**7COM1079-0901-2024 - Team Research and Development Project**

**Final Report Tittle: Correlation Between Unemployment rate and Economic changes in Cities**

**Group ID:** A282

**Dataset number:** DS282

**Prepared by:** Mahesh Gangshettiwar, Student ID: 23096105

**Other Group Members:**  Vinay Gudepu, Student ID: 23067663

Karthik Krishnan, Student ID: 23035018

Arshad Abdulkadir, Student ID:23096759

Mujahid Ikram, Student ID: 23011500

**University of Hertfordshire**

**Hatfield, 2024**

Table of Contents

[1. Introduction 3](#_Toc186824198)

[1.1 Problem statement and research motivation 3](#_Toc186824199)

[1.2 The data set 3](#_Toc186824200)

[1.3 Research question 3](#_Toc186824201)

[1.4 Null hypothesis and alternative hypothesis 3](#_Toc186824202)

[2. Background research 4](#_Toc186824203)

[2.1 Research papers 4](#_Toc186824204)

[2.2 Why RQ is of interest 4](#_Toc186824205)

[3. Visualisation 5](#_Toc186824206)

[3.1 Appropriate plot for the RQ 5](#_Toc186824207)

[3.2 Additional information relating to understanding the data 7](#_Toc186824208)

[3.3 Useful information for the data understanding 7](#_Toc186824209)

[4. Analysis 8](#_Toc186824210)

[4.1 Statistical test used to test the hypotheses and output 8](#_Toc186824211)

[4.2 The null hypothesis is rejected /not rejected based on the p-value 8](#_Toc186824212)

[5. Evaluation – group’s experience at 7COM1079 8](#_Toc186824213)

[5.1 What went well 8](#_Toc186824214)

[5.2 Points for improvement 9](#_Toc186824215)

[5.3 Group’s time management 9](#_Toc186824216)

[5.4 Project’s overall judgement 9](#_Toc186824217)

[5.5 Comment on github log output 9](#_Toc186824218)

[6. Conclusions 10](#_Toc186824219)

[6.1 Result explained: 10](#_Toc186824220)

[6.2 Interpretation of outcome: 10](#_Toc186824221)

[6.3 Future work and limitation: 11](#_Toc186824222)

[7. Reference list 12](#_Toc186824223)

[8. Appendices 13](#_Toc186824224)

[A. R code used for analysis and visualisation 13](#_Toc186824225)

[B. GitHub log output. 14](#_Toc186824226)

# 1. Introduction

## 1.1 Problem statement and research motivation

This study has been reported by the growing concern of unemployment as an economic and social challenge in miscellaneous societies. Thus, in organizing efforts to decrease joblessness, it is achievable to note that important differences have stayed constant between various cities and states, as well as discussing different years. Through the examination of the causative factors between unemployment rates and elements including geographical region and time, this study aims to analyse typical characteristics of labour market status (Kurnia and Septiani, 2021). Cognitive awareness of such connections that policymakers may benefit from these insights may allow them to design remedial measures, establish fair employment options, and develop lasting economic recovery in depression-plagued areas of high unemployment rates.

## 1.2 The data set

The designated data source and the topic are meant to consider tendencies, patterns and factors that contribute to excessive unemployment rates giving the essential background of the economic problem. Sensing these dynamics is beneficial for policy-making, the degree of intervention in the regions that require job creation and otherwise. The independent variables in this study are city, state, year, and month, whereby the type of data is interval, which allows the analysis of material dynamics and regional differentiation. The dependent variable, the unemployment rate also allows in the interval data type as it is a continuous variable hence easier to estimate. With such variables associated, it is possible to resolve factors which charm unemployment and consequently, composition policies adjusting to labour market deficiencies.

## 1.3 Research question

What is the correlation between the unemployment rate and various factors such as city, state, and year?

## 1.4 Null hypothesis and alternative hypothesis

**Null hypothesis (H0):**

There doesn't appear to be a strong connection between the unemployment rate and factors like city, state, or year.

**Alternative hypothesis (H1):**

A clear correlation exists between the unemployment rate and factors like city, state, and year.

# 2. Background research

## 2.1 Research papers

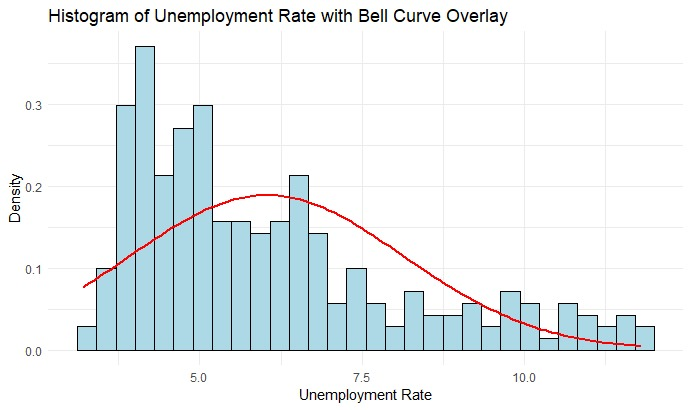
This paper estimates and analyses average unemployment rates for German cities and explores the influence of communities’ urban characteristics on unemployment (Heuermann and Vom Berge, 2024). On the other hand, the present study investigates the association of unemployment rates measured at the county level with individual-level health status, while regulating education, income, and region (Majeed *et al.,* 2023). As cited by Di Paolo and Ferrer‐i‐Carbonell, (2022) the tendency is being investigated by scholars in terms of how regional rates of unemployment affect the equation of life satisfaction, significantly when it comes to regional borders and local requirements. These studies demonstrate the value of the dataset where different social factors have corresponded to the unemployment rate of certain regions during separate periods of time.

## 2.2 Why RQ is of interest

From the review of the literature, it was found that there is a noteworthy gap in the literature regarding long-term quantities of regional economic approaches to unemployment in a variety of urban and pastoral regions. Similarly, although miscellaneous authors have explored the connection between unemployment and socio-economic factors, no significant intensity has been placed on understanding how these methods evolve with the help of new technologies and modifications in the world economy. Successive studies must incorporate real-time information; investigate relations between unemployment and rejuvenated forms of work; and consider the effectiveness of confident policies. Accordingly, further analysis of other regional variables like areal differentiation, urban and rural employment differences and their influence on unemployment will be more enlightening.

# 3. Visualisation

## 3.1 Appropriate plot for the RQ

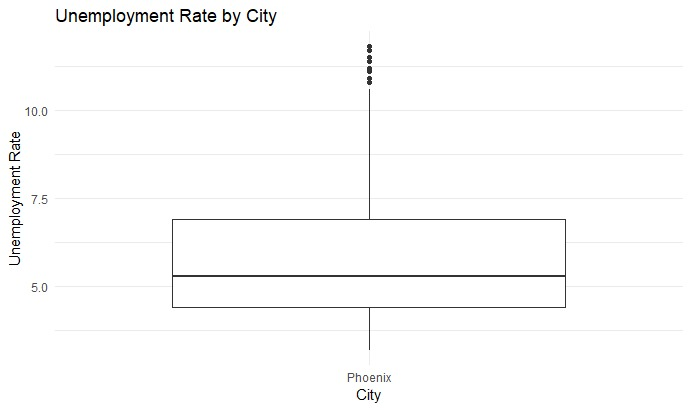
****

**Figure 1: Histogram of unemployment rate with bell curve overlay**

(Source: R-software)

This histogram shows the distribution of the unemployment rates The distribution of the data is roughly normal as shown by the bell curve (the red line). They used the chart to present the distribution of the employment rates whereby it indicated that most of the rates were in the

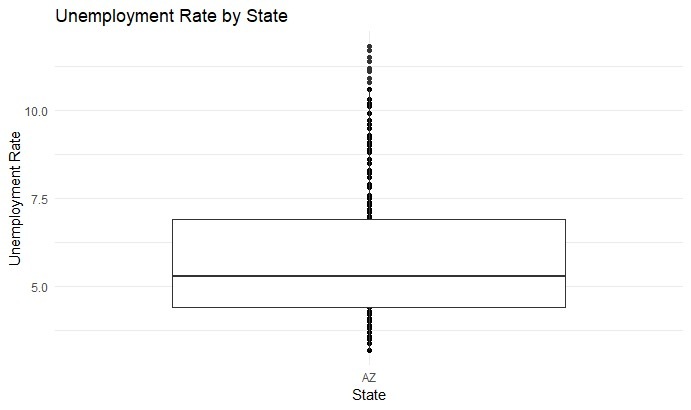
4-6% range and thereafter they decreased gradually as the dots got less dense and they indicated deviations from the normal rates.

****

**Figure 2: Unemployment rate by city**

(Source: R-software)

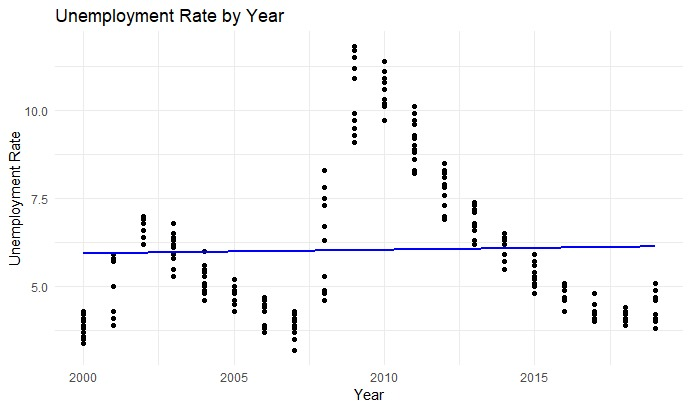
As can be observed from this boxplot below, it displays the distribution of the unemployment rate in Phoenix City. Thus, the central box of the figure exemplifies the IQR and the median is indicated by the horizontal line. Peculiarities are represented as a value higher than the upper whisker, where an asterisk denotes exceptional rates of unemployment. When it comes to local unemployment data variability and possible distinctions are illustrated in the plot.

****

**Figure 3: Unemployment rate by state**

(Source: R-software)

The following box plot demonstrates information on the unemployment rates of Arizona State abbreviated as AZ. The whisker expands to 1.5 times the IQR beyond this and there is an important horizontal line showing the median. There is much divergence in figures and many are dispersed below and above the whiskers which point out the high variability of unemployment rates throughout the condition.

****

**Figure 4: Unemployment rate by year**

(Source: R-software)

This graph displays the unemployment rate in terms of yearly data point computations for the years between 2000 and 2017. The blue line is the fitted trend line and overall unemployment has stayed reasonably stable, with points during the economic downturn of 2008-2010 and steady recovery post-2015 in tandem with rising GDP.

## 3.2 Additional information relating to understanding the data

These visualizations are desired to identify changes in the pattern of unemployment rate in Phoenix City and Arizona State between 2000 and 2017. Climactic insights show near symmetrical, normal distribution, local and state fluctuations, a dramatic decline during the economic 2008-2010 period and a gradual steady rise after 2015 which connects with GDP increases.

## 3.3 Useful information for the data understanding

Within these ones, norm distribution, local and state variations, and modifications during the 2008/2010 crisis are described for unemployment 2000-2017. Some of the insights received are; that most rates fall within 4–6%; abnormality or peculiarity which are regarded as exceptions; post-2015 upturn that corresponds with economic recovery; focus on past employment trends; and disparity between Arizona and Phoenix City.

# 4. Analysis

## 4.1 Statistical test used to test the hypotheses and output

**Regression analysis:**

Based on the dataset this analysis delivers key insights:

***Coefficients and p-value:***

Year: Coefficient: 0.0104, p-value: 0.667

Is\_phoenix: Coefficient: -14.7852, p-value: 0.760

***Model summary:***

The value is 0.001 which means a very small variance in unemployment rates.

The P-value is 0.667 which shows the model is not statistically significant.

## 4.2 The null hypothesis is rejected /not rejected based on the p-value

By directing to the regression results, the computed p values for year and city are 0.667 and 0.760 respectively which are significantly higher than the standard statistically significant level of 0.05. From this, the coefficients given under unemployment rate, year, and city are insignificant to give a base to the hypothesis. Moreover, F-statistics and p-value = 0.667 expose that the overall model does not significantly explain variations in the unemployment rate.

# 5. Evaluation – group’s experience at 7COM1079

## 5.1 What went well

The group was effective in a collaborative approach by observing close and timely communication, especially on improvement and goals of tasks. Concerning the third research question, team meetings and overall team discussions in particular allowed the participants to examine. GitHub also guaranteed easy control over versions and included individual contributions to a standard work without any problems. Also, the group conducted positive patterns when it came to problem-solving as well as problem-solving strategies were implemented to address technical concerns and where the group adhered to the timelines of the project. In general, the issue of coordination and efficient resource utilization was a compulsory factor in reaching landmarks within the group.

## 5.2 Points for improvement

However, some sections required advancements; otherwise, the results were satisfactory in the greater picture. There might have been better time division because at times there were some days of disorganization and some tasks were paused. More precisely, communication could be organised even more frequently, with regular meetings dedicated to progress tracking and possible blockers’ removal. The officials also know that technical documentation is an area where they could enhance as some parts are not detailed. Last but not least, the group could benefit from a better definition of the roles and responsibilities to help separate the work more reasonably and to dissuade any issues connected with the subject of the main responsibility for specific readings.

## 5.3 Group’s time management

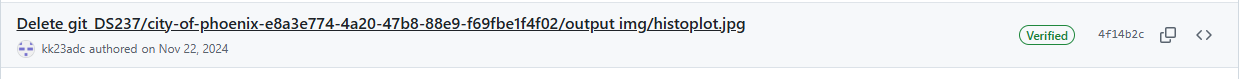
The group was able to endure time appropriately but in some instances, some were late due to a lack of understanding of deadlines and the problem of prioritization of tasks. For this, better organization of the project’s time and weekly Gantt charts should be operated in the subsequent assignments. Nevertheless, it made me confident that those important activities had to be processed on time and the team was able to complete the project.

## 5.4 Project’s overall judgement

The project was successfully completed to its preferred goals and it established good output. The group demonstrated all elements of teamwork, mastery of the technical skills involved and the ability to think creatively. There were also insignificant time management and communication issues which however could not be said to have dissembled the result drastically. In general, the project encounters the group’s commitment and work collaboration.

## 5.5 Comment on github log output

After the completion of Assignment 1 the members of the group were altered by the inclusion of a new member. Overall this inclusion was positive towards the capacity and knowledge within the team. Some transformations were made to the project plan to guarantee that all the project activities could be intercommunicated and assigned to the new member. There were no obvious complications in working in a group with the newly added member; this might have been due to the new member’s fast adjustment to the flow of operations. Since then we did not observe any other adaptations that occurred.

* 1. Comment on the GitHub log output
  2. ****

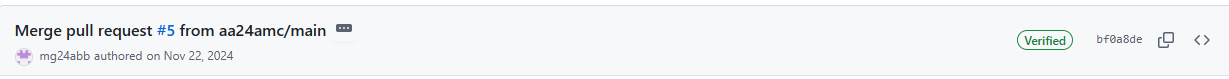
**Figure 5: Commit message 1**

(Source: Github)

* 1. ****

**Figure 6: Commit message 2**

(Source: Github)

* 1. ****

**Figure 7: Commit message 3**

(Source: Github)

The links to GitHub are given to present the latest commit and a pull request for the repository DS237\_unemployment. The fourth commit (4f14b2c) modified the code and is now potential for data processing or analysis connected with unemployment. The second commit organises more changes (495c69) to enhance the code even more. Interestingly, the pull request (number 5) seems to reverse these changes into the main branch meaning that some cooperation in refining or concluding the project concerning the analysis or research on the unemployment rate is going on.

# 6. Conclusions

## 6.1 Result explained:

The computed p values for year and city are 0.667 and 0.760 respectively which are significantly higher than the standard statistically significant level of 0.05.

## 6.2 Interpretation of outcome:

The overall model does not significantly explain variations in the unemployment rate.

## 6.3 Future work and limitation:

In the future this project helps to enhance developer productivity and the limitation of this project is that GitHub can be harder to work with and maintain the large dataset.

# 7. Reference list

Kurnia, R.E. and Septiani, Y., 2021. Social and economic factors determining the Unemployment rate in the Bregasmalang region 2010-2020. Eko-Regional: Jurnal Pembangunan Ekonomi Wilayah, 16(1).

Majeed, H., Baumann, S. and Majeed, H., 2023. Understanding the association between county-level unemployment and health stratified by education and income in the southwestern United States. Scientific Reports, 13(1), p.21988.

Di Paolo, A. and Ferrer‐i‐Carbonell, A., 2022. Regional borders, local unemployment, and life satisfaction. Journal of Regional Science, 62(2), pp.412-442.

Heuermann, D.F. and Vom Berge, P., 2024. Unemployment, Segregation, and the Structure of Cities (No. 17058). IZA Discussion Papers.

# Appendices

## R code used for analysis and visualisation

Analysis.

# Load necessary libraries

library(tidyverse)

library(ggplot2)

library(dplyr)

# Step 1: Load the dataset

file\_path <- "original/unemployment-rate-1.csv" # Replace with the path to your dataset

data <- read.csv(file\_path)

# Step 2: Data Cleaning

# Ensure column names are consistent and remove rows with missing values

cleaned\_data <- data %>%

rename\_all(tolower) %>%

filter(complete.cases(.))

# Convert relevant columns to appropriate data types

cleaned\_data$city <- as.factor(cleaned\_data$city)

cleaned\_data$state <- as.factor(cleaned\_data$state)

cleaned\_data$year <- as.numeric(cleaned\_data$year)

cleaned\_data$unemployment\_rate <- as.numeric(cleaned\_data$unemployment\_rate)

# View cleaned data

print("Cleaned Data Sample:")

print(head(cleaned\_data))

# Step 3: Histogram with bell curve overlay for unemployment\_rate

ggplot(cleaned\_data, aes(x = unemployment\_rate)) +

geom\_histogram(aes(y = ..density..), bins = 30, color = "black", fill = "lightblue") +

stat\_function(fun = dnorm, args = list(mean = mean(cleaned\_data$unemployment\_rate, na.rm = TRUE),

sd = sd(cleaned\_data$unemployment\_rate, na.rm = TRUE)),

color = "red", size = 1) +

labs(title = "Histogram of Unemployment Rate with Bell Curve Overlay",

x = "Unemployment Rate",

y = "Density") +

theme\_minimal()

# Step 4: Scatterplots for correlation

# Unemployment rate vs. City

ggplot(cleaned\_data, aes(x = city, y = unemployment\_rate)) +

geom\_boxplot() +

labs(title = "Unemployment Rate by City",

x = "City",

y = "Unemployment Rate") +

theme\_minimal()

# Unemployment rate vs. State

ggplot(cleaned\_data, aes(x = state, y = unemployment\_rate)) + geom\_point() + geom\_smooth(method = "lm", color = "blue", se = FALSE) +

labs(title = "Unemployment Rate by State",

x = "State",

y = "Unemployment Rate")

# Unemployment rate vs. Year

ggplot(cleaned\_data, aes(x = year, y = unemployment\_rate)) +

geom\_point() +

geom\_smooth(method = "lm", color = "blue", se = FALSE) +

labs(title = "Unemployment Rate by Year",

x = "Year",

y = "Unemployment Rate") +

theme\_minimal()

# Step 5: Calculate correlations

correlation <- cor(cleaned\_data$unemployment\_rate, as.numeric(cleaned\_data$year), use = "complete.obs")

print(paste("Correlation between unemployment rate and year:", round(correlation, 2)))

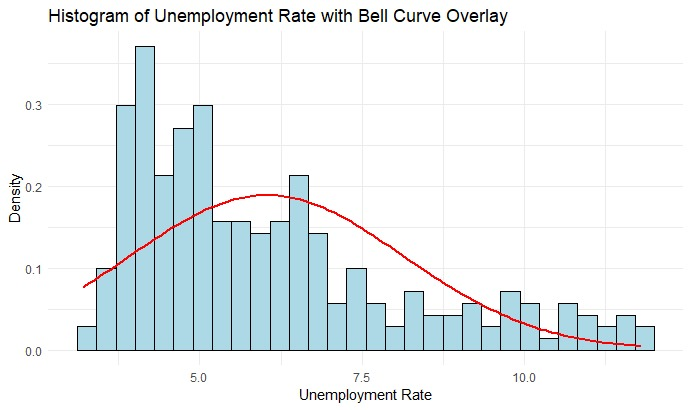
# Hypothesis testing

# Run correlation test for numeric variable (year)

cor\_test <- cor.test(cleaned\_data$unemployment\_rate, cleaned\_data$year, method = "pearson")

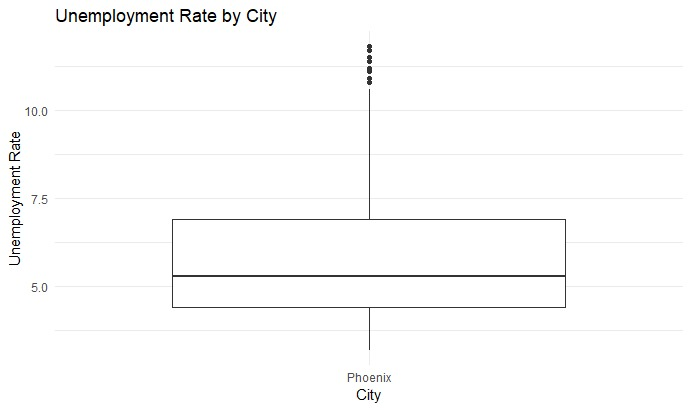
print(cor\_test)

## GitHub log output.

****

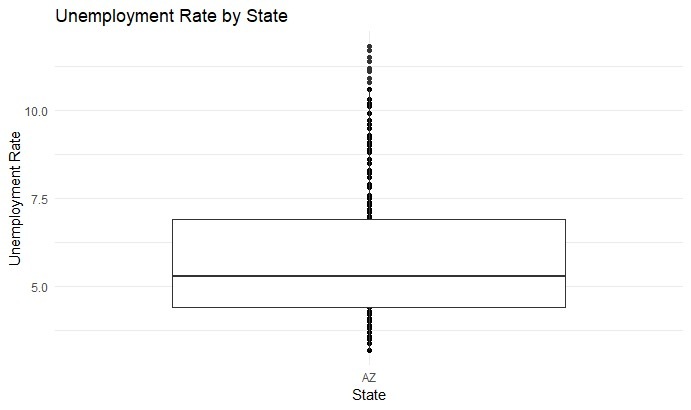
**Figure 1: Histogram of unemployment rate with bell curve overlay**

(Source: R-software)

****

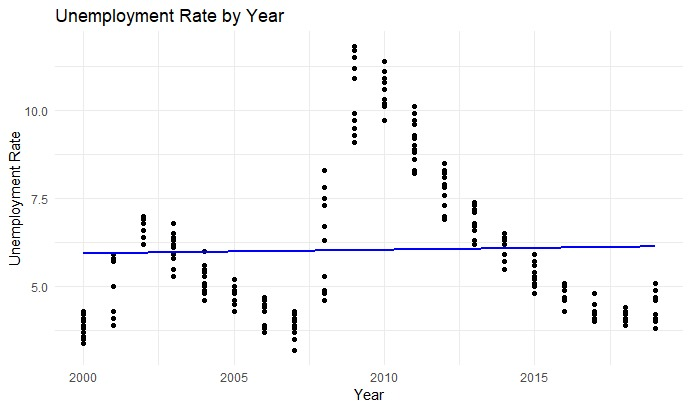
**Figure 2: Unemployment rate by city**

(Source: R-software)

****

**Figure 3: Unemployment rate by state**

(Source: R-software)

****

**Figure 4: Unemployment rate by year**

(Source: R-software)