

HTW - Hochschule für Technik und Wirtschaft

Schriftliche Ausarbeitung für Grundlagen Data Science

# Sozioökonomische Indikatoren & Fettleibigkeit in den USA

---

GitHub Repository: [https://github.com/rendermat/datafriends\\_workshop](https://github.com/rendermat/datafriends_workshop)

Fabian Georgi s0563263

loakeim loakeim s0564778

Matthias Titze s0564680

<b>1 Einleitung</b>	<b>2</b>
<b>2 Food Atlas - Datenexploration</b>	<b>3</b>
2.1 Introduction Data Source	3
2.2 Geoplotting	3
2.2.1 State Level Obesity	3
2.2.2 County Level Obesity	4
2.3 Univariate Correlations	6
2.3.1 Obesity	6
2.3.2 Diabetes	8
2.3.3 Fast Food Restaurants	9
2.3.4 Fast Food Expenditure	10
2.3.5 Households without a car and low access to stores	12
2.4 Conclusion	12
<b>3 Food Atlas - Vorhersagemodelle</b>	<b>13</b>
3.1 Einführung	13
3.2 Mathematische Grundlagen der Korrelation	13
3.2 Lineare Regression	16
3.3 Multivariable lineare Regression	17
3.4 Auswertungsmaßzahlen	18
<b>4 NHANES National Health and Nutrition Examination Survey</b>	<b>20</b>
4.1 Einführung	20
4.2 BMI und Einkommen	20
4.2.2 Chi2-Test für BMI und Einkommensgruppe	22
4.2.3 Lineare Regression und p-Wert für BMI und Einkommensgruppe	23
4.2.4 Fazit: BMI und Einkommensgruppe	25
4.3 BMI und Bildung	26
4.3.1 Einkommen und Bildung	26
4.3.2 BMI Verteilung nach Bildungslevel	27
4.4 BMI und Ethnische Gruppierung	27
4.4.1 Plotanalyse zur BMI Verteilung in ethnischen Gruppen	28
4.5 Vorhersagemodell - Supervised Machine Learning	29
4.5.1 Klassifizierung mit Naive Bayes (Gauss Version)	30
4.5.2 Bewertung des Modells	30
4.5.3 Fazit Klassifikator	31
<b>5 Fazit</b>	<b>32</b>
<b>6 Anhang</b>	<b>33</b>
6.1 Quellen	33
6.2 Abbildungen	33
6.3 Tabelleverzeichnis	33

# 1 Einleitung

„Armut macht dick, unbeweglich und abhängig.“ (1: Frankfurter Allgemeine 29.01.2008)

„Zu viel Fast-Food: Fettleibigkeit steigt in Deutschland.“ (2: Deutsche Wirtschaftsnachrichten 15.5.2015).

Die Fettleibigkeit hat in vielen Ländern der Welt schon epidemische Ausmaße angenommen. Je nach Studie unterscheiden sich die Zahlen zwischen jeder dritten und jeder vierten Adipos - Tendenz steigend. Die gesundheitlichen Risiken von extremem Übergewicht sind keinesfalls zu unterschätzen. Viele Betroffene leiden unter erhöhtem Krebs- und Sterberisiko, Hypertonie oder Schlafapnoe. Die Fettleibigkeit ist längst zur Volkskrankheit geworden und hat durch die damit verbundene Belastung der Krankenkassen auch deutliche, negative Auswirkungen auf die Sozialsysteme von Volkswirtschaften. Adipositas ist insbesondere in den Industriestaaten zu einer zentralen, medizinischen Herausforderung aufgestiegen. Die kritische Bewertung von Übergewicht ist in diesem Kontext also nicht mit einer Diskussion um Schönheitsideale zu verwechseln.

Doch was sind jenseits von der banalen Vereinfachung, zu viel Fast-Food, eigentlich die Gründe für das Grassieren der Fettleibigkeit? Trifft das Klischee, dass die armen und folglich zum Konsum von billigerem Junk-Food gezwungenen Bevölkerungsanteile stärker unter Übergewicht leiden, überhaupt zu?

Das Land der unbegrenzten Möglichkeiten scheint nicht nur Millionäre, sondern auch überdurchschnittlich viele fettleibige Menschen hervorzubringen. Die Vereinigten Staaten von Amerika führen die Liste der Adipositas an, sodass dieses Land optimalen Nährboden zur Untersuchung der Fettleibigkeit bietet. Die staatliche Verwaltung hat das Problem zudem schon seit einigen Jahren erkannt und Programme zur Eindämmung von Übergewicht initiiert. In diesem Rahmen wurden auch eine Reihe von Erhebungen durchgeführt, die in umfassenden Datensätzen gebündelt und veröffentlicht wurden. Das „United States Department of Agriculture“ stellt den „Food Environment Atlas zur Verfügung“, während die „Centers for Disease Control and Prevention“ die „National Health and Nutrition Examination Survey“ anbieten.

Das Ziel dieser Ausarbeitung ist, mögliche sozioökonomische Faktoren zu finden, welche krankhaftes Übergewicht in den USA eventuell begünstigen. Eine Reihe von Indikatoren werden auf mögliche Korrelationen zur krankhaften Fettleibigkeit geprüft. Aus der Untersuchung sind zunächst keine Kausalitäten abzuleiten; sie kann jedoch als Grundlage für weiterführende Studien dienen, die nach den Gründen für Übergewicht suchen.

---

## Forschungsfrage

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

# 2 Food Atlas - Datenexploration

## 2.1 Introduction Data Source

The dataset used for this project had to meet several criteria, including: a reliable source, an adequate size, fine data granularity, recent and diverse data. Based on these criteria, the „Food Access Research Atlas“ dataset was chosen as the main source of information. This initial dataset is provided by the United States Department of Agriculture, which should be a reliable source of information, since it is a government entity. All the indicators used in this study are directly derived from their corresponding tables. These include medical terms such as diabetes and obesity. The collection of these data stems from public entities. The data is provided as an Excel file with multiple sheets. The values of the food atlas are collected for each US county which are identified by their FIPS-code. The Version used was 5/18/2017, being the most recent version at the time the project was done. The data span between 2001 and 2016, with most data ranging from 2007 to 2015, making them recent enough for the purpose of this research. The data is also mostly complete, including all 3142 counties, which belong to the 50 States. The District of Columbia, which is not located in any state but is instead a territory of its own, is also included. In the data it is included as an additional county as well as a state. U.S. territories and their county equivalents are excluded. Required Data was extracted from the original raw data source, it was then cleaned and in some cases refined.

## 2.2 Geoplotting

Due to the nature of the dataset, an atlas, a geospatial analysis was the first statistical analysis that was carried out. The goal here is to investigate, whether different parts of the United States have varying obesity rates. If that is the case, then the next step would be to identify the underlying reason for these differences. For this investigation, the U.S. map was plotted, divided into its states, with each state showing the median value of obesity. The lighter, greener colour indicates a lower rate, whereas the darker, purple areas are those with a higher value. A higher median hints at a higher percentage of people with obesity.

### 2.2.1 State Level Obesity

Despite coarse granularity, there are some distinct differences between states and regions. The regions and divisions are based on the United States Census Bureau, which is widely used for data collection and analysis, and is the most commonly used classification system. The map shows that the Pacific, Mountain and New England divisions, in other words, mostly the West and Northeast regions, have the lowest obesity rates. On the other hand South Atlantic, West North Central and West South Central, East North Central as well as East South Central divisions; that is, mostly the Midwest and southern regions, have the highest obesity rates.

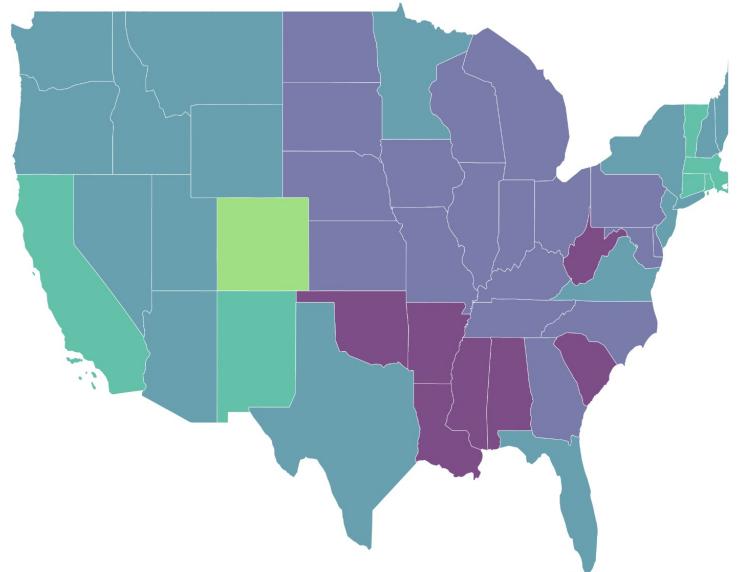
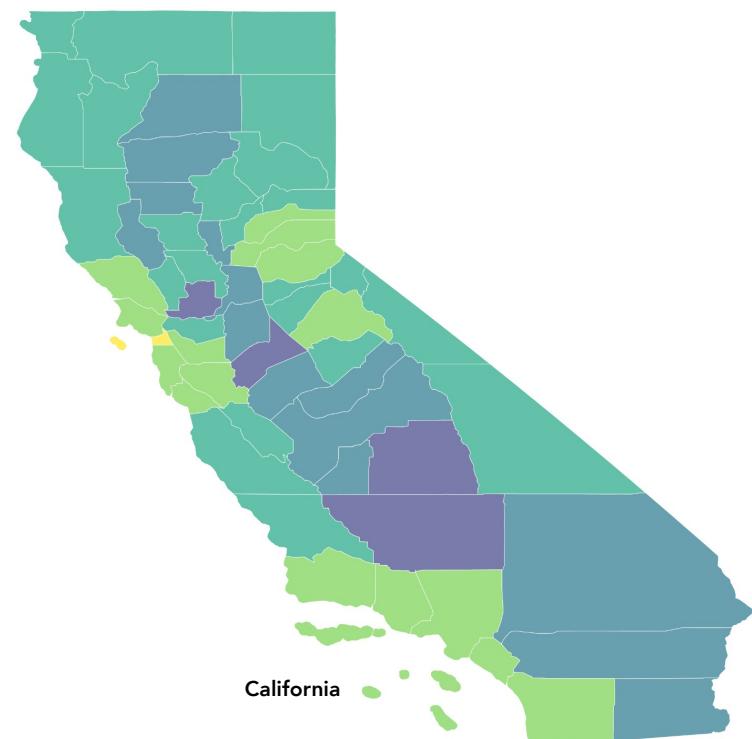
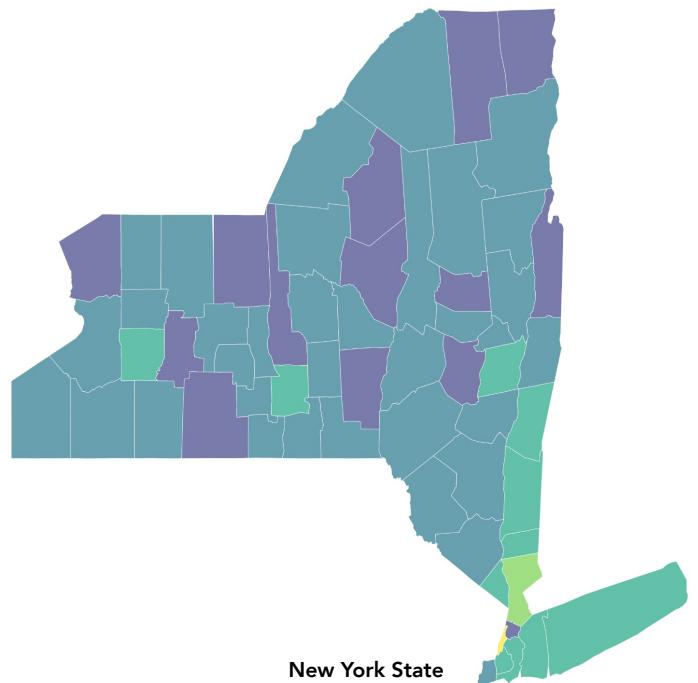
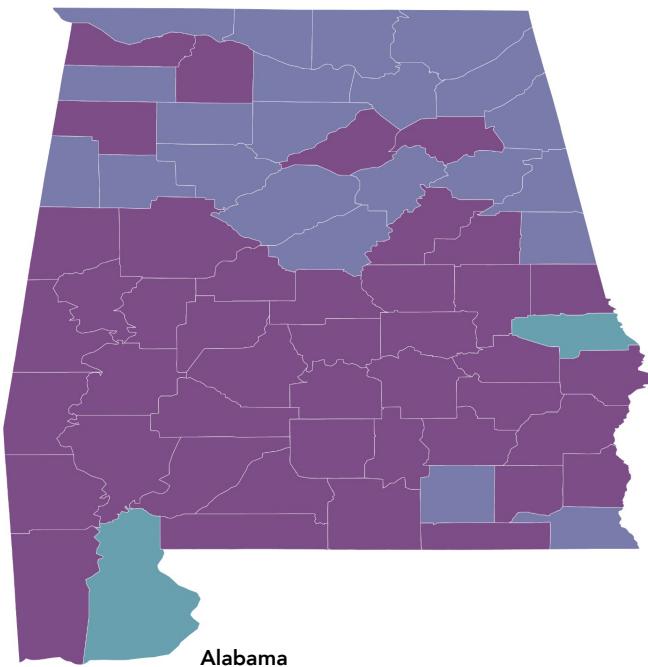


Abb. 2.2.1 Geoplot: US-States >> Obesity Adults US 2013

The lowest obesity rate is found in the state of Colorado, with a median value of 20,5%. In contrast to that, the state of Mississippi has the highest obesity levels, with a median value of 36.85%, that's more than one third of the population. The most important observation here is the relationship between income and obesity: Mississippi, Arkansas, West Virginia and Alabama are among the states with the highest obesity rates. At the same time these are the 4 States with the lowest median family income in 2013. Alaska and the state of Hawaii with 30.6% and 22.3% obesity rates respectively, are not being shown on the map. More refined insights can be gained by zooming in on the county level.

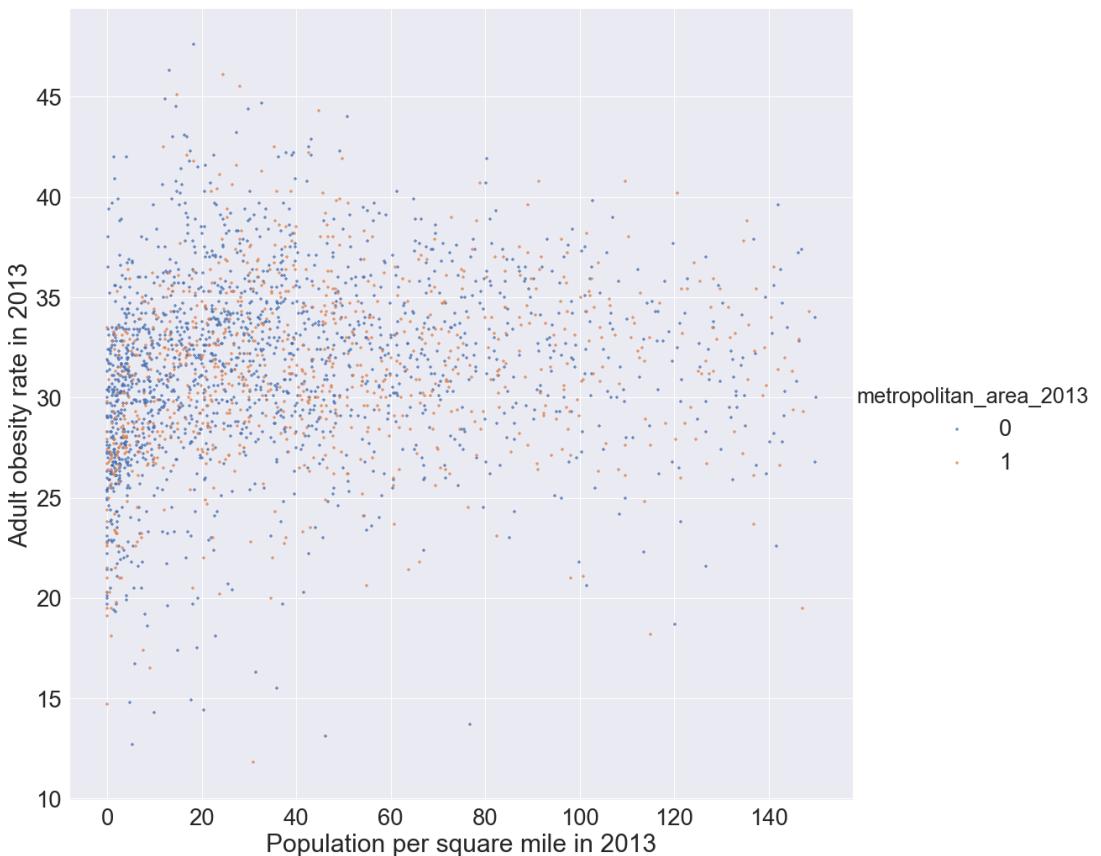
## 2.2.2 County Level Obesity

**Abb. 2.2.2** Geoplot: US Counties >> Obesity Rates US 2013



As an example for the county level analysis, the focus is set on the states of California, Alabama and New York. Looking at Alabama, the first clear observation to be made, is the fact that apart from two counties, Baldwin and Lee, all other counties have an obesity rate higher than 30 percent. Greene has a value of 46.3 percent, which is among the counties with the highest obesity rates in the whole U.S. In the case of New York State, New York City has the lowest rate with 14.7 percent, followed by Westchester with 20.3 percent. The counties with the lowest obesity rates are also grouped together near New York City which would suggest the geographical location has a role to play. Looking at the state of California, we can observe that coastal areas as well as the islands have very low obesity rates. These areas have a rate, which is lower than 26 percent. Additionally the span of obesity is also very low. It only ranges between 19 and 33 percent , when San Francisco is excluded. San Francisco has an obesity rate of 16.1 percent. When all three states are taken into consideration, it appears that specific metropolitan areas have lower obesity rates as opposed to non-metropolitan ones. San Francisco, New York City as well as San Diego to name a few, are examples of metropolitan areas, which have relatively low obesity rates.

Therefore the next step would be to plot the county obesity versus the population density, while differentiating between metropolitan and non metropolitan areas:



**Abb. 2.2.3** Scatterplot: Population Density & Metropolitan Areas >> Obesity

The graph shows the obesity rates from 2013 against population per square mile in 2013. Metropolitan areas are shown as orange dots whereas non-metropolitan ones are shown as blue markers. There is no noticeable difference in obesity rates between metropolitan and non-metropolitan areas. It also becomes clear that metro-areas are not identified by population density. At the same time, population per square mile has no strong dependency on the obesity rate. Therefore the metro-effect on obesity seems to be limited to a handful of mega-cities.

**Tab. 2.2.1** Description: Metropolitan vs. Non-Metropolitan Areas

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

However based on the descriptive table (left), there is a difference in obesity albeit a small one. There are almost twice as many non-metropolitan counties compared to metropolitan ones. Additionally the average obesity rate of metropolitan areas is 0.35 percent lower than the non-metropolitan one. The lowest metropolitan obesity rate is also 0.9 percent lower, the first quartile 0.3 lower, the median 0.3 lower, the third quartile 0.2 lower and finally, the highest metropolitan obesity rate is 1.5 percent lower compared to the non-metropolitan ones. The differences are only slight, which is mainly why they did not show on the graph.

Further analysis of obesity rates was carried out in isolation. Since obesity is the main subject of the paper, its data from different years was looked into.

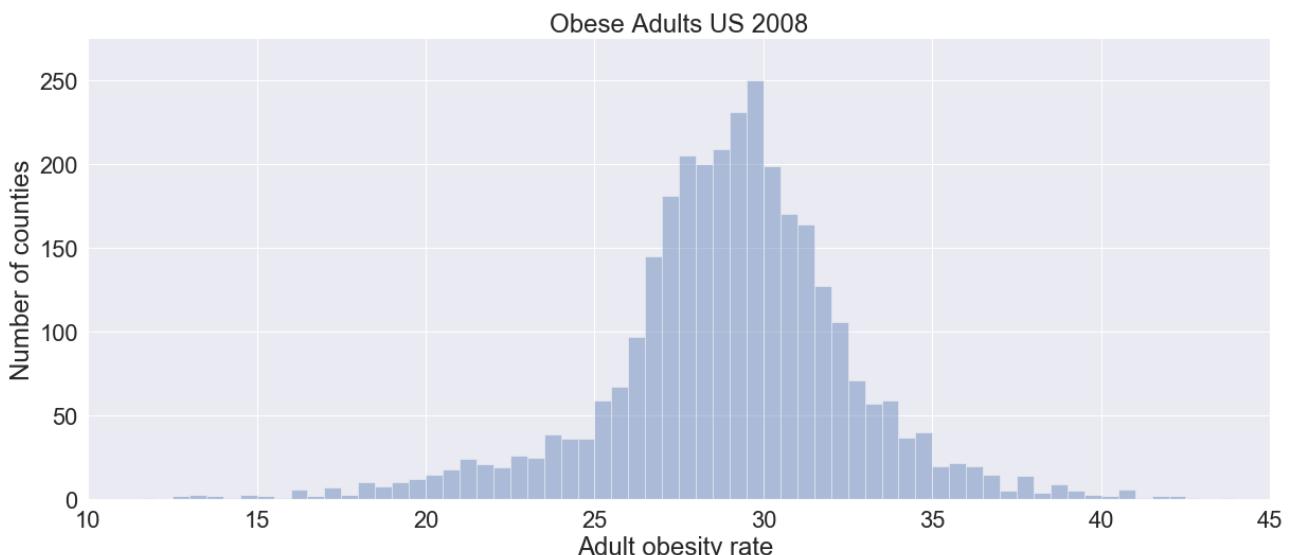
## 2.3 Univariate Correlations

### 2.3.1 Obesity

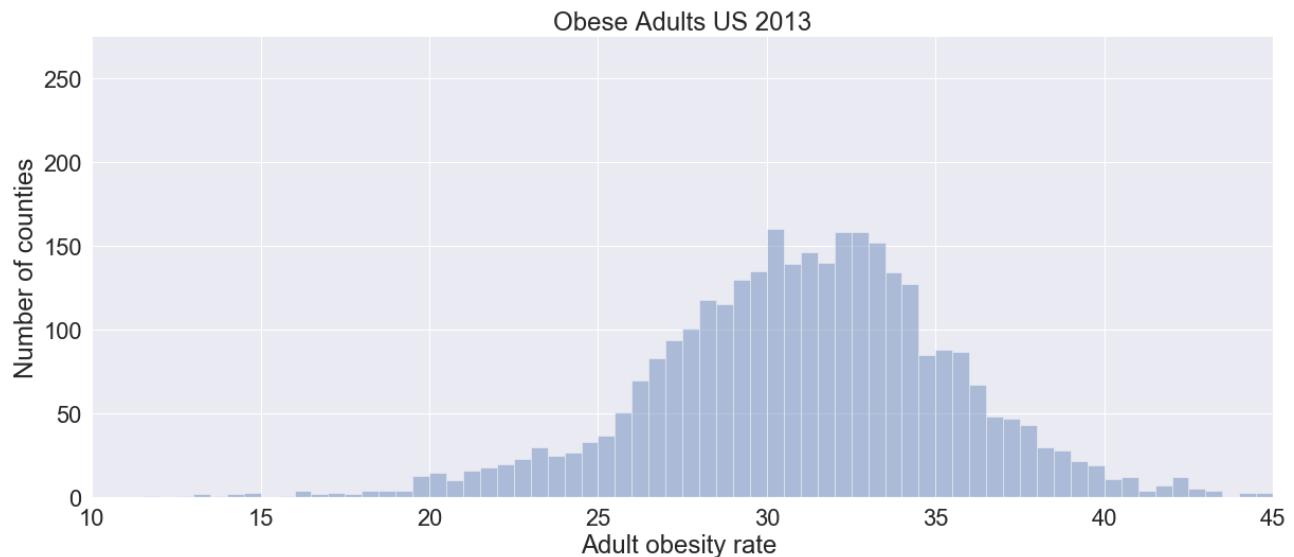
**Definition:** Estimate of age-adjusted percentage of persons age 20 and older who are obese, where obesity is Body Mass Index (BMI) greater than or equal to 30 kilograms per meters squared.

**Source:** Estimates are from Centers for Disease Control and Prevention (CDC). CDC used data from the Behavioural Risk Factor Surveillance System (BRFSS) for 2008, 2009, and 2010 and from the U.S. Census Bureau.

The purpose of this analysis is to explore the differences in Obesity rates from different years and how strong these differences are. For clarity purposes the counties are being categorised into buckets. An accuracy of 0.5 percent was chosen as the most adequate one.

**Abb. 2.3.1** Histogramm: US Adult Obesity Rate 2008

The histogram shows the different obesity rates and the number of counties which fall into these obesity buckets. These data are from 2008 and they show a typical gaussian (normal) distribution, where the bucket with most counties, 250 to be exact, has an obesity rate of 30 percent.

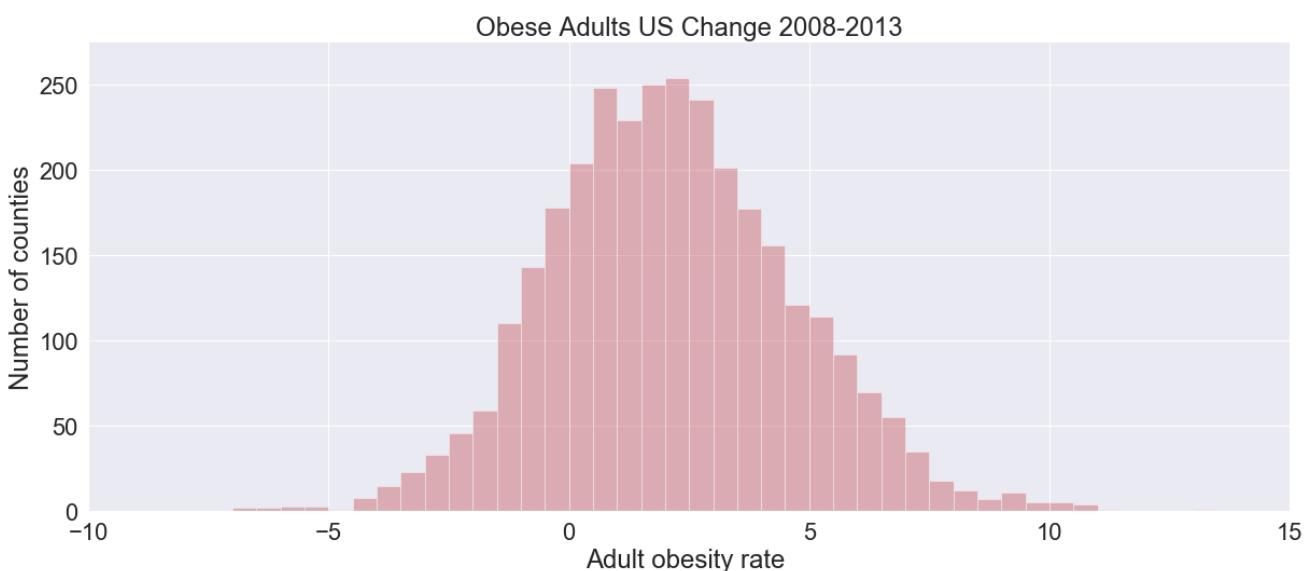


**Abb. 2.3.2** Histogramm: US Adult Obesity Rate 2013

Data recorded in 2013 on the other hand, show that the distribution is more widely spread, in other words, the standard deviation has grown: from 3.711 in 2008 to 4.523 in 2013. The biggest bucket here is about 160 counties with an obesity rate of 30.5 percent. For more accurate information, descriptive statistics were also generated including averages of obesity rates from 2008 and 2013 and the percentage change from 2008 to 2013:

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

**Tab. 2.3.1** Description: US adult obesity rate



**Abb. 2.3.3** Histogramm: US Adult Obesity Rate 2008-2013

The most noteworthy value here is the low standard deviation value of the percentage change in obesity (last row). To better understand what this value is implying a histogram of the percentage change of obesity from 2008 to 2013 was plotted. The first observation is that, obesity rates throughout these years have slightly increased for the most part. The second observation is that, the percentage increase between counties is similar, with the majority of these ranging between 0% and 4%, which is what the low standard deviation value showed previously. This means the data have stability, which would justify a linear interpolation between the years as a means of higher accuracy, in case one is required at a later stage.

For further data exploration, the variable list spreadsheet was thoroughly examined for possible factors that might (as a single variable) correlate with obesity.

### 2.3.2 Diabetes

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

**Source:** Estimates are from Centers for Disease Control and Prevention (CDC). CDC used data from the Behavioural Risk Factor Surveillance System (BRFSS) for 2008, 2009, and 2010 and from the U.S. Census Bureau. The methodology is described on the CDC's County Estimates page.

As a baseline obesity rate was put up against diabetes rate, which is more of a ramification/result rather than a cause for obesity. The aim was to examine how an assumed strong correlation between variables which are directly related to each other would manifest in the data. This can be used as a standard to match other correlations against.

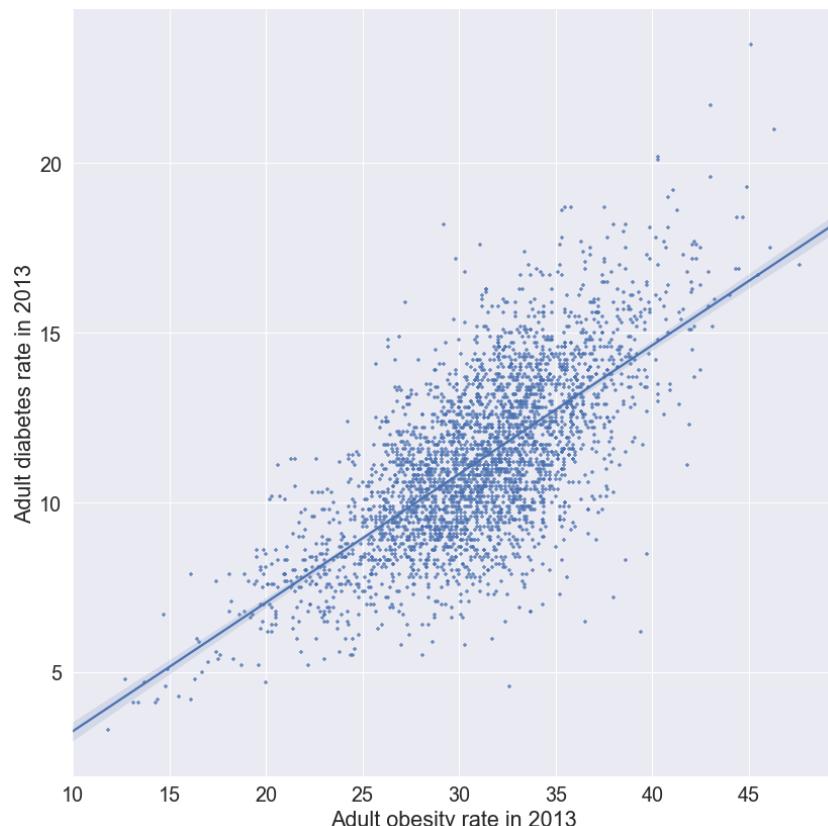
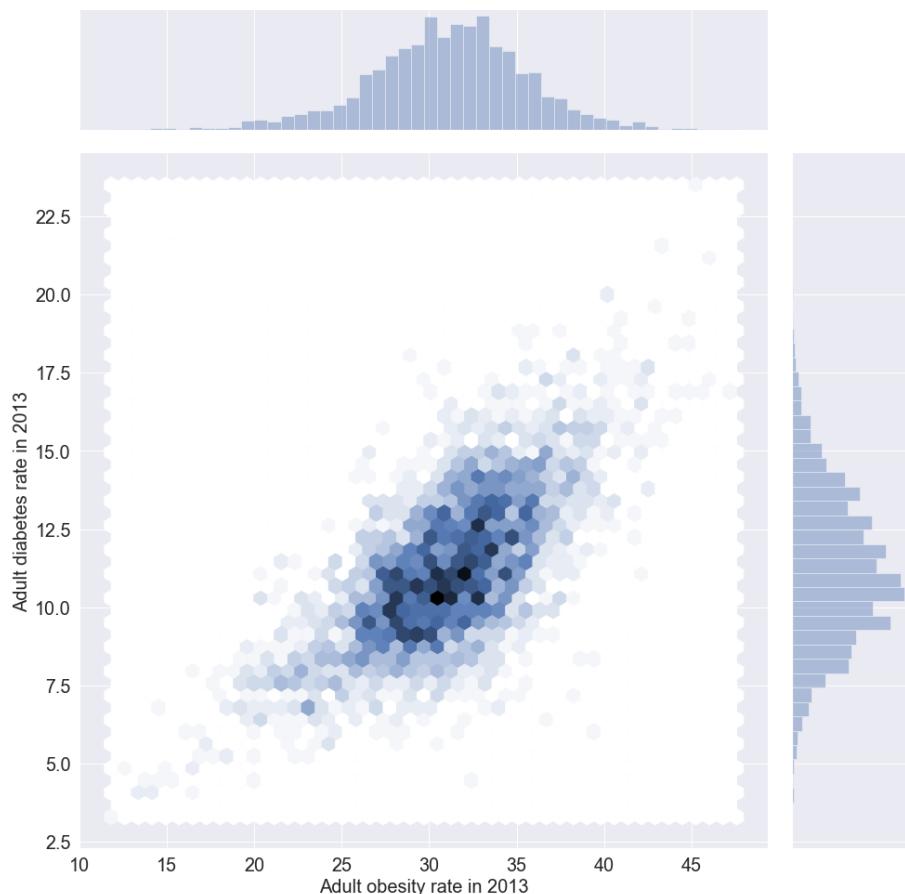


Abb. 2.3.4 Scatterplot: Obesity >> US Adults Diabetes 2013

The scatter graph shows a positive relationship between obesity and diabetes, but the graph is not very clear due to the fact that too many points overlap with each other. To solve this, a hex graph better shows where the points lie, by using darker colouring on the graph, when points are densely packed together.

The hex graph reveals that the relationship is not as strong as one would have expected. There are multiple reasons as to why this is. Firstly, obesity is known as a main cause of type II diabetes, but not type I. Secondly there are obese people, who do not suffer from diabetes and vice versa, there are diabetics who are not obese. Additionally, there is the fact that the data is county based and not on an individual level, which means that the linkage between obesity and diabetes between individuals gets lost.



**Abb. 2.3.5** Hexplot: Obesity >> US Adults Diabetes 2013

### 2.3.3 Fast Food Restaurants

**Definition:** The number of limited-service restaurants in the county. Limited-service restaurants (defined by North American Industry Classification System (NAICS) code 722211) include establishments primarily engaged in providing food services (except snack and nonalcoholic beverage bars) 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. Some establishments in this industry may provide these food services in combination with alcoholic beverage sales.

**Source:** Restaurant data are from the U.S. Census Bureau, County Business Patterns.

Studies have shown that rises in obesity rates among the world population could be attributed to an increase in calorie intake coupled with lack of adequate physical activity. Fast food is regarded as a low quality, calorically dense food. For this reason, the number of fast food restaurants, and its effect on obesity was assessed. As a first step the number of empty values, in other words the number of counties with no

record of the number of fast food restaurants, was checked: The selected data, which are the records from 2014 has no empty entries. Relevant data from the atlas was recorded in 2009 and in 2014. This gives rise to the problem that obesity and fast food data are recorded in different years. However, due to the low standard deviation value of the percentage change of obesity, as well as the assumption that realistically the number of people who are cured from obesity within one year's time, is extremely low, the two can still be plotted against each other with relatively high accuracy.

Despite some overlapping between the points, the graph gives a clear message: There is no linear correlation between obesity and the number of fast food restaurants. That is to be half expected, since the

number of restaurants says basically nothing about the customers who go there. It remains unclear how often they dine at such places or how much food they consume. This means that the expenditure at fast food restaurants would make a better candidate to evaluate whether obesity and fast food consumption have a relationship. Before this test was run, fast food expenditure and fast food restaurant count were plotted against each other. It is to be expected, that the two are related since from a rational perspective, more money is spent somewhere, where there are a lot of fast food restaurants as oppose to a place where there are only a few. The purpose of this comparison is once again experimental, where the goal is to examine the relationship between two variables that are expected to be tightly connected.

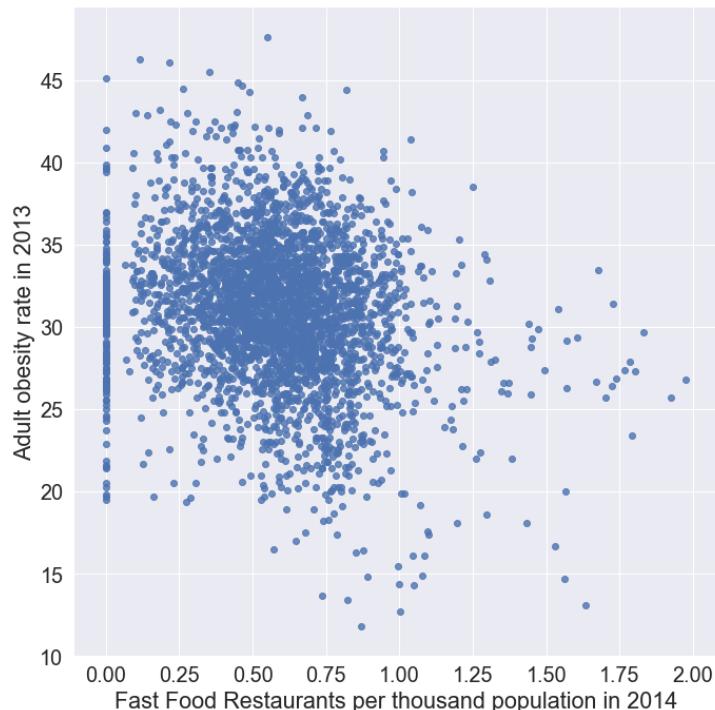


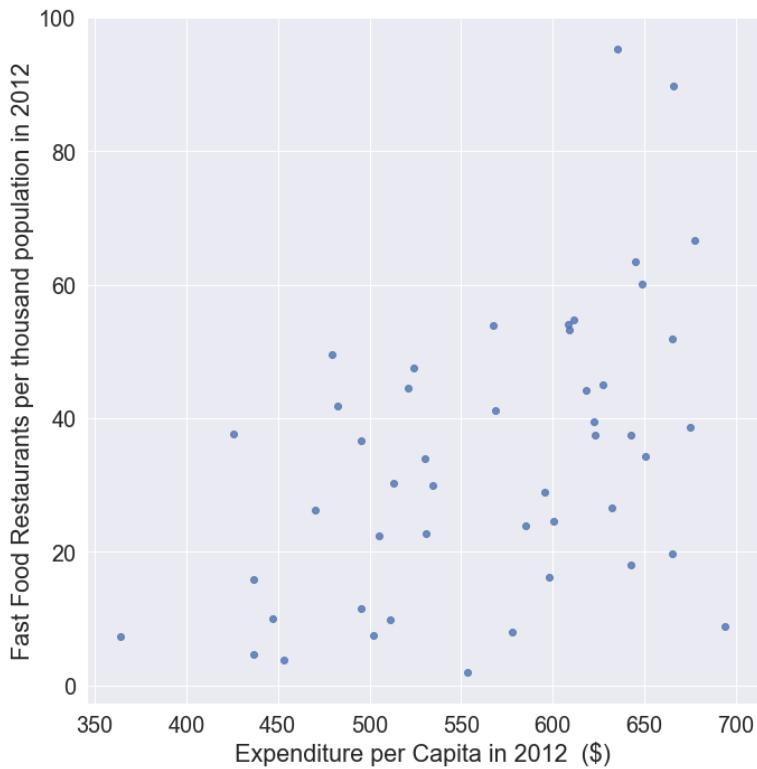
Abb. 2.3.6 Scatterplot: US Fast Food Restaurants / 1000 Pop >> Obesity

### 2.3.4 Fast Food Expenditure

**Definitions:** Average expenditures (in current dollars) on food purchased at limited-service restaurants (defined by North American Industry Classification System (NAICS) code 7222) by county residents. Limited-service restaurants include establishments primarily engaged in providing food services where patrons generally order or select items and pay before eating (see 2.3.3 for reference).

**Source:** Economic Census, Accommodation and Food Services: Geographic Area Series, accessed at U.S. Census Bureau, American Factfinder. Population data are drawn from the U.S. Census Bureau, Population Estimates.

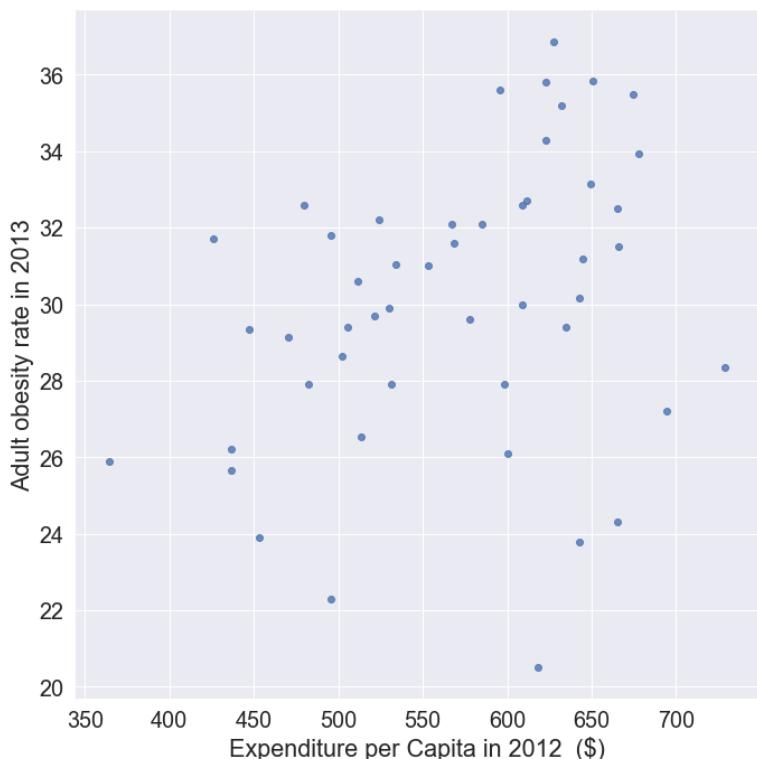
Plotting Fast food expenditure per capita against fast food restaurants per 1000 population gave rise to two challenges. One was the fact that Expenditure was recorded on a state level as oppose to fast food restaurant count which was on a county level. To bring the scope of the two variables on the same level, the sum of fast food restaurants of all counties in a given state was calculated. Secondly, the expenditure was recorded in 2007 and 2012 but restaurant numbers were recorded in 2009 and 2014. To rectify this problem, the number of fast food restaurants was linearly interpolated to the year 2012.

**Abb. 2.3.7** Scatterplot: US Fast Food Expenditure >> Restaurant Number

Unexpectedly, there is no correlation to be seen from the graph. One of the main reasons for this could be the scope of the data. This is seen again, when we calculate obesity on a state level, in this case, the median value of all counties that belong to a given state.

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

When we combine the obesity data of counties in their respective states, we get values which are very close to the average. This is verified by the small standard deviation value in the description frame. This means working with data on a state level does not make much sense, because we lose too much information resolution along the way.



For demonstration purposes, obesity against expenditure is graphed, to highlight the points, which were previously mentioned. The fact that obesity was recorded in 2013 and expenditure in 2012 has no significant impact on the accuracy of the graph since realistically, it is highly unlikely that obesity rates will change a lot in a year's time. This was also statistically verified above.

**Abb. 2.3.8** Scatterplot: US Fast Food Expenditure >> Obesity

### 2.3.5 Households without a car and low access to stores

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

**Source:** Data are from the 2012 report, Access to Affordable and Nutritious Food: Updated Estimates of Distances to Supermarkets Using 2015 Data. In this report, a directory of supermarkets, supercenters and large grocery stores within the United States, including Alaska and Hawaii, was derived from merging the 2015 STARS directory of stores authorised to accept SNAP benefits and the 2015 Trade Dimensions TDlinx directory of stores. Stores met the definition of a supermarket, supercenter, or large grocery store if they reported at least \$2 million in annual sales and contained all the major food departments found in a traditional supermarket, including fresh meat and poultry, dairy, dry and packaged foods, and frozen foods. The combined list of supermarkets and large grocery stores was converted into a GIS-usable format by geocoding the street address into store-point locations. Population data are reported at the block level from the 2015 Census of Population and Housing. These population data were aerially allocated down to ½-kilometer-square grids across the United States. For each ½-kilometer-square grid cell, the distance was calculated from its geographic center to the center of the grid cell with the nearest supermarket. Rural or urban status is designated by the Census Bureau's Urban Area definition.

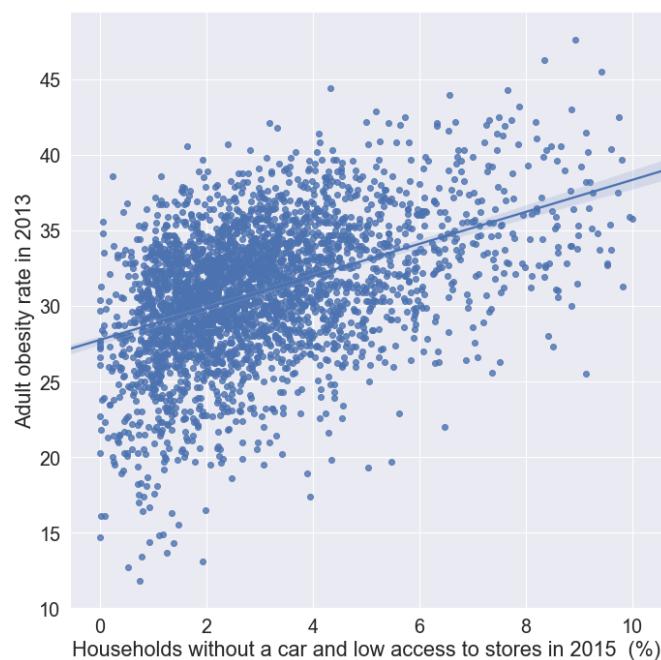


Abb. 2.3.9 Scatterplot: Low store access (no car) > Obesity

This time the graph shows a very weak connection between the two variables, but a connection nevertheless. The variables might be two years apart from each other but obesity has already been tested for data stability. It would have been interesting to see whether spending in fast food restaurants has a connection to low access in stores, in other words, whether people with low access to stores eat at fast food restaurants more frequently, but due to the fact that expenditure is on a state level, and we already know data on a state level are too generalised, this would unfortunately give no relevant information.

## 2.4 Conclusion

The state level geo data show a definite distinction between states. Those with the lowest income are also those who suffer from the highest obesity rates. Oklahoma, Louisiana, Arkansas, Mississippi, Alabama, South Carolina and West Virginia are the states with the highest obesity rates. The county level geo data show that metropolitan areas have lower obesity rates than non-metropolitan ones. The population density graph on the other hand shows that there are no differences between metropolitan and non-metropolitan areas. Only cities like San Francisco and New York City have significantly lower rates. 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, which is a result of obesity, and obesity, are not as strong as expected. Plotting single possible factors against obesity never yields the expected results. This is probably due to the fact that obesity is a chronic disease and hence is influenced by many factors.

# 3 Food Atlas - Vorhersagemodelle

## 3.1 Einführung

Bisher wurden einzelne sozioökonomische Merkmale aus dem Food Atlas ausgewählt und diese Faktoren jeweils einzeln in Verbindung zur Fettleibigkeit in den USA gesetzt. Die paarweisen Untersuchungen mittels Streudiagramm sind für den Einstieg unabdingbar, jedoch visualisieren sie nur Relationen und bieten keine Möglichkeit zur Prognose von Übergewicht. Die die Betrachtung war zudem bisher auf wenige Faktoren beschränkt.

Wir versuchen nun in einer breiter angelegten Studie den Großteil aller sinnvollen Variablen des Food-Atlas zu erfassen und kreuzweise die Beziehungen aller Variablenpaare zu prüfen. Dies erfolgt über eine Reihe von Hitzekarten, die für jedes Variablenpaar den Korrelationskoeffizienten darstellen. So kann ein Überblick über alle eventuellen Korrelationen gewonnen werden.

## 3.2 Mathematische Grundlagen der Korrelation

Einen Zusammenhang zwischen zwei Größen lässt sich mathematisch mit Hilfe der Korrelation nachweisen. Mit dem Pearsonschen Maßkorrelationskoeffizienten kann man den Zusammenhang zwischen zwei Merkmalen quantifizieren. Er ist wie folgt definiert:

---


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

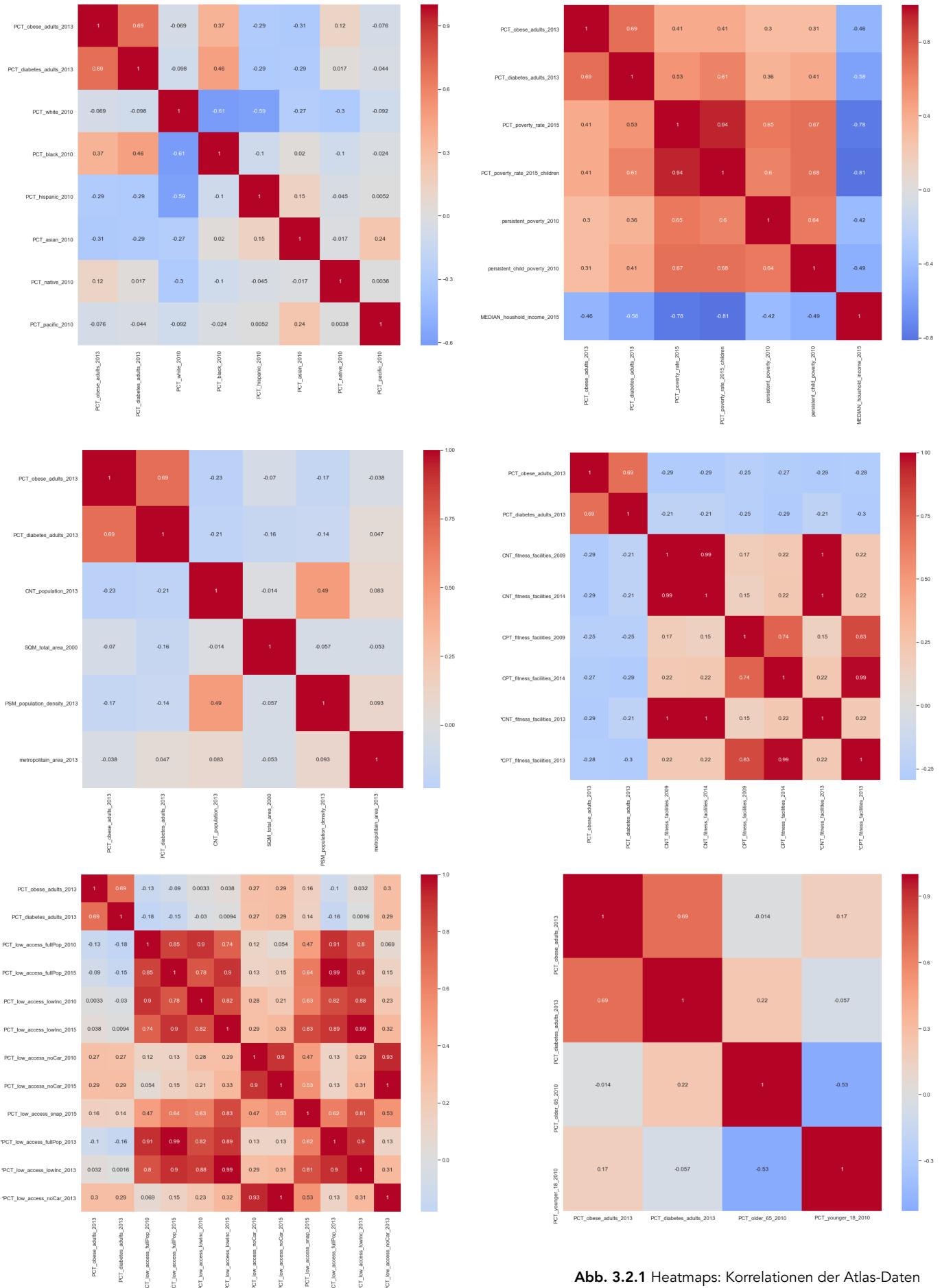

---

Zur Einordnung dieses zwischen -1 und 1 standardisierten Wertes helfen uns folgende Richtlinien:

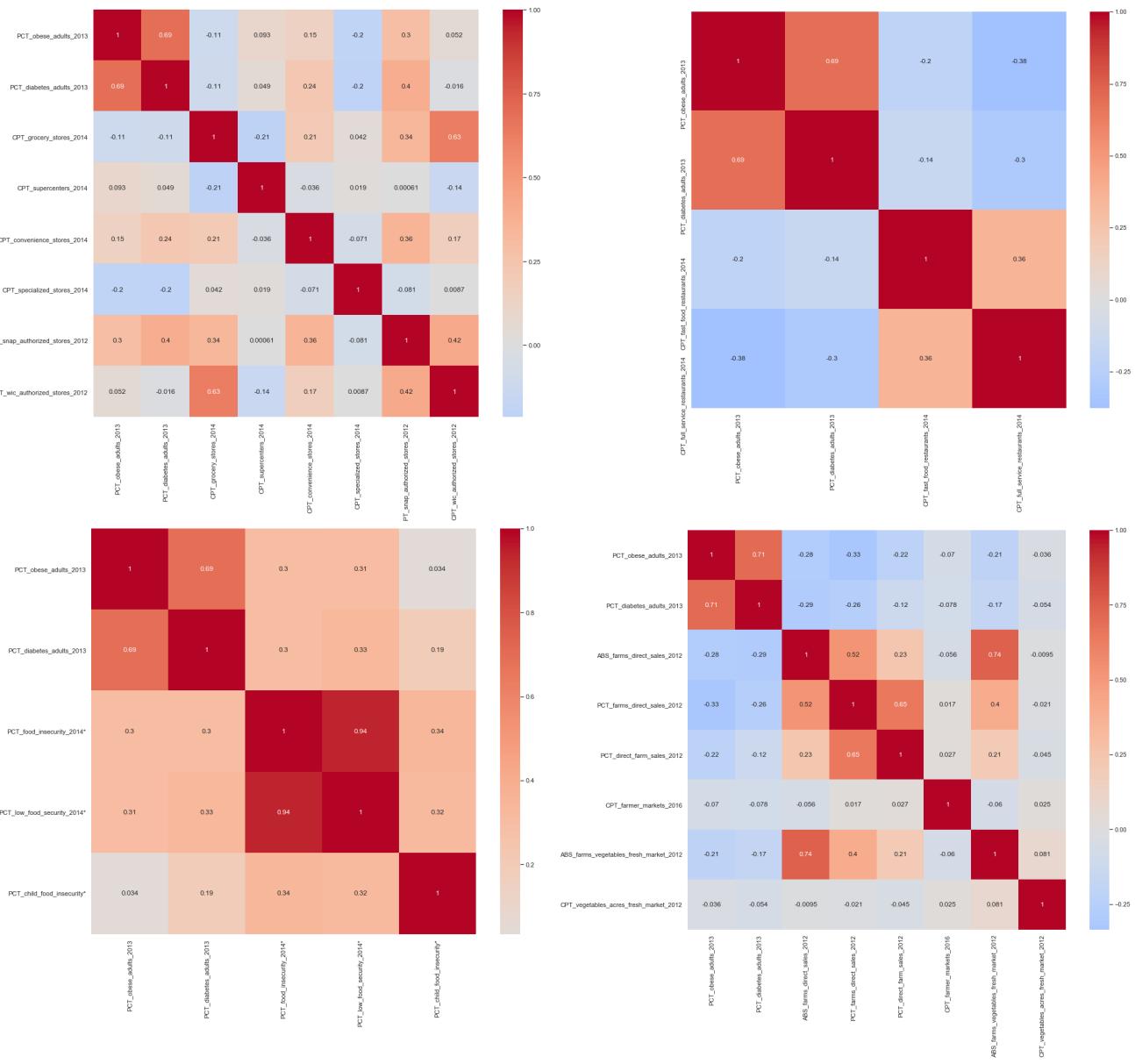
- 
- |  |
|--|
| $\rho \approx 0$ : zu vernachlässigende lineare Abhängigkeit zwischen X und Y,<br>$0.3 <  \rho  < 0.7$ : schwacher linearer Zusammenhang zwischen X und Y,<br>$ \rho  > 0.7$ : starker linearer Zusammenhang zwischen X und Y. |
|--|
- 

Dieser Korrelationskoeffizient misst nur lineare, keine logarithmischen oder quadratischen Abhängigkeiten und gibt auch keine mengenmäßige Änderungsqualität. Eine hohe Korrelation sagt lediglich aus, dass ein tendenziell überdurchschnittlicher hoher/niedriger Wert von X mit einem überdurchschnittlich hohen/niedrigen Wert von Y einhergeht.

Um jetzt nicht wie im vorherigen Teil intuitiv Indikatoren auszuwählen, also zu vermuten, welche Daten anhand eines Graphs korrelieren könnten, werden wir die Abhängigkeit der Fettleibigkeit von allen sinnvollen Merkmalserhebungen des Datensatzes mathematisch berechnen. Anschließend können wir direkt nur relevante (stärker korrelierende) Daten für eine Vorhersagemodell selektieren. Nach der Brute-Force-Methode werden die Korrelation der einzelnen Merkmale des Atlas zueinander ausgerechnet. Im Folgenden ist ein Ausschnitt der Korrelationsanalyse zu sehen, welche die Dimension und den Ausmaß der Exploration verdeutlichen soll:

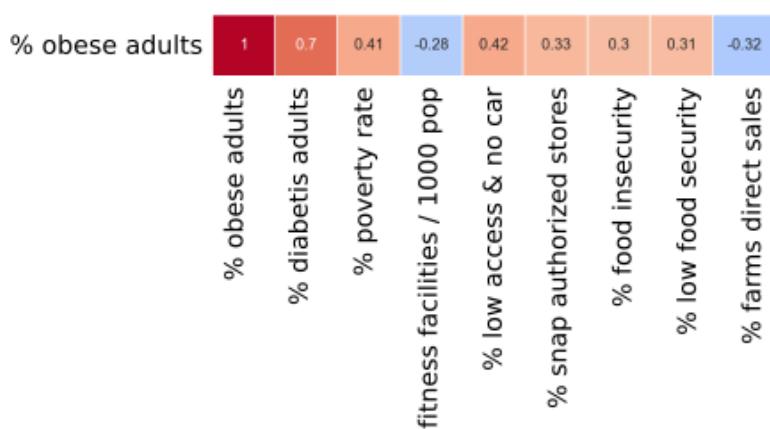


**Abb. 3.2.1** Heatmaps: Korrelationen der Atlas-Daten



Nicht alle dieser Daten sind brauchbar und sinnvoll zum weiteren Fortschreiten, sodass eine Reinigung und Konzentration auf die wichtigsten Ausprägungen nötig ist. Durch Herausfiltern der nicht korrelierenden beziehungsweise nur sehr schwach korrelierenden Merkmale, entsteht folgende Grafik, welche den Ausgangspunkt der weiteren Arbeit darstellen wird.

**Abb. 3.2.2 Heatmap: Signifikante Korrelationen mit Fettleibigkeit**



Man kann an dieser Heatmap sehr gut erkennen, welche Merkmale die Fettleibigkeit beeinflussen. Eine Steigung eines roten Merkmals bewirkt auch eine positive Entwicklung der prozentualen krankhaften Übergewichtigkeit, während eine bläuliche Färbung eine Senkung jener beschreibt. Die Intensität der Verfärbung beschreibt die Stärke der Korrelation in Vergleich zu den anderen Messgrößen. Weiter sei angemerkt, dass nicht alle Merkmale der Statistik auf Grundlage desselben Jahres erhoben worden. Trotzdem bleibt die Aussage der Daten bestehen, da eine Interpolation – also eine approximative Annäherung der Daten – keine bessere Genauigkeit ergeben würde. Des Weiteren handelt es sich bei diesen Merkmalserhebungen um Ausprägungen, welche über mehrere Jahre relativ konstant bleiben, sodass eine Differenz um ein Jahr keinesfalls zu einer Verfälschung der Aussage führt.

Viele Aspekte des analysierten Datensatzes lassen sich durch einen Mangel an finanziellen Mitteln erklären. So korreliert die Armutssquote (relative Armut) in den USA mit der Fettleibigkeit. Aber auch die Ernährungssicherheit – also die Verfügbarkeit von Nahrung und der Zugang zu Lebensmitteln – hat einen direkten Einfluss auf die adipöse Entwicklung eines Menschen. Eine lokale Nähe zu Farmen, Fitnessstudios oder öffentlichen Freizeiteinrichtungen geht mit einer Minderung des Übergewichtes der dort ansässigen Menschen einher. Dies sieht man auch daran, dass eine fehlende regionale fehlende oder mangelhafte Verfügbarkeit von Lebensmitteln nur einen messbaren Einfluss hat, wenn die Leute dort auch kein Auto besitzen. Ein räumliches Vorhandensein von Sporteinrichtungen und Verkäufern von Nahrungsmitteln, scheint also ein wichtiges Kriterium zur Verhinderung von Fettleibigkeit zu sein.

## 3.2 Lineare Regression

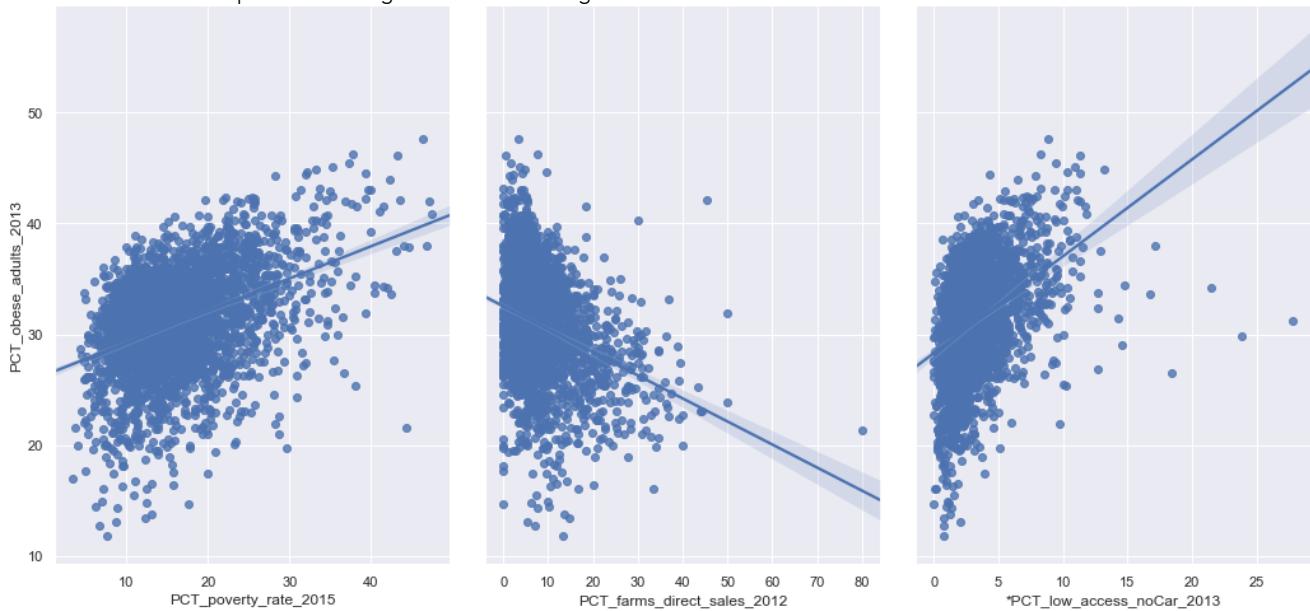
Einen Zusammenhang zwischen zwei Größen zu finden ist ein wichtiger Schritt, jedoch haben wir dadurch nur eine sehr eingeschränkte Vorhersagemöglichkeit. Ziel ist es also ein Vorhersagemodell zu schaffen, welches möglichst präzise Aussagen zur Wahrscheinlichkeit zur Adipositas eines Menschen liefert, wenn sein Umfeld bekannt ist. Mithilfe einer festgelegten exogenen Variable versuchen wir nun also eine Prognose über den Regressanden, also die Fettleibigkeit, zu erreichen.  $f(X) \rightarrow Y$

Eine Herangehensweise an diese Vorhersagemöglichkeit ist die einfache lineare Regression. Wie der Name schon vermuten lässt, wird eine lineare Funktion gesucht, dessen Gerade den Zusammenhang möglichst gut beschreibt:  $h_{\theta}(x) = \theta_0 + \theta_1 x$

Die Funktion der linearen Regression kann mit Hilfe der Methode der kleinsten Quadrate ermittelt werden. Man versucht die quadrierte Abweichung der einzelnen Punkte von der Geraden zu klein wie möglich zu halten. Das Quadrieren bewirkt, dass weit entfernte Punkte eine starke Gewichtung in der Korrektur der Geraden darstellen, sodass die Annäherungsfunktion eine breite Masse von Punkten möglichst genau abbildet. Dieses Verfahren ist in der Implementierung ein Supervised Machine-Learning-Algorithmus, da bereits existierende Daten genutzt werden, um das Modell immer weiter zu präzisieren. Sollte der Algorithmus mit genug Daten trainiert worden sein, so ist er – wenn der Zusammenhang zwischen den beiden Merkmalen annähernd linear ist – in der Lage, eine verlässliche Aussage zu treffen.

Beispielhaft sind unten einige Streudiagramme der linearen Regression dargestellt. Dabei sind die Merkmale Armutssquote, Anzahl der Farmen, welche direkt am lokalen Markt verkaufen, und der Leute, welche nur wenig Zugang zu Verkäufern haben und gleichzeitig kein Auto besitzen, mit dem prozentualen adipösen Teil der Gesellschaft in Verbindung gesetzt.

Abb. 3.2.3 Scatterplot: lineare Regression zur Fettleibigkeit



Betrachtet man die Datenpunkte und die daraus resultierende Regressionsgeraden, so sieht man, dass diese nur bedingt fähig sind, eine zuverlässige Aussage zu treffen. Nur ein äußerst geringer Anteil der Abweichung würde durch den linearen Zusammenhang erklärt. Fettleibigkeit lässt sich demnach nicht auf eine einzelne Ursache zurückzuführen, sondern muss zumindest als Ergebnis einer Menge von Ursachen angesehen werden. Wir müssen also den zweidimensionalen Raum verlassen, da die univariablen lineare Regression keine ausreichend prägnanten Erkenntnisse produziert. Der nächste Schritt zu einem zuverlässigeren Vorhersagemodell führt zur multivariablen linearen Regression.

### 3.3 Multivariable lineare Regression

Die multivariable lineare Regression unterscheidet sich von der univariablen linearen Regression insofern, dass wir nicht nur ein Merkmal in die Betrachtung einwirken lassen, sondern eine Menge von Merkmalen.

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

Die Vorgehensweise zur Ermittlung der Geraden bleibt dieselbe, jedoch wirken jetzt mehrere Dimensionen auf die Anpassung der Koeffizienten ein. Daher ist eine Darstellung als Geraden nicht im menschlich wahrnehmbaren dreidimensionalen Raum möglich. Jedoch kann man die über die Testdaten vorhergesagten Werte mit den tatsächlich gemessenen Werten vergleichen und diese zur Bewertung der Qualität des Modells kartieren.

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_snapAuthorized_stores_2012	-0.305264
PCT_food_insecurity_2014*	0.141917
PCT_low_food_security_2014*	0.015442
PCT_farms_direct_sales_2012	-0.087378

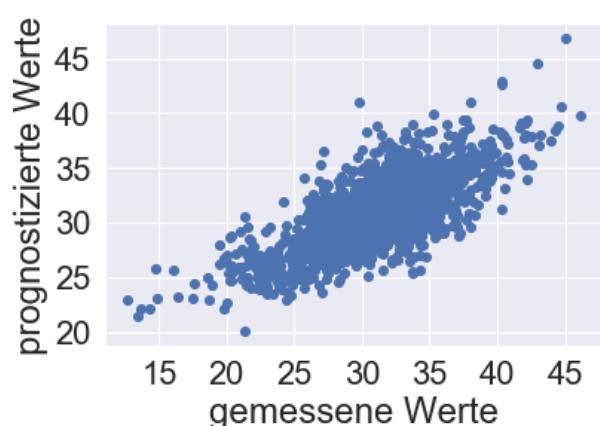


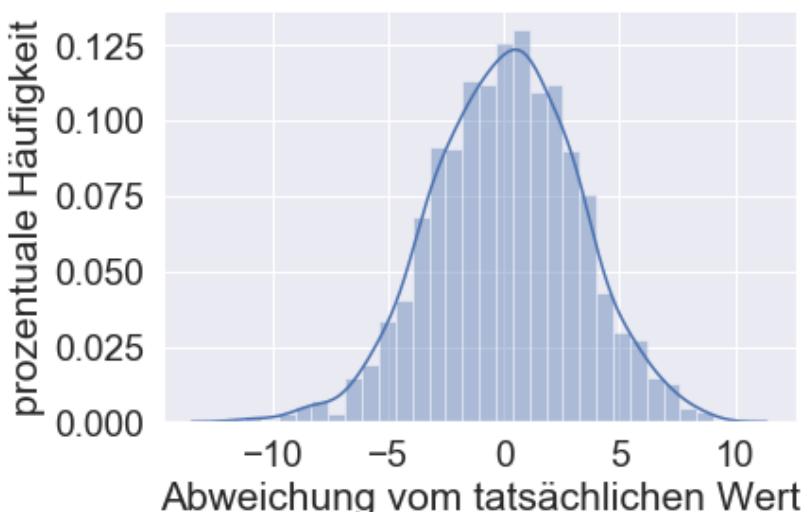
Abb. 3.2.4 Scatterplot: Abweichung der prognostizierten Werte

Unsere Regressions-Algorithmus wurde mit 50% (ca. 1750 Datensätze) der Datenquelle trainiert und mit der anderen Hälfte ausgewertet. Jede Merkmalsausprägung fließt mit einem Koeffizienten in die Prognosefunktion ein. Sollte das Vorhersagemodell optimal arbeiten, wäre im Scatterplot eine Ursprungsgerade zu sehen, da die prognostizierten Werte dann mit den tatsächlich gemessenen Werten übereinstimmen. In Fall des in dieser Studie erzeugten Modells sieht man, dass dies nur annähernd der Fall ist. Ist dieses Modell also nicht brauchbar? Im Folgenden werden Auswertungsmetriken betrachtet, welche eine Bewertung der Güte des Modells erlauben.

## 3.4 Auswertungsmetriken

Zunächst betrachten wir einmal die Abweichung einzelnen Werte vom tatsächlichen Wert. Wir sehen, dass dieses Modell über einen Großteil der Daten eine gute Aussage mit nur einer sehr geringen Abweichung liefert. Jedoch streut das Modell leider zu beiden Seiten sehr stark.

Abb. 3.2.5 Histogramm: Verteilung der Abweichungen



Dieser Graph reicht jedoch zur Auswertung nicht aus. Wir betrachten einige mathematische Kennzahlen.

$$\text{Mean Absolute Error (MAE): } \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| = 2.5016$$

$$\text{Mean Squared Error (MSE): } \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 9.8353$$

$$\text{Root Mean Squared Error (RMSE): } \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = 3.1361$$

Anhand dieser Werte ist zu erkennen, dass unser Modell im Schnitt nur rund 2,5% vom tatsächlichen Prozentsatz der Fettleibigkeit innerhalb eines Countys abweicht. Problematisch wird es jedoch, wenn wir die Ausreißer durch das Quadrieren stärker gewichten, sodass dieses Modell auf ganzer Linie versagt. Eine durchschnittliche Abweichung von 10% ist eindeutig zu viel, um es als zuverlässig prognosefähig einzuschätzen. Anderseits streuten die Ursprungsdaten aus der Datenquelle auch bei einigen Merkmalen

relativ stark, sodass auch eine normalisierende Darstellung mittels des RSME-Wertes mit ca. 3,1% Abweichung als gar nicht schlecht erscheinen mag.

Als letzte Bewertungsmethode werten wir das Bestimmtheitsmaß  $R^2$  aus. Dieser Determinationskoeffizient beschreibt in der Statistik die Anpassungsgüte einer Regression - konkret gesagt also, wie gut die Messwerte zu dem aufgestellten Modell passen. Es ist wie folgt definiert:

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

Dies bedeutet, dass unser Modell über 50% aller getesteten Daten erklären kann. Im Umkehrschluss heißt dies, dass die Hälfte der Daten zu sehr streuen, um von dem aufgebauten Modell beschrieben werden zu können. Inwiefern das Modell jetzt nutzlos ist oder nicht, kommt auf den Anwendungsbereich an. Es eignet sich gut um Tendenzen einiger Einflussfaktoren auf die Fettleibigkeit zu erkennen, jedoch um präzise Vorhersagen über eine prozentuale Adipositas-Rate im County zu erhalten taugt es kaum. Dafür streuen einige Messwerte zu stark, sodass sie generell nicht gut auf einer linearen Funktion darstellbar sind. Grundsätzlich belegt das Modell jedoch einen Zusammenhang zwischen sozioökonomischen Faktoren und Übergewicht.



Abb. 4.1.1 NHANES Examination Trucks

#### 4 NHANES National Health and Nutrition Examination Survey

## 4.1 Einführung

**Limitationen Food-Atlas:** Innerhalb der Untersuchung des „Food-Atlas“ konnten wenig starke Korrelationen zwischen Übergewicht (bzw. Adipositas) und sozioökonomischen Faktoren gefunden werden. Dies könnte aber auch in der Natur der Daten begründet liegen: Der Atlas bildet durch Bündelung der Erhebungen auf County-Ebene nur prozentuale Bevölkerungsanteile für eine Vielzahl von Indikatoren. Über Haushalte oder Individuen wird keine Aussage möglich. Dies spiegelt sich insbesondere in der Tatsache, dass die einzelnen Merkmale strukturell nicht verknüpft sind. Wir können zwangsläufig nicht wissen, ob etwa bei einem höheren Prozentsatz an einkommensschwachen Familien und einem gleichsam gesteigerten Anteil an fettleibigen Patienten auch tatsächlich die Ärmeren stärker zu Übergewicht tendieren. Die Beziehung bleibt unbestimmt.

**Alternativer Datensatz:** Aus diesem Grund wird im Folgenden ein zweiter Datensatz untersucht: NHANES, eine großflächig angelegte Studie zur Ernährung und Gesundheit in den USA. Im Rahmen dieser Erhebung wurden freiwillige Teilnehmer klinisch untersucht und befragt. Dies erfolgte in speziellen Trucks, welche als mobile Kliniken durch die Staaten reisten. Da innerhalb dieser Studie auch demographische und sozioökonomische Kennwerte erfasst wurden, verfügen wir somit über einen Datensatz, mit dem sich Wechselbeziehungen realistisch erforschen lassen. Die Daten sind anonymisiert, aber über einen durchlaufenden Patientencode verknüpft. Da NHANES eine Reihe von zweijährlich durchgeföhrten Studien umfasst, wurden hier fünf Jahrgänge ausgewählt, die sich grob mit den schon im Atlas sondierten Zeiträumen decken: 2007-2015. Jeder dieser Abschnitte enthält rund 10.000 Untersuchungsdaten. Wir erhalten so einen umfassenden Datensatz von mehr als 48.000 Zeilen.

### Bereinigen und kategorisieren der Daten

Die Daten wurden weiterhin auf Aussagekraft, statistische Standards und fehlende Daten hin überprüft. Das entsprechende Online Portal stellt die dafür notwendige Dokumentation bereit. Im Rahmen dieser Arbeit wurden die Merkmale entschlüsselt, umbenannt und ergänzt. Schließlich wurden stetige Daten zu diskreten oder kategorischen Variablen konvertiert. In diesem Rahmen erfolgte auch eine Prüfung der Verteilung der Daten, mit dem Ziel widersprüchliche oder wenig representative Merkmale zu erfassen. Auf eine detaillierte Beschreibung dieses Prozesses soll hier verzichtet werden. Die vollständige Dokumentation liegt jedoch im Anhang (Source Code) vor.

## 4.2 BMI und Einkommen

Die Untersuchung der Korrelation von Body-Mass-Index und Einkommen ist der Ausgangspunkt für die Erforschung von sozioökonomischen Faktoren, welche in einem Zusammenhang mit der grassierenden

Fettleibigkeit in den USA stehen könnten. Ähnlich wie zuvor beim Food Atlas, wird hier nun über den NHANES-Datensatz das Einkommen der Haushalte als Kernindikator herausgegriffen und in Beziehung zum BMI gesetzt.

## Hypothesen

**Nullhypothese H<sub>0</sub>:** Übergewicht und Fettleibigkeit korrelieren negativ mit Einkommen. Das heißt, mit sinkenden Einkommen steigt die Wahrscheinlichkeit für ungesundes Körpermass.

**Alternativhypothese H<sub>1</sub>:** Übergewicht und Einkommen korrelieren nicht. Es besteht kein erkennbarer Zusammenhang zwischen Fettleibigkeit und diesem sozioökonomischen Indikator.

**Erweiterung:** Dieses sehr grob gefasste Hypothesenpaar muss je nach Aussagekraft der verwendeten Tests angepasst und zum Teil umgestellt werden. Das Erkenntnisziel ändert sich durch diesen Vorgang nicht.

## Selektion und Aufbereitung der Daten

Wir wählen zunächst die kategorisierten Tabellenspalten für BMI und Haushaltseinkommen aus.

**BMI:** Es stehen folgende Merkmalsausprägungen für den Body-Mass-Index zur Verfügung: {10-18; 18-25; 25-30; 30-35; 35-40; 40+; unbekannt}. Diese Ausprägungen sind geordnet aber nicht in streng gleichförmigen Intervallen. Sie repräsentieren die gesundheitlich relevanten Kategorien: Untergewicht, Normalgewicht, Übergewicht und Fettleibigkeit 1-3.

**Einkommen:** Die Einkommensklassen beschreiben Intervalle mit unterschiedlicher (ansteigender) Breite. Dazu kommt ein nach oben offener Maximalwert "100+K" sowie die weniger präzisen Ausweichkategorien "<20K" und ">20K". Also: {000-005K; 005-010K; 010-015K; 015-020K; 020-025K; 025-035K; 035-045K; 045-055K; 055-065K; 065-075K; 075-100K; 100+K; <20K; >20K; unbekannt}.

**Fehlende Daten:** Von den 47873 Datenreihen enthalten einige die nicht aussagekräftigen "unbekannt" Einträge, welche ausgeschlossen werden. Damit reduziert sich der Datensatz auf 41800 Stichproben.

**Kontingenztabelle:** Der selektierte Datensatz kann über Pivotieren in eine Kontingenztabelle umgewandelt werden. Als Reihen werden nun die BMI-Klassen und als Spalten die Einkommensgruppen abgebildet. Die jeweiligen Ausprägungen werden dabei zu Zählungen der absoluten Häufigkeiten aggregiert. Eine weitere Tabelle der relativen Häufigkeiten erhalten wir über Division durch die Spaltensumme. Dieser Tabelle ist schließlich zu entnehmen, wie in jeder Einkommensgruppe die BMI-Klassen prozentual verteilt sind.

household_income	000-005K	005-010K	010-015K	015-020K	020-025K	025-035K	035-045K	045-055K	055-065K	065-075K	075-100K	100+K	All
bmi													
10-18	282	406	577	647	752	1050	770	593	442	368	736	1412	8035
18-25	325	548	801	837	999	1505	1148	993	731	531	1167	2250	11835
25-30	210	393	732	704	854	1202	1004	810	605	520	912	1587	9533
30-35	135	284	437	462	500	759	625	472	363	317	566	845	5765
35-40	52	116	243	206	242	354	282	235	183	143	245	322	2623
40+	59	131	156	162	197	312	209	163	105	90	167	210	1961
All	1063	1878	2946	3018	3544	5182	4038	3266	2429	1969	3793	6626	39752

Tab. 4.2.1 Kontingenztabelle: BMI-Klassen vs. Haushaltseinkommen (absolute Häufigkeiten)

## Visuelle Analyse von BMI in Einkommensgruppen

Für die Untersuchung der Fettleibigkeit sind vor allem die BMI-Gruppen "25-30" (Übergewicht), "30-35" (Adipositas 1), 35-40" (Adipositas 2) und "40+" (Adipositas 3) interessant. Sie werden aus der Kontingenztabelle der relativen Häufigkeiten gelesen und als gestapelte Balken für jede Einkommensgruppe dargestellt.

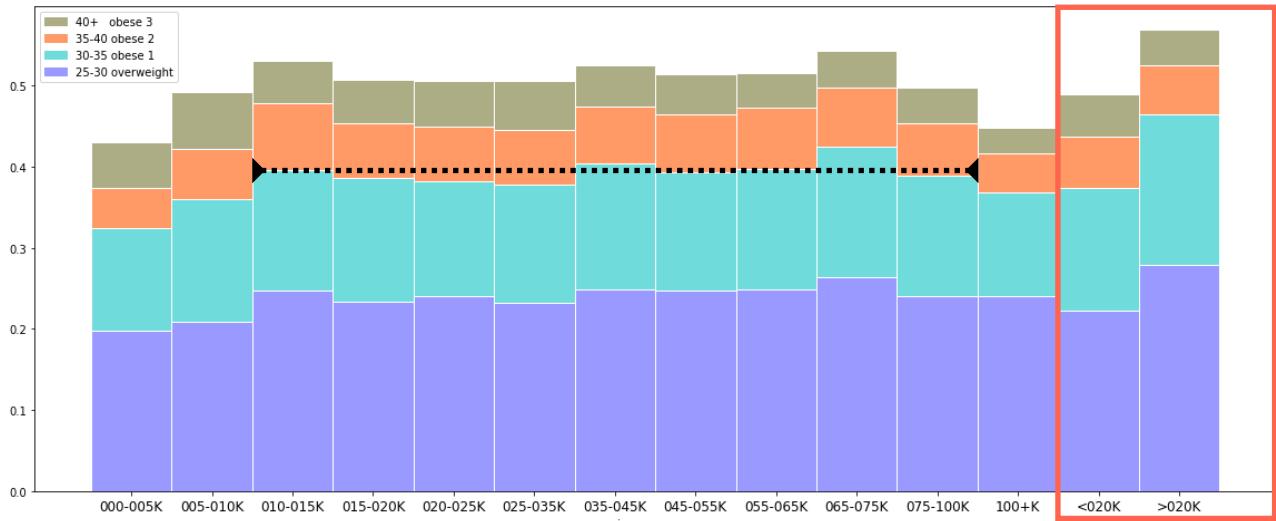


Abb. 4.2.1 Balkendiagramm: Prozentualer Anteil der höheren BMI-Klassen nach Haushaltseinkommen

**Das Resultat überrascht:** Zunächst wird das Ausmaß des Problems deutlich - in den USA liegt der Anteil an Menschen mit krankhafter Fettleibigkeit im Schnitt bei knapp 30%. Werden übergewichtige Personen hinzugerechnet, erreicht der Anteil fast 50%. Eine Anstieg des prozentualen Anteils von übergewichtigen oder fettleibigen Menschen bei sinkendem Einkommen ist in keiner Form zu erkennen. Stattdessen finden wir in den untersten Einkommensklassen sogar einen Rückgang. In der Spanne von 10.000\$-75.000\$ scheint die prozentuale Verteilung jedoch gleichförmig zu verlaufen. Jenseits von kleineren Schwankungen ist kein Trend auszumachen. Erst ab 100.000\$ sehen wir einen deutlichen Rückgang. Besonderheiten sind in den übergreifenden Einkommensgruppen, " $<20K$ " und „ $>20K$ “, zu finden. Hier scheint sich der vermutete Zusammenhang eher abzubilden. Bei jedoch gerade einmal 141 dargestellten Stichproben in der viel breiteren Einkommensklasse " $<20K$ ", sollte dieser Beobachtung jedoch nicht zu viel Bedeutung beigemessen werden.

## 4.2.2 Chi2-Test für BMI und Einkommensgruppe

Um den Zusammenhang zwischen Fettleibigkeit und Einkommensklasse weiter mit statistischen Methoden zu erforschen, soll der Chi-Quadrat-Test zur Anwendung kommen. Dieser bietet sich zunächst an, da hier mit kategorischen Daten und einer Kontingenztabelle gearbeitet wurde. Er kann jedoch nur nachweisen, ob die zwei Merkmale stochastisch unabhängig sind. Die Hypothesen sind entsprechend anzupassen.

**Nullhypothese  $H_0$ :** BMI und Haushaltseinkommen sind stochastisch unabhängig.

**Alternativhypothese  $H_1$ :** BMI und Einkommensklasse sind voneinander abhängige Variablen.

**Sigma:** Gewählt wird ein Signifikanzniveau von 5% ( $p = 0.05$ ).

Der Test basiert auf der Überlegung, dass die zu erwartenden Werte innerhalb einer Kontingenztabelle über die Randsummen berechnet werden können. Wir betrachten dann die Differenzen zu den realen Werten und berechnen Chi-Quadrat nach folgender Formeln:

$$\chi^2 = \sum_{i=1}^j \sum_{k=1}^m \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} = 301,21 , \quad \tilde{h}_{ij} = \frac{h_i * h_{*j}}{n}$$

## Kontingenzkoeffizient für BMI und Einkommensgruppe (nach Pearson)

Der berechnete Wert für Chi-Quadrat (301.21) ist noch nicht besonders aussagekräftig. Daher werden weitere Kenngrößen ermittelt bevor die Thesenprüfung erfolgt. Der korrigierte Kontingenzkoeffizient bringt den Wert von Chi-Quadrat in einen normalisierten Wertebereich von 0-1. Dafür wird zunächst der normierte Kontingenzkoeffizient berechnet, wobei n=9752 die Anzahl unserer Merkmalsausprägungen ist. Dieser Wert wird durch einen Korrekturfaktor dividiert, um den Wertebereich voll auszuschöpfen. In diesem Fall ist die minimale Tabellendimension k = 6 (die Zahl der BMI-Ausprägungen). Für die Auswertung des Chi-Quadrat-Test wird schließlich noch die Zahl der Freiheitsgrade (DF) benötigt.

Sie ist für die reduzierte Kontingenztabelle 55.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad C_{corr} = \frac{C}{C_{max}}, \quad C_{max} = \sqrt{\frac{k-1}{k}} \quad \text{mit } k = \min(x, y)$$

$$DF = (k - 1)(m - 1) \quad \text{mit } k = 6 \quad \text{und } m = 12$$

	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

Tabelle 4.2.2 Resultate Chi2-Test

## Auswertung Chi-Quadrat Test

Mit 55 Freiheitsgraden überschreiten wir nach Chi-Quadrat-Tabelle bei einem Wert von 73,311 das gewählte Signifikanzniveau von 5%. Das heißt die Wahrscheinlichkeit, dass der Wert für Chi-Quadrat 73,311 erreicht, liegt gerade bei 5%. Bei dieser geringen Wahrscheinlichkeit sprechen wir von einer signifikanten Beobachtung und verwerfen die Nullhypothese. Da der berechnete Wert von Chi-Quadrat für die Merkmale BMI und Einkommensgruppe bei 301,213 und somit weit über 73,311 liegt, ist dies der Fall:

*Die Nullhypothese, dass die untersuchten Merkmale stochastisch unabhängig sind, ist zu verwerfen.*

Dies scheint den Ergebnissen der visuellen Analyse entgegen zu laufen. Es ist jedoch zu beachten, dass aus dem Chi-Quadrat-Test nur ein vager Zusammenhang abzuleiten ist. Es wird keine Aussage über die Form der Abhängigkeiten getroffen. Eine Betrachtung des korrigierten Kontingenzkoeffizienten gibt weitere Auskunft: Der Wert liegt mit 0,095 relativ dicht bei 0. Damit ist der Chi-Quadrat zu hinterfragen.

## 4.2.3 Lineare Regression und p-Wert für BMI und Einkommensgruppe

In einem weiteren Schritt mittels Lineare Regression und p-Wert auf eine lineare Korrelation getestet.

**Nullhypothese H<sub>0</sub>:** Es besteht keine lineare Korrelation zwischen BMI und Einkommensgruppe.

**Alternativhypothese H<sub>1</sub>:** BMI und Einkommen antikorrelieren. Bei steigendem Einkommen sinkt der prozentuale Anteil an Testpersonen mit Übergewicht und somit der durchschnittliche Body-Mass-Index.

**Alpha-Level:** Das gewählte Signifikanzniveau bleibt bei 5% (p = 0.05).

---

## Einschränkungen und modifizierte Datenselektion

Da für die lineare Regression und den damit verbundenen p-Test stetige Daten, oder zumindest Daten auf Intervallskalen, benötigt werden, ist diese Untersuchung nur eingeschränkt möglich:

**Der BMI** kann aus den ursprünglichen Tabellen als genauer Wert (stetige Verteilung) extrahiert werden.

**Das Einkommen** der Haushalte wurde hingegen kategorisch erfasst und mit einem numerisch geordneten Code von 1-12 betitelt. Dieser Code lässt sich auf einer Achse abbilden. Jedoch bleibt die Einschränkung bestehen, dass die codierten Einkommensintervalle bei höherem Einkommen an Breite gewinnen. Eine

mögliche Lösung ist die Annäherung an eine stetige Verteilung durch Zuweisung eines Zufallswertes innerhalb des codierten Intervalls. Alternativ kann auch der Mittelwert des Intervalls gewählt werden. Eine derartige Transformation der Daten birgt natürlich Gefahren, die eine Gegenprüfung erfordern.

## Analyse Streudiagramm & Boxplot

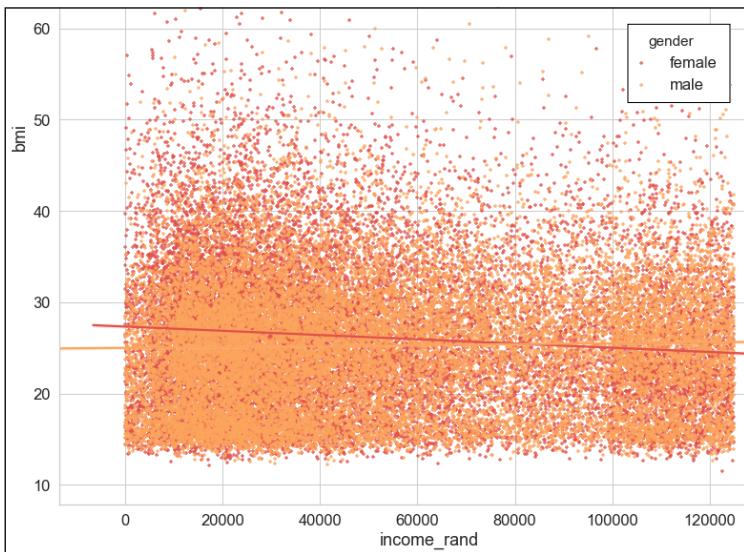


Abb. 4.2.3 Scatterplot: BMI vs. Einkommen

Der Scatterplot mit bietet einen Überblick über die Verteilung der Daten: Die Regressionslinien verlaufen um den mittleren BMI von 25 eher horizontal. Es scheint also auch hier wenig Korrelation zwischen BMI und Einkommen zu geben. Durch die Trennung der Datenpunkte nach Geschlecht ist zumindest bei den Frauen eine leichte Antikorrelation zu erahnen - die Linie fällt. Im Allgemeinen sind die Punkte aber für jedes Einkommen ähnlich (normal-) verteilt. Dies bestätigt auch ein Box-Whisker-Plot, der zur Überprüfung auf eine Intervall-Randomisierung der Einkommensdaten verzichtet.

## Auswertung Lineare Regression und p-Test

Beim p-Test betrachten wir die Überschreitungswahrscheinlichkeit  $p$ , welche im Rahmen der linearen Regression ermittelt wird. Der  $p$ -Wert für die lineare Korrelation von BMI und Einkommen ist mit  $1.25\text{e-}19$  erstaunlich niedrig und liegt deutlich unter der Grenze von 5%. Bei isolierter Betrachtung wäre also in Übereinstimmung mit dem Chi-Quadrat-Test die Nullthese abzulehnen. Es gibt demnach eine lineare Korrelation zwischen BMI und Einkommen. Gleichzeitig macht diese Aussage nach der visuellen Einschätzung der Daten intuitiv wenig Sinn. Der  $r^2$ -Wert ist mit 0.002 außerordentlich klein - gerade 0.2% der Abweichung werden durch das Regressionsmodell erklärt. Nehmen wir nun noch die Steigung der Regressionslinie hinzu, welche fast bei 0 liegt, dann muss festgestellt werden, dass eigentlich keine Aussage getroffen wurde. Der geringe  $p$ -Wert täuscht:

Tabelle. 4.2.3 Resultate 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

*Die Nullhypothese kann nicht direkt durch den  $p$ -Wert verworfen werden. Die Alternativhypothese einer Antikorrelation von BMI und Einkommen ist aber auch nicht anzunehmen.*

**Korrelationskoeffizient:** Abschließend werden Korrelationskoeffizienten (Bravais-Pearson) ermittelt. Sie sind ein Maß für die Stärke des Zusammenhangs. Die Berechnung erfolgt durch Division der Kovarianz der beiden betrachteten Variablen durch das Produkt ihrer Standardabweichungen.

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

Werden nur die Frauen betrachtet, dann liegt der Wert bei -9.98%. Demnach bestätigt sich die im Scatterplot ausgemachte leichte Antikorrelation. Insgesamt muss festgestellt werden, dass der Wert für alle Stichproben, von -0,0437, sehr dicht bei 0 liegt. Der Zusammenhang ist nachweislich zu schwach ( $p < 0,3$ ).

#### 4.2.4 Fazit: BMI und Einkommensgruppe

Für die Ausgangsfrage nach einem Zusammenhangs von BMI und Einkommen kann die These:

„Übergewicht und Fettleibigkeit korrelieren negativ mit Einkommen. Das heißt, mit sinkenden Einkommen steigt die Wahrscheinlichkeit für ungesundes Körpergewicht.“ nicht bestätigt werden.

Auch wenn sowohl der Chi-Quadrat als auch der p-Test einen Zusammenhang nicht grundsätzlich ausschließen, zeigen weitere Untersuchungen doch offensichtlich, dass die Datenlage zu diffus ist, um eine Annahme der These zu rechtfertigen. Damit widerspricht diese Untersuchung der Mehrzahl der Studien, die Fettleibigkeit durch geringes Einkommen begünstigt sehen. Vermutlich müssen weitere sozioökonomische Faktoren hinzugezogen werden, um eine besseres Bild zu gewinnen.

#### Präzisierung des Einkommens durch Mitglieder im Haushalt

Das Haushaltseinkommen kann noch weiter präzisiert werden, indem die Anzahl der Mitglieder im jeweiligen Haushalt einbezogen wird. Die Vermutung wäre, dass Großfamilien mit geringem Einkommen noch zusätzlich benachteiligt würden. Tragen wir aber das berechnete Einkommen pro Kopf gegen den BMI in einem Scatterplot auf, so verändert sich der Eindruck nicht wesentlich.



Abb. 4.2.4 Heatmap: Prozentuale Verteilung der Anzahl vom Haushaltmitgliedern nach Haushaltseinkommen

**Heatmap für Einkommen und Zahl der Haushaltmitglieder:** Eine Untersuchung des Zusammenhangs von Einkommen und Mitgliederzahl mittels Heatmap, macht deutlich, dass gering verdienende Familien keineswegs mehr Personen (Großfamilien) zu versorgen haben. Stattdessen überwiegen in den unteren Einkommensklassen Ein- und Zweipersonenhaushalte, während vier Personen bei den Großverdienern Vorrang erhalten. Dies erklärt, warum die Mitgliederzahl im Haushalt, die Antikorrelation von BMI und Einkommen eher schwächt: Das Einkommen wird etwas ausgeglichen.

**Anmerkung:** Die Methodik der statistischen Nachweise mit Hypothesenpaar wurde nun ausführlich beleuchtet. In den folgenden Abschnitten wird auf diese Präzision weniger Wert gelegt, um einen breiteren Überblick darzustellen. Zur genaueren Prüfung der Aussagen sei auf den Source-Code Anhang verwiesen.

## 4.3 BMI und Bildung

Ein gemeinhin mit dem Einkommen assoziierter sozioökonomischer Faktor ist die Bildung. Durch Einbezug dieses Feldes hoffen wir, die bisher diffusen Aussagen zu schärfen. In einem ersten Schritt wird geklärt, ob der NHANES-Datensatz die positive Korrelation von Bildung und Einkommensgruppe auch tatsächlich abbildet. Danach kann dann geprüft werden, welche Zusammenhänge zwischen Fettleibigkeit und Bildung bestehen.

### 4.3.1 Einkommen und Bildung

#### Chi-Quadrat Test

**Nullhypothese  $H_0$ :** Einkommen und Bildung stochastisch unabhängig.

**Alternativhypothese  $H_1$ :** Einkommen und Bildung sind voneinander abhängige Variablen.

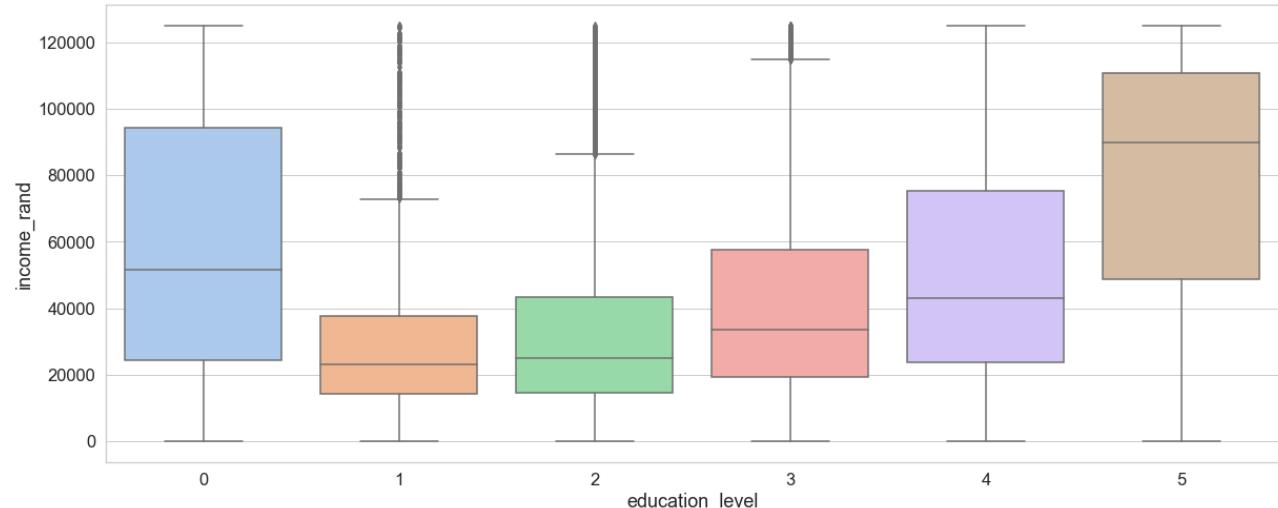
**Sigma:** Gewählt wird ein Signifikanzniveau von 5% ( $p = 0.05$ ).

Der Chi-Quadrat-Test wurde wiederum über eine Kontingenztabelle mit den Merkmalen Bildungslevel und Einkommensklasse ermittelt. Der berechnete Wert ist **10.133**. Bei 44 Freiheitsgraden würde die Wahrscheinlichkeit, dass Chi-Quadrat einen Wert von 60,46 erreicht bei 5% liegen. Dies ist das gewählte Signifikanzniveau. Da der berechnete Wert weit darüber liegt, ist die Nullhypothese zu verwerfen.

*Bildung und Einkommen sind somit stochastisch abhängig.*

#### Korrelationkoeffizient und Plotanalyse

**Abb. 4.3.1** Boxplot: Einkommen vs. Bildungsstufe



Über den Kontingenzkoeffizienten lässt sich für Einkommen und Bildung klar eine positive Korrelation nachweisen. Der entsprechende Wert liegt bei 0,3983 ( $> 0,3$ ), also 39,8%. Dieser Zusammenhang ist auch im Boxplot abzulesen: Der Anstieg scheint nicht linear sondern quadratisch bis exponentiell zu erfolgen. Da es sich bei den Bildungsstufen aber nicht um Werte auf einer Kardinalskala handelt, ist die Herstellung einer mathematischen Funktion jedoch nicht möglich. Eine alternative Darstellung, die die positive Korrelation von Bildung und Einkommen offen legt, ist die "Heatmap". Diese kann aus der zuvor erstellten Kontingenztabelle abgeleitet werden. Sie zeigt die für jede Einkommensgruppe prozentual, aus welchen Bildungsklassen sich die Testpersonen speisen. Das diagonale „Hitzeband“ bestätigt die Ergebnisse.

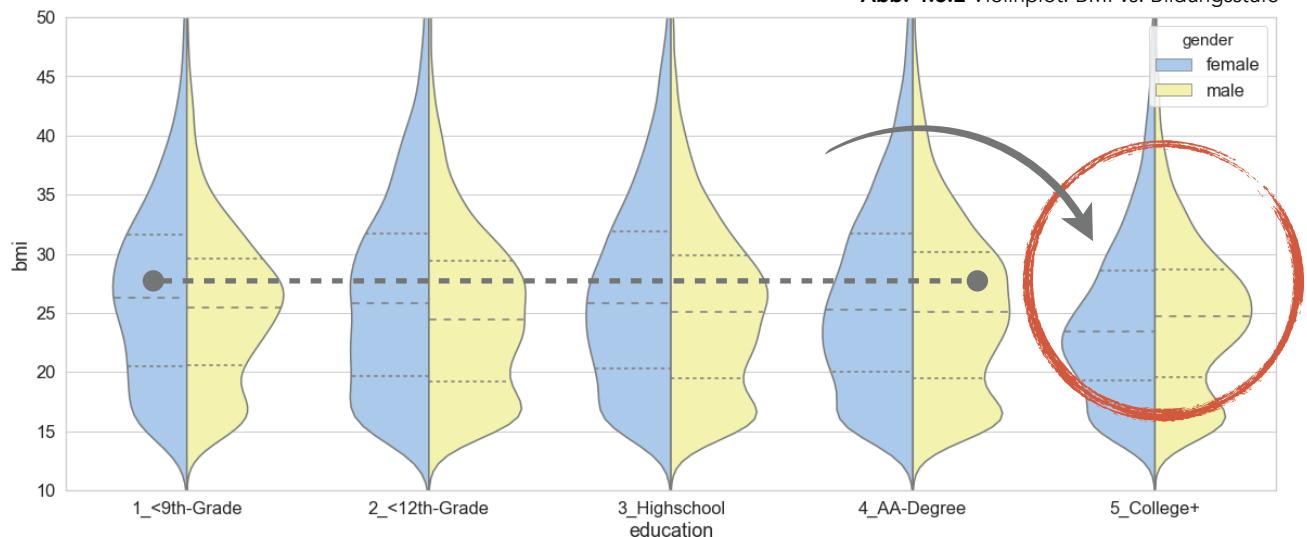
### 4.3.2 BMI Verteilung nach Bildungslevel

Da nun der triviale Zusammenhang zwischen Bildung und Einkommen nachgewiesen wurde, ist zumindest klar gestellt, dass die NHANES-Daten überhaupt eine Aussagekraft besitzen. Um so mehr ist demnach auch den der Intuition entgegen laufenden Erkenntnissen zu trauen. In diesem Schritt wird nun die Beziehung zwischen Bildungslevel und BMI beleuchtet. Als Hypothese anzunehmen wäre eine negative Korrelation zwischen diesen Variablen. Bei einem höheren Bildungslevel müsste demnach der Anteil an Übergewichtigen und Fettleibigen deutlich sinken.

#### Violinplot der BMI-Verteilung nach Bildungslevel

Der Violinplot stellt die Dichtefunktion des BMI für jedes Bildungslevel nach Geschlecht differenziert dar. Auch hier zeigt sich ein Bild, dass der zuvor erfolgten Gegenüberstellung von BMI und Einkommen entspricht: Der BMI liegt bei den Frauen zunächst etwas höher. Über die ersten vier Bildungsstufen liegen die Median- und Quartilslinien in etwa auf gleicher Höhe. Dies würde implizieren, dass das Übergewicht als Breitenphänomen vollkommen losgelöst vom Bildungsstand ist. Das Bildungslevel 5 (College und Universität) weist jedoch einen starken Rückgang in den höheren BMI-Klassen auf. Insbesondere bei den Frauen liegen die Kennlinien dort deutlich niedriger und erstmals sogar unter den Werten der männlichen Testpersonen. Wie schon beim Einkommen gilt für Frauen eine negative Korrelation von Bildung und BMI, die allerdings erst auf dem höchsten Level signifikant greift. Bei den Männern ist diese Tendenz weniger stark ausgeprägt. Der Rückgang betrifft vor allem die Bereiche krankhafter Fettleibigkeit (oberes Quartil).

Abb. 4.3.2 Violinplot: BMI vs. Bildungsstufe



#### Korrelationkoeffizient für BMI und Bildungslevel

Der Korrelationskoeffizient für BMI und Bildung liegt bei -0,049 und bestätigt somit die aus den Violinplots abgelesene minimale Antikorrelation. Gerade im Vergleich zu den bei der Verknüpfung von Einkommen und Bildung erreichten 0,39 wird deutlich, wie dünn der Hinweis auf eine lineare Beziehung ist. Der Wert von |-0,049| liegt weit unter der 0,3 Grenze. Die Beziehung ist somit nicht signifikant. Die Restspur einer Antikorrelation speist sich vor allem aus dem deutlichen Abfall des starken Übergewichts im höchsten Bildungslevel bei gleichzeitig ausbleibender Gegentendenz in allen anderen Stufen.

## 4.4 BMI und Ethnische Gruppierung

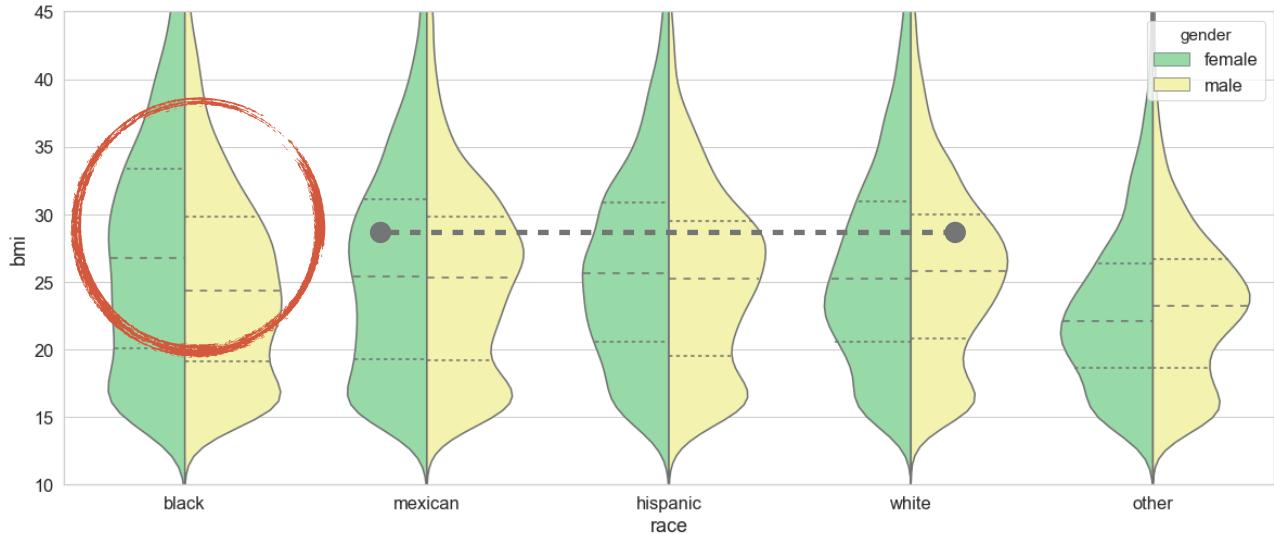
Für das Einwanderungsland USA stellt die Zugehörigkeit zu einer ethnischen Gruppierung einen wichtigen sozioökonomischen Faktor dar. Daher soll dieser hier nicht vernachlässigt werden.

**Selektion der Daten:** Für die Betrachtung der ethnischen Zugehörigkeit stehen im NHANES-Datensatz vier Klassen zur Verfügung: {white; black; mexican; hispanic; other}. Etwas problematisch ist die Gruppe „andere“. Sie umfasst viele marginale Minderheiten, sowie auch die größere Gruppe der Asiaten. Eine weitere Differenzierung wäre hier wünschenswert, wurde im Rahmen des NHANES-Programms aber erst in den letzten drei der insgesamt fünf verwendeten Jahrgänge vorgenommen.

#### 4.4.1 Plotanalyse zur BMI Verteilung in ethnischen Gruppen

Ein Box-Whisker-Plot, der die Verteilung des BMI in den verschiedenen ethnischen Gruppen darstellt, ist durchaus aufschlussreich: Die Tendenz zu einem erhöhten BMI ist bei Afroamerikanern und mexikanischen Einwanderern deutlich ausgeprägter. Insbesondere bei den Afroamerikanern gibt es eine markante Verschiebung des oberen Quartils hinein in den Bereich der krankhaften Fettleibigkeit (BMI über 30). Obwohl also die Mediane bei vier der fünf ethnischen Gruppen gleichauf liegen, gibt es bei Afroamerikanern und auch Mexikanern einen entschieden größeren Anteil mit gesundheitlich kritischer Fettleibigkeit. Das Gegenteil findet sich in der Sammelgruppe „andere“. Hier liegt der Median im Normalgewicht. Dies ist stark bedingt durch den großen Anteil an Asiaten in dieser Gruppe.

Abb. 4.4.1 Violinplot: BMI vs. Ethnische Gruppe

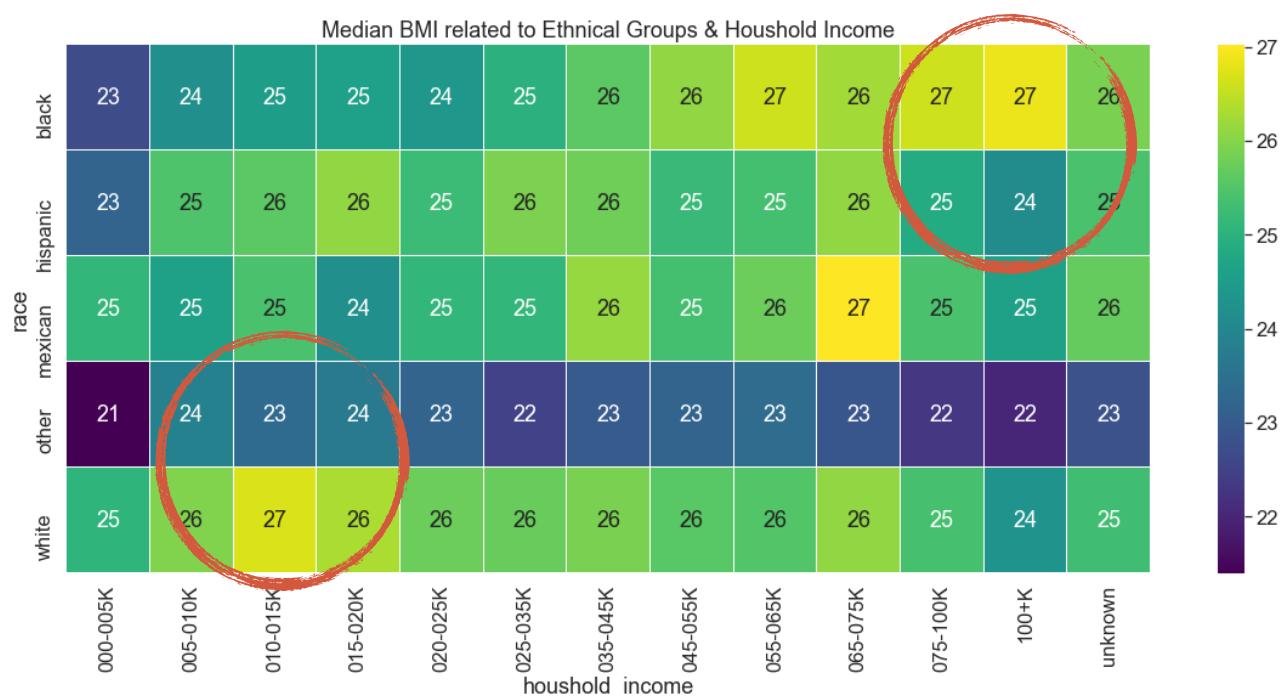


**Weitere Differenzierung nach Geschlecht:** wird Mittels des geteilten Violinplots möglich. Hier überrascht die Rolle der weiblichen Testpersonen innerhalb der afroamerikanischen Gruppe. Während sich die Verteilung des BMI der männlichen Teilnehmer nicht allzu stark von Weißen, Mexikanern und Lateinamerikanern unterscheidet, wird der zuvor ausgemachte BMI-Anstieg in den krankhaften Bereich vor allem durch Frauen realisiert. Die obere Quartilsline nähert sich stark einem BMI, der einer Adipositas Grad II entspricht. In der ethnischen Sammelgruppe „andere“ kehrt sich diese Struktur hingegen um.

**Heatmap des BMI-Median:** Die Betrachtung des BMI in Hinblick auf die ethnische Gruppe, kann zusätzlich nach Einkommen differenziert werden. Dies wird durch die Bildung einer Heatmap ermöglicht, die den median des BMI für jede Ethnie und Einkommensklasse kartiert. Ein erhöhter Median weist auch auf größere prozentuale Anteile in den oberen BMI-Klassen hin. Interessant sind hier vor allem zwei Beobachtungen: Sowohl bei den Mexikanern als auch bei den Afroamerikanern liegen die BMI-Spitzen in den oberen Einkommensklassen. Die Antikorrelation von BMI und Haushaltseinkommen gilt hier gerade nicht. Nur bei den weißen Amerikanern gibt es einen minimalen Schwerpunkt in einem Intervall von 10-15

**Abb. 4.4.2** Heatmap: BMI-Median vs. Ethnische Gruppe und Einkommen

Matthias Titze, s0563413



Tausend Dollar. Da der Anteil der weißen Bevölkerung aber in den USA mit 36,1% überwiegt, nivellieren sich die widersprüchlichen Tendenzen aus.

## 4.5 Vorhersagemodell - Supervised Machine Learning

Im Folgenden soll versucht werden, ein Vorhersagemodell (Klassifizierung) für die BMI-Klasse zu erstellen. Dazu verwenden wir Supervised-Machine-Learning Algorithmen, d.h. die vorhandenen Daten werden in Trainingsdaten und Testdaten unterteilt. Mit den ersten Trainieren wir den Prognose-Algorithmus. Anschließend lassen wir diesen über die Testdaten eine Vorhersagetreffen treffen und vergleichen das Resultat mit den für die Testdaten vorliegenden, realen Merkmalsausprägungen.

**Limitationen:** Da die bisherigen Untersuchungen gezeigt haben, dass es für die Tendenz zu Übergewicht und Fettleibigkeit keine entscheidenden sozioökonomischen Faktoren gibt, muss vermutet werden, dass auch eine Machine-Learning-Algorithmus keine klaren Aussagen treffen kann. Dies ist insbesondere für den Ansatz der Klassifizierung des Körperfanges, einer Eigenschaft, die eher normal verteilt ist, anzunehmen: Auch wenn es Faktoren gäbe, die eine Tendenz zur Adipositas leicht begünstigen würden, wäre die so gesteigerte Wahrscheinlichkeit immer noch geringer als der Druck hin zum Erwartungswert der Normalverteilung. Klassifizierung als Vorhersagemodell ist somit eher nicht der treffende Ansatz. Günstiger wäre ein Modell, welches eine Aussage darüber treffen würde, in welchem Maß die Wahrscheinlichkeit durch entsprechende Einflussfaktoren beeinflusst wird. Trotz dieser grundsätzlichen Problematik soll, die Klassifizierung mit Naive-Bayes getestet werden. Treffen wir dort auf überraschende Ergebnisse, so findet sich im Umkehrschluss in den Daten doch eine Kombination von Eigenschaften, die kritisch für die Ausbildung von Adipositas ist.

**Vorherzusagen (y):** BMI-Klasse {Untergewicht, Normal, Übergewicht, Adipositas 1-3}

**Genutzte Merkmale (xi):** {Geschlecht, Ethnie, Mitglieder im Haushalt, Einkommen, Bildung, Alter}

## 4.5.1 Klassifizierung mit Naive Bayes (Gauss Version)

Der Naive Bayes Klassifikator basiert auf Bayes Theorem. Er wird als "naive" bezeichnet, da er die Unabhängigkeit aller einbezogenen Merkmalspaare im Datensatz voraussetzt.  $V = (x_1, x_2, \dots, x_n)$  sei der Vektor der Merkmale und  $y$  die Eigenschaft, die durch den Vektor voraus gesagt werden soll. Dann gilt:

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)}$$

und wenn  $x_1, x_2, \dots, x_n$  weiterhin als stochastisch unabhängig angenommen werden:

$$P(y|x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^n P(x_i|y)}{P(x_1, \dots, x_n)} \quad \text{bzw.} \quad P(y|x_1, \dots, x_n) = P(y)\prod_{i=1}^n P(x_i|y)$$

Diese Formel reduziert sich für die Klassifizierung, da die errechneten Wahrscheinlichkeiten nur auf das Maximum ausgewertet werden.  $P(x_i)$  ist eine Konstante, die somit ignoriert werden kann. Im Falle des hier verwendeten Gauss-Naive-Klassifikators wird eine Normalverteilung jedes Merkmals angenommen. Dann ergeben sich die Wahrscheinlichkeiten dafür zu:

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

**Potenzielle Fehlerquellen:** Schon vorab muss der gewählte Klassifikator kritisch betrachtet werden. Die Merkmale, Einkommen und Ausbildung, sind zum Beispiel keinesfalls unabhängig. Dies wurde in der Datenexploration nachgewiesen. Des Weiteren ist eigentlich nicht davon Auszugehen, dass die zur Vorhersage genutzten Eigenschaften normal verteilt sind.

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

Tab. 4.5.1 Datenselektion für Naive Bayes

### Modell Trainieren

Aus dem NHANES-Datensatz werden die oben genannten Merkmale als Grundlage für die Klassifizierung extrahiert. Voraussetzung ist die BMI-Gruppe. Alle genutzten Daten liegen als kategorisierte Eigenschaften vor, die hier in Form von Integer-Zahlen codiert

wurden. Auf fehlende Daten wurde der Einfachheit halber mit Ausschluss reagiert. Die Anzahl der Zeilen schrumpft so auf 40.686, was für den angestrebten Rahmen genügen soll. Der so extrahierte Datensatz wird anschließend zufällig in 75% Trainings- und 25% Testdaten unterteilt. Mit den ersten trainieren wir den Klassifikator aus der Python Bibliothek. Anschließend erfolgt die Vorhersage auf den Trainingsdaten.

## 4.5.2 Bewertung des Modells

Die erhaltenen Vorhersagen können mit den tatsächlich vorliegenden BMI-Daten abgeglichen werden.

**Wertebereich:** Schon bei der Überprüfung des Wertebereichs tritt Bedenkliches zu Tage: Während die tatsächlichen Werte, alle BMI-Klassen umfassen, liegen die vorhergesagten Ausprägungen nur in den Kategorien -1 (Untergewicht) bis 1 (leichteres Übergewicht). Der Klassifikator hat also versagt, Personen mit Adipositas 1-3 zu erkennen. Davon war jedoch von Beginn an auszugehen, da die betrachteten Faktoren nur einen leichten Einfluss auf die Tendenz zur Fettleibigkeit haben. Der Naive Bayes wählt, die

wahrscheinlichste Klasse aus. Auch wenn also Adipositas durch einige Merkmale etwas wahrscheinlicher würde, so bleibt der Erwartungswert der Normalverteilung eben doch dominant.

**Modifikation der BMI-Klassen:** Auch nach Verschmelzung der oberen BMI-Klassen bleibt das Problem erhalten: Es besteht keine Möglichkeit, krankhafte Fettleibigkeit vorher zu sagen. Dies ist vielleicht auch beruhigend, da somit klar wird, dass sozioökonomische Faktoren eben nicht allein über das Körpergewicht entscheiden, sondern maximal eine Tendenz fördern könnten. Der Naive-Bayes-Klassifikator ist jedoch nicht dafür geeignet, eine derartige Verschiebung abzubilden.

**Die Genauigkeit** ist mit 41,17 Prozent schwach und könnte schon damit erklärt werden, dass wir die Personen richtig vorher sagen, die ohnehin im Mittelfeld liegen.

## Konfusionsmatrix

Um die Trefferquote genauer zu beleuchten, hilft eine Matrix, welche die realen und die vorher gesagten Merkmalsausprägungen gegenüber stellt. Wir können also erkennen, wie viele Treffer in jeder Klasse gefunden wurden und wohin sich die Alpha-Fehler bewegt haben. In normalisierter Form ist der prozentuale Trefferanteil erfasst. In unserem Modell wird somit eine erstaunlich präzise Aussage über die Testpersonen mit Untergewicht getroffen. Von 2114 Teilnehmern werden 1960 richtig eingeschätzt. Das ist eine Trefferquote von 93%. Dieser gute Wert verlangt eigentlich eine Folgeuntersuchung, die ergründet, ob es markante Einzelindikatoren wie Alter (Jugendliche), Ethnie (Asiaten) und Einkommen (extreme Armut) gibt. Dies lag aber nicht im Fokus dieser Arbeit. Die Vorhersage der Normalgewichtigen ist sehr ungenau - nur 32% werden richtig erfasst. Viel zu viele Personen wurden ins Untergewicht verschoben (wo dann ein großer Beta-Fehler vorliegt). Die Vorhersage der leicht Übergewichtigen ist mit 52 Prozent etwas genauer. Immerhin wird dort der Schwerpunkt richtig gesetzt. Das Hauptproblem liegt bei der Kategorie "Adipös": Die Vorhersage dazu hat sich komplett in den über- und normalgewichtigen Teil verschoben.

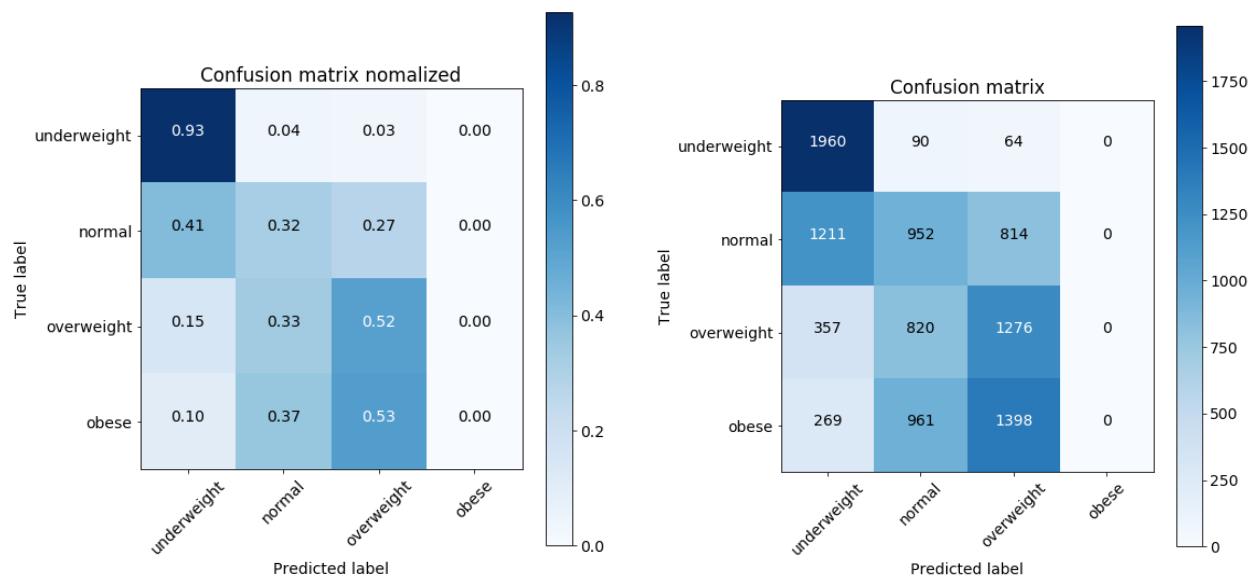


Abb. 4.5.3 Konfusionsmatrix: Gauss Naive Bayes

## 4.5.3 Fazit Klassifikator

Das Klassifikator-Modell ist sehr schwach. Bei guter Aussagekraft erwarten wir die Konfusionsmatrix eine deutlich ausgeprägte Hauptdiagonale. Diese ist jedoch nur schwach zu erkennen - die Verteilung der Schätzwerte ist zumindest nicht zufällig. Dass bedeutet vielleicht, dass die angedachten Zusammenhänge nicht ganz abzustreiten sind, durch den Naive-Bayes aber auch schlecht nachgewiesen werden können.

# 5 Fazit

Die Arbeit kommt zu Fragestellung, ob sozioökonomische Faktoren zur Prognose von Fettleibigkeit herangezogen werden können, zu einem ambivalenten Schluss:

Die Auswertung der geographischen Daten hat gezeigt, dass die Fettleibigkeit im Zentrum der USA und dort im Südosten besonders stark ist. Die betroffenen Staaten zählen zu den Ärmsten. Hingegen weißt die Gesamtheit der Menschen in der Nähe der Küstenregion einen deutlich geringeren prozentualen Anteil mit Adipositas auf. Des Weiteren ist die Bevölkerung einiger Metropolen - wo die Mietpreise und die Lebenserhaltung durchschnittlich teurer sind - nur wenig übergewichtig. Diese Beobachtungen stützen zunächst die These, dass Armut und Übergewicht stark korrelieren.

Der Versuch einer Verallgemeinerung dieser lokalen Phänomene mittels einer statistischen Gegenüberstellung von Einkommen und Fettleibigkeit schlägt jedoch fehl. Betrachtet man alle Countys der USA so findet man nur unzureichende Korrelationen zwischen diesen zwei Kernvariablen. Die Untersuchung der Forschungsfrage verlangt also eine differenziertere Betrachtung. Grundsätzlich ist festzustellen, dass kein einzelnes Merkmal eine ausschlaggebende Abhängigkeit produziert.

Das Zusammenführen mehrerer Variablen in das Prognosemodell der linearen Regression liefert gemischte Resultate: Die Kombination von Einzelfaktoren fördert tatsächlich die Tendenz zum Übergewicht. Die Prognose liegt nicht vollkommen daneben, jedoch ist die Aussage sehr grob und durch Ausreißer verfälscht. Dies kann unter anderem im Maßstab der Atlas-Daten begründet liegen. Insbesondere fehlt hier die statistische Verknüpfung der Variablen zueinander.

Dieser Schwachpunkt kann grundsätzlich durch den weiteren Datensatz, NHANES, ausgeräumt werden. Dieser führt die sozioökonomischen Faktoren auf Patientenebene zusammen. Eindeutige Beziehungen sind jedoch auch hier nicht zu finden: Durch die Korrelationsanalyse des Datensatzes ist zwar klar geworden, dass einige Merkmalsausprägungen mit der Fettleibigkeit schwach korrelieren, jedoch ergibt sich kein linearer Zusammenhang. Nur die obersten Einkommens- und Bildungsschichten zeigen einen deutlichen Rückgang des krankhaften Übergewichtes. Die markantesten Unterschiede sind bei der Ethnie zu finden. Afroamerikanische Frauen leiden statistisch öfter an Adipositas als der Durchschnitt. Diese Übereinstimmung ist aber nicht an das Einkommen geknüpft. Allerdings deckt sich die ethnische Verteilung mit den geografischen Untersuchungen aus Teil 1.

Der Versuch, die kategorischen Daten des NHANES für ein weiteres Prognosemodell (Naive-Bayes) zu verwenden, war nur bedingt erfolgreich. Wie schon bei der linearen Regression, sind die Ergebnisse nicht widersprüchlich oder zufällig. Die Trefferquote des Modells ist aber unzureichend.

Zusammenfassend lässt sich feststellen, dass ein Zusammenhang zwischen Fettleibigkeit und sozioökonomische Faktoren nicht grundsätzlich auszuschließen ist. Die Untersuchung liefert einige Hinweise für entsprechende Korrelationen. Eine Reduktion des Problems auf die in dieser Studie ausgewählten Faktoren ist jedoch unzureichend. Die Behauptung, dass Armut allein schon einen entscheidenden Einfluss auf die Tendenz zum Übergewicht hätte, ist nicht durch die Daten zu belegen. Kulturelle und individuelle Aspekte scheinen einen weit größeren Einfluss zu haben als zunächst vermutet. Adipositas ist in den USA ein Breitenphänomen, das sich durch alle Einkommens- und Bildungsschichten zieht. Die Verdichtung der Fettleibigkeit auf bestimmte Regionen der USA, weist auf eine Bündelung von strukturellen Faktoren hin, die erst zusammen genommen entscheidend werden. Diese Arbeit konnte die Komplexität dieser Zusammenhänge noch nicht entschlüsseln.

## 6.1 Quellen

- ❖ United States Department of Agriculture - Food Atlas:  
[https://www.ers.usda.gov/data-products/food-environment-atlas/  
data-access-and-documentation-downloads/#Current%20Version.](https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/#Current%20Version)
- ❖ List of U.S. states and territories by income:  
[https://en.wikipedia.org/wiki/List\\_of\\_U.S.\\_states\\_and\\_territories\\_by\\_income](https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_income)
- ❖ Centers for Disease Control and Prevention - National Health and Nutrition Examination Survey:  
<https://www.cdc.gov/nchs/nhanes/index.htm>
- ❖ NHANES Examination Trucks: <https://www.cdc.gov/nchs/participant/participant-about.html>

## 6.2 Abbildungen

Seite

<b>Abb. 2.2.1</b> Geoplot: US States >> Obesity Rate Adults US 2013	03
<b>Abb. 2.2.2</b> Geoplot: US Counties >> Obesity Rate Adults US 2013	04
<b>Abb. 2.2.3</b> Scatterplot: Population Density & Metropolitan Areas >> Obesity	05
<b>Abb. 2.3.1</b> Histogramm: US Adult Obesity Rate 2008	06
<b>Abb. 2.3.2</b> Histogramm: US Adult Obesity Rate 2013	07
<b>Abb. 2.3.3</b> Histogramm: US Adult Obesity Rate Change 2008-2013	07
<b>Abb. 2.3.4</b> Scatterplot: Obesity >> US Adults Diabetes 2013	08
<b>Abb. 2.3.5</b> Hexplot: Obesity >> US Adults Diabetes 2013	09
<b>Abb. 2.3.6</b> Scatterplot: US Fast Food Restaurants / 1000 Pop >> Obesity	10
<b>Abb. 2.3.7</b> Scatterplot: US Expenditure >> Fast Food Restaurant Number	11
<b>Abb. 2.3.8</b> Scatterplot: US Expenditure >> Obesity	11
<b>Abb. 2.3.9</b> Scatterplot: Low Store Access (no car) > Obesity	12
<b>Abb. 3.2.1</b> Heatmap: Korrelationen der Atlas-Daten	14
<b>Abb. 3.2.2</b> Heatmap: Signifikante Korrelationen mit Fettleibigkeit	15
<b>Abb. 3.2.3</b> Scatterplot: lineare Regression zur Fettleibigkeit	17
<b>Abb. 3.2.4</b> Scatterplot: Abweichung der prognostizierten Werte	17
<b>Abb. 3.2.5</b> Histogramm: Verteilung der Abweichungen bei der multivariablen Regression	18
<b>Abb. 4.1.1</b> NHANES Examination Trucks	20
<b>Abb. 4.2.1</b> Balkendiagramm: Prozentualer Anteil der höheren BMI-Klassen nach Einkommen	22
<b>Abb. 4.2.3</b> Scatterplot: BMI vs. Einkommen	24
<b>Abb. 4.2.4</b> Heatmap: Anzahl von Haushaltmitgliedern nach Haushaltseinkommen	25
<b>Abb. 4.2.4</b> Boxplot: Einkommen vs. Bildungsstufe	26
<b>Abb. 4.3.2</b> Violinplot: BMI vs. Bildungsstufe	27
<b>Abb. 4.4.1</b> Violinplot: BMI vs. Ethnische Gruppe	28
<b>Abb. 4.4.2</b> Heatmap: BMI-Median vs. Ethnische Gruppe und Einkommen	29
<b>Abb. 4.5.3</b> Konfusionsmatrix: Gauss Naive Bayes	31

## 6.3 Tabelleverzeichnis

<b>Tab. 2.2.1</b> Description: Metropolitan vs. Non-Metropolitan Areas	05
<b>Tab. 2.3.1</b> Description: US Adult Obesity Rate	06
<b>Tab. 4.2.1</b> Kontingenztabelle BMI vs. Haushaltseinkommen (absolute Häufigkeiten)	21
<b>Tab. 4.2.2</b> Resultate Chi2-Test: Einkommen >> Übergewicht	23
<b>Tab. 4.5.1</b> Datenselektion für Naive Bayes	30