DATA9001

Fundamentals of Data Science

Assignment 2

Assignment Details:

- This assignment is due **Friday 18th July 2025, 5pm** and must be uploaded to Moodle.
- Make sure you submit your work in one PDF format using the following convention using your zID, First name and Surname.

A2-zXXXXXXX-FirstName-Surname.pdf

- Assignments without signed plagiarism declaration (below) will not be accepted and late assignments will not be accepted unless accompanied by medical certificates.
- This assignment weights for 15% of the final mark.
- There is a total of 3 components and 15 marks.

Plagiarism Statement:

I declare that this assessment item is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere. I acknowledge that the assessor of this item may, for the purpose of assessing this item reproduce this assessment item and provide a copy to another member of the University; and/or communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct.

Name _____
Student No _____
Signature _____
Date _____

Problem Statement:

You are a data scientist working for a large real estate development company. The company is looking to invest in and develop new apartments in the state of Victoria, Australia. The market there has been quite difficult the last few years and so they have tasked you to help them identify where will be a good location to invest. You have been given 3 datasets which are typically used by the company to analyse where to invest.

Report Style:

The report should be approximately 500 words. And be divided into 3 clear sections. 1) Data summary 2) Model Estimation 3) Model interpretation. With multiple appropriate graphs.

Datasets:

A) Apartment_prices.csv
    a. Median apartment prices by Victorian suburb for past year
B) Historical_demographic.csv
    a. Historical population growth rates for the past year
    b. Historical median income for the past year
    c. Historical unemployment rate for the past year
    d. Historical priority growth area (true/false)
C) Projected_demographic.csv
    a. Projected population growth rates for the next year
    b. Projected median income for the next year
    c. Projected unemployment rate for the next year
    d. Projected priority growth area (true/false)

Using this data write a report that provides a recommendation on the best suburb to invest in for next year. Use as a stimulus for style and approach, the news report we presented in week 4 where data visualisation tools were used to provide a clear story justifying a particular conclusion. For your modelling approach use the material that we covered in week 5 which discussed how to estimate and interpret a linear regression. **Make sure to attach your R code to the end of your report.**

Notes:

- Note that the datasets provided have issues that need to be addressed some data cleaning and processing is necessary. As per examples discussed in Tutorial week 4 pertaining to outliers, missing data and incorrectly input data.
- Data visualisation is important to communicate your extract, transform and load decisions with respect to data issues
- When building your model with linear regression make sure to justify why you include/exclude certain variable. This justification can be through words or data visualisation.
- When estimating your model make sure to write out the exact functional relationship you are estimating so it is easier to interpret e.g. $B_0 + B_1 Wage + B_2 Wage^2$
- When interpreting your linear regression make sure to reference both economic and statistical significance.
- Note that the data for future projection should be viewed as testing feature data, you are expected to make your prediction using that forecast data

Marking Guide:

Criteria	Value	Assessment		
		Poor (0-2)	Average (2-3)	Excellent (4-5)
Data Handling & Visualisation	5 marks	The student failed to correctly extract, transform and load the dataset. Examples includes failing to handle incorrect data, missing data and outliers. Failing to correctly merge and use datasets. There was little to no use of data visualisation tools to interpret the data or explain decisions made. Data visualisation presented was of poor quality not reflecting style guidelines as discussed in lectures.	The student successfully extracted, transformed and loaded the dataset. They systematically handled data issues such as incorrect data, missing data and outliers but with little to no explanation. There was adequate use of data visualisation to explain the underlying data types highlighting key data features through the use of bar charts, histograms, box plots and scatterplots. Presented graphs was of adequate quality reflecting the style guidelines discussed in lectures.	The student successfully extracted, transformed and loaded the dataset. They systematically handled data issues such as incorrect data, missing data and outliers providing a simple but effective explanation justifying strategies used. There was excellent use of data visualisation to explain the underlying data types highlighting key data features such as through the use of bar charts, histograms, box plots and scatterplots. Presented graphs was of exceptional quality reflecting best practice of style guidelines as discussed in lectures.
Model Estimation	5 marks	The student failed to provide any justification or explanation of their model estimation strategy. Either failing to or incorrectly estimating an appropriate linear regression from the dataset.	The student successfully estimated a linear regression with a clearly stated functional relationship between historical feature and label data. They transformed data as appropriate in order to identify alternative potential relationships without detailed explanation. The student provided some statistical and practical justification for their estimation strategy.	The student estimated a linear regression with a clearly stated functional relationship between historical feature and label data. The functional relationship included correctly transformed feature/label data with an excellent justification and/or visualisation to provide a compelling case for modelling decisions.
Model Interpretation	5 marks	The student provided little to no explanation of the implications of their model. The student did not or poorly answered the initial problem statement of providing a recommendation of a suburb to invest in. If a recommendation was provided it lacked any logical flow from the previous analysis done and reported. The recommendation failed to use any data visualisation tools. No mention was made of any limitations of the model with respect to Gauss Markov assumptions.	The student provided some explanation of the implications of the model. The student adequately answered the initial problem statement by providing a clear recommendation of a suburb to invest in. The recommendation naturally followed the earlier analysis done and was simply justified from the model estimated using forecast data. The recommendation used some data visualisation of adequate quality. Some mention was made of the limitations with respect to Gauss Markov assumptions.	The student provided an excellent explanation of the implications of the model. The student compellingly answered the initial problem statement by providing a clear recommendation of a suburb to invest in. The recommendation was easy to interpret from the analysis done and was difficult to contest. The recommendation used excellent logic and data visualisation to build the case for investment. The limitations of the model were also comprehensively addressed with respect to Gauss Markov assumptions.