PST PROJECT

SIMPHIWE MNGADI&LUNGISANI SIKHOSANA

2023-10-29

Introduction

The Quarterly Labour Survey (QLS) is a survey conducted regularly by statistical agencies or government bodies in many countries to collect data on key labor market indicators. The survey is typically conducted on a quarterly basis, hence its name.The purpose of the Quarterly Labour Force Survey (QLFS) is to collect and provide data on labor market indicators at the national level. It aims to measure and track the dynamics of employment, unemployment, and other related labor market indicators on a quarterly basis.The analysis for this report will focus on QLFS conducted by Statistics South Africa.R statistics software will be used to generates codes and analysis to give an insights on identifying any changes or fluctuations in employment and unemployment rates.

Part One(A brief explanation why we chose R sowftware over SAS)

we both chose R because,it is an open-source programming language, has gained popularity among statisticians and data scientists due to its flexibility and extensive range of statistical packages. It offers a wide variety of statistical techniques and data visualization capabilities, making it a powerful tool for exploratory data analysis. R's syntax is relatively straightforward and encourages experimentation and customization, allowing programmers to easily develop complex statistical models.One advantage of R is its active and vast online community. This community contributes to the continuous development of new packages and updates, ensuring that R remains at the forefront of statistical programming.The R community boasts an extensive ecosystem of user-contributed packages covering a wide range of domains, from machine learning and data mining to econometrics and genomics. These packages enable users to access and leverage state-of-the-art algorithms and statistical techniques easily. Additionally, the active and supportive R community ensures that these packages are well-maintained, regularly updated, and readily available.Another remarkable feature of R is its robust data visualization capabilities. The ggplot2 library is a shining example of the power and flexibility of R for creating stunning and informative visual representations of data. With just a few lines of code, users can generate beautiful and customizable plots, facilitating data exploration, analysis, and communication.In conclusion, R programming software is cherished by programmers and data scientists for its flexibility, extensive package ecosystem, powerful visualization capabilities, and collaborative nature. Its versatility makes it suitable for a wide range of applications and enables users to stay at the forefront of statistical techniques and data analysis. The love for R is firmly rooted in its ability to empower

individuals and organizations to extract valuable insights from data, further advancing the field of data science.

Part two

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
load("qlfs.RData") #load qlfs data file
nrow(qlfs)  # number of rows in a data
```

```
## [1] 2659961
```

```
ncol(qlfs)  # number of columns in a data
```

```
## [1] 161
```

Part three

```
set.seed(04)
# finding the `10% rows of the data
qlfs_sample <- qlfs %>%
  group_by(QDate) %>%
  slice_sample(prop = 0.1, replace = FALSE)
qlfs_sample
```

```
## # A tibble: 265,977 × 161
## # Groups:   QDate [40]
##     YEAR QUARTER QDate      PERSONNO Q12NIGHTS Q13GENDER Q15POPULATION
##    <int>   <int> <date>     <fct>    <fct>     <fct>     <fct>
##  1  2013       3 2013-09-30 4        Yes       Male      African/Black
##  2  2013       3 2013-09-30 2        Yes       Female    African/Black
##  3  2013       3 2013-09-30 2        Yes       Male      African/Black
##  4  2013       3 2013-09-30 5        Yes       Female    African/Black
##  5  2013       3 2013-09-30 4        Yes       Female    Coloured
##  6  2013       3 2013-09-30 7        Yes       Male      African/Black
##  7  2013       3 2013-09-30 1        Yes       Male      African/Black
```

```
##  8  2013       3 2013-09-30 1        Yes       Male      African/Black
##  9  2013       3 2013-09-30 1        Yes       Male      African/Black
## 10  2013       3 2013-09-30 8        Yes       Female    African/Black
## # ℹ 265,967 more rows
## # ℹ 154 more variables: Q16MARITALSTATUS <fct>, Q17EDUCATION <fct>,
## #   Q18FIELD <fct>, Q19ATTE <fct>, Q110EDUI <fct>, Q20SELFRESPOND <fct>,
## #   Q24APDWRK <fct>, Q24BOWNBUSNS <fct>, Q24CUNPDWRK <fct>, Q25APDWRK
<fct>,
## #   Q25BOWNBUSNS <fct>, Q25CUNPDWRK <fct>, Q27RSNABSENT <fct>, Q27ATIME
<fct>,
## #   Q27BRECPAY <fct>, Q31ALOOKWRK <fct>, Q31BSTARTBUSNS <fct>,
## #   Q31CTYPWRK <fct>, Q3201REGISTER <fct>, Q3202ENQUIRE <fct>, …
```

```
#  summary responses for each of the 40 Quaters
summary_responses <- qlfs_sample %>%
  count(QDate)
print(summary_responses)
```

```
## # A tibble: 40 × 2
## # Groups:   QDate [40]
##    QDate          n
##    <date>     <int>
##  1 2013-09-30  8705
##  2 2013-12-31  8752
##  3 2014-03-31  8699
##  4 2014-06-30  8474
##  5 2014-09-30  8530
##  6 2014-12-31  8427
##  7 2015-03-31  7256
##  8 2015-06-30  7200
##  9 2015-09-30  7186
## 10 2015-12-31  7034
## # ℹ 30 more rows
```

Part four

```
# Drop the missing values in Variable Province and NEET
qlfs_sample<-qlfs_sample|>
  drop_na(PROVINCE,NEET)

# Relative frequency table of NEET variable
qlfs_sample |>
  group_by(PROVINCE,NEET)|>
  summarise(neet_count= n())|>
  mutate(relative_frequency = (neet_count /sum(neet_count) ) * 100)|>
  filter(NEET=="Yes")->freq_neet
```

```
## `summarise()` has grouped output by 'PROVINCE'. You can override using the
## `.groups` argument.
```

```
freq_neet
```

```
## # A tibble: 9 × 4
## # Groups:   PROVINCE [9]
##   PROVINCE      NEET  neet_count relative_frequency
##   <fct>         <fct>      <int>              <dbl>
## 1 Western Cape  Yes         8186               40.7
## 2 Eastern Cape  Yes        12938               55.1
## 3 Northern Cape Yes         4385               54.0
## 4 Free State    Yes         5941               47.8
## 5 KwaZulu-Natal Yes        16382               50.4
## 6 North West    Yes         6267               53.5
## 7 Gauteng       Yes        17536               43.6
## 8 Mpumalanga    Yes         7704               48.1
## 9 Limpopo       Yes        10242               49.8
```

With 55.1% and 54.0%, respectively, the Eastern and Northern Cape had the highest relative frequencies of NEET persons. This implies that a substantial portion of the population in these provinces is not currently involved in education, employment, or training. These provinces' high NEET percentages may reflect economic and educational issues.

NEET rates in KwaZulu-Natal and North West are likewise relatively high, at 50.4% and 53.5%, respectively. These provinces may suffer similar economic and educational challenges, resulting in a sizable NEET population.

NEET rates in the Western Cape and Gauteng, which contain large urban centers such as Cape Town and Johannesburg, are lower, with 40.7% and 43.6%, respectively. These provinces may offer a broader range of career and educational possibilities, which may contribute to lower NEET rates.

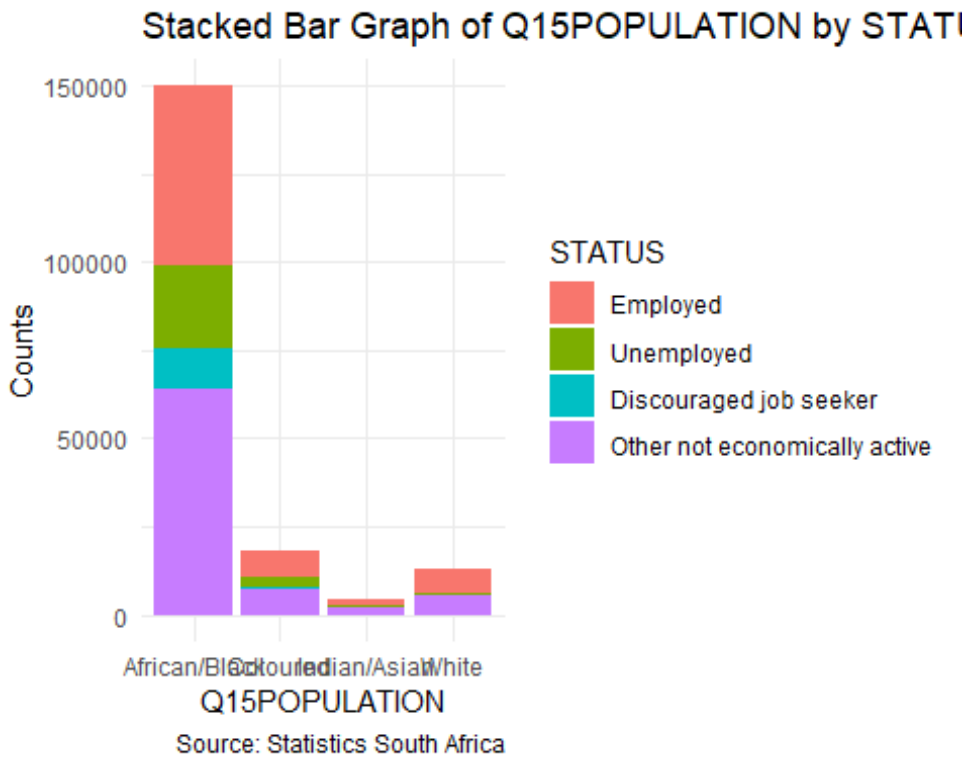
NEET rates in Mpumalanga and Limpopo are mild, with 48.1% and 49.8%, respectively. These provinces are in the country's northeast and may confront particular economic and educational issues.

With a NEET rate of 47.8%, Free State is in the middle of the provinces in terms of NEET rates.

Part Five

```
# drop the missing values in Q15POPULATION and STATUS variables
data_filtered<-qlfs_sample %>%
  drop_na(Q15POPULATION,STATUS)

# show the stacked bar graph of Q15POPULATION and STATUS
ggplot(data_filtered, aes(x = Q15POPULATION, fill = STATUS)) +
  geom_bar(position = "stack") +
  labs(
    x = "Q15POPULATION",  # X-axis label
```

```
    y = "Counts",          # Y-axis label
    title = "Stacked Bar Graph of Q15POPULATION by STATUS",
    caption = "Source: Statistics South Africa"
  ) +
  theme_minimal() +
  theme(legend.position = "right")
```



In the African/Black population there are more people who are not active economically and in comparison to other population group it seem to have a high proportion of unemployment followed by coloured, while both Indian/Asia and White have lower proportion of unemployment.

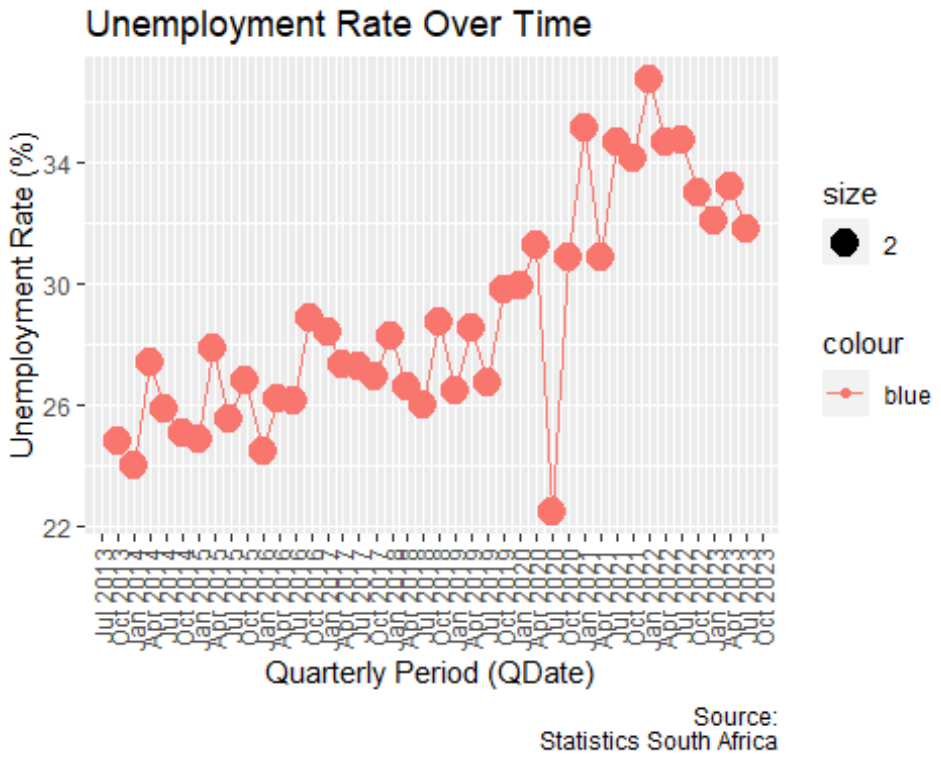
Part six

```
#drop missing values in status
qlfs_sample <- qlfs_sample|>
  drop_na(STATUS)
#frequency table of Status
qlfs_sample |>filter(STATUS %in% c("Employed", "Unemployed"))|>
  group_by(QDate,STATUS)|>
  summarize(Count = n()) |>
  mutate(rate = (Count / sum(Count)) * 100)|>
  filter(STATUS == "Unemployed")->freq_unemployed

## `summarise()` has grouped output by 'QDate'. You can override using the
## `.groups` argument.
```

```
freq_unemployed
```

```
## # A tibble: 40 × 4
## # Groups:   QDate [40]
##    QDate      STATUS     Count  rate
##    <date>     <fct>      <int> <dbl>
##  1 2013-09-30 Unemployed   748  24.8
##  2 2013-12-31 Unemployed   699  24.0
##  3 2014-03-31 Unemployed   823  27.4
##  4 2014-06-30 Unemployed   746  25.9
##  5 2014-09-30 Unemployed   726  25.1
##  6 2014-12-31 Unemployed   689  24.9
##  7 2015-03-31 Unemployed   728  27.9
##  8 2015-06-30 Unemployed   655  25.5
##  9 2015-09-30 Unemployed   714  26.8
## 10 2015-12-31 Unemployed   621  24.5
## # ℹ 30 more rows
```

```
#graph of unemployment rate over time
ggplot(freq_unemployed)+
  geom_line(mapping=aes(x=QDate,y=rate,
                        color="blue"))+
  geom_point(mapping=aes(x=QDate,y=rate,color="blue",size=2))+

  scale_x_date(date_breaks = "3 months", date_labels = "%b %Y") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  labs(
    x = "Quarterly Period (QDate)",
    y = "Unemployment Rate (%)",
    title = "Unemployment Rate Over Time",
    caption = "Source:
Statistics South Africa")
```



what is obviously notable in the graph of unemployment rate over time is that from second quarter of 2020 there was huge decrease of unemployment rate to 24.29% and then after third quarter unemployment rate increased extremely till the fourth quarter of 2021 where it was 36.76% and again it started to decrease in the first quarter of 2022 till fourth of 2023.

Part Seven

```
# unemployment rate per quarter
qlfs_sample %>%
  group_by(QDate, QUARTER, STATUS) %>%
  summarise(Frequency = n()) %>%
  mutate(Percent = (Frequency / sum(Frequency)) * 100) %>%
  filter(STATUS == "Unemployed")->qlfs_rate

## `summarise()` has grouped output by 'QDate', 'QUARTER'. You can override
using
## the `.groups` argument.

qlfs_rate

## # A tibble: 40 × 5
## # Groups:   QDate, QUARTER [40]
##    QDate      QUARTER STATUS     Frequency Percent
##    <date>       <int> <fct>          <int>   <dbl>
##  1 2013-09-30       3 Unemployed       748    12.3
```

```
##  2 2013-12-31       4 Unemployed       699    11.6
##  3 2014-03-31       1 Unemployed       823    13.5
##  4 2014-06-30       2 Unemployed       746    12.7
##  5 2014-09-30       3 Unemployed       726    12.3
##  6 2014-12-31       4 Unemployed       689    12.0
##  7 2015-03-31       1 Unemployed       728    14.7
##  8 2015-06-30       2 Unemployed       655    13.2
##  9 2015-09-30       3 Unemployed       714    14.4
## 10 2015-12-31       4 Unemployed       621    12.8
## # ℹ 30 more rows
```

```
library(readxl)

inflat <- read_excel("inflation.xlsx") # reading an inflation data from excel

#converting quarter from being an integer to factor
qlfs_rate$QUARTER <- as.factor(qlfs_rate$QUARTER)
qlfs_rate
```

```
## # A tibble: 40 × 5
## # Groups:   QDate, QUARTER [40]
##    QDate      QUARTER STATUS     Frequency Percent
##    <date>     <fct>   <fct>          <int>   <dbl>
##  1 2013-09-30 3       Unemployed       748    12.3
##  2 2013-12-31 4       Unemployed       699    11.6
##  3 2014-03-31 1       Unemployed       823    13.5
##  4 2014-06-30 2       Unemployed       746    12.7
##  5 2014-09-30 3       Unemployed       726    12.3
##  6 2014-12-31 4       Unemployed       689    12.0
##  7 2015-03-31 1       Unemployed       728    14.7
##  8 2015-06-30 2       Unemployed       655    13.2
##  9 2015-09-30 3       Unemployed       714    14.4
## 10 2015-12-31 4       Unemployed       621    12.8
## # ℹ 30 more rows
```

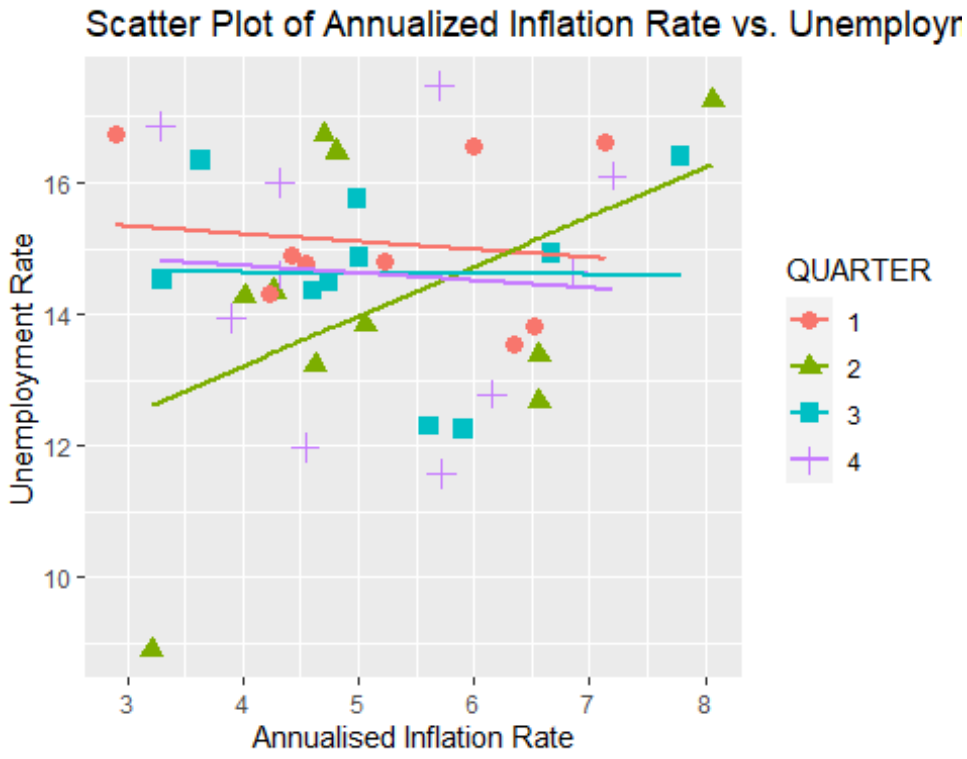
```
# combine two variables from different data sets
joined_data <- left_join(qlfs_rate, inflat, by = c("QDate" = "Date"))
joined_data
```

```
## # A tibble: 40 × 6
## # Groups:   QDate, QUARTER [40]
##    QDate               QUARTER STATUS     Frequency Percent
Annualised_Inflation
##    <dttm>              <fct>   <fct>          <int>   <dbl>
<dbl>
##  1 2013-09-30 00:00:00 3       Unemployed       748    12.3
5.62
##  2 2013-12-31 00:00:00 4       Unemployed       699    11.6
5.72
##  3 2014-03-31 00:00:00 1       Unemployed       823    13.5
6.35
```

```
##  4 2014-06-30 00:00:00 2       Unemployed       746    12.7
6.57
##  5 2014-09-30 00:00:00 3       Unemployed       726    12.3
5.91
##  6 2014-12-31 00:00:00 4       Unemployed       689    12.0
4.54
##  7 2015-03-31 00:00:00 1       Unemployed       728    14.7
4.55
##  8 2015-06-30 00:00:00 2       Unemployed       655    13.2
4.63
##  9 2015-09-30 00:00:00 3       Unemployed       714    14.4
4.61
## 10 2015-12-31 00:00:00 4       Unemployed       621    12.8
6.15
## # ℹ 30 more rows
```

```r
# scatter plot showing the correlation between Annualized inflation rate and
Unemployment rate
plot <- ggplot(joined_data, aes(x = Annualised_Inflation, y = Percent, shape
= QUARTER,color=QUARTER)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm",se=FALSE)+
  # Set axis labels and a title
  labs(
    x = "Annualised Inflation Rate",
    y = "Unemployment Rate",
    title = "Scatter Plot of Annualized Inflation Rate vs. Unemployment Rate"
  )

plot
```

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



In these results shown on the plot ,it is noticed that quarter 1 ,3 and 4 have a negative slope which implies that there is negative correlation between annualized rate and unemployment rate .This suggest that when the unemployment rate falls ,the inflation rate tends to rise.As the labor markets show low unemployment ,there is more upward pressure on wages,leading to higher production costs and potentially higher prices.In Quarter 2 shows the positive correlation and this implies that when the unemployment rate rises ,the inflation rate tends to rise as well,further more there is an outlier which shows the representation of low unemployment and low inflation rate which is often seen as healthy economic situation.

Part eight

```
# filtering data with both Quarter 1 and Year 2015
qlfs_sample|>

  filter(QUARTER==1 & year(QDate)==2015)|>
  select(matches("^Q419.*WRK$"))->dat

## Adding missing grouping variables: `QDate`

dat

## # A tibble: 4,943 × 8
## # Groups:   QDate [1]
##    QDate      Q419MONHRSWRK Q419TUEHRSWRK Q419WEDHRSWRK Q419THUHRSWRK
##    <date>             <dbl>         <dbl>         <dbl>         <dbl>
##  1 2015-03-31            NA            NA            NA            NA
```

```
##  2 2015-03-31            NA            NA            NA            NA
##  3 2015-03-31            NA            NA            NA            NA
##  4 2015-03-31            NA            NA            NA            NA
##  5 2015-03-31            NA            NA            NA            NA
##  6 2015-03-31            NA            NA            NA            NA
##  7 2015-03-31            11            11            11            11
##  8 2015-03-31             8             8             8             8
##  9 2015-03-31             9             9             9             9
## 10 2015-03-31            NA            NA            NA            NA
## # ℹ 4,933 more rows
## # ℹ 3 more variables: Q419FRIHRSWRK <dbl>, Q419SATHRSWRK <dbl>,
## #   Q419SUNHRSWRK <dbl>
```

```
#pivot data to put all selected columns in one column called Hours
pivoted_data <- dat %>%
  pivot_longer(cols = starts_with("Q419"), names_to = "WEEKDAY", values_to =
"HOURS")
pivoted_data
```

```
## # A tibble: 34,601 × 3
## # Groups:   QDate [1]
##    QDate      WEEKDAY       HOURS
##    <date>     <chr>         <dbl>
##  1 2015-03-31 Q419MONHRSWRK    NA
##  2 2015-03-31 Q419TUEHRSWRK    NA
##  3 2015-03-31 Q419WEDHRSWRK    NA
##  4 2015-03-31 Q419THUHRSWRK    NA
##  5 2015-03-31 Q419FRIHRSWRK    NA
##  6 2015-03-31 Q419SATHRSWRK    NA
##  7 2015-03-31 Q419SUNHRSWRK    NA
##  8 2015-03-31 Q419MONHRSWRK    NA
##  9 2015-03-31 Q419TUEHRSWRK    NA
## 10 2015-03-31 Q419WEDHRSWRK    NA
## # ℹ 34,591 more rows
```

```
# mean Hours per weekday
mean_hours_per_weekday <- pivoted_data %>%
  group_by(WEEKDAY) %>%
  summarise(mean_hours = mean(HOURS, na.rm = TRUE)) %>%
  arrange(desc(mean_hours))
mean_hours_per_weekday
```

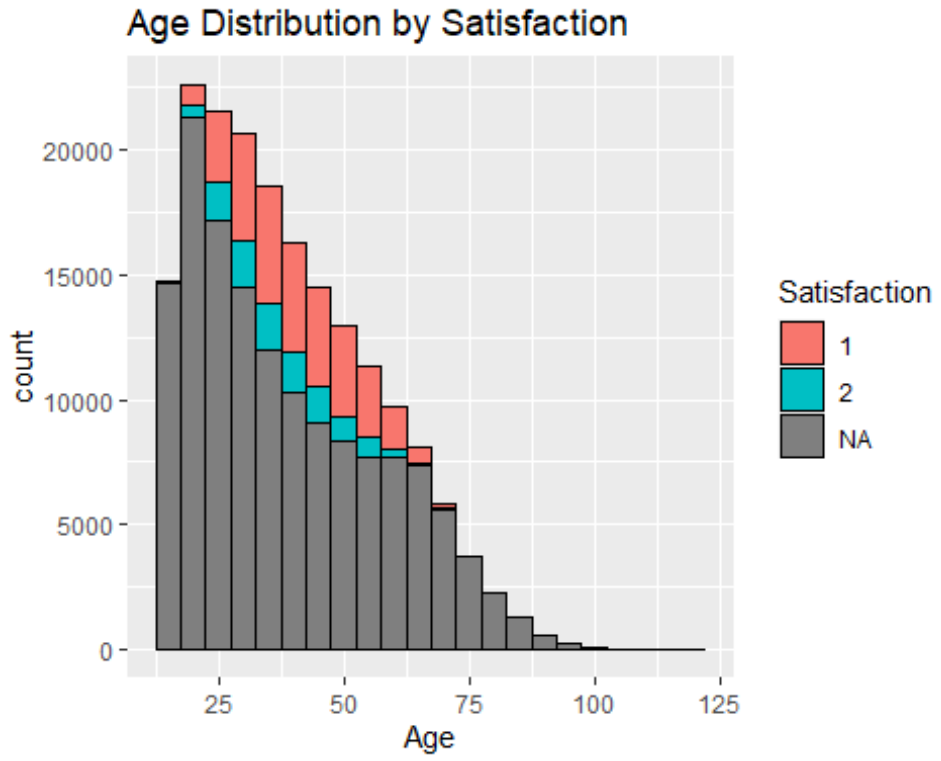
```
## # A tibble: 7 × 2
##   WEEKDAY       mean_hours
##   <chr>              <dbl>
## 1 Q419MONHRSWRK       7.74
## 2 Q419WEDHRSWRK       7.73
## 3 Q419TUEHRSWRK       7.68
## 4 Q419THUHRSWRK       7.64
## 5 Q419FRIHRSWRK       7.60
```

```
## 6 Q419SATHRSWRK       2.51
## 7 Q419SUNHRSWRK       1.21
```

The respondents worked most hours in weekday Q419WEDHRSWRK.

Part Nine

```
# histogram plot shows satisfaction per age group
histogram_plot <- ggplot(qlfs_sample, aes(x = Q14AGE, fill = Q416ASATISFIED))
+
  geom_histogram(binwidth = 5, color = "black") +

  labs(
    x = "Age",
    fill = "Satisfaction",
    title = "Age Distribution by Satisfaction"
  )
histogram_plot
```



Here most people in most age groups reserved their comment on satisfaction in their job(represented by NA on a plot).Again most people indicated that they are happy in their jobs in most age groups(represented by 1 in the plot) and few people said they are not satisfied with their job(represented by 2 in the plot).

Part Ten

```
# Preparing data for logistic regression
library(forcats)
qlfs_sample|>filter(STATUS %in% c("Employed", "Unemployed"))|>
  filter(Q14AGE >= 15 & Q14AGE <= 64)|>
  mutate(STATUS=fct_relevel( "Unemployed"))->dat3
dat3
```

```
## # A tibble: 93,010 × 161
## # Groups:   QDate [40]
##     YEAR QUARTER QDate      PERSONNO Q12NIGHTS Q13GENDER Q15POPULATION
##    <int>   <int> <date>     <fct>    <fct>     <fct>     <fct>
##  1  2013       3 2013-09-30 4        Yes       Female    Coloured
##  2  2013       3 2013-09-30 1        Yes       Male      African/Black
##  3  2013       3 2013-09-30 1        Yes       Male      African/Black
##  4  2013       3 2013-09-30 1        Yes       Male      African/Black
##  5  2013       3 2013-09-30 2        Yes       Female    White
##  6  2013       3 2013-09-30 1        Yes       Female    African/Black
##  7  2013       3 2013-09-30 2        Yes       Male      African/Black
##  8  2013       3 2013-09-30 1        Yes       Male      African/Black
##  9  2013       3 2013-09-30 1        Yes       Male      African/Black
## 10  2013       3 2013-09-30 4        Yes       Male      African/Black
## # ℹ 93,000 more rows
## # ℹ 154 more variables: Q16MARITALSTATUS <fct>, Q17EDUCATION <fct>,
## #   Q18FIELD <fct>, Q19ATTE <fct>, Q110EDUI <fct>, Q20SELFRESPOND <fct>,
## #   Q24APDWRK <fct>, Q24BOWNBUSNS <fct>, Q24CUNPDWRK <fct>, Q25APDWRK
<fct>,
## #   Q25BOWNBUSNS <fct>, Q25CUNPDWRK <fct>, Q27RSNABSENT <fct>, Q27ATIME
<fct>,
## #   Q27BRECPAY <fct>, Q31ALOOKWRK <fct>, Q31BSTARTBUSNS <fct>,
## #   Q31CTYPWRK <fct>, Q3201REGISTER <fct>, Q3202ENQUIRE <fct>, …
```

```
dat3$STATUS |>
  fct_relevel("Unemployed")->dat3$STATUS #Unemployed as reference level of
STATUS

dat3$EDUCATION_STATUS|>fct_lump( prop = 0.05, other_level = "Other")-
>dat3$EDUCATION_STATUS


dat3$GEO_TYPE|>fct_lump( n = 1, other_level = "other")->dat3$GEO_TYPE

dat3$GEO_TYPE|>fct_relevel("Urban")->dat3$GEO_TYPE

dat3$Q13GENDER|>fct_relevel("Female")->dat3$Q13GENDER

dat3$PROVINCE|>fct_relevel("Western Cape")->dat3$PROVINCE
```

```r
dat3|>
  group_by(QUARTER,QDate)
```

```
## # A tibble: 93,010 × 161
## # Groups:   QUARTER, QDate [40]
##     YEAR QUARTER QDate      PERSONNO Q12NIGHTS Q13GENDER Q15POPULATION
##    <int>   <int> <date>     <fct>    <fct>     <fct>     <fct>
##  1  2013       3 2013-09-30 4        Yes       Female    Coloured
##  2  2013       3 2013-09-30 1        Yes       Male      African/Black
##  3  2013       3 2013-09-30 1        Yes       Male      African/Black
##  4  2013       3 2013-09-30 1        Yes       Male      African/Black
##  5  2013       3 2013-09-30 2        Yes       Female    White
##  6  2013       3 2013-09-30 1        Yes       Female    African/Black
##  7  2013       3 2013-09-30 2        Yes       Male      African/Black
##  8  2013       3 2013-09-30 1        Yes       Male      African/Black
##  9  2013       3 2013-09-30 1        Yes       Male      African/Black
## 10  2013       3 2013-09-30 4        Yes       Male      African/Black
## # ℹ 93,000 more rows
## # ℹ 154 more variables: Q16MARITALSTATUS <fct>, Q17EDUCATION <fct>,
## #   Q18FIELD <fct>, Q19ATTE <fct>, Q110EDUI <fct>, Q20SELFRESPOND <fct>,
## #   Q24APDWRK <fct>, Q24BOWNBUSNS <fct>, Q24CUNPDWRK <fct>, Q25APDWRK
<fct>,
## #   Q25BOWNBUSNS <fct>, Q25CUNPDWRK <fct>, Q27RSNABSENT <fct>, Q27ATIME
<fct>,
## #   Q27BRECPAY <fct>, Q31ALOOKWRK <fct>, Q31BSTARTBUSNS <fct>,
## #   Q31CTYPWRK <fct>, Q3201REGISTER <fct>, Q3202ENQUIRE <fct>, …
```

```r
periods <- c("Q3 2013", "Q4 2013", "Q1 2014", "Q2 2014", "Q3 2014", "Q4
2014", "Q1 2015", "Q2 2015", "Q3 2015", "Q4 2015",
             "Q1 2016", "Q2 2016", "Q3 2016", "Q4 2016", "Q1 2017", "Q2
2017", "Q3 2017", "Q4 2017", "Q1 2018", "Q2 2018",
             "Q3 2018", "Q4 2018", "Q1 2019", "Q2 2019", "Q3 2019", "Q4
2019", "Q1 2020", "Q2 2020", "Q3 2020", "Q4 2020",
             "Q1 2021", "Q2 2021", "Q3 2021", "Q4 2021", "Q1 2022", "Q2
2022", "Q3 2022", "Q4 2022", "Q1 2023", "Q2 2023")

  # Convert QDATE to Date format
library(lubridate)
dat3$QDate <- as.Date(dat3$QDate, format = "%Y-%m-%d")

# Extract year and quarter from QDATE
dat3$YEAR <- lubridate::year(dat3$QDate)
dat3$QUARTER <- lubridate::quarter(dat3$QDate)

# Create PERIOD variable using YEAR and QUARTER
dat3$PERIOD <- match(paste0("Q", dat3$QUARTER, " ", dat3$YEAR), periods)

library(broom)
library(glm2)
```

```r
logistic_model <- glm(STATUS ~ Q14AGE + EDUCATION_STATUS + Q13GENDER +
PROVINCE +GEO_TYPE + PERIOD, data = dat3, family = "binomial")

## Warning: glm.fit: algorithm did not converge

tidy_output <- tidy(logistic_model)
print(tidy_output)

## # A tibble: 17 × 5
##    term                                    estimate std.error statistic
p.value
##    <chr>                                      <dbl>     <dbl>     <dbl>
<dbl>
##  1 (Intercept)                            -2.66e+ 1     7777. -3.42e- 3
0.997
##  2 Q14AGE                                 -5.41e-16      111. -4.88e-18
1
##  3 EDUCATION_STATUSSecondary not completed -2.07e-15     5066. -4.09e-19
1
##  4 EDUCATION_STATUSSecondary completed      8.89e-15     5204.  1.71e-18
1
##  5 EDUCATION_STATUSTertiary                 1.33e-16     5533.  2.41e-20
1
##  6 EDUCATION_STATUSOther                  -1.39e-16     6408. -2.17e-20
1
##  7 Q13GENDERMale                          -5.38e-15     2352. -2.29e-18
1
##  8 PROVINCEEastern Cape                   -2.87e-14     4909. -5.85e-18
1
##  9 PROVINCENorthern Cape                  -3.18e-14     6721. -4.74e-18
1
## 10 PROVINCEFree State                     -3.01e-14     5453. -5.51e-18
1
## 11 PROVINCEKwaZulu-Natal                  -3.00e-14     4576. -6.56e-18
1
## 12 PROVINCENorth West                     -2.85e-14     6170. -4.63e-18
1
## 13 PROVINCEGauteng                        -2.77e-14     3935. -7.04e-18
1
## 14 PROVINCEMpumalanga                     -2.79e-14     5279. -5.29e-18
1
## 15 PROVINCELimpopo                        -3.01e-14     5603. -5.38e-18
1
## 16 GEO_TYPEother                          -5.75e-15     3192. -1.80e-18
1
## 17 PERIOD                                 -5.09e-16      102. -4.99e-18
1
```

(Intercept): The intercept, often represented as (Intercept), is the estimated log-odds of being unemployed when all other predictor variables are zero. In this case, the estimate is

approximately -2.66e+1. This estimate represents the log-odds of being unemployed for the reference categories of the categorical variables.

Q14AGE: The coefficient estimate for Q14AGE is approximately -5.41e-16. This suggests that for a one-unit change in Q14AGE, the log-odds of being unemployed change by this amount. However, the p-value is very high (1), indicating that Q14AGE is not statistically significant in predicting employment status. It might be considered not relevant in this model.

EDUCATION_STATUS: There are multiple levels for EDUCATION_STATUS (e.g., secondary not completed, secondary completed, tertiary, other). Each coefficient estimate represents the change in log-odds of being unemployed when compared to the reference category "Other." All p-values for these variables are very high (1), suggesting that none of the education status levels are statistically significant in predicting employment status.

Q13GENDER: The coefficient estimate for Q13GENDER (Male) is approximately -5.38e-15, which suggests that being male, when compared to the reference category "Female," results in a change in the log-odds of being unemployed. However, the p-value is very high (1), indicating that gender is not statistically significant in predicting employment status.

PROVINCE: There are multiple levels for PROVINCE (e.g., Eastern Cape, Northern Cape, etc.). Each coefficient estimate represents the change in log-odds of being unemployed when compared to the reference category "Western Cape." All p-values for these variables are very high (1), suggesting that none of the provinces are statistically significant in predicting employment status.

GEO_TYPEother: The coefficient estimate for GEO_TYPEother is approximately -5.75e-15. This suggests that having a different geo-type (other than the reference level) results in a change in the log-odds of being unemployed. However, the p-value is very high (1), indicating that geo-type is not statistically significant in predicting employment status.

PERIOD: The coefficient estimate for PERIOD is approximately -5.09e-16. This variable appears to represent the time period of the data. The p-value is very high (1), suggesting that the time period is not statistically significant in predicting employment status.

all in all, based on the provided results, none of the predictor variables (age, education status, gender, province, geo-type, or time period) appear to be statistically significant in predicting employment status. This may suggest that the logistic regression model with these variables does not provide a strong fit for the data, or it may indicate that additional relevant factors are missing from the model.