Nutrition Changes in South Korea

Bachelor's Thesis

by

Darius Jonda

(544727)

Humboldt-Universität zu Berlin

School of Business and Economics

Ladislaus von Bortkiewicz Chair of Statistics

First Examiner: Prof. Dr. Wolfgang Härdle

Second Examiner: Prof. Dr. Dr. h. c. Dieter Kirschke

Thesis Supervisor: Dr. rer. nat. Sigbert Klinke

in partial fulfillment of the requirements

for the degree of

Bachelor of Science

Berlin, February 20, 2017

Abstract

Rapid diet changes can be the cause of negative health impacts on people. This might bring up new challenges for countries experiencing rapid economic growth and westernization, like South Korea. This paper uses data from the Korean National Health and Nutrition Examination Survey (KNHANES), to examine the nutrition intake by 23 food groups from 1998 until 2015 to reveal any dietary changes.

Evidence has been found using the principal component analysis, that the traditional Korean diet, which is considered to be high in white rice and vegetables, but low in fat and meat, is slowly diminishing. The western diet (high in meat, fats and processed foods) however was found in the major dietary patterns for both 1998 and 2015. The analysis of the 23 food groups showed a significant increase in meat, processed foods and alcohol consumption and a major decrease in white rice consumption, supporting the thesis of a potential nutrition transition in Korea.

Contents

List of Abbreviations

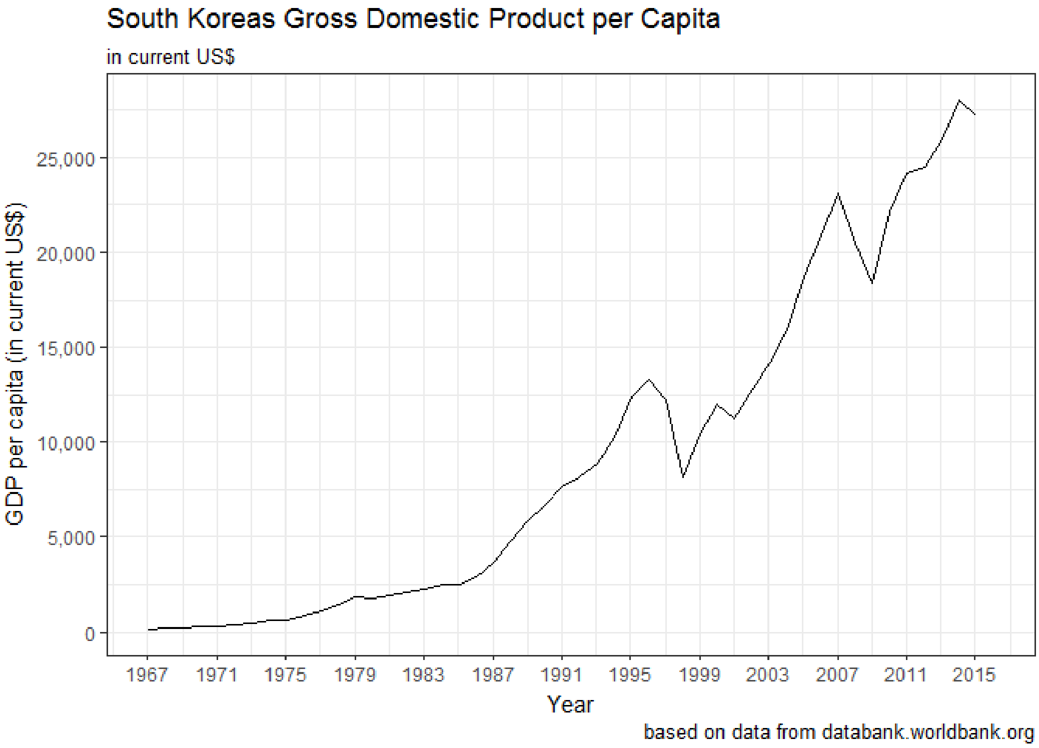
KNHANES	Korean National Health and Examination Survey
KCDC	The Korean Center for Disease Control and Prevention
MEC	Mobile Examination Center
PSU	Primary Sampling Unit
kcal	kilocalories
PCA	Principal Component Analysis
PCs	Principal Components
g/d	gram(s) per day
kcal/d	kilocalories per day
CA	Cluster Analysis
BMI	Body Mass Index
GDP	Gross Domestic Product
kg	kilogram(s)
m	meter(s)
g	grams(s)
WHO	World Health Organization
US$	US Dollar(s)

List of Figures

List of Tables

1 Introduction

The Republic of Korea (hereafter ”Korea”) is one of the so called ”Four Asian Tiger” countries[1], that experienced rapid technological development and industrialization while maintaining exceptionally high economic growth rates between the early 1960s and 1990s. Today Korea counts as one of the world leaders in technology manufacturing, with an economic success story, that has been seen as a role model for other developing countries (Lee et al. (2008)). The foundation of the economic success story took place before the 21 century, but as seen in figure 1 the economy continued to grow during the past 20 years up until today.


South Koreas Gross Domestic Product per Capita

in current US$

based on data from databank.worldbank.org

Figure 1: Koreas GDP per capita development from 1967 until 2015. (The World Bank (2017)) PlotGDPpCapita

While focusing almost entirely on the economic side, little has been known about the impact of these economic shifts on the dietary habits of Koreans. It is recognized that rapid changes in diet might be the cause of so-called diseases of civilization, including coronary heart disease, obesity, type 2 diabetes (Carrera-Bastos et al. (2011)). These

[1] the remaining three Tiger countries are Hong Kong, Singapore and Taiwan

might be aspects to be concerned about when observing quick changes in the traditional diet.

For this reason, the purpose of this thesis is to study the transition in daily food consumption, as well as the derivation of key dietary patterns in 1998 and 2015 of the Koreans population, based on data from the Korean National Health and Nutrition Examination Survey (KNHANES).

This thesis is organized as follows. Section 2 describes the underlying data in detail. Section 3 gives a brief introduction about the background of the procedures, that are being used during the analysis, following by section 4, where the results are being presented. Finally, Section 5 concludes the thesis.

2 Data

2.1 KNHANES Background

The data analyzed in this thesis is from the Korean National Health and Nutrition Examination Survey (KNHANES)[2], which is a representative cross-sectional health and examination survey, targeted at non-institutionalised Korean civilians aged 1 and older. The Korean Center for Disease Control and Prevention (KCDC) conducted the survey and includes approximately 10,000 individuals each year as a survey sample.

KNHANES is an ongoing survey in Korea and was first established in 1998, based on Article 16 of the National Health Promotion Act proclaimed in 1995. According to the available information provided by the KCDC, the study was conducted triennially beginning in 1998 (KNHANES I, II and III) and yearly since 2007 (KNHANES IV, V and VI). The first and second survey were conducted in November and December, the third survey in April and June and the following surveys from 2007 on were collected over a 12-month seasonally adjusted sample.

The survey consists of three main component surveys: a health interview, health examination and nutrition survey. (KNHANES Survey Contents (2016))

- The health interview collects information on a household and individual level. The household component consists of data about the general demographic profile of all members of the sampled household, including the income provided by an adult respondent aged 19 years and older.(Kweon et al. (2014)). The individual component is self administrated and aggregates information about general demographics, like education and occupation, quality of life, injury, mental condition, smoking and drinking habits, physical activity, oral health, weight control and general safety.

- The health examination survey addresses health related facts, like obesity, hypertension, various diseases like eye disease or chronic obstructive pulmonary disease, osteoarthritis and osteoporosis. Since it is a necessity to follow standardized procedures gathering these information, the health examination is performed by trained medical personnel using properly calibrated equipment.

[2] available for download from `https://knhanes.cdc.go.kr/knhanes/sub03/sub03_02_02.do`

- The nutrition survey is a follow up of the health interview and health examination survey. It is being performed by nutritionists at the homes of the study participants and aims to gather facts about the dietary behaviors and past 24 hour food intake of the Korean citizen, which includes information about eating home with the family, eating out, the intake of dietary supplements and food security. To assist recall of portion sizes and food details, especially for food that has not been cooked at home, food models, food shapes and two-dimensional models of different sizes of traditional pots or bowls were used.

Survey	Information	Methodology
Health Interview	Household survey (household income, number of householders etc.), individual survey (medical conditions, physical activity, mental health, education etc.),	Interview and self-reported at MEC (mobile examination center)
Health examination	Antrophometry, Blood pressure measurements, Muscular strength test, Blood test, Urine test, Oral health examination, Eye examination and more.	Examination at MEC (mobile examination center)
Nutrition survey	Dietary behaviour, food intake, food frequency, food security	Interview at participants' homes

Table 1: General overview of the KNHANES survey design.

2.2 Nutrition Survey Data Set

This thesis aims to gather insights about nutritional habits, therefore the nutrition survey is being used as the main source of data for most of the upcoming analysis. Each observation of the data set represents a food item, that has been consumed by one of the survey participants within the past 24 hours of the questionnaire. The variables give further information to each item, such as the ID of the consumer, the time of consumption, food codes, the weight / volume of the food, the dish name, as well macro and micro nutritional information.

2.3 Sampling Design

According to information regarding the sampling design of KNHANES, which are available on the official website (KNHANES Sampling Design (2016)), the survey is based on a multi-stage clustered probability sample of non-institutionalized Korean households. In the 2010 survey for example 192 primary sampling units (PSUs) were drawn out of a pool of around 200,000 geographically defined PSUs for the entire country. Each PSU averages approximately 60 household, from which 23 final target households were determined using a systematic sampling approach. Due to the inherent nature of this survey, it is required to use special analysis that are not being used in ordinary statistical procedures, like sample weights, stratification and clustering. (Kim et al. (2013))

3 Methods

This section is going to briefly introduce the statistical methods used for the later analysis. First the two-sample t-test, being a statistical hypothesis test to find out whether two sets of data are significantly different from each other. Second the principal component analysis, which helps to reduce large data sets into smaller components and third the cluster analysis, outlining certain characteristics in high-dimensional data by separating them into smaller clusters.

3.1 Two-Sample t-Test

The two-sample t-test is used to test the differences between two population means, aiming to find out whether two sets of data are significantly different from each other.

In order to apply the correct test, the variances of both samples need to be tested first using the f-test. Additionally, the samples must be independent and normally distributed (Wackerly et al. (2007).

- In case the variances are equal, the following equation will be used (Wackerly et al. (2007):

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - D_0}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{1}$$

  where $S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$ and $D_0 = \mu_1 - \mu_2$.

- In case the variances are unequal, the Welch's t-test will be applied (Ott and Longnecker (2008)):

$$T = \frac{\overline{y}_1 - \overline{y}_2 - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2}$$

  with s_1^2 and s_2^2 being the sample variances.

3.2 Principal Component Analysis

Principal component analysis (PCA) is a multivariate technique, that analyzes a data table, in which observations are described by several inter-correlated quantitative dependent variables (Abdi and Williams (2010)).

The idea of PCA is to explain a high-dimensional data set by a linear low-dimensional subspace, achieved by an orthogonal rotation of the coordinate system (Klinke et al. (2010)).

The PCA computes a subspace of principal components (PCs), which are linear combinations of the original variables. The first principal component requires to have the highest possible variance, the second component is then computed under the constraint of being orthogonal to the first component, again having the highest possible variance. This procedure is being repeated for the other components likewise. These new variables for the observations are called factor scores and can be interpreted as projections of the observations on the principal components (Abdi and Williams (2010)).

To find the components, they must be obtained from the single value decomposition of the data table X.

With $X = P\Delta Q^T$, the IxL matrix of factor scores (denoted F) is obtained as

$$F = P\Delta \tag{3}$$

with

- P: being the IxL matrix of of left singular vectors
- Q: as the JxL matrix of the right singular vectors
- Δ: being the diagonal matrix of singular values, $\Delta = \Lambda^{\frac{1}{2}}$, with Λ being the diagonal matrix of the eigenvalues of matrix XX^T and the matrix A^TA.

The variance of a column is defined as the sum of squared elements of this column and is calculated as

$$\gamma_j^2 = \sum_i^I x_{i,j}^2 \tag{4}$$

The sum of all the γ_j^2 is denoted as I (also called the total inertia or inertia of the data table) and equals the sum of squared singular values of the data table.

The coefficients of the linear combination used to compute the factors score are given in matrix Q. This matrix can also be interpreted as the projection matrix because multiplying X by Q results in the values of the projections of the observation on the PCs (Abdi and Williams (2010)).

$$F = P\Delta = P\Delta QQ^T = XQ \tag{5}$$

3.3 Cluster Analysis

Finding similarities in high-dimensional sets of data can often times be challenging, another set of procedures for bundling high-dimensional data sets into homogeneous groups is called cluster analysis. The goal of the cluster analysis is to find objects in a group, that are similar to one another and different to objects in other groups, the greater the difference between groups, the better or more distinct is the clustering (Tan et al. (2005)).

There are different variations of clusterings, the most commonly discussed distinctions among different types of clusterings is whether it is hierarchical or partitional. Hierarchical clusterings are characterized by the fact, that each cluster may have sub clusters, partitional clustering on the other hand divide the data only into non-overlapping subsets (Tan et al. (2005)).

This form of analysis is being used in a variety of science disciplines, some examples are:

- **Biology:** For finding groups of genes that have similar functions or creating a taxonomy (hierarchical classification) of living things.

- **Computer Science:** Search engines are constantly crawling the World Wide Web to gather every bit of information available. To combine this huge amount of data into small groups, clustering procedures are being used.

- **Medicine:** Since diseases may have different variations, clustering can help in categorizing certain symptoms to a special form of condition.

- **Business:** Businesses often use clustering algorithms to group their customers into a number of sub groups to improve their customer relationship. One common use case is targeted advertisement to each individual customer cluster.

3.3.1 Introducing the K-Means Algorithm

One procedure of applying a cluster analysis on a variety of data sets is the k-means algorithm. This algorithm is one of the most known and used clustering algorithm available, since it is considered to be fast and easy to implement (Tan et al. (2005)).

The first step is to define a number of clusters k in advance, the algorithm then randomly chooses k amount of cluster centers (centroids) and assigns each nearest data point to its cluster. After every observation has been assigned to one of the k clusters, the algorithm calibrates the new centroids and repeats the process of assigning the data points to one of the clusters. This process is being repeated, until the clusters won't show any significant change anymore. This procedure can be summarized using the following four steps:

1. Select k number of clusters.
2. Assign each object to its nearest centroid.
3. New calculation of each cluster centroid.
4. Repeat step 2-3 until the centroids do not change.

More formally speaking the k-means algorithm aims to minimize the intra-cluster variance or the squared error function as formalized in equation 6. The objective function C can be interpreted as the optimal cluster assignment, N being the number of cases in cluster k, and $\bar{x_k}$ representing the mean vector of cluster x (Tan et al. (2005)).

$$C = \min_{c} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \parallel x_i - \bar{x}_k \parallel^2 \tag{6}$$

3.3.2 Determining the Number of Clusters in K-Means

Even though the implementation of the k-means algorithm is fairly simple, there is obviously one major problem: how to decide on the right number of k? Unfortunately, there is no simple answer to this question, but there are ways to interpret and validate the consistency withing clusters of data.

- **The elbow critertion:** This criterion looks at the percentage of the explained variance as a function of the number of clusters. This method is based on the idea, that adding another cluster doesn't add much better modelling of the data (Bholowalia and Kumar (2014)).

- **Silhouette:** The silhouette measures the similarity of a data element in its own cluster to elements in another cluster. The silhouette index of element x^i of cluster S^j is defined as follows (Kaufman and Rousseeuw (2009))

  $$q_i = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, -1 \leq q_i \leq 1, \tag{7}$$

  with $a(i)$ being the average similarity between x^i to the other objects in cluster S^j and $b(i)$ is the minimum average similarity between object x_i and the rest of the objects in all the clusters (Rousseeuw (1987)).

  The silhouette index ranges from -1 to 1, with a higher value being a better indicator of correct cluster assignment.

  - $q_i \approx -1$: assignment to neighbor cluster is better.
  - $q_i \approx 0$: assignment neutral, i lies between two clusters.
  - $q_i \approx +1$: good assignment.

  Running the silhouette measure on multiple clusterings, each with different numbers of centroids, returns a measure for the optimal assignment.

4 Results

4.1 Background

Since the Korean National Health and Nutrition Examination Survey (KNHANES) is based on a multi-stage clustered probability sample, sample weights need to be incorporated to the sample (Kim et al. (2013)). The analysis has been performed using R version 3.3.1 and source codes are available in the appendix of this document. The data set has been filtered by outliers that reported back implausibilities in their nutrient intake (consuming less than 500 kcal/day and more than 5000 kcal/day).

4.1.1 Demographics

Table 2 shows a comparison of the demographic characteristics between the KNHANES from 1998 and 2015. The fact that the population of Korea is aging rapidly (Moon (2015)) is clearly visible when comparing the numbers of the two sets of data. But also the number of people with High school diplomas showed a noticeable increase.

	Demographics		
Group		KNHANES 1998	KNHANES 2015
Sample size(n)	Number of individuals	10,400	6,628
	Number of households	3,475	2,910
Gender (in %)	Female	51.1	51.2
	Male	48.9	48.8
Age groups (in %)	<2 years	2.9	1.0
	2-6 years	7.8	5.2
	7-12 years	9.0	6.6
	13-18 years	10.1	6.5
	19-39 years	34.8	22.2
	40-59 years	24.1	30.4
	60+ years	11.3	28.1
Education (19y+ in %)	<High school diploma	49.8	42.7
	≥High school diploma	50.2	57.3

Table 2: Overview of the 1998 and 2015 KNHANES demographics. QDASampleWeights

4.1.2 Top food groups

The food was initially grouped into 18 food groups, consisting of grains, potatoes and starch, sugars, legumes, seeds and nuts, vegetables, mushrooms, fruits, meats, eggs, fish, seaweeds, milk and dairy products, fats and oils, beverages, seasonings, processed foods and others. Grains have been separated into white rice, breads and other grains and kimchi has been extracted from the vegetable food group, due to the distinct characteristics of these foods. Lastly alcoholic beverages, as well as coffee have been separated from the beverages group resulting in a total number of 23 food groups, which are going to be examined in later analysis.

Table 3 presents a brief overview of the kcal intake per capita per day (kcal/capita/d) for each individual food group in 1998 and 2015.

	Consumption in kcal/capita/day	
Food Group	KNHANES 1998	KNHANES 2015
White Rice	785.3	550.2
Grains	264.0	324.3
Meats	144.7	185.5
Fruits	94.7	104.3
Fish	83.0	60.2
Fat and Oils	57.6	60.4
Milk and Dairy Products	50.6	40.2
Potatoes and Starch	44.9	44.0
Other Vegetables	42.9	59.9
Alcohol	37.1	42.5
Seasonings	36.0	56.8
Legumes	35.7	44.0
Eggs	33.4	41.2
Kimchi	26.9	25.4
Sugars	25.5	32.4
Beverages	17.2	20.1
Bread	10.4	13.2
Seaweeds	8.6	2.3
Seeds and Nuts	8.0	22.9
Coffee	5.9	5.8
Processed Foods	4.0	20.1
Mushrooms	1.5	1.9
Others	0.4	0.5

Table 3: Comparison of energy intake per capita per day by 23 selected food groups. DATopFoodGroups

As shown in table 3, white rice is the food group that contributes the most amount of energy in a daily average korean diet, even though consumption went down significantly from 758 kcal/capita/day in 1998 to 550 kcal/capita/day in 2015. However, the negative trend of rice consumption is not a phenomenon from the recent 20 years, it's consumption has been estimated to peak at 1300 kcal/capita/day in 1970, following a steady decline up until now, causing the Korean traditional diet, which is described as being high in vegetables and white rice, as well as low in fat, to be less present (Choi et al. (2016); Lee et al. (2002)).

Some factors contributing to this decline are described by Choi et al. (2016):

- **Urbanization:** The World Bank estimates around 80 percent of the Korean population to life in urban areas, where a wide variety of foods are more available and advertised by the restaurants.

- **Eating-Out:** As stated by Choi et al. (2016), expenditures on eating away from home rose from 2.1 percent in the 1960s to 45.6 percent in 2006, which disrupts the old pattern of the traditional rice-centered cooking.

- **Low priced alternatives:** Low priced foods that are based on imported wheat, like instant noodles or breads are another alternative to rice. Especially post-World War II these types of food entered the Korean diet with the consumption level being static from 1980 until now.

- **High rice prices:** The Producer rice price trebled since 1980 according to data from the Korean Statistical Information Service (KOSIS) database, which has a direct impact on rice consumption.

Another indicator for the drift away from the traditional Korean diet is the minor decline in kimchi consumption from 26.9 kcal/capita/day in 1998 to 25.4 kcal/capita/day in 2015. During the same time period the meat consumption increased from 144 kcal/capita/day to 185 kcal/capita/day, as well as processed foods increasing five-fold from 4 kcal/capita/day to 20 kcal/capita/day. This might be another indicator for a newly developed dietary habit, that might be caused by the recent economic boom and the resulting cultural westernization.

Variable	1998	2015	p-value[a]
Carbohydrates	1249	1219	< 0.01
Proteins	285	276	< 0.01
Fats	352	403	< 0.001
Total kcal	1904	1989	< 0.001

Table 4: Changes in macronutrients and energy intake from 1998 to 2015. SAMacronutrientChanges

[a] the p-value was computed using Welch's Two Sample t-test for unequal variances, and the Student t-test for equal variances.

4.1.3 Macronutrient intake

Since the traditional Korean diet has been described as being high in vegetables and white rice and low in meat and fats, the daily intake grouped by the three macronutrients (carbohydrates, proteins, fats) is being examined further, to reveal any potential shifts in this regard.

The stacked area chart in Figure 2 shows the amount each macronutrient contributes to the total daily intake in kcal from 1998 to 2015. Apart from a visible drop in total kcal intake from 2005 to 2007, including the consumption of each macronutrient, the graphical analysis does not reveal much detail.

Comparing the means using the t-test[3] however shows a significant increase in total kcal intake from 1904 kcal/d in 1998 to 1989 kcal/d in 2015 (p-value < 0.001) and fat from 352 kcal/d in 1998 up to 403 kcal/d in 2015 (p-value < 0.001). During the same period the intake of proteins and carbohydrates went slightly down from 285 kcal/d to 276 kcal/d for proteins (p-value < 0.01) and from 1249 kcal/d to 1219 kcal/d (p-value < 0.01) for carbohydrates respectively (see table 4).

4.1.4 Obesity

An increase in fat intake is often associated with an increase in obesity (Golay and Bobbioni (1997)). To compare the insights from the previous analysis regarding macronu-

[3] Welch's Two Sample t-test for unequal variances, Student t-test for equal variances

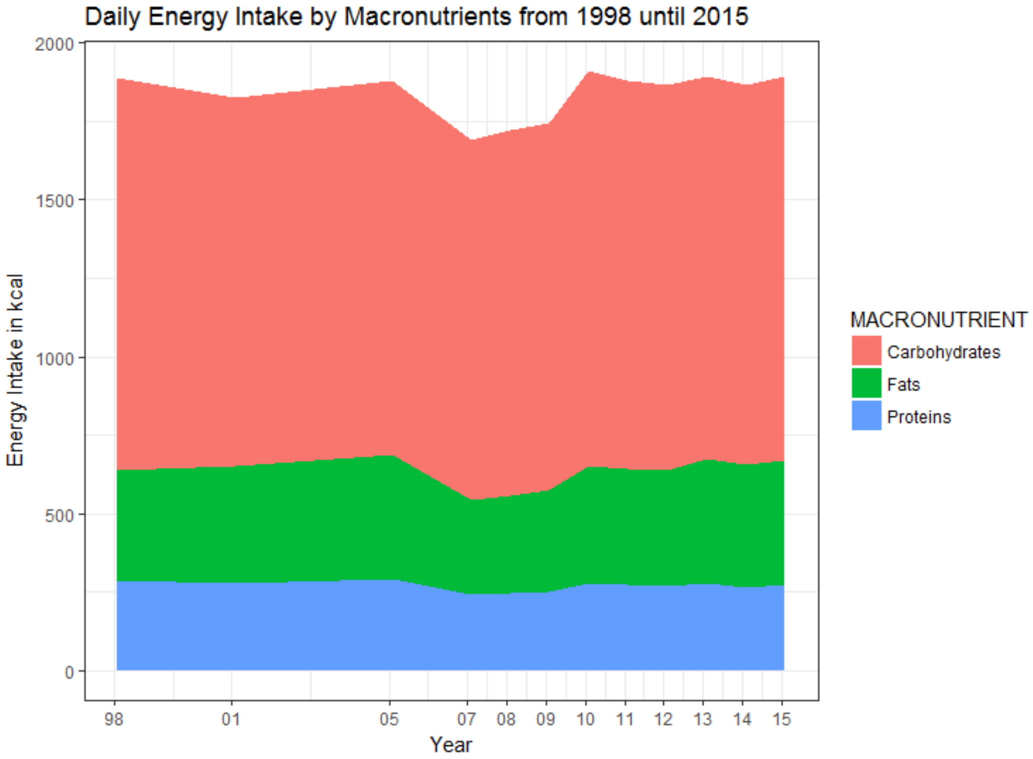


Figure 2: Macronutrient Intake in kcal from 1998 to 2015 PlotMacronutrientIntake

trients with the prevelance of obesity, the body mass index is going to be used as an indicator. The Body Mass Index (BMI) is defined as the weight in kilograms divided by the square of the height in metres ($\frac{kg}{m^2}$). The resulting values are age-independant and the same for both sexes. The World Health Organization (WHO) classified the Body Mass Index (BMI) using the groups listed in table 5 (WHO (2017)).

Applying this categorization onto the KNHANES data set, resulted in a significant increase of the average BMI from 22.61 kg/m^2 in 1998 to 22.97 kg/m^2 in 2015 (p-value < 0.001[4]). Also the relative number of overweight and obese people went up (from 21.1 percent in 1998 to 25.3 percent in 2015 and from 2.2 percent to 4.5 percent respectively).

4.2 Dietary patterns

To go into more detail about the food and nutrient intake, the cluster analysis and principal component analysis are being applied, aiming to identify key dietary patterns.

[4] p-value was computed using the Welch Two Sample t-test

Classification	BMI (kg/m^2)	KNHANES 1998 (in %)	KNHANES 2015 (in %)
Underweight	< 18.5	10.3	13.8
Normal range	18.5-24.9	66.5	56.3
Overweight	≥ 25	21.1	25.3
Obese	≥ 30	2.2	4.5

Table 5: A comparison by BMI classifications for the 1998 and 2015 KNHANES data sets. QDABmi

These two procedures specifically have been selected, since they have shown stable results in similar studies finding dietary patterns (Sauvageot et al. (2017)). Before deciding on a proper measure, five options have been examined: the weight consumed of each food group per day (g/d), kilocalorie intake per day (kcal/d), the relative daily weight and energy intake respectively, as well as a binary measure (food group has been consumed or not consumed).

Since the stability of the tests didn't show major differences and the reasoning behind the relative kcal intake seems superior in emphasizing the nutrition habits, it has been concluded to look at the percentage of calories consumed each day per food group.

Before analyzing the selected measure on the 23 previously defined food groups for each individual, outliers with implausible dietary patterns, consuming more than 5000 kcal a day or less than 500 kcal a day, have been removed from the data set.

Both the cluster analysis and principal component analysis intend to reduce the dimensionality of a data set with myriad variables, but are different techniques in achieving the goal. For that reason, the results of both procedures are being presented first separately for each year and evaluated afterwards.

4.2.1 Principal Component Analysis

The principal component analysis is being used to find out, if food groups can be categorized into certain dietary patterns.

The daily food intake, grouped by the 23 earlier defined food groups, has been calculated for each individual and is being used as input data for the model. The variables have been standardized before applying the principal component analysis and 170 of 10,400 subjects from the 1998 survey, as well as 149 of 6,628 from the 2015 survey have been excluded because of an implausible dietary intake (consumption of less than 500 kcal a day or more than 5000 kcal a day).

The scree plot in figure 3 shows the variance explained by each principal component. According to the methodology behind the PCA in section 3.2, the first principal components are describing a relative high amount of the total variance of the data table, compared to the following PCs. However, looking at the numbers more closely in table 7 reveals, that the cumulative proportion is rather low, with only around 21.1% of the total variance explained by the principal components one to three.

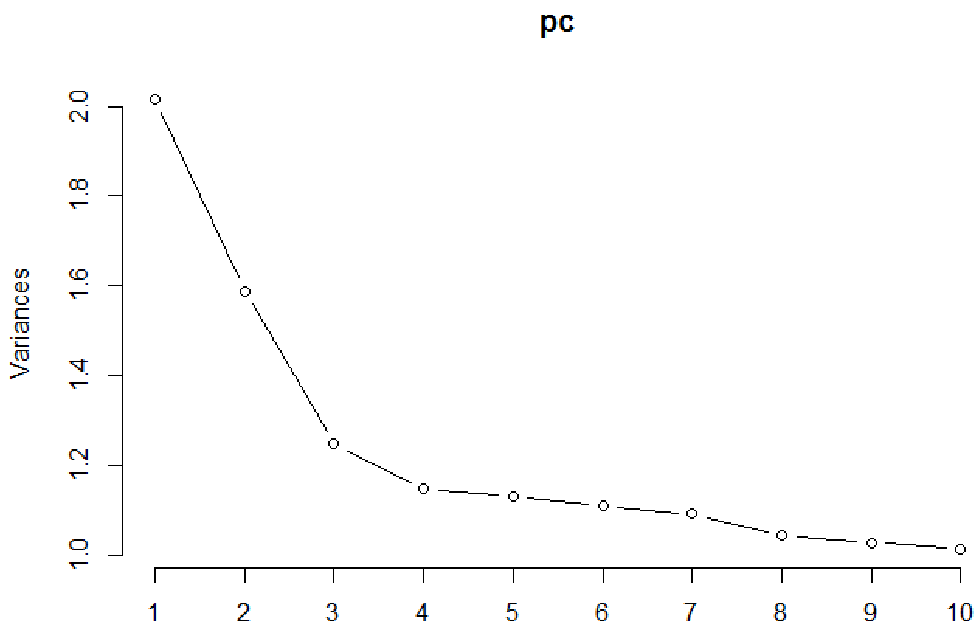


Figure 3: The scree plot for the 1998 data set shows the variance explained by each principal component. SAPrincipalComponentAnalysis

To focus more on the key dietary patterns, the first three principal components are being used for further analysis, taking into account that the amount of total variance being explained using the first three PCs is rather limited.

Table 8 summarizes the loadings of the top three principal components, which present the following characteristics:

PC1 shows a relatively high consumption of white rice, kimchi, seaweeds and legumes, with a negative loading on meats, milk and dairy products and processed foods, indicating that this might be similar to the previously defined traditional Korean diet.

PC2 has positive loadings on alcohol, coffee, fat and oils, meats, fish and sugars and the lowest loading on white rice among the three PCs. This principal component might indicate a western diet.

PC3 Is relatively high in sugars, eggs, fish, fat and oils, coffee and grains and relatively low in alcohol and meats compared to the other two principal components.

Applying the same procedure on the 2015 KNHANES data set results in the following three principal components (see table 9):

PC1 has high loadings in fats and oils, meats, grains, seasonings, coffee, sugars and processed foods with negative loadings on white rice and kimchi indicating this to be closer to a western diet.

PC2 shows high loadings on vegetables, kimchi and white rice and negative loadings on milk and dairy products, as well as eggs, bread, grains and processed foods. Even though white rice, vegetables and kimchi are loaded relatively high, the high loadings on meats, as well as fats and oils keep this principal component from being interpreted as the traditional Korean diet.

PC3 shows high loadings on fruits, seeds and nuts, vegetables, legumes, fish and seaweeds, as well as negative loadings on meats, processed foods, alcohol and white rice. This principal component indicates a vegetarian-like diet, low in animal products and high in plant based foods.

Comparing both of the three-principal component solutions for the 1998 and 2015 KNHANES data sets indicates a subtle disappearance of the traditional Korean diet, which showed similar characteristics as the first principal component of the 1998 data set, but could not be categorized into any of the three PCs in the latest data from 2015. In addition to that the first principal component of the 2015 data showed characteristics similar to the western diet (high in meat, fat and oils and low in white rice according to Carrera-Bastos et al. (2011)) and was also similar to the second PC in the 1998 data. These findings might support the idea of a potential nutrition transition in Korea, however it should not be forgotten that both three-principal component solutions only explain a limited amount of the total variance in the data.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	1.4200	1.2603	1.1166	1.0715	1.0632	1.0530	1.0447	1.0217	1.0141	1.0066	1.0043	0.9948	0.9903
Proportion of Variance	0.0877	0.0691	0.0542	0.0499	0.0491	0.0482	0.0474	0.0454	0.0447	0.0440	0.0438	0.0430	0.0426
Cumulative Proportion	0.0877	0.1567	0.2109	0.2609	0.3100	0.3582	0.4057	0.4510	0.4958	0.5398	0.5837	0.6267	0.6693

Table 6: 1998 KNHANES: Summary of the first 13 PCs of the PCA. SAPrincipalComponentAnalysis

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	1.3958	1.2456	1.1573	1.0986	1.0770	1.0536	1.0475	1.0361	1.0172	1.0117	1.0053	0.9951	0.9870
Proportion of Variance	0.0847	0.0675	0.0582	0.0525	0.0504	0.0483	0.0477	0.0467	0.0450	0.0445	0.0439	0.0430	0.0423
Cumulative Proportion	0.0847	0.1522	0.2104	0.2629	0.3133	0.3616	0.4093	0.4560	0.5009	0.5454	0.5894	0.6324	0.6748

Table 7: 2015 KNHANES: Summary of the first 13 PCs of the PCA. SAPrincipalComponentAnalysis

	PC1	PC2	PC3
Alcohol	-0.02	0.12	-0.29
Beverages	-0.13	0.00	-0.14
Bread	-0.07	-0.08	-0.06
Coffee	-0.09	0.22	0.26
Eggs	-0.20	-0.10	0.29
Fat and Oils	-0.32	0.33	0.31
Fish	-0.00	0.14	0.31
Fruits	-0.10	-0.22	-0.00
Grains	-0.42	-0.23	0.02
Kimchi	0.30	0.11	0.03
Legumes	0.09	0.00	0.01
Meats	-0.18	0.19	-0.53
Milk and Dairy Products	-0.19	-0.36	0.06
Mushrooms	-0.06	0.09	-0.13
Other Vegetables	-0.05	0.45	-0.07
Others	-0.05	-0.04	-0.01
Potatoes and Starch	-0.06	-0.05	-0.22
Processed Foods	-0.04	-0.06	0.06
Seasonings	-0.04	0.43	-0.18
Seeds and Nuts	-0.07	0.07	-0.02
Seeweads	0.06	-0.07	0.17
Sugars	-0.23	0.31	0.32
White Rice	0.63	0.06	0.15

Table 8: 1998 KNHANES: Loadings of the first three PCs. QSAPrincipalComponentAnalysis

	PC1	PC2	PC3
Alcohol	0.09	0.10	-0.16
Beverages	0.07	-0.04	-0.06
Bread	0.09	-0.09	-0.11
Coffee	0.16	0.17	-0.13
Eggs	0.06	-0.14	-0.03
Fat and Oils	0.37	0.28	-0.02
Fish	0.01	0.34	0.25
Fruits	-0.03	-0.16	0.37
Grains	0.26	-0.40	0.30
Kimchi	-0.35	0.09	-0.03
Legumes	-0.21	-0.02	0.27
Meats	0.26	0.13	-0.41
Milk and Dairy Products	0.17	-0.28	-0.02
Mushrooms	0.07	0.12	0.12
Other Vegetables	-0.01	0.41	0.32
Others	-0.00	0.00	0.04
Potatoes and Starch	0.06	-0.12	0.05
Processed Foods	0.13	-0.11	-0.15
Seasonings	0.24	0.43	0.13
Seeds and Nuts	0.01	0.02	0.36
Seeweads	-0.11	0.07	0.19
Sugars	0.22	0.14	-0.08
White Rice	-0.59	0.16	-0.28

Table 9: 2015 KNHANES: Loadings of the first three PCs. QSAPrincipalComponentAnalysis

4.2.2 Cluster Analysis using K-Means

The cluster analysis offers a useful set of procedures in identifying key patterns in various kinds of data (Sauvageot et al. (2017)), therefore the k-means algorithm is going to be applied to find out, how the food groups can be categorized into clusters, based on the nutrient intake documented in the surveys.

Analogous to the principal component analysis, the cluster analysis is going to use the same input data table measuring the relative daily kcal intake based on the 23 food groups, filtered by the individuals consuming more than 5000 kcal a day or less than 500 kcal a day.

As mentioned in Section 3.3.1, the first step in applying the k-means algorithm is to determine the appropriate number of clusters k. Two solutions to this problem were presented: the elbow criterion and the silhouette measure.

The elbow criterion in figure 4 shows no noticeable bend, indicating that there might be no clear solution with a low number of clusters minimizing the within-cluster variance.

Applying the silhouette measure (see figure 5) reveals a relatively high silhouette index for two to three clusters, with a value of around 0.25, which indicates a decent assignment. However, a good assignment would ideally return a silhouette index of 0.5 and more.

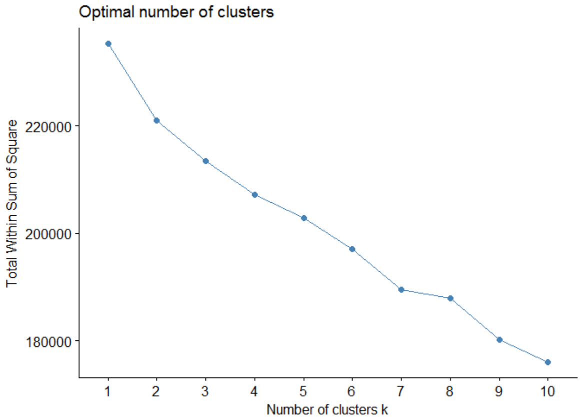


Figure 4: Elbow Criterion. SAClusterAnalysis

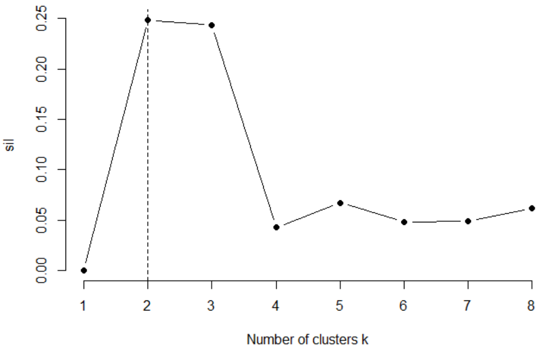


Figure 5: Silhouette Index. SAClusterAnalysis

Again it has been decided to proceed further analysis with a three cluster solution, for the following reasons: First to have more comparable results to the principal component analysis, which also resulted in three principal components. Second due to the silhouette measure indicating a two to three cluster solution as the optimal amount. And third, because it has been found in similar studies, that the k-means clustering solution with three clusters was chosen to be the most stable one in analyzing dietary patterns (Sauvageot et al. (2017)).

Applying the k-means analysis onto the 1998 KNHANES data set using a three cluster solution, resulted in dietary patterns as displayed in table 10, showing the following characteristics:

- 3,663 subjects were assigned to cluster 1, with 68.5 percent of the daily energy intake solely from white rice, the group is also relatively high in kimchi and relatively low in meats, fats, eggs and milk and dairy products compared to the other two groups.

- Cluster two represents around 4,417 people from our sample, which is the biggest group of all three clusters with a more spread out relative kcal intake across the 23 food groups. This group is relatively high in meats (10.1 percent of daily kcal), fats and oils (3.6 percent of daily kcal), as well as alcohol (2.4 percent of daily kcal) and milk and dairy products (4.3 percent of daily kcal).

- The third cluster incorporates 2,150 people and includes the least amount of subjects from all three groups. The majority of energy is taken from grains with 38.4 percent of daily kcal, white rice adds another 21.8 percent of daily kcal and meat consumption lies between cluster 1 and 2 with 5.7 percent of daily kcal.

Among the three described clusters, the first one comes close to the earlier defined "traditional Korean diet", being relatively high in white rice, kimchi, vegetables and low in meat. The second cluster reminds of the western diet, with a relative high intake of meats, fat and oils, as well as milk and dairy products consumption.

Looking at the three cluster solution for the 2015 survey data (see table 11) draws the following picture:

- Cluster 1 is the largest group with 1,870 subjects having a high white rice consumption with 54.9 percent of the daily energy intake, the highest amount of kimchi and vegetable consumption across all three groups and the lowest amount in processed foods, meats and alcohol, similar to the "traditional Korean diet".
- Cluster 2 is the smallest group of all three with 1,124 individuals and gets the majority of the daily kcal intake from grains. White rice shows the lowest consumption level among the three clusters.
- Cluster 3 is the second largest group with 1,815 individuals and is relatively high in meats with 15.8 percent of daily kcal, fats and oils with 4.1 percent of daily kcal and alcohol with 3.2 percent of daily kcal. Further this group has the highest intake of processed foods, coffee, alcohol, sugars, fruits and bread, which are similar characteristics to the "modern Korean diet".

The k-means cluster analysis shows, that both the traditional Korean diet, as well as the western diet could be identified in one of the three clusters for each year.

4.2.3 Comparing the Principal Component Analysis and Cluster Analysis

In summary, both the principal component analysis, as well as the k-means cluster analysis were able to identify the western diet for each 1998 and 2015, however the traditional Korean diet could not be categorized according to the results of the 2015 principal component analysis, indicating a slight disappearance. These differences might be due to the fact that neither the principal component analysis, nor the cluster analysis were able to produce results with high stability.

1998 KNHANES: Dietary Patterns by Energy Intake			
Food Group	C1	C2	C3
White Rice	68.5%	37.9%	21.8%
Grains	4.0%	10.3%	38.4%
Fish	3.6%	5.5%	3.6%
Meats	3.5%	10.1%	5.7%
Fruits	3.2%	7.0%	5.4%
Other Vegetables	2.2%	2.5%	2.3%
Fat and Oils	2.1%	3.6%	3.4%
Kimchi	2.0%	1.4%	1.2%
Legumes	1.9%	2.3%	1.7%
Seasonings	1.8%	2.1%	1.9%
Milk and Dairy Products	1.3%	4.3%	4.2%
Eggs	1.2%	2.1%	2.4%
Potatoes and Starch	1.2%	3.3%	2.0%
Sugars	1.0%	1.6%	1.5%
Alcohol	0.7%	2.4%	1.1%
Seaweeds	0.6%	0.5%	0.4%
Beverages	0.4%	1.1%	1.2%
Seeds and Nuts	0.3%	0.5%	0.4%
Coffee	0.2%	0.4%	0.3%
Processed Foods	0.1%	0.3%	0.3%
Bread	0.1%	0.8%	0.5%
Mushrooms	0.1%	0.1%	0.1%
Others	0.0%	0.0%	0.0%

Table 10: Dietary Patterns for the 1998 survey data derived using the k-means cluster analysis with 3 predefined centroids. SAClusterAnalysis

2015 KNHANES: Dietary Patterns by Energy Intake			
Food Group	C1	C2	C3
White Rice	54.9%	17.9%	23.6%
Grains	9.0%	41.7%	12.0%
Meats	5.3%	6.1%	15.8%
Fruits	4.6%	5.1%	7.7%
Other Vegetables	3.5%	3.1%	3.4%
Fish	3.2%	2.4%	3.7%
Legumes	2.9%	2.2%	2.3%
Seasonings	2.3%	2.6%	3.8%
Fat and Oils	2.2%	3.4%	4.1%
Eggs	2.2%	2.3%	2.8%
Kimchi	2.1%	1.3%	1.3%
Potatoes and Starch	1.6%	2.0%	3.6%
Milk and Dairy Products	1.4%	2.5%	3.5%
Sugars	1.3%	1.7%	2.3%
Beverages	0.9%	1.3%	1.4%
Seeds and Nuts	0.9%	1.0%	1.4%
Processed Foods	0.5%	1.2%	1.7%
Alcohol	0.5%	0.8%	3.2%
Bread	0.3%	0.7%	1.4%
Coffee	0.3%	0.3%	0.5%
Seaweeds	0.2%	0.1%	0.1%
Mushrooms	0.1%	0.1%	0.1%
Others	0.0%	0.0%	0.0%

Table 11: Dietary Patterns for the 2015 survey data derived using the k-means cluster analysis with 3 predefined centroids. SAClusterAnalysis

5 Conclusions

The nutrition intake by food groups and macronutrients has been examined for 1998 and 2015, finding evidence in an increase in meat consumption and an overall increase in fat as well as total caloric intake, resulting in higher obesity rates. White rice is still staple food in Korea, but with a significant decline during our examined time period. Three dietary patterns have been derived using the principal component analysis and the k-means cluster analysis. Whereas the k-means cluster analysis was able to identify characteristics of the traditional Korean diet (high in white rice, vegetables and low in meat) in both the 1998 and 2015 data, the principal component analysis was not able to do so for the 2015 data, indicating a slight disappearance. The western diet (high in meats and processed foods) was found in both procedures for 1998 and 2015.

The derivation of dietary patterns using both the principal component analysis and k-means cluster analysis however did not show high stability and the selection of the three dietary patterns was based on assumptions, which should be taken into account, while evaluating the results.

Future work might consider a more distinct variable selection or alternative statistical procedures in deriving dietary patterns for upcoming KNHANES data sets.

References

Abdi, H. and L. J. Williams (2010): "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, 2, 433–459.

Bholowalia, P. and A. Kumar (2014): "EBK-means: A clustering technique based on elbow method and k-means in WSN," *International Journal of Computer Applications*, 105.

Carrera-Bastos, P., M. Fontes-Villalba, J. H. O?Keefe, S. Lindeberg, and L. Cordain (2011): "The western diet and lifestyle and diseases of civilization," *Res Rep Clin Cardiol*, 2, 15–35.

Choi, S., J. Dyck, and N. Childs (2016): "The Rice Market in South Korea," .

Golay, A. and E. Bobbioni (1997): "The role of dietary fat in obesity." *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity*, 21, S2–11.

Kaufman, L. and P. J. Rousseeuw (2009): *Finding groups in data: an introduction to cluster analysis*, vol. 344, John Wiley & Sons.

Kim, Y., S. Park, N.-S. Kim, and B.-K. Lee (2013): "Inappropriate Survey Design Analysis of the Korean National Health and Nutrition Examination Survey May Produce Biased Results," *J Prev Med Public Health*, 46, 96–104, 23573374[pmid].

Klinke, S., A. Mihoci, and W. Härdle (2010): "Exploratory factor analysis in MPlus, R and SPSS," *Invited paper ICOTS8 of the International Association of Statistical Education.*

KNHANES Sampling Design (2016): "KNHANES Sampling Design," Information retrieved from the official KNHANES website, `https://knhanes.cdc.go.kr/knhanes/eng/sub02/sub02_01.do#s6_01`.

KNHANES Survey Contents (2016): "KNHANES Survey Contents," Information retrieved from the official KNHANES website, `https://knhanes.cdc.go.kr/knhanes/eng/sub02/sub02_03.do#s8_02`.

Kweon, S., Y. Kim, M.-j. Jang, Y. Kim, K. Kim, S. Choi, C. Chun, Y.-H. Khang, and K. Oh (2014): "Data resource profile: the Korea national health and nutrition examination survey (KNHANES)," *International journal of epidemiology*, 43, 69–77.

Lee, J., P. LaPlaca, and F. Rassekh (2008): "Korean economic growth and marketing practice progress: A role model for economic growth of developing countries," *Industrial Marketing Management*, 37, 753–757.

Lee, M.-J., B. M. Popkin, and S. Kim (2002): "The unique aspects of the nutrition transition in South Korea: the retention of healthful elements in their traditional diet," *Public health nutrition*, 5, 197–203.

Moon, K. H. (2015): "South Korea?s Demographic Changes and their Political Impact," *East Asia Policy Paper*, 6, 1–24.

Ott, R. L. and M. T. Longnecker (2008): *An introduction to statistical methods and data analysis*, Cengage Learning.

Rousseeuw, P. J. (1987): "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, 20, 53 – 65.

Sauvageot, N., A. Schritz, S. Leite, A. Alkerwi, S. Stranges, F. Zannad, S. Streel, A. Hoge, A.-F. Donneau, A. Albert, et al. (2017): "Stability-based validation of dietary patterns obtained by cluster analysis," *Nutrition Journal*, 16, 4.

Tan, P.-N., M. Steinbach, and V. Kumar (2005): *Introduction to Data Mining, (First Edition)*, Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

The World Bank (2017): "GDP per Capita for South Korea from 1967-2015," Data retrieved from the World Bank database, http://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=KR.

Wackerly, D., W. Mendenhall, and R. Scheaffer (2007): *Mathematical statistics with applications*, Nelson Education.

WHO (2017): "WHO BMI classification," Information retrieved from the WHO website, http://apps.who.int/bmi/index.jsp?introPage=intro_3.html.

A Source Code

```r
########################
## READ IN LIBRARIES ###
########################
library(foreign)
library(dplyr)
library(tidyr)
library(psych)
library(scales)
library(ggplot2)
library(survey)
library(factoextra)
library(haven)
library(stringr)
library(cluster)
library(survey)

########################
## FUNCTIONS ###########
########################
FoodGroupAdd <- function(df = NULL, year = NULL, agefilter = NULL) {
  df_fg <- upper(df) %>%
    mutate(FOOD_GROUP = as.numeric(substr(N_FCODE, 1, 2)))

  if(!is.null(agefilter)) {
    df_fg <- df_fg %>%
      filter(AGE > agefilter)
  }

  df_fg_wname <- merge(df_fg, foodgroup_db,
                       by.x = "FOOD_GROUP",
                       by.y = "ID") %>%
    mutate(NAME = as.character(NAME))


  if (year == 98) {
    alc <- as.factor(c(15031:15062))
    kimchi <- paste0("0", c(6045:6057))
    whiterice <- paste0("0", c(1157:1167))
    coffee <- as.character(c(15012:15018))
    bread <- paste0("0", c(1049:1071))
    df_fg_wname$NAME[trim(df_fg_wname$N_FCODE) %in% alc] <- "Alcohol"
    df_fg_wname$NAME[trim(df_fg_wname$N_FCODE) %in% kimchi] <- "Kimchi"
    df_fg_wname$NAME[trim(df_fg_wname$N_FCODE) %in% whiterice] <- "White␣Rice"
    df_fg_wname$NAME[trim(df_fg_wname$N_FCODE) %in% coffee] <- "Coffee"
    df_fg_wname$NAME[trim(df_fg_wname$N_FCODE) %in% bread] <- "Bread"
  } else if (year == 14) {
    alc <- as.character(c(15026:15060))
    kimchi <- paste0("0", c(6057:6070))
    whiterice <- paste0("0", c(1173:1182))
    coffee <- as.character(c(15083:15088))
    bread <- paste0("0", c(1053:1076))
    df_fg_wname$NAME[trim(df_fg_wname$N_FCODE) %in% alc] <- "Alcohol"
    df_fg_wname$NAME[trim(df_fg_wname$N_FCODE) %in% kimchi] <- "Kimchi"
    df_fg_wname$NAME[trim(df_fg_wname$N_FCODE) %in% whiterice] <- "White␣Rice"
    df_fg_wname$NAME[trim(df_fg_wname$N_FCODE) %in% coffee] <- "Coffee"
    df_fg_wname$NAME[trim(df_fg_wname$N_FCODE) %in% bread] <- "Bread"
  }
  return(df_fg_wname)
}

FoodGroupRank <- function(df = NULL, year = NULL) {
  df1 <- FoodGroupAdd(df, year) %>%
    group_by(ID, FOOD_GROUP, NAME) %>%
    summarise(DAILY_INTAKE_KCAL = sum(NF_EN))

  dfreturn <- df1 %>%
    group_by(FOOD_GROUP, NAME) %>%
    summarise(DailYIntGram = sum(DAILY_INTAKE_KCAL)/length(unique(df1$ID))) %>%
    arrange(desc(DailYIntGram))

  return(dfreturn)
}


baseNutrientsSummary <- function(df, groupby = NULL) {
  if (is.null(groupby)) {
  out <- df %>%
  summarise(NF_EN = mean(NF_EN, na.rm = T),
            NF_PROT = mean(NF_PROT, na.rm = T),
            NF_FAT = mean(NF_FAT, na.rm = T),
            NF_CHO = mean(NF_CHO, na.rm = T))
  } else {
  out <- df %>%
  group_by_(groupby) %>%
  summarise(NF_EN = mean(NF_EN, na.rm = T),
            NF_PROT = mean(NF_PROT, na.rm = T),
            NF_FAT = mean(NF_FAT, na.rm = T),
            NF_CHO = mean(NF_CHO, na.rm = T))
  }
  return(out)
}

upper <- function(df) {
  names(df) <- toupper(names(df))
  df
}
lower <- function(df) {
  names(df) <- tolower(names(df))
  df
}

trim <- function(x) gsub("^\\s+|\\s+$", "", x)



########################
## READ IN DATA ########
########################
# read in all 24h recall nutrition survey
files <- list.files("data/", pattern = "24RC")

for (file in files) {
  td <- as.data.frame(read.spss(paste0("data/", file)), stringsAsFactors = F)
  td_name <- substr(file, 0, 9)

  assign(td_name, td)
}

# read in health examination data
files_all <- list.files(path = "data/", pattern = "_ALL")
for (file in files_all) {
  td <- as.data.frame(read.spss(paste0("data/", file), reencode = "UTF-8"))
  td_name <- substr(file, 0, 8)

  assign(td_name, td)
}


########################
## MISC ################
########################
# Read in GDP DATA PER CAPITA
gdp <- read_excel("data/gdp/gdp.xlsx") %>%
  select(-c(1:3)) %>%
  gather(YEAR, GDPPC, -CountryCode) %>%
  mutate(YEAR = as.Date(substr(YEAR, 1, 4), format = "%Y"),
         GDPPC = as.numeric(as.character(GDPPC)))

# GDP per capita plot
ggplot(gdp, aes(x = YEAR, y = GDPPC)) +
  geom_line(group = 1) +
  labs(title = "South␣Koreas␣Gross␣Domestic␣Product␣per␣Capita",
       subtitle = "in␣current␣US$",
       caption = "based␣on␣data␣from␣databank.worldbank.org",
       x = "Year",
       y = "GDP␣per␣capita␣(in␣current␣US$)") +
  theme_bw() +
  scale_y_continuous(breaks = seq(0, 30000, by = 5000), labels = comma) +
  scale_x_date(breaks =  seq(as.Date("1967-01-05"), as.Date("2016-01-05"),
                             by = "4␣years"),
               labels = date_format("%Y"))


#########################
## DEMOGRAPHIC ANALYSIS #
#########################
# application of survey weights
########
# 1998 #
########
# testsvy <- svydesign(ids = ~kstrata, data = HN98_24RC, weights = ~wt_24rc)
# Mutate new variables for further analysis
HN98_ALL$BMI <- HN98_ALL$HE_WT/(HN98_ALL$HE_HT/100)^2
HN98_ALL$EDUC_GROUP <- ifelse(HN98_ALL$educ %in% c(4:7),
                              ">=High␣school␣diploma",
                              "<High␣school␣diploma")
HN98_ALL <- HN98_ALL %>%
  mutate(AGEGROUP = ifelse((age > 6 & age < 13), "7-12␣y",
                    ifelse((age > 12 & age < 19), "13-18␣y",
                    ifelse((age > 18 & age < 40), "19-39␣y",
                    ifelse((age > 39 & age < 60), "40-59␣y",
                    ifelse(age > 59, "60+␣y",
                    ifelse((age < 7 & age > 1), "2-6␣y", "<2y")))))))
# initialize survey
svy98 <- svydesign(ids = ~kstrata, data = HN98_ALL, weights = ~wt_itv)

# Age
svymean(~age, design = svy98)
# Sex
prop.table(svytable(~sex, design = svy98))
# Agegroup
prop.table(svytable(~AGEGROUP, design = svy98))
# bmi
svymean(~BMI, design = svy98, na.rm = T)
# bmi by agegroup
svyttest(BMI ~ AGEGROUP, svy98)
# bmi by sex
svymean(~BMI, design = subset(svy98, sex == "1"), na.rm = T)
svymean(~BMI, design = subset(svy98, sex == "2"), na.rm = T)
svyttest(BMI ~ sex, svy98)
# bmi by EDUCGROUP
prop.table(svytable(~EDUC_GROUP, design = svy98))
svymean(~BMI, design = subset(svy98,
                              EDUC_GROUP == ">=High␣school␣diploma"), na.rm = T)
svymean(~BMI, design = subset(svy98,
                              EDUC_GROUP == "<High␣school␣diploma"), na.rm = T)
svyttest(BMI ~ EDUC_GROUP, svy98)
```

```
########
# 2015 #
########
HN15_ALL$BMI <- HN15_ALL$HE_wt/(HN15_ALL$HE_ht/100)^2
HN15_ALL$EDUC_GROUP <- ifelse(HN15_ALL$educ %in% c(4:7),
                              ">=High␣school␣diploma",
                              "<High␣school␣diploma")
HN15_ALL <- HN15_ALL %>%
  mutate(AGEGROUP = ifelse((age > 6 & age < 13), "7-12␣y",
                    ifelse((age > 12 & age < 19), "13-18␣y",
                    ifelse((age > 18 & age < 40), "19-39␣y",
                    ifelse((age > 39 & age < 60), "40-59␣y",
                    ifelse(age > 59, "60+␣y",
                    ifelse((age < 7 & age > 1), "2-6␣y", "<2y")))))))

# initialize survey
svy15 <- svydesign(ids = ~kstrata, data = HN15_ALL, weights = ~wt_hs)

# Age
svymean(~age, design = svy15)
# Sex
prop.table(svytable(~sex, design = svy15))

# Agegroup
prop.table(svytable(~AGEGROUP, design = svy15))
# bmi
svymean(~BMI, design = svy15, na.rm = T)
# educgroup
prop.table(svytable(~EDUC_GROUP, design = svy15))
# bmi by agegroup
svyttest(BMI ~ AGEGROUP, svy15)
# bmi by sex
svymean(~BMI, design = subset(svy15, sex == "1"), na.rm = T)
svymean(~BMI, design = subset(svy15, sex == "2"), na.rm = T)
svyttest(BMI ~ sex, svy15)
# bmi by EDUCGROUP
svymean(~BMI, design = subset(svy15,
                              EDUC_GROUP == ">=High␣school␣diploma"), na.rm = T)
svymean(~BMI, design = subset(svy15,
                              EDUC_GROUP == "<High␣school␣diploma"), na.rm = T)
svyttest(BMI ~ EDUC_GROUP, svy15)


########################
## MACRO ANALYSIS #######
########################
ordervec <- as.character(c(1998, 2001, 2005, 2007, 2008, 2009, 2010:2015))
```

```
df_macros <- data.frame()
for (vec in ordervec) {
  print(vec)
  df <- get(paste0("HN_", vec))
  sampsize <- length(unique(df$id))
  print(sampsize)

  df2 <- baseNutrientsSummary(df = df)
  df2$YEAR <- vec
  df_macros <- rbind(df_macros, df2)
}

df_m2 <- df_macros %>%
  group_by(YEAR)  %>%
  summarise(NF_EN = mean(NF_EN, na.rm = T)*4,
            NF_PROT = mean(NF_PROT, na.rm = T)*4,
            NF_FAT = mean(NF_FAT, na.rm = T)*9,
            NF_CHO = mean(NF_CHO, na.rm = T)*4) %>%
  mutate(YEAR = as.Date(YEAR, format = "%Y")) %>%
  gather(MACRONUTRIENT, GRAMS_DAILY, NF_PROT:NF_CHO)

# Macronutrient Plot
df_m2 %>%
  mutate(MACRONUTRIENT = ifelse(MACRONUTRIENT == "NF_PROT", "Proteins",
                         ifelse(MACRONUTRIENT == "NF_FAT", "Fats",
                         ifelse(MACRONUTRIENT == "NF_CHO",
                                "Carbohydrates", "")))) %>%
ggplot(., aes(x = YEAR, y = GRAMS_DAILY, group = MACRONUTRIENT,
              fill = MACRONUTRIENT)) +
  geom_area() +
  scale_x_date(breaks = c(as.Date(c("1998-01-05", "2001-01-05", "2005-01-05")),
                          seq(as.Date("2007-01-05"), as.Date("2014-01-05"),
                              by = "1␣years")),
               labels = date_format("%Y")) +
  labs(title = "Daily␣Energy␣Intake␣by␣Macronutrients␣from␣1998␣until␣2015",
       x = "Year",
       y = "Energy␣Intake␣in␣kcal") +
  theme_bw()


## T-TEST
# EN
en98 <- upper(HN98_24RC) %>%
  group_by(ID) %>%
  summarise(SUM_FAT = sum(NF_EN)) %>%
  select(SUM_FAT) %>%
  unlist(.)

en15 <- upper(HN15_24RC) %>%
  group_by(ID) %>%
  summarise(SUM_FAT = sum(NF_EN)) %>%
  select(SUM_FAT) %>%
  unlist(.)

var.test(en98, en15)
t.test(en98, en15, var.equal = F)

# FAT
fat98 <- upper(HN98_24RC) %>%
  group_by(ID) %>%
  summarise(SUM_FAT = sum(NF_FAT*9)) %>%
  select(SUM_FAT) %>%
  unlist(.)

fat15 <- upper(HN15_24RC) %>%
  group_by(ID) %>%
  summarise(SUM_FAT = sum(NF_FAT*9)) %>%
  select(SUM_FAT) %>%
  unlist(.)

var.test(fat98, fat15)
t.test(fat98, fat15, var.equal = F)

# PROTEIN
PROT98 <- upper(HN98_24RC) %>%
  group_by(ID) %>%
  summarise(SUM_PROT = sum(NF_PROT*4)) %>%
  select(SUM_PROT) %>%
  unlist(.)

PROT15 <- upper(HN15_24RC) %>%
  group_by(ID) %>%
  summarise(SUM_PROT = sum(NF_PROT*4)) %>%
  select(SUM_PROT) %>%
  unlist(.)

var.test(PROT98, PROT15)
t.test(PROT98, PROT15, var.equal = F)


# CARBS
CHO98 <- upper(HN98_24RC) %>%
  group_by(ID) %>%
  summarise(SUM_CHO = sum(NF_CHO*4)) %>%
  select(SUM_CHO) %>%
  unlist(.)

CHO15 <- upper(HN15_24RC) %>%
  group_by(ID) %>%
  summarise(SUM_CHO = sum(NF_CHO*4)) %>%
  select(SUM_CHO) %>%
  unlist(.)

var.test(CHO98, CHO15)
# p value = 0.2313 > 0.01 -> null hyp. not rejected: variances equal
t.test(CHO98, CHO15, var.equal = T)
# -> p value = 0.00139 < 0.01 -> null hyp rejected: means not equal

# BMI
BMI98 <- upper(HN98_ALL) %>%
  group_by(ID) %>%
  summarise(BMI = (HE_WT/(HE_HT/100)^2)) %>%
  select(BMI) %>%
  unlist(.)

BMI15 <- upper(HN15_ALL) %>%
  group_by(ID) %>%
  summarise(BMI = (HE_WT/(HE_HT/100)^2)) %>%
  select(BMI) %>%
  unlist(.)

var.test(BMI98, BMI15)
# p value < 0.01 -> reject null hypothesis: variance is not equal
t.test(BMI98, BMI15, var.equal = F)
# null hypothesis: difference in means is 0 | no difference in means
# alt. hypothesis: difference in means is not 0 | difference in means
# p value < 0.01 -> reject null hypothesis -> means are signif. different


#########################
## FOOD GROUP ANALYSIS ##
#########################
foodgroup_db <- data.frame(
  ID = 1:18,
  NAME = c("Grains", "Potatoes␣and␣Starch", "Sugars", "Legumes",
           "Seeds␣and␣Nuts", "Other␣Vegetables", "Mushrooms",
           "Fruits", "Meats", "Eggs", "Fish", "Seeweads",
           "Milk␣and␣Dairy␣Products", "Fat␣and␣Oils",
           "Beverages", "Seasonings", "Processed␣Foods", "Others"))

foodgroup98 <- FoodGroupRank(df = HN98_24RC, year = 98)
foodgroup15 <- FoodGroupRank(df = HN15_24RC, year = 14)


#########################
# DIETARY PATTERNS ######
#########################
# 1998
df1 <- FoodGroupAdd(HN98_24RC, 98) %>%
  group_by(ID, NAME) %>%
  summarise(DAILY_INTAKE_KCAL = sum(NF_EN)) %>%
  mutate(DAILY_INTAKE_RELATIVE = DAILY_INTAKE_KCAL/sum(DAILY_INTAKE_KCAL)) %>%
  select(ID, NAME, INTK_GRAM = DAILY_INTAKE_RELATIVE)

filternames <- upper(HN98_24RC) %>%
  group_by(ID) %>%
  summarise(NF_EN = sum(NF_EN, na.rm = T)) %>%
  filter(NF_EN > 500,
         NF_EN < 5000) %>%
  select(ID) %>%
  unlist(.)

allnames <- upper(HN98_24RC) %>%
  group_by(ID) %>%
  summarise(NF_EN = sum(NF_EN, na.rm = T)) %>%
  select(ID) %>%
  unlist(.)

df2 <- df1 %>%
  spread(NAME, INTK_GRAM) %>%
  filter(ID %in% filternames)

dfana <- df2[ ,-1]
dfana[is.na(dfana)] <- 0

# SILHOUETTE AND ELBOW CRITERION
mydata <- scale(dfana)
fviz_nbclust(mydata, kmeans, method = "wss")
fviz_nbclust(mydata, kmeans, method = "silhouette")


# 2015
df1 <- FoodGroupAdd(HN15_24RC, 14) %>%
  group_by(ID, NAME) %>%
  summarise(DAILY_INTAKE_KCAL = sum(NF_EN)) %>%
  mutate(DAILY_INTAKE_RELATIVE = DAILY_INTAKE_KCAL/sum(DAILY_INTAKE_KCAL)) %>%
  select(ID, NAME, INTK_GRAM = DAILY_INTAKE_RELATIVE)

filternames <- upper(HN15_24RC) %>%
  group_by(ID) %>%
  summarise(NF_EN = sum(NF_EN, na.rm = T)) %>%
  filter(NF_EN > 500,
         NF_EN < 5000) %>%
  select(ID) %>%
  unlist(.)

allnames <- upper(HN15_24RC) %>%
  group_by(ID) %>%
  summarise(NF_EN = sum(NF_EN, na.rm = T)) %>%
  select(ID) %>%
  unlist(.)

length(allnames) - length(filternames)

df2 <- df1 %>%
  spread(NAME, INTK_GRAM) %>%
  filter(ID %in% filternames)

dfana <- df2[ ,-1]
dfana[is.na(dfana)] <- 0

# SILHOUETTE AND ELBOW CRITERION
mydata <- scale(dfana)
fviz_nbclust(mydata, kmeans, method = "wss")
fviz_nbclust(mydata, kmeans, method = "silhouette")

## PCA
pc <- prcomp(dfana, center = T, scale. = T)
screeplot(pc, type = "l")

## K MEANS CLUSTER ANALYSIS
kmcl <- kmeans(dfana, 2, 100)
kmcl
```

Declaration of Authorship

I, Darius Jonda, hereby confirm that I have authored this Bachelor's thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, February 20, 2017

Darius Jonda