1. Read the data into a data frame. Check for typographical errors. Correct as necessary. (4)

We load the data using the command `read.csv`. The first 23 lines are explanations from the instigators of the analysis; we remove these using the option `skip`. The options `header` and `fill` are set to `TRUE` by default in `read.csv`. The pairs plot in Figure 1 shows that there seem to be some typographical errors in the variables `latitude` and `longitude`. We look at them closer using the range we observe in the plot and see that for these points the decimal point has moved one position to the left. We correct this by multiplying the points by 10. The provided file `housing200correct.csv` contains the true values without the typographical errors. We can use this to check whether we have found all typographical errors.

```r
# Read data
housing.df <- read.csv("housing200.csv",skip=23)

# Create pairs plot
pairs(housing.df)

# Identify and amend typographical errors
housing.df[housing.df$latitude<10,1:2] <- 10*housing.df[housing.df$latitude<10,1:2]

# Compare to original data
correct.housing.df <- read.csv("housing200correct.csv",skip=23)
all(housing.df==correct.housing.df)
```

In this case the command returned `FALSE` instead of `TRUE` due to a rounding error. We can verify the truthfulness of our claim by looking at `housing.df-correct.housing.df`. If every value is approximately 0, we are done.

**To attain marks for Question 1:**

- One (1) point for loading the data and using `skip=23`. I also accept when the students removed the first 23 lines using Excel or similar programs and loaded the modified file into R.

- One (1) point for identifying the typographic errors in a plot. I also accept a screenshot from Excel or similar program showing the culprits.

- Two (2) points for changing the typographic errors, either by multiplying by 10, or by replacing them with the true values. Also award points if student removed all three points. I also accept if the student adapted the points in Excel.
Figure 1: Pairs plot and absolute of correlation for housing data.
2. Fit a regression model to these data, using `medianHouseValue` as the response. Do all variables appear to have an effect on the house value? What is the nature of these effects, if any? Give reasons.

We fit the model and look at the summary. We find three non significant variables. Interpreting the signs of the coefficients we find that `housingMedianAge` and `medianIncome` appear to have a positive relationship with the response while `longitude`, `latitude` and `population` have a negative relationship with the response. We can assume that the relationships with `longitude` and `latitude` to the response are of geographical nature. That the housing value increases with `medianIncome` is not too surprising since people with higher income tend to live in more expensive housing.

Having said all that, interpreting the coefficients before checking the model is somewhat premature, because every violation of the model assumption can lead to wrongly estimated coefficients.

```r
# Linear Model
housing.lm <- lm(medianHouseValue~.,data=housing.df)
summary(housing.lm)
```

```
Call:
  lm(formula = medianHouseValue ~ ., data = housing.df)

Residuals:
   Min     1Q Median     3Q    Max
-348100 -118100    1600  136100  772000

Coefficients:             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.130e+06   7.789e+05   -5.30 3.14e-07  ***
longitude      -5.059e+04   9.026e+03    -5.61 7.20e-08  ***
latitude       -5.340e+04   8.700e+03    -6.14 4.73e-09  ***
housingMedianAge  1.868e+03   4.833e+02     3.87 0.000152  ***
totalRooms      1.237e+01   9.614e+00     1.29 0.199780
totalBedrooms   9.346e+01   1.361e+02     0.69 0.493234
population      -7.725e+01   1.840e+01    -4.19 4.11e-05  ***
households      8.209e+01   1.550e+02     0.53 0.596874
medianIncome    3.324e+04   3.433e+03     9.68 < 2e-16  ***

Residual standard error: 75300 on 191 degrees of freedom
Multiple R-squared: 0.6577,   Adjusted R-squared: 0.6433
F-statistic: 45.87 on 8 and 191 DF,  p-value: < 2.2e-16
```

**To attain marks for Question 2:**

- One (1) point for fitting the linear model of all eight variables against the response. This can be done by using the dot or writing out all variable names.
- One (1) point for presenting the output of the `summary` function.
- Three (3) points for interpreting the `summary` output. Give full points if three of the following statements are made:
- $R^2$ is mentioned in some way,
- Identification of non-significant variables
- Identification of type of relationship of significant variables
- Interpretation of significant relationships

- If a student mentions that interpretation is premature due to potential model violations, he gets the three (3) points irrespective of her or his interpretations of coefficients and model.

3. Do you think these data have collinearity problems? Give reasons. Are all variables necessary? Based on your observations suggest a submodel and compare it to the full model. Continue with the model that you consider appropriate.

We found for question 2 that there were three non-significant variables. The pairs plot in Fig. [1] indicates that there might be a collinearity problem in the data. We compute the Variance Inflation Factors (VIF) for the covariates to find evidence of collinearity. We see that inflated VIFs for households (83.027023) and totalBedrooms (71.389478). Since neither had a significant coefficient, we start by removing households, refit and use anova to test the submodel against the full model. anova returns a $p$-value of 0.5969, so we have no evidence against the null hypothesis that the submodel is as good as the full model. Let us have a look at the summary for the submodel. The summary of the submodel shows that the variable totalRooms is still not significant. We look at the VIFs and find that while the VIFs are still slightly raised they are no indication of having collinearity in totalBedrooms. So we remove this variable as well. An anova test of the new submodel against the original model is not significant with a $p$-value of 0.3684. This model is significant in all variables, and an added.variable.plot indicates that all are needed to explain the variation in the data.

```r
# Computing and outputting VIFs
housing.covariates <- housing.df[, -9]
VIF.housing.cov <- diag(solve(cor(housing.covariates)))
VIF.housing.cov

longitude    latitude housingMedianAge  totalRooms
12.396514     12.632627      1.187696    11.840692

# Checking submodel appropriateness
housing.sub1.df <- housing.df[-7]
housing.sub1.lm <- lm(medianHouseValue ~ ., data=housing.sub1.df)
anova(housing.sub1.lm, housing.lm)

Analysis of Variance Table

Model 1: medianHouseValue ~ longitude + latitude + housingMedianAge +
          totalRooms + totalBedrooms + population + medianIncome
Model 2: medianHouseValue ~ longitude + latitude + housingMedianAge +
          totalRooms + totalBedrooms + population + households + medianIncome
```
### Summary for submodel

Call: `lm(formula = medianHouseValue ~ ., data = housing.sub1.df)`

---

Coefficients:

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | -4.244e+06 | 7.470e+05 | -5.682 | 4.88e-08 *** |
| longitude | -5.194e+04 | 8.645e+03 | -6.008 | 9.25e-09 *** |
| latitude | -5.475e+04 | 8.305e+03 | -6.592 | 4.08e-10 *** |
| housingMedianAge | 1.885e+03 | 4.814e+02 | 3.916 | 0.000125 *** |
| totalRooms | 1.262e+01 | 9.584e+00 | 1.317 | 0.189523 |
| totalBedrooms | 1.587e+02 | 5.776e+01 | 2.748 | 0.006563 ** |
| population | -7.297e+01 | 1.649e+01 | -4.424 | 1.62e-05 *** |
| medianIncome | 3.336e+04 | 3.420e+03 | 9.754 | < 2e-16 *** |

---

Residual standard error: 75160 on 192 degrees of freedom
Multiple R-squared: 0.6572, Adjusted R-squared: 0.6447
F-statistic: 52.58 on 7 and 192 DF, p-value: < 2.2e-16

# Removing, refitting and retesting

```
housing.sub2.df <- housing.df[-c(4,7)]
housing.sub2.lm <- lm(medianHouseValue ~., data = housing.sub2.df)
anova(housing.sub2.lm, housing.lm)
summary(housing.sub2.lm)
```

To attain marks for Question 3:

- The criteria for marks for this questions are a bit vague due to the wide variety of approaches one can make.
- Three (3) marks for computing VIFs, interpreting them accordingly (collinearity), and relating them to observed non-significance of coefficients.
- Two (2) marks for subsequent removal of variables. Removing multiple variables at once might overcompensate for collinearity situations.
- One (1) mark for using `anova` as a decision tool for appropriateness of removal.
- Give the five (5) remaining points for appropriate use of the tools provided (e.g. `added.variable.plots`, `summary`) and good justification of removal.

4. Do you think that these data are suitable to be modeled by a linear regression model? If not, what action should be taken? Is the fit improved if this action is taken? (10)
Figure 2: The contribution of the remaining variables after removal of two variables due to collinearity and non-significance.
We start by looking at the diagnostic plots provided for the `lm`-function. Figure 3 indicates that there is still some signal in the residuals. We use the GAM model approach to test whether any of our variables needs transforming. Figure 4 shows the curves for all variables. We see that no variable seems really in need of transformation, the curvature in the right column seem to be driven by outliers rather than signal. However, since we still need to find the reason for the signal in the residual plot in Figure 3 we check to see a transformation of the response using the Box-Cox plot. I chose the function `boxcox` from the R-package `MASS` to compute the plot. Figure 5(a) indicates that a power of 1/3 seems appropriate. We refit the model using this power, and recheck the Box-Cox plot to find that the revised model is optimal at $p = 1$, i.e. we found the right power for the response, see Figure 5(b). Figure 6 shows, that apart from a few potential outliers (see next question), the residuals have no signal left and are reasonable uniform around 0. Therefore, we refrain from transforming the covariates because the residuals already indicate planarity.

Following a summary of the commands for the plots used in this section.

```r
# Diagnostic plot
par(mfrow=c(2,2))
plot(housing.sub2.lm)

# GAM plot
library(mgcv)
par(mfrow=c(3,2))
plot(gam(medianHouseValue~s(longitude)+s(latitude)+s(housingMedianAge)+s(totalBedrooms)+s(population)+s(medianIncome),data=housing.sub2.df))

# Box-Cox plot
library(MASS)
boxcox(housing.sub2.lm)

# New model
housing.sub3.lm <- lm(medianHouseValue^(1/3)~.,data=housing.sub2.df)
boxcox(housing.sub3.lm)

# Diagnostic plots 2
par(mfrow=c(2,2))
plot(housing.sub3.lm)
```

To attain marks for Question 4:

- Marks are mainly given for using diagnostic tools, identifying problems with planarity and justifying transformations.
- Two (2) marks for providing diagnostic plots for the model taken from Question 3 (whichever model has been forwarded), and pointing out that there is some signal left in the residuals (if there is).
- Two (2) marks for checking the GAM plots OR/AND Box-Cox plots and the subsequent decision. One point for plots, and one point for the reasoning.
Figure 3: Diagnostic plots for our reduced model from Question 3.
Figure 4: GAM plots for reduced model from Question 3.
Figure 5: Box-Cox plot to find optimal power for response before and after transformation.

- Four (4) marks for executing the subsequent decision. The student can decide (as I have) to not put powers on the variables and look at the response instead, or tries to fit polynomials or splines. If student tries polynomials or splines, distribute the points on their revised model (deciding on power by looking at plot or choosing degrees of freedom for splines according to number suggested in GAM), and their decision for which power to go with when looking at the summary of the new model. If student goes for Box-Cox, give points for choice of power (if they have chosen anything between log and 1/2, that is fine by me), refitting the model and rechecking Box-Cox.

- Two (2) points for stopping because linearity has been achieved.

5. Are there any data points that have high leverage or large studentised residuals? If so, are these points having an undue influence on any aspect of the fitted model? Give a full discussion with reasons.

We start by looking at the LR function (Figure 7). The lines indicate that there are quite a few points with very high hat values, and a few points with residuals, and two points which light up as high-leverage outliers. We compute the hat values in detail noting that the threshold value is at $3(k + 1)/n = 3 \cdot 7/200 = 21/200 = 0.105$

```
plot(housing.sub3.lm,which=5)
abline(h=c(-2,2),v=HMD.lower)
hatvalues(housing.sub3.lm)[hatvalues(housing.sub3.lm)>HMD.lower]
```

<table>
<thead>
<tr>
<th>35</th>
<th>84</th>
<th>113</th>
<th>115</th>
<th>166</th>
<th>168</th>
<th>194</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1407244</td>
<td>0.1938485</td>
<td>0.1506109</td>
<td>0.1205688</td>
<td>0.1473453</td>
<td>0.1636280</td>
<td>0.1155329</td>
</tr>
</tbody>
</table>
Figure 6: Diagnostic plots for the new model with transformed response.
In addition we compute the standardised and studentised residuals and record the three most extreme values to either side of 0.

```r
# Standardised residuals
housing.rstand <- stdres(housing.sub3.lm)
housing.rstand[order(housing.rstand)][c(1:3,198:200)]

166 168 70 87 57 40
-2.775766 -2.755019 -2.607945 2.712343 2.724675 2.894792

# Studentised residuals
housing.rstud <- studres(housing.sub3.lm)
housing.rstud[order(housing.rstud)][c(1:3,198:200)]

166 168 70 87 57 40
-2.825540 -2.803553 -2.648261 2.758391 2.771432 2.952082
```

To test whether these values are unusual we go back to the assumptions and remember that the standardised residuals should be standard normal distributed. To this end, we simulate 1,000 samples with 200 observations each, and store the maximum (in absolute) of the recorded values. If our outliers would be unusual we would expect to see a value that is unusual for the distribution of these maximum values. Figure 8 shows the histogram from one such simulation. We see that our maximum value of 2.9 is well within this distribution.

```r
nsim = 1000
result = numeric(nsim)
for(i in 1:nsim){
  result[i] = max(abs(rnorm(200)))
}
hist(result)
abline(v=max(housing.rstud),lwd=2,col="violet")
```

We see that 166 and 168 appear to be the high-leverage outliers appearing in both lists (for hat values and residuals). We look at them in detail, remove them from our data set and refit the model.

```r
> housing.sub2.df[c(166,168),]

  longitude latitude housingMedianAge totalBedrooms population
166  -122.21   37.46             40          207           577
168  -122.14   37.50             46           4            13

  medianIncome medianHouseValue
166      15.0001         500001
168      15.0001         500001

> housing.rem.df <- housing.sub2.df[-c(166,168),]
> housing.rem.lm <- lm(medianHouseValue^(1/3)~.,data=housing.rem.df)
```

12
> summary(housing.rem.lm)
Call:
  lm(formula = medianHouseValue^(1/3) ~ ., data = housing.rem.df)
---
Coefficients: 
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.165e+02  6.305e+01  -6.606  3.82e-10 ***
longitude      -5.513e+00  7.261e-01  -7.592  1.35e-12 ***
latitude       -5.785e+00  6.950e-01  -8.323  1.61e-14 ***
housingMedianAge  1.821e+01  4.215e-02   4.321  2.49e-05 ***
totalBedrooms   1.966e-02  3.519e-03   5.586  7.90e-08 ***
population     -6.430e-03  1.382e-03  -4.652  6.14e-06 ***
medianIncome    3.745e+00  2.577e-01  14.531  < 2e-16 ***
---
Residual standard error: 6.478 on 191 degrees of freedom
Multiple R-squared:  0.6941 ,  Adjusted R-squared:  0.6845
F-statistic: 72.23 on 6 and 191 DF,  p-value: < 2.2e-16

We see that the removal of these points improved the fit quite a bit. It looks like these points are the two points with the most inflated medianHouseValue and the highest medianIncome and thus might not fit the rest of the data very well. Figure 9 shows the influence functions after the removal of our main culprits.

We finish at this point having convinced ourselves that this model helps explain the relationship between the variables reasonably and in a linear model fashion, and satisfies most of the assumption on multiple linear regression models. Our model housing.rem.lm has a multiple $R^2$-squared of 0.6941.

To attain marks for Question 5:

- Marks are mainly given for using diagnostic tools, identifying problems with planarity and justifying transformations.
- Four (4) points for providing an LR plot.
- Three (3) points each for identifying the largest HMD’s and residuals.
- Five (5) points for discussing the effect the points have on the fit.
- Actual results will depend on the model fitted.
- Do not deduct points if student has removed extreme outliers (166 and 168) in Question 4.

Total: (45)
Figure 7: LR-plot for our model.

Figure 8: Histogram for simulated extreme values.
Figure 9: Influence functions for our final model.