COMP6200 - Mid-Sem Exam

**QUESTION 1:**

How much gas oil is consumed by the family cars in Australia in a year? Use Fermi estimation to get an estimate of the number. Your estimate doesn't need to be precise. Document any assumptions that you make and identify any that might be debatable. (Try to look up as few quantities as possible -- we're more interested in the method than the figure you end up with.)

Write up to 200 words.

To begin, we consider each element that makes up for the calculation of the total gas oil consumption for all cars in Australia.

1. Number of households: As of 2021 (Qu et al., 2023), Australia has approximately 9.275 million households, which we will round up to 9.5 million to account for growth up until 2024.
2. Number of cars per household: Accounting for both families with one single car and more than two cars, we take an average of 1.5 cars per household.
3. Average fuel consumption per car: Although this variable is highly dependent on the type of car, as a rule of thumb, the range is from 6 to 10 liters per 100 kilometers (Morley, 2022). Thus, we use the range average, which consists of 8 liters per 100 kilometers.
4. Average kilometers driven per year per car: Once again, this can vary greatly, but let’s estimate about 10,000 kilometers per year for each car.

Finally, the total fuel consumption for all cars equals: 9.5 million households x 1.5 cars/household) x (10,000 km/car/year x 8 liters/100 km. This calculation yields approximately 1.14 billion liters consumed annually.

# **References (Q1)**

Morley, D. (2022, December 31). *What is average fuel consumption?* CarsGuide. https://www.carsguide.com.au/car-advice/what-is-average-fuel-consumption-88469#:~:text=However%2C%20as%20a%20rule%20of

Qu, L., Baxter, J., & Gorniak, M. (2023, July). *Population, households and families*. Aifs.gov.au. https://aifs.gov.au/research/facts-and-figures/population-households-and-families

**QUESTION 2:**

Is there a correlation between the amount of time spent exercising each week and a person's resting heart rate?

Explain your answer and the possible implications of such a correlation. What would it take to establish a causal relationship in this case and how is that different from just observing correlation?

Answer within 300 words

To begin, there is often a correlation between the amount of time spent exercising each week and a person's resting heart rate (RHR). Regular exercise tends to lower RHR over time. This correlation occurs because regular physical activity strengthens the heart, allowing it to pump more blood with each beat. Consequently, the heart does not need to work as hard while resting, which results in a lower RHR.

The significance of this correlation lies in its establishment of a relationship between health and fitness. A lower RHR is typically associated with better cardiovascular health and higher fitness levels. It suggests that the heart operates more efficiently, potentially decreasing the risk of cardiovascular diseases. Additionally, a lower RHR is frequently viewed as an indicator of overall fitness.

According to the Australian Bureau of Statistics (2023), employing a controlled study is considered the most efficient way for establishing causation between variables. Thus, a controlled experiment may need to be conducted in this case. In such studies, participants are randomly assigned to an exercise group or a control group to prevent biases. Comparing changes in RHR between these groups allows researchers to determine if exercising causes observable changes.

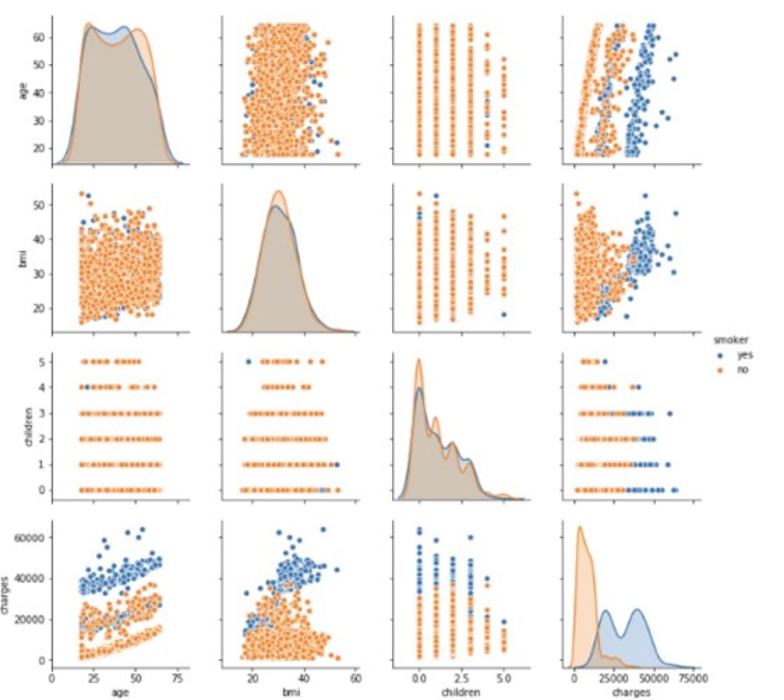
Finally, looking at correlation alone does not always indicate causation and vice versa. Although a correlation between exercise and RHR is suggested, it does not mean that one variable causes changes in the other. Other factors, such as diet, genetics, and overall lifestyle, could also influence resting heart rate. Therefore, establishing causation requires thorough experimentation to control for these potential confounding variables.

# **References (Q2)**

Australian Bureau of Statistics. (2023, February 2). *Correlation and causation*. Www.abs.gov.au. <https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/correlation-and-causation#:~:text=The%20use%20of%20a%20controlled>

**QUESTION 3:**

The pairplot below contains 5 variables: ‘age’, ‘bmi’, ‘children’, ‘charges’ and ’smoker’. The variable is a categorical variable with two responses ‘yes’ and ‘no’.



1. Comment on the shape of each variable

The distribution of age has no one distinct mode, and people from a wide range of ages are included in the dataset. Thus, the shape of this distribution is neither skewed nor symmetric.

Among these variables, the distribution of 'bmi' is the most symmetric, while the distribution shape of children shows positive skewness with a longer right tail, a pattern also observed in the 'charges' distribution. However, there is a difference in the shape of the 'charges' distribution depending on whether or not they are smokers. For people who are not smokers, the shape of their 'charges' distribution is significantly positively skewed, while for those who smoke, it is bimodal.

1. Which pairs of variables show a significant correlation

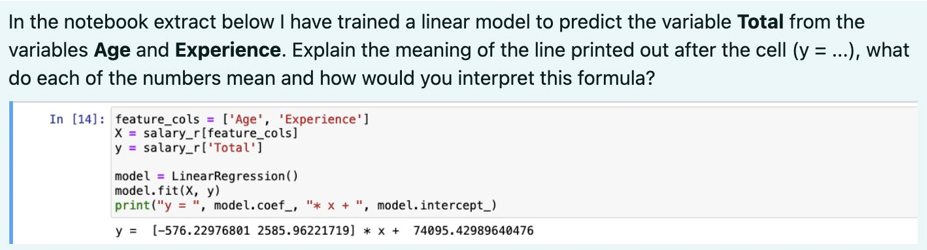
The 'age' of a person and their 'charges' are observed to be the most correlated among the presented pairs of variables. Firstly, a clear linear trend is observed between them, suggesting a linear relationship. Indeed, the data points between the 'age' and 'charges' variables in the scatter plot trend upwards from left to right, indicating a positive correlation between them. This means that if the age of the person increases, their 'charges' tend to increase as well.

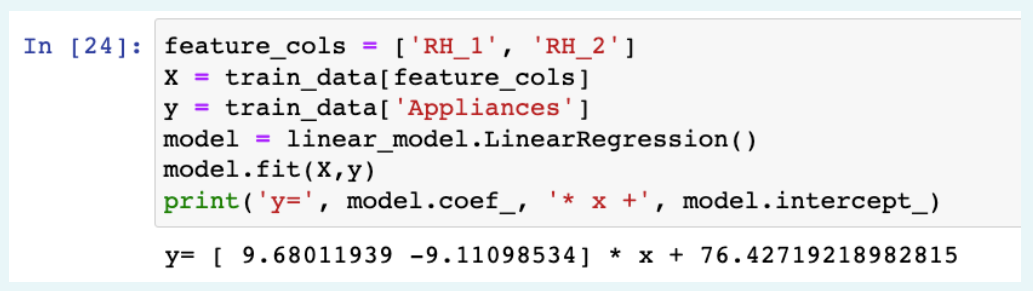
The correlation between ‘bmi’ and ‘charges’ might also be present for those who are smokers. However, it is not very clear for those who are not smokers.

1. If you were to build a model to predict the ‘smoker’, which variable would be the most useful? Why?

The most useful variable for predicting whether someone is a smoker would likely be 'charges'. The first reason is because of the difference in distribution based on smoking status. The 'charges' distribution exhibits a clear difference depending on whether individuals are smokers or non-smokers. For non-smokers, the distribution is significantly positively skewed, while for smokers, it is bimodal. This indicates distinct patterns of healthcare costs based on smoking status, making 'charges' potentially informative for predicting smoking behavior. Additionally, 'charges' may indirectly capture information about age, as indicated by its positive correlation with age. Since age is often associated with healthcare costs and may also be related to smoking behavior, 'charges' can potentially capture the influence of age on smoking status.

**QUESTION 4:**





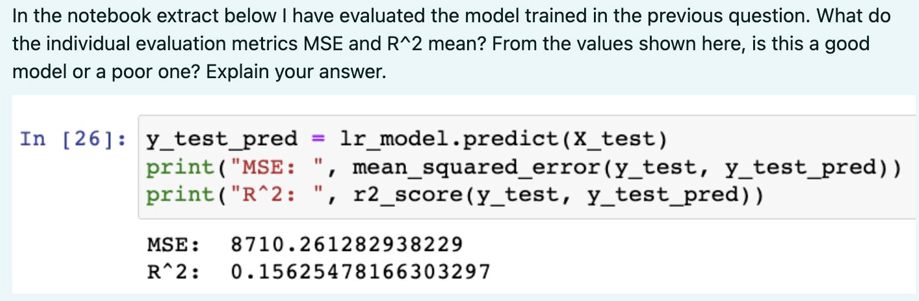
The line is the linear regression model built to predict the variable ‘Appliances’ from the variables RH\_1 and RH2. The equation of the model (with two decimals rounded up) can be rewrite as follows:

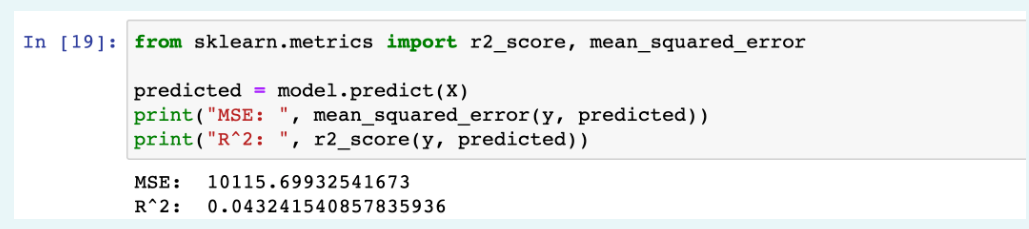
Predicted ‘Appliances’ = 9.68 x ‘RH\_1’ - 9.11 x ‘RH\_2’ + 76.43

In terms of the coefficients, we have (9.68 ) for 'RH\_1' and - 9.11 for 'RH\_2'. These represent the estimated change in the predicted value of 'Appliances' for a one-unit increase in the respective predictor variable, holding all other variables constant. Additionally, the coefficients can also indicate the strength and direction of the relationship between each predictor variable and the outcome variable 'Appliances'. A negative coefficient for 'RH\_2' suggests a negative correlation (holding 'RH\_1' constant), while a positive coefficient for ‘RH\_1’ indicates a positive correlation between it and ‘Appliances'’ (holding 'RH\_2' constant).

Regarding the intercept (76.43), it represents the estimated value of 'Appliances' when both RH\_1 and RH\_2 are zero. Thus, to predict the value of 'Appliances', we will input the values of 'age' and 'experience' into the equation.

**QUESTION 5:**





Mean Squared Error (MSE) is the average squared difference between the actual and predicted values of the target variable. A lower MSE implies better performance from the model. However, the interpretation of MSE also depends on the scaling of each context. In this case, the MSE value of ~ 10115.70 suggests that, on average, the squared difference between the actual and predicted values is approximately 10115.70.

Meanwhile, R-squared (R^2) represents the proportion of the dependent variable explained by the independent variables in the model. With its range from 0 to 1, a value closer to 1 indicates a higher proportion of the target variable explained by the model. In this case, the R^2 value of ~ 0.0432 indicates that approximately 4.32% of the variance in the target variable is explained by the model.

For this model, the relatively high MSE value suggests that the model's predictions have some degree of error, while the relatively low R^2 value indicates that only a small proportion of the variance in the target variable is explained by the independent variables in the model. Thus, based on these evaluation metrics, the model might not be considered very good.

Some Information Security Policy that Macquarie University Hospital can employ:

1. Ongoing risk assessments to mitigate system vulnerabilities (e.g. unauthorized access).
2. Patient data will need to be encrypted to prevent unauthorized access.

**QUESTION 6:**

A recent infographic circulated among the Macquarie University community, representing the average study hours of students across different faculties, suggests significant disparities in academic effort. However, upon closer inspection, the infographic employs misleading data visualization techniques (e.g. non-linear scales and selective data presentation).

Discuss the potential impact of misleading infographics on students and public perception of academic rigour across faculties.

Limit to 200 words.

To begin with, according to Bresciani and Eppler (2015), misinformation comes in two types: intentional, when it is done on purpose, and unintentional, when it is accidental. Regardlessly, it is undeniable that misleading infographics negatively impact the university by fostering false perceptions of academic effort and disparities across faculties. For students from undervalued faculties, it can lead to discouragement regarding their studies, indirectly affecting academic performance. Misleading information can also influence unfair funding allocations or policy decisions for the university's decision-makers. Additionally, it can foster unhealthy competition among university faculties, hindering collaboration and mutual support. Therefore, ensuring transparency and accuracy in data visualization is crucial due to its serious and unpredictable consequences. To do that, as Wolfe (2015) emphasizes, it is important to return to the data and reconsider the selection, summarization, and synthesis of data for creating compelling visual arguments.

# **References (Q6)**

Bresciani, S., & Eppler, M. J. (2015). The Pitfalls of Visual Representations. *SAGE Open*, *5*(4), 215824401561145. https://doi.org/10.1177/2158244015611451

Wolfe, J. (2015). Teaching Students to Focus on the Data in Data Visualization. *Journal of Business and Technical Communication*, *29*(3), 344–359. https://doi.org/10.1177/1050651915573944

1. Comment on the shape of each variable

The distribution shape of all present variables for ‘phi’ blue shows positive skewness with a longer right tail, while for ‘phi’ red it varies from one to another variable. Regarding ‘phi’ red, 'alpha' and 'beta' distribution shapes have negative skewness with a longer left tail, while for 'gamma' and 'delta' distribution shapes are mostly symmetrical.

1. Which pairs of variables show a significant correlation

The 'gamma' and 'delta' are observed to be the most correlated among the presented pairs of variables. Firstly, a clear linear trend is observed between them, suggesting a linear relationship. Indeed, the data points between the 'gamma' and 'delta' variables in the scatter plot trend upwards from left to right, indicating a positive correlation between them. This means that if ‘gamma’ increases, ‘delta’ tends to increase as well.

The correlation between ‘beta’ and ‘gamma’ might also be present.

1. If you were to build a model to predict the ‘smoker’, which variable would be the most useful? Why?

The most useful variable for predicting ‘phi’ would likely be 'gamma'. The first reason is because of the difference in distribution based on ‘phi’ status. The 'gamma' distribution exhibits a clear difference depending on whether ‘phi’ are red or blue.

For ‘phi’ red, the distribution is significantly positively skewed, while for ‘phi’ blue, it is symmetrical. This indicates distinct patterns of phi’ based on ‘gamma’, making 'gamma' potentially informative for predicting ‘phi’.

Additionally, 'gamma' may indirectly capture information about ‘delta’ and ‘beta’, as indicated by its positive correlation with both ‘delta’ and ‘beta’.