7COM1079-0901-2024 - Team Research and Development Project

Final report title: Analyzing Weather Data in R

Group ID: A82

Dataset number: DS031

Prepared by: Muhammad Hamza Zubair

University of Hertfordshire
Hatfield, 2024

Table of Contents

1. Visualisation
   1.1. Appropriate plot for the RQ

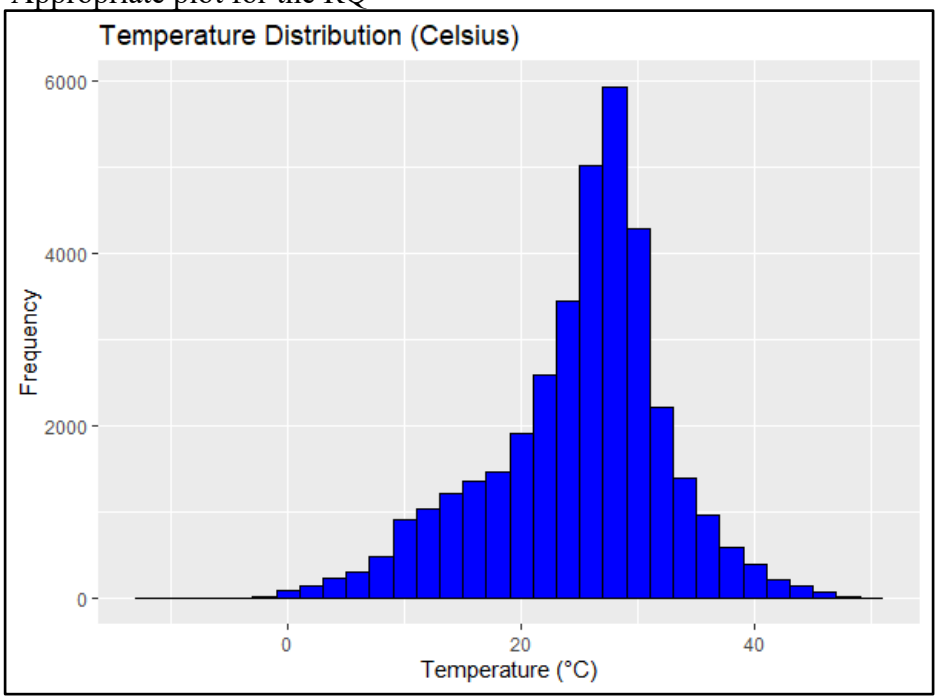

Figure 1: Histogram for temperature in Celsius

In the above image a histogram of temperature distribution in Celsius with a positive skewness of a normal distribution curve. The peak frequency ranges from 20 to 25°C while the other temperatures range from 0 up to 40°C.

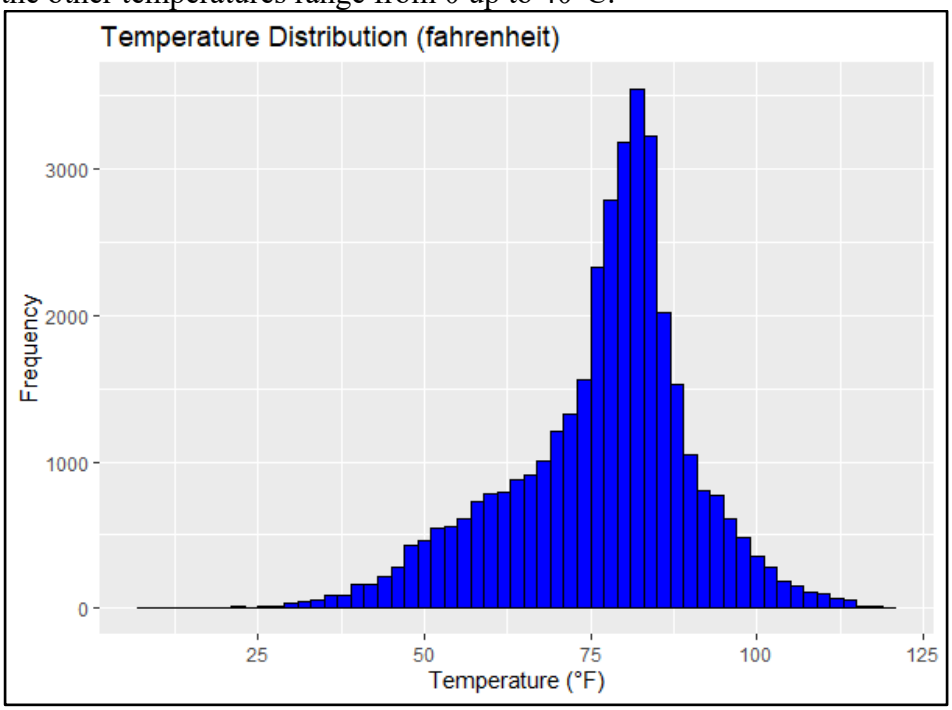

Figure 2: Histogram for temperature in Fahrenheit

The above image shows the temperature distribution in Fahrenheit form and is also right skewed with the peak of the bell curve. This is approximately equivalent to 24-51°C and the peak frequency is at about 38-27°C, maximum and minimum range respectively.

1.2. Additional information relating to understanding the data (optional)

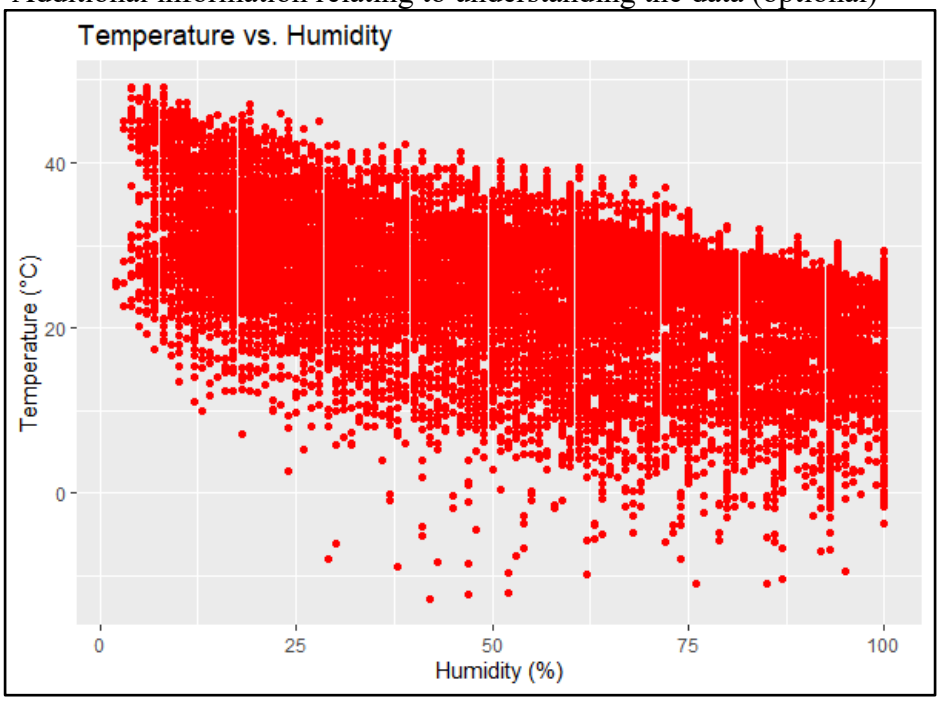

Figure 3: Scatter plot for temperature vs. humidity

The above image shows the temperatures against the humidity, a negative correlation is depicted. It also shows how the temperature decreases generally as the relative humidity increases from 0% to 100% with significant fluctuation in the data points.

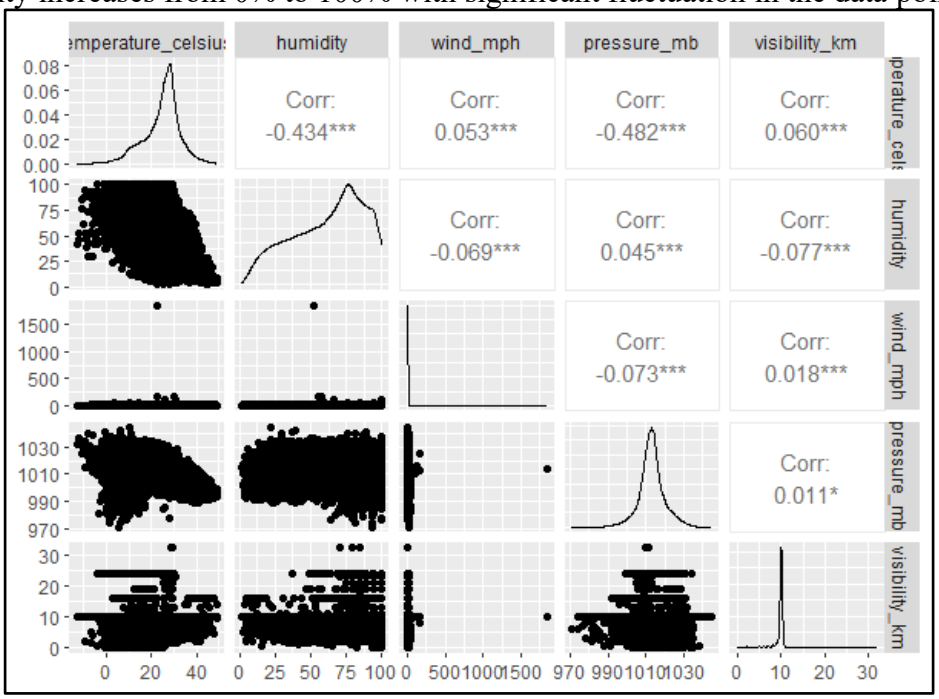

Figure 4: Pairplot of Temperature, humidity, wind_mph, pressure_mb, visdibility_km

The above image is a pair plot that captures the information about the relationships between the different weather measures such as temperature, humidity, wind speed, pressure, and visibility, all represented with distribution curves.

1.3. Useful information for the data understanding

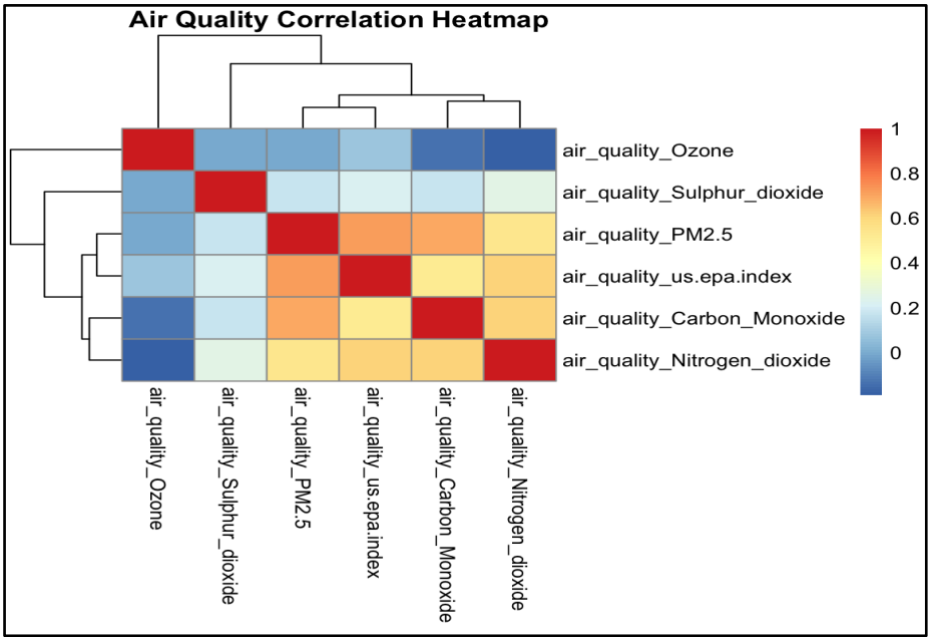

Figure 5: Air Quality Correlation Heatmap

The above figure is an image that shows a heatmap of air quality correlation using the red-blue color gradient. This one demonstrates a correlation between the different air pollutants such as Ozone, Sulphur dioxide, Nitrogen dioxide, Carbon Monoxide, and PM2.5, air_quality_us-epa-index.

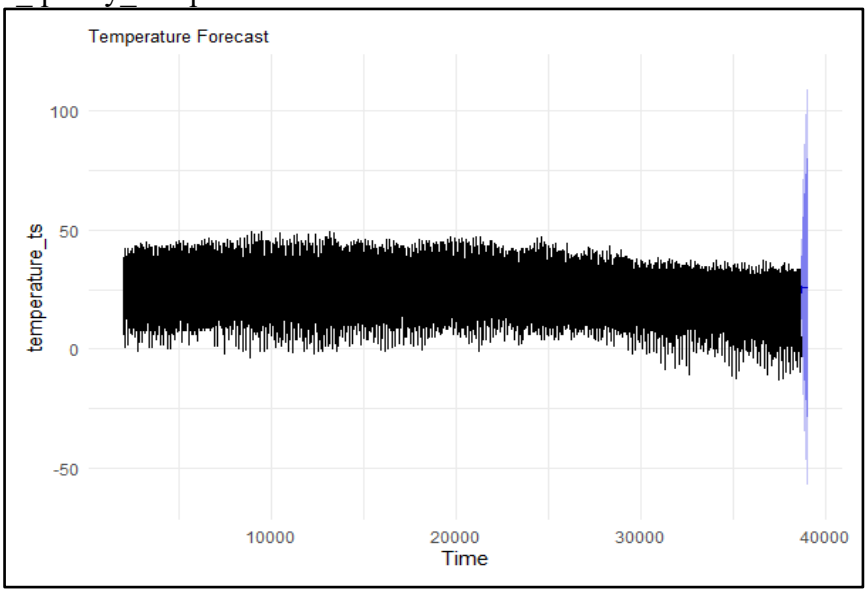

Figure 6: Temperature Forecasting

The above figure shows a time series of temperature forecasts having a large amount of variation. Black scatters vary around and lie between 0–50°F and there is an isolated blue hump at the last point.

2. Analysis

2.1. Statistical test used to test the hypotheses and output

4.1.1. Correlation Test

Source Code

```r
## Test 1: Pearson correlation
correlation_test <- cor.test(
  data$temperature_celsius,
  data$air_quality_us.epa.index,
  method = "pearson"
  )

cat("Pearson Correlation Coefficient:", correlation_test$estimate, "\n")
cat("p-value:", correlation_test$p.value, "\n")

if (correlation_test$p.value < 0.05) {
  cat("As there exists a statistical significance between temperature and air_quality_us-epa-index,\nthese variables are correlated.\n")
} else {
  cat("As there is no statistical significance between temperature and air_quality_us-epa-index,\nthese variables are not correlated.\n")
}
```

Formulation

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Figure 7: Pearson correlation test

The Pearson correlation method provides the measure of such correlation between the **temperature_celsius** and **air_quality_us-epa-index**. Covariance is calculated between the variables of interest and standard deviation is the square root of the average of the squared difference of a variable from the mean (Talbot *et al.,* 2021). It also focuses on the analysis of the relationship between fluctuation in temperature with fluctuations in air_quality_us-epa-index in the data set.

Hence, the selected variables for this test are **temperature_celsius** (x) and **air_quality_us-epa-index** (y). The formula implemented in the program looks to determine the correlation coefficient of x and y along with the p_value for comparison. The x_i and y_i represents each data points contained by the variables x and y while the terms $\bar{x}$ and $\bar{y}$ denote the mean value of corresponding columns

4.1.2. Chi-square Test

Source Code

```r
### Performing the Chi-Square Test
chi_square_test <- chisq.test(contingency_table)

cat("Chi-Square Test Statistic:", chi_square_test$statistic, "\n")
cat("p-value:", chi_square_test$p.value, "\n")

if (chi_square_test$p.value < 0.05) {
  cat("As there exists a statistical significance between temperature categories and air_quality_us-epa-index,\nthese variables are dependent (i.e., correlated).\n")
} else {
  cat("As there is no statistical significance between temperature categories and air_quality_us-epa-index,\nthese variables are independent (i.e., not-correlated).\n")
}
```

Formulation

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$ [Test statistics]

and,

$$E_{ij} = \frac{(row\ total)_i \times (column\ total)_j}{grand\ total}$$ [Expected frequency]

Figure 8: Chi-square test

The validity of the relationship between the nominal variable, temperature_celsius, and the interval variable, air_quality_us-epa-index, is tested using the Chi-Square Test of Independence by comparing the observed frequencies, O_{ij}, in the contingency table with expected frequencies, E_{ij}, considering the variables are independent. The test statistic is calculated with the values of expected frequencies obtained from marginal totals (Sun and Khayatnezhad, 2021).

4.1.3. Augmented Dicky-Fuller Test

```
> print(adf_result_temp)

        Augmented Dickey-Fuller Test

data:  data$temperature_celsius
Dickey-Fuller = -21.076, Lag order = 33, p-value = 0.01
alternative hypothesis: stationary
```

Figure 9: ADF testing on temperature

The Augmented Dickey-Fuller (ADF) test has been used and it determines the order of Integration for both temperature and humidity time series data. For temperature, the ADF test statistic was -21.076 with a lag order of 33, a significance level of 0.01.

```
> print(adf_result_humidity)

        Augmented Dickey-Fuller Test

data:  data$humidity
Dickey-Fuller = -31.248, Lag order = 33, p-value = 0.01
alternative hypothesis: stationary
```

Figure 10: ADF testing on Humidity

For the humidity, the test statistic was -31.248 with the lag order productive of the same p-value. PEMST and BDQ-TEST both fail to support the null hypothesis of non-stationary which suggests that both the time series are stationary.

2.2. The null hypothesis is rejected /not rejected based on the p-value

4.2.1. To investigate how temperature and air_quality_us-epa-index **are correlated**

Test 1: Pearson correlation

```
> cat("Pearson Correlation Coefficient:", correlation_test$estimate, "\n")
Pearson Correlation Coefficient: 0.0318125
> cat("p-value:", correlation_test$p.value, "\n")
p-value: 1.10705e-09
```

```
> if (correlation_test$p.value < 0.05) {
+   cat("As there exists a statistical significance between temperature and air_quality_us-epa-index,\nthese variables are correlated.\n")
+ } else {
+   cat("As there is no statistical significance between temperature and air_quality_us-epa-index,\nthese variables are not correlated.\n")
+ }
As there exists a statistical significance between temperature and air_quality_us-epa-index,
these variables are correlated.
```

Figure 11: Result of correlation test

The Pearson correlation test conducted involving the variables “temperature” and “air_quality_us-epa-index” generates a p_value of “1.10705e-09” that is less than the standard significance level of 0.05. This suggests the alternative hypothesis to be true. Hence, the aforementioned research question can be answered by the statement – temperature and air_quality_us-epa-index are correlated.

Test 2: Chi-Square Test of Independence

```
> cat("Chi-Square Test Statistic:", chi_square_test$statistic, "\n")
Chi-Square Test Statistic: 686.4515
> cat("p-value:", chi_square_test$p.value, "\n")
p-value: 2.990838e-147
```

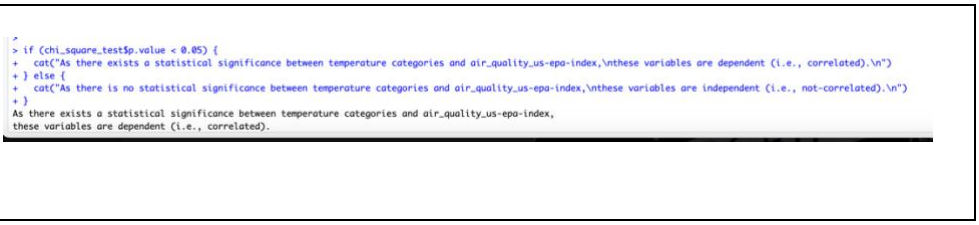

Figure 12: Result of Chi-square test

Similarly, the p_value derived from the Chi-square test implies the null hypothesis to be false as it (2.990838e-147) appears to be smaller than the standard significance level of 0.05. Thus, to answer the research question, it can be said that the variables temperature and air_quality_us-epa-index are correlated.

4.2.2. To check data stationarity

In the Augmented Dickey-Fuller test, p-values are obtained and these are 0.01 for both the temperature and humidity time series data analyzed. For both of these p-values, it obtains lower results than the conventional significance level of 0.05 and thus rejects the null hypothesis that the series is non-stationary. The null hypothesis is rejected with test statistics of -21.076 for temperature, and -31.248 for humidity. The first analysis of the data shows that they are not non-stationary and they do not exhibit unit root hence they have stable statistical properties in the time series.

3. Reference list

Bamal, A., Uddin, M.G. and Olbert, A.I., 2024. Harnessing machine learning for assessing climate change influences on groundwater resources: a comprehensive review. *Heliyon*. Harvard (author, date) format.

Ji, X., Dong, W., Wang, W., Dai, X. and Huang, H., 2024. Impacts of Climate Change on Extreme Precipitation Events and Urban Waterlogging: A Case Study of Beijing. *Natural Hazards Review*, *25*(1), p.05023014.

Jihan, M.A.T., Popy, S., Kayes, S., Rasul, G. and Rahman, M.M., 2024. Climate Change Scenario in Bangladesh: Historical Data Analysis and Future Projection Based on CMIP6 Model.

Krogh, A., Junginger, M., Shen, L., Grue, J. and Pedersen, T.H., 2024. Climate change impacts of bioenergy technologies: A comparative consequential LCA of sustainable fuels production with CCUS. *Science of The Total Environment*, p.173660.

Mousavi, S.M., Mobarghaee Dinan, N., Ansarifard, S., Darvishi, G., Borhani, F. and Naghibi, A., 2024. Assessing the impact of global carbon dioxide changes on atmospheric fluctuations in Iran through satellite data analysis. *Journal of Water and Climate Change*, p.jwc2024702.

Sun, X. and Khayatnezhad, M., 2021. Fuzzy-probabilistic modeling the flood characteristics using bivariate frequency analysis and α-cut decomposition. *Water Science & Technology*, *21*(8), p.4391.

Talbot, D., Diop, A., Lavigne-Robichaud, M. and Brisson, C., 2021. The change in estimate method for selecting confounders: A simulation study. *Statistical methods in medical research*, *30*(9), pp.2032-2044.

Tyystjärvi, V., Markkanen, T., Backman, L., Raivonen, M., Leppänen, A., Li, X., Ojanen, P., Minkkinen, K., Hautala, R., Peltoniemi, M. and Anttila, J., 2024. Future methane fluxes of peatlands are controlled by management practices and fluctuations in hydrological conditions due to climatic variability. *Biogeosciences*, *21*(24), pp.5745-5771.

4. Appendices

R code used for analysis and visualisation

```
# Research Question
# Is there a correlation between temperature and the air quality index (US EPA)
# in different global locations?

# Hypotheses
# Null Hypothesis (H0): As there is no statistical significance between
# temperature and air quality index (US EPA) these variables are not correlated.
# Alternative Hypothesis (H1): As there exists a statistical significance between
# temperature and air quality index (US EPA) these variables are correlated.

# Selected Hypothesis Tests
# Test 1: Pearson correlation
# Test 2: Chi-Square Test of Independence

# Loading necessary libraries
library(ggplot2)
library(dplyr)
library(forecast)
library(zoo)
library(lubridate)
library(tseries)
library(GGally)
library(pheatmap)

# Loading the data
file_path <- "GlobalWeatherRepository.csv"
data <- read.csv(file_path)

# Previewing the dataset
head(data)
str(data)
summary(data)


# Checking for missing values
sum(is.na(data))

# Handling missing values (e.g., remove or impute)
data <- na.omit(data) # Removing rows with missing values

# Checking for duplicates
duplicates <- data[duplicated(data), ]
print(duplicates)
```

```
# Removing duplicates
data <- data[!duplicated(data), ]

# Checking data types and convert if necessary
str(data)

data$humidity <- as.numeric(data$humidity)
data$temperature_celsius <- as.numeric(data$temperature_celsius)

# Histogram for temperature in Celsius
ggplot(data, aes(x = temperature_celsius)) +
  geom_histogram(binwidth = 2, fill = "blue", color = "black") +
  labs(title = "Temperature Distribution (Celsius)", x = "Temperature (°C)", y = "Frequency")

# Histogram for temperature in Fahrenheit
ggplot(data, aes(x = temperature_fahrenheit)) +
  geom_histogram(binwidth = 2, fill = "blue", color = "black") +
  labs(title = "Temperature Distribution (Fahrenheit)", x = "Temperature (°F)", y =
"Frequency")

# Boxplot for temperature by country
ggplot(data, aes(x = country, y = temperature_celsius, fill = country)) +
  geom_boxplot() +
  labs(title = "Temperature by Country", x = "Country", y = "Temperature (°C)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Scatter plot for temperature vs. humidity
ggplot(data, aes(x = humidity, y = temperature_celsius)) +
  geom_point(color = "red") +
  labs(title = "Temperature vs. Humidity", x = "Humidity (%)", y = "Temperature (°C)")

# Performing visual inspection as well as hypothesis testing (Shapiro-Wilk test)
# to answer the research question
## Visual inspection
ggplot(data, aes(x = temperature_celsius, y = air_quality_us.epa.index)) +
  geom_point() +
  labs(title = "Scatter Plot of Temperature vs air_quality_us-epa-index",
       x = "Temperature (°C)", y = "air_quality_us-epa-index") +
  theme_minimal()

## Test 1: Pearson correlation
correlation_test <- cor.test(
  data$temperature_celsius,
  data$air_quality_us.epa.index,
  method = "pearson"
  )

cat("Pearson Correlation Coefficient:", correlation_test$estimate, "\n")
cat("p-value:", correlation_test$p.value, "\n")

if (correlation_test$p.value < 0.05) {
  cat("As there exists a statistical significance between temperature and air_quality_us-epa-
index,\nthese variables are correlated.\n")
} else {
```

```
  cat("As there is no statistical significance between temperature and air_quality_us-epa-
index,\nthese variables are not correlated.\n")
}

## Test 2: Chi-Square Test of Independence
### Data Preparation
data$temperature_category <- cut(
  data$temperature_celsius,
  breaks = 3,
  labels = c("Low", "Medium", "High")
) #categorizing the "temperature_celsius" column

data$pm_category <- cut(
  data$air_quality_us.epa.index,
  breaks = 3,
  labels = c("Low", "Medium", "High")
) #categorizing the "air_quality_us-epa-index" column

contingency_table <- table(data$temperature_category, data$pm_category)
print(contingency_table)

### Performing the Chi-Square Test
chi_square_test <- chisq.test(contingency_table)

cat("Chi-Square Test Statistic:", chi_square_test$statistic, "\n")
cat("p-value:", chi_square_test$p.value, "\n")

if (chi_square_test$p.value < 0.05) {
  cat("As there exists a statistical significance between temperature categories and
air_quality_us-epa-index,\nthese variables are dependent (i.e., correlated).\n")
} else {
  cat("As there is no statistical significance between temperature categories and
air_quality_us-epa-index,\nthese variables are independent (i.e., not-correlated).\n")
}

# Performing ADF test on temperature_celsius (to check for stationarity in the time series
data)
adf_result_temp <- adf.test(data$temperature_celsius, alternative = "stationary")
print(adf_result_temp)

# Performing ADF test on humidity (for stationarity)
adf_result_humidity <- adf.test(data$humidity, alternative = "stationary")
print(adf_result_humidity)

# Selecting variables for pair plot
weather_vars <- data %>% select(temperature_celsius, humidity, wind_mph, pressure_mb,
visibility_km)
ggpairs(weather_vars)

# Analyzing summary statistics for key weather variables
summary(data$temperature_celsius)
summary(data$humidity)
summary(data$wind_mph)
summary(data$visibility_km)
```

```
# Creating a heatmap for air quality parameters
air_quality_data <- data %>% select(air_quality_Carbon_Monoxide, air_quality_Ozone,
air_quality_Nitrogen_dioxide, air_quality_Sulphur_dioxide, air_quality_PM2.5,
air_quality_us.epa.index)
correlation_air_quality <- cor(air_quality_data, use = "complete.obs")

# Creating heatmap for air quality correlations
pheatmap(correlation_air_quality, cluster_rows = TRUE, cluster_cols = TRUE, main = "Air
Quality Correlation Heatmap")

# Converting 'last_updated' to a POSIXct date-time format
data$last_updated <- mdy_hms(data$last_updated)  # Adjust format if necessary
data$last_updated <- as.POSIXct(data$last_updated, format="%m/%d/%Y %H:%M",
tz="UTC")

# Converting 'temperature_celsius' to a time series (daily data assumption)
temperature_ts <- ts(data$temperature_celsius, frequency=1, start=c(2024, 1))
print(data$temperature_celsius)

# Fitting ARIMA model to the time series data
model <- auto.arima(temperature_ts)

# Forecasting the next 7 days
forecasted_values <- forecast(model, h=365)

# Plotting the forecasted values
autoplot(forecasted_values) +
  theme_minimal() +
  theme(plot.margin = margin(10, 10, 10, 10)) +
  ggtitle("Temperature Forecast") +
  theme(plot.title = element_text(size = 10))
```
