University of Cape Coast

College of Humanities and Legal Studies

School of Economics

Department of Data Science and Economic Policy

DMA 803: Data Literacy and Mining

**Instructions**

Please you are required to use Stata and Weka for the work. You must Print and submit a well-explained as part of the Term paper. You must also provide detailed explanations of tables, figures and any other output generated using Stata or Weka using intuition, theory, and empirical explanations in a printed Word document that contains the tables and figures. Where data cleaning is involved, you must also provide the reason(s) why the action has been taken in the Word document to be submitted.

**Attachment to the Questions**

Attached to this question is an Excel file that contains the data to be used for the exercise. The data has the file named IBESDATA. The variable definitions can be found in the same Excel file. The data. The IBESDATA is data extracted from the 2013/2014 Integrated Business Establishment Survey by the Ghana Statistical Service. The definitions of the variables are in Sheet 2 of the file. The MBA is data on a market basket of goods purchased from ShopRite shops across Africa.

**Answer All Questions**

**Data Cleaning [20 Marks]**

1. Use the IBESDATA and generate the age variable using 2016 as the reference year for the computation of the age of firms surveyed in the 2013/2014 IBES Data. Please note that all firms reporting as having been established before 1950 must not be considered part of this data and all firms indicating they were established after 2014 must also not be included.
2. Generate a boxplot of the profit variable (the quantitative measure of profit) and comment on the result.
3. Assuming you have been informed that firms that reported zero profit and firms that reported profits above GH₵19,561.91 reported figures that were in error. Explain how you would deal with this problem and generate a new boxplot after dealing with the problem and explain.

**Data Exploration [10Marks]**

1. Generate a scatter plot of age and profit (with proper labelling). Explain the relationship you see.
2. Generate another scatter plot of age and profit by region (with proper labelling)
3. Generate a bar graph of profit (using the qualitative measure of profit) and region first, using column percentages and second using row percentages. Please make sure your graph is well labelled and a proper explanation is provided.

**Hypothesis Testing and Linear Regression [30 Marks]**

1. Using the worker\_sex variable, can you conclude that men and women have equal job opportunities in the Ghanaian business environment? Please indicate the appropriate test for this and explain why. Please explain the results generated from the test.
2. A business owner wants to know whether his profit would be high if he employs more males or when he employs more females.
3. A foreign firm wants to identify whether male-dominated employees are more efficient than female-dominated employees.
4. A foreign investor wishes to know the relationship that exists between the region of operation and the profitability of a firm.
5. Government wishes to identify the effect of age and the number of male and female employees, on profitability.

|  |
| --- |
| Group Obs Mean Std. err. Std. dev. [95% conf. interval] |
| Female Dom 3,600 6610.013 84.48586 5069.151 6444.368 6775.658 |
| Male Dom 6,645 7226.923 63.8596 5205.634 7101.737 7352.108 |
| Combined 10,245 7010.146 51.04095 5166.242 6910.096 7110.197 |
| diff | -616.9097 106.7447 -826.1501 -407.6692 |

Group Mean Std. err. z P>|z| [95% conf. interval]

Female 0.3175179 0.0032591 0.3111302 0.3239055

Male 0.6824821 0.0032591 0.6760945 0.6888698

diff -0.3649642 0.004609 -0.3739977 -0.3559307

under H0: 0.049505 -73.72 0.000

|  |
| --- |
| Profit Coefficient Std. Err. T P>|T| [95% Conf. Interval] |
| Age 10.98657 5.111343 2.15 0.032 0.9673395 21.0058 |
| Male Employees 74.7787 6.982796 10.71 0.000 61.09106 88.46635 |
| Female Employees -13.98786 7.62664 -1.83 0.067 -28.93756 .9618504 |
| Constant 6654.627 78.60527 84.66 0.000 6500.545 6808.708 |







