., If I see the same comments, variable names, etc. from class exercises being using in your code).**
- **If cheating is determined (i.e., you shared your work with another student in the class), your work will a ZERO mark and you will face further consequences.**
- **Make sure to include all the necessary files to make sure that the code can run properly without producing any error**

In this lab, we will practice how to prepare and explore an unclean dataset. You need study the demo code and do your own research to make sure that you can perform all the tasks describe below.

1. Create a python notebook named as **Lab1_FLaXXX** with F signifies the first letter of your **first name**, La signifies the first two letters of your **last name** and XXX denotes the last three digits of your **student ID**.

2. Create a markdown cell at the top of the Jupyter notebook to state the lab, **your name and student ID** with the correct heading.

3. For each of the following section, you need to create a **markdown heading cell** followed by a few code cells to complete the tasks. Please also put some comments in each code cell.
   a. **Load the python library**. Please load all the required python libraries in this section
   b. **Read the data**. Please load the provided csv file and have a peek at the data by using the head() function. Then display the column information. Also display the summary of datatypes of your dataset.
   c. **Checking duplicates**. Please drop duplicates by keeping the last of the duplicated row
   d. **Column organization**.
      - Split the column **Make/model** into two columns named as **make** and **model**
      - **Change the column names** such that it is all in lower-case letter and that all whitespaces are replaced by underscore
      - **Work on the price column**. Remove any appearances of **$** sign and **comma** from the price column. Then change the datatype into numerical
   e. Dropping and Filling data
      - **Drop with threshold**. We have many null values, but we do not want to drop too many rows of data. Firstly, we want to drop any rows that have more than three nulls, i.e., we will drop with threshold parameter equal to seven.
      - **Make sure that the price column does not have any null** since it is our target column. Use dropna with subset parameter to drop any null within the price column.
      - **Work on kilometer and mileage columns**. You should check that column kilometer has more null than mileage. Fill the missing values in mileage with the following formula, mileage = kilometer/1.609. Then drop the column kilometer. At this point, you should still have 72190 rows of data
      - **Fill the fuel_type column with its mode**. Check using value_counts() the count of each category in fuel_type column. Fill the missing value in fuel_type column with the most common value, i.e., ***Petrol***.

- **Fill the transmission column with its mode**. Check using value_counts() the count of each category in transmission column. Fill the missing value in transmission column with the most common value, i.e., ***Manual***.
- **Fill the mpg column with its mean**. If you check the distribution of mpg column (you can plot it if you know how to do it), you will see that the mpg column has a rather nice ***normal*** distribution. Hence, we will fill the missing value in the mpg column with its ***mean***.
- **Fill the mileage column with its median**. If you check the distribution of mileage column (you can plot it if you know how to do it), you will see that the mpg column has a rather ***skewed*** distribution. Hence, we will fill the missing value in the mileage column with its ***median***.
- **Fill the tax column with zero**. We will assume that if there is no tax information, the tax for that specific car is zero.
- Try to check using the head() and info() again to make sure that the changes took place.

f. Formatting the categorical columns

- Find all the categorical columns. Within each categorical column, replace any hypen “-“ with underscore “_” and replace any whitespace “ “ with underscore “_”. Change the text into lower case letter as well.

4. Remove the remaining NaN and save the cleaned file

Use dropna to drop the remaining null. You should still have 69575 rows of data. Reset the index of the dataset. Save the dataframe into a file named as Lab1_cleaned_JDoXXX.csv without any index using index=False parameter.

Note on submission:

- Create a folder named as Lab1_FLaXXX following the naming convention.
- Put your Jupyter notebook and the original and cleaned dataset in this folder.
- Zip the file and submit it through the blackboard

LAB/ASSIGNMENT PRE-SUBMISSION CHECKLIST

- Did you follow the naming convention for your files?!
- Did you follow the naming convention for your folder?!
- Does your submission work on another computer?!
- Double check **before** submitting