MULTIVARIATE STATISTICS
The Impact of COVID-19 on Employment

Author:
ABCXYZ001

Student Number:
ABCXYZ001

24 March 2025

Introduction

The COVID-19 pandemic has had a huge impact on labor markets around the world, reshaping employment trends across various industries. The economic disruptions caused by lockdowns, social distancing, and business closures led to a significant number of layoffs, reduced working hours, and noticeable changes in labor market participation. While some industries, such as healthcare, experienced surging demand and workforce expansion, others, particularly hospitality and retail, faced severe contractions. The pandemic's impact on employment has also been uneven across demographic groups, with variations based on gender and geographic location. Understanding these dynamics is crucial for designing effective policy interventions to support labor market recovery.

This research explores the employment impacts of the COVID-19 pandemic through an examination of high-frequency survey data from the Impact of COVID-19 on Employment dataset, which has been made available by the International Labour Organization (ILO) via the ILOSTAT database and accessed through Kaggle. The analysis will focus on employment rates, work hours and labor force participation across different regions. By applying advanced multivariate statistical techniques, this research seeks to uncover the complex relationships between demographic factors and employment outcomes during the pandemic.

Motivation

The pandemic of COVID-19 has disturbed the labor markets across the globe, leading to historic job losses and changes in the trend of employment. Literature recognizes that the pandemic affected industries such as hospitality more significantly, which was under the strictest lockdowns and restrictions (Bartik, 2020). In the same way, workers in the health care sector experienced greater demand along with heightened exposure (Liu, 2020), whereas retail workers had to contend with both labor shortages and the closure of retail establishments (Bartik, 2020). The reason for carrying out this research is so that we can know how these various variables interconnected and impacted labor market recovery. The overall impacts of employment have been considered in some research papers; nonetheless, there is a lack of adequate focus on the interactions between gender, age, and industry influencing employment experiences.

The COVID-19 pandemic reshaped labor markets worldwide to a fundamental degree, leading to unprecedented job loss and significant changes in work patterns. The magnitude and nature of these disruptions varied across industries. For example:

- Hospitality and retail sectors faced some of the most stringent lockdown restrictions, leading to business closures, mass layoffs, and financial insecurity for workers (Bartik, 2020).
- Healthcare workers experienced an increase in demand coupled with heightened exposure risks and burnout (Liu, 2020).
- Retail employees navigated a mix of labor shortages and store shutdowns, leading to fluctuating job stability.

This project will bridge these gaps by employing high-frequency International Labour Organization (ILOSTAT) data to analyze how employment trends have varied by demographic categories such as gender, and geographic location, as well as by key employment indicators like total weekly hours worked and percentage of working hours lost.

Research Questions

1. What has been the impact of COVID-19 on employment rates across different regions?

2. How have work hours and labor force participation changed during the pandemic, and how do these trends vary by gender and region?
3. Which demographic groups have been most vulnerable to employment losses, and how can policy interventions address these disparities?

Hypotheses

1. Countries with higher labor dependency ratios experienced more significant employment disruptions due to COVID-19.
2. Women experienced greater employment and work-hour losses compared to men during the pandemic.
3. The impact of COVID-19 on employment was more severe in younger workers, with higher job loss rates and reduced participation in the labor force.
4. Regional disparities exist in employment recovery, with certain regions recovering faster than others due to varying economic structures and government interventions.

Significance of the Study

This research contributes to the understanding of labor market dynamics during a global crisis by:

- Providing empirical evidence on employment disruptions across key demographic groups.
- Offering insights into the effectiveness of government responses and economic resilience factors.
- Supporting policymakers in designing targeted labor market recovery strategies to address employment disparities and ensure equitable workforce reintegration.

Through rigorous statistical analysis, this study will present a comprehensive picture of COVID-19's impact on employment and contribute to shaping post-pandemic labor policies.

Methodology

This study employs a rigorous quantitative approach to analyze the impact of COVID-19 on employment across different demographic groups and regions using 2019 data. The methodology is structured to provide insights into the pre-pandemic employment landscape and identify underlying factors influencing employment trends.

Data Preprocessing

The data used in this study comes from the International Labour Organization (ILOSTAT). The following preprocessing steps will be undertaken:

- **Handling Missing Data:** Missing values will be addressed using mean or median imputation for continuous variables. If missingness is substantial, multiple imputation techniques will be applied.
- **Categorical Variable Encoding:** Categorical variables such as region will be appropriately encoded using factor variables or dummy variables where necessary.

The detailed explanation and data dictionary for the dataset can be found through this link: https://github.com/ditiroletsoalo/Multivariate-Statistics-Project/tree/main/data.

Response Variables

The primary response variables in this study are:

- **Total Weekly Hours Worked:** The total number of hours worked per week, estimated in thousands.
- **Percentage of Working Hours Lost:** The proportion of total working hours lost due to the COVID-19 pandemic.
- **Percent Hours Lost (40hrs per week):** The percentage of normal work hours lost for individuals working a typical 40-hour week.
- **Percent Hours Lost (48hrs per week):** The percentage of normal work hours lost for individuals working a typical 48-hour week.

Exploratory Data Analysis (EDA)

Before delving into the analysis, an exploratory data analysis (EDA) will be conducted to better understand the distributions, relationships, and potential outliers in the data. The following steps will be taken:

Relationship Between Total Weekly Hours Worked and Percentage of Hours Lost

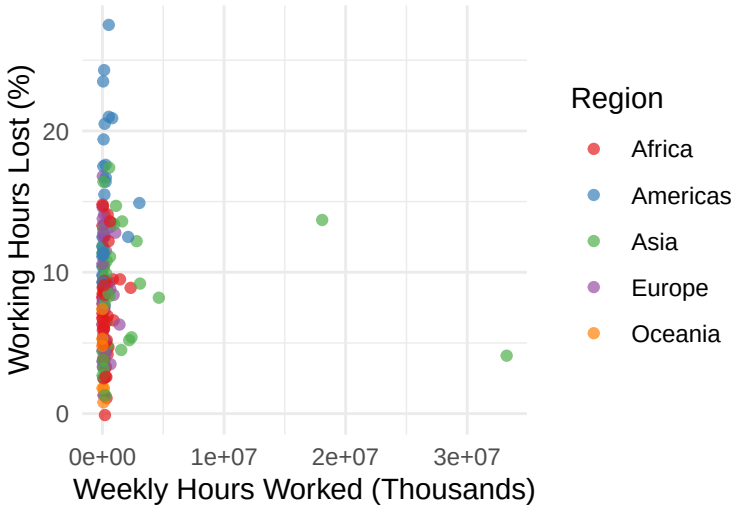

Figure 1: Weekly Hours Worked vs. Percentage of Hours Lost by Region

The scatter plot above is illustrating the relationship between total weekly hours worked and the percentage of working hours lost due to COVID-19, categorized by region, which reveals significant insights. It shows that regions with higher total weekly hours worked tend to have a lower percentage of working hours lost. This trend suggests that regions with more robust labor markets were better able to sustain working hours during the pandemic. For instance, regions like Europe and Asia, which typically have higher total weekly hours worked, experienced relatively lower percentages of working hours lost. This could be attributed to stronger economic structures and more effective government interventions in these regions.

Distribution of Total Weekly Hours Worked and Percentage of Working Hours Lost

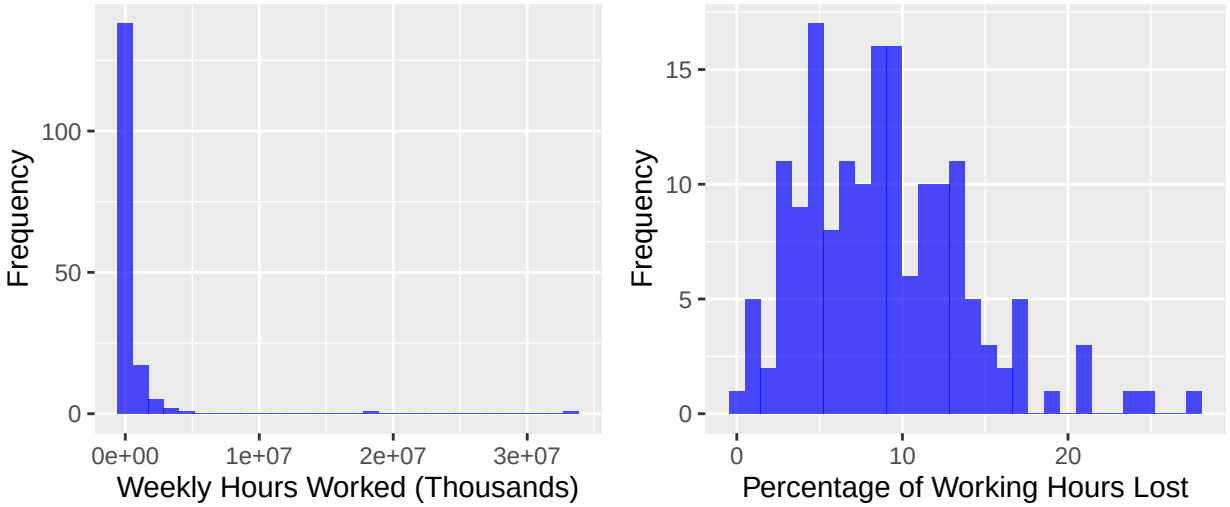

Figure 2: Distribution of Total Weekly Hours Worked and Percentage of Working Hours Lost

From the above plot it can be seen that total weekly hours worked has most observations clustered around zero, indicating that a significant portion of the workforce experienced minimal working hours during the pandemic. This suggests widespread job losses or severe reductions in working hours. The histogram for the percentage of working hours lost reveals that most observations fall within a moderate range, with a few outliers experiencing very high losses. This distribution highlights the varying impact of the pandemic on working hours, with some groups facing more severe disruptions than others. These histograms provide a clear visual representation of the central tendencies and variability in key employment indicators during the COVID-19 pandemic.

Descriptive Statistics of Key Variables

Table 1: Descriptive statistics of the Response variables

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Weekly hours worked	995.74	42090.18	136984.1	672659.70	391194.3	33243962.4
Weekly hours Lost (%)	-0.10	5.20	8.6	8.97	12.2	27.5

The table of descriptive statistics above gives a complete overview of the key response variables, indicating their central tendencies and dispersion. In the case of Weekly Hours Worked, the lowest value ever observed is 995.74 while the highest is 33,243,962.4, indicating a very wide range. The mean is considerably higher than the median, indicating the existence of outliers corresponding to very high weekly hours worked. The Weekly Hours Lost (%) percentages vary between -0.10% and 27.5%, and the mean is near the median, suggesting a relatively symmetric distribution.

Weekly Hours and Employment Ratio by Region

The plot shows employment ratios and weekly hours worked by region, disaggregated by gender, highlights regional and gender inequalities in employment. Some regions, such as Asia, register more weekly hours

worked than other regions, an indicator of tighter labor market involvement in these countries. The plot also indicates some gender disparities in employment ratios, with men consistently registering higher employment ratios than women. This gender differential indicates that women were negatively impacted by the pandemic more than men in the context of employment, most likely because of their higher numbers in the affected sectors, namely hospitality and retail.

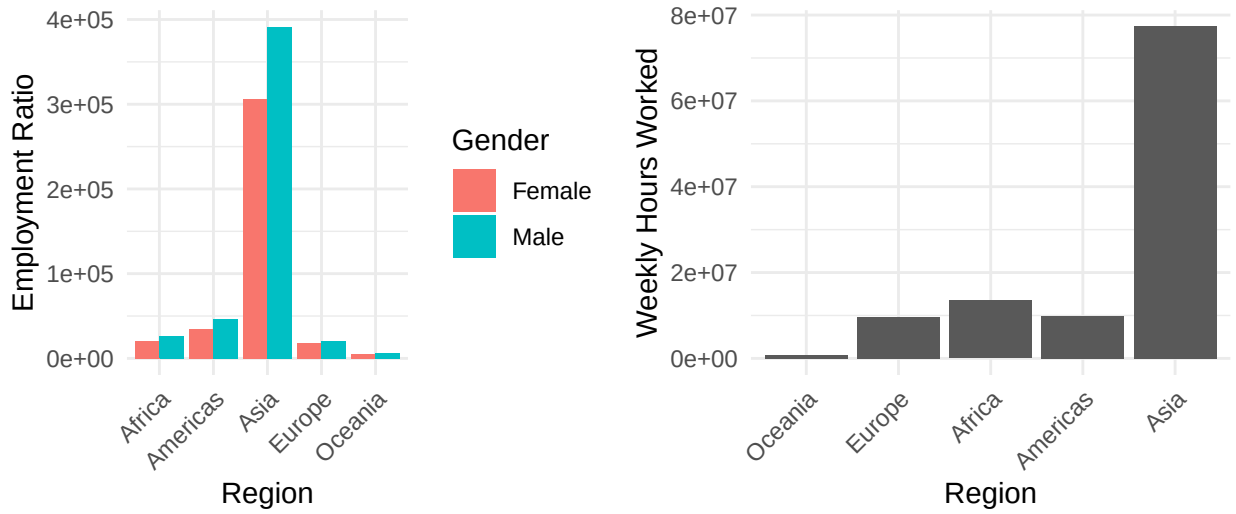

Figure 3: Comparison of Employment Ratios and Weekly Hours Worked Across Regions

Factor Analysis

Factor analysis will be employed to unearth the underlying latent factors that affect employment patterns across all various demographic groups. Since some of the variables in the dataset are categorical, an appropriate version such as **Factor Analysis for Mixed Data (FAMD)** will be used.

Structural Equation Modeling (SEM)

To examine the relationships between observed and latent variables, Structural Equation Modeling (SEM) will be applied. This approach is particularly useful in assessing:

- The direct and indirect effects of demographic variables (such as gender and region) on employment status.
- The mediating role of labor dependency ratio and employment rates by gender in shaping employment outcomes.

SEM will allow for testing hypothesized relationships while accounting for measurement error.

Results and Discussion

Multiple Linear Regression Analysis

The purpose of the multiple linear regression analysis is to examine the relationship between various predictors and the Total Weekly Hours Worked during COVID-19.

Model Summary

- **Dependent Variable**: Total Weekly Hours Worked
- **Independent Variables**: Percentage Working Hrs Lost, Labour Dependency Ratio , Employed Female, Employed Male, America, Asia, Europe, Oceania

Table 2: Model Coefficients and Statistical Values

Variable	Estimate	Std_Error	t_value	P_value
Intercept	-52460.00	4.222e+04	-1.242	0.2160
% Working Hrs Lost	-3291.00	2.445e+03	-1.346	0.1803
Labour dependency ratio	34050.00	1.670e+04	2.039	0.0432
Employed Female	56.12	9.741e-01	57.619	0.0000
Employed Male	40.32	6.113e-01	65.957	0.0000
America	-20520.00	3.606e+04	-0.569	0.5701
Asia	25270.00	2.900e+04	0.872	0.3847
Europe	-37870.00	2.857e+04	-1.326	0.1869
Oceania	-4606.00	4.974e+04	-0.093	0.9263

In this model,it can be seen from the above table that Africa is the reference region, and none of the geographic variables (America, Asia, Europe, Oceania) are statistically significant, meaning no notable differences in total weekly hours worked compared to Africa and region did not have an impact on the hours worked. The intercept is also not significant, suggesting no significant baseline difference when all predictors are zero. The percentage of working hours lost has a negative but non-significant association with total weekly hours worked. The labour dependency ratio is significantly positive, indicating that higher ratios are linked to more hours worked. Similarly, employed female and employed male variables are both highly significant and positively associated with more hours worked.

Hypothesis Testing

To determine if the coefficients of the predictors are significantly different from zero, we perform hypothesis testing:

- Null Hypothesis (H_0): The coefficient is equal to zero (no effect).
- Alternative Hypothesis (H_1): The coefficient is not equal to zero (there is an effect).

We use the t-value and p-value to test these hypotheses. A p-value less than 0.05 indicates that we reject the null hypothesis and conclude that the coefficient is significantly different from zero. The significant predictors in the model include the **Labour Dependency Ratio** (p-value = 0.0432), **Employed Female** (p-value = 0.0000), and **Employed Male** (p-value = 0.0000). These results highlight that the labour dependency ratio, along with the employment rates for females and males, are key factors influencing total weekly hours worked during the COVID-19 pandemic.

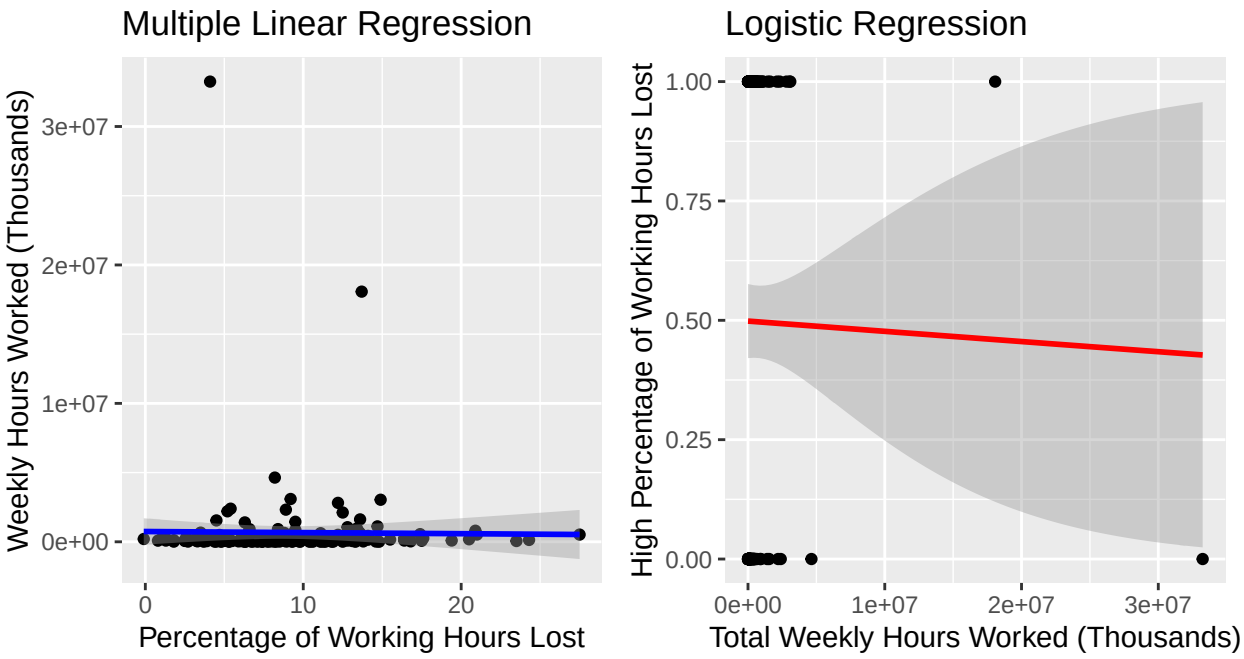

Figure 4: Regression Analysis: Multiple Linear vs Logistic Regression

Figure 4 shows the comparison of the result of multiple linear regression and logistic regression models using the relationship between the percentage of working hours lost and total weekly hours worked. The linear regression plot indicates a negative correlation, indicating that as the percentage of working hours lost increases as total weekly hours worked decreases. This is supported by the regression analysis, in which the estimate of working hours lost was negative. Conversely, however, the logistic regression plot shows that an increase in total weekly working hours is related to a decreased probability of having a high proportion of working hours lost (greater than the median). These plots together provide a complete picture of the relation between lost working hours and total weekly working hours, both analyzed continuously and as a binary.

Factor Analysis

Factor analysis was conducted to identify underlying relationships between variables by grouping them into factors, thereby reducing the dimensionality of the data while retaining most of the original variability.

The scree plot was obtained by Factor Analysis for Mixed Data and is utilized to decide on the correct number of factors to retain. The eigenvalues of each factor in decreasing order are plotted. The "elbow" point, where the plot flattens out, is the ideal number of factors. In this research, the scree plot showed an elbow after the second component, indicating that a large proportion of the variance is explained by the first two components. We therefore retained two components and hence simplified the analysis without losing the most significant information. This will enable us to concentrate on the most significant factors influencing employment patterns in the COVID-19 pandemic.

- **Factor 1: Economic Resilience** - This factor is primarily influenced by the labour dependency ratio and total weekly hours worked.
- **Factor 2: Gender Employment Disparity** - This factor is driven by the employment rates of females and males aged 25 and above.

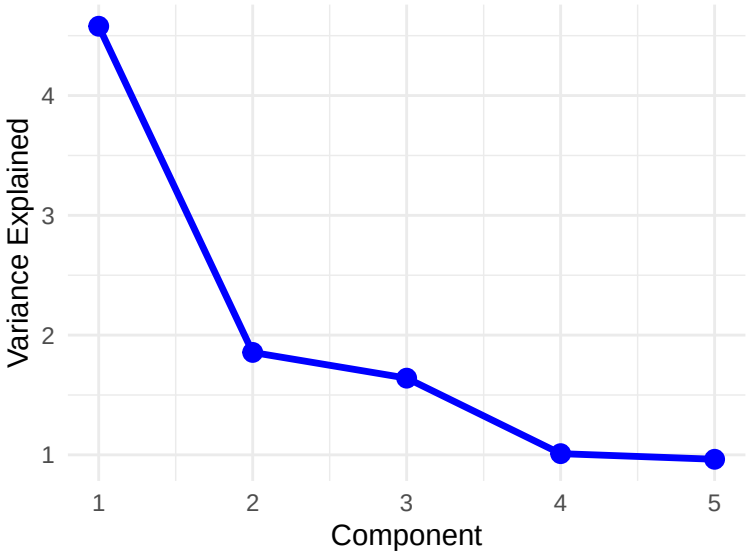

Figure 5: The Scree Plot

The Mclust analysis was conducted to identify the optimal number of components for clustering the data based on the Bayesian Information Criterion (BIC) values.

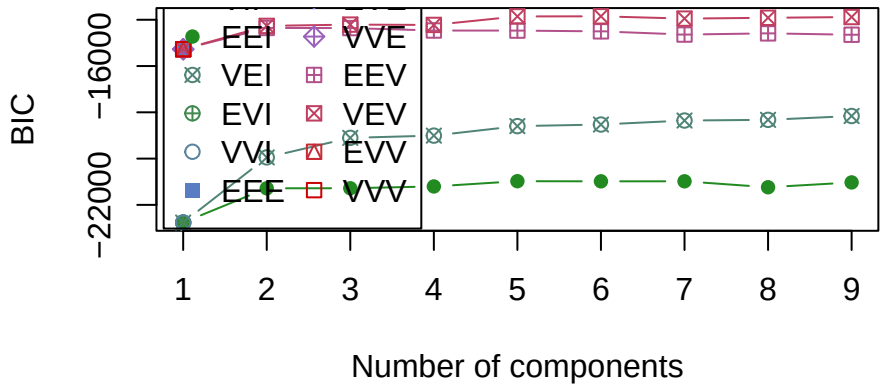

Figure 6: BIC Plot for Model Selection

The above BIC figure assists in determining the optimal model and the number of components that are best suited for the task of clustering. The graph indicates that the VEV model produces the minimum BIC value, which implies that it is the best fitting of the considered models. The plot also indicates that the 5-component point has the maximum BIC value, which is where the elbow point is located, after which the curve flattens. This indicates that the VEV model with 5 components offers a trade-off between goodness of fit and model complexity and is thus the most suitable model for clustering the data.

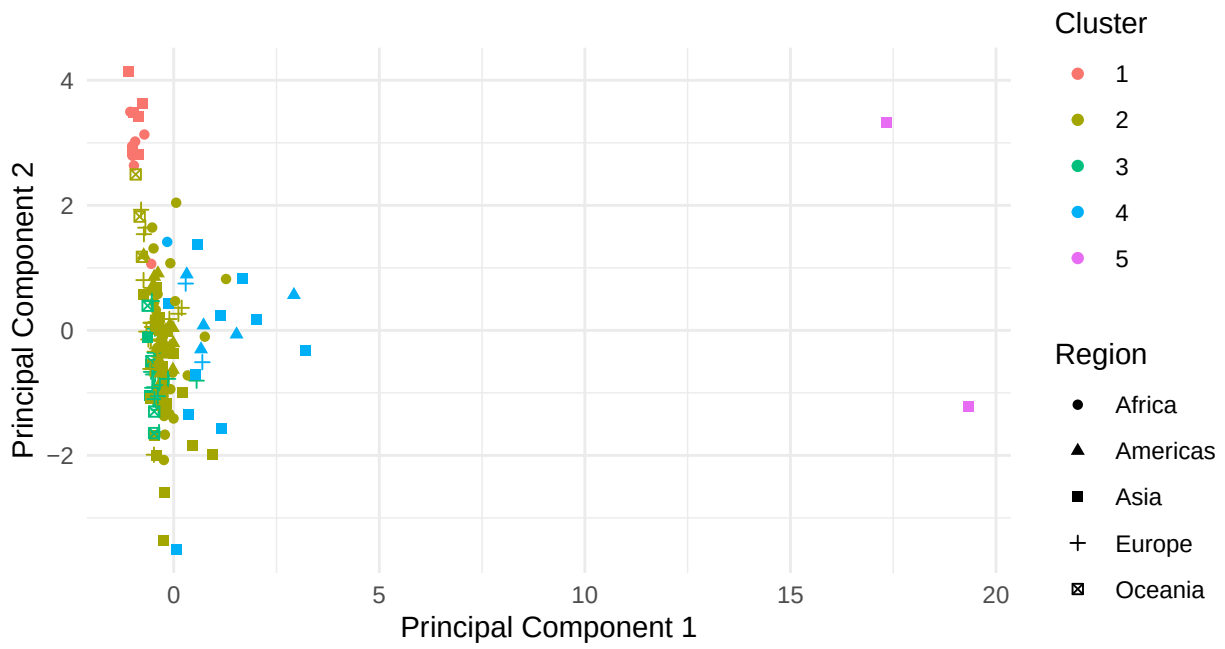

Figure 7: PCA of Data with Clustering, Categorized by Region

The plot above shows the relations between individuals in terms of two main dimensions, economic resilience and gender employment gap. PC1 is most influenced by the labour dependency ratio and total weekly working hours, and what is indicated here is that individuals from regions with greater labour dependency ratios and more resilient labour markets have greater scores on this dimension. PC2 is influenced by the employment of males and females aged 25 and older, indicating areas of high employment rates for both men and women. The graph splits individuals into groups according to region (Africa, Americas, Asia, Europe, Oceania), with the clear clusters indicative of regional patterns and differences in employment during the COVID-19 pandemic. This plot assists in determining the underlying determinants of the employment trends and makes regional variation in economic resilience and gender employment gap apparent. The clustering was achieved through the VEV model with 5 components, which assists in accounting for nearly all the variance within the data.

Structural Equation Modeling (SEM)

The Structural Equation Modeling (SEM) provides a comprehensive analysis of the relationships between observed and latent variables, focusing on the direct and indirect effects of demographic variables on employment status.

This is how the SEM looks like, with other variables that are correlating with each other not included:

- **Measurement Model:**
  - *Economic Resilience* is measured by the labour dependency ratio and total weekly hours worked.
  - *Gender Employment Disparity* is measured by the employment rates of females and males, and the percentage of working hours lost.
  - *Regional Employment Trends* is measured by the region, total weekly hours worked, and the percentage of working hours lost.
- **Structural Model:**

- Total weekly hours worked is influenced by Economic Resilience, Gender Employment Disparity, and Regional Employment Trends.
- Percentage of working hours lost is influenced by Economic Resilience, Gender Employment Disparity, and Regional Employment Trends.

The following are the classification of the SEM model:

- **Latent Exogenous Variables**: Economic Resilience, Gender Employment Disparity, Regional Employment Trends
- **Latent Endogenous Variables**: None (Not exactly cause I have not yet added the error terms in the model)
- **Manifest Exogenous Variables**: Labour Dependency Ratio, Employed Female, Employed Male, Region
- **Manifest Endogenous Variables**: Total Weekly Hours Worked, Percentage of Working Hours Lost

SEM analysis is not yet complete. Regarding the graphical representation of the SEM model illustrated above, it is important to note that error terms are still to be included in some of the variables. They encapsulate the measurement error and residual variance in observed variables, leading to a more specific and detailed explanation of the model. Following the inclusion of error terms, the model can now be fully executed and run within R to validate its structure and relationships, where hypothesis testing is conducted to validate if data covariance observed is indeed equal to the covariance described by the model. This will involve the testing of multiple metrics through SEM procedures to validate the trueness and validity of the model.

References:

- Bartik, A. W., Bertrand, M., Lin, F., Rothstein, J., & Unlu, F. (2020). How are women and racial minorities represented in the essential workforce? Evidence from the COVID-19 pandemic. *Brookings Institution*.
- Choi, E. K., Lee, S. H., & Lee, J. (2020). Gender inequality in COVID-19 labor markets. *OECD Economic Studies*.
- Liu, K., Chen, Y., Lin, R., & Han, D. (2020). COVID-19 in healthcare workers: A systematic review and meta-analysis. *International Journal of Nursing Studies*.
- OECD (2020). COVID-19 and the labour market: Impacts and policy responses. *OECD Policy Responses to Coronavirus (COVID-19)*.
- Izenman, A. J. (2013). *Modern Multivariate Statistical Techniques*. Springer.
- International Labour Organization (2023). Impact of COVID-19 on Employment. *ILOSTAT*. https://ilostat.ilo.org/

Appendix: R-code

```
knitr::opts_chunk$set(echo = TRUE, cache=T)

suppressPackageStartupMessages(library(dbscan))
library(ggplot2)
library(tibble)
library(cowplot)
library(mclust)  # Load the mclust library
library(cluster)
library(MASS)
library(dbscan)
library(ggplot2)
library(dplyr)
library(kableExtra)
library(readxl)
library(tidyr)
library(ggplot2)
library(knitr)
library(gridExtra)




dat=read.csv("Impact of Covid-19 on employment.csv", h=T)
dat=na.omit(dat)
dat <- dat[!duplicated(dat), ]




ggplot(dat, aes(x = total_weekly_hours_worked.estimates_in_thousands.,
                y = percentage_of_working_hrs_lost,
                color = Region)) +
  geom_point(alpha = 0.7) +  # Scatter plot with some transparency
  labs(title = "",
       x = "Weekly Hours Worked (Thousands)",
       y = "Working Hours Lost (%)") +
  theme_minimal() +  # Clean theme
  scale_color_brewer(palette = "Set1")  # Color palette for regions



par(mfrow=c(1, 2))
# Histograms: Distribution of continuous variables such as total hours worked and percentage of working
p1=ggplot(dat, aes(x = total_weekly_hours_worked.estimates_in_thousands.)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.7) +
  labs(title = "", x = "Weekly Hours Worked (Thousands)", y = "Frequency")

p2=ggplot(dat, aes(x = percentage_of_working_hrs_lost)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.7) +
  labs(title = "", x = "Percentage of Working Hours Lost", y = "Frequency")
```

```r
grid.arrange(p1, p2, ncol = 2)



# Descriptive Statistics: Means, medians, and standard deviations for continuous response variables
descriptive_stats_worked <- summary(dat$total_weekly_hours_worked.estimates_in_thousands.)
descriptive_stats_lost <- summary(dat$percentage_of_working_hrs_lost)

statistics=rbind(descriptive_stats_worked, descriptive_stats_lost)
rownames(statistics)=c("Weekly hours worked", "Weekly hours Lost (%)")

kable(round(statistics, 2), caption= "Descriptive statistics of the Response variables")%>%
  kable_styling()

# Load necessary libraries
 # For arranging multiple plots

# Reshape the data from wide to long format
dat_long <- dat %>%
  pivot_longer(cols = c("employed_female_25._2019", "employed_male_25._2019"),
               names_to = "gender",
               values_to = "employment_ratio") %>%
  mutate(gender = recode(gender,
                         "employed_male_25._2019" = "Male",
                         "employed_female_25._2019" = "Female"))

# Create the grouped bar plot
plot1 <- ggplot(dat_long, aes(x = Region, y = employment_ratio, fill = gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "",
    x = "Region",
    y = "Employment Ratio"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  guides(fill = guide_legend(title = "Gender",
                             labels = c("Male" = "Employed Males", "Female" = "Employed Females")))

# Bar Chart: Distribution of total weekly hours worked across different regions
plot2 <- ggplot(dat, aes(x = reorder(Region, total_weekly_hours_worked.estimates_in_thousands.), y = to
  geom_bar(stat = "identity") +
  labs(title = "", x = "Region", y = "Weekly Hours Worked")+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


# Arrange the plots side by side
grid.arrange(plot1, plot2, ncol = 2)



# Multiple Linear Regression: Model the relationship between total weekly hours worked and other predic
```

```r
model1 <- lm(total_weekly_hours_worked.estimates_in_thousands. ~ percentage_of_working_hrs_lost + labou
# summary(model1)

# Create a data frame using the provided values

model_results <- data.frame(
  Variable = c('Intercept',
               '% Working Hrs Lost',
               'Labour dependency ratio',
               'Employed Female',
               'Employed Male',
               'America',
               'Asia',
               'Europe',
               'Oceania'),
  Estimate = c(-5.246e+04,
               -3.291e+03,
               3.405e+04,
               5.612e+01,
               4.032e+01,
               -2.052e+04,
               2.527e+04,
               -3.787e+04,
               -4.606e+03),
  Std_Error = c(4.222e+04,
                2.445e+03,
                1.670e+04,
                9.741e-01,
                6.113e-01,
                3.606e+04,
                2.900e+04,
                2.857e+04,
                4.974e+04),
  t_value = c(-1.242,
              -1.346,
              2.039,
              57.619,
              65.957,
              -0.569,
              0.872,
              -1.326,
              -0.093),
  P_value = c(0.2160,
              0.1803,
              0.0432,
              2e-16,
              2e-16,
              0.5701,
              0.3847,
              0.1869,
              0.9263)
)
```

```r
kable(model_results, caption="Model Coefficients and Statistical Values", escape = TRUE) %>%
  kable_styling()
# Logistic Regression: Model the likelihood of being in a high percentage of working hours lost categor
# Create a binary variable for high percentage of working hours lost
dat1=dat
dat1$high_percentage_hours_lost <- ifelse(dat1$percentage_of_working_hrs_lost > median(dat1$percentage_

model2 <- glm(high_percentage_hours_lost ~ total_weekly_hours_worked.estimates_in_thousands. + labour_d
#summary(model2)
# Visualize the regression results
# Plot for Multiple Linear Regression
p1 = ggplot(dat, aes(x = percentage_of_working_hrs_lost, y = total_weekly_hours_worked.estimates_in_thou
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, col = "blue") +  # Explicit formula
  labs(title = "Multiple Linear Regression", x = "Percentage of Working Hours Lost", y = "Weekly Hours


# Plot for Logistic Regression
p2 = ggplot(dat1, aes(x = total_weekly_hours_worked.estimates_in_thousands., y = high_percentage_hours_
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), formula = y ~ x, col = "red") +
  labs(title = "Logistic Regression", x = "Total Weekly Hours Worked (Thousands)", y = "High Percentage

grid.arrange(p1, p2, ncol = 2)



library(FactoMineR)
library(factoextra)

dat$Region <- as.factor(dat$Region)
famd_result <- FAMD(dat[,-1], ncp = 5, graph = FALSE)

#par(mfrow=c(1, 2))

# Visualize the eigenvalues (scree plot)
#fviz_screeplot(famd_result, addlabels = TRUE, ylim = c(0, 40))

# plot(famd_result$eig[, 1], type = "b",
#      xlab = "Component", ylab = "Variance Explained",
#      pch = 19, col = "blue", lwd = 2)
#




eig_data <- data.frame(
  Component = 1:length(famd_result$eig[, 1]),
  Variance_Explained = famd_result$eig[, 1]
)

ggplot(eig_data, aes(x = Component, y = Variance_Explained)) +
  geom_point(size = 3, color = "blue") +  # Points
```

```
  geom_line(linewidth = 1.2, color = "blue") +  # Line connecting points
  labs(x = "Component",
       y = "Variance Explained") +
  theme_minimal()

# Visualize the factor map
#fviz_famd_ind(famd_result, repel = TRUE, habillage = dat$Region)

# Get the factor loadings

famd_result <- FAMD(dat[,-1], ncp = 2, graph = FALSE)

fviz_famd_ind(famd_result)

fviz_famd_var(famd_result)
fviz_famd_var(famd_result, repel = TRUE)



component_scores <- famd_result$ind$coord
plot(component_scores)
variable_contributions <- famd_result$var$contrib
rownames(variable_contributions)=c( 'Weekly HRS Worked', '% of Working Hrs Lost', "% of Working Hrs Lost


kable(round(variable_contributions,3), caption="Variables in the Factorial Analysis of Mixed Data Space
  kable_styling()



X <- data.matrix(dat[,c(-1)])
mod <- Mclust(X)
#summary(mod$BIC)
# Best BIC values:
#   VVE,3       EVE,4       VVE,4
# BIC      -6849.391 -6873.61648 -6885.47222
# BIC diff     0.000   -24.22499   -36.08073

# Since the EII and VII are super small with -27k value, exclude them and plot again

plot(mod, what = "BIC", ylim = range(mod$BIC[,-(1:2)],
                                     na.rm = TRUE),
       legendArgs = list(x = "bottomleft"))




mod1 <- Mclust(X, modelNames = "VEV", G = 5, x = mod$BIC)

x=summary(mod1)

log_likelihood <- x$loglik
```

```r
n <- x$n
df <- x$df
BIC <- x$bic
ICL = x$icl


# Create a dataframe with the extracted values
model_metrics <- data.frame(
  Log_Likelihood = log_likelihood,
  Number_of_Observations = n,
  Degrees_of_Freedom = df,
  BIC = BIC,
  ICL=ICL
)

colnames(model_metrics)=c("Log Likelihood", "Number of Observations", "Degrees of Freedom", "BIC", "ICL

# kable(model_metrics, caption="Model Evaluation Metrics")%>%
#   kable_styling()


# Assuming `dat` is your original data and `mod` is the result from Mclust
# Add the cluster assignments from Mclust to your data
dat2=dat
dat2$cluster <- mod$classification

# Plot the data points, colored by their cluster assignment
library(ggplot2)
pca <- prcomp(dat2[, c(-1, -10)], scale = TRUE)  # Assuming the categorical variable is the first colum

# Create a new data frame with PCA results, cluster assignments, and the categorical variable
pca_data <- data.frame(pca$x[, 1:2],
                       cluster = factor(mod1$classification),
                       Category = factor(dat2$Region))


ggplot(pca_data, aes(x = PC1, y = PC2, color = cluster, shape = Category)) +
  geom_point() +
  labs(title = "",
       x = "Principal Component 1", y = "Principal Component 2", color = "Cluster", shape = "Region") +
  theme_minimal() +
  guides(color = guide_legend(order = 1),
         shape = guide_legend(order = 2)) +
  theme(legend.position = "right",
        legend.direction = "vertical",
        legend.box = "vertical")



# Load necessary libraries
library(lavaan)
library(semPlot)
```

```
#
# # Ensure the 'Region' column is a factor
# dat$Region <- as.factor(dat$Region)
#
# # Select relevant columns for SEM
# df_sem <- dat %>%
#   select(total_weekly_hours_worked.estimates_in_thousands., percentage_of_working_hrs_lost,
#          percent_hours_lost_40hrs_per_week, percent_hours_lost_48hrs_per_week, labour_dependency_rati
#          employed_female_25._2019, employed_male_25._2019, ratio_of_weekly_hours_worked_by_population_
# df_sem[, -9]=scale(df_sem[,-9])
# df_sem$Region <- factor(df_sem$Region, ordered = TRUE)
#
#
# # Define the SEM model
# sem_model <- "
#   # Measurement model
#   EconomicResilience =~ labour_dependency_ratio + total_weekly_hours_worked.estimates_in_thousands.
#   GenderEmploymentDisparity =~ employed_female_25._2019 + employed_male_25._2019 + percentage_of_work
#   RegionalEmploymentTrends =~ Region + total_weekly_hours_worked.estimates_in_thousands. + percentage_
#
#   # Structural model
#   total_weekly_hours_worked.estimates_in_thousands. ~ EconomicResilience + GenderEmploymentDisparity +
#   percentage_of_working_hrs_lost ~ EconomicResilience + GenderEmploymentDisparity + RegionalEmploymen
#
#
# # Simplified SEM model
# sem_model <- "
#   # Measurement model
#   EconomicResilience =~ labour_dependency_ratio + total_weekly_hours_worked.estimates_in_thousands.
#
#   # Structural model
#   percentage_of_working_hrs_lost ~ EconomicResilience
#   total_weekly_hours_worked.estimates_in_thousands. ~ EconomicResilience
# "
#
#
# standardizedsolution(sem_model, type="std.all")
#
# # Fit the SEM model
# fit <- sem(sem_model, data = df_sem)
#
# fitmeasures(fit)
# Remove the highly correlated variables based on the correlation matrix
# df_sem_cleaned <- df_sem %>%
#   select(total_weekly_hours_worked.estimates_in_thousands.,
#          percentage_of_working_hrs_lost,
#          labour_dependency_ratio,
#          employed_female_25._2019,  # Keeping employed_female_25._2019
#          ratio_of_weekly_hours_worked_by_population_age_15.64,
#          Region)
#
# # Scale the continuous variables, excluding Region
# #df_sem_cleaned[, -ncol(df_sem_cleaned)] <- scale(df_sem_cleaned[, -ncol(df_sem_cleaned)])
```

```
#
#
# # Define the simplified SEM model
# sem_model <- "
#   # Measurement model
#   EconomicResilience =~ labour_dependency_ratio + total_weekly_hours_worked.estimates_in_thousands.
#
#   # Structural model
#   percentage_of_working_hrs_lost ~ EconomicResilience
#   total_weekly_hours_worked.estimates_in_thousands. ~ EconomicResilience
# "
#
# # Fit the SEM model
# fit <- sem(sem_model, data = df_sem_cleaned, ordered="Region")
#
# # Check the summary of the fit
# summary(fit, fit.measures = TRUE)

# Summary of the SEM model
#summary(fit, fit.measures = TRUE, standardized = TRUE)

# # Visualize the SEM model
# semPaths(fit, "std", layout = "tree", whatLabels = "std", edge.label.cex = 0.8, sizeMan = 5, sizeLat =
#
#
```