SOCIOECONOMIC FACTORS &

OBESITY USA

Gliederung

Ioakeim Ioakeim

Fabian Georgi

Matthias Titze

01 Forschungsfrage

Food Atlas

02 Geoplotting & Daten-Exploration

03 Vorhersagemodell: Lineare Regression

NHANES

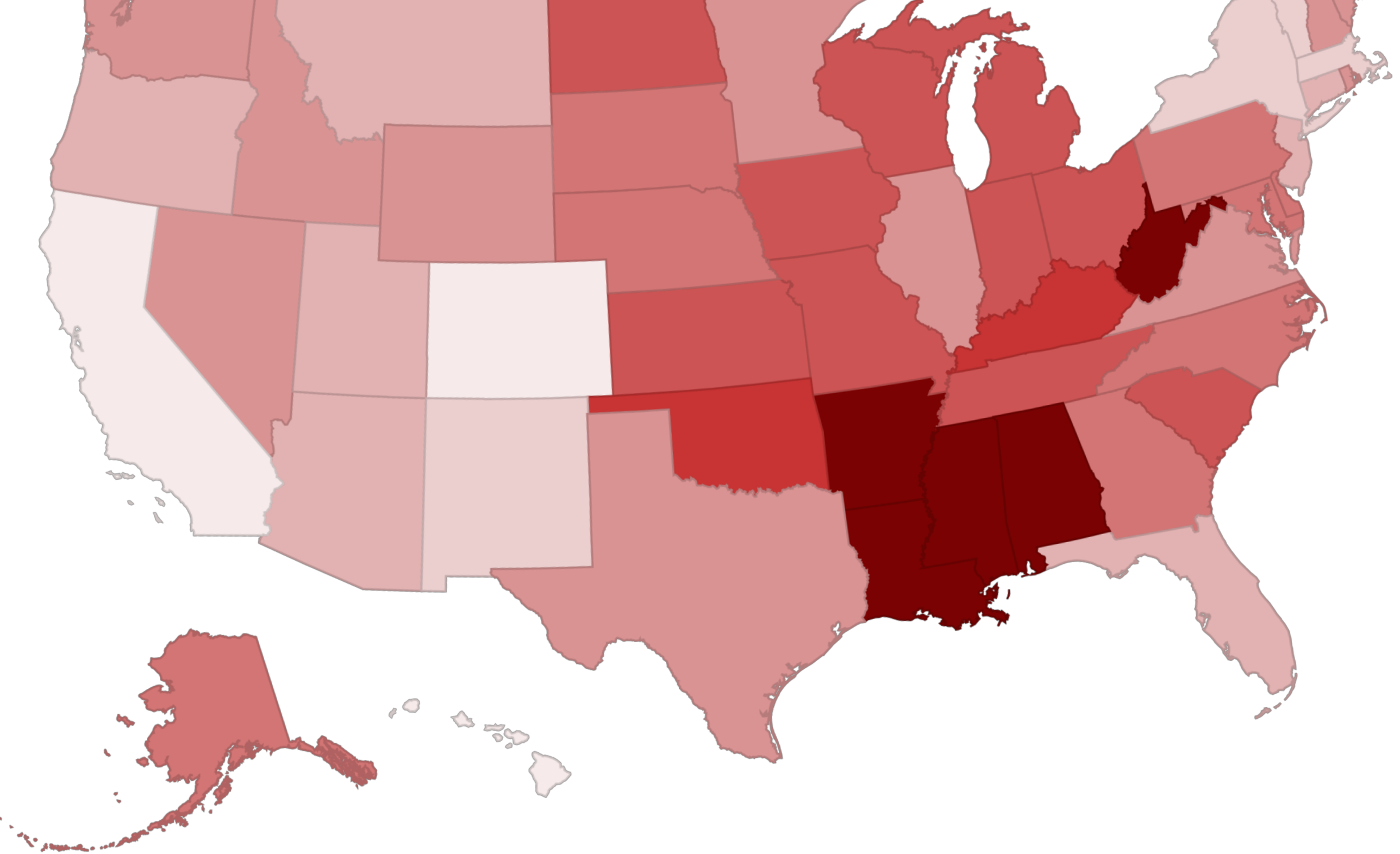
04 Daten-Exploration & Korrelationstests

05 Prognosemodell: Naive Bayes

06 Fazit & Ausblick

01 Forschungsfrage

Können sozioökonomische Faktoren genutzt werden, um Fettleibigkeit in den USA zu prognostizieren?



FOOD ATLAS

ATLAS 02 Data Exploration

02 .1 Data Source

Abb. 2.1.1 Online-Portal: USDA Food Atlas

United States Department of Agriculture
Economic Research Service

Due to a lapse in federal funding, this USDA website will not be actively updated. Once funding has been reestablished, online operations will continue.

Topics | Data Products | Publications | Newsroom | Calendar | Amber Waves Magazine | ERS Info

Home / Data Products / Food Environment Atlas / Documentation

Food Environment Atlas

- Overview
- Go to the Atlas
- About the Atlas
- Documentation
- Data Access and Documentation Downloads

Related Topics

- WIC Program
- Food Choices & Health
- Child Nutrition Programs
- News Release: 2018 Research Hightlights
- Fruit & Tree Nuts

Documentation

Definitions and Data Sources

This page provides definitions and data sources for the Food Environment Atlas indicators, grouped under the following categories:

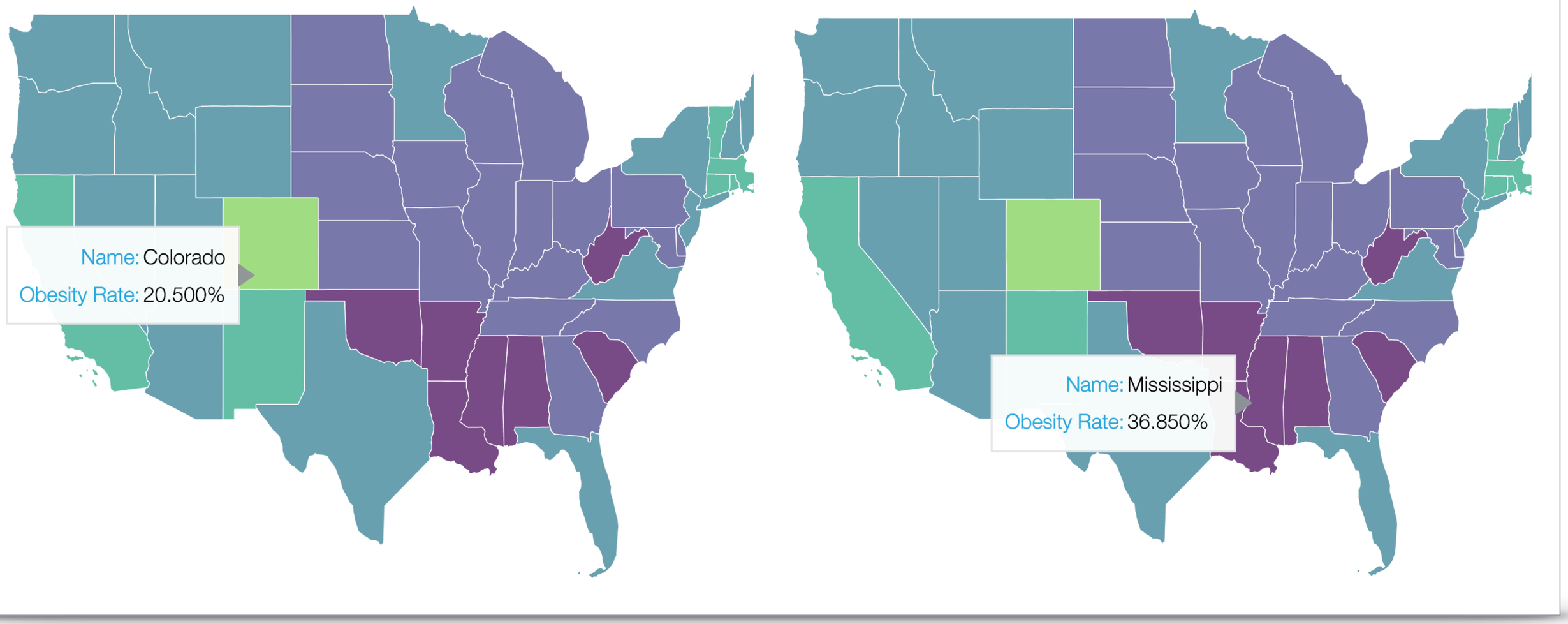
- Access and Proximity to Grocery Store
- Store Availability
- Restaurant Availability and Expenditures
- Food Assistance
- State Food Insecurity
- Food Prices and Taxes
- Local Foods
- Health and Physical Activity
- Socioeconomic Characteristics

Indicators are county-level measures unless otherwise noted as follows:

02 .2 Geoplotting

State Level - Obese Adults 2013

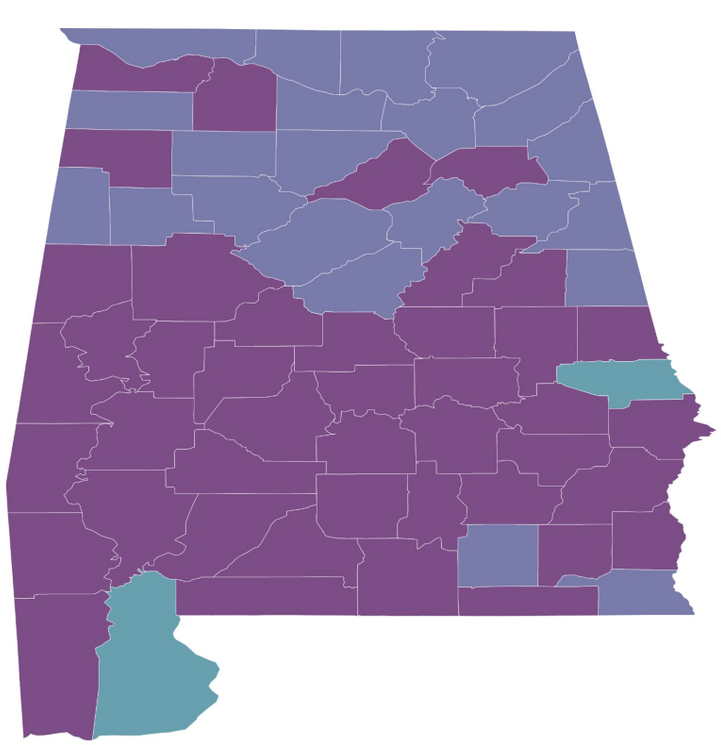
Abb. 2.2.1 Obesity Adults 2013 (US States)



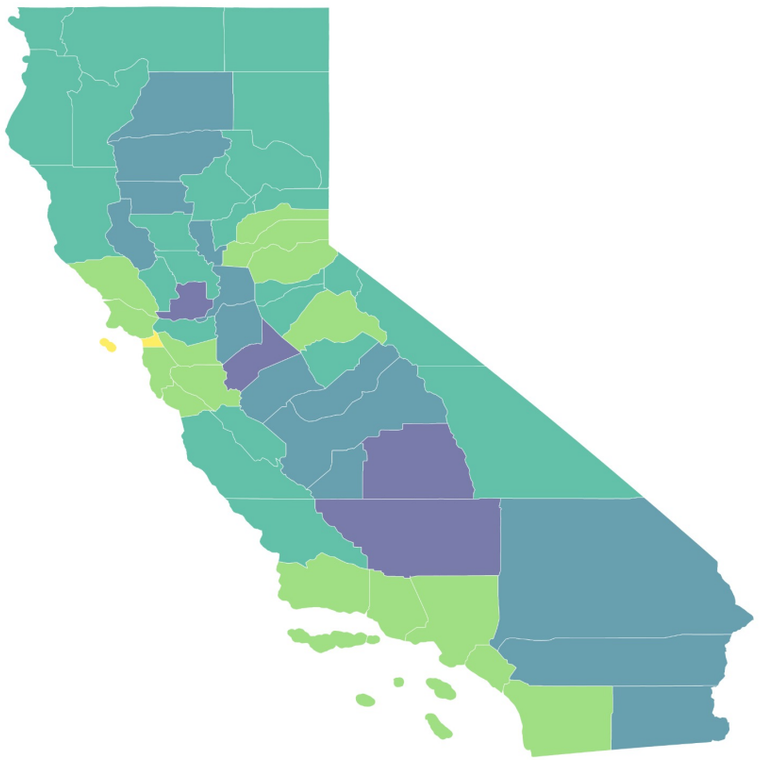
02 .2 Geoplotting

County Level - Obese Adults 2013

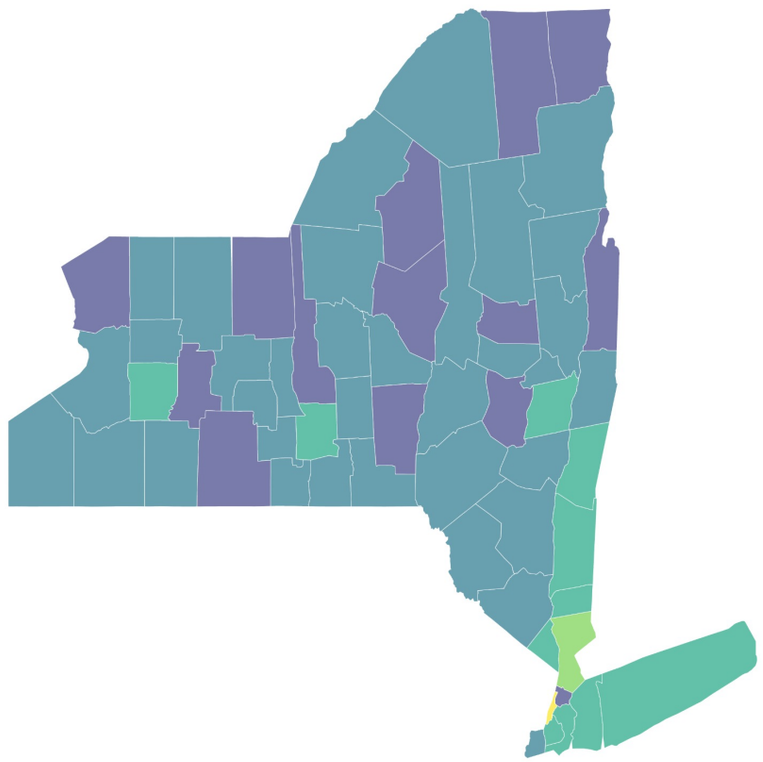
Abb. 2.2.2 Obesity Adults 2013 (US States)



Alabama



California



New York (State)

02 .2 Metropolitan Areas

Population Density vs. Obesity

```
df_non_metro_describe['PCT_obese_adults_2013'].describe()
```

```
count    1973.000000
mean       31.144450
std         4.529695
min        12.700000
25%        28.400000
50%        31.300000
75%        33.900000
max        47.600000
Name: PCT_obese_adults_2013, dtype: float64
```

```
df_metro_describe['PCT_obese_adults_2013'].describe()
```

```
count    1167.000000
mean       30.790231
std         4.495602
min        11.800000
25%        28.100000
50%        31.000000
75%        33.700000
max        46.100000
Name: PCT_obese_adults_2013, dtype: float64
```

Abb. 2.2.4 Dataframe Descriptive Values

Abb. 2.2.3 Population Density & Metro-Area vs. Obesity

02 .3 Obesity

Definitions

➤ Description: Estimate of age-adjusted percentage of persons age 20 and older who are obese, where obesity is Body Mass Index (BMI) greater than or equal to 30 kilograms per meters squared.

- ➤ BMI < 18.5 underweight
- ➤ BMI 18.5 to < 25 normal weight
- ➤ BMI 25.0 to < 30 overweight
- ➤ BMI > 30.0 obese

Obesity

- ➤ Class 1: BMI 30 to < 35,
- ➤ Class 2: BMI 35 to < 40,
- ➤ Class 3: BMI > 40.

Class 3 is considered to be “extreme” Obesity.

02 .3 Obesity

Histograms

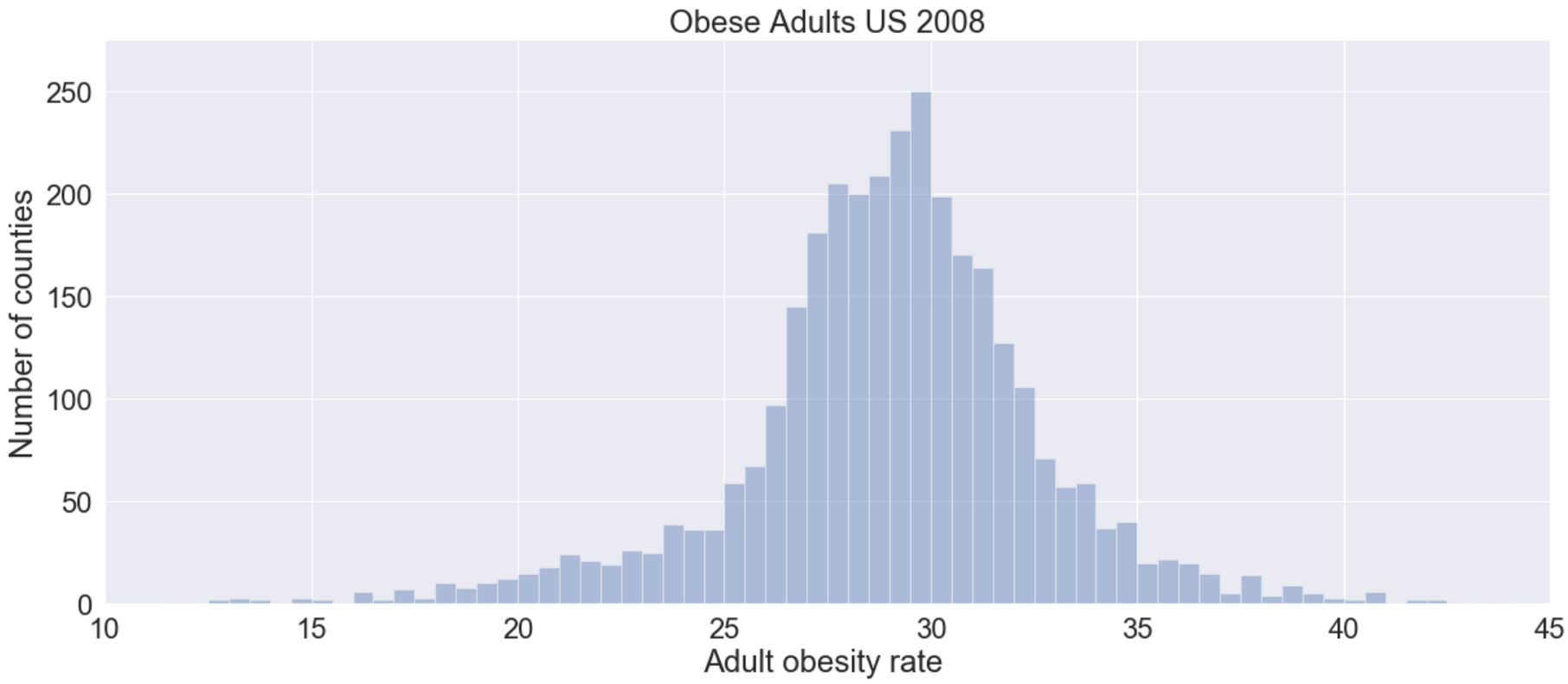


Abb. 2.3.1 Histogram: Obesity 2008

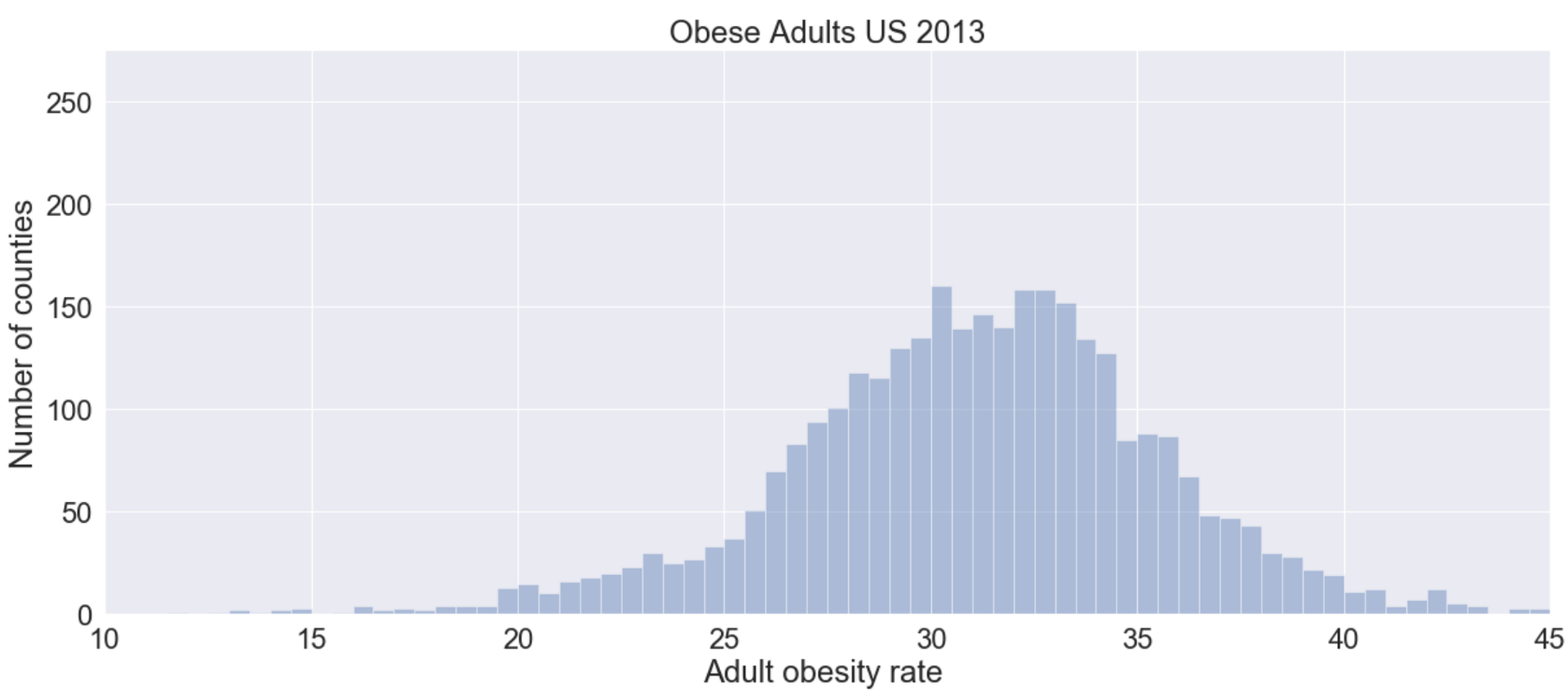


Abb. 2.3.2 Histogram: Obesity 2013

02 .3 Obesity

Histograms

	count	mean	std	min	25%	50%	75%	max
Adult obesity rate 2008	3138.0	28.931	3.711	11.70	27.20	29.10	31.00	43.7
Adult obesity rate 2013	3142.0	31.017	4.523	11.80	28.30	31.20	33.80	47.6
Average of adult obesity rate from 2008 & 2013	3137.0	29.974	3.935	12.35	27.85	30.25	32.35	45.0
Percentage change of adult obesity rate from 2008 to 2013	3137.0	2.086	2.566	-6.90	0.40	2.00	3.70	13.0

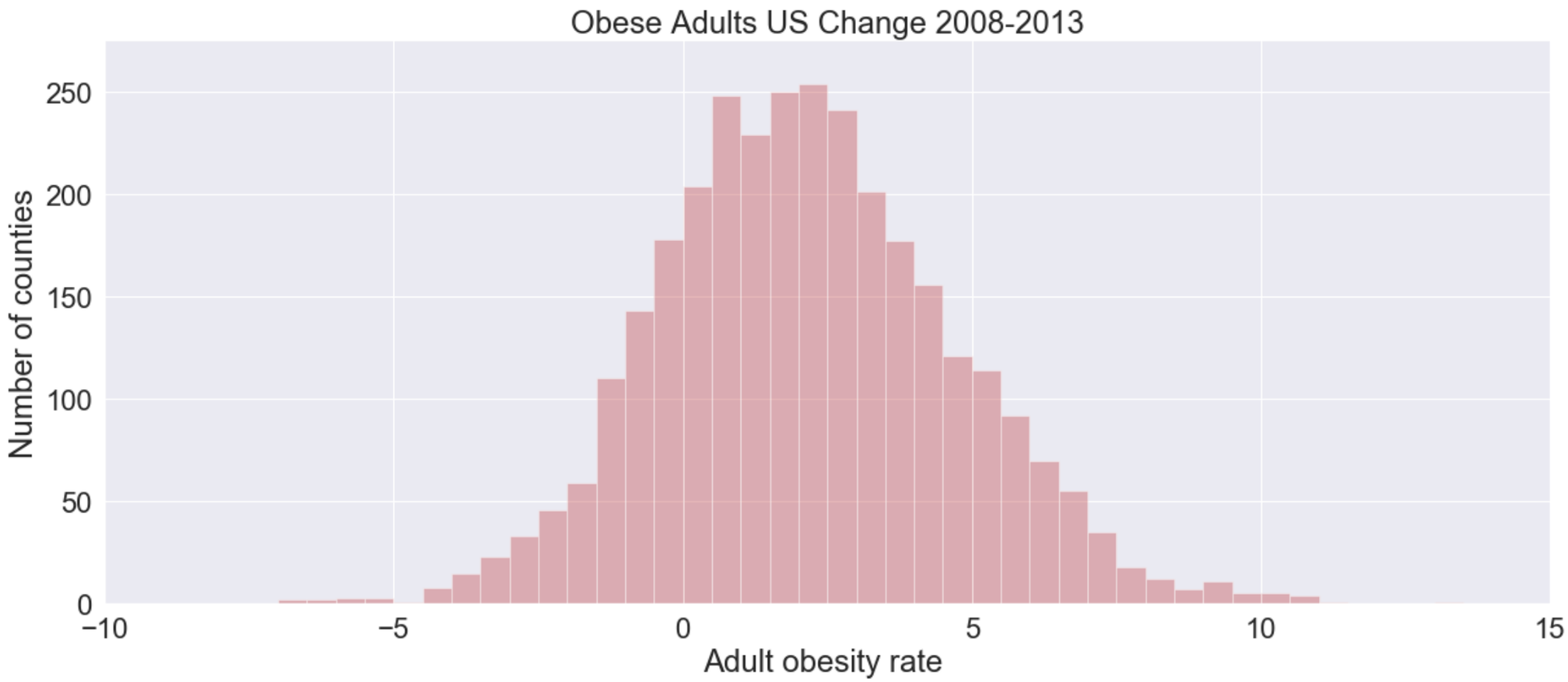


Abb. 2.3.3 Histogram: Obesity Change 2008 to 2013

02 .4 Univariant Correlations

Diabetes > Obesity

➤ **Diabetes Description:** Estimates of age-adjusted percentage of persons age 20 and older with diabetes (gestational diabetes excluded).

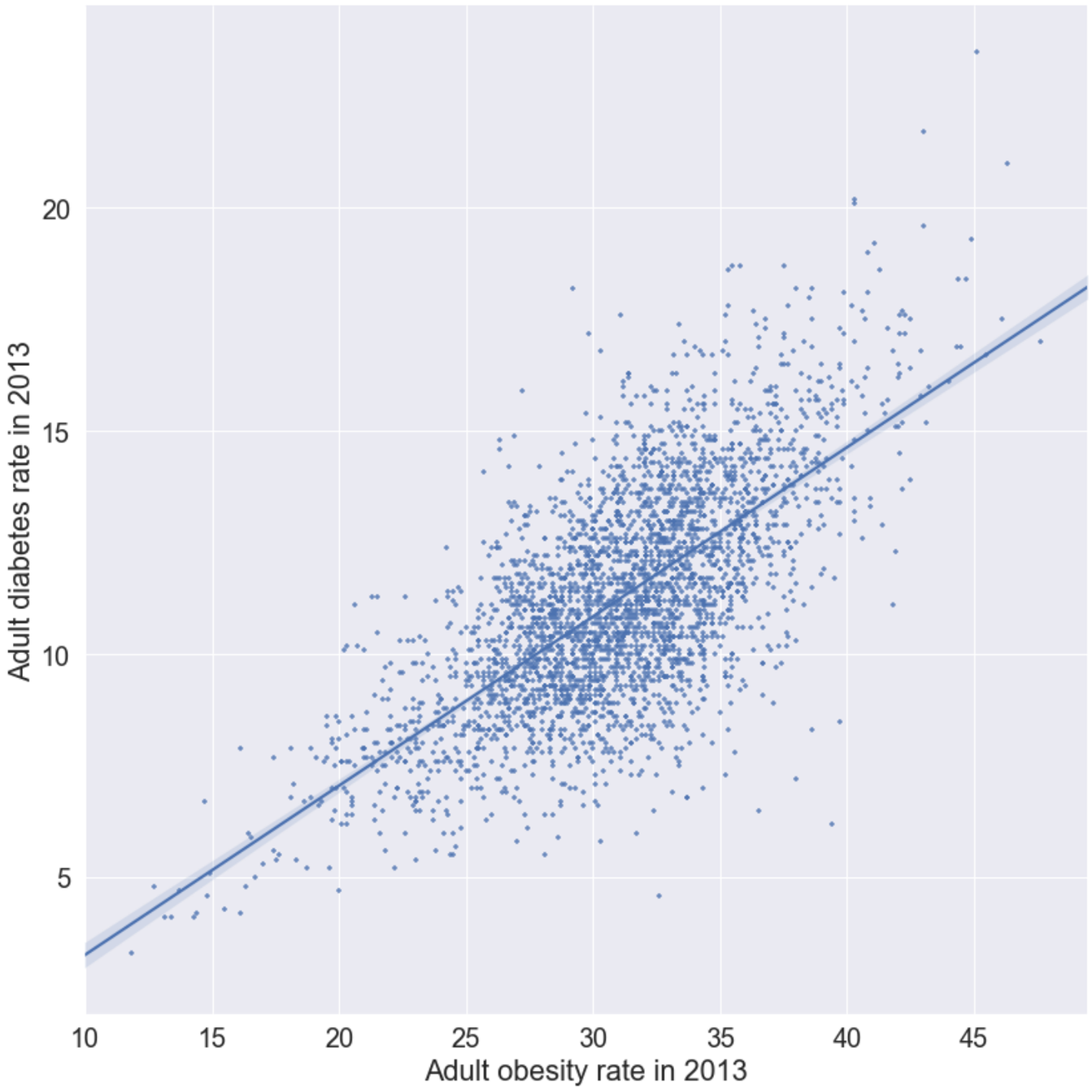


Abb. 2.4.1 Scatterplot: Adult Obesity vs. Diabetes

02 .4 Univariant Correlations

Diabetes > Obesity

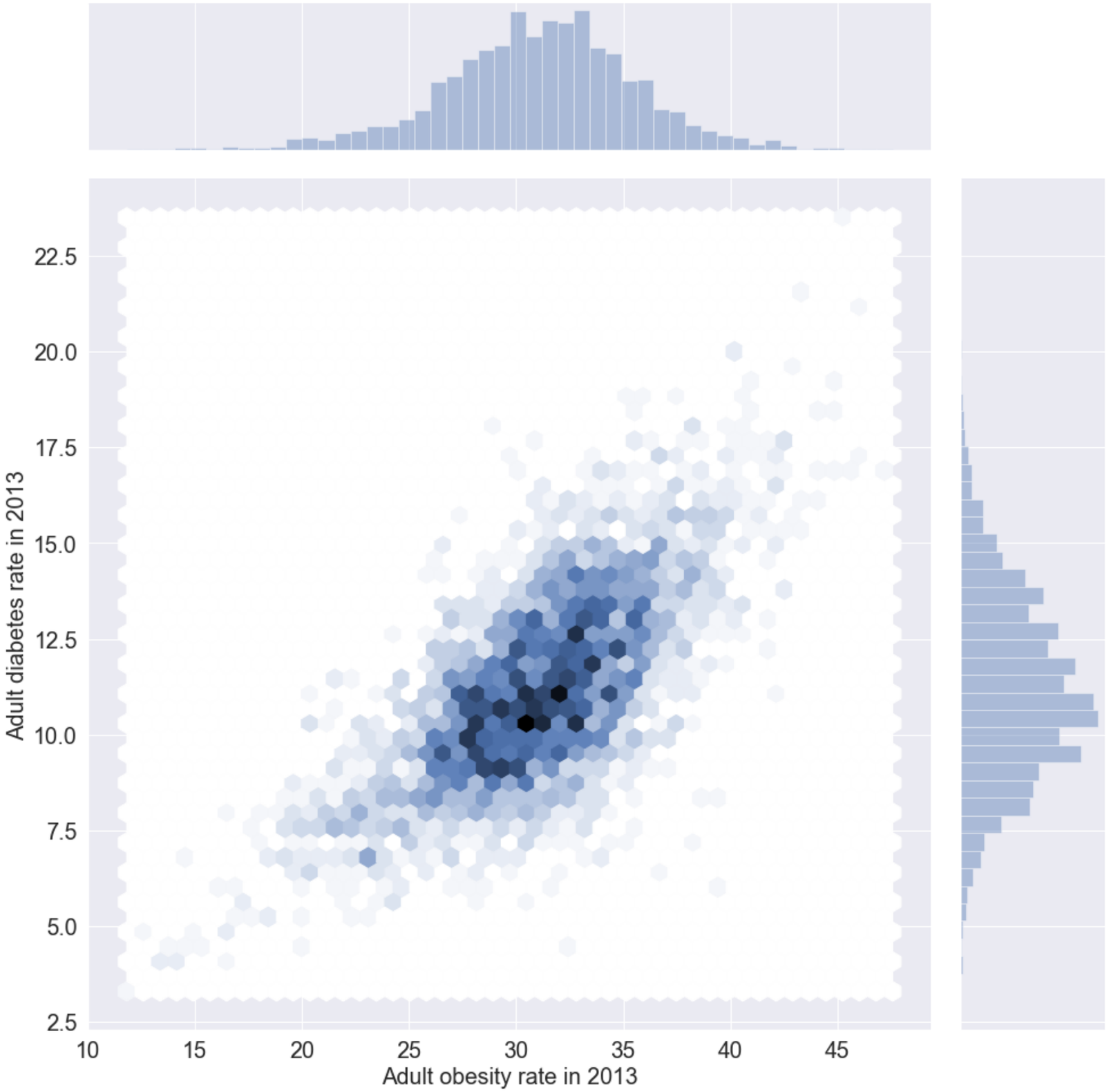


Abb. 2.4.2 Hexplot: Obesity vs. Diabetes

02 .4 Univariant Correlations

Fast Food Restaurants > Obesity

➤ Restaurants Description: The number of limited-service restaurants in the county. This includes establishments primarily engaged in providing food services, where patrons generally order or select items and pay before eating. Food and drink may be consumed on premises, taken out, or delivered to the customer's location.

02 .4 Univariant Correlations

Fast Food Restaurants > Obesity

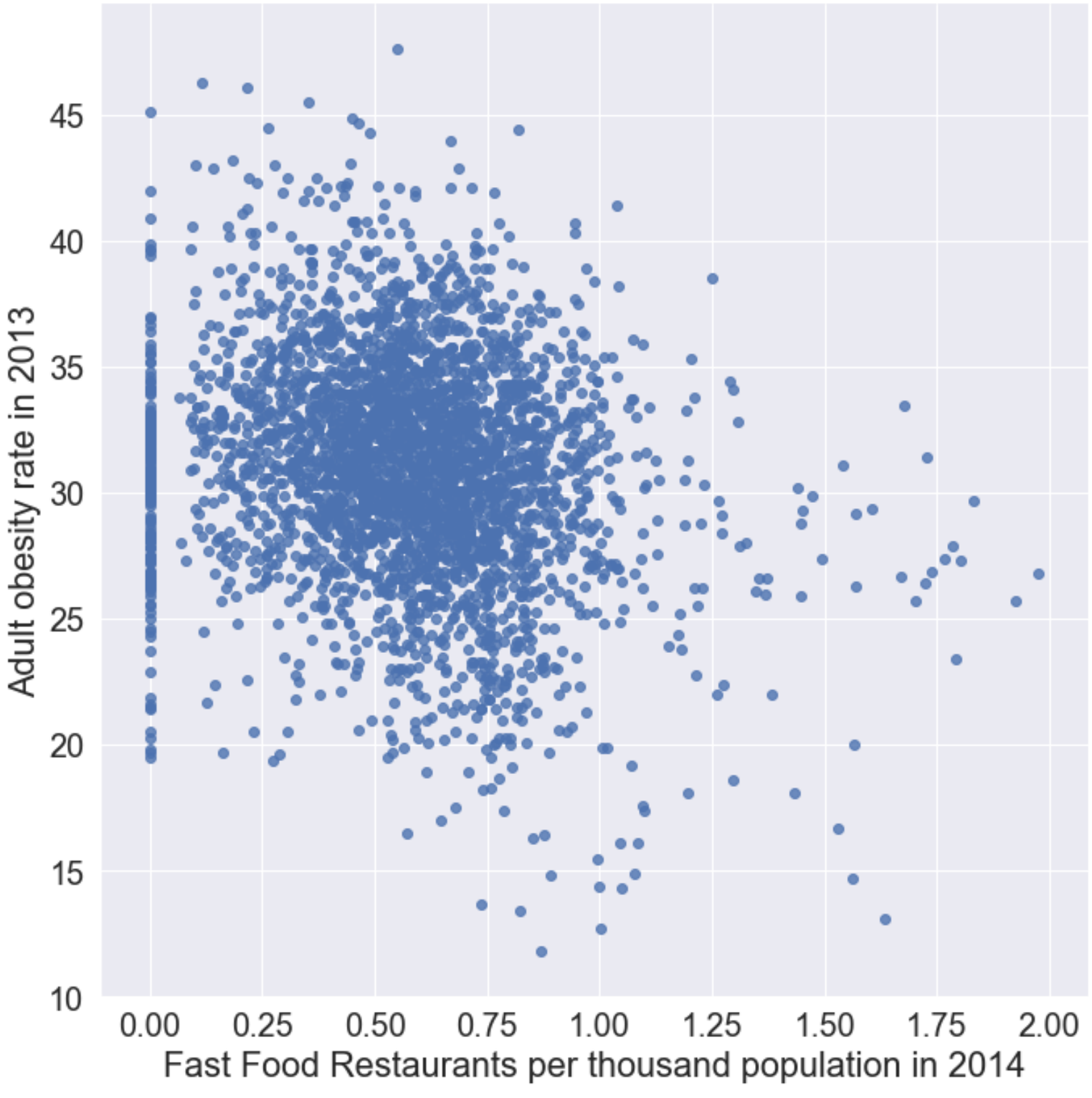


Abb. 2.4.3 Scatterplot: Obesity vs. Fast-Food Restaurants

02 .4 Univariant Correlations

Fast Food Expenditure > Restaurants per Thousands

➤ **Expenditure Description:** Average expenditures (in current dollars) on food purchased at limited-service restaurants in the county.

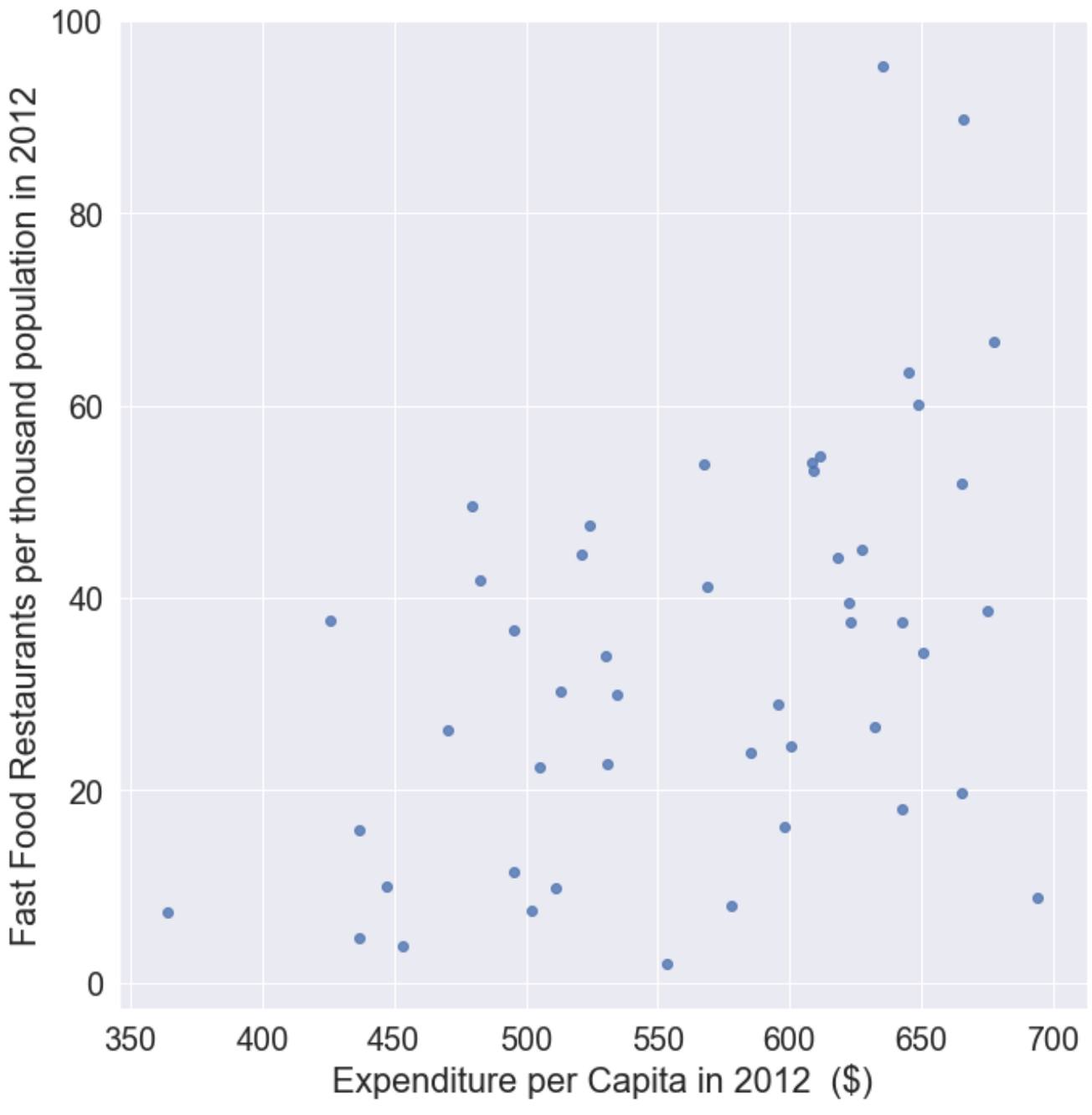


Abb. 2.4.4 Scatterplot: Expenditure vs. Fast-Food Restaurants per Thousand

02 .4 Univariant Correlations

Fast Food Expenditure > Obesity

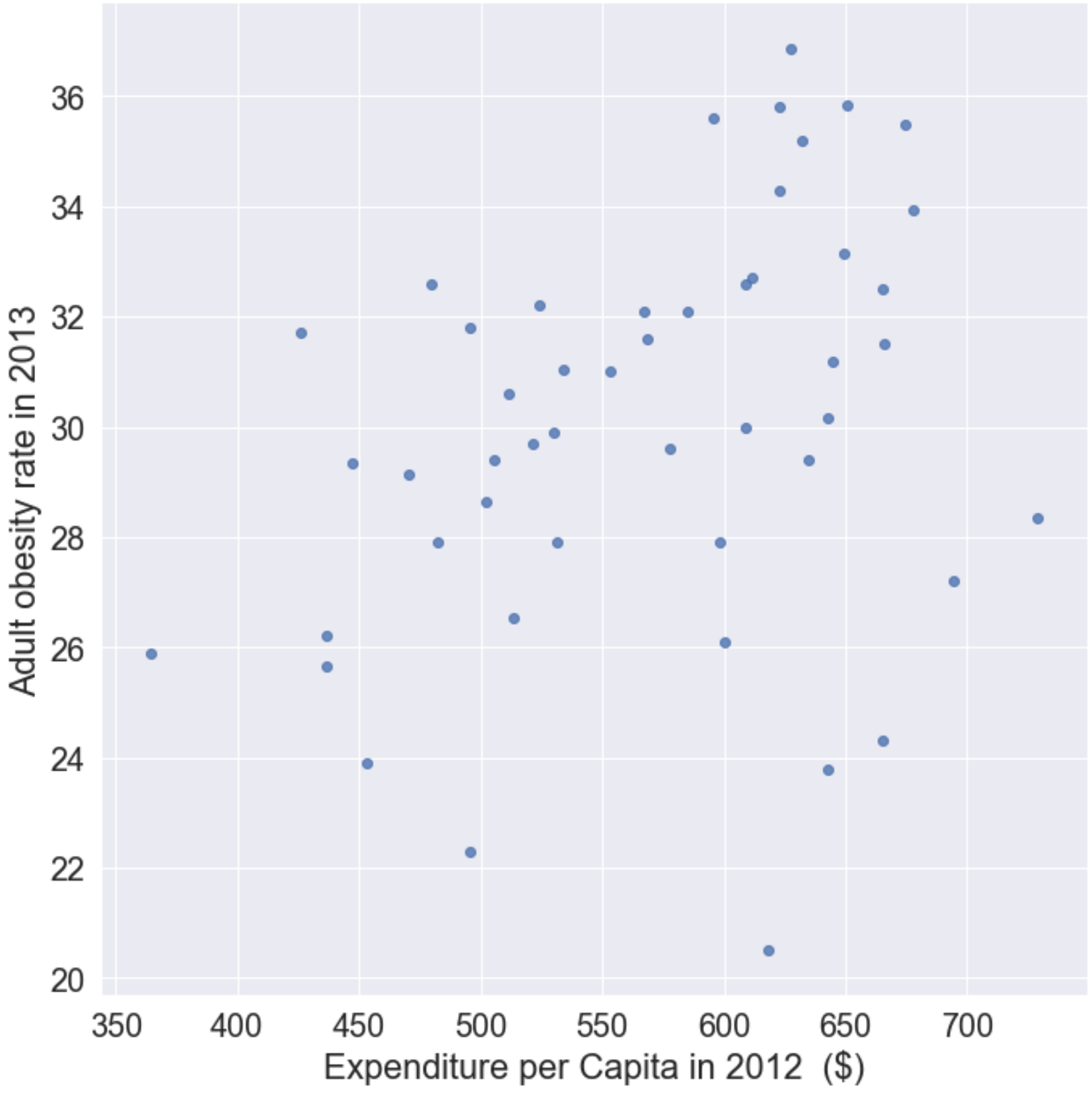


Abb. 2.4.5 Scatterplot: Obesity vs. Fast-Food Expenditure

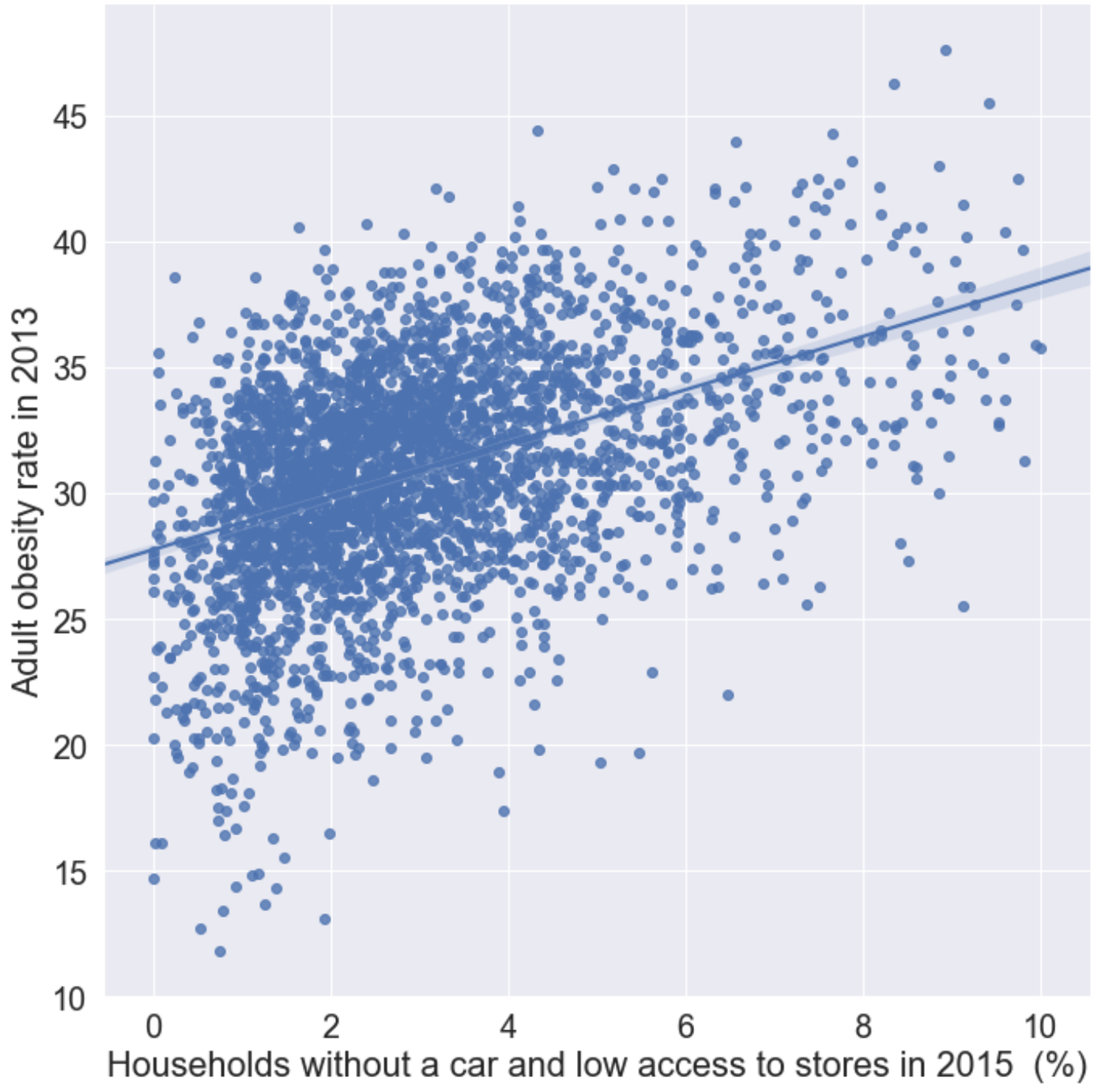
count	51.000000
mean	29.762745
std	3.949479
min	20.500000
25%	27.550000
50%	30.000000
75%	32.350000
max	36.850000

Abb. 2.4.6 Obesity States 2013

02 .4 Univariant Correlations

Low Store Access > Obesity

Abb. 2.4.7 Scatterplot: Obesity vs. Low Store Access (no car)



➤ Description: Percentage of housing units in a county without a car and more than 1 mile from a supermarket, supercenter or large grocery store.

02 .5 Conclusions
of the data-exploration

- The state level geo data show a definite distinction between states.
- Lower income equals higher obesity rates.
- The county level geo data findings are incoherent.
- This suggests that megacities like the ones just mentioned are in a category of their own.
- Obesity rates have seen a slight increase through the years 2008 and 2013.
- Observational results from graphs with diabetes and obesity, are not as strong as expected.
- Plotting single possible factors against obesity never yields the desired or expected results.
- This is probably due to the fact that obesity is influenced by many factors.

ATLAS 03

Prognosemodell

L-Regression

03 .1 Korrelationsanalyse

des gesamten Food-Atlas

- Brute-Force-Methode zur Analyse der Daten
- Auswertung folgender großer Datengruppen
  - Ethnische Gruppen
  - Armut und Einkommen
  - Bevölkerung
  - Fitness
  - Arten von Geschäften/Restaurants
  - Erreichbarkeit von Geschäften
  - Ernährungsunsicherheit
  - Altersgruppen
  - lokale Farmen

03 .1 Korrelationsanalyse

des gesamten Food-Atlas

➤ Pearson Maßkorrelationskoeffizient

$$\rho := \rho(X,Y) := \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \in [-1;1]$$

➤ Einordnung:

$$\begin{aligned} \rho \approx 0: &\quad \text{vernachlässigbare lineare Abhängigkeit zwischen X und Y} \\ 0.3 < |\rho| < 0.7: &\quad \text{schwacher linearer Zusammenhang zwischen X und Y} \\ |\rho| > 0.7: &\quad \text{starker linearer Zusammenhang zwischen X und Y} \end{aligned}$$

03 .1 Korrelationsanalyse des gesamten Food-Atlas

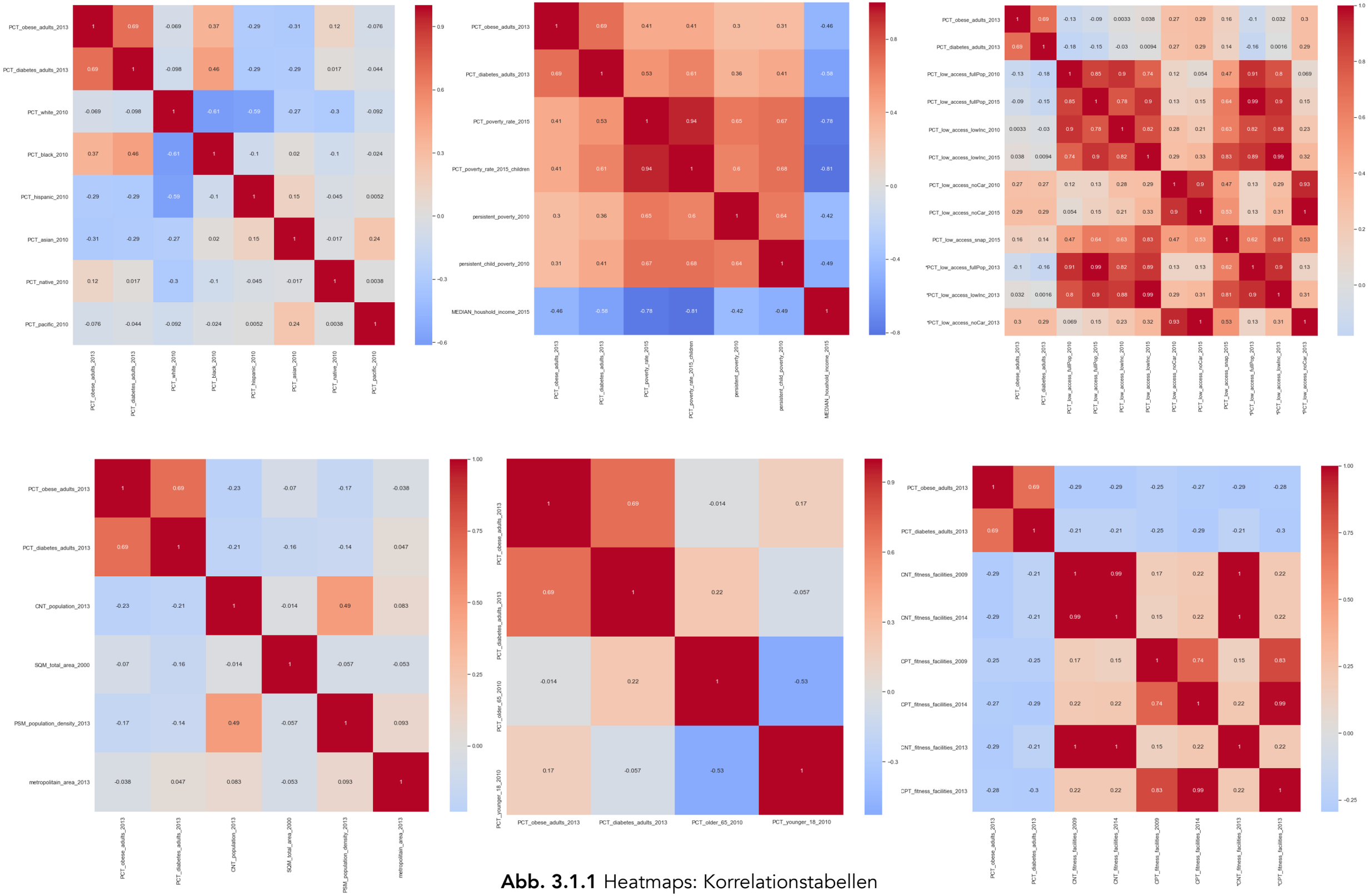
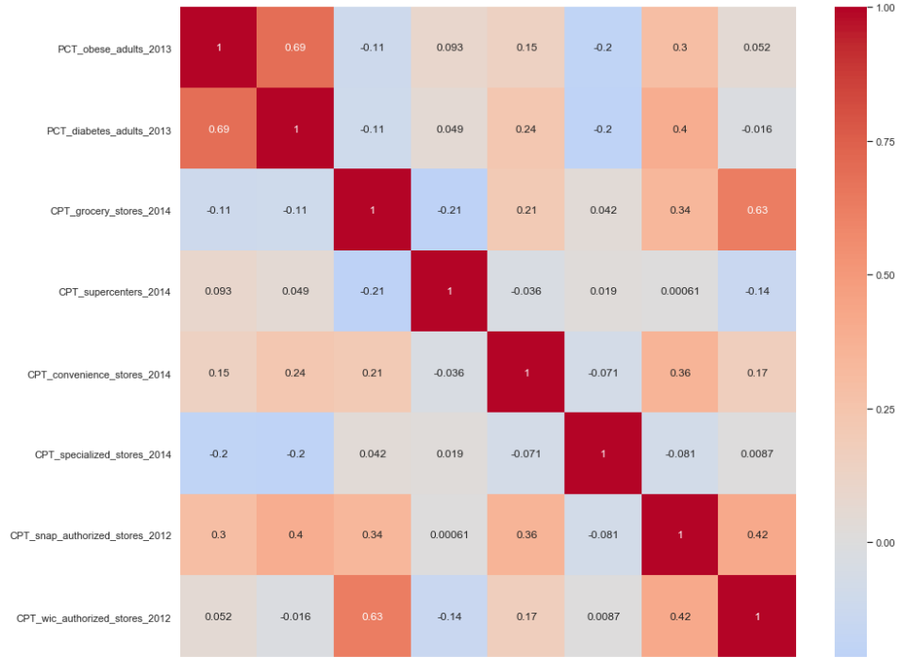
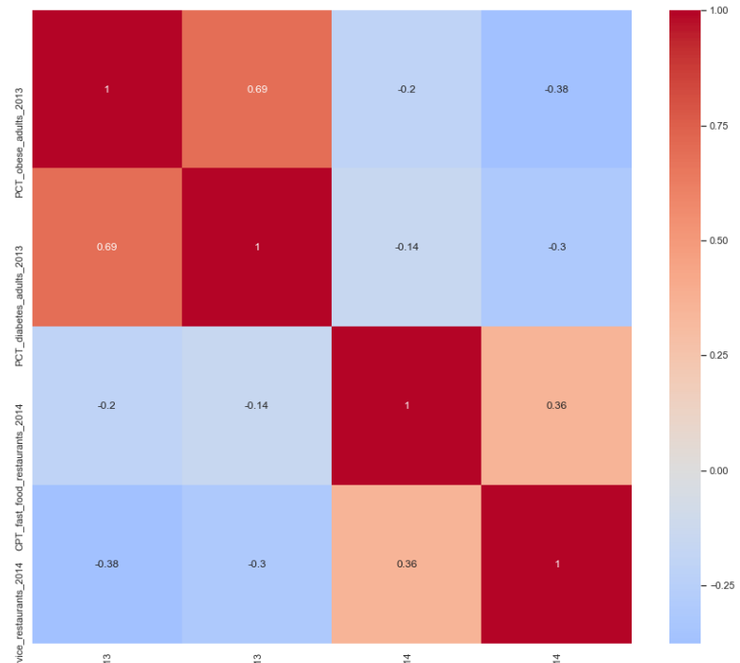
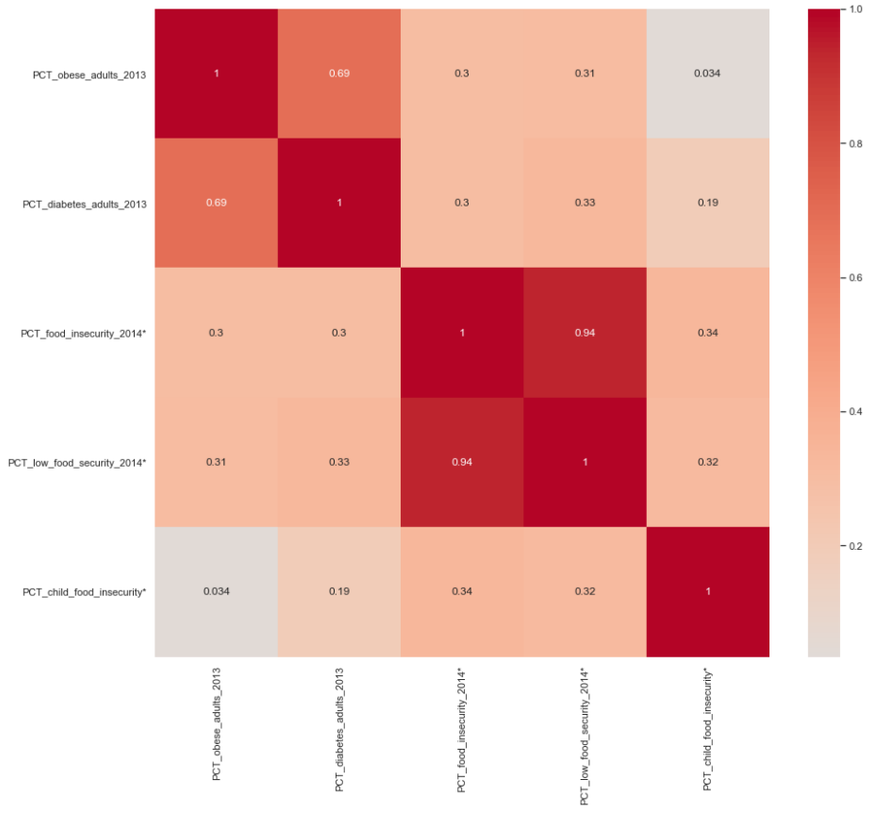


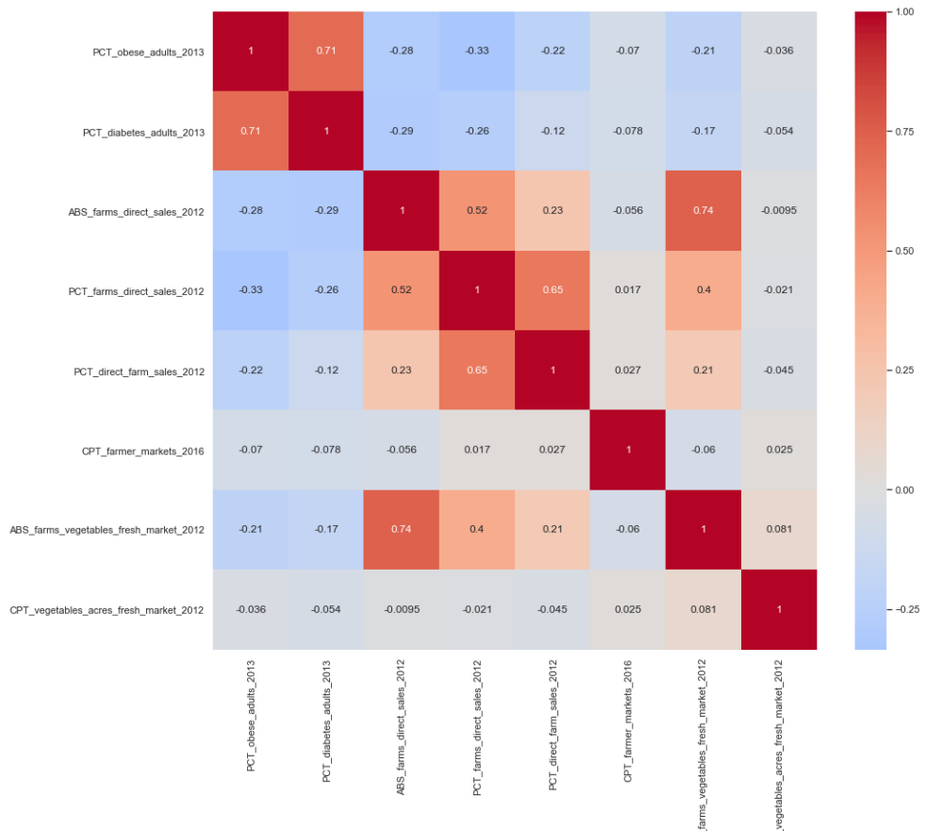
Abb. 3.1.1 Heatmaps: Korrelationstabellen

03 .1 Korrelationsanalyse des gesamten Food-Atlas









03 .1 Korrelationsanalyse

Relevante und korrelierende Daten

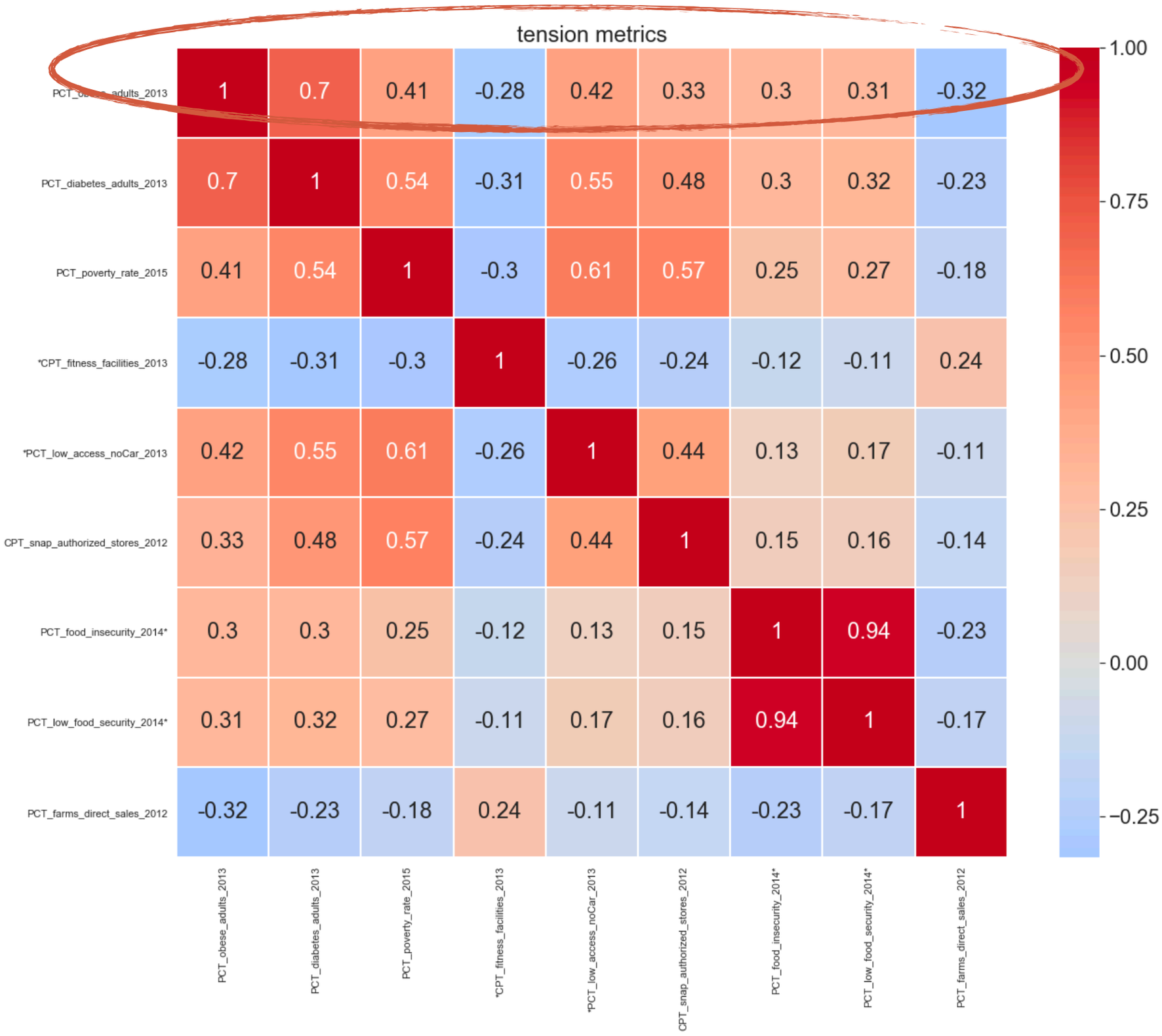


Abb. 3.1.2 Heatmap: korrelierende Daten zur Fettleibigkeit

03 .2 univariable lineare Regression

zwischen zwei Merkmalen

➤ **Vorhersagemodell:** von X auf Y schließen

$$f(X) \rightarrow Y$$

➤ **Methodik:** Kleinste-Quadrate-Methode

$$h_\theta(x) = \theta_0 + \theta_1 x$$

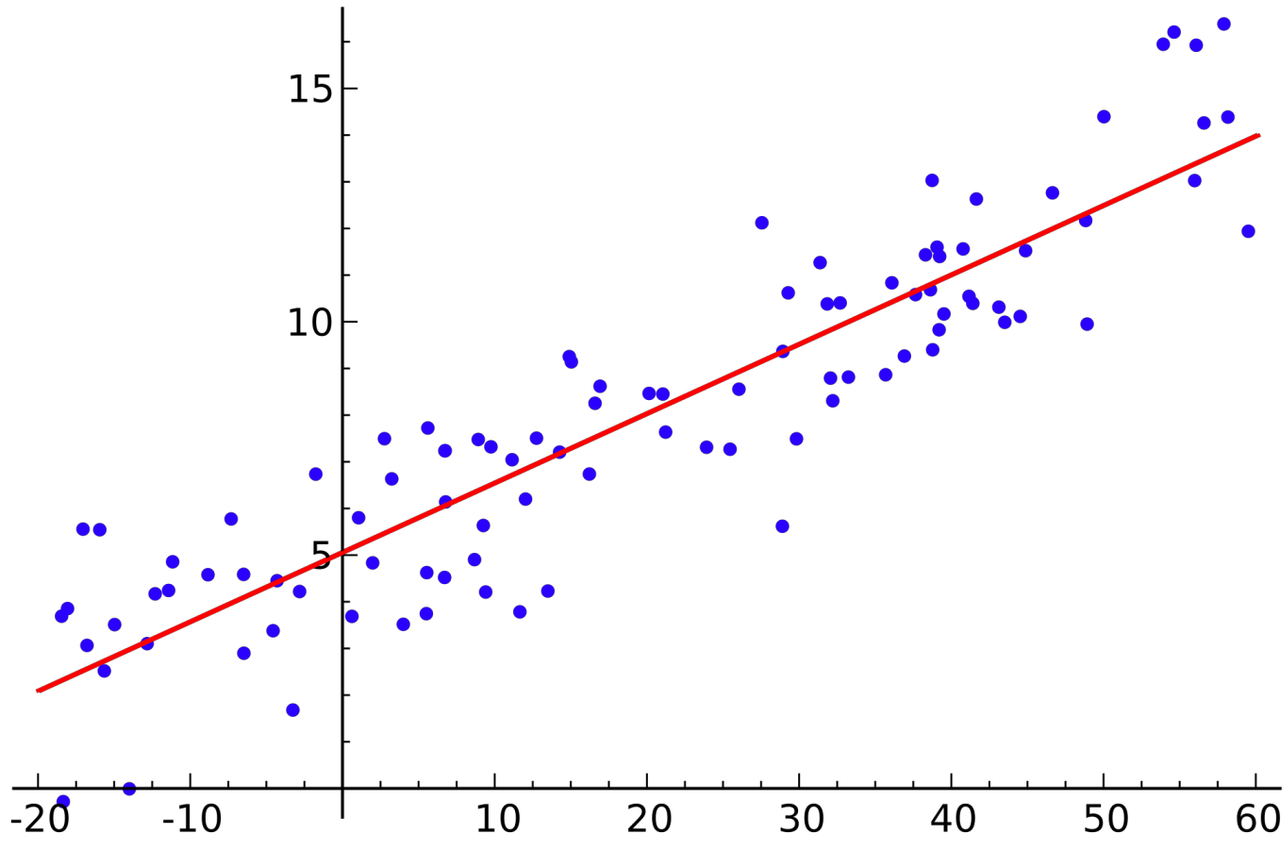
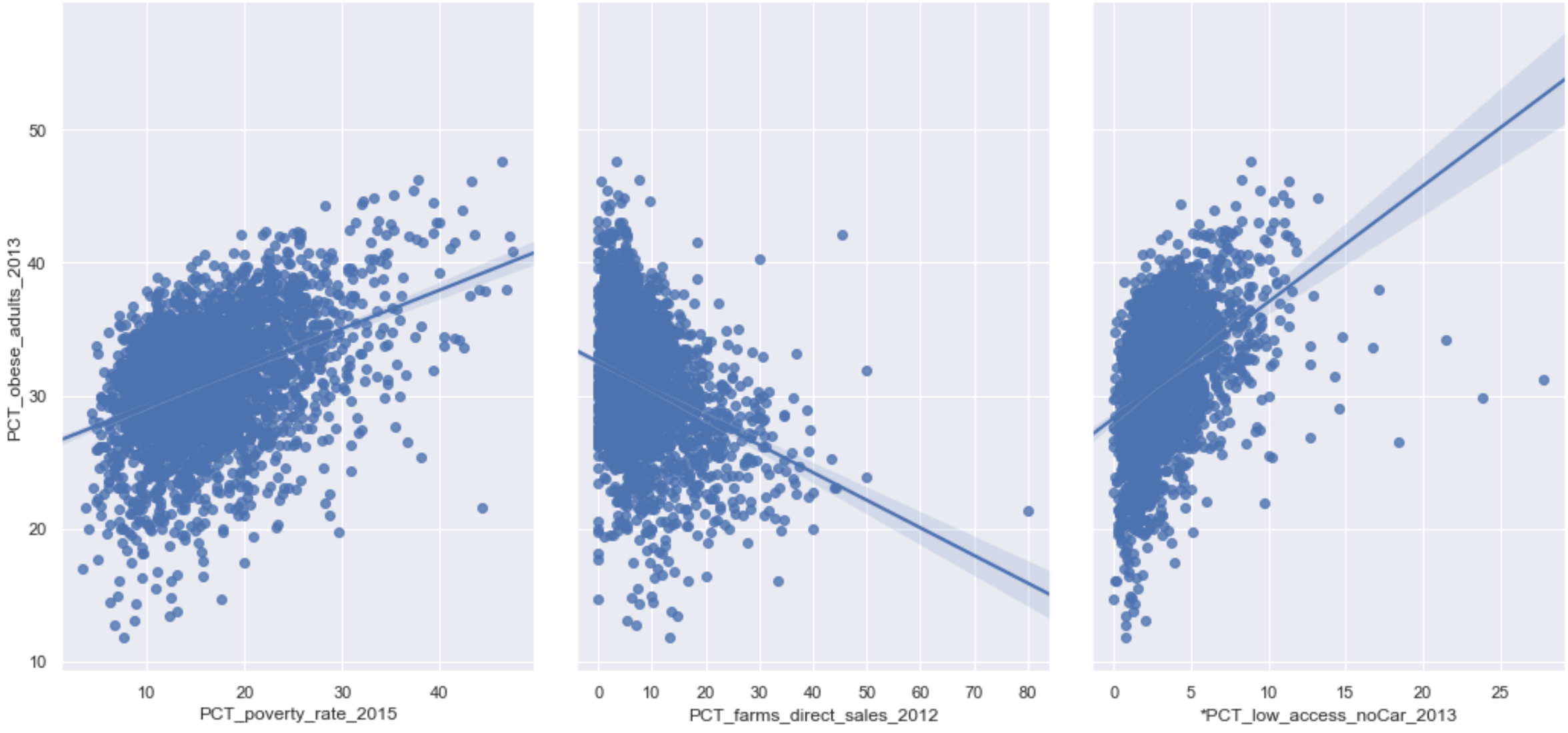


Abb. 3.2.1 Graph: Beispiel lineare Regression

03 .2 univariable lineare Regression

zwischen zwei Merkmalen

Abb. 3.2.2 Scatterplot: lineare Regression zur Fettleibigkeit



- keine präzise Aussage

03 .3 multivariable lineare Regression

der korrelierenden Merkmale

- Fettleibigkeit durch mehrere Merkmalsausprägungen bedingt
  - also auch mehrere Regressoren

$$f(X_1, X_2, \ldots, X_n) \rightarrow Y$$

- Bau eines Vorhersagemodells durch Training mit Testdaten
  - 50% (ca. 1750 Daten) des Datensatzes zum Trainieren des Modells
  - 50% des Datensatzes zur Auswertung des Modells

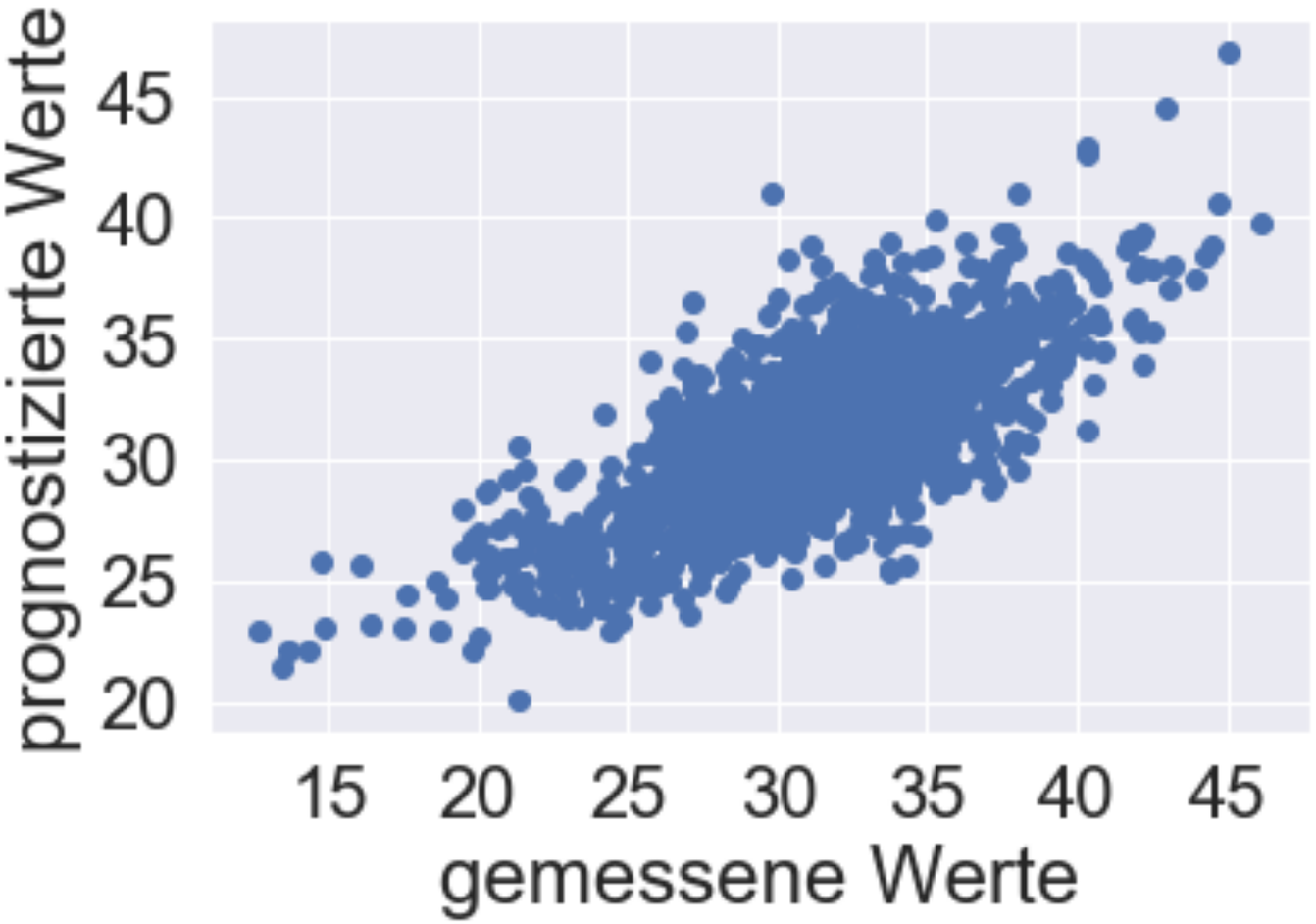
03 .3 multivariable lineare Regression

der korrelierenden Merkmale

	Coefficients
PCT_diabetes_adults_2013	1.168278
PCT_poverty_rate_2015	0.000418
***CPT_fitness_facilities_2013**	-0.300320
***PCT_low_access_noCar_2013**	0.126281
CPT_snap_authorized_stores_2012	-0.305264
PCT_food_insecurity_2014*	0.141917
PCT_low_food_security_2014*	0.015442
PCT_farms_direct_sales_2012	-0.087378

Abb. 3.3.1 Tabelle: Ausschnitt der Koeffizienten

Abb. 3.3.2 Scatterplot: Abweichung der Prognose



03 .4 Auswertungsmetriken

der multivariablen linearen Regression

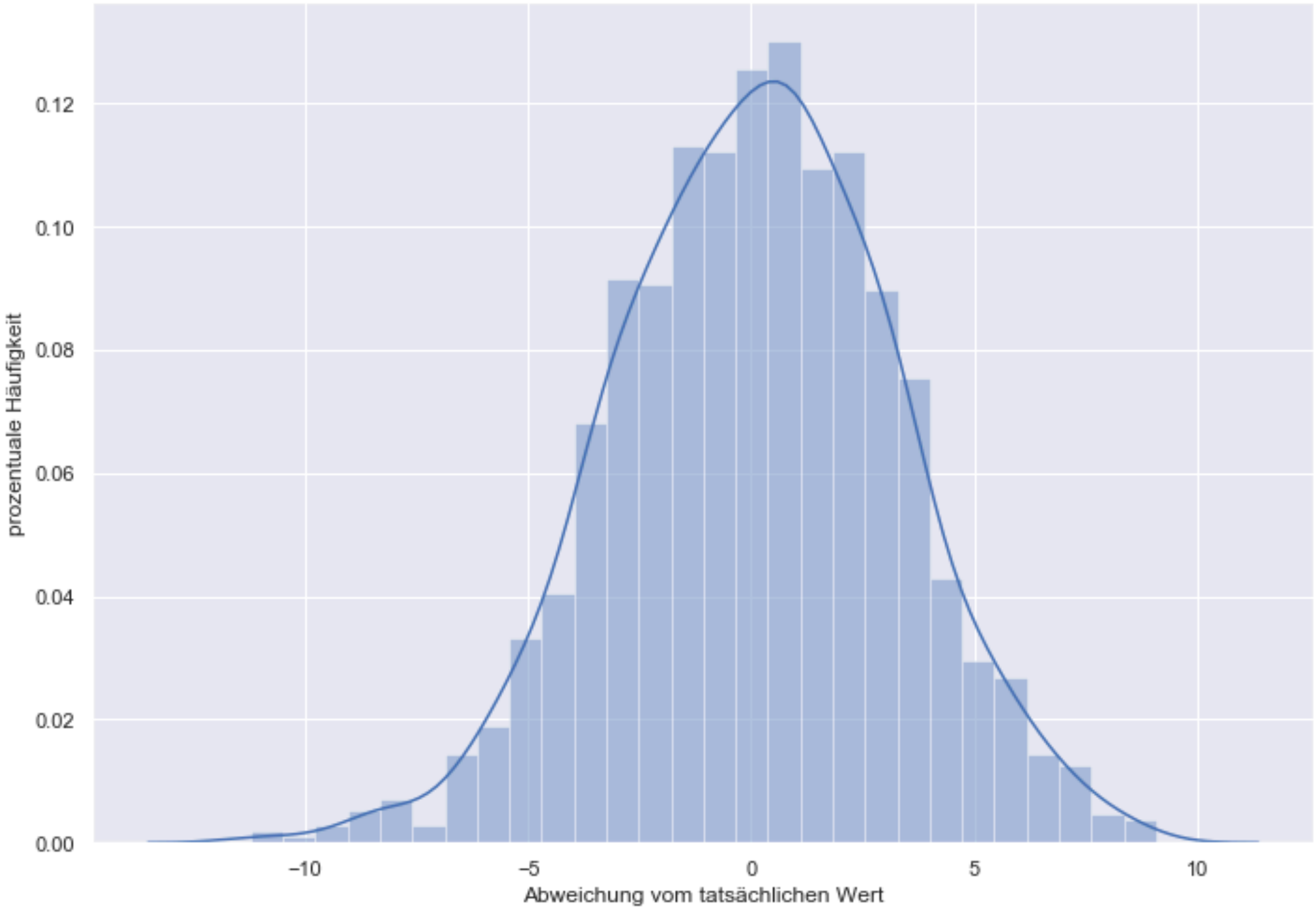


Abb. 3.4.1 Histogramm: Häufigkeiten der Abweichungen

03 .4 Auswertungsmetriken

der multivariablen linearen Regression

- Mean Absolute Error (MAE)

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad = 2.5016$$

- Mean Squared Error (MSE)

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \quad = 9.8353$$

- Root Mean Squared Error (RSME)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} = 3.1361$$

03 .4 Auswertungsmetriken

der multivariablen linearen Regression

- Bestimmtheitsmaß R^2
- Anpassungsgüte einer Regression

$$R^2 := \frac{\sum \left(\hat{y}_i - \bar{y}\right)^2}{\sum \left(y_i - \bar{y}\right)^2} = 0.5133$$

- über die Hälfte der Daten können durch unser Modell erklärt werden
- Daten des Food-Atlas reichen nicht aus zur präzisen Prognose

National Health and Nutrition Examination Survey

NHANES

NHANES 04

Daten-Exploration

04 .1 Datenquelle

Demographische & Untersuchungsdaten

Abb. 4.1.1 Online-Portal: CDC NHANES

Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

Search NCHS | SEARCH

CDC A-Z INDEX

National Center for Health Statistics

National Health and Nutrition Examination Survey

National Health and Nutrition Examination Survey

CDC > National Health and Nutrition Examination Survey > Questionnaires, Datasets, and Related Documentation

- About NHANES +
- What's New +
- Questionnaires, Datasets, and Related Documentation −
  - Survey Methods and Analytic Guidelines
  - Search Variables
  - All Continuous NHANES +
  - NHANES 2017-2018 +
  - **NHANES 2015-2016** −
    - Demographics Data
    - Dietary Data
    - Examination Data
    - Laboratory Data
    - Questionnaire Data

NHANES 2015-2016

Data, Documentation, Codebooks, SAS Code

- Demographics Data
- Dietary Data
- Examination Data
- Laboratory Data
- Questionnaire Data
- Limited Access Data

Contents in Detail

- Questionnaire Instruments

Using the Data

- Overview
- Release Notes
- Laboratory Data Overview
- Questionnaire Data Overview
- Examination Data Overview
- Survey Methods and Analytic Guidelines
- Response Rates and Population Totals
- NHANES Web Tutorial

04 .1 Datenquelle

Selektierte Indikatoren

- BMI (Zielvariable)
- Einkommen der Haushalte
- Mitglieder im Haushalt
- Bildungsstufe (Referenz)
- Ethnische Gruppierung
- Alter / Geschlecht

```
RangeIndex: 47873 entries, 0 to 47872
Data columns (total 26 columns):
id                          47873 non-null int64
CODE_status                 47873 non-null int64
CM_height                   43969 non-null float64
KG_weight                   47374 non-null float64
KG_M2_bmi                   43912 non-null float64
CODE_bmi_category_youth     10155 non-null float64
BINARY_gender               47873 non-null int64
YEARS_age                   47873 non-null int64
CODE_race                   47873 non-null int64
CODE_race_extended          28524 non-null float64
CODE_education_youth        12493 non-null float64
CODE_education_adults       27662 non-null float64
CODE_marital                27662 non-null float64
CNT_houshold_members        47873 non-null int64
CNT_family_members          47873 non-null int64
CNT_children_age_to5        28524 non-null float64
CNT_children_age_6to17      28524 non-null float64
CNT_adults_older_60         28524 non-null float64
BINARY_gender_hh            47873 non-null int64
CODE_education_hh           46480 non-null float64
YEARS_age_hh                47873 non-null int64
CODE_marital_hh             47139 non-null float64
CODE_income_houshold        47292 non-null float64
CODE_income_family          47345 non-null float64
RATIO_poverty               43720 non-null float64
year                        47873 non-null int64
```

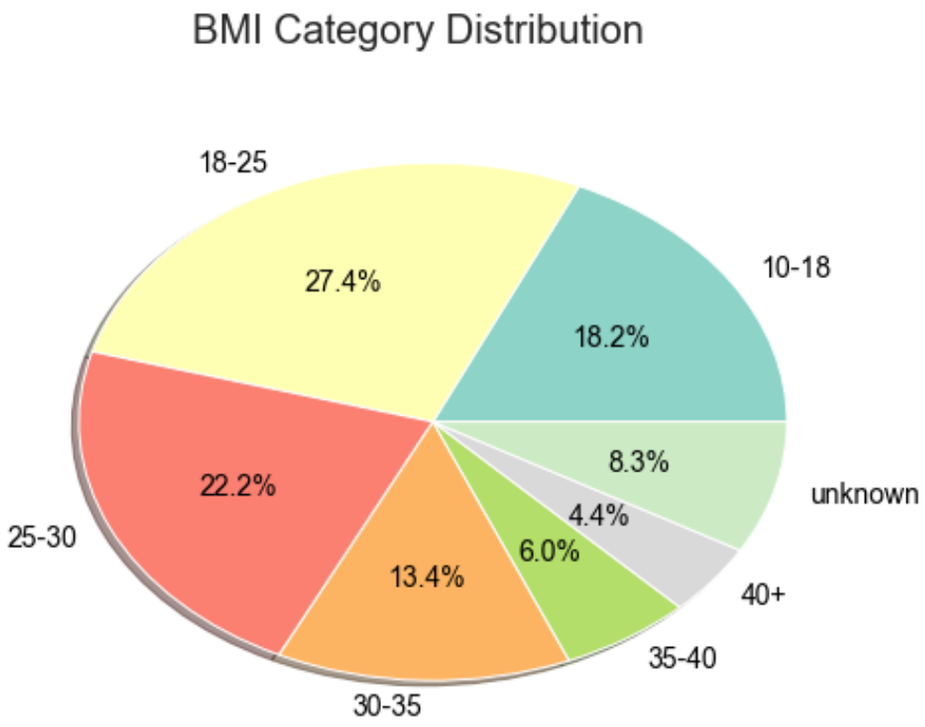
Abb. 4.2.2 Bereinigen Dataframe

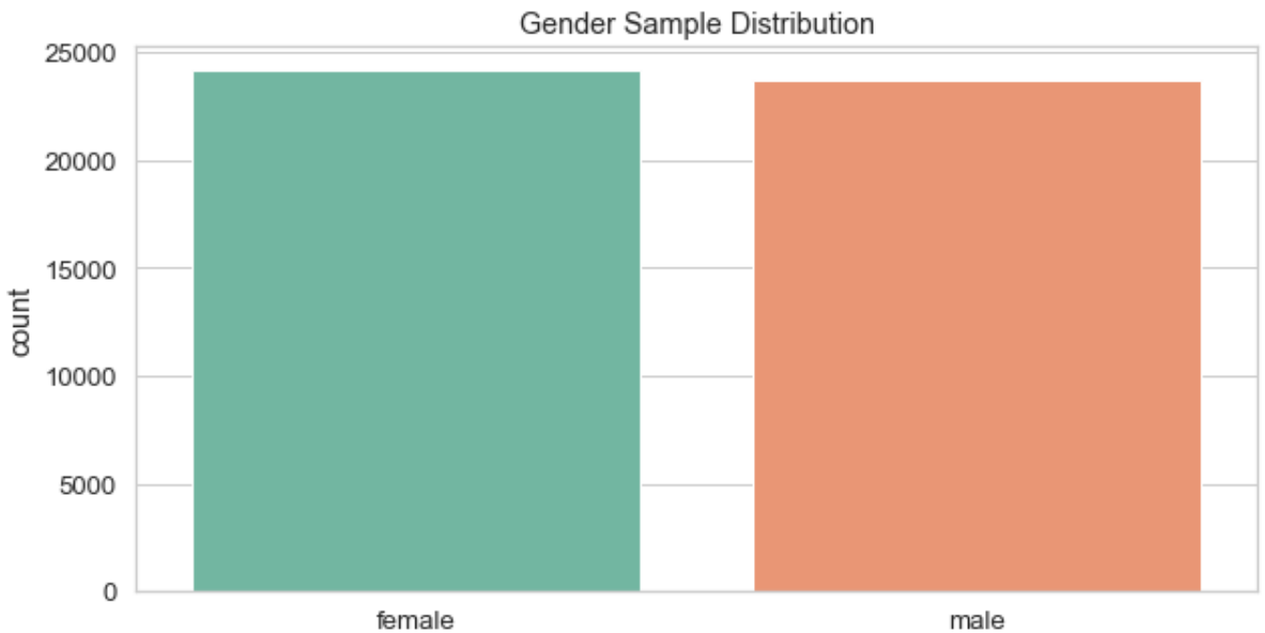
	id	CODE_status	CM_height	KG_weight	KG_M2_bmi	CODE_bmi_category_youth	BINARY_gender	YEARS_age	CODE_race
0	41475	3	154.7	138.9	58.04	NaN	2	62	5
1	41476	1	120.4	22.0	15.18	NaN	2	6	5
2	41477	1	167.1	83.9	30.05	NaN	1	71	3
3	41478	1	NaN	11.5	NaN	NaN	2	1	3
4	41479	1	154.4	65.7	27.56	NaN	1	52	1
5	41480	1	122.7	27.0	17.93	NaN	1	6	1
6	41481	1	182.7	77.9	23.34	NaN	1	21	4
7	41482	1	173.8	101.6	33.64	NaN	1	64	1
8	41483	3	173.8	133.1	44.06	NaN	1	66	4
9	41484	1	NaN	9.3	NaN	NaN	1	0	3
10	41485	1	157.9	64.8	25.99	NaN	2	30	2

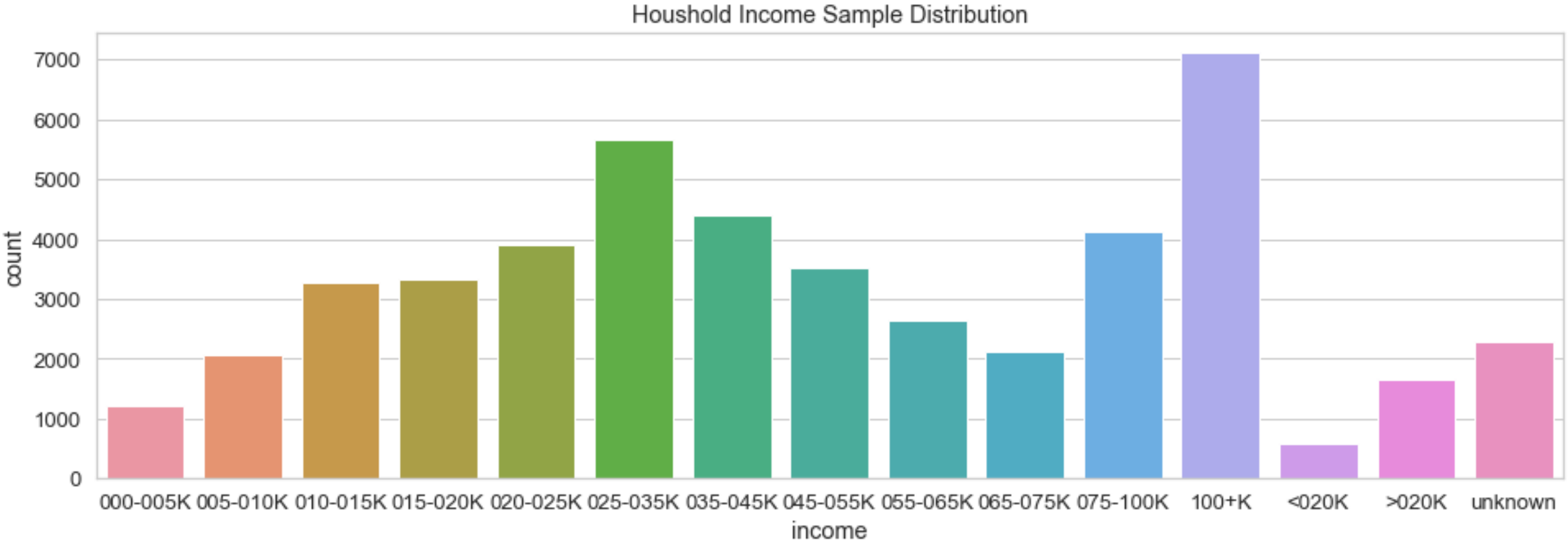
04 .1 Datenquelle

Aufbereitung und Kategorisierung

Abb. 4.1.3 NHANES: Verteilung der Daten







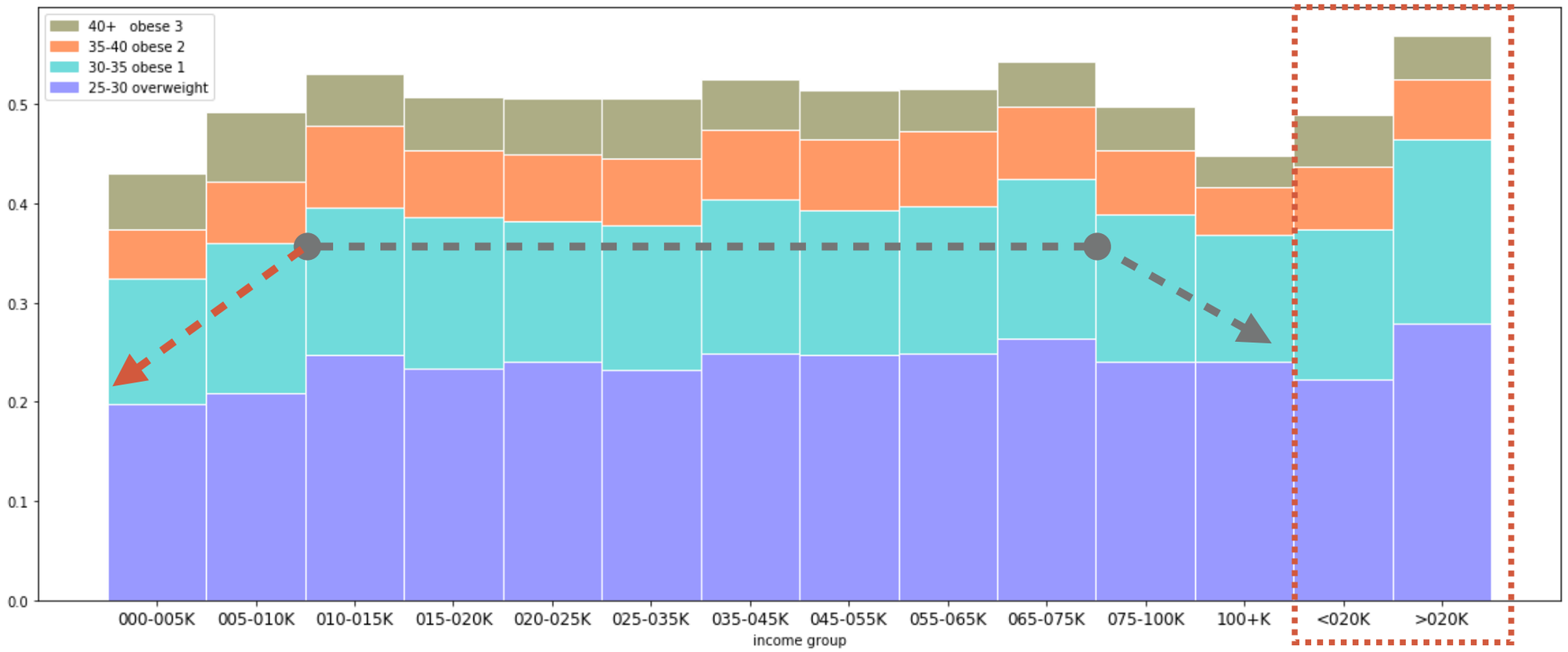
04 .2 BMI ➠ Einkommen

Prozentuale Anteile Übergewicht ++

Abb. 4.2.1 Kontingenztabelle: BMI Anteile nach Einkommen

houshold_income bmi	000- 005K	005- 010K	010- 015K	015- 020K	020- 025K	025- 035K	035- 045K	045- 055K	055- 065K	065- 075K	075- 100K	100+K	<020K	>020K	row_total
10-18	282	406	577	647	752	1050	770	593	442	368	736	1412	121	224	8380
18-25	325	548	801	837	999	1505	1148	993	731	531	1167	2250	149	429	12413
25-30	210	393	732	704	854	1202	1004	810	605	520	912	1587	118	424	10075
30-35	135	284	437	462	500	759	625	472	363	317	566	845	80	282	6127
35-40	52	116	243	206	242	354	282	235	183	143	245	322	34	91	2748
40+	59	131	156	162	197	312	209	163	105	90	167	210	27	69	2057
col_total	1063	1878	2946	3018	3544	5182	4038	3266	2429	1969	3793	6626	529	1519	41800

Abb. 4.2.2 Barplot: BMI Anteile nach Einkommen



04 .2 BMI ➠ Einkommen

Chi-Quadrat-Test

$$\tilde{h}_{ij} = \frac{h_{i*}h_{*j}}{n} \qquad \chi^2 = \sum_{i=1}^{j}\sum_{k=1}^{m}\frac{(h_{ij}-\tilde{h}_{ij})^2}{\tilde{h}_{ij}}$$

Abb. 4.2.3 Kontingenztabelle: BMI x Einkommen - Erwartungswerte

houshold_income bmi	000-005K	005-010K	010-015K	015-020K	020-025K	025-035K	035-045K	045-055K	055-065K	065-075K	075-100K	100+K
10-18	214.86	379.60	595.47	610.02	716.34	1047.43	816.19	660.15	490.97	397.99	766.67	1339.30
18-25	316.48	559.12	877.09	898.52	1055.12	1542.79	1202.20	972.36	723.16	586.21	1129.26	1972.70
25-30	254.92	450.37	706.49	723.75	849.89	1242.70	968.36	783.23	582.50	472.19	909.61	1588.99
30-35	154.16	272.36	427.24	437.68	513.97	751.52	585.61	473.65	352.26	285.55	550.08	960.93
35-40	70.14	123.92	194.39	199.14	233.85	341.93	266.44	215.50	160.28	129.92	250.28	437.21
40+	52.44	92.64	145.33	148.88	174.83	255.63	199.20	161.11	119.82	97.13	187.11	326.87
All	1063.00	1878.00	2946.00	3018.00	3544.00	5182.00	4038.00	3266.00	2429.00	1969.00	3793.00	6626.00

Abb. 4.2.4 Kontingenztabelle: BMI x Einkommen - Abweichungen

houshold_income bmi	000-005K	005-010K	010-015K	015-020K	020-025K	025-035K	035-045K	045-055K	055-065K	065-075K	075-100K	100+K
10-18	67.14	26.40	-18.47	36.98	35.66	2.57	-46.19	-67.15	-48.97	-29.99	-30.67	72.70
18-25	8.52	-11.12	-76.09	-61.52	-56.12	-37.79	-54.20	20.64	7.84	-55.21	37.74	277.30
25-30	-44.92	-57.37	25.51	-19.75	4.11	-40.70	35.64	26.77	22.50	47.81	2.39	-1.99
30-35	-19.16	11.64	9.76	24.32	-13.97	7.48	39.39	-1.65	10.74	31.45	15.92	-115.93
35-40	-18.14	-7.92	48.61	6.86	8.15	12.07	15.56	19.50	22.72	13.08	-5.28	-115.21
40+	6.56	38.36	10.67	13.12	22.17	56.37	9.80	1.89	-14.82	-7.13	-20.11	-116.87

04 .2 BMI ➠ Einkommen

Chi-Quadrat-Test

➤ Korrigierter Kontingenzkoeffizient:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

$$C_{korr} = \frac{C}{C_{max}}, \quad C_{max} = \sqrt{\frac{k-1}{k}} \quad mit \quad k = min(x, y)$$

$$DF = (k-1)(m-1) = 55 \quad mit \quad k = 6 \quad und \quad m = 12$$

	values
Chi2	301.213
C_norm	0.0867198
C_max	0.912871
C_corr	0.0949968
n_samples	39752
k_min	6
degree_free	55
sigma	0.05
Chi2_significant	73.3115

➤ Variablen sind stochastisch abhängig!

$Chi^2 = 301{,}23$ Signifikanz 5% , DF = 55 : $Chi^2 = 73{,}311$

Die Wahrscheinlichkeit für den Chi^2 Wert liegt unter 5%.

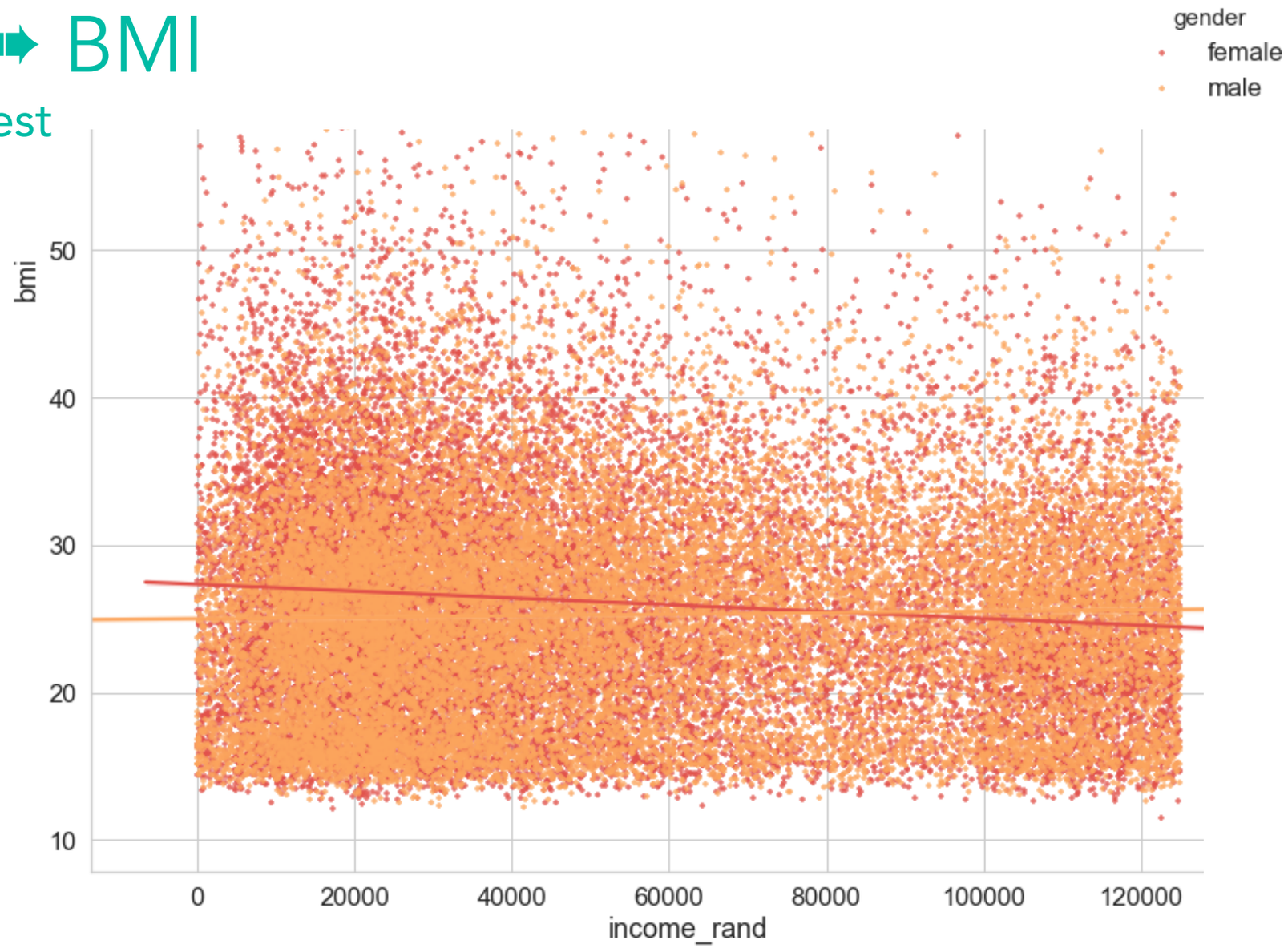
➤ Der korrigierte Kontingenzkoeffizient liegt jedoch dicht bei 0.

$C_{korr} = 0{,}095$ 9,5%

04 .2 Einkommen ➠ BMI

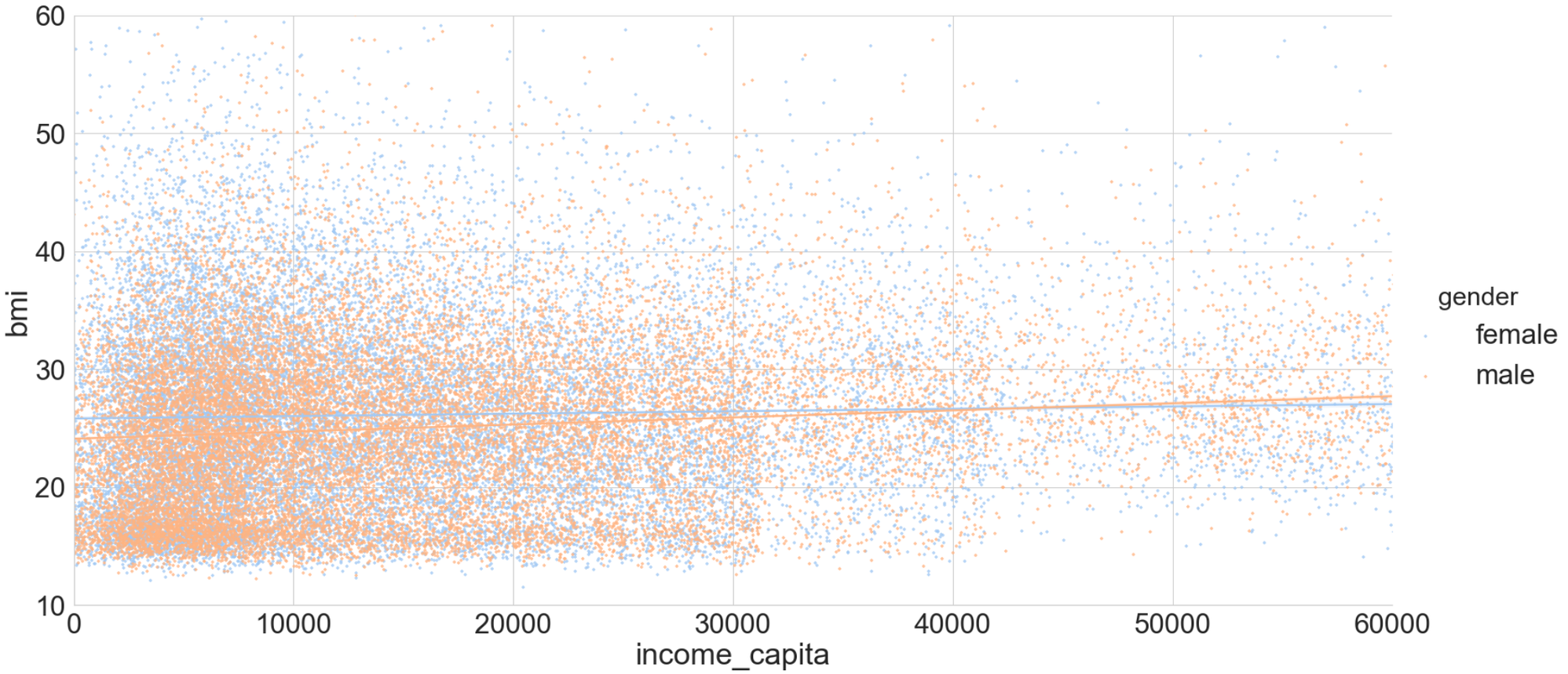
Lineare Regression & p-Test

	values
slope	-9.57863e-06
intercept	26.2167
r	-0.0437309
r2	0.00191239
p	2.71115e-18
std_error	1.09757e-06



	id	income_group	income_code	bmi	gender	members	income_mean	income_rand	income_capita
0	41475	025-035K	6	58.04	female	2	30000.0	26002.0	13001.000000
1	41476	100+K	12	15.18	female	6	112500.0	113606.0	18934.333333
2	41477	020-025K	5	30.05	male	2	22500.0	22955.0	11477.500000
3	41479	045-055K	8	27.56	male	5	50000.0	51288.0	10257.600000
4	41480	035-045K	7	17.93	male	4	40000.0	42234.0	10558.500000

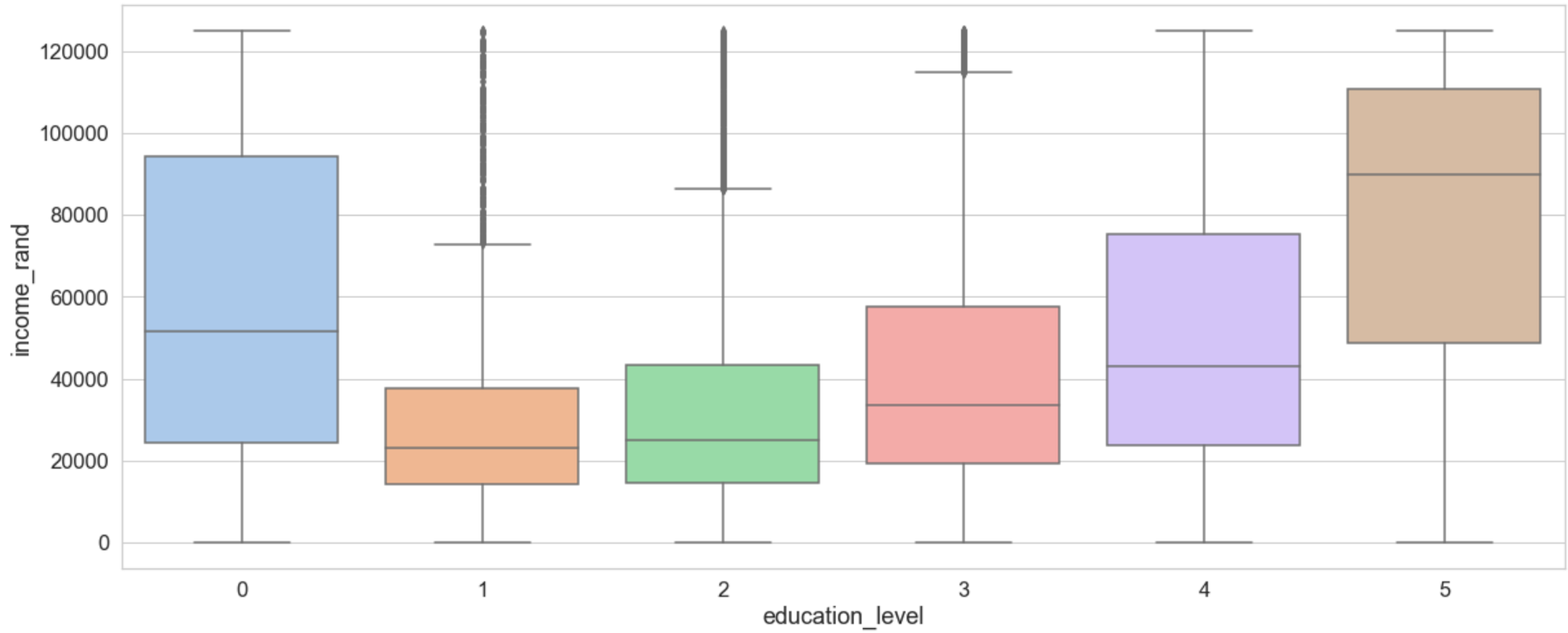
04 .2 Einkommen ➠ BMI mit Anzahl Haushaltsmitglieder



04 .3 Bildung ➠ Einkommen

Boxplot & Heatmap

Abb. 4.3.1 Boxplot: Bildung > Einkommen



- r = 39,83%

  |-0,3988| > 0,3

	values
Chi2	10133.5
C_norm	0.455175
C_max	0.894427
C_corr	0.508901
n_samples	38777
k_min	5
degree_free	44
sigma	0.05
Chi2_significant	60.4809

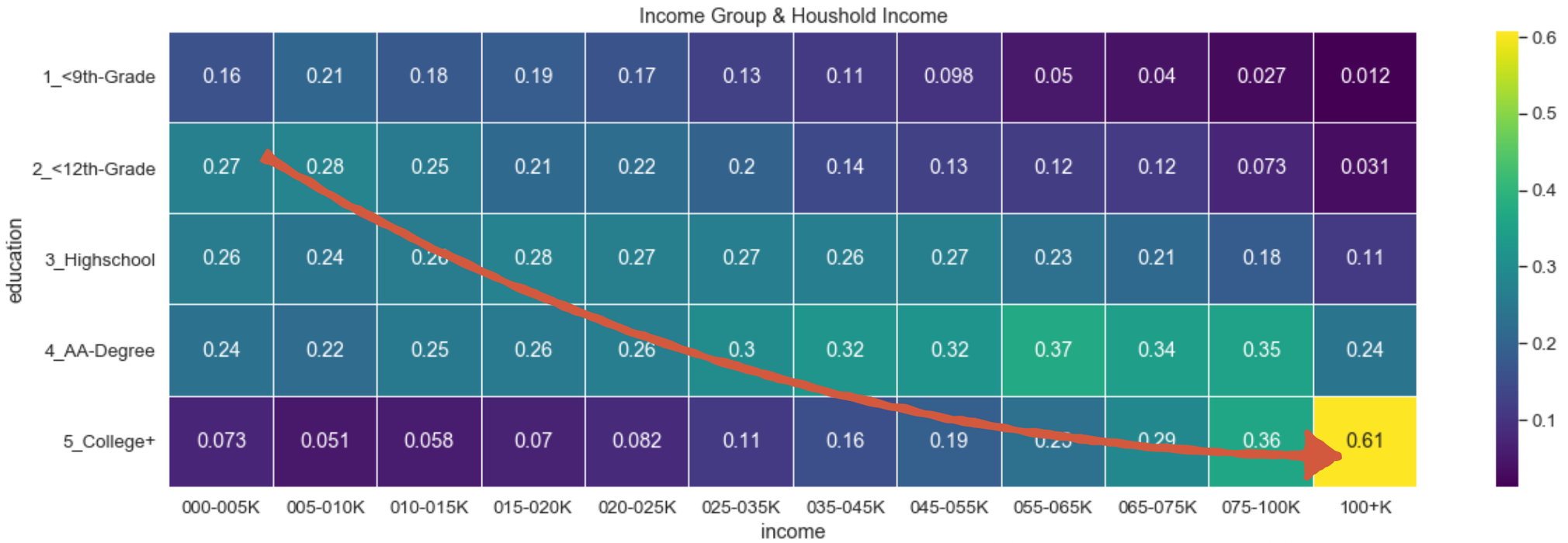


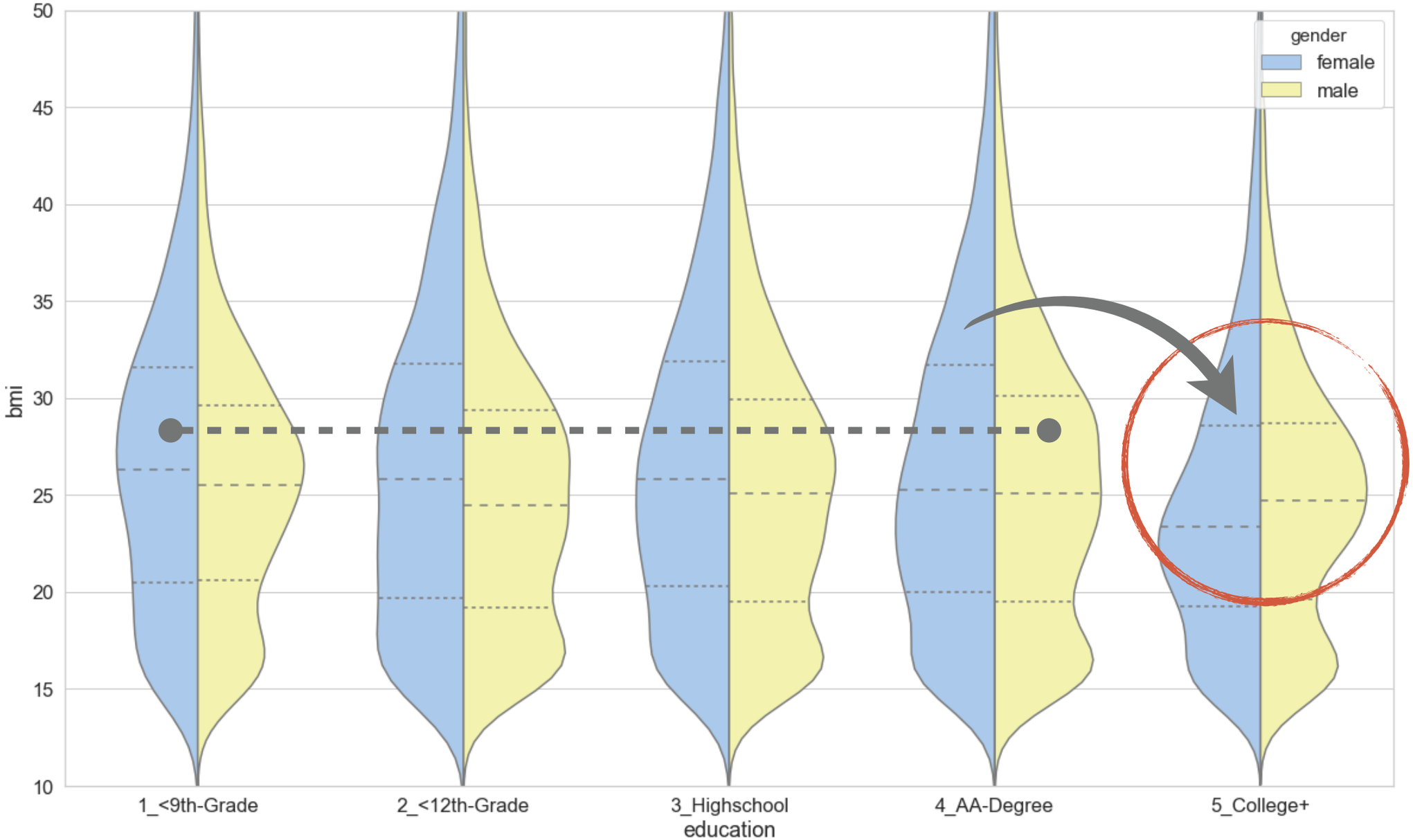
Abb. 4.3.2 Heatmap: Bildung x Einkommen

04 .3 Bildung ➠ BMI

Violinplot

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

Abb. 4.3.2 Violinplot: BMI vs. Bildungsstufe



	education_level	bmi
education_level	1.000000	-0.049918
bmi	-0.049918	1.000000

➤ r = -4,99%

|-0,0499| < 0,3

04 .3 Ethnische Gruppe ➠ BMI

Violinplot

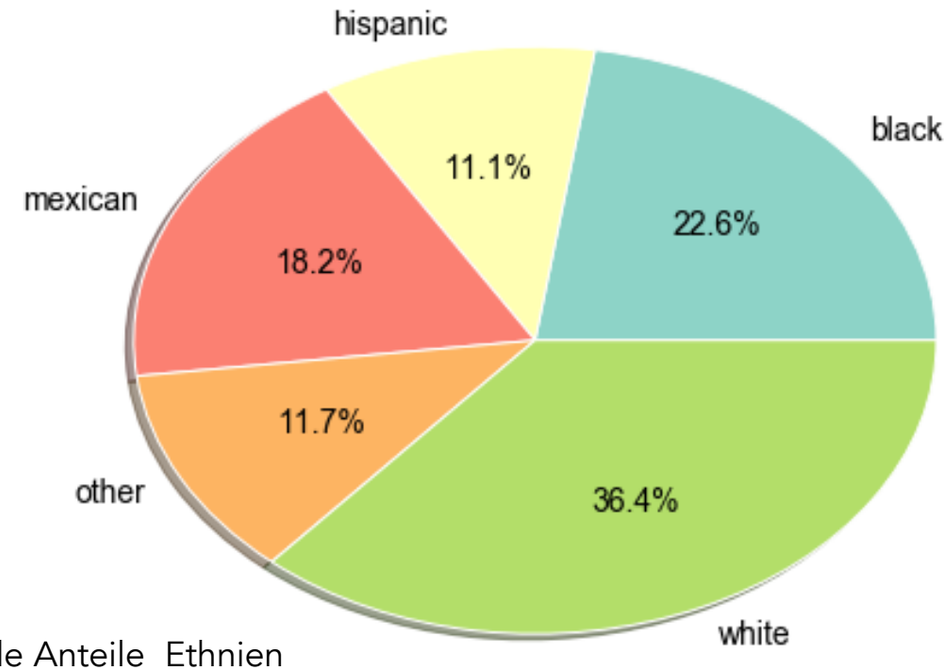
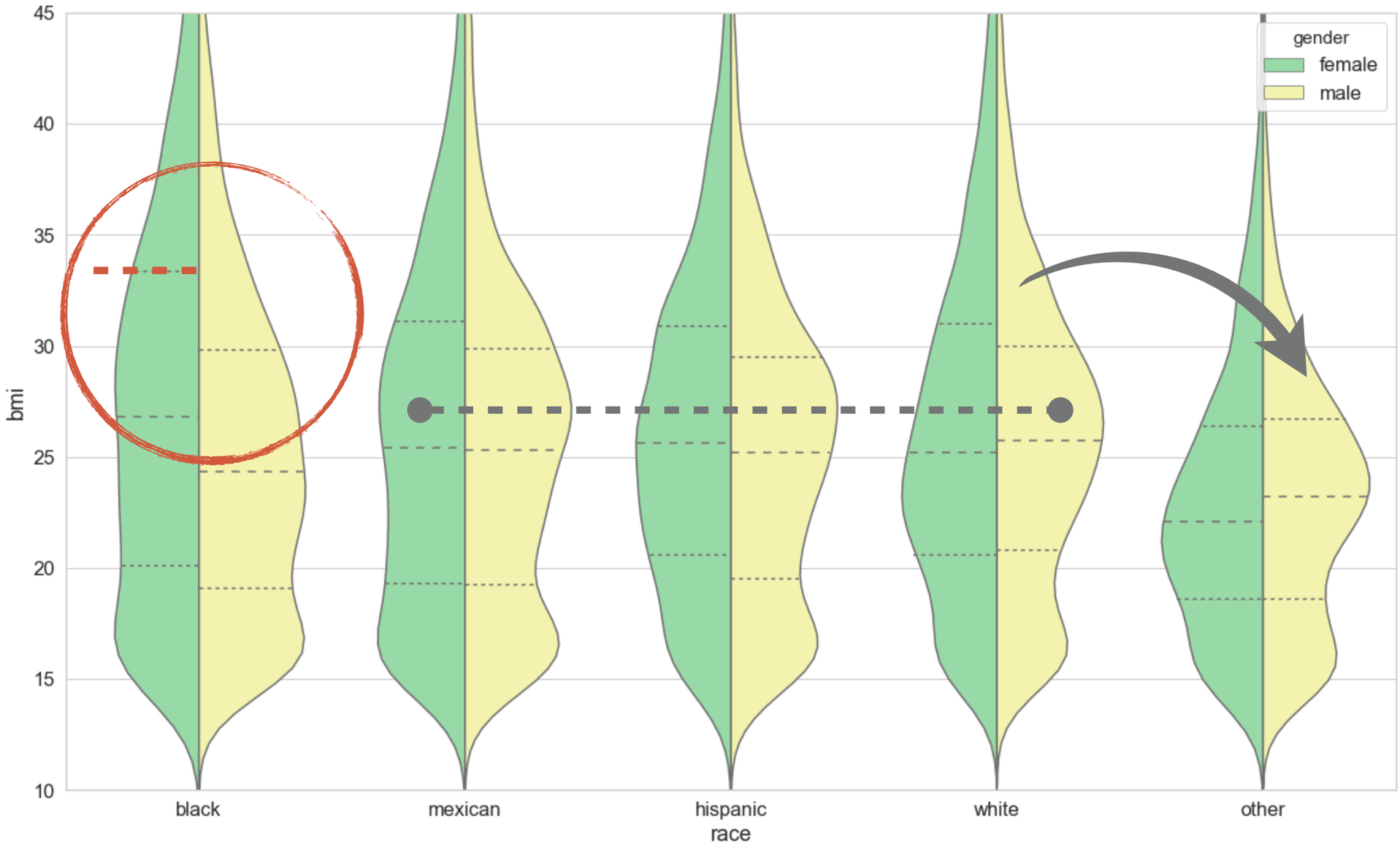


Abb. 4.3.2 Violinplot: Ethnie > BMI

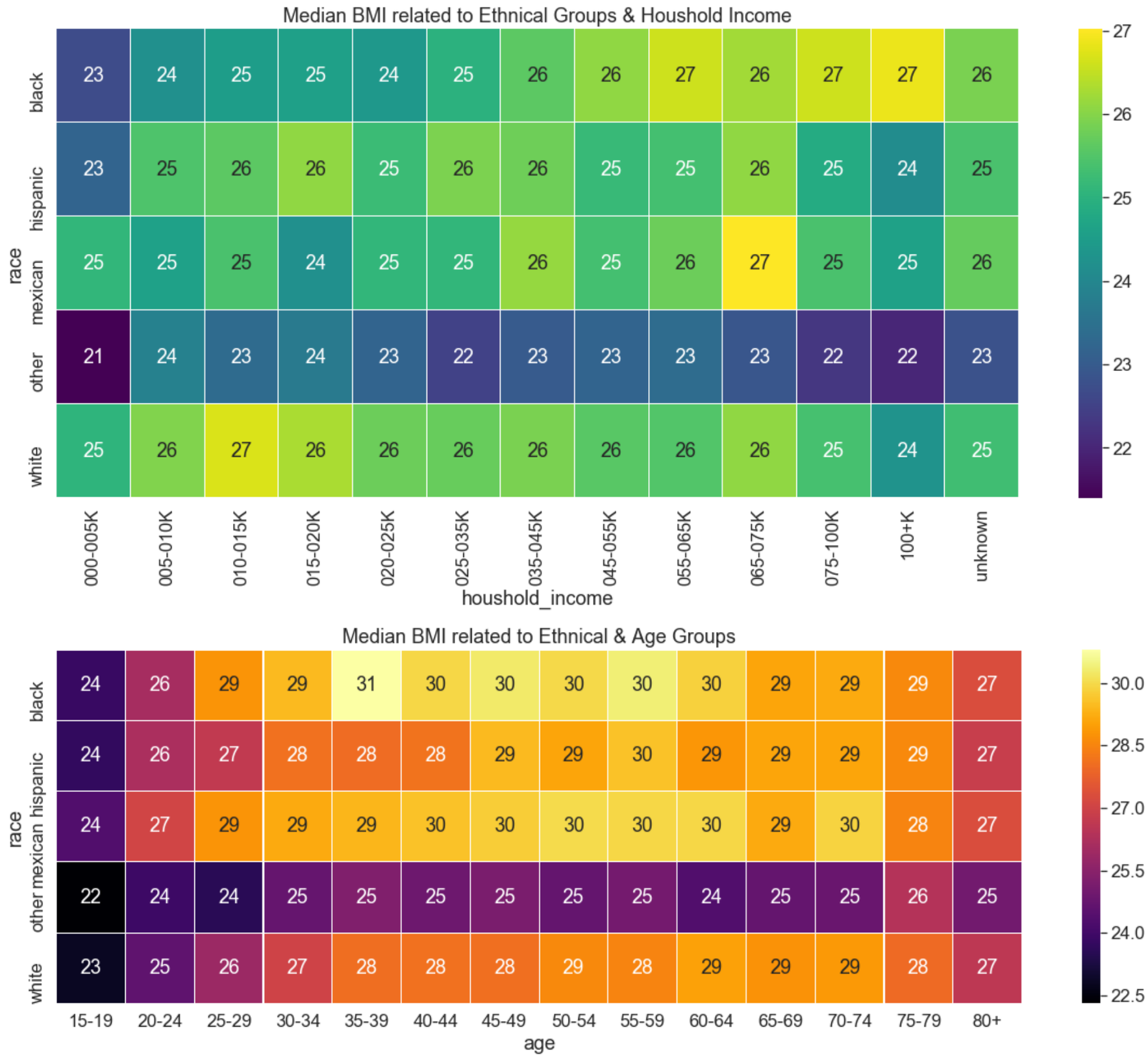
Abb. 4.3.3 prozentuale Anteile Ethnien



04 .3 Ethnische Gruppe ➠ BMI

weitere Differenzierung

Abb. 4.3.4 Heatmap: Medan BMI



nach Einkommen

nach Alter

04 .4 Auswertung
der Daten-Exploration

➤ BMI und Einkommen sind stochastisch nicht unabhängig.

- ❖ Beziehung jedoch wenig aussagekräftig, da C_{korr} dicht bei 0,
- ❖ Abfall des BMI nur in den höchsten Einkommensschichten,
- ❖ Lineare Antikorrelation ist nicht von signifikanter Bedeutung,
- ❖ Differenzierung nach Mitgliederzahl im Haushalt nicht relevant.

➤ BMI und Bildung sind stochastisch nicht unabhängig.

- ❖ Beziehung jedoch nur in den obersten Bildungslevel stark.
- ❖ Bildung und Einkommen stehen in starker Korrelation.

➤ Ethnische Gruppierung hat einen stärkeren Einfluss auf BMI.

- ❖ Starke Steigerung des BMI bei afroamerikanischen Frauen.
- ❖ Dies kann aber nicht auf Einkommen zurück geführt werden.

➠ Weitere Untersuchungen noch verborgenen Zusammenhängen!

NHANES 05

Prognosemodell

Naive Bayes

05 Naive Bayes

Python Gaussian

```
# Machine leraning libraries

from sklearn import datasets

from sklearn.metrics import
classification_report

from sklearn.metrics import
confusion_matrix

from sklearn.model_selection
import train_test_split

from sklearn.naive_bayes import
GaussianNB
```

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

$$P(y \mid x_1, \ldots, x_n) = P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2})$$

Problemstellungen

- Annahme: Merkmale unabhängig ➠ in der Realität fragwürdig!
- Annahme: Merkmalsausprägung normal Verteilt ➠ ebenso fragwürdig!
- Wahl des wahrscheinlichsten Merkmals ➠ hier Tendenz zum Median.

05 Naive Bayes

Modified Data Selection

predict

	id	gender	race	members	income	education	age	bmi	bmi_redux
0	41475	2	1	2	6	4	13	4	2
1	41476	2	1	6	12	5	2	-1	-1
2	41477	1	1	2	5	3	15	2	2
3	41479	1	1	5	8	1	11	1	1
4	41480	1	1	7	7	2	2	-1	-1

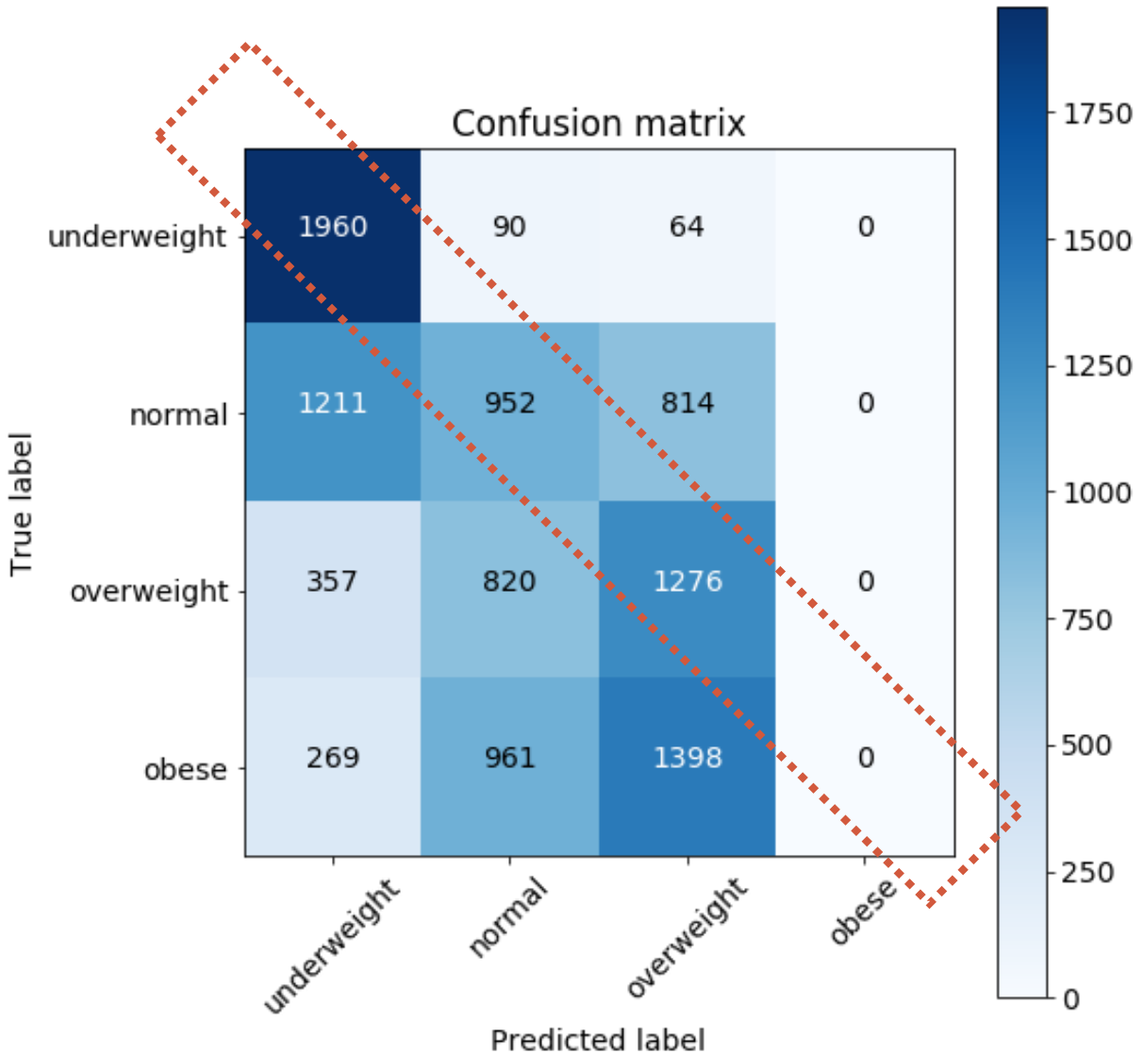
Train | Test Data Split

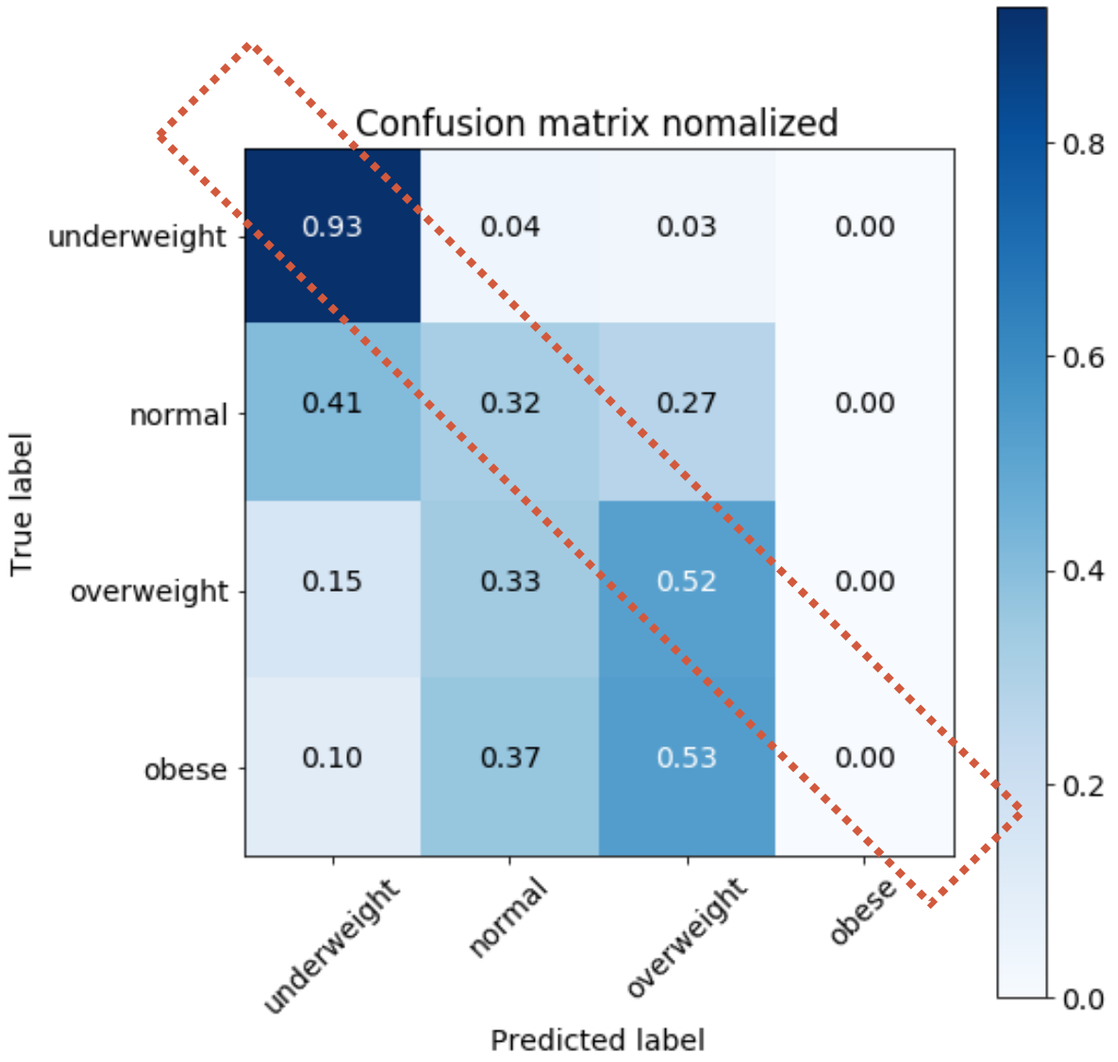
```
# Dividing x, y into train and test data.
x_train, x_test, y_train, y_test = train_test_split(x, y,random_state = 120)
# Training a Naive Bayes classifier.
gnb = GaussianNB().fit(x_train, y_train)
gnb_predictions = gnb.predict(x_test)
# Accuracy on x_test.
accuracy = gnb.score(x_test, y_test)
```

05 Naive Bayes

Auswertung via Konfusionsmatrix

Abb. 5.1.1 Konfusionsmatrix NB





- Diagonale schwach ausgeprägt ➠ starkes „Ausbluten“ zum Median
- keine Erfassung der Kategorie „Obese“ ➠ nach „Overweight“ verschoben
- gute Trefferquote bei „Underweight“ ➠ aber hoher β-Fehler

05 Naive Bayes

Bewertung und Fazit

```
classification_report(y2_test, gnb2_predictions)
```

	precision	recall	f1-score	support
-1	0.52	0.93	0.66	2114
0	0.34	0.32	0.33	2977
1	0.36	0.52	0.42	2453
2	0.00	0.00	0.00	2628
avg / total	0.29	0.41	0.34	10172

- Score: 41,17% ➠ Dem Klassifikator-Modell fehlt es deutlich an Präzision..
- Diagonal immerhin zu erkennen ➠ Schätzwerte nicht zufällig.
- Problem: Leicht gesteigerte Wahrscheinlichkeiten finden bei der Klassifikation keinen Ausdruck.

  Besser wäre ein Modell was über Wahrscheinlichkeiten Aussagen macht.
- Bei den minimalen Korrelationen war das Ergebnis zu erwarten.

FAZIT 06

06 Fazit

„Können sozioökonomische Faktoren genutzt werden, um Fettleibigkeit in den USA zu prognostizieren?“

- Grundsätzlich gibt es Hinweise auf Beziehungen zwischen Übergewicht und sozioökonomischen Faktoren. Korrelation ist nicht Kausalität!
- Kein Merkmal allein sticht heraus. An vielen Stellen finden sich kleinere Beziehungen.
- Diese Korrelationen sind jedoch wenig linear. Die Multivariable Lineare Regression ist wenig präzise.
- Auch ein Zusammenwirken von Faktoren erhöht nur die Wahrscheinlichkeit. Die Klassifikation ist sehr grob.

- Alle Ergebnisse gehen aber eindeutig in eine Richtung. Der Ansatz einer Prognose ist gegeben.

Vielen Dank für die Aufmerksamkeit!

FRAGEN ?