

Case Study für Deutschland

Alexander Rieber - 30 April 2020

Zweiter Teil der Case Study

Im zweiten Teil der Case Study werden Sie die eingelesenen und aufgearbeiteten Daten aus Teil 1 deskriptiv untersuchen. Hierbei erhalten Sie einen Eindruck von den Daten und können mögliche Zusammenhänge entdecken, indem Sie unterschiedliche Informationen visualisieren und auch in Tabellenform auswerten. Ziele des zweiten Teils der Case Study:

- Daten visualisieren und Zusammenhänge grafisch veranschaulichen
- Deskriptive Analysen mittels Korrelationstabellen und deskriptiven Tabellen anfertigen
- Das Verständnis wie Sie ihre Informationen zu bestimmten Fragestellungen möglichst effektiv aufbereiten
- Interaktive Grafiken erstellen

Sie erhalten durch deskriptive Analysen einen sehr guten Eindruck von den regionalen Unterschieden innerhalb Deutschlands. Das begleitende 3. RTutor Problem Set gibt Ihnen einen sehr guten Eindruck davon, wie die Unterschiede zwischen den einzelnen Ländern auf europäischer Ebene aussehen.

Daten und Pakete laden

Nachdem wir uns im ersten Teil der Case Study alle Daten aus verschiedenen Datenquellen zusammengetragen und in R eingelesen haben, wollen wir in diesem zweiten Teil die darin enthaltenen Informationen analysieren, insbesondere visualisieren.

Hierzu laden wir uns die aus Teil 1 erstellten Datensätze:

```
library(tidyverse)
library(skimr)
library(sf)
library(viridis)
library(plotly)

# Daten einlesen
einkommen <- readRDS("data/einkommen.rds")
bundesland <- readRDS("data/bundesland.rds")
landkreise <- readRDS("data/landkreise.rds")
bip_zeitreihe <- readRDS("data/bip_zeitreihe.rds")
gemeinden <- readRDS("data/gemeinden.rds")
gesamtdaten <- readRDS("data/gesamtdaten.rds")
schulden_bereinigt <- readRDS("data/schulden_bereinigt.rds")
```

Deskriptive Analysen

Summary Statistics

Wir wollen uns zu Beginn unserer Analysen einen Überblick über die Daten verschaffen. D.h. wie viele Landkreise haben wir in den Daten? Wie ist die Verteilung der Schulden, Arbeitslosigkeit und des BIP?

Hierfür wollen wir uns im ersten Schritt die Arbeitslosenquote berechnen. Die Schulden pro Kopf und das BIP pro Kopf hatten wir bereits in dem ersten Teil der Case-Study berechnet. Die Arbeitslosenquote wollen wir als $Arbeitslosenquote = Erwerbslose / (Erwerbstätige + Erwerbslose)$ berechnen. Bei der Berechnung der Arbeitslosenquote beziehen wir also das komplette Potential an erwerbsfähigen Personen ein.

In den nächsten Abschnitten wollen wir uns die Parameter für die einzelnen Variablen dann genauer anschauen.

```
# Zuerst wollen wir uns noch die Arbeitslosenquote pro Landkreis berechnen
gesamtdaten <- gesamtdaten %>%
  mutate(alo_quote = (total_alo / (erw+total_alo))*100)
```

Anzahl an Beobachtungen

Wir wollen zuerst einen Blick auf die Anzahl an Erwerbstägigen und Einwohnern in Deutschland werfen. Hier haben wir 41 Mio. Erwerbstägige und 76,5 Mio. Einwohner in Deutschland. Dies sollte stimmen, da wir Hamburg (1.8 Mio.), Berlin (3.75 Mio.) und Bremen (0.7 Mio.), sowie Bremerhaven (0.1 Mio.) nicht in unserem Datensatz haben.

```
# Wie viele Erwerbstägige und Einwohner (ohne Berlin, Hamburg, Bremen und Bremerhaven) hat Deutschland?
gesamtdaten %>%
  summarise(total_erw = sum(erw, na.rm=TRUE), total_einwohner = sum(Einwohner, na.rm=TRUE))

## # A tibble: 1 x 2
##   total_erw total_einwohner
##       <dbl>        <dbl>
## 1     41068448      76573483
```

Nun wollen wir uns die Variablen im Datensatz genauer anschauen:

```
# Anschließend wollen wir eine Summary Statistic für alle Variablen ausgeben lassen
# Entfernen der Histogramme, damit alles auch schön in PDF gedruckt werden kann
skim_without_charts(select(gesamtdaten, alo_quote, Schulden_pro_kopf_lk, bip_pro_kopf, landkreis_name))
```

Table 1: Data summary

Name	select(...)
Number of rows	401
Number of columns	4
Column type frequency:	
factor	1
numeric	3
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
landkreis_name	0	1	FALSE	379	Ans: 2, Asc: 2, Aug: 2, Bam: 2

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
alo_quote	2	1.00	5.33	2.36	1.66	3.38	4.99	6.88

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
Schulden_pro_kopf_lk	4	0.99	2742.91	2147.50	264.28	1295.03	2080.59	3447.76
bip_pro_kopf	2	1.00	37086.95	16127.06	16398.46	27850.57	33105.40	40458.36

Wir haben 401 individuelle Beobachtungen in unserem Datensatz. Hierbei handelt es sich um alle Landkreise und kreisfreien Städte in Deutschland. Stimmen diese Angaben? Für einen kurzen Konsistenzcheck wollen wir uns Wikipedia bedienen.

In Deutschland gibt es 294 Landkreise. Die Anzahl der Landkreise pro Bundesland finden wir [hier](#). Weiterhin gibt es in Deutschland 107 kreisfreie Städte, die genaue Auflistung finden wir [hier](#). D.h. unsere 401 Landkreise und kreisfreien Städte sollten stimmen.

Jedoch gibt es nur 379 unterschiedliche Landkreis Namen in unserem Datensatz mit 401 unterschiedlichen Beobachtungen (Regionalschlüsseln). Dies kommt daher, dass z.B. die Stadt München eine Beobachtung ist und der Landkreis München eine weitere Beobachtung mit anderem Regionalschlüssel. D.h. der "landkreis_name" ist der gleiche, jedoch ist der Regionalschlüssel ein anderer.

Für die Schulden und die Einwohnerzahlen fehlen uns leider Daten für *vier* Landkreise, für das BIP fehlen uns Daten für *zwei* Landkreise:

```
gesamtdaten %>%
  filter(is.na(Einwohner)) %>%
  select(landkreis_name)

## # A tibble: 4 x 1
##   landkreis_name
##   <fct>
## 1 Hamburg
## 2 Bremen
## 3 Bremerhaven
## 4 Berlin
```

Leider haben wir hier in den Originaldaten keine Informationen zu Schulden und BIP für diese Städte, daher können wir sie nicht mit in unsere Analysen einbeziehen.

Beschreibung der Tabelle

Arbeitslosenquote

Im Durchschnitt liegt die Arbeitslosenquote bei 5,33 Prozent. Dies mag uns zuerst etwas hoch erscheinen, jedoch sollten wir bedenken, dass wir **alle** Arbeitslosen mit in unsere Analyse einbezogen haben, d.h. Bezieher von SGB II und SGB III. Ein kurzer Konsistenzcheck auf [Statista](#) zeigt uns die dort gemeldete Arbeitslosenquote von 5,7% für 2017. Unsere niedrigere Quote könnte insbesondere daran liegen, dass die Großstädte Berlin und Hamburg nicht in unserer Analyse enthalten sind. Die Standardabweichung beträgt 2,36 und zeigt damit, dass es in Deutschland deutliche regionale Unterschiede bzgl. der Arbeitslosenquote gibt. Ein Blick auf die Verteilung zeigt, dass der Landkreis mit der geringsten Arbeitslosenquote nur eine Arbeitslosenquote von 1,66% aufweist und der Landkreis mit der höchsten Arbeitslosenquote von 13,5%. Zwar sind die Werte noch ein ganzes Stück von dem Durchschnitt der Arbeitslosenquote in Spanien entfernt ([17% im Jahr 2017](#)), zeigen jedoch schon, dass es auch in Deutschland durchaus Regionen mit einer sehr hohen Arbeitslosenquote gibt.

Verschuldung pro Kopf

Bei der Verschuldung der Landkreise ergibt sich ein ähnliches Bild. Durchschnittlich beträgt die Verschuldung der Landkreise 2743€, mit einer Standardabweichung von 2148€. D.h. auch hier gibt es eine große Bandbreite bzgl. der Verschuldung einzelner Landkreise.

BIP pro Kopf

Wie schon bei der Arbeitslosenquote und der Verschuldung sehen wir auch bei dem BIP pro Kopf deutliche Unterschiede zwischen den einzelnen Landkreisen in Deutschland. Im Durchschnitt liegt das BIP pro Kopf bei 37087€, jedoch haben wir eine Standardabweichung von 16127€, was zuerst nach sehr viel aussieht. Dies könnte jedoch an einzelnen Landkreisen liegen (so hat ein Landkreis bspw. ein BIP pro Kopf von 172437€). Da der Median des BIP pro Kopf bei 33105€ liegt haben wir hier schon einen Hinweis, dass das BIP pro Kopf nicht unbedingt normalverteilt über alle Landkreise ist und es wohl einzelne Ausreißer in den Daten gibt.

Summary Statistics auf Bundeslandebene für die Arbeitslosigkeit

Nachdem wir im vorherigen Abschnitt bereits gesehen haben, dass es wohl deutliche regionale Unterschiede bei allen Variablen geben muss, wollen wir uns nun noch die Arbeitslosenquote auf Bundeslandebene abschauen:

`gesamtdaten %>%`

```
group_by( bundesland_name ) %>%
  select(alo_quote, bundesland_name) %>%
  skim_without_charts()
```

Table 4: Data summary

Name	Piped data
Number of rows	401
Number of columns	2
Column type frequency:	
numeric	1
Group variables	bundesland_name

Variable type: numeric

skim_variable	bundesland_name	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
alo_quote	Baden-Württemberg	0	1	3.31	0.64	2.08	2.90	3.33	3.57	4.99
alo_quote	Bayern	0	1	3.04	0.77	1.66	2.52	2.98	3.41	5.83
alo_quote	Berlin	1	0	NaN	NA	NA	NA	NA	NA	NA
alo_quote	Brandenburg	0	1	7.98	1.93	4.57	6.66	8.27	8.99	12.85
alo_quote	Bremen	0	1	8.61	2.02	7.18	7.90	8.61	9.33	10.04
alo_quote	Hamburg	1	0	NaN	NA	NA	NA	NA	NA	NA
alo_quote	Hessen	0	1	5.02	1.34	2.64	3.92	5.11	5.59	9.14
alo_quote	Mecklenburg-Vorpommern	0	1	8.41	1.56	6.30	7.35	7.95	9.98	10.44
alo_quote	Niedersachsen	0	1	6.14	1.76	2.47	5.10	6.03	7.19	10.73
alo_quote	Nordrhein-Westfalen	0	1	7.14	2.42	3.49	5.42	6.84	8.75	13.52
alo_quote	Rheinland-Pfalz	0	1	5.29	1.47	3.29	3.97	5.20	6.28	8.75
alo_quote	Saarland	0	1	5.88	1.75	4.29	4.69	5.34	6.55	8.90
alo_quote	Sachsen	0	1	6.59	1.05	5.45	5.92	6.25	6.62	9.18
alo_quote	Sachsen-Anhalt	0	1	9.00	1.56	7.03	7.98	8.87	9.32	12.66
alo_quote	Schleswig-Holstein	0	1	6.34	0.99	3.66	5.88	6.86	7.01	7.36
alo_quote	Thüringen	0	1	6.26	1.79	3.76	5.01	5.68	6.99	10.43

Hier sehen wir insbesondere für Bayern und Baden-Württemberg Arbeitslosenquoten unter 4% und für Sachsen-Anhalt, Bremen und Mecklenburg-Vorpommern Arbeitslosenquoten von über 8%. Es fällt weiterhin

auf, dass die ehemaligen Ostdeutschen Bundesländer alle sehr hohe Arbeitslosenquoten aufweisen. Um dem nachzugehen wollen wir uns eine Dummyvariable "ost" generieren, welche 0 für alle ehemaligen westdeutschen und 1 für alle ehemaligen ostdeutschen Bundesländer ist:

```
gesamtdaten <- gesamtdaten %>%
  mutate( ost = as.factor(ifelse(bundesland_name %in% c("Brandenburg", "Mecklenburg-Vorpommern", "Sachsen-Anhalt", "Sachsen", "Thüringen"),
                                ost_name = ifelse(ost == 1, "Ostdeutschland", "Westdeutschland")))
```

Durch diese Aufteilung treten die Unterschiede in der Arbeitslosenquote zwischen den ehemaligen ost- und westdeutschen Bundesländern besonders stark zutage. Insbesondere wenn wir uns die Quantile anschauen: Im **25% Quantil in Ostdeutschland** ist die Arbeitslosenquote bei **6,00%**, in **Westdeutschland** ist das **75% Quantil** bei einer Arbeitslosenquote von **6,17%**!

```
gesamtdaten %>%
  group_by(ost) %>%
  select(alo_quote, ost) %>%
  skim_without_charts()
```

Table 6: Data summary

Name	Piped data
Number of rows	401
Number of columns	2
Column type frequency:	
numeric	1
Group variables	
	ost

Variable type: numeric

skim_variable	ost	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
alo_quote	0	2	0.99	4.83	2.17	1.66	3.23	4.27	6.17	13.52
alo_quote	1	0	1.00	7.46	1.95	3.76	6.00	7.30	8.85	12.85

Visualisierung der Unterschiede

Tabellen sind schön um zusammenfassende Informationen kompakt zu präsentieren. Jedoch ist es oft ebenso wichtig (und manchmal für ihre Argumentation umso wichtiger) Erkenntnisse visuell zu veranschaulichen. In diesem Abschnitt wollen wir mehr über die Zusammensetzung jeder Variablen erfahren. Hierfür eignen sich Grafiken besonders gut.

Arbeitslosenquote

Die Variable, welche uns besonders interessiert ist die Arbeitslosenquote, insbesondere da ihr Cousin gemeint hat, dass Deutschland kein Problem mit der Arbeitslosigkeit hat. Wir wollen hier alle Datenpunkte zeigen, d.h. die Arbeitslosenquote eines jeden Landkreises für das Jahr 2017, getrennt nach Ost- und Westdeutschland. Weiterhin wollen wir unsere Grafik um einen Boxplot erweitern um einen Vergleich des Medians der Arbeitslosenquote in Ost- und Westdeutschland zu ermöglichen.

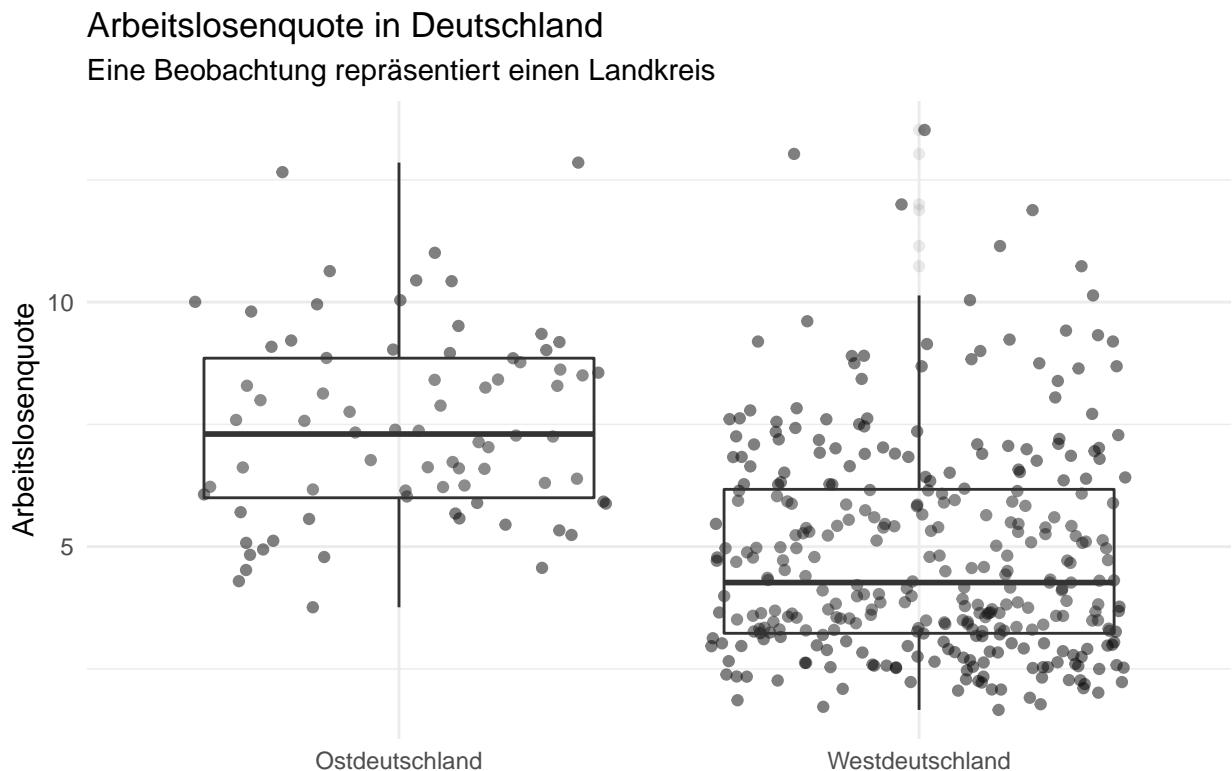
```
alo_quote_jitter <- gesamtdaten %>%
  select(alo_quote, landkreis_name, bundesland_name, ost_name) %>%
  ggplot(aes(x = ost_name, y=alo_quote)) +
```

```

geom_jitter(alpha=0.5) +
geom_boxplot(alpha = 0.1) +
theme_minimal() +
labs(title = "Arbeitslosenquote in Deutschland",
subtitle = "Eine Beobachtung repräsentiert einen Landkreis",
x = "",
y = "Arbeitslosenquote",
caption = "Quelle: Daten der Bundesagentur Agentur für Arbeit aus dem Jahr 2017")

alo_quote_jitter

```



Quelle: Daten der Bundesagentur Agentur für Arbeit aus dem Jahr 2017

Das Schaubild zeigt uns zum Einen, dass es deutlich mehr Westdeutsche, wie Ostdeutsche Landkreise gibt (nicht verwunderlich), aber auch, dass diese westdeutschen Landkreise zu einem sehr großen Teil weniger als 5% Arbeitslosigkeit aufweisen, wohingegen der größte Teil aller ostdeutschen Landkreise mehr als 5% Arbeitslosigkeit aufweist. Selbst der ostdeutsche Landkreis mit der niedrigsten Arbeitslosenquote ist nur leicht unter dem Median in Westdeutschland. Jedoch können wir konstatieren, dass ihr Cousin recht hatte mit seiner Behauptung, denn es gibt sowohl in Ost als auch in Westdeutschland nur sehr wenige Landkreise welche eine Arbeitslosenquote von mehr als 10% haben. In Spanien gibt es **fast keine Region mit einer Arbeitslosenquote unter 10%**! Nichtsdestotrotz sind auch in Deutschland regionale Unterschiede erkennbar, welche wir insbesondere im dritten Teil der Case Study zu erklären versuchen. Als mögliche Faktoren, welche die Arbeitslosenquote erklären könnten wollen wir hier das BIP pro Kopf und die Pro-Kopf-Verschuldung näher untersuchen..

Hinweis zur vorherigen Grafik: Falls wir dem Leser etwas mehr Freiheiten einräumen möchten und unsere Grafik als HTML Datei übergeben, so können wir auch eine interaktive Grafik nutzen. So kann der Leser mit unserer Ausarbeitung interagieren:

```

# Zusätzlich Info um welchen Landkreis es sich handelt
alo_quote_jitter_plotly <- alo_quote_jitter <- gesamtdaten %>%
  select(alo_quote, landkreis_name, bundesland_name, ost_name) %>%
  ggplot(aes(x = ost_name, y=alo_quote, label = landkreis_name)) +
  geom_jitter(alpha=0.5) +
  geom_boxplot(alpha = 0.1) +
  theme_minimal() +
  labs(title = "Arbeitslosenquote in Deutschland",
       subtitle = "Eine Beobachtung repräsentiert einen Landkreis",
       x = "",
       y = "Arbeitslosenquote",
       caption = "Quelle: Daten der Bundesagentur Agentur für Arbeit aus dem Jahr 2017.")

```

ggplotly(alo_quote_jitter_plotly)

Bruttoinlandsprodukt pro Kopf

Wir haben im vorherigen Abschnitt gesehen, dass es durchaus deutliche regionale Unterschiede in der Arbeitslosenquote in 2017 gibt. Doch was könnten die Treiber dafür sein?

In diesem Unterabschnitt wollen wir uns das Bruttoinlandsprodukt pro Kopf auf Landkreisebene näher anschauen. Hierbei haben wir nicht nur die Daten für 2017, sondern können uns einer längeren Zeitreihe bedienen. Somit können wir uns die Entwicklung des BIP pro Kopf in den ehemaligen ost- und westdeutschen Bundesländern näher anschauen. Diese Zeitreihen sind für uns nützliche Hinweise, denn dadurch können wir sehen:

- ob es auch regionale Unterschiede im BIP pro Kopf gibt und nicht nur in der Arbeitslosenquote
- ob die regionalen Unterschiede schon längere Zeit bestehen
- ob die regionalen Unterschiede sich vergrößern oder verkleinern

Das Bruttoinlandsprodukt stellt die wichtigste gesamtwirtschaftliche Kenngröße dar. Es gibt Aufschluss darüber, wie viele Güter und Dienstleistungen in dem jeweiligen Landkreis produziert wurden. Falls das BIP in einem Landkreis hoch ist könnte dies daran liegen, dass

- viele Personen in diesem Landkreis erwerbstätig sind,
- oder das die Erwerbstätigen in Branchen mit hoher Produktivität arbeiten.

Falls der erste Punkt zutrifft sollte ein hohes BIP pro Kopf (man beachte das hier das BIP pro Einwohner berechnet wird) auch mit einer niedrigeren Arbeitslosenquote einhergehen.

```

# Zuerst die Information zu landkreis_name und bundesland_name zum Datensatz big_long hinzumergen
namen <- gesamtdaten %>%
  select(Regionalschluessel, bundesland_name, landkreis_name, ost_name)

bip_zeitreihe <- bip_zeitreihe %>%
  filter( nchar(Regionalschluessel) == 5) %>%
  left_join(namen, by="Regionalschluessel")

options(scipen = 5)
plot_bip <- bip_zeitreihe %>%
  filter( Jahr >= 2000 ) %>%
  group_by(ost_name, Jahr) %>%
  mutate( durchschnitt = mean(bip_pro_kopf)) %>%
  ggplot() +
  geom_line(aes(x = Jahr, y = bip_pro_kopf, group = Regionalschluessel), color = "grey") +
  geom_line(aes(x = Jahr, y = durchschnitt, group = Regionalschluessel), color = "darkblue") +

```

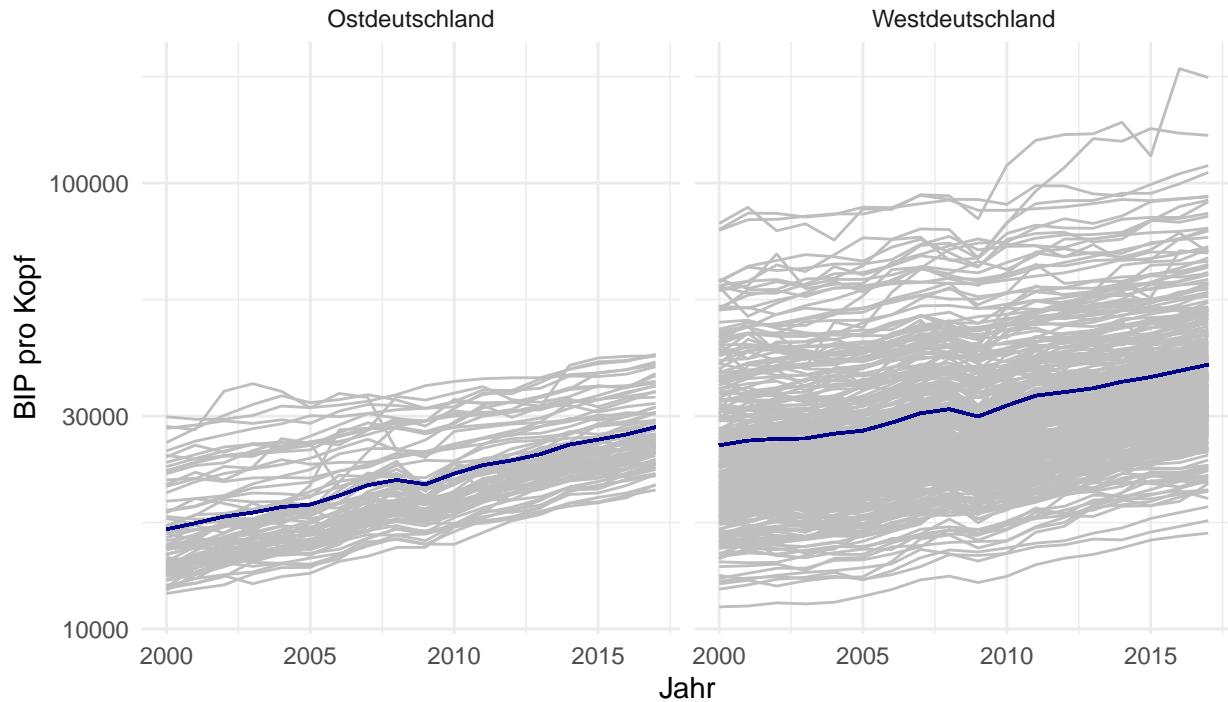
```

scale_y_continuous(trans = "log10") +
theme_minimal() +
facet_wrap(ost_name ~ .) +
theme(legend.position = "none") +
labs(title = "Ein Vergleich des BIP pro Kopf von ost- und westdeutschen Landkreisen",
subtitle = "Durchschnittswerte in Dunkelblau",
caption = "Quelle: Daten der Statistischen Ämter der Länder und des Bundes.",
x = "Jahr",
y = "BIP pro Kopf")

plot_bip

```

Ein Vergleich des BIP pro Kopf von ost– und westdeutschen Landkreisen Durchschnittswerte in Dunkelblau



Quelle: Daten der Statistischen Ämter der Länder und des Bundes.

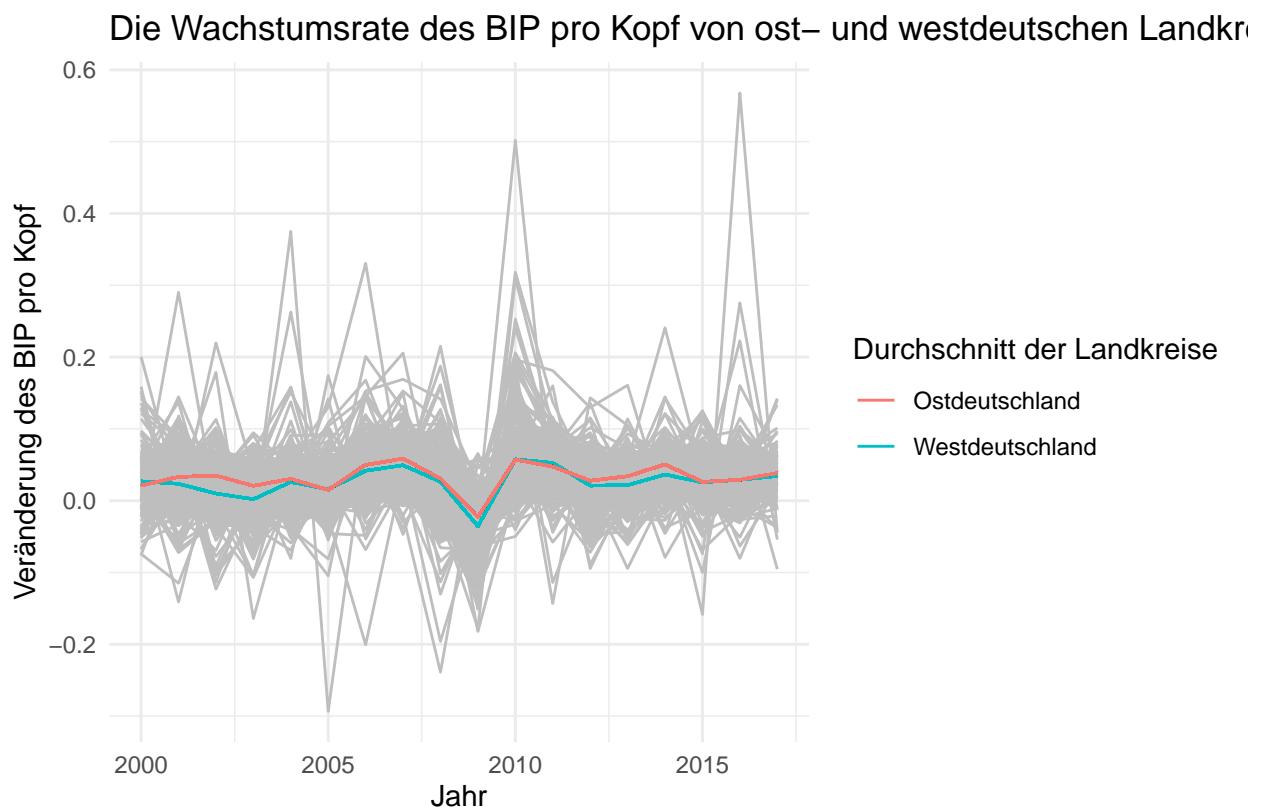
Das Niveau des BIP pro Kopf ist in den ostdeutschen Bundesländern deutlich niedriger als in den westdeutschen. Insbesondere wenn wir uns die logarithmische Skalierung der y-Achse anschauen, so können wir doch deutliche Unterschiede erkennen. Beispielsweise ist das durchschnittliche BIP pro Kopf im ehemaligen Ostdeutschland mit 28338€ in 2017 so hoch wie in Westdeutschland im Jahr 2006 (zu diesem Zeitpunkt lag das durchschnittliche BIP pro Kopf in Westdeutschland bei 29019€)! In beiden Grafiken sehen wir eine Delle im Jahr 2009 als die globale Finanzkrise zuschlug. Noch besser können wir den Effekt der Finanzkrise an der folgenden Grafik sehen, bei welcher die Veränderung des BIP pro Kopf zum Vorjahr abgetragen wird:

```

bip_wachstum <- bip_zeitreihe %>%
  group_by(Regionalschlüssel) %>%
  arrange(Regionalschlüssel, Jahr) %>%
  mutate( bip_wachstum = (bip_pro_kopf / lag(bip_pro_kopf))-1) %>%
  ungroup() %>%
  group_by(ost_name, Jahr) %>%
  mutate( durchschnitt = mean(bip_wachstum, na.rm=TRUE))

```

```
bip_wachstum %>%
  filter( Jahr >= 2000 ) %>%
  ggplot() +
  geom_line(aes(x = Jahr, y = bip_wachstum, group = Regionalschlüssel), color = "grey") +
  geom_line(aes(x = Jahr, y = durchschnitt, group = Regionalschlüssel, color = ost_name)) +
  theme_minimal() +
  labs(color = "Durchschnitt der Landkreise",
       title = "Die Wachstumsrate des BIP pro Kopf von ost- und westdeutschen Landkreisen",
       caption = "Quelle: Daten der Statistischen Ämter der Länder und des Bundes.",
       x = "Jahr",
       y = "Veränderung des BIP pro Kopf")
```



Quelle: Daten der Statistischen Ämter der Länder und des Bundes.

Insgesamt haben ost- und westdeutsche Landkreise im Durchschnitt eine sehr ähnliche Wachstumsrate des BIP pro Kopf. In den westdeutschen Landkreisen ist sie stellenweise etwas niedriger, insbesondere vor der Finanzkrise.

Durch diese zwei Schaubilder sehen wir, dass die Unterschiede zwischen den ost- und westdeutschen Landkreisen schon lange bestehen (mind. seit 2000) und das es keine große Annäherung der ostdeutschen Landkreise an das westdeutsche Niveau gibt. Die Wachstumspfade sind hierfür zu ähnlich. Wenn das BIP pro Kopf ein signifikanter Faktor für die Arbeitslosenquote in einem Landkreis ist, dann verheit dies für die ostdeutschen Landkreise, welche aktuell deutlich höhere Arbeitslosenquoten als die westdeutschen Landkreise aufweisen, nichts Gutes.

Wenn wir uns nur auf das Jahr 2017 beschränken, so sehen wir auch hier im BIP pro Kopf deutliche Unterschiede. Im folgenden Schaubild ist die Verteilung des BIP pro Kopf über alle Landkreise für Ost- und Westdeutschland im Jahr 2017 abgetragen, für alle Landkreise mit einem BIP pro Kopf von unter 100 000€. Zusätzlich sind die jeweiligen Mittelwerte des BIP für Ost- und Westdeutschland eingetragen.

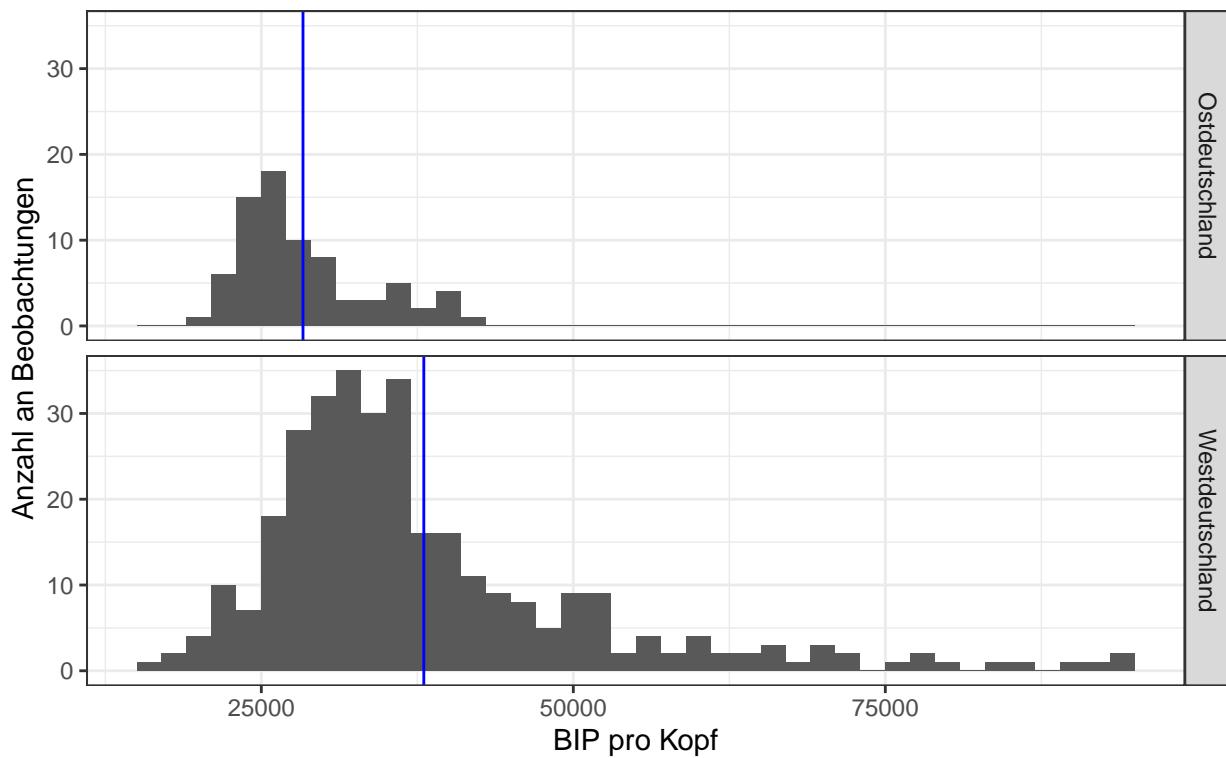
```

ggplot(gesamtdaten %>%
  filter(bip_pro_kopf<100000) %>%
  group_by(ost_name) %>%
  summarise(durchschnitt = mean(bip_pro_kopf)), aes(x = bip_pro_kopf)) +
  geom_histogram(data = filter(gesamtdaten, bip_pro_kopf<100000), binwidth = 2000) +
  facet_grid(ost_name~.) +
  geom_vline(aes(xintercept = durchschnitt), color = "blue") +
  theme_bw() +
  labs(title = "Verteilung des BIP pro Kopf für Ost- und Westdeutschland",
       subtitle = "Beobachtungen auf Landkreisebene in 2017",
       x = "BIP pro Kopf",
       y = "Anzahl an Beobachtungen")

```

Verteilung des BIP pro Kopf für Ost– und Westdeutschland

Beobachtungen auf Landkreisebene in 2017



Auch die Analyse des BIP pro Kopf im Jahr 2017 stützt unsere bisherigen Erkenntnisse. Die Ostdeutschen Landkreise haben in 2017 ein durchschnittliches BIP pro Kopf von 28338€, die westdeutschen Landkreise von 39146€.

Verschuldung der einzelnen Landkreise

Ein weiterer Faktor, welcher die Arbeitslosenquote bestimmen kann ist die Verschuldung des öffentlichen Haushalts. In strukturschwachen Landkreisen mit einer hohen Verschuldung sind tendenziell weniger Jobs vorhanden und die Gemeinden haben niedrigere Gewerbesteuereinnahmen. Für Investitionen müssen sich diese Gemeinden beim nicht-öffentlichen Sektor verschulden.

Um uns die Verschuldung des öffentlichen Haushalts besser vor Augen zu führen wollen wir diese auf einer Deutschlandkarte darstellen:

```

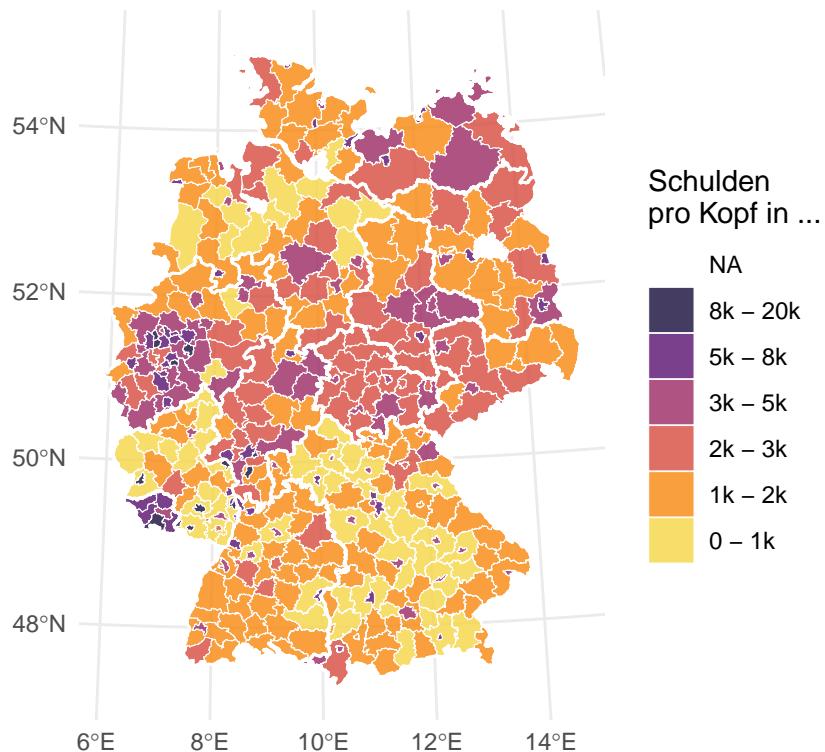
schulden_landkreise <- left_join(landkreise, gesamtdaten, by="Regionalschluessel") %>%
  mutate(schulden_pro_kopf = as.factor(case_when(
    Schulden_pro_kopf_lk <= 1000 ~ "0 - 1k",
    Schulden_pro_kopf_lk > 1000 & Schulden_pro_kopf_lk <= 2000 ~ "1k - 2k",
    Schulden_pro_kopf_lk > 2000 & Schulden_pro_kopf_lk <= 3000 ~ "2k - 3k",
    Schulden_pro_kopf_lk > 3000 & Schulden_pro_kopf_lk <= 5000 ~ "3k - 5k",
    Schulden_pro_kopf_lk > 5000 & Schulden_pro_kopf_lk <= 8000 ~ "5k - 8k",
    Schulden_pro_kopf_lk > 8000 & Schulden_pro_kopf_lk <= 20000 ~ "8k - 20k"
  )))

plot_schulden_lk <- ggplot(
  # define main data source
  data = schulden_landkreise
) +
  geom_sf(
    mapping = aes(
      fill = schulden_pro_kopf
    ),
    color = "white",
    size = 0.1
  ) +
  # use the Viridis color scale
  scale_fill_viridis_d(
    option = "inferno",
    name = "Schulden\npro Kopf in €",
    alpha = 0.8, # make fill a bit brighter
    begin = 0.1,
    end = 0.9,
    direction = -1,
    guide = guide_legend(reverse = T)) +
  # use thicker white stroke for cantonal borders
  geom_sf(
    data = bundesland,
    fill = "transparent",
    color = "white",
    size = 0.5
  ) +
  # add titles
  labs(x = NULL,
       y = NULL,
       title = "Wie verschuldet sind die deutschen Landkreise?",
       subtitle = "Öffentliche Schulden pro Kopf in 2017") +
  theme_minimal()

plot_schulden_lk

```

Wie verschuldet sind die deutschen Landkreise? Öffentliche Schulden pro Kopf in 2017



Wir sehen besonders in Bayern und Baden-Württemberg sehr viele Landkreise mit einer geringen Pro-Kopf Verschuldung bis maximal 2000€, gleiches gilt für Reinland-Pfalz und zum Großteil auch Niedersachsen. In Nordrhein-Westfalen und dem Saarland hingegen haben wir eine sehr hohe Verschuldung pro Kopf. Thüringen, Mecklenburg-Vorpommern und Hessen sind im Mittelfeld. Um jedoch beurteilen zu können, in wie fern diese Pro-Kopf Verschuldung der öffentlichen Haushalte mit der Arbeitslosigkeit in den jeweiligen Landkreisen zusammenfällt wollen wir uns auch noch eine Karte zur Arbeitslosigkeit in den jeweiligen Landkreisen anschauen:

```
schulden_landkreise <- left_join(landkreise, gesamtdaten, by="Regionalschlüssel") %>%
  mutate(alo_quote_lk = as.factor(case_when(
    alo_quote <= 2 ~ "0 - 2%",
    alo_quote > 2 & alo_quote <= 4 ~ "2 - 4%",
    alo_quote > 4 & alo_quote <= 6 ~ "4 - 6%",
    alo_quote > 6 & alo_quote <= 8 ~ "6 - 8%",
    alo_quote > 8 & alo_quote <= 10 ~ "8 - 10%",
    alo_quote > 10 & alo_quote <= 13.5 ~ "10 - 13.5%"
  )))
  
plot_alo_lk <- ggplot(
  # define main data source
  data = schulden_landkreise
) +
  geom_sf(
    mapping = aes(
      fill = fct_relevel(alo_quote_lk, "10 - 13.5%", after = 5)
    ),
    color = "white",
    size = 0.1
)
```

```

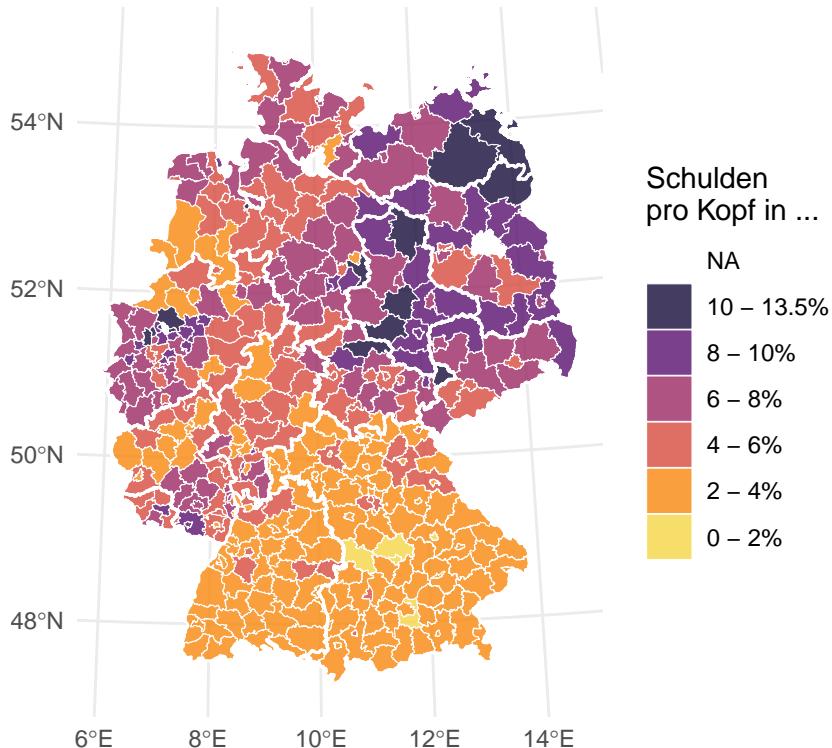
) +
# use the Viridis color scale
scale_fill_viridis_d(
  option = "inferno",
  name = "Schulden\npro Kopf in €",
  alpha = 0.8, # make fill a bit brighter
  begin = 0.1,
  end = 0.9,
  direction = -1,
  guide = guide_legend(reverse = T)) +
# use thicker white stroke for cantonal borders
geom_sf(
  data = bundesland,
  fill = "transparent",
  color = "white",
  size = 0.5
) +
# add titles
labs(x = NULL,
     y = NULL,
     title = "Arbeitslosigkeit in Deutschland",
     subtitle = "Dargestellt ist die Arbeitslosenquote für alle Landkreise in 2017") +
theme_minimal()

```

plot_alo_lk

Arbeitslosigkeit in Deutschland

Dargestellt ist die Arbeitslosenquote für alle Landkreise in 2017



In dieser Karte sehen wir einen deutlichen Unterschied für ost- und westdeutsche Landkreise. In Ostdeutschland ist die Arbeitslosigkeit deutlich höher als in Westdeutschland, d.h. das Kartenpendant zu unseren vorherigen Analysen. Jedoch sind die Landkreise in Westdeutschland, welche eine hohe Verschuldung aufweisen, wie beispielsweise im Saarland oder Nordrhein-Westfalen, auch die Landkreise mit der höchsten Arbeitslosenquote in Westdeutschland. Daher könnte die Verschuldung des öffentlichen Haushalts durchaus auch ein Faktor sein, welcher die Arbeitslosenquote mit bestimmt und wir sollten diesen Faktor in unserer Analyse nicht außen vor lassen.

Zusammenfassung und Ausblick

Nun haben wir uns einen Überblick über unsere Daten verschafft und schon einiges zu den regionalen Unterschieden in Deutschland gelernt. Sowohl die deskriptive Analyse mittels Tabellen als auch die grafische Analyse durch verschiedene Plots zeigen uns auch 28 Jahre nach dem Mauerfall (wir analysieren Daten aus 2017) einen deutlichen Unterschied zwischen dem ehemaligen Ostdeutschland und Westdeutschland. Jedoch sind die Unterschiede in den Arbeitslosenquoten nicht nur zwischen Ost- und Westdeutschland erkennbar, sondern es gibt auch einige Regionen in Westdeutschland, wie das Saarland oder Nordrhein-Westfalen, welche eine recht hohe Arbeitslosenquote aufweisen.

In dem letzten Teil der Case Study wollen wir uns die Korrelation der einzelnen Variablen genauer anschauen und mittels linearer Regression ein tiefgreifendes Verständnis von den möglichen Einflussfaktoren auf die Arbeitslosenquote erhalten.

Übungsaufgaben

Im ersten Teil der Case Study hatten Sie sich noch die durchschnittlichen Einkommen auf Landkreisebene in R eingelesen. Nun sollten Sie diese Tabelle deskriptiv analysieren:

- Erstellen Sie eine deskriptive Tabelle, welche das Einkommen für das Jahr 2017 darstellt. Wie ist hier die Verteilung der Einkommen?
 - Beschreiben Sie Mittelwert, Standardabweichung, sowie Median
- Erstellen Sie ein Liniendiagramm zu der Entwicklung des Einkommensniveaus in den einzelnen Landkreisen. Sie können sich hierbei an dem Diagramm zum BIP pro Kopf orientieren.
 - Hinweis: Mergen Sie zu dem Datensatz "Einkommen" zuerst noch die Information zu "Landkreis_name, Bundesland_name und ost_name" hinzu (siehe auch hierzu [diesen Abschnitt](#))
- Erstellen Sie eine Karte zum Einkommensniveau der einzelnen Landkreise. Sie können sich hierbei an der Karte zur Verschuldung orientieren.

Weitere Quellen:

- [Data to Viz](#)
- [A ggplot2 Tutorial for Beautiful Plotting in R](#)
- [The Evolution of a ggplot](#)
- [Switzerland's regional income \(in-\)equality](#)