– Business Analytics – Trimester 2, 2025 Assessment Task 3 – Individual

DUE DATE: Monday, 29 September 2025, by 8:00pm (Melbourne time)

# PERCENTAGE OF FINAL GRADE: 40%

SUBMISSION: You will submit to unit site:

* one Excel file, with your analysis, and
* one Word file, with your written report

# Description

The assignment requires that you analyse a data set, interpret, and draw conclusions from your analysis, and then convey your conclusions in a written report. The assignment must be completed individually and must be submitted electronically in CloudDeakin by the due date. When submitting electronically, you must check that you have submitted the work correctly by following the instructions provided in CloudDeakin. Hard copies or assignments submitted via email will NOT be accepted.

The assignment uses the file 2025 T2 MIS171 Assignment 3 Data.xlsx which can be downloaded from CloudDeakin. The assignment focuses on materials presented up to and including Week 11. The Excel file which has been provided has different worksheets explaining and containing the Chronic Kidney Disease (CKD) dataset. Following is an introduction to this scenario and detailed guidelines.

# Context/Scenario: Chronic Kidney Disease (CKD)

Chronic Kidney Disease (CKD) is a major global health challenge, impacting millions of individuals each year. The condition is marked by a gradual loss of kidney function, which often remains undetected until it progresses to advanced stages. Early identification of risk factors is crucial, as it can help delay or prevent disease onset, minimize complications, reduce healthcare costs, and enhance patients’ quality of life. This case study explores how data analytics can be applied to uncover and understand CKD risk factors, with the aim of developing an effective model for early detection and prevention.

IMPORTANT: Considerable effort has been made to ensure this case study resembles realistic circumstances. However, it is NOT a formal or medically authoritative case-study. You MUST NOT draw any conclusions or take any action regarding your health or personal circumstances based on this case-study. If you, or anyone known to you, has concerns you must seek a suitably qualified medical professional.

## Problem Statement and Data Description

CKD doesn’t usually show symptoms in its early stages. By the time signs appear, permanent kidney damage may have already occurred. Traditional diagnostic methods rely on clinical evaluations and lab tests, which are often done after symptoms arise. The challenge is to proactively identify individuals at high risk by using available patient data before the disease advances. This case-study aims to pinpoint key risk factors for CKD from a dataset of 250 randomly selected Australians. The dataset includes both continuous (numeric) and categorical variables commonly linked to kidney function, general health, and lifestyle.

## Assignment Objective

This assignment is designed to engage your critical thinking, problem-solving, and analytical skills through the use of predictive analytics on the given dataset. The objective is to conduct a multiple linear regression analysis to explore the factors that potentially contribute to maximising CKD Risk. Building upon Assignment 1's interactive dashboard/data visualisation and Assignment 2’s descriptive analytics, your challenge is to explore the dataset to uncover meaningful insights and patterns that illustrate the progress made and challenges faced in enhancing CKD Risk.

The assessment task, accompanied by guidelines, highlighted in blue, are presented below. You are required to submit your Excel file containing your data analysis, along with a report that explains the outcomes of your analysis and two recommendations. Given that your audience may not have training in business analytics, your report must present the results in plain, straightforward business language. A template has been provided for your use.

## Multiple Linear Regression Modelling (consider α = 5%)

CKD Risk is an important measure for the Deakin University Hospital’s Renal Unit, as this will influence clinical advice and public education programs.

Build a multiple regression model to predict CKD Risk. Your model should provide insights into which factors have a significant influence on CKD Risk, as well as the ability to predict CKD Risk for various scenarios.

For this analysis, you will need to build a multiple regression model using the numerical variable CKD Risk as the dependent variable. All other variables in the CKD Risk dataset should be included in the model, except Patient ID. There is no need to create the categorical variable CKD Risk (i.e. exclude Patient ID and CKD Risk from your regression model).

Follow the model building process introduced in the lecture and seminars.

Carefully consider the following:

1. Transform categorical variables into suitable dummy variables

(i.e. Residence, Diabetes Mellitus, Hypertension, Alcohol Intake, and Physical Activity).

Copy the CKD Risk Dataset to the “Correlation” spreadsheet in the Excel file that has been provided (no earlier than Column AI - be careful not to overwrite the Conclusion, Correlation Table and Scatter Diagram frames).

* + 1. When transforming Residence into dummy variables, consider Rural as the baseline category.
    2. When transforming Diabetes Mellitus into dummy variables, consider No as the baseline category.
    3. When transforming Hypertension into dummy variables, consider No as the baseline category.
    4. When transforming Alcohol Intake into dummy variables, consider High as the baseline category; meaning the created dummy variables for Alcohol Intake should only include Moderate (Yes and No), and No Alcohol (Yes and No).
    5. When transforming Physical Activity into dummy variables, consider Active as the baseline category.

Complete the Dummy Variables Summary table which is in the Conclusion section of the Correlation worksheet. The table summarises the results of your transformation of categorical variables into dummy variables.

1. Using the CKD Risk dataset (which now includes transformed dummy variables) as your reference, complete the following steps:
   * 1. Correlation – in the section marked “Correlation Table” (below the Conclusion section on the “Correlation” worksheet) generate a correlation table. Use the “Correlation” option in Excel’s Data Analysis ToolPak.
     2. On the correlation table, identify and clearly indicate the Independent Variables which are (virtually) uncorrelated with the Dependent Variable (i.e. all IVs which have a correlation coefficient with the DV of between -0.050 and 0.050). These IVs are to be removed from the model prior to running the first iteration of the regression model. iii. Complete the Uncorrelated Independent Variables summary table which is in the Conclusion section of the Correlation worksheet. This table summarises which Independent Variables are to be eliminated from the regression model due to being (virtually) uncorrelated with CKD Risk (DV).
     3. Multi-collinearity - review the correlation table for instances of multi-collinearity between Independent Variables (IV). In cases of multicollinearity (please consider correlation between IVs greater than 0.7 or less than -0.7), identify and clearly indicate the IVs with the weakest correlation with the Dependent Variable.

These IVs are to be removed from the model prior to running the first iteration of the regression model.

* + 1. Complete the Multi-Collinearity summary table which is in the Conclusion section of the Correlation worksheet. This table summarises which Independent Variables are to be eliminated from the regression model due to multi-collinearity.
    2. Scatter diagrams - in the section marked “Scatter Diagrams” (below the Correlation Table section on the “Correlation” worksheet) generate three scatter diagrams, for:
       - CKD Risk (Dependent Variable, DV) and the numerical (not dummy categorical) Independent Variable (IV) which has the highest correlation with the DV. Include a calculation of the correlation coefficient. Format the diagram, and include a linear trendline, with the equation and the coefficient of determination.
       - CKD Risk (DV) and the numerical (not dummy categorical) Independent Variable (IV) which has the highest inverse (i.e. most negative) correlation with the DV. Include a calculation of the correlation coefficient. Format the diagram, and include a linear trendline, with the equation and the coefficient of determination.
       - CKD Risk (DV) and the numerical (not dummy categorical) Independent Variable (IV) that is closest to being uncorrelated with the DV (i.e. correlation coefficient closest to zero). Include a calculation of the correlation coefficient. Format the diagram and include a linear trendline, with the equation and the coefficient of determination.

1. On the “Regression Model” spreadsheet in the Excel file that has been provided (the data set includes the dummy variables you have created and excludes the Independent Variables which have been eliminated due to multi-collinearity or being uncorrelated with the Dependent Variable), complete the following steps:

i. Using the “Regression” option in Excel’s Data Analysis ToolPak build a multiple regression model.

 Assess the model for overall significance (F test with alpha set at 0.05, i.e.

Confidence Level = 95%).

ii. If your first iteration of the overall model is found to be significant, in a step-wise fashion, sequentially (one at a time) remove the Independent Variables that are least likely to be contributing to any significant change in the Dependent Variable.

 You will need to conduct t-tests (i.e. check p values) with alpha set at 0.05 to determine the significance of the various IVs you exclude and include in your model.

1. Once you have created a regression model where all the remaining Independent Variables are contributing significantly to a change in CKD Risk, copy the Summary Output of your final multiple regression model and paste it into the Output section of the “Regression

Model” spreadsheet in the Excel file that has been provided,

i. In the Conclusion section of the “Regression Model” spreadsheet,

* + - * + Write the (final) multiple regression equation.

Use the format: Ŷ = ƅ0 + ƅ1X1 + ƅ2X2…

* + - * + Explain (interpret) the (final) multiple regression equation/model.

1. Using the final multiple regression equation (from the previous step),

i. In the Predictions section of the “Regression Model” spreadsheet in the Excel file that has been provided, for the scenario outlined below:  Calculate a Point Estimate for CKD Risk (DV),

* + - * + Calculate a Prediction Interval for CKD Risk (DV),
        + Calculate a Confidence Interval for CKD Risk (DV),

ii. In the Conclusion section of the “Regression Model” spreadsheet in the Excel file that has been provided, for the scenario outlined below:

* + - * + Interpret the Point Estimate calculation
        + Interpret the Prediction Interval calculation
        + Interpret the Confidence Interval calculation

|  |  |
| --- | --- |
| Independent Variables | Scenario |
| Age | 50 years |
| Residence | Urban |
| Diabetes Mellitus | Yes |
| Blood Pressure | 120 mmHg |
| Hypertension | Yes |
| Blood Sugar | 5.0mmol/L |
| Blood Glucose | 260 mg/dL |
| Reticulocyte Count | 2.25 |
| Packed Cell Volume | 35 |
| Haemoglobin | 12.5 g/dL |
| Red Blood Cell Count | 4.75 million/cmm |
| White Blood Cell Count | 9,250 cells/cmm |
| Blood Urea | 3.25 mmol/L |
| Serum Creatine | 125 µmol/L |
| Alcohol Intake | Moderate |
| Physical Activity | Moderate |

## Data description

The provided Excel file includes multiple sheets, labelled “Data Description”, “CKD Risk Data” and several other worksheets for the above questions. The “Data Description” sheet describes all the variables used in the “CKD Risk Data” and is copied below for your convenience.

|  |  |
| --- | --- |
| Variable | Description |
| Patient ID | Unique identifier for each patient |
| Age | Age of the survey participant |
| Residence | Location of the survey participant: Rural or Urban |
| Diabetes Mellitus | Diabetes indicator present: Yes or No |
| CKD Risk Score | Risk of developing Chronic Kidney Disease (CKD Score from 0 to 100) |
| Blood Pressure (mm Hg) | Average blood pressure measured as millimeters of mercury |
| Blood Sugar (mmol/L) | The amount of sugar in the Participant’s blood, measured in millimoles per litre |
| Blood Glucose (mg/dL) | The amount of glucose in the Participant’s blood, measured in milligrams per decilitre |
| Reticulocyte Count (%) | Measurement of immature red blood cells in the bloodstream, measured as a percentage |
| Packed Cell Volume (%) | The percentage of red blood cells in the total blood volume |
| Haemoglobin (g/dL) | Grams of haemoglobin per deciliter of blood. |
| Red Blood Cell Count (million/cmm) | The number of red blood cells per unit of blood, measured as million cells per cubic millimetre of blood |
| White Blood Cell Count (cells/cmm) | The number of white cells in the blood sample. Measured as cells per cubic millimetre of blood |
| Blood Urea (mmol/L) | The amount of urea nitrogen (BUN) in the blood sample, measured in millimoles per litre |
| Serum Creatinine (µmol/L) | The amount of serum creatinine in the blood sample, measured in micromoles per litre |
| Hypertension | Whether/not the participant suffers from hypertension: Yes, or No |
| Alcohol Intake | The average amount of alcohol consumed by the Participant, classified as above average (High), below average (Moderate) and Abstainer (No Alcohol) |
| Physical Activity | Average amount of exercise taken by the participant, classified as above average (Active), average (Typical), and Passive (Inactive) |

Assignment instructions

The assignment consists of two parts.

## Part 1: Data Analysis

Your data analysis must be performed on the Assignment 3 Excel file. The file includes tabs (spreadsheets) for:

* Data Description
* CKD Risk Dataset
* Correlation, which includes:
  + creating dummy variables, o creating correlation table, o eliminating uncorrelated independent variables (IVs), and
  + eliminating IVs where multi-collinearity is present
* Regression Model – building the regression model, including multiple iterations, and o reporting the summary output of the final regression model, o identifying the final equation, and explaining/interpreting the final equation, and
  + calculating and explaining the point estimate, prediction interval, and confidence interval for the scenario provided.

When conducting the analysis, you need to apply techniques learnt in the lectures and seminars. The analysis section you submit should be limited to the Correlation and Regression Model worksheets of the Excel file. These are the only worksheets which will be marked. Your analysis should be clearly labelled and grouped around each question. Poorly presented, unorganised analysis or excessive output will be penalised.

In the Conclusion section of each worksheet there is space allocated for you to write a succinct response to the questions. When drafting your Conclusion, make sure that you directly answer the questions asked. State the important features of the analysis in your Output section. Responses in the Conclusion section will be marked.

Use the Output section for your analysis to complete the analysis as directed and supports your response to the questions (which you will write in the Conclusion section). Analysis in the Output section will be marked, please make sure your analysis and process complete, clear, and easy to follow. You may need to add (or widen/narrow) rows or columns to present your analysis clearly and completely. Poorly presented, disorganised analysis or excessive output will be penalised. It is useful to produce both numerical and graphical analysis. Sometimes something is revealed in one that is not obvious in the other.

Use the Workings section for calculations and workings that support your analysis. The Workings section will not be marked.

## Part 2: Report

Having analysed the data, including answers (in technical terms) to the Data Analysis questions from Part 1 you are required to provide a formal report. Your audience is the Board of the Deakin University Renal Unit. They are highly educated and intelligent medical professionals. However, they do not have training in business analytics. Your report must present the results in appropriate, straight-forward business language. The audience will only be familiar with broad generally understood terms (e.g. average, correlation, proportion, and probability). They will need you to explain more technical terms, such as dependent variable, independent variable, standard deviation, coefficient of determination, correlation coefficient, and confidence interval, etc.

In section 1 of the report, provide a brief interpretation of your findings of the Correlation and Regression analyses.

In section 2 of the report, Make TWO (2) recommendations that the Deakin University Renal Unit Board could consider when drafting clinical guidance (and/or patient advice) for minimising CKD Risk. Your recommendations should be based on analysis in this assignment, analysis from previous assignments, and any additional relevant analysis that enhances the impact of your recommendations.

Consider the following in framing your recommendations:

* Specific actions the DU Renal Unit could take to minimise CKD Risk based on the outcomes of your regression model.
* Specific actions the DU Renal Unit could take to minimise CKD Risk based on the outcomes of your analysis from Assignment 1 and Assignment 2.
* Specific actions the DU Renal Unit could take to minimise CKD Risk based on the outcomes of any additional analysis you perform.
* Recommending targeting a group that the DU Renal Unit could pursue that is exposed to CKD Risk.
* The impact of other important measures such as Hypertension, Red and White Blood Cell Count, Physical Activity, and Alcohol Consumption on CKD Risk.
* Considering the impact on CKD Risk of the variables not specifically included in your regression model.
* Recommending strategies for targeting specific variables that could significantly improve the CKD Risk Score.

Ensure that all your recommendations are directly informed by your data analysis. Do not include any commentary that is not supported by your data analysis.

Highest marks will be awarded to students who draft distinct (i.e. different) recommendations, and whose recommendations take into account a broad range of (data-supported) significant considerations.

When exploring data, we often produce more results than we eventually use in the final report, but by investigating the data from different angles, we can develop a much deeper understanding of the data. This will be valuable when drafting your written report.

It is useful to produce both numerical and graphical statistical summaries. Sometimes something is revealed in one that is not obvious in the other.

You are allowed approximately 1,000 words (950 to 1,050 words) for your report. Remember you should use font size 11 and leave margins of 2.54 cm.

A template is provided for your convenience. Carefully consider the following points:

* Your report is to be written as a stand-alone document.
* Keep the English simple and the explanations clear. Avoid the use of technical statistical jargon. Your task is to convert your analysis into plain, simple, easy to understand business language.
* Follow the format of the template when writing your report. Delete the report template instructions (in purple) when drafting your report.
* Do not include any charts, graphs, or tables into your Report.
* Include a succinct introduction at the start of your report, and a conclusion that clearly summarises your findings.
* Marks will be deducted for the inclusion of irrelevant material, poor presentation, poor organisation, poor formatting, and reports that exceed the word limit.

When you have completed drafting your report, it is a useful exercise to leave it for a day, and then return to it and re-read it as if you knew nothing about the analysis. Does it flow easily? Does it make sense? Can someone without prior knowledge follow your written conclusions? Often when rereading, you become aware that you can edit the report to make it more direct and clearer.