Literature Review and Exploratory Data Analysis

The effect of dietary patterns in general population on the mortality and survival of Chronic Kidney Disease (CKD) patients

Data Analytics: Major Research Project

Sayed Ahmed
MSc in Data Science and Analytics
Ryerson University
500869723

Supervisor
Dr. Youcef Derbal
Information Technology Department
Ted Rogers School of ITM
Ryerson University

Abstract

Chronic kidney disease (CKD) is very prevalent in today's world, and CKD incidents are continually increasing such as over 30 millions of Americans have CKD [30]. CKD can result in End Stage Renal Disease (ESRD) i.e. complete loss of kidney function. CKD/ESRD and other interrelated diseases such as Hypertension, Heart Diseases, and Diabetes cause a majority of the early deaths [31]. In addition to kidney failure, CKD is also a major cause of death from stroke, and heart diseases. On the other hand, hypertension and diabetes also cause CKD. Studies show that drugs as well as lifestyle choices can prevent CKD, slow the progression of CKD [29], delay dialysis and kidney transplantation; consequently can prevent early deaths. Though there are many studies on the effect of drugs to control CKD and related complications, there are few studies on the effect of diets and lifestyles [1]. This research has identified the association between dietary patterns and mortality/survival of CKD patients. Dietary pattern data provided by the Centers for Disease Control and Prevention (CDC) and Health.gov as well as CKD related mortality and survival data provided by the United States Renal Data System (USRDS) [22] is used to study the effect of the dietary patterns in general population on the mortality/survival of patients with CKD. Machine Learning approaches such as Regression, and Principal Component Analysis are utilized for initial analysis to identify some of the affecting features (i.e. food groups/subgroups). For data exploration, Univariate and Bivariate Analysis, Pearson correlation, Heatmap, and Data Visualizations are used. However, approaches such as Clustering, Decision Trees, Random Forests, SVM, Ensemble Methods, Deep Learning and/or others will be used as appropriate to find out and analyze the relations between dietary patterns and survival/mortality of CKD and ESRD patients.

Introduction

Chronic kidney disease (CKD) is very prevalent in today's world and CKD incidents are continually increasing such as 10 to 13% of the US population get affected by Chronic Kidney Disease. CKD is not reversible and is progressive that gradually reduces kidney function. CKD is identified with a blood test such as Glomerular Filtration Rate (GFR) or a urine test such as Albumin Creatinine Ratio (ACR). GFR is measured in ml/min/1.73 m^2 . CKDs are described in stages such as **Stage** 1 with normal or high GFR (GFR > 90 mL/min), **Stage** 2 with Mild **CKD** (GFR = 60-89 mL/min), **Stage** 3A with Moderate **CKD** (GFR = 45-59 mL/min), **Stage** 3B with Moderate **CKD** (GFR = 30-44 mL/min), **Stage** 4 with Severe **CKD** (GFR = 15-29 mL/min), **Stage** 5 with End **Stage CKD** (GFR <15 mL/min) [5]. At stage 5, patients loss complete kidney function then either require dialysis or transplantation to survive.

CKD/ESRD and other interrelated diseases such as Hypertension, Heart Diseases, and Diabetes cause a majority of the early deaths [31]. In addition to kidney failure, CKD is also a major cause of death from stroke, and heart diseases. On the other hand, hypertension and diabetes are also major causes of CKD.
As CKDs are not curable and reversible controlling diabetes and blood pressure with or without medication can slow the progress of CKDs. As Kidneys filter waste products and our diet produce those waste products controlling diet have an effect on how much work kidney has to perform and how well the kidney will function. Studies show that drugs as well as lifestyle choices (diet, exercise) can prevent CKD, slow the progression of CKD [29], delay dialysis and kidney transplantation; consequently can prevent early deaths. Though there are many studies on the effect of drugs to control CKD and related complications, there are few studies on the effect of diets and lifestyles [29]. There are studies in how

controlling nutrients/chemicals in food items can help prevent or slow the progression of CKD. However, adhering to the recommended amount of nutrients is challenging. Hence, there is an emerging trend where the effect is studied utilizing dietary patterns with food groups and subgroups rather than nutrients/chemicals in food. This research analyzes the effect of dietary patterns using food groups and subgroups on the mortality and survival of CKD patients.

Literature Review

Kidney patients commonly are given dietary advice based on individual nutrients or chemicals primarily or sometimes on food items instead of whole eating patterns. However, that advice is challenging to adhere to for the majority of the patients [2]. Also, there is limited evidence that adherence to such advice prevents clinical complications [23]. Hence, studying the whole dietary patterns rather than single nutrient or food group restrictions is an emerging trend for CKD/ESRD patient diets [2] [24-26]. This is also easier to adhere to. There are several studies on analyzing the relation between dietary patterns and clinical outcomes for CKD patients [3, 4, 5, 6, 7, 8, 9, 26].

Chen at al [3] studied the association of plant protein intake for all cause mortality in CKD. In the study higher plant protein ratio was found to cause lower mortality for CKD patients in stage 3 or higher (eGFR cys < 60 ml/min/1.73 m2) though not for others (stage 1 and 2) [3]. This study primarily used statistical methods and Regression Models such as Cox regression models to find the association [3]. Hao-Wen et al [26] studied the association between vegetarian diets and CKD. The study found that vegetarian diets including vegan and ovo-lacto vegetarian diets were possible protective factors. The study utilized The multivariable logistic regression analysis [26].

Gutiérrez et al [4] studied 5 empirically derived dietary patterns such as "convenience" (Chinese and Mexican foods, pizza, and other mixed dishes), "plant-based" (fruits and vegetables), "sweets/fats" (sugary foods), "Southern" (fried foods, organ meats, and sweetened beverages), and "alcohol/salads" (alcohol, green-leafy vegetables, and salad dressing) [4]. The study found that dietary pattern rich in processed and fried foods was associated with higher mortality in persons with CKD. On the other hand, a diet rich in fruits and vegetables was found to be protective [4].

Huang et al [5] studied whether Mediterranean diet can preserve kidney function along with maintaining favorable cardiometabolic profile with reduced mortality risk for individuals with CKD. The study found that adhering to Mediterranean diet has a lower likelihood of having CKD in elderly men. The study also found that a greater adherence to this diet can improve survival for CKD patients [5]. Huang et al [5] in the above study, used unpaired *t* test, nonparametric Mann–Whitney test, or $\chi 2$ test as appropriate for Comparisons between CKD and non-CKD men. To evaluate the association of Mediterranean diet with the presence of CKD, Crude and multiple adjusted logistic regression models were fitted. All tests were two-tailed, and $P < 0.05$ was considered significant [5].

One aspect of Muntner et al [6] study was to find out how Life's Simple 7 factors (Smoke, Activity, BMI, Diet, Blood Pressure, Cholesterol, and Glucose) affect in getting ESRD. The study shows that people who have high/ideal scores in more of these factors have lower likelihood of getting ESRD. This study utilized Cox proportional hazards models. Adjustment were made for age, race, sex, stroke-based geographic region of residence, income, education, and history of stroke or coronary heart disease [6].

Ricardo et al [7] studied the association of death to healthy lifestyles esp. in relation to CKD. The study found that adherence to healthy lifestyles was associated with lower risk of all cause mortality in CKD

patients. In this study, to determine the association between a healthy lifestyle and survival among individuals with CKD, Cox proportional hazards models were used while also adjusting for important covariates. Stratified survival analyses by eGFR and UACR was performed for Sensitivity analyses [7].

Suruya et al. [8] studied dietary patterns in hemodialysis patients in Japan and researched associations between dietary patterns and clinical outcomes. The study found that patients with unbalanced diet were more likely to have adverse clinical outcomes. Hence, such patients when in addition to portion control, maintains a well-balanced diet esp. for the food groups meat, fish, and vegetables will have less adverse clinical outcomes [8]. Suruya et al [8] utilized a principal components analysis (PCA) with Promax rotation to reduce to a smaller set of food groups for analysis. PCA was used to find food groups eaten with equal frequencies [8]. Cox regression model was used for the analysis with multiple models where each model had a different combination of covariants [8].

Another study by Ricardo et al [9] estimated the degree of adherence to a healthy lifestyle that decreases the risk of renal and cardiovascular events among adults with chronic kidney disease (CKD). The study found that adherence to a healthy lifestyle was associated with lower all-cause mortality risk in CKD. The greatest reduction in all-cause mortality was related to nonsmoking [9]. This study by Ricardo et al [9], to compare categorical and continuous variables used Chi-squared and analysis of variance tests respectively. To examine the association between healthy lifestyle and outcomes, Cox proportional hazards models were used. Death was treated as a censoring event. Three nested Cox proportional hazards models were fitted and were adjusted sequentially for potential explanatory variables [9].

G. Asghari et al studied the association of population-based dietary pattern with the risk of incident CKD The study concluded that high fat and high sugar diet pattern is associated with significantly increased (46%) odds of incident CKD where a lacto-vegetarian diet can be protective of CKD by 43%. The study utilized multivariable logistic regression to calculate odds ratio for the association.

One of the studies above utilized the dietary pattern data from CDC and NHANES as this study will also use. However, this study will differ in the methodology, exploration, and analysis. This study is finding relations between datasets from multiple sources and is focused on finding patterns and relations in general population than specific/selected individuals. Most of the studies above utilized primarily statistical methods and sensitivity analysis where primarily regression models esp. Cox regression models were used. In a couple of cases, Principal Component Analysis (PCA) was used. There is a lack of study that utilized AI approaches including Machine Learning, and/or Deep Learning to find the association between dietary patterns and CKD/ESRD mortality/survival. In this study, Regression/Cox's Regression as well as PCA is also used. However, in this study, approaches such as AI, ML, and Deep Learning will be explored to find and analyze the association between dietary patterns and CKD/ESRD mortality.

Methodology and Exploratory Data Analysis

The primary purpose of this research is to assess the effect of dietary habits of the general population on the mortality and survival of chronic kidney disease (CKD)/End Stage Renal Disease (ESRD) patients. The dietary habits of different age groups as well as age group based mortality and survival of CKD/ESRD patients are studied; Afterwards, machine learning approaches are applied to find the

relation between dietary habits and the mortality and survival of CKD/ESRD patients.

The aim is to find out if deviation or compliance with current food intake recommendations [15] by age groups has any effect on mortality and survival. Current food intake recommendations from health.gov [15] is used. Additionally, a study [11] on shifting from current recommendations conducted by health.gov is considered. Whether the recommended shift [11] from current diet style [15] can have an improved outcome or not is also studied * (i.e. the difference between current style and shift style when big, do we see a more negative effect in that population).

The primary aim is to study the effect of food groups and subgroups based dietary patterns [14, 12, 11] in American population on the mortality and survival of CKD/ESRD patients.

Study Selection

For mortality and survival, data from the United States Renal Data System (USRDS) on CKD and ESRD [16, 17] was studied. "USRDS investigates the transition of care from CKD to ESRD and end-of-life care for those with advanced kidney disease" [19]. USRDS also releases data on the Incidence, Prevalence, Patient Characteristics, and Treatment Modalities on CKD, and ESRD patients. USRDS reports the survival and mortality using metrics such as 90 day survival, 5 year survival and/or 10 year survival, Mortality rates: ESRD patients, Avg. Expected remaining lifetime with or without pre-condition and treatment options used. The data released are either aggregated or patient specific detail data. However, only aggregated data are public where patient specific data requires special request and permission.

For dietary data, the **National Health and Nutrition Examination Survey** on dietary habits as conducted by the Centers for Disease Control and Prevention (CDC) [10] was used. The survey has data from 1996 to 2016 [10]. The survey recorded 24 hours intake amount. Two surveys were taken within 3 to 10 days after the first survey. The survey data provided intake amount by food groups and subgroups, also mentioned the diet style, diet-restriction, and isolated nutrient intake (such as sodium, and sugar).

For this study, primary focus is 2015 - 2016 data (diet and mortality) where previous years' (1996 to 2014) data is used to find out the changes in dietary habits over the years and whether that pattern change have any relation on the survival and mortality of CKD/ESRD patients. Additionally, the effect/relation data for each year is used as one vote, and hence, food groups/subgroups that are found to affect in multiple years will got more votes to be the dominant actor. (* the approaches as said are subject to change)

For dietary patterns, the food groups and subgroups as used in the study/article by health.gov [11] on recommending shift/changes to existing recommendations on diet styles is used. The research [11] studied and recommended where shifts will be important to maintain a good health as well as what can be easily followed/adhered to by the population.

The dietary survey data (NHANES) represented the food items taken by the participants using USDA food codes [14, 12, 13]. Hence, USDA food codes [14, 12, 13] are used to assign food groups and subgroups to the NHANES [10] survey data to properly group/subgroup the dietary intake of the participants. Proper subgroups were assigned to the foods taken as closely/completely possible to match the shift recommendation article [11]. More subgroups are there in the NHANES survey data than the shift recommendation article [11]. When no matching subgroup from survey data was found in the article,

that subgroup is used as a new subgroup in this study. The same methodology as applied on the matching subgroups is applied on these new subgroups to study the effect.

Adjustments were required for primary groups as well such as NHANES/USDA/CDC [14] used Legumes, and Eggs as primary food groups. However, in the shift recommendation article [11] groups such as Legumes, and Eggs are not primary groups rather Legumes (Lentils and Peas) are part of Vegetables group, and Eggs are under Protein. The approach taken by the shift recommendation article is used in this research.

The primary focus is to study the effect of dietary patterns. However, the effect of the recommended shift is also studied (provided I can find a methodology/algorithm for that); The same algorithms are applied on the individual nutrients to understand their effects.

In this study, survey data from two different days are utilized i.e averaged intake values are used. At this point, this study did not exclude any survey data based on dietary restrictions or for health pre conditions. All survey data were used irrespective. [measures such as dietary restrictions or for health pre conditions are subject to be considered in future work]

Data Aggregation strategy used to combine multiple survey data

1. Combined all survey data into one dataset
2. First aggregated the combined data separately for each day (each survey day)
3. Then divided the sum of food intake by the sum of participants for that food group/subgroup

Data Extraction and Quality Assessment

For mortality and survival study, the data from the United States Renal Data System (USRDS) on CKD and ESRD statistics [16, 17] was used. The age grouping as utilized in the shift recommendation study by health.gov article were different than USRDS data; (health.gov also used NHANES survey data). NHANES dietary intake survey data as provided by each participant was customized to reflect the age groups of USRDS.

Neither NHANES (diet) nor USRDS (mortality/survival) provide aggregated data based on age groups used by the other party. Getting individual patient data or customized (such as age groups) aggregated data from USRDS required a request and permission procedure that could affect the timely completion of this project. As dietary intake data for each participant from NHANES was available publicly, dietary intake data was aggregated based on the age groups used by USRDS.

The recommended food group intake used in the shift recommendation article by health.gov (also in general by health.gov) uses age groups that differs from age groups used in USRDS data and the aggregated average intake data. Recommended amounts are also regrouped to reflect USRDS age groups by evenly distributing the amount to each age and then calculating average recommendation amounts for USRDS age groups. Specific age based recommendation data is also generated along with mortality data is also calculated for each age. As a first step, the association between dietary pattern and mortality is studied. As a second step, the association between deviations from the recommendations and mortality is studied (age or age-group based provided the recommendation data generation is found to be appropriate).

Data Synthesis and Analysis

The shift recommendation article utilized gender based data and gender based recommendations. The USRDS data for mortality used gender neutral data where remaining life data provided both gender based and gender neutral data. Hence, experiments are designed based on the data availability considering gender or gender-neutral.

Effects are studied for groups and subgroups separately. The new (not in the shift recommendation study) subgroups as we found in the data utilizing USDA food codes are also studied. For the mortality/survival measures such as 90 day survival, 5 year survival and/or 10 year survival, Mortality rates: ESRD patients, Avg. Expected remaining lifetime, ESRD patients: Total (or %) deaths for target year, ESRD patients: Avg. Annual Mortality rates, Dialysis patients: Total (or %) deaths for target year, Dialysis patients: Avg. Annual Mortality rates, primary cause of mortality, Avg. Expected remaining lifetime (Optional), 90 day survival probabilities, 1 year survival probabilities, 3 years survival probabilities , 10 years survival probabilities and/or similar are used.

On another note, the target variables will be from USRDS data where varying age groupings are utilized to report data from varying studies. As mentioned before, dietary data is customized depending on what age groups are used in the USRDS study under consideration.

Data Exploration/Exploratory Analysis

Data Description

The dietary intake data as used from NHANES [10] (National Health and Nutrition Examination Survey) provides the demographics of the survey participants, food item names used for the survey, associated USDA food code for each food item taken, survey on food items taken by participants on two different days (within 3 to 10 days of first day), nutrients taken on two days, characteristics (such as diet-restriction) of the patients.

Additionally, this study used some other data (meta-data, tables) (file in the submission: data_helping_with_food_grouping_subgrouping.zip or kept on Google Drive) to help with assigning groups and subgroups to dietary intake data such as USDA primary food code grouping strategy such as Key Concepts About the USDA Food Coding Scheme [14] (File in the submission: usda_primary_food_groups). Food subgrouping scheme in this study used the information from Food Code Numbers and the Food Coding Scheme [12] and VEGETABLE SUBGROUPS [13]. The food groups and subgroups as used on a-closer-look-at-current-intak [11] are the target food groups and sub-groups for this research. A data table is created to keep and map USDA Food groups and subgroups to heath.gov food groups and subgroups. For two to four digits of the USDA food codes were used to assign subgroups. (The mapping file: **food_groups_shift_recommendation).** Additionally, every food item taken by the survey participants is mapped to a corresponding group and subgroup (File: map_food_to_groups_sub_groups)

Age groups as used by USRDS for mortality study [22] is used as primary target age groups (the file on google drive: age-groups stores the age groups). However, studies by USRDS have used larger group sizes in other studies such as remaining life study used age groups starting from 21. Customized age

groups are used in this study depending on the aspects measured. (the file age-groups_remaing stores the age groups used by the remaining life study).

Aggregated Data for Analysis

Dietary intake data by age groups, food groups/subgroups, and by gender are kept on Google Drive (also under: multi-day-aggregated-dietary-data.zip). Though the study shift-recommendation utilized gender based food amount recommendations, however, the Mortality/Survival data does not provide details breakdown by gender (for mortality). USRDS requires a long permission procedure to get access to individual patient level data to generate data at custom age groups or gender level that seemed not practical time wise. Hence, initially gender neutral aggregated data as stored on Google Drive (also under: multi-day-aggregated-dietary-data.zip) is used for this analysis.

To relate dietary data to mortality/survival data, related data are put together. Age-group based dietary intake and age group based mortality/survival data are kept side by side on the same excel sheet as shown in the image below also kept on Google drive at age group based dietary intake at subgroup level and mortality, age group based dietary intake at food group level and mortality (Files as submitted: mortality_recom_added_group_data_june_9th_gender_based_data_after_processing, mortality_group_data_june_9th_gender_based_data_after_processing.xlsx, mortality_subgroup_data_june_9th_gender_based_data_after_processing.xlsx)

Age-group – USRDS		Gender	Recommended Vegetable Intake	Actual **Vegetable** Intake	Recommended Protein Intake	Actual **Protein** Intake	Recommended Grain Intake	Actual **Grain** Intake	Recommended Dairy Intake	Actual **Dairy** Intake	Recommended Fruit category intakes	Actual **Fruit** intakes	ommended Sugars, sweets, and beverages amount	Actual Taken **Sugars, sweets, and beverages** amount	ecommended Fats, oils, and salad dressings intake	Avg **Fats, oils, and salad dressings** taken	Patients went for Kidney Transplantation	ESRD patients: Total (or %) deaths for target year	ESRD patients: Avg. Annual Mortality rates	Dialysis patients: Total (or %) deaths for target year	Dialysis patients: Avg. Annual Mortality rates
0	4	Neutral		113.08		125.		196		477.9		264.7		430.64		8.77		32	4.39	27.5	6.665
5	9	Neutral		136.7		179.6		328		368.5		276.3		759.43		14.5		12.5	0.925	8	4.065
10	13	Neutral		150.48		215.6		371		355.6		254.9		941.4		21.3		8	0.505	3.5	1.115
14	17	Neutral		154.13		237.		377		324.6		291.6		1277.5		24.3		11.5	0.565	6	1.215
18	21	Neutral		185.35		279.		378		241.7		287.4		1699.9		25.5		57	1.45	41.5	3.13
22	24	Neutral		216.22		297.		402		250.9		296.2		2059.3		28.9		95	1.995	81	3.455

Figure: Avg Food Group intake and mortality data by age group

Similarly, data are generated and put together for the remaining life study [remaining life].

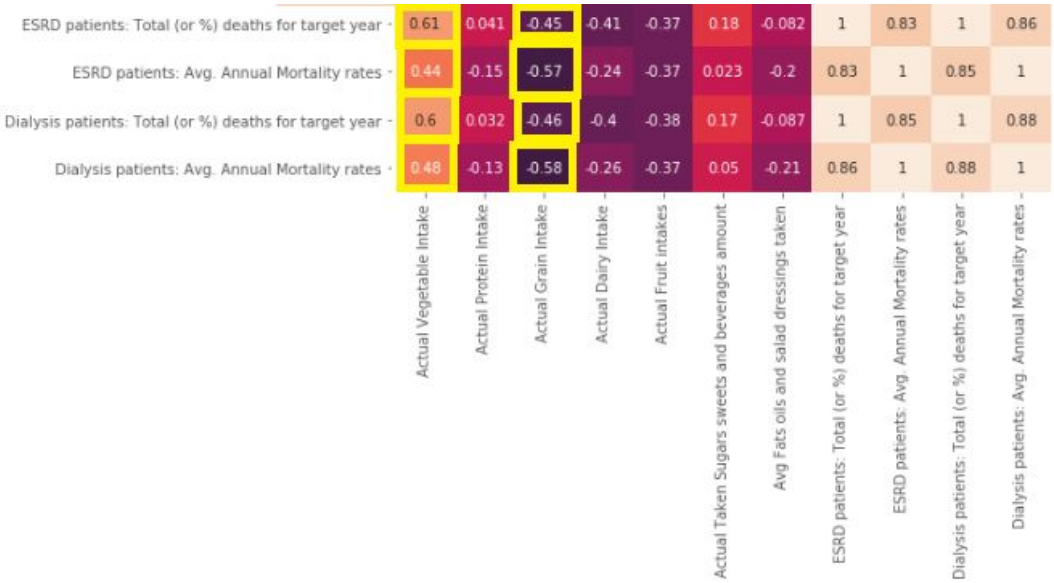
Age-group – USRDS		Gender	Recommended Vegetable Intake	Actual **Vegetable** Intake	Recommended Protein Intake	Actual **Protein** Intake	Recommended Grain Intake	Actual **Grain** Intake	Recommended Dairy Intake	Actual **Dairy** Intake	Recommended Fruit category intakes	Actual **Fruit** intakes	ommended Sugars, sweets, and beverages amount	Actual Taken **Sugars, sweets, and beverages** amount	Recommended Fats, oils, and salad dressings intake	Avg **Fats, oils, and salad dressings** taken	% Population got CKD	People (or %) progressed to Stage 3 CKD	Patients went for Kidney Transplantation	General population: Expected remaining lifetimes:	Dialysis patients: Expected remaining lifetimes	Transplant patients: Expected remaining lifetimes	Dialysis patients with diabetes: Expected lifetimes
0	14	Neutral		132.65		173.		294		410.9		269.7		722.03		15.3				73.0	4.39	60.0	8.7
15	19	Neutral		165.72		257.		378		300.5		270.5		1406.4		24.3				62.0	0.925	48.2	8.0
20	24	Neutral		203.97		288.		392		241.1		315.2		2015.6		29.8				57.2	0.505	44.3	3.0
25	29	Neutral		222.39		302.		405		252.1		275.7		2256.6		25.7				52.5	0.565	40.2	10.0
30	34	Neutral		224.26		295.		416		262.7		279.4		2224.9		22.6				47.8	1.45	36.0	10.2
35	39	Neutral		240.38		273.		398		213.0		301.4		2373.0		27.8				43.1	1.995	32.0	9.6

Association Analysis

Sample Exploratory Analysis: Food Groups

When actual intake amount is utilized, **Vegetables and Grains** are found to be correlated to Mortality or similar target variables. When 'ESRD patients: Total (or %) deaths for target yea' is used as target Vegetable shows more correlation. When '**ESRD patients: Avg. Annual Mortality rates**' is used as the target variable Grain shows better correlation. '**ESRD patients: Avg. Annual Mortality rates**' is in %, where the other one shows count.

For Normalized Data:
Heatmap



Correlation Matrix:

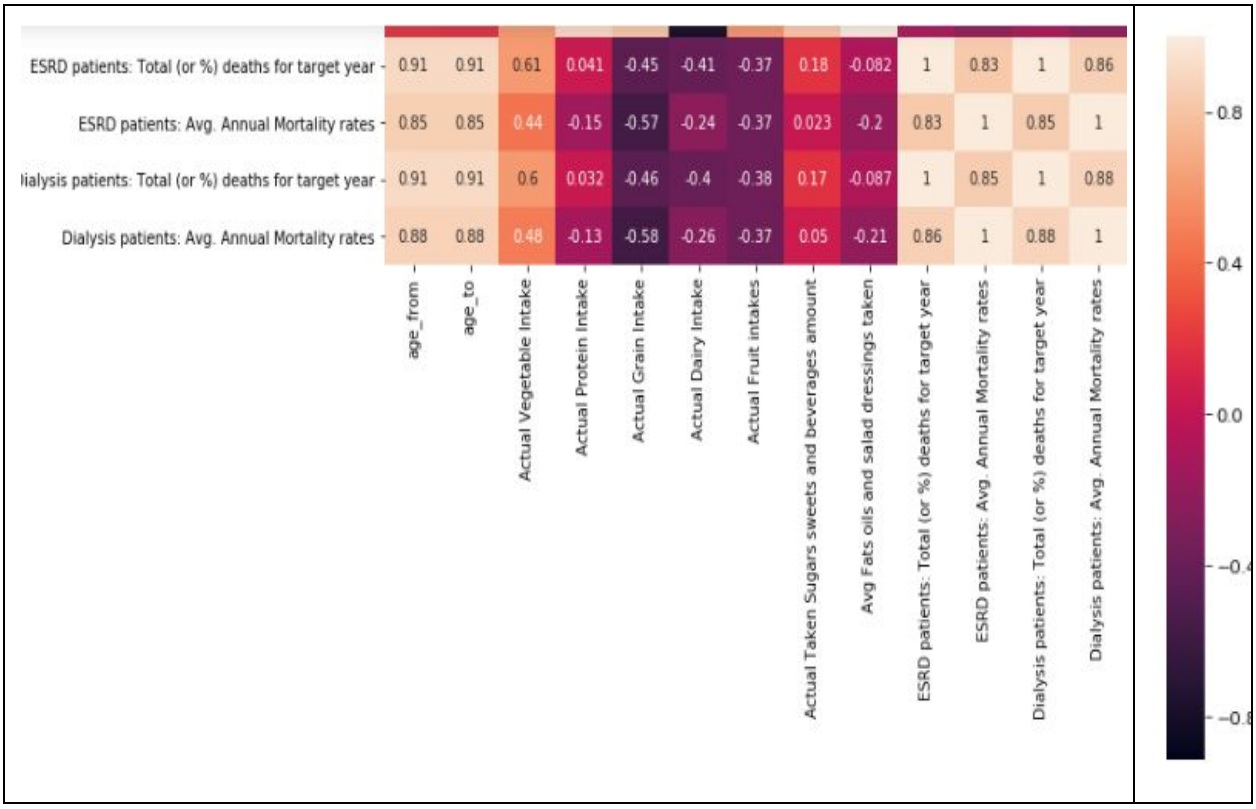
	Vegetable	Protein	Grain	Dairy	Fruit s	Sugars sweets	Fats oils
ESRD patients: Total (or %) deaths for target year	0.607	0.041	-0.453	-0.409	-0.373	0.179	-0.081677
ESRD patients: Avg. Annual Mortality rates	0.445	-0.151	-0.572	-0.238	-0.370	0.023	-0.202818
Dialysis patients: Total (or %) deaths for target year	0.603	0.032	-0.463	-0.404	-0.375	0.173	-0.087055
Dialysis patients: Avg. Annual Mortality rates	0.481	-0.134	-0.581	-0.264	-0.371	0.050	-0.208805

Data Not Normalized:

Correlation Matrix

	Vegetable	Protein	Grain	Dairy	Fruit s	Sugars sweets	Fats oils
ESRD patients: Total (or %) deaths for target yea	**0.61**	0.04	-0.45	-0.41	-0.37	0.18	-0.08
ESRD patients: Avg. Annual Mortality rates	0.44	-0.15	**-0.57**	-0.24	-0.37	0.02	-0.20
Dialysis patients: Total (or %) deaths for target year	**0.60**	0.03	-0.46	-0.40	-0.38	0.17	-0.09
Dialysis patients: Avg. Annual Mortality rates	0.48	-0.13	**-0.58**	-0.26	-0.37	0.05	-0.21
	Positively Related		**Negatively Related**		Less Correlated	Less Correlated	Less Correlated

Heatmap:

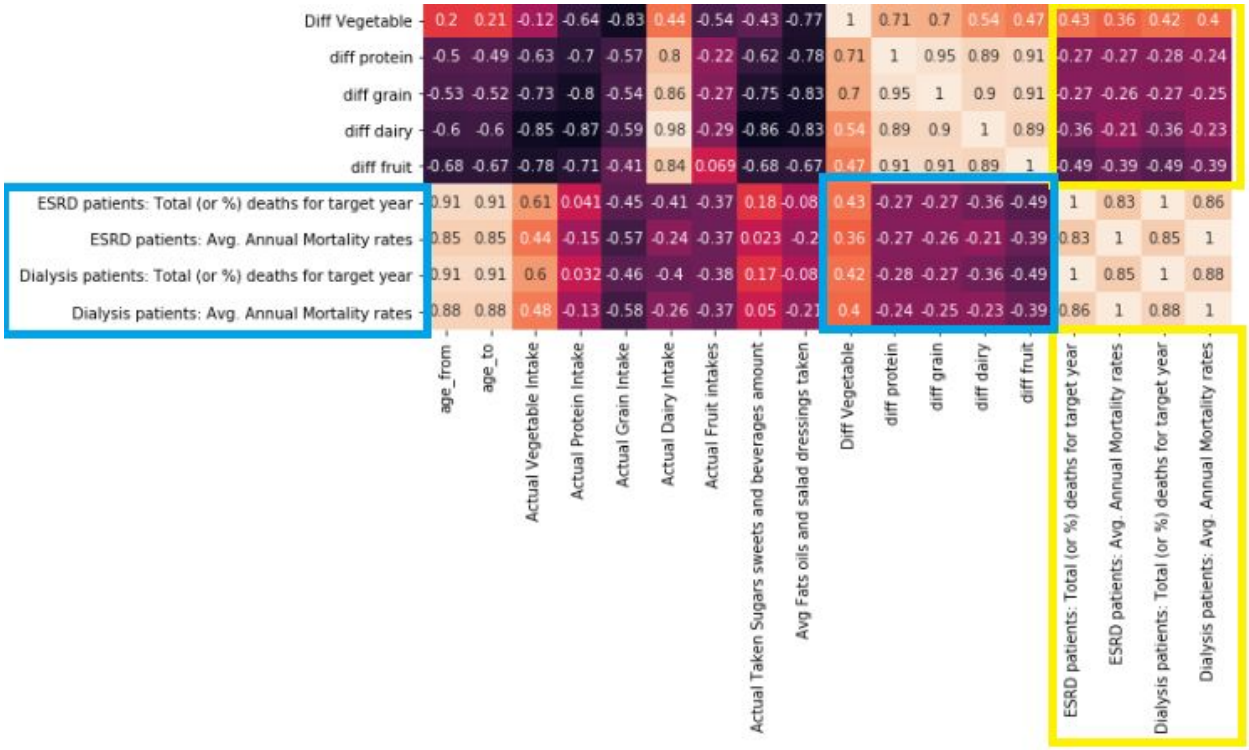


Deviation from Average Recommendation

Deviation from average recommended intake amount for **Fruits and Vegetables** show correlations. Both normalized and not normalized data show very similar (found same) correlation matrix and heatmaps. Hence, showing images only for not-normalized data. The Python file (foodgroup-ckd-mortality.ipynb) will have other plots.

	Diff Vegetable	diff protein	diff grain	diff dairy	diff fruit
ESRD patients: Total (or %) deaths for target year	0.426499	-0.273596	-0.268453	-0.364665	-0.488155
ESRD patients: Avg. Annual Mortality rates	0.35959	-0.266776	-0.263227	-0.213864	-0.39209
Dialysis patients: Total (or %) deaths for target year	0.42434	-0.278817	-0.272938	-0.36214	-0.489873
Dialysis patients: Avg. Annual Mortality rates	0.401294	-0.244088	-0.246893	-0.229239	-0.387759

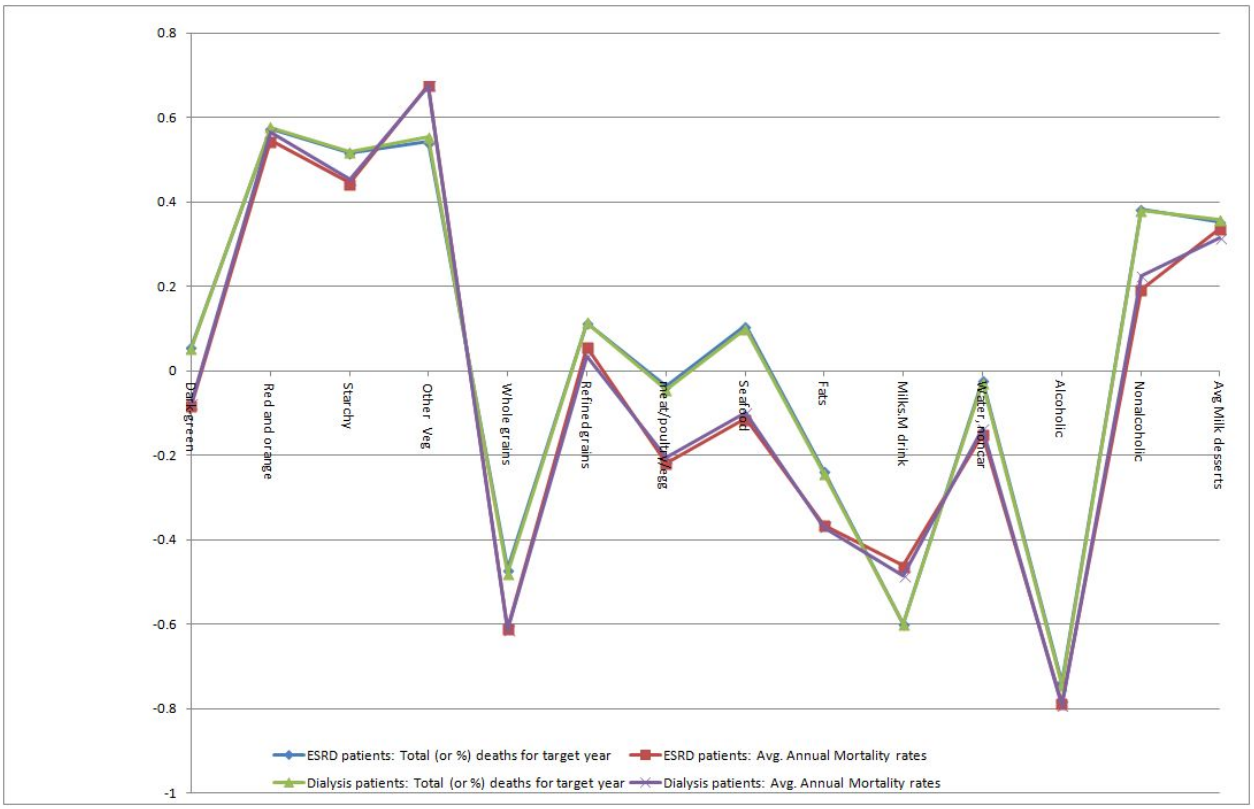
Heatmaps for the Correlation:

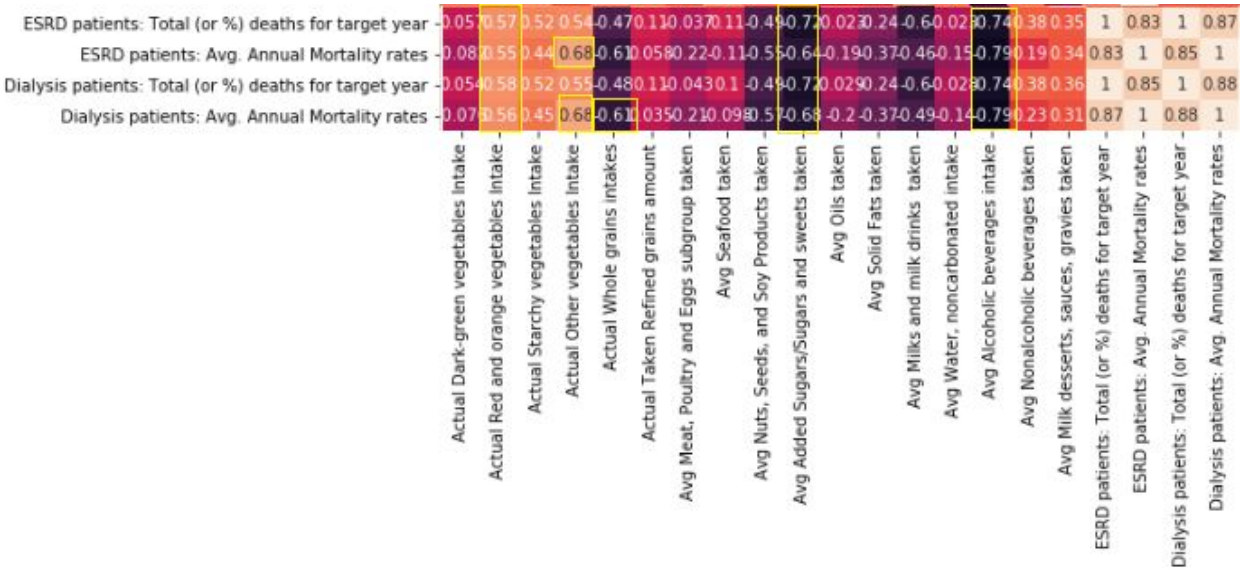


Also, Python file: foodgroup-ckd-mortality.ipynb (or foodgroup-ckd-mortality.pdf file) shows the correlation, heatmaps, univariate and bivariate analysis including PCA based exploration for food groups.

Sample Exploratory Analysis: Food Sub Groups

From correlation data, some of the correlated food subgroups with mortality are: **Alcoholic beverages, Milks/Milk Drinks (lower correlation than Alcoholic beverages), Whole Grains, Other Vegetables. Red and Green Vegetables, and Starchy Vegetables also found to be correlated.** Correlation data and plots are on the file: food-subgroup-mortality-correlation.xlsx. Also, Python file: food-subgroup-ckd-mortality.ipynb shows the correlation, heatmaps, univariate and bivariate analysis including PCA based exploration.


0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Dark green
Red and orange
Starchy
Other Veg
Whole grains
Refined grains
Meat/poultry/eggs
Seafood
Fats
Milks/M drink
Water, noncar
Alcoholic
Nonalcoholic
Avg Milk desserts
ESRD patients: Total (or %) deaths for target year
ESRD patients: Avg. Annual Mortality rates
Dialysis patients: Total (or %) deaths for target year
Dialysis patients: Avg. Annual Mortality rates

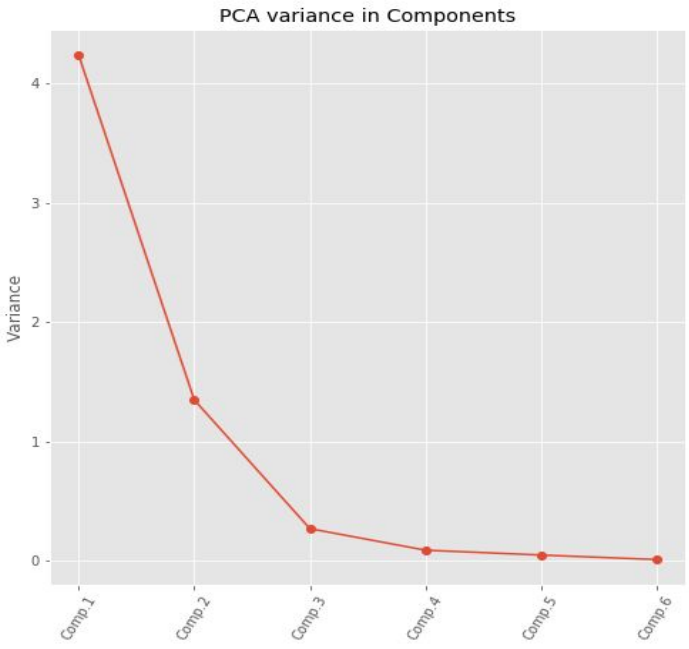

ESRD patients: Total (or %) deaths for target year
ESRD patients: Avg. Annual Mortality rates
Dialysis patients: Total (or %) deaths for target year
Dialysis patients: Avg. Annual Mortality rates
Actual Dark-green vegetables Intake
Actual Red and orange vegetables Intake
Actual Starchy vegetables Intake
Actual Other vegetables Intake
Actual Whole grains intakes
Actual Taken Refined grains amount
Avg Meat, Poultry and Eggs subgroup taken
Avg Seafood taken
Avg Nuts, Seeds, and Soy Products taken
Avg Added Sugars/Sugars and sweets taken
Avg Oils taken
Avg Solid Fats taken
Avg Milks and milk drinks taken
Avg Water, noncarbonated intake
Avg Alcoholic beverages intake
Avg Nonalcoholic beverages taken
Avg Milk desserts, sauces, gravies taken
ESRD patients: Total (or %) deaths for target year
ESRD patients: Avg. Annual Mortality rates
Dialysis patients: Total (or %) deaths for target year
Dialysis patients: Avg. Annual Mortality rates


0.8
0.4
0.0
−0.4
−0.8

The above plots used not normalized data. However, data normalization give the same/very similar output. Python file, food-subgroup-ckd-mortality.ipynb has the plots for normalized data

Principal Component Analysis and the Affecting Variables

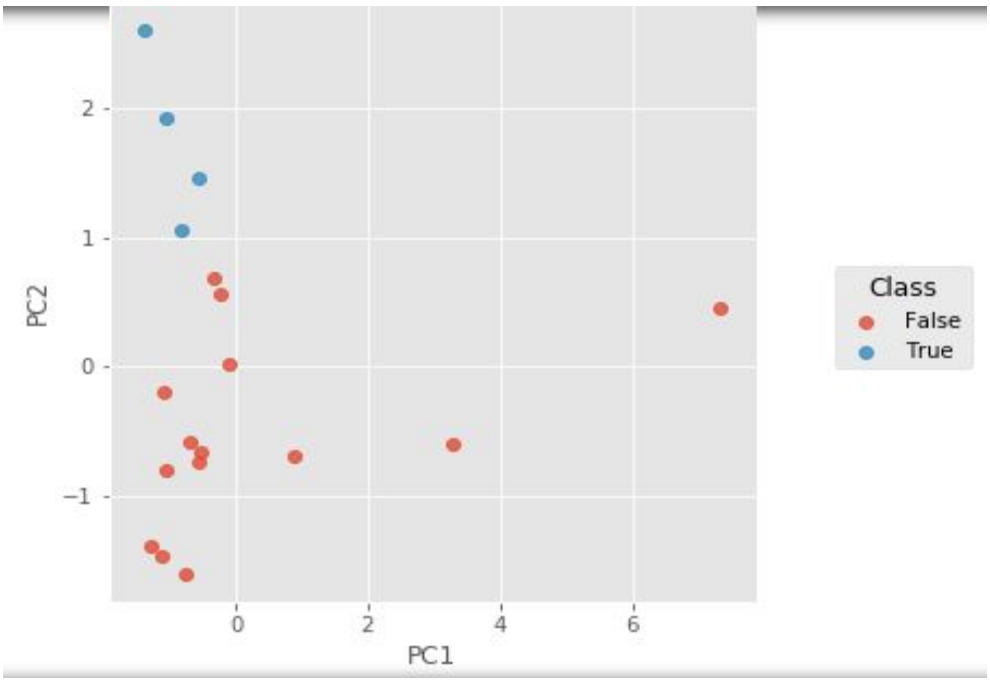
Food Groups



Importance of components:

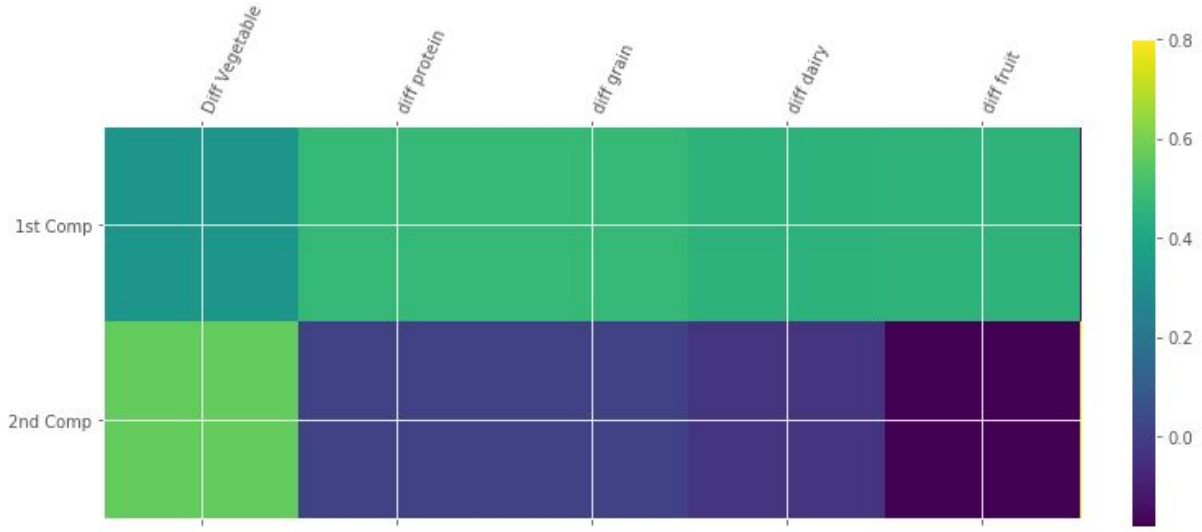
	Std Dev	Propo Vari	Cum Prop
PC1	2.058	0.705	0.705
PC2	1.1607	0.224	0.930
PC3	0.518	0.044	0.975
PC4	0.297	0.0147	0.990
PC5	0.219	0.008	0.998
PC6	0.102	0.0017	1.000

First two component can define over 95%



PC1 and PC2 can separate high and low mortality
Normalized on Avg Mortality rates > 0.5 = High Mortality = True

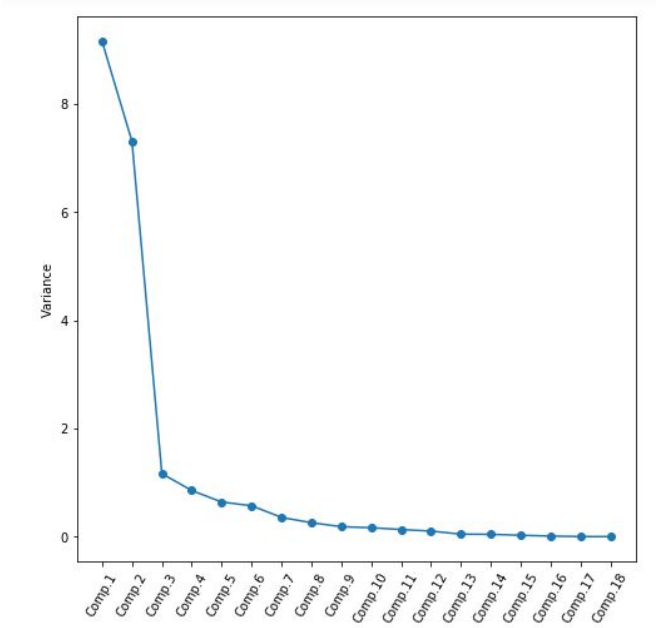
Contributing factors to the First two components



from the above plot, Vegetable, Grain, Protein contribute the most to the 1st component

Food Subgroups

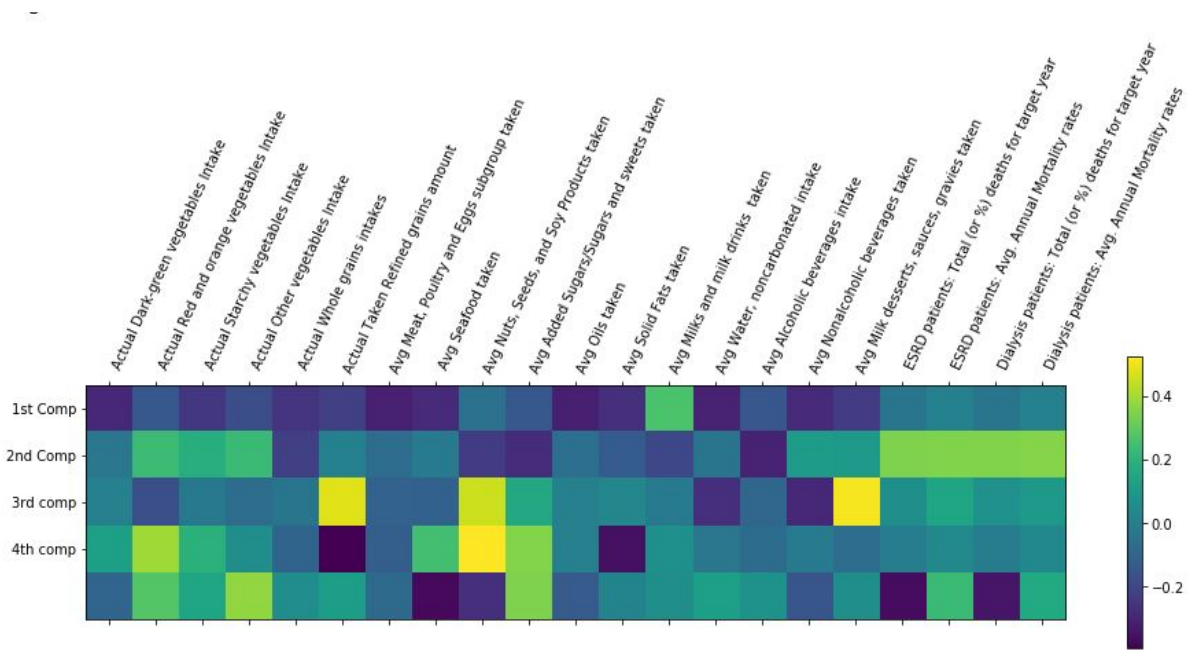
Importance of Components



	sdev	varprop	cumprop
	Standard deviation	**Proportion of Variance**	**Cumulative Proportion**
PC1	3.025508e+00	4.358905e-01	0.435891
PC2	2.702769e+00	3.478552e-01	0.783746
PC3	1.079007e+00	5.544077e-02	0.839186
PC4	9.229833e-01	4.056658e-02	0.879753
PC5	7.996784e-01	3.045169e-02	0.910205
PC6	7.554382e-01	2.717557e-02	0.937380

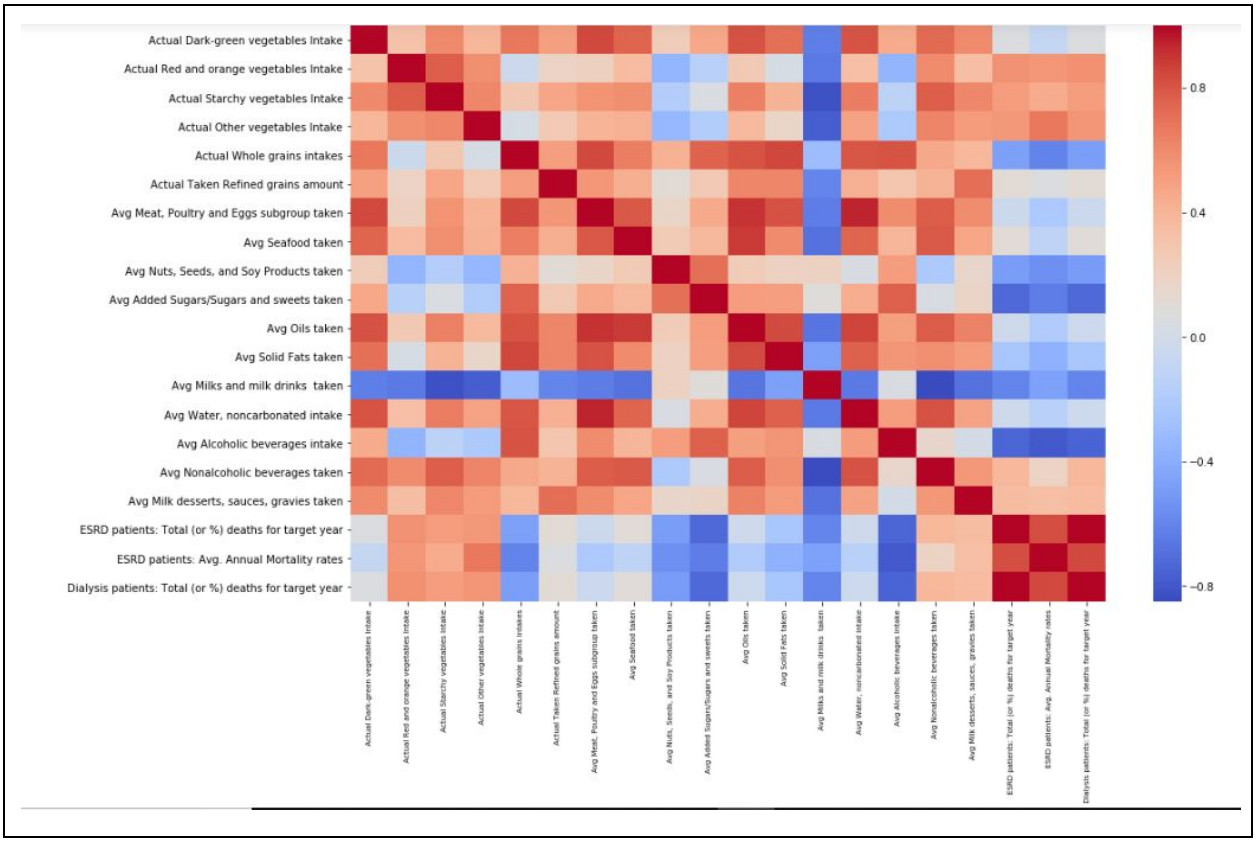
PC7 .... PC18

- Comp 3 to comp 4 has the most change for slope
- First three or at best first 4 components can be retained



Food subgroups contributing to the first four principal components

Correlation of all the features using heatmaps are given below. From the above figure, only features contributing highly to the first four components could be shown in heatmaps.



Regression Analysis with Excel

Regression on Food Groups

Related attached file: june-19-regression_mr_and_analysis.xlsx
Related Worksheet Sheet: food group 95

Regression Statistics

Multiple R	0.880156954
R Square	0.774676264
Adjusted R Square	0.616949648
Standard Error	6.223447127
Observations	18

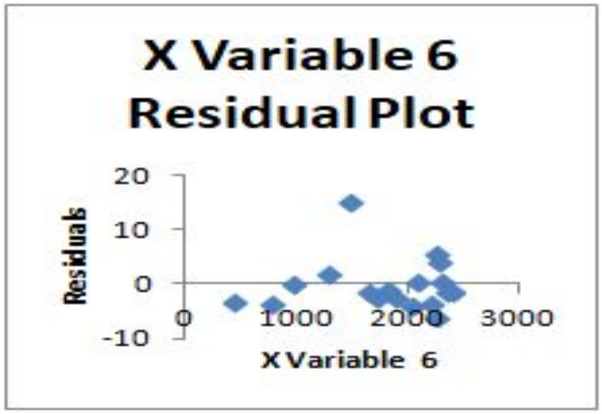
R square can be seen as significant i.e. explains relations and variations

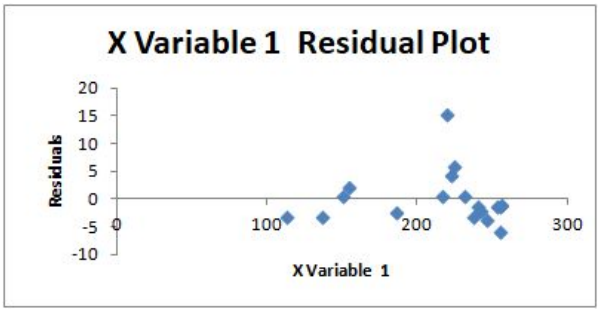
Coefficients and p values

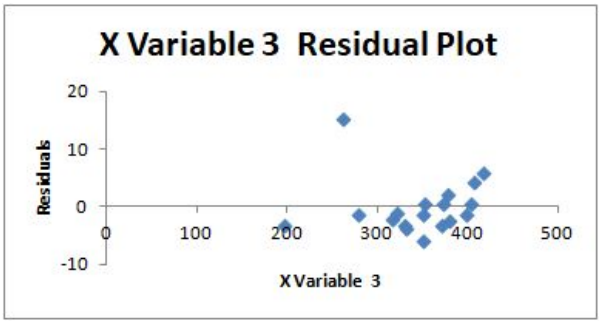
		Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
	Intercept	44.479	95.829	0.464	0.652	-169.041	257.999	-169.041	257.999
Actual Vegetable Intake	**X Variable 1**	**0.191**	**0.165**	**1.155**	**0.275**	**-0.177**	**0.559**	**-0.177**	**0.559**
Actual Fruit intakes	X Variable 5	0.031	0.169	0.183	0.858	-0.345	0.407	-0.345	0.407
Avg Fats oils and salad dressings take	X Variable 7	0.012	0.978	0.013	0.990	-2.168	2.192	-2.168	2.192
Actual Taken Sugars sweets and beve	**X Variable 6**	**-0.015**	**0.013**	**-1.118**	**0.290**	**-0.045**	**0.015**	**-0.045**	**0.015**
Actual Protein Intake	X Variable 2	-0.023	0.154	-0.152	0.882	-0.367	0.320	-0.367	0.320
Actual Dairy Intake	X Variable 4	-0.085	0.099	-0.852	0.414	-0.306	0.137	-0.306	0.137
Actual Grain Intake	**X Variable 3**	**-0.086**	**0.066**	**-1.298**	**0.223**	**-0.233**	**0.062**	**-0.233**	**0.062**

Grain, Sugar, Vegetables seem to affect mortality based on Coefficients and P values [33-48] . The interpretation from the references were used. P value though does not indicate strong significance.

Residual Plots for the affecting variables







Regression on Food Subgroup

Related attached file: june-19-regression_mr_and_analysis.xlsx
Related Worksheet Sheet: subgroup 95 affecting

Regression Statistics

Multiple R	0.999378722
R Square	**0.99875783**
Adjusted R Square	0.978883104
Standard Error	1.461167293
Observations	18

R Square is 99%

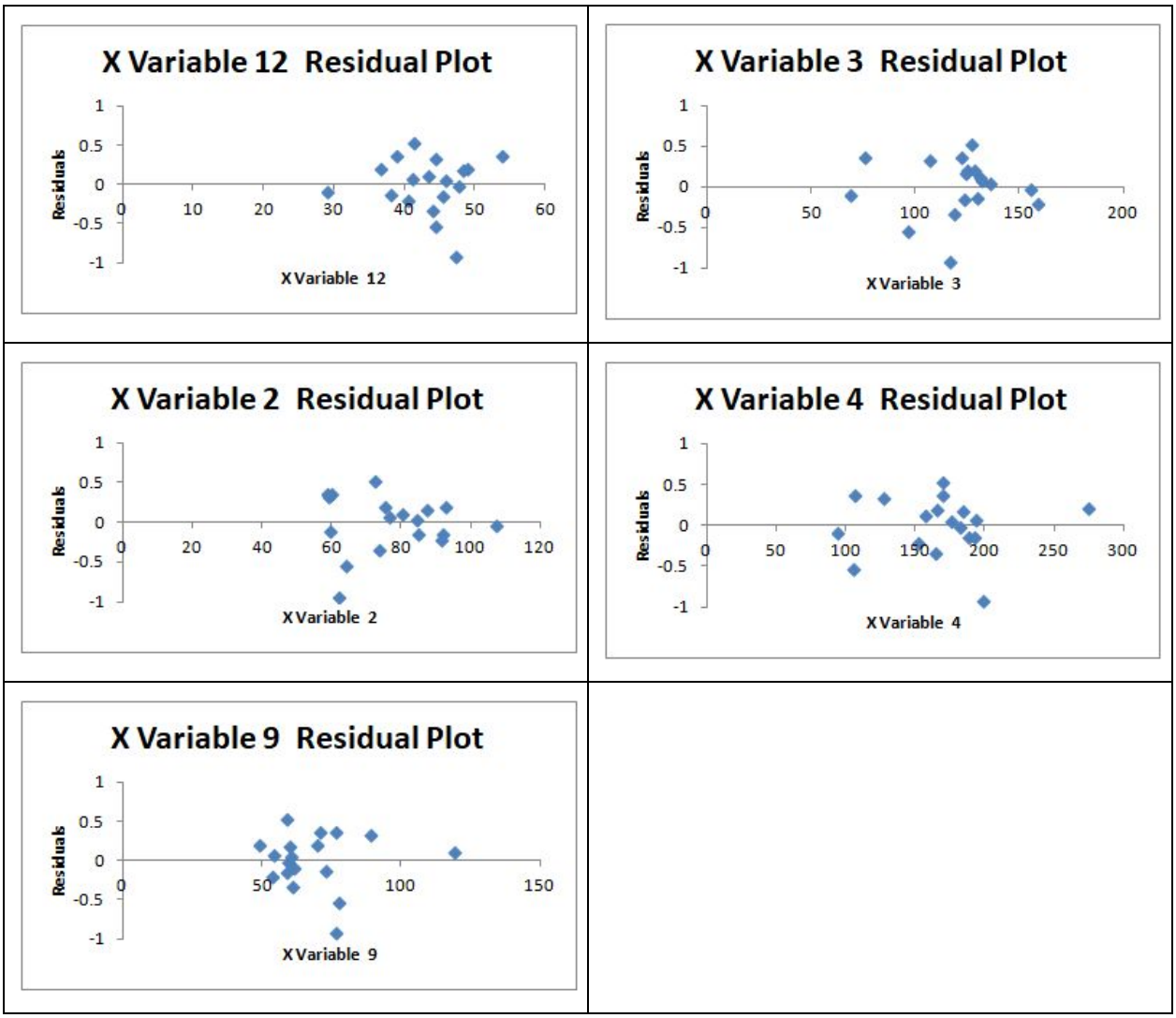
Regression Coefficients

		Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
	Intercept	1.372	54.715	0.025	0.984	-693.852	696.595	-693.852	696.595
Actual Starchy vegetables Intake	**X Variable 3**	**0.310**	**0.135**	**2.299**	**0.261**	**-1.404**	**2.024**	**-1.404**	**2.024**
Actual Other vegetables Intake	**X Variable 4**	**0.159**	**0.069**	**2.287**	**0.262**	**-0.722**	**1.039**	**-0.722**	**1.039**
Avg Meat, Poultry and Eggs subgroup taken	X Variable 7	0.144	0.111	1.304	0.416	-1.263	1.552	-1.263	1.552
Actual Dark-green vegetables Intake	X Variable 1	0.098	0.121	0.807	0.568	-1.443	1.639	-1.443	1.639
Avg Added Sugars/Sugars and sweets taken	X Variable 10	0.047	0.091	0.520	0.695	-1.110	1.205	-1.110	1.205
Actual Whole grains intakes	X Variable 5	0.036	0.114	0.312	0.807	-1.416	1.488	-1.416	1.488
Actual Taken Refined grains amount	X Variable 6	0.027	0.108	0.255	0.841	-1.339	1.394	-1.339	1.394
Avg Nonalcoholic beverages taken	X Variable 16	0.014	0.010	1.414	0.392	-0.115	0.143	-0.115	0.143
Avg Milks and milk drinks taken	X Variable 13	-0.001	0.079	-0.008	0.995	-1.009	1.008	-1.009	1.008
Avg Alcoholic beverages intake	**X Variable 15**	**-0.016**	**0.008**	**-1.949**	**0.302**	**-0.118**	**0.087**	**-0.118**	**0.087**
Avg Water, noncarbonated intake	**X Variable 14**	**-0.032**	**0.014**	**-2.288**	**0.262**	**-0.211**	**0.147**	**-0.211**	**0.147**
Avg Seafood taken	X Variable 8	-0.052	0.071	-0.742	0.594	-0.951	0.846	-0.951	0.846
Avg Nuts, Seeds, and Soy Products taken	**X Variable 9**	**-0.153**	**0.068**	**-2.245**	**0.267**	**-1.016**	**0.711**	**-1.016**	**0.711**
Actual Red and orange vegetables Intake	**X Variable 2**	**-0.245**	**0.075**	**-3.248**	**0.190**	**-1.202**	**0.712**	**-1.202**	**0.712**
Avg Solid Fats taken	**X Variable 12**	**-0.583**	**0.389**	**-1.497**	**0.375**	**-5.528**	**4.363**	**-5.528**	**4.363**
Avg Oils taken	X Variable 11	-0.594	0.668	-0.889	0.538	-9.087	7.899	-9.087	7.899

Affecting Factors

Affecting factors		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Avg Solid Fats taken	**X Variable 12**	**-0.583**	**0.389**	**-1.497**	**0.375**	**-5.528**	**4.363**	**-5.528**	**4.363**
Actual Starchy vegetables Intake	**X Variable 3**	**0.310**	**0.135**	**2.299**	**0.261**	**-1.404**	**2.024**	**-1.404**	**2.024**
Actual Red and orange vegetables Intake	**X Variable 2**	**-0.245**	**0.075**	**-3.248**	**0.190**	**-1.202**	**0.712**	**-1.202**	**0.712**
Actual Other vegetables Intake	**X Variable 4**	**0.159**	**0.069**	**2.287**	**0.262**	**-0.722**	**1.039**	**-0.722**	**1.039**
Avg Nuts, Seeds, and Soy Products taken	**X Variable 9**	**-0.153**	**0.068**	**-2.245**	**0.267**	**-1.016**	**0.711**	**-1.016**	**0.711**

Residual Plots for the affecting factors



Note: One predictor variable is not used here as limit for Excel 2007 is 16 predictors.

Future work for Association and Exploratory Analysis

Principal Component Analysis (PCA) is being explored on the dataset. Further work are subject to be done with PCA.

Deep Learning might be explored to find association and compare with other methods used

- Put all dietary intake data from all the years and provide data for each age as well as keep all subgroups and then let the Deep Learning methods to identify the most affecting factors

Clustering algorithms or Decision trees might be an option to classify the data into clusters of affecting factors or to find a food group/subgroup combinations that affect mortality. This research might explore that and compare with other approaches as used.

Ensemble/Boosting methods with voting where each year can be studied separately will be an option to explore

More Univariate, Bivariate, and multivariate exploration/analysis might be done in future work.

- Some are also provided/done as part of the Python scripts attached
- Some are also provided/done as part of the SQL scripts and Stored Procedures attached

Cox's regression (hazard model or survival model) is an option to explore on the dataset

Appendix

Address: Age groups are different in the intake recommendation and USRDS/NHANES data

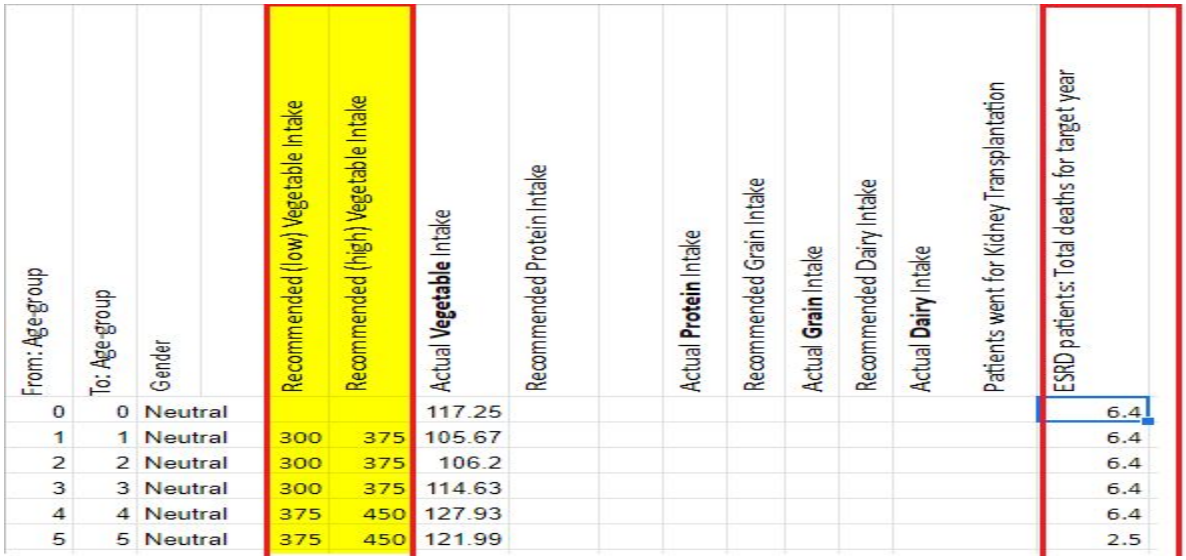
At this point, as a challenge to find out the recommended intake amount by USRDS age groups is there, a different approach can be taken as follows as given in the image below.

Age Based Approach

Here, no age groups is used, only individual ages are used. NHANES survey data is grouped by each age for food intake. In the same way, recommended intake amount as provided in the shift recommendation article [11] is also converted for each age such as the recommended amount for ages 1 to 3 is kept the same for each age 1, 2, and 3. For the USRDS mortality data a similar approach was taken. When the mortality is given as total count for the age groups, the count is divided equally to each age in that range. For example the mortality count for ages 0 to 4 is divided by 5 to get the count for each age from 0 to 4.

The recommended intake amount by health.gov in the shift recommendation study [11] and in general is given by genders. An average of the gender based amount are utilized. The amounts for each gender also appeared to be the same in the recommendation. The measure cup was used by health gov. 1 cup = 150gm is used in the image below.

As data are used from different sources where the participants are different for NHANES and USRDS studies. The total count in mortality will be normalized using percentages.



From: Age-group	To: Age-group	Gender	Recommended (low) Vegetable Intake	Recommended (high) Vegetable Intake	Actual **Vegetable** Intake	Recommended Protein Intake	Actual **Protein** Intake	Recommended Grain Intake	Actual **Grain** Intake	Recommended Dairy Intake	Actual **Dairy** Intake	Patients went for Kidney Transplantation	ESRD patients: Total deaths for target year
0	0	Neutral			117.25								6.4
1	1	Neutral	300	375	105.67								6.4
2	2	Neutral	300	375	106.2								6.4
3	3	Neutral	300	375	114.63								6.4
4	4	Neutral	375	450	127.93								6.4
5	5	Neutral	375	450	121.99								2.5

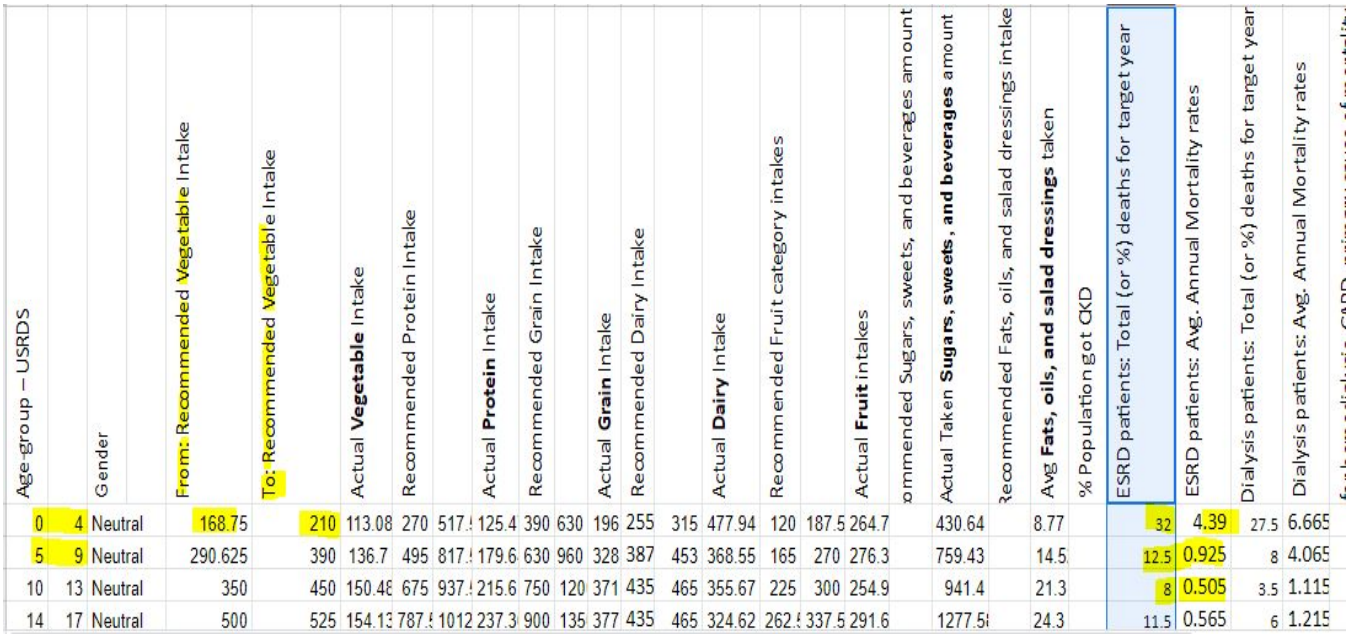
The complete data for the image above can be seen on Google Drive

Food group based dietary intake data for each age can be seen at: age-based-avg-intake

USRDS mortality data as used in the image above can be seen at: Age-based-USRDS-Mortality . i.e. H1 sheet from the excel file [22] at : USRDS Mortality

Age Group Based Approach

In another approach: the recommended amount by age groups was distributed to each age, then regrouped those to reflect the age groups of USRDS. The file list table will show where those data are kept (Files as submitted: mortality_recom_added_group_data_june_9th_gender_based_data_after_processing).



Age-group – USRDS		Gender	From: Recommended Vegetable Intake	To: Recommended Vegetable Intake	Actual **Vegetable** Intake	Recommended Protein Intake		Actual **Protein** Intake	Recommended Grain Intake		Actual **Grain** Intake	Recommended Dairy Intake		Actual **Dairy** Intake	Recommended Fruit category intakes		Actual **Fruit** intakes	ommended Sugars, sweets, and beverages amount	Actual Taken **Sugars, sweets, and beverages** amount	Recommended Fats, oils, and salad dressings intake	Avg **Fats, oils, and salad dressings** taken	% Population got CKD	ESRD patients: Total (or %) deaths for target year	ESRD patients: Avg. Annual Mortality rates	Dialysis patients: Total (or %) deaths for target year	Dialysis patients: Avg. Annual Mortality rates
0	4	Neutral	168.75	210	113.08	270	517.5	125.4	390	630	196	255	315	477.94	120	187.5	264.7		430.64		8.77		32	4.39	27.5	6.665
5	9	Neutral	290.625	390	136.7	495	817.5	179.6	630	960	328	387	453	368.55	165	270	276.3		759.43		14.5		12.5	0.925	8	4.065
10	13	Neutral	350	450	150.48	675	937.5	215.6	750	120	371	435	465	355.67	225	300	254.9		941.4		21.3		8	0.505	3.5	1.115
14	17	Neutral	500	525	154.13	787.5	1012	237.3	900	135	377	435	465	324.62	262.5	337.5	291.6		1277.5		24.3		11.5	0.565	6	1.215

Details on Food subgroup assignment

Note: I will continuously seek for improvement in assignment. Also, in some cases experiment might be done by using the food item to different subgroups esp. when it is a mixed food item, or assignment varies across studies

From food code subgroups [12] if I could find that in the shift recommendation - I used that subgroup. Otherwise, I found out the closest matching subgroup, or I divided that food item to sub-subgroup to assign to subgroups as appropriate or created a new subgroup as the last resort. Using these, I could assign all NHANES survey [10] rows to a corresponding subgroup (whether used in the shift recommendation article or not).

Some subgroup assignments when exact match not found
[Note: For USDA food codes couple of initial digits indicate the subgroup]

- Subgroup: 27 and 28 assigned to protein though they are mixed food
    - 27 Meat, poultry, fish with non meat items
    - 271 Meat, poultry, fish in gravy or sauce
    - 28 Frozen and shelf-stable plate meals, soups, and gravies with meat, poultry, fish base; gelatin and gelatin-based drinks
- 77, and 78 currently assigned to other vegetables even though they are actually mixed food item
- 12 can be assigned to saturated Fats and Solid Fats. Used Solid Fat as the shift recommendation article used this measure.
- I used 95 for added sugars subgroup : Formulated nutrition beverages, energy drinks, sports drinks, functional beverages
- 92: created new subgroup: Nonalcoholic beverages: It could be part of Added sugar subgroup
- 81 : is kept under : solid fats: it could be Saturated Fats as well though Saturated Fats is not part of shift recommendation article

- 14 - cheese is used for solid fat, can be saturated fat as well
- 99991400: cheese as an ingredient in sandwich : is assigned under solid fat subgroup: it might have a side effect. I need to check the gm (amount) - does it make sense
- 99998 : Assigned to oils
    - 99998130 sauce as ingredient in hamburg
    - 99998210 industrial oil as ingredient
- 99995: assigned to whole grain
    - 99995000 breading or batter as ingredie
    - 99995130 wheat bread as an ingredient in s

Food Group Assignments: to experiment at Analysis

12 Creams and cream substitutes
I assigned Solid Fats as the subgroup name for the following. **Though the Group name is kept Dairy according to food code.**

The group can be changed to Fats as well. Fat is a shift recommendation primary group. Calculation will also get affected - the intake will count towards dairy when analysis use group based, and then towards solid fats when calculating subgroup based. I am biased towards changing the group to Fats so that amounts are calculated properly

12 Creams and cream substitutes
    121 Sweet dairy cream
    122 Cream substitutes
    123 Sour cream

74 Tomatoes and tomato mixtures
was assigned to other vegetables

Nuts, Seeds, and Soy Products
I have used them as subgroups of Protein - same with shift recommendation though I used Vegetable as the group based on the food code. I am biased to change the primary group to Protein

Though I used group name to be vegetables as usda code starting with 4 belongs to Vegetables (Legumes and Peas). I am biased to use Protein as Group name for nuts/seeds subgroup. I can/want to keep legumes/peas under Vegetable group and **Other vegetable** subgroup.

95 Formulated nutrition beverages, energy drinks, sports drinks, functional beverages
Assigned to Added Sugars. It can be its own subgroup though shift recommendation article does not have this (i.e Formulated nutrition beverages, energy drinks, sports drinks, functional beverages) subgroup

Sample Data

Subgroup based average food item intake:

Grouped as By Age Groups and then by Food SubGroups:

File: https://drive.google.com/file/d/1rkzyDVK9034HHVV7o1gBJzNtOWWRK4wE/view?usp=sharing

no_of_participants	min_age_for_group	max_age_for_group	age_group_id	food_subgroup_id	food_subgroup_name	sum_taken_in_gms	avg_taken_in_gms	m
647	0	4	1	6	Starchy vegetables	61265.23	94.69	S
929	0	4	1	10	Whole grains	182314.02	196.25	W
514	0	4	1	14	Seafood	39861.87	77.55	S
189	0	4	1	15	Nuts, Seeds, and Soy Products	13837	73.21	N
196	0	4	1	18	Egg	13765.54	70.23	E
542	5	9	2	6	Starchy vegetables	58368.35	107.69	S
838	5	9	2	10	Whole grains	281464.51	335.88	W
534	5	9	2	14	Seafood	55011.87	103.02	S
193	5	9	2	15	Nuts, Seeds, and Soy Products	17887.79	92.68	N
149	5	9	2	18	Egg	13969.27	93.75	E

Grouped as food subgroup and then by age group
File: https://drive.google.com/file/d/1aKYozwBrXxyLe4HyGSjGjnuG6EbU_4Ul/view?usp=sharing

no_of_participants	min_age_for_group	max_age_for_group	age_group_id	food_subgroup_id	food_subgroup_name	sum_taken_in_gms	avg_taken_in_gms	map_subgroup
647	0	4	1	6	Starchy vegetables	61265.23	94.69	Starchy vegetables
542	5	9	2	6	Starchy vegetables	58368.35	107.69	Starchy vegetables
417	10	13	3	6	Starchy vegetables	49919.81	119.71	Starchy vegetables
388	14	17	4	6	Starchy vegetables	50120.8299999999	129.18	Starchy vegetables
248	18	21	5	6	Starchy vegetables	37862.28	152.67	Starchy vegetables
165	22	24	6	6	Starchy vegetables	32470.57	196.79	Starchy vegetables
317	25	29	7	6	Starchy vegetables	57938.48	182.77	Starchy vegetables
307	30	34	8	6	Starchy vegetables	55190.88	179.77	Starchy vegetables
305	35	39	9	6	Starchy vegetables	59012.34	193.48	Starchy vegetables
307	40	44	10	6	Starchy vegetables	55555.76	180.96	Starchy vegetables
304	45	49	11	6	Starchy vegetables	59665.88	196.27	Starchy vegetables
314	50	54	12	6	Starchy vegetables	55387.69	176.39	Starchy vegetables
270	55	59	13	6	Starchy vegetables	53890.35	199.59	Starchy vegetables
353	60	64	14	6	Starchy vegetables	63456.06	179.76	Starchy vegetables
272	65	69	15	6	Starchy vegetables	52964.46	194.72	Starchy vegetables
215	70	74	16	6	Starchy vegetables	44185.84	205.52	Starchy vegetables
156	75	79	17	6	Starchy vegetables	33508.78	214.8	Starchy vegetables
226	80	80	18	6	Starchy vegetables	40159.27	177.7	Starchy vegetables

Notes and Process used to generate the data above:

Used food subgroup codes from:
https://reedir.arsnet.usda.gov/codesearchwebapp/(gcp3kq55ssdyc445ry2k2rus)/coding_scheme.pdf

Mapping shift recommendation subgroups to USDA subgroup codes:

id	food_group_name	is_parent	parent_gr...	usda_grou...	is_in_usda	usda_subgro...	usda_subg...	usda_...	usda_s...	usda_subgr...	usda_subg...
1	Vegetables	1	1	7	1	7	NULL	NULL	NULL	NULL	NULL
2	Dark-green vegetables	0	1	7	0	72	NULL	NULL	NULL	NULL	NULL
3	Red and orange vegeta...	0	1	7	0	73	NULL	NULL	NULL	NULL	NULL
4	Legumes (beans and p...	0	1	4	1	41	NULL	NULL	NULL	NULL	NULL
5	Starchy vegetables	0	1	7	0	71	NULL	NULL	NULL	NULL	NULL
6	Other vegetables	0	1	7	0	75	76	78	NULL	NULL	NULL
7	Fruits	1	7	6	1	6	NULL	NULL	NULL	NULL	NULL
8	Grains	1	8	5	1	5	NULL	NULL	NULL	NULL	NULL
9	Whole grains	0	8	5	0	50	51	56	57	58	59
10	Refined grains	0	8	5	0	52	53	54	55	NULL	NULL
11	Dairy	1	11	1	1	1	NULL	NULL	NULL	NULL	NULL
12	Protein	1	12	2	1	2	NULL	NULL	NULL	NULL	NULL
13	Meat, Poultry and Eggs...	0	12	2	0	20	21	22	23	24	25
14	Seafood	0	12	2	0	26	NULL	NULL	NULL	NULL	NULL
15	Nuts, Seeds, and Soy Pr...	0	12	2	0	42	43	NULL	NULL	NULL	NULL
16	Sugars, sweets, and bev...	1	16	9	1	9	NULL	NULL	NULL	NULL	NULL
17	Fats, oils, and salad dre...	1	17	8	1	8	NULL	NULL	NULL	NULL	NULL
18	Egg	0	12	3	1	31	32	33	35	NULL	NULL

Sample Mapping of Foods to Subgroups (Will adjust for better mapping)

id	usda_food_code	food_name	group_id	sub_group_id	group_name	subgroup_name
6118	72101100	BEET GREENS, RAW	1	2	Vegetables	Dark-green vegetables
6119	72101200	BEET GREENS, COOKED, NS AS TO	1	2	Vegetables	Dark-green vegetables
6120	72101210	BEET GREENS, COOKED, FAT NOT A	1	2	Vegetables	Dark-green vegetables
6446	73102202	CARROTS, COOKED, FROM FROZEN,	1	3	Vegetables	Red and orange veg...
6447	73102203	CARROTS, COOKED, FROM CANNED,	1	3	Vegetables	Red and orange veg...
6448	73102210	CARROTS, COOKED, NS AS TO FORM	1	3	Vegetables	Red and orange veg...
6449	73102211	CARROTS, COOKED, FROM FRESH, F	1	3	Vegetables	Red and orange veg...
2802	58146302	PASTA WITH TOMATO-BASED SAUCE,	8	9	Grains	Whole grains
2803	58146303	PASTA WITH TOMATO-BASED SAUCE,	8	9	Grains	Whole grains
2804	58146315	PASTA WITH SAUCE AND MEAT, FRO	8	9	Grains	Whole grains
2805	58146321	PASTA WITH TOMATO-BASED SAUCE	8	9	Grains	Whole grains

SQL Code as used to generate the SubGroup Based data:
Folder: https://drive.google.com/drive/folders/1HmvdmEILYQDUDvxwnIHyl8M404iv-3FY?usp=sharing
For Mapping: assign_subgroup_to_food_items.sql
Data Aggregation: get_food_subgroup_based_dietary_intake_by_participants.sql

Steps taken in Data Exploration

1. Explored USRDS data on CKD and ESRD patients such as
    a. Patient characteristics, Mortality, Survival, Dietician care or not
    b. Adjusted data and non-adjusted data
2. Dietician Care received or not and mortality/survival/remaining life/Got into ESRD got attention of focus
    a. However, the target variables such as mortality/survival linked to patients under consideration. No link is there between dietician care data and target/censoring variables

    b. Corresponding visualizations are provided on: **data exploration on dietician care.docx**
3. USRDS public data was aggregated; individual patient data and characteristics including mortality/survival and link to dietician care or not was missing
4. Requested individual patient data (where each patient whether received dietician care or not will also be mentioned; also links to target variable can be made for individual patients)
    a. Though got the data format that showed all required data will be there; however, that required special request and permission including involving ethical bodies. That also require couple of months of time to fulfill the request.
    b. Hence, abandoned the idea to get those data
5. Then Dietary shift recommendation dataset explored where recommendation for each age group including average intake amount by age groups and food groups/subgroups were provided.
    a. However, there was a missing link: Age groups between Dietary shift recommendation dataset and USRDS mortality/target variable data were not aligned
6. Then dietary intake data from NHANES survey for each participant were explored. The same survey data was used by Dietary shift recommendation dataset. Hence, regrouped NHANES survey data to reflect the USRDS age groups
    a. USDS codes was used for Food groups/subgroups/intake food for NHANES survey. Hence, survey food intake was mapped using USDA food codes.
7. Still a missing link was there. The age groups used in the recommended amount of intake by CDC/health.gov/Dietary shift recommendation dataset did not match with USRDS i.e. also with age-groups in the newly regrouped average intake
    a. Then survey data was averaged for each age. Also, mortality/survival data was divided for each age - this may or may not be used. As analysis using recommendation will be a secondary analysis in the research
    b. In another approach: the recommended amount by age groups was distributed to each age, then regrouped those to reflect the age groups of USRDS. The file list table will show where those data are kept.

List of files submitted

File Name	**Purpose**
data exploration on dietician care.docx	Plotted data on patients received dietician data based on age groups, races, gender and similar. At this point, this exploration does not seem to be important for our final analysis
grouped-diet-data	Shows the SQLs used and the grouped data from NHANES survey. Not that important, just shows the steps in SQLs and sample data aggregation. This has changed a lot.
csvdietfiles.zip	Has csv files that provided recommended intake amount for food groups and subgroups. Data are

	provided by health.gov/CDC Python code used: extract_data_for_diets.ipynb
dietfiles.zip	Txt files that provided recommended intake amount for food groups and subgroups. Data are provided by health.gov/CDC
agescsvdietfiles.zip	csv files that provides average recommended amount for each age. Male and Female recommendations are averaged. Original data was for age groups, here data are converted for each age. This will help to analyze for each age or with custom age-groups to match with USRDS age groups. Original data (dietfiles.zip) was in cups, converted to gms 1 cup = 150 gms. The Male or Female in the csv file name has no significance .. averaged with Male and Female data **Python code used:** ages_extract_data_for_diets.ipynb
recommended_for_each_age.csv	For each age, recommended average intake amount, all food groups are here, recommendation low to high amount in gms
regroup_ages_food_intake_recommendations	Age groups do not match between CDC (health.gov) and USRDS. Hence, rearranged age groups and calculated average recommended intake to reflect USRDS age groups. **Calculated over** recommended_for_each_age.csv . The methodology whether appropriate or not will be justified. And this will be used for the second step to find complying with or not with the recommendation - how that affects. First step is to find the affecting food groups and subgroups **Related SQL Server Stored Procedure:** regroup_ages_food_intake_recommendations
mortality_recom_added_group_data_june_9th_gender_based_data_after_processing.xlsx	Data for food group based analysis Actual intake data by the population from NHANES survey

	USRDS mortality Data Recommended intake data from CDC/Health.gov (age groups aligned with USRDS data)
data_helping_with_food_grouping_subgrouping.zip	Meta data on age groups, usda food codes, mapping NHANES survey food items to USDA food groups and subgroups.
multi-day-aggregated-dietary-data.zip	Gender based and Gender Neutral: Dietary intake data by age groups, food groups/subgroups
mortality_recom_added_group_data_june_9th_gender_based_data_after_processing, mortality_group_data_june_9th_gender_based_data_after_processing.xlsx, mortality_subgroup_data_june_9th_gender_based_data_after_processing.xlsx	To relate dietary data to mortality/survival data, related data are put together. Age-group based dietary intake and age group based mortality/survival data are kept side by side
remaining_group_data_june_9th_gender_based_data_after_processing.xlsx	To relate dietary data to remaining life data measures, related data are put together. Age-group based dietary intake and age group based mortality/survival data are kept side by side Might not be analyzed
SQL scripts for tables and stored procedures.zip	All database tables, views, stored procedures as used for data exploration and data generation
Python scripts.zip	Python scripts as used for data collection, processing, data cleaning, data adjustments, and data exploration
foodgroup-ckd-mortality.ipynb	Exploratory Analysis. Regression, Correlation, Heatmaps for Food Group based analysis
food-subgroup-ckd-mortality.ipynb	Exploratory Analysis. Regression, Correlation, Heatmaps for Food Sub Group based analysis

Exploratory Analysis.zip	Input/output csv/excel files for Python scripts foodgroup-ckd-mortality.ipynb, food-subgroup-ckd-mortality including ipynb files
Foodgroup-ckd-mortality.ipynb foodgroup-ckd-mortality.pdf	Univariate/Bivariate analysis and visualizations for food groups
Food-subgroup-ckd-mortality.ipynb food-subgroup-ckd-mortality.pdf	Univariate/Bivariate analysis and visualizations For Food Subgroups
pca_univariate_bivariate.zip	png images as saved from univariate, bivariate, and PCA exploration. However, only a few images are here, all the output images can be seen as part of the ipynb files

References

1. What Is Chronic Kidney Disease?
2. Jaimon T. Kelly, Suetonia C. Palmer, Shu Ning Wai, Marinella Ruospo, Juan-Jesus Carrero, Katrina L. Campbell, and Giovanni F. M. Strippol Healthy Dietary Patterns and Risk of Mortality and ESRD in CKD: A Meta-Analysis of Cohort Studies
3. Chen X, Wei G, Jalili T, Metos J, Giri A, Cho ME, Boucher R, Greene T, Beddhu S: The associations of plant protein intake with all-cause mortality in CKD. Am J Kidney Dis 67: 423–430, 2016 (26)
4. Gutie´rrez OM, Muntner P, Rizk DV, McClellan WM, Warnock DG, Newby PK, Judd SE: Dietary patterns and risk of death and progression to ESRD in individuals with CKD: A cohort study. Am J Kidney Dis 64: 204–213, 2014 (27)
5. Huang X, Jime´nez-Moleo´n JJ, Lindholm B, Cederholm T, Arnold ¨v J, Rise´rus U, Sjo¨gren P, Carrero JJ: Mediterranean diet, kidney function, and mortality in men with CKD. Clin J Am Soc Nephrol 8: 1548–1555, 2013 (28)
6. Muntner P, Judd SE, Gao L, Gutie´rrez OM, Rizk DV, McClellanW, Cushman M, Warnock DG: Cardiovascular risk factors in CKD associated with both ESRD and mortality. J Am Soc Nephrol 24: 1159–1165, 2013 (29)
7. Ricardo AC, Madero M, Yang W, Anderson C, Menezes M, Fischer MJ, Turyk M, Daviglus ML, Lash JP: Adherence to a healthy lifestyle and all-cause mortality in CKD. Clin J Am Soc Nephrol 8: 602–609, 2013 (30)
8. Tsuruya K, Fukuma S, Wakita T, Ninomiya T, Nagata M, Yoshida H, Fujimi S, Kiyohara Y, Kitazono T, Uchida K, Shirota T, Akizawa T, Akiba T, Saito A, Fukuhara S: Dietary patterns and clinical outcomes in hemodialysis patients in Japan: A cohort study. PLoS One 10: e0116677, 2015 (31)
9. Ricardo AC, Anderson CA, Yang W, Zhang X, Fischer MJ, Dember LM, Fink JC, Frydrych A, Jensvold NG, Lustigova E, Nessel LC, Porter AC, Rahman M, Wright Nunes JA, Daviglus ML, Lash JP; CRIC Study Investigators: Healthy lifestyle and risk of kidney disease progression, atherosclerotic events, and death in CKD: Findings from the Chronic Renal Insufficiency Cohort (CRIC) Study. Am J Kidney Dis 65: 412–424, 2015 (17)
10. National Health and Nutrition Examination Survey

11. https://health.gov/dietaryguidelines/2015/guidelines/chapter-2/a-closer-look-at-current-intakes-and-recommended-shifts/
12. Food Code Numbers and the Food Coding Scheme
13. VEGETABLE SUBGROUPS
14. Key Concepts About the USDA Food Coding Scheme:
15. Appendix 3. USDA Food Patterns: Healthy U.S.-Style Eating Pattern.
16. United states Renal Data System (USRDS). 2018 ADR Reference Tables:
17. 2018 ADR Chapters.
18. Documentation and Dataset:
19. 2018 USRDS Annual Data Report: Executive Summary
20. DIETARY_GUIDELINES_FOR AMERICANS_2015-2020
21. https://health.gov/dietaryguidelines/dga95/9DIETGUI.HTM
22. Mortality and Causes of Death
23. Tong A, Chando S, Crowe S, Manns B, Winkelmayer WC, Hemmelgarn B, Craig JC: Research priority setting in kidney disease: A systematic review. Am J Kidney Dis 65: 674–683, 2015
24. Lin J, Fung TT, Hu FB, Curhan GC: Association of dietary patterns with albuminuria and kidney function decline in older white women: A subgroup analysis from the Nurses' Health Study. Am J Kidney Dis 57: 245–254, 2011
25. Taylor EN, Fung TT, Curhan GC: DASH-style diet associates with reduced risk for kidney stones. J Am Soc Nephrol 20: 2253–2259, 2009
26. Liu, Hao-Wen; Tsai, Wen-Hsin; Liu, Jia-Sin; Kuo, Ko-Lin. 2019. "Association of Vegetarian Diet with Chronic Kidney Disease." *Nutrients* 11, no. 2: 279.
27. Golaleh Asghari, Mehrnaz Momenan, Emad Yuzbashian, Parvin MirmiranEmail author and Fereidoun Azizi. Dietary pattern and incidence of chronic kidney disease among adults: a population-based study
28. Tanushree Banerjee1 , Deidra C. Crews2 , Delphine S. Tuot3 , Meda E. Pavkov4 , Nilka Rios Burrows4 , Austin G. Stack5 , Rajiv Saran6,7 , Jennifer Bragg-Gresham6 and Neil R. Powe1,8 ; for the Centers for Disease Control and Prevention Chronic Kidney Disease Surveillance Team9 Poor accordance to a DASH dietary pattern is associated with higher risk of ESRD among adults with moderate chronic kidney disease and hypertension
29. Jacek R., Beata F., Aleksandra C., Anna G.The Effect of Diet on the Survival of Patients with Chronic Kidney Disease. *Nutrients* 2017, *9*(5), 495; https://doi.org/10.3390/nu9050495
30. https://www.kidney.org/news/one-seven-american-adults-estimated-to-have-chronic-kidney-disease
31. https://www.medicalnewstoday.com/articles/282929.php
32. Five Stages of CKD. https://www.davita.com/education/kidney-disease/stages
33. Goal: How to Identify the Most Important Predictor Variables in Regression Models https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-identify-the-most-important-predictor-variables-in-regression-models
34. How to Interpret Regression Analysis Results: P-values and Coefficients https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients
35. Regression Analysis: How to Interpret the Constant (Y Intercept) https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-to-interpret-the-constant-y-intercept

36. How to Compare Regression Slopes: How to statistically test the difference between regression slopes and constants https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-compare-regression-lines-between-different-models
37. How Do I Interpret R-squared and Assess the Goodness-of-Fit? https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
38. How High Should R-squared Be in Regression Analysis? https://blog.minitab.com/blog/adventures-in-statistics-2/how-high-should-r-squared-be-in-regression-analysis
39. How to Interpret a Regression Model with Low R-squared and Low P values https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values
40. Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables https://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables
41. How to Interpret S, the Standard Error of the Regression https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-to-interpret-s-the-standard-error-of-the-regression
42. What Is the F-test of Overall Significance in Regression Analysis? https://blog.minitab.com/blog/adventures-in-statistics-2/what-is-the-f-test-of-overall-significance-in-regression-analysis
43. Understanding Analysis of Variance (ANOVA) and the F-test https://blog.minitab.com/blog/adventures-in-statistics-2/understanding-analysis-of-variance-anova-and-the-f-test
44. How to Compare Regression Slopes https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-compare-regression-lines-between-different-models
45. How to Present and Use the Results to Avoid Costly Mistakes, part 1 https://blog.minitab.com/blog/adventures-in-statistics-2/applied-regression-analysis-how-to-present-and-use-the-results-to-avoid-costly-mistakes-part-1
46. How to Identify the Most Important Predictor Variables in Regression Models https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-identify-the-most-important-predictor-variables-in-regression-models
47. **How to Interpret your Regression Results** http://sitestree.com/how-to-interpret-your-regression-results/
48.