Unemployement Rate Analysis by Sriram Amruthur and Yohana Pandian

Gitlink: [Unemployement Rate Analysis code file](https://github.com/yohana29/Unemployement-rate-analysis.git)

#### The GIT repository has the code file, datasets, README file and the markdown file as well

# Motivation and Overview

Firstly, we developed an inclination after learning how machine learning models are employed in various fields to solve real-life problems.We were inspired to dig deep into our analysis after attending the guest lecture2.Exploring on concepts discussed in the class helped us prepare for this project and inject more confidence for us to deliver.Ultimately, we zeroed in on this topic of unemployment as there has been a steady rise in job losses across the world, especially after Covid-19.

The main goal of the project is to find the factors that influence the unemployement rate. Further, to find if unemployement truly is a factor that affects the country’s economy (GDP) and crime rates.

# Question we are trying to answer

Initially the goal of our project was to only find the factors that are responsible for unemployement. Over the course of time, on reading about unemployement and how it affects the society we wanted to analyse and see if the results claimed in the internet where GDP,crimes rates are effected by unemployed could be statistically proven with the available data which eventually led us to extend the project. These are the questions we are trying to answer. 1.Does race,gender,age,marital status have an impact on the unemployement? 2.If yes, what factors among those significantly affect the unemployement? 3.Out of all the Race, which has the most to least unemployement? 4.Are males more unemployed or females? 5.Does marital status have an effect on employement? 6.Which year has recorded the least and most unemployement rates? 7.Does age have an effect on employement? 8.Does as country’s GDP depend on unemployement? 9.Does unemployement have an effect on the increasing/decreasing crime rates? 10.Can the above questions be statistically proven and visualized?

# Exploratory data analysis

We performed exploratory data analysis using data visualization techniques to gather some insightful information between unemployment rate and other features in our dataset.First, we used the merged dataset and plotted age as our response variable and employment rate as the target variable to see which category had a predominant effect.Rest of the other features like gender, marital status, race and year was also compared with the unemployment rate.

# Data Analysis and statistical methods used

We created a linear regression model to forecast the unemployment rate for the year 2021 using the inbuilt lm() function in R.We received a multiple R square value of 0.4853.Since the goal of our project was to perform analysis based on the unemployment rate, we decided to stick with the linear regression model as we were content with the results obtained.We also built an ANOVA model to check for significance for our features.

# Narrative and Summary

We got the answers to all the questions we were trying to answer 1.Yes,race,gender,age,marital status has an impact on the unemployement. 2.All the factors above has a significant p-value from the anova table. 3.The order of unemployement among the races is Black or African > Hispanic or latino > White > Asian 4.Males are more unemployed than females. 5.Since people who are married has a commitment we can see that they are more employed than individuals who are unmarried or widowed 6.The year 2010 has the lowest employement and 2021 the highest. 7.Individuals who are 25 and over are more employed than 16 and over. 8.Yes, GDP is affected by unemployement rate and the relationship is inversely proportinal. 9. Astonishingly, we are not able to statistically prove that crime rates and unemployement are related. This may be due to the fact that a lots of other factors contirbute to crime rates as well such as easy access to guns and the level of literacy. 10.Yes the above questions we statitically infered and visualized below.

## Required Libraries

library(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

library(tidyr)  
library(ggplot2)

## Creating dummy dataframe to save the compiled single dataframe for later

testing\_merge <- data.frame(matrix(ncol = 6, nrow = 0))  
colnames(testing\_merge) <- c('Year','Gender','Age','Race', 'Marital.status','Unemployement\_Rate')

## Code for dynamic reading of files

file\_names<-c('yr\_2021\_test.csv','yr\_2020\_test.csv','yr\_2019\_test.csv'  
 ,'yr\_2018\_test.csv','yr\_2017\_test.csv','yr\_2016\_test.csv'  
 ,'yr\_2015\_test.csv','yr\_2014\_test.csv','yr\_2013\_test.csv'  
 ,'yr\_2012\_test.csv','yr\_2011\_test.csv','yr\_2010\_test.csv'  
 ,'yr\_2009\_test.csv','yr\_2008\_test.csv','yr\_2007\_test.csv'  
 ,'yr\_2006\_test.csv','yr\_2005\_test.csv','yr\_2004\_test.csv'  
 ,'yr\_2003\_test.csv','yr\_2002\_test.csv')

## Recurssive function to dynamically read files for data preprocessing

for (val in file\_names){  
 #Read the CSV file  
 yr\_2021<-read.csv(paste0("D:/Data Science 2021-2023/Spring 2022/Advanced R/Project/Datasets/Excel/Test\_CSV/",val))  
   
 #introduce identity column  
 yr\_2021 <- tibble::rowid\_to\_column(yr\_2021, "index")  
   
 #create two subsets with the identity   
 #1) identity,men,women   
 yr\_2021\_gender<-yr\_2021%>%  
 select(index,Men,Women)  
   
 #2)identity,Year,Race, Age, Marital.status  
 yr\_2021\_othercol<-yr\_2021%>%  
 select(index,Year,Race, Age, Marital.status)  
   
 ## Pivot the Gender and unemployement rates  
 yr\_2021\_gender <- yr\_2021\_gender %>%  
 select(index,Men,Women) %>%  
 gather(key = Gender, value = Unemployement\_Rate, -index)  
  
 #join the othercol and gender df wrt to the index column  
 yr\_2021\_fin<-yr\_2021\_gender%>%  
 inner\_join(yr\_2021\_othercol, by = c("index" = "index")) %>%   
 select(Year,Gender,Age,Race, Marital.status,Unemployement\_Rate)  
   
 #Append all the datasets together  
 testing\_merge<-rbind(testing\_merge,yr\_2021\_fin)  
   
}  
View(testing\_merge)

## Data validation

#Checking if all the year is read  
testing\_merge %>% distinct(Year,Marital.status)

## Year Marital.status  
## 1 2021 Married, spouse present  
## 2 2021 Widowed, divorced, or separated  
## 3 2021 Never married  
## 4 2020 Married, spouse present  
## 5 2020 Widowed, divorced, or separated  
## 6 2020 Never married  
## 7 2019 Married, spouse present  
## 8 2019 Widowed, divorced, or separated  
## 9 2019 Never married  
## 10 2018 Married, spouse present  
## 11 2018 Widowed, divorced, or separated  
## 12 2018 Never married  
## 13 2017 Married, spouse present  
## 14 2017 Widowed, divorced, or separated  
## 15 2017 Never married  
## 16 2016 Married, spouse present  
## 17 2016 Widowed, divorced, or separated  
## 18 2016 Never married  
## 19 2015 Married, spouse present  
## 20 2015 Widowed, divorced, or separated  
## 21 2015 Never married  
## 22 2014 Married, spouse present  
## 23 2014 Widowed, divorced, or separated  
## 24 2014 Never married  
## 25 2013 Married, spouse present  
## 26 2013 Widowed, divorced, or separated  
## 27 2013 Never married  
## 28 2012 Married, spouse present  
## 29 2012 Widowed, divorced, or separated  
## 30 2012 Never married  
## 31 2011 Married, spouse present  
## 32 2011 Widowed, divorced, or separated  
## 33 2011 Never married  
## 34 2010 Married, spouse present  
## 35 2010 Widowed, divorced, or separated  
## 36 2010 Never married  
## 37 2009 Married, spouse present  
## 38 2009 Widowed, divorced, or separated  
## 39 2009 Never married  
## 40 2008 Married, spouse present  
## 41 2008 Widowed, divorced, or separated  
## 42 2008 Never married  
## 43 2007 Married, spouse present  
## 44 2007 Widowed, divorced, or separated  
## 45 2007 Never married  
## 46 2006 Married, spouse present  
## 47 2006 Widowed, divorced, or separated  
## 48 2006 Never married  
## 49 2005 Married, spouse present  
## 50 2005 Widowed, divorced, or separated  
## 51 2005 Never married  
## 52 2004 Married, spouse present  
## 53 2004 Widowed, divorced, or separated  
## 54 2004 Never married  
## 55 2003 Married, spouse present  
## 56 2003 Widowed, divorced, or separated  
## 57 2003 Never married  
## 58 2002 Married, spouse present  
## 59 2002 Widowed, divorced, or separated  
## 60 2002 Never married

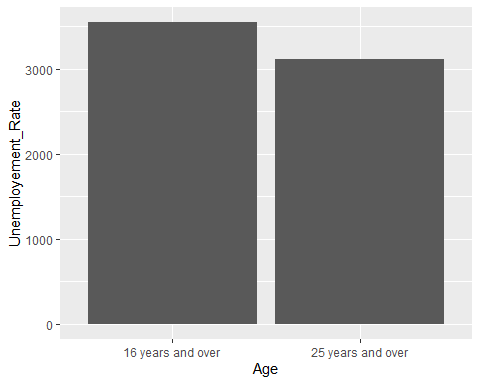
#Each year has 48 records  
testing\_merge %>% count()

## n  
## 1 960

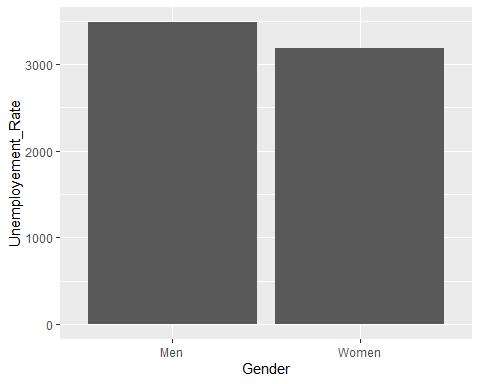
## Exploratory Analysis

### Visualization using ggplot

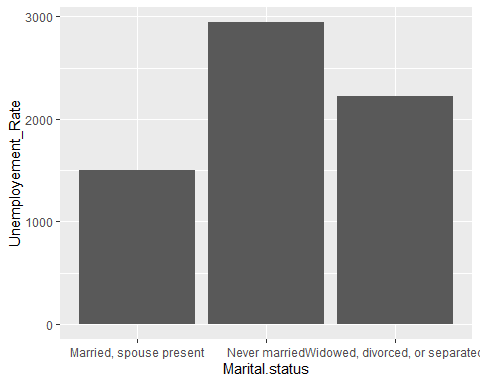
#bar chart  
ggplot(data = testing\_merge) +  
 geom\_col(mapping = aes(x = Age, y = Unemployement\_Rate))



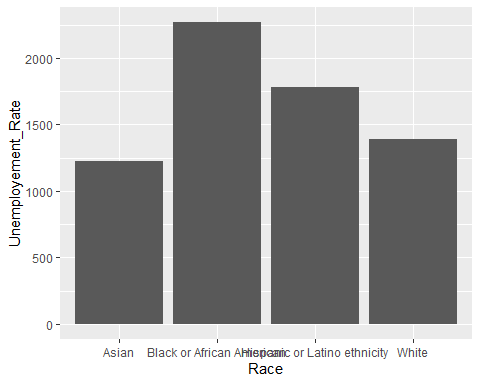
ggplot(data = testing\_merge) +  
 geom\_col(mapping = aes(x = Gender, y = Unemployement\_Rate))



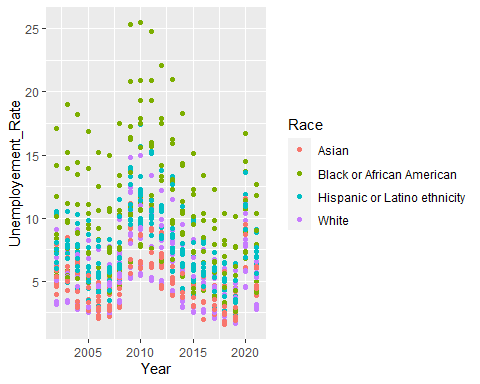
ggplot(data = testing\_merge) +  
 geom\_col(mapping = aes(x = Marital.status, y = Unemployement\_Rate))



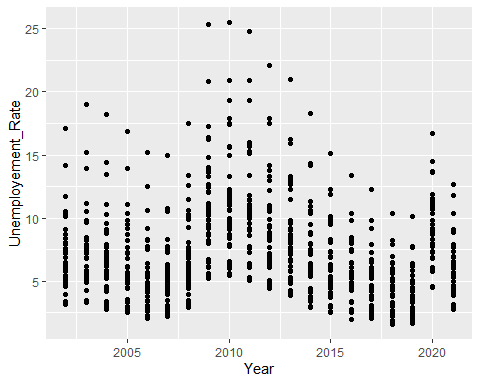
ggplot(data = testing\_merge) +  
 geom\_col(mapping = aes(x = Race, y = Unemployement\_Rate))



#Coloured point chart  
ggplot(data = testing\_merge) +  
 geom\_point(mapping = aes(x = Year, y = Unemployement\_Rate, color = Race))



ggplot(data = testing\_merge) +  
 geom\_point(mapping = aes(x = Year, y = Unemployement\_Rate))



### Analysis of Data using Regression Model

#Building a prediction model  
colnames(testing\_merge)

## [1] "Year" "Gender" "Age"   
## [4] "Race" "Marital.status" "Unemployement\_Rate"

model1= lm(Unemployement\_Rate ~ Gender + Age + Race + Marital.status, data=testing\_merge)  
anova(model1)#all the variables are significant (to check significance)

## Analysis of Variance Table  
##   
## Response: Unemployement\_Rate  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Gender 1 97.3 97.35 13.974 0.0001964 \*\*\*  
## Age 1 198.5 198.47 28.491 1.177e-07 \*\*\*  
## Race 3 2695.2 898.39 128.964 < 2.2e-16 \*\*\*  
## Marital.status 2 3263.0 1631.50 234.202 < 2.2e-16 \*\*\*  
## Residuals 952 6631.8 6.97   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(model1) #to interpret the effect of each factor

##   
## Call:  
## lm(formula = Unemployement\_Rate ~ Gender + Age + Race + Marital.status,   
## data = testing\_merge)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.3897 -1.8415 -0.3936 1.4130 13.0081   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 3.6255 0.2409 15.047  
## GenderWomen -0.6369 0.1704 -3.738  
## Age25 years and over -0.9094 0.1704 -5.338  
## RaceBlack or African American 4.3504 0.2409 18.056  
## RaceHispanic or Latino ethnicity 2.3167 0.2409 9.615  
## RaceWhite 0.6933 0.2409 2.878  
## Marital.statusNever married 4.5159 0.2087 21.643  
## Marital.statusWidowed, divorced, or separated 2.2506 0.2087 10.786  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## GenderWomen 0.000196 \*\*\*  
## Age25 years and over 1.18e-07 \*\*\*  
## RaceBlack or African American < 2e-16 \*\*\*  
## RaceHispanic or Latino ethnicity < 2e-16 \*\*\*  
## RaceWhite 0.004096 \*\*   
## Marital.statusNever married < 2e-16 \*\*\*  
## Marital.statusWidowed, divorced, or separated < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.639 on 952 degrees of freedom  
## Multiple R-squared: 0.4853, Adjusted R-squared: 0.4816   
## F-statistic: 128.3 on 7 and 952 DF, p-value: < 2.2e-16

## Additonal analysis to analyse the effect of Unemployement on GDP

#reading GDP data  
gdpdata<-read.csv("D:/Data Science 2021-2023/Spring 2022/Advanced R/Project/Datasets/GDP.csv")  
   
#Summarizing unemployement data for each year by taking the means  
yr\_groupby<-testing\_merge%>%  
 group\_by(Year) %>%   
 summarise\_at(vars(Unemployement\_Rate), list(Unemployement\_Rate = mean))  
  
yr\_groupby

## # A tibble: 20 x 2  
## Year Unemployement\_Rate  
## <int> <dbl>  
## 1 2002 6.97  
## 2 2003 7.21  
## 3 2004 6.59  
## 4 2005 5.91  
## 5 2006 5.09  
## 6 2007 5.09  
## 7 2008 6.55  
## 8 2009 10.5   
## 9 2010 10.9   
## 10 2011 10.4   
## 11 2012 9.11  
## 12 2013 8.11  
## 13 2014 6.94  
## 14 2015 5.93  
## 15 2016 5.38  
## 16 2017 4.75  
## 17 2018 4.18  
## 18 2019 3.84  
## 19 2020 9.32  
## 20 2021 6.18

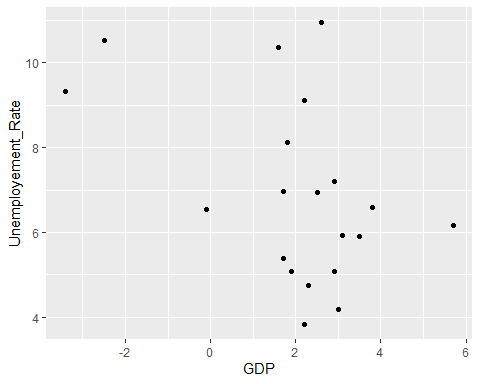
#Joining the unemployement data with GDP data for analysis  
yr\_unemp\_gdp<-yr\_groupby%>%  
 inner\_join(gdpdata, by = c("Year" = "Year")) %>%   
 select(Year,GDP,Unemployement\_Rate)  
  
yr\_unemp\_gdp

## # A tibble: 20 x 3  
## Year GDP Unemployement\_Rate  
## <int> <dbl> <dbl>  
## 1 2002 1.7 6.97  
## 2 2003 2.9 7.21  
## 3 2004 3.8 6.59  
## 4 2005 3.5 5.91  
## 5 2006 2.9 5.09  
## 6 2007 1.9 5.09  
## 7 2008 -0.1 6.55  
## 8 2009 -2.5 10.5   
## 9 2010 2.6 10.9   
## 10 2011 1.6 10.4   
## 11 2012 2.2 9.11  
## 12 2013 1.8 8.11  
## 13 2014 2.5 6.94  
## 14 2015 3.1 5.93  
## 15 2016 1.7 5.38  
## 16 2017 2.3 4.75  
## 17 2018 3 4.18  
## 18 2019 2.2 3.84  
## 19 2020 -3.4 9.32  
## 20 2021 5.7 6.18

#Relationship between unemployement and GDP, (HAS A NEGATIVE ESTIMATE)  
model\_test= lm(GDP ~ Unemployement\_Rate ,data=yr\_unemp\_gdp)  
summary(model\_test)

##   
## Call:  
## lm(formula = GDP ~ Unemployement\_Rate, data = yr\_unemp\_gdp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3355 -0.8994 0.2289 1.0877 3.3924   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.0045 1.4402 3.475 0.0027 \*\*  
## Unemployement\_Rate -0.4367 0.1986 -2.199 0.0412 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.848 on 18 degrees of freedom  
## Multiple R-squared: 0.2118, Adjusted R-squared: 0.168   
## F-statistic: 4.838 on 1 and 18 DF, p-value: 0.04115

#plots to determine the relationship  
ggplot(data = yr\_unemp\_gdp) +  
 geom\_point(mapping = aes(x = GDP, y = Unemployement\_Rate))



ggplot(data = yr\_unemp\_gdp) +  
 geom\_point(mapping = aes(x = Unemployement\_Rate, y =GDP ))



## Additonal analysis to analyse the effect of Unemployement on Crime Rates

crdata<-read.csv("D:/Data Science 2021-2023/Spring 2022/Advanced R/Project/Datasets/CrimeRates.csv")  
  
yr\_groupby\_cr<-testing\_merge%>%  
 group\_by(Year) %>%   
 summarise\_at(vars(Unemployement\_Rate), list(Unemployement\_Rate = mean))  
  
yr\_groupby\_cr

## # A tibble: 20 x 2  
## Year Unemployement\_Rate  
## <int> <dbl>  
## 1 2002 6.97  
## 2 2003 7.21  
## 3 2004 6.59  
## 4 2005 5.91  
## 5 2006 5.09  
## 6 2007 5.09  
## 7 2008 6.55  
## 8 2009 10.5   
## 9 2010 10.9   
## 10 2011 10.4   
## 11 2012 9.11  
## 12 2013 8.11  
## 13 2014 6.94  
## 14 2015 5.93  
## 15 2016 5.38  
## 16 2017 4.75  
## 17 2018 4.18  
## 18 2019 3.84  
## 19 2020 9.32  
## 20 2021 6.18

yr\_unemp\_cr<-yr\_groupby\_cr%>%  
 right\_join(crdata, by = c("Year" = "Year")) %>%   
 select(Year,Crime\_Rate,Unemployement\_Rate)  
  
yr\_unemp\_cr

## # A tibble: 18 x 3  
## Year Crime\_Rate Unemployement\_Rate  
## <int> <dbl> <dbl>  
## 1 2002 5.65 6.97  
## 2 2003 5.7 7.21  
## 3 2004 5.52 6.59  
## 4 2005 5.67 5.91  
## 5 2006 5.81 5.09  
## 6 2007 5.7 5.09  
## 7 2008 5.43 6.55  
## 8 2009 5.03 10.5   
## 9 2010 4.76 10.9   
## 10 2011 4.71 10.4   
## 11 2012 4.73 9.11  
## 12 2013 4.53 8.11  
## 13 2014 4.44 6.94  
## 14 2015 4.95 5.93  
## 15 2016 5.39 5.38  
## 16 2017 5.32 4.75  
## 17 2018 4.96 4.18  
## 18 2019 4.46 3.84

#no significance  
model\_test\_cr= lm( Crime\_Rate ~ Unemployement\_Rate ,data=yr\_unemp\_cr)  
summary(model\_test\_cr)

##   
## Call:  
## lm(formula = Crime\_Rate ~ Unemployement\_Rate, data = yr\_unemp\_cr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.9137 -0.2681 0.0706 0.4000 0.5720   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.65366 0.37001 15.280 5.79e-11 \*\*\*  
## Unemployement\_Rate -0.07294 0.05156 -1.415 0.176   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4618 on 16 degrees of freedom  
## Multiple R-squared: 0.1112, Adjusted R-squared: 0.05564   
## F-statistic: 2.002 on 1 and 16 DF, p-value: 0.1763

ggplot(data = yr\_unemp\_cr) +  
 geom\_point(mapping = aes(x = Crime\_Rate, y = Unemployement\_Rate))

