CIS 5357

Assignment # 8

(35 points)

**Due Date: by 11:59 pm on Sunday, November 15, 2020**

**Caution:**

**The instructor has found many submissions with duplicate or similar code. The instructor expects individual effort on each and every submission. Assignments and exams assigned in this course are neither group projects nor any kind of group activity or collaboration are sanctioned. Such activities will be treated per the Academic Dishonesty policy as stated in the course syllabus. Each submission will be closely examined for plagiarism**. **A slightest hint of duplicate or similar code will be examined for academic integrity violation. Such submissions will automatically receive a grade of zero and reported to the Graduate College for further disciplinary action. To avoid such scrutiny, please do your own work. Please consult your instructor if you any questions or need clarification about assignments and examinations.**

1. **Requirements for Assignment 8:**
2. Name your Jupyter Notebook ‘YourName-Assignment8.ipynb’
3. Include your name and submission date as level 2 headings in the first cell of the notebook.
4. Insert a markdown cell with level 3 heading “Grading Comments – Totality of all points noted below resulted in a reduction of xx Points”
5. Insert a markdown cell with level 3 heading for the Program Name and its objective
6. Please store the AutoMPG.CSV file in the following path: **“/Users/cis\_developer/CIS5357Fall2020/AutoMPG.CSV”.** If you are not sure how to create this path on your computer (windows or macs), now is the time to learn it. The movie database file **MUST** be in this path. Do not replace your account name in place of cis-developer.
7. Upload your source code file (.ipynb) to the Assignments section of Canvas BEFORE 11:59 pm on Sunday, November 15, 2020. No other files need to be uploaded.
8. **NO LATE ASSIGNMENTS WILL BE ACCEPTED. ASSIGNMENTS SENT VIA EMAIL AS ATTACHMENTS WILL ALSO NOT BE ACCEPTED. YOU ARE ALLOWED A MAXIMUM OF THREE ATTEMPTS TO SUBMIT YOUR FILE BEFORE THE DUE DEADLINE. YOU MISS THE DEADLINE, YOU LOSE IT. So, please start early to have a chance at getting any problem resolved before the submission deadline.**

**Introduction**:

You will use the data contained in the AutoMPG.csv file to prepare answers to the following questions. You will use the Linear Regression algorithm from the Statsmodels statistical package library and Pandas framework for this project.

1. Download and copy the AutoMPG.csv file to the folder indicated in requirement # 5 above. Your code should read the data from this location. Do not put this file in local path on your computer. I will execute your notebook such that it will access the AutoMPG.csv file from this path on my computer. **The instructor will stop grading and a grade of zero will be assigned if your submission cannot find the autompg.csv in the specified location.**
2. *The objective is to apply Linear Regression algorithm to predict mpg given various other factors known to impact mpg rating.*
3. Create as many Dataframes as you need, especially if you are making transformation that will change the schema of AutoMPG.csv (adding OR deleting columns will change the schema).
4. **The Dataframes you create must have meaningful names,** not DF, DF1, etc. The name you assign must have the letters **df** appended to it e.g., actors\_df or movies\_df
5. Where questions are asking you to provide a written response, please include your answers in markdown cells.

**Schema of the Dataset:**

1. The AutoMPG dataset is saved in CSV (Comma Separated Values) format. It includes 9 attributes/features as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column # | Column Header | Data Type | Numeric DataType | Description |
| 1 | mpg | Float | continuous | Miles per gallon |
| 2 | cylinders | integer | discrete | # of cylinders |
| 3 | displacement | Float | continuous | Displacement in CCs |
| 4 | hp | Float | continuous | horsepower |
| 5 | weight | Float | continuous | Car’s weight |
| 6 | acceleration | Float | continuous | Number of seconds to go from 0 to 60 mph |
| 7 | model\_year | Integer | discrete | Represents last two digits of 70’s |
| 8 | origin | Integer | discrete | Numeric representation of where the car was manufactured |
| 9 | car\_name | string | unique | Car make/model |

1. The first row in the data set is column header text (or variable name)

**Questions to be Answered (35 points).**

**In each of the following, you must support your answer with output generated by appropriate python code. Where possible and appropriate, python code to support each question should be in its own cell. That is, for 10 questions, you should have 10 cells other than cells for importaing packages and markdown cells containing your typed responses.**

1. What is the shape of the data contained in AutoMPG.csv? What do the numbers mean?
2. What features (or attributes or variables) are recorded for each automobile?
3. Provide a schema of the AutoMPG data set to verify that all relevant features contain numeric data type. Are there any columns/features that is not applicable in developing a Linear Regression algorithm? That is, does not meet the requirements/assumptions to use a Linear Regression model. If so, eliminate those columns from further analysis and regenerate the schema to ensure that the ‘offending’ column(s) is/are removed from further analysis. Remember, it should not be permanently removed from the dataset.
4. Evaluate the correlation between mpg and each of the independent variables (pairwise mpg and cylinders, mpg and displacement, etc.). On the basis of individual correlation coefficients, can you determine which independent variables are useful in predicting mpg? Write response to your analysis as comments in your source code/notebook. Remember, Correlation coefficient value ranges from -1 to +1; closer to 1, stronger the relationship. \
5. Bases on correlation analysis performed in step 4, identify only one independent variable that is the most highly correlated variable with the dependent variable. Generate a new dataframe that contains only these two variables. Display this dataframe contents.
6. Provide a listing of summary descriptive statistics such as average and standard deviation, etc. for the independent variable that you selected. Round all float data to two decimal places.
7. Generate a scatter plot to display the relationship between your selected independent variable and the dependent variable. Label all axes properly and appropriately (you may use the plotting capabilities of the Pandas dataframe or matplotlib library that used in Assignment 7). Does your data visualization display any underlying linear trend?
8. Run a linear regression analysis algorithm (using the ordinary least squares or OLS method) on your selected independent and dependent variables. Display the output of applying this algorithm to your data. Based on this output:
   1. write the linear model that is suggested in the form y= a + bx, where values for a and b will be extracted from the results of your model.
   2. How well does the linear model identified by the algorithm fit your data? To answer this question, look at the R2 or adjusted R2 value and base your answer on this value. You must provide a correct interpretation of this value in terms of what it means within the context of the problem.
   3. Are the values of regression coefficient(s) and constant statistically significant and worthy of using to predict the MPG?
9. Generate a scatter plot as before but superimpose the regression line on top so you can see how well the model fits the data.
10. Select one unknown value (value that is not included in your dataset) for the independent variable that you selected in your analysis, and predict the MPG for that value. Display the answer. This has to be done via python code (and not hand calculation). This is a topic you will have to research.