Applied Multivariate Data Analysis

Cristel Kaye Billones

Formative Assessment 5

```
library(readr)
library(ggplot2)
library(dplyr)
library(tidyr)
```

```
file_path <- "C:/Users/Cipher/Desktop/AMDA/employee_attrition_train.csv"
# Load the dataset
df <- read_csv(file_path)
```

```
## Rows: 1029 Columns: 35
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (9): Attrition, BusinessTravel, Department, EducationField, Gender, Job...
## dbl (26): Age, DailyRate, DistanceFromHome, Education, EmployeeCount, Employ...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
# View the first few rows of the dataset
head(df)
```

```
## # A tibble: 6 x 35
##     Age Attrition BusinessTravel DailyRate Department DistanceFromHome Education
##   <dbl> <chr>     <chr>              <dbl> <chr>                 <dbl>     <dbl>
## 1    50 No        Travel_Rarely       1126 Research ~                1         2
## 2    36 No        Travel_Rarely        216 Research ~                6         2
## 3    21 Yes       Travel_Rarely        337 Sales                     7         1
## 4    50 No        Travel_Freque~      1246 Human Res~               NA         3
## 5    52 No        Travel_Rarely        994 Research ~                7         4
## 6    33 Yes       Travel_Rarely       1277 Research ~               15         1
## # i 28 more variables: EducationField <chr>, EmployeeCount <dbl>,
## #   EmployeeNumber <dbl>, EnvironmentSatisfaction <dbl>, Gender <chr>,
## #   HourlyRate <dbl>, JobInvolvement <dbl>, JobLevel <dbl>, JobRole <chr>,
## #   JobSatisfaction <dbl>, MaritalStatus <chr>, MonthlyIncome <dbl>,
## #   MonthlyRate <dbl>, NumCompaniesWorked <dbl>, Over18 <chr>, OverTime <chr>,
## #   PercentSalaryHike <dbl>, PerformanceRating <dbl>,
## #   RelationshipSatisfaction <dbl>, StandardHours <dbl>, ...
```

```r
colnames(df)
```

```
##  [1] "Age"                      "Attrition"
##  [3] "BusinessTravel"           "DailyRate"
##  [5] "Department"               "DistanceFromHome"
##  [7] "Education"                "EducationField"
##  [9] "EmployeeCount"            "EmployeeNumber"
## [11] "EnvironmentSatisfaction"  "Gender"
## [13] "HourlyRate"               "JobInvolvement"
## [15] "JobLevel"                 "JobRole"
## [17] "JobSatisfaction"          "MaritalStatus"
## [19] "MonthlyIncome"            "MonthlyRate"
## [21] "NumCompaniesWorked"       "Over18"
## [23] "OverTime"                 "PercentSalaryHike"
## [25] "PerformanceRating"        "RelationshipSatisfaction"
## [27] "StandardHours"            "StockOptionLevel"
## [29] "TotalWorkingYears"        "TrainingTimesLastYear"
## [31] "WorkLifeBalance"          "YearsAtCompany"
## [33] "YearsInCurrentRole"       "YearsSinceLastPromotion"
## [35] "YearsWithCurrManager"
```

```r
# Drop rows with any missing values
df_clean <- df %>% drop_na()
# Check the cleaned data
summary(df_clean)
```

```
##       Age        Attrition          BusinessTravel       DailyRate
##  Min.   :18.00   Length:775         Length:775         Min.   : 102.0
##  1st Qu.:31.00   Class :character   Class :character   1st Qu.: 431.5
##  Median :37.00   Mode  :character   Mode  :character   Median : 750.0
##  Mean   :38.05                                         Mean   : 786.4
##  3rd Qu.:44.00                                         3rd Qu.:1148.5
##  Max.   :60.00                                         Max.   :1495.0
##   Department        DistanceFromHome   Education     EducationField
##  Length:775         Min.   : 1.00    Min.   :1.000   Length:775
##  Class :character   1st Qu.: 2.00    1st Qu.:2.000   Class :character
##  Mode  :character   Median : 8.00    Median :3.000   Mode  :character
##                     Mean   : 9.68    Mean   :2.917
##                     3rd Qu.:15.00    3rd Qu.:4.000
##                     Max.   :29.00    Max.   :5.000
##  EmployeeCount EmployeeNumber   EnvironmentSatisfaction    Gender
##  Min.   :1     Min.   :   1.0   Min.   :1.00            Length:775
##  1st Qu.:1     1st Qu.: 499.5   1st Qu.:2.00            Class :character
##  Median :1     Median :1025.0   Median :3.00            Mode  :character
##  Mean   :1     Mean   :1027.1   Mean   :2.68
##  3rd Qu.:1     3rd Qu.:1554.5   3rd Qu.:4.00
##  Max.   :1     Max.   :2068.0   Max.   :4.00
##    HourlyRate     JobInvolvement     JobLevel       JobRole
##  Min.   : 30.00   Min.   :1.000   Min.   :1.000   Length:775
##  1st Qu.: 49.00   1st Qu.:2.000   1st Qu.:1.000   Class :character
##  Median : 68.00   Median :3.000   Median :2.000   Mode  :character
##  Mean   : 67.07   Mean   :2.729   Mean   :2.124
##  3rd Qu.: 85.00   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :100.00   Max.   :4.000   Max.   :5.000
##  JobSatisfaction MaritalStatus      MonthlyIncome    MonthlyRate
##  Min.   :1.000   Length:775         Min.   : 1009   Min.   : 2094
##  1st Qu.:2.000   Class :character   1st Qu.: 2908   1st Qu.: 7744
##  Median :3.000   Mode  :character   Median : 4963   Median :14115
##  Mean   :2.735                      Mean   : 6797   Mean   :14198
##  3rd Qu.:4.000                      3rd Qu.: 9302   3rd Qu.:20379
##  Max.   :4.000                      Max.   :19999   Max.   :26999
##  NumCompaniesWorked    Over18             OverTime          PercentSalaryHike
##  Min.   :0.000      Length:775         Length:775         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   :2.759                                            Mean   :15.29
##  3rd Qu.:4.000                                            3rd Qu.:18.00
##  Max.   :9.000                                            Max.   :25.00
##  PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
##  Min.   :3.00      Min.   :1.000            Min.   :80    Min.   :0.0000
##  1st Qu.:3.00      1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000
##  Median :3.00      Median :3.000            Median :80    Median :1.0000
##  Mean   :3.16      Mean   :2.679            Mean   :80    Mean   :0.8452
##  3rd Qu.:3.00      3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000
##  Max.   :4.00      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   :11.99     Mean   :2.748         Mean   :2.765   Mean   : 7.355
##  3rd Qu.:17.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.:10.000
##  Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :37.000
##  YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000     Min.   : 0.00           Min.   : 0.000
##  1st Qu.: 2.000     1st Qu.: 0.00           1st Qu.: 2.000
##  Median : 3.000     Median : 1.00           Median : 3.000
##  Mean   : 4.365     Mean   : 2.27           Mean   : 4.195
##  3rd Qu.: 7.000     3rd Qu.: 3.00           3rd Qu.: 7.000
##  Max.   :18.000     Max.   :15.00           Max.   :17.000
```

```r
# Ensure JobSatisfaction is treated as a factor
df_clean$JobSatisfaction <- as.factor(df_clean$JobSatisfaction)

# Perform MANOVA, handling missing data with na.omit
manova_test <- manova(cbind(Age, DailyRate, MonthlyIncome) ~ JobSatisfaction, data = df_clean, na.actio

# Wilks' Lambda for overall significance
manova_wilks <- summary(manova_test, test = "Wilks")

# Follow-up ANOVAs for each dependent variable
anova_results <- summary.aov(manova_test)

# Create a table for the results
result_table <- data.frame(
  Dependent_Variable = c("Age", "Daily Rate", "Monthly Income"),
  MANOVA_p_value = c(manova_wilks$stats[1, "Pr(>F)"]),
  ANOVA_p_value = c(anova_results[[1]]$`Pr(>F)`[1],
                    anova_results[[2]]$`Pr(>F)`[1],
                    anova_results[[3]]$`Pr(>F)`[1]),
  Interpretation = c("No significant effect of job satisfaction on age.",
                     "No significant effect of job satisfaction on daily rate.",
                     "No significant effect of job satisfaction on monthly income.")
)

# Print the table
print(result_table)
```

```
##   Dependent_Variable MANOVA_p_value ANOVA_p_value
## 1                Age      0.6245327     0.8953227
## 2         Daily Rate      0.6245327     0.1301764
## 3     Monthly Income      0.6245327     0.9047823
##                                                    Interpretation
## 1            No significant effect of job satisfaction on age.
## 2     No significant effect of job satisfaction on daily rate.
## 3 No significant effect of job satisfaction on monthly income.
```

Interpretation of Results

1. The results from the MANOVA and subsequent ANOVAs suggest that job satisfaction levels do not have a significant effect on employees' age, daily rate, or monthly income. The overall MANOVA test shows no significant difference between the groups with a p-value of 0.6245. Furthermore, the individual ANOVA tests for each dependent variable—age (p = 0.8953), daily rate (p = 0.1302), and monthly income (p = 0.9048)—also fail to show any statistically significant differences across the levels of job satisfaction. Therefore, the analysis concludes that job satisfaction does not significantly impact these factors in this dataset.

2. The ANOVA result for age across different job satisfaction levels shows no significant difference, with a p-value of 0.8953, indicating that age does not vary significantly with job satisfaction levels.

3. The ANOVA result for daily rate across different job satisfaction levels shows a p-value of 0.1302, which is greater than the typical significance level of 0.05, indicating that there is no significant difference in daily rate across job satisfaction levels.

4. The ANOVA result for monthly income across different job satisfaction levels shows a p-value of 0.9048, which is much higher than the typical significance level of 0.05, indicating that there is no significant difference in monthly income across job satisfaction levels.

5. Based on the ANOVA results, none of the dependent variables (age, daily rate, or monthly income) show significant differences across job satisfaction levels. However, if we look at the p-values, daily rate (p = 0.1302) is the closest to being significant, though still not significant at the 0.05 level, suggesting it might contribute slightly more to the differences between job satisfaction levels compared to age (p = 0.8953) and monthly income (p = 0.9048).

6. The MANOVA results show no significant interactions or patterns in the relationships between age, daily rate, and monthly income when grouped by job satisfaction levels, as indicated by the lack of significant results in both the Wilks' Lambda test (p = 0.6245) and the individual ANOVAs for each dependent variable.

```r
# 1. Pearson Correlation Test between Age, DailyRate, and MonthlyIncome
# Checking correlation between continuous variables
cor_test_age_daily <- cor.test(df$Age, df$DailyRate, method = "pearson")
cor_test_age_income <- cor.test(df$Age, df$MonthlyIncome, method = "pearson")
cor_test_daily_income <- cor.test(df$DailyRate, df$MonthlyIncome, method = "pearson")

# Print Pearson correlation results
print(cor_test_age_daily)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$Age and df$DailyRate
## t = 0.70272, df = 868, p-value = 0.4824
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04268826  0.09016796
## sample estimates:
##        cor
## 0.02384513
```

```r
print(cor_test_age_income)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$Age and df$MonthlyIncome
## t = 16.885, df = 891, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4410003 0.5405046
## sample estimates:
##       cor
## 0.4923595
```

```r
print(cor_test_daily_income)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$DailyRate and df$MonthlyIncome
## t = 0.72927, df = 1000, p-value = 0.466
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03893129  0.08486544
## sample estimates:
##        cor
## 0.02305546
```

```r
# 2. Kruskal-Wallis Test for non-parametric comparison across job satisfaction levels
kruskal_age <- kruskal.test(Age ~ JobSatisfaction, data = df)
```

```r
kruskal_daily_rate <- kruskal.test(DailyRate ~ JobSatisfaction, data = df)
kruskal_monthly_income <- kruskal.test(MonthlyIncome ~ JobSatisfaction, data = df)

# Print Kruskal-Wallis results
print(kruskal_age)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Age by JobSatisfaction
## Kruskal-Wallis chi-squared = 0.28755, df = 3, p-value = 0.9624
```

```r
print(kruskal_daily_rate)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  DailyRate by JobSatisfaction
## Kruskal-Wallis chi-squared = 5.0732, df = 3, p-value = 0.1665
```

```r
print(kruskal_monthly_income)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  MonthlyIncome by JobSatisfaction
## Kruskal-Wallis chi-squared = 0.092489, df = 3, p-value = 0.9927
```

These interpretations are aligned with the output in the `result_table`, showing the correct p-values and their corresponding meanings.

Supporting test using pearson correlation and kruskal:

1. **Do different levels of job satisfaction affect employees' age, daily rate, and monthly income?**
   - The Pearson correlation and Kruskal-Wallis tests suggest that job satisfaction levels do not significantly affect age, daily rate, or monthly income, as there are no significant correlations or differences found.
2. **Is there a significant difference in age across different job satisfaction levels?**
   - The Kruskal-Wallis test for age shows no significant difference across job satisfaction levels, with a p-value of 0.9624, indicating no notable variation.
3. **Is there a significant difference in daily rate across different job satisfaction levels?**
   - The Kruskal-Wallis test for daily rate shows a p-value of 0.1665, suggesting that job satisfaction levels do not significantly affect daily rates.
4. **Is there a significant difference in monthly income across different job satisfaction levels?**
   - The Kruskal-Wallis test for monthly income shows a p-value of 0.9927, indicating no significant differences in monthly income across job satisfaction levels.
5. **Which of the dependent variables (age, daily rate, or monthly income) contributes most to the differences between job satisfaction levels?**
   - None of the dependent variables (age, daily rate, or monthly income) show significant contributions to differences in job satisfaction levels, as supported by both the Pearson correlation and Kruskal-Wallis test results.
6. **Are there any interactions or patterns in the relationships between age, daily rate, and monthly income when grouped by job satisfaction levels?**
   - The Pearson correlation results show weak correlations between variables, and the Kruskal-Wallis tests show no significant differences, indicating no clear interactions or patterns in the relationships between age, daily rate, and monthly income based on job satisfaction levels.

Combining the results of MANOVA, ANOVA, PEARSON, AND KRUSKAL.

Here are the updated explanations, incorporating the results from the Pearson correlation and Kruskal-Wallis tests:

1. **Overall MANOVA and ANOVA Results**: The results from the MANOVA and subsequent ANOVAs suggest that job satisfaction levels do not have a significant effect on employees' age, daily rate, or monthly income. The overall MANOVA test, as indicated by Wilks' Lambda, shows no significant difference between the groups (p = 0.6245). Furthermore, the individual ANOVA tests for each dependent variable—age (p = 0.8953), daily rate (p = 0.1302), and monthly income (p = 0.9048)—also fail to show any statistically significant differences across the levels of job satisfaction. Therefore, the analysis concludes that job satisfaction does not significantly impact these factors in this dataset.

2. **Age Across Job Satisfaction Levels**: The ANOVA result for age across different job satisfaction levels shows no significant difference, with a p-value of 0.8953, indicating that age does not vary significantly with job satisfaction levels. Additionally, the Kruskal-Wallis test confirms this result, with a p-value of 0.9624, suggesting that there are no significant differences in age across job satisfaction levels using a non-parametric approach.

3. **Daily Rate Across Job Satisfaction Levels**: The ANOVA result for daily rate across different job satisfaction levels shows a p-value of 0.1302, which is greater than the typical significance level of 0.05, indicating that there is no significant difference in daily rate across job satisfaction levels. The Kruskal-Wallis test results support this, with a p-value of 0.1665, further confirming that job satisfaction does not significantly affect daily rate.

4. **Monthly Income Across Job Satisfaction Levels**: The ANOVA result for monthly income across different job satisfaction levels shows a p-value of 0.9048, which is much higher than the typical significance level of 0.05, indicating that there is no significant difference in monthly income across job satisfaction levels. Similarly, the Kruskal-Wallis test shows a p-value of 0.9927, further supporting the conclusion that job satisfaction has no significant impact on monthly income.

5. **Contributions of Dependent Variables**: Based on the ANOVA results, none of the dependent variables (age, daily rate, or monthly income) show significant differences across job satisfaction levels. However, the Pearson correlation test results provide additional insight. The correlation between **Age** and **DailyRate** is very weak, with a correlation coefficient of 0.0238 and a p-value of 0.4824, indicating no significant linear relationship. The correlation between **Age** and **MonthlyIncome** is moderate, with a correlation coefficient of 0.4924 and a highly significant p-value (< 2.2e-16), suggesting a moderate positive relationship. The correlation between **DailyRate** and **MonthlyIncome** is also weak, with a correlation coefficient of 0.0231 and a p-value of 0.4660, indicating no significant linear relationship. While **Age** and **MonthlyIncome** have a moderate correlation, no variables show significant differences across job satisfaction levels in either the ANOVA or Kruskal-Wallis tests.

6. **Interactions and Patterns**: The MANOVA results show no significant interactions or patterns in the relationships between age, daily rate, and monthly income when grouped by job satisfaction levels, as indicated by the lack of significant results in both the Wilks' Lambda test (p = 0.6245) and the individual ANOVAs for each dependent variable. The Pearson correlation results further support this, as there are no significant linear relationships between the variables (except for a moderate correlation between age and monthly income). The Kruskal-Wallis tests also show no significant differences in the distributions of the variables across job satisfaction levels, with p-values of 0.9624 for age, 0.1665 for daily rate, and 0.9927 for monthly income, reinforcing the conclusion that there are no patterns or significant interactions between the variables when grouped by job satisfaction.