Reliability Class Activity

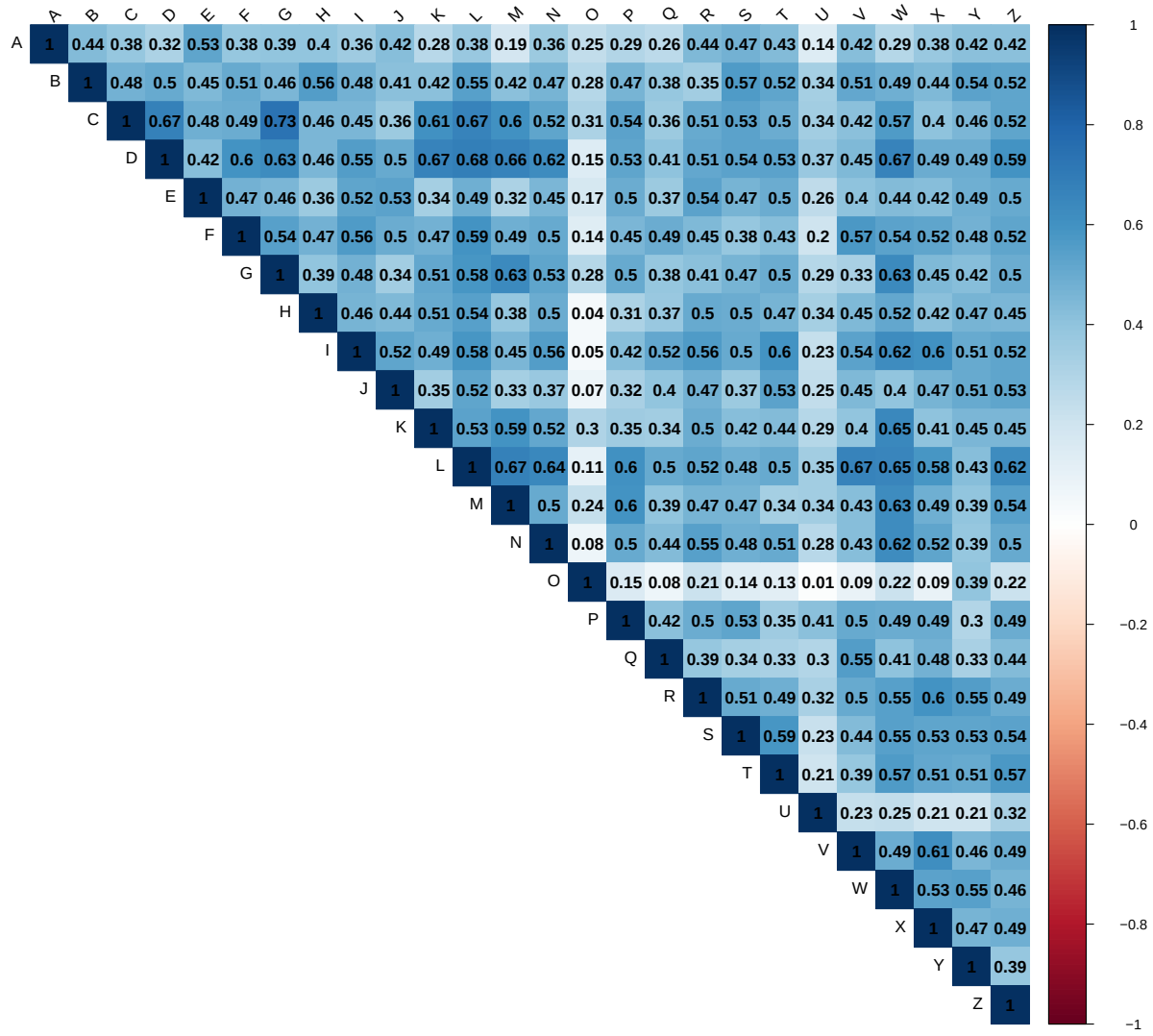
Vedanivas

20/02/2024

Contents

1 Advert Rating: Outlier Detection

```
library(openxlsx)
library(corrplot)
library(ggplot2)

file_path <- "BRSM_Assignment_2_datasets.xlsx"
sheet_name <- "Advert Rating"
advert_ratings <- read.xlsx(file_path, sheet = sheet_name)

corr_matrix <- cor(advert_ratings)

corrplot(corr_matrix, method = "color", type= "upper",
         tl.col = "black", tl.srt = 45, addCoef.col = "black")
```




```
avg_correlations <- colMeans(corr_matrix, na.rm = TRUE)
outlier <- which.min(avg_correlations)

cat("The outlier is the participant", names(outlier))
```

```
## The outlier is the participant O
```

2 Reliable Job: Internal Consistency

```
library(openxlsx)
library(psych)
```

```r
file_path <- "BRSM_Assignment_2_datasets.xlsx"
sheet_name <- "Reliable Job"
data <- read.xlsx(file_path, sheet = sheet_name)

calculate_cronbach_alpha <- function(df) {
  corr_matrix <- cor(df, method = "spearman")
  items <- ncol(df)
  mean_corr <- mean(corr_matrix[lower.tri(corr_matrix)])
  alpha <- (items * mean_corr) / (1 + ((items - 1) * mean_corr))
  return(alpha)
}

js_columns <- grep("JS", names(data), value = TRUE)
jp_columns <- grep("JP", names(data), value = TRUE)

js_columns <- js_columns[-5]
jp_columns <- jp_columns[-5]

js_data <- data[, js_columns]
jp_data <- data[, jp_columns]

cronbach_alpha_js <- calculate_cronbach_alpha(js_data)
cronbach_alpha_jp <- calculate_cronbach_alpha(jp_data)

cat("Cronbach's Alpha for Job Satisfaction (JS):", cronbach_alpha_js,
    "\nCronbach's Alpha for Job Performance (JP):", cronbach_alpha_jp)
```

```
## Cronbach's Alpha for Job Satisfaction (JS): 0.8584497
## Cronbach's Alpha for Job Performance (JP): 0.5242351
```

Based on the calculated Cronbach's alpha values for Job Satisfaction (JS) and Job Performance (JP):

- **Job Satisfaction (JS) with a Cronbach's Alpha of 0.8584497**: This value is well above the commonly accepted threshold of 0.7 for acceptable internal consistency, indicating excellent internal consistency among the items measuring job satisfaction. A Cronbach's alpha of approximately 0.858 suggests that the items included in the Job Satisfaction scale are very well aligned and effectively measure the same underlying construct. This high level of coherence among the items suggests that the Job Satisfaction scale is a reliable measure for assessing job satisfaction.
- **Job Performance (JP) with a Cronbach's Alpha of 0.5242351**: This value is below the commonly accepted threshold of 0.7 for acceptable internal consistency. A Cronbach's alpha of approximately 0.524 indicates poor internal consistency among the items measuring job performance. This suggests that the items included in the Job Performance scale may not all be effectively measuring the same underlying construct, or there could be a lack of cohesion among the items. It may be necessary to review and possibly revise the items in this scale to improve its reliability. Adjustments could include refining the existing items or introducing new ones that are more aligned with the overall construct of job performance.

Commentary on Internal Consistency and Acceptability:

- The **Job Satisfaction** measure demonstrates excellent reliability and internal consistency, making it a robust tool for assessing job satisfaction. The high Cronbach's alpha value indicates that the scale

is composed of items that cohesively measure the construct of job satisfaction, providing confidence in the scale's use for research or practical applications.

- In contrast, the **Job Performance** measure exhibits a level of reliability that falls short of the acceptable standard, indicating a need for improvement. The lower Cronbach's alpha value suggests that the scale may benefit from a thorough review of its items to ensure they collectively measure the intended construct of job performance. Enhancing the internal consistency of this scale is crucial for obtaining reliable and valid assessments of job performance.

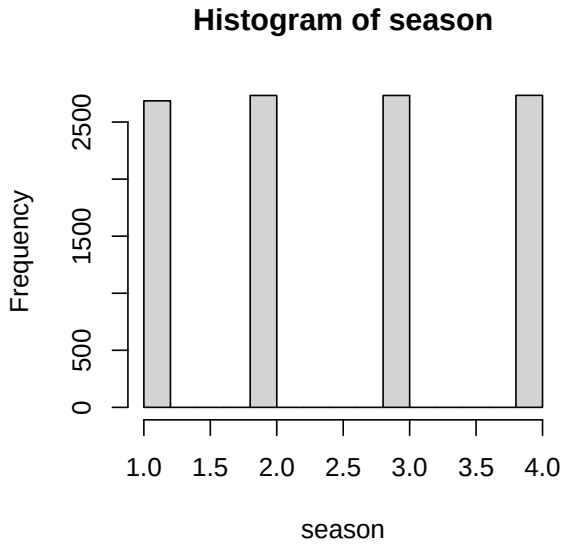
Conclusion:

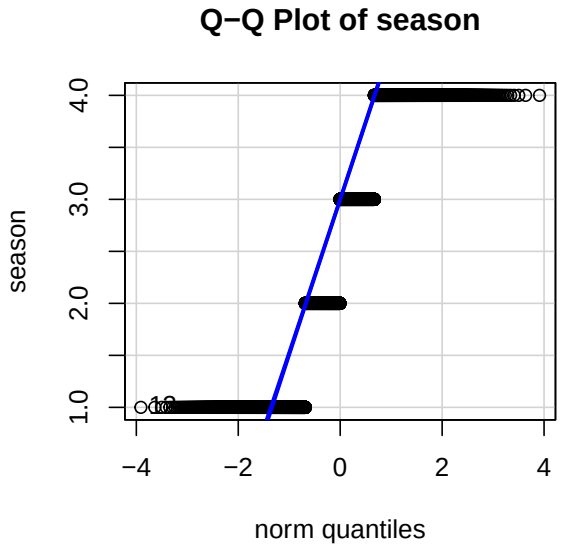
The analysis highlights a disparity in the reliability of the two scales. While the Job Satisfaction scale shows an excellent level of reliability and is deemed a highly consistent and effective tool for measuring job satisfaction, the Job Performance scale requires significant improvements to reach an acceptable level of reliability. Efforts to refine the Job Performance scale should focus on increasing the Cronbach's alpha value by ensuring that all items are relevant and contribute positively to the measurement of job performance.

3 Yulu: Normality Testing

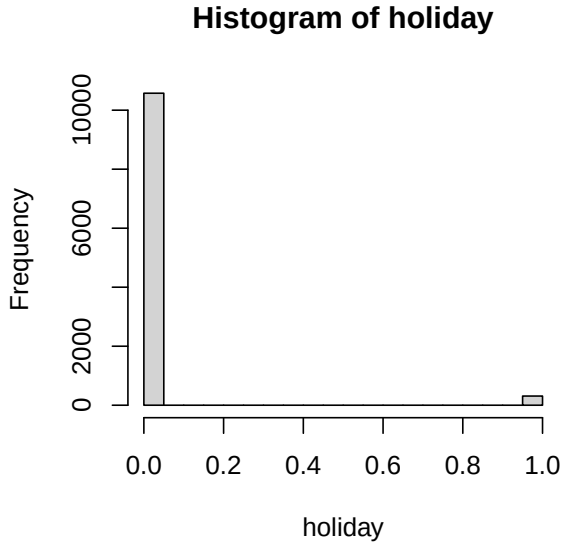
```
library(openxlsx)
library(ggplot2)
library(car)

explore_variable <- function(variable_name, data) {
  par(mfrow=c(1, 2))
  hist(data[[variable_name]], main=paste("Histogram of", variable_name),
       xlab=variable_name)

  qqPlot(data[[variable_name]], main=paste("Q-Q Plot of", variable_name),
         ylab=variable_name)

  if (length(data[[variable_name]]) > 5000) {
    sampled_data <- sample(data[[variable_name]], 5000, replace = FALSE)
  } else {
    sampled_data <- data[[variable_name]]
  }

  shapiro_result <- shapiro.test(sampled_data)
  cat("Shapiro-Wilk Test for", variable_name, ":\n")
  print(shapiro_result)
  cat("\n")
}

file_path <- "BRSM_Assignment_2_datasets.xlsx"
data <- read.xlsx(file_path, sheet = "Yulu")

# Excluding the datetime column and the categorical columns
variables <- names(data)[-1]

for(variable in variables) {
```

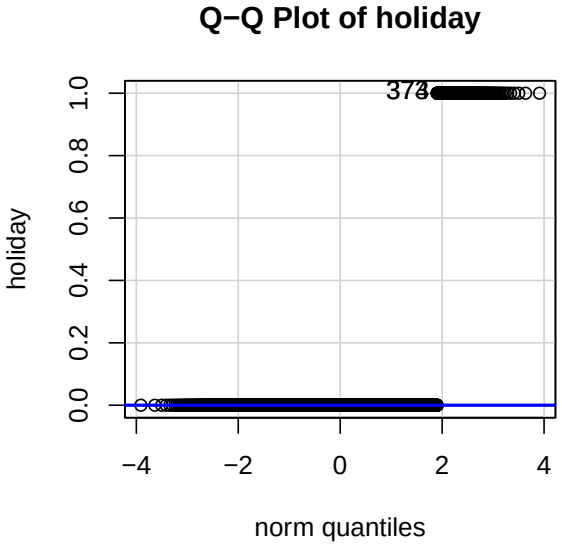
```
  explore_variable(variable, data)
}
```





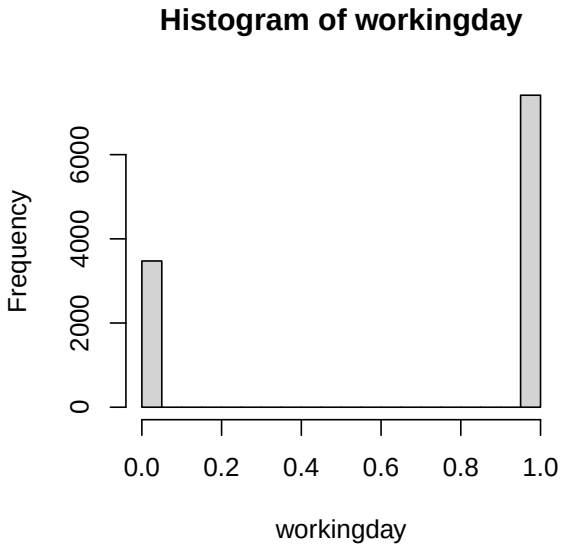

```
## Shapiro-Wilk Test for season :
##
##  Shapiro-Wilk normality test
##
## data:  sampled_data
## W = 0.85602, p-value < 2.2e-16
```

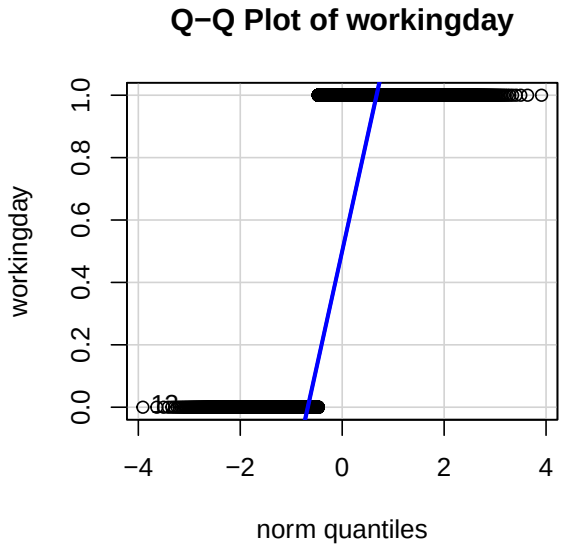




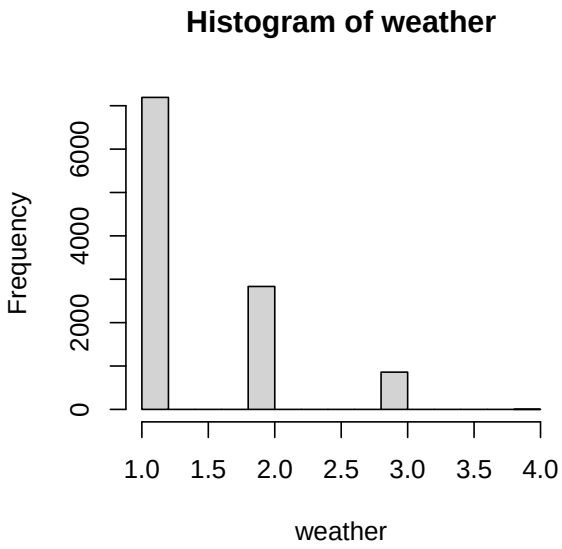

```
## Shapiro-Wilk Test for holiday :
```

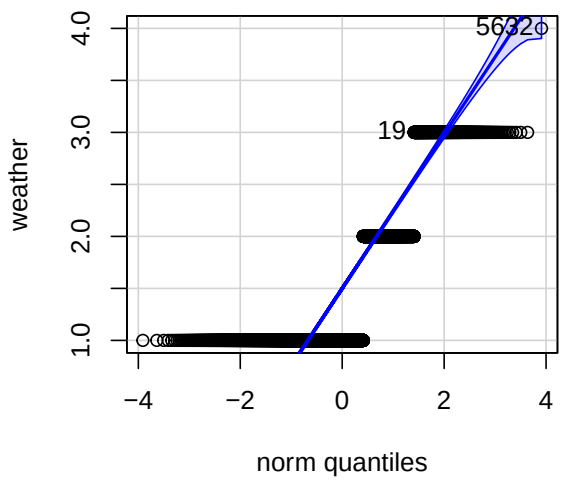
```
##
##  Shapiro-Wilk normality test
##
## data:  sampled_data
## W = 0.15314, p-value < 2.2e-16
```



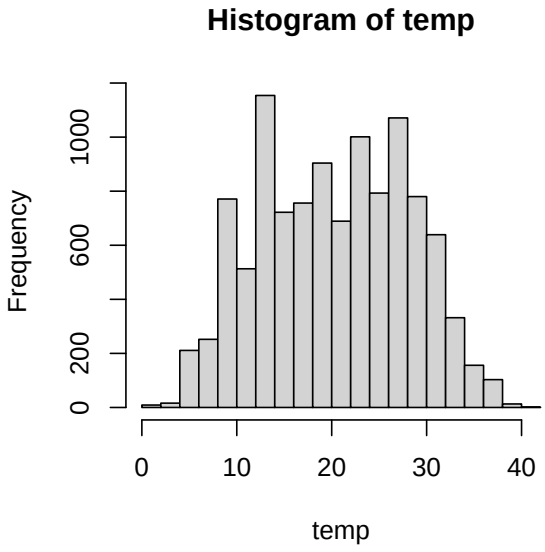


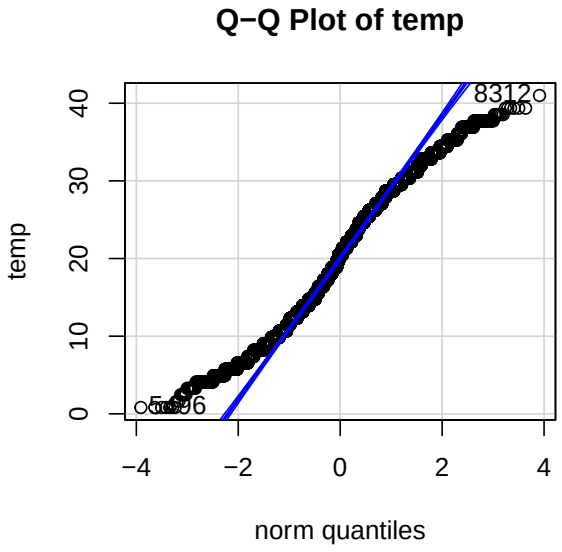

```
## Shapiro-Wilk Test for workingday :
##
##  Shapiro-Wilk normality test
##
## data:  sampled_data
## W = 0.57943, p-value < 2.2e-16
```



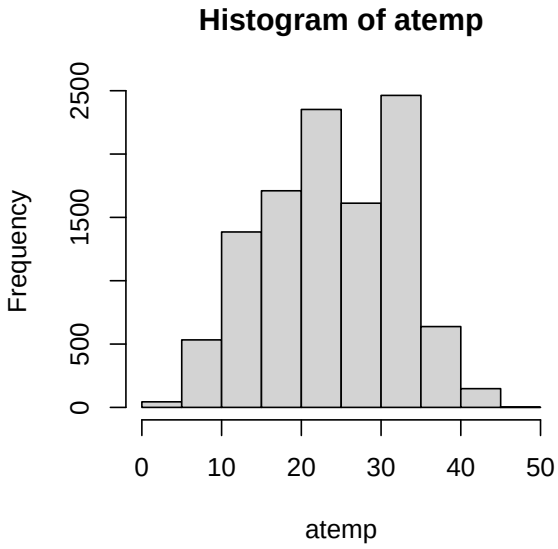


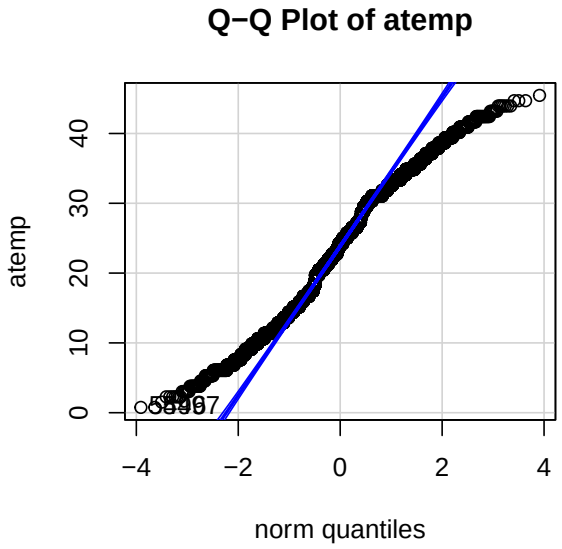
```
## Shapiro-Wilk Test for weather :
##
##  Shapiro-Wilk normality test
##
## data:  sampled_data
## W = 0.6556, p-value < 2.2e-16
```



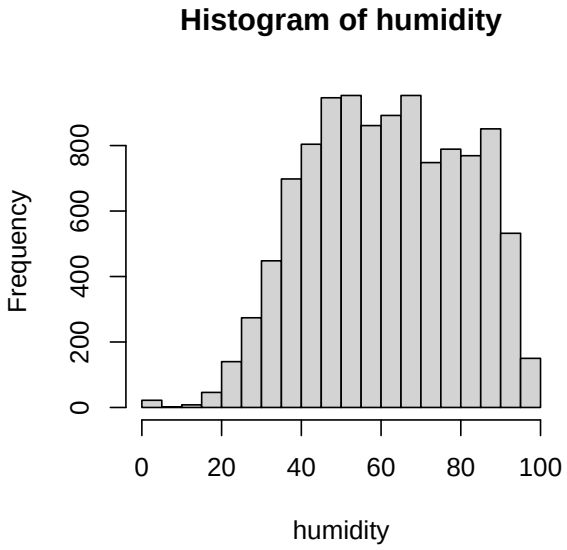


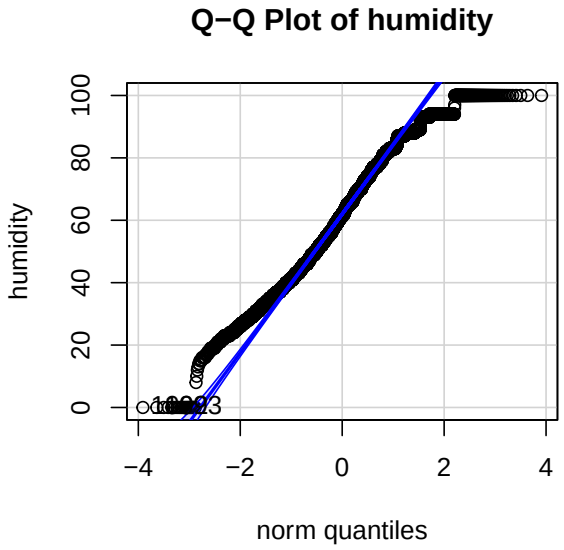

```
## Shapiro-Wilk Test for temp :
##
##  Shapiro-Wilk normality test
##
## data:  sampled_data
## W = 0.97946, p-value < 2.2e-16
```



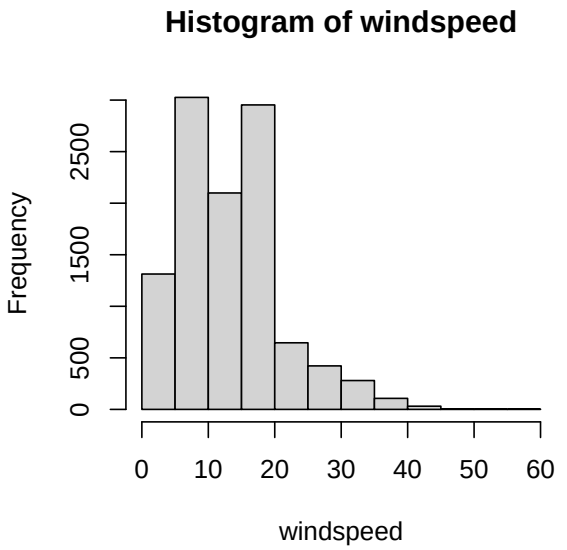


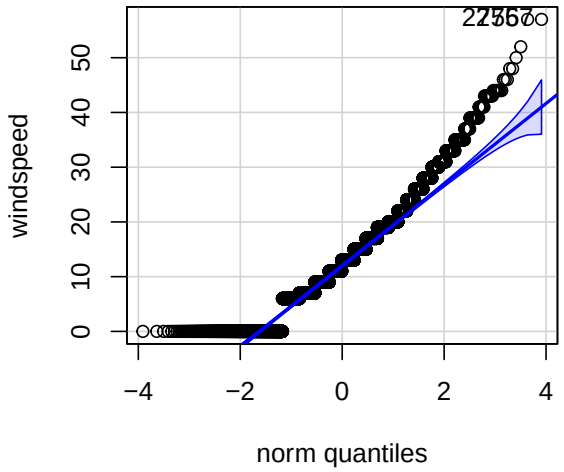
```
## Shapiro-Wilk Test for atemp :
##
##  Shapiro-Wilk normality test
##
## data:  sampled_data
## W = 0.98192, p-value < 2.2e-16
```



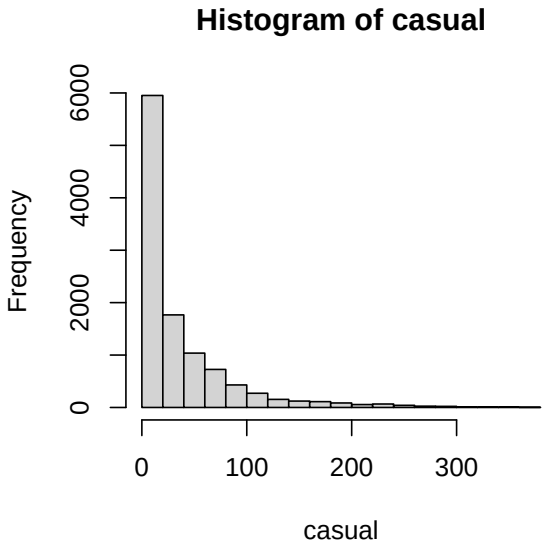


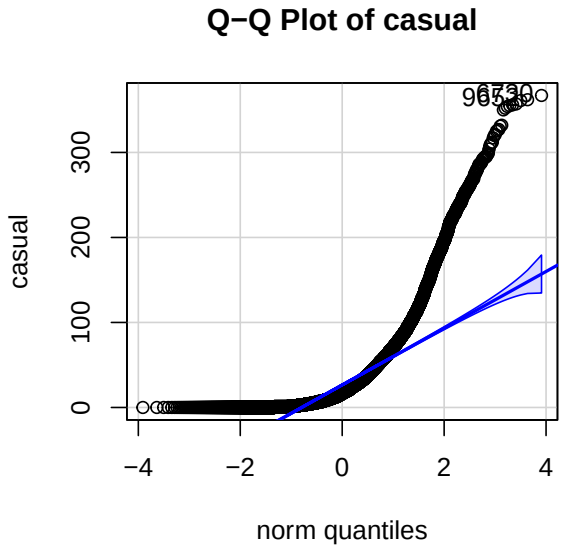

```
## Shapiro-Wilk Test for humidity :
##
##  Shapiro-Wilk normality test
##
## data:  sampled_data
## W = 0.98243, p-value < 2.2e-16
```



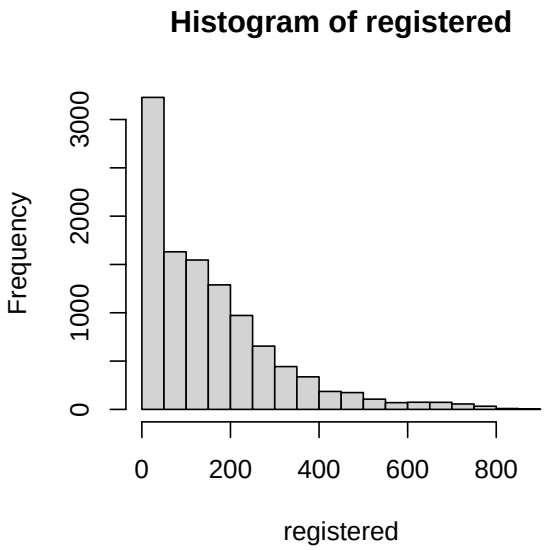


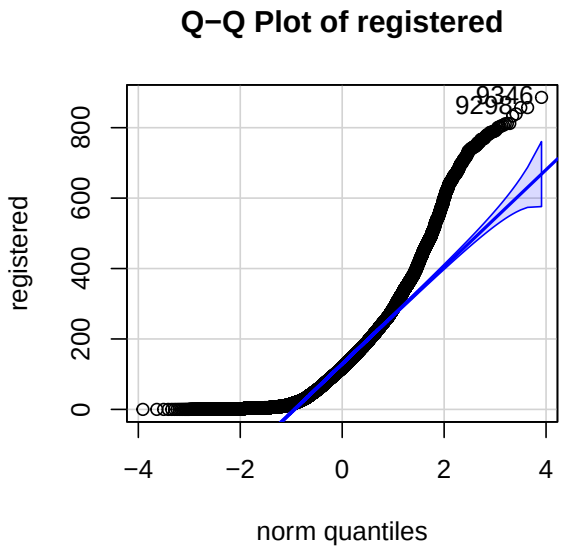
```
## Shapiro-Wilk Test for windspeed :
##
##  Shapiro-Wilk normality test
##
## data:  sampled_data
## W = 0.95967, p-value < 2.2e-16
```



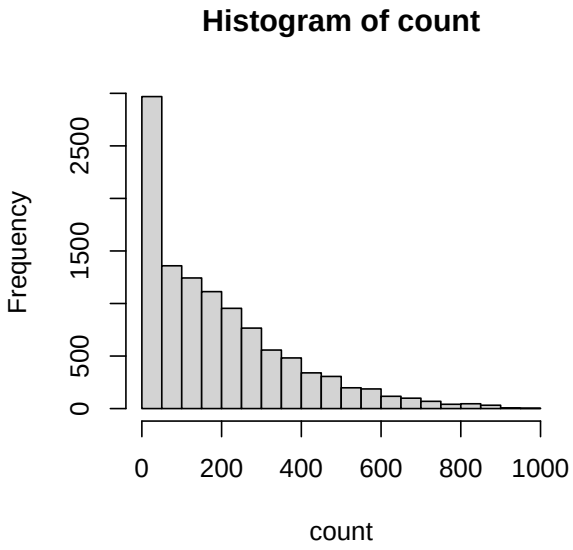


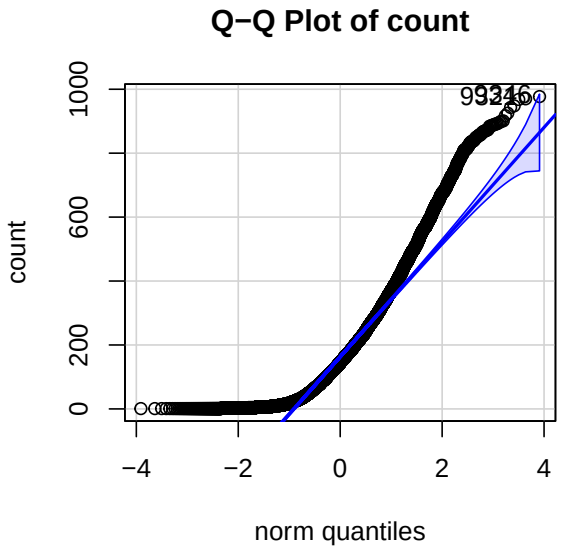

```
## Shapiro-Wilk Test for casual :
##
##  Shapiro-Wilk normality test
##
## data:  sampled_data
## W = 0.69504, p-value < 2.2e-16
```





```
## Shapiro-Wilk Test for registered :
##
##  Shapiro-Wilk normality test
##
## data:  sampled_data
## W = 0.86018, p-value < 2.2e-16
```






```
## Shapiro-Wilk Test for count :
##
##  Shapiro-Wilk normality test
##
## data:  sampled_data
## W = 0.88114, p-value < 2.2e-16
```

The variables season, holiday, workingday, weather are categorical so they can't be normal distribution. From the plots and the W-values of the Shapiro-Wilk test, the variables windspeed, casual, registered, and count are comparatively more deviated from normal distribution so they have to be transformed using Box-Cox.

```r
library(MASS)

transform_variable <- function(variable_name, data) {
  data[[variable_name]] <- data[[variable_name]] +
    abs(min(data[[variable_name]])) + 1
  transformed_data <- boxcox(data[[variable_name]] ~ 1, plot=FALSE)
  lambda <- transformed_data$x[which.max(transformed_data$y)]

  return (data[[variable_name]]^lambda - 1) / lambda
}

non_normal_variables <- c("windspeed", "casual", "registered", "count")
for (variable in non_normal_variables) {
  data[[variable]] <- transform_variable(variable, data)
}
```

```r
library(ppcor)
library(knitr)

main_variables <- c("count", "registered", "casual")
control_variables <- c("season", "holiday", "workingday", "weather",
                       "temp", "atemp", "windspeed", "humidity")

results_df <- data.frame(
  "Var1" = character(),
  "Var2" = character(),
  "PartialCorr" = numeric(),
  "PartialP" = numeric(),
  "SemiPartialCorr" = numeric(),
  "SemiPartialP" = numeric(),
  stringsAsFactors = FALSE
)

for (main_var in main_variables) {
  for (control_var in control_variables) {
    current_controls <- setdiff(control_variables, control_var)
    pcor_test <- pcor.test(data[[main_var]], data[[control_var]],
                           data[,current_controls], method = "pearson")
    spcor_test <- spcor.test(data[[main_var]], data[[control_var]],
                             data[,current_controls], method = "pearson")

    results_df <- rbind(results_df, data.frame(
      "Var1" = main_var,
      "Var2" = control_var,
      "PartialCorr" = pcor_test$estimate,
      "PartialP" = pcor_test$p.value,
      "SemiPartialCorr" = spcor_test$estimate,
      "SemiPartialP" = spcor_test$p.value
    ))
  }
}

kable(results_df, format = "latex",
      caption = "Partial and Semi-Partial Correlation Results")
```

Based on the provided table of partial and semi-partial correlation results:

Existence of Correlations:

- **Partial Correlations**: There are statistically significant partial correlations between many of the pairs of variables. For instance, `count` has a strong negative partial correlation with `humidity`, and `casual` has a strong negative partial correlation with `workingday`. These correlations persist even after controlling for the effects of the other variables in the model.
- **Semi-Partial Correlations**: Similarly, there are significant semi-partial correlations. The semi-partial correlation gives us insight into the unique contribution of one variable to the dependent variable while controlling for other variables. For example, `casual` has a significant negative semi-partial correlation with `humidity`, suggesting that `casual` uniquely predicts the dependent variable while accounting for other factors.

Table 1: Partial and Semi-Partial Correlation Results

Var1	Var2	PartialCorr	PartialP	SemiPartialCorr	SemiPartialP
count	season	0.1686774	0.0000000	0.1440376	0.0000000
count	holiday	-0.0115566	0.2280947	-0.0097277	0.3103306
count	workingday	-0.0198535	0.0383836	-0.0167137	0.0812978
count	weather	0.0450623	0.0000026	0.0379669	0.0000746
count	temp	0.0065631	0.4936764	0.0055242	0.5645318
count	atemp	0.0624630	0.0000000	0.0526772	0.0000000
count	windspeed	0.0576169	0.0000000	0.0485762	0.0000004
count	humidity	-0.3399263	0.0000000	-0.3042279	0.0000000
registered	season	0.1744950	0.0000000	0.1544654	0.0000000
registered	holiday	-0.0035409	0.7119167	-0.0030864	0.7475427
registered	workingday	0.0594211	0.0000000	0.0518851	0.0000001
registered	weather	0.0424800	0.0000093	0.0370605	0.0001105
registered	temp	0.0007737	0.9356908	0.0006744	0.9439314
registered	atemp	0.0542599	0.0000000	0.0473645	0.0000008
registered	windspeed	0.0574738	0.0000000	0.0501790	0.0000002
registered	humidity	-0.3109899	0.0000000	-0.2852119	0.0000000
casual	season	0.1298880	0.0000000	0.0914316	0.0000000
casual	holiday	-0.0430701	0.0000070	-0.0300893	0.0016968
casual	workingday	-0.3257761	0.0000000	-0.2404999	0.0000000
casual	weather	0.0435527	0.0000055	0.0304271	0.0015036
casual	temp	0.0273464	0.0043376	0.0190940	0.0464259
casual	atemp	0.1007963	0.0000000	0.0707123	0.0000000
casual	windspeed	0.0607581	0.0000000	0.0424854	0.0000093
casual	humidity	-0.4068399	0.0000000	-0.3108480	0.0000000

Implications:

- The presence of significant partial correlations indicates that the relationships between the pairs of variables are not merely due to shared variance with the control variables. For example, the strong negative correlation between `count` and `humidity` implies that as humidity increases, the total count of vehicle rentals decreases, independent of the season, temperature, windspeed, and other factors considered.
- Significant semi-partial correlations indicate that a variable has a unique relationship with the dependent variable, which is not explained by the control variables. For example, the negative semi-partial correlation between `casual` and `workingday` suggests that there are fewer casual riders on working days, independently of the weather conditions or the season.

Usage of Partial Correlations:

- Partial correlations are used to understand the direct relationship between two variables by removing the influence of one or more other variables. This is particularly useful in complex datasets where many interrelated factors may influence the variables of interest. It helps in identifying the strength and direction of associations that are not confounded by other variables.

Observations:

- High correlations (both partial and semi-partial) involving `humidity` suggest that this variable has a strong relationship with the total count and types of users, potentially more so than other environmental factors.
- The correlations involving `workingday` indicate that the patterns of vehicle usage are different on working days compared to holidays or weekends, even after accounting for other factors.
- The relatively low correlations with `temp` and `atemp` suggest that the temperature has a less direct influence on vehicle rental behaviors when other factors are considered.

Conclusion:

The observed correlations provide valuable insights into the factors affecting vehicle rental behaviors. High partial and semi-partial correlations signify that despite the influence of other variables, certain factors like `humidity` and `workingday` have strong and unique relationships with vehicle rental counts and types of users. These findings can inform strategies to improve vehicle rental services by addressing the specific needs and patterns of users based on different conditions and times.