Project_FinalReport

Sumanth Mungi

2022-12-06

The dataset that I have chosen is "Quarterly_Census_of_Employment_and_Wages_QCEW_" from the California State Government's "Open Data Portal" website. The link to the dataset is https://data.ca.gov/dataset/quarterly-census-of-employment-and-wages-qcew/resource/efdcc006-bcaf-4066-a763-aef58514a7dd

The Quarterly Census of Employment and Wages (QCEW) Program is a Federal-State cooperative program between the U.S. Department of Labor's Bureau of Labor Statistics (BLS) and the California EDD's Labor Market Information Division (LMID). The QCEW program produces a comprehensive tabulation of employment and wage information for workers covered by California Unemployment Insurance (UI) laws and Federal workers covered by the Unemployment Compensation for Federal Employees (UCFE) program.

1. Perform checks to determine the quality of the data (missing values, outliers, etc.)

```
library(readr)
rawdata <-
read.csv("C:/Users/Checkout/Downloads/Quarterly_Census_of_Employment_and_Wages__QCEW_.csv")
```

I assigned the selected dataset to the 'rawdata' so as to not risk any changes to the original dataset.

```
library(psych)
dim(rawdata)

## [1] 4555904      15

head(rawdata)

##   Area.Type      Area.Name Year Quarter        Ownership NAICS.Level
NAICS.Code
## 1    County Alameda County 2004 1st Qtr State Government           5
92212
## 2    County Alameda County 2004 1st Qtr          Private           6
325612
## 3    County Alameda County 2004 1st Qtr          Private           5
33341
## 4    County Alameda County 2004 1st Qtr          Private           4
4441
## 5    County Alameda County 2004 1st Qtr          Private           6
711130
## 6    County Alameda County 2004 1st Qtr          Private           2
```

```
1027
##                             Industry.Name Establishments
## 1                          Police Protection             10
## 2 Polish and Sanitation Good Manufacturing              7
## 3  HVAC and Commercial Refrigeration Equip             15
## 4   Building Material and Supplies Dealers            225
## 5               Musical Groups and Artists             35
## 6                            Other Services          15655
##   Average.Monthly.Employment X1st.Month.Emp X2nd.Month.Emp X3rd.Month.Emp
## 1                        354            356            353            353
## 2                        638            636            639            641
## 3                        354            356            354            352
## 4                       5172           5163           5132           5221
## 5                        306            252            261            406
## 6                      29553          29158          29533          29969
##   Total.Wages..All.Workers. Average.Weekly.Wages
## 1                   2532357                  550
## 2                  53376354                 6429
## 3                   7828534                 1701
## 4                  42069506                  626
## 5                   1053787                  265
## 6                 201730468                  525

#describe(rawdata)

summary(rawdata)

##   Area.Type          Area.Name              Year        Quarter
##  Length:4555904     Length:4555904     Min.   :2004   Length:4555904
##  Class :character   Class :character   1st Qu.:2008   Class :character
##  Mode  :character   Mode  :character   Median :2012   Mode  :character
##                                        Mean   :2012
##                                        3rd Qu.:2017
##                                        Max.   :2021
##
##   Ownership          NAICS.Level     NAICS.Code        Industry.Name
##  Length:4555904     Min.   :0.00    Length:4555904     Length:4555904
##  Class :character   1st Qu.:4.00    Class :character   Class :character
##  Mode  :character   Median :5.00    Mode  :character   Mode  :character
##                     Mean   :4.75
##                     3rd Qu.:6.00
##                     Max.   :6.00
##                     NA's   :50296
##  Establishments     Average.Monthly.Employment X1st.Month.Emp
##  Min.   :       0   Min.   :        0          Min.   :        0
##  1st Qu.:       7   1st Qu.:       69          1st Qu.:       12
##  Median :      20   Median :      288          Median :      143
##  Mean   :    2322   Mean   :    32731          Mean   :    26048
##  3rd Qu.:      87   3rd Qu.:     1525          3rd Qu.:      922
##  Max.   :11178274   Max.   :149931099          Max.   :149527674
```

```
##
##  X2nd.Month.Emp      X3rd.Month.Emp      Total.Wages..All.Workers.
##  Min.   :        0   Min.   :        0   Min.   :0.000e+00
##  1st Qu.:       12   1st Qu.:       12   1st Qu.:8.442e+05
##  Median :      143   Median :      144   Median :4.149e+06
##  Mean   :    26195   Mean   :    26311   Mean   :6.861e+08
##  3rd Qu.:      925   3rd Qu.:      926   3rd Qu.:2.577e+07
##  Max.   :150260321   Max.   :150005303   Max.   :9.721e+12
##
##  Average.Weekly.Wages
##  Min.   :     0.0
##  1st Qu.:   565.0
##  Median :   841.0
##  Mean   :   967.9
##  3rd Qu.:  1195.0
##  Max.   :105149.0
##

is.data.frame(rawdata)

## [1] TRUE
```

From the above results, we can see that this is a very big dataset, and has multiple variables, that in many ways influence each other. There are 15 variables in the dataset - some categorical and some continuous numeric variables.

rows with missing values

```
sum(is.na(rawdata))

## [1] 50296
```

columns with missing values

```
colSums(is.na(rawdata))

##                   Area.Type                   Area.Name
##                           0                           0
##                        Year                     Quarter
##                           0                           0
##                   Ownership                 NAICS.Level
##                           0                       50296
##                  NAICS.Code               Industry.Name
##                           0                           0
##             Establishments Average.Monthly.Employment
##                           0                           0
##             X1st.Month.Emp              X2nd.Month.Emp
##                           0                           0
##             X3rd.Month.Emp  Total.Wages..All.Workers.
##                           0                           0
##       Average.Weekly.Wages
##                           0
```

```
cleaned_data <- na.omit(rawdata)
dim(cleaned_data)

## [1] 4505608      15

clean_data <- cleaned_data
```

On removing the rows with missing values, we get the above dimensions of the data and the dataset is still huge.

```
summary(clean_data)

##   Area.Type          Area.Name              Year        Quarter
##  Length:4505608     Length:4505608     Min.   :2004   Length:4505608
##  Class :character   Class :character   1st Qu.:2008   Class :character
##  Mode  :character   Mode  :character   Median :2013   Mode  :character
##                                        Mean   :2013
##                                        3rd Qu.:2017
##                                        Max.   :2021
##   Ownership          NAICS.Level     NAICS.Code        Industry.Name
##  Length:4505608     Min.   :0.000   Length:4505608     Length:4505608
##  Class :character   1st Qu.:4.000   Class :character   Class :character
##  Mode  :character   Median :5.000   Mode  :character   Mode  :character
##                     Mean   :4.747
##                     3rd Qu.:6.000
##                     Max.   :6.000
##  Establishments     Average.Monthly.Employment X1st.Month.Emp
##  Min.   :       0   Min.   :        0          Min.   :        0
##  1st Qu.:       7   1st Qu.:       69          1st Qu.:       11
##  Median :      20   Median :      288          Median :      141
##  Mean   :    2323   Mean   :    32745          Mean   :    25989
##  3rd Qu.:      87   3rd Qu.:     1525          3rd Qu.:      916
##  Max.   :11178274   Max.   :149931099          Max.   :149527674
##  X2nd.Month.Emp      X3rd.Month.Emp      Total.Wages..All.Workers.
##  Min.   :        0   Min.   :        0   Min.   :0.000e+00
##  1st Qu.:       12   1st Qu.:       12   1st Qu.:8.474e+05
##  Median :      142   Median :      142   Median :4.168e+06
##  Mean   :    26135   Mean   :    26251   Mean   :6.896e+08
##  3rd Qu.:      919   3rd Qu.:      920   3rd Qu.:2.589e+07
##  Max.   :150260321   Max.   :150005303   Max.   :9.721e+12
##  Average.Weekly.Wages
##  Min.   :     0.0
##  1st Qu.:   566.0
##  Median :   842.0
##  Mean   :   969.2
##  3rd Qu.:  1196.0
##  Max.   :105149.0

head(clean_data)
```

```
##   Area.Type      Area.Name Year Quarter        Ownership NAICS.Level
NAICS.Code
## 1    County Alameda County 2004 1st Qtr State Government           5
92212
## 2    County Alameda County 2004 1st Qtr          Private           6
325612
## 3    County Alameda County 2004 1st Qtr          Private           5
33341
## 4    County Alameda County 2004 1st Qtr          Private           4
4441
## 5    County Alameda County 2004 1st Qtr          Private           6
711130
## 6    County Alameda County 2004 1st Qtr          Private           2
1027
##                             Industry.Name Establishments
## 1                          Police Protection             10
## 2 Polish and Sanitation Good Manufacturing              7
## 3  HVAC and Commercial Refrigeration Equip             15
## 4   Building Material and Supplies Dealers            225
## 5               Musical Groups and Artists             35
## 6                           Other Services          15655
##   Average.Monthly.Employment X1st.Month.Emp X2nd.Month.Emp X3rd.Month.Emp
## 1                        354            356            353            353
## 2                        638            636            639            641
## 3                        354            356            354            352
## 4                       5172           5163           5132           5221
## 5                        306            252            261            406
## 6                      29553          29158          29533          29969
##   Total.Wages..All.Workers. Average.Weekly.Wages
## 1                   2532357                  550
## 2                  53376354                 6429
## 3                   7828534                 1701
## 4                  42069506                  626
## 5                   1053787                  265
## 6                 201730468                  525
```

2. Proposal on what questions you are interested in answering from the data.

This dataset looks fascinating to me. It gives us a lot of information about the wages of employees in different industries across California. I could deduce where the wages of people are less and where they are more.

If I were an investor and I am given this data, I would understand where and in which industries to put my investments in for more returns. If I am the Governor of California, this dataset would be of tremendous help to me. I could direct the social welfare schemes to cater the necessary people, invest better in the public transport system in the State, frame new and better social welfare schemes, etc.

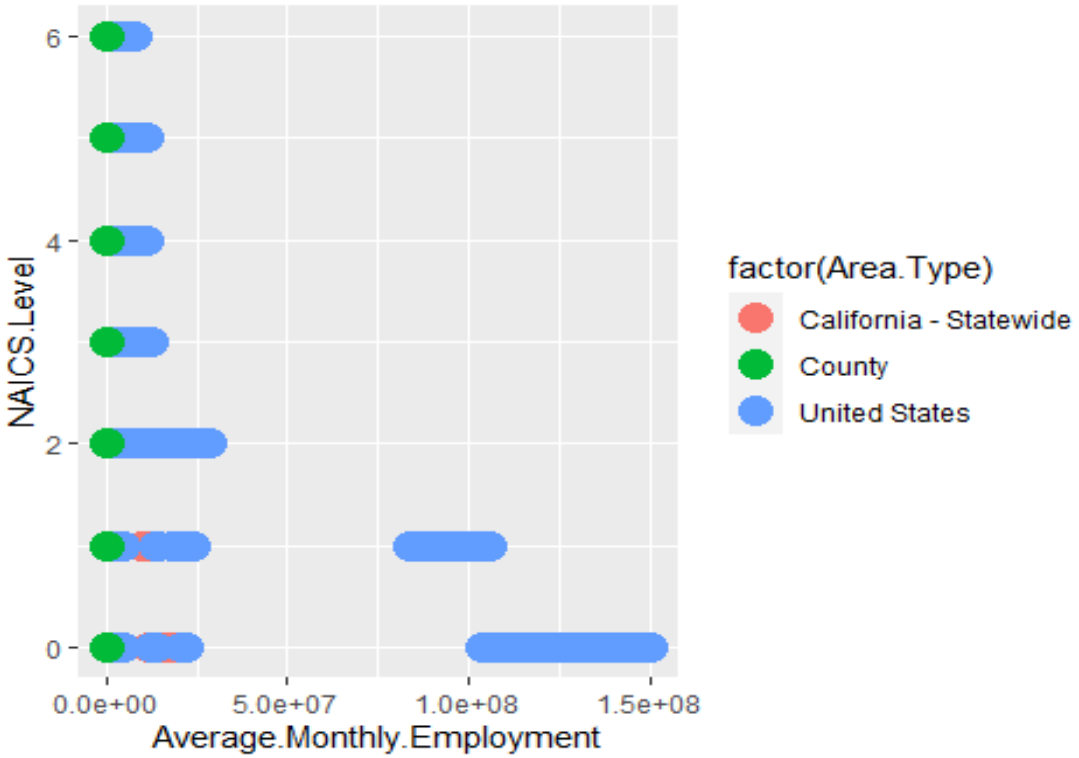
Some questions I would be asking on seeing this dataset is-

1. Which industries have employees with less wages and where are such industries located?
2. Which industries have employees with more wages so I could better redirect the public services and cater the needy?
3. Which industries are easier to manage and come out efficient?

3. Initial visualizations and if required, transform to get the data ready.

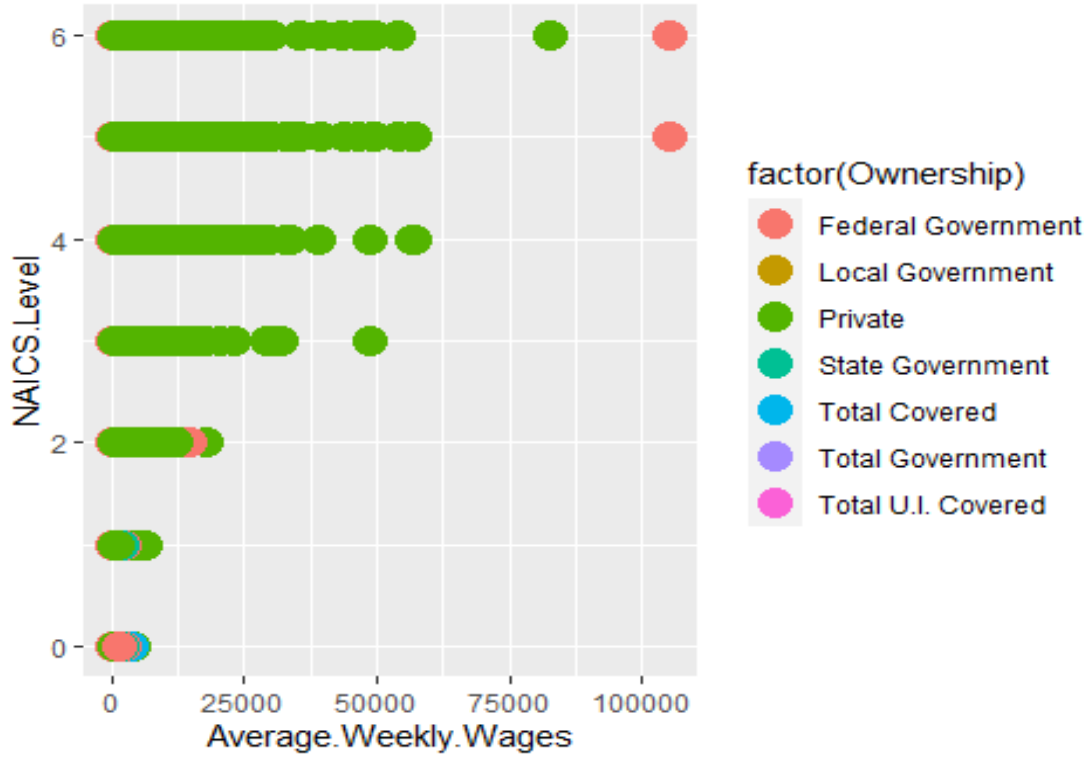
Done visualization and transformation in the later section in EDA. I have added initial visualization of my data sets, just to give an idea about the data.

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.2

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

ggplot(data = clean_data) +
  geom_point(mapping = aes(x = Average.Monthly.Employment, y = NAICS.Level,
color = factor(Area.Type)), size = 5)
```



We have factored the data on Area type. We can see that as the NAICS level of the industry goes up, the wages would go down- meaning, the lesser the employees, the more are the chances for better and stable wages. And also, we can observe that the employees working for federal industries earn much higher than those working for state or county funded industries.

```
ggplot(data = clean_data) + geom_point(mapping = aes(x =
Average.Weekly.Wages, y = NAICS.Level, color = factor(Ownership)), size=5)
```



We can see that the weekly wages of federal government is far greater than that of private employees for industries with NAICS levels 5 and 6. But for companies with NAICS levels 0,1 and 2, it is more or less the same. But, from the data above, we can see that private companies contribute a lot towards job creation than the government. Enhancing and improving private sector and empowering them is a must for improving living standard in California.

Data Information:

1. Background or the context of data selected -sources, description of how it was collected, time period it represents, context in it was collected if available, perhaps why you selected it.

The dataset that I have chosen is “Quarterly_Census_of_Employment_and_Wages_QCEW_” from the California State Government’s “Open Data Portal” website. The link to the dataset

is https://data.ca.gov/dataset/quarterly-census-of-employment-and-wages-qcew/resource/efdcc006-bcaf-4066-a763-aef58514a7dd

The Quarterly Census of Employment and Wages (QCEW) Program is a Federal-State cooperative program between the U.S. Department of Labor's Bureau of Labor Statistics (BLS) and the California EDD's Labor Market Information Division (LMID). The QCEW program produces a comprehensive tabulation of employment and wage information for workers covered by California Unemployment Insurance (UI) laws and Federal workers covered by the Unemployment Compensation for Federal Employees (UCFE) program.

2. Description of the data - how big is it(number of observations, variables), how many numeric variables, how many categorical variables, description of the variables.

This is a very big dataset, and has multiple variables, that in many ways influence each other. There are 15 variables in the dataset - some categorical and some continuous numeric variables.

-> Area Type - This variable defines what type of area the row sample is collected from and takes in only specific string values like mostly 'County' and sometimes 'District' etc. So, I find it to be a categorical variable.

-> Area Name - This variable simply contains the name of the particular area from where the sample is collected from.

-> Year - contains the year from when the sample was collected.

-> Quarter - contains the quarter of the year when the sample was collected. This too is a categorical variable and takes in only specific values like '1st Qtr', '2nd Qtr', 'Annual' etc.

-> Ownership - this is a categorical variable that specifies who is the owner of the firm from whom the data is collected. It takes in values like 'State Government', 'Private' etc.

-> NAICS level - contains the NAICS level of the industry from whom the data is collected.

-> NAICS code - contains the NAICS code of that particular industry.

-> Industry Name -contains the name of the firm/industry.

-> Establishments - contains the number of establishments of that particular industry.

-> Average Monthly Employment - contains the average monthly salary of the employees. This is a numeric variable.

-> 1st Month Emp, 2nd Month Emp, 3rd Month Emp - contains the average salary of the employees for the first 3 months. This is a numeric variable.

-> Total Wages(All workers) - contains the total salaries of all the workers in that industry. This is a numeric variable.

-> Average Weekly Wages - contains the average weekly wages of all the employees. This is a numeric variable.

3. Goal - What questions you plan to understand from the data.

I will be doing analysis on all the features but mainly trying to focus on Area Type, Average Weekly Wages, Average Monthly Employment, NAICS level(that may tell us the type and size of industry) and Ownership which tells us which firm is paying how much, so that we get an overall picture of the industries and the financial condition of their employees in California.

3. Analysis - Descriptive statistics and visualization of key variables.

```
print(summary(rawdata))
```

```
##   Area.Type          Area.Name              Year        Quarter
##  Length:4555904     Length:4555904     Min.   :2004   Length:4555904
##  Class :character   Class :character   1st Qu.:2008   Class :character
##  Mode  :character   Mode  :character   Median :2012   Mode  :character
##                                        Mean   :2012
##                                        3rd Qu.:2017
##                                        Max.   :2021
##
##   Ownership          NAICS.Level     NAICS.Code        Industry.Name
##  Length:4555904     Min.   :0.00    Length:4555904     Length:4555904
##  Class :character   1st Qu.:4.00    Class :character   Class :character
##  Mode  :character   Median :5.00    Mode  :character   Mode  :character
##                     Mean   :4.75
##                     3rd Qu.:6.00
##                     Max.   :6.00
##                     NA's   :50296
##  Establishments     Average.Monthly.Employment X1st.Month.Emp
##  Min.   :       0   Min.   :        0          Min.   :        0
##  1st Qu.:       7   1st Qu.:       69          1st Qu.:       12
##  Median :      20   Median :      288          Median :      143
##  Mean   :    2322   Mean   :    32731          Mean   :    26048
##  3rd Qu.:      87   3rd Qu.:     1525          3rd Qu.:      922
##  Max.   :11178274   Max.   :149931099          Max.   :149527674
##
##  X2nd.Month.Emp      X3rd.Month.Emp      Total.Wages..All.Workers.
##  Min.   :        0   Min.   :        0   Min.   :0.000e+00
##  1st Qu.:       12   1st Qu.:       12   1st Qu.:8.442e+05
##  Median :      143   Median :      144   Median :4.149e+06
##  Mean   :    26195   Mean   :    26311   Mean   :6.861e+08
##  3rd Qu.:      925   3rd Qu.:      926   3rd Qu.:2.577e+07
##  Max.   :150260321   Max.   :150005303   Max.   :9.721e+12
##
##  Average.Weekly.Wages
##  Min.   :     0.0
##  1st Qu.:   565.0
```

```
##  Median :   841.0
##  Mean   :   967.9
##  3rd Qu.:  1195.0
##  Max.   :105149.0
##

#print(rawdata %>% describe())
```

4. Summary of findings from the analysis and further questions for future analysis.

So from the analysis, I found that there are few outliers we shouldn't remove as it may add value to dataset as its real data, and there might be some exception where some industries may act untraditionally compared to their counterparts and we have to study its effects on the employees.

5. References - link to data or analysis sources you have referenced for the report.

The dataset that I have chosen is "Quarterly_Census_of_Employment_and_Wages_QCEW_" from the California State Government's "Open Data Portal" website. The link to the dataset is https://data.ca.gov/dataset/quarterly-census-of-employment-and-wages-qcew/resource/efdcc006-bcaf-4066-a763-aef58514a7dd

Information of the cleaned data

```
head(clean_data)

##   Area.Type      Area.Name Year Quarter        Ownership NAICS.Level
NAICS.Code
## 1    County Alameda County 2004 1st Qtr State Government           5
92212
## 2    County Alameda County 2004 1st Qtr          Private           6
325612
## 3    County Alameda County 2004 1st Qtr          Private           5
33341
## 4    County Alameda County 2004 1st Qtr          Private           4
4441
## 5    County Alameda County 2004 1st Qtr          Private           6
711130
## 6    County Alameda County 2004 1st Qtr          Private           2
1027
##                             Industry.Name Establishments
## 1                         Police Protection             10
## 2 Polish and Sanitation Good Manufacturing              7
## 3  HVAC and Commercial Refrigeration Equip             15
## 4   Building Material and Supplies Dealers            225
## 5               Musical Groups and Artists             35
## 6                           Other Services          15655
##   Average.Monthly.Employment X1st.Month.Emp X2nd.Month.Emp X3rd.Month.Emp
```

```
## 1                          354             356             353             353
## 2                          638             636             639             641
## 3                          354             356             354             352
## 4                         5172            5163            5132            5221
## 5                          306             252             261             406
## 6                        29553           29158           29533           29969
##   Total.Wages..All.Workers. Average.Weekly.Wages
## 1                   2532357                  550
## 2                  53376354                 6429
## 3                   7828534                 1701
## 4                  42069506                  626
## 5                   1053787                  265
## 6                 201730468                  525
```

```
print('Dimensions: ')
```

```
## [1] "Dimensions: "
```

```
dim(clean_data)
```

```
## [1] 4505608      15
```

```
print('Checking if the data has null values: ')
```

```
## [1] "Checking if the data has null values: "
```

```
sum(is.na(clean_data))
```

```
## [1] 0
```

Final Write-up

1. Introduction: What is your research question? Why do you care? Why should others care? If you know of any other related work done by others, please include a brief description.

If I were an investor and I am given this data, I would understand where and in which industries to put my investments in for more returns. If I am the Governor of California, this dataset would be of tremendous help to me. I could direct the social welfare schemes to cater the necessary people, invest better in the public transport system in the State, frame new and better social welfare schemes, etc.

Some questions I would be asking on seeing this dataset is-

1. Which industries have employees with less wages and where are such industries located?
2. Which industries have employees with more wages so I could better redirect the public services and cater the needy?
3. Which industries are easier to manage and come out efficient?

These are important questions to answer because there is a lot that this data set can offer and by answering these questions, I feel we can actually extract the most important details from this data- from the California Governor's point of view and from the investor's point of view. And also, we can also try to predict the future wages of the employees based on the past data of their salaries. This can help the Government manage inflation under normal conditions.

2. Data: Include context about the data covering:

a. Data source: Include the citation for your data, and provide link to the source.

The dataset that I have chosen is "Quarterly_Census_of_Employment_and_Wages_QCEW_" from the California State Government's "Open Data Portal" website. The link to the dataset is https://data.ca.gov/dataset/quarterly-census-of-employment-and-wages-qcew/resource/efdcc006-bcaf-4066-a763-aef58514a7dd

b. Data collection: Context on how the data was collected?

The data tells us about the condition of employment and wages in California state across different counties. This is a historical data that covers the information from years 2004 to 2021. It is an observational study based on the data collected by surveys from different counties and industries.

c. Cases: What are the cases (units of observation or experiment)? What do the rows represent in your dataset?

-> Area Type - This variable defines what type of area the row sample is collected from and takes in only specific string values like mostly 'County' and sometimes 'District' etc. So, this is a categorical variable.

-> Area Name - This variable simply contains the name of the particular area from where the sample is collected from. This is a string datatype.

-> Year - contains the year from when the sample was collected. This is an integer data type.

-> Quarter - contains the quarter of the year when the sample was collected. This too is a categorical variable and takes in only specific values like '1st Qtr', '2nd Qtr', 'Annual' etc. This is a string data type.

-> Ownership - this is a categorical string data type variable that specifies who is the owner of the firm from whom the data is collected. It takes in values like 'State Government', 'Private' etc.

-> NAICS level - contains the NAICS level of the industry from whom the data is collected. This is an integer data type.

-> NAICS code - contains the NAICS code of that particular industry. This is a double data type (continuous).

-> Industry Name -contains the name of the firm/industry. This is a string data type.

-> Establishments - contains the number of establishments of that particular industry. This is an integer data type.

-> Average Monthly Employment - contains the average monthly salary of the employees. This is a numeric double data type variable.

-> 1st Month Emp, 2nd Month Emp, 3rd Month Emp - contains the average salary of the employees for the first 3 months. This is a numeric double data type variable.

-> Total Wages(All workers) - contains the total salaries of all the workers in that industry. This is a numeric double data type variable.

-> Average Weekly Wages - contains the average weekly wages of all the employees. This is a numeric double data type variable.

d. Variables: What are the variables you will be studying?

I will be doing analysis on all the features but mainly trying to focus on Area Type, Average Weekly Wages, Average Monthly Employment, NAICS level(that may tell us the type and size of industry) and Ownership which tells us which firm is paying how much, so that we get an overall picture of the industries in California.

e. Type of study: was it an observational study or an experiment?

It is an observational study based on the data collected from various surveys conducted across different counties in California and across different industries.

f. Data clean-up: (Optional) If you had to do any data clean up (missing values, outliers, transformation), include a very brief description of your steps.

From the below code we conclude that we don't have any missing values but we do have some outliers in our data set.

```
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

head(clean_data)

##   Area.Type      Area.Name Year Quarter        Ownership NAICS.Level
NAICS.Code
```

```
## 1    County Alameda County 2004 1st Qtr State Government           5
92212
## 2    County Alameda County 2004 1st Qtr          Private           6
325612
## 3    County Alameda County 2004 1st Qtr          Private           5
33341
## 4    County Alameda County 2004 1st Qtr          Private           4
4441
## 5    County Alameda County 2004 1st Qtr          Private           6
711130
## 6    County Alameda County 2004 1st Qtr          Private           2
1027
##                             Industry.Name Establishments
## 1                         Police Protection             10
## 2 Polish and Sanitation Good Manufacturing              7
## 3  HVAC and Commercial Refrigeration Equip             15
## 4   Building Material and Supplies Dealers            225
## 5               Musical Groups and Artists             35
## 6                           Other Services          15655
##   Average.Monthly.Employment X1st.Month.Emp X2nd.Month.Emp X3rd.Month.Emp
## 1                        354            356            353            353
## 2                        638            636            639            641
## 3                        354            356            354            352
## 4                       5172           5163           5132           5221
## 5                        306            252            261            406
## 6                      29553          29158          29533          29969
##   Total.Wages..All.Workers. Average.Weekly.Wages
## 1                   2532357                  550
## 2                  53376354                 6429
## 3                   7828534                 1701
## 4                  42069506                  626
## 5                   1053787                  265
## 6                 201730468                  525
```

```
#clean_data %>% describe()
print('dimension: ')
```

```
## [1] "dimension: "
```

```
dim(clean_data)
```

```
## [1] 4505608      15
```

```
print('Checking if the data has Null values:')
```

```
## [1] "Checking if the data has Null values:"
```

```
sum(is.na(clean_data))
```

```
## [1] 0
```

Removing outliers

To get an initial look at how the data is distributed. We can use R's built in summary() on the data set, seen below:

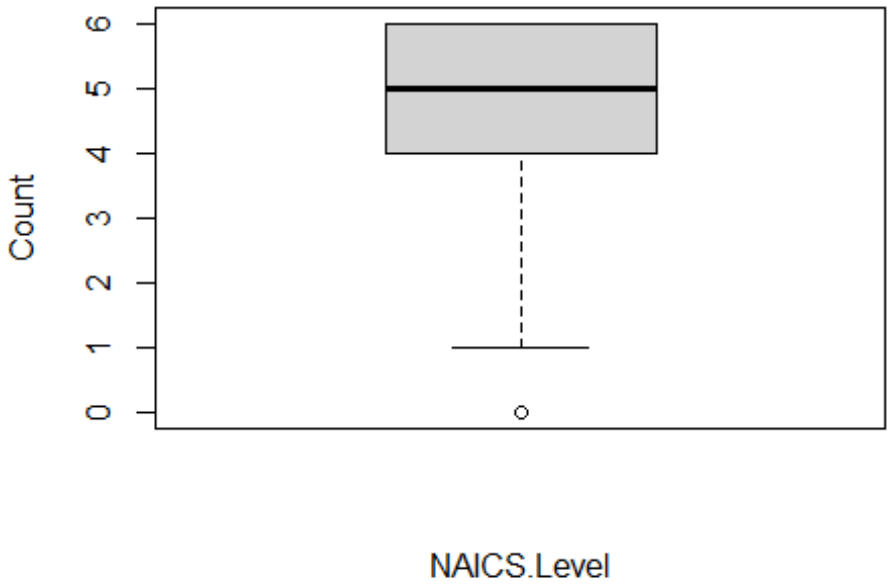
```
summary(clean_data)

##   Area.Type          Area.Name              Year        Quarter
##  Length:4505608     Length:4505608     Min.   :2004   Length:4505608
##  Class :character   Class :character   1st Qu.:2008   Class :character
##  Mode  :character   Mode  :character   Median :2013   Mode  :character
##                                        Mean   :2013
##                                        3rd Qu.:2017
##                                        Max.   :2021
##   Ownership          NAICS.Level     NAICS.Code        Industry.Name
##  Length:4505608     Min.   :0.000   Length:4505608     Length:4505608
##  Class :character   1st Qu.:4.000   Class :character   Class :character
##  Mode  :character   Median :5.000   Mode  :character   Mode  :character
##                     Mean   :4.747
##                     3rd Qu.:6.000
##                     Max.   :6.000
##  Establishments     Average.Monthly.Employment X1st.Month.Emp
##  Min.   :       0   Min.   :        0          Min.   :        0
##  1st Qu.:       7   1st Qu.:       69          1st Qu.:       11
##  Median :      20   Median :      288          Median :      141
##  Mean   :    2323   Mean   :    32745          Mean   :    25989
##  3rd Qu.:      87   3rd Qu.:     1525          3rd Qu.:      916
##  Max.   :11178274   Max.   :149931099          Max.   :149527674
##  X2nd.Month.Emp      X3rd.Month.Emp      Total.Wages..All.Workers.
##  Min.   :        0   Min.   :        0   Min.   :0.000e+00
##  1st Qu.:       12   1st Qu.:       12   1st Qu.:8.474e+05
##  Median :      142   Median :      142   Median :4.168e+06
##  Mean   :    26135   Mean   :    26251   Mean   :6.896e+08
##  3rd Qu.:      919   3rd Qu.:      920   3rd Qu.:2.589e+07
##  Max.   :150260321   Max.   :150005303   Max.   :9.721e+12
##  Average.Weekly.Wages
##  Min.   :     0.0
##  1st Qu.:   566.0
##  Median :   842.0
##  Mean   :   969.2
##  3rd Qu.:  1196.0
##  Max.   :105149.0
```

As we can see, R outputs statistics for each column. These statistics being: Min,Median,Mean 1st & 3rd Quantile, and Maximum. We know to identify outliers data below this is Q1 - 1.5 * IQR is the outlier at lower end. While an observation that is above Quantile Q3 + 1.5 *IQR is one that is high. In this dataset we cannot conclude an observation is an outlier just because it is higher than Quantile 3 + 1.5* IQR for a reason being they are data of true outcome. And also, given the size of the data, it is difficult for us to show mean and quartile values clearly at scale on the graphs. So, outliers play a key role in the analysis of

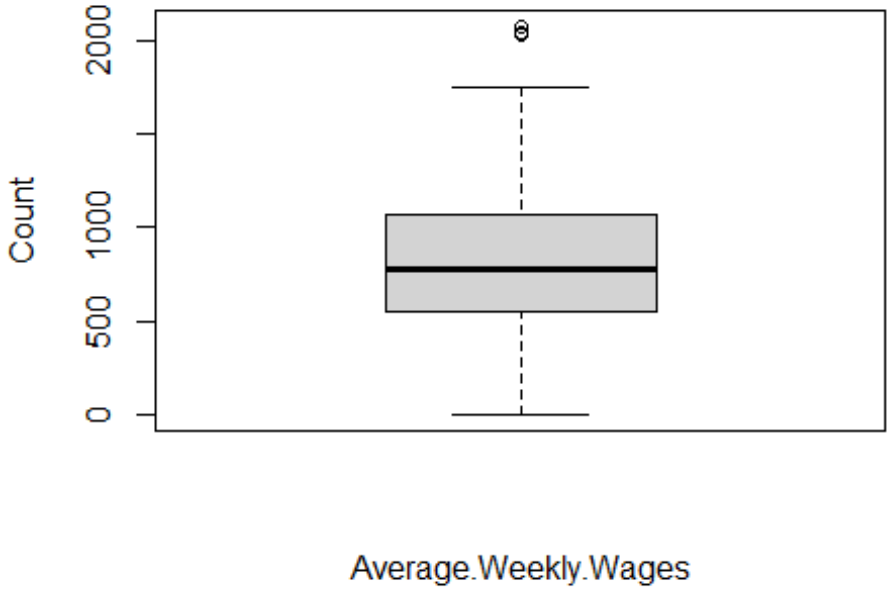
the data. We should know which industry is paying too high and too low, so that the Government would know about it.

Below boxplot shows the outliers in the dataset.

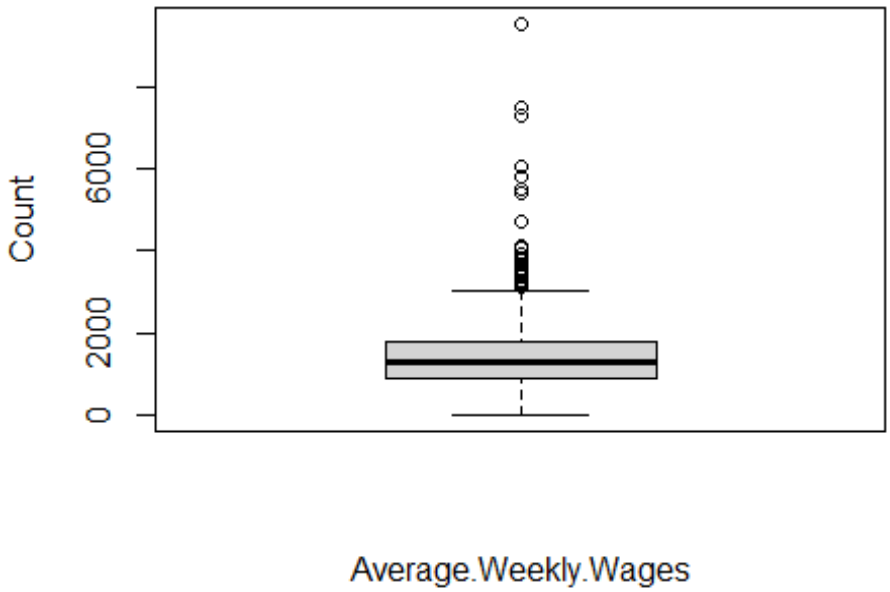
```
clean_data$Area.Name<- gsub(" ", "", tolower(clean_data$Area.Name))
clean_data$Quarter <- gsub(" ","", tolower(clean_data$Quarter))
clean_data$Ownership <- gsub(" ","", tolower(clean_data$Ownership))
clean_data1 <- clean_data[clean_data$Year == 2020 & clean_data$Area.Name ==
'lakecounty' &clean_data$Quarter == '4thqtr',]
boxplot(clean_data$NAICS.Level, xlab = "NAICS.Level",
 ylab = "Count")
```




```
clean_data2 <- clean_data[clean_data$Year == 2019 & clean_data$Area.Name ==
'kingscounty' &clean_data$Quarter == '3rdqtr',]
boxplot(clean_data2$Average.Weekly.Wages, xlab = "Average.Weekly.Wages",
 ylab = "Count")
```




```
#unique(clean_data$Quarter)

clean_data3 <- clean_data[clean_data$Year == 2019 & clean_data$Area.Name ==
'alamedacounty' &clean_data$Quarter == '3rdqtr',]
boxplot(clean_data3$Average.Weekly.Wages, xlab = "Average.Weekly.Wages",
 ylab = "Count")
```




```
#unique(clean_data$Industry.Name)
```

3. Exploratory Data Analysis: summarize your data using descriptive statistics / summary statistics and visualizations relevant to your questions or ones that highlight some interesting insight.

1. Which areas have less average wages?

```
library(dplyr)
library(ggplot2)
clean_data4 <- clean_data[clean_data$Year == 2021,]
clean_data4 %>%
  group_by(Area.Name) %>%
  summarise(average = mean(Average.Weekly.Wages)) #%>%
```

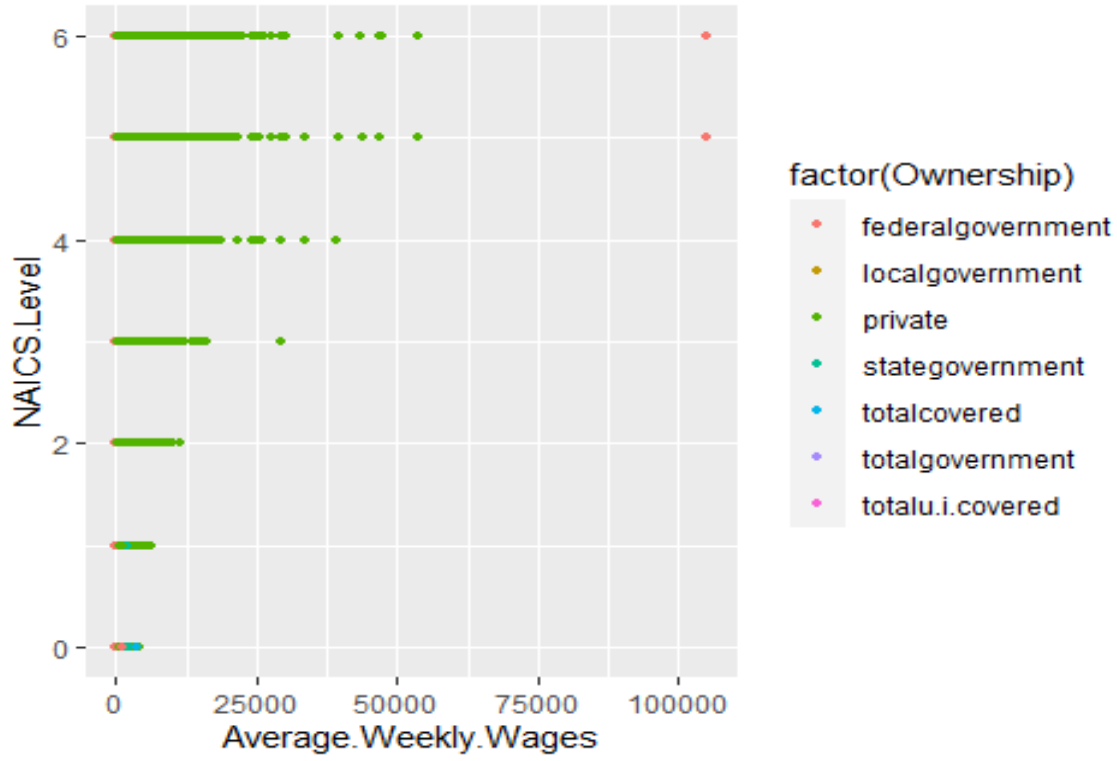
```
## # A tibble: 60 × 2
##    Area.Name         average
##    <chr>               <dbl>
##  1 alamedacounty       1603.
##  2 alpinecounty         573.
##  3 amadorcounty         946.
##  4 buttecounty         1004.
##  5 calaverascounty      856.
##  6 california          1592.
##  7 colusacounty         913.
##  8 contracostacounty   1539.
##  9 delnortecounty       840.
```

```
## 10 eldoradocounty      1111.
## # … with 50 more rows

  #sort(????)

  #geom_line(mapping = aes(x = Average.Weekly.Wages, y = Area.Name))
```

From the above summary, we can see that for the year 2021, Alpine county has the lowest average weekly wages followed by Sierra county, Delnorte County and Mariposa county. San Francisco county has the highest average weekly wages at around $2325.76, followed by SanMateo County and Santa Clara county. So, based on the above data, we can assume that the influx of people into San Francisco, Santa Clara and SanMateo counties would grow in future and the Government should develop the infrastructure in those areas to accommodate more people. And, in counties with low average wages, the Government should improve public transport infrastructure and initiate more social welfare schemes to support the needy. And also, the Government should encourage investors to invest more into these areas to increase the financial prospects of those places.

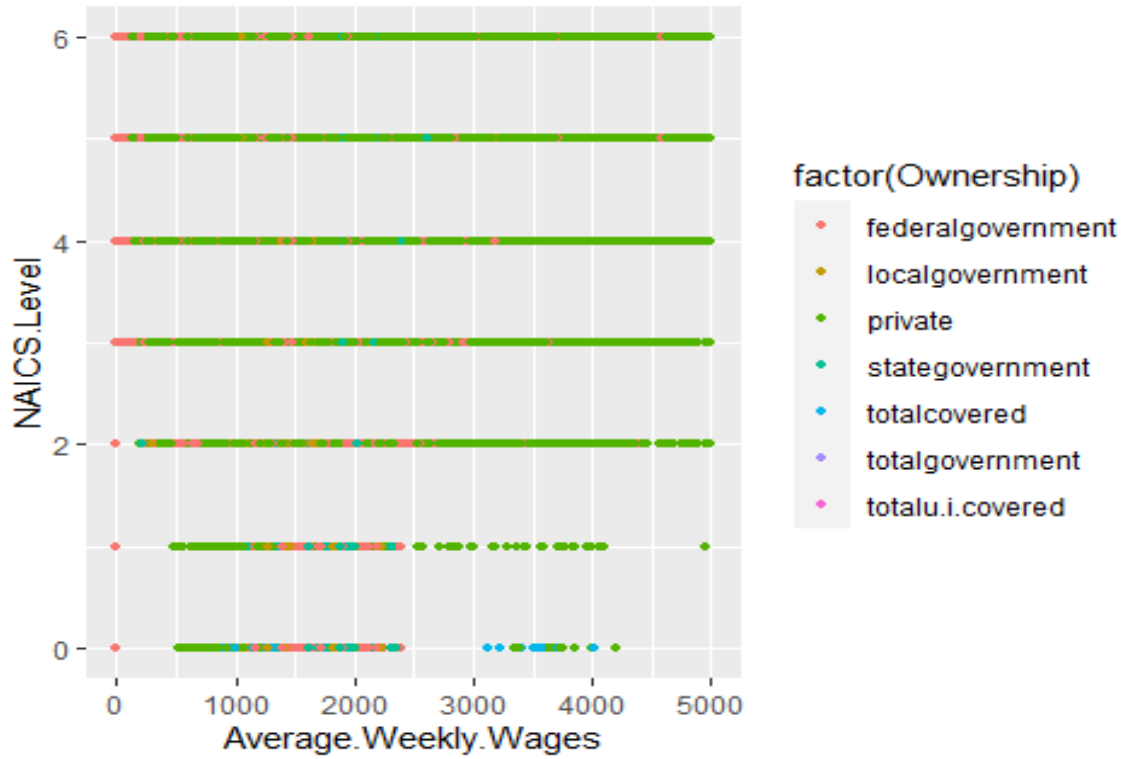
2. Which industries have employees with less annual wages?

```
clean_data5 <- clean_data[clean_data$Year == 2021 ,]
ggplot(data = clean_data4) +
  geom_point(mapping = aes(x = Average.Weekly.Wages, y = NAICS.Level, color =
factor(Ownership)), size = 1)
```




```
clean_data5 <- clean_data4[clean_data4$Average.Weekly.Wages < 5000,]
ggplot(data = clean_data5) +
```

```
  geom_point(mapping = aes(x = Average.Weekly.Wages, y = NAICS.Level, color =
factor(Ownership)), size = 1)
```



These graphs are for the year 2021. I have elaborated the 0-5000$ Average weekly wages segment from the first graph to have an even clear picture of that small segment. From the above two graphs, we can see that all the major players in the industry employ people across all NAICS levels. Private players seem to employ a large number of people and their average weely wages are pretty evenly spread out. High pays are offered for high skilled jobs and low pays are offered for low skilled jobs. Federal Government offers both low paying jobs and pretty high paying jobs. But the average pay of the feredal government is the highest, followed by State Government and other government agencies. We can confirm these claims from the below summary also.

```
library(dplyr)
clean_data4 %>%
  group_by(Ownership) %>%
  summarise(average = mean(Average.Weekly.Wages))
```

```
## # A tibble: 7 × 2
##   Ownership         average
##   <chr>               <dbl>
## 1 federalgovernment   1473.
## 2 localgovernment     1303.
## 3 private             1278.
## 4 stategovernment     1459.
## 5 totalcovered        1230.
```

```
## 6 totalgovernment     1403.
## 7 totalu.i.covered    1291
```

4. Data Analysis: Pick and perform two of the following techniques we have learned in class and that helps answer your question about the dataset: PCA, hypothesis testing / confidence interval, regression analysis (linear /logistic).

PCA

To perform PCA, I am eliminating some columns that have categorical values like names of places, ownership, area name etc. I will be analysing the remaining variables.

```
summary(clean_data)
```

```
##   Area.Type          Area.Name              Year        Quarter
##  Length:4505608     Length:4505608     Min.   :2004   Length:4505608
##  Class :character   Class :character   1st Qu.:2008   Class :character
##  Mode  :character   Mode  :character   Median :2013   Mode  :character
##                                        Mean   :2013
##                                        3rd Qu.:2017
##                                        Max.   :2021
##   Ownership          NAICS.Level     NAICS.Code         Industry.Name
##  Length:4505608     Min.   :0.000   Length:4505608     Length:4505608
##  Class :character   1st Qu.:4.000   Class :character   Class :character
##  Mode  :character   Median :5.000   Mode  :character   Mode  :character
##                     Mean   :4.747
##                     3rd Qu.:6.000
##                     Max.   :6.000
##  Establishments     Average.Monthly.Employment X1st.Month.Emp
##  Min.   :       0   Min.   :         0         Min.   :        0
##  1st Qu.:       7   1st Qu.:        69         1st Qu.:       11
##  Median :      20   Median :       288         Median :      141
##  Mean   :    2323   Mean   :     32745         Mean   :    25989
##  3rd Qu.:      87   3rd Qu.:      1525         3rd Qu.:      916
##  Max.   :11178274   Max.   :149931099         Max.   :149527674
##  X2nd.Month.Emp      X3rd.Month.Emp      Total.Wages..All.Workers.
##  Min.   :        0   Min.   :        0   Min.   :0.000e+00
##  1st Qu.:       12   1st Qu.:       12   1st Qu.:8.474e+05
##  Median :      142   Median :      142   Median :4.168e+06
##  Mean   :    26135   Mean   :    26251   Mean   :6.896e+08
##  3rd Qu.:      919   3rd Qu.:      920   3rd Qu.:2.589e+07
##  Max.   :150260321   Max.   :150005303   Max.   :9.721e+12
##  Average.Weekly.Wages
##  Min.   :     0.0
##  1st Qu.:   566.0
##  Median :   842.0
##  Mean   :   969.2
##  3rd Qu.:  1196.0
##  Max.   :105149.0
```

```r
clean_datae <- clean_data[,10:15]
head(clean_datae)
```

```
##   Average.Monthly.Employment X1st.Month.Emp X2nd.Month.Emp X3rd.Month.Emp
## 1                        354            356            353            353
## 2                        638            636            639            641
## 3                        354            356            354            352
## 4                       5172           5163           5132           5221
## 5                        306            252            261            406
## 6                      29553          29158          29533          29969
##   Total.Wages..All.Workers. Average.Weekly.Wages
## 1                   2532357                  550
## 2                  53376354                 6429
## 3                   7828534                 1701
## 4                  42069506                  626
## 5                   1053787                  265
## 6                 201730468                  525
```

```r
cov_data <- cov(clean_datae[,c(1:6)])
dim(cov_data)
```

```
## [1] 6 6
```

```r
cov_data
```

```
##                            Average.Monthly.Employment X1st.Month.Emp
## Average.Monthly.Employment               1.287419e+12   1.022122e+12
## X1st.Month.Emp                           1.022122e+12   1.017216e+12
## X2nd.Month.Emp                           1.027760e+12   1.022786e+12
## X3rd.Month.Emp                           1.031952e+12   1.026881e+12
## Total.Wages..All.Workers.                2.625137e+16   1.297730e+16
## Average.Weekly.Wages                     1.292178e+06   1.046640e+06
##                            X2nd.Month.Emp X3rd.Month.Emp
## Average.Monthly.Employment   1.027760e+12   1.031952e+12
## X1st.Month.Emp               1.022786e+12   1.026881e+12
## X2nd.Month.Emp               1.028424e+12   1.032590e+12
## X3rd.Month.Emp               1.032590e+12   1.036905e+12
## Total.Wages..All.Workers.    1.304931e+16   1.310068e+16
## Average.Weekly.Wages         1.038762e+06   1.025269e+06
##                            Total.Wages..All.Workers. Average.Weekly.Wages
## Average.Monthly.Employment              2.625137e+16         1.292178e+06
## X1st.Month.Emp                          1.297730e+16         1.046640e+06
## X2nd.Month.Emp                          1.304931e+16         1.038762e+06
## X3rd.Month.Emp                          1.310068e+16         1.025269e+06
## Total.Wages..All.Workers.               8.652706e+20         1.882541e+11
## Average.Weekly.Wages                    1.882541e+11         4.744888e+05
```

```r
cor_data <- cor(clean_datae[,c(1:6)])
dim(cor_data)
```

```
## [1] 6 6
```

```
cor_data
```

```
##                             Average.Monthly.Employment X1st.Month.Emp
## Average.Monthly.Employment                1.000000000     0.89317461
## X1st.Month.Emp                            0.893174606     1.00000000
## X2nd.Month.Emp                            0.893194147     0.99998181
## X3rd.Month.Emp                            0.893161508     0.99987156
## Total.Wages..All.Workers.                 0.786531363     0.43742328
## Average.Weekly.Wages                      0.001653291     0.00150653
##                             X2nd.Month.Emp X3rd.Month.Emp
## Average.Monthly.Employment     0.89319415     0.89316151
## X1st.Month.Emp                 0.99998181     0.99987156
## X2nd.Month.Emp                 1.00000000     0.99993686
## X3rd.Month.Emp                 0.99993686     1.00000000
## Total.Wages..All.Workers.      0.43744714     0.43736941
## Average.Weekly.Wages           0.00148702     0.00146169
##                             Total.Wages..All.Workers. Average.Weekly.Wages
## Average.Monthly.Employment                0.786531363          0.001653291
## X1st.Month.Emp                            0.437423279          0.001506530
## X2nd.Month.Emp                            0.437447145          0.001487020
## X3rd.Month.Emp                            0.437369405          0.001461690
## Total.Wages..All.Workers.                 1.000000000          0.009290849
## Average.Weekly.Wages                      0.009290849          1.000000000
```

Computing eigen values and eigen vectors for covariance and correlation matrices.

```
cov_eigen <- eigen(cov_data)
cor_eigen <- eigen(cor_data)
print(cov_eigen)
```

```
## eigen() decomposition
## $values
## [1] 8.652706e+20 2.974483e+12 9.114204e+09 1.378548e+08 6.068793e+06
## [6] 4.735783e+05
##
## $vectors
##               [,1]          [,2]          [,3]          [,4]          [,5]
## [1,] -3.033891e-05 -4.031105e-01  0.9151481874  2.047423e-03 -1.210896e-03
## [2,] -1.499797e-05 -5.257028e-01 -0.2321983201  6.074853e-01  5.483443e-01
## [3,] -1.508119e-05 -5.285975e-01 -0.2342820327  1.727837e-01 -7.973968e-01
## [4,] -1.514056e-05 -5.307760e-01 -0.2317316780 -7.753084e-01  2.519405e-01
## [5,] -1.000000e+00  3.612253e-05 -0.0000172403 -4.034179e-08  2.386012e-08
## [6,] -2.175667e-10  1.558532e-06 -0.0003058024  1.124692e-04  1.224036e-03
##               [,6]
## [1,]  2.817349e-04
## [2,] -8.097047e-04
## [3,]  8.857901e-04
## [4,] -2.912230e-04
## [5,] -5.570662e-09
## [6,]  9.999992e-01
```

```
print(cor_eigen)
```

```
## eigen() decomposition
## $values
## [1] 4.185391e+00 1.000266e+00 8.081576e-01 6.045830e-03 1.338897e-04
## [6] 5.890627e-06
##
## $vectors
##              [,1]        [,2]        [,3]         [,4]          [,5]
## [1,] -0.476340017  0.00516294  0.23852382  0.846269821 -2.335062e-03
## [2,] -0.474376516 -0.01014138 -0.26736608 -0.192387321 -6.067501e-01
## [3,] -0.474387917 -0.01016066 -0.26735584 -0.193352464 -1.684949e-01
## [4,] -0.474365880 -0.01018754 -0.26740987 -0.189022529  7.768249e-01
## [5,] -0.313044565  0.03252485  0.85279230 -0.416759662  1.191322e-03
## [6,] -0.001823776  0.99930256 -0.03714655  0.003346803  2.196056e-05
##                [,6]
## [1,]  1.362754e-03
## [2,] -5.460961e-01
## [3,]  7.984973e-01
## [4,] -2.533352e-01
## [5,] -6.961297e-04
## [6,]  9.840588e-06
```

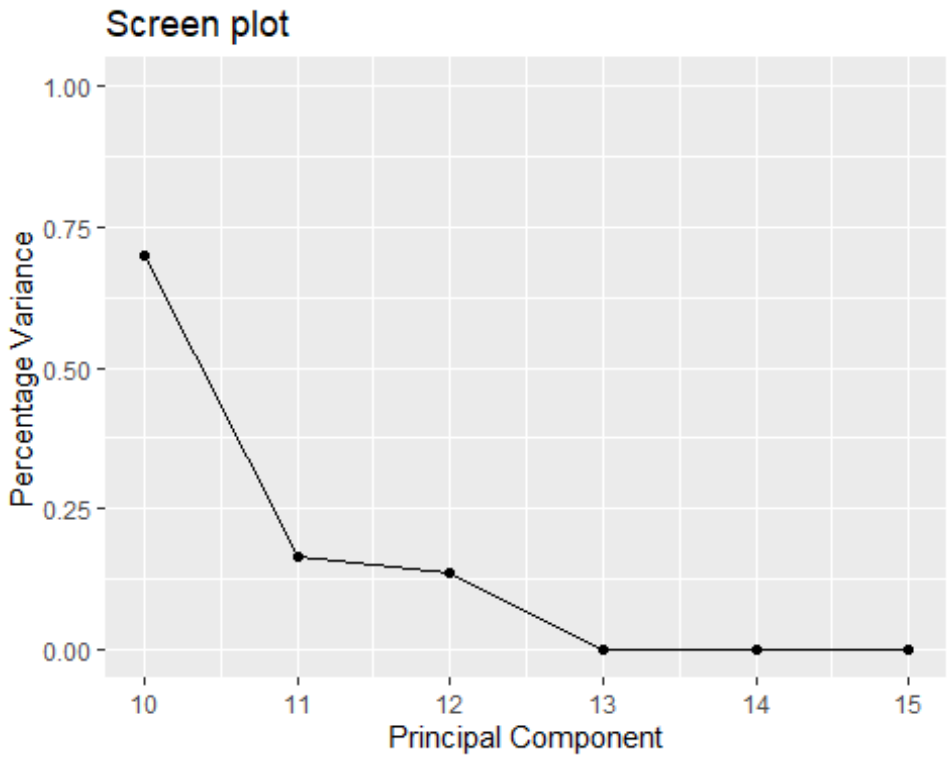
Determining the number of principal components we would require to reduce feature dimensions yet capture atleast 85% of the variability in the data.

```
cor_eigen$values
```

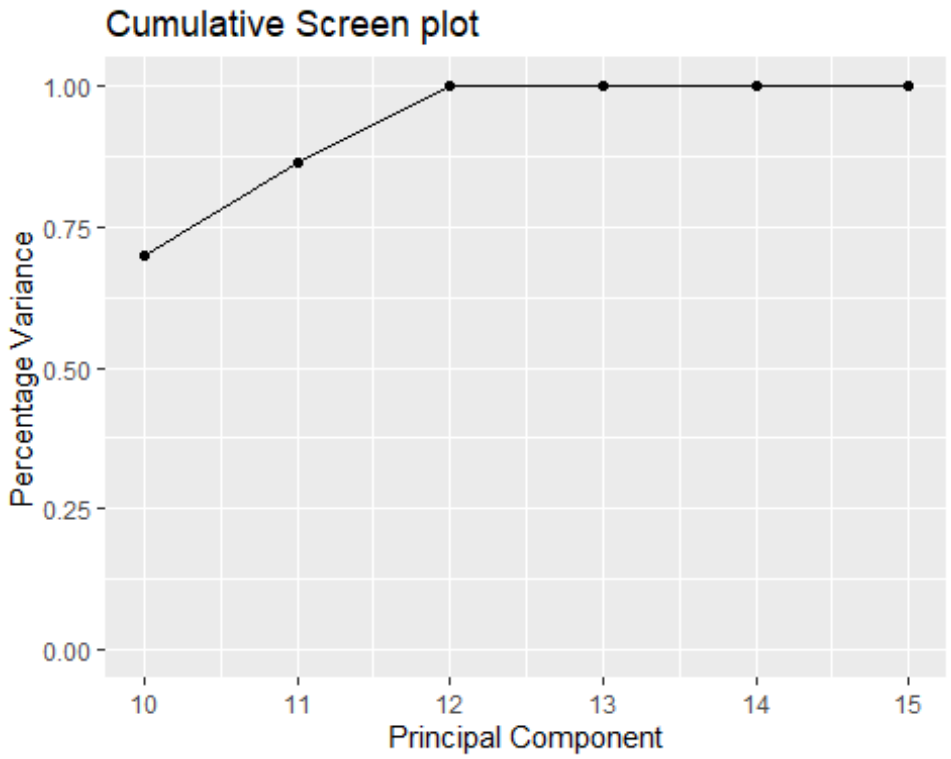
```
## [1] 4.185391e+00 1.000266e+00 8.081576e-01 6.045830e-03 1.338897e-04
## [6] 5.890627e-06
```

```
var_analyze <- cor_eigen$values / sum(cor_eigen$values)
cumulative_var <- cumsum(var_analyze)

#create a screen plot
library(ggplot2)
qplot(c(10:15), var_analyze) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Percentage Variance") +
  ggtitle("Screen plot") +
  ylim(0, 1)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```




```
#creating a cumulative screen plot
qplot(c(10:15), cumulative_var) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Percentage Variance") +
  ggtitle("Cumulative Screen plot") +
  ylim(0, 1)
```



Upon observation from both the screen plot and cumuative screen plot, we can observe that the first 3 principal components capture almost 98% of the variability in the data. So, as per the requirement, we can take in the first 3 principal components and omit the others, and still have more than 90% of the variance in the data.

```
#plot the number of eigen values to be selected
evecs <- cor_eigen$vectors[, 1:3]
colnames(evecs) <- c("e1", "e2", "e3")
row.names(evecs) <- colnames(clean_datae)
evecs
```

```
##                                     e1          e2          e3
## Average.Monthly.Employment -0.476340017  0.00516294  0.23852382
## X1st.Month.Emp             -0.474376516 -0.01014138 -0.26736608
## X2nd.Month.Emp             -0.474387917 -0.01016066 -0.26735584
## X3rd.Month.Emp             -0.474365880 -0.01018754 -0.26740987
## Total.Wages..All.Workers.  -0.313044565  0.03252485  0.85279230
## Average.Weekly.Wages       -0.001823776  0.99930256 -0.03714655
```

```
pc1 <- as.matrix(clean_datae) %*% evecs[,1]
pc2 <- as.matrix(clean_datae) %*% evecs[,2]
pc3 <- as.matrix(clean_datae) %*% evecs[,3]
pc <- data.frame(pc1, pc2, pc3)
head(pc)
```

```
##           pc1        pc2         pc3
## 1   -793414.0   82905.18   2159354.6
```

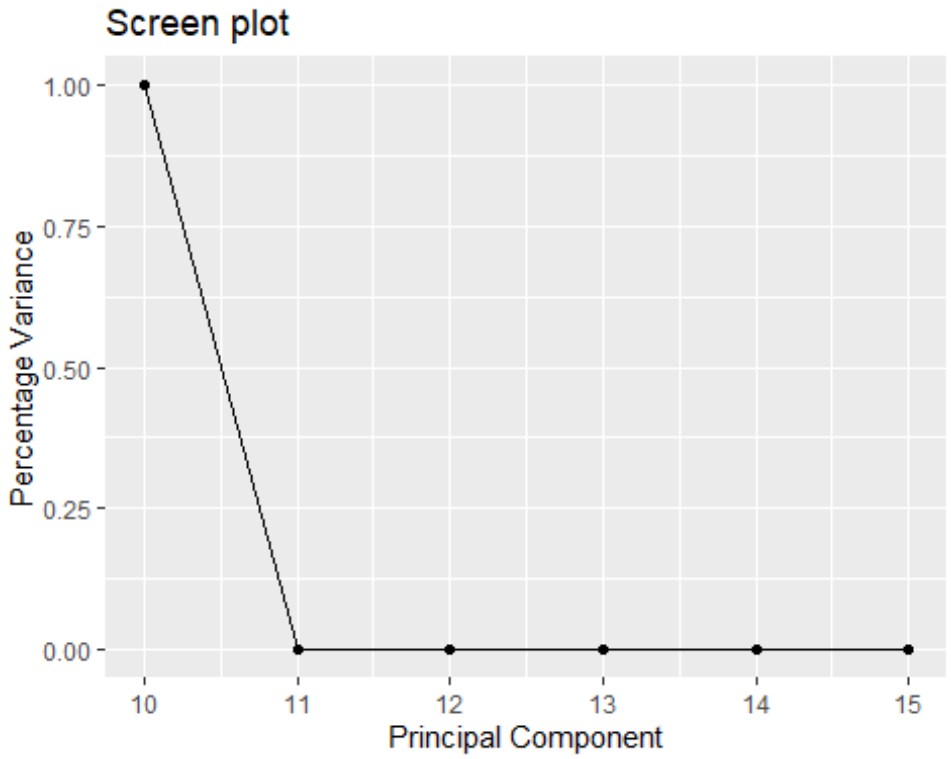
```
## 2 -16710402.0 1742466.16  45518345.0
## 3  -2451355.5  256312.73   6675850.8
## 4 -13179455.4 1368798.88  35873612.7
## 5   -330464.5   34531.31    898478.9
## 6 -63206763.0 6561029.02 172017514.6
```

Now, we do the same for covariance matrix.

```
cov_eigen$values
```

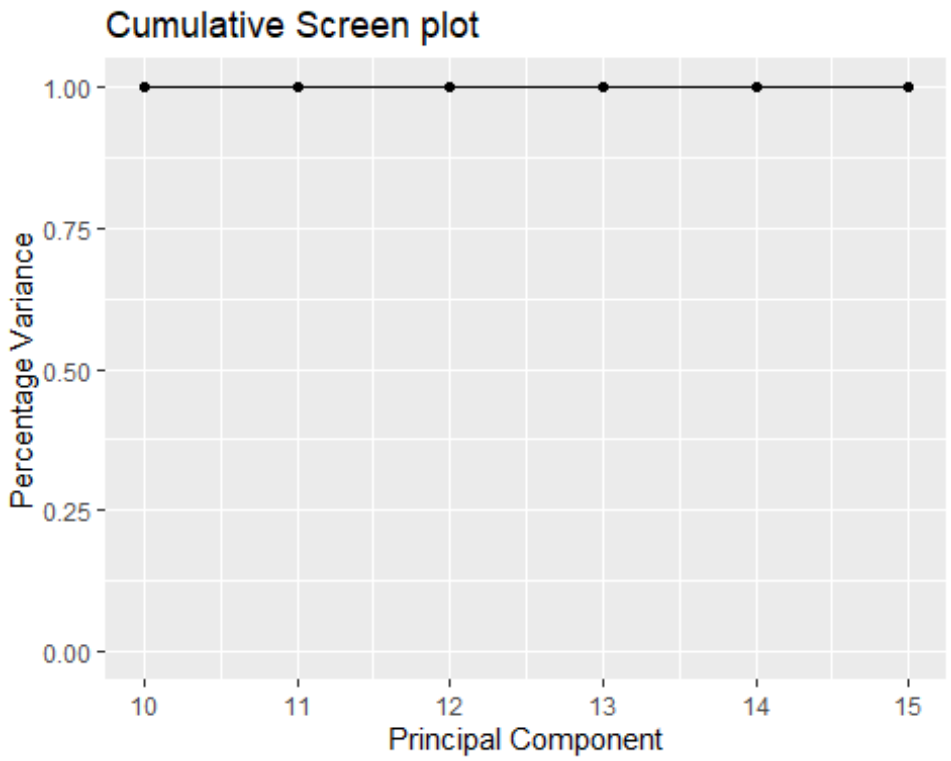
```
## [1] 8.652706e+20 2.974483e+12 9.114204e+09 1.378548e+08 6.068793e+06
## [6] 4.735783e+05
```

```
variance_analyze <- cov_eigen$values / sum(cov_eigen$values)
cumulative_variance <- cumsum(variance_analyze)

#create a screen plot
library(ggplot2)
qplot(c(10:15), variance_analyze) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Percentage Variance") +
  ggtitle("Screen plot") +
  ylim(0, 1)
```




```
#creating a cumulative screen plot
qplot(c(10:15), cumulative_variance) +
  geom_line() +
```

```
  xlab("Principal Component") +
  ylab("Percentage Variance") +
  ggtitle("Cumulative Screen plot") +
  ylim(0, 1)
```



Upon observation, we find that the first principal component alone is capturing almost 98% of the variability in the data. Since the assumed requirement is 90%, we can proceed with considering just the first principal component - in case of the covariance matrix. Yes, the interpretation of the components differ by a very big margin when comparing with correlation matrix. Correlation matrix is allowing us to take multiple principal components into consideration, thereby giving us more scope to see how different attributes in the data are interacting and influencing each other.

For the below process, we are considering the 2nd principal component also, in addition to the first principal component.

```
#computing the principal component vectors based on our selection above
evecs1 <- cov_eigen$vectors[, 1:2]
colnames(evecs1) <- c("e1", "e2")
row.names(evecs1) <- colnames(clean_datae)
evecs1
```

```
##                                      e1            e2
## Average.Monthly.Employment -3.033891e-05 -4.031105e-01
## X1st.Month.Emp             -1.499797e-05 -5.257028e-01
## X2nd.Month.Emp             -1.508119e-05 -5.285975e-01
## X3rd.Month.Emp             -1.514056e-05 -5.307760e-01
```

```
## Total.Wages..All.Workers.  -1.000000e+00  3.612253e-05
## Average.Weekly.Wages       -2.175667e-10  1.558532e-06
```

```
pc1 <- as.matrix(clean_datae) %*% evecs1[,1]
pc2 <- as.matrix(clean_datae) %*% evecs1[,2]
pc <- data.frame(pc1, pc2)
head(pc)
```

```
##          pc1         pc2
## 1   -2532357   -612.3342
## 2  -53376354    658.5664
## 3   -7828534   -421.0189
## 4  -42069506  -8763.3769
## 5   -1053787   -571.2221
## 6 -201730470 -51472.4467
```

The 'colnames' builtin method is not accepting inputs with less than 2 dimensions. So, for the sake of convenience, I am considering the 2nd principal component also, which will take the variability coverage to almost 98%. This is only for the sake of convenience and keeping in mind the limitations that we have regarding some inbuilt methods in R.

Hypothesis Testing

```
clean_data$Area.Name<- gsub(" ", "", tolower(clean_data$Area.Name))
clean_data$Quarter <- gsub(" ","", tolower(clean_data$Quarter))
#clean_data$Ownership <- gsub(" ","", tolower(clean_data$Ownership))
dim(clean_data[clean_data$Area.Name == 'lakecounty',])
```

```
## [1] 28859    15
```

HYPOTHESIS 1:

Null hypothesis H0: Average weekly wages in Lake County is $912 for 4th quarter of 2020 This statement has been picked up from the below source.
https://www.bls.gov/regions/west/news-release/2021/countyemploymentandwages_california_20210806.htm#:~:text=Lake%20County%20%28%24912%29%20had%20the%20lowest%20weekly%20wage,wages%20of%20%241%2C400%20or%20higher.%20%28See%20chart%203.%29

Alternate hypothesis H1: Average weekly wages in Lake county is not $912 for 4th quarter of 2020.

Test statistic has been calculated below and it is 0.2894076 Reference distribution is normal distribution of the chosen data for Lake county. Rejection criteria is if the data is above 95% of the data on the normal distribution.

```
print('Hypothesis 1 : Average weekly wages in Lake county is $912 for 4th quarter of 2020')
```

```
## [1] "Hypothesis 1 : Average weekly wages in Lake county is $912 for 4th quarter of 2020"
```

```
library(dplyr)
clean_data1 <- clean_data[clean_data$Year == 2020 & clean_data$Area.Name ==
'lakecounty' &clean_data$Quarter == '4thqtr',]
dim(clean_data1)
```

```
## [1] 318  15
```

```
if(mean(clean_data1$Average.Weekly.Wages) != 912)
{
  print('Hypothesis is wrong')
}
```

```
## [1] "Hypothesis is wrong"
```

```
std_dev <- sd(clean_data1$Average.Weekly.Wages)
mean1 <- mean(clean_data1$Average.Weekly.Wages)
alpha <- 0.05
test_statistic <- (912 -
mean(clean_data1$Average.Weekly.Wages))/(std_dev/sqrt(318))
test_statistic
```

```
## [1] 0.2894076
```

```
z_0 <- (912 - mean1)/(std_dev/sqrt(318))
z_c <- -qnorm(alpha)
print('z_0 is:')
```

```
## [1] "z_0 is:"
```

```
z_0
```

```
## [1] 0.2894076
```

```
print('z_c is:')
```

```
## [1] "z_c is:"
```

```
z_c
```

```
## [1] 1.644854
```

```
#z_0 < z_c, fail to reject null hypothesis, type II error
```

From the above test, we can see that z_0 is less than z_c. So, we fail to reject the null hypothesis and this is a type II error.

```
clean_data6 <- clean_data[clean_data$Year == 2020,]
clean_data6 %>%
  group_by(Area.Name) %>%
  summarise(average = mean(Average.Weekly.Wages))
```

```
## # A tibble: 60 × 2
##    Area.Name         average
##    <chr>               <dbl>
```

```
##  1 alamedacounty       1523.
##  2 alpinecounty         764.
##  3 amadorcounty         887.
##  4 buttecounty          950.
##  5 calaverascounty      828.
##  6 california          1484.
##  7 colusacounty         910.
##  8 contracostacounty   1480.
##  9 delnortecounty       773.
## 10 eldoradocounty      1067.
## # … with 50 more rows
```

From the above data, we can see that Sierra County has the least average weekly wages in all of California and the average weekly wages in lake county is only 840.23, not $912 as said in the above article.

HYPOTHESIS 2:

NUll hypothesis H0: Private sector average weekly salary is $1627 This statement has been picked up from the below source. https://www.bls.gov/charts/county-employment-and-wages/percent-change-aww-by-state.htm

Alternative hypothesis H1: Private sector average weekly salary is not $1627.

Test statistic has been calculated below and it is 71.74951. Reference distribution is normal distribution of the chosen data for Lake county. Rejection criteria is if the data is above 95% of the data on the normal distribution.

```
print('Hypothesis: Private sector avg week salary is $1627 ')
```

```
## [1] "Hypothesis: Private sector avg week salary is $1627 "
```

```
clean_data2 <- clean_data[clean_data$Year == 2021 & clean_data$Quarter ==
'annual' & clean_data$Ownership == 'private',]
dim(clean_data2)
```

```
## [1] 42529    15
```

```
std_dev2 <- sd(clean_data2$Average.Weekly.Wages)
mean2 <- mean(clean_data2$Average.Weekly.Wages)
alpha2 <- 0.05
test_statistic2 <- (1627 - mean2)/(std_dev2/sqrt(42529))
z_0 <- test_statistic2
z_c <- -qnorm(alpha2)
print('z_0 is:')
```

```
## [1] "z_0 is:"
```

```
z_0
```

```
## [1] 71.74951
```

```
print('z_c is:')
```

```
## [1] "z_c is:"
```

```
z_c
```

```
## [1] 1.644854
```

```
#z_0 > z_c, reject H0, type I error
```

From the above test, we can see that z_0 is greater than z_c. So, we reject the null hypothesis when it is false and this is a type I error.

```
clean_data7 <- clean_data[clean_data$Year == 2021 & clean_data$Ownership ==
'private',]
print(mean(clean_data7$Average.Weekly.Wages))
```

```
## [1] 1278.411
```

So, the actual weekly wages in the private industry is way less, at just 1278.411$ for 2021.

Regression

Linear Regression

```
library(ggpubr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✔ tibble  3.1.8     ✔ stringr 1.4.1
## ✔ tidyr   1.2.1     ✔ forcats 0.5.2
## ✔ purrr   0.3.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ ggplot2::%+%()   masks psych::%+%()
## ✖ ggplot2::alpha() masks psych::alpha()
## ✖ dplyr::filter()  masks stats::filter()
## ✖ dplyr::lag()     masks stats::lag()
```

```
library(confintr)
```

```
## Warning: package 'confintr' was built under R version 4.2.2
```

```
clean_dataa <- clean_data[clean_data$Year == 2021,]
clean_datae <- clean_data[,10:15]
linearmodel = lm(Average.Weekly.Wages ~ Average.Monthly.Employment +
X1st.Month.Emp + X2nd.Month.Emp + X3rd.Month.Emp + Total.Wages..All.Workers.,
data = clean_datae)
print(linearmodel)
```

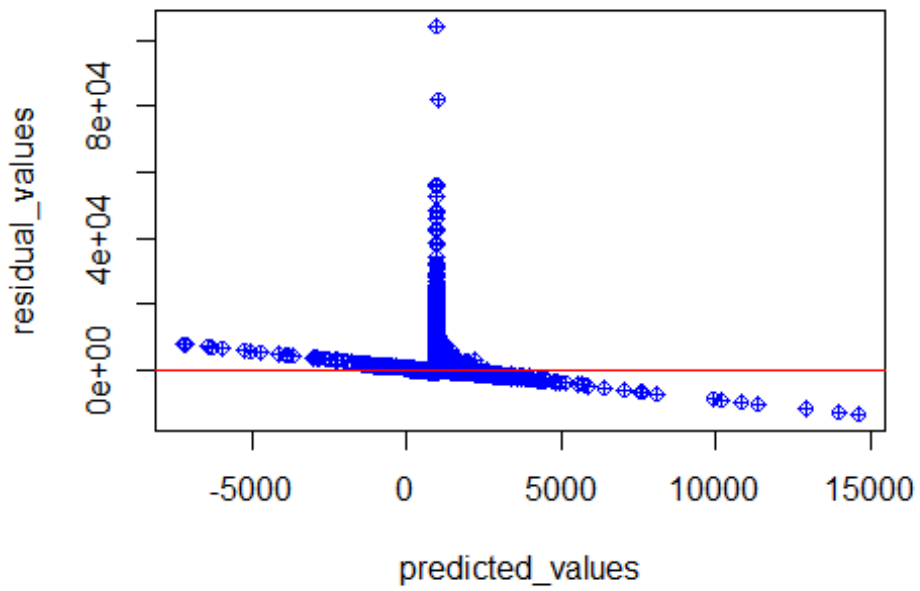
```
##
## Call:
```

```
## lm(formula = Average.Weekly.Wages ~ Average.Monthly.Employment +
##     X1st.Month.Emp + X2nd.Month.Emp + X3rd.Month.Emp +
Total.Wages..All.Workers.,
##     data = clean_datae)
##
## Coefficients:
##                (Intercept)  Average.Monthly.Employment
##                  9.691e+02                   -2.816e-04
##             X1st.Month.Emp              X2nd.Month.Emp
##                  7.607e-04                   -8.146e-04
##             X3rd.Month.Emp   Total.Wages..All.Workers.
##                  2.688e-04                    5.568e-09
```

```
summary(linearmodel)
```

```
##
## Call:
## lm(formula = Average.Weekly.Wages ~ Average.Monthly.Employment +
##     X1st.Month.Emp + X2nd.Month.Emp + X3rd.Month.Emp +
Total.Wages..All.Workers.,
##     data = clean_datae)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13308   -402   -127    227 104180
##
## Coefficients:
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  9.691e+02  3.243e-01 2987.780  < 2e-16 ***
## Average.Monthly.Employment -2.816e-04  3.113e-06  -90.454  < 2e-16 ***
## X1st.Month.Emp               7.607e-04  7.427e-05   10.242  < 2e-16 ***
## X2nd.Month.Emp              -8.146e-04  1.053e-04   -7.737 1.02e-14 ***
## X3rd.Month.Emp               2.688e-04  3.952e-05    6.801 1.04e-11 ***
## Total.Wages..All.Workers.    5.568e-09  6.005e-11   92.721  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 688.2 on 4505602 degrees of freedom
## Multiple R-squared:  0.001918,   Adjusted R-squared:  0.001917
## F-statistic:  1731 on 5 and 4505602 DF,  p-value: < 2.2e-16
```

```
predicted_values <- predict(linearmodel)
residual_values <- resid(linearmodel)
plot(predicted_values, residual_values, col="blue", pch = 10)+abline(0,0,col
= 'red')
```




```
## integer(0)
```

Based on the R-squared, we come to know that only 0.19% of average wages can be predicted by our model, which is not a good statistic. So, we will try fitting in logistic regression to the data and see how that works.

Logistic Regression

```
logit = glm(Average.Weekly.Wages ~ Average.Monthly.Employment +
X1st.Month.Emp + X2nd.Month.Emp + X3rd.Month.Emp + Total.Wages..All.Workers.,
data = clean_datae)
summary(logit)

##
## Call:
## glm(formula = Average.Weekly.Wages ~ Average.Monthly.Employment +
##     X1st.Month.Emp + X2nd.Month.Emp + X3rd.Month.Emp +
Total.Wages..All.Workers.,
##     data = clean_datae)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -13308    -402    -127     227  104180
##
## Coefficients:
##                                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                   9.691e+02  3.243e-01 2987.780  < 2e-16 ***
## Average.Monthly.Employment -2.816e-04  3.113e-06  -90.454  < 2e-16 ***
```

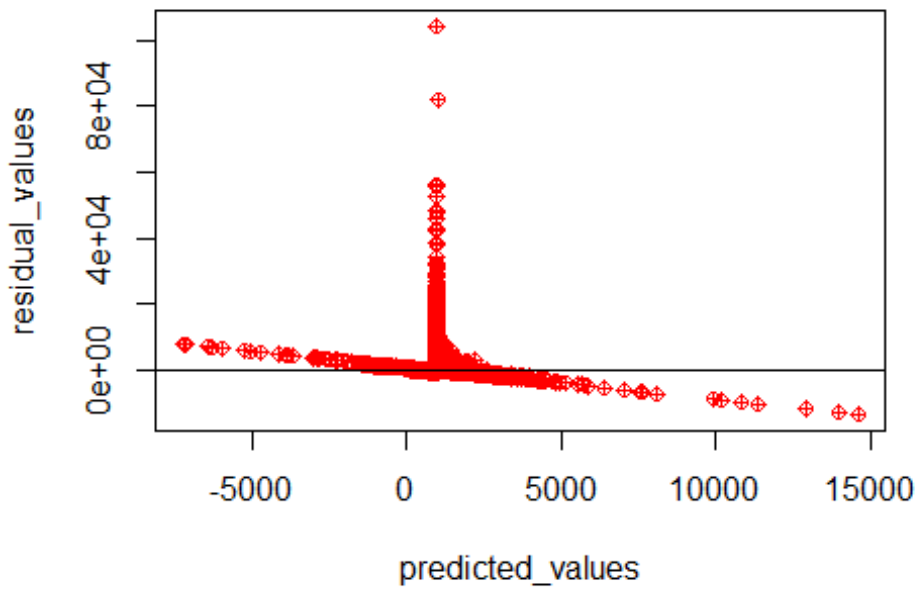
```
## X1st.Month.Emp              7.607e-04  7.427e-05   10.242  < 2e-16 ***
## X2nd.Month.Emp             -8.146e-04  1.053e-04   -7.737 1.02e-14 ***
## X3rd.Month.Emp              2.688e-04  3.952e-05    6.801 1.04e-11 ***
## Total.Wages..All.Workers.   5.568e-09  6.005e-11   92.721  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 473579.4)
##
##     Null deviance: 2.1379e+12  on 4505607  degrees of freedom
## Residual deviance: 2.1338e+12  on 4505602  degrees of freedom
## AIC: 71665993
##
## Number of Fisher Scoring iterations: 2
```

```
confint(logit)
```

```
## Waiting for profiling to be done...
```

```
##                                    2.5 %        97.5 %
## (Intercept)                 9.684292e+02  9.697006e+02
## Average.Monthly.Employment -2.877163e-04 -2.755122e-04
## X1st.Month.Emp              6.151334e-04  9.062618e-04
## X2nd.Month.Emp             -1.020971e-03 -6.082667e-04
## X3rd.Month.Emp              1.913320e-04  3.462502e-04
## Total.Wages..All.Workers.   5.450591e-09  5.685999e-09
```

```
predicted_values <- predict(logit)
residual_values <- resid(logit)
plot(predicted_values, residual_values, col="red", pch = 10)+abline(0,0,col =
'black')
```




```
## integer(0)
```

The logistic regression model, like the linear regression model, doesn't seem to give good predictions for the data.So, we cannot use them for realtime predictions. We can see that for some predicted values, we see the residual error is increasing as the value of predicted values and the points are scattered above and below the line. For the constant variance, the points should ideally look like a uniform cluster of points. So, constant variance assumption is not met for some data points. We have to design even more complex regression hypothesis to fit into the data perfectly.

Going forward, I would like to do this and design a perfect hypothesis that could give us the best possible results.

5. Conclusion: Summarize your findings and include a discussion of what you have learned about your data through this project. You may also want to include limitations of your approach and include ideas for possible future work.

Based on the data, I have learnt how different industries offer different employment wages to people across California, how different counties across California have different average wages that directly impact their quality of life. Based on the data, the State Government of California can utilize their resources in an even better way to facilitate the needy and the downtrodden. I have also taken some hypothesis into consideration and have seen how true they hold against the data that we have. Through correlation of the data we came to know how the dependent and independent variables were related. Also I learnt how categorical variables need to be converted to the numeric types before passing it to the

model. I also learnt how to carve out the specific data that we want to process the required task. So, using all the above methods, I was able to see the economic condition of people across different counties of california. The limitation would be that always the data is not going to be linearly related to the target variable so that it would be difficult to fit some linear regression model in that case. Future scope is that I would like to make the graphs a bit more interactive where I could work with the live datasets. We can use tableau and PowerBI for more interactive graphs. I would also add more models which fits the points very perfectly compared to the above used linear and logistic regressions.Developing an even complex and better hypothesis that could fit into the data would be of high priority for me.

6. References: Include links that you have referenced for this project.

https://data.ca.gov/dataset/quarterly-census-of-employment-and-wages-qcew/resource/efdcc006-bcaf-4066-a763-aef58514a7dd

https://www.bls.gov/regions/west/news-release/2021/countyemploymentandwages_california_20210806.htm#:~:text=Lake%20County%20%28%24912%29%20had%20the%20lowest%20weekly%20wage,wages%20of%20%241%2C400%20or%20higher.%20%28See%20chart%203.%29

https://www.bls.gov/charts/county-employment-and-wages/percent-change-aww-by-state.htm