OPS 804 Midterm Exam

Fall 2019

Exam time: 2 hours

Problem 1:

- Load the anomaly dataset.
- The data shows the number of customers who have purchased one of the products from a company's Ecommerce site.
- Determine if there are any outliers in this dataset using two methods:
    - IQR method
    - By determining the 95% confidence interval.

Problem 2:

- Load the Baseball_salary.csv as your source.
- Consider Salary as the output and the rest of the columns as input.
- Perform EDA on this dataset using the following steps:
    - Clean the data by removing nulls and text data (categorical data if any)
    - Find descriptive statistics and comment on what columns may need further pre-processing such as normalization or standardization.
    - plot the histograms for all columns.
    - Does any column need to be log transformed? Explain the reason for your decision.
    - Choose the log of salary as your output instead of salary and insert the logSalary in the dataframe in place of the salary column.
    - Perform normalization and standardization on the input columns and determine if any of these transformations could potentially improve the modeling phase.
    - Perform correlation analysis and identify the independent and correlated columns.
    - Plot the scatter plots and determine if there is any linear and nonlinear relationship between the columns of the data.
- For both problems upload your code and the results and explanations as a word file onto SMCMBA website.