1. The dataset `train.csv` contains 79 explanatory variables. The data description and csv file can be downloaded directly from kaggle. Your task, as suggested on the kaggle website, is to build a model to predict final home prices. Note, this is part of a kaggle competition which you might consider participating in later on. Before you start the parts below, identify any 10 variables of your choice and write a brief paragraph of why you selected them. These are the predictors you will use for solving the problem.

   (a) Provide a descriptive analysis of your variables. This should include histograms and fitted distributions, quantile plots, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five-number summary). All figures must include comments.

   (b) For each variable (except indicator ones), test if a transformation to linearity is appropriate, and if so, apply the respective transformation, and comment on the transformed predictor(s).

   (c) Estimate a multiple linear regression model that includes all the main effects only (i.e., no interactions nor higher order terms). We will use this model as a baseline. Comment on the statistical and economic significance of your estimates. Also, make sure to provide an interpretation of your estimates. *Note:* You can use any combination of transformed and untransformed variables from the model in part (b).

   (d) In your model from part (c), identify if there are any outliers worth removing. If so, remove them but justify your reason for doing so and re-estimate your model from part (c)

   (e) Use Mallows Cp for identifying which terms you will keep from the model in part (d) and also test for multicollinearity. Based on your findings estimate a new model.

   (f) For your model in part (e) plot the respective residuals vs. y, and comment on your results.

   (g) Using AIC and BIC for model comparison, identify which model is better, (c) or (e). Why?

   (h) Estimate a model based on (g) that includes interaction terms and if needed, any higher power terms. Comment on the performance of this model compared to your other two models.

   (i) Lastly, choose you favorite model from all the ones estimated and perform a five-fold cross validation test on it. Then use the `test.csv` dataset to evaluate how well your model predicts home prices for out of sample data, and comment on your overall findings.

2. Assume a healthcare insurance company hired you as a consultant to develop an econometric model to estimate the number of doctor visits a patient has over a 3 month period. The rational behind this study is that patients with a higher number of doctors visits wold pose a higher liability in terms of insurance expenses, and therefore, this may be mitigated via a higher insurance premium. The panel data are from the *German Health Care Usage Dataset*, and consist of 7,293 individuals across varying numbers of periods with a total of 27,326 observations.

   (a) Build a multiple regression model with a subset of 10 predictors (at most), including interaction and non-linear transformations if appropriate. For this part you only need to briefly discuss a justification for the model chosen, and discuss the respective regression output.

   (b) *Differences in Differences:* In 1987 the German Government passed a series of legislations to improve healthcare access for unemployed people and women.

      i. Determine whether or not the policy worked for women.

      ii. Determine whether or not the policy worked for unemployed.

   (c) Test the hypothesis that the number of doctor visits a patient has over a 3 month period is greater for women than for men.

   (d) Based on your findings propose and test your own hypothesis of interest using the linear functional form: $\lambda = c_1\beta_1 + c_2\beta_2 + \cdots$.

Data Description (For Problem 2)

This is a large data set. There are altogether 27,326 observations. The number of observations ranges from 1 to 7. (Frequencies are: 1=1525, 2=2158, 3=825, 4=926, 5=1051, 6=1000, 7=987). Note, the variable NUMOBS below tells how many observations there are for each person. This variable is repeated in each row of the data for the person. Below are the variables definitions. Note: You can ignore the variables `TI` and `INCOME` (this one is just a copy of HHINC).

`ID` = person - identification number
`FEMALE` = female = 1; male = 0
`YEAR` = calendar year of the observation
`AGE` = age in years
`HSAT` = health satisfaction, coded 0 (low) - 10 (high) Note, this variable has 40 coding errors. Variable `NEWHSAT` below fixes them.
`HANDDUM` = handicapped = 1; otherwise = 0
`HEALTHY` = self reported to be healthy = 1; otherwise = 0
`ALC` = average alcohol consumption in the last 3 months
`TRAVEL` = traveled in the last 3 months abroad = 1; otherwise = 0
`HANDPER` = degree of handicap in percent (0 - 100)

HHNINC = household nominal monthly net income in German marks / 10000
LOGINC = Natural log (ln) of household nominal monthly net income in German marks / 10000
HHKIDS = children under age 16 in the household = 1; otherwise = 0
EDUC = years of schooling
MARRIED = married = 1; otherwise = 0
HAUPTS = highest schooling degree is Hauptschul degree = 1; otherwise = 0
REALS = highest schooling degree is Realschul degree = 1; otherwise = 0
FACHHS = highest schooling degree is Polytechnical degree = 1; otherwise = 0
ABITUR = highest schooling degree is Abitur = 1; otherwise = 0
UNIV = highest schooling degree is university degree = 1; otherwise = 0
WORKING = employed = 1; otherwise = 0
BLUEC = blue collar employee = 1; otherwise = 0
WHITEC = white collar employee = 1; otherwise = 0
SELF = self employed = 1; otherwise = 0
BEAMT = civil servant = 1; otherwise = 0
DOCVIS = number of doctor visits in last three months
HOSPVIS = number of hospital visits in last calendar year
UNEMPLOY = unemployed = 1; otherwise = 0
DOCTOR = dummy variable = 1 if DOCVIS > 0, 0 otherwise.
HOSPITAL = dummy variable = 1 if HOSPVIS > 0, 0 otherwise.
PUBLIC = insured in public health insurance = 1; otherwise = 0
ADDON = insured by add-on insurance = 1; otherswise = 0
NUMOBS = number of observations for this person. Repeated in each row of data.
NEWHSAT = recoded value of HSAT with coding errors corrected.
PRESCRIP = number of prescriptions in last three months