INFS 324 Project: Exploratory analysis using R R Malatjie student number: 37820982

2024-10-08

This Document Demonstrates the Exploratory Data Analysis of the turnover data provided for the INFS324 Project using R. Below are the observations and results of each analysis.First, the necessary packages had to be installed and loaded.

# 1. Import data turnover excel file

turnoverdata <-read\_excel("Turnover.xlsx")

Next, we proceed to inspect the number of rows and columns

## inspection of no. of rows and columns

#Code to inspect rows and columns  
rowsandcolumns<- dim(turnoverdata)  
print(rowsandcolumns)

## [1] 1129 16

cont\_variables<-str(turnoverdata)

## tibble [1,129 × 16] (S3: tbl\_df/tbl/data.frame)  
## $ experience : num [1:1129] 7.03 22.97 15.93 15.93 8.41 ...  
## $ event : num [1:1129] 1 1 1 1 1 1 1 1 1 1 ...  
## $ gender : chr [1:1129] "m" "m" "f" "f" ...  
## $ age : num [1:1129] 35 33 35 35 32 42 42 28 29 30 ...  
## $ industry : chr [1:1129] "Banks" "Banks" "PowerGeneration" "PowerGeneration" ...  
## $ profession : chr [1:1129] "HR" "HR" "HR" "HR" ...  
## $ traffic : chr [1:1129] "rabrecNErab" "empjs" "rabrecNErab" "rabrecNErab" ...  
## $ coach : chr [1:1129] "no" "no" "no" "no" ...  
## $ head\_gender : chr [1:1129] "f" "m" "m" "m" ...  
## $ greywage : chr [1:1129] "white" "white" "white" "white" ...  
## $ way : chr [1:1129] "bus" "bus" "bus" "bus" ...  
## $ extraversion: num [1:1129] 6.2 6.2 6.2 5.4 3 6.2 6.2 3.8 8.6 5.4 ...  
## $ independ : num [1:1129] 4.1 4.1 6.2 7.6 4.1 6.2 6.2 5.5 6.9 5.5 ...  
## $ selfcontrol : num [1:1129] 5.7 5.7 2.6 4.9 8 4.1 4.1 8 2.6 3.3 ...  
## $ anxiety : num [1:1129] 7.1 7.1 4.8 2.5 7.1 5.6 5.6 4 4 7.9 ...  
## $ novator : num [1:1129] 8.3 8.3 8.3 6.7 3.7 6.7 6.7 4.4 7.5 8.3 ...

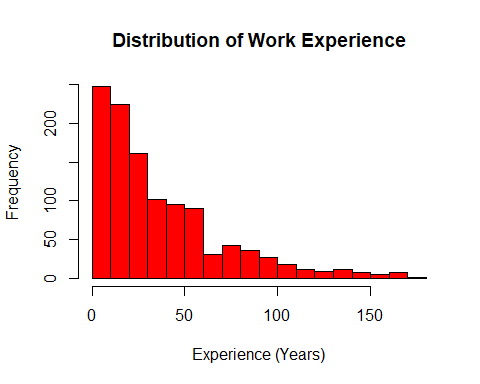
From this information, the variables found are numbers(num) and characters(chr), along with the column names and sample of the data variables.

There are 1129 rows, and 16 columns

# 2.Histogram showing distribution of Employee’s work experience

To investigate the distribution of employees’ work experience, a histogram has to be created to display the distribution for proper analysis.

#histogram creation  
hist(turnoverdata$experience,breaks = 15,   
 col = "red", # Color for the bars according the question  
 border = "black", # Border color  
 main = "Distribution of Work Experience", # Title  
 xlab = "Experience (Years)", # X-axis label  
 ylab = "Frequency") # Y-axis label



# 3.Descriptive stats for Average levels of extraversion and anxiety

Here, we calculate the average (mean) and the middle value (median) for two personality traits: extraversion and anxiety. This helps us understand the avarege levels and general tendencies of employees regarding these traits.

## Calculate mean and median for extraversion

extrav\_mean<-mean(turnoverdata$extraversion, na.rm= TRUE)  
extrav\_median<-median(turnoverdata$extraversion, na.rm= TRUE)  
  
#calculate mean and median for anxiety  
anxiety\_mean<-mean(turnoverdata$anxiety, na.rm= TRUE)  
anxiety\_median<-median(turnoverdata$anxiety, na.rm= TRUE)  
  
print(extrav\_mean)

## [1] 5.592383

print(extrav\_median)

## [1] 5.4

print(anxiety\_mean)

## [1] 5.665633

print(anxiety\_median)

## [1] 5.6

From the Results of the code:

## Extraversion:

Mean= 5.592383

Findings: On average, employees show a moderately high level of extraversion.

Median: 5.4

Findings: Slightly lower than the mean, implying a slight right skew in the extraversion scores because few employees may have a higher score.

## Anxiety:

Mean= 5.665633

Findings: On average, the employees have reported a moderate level of anxiety.

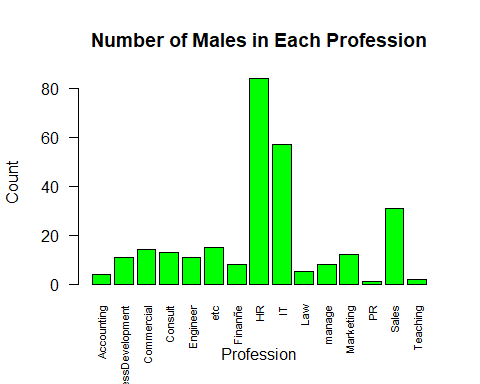
Median= 5.6

Findings: Being closely similar to the mean suggests a symmetric distribution of anxiety scores among the employees.

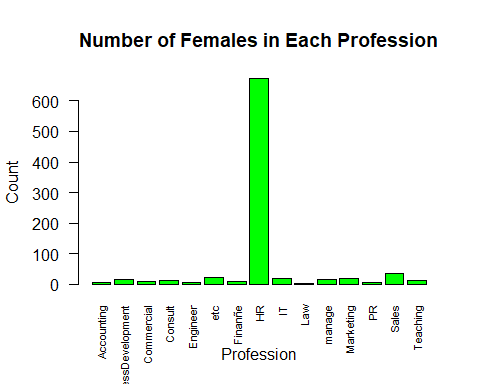
# 4.Gender Distribution across professions

Here, we calculate the average (mean) and the middle value (median) for two personality traits: extraversion and anxiety. This helps us understand the general tendencies of employees regarding these traits and can indicate how they might affect their work behavior.

genderProfession\_table <- table(turnoverdata$profession, turnoverdata$gender)  
  
barplot(genderProfession\_table[, "m"],# Select counts for males  
 col = "green", # Color for male bars  
 main = "Number of Males in Each Profession",   
 xlab = "Profession",   
 ylab = "Count",   
 names.arg = rownames(genderProfession\_table), # Profession names  
 las = 2, # Rotate x-axis labels to be vertical  
 cex.names = 0.7, # Font size for profession names  
 border = "black")



# Bar chart for females  
barplot(genderProfession\_table[,"f"], # Select counts for females  
 col = "green", # Color for female bars  
 main = "Number of Females in Each Profession",   
 xlab = "Profession",   
 ylab = "Count",   
 names.arg = rownames(genderProfession\_table), # Profession names  
 las = 2.5,# Rotate x-axis labels to be vertical  
 cex.names = 0.7, # Font size for profession names  
 border = "black")



# 5.Cross tabulation: Relationship between industry and experience

Here, We categorize experience into groups (like 0-5 years, 6-10 years, etc.) and create a cross-tabulation to observe how experience levels are distributed across different industries. This analysis reveals trends that might exist in hiring or promotion practices.

##Investigate relationships(experience&industry)

# Finding max value to evaluate categorizing  
  
library(knitr)  
maxexperience <- max(turnoverdata$experience, na.rm = TRUE)  
  
# Categorizing the experience into groups  
turnoverdata$experiencegroup <- cut(turnoverdata$experience,  
 breaks = c(0, 5, 10, 20, 30, 50, Inf),  
 labels = c("0-5 years", "6-10 years", "11-20 years",   
 "21-30 years", "31-50 years", "50+ years"),  
 right = FALSE)  
  
# Create a contingency table for industry and experience groups  
industryAndexperience\_table <- table(turnoverdata$industry, turnoverdata$experiencegroup)  
  
kable(industryAndexperience\_table, format = "markdown",   
 caption = "Contingency Table of Industry and Experience Groups")

Contingency Table of Industry and Experience Groups

|  | 0-5 years | 6-10 years | 11-20 years | 21-30 years | 31-50 years | 50+ years |
| --- | --- | --- | --- | --- | --- | --- |
| Agriculture | 2 | 1 | 5 | 5 | 1 | 1 |
| Banks | 20 | 15 | 23 | 14 | 16 | 26 |
| Building | 5 | 2 | 10 | 10 | 3 | 11 |
| Consult | 6 | 11 | 14 | 17 | 15 | 11 |
| etc | 9 | 7 | 22 | 14 | 18 | 24 |
| HoReCa | 2 | 1 | 2 | 1 | 1 | 4 |
| IT | 18 | 20 | 16 | 19 | 23 | 26 |
| manufacture | 11 | 17 | 23 | 26 | 29 | 39 |
| Mining | 1 | 2 | 6 | 1 | 6 | 8 |
| Pharma | 3 | 0 | 6 | 2 | 4 | 5 |
| PowerGeneration | 2 | 3 | 10 | 1 | 12 | 10 |
| RealEstate | 0 | 2 | 2 | 2 | 1 | 6 |
| Retail | 25 | 33 | 58 | 34 | 51 | 88 |
| State | 4 | 11 | 9 | 6 | 9 | 16 |
| Telecom | 5 | 2 | 10 | 1 | 1 | 17 |
| transport | 2 | 6 | 9 | 8 | 7 | 6 |

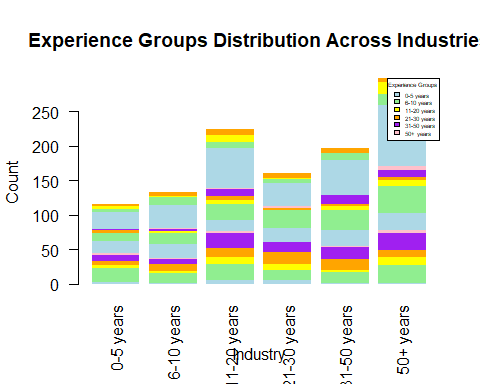
### Summary:

The table shows how experience levels are distributed across 14 different industries, highlighting the number of employees in each experience category for each industry.

Retail and manufacturing show to have a higher representation of experienced employees, whereas telecom have a more youthful group of employees

## Bar Plot visualizing the cross-tabulation results:

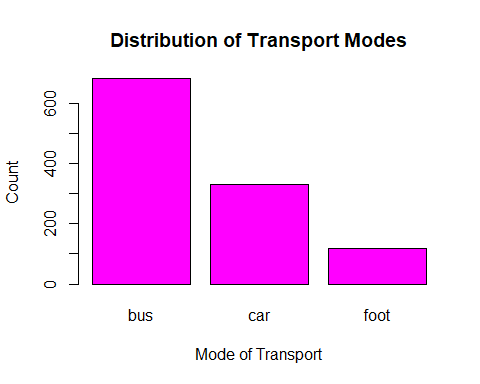
#Bar plot for visualiztion of table  
barplot(industryAndexperience\_table,  
 main = "Experience Groups Distribution Across Industries",  
 xlab = "Industry",  
 ylab = "Count",  
 col = c("lightblue", "lightgreen", "yellow", "orange", "purple", "pink"), # Colors for different experience groups  
 beside = FALSE, # Stacked bar graph look  
 las = 2, # Rotate x-axis labels  
 border = NA)  
legend("topright",  
 legend = colnames(industryAndexperience\_table), # Experience group labels  
 fill = c("lightblue", "lightgreen", "yellow", "orange", "purple", "pink"),  
 title = "Experience Groups",  
 cex = 0.4)



# 6. Distribution of Employees based on mode of Transport

This visualization helps understand employees’ commuting preferences and can inform company policies on transport benefits or initiatives.

wayTable <- table(turnoverdata$way)  
  
# Bar plot for transport modes  
barplot(wayTable,  
 col = "magenta", # magenta  
 main = "Distribution of Transport Modes",  
 xlab = "Mode of Transport",  
 ylab = "Count")



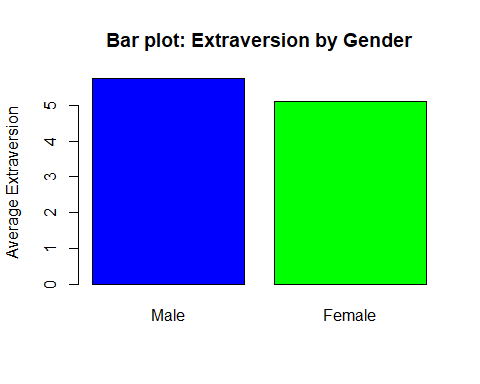
### Summary

Most used mode of transport amongst the employees is the bus, followed by those who use cars and those walking being a small minority of the employees.

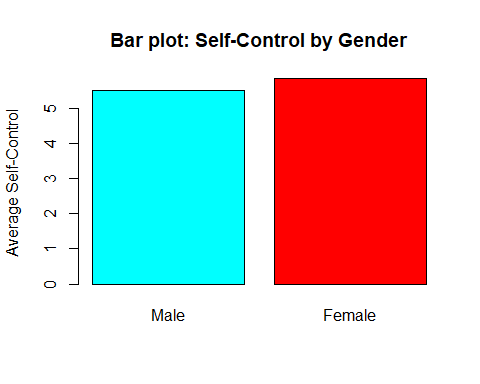
# 7.Examine personality traits(extraversion and Self-Control) by gender

#Descriptive Statistics by gender We calculate the average for each trait and create bar plots to visually compare the two genders. This comparison provides insights into how personality traits may influence work dynamics based on gender.

# Calculate the average extraversion by gender  
avg\_extraversion <- tapply(turnoverdata$extraversion, turnoverdata$gender, mean, na.rm = TRUE)  
  
# Create a bar plot for average extraversion by gender  
barplot(avg\_extraversion,  
 main = "Bar plot: Extraversion by Gender",  
 ylab = "Average Extraversion",  
 col = c("blue", "green"),  
 names.arg = c("Male", "Female"), # Change labels if needed  
 beside = TRUE)



# Calculate the average self-control by gender  
avg\_selfcontrol <- tapply(turnoverdata$selfcontrol, turnoverdata$gender, mean, na.rm = TRUE)  
  
# Create a bar plot for average self-control by gender  
barplot(avg\_selfcontrol,  
 main = "Bar plot: Self-Control by Gender",  
 ylab = "Average Self-Control",  
 col = c("cyan", "red"),  
 names.arg = c("Male", "Female"), # Change labels if needed  
 beside = TRUE)

 ## Findings Females show a lower avarage of extraversion while showing a higher average in self-control, while the males show a higher average in extraversion , but showing to have a lower average in self-control.

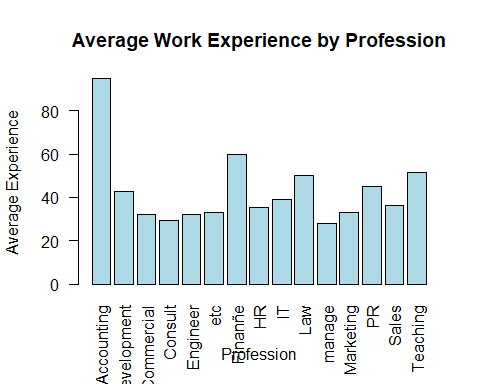
# 8. Finding the highest average work experience by profession

This analysis helps identify which roles may require more experience and can assist in workforce planning and development.

avgExperiencebyProfession <-   
 aggregate(experience ~ profession, data = turnoverdata, FUN = mean)  
  
print(avgExperiencebyProfession) #to show the tabular results

## profession experience  
## 1 Accounting 94.75483  
## 2 BusinessDevelopment 42.93437  
## 3 Commercial 32.18713  
## 4 Consult 29.44657  
## 5 Engineer 32.08323  
## 6 etc 33.00960  
## 7 Finanñe 59.74441  
## 8 HR 35.53324  
## 9 IT 39.15423  
## 10 Law 50.05104  
## 11 manage 28.08587  
## 12 Marketing 32.91992  
## 13 PR 44.95003  
## 14 Sales 36.05700  
## 15 Teaching 51.38946

## Bar chart for average experience by profession  
# code to calculate the average experience by profession  
avgExperiencebyProfession <- tapply(turnoverdata$experience, turnoverdata$profession, mean, na.rm = TRUE)  
  
# bar plot code for average experience by profession  
barplot(avgExperiencebyProfession,  
 main = "Average Work Experience by Profession",  
 xlab = "Profession",  
 ylab = "Average Experience",  
 col = "lightblue",  
 las = 2)

 ### Results By observing the bar chart, it shows Highest average work experience to be in the Accounting profession.

# 9.Proportion of the employees getting coaching by gender

We investigate how many employees receive coaching and compare this between genders using a cross-tabulation, then visualizing it using a bar plot

## Cross-tabulation of employees who receive coaching and comparing between genders   
  
coachingGender <- table(turnoverdata$gender, turnoverdata$coach)  
  
print(coachingGender)

##   
## my head no yes  
## f 248 498 107  
## m 66 185 25

# Proportion of employees receiving coaching by gender, in percentage format  
coachingGenderProp <- prop.table(coachingGender, margin = 1)\*100  
  
# Print the cross-tabulation table  
library(knitr)  
kable(coachingGenderProp, format = "pipe",   
 caption = " Employees Coaching count by Gender(%)")

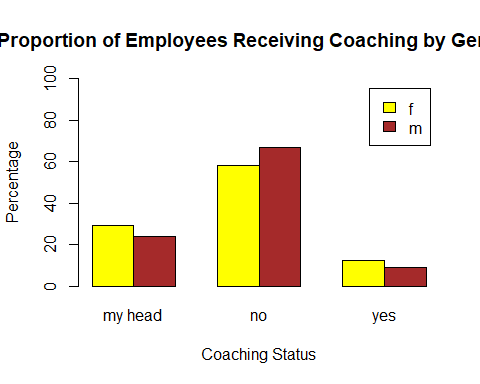
Employees Coaching count by Gender(%)

|  | my head | no | yes |
| --- | --- | --- | --- |
| f | 29.07386 | 58.38218 | 12.543962 |
| m | 23.91304 | 67.02899 | 9.057971 |

###Findings: 58.38% of female employees and 67.03% of male employees receive coaching, indicating higher participation among males. While 29.07% of females and 23.91% of males do not receive coaching, females show a slightly higher rate of non-participation. The smaller percentages in the “uncertain or other status” category (12.54% for females and 9.06% for males) suggest that most employees clearly identify as either receiving or not receiving coaching.

# Create a bar plot to visualize the proportions of coaching by gender

barplot(coachingGenderProp,  
 main = "Proportion of Employees Receiving Coaching by Gender",  
 xlab = "Coaching Status",  
 ylab = "Percentage",  
 col = c("yellow", "brown"),  
 legend = rownames(coachingGenderProp), #the legend for gender  
 beside = TRUE, # to Separate bars for each gender  
 ylim = c(0, 100))



By implementing R functions and visualizations, addressing key research questions is successful in conducting an exporatory data analysis of the Turnover Dataset, shedding light on the employee’s characteristics and trends.

End