

CDS-101 Checkpoint #2

Eunho Cha, Daeun Choi, Songlee Jun, Hyun Woo Kang
Dawon Kyoung, Byungwook Oh, Duy Tran

2024-06-03

#GHG = Greenhouse Gases #FJO = Female Job Occupation #YOY = Year-on-Year
#Outlier function

```
find_outlier <- function(x) {  
  return(x < quantile(x, .25) - 1.5*IQR(x) | x > quantile(x, .75) + 1.5*IQR(x))  
}
```

Organizing Dataset (Hyunwoo Kang)

Summarize

```
summary(Group_dataset)
```

```
##      Year      Birth_per_1000      Birth_YOY      FJO  
## Min.      :1991      Min.      : 7.036      Min.      : -5.870      Min.      : 7529000  
## 1st Qu.:1998      1st Qu.: 8.795      1st Qu.: -3.555      1st Qu.: 8616500  
## Median :2006      Median : 9.786      Median : -2.450      Median : 9707000  
## Mean      :2006      Mean      :10.835      Mean      : -2.499      Mean      : 9649613  
## 3rd Qu.:2014      3rd Qu.:13.431      3rd Qu.: -1.305      3rd Qu.:10697000  
## Max.      :2021      Max.      :16.002      Max.      : 1.080      Max.      :11725000  
##      FJO_YOY      GHG      GHG_YOY      INF_Rate  
## Min.      : -7.000      Min.      :281.4      Min.      : -17.080      Min.      :0.383  
## 1st Qu.: 1.000      1st Qu.:438.5      1st Qu.: 1.065      1st Qu.:1.710  
## Median : 2.000      Median :509.5      Median : 2.540      Median :2.756  
## Mean      : 1.581      Mean      :519.4      Mean      : 3.181      Mean      :3.232  
## 3rd Qu.: 3.000      3rd Qu.:638.3      3rd Qu.: 6.940      3rd Qu.:4.460  
## Max.      : 5.000      Max.      :684.7      Max.      :11.860      Max.      :9.333  
## Inflation_YOY  
## Min.      : -6.7000  
## 1st Qu.: -0.9900  
## Median : 0.0800  
## Mean      : -0.1968
```

```
## 3rd Qu.: 0.8650
## Max.    : 3.0700
```

Select

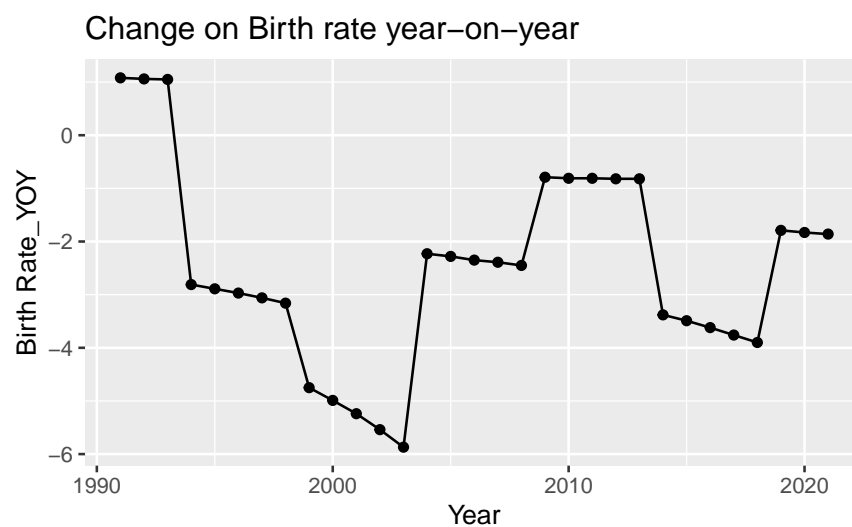
```
Envo_ft <- Group_dataset %>%
  select(Year, Birth_per_1000, Birth_YOY, GHG, GHG_YOY)
```

```
Econ_ft <- Group_dataset %>%
  select(Year, Birth_per_1000, Birth_YOY, INF_Rate, Inflation_YOY)
```

```
Soci_ft <- Group_dataset %>%
  select(Year, Birth_per_1000, Birth_YOY, FJO, FJO_YOY)
```

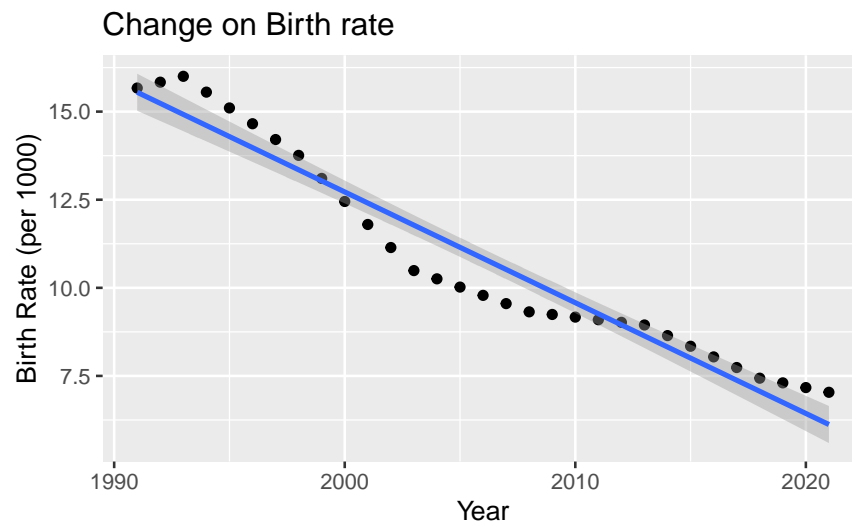
Variation of Birth Rate change Year-on-Year (Hyunwoo Kang)

```
Group_dataset %>%
  ggplot() +
  geom_line(mapping = aes(y = Birth_YOY, x= Year)) +
  geom_point(mapping = aes(y = Birth_YOY, x= Year))+
  labs(title = "Change on Birth rate year-on-year",
       y = "Birth Rate_YOY",
       x = "Year")
```



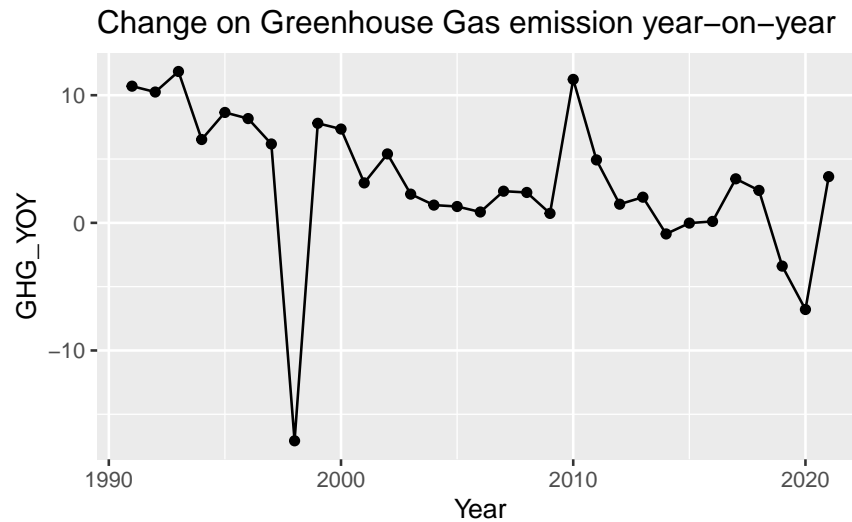
```
Group_dataset %>%
  ggplot() +
  geom_point(mapping = aes(y = Birth_per_1000, x= Year))+
  geom_smooth(mapping = aes(y = Birth_per_1000, x = Year), method = "lm") +
  labs(title = "Change on Birth rate",
       y = "Birth Rate (per 1000)",
       x = "Year")
```

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



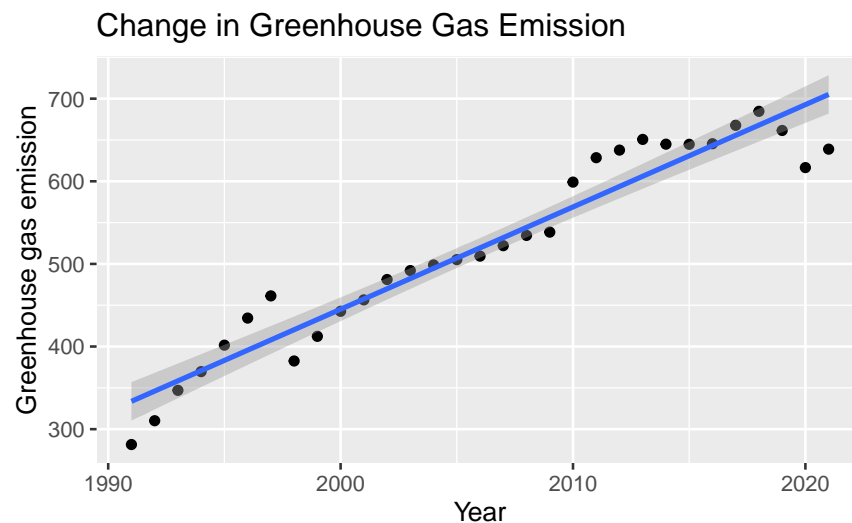
Variation and Covariation - Envo_ft (Songlee Jun)

```
Group_dataset %>%
  ggplot() +
  geom_line(mapping = aes(y = GHG_YOY, x= Year)) +
  geom_point(mapping = aes(y = GHG_YOY, x= Year))+
  labs(title = "Change on Greenhouse Gas emission year-on-year",
       y = "GHG_YOY",
       x = "Year")
```



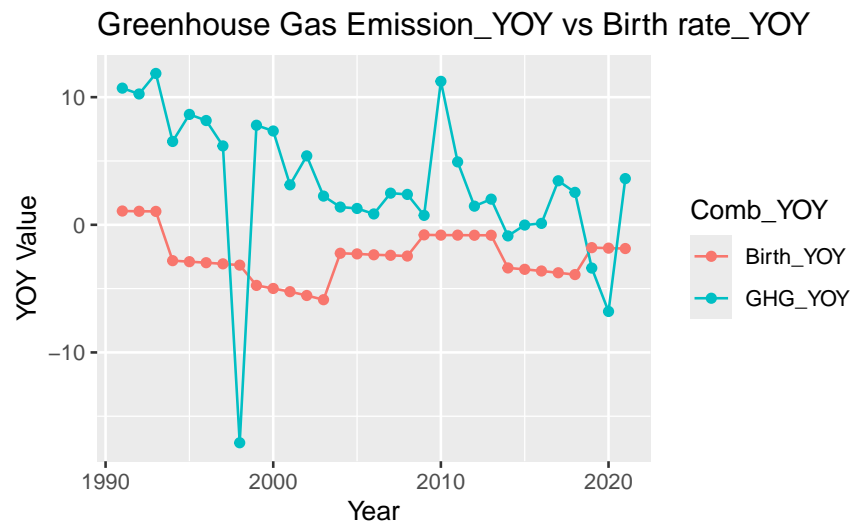
```
Envo_ft %>%
  ggplot() +
  geom_point(mapping = aes(x = Year, y = GHG)) +
  geom_smooth(mapping = aes(x = Year, y = GHG), method = "lm") +
  labs(title = "Change in Greenhouse Gas Emission",
       x = "Year",
       y = "Greenhouse gas emission")
```

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



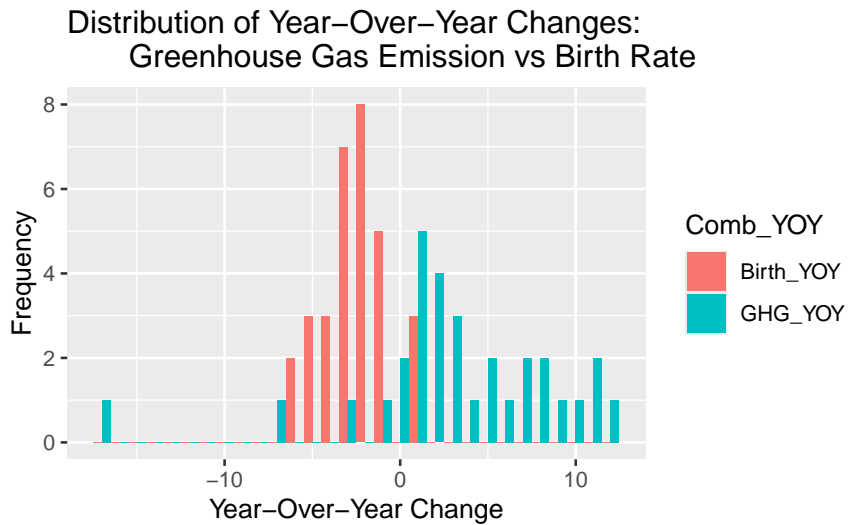
```
Group_dataset%>%
  pivot_longer(cols = c('Birth_YOY', 'GHG_YOY'),
               names_to = 'Comb_YOY',
               values_to = 'Val_YOY')%>%
```

```
ggplot(aes(y = Val_YOY, x = Year)) +
  geom_line(aes(color = Comb_YOY)) +
  geom_point(aes(color = Comb_YOY)) +
  labs(title = "Greenhouse Gas Emission_YOY vs Birth rate_YOY",
       x = "Year",
       y = "YOY Value")
```



```
Group_dataset %>%
  pivot_longer(cols = c('Birth_YOY', 'GHG_YOY'),
               names_to = 'Comb_YOY',
               values_to = 'Val_YOY') %>%

ggplot(aes(x = Val_YOY, fill = Comb_YOY)) +
  geom_histogram(position = 'dodge', bins = 30) +
  labs(title = "Distribution of Year-Over-Year Changes:
             Greenhouse Gas Emission vs Birth Rate",
       x = "Year-Over-Year Change",
       y = "Frequency")
```

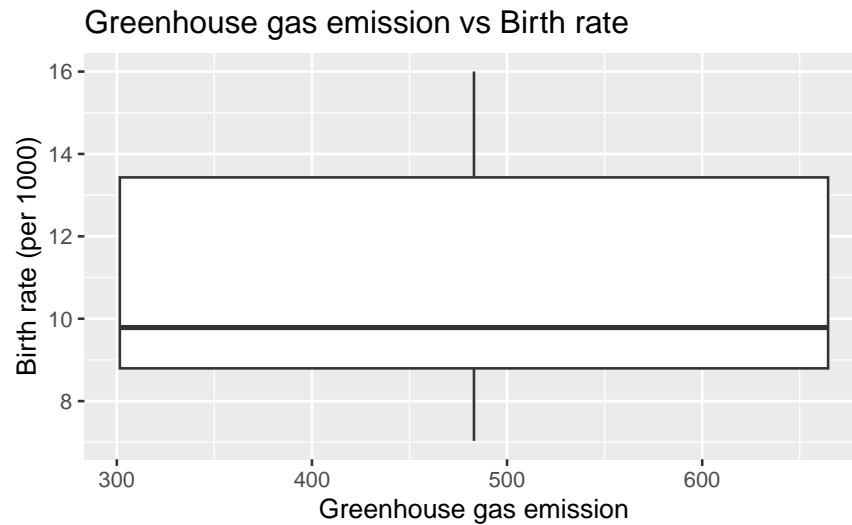


```
Envo_ft %>%
  ggplot(mapping = aes(x = GHG, y = Birth_per_1000)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16) +
  stat_summary(
    aes(label = round(..y.., 1)),
    geom = "text",
    fun.data = function(y) {
      out <- boxplot.stats(y)$out
      if(length(out) == 0) return(NULL)
      data.frame(y = out)
    },
    hjust = -0.5
  ) +
  labs(
    title = "Greenhouse gas emission vs Birth rate",
    x = 'Greenhouse gas emission',
    y = "Birth rate (per 1000)"
  )
)
```

```
## Warning: The dot-dot notation ('..y..') was deprecated in ggplot2 3.4.0.
## i Please use 'after_stat(y)' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

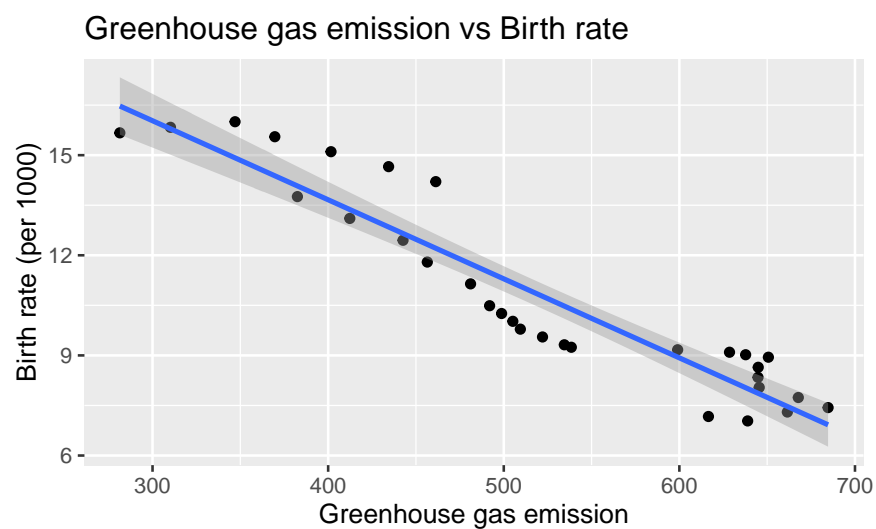
```
## Warning: Continuous x aesthetic
## i did you forget 'aes(group = ...)'?
```

```
## Warning: Computation failed in 'stat_summary()'.
## Caused by error in 'fix.by()':
## ! 'by' must specify uniquely valid columns
```



```
Envo_ft %>%
  ggplot()+
  geom_point(mapping = aes(x = GHG, y = Birth_per_1000)) +
  geom_smooth(mapping = aes(x = GHG, y = Birth_per_1000), method="lm")+
  labs(
    title = "Greenhouse gas emission vs Birth rate",
    x = 'Greenhouse gas emission',
    y = "Birth rate (per 1000)")
```

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



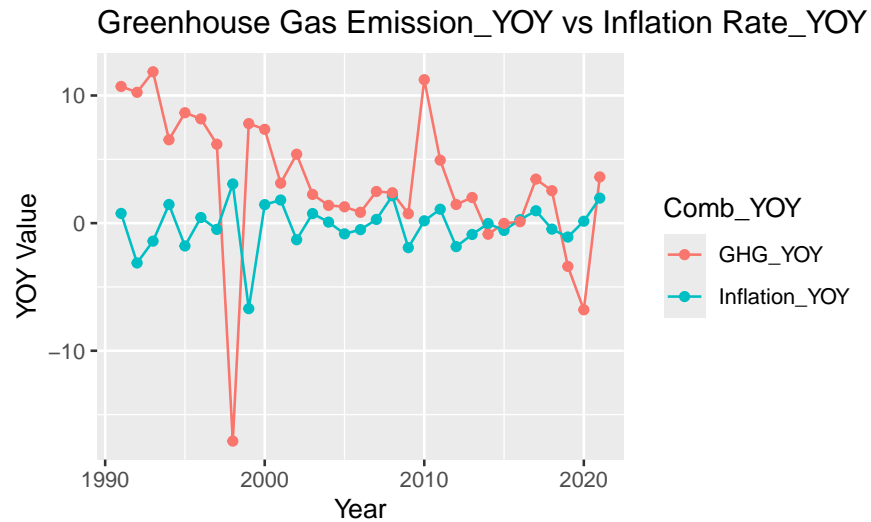
```
Group_dataset%>%
  pivot_longer(cols = c('Inflation_YOY', 'GHG_YOY'),
    names_to = 'Comb_YOY',
```

```

values_to = 'Val_YOY') %>%

ggplot(aes(y = Val_YOY, x = Year)) +
  geom_line(aes(color = Comb_YOY)) +
  geom_point(aes(color = Comb_YOY)) +
  labs(title = "Greenhouse Gas Emission_YOY vs Inflation Rate_YOY",
        x = "Year",
        y = "YOY Value")

```



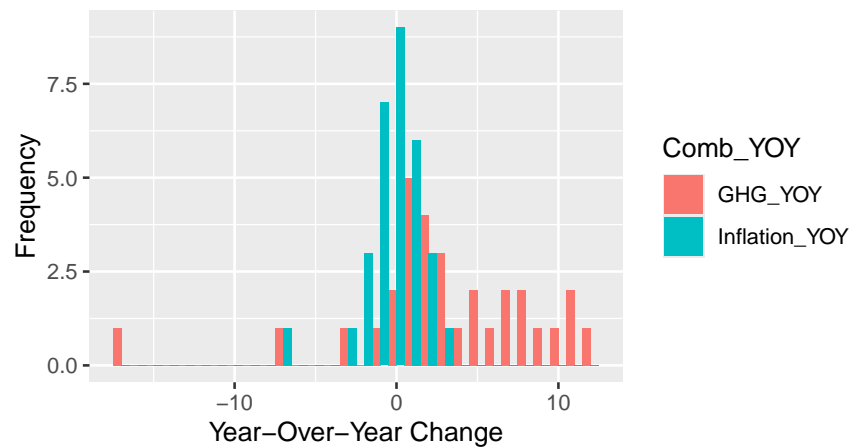
```

Group_dataset %>%
  pivot_longer(cols = c('Inflation_YOY', 'GHG_YOY'),
               names_to = 'Comb_YOY',
               values_to = 'Val_YOY') %>%

ggplot(aes(x = Val_YOY, fill = Comb_YOY)) +
  geom_histogram(position = 'dodge', bins = 30) +
  labs(title = "Distribution of Year-Over-Year Changes:
               Greenhouse Gas Emission vs Inflation rate",
        x = "Year-Over-Year Change",
        y = "Frequency")

```

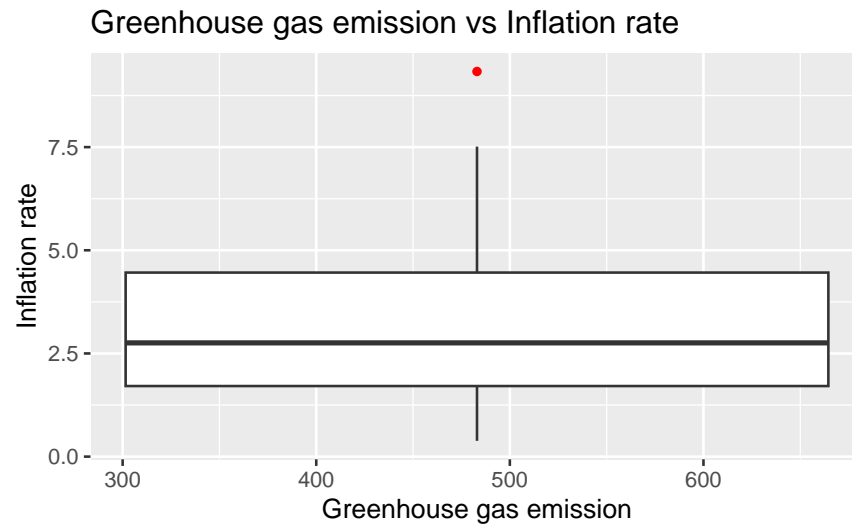

Distribution of Year-Over-Year Changes:
Greenhouse Gas Emission vs Inflation rate



```
Group_dataset %>%
  ggplot(mapping = aes(x = GHG, y = INF_Rate)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16) +
  stat_summary(
    aes(label = round(..y.., 1)),
    geom = "text",
    fun.data = function(y) {
      out <- boxplot.stats(y)$out
      if (length(out) == 0) return(NULL)
      data.frame(y = out)
    },
    hjust = -0.5
  ) +
  labs(
    title = "Greenhouse gas emission vs Inflation rate",
    x = 'Greenhouse gas emission',
    y = "Inflation rate"
  )
)
```

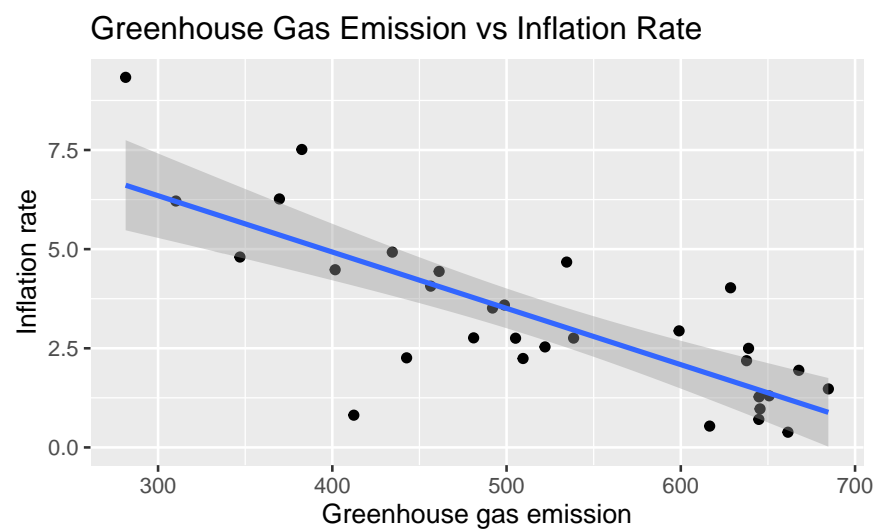
```
## Warning: Continuous x aesthetic
## i did you forget 'aes(group = ...)'?
```

```
## Warning: Computation failed in 'stat_summary()'.
## Caused by error in 'fix.by()':
## ! 'by' must specify uniquely valid columns
```



```
Group_dataset %>%
  ggplot() +
  geom_point(mapping = aes(x = GHG, y = INF_Rate)) +
  geom_smooth(mapping = aes(x = GHG, y = INF_Rate), method = "lm") +
  labs(
    title = "Greenhouse Gas Emission vs Inflation Rate",
    x = "Greenhouse gas emission",
    y = "Inflation rate")
```

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



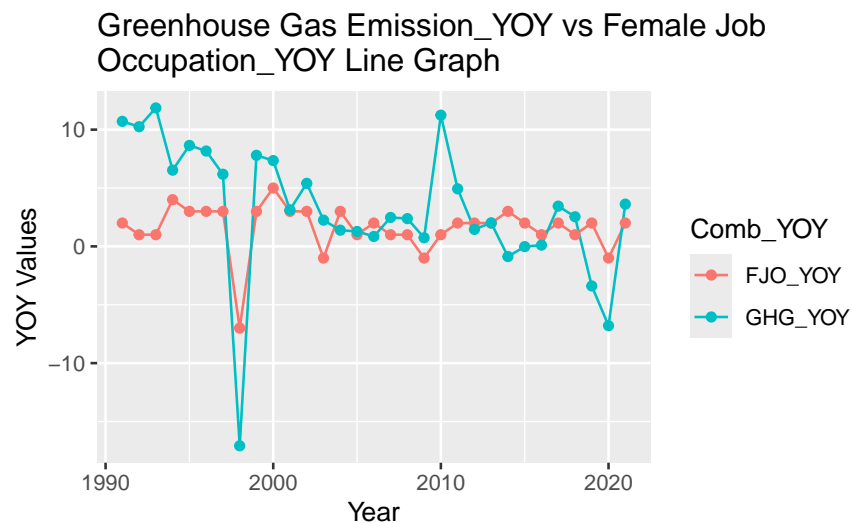
```
Group_dataset %>%
  pivot_longer(cols = c('FJO_YOY', 'GHG_YOY'),
    names_to = 'Comb_YOY',
```

```

values_to = 'Val_YOY') %>%

ggplot(aes(y = Val_YOY, x = Year)) +
  geom_line(aes(color = Comb_YOY)) +
  geom_point(aes(color = Comb_YOY)) +
  labs(title = "Greenhouse Gas Emission_YOY vs Female Job
Occupation_YOY Line Graph",
       x = "Year",
       y = "YOY Values")

```



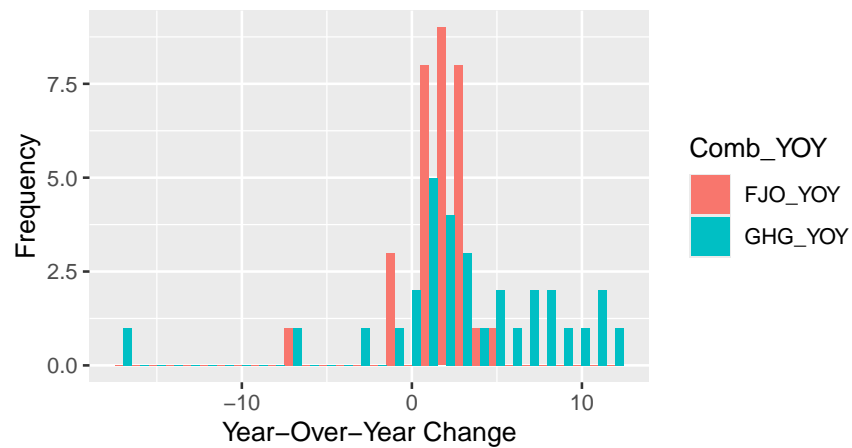
```

Group_dataset %>%
  pivot_longer(cols = c('FJO_YOY', 'GHG_YOY'),
               names_to = 'Comb_YOY',
               values_to = 'Val_YOY') %>%

ggplot(aes(x = Val_YOY, fill = Comb_YOY)) +
  geom_histogram(position = 'dodge', bins = 30) +
  labs(title = "Distribution of Year-Over-Year Changes:
GHG Emission vs Female Job Occupancy",
       x = "Year-Over-Year Change",
       y = "Frequency")

```

Distribution of Year-Over-Year Changes:
GHG Emission vs Female Job Occupancy

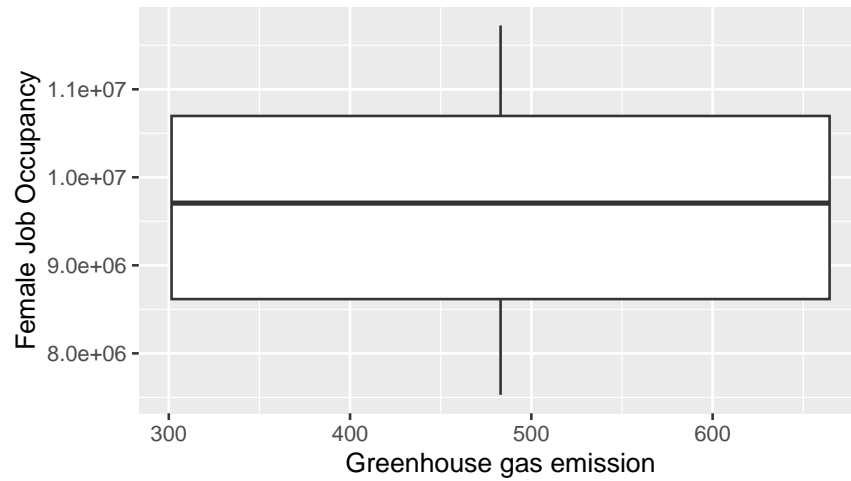


```
Group_dataset %>%
  ggplot(mapping = aes(x = GHG, y = FJO)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16) +
  stat_summary(
    aes(label = round(..y.., 1)),
    geom = "text",
    fun.data = function(y) {
      out <- boxplot.stats(y)$out
      if (length(out) == 0) return(NULL)
      data.frame(y = out)
    },
    hjust = -0.5
  ) +
  labs(
    title = "Greenhouse gas emission vs Female Job Occupancy ",
    x = 'Greenhouse gas emission',
    y = "Female Job Occupancy"
  )
)
```

```
## Warning: Continuous x aesthetic
## i did you forget 'aes(group = ...)'?
```

```
## Warning: Computation failed in 'stat_summary()'.
## Caused by error in 'fix.by()':
## ! 'by' must specify uniquely valid columns
```

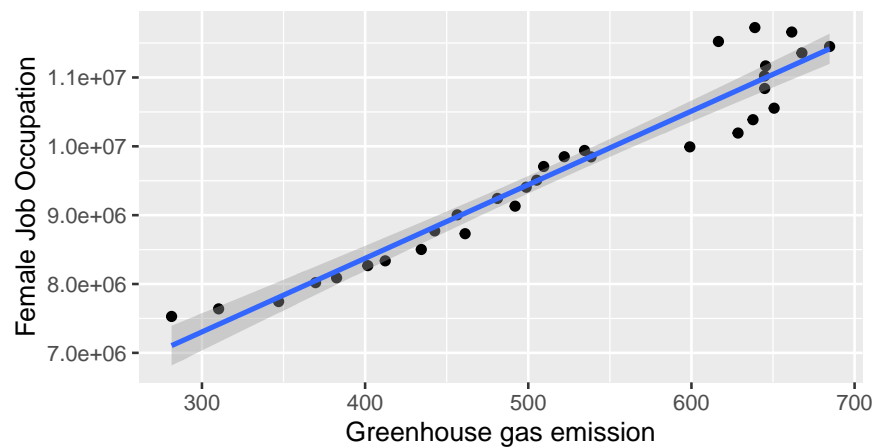
Greenhouse gas emission vs Female Job Occupancy



```
Group_dataset%>%
  ggplot() +
  geom_point(mapping = aes(x = GHG, y = FJO)) +
  geom_smooth(mapping=aes(x = GHG, y = FJO), method="lm")+
  labs(
    title = "Greenhouse Gas Emission vs Female Job
Occupation",
    x = "Greenhouse gas emission",
    y = "Female Job Occupation")
```

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

Greenhouse Gas Emission vs Female Job Occupation



Modeling & Hypothesis test- Envo_ft year-on-year (Byungwook Oh)

```
# Model
Envo_ft_model <- lm(Birth_YOY ~ GHG_YOY, data = Envo_ft)
```

```
# Tidy model
Envo_ft_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-2.7278213	0.3752186	-7.269953	0.0000001
GHG_YOY	0.0719243	0.0580149	1.239755	0.2250047

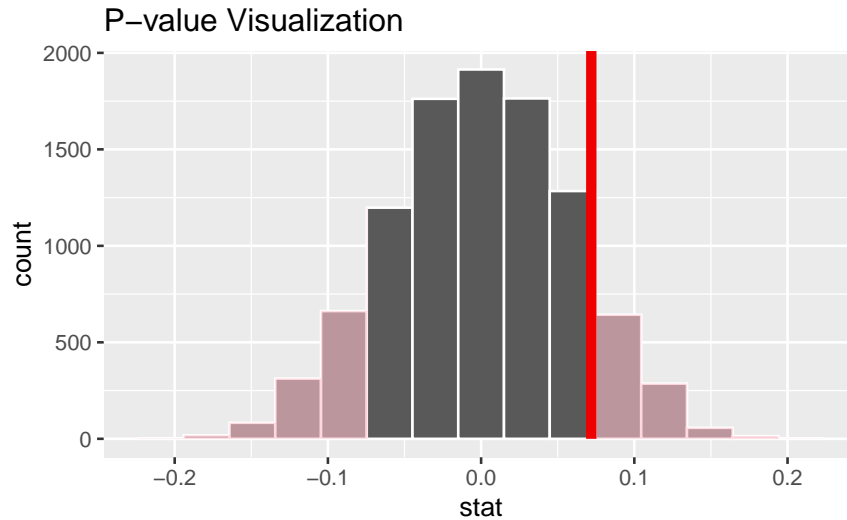
```
# Null distribution
Envo_null_distribution <- Envo_ft %>%
  specify(Birth_YOY ~ GHG_YOY) %>%
  hypothesize(null="independence") %>%
  generate(reps=10000, type="permute") %>%
  calculate(stat="slope")
```

```
# Observed stat
Observed_stat <- Envo_ft %>%
  specify(Birth_YOY ~ GHG_YOY) %>%
  calculate(stat="slope")
```

```
# P-value
Envo_null_distribution %>%
  get_p_value(obs_stat=Observed_stat, direction="both")
```

p_value
0.218

```
# P-value visualization
Envo_null_distribution %>%
  visualize() +
  shade_p_value(obs_stat=Observed_stat, direction= "both") +
  labs(title = "P-value Visualization")
```



Modeling & Hypothesis test- Envo_ft with actual value (Byungwook Oh)

```
# Model
Envo_ft_model <- lm(Birth_per_1000 ~ GHG, data = Envo_ft)
```

```
# Null distribution
Envo_null_distribution_rv <- Envo_ft %>%
  specify(Birth_per_1000 ~ GHG) %>%
  hypothesize(null="independence") %>%
  generate(reps=10000, type="permute") %>%
  calculate(stat="slope")
```

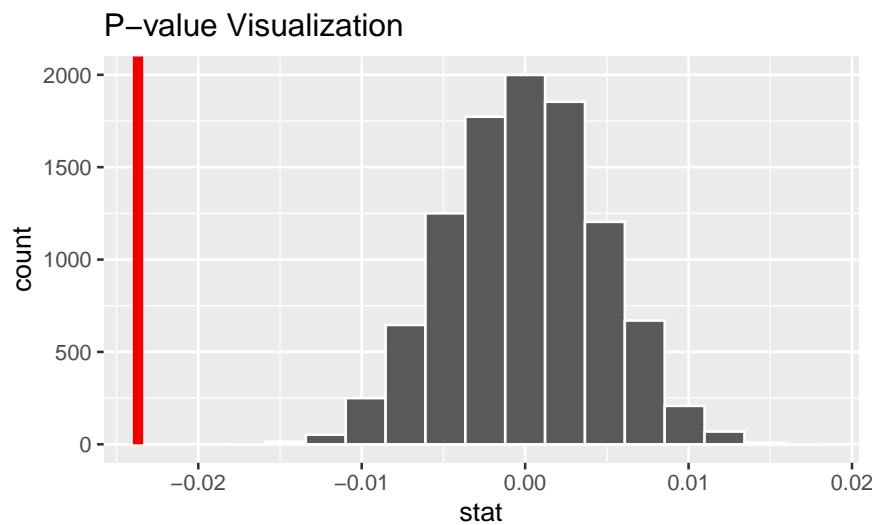
```
# Observed stat
Observed_stat <- Envo_ft %>%
  specify(Birth_per_1000 ~ GHG) %>%
  calculate(stat="slope")
```

```
# P-value
Envo_null_distribution_rv %>%
  get_p_value(obs_stat=Observed_stat, direction="both")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an approximation
## based on the number of 'reps' chosen in the 'generate()' step.
## i See 'get_p_value()' ('?infer::get_p_value()') for more information.
```

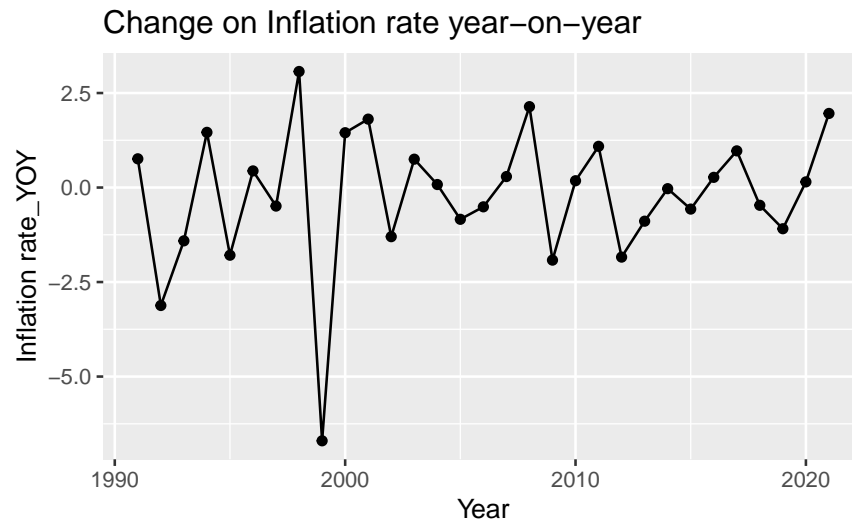
p_value
0

```
# P-value visualization
Envo_null_distribution_rv %>%
  visualize() +
  shade_p_value(obs_stat=Observed_stat, direction= "both") +
  labs(title = "P-value Visualization")
```



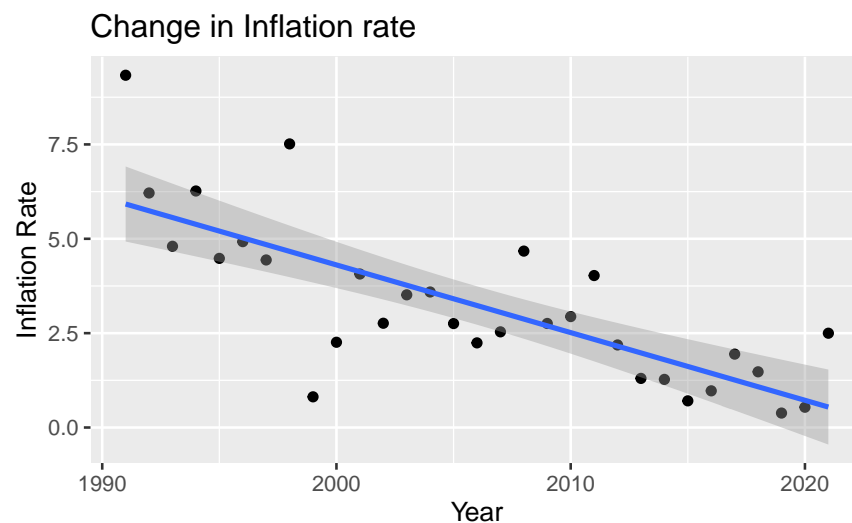
Variation and Covariation - Econ_ft (Dawon Kyoung)

```
Group_dataset %>%
  ggplot() +
  geom_line(mapping = aes(y = Inflation_YOY, x= Year)) +
  geom_point(mapping = aes(y = Inflation_YOY, x= Year)) +
  labs(title = "Change on Inflation rate year-on-year",
       y = "Inflation rate_YOY", x = "Year")
```

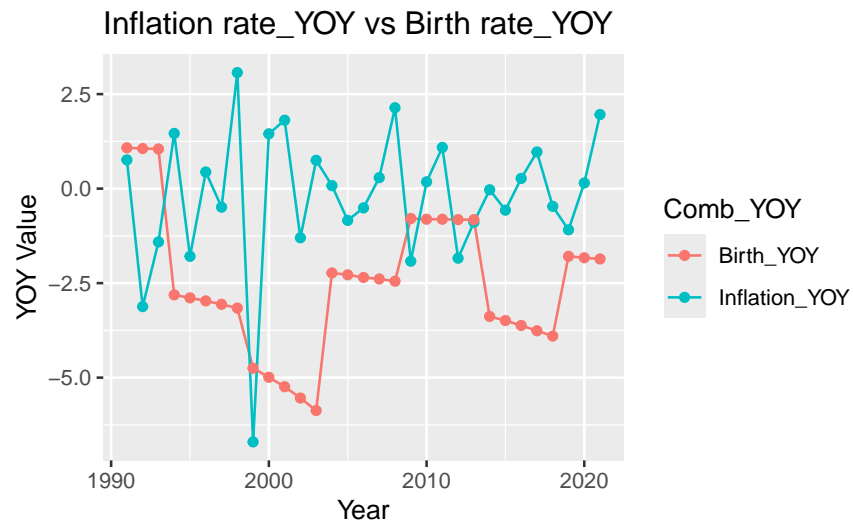
```
Econ_ft %>%
  ggplot() +
  geom_point(mapping = aes(y = INF_Rate, x= Year))+
  geom_smooth(mapping = (aes(y = INF_Rate, x = Year)), method = 'lm') +
  labs(title = "Change in Inflation rate",
       y = "Inflation Rate",
       x = "Year")
```

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



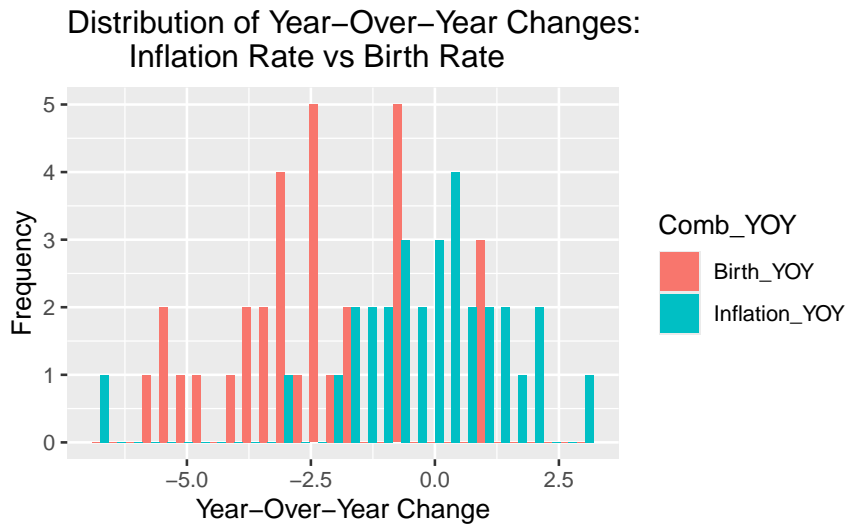
```
Group_dataset%>%
  pivot_longer(cols = c('Inflation_YOY', 'Birth_YOY'),
               names_to = 'Comb_YOY',
               values_to = 'Val_YOY')%>%
```

```
ggplot(aes(y = Val_YOY, x = Year)) +
  geom_line(aes(color = Comb_YOY)) +
  geom_point(aes(color = Comb_YOY)) +
  labs(
    title= 'Inflation rate_YOY vs Birth rate_YOY',
    x= 'Year',
    y= "YOY Value")
```



```
Group_dataset %>%
  pivot_longer(cols = c('Inflation_YOY', 'Birth_YOY'),
    names_to = 'Comb_YOY',
    values_to = 'Val_YOY') %>%

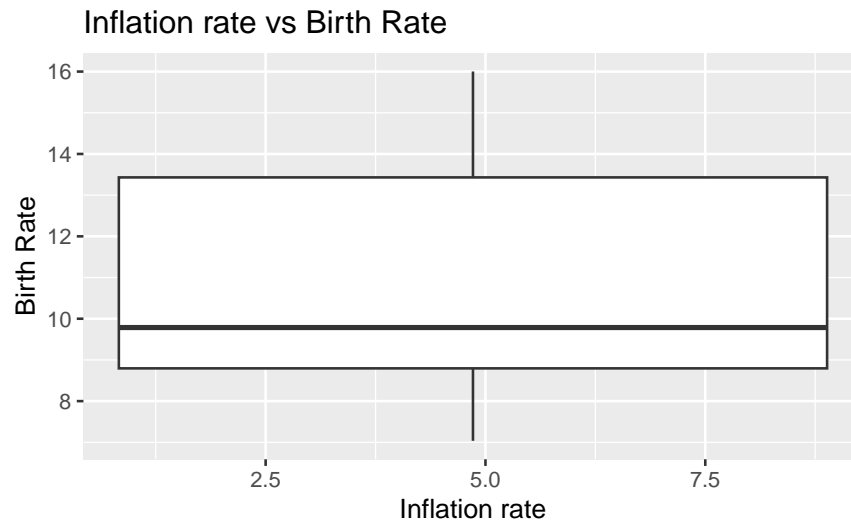
ggplot(aes(x = Val_YOY, fill = Comb_YOY)) +
  geom_histogram(position = 'dodge', bins = 30) +
  labs(title = "Distribution of Year-Over-Year Changes:
    Inflation Rate vs Birth Rate",
    x = "Year-Over-Year Change",
    y = "Frequency")
```



```
Econ_ft %>%
  ggplot(mapping = aes( x = INF_Rate, y = Birth_per_1000)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16) +
  stat_summary(
    aes(label = round(..y.., 1)),
    geom = "text",
    fun.data = function(y) {
      out <- boxplot.stats(y)$out
      if (length(out) == 0) return(NULL)
      data.frame(y = out)
    },
    hjust = -0.5
  ) +
  labs(
    title = "Inflation rate vs Birth Rate",
    x = 'Inflation rate',
    y = "Birth Rate")
```

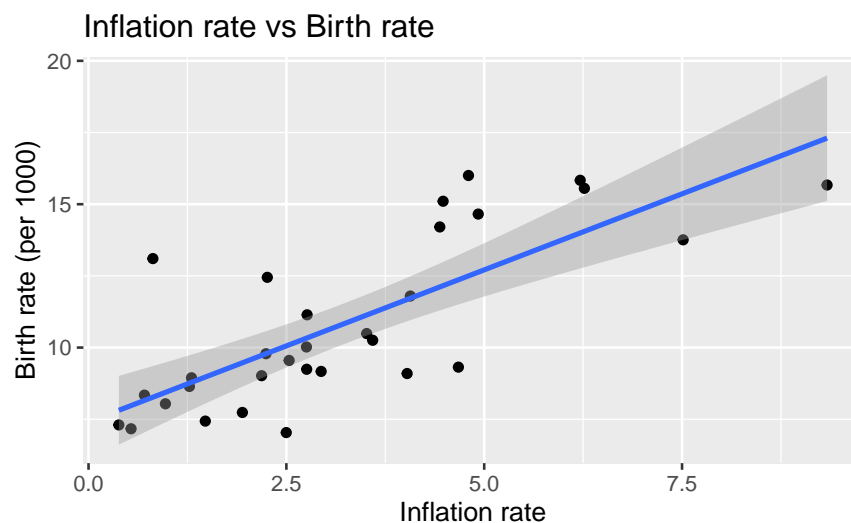
```
## Warning: Continuous x aesthetic
## i did you forget 'aes(group = ...)'?
```

```
## Warning: Computation failed in 'stat_summary()'.
## Caused by error in 'fix.by()':
## ! 'by' must specify uniquely valid columns
```



```
Econ_ft %>%
  ggplot() +
  geom_point(mapping=aes(x = INF_Rate, y = Birth_per_1000)) +
  geom_smooth(mapping = aes(x = INF_Rate, y = Birth_per_1000), method = 'lm') +
  labs(
    title= 'Inflation rate vs Birth rate',
    x= 'Inflation rate',
    y= "Birth rate (per 1000)")
```

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



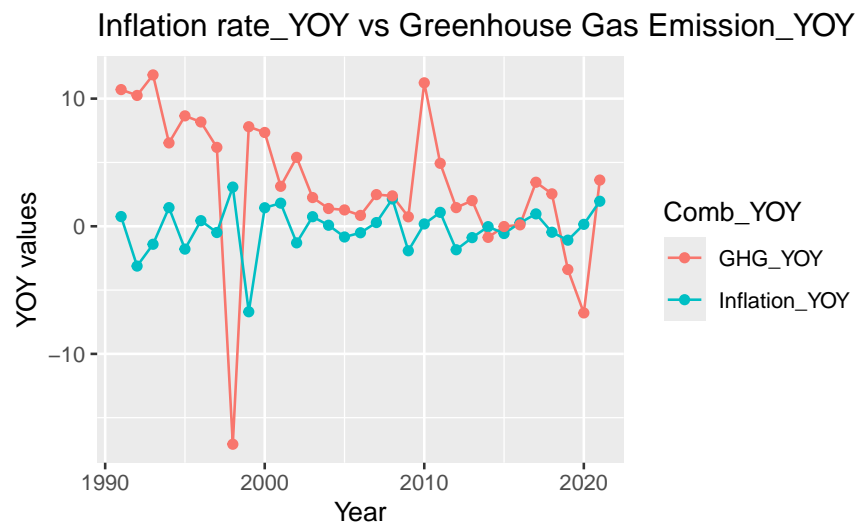
```
Group_dataset%>%
  pivot_longer(cols = c('Inflation_YOY', 'GHG_YOY'),
    names_to = 'Comb_YOY',
```

```

values_to = 'Val_YOY') %>%

ggplot(aes(y = Val_YOY, x = Year)) +
  geom_line(aes(color = Comb_YOY)) +
  geom_point(aes(color = Comb_YOY)) +
  labs(
    title= 'Inflation rate_YOY vs Greenhouse Gas Emission_YOY',
    x= 'Year',
    y= 'YOY values')

```



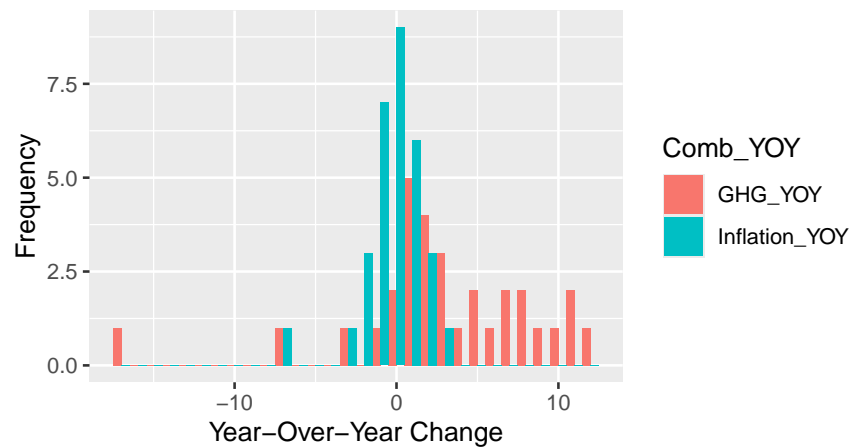
```

Group_dataset %>%
  pivot_longer(cols = c('Inflation_YOY', 'GHG_YOY'),
    names_to = 'Comb_YOY',
    values_to = 'Val_YOY') %>%

ggplot(aes(x = Val_YOY, fill = Comb_YOY)) +
  geom_histogram(position = 'dodge', bins = 30) +
  labs(title = "Distribution of Year-Over-Year Changes:
    Inflation Rate vs Greenhouse Gas Emission",
    x = "Year-Over-Year Change",
    y = "Frequency")

```

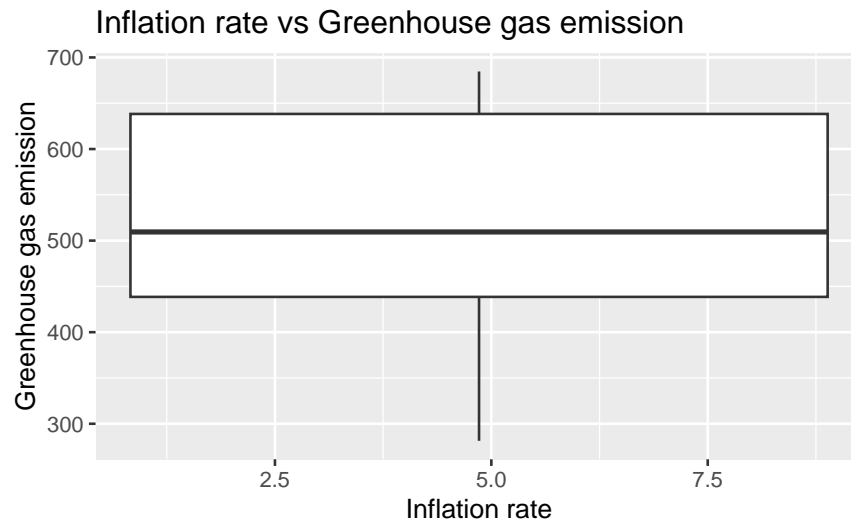
Distribution of Year-Over-Year Changes:
Inflation Rate vs Greenhouse Gas Emission



```
Group_dataset %>%
  ggplot(mapping = aes(x = INF_Rate, y = GHG)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16) +
  stat_summary(
    aes(label = round(..y.., 1)),
    geom = "text",
    fun.data = function(y) {
      out <- boxplot.stats(y)$out
      if (length(out) == 0) return(NULL)
      data.frame(y = out)
    },
    hjust = -0.5
  ) +
  labs(
    title = "Inflation rate vs Greenhouse gas emission",
    x = 'Inflation rate',
    y = "Greenhouse gas emission")
```

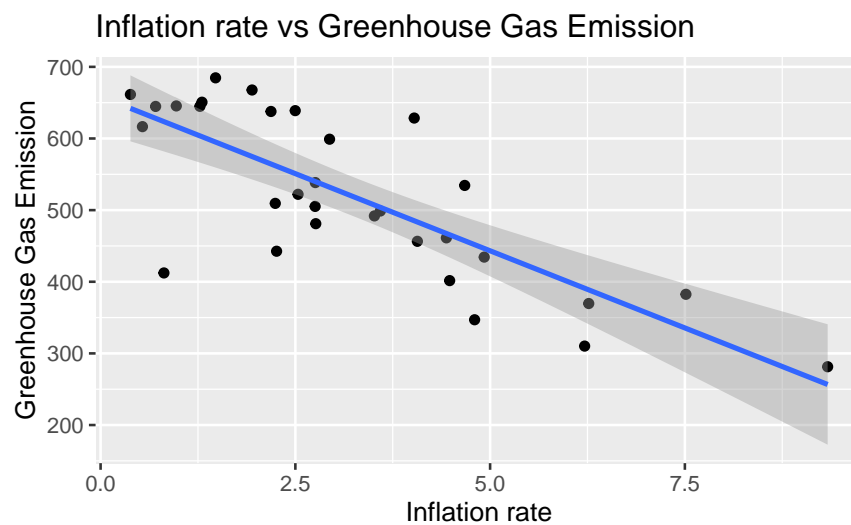
```
## Warning: Continuous x aesthetic
## i did you forget 'aes(group = ...)'?
```

```
## Warning: Computation failed in 'stat_summary()'.
## Caused by error in 'fix.by()':
## ! 'by' must specify uniquely valid columns
```



```
Group_dataset %>%
  ggplot() +
  geom_point(mapping= aes( x = INF_Rate, y = GHG))+
  geom_smooth(mapping=aes(x = INF_Rate, y = GHG),
method= 'lm')+
  labs(
    title= 'Inflation rate vs Greenhouse Gas Emission',
    x= 'Inflation rate',
    y= 'Greenhouse Gas Emission')
```

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



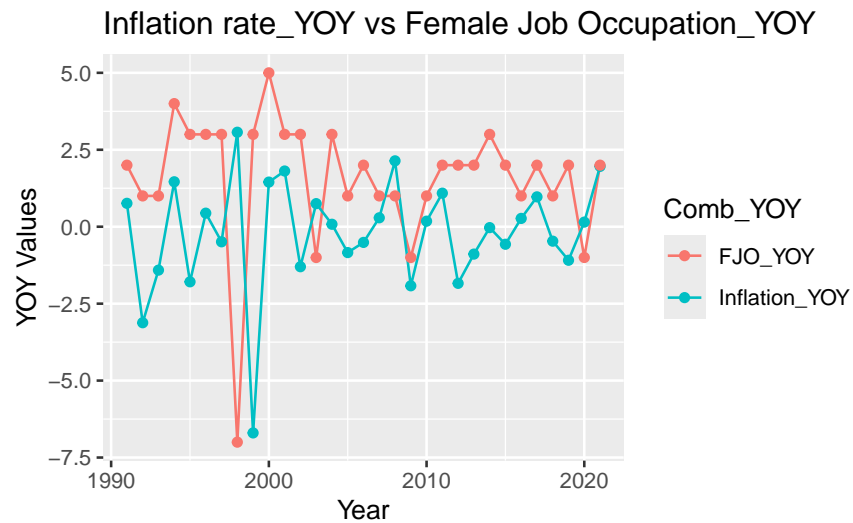
```
Group_dataset %>%
  pivot_longer(cols = c('Inflation_YOY', 'FJO_YOY'),
```

```

names_to = 'Comb_YOY',
values_to = 'Val_YOY') %>%

ggplot(aes(x = Year, y = Val_YOY)) +
  geom_line(aes(color = Comb_YOY)) +
  geom_point(aes(color = Comb_YOY)) +
  labs(
    title = 'Inflation rate_YOY vs Female Job Occupation_YOY',
    x = 'Year',
    y = 'YOY Values')

```



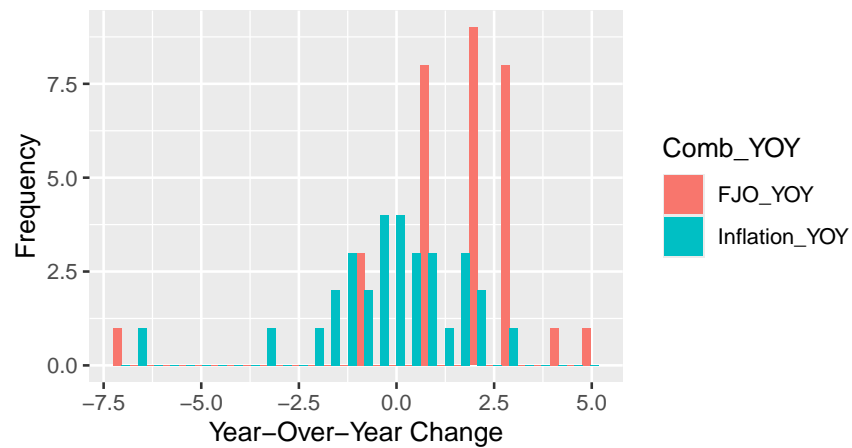
```

Group_dataset %>%
  pivot_longer(cols = c('FJO_YOY', 'Inflation_YOY'),
    names_to = 'Comb_YOY',
    values_to = 'Val_YOY') %>%

ggplot(aes(x = Val_YOY, fill = Comb_YOY)) +
  geom_histogram(position = 'dodge', bins = 30) +
  labs(title = "Distribution of Year-Over-Year Changes:
    Inflation Rate vs Female Job Occupancy",
    x = "Year-Over-Year Change",
    y = "Frequency")

```

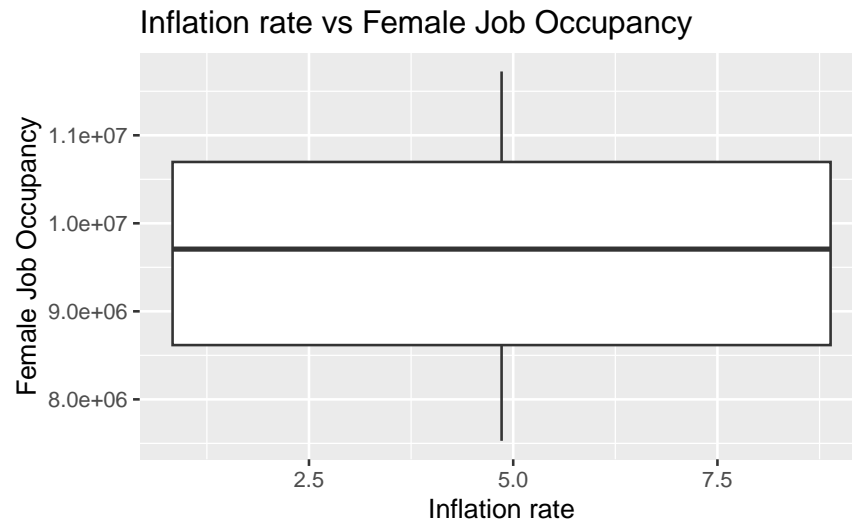

Distribution of Year-Over-Year Changes:
Inflation Rate vs Female Job Occupancy



```
Group_dataset %>%
  ggplot(mapping = aes(x = INF_Rate, y = FJO)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16) +
  stat_summary(
    aes(label = round(..y.., 1)),
    geom = "text",
    fun.data = function(y) {
      out <- boxplot.stats(y)$out
      if (length(out) == 0) return(NULL)
      data.frame(y = out)
    },
    hjust = -0.5
  ) +
  labs(
    title = "Inflation rate vs Female Job Occupancy ",
    x = 'Inflation rate',
    y = "Female Job Occupancy")
```

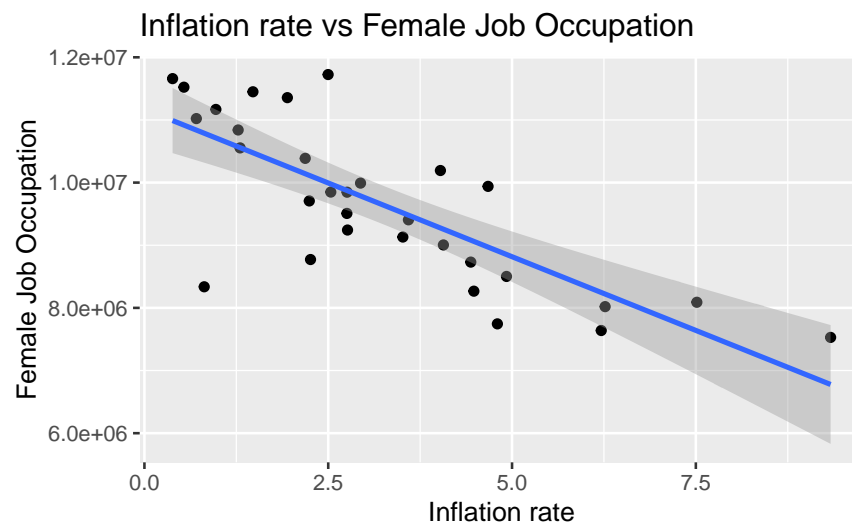
```
## Warning: Continuous x aesthetic
## i did you forget 'aes(group = ...)'?
```

```
## Warning: Computation failed in 'stat_summary()'.
## Caused by error in 'fix.by()':
## ! 'by' must specify uniquely valid columns
```



```
Group_dataset %>%
  ggplot() +
  geom_point(mapping = aes(x= INF_Rate, y= FJO))+
  geom_smooth(mapping=aes(x= INF_Rate, y= FJO),
method= 'lm')+
  labs(
    title= 'Inflation rate vs Female Job Occupation',
    x= 'Inflation rate',
    y= 'Female Job Occupation')
```

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



Modeling & Hypothesis test- Econ_ft year on year (Daeun Choi)

```
# Model
Econ_ft_model <- lm(Birth_YOY ~ Inflation_YOY, data=Econ_ft)
```

```
# Tidy model
Econ_ft_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-2.5239630	0.3345546	-7.5442476	0.0000000
Inflation_YOY	-0.1266971	0.1850287	-0.6847431	0.4989426

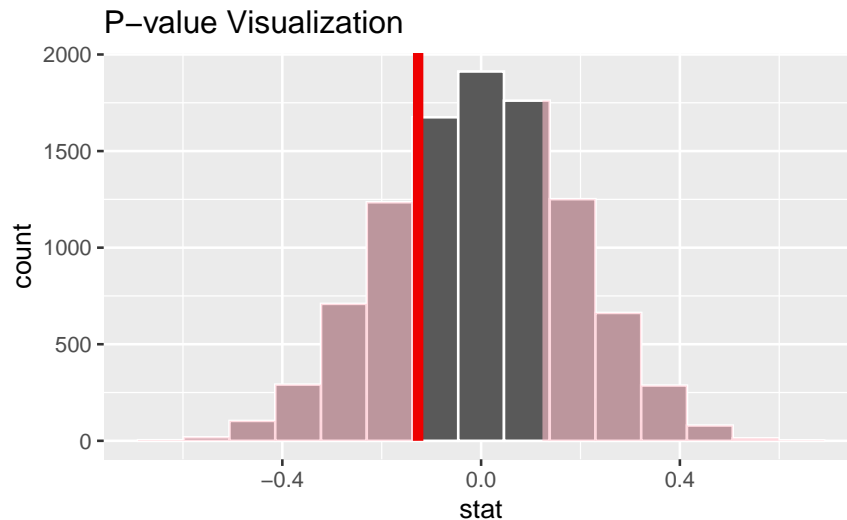
```
# Null distribution
Econ_null_distribution <- Econ_ft %>%
  specify(Birth_YOY ~ Inflation_YOY) %>%
  hypothesize(null="independence") %>%
  generate(reps=10000, type="permute") %>%
  calculate(stat="slope")
```

```
# Observed stat
Observed_stat <- Econ_ft %>%
  specify(Birth_YOY ~ Inflation_YOY) %>%
  calculate(stat="slope")
```

```
# P-value
Econ_null_distribution %>%
  get_p_value(obs_stat=Observed_stat, direction="both")
```

p_value
0.5108

```
# P-value visualization
Econ_null_distribution %>%
  visualize() +
  shade_p_value(obs_stat=Observed_stat, direction= "both") +
  labs(title = "P-value Visualization")
```



Modeling & Hypothesis test - Econ_ft with actual value (Daeun Choi)

```
# Model
Econ_ft_model <- lm(Birth_per_1000 ~ INF_Rate, data = Econ_ft)
```

```
# Null distribution
Econ_null_distribution_rv <- Econ_ft %>%
  specify(Birth_per_1000 ~ INF_Rate) %>%
  hypothesize(null="independence") %>%
  generate(reps=10000, type="permute") %>%
  calculate(stat="slope")
```

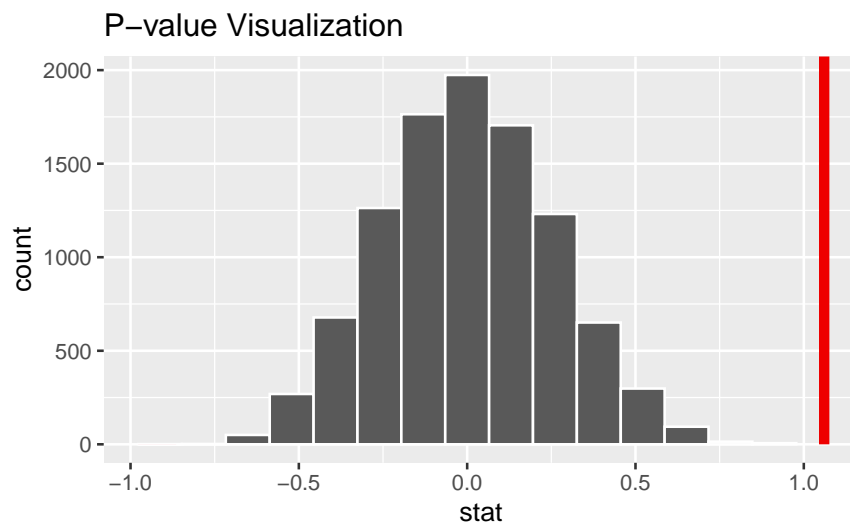
```
# Observed stat
Observed_stat <- Econ_ft %>%
  specify(Birth_per_1000 ~ INF_Rate) %>%
  calculate(stat="slope")
```

```
# P-value
Econ_null_distribution_rv %>%
  get_p_value(obs_stat=Observed_stat, direction="both")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an approximation
## based on the number of 'reps' chosen in the 'generate()' step.
## i See 'get_p_value()' ('?infer::get_p_value()') for more information.
```

p_value
0

```
# P-value visualization
Econ_null_distribution_rv %>%
  visualize() +
  shade_p_value(obs_stat=Observed_stat, direction= "both") +
  labs(title = "P-value Visualization")
```



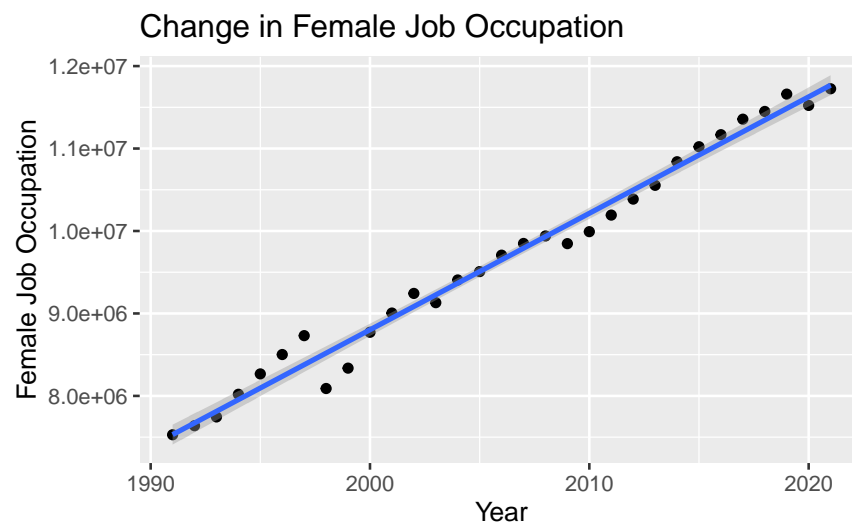
Variation and Covariation - Soci_ft (Eunho Cha)

```
Group_dataset %>%
  ggplot() +
  geom_line(mapping = aes(y = FJO_YOY, x= Year)) +
  geom_point(mapping = aes(y = FJO_YOY, x= Year))+
  labs(title = "Change on Female Job Occupation year-on-year",
       y = "FJO_YOY",
       x = "Year")
```



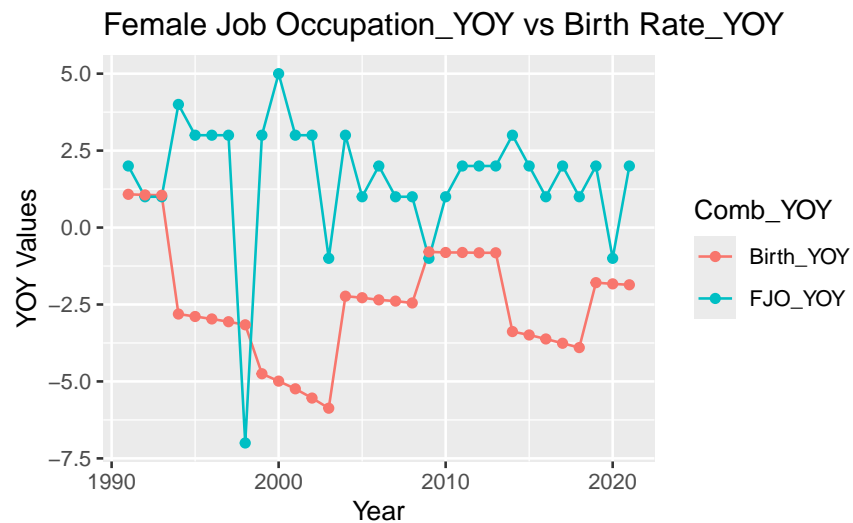
```
Soci_ft %>%
  ggplot() +
  geom_point(mapping = aes(y = FJO, x = Year)) +
  geom_smooth(mapping = aes(y = FJO, x = Year), method = 'lm') +
  labs(title = "Change in Female Job Occupation",
       y = "Female Job Occupation",
       x = "Year")
```

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



```
Group_dataset%>%
  pivot_longer(cols = c('FJO_YOY', 'Birth_YOY'),
               names_to = 'Comb_YOY',
               values_to = 'Val_YOY')%>%
```

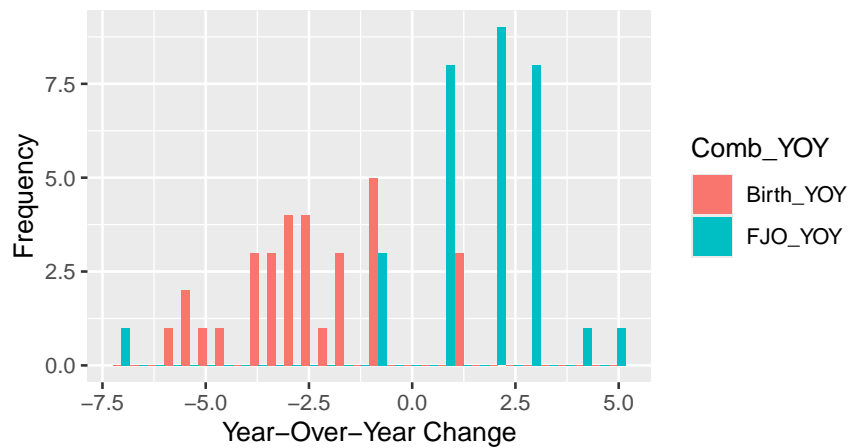
```
ggplot(aes(y = Val_YOY, x = Year)) +
  geom_line(aes(color = Comb_YOY)) +
  geom_point(aes(color = Comb_YOY)) +
  labs(
    title="Female Job Occupation_YOY vs Birth Rate_YOY",
    x= "Year",
    y= "YOY Values"
  )
)
```



```
Group_dataset %>%
  pivot_longer(cols = c('FJO_YOY', 'Birth_YOY'),
    names_to = 'Comb_YOY',
    values_to = 'Val_YOY') %>%

ggplot(aes(x = Val_YOY, fill = Comb_YOY)) +
  geom_histogram(position = 'dodge', bins = 30) +
  labs(title = "Distribution of Year-Over-Year Changes:
    Female Job Occupancy vs Birth rate",
    x = "Year-Over-Year Change",
    y = "Frequency")
```

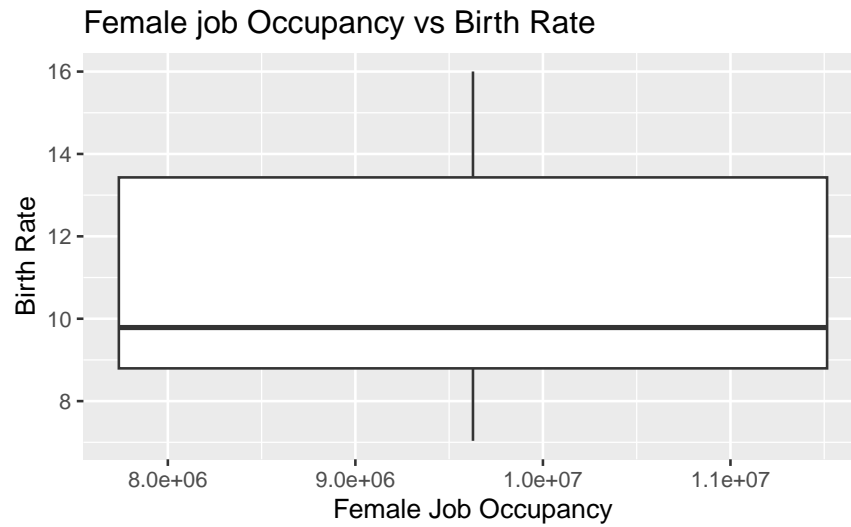
Distribution of Year-Over-Year Changes:
Female Job Occupancy vs Birth rate



```
Group_dataset %>%
  ggplot(mapping = aes(x = FJO, y = Birth_per_1000)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16) +
  stat_summary(
    aes(label = round(..y.., 1)),
    geom = "text",
    fun.data = function(y) {
      out <- boxplot.stats(y)$out
      if (length(out) == 0) return(NULL)
      data.frame(y = out)
    },
    hjust = -0.5
  ) +
  labs(
    title = "Female job Occupancy vs Birth Rate",
    x = 'Female Job Occupancy',
    y = "Birth Rate")
```

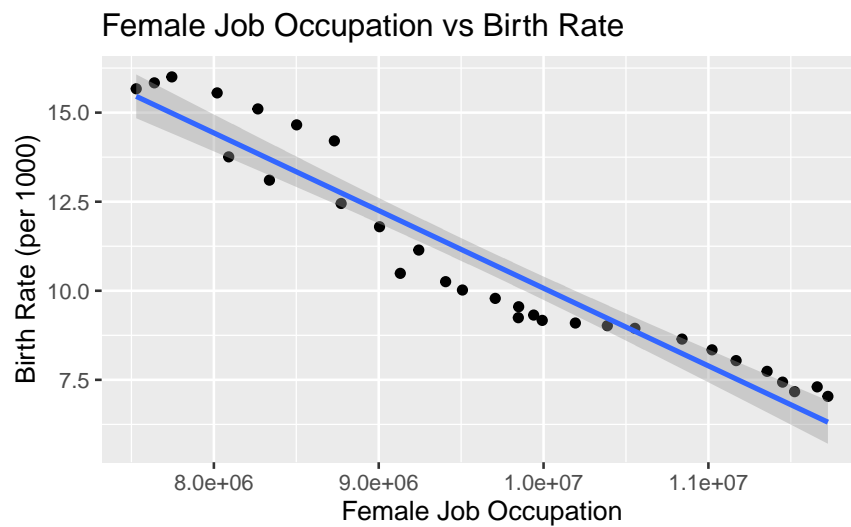
```
## Warning: Continuous x aesthetic
## i did you forget 'aes(group = ...)'?
```

```
## Warning: Computation failed in 'stat_summary()'.
## Caused by error in 'fix.by()':
## ! 'by' must specify uniquely valid columns
```

```
Soci_ft %>%
  ggplot() +
  geom_point(mapping = aes(x= FJO, y= Birth_per_1000))+
  geom_smooth(mapping=aes(x = FJO, y = Birth_per_1000), method="lm")+
  labs(
    title="Female Job Occupation vs Birth Rate",
    x= "Female Job Occupation",
    y= "Birth Rate (per 1000)"
  )
```

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



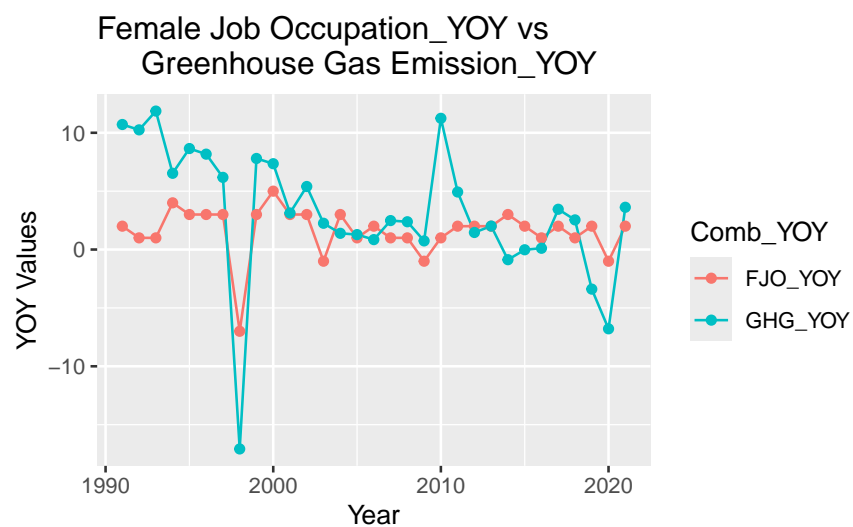
```
Group_dataset%>%
  pivot_longer(cols = c('FJO_YOY', 'GHG_YOY'),
```

```

names_to = 'Comb_YOY',
values_to = 'Val_YOY') %>%

ggplot(aes(y = Val_YOY, x = Year)) +
  geom_line(aes(color = Comb_YOY)) +
  geom_point(aes(color = Comb_YOY)) +
  labs(
    title = "Female Job Occupation_YOY vs
    Greenhouse Gas Emission_YOY",
    x = "Year",
    y = "YOY Values"
  )

```



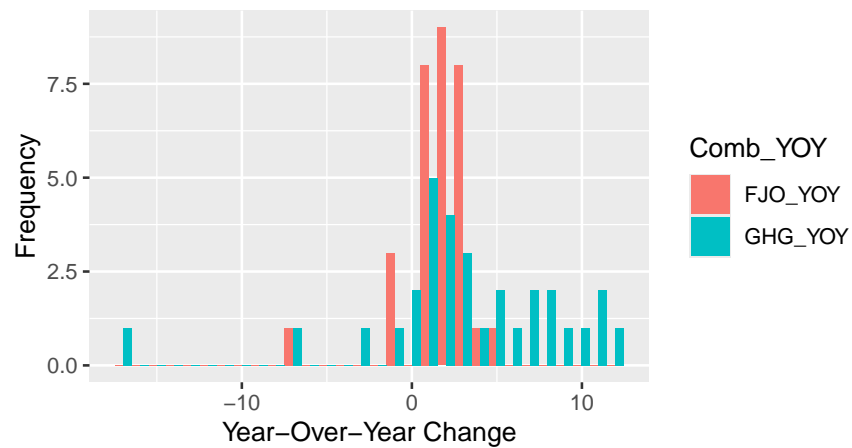
```

Group_dataset %>%
  pivot_longer(cols = c('FJO_YOY', 'GHG_YOY'),
    names_to = 'Comb_YOY',
    values_to = 'Val_YOY') %>%

ggplot(aes(x = Val_YOY, fill = Comb_YOY)) +
  geom_histogram(position = 'dodge', bins = 30) +
  labs(title = "Distribution of Year-Over-Year Changes:
    Female Job Occupancy vs Greenhouse Gas Emission",
    x = "Year-Over-Year Change",
    y = "Frequency")

```

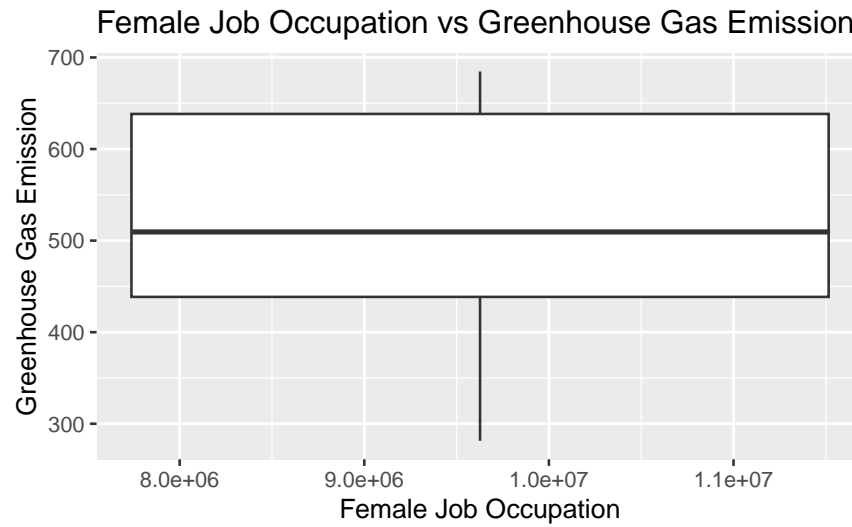
Distribution of Year–Over–Year Changes:
Female Job Occupancy vs Greenhouse Gas Emission



```
Group_dataset %>%
  ggplot(mapping = aes(x = FJO, y = GHG)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16) +
  stat_summary(
    aes(label = round(..y.., 1)),
    geom = "text",
    fun.data = function(y) {
      out <- boxplot.stats(y)$out
      if (length(out) == 0) return(NULL)
      data.frame(y = out)
    },
    hjust = -0.5
  ) +
  labs(
    title= "Female Job Occupation vs Greenhouse Gas Emission",
    x= "Female Job Occupation",
    y= "Greenhouse Gas Emission"
  )
)
```

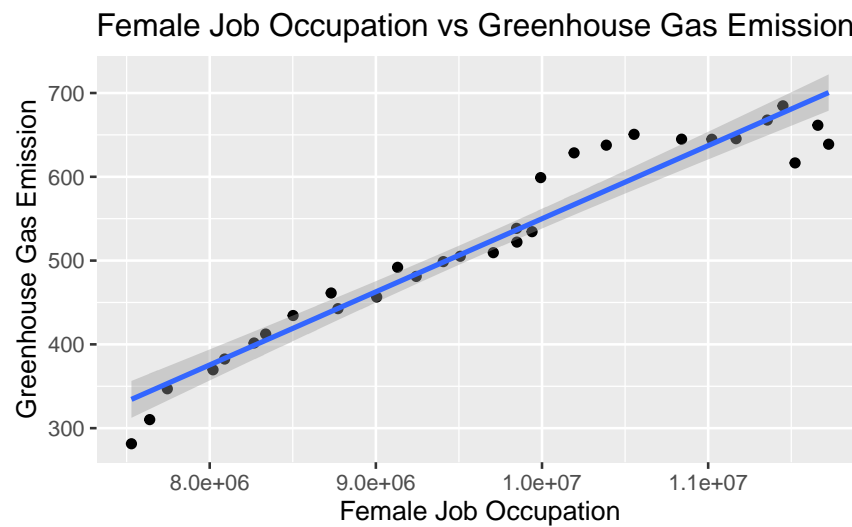
```
## Warning: Continuous x aesthetic
## i did you forget 'aes(group = ...)'?
```

```
## Warning: Computation failed in 'stat_summary()'.
## Caused by error in 'fix.by()':
## ! 'by' must specify uniquely valid columns
```



```
Group_dataset %>%
  ggplot() +
  geom_point(mapping= aes (x= FJO, y= GHG)) +
  geom_smooth(mapping=aes (x= FJO, y=GHG), method="lm")+
  labs(
    title= "Female Job Occupation vs Greenhouse Gas Emission",
    x= "Female Job Occupation",
    y= "Greenhouse Gas Emission"
  )
```

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



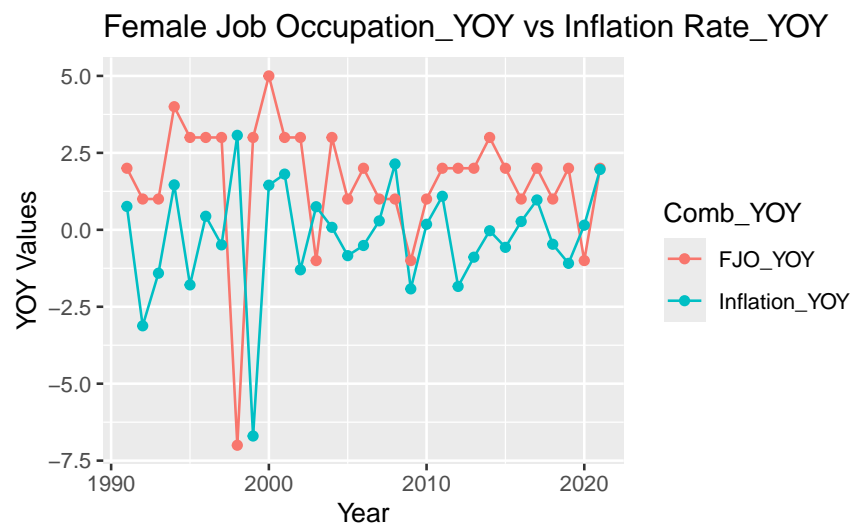
```
Group_dataset%>%
  pivot_longer(cols = c('FJO_YOY', 'Inflation_YOY'),
```

```

names_to = 'Comb_YOY',
values_to = 'Val_YOY') %>%

ggplot(aes(y = Val_YOY, x = Year)) +
  geom_line(aes(color = Comb_YOY)) +
  geom_point(aes(color = Comb_YOY)) +
  labs(
    title= "Female Job Occupation_YOY vs Inflation Rate_YOY",
    x= "Year",
    y= "YOY Values"
  )

```



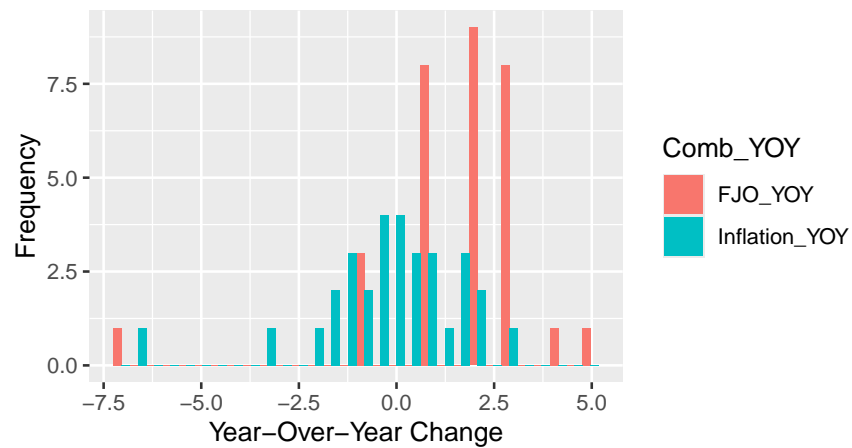
```

Group_dataset %>%
  pivot_longer(cols = c('FJO_YOY', 'Inflation_YOY'),
    names_to = 'Comb_YOY',
    values_to = 'Val_YOY') %>%

ggplot(aes(x = Val_YOY, fill = Comb_YOY)) +
  geom_histogram(position = 'dodge', bins = 30) +
  labs(title = "Distribution of Year-Over-Year Changes:
    Female Job Occupancy vs Inflation Rate",
    x = "Year-Over-Year Change",
    y = "Frequency")

```

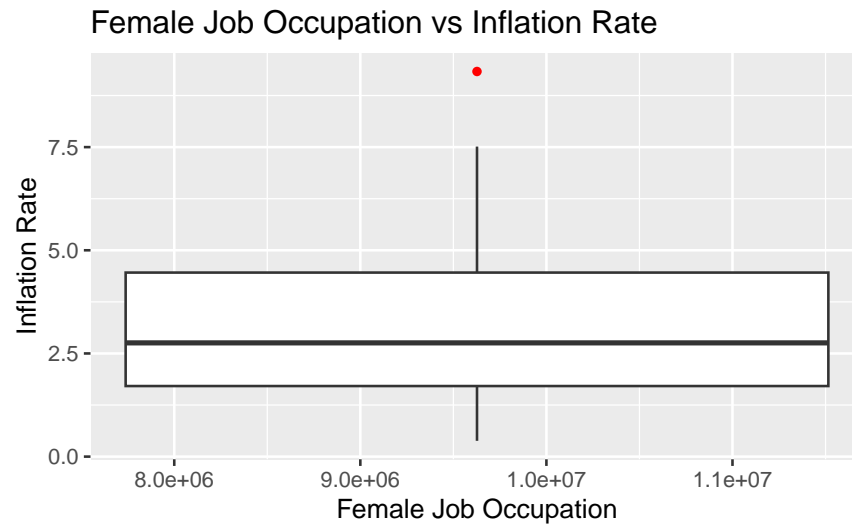
Distribution of Year-Over-Year Changes:
Female Job Occupancy vs Inflation Rate



```
Group_dataset %>%
  ggplot(mapping = aes(x = FJO, y = INF_Rate)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16) +
  stat_summary(
    aes(label = round(..y.., 1)),
    geom = "text",
    fun.data = function(y) {
      out <- boxplot.stats(y)$out
      if (length(out) == 0) return(NULL)
      data.frame(y = out)
    },
    hjust = -0.5
  ) +
  labs(
    title = "Female Job Occupation vs Inflation Rate",
    x = "Female Job Occupation",
    y = "Inflation Rate"
  )
```

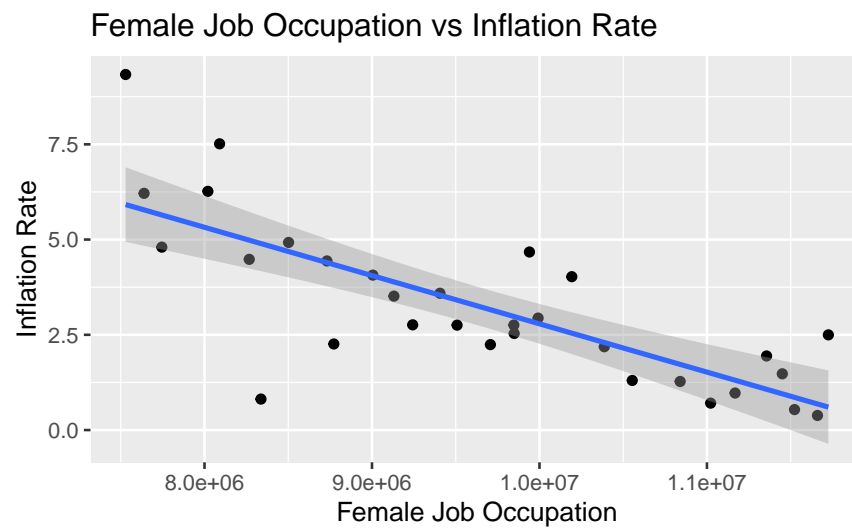
```
## Warning: Continuous x aesthetic
## i did you forget 'aes(group = ...)'?
```

```
## Warning: Computation failed in 'stat_summary()'.
## Caused by error in 'fix.by()':
## ! 'by' must specify uniquely valid columns
```



```
Group_dataset %>%
  ggplot() +
  geom_point(mapping= aes (x= FJO, y=INF_Rate)) +
  geom_smooth(mapping=aes (x= FJO, y=INF_Rate), method="lm")+
  labs(
    title= "Female Job Occupation vs Inflation Rate",
    x= "Female Job Occupation",
    y= "Inflation Rate"
  )
```

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



Modeling & Hypothesis test - Soci_ft year on year (Duy Tran)

```
# Model
Soci_ft_model <- lm(Birth_YOY ~ FJO_YOY, data = Soci_ft)
```

```
# Tidy model
Soci_ft_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-2.3081339	0.4185330	-5.5148199	0.0000061
FJO_YOY	-0.1207724	0.1611897	-0.7492564	0.4597376

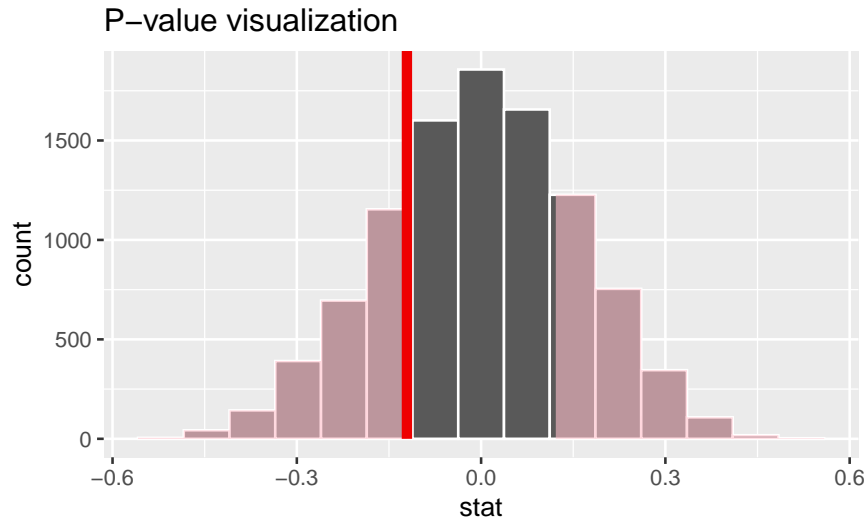
```
# Null distribution
Soci_null_distribution <- Soci_ft %>%
  specify(Birth_YOY ~ FJO_YOY) %>%
  hypothesize(null="independence") %>%
  generate(reps=10000, type="permute") %>%
  calculate(stat="slope")
```

```
# Observed stat
Observed_stat <- Soci_ft %>%
  specify(Birth_YOY ~ FJO_YOY) %>%
  calculate(stat="slope")
```

```
# P-value
Soci_null_distribution %>%
  get_p_value(obs_stat=Observed_stat, direction="both")
```

p_value
0.4508

```
# P-value visualization
Soci_null_distribution %>%
  visualize() +
  shade_p_value(obs_stat=Observed_stat, direction="both") +
  labs(title = "P-value visualization")
```

Modeling & Hypothesis test- Soci_ft with actual value (Duy Tran)

```
# Model
Soci_ft_model <- lm(Birth_per_1000 ~ FJO, data = Soci_ft)
```

```
# Null distribution
Soci_null_distribution_rv <- Soci_ft %>%
  specify(Birth_per_1000 ~ FJO) %>%
  hypothesize(null="independence") %>%
  generate(reps=10000, type="permute") %>%
  calculate(stat="slope")
```

```
# Observed stat
Observed_stat <- Soci_ft %>%
  specify(Birth_per_1000 ~ FJO) %>%
  calculate(stat="slope")
```

```
# P-value
Soci_null_distribution_rv %>%
  get_p_value(obs_stat=Observed_stat, direction="both")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an approximation
## based on the number of 'reps' chosen in the 'generate()' step.
## i See 'get_p_value()' ('?infer::get_p_value()') for more information.
```

p_value
0

```
# P-value visualization
Soci_null_distribution_rv %>%
  visualize() +
  shade_p_value(obs_stat=Observed_stat, direction= "both") +
  labs(title = "P-value Visualization")
```

