Final Project Part 2

Kristie Kooken

2022-08-07

```r
# read in data

setwd("C:/Users/kkooken/Documents/EDU/520/R/dsc520-1")

# Setting libraries
library(readxl)
library(dplyr)
library(ggplot2)
library(Hmisc)

options(scipen = 999)

# Function needed later
"%!in%" <- Negate("%in%")

# Reading in all data sources

# #5, Real Estate Witch House Price to Income Ratio Study, 2021
home_in_df <- read.csv("data/finalproject/homechg_incomechg_f.csv")
names(home_in_df) <- c("Year", "Income_change", "Home_change")

# creating numeric as needed
home_in_df$Year <- as.numeric(home_in_df$Year)
str(home_in_df)
```

```
## 'data.frame':    49 obs. of  3 variables:
##  $ Year         : num  1965 1970 1975 1976 1977 ...
##  $ Income_change: num  0 0.03 15.32 23.47 28.36 ...
##  $ Home_change  : num  0 2.7 0.13 1.69 2.23 5.57 3.54 -1.89 -4.15 -4.59 ...
```

```r
# #1, This is Federal Reserve Economic Data creating yr variable, creating
# average from quarter data
fred_df <- read_excel("data/finalproject/ASPUS_f.xls", sheet = "Newfred")
str(fred_df)
```

```
## tibble [59 x 2] (S3: tbl_df/tbl/data.frame)
##  $ Average_home_sale: num [1:59] 19375 20300 21450 22925 24125 ...
##  $ Year             : num [1:59] 1963 1964 1965 1966 1967 ...
```

```r
# #2, Economic Policy Institute from February 20, 2020
ecopoly_df <- read_excel("data/finalproject/economicpolicy_prod_work_f.xlsx")
# renaming and creating numeric as needed
names(ecopoly_df) <- c("Year", "Hr_comp", "Net_prod")
ecopoly_df$Year <- as.numeric(ecopoly_df$Year)
ecopoly_df$Hr_comp <- as.numeric(ecopoly_df$Hr_comp)
ecopoly_df$Net_prod <- as.numeric(ecopoly_df$Net_prod)
str(ecopoly_df)
```

```
## tibble [71 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Year    : num [1:71] 1948 1949 1950 1951 1952 ...
##  $ Hr_comp : num [1:71] 0 6.24 10.46 11.74 15.02 ...
##  $ Net_prod: num [1:71] 0 1.55 9.34 12.24 15.49 ...
```

```r
# #3 U.S. Inflation Rate History and Forecast
infl_df <- read_excel("data/finalproject/inflation_overtime_f.xlsx", sheet = "no_fmt")
# renaming and creating numeric as needed
names(infl_df) <- c("Year", "Inflation_chg", "FedRate_x5", "BusCyc", "Events")
infl_df$Year <- as.numeric(infl_df$Year)
str(infl_df)
```

```
## tibble [94 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Year         : num [1:94] 1929 1930 1931 1932 1933 ...
##  $ Inflation_chg: num [1:94] 0.6 -6.4 -9.3 -10.3 0.8 1.5 3 1.4 2.9 -2.8 ...
##  $ FedRate_x5   : num [1:94] NA NA NA NA NA NA NA NA NA NA ...
##  $ BusCyc       : chr [1:94] "August peak" "Contraction (-8.5%)" "Contraction (-6.4%)" "Contraction
##  $ Events       : chr [1:94] "Market crash" "Smoot-Hawley" "Dust Bowl" "Hoover tax hikes" ...
```

```r
# new inflation that starts at 1967
infl_df2 <- filter(infl_df, Year > 1966)

# creating a dichotomous variable for inflation, based on sources 5% or lower
# is 'normal' while greater can signify a higher inflation
infl_df2$InflatiYN <- as.numeric(infl_df2$Inflation_chg >= 5)
infl_df2$InflatCYN <- infl_df2$Inflation_chg >= 5

# #4 Historical Median Income Using Alternative Price Indices: 1967 to 2020
tabd_df <- read_excel("data/finalproject/medianincome_tableD1_f.xlsx")
# creating numeric as needed
names(tabd_df) <- c("Year", "MedIncome", "MarginError")
str(tabd_df)
```

```
## tibble [56 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Year       : num [1:56] 1967 1968 1969 1970 1971 ...
##  $ MedIncome  : num [1:56] 7143 7743 8389 8734 9028 ...
##  $ MarginError: num [1:56] 43 46 51 53 58 61 66 71 79 77 ...
```

```r
# #6 Cumulative percent change in real annual earnings, by earnings group,
# 1979-2018.
tmb_df <- read_excel("data/finalproject/topmidbot_f.xlsx")
# renaming and creating numeric as needed
```

```r
tmb_df$Year <- as.numeric(tmb_df$Year)
names(tmb_df) <- c("Year", "Bot90p", "Top1p", "Top_tenthp")
str(tmb_df)
```

```
## tibble [40 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Year      : num [1:40] 1979 1980 1981 1982 1983 ...
##  $ Bot90p    : num [1:40] 0 -2.2 -2.6 -3.9 -3.7 -1.8 -1 1.1 2.1 2.2 ...
##  $ Top1p     : num [1:40] 0 3.4 3.1 9.5 13.6 20.7 23 32.6 53.5 68.7 ...
##  $ Top_tenthp: num [1:40] 0 5.8 7.3 17.4 28.7 ...
```

```r
# late breaking data - I found a source for minimum wage in 2019 dollars
minw_df <- read_excel("data/finalproject/latebreaker_minwage_f.xlsx")
# creating numeric as needed
names(minw_df) <- c("Min_Wage", "Year")

# creating a dichotomous variable for when less than or equal to 7.25 and
# greater than 7.25 of Min wage based on 2019 dollar valuation - where 7.25 is
# forwarded to 2022 and multiplied by a factor of 1.16.
minw_df$CURRYNN <- as.numeric(minw_df$Min_Wage >= 8.41)
minw_df$CURRYNC <- minw_df$Min_Wage >= 8.41
str(minw_df)
```

```
## tibble [45 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Min_Wage: num [1:45] 8.52 8.64 9.82 9.97 10.47 ...
##  $ Year    : num [1:45] 1968 1970 1972 1976 1979 ...
##  $ CURRYNN : num [1:45] 1 1 1 1 1 1 1 1 1 0 ...
##  $ CURRYNC : logi [1:45] TRUE TRUE TRUE TRUE TRUE TRUE ...
```

```r
# merging the data together for these initial merges - using outer joins
df1 = merge(x = ecopoly_df, y = fred_df, by = "Year", all = TRUE)
df2 = merge(x = df1, y = home_in_df, by = "Year", all = TRUE)
df3 = merge(x = df2, y = infl_df2, by = "Year", all = TRUE)
df4 = merge(x = df3, y = tabd_df, by = "Year", all = TRUE)
df5 = merge(x = df4, y = minw_df, by = "Year")
All_Outer0 = merge(x = df5, y = tmb_df, by = "Year", all = TRUE)
str(All_Outer0)
```

```
## 'data.frame':    47 obs. of  20 variables:
##  $ Year             : num  1968 1970 1972 1976 1979 ...
##  $ Hr_comp          : num  71 76.8 91.3 89.3 93.2 ...
##  $ Net_prod         : num  77.1 80.3 92.2 103.6 108.1 ...
##  $ Average_home_sale: num  26425 26650 30075 48050 71900 ...
##  $ Income_change    : num  NA 0.03 NA 23.47 38.32 ...
##  $ Home_change      : num  NA 2.7 NA 1.69 3.54 -1.89 -4.15 -4.59 -4.22 -1.56 ...
##  $ Inflation_chg    : num  4.7 5.6 3.4 4.9 13.3 12.5 8.9 3.8 3.8 3.9 ...
##  $ FedRate_x5       : num  6 5 5.75 4.75 12 18 12 8.5 9.25 8.25 ...
##  $ BusCyc           : chr  "Expansion (4.9%)" "Nov. trough (0.2%)" "Expansion (5.3%)" "Expansion (5.4
##  $ Events           : chr  "Moon landing" "Recession" "Stagflation" NA ...
##  $ InflatiYN        : num  0 1 0 0 1 1 1 0 0 0 ...
##  $ InflatCYN        : logi  FALSE TRUE FALSE FALSE TRUE TRUE ...
##  $ MedIncome        : num  7743 8734 9697 12686 16461 ...
##  $ MarginError      : num  46 53 61 77 128 150 165 150 157 168 ...
##  $ Min_Wage         : num  8.52 8.64 9.82 9.97 10.47 ...
##  $ CURRYNN          : num  1 1 1 1 1 1 1 1 1 0 ...
##  $ CURRYNC          : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ Bot90p           : num  NA NA NA NA 0 -2.2 -2.6 -3.9 -3.7 -1.8 ...
##  $ Top1p            : num  NA NA NA NA 0 3.4 3.1 9.5 13.6 20.7 ...
##  $ Top_tenthp       : num  NA NA NA NA 0 5.8 7.3 17.4 28.7 44 ...
```

```
All_Outer <- subset(All_Outer0, select = -c(BusCyc, Events, FedRate_x5, MarginError))

str(All_Outer)
```

```
## ’data.frame’:    47 obs. of  16 variables:
##  $ Year             : num  1968 1970 1972 1976 1979 ...
##  $ Hr_comp          : num  71 76.8 91.3 89.3 93.2 ...
##  $ Net_prod         : num  77.1 80.3 92.2 103.6 108.1 ...
##  $ Average_home_sale: num  26425 26650 30075 48050 71900 ...
##  $ Income_change    : num  NA 0.03 NA 23.47 38.32 ...
##  $ Home_change      : num  NA 2.7 NA 1.69 3.54 -1.89 -4.15 -4.59 -4.22 -1.56 ...
##  $ Inflation_chg    : num  4.7 5.6 3.4 4.9 13.3 12.5 8.9 3.8 3.8 3.9 ...
##  $ InflatiYN        : num  0 1 0 0 1 1 1 0 0 0 ...
##  $ InflatCYN        : logi  FALSE TRUE FALSE FALSE TRUE TRUE ...
##  $ MedIncome        : num  7743 8734 9697 12686 16461 ...
##  $ Min_Wage         : num  8.52 8.64 9.82 9.97 10.47 ...
##  $ CURRYNN          : num  1 1 1 1 1 1 1 1 1 0 ...
##  $ CURRYNC          : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ Bot90p           : num  NA NA NA NA 0 -2.2 -2.6 -3.9 -3.7 -1.8 ...
##  $ Top1p            : num  NA NA NA NA 0 3.4 3.1 9.5 13.6 20.7 ...
##  $ Top_tenthp       : num  NA NA NA NA 0 5.8 7.3 17.4 28.7 44 ...
```

```
# for this merges - using inner joins - this is the dataframe I would use to
# run any combined analysis however I am interested in the outer join to better
# see the gaps. This dataframe is not that large so I can look at the gaps and
# some understanding of what is the same across variables or not

df1 = merge(x = ecopoly_df, y = fred_df, by = "Year")
df2 = merge(x = df1, y = home_in_df, by = "Year")
df3 = merge(x = df2, y = infl_df2, by = "Year")
df4 = merge(x = df3, y = tabd_df, by = "Year")
df5 = merge(x = df4, y = minw_df, by = "Year")
All_Inner0 = merge(x = df5, y = tmb_df, by = "Year")
str(All_Inner0)
```

```
## ’data.frame’:    42 obs. of  20 variables:
##  $ Year             : num  1979 1980 1981 1982 1983 ...
##  $ Hr_comp          : num  93.2 88 87.4 87.7 88.5 ...
##  $ Net_prod         : num  108 107 110 108 115 ...
##  $ Average_home_sale: num  71900 76375 83175 83850 89775 ...
##  $ Income_change    : num  38.3 23.6 21.5 14.7 19.7 ...
##  $ Home_change      : num  3.54 -1.89 -4.15 -4.59 -4.22 -1.56 0.2 3.68 4.68 5.05 ...
##  $ Inflation_chg    : num  13.3 12.5 8.9 3.8 3.8 3.9 3.8 1.1 4.4 4.4 ...
##  $ FedRate_x5       : num  12 18 12 8.5 9.25 8.25 7.75 6 6.75 9.75 ...
##  $ BusCyc           : chr  "Expansion (3.2%)" "Jan. peak (-0.3%)" "July trough (2.5%)" "November (-1
##  $ Events           : chr  NA "Recession" "Reagan tax cut" "Recession ended" ...
##  $ InflatiYN        : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ InflatCYN        : logi  TRUE TRUE TRUE FALSE FALSE FALSE ...
##  $ MedIncome        : num  16461 17710 19074 20171 20885 ...
##  $ MarginError      : num  128 150 165 150 157 168 211 212 203 219 ...
##  $ Min_Wage         : num  10.47 9.86 9.59 8.93 8.93 ...
##  $ CURRYNN          : num  1 1 1 1 1 0 0 0 0 0 ...
##  $ CURRYNC          : logi  TRUE TRUE TRUE TRUE TRUE FALSE ...
##  $ Bot90p           : num  0 -2.2 -2.6 -3.9 -3.7 -1.8 -1 1.1 2.1 2.2 ...
##  $ Top1p            : num  0 3.4 3.1 9.5 13.6 20.7 23 32.6 53.5 68.7 ...
##  $ Top_tenthp       : num  0 5.8 7.3 17.4 28.7 ...
```

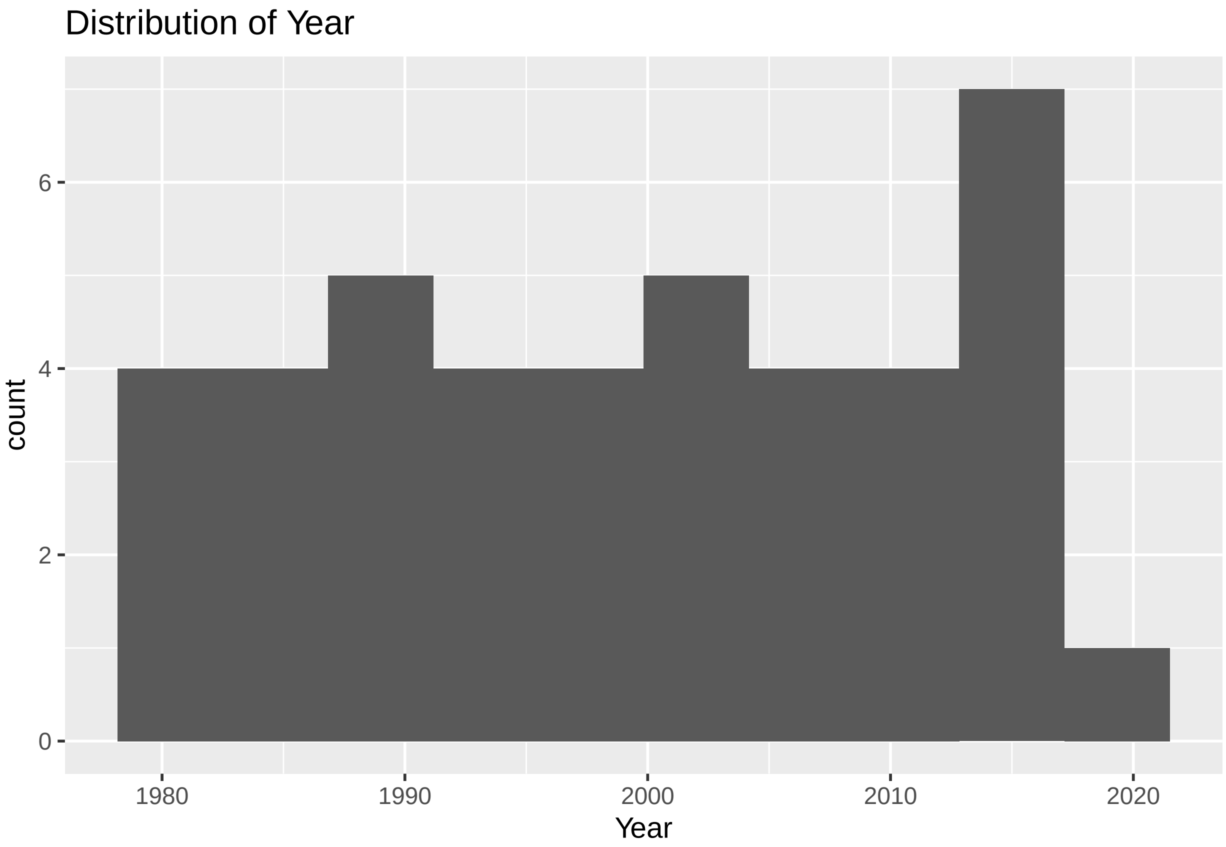
```
# removing variables Events & BusCyc at this stage, minus c

All_Inner <- subset(All_Inner0, select = -c(BusCyc, Events, FedRate_x5, MarginError))
str(All_Inner)
```

```
## 'data.frame':    42 obs. of  16 variables:
##  $ Year             : num  1979 1980 1981 1982 1983 ...
##  $ Hr_comp          : num  93.2 88 87.4 87.7 88.5 ...
##  $ Net_prod         : num  108 107 110 108 115 ...
##  $ Average_home_sale: num  71900 76375 83175 83850 89775 ...
##  $ Income_change    : num  38.3 23.6 21.5 14.7 19.7 ...
##  $ Home_change      : num  3.54 -1.89 -4.15 -4.59 -4.22 -1.56 0.2 3.68 4.68 5.05 ...
##  $ Inflation_chg    : num  13.3 12.5 8.9 3.8 3.8 3.9 3.8 1.1 4.4 4.4 ...
##  $ InflatiYN        : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ InflatCYN        : logi  TRUE TRUE TRUE FALSE FALSE FALSE ...
##  $ MedIncome        : num  16461 17710 19074 20171 20885 ...
##  $ Min_Wage         : num  10.47 9.86 9.59 8.93 8.93 ...
##  $ CURRYNN          : num  1 1 1 1 1 0 0 0 0 0 ...
##  $ CURRYNC          : logi  TRUE TRUE TRUE TRUE TRUE FALSE ...
##  $ Bot90p           : num  0 -2.2 -2.6 -3.9 -3.7 -1.8 -1 1.1 2.1 2.2 ...
##  $ Top1p            : num  0 3.4 3.1 9.5 13.6 20.7 23 32.6 53.5 68.7 ...
##  $ Top_tenthp       : num  0 5.8 7.3 17.4 28.7 ...
```

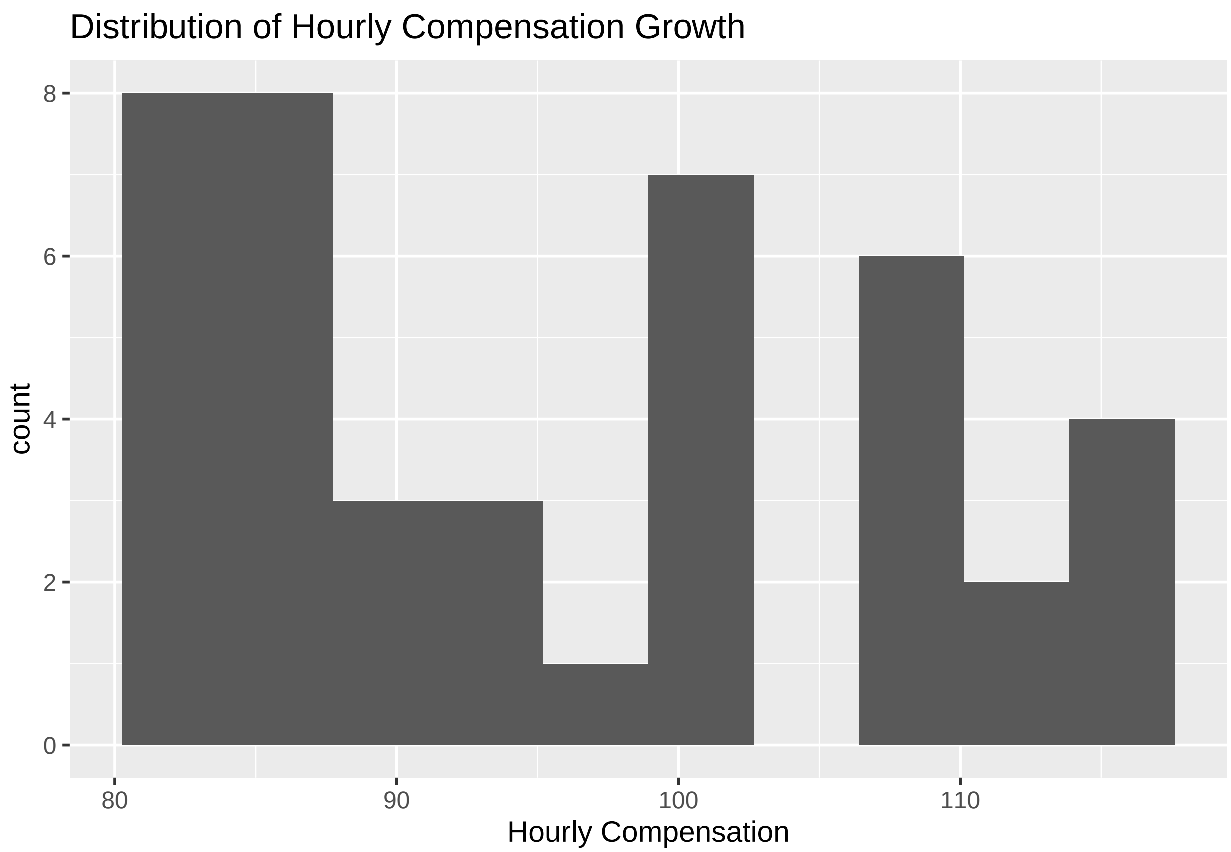
```
########################## Final Project Part 2 questions start #########

# QUESTION 1, showing final dataframe in concise format
All_Inner %>%
    slice_head(n = 5) %>%
    print.data.frame
```

```
##   Year Hr_comp Net_prod Average_home_sale Income_change Home_change
## 1 1979   93.25   108.11             71900         38.32        3.54
## 2 1980   88.05   106.77             76375         23.56       -1.89
## 3 1981   87.36   110.50             83175         21.53       -4.15
## 4 1982   87.70   108.37             83850         14.72       -4.59
## 5 1983   88.49   114.51             89775         19.71       -4.22
##   Inflation_chg InflatiYN InflatCYN MedIncome Min_Wage CURRYNN CURRYNC Bot90p
## 1          13.3         1      TRUE     16461    10.47       1    TRUE    0.0
## 2          12.5         1      TRUE     17710     9.86       1    TRUE   -2.2
## 3           8.9         1      TRUE     19074     9.59       1    TRUE   -2.6
## 4           3.8         0     FALSE     20171     8.93       1    TRUE   -3.9
## 5           3.8         0     FALSE     20885     8.93       1    TRUE   -3.7
##   Top1p Top_tenthp
## 1   0.0        0.0
## 2   3.4        5.8
## 3   3.1        7.3
## 4   9.5       17.4
## 5  13.6       28.7
```

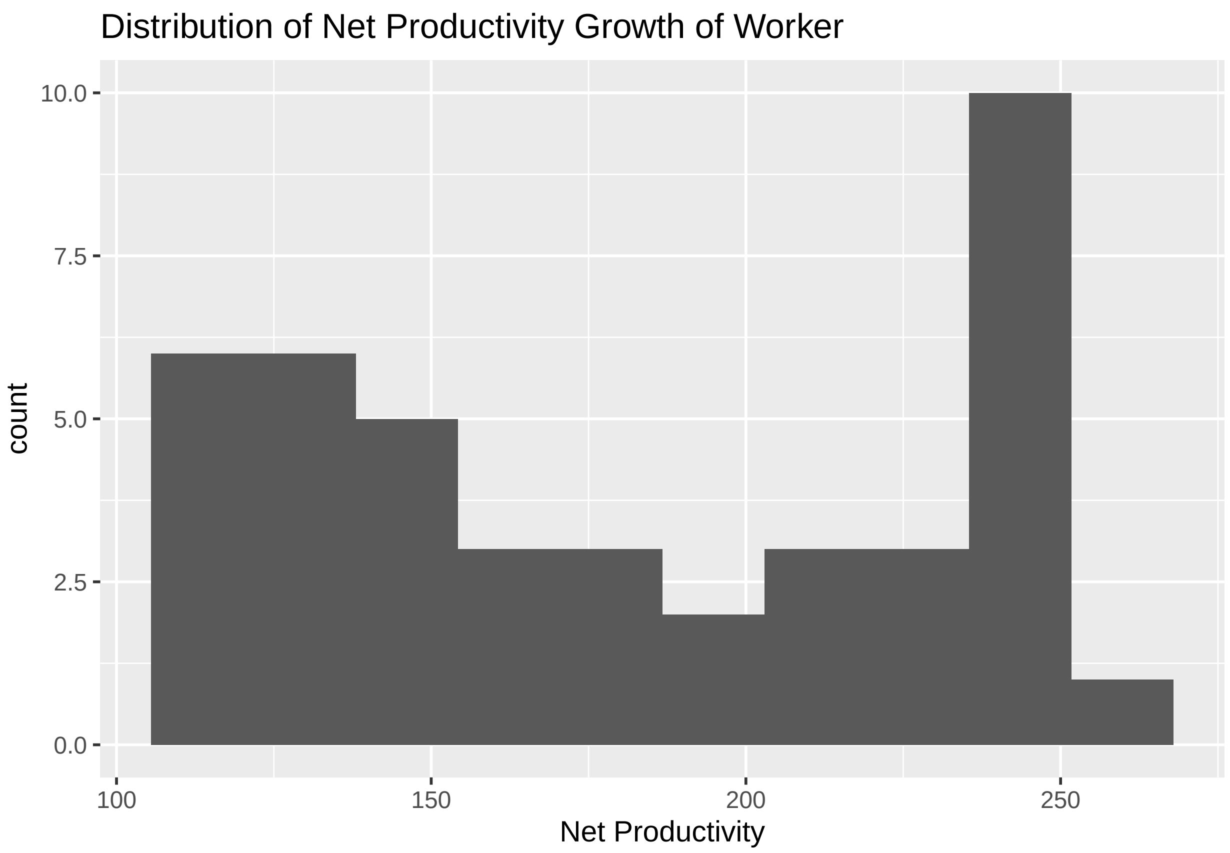
```
ggplot(All_Inner, aes(Year)) + geom_histogram(bins = 10) + ggtitle("Distribution of Year") +
    xlab("Year")
```



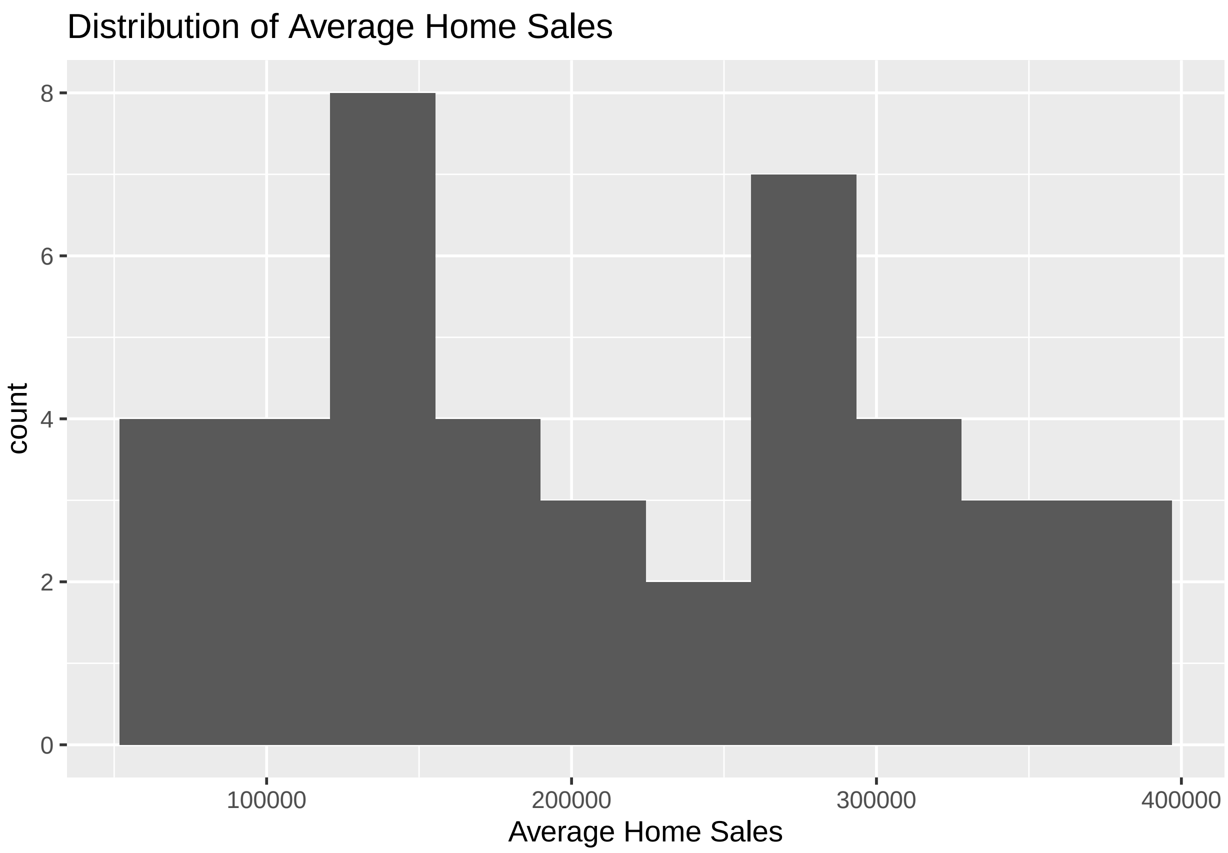
```
ggplot(All_Inner, aes(Hr_comp)) + geom_histogram(bins = 10) + ggtitle("Distribution of Hourly Compensati
    xlab("Hourly Compensation")
```



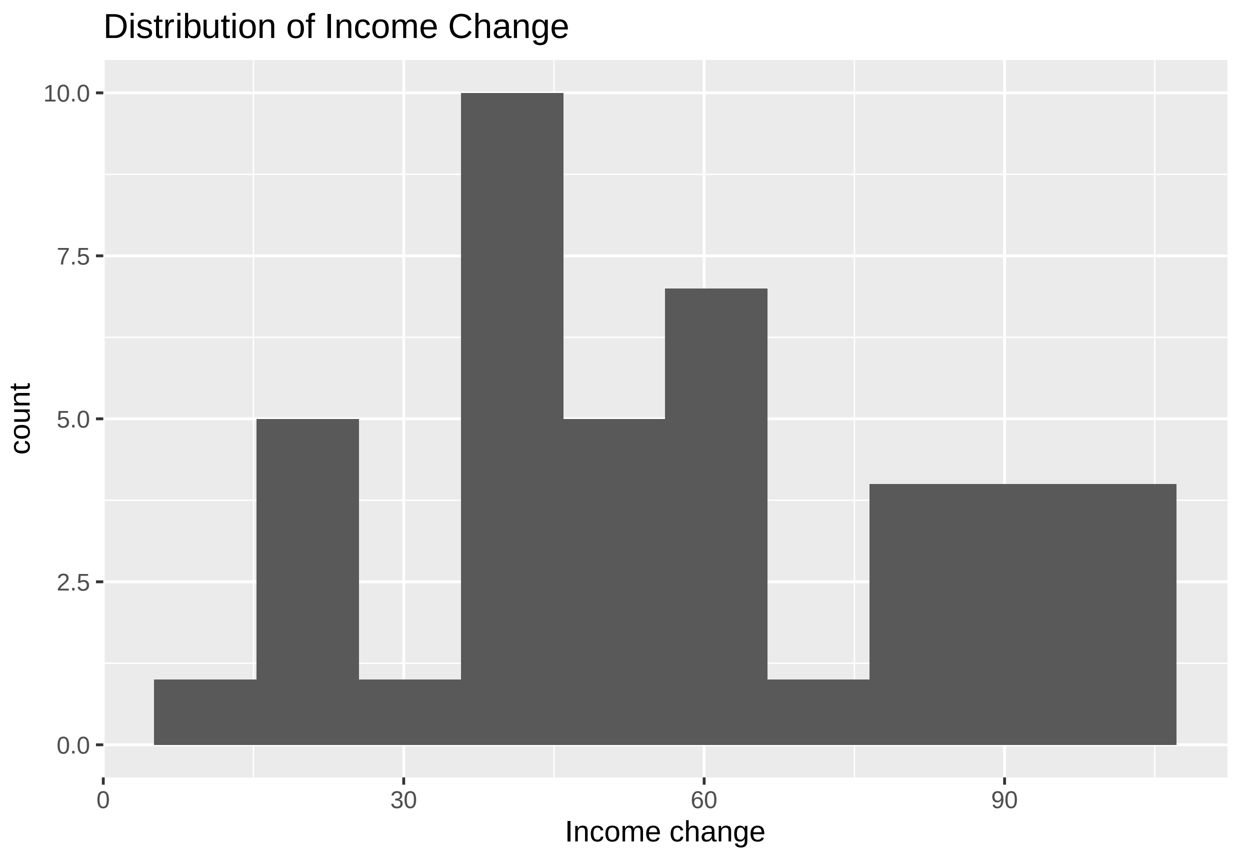
```
ggplot(All_Inner, aes(Net_prod)) + geom_histogram(bins = 10) + ggtitle("Distribution of Net Productivity
    xlab("Net Productivity")
```



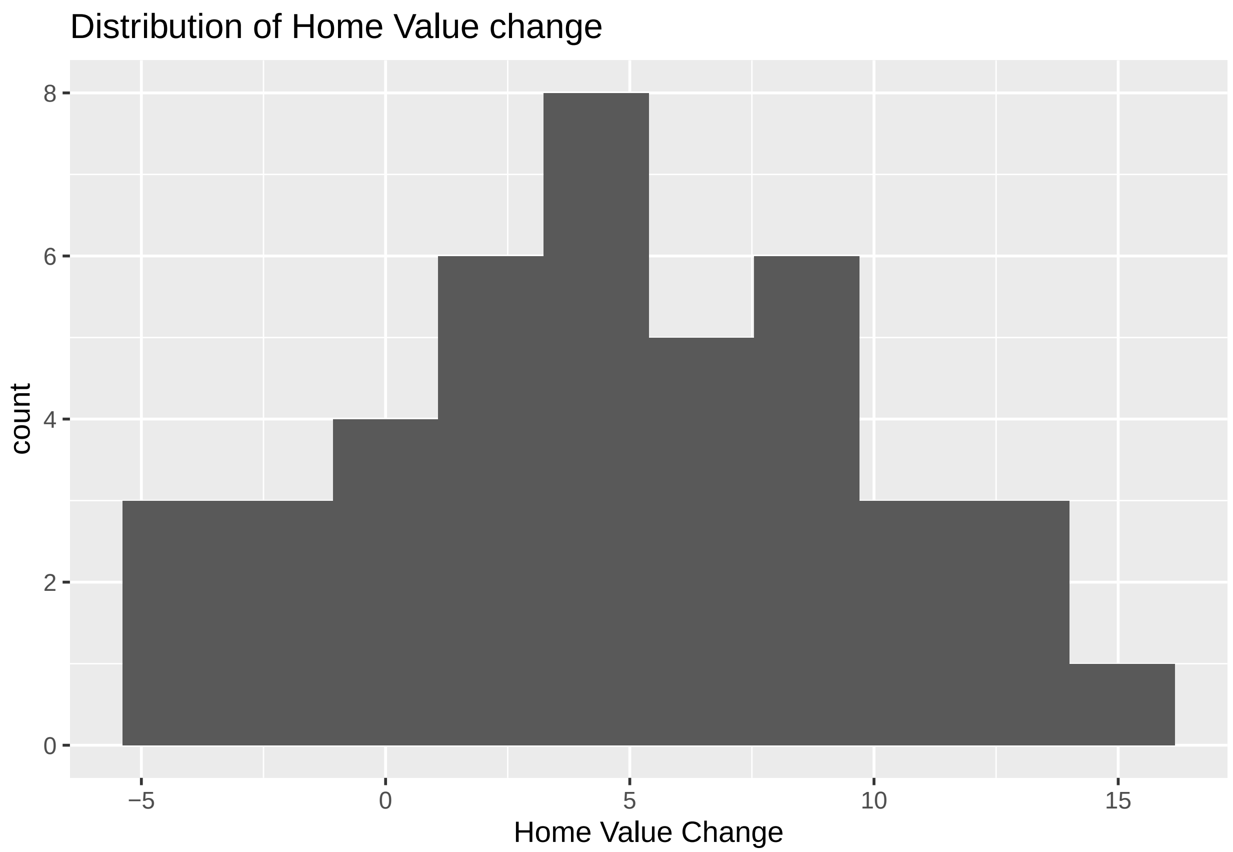
```
ggplot(All_Inner, aes(Average_home_sale)) + geom_histogram(bins = 10) + ggtitle("Distribution of Averag
    xlab("Average Home Sales")
```



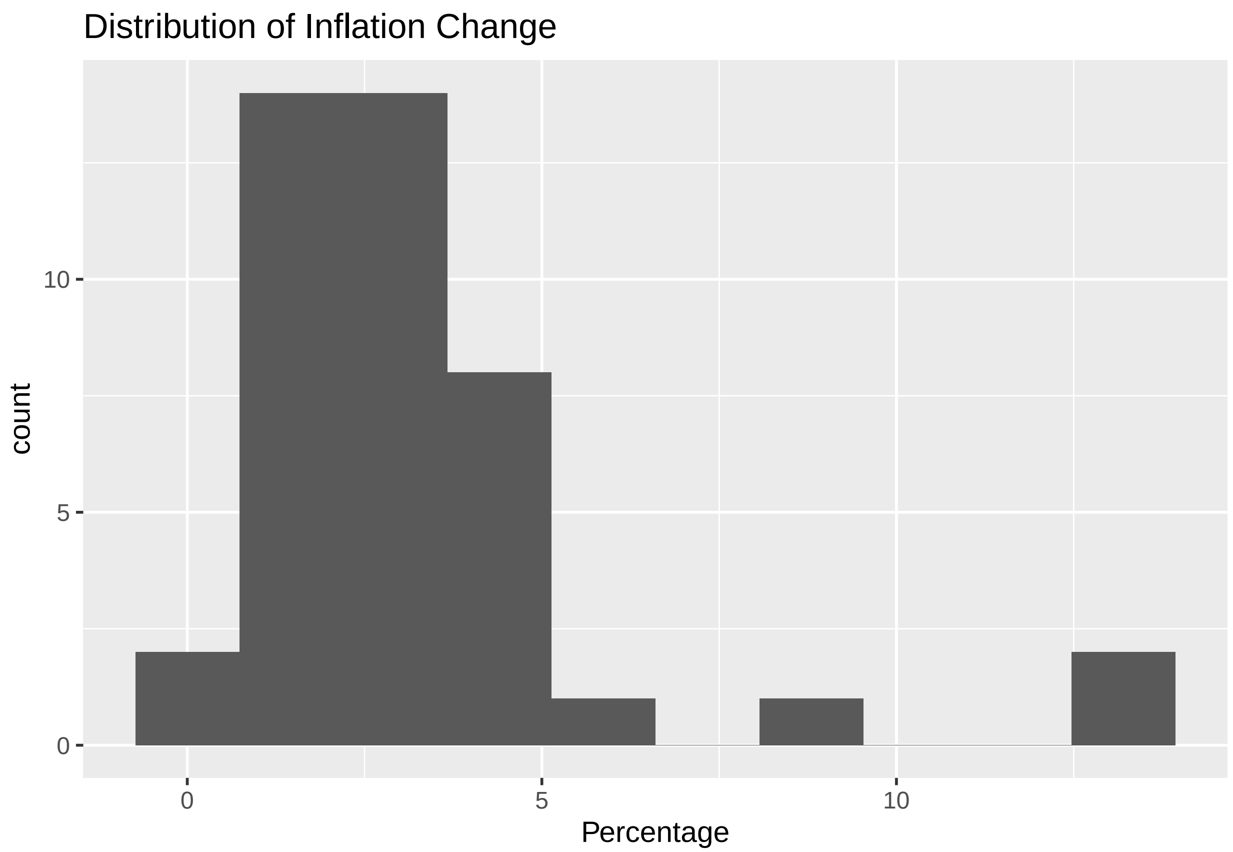
```
ggplot(All_Inner, aes(Income_change)) + geom_histogram(bins = 10) + ggtitle("Distribution of Income Cha
    xlab("Income change")
```



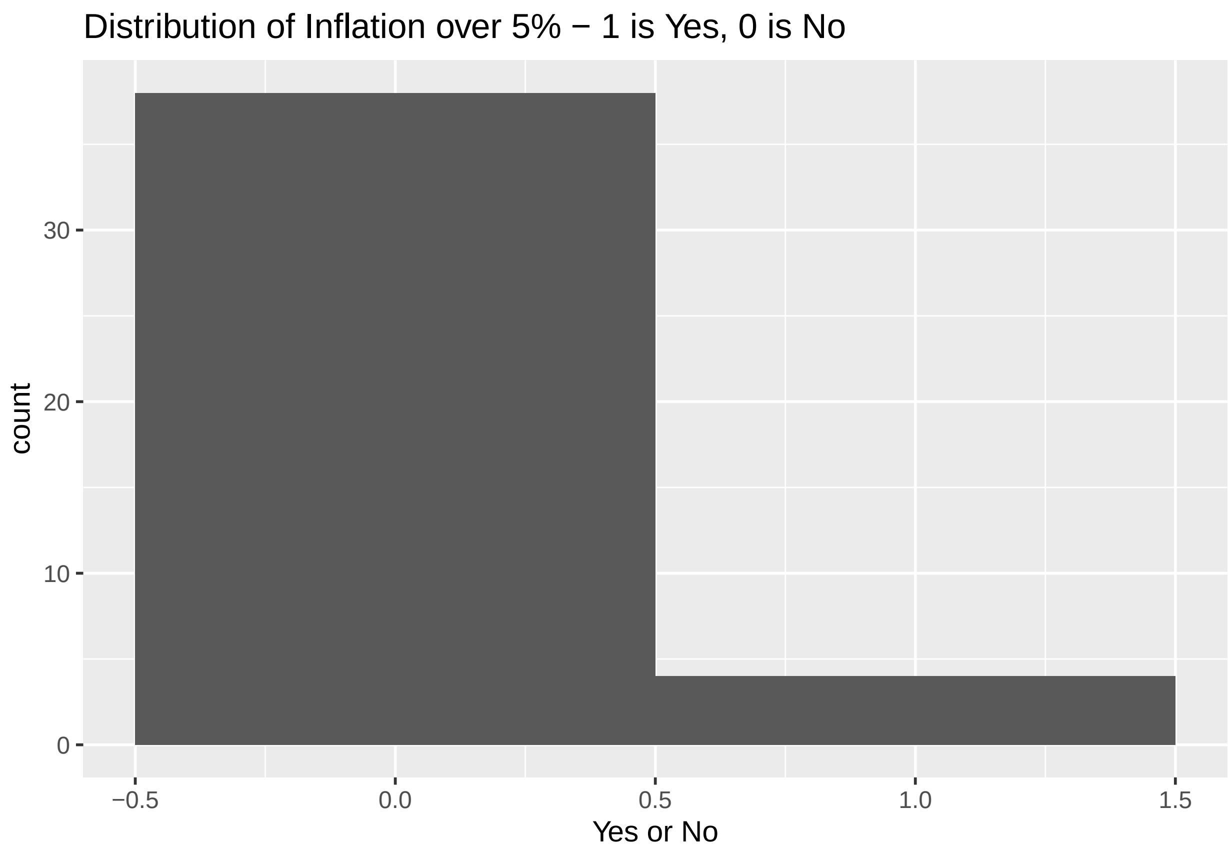
```
ggplot(All_Inner, aes(Home_change)) + geom_histogram(bins = 10) + ggtitle("Distribution of Home Value ch
    xlab("Home Value Change")
```



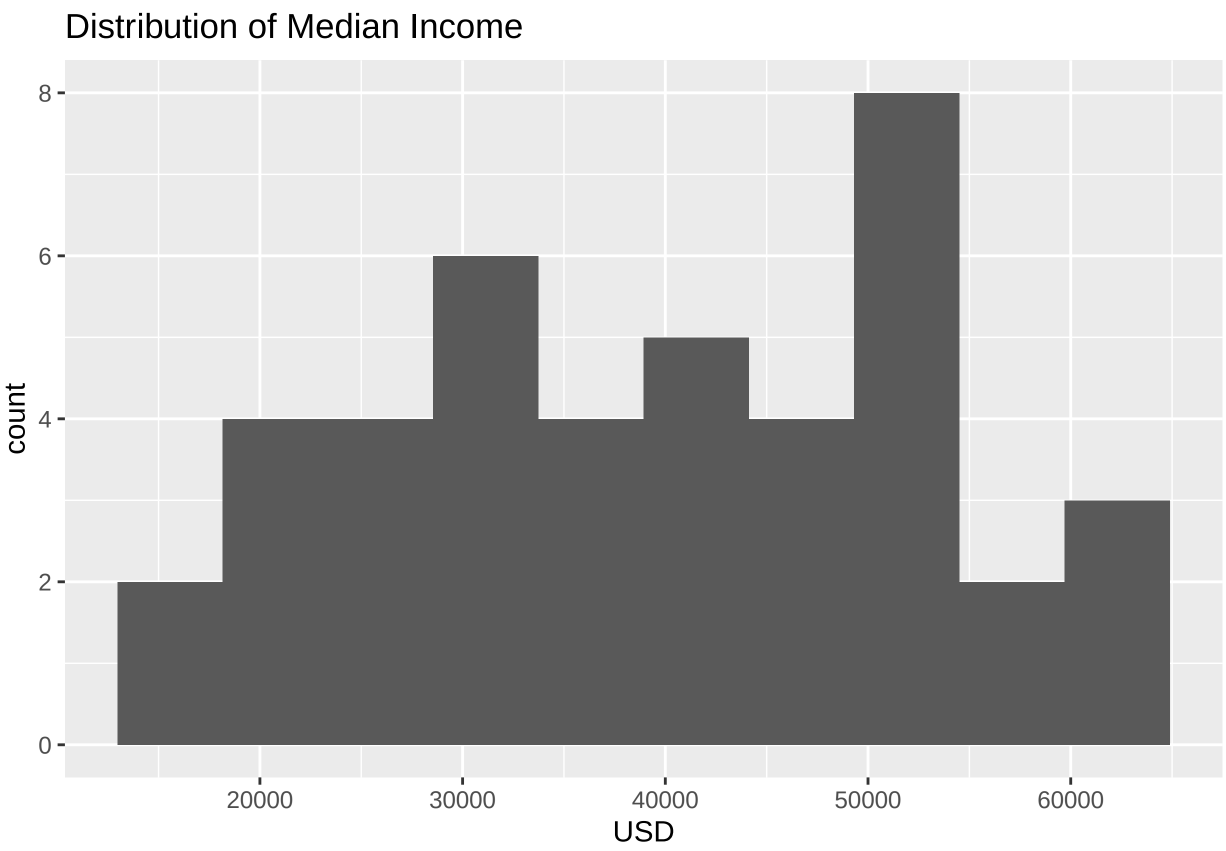
```
ggplot(All_Inner, aes(Inflation_chg)) + geom_histogram(bins = 10) + ggtitle("Distribution of Inflation
    xlab("Percentage")
```



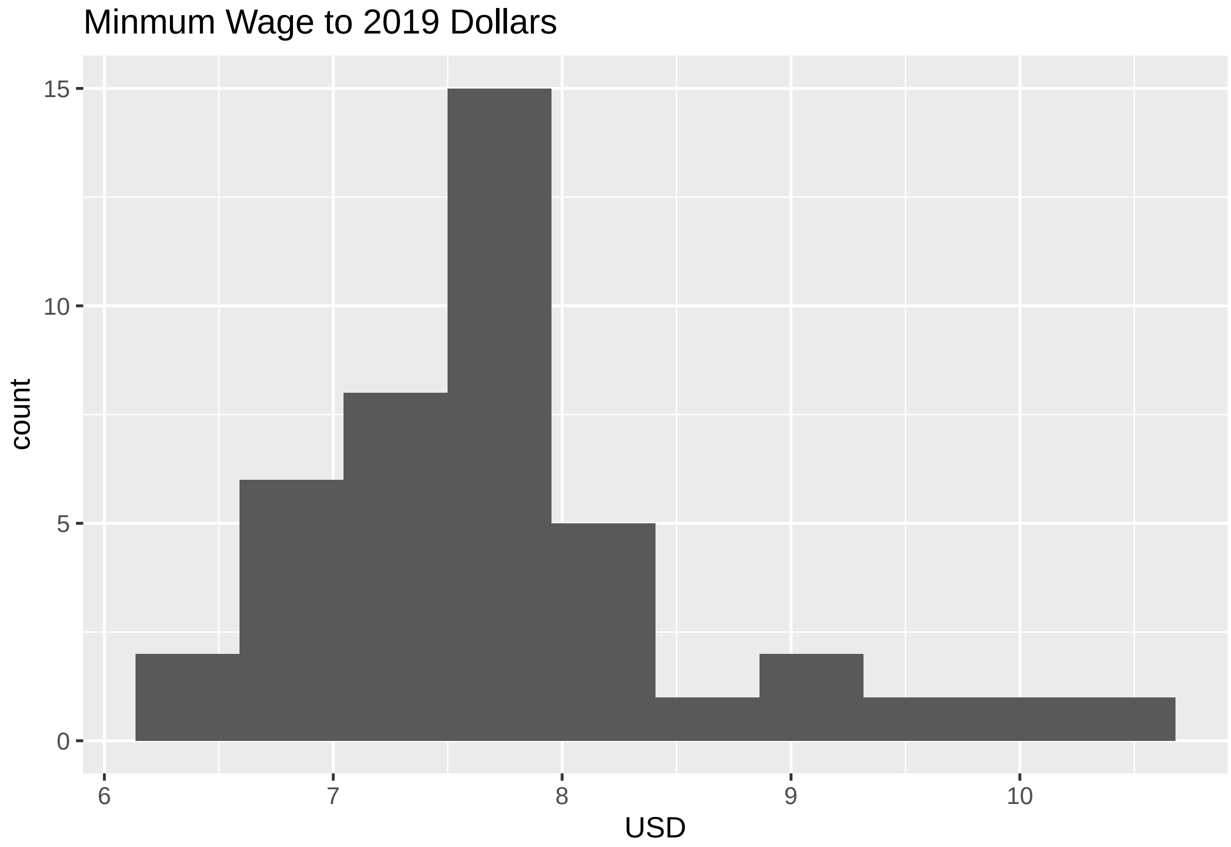
```
ggplot(All_Inner, aes(InflatiYN)) + geom_histogram(bins = 2) + ggtitle("Distribution of Inflation over 5
    xlab("Yes or No")
```



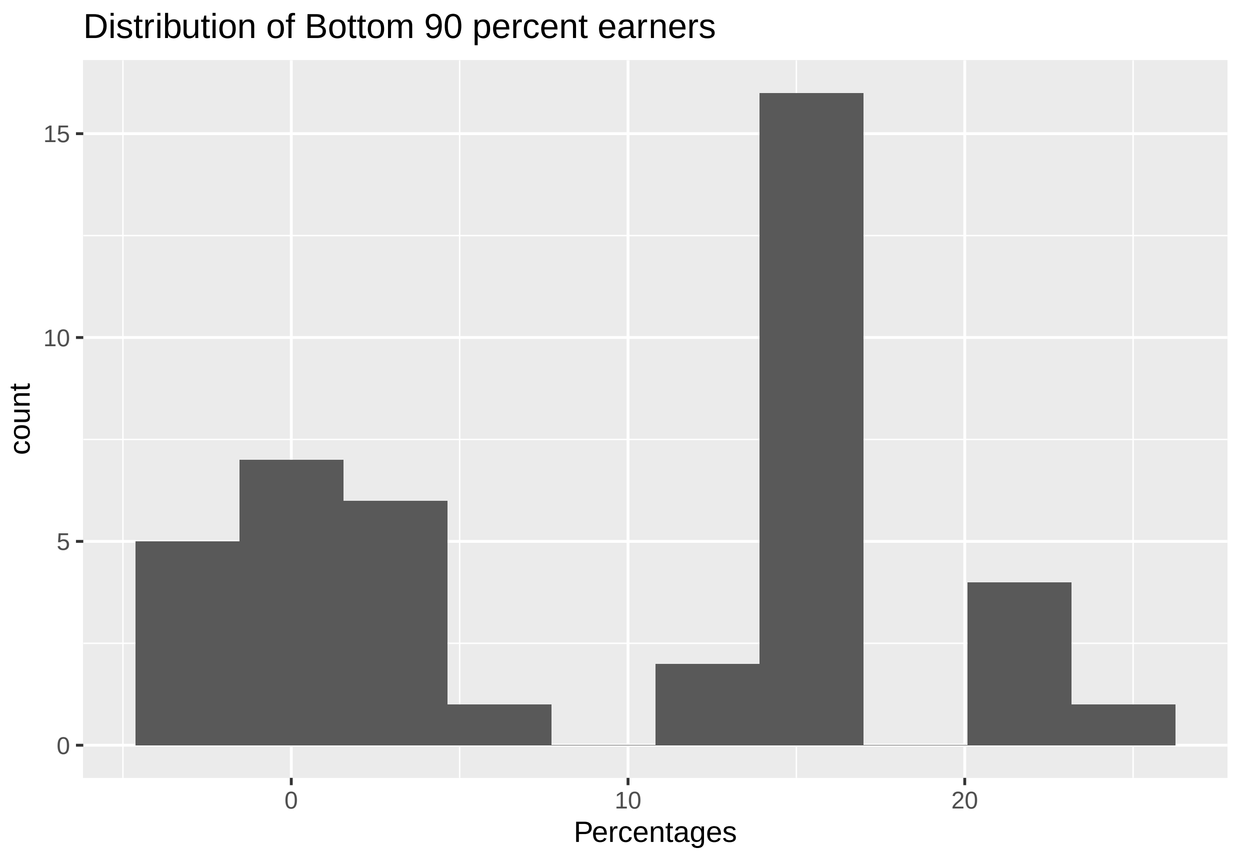
```
ggplot(All_Inner, aes(MedIncome)) + geom_histogram(bins = 10) + ggtitle("Distribution of Median Income"
    xlab("USD")
```



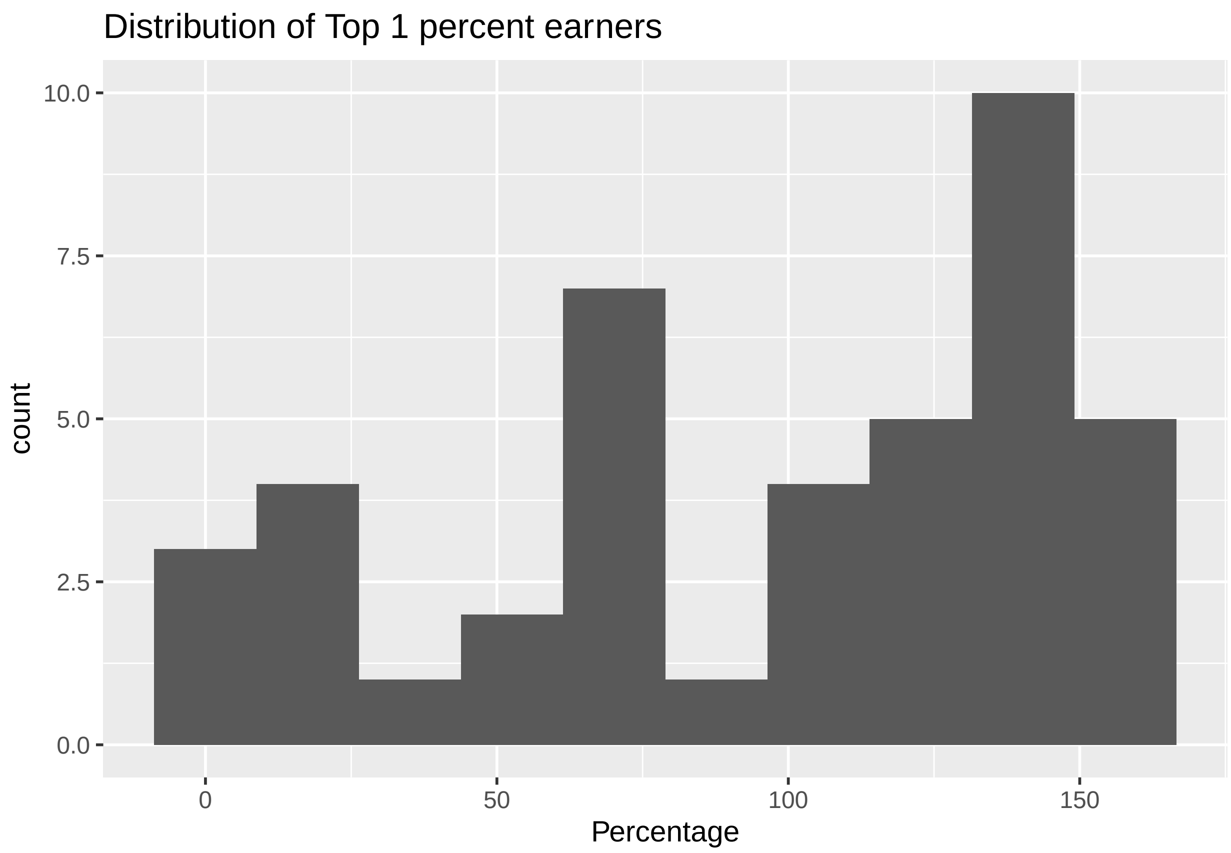
```
ggplot(All_Inner, aes(Min_Wage)) + geom_histogram(bins = 10) + ggtitle("Minmum Wage to 2019 Dollars") +
    xlab("USD")
```



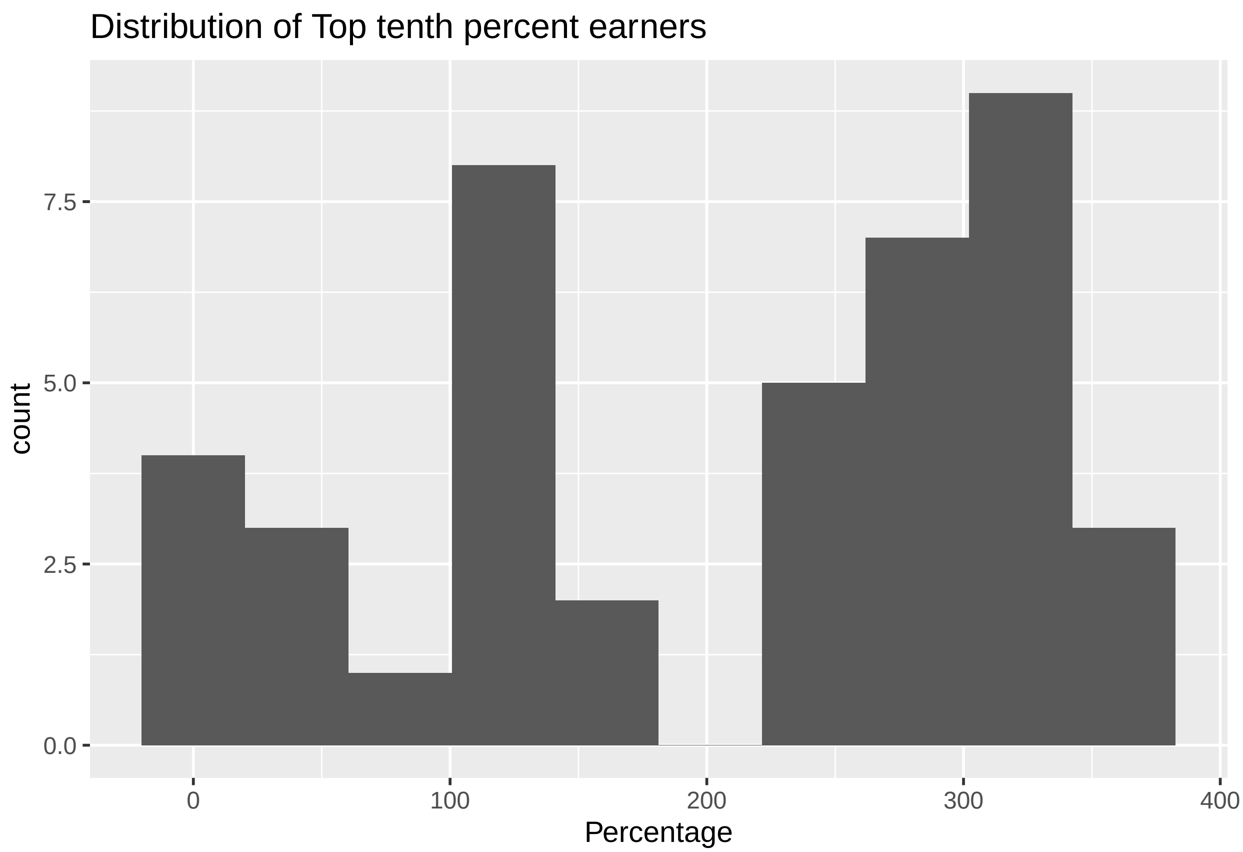
```
ggplot(All_Inner, aes(Bot90p)) + geom_histogram(bins = 10) + ggtitle("Distribution of Bottom 90 percent
    xlab("Percentages")
```



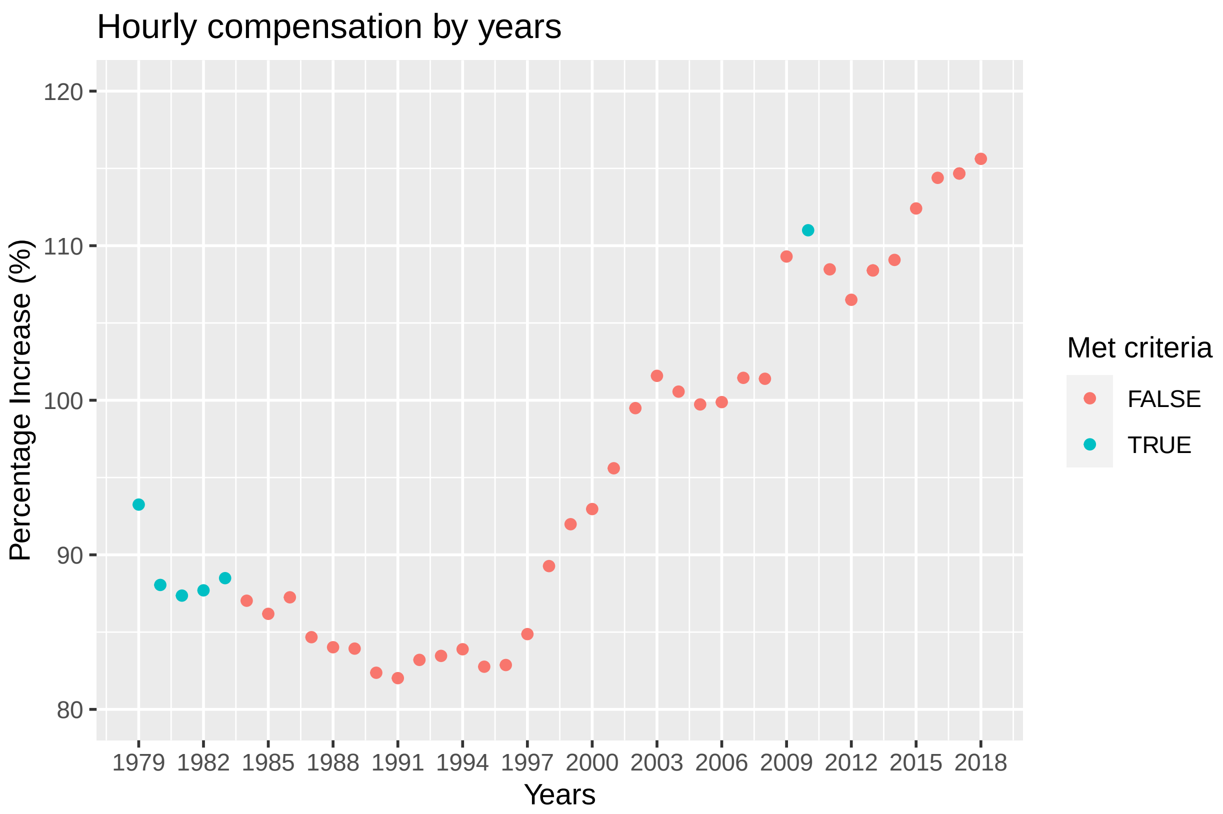
```
ggplot(All_Inner, aes(Top1p)) + geom_histogram(bins = 10) + ggtitle("Distribution of Top 1 percent earne
    xlab("Percentage")
```



```
ggplot(All_Inner, aes(Top_tenthp)) + geom_histogram(bins = 10) + ggtitle("Distribution of Top tenth per
    xlab("Percentage")
```

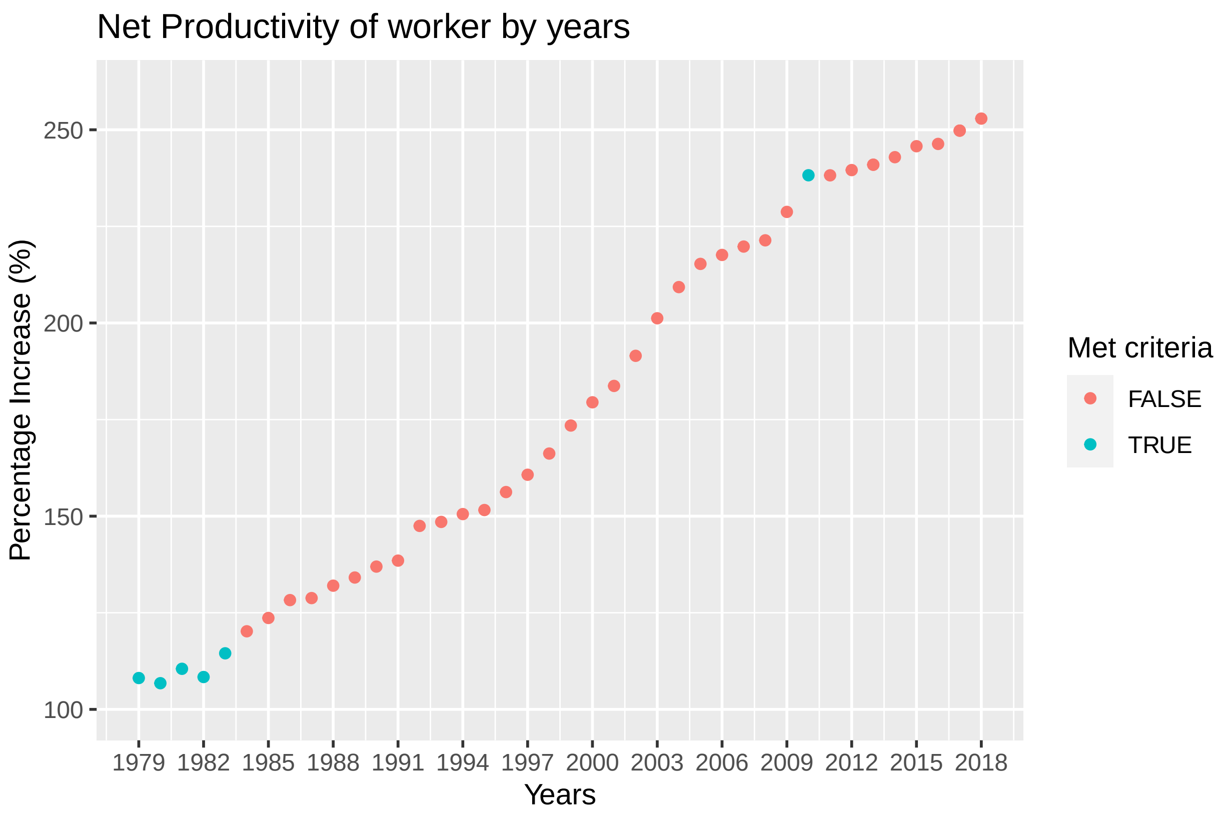


```
## EDA year x other variables

plot1 <- ggplot(All_Inner, aes(x = Year)) + geom_point(aes(y = Hr_comp, col = CURRYNC)) +
    ggtitle("Hourly compensation by years") + labs(caption = "False = Less than $7.25 in 2022 dollars, '
    color = "Met criteria")

# Change x and y axis labels, and limits
plot1 + scale_x_continuous(name = "Years", breaks = seq(1979, 2018, 3)) + scale_y_continuous(name = "Pe
    limits = c(80, 120))
```



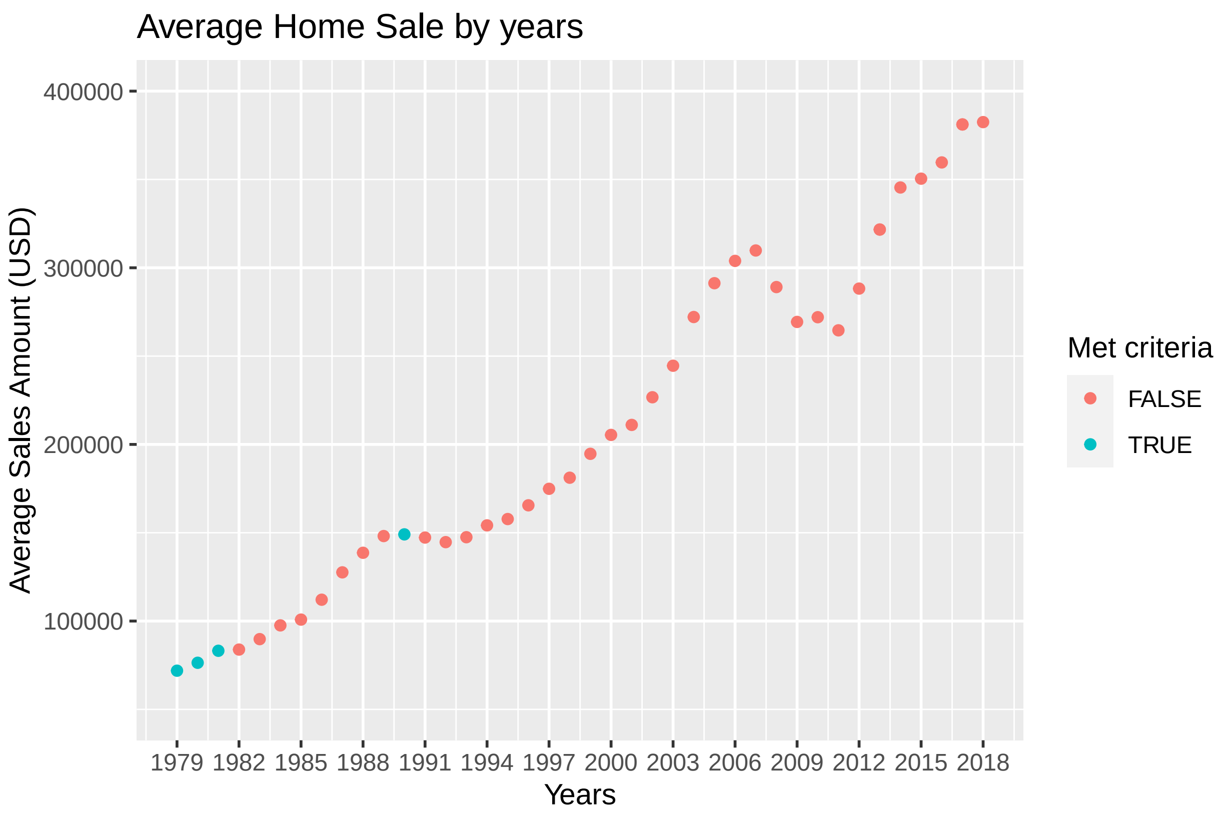
False = Less than $7.25 in 2022 dollars, True = Higher or equal to than $7.25 in 2022 dollars

```
plot2 <- ggplot(All_Inner, aes(x = Year)) + geom_point(aes(y = Net_prod, col = CURRYNC)) +
    ggtitle("Net Productivity of worker by years") + labs(caption = "False = Less than $7.25 in 2022 dol
    color = "Met criteria")

# Change x and y axis labels, and limits
plot2 + scale_x_continuous(name = "Years", breaks = seq(1979, 2018, 3)) + scale_y_continuous(name = "Pe
    limits = c(100, 260))
```



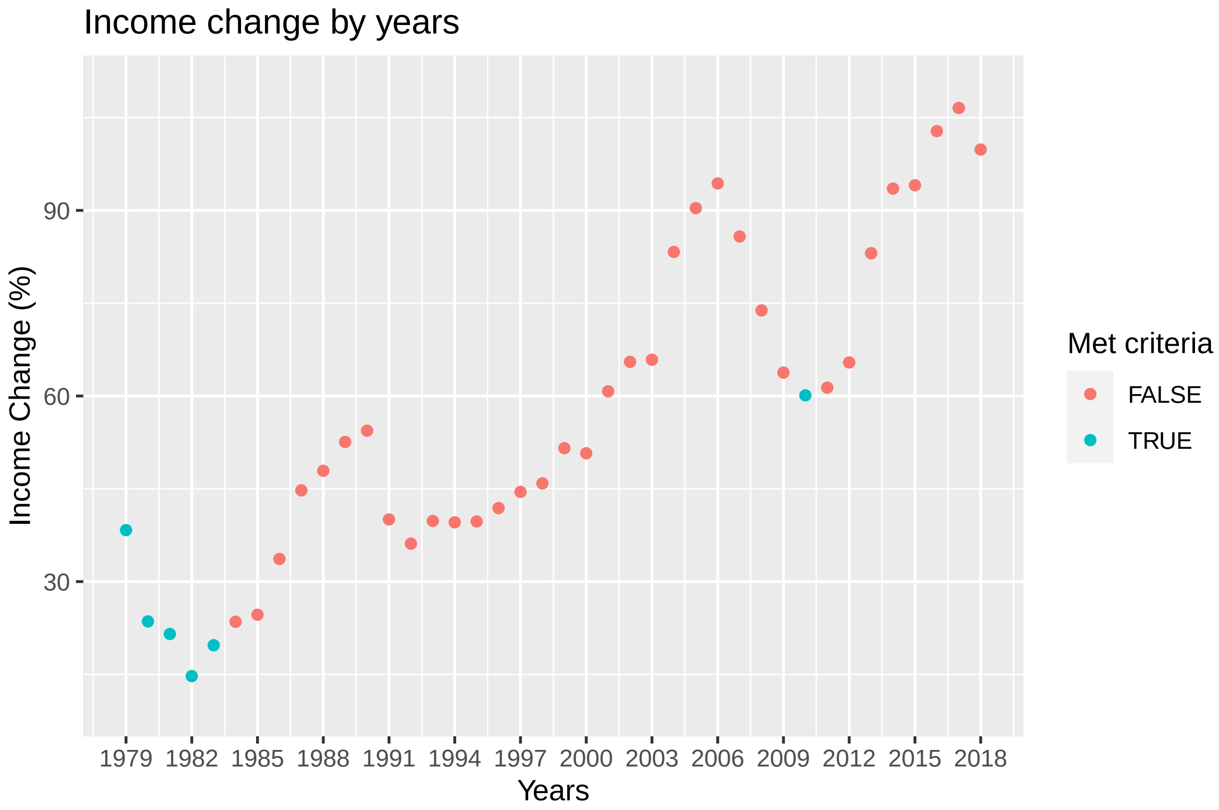
False = Less than $7.25 in 2022 dollars, True = Higher or equal to than $7.25 in 2022 dollars

```
plot3 <- ggplot(All_Inner, aes(x = Year)) + geom_point(aes(y = Average_home_sale,
    col = InflatCYN)) + ggtitle("Average Home Sale by years") + labs(caption = "False = Inflation chang
    color = "Met criteria")

# Change x and y axis labels, and limits
plot3 + scale_x_continuous(name = "Years", breaks = seq(1979, 2018, 3)) + scale_y_continuous(name = "Av
    limits = c(50000, 4e+05))
```



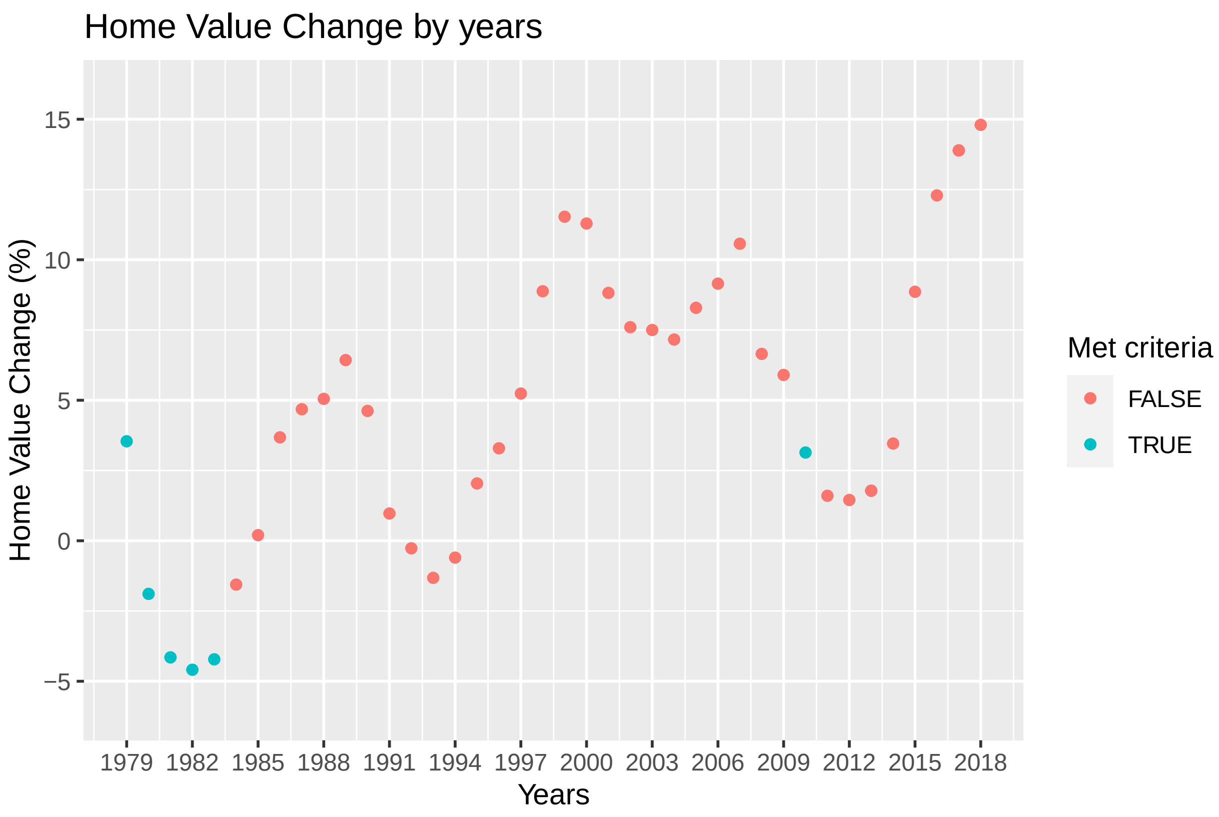
False = Inflation change less than 5%, True = Inflation change greater or equal to 5%

```
plot4 <- ggplot(All_Inner, aes(x = Year)) + geom_point(aes(y = Income_change, col = CURRYNC)) +
    ggtitle("Income change by years") + labs(caption = "False = Less than $7.25 in 2022 dollars, True =
    color = "Met criteria")

# Change x and y axis labels, and limits
plot4 + scale_x_continuous(name = "Years", breaks = seq(1979, 2018, 3)) + scale_y_continuous(name = "Inc
    limits = c(10, 110))
```



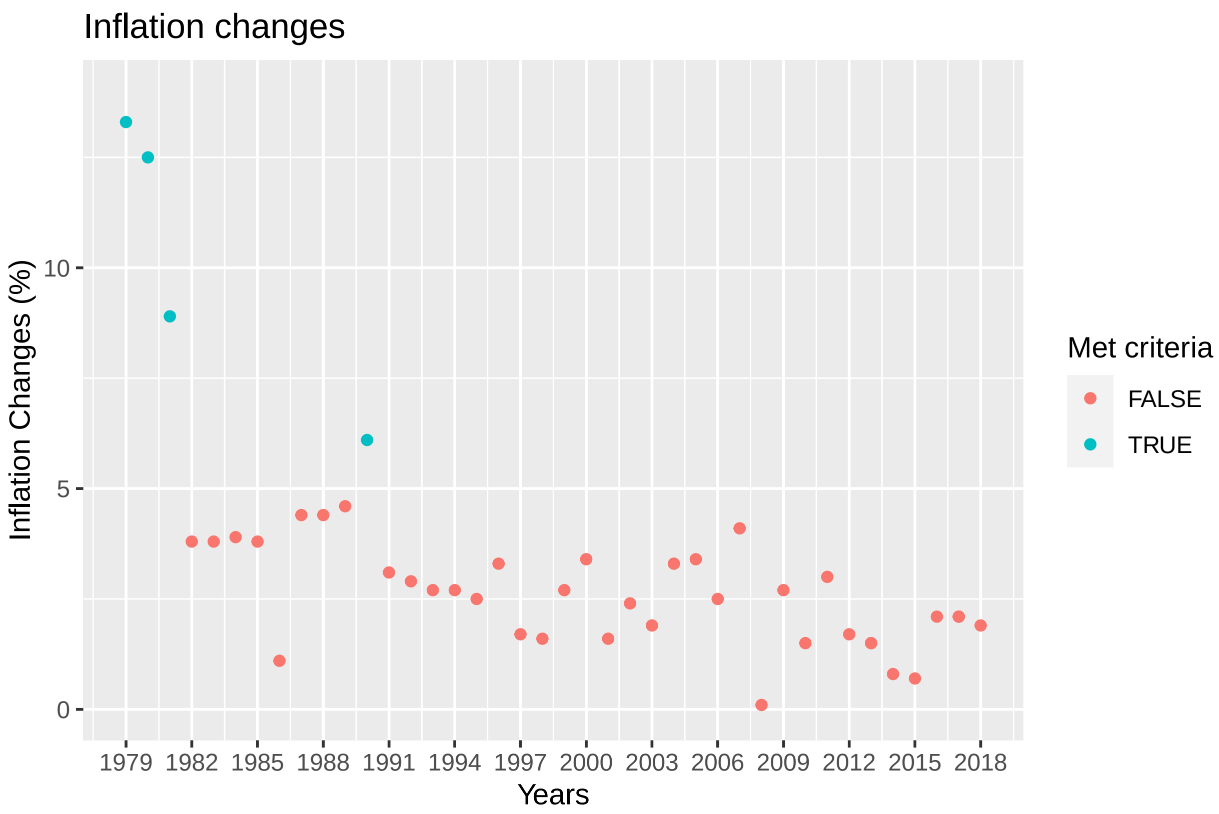
False = Less than $7.25 in 2022 dollars, True = Higher or equal to than $7.25 in 2022 dollars

```
plot5 <- ggplot(All_Inner, aes(x = Year)) + geom_point(aes(y = Home_change, col = CURRYNC)) +
    ggtitle("Home Value Change by years") + labs(caption = "False = Less than $7.25 in 2022 dollars, Tr
    color = "Met criteria")

# Change x and y axis labels, and limits
plot5 + scale_x_continuous(name = "Years", breaks = seq(1979, 2018, 3)) + scale_y_continuous(name = "Hom
    limits = c(-6, 16))
```



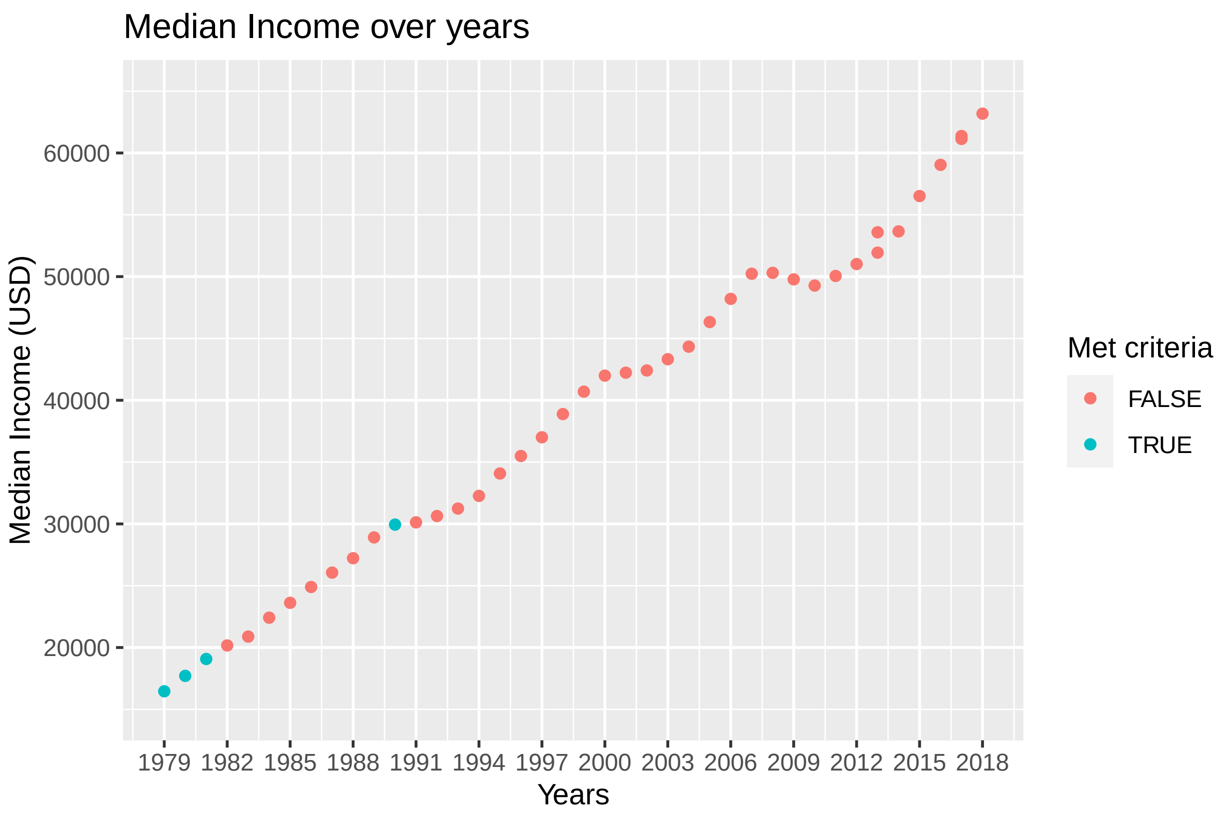
False = Less than $7.25 in 2022 dollars, True = Higher or equal to than $7.25 in 2022 dollars

```
plot6 <- ggplot(All_Inner, aes(x = Year)) + geom_point(aes(y = Inflation_chg, col = InflatCYN)) +
    ggtitle("Inflation changes") + labs(caption = "False = Inflation change less than 5%, True = Inflat
    color = "Met criteria")

# Change x and y axis labels, and limits
plot6 + scale_x_continuous(name = "Years", breaks = seq(1979, 2018, 3)) + scale_y_continuous(name = "In
    limits = c(0, 14))
```



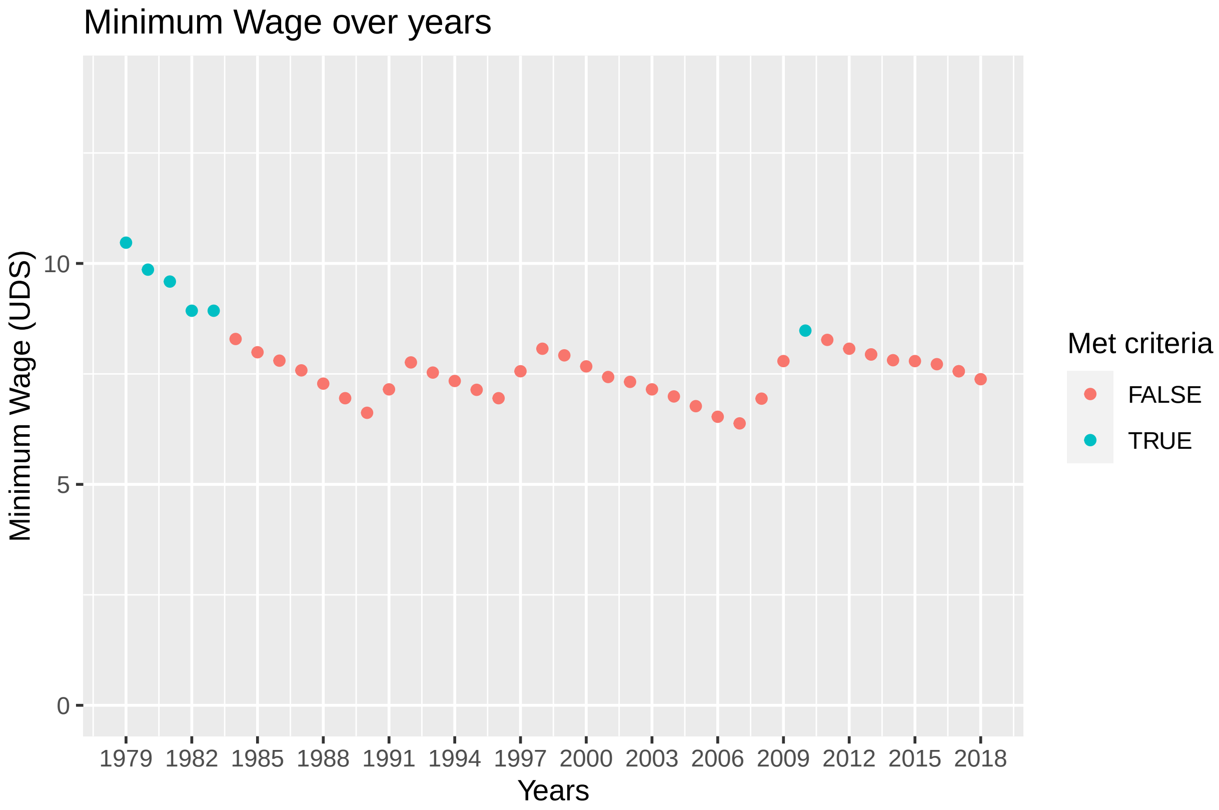
False = Inflation change less than 5%, True = Inflation change greater or equal to 5%

```
plot7 <- ggplot(All_Inner, aes(x = Year)) + geom_point(aes(y = MedIncome, col = InflatCYN)) +
    ggtitle("Median Income over years") + labs(caption = "False = Inflation change less than 5%, True =
    color = "Met criteria")

# Change x and y axis labels, and limits
plot7 + scale_x_continuous(name = "Years", breaks = seq(1979, 2018, 3)) + scale_y_continuous(name = "Med
    limits = c(15000, 65000))
```



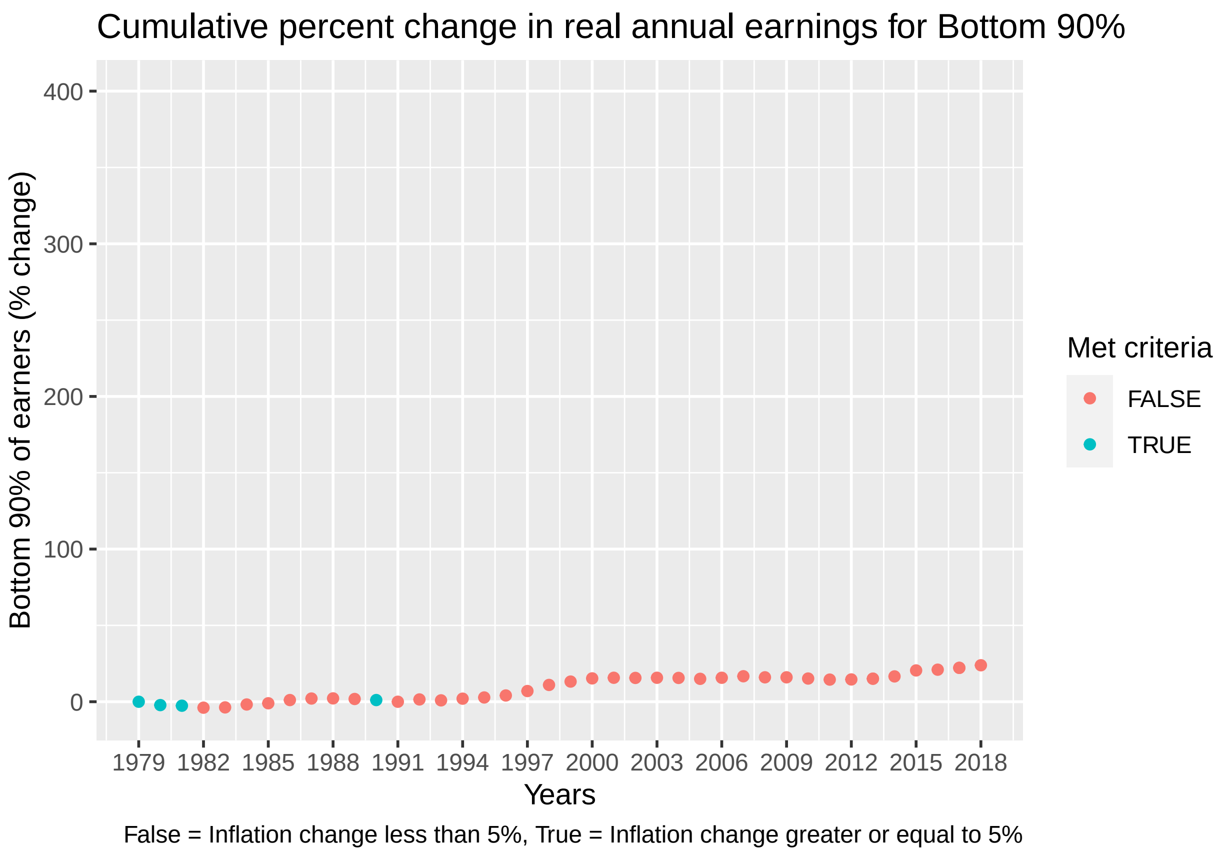
False = Inflation change less than 5%, True = Inflation change greater or equal to 5%

```
plot8 <- ggplot(All_Inner, aes(x = Year)) + geom_point(aes(y = Min_Wage, col = CURRYNC)) +
    ggtitle("Minimum Wage over years") + labs(caption = "False = Less than $7.25 in 2022 dollars, True =
    color = "Met criteria")

# Change x and y axis labels, and limits
plot8 + scale_x_continuous(name = "Years", breaks = seq(1979, 2018, 3)) + scale_y_continuous(name = "Mi
    limits = c(0, 14))
```

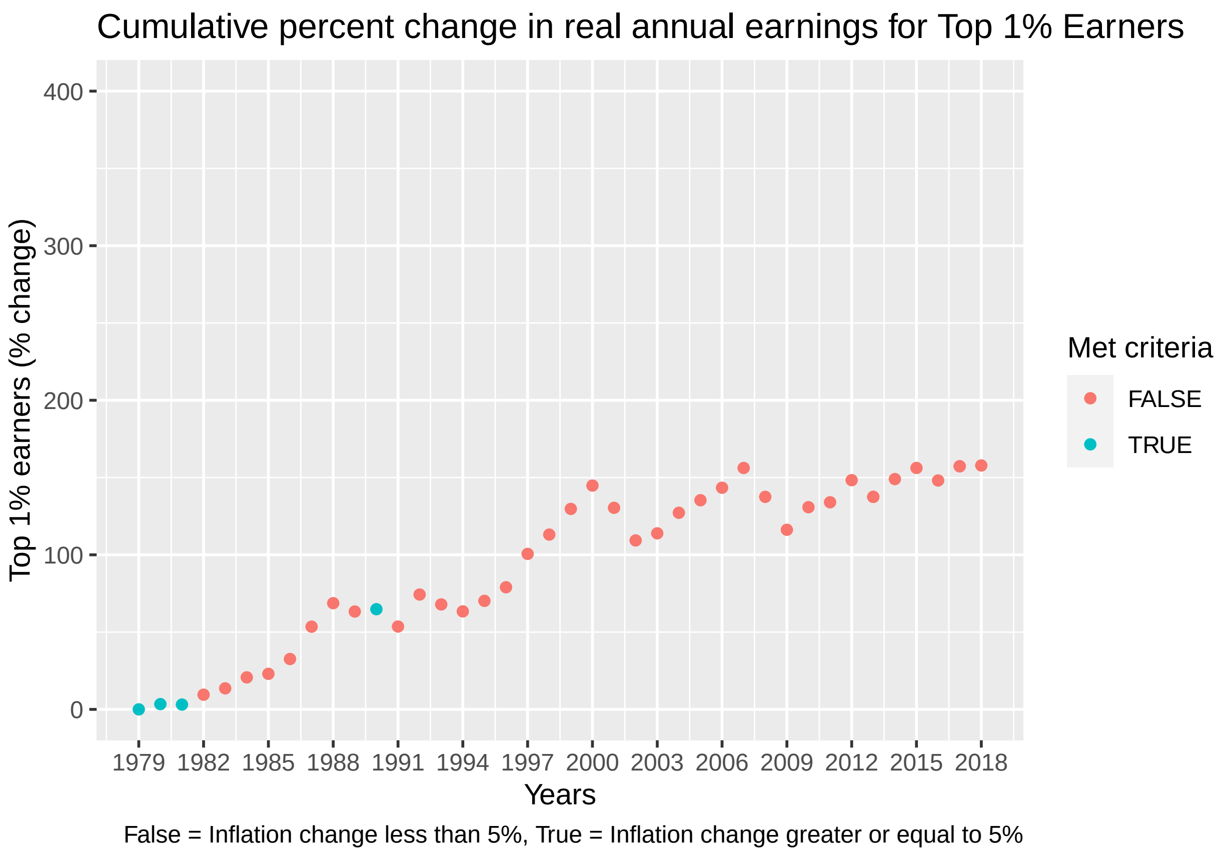


False = Less than $7.25 in 2022 dollars, True = Higher or equal to than $7.25 in 2022 dollars

```
plot9 <- ggplot(All_Inner, aes(x = Year)) + geom_point(aes(y = Bot90p, col = InflatCYN)) +
    ggtitle("Cumulative percent change in real annual earnings for Bottom 90%") +
    labs(caption = "False = Inflation change less than 5%, True = Inflation change greater or equal to 
        color = "Met criteria")

# Change x and y axis labels, and limits
plot9 + scale_x_continuous(name = "Years", breaks = seq(1979, 2018, 3)) + scale_y_continuous(name = "Bo
    limits = c(-5, 400))
```

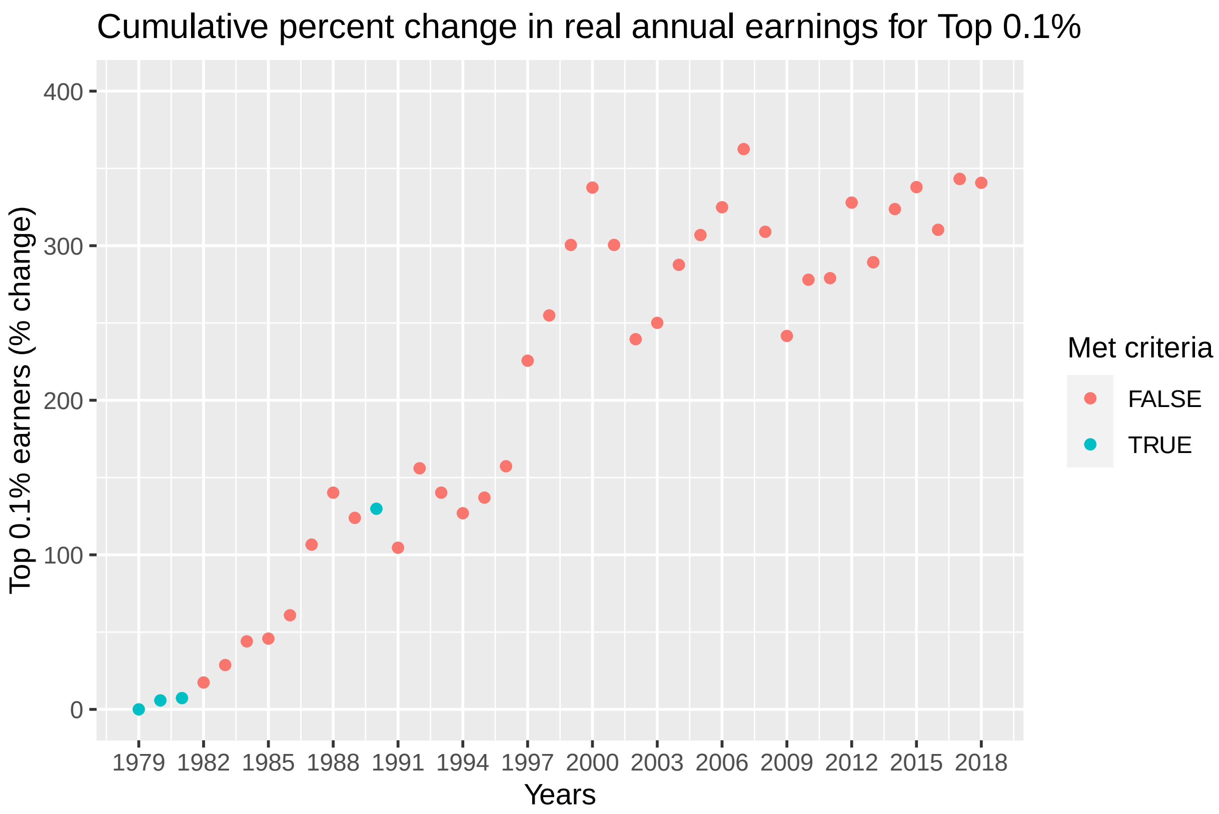


```
plot10 <- ggplot(All_Inner, aes(x = Year)) + geom_point(aes(y = Top1p, col = InflatCYN)) +
    ggtitle("Cumulative percent change in real annual earnings for Top 1% Earners") +
    labs(caption = "False = Inflation change less than 5%, True = Inflation change greater or equal to 5
        color = "Met criteria")

# Change x and y axis labels, and limits
plot10 + scale_x_continuous(name = "Years", breaks = seq(1979, 2018, 3)) + scale_y_continuous(name = "To
    limits = c(0, 400))
```



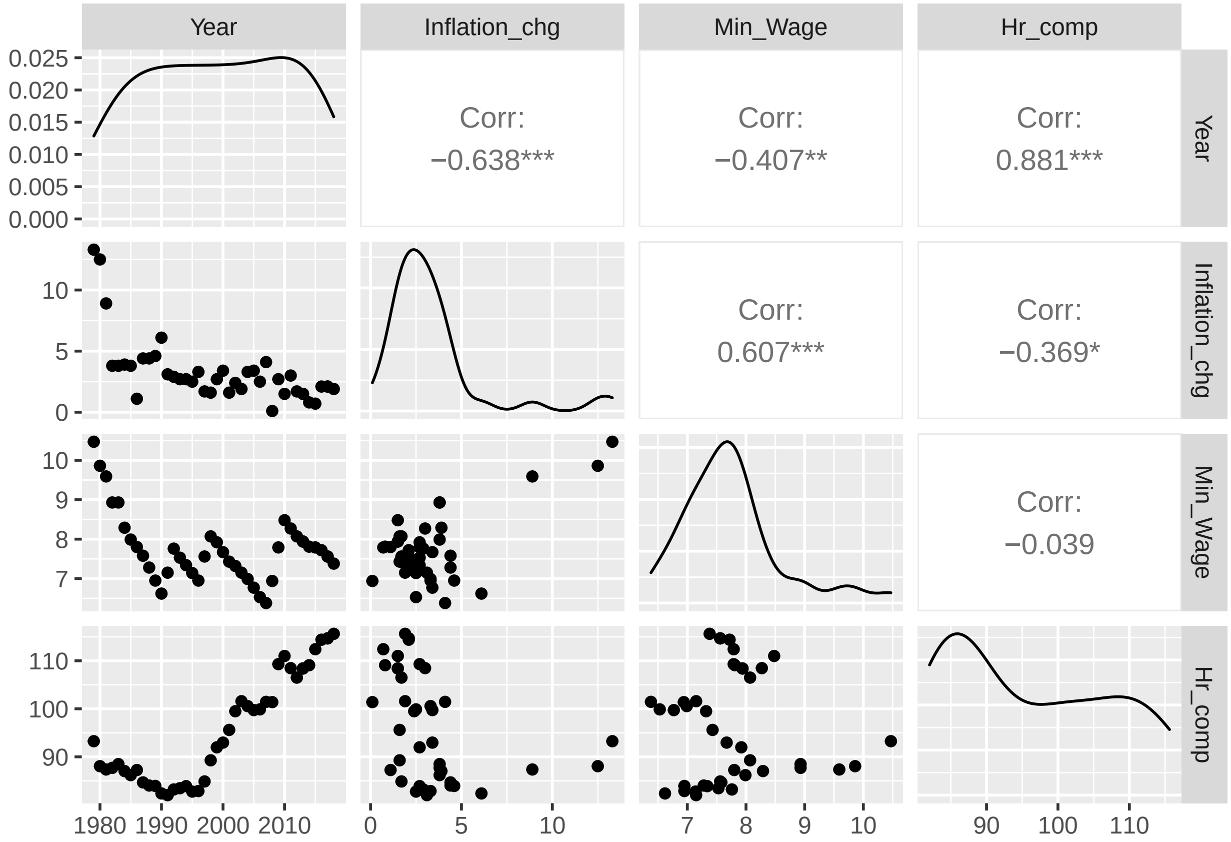
```
plot11 <- ggplot(All_Inner, aes(x = Year)) + geom_point(aes(y = Top_tenthp, col = InflatCYN)) +
    ggtitle("Cumulative percent change in real annual earnings for Top 0.1%") + labs(caption = "False =
    color = "Met criteria")

# Change x and y axis labels, and limits
plot11 + scale_x_continuous(name = "Years", breaks = seq(1979, 2018, 3)) + scale_y_continuous(name = "To
    limits = c(0, 400))
```

Cumulative percent change in real annual earnings for Top 0.1%

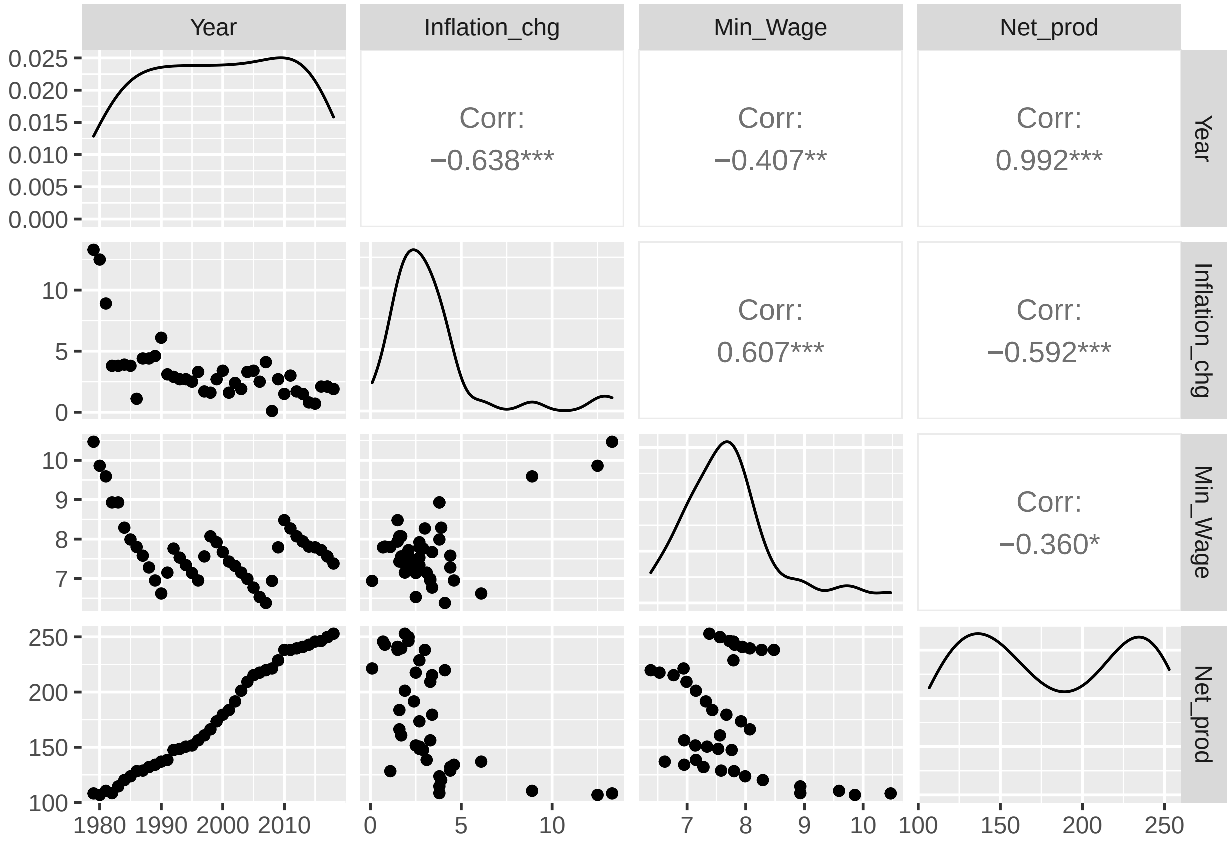


False = Inflation change less than 5%, True = Inflation change greater or equal to 5%

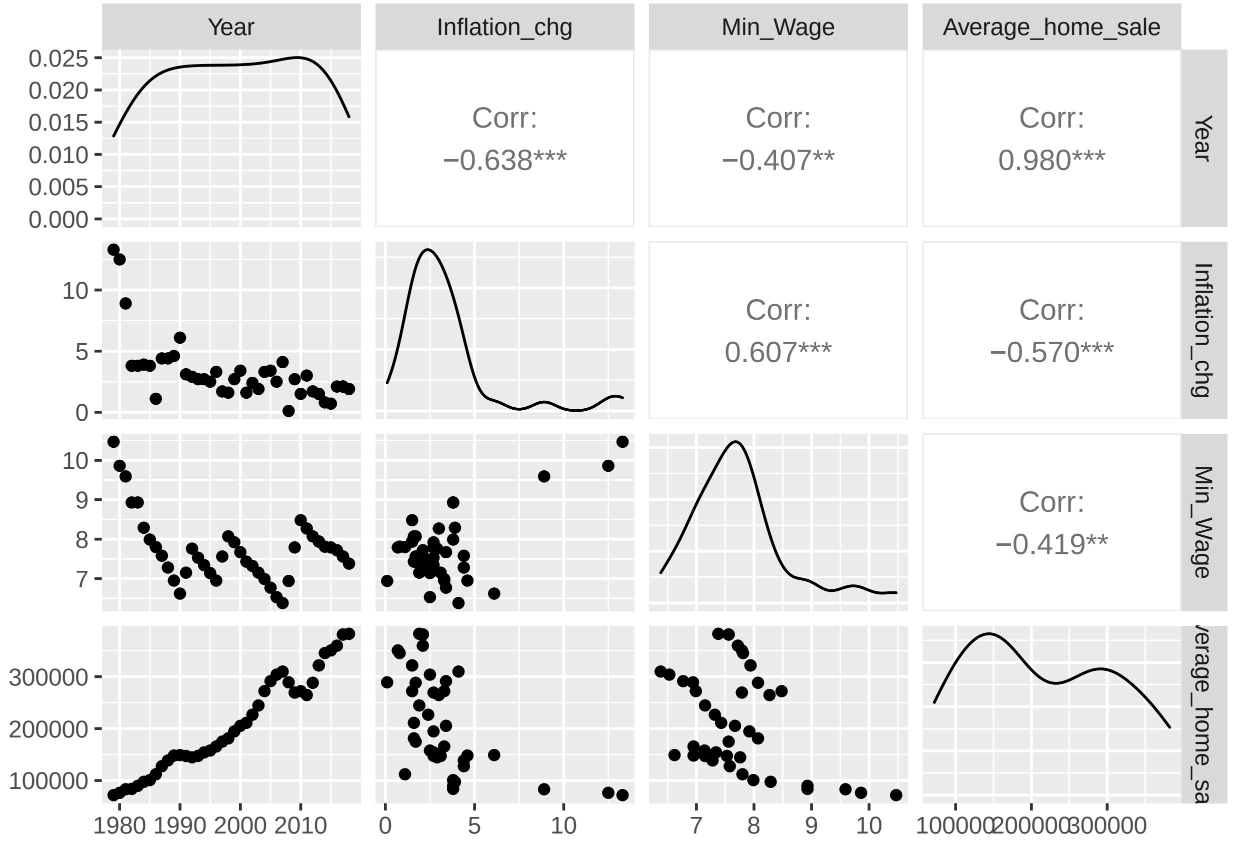
```r
# for correlation - dropping T/F variables
All_InnerN <- subset(All_Inner, select = -c(InflatiYN, InflatCYN, CURRYNN, CURRYNC))

library("GGally")
# 1 is year, 7 inflation chg, 10 is min wage, I am keeping these will all runs
# adding hourly compensation
GGally::ggpairs(All_InnerN, columns = c(1, 7, 9, 2))
```



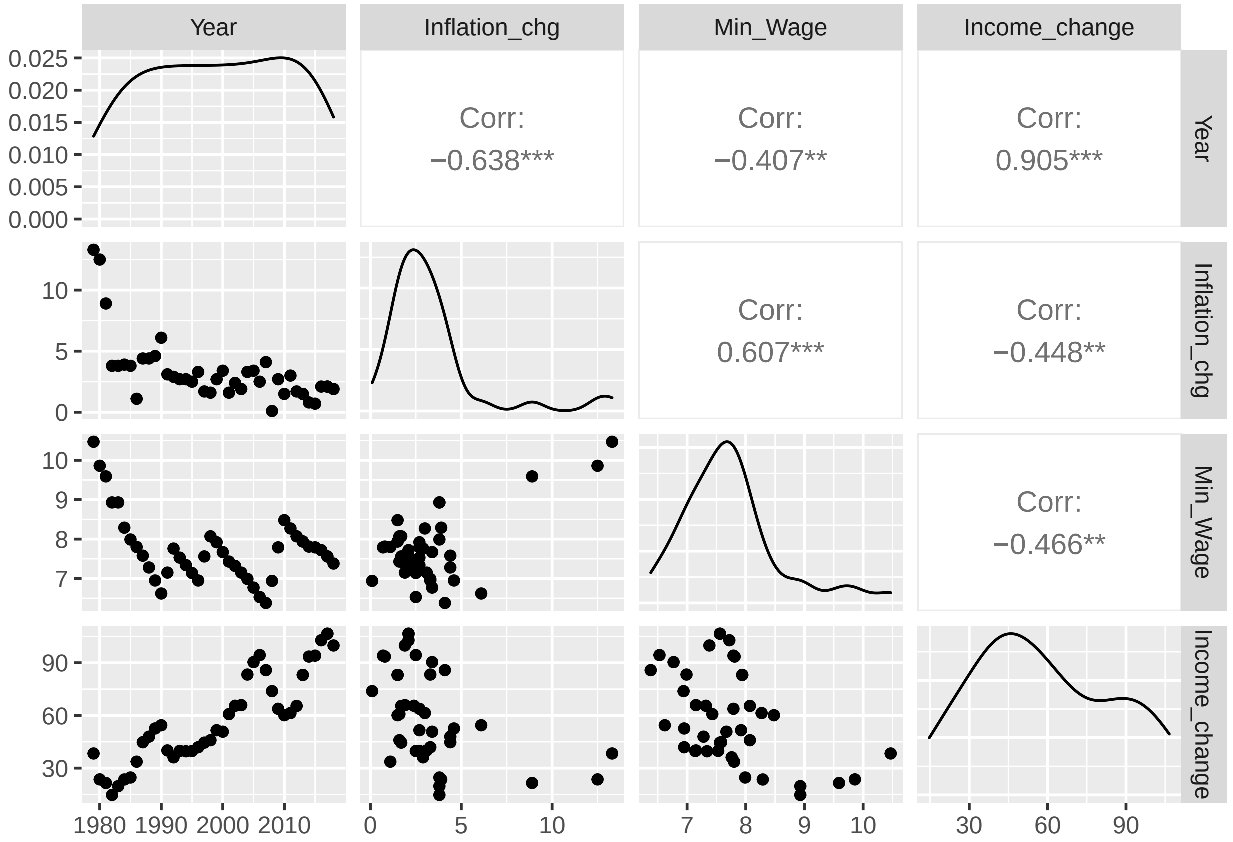
```
# adding productivity change
GGally::ggpairs(All_InnerN, columns = c(1, 7, 9, 3))
```



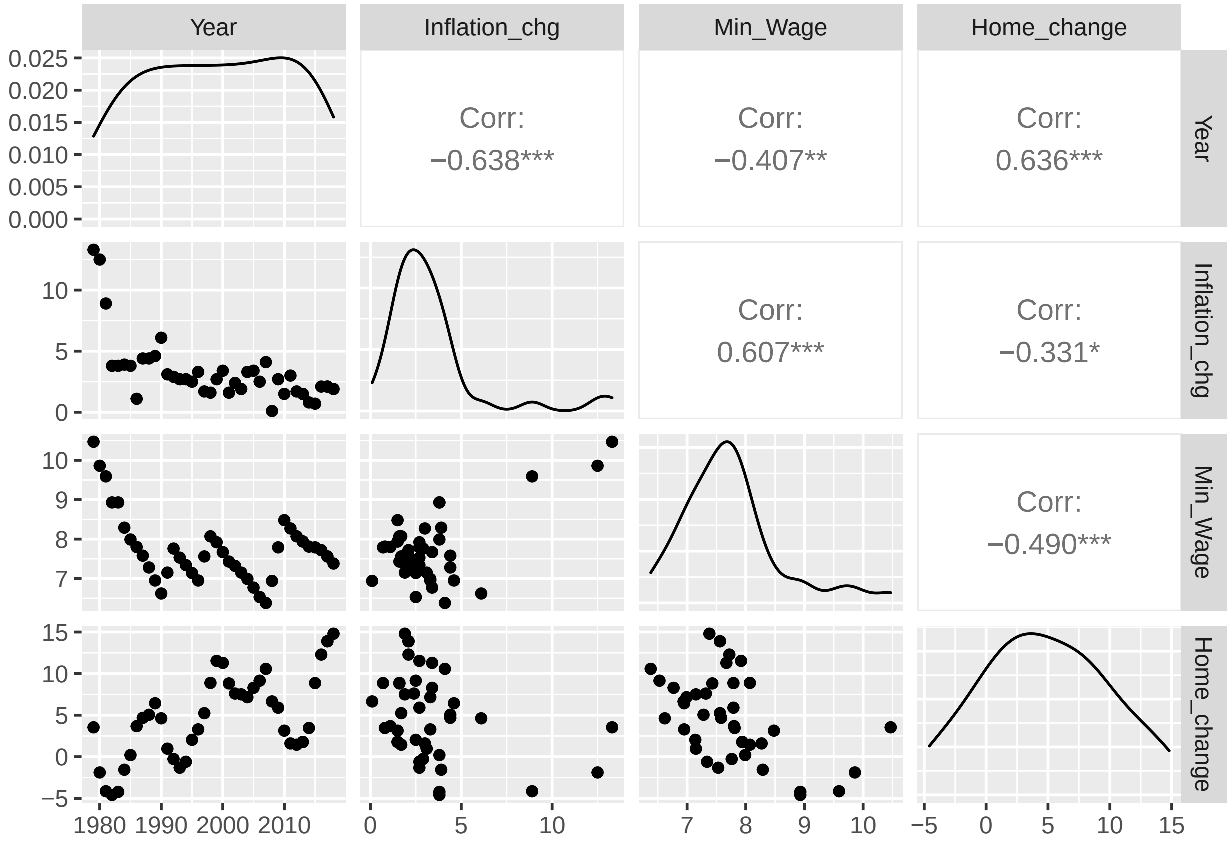
```
# adding in average home sales
GGally::ggpairs(All_InnerN, columns = c(1, 7, 9, 4))
```



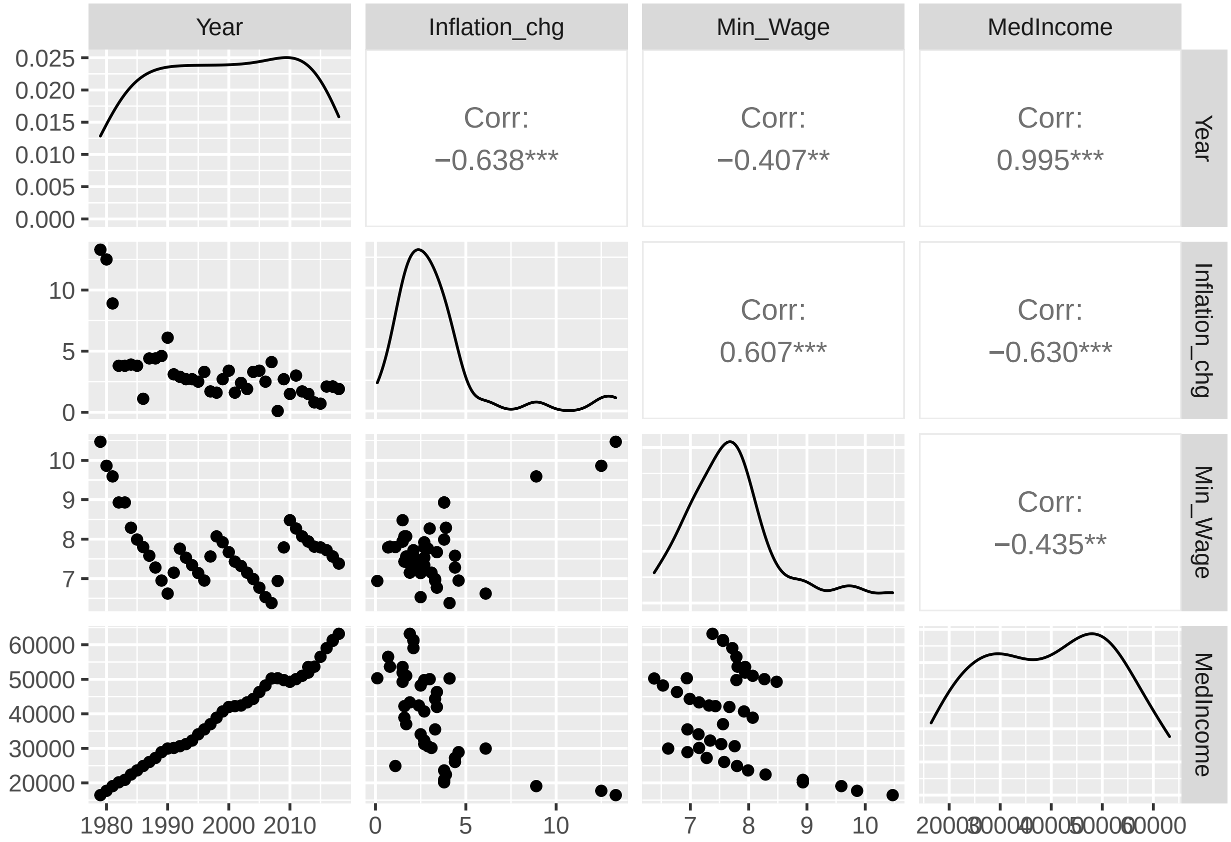
```r
# adding in ave home sales
GGally::ggpairs(All_InnerN, columns = c(1, 7, 9, 5))
```



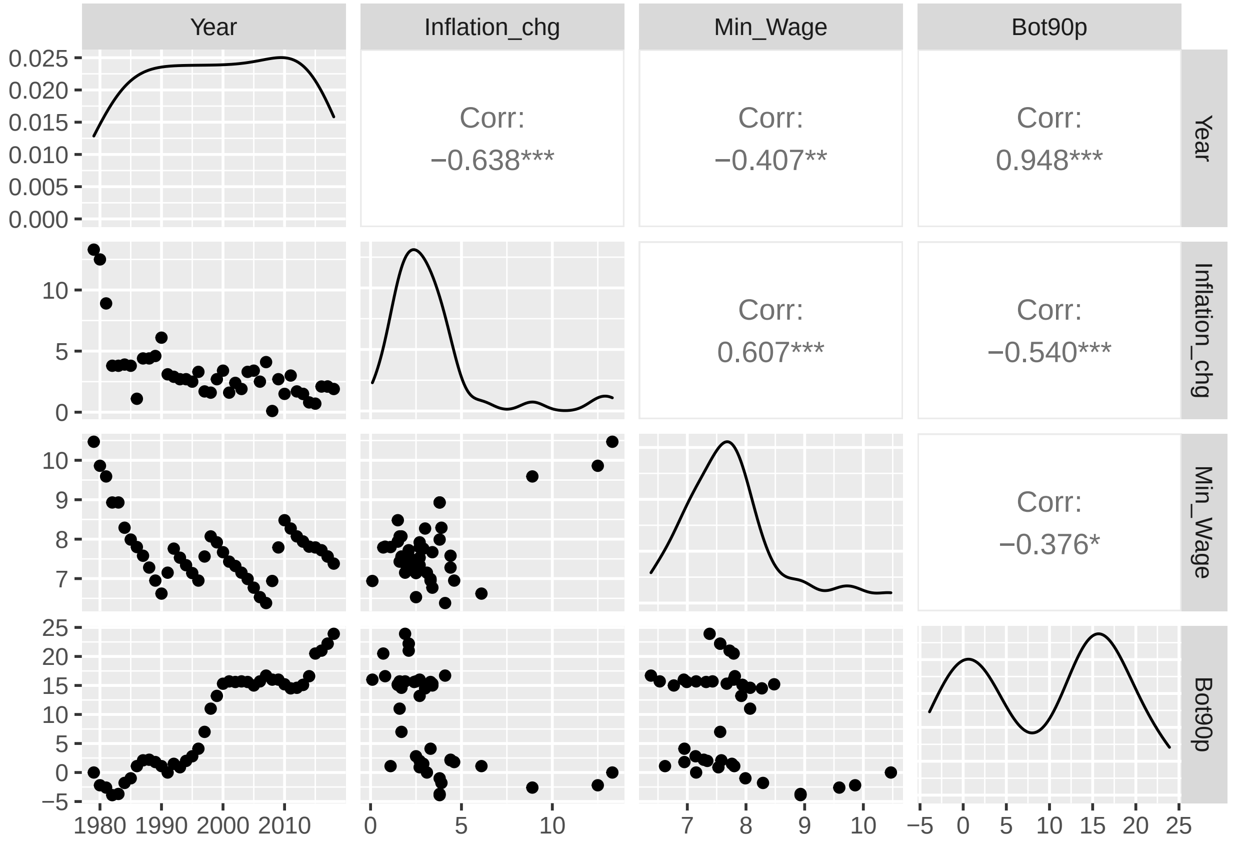
```
# adding in income chg
GGally::ggpairs(All_InnerN, columns = c(1, 7, 9, 6))
```



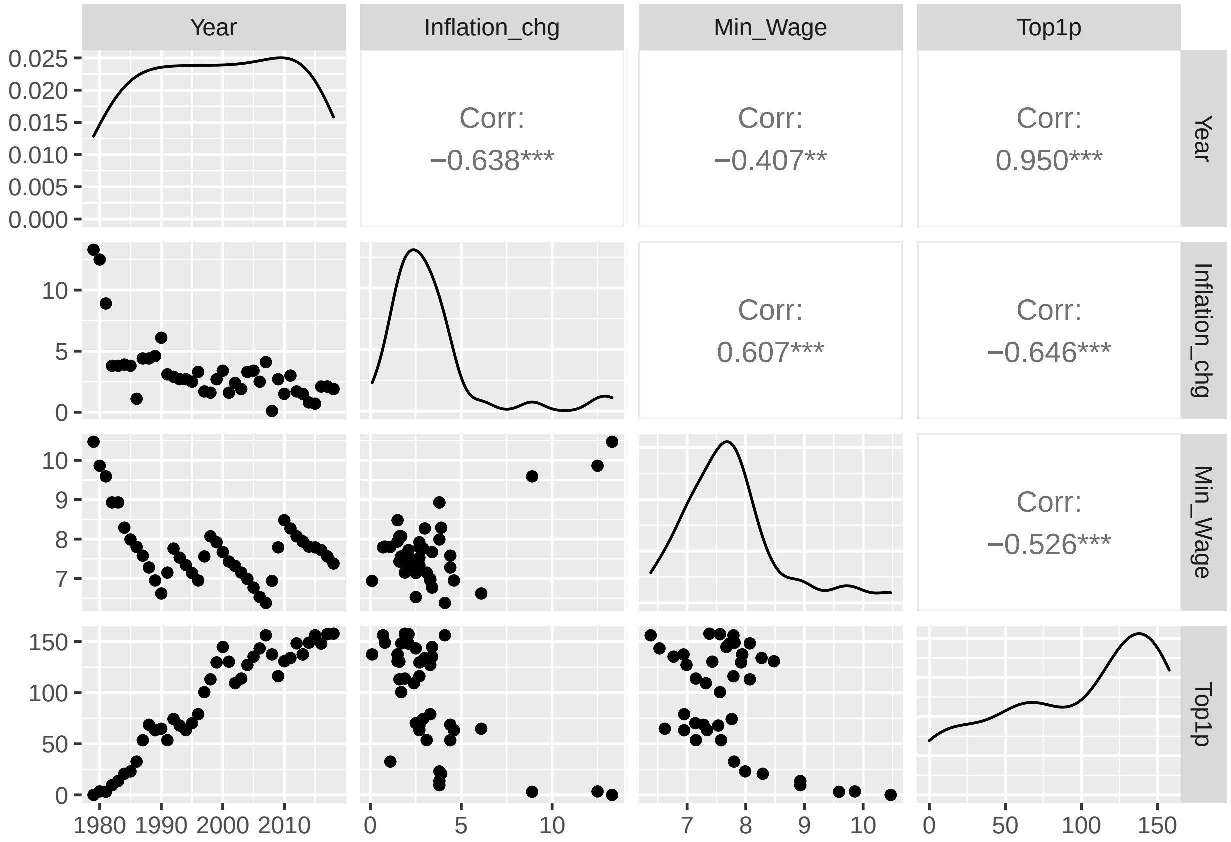
```
# adding in median income
GGally::ggpairs(All_InnerN, columns = c(1, 7, 9, 8))
```



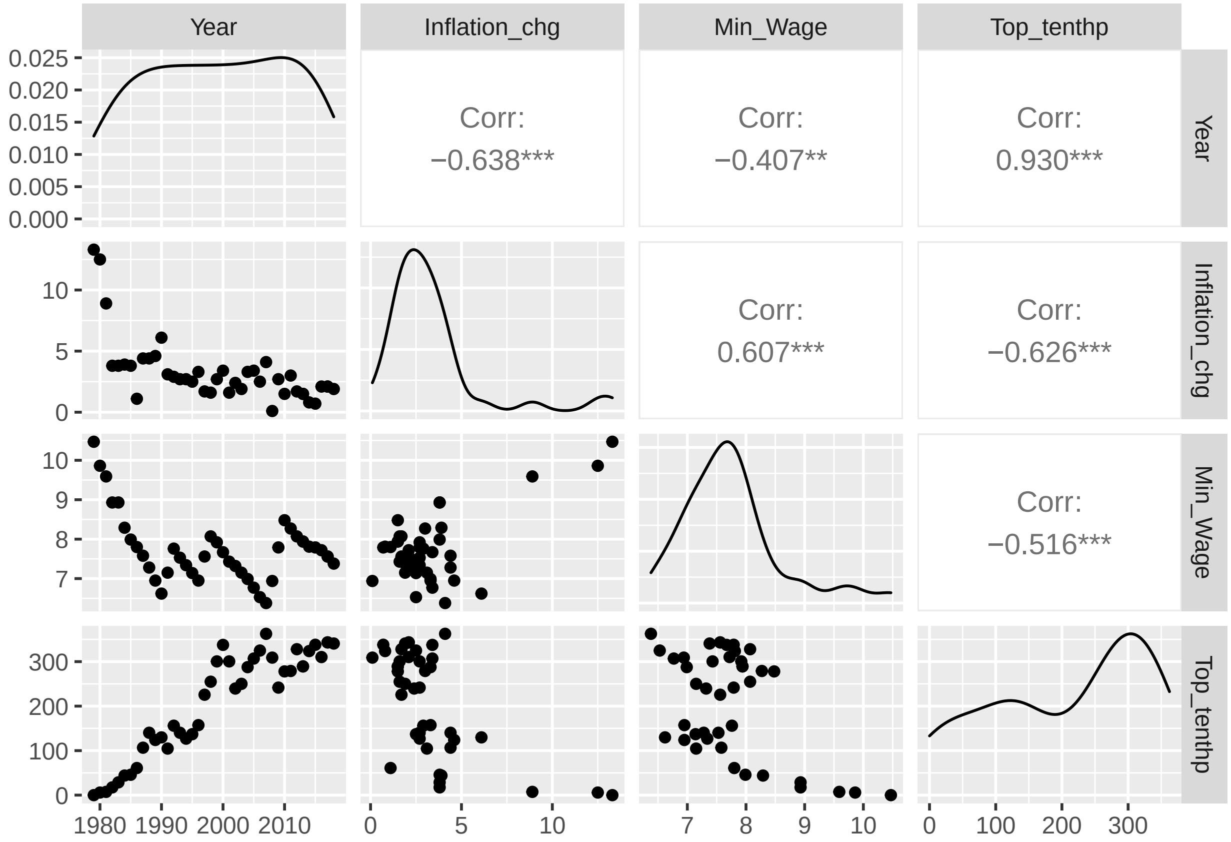
```
# adding in bottom 90%
GGally::ggpairs(All_InnerN, columns = c(1, 7, 9, 10))
```



```r
# adding in top 1%
GGally::ggpairs(All_InnerN, columns = c(1, 7, 9, 11))
```



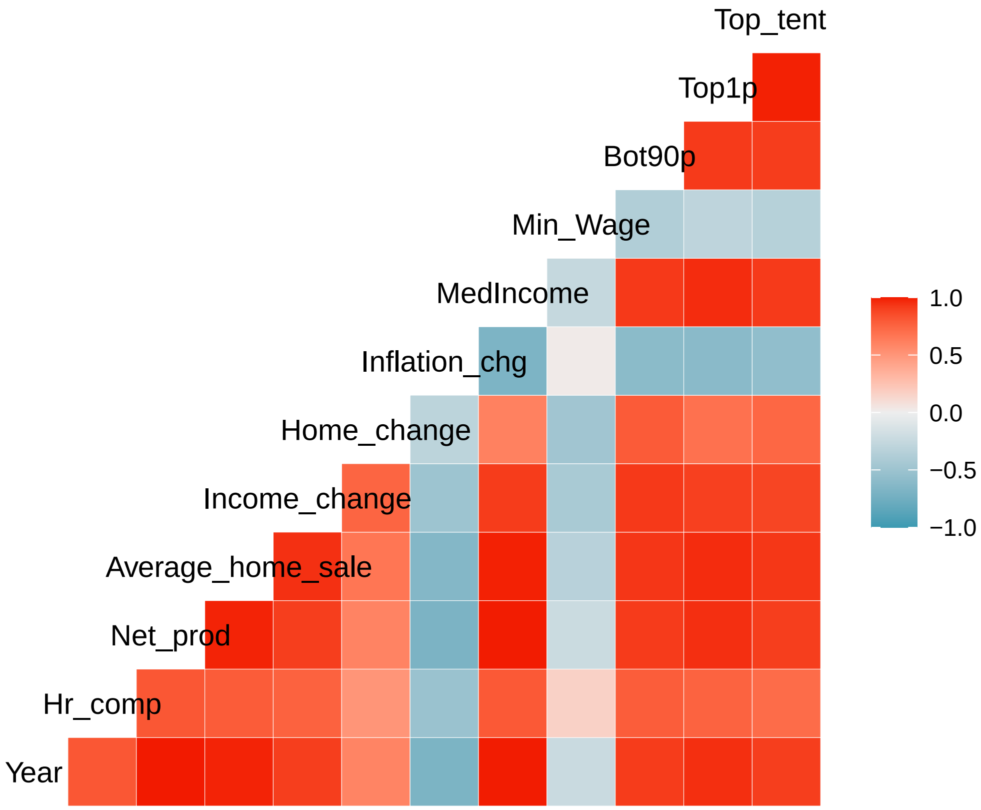
```r
# adding in top 0.1%
GGally::ggpairs(All_InnerN, columns = c(1, 7, 9, 12))
```



```
# cor(All_InnerN, use='all', method = c('spearman'))

# using this from Hmisc as I want to have the pvalues
Corr_tbl <- rcorr(as.matrix(All_InnerN, type = c("spearman")))
Corr_tbl
```

```
##                    Year Hr_comp Net_prod Average_home_sale Income_change
## Year               1.00    0.88     0.99              0.98          0.90
## Hr_comp            0.88    1.00     0.92              0.90          0.84
## Net_prod           0.99    0.92     1.00              0.98          0.91
## Average_home_sale  0.98    0.90     0.98              1.00          0.96
## Income_change      0.90    0.84     0.91              0.96          1.00
## Home_change        0.64    0.54     0.61              0.68          0.76
## Inflation_chg     -0.64   -0.37    -0.59             -0.57         -0.45
## MedIncome          1.00    0.88     0.98              0.99          0.92
## Min_Wage          -0.41   -0.04    -0.36             -0.42         -0.47
## Bot90p             0.95    0.88     0.95              0.95          0.91
## Top1p              0.95    0.77     0.94              0.93          0.88
## Top_tenthp         0.93    0.75     0.92              0.91          0.87
##                   Home_change Inflation_chg MedIncome Min_Wage Bot90p Top1p
## Year                     0.64         -0.64      1.00    -0.41   0.95  0.95
## Hr_comp                  0.54         -0.37      0.88    -0.04   0.88  0.77
## Net_prod                 0.61         -0.59      0.98    -0.36   0.95  0.94
## Average_home_sale        0.68         -0.57      0.99    -0.42   0.95  0.93
## Income_change            0.76         -0.45      0.92    -0.47   0.91  0.88
## Home_change              1.00         -0.33      0.70    -0.49   0.79  0.74
## Inflation_chg           -0.33          1.00     -0.63     0.61  -0.54 -0.65
## MedIncome                0.70         -0.63      1.00    -0.44   0.96  0.96
## Min_Wage                -0.49          0.61     -0.44     1.00  -0.38 -0.53
## Bot90p                   0.79         -0.54      0.96    -0.38   1.00  0.95
## Top1p                    0.74         -0.65      0.96    -0.53   0.95  1.00
## Top_tenthp               0.75         -0.63      0.94    -0.52   0.95  1.00
##                   Top_tenthp
## Year                    0.93
## Hr_comp                 0.75
## Net_prod                0.92
## Average_home_sale       0.91
## Income_change           0.87
## Home_change             0.75
## Inflation_chg          -0.63
## MedIncome               0.94
## Min_Wage               -0.52
## Bot90p                  0.95
## Top1p                   1.00
## Top_tenthp              1.00
##
## n= 42
##
##
## P
##                   Year   Hr_comp Net_prod Average_home_sale Income_change
## Year                     0.0000  0.0000   0.0000            0.0000
## Hr_comp           0.0000         0.0000   0.0000            0.0000
## Net_prod          0.0000 0.0000           0.0000            0.0000
## Average_home_sale 0.0000 0.0000  0.0000                     0.0000
## Income_change     0.0000 0.0000  0.0000   0.0000
## Home_change       0.0000 0.0003  0.0000   0.0000            0.0000
## Inflation_chg     0.0000 0.0162  0.0000   0.0000            0.0029
## MedIncome         0.0000 0.0000  0.0000   0.0000            0.0000
## Min_Wage          0.0075 0.8052  0.0193   0.0058            0.0019
## Bot90p            0.0000 0.0000  0.0000   0.0000            0.0000
## Top1p             0.0000 0.0000  0.0000   0.0000            0.0000
## Top_tenthp        0.0000 0.0000  0.0000   0.0000            0.0000
##                   Home_change Inflation_chg MedIncome Min_Wage Bot90p Top1p
## Year              0.0000      0.0000        0.0000    0.0075   0.0000 0.0000
## Hr_comp           0.0003      0.0162        0.0000    0.8052   0.0000 0.0000
## Net_prod          0.0000      0.0000        0.0000    0.0193   0.0000 0.0000
## Average_home_sale 0.0000      0.0000        0.0000    0.0058   0.0000 0.0000
## Income_change     0.0000      0.0029        0.0000    0.0019   0.0000 0.0000
## Home_change                   0.0323        0.0000    0.0010   0.0000 0.0000
## Inflation_chg     0.0323                    0.0000    0.0000   0.0002 0.0000
## MedIncome         0.0000      0.0000                  0.0040   0.0000 0.0000
## Min_Wage          0.0010      0.0000        0.0040             0.0141 0.0004
## Bot90p            0.0000      0.0002        0.0000    0.0141          0.0000
## Top1p             0.0000      0.0000        0.0000    0.0004   0.0000
## Top_tenthp        0.0000      0.0000        0.0000    0.0005   0.0000 0.0000
##                   Top_tenthp
## Year              0.0000
## Hr_comp           0.0000
## Net_prod          0.0000
## Average_home_sale 0.0000
## Income_change     0.0000
## Home_change       0.0000
## Inflation_chg     0.0000
## MedIncome         0.0000
## Min_Wage          0.0005
## Bot90p            0.0000
## Top1p             0.0000
## Top_tenthp
```

```
# visual of the correlation matrix
ggcorr(All_InnerN, method = c("pairwise.complete.obs", "spearman"))
```



# Questions for part 2 of final project

Late data source added

I added one additional data source because my inflation rate data is through 2022 while other sources are only to 2018. I hand entered this data from the plot on website. The plot and article use govt data as the source. How this data is converted to 2019 dollars was not explicitly defined on the site.

7. latebreaker_minwage.xlsx. USA Facts, September 18, 2019. website: https://usafacts.org/articles/minimum-wage-america-how-many-people-are-earning-725-hour/?utm_source=bing&utm_

medium=cpc&utm_campaign=ND-Economy&msclkid=e828e5a0aeda14c17aa9874fedeec55f. Minimum wage in America: How many people are earning $7.25 an hour?

There are two variables in the data source, Year (numeric, cumulative), Minimum wage that has been adjusted to 2019 dollars (numeric, dollar value).

Note, because this value is from 2019, I wanted to have this value be at 2022 but was not able to find it. However, I did use a online calculator to determine what $7.25 would be in 2022, which is $8.41 (https://www.in2013dollars.com/us/inflation/2019?amount=7.25). I used this value of 8.41 to create a True/False value for each year as meeting the base requirement of minimum wage or not. I am investigating if this is the right approach. The past few years, due to the pandemic and other factors, have impacted the value of the dollar and I wanted to capture that.

Q1: How to import and clean my data

For most of my data, I imported excel workbooks. Because many of my sources were from the government sites or derived from govenment sites, there were footnotes and titles in the source data itself. I have original copies but imported workbooks where I removed those. Also, I had a few instances where there was a single cell that was not importing due to some entry difference. In these cases, I fixed the source so that it would import as expected (and not as a missing value).

In terms of cleaning, I did ensure that my key field of Year was a numeric variable in all dataframes. Another issue I had was with dollar signs, at times Excel displays a dollar sign for a numeric and at times the Excel field is a character with a dollar sign and comma. I removed the characters signs when appropriate. I had this same issue with the percent symbol and removed this as well. This reminds me that I need to work on a few things in R to get more fluid and knowledgeable with how the code works (like regular expressions or sub-stringing, etc.).

Initially I did an outer merge so I could see what was missing and where or when. I did subset one data source to the year value greater than 1966 as this was similar to other data sources' starting points. For the quarterly median income data, I took the mean over the year, I did this in excel and then used a formula to get the year from the date as another column. I then made this a new spreadsheet in the workbook with a new name. At times, I am concerned that my R skills are too new to do heavy data manipulation thus I did this outside of R.

I renamed a lot of variables in R from how they were read in and then checked values to ensure a numeric was remained a numeric.

Q2: What does the final data set look like?

The final dataframe has 16 variables, 4 of which are created as dichotomous variables for TRUE/FALSE (1/0) categories for Inflation rate being over 5% and if minimum wage met the base criteria of today's minimum wage in today's dollars.

The dataframe I used for correlation analysis had 12 variables, I dropped the 4 dichotomous variables prior to using in analyses. This dataframe is printed above in this pdf with the R output section titled: # QUESTION 1, showing final dataframe in concise format.

Q3: Questions for future steps.

Questions for future steps include better understanding using percentages with correlation analyses. The plots are as expected but I am not clear if I can use a percentage in a correlation analysis. I read mixed information online but because these percentages are not adding up to 100% (as in part of a whole where all the column (variable) added together would equal 100%) I conducted the correlation analysis without tranforming the variables, from what I read, this was appropriate.

Also, it would be great to explore the relationship between the variables more and I would add additional visualizations. I wanted to tap into this more, I think there is more in the data that can demonstrate how minimum wage and inflation rate can affect or not affect items like housing cost, wages, and other key factors. Did I have the right data sources to look at the issue I wanted to? This is an unanswered question to me; the topic is very large and there is quite a bit of data, the government sites are data heavy and not always easy to navigate.

Once I am sure using percentage values is appropriate for the correlation analysis, I would be interested in understand what these data would look like modeled.

In terms of what did I need to learn for data cleaning and data munging, there are a few items.

I still need to get a handle on loops, I have been avoiding them and going the long route of not using them however, they can save time and create efficiency. I do prefer to use shorter code if possible. Also, I wanted to have a better handle on vectors and how data imports and how to quickly check for values that shouldn't be there (frequency counts). I did many of these checks visually or using excel but I would prefer to use R for this.

Q4: What information is not self-evident?

Creating the 2 dichotomous variables was done in order to better explore information that was not evident from the continuous variables. I planned to look at the other variables from my main dataframe and see if there was patterns that were not known or, if by combining these information can I investigate why minimum wage has not been raised and potentially to take note of the affects of not raising it.

Q5: What are different ways you could look at this data?

Visualization will be a key way to review this data, I created a limited number of visualizations for this project (at this stage). I wanted to investigate the relationship of the each key economic indicator against the Year variable, as a primary way to review the data and break up the key indicators by the dichotomous variables for inflation rate and minimum wage.

There are other visualizations that would be also important to investigate the relationship between pairs of variables (and using year). Histograms were created to explore the distribution of the data - but ideally, I would create scatter plots of the paired variables (for example, looking at house value increase and income changes) as well as plotting these as line plots over time and using overlay.

Most of the questions detailed in Part 1 were to explore the relationship between different variables. For this, I ran spearman correlations with p values to better understand the underlying relationships between the variables in my data frame.

Q6: How do you plan to slice and dice the data?

I imported seven data sources, cleaned these, created new variables, in some cases dropped variables that seemed like they would not be useful in analysis and then merged the data.

I was interested in having one single merged dataset that had all the years of economic data included.

I created two merged dataframes previously described by joining these on the variable Year. This allowed me to more easily create one large dataset. However, as mentioned, merging all sources together did result in data lost.

It would be good to review if the different initial merges resulted in fuller data however, I did also create an outer join dataframe and initially did explore correlations and plots using the "GGally" library. The correlation values were slightly different but not enough to warrent switching my approach.

An approach I did not take but would like to would be to describe the data more fully. I created histograms to investigate distribution to determine which correlation method should be use. However, I would have liked to include descriptive statistics on each variable and that is an oversight on my part. I think this would have helped understand initial trends with the data.

Key ways that I am slicing and dicing the data are by the Inflation rate Y/N variable which was created by any year where the inflation rate was 5% or higher. This rate of 5% was mentioned in various sources as a marker of acceptable inflation however, there are many factors which contribute to this determination. I used online sources for the US economy and what is considered an acceptable amount of inflation in the US. Another key way to slice the data was to create a Y/N variable of minimum wage to determine if each year met that criteria based on the current minimum wage of $7.25. I attempted to bring this value of 7.25 forward to 2022 dollars (where the dollar has increased by a factor of 1.16).

Q7: How could you summarize your data to answer key questions?

I could create correlation tables with p values and answer the key questions from the Part 1 of this project.

Likewise, I would also want to create descriptive statistics tables for the variables in order to describe the data first.

If the models are run, I would create tables to describe the model statistics.

Q8: What types of plots and tables will help you to illustrate the findings to your questions?

Time series plots for each variable will help to illustrate the findings as well as plots of paired variables (not time series). Currently, I have developed scatter plots for time series. For plots that would be paired variables, I would create scatter plots as well. I am exploring if overlayed line plots would help illustrate relationships for these data.

Q9: Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

It would be possible to incorporate machine learning techniques to improve the answer for this research question. All my current data sources are pared down from the original data sources at the US government level. These original data are quite large and typically based on different surveys. However, there are other sources from the web such as YouTube, Google Trends, Twitter data which could be explored to analyze key words on minimum wage.These could then be added a model, and then train the model, and then creating predictions using the model.

Q10: Questions for future steps.

I used the pipe feature at times and find this very helpful; I did this to create a concise display of my final dataframe. Also, I reduced variables on my dataset as mentioned in order to be able to run correlation analysis with more ease. Though I create two variables, I wonder if I could have created more categories from these continuous variables that I have. I think this data has a lot of possibilities for exploration and I am at the beginning phases.