Queen Mary
University of London

MTH781P: Data Analytics **2022/23**
Assessed Coursework **A. Baule**

The subject of this coursework is the investigation of a historical dataset containing data on property sales[1]. A summary of the variables of interest is given in the following table:

MSZoning	The general zoning classification
LotArea	Lot size in square feet
BldgType	Type of dwelling
HouseStyle	Style of dwelling
OverallQual	Overall material and finish quality
OverallCond	Overall condition rating
YearBuilt	Original construction date
CentralAir	Central air conditioning
GrLivArea	Above grade (ground) living area square feet
FullBath	Full bathrooms above grade
HalfBath	Half baths above grade
BedroomAbvGr	Number of bedrooms above basement level
KitchenAbvGr	Number of kitchens
KitchenQual	Kitchen quality
Fireplace	Fireplace
GarageArea	Size of garage in square feet
SaleCondition	Condition of sale
SalePrice	Sale price of property in US$

The dataset is available on QMplus as *property-sales.csv*. Further details on the variables can be found in the file *description.txt*. Use R to analyze the dataset and address the following tasks:

1. Explore the data. Plot and produce summary statistics to identify the key characteristics of the data and produce a report of your findings. We would expect between 5 and 10 tables or figures accompanied by a description of your main findings. Interpret your statistical observations in the business context of the dataset. (40 marks)

2. Develop a regression model to predict **SalePrice** from one or more of the other variables. Discuss your methodology including, for example, variable selection, goodness of fit, performance. Consider both linear and nonlinear models. Produce a report of your findings supported by plots and statistical analysis. (35 marks)

3. Develop a classification model to predict whether a property has a fireplace or not. The relevant variable is: **Fireplace**=Y,N. Explore various subsets of predictors and discuss the performance of your model. (15 marks)

Additional marks will be given for the overall presentation of the coursework, the quality of figures and writing. (10 marks)

[1]The original dataset is available on Kaggle and has been slightly modified for this assessment.