**Homework One**

**Instruction:** The data file for this homework is ToyotaCorolla.xls, which can be downloaded from Canvas. Create a new Word document and save it as HW1Answers\_X (where X is your team number). Where required, write your answers/paste screenshots into this Word document. Your response should not exceed 100 words for each below question. Write every member’s full name and participation on the first page of the Word document as follows. You need to submit this Word document and XLMiner solution.

|  |  |  |  |
| --- | --- | --- | --- |
| Participant | Complete the Assignment before the Meeting (Y/N) | Percentage of Contribution | Justification |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

The dataset contains data on used cars for sale during the late summer of 2004 in The Netherlands. It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other speciﬁcations.

Before building models using various data analytics techniques, we need to verify that the data are in reasonable condition. Exploring data techniques are utilized to help us understanding the global landscape of the collected data, detecting unusual values, cleaning and manipulating data, visual discovery and “hypothesis generation”.

Methods for exploring data include looking at various data aggregations and summarizes, both numerically and graphically. This includes looking at each variable separately as well as looking at relationships between variables. The purpose is to discovery patterns and exceptions. For numerical variables, histograms and box plots can be used to learn about the distributions of their value, to detect outliers, and to find other information that is relevant to the analysis task. Similarly, for categorical variables, we can use bar charts. Scatter plots of pairs of numerical variable can be used to learn about possible relationships, the type of relationship, and to detect outliers. In addition to scatter matrix plot, we can use correlation tables to identify the possible relationship between the variables.

**Tasks:**

1. Which variables are numerical? What are binary? Which are ordinal? Which are nominal?
2. Create a table with the average, median, min, max, standard deviation, count blank(the number of records with missing values) for each of the quantitative variables. This can be done through 1) Excel’s functions or 2)Excel’s Data 🡪 Data Analysis 🡪 Descriptive Statistics menu and then use a excel function for counting missing values.
3. Use XLMiner to plot a histogram for each of the quantitative variables. Based on the histogram and summary statistics, answer thee following questions:
   1. Which variable have the largest variability?
   2. Which variables seem skewed?
   3. Are there any values that seem extreme (outliers)?
4. Compute the correlation table for the quantitative variable (use Excel’s Data 🡪 Data Analysis 🡪 Correlation menu). In addition, use XLMiner to generate a matrix plot for these variables.
   * Which pair of variables is most strongly correlated?
   * How can we reduce the number of variables based on these correlations?
   * How would the correlations change if we normalized the data first?
5. Explore the relationship between price and some of the numerical variables by using scatter plot matrix. What does this plot show us? Present your insights within 100 words
6. Explore the relationship between price and the categorical variables by using **box plots** and Pivot Charts. Which of the pair among variables seem to be correlated? What do these plots and charts show us? Present your insights within 100 words
7. Describe how you would handle records with missing values.
8. Describe how you would handle records with outliers.
9. Describe how you would convert Fuel Type and Color to binary variables. Conﬁrm this using XLMiner’s utility to transform categorical data into dummies.
10. Conduct a principal components analysis on the cleaned data and comment on the results. Should the data be normalized? Discuss what characterizes the components you consider key. Use the “principal components utility in XLMiner”.
11. Present your insights in a word document. Your insights should not exceed 100 words for each above question.
12. Modify the worksheets’ titles and names appropriately to reflect the contents.

**Important submission instructions**

Save your Word file and XLMiner solution. Use the link “Homework 1” to upload these files. **Due by 11.59 P.M. Jan. 29, 2019.**