# Semesterarbeit - Statistische Datenanalyse

Analyse von Einflussfaktoren im Strommarkt Schweiz

Stefan Bieri

20.06.2025

## Contents

# 1 Explorative Datenanalyse

## 1.1 Theoretische Fundierung

Im Rahmen dieser Arbeit sollen mögliche Zusammenhänge zwischen dem Strompreis und unterschiedlichen Einflussgrössen wie Temperatur, Windgeschwindigkeit, Energieverbrauch sowie verschiedenen erneuerbaren Energieträgern untersucht werden. Dazu werden grafische Methoden (Histogramme, Streudiagramme, Boxplots und QQ-Plots) sowie statistische Kennzahlen verwendet, um Verteilungseigenschaften und Abhängigkeiten in den Daten zu untersuchen.

Zur Quantifizierung möglicher linearer Zusammenhänge zwischen dem Strompreis und den anderen Variablen wird ergänzend eine lineare Regressionsanalyse durchgeführt. Diese Analyse verlangt spezifische Voraussetzungen wie Normalverteilung der Residuen, Linearität der Zusammenhänge sowie Varianzhomogenität (Homoskedastizität). Zur Prüfung dieser Annahmen werden die erforderlichen Analysen durchgeführt.Die theoretische Fundierung dieser Analyse basiert auf den Ausführungen bei Andreas Handl und Torben Kuhlenkasper, Multivariate Analysemethoden, 3. Aufl. 2017.

Aufgrund des umfangreichen Platzbedarfs und der besseren Lesbarkeit wurden die Visualisierungen im Anhang dargestellt. In den folgenden Kapiteln erfolgt eine ausführliche Beschreibung sowie Interpretation der entsprechenden Daten.

## 1.2 Beschreibung der Ausgangsvariablen

Die Analyse basiert auf einem Datensatz, der meteorologische und energiewirtschaftliche Kennzahlen im Zusammenhang mit dem Strompreis umfasst. Alle Grössen sind auf die Schweiz bezogen. Die abhängige Variable „Strompreis“ gibt den Preis pro Megawattstunde (CHF/MWh) an und bildet die zentrale Zielgrösse der Untersuchung. Als erklärende Einflussgrössen dienen:

Temperatur [°C]: Repräsentiert die Umgebungstemperatur (Luzern) zum jeweiligen Messzeitpunkt und steht in engem Zusammenhang mit der Heiz- und Kühlnachfrage.

Windgeschwindigkeit [m/s]: Gibt die durchschnittliche Windgeschwindigkeit (Schweiz) an und beeinflusst massgeblich das Potenzial zur Stromproduktion aus Windkraft.

Windenergieerzeugung [MW]: Misst die tatsächlich ins Netz eingespeiste Windenergieleistung in Megawatt.

PV-Leistung [MW]: Beschreibt die eingespeiste elektrische Leistung aus Photovoltaikanlagen.

Erneuerbare Energien [MW]: Umfasst die summierte Einspeisung aus allen erneuerbaren Energiequellen (u.a. Wind, PV, Wasser).

Netzbezogene Leistung [MW]: Gibt den Stromverbrauch bzw. die Netznachfrage an und bildet die aggregierte Last auf der Verbraucherseite ab.

## 1.3 Daten vorbereiten

```r
file_path <- paste0(
  "C:/Users/41788/OneDrive/CAS Statische Datenalyse & ",
  "Datenvisualisierung/Skript R/01 FFHS Semesterarbeit ",
  "Statistische Datenanalyse/Data.xlsx"
)
data <- read_excel(file_path, sheet = "alle", na = "", trim_ws = TRUE)
```

Spaltennamen bereinigen

```r
colnames(data) <- gsub(" ", "_", colnames(data))
colnames(data) <- gsub("[^[:alnum:]_]", "", colnames(data))
```

Umwandlung in numerische Werte

```r
data$Strompreis <- as.numeric(gsub("[^0-9.-]", "", data$Strompreis))
data$Temperatur <- as.numeric(gsub("[^0-9.-]", "", data$Temperatur))
data$Windgeschwindigkeit <- as.numeric(gsub("[^0-9.-]", "", data$Windgeschwindigkeit))
data$Windenergieerzeugung <- as.numeric(gsub("[^0-9.-]", "", data$Windenergieerzeugung))
data$PV_Leistung <- as.numeric(gsub("[^0-9.-]", "", data$PV_Leistung))
data$Erneuerbare_Energien <- as.numeric(gsub("[^0-9.-]", "", data$Erneuerbare_Energien))
data$Netzbezogene_Leistung <- as.numeric(gsub("[^0-9.-]", "", data$Netzbezogene_Leistung))
```

Fehlende Werte entfernen

```r
data <- na.omit(data)
```

Numerische Daten filtern

```r
numeric_data <- data[, sapply(data, is.numeric)]
numeric_data <- numeric_data[, !colnames(numeric_data) %in% c("Datum_MEZ")]
```

Zeitspalte umwandeln und auf Stunde filtern (nur 08–16 Uhr)

```r
data$Datum <- as.POSIXct(data$Datum_MEZ, format = "%d.%m.%y %H:%M")
data$Stunde <- lubridate::hour(data$Datum)
data_f <- dplyr::filter(data, Stunde >= 8 & Stunde <= 16)
```

## 1.4 Korrelationsmatrix

```r
colnames(numeric_data) <- c("Preis", "Temp", "Wind", "Wind_Erz", "PV", "EE", "Netz")
cor_matrix <- cor(numeric_data, use = "complete.obs")
knitr::kable(cor_matrix, digits = 2, caption = 'Korrelationsmatrix der numerischen Variablen') %>%
  kable_styling(latex_options = c("scale_down"))
```

Table 1: Korrelationsmatrix der numerischen Variablen

| | Preis | Temp | Wind | Wind_Erz | PV | EE | Netz |
|---|---|---|---|---|---|---|---|
| Preis | 1.00 | -0.49 | -0.07 | 0.05 | -0.47 | -0.05 | 0.22 |
| Temp | -0.49 | 1.00 | -0.09 | -0.24 | 0.47 | 0.56 | -0.46 |
| Wind | -0.07 | -0.09 | 1.00 | 0.21 | 0.24 | 0.01 | 0.32 |
| Wind_Erz | 0.05 | -0.24 | 0.21 | 1.00 | -0.17 | -0.25 | 0.21 |
| PV | -0.47 | 0.47 | 0.24 | -0.17 | 1.00 | 0.39 | 0.14 |
| EE | -0.05 | 0.56 | 0.01 | -0.25 | 0.39 | 1.00 | -0.05 |
| Netz | 0.22 | -0.46 | 0.32 | 0.21 | 0.14 | -0.05 | 1.00 |

- Preis und Temperatur ist negativ korreliert. Dies ist in den Daten gut ersichtlich.
- Preis und Wind ist nur leicht negativ korreliert. Dies ist aber auch in den Daten zu erkennen.
- Preis und Winderzeugung ist die korrelation sehr klein und mit dem Auge nicht zu erkennen.
- Preis und PV-Erzeugung ist die klare negative Korrelation gut ersichtlich.
- Preis und Erneuerbare Energien ist die Korrelation sehr klein.
- Preis und Netz(Verbrauch) sieht man die positive Korrelation gut.

## 1.5 Statistische Auswertung

```r
summary(data)
```

```
##    Datum_MEZ                        Strompreis       Temperatur
##  Min.   :2024-01-01 00:00:00.00   Min.   :-416.69   Min.   :-4.99
##  1st Qu.:2024-04-01 08:26:15.00   1st Qu.:  53.23   1st Qu.: 7.16
##  Median :2024-07-01 16:52:30.00   Median :  72.67   Median :12.06
##  Mean   :2024-07-01 21:06:01.00   Mean   :  72.26   Mean   :12.76
##  3rd Qu.:2024-10-01 09:18:45.00   3rd Qu.:  92.23   3rd Qu.:18.38
##  Max.   :2024-12-31 23:45:00.00   Max.   : 291.54   Max.   :34.01
##  Windgeschwindigkeit Windenergieerzeugung  PV_Leistung     Erneuerbare_Energien
##  Min.   :0.2000      Min.   : 0.00        Min.   :   0.1   Min.   : 1113
##  1st Qu.:0.5200      1st Qu.: 5.60        1st Qu.:   0.8   1st Qu.: 3778
##  Median :0.6800      Median :12.60        Median :   6.3   Median : 5620
##  Mean   :0.7437      Mean   :18.35        Mean   : 666.1   Mean   : 5534
##  3rd Qu.:0.9200      3rd Qu.:27.20        3rd Qu.: 985.4   3rd Qu.: 7081
##  Max.   :2.3400      Max.   :78.90        Max.   :4626.8   Max.   :12510
##  Netzbezogene_Leistung     Datum                            Stunde
##  Min.   : 2743         Min.   :2024-01-01 00:00:00.00   Min.   : 0.0
##  1st Qu.: 6021         1st Qu.:2024-04-01 08:26:15.00   1st Qu.: 6.0
##  Median : 6813         Median :2024-07-01 16:52:30.00   Median :12.0
##  Mean   : 6788         Mean   :2024-07-01 21:06:01.00   Mean   :11.5
##  3rd Qu.: 7564         3rd Qu.:2024-10-01 09:18:45.00   3rd Qu.:18.0
##  Max.   :10437         Max.   :2024-12-31 23:45:00.00   Max.   :23.0
```

Die Auswertung zeigt:

- Strompreis: Median und Mittelwert sind nahe beieinander → symmetrische Verteilung
- Temperatur: Plausibel für das Schweizer Mittelland
- Windgeschwindigkeit: Eher tief, aufgrund der 15 min - Mittelwerte. Angaben sind in m/s.
- PV-Leistung: Extreme Unterschiede zwischen Median und Mittelwert
- Preis und Temperatur ist negativ korreliert. Dies ist in den Daten gut ersichtlich.
- Preis und Wind ist nur leicht negativ korreliert. Dies ist aber auch in den Daten zu erkennen.
- Preis und Winderzeugung ist die korrelation sehr klein und mit dem Auge nicht zu erkennen.
- Preis und PV-Erzeugung ist die klare negative Korrelation gut ersichtlich.
- Preis und Erneuerbare Energien ist die Korrelation sehr klein.
- Preis und Netz(Verbrauch) sieht man die positive Korrelation gut.

## 1.6 Streudiagramm-Matrix

```r
output_dir <- paste0(
  "C:/Users/41788/OneDrive/CAS Statische Datenalyse & Datenvisualisierung/",
  "Skript R/01 FFHS Semesterarbeit Statistische Datenanalyse"
)
pdf(file.path(output_dir, "1.Teil Streudiagramm_Matrix.pdf"), width = 12, height = 20)
ggpairs(numeric_data, title = "Streudiagramm Matrix mit Korrelation")
dev.off()
```

```
## pdf
##   2
```

Siehe PDF Streudiagramm_Matrix im Anhang:

- Variablen Temperatur und Netzbezogene Leistung sehen annäherend Normalverteilt aus.
- Beim Strompreis sind Extremwerte häufiger als bei der Normalverteilung zu erwarten wäre (heavy tailed), die Werte sind zudem "linksschief".
- Windgeschwindigkeit und somit auch Windenergie sind rechtsschief. Durchschnitt ist höher als Median.
- Korrelation mit Strompreis ist im Kapitel Korrelationsmatrix zu entnehmen.

## 1.7 Prüfung der Residuen und Modelvoraussetzung

```r
model <- lm(Strompreis ~ Temperatur + Windgeschwindigkeit +
              Windenergieerzeugung + PV_Leistung +
              Erneuerbare_Energien + Netzbezogene_Leistung,
            data = data)
pdf(file.path(output_dir, "1.Teil Modellvoraussetzungs_Plots.pdf"))
par(mfrow = c(2, 2))
plot(model)
dev.off()
```

```
## pdf
##   2
```

Siehe PDF Modellvoraussetzungs_Plots im Anhang.

Residuals vs Fitted Plot: - Die Residuen weisen im Allgemeinen eine horizonalte Verteilung auf, allerdings zeigt sich mit zunehmenden Werten (fitted values) eine leicht systematische Abnahme der Varianzhomogenität.
- Es gibt keine starken Ausreisser oder Klumpen, die auf Modellierungsfehler oder Gruppenstrukturen hinweisen würden.

Scale-Location Plot: - Leichtes Muster (keine horizontale Linie) ersichtlich. Abweichung nimmt zu bei zunehmenden Werten (fitted values), wahrscheinlich durch nicht lineare Zusammenhänge. - Die Residuen sind standardisiert, was den Vergleich erleichtert. Es sind keine extremen Abweichungen sichtbar.

Q-Q Plot: - Abweichung an beiden Enden. Wie bereits im Streudiagramm erkannt zeigen die Residuen eine "heavy-tailed" Verteilung.

Residuals vs Leverage: - Keine Werte ausserhalb der Cook's Distance und somit keinen aussergewöhnliche starken Einfluss auf die Schätzung der Regression. - Die zentrale Verteilung ist relativ gut approximiert, die Abweichungen betreffen primär die Extrembereiche (Tail-Verhalten).

## 1.8 Histogramme

```
pdf(file.path(output_dir, "1. Teil Histogramme.pdf"), width = 10, height = 8)
par(mfrow = c(3, 2))

for (i in 1:ncol(numeric_data)) {
  x <- numeric_data[[i]]

  # Histogramm mit Dichte-Skalierung
  hist(x,
       main = paste("Histogramm von", colnames(numeric_data)[i]),
       col = "lightblue",
       probability = TRUE,
       xlab = "",
       ylab = "Dichte")

  # Normalverteilungslinie
  x_vals <- seq(min(x, na.rm = TRUE), max(x, na.rm = TRUE), length.out = 100)
  y_vals <- dnorm(x_vals, mean = mean(x, na.rm = TRUE), sd = sd(x, na.rm = TRUE))
  lines(x_vals, y_vals, col = "red", lwd = 2)
}

par(mfrow = c(1, 1))
dev.off()
```

```
## pdf
##   2
```

Siehe PDF Histogramme im Anhang.

Wie bereits beim Streudiagramme aufgeführt:

- Variablen Temperatur und Netzbezogene Leistung sehen annähernd Normalverteilt aus.
- Beim Strompreis sind die Anzahl Werte um den Durchschnittswerte viel häufiger, als bei einer Normalverteilung zu erwarten wäre (heavy tailed).
- Windgeschwindigkeit und somit auch Windenergie sind rechtsschief. Durchschnitt ist höher als Median.
- Verteilung der Werte PV-Leistung ist sehr stark konzentriert und extrem rechtsschief.

## 1.9 QQ-Plots

```
pdf(file.path(output_dir, "1. Teil QQ_Plots.pdf"), width = 10, height = 8)
par(mfrow = c(3, 2))
for (i in 1:ncol(numeric_data)) {
  qqnorm(numeric_data[[i]], main = paste("QQ-Plot von", colnames(numeric_data)[i]))
  qqline(numeric_data[[i]], col = "red")
}
par(mfrow = c(1, 1))
dev.off()
```

```
## pdf
##   2
```

Siehe PDF QQ-Plots im Anhang.

- Strompreis: Abweichung im unteren und oberen Bereich.
- Temperatur: Abweichung im unteren und oberen Bereich.
- Windgeschwindigkeit: Rechtsschief
- Winderzeugung:S-Kurve deutet auf nicht linearen Zusammenhang hin.
- PV_Leistung: Wie bereits erwC$hnt. Werte sind extrem rechtsschief und nciht normalverteilt.
- Eneruerbare_Energien: Abweichung im unteren und oberen Bereich.
- Netzbezogene_Leistung: AnnC$herend Normalverteilt.

## 1.10 Boxplots

```
pdf(file.path(output_dir, "1.Teil Boxplots.pdf"), width = 10, height = 8)
par(mfrow = c(3, 2))
for (i in 1:ncol(numeric_data)) {
  boxplot(numeric_data[[i]], main = colnames(numeric_data)[i], col = "lightblue")
}
par(mfrow = c(1, 1))
dev.off()
```

```
## pdf
##   2
```

Siehe PDF Boxplots im Anhang.

- Zeigen die selben Merkmale wie bereits genannt.

## 1.11 Interpretation der Ergebnisse

Die durchgeführte explorative Datenanalyse zeigt, dass der Strompreis mit den untersuchten Variablen Temperatur, Windgeschwindigkeit, Energieverbrauch und erneuerbaren Energieträgern lineare, jedoch nicht sehr starke Zusammenhänge aufweist.

Temperatur und netzbezogene Leistung weisen dabei annähernd normale Verteilungen auf. Allerdings zeigen Windgeschwindigkeit und PV-Leistung eine stark rechtsschiefe Verteilung.

Die lineare Regression bestätigt diese Beobachtungen und zeigt leichte Verletzungen der Annahmen bzgl. Homoskedastizität und Normalverteilung der Residuen.

→ Für einzelne Variablen könnten Transformationen oder komplexere Modelle sinnvoll sein.

# 2 Multiple lineare Regression zur Strompreisvorhersage

## 2.1 Theoretische Fundierung

Die theoretische Fundierung dieser Analyse basiert auf den Ausführungen bei Andreas Handl und Torben Kuhlenkasper, Multivariate Analysemethoden, 3. Aufl. 2017. Die multiple lineare Regression ist ein statistisches Verfahren zur Analyse des Einflusses mehrerer unabhängiger Variablen auf eine metrische Zielgrösse. In der vorliegenden Untersuchung dient der Strompreis in CHF/MWh als abhängige Variable. Ziel ist es, zu quantifizieren, in welchem Ausmass verschiedene Einflussgrössen aus den Bereichen Wetter und erneuerbarer Energieerzeugung mit dem Strompreis zusammenhängen.

Die Methode beruht auf dem Prinzip der kleinsten Quadrate: Es wird jene Regressionsgerade geschätzt, bei der die Abweichungen zwischen den beobachteten und den durch das Modell prognostizierten Werten minimiert werden. Die geschätzten Regressionskoeffizienten zeigen, wie stark und in welche Richtung eine unabhängige Variable den Strompreis beeinflusst – unter Konstanthaltung aller anderen Einflussgrössen.

Insbesondere im Kontext eines liberalisierten Strommarkts ist der Strompreis starken kurzfristigen Schwankungen unterworfen. Diese ergeben sich aus einem komplexen Zusammenspiel von Angebot und Nachfrage, meteorologischen Bedingungen sowie technischen Rahmenbedingungen des Stromnetzes. Mittels der multiplen linearen Regression sollen diese Zusammenhänge auf quantitativer Ebene modelliert werden. Dabei wird auch geprüft, ob einzelne Einflussfaktoren statistisch signifikant zur Erklärung der Strompreisentwicklung beitragen.

Zur Überprüfung möglicher Multikollinearitäten zwischen den unabhängigen Variablen werden zusätzlich Varianzinflationsfaktoren (VIF) berechnet.

Das übergeordnete Ziel dieser Analyse besteht darin, zentrale Einflussfaktoren der Strompreisbildung zu identifizieren und ihre Wirkungszusammenhänge datenbasiert abzuleiten.

## 2.2 Daten vorbereiten

```r
file_path <- (
  "C:/Users/41788/OneDrive/CAS Statische Datenalyse & Datenvisualisierung/" %>%
  paste0("Skript R/01 FFHS Semesterarbeit Statistische Datenanalyse/Data.xlsx")
)
data <- read_excel(file_path, sheet = "alle")
```

Relevante Spalten auswählen

```r
data <- data[, c("Strompreis", "Temperatur", "Windgeschwindigkeit",
                 "Windenergieerzeugung", "PV_Leistung",
                 "Erneuerbare_Energien", "Netzbezogene_Leistung")]
```

Sicherstellen, dass alle Werte numerisch sind

```r
data <- data.frame(lapply(data, as.numeric))
```

Fehlende Werte entfernen

```r
data <- na.omit(data)
```

## 2.3 Multiple lineare Regression

```r
model_lm <- lm(Strompreis ~ Temperatur + Windgeschwindigkeit + Windenergieerzeugung +
                 PV_Leistung + Erneuerbare_Energien + Netzbezogene_Leistung, data = data)
```

```r
summary_model1 <- summary(model_lm)

# Tabelle mit Regressionskoeffizienten
knitr::kable(as.data.frame(summary_model1$coefficients), caption = "Regressionskoeffizienten (summary_mo
```

Table 2: Regressionskoeffizienten (summary_model1)

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 62.0182885 | 1.3195355 | 47.000093 | 0 |
| Temperatur | -2.3577617 | 0.0313462 | -75.216828 | 0 |
| Windgeschwindigkeit | -5.3847481 | 0.5496078 | -9.797437 | 0 |
| Windenergieerzeugung | -0.1189898 | 0.0097999 | -12.141893 | 0 |
| PV_Leistung | -0.0133994 | 0.0001799 | -74.478545 | 0 |
| Erneuerbare_Energien | 0.0065504 | 0.0000894 | 73.232581 | 0 |
| Netzbezogene_Leistung | 0.0028097 | 0.0001759 | 15.970190 | 0 |

```r
# R², Adjusted R² und p-Wert der Gesamt-F-Statistik anzeigen
r2 <- round(summary_model1$r.squared, 3)
adj_r2 <- round(summary_model1$adj.r.squared, 3)

# F-Statistik und p-Wert berechnen
fstat <- summary_model1$fstatistic
f_pval <- pf(fstat[1], fstat[2], fstat[3], lower.tail = FALSE)

# Ausgabe als formatierten Text
cat("\n\n**Bestimmtheitsmaß (R²):**", r2,
    "\n\n**Adjustiertes R²:**", adj_r2,
    "\n\n**Gesamtmodell:** F(", fstat[2], ", ", fstat[3], ") = ", round(fstat[1], 2),
    ", p ", ifelse(f_pval < 0.001, "< 0.001", paste0("= ", round(f_pval, 3))), "\n")
```

```
##
##
## **Bestimmtheitsmaß (R²):** 0.432
##
## **Adjustiertes R²:** 0.432
##
## **Gesamtmodell:** F( 6 ,  35129 ) =  4446.41 , p  < 0.001
```

Interpretation der Regressionskoeffizienten

(Intercept): 62.02 CHF/MWh – geschätzter Strompreis, wenn alle Einflussgrössen 0 wären. In der Praxis ohne inhaltliche Bedeutung (nur mathematischer Referenzpunkt).

Temperatur [°C]: Für jede Erhöhung um 1 °C sinkt der Strompreis um ca. 2.36 CHF. → Höhere Temperaturen senken tendenziell die Strompreise. Temperatur ist wichtige Einflussgrösse auf Stromnachfrage im Herbst und Frühling.

Windgeschwindigkeit [m/s]: Eine Zunahme um 1 m/s senkt den Strompreis im Schnitt um 5.39 CHF. → Mehr Wind → günstigerer Strom.

Windenergieerzeugung [MW]: Je 1 MW mehr Windstrom führt zu einem Rückgang des Strompreises um 0.119 CHF.

PV_Leistung [MW]: Pro 1 MW zusätzlicher Photovoltaik-Leistung sinkt der Strompreis um ca. 0.0134 CHF.

Erneuerbare_Energien [MW]: Unerwartet – pro 1 MW mehr Gesamtleistung aus erneuerbaren Quellen steigt der Strompreis um etwa 0.00655 CHF. → Möglicherweise durch indirekte Marktmechanismen oder Netzbelastung erklärbar.

Netzbezogene_Leistung [MW]: Pro zusätzlichem MW Netzlast steigt der Strompreis um ca. 0.00281 CHF.

Alle Effekte sind hochsignifikant ($p < 0.001$), also statistisch sehr zuverlässig. $R^2 = 0.43$ → Das Modell erklärt rund 43 % der Schwankungen im Strompreis.

## 2.4 VIF-Berechnung

```r
vif_values <- vif(model_lm)
knitr::kable(as.data.frame(vif_values), caption = "VIF-Werte")
```

Table 3: VIF-Werte

| | vif_values |
|---|---|
| Temperatur | 2.683137 |
| Windgeschwindigkeit | 1.223244 |
| Windenergieerzeugung | 1.180089 |
| PV_Leistung | 1.741778 |
| Erneuerbare_Energien | 1.647279 |
| Netzbezogene_Leistung | 1.809081 |

Alle Werte liegen deutlich unter 5 → Es besteht keine problematische Multikollinearität im Modell. Die erklärenden Variablen liefern weitgehend unabhängige Informationen.

## 2.5 Reduziertes Modell (ohne Windenergieerzeugung)

```r
model_lm_reduced <- lm(Strompreis ~ Temperatur + Windgeschwindigkeit +
                         PV_Leistung + Erneuerbare_Energien + Netzbezogene_Leistung,
                       data = data)
```

## 2.6 F-Test zum Modellvergleich

```r
f_test_result <- anova(model_lm_reduced, model_lm)
print("F-Test Ergebnis zum Modellvergleich:")
```

```
## [1] "F-Test Ergebnis zum Modellvergleich:"
```

```r
print(f_test_result)
```

```
## Analysis of Variance Table
##
## Model 1: Strompreis ~ Temperatur + Windgeschwindigkeit + PV_Leistung +
##     Erneuerbare_Energien + Netzbezogene_Leistung
## Model 2: Strompreis ~ Temperatur + Windgeschwindigkeit + Windenergieerzeugung +
##     PV_Leistung + Erneuerbare_Energien + Netzbezogene_Leistung
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1  35130 26667717
## 2  35129 26556269  1    111448 147.43 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Der F-Wert ist sehr hoch (147.43) → Das deutet auf eine deutliche Verbesserung hin. Der p-Wert ist extrem klein ($< 0.001$) → Statistisch hochsignifikant. Die Hinzunahme von „Windenergieerzeugung“ verbessert das Modell signifikant.

## 2.7 Interpretation der Ergebnisse

Das lineare Regressionsmodell erklärt rund 43 % der Varianz des Strompreises ($R^2 = 0.4316$).

Alle im Modell enthaltenen Prädiktoren zeigen hochsignifikante Effekte ($p < 0.05$), wobei Temperatur, Windgeschwindigkeit sowie die Einspeisung aus Photovoltaik und Wind erwartungsgemäss negativ mit dem Strompreis zusammenhängen – also preisreduzierend wirken.

Auffällig ist der positive Effekt der Gesamtleistung aus erneuerbaren Energien, was auf indirekte Marktmechanismen oder systemische Zusammenhänge wie Prognoseabweichungen hindeuten könnte.

Die Variablen weisen keine problematische Multikollinearität auf (alle VIF-Werte deutlich $< 5$).

Ein F-Test zum Vergleich eines reduzierten Modells (ohne „Windenergieerzeugung“) mit dem vollständigen Modell ergibt einen signifikanten Unterschied ($F = 147.43$, $p < 2.2e-16$). Die Variable „Windenergieerzeugung“ trägt somit signifikant zur Modellgüte bei und sollte im finalen Modell enthalten bleiben.

Für eine weitere Verbesserung des Modells könnten zusätzliche Datenquellen berücksichtigt werden. Dazu zählen insbesondere: - Brennstoffpreise (Gas, Kohle), da sie den Börsenpreis stark beeinflussen (Merit-Order) - Import-/Exportmengen im Stromnetz (grenzüberschreitender Handel) - Prognoseabweichungen (Differenz zwischen geplanter und tatsächlicher Einspeisung) - Tageszeit, Wochentag oder saisonale Effekte

# 3 Logistische Regression zur Vorhersage extremer Strompreise

## 3.1 Theoretische Fundierung

Die theoretische Fundierung dieser Analyse basiert auf den Ausführungen bei Andreas Handl und Torben Kuhlenkasper, Multivariate Analysemethoden, 3. Aufl. 2017. Die logistische Regression ist ein statistisches Verfahren zur Modellierung binärer Zielgrössen. In dieser Analyse wird sie verwendet, um die Wahrscheinlichkeit für das Auftreten hoher oder tiefer Strompreise zu schätzen. Hohe Strompreise sind dabei als Werte oberhalb des 90 %-Quantils, tiefe Strompreise als Werte unterhalb des 10%-Quantils definiert.

Im Gegensatz zur linearen Regression basiert die logistische Regression auf einer nicht-linearen Verknüpfung zwischen Prädiktoren und Zielvariable. Anstelle eines stetigen Zielwerts wird hier die logit-transformierte Wahrscheinlichkeit modelliert, also der natürliche Logarithmus der Odds (Wahrscheinlichkeit / Gegenwahrscheinlichkeit). Dadurch wird sichergestellt, dass die vorhergesagten Werte stets im Intervall zwischen 0 und 1 liegen und somit als Wahrscheinlichkeiten interpretierbar sind.

Als erklärende Variablen dienen verschiedene Einflussgrössen aus den Bereichen Wetter und Energieerzeugung. Dazu zählen Temperatur, Windgeschwindigkeit, Windenergieerzeugung, PV-Leistung, die Gesamtleistung erneuerbarer Energien sowie die netzbezogene Leistung.

Die binäre Zielvariable ermöglicht den Einsatz der logistischen Regression, um Zusammenhänge zwischen den Einflussgrössen und dem Auftreten extremer Strompreise zu modellieren. Zur Beurteilung der Modellgüte werden unter anderem der AIC-Wert (Mass für Modellkomplexität und Anpassung) sowie die AUC der ROC-Kurve (Mass für die Klassifikationsgüte) herangezogen.

## 3.2 Daten vorbereiten

Datei einlesen

```r
file_path <- (
  "C:/Users/41788/OneDrive/CAS Statische Datenalyse & Datenvisualisierung/" %>%
  paste0("Skript R/01 FFHS Semesterarbeit Statistische Datenanalyse/Data.xlsx")
)
data <- read_excel(file_path, sheet = "alle")
```

Relevante Spalten auswählen

```r
data <- data[, c("Strompreis", "Temperatur", "Windgeschwindigkeit",
                 "Windenergieerzeugung", "PV_Leistung",
                 "Erneuerbare_Energien", "Netzbezogene_Leistung")]
```

Sicherstellen, dass alle Werte numerisch sind

```r
data <- as.data.frame(sapply(data, function(x) as.numeric(as.character(x))))
```

Fehlende Werte entfernen

```r
data <- na.omit(data)
```

### 3.3 90 %-Quantil berechnen und binäre Variable erstellen

```r
quantil_90 <- quantile(data$Strompreis, 0.90, na.rm = TRUE)
print(paste("Das 90%-Quantil des Strompreises beträgt:", round(quantil_90, 2)))
```

```
## [1] "Das 90%-Quantil des Strompreises beträgt: 115.15"
```

```r
data$Hoher_Strompreis <- ifelse(data$Strompreis > quantil_90, 1, 0)
```

### 3.4 Logistische Regression mit allen Variablen

```r
model_logit <- glm(Hoher_Strompreis ~ Temperatur + Windgeschwindigkeit +
                     Windenergieerzeugung + PV_Leistung +
                     Erneuerbare_Energien + Netzbezogene_Leistung,
                   data = data,
                   family = binomial)
```

### 3.5 AIC-Wert ausgeben

```r
aic_value <- AIC(model_logit)
print(paste("AIC-Wert des Modells:", round(aic_value, 2)))
```

```
## [1] "AIC-Wert des Modells: 16931.16"
```

### 3.6 Ergebnisse des Modells anzeigen

```r
summary_model <- summary(model_logit)
print(summary_model)
```

```
##
## Call:
## glm(formula = Hoher_Strompreis ~ Temperatur + Windgeschwindigkeit +
##     Windenergieerzeugung + PV_Leistung + Erneuerbare_Energien +
##     Netzbezogene_Leistung, family = binomial, data = data)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -5.038e+00  1.840e-01 -27.380   <2e-16 ***
## Temperatur            -1.403e-01  4.662e-03 -30.086   <2e-16 ***
## Windgeschwindigkeit   -7.951e-01  7.566e-02 -10.508   <2e-16 ***
## Windenergieerzeugung   1.802e-03  1.258e-03   1.432    0.152
## PV_Leistung           -8.762e-04  4.474e-05 -19.586   <2e-16 ***
## Erneuerbare_Energien   5.694e-04  1.325e-05  42.957   <2e-16 ***
## Netzbezogene_Leistung  2.688e-04  2.412e-05  11.144   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22837  on 35135  degrees of freedom
## Residual deviance: 16917  on 35129  degrees of freedom
## AIC: 16931
##
## Number of Fisher Scoring iterations: 7
```

### 3.7 Beschreibung der Regressionskoeffizienten (basierend auf z-Werten):

(Intercept): Der Achsenabschnitt von –5.038 entspricht dem geschätzten Logit-Wert, wenn alle Prädiktoren den Wert 0 annehmen. Er dient rein der Modellkalibrierung und ist inhaltlich nicht interpretierbar.

Erneuerbare Energien: Mit einem z-Wert von 42.96 zeigt diese Variable den mit Abstand stärksten statistischen Einfluss im Modell. Ein Anstieg der erneuerbaren Gesamtleistung geht mit einer erhöhten Wahrscheinlichkeit für hohe Strompreise einher – ein hochsignifikanter und dominanter Effekt.

Temperatur: Der z-Wert von –30.09 weist ebenfalls auf einen sehr starken Einfluss hin. Mit steigender Temperatur sinkt die Wahrscheinlichkeit hoher Strompreise signifikant. Temperatur ist damit ein zentraler erklärender Faktor.

PV-Leistung: Der Effekt ist mit einem z-Wert von –19.59 hochsignifikant. Eine höhere PV-Leistung geht mit sinkender Wahrscheinlichkeit hoher Preise einher. Der Einfluss ist klar nachweisbar, auch wenn der Koeffizient klein ist.

Netzbezogene Leistung: Der z-Wert von 11.14 zeigt einen signifikanten, aber im Vergleich schwächeren Effekt. Eine höhere Netzlast erhöht die Wahrscheinlichkeit hoher Strompreise leicht, aber konsistent.

Windgeschwindigkeit: Mit einem z-Wert von –10.51 ist auch dieser Effekt statistisch sehr deutlich. Mehr Wind senkt die Wahrscheinlichkeit hoher Strompreise, was auf den preisdämpfenden Einfluss der Windkraft hindeutet.

Windenergieerzeugung: Der z-Wert von 1.43 liegt deutlich unter der Signifikanzschwelle ($p = 0.152$). Es gibt keinen belastbaren Hinweis, dass die tatsächliche Einspeisung von Windenergie einen eigenständigen Einfluss auf hohe Strompreise hat. Eine höhere netzbezogene Leistung geht mit einer leicht erhöhten Wahrscheinlichkeit hoher Preise einher.

### 3.8 AIC-Wert ohne Windenergieerzeugung

```r
model_logit_reduced <- glm(Hoher_Strompreis ~ Temperatur + Windgeschwindigkeit +
                             PV_Leistung + Erneuerbare_Energien +
                             Netzbezogene_Leistung,
                           data = data,
                           family = binomial)

aic_value_reduced <- AIC(model_logit_reduced)

print(paste("AIC-Wert mit Windenergieerzeugung:", round(aic_value, 2)))
```

```
## [1] "AIC-Wert mit Windenergieerzeugung: 16931.16"
```

```r
print(paste("AIC-Wert ohne Windenergieerzeugung:", round(aic_value_reduced, 2)))
```

```
## [1] "AIC-Wert ohne Windenergieerzeugung: 16931.2"
```

Der Unterschied zwischen den beiden AIC-Werten beträgt nur 0.04 – also praktisch vernachlässigbar.
Das bedeutet: Die Variable "Windenergieerzeugung" hat keinen relevanten Einfluss auf die Modellgüte.

### 3.9 ROC-Kurve erstellen & AUC berechnen

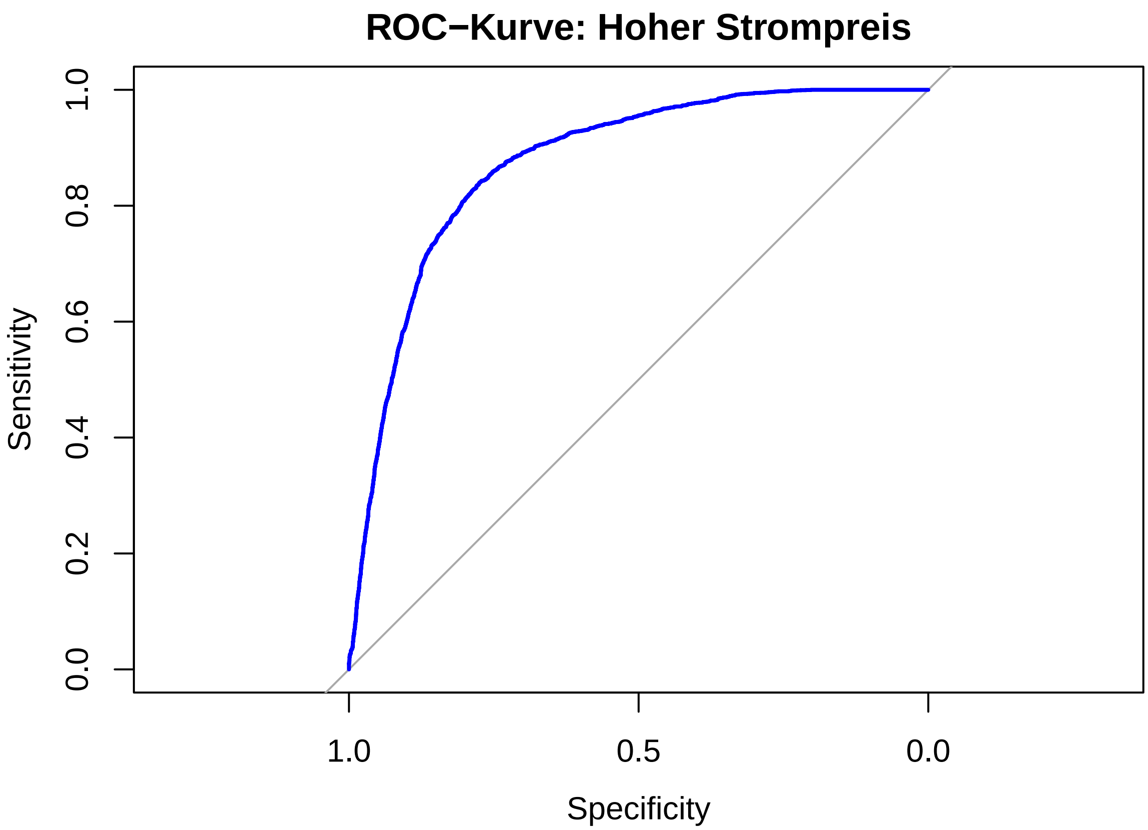
```r
predicted_probs <- predict(model_logit, type = "response")
roc_curve <- roc(data$Hoher_Strompreis, predicted_probs, levels = c(0, 1))
auc_value <- auc(roc_curve)
print(paste("AUC-Wert der ROC-Kurve:", round(auc_value, 2)))
```

```
## [1] "AUC-Wert der ROC-Kurve: 0.87"
```

Mit einem Wert von 0.87 zeigt das Modell eine gute Klassifikationsfähigkeit.
Es kann also zuverlässig vorhersagen, wann ein hoher Strompreis vorliegt.
Ab einem Wert von 0.9 gilt die Vorhersage als sehr gut.

### 3.10 Grafik für hohe und tiefe Strompreise

```r
plot(roc_curve, main = "ROC-Kurve: Hoher Strompreis", col = "blue", lwd = 2)
```




### 3.11 10%-Quantil berechnen und binäre Variable erstellen

```
quantil_10 <- quantile(data$Strompreis, 0.1, na.rm = TRUE)
print(paste("Das 10%-Quantil des Strompreises beträgt:", round(quantil_10, 2)))
```

```
## [1] "Das 10%-Quantil des Strompreises beträgt: 29.61"
```

```
data$Tiefer_Strompreis <- ifelse(data$Strompreis < quantil_10, 1, 0)
```

### 3.12 Logistische Regression für tiefe Strompreise

```
model_logit_low <- glm(Tiefer_Strompreis ~ Temperatur + Windgeschwindigkeit +
                         Windenergieerzeugung + PV_Leistung +
                         Erneuerbare_Energien + Netzbezogene_Leistung,
                       data = data,
                       family = binomial)
```

### 3.13 AIC-Wert für tiefe Strompreise

```r
aic_value_low <- AIC(model_logit_low)
print(paste("AIC-Wert des Modells für tiefe Strompreise:", round(aic_value_low, 2)))
```

```
## [1] "AIC-Wert des Modells für tiefe Strompreise: 15709.12"
```

Im Vergleich zum Modell für hohe Strompreise ist der AIC deutlich niedriger (16931 zu 15709). Das weist darauf hin, dass das Modell zur Vorhersage tiefer Strompreise
eine bessere Erklärungskraft aufweist – bezogen auf die gegebenen Daten.

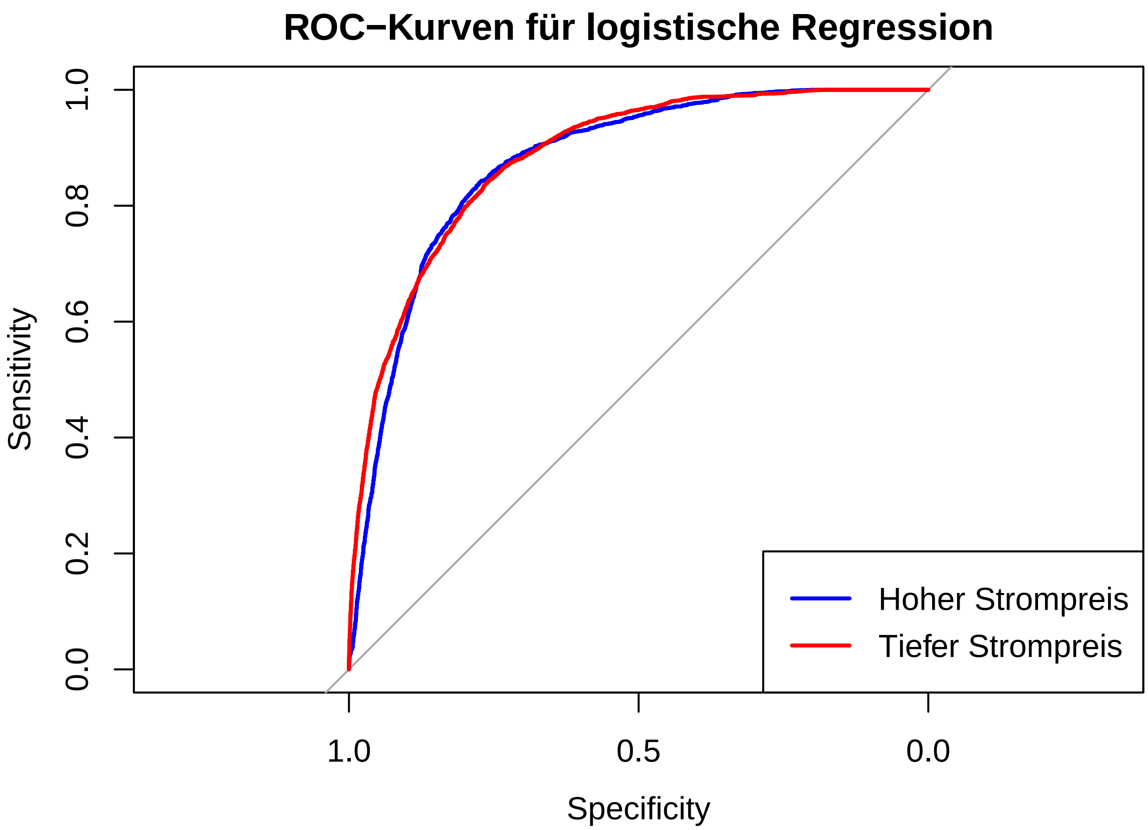
### 3.14 ROC-Kurve & AUC für tiefe Strompreise

```r
predicted_probs_low <- predict(model_logit_low, type = "response")
roc_curve_low <- roc(data$Tiefer_Strompreis, predicted_probs_low, levels = c(0, 1))
auc_value_low <- auc(roc_curve_low)
print(paste("AUC-Wert der ROC-Kurve für tiefe Strompreise:", round(auc_value_low, 2)))
```

```
## [1] "AUC-Wert der ROC-Kurve für tiefe Strompreise: 0.88"
```

### 3.15 Beide ROC-Kurven in einer Grafik

```r
plot(roc_curve, main = "ROC-Kurven für logistische Regression", col = "blue", lwd = 2)
lines(roc_curve_low, col = "red", lwd = 2)
legend("bottomright", legend = c("Hoher Strompreis", "Tiefer Strompreis"),
       col = c("blue", "red"), lwd = 2)
```




Das Modell zur Vorhersage tiefer Strompreise (rote Kurve) steigt in der Frühphase etwas schneller an → höhere Sensitivität bei hoher Spezifität, d.h. es erkennt „tiefe Preise" im kritischen Schwellenbereich etwas zuverlässiger.

Das Modell für hohe Strompreise (blaue Kurve) liegt in späteren Bereichen leicht über der roten → etwas bessere Leistung bei geringerer Spezifität.

### 3.16 Interpretation der Ergebnisse

Das Modell für hohe Strompreise zeigt gute Vorhersagekraft (AUC = 0.87).
Die Trennung zwischen hohen und nicht hohen Preisen gelingt gut.

Der AIC-Wert des Modells liegt bei 16931.16 – das Modell passt gut,
ist aber nicht wesentlich besser als das reduzierte Modell ohne Windenergieerzeugung (AIC = 16931.2).
→ Windenergieerzeugung trägt kaum zur Erklärung bei.

Die wichtigsten Einflussfaktoren auf hohe Strompreise sind:

- Niedrige Temperatur (stark negativer Einfluss)
  → Intuitiv vorhersehbar, da Stromnachfrage zu grossem Teil von Temp. abhängig ist.

- Höherer Anteil erneuerbarer Energien geht mit höherer Preiswahrscheinlichkeit einher.

Bemerkenswert ist der signifikant positive Einfluss des Anteils erneuerbarer Energien auf die Wahrscheinlichkeit hoher Strompreise. Dieses Ergebnis steht im Widerspruch zur gängigen Erwartung, dass ein höherer Anteil erneuerbarer Erzeugung – aufgrund niedrigerer Grenzkosten – preisdämpfend wirkt.

Ein möglicher Erklärungsansatz liegt in systemischen Wechselwirkungen, etwa durch Netzengpässe, Produktionsextreme (z.B. PV-Spitzen zur Mittagszeit) oder nicht planbare Einspeisung.Zudem können Prognoseabweichungen zwischen geplanter und tatsächlich eingespeister Leistung zu Marktunsicherheit führen und Preisausschläge verstärken.

Obwohl erneuerbare Energien grundsätzlich geringere Gestehungskosten aufweisen, können diese indirekten Effekte die Marktpreise kurzfristig auch erhöhen – insbesondere bei hoher Gesamteinspeisung ohne entsprechende Netzflexibilität.

Das Modell für tiefe Strompreise zeigt noch leicht bessere Modellgüte (AIC = 15709.12, AUC = 0.88).
→ Die Vorhersage tiefer Preise gelingt statistisch etwas besser.

Beide Modelle liefern Ergebnisse mit guter Trennschärfe.Temperatur, Wind und PV-Leistung sind zentrale Einflussfaktoren –
die Rolle der erneuerbaren Energien ist komplexer und differenziert zu betrachten.

# 4 Hypothesentest und ANOVA zur Strompreisvariation

## 4.1 Theoretische Fundierung

Die theoretische Fundierung dieser Analyse basiert auf den Ausführungen bei Andreas Handl und Torben Kuhlenkasper, Multivariate Analysemethoden, 3. Aufl. 2017. Um den Einfluss der Tageszeit (Stunde) und des Wochentags auf die Strompreise zu untersuchen, wird die einfaktorielle Varianzanalyse (ANOVA) angewendet. Dabei handelt es sich um einen statistischen Hypothesentest, mit dem überprüft wird, ob sich die Mittelwerte einer metrischen Zielvariable (hier: Strompreis in CHF/MWh) zwischen mehreren Gruppen (hier: Stunden bzw. Wochentage) signifikant unterscheiden.

Die Nullhypothese (H0) lautet: Es gibt keinen Unterschied in den mittleren Strompreisen zwischen den Gruppen. Die Alternativhypothese (H1) geht davon aus, dass mindestens eine Gruppe einen signifikant anderen Mittelwert aufweist.

Wird die Nullhypothese durch die ANOVA verworfen (p-Wert $< 0.05$), folgt ein Post-hoc-Test – in diesem Fall der TukeyHSD-Test – um zu bestimmen, welche Gruppen sich voneinander unterscheiden.

In der folgenden Analyse wird daher zunächst geprüft, ob der Strompreis je nach Stunde (ANOVA Stunde) oder Wochentag (ANOVA Wochentag) signifikant variiert. Anschliessend werden mit dem Tukey-Test die konkreten Unterschiede zwischen den Gruppen identifiziert.

## 4.2 Datenvorbereiten

Daten einlesen

```r
file_path <- (
  "C:/Users/41788/OneDrive/CAS Statische Datenalyse & Datenvisualisierung/" %>%
  paste0("Skript R/01 FFHS Semesterarbeit Statistische Datenanalyse/Data.xlsx")
)
data <- read_excel(file_path, sheet = "alle", skip = 1,
                   col_names = c("Datum_MEZ", "Strompreis", "Temperatur", "Windgeschwindigkeit",
                                 "Windenergieerzeugung", "PV_Leistung", "Erneuerbare_Energien",
                                 "Netzbezogene_Leistung"))
```

Datum umwandeln

```r
data$Datum <- parse_date_time(data$Datum_MEZ, orders = c("dmy HMS", "ymd HMS"))
```

Spalten bereinigen und neue Variablen erstellen

```r
data <- data %>%
  select(Datum, Strompreis) %>%
  mutate(
    Strompreis = as.numeric(Strompreis),
    Stunde = hour(Datum),
    Wochentag = wday(Datum, label = TRUE)
  ) %>%
  na.omit()
```

## 4.3 ANOVA nach Stunden

```r
anova_stunde <- aov(Strompreis ~ factor(Stunde), data = data)
summary(anova_stunde)
```

```
##                   Df   Sum Sq Mean Sq F value Pr(>F)
## factor(Stunde)    23  4801101  208744   173.8 <2e-16 ***
## Residuals      34746 41735489    1201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(anova_stunde)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Strompreis ~ factor(Stunde), data = data)
##
## $`factor(Stunde)`
##              diff         lwr          upr     p adj
## 1-0    -3.0936346  -8.1262256   1.93895630 0.8523380
## 2-0    -5.1163749 -10.1489659  -0.08378401 0.0407583
## 3-0    -6.2750665 -11.3106111  -1.23952184 0.0014319
## 4-0    -6.3811038 -11.4136947  -1.34851286 0.0009941
## 5-0    -2.8760767  -7.9057283   2.15357492 0.9220279
## 6-0     5.6804117   0.6478208  10.71300265 0.0089362
## 7-0    13.8670313   8.8344403  18.89962218 0.0000000
## 8-0    13.5832014   8.5506105  18.61579234 0.0000000
## 9-0     4.4830765  -0.5495144   9.51566741 0.1643827
## 10-0   -3.7717711  -8.8043620   1.26081983 0.4952490
## 11-0   -9.5469292 -14.5795201  -4.51433826 0.0000000
## 12-0  -17.2059172 -22.2385081 -12.17332624 0.0000000
## 13-0  -19.5272223 -24.5598132 -14.49463134 0.0000000
## 14-0  -18.2660128 -23.2986038 -13.23342190 0.0000000
## 15-0  -10.5043844 -15.5369753  -5.47179344 0.0000000
## 16-0   -0.8304085  -5.8629994   4.20218241 1.0000000
## 17-0   10.4493202   5.4167293  15.48191115 0.0000000
## 18-0   20.1493323  15.1167413  25.18192320 0.0000000
## 19-0   22.4801514  17.4475605  27.51274232 0.0000000
## 20-0   17.4076046  12.3750137  22.44019553 0.0000000
## 21-0   11.2234705   6.1908795  16.25606139 0.0000000
## 22-0    8.4437150   3.4111241  13.47630590 0.0000003
## 23-0    3.1519812  -1.8806097   8.18457218 0.8287083
## 2-1    -2.0227403  -6.6820141   2.63653352 0.9971135
## 3-1    -3.1814318  -7.8438959   1.48103220 0.6899869
## 4-1    -3.2874692  -7.9467430   1.37180468 0.6240063
## 5-1     0.2175579  -4.4385409   4.87365680 1.0000000
## 6-1     8.7740463   4.1147725  13.43332018 0.0000000
## 7-1    16.9606659  12.3013921  21.61993972 0.0000000
## 8-1    16.6768360  12.0175622  21.33610988 0.0000000
## 9-1     7.5767111   2.9174373  12.23598495 0.0000009
## 10-1   -0.6781365  -5.3374103   3.98113736 1.0000000
```

```
## 11-1   -6.4532946 -11.1125684  -1.79402073 0.0001236
## 12-1  -14.1122825 -18.7715564  -9.45300871 0.0000000
## 13-1  -16.4335876 -21.0928615 -11.77431380 0.0000000
## 14-1  -15.1723782 -19.8316520 -10.51310436 0.0000000
## 15-1   -7.4107497 -12.0700236  -2.75147590 0.0000020
## 16-1    2.2632261  -2.3960477   6.92249995 0.9870748
## 17-1   13.5429549   8.8836810  18.20222869 0.0000000
## 18-1   23.2429669  18.5836931  27.90224074 0.0000000
## 19-1   25.5737860  20.9145122  30.23305986 0.0000000
## 20-1   20.5012392  15.8419654  25.16051307 0.0000000
## 21-1   14.3171051   9.6578313  18.97637892 0.0000000
## 22-1   11.5373496   6.8780758  16.19662344 0.0000000
## 23-1    6.2456159   1.5863420  10.90488971 0.0002801
## 3-2    -1.1586915  -5.8211555   3.50377250 0.9999998
## 4-2    -1.2647288  -5.9240027   3.39454498 0.9999990
## 5-2     2.2402982  -2.4158006   6.89639710 0.9885248
## 6-2    10.7967867   6.1375128  15.45606049 0.0000000
## 7-2    18.9834062  14.3241324  23.64268003 0.0000000
## 8-2    18.6995764  14.0403025  23.35885019 0.0000000
## 9-2     9.5994514   4.9401776  14.25872526 0.0000000
## 10-2    1.3446038  -3.3146700   6.00387767 0.9999969
## 11-2   -4.4305543  -9.0898281   0.22871958 0.0882647
## 12-2  -12.0895422 -16.7488161  -7.43026840 0.0000000
## 13-2  -14.4108473 -19.0701212  -9.75157349 0.0000000
## 14-2  -13.1496379 -17.8089117  -8.49036406 0.0000000
## 15-2   -5.3880094 -10.0472833  -0.72873559 0.0059276
## 16-2    4.2859664  -0.3733074   8.94524026 0.1229746
## 17-2   15.5656952  10.9064213  20.22496900 0.0000000
## 18-2   25.2657072  20.6064334  29.92498105 0.0000000
## 19-2   27.5965263  22.9372525  32.25580017 0.0000000
## 20-2   22.5239795  17.8647057  27.18325338 0.0000000
## 21-2   16.3398454  11.6805716  20.99911923 0.0000000
## 22-2   13.5600899   8.9008161  18.21936375 0.0000000
## 23-2    8.2683562   3.6090824  12.92763002 0.0000000
## 4-3    -0.1060373  -4.7685014   4.55642670 1.0000000
## 5-3     3.3989898  -1.2603015   8.05828099 0.5538363
## 6-3    11.9554782   7.2930142  16.61794220 0.0000000
## 7-3    20.1420977  15.4796337  24.80456174 0.0000000
## 8-3    19.8582679  15.1958038  24.52073190 0.0000000
## 9-3    10.7581429   6.0956789  15.42060697 0.0000000
## 10-3    2.5032954  -2.1591687   7.16575938 0.9587727
## 11-3   -3.2718627  -7.9343268   1.39060129 0.6350881
## 12-3  -10.9308507 -15.5933147  -6.26838669 0.0000000
## 13-3  -13.2521558 -17.9146198  -8.58969178 0.0000000
## 14-3  -11.9909464 -16.6534104  -7.32848235 0.0000000
## 15-3   -4.2293179  -8.8917819   0.43314612 0.1400568
## 16-3    5.4446579   0.7821939  10.10712197 0.0049894
## 17-3   16.7243867  12.0619227  21.38685071 0.0000000
## 18-3   26.4243987  21.7619347  31.08686276 0.0000000
## 19-3   28.7552179  24.0927538  33.41768188 0.0000000
## 20-3   23.6826711  19.0202070  28.34513509 0.0000000
## 21-3   17.4985369  12.8360729  22.16100094 0.0000000
## 22-3   14.7187814  10.0563174  19.38124546 0.0000000
## 23-3    9.4270477   4.7645837  14.08951173 0.0000000
```

```
## 5-4     3.5050271  -1.1510718   8.16112595 0.4856170
## 6-4    12.0615155   7.4022417  16.72078934 0.0000000
## 7-4    20.2481350  15.5888612  24.90740887 0.0000000
## 8-4    19.9643052  15.3050314  24.62357903 0.0000000
## 9-4    10.8641803   6.2049064  15.52345411 0.0000000
## 10-4    2.6093327  -2.0499411   7.26860652 0.9363254
## 11-4   -3.1658254  -7.8250992   1.49344843 0.6979256
## 12-4  -10.8248134 -15.4840872  -6.16553955 0.0000000
## 13-4  -13.1461185 -17.8053923  -8.48684465 0.0000000
## 14-4  -11.8849090 -16.5441829  -7.22563521 0.0000000
## 15-4   -4.1232806  -8.7825544   0.53599326 0.1738346
## 16-4    5.5506953   0.8914214  10.20996911 0.0034656
## 17-4   16.8304240  12.1711502  21.48969785 0.0000000
## 18-4   26.5304361  21.8711622  31.18970990 0.0000000
## 19-4   28.8612552  24.2019814  33.52052902 0.0000000
## 20-4   23.7887084  19.1294346  28.44798223 0.0000000
## 21-4   17.6045742  12.9453004  22.26384808 0.0000000
## 22-4   14.8248188  10.1655449  19.48409260 0.0000000
## 23-4    9.5330850   4.8738112  14.19235887 0.0000000
## 6-5     8.5564884   3.9003896  13.21258727 0.0000000
## 7-5    16.7431079  12.0870091  21.39920681 0.0000000
## 8-5    16.4592781  11.8031792  21.11537697 0.0000000
## 9-5     7.3591532   2.7030543  12.01525204 0.0000024
## 10-5   -0.8956944  -5.5517933   3.76040445 1.0000000
## 11-5   -6.6708525 -11.3269514  -2.01475364 0.0000499
## 12-5  -14.3298405 -18.9859393  -9.67374162 0.0000000
## 13-5  -16.6511456 -21.3072444 -11.99504671 0.0000000
## 14-5  -15.3899361 -20.0460350 -10.73383727 0.0000000
## 15-5   -7.6283077 -12.2844065  -2.97220881 0.0000007
## 16-5    2.0456682  -2.6104307   6.70176704 0.9965829
## 17-5   13.3253969   8.6692981  17.98149578 0.0000000
## 18-5   23.0254090  18.3693101  27.68150783 0.0000000
## 19-5   25.3562281  20.7001292  30.01232695 0.0000000
## 20-5   20.2836813  15.6275824  24.93978016 0.0000000
## 21-5   14.0995472   9.4434483  18.75564601 0.0000000
## 22-5   11.3197917   6.6636928  15.97589053 0.0000000
## 23-5    6.0280579   1.3719591  10.68415680 0.0006294
## 7-6     8.1866195   3.5273457  12.84589337 0.0000000
## 8-6     7.9027897   3.2435159  12.56206353 0.0000002
## 9-6    -1.1973352  -5.8566091   3.46193860 0.9999997
## 10-6   -9.4521828 -14.1114567  -4.79290899 0.0000000
## 11-6  -15.2273409 -19.8866147 -10.56806708 0.0000000
## 12-6  -22.8863289 -27.5456027 -18.22705505 0.0000000
## 13-6  -25.2076340 -29.8669078 -20.54836015 0.0000000
## 14-6  -23.9464245 -28.6056984 -19.28715071 0.0000000
## 15-6  -16.1847961 -20.8440699 -11.52552225 0.0000000
## 16-6   -6.5108202 -11.1700941  -1.85154640 0.0000980
## 17-6    4.7689085   0.1096347   9.42818234 0.0373942
## 18-6   14.4689206   9.8096467  19.12819439 0.0000000
## 19-6   16.7997397  12.1404658  21.45901351 0.0000000
## 20-6   11.7271929   7.0679191  16.38646672 0.0000000
## 21-6    5.5430587   0.8837849  10.20233257 0.0035557
## 22-6    2.7633033  -1.8959706   7.42257709 0.8910256
## 23-6   -2.5284305  -7.1877043   2.13084336 0.9538341
```

```
## 8-7    -0.2838298  -4.9431037   4.37544399 1.0000000
## 9-7    -9.3839548 -14.0432286  -4.72468093 0.0000000
## 10-7  -17.6388024 -22.2980762 -12.97952852 0.0000000
## 11-7  -23.4139604 -28.0732343 -18.75468661 0.0000000
## 12-7  -31.0729484 -35.7322223 -26.41367459 0.0000000
## 13-7  -33.3942535 -38.0535274 -28.73497969 0.0000000
## 14-7  -32.1330441 -36.7923179 -27.47377025 0.0000000
## 15-7  -24.3714156 -29.0306895 -19.71214178 0.0000000
## 16-7  -14.6974398 -19.3567136 -10.03816593 0.0000000
## 17-7   -3.4177110  -8.0769849   1.24156281 0.5419908
## 18-7    6.2823010   1.6230272  10.94157486 0.0002429
## 19-7    8.6131201   3.9538463  13.27239398 0.0000000
## 20-7    3.5405734  -1.1187005   8.19984719 0.4650717
## 21-7   -2.6435608  -7.3028346   2.01571304 0.9276575
## 22-7   -5.4233163 -10.0825901  -0.76404244 0.0052852
## 23-7  -10.7150500 -15.3743238  -6.05577617 0.0000000
## 9-8    -9.1001249 -13.7593988  -4.44085109 0.0000000
## 10-8  -17.3549725 -22.0142463 -12.69569868 0.0000000
## 11-8  -23.1301306 -27.7894044 -18.47085677 0.0000000
## 12-8  -30.7891186 -35.4483924 -26.12984475 0.0000000
## 13-8  -33.1104237 -37.7696975 -28.45114985 0.0000000
## 14-8  -31.8492142 -36.5084881 -27.18994041 0.0000000
## 15-8  -24.0875858 -28.7468596 -19.42831195 0.0000000
## 16-8  -14.4136099 -19.0728838  -9.75433609 0.0000000
## 17-8   -3.1338812  -7.7931550   1.52539265 0.7165294
## 18-8    6.5661309   1.9068570  11.22540469 0.0000783
## 19-8    8.8969500   4.2376761  13.55622381 0.0000000
## 20-8    3.8244032  -0.8348706   8.48367703 0.3033212
## 21-8   -2.3597310  -7.0190048   2.29954288 0.9785841
## 22-8   -5.1394864  -9.7987603  -0.48021260 0.0129154
## 23-8  -10.4312202 -15.0904940  -5.77194633 0.0000000
## 10-9   -8.2548476 -12.9141214  -3.59557375 0.0000000
## 11-9  -14.0300057 -18.6892795  -9.37073184 0.0000000
## 12-9  -21.6889937 -26.3482675 -17.02971982 0.0000000
## 13-9  -24.0102988 -28.6695726 -19.35102492 0.0000000
## 14-9  -22.7490893 -27.4083631 -18.08981548 0.0000000
## 15-9  -14.9874609 -19.6467347 -10.32818702 0.0000000
## 16-9   -5.3134850  -9.9727588  -0.65421117 0.0075268
## 17-9    5.9662437   1.3069699  10.62551758 0.0008030
## 18-9   15.6662558  11.0069820  20.32552962 0.0000000
## 19-9   17.9970749  13.3378011  22.65634874 0.0000000
## 20-9   12.9245281   8.2652543  17.58380195 0.0000000
## 21-9    6.7403940   2.0811201  11.39966781 0.0000380
## 22-9    3.9606385  -0.6986353   8.61991232 0.2384490
## 23-9   -1.3310952  -5.9903691   3.32817860 0.9999974
## 11-10  -5.7751581 -10.4344319  -1.11588426 0.0015983
## 12-10 -13.4341461 -18.0934199  -8.77487224 0.0000000
## 13-10 -15.7554512 -20.4147250 -11.09617733 0.0000000
## 14-10 -14.4942417 -19.1535156  -9.83496789 0.0000000
## 15-10  -6.7326133 -11.3918871  -2.07333943 0.0000393
## 16-10   2.9413626  -1.7179112   7.60063642 0.8179637
## 17-10  14.2210913   9.5618175  18.88036516 0.0000000
## 18-10  23.9211034  19.2618295  28.58037721 0.0000000
## 19-10  26.2519225  21.5926487  30.91119633 0.0000000
```

```
## 20-10  21.1793757  16.5201019  25.83864954 0.0000000
## 21-10  14.9952416  10.3359677  19.65451539 0.0000000
## 22-10  12.2154861   7.5562122  16.87475991 0.0000000
## 23-10   6.9237523   2.2644785  11.58302618 0.0000174
## 12-11  -7.6589880 -12.3182618  -2.99971415 0.0000006
## 13-11  -9.9802931 -14.6395669  -5.32101924 0.0000000
## 14-11  -8.7190836 -13.3783575  -4.05980980 0.0000000
## 15-11  -0.9574552  -5.6167290   3.70181866 1.0000000
## 16-11   8.7165207   4.0572468  13.37579451 0.0000000
## 17-11  19.9962494  15.3369756  24.65552325 0.0000000
## 18-11  29.6962615  25.0369876  34.35553530 0.0000000
## 19-11  32.0270806  27.3678068  36.68635442 0.0000000
## 20-11  26.9545338  22.2952600  31.61380763 0.0000000
## 21-11  20.7703997  16.1111258  25.42967348 0.0000000
## 22-11  17.9906442  13.3313703  22.64991800 0.0000000
## 23-11  12.6989104   8.0396366  17.35818427 0.0000000
## 13-12  -2.3213051  -6.9805789   2.33796873 0.9823782
## 14-12  -1.0600957  -5.7193695   3.59917817 1.0000000
## 15-12   6.7015328   2.0422590  11.36080664 0.0000448
## 16-12  16.3755087  11.7162348  21.03478249 0.0000000
## 17-12  27.6552374  22.9959636  32.31451123 0.0000000
## 18-12  37.3552494  32.6959756  42.01452328 0.0000000
## 19-12  39.6860686  35.0267947  44.34534240 0.0000000
## 20-12  34.6135218  29.9542479  39.27279561 0.0000000
## 21-12  28.4293876  23.7701138  33.08866146 0.0000000
## 22-12  25.6496321  20.9903583  30.30890598 0.0000000
## 23-12  20.3578984  15.6986246  25.01717225 0.0000000
## 14-13   1.2612094  -3.3980644   5.92048327 0.9999991
## 15-13   9.0228379   4.3635641  13.68211174 0.0000000
## 16-13  18.6968138  14.0375399  23.35608759 0.0000000
## 17-13  29.9765425  25.3172687  34.63581633 0.0000000
## 18-13  39.6765545  35.0172807  44.33582838 0.0000000
## 19-13  42.0073737  37.3480998  46.66664750 0.0000000
## 20-13  36.9348269  32.2755530  41.59410071 0.0000000
## 21-13  30.7506927  26.0914189  35.40996656 0.0000000
## 22-13  27.9709372  23.3116634  32.63021108 0.0000000
## 23-13  22.6792035  18.0199297  27.33847735 0.0000000
## 15-14   7.7616285   3.1023546  12.42090230 0.0000004
## 16-14  17.4356043  12.7763305  22.09487815 0.0000000
## 17-14  28.7153331  24.0560592  33.37460689 0.0000000
## 18-14  38.4153451  33.7560713  43.07461894 0.0000000
## 19-14  40.7461642  36.0868904  45.40543806 0.0000000
## 20-14  35.6736174  31.0143436  40.33289127 0.0000000
## 21-14  29.4894833  24.8302095  34.14875712 0.0000000
## 22-14  26.7097278  22.0504540  31.36900164 0.0000000
## 23-14  21.4179941  16.7587202  26.07726791 0.0000000
## 16-15   9.6739759   5.0147020  14.33324968 0.0000000
## 17-15  20.9537046  16.2944308  25.61297843 0.0000000
## 18-15  30.6537166  25.9944428  35.31299047 0.0000000
## 19-15  32.9845358  28.3252619  37.64380959 0.0000000
## 20-15  27.9119890  23.2527151  32.57126280 0.0000000
## 21-15  21.7278548  17.0685810  26.38712866 0.0000000
## 22-15  18.9480993  14.2888255  23.60737317 0.0000000
## 23-15  13.6563656   8.9970918  18.31563945 0.0000000
```

```
## 17-16  11.2797287   6.6204549  15.93900257 0.0000000
## 18-16  20.9797408  16.3204670  25.63901462 0.0000000
## 19-16  23.3105599  18.6512861  27.96983374 0.0000000
## 20-16  18.2380131  13.5787393  22.89728695 0.0000000
## 21-16  12.0538790   7.3946051  16.71315281 0.0000000
## 22-16   9.2741235   4.6148497  13.93339732 0.0000000
## 23-16   3.9823898  -0.6768841   8.64166360 0.2289894
## 18-17   9.7000120   5.0407382  14.35928588 0.0000000
## 19-17  12.0308312   7.3715573  16.69010500 0.0000000
## 20-17   6.9582844   2.2990105  11.61755821 0.0000150
## 21-17   0.7741502  -3.8851236   5.43342406 1.0000000
## 22-17  -2.0056053  -6.6648791   2.65366858 0.9974435
## 23-17  -7.2973390 -11.9566128  -2.63806515 0.0000033
## 19-18   2.3308191  -2.3284547   6.99009295 0.9814930
## 20-18  -2.7417277  -7.4010015   1.91754616 0.8983680
## 21-18  -8.9258618 -13.5851356  -4.26658798 0.0000000
## 22-18 -11.7056173 -16.3648911  -7.04634347 0.0000000
## 23-18 -16.9973510 -21.6566249 -12.33807719 0.0000000
## 20-19  -5.0725468  -9.7318206  -0.41327296 0.0157901
## 21-19 -11.2566809 -15.9159548  -6.59740710 0.0000000
## 22-19 -14.0364364 -18.6957103  -9.37716259 0.0000000
## 23-19 -19.3281701 -23.9874440 -14.66889631 0.0000000
## 21-20  -6.1841341 -10.8434080  -1.52486031 0.0003548
## 22-20  -8.9638896 -13.6231635  -4.30461580 0.0000000
## 23-20 -14.2556234 -18.9148972  -9.59634952 0.0000000
## 22-21  -2.7797555  -7.4390293   1.87951835 0.8852049
## 23-21  -8.0714892 -12.7307630  -3.41221538 0.0000001
## 23-22  -5.2917337  -9.9510076  -0.63245989 0.0080634
```

Der p-Wert $< 0.001$ ist extrem klein → die Unterschiede im Strompreis zwischen den Stunden sind also hochsignifikant. Die Nullhypothese kann klar verworfen werden.

Die F-Statistik (173.8) zeigt eine starke Varianz zwischen den Stunden im Vergleich zur Varianz innerhalb der Gruppen. Die Uhrzeit hat, wie zu erwarten, einen signifikanten Einfluss auf den Strompreis.

Strompreis ist um 13 Uhr am günstigsten und für 19 Uhr am teuersten. Grundsätzlich ist der Preis am Abend (18–20 Uhr) am teuersten und am frühen Nachmittag (13–15 Uhr) am günstigsten.

## 4.4 ANOVA nach Wochentagen

```r
anova_wochentag <- aov(Strompreis ~ Wochentag, data = data)
anova_summary <- summary(anova_wochentag)
knitr::kable(as.data.frame(anova_summary[[1]]), caption = "ANOVA-Ergebnisse nach Wochentagen")
```

Table 4: ANOVA-Ergebnisse nach Wochentagen

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Wochentag | 6 | 3352581 | 558763.433 | 449.8029 | 0 |
| Residuals | 34763 | 43184010 | 1242.241 | NA | NA |

```r
tukey_result <- TukeyHSD(anova_wochentag)
knitr::kable(as.data.frame(tukey_result$Wochentag), caption = "TukeyHSD Post-Hoc-Vergleich: Wochentage")
```

Table 5: TukeyHSD Post-Hoc-Vergleich: Wochentage

| | diff | lwr | upr | p adj |
|---|---|---|---|---|
| Mo-So | 25.0115636 | 22.9305701 | 27.0925571 | 0.0000000 |
| Di-So | 27.0933368 | 25.0123433 | 29.1743303 | 0.0000000 |
| Mi-So | 28.4887164 | 26.3978369 | 30.5795959 | 0.0000000 |
| Do-So | 25.4313240 | 23.3404445 | 27.5222035 | 0.0000000 |
| Fr-So | 24.3610939 | 22.2702144 | 26.4519734 | 0.0000000 |
| Sa-So | 10.9689592 | 8.8780797 | 13.0598387 | 0.0000000 |
| Di-Mo | 2.0817732 | 0.0107128 | 4.1528335 | 0.0478448 |
| Mi-Mo | 3.4771528 | 1.3961593 | 5.5581463 | 0.0000173 |
| Do-Mo | 0.4197603 | -1.6612332 | 2.5007539 | 0.9969845 |
| Fr-Mo | -0.6504697 | -2.7314632 | 1.4305238 | 0.9692450 |
| Sa-Mo | -14.0426045 | -16.1235980 | -11.9616109 | 0.0000000 |
| Mi-Di | 1.3953796 | -0.6856139 | 3.4763731 | 0.4294063 |
| Do-Di | -1.6620128 | -3.7430063 | 0.4189807 | 0.2181797 |
| Fr-Di | -2.7322429 | -4.8132364 | -0.6512494 | 0.0020876 |
| Sa-Di | -16.1243776 | -18.2053711 | -14.0433841 | 0.0000000 |
| Do-Mi | -3.0573924 | -5.1482719 | -0.9665129 | 0.0003262 |
| Fr-Mi | -4.1276225 | -6.2185020 | -2.0367429 | 0.0000001 |
| Sa-Mi | -17.5197572 | -19.6106367 | -15.4288777 | 0.0000000 |
| Fr-Do | -1.0702301 | -3.1611096 | 1.0206495 | 0.7395082 |
| Sa-Do | -14.4623648 | -16.5532443 | -12.3714853 | 0.0000000 |
| Sa-Fr | -13.3921347 | -15.4830143 | -11.3012552 | 0.0000000 |

Auch hier ist der p-Wert extrem klein → starke, signifikante Unterschiede zwischen den Wochentagen. Die Nullhypothese kann auch hier klar verworfen werden.

Der Wochentag hat einen klaren Einfluss auf den Strompreis. Sonntag ist mit Abstand am günstigsten. Montag bis Freitag sind signifikant teurer. Samstag liegt dazwischen, aber immer noch günstiger als Werktage. Preisniveau: Mittwoch/Donnerstag/Dienstag > Montag > Freitag > Samstag > Sonntag.

### 4.5 Interpretation der Ergebnisse: Hypothesentest und ANOVA

Die ANOVA zeigt, dass es signifikante Unterschiede der Strompreise je nach Stunde gibt ($p < 0.001$). Die anschliessende TukeyHSD-Analyse zeigt, dass insbesondere die Abendstunden (18–20 Uhr) signifikant höhere Strompreise aufweisen als fast alle anderen Stunden des Tages.

Der grösste Unterschied besteht zwischen 19 Uhr und 13 Uhr, mit einem durchschnittlichen Preisunterschied von über 42 CHF/MWh. Dies lässt sich durch den Rückgang auf der Angebotsseite (erneuerbare Energien) erklären.

Im Gegensatz dazu sind die Preise zwischen 13–15 Uhr am tiefsten. Dies lässt sich aus der erhöhten Einspeisung von Photovoltaik-Strom zur Mittagszeit begründen.

Insgesamt zeigen die Ergebnisse, dass die Tageszeit ein klarer Einflussfaktor auf den Strompreis ist.

# 5 Hauptkomponenten- und Faktoranalyse

## 5.1 Theoretische Fundierung

Die Hauptkomponentenanalyse (PCA) ist ein Verfahren zur Dimensionsreduktion, das stark korrelierte Variablen durch eine kleinere Zahl unkorrelierter Hauptkomponenten ersetzt. Ziel ist es, die Varianz der Ausgangsdaten mit möglichst wenigen Komponenten zu erklären (vgl. Handl & Kuhlenkasper, Multivariate Analysemethoden, 3. Aufl. 2017).

Die Hauptkomponenten sind lineare Kombinationen der ursprünglichen Variablen und entstehen durch Eigenwertzerlegung der Kovarianz- oder Korrelationsmatrix. Die Eigenwerte geben an, wie viel Varianz durch die jeweilige Komponente erklärt wird. Zur Vermeidung von Verzerrungen werden die Variablen zuvor standardisiert.

In dieser Analyse dient die PCA dazu, Korrelationen und redundante Informationen zwischen den Einflussgrössen auf den Strompreis zu erkennen. Die Auswahl relevanter Komponenten erfolgte anhand des Scree-Plots und des Kaiser-Kriteriums. Die PCA unterstützt so die Strukturierung der Daten und liefert Hinweise auf potenzielle Multikollinearitäten.

## 5.2 Daten vorbereiten

Daten einlesen

```r
file_path <- (
  "C:/Users/41788/OneDrive/CAS Statische Datenalyse & Datenvisualisierung/" %>%
  paste0("Skript R/01 FFHS Semesterarbeit Statistische Datenanalyse/Data.xlsx")
)
data <- read_excel(file_path, sheet = "alle", skip = 1,
                   col_names = c("Datum_MEZ", "Strompreis", "Temperatur", "Windgeschwindigkeit",
                                 "Windenergieerzeugung", "PV_Leistung", "Erneuerbare_Energien",
                                 "Netzbezogene_Leistung"))
```

Relevante numerische Variablen vorbereiten

```r
df_pca <- data %>%
  select(Temperatur, Windgeschwindigkeit, Windenergieerzeugung,
         PV_Leistung, Erneuerbare_Energien, Netzbezogene_Leistung) %>%
  mutate(across(everything(), ~ as.numeric(gsub(",", ".", gsub("[^0-9.,-]", "", .))))) %>%
  na.omit()
```

## 5.3 Hauptkomponentenanalyse (PCA)

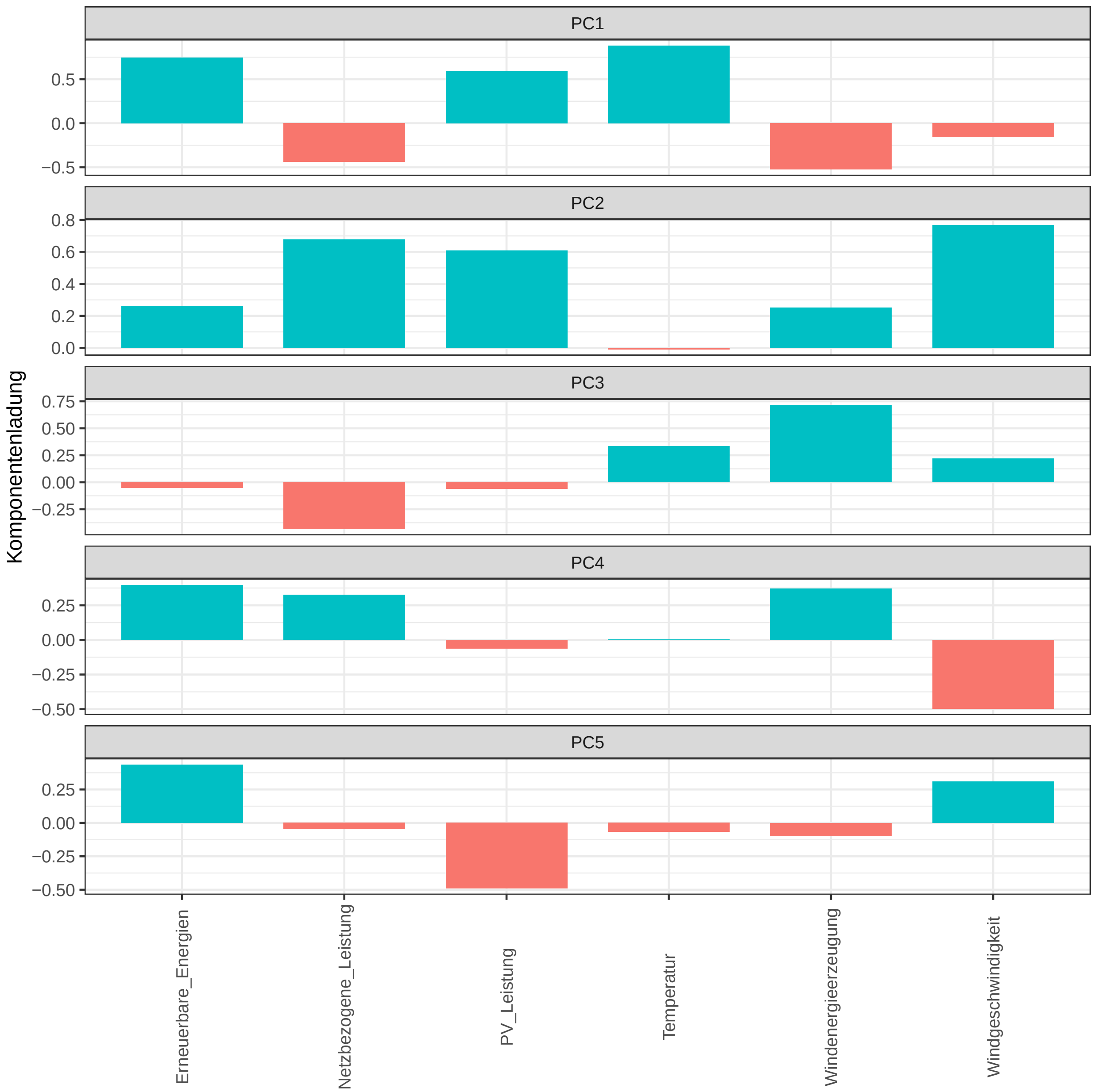
```r
pca_result <- principal(df_pca, nfactors = ncol(df_pca), rotate = "none")
knitr::kable(as.data.frame(unclass(pca_result$loadings)), caption = "PCA Loadings")
```

Table 6: PCA Loadings

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|---|---|---|---|---|---|
| Temperatur | 0.8807098 | -0.0086515 | 0.3324286 | 0.0034255 | -0.0669040 | 0.3305734 |
| Windgeschwindigkeit | -0.1531535 | 0.7657097 | 0.2179282 | -0.4968080 | 0.3093519 | 0.0149380 |
| Windenergieerzeugung | -0.5259108 | 0.2519662 | 0.7143668 | 0.3707535 | -0.0986247 | -0.0492542 |
| PV_Leistung | 0.5921377 | 0.6081431 | -0.0598431 | -0.0635193 | -0.4901524 | -0.1779600 |
| Erneuerbare_Energien | 0.7466387 | 0.2632884 | -0.0521668 | 0.3976105 | 0.4361262 | -0.1489577 |
| Netzbezogene_Leistung | -0.4406809 | 0.6791384 | -0.4326980 | 0.3253617 | -0.0432116 | 0.2227471 |

## 5.4 Visualisierung der Loadings für die ersten 5 Hauptkomponenten

```r
loadings <- as.data.frame(pca_result$loadings[, 1:5])
loadings$Variable <- rownames(loadings)
loadings_long <- gather(loadings, key = "Komponente", value = "Ladung", -Variable)

ggplot(loadings_long, aes(x = Variable, y = Ladung, fill = Ladung > 0)) +
  geom_bar(stat = 'identity', width = 0.75) +
  facet_wrap(~ Komponente, ncol = 1, scales = 'free_y') +
  guides(fill = 'none') +
  ylab('Komponentenladung') +
  theme_bw() +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_text(angle = 90, vjust = 0.5))
```




Die ersten drei PCs bilden klar inhaltlich unterscheidbare Energiesystem-Aspekte ab:

PC1 → solar / Temperatur) / erneuerbar (ohne Wind)

PC2 → Netzlast / Windgeschwindigkeit

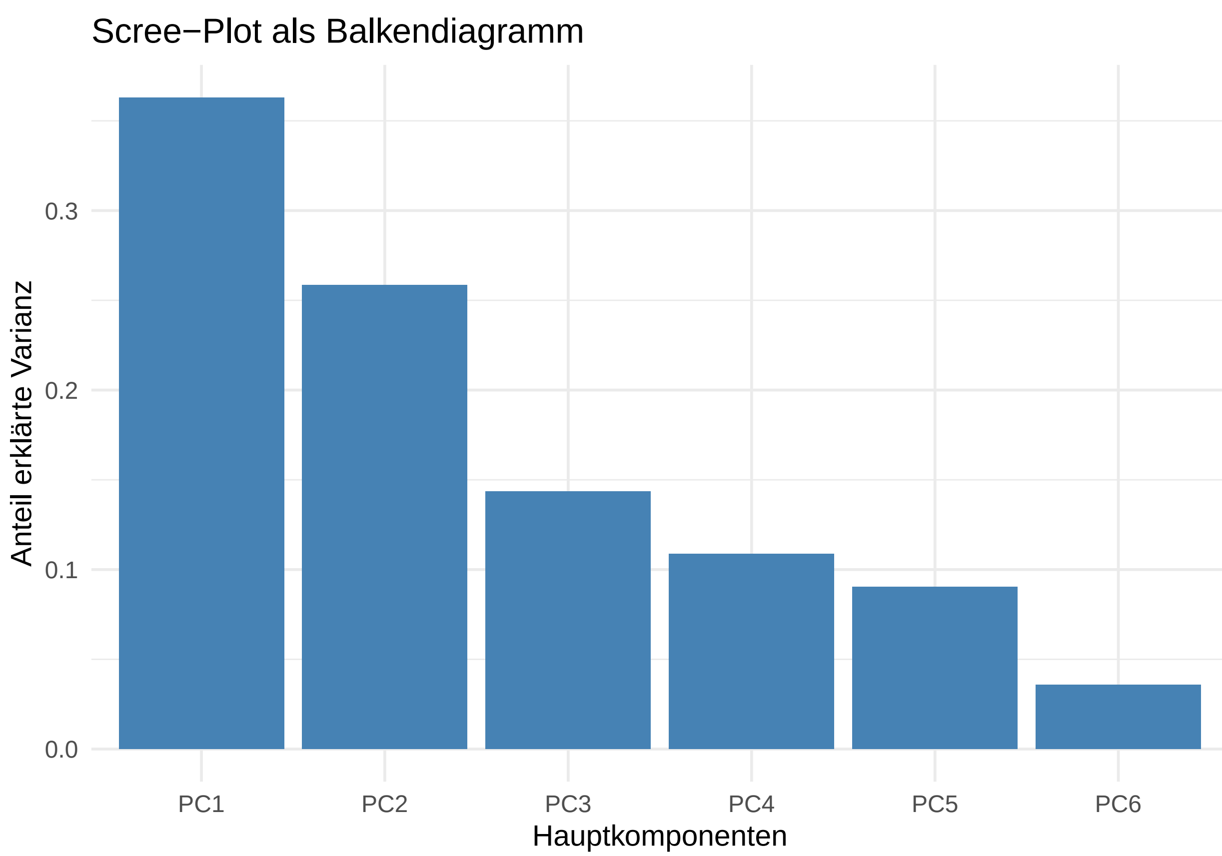
PC3 → Windkomponenten

PC4 & PC5 tragen weniger Varianz bei und zeigen keine eindeutige Struktur.

## 5.5 Scree-Plot als Balkendiagramm: Erklärte Varianz pro Hauptkomponente

```
pca_variance <- pca_result$values / sum(pca_result$values)
variance_df <- data.frame(
  Komponente = paste0("PC", 1:length(pca_variance)),
```

```r
  Varianzanteil = pca_variance
)

plot_scree <- ggplot(variance_df, aes(x = Komponente, y = Varianzanteil)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  ylab("Anteil erklärte Varianz") +
  xlab("Hauptkomponenten") +
  ggtitle("Scree-Plot als Balkendiagramm") +
  theme_minimal()

print(plot_scree)
```




PC1 erklärt den grössten Anteil der Gesamtvarianz, nämlich über 35 %. → Diese Komponente enthält also den wichtigsten gemeinsamen Informationsgehalt der Variablen.

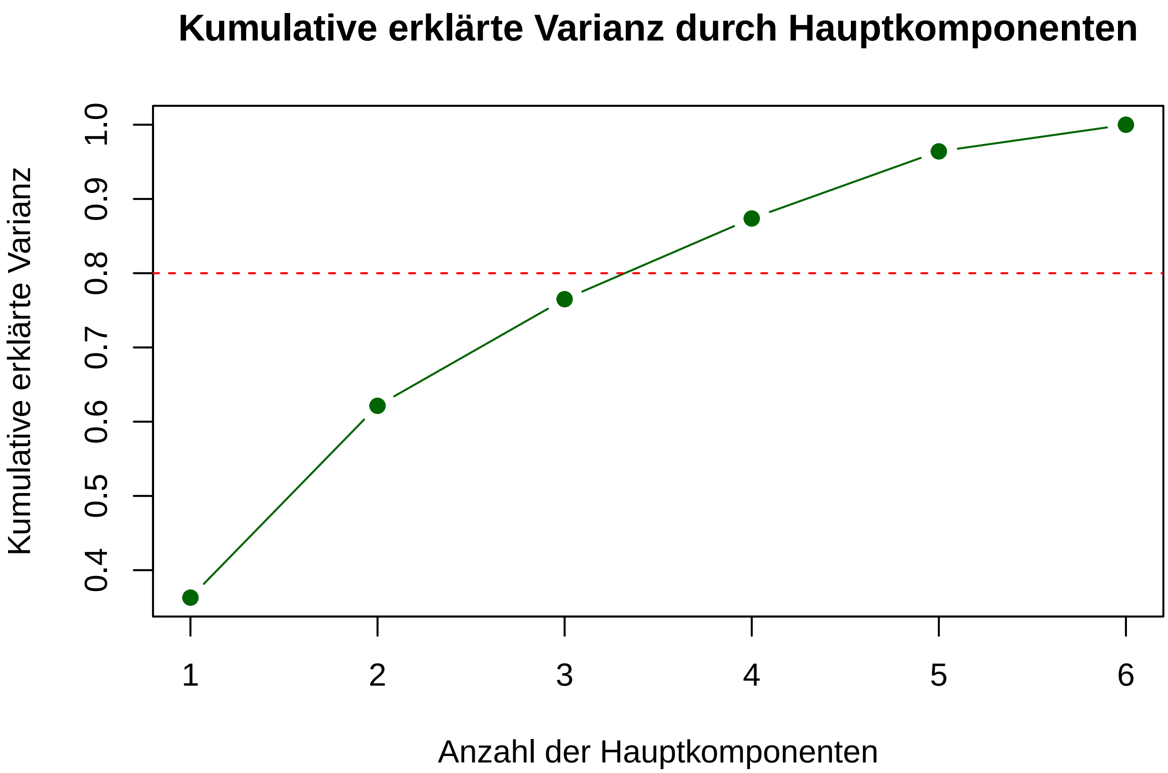
PC2 trägt mit etwa 26 % ebenfalls wesentlich zur Gesamtvarianz bei.

PC3 bis PC5 tragen jeweils nur noch zwischen 8 % und 15 % zur Varianz bei.

PC6 erklärt nur einen sehr geringen Anteil (< 5 %) und ist vermutlich nicht mehr interpretativ relevant.

```r
explained_var_ratio <- pca_result$values / sum(pca_result$values)
cum_explained_var <- cumsum(explained_var_ratio)

plot(cum_explained_var, type = "b", pch = 19, col = "darkgreen",
     xlab = "Anzahl der Hauptkomponenten",
     ylab = "Kumulative erklärte Varianz",
```

```
    main = "Kumulative erklärte Varianz durch Hauptkomponenten")
abline(h = 0.8, col = "red", lty = 2)
```




Die Hauptkomponentenanalyse zeigt, dass die ersten drei Hauptkomponenten (PC1–PC3) zusammen 77 % der Gesamtvarianz in den Daten erklären. Besonders PC1 trägt mit 36 % den grössten Anteil zur Datenstruktur bei und ist stark mit Temperatur (0.88), Erneuerbare Energien (0.75) und PV-Leistung (0.59) assoziiert – was auf eine „Solar-Komponente“ hinweist.

PC2 wird stark von Windgeschwindigkeit (0.77) und Netzbezogene Leistung (0.68) geprägt und kann als „Wind-/Last-Komponente“ interpretiert werden. Da die Anzahl der Variablen bereits klein ist, ist die Hauptkomponentenanalyse nicht sehr aussagekräftig.

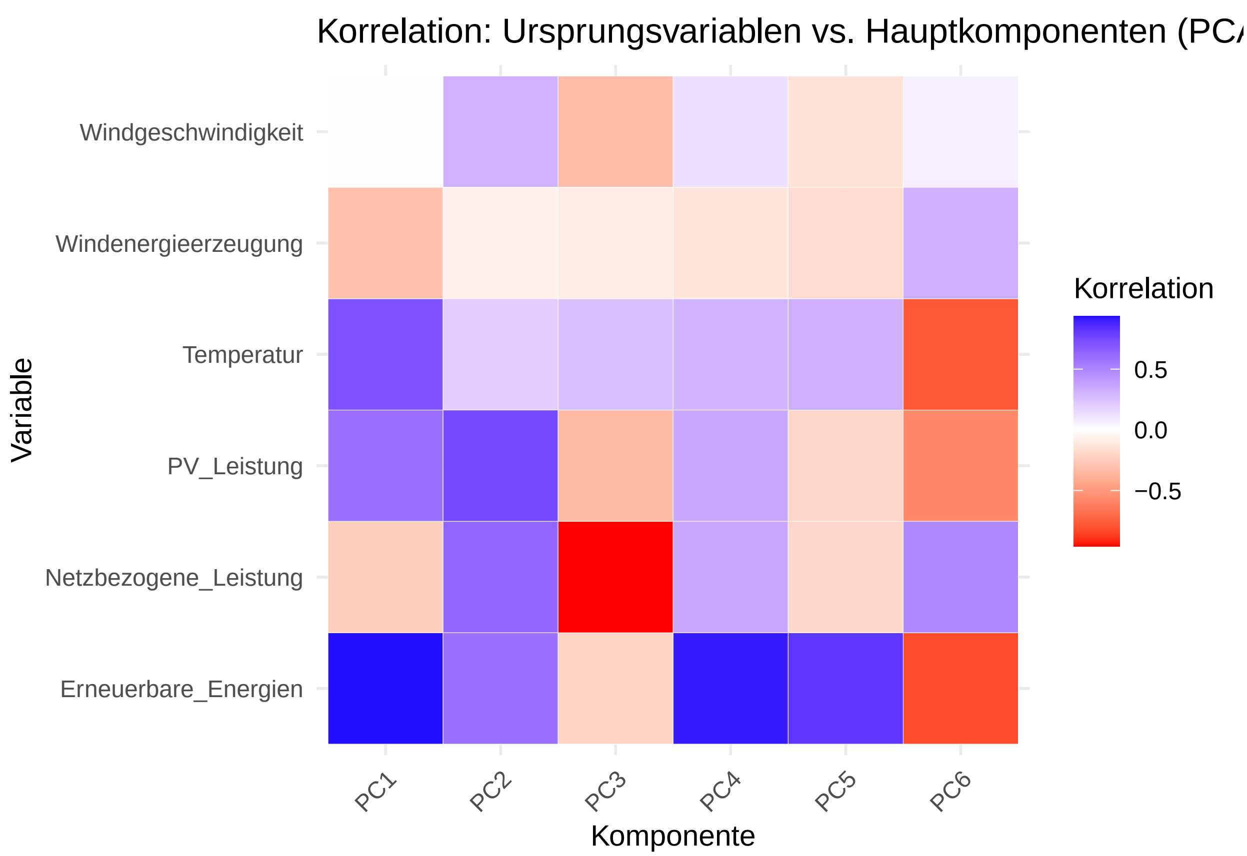
### 5.6 Projektion auf Hauptkomponenten berechnen

```
df_zentriert <- scale(df_pca, center = TRUE, scale = FALSE)
scores <- df_zentriert %*% pca_result$loadings
pc1 <- scores[, 1]
```

### 5.7 Korrelation mit Originalvariablen

```
scores_df <- as.data.frame(scores)
colnames(scores_df) <- paste0("PC", 1:ncol(scores_df))
```

```r
cor_matrix <- cor(df_pca, scores_df)

cor_df <- as.data.frame(cor_matrix)
cor_df$Variable <- rownames(cor_df)
cor_long <- gather(cor_df, key = "Komponente", value = "Korrelation", -Variable)

ggplot(cor_long, aes(x = Komponente, y = Variable, fill = Korrelation)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "red", high = "blue", mid = "white", midpoint = 0) +
  theme_minimal() +
  ggtitle("Korrelation: Ursprungsvariablen vs. Hauptkomponenten (PCA Scores)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




Diese Zuordnungen und Interpretation der Komponenten wurden bereits in den vorhergehenden Kapiteln auf Basis der Komponentenladungen und des Scree-Plots ausführlich beschrieben und werden hier durch die Korrelationsdarstellung nochmals grafisch bestätigt.

### 5.8 Explorative Faktoranalyse mit 5 Faktoren (Varimax-Rotation)

```r
fa_result <- fa(df_pca, nfactors = 5, rotate = "varimax")

# Rotierte Loadings als saubere Tabelle (nur Ausgabe über kable)
loadings_df <- as.data.frame(unclass(fa_result$loadings))
knitr::kable(loadings_df, digits = 2, caption = "Faktoranalyse: Rotierte Loadings")
```

Table 7: Faktoranalyse: Rotierte Loadings

| | MR1 | MR5 | MR4 | MR3 | MR2 |
|---|---|---|---|---|---|
| Temperatur | 0.76 | -0.51 | 0.38 | -0.02 | -0.08 |
| Windgeschwindigkeit | 0.02 | 0.20 | 0.13 | 0.18 | 0.56 |
| Windenergieerzeugung | -0.17 | 0.10 | -0.09 | 0.63 | 0.17 |
| PV_Leistung | 0.40 | 0.09 | 0.62 | -0.17 | 0.29 |
| Erneuerbare_Energien | 0.68 | 0.01 | 0.12 | -0.22 | 0.04 |
| Netzbezogene_Leistung | -0.07 | 0.81 | 0.07 | 0.13 | 0.23 |

```r
# Erklärte Varianz als zusätzliche Tabelle
varianz_df <- as.data.frame(t(fa_result$Vaccounted))
knitr::kable(varianz_df, digits = 2, caption = "Erklärte Varianz pro Faktor (SS Loadings, Anteil und ku
```

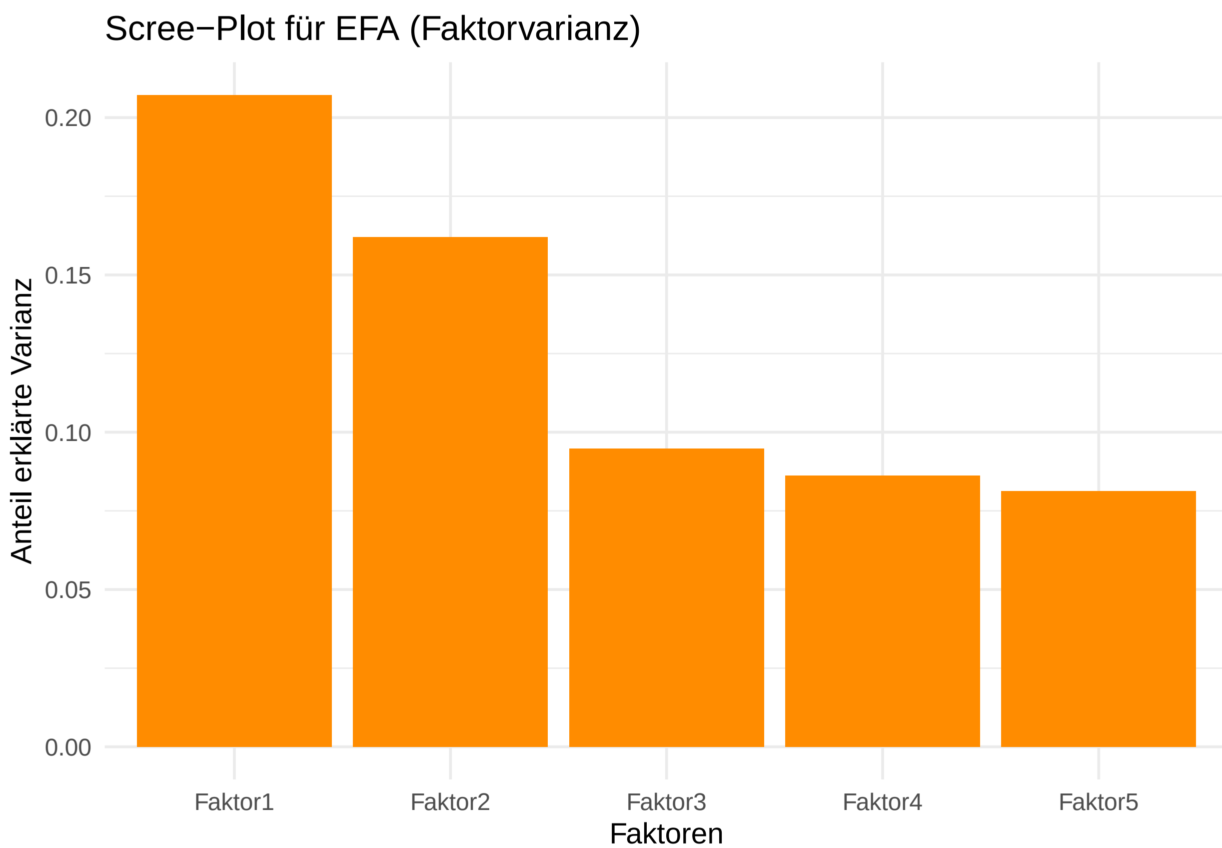
Table 8: Erklärte Varianz pro Faktor (SS Loadings, Anteil und kumuliert)

| | SS loadings | Proportion Var | Cumulative Var | Proportion Explained | Cumulative Proportion |
|---|---|---|---|---|---|
| MR1 | 1.24 | 0.21 | 0.21 | 0.33 | 0.33 |
| MR5 | 0.97 | 0.16 | 0.37 | 0.26 | 0.58 |
| MR4 | 0.57 | 0.09 | 0.46 | 0.15 | 0.73 |
| MR3 | 0.52 | 0.09 | 0.55 | 0.14 | 0.87 |
| MR2 | 0.49 | 0.08 | 0.63 | 0.13 | 1.00 |

MR1 trägt mit 21 % (bzw. 33 % der erklärten Varianz) am meisten bei, gefolgt von MR5 mit 16 %. Die restlichen Faktoren liefern kleinere, aber ergänzende Beiträge. Die kumulative Varianz steigt gleichmässig an, was auf eine sinnvolle Aufteilung der Datenstruktur auf mehrere Komponenten hinweist.

### 5.9 Screeplot für EFA

```r
efa_variance <- fa_result$Vaccounted["Proportion Var", ]
efa_df <- data.frame(
  Faktor = paste0("Faktor", 1:5),
  Varianzanteil = as.numeric(efa_variance)
)

ggplot(efa_df, aes(x = Faktor, y = Varianzanteil)) +
  geom_bar(stat = "identity", fill = "darkorange") +
  ylab("Anteil erklärte Varianz") +
  xlab("Faktoren") +
  ggtitle("Scree-Plot für EFA (Faktorvarianz)") +
  theme_minimal()
```




Grafische Darstellung der im vorherigen Kapitel aufgeführten Faktorvarianz.

### 5.10 Interpretation der Ergebnisse:

MR1 steht wie bei PCA für Temperatur- /Solar-/Erneuerbare-Energie-Komponenten: Hohes Gewicht bei Temperatur (0.76), Erneuerbare_Energien (0.68), PV_Leistung (0.40). MR5 ist stark geprägt von Netzbezogene_Leistung (0.81) und negativ von Temperatur (–0.51):

MR1 ist der wichtigste Faktor und erklärt alleine 21 % der Gesamtvarianz. Die ersten zwei Faktoren (MR1 und MR5) erklären zusammen bereits 58 % der Varianz. Ab MR3 nimmt der Erklärungsanteil deutlich ab, diese Faktoren tragen nur noch wenig neue Information bei. Die Gesamtvarianz ist gut verteilt, es liegt keine Dominanz weniger Faktoren vor.

# 6 Clusteranalyse zur Mustererkennung im Strommarkt

## 6.1 Theoretische Fundierung

Die theoretische Fundierung dieser Analyse basiert auf den Ausführungen bei Andreas Handl und Torben Kuhlenkasper, Multivariate Analysemethoden, 3. Aufl. 2017. Die Clusteranalyse dient der Identifikation von Gruppen ähnlicher Beobachtungen – hier: Zeitpunkte mit vergleichbaren meteorologischen und energiewirtschaftlichen Bedingungen. Ziel ist es, typische Tagesmuster in Bezug auf die Strompreisbildung zu identifizieren.

Verwendet wurde die K-Means-Methode, ein partitionierendes Verfahren, das besonders gut zur Gruppierung metrischer Variablen geeignet ist. Dabei werden die Beobachtungen so gruppiert, dass die Varianz innerhalb der Cluster minimiert und die Varianz zwischen den Clustern maximiert wird. Mathematisch geschieht dies durch Minimierung der Summe der quadrierten Abstände der Datenpunkte zu ihren jeweiligen Clusterzentren.

Da K-Means auf Distanzmassen (euklidischer Abstand) basiert, ist eine Standardisierung (Z-Transformation) der Variablen notwendig, um zu verhindern, dass Variablen mit grösseren Skalen dominieren. Diese Skalierung stellt sicher, dass alle Eingangsgrössen gleichgewichtet in die Clusterbildung eingehen.

Die Anzahl der Cluster wurde mithilfe der Elbow-Methode bestimmt, bei der der Verlauf der Gesamtvarianz (WSS) in Abhängigkeit von der Clusteranzahl betrachtet wird. Der „Knickpunkt" im Diagramm deutet auf eine sinnvolle Clusterzahl hin – in diesem Fall zwei Cluster.

In dieser Analyse wurden somit zwei Cluster gebildet: Cluster 1 beschreibt Tage mit hohen Strompreisen, die durch geringe Wind- und PV-Leistung gekennzeichnet sind. Cluster 2 repräsentiert Tage mit niedrigen Strompreisen, an denen vor allem hohe PV-Leistung zu einem Überangebot führt.

Die Qualität der Clusterlösung wurde zusätzlich durch den Silhouetten-Koeffizienten beurteilt, der misst, wie gut einzelne Beobachtungen ihrer zugewiesenen Gruppe zugeordnet sind. Werte nahe 1 deuten auf eine klare Zuordnung hin, während Werte nahe 0 auf Überschneidungen zwischen Clustern hinweisen.

## 6.2 Datenvorbereiten

Daten einlesen und vorbereiten

```
setwd(paste0(
  "C:/Users/41788/OneDrive/CAS Statische Datenalyse & Datenvisualisierung/",
  "Skript R/01 FFHS Semesterarbeit Statistische Datenanalyse"
))
data <- read_excel("C:/Users/41788/OneDrive/CAS Statische Datenalyse & Datenvisualisierung/Skript R/01 F
names(data) <- c("Datum", "Strompreis", "Temperatur", "Windgeschwindigkeit",
                 "Windenergieerzeugung", "PV_Leistung", "Erneuerbare_Energien", "Netzbezogene_Leistung")
data <- data %>%
  mutate(
    Datum = as.POSIXct(Datum, format = "%d.%m.%y %H:%M"),
    Stunde = hour(Datum)
  ) %>%
  filter(Stunde >= 8 & Stunde <= 17)
```

Datenbereinigung und Umwandlung

```
data <- data %>%
  mutate(across(c(Strompreis, Temperatur, Windgeschwindigkeit,
                  Windenergieerzeugung, PV_Leistung,
```

```r
                  Erneuerbare_Energien, Netzbezogene_Leistung),
                ~ as.numeric(gsub(",", ".", as.character(.))))) %>%
  na.omit()

mittelwert_gesamt <- round(mean(data$Strompreis, na.rm = TRUE), 2)
cat("\nDurchschnittlicher Strompreis über alle Tagesstunden (8-17 Uhr):", mittelwert_gesamt, "CHF/MWh\n
```

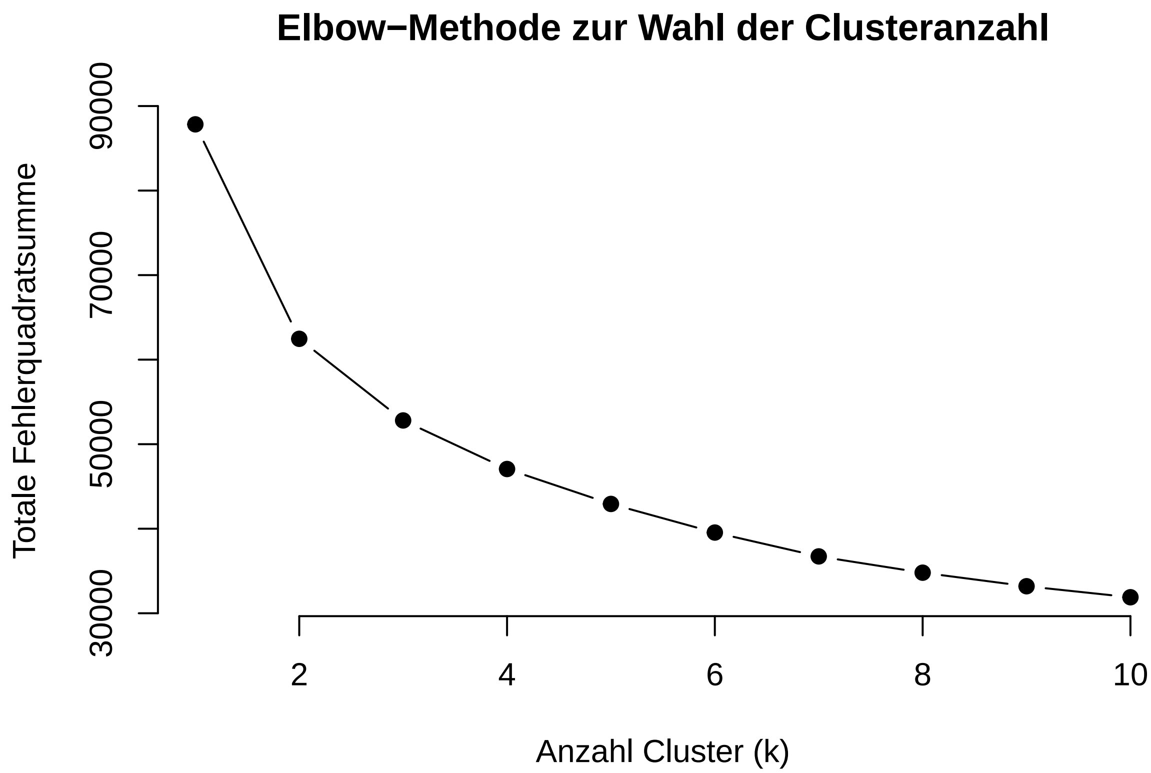
```
##
## Durchschnittlicher Strompreis über alle Tagesstunden (8-17 Uhr): 65.81 CHF/MWh
```

## 6.3 Durchführung der K-Means-Clusteranalyse

```r
cluster_vars <- data %>%
  select(Temperatur, Windgeschwindigkeit, Windenergieerzeugung,
         PV_Leistung, Erneuerbare_Energien, Netzbezogene_Leistung)
cluster_scaled <- scale(cluster_vars)
```

## 6.4 Elbow-Methode zur Clusterwahl

```r
wss <- numeric(10)
for (i in 1:10) {
  wss[i] <- sum(kmeans(cluster_scaled, centers = i, nstart = 10)$withinss)
}

plot(1:10, wss, type = "b", pch = 19, frame = FALSE,
     xlab = "Anzahl Cluster (k)", ylab = "Totale Fehlerquadratsumme",
     main = "Elbow-Methode zur Wahl der Clusteranzahl")
```




Anzahl Cluster sollte auf 2 festgelegt werden aufgrund des Knicks in Plot.

### 6.5 Clusteranalyse & Clusterzuweisung

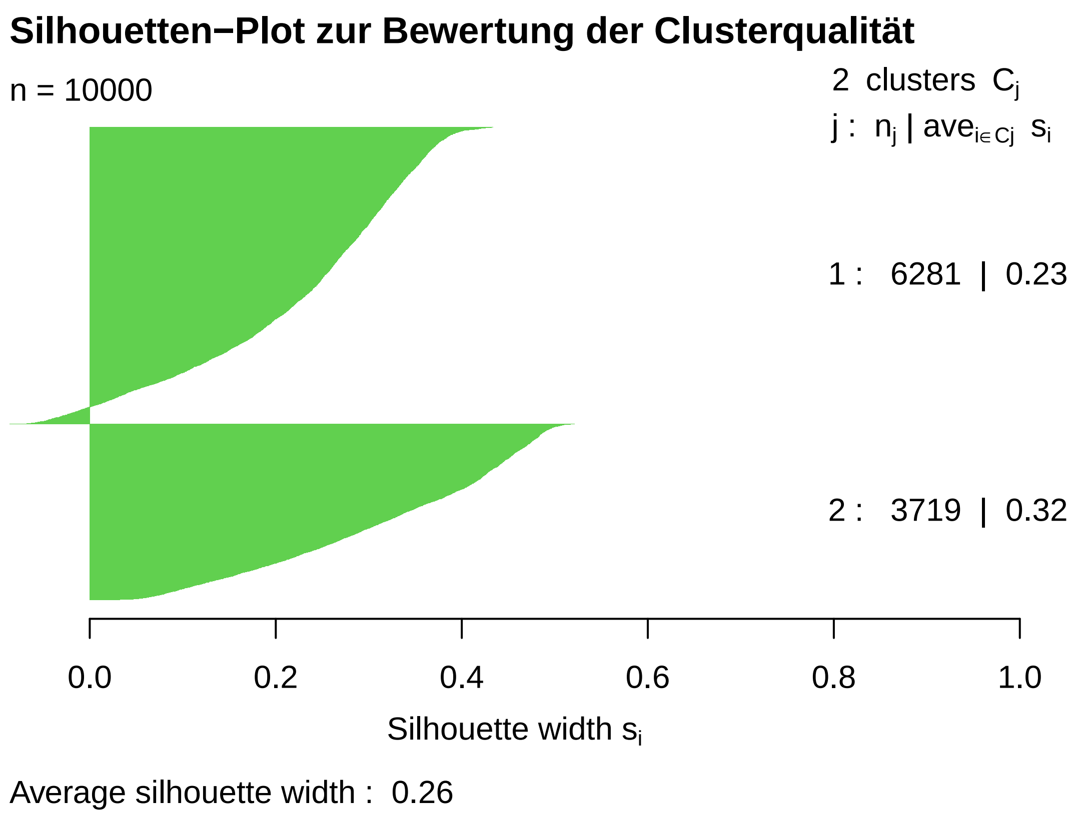
```
unique_rows <- nrow(unique(cluster_scaled))
if (unique_rows >= 2) {
  set.seed(123)
  kmeans_result <- kmeans(cluster_scaled, centers = 2, nstart = 25)
  data$Cluster <- as.factor(kmeans_result$cluster)
}
```

### 6.6 Bewertung der Clusterqualität mit dem Silhouetten-Score

```
# Nur 10000 zufällige Punkte für schnellere Silhouetten-Berechnung
sample_size <- min(10000, nrow(cluster_scaled))
sample_idx <- sample(nrow(cluster_scaled), sample_size)
silhouette_values <- silhouette(kmeans_result$cluster[sample_idx],
                                 dist(cluster_scaled[sample_idx, ]))

avg_silhouette_score <- mean(silhouette_values[, 3])
cat("\nDurchschnittlicher Silhouetten-Score:", round(avg_silhouette_score, 3), "\n")
```

```
##
## Durchschnittlicher Silhouetten-Score: 0.262
```

```r
plot(silhouette_values, border = NA,
     col = 2:length(unique(kmeans_result$cluster)) + 1,
     main = "Silhouetten-Plot zur Bewertung der Clusterqualität")
```




Die Silhouettenanalyse zeigt, dass die Clusterqualität insgesamt mässig ist – mit einem durchschnittlichen Silhouettenwert von 0.26. Cluster 2 ist klarer abgegrenzt (Ø = 0.32) als Cluster 1 (Ø = 0.22), was auf eine bessere Kohärenz innerhalb von Cluster 2 hindeutet. Insgesamt lässt sich erkennen, dass die gewählte Clusterlösung brauchbar, aber noch verbesserungswürdig ist.

### 6.7 Cluster-Zusammenfassung

```r
cluster_summary <- data %>%
  group_by(Cluster) %>%
  summarise(
    Mittelwert_Strompreis = round(mean(Strompreis), 2),
    Mittelwert_Wind = round(mean(Windgeschwindigkeit), 2),
    Mittelwert_PV = round(mean(PV_Leistung), 2)
  )
knitr::kable(cluster_summary, caption = "Cluster-Zusammenfassung")
```

Table 9: Cluster-Zusammenfassung

| Cluster | Mittelwert_Strompreis | Mittelwert_Wind | Mittelwert_PV |
|---|---|---|---|
| 1 | 82.88 | 0.84 | 848.12 |
| 2 | 36.94 | 0.82 | 2661.96 |

### 6.8 Clusterzentren (Z-standardisiert)

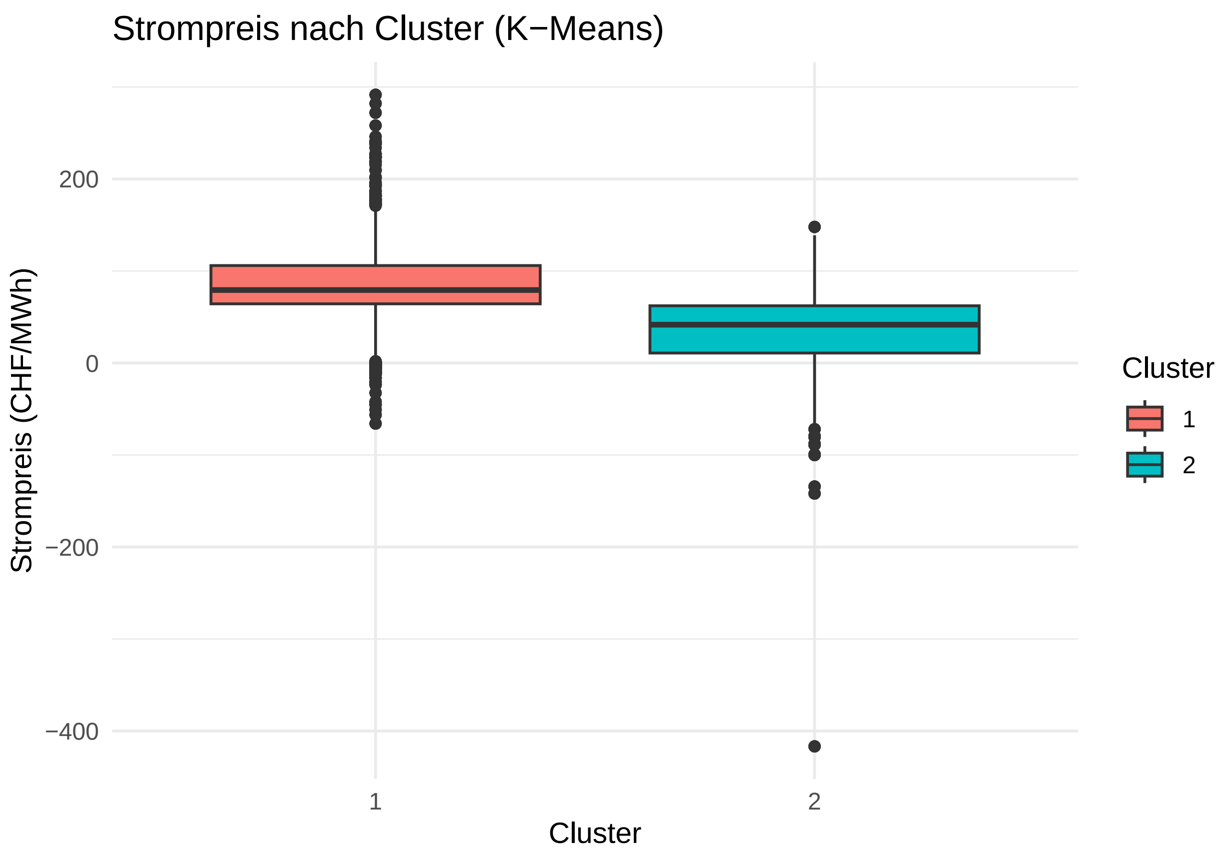
```
cluster_means <- aggregate(cluster_scaled, by = list(Cluster = kmeans_result$cluster), FUN = mean)
colnames(cluster_means) <- c("Cluster", "Temp", "Wind", "Wind_Erz", "PV", "EE", "Netz")
knitr::kable(cluster_means, caption = "Z-standardisierte Clusterzentren")
```

Table 10: Z-standardisierte Clusterzentren

| Cluster | Temp | Wind | Wind_Erz | PV | EE | Netz |
|---|---|---|---|---|---|---|
| 1 | -0.5968221 | 0.0287688 | 0.2427667 | -0.5553485 | -0.4356846 | 0.3328086 |
| 2 | 1.0090364 | -0.0486388 | -0.4104412 | 0.9389176 | 0.7366040 | -0.5626735 |

### 6.9 Boxplot Strompreis je Cluster

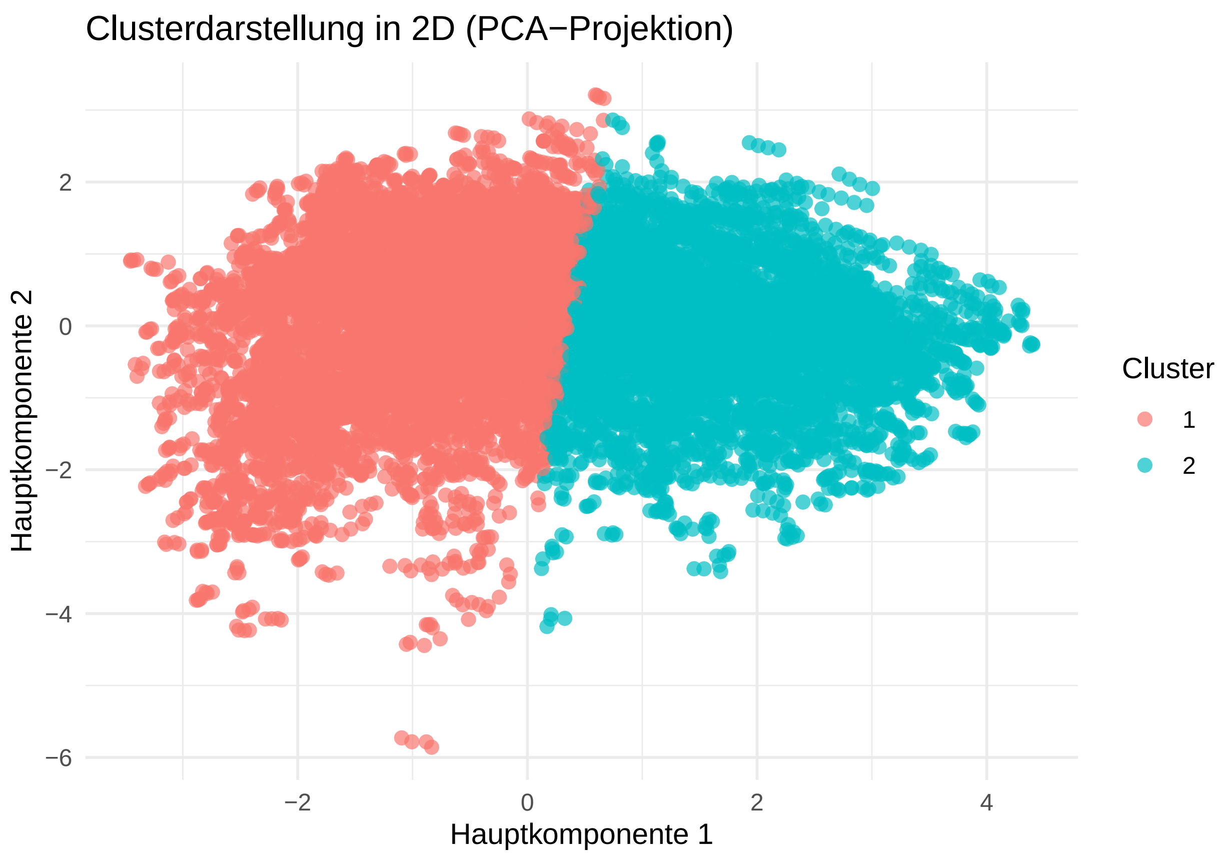
```
ggplot(data, aes(x = Cluster, y = Strompreis, fill = Cluster)) +
  geom_boxplot() +
  labs(title = "Strompreis nach Cluster (K-Means)",
       x = "Cluster", y = "Strompreis (CHF/MWh)") +
  theme_minimal()
```




Auch wenn in der Grafik nicht so gut ersichtlich. Strompreismedian von Cluster 2 ist deutlich tiefer. Cluster 1: 82.88 CHF/MWh Cluster 2: 36.94 CHF/MWh

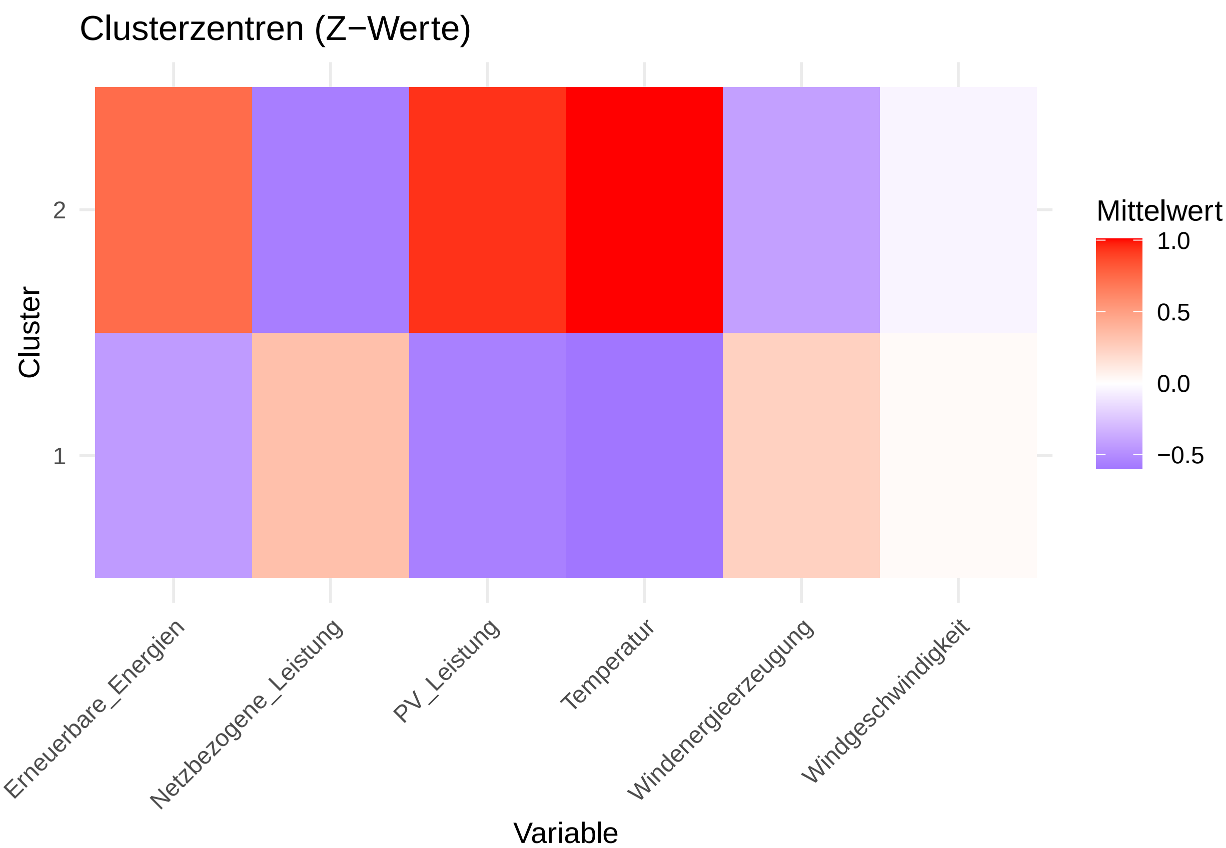
## 6.10 Clusterdarstellung in 2D (PCA)

```
pca_result <- prcomp(cluster_scaled, center = TRUE, scale. = TRUE)
pca_data <- as.data.frame(pca_result$x[, 1:2])
pca_data$Cluster <- data$Cluster

ggplot(pca_data, aes(x = PC1, y = PC2, color = Cluster)) +
  geom_point(alpha = 0.7, size = 2) +
  labs(title = "Clusterdarstellung in 2D (PCA-Projektion)",
       x = "Hauptkomponente 1", y = "Hauptkomponente 2") +
  theme_minimal()
```




## 6.11 Clusterzentren als Heatmap (Z-Werte)

```
cluster_scaled_df <- as.data.frame(cluster_scaled)
cluster_scaled_df$Cluster <- as.factor(kmeans_result$cluster)

heat_data <- cluster_scaled_df %>%
  group_by(Cluster) %>%
  summarise(across(where(is.numeric), mean)) %>%
  pivot_longer(-Cluster, names_to = "Variable", values_to = "Mittelwert")

ggplot(heat_data, aes(x = Variable, y = Cluster, fill = Mittelwert)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
  labs(title = "Clusterzentren (Z-Werte)", x = "Variable", y = "Cluster") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




## 6.12 Beschreibung der Cluster

**Cluster 1: Hoher Strompreis – wenig Wind und Sonne -** Unterdurchschnittlicher Wind und PV-Leistung
- Netzbezogene Leistung ist durchschnittlich
- Hohes Preisniveau, da auf der Erzeugerseite geringe Einspeisung

**Cluster 2: Tiefer Strompreis – viel PV -** Überdurchschnittliche PV-Leistung
- Windleistung leicht unterdurchschnittlich
- Netzbezogene Leistung gering
- Tiefer Strompreis → Überangebot

## 6.13 Interpretation der Ergebnisse:

Die Clusteranalyse hat erfolgreich zwei Tagesmuster im Strommarkt identifiziert. Diese Muster unterscheiden sich deutlich hinsichtlich der erneuerbaren Einspeisung (Wind und PV) und wirken sich sichtbar auf das Strompreisniveau aus. Cluster 2 liegt mit einem Mittelwert von nur 36.9 CHF/MWh deutlich unter dem Gesamtdurchschnitt von 65.81 CHF/MWh, was den preisdämpfenden Effekt von PV-Einspeisung eindrucksvoll zeigt. Die Analyse zeigt, dass erneuerbare Energien, vor allem PV, eine zentrale Rollein der kurzfristigen Strompreisbildung spielen. Gleichzeitig macht Cluster 1 deutlich, dass hohe Windleistung allein nicht zwangsläufig zu günstigen Preisen führt.

# 7 Zeitreihenanalyse

## 7.1 Theoretische Fundierung: Zeitreihenanalyse im Strommarkt

Die theoretische Fundierung dieser Analyse basiert auf den Ausführungen bei Andreas Handl und Torben Kuhlenkasper, Multivariate Analysemethoden, 3. Aufl. 2017.

Strompreise zeigen typische Zeitreihenmuster: langfristige Trends, saisonale Zyklen (z. B. Tagesverlauf) sowie kurzfristige Volatilität durch Angebot-Nachfrage-Schwankungen. Zur Modellierung solcher Daten werden stochastische Modelle eingesetzt, die Strukturen im zeitlichen Verlauf erfassen können.

Klassische AR- (autoregressive), MA- (moving average) und ARMA-Modelle eignen sich für stationäre Prozesse, bei denen Mittelwert und Varianz über die Zeit konstant bleiben. Für nicht-stationäre Zeitreihen – wie Strompreise mit erkennbaren Trends – wird das ARIMA-Modell verwendet. Dieses erweitert ARMA um Differenzierungsschritte, um Trends aus den Daten zu entfernen und Stationarität herzustellen.

Da Strompreise zudem ausgeprägte saisonale Muster (z.B. tägliche oder wöchentliche Rhythmen) aufweisen, kommt in dieser Analyse das SARIMA-Modell (Seasonal ARIMA) zum Einsatz. Dieses Modell integriert zusätzlich saisonale AR- und MA-Terme und erlaubt somit die Modellierung periodischer Schwankungen – hier z.B. auf Basis von 96 Viertelstunden pro Tag (entsprechend der saisonalen Frequenz).

Mathematisch basiert SARIMA auf der Kombination mehrerer differenzierter autoregressiver Komponenten mit saisonalen Effekten. Die Modellparameter werden datenbasiert geschätzt und über bewertet.

Ein ergänzender Bestandteil der Analyse ist die Zerlegung der Zeitreihe in ihre Komponenten (Trend, Saisonalität, Rest), um die Struktur visuell und analytisch zu untersuchen. Die Auswahl geeigneter Modellparameter erfolgt dabei automatisiert (via auto.arima()), wobei sowohl Drift als auch saisonale Differenzierung berücksichtigt werden.

## 7.2 Datenvorbereiten

Daten einlesen

```
setwd(paste0(
  "C:/Users/41788/OneDrive/CAS Statische Datenalyse & Datenvisualisierung/",
  "Skript R/01 FFHS Semesterarbeit Statistische Datenanalyse"
))
data <- read_excel("C:/Users/41788/OneDrive/CAS Statische Datenalyse & Datenvisualisierung/Skript R/01 F
```

Spaltennamen bereinigen

```
colnames(data) <- gsub(" ", "_", colnames(data))
colnames(data) <- gsub("[^[:alnum:]_]", "", colnames(data))
```

Umwandeln relevanter Spalten in numerisch

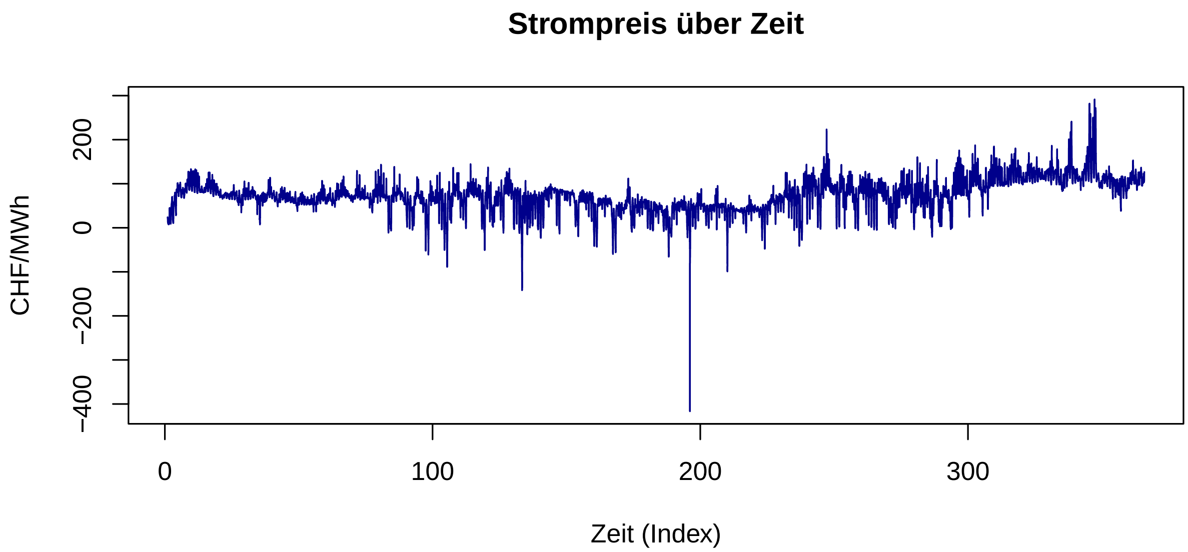
```
numeric_vars <- c("Strompreis", "Temperatur", "Windgeschwindigkeit",
                  "Windenergieerzeugung", "PV_Leistung", "Erneuerbare_Energien")
data[numeric_vars] <- lapply(data[numeric_vars], function(x) as.numeric(gsub("[^0-9.-]", "", x)))

data <- na.omit(data)
data$Zeitindex <- 1:nrow(data)
```

Zeitreihe erzeugen

```r
strom_ts <- ts(data$Strompreis, frequency = 96)
```

### 7.3 Zeitverlauf des Strompreises

```r
plot(strom_ts,
     main = "Strompreis über Zeit",
     ylab = "CHF/MWh", xlab = "Zeit (Index)",
     col = "darkblue")
```

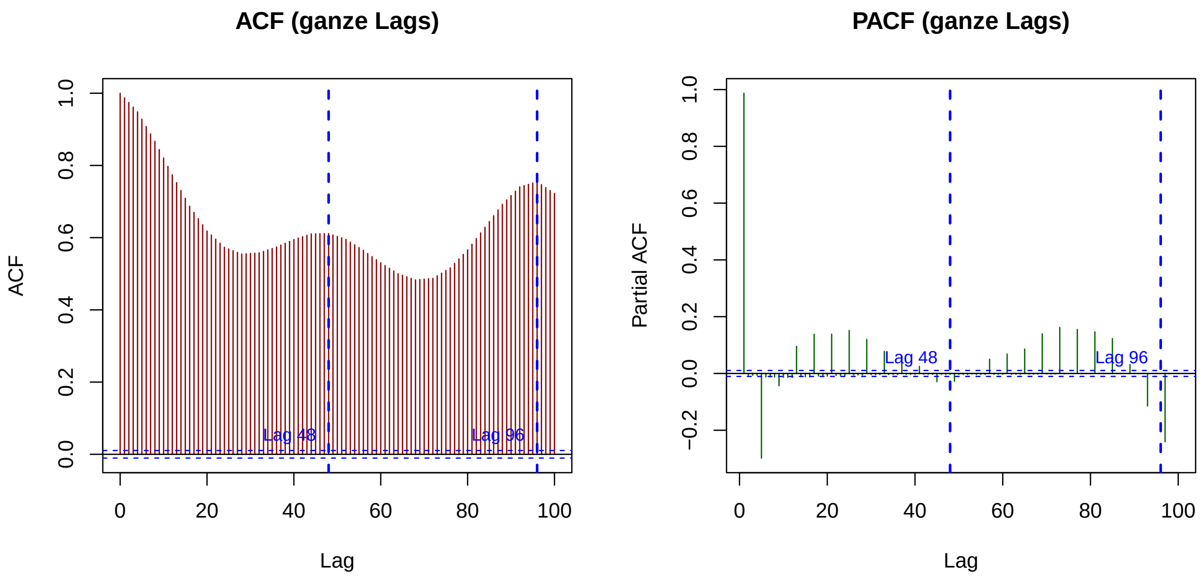



Die Abbildung zeigt den zeitlichen Verlauf des Strompreises in CHF/MWh über einen kontinuierlichen Zeitraum mit 15-Minuten-Intervallen.
Im ersten Abschnitt der Zeitreihe zeigt der Strompreis eine relativ stabile Entwicklung mit kleineren Schwankungen. Etwa in der Mitte fällt ein deutlicher negativer Ausreisser auf, bei dem der Strompreis unter –400 CHF/MWh sinkt.
Gegen Ende der Zeitreihe ist der Durchschnitt höher als sonst.

### 7.4 Autokorrelation und partielle Autokorrelation (nicht sinvoll)

```r
strom_ts_acf <- ts(data$Strompreis)  # frequency weglassen
par(mfrow = c(1,2))

## ACF mit Beschriftung bei Lag 48 und 96
acf(strom_ts_acf, lag.max = 100, main = "ACF (ganze Lags)", col = "darkred")
abline(v = c(48, 96), col = "blue", lty = 2, lwd = 2)
text(x = 48, y = 0.05, labels = "Lag 48", col = "blue", pos = 2, cex = 0.8)
text(x = 96, y = 0.05, labels = "Lag 96", col = "blue", pos = 2, cex = 0.8)

## PACF mit Beschriftung bei Lag 48 und 96
pacf(strom_ts_acf, lag.max = 100, main = "PACF (ganze Lags)", col = "darkgreen")
```

```r
abline(v = c(48, 96), col = "blue", lty = 2, lwd = 2)
text(x = 48, y = 0.05, labels = "Lag 48", col = "blue", pos = 2, cex = 0.8)
text(x = 96, y = 0.05, labels = "Lag 96", col = "blue", pos = 2, cex = 0.8)
```

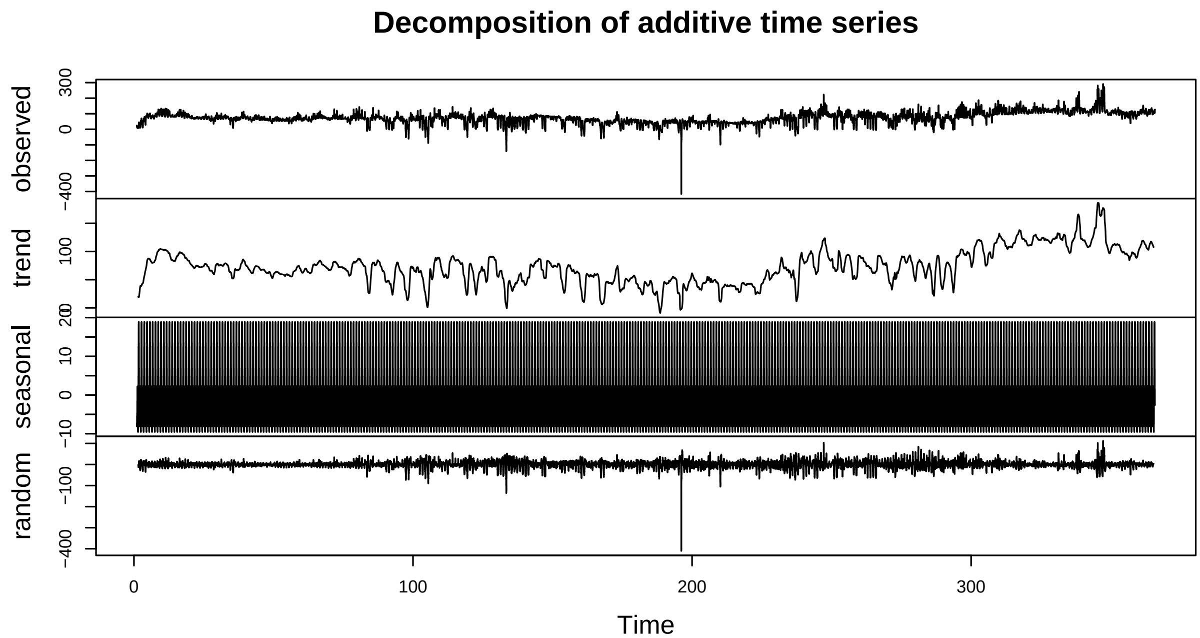



```r
par(mfrow = c(1, 2), mar = c(5, 4, 4, 4))
```

Die obigen Abbildungen zeigen die Autokorrelationsfunktion (ACF) und die partielle Autokorrelationsfunktion (PACF) der Strompreiszeitreihe bei 15-minütiger Auflösung. Die X-Achse zeigt die Verzögerung (Lag) in Anzahl der Zeitpunkte, wobei jeweils 96 Lags einem vollen Tag entsprechen.
In der ACF fällt auf, dass die Korrelation sehr langsam abnimmt und auch bei höheren Lags signifikant bleibt. Da es sich um nicht-stationäre Zeitreihe handelt, ist diese zu erwarten. Zudem ist ein deutlich periodisches, wellenförmiges Muster erkennbar. Besonders auffällig sind die Autokorrelationsspitzen bei Lag 48 und Lag 96. Diese entsprechen 12 bzw. 24 Stunden und zeigen, dass der Strompreis einem typischen Tageszyklus unterliegt. Das ist bei Strompreisen üblich, da die Nachfrage sowie die Einspeisung aus erneuerbaren Energien (z.B. Photovoltaik) einem stark tageszeitabhängigen Verlauf folgen.
Die PACF zeigt eine hohe partielle Korrelation bei Lag 1, was auf einen autoregressiven Zusammenhang erster Ordnung hinweist. Bei höheren Lags, insbesondere bei Lag 48 und 96 (12 bzw. 24 Stunden), sind die Werte hingegen nicht signifikant. Es besteht somit kein direkter saisonaler autoregressiver Zusammenhang.

## 7.5 Komponentenmodell: Zerlegung in Trend, Saison, Rest

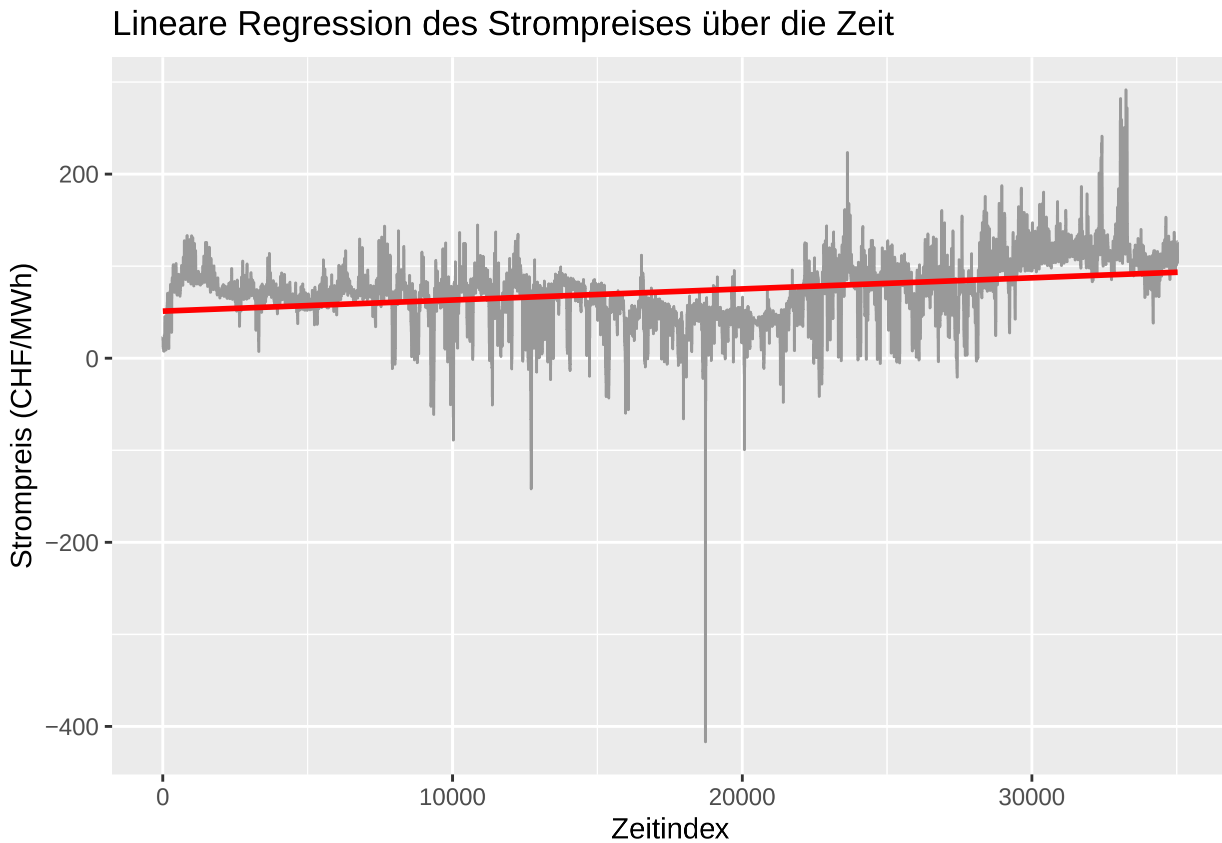
```r
strom_ts_daily <- ts(data$Strompreis, frequency = 96)
decomposed <- decompose(strom_ts_daily, type = "additive")
plot(decomposed)
```




Die Strompreisreihe ist nicht stationär, weil sie sowohl einen klaren Trend als auch starke saisonale Schwankungen aufweist. Das bedeutet, dass die Daten vor einer Modellierung angepasst werden müssen – z.B. durch Differenzierung (für den Trend) und Saisonbereinigung (für das Tagesmuster).

### 7.6 Zeittrendanalyse mit linearer Regression

```r
modell_trend <- lm(Strompreis ~ Zeitindex, data = data)
library(ggplot2)
ggplot(data, aes(x = Zeitindex, y = Strompreis)) +
  geom_line(color = "grey60") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Lineare Regression des Strompreises über die Zeit",
       x = "Zeitindex", y = "Strompreis (CHF/MWh)")
```




Im Mittel nimmt der Strompreis im Jahresverlauf zu. Die tatsächlichen Strompreise schwanken jedoch stark und weisen eine hohe Volatilität auf. Dies lässt vermuten, dass neben einem leichten Trend auch kurzfristige und saisonale Effekte eine wichtige Rolle spielen.

## 7.7 SARIMA-Modell

Die Berechnung des SARIMA-Modells wurde nicht erneut ausgeführt, da sie sehr lange dauert.Da Strompreise starke saisonale Muster bspw. tägliche Schwankungen zeigen, reicht ein klassisches ARIMA-Modell nicht aus. SARIMA erweitert ARIMA um saisonale Komponenten und kann so regelmässige Zyklen explizit abbilden.Für Strompreisdaten mit starker Tagesstruktur ist SARIMA daher besser geeignet.

## 7.8 Automatische SARIMA-Modellierung

```
#strom_sarima_full <- auto.arima(
 # strom_ts,
  #seasonal = TRUE,
  #stepwise = TRUE,
  #approximation = TRUE
)

Ergebnisse:
ma1 = 0.8714: Der Strompreis reagiert stark auf zufällige Schocks der Vorperiode.
drift = 0.0025: Es besteht ein leichter langfristiger Aufwärtstrend im Preis.
RMSE = 14.64 / MAE = 8.96: Die Prognosefehler sind moderat, aber noch deutlich.
ACF1 = 0.69: Die Fehler sind stark autokorreliert - das Modell erklärt nicht alle Strukturen.
```

MA1 = 0.8714 Das Modell weist einen starken Moving-Average-Effekt erster Ordnung auf. Der aktuelle Strompreis reagiert also deutlich auf kurzfristige Zufallsschwankungen aus der Vorperiode.

Drift = 0.0025 Es besteht ein leichter langfristiger Aufwärtstrend im Strompreis.

RMSE = 14.64 / MAE = 8.96 → Die Prognosegüte ist moderat: Die Fehler liegen durchschnittlich bei ca. 9 CHF/MWh (MAE). Die RMSE ist etwas höher, was auf gelegentliche grössere Ausreisser hindeutet.

ACF1 = 0.69 (Autokorrelation der Residuen bei Lag 1) → Die Residuen sind deutlich autokorreliert, was darauf hinweist, dass das Modell nicht alle zeitlichen Strukturen vollständig erklärt.

Es könnten noch saisonale oder autoregressive Komponenten fehlen, die das Modell verbessern würden. Das Modell bildet saisonale Effekte gut ab, lässt aber noch Verbesserungspotenzial offen (SARIMAX).

## 7.9 Interpretation der Ergebnisse

Die Analyse untersucht Strompreise im 15-Minuten-Takt über ein Jahr. Die Visualisierung zeigt klare Strukturen: einen extremen negativen Ausreisser, einen leichten langfristigen Aufwärtstrend sowie deutlich ausgeprägte tägliche Schwankungen.

Die Autokorrelationsfunktion (ACF) bestätigt diese Muster. Insbesondere die signifikanten Peaks bei Lag 48 und 96 (entsprechend 12 bzw. 24 Stunden) deuten auf einen starken saisonalen Tageszyklus hin – typisch für strompreisrelevante Nachfragemuster und PV-Erzeugung.

Auch die additive Zerlegung der Zeitreihe macht sowohl einen Trend als auch eine Tagesperiodik sichtbar. Eine einfache lineare Regression bestätigt den Aufwärtstrend, jedoch mit hoher Volatilität im Tagesverlauf.

Das automatisch geschätzte SARIMA-Modell weist einen deutlichen MA(1)-Effekt (0.871) auf – der Strompreis reagiert also stark auf Zufallsschwankungen der Vorperiode. Der geschätzte Drift (0.0025) bestätigt den beobachteten langfristigen Preisanstieg. Die Prognosegüte ist moderat (RMSE = 14.64, MAE = 8.96), jedoch zeigt die Autokorrelation der Residuen (ACF1 = 0.69), dass noch nicht alle zeitlichen Strukturen erfasst werden.

# 8 Redlichkeitserklärung

Der Verfasser erklärt hiermit, dass die vorliegende Arbeit selbstständig und ohne unzulässige Hilfe Dritter verfasst wurde;
dass alle wörtlich oder sinngemäss übernommenen Textstellen aus fremden Quellen (einschliesslich elektronischer Quellen)
als solche kenntlich gemacht und korrekt zitiert wurden;
dass alle mit Hilfsmitteln erbrachten Teile der Arbeit vollständig deklariert wurden und keine anderen als die im Verzeichnis
der Hilfsmittel aufgeführten Quellen oder Werkzeuge verwendet wurden;
dass das Thema sowie die Inhalte dieser Arbeit oder Teile davon nicht bereits als Leistungsnachweis in einem anderen Modul
eingereicht wurden – ausser in Fällen, in denen dies im Vorfeld ausdrücklich mit der betreuenden Lehrperson vereinbart wurde.

Der Verfasser ist sich darüber hinaus bewusst, dass die Arbeit elektronisch auf Plagiate sowie auf Drittautorschaft –
menschlichen oder technischen Ursprungs – überprüft werden kann, und räumt der FFHS hierfür die notwendigen Nutzungsrechte ein.

**Hinweis zur Hilfsmittelnutzung:**
Zur Unterstützung bei der sprachlichen Ausarbeitung, Strukturierung und Erstellung des R-Codes wurde das KI-gestützte Assistenzsystem *ChatGPT* von *OpenAI* verwendet.

Freitag, 20. Juni 2025

# 9 Literaturverzeichnis

Handl, A., & Kuhlenkasper, T. (2017). *Multivariate Analysemethoden: Theorie und Praxis mit R* (3., wesentlich überarb. Aufl.). Springer Gabler.

# 10 Anhang

## 10.1 Explorative Datenanalyse