R Exercise 1: INTA 6450 - Data Analytics and Security

R Exercise 1

Prompt

1. Download `exercise_1.r`
2. Open the file in RStudio and run the commands in the exercise and look at the output.
3. Read the comments to try to understand how each command is working, and look at the output. Make some changes to the code to explore how the output differs.
4. Then, submit the output for running these commands for this assignment. You may copy/paste output from your terminal into a word or text document. You can convert the notebook to PDF submission:File->Compile Report->Report output format as HTML/PDF.
5. Write and submit 100-150 words describing the extension you made to the code. You may post the change description as a comment at the beginning of your code, or submit the write-up in a separate document. Make sure to clearly mark where in your code the changes were made either by commenting, highlighting, or noting the command lines where the changes occur. **NOTE:** Exported code does not retain line numbers. Please refrain from phrases such as "code change begins on line 62".

Submissions

You will submit the following:

1. Original code output (R file code and output)
2. Your code change and output from that code change (comment where your code changes start and end)
3. Your code change description (this should be between 100 to 150 words)

An example of what the change description should look like:

```
############################
# CHANGE DESCRIPTION
# 100-150 words on the changes I've made, including a reference to which
# initial command I've changed >
# END OF CHANGE DESCRIPTION
############################
```

Files must be in one of the following formats: pdf, doc, html

Please note some of these exercises have graphs/charts - be sure when you export the original output to also check that the file includes graphs/charts that may have been present. If it does not, you will need to re-export or manually include the graphs in your final pdf, doc or html file

Solution

Wages

Blackburn and Neumark (1992)

- wage
- hours
- IQ
- KWW
- educ
- exper
- tenure
- age
- married
- black
- south
- urban
- sibs
- brthord
- meduc
- feduc
- lwage

Predicted wage based on simple linear regression model using `lm` in the `stats` package. Using `wage` as the response variable and `educ` or education as the predictor variable. Second try is using `wage` as the response variable and both `educ` and `IQ` as predictors.

- `lm` regression model
- Use `ggplot2` to evaluate the relationship between `wage` and all predictors.
  - See HW5 ISYE6501 Question 8.2
- Model data is adequate as we typically look for 10:1 ratio on data points vs. factors
- cv for `lm` uses `DAAG` and `glm` uses `boot`

To observe the relationship between `wage` and each of the 17 predictors, a matrix scatter plots is developed with `ggplot` and `gridExtra`. A function is created to create a single scatter plot so that mapping can be utilized to create all 17 plots at once and be placed in a matrix for a single image.

Load and Inspect the Data

```
# Import libraries
# lm and glm are in the stats package loaded by default
library(ggplot2) # Plot functions
library(reshape)

# Wages ------------------------------------------------------------------------
# Load wages data from AWS .csv
wages <- read.csv('http://inta.gatech.s3.amazonaws.com/wage2.csv')
# Write the wages data to a .csv in local directory
# write.csv(wages, "exercises/M5_regression/data/wage2.csv", row.names = FALSE)
```

```r
# Summary of wages statistics from data frame
summary(wages)
```

```
##       wage             hours             IQ             KWW
##  Min.   : 115.0   Min.   :20.00   Min.   : 50.0   Min.   :12.00
##  1st Qu.: 669.0   1st Qu.:40.00   1st Qu.: 92.0   1st Qu.:31.00
##  Median : 905.0   Median :40.00   Median :102.0   Median :37.00
##  Mean   : 957.9   Mean   :43.93   Mean   :101.3   Mean   :35.74
##  3rd Qu.:1160.0   3rd Qu.:48.00   3rd Qu.:112.0   3rd Qu.:41.00
##  Max.   :3078.0   Max.   :80.00   Max.   :145.0   Max.   :56.00
##
##       educ           exper           tenure           age
##  Min.   : 9.00   Min.   : 1.00   Min.   : 0.000   Min.   :28.00
##  1st Qu.:12.00   1st Qu.: 8.00   1st Qu.: 3.000   1st Qu.:30.00
##  Median :12.00   Median :11.00   Median : 7.000   Median :33.00
##  Mean   :13.47   Mean   :11.56   Mean   : 7.234   Mean   :33.08
##  3rd Qu.:16.00   3rd Qu.:15.00   3rd Qu.:11.000   3rd Qu.:36.00
##  Max.   :18.00   Max.   :23.00   Max.   :22.000   Max.   :38.00
##
##     married          black            south            urban
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   :0.893   Mean   :0.1283   Mean   :0.3412   Mean   :0.7176
##  3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##       sibs           brthord          meduc           feduc
##  Min.   : 0.000   Min.   : 1.000   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 8.00   1st Qu.: 8.00
##  Median : 2.000   Median : 2.000   Median :12.00   Median :10.00
##  Mean   : 2.941   Mean   : 2.277   Mean   :10.68   Mean   :10.22
##  3rd Qu.: 4.000   3rd Qu.: 3.000   3rd Qu.:12.00   3rd Qu.:12.00
##  Max.   :14.000   Max.   :10.000   Max.   :18.00   Max.   :18.00
##                   NA's   :83       NA's   :78      NA's   :194
##      lwage
##  Min.   :4.745
##  1st Qu.:6.506
##  Median :6.808
##  Mean   :6.779
##  3rd Qu.:7.056
##  Max.   :8.032
##
```

Form Linear Regression Model with All Predictors

```r
# Predict observed crime rate
# wage~. separates wage (response variable) from predictors
model <- lm(wage~., data=wages)
summary(model)
```

Observing the Relationship between Wage and Predictors

To observe the relationship between `wage` and each of the 17 predictors, a matrix scatter plots is developed with `ggplot` and `gridExtra`. Remember, `ggplot2` was imported earlier, so two additional libraries need to be loaded. The library `gridExtra` is to arrange a grid, and `purrr` is used to map the plots. First, a list of the predictor variables is defined as `predictors`. The list of predictors is then utilized to generate scatter plots. A function is created to create a single scatter plot so that mapping can be utilized to create all 17 plots at once and be placed in a matrix for a single image.

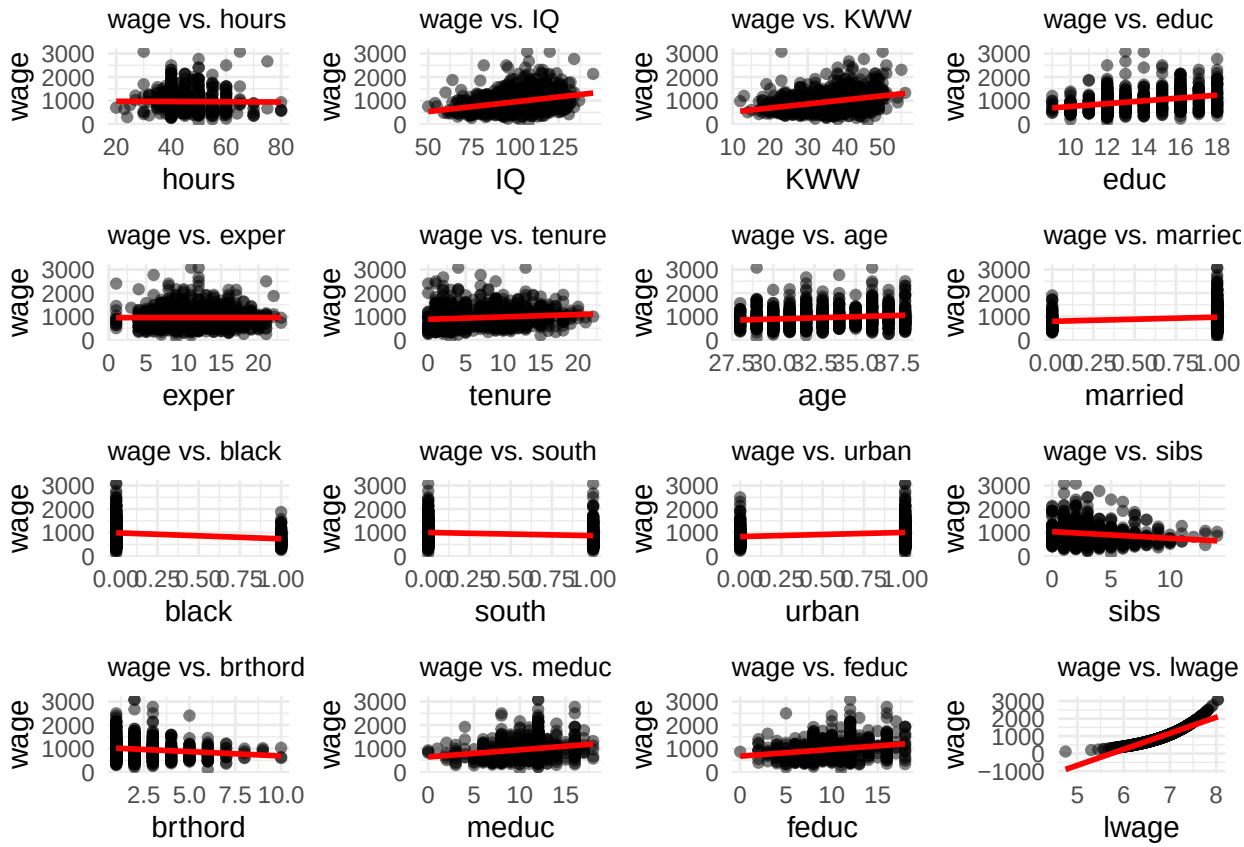
```
library(gridExtra)
library(purrr)

# Create a list of predictor variables
predictors <- c("hours", "IQ", "KWW", "educ", "exper", "tenure", "age",
                "married", "black", "south", "urban", "sibs", "brthord",
                "meduc", "feduc", "lwage")

# Function to create a scatter plot with regression line for each predictor
plot_predictor <- function(predictor) {
  ggplot(wages, aes_string(x = predictor, y = "wage")) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    theme_minimal() +
    labs(title = paste("wage vs.", predictor))
}

# Create plots for all predictors
plots <- map(predictors, plot_predictor)

# Arrange plots in a grid
scatter <- grid.arrange(grobs = plots, ncol = 4)
scatter
```


wage vs. hours
wage vs. IQ
wage vs. KWW
wage vs. educ
wage vs. exper
wage vs. tenure
wage vs. age
wage vs. married
wage vs. black
wage vs. south
wage vs. urban
wage vs. sibs
wage vs. brthord
wage vs. meduc
wage vs. feduc
wage vs. lwage
wage
hours
IQ
KWW
educ
exper
tenure
age
married
black
south
urban
sibs
brthord
meduc
feduc
lwage