Big Data Analytics

Assignment 2

# Description

In the previous assignment you have performed an exploratory analysis of the housing dataset. You will work with similar housing dataset for as- signment 2 and it contains the prices and other attributes of more than 48k houses in the city of Melbourne. This dataset does not contain any missing values and it is made available on Moodle for download. Your assignment requires you to perform hypothesis testing, regression and classification tasks on the given dataset.

# Tasks

Complete the following tasks.

## Task A

Before you start statistical analysis, you have to define hypotheses, which will be tested. You should state at least 2 different hypotheses, each to test different data (so not all hypotheses should be checking the same statement just on different variables). Remember that there are different types of tests and you should use as many as you can (given if they are valid and make sense). Your ultimate goal is to report some findings. You should also prove that these findings are statistically correct. Take the below points as hints but do not limit yourself to these:

* + - Look at different plots you have created during exploratory analysis. What conclusions can be drawn based on these? These could become your hypotheses.
    - If you focus on one attribute, what is your intuition about the distri- bution that could explain such results? You can check and measure how well the data fits some distribution.

Remember that data analysis is not only about finding and proving hypotheses but also about summarising data and communicating it. It is not a failure if you do not get ”significant” results, you still have to report that. If your analysis makes sense (e.g. it is valid from the sta- tistical point of view), there is no such thing as a bad result. Present your analysis in the form of a report. Each hypothesis should be de- scribed, you should state what you want to prove. If you are claiming that groups have different characteristics, first show these on plots and comment on them. Report should be written in a way that a person without prior knowledge of the data is able to follow it.

## Task B

* + - Divide the dataset into training and test data. Use 75/25 split.
    - Perform Linear Regression with Multiple Variables to predict the house price.
    - Report adjusted R squared (on training data). Use RMSE and correla- tion to report the prediction accuracy of the model on the test data.
    - Normalize the data and repeat the process of performing Linear Re- gression with Multiple Variables on normalized data to predict the house price.
    - Highlight the difference in prediction accuracy of both models.
    - Write your findings in this section. Each valid iteration Linear regres- sion, will get you 15 marks.

## Task C

* + - Divide the dataset into training and test data. Use 80/20 split.
    - Use kNN to classify houses into appropriate types based on their fea- tures.
    - Use C5.0 to classify houses into appropriate types based on their fea- tures.
    - Use ANN (hidden=5) to classify houses into appropriate types based on their features.
    - Evaluate and compare the (best) performance of each classifier.
    - Write your findings in this section.