ECON 282, Professor Hogendorn

Assignment 4

For this assignment, we will use the US Zip Code demographics dataset included in the DataComputing package. To load this data, type:

```
> library(DataComputing)
> data(ZipDemography)
> Places <- tbl_df(ZipDemography)
> Places <- select(Places,-Medianofselectedmonthlyownercosts)
```

(The last line is needed because that variable isn't really present, and it took me a while to figure it out. Ahh, data wrangling....)

To get summary statistics on this data:

`install.package("stargazer")` (Only needs to be done once)

```
library(stargazer)
stargazer(Places,type="text")
```

Note: If you don't edit pretty heavily, the printout of this assignment would be crazy long. You can delete lots of information from the regression summaries and still have useful results.

1. Divide the dataset into a 70% training sample and a 30% testing sample. Here the ZIP variable provides a unique key for each observation. Some nice code for dividing the sample would be

   ```
   > train_index <- sample(Places$ZIP,29918)
   > Places_train <- filter(Places,ZIP %in% train_index)
   > Places_test <- filter(Places,!(ZIP %in% train_index))
   ```

2. Let's try to explain the `Medianvaluedollars` (the median price of a home) by using the rest of the data. Is there a variable you think should not be included as an X variable for theoretical reasons? (There may be many, but 1 is enough.)

3. Now make a linear model regressing `Medianvaluedollars` on everything except whatever you omitted from 2. There's a great syntax for this: say you want to omit ZIP (which is just a key variable) and variable X, but otherwise use every variable. Type:
   `> Regression <- lm(Medianvaluedollars ~ . - ZIP - X, data=Places_train)`
   You can read this as "regress on everything minus Zip and X."

4. See anything with a really high p-value that you think might also be good to omit? If so, omit it.

5. Now visualize the residuals from your regression as a function of the *Totalpopulation*. See any nonlinearities? If so, add a squared and or cubed term for that variable.

6. Run your final regression on your training data, and then give it a try on the testing data to see how well it really works.