General Information

- **Keep your work to 4 decimals, unless otherwise stated**
- If you use any sources outside of the course (i.e., not from course notes or the textbook), it is expected that you properly cite them. See Outside sources for what is expected. If you are using citations directly in your work, then both in-text citations and the reference list need to be done in APA style: https://library.mtroyal.ca/citations

Model Building in Linear Regression

Background

With the onset of the tariff war between Canada and the US, prices of many commodities are expected to rise. New cars are one such commodity. But a downstream effect of the tariffs may also increase the price of used cars. The owners of a start-up—***AutoConnect***—are building an online business that connects used car buyers and sellers. They are exploring ways to accurately and competitively estimate the price of used cars. To this end, they wish to initially develop two models: one model will incorporate only categorical data such as make, model, colour etc. (A common approach to working with categorical data is logistic regression.) The other model will focus purely on quantitative data. Your job for this assignment is to create the best model using only the *quantitative data* on the included Excel file. You will be using a multiple linear regression model to accomplish this task.

Dataset and Data Description

Here is the data - Raw Data

Be careful: this dataset contains both categorical data and quantitative data. Before beginning, be sure to carefully identify the variables that qualify as quantitative data. (Do not include any categorical data in your model building.)

Goals of the Study

Your goal is to determine which quantitative features are most closely associated with the used car selling prices. Your model will then be incorporated into another model focusing on the categorical data as a basis for suggested asking/purchasing prices for both buyers and sellers of used cars.

> ***You have been tasked with doing the initial research, developing a good model, and writing up your findings in a report and creating a video presentation to AutoConnect.***

Steps to Complete the Multiple Regression Analysis and Model Building

For these steps, use ***VIF > 5*** to indicate multicollinearity**.**

- **Use the following template to set up your Excel work.** In the instructions below, this will be referred to as "Excel". The Excel template provides a general guide on how to set up your work. You may need to add additional sheets.
- **Use the following template to write up your analysis.** In the instructions below, this will be referred to as "Word". Ensure that the reader can understand your analysis without having to refer to your Excel document. That is, include tables of relevant values, visualisations (if relevant), and any other information you are using from the above analysis, explaining the model you have developed and chosen.

Training and testing data

Step 1. In Excel, divide original data into training (90%) and test (10%) data using a *simple random sampling technique*. Put your training data in the sheet entitled "Step 1_Training Data" and your testing data in the sheet entitled "Step 1_Testing Data".

- Random sampling was covered in the first unit of the course, so you may wish to review the technique before proceeding. This way each student will have a unique set of data. (Identical data sets between students will be considered as strong evidence of cheating). Support: Here is a video that shows you how to do a simple random sample (apologies for the dogs barking in the background). In case you want to follow along, here is the Excel file used in the video: Clear Mountain data(1).xlsx.

Choose the best model: Unless otherwise stated, use the training data to perform the analysis.

Step 2. In Excel, perform exploratory data analysis (EDA) on your training data by producing appropriate visuals and descriptive statistics, and by creating a correlation matrix. In Word, interpret the results of your initial EDA to support a discussion of your first impressions on the strength of the initial model and on which independent variables appear to look like promising candidates and why.

Step 3. Prior to building your model, choose the level of significance you want to use. When determining the level of significance, assume that all of the hypothesis tests are attempting to determine if the model is effective. In Word, state and support your choice by providing thorough reasoning for how you arrived at your choice. Utilize the level of significance for the remainder of the questions/steps.

Step 4. In Excel, explore multicollinearity for the promising candidates for the independent variables. In Word, discuss the process by which you reduced the model by eliminating independent variables that exhibit multicollinearity. Briefly discuss each iteration of the model, which independent variables were eliminated and why.

Step 5. Using the best subsets regression approach, in Excel, develop all possible models based on the remaining variables (excluding simple linear regression). Then use the root mean squared error (RMSE) and the adjusted R-squared statistics to compare the performance of your models. Ensure you check the *p*-values to ensure your best model only has significant independent variables. You may create a table here that summarises the RMSE and adjusted R-squared. Based on your work, decide on a final regression model. In Word, state the best model and support the decision based on the work you did in this step.

Step 6. In Excel and using your test data, predict the car prices using the regression equation you developed using the training data – this will tell you something important about the predictive quality of your final model.

Compare the predicted car price with the actual car price using RMSE (root mean square error).

Step 7. In Word, evaluate the best model by discussing both the accuracy and appropriateness of the best model, and by discussing how well the model worked on the test data.

The objectives of the Word document are:

- Can you demonstrate the ability to complete all the steps of the model-building process?
- Can you demonstrate a robust understanding of the meaning of the numbers from each step and how they are used to develop the best model?

Here are some tips on formatting and correctness in business writing:

- Include a space after each paragraph and before and after each heading.
- Do not indent your paragraphs.
- Do not double-space the summary.
- Pay careful attention to style and tone. You are writing to a board of executives.
- Use coherent and unified paragraphs---not point form or big chunks of rambling prose.
- Use grammatically correct sentences.
- Use correct word choice (note the difference between "then" and "than", for example).
- Use correct punctuation, including apostrophes (note the difference between "it's" and "its", for example)

Communicating the results of the analysis to AutoConnect through a video presentation

Step 8. Finally, communicate the story of your analysis to the owners of AutoConnect through **a 5-minute video presentation** and slide deck. Include the following (in this order):

- Introduce us to the problem.
- Provide some descriptive statistics to help us better understand the data.
- Then present the final model and give a brief overview of why the model is effective and appropriate to use.
- Then provide a thorough interpretation of the model that explains how AutoConnect can use the model to predict car prices for better pricing. This can include:
  - For each independent variable, how does the dependent variable change?
  - Examples of predictions with commentary on their accuracy.
  - Independent variables that contribute the most to the car price.
- Finally: Clearly communicate potential additional quantitative independent variables that could be explored to improve the model.
- After finishing your interpretation of the model, answer this more theoretical question: "Calculate the 90%, 95%, and 99% confidence intervals for a slope. What impact do these different confidence levels have on the margin of error? What does this help us understand about the model?" Choose one independent variable from your final model and focus on that slope to answer the question.

It is expected that you show *your face* during the video so that we can verify that it is you giving the presentation. If you are uncomfortable showing your face, please discuss this with your instructor immediately to discuss alternatives.

The objectives of the video are:

- Can you demonstrate that you can communicate the results of your analysis using business language (rather than statistical language)?
- Can you demonstrate that you can articulate "why should someone care?" about the results of your analysis?

Tips for a good slide deck:

- The slides should be mainly visual, so keep the text to a minimum. (The slides are not your script!)
- Avoid clipart and other gimmicks.
- Use a consistent font style and colour scheme.
- Keep your audience in mind. What are their needs and interests?

Tips for a good presentation:

- You can use Google Meet or similar software.
- Do plan and rehearse! Practise your pacing and your transitions.
- Use good software. Here are two options:
  - PC: https://www.ellenfinkelstein.com/pptblog/create-screenshot-and-screen-recording-in-powerpoint/
  - Mac: https://support.apple.com/en-ca/HT208721
- Ensure the audio is of good quality.
- Review the final product and do it again if it's not professional sounding and looking.

What do I submit? - Do not submit the link to Excel & Word - it is expected you submit the file itself.

1. Completed Excel following the template provided. Add additional tabs for Steps 4 and 5 as needed using the same naming conventions.
2. PDF following the template provided.
3. Video recording of your presentation. If you submit a link, ensure that you share it with your instructor. Failure to do so could result in a 0 for the video portion.
4. List of "Outside sources" or a disclaimer (see below)