**MSBA 250 — Applied Business Analytics**

**University of the Pacific**

**Spring 2024**

**Assignment 4**

**Due: Saturday April 6, 2024**

**Instructions:**

This problem set has 3 questions. The full credit is 20.

**Problem 1 (7 points):**

Please find the EXCEL file named “Employee Retention,” and please find the determinants (either positive or negative) for employee retention (measured by YearsPLE).

1. Please first code the categorical variables into numerical variables. A screenshot of a computer

   Description automatically generated
2. Please run multiple linear regression and attach the regression result Table.

A screenshot of a computer

Description automatically generated

1. Based on the regression results, please discuss the determinants (either positive or negative) for employee retention (measured by YearsPLE).

**YrsEducation:** Positive coefficient, 0.083979089, but p-value, 0.863119116, is insignificant. Suggesting the coefficient is unreliable.

**College GPA:** Negative coefficient, -0.745109069, but p-value, 0.476600602, is insignificant. Suggesting the coefficient is unreliable.

**Age:** Positive coefficient, 0.181607005, but p-value, 0.13388154, is insignificant. Suggesting the coefficient is unreliable.

**Gender:** Negative coefficient, -0.327410634, but p-value, 0.676697966, is insignificant. Suggesting the coefficient is unreliable.

**College Grad:** Positive coefficient, 1.391804392, but p-value, 0.336176258, is insignificant. Suggesting the coefficient is unreliable.

**Local** Positive coefficient, 3.492096285, and p-value, 0.000105319, is significant. Suggesting the coefficient is reliable.

1. Please discuss the insights based on your findings in Step (3).

The value of R2 indicates that 47% of the variation in the dependent value is explained by local variable and being local is a factor associated with longer employee retention.

**Problem 2 (6 points):**

Please find the EXCEL file named “Engines,” and then please use a regression model to fit the relationship between sample and production time. Please:

1. List your regression model
   1. Logarithmic
   2. Polynomial
   3. Power
2. Attach a screenshot of the scatter plot of these two variables (sample and production time) and the trend line of your regression.

A screen shot of a graph

Description automatically generated

1. How about your model fit? Please provide some evidence.

Out of the models, the Logarithmic has the largest R2, 0.9959. Therefore, it is significant.

1. Please discuss the insights based on your findings in Steps (2) and (3). What do you think about the relationship of these two variables (sample and production time)

The scatterplot displays a negative correlation, therefore, there’s a significant relationship of the two variables, Sample and Production Time, with the R2 is 0.9959.

**Problem 3 (6 points):**

Please find the EXCEL file named “Students,” and then cluster these students based on your selected features.

1. Please the features you selected.

* Math Score
* Reading Score
* Parental Level of Education

A screenshot of a computer

Description automatically generated

1. Please use statistics to show why your selected features are appropriate.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Please discuss how many groups you clustered and for each group (cluster), what are the features of the students?

There are 7 groups of clusters , each group is a sum of math and reading score with a different parental level of education. Group 1 has the lowest average score of math and reading with most parents have high school or some high school education. While group 7 has the highest average score of math and reading and most parents who have master’s degrees. There’s a correlation between score and parental’s level of education, students with parents who have higher level of education have better score of math and reading.