DELIVERABLE 1

Due by Wednesday, September 17th 2025, 23:59.

This is a team deliverable. Work with your assigned team members only. Each team needs to upload its solutions to Canvas as a single PDF file (any team member can do so through their Canvas account). All team members should contribute to the deliverable. If a team member did not contribute, then their name should not appear on the cover page. Scripts should be written in R.

Problem 1 (25 points)

Cambridge Computers asked for your help in a project that seeks to better inform their pricing strategy for laptops. You have been provided a dataset with the following variables:

Variable	Definition
`InventoryID`	Inventory ID of laptop for internal stocking purposes
`Company`	Manufacturer of laptop: Asus, Dell, HP, or Lenovo
`TypeName`	Type of laptop: Gaming, Notebook, or Ultrabook
`GPU`	Provider of GPU: AMD, Intel, Nvidia
`Screen`	Screen size of laptop in inches
`Memory`	Laptop's RAM in GBs
`Weight`	Weight of laptop in KG
`Rating`	The average customer rating of the laptops from an online review site
`Price`	Price of laptop [in Euros]

Table 1: Variables and definitions for Cambridge Computers dataset.

You have been tasked with the job of building a good model to predict the price of a laptop using the variables provided. The training and test datasets are available on canvas: <**laptop_train.csv**> and <**laptop_test.csv**>. In this dataset, you should treat the variables `Screen`, `Memory`, `Weight`, `Rating` and `Price` as numeric variables.

Answer the following questions.

a) Build a good linear statistical model to predict the price of a laptop using the data provided. Outline the procedure you followed to arrive at your final model. Provide a screenshot of the R output of your final regression model.

b) Discuss which of the independent variables in your final model are managerially sensible to you, and which if any are worthy of further investigation.

c) Based on your model, which Laptop manufacturer has the highest effect on the price of the laptop? Which Laptop manufacturer has the smallest effect?

d) What is the out-of-sample R-squared of your model? Give a precise interpretation of it.

e) Consider a laptop with the following characteristics:

`InventoryID` = 950, `Company` = Asus, `TypeName` = Ultrabook, `GPU` = Intel, `Screen` = 15.6, `Memory` = 6, `Weight` = 3, `Rating` = 8

For this laptop, what is the probability that the predicted price will exceed 1,100 Euros using your final model? Please state clearly the assumptions you are making, if any, to arrive at your answer.

In linear regression models, an interaction term or effect between two predictors x1 and x2 denotes the product of these two predictors i.e. x1*x2. You can fit a linear regression model to response y using features x1, x2 and their interaction effect x1*x2 in **R** with the following function call:

```
lm(data = your_data, y ~ x1 + x2 + x1:x2)
```

f) Modify your model from part (a) by including an interaction term of `Memory` and `GPU`. What is the interpretation of the coefficient corresponding to this interaction effect? Comment briefly (at most 2 sentences) on how this new model differs from the part (a) model. [Make sure categorical variables are appropriately specified using **R**'s **`factor()`** function].

g) Modify your model from part (a) by including an interaction term of `GPU` with `Company`. What is the interpretation of the coefficient corresponding to this interaction effect? Comment briefly (at most 2 sentences) on how this new model differs from model in part (a). [Make sure categorical variables are appropriately specified using **R**'s **`factor()`** function].

Problem 2 (30 points)

Inference in statistical models is a crucial aspect of data-driven decision-making, particularly in industries such as insurance, where accurate risk assessment and pricing are essential. However, the misuse or misinterpretation of these models can lead to serious consequences, such as unfair pricing practices or misjudgment of risk exposure.

To better understand the factors affecting health insurance premiums, you have been provided with a dataset <**insurance.csv**> on Canvas containing information on 1,338 insured individuals. The dataset includes the following attributes for each policyholder: Age, Sex, BMI, Number of Children, Smoker status (Y/N), and Region, along with their corresponding insurance charges.

Your task is to analyze the dataset and build models to gain insights into the risk underwriting process and the relationship between policyholder attributes and insurance charges. This analysis will involve examining the distribution of insurance charges, assessing the fit of various probability distributions, and building a regression model to predict charges based on policyholder attributes.

a) Load the insurance dataset (**insurance.csv**) as a dataframe in **R** using **`read_csv()`**.

b) Calculate and report the mean and standard deviation of the insurance 'charges' variable using the **`mean()`** and **`sd()`** command.

c) Assuming the charges for the individuals follow a normal distribution with your mean and standard deviation from part b), use the **`pnorm`** function (in **R**) to determine the probability that a randomly selected policyholder has an insurance charge between $8,000 and $14,000.

d) Using the actual historical data, calculate the proportion of individuals with an insurance charge between $8,000 and $14,000. Does this align with the probability you got when

assuming the charges follow a normal distribution in part (c)? Provide justification as to why this is the case or not.

e) The insurance company is curious to explore different probability models for the insurance charges. Determine whether a normal, lognormal, uniform, or Cauchy distribution is a good fit for the insurance charges data using the **`fitdist()`** command. To run this command, install and load the **`fitdistrplus`** library using the following lines:

```
install.packages("fitdistrplus")
library(fitdistrplus)
```

Which distribution fits the best? Justify your choice. How can you refine your conclusion from part (d) in light of this?

[Hint: Consider using the **R** functions **`qlnorm, qunif,qnorm,`**… etc or **`plnorm, punif, pnorm,`** … etc]

f) Using the **`lm()`** command, build an appropriate regression model to predict the insurance charges using all of the given attributes of the insured (no variable selection is needed, no variable transformation is needed). Provide the regression output of the **`summary()`** command.

[Hint: If there are any categorical variables in your dataset, you can specify them using the **R** function **`factor()`**]

g) Interpret the regression coefficients of your model from part (f) in order to explain the insurance company's pricing strategy to a regulator. In particular, explain how the company is setting their charges as a function of each of the independent variables included in the model. For every variable in the model, interpret the coefficient sign and magnitude, as well as their statistical significance.

h) Due to new regulations, age is now considered a sensitive attribute and cannot be used in any predictive modelling. Re-fit your model from part (f) with this variable omitted. How does this change affect the model's coefficient of determination (R^2) evaluated on the training dataset? Comment on whether you think some age groups are more likely to be undercharged or overcharged when age is not considered as an independent variable.

i) In light of your findings in part (e), would you recommend doing a transformation to the response variable for part (f)? If yes, outline in one sentence how you can redo your analysis in part (f). (No need to carry out the full analysis here.)

Problem 3 (25 points)

In this problem we will be using the **`mtcars`** dataset in **R**, which contains data extracted from the 1974 Motor Trend US magazine and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). You can load the dataset in **R** using the command: **`data(mtcars)`**

Suppose we are interested in predicting mpg (miles per gallon) using a linear model with the independent variables

- hp (horsepower)
- wt (weight), and
- the interaction between hp and wt.

We will denote this as Model 1.

a) Fit Model 1 to the **`mtcars`** dataset and show the fitted model using the **`summary()`** command.

b) In Model 1, what is the interpretation of the coefficients of the main effects $hp$and wt? What is the interpretation of the coefficient of the interaction effect $hp * wt$?

Let$hp_c = hp - \overline{hp}$ and $wt_c = wt - \overline{wt}$ denote the mean-centered versions of the predictors hp, and wt respectively.

c) You now consider Model 2 that fits response mpg using the independent variables hp_c , wt_c and their interaction effect hp_c*wt_c. Fit Model 2 using **R**, show your output, and comment on how the coefficients of Model 2 compare to those of Model 1.

d) You now consider Model 3 that fits response mpg using only the interaction term hp_c*wt_c. In other words, Model 3 only considers the interaction effect but does not include the main effects hp_c , wt_c. Fit Model 3 using **R**, show your output, and comment on how the coefficients of Model 3 compare to those of Models 1 and 2.

e) Can you explain mathematically (algebraically) your findings in parts (c) and (d)? [Hint: expand out the expressions for the mean-centered variables.]

Problem 4 (20 points)

Let's go back to Problem 1 (Cambridge Computers). Instead of predicting the raw price, the management wants to determine whether a particular computer will be high-priced (expensive) or not. They consider the price of a computer **`high`** if it is 500 Euros or higher, and **`low`** if it is 499.99 Euros or lower. (Start by creating a new binary column **`high`** for both the test and train dataframes which is 1 if the price is 500 Euros or higher, and 0 otherwise). Our goal is to build a logistic regression model predicting whether the price is high based on the independent variables of Table 1.

a) Build a logistic regression model using the `train` dataset where you predict whether the price is high or low using the other independent variables, and show the output of the **R** summary command. Do not perform any variable selection.

Answer questions (b)-(g) using your answer for (a):

b) Which variables are significant in predicting the probability of a price's being high?

c) Comment on whether the significant variables of the logistic regression model are the same as the significant variables in the linear regression model from Problem 1.

d) For each significant variable, comment on if an increase in its value increases or decreases the probability of a computer's being high-priced. Comment on if this makes sense.

e) For which independent variables are the signs of the coefficients the same between the logistic regression model and the linear model? different?

f) Consider a laptop with the following characteristics:

- `inventoryID: 4096`
- `company:` Lenovo
- `typename:` Ultrabook
- `GPU:` Intel
- `screen: 8`
- `memory: 8`
- `weight: 4.2`
- `rating: 7`

Write down the equation for the probability of its being high-priced and calculate that probability.

g) Now, apply your model to the test dataset. Using a probability cutoff of 0.5, what is the accuracy of your model on the test dataset?