109006240A2

Jansen Reynaldi Gautama

2023-11-23

Package Loading

```r
require(lubridate)
```

```
## Loading required package: lubridate
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
require(gapminder)
```

```
## Loading required package: gapminder
```

```r
require(readr)
```

```
## Loading required package: readr
```

```r
require(knitr)
```

```
## Loading required package: knitr
```

```r
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v stringr 1.5.0
## v ggplot2 3.4.4     v tibble  3.2.1
## v purrr   1.0.2     v tidyr   1.3.0
## -- 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
```

```
require(tidyr)
```

Part 1

Question 1

red csv

```
air99 <- read_csv('pm99.csv')
```

```
## Rows: 117421 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): X..RD, Action.Code, State.Code, County.Code, Site.ID
## dbl (7): Parameter, POC, Sample.Duration, Unit, Method, Date, Sample.Value
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
air12 <- read_csv('pm12.csv')
```

```
## Rows: 1304287 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): X..RD, Action.Code, State.Code, County.Code, Site.ID
## dbl (7): Parameter, POC, Sample.Duration, Unit, Method, Date, Sample.Value
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

do 3 task asked in pipe sequence then combine and drop row with na

```
air99 <- air99 %>%
  mutate(Year = 1999, PM2.5 = Sample.Value) %>%
  select(-Sample.Value)

air12 <- air12 %>%
  mutate(Year = 2012, PM2.5 = Sample.Value) %>%
  select(-Sample.Value)

air_combined <- bind_rows(air99, air12) %>%
  drop_na()
```

now we can check the dataframe using glimpse()

```
glimpse(air_combined)
```

```
## Rows: 1,335,358
## Columns: 13
## $ X..RD           <chr> "RD", "RD", "RD", "RD", "RD", "RD", "RD", "RD", "RD", ~
## $ Action.Code     <chr> "I", "I", "I", "I", "I", "I", "I", "I", "I", "I", "I",~
## $ State.Code      <chr> "01", "01", "01", "01", "01", "01", "01", "01", "01", ~
## $ County.Code     <chr> "027", "027", "027", "027", "027", "027", "027", "027"~
## $ Site.ID         <chr> "0001", "0001", "0001", "0001", "0001", "0001", "0001"~
## $ Parameter       <dbl> 88101, 88101, 88101, 88101, 88101, 88101, 88101, 88101~
## $ POC             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ Sample.Duration <dbl> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ~
## $ Unit            <dbl> 105, 105, 105, 105, 105, 105, 105, 105, 105, 105, 105,~
## $ Method          <dbl> 120, 120, 120, 120, 120, 120, 120, 120, 120, 120, 120,~
## $ Date            <dbl> 19990112, 19990115, 19990118, 19990121, 19990124, 1999~
## $ Year            <dbl> 1999, 1999, 1999, 1999, 1999, 1999, 1999, 1999, 1999, ~
## $ PM2.5           <dbl> 8.841, 14.920, 3.878, 9.042, 5.464, 20.170, 11.560, 13~
```

to make sure that the NA dropped i also check using is.na()

```
any_na <- any(is.na(air99))
any_na
```

```
## [1] TRUE
```

```
any_na <- any(is.na(air12))
any_na
```

```
## [1] TRUE
```

```
any_na <- any(is.na(air_combined))
any_na
```

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

Question 2

group by and summarize the statistic

```
pm_summary <- air_combined %>%
  group_by(Year) %>%
  summarize(
    mean_pm25 = mean(PM2.5),
    median_pm25 = median(PM2.5),
    min_pm25 = min(PM2.5),
    max_pm25 = max(PM2.5)
  )
```

filter the PM2.5>0

```
air_combined <- air_combined %>%
  filter(PM2.5 > 0)
```

make sure the data is good

```
pm_summary
```

```
## # A tibble: 2 x 5
##    Year mean_pm25 median_pm25 min_pm25 max_pm25
##   <dbl>     <dbl>       <dbl>    <dbl>    <dbl>
## 1  1999     13.7        11.5         0     157.
## 2  2012      9.14        7.63      -10     909.
```

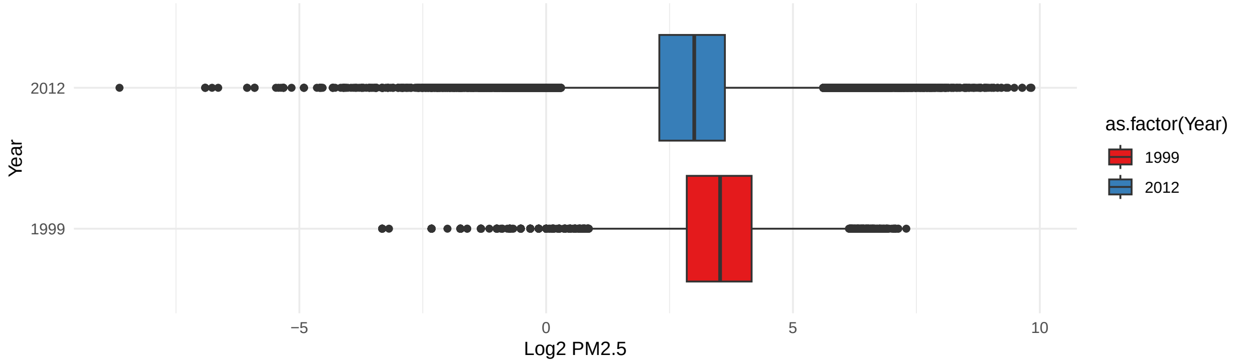
```
head(air_combined)
```

```
## # A tibble: 6 x 13
##   X..RD Action.Code State.Code County.Code Site.ID Parameter   POC
##   <chr> <chr>       <chr>      <chr>       <chr>       <dbl> <dbl>
## 1 RD    I           01         027         0001        88101     1
## 2 RD    I           01         027         0001        88101     1
## 3 RD    I           01         027         0001        88101     1
## 4 RD    I           01         027         0001        88101     1
## 5 RD    I           01         027         0001        88101     1
## 6 RD    I           01         027         0001        88101     1
## # i 6 more variables: Sample.Duration <dbl>, Unit <dbl>, Method <dbl>,
## #   Date <dbl>, Year <dbl>, PM2.5 <dbl>
```

Question 3

do all the things mentioned and make the box plot

```
ggplot(air_combined, aes(x = as.factor(Year), y = log2(PM2.5), fill = as.factor(Year))) +
  geom_boxplot() +
  labs(x = "Year", y = "Log2 PM2.5") +
  scale_fill_brewer(palette = "Set1") +
  theme_minimal() +
  coord_flip()
```



as we can see that in 2012 the distribution are spread widely and there are time where the polution is really bad and where the polution is lower, in 1999 the max polution is lower and have smaller spread than 2012, but overall the polution in 2012 is lower and we can see it by the median of the box plot where its smaller than 1999 and we can also verify it by looking at the pm_summary.

and the reason this happened as mentioned in the question the 2012 data are more concentrated in cleaner areas and that explained why the 2012 is lower than in 1999

Question 4

create subset to get the NYC data

```
ny_data <- subset(air_combined, State.Code == 36)
```

use paste0 to create site.code

```
ny_data$site.code <- paste0(ny_data$County.Code, ".", ny_data$Site.ID)
```

```
head(ny_data)
```

```
## # A tibble: 6 x 14
##   X..RD Action.Code State.Code County.Code Site.ID Parameter   POC
##   <chr> <chr>       <chr>      <chr>       <chr>       <dbl> <dbl>
## 1 RD    I           36         001         0005        88101     1
## 2 RD    I           36         001         0005        88101     1
## 3 RD    I           36         001         0005        88101     1
## 4 RD    I           36         001         0005        88101     1
## 5 RD    I           36         001         0005        88101     1
## 6 RD    I           36         001         0005        88101     1
## # i 7 more variables: Sample.Duration <dbl>, Unit <dbl>, Method <dbl>,
## #   Date <dbl>, Year <dbl>, PM2.5 <dbl>, site.code <chr>
```

Question 5

Group by site.code then use n_distinct in filter to make monitor

```
ny_data_group <- ny_data %>%
  group_by(site.code) %>%
  filter(n_distinct(Year)==2)
```

```
active_both_year<- pull(ny_data_group, site.code)
```

now we can use unique to check how many site.code existed in 1999 and 2012 and we can see there are 10 out of 41 site.code that existed booth in 2012 and 1999

```
unique(active_both_year)
```

```
##  [1] "001.0005" "001.0012" "005.0080" "013.0011" "029.0005" "031.0003"
##  [7] "063.2008" "067.1015" "085.0055" "101.0003"
```

```
unique(ny_data$site.code)
```

```
##  [1] "001.0005" "001.0012" "005.0073" "005.0080" "005.0083" "005.0110"
##  [7] "013.0011" "027.1004" "029.0002" "029.0005" "029.1007" "031.0003"
## [13] "047.0011" "047.0076" "055.6001" "059.0005" "059.0008" "059.0011"
## [19] "061.0010" "061.0056" "061.0062" "063.2008" "065.2001" "067.0019"
## [25] "067.1015" "081.0094" "081.0097" "085.0055" "085.0067" "089.3001"
## [31] "093.0003" "101.0003" "103.0001" "005.0133" "047.0122" "055.1007"
## [37] "061.0079" "061.0134" "071.0002" "081.0124" "103.0002"
```

Question 6

filter the ny data extracting active_both_year

```
filtered_nyc_active_both_year <- ny_data %>%
  filter(site.code %in% active_both_year)
```

count the site.code and sort it in decending order

```
most_active <- filtered_nyc_active_both_year %>%
  count(site.code, sort = TRUE)
```

```
most_active
```

```
## # A tibble: 10 x 2
##    site.code     n
##    <chr>     <int>
##  1 101.0003    138
##  2 063.2008    117
##  3 031.0003    116
##  4 001.0005    114
##  5 067.1015    110
##  6 029.0005     79
##  7 013.0011     67
##  8 005.0080     64
##  9 001.0012     45
## 10 085.0055     30
```

Question 7

filter the ny_data acording to site.code and get the 101.0003 only, then mutate the date and get the date of year

```
air101.0003 <- ny_data %>%
  filter(site.code == "101.0003") %>%
  mutate(Date = ymd(Date), dayofyear = yday(Date))
```

```
air101.0003
```

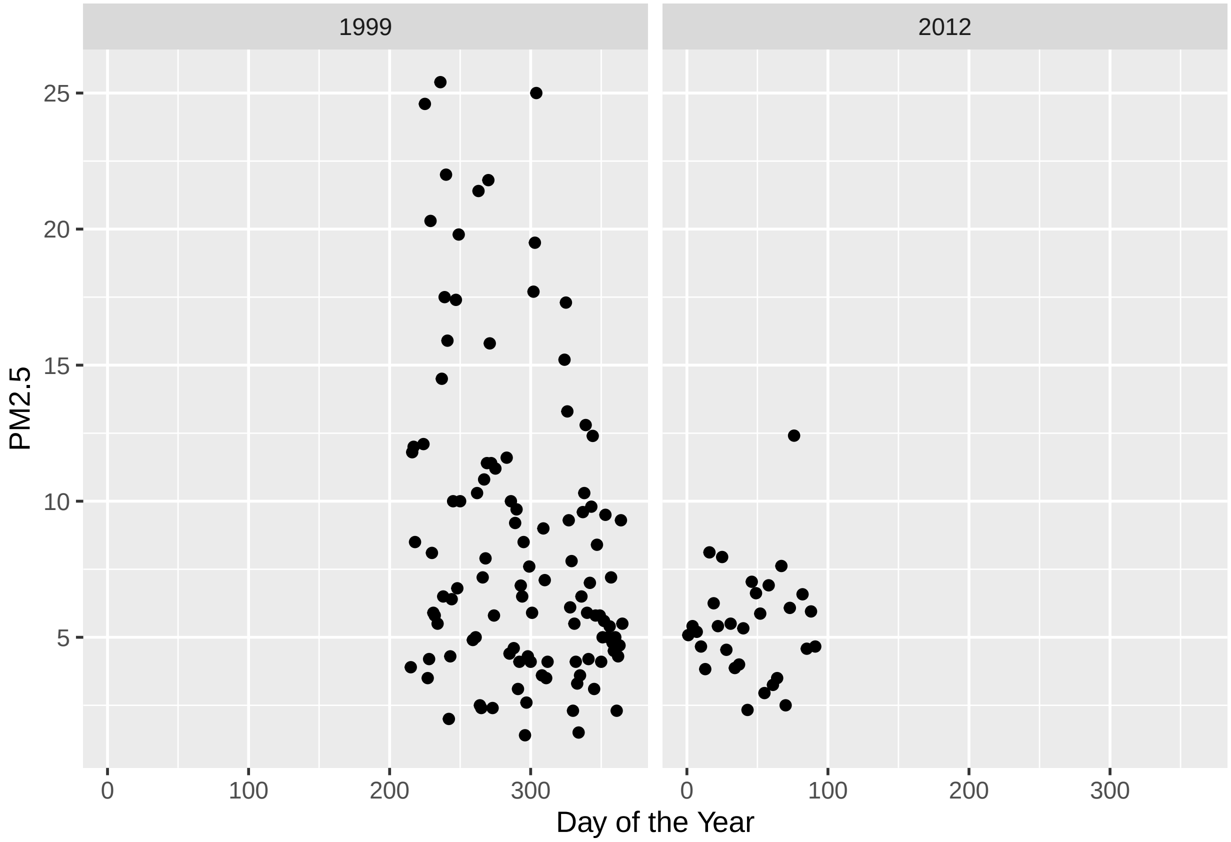
```
## # A tibble: 138 x 15
##    X..RD Action.Code State.Code County.Code Site.ID Parameter   POC
##    <chr> <chr>       <chr>      <chr>       <chr>       <dbl> <dbl>
##  1 RD    I           36         101         0003        88101     1
##  2 RD    I           36         101         0003        88101     1
##  3 RD    I           36         101         0003        88101     1
##  4 RD    I           36         101         0003        88101     1
##  5 RD    I           36         101         0003        88101     1
##  6 RD    I           36         101         0003        88101     1
##  7 RD    I           36         101         0003        88101     1
##  8 RD    I           36         101         0003        88101     1
##  9 RD    I           36         101         0003        88101     1
## 10 RD    I           36         101         0003        88101     1
## # i 128 more rows
## # i 8 more variables: Sample.Duration <dbl>, Unit <dbl>, Method <dbl>,
## #   Date <date>, Year <dbl>, PM2.5 <dbl>, site.code <chr>, dayofyear <dbl>
```

Question 8

we can see that acording this data the air quality in first 3 month of 2012 is better than last 5 month of 1999 where in 1999 it reach max above 25 and in 2012 it almost reach 12.5.

just by seeing the plot we can also know that the average PM2.5 is lower in 2012 than in 1999 which means quality improves and acording to google PM2.5 is saves under 12.5 which happened in 2012 and it means at some time air quality in some part of NY is really bad and reach a dagerous level

```
ggplot(air101.0003, aes(x = dayofyear, y = PM2.5)) +
  geom_point() +
  labs(x = "Day of the Year", y = "PM2.5") +
  facet_wrap(~Year, ncol = 2)
```



Part 2

Question 9

in this part we can see how the dataset is distributed from the average so for example we can see the age average is 24, most of the participant is a HS dropout, etc.

first read the csv

```
lalonde <- read_csv('lalonde.csv')
```

```
## Rows: 722 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): data_id
## dbl (9): treat, age, education, black, hispanic, married, nodegree, re75, re78
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
lalonde
```

```
## # A tibble: 722 x 10
##    data_id    treat   age education black hispanic married nodegree  re75   re78
##    <chr>      <dbl> <dbl>     <dbl> <dbl>    <dbl>   <dbl>    <dbl> <dbl>  <dbl>
##  1 Lalonde S~     1    37        11     1        0       1        1     0  9930.
##  2 Lalonde S~     1    22         9     0        1       0        1     0  3596.
##  3 Lalonde S~     1    30        12     1        0       0        0     0 24909.
##  4 Lalonde S~     1    27        11     1        0       0        1     0  7506.
##  5 Lalonde S~     1    33         8     1        0       0        1     0   290.
##  6 Lalonde S~     1    22         9     1        0       0        1     0  4056.
##  7 Lalonde S~     1    23        12     1        0       0        0     0     0
##  8 Lalonde S~     1    32        11     1        0       0        1     0  8472.
##  9 Lalonde S~     1    22        16     1        0       0        0     0  2164.
## 10 Lalonde S~     1    33        12     0        0       1        0     0 12418.
## # i 712 more rows
```

we can create the balancetable by grouping by treat then use summarize and get the mean/avg of all the variables then we can use kable to get a table

```
balance_table <- lalonde %>%
  group_by(treat) %>%
  summarize(
    age_avg = mean(age),
    education_avg = mean(education),
    black_avg = mean(black),
    hispanic_avg = mean(hispanic),
    married_avg = mean(married),
    nodegree_avg = mean(nodegree)
  )
knitr::kable(balance_table, caption ="Covariate Balance",
             col.names = c("Group", "Age", "Education", "Black", "Hispanic", "Married", "No Degree"))
```

Table 1: Covariate Balance

Group	Age	Education	Black	Hispanic	Married	No Degree
0	24.44706	10.18824	0.8000000	0.1129412	0.1576471	0.8141176
1	24.62626	10.38047	0.8013468	0.0942761	0.1683502	0.7306397

Question 10

we can see below that people that get treatment have a higher change when they get treatment and the difference are pretty high as we can see that people who get treatment they improve between 40-50% more than people that didnt get the treatment

first create the change

```
lalonde <- lalonde %>%
  mutate(change = re78 - re75)
```

then get the avg of changes of treated group and ctr group and we can see that treated group has higher change avg

```
trt_change <- lalonde %>%
  filter(treat == 1) %>%
  summarize(avg_change = mean(change))
trt_change
```

```
## # A tibble: 1 x 1
##   avg_change
##        <dbl>
## 1      2910.
```

```
ctr_change <- lalonde %>%
  filter(treat == 0) %>%
  summarize(avg_change = mean(change))
ctr_change
```

```
## # A tibble: 1 x 1
##   avg_change
##        <dbl>
## 1      2063.
```

and we can see ate which is the avg treatment effect

```
ate <- trt_change - ctr_change
ate
```

```
##   avg_change
## 1   846.8883
```

Question 11

In this experiment there are 2 group, 1 group where they get a special treatment/course and the other don't get the treatment which is the control group. and this experiment give us insight of the effect of the treatment to a bunch of people and we can see that the group who get the treatment their earning grows higher than the control group. and i think there are some bias in the data since in my opinion we can't only see whether they get treatment or not because in my opinion there are more factors such as people characteristic and other uncontrolable variables

Question 12

in my opinion we should use the difference between the begining(re75) and the end(re78) because we want to see the growth not what they earn before of after, because re75 depends on the person it self and re78 also depends on re75.

Question 13

we can see that people that finished highschool have higher ATE which means that people that finished HS get higher impact from the treatment than people who dropped out from HS

for this part i just follow the direction.

```r
ate_dropout <- lalonde %>%
  mutate(
    dropout = ifelse(nodegree == 1, "Dropped out", "Finished HS"),
    treatment_group = ifelse(treat == 1, "Treated", "Control")
  ) %>%
  group_by(dropout, treatment_group) %>%
  summarize(mean_change = mean(change)) %>%
  pivot_wider(names_from = treatment_group, values_from = mean_change) %>%
  mutate(ATE = Treated - Control) %>%
  select(dropout, Treated, Control, ATE)
```

```
## 'summarise()' has grouped output by 'dropout'. You can override using the
## '.groups' argument.
```

```r
ate_dropout
```

```
## # A tibble: 2 x 4
## # Groups:   dropout [2]
##   dropout     Treated Control   ATE
##   <chr>         <dbl>   <dbl> <dbl>
## 1 Dropped out   2623.   2173.  450.
## 2 Finished HS   3689.   1584. 2105.
```

```r
knitr::kable(ate_dropout, caption ="ATE by DO Status", col.names = c("Dropout Status", "Mean Change (Tr
```

Table 2: ATE by DO Status

Dropout Status	Mean Change (Treated)	Mean Change (Control)	ATE
Dropped out	2623.151	2172.871	450.2804
Finished HS	3689.019	1583.760	2105.2599

Question 14

as we can see below that the treatment give more impact to people between age 31-40

first we use case when to group the data according to the age range

```r
lalonde <- lalonde %>%
  mutate(
    age_group = case_when(
      age <= 30 ~ "30 and under",
      age > 30 & age <= 40 ~ "31 - 40",
      age > 40 ~ "Over 40"
    )
  )
```

then we treat the data just like the previous data

```
ate_age <- lalonde %>%
  mutate(
    treatment_group = ifelse(treat == 1, "Treated", "Control")
  ) %>%
  group_by(age_group, treatment_group) %>%
  summarize(mean_change = mean(change, na.rm = TRUE)) %>%
  pivot_wider(names_from = treatment_group, values_from = mean_change) %>%
  mutate(ATE = Treated - Control) %>%
  select(age_group, Treated, Control, ATE)
```

```
## ‘summarise()‘ has grouped output by ’age_group’. You can override using the
## ‘.groups‘ argument.
```

```
ate_age
```

```
## # A tibble: 3 x 4
## # Groups:   age_group [3]
##   age_group    Treated Control   ATE
##   <chr>          <dbl>   <dbl> <dbl>
## 1 30 and under   2736.   2137.  600.
## 2 31 - 40        3951.   1193. 2758.
## 3 Over 40        3915.   3786.  128.
```

```
ate_age
```

```
## # A tibble: 3 x 4
## # Groups:   age_group [3]
##   age_group    Treated Control   ATE
##   <chr>          <dbl>   <dbl> <dbl>
## 1 30 and under   2736.   2137.  600.
## 2 31 - 40        3951.   1193. 2758.
## 3 Over 40        3915.   3786.  128.
```

```
knitr::kable(ate_age, caption ="ATE by Age Group", col.names = c("Age Group", "Mean Change (Treated)",
```

Table 3: ATE by Age Group

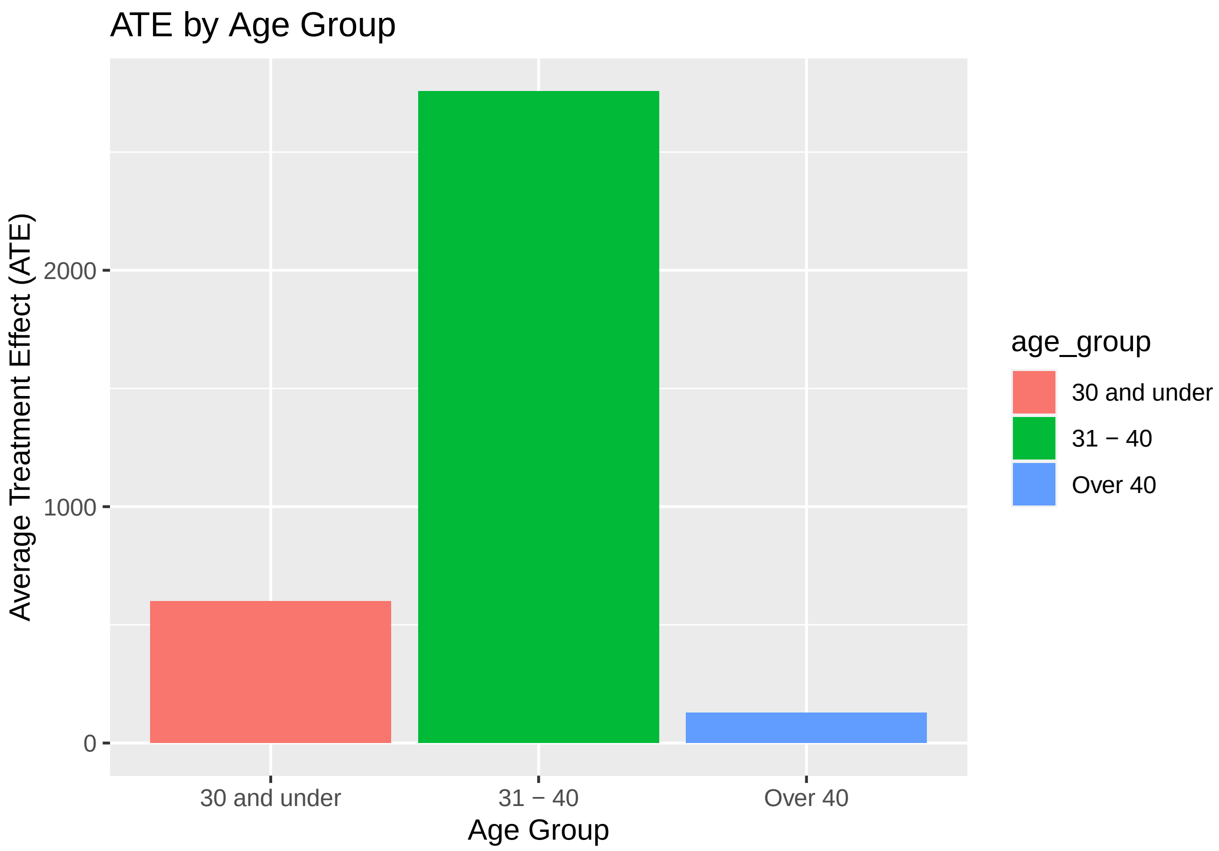
Age Group	Mean Change (Treated)	Mean Change (Control)	ATE
30 and under	2736.367	2136.730	599.6373
31 - 40	3950.871	1192.721	2758.1497
Over 40	3914.669	3786.459	128.2100

Question 15

as we can see that the ATE improve way better on people on age 31-40 so it means it have better impact on people between age of 31-40

```r
age_plot <- ggplot(ate_age, aes(x = age_group, y = ATE, fill = age_group)) +
  geom_bar(stat = "identity") +
  labs(title = "ATE by Age Group",
       x = "Age Group",
       y = "Average Treatment Effect (ATE)")
```

```r
age_plot
```


ATE by Age Group
Average Treatment Effect (ATE)
30 and under
Over 40
Age Group
age_group
30 and under
Over 40