MMA 860 Assignment 2

**General Instructions**

This assignment consists of the following four questions and the data found in the workbook “MMA860\_Assignment2\_Data.xlsx”.

Please upload a single Word or Jupyter Notebook (as PDF) document that includes **your code, answers, and visualizations as required.**

When asked to explain something to your manager, assume that person has a non-technical background and requires a brief but complete explanation of the issue. Use this as an opportunity to demonstrate your thorough understanding of the material and your ability to communicate it effectively.

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. If his suggestions / comments are good ones, explain why; if they are bad ones, explain why and what you should do to fix them or do better. Make sure your answer demonstrates that you have a sophisticated understanding of the issues involved.

* 1. “If the p-value for the ‘joint F-test’ is too large – say greater than 0.05 – there is really no point in looking at the rest of the regression output. The model is just plain crap.” **3 marks**
  2. “When you detect an outlier, you should just replace it with the mean value of the dataset.” **3 marks**
  3. “When conducting a hypothesis test, like the t-test, you should always set alpha = 0.05.” **3 marks**
  4. “Our hotel did a customer satisfaction survey for all the guests who visited last weekend. The results were plagued by heteroskedasticity, so we basically had to scrap the results and start again.” **3 marks**

1. Your client has asked you to perform some analysis on the data found on the tab Missing.
   1. Explain in language that your manager is likely to understand how multiple imputation deals with missing values. **2 marks**
   2. Under what condition(s) could multiple imputation be used reliably to deal with missing values. Provide an original example (i.e. not ones that I have provided or that we have discussed in class) to illustrate when multiple imputation could reliably be used – if there are more than one condition, be sure to illustrate them all and describe how they apply in your example. **4 marks**
   3. If the conditions you describe above were not met, what else could you do? What might some problems / concerns be with such an approach? **2 marks**
   4. Estimate the model y = B0 + B1 X1 + B2 X2 + … B5 X5 using multiple imputation to correct for missing values. **2 marks**
   5. According to your results, does X2 belong in that model? Explain why / why not **2 marks**
2. I am interested in collecting wine, and typically buy Italian wine with a cork (because no good wine can come with a screw top, right?). I rely very heavily on expert ratings to determine which wines I buy. However, whenever a wine with a high rating is released, it immediately sells out. Despite many complaints, the LCBO does not want to help. Using the data set provided on historical wine ratings and their characteristics, answer the following questions with a model that predicts expert ratings. Note: you should read all parts of the question before answering, and build a single model to answer all 4 parts.
3. Present your final model and the estimated parameters. What steps did you go through to develop this model? **2 marks**
4. Does the data appear to be heteroskedastic? Why or why not? Show evidence. **2 marks**
5. Assuming there are no data problems, what would a wine be rated if it comes from France, has a price of $39.99, sulphates of 1.1, alcohol of 13.9%, residual sugar of 1.83 and a pH of 2.1? **1 mark**
6. Would increasing the price of a wine increase its expert rating? Be sure to clearly explain your thinking. **3 marks**
7. You might not know this about me, but I am pretty into ‘epic cooking’. I have, at times, cooked over 35 liters of curry in one batch (some for eating and some for preserving). I like to believe my curry is pretty good. In fact, I believe it is so good I may be able to sell it. I have simulated some data based on my understanding of the North American curry market in relation to price, advertising budget, how far away people live from the nearest retailer that might stock my curry, and what country people reside in. Of particular interest to me is whether Canadians and Americans feel differently about curry.
   1. Build a model to predict sales. Include your model results and explain your process. **2 marks**
   2. I believe that US consumers respond more to changes in Ad\_Budget than do Canadians. Conduct the most powerful test you can to see if this is the case. Briefly explain the test and your results (i.e. state the null, alternative, the p-value of the result and what it means.) **2 marks**
   3. Now conduct a Chow test to determine if Canadians are different from Americans. Briefly explain these results, particularly in light of your test above. **3 marks**