NATIONAL INSTITUTE OF BUSINESS MANAGEMENT

In collaboration with **COVENTRY UNIVERSITY**

Diploma in Data Science 21.1P

July – September 2021

DATA VISUALIZATION WITH PYTHON

Final Assignment (15 %; Due 26th October 2021)

Submission Points:

1. Your final sub missions should be a html file and a Jupiter notebook file.
2. All the interpretations should be inside markdown mode in your Jupiter notebook.
    - On top of 100 Marks for this assignment, I will evaluate how you have presented this notebook using Markdown to design the HTML document. From heading to interpretation styles and everything. It will be considered as an extra quiz worth of 5 marks and that will be included inside your quiz pool. Where top 5 will be considered.
    - Please Use 3rd lecture notes containing Markdown notes for this purpose.
3. While creating the Final HTML file, please use below instructions:
    - Your HTML file will be used to evaluate the final marks and it is mandatory that the final HTML file should not have any python codes.
    - To create a HTML file by hiding codes:
        - Save your final Jupiter notebook in any directory.
        - Open the anaconda prompt
        - Check whether nbconvert installed.
            - If not **install it**
        - Run below line
            - jupyter nbconvert --to html --TemplateExporter.exclude_input=True **"file path of Jupiter notebook"**
        - An HTML file will be saved as final document in the same directory of your Jupiter notebook.
        - Use semi colons at the end of the each code cell to avoid displaying **out[ ]** in Jupiter notebook which will give unintended output in your final HTML file.+

Notes:

4. Dataset given is obtained from Kaggle.
5. Dataset and description of the dataset is given in assignment link.
6. Marks will be given for selecting suitable graphs for the datatypes in the questions and adding extra information inside the graphs.
7. **Specially give importance to title, axis names, fig size, color maps and adding extra dimension by introducing another visual variable. (Lecture 4 >> color, size, shape etc)**
8. plagiarism, copying, giving, or receiving aid are strictly not permitted and considered as exam violation. (Jupiter notebook codes will be examined to validate your assignment) but of course you can discuss with your friends and get their help.
9. Selection of graphs is completely arbitrary, but it should explain the question properly.
10. You can use Seaborn, Matplotlib or both modules in python to answer the questions. (But for some charts use of seaborn is mandatory)

Questions: for each questions Variable Names of interest is given inside curly brackets {}

1. Interpret the main aspects of the data (25 to 50 words)? (Count of dataset, important variables) [5 marks]

2. 81 variables in this data. Select any 6 variables you feel important including {**SalePrice**} and create a Seaborn pairplot. Based on this plot interpret your initial idea about the data. (75 – 125 words) [10 marks]

3. Price is the important variable and understanding the distribution of price is very important to understand the data. [15 marks >> 5 each]
   a) Create a simple chart to identify the overall distribution of price variable and interpret it. {**SalePrice**}
   b) Identify the distribution of price with respect to different Types of foundation inside a single chart figure {**SalePrice, Foundation**}
   c) Create a chart or facet grid to identify the distribution of price with respect to alley access and condition of the material on the exterior. (Use kernel density estimation plots for smooth charts) **{SalePrice, Alley, ExterCond}**

   interpret your findings for each data after creating proper visualization chart for each question.

4. Create a simple chart to identify the **median** Sale price for each year and interpret the results. [10 marks] {**SalePrice, YrSold**}

5. By creating necessary graphs interpret which of the following 2 variable have high relationship with Sale price. [20 marks >> 5,5,10]
   a) Relationship between price and living area square feet {**GrLivArea**, **SalePrice** }
   b) Relationship between price and square feet of basement area {**TotalBsmtSF**, **SalePrice**}
   c) Finally comment a or b have high relationship with sales price by mentioning graphical reasons. (Use regression plot to create both regression line in the same graph to validate your arguments)
   (Interpret your finding for all the questions)

6. In an effort to visualize all the points of Sales price with respect to rating of basement, a student tried run below code to create a chart: {**SalePrice, BsmtFinType1**} [15 marks >> 5 each]

```
sns.relplot(data = df, x='SalePrice', y='BsmtFinType1' )
```

(**df** >> pandas data frame name of the given housing price dataset)

   a) Run the above code, create the chart, and explain what is the visual issue in this chart?
   b) Suggest a modern type of chart in seaborn we can use to visualize the same information but avoiding the issue you mentioned in question 6.a.
   c) Based on this chart interpret your findings.

7. Using the same variables in 3.c create a boxplot and violin plot in seaborn. Main category Should be Alley type and distinguish exterior condition using color. Interpret and discuss pros and cons of charts in 7 comparing it with chart created in 3.c (50 – 100 words) **{SalePrice, Alley, ExterCond}** [15 marks]

8. Do you think, average sales price is changing with respect to month they sell? Create necessary plot and interpret the results [10 marks] {**SalePrice**, **MoSold**}