Project

2023-05-25

Students Name: Students ID : Section:

Monerah almobarak 442002988 91S

Nada Alotaibi 442003374 91S

Sarah Altaweel 442000786 91S

Sarah Aljuhani 442005104 91S

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.1     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── 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
```

```
Data<-read.csv("C:\\PNU\\semester9\\BI\\Projuct\\Faculty_Data.csv")
Data
```

ID <int>	Rank <chr>	Experience <int>	Salary <int>
1	Prof	18	139750
2	Prof	16	173200
3	AsstProf	3	79750
4	Prof	39	115000
5	Prof	41	141500
6	AssocProf	6	97000
7	Prof	23	175000
8	Prof	45	147765
9	Prof	20	119250
10	Prof	18	129000

```
view(Data)
```

1.Provide a brief description of the dataset (population, observations, variables' types).

```
#str() function: This function provides a compact way to display the structure of an R object, including it
s type, dimensions, and contents.
str(Data)
```

```
## 'data.frame':    403 obs. of  4 variables:
##  $ ID        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Rank      : chr  "Prof" "Prof" "AsstProf" "Prof" ...
##  $ Experience: int  18 16 3 39 41 6 23 45 20 18 ...
##  $ Salary    : int  139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ...
```

it is a data frame withe '403' observations and '4 'variables.

```
#Using the summary() function: This function provides a summary of the main characteristics of the variable
s in a dataset, such as minimum and maximum values, mean, median, and quartiles.
summary(Data)
```

```
##        ID            Rank             Experience        Salary
##  Min.   :  1.0   Length:403         Min.   :  0.00   Min.   : 57800
##  1st Qu.:100.5   Class :character   1st Qu.:  7.00   1st Qu.: 91025
##  Median :201.0   Mode  :character   Median : 16.50   Median :107175
##  Mean   :200.4                      Mean   : 17.95   Mean   :113478
##  3rd Qu.:300.5                      3rd Qu.: 27.00   3rd Qu.:133975
##  Max.   :400.0                      Max.   :150.00   Max.   :231545
##                                     NA's   :3        NA's   :1
```

```
#Using the head() function: This function displays the first few rows of a dataset.
head(Data)
```

	ID	Rank	Experience	Salary
	<int>	<chr>	<int>	<int>
1	1	Prof	18	139750
2	2	Prof	16	173200
3	3	AsstProf	3	79750
4	4	Prof	39	115000
5	5	Prof	41	141500
6	6	AssocProf	6	97000

6 rows

```
#Using the dim() function: This function provides the dimensions of a dataset, which can be useful for dete
rmining how many rows and columns the data has.
dim(Data)
```

```
## [1] 403   4
```

```
#Using the names() function: This function displays the names of the variables in a dataset, which can be u
seful for identifying which columns contain which information.
names(Data)
```

```
## [1] "ID"         "Rank"       "Experience" "Salary"
```

2.Consider the quality factors and provide a quality report on the raw data.

We considered six quality factors: 1- uniqueness 2- completeness 3- validity 4- consistency 5- relevancy 6- timeliness

the next code targets the quality dimension of uniqueness by evaluating whether there are any duplicate entries in the dataset. This is crucial because duplicates directly affect the distinctiveness and uniqueness of individual data instances within the dataset.

```
# Check for duplicates
has_duplicates <- duplicated(Data)
# Print the result
if (any(has_duplicates)) {
  print("The dataset has duplicates.")
} else {
  print("The dataset does not have duplicates.")
}
```

```
## [1] "The dataset has duplicates."
```

```
# Print the number of rows before removing duplicates
print(paste("Number of rows before removing duplicates:", nrow(Data)))
```

```
## [1] "Number of rows before removing duplicates: 403"
```

After running the code, it becomes clear that the dataset contains duplicate entries, indicating a lack of complete uniqueness. To ensure full uniqueness, additional steps such as thorough data cleaning may be necessary.

```r
# Check for duplicates in the ID column
has_duplicates_ID <- duplicated(Data$ID)
# Print the result
if (any(has_duplicates_ID)) {
  print("The dataset contains duplicate IDs.")
} else {
  print("The dataset does not have any duplicate IDs.")
}
```

```
## [1] "The dataset contains duplicate IDs."
```

```r
# Print the number of rows before removing duplicate ID
print(paste("Number of rows before removing duplicate ID:", nrow(Data)))
```

```
## [1] "Number of rows before removing duplicate ID: 403"
```

the next code targets the data quality dimension of completeness by evaluating whether the dataset contains any missing values.

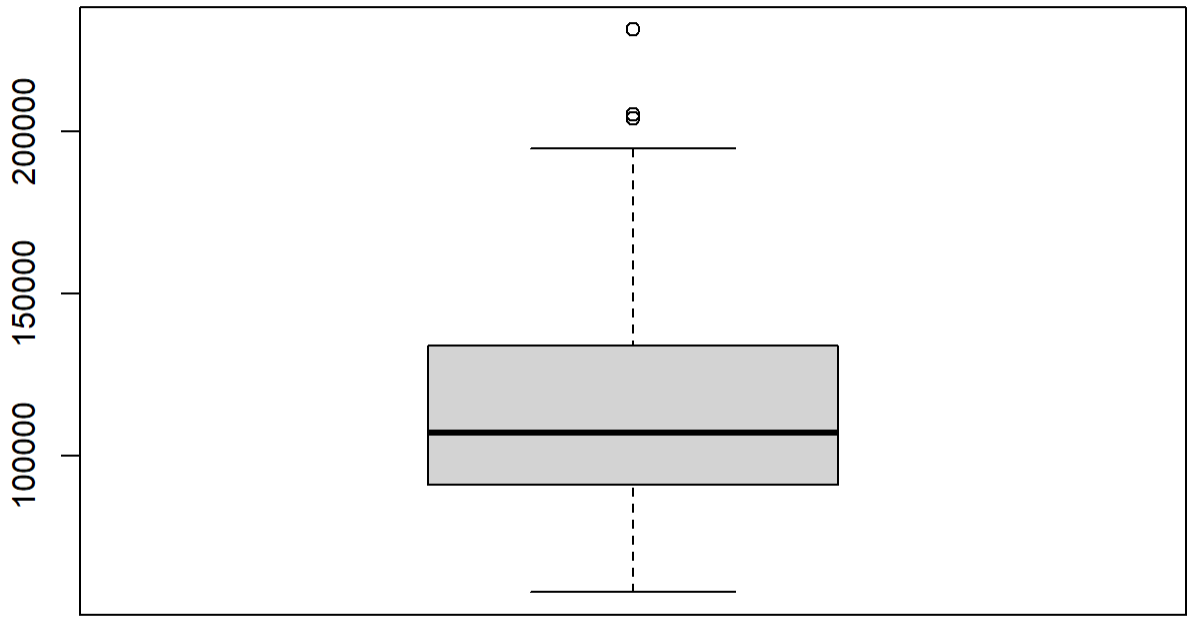
```r
# Check for missing values
has_missing_values <- sum(is.na(Data))
# Print the result
if (has_missing_values > 0) {
print(paste("The dataset contains", has_missing_values, "missing value(s)."))
} else {
  print("The dataset does not have any missing values.")
}
```

```
## [1] "The dataset contains 4 missing value(s)."
```

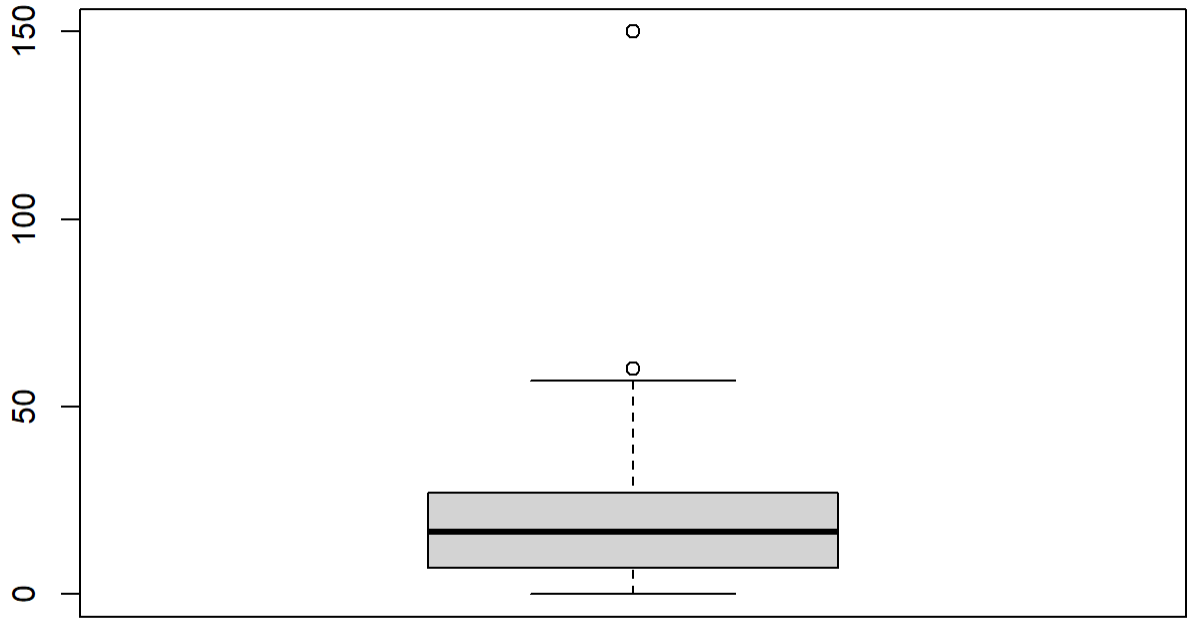
After running the code, it becomes apparent that the dataset contains missing values. This suggests the presence of incomplete data, requiring further data cleaning.

the next code focuses on the quality dimension of data validity, specifically addressing outliers. By utilizing a box plot, it visually identifies potential outliers in the dataset. The Tukey's fences method is then applied to define the range within which values are considered typical, allowing for the detection and handling of outliers that significantly deviate from the norm.

```r
# Check for outliers in Salary using box plot
boxplot(Data$Salary)
```




```
# Calculate outliers in Salary using the Tukey's fences method
Q1_Salary <- quantile(Data$Salary, 0.25, na.rm = TRUE)  # Calculate the 1st quartile (Q1)
Q3_Salary <- quantile(Data$Salary, 0.75, na.rm = TRUE)  # Calculate the 3rd quartile (Q3)
IQR_Salary <- Q3_Salary - Q1_Salary  # Calculate the interquartile range (IQR)
lower_fence_Salary <- Q1_Salary - 1.5 * IQR_Salary  # Calculate the lower fence
upper_fence_Salary <- Q3_Salary + 1.5 * IQR_Salary  # Calculate the upper fence
# Identify outliers in Salary
outliers_Salary <- Data$Salary < lower_fence_Salary | Data$Salary > upper_fence_Salary
# Check for outliers in Experience using box plot
boxplot(Data$Experience)
```




```
# Calculate outliers in Experience using the Tukey's fences method
Q1_Experience <- quantile(Data$Experience, 0.25, na.rm = TRUE)  # Calculate the 1st quartile (Q1)
Q3_Experience <- quantile(Data$Experience, 0.75, na.rm = TRUE)  # Calculate the 3rd quartile (Q3)
IQR_Experience <- Q3_Experience - Q1_Experience  # Calculate the interquartile range (IQR)
lower_fence_Experience <- Q1_Experience - 1.5 * IQR_Experience  # Calculate the lower fence
upper_fence_Experience <- Q3_Experience + 1.5 * IQR_Experience  # Calculate the upper fence
# Identify outliers in Experience
outliers_Experience <- Data$Experience < lower_fence_Experience | Data$Experience > upper_fence_Experience
# Print the result for Salary
if (any(outliers_Salary)) {
  print("The Salary attribute contains outliers.")
} else {
  print("The Salary attribute does not have any outliers.")
}
```

```
## [1] "The Salary attribute contains outliers."
```

```
# Print the result for Experience
if (any(outliers_Experience)) {
  print("The Experience attribute contains outliers.")
} else {
  print("The Experience attribute does not have any outliers.")
}
```

```
## [1] "The Experience attribute contains outliers."
```

Upon running the code, it becomes apparent that there are outliers present in the Salary attribute, Similarly, the Experience attribute exhibits outliers, which may raise concerns about the data's validity due to the presence of significantly deviating data points.

Although outliers have been identified in the Salary data, it is important to consider their significance in the context of analyzing inequities. Removing these outliers may not be the most appropriate course of action as they can provide valuable insights into potential wage gaps or discriminatory practices within the dataset. These outliers represent extreme values that can help uncover salary disparities and contribute to a comprehensive understanding of the distribution and potential inequities within the salary data. By retaining the outliers, we can conduct further analysis and investigation to identify the underlying factors contributing to these discrepancies and formulate strategies to address them effectively.

On the other hand, the outliers in the Experience data require careful consideration due to their potential impact on our analysis. While outliers can provide valuable insights into unusual cases or data entry errors, extreme outliers can introduce significant distortions and affect the reliability of our results. For instance, the presence of an instance with an experience of 150 years is highly unlikely and is likely an erroneous data point. If not addressed, this outlier can adversely influence our analysis and compromise the accuracy of the relationship between experience and other variables. Therefore, removing such extreme outliers is necessary to ensure the integrity and robustness of our analysis. By eliminating these outliers, we can mitigate their adverse effects and obtain more accurate insights into the relationship between experience and other factors within the dataset.

the next code targets the quality dimension of consistency and correctness by identifying potentially misspelled ranks in the dataset. Ensuring consistent and correct attribute values is crucial for accurate data analysis and reporting. Identifying misspelled ranks allows further investigation and potential correction to maintain data integrity and quality.

```
# Specify the attribute to check for misspelled words (e.g., "Rank")
attribute_to_check <- "Rank"

# Define the expected abbreviations for each rank
expected_abbreviations <- c("Prof", "AssocProf", "AsstProf")

# Perform spell checking for the specified attribute
rank_vector <- Data[[attribute_to_check]]

misspelled_words <- rank_vector[!rank_vector %in% expected_abbreviations]

# Print the misspelled words, if any
if (length(misspelled_words) > 0) {
  print("The 'Rank' attribute contains misspelled words or unexpected abbreviations:")
  print(misspelled_words)
} else {
  print("The 'Rank' attribute does not have any misspelled words or unexpected abbreviations.")
}
```

```
## [1] "The 'Rank' attribute contains misspelled words or unexpected abbreviations:"
## [1] "AssstProf"  "AssocProff" ""
```

After running the code, it is evident that the dataset contains some misspelled ranks. Identifying these misspelled entries is crucial for data quality and consistency.

The data in this dataset possesses relevance, which is a fundamental quality dimension. The attributes, including ID, Salary, Experience, and Rank, capture essential information about faculty members. This data is provided to us by a reliable source, ensuring its credibility and relevancy. These attributes provide valuable insights into faculty-related aspects such as unique identifiers, salary levels, years of service, and position within the academic hierarchy. The relevance of the data lies in its ability

to contribute meaningful information for analyzing salary distributions, identifying experience trends, and understanding the composition of different ranks among faculty members. By ensuring the relevance of the data, sourced from a reliable provider, we can leverage its significance and applicability in various analyses and decision-making processes.

The timeliness of the data is another critical quality dimension to consider in this dataset. This data lacks timeliness and may not capture the most recent or up-to-date information. Timeliness is an essential aspect of data quality as it ensures the relevance and accuracy of the dataset. By recognizing the timeliness dimension, we acknowledge that the dataset provides a snapshot of the attributes (ID, Salary, Experience, and Rank) at a specific point in time.

3. Apply required operations for data cleansing.

```
# Remove duplicate rows from the dataset
Data <- unique(Data)
# Print the updated dataset
print(Data)
```

```
##      ID       Rank Experience Salary
## 1     1       Prof         18 139750
## 2     2       Prof         16 173200
## 3     3   AsstProf          3  79750
## 4     4       Prof         39 115000
## 5     5       Prof         41 141500
## 6     6  AssocProf          6  97000
## 7     7       Prof         23 175000
## 8     8       Prof         45 147765
## 9     9       Prof         20 119250
## 10   10       Prof         18 129000
## 11   11  AssocProf          8 119800
## 12   12   AsstProf          2  79800
## 13   13   AsstProf          1  77700
## 14   14   AsstProf          0  78000
## 15   15       Prof         18 104800
## 16   16       Prof          3 117150
## 17   17       Prof         20 101000
## 18   18       Prof         34 103450
## 19   19       Prof         23 124750
## 20   20       Prof         36 137000
## 21   21       Prof         26  89565
## 22   22       Prof         31 102580
## 23   23       Prof         30  93904
## 24   24       Prof         19 113068
## 25   25  AssocProf          8  74830
## 26   26       Prof          8 106294
## 27   27       Prof         23 134885
## 28   28   AsstProf          3  82379
## 29   29   AsstProf          0  77000
## 30   30       Prof          8 118223
## 31   31       Prof          4 132261
## 32   32   AsstProf          2  79916
## 33   33       Prof          9 117256
## 34   34   AsstProf          2  80225
## 35   35   AsstProf          2  80225
## 36   36   AsstProf          0  77000
## 37   37       Prof         21 155750
## 38   38   AsstProf          4  86373
## 39   39       Prof         31 125196
## 40   40  AssocProf          9 100938
## 41   41       Prof          2 146500
## 42   42  AssocProf         23  93418
## 43   43       Prof         27 101299
## 44   44       Prof         38 231545
## 45   45       Prof         19  94384
## 46   46       Prof         15 114778
## 47   47       Prof         28  98193
## 48   48       Prof         19 151768
## 49   49       Prof         25 140096
## 50   50   AsstProf          1  70768
## 51   51       Prof         28 126621
## 52   52       Prof         11 108875
## 53   53   AsstProf          3  74692
```

```
## 54   54       Prof          9 106639
## 55   55  AssocProf         11 103760
## 56   56  AssocProf          5  83900
## 57   57       Prof         21 117704
## 58   58  AssocProf          8  90215
## 59   59  AssocProf          9 100135
## 60   60   AsstProf          3  75044
## 61   61  AssocProf          8  90304
## 62   62   AsstProf          2  75243
## 63   63       Prof         31 109785
## 65   64  AssocProf         11 103613
## 66   65   AsstProf          3  68404
## 67   66  AssocProf          8 100522
## 68   67       Prof         12 101000
## 69   68       Prof         31  99418
## 70   69       Prof         17 111512
## 71   70       Prof         36  91412
## 72   71       Prof          2 126320
## 73   72       Prof         45 146856
## 74   73       Prof         19 100131
## 75   74       Prof         34  92391
## 76   75       Prof         23 113398
## 77   76   AsstProf          3  73266
## 78   77       Prof          3 150480
## 79   78       Prof         19 193000
## 80   79   AsstProf          1  86100
## 81   80  AssstProf          2  84240
## 82   81       Prof         28 150743
## 83   82       Prof         16 135585
## 84   83       Prof         20 144640
## 85   84       Prof         NA     NA
## 86   85   AsstProf          2  88825
## 87   86       Prof         18 122960
## 88   87       Prof         14 132825
## 89   88       Prof         37 152708
## 90   89   AsstProf          2  88400
## 91   90       Prof         25 172272
## 92   91  AssocProf          7 107008
## 93   92   AsstProf          5  97032
## 94   93 AssocProff          7 105128
## 95   94  AssocProf          7 105631
## 96   95       Prof         38 166024
## 97   96       Prof         20 123683
## 98   97   AsstProf          0  84000
## 99   98  AssocProf         12  95611
## 100  99       Prof          7 129676
## 101 100       Prof         14 102235
## 102 101       Prof         26 106689
## 103 102       Prof         25 133217
## 104 103       Prof         23 126933
## 105 104       Prof          5 153303
## 106 105  AssocProf         NA 100012
## 107 106       Prof         14 127512
## 108 107  AssocProf         10  83850
```

```
## 109 108       Prof         28 113543
## 110 109  AssocProf          8  82099
## 111 110  AssocProf          8  82600
## 112 111  AssocProf          8  81500
## 113 112       Prof         31 131205
## 114 113       Prof         16 112429
## 115 114  AssocProf         16  82100
## 116 115   AsstProf          1  72500
## 117 116       Prof         37 104279
## 118 117       Prof          0 105000
## 119 118       Prof          9 120806
## 120 119       Prof         29 148500
## 121 120       Prof         36 117515
## 122 121   AsstProf          1  72500
## 123 122   AsstProf          3  73500
## 124 123       Prof         14 115313
## 125 124       Prof         32 124309
## 126 125       Prof         22  97262
## 127 126  AssocProf         22  62884
## 128 127       Prof         22  96614
## 129 128       Prof         49  78162
## 130 129       Prof         26 155500
## 131 130   AsstProf          0  72500
## 132 131       Prof         30 113278
## 133 132   AsstProf          2  73000
## 134 133  AssocProf          9  83001
## 135 134       Prof         57  76840
## 136 135  AssocProf          8  77500
## 137 136   AsstProf          1  72500
## 138 137       Prof         25 168635
## 139 138       Prof         18 136000
## 140 139       Prof         14 108262
## 141 140       Prof         14 105668
## 142 141  AssocProf          7  73877
## 143 142       Prof         18 152664
## 144 143  AssocProf          8 100102
## 145 144  AssocProf         10  81500
## 146 145       Prof         11 106608
## 147 146   AsstProf          3  89942
## 148 147       Prof         27 112696
## 149 148       Prof         28 119015
## 150 149   AsstProf          4  92000
## 151 150       Prof         27 156938
## 152 151       Prof         26 144651
## 153 152   AsstProf          3  95079
## 154 153       Prof         12 128148
## 155 154   AsstProf          4  92000
## 156 155       Prof          9 111168
## 157 156  AssocProf         10 103994
## 158 157   AsstProf          0  92000
## 159 158       Prof         21 118971
## 160 159  AssocProf         18 113341
## 161 160   AsstProf          0  88000
## 162 161  AssocProf          6  95408
```

```
## 163 162       Prof         16 137167
## 164 163   AsstProf          2  89516
## 165 164       Prof         19 176500
## 166 165  AssocProf          7  98510
## 167 166   AsstProf          3  89942
## 168 167   AsstProf          0  88795
## 169 168       Prof          8 105890
## 170 169       Prof         16 167284
## 171 170                    NA  80000
## 172 171       Prof         19 130664
## 173 172  AssocProf          6 101210
## 174 173       Prof         18 181257
## 175 174   AsstProf          5  91227
## 176 175       Prof         19 151575
## 177 176       Prof         24  93164
## 178 177       Prof         20 134185
## 179 178  AssocProf          6 105000
## 180 179       Prof         25 111751
## 181 180  AssocProf          7  95436
## 182 181  AssocProf          9 100944
## 183 182       Prof         14 147349
## 184 183   AsstProf          3  92000
## 185 184       Prof         11 142467
## 186 185       Prof          5 141136
## 187 186  AssocProf          8 100000
## 188 187       Prof         22 150000
## 189 188       Prof         23 101000
## 190 189       Prof         30 134000
## 191 190  AssocProf         10 103750
## 192 191       Prof         10 107500
## 193 192  AssocProf         28 106300
## 194 193       Prof         19 153750
## 195 194       Prof          9 180000
## 196 195       Prof         22 133700
## 197 196       Prof         18 122100
## 198 197  AssocProf         19  86250
## 199 198  AssocProf         53  90000
## 200 199  AssocProf          7 113600
## 201 200   AsstProf          4  92700
## 202 201   AsstProf          4  92700
## 203 201   AsstProf          4  92000
## 204 202       Prof         33 189409
## 205 203       Prof         22 114500
## 206 204   AsstProf          4  92700
## 207 205       Prof         40 119700
## 208 206       Prof         17 160400
## 209 207       Prof         17 152500
## 210 208       Prof          5 165000
## 211 209       Prof          2  96545
## 212 210       Prof         33 162200
## 213 211       Prof         18 120000
## 214 212   AsstProf          2  91300
## 215 213       Prof         20 163200
## 216 214   AsstProf          3  91000
```

```
## 217 215       Prof         39 111350
## 218 216       Prof          7 128400
## 219 217       Prof         19 126200
## 220 218  AssocProf          1 118700
## 221 219       Prof         11 145350
## 222 220       Prof         11 146000
## 223 221  AssocProf         22 105350
## 224 222  AssocProf          7 109650
## 225 223       Prof         11 119500
## 226 224       Prof         21 170000
## 227 225       Prof         10 145200
## 228 226  AssocProf          6 107150
## 229 227       Prof         20 129600
## 230 228       Prof         35  87800
## 231 229       Prof         20 122400
## 232 230   AsstProf          1  63900
## 233 231  AssocProf          7  70000
## 234 232       Prof         11  88175
## 235 233       Prof         38 133900
## 236 234       Prof         27  91000
## 237 235  AssocProf         24  73300
## 238 236       Prof         19 148750
## 239 237       Prof         19 117555
## 240 238   AsstProf          3  69700
## 241 239       Prof         17  81700
## 242 240       Prof         25 114000
## 243 241   AsstProf          6  63100
## 244 242       Prof         40  77202
## 245 243       Prof          6  96200
## 246 244   AsstProf          3  69200
## 247 245       Prof         30 122875
## 248 246       Prof         37 102600
## 249 247       Prof         23 108200
## 250 248       Prof         23  84273
## 251 249       Prof         11  90450
## 252 250       Prof         23  91100
## 253 251       Prof         18 101100
## 254 252       Prof         23 128800
## 255 253       Prof          7 204000
## 256 254       Prof         39 109000
## 257 255       Prof          8 102000
## 258 256       Prof         12 132000
## 259 257   AsstProf          2  77500
## 260 258       Prof          7 116450
## 261 259  AssocProf          8  83000
## 262 260       Prof         22 140300
## 263 261  AssocProf         23  74000
## 264 262   AsstProf          3  73800
## 265 263       Prof         30  92550
## 266 264  AssocProf         33  88600
## 267 265       Prof         45 107550
## 268 266       Prof         26 121200
## 269 267       Prof         31 126000
## 270 268       Prof         35  99000
```

```
## 271 269       Prof         30 134800
## 272 270       Prof         43 143940
## 273 271       Prof         10 104350
## 274 272       Prof         44  89650
## 275 273       Prof          7 103700
## 276 274       Prof         40 143250
## 277 275       Prof         18 194800
## 278 276   AsstProf          1  73000
## 279 277   AsstProf          4  74000
## 280 278   AsstProf          3  78500
## 281 279       Prof          6  93000
## 282 280       Prof         48 107200
## 283 281       Prof         27 163200
## 284 282       Prof         18 107100
## 285 283       Prof         46 100600
## 286 284       Prof         38 136500
## 287 285       Prof         27 103600
## 288 286       Prof         51  57800
## 289 287       Prof         43 155865
## 290 288  AssocProf          6  88650
## 291 289  AssocProf         49  81800
## 292 290       Prof         27 115800
## 293 291   AsstProf          0  85000
## 294 292       Prof         27 150500
## 295 293   AsstProf          5  74000
## 296 294       Prof          7 174500
## 297 295       Prof         28 168500
## 298 296       Prof          9 183800
## 299 297  AssocProf          1 104800
## 300 298       Prof          7 107300
## 301 299       Prof         36  97150
## 302 300       Prof         18 126300
## 303 301       Prof         11 148800
## 304 302       Prof         43  72300
## 305 303  AssocProf         39  70700
## 306 304       Prof         36  88600
## 307 305       Prof         16 127100
## 308 306       Prof         13 170500
## 309 307       Prof          4 105260
## 310 308       Prof         44 144050
## 311 309       Prof         31 111350
## 312 310   AsstProf          4  74500
## 313 311       Prof         28 122500
## 314 312   AsstProf          0  74000
## 315 313       Prof         15 166800
## 316 314       Prof          7  92050
## 317 315       Prof          9 108100
## 318 316       Prof         19  94350
## 320 317       Prof         35 100351
## 321 318       Prof          6 146800
## 322 319   AsstProf          3  84716
## 323 320  AssocProf          9  71065
## 324 321       Prof         45  67559
## 325 322       Prof         16 134550
```

```
## 326 323       Prof        150 135027
## 327 324       Prof         23 104428
## 328 325  AssocProf          9  95642
## 329 326  AssocProf         11 126431
## 330 327       Prof         15 161101
## 331 328       Prof         31 162221
## 332 329   AsstProf          4  84500
## 333 330       Prof         15 124714
## 334 331       Prof         37 151650
## 335 332  AssocProf         10  99247
## 336 333       Prof         23 134778
## 337 334       Prof         60 192253
## 338 335       Prof          9 116518
## 339 336       Prof         10 105450
## 340 337       Prof         19 145098
## 341 338  AssocProf          6 104542
## 342 339       Prof         38 151445
## 343 340       Prof         23  98053
## 344 341       Prof         12 145000
## 345 342       Prof         25 128464
## 346 343       Prof         15 137317
## 347 344       Prof         11 106231
## 348 345       Prof         17 124312
## 349 346       Prof         38 114596
## 350 347       Prof         31 162150
## 351 348       Prof         35 150376
## 352 349       Prof         10 107986
## 353 350       Prof         27 142023
## 354 351       Prof         33 128250
## 355 352   AsstProf          3  80139
## 356 353       Prof         28 144309
## 357 354       Prof         49 186960
## 358 355       Prof         38  93519
## 359 356       Prof         27 142500
## 360 357       Prof         20 138000
## 361 358   AsstProf          1  83600
## 362 359       Prof         21 145028
## 363 360       Prof         40  88709
## 364 361       Prof         35 107309
## 365 362       Prof         14 109954
## 366 363   AsstProf          4  78785
## 367 364       Prof         11 121946
## 368 365       Prof         15 109646
## 369 366       Prof         30 138771
## 370 367  AssocProf         17  81285
## 371 368       Prof         43 205500
## 372 369       Prof         40 101036
## 373 370       Prof         10 115435
## 374 371  AssocProf          1 108413
## 375 372       Prof         30 131950
## 376 373       Prof         31 134690
## 377 374  AssocProf          8  78182
## 378 375       Prof         20 110515
## 379 376       Prof          7 109707
```

```
## 380 377       Prof         26 136660
## 381 378       Prof         19 103275
## 382 379       Prof         26 103649
## 383 380   AsstProf          1  74856
## 384 381   AsstProf          3  77081
## 385 382       Prof         38 150680
## 386 383  AssocProf          8 104121
## 387 384   AsstProf          3  75996
## 388 385       Prof         23 172505
## 389 386  AssocProf          5  86895
## 390 387       Prof         44 105000
## 391 388       Prof         21 125192
## 392 389       Prof          9 114330
## 393 390       Prof         27 139219
## 394 391       Prof         15 109305
## 395 392       Prof         36 119450
## 396 393       Prof         18 186023
## 397 394       Prof         19 166605
## 398 395       Prof         19 151292
## 399 396       Prof         30 103106
## 400 397       Prof         19 150564
## 401 398       Prof         25 101738
## 402 399       Prof         15  95329
## 403 400   AsstProf          4  81035
```

```
# Print the number of rows after removing duplicates
print(paste("Number of rows after removing duplicates:", nrow(Data)))
```

```
## [1] "Number of rows after removing duplicates: 401"
```

```
# Remove duplicate rows based on ID column
if (any(has_duplicates_ID)) {
  # Print the result
  Data <- Data[!duplicated(Data$ID), ]
  print("Duplicate ID removed.")
} else {
  print("No duplicate ID found.")
}
```

```
## [1] "Duplicate ID removed."
```

```
# Print the number of rows after removing duplicate ID
print(paste("Number of rows after removing duplicate ID:", nrow(Data)))
```

```
## [1] "Number of rows after removing duplicate ID: 400"
```

```r
# Remove missing values from Data
cleaned_data <- na.omit(Data)
# Check if any rows with missing values were removed
if (nrow(cleaned_data) < nrow(Data)) {
  print("Rows with missing values were removed.")
} else {
  print("No rows with missing values were found.")
}
```

```
## [1] "Rows with missing values were removed."
```

```r
# Assign the cleaned data back to the Data variable
Data <- cleaned_data
```

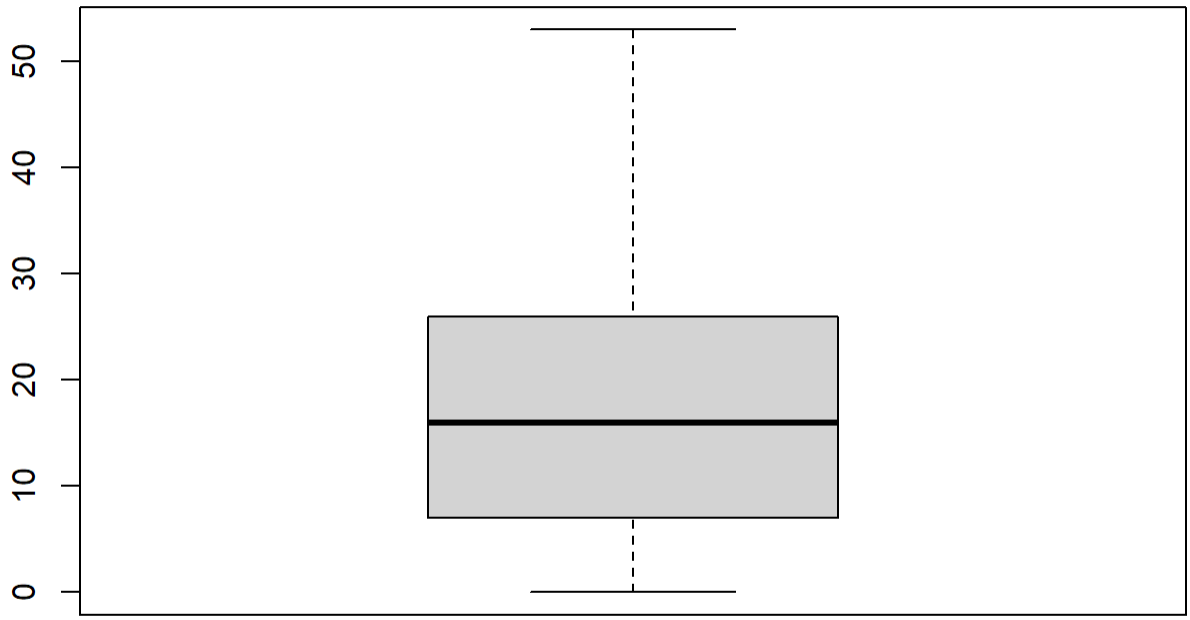
```r
#After Handling missing values .
dim(Data)
```

```
## [1] 397   4
```

```r
# Calculate the median and IQR for Experience
median_Experience <- median(Data$Experience, na.rm = TRUE)
IQR_Experience <- IQR(Data$Experience, na.rm = TRUE)
# Calculate the upper and lower thresholds for outliers
lower_threshold_Experience <- median_Experience - 2* IQR_Experience
upper_threshold_Experience <- median_Experience + 2* IQR_Experience
# Remove outliers from Experience (ignoring missing values)
Data <- subset(Data, Experience >= lower_threshold_Experience & Experience <= upper_threshold_Experience)
# Check if outliers were removed
outliers_Experience_updated <- Data$Experience < lower_threshold_Experience | Data$Experience > upper_thres
hold_Experience
if (sum(outliers_Experience_updated, na.rm = TRUE) == 0) {
  cat("Outliers have been removed from the Experience attribute.\n")
} else {
  cat("Outliers have not been removed from the Experience attribute.\n")
}
```

```
## Outliers have been removed from the Experience attribute.
```

```r
# Boxplot after removing outliers from the "Experience" variable
boxplot(Data$Experience)
```



the next code addresses the quality dimension of consistency and correctness by identifying and correcting misspelled ranks in the dataset. Ensuring consistent and correct attribute values is crucial for accurate data analysis and reporting. By leveraging similarity-based correction, misspelled ranks can be automatically replaced with the most appropriate and accurate values, promoting data integrity and quality.

```
# Load the stringdist library for string distance calculations
library(stringdist)
```

```
##
## Attaching package: 'stringdist'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
# Specify the attribute to correct misspellings (e.g., "Rank")
attribute_to_correct <- "Rank"
# Define the correct spellings for each rank
correct_spellings <- c("Prof", "AssocProf", "AsstProf")
# Create a copy of the attribute to correct
corrected_ranks <- Data[[attribute_to_correct]]
# Identify the misspelled ranks
misspelled_indices <- !is.na(corrected_ranks) & corrected_ranks != "" & !corrected_ranks %in% correct_spell
ings
# Initialize a vector to store the old misspelled ranks and the corresponding corrected ranks
corrections <- data.frame(Old_Rank = character(0), Corrected_Rank = character(0))
# Loop through the misspelled ranks
for (i in which(misspelled_indices)) {
  rank <- corrected_ranks[i]
  # Calculate the Jaro-Winkler distance between the rank and the correct spellings
  distances <- stringdist::stringdist(rank, correct_spellings, method = "jw")
  # Find the index of the correct spelling with the minimum distance
  closest_index <- which.min(distances)
  # Replace the misspelled rank with the closest correct spelling
  corrected_rank <- correct_spellings[closest_index]
  # Store the old misspelled rank and the corrected rank
  corrections <- rbind(corrections, data.frame(Old_Rank = rank, Corrected_Rank = corrected_rank))
  # Update the attribute with the corrected value
  corrected_ranks[i] <- corrected_rank
}
# Update the attribute with the corrected values
Data[[attribute_to_correct]][misspelled_indices] <- corrected_ranks[misspelled_indices]
# Print the corrections if any misspelled ranks are found
if (nrow(corrections) > 0) {
  cat("The following ranks have been corrected:\n")
  print(corrections)
} else {
  cat("No misspelled ranks found.\n")
}
```

```
## The following ranks have been corrected:
##     Old_Rank Corrected_Rank
## 1  AssstProf       AsstProf
## 2 AssocProff      AssocProf
```

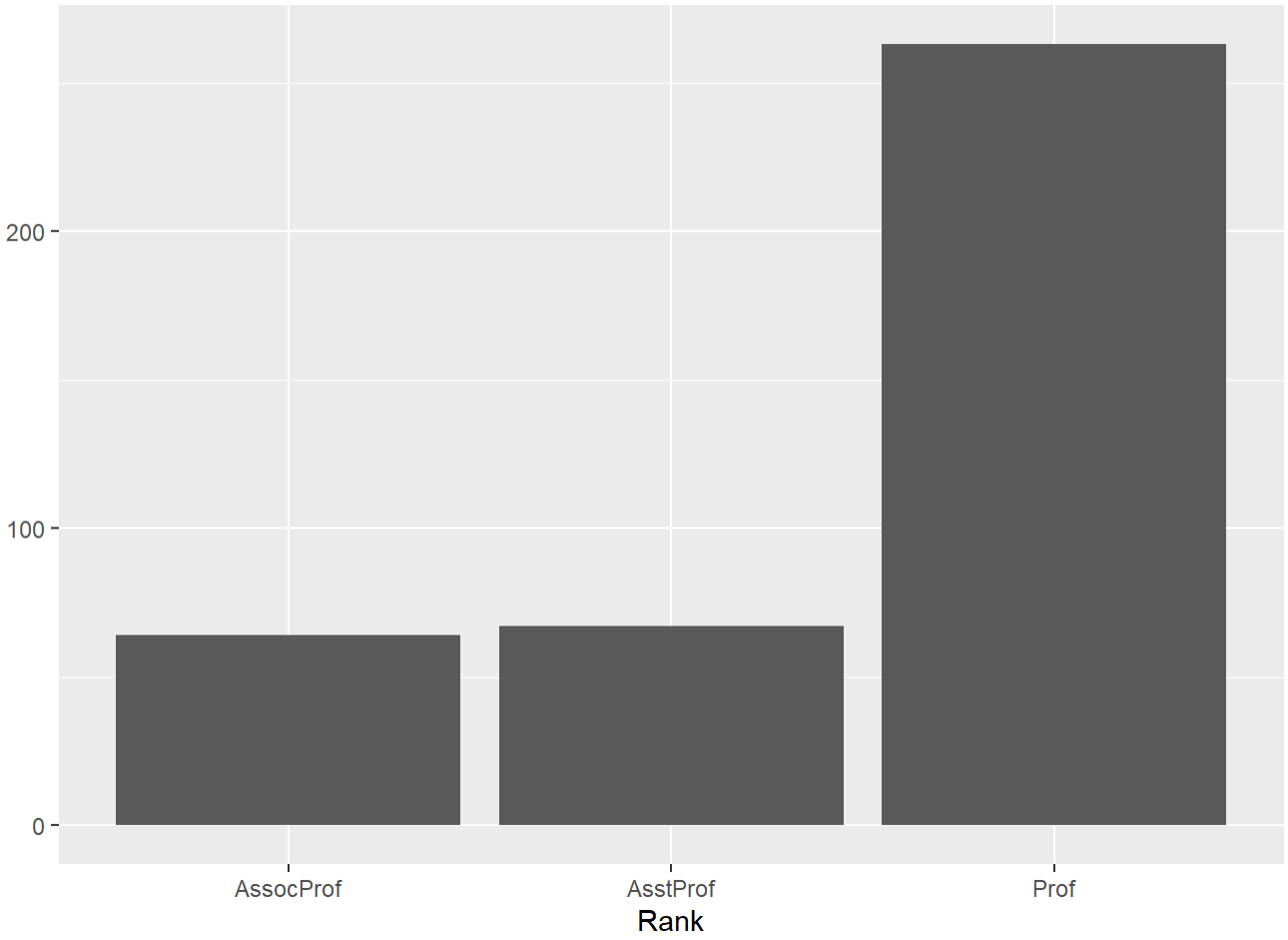
After running the code, it is observed that the dataset contains misspelled ranks, which have now been corrected. By utilizing similarity-based correction, the misspelled ranks were replaced with the most appropriate and accurate values. This correction enhances the consistency and correctness of attribute values.

#4. Provide appropriate plots for each attribute or variable.

```
#Plotting Rank attribute by using bar chart plot
qplot(Rank, data = Data, bins = 10)
```

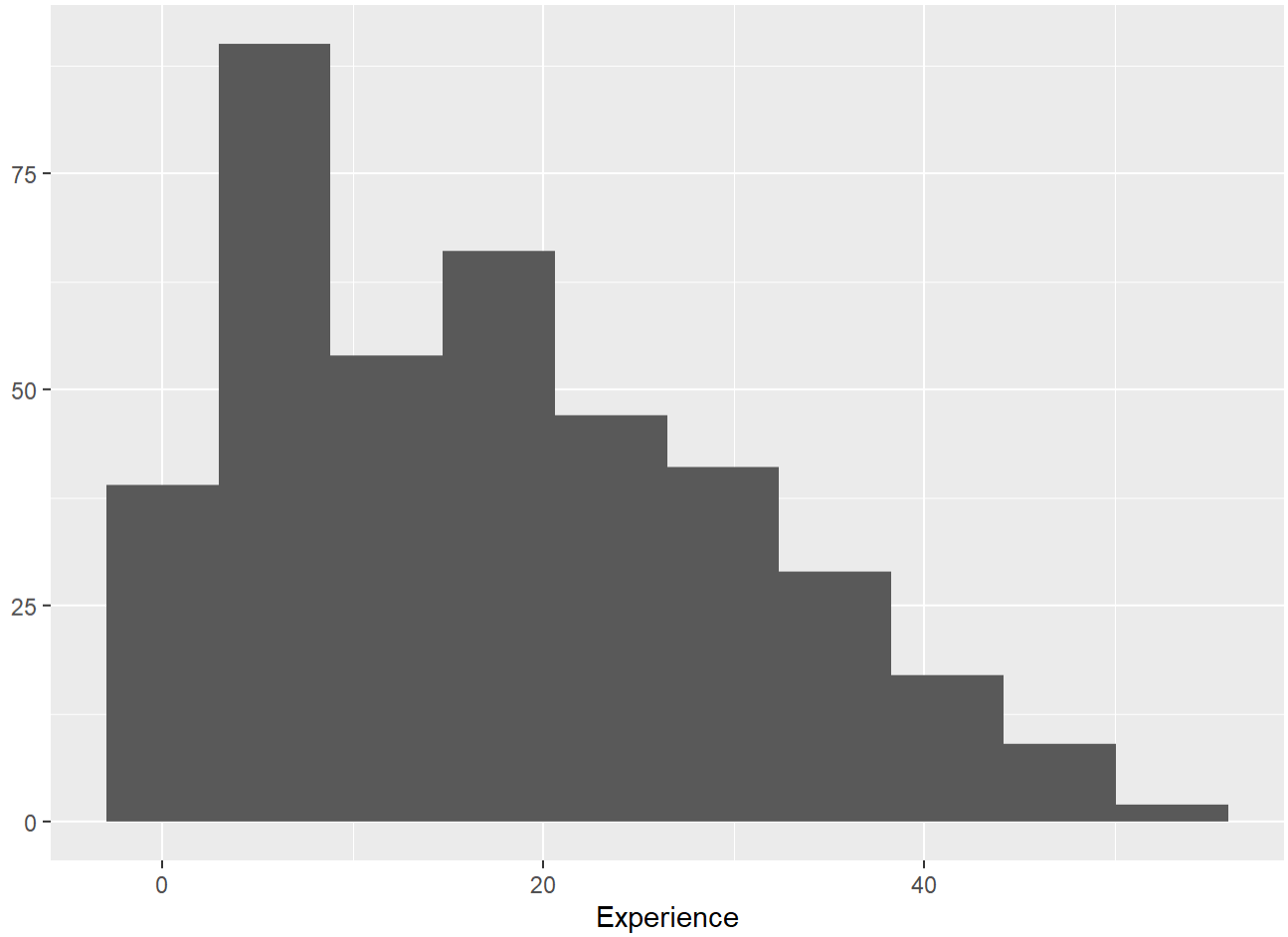
```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning in geom_bar(bins = 10): Ignoring unknown parameters: `bins`
```



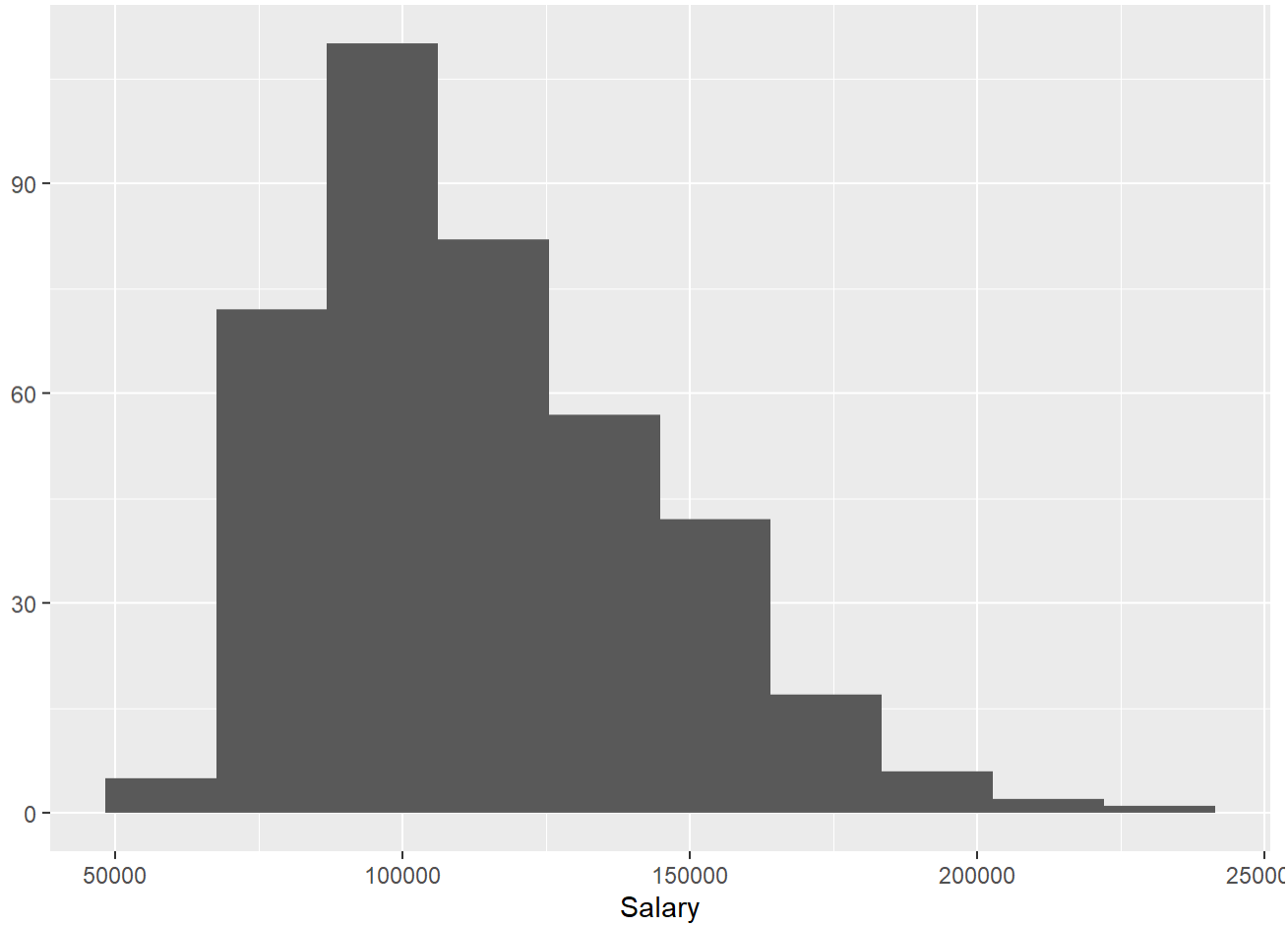
It became clear to us that most of the faculty college members are professors.

```
#Plotting Experience attribute by using histogram plot
qplot(Experience, data = Data, bins = 10)
```



The experience has appeared that have right-skewed distributions.

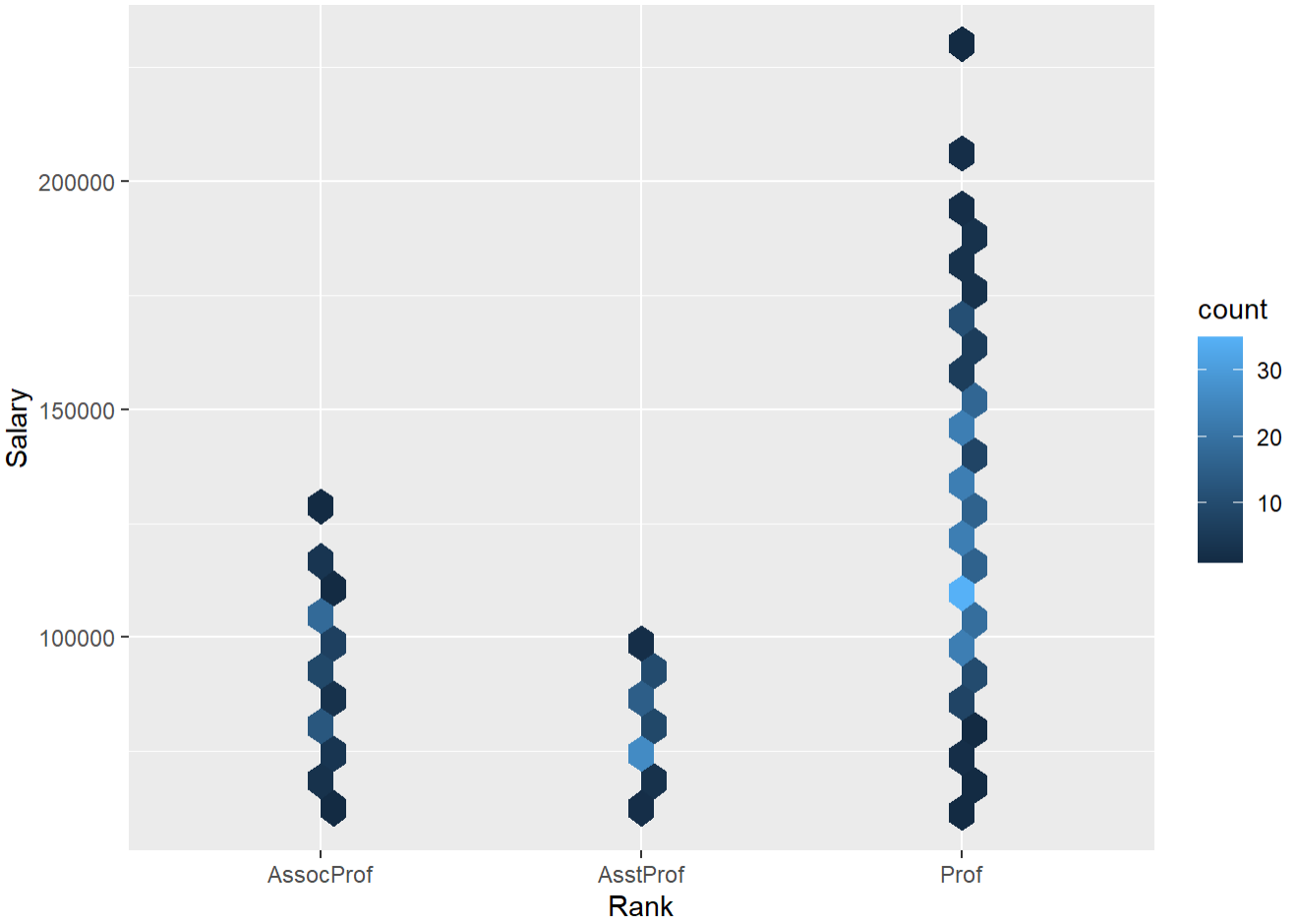
```
# Plotting Salary attribute by using histogram plot
qplot(Salary, data = Data, bins = 10)
```



The salary has appeared that have right-skewed distributions.

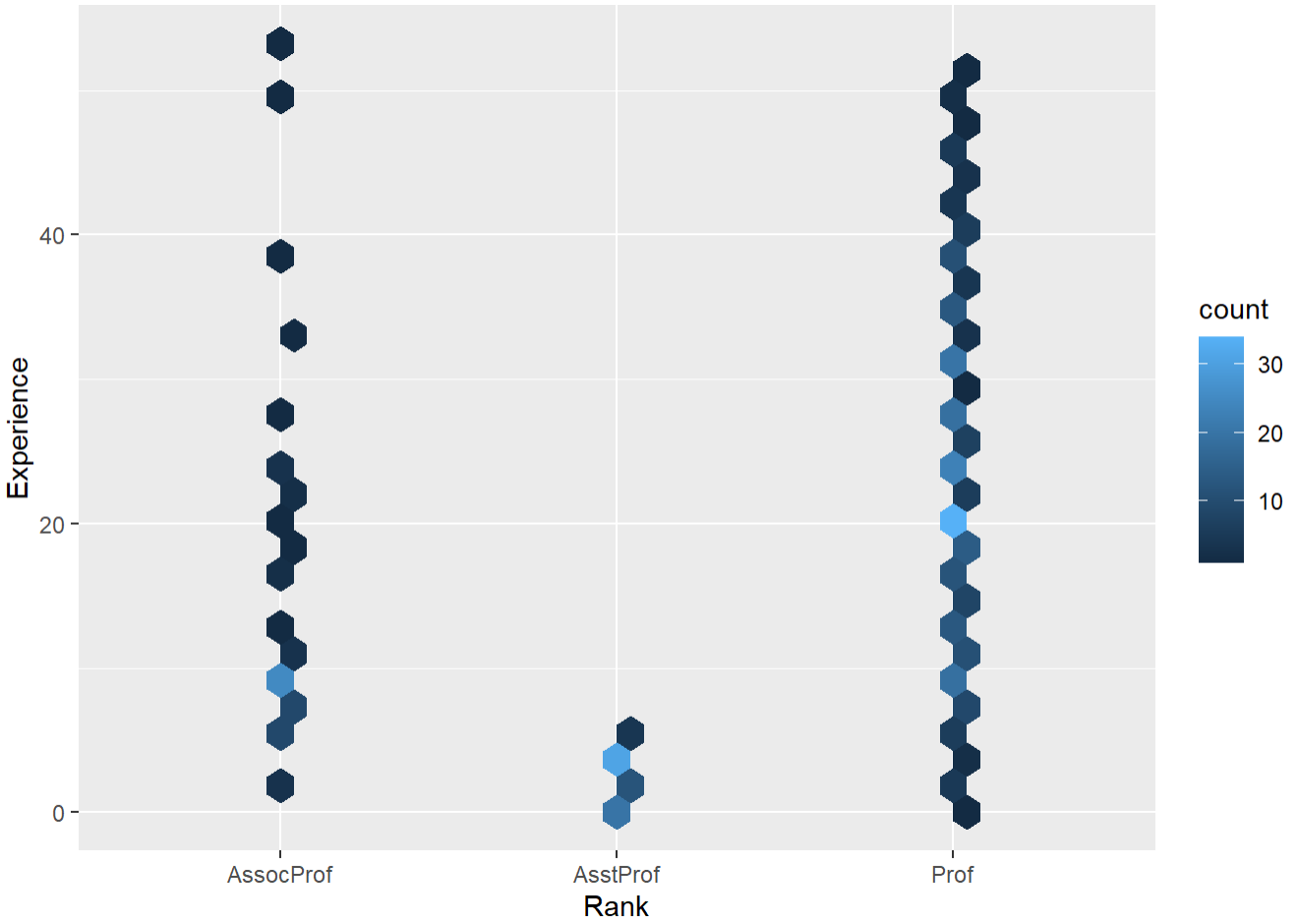
#5. Provide appropriate plots that visualize relations or associations between each pair of variables

```
# Plotting associations between Rank and Salary attributes by using a Hexagonal heatmap plot
ggplot(Data, aes(x = Rank, y = Salary)) +
  geom_hex(bins = 25)
```



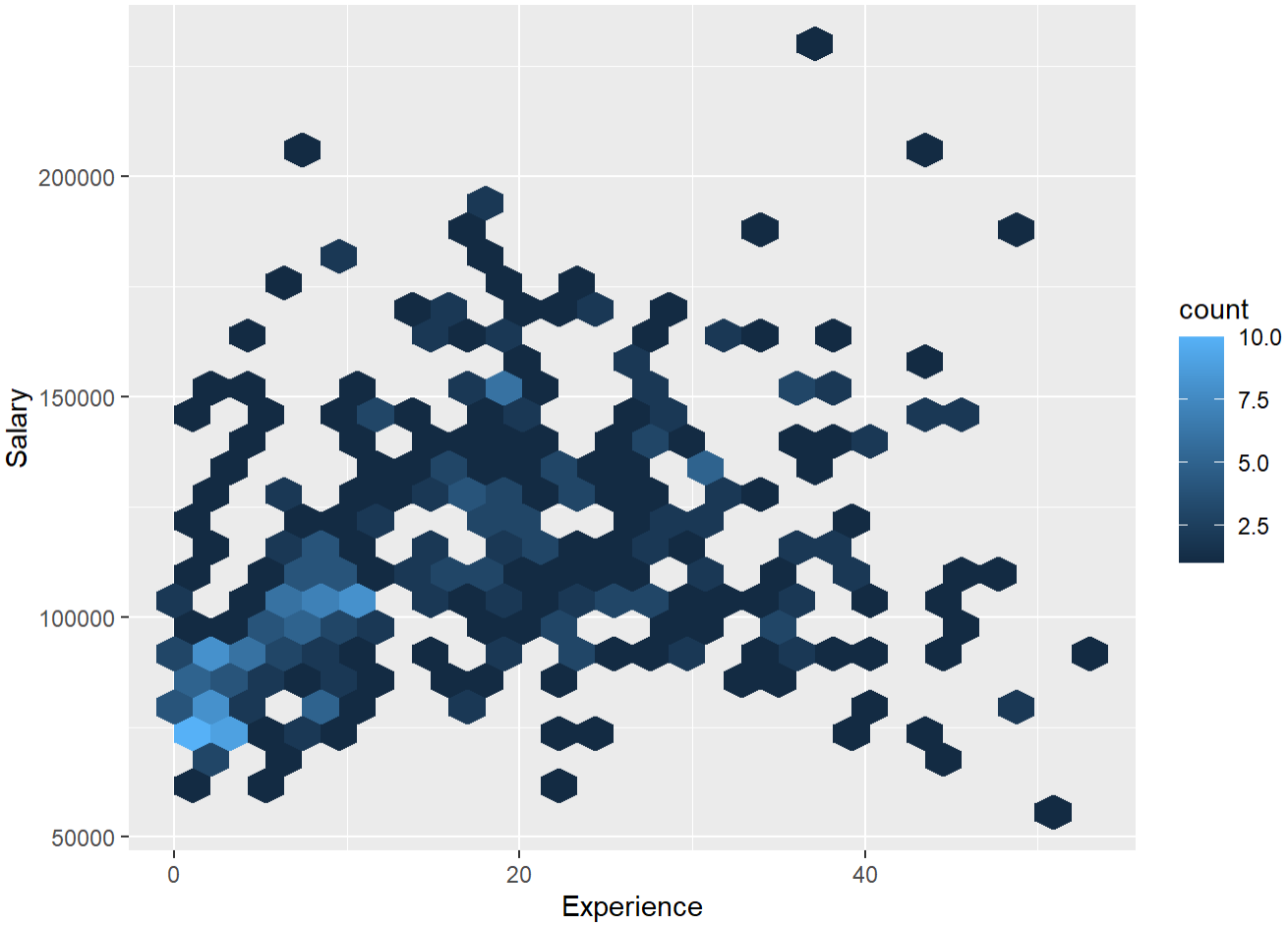
It turns out that the Professor has the highest salary and Associate Professor has less than the Professor amount of salary and the Assistant professor has the lowest salary.

```
# Plotting associations between Rank and Experience attributes by using a Hexagonal heatmap plot
ggplot(Data, aes(x = Rank , y =Experience )) +
  geom_hex(bins = 25)
```



As it turns out, the assistant professor has the shortest experience time, followed by the associate professor and the professor with the longest experience.

```
# Plotting associations between Experience and Salary attributes by using a Hexagonal heatmap plot
ggplot(Data, aes(x = Experience, y = Salary)) +
  geom_hex(bins = 25)
```



The Hexagonal heatmap graphic indicates that the majority of the faculty college members are paid almost between $50,000 and $150,000.

```
#Plotting associations between the whole three attributes Experience, Salary, and Rank attributes by using a scatter plot.
ggplot(Data,aes(x=Experience,y=Salary,col=Rank))+geom_point()
```

If we examine the incomes of the three rank category, we discover that :

1- The assistant professor position is centered on a certain number of experiences, with the salary being broadly rounded from those with no experience to those who have at least five years of experience and a salary of up to 100,000.

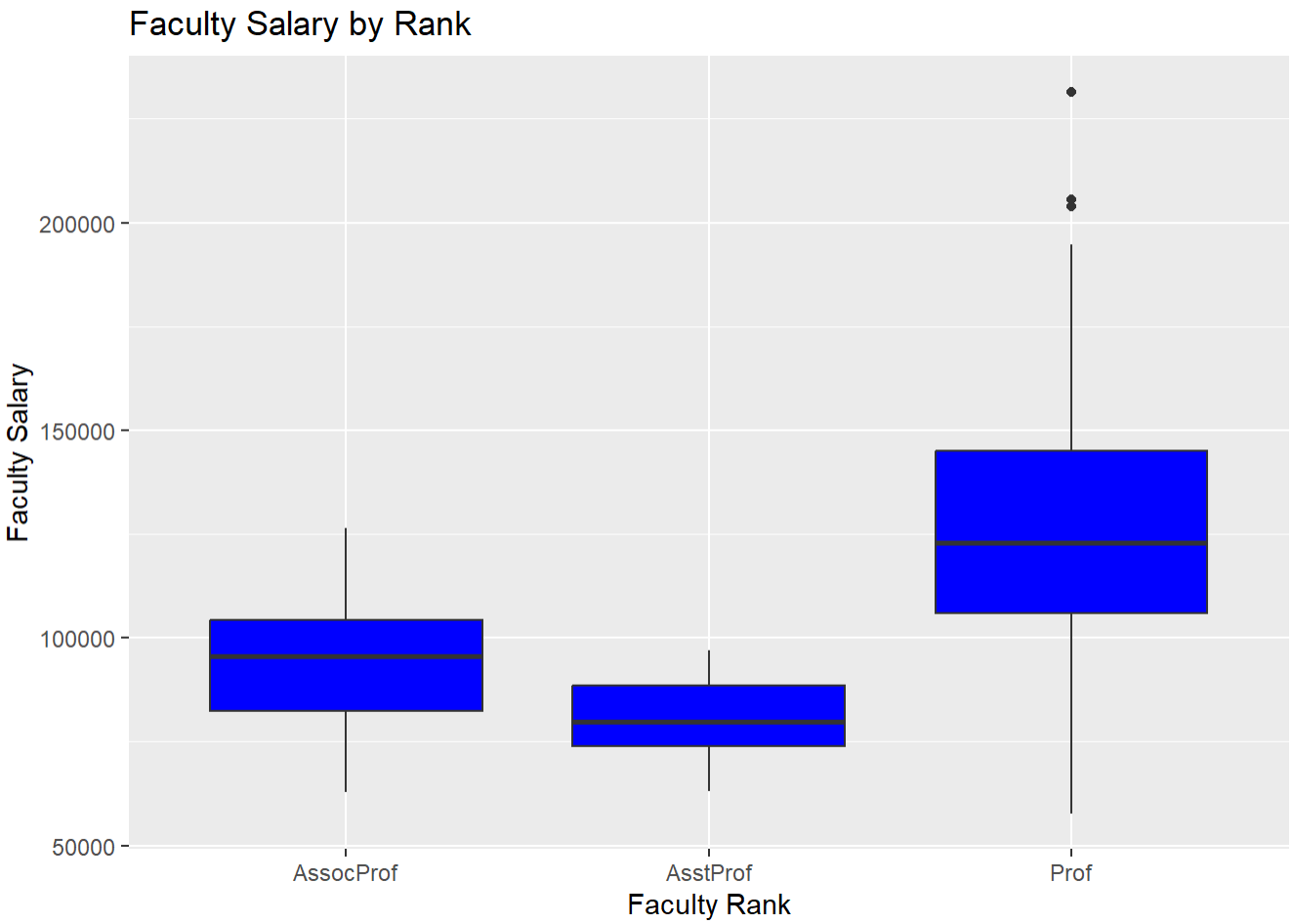
2- The associate Professor has a variety of experiences, but most of them are based on having fewer than 20 years of experience, and they earn around 103,000 each.

3- The professors with the longest tenures and most experience earn approximately between 100,000 and 153,000.

#6. Are there any discriminations or wage gaps that are not justified? a. Does “faculty-rank” affect “faculty-salary”? Justify

yes, faculty rank seems to influence faculty salary. professors earn higher salaries compared to associate professors and assistant professors. This observation is supported by the box plot, which illustrates the differences in salary levels among various faculty ranks.

```
plot1 <- ggplot(Data, aes(x = Rank, y = Salary)) +
         geom_boxplot(fill = "blue") +
         xlab("Faculty Rank") +
         ylab("Faculty Salary") +
         ggtitle("Faculty Salary by Rank")
plot1
```

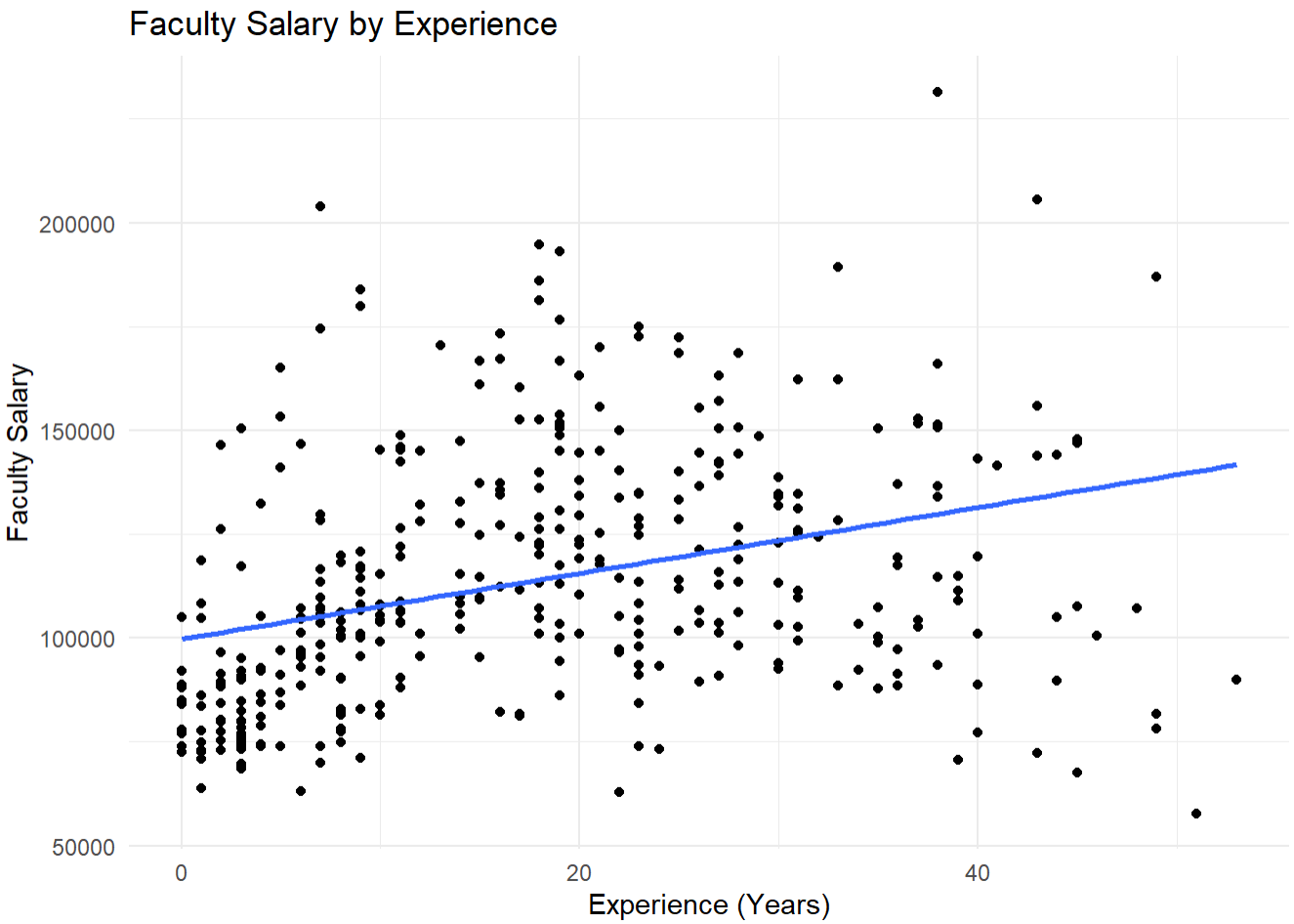


b. Does “Faculty-Experience” impact “faculty salary”? Justify

yes, faculty experience seems to have an impact on faculty salary. As experience increases, faculty members generally tend to earn higher salaries. This relationship can be effectively visualized using a scatter plot and a linear regression line, which helps to demonstrate the positive correlation between experience and salary.

```
plot2 <-ggplot(Data, aes(x = Experience, y = Salary)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Faculty Salary by Experience", x = "Experience (Years)", y = "Faculty Salary") +
  theme_minimal()
plot2
```

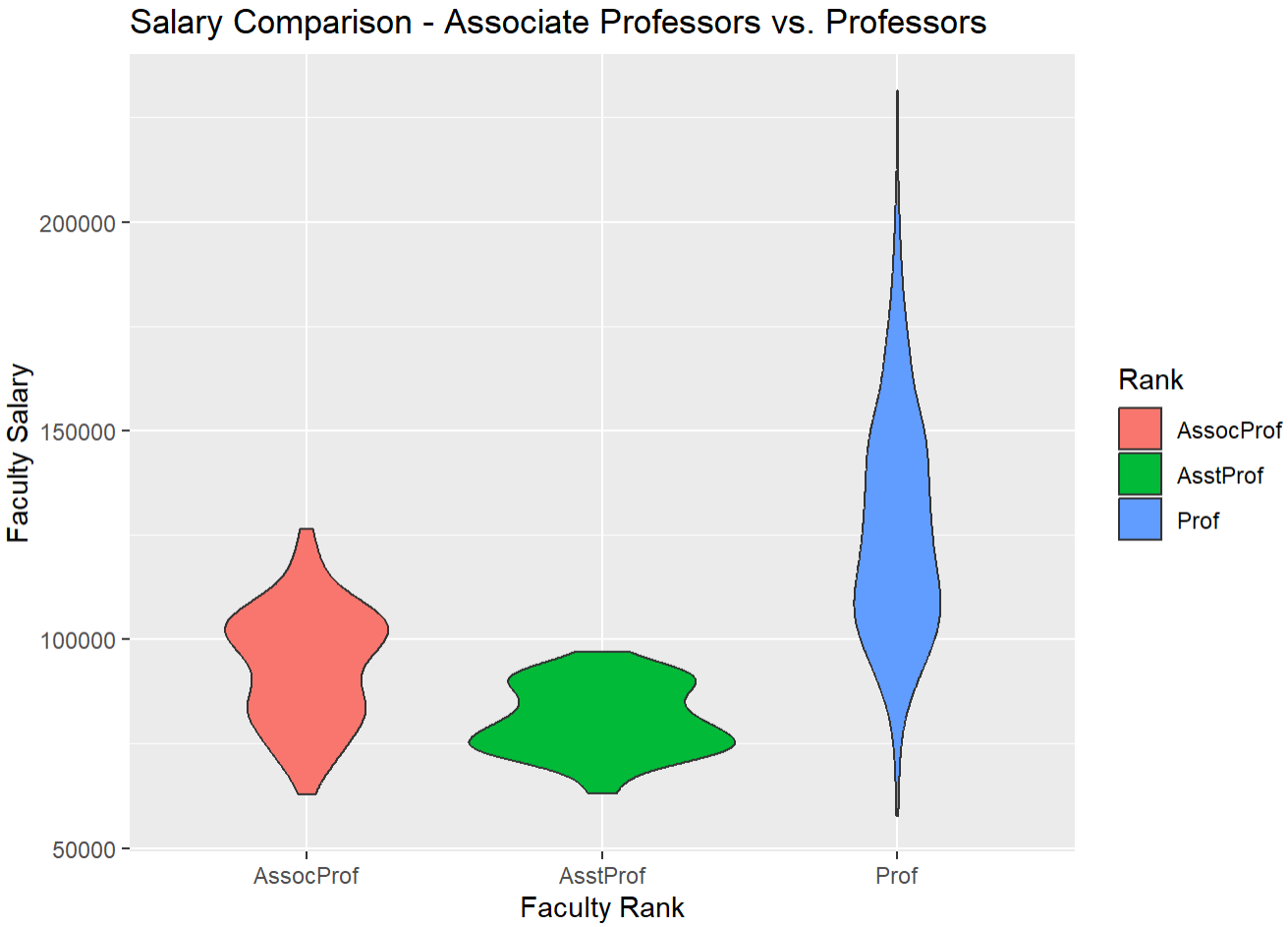
```
## `geom_smooth()` using formula = 'y ~ x'
```



c. Is the difference between associate professors' salaries and professors' salaries significant?

Yes, there is a difference between the salaries of associate professors and professors. This difference is significant, as demonstrated by the violin plot, which shows that professors generally earn higher salaries than associate professors.

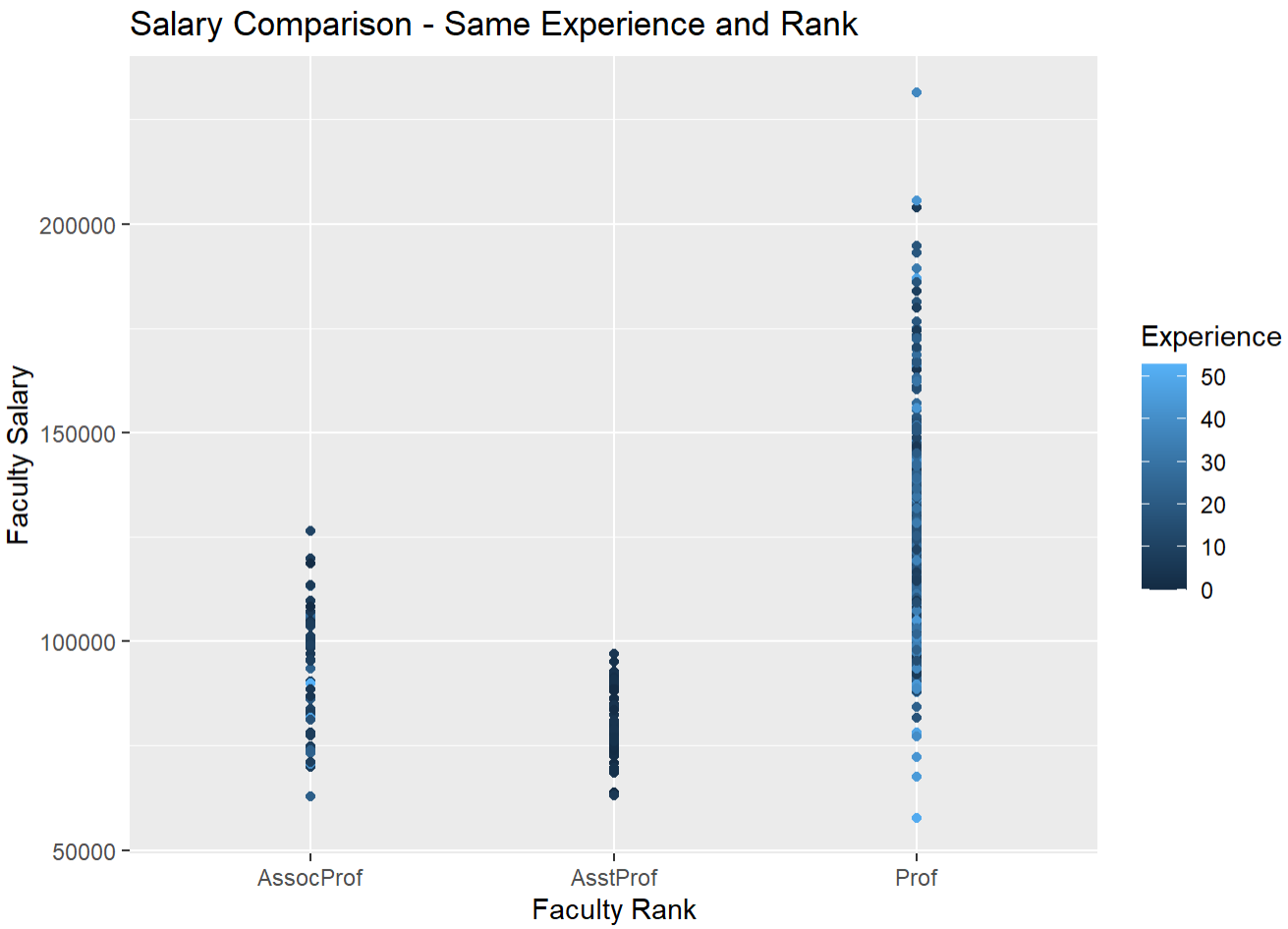
```
plot3 <- ggplot(Data, aes(x = Rank, y = Salary, fill = Rank)) +
         geom_violin() +
         xlab("Faculty Rank") +
         ylab("Faculty Salary") +
         ggtitle("Salary Comparison - Associate Professors vs. Professors")
plot3
```



d. Are there any wage gaps for employees with the same experience and rank? Yes, wage gaps exist among employees with the same experience and rank. The scatter plot clearly illustrates varying salary levels for individuals with the same experience and rank, which indicates the presence of wage gaps. This observation could imply potential unjustified disparities or discrimination in salary distribution.

```
plot4 <- ggplot(Data, aes(x =Rank, y = Salary, color = Experience)) +
         geom_point() +
         xlab("Faculty Rank") +
         ylab("Faculty Salary") +
         ggtitle("Salary Comparison - Same Experience and Rank")

plot4
```



#7. If inequities exist, what are the suggested adjustment strategies that solve or improve the situation?

Inequities or wage gaps are identified within the dataset, several suggested adjustment strategies can help address or improve the situation:

1. Conduct a thorough analysis: Further investigate the factors contributing to the wage gaps or inequities, such as considering additional variables like gender, ethnicity, or department affiliation. This analysis can provide more insights into the underlying causes and help guide appropriate adjustment strategies.
2. Implement pay equity policies: Establish and enforce policies that ensure fair and equal compensation for employees with similar qualifications, experience, and responsibilities. These policies should be designed to address any wage gaps or discriminatory practices and promote pay equity within the organization.
3. Review and revise salary structures: Evaluate the current salary structures and consider adjustments that align with industry standards and best practices. This may involve revising salary scales, implementing performance-based pay systems, or addressing any discrepancies in pay levels based on rank or experience.
4. Provide professional development opportunities: Offer training, mentorship, and career advancement programs to employees to enhance their skills, knowledge, and qualifications. This can help create a more equitable environment by enabling employees to progress in their careers and increase their earning potential based on merit and achievement.
5. Foster transparency and communication: Ensure transparency in the compensation process by clearly communicating salary structures, criteria for promotions, and other relevant information to all employees. Encourage open dialogue and feedback mechanisms to address any concerns or perceptions of inequity and provide opportunities for employees to voice their opinions.
6. Regularly monitor and evaluate: Continuously monitor salary data, analyze trends, and conduct periodic reviews to identify any emerging inequities or wage gaps. Regular evaluations will help assess the effectiveness of adjustment strategies and allow for timely interventions when necessary.

It is important to note that specific adjustment strategies will depend on the organization’s policies, legal requirements, and the unique characteristics of the workforce. Consulting with HR professionals, legal experts, or relevant stakeholders can provide valuable guidance in developing and implementing appropriate strategies to address wage gaps and promote equity. By implementing these adjustment strategies and continuously striving for fair and equitable compensation practices, organizations can work towards eliminating inequities and fostering a more inclusive and equitable work environment.