

Lab5.R

anjaney

2024-11-21

```
library(ggplot2)
library(plyr)
library(tidyr)

# Read Employee Attrition Dataset
data <- read.csv("WA_Fn-UseC_-HR-Employee-Attrition.csv")

# Single-variable visualizations: Using 'Age' column
# 1. Bar plot
p1 <- ggplot(data, aes(x = Age)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Bar Plot of Age", x = "Age", y = "Count")

# 2. Pie chart
age_counts <- data.frame(table(data$Age))
colnames(age_counts) <- c("age", "freq")
age_counts$percentage <- (age_counts$freq / sum(age_counts$freq)) * 100
p2 <- ggplot(age_counts, aes(x = "", y = percentage, fill = as.factor(age))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  labs(title = "Pie Chart of Age", fill = "Age")

# 3. Box plot
p3 <- ggplot(data, aes(x = "", y = Age)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Box Plot of Age", y = "Age", x = "")

# 4. Density plot
p4 <- ggplot(data, aes(x = Age)) +
  geom_density(fill = "purple", alpha = 0.5) +
  labs(title = "Density Plot of Age", x = "Age", y = "Density")

# 5. Histogram
p5 <- ggplot(data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "orange", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Age", x = "Age", y = "Frequency")

# Two-variable visualizations:
# 1. Bar plot (Count of Attrition per Department)
p6 <- ggplot(data, aes(x = Department, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Bar Plot of Department by Attrition", x = "Department", y = "Count") +
```

```

    scale_fill_manual(values = c("steelblue", "lightgreen"))

# 2. Scatter plot (Age vs. Monthly Income)
p7 <- ggplot(data, aes(x = Age, y = MonthlyIncome, color = Attrition)) +
  geom_point(alpha = 0.6) +
  labs(title = "Scatter Plot of Age vs Monthly Income", x = "Age", y = "Monthly Income") +
  scale_color_manual(values = c("steelblue", "lightgreen"))

# 3. Violin plot (Age distribution by Job Level)
p8 <- ggplot(data, aes(x = factor(JobLevel), y = Age, fill = factor(JobLevel))) +
  geom_violin(trim = TRUE) +
  labs(title = "Violin Plot of Age Distribution by Job Level", x = "Job Level", y = "Age") +
  scale_fill_brewer(palette = "Set3")

# 4. Box plot (Monthly Income by Department)
p9 <- ggplot(data, aes(x = Department, y = MonthlyIncome, fill = Department)) +
  geom_boxplot() +
  labs(title = "Box Plot of Monthly Income by Department", x = "Department", y = "Monthly Income") +
  scale_fill_brewer(palette = "Set2")

# Multi-variable visualizations:
# Selecting numerical columns
numerical_data <- data[c("Age", "MonthlyIncome", "TotalWorkingYears", "YearsAtCompany")]

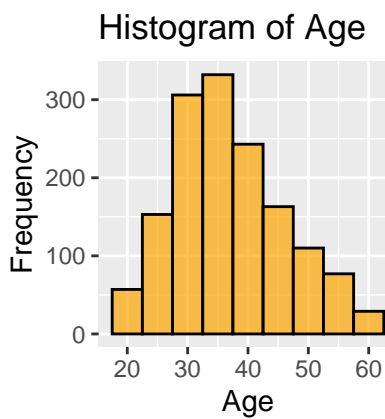
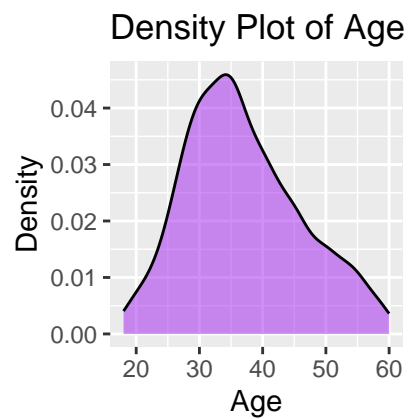
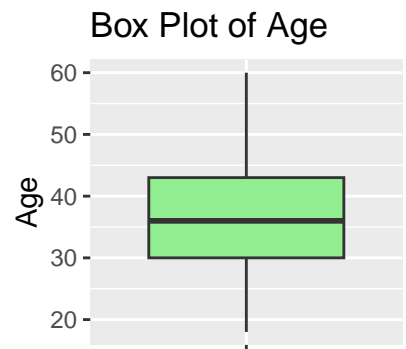
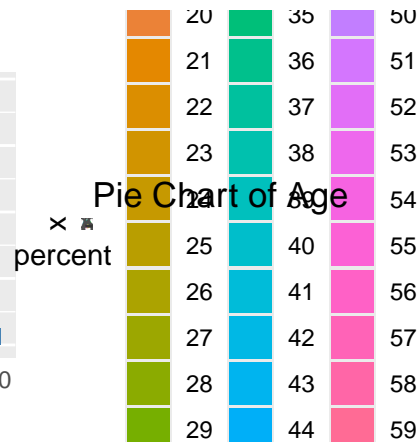
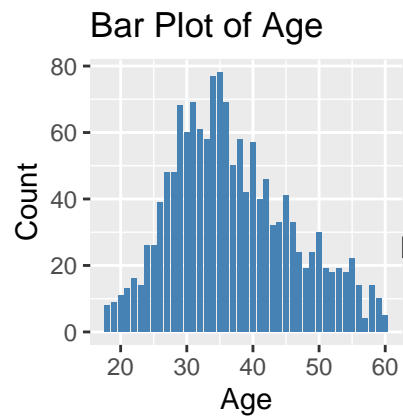
# Reshape data to long format
long_data <- ldply(names(numerical_data), function(var) {
  data.frame(Age = numerical_data$Age,
             value = numerical_data[[var]],
             variable = var)
})

# 1. Bar Plot (Job and Attrition)
p10 <- ggplot(data, aes(x = JobRole, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Bar Plot of Job Role by Attrition", x = "Job Role", y = "Count") +
  scale_fill_manual(values = c("steelblue", "lightgreen")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# 2. Pair Plot (Numerical Variables)
p11 <- ggplot(long_data, aes(x = Age, y = value, color = variable)) +
  geom_point(alpha = 0.6) +
  facet_wrap(~variable, scales = "free") +
  labs(title = "Pair Plot of Numerical Variables", x = "Age", y = "Value") +
  scale_color_brewer(palette = "Set1")

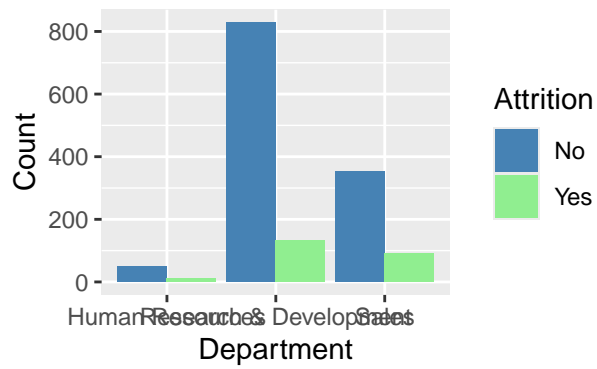
# Arrange and display plots
library(gridExtra)
grid.arrange(p1, p2, p3, p4, p5, ncol = 3)

```

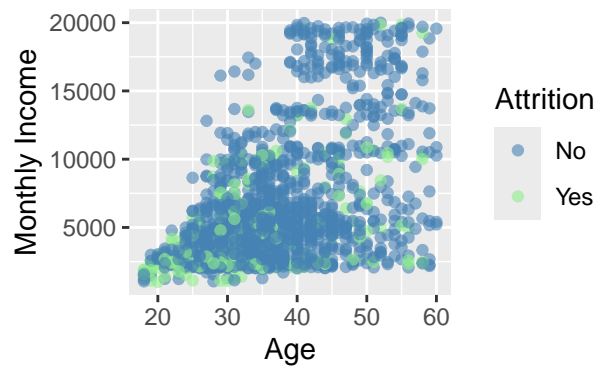


```
grid.arrange(p6, p7, p8, p9, ncol = 2)
```

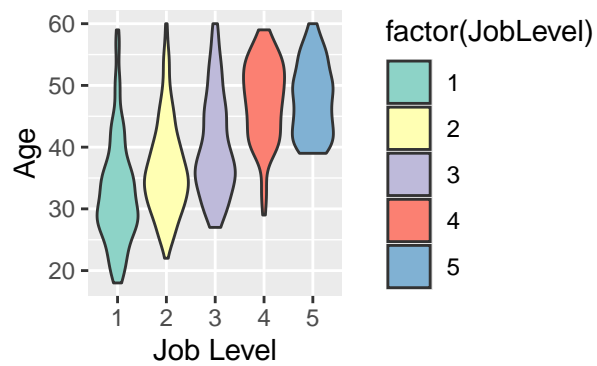
Bar Plot of Department by Attrition



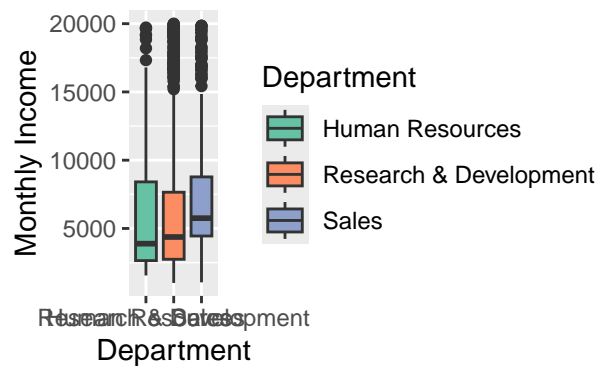
Scatter Plot of Age vs Monthly Income



Violin Plot of Age Distribution by Job Level



Box Plot of Monthly Income by Department



```
grid.arrange(p10, p11, ncol = 2)
```



```
# Metadata and Summary
print(summary(data))
```

```
##      Age      Attrition      BusinessTravel      DailyRate
##  Min.   :18.00  Length:1470  Length:1470      Min.    : 102.0
##  1st Qu.:30.00  Class :character  Class :character  1st Qu.: 465.0
##  Median :36.00  Mode  :character  Mode  :character  Median : 802.0
##  Mean   :36.92                                     Mean   : 802.5
##  3rd Qu.:43.00                                     3rd Qu.:1157.0
##  Max.    :60.00                                     Max.    :1499.0
##  Department      DistanceFromHome      Education      EducationField
##  Length:1470      Min.    : 1.000  Min.    :1.000  Length:1470
##  Class :character  1st Qu.: 2.000  1st Qu.:2.000  Class :character
##  Mode  :character  Median : 7.000  Median :3.000  Mode  :character
##                                     Mean   : 9.193  Mean   :2.913
##                                     3rd Qu.:14.000  3rd Qu.:4.000
##                                     Max.    :29.000  Max.    :5.000
##  EmployeeCount  EmployeeNumber  EnvironmentSatisfaction  Gender
##  Min.    :1      Min.    : 1.0  Min.    :1.000          Length:1470
##  1st Qu.:1      1st Qu.: 491.2  1st Qu.:2.000          Class :character
##  Median :1      Median :1020.5  Median :3.000          Mode  :character
##  Mean    :1      Mean   :1024.9  Mean   :2.722
##  3rd Qu.:1      3rd Qu.:1555.8  3rd Qu.:4.000
##  Max.    :1      Max.    :2068.0  Max.    :4.000
##  HourlyRate      JobInvolvement      JobLevel      JobRole
##  Min.    : 30.00  Min.    :1.00  Min.    :1.000  Length:1470
```

```

## 1st Qu.: 48.00    1st Qu.:2.00    1st Qu.:1.000    Class :character
## Median : 66.00    Median :3.00    Median :2.000    Mode  :character
## Mean   : 65.89    Mean   :2.73    Mean   :2.064
## 3rd Qu.: 83.75    3rd Qu.:3.00    3rd Qu.:3.000
## Max.   :100.00    Max.   :4.00    Max.   :5.000
## JobSatisfaction MaritalStatus    MonthlyIncome    MonthlyRate
## Min.    :1.000    Length:1470    Min.    : 1009    Min.    : 2094
## 1st Qu.:2.000    Class :character    1st Qu.: 2911    1st Qu.: 8047
## Median :3.000    Mode  :character    Median : 4919    Median :14236
## Mean   :2.729                    Mean   : 6503    Mean   :14313
## 3rd Qu.:4.000                    3rd Qu.: 8379    3rd Qu.:20462
## Max.   :4.000                    Max.   :19999    Max.   :26999
## NumCompaniesWorked    Over18    OverTime    PercentSalaryHike
## Min.    :0.000    Length:1470    Length:1470    Min.    :11.00
## 1st Qu.:1.000    Class :character    Class :character    1st Qu.:12.00
## Median :2.000    Mode  :character    Mode  :character    Median :14.00
## Mean   :2.693                    Mean   :15.21
## 3rd Qu.:4.000                    3rd Qu.:18.00
## Max.   :9.000                    Max.   :25.00
## PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
## Min.    :3.000    Min.    :1.000    Min.    :80    Min.    :0.0000
## 1st Qu.:3.000    1st Qu.:2.000    1st Qu.:80    1st Qu.:0.0000
## Median :3.000    Median :3.000    Median :80    Median :1.0000
## Mean   :3.154    Mean   :2.712    Mean   :80    Mean   :0.7939
## 3rd Qu.:3.000    3rd Qu.:4.000    3rd Qu.:80    3rd Qu.:1.0000
## Max.   :4.000    Max.   :4.000    Max.   :80    Max.   :3.0000
## TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
## Min.    : 0.00    Min.    :0.000    Min.    :1.000    Min.    : 0.000
## 1st Qu.: 6.00    1st Qu.:2.000    1st Qu.:2.000    1st Qu.: 3.000
## Median :10.00    Median :3.000    Median :3.000    Median : 5.000
## Mean   :11.28    Mean   :2.799    Mean   :2.761    Mean   : 7.008
## 3rd Qu.:15.00    3rd Qu.:3.000    3rd Qu.:3.000    3rd Qu.: 9.000
## Max.   :40.00    Max.   :6.000    Max.   :4.000    Max.   :40.000
## YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
## Min.    : 0.000    Min.    : 0.000    Min.    : 0.000
## 1st Qu.: 2.000    1st Qu.: 0.000    1st Qu.: 2.000
## Median : 3.000    Median : 1.000    Median : 3.000
## Mean   : 4.229    Mean   : 2.188    Mean   : 4.123
## 3rd Qu.: 7.000    3rd Qu.: 3.000    3rd Qu.: 7.000
## Max.   :18.000    Max.   :15.000    Max.   :17.000

```