MIS770 – Foundation Skills in Data Analysis - Trimester 3 2020
Assessment Task 2 – Analysis of US Health Insurance data – Individual Assignment

DUE DATE AND TIME:	Thursday, 21st January 2021, by 8:00pm (AEST)
PERCENTAGE OF FINAL GRADE:	30%
WORD COUNT:	2,000 words or approximate *equivalent*

Description

The purpose of this assignment is to investigate a dataset using the knowledge learned in Modules 1 and 2. This will enable conclusions to be drawn that ultimately assist in decision making.

The assignment requires you to analyse a given dataset, interpret the results, and then draw conclusions such that you are able to reply to specific questions being asked of you in the form of a business report. (These questions are asked in the following email).

The aims of the assignment are to:

- provide you with some examples of the application of data analysis
- test your understanding of the material presented in the relevant topics
- test your ability to analyse data and interpret your results
- test your ability to effectively communicate your results to others

Before attempting the assignment, make sure that you have prepared yourself well by reading the relevant sections of the prescribed textbook and reviewing the materials provided in Modules 1 and 2 (i.e. Topics 1 to 7).

Specific Requirements

The UnitedHealth Group is America's most prominent health insurance provider. They want to better understand certain population characteristics that might contribute to the high medical costs being billed to insurance providers. They have access to a random sample of US Health Insurance data containing 1338 insured personnel with their Age, Gender, Body Mass Index (BMI), Number of Children, Smoking status, Region and Charges.

You are a Data Analyst working for UnitedHealth Group. Your Manager, Daisy Pearce, has asked you to conduct a preliminary analysis. In particular, you are expected to apply a series of statistical techniques and produce a report based on your findings.

Daisy's email is reproduced on the next page.

Email from Daisy Pearce

To: <Your name>

From: Daisy Pearce

Subject: Analysis of US Health Insurance data

Hi,

As per our conversation, I have spoken with our reporting team and we have THE following questions relating to the US health insurance data (contained in the file Insurance.xlsx). Please complete the required analysis and prepare a report for me containing answers to the following questions:

Q1. An Overall View of both "Charges" and "Smoking"

Can you provide me with overall summaries of

a) Individual medical cost billed by health insurance
b) Smoking status

Q2. Relationships

a) Is there a relationship between the age of the primary beneficiary, their body mass index (BMI), number of children and medical cost?
b) We would also like to know is there a gender bias in the smoking behaviour of the beneficiary.
c) Can you further analyse to see whether the beneficiary's residential area/region in the US affect how health insurance provider bill their medical costs?

I realise that the US Health Insurance data contain a random sample of 1338 insured personnel, and that this information can be used to draw inferences about the specific attributes of the whole insured population and charges billed by health insurance providers. With that in mind, Please provide me with answers to the following questions:

Q3. The UnitedHealth Group would like estimates of the following.

a) Average medical cost for an older beneficiary (older adulthood: 56 years and older)
b) Proportion of smokers who are obese (BMI of at least 30)

Q4. The UnitedHealth Group would like a comparison between this year's medical cost and the industry average.

a) The industry average medical cost for a single adult (i.e. without children) is at least $10,000. Is there any evidence to support this assertion?
b) Based on the industry average, less than 50% of beneficiaries are female. Can this claim also be substantiated?

Q5. Appropriate Sample Size

One of the company's overall goals is to estimate the average medical cost for all insured personnel to within $1000 (±1000) and the proportion of all insured smokers to within 3%, Will a sample size of 1338 be large enough? If not, what size sample should be taken? What other factors should be taken into account when sampling?

I look forward to your response on or before 21st January 2021.

Sincerely,

Daisy
Chief Data Scientist – UnitedHealth Group

Business Report Requirements

- Your report should be no longer than 4 pages and should not include any charts, tables, or appendices in the report. Charts/graphics and tables are only to be placed in the Data Analysis file i.e. the Excel spreadsheet and not reproduced in the report.
- Suggested formatting for the report: single-line spacing; no smaller than 10- point font; page margins approx. 25mm, and good use of white space.
- Your report must have a cover sheet containing your particulars and Unit details.
- The report is to be written as a stand-alone document (assume Daisy will only read your report). Thus, you should not have any references in the report to your data analysis output. Eg. “According to Table 1 in the analysis…”
- Your report must contain an executive summary that explains in plain language the purpose of the report and summarises the main findings. The executive summary should be no more than 300 words long.
- The body of your report must be set out in the same order as in the originating email from Daisy, with each section (question) clearly marked
- Use plain language and succinct explanations. Avoid the use of technical or statistical jargon as Daisy cannot be expected to understand statistical terminology. As a guide to the meaning of “Plain Language”, imagine you are explaining your findings to a person without any statistical training (e.g. someone who has not studied this unit). What type of language would you use in this case?
- Marks will be lost if you use unexplained technical terms, irrelevant material, or have poor presentation/ organization
- All Microsoft Excel data analysis output associated with each question in the Email are to be placed in the corresponding tab in the T22020MIS770_A2_yourstudentid.xlsx file

Data Analysis Instructions/Guidelines

In order to prepare a reply to Daisy’s email, you will need to examine and analyse the dataset Insurance.xlsx thoroughly.
Daisy has asked a number of questions and your Data Analysis output (i.e. your charts/tables/graphs) should be structured such that you answer each question on the separate tab/worksheet provided in your Excel document. There are also three extra tabs in Insurance.xlsx called CI, HT and SampleSize and you should use the various templates contained in these tabs arriving at your “Confidence Interval”, “Hypothesis” and “Sample Size” answers.

Q1. An overall summary of Charges (in dollars) and summary of Smoking status
You are required to comprehensively describe the variable ‘Charge’ by itself and the variable ‘Smoking’ by itself using the most appropriate techniques from Module 1.
Your analysis should include numerical summaries, graphs and tables. The importance of other variables is considered in other questions. You should thoroughly investigate relevant summary measures (and their reliability) for these two variables. Also, there may well be suitable tables and charts/graphs that will illustrate more clearly other important features of charges and smoking. (See Topics 1-3 learning materials)

Q2. Descriptive measures and insights
Your course notes (Module One) give methods (numerical summaries/tables/graphs/charts) for summarising a single variable and investigating the relationships (dependencies) between two variables for these situations. For example
• Pie/Bar charts
• Summary/Frequency Distribution tables
• Comparative summary measures including quartiles and percentiles
• Scatter diagrams
• Coefficient of correlation, r value
• Contingency tables/Cross tabs
• Stack bar charts, side-by-side bar charts
• Histograms/Frequency polygons/Ogives
• Single/Multiple box and whisker plots etc. (See Module One learning materials)
Use whatever techniques you have studied in Module 1 to investigate the associations/relationships. Generate suitable visualisations (Tables/Graphs/Charts) and numerical measure(s) demonstrating the existence or otherwise of a relationship. Remember to provide a brief overall summary when concluding these questions.

Q3-Q4 The analysis required involves inferential statistics, which are covered in Module 2.
Use the relevant Excel templates (CI and HT) provided in the Data file.
These questions will require you to complete either a confidence interval or a hypothesis test. Go through each of the questions asked by Daisy and decide which technique is the most appropriate. Below are some hints regarding the most appropriate technique:
• Do we have to make an estimate, and therefore need a confidence interval?
• Are we testing a theory/claim/ or comparing values… and therefore need a hypothesis test?
So decide which you think is the most appropriate technique (tutorials for topics 6 and 7 help here).
• You can assume that a 95% confidence level is appropriate.
• Use 5% significance in any hypothesis tests you perform, and provide a summary of your conclusions.

Q5. Use the relevant Excel templates provided in the Data file.
• You can assume that a 95% confidence level is appropriate.

Note: There is an Appendix at the end of each Chapter of the Prescribed Textbook which describes the basic Excel steps associated with that Topic. Chapters 1 to 9 are applicable for this assessment.

Other Guidelines:

- To answer some questions, you may need to make certain assumptions about the data set we are using. Mention these in your data analysis, where relevant. There is no need to mention this in the report.
- Please ensure you analyse the data thoroughly but do not go beyond what the question asks – for example, if one question requires comparison for the Gender classification, it does not mean you must do it for any other question, unless specifically asked to do so.
- We assume you will be using Excel to perform you data analysis.
- Detailed algebraic responses are not expected. Thus, avoid including extensive derivations, formulae, etc. You are to use Excel where possible to complete your answers.
- In your data analysis output you may include, as annotations, appropriate comments in either plain or technical language. These will be assumed to be your personal working comments.
- Overall, you will generate a great deal of output for this assignment, but you should trim it down to only show the most relevant results. Superfluous output will be penalised, so ensure you only include relevant materials (graphs, tables, calculations, tests, etc.) that are essential for writing up your report.
- Any of your computer work that is not useful should be discarded – your data analysis should only include computer output that is relevant to your report.
- Save your computer analysis frequently (every 10 to 15 minutes).

Learning Outcomes

This task allows you to demonstrate achievement towards the unit learning outcomes. The ULOs are aligned with specific graduate learning outcomes – that is, the skills and knowledge graduates are expected to have upon completion of their studies – and this assessment task is an important tool in determining achievement of those outcomes.

If you do not demonstrate achievement of the unit learning outcomes, you will not be successful in this unit.

It is good practice to familiarise yourself with the ULOs and GLOs as they provide guidance on the knowledge, understanding and skills you're expected to demonstrate upon completion of the unit. In this way they can be used to guide your study.

Unit Learning Outcomes (ULO)	Graduate Learning Outcomes (GLO)
ULO2: Manipulate and summarise data that accurately represents real world problems **ULO3:** Interpret and appraise statistical output to assist in real-world decision making	**GLO4:** Critical thinking: evaluating information using critical and analytical thinking and judgment

Submission

You are to submit your assignment in the individual Assignment Dropbox in the MIS770 CloudDeakin unit site on or before the due date.

Your completed assignment should be submitted in two separate files:

- Business report (Part A): A Word document of no more than 4 pages that is **not** to contain any charts/tables/graphs. (Note: Do not submit a pdf document in lieu.). Please name your Word document T32020***MIS770_A2_yourstudentid.docx***
- Data Analysis (Part B): An Excel document containing separate tabs/worksheets with charts/tables/graphs for each question. Please note that all interpretations should be presented in your "Business Report" and the Excel document should only contain your intermediate analysis and final output. Please name your Excel document T32020***MIS770_A2_yourstudentid.***xlsx

Please ensure that you include your name and student details in your Word document as well following the above file naming convention. Failure to follow this convention may lead to a delay in receiving feedback and marks.

Submitting a hard copy of this assignment is not required.

You must keep a backup copy of every assignment you submit, until the marked assignment has been returned to you. In the unlikely event that one of your assignments is misplaced, you will need to submit your backup copy.

Any work you submit may be checked by electronic or other means for the purposes of detecting collusion and/or plagiarism.

When you submit an assignment through your CloudDeakin unit site, you will receive an email to your Deakin email address confirming that it has been submitted. You should check that you can see your assignment in the Submissions view of the Assignment Dropbox folder after upload, and check for, and keep, the email receipt for the submission.

Marking and feedback

The marking rubric for this task is below and also available in the MIS770 CloudDeakin unit site - in the Assessment folder (under Assessment Resources).

It is always a useful exercise to familiarise yourself with the criteria before completing any assessment task. Criteria act as a boundary around the task and help identify what assessors are looking for specifically in your submission. The criteria are drawn from the unit's learning outcomes ensuring they align with appropriate graduate attribute/s.

Identifying the standard you aim to achieve is also a useful strategy for success and to that end, familiarising yourself with the descriptor for that standard is highly recommended.

Students who submit their work by the due date will receive their marks and feedback on CloudDeakin 15 working days after the submission date.

Extensions

Extensions will only be granted for exceptional and/or unavoidable circumstances outside the student's control.

Students seeking an extension for an assignment prior to the due date should apply directly to the Unit Chair by completing the Assignment and Online Test Extension Application Form. Requests for extensions will not be considered after 12 noon, Thursday 21st January 2021. Applications for special consideration after the due date must be submitted via StudentConnect.

Late submission

The following marking penalties will apply if you submit an assessment task after the due date without an approved extension: 5% will be deducted from available marks for each day up to five days, and work that is submitted more than five days after the due date will not be marked and will receive 0% for the task.

'Day' means working day for paper submissions and calendar day for electronic submissions. The Unit Chair may refuse to accept a late submission where it is unreasonable or impracticable to assess the task after the due date.

Calculation of the late penalty is as follows: *this is based on the assignment being due on a Thursday at 8:00pm*

- 1 day late: submitted after Thursday 11:59pm and before Friday 11:59pm– 5% penalty.
- 2 days late: submitted after Friday 11:59pm and before Saturday 11:59pm – 10% penalty.
- 3 days late: submitted after Saturday 11:59pm and before Sunday 11:59pm – 15% penalty.
- 4 days late: submitted after Sunday 11:59pm and before Monday 11:59pm – 20% penalty.
- 5 days late: submitted after Monday 11:59pm and before Tuesday 11:59pm – 25% penalty.

Dropbox closes the Tuesday after 11:59pm AEST time.

Support

The Division of Student Life (see link below) provides all students with editing assistance. Students who wish to take advantage of this service must be organized and plan ahead and contact the Division of Student Life in order to schedule a booking, well in advance of the due date of this assignment.
http://www.deakin.edu.au/about-deakin/administrative-divisions/student-life

Referencing

Any material used in this assignment that is not your original work must be acknowledged as such and appropriately referenced. You can find information about plagiarism and other study support resources at the following website: http://www.deakin.edu.au/students/study-support

Academic misconduct

For information about academic misconduct, special consideration, extensions, and assessment feedback, please refer to the document ***Your rights and responsibilities as a student*** in this Unit in the first folder next to the Unit Guide in the Resources area of the CloudDeakin unit site.

Marking Rubric

	Poor	Needs Improvement	Satisfactory	Good	Very Good	Excellent
Executive summary **(Marks: 10)**	0 points Does not communicate any of the main findings of the analysis in an accurate or useful way, or the findings are basic. 0 – 2.9 Marks	3 points Explains some main findings of the analysis accurately and enables reader to draw a few conclusions. 3 – 4.9 Marks	5 points Explains most of the main findings of the analysis accurately and enables reader to draw some reasonable conclusions. 5 – 5.9 Marks	6 points Explains nearly all of the main findings of the analysis accurately and enables reader to draw mostly reasonable conclusions. 6 – 6.9 Marks	7 points Provides detailed and accurate descriptions of the most important features of the analysis along with appropriately qualified conclusions. 7 – 7.9 Marks	10 points Provides outstanding descriptions and reaches conclusions that are carefully considered and insightful. 8 – 10 Marks
Data Analysis **(Marks: 40)** *This part relates to the various visualisations in the form of charts, tables & graphs etc. created by Ellyse which formed the basis of her response to Daisy.*	0 points Uses irrelevant or inappropriate techniques to analyse the data, or the Data Analysis and visualisation tools were used to analyse the data but in an incomplete or inaccurate manner. A very poor presentation of the analysis, or the analysis does not follow principles of good graphical display. 0 – 15.9 Marks	16 points Uses some appropriate data analysis and visualisation tools to analyse the data but there are many errors in the analysis. The presentation of the analysis needs improvement. 16 – 19.9 Marks	20 points Uses appropriate data analysis and visualisation tools to analyse the data but there are several errors in the analysis. The presentation of the analysis is satisfactory. 20 – 23.9Marks	24 points Uses appropriate data analysis and visualisation tools to analyse the data but there are some errors in the analysis. The presentation of the analysis is of a respectable standard. 24 – 27.9Marks	28 points Comprehensive analysis of the data using appropriate techniques, but there are some minor errors in the analysis. Uses data visualisations to understand the patterns in data. The analysis is well organised and follows principles of good graphical display. 28 – 31.9Marks	40 points Skilful and comprehensive analysis of data using many different techniques. Uses data visualisations to produce novel insights. An excellent presentation of the analysis. 32 – 40 Marks
Business Report **(Marks: 40)** *This part is the written response by Ellyse to the questions posed by Daisy.*	0 points Does not communicate any of the main findings of the analysis in an accurate and/or useful way, or the interpretation and communication of findings is at a basic level. The written communication is unprofessional or difficult to follow and contains numerous errors. 0 – 15.9 Marks	16 points Explains some of the main findings of the analysis accurately which only enables the reader to draw a few reasonable conclusions. The written communication is not very easy to follow and/or it contains too many errors. 16 – 19.9 Marks	20 points Explains most of the main findings of the analysis accurately and enables the reader to draw several reasonable conclusions. The written communication is clear and easy to follow but it contains minor errors. 20– 23.9 Marks	24 points Explains nearly all of the main findings of the analysis accurately and enables the reader to draw mostly reasonable conclusions. The written communication is clear and easy to follow and generally free of errors. 24 – 27.9Marks	28 points Provides detailed and accurate descriptions of the most important features of the analysis along with appropriately qualified conclusions. The written communication is professional, easy to follow and has a good structure. 28 – 31.9Marks	40 points Provides outstanding descriptions and conclusions that are carefully considered and insightful. The written communication is very professional, logical and easy to follow. 32 – 40Marks

Overall Assignment Presentation (Marks: 10)	0 points No attempt has been made to follow the assignment Requirements/ Instructions/ Guidelines. Poorly presented 0 – 2.9 Marks	3 points Little attempt has been made to follow the assignment Requirements/ Instructions/ Guidelines. Unsatisfactorily presented 3 – 4.9 Marks	5 point Majority of the assignment Requirements/ Instructions/ Guidelines have been followed. Satisfactorily presented 5 – 5.9 Marks	6 point Nearly all of the assignment Requirements/ Instructions/ Guidelines have been followed. Good presentation 6 – 6.9 Marks	7 points All of the assignment Requirements/ Instructions/ Guidelines have been followed. Very good presentation 7 – 7.9 Marks	10 points All of the assignment Requirements/ Instructions/ Guidelines have been dealt with meticulously. Faultless assignment presentation 8 – 10 Marks