DATA SCIENCE 141

A1: Take home assessment

2023

IMPORTANT INFORMATION

Timing and due date

The assessment data and questions are available from **Wednesday 30 August 2023 at 07:00**. You should work on this assessment in your own time. Completing the assessment should not take much more than three to four hours of your time **if** your R skills are good, but please allow yourself much more time than this and do not leave it until the last minute. Make provision for the fact that the assessment could potentially take you much longer than three hours to complete if you struggle with R and must consult notes / help files or obtain help from a teaching assistant / lecturer. You should also allow for the fact that you can run into unforeseen computer problems or encounter errors and in this case will need sufficient time to deal with it. Unforeseen events such as illness or other commitments might also impact the time you have available to work on the assessment. You should therefore not leave the assessment until the last day. **The due date for the assessment is Tuesday 19 September at 12:00 (midday).** Please note that very strict late submission penalties will be applied (refer to the module framework for details of these) and **no** exceptions will be made. Also remember that submission of A1 is compulsory, and that you therefore cannot pass the module if you do not submit this assessment.

Required submissions

For the assessment, you will have to work in R Markdown. You must submit your Markdown file (.Rmd) as well as the rendered HTML version thereof; both these files must be uploaded to SUNLearn. Note that for the SUNLearn submissions, only a single attempt is allowed. This means that, once you have submitted your assessment via SUNLearn, you will NOT be able to go back and change the uploaded documents. It is therefore your responsibility to make sure that you are satisfied that everything is complete and correct before you submit your files on SUNLearn. Please also note that after uploading the files, you are required to Submit the files too (you will see a Submit button after the files have been uploaded); if not, the files will be regarded as Draft Submissions on SUNLearn and the assessment will not be regarded as Completed.

Help available

Detailed instructions are provided in this document for the completion of important preliminary steps (such as importing the data and taking a sample of the data). Help can be sought from the lecturer or teaching assistants with the *preliminary steps* if required – but within reason. If any of the *assessment questions* are unclear, clarification can be sought from the lecturer on the module discussion forum on SUNLearn only. Specific help (such as assistance with the correct R code to use) will not be provided. Everything needed to successfully complete the assessment was covered in the practicals and tutorials.

Additional information

All data analysis and visualisations must be done in R. Plots and summaries created using other software packages will not be marked.

PLAGIARISM

This take-home assessment constitutes one of the formal assessments for the module Data Science 141. While you are allowed to work on the assessment in your own time and may refer to notes and other sources as described below, it is important that you must work as an individual and may not obtain help from any individuals or sources other than those specified below. You may also not upload the assessment paper / questions or data online and may not ask for help with the assessment on any online forums either – this implies that you also may not use ChatGPT or other AI sources to assist you with the completion of the assessment.

Your assessment submissions should be entirely your own work. You will be required to submit a signed plagiarism declaration (see Appendix 1 to this document; also available as a stand-alone form on SUNLearn). The signed plagiarism declaration can be uploaded to SUNLearn or a printed copy can be submitted to the lecturer in person by the due date. **Failure to submit a signed plagiarism declaration form means that your assessment will not be marked.**

Please take note that contravention of the academic honesty policy could lead to disciplinary action.

Allowed sources:

- Class notes.
- Lecture material (including practicals and tutorials).
- Textbooks and similar material.
- Online sources such as R help files.
- Help with preliminary steps may be sought from teaching assistants / lecturer only.
- Clarification of assessment questions may be sought online on the module discussion forum only.

What is not allowed?

- You are not allowed to seek help from anyone with any of the assessment questions, including your classmates, teaching assistants or lecturer.
- You are not allowed to seek direct help online. This means that you may not post the assessment dataset or any of the questions online on public forums in order to obtain assistance.
- You may not use ChatGPT or similar AI sources to generate code and / or answers to the assessment questions.

ADVICE / TIPS FOR SUCCESSFUL COMPLETION OF THIS ASSESSMENT

1. Start by carefully reading through this whole document – including the questions – before you start working with the data.

2. Once you have downloaded the dataset and imported it into RStudio, you should spend sufficient time exploring and getting to know the data before you start working on the assessment questions.

3. The preliminary steps are important, as they are there to ensure that you are doing the basics right before you start answering the questions. You should use the "Preliminary check" link on SUNLearn to help you ascertain whether you are on the right track. If you struggle with any of these preliminary steps, please ask for help as soon as possible. Remember that the week before the assessment is due, is the recess week, so there will be no practical lectures or tutorial sessions during which you can ask for help in person. The practical lectures on 30 and 31 August 2023 will be used to explain the basics of the assessment, so the tutorial session on 1 September, the practical lectures on 6 and 7 September 2023 as well as the tutorial session on 8 September 2023 can be used as opportunities to ask for help with the preliminary steps in person.

4. In the discussion type questions, make sure that you think about the domain / business context of the data. Remember that data science is more than just producing plots and building models. So if you create plots, think carefully about the information you want to convey in your plot and how best to achieve that. Also think about what makes sense in the context of your data. You are not being assessed purely on your practical R skills, but also on how you can "think" with data.

5. Many of the assessment questions do not necessarily have only one correct answer. There can be several suitable ways to visualise certain data for instance; in terms of the discussion, your own interpretation / opinion is important. Therefore, make sure that you are clear in describing any conclusions you draw from the data and your analyses thereof, and always substantiate your answers with numbers and / or plots as required.

6. Use suitable headings so that it is clear what question is answered by each piece of code and output. Interpretations of plots and statistics should follow after the relevant code and output. If it is not clear what question is being answered by the code / output, it will not be marked. A mock template has been provided for you on SUNLearn (file `A1 mock template`) to show what is expected in terms of question headings.

7. Think carefully about the output you want to produce with your code. It should be clear from the rendered version of the file what the answer to a particular question is. You shouldn't produce lots of output and expect the assessor to find the relevant piece of output to answer a question. As an example, if you are asked to calculate an average price, it will not be sufficient to simply produce the output of the `summary()` function to answer the question. Be purposeful about the code you are writing, and make sure you give sufficient interpretations and discussions.

8. Make sure your plots have suitable titles and labels. You should pay attention to the interpretability of your plots. Awarding of marks for plots will take this into account.

START OF ASSESSMENT

Background scenario

Usually, when the owner(s) of a house would like to sell the property, they will contact an estate agent. Often, the owner(s) themselves will not have a realistic idea of what the property is worth or could potentially sell for, so they would depend on the agent to suggest an asking price for the property. The estate agent would view the property and suggest a suitable selling price.

To suggest a selling price for a property, the estate agent would typically rely on his / her experience and knowledge, including what similar houses in the area have sold for. The agent will also take property market movements into account. However, there can still be considerable variation in the price different estate agents would assign to the same property.

Underestimating what a property is worth could have negative consequences for the seller, as the property could then sell for much less than it is actually worth. Overestimating the price of a property could also have negative consequences, as this could mean that a property struggles to sell, or that a buyer overpays for the property. It would therefore be beneficial for estate agents to have a better understanding of what the underlying features of a property are that could influence its selling price.

A dataset consisting of almost 3000 properties in a certain area was constructed. This dataset contains the selling price of each property, together with measurements of many different features of each property.

You – as a data scientist – will explore this dataset in R, with the aim of understanding what the different factors are that could potentially influence the price of properties.

The data

In a certain area, sales were recorded for the period from January 2016 to July 2020. For each property, the agent responsible for the sale recorded the price the property sold for, together with many other features of each property. The variables included in the dataset are as follows:

Variable name	Variable description
`Identifier`	This is a unique database identifier, allocated internally to each record. *These identifiers are not allocated sequentially.*
`Price`	Price (in Rands) for which the property was sold
`Price_cat`	Categorisation of `Price` into one of three possible categories: `Low` = selling price less than R1.3 million. `Average` = selling price between R1.3 million (inclusive) and R2.1 million (exclusive). `High` = selling price of R2.1 million or more.
`Month_sold`	Month in which the property sale was finalised. Months are coded by number (e.g. January is represented by the number 1 and December by the number 12).
`Year_sold`	Year in which the property sale was finalised.
`Sale_type`	Sales were classified as "Normal" or "Other". `Normal` sales are sales which were concluded in the typical way; in other words, an existing property was sold by the current owner to a new owner. `Other` represents all other sales and include scenarios where a brand-new home was sold (i.e. a house that did not have a previous owner), transfers of properties between family members, and properties that were sold on auction due to foreclosures.
`Type`	Description of the type of house on the property. `Single storey` = single level house. `Double storey` = multiple level house. `Estate` = house in an access-controlled estate. These houses can be single storey or double storey. `Duplex` = multiple family homes. Typically these would be semi-detached homes, or townhouses, where different floors are occupied by different owners, or homes share a common wall.
`Area`	Indication of the type of area the property is located in. Since the dataset contains house sales, most of the properties are in residential areas, but some are also included in other areas. `Residential` = property falls in a residential zoning area. Also includes properties in mixed-use developments. `Commercial` = property falls in a commercial zoning area. `Industrial` = property falls in an industrial zoning area. `Agricultural` = property is located in an agricultural area.
`Size`	Size of the erf (plot), in square meters.
`Street_front`	Measurement of the street frontage of the property, in meters.
`Plot_shape`	Shape of the erf (plot): classified as `Regular` or `Irregular`.

Variable name	Variable description
Incline	Position of the house relative to street level. Flat = house is level / near level with street. Up = position of house is significantly higher than street level. Down = position of house is lower than street level.
SlopeType	Slope of the erf (plot). Classified as Gentle, Moderate or Severe.
Condition	A rating of the overall condition of the house. Measured on a scale of 1 to 10, where 10 represents Excellent and 1 represents Extremely poor. This rating is assigned by the estate agent.
Year_Built	Year in which the house was initially constructed.
Year_Renov	Year in which the house was last renovated. This only includes major renovations, and not minor renovations. If this value is the same as the year the house was built, it means that the house has never been renovated.
Water	Whether the house has a municipal water connection. 1 means the house is connected to municipal water supply; 0 means it is not.
Tarred_road	Whether the street on which the property is located is tarred or not.
Culdesac	Whether the property is located in a cul-de-sac. 1 = property is located in a cul-de-sac; 0 = property is not located in a cul-de-sac.
Corner	Whether the house is situated on a corner plot (coded as 1) or not (coded as 0).
Roof_Type	Whether the house has a traditional roof or flat roof.
Bathrooms	Number of bathrooms. (Half bathrooms are bathrooms that don't contain a shower and/or bath.)
Bedrooms	Number of bedrooms.
KitchenCondition	The condition of the kitchen. 1 means the kitchen is in an excellent condition (typically a kitchen that has already been renovated). -1 means the kitchen is in a poor condition and needs renovation. 0 means the kitchen is in a standard / average condition.
Braais_Fireplaces	The number of built-in braais and/or fireplaces that the property has.
Garages	The number of garages that the property has.
Driveway	Indicates whether the property's driveway is paved or not. (Missing values imply that the driveway is not paved; this could mean that there is no driveway, or it could mean that the driveway is gravel or other material).
Pool	Whether the property has a pool or not.
BusyStreet	Whether the property is located on or near a busy street / highway (coded as 1) or not (coded as 0).
NearRail	Whether the property is located near a railway line (coded as 1) or not (coded as 0).
NearGreen	Whether the property is located near a green area such as a park or urban forest (coded as 1) or not (coded as 0).

Preliminary steps

These steps are important and should be followed carefully. Pay special attention to what the different files and objects should be named. If you don't follow the prescribed naming conventions, it could result in your assessment not being marked.

If you struggle with any of these steps, you are allowed to seek assistance from the lecturer (during practical lecture sessions or by making an appointment) or from a teaching assistant (during tutorial slots).

a) The file `A1_data.csv` contains the data required for this assessment. The file is available on SUNLearn under the Assessment 1 topic. Download this file and save it to the computer you will be working on. Make a note of the location where you saved the file, as you will need to access it later.

b) Open RStudio and create a new R Markdown file. This file should be a Document, and the title of the file should be A1 submission. The Author name should be your student number. In the date field, type only the year (2023). Make sure that you choose HTML as your output format.

c) In the markdown file you created in Step (b), delete everything from `## R Markdown` onwards. In other words, leave your header in place as well as the first code chunk named `setup`; everything else can be deleted.

d) In your markdown file, create a new code chunk. In this code chunk, write code to load the `tidyverse` group of packages. as well as the `tree` package. Include code that suppresses both the code and the output from this specific code chunk in your final document.

e) Before you continue, Knit your markdown document. Save it in the same location where you saved the `A1_data.csv` file. When prompted to provide a file name, the format of the name you provide should be `Surname_studentnumber`. For instance, if your surname is Smith and your student number is 12345678, you should name your file `Smith_12345678`.

f) Create another new code chunk below the one you created in Step (d). Within this code chunk, write code to import the dataset you downloaded in Step (a) into RStudio and store it in an object called **`main_data`**. Note that the dataset contains the variable (column) names in the first row, so you should include the argument **`header = TRUE`** in your **`read.csv()`** function.

g) Check the dimensions of the dataset you imported in Step (f) – you should have 2 930 observations and 31 variables.

h) Check the variable types (remember that you can use the **`str()`** function to do this). After importing, the variable types should correspond to the types indicated in the table on the next page. If any of the variable types are incorrect, you will have to write code to fix this.

Variable name	Variable type
Identifier	int
Price	int
Price_cat	Factor
Month_sold	int
Year_sold	int
Sale_type	Factor
Type	Factor
Area	Factor
Size	int
Street_front	int
Plot_shape	Factor
Incline	Factor
SlopeType	Factor
Condition	int
Year_Built	int
Year_Renov	int
Water	Factor / int*
Tarred_road	Factor
Culdesac	Factor / int*
Corner	Factor / int*
Roof_Type	Factor
Bathrooms	num
Bedrooms	int
KitchenCondition	int
Braais_Fireplaces	int
Garages	int
Driveway	Factor
Pool	Factor
BusyStreet	Factor / int*
NearRail	Factor / int*
NearGreen	Factor / int*

* *these can be factors or integers*

i) You will need to take a sample (without replacement) of the observations in the dataset before you start working on the assessment questions. To ensure reproducibility of your work, you need to set a seed before writing the code to draw a sample. ***The seed should be set as the last digit of your student number.*** For example, if your student number is 12345678, you should use the command `set.seed(8)`.

j) Draw a sample of size 2 500 from the original dataset. Store this sample in an object called `test_data`. Make sure that you set `replace = FALSE` for a sample without replacement. (You can refer to the instructions from Practical 4 if you are unsure of how to draw the sample.)

k) To check whether your sample is correct, please sort the records in the `test_data` object ascending according to `Identifier` and enter the 1st, 10th, 20th and 100th `Identifier` values in your sorted `test_data` object on the provided link on SUNLearn. If your sample appears to be incorrect (in other words, if the numbers you entered are marked as incorrect on SUNLearn), please check where you went wrong and correct it, or reach out to the lecturer or a teaching assistant for help. Once you are satisfied that these steps have been followed exactly and that you are working with the correct dataset, you can proceed with answering the assessment questions.

Marks for preliminary steps: 7

The allocated marks for the preliminary steps will take the following into account:

- Correct header
- Packages correctly loaded and code chunk suppressed as specified
- Data correctly imported
- Correct variable types
- Sample correctly drawn
- Markdown file runs correctly (i.e. reproducible)

Questions

This constitutes the formal assessment portion. Please remember that you are not allowed to seek help from anyone with any of these questions, including your classmates, teaching assistants or lecturer. Your assessment submission should be entirely your own work.

For all of the questions below, unless otherwise stated, use the sample you created in preliminary step (j); in other words, work with the `test_data` object.

1. Calculate the mean (average) price of the properties in the dataset.

(2)

2. Create a suitable plot to show the distribution of the `Price` variable. Based on this plot, comment on the distribution of the price of the properties in the dataset.

(5)

3. Create a bar chart showing the number of properties in the dataset for each `Type` of property (i.e. Single Storey, Double Storey, Estate and Duplex).

(3)

4. Create side-by-side boxplots to show how the distribution of `Price` differs according to the property `Type`.

(5)

5. Create a table showing the number of properties in each `Area` category as well as the average price of properties in each category. In other words, you should have the following information in a table (your table does not have to be formatted precisely like this though):

Area	**Number of properties**	**Average price of properties**
Residential		
Commercial		
Industrial		
Agricultural		

Based on the information in this table, what can you say about the difference in selling price between the different `Area` categories?

(6)

6. This question will focus on the `Street_front` variable.

   6.1 There are missing values in the `Street_front` variable. Write suitable R code to determine how many observations in your dataset have missing values for this variable.

(2)

   6.2 Create a new dataset called `cordata`, which excludes the observations with missing values for the `Street_front` variable.

(2)

6.3 Using the **cordata** object you created in Question 6.2, calculate the correlation between `Street_front` and `Size`. Display your answer showing only 3 decimal places.

(3)

6.4 Create a scatterplot showing `Street_front` on the x-axis and `Size` on the y-axis. You should use the **cordata** object you created in Question 6.2. Based on this plot, do you think there are any outliers present in these two variables? If so, describe how you would handle these observations in the dataset.

(7)

For Question 7 you should work with the full `test_data` object again, and <u>not</u> the `cordata` object you used in Question 6.

7. Examine the `Water` variable (you can do this using plots and / or summary statistics). What does this suggest to you about the use of this variable to help determine property prices?

(3)

8. For this question, you are going to create another new data object.

8.1 Create a new data object, called **newdata**. This dataset should contain all of the original data in the **test_data** object, but also two new variables called `transport_proximity` and `not_renovated`. The value of these two new variables should be determined as follows:

<u>transport proximity</u>
Coded as 1 if the property is on or near a busy street and / or near a railway line; 0 in all other cases.
(In other words, if `BusyStreet = 1` OR `NearRail = 1`, then `transport_proximity = 1`).

<u>not_renovated</u>
This variable should have a logical value of TRUE if the house has never been renovated and FALSE otherwise. (<u>Hint:</u> to determine whether a house has been renovated or not, you should compare the `Year_Built` and `Year_Renov` variables…)

(5)

8.2 How many properties are on or near a busy street or near a railway line?

(2)

8.3 How many properties have been renovated?

(2)

8.4 Create a suitable plot to show whether there is a relationship between the price of a property and whether the property has been renovated or not. Based on your plot, do you think that renovating a property has an impact on price?

(5)

For Question 9 onwards you should work with the full `test_data` object again, and not the `newdata` object you used in Question 8.

9. South Africans love a braai and the outdoor lifestyle! But does the presence of braais/fireplaces and pools have an impact on the price of a property? Create suitable data visualisations to motivate your answer.

(9)

10. Examine the relationship between the date a property was sold and its price, by creating a suitable data visualisation. What do you observe? In your answer, specifically consider the overall trend of prices over the period but also whether there appears to be any seasonality present.
Hint: You might want to consider creating a new variable, which combines the `Year_sold` and `Month_sold` variables. For working with dates in the `tidyverse`, you could look at the `lubridate` cheatsheet, which is available here: https://rawgit.com/rstudio/cheatsheets/main/lubridate.pdf

(9)

11. Which factors do you think might be the most significant in predicting house prices?

    11.1 Pick a maximum of 5 variables that you feel might have a significant impact on property prices (you can pick less than 5, but not more than 5). State what these variables are and explain why you think they might be significant.

    (4)

    11.2 Conduct suitable data analyses to show whether your expectation(s) in Question 11.1 are confirmed or not. Your analysis can include summary statistics and tables and / or data visualisations. Clearly state whether your analysis confirms your supposition in Question 11.1.

    (8)

12. For this question, you are going to construct a decision tree for classification. You should use the **`tree()`** function from the **`tree`** package in R to construct the tree.

    12.1 Construct a decision tree to classify the price category (`Price_cat`) of a property. Your decision tree may only consider the following variables:
    `Area`, `Type`, `Size`, `Corner`, `Bathrooms`, `Bedrooms`, `KitchenCondition`, `Braais_Fireplaces`, `Garages`, `Driveway` and `Pool`.
    This means that, in the formula specification for the tree, you should specify that only these variables may be considered. (Since `Price_cat` is derived from the `Price` variable, including it in the tree will give you a tree with perfect prediction – a prime example of data leakage!)
    Hint: To model a variable Z using only variables X and Y, you would use `formula = Z~X + Y`

    (4)

    12.2 Plot the tree that was constructed in Question 12.1.

    (3)

12.3 For the tree constructed in Question 12.1, what is the misclassification error on the training data?

(2)

12.4 Use the tree that you constructed in Question 12.1 to predict the price category for the following property. You can write code to determine your prediction, or you can use the tree plot from Question 12.2 to "visually" make a prediction.

Variable	**Value**
`Area`	Residential
`Type`	Single storey
`Size`	1300
`Corner`	0
`Bathrooms`	4
`Bedrooms`	5
`KitchenCondition`	1
`Braais_Fireplaces`	3
`Garages`	2
`Driveway`	Paved
`Pool`	Yes

(2)

Marks for questions: 93

TOTAL MARKS: 100

Before you submit:

- Make sure that your rendered HTML file contains headings to indicate the different question numbers.
- Check that your rendered HTML file does not include unnecessary output. For instance, including a print-out of the entire dataset will be considered unnecessary output! **Marks will be deducted for this.**

Remember that you should submit the following documents:

1. RMarkdown file, in the format Surname_studentnumber.Rmd. (Uploaded to SUNLearn.)
2. Rendered HTML version of the RMarkdown file, in the format Surname_studentnumber.html. (Uploaded to SUNLearn.)
3. Signed plagiarism declaration. (Scanned copy uploaded to SUNLearn or physical copy handed in to the lecturer by the due date.)

APPENDIX 1

Plagiarism declaration

1. Plagiarism is the use of ideas, material and other intellectual property of another's work and to present it as my own.

2. I understand what plagiarism is and I am aware of Stellenbosch University's policy in this regard.

3. I agree that plagiarism is a punishable offence because it constitutes theft.

4. I also understand that direct translations are plagiarism.

5. All quotations and contributions from any sources whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism.

6. I declare that the work contained in this assessment is my own original work and I have not obtained help from anyone. I acknowledge that copying someone else's assessment and / or code, or part thereof, is wrong, and that submitting identical work to others constitutes a form of plagiarism.

7. I declare that I have not permitted anyone else to copy my answers and / or code or provided help to another student in completing this assessment.

8. I have not copied this assessment or created images of the screen showing any part of the assessment (including, but not limited to, images obtained by means of screenshots, photographs and any software applications). I have also not shared the data or any part of the assessment questions with anyone else, including uploading it online.

First name: …….

Surname: ……

Student number: ……

Date: ……

Signature: ……