**Biostatistics 202: Opportunities and Challenges of Complex Biomedical Data: Introduction to the Science of “Big Data” Final Project**

Research Question: What factors contribute to differences in reported daily whole grain intake in the NHANES Dataset?

Description: The National Health and Nutrition Examination Survey (NHANES) is a program designed to assess the health and nutritional status of adults and children in the United States of America. The aim of this project is to explore the factors that contribute to differences in reported daily whole grain intake in the NHANES Dataset. This research question stemmed from an article relating nutrition and SARS-Cov/Covid-19 (1). The article suggests that the health status of individuals is based on their diet. A good diet gives people a fighting chance once infected by Covid-19.

In this project, we will develop a model that uses unsupervised learning methods to explore some underlying structure in the NHANES data by capturing the variables that best explain the top clusters that are created. The clusters are then related to daily whole grain intake.

Details:

The following are the steps taken to clean and analyze the NHANES dataset.

**Data Importation and Cleaning**

1. Import the “Project 3 – NHANES Data.csv” file into Orange Data Mining Software using the “File” widget. Of the 50 variables in the NHANES Dataset, 17 variables including the predwhgr variable are set as features and the rest are set as meta variables. (See appendix for the list of variables used) All variables selected are demographic features that will be used to determine the clusters. Non-demographic variables are excluded from analysis. The age (ridageyr), predwhgr and ratio of family income to poverty (indfmpir) variables are numeric features and the other 14 are categorical features.
2. In the “File” widget, we assign the variables to the measurement level consistent with the data dictionary. The categorical variables are assigned as such and the numeric variables assigned as numeric.
3. Attach a “Select Columns” widget to the “File” widget. A majority of the meta attributes are removed from analysis. Note: we are considering the predwhgr variable (daily whole grain intake) as a numeric feature for cleaning purposes and will be reassigned meta attributes later.
4. Attach the “Feature Statistics” widget to the “Select Columns” widget. In this step missing data is observed. Features such as dmdyrsus, dmdschol and dmdeduc3 are missing more than half of the data required in this field. It makes little sense to use less than half of the data collected to determine averages to impute the missing data points on the majority of the samples. For this reason, we exclude these variables from the analysis by using the “Select Columns” Widget connected in step 3.
5. Attach the “Impute” widget to the “Select Columns” widget. We impute the missing data in the feature variables by setting the default method to the “average/most frequent” option.
6. Attach another “Select Columns” widget to set the “predwhgr” variable as a meta variable. For now, we do not want to directly model the daily whole grain intake data rather we want to learn the substructures that are related to daily whole grain intake later.

**Clustering**

The cluster analysis hypothesizes some characteristics which generate clusters and we assign a label to the clusters based on these characteristics. We use hierarchical clustering and k-means method to carry out the cluster analysis.

k-means Clustering Method

1. Attach a “k-means” widget to the “Select Columns” widget. Select the ‘number of clusters’ option to be from 2 to 8. Normalize the columns with 10 re-runs and 300 maximum iterations. We observe that the highest silhouette score of 0.127 is associated with cluster size of 4. The “k-means” widget shows an error message stating “silhouette scores are not computed for >5000 samples.” We will carry on with the analysis and compare this with the Hierarchical clustering method later.
2. Attach a “Silhouette Plot” widget to the “k-means” widget to show the individual silhouette score assigned to each participant.
3. Attach a “Scatter Plot” widget to the “k-means” widget. Select “cluster” as the color variable and select “Find Informative Projections” option and click “Start”. Orange will attempt to find the pair of variables that best explain the cluster behavior. Now select the highest ranked projection in the “Score Plot” window as shown in the image below. Notice the four clusters in four different colors shown on the scatter plot. According to the score plot, the pair of variables, ratio of income to poverty (indfmpir) and age (ridageyr), best explains the four clusters. From the scatter plot, we notice that cluster 1 colored blue comprises of people that have a higher ratio of income to poverty spanning all ages. Cluster 4 contains young people with lower ratio of income to poverty value. It is important to note that the categorical variables were not considered in the ranking of pairs because this is a scatterplot that only considers numerical features.

A screenshot of a computer

Description automatically generated

1. Attach a “Distribution” widget to the “k-Means” widget and examine the relationship between cluster assignment and the predwhgr variable.

1. Attach a “Box Plot” widget to the “k-Means” widget to examine this relationship.

A screenshot of a computer

Description automatically generated

Notice that cluster 1 has the highest mean daily intake of wholegrains followed by cluster 3 and cluster 2. Cluster 4 has the lowest mean daily intake of wholegrains.

The Box plot and statistical analysis shows that the differences in means of daily wholegrain intake across clusters is significant with a P-Value of 0.000

Hierarchical Clustering Method

1. Attach a “Distances” widget to the “Select Columns” widget created in step 6 to calculate the distances among the data points. We select distances between rows and use the normalized Euclidean distance metric options within the “Distances” widget.
2. Attach a “Hierarchical Clustering” widget to the “Distances” widget. A vertical line is drawn on the dendrogram at value 4 to form seven clusters. The seven clusters are found by trial and error using the scale above the dendrogram to create a reasonably sized cluster.
3. Attach a “Scatter Plot” widget to the “Hierarchical Clustering” widget. Select “Cluster” as the color option and select “Find Informative Projections” button and clicked “start”. From the scatter plot, Orange shows the seven clusters represented by seven different colors with the top three pair of variables that best explain the cluster behavior in the “score plot” window. The screenshot below shows the seven clusters on a scatterplot.

A screenshot of a cell phone

Description automatically generated

From our cluster analysis, we observe that indfmpir variable and ridageyr form the top pair that best explains the cluster behavior. Notice cluster 1 and 3 on the top right of the scatter plot with ages ranging from approximately 21 to 69 years old and the ratio of family income to poverty variable ranging pretty high.

1. Attach the “Distributions” widget to examine the relationship between the cluster assignment and the predwhgr variable.
2. Attach the “Box Plot” widget to “Hierarchical Clustering” widget and examine the relation

A screenshot of a computer

Description automatically generated

Notice that cluster 3 has the highest mean daily intake of wholegrains and Cluster 7 has the lowest mean daily intake of wholegrains with the other cluster means ranging in between.

We see the box plots corresponding to the various clusters and a statistical summary that shows that the differences in means of daily wholegrain intake across clusters is significant with a P-Value of 0.00

**Results**

From this cluster analysis, it is observed that age and ratio of family income to poverty variables are the most important features that determined the seven clusters in the hierarchical clustering method and four clusters in the k-means method. The clusters were then associated with the wholegrain variable in the “Distributions” and “Box Plot” widgets. In both clustering methods it is that older and financially better off participants who took the National Health and Nutrition Examination Survey of the NHANES have a high probability of falling into a cluster and would have a higher mean of daily whole grain intake in ounces.

**Conclusion**

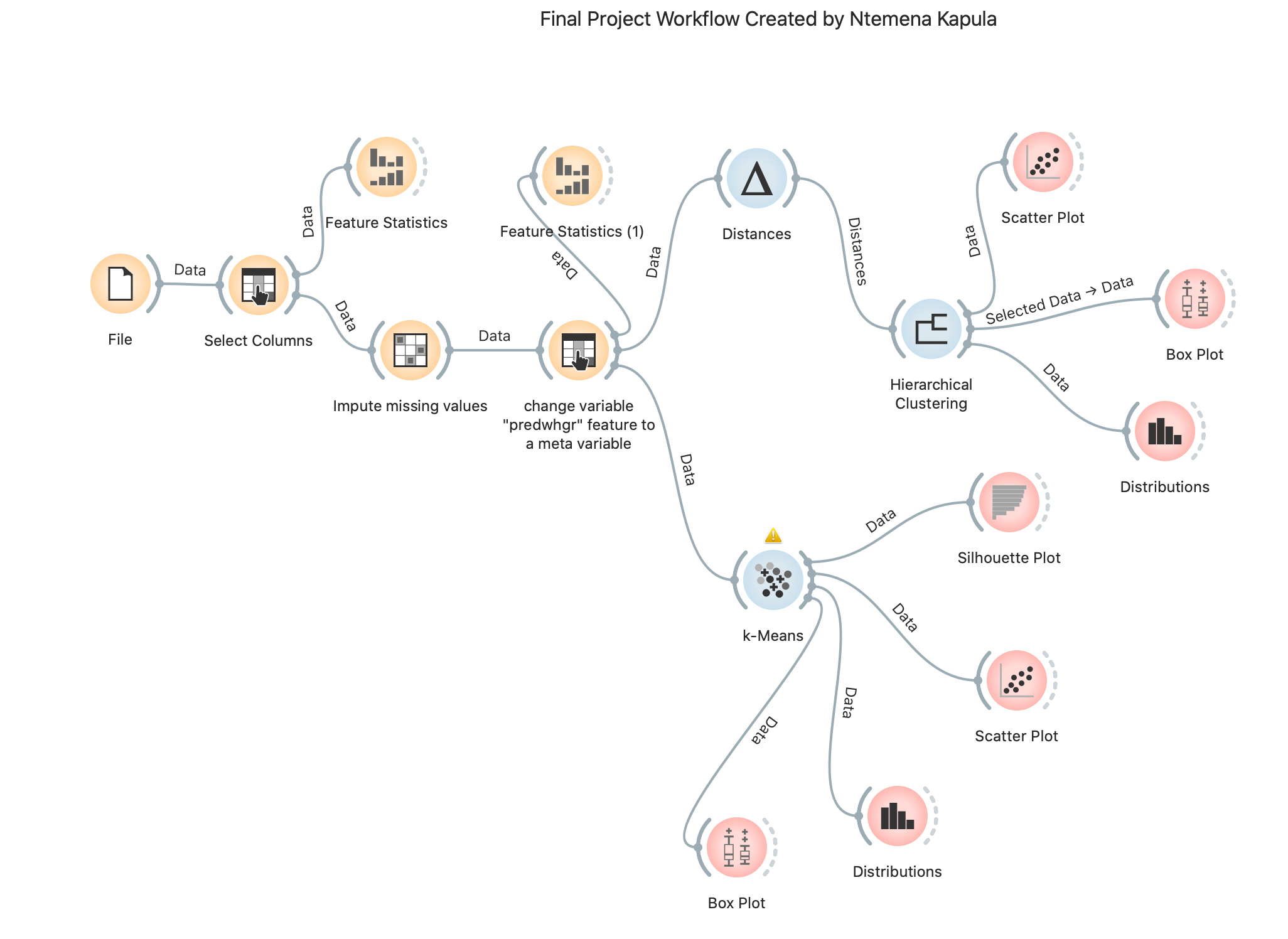
The research question, *what factors contribute to differences in reported daily whole grain intake in the NHANES Dataset*? was addressed using unsupervised learning methods. Age and family income to poverty ratio are some of the factors that are estimated by the model to contribute to differences in reported daily whole grain intake.

References

1. ago MR• 5 months. Health Status and the Role of Nutrition on SARS-CoV / Covid-19 [Internet]. Naked Food Magazine. 2020 [cited 2020 Aug 16]. Available from: https://nakedfoodmagazine.com/health-status-covid-19/

**Appendix**

Figure1: Final Workflow

Table1: Variables used in the cluster analysis

|  |  |  |
| --- | --- | --- |
| **Label** | **Variable** | **Measurement Level** |
| Gender | riagendr | Nominal |
| Age | ridageyr | Scale |
| Race/Ethnicity | ridreth1 | Nominal |
| Served in the US Armed Forces | dmqmilit | Nominal |
| Country of Birth | dmdborn2 | Nominal |
| Citizenship Status | dmdcitzn | Nominal |
| Length of time in US | dmdyrsus | Nominal |
| Education level - children | dmdeduc3 | Nominal |
| Education level - Adults | dmdeduc2 | Nominal |
| Now attending school | dmdschol | Nominal |
| Marital Status | dmdmartl | Nominal |
| People in household | dmdhhsiz | Nominal |
| People in family | dmdfmsiz | Nominal |
| Household income | indhhin2 | Nominal |
| Family income | indfmin2 | Nominal |
| Ratio of income to poverty | indfmpir | Scale |
| Daily whole grain intake | predwhgr | Scale |