MMA 860 Final Practice Questions

1. Your manager doesn't know much about analytics. That is too bad, but don't worry, you'll have his job soon enough. In the meantime, he has made the following comments. Here are your tasks for each of the below quotes: Identify whether his suggestions /comments are good or bad.

    - If they are good, explain why.
    - If they are bad, explain why and what you should do to fix them or make them better.

Make sure your answer demonstrates that you have a sophisticated understanding of the issues involved.

a. "Ten percent of our customers did not complete a satisfaction survey. We can get most of their demographic information from an existing database, so we decided to complete our analysis substituting in the average response from other people for any missing values."
b. I was running a regression model – variable X1 worked fine until I added X2, then neither were significant, so I took X2 back out and left X1 by itself.
c. "The R-squared of a regression is generally a good indicator of how well the overall model fits the data."
d. "When developing a model from theory, it doesn't matter whether you start with a small model and build up, or a large model and build down, you get to the same place – somewhere in the middle – either way."

2. Many university and college students work during their summer vacations. One particular school was very interested in understanding student employment prospects. To develop insights, a confidential survey was conducted for second-year students and that data was linked to records on their academic performance and other demographic information for those who worked in the summer.

   You have been hired by the school's principal to analyze the data. In general he would like you to analyze student earnings in terms of their grade performance and demographic characteristics. He has provided you with a list of questions below that he would like to have answered by your analysis. For each question, he would like you to **do whatever analysis is required to answer the questions and provide an explanation in language he is likely to understand.** He encourages you to read all of the parts of the question first.

    a. Develop a model of student earnings based on the data on tab 'Students'. Explain why you chose the model you did and what the results tell you.
    b. Is heteroskedasticity an issue with this model? Regardless of your findings, assume that heteroskedasticity is not an issue for the balance of your analysis.
    c. The principal claims that summer earnings increase as grades in calculus increase. If so, he would like to offer free tutoring support to students to increase their calculus grades so that they will earn more. Does the evidence support the existence of such a relationship? If such a relationship existed, would it justify his strategy?

d. Is there evidence of any difference in performance between male and female students in terms of the relationship between at least some of their grades and their summer earnings?

3. A Kingston real estate company believes that the most important factor in selling a house is setting the initial price correctly. This company has hired you as a consultant to develop a model to predict housing prices so they can set initial prices accordingly. A collection of data has been provided to you on the tab 'housing'. The client insists that the model should include all the explanatory variables they have provided: Price = B0 + B1 N_Bedrooms + B2 N_Bathrooms + B3 House_Size + B4 Age + B5 Renovated_Kitchen + B6 Finished_Basement + B7 Close_to_Campus.
    a. In principle, how would you look for outliers in the data? If you found any how would you recommend that the client deal with them? Note: you do not need to actually do anything for this question – there are no outliers in the data.
    b. If there had been missing data elements, how would you recommend the client deal with that in their data? Note: you do not need to actually do anything for this question – there are no missing values in the data.
    c. Aside from the size variables House_Size and N_Bathrooms, do all the variables appear to belong? If not, which would you remove if the client did not insist on including them? What is the consequence of including irrelevant variables?
    d. The client believes that the variables House_Size and N_Bathrooms should both be important in determining the selling price of homes.
        i. Do these variables appear to belong in the model according to the individual t-tests?
        ii. Perform a test to see if any 'questionable' variables jointly belong in the model. What is the most likely explanation for the results?
        iii. If you ran the model without House_Size, what happens to the apparent significance of N_Bedrooms? Would this make for a better model?
    e. Suppose I want to sell my house. Does your first model suggest that I should add a bathroom at a cost of $3,000 and that doing so would be justified by the resulting increase in selling price? Be sure to justify your answer.

4. The tab 'Sales' contains 12 months of sales data for a sample of stores from across Canada, along with their location, prices, and advertising budget.
    a. Consider the simple linear regression model: $Sales = \beta_0 + \beta_1 Ad\_Budget + \beta_2 \text{Price}$ (i.e. ignoring the location and time data). Comment on the significance of Ad_Budget and Price in this model? Based on this model, can you say with confidence that an increase in price of one unit is associated with a reduction in sales of more than 400 units?
    b. A new corporate strategy came into effect in month seven. Is there any evidence that the relationship between Sales and Ad_Budget changes at that point going forward? Run an appropriate test and explain your results.

5. It is widely believed that the transit system's revenue has been trending up over the course of the year, subject to the fact that sales are normally lower in the months of July and August, however, there are those that believe that somewhere around the beginning of May, the rate of growth

increased significantly. Like most businesses, advertising budget and the cost of alternatives (i.e. gasoline for car commutes) should also play a role.

Using the data on the 'TransitData' tab, consider each of the issues below in isolation and construct a fairly complete test of each them alone (i.e. do not attempt to build a single model to test all of these things simultaneously.)

- That there is an upwards trend in transit revenue as a function of time and that this trend has increased since early May.
- Do advertising budget and gasoline each have the anticipated impact on transit revenue?
- Is there evidence that the impact of gasoline price and advertising budget are somehow different in the summer, controlling for the 'summer effect' with dummy variables?

6. Consider the data on the CANvUS tab, which has data on Canadian and American respondents. Use this data to answer the following questions.

   Build a model to explain Y using the X1..X4 variables and the US dummy.

   a. Does being a US respondent appear to explain the results?
   b. If being a US respondent does not explain the results, does that mean that Canadians and Americans are the same?
   c. Run a model to explain Y using X1..X4 then perform a test to determine if Canadians and Americans are in fact the same based or different. Explain the results.
   d. Without checking the data, what differences, in principle, could still exist between Canadians and Americans?