Assignment - 1 : R Programming

2023-02-04

Team Number: 08

- Ashwin Bezliel Mathew (101415428)
- Zeel Malaviya (101420088)
- Aditya Anupam Shukla (101421678)
- Areeba Zubair (101455510)
- Jasleen Kaur Arora (101412147)
- Ishita Saha (101396418)

DATASET

The Data set Choose for the analysis is Employee data set named “Employee_Data”

To import and view the data set

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   1.0.1
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.3.0      ✓ stringr 1.5.0
## ✓ readr   2.1.3      ✓ forcats 1.0.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(readxl)
Employee_data <- read_excel("Employee_data.xlsx")
View(Employee_data)
```

ANALYSIS

1) Print the structure of your dataset

Ans: To print the Structure of the data set

```
str(Employee_data)
```

```
## tibble [1,000 × 13] (S3: tbl_df/tbl/data.frame)
##  $ EEID         : chr [1:1000] "E02387" "E04105" "E02572" "E02832" ...
##  $ Full Name    : chr [1:1000] "Emily Davis" "Theodore Dinh" "Luna
```

```
Sanders" "Penelope Jordan" ...
##  $ Job Title    : chr [1:1000] "Sr. Manger" "Technical Architect"
"Director" "Computer Systems Manager" ...
##  $ Department   : chr [1:1000] "IT" "IT" "Finance" "IT" ...
##  $ Business Unit: chr [1:1000] "Research & Development" "Manufacturing"
"Speciality Products" "Manufacturing" ...
##  $ Gender       : chr [1:1000] "Female" "Male" "Female" "Female" ...
##  $ Ethnicity    : chr [1:1000] "Black" "Asian" "Caucasian" "Caucasian" ...
##  $ Age          : num [1:1000] 55 59 50 26 55 57 27 25 29 34 ...
##  $ Hire Date    : POSIXct[1:1000], format: "2016-04-08" "1997-11-29" ...
##  $ Annual Salary: num [1:1000] 141604 99975 163099 84913 95409 ...
##  $ Bonus %      : num [1:1000] 0.15 0 0.2 0.07 0 0 0.1 0 0.06 0 ...
##  $ Country      : chr [1:1000] "United States" "China" "United States"
"United States" ...
##  $ City         : chr [1:1000] "Seattle" "Chongqing" "Chicago" "Chicago"
...
```

2) List the variables in your dataset

Ans:

```
colnames(Employee_data)

##  [1] "EEID"          "Full Name"     "Job Title"     "Department"
##  [5] "Business Unit" "Gender"        "Ethnicity"     "Age"
##  [9] "Hire Date"     "Annual Salary" "Bonus %"       "Country"
## [13] "City"
```

3) Print the top 15 rows of your dataset

Ans:

```
head(Employee_data,15)

## # A tibble: 15 × 13
##    EEID   `Full Name`     `Job Title`       Depar…¹ Busin…² Gender Ethni…³
Age
##    <chr>  <chr>           <chr>             <chr>   <chr>   <chr>  <chr>
<dbl>
##  1 E02387 Emily Davis     Sr. Manger        IT      Resear… Female Black
55
##  2 E04105 Theodore Dinh   Technical Archit… IT      Manufa… Male   Asian
59
##  3 E02572 Luna Sanders    Director          Finance Specia… Female Caucas…
50
##  4 E02832 Penelope Jordan Computer Systems… IT      Manufa… Female Caucas…
26
##  5 E01639 Austin Vo       Sr. Analyst       Finance Manufa… Male   Asian
55
##  6 E00644 Joshua Gupta    Account Represen… Sales   Corpor… Male   Asian
57
```

```
##  7 E01550 Ruby Barnes     Manager           IT      Corpor… Female Caucas…
27
##  8 E04332 Luke Martin     Analyst           Finance Manufa… Male   Black
25
##  9 E04533 Easton Bailey   Manager           Accoun… Manufa… Male   Caucas…
29
## 10 E03838 Madeline Walker Sr. Analyst       Finance Specia… Female Caucas…
34
## 11 E00591 Savannah Ali    Sr. Manger        Human … Manufa… Female Asian
36
## 12 E03344 Camila Rogers   Controls Engineer Engine… Specia… Female Caucas…
27
## 13 E00530 Eli Jones       Manager           Human … Manufa… Male   Caucas…
59
## 14 E04239 Everleigh Ng    Sr. Manger        Finance Resear… Female Asian
51
## 15 E03496 Robert Yang     Sr. Analyst       Accoun… Specia… Male   Asian
31
## # … with 5 more variables: `Hire Date` <dttm>, `Annual Salary` <dbl>,
## #   `Bonus %` <dbl>, Country <chr>, City <chr>, and abbreviated variable
names
## #   ¹Department, ²`Business Unit`, ³Ethnicity
```

4) Write a user defined function using any of the variables from the data set

Ans: We created a function to calculate average age

```
calculate_average_Age <- function(Age)
  {
   mean(Age)
  }

average_Age <- calculate_average_Age(Employee_data$Age)
average_Age
```

```
## [1] 44.382
```

5) Use data manipulation techniques and filter rows based on any logical criteria that exist in

Ans: We are filtering out Employees with high bounus i.e. 40% or above

```
library(dplyr)
High_Bonus <- filter(Employee_data,`Bonus %`> .39)
High_Bonus
```

```
## # A tibble: 8 × 13
##   EEID  Full …¹ Job T…² Depar…³ Busin…⁴ Gender Ethni…⁵   Age `Hire Date`
##   <chr> <chr>   <chr>   <chr>   <chr>   <chr>  <chr>   <dbl> <dttm>
## 1 E007… Thomas… Vice P… Market… Resear… Male   Latino     57 2003-07-26
00:00:00
```

```
## 2 E024… Sophie… Vice P… Engine… Corpor… Female Latino     28 2017-07-06
00:00:00
## 3 E000… Isla W… Vice P… Accoun… Corpor… Female Asian      56 2014-03-16
00:00:00
## 4 E014… Mason … Vice P… Accoun… Resear… Male   Asian      59 2011-05-18
00:00:00
## 5 E047… Kinsle… Vice P… Accoun… Corpor… Female Latino     33 2020-12-16
00:00:00
## 6 E049… Elena … Vice P… Engine… Manufa… Female Asian      50 2008-10-13
00:00:00
## 7 E025… Emily … Vice P… Accoun… Corpor… Female Caucas…    36 2020-01-13
00:00:00
## 8 E032… Christ… Vice P… Accoun… Manufa… Male   Asian      64 2013-03-29
00:00:00
## # … with 4 more variables: `Annual Salary` <dbl>, `Bonus %` <dbl>,
## #   Country <chr>, City <chr>, and abbreviated variable names ¹`Full
Name`,
## #   ²`Job Title`, ³Department, ⁴`Business Unit`, ⁵Ethnicity
```

6) Identify the dependent & independent variables and use reshaping techniques and create a new data frame by joining those variables from your dataset.

Ans: Identify the dependent & independent variables. Let's say, dependent variable is Annual Salary and independent variables are Age and Bonus %

```
dep_var <- Employee_data$`Annual Salary`
indep_vars <- Employee_data[c("Age", "Bonus %")]
indep_vars
```

```
## # A tibble: 1,000 × 2
##      Age `Bonus %`
##    <dbl>     <dbl>
##  1    55      0.15
##  2    59      0
##  3    50      0.2
##  4    26      0.07
##  5    55      0
##  6    57      0
##  7    27      0.1
##  8    25      0
##  9    29      0.06
## 10    34      0
## # … with 990 more rows
```

7) Create a new data frame by joining dependent and independent variables

Ans

```
Employee_data_new <- cbind(dep_var, indep_vars)
head(Employee_data_new,5)
```

```
##   dep_var Age Bonus %
## 1  141604  55    0.15
## 2   99975  59    0.00
## 3  163099  50    0.20
## 4   84913  26    0.07
## 5   95409  55    0.00
```

8) Remove missing values in your dataset.

Ans:

```
Employee_data_new_clean <-
Employee_data_new[complete.cases(Employee_data_new),]
```

9) Identify and remove duplicated data in your dataset

Ans:

```
Employee_data_new_clean <- unique(Employee_data_new_clean)
```

10) Reorder multiple rows in descending order

Ans:

```
Employee_data_new_clean %>% head(15,) %>%arrange(desc(Age))

##    dep_var Age Bonus %
## 1    99975  59    0.00
## 2   105086  59    0.09
## 3    50994  57    0.00
## 4   141604  55    0.15
## 5    95409  55    0.00
## 6   146742  51    0.10
## 7   163099  50    0.20
## 8   157333  36    0.15
## 9    77203  34    0.00
## 10   97078  31    0.00
## 11  113527  29    0.06
## 12  119746  27    0.10
## 13  109851  27    0.00
## 14   84913  26    0.07
## 15   41336  25    0.00
```

11) Rename some of the column names in your dataset

Ans:

```
colnames(Employee_data_new_clean) <- c("Annual_Salary", "Age",
"Bonus_Percentage")
```

12) Add new variables in your data frame by using a mathematical function (for e.g. – multiply an existing column by 2 and add it as a new variable to your data frame)

Ans:

```
Employee_data_new_clean$Double_Annual_Salary <- 2 *
Employee_data_new_clean$Annual_Salary
str(Employee_data_new_clean)
```

```
## 'data.frame':    1000 obs. of  4 variables:
##  $ Annual_Salary       : num  141604 99975 163099 84913 95409 ...
##  $ Age                 : num  55 59 50 26 55 57 27 25 29 34 ...
##  $ Bonus_Percentage    : num  0.15 0 0.2 0.07 0 0 0.1 0 0.06 0 ...
##  $ Double_Annual_Salary: num  283208 199950 326198 169826 190818 ...
```

13) Create a training set using random number generator engine.

Ans:

```
set.seed(123)
training_set_index <- sample(1:nrow(Employee_data_new_clean), 0.8 *
nrow(Employee_data_new_clean))
training_set <- Employee_data_new_clean[training_set_index, ]
```

14) Print the summary statistics of your dataset

Ans:

```
summary(Employee_data_new_clean)
```

```
##  Annual_Salary         Age        Bonus_Percentage  Double_Annual_Salary
##  Min.   : 40063   Min.   :25.00   Min.   :0.00000   Min.   : 80126
##  1st Qu.: 71430   1st Qu.:35.00   1st Qu.:0.00000   1st Qu.:142861
##  Median : 96557   Median :45.00   Median :0.00000   Median :193114
##  Mean   :113217   Mean   :44.38   Mean   :0.08866   Mean   :226435
##  3rd Qu.:150782   3rd Qu.:54.00   3rd Qu.:0.15000   3rd Qu.:301565
##  Max.   :258498   Max.   :65.00   Max.   :0.40000   Max.   :516996
```

15) Use any of the numerical variables from the dataset and perform the following statistical functions • Mean • Median • Mode • Range

Ans:

Mean

```
mean(Employee_data_new_clean$Annual_Salary)
```

```
## [1] 113217.4
```

Median

```
median(Employee_data_new_clean$Annual_Salary)
```

```
## [1] 96557
```

Mode

```
mode(Employee_data_new_clean$Annual_Salary)
```

```
## [1] "numeric"
```

Range

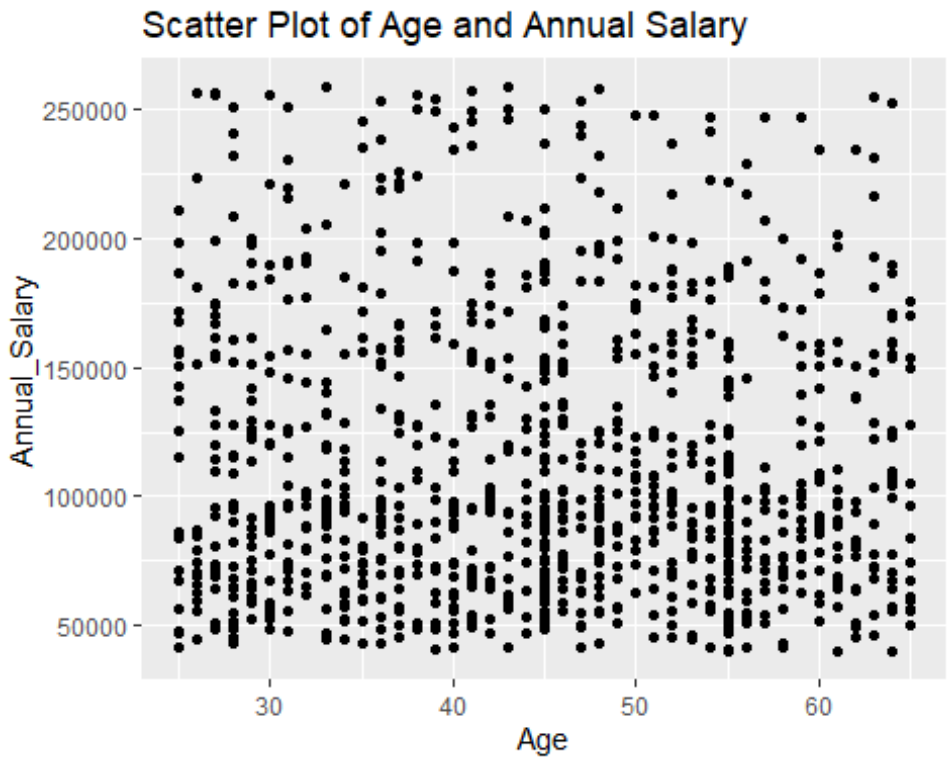
```
range(Employee_data_new_clean$Annual_Salary)

## [1]  40063 258498
```

16) Plot a scatter plot for any 2 variables in your dataset

Ans:

```
ggplot(Employee_data_new_clean, aes(x = Age, y = Annual_Salary)) +
  geom_point() +
  ggtitle("Scatter Plot of Age and Annual Salary")
```



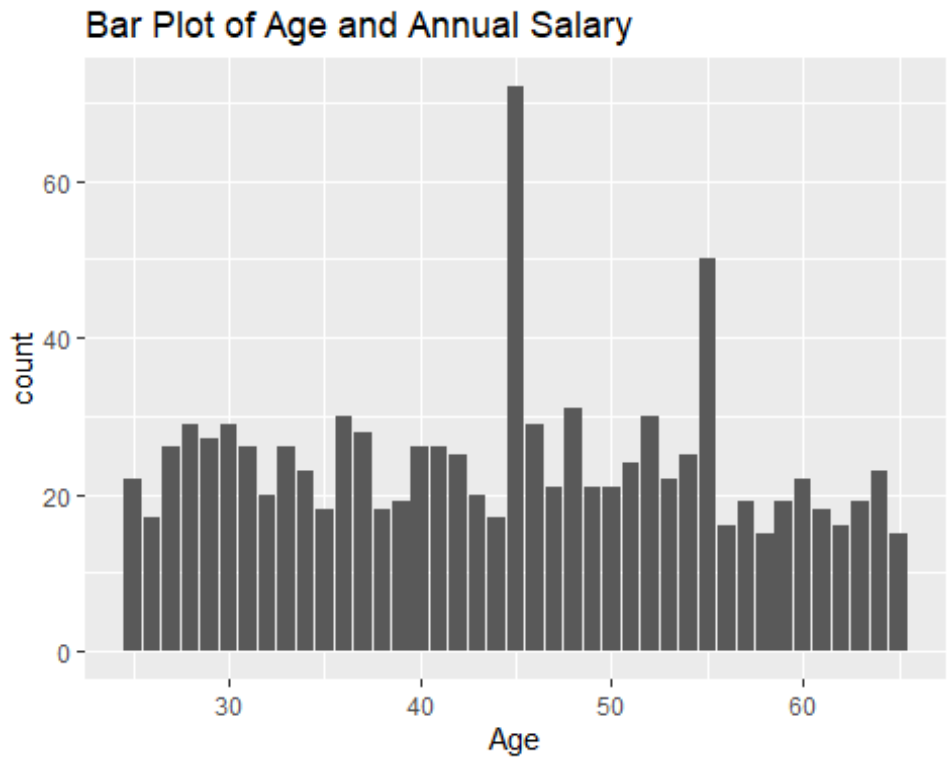
17) Plot a bar plot for any 2 variables in your dataset

Ans:

```
ggplot(Employee_data_new_clean, aes(x = Age, fill = Annual_Salary)) +
  geom_bar(position = "dodge") +
  ggtitle("Bar Plot of Age and Annual Salary") +
scale_color_brewer(palette="Accent")

## Warning: The following aesthetics were dropped during statistical
transformation: fill
## ℹ This can happen when ggplot fails to infer the correct grouping
structure in
```

```
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



18) Find the correlation between any 2 variables by applying least square linear regression model

Ans:

```
model <- lm(Annual_Salary ~ Age, data = Employee_data_new_clean)
summary(model)

##
## Call:
## lm(formula = Annual_Salary ~ Age, data = Employee_data_new_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -74050 -41947 -16785  37459 145268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 118182.4     6897.9  17.133   <2e-16 ***
## Age           -111.9      150.7  -0.743    0.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53560 on 998 degrees of freedom
```

```
## Multiple R-squared:  0.0005521,  Adjusted R-squared:  -0.0004493
## F-statistic: 0.5513 on 1 and 998 DF,  p-value: 0.4579
```