Take Home Examination

Modul: Multivariate Geodatenanalyse WS 24/25
Bearbeitet von: Arvo Klöck | 909003
Variante: „Beschäftigte Tertiärer Sektor” // Bandbreitenverfügbarkeit mindestens 50 Mbit/s“

Vorbereitung der Entwicklungsumgebung

- **a) Setzen des Arbeitsverzeichnisses**
  Bitte vor dem Ausführen auf den Ordner anpassen, in dem diese Datei liegt.
- **b) Benötigten Pakete laden**
  Die Funktion `install_and_load` installiert/läd ein Paket, falls es noch nicht installiert ist.

```
knitr::opts_chunk$set(encoding = "UTF-8")

# a)
setwd("E:/srv/repos/spatial_correlation")

# b)
install_and_load <- function(package) {
  if (!require(package, character.only = TRUE)) {
    install.packages(package, dependencies = TRUE)
    library(package, character.only = TRUE)
  }
}
packages <- c("sf", "tmap", "dplyr", "classInt", "ggplot2","readr","tidyr")
invisible(lapply(packages, install_and_load))

print_installed_packages <- function(packages) {
  installed <- installed.packages()[, c("Package", "Version")]
  installed <- installed[installed[, "Package"] %in% packages, ]
  print(as.data.frame(installed))
}
print_installed_packages(packages)
```

```
##           Package Version
## classInt classInt  0.4-10
## dplyr       dplyr   1.1.4
## ggplot2   ggplot2   3.5.1
## readr       readr   2.1.5
## sf             sf  1.0-19
## tidyr       tidyr   1.3.1
## tmap         tmap   3.3-4
```

1. Aufgabenteil

- **a) CSV als Dataframe Objekt importieren**
- **b) Tabellenspalten Definieren**
- **c) Dataframe als RDS im Arbeitsverzeichnis speichern**

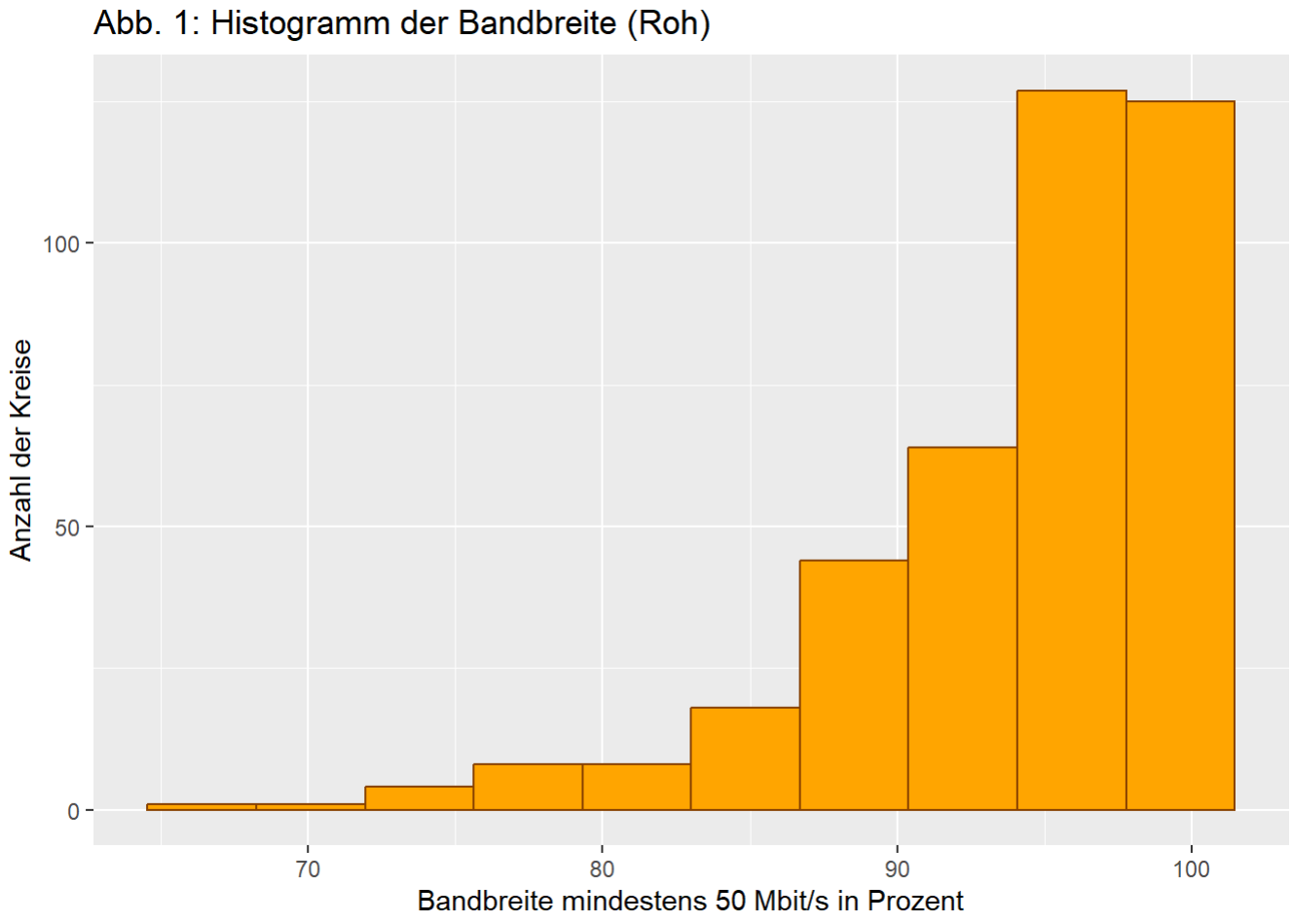
```
# a)
rohdaten <- as.data.frame(read_csv2("909003.csv", col_names = TRUE))
# b)
colnames(rohdaten) <- c("ID", "Region", "Typ", "Beschaeftigte", "Bandbreite")
# c)
saveRDS(rohdaten, file = "909003.rds")
```

2. Aufgabenteil

- **a) Klassenanzahl berechnen: Sturges-Formel**
  Angemessene Balance zwischen Übersichtlichkeit und Detailtreue
- **b) Histogramm der Bandbreite**
  `bins` ist die Anzahl der Klassen im Histogramm.

```
# a)
k_bandbreite <- ceiling(log2(length(rohdaten$Bandbreite)) + 1)

# b)
ggplot(rohdaten, aes(x = Bandbreite)) +
  geom_histogram(bins = k_bandbreite, fill = "orange", color = "darkorange4") +
  labs(title = "Abb. 1: Histogramm der Bandbreite (Roh)", x = "Bandbreite mindestens 50 Mbit/
s in Prozent", y = "Anzahl der Kreise")
```



- **c) Klassenanzahl berechnen: Sturges-Formel**
  Angemessene Balance zwischen Übersichtlichkeit und Detailtreue (Eigentlich redundant)
- **d) Histogramm der Beschäftigten**
  `bins` ist die Anzahl der Klassen im Histogramm.

```
# c)
k_beschaeftigte <- ceiling(log2(length(rohdaten$Beschaeftigte)) + 1)
# d)
ggplot(rohdaten, aes(x = Beschaeftigte)) +
  geom_histogram(bins = k_beschaeftigte, fill = "blue", color = "blue4") +
  labs(title = "Abb. 2: Histogramnm Beschäftigte Tertiärer Sektor ", x = "Beschäftigte Tertiä
rer Sektor", y = "Anzahl der Kreise")
```

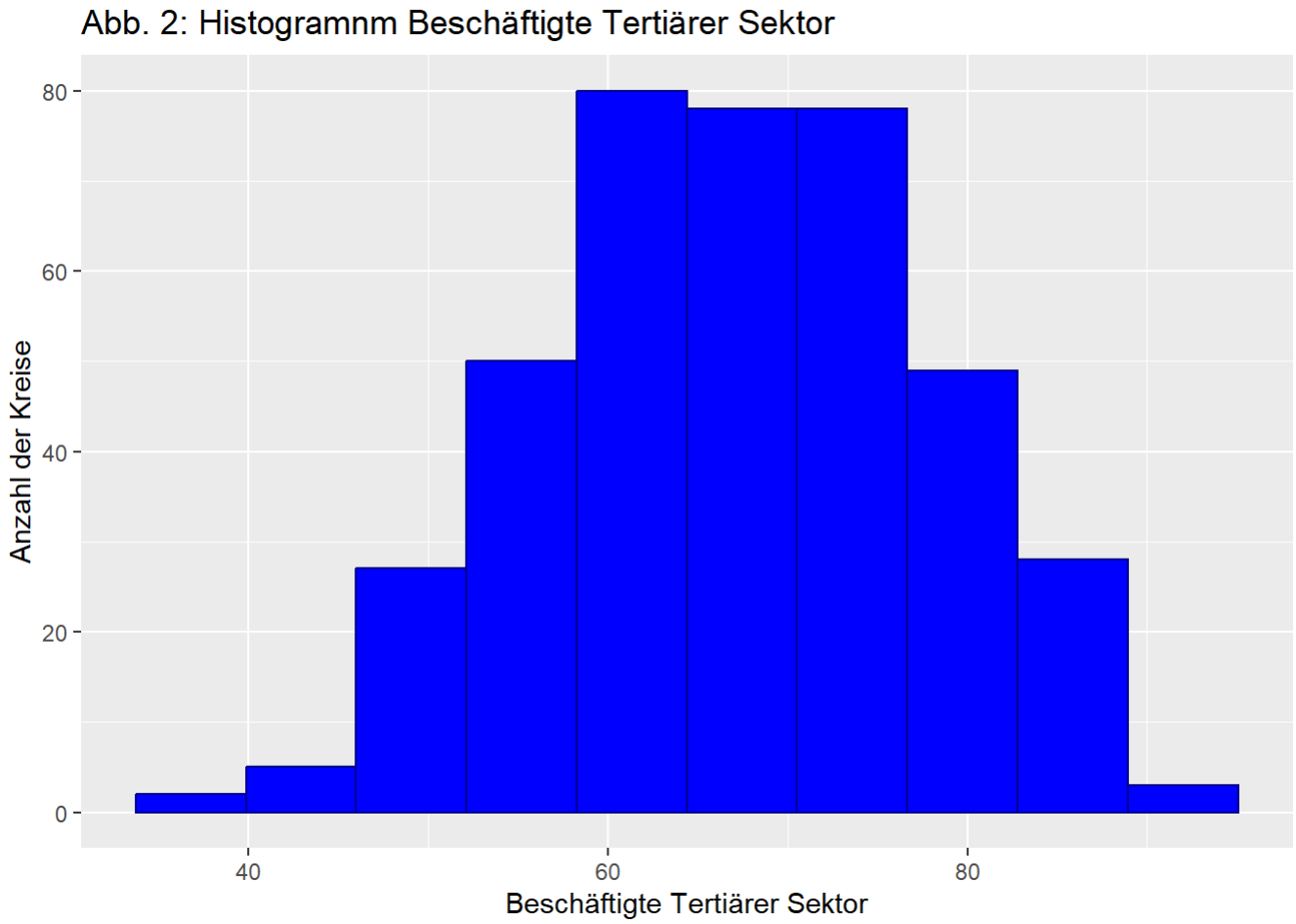


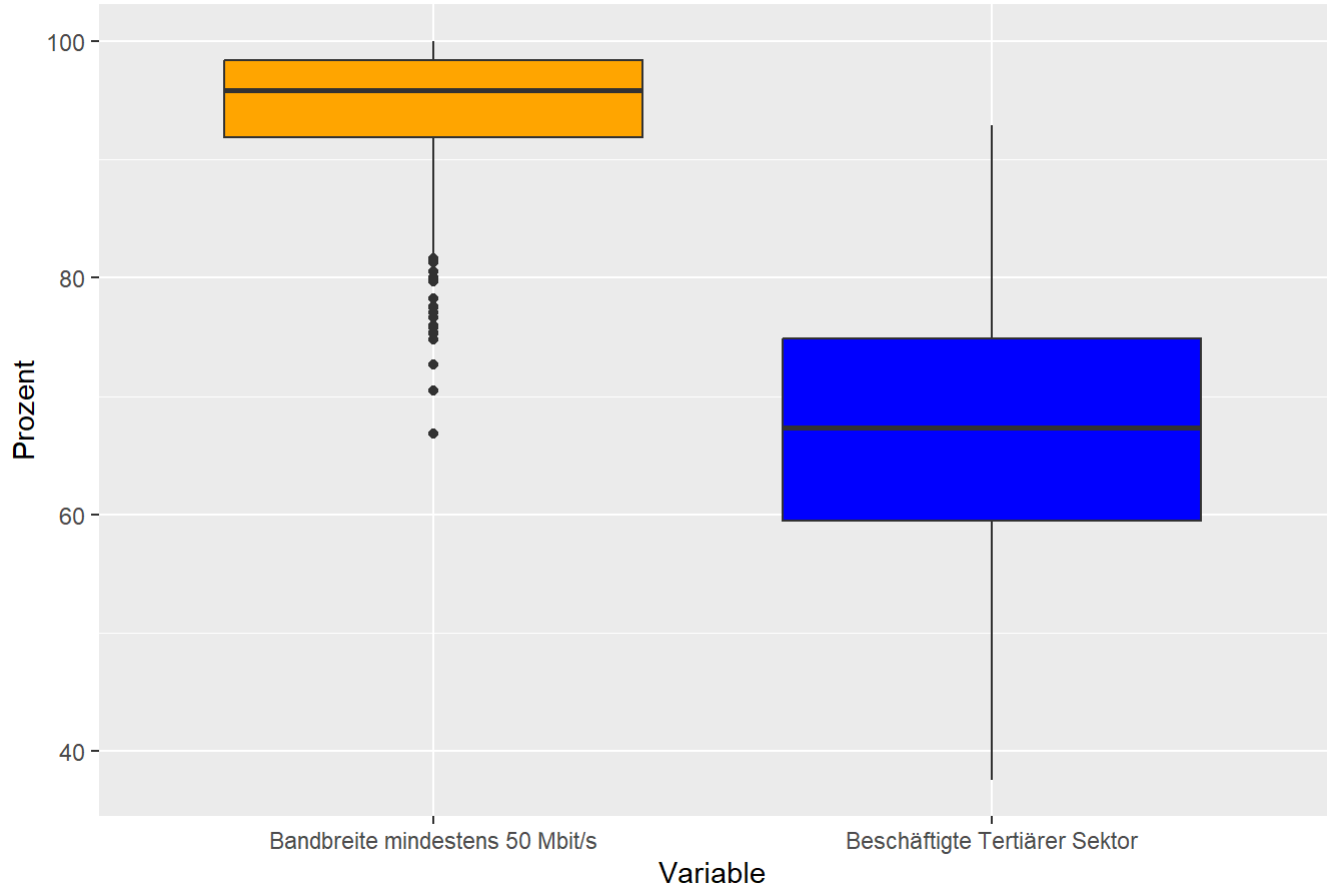
Abb 2.

- Normalverteilt
- Leicht rechtsschief

- **e) Boxplot der beiden Variablen**

```
# e)
ggplot(rohdaten %>% pivot_longer(cols = c(Bandbreite, Beschaeftigte), names_to = "Variable",
values_to = "Wert"), aes(x = Variable, y = Wert)) +
  geom_boxplot(fill = c("orange", "blue")) +
  labs(title = "Abb. 3: Boxplots der Rohdaten", x = "Variable", y = "Prozent") +
  scale_x_discrete(labels = c("Bandbreite mindestens 50 Mbit/s", "Beschäftigte Tertiärer Sekt
or"))
```

Abb. 3: Boxplots der Rohdaten



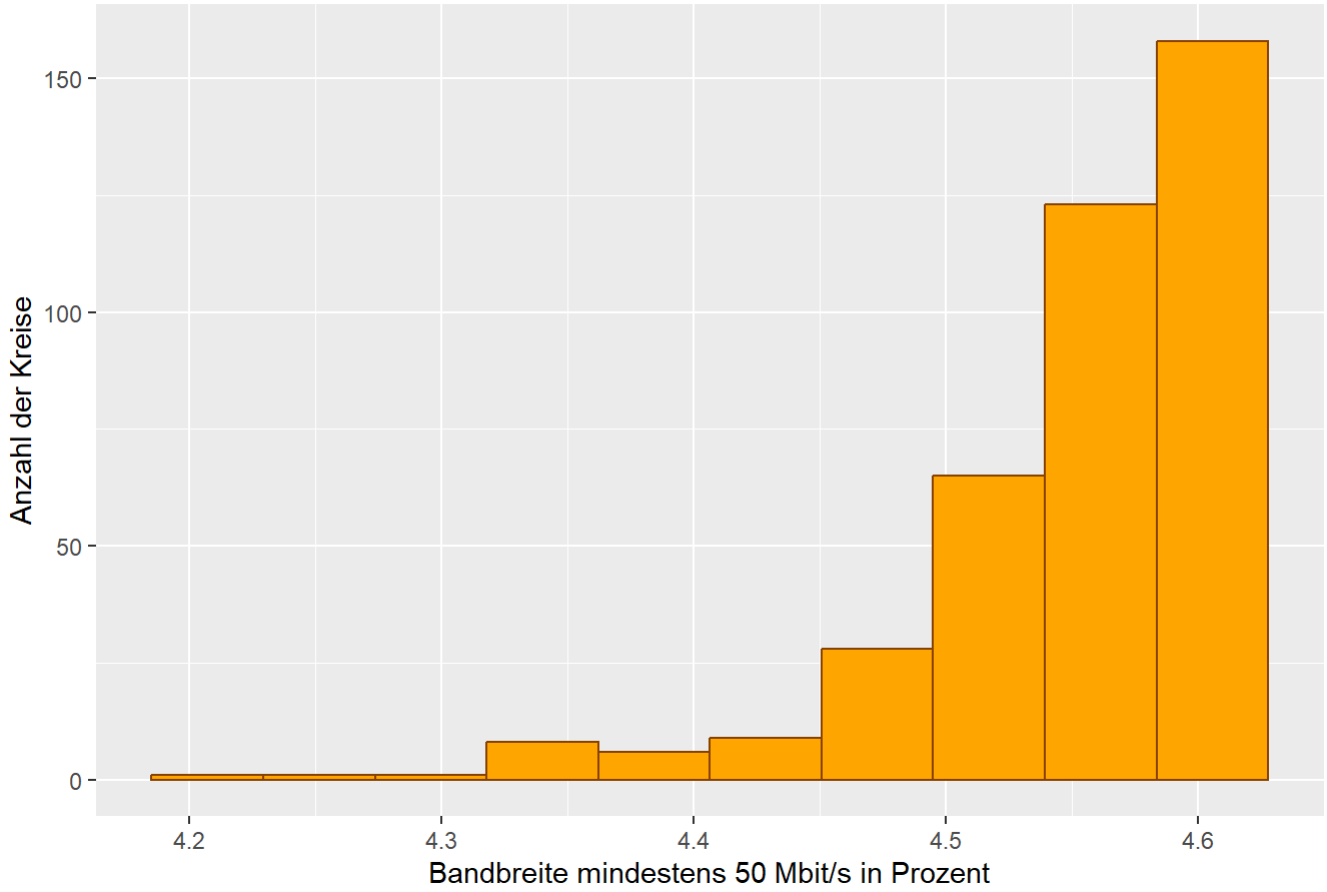
3. Aufgabenteil (Zusatzaufgabe)

a) Rechtsschiefe:

- Eine logarithmische Transformation reduziert die Schiefe reduzieren, indem sie die größeren Werte relativ stärker verkleinert als die kleineren Werte. **b) Daten weiter verwenden:**

```
# a)
ggplot(rohdaten, aes(x = log(Bandbreite+1))) +
  geom_histogram(bins = k_bandbreite, fill = "orange", color = "darkorange4") +
  labs(title = "Abb. 4: Korrekturversuch Rechtsschiefe", x = "Bandbreite mindestens 50 Mbit/s
in Prozent", y = "Anzahl der Kreise")
```

Abb. 4: Korrekturversuch Rechtsschiefe




```
# b)
#rohdaten <- rohdaten %>%
#  mutate(Bandbreite = log(Bandbreite + 1))
```

Abb. 4

> Die Schiefe kann durch die logarithmische Transformation reduziert werden.

4. Aufgabenteil

- **a) Prüfen auf NA-Werte**
  Bereinigung der Daten nicht notwendig (keine NA-Werte)
- **b) Bereinigung der Daten**
  Entfernen von Ausreißern (IQR-Methode)
- **c) Boxplot der bereinigten Daten**
  Boxplot der bereinigten Daten
- **d) Dataframe als RDS im Arbeitsverzeichnis speichern**
  Speichern des bereinigten Dataframes (Subset)

```
#  a)
na_columns <- sum(colSums(is.na(rohdaten)) > 0)
print(paste("Number of NA columns :", na_columns))
```

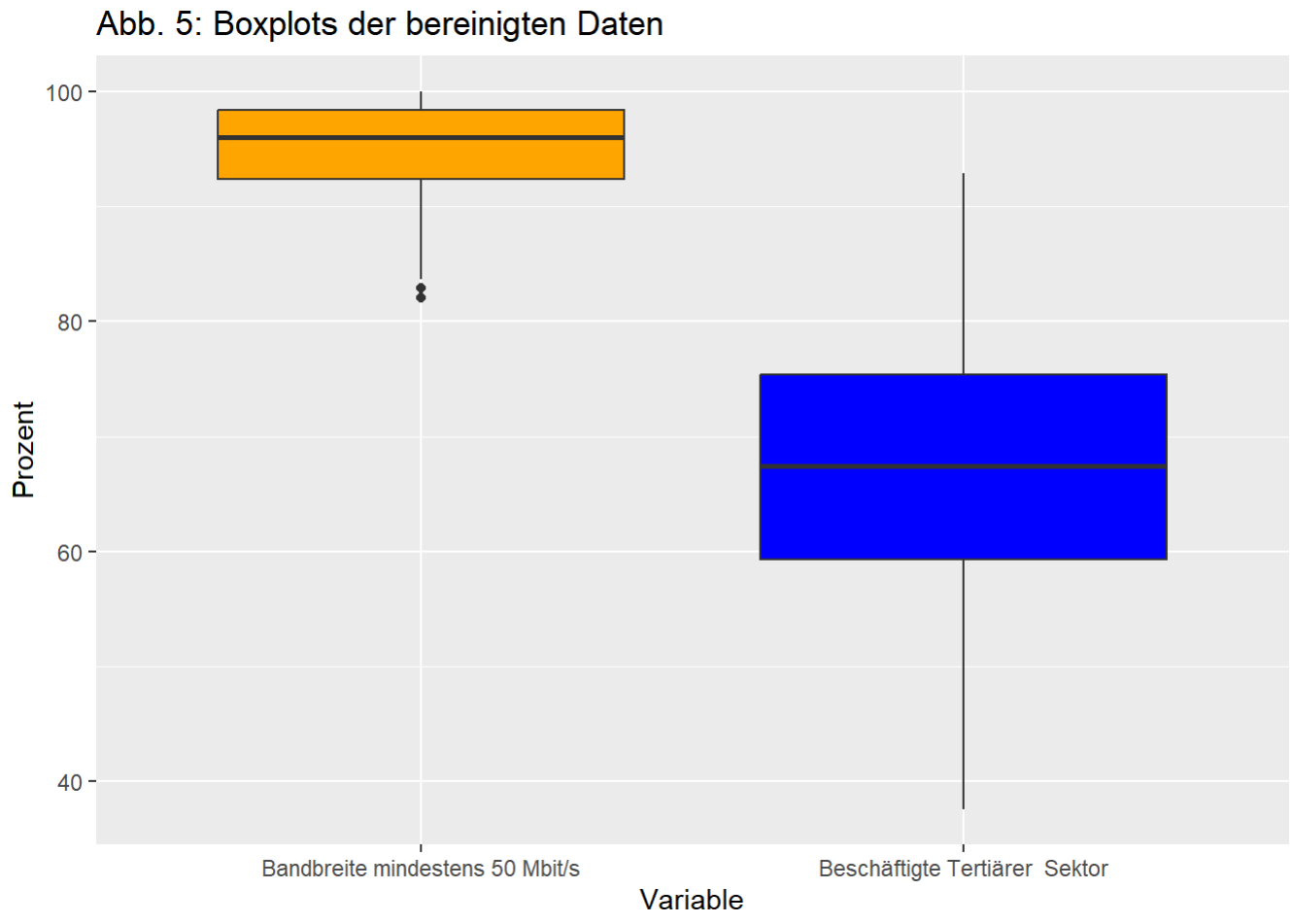
```
## [1] "Number of NA columns : 0"
```

```r
# b)
remove_outliers <- function(df, column) {
  q1 <- quantile(df[[column]], 0.25)
  q3 <- quantile(df[[column]], 0.75)
  iqr <- q3 - q1
  lower_bound <- q1 - 1.5 * iqr
  upper_bound <- q3 + 1.5 * iqr
  df %>%
    filter(.data[[column]] >= lower_bound & .data[[column]] <= upper_bound)
}

rohdaten_cleaned <- rohdaten %>%
  remove_outliers("Bandbreite") %>%
  remove_outliers("Beschaeftigte")


rohdaten_cleaned_long <- rohdaten_cleaned %>%
  pivot_longer(cols = c(Bandbreite, Beschaeftigte), names_to = "Variable", values_to = "Wer
t")

# c)
ggplot( rohdaten_cleaned %>% pivot_longer(cols = c(Bandbreite, Beschaeftigte), names_to = "Va
riable", values_to = "Wert"), aes(x = Variable, y = Wert)) +
  geom_boxplot(fill = c("orange", "blue")) +
  labs(title = "Abb. 5: Boxplots der bereinigten Daten", x = "Variable", y = "Prozent") +
  scale_x_discrete(labels = c("Bandbreite mindestens 50 Mbit/s", "Beschäftigte Tertiärer  Sek
tor"))
```

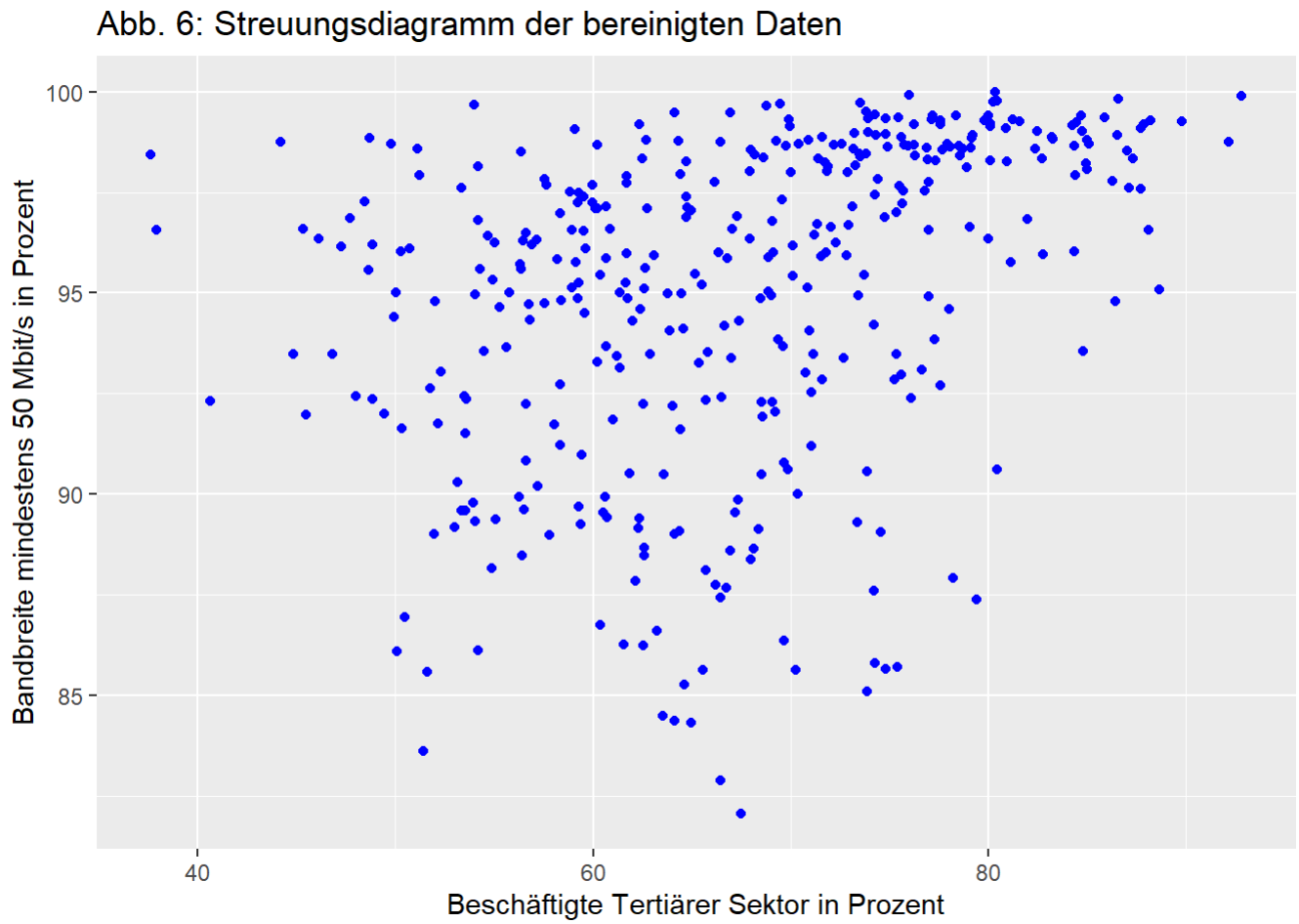


```
# d)
saveRDS(rohdaten_cleaned, file = "909003_clean.rds")
```

5. Aufgabenteil

- **a) Puntkwolke Plotten**

```
# a)
ggplot(rohdaten_cleaned, aes(x = Beschaeftigte, y = Bandbreite)) +
  geom_point(color = "blue") +
  labs(title = "Abb. 6: Streuungsdiagramm der bereinigten Daten",
       x = "Beschäftigte Tertiärer Sektor in Prozent",
       y = "Bandbreite mindestens 50 Mbit/s in Prozent")
```



Einordnung der Variablen

> Ich möchte wissen, ob es in Deutschland einen Trend gibt, dass Kreise mit einem hohen Anteil an Beschäftigten im tertiären Sektor auch einen tendenziell hohen Ausbau der Bandbreite bedeutet.
> Demnach ist die Bandbreite meine abhängige Variable .

Interpretation der Punktwolke

Abb. 5

- **Anordnung der Punkte**:
- Die Punkte zeigen, dass es eine leichte Tendenz gibt, dass Kreise mit einem hohen Anteil an Beschäftigten im tertiären Sektor auch einen tendenziell hohen Ausbau der Bandbreite haben.
- **Cluster**:
- Gerade im hohen Prozentbereich bei dem Breitbandausbau gibt es eine hohe Dichte an Punkten. Das bedeutet, dass es viele Kreise gibt, die einen hohen Anteil an Beschäftigten im tertiären Sektor über überdurchschnittliche Bandbreite verfügen.
  Dies trifft aber nicht auf die unteren Prozente zu, gerade hier sind die Punkte sehr verstreut und es gibt keine klare aussage.
- **Fazit**
- Es gibt **keine** klare lineare Beziehung zwischen den beiden Variablen jedoch gibt es einen Aufwärtstrend, welcher sich an die 100 % Breitbandausbau annähert.

6. Aufgabenteil

- **a) Korrelation plotten**
  Korrelationskoeffizient zwischen Bandbreite und Beschäftigten

```
ggplot(rohdaten_cleaned, aes(x = Beschaeftigte, y = Bandbreite)) +
  geom_point(size = 0.2, alpha = 0.8, na.rm = TRUE) +
  geom_smooth(method = "lm", se = FALSE, color = "green") +
  labs(
    title = "Abb. 7: Regressionsgerade der bereinigten Daten",
    x = "Beschäftigte Tertiärer Sektor in Prozent",
    y = "Bandbreite mindestens 50 Mbit/s in Prozent"
  ) +
  theme_minimal()
```

Abb. 7: Regressionsgerade der bereinigten Daten

Interpretation der Korrelation

`Abb. 6`

> Das Modell kann die Daten nicht erklären. Eine Homoskedastizität ist nicht gegeben, da die Residuen nicht gleichmäßig verteilt sind und einen hohen und ungleichmäßigen Streuungsbereich aufweisen. Somit ist eine Regression nicht sinnvoll, da die Daten nicht linear sind und die Residuen nicht gleichmäßig verteilt sind.

7. Aufgabenteil

- **a) Korrelation berechnen**
  Lineares Modell zwischen Bandbreite und Beschäftigten erstellen
- **b) R^2, Residuale Standardabweichung und P-Werte der Koeffizienten abspeichern und printen**

```r
# a)
model <- lm(Bandbreite ~ Beschaeftigte, data = rohdaten_cleaned)

# b)
summary_model <- summary(model)

r_squared <- summary_model$r.squared
adjusted_r_squared <- summary_model$adj.r.squared

rse <- summary_model$sigma

p_values <- summary_model$coefficients[, 4]

# Print the results
# Ausgabe der Ergebnisse
cat("R-Quadrat: ", r_squared, "\n")
```

```
## R-Quadrat:  0.1184023
```

```r
cat("Residuale Standardabweichung: ", rse, "\n")
```

```
## Residuale Standardabweichung:  3.809914
```

```r
cat("P-Werte der Koeffizienten:",p_values,"\n")
```

```
## P-Werte der Koeffizienten: 1.522802e-218 5.299159e-12
```

R^2: Gibt die generelle Modellgüte an.
Ein Wert von ~0,12 deutet auf ein schwach erklärendes Modell hin. Das bedeutet:
- Es gibt nur eine sehr schwache Beziehung zwischen den erklärenden Variablen und der Zielvariable.
- Das Modell ist möglicherweise nicht ausreichend, um zuverlässige Vorhersagen zu treffen.
- Die Beziehung zwischen den Variablen nicht linear sein.

Standardabweichung der Residuen: Zeigt, wie genau das Modell die Daten beschreibt eine Abweichung von 3.8%

P-Werte: Bewerten die Signifikanz der einzelnen Modellparameter
Diese Werte sind extrem klein (weit unter dem üblichen Signifikanzniveau von 0,05). Das bedeutet:
-Die unabhängigen Variablen haben einen signifikanten Einfluss auf die abhängige Variable.
-Es ist sehr unwahrscheinlich, dass diese Ergebnisse zufällig zustande gekommen sin

9. Aufgabenteil

- **a) Shapefile wird in R-Variable abgespeichert**
- **b) Beschriftung der Bundesländer als Dictonary anlegen**

```
# a)
shapefile <- sf::st_read("shp/VZ250_GEM.shp")
```

```
## Reading layer `VZ250_GEM' from data source
##   `E:\srv\repos\spatial_correlation\shp\VZ250_GEM.shp' using driver `ESRI Shapefile'
## Simple feature collection with 11126 features and 28 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 280371.1 ymin: 5235856 xmax: 921292.4 ymax: 6101487
## Projected CRS: ETRS89 / UTM zone 32N
```

```
# b)
beschriftung_bundeslaender <- c(
  "01" = "Schleswig-Holstein",
  "02" = "Hamburg",
  "03" = "Niedersachsen",
  "04" = "Bremen",
  "05" = "Nordrhein-Westfalen",
  "06" = "Hessen",
  "07" = "Rheinland-Pfalz",
  "08" = "Baden-Wuettemberg",
  "09" = "Bayern",
  "10" = "Saarland",
  "11" = "Berlin",
  "12" = "Brandenburg",
  "13" = "Mecklenburg-Vorpommern",
  "14" = "Sachsen",
  "15" = "Sachsen-Anhalt",
  "16" = "Thüringen"
)
```

- **c) R-Variable nach Gemeinden Gruppiert abspeichern**
- **d) Shapefile nach Bundesländern Gruppieren nach ARS**
- **e) Daten verknüpfen (Raumbezug herstellen)**
- **f) Identifizieren der fehlenden Kreise um auf der Karte die Ausreißer zu benennen**

```
# c)
kreise <- shapefile %>%
  dplyr::group_by(ARS_K, GEN_K) %>%
  dplyr::summarize(Gemeinden = n())

# d)
bundeslaender <- shapefile %>%
  dplyr::mutate(Bundesland = recode(substr(ARS_K, 1, 2), !!!beschriftung_bundeslaender)) %>%
  dplyr::group_by(Bundesland) %>%
  dplyr::summarize(Gemeinden = n())

# e)
data_cleaned_geom_kreise <- merge(x = kreise, y = rohdaten_cleaned, by.x = "ARS_K", by.y = "I
D")

# f)
fehlende_kreise <- setdiff(kreise$ARS_K, data_cleaned_geom_kreise$ARS_K)
fehlende_kreise_df <- kreise[kreise$ARS_K %in% fehlende_kreise, ]
```

10. Aufgabenteil

- **a) Differenz zwischen Bandbreite und Beschäftigten berechnen**
- **b) Karte der Differenz plotten**

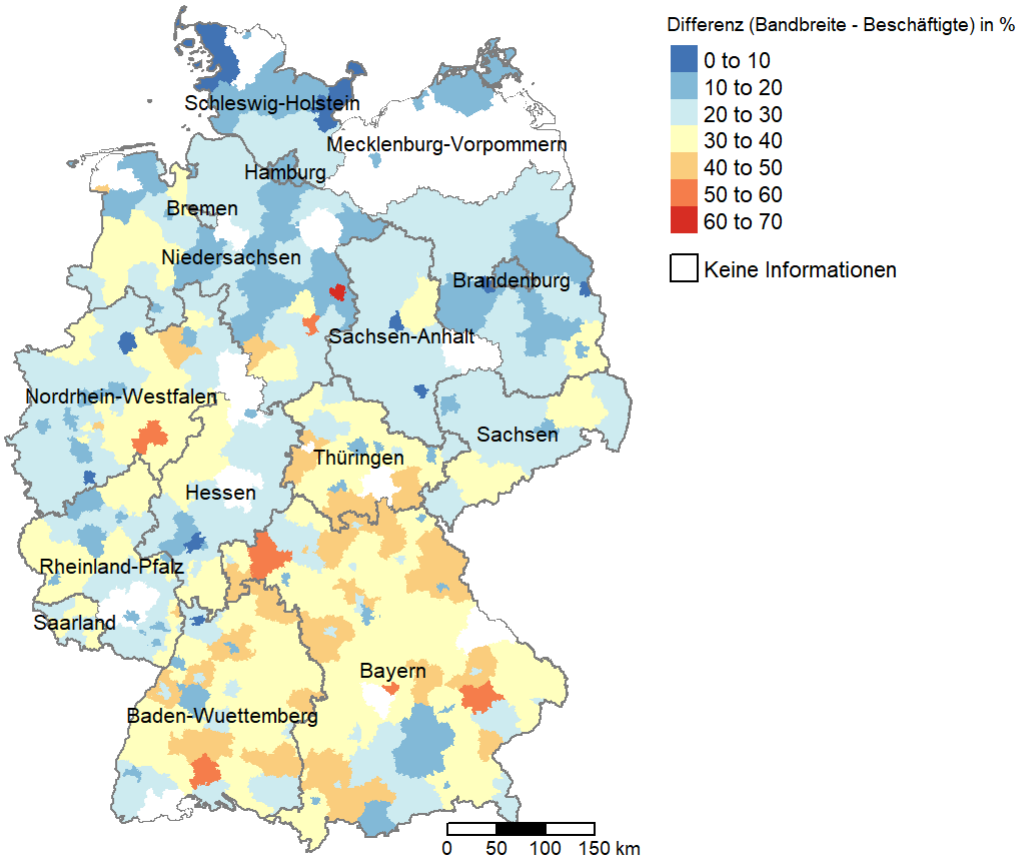
```
# a)
data_cleaned_geom_kreise_01 <- data_cleaned_geom_kreise %>%
  mutate(Differenz = Bandbreite - Beschaeftigte)

# b)
map <- tm_shape(data_cleaned_geom_kreise_01) +
  tm_fill(
    "Differenz",
    title = "Differenz (Bandbreite - Beschäftigte) in %",
    palette = "-RdYlBu"
  ) +
  tm_shape(bundeslaender) +
  tm_borders(col = 'grey50') +
  tm_scale_bar(
    width = 0.18,
    position = c("RIGHT", "BOTTOM"),
    text.size = 0.6,
    color.dark = "black",
    color.light = "white",
    just = c("RIGHT", "BOTTOM"),
    lwd = 1
  ) +
  tm_layout(
    frame = FALSE,
    main.title = "Abb. 7: Gegenüberstellung Bandbreite - Beschäftigungsanteil Tertiärer Sekto
r",
    main.title.position = "center",
    main.title.size = 0.9,
    title.position = c("center", "top"),
    legend.title.size = 0.8,
    legend.outside = TRUE,
    title.snap.to.legend = FALSE
  ) +
  tm_shape(fehlende_kreise_df) +
  tm_fill(col = "white") +
  tm_add_legend(type = "fill", labels = "Keine Informationen", col = "white") +
  tm_shape(sf::st_centroid(bundeslaender)) +
  tm_text(
    "Bundesland",
    remove.overlap = TRUE,
    size = 0.65,
    col = "black",
    fontface = "bold",
    shadow = FALSE
  )
```

```
## Warning: st_centroid assumes attributes are constant over geometries
```

```
tmap_mode("plot")
invisible(print(map))
```

Abb. 7: Gegenüberstellung Bandbreite - Beschäftigungsanteil Tertiärer Sektor



Interpretation der Karte

Abb. 8

- Wir können sehen das es keinen Kreis gibt, wo es Prozentual mehr beschäftigte im tertiären Sektor arbeiten als die Bandbreite über 50 mb/s gewährleistet sind.
- Nach Süden hin nimmt die Differenz zu, dort gibt es mehr gutes Internet, auch wenn es weniger Beschäftigte im tertiären Sektor gibt.
- Im Ost-West Vergleich gibt es keine klare Tendenz

11. Aufgabenteil

- **a) Lineare Regression durchführen**
- **b) Klassen nach Sturges definieren**
- **c) Klassenbeschriftungen definieren**
- **d) Karte der Residuen plotten**

```r
# a)
model <- lm(Beschaeftigte ~ Bandbreite, data = data_cleaned_geom_kreise)
data_cleaned_geom_kreise$residuals <- as.numeric(residuals(model))


# b)
residuals_class <- classInt::classIntervals(
  data_cleaned_geom_kreise$residuals,
  n = nclass.Sturges(data_cleaned_geom_kreise$residuals),
  style = 'pretty'
)

# c)
breaks <- residuals_class$brks
labels <- c(
  paste0("<", breaks[2]),
  paste0(breaks[-c(1, length(breaks))], " to <", breaks[-c(1, 2)]),
  paste0(">", breaks[length(breaks) - 1])
)

# d)
res_map <- tm_shape(data_cleaned_geom_kreise) +
  tm_fill(
    "residuals",
    breaks = residuals_class$brks,
    title = "Residuen in %",
    labels = labels,
    midpoint = NA,
    palette = "-RdYlBu"
  ) +
  tm_shape(bundeslaender) +
  tm_borders(col = 'grey50') +
  tm_scale_bar(
    width = 0.18,
    position = c("RIGHT", "BOTTOM"),
    text.size = 0.6,
    color.dark = "black",
    color.light = "white",
    just = c("RIGHT", "BOTTOM"),
    lwd = 1
  ) +
  tm_layout(
    frame = FALSE,
    main.title = "Abb. 9: Abweichungen vom Erwartungswert",
    main.title.position = "center",
    main.title.size = 0.9,
    title.position = c("center", "top"),
    legend.title.size = 0.8,
    legend.outside = TRUE,
    title.snap.to.legend = FALSE
  ) +
  tm_shape(fehlende_kreise_df) +
  tm_fill(col = "white") +
  tm_add_legend(type = "fill", labels = "Keine Informationen", col = "white") +
  tm_shape(sf::st_centroid(bundeslaender)) +
```

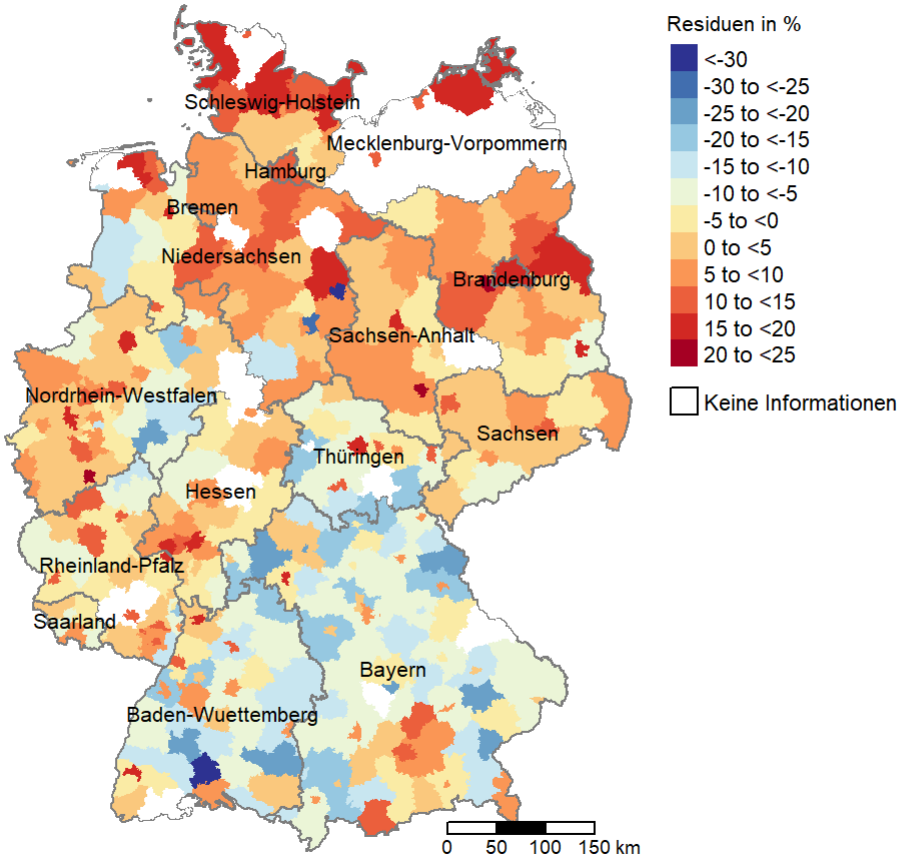
```
  tm_text(
    "Bundesland",
    remove.overlap = TRUE,
    size = 0.65,
    col = "black",
    fontface = "bold",
    shadow = FALSE
  )
```

```
## Warning: st_centroid assumes attributes are constant over geometries
```

```
tmap_mode("plot")
invisible(print(res_map))
```

```
## Warning: number of legend labels should be 12
```

Abb. 9: Abweichungen vom Erwartungswert



Interpretation der Karte

Abb. 9

- Im Norden und Nord-Osten sehen wir viele Ausreißer. Das deutet auf eine hohe Beschäftigungsquote im Tertiären Setor hin, die aber nicht vom Internetausbau getragen wird.
- Im Süden und Westen gibt es weniger Ausreißer, wenn sie vorkommen sind diese aber Positiv. Es wird dort allgemein mehr auf den Breitbandausbau wert gelegt, somit befinden sich die Residuen nah an dem Cluster oberhalb der 90% des Breitbandausbaus.
- Allerdings lässt sich der einfluss des Breitbandausbaus auf die Beschäftigungsquote mit der Regression auch in diesen Regionen nicht bestätigen, dar auch in Gebieten mit sehr ausgehprägen primären Sektor über hohen Bandbreitenanteil verfügt.

Ähnlichkeiten der beiden Karten

Abb. 8, Abb. 9

- Extremwerte zeichnen sich räumlich ähnlich ab
- Nord/Süd gefälle
- Sonst keinen weiteren zusammenhang zwischen den Karten zu erkennen.

Fazit

Abb. 8, Abb. 9, Abb. 4

In Mecklenburg-Vorpommern waren die Abweichungen so stark, dass sie als Ausreißer aus den Daten entfernt werden mussten. Dies könnte darauf zurückzuführen sein, dass hier dicht besiedelte Gebiete mit ländlichen Regionen verglichen werden, wo die Herausforderung, flächendeckend Internet bereitzustellen, deutlich größer ist. Die infrastrukturellen Unterschiede zwischen urbanen und ruralen Gebieten spielen hier eine entscheidende Rolle.
Die Ausreißer stammen vor allem aus ländlich geprägten Regionen mit einem hohen Anteil an Beschäftigten im tertiären Sektor. Um das Modell robuster gegenüber solchen Extremwerten zu gestalten, wurden die Daten entsprechend bereinigt.
Leider ließ sich kein linearer Zusammenhang zwischen den beiden Variablen feststellen, was die Interpretation der Daten erschwert. Dennoch war es möglich, aussagekräftige Schlussfolgerungen zu ziehen und zu verdeutlichen, warum in diesem Fall kein linearer Zusammenhang zu erwarten ist.
Dies unterstreicht die Komplexität der Thematik und die Notwendigkeit, weitere Faktoren wie infrastrukturelle Gegebenheiten, politische Maßnahmen und sozioökonomische Bedingungen in zukünftigen Analysen zu berücksichtigen.