QM3 Assignment 2 Code

2025-01-11

QM3 Assignment 2 Markdown File

England's Third Spaces

Investigating the Effectiveness of London Pubs in Maintaining Good Mental Health.

Abstract

This project investigates the posited positive impact of pubs on mental health in London.

We well begin with the null hypothesis, H0, that there is no correlation between the number of pubs in a borough, and the number of mental health issues in the same borough. If this is found to not be the case at the $p<0.01$ significance level, we will reject the null hypothesis, H0, and instead accept an alternative hypothesis, H1.

Code

1. Setting Up

Set up packages.

```
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(stargazer)
```

```
##
## Please cite as:
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
library(ggplot2)
library(dplyr)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(HistData)
library(performance)
```

Read data files.

```r
getwd()
```

```
## [1] "/Users/jacob/Desktop/Admin/UCL/Degree/3rd Year Modules/Term 1/BASC0056/Assessments/Assessment 2/
```

```r
esa_data <-
  read.csv('esa-mental-behavioural-disorders-benefit-claimants-borough copy.csv')
hours_data <- read.csv('hours-worked.csv')
income_data <- read.csv('ons-model-based-income-estimates-msoa.csv')
pubs_data <- read.csv('Pubs.csv')
population_data <- read.csv('historical-census-tables-2021.csv')
```

2. Data Cleaning

Clean census dataset to extract relevant population data.

```r
pop_data_clean <- population_data[,c(colnames(population_data)[1],'X.20')]
colnames(pop_data_clean)<-c("Area","Pop 2011 (per 1000)")
pop_data_clean <- pop_data_clean[-c(0:3),]
rownames(pop_data_clean) <- NULL
pop_data_clean <- pop_data_clean[-34,]
rownames(pop_data_clean) <- NULL
pop_data_clean <- pop_data_clean[-c(34:92),]
```

Clean Mental Health ESA dataset to extract relevant mental health data.

```
esa_data_clean <- esa_data[,c('Area',
                              'May.2011',
                              'May.2012',
                              'May.2013',
                              'May.2014',
                              'May.2015',
                              'May.2016',
                              'May.2017',
                              'May.2018')]
colnames(esa_data_clean)<-c("Area",
                            "ESA 2011",
                            "ESA 2012",
                            "ESA 2013",
                            "ESA 2014",
                            "ESA 2015",
                            "ESA 2016",
                            "ESA 2017",
                            "ESA 2018")
esa_data_clean <- esa_data_clean[-c(34:53),]
```

Clean working hours dataset.

```
hours_data_clean <- hours_data[,
                               c('X.1', colnames(hours_data)[c(3, 7, 11, 15)])]
colnames(hours_data_clean) <- c("Area",
                                "Hours 10",
                                "Hours 10-34",
                                "Hours 34-44",
                                "Hours 45")
hours_data_clean <- hours_data_clean[-c(0:2),]
rownames(hours_data_clean) <- NULL
hours_data_clean <- hours_data_clean[-c(34:52),]
```

Clean income dataset.

```
income_data_clean <- income_data %>% filter(grepl("London",
                                                  Region.name,
                                                  ignore.case = TRUE))
income_data_clean <- income_data_clean[, -c(1:3, 5:18, 20:22)]
```

Convert income data into a usable form.

```
income_matrix <- matrix(ncol=2)
cum_income <- 0
counter <- 0
prev_region <- as.character(income_data_clean[1, 1])

for(i in 1:dim(income_data_clean)[1]) {
  region <- as.character(income_data_clean[i, 1])
  if(region == prev_region){
    cum_income <- cum_income + income_data_clean[i, 2]
    counter <- counter + 1
```

```
  }
  else {
    av_income <- cum_income/counter
    new_row <- c(prev_region, av_income)
    income_matrix <- rbind(income_matrix, new_row)
    cum_income <- 0 + income_data_clean[i, 2]
    counter <- 1
  }
  prev_region <- region
}
```

Finish cleaning income data.

```
income_matrix <- income_matrix[-1,]
rownames(income_matrix) <- NULL
colnames(income_matrix) <- c("Area", "Average Income after Housing Cost")

income_data_cleaned <- as.data.frame(income_matrix, stringsAsFactors = FALSE)
```

Clean pubs data and infer number of pubs in each borough.

```
pubs_vector <- pubs_data[, 5]
pubs_matrix <- matrix(data = c("None",0), ncol = 2)
colnames(pubs_matrix) <- c("Area", "Pubs")

for(i in 1:length(pubs_vector)) {
  region <- pubs_vector[i]
  duplicate <- FALSE
  for(j in 1:dim(pubs_matrix)[1]){
    if(region == pubs_matrix[j, 1]){
      pubs_matrix[j, 2] <- as.numeric(pubs_matrix[j, 2]) + 1
      duplicate <- TRUE
    }
  }
  if(duplicate == FALSE){
    new_row <- c(region, as.numeric(1))
    pubs_matrix <- rbind(pubs_matrix, new_row)
  }
}

pubs_matrix <- pubs_matrix[-1,]
rownames(pubs_matrix) <- NULL
pubs_matrix[19, 1] <- "City of London"
pubs_matrix[30, 1] <- "Westminster"


pubs_data_cleaned <- as.data.frame(pubs_matrix, stringsAsFactors = FALSE)
```

Merge data into one dataset for convenience.

```
data_set <- merge(esa_data_clean, hours_data_clean, by = "Area", all = TRUE)
data_set <- merge(data_set, income_data_cleaned, by = "Area", all = TRUE)
```

```r
data_set <- merge(data_set, pop_data_clean, by = "Area", all = TRUE)
data_set <- merge(data_set, pubs_data_cleaned, by = "Area", all = TRUE)
```

Remove null values.

```r
data_set[-which(names(data_set) == "Area")] <-
  lapply(data_set[-which(names(data_set) == "Area")],
         function(x) as.numeric(gsub(",", "", x)))
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```r
data_set <- na.omit(data_set)
head(data_set)
```

```
##                    Area ESA 2011 ESA 2012 ESA 2013 ESA 2014 ESA 2015 ESA 2016
## 1 Barking and Dagenham      890     1540     2310     3010     3520     3350
## 2                Barnet     1130     2250     3470     4120     4760     4820
## 3                Bexley      750     1430     2120     2610     2920     2980
## 4                 Brent     1090     2260     3470     4280     4850     4970
## 5               Bromley      960     1720     2840     3500     3930     3930
## 7        City of London       10       10       30       50       60       60
##   ESA 2017 ESA 2018 Hours 10 Hours 10-34 Hours 34-44 Hours 45
## 1     3290     3190     6200       41600       73100    53000
## 2     4890     4960     5100       25500       51000    24300
## 3     2990     3160     2600       39500       68000    33900
## 4     4910     5090     5100       34900       68500    41000
## 5     4060     4210     4700       21300       36900    38100
## 7       60       60     1900       22200       38500    14800
##   Average Income after Housing Cost Pop 2011 (per 1000) Pubs
## 1                          410.4545              185911   29
## 2                          581.4634              356386   77
## 3                          557.5000              231997   85
## 4                          475.0000              311215  109
## 5                          630.2564              309392  108
## 7                          760.0000                7375  215
```

3. Removing Outliers

Take a first look at our dataset.

```r
stargazer(data_set,
          type="text",
          digits=1,
          title ="Table 1: summary statistics",
          summary=TRUE)
```

```
##
```

```
## Table 1: summary statistics
## ======================================================================
## Statistic                           N    Mean    St. Dev.  Min     Max
## ----------------------------------------------------------------------
## ESA 2011                           31   997.7    362.6     10    1,510
## ESA 2012                           31  1,868.1   669.8     10    2,800
## ESA 2013                           31  2,942.3  1,076.2    30    4,580
## ESA 2014                           31  3,619.4  1,352.7    50    5,630
## ESA 2015                           31  4,131.6  1,566.9    60    6,420
## ESA 2016                           31  4,078.7  1,546.6    60    6,290
## ESA 2017                           31  3,972.9  1,535.2    60    6,290
## ESA 2018                           31  3,932.9  1,521.6    60    6,440
## Hours 10                           31  3,512.9  1,511.5  1,500   6,200
## Hours 10-34                        31 28,696.8  8,496.6  14,900 54,100
## Hours 34-44                        31 53,564.5  13,665.2 23,100 80,400
## Hours 45                           31 36,206.5  10,761.5 14,800 71,200
## Average Income after Housing Cost 31   548.2     86.8    383.5   760.0
## Pop 2011 (per 1000)                31 248,237.2 71,568.6 7,375  363,378
## Pubs                               31   121.3     81.7     29     457
## ----------------------------------------------------------------------
```

Define a function to remove outliers.

```r
remove_outliers <- function(input_data, column_name) {
  working_matrix <- data.frame(
    Values = input_data[[column_name]],
    Mean = mean(input_data[[column_name]]),
    SD = sd(input_data[[column_name]])
  )
  input_data$z_scores <-
    (working_matrix$Values - working_matrix$Mean) / working_matrix$SD
  outliers_removed <- input_data[(input_data$z_scores)^2 <= 9, ]
  rownames(outliers_removed) <- NULL
  outliers_removed$z_scores <- NULL
  return(outliers_removed)
}
```

Turning attention to the ESA data, check whether the distribution of ESA claims across boroughs is similar year to year.

```r
ggplot() +
  geom_line(data = data_set,
            aes(x = Area, y = `ESA 2011`, group = 1),
            color = "blue",
            size = 0.5) +
  geom_line(data = data_set,
            aes(x = Area, y = `ESA 2012`, group = 1),
            color = "red",
            size = 0.5) +
  geom_line(data = data_set,
            aes(x = Area, y = `ESA 2013`, group = 1),
            color = "yellow",
            size = 0.5) +
```

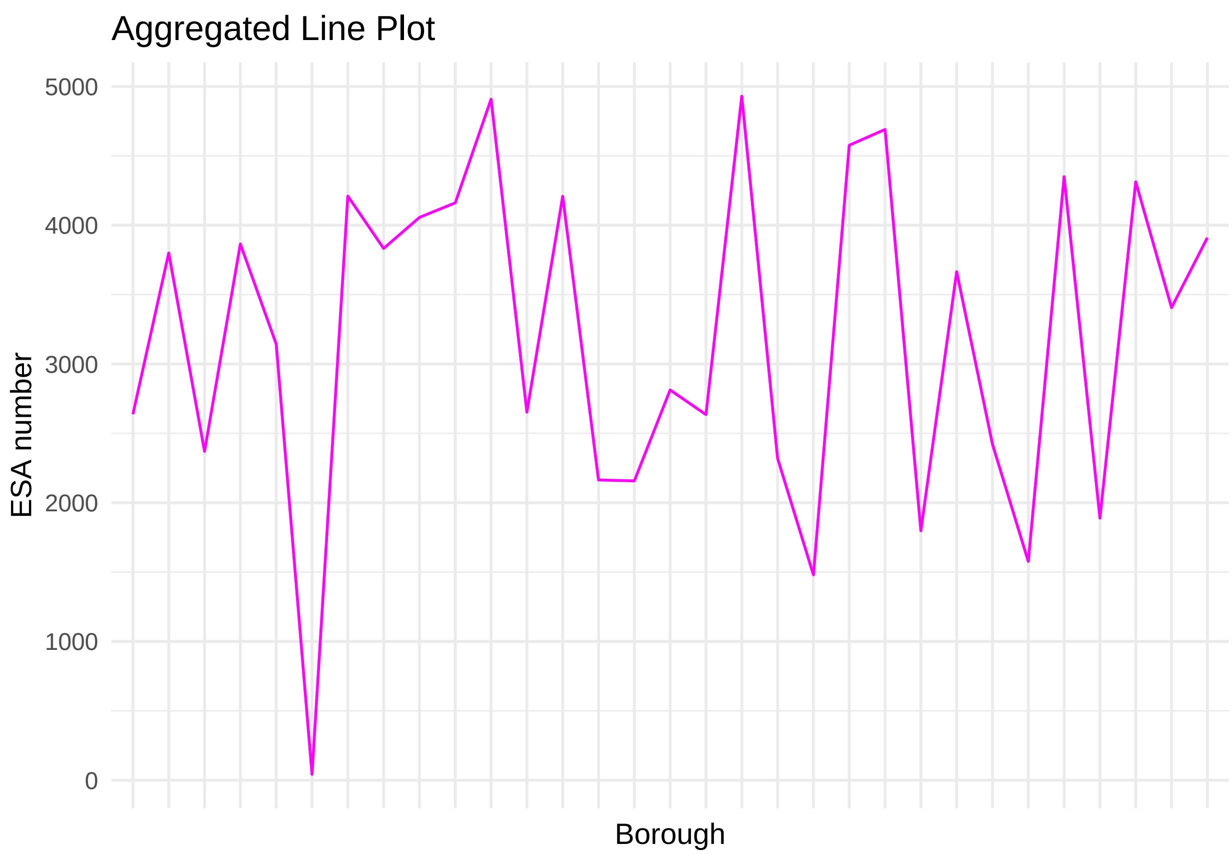
```r
  geom_line(data = data_set,
            aes(x = Area, y = `ESA 2014`, group = 1),
            color = "green",
            size = 0.5) +
  geom_line(data = data_set,
            aes(x = Area, y = `ESA 2015`, group = 1),
            color = "purple",
            size = 0.5) +
  geom_line(data = data_set,
            aes(x = Area, y = `ESA 2016`, group = 1),
            color = "orange",
            size = 0.5) +
  geom_line(data = data_set,
            aes(x = Area, y = `ESA 2017`, group = 1),
            color = "pink",
            size = 0.5) +
  geom_line(data = data_set,
            aes(x = Area, y = `ESA 2018`, group = 1),
            color = "grey",
            size = 0.5) +
  labs(
    title = "Overlayed ESA Line Plots",
    x = "Borough",
    y = "ESA number"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_blank()
  )
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

ESA score varies across boroughs in roughly the same shape. Therefore it is appropriate to aggregate this and create an average ESA column which is representative of the period of data.

```
cum_ESA <- rowSums(data_set[, c("ESA 2011",
                                "ESA 2012",
                                "ESA 2013",
                                "ESA 2014",
                                "ESA 2015",
                                "ESA 2016",
                                "ESA 2017",
                                "ESA 2018")])
av_ESA <- cum_ESA/8
data_set$Av.ESA <- av_ESA
data_set <- data_set[, !colnames(data_set) %in% c("ESA 2011",
                                                  "ESA 2012",
                                                  "ESA 2013",
                                                  "ESA 2014",
                                                  "ESA 2015",
                                                  "ESA 2016",
                                                  "ESA 2017",
                                                  "ESA 2018")]

ggplot() +
  geom_line(data = data_set,
            aes(x = Area, y = Av.ESA, group = 1),
            color = "magenta",
```

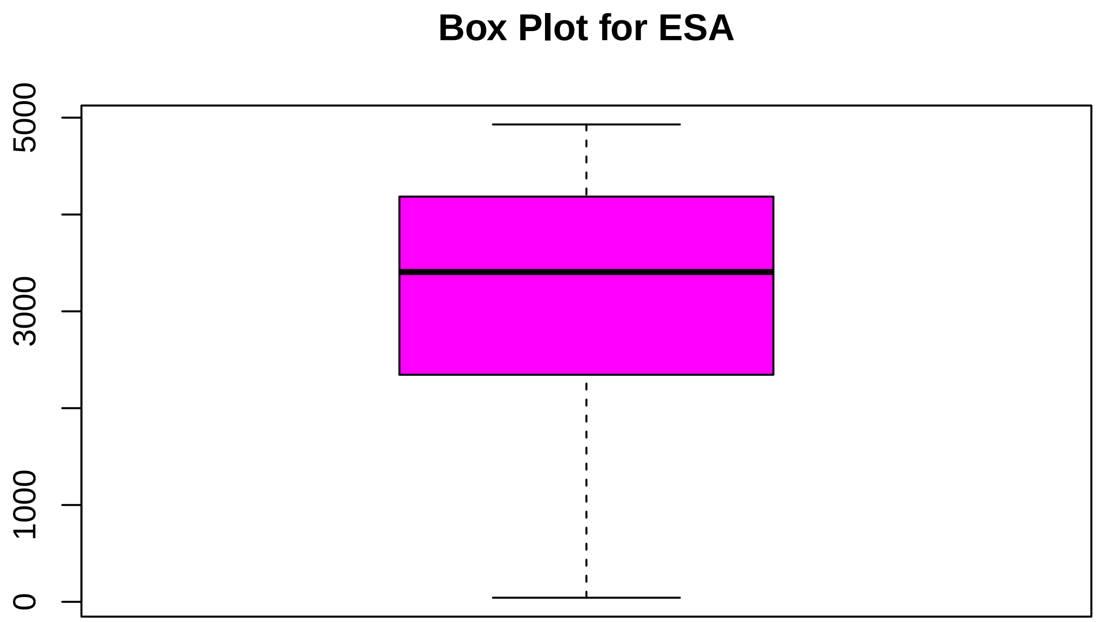
```
              size = 0.5) +
  labs(
    title = "Aggregated Line Plot",
    x = "Borough",
    y = "ESA number"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_blank()
  )
```



This aggregated data has a very similar shape to each of the years individually, so is suitable for use.

Now, need to check aggregated ESA data for outliers.

```
boxplot(data_set$Av.ESA,
        main = "Box Plot for ESA",
        col = "magenta")
```



No points lie outside the box plot, so there are no outliers to remove.

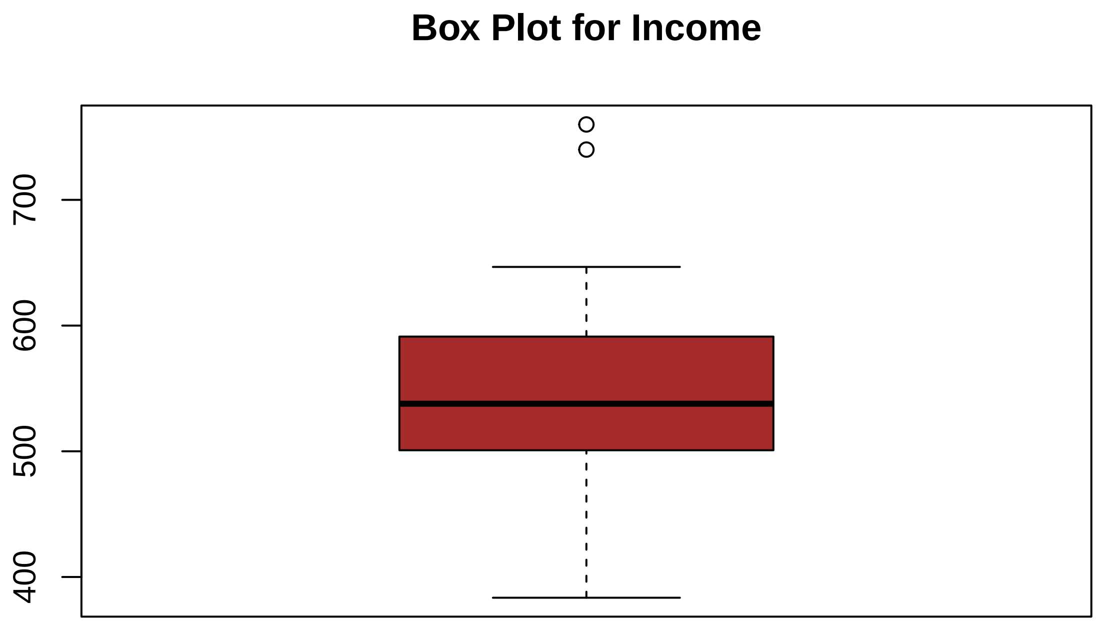
Now it's time to separate the hours worked dataset into bins based on the modal average.

```
columns_to_check <- c("Hours 10", "Hours 10-34", "Hours 34-44", "Hours 45")
Mode.Hours <- colnames(data_set)[max.col(data_set[, columns_to_check],
                                          ties.method = "first")]
data_set$Mode.Hours <- Mode.Hours
```

Since there are no actual values given in this data, just ranges, it is impossible to identify outliers.

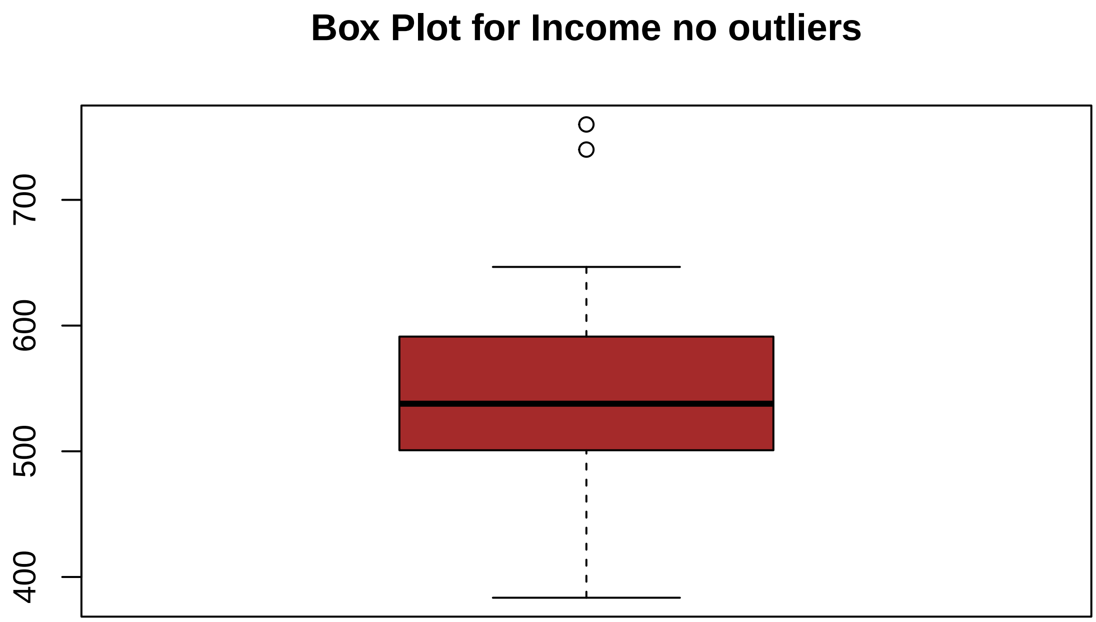
Move on to considering the income data.

```
boxplot(data_set$`Average Income after Housing Cost`,
        main = "Box Plot for Income",
        col = "brown")
```



A couple of points might be outliers, so check this.

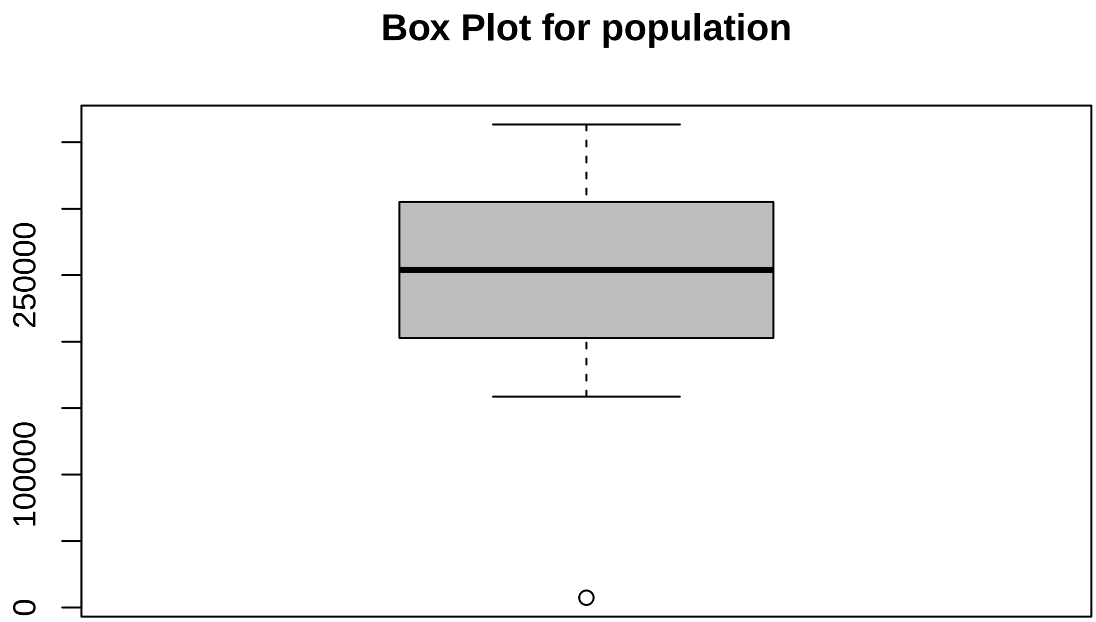
```
data_set <- remove_outliers(data_set, "Average Income after Housing Cost")
boxplot(data_set$`Average Income after Housing Cost`,
        main = "Box Plot for Income no outliers",
        col = "brown")
```



It turns out those points were not outliers according to our function, so that is good.

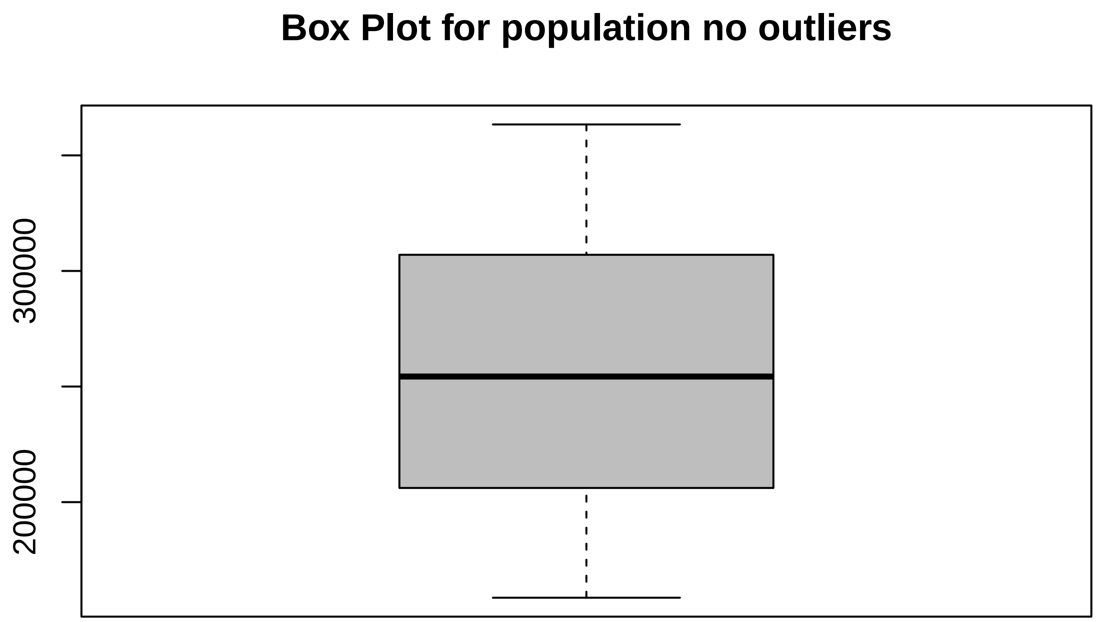
Now it's time to look at the population data.

```
boxplot(data_set$`Pop 2011 (per 1000)`,
        main = "Box Plot for population",
        col = "grey")
```



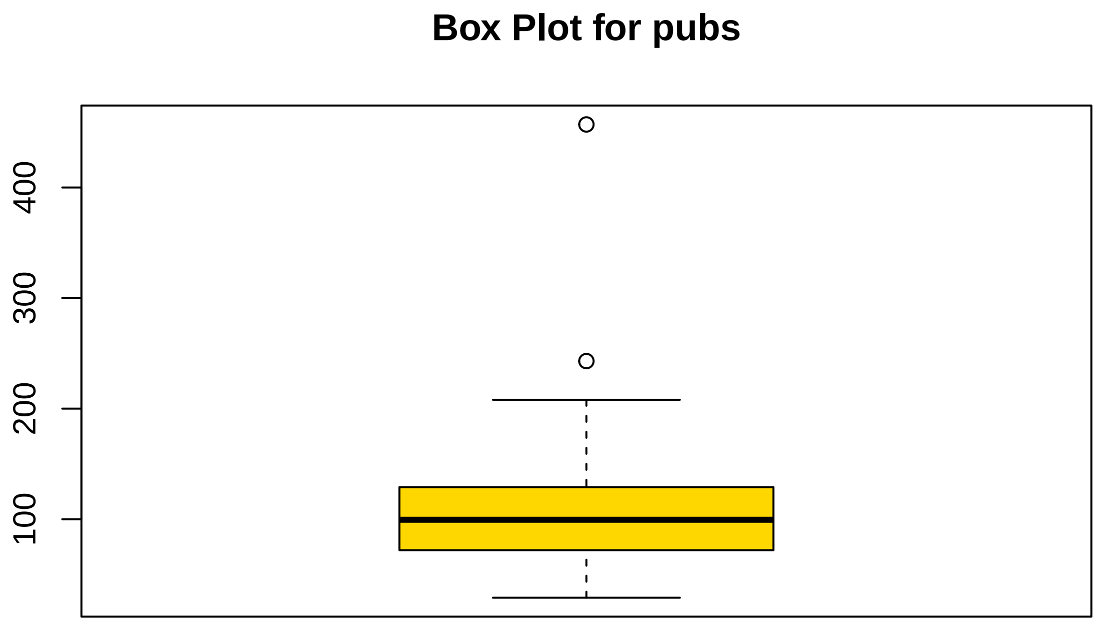
There definitely seems to be an outlier here, so remove it.

```r
data_set <- remove_outliers(data_set, "Pop 2011 (per 1000)")
boxplot(data_set$`Pop 2011 (per 1000)`,
        main = "Box Plot for population no outliers",
        col = "grey")
```



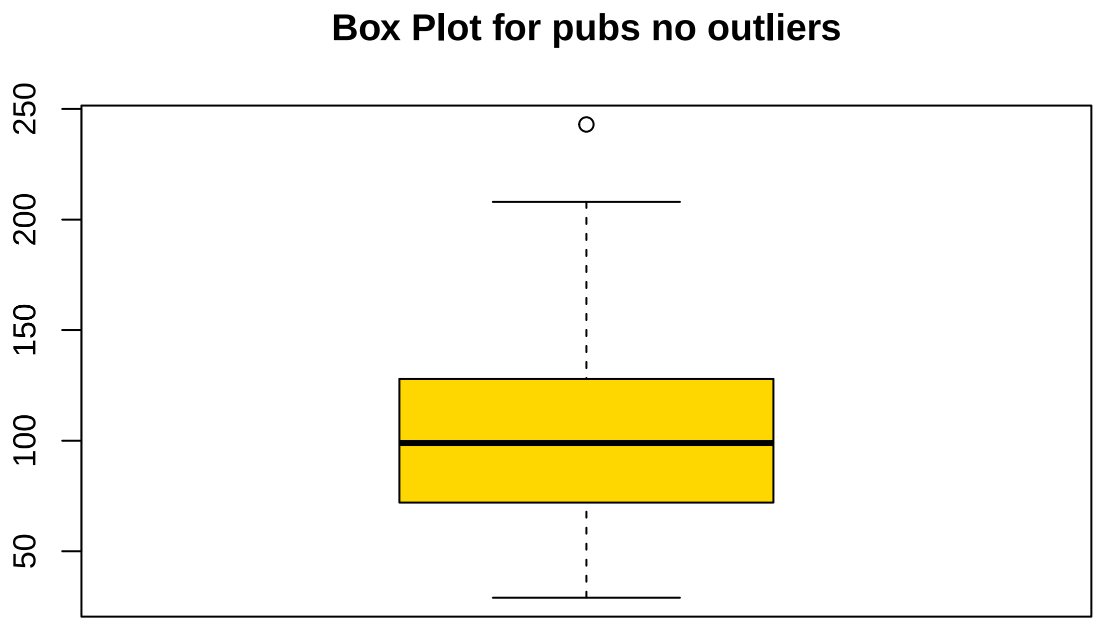
Finally, we turn our attention to the pubs data.

```
boxplot(data_set$Pubs,
        main = "Box Plot for pubs",
        col = "gold")
```



There looks like there might be an outlier, so let's check.

```
data_set <- remove_outliers(data_set, "Pubs")
boxplot(data_set$Pubs,
        main = "Box Plot for pubs no outliers",
        col = "gold")
```



Again, it turned out not to be an outlier, which is good.

Now we've removed these outliers, let's take another overview of our data.

```
stargazer(data_set,
          type="text",
          digits=1,
          title ="Table 2: summary statistics",
          summary=TRUE)
```

```
##
## Table 2: summary statistics
## =======================================================================
## Statistic                          N    Mean    St. Dev.   Min     Max
## -----------------------------------------------------------------------
## Hours 10                           29  3,613.8  1,510.8   1,500   6,200
## Hours 10-34                        29 29,300.0  8,434.8  14,900  54,100
## Hours 34-44                        29 54,417.2  13,707.0 23,100  80,400
## Hours 45                           29 36,620.7  10,217.1 23,700  71,200
## Average Income after Housing Cost  29   539.3     79.5    383.5   740.0
## Pop 2011 (per 1000)                29 257,537.3 57,416.7 158,649 363,378
## Pubs                               29   106.4     50.9     29      243
## Av.ESA                             29  3,276.9  1,071.9  1,480.0 4,930.0
## -----------------------------------------------------------------------
```

4. Running the Regression Models

Now we're happy with our data, it's time to run some OLS models on it.

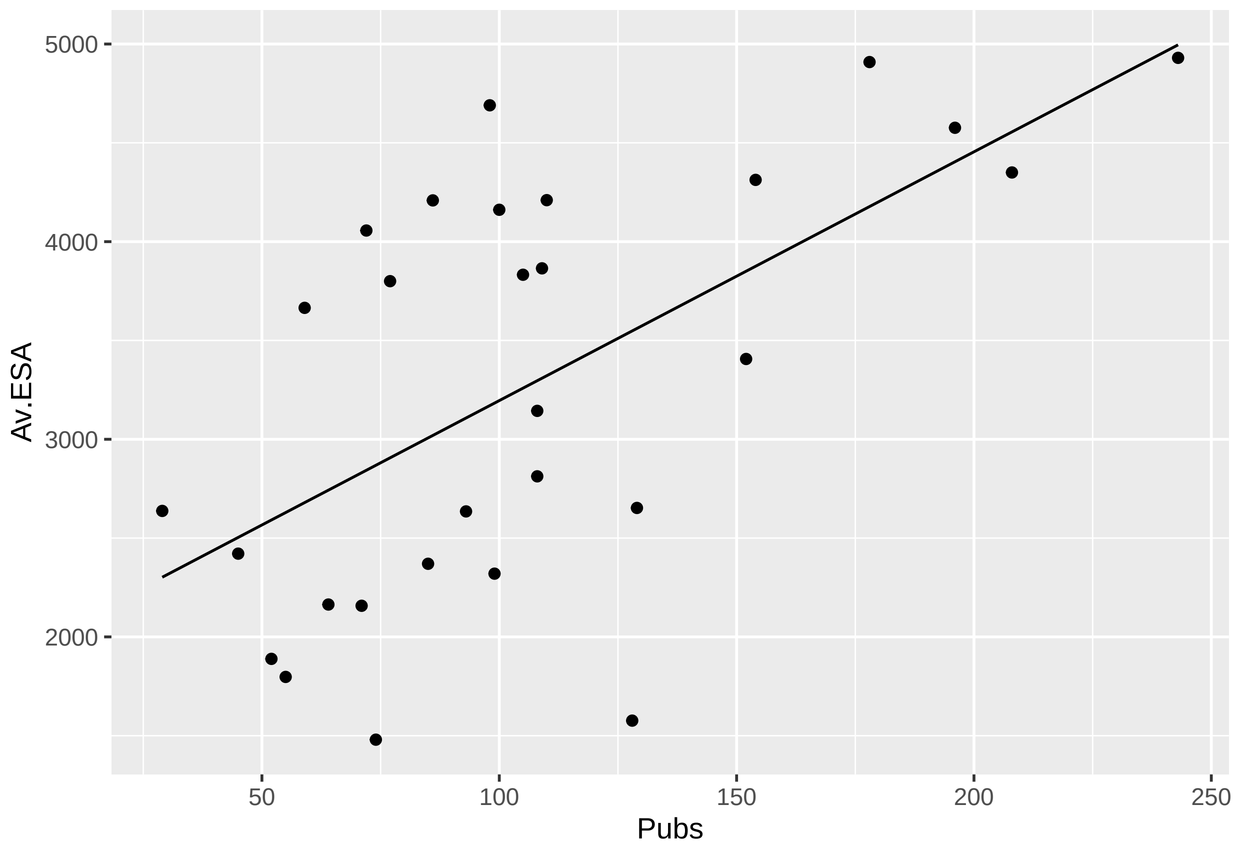
Because the sample size is so low (convention suggests that 30 data points is the minimum, so we're on shaky territory), we will only accept results at the p<0.01 significance level.

We'll start with the most simple OLS model, where we just consider ESA claims against the number of pubs.

```r
model_1 <- lm(Av.ESA ~ Pubs, data = data_set)

stargazer(model_1, type = "text",
          title ="Table 3: ESA against Pubs",
          dep.var.labels = "ESA Score",
          covariate.labels = "Pubs")
```

```
##
## Table 3: ESA against Pubs
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                              ESA Score
## ------------------------------------------------
## Pubs                         12.588***
##                               (3.252)
##
## Constant                   1,936.837***
##                              (382.419)
##
## ------------------------------------------------
## Observations                     29
## R2                             0.357
## Adjusted R2                    0.333
## Residual Std. Error      875.361 (df = 27)
## F Statistic           14.986*** (df = 1; 27)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
ggplot(model_1, aes(x = Pubs, y=Av.ESA)) +
  geom_point() +
  geom_line(aes(y=.fitted))
```



Suggests a significant positive correlation, which we expect, since highly populated areas will likely have more pubs and more mental health issues, simply because there are more people. We expect the correlation to flip when we account for the confounder which is population per borough, as we do in model 2.

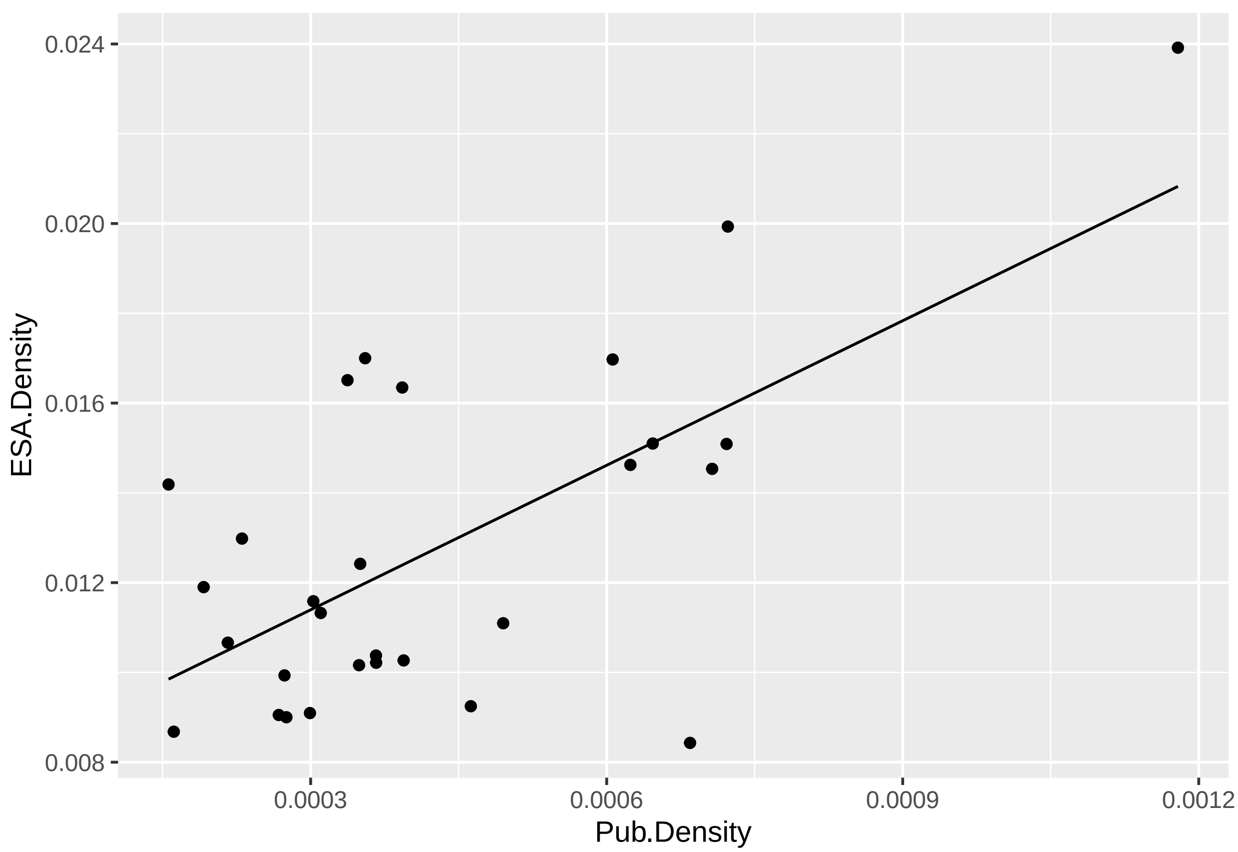
Rather than including population as another independent variable in the model, we will simply divide it into the variables that are currently in the model, as a dataset with such a small sample size will break down with too many variables.

```
pub.density <- data_set$Pubs / data_set$`Pop 2011 (per 1000)`
esa.density <- data_set$Av.ESA / data_set$`Pop 2011 (per 1000)`
data_set$ESA.Density <- esa.density
data_set$Pub.Density <- pub.density

model_2 <- lm(ESA.Density ~ Pub.Density, data = data_set)

stargazer(model_2, type = "text",
          title ="Table 3: ESA density against Pub density",
          dep.var.labels = "ESA density",
          covariate.labels = "Pubs per person")
```

```
##
## Table 3: ESA density against Pub density
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                             ESA density
```

```
## -------------------------------------------------
## Pubs per person              10.733***
##                               (2.405)
##
## Constant                      0.008***
##                               (0.001)
##
## -------------------------------------------------
## Observations                     29
## R2                              0.424
## Adjusted R2                     0.403
## Residual Std. Error        0.003 (df = 27)
## F Statistic            19.915*** (df = 1; 27)
## =================================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
```

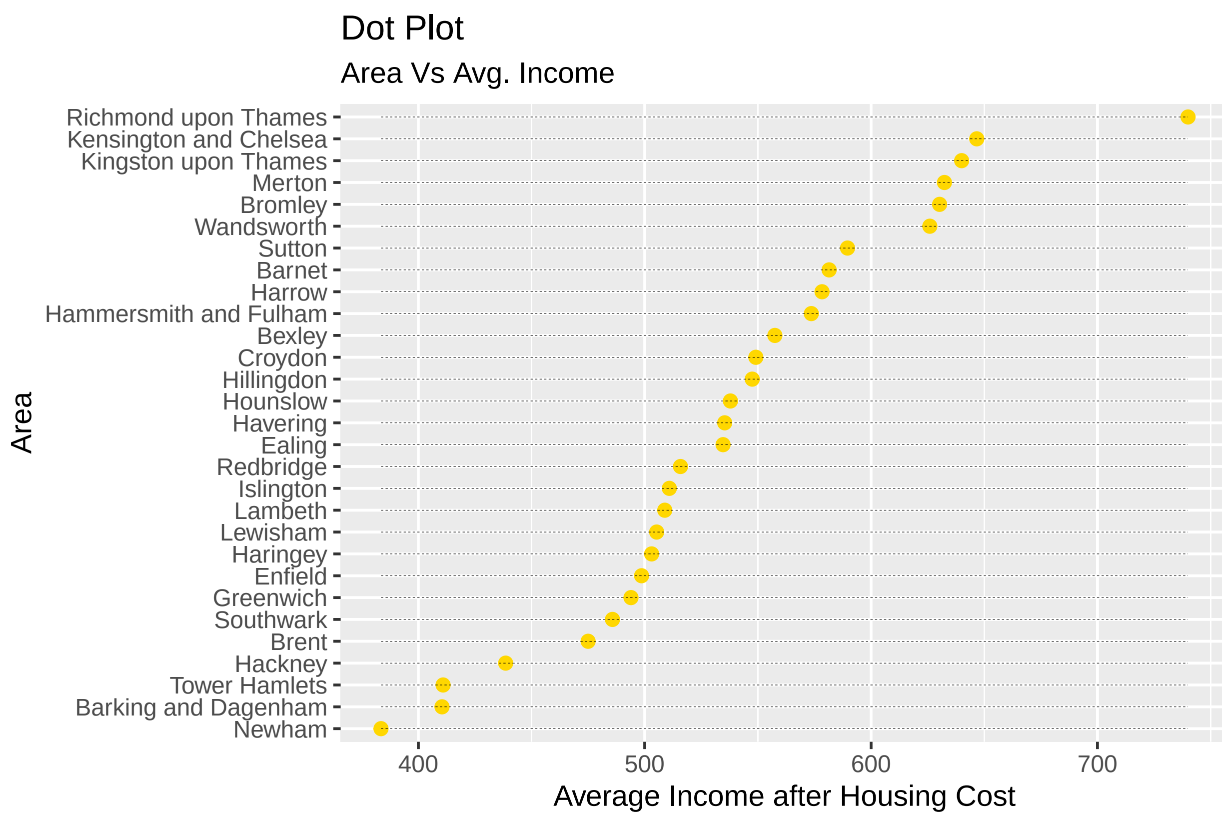
```r
ggplot(model_2, aes(x = Pub.Density, y=ESA.Density)) +
  geom_point() +
  geom_line(aes(y=.fitted))
```



Surprisingly, there continues to be a significant positive correlation, which may be due to another confounding variable: each borough's wealth. A richer area may have fewer mental health issues, along with fewer institutions considered to be common, a category into which pubs may fall.

This time, consider the average inhabitant's income as another independent variable, rather than dividing it in. To do this, we will try and stratify the income data into sensible bins.

```r
data_set_ordered <-
  data_set[order(data_set$`Average Income after Housing Cost`), ]
data_set_ordered$Area <-
  factor(data_set_ordered$Area, levels = data_set_ordered$Area)

ggplot(data_set_ordered, aes(x=Area, y=`Average Income after Housing Cost`)) +
  geom_point(col="gold", size=2) +
  geom_segment(aes(x=Area,
                   xend=Area,
                   y=min(`Average Income after Housing Cost`),
                   yend=max(`Average Income after Housing Cost`)),
               linetype="dashed",
               size=0.1) +
  labs(title="Dot Plot",
       subtitle="Area Vs Avg. Income",
       caption="") +
  coord_flip()
```



There don't seem to be obvious clusters unfortunately. Let's see if the code can come up with reasonable suggestions.

```r
set.seed(123)
km.res <- kmeans(data_set$`Average Income after Housing Cost`, 3, nstart = 550)
print(km.res)
```

```
## K-means clustering with 3 clusters of sizes 4, 10, 15
```

```
##
## Cluster means:
##       [,1]
## 1 410.8692
## 2 623.8249
## 3 517.2714
##
## Clustering vector:
##  [1] 1 2 3 3 2 3 3 3 3 1 2 3 2 3 3 3 3 2 2 3 3 2 1 3 2 3 2 1 2
##
## Within cluster sum of squares by cluster:
## [1]  1515.924 21958.433  8727.857
##  (between_SS / total_SS =  81.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

These clusters don't look sensible. It's probably just as effective to eyeball the clusters according to what looks like it makes sense.

```
ealing <- data_set$`Average Income after Housing Cost`[7]
wansworth <- data_set$`Average Income after Housing Cost`[29]

Strat.Income <- numeric(29)
strat1 <- (min(data_set$`Average Income after Housing Cost`) + ealing)/13
strat2 <- (ealing + wansworth)/10
strat3 <- (wansworth + max(data_set$`Average Income after Housing Cost`))/6

for (i in 1:29) {
  if (data_set$`Average Income after Housing Cost`[i] < ealing) {
    Strat.Income[i] <- strat1
  } else if (data_set$`Average Income after Housing Cost`[i] < wansworth) {
    Strat.Income[i] <- strat2
  } else {
    Strat.Income[i] <- strat3
  }
}

data_set$Strat.Income <- as.factor(Strat.Income)
```

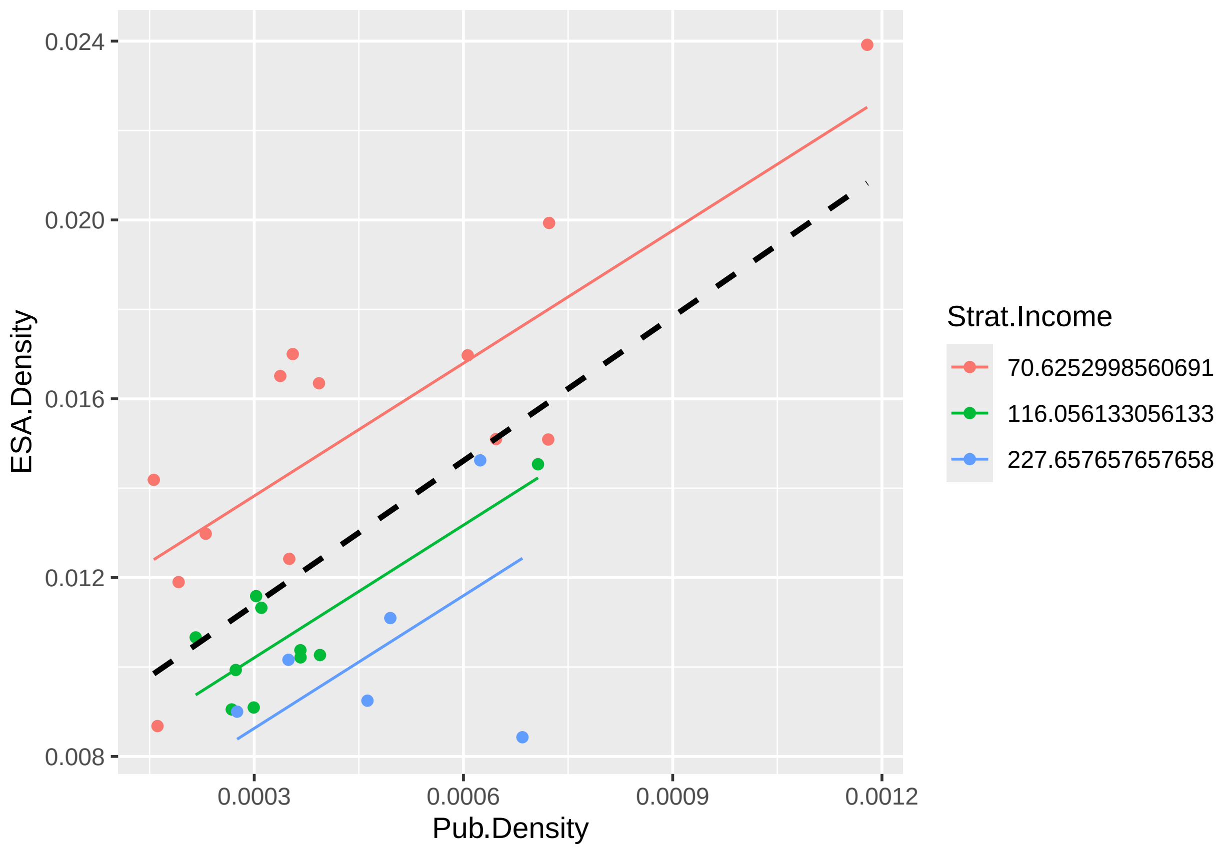
```
model_3 <-lm(ESA.Density ~ Pub.Density + Strat.Income, data=data_set)

stargazer(model_3, type = "text",
          title ="Table 4: ESA density & Pub density with income",
          dep.var.labels = "ESA density",
          covariate.labels = "Pubs per person", "Income")
```

```
##
## Table 4: ESA density & Pub density with income
## ========================================================
##                                  Dependent variable:
```

```
##                               ---------------------------
##                                       ESA density
## ---------------------------------------------------------
## Pubs per person                        9.894***
##                                         (1.633)
##
## Strat.Income116.056133056133          -0.004***
##                                         (0.001)
##
## Strat.Income227.657657657658          -0.005***
##                                         (0.001)
##
## Constant                               0.011***
##                                         (0.001)
##
## ---------------------------------------------------------
## Observations                              29
## R2                                       0.771
## Adjusted R2                              0.743
## Residual Std. Error                0.002 (df = 25)
## F Statistic                   27.988*** (df = 3; 25)
## =========================================================
## Note:                         *p<0.1; **p<0.05; ***p<0.01
##
## Table 4: ESA density & Pub density with income
## ======
## Income
## ------
```

```
ggplot(model_3, aes(x = Pub.Density, y=ESA.Density, color=Strat.Income)) +
  geom_point() +
  geom_line(aes(y=.fitted))+
  geom_smooth(method="lm", se=F,
              aes(group=1),
              color="black",
              linetype = "dashed")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Interestingly, very little changes, but there is still a statistically significant relationship. The R^2 value is higher, suggesting adding this confounder improves the model. It is surprising that the relationship is still positive.

Now we apply the moderating variable, which again needs to be stratified.

For now, we will ignore the confounding income variable, due to the small sample size, but we will add it back in later.

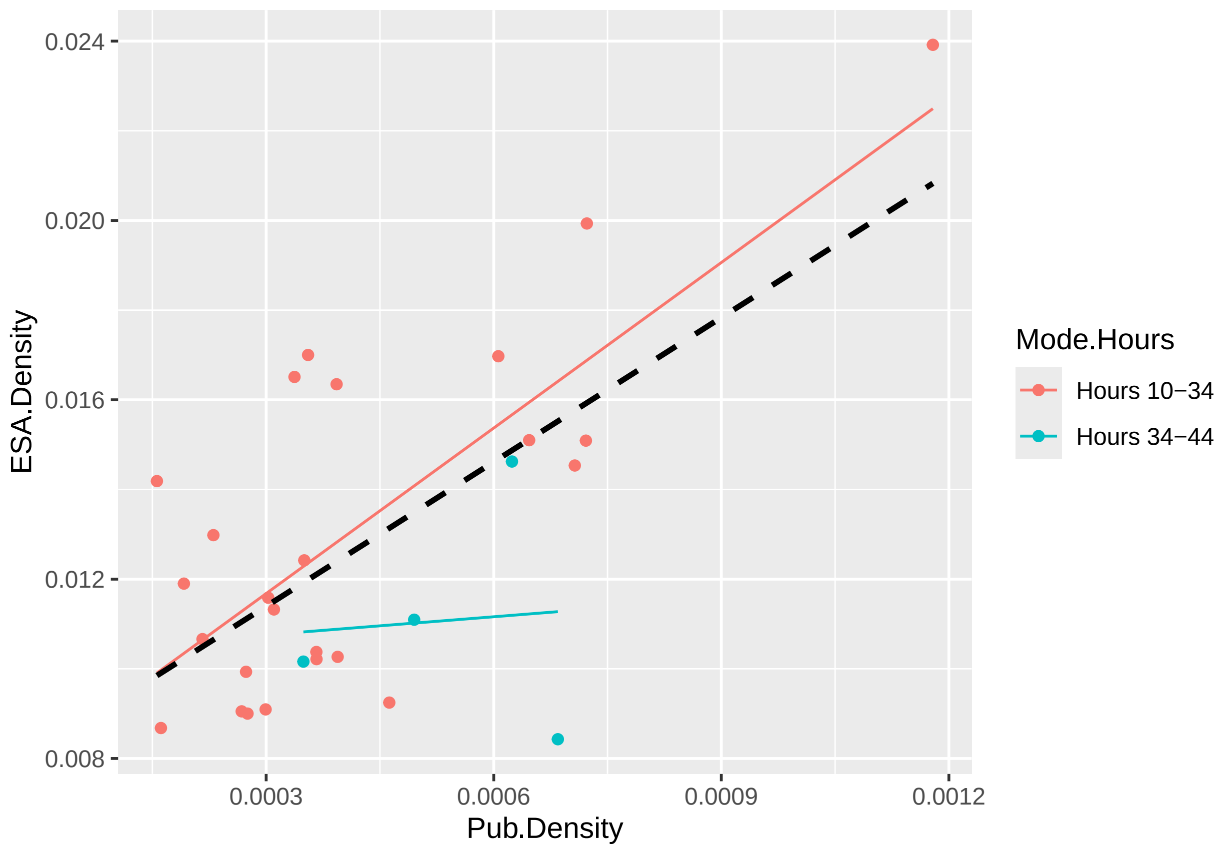
```
model_4 <-lm(ESA.Density ~ Pub.Density + Mode.Hours + Pub.Density*Mode.Hours,
             data=data_set)

stargazer(model_4, type = "text",
          title ="Table 5: ESA density & Pub density with free time",
          dep.var.labels = "ESA density",
          covariate.labels = "Pubs per person", "Hours Worked")
```

```
##
## Table 5: ESA density & Pub density with free time
## =============================================================
##                                       Dependent variable:
##                                   ---------------------------
##                                           ESA density
## -------------------------------------------------------------
## Pubs per person                            12.304***
##                                             (2.321)
##
```

```
## Mode.HoursHours 34-44                          0.002
##                                               (0.006)
##
## Pub.Density:Mode.HoursHours 34-44            -10.958
##                                              (10.614)
##
## Constant                                     0.008***
##                                               (0.001)
##
## -------------------------------------------------------------
## Observations                                    29
## R2                                             0.546
## Adjusted R2                                    0.491
## Residual Std. Error                       0.003 (df = 25)
## F Statistic                          10.003*** (df = 3; 25)
## =============================================================
## Note:                             *p<0.1; **p<0.05; ***p<0.01
##
## Table 5: ESA density & Pub density with free time
## ============
## Hours Worked
## ------------
```

```
ggplot(model_4, aes(x = Pub.Density, y=ESA.Density, color=Mode.Hours)) +
  geom_point() +
  geom_line(aes(y=.fitted))+
  geom_smooth(method="lm", se=F,
              aes(group=1),
              color="black",
              linetype = "dashed")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



At this point I was certain the relationship would switch directions but no, it's still a positive correlation.

Further, the effects of the moderator are not statistically significant, so we will see if we can make the data better by approximating the number of hours worked in each borough rather than taking the modal average, thus treating the moderator as a continuous variable rather than a categorical one. This is a very rough estimate for the hours worked, however, as the data does not give us much of an insight.

```
(10+34)/2
```

```
## [1] 22
```

```
(34+44)/2
```

```
## [1] 39
```

```
Apx.Hours <- (data_set$`Hours 10`*10 +
                data_set$`Hours 10-34`*22 +
                data_set$`Hours 34-44`*39 +
                data_set$`Hours 45`*45) / (
                  data_set$`Hours 10` +
                    data_set$`Hours 10-34` +
                    data_set$`Hours 34-44` +
                    data_set$`Hours 45`
                )
data_set$Apx.Hours <- Apx.Hours
```

Run a model using the approximate hours worked data rather than the modal average.

```
model_5 <-lm(ESA.Density ~ Pub.Density + Apx.Hours + Pub.Density*Apx.Hours,
             data=data_set)

stargazer(model_5, type = "text",
          title ="Table 6: ESA density & Pub density with continous free time",
          dep.var.labels = "ESA density",
          covariate.labels = "Pubs per person", "Hours Worked")
```

```
##
## Table 6: ESA density & Pub density with continous free time
## ==================================================
##                           Dependent variable:
##                       ---------------------------
##                               ESA density
## --------------------------------------------------
## Pubs per person                 -36.781
##                                (100.863)
##
## Apx.Hours                       -0.001
##                                 (0.001)
##
## Pub.Density:Apx.Hours            1.357
##                                 (2.780)
##
## Constant                         0.057
##                                 (0.051)
##
## --------------------------------------------------
## Observations                      29
## R2                               0.461
## Adjusted R2                      0.396
## Residual Std. Error         0.003 (df = 25)
## F Statistic              7.115*** (df = 3; 25)
## ==================================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
##
## Table 6: ESA density & Pub density with continous free time
## ============
## Hours Worked
## ------------
```

Any new insight is still statistically insignificant, and the correlation remains positive.

Perhaps re-including the income confounding variable might improve the model.

```
model_6 <-lm(ESA.Density ~
               Pub.Density + Apx.Hours + Pub.Density*Apx.Hours + Strat.Income,
             data=data_set)

stargazer(model_6, type = "text",
          title ="Table 7: Full model",
```

```r
          dep.var.labels = "ESA density",
          covariate.labels = "Pubs per person", "Hours Worked", "Income")
```

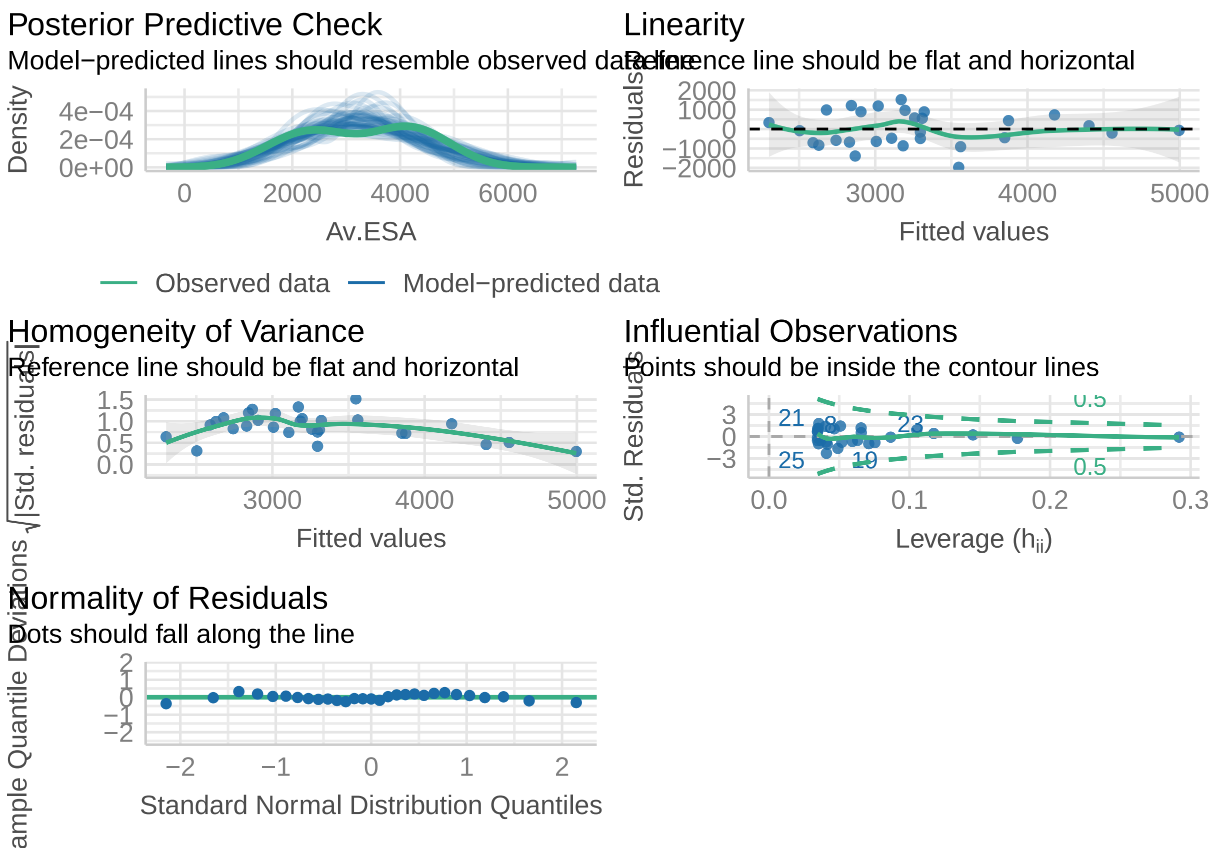
```
##
## Table 7: Full model
## ========================================================
##                                Dependent variable:
##                             ---------------------------
##                                    ESA density
## --------------------------------------------------------
## Pubs per person                      -17.897
##                                     (70.187)
##
## Apx.Hours                           -0.00005
##                                      (0.001)
##
## Strat.Income116.056133056133        -0.004***
##                                      (0.001)
##
## Strat.Income227.657657657658        -0.005***
##                                      (0.001)
##
## Pub.Density:Apx.Hours                 0.748
##                                      (1.935)
##
## Constant                              0.013
##                                      (0.037)
##
## --------------------------------------------------------
## Observations                           29
## R2                                    0.776
## Adjusted R2                           0.727
## Residual Std. Error              0.002 (df = 23)
## F Statistic                  15.945*** (df = 5; 23)
## ========================================================
## Note:                        *p<0.1; **p<0.05; ***p<0.01
##
## Table 7: Full model
## ============
## Hours Worked
## ------------
##
## Table 7: Full model
## ======
## Income
## ------
```

This model is superior to the previous two, but no better than model 3, suggesting that actually the confounding variable is much more important in the relationship than the posited moderator.

5. Model Diagnostics

The quickest way to consider these models' trustworthiness is via a package for Gauss-Markov diagnostics.

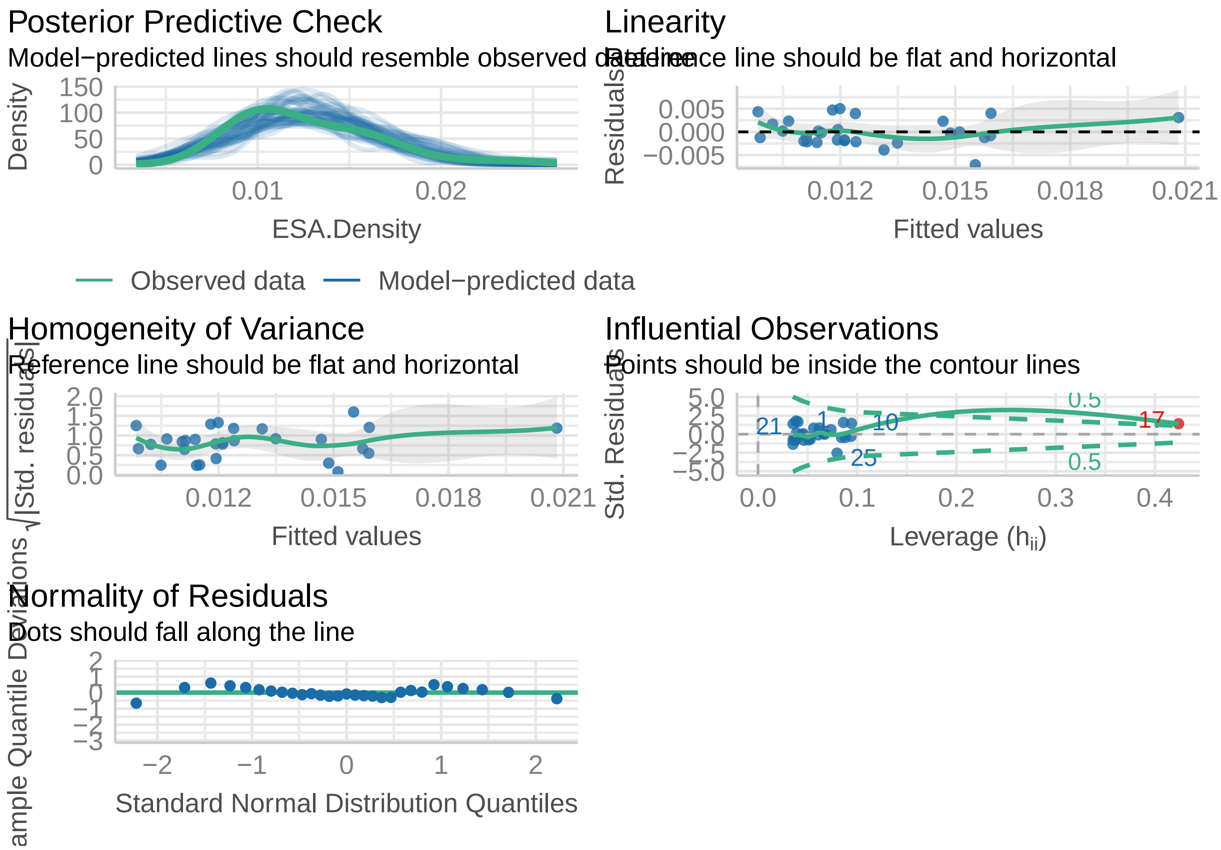
```r
check_model(model_1)
```



Posterior predictive check: alright, has a bit of a dip so not perfect. Linearity: line is fairly flat. Homogeneity of Variance: has a downward curve, not great. Influential Observations: all points are withing contour lines. Normality of residuals: dots are mostly along the line.

So far, data is not bad, so model can be trusted.

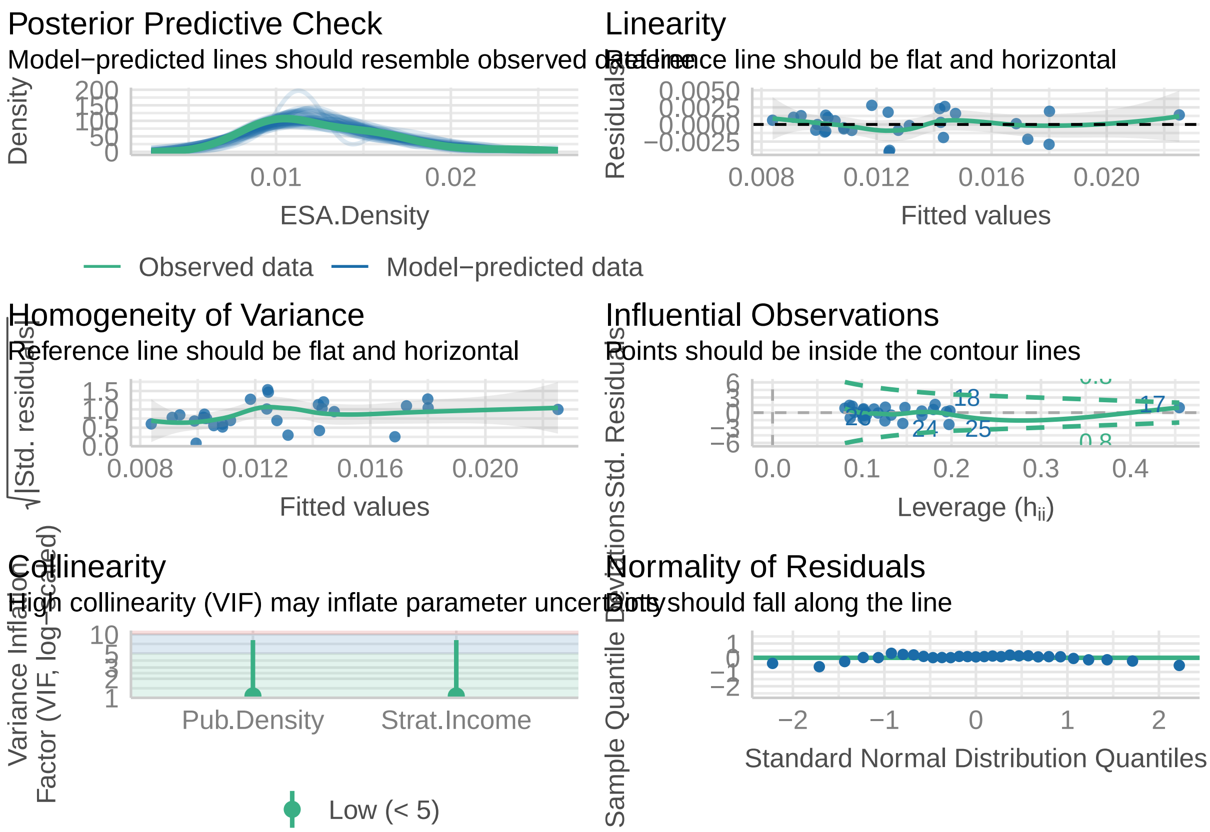
```r
check_model(model_2)
```



Posterior predictive check: alright, slightly skewed to the left. Linearity: bit of an upward curve. Homogeneity of Variance: too wavy to be considered great. Influential Observations: point outside countour lines suggests there is an outlier in this data. Normality of residuals: dots are mostly along the line, a little less than model 1.

Inclusion of second parameter actually makes the data - and the model - worse.

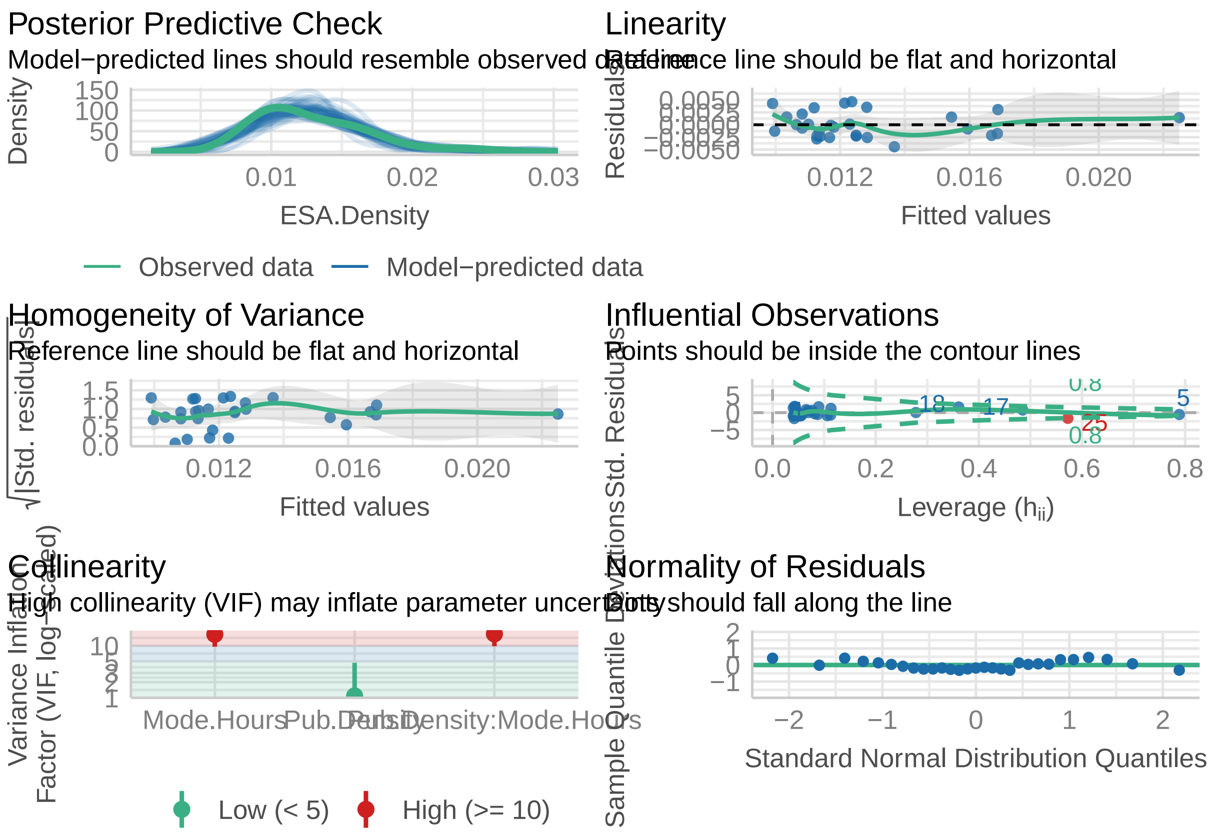
```
check_model(model_3)
```



Posterior predictive check: alright, slightly skewed to the left. Linearity: roughly on the line. Homogeneity of Variance: improvement from model 2. Influential Observations: all points are within contour lines again. Collinearity: low. Normality of residuals: dots are mostly along the line.

Adding the second confounding variable improves the model again.

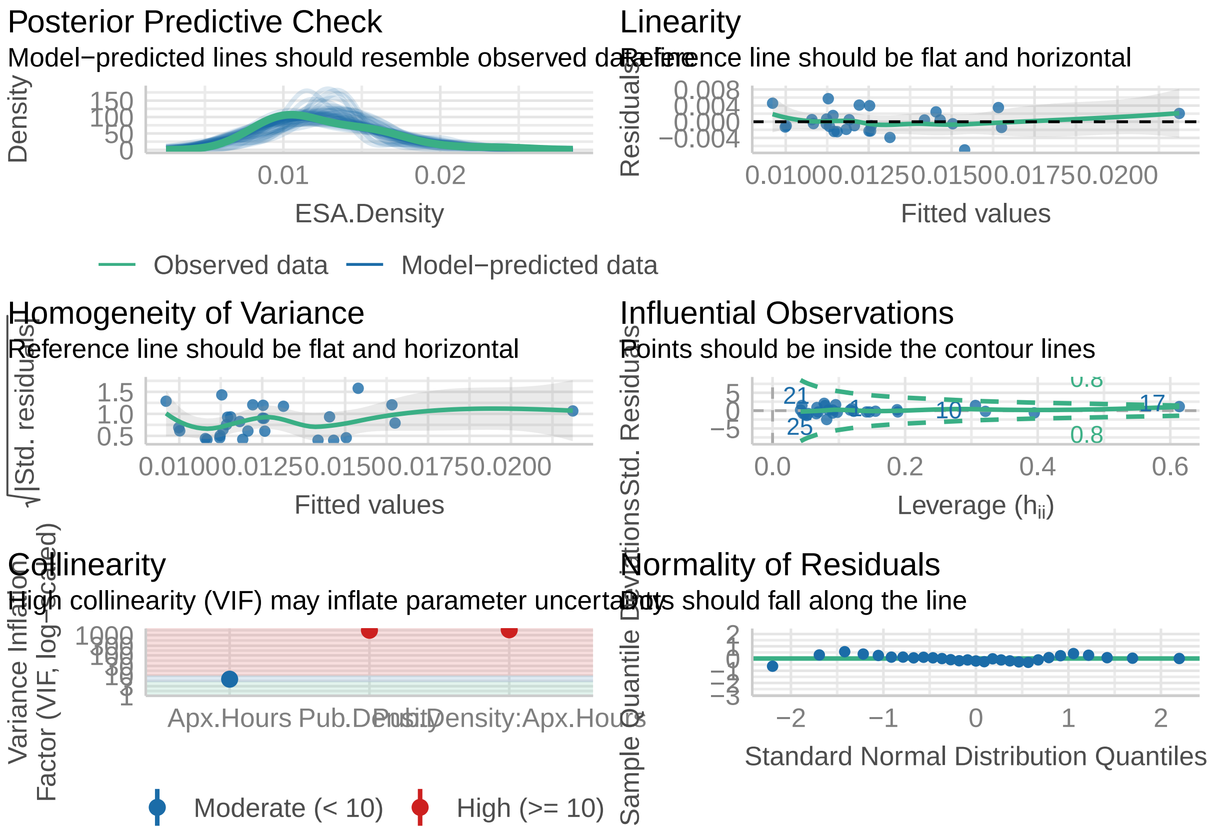
```
check_model(model_4)
```



Posterior predictive check: alright, slightly skewed to the left. Linearity: more wavy. Homogeneity of Variance: more wavy. Influential Observations: there is a point outside the contour lines, acting as an outlier. Collinearity: high Normality of residuals: dots are mostly along the line.

Adding the moderator makes the model worse.

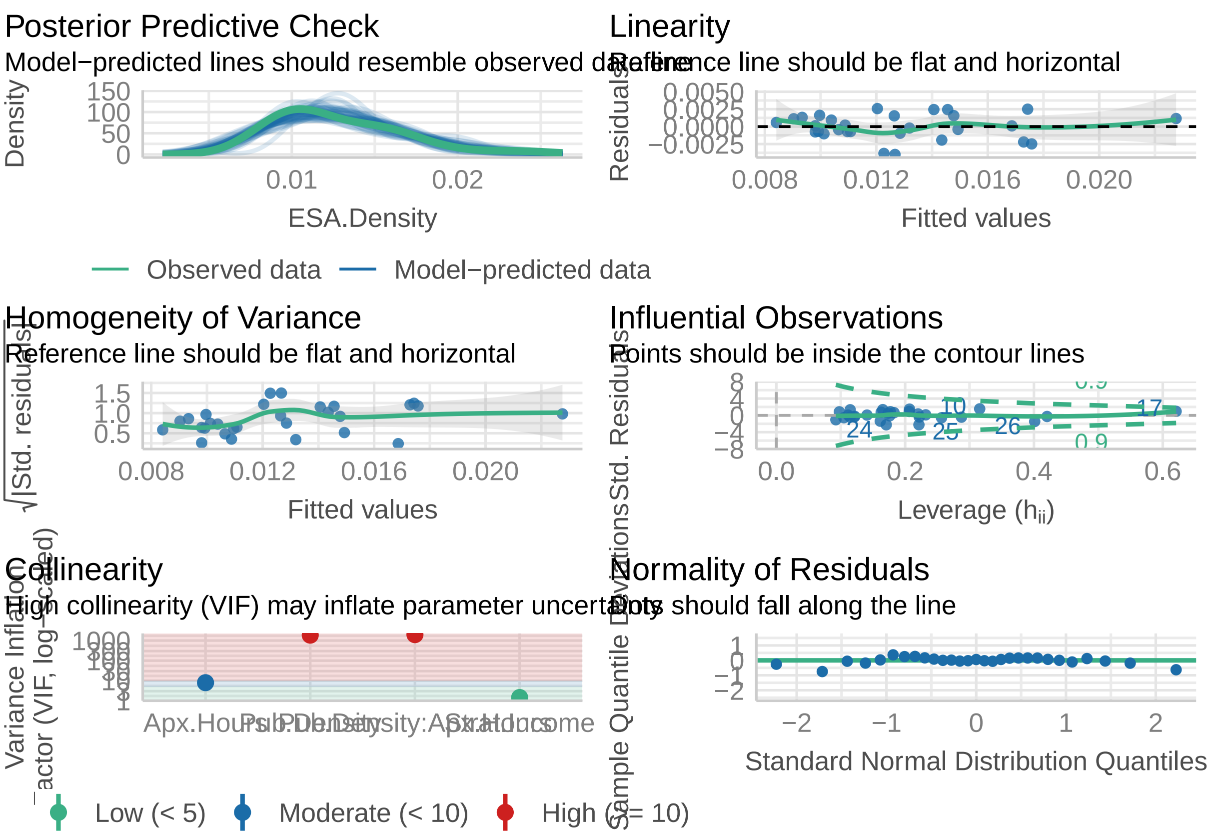
```
check_model(model_5)
```



Posterior predictive check: alright, slightly skewed to the left. Linearity: less wavy. Homogeneity of Variance: more wavy. Influential Observations: no points outside the contour lines. Collinearity: high Normality of residuals: dots are mostly along the line.

Using approximate hours rather than modal hours slightly improves the model.

```
check_model(model_6)
```



Posterior predictive check: alright, slightly skewed to the left. Linearity: about as wavy as model 3. Homogeneity of Variance: more wavy than model 3. Influential Observations: no points outside the contour lines. Collinearity: high Normality of residuals: dots are mostly along the line.

Combining confounding variable with moderator improves model 5, but doesn't drastically change model 3.

```
compare_performance(model_1, model_2, model_3, model_4, model_5, model_6)
```

```
## When comparing models, please note that probably not all models were fit
##   from same data.
```

```
## # Comparison of Model Performance Indices
##
## Name    | Model |  AIC (weights) | AICc (weights) |  BIC (weights) |    R2 | R2 (adj.) |    RMSE |
## ------------------------------------------------------------------------------------------------------
## model_1 |    lm |  479.2 (<.001) |  480.1 (<.001) |  483.3 (<.001) | 0.357 |     0.333 | 844.637 | 8
## model_2 |    lm | -252.9 (<.001) | -251.9 (<.001) | -248.8 (<.001) | 0.424 |     0.403 |   0.003 |
## model_3 |    lm | -275.5 (0.838) | -272.9 (0.953) | -268.7 (0.953) | 0.771 |     0.743 |   0.002 |
## model_4 |    lm | -255.7 (<.001) | -253.1 (<.001) | -248.9 (<.001) | 0.546 |     0.491 |   0.002 |
## model_5 |    lm | -250.7 (<.001) | -248.1 (<.001) | -243.9 (<.001) | 0.461 |     0.396 |   0.003 |
## model_6 |    lm | -272.2 (0.162) | -266.9 (0.047) | -262.7 (0.047) | 0.776 |     0.727 |   0.002 |
```

Models 3 and 6 are similar in AIC, BIC and R2 (adj.), but model 3 is slightly better in these areas, and has no statistically insignificant elements. Therefore we take model 3 to be our most authoritative model.

6. Conclusion

Using Model 3, we can conclude that, after adjusting for population and income, there is sufficient evidence to reject the null hypothesis H0, that there is no correlation between the number of pubs and the number of mental health cases, specifically within London boroughs. We therefore accept the alternative hypothesis H1, that the number of pubs and mental health cases are positively correlated. We cannot, however, conclude anything about the causality of these two variables, as the theory suggested that they should be negatively correlated, so there are clearly factors that have not been accounted for in this situation, and further research would be required to unearth them.