Homework 5

Hannah Marr

II - R

You will need to download the housing.csv dataset. Read the dataset into R and clean it before proceeding. The response variable of interest will be: Y = median house value The predictor variables we are interested in are: X1 = housing median age X2 = population X3 = median income

```
options(repos = c(CRAN = "https://cloud.r-project.org/"))
```

```
# Install necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(dplyr)
```

```
# Read the housing data from a CSV file into a dataframe.
cal_housing_raw_data <- read.csv("/Users/hannahmarr/Desktop/Tufts/DATA200/Labs/housing.csv")

# Display the first few rows of the dataframe to inspect the data.
head(cal_housing_raw_data)
```

```
##   longitude latitude housing_median_age total_rooms total_bedrooms population
## 1   -122.23    37.88                 41         880            129        322
## 2   -122.22    37.86                 21        7099           1106       2401
## 3   -122.24    37.85                 52        1467            190        496
## 4   -122.25    37.85                 52        1274            235        558
## 5   -122.25    37.85                 52        1627            280        565
## 6   -122.25    37.85                 52         919            213        413
##   households median_income median_house_value ocean_proximity
## 1        126        8.3252             452600        NEAR BAY
## 2       1138        8.3014             358500        NEAR BAY
## 3        177        7.2574             352100        NEAR BAY
## 4        219        5.6431             341300        NEAR BAY
## 5        259        3.8462             342200        NEAR BAY
## 6        193        4.0368             269700        NEAR BAY
```

```r
# Get the dimensions of the dataframe (number of rows and columns).
dim(cal_housing_raw_data)
```

```
## [1] 20640    10
```

```r
# Check for missing values in each column.
# 'colSums(is.na())' will return the count of missing values for each column.
colSums(is.na(cal_housing_raw_data))
```

```
##          longitude           latitude housing_median_age        total_rooms
##                  0                  0                  0                  0
##     total_bedrooms         population         households      median_income
##                207                  0                  0                  0
## median_house_value    ocean_proximity
##                  0                  0
```

```r
# Drop rows with any NA values
cal_housing <- na.omit(cal_housing_raw_data)

# Get the dimensions of the dataframe (number of rows and columns).
dim(cal_housing)
```

```
## [1] 20433    10
```

```r
# Check for missing values in each column.
# 'colSums(is.na())' will return the count of missing values for each column.
colSums(is.na(cal_housing))
```

```
##          longitude           latitude housing_median_age        total_rooms
##                  0                  0                  0                  0
##     total_bedrooms         population         households      median_income
##                  0                  0                  0                  0
## median_house_value    ocean_proximity
##                  0                  0
```

```r
# check if there are any null values left in the dataset
any(is.na(cal_housing))
```

```
## [1] FALSE
```

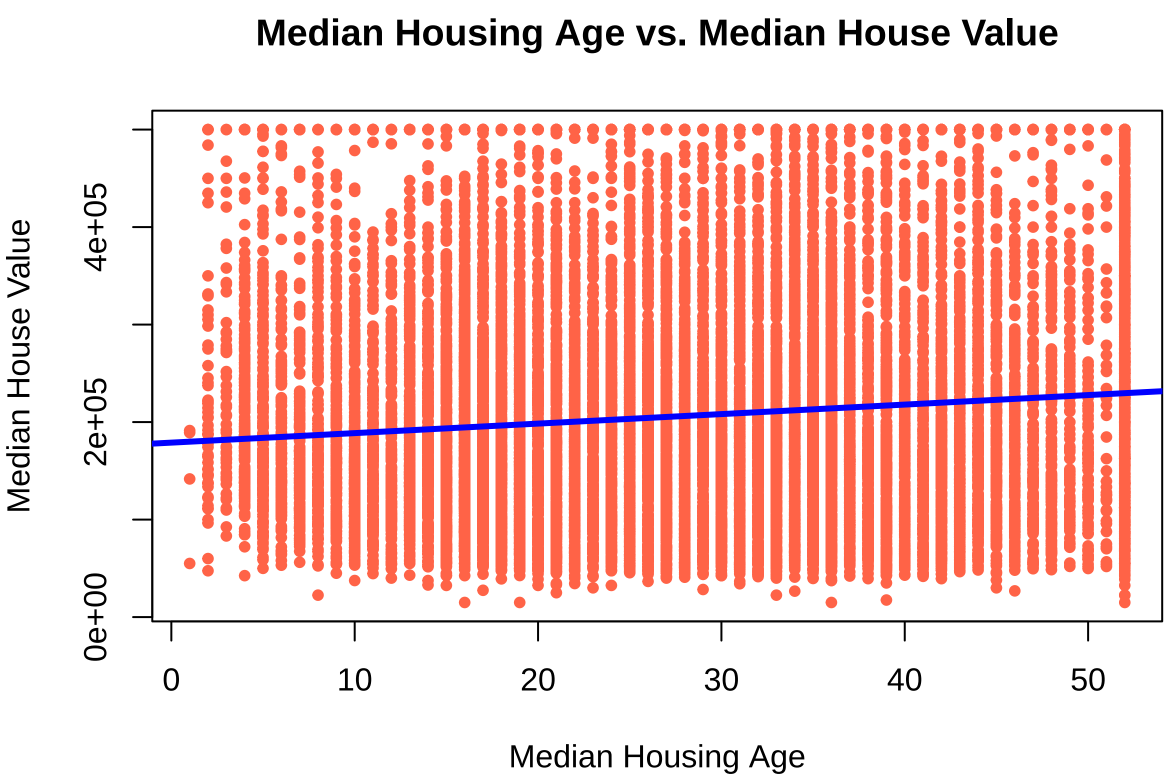
1. Fit a simple linear regression model for each predictor(X1, X2, X3) to predict the response. Determine if there is a statistically significant association between the predictor and the response. Create plots for each simple linear regression to visualize the relationships. Provide the R code. (2 points)

```r
# Fit a simple linear regression model with 'median_house_value' as the dependent variable and 'housing
lm_median_age <- lm(median_house_value ~ housing_median_age, data = cal_housing)

# Summary of the linear model to display the coefficients, R-squared value, and significance levels.
summary(lm_median_age)
```

```
##
## Call:
## lm(formula = median_house_value ~ housing_median_age, data = cal_housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -214665  -85114  -25771   58290  319123
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        178926.58    1994.76    89.7   <2e-16 ***
## housing_median_age    975.72      63.77    15.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114800 on 20431 degrees of freedom
## Multiple R-squared:  0.01133,    Adjusted R-squared:  0.01128
## F-statistic: 234.1 on 1 and 20431 DF,  p-value: < 2.2e-16
```

```r
# Plot the relationship between 'housing_median_age' and 'median_house_value'
plot(cal_housing$housing_median_age, cal_housing$median_house_value,
    xlab = "Median Housing Age", # Label for the x-axis
    ylab = "Median House Value", # Label for the y-axis
    main = "Median Housing Age vs. Median House Value", # Title of the plot
    pch = 20, # Shape of the plot points (filled circle)
    col = 'tomato')  # Color of the plot points

# Add the regression line to the plot
abline(lm_median_age, lwd = 3, col = 'blue')
```

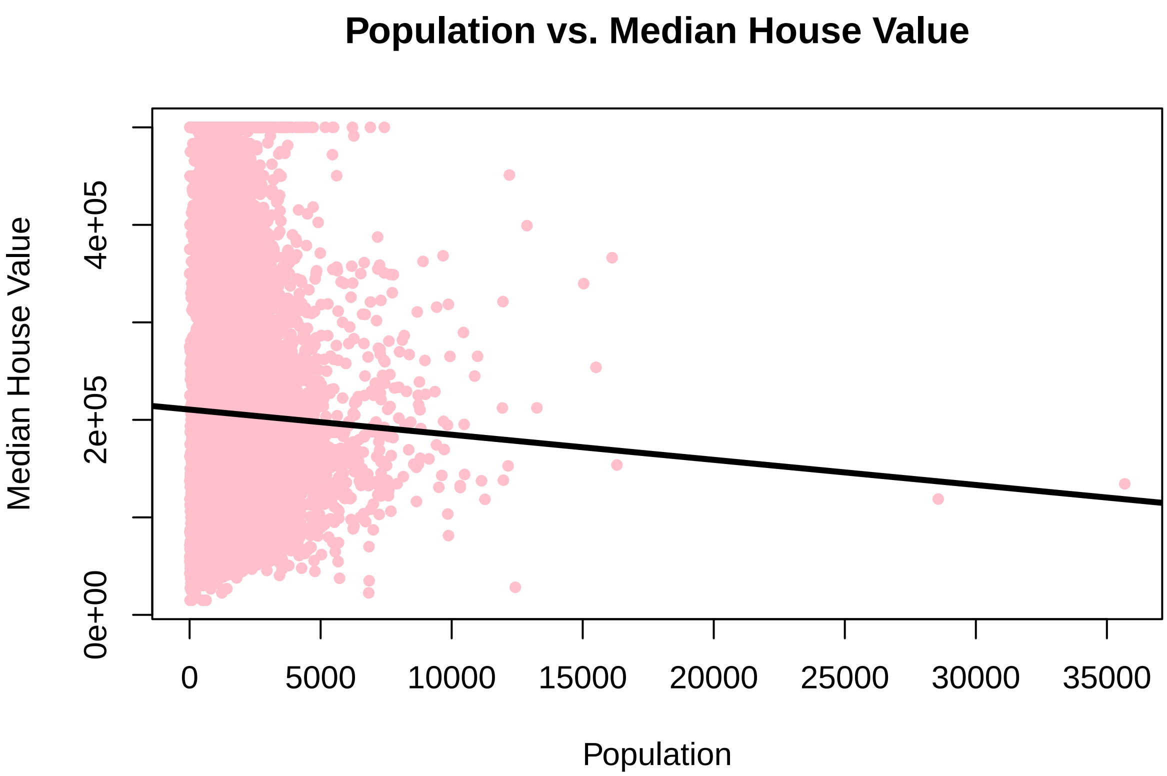


There is a statistically significant relationship between housing_median_age and median_house_value based on the small p-value (< 2e-16) and large t-statistic (15.3). The estimated coefficient for housing_median_age is 975.72, meaning that for each additional year of housing median age, the median house value increase by approximately $976, holding other factors constant. However, the R-squared value of 0.01133 (adjusted R-squared: 0.01128) is very low, indicating that only about 1.13% of the variability in median house value is explained by housing median age. While the relationship is statistically significant, it is not practically significant in terms of explaining much of the variation in median house value.

```r
# Fit a simple linear regression model with 'median_house_value' as the dependent variable and 'populat
lm_population <- lm(median_house_value ~ population, data = cal_housing)

# Summary of the linear model to display the coefficients, R-squared value, and significance levels.
summary(lm_population)
```

```
##
## Call:
## lm(formula = median_house_value ~ population, data = cal_housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -195491  -86980  -26885   58117  308615
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.105e+05  1.297e+03 162.318  < 2e-16 ***
## population  -2.577e+00  7.124e-01  -3.617 0.000298 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 115400 on 20431 degrees of freedom
## Multiple R-squared:  0.0006401,  Adjusted R-squared:  0.0005912
## F-statistic: 13.09 on 1 and 20431 DF,  p-value: 0.0002983
```

```r
# Plot the relationship between 'population' and 'median_house_value'
plot(cal_housing$population, cal_housing$median_house_value,
    xlab = "Population", # Label for the x-axis
    ylab = "Median House Value", # Label for the y-axis
    main = "Population vs. Median House Value", # Title of the plot
    pch = 20, # Shape of the plot points (filled circle)
    col = 'pink')  # Color of the plot points

# Add the regression line to the plot
abline(lm_population, lwd = 3, col = 'black')
```



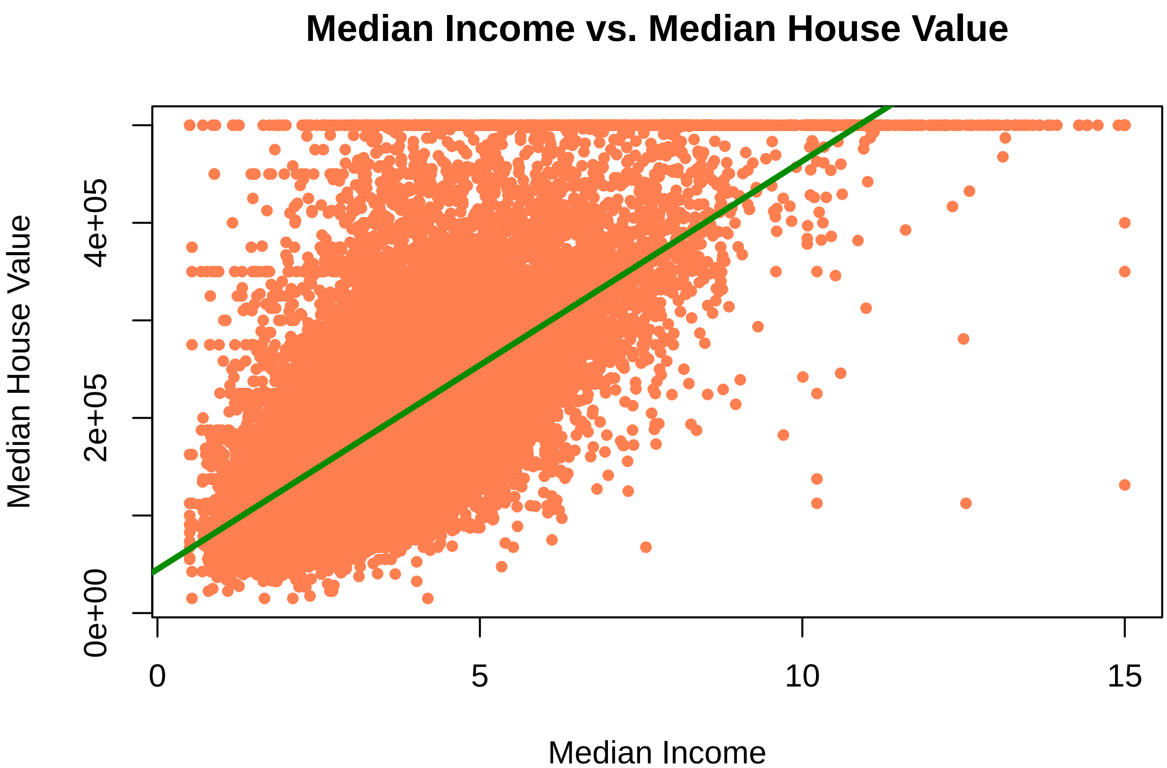
There is a statistically significant relationship between population and median house value based on the small p-value (0.00298). The t-value for population is also sufficiently large (-3.617) to indicate a statistically significant relationship. The estimated coefficient for population is -2.577, meaning that for each additional unit increase in population, the median house value decreases by approximately $2.58, holding other factors constant. However, the R-squared value of 0.0006401 (adjusted R-squared: 0.005912) is extremely low, indicating that only about 0.06% of the variability in median house value is explained by population. This indicates that population is a very weak predictor of housing values.

```
# Fit a simple linear regression model with 'median_house_value' as the dependent variable and 'median_
lm_income <- lm(median_house_value ~ median_income, data = cal_housing)

# Summary of the linear model to display the coefficients, R-squared value, and significance levels.
summary(lm_income)
```

```
##
## Call:
## lm(formula = median_house_value ~ median_income, data = cal_housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -541167  -55858  -16955   36895  434180
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    44906.4     1330.0   33.77   <2e-16 ***
## median_income  41837.1      308.4  135.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83740 on 20431 degrees of freedom
## Multiple R-squared:  0.4738, Adjusted R-squared:  0.4738
```

```
## F-statistic: 1.84e+04 on 1 and 20431 DF,  p-value: < 2.2e-16
```

```r
# Plot the relationship between 'median_income' and 'median_house_value'
plot(cal_housing$median_income, cal_housing$median_house_value,
    xlab = "Median Income", # Label for the x-axis
    ylab = "Median House Value", # Label for the y-axis
    main = "Median Income vs. Median House Value", # Title of the plot
    pch = 20, # Shape of the plot points (filled circle)
    col = 'coral')  # Color of the plot points

# Add the regression line to the plot
abline(lm_income, lwd = 3, col = 'green4')
```



There is a statistically significant relationship between median income and median house value based on the small p-value (< 2e-16) and very large t-statistic (135.64). The estimated coefficient for median income is 41,837.1, meaning that for each additional unit increase in median income, the median house value increases by approximately $41,837.10, holding other factors constant. The R-squared value of 0.4738 (adjusted R-squared: 0.4738) is relatively high, meaning that about 47.38% of the variability in median house value is explained by median income. This indicates that median income explains a substantial portion of the variation in median housing values, and can be seen as a practically significant relationship.

2. Implement a forward variable selection method using R2 as the metric. The stopping condition is when the R2 value decreases compared to the largest R2 value from the previous step. Note: Do not use any libraries that provide automatic functions for forward variable selection. Provide the R code. (2 points)

```r
# Define the response and predictors
response <- cal_housing$median_house_value
X1 <- cal_housing$housing_median_age
X2 <- cal_housing$population
X3 <- cal_housing$median_income

# Initialize variables
```

```r
predictors <- list(X1, X2, X3)
predictor_names <- c("housing_median_age", "population", "median_income")
selected_vars <- c()  # To store selected variables
best_r2 <- 0          # Store the best R-squared value
current_r2 <- 0       # R-squared in the current step

# Forward selection loop
for (step in 1:length(predictors)) {

  best_step_r2 <- 0   # Best R-squared for this step
  best_var <- NULL    # Best variable to add in this step

  # Test adding each remaining variable to the model
  for (i in 1:length(predictors)) {

    # Check if the variable is already selected
    if (predictor_names[i] %in% selected_vars) {
      next  # Skip if the variable is already selected
    }

    # Build the formula with the current variables plus the new one
    current_vars <- paste(selected_vars, collapse = " + ")
    if (current_vars == "") {
      formula <- as.formula(paste("response ~", predictor_names[i]))
    } else {
      formula <- as.formula(paste("response ~", current_vars, "+", predictor_names[i]))
    }

    # Fit the linear model
    model <- lm(formula, data = cal_housing)

    # Calculate R-squared
    r2 <- summary(model)$r.squared

    # Keep track of the best R-squared and variable in this step
    if (r2 > best_step_r2) {
      best_step_r2 <- r2
      best_var <- predictor_names[i]
    }
  }

  # Check if R-squared decreases compared to the previous step
  if (best_step_r2 < best_r2) {
    cat("Stopping: R-squared decreased.\n")
    break
  }

  # Update the selected variables and best R-squared
  selected_vars <- c(selected_vars, best_var)
  best_r2 <- best_step_r2

  # Output the selected variable and current R-squared
  cat("Step", step, ": Added", best_var, "with R-squared =", best_r2, "\n")
```

```
}
```

```
## Step 1 : Added median_income with R-squared = 0.4738333
## Step 2 : Added housing_median_age with R-squared = 0.5096212
## Step 3 : Added population with R-squared = 0.510447
```

```
# Final selected model
cat("Selected variables:", paste(selected_vars, collapse = ", "), "\n")
```

```
## Selected variables: median_income, housing_median_age, population
```

3. Check for multicollinearity issues in your best model. Provide the R code. (1 points)

```
# Fit the final model including all 3 predictors, as that resulted in the highest R-squared
final_model <- lm(median_house_value ~ median_income + housing_median_age + population, data = cal_hous

# Load car package for calculating VIF
install.packages("car")
```

```
##
## The downloaded binary packages are in
##  /var/folders/1p/m5frxr_n1c19zhr2wxwwrclh0000gn/T//Rtmpea05wM/downloaded_packages
```

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```
# Calculate VIF for the final model
vif_values <- vif(final_model)

# Print VIF values
print(vif_values)
```

```
##      median_income housing_median_age         population
##           1.015197           1.112504           1.096968
```

```
# Check if any VIF value exceeds the common threshold of 5 or 10
if (any(vif_values > 5)) {
  cat("Warning: Potential multicollinearity issue detected (VIF > 5).\n")
} else {
  cat("No multicollinearity issues detected (VIF < 5).\n")
}
```

```
## No multicollinearity issues detected (VIF < 5).
```