STAT 231 Spring 2021 Final Project

The Final Project is due on **Tuesday August 10th at 2:00pm** EDT.

Your Final Project should be written in a style similar to a report that you would write for an employer, that is, it should be well organized and well-written in **complete sentences**.

However, for ease of marking, **you are to fill in the Word Document template "Final Project Template.docx" posted on LEARN and then save it as a pdf for submission to Crowdmark. Submissions that do not populate the provided template will receive a 10% penalty.** Please keep the question labelling and other structure included in the template – you are only to add your solutions and make no other modifications. Your project must be typed. Any submitted project which is not typed will not be marked and given a mark of zero.

Numerical and graphical summaries must be obtained using R. However, **DO NOT submit any R code in your report.** Your report must be professionally formatted and written as if presented to an audience without detailed R knowledge. **Submissions that include R code will receive a 10% penalty.**

Upload your final project to **Crowdmark as a pdf file for marking.** You can upload your project as one document or individually for each problem. If you upload one document, then you must drag and drop the pages for each problem to the appropriate question as indicated in Crowdmark. **Projects which are left as a single document and not uploaded to the appropriate places in Crowdmark will be assigned a 10% penalty.**

In addition to submitting in Crowdmark for marking, you must submit your materials to the following 3 dropboxes on LEARN:

- Report pdf to the **Final Project LEARN dropbox** to facilitate the running of your midterm through plagiarism detection software.
- R code file to the **Final Project R file LEARN dropbox**
- Your individual dataset to the **Final Project Dataset LEARN dropbox**

A penalty of 10% per hour is applied for late submissions.

Checklist to complete for this project:

- ☐ Complete the project analysis and **create the report within the template provided**.
- ☐ Upload the PDF of your completed project report to Crowdmark by the deadline.
- ☐ Upload the PDF of your project report to the appropriate learn dropbox by the deadline.
- ☐ Upload the file with your R code to the appropriate learn dropbox by the deadline.
- ☐ Upload your unique dataset to the appropriate learn dropbox by the deadline.

Marks for Final Project:

Section	Marks
Analysis Q1	23
Analysis Q2	15
Analysis Q3	22
Analysis Q4	17
Conclusion	10
PDF, R file, Dataset LEARN dropboxes	3
Total	90

For this project, we wish to model the price of residential real-estate in Taiwan. Specifically, we want to examine whether the price of a house can be explained by factors such as the age of the house, the distance to the nearest public transit, or the number of convenience stores within walking distance, and whether a linear regression model is appropriate for this purpose.

The data are provided in the dataset "real_estate2.csv" posted on the LEARN page. The following is some information about the data:

The market value of real estate were collected from Sindian Dist., New Taipei City, Taiwan from approximately late 2012 to early 2014[1]. The variates recorded were:

- date = the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
- age = the house age (unit: year)
- MRT_distance = the distance to the nearest metro station (unit: meter)
- stores =the number of convenience stores in the living circle (I.e. within walking distance) on foot (integer)
- latitude = the geographic coordinate, latitude. (unit: degree)
- longitude = the geographic coordinate, longitude. (unit: degree)
- price = house price per unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared) **Note:** You can refer to it as house price per unit area and ignore the local currency/area unit

Note that the dataset is complete, I.e. there are no missing values.

IMPORTANT: you must create your own unique dataset with 300 observations from the full dataset posted on the LEARN page. To do so, complete each of the following steps:

i. Download the CSV "real_estate2.csv" from the LEARN page

ii. Import the dataset into R studio

iii. Copy the following code, replace the student ID with your own ID, and then run it to create your dataset:
```
set.seed(2033074) #replace with your student ID
obs <- sample(1:413, 300, replace = FALSE) #uniquely sample observations
dataset <- real_estate2[obs,] #create new dataset with 300 observations
```

iv. You must **save your dataset on your computer and use this dataset for all analysis**. To do so use the following R code but replace the file path with something appropriate for you:
```
write.csv(dataset,"C:/Users/Emily/Documents/dataset.csv", row.names = FALSE)
```

v. Upload **your dataset** to the **LEARN dropbox** titled "Final Project Dataset"

---

1. Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271

ANALYSIS (77 marks):

Do the analysis given below for each question as indicated. Be sure to explain clearly what you are doing and write complete sentences. All plots should be titled with axes labelled for full marks.

Question 1 (23 marks) - **Is distance to nearest metro station well described by an exponential model?**

We wish to assess whether the variate X_i = MRT_distance can be modelled as a random observation from an Exponential(θ) for unknown parameter θ.

a) (3 marks) Use R to solve for the numerical summaries listed in the table below. Present your results using the table below.

Minimum		IQR	
1st Quartile		Range	
Sample Median		Sample Mean = $\bar{y}$	
3rd Quartile		Sample Standard Deviation = s	
Maximum		Sample Skewness	

b) (2 marks) Assuming an exponential model is reasonable, give the maximum likelihood estimate for θ based on your dataset and describe in words what attribute of the study population this is estimating.

c) (2 marks) Create and insert a plot of the relative frequency histogram for metro distance with superimposed Exponential($\hat{\theta}$) probability density function.

d) (7 marks) Use R to complete the following table of observed and expected frequencies within the given distance bins. The expected frequencies should be calculated assuming an Exponential($\hat{\theta}$) model. For the expected frequencies, provide **typed (or a screenshot of handwritten inserted into word doc) formulas and steps** that explain how they would be calculated by hand, even if you calculated them in R.

j	Distance (meters) to Public Transit Group x_j	Observed Frequency f_j	Expected Frequency e_j
1	[0, 200)		
2	[200, 400)		
3	[400, 600)		
4	[600, 800)		
5	[800, 1000)		
6	[1000, 1200)		
7	≥ 1200		
Total		300	300

e) (4 marks) Conduct a **likelihood ratio** goodness of fit test for the exponential model by following the steps below:
Step 1: State the hypothesis you are testing (both null and alternate) **(must be typed)**
Step 2: Write out the formula of the test statistic being calculated and give its value **(can be typed or screenshot of handwritten inserted into word doc)**
Step 3: State the formula of the p-value being solved for (including degrees of freedom) and include the appropriate value from the R output **(can be typed or screenshot of handwritten inserted into word doc)**
Step 4: Use your findings to make a conclusion statement (use Table 5.1 from the course notes) **(must be typed)**

f) (5 marks) In full sentences, write a short paragraph that explains your results for parts a)-e) and make a conclusion with support from the numerical summaries, graphical summaries, and goodness of fit test on whether distance to the nearest metro station is well described by an Exponential model.

Question 2 (15 marks) – Is price explained by age of the house?

In this question we examine whether there is a relationship between age of the house and price.

a) (1 marks) Create and insert a scatterplot of the price per unit area (y) versus house age (x).

b) (1 mark) Provide the sample correlation for price and age of house.

c) (2 marks) We will create a two-way (contingency) table for price and age of house that can be used to test for independence. To do so, we will classify age into 3 categories (< 15 years old, [15, 30) years old, and ≥ 30 years old) and classify price into 3 categories (low, medium, high) which corresponds to < 20 dollars/unit area, [20, 40) dollars/unit area, and ≥ 40 dollars/unit area.

Use R to complete the table below with the observed frequencies of the intersections.

Observed Frequencies	Price Category (per unit area)			
Age Group (in yrs)	Low (< 20)	Medium ([20,40))	High (≥ 40)	Total
< 15				
[15,30)				
≥ 30				
Total				300

d) (3 marks) Use R to complete the following table of expected frequencies under the null hypothesis of independence between the two classifications. Show in your report with **typed formulas** how the value in the box labelled x is calculated.

Expected Frequencies	Price Category (per unit area)			
Age Group (in yrs)	Low (< 20)	Medium ([20,40))	High (≥ 40)	Total
< 15	x			
[15,30)				
≥ 30				
Total				300

e) (4 marks) Conduct a formal test of independence by following the following steps:
Step 1: State the hypothesis you are testing (both null and alternate) **(must be typed)**
Step 2: Write out the formula of the test statistic being calculated and give its value **(can be typed or screenshot of handwritten inserted into word doc)**
Step 3: State the formula of the p-value being solved for (including degrees of freedom) and include the appropriate value from the R output **(can be typed or screenshot of handwritten inserted into word doc)**
Step 4: Use your findings to make a conclusion statement (use Table 5.1 from the course notes) **(must be typed)**

f) (4 marks) In full sentences, write a short paragraph that explains your results for parts a)-e) and make a conclusion on whether there is a relationship between age and price of house with support from the graphical and numerical summaries, and the hypothesis test. If there is a relationship, describe what type of relationship.

Question 3 (22 marks) – **Is price explained by distance to nearest metro station?**

The purpose of this question is to investigate the possible linear association that may exist between these two variates.

a) (2 marks) Create a scatter plot and solve for the correlation coefficient between price per unit area (y) and distance to nearest metro station (x). Insert the scatter plot and provide the correlation coefficient. In the context of regression analysis comment on whether the scatter plot and correlation coefficient suggest a linear association. Explain.

b) (3 marks) Assume that a linear model was appropriate with MRT_distance as the explanatory variate x and price as the response variate y. Use R to fit the simple linear regression model

$$Y_i \sim G(\alpha + \beta x_i, \sigma), i = 1,2, \dots ,300 \text{ independently}$$

where the $x_i's$ are assumed to be known constants to your data.

Use your fitted model summary to complete the following table based on the output from R:

maximum likelihood estimate of the intercept α	
maximum likelihood estimate of the slope β	
unbiased estimate of σ	
estimate of the standard deviation of $\tilde{\beta}$	
estimate of the standard deviation of $\tilde{\alpha}$	

c) (4 marks) Use the output from the R function summary() to test the hypothesis H_0: β = 0 (the hypothesis of no relationship). Your answer should be displayed in the following four step structure:
Step 1: State the hypothesis you are testing (both null and alternate) in words that are related to the variates of interest and study population **(must be typed)**
Step 2: Write out the formula of the test statistic being calculated and give its value **(can be typed or screenshot of handwritten inserted into word doc)**
Step 3: State the formula of the p-value being solved for (including degrees of freedom) and include the appropriate value from the R output **(can be typed or screenshot of handwritten inserted into word doc)**
Step 4: Use your findings to make a conclusion statement (use Table 5.1 from the course notes) **(must be typed)**

d) (1 mark) Use R to construct a 90% confidence interval for the slope β. Provide the result in a full sentence.

e) (2 marks) Use R to construct an estimate of, and the 90% confidence interval for, the mean response $\mu(x) = \alpha + \beta x$ for x = 1000m. Provide the results in a full sentence that uses terminology based on the variate descriptions.

f) (2 marks) Use R to construct an estimate of, and the 90% prediction interval for, the predicted response for x = 1000m. Provide the results in a full sentence that uses terminology based on the variate descriptions.

g) (2 marks) Insert the following plots in your assignment:

- scatterplot of your data with the fitted line superimposed
- plot of the standardized residuals versus the explanatory variate
- plot of the standardized residuals versus the fitted values
- qqplot with line of the standardized residuals

h) (4 marks) Referring to the plots in part g), make a conclusion regarding the fit of the simple linear regression model to your data? If the linear regression model is not a good fit, then what do these plots suggest is wrong with the model?

i) (2 marks) Discuss what the results of your model fit tell you about the relationship between these two variates. Your answer should be phrased in 'real-world' terms, that is, discussing the actual variate names and not simply referring to algebraic notation. Your discussion should include an interpretation of $\hat{\beta}$. Are you surprised by your results? Why or why not?

Question 4 (17 marks) – **Is price explained by number of convenience stores within walking distance?**

The purpose of this question is to investigate the possible linear association that may exist between these two variates.

a) (2 marks) Create a scatter plot and solve for the correlation coefficient between price per unit area (y) and number of nearby stores (x). Insert the scatter plot and provide the correlation coefficient. In the context of regression analysis comment on whether the scatter plot and correlation coefficient suggest a linear association. Explain.

b) (3 marks) Assume that a linear model was appropriate with number of stores as the explanatory variate x and price as the response variate y. Use R to fit the simple linear regression model

$$Y_i \sim G(\alpha + \beta x_i, \sigma), i = 1,2, \dots ,300 \text{ independently}$$

where the $x_i's$ are assumed to be known constants to your data.

Use your fitted model summary to complete the following table based on the output from R:

maximum likelihood estimate of the intercept α	
maximum likelihood estimate of the slope β	
unbiased estimate of σ	
estimate of the standard deviation of $\tilde{\beta}$	
estimate of the standard deviation of $\tilde{\alpha}$	

c) (4 marks) Use the output from the R function summary() to test the hypothesis H_0: β = 0 (the hypothesis of no relationship). Your answer should be displayed in the following four step structure:
Step 1: State the hypothesis you are testing (both null and alternate) in words that are related to the variates of interest and study population **(must be typed)**

Step 2: Write out the formula of the test statistic being calculated and give its value **(can be typed or screenshot of handwritten inserted into word doc)**
Step 3: State the formula of the p-value being solved for (including degrees of freedom) and include the appropriate value from the R output **(can be typed or screenshot of handwritten inserted into word doc)**
Step 4: Use your findings to make a conclusion statement (use Table 5.1 from the course notes) **(must be typed)**

d) (2 marks) Insert the following plots in your assignment:

- scatterplot of your data with the fitted line superimposed
- plot of the standardized residuals versus the explanatory variate
- plot of the standardized residuals versus the fitted values
- qqplot with line of the standardized residuals

e) (4 marks) Referring to the plots in part d), make a conclusion regarding the fit of the simple linear regression model to your data? If the linear regression model is not a good fit, then what do these plots suggest is wrong with the model?

f) (2 marks) Discuss what the results of your model fit tell you about the relationship between these two variates. Your answer should be phrased in 'real-world' terms, that is, discussing the actual variate names and not simply referring to algebraic notation. Your discussion should include an interpretation of $\hat{\beta}$. Are you surprised by your results? Why or why not?

CONCLUSION (10 marks):

Write a two-paragraph conclusion for your report.

In paragraph one, summarize the key results from your analysis that address the main purpose of this study: whether the price of a house is associated with factors such as the age of the house, the distance to the nearest public transit, or the number of convenience stores within walking distance, and whether a simple linear regression model is appropriate for these relationships.

In paragraph two, discuss whether or not we can conclude there are casual relationships between the explanatory variates and the response of house price. Explain why or why not with at least 3 different points.