STA301-Homework **Due: Monday, May 5 at 11:00 p.m.**

Please submit both a hard copy and an electronic copy via email. Make sure to include a title page listing all group members' names and student IDs.

Part I: To be completed by hand:

Q1: When a student performs poorly on a quiz, the student sometimes is convinced that their score is an anomaly and that they will do much better on the next quiz. The following data show the quiz scores (out of 20 points) for 4 students in STA301 in Spring2024. X= quiz1 score and Y=quiz 2 score.

(8, 10), (12, 12), (8, 15), (11, 10)

a) Write the residual sum of squares (RSS) function and use it to derive the least squares estimates of the regression line.

b) Give an estimate of the pure error sum of squares. What are the degrees of freedom for the pure error?

c) Give an estimate of the common variance, σ^2.

d) Complete the following table for the calculations of expected residuals under Normality assumption:

Rank	$\mathbf{e}$	$\boldsymbol{P}$	$\hat{\boldsymbol{e}}$
1			
2			
3			
4			

e) Use a software to calculate the correlation coefficient between $\mathbf{e}$ and $\hat{\boldsymbol{e}}$ and use it to make a conclusion on whether the normality assumption is possibly satisfied for this small sample of instances.

Part II: To be completed with the help of software:

The dataset "*dubai_properties*" provides a detailed snapshot of rental property listings across major cities in the United Arab Emirates, including Abu Dhabi, Dubai, Sharjah, Ajman, Ras Al Khaimah, Umm Al Quwain, and Al Ain. Gathered from bayut.com, it includes various attributes such as property type, size, rent, and location details, making it a valuable resource for data-driven insights into the UAE rental market. Ideal for data analysts, researchers, and real estate professionals, this dataset allows for comprehensive analysis of trends, pricing, and market dynamics in the UAE's diverse rental landscape.

a) Is the study an example of experimental design, observational study or mixed?

Randomly sample 5000 entries from the data set (without replacement) and use the resulting data set to answer the below.

Probability calculations:

b) Construct the frequency table of the "Furnishing" column and use the table to answer the following questions:
   i. Calculate the percentage of "furnished" properties.
   ii. Suppose you survey 50 renting properties across major cities in the UAE, what is the probability that 35 of them are furnished?

c) Construct a frequency table of the "Rent_category" columns and use the results to answer the following questions:
   i. Calculate the percentages of the three values "High", "Medium" and "Low".
   ii. Suppose you survey 50 properties across major cities in the UAE, what is the probability 15 are Low rent, 10 are Medium rent and the rest are of High rent?

d) Suppose the average number of bedrooms for the rental properties in Dubai is 2. Suppose you survey some rental properties in Dubai, what is the probability that you find at most 1 of them with 2 bedrooms.

e) Construct box plot chart of "*Age_of_listing_in_days*" by the variable "*Furnishing*". Discuss the patterns you see in the plots and compare the central values, variability and skewness across the two Furnishing categories

f) Give the summary statistics of the "*Age_of_listing_in_days*" by the variable "*Furnishing*"?
   i. Compare the mean values of the age of listing for furnished and unfurnished properties. Which type of properties has longer listing.
   ii. Calculate the coefficient of variation of the age of listing for the two furnishing categories and discuss the results.
   iii. Construct the 95% confidence interval for the ratio of variances of the age of listing for furnished and unfurnished properties. What do you conclude?
   iv. Based on the result in part iii., construct the 95% confidence interval for the difference in the mean age of listing of the two groups of properties ($\mu_1 - \mu_2$). Based on the interval, can you conclude that *mean age of listing is significantly different for the two groups?* Explain how it is different.

g) Construct the Normal probability plot (Q-Q plot) for the rent? Does the plot indicate that the rent is normally distributed.

h) Construct the correlation coefficient matrix for the features: rent, beds, baths, area_in_sqft, rent_per_sqft and age of listing . Comment on the strength of the correlations? Which pair of features is the most correlated? Which pair of features is the least correlated?

Hint: consider depicting the correlations in a heatmap.

i) Construct a two-way classification table for the variables "*rent_category*" and "*city*". Use the table to test if the two variables are independent? State the hypotheses and conclusions clearly and discuss any pattern you find in the table. Use α=0.05.

j) Construct a two-way classification table for the variables "*rent_category*" and "*Furnishing*". Test for independence? State the hypotheses and conclusions clearly and discuss any pattern you find in the table. Which rent category is more likely to be unfurnished. Use α=0.05.

Naïve Bayes

k) Split your sampled data set into 70% training data set and 30% testing data set.

l) Apply Naïve Bayes classifier to the training data set in order to develop a model that predicts the target variable "Furnishing" using the features: rent, beds, baths, type, area_in_sqft, rent category and age of listing

m) Apply your model in part l) to the testing data. Construct the confusion matrix and calculate the misclassification rate.

n) Does the data suffer from imbalanced classification of the target variable? If so, how does this imbalance affect the accuracy of your model.

Simple Regression

o) Construct a scatter plot of the variable "Y=rent" against "X= area_in_sqft" using the training dataset. Comment on the type of relationship between the two variables? Are there outliers?

p) Fit a simple linear regression function relating Y to X.

q) Explain what does the value of R-sq represent and use it to obtain the correlation coefficient between Y and X? Is the correlation significant from zero?

r) Explain what the lack of fit test tells you about the fitted equation in part n). Conduct the test and state your conclusions clearly.

s) Evaluate the appropriateness of the regression equation and use residual analyses to check model assumptions. Are all model assumptions satisfied?

t) Apply one of the remedies discussed in the class to improve the fitted equation.

u) Use the improved regression equation in part r) to find the 95% prediction and confidence intervals for X=350 sqft, and X=2200 sqft. Which prediction seems to be most accurate? Why?

v) Use the regression equation in part p) to predict the Y for the values of X in the testing dataset. Calculate the sum of squared prediction error (SSE) and compare it the SSE obtained in the model fit in part p)?

Note: You must submit the programing code, as a separate file, with your HW answers

Project Data set:

You must consider implementing all or most of the analysis questions in Part II of the HW to your project data set.