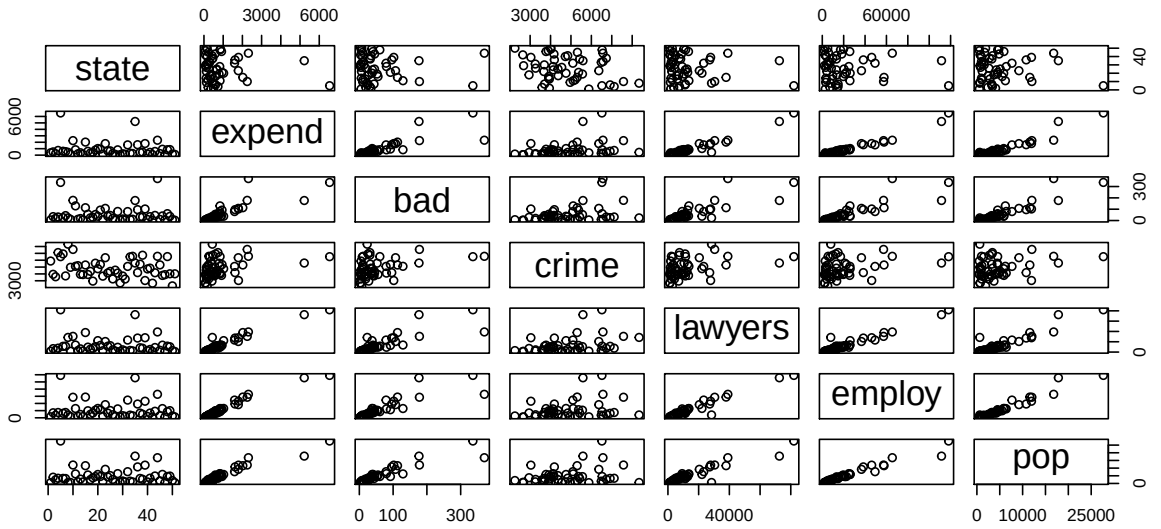
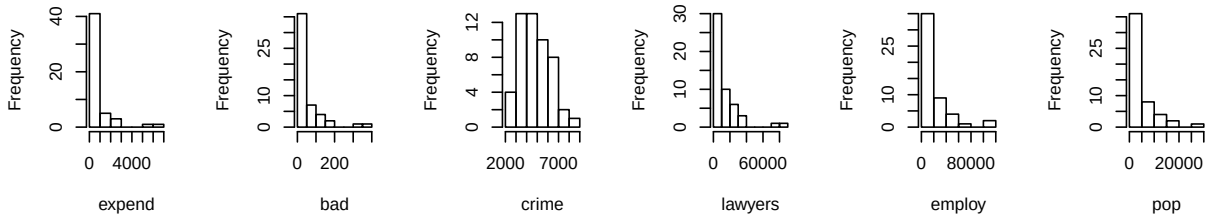
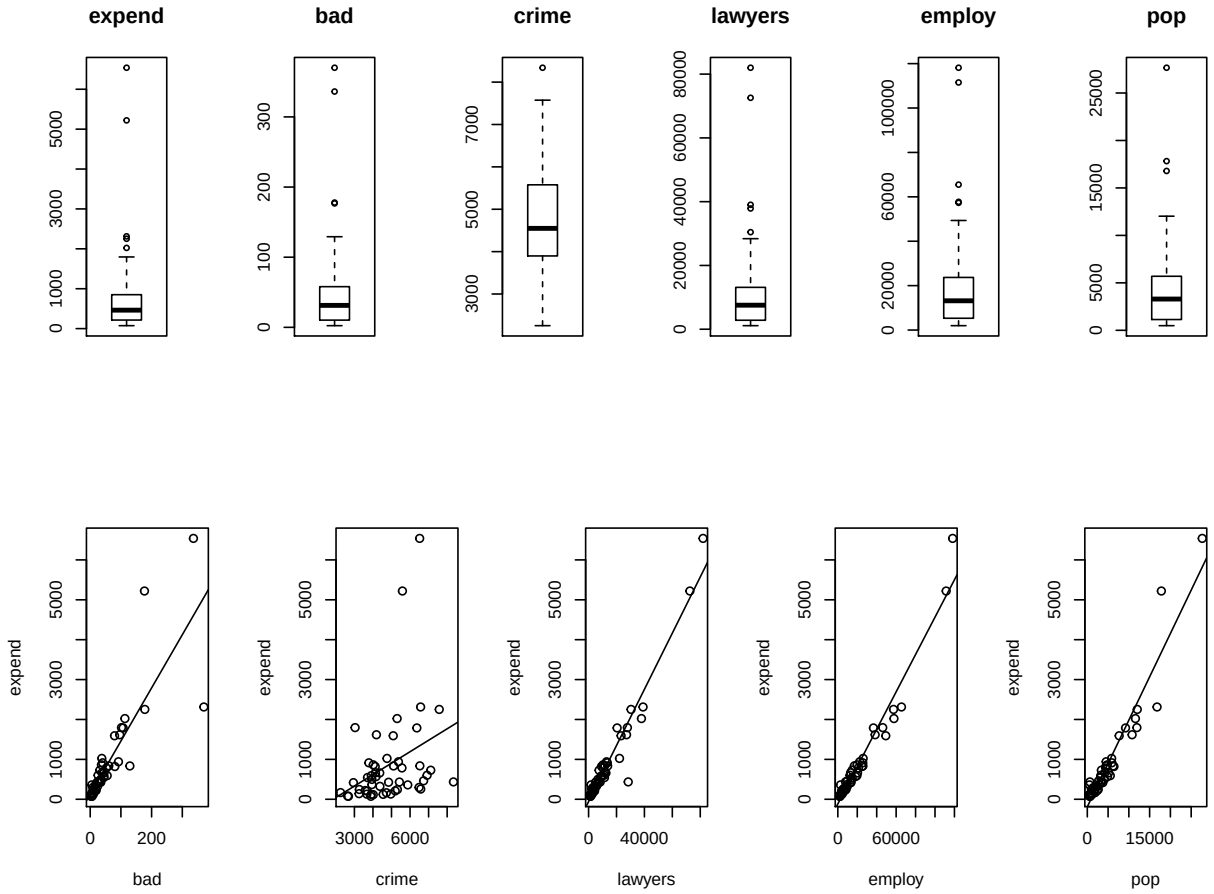
Exercise 5

In our regression analysis, the response variable is "expend" and the explanatory variables are: "bad, crime, lawyers, employ and pop". The purpose is to explain expend by a numerical function of the explanatory variables.

a) First, we make a graphical summary of the data by plotting each variable against the others in scatterplots. Looking at the plots, we observe that expend, lawyers, employ and pop all approximate a linear relationship with each other. Furthermore, state and crime have nonlinear relationhsips with all the other variables. Lastly, the variable bad can be argued to have a weak linear relationship with the variables expend, lawyers, employ and pop. Thereafter, we constructed histograms of the numerical data. Looking at the histograms, it is interesting to see that almost all variables (expend, bad, lawyers, employ and pop) follow a similar pattern, namely, the lowest value appears frequently and as the value increases, the frequency decreases steeply. Except for a few outliers of frequently occuring high values. In contrast, the variable crime shows a different pattern. Namely, the values in the middle occur also relatively frequently. But the rule: as the value increases, the frequency decreases, applies as well. Afterwards, we constructed boxplots of the data. Again, we observe a similar pattern of all variables except the variable crime. Crime is more evenly distributed and contains fewer outliers. The other variables are skewed towards the lower values combined with outlying higher values. To explore these outliers further and in order to build an intuition of the linear relationship between the response variable (expand) and the explanatory variables, we zoomed in on the relevant scatter plots that were presented above. From these plots, we observe a strong linear relationship between the response variable expend and the variables: bad, lawyers, employ and pop. When plotting the simple regression models in these scatterplots, we observe that the outliers follow the linear pattern.







A potential point is an outlier in an explanatory variable. The effect can be studied by fitting the model with and without the potential point. If the estimated parameters change drastically when removing the potenetial point, the observation is called an influence point. Using the Cook's formula, the distance of an observation on the predictions can be calculated. Whenever the Cook's distance for an observation approximates or is larger than 1, the observation can be considered to be an influence point. As we have not constructed a model yet, we analyse the potential and influence points of our chosen model in c). Another relevent concept is collinearity. This is the problem of linear relations between explanatory variables. Collinearity can be detected by a straight line in a scatter plot or by calculating the correlation coefficient. Looking at the scatter plots of the data, we suspect collinearity between the variables bad, lawyers, employ and pop. We confirm this by calculating the correlaction coeffiecients of all possible variable combinations. Looking at the output below, we observe that all the combinations of the variables lawyers, employ and pop have a correlation coefficient above 93. Thus, we conclude that these variables have a collinear relation. The variable bad has a weaker collinear relation with the variables lawyers, employ and pop, namely, ranging from 0.83 to 0.93. Lastly, the variable crime has no collinear relation with any of the other variables. When collinearity is detected among variables, we should avoid having both explanatory variables in the model.

```
##           bad crime lawyers employ  pop
## bad     1.00  0.37    0.83   0.87 0.92
## crime   0.37  1.00    0.38   0.31 0.28
## lawyers 0.83  0.38    1.00   0.97 0.93
## employ  0.87  0.31    0.97   1.00 0.97
## pop     0.92  0.28    0.93   0.97 1.00
```

b) To fit a linear regression model to the data, first, we start with the step-up method. Using this method, we start by fitting all possible simple linear regression models and calculate the determination coefficient

(R^2). The results are shown in the table below (Round 1). Looking at this table, we observe that employ has the largerst value of R^2 (0.954) and is thus selected. Therefore, we add the remaining variables to contruct a model with the variable employ. The results are shown in the table below (Round 2). We observe that the model that is constructed using the variables employ and lawyers has the highest value of R^2 (0.9632). This value is also higher than the R^2 value of the previous model (0.954) and is significant (the p-values are smaller than the significance level of 0.05). For this reason, we extend the model with the remaining variables. The results are shown in the table below (Round 3). The table constructed using the variables employ, lawyers and bad has the highest value of R^2 (0.9639). However, the result in insignificant (p-value is equal to 0.34496), therefore, the method stops. The resulting model (expend ~ lawyers + employ) is expend = -1.107ee^{+02} + 2.686e^{-02} * lawyers + 2.971e^{-02} * employ + error.

Round 1:

Explantory Variable(s)	bad	crime	lawyers	employ	pop
Multiple R-squared	0.6964	0.1119	0.9373	0.954	0.9073

Round 2:

Explantory Variable(s)	employ, bad	employ, crime	employ, lawyers	employ, pop
Multiple R-squared	0.9551	0.9551	0.9632	0.9543

Round 3:

Explantory Variable(s)	employ, lawyers, bad	employ, lawyers, crime	employ, lawyers, pop
Multiple R-squared	0.9639	0.9632	0.9637

Second, we use the step-down method. This method starts with fitting all explanatory variables in the so-called full model. In each iteration, one explanatory variable is removed. In Round 1, we observe that the variable crime has the highest p-value, $0.25534 > 0.05$, therefore, the variable crime will be removed. In Round 2, pop has the highest p-value, $0.06012 > 0.05$, therefore, the variable pop will be removed. In Round 3, bad has the highest p-value, $0.34496 > 0.05$, therefore, the variable bad will be removed. In Round 4, lawyers has the highest p-value, $0.00113 < 0.05$, therefore, the variable will not be removed and the method stops. This results in the model expend = -1.107e^{+02} + 2.686e^{-02} * lawyers + 2.971e^{-02} * employ + error. The step-up and step-down model resulted in the same model. Note that the explanatory variables in this model are linearly correlated, this is bad practise for a regression model. If the step-up and step-down resulted in different models, the one with the higest value of R^2, the lowest number of explanatory variables and no collinear variables would be prefered.

Round 1: expend ~ bad, crime, lawyers, employ, pop

Explantory Variables	bad	crime	lawyers	employ	pop
p-value	0.02719	0.25534	0.00592	0.00354	0.03184

Round 2: expend ~ bad, lawyers, employ, pop

Explantory Variables	bad	lawyers	employ	pop
p-value	0.05402	0.00106	0.00380	0.06012

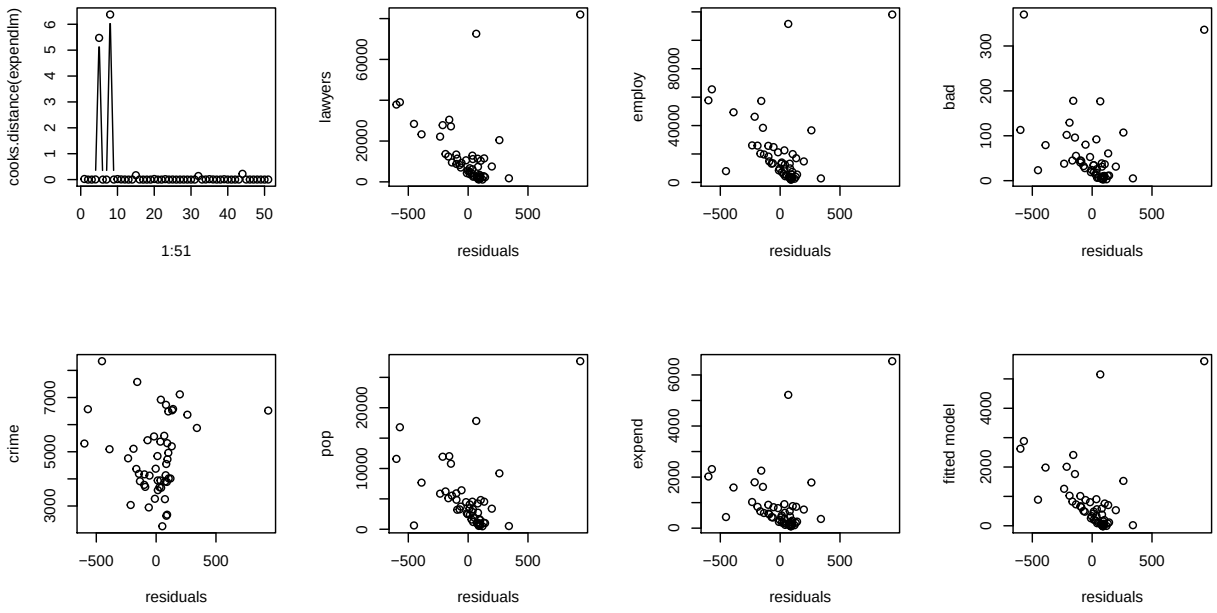
Round 3: expend ~ bad, lawyers, employ

Explantory Variables	bad	lawyers	employ
p-value	0.34496	0.00147	$1.2e^{-06}$

Round 4: expend ~ lawyers, employ

Explantory Variables	lawyers	employ
p-value	0.00113	$4.89e^{-07}$

c) Coming back to question a), we investigate the potential and influence points of the model (expend ~ lawyers + employ). When ordering the residuals of the model, we observe the presence of outliers. Thus, we test if these potential points are influence points using the Cook's distance. We calculate that the data on the 5th and 8th position have a Cook's distance larger than 1. This is also visible from the plot of the Cook's distances. Hence, we conclude that the data on position 5 and 8 are influence points. Now, we check the model assumptions (linearity of the relation and normality of the errors) using both graphical and numerical tools. First, we construct the scatter plot of the residuals against each explanatory variable that is in the model (lawyers and employ) seperately. We observe a cluster around zero and relatively little spread, except for two outliers. Second, we construct the scatter plots of the residuals against each explanatory variable that is not in the model (bad, crime and pop) seperatly. The outputs of bad and pop have a similar pattern as the previous plots. This can be explained by the fact that these variables are all collinear (see a). Except for the variable crime, this plot shows more spread. When looking at the patterns, we do not observe a linear relation and therefore should not include more variables into the model. Third, we construct the scatter plot of the reduals against the response variable (expend) and the fitted model. We observe little spread as in both plots there is a cluster around zero except the two reoccuring outliers.



Afterwards, we construct the added variable plots. In these plots, the residuals of the explanatory variables are plotted against the residuals of the model without that specific variable. This shows the effect of adding an explanatory variable to the model. Looking at the figure, we observe that the plots ... Lastly, we check

the normality assumption by constructing a normal QQ plot. We cannot assume normality as the QQ plot does not approximate a straight line, this means that the model is invalid. In conclusion, the model is flawed. First of all, the explanatory variables lawyers and employ are collinear, the scatter plots of the residuals show little spread and variance, and are clustered around zero. Furthermore, the residuals are not normally distributed. Therefore, in a next iteration the model should be adapted.