The Axis of Smoking, Smokers, and Health Standards

Understanding Health Effects of Smoking Through SAS Studio

Aaditya Alekar Medrik Minassian Bita Faraji

Master of Science Information Systems

Submitted to Dr. Shilpa Balan
In requirement for the CIS 5210
Department of Information Systems
College of Business & Economics
California State University, Los Angeles

Contents

Abstract

The following research work uses SAS Studio to Analyze the ‘Healthcare Data Set Stroke Data’ from www.Kaggle.com, an online open database depository. The researchers have created six elaborate visuals to explain the relation between smoking and health factor such as BMI, Hypertension, and other relations such as sex, age, and employment category. This work opens avenues to understanding smoking and its relations with factors that may not previously been explored. It explains the steps that are taken in SAS Studio to analyze a data set.

Data Sets Used

https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data

Figure 1 Knowing why you smoke can be the first step in helping you to quit image source: Getty images

Data Description

The Data used from Kaggle is a healthcare related dataset that has information about stroke and other health factors such as hypertension, smoking, and BMI. The dataset has more than 43000 rows and 12 columns which include the following numeric and string categorical data.

Numeric	String
ID	Gender
Age	Ever married
Hypertension	Work Type
Heart Diseases	Residence Type
Average Glucose Level	BMI
	Smoking Status

We see that the columns in our data are divided into two categories Numeric and String.

Numerical elements are the variables such as Hypertension, which is defined by 0 and 1. Zero stands for no hypertension and One stands for hypertension. Heart disease is represented in the same format as well. Average Glucose Level is shown in actual values.

String values make the data analysis very flexible. Our analysis has used all string categories in relation to numeric variables to create the best visuals in description. Heart disease is a central element in overall health of every individual irrespective of their life style, gender and age. Hence, it is important to know how are commonly known heart related elements such as BMI and Smoking Status related to each other. Our data set allows us to identify smokers and compare their employment types, age and even BMI. We can see that how prevalent is heart disease among people with different smoking status or even employment category.

For example, it is a common researched result that smoking increases the chances of coronary heart disease and even stroke. (British Heart Foundation, 2017) Hence, our dataset allows us to find the relation between smoking and other factors such as BMI and work type.

Work Type, refers to employment category. In our dataset we can differentiate individuals based on their employment, whether private, government, self-employed or even never worked. Categorical data column creates opportunity to conduct deeper research and analysis.

An important insight into our data shows that the sample size of 43000+ individuals has more female samples than male samples. Therefore, an analysis based solely on gender may show results leaning toward females rather than males. When conducting analysis based on gender and age it is recommended to take the same number of samples from the given categories to refute further doubt and unwanted critical responses to results and even to the dataset.

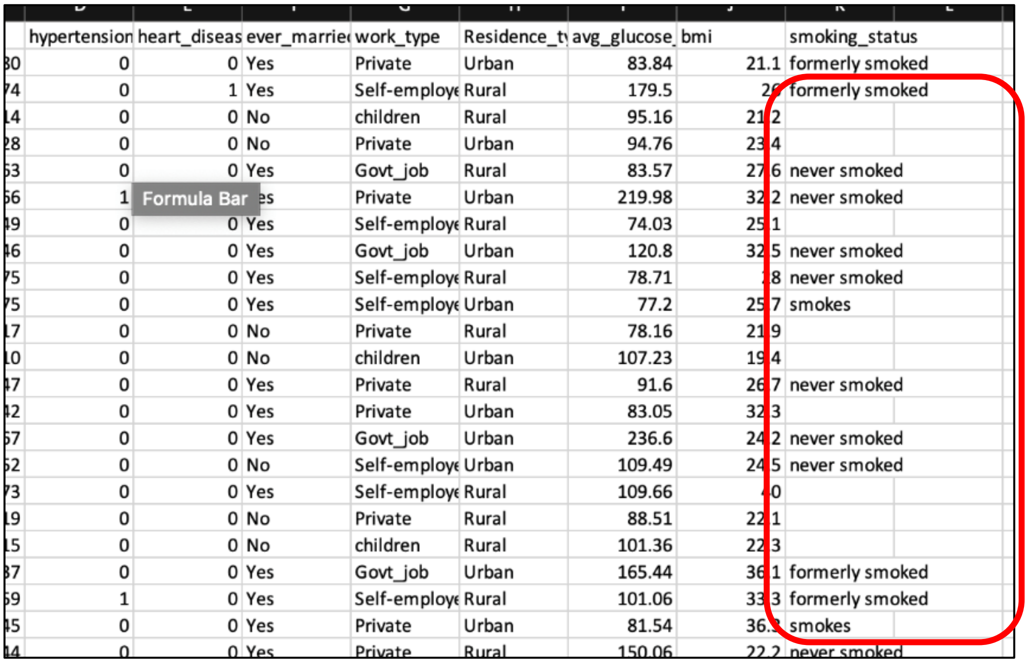
Data Refinement

We have used four cleaning methods in our dataset refinement process. Which are the following;

Select and treat Blank cells:

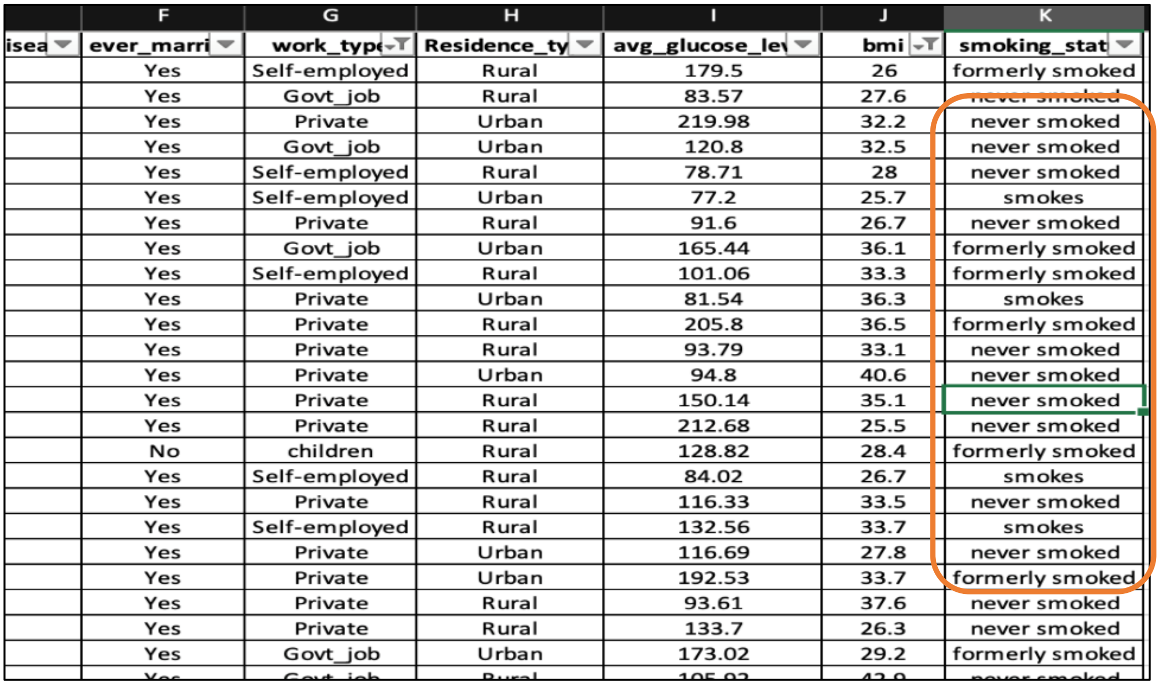
The screenshots below show how spaces are cleared between cells in excel. Data has been cleaned in excel before being imported to SAS Studio. This allows uniformity in data and makes further processing easier.

Before:



	hypertension	heart_diseas	ever_marrie	work_type	Residence_t	avg_glucose_	bmi	smoking_status	
30	0	0	Yes	Private	Urban	83.84	21.1	formerly smoked	
74	0	1	Yes	Self-employe	Rural	179.5	26	formerly smoked	
14	0	0	No	children	Rural	95.16	21.2		
28	0	0	No	Private	Urban	94.76	23.4		
53	0	0	Yes	Govt_job	Rural	83.57	27.6	never smoked	
56	1	Formula Bar	es	Private	Urban	219.98	32.2	never smoked	
49	0	0	Yes	Self-employe	Rural	74.03	25.1		
46	0	0	Yes	Govt_job	Urban	120.8	32.5	never smoked	
75	0	0	Yes	Self-employe	Rural	78.71	28	never smoked	
75	0	0	Yes	Self-employe	Urban	77.2	25.7	smokes	
17	0	0	No	Private	Rural	78.16	21.9		
10	0	0	No	children	Urban	107.23	19.4		
47	0	0	Yes	Private	Rural	91.6	26.7	never smoked	
42	0	0	Yes	Private	Urban	83.05	32.3		
57	0	0	Yes	Govt_job	Urban	236.6	24.2	never smoked	
52	0	0	No	Self-employe	Urban	109.49	24.5	never smoked	
73	0	0	Yes	Self-employe	Rural	109.66	40		
19	0	0	No	Private	Rural	88.51	22.1		
15	0	0	No	children	Rural	101.36	22.3		
37	0	0	Yes	Govt_job	Urban	165.44	36.1	formerly smoked	
59	1	0	Yes	Self-employe	Rural	101.06	33.3	formerly smoked	
45	0	0	Yes	Private	Urban	81.54	36.[illegible]	smokes	
44	0	0	Yes	Private	Rural	150.06	22.2	never smoked	

After:

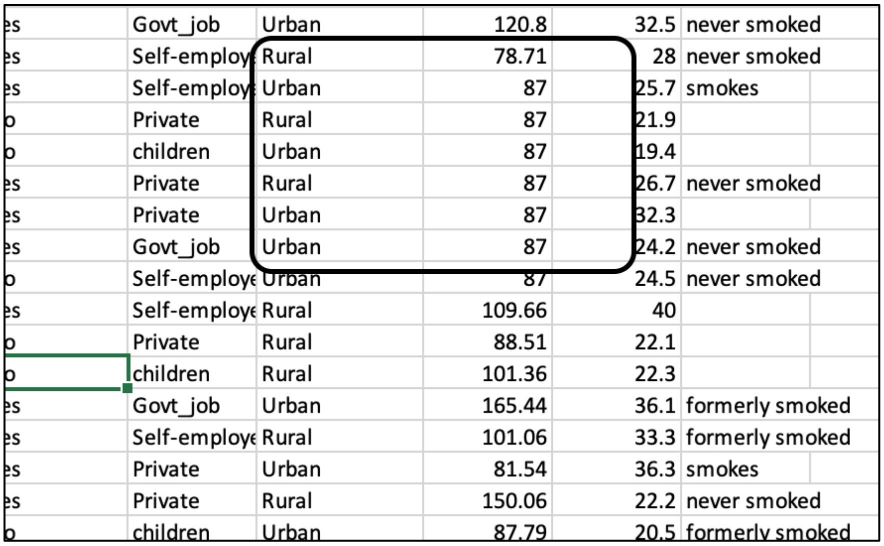


	F	G	H	I	J	K
isea	ever_marri	work_type	Residence_ty	avg_glucose_lev	bmi	smoking_stat
	Yes	Self-employed	Rural	179.5	26	formerly smoked
	Yes	Govt_job	Rural	83.57	27.6	never smoked
	Yes	Private	Urban	219.98	32.2	never smoked
	Yes	Govt_job	Urban	120.8	32.5	never smoked
	Yes	Self-employed	Rural	78.71	28	never smoked
	Yes	Self-employed	Urban	77.2	25.7	smokes
	Yes	Private	Rural	91.6	26.7	never smoked
	Yes	Govt_job	Urban	165.44	36.1	formerly smoked
	Yes	Self-employed	Rural	101.06	33.3	formerly smoked
	Yes	Private	Urban	81.54	36.3	smokes
	Yes	Private	Rural	205.8	36.5	formerly smoked
	Yes	Private	Rural	93.79	33.1	never smoked
	Yes	Private	Urban	94.8	40.6	never smoked
	Yes	Private	Rural	150.14	35.1	never smoked
	Yes	Private	Rural	212.68	25.5	never smoked
	No	children	Rural	128.82	28.4	formerly smoked
	Yes	Self-employed	Rural	84.02	26.7	smokes
	Yes	Private	Rural	116.33	33.5	never smoked
	Yes	Self-employed	Rural	132.56	33.7	smokes
	Yes	Private	Urban	116.69	27.8	never smoked
	Yes	Private	Urban	192.53	33.7	formerly smoked
	Yes	Private	Rural	93.61	37.6	never smoked
	Yes	Private	Rural	133.7	26.3	never smoked
	Yes	Govt_job	Urban	173.02	29.2	formerly smoked
	Yes	Govt_job	Rural	105.92	42.9	never smoked

Remove Duplicate Values

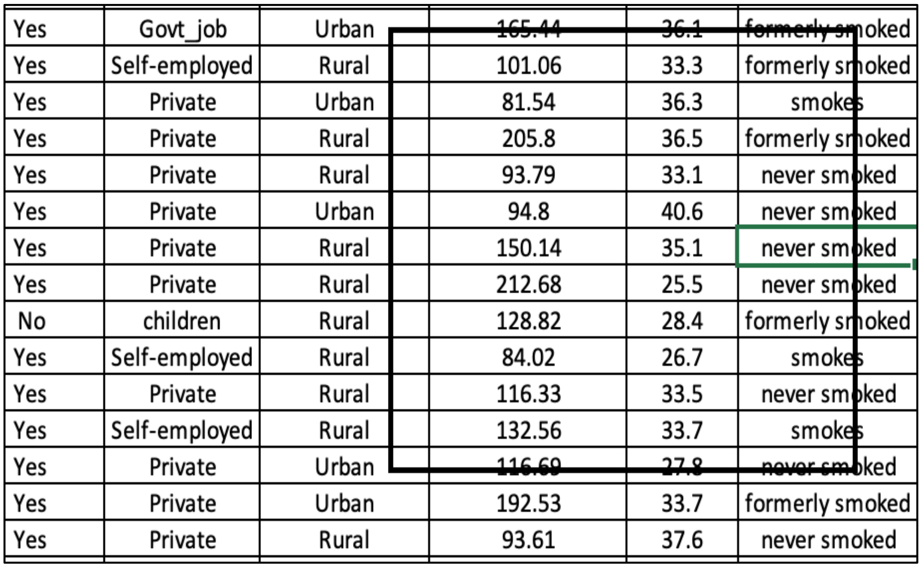
The cleaning process below removed duplicate values from the data set to allow more accuracy.

Before:



es	Govt_job	Urban	120.8	32.5	never smoked
es	Self-employ	Rural	78.71	28	never smoked
es	Self-employ	Urban	87	25.7	smokes
o	Private	Rural	87	21.9	
o	children	Urban	87	19.4	
es	Private	Rural	87	26.7	never smoked
es	Private	Urban	87	32.3	
es	Govt_job	Urban	87	24.2	never smoked
o	Self-employe	Urban	87	24.5	never smoked
es	Self-employe	Rural	109.66	40	
o	Private	Rural	88.51	22.1	
o	children	Rural	101.36	22.3	
es	Govt_job	Urban	165.44	36.1	formerly smoked
es	Self-employe	Rural	101.06	33.3	formerly smoked
es	Private	Urban	81.54	36.3	smokes
es	Private	Rural	150.06	22.2	never smoked
o	children	Urban	87.79	20.5	formerly smoked

After:

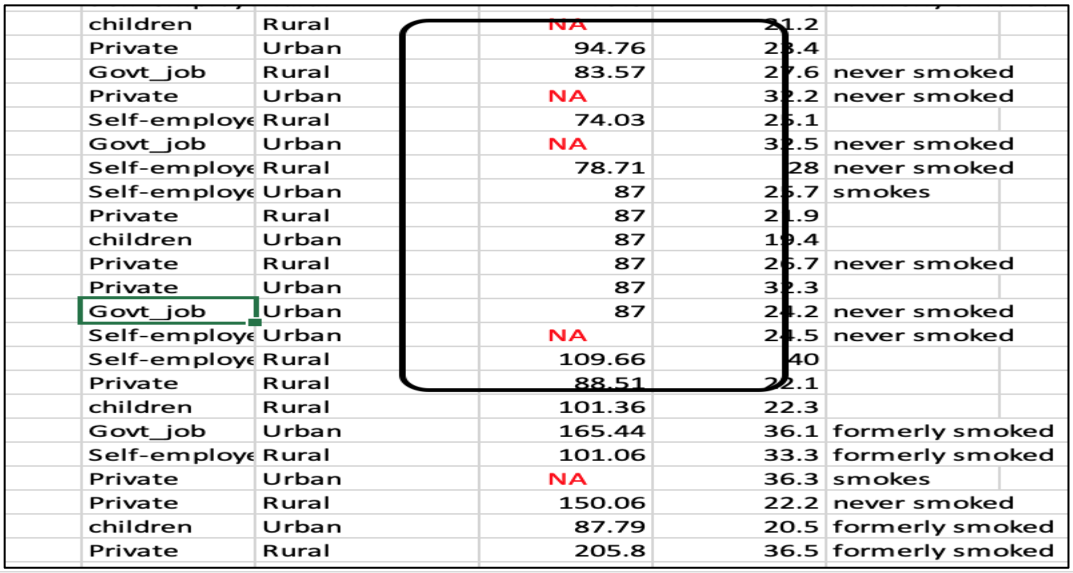


Yes	Govt_job	Urban		165.44	36.1	formerly smoked
Yes	Self-employed	Rural		101.06	33.3	formerly smoked
Yes	Private	Urban		81.54	36.3	smokes
Yes	Private	Rural		205.8	36.5	formerly smoked
Yes	Private	Rural		93.79	33.1	never smoked
Yes	Private	Urban		94.8	40.6	never smoked
Yes	Private	Rural		150.14	35.1	never smoked
Yes	Private	Rural		212.68	25.5	never smoked
No	children	Rural		128.82	28.4	formerly smoked
Yes	Self-employed	Rural		84.02	26.7	smokes
Yes	Private	Rural		116.33	33.5	never smoked
Yes	Self-employed	Rural		132.56	33.7	smokes
Yes	Private	Urban		116.69	27.8	never smoked
Yes	Private	Urban		192.53	33.7	formerly smoked
Yes	Private	Rural		93.61	37.6	never smoked

Removing NA values

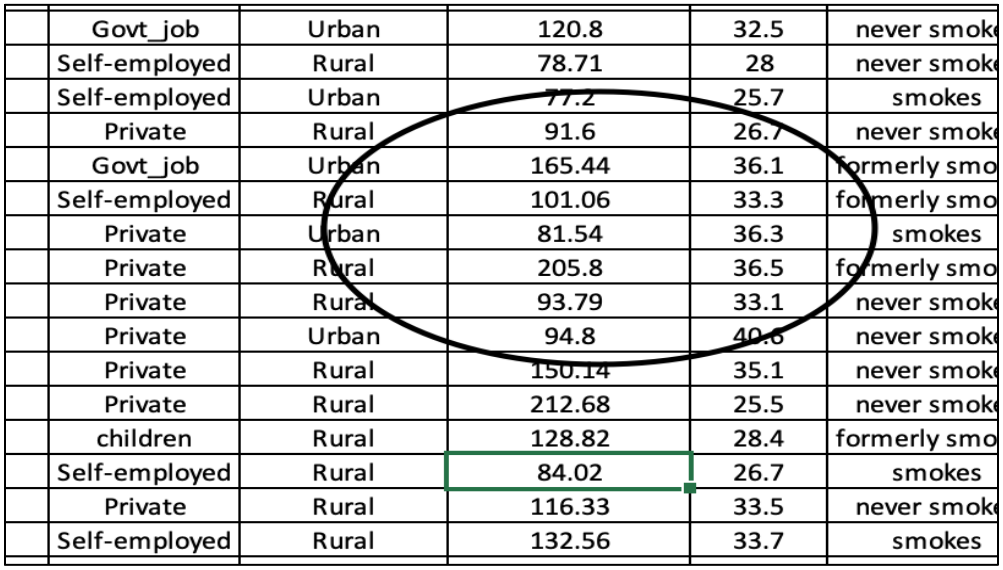
Not Applicable values have been removed in the below screenshots. The images show how the data looks before and after the step was implemented. This is very vital while conducting statistical analysis and tests in SAS Studio.

Before:



	children	Rural		NA	21.2	
	Private	Urban		94.76	23.4	
	Govt_job	Rural		83.57	27.6	never smoked
	Private	Urban		NA	32.2	never smoked
	Self-employe	Rural		74.03	25.1	
	Govt_job	Urban		NA	32.5	never smoked
	Self-employe	Rural		78.71	28	never smoked
	Self-employe	Urban		87	25.7	smokes
	Private	Rural		87	21.9	
	children	Urban		87	19.4	
	Private	Rural		87	26.7	never smoked
	Private	Urban		87	32.3	
	Govt_job	Urban		87	24.2	never smoked
	Self-employe	Urban		NA	24.5	never smoked
	Self-employe	Rural		109.66	40	
	Private	Rural		88.51	22.1	
	children	Rural		101.36	22.3	
	Govt_job	Urban		165.44	36.1	formerly smoked
	Self-employe	Rural		101.06	33.3	formerly smoked
	Private	Urban		NA	36.3	smokes
	Private	Rural		150.06	22.2	never smoked
	children	Urban		87.79	20.5	formerly smoked
	Private	Rural		205.8	36.5	formerly smoked

After:

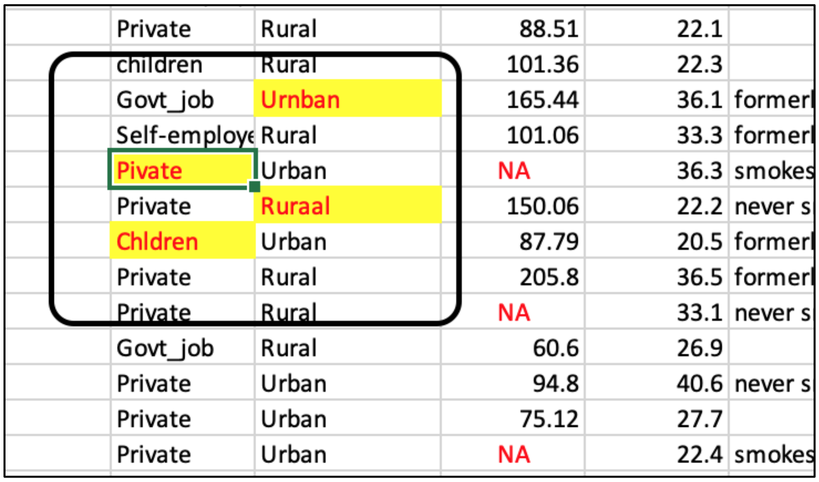


	Govt_job	Urban	120.8	32.5	never smok
	Self-employed	Rural	78.71	28	never smok
	Self-employed	Urban	77.2	25.7	smokes
	Private	Rural	91.6	26.7	never smok
	Govt_job	Urban	165.44	36.1	formerly smo
	Self-employed	Rural	101.06	33.3	formerly smo
	Private	Urban	81.54	36.3	smokes
	Private	Rural	205.8	36.5	formerly smo
	Private	Rural	93.79	33.1	never smok
	Private	Urban	94.8	40.6	never smok
	Private	Rural	150.14	35.1	never smok
	Private	Rural	212.68	25.5	never smok
	children	Rural	128.82	28.4	formerly smo
	Self-employed	Rural	84.02	26.7	smokes
	Private	Rural	116.33	33.5	never smok
	Self-employed	Rural	132.56	33.7	smokes

Spell Check

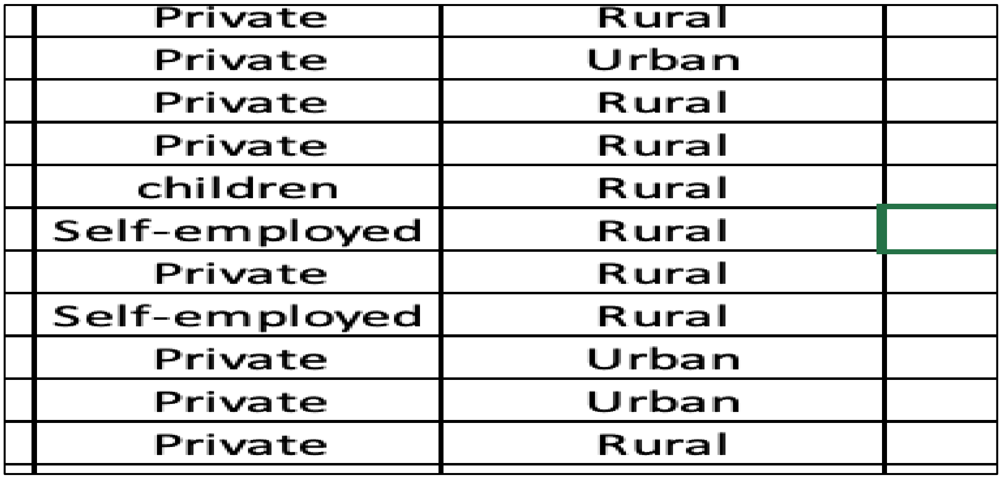
Spellcheck is one of the most frequently used methods of data cleaning and refinement. Below words have been identified with errors and uniformed with the rest of the data in order to assist the accuracy of the dataset and the analysis.

Before:



	Private	Rural	88.51	22.1	
	children	Rural	101.36	22.3	
	Govt_job	Urnban	165.44	36.1	formerl
	Self-employe	Rural	101.06	33.3	formerl
	Pivate	Urban	NA	36.3	smokes
	Private	Ruraal	150.06	22.2	never s
	Chldren	Urban	87.79	20.5	formerl
	Private	Rural	205.8	36.5	formerl
	Private	Rural	NA	33.1	never s
	Govt_job	Rural	60.6	26.9	
	Private	Urban	94.8	40.6	never s
	Private	Urban	75.12	27.7	
	Private	Urban	NA	22.4	smokes

After:

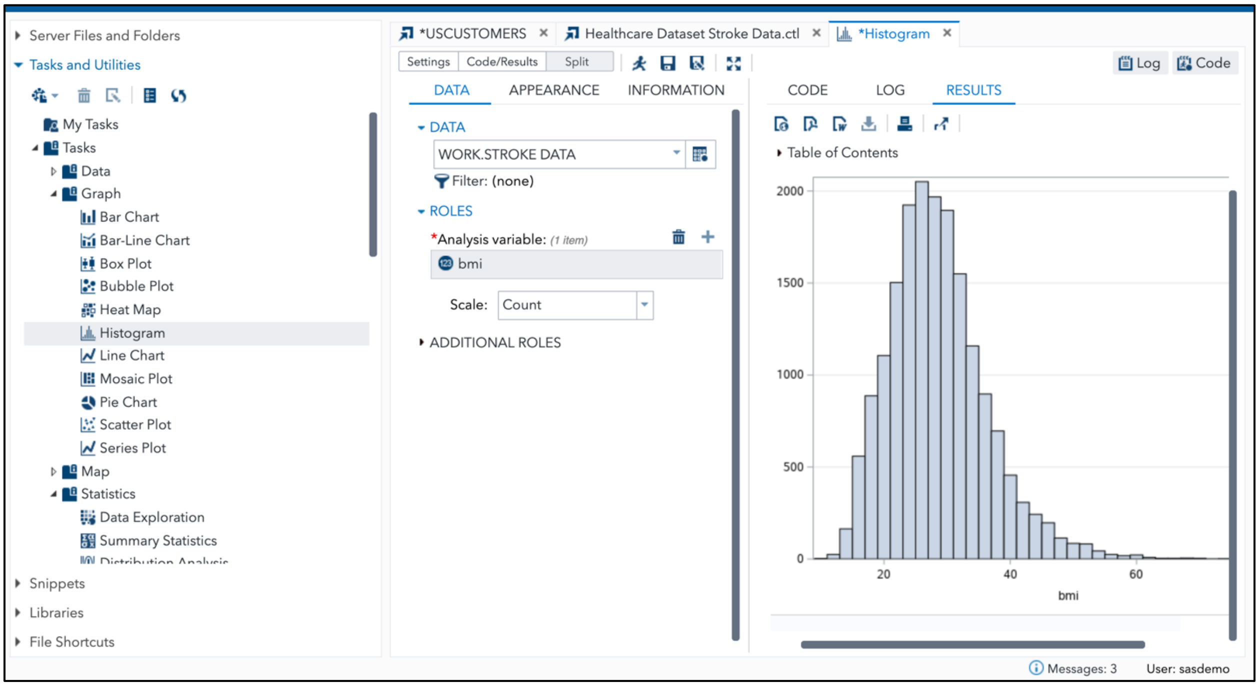


	Private	Rural	
	Private	Urban	
	Private	Rural	
	Private	Rural	
	children	Rural	
	Self-employed	Rural	
	Private	Rural	
	Self-employed	Rural	
	Private	Urban	
	Private	Urban	
	Private	Rural	

Data Analysis and Visualizations

The following analysis has six different visualizations, which answer different questions. Every question is answered with relevant visuals from SAS Studio.

What is the Count of People in different Body Mass Index categories?



DATA

WORK.STROKE DATA

Filter: (none)

ROLES

*Analysis variable: (1 item)

bmi

Scale: Count

User: sasdemo

Figure 2 Count of Individuals and BMI, Screenshot from SAS Studio

The figure above shows the raw chart from SAS Studio. We see that it is a histogram and has BMI on the X axis and count of individuals on the Y axis. The following figure shows a closer look and a detailed explanation.

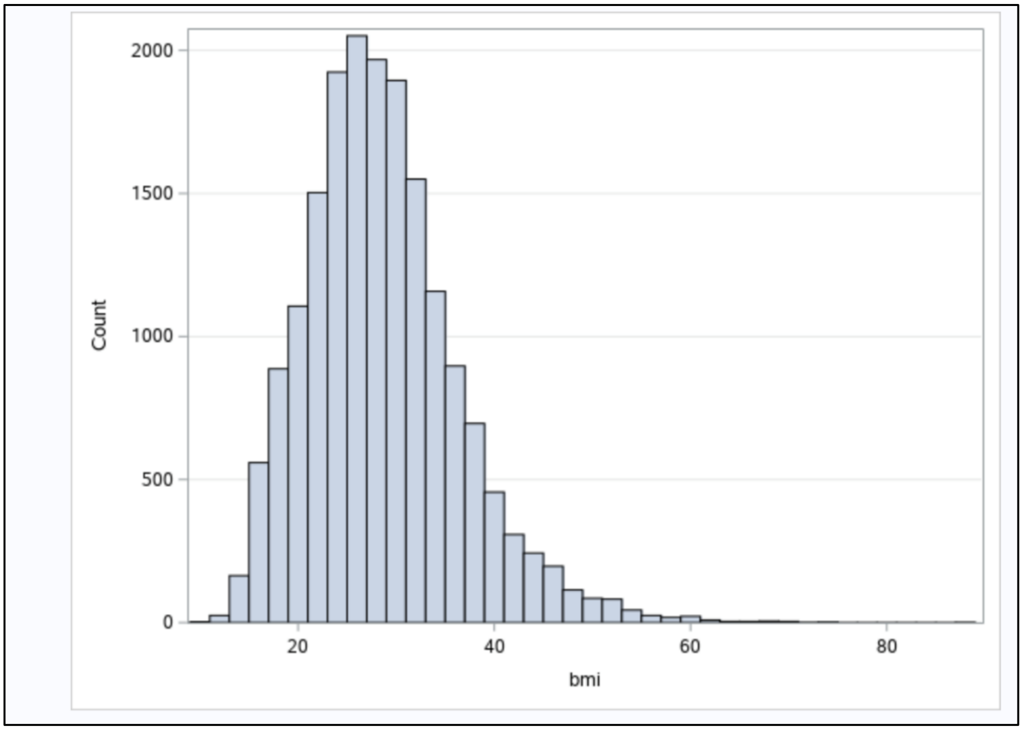
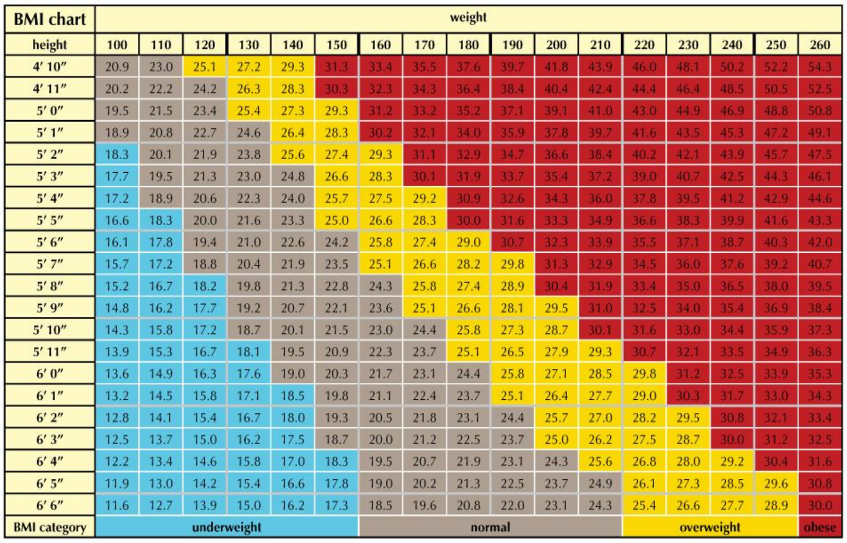


Figure 3 Count of BMI

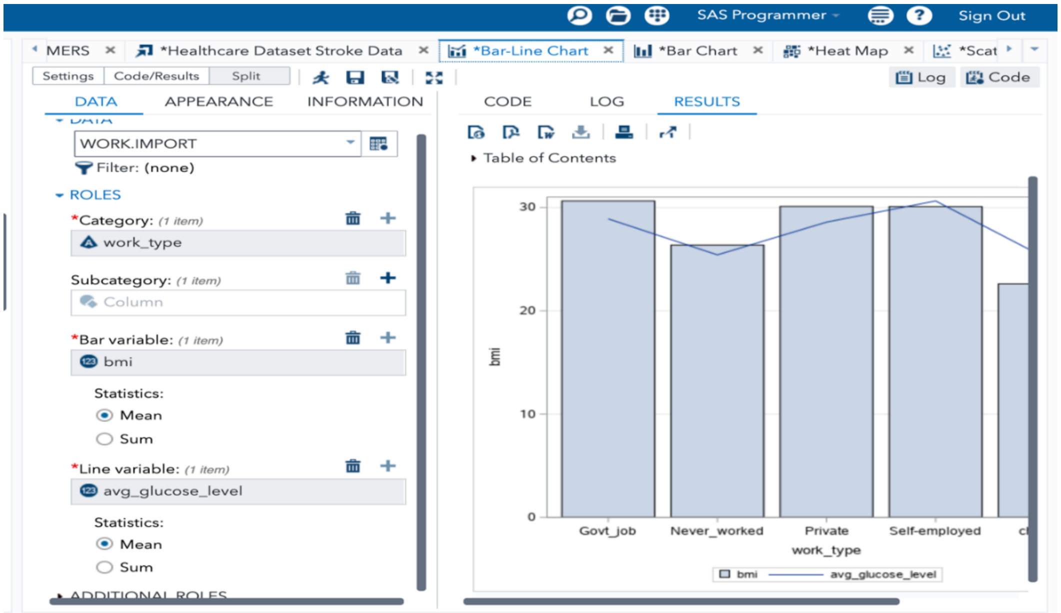
The figure above shows count of Individuals on the Y axis and the BMI levels on the X axis. We can therefore see that the highest count is over 2000 for a single level of body mass index 28. It is also clear that the highest levels of BMI are 25 to 33. According to global health standard most individuals in our dataset are overweight as BMI levels above 25 are usually considered to be overweight for all height and weight categories. Below is a standard BMI Chart, which will allow us to better understand the visualization above.



BMI chart	weight																
height	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260
4′ 10″	20.9	23.0	25.1	27.2	29.3	31.3	33.4	35.5	37.6	39.7	41.8	43.9	46.0	48.1	50.2	52.2	54.3
4′ 11″	20.2	22.2	24.2	26.3	28.3	30.3	32.3	34.3	36.4	38.4	40.4	42.4	44.4	46.4	48.5	50.5	52.5
5′ 0″	19.5	21.5	23.4	25.4	27.3	29.3	31.2	33.2	35.2	37.1	39.1	41.0	43.0	44.9	46.9	48.8	50.8
5′ 1″	18.9	20.8	22.7	24.6	26.4	28.3	30.2	32.1	34.0	35.9	37.8	39.7	41.6	43.5	45.3	47.2	49.1
5′ 2″	18.3	20.1	21.9	23.8	25.6	27.4	29.3	31.1	32.9	34.7	36.6	38.4	40.2	42.1	43.9	45.7	47.5
5′ 3″	17.7	19.5	21.3	23.0	24.8	26.6	28.3	30.1	31.9	33.7	35.4	37.2	39.0	40.7	42.5	44.3	46.1
5′ 4″	17.2	18.9	20.6	22.3	24.0	25.7	27.5	29.2	30.9	32.6	34.3	36.0	37.8	39.5	41.2	42.9	44.6
5′ 5″	16.6	18.3	20.0	21.6	23.3	25.0	26.6	28.3	30.0	31.6	33.3	34.9	36.6	38.3	39.9	41.6	43.3
5′ 6″	16.1	17.8	19.4	21.0	22.6	24.2	25.8	27.4	29.0	30.7	32.3	33.9	35.5	37.1	38.7	40.3	42.0
5′ 7″	15.7	17.2	18.8	20.4	21.9	23.5	25.1	26.6	28.2	29.8	31.3	32.9	34.5	36.0	37.6	39.2	40.7
5′ 8″	15.2	16.7	18.2	19.8	21.3	22.8	24.3	25.8	27.4	28.9	30.4	31.9	33.4	35.0	36.5	38.0	39.5
5′ 9″	14.8	16.2	17.7	19.2	20.7	22.1	23.6	25.1	26.6	28.1	29.5	31.0	32.5	34.0	35.4	36.9	38.4
5′ 10″	14.3	15.8	17.2	18.7	20.1	21.5	23.0	24.4	25.8	27.3	28.7	30.1	31.6	33.0	34.4	35.9	37.3
5′ 11″	13.9	15.3	16.7	18.1	19.5	20.9	22.3	23.7	25.1	26.5	27.9	29.3	30.7	32.1	33.5	34.9	36.3
6′ 0″	13.6	14.9	16.3	17.6	19.0	20.3	21.7	23.1	24.4	25.8	27.1	28.5	29.8	31.2	32.5	33.9	35.3
6′ 1″	13.2	14.5	15.8	17.1	18.5	19.8	21.1	22.4	23.7	25.1	26.4	27.7	29.0	30.3	31.7	33.0	34.3
6′ 2″	12.8	14.1	15.4	16.7	18.0	19.3	20.5	21.8	23.1	24.4	25.7	27.0	28.2	29.5	30.8	32.1	33.4
6′ 3″	12.5	13.7	15.0	16.2	17.5	18.7	20.0	21.2	22.5	23.7	25.0	26.2	27.5	28.7	30.0	31.2	32.5
6′ 4″	12.2	13.4	14.6	15.8	17.0	18.3	19.5	20.7	21.9	23.1	24.3	25.6	26.8	28.0	29.2	30.4	31.6
6′ 5″	11.9	13.0	14.2	15.4	16.6	17.8	19.0	20.2	21.3	22.5	23.7	24.9	26.1	27.3	28.5	29.6	30.8
6′ 6″	11.6	12.7	13.9	15.0	16.2	17.3	18.5	19.6	20.8	22.0	23.1	24.3	25.4	26.6	27.7	28.9	30.0
BMI category	underweight						normal						overweight				obese

Figure 4 BMI Chart Source: www.geneticsandfertility.com

What are the different BMI levels among different employment categories?



DATA

WORK.IMPORT

Filter: (none)

ROLES

*Category: (1 item)
work_type

Subcategory: (1 item)
Column

*Bar variable: (1 item)
bmi

Statistics:
● Mean
○ Sum

*Line variable: (1 item)
avg_glucose_level

Statistics:
● Mean
○ Sum

ADDITIONAL ROLES

Figure 5 SAS Studio Screenshot Bar-Line Chart

The figure above is a screenshot from SAS Studio where we have used a Bar-Line Chart to explain the relation between work type and BMI and glucose levels. The figure below clearly shows the complete chart in detail.

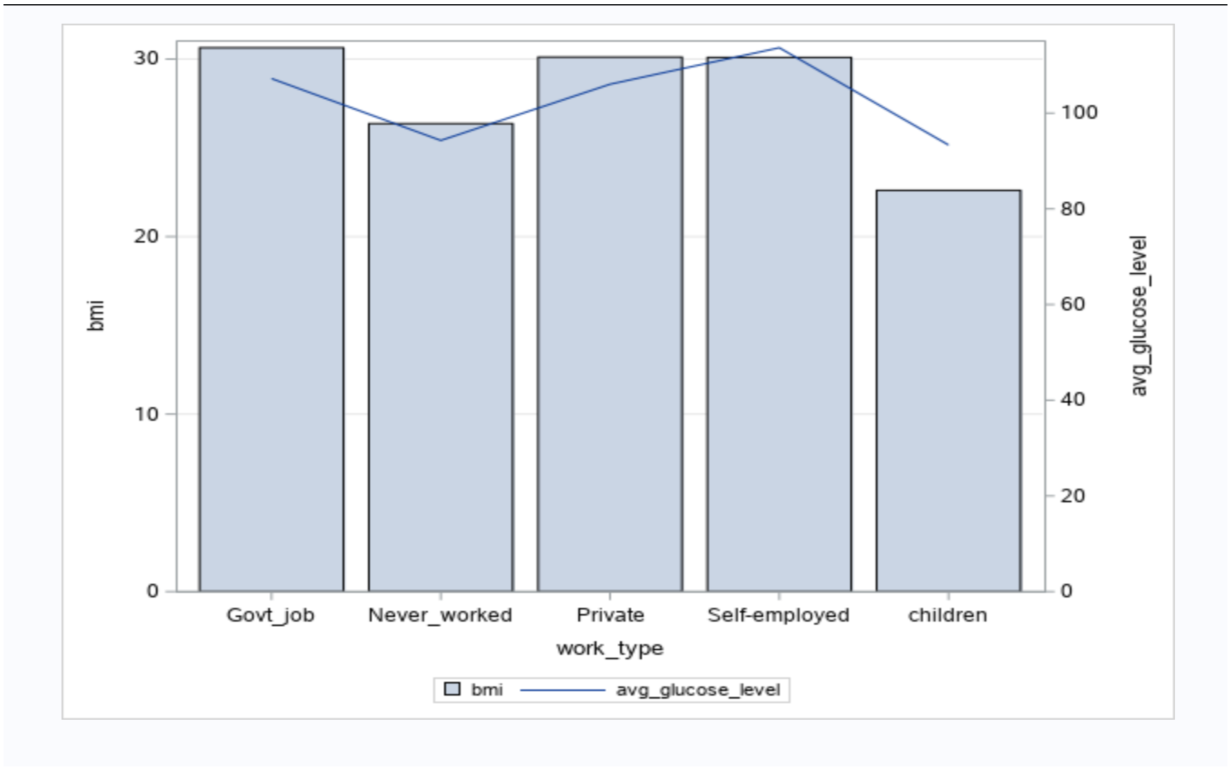
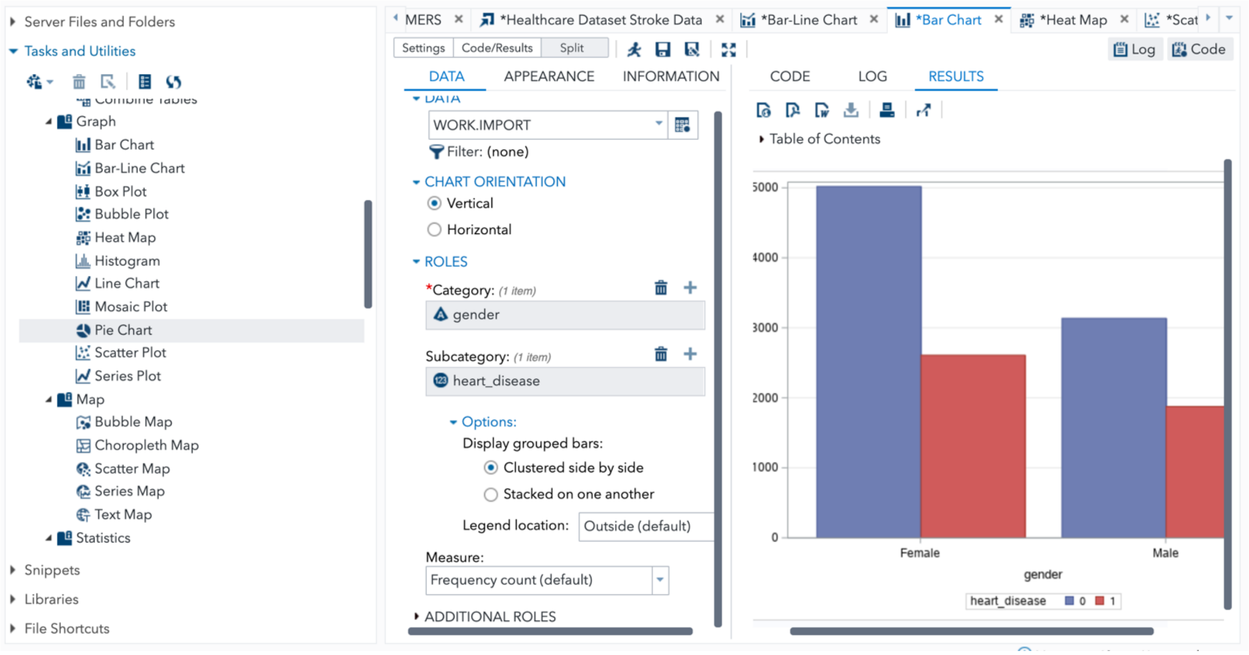


Figure 6 Bar-Line Chart showing BMI Levels and average glucose levels according to Employment Type

The Bar-Line Chart above shows different work types and their levels of BMI and glucose levels. It is seen that people having Government job also have higher BMI levels, while their glucose levels are lower than that of self-employed people. Self-employed individuals, however, have the highest average glucose levels. We can also see that children are significantly different from other categories where age can be a major factor. Over all people that are working have higher glucose and BMI levels. Another interesting information can be derived from this chart that the higher the BMI the higher the average glucose level will be. An article published in 2014 on the National Library of Medicine website by the US National Institute of Health explains that higher glucose levels are associated with Body Mass Index with elders. (National Library of Medicine , 2014) Further research can prove this for all age categories.

How is heart disease spread between males and females?



DATA

WORK.IMPORT

Filter: (none)

CHART ORIENTATION

● Vertical

○ Horizontal

ROLES

*Category: (1 item)

gender

Subcategory: (1 item)

heart_disease

Options:

Display grouped bars:

● Clustered side by side

○ Stacked on one another

Legend location: Outside (default)

Measure:

Frequency count (default)

Figure 7 SAS Studio screenshot showing bar chart

The figure above shows the details of a bar chart from SAS Studio that were input to create a bar chart showing distinct counts of people with heart disease among males and females.

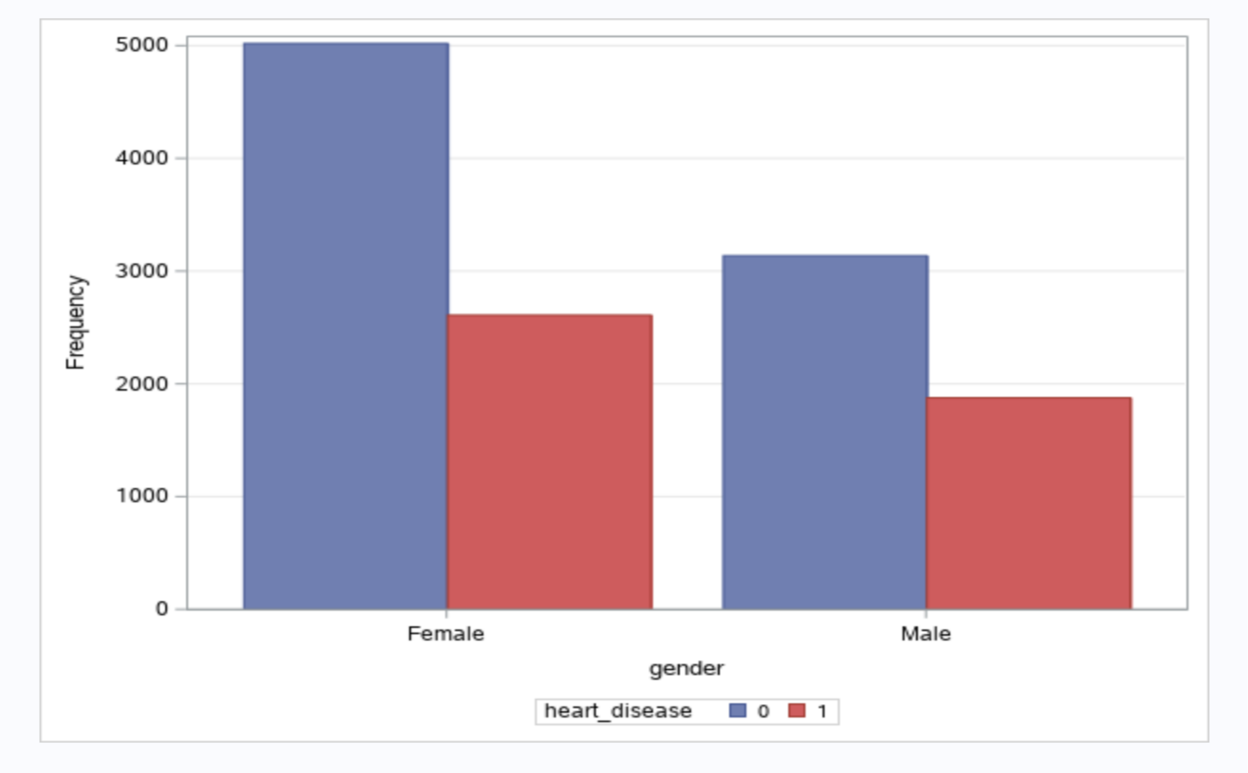
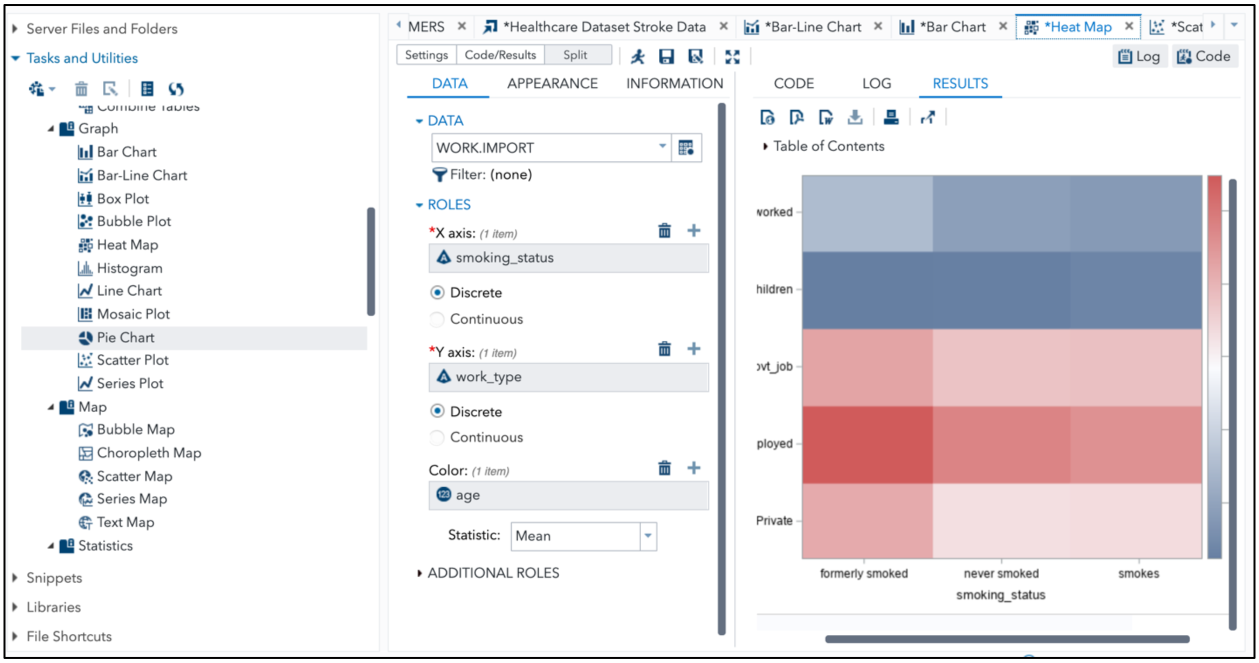


Figure 8 Bar Chart showing heart disease among males and females

The Bar Chart above shows the count of heart disease among Males and Females. We can see that in both categories people with heart disease are less. It is clear that more woman has heart diseases compared to men, but this is because the dataset includes more female samples than male. However, another closer look shows that the proportion of those with heart disease to those without one is more among men. Hence, we can see that men have more chance of having heart disease than women.

What is the nexus between work type, smoking status and age?



DATA

WORK.IMPORT

Filter: (none)

ROLES

*X axis: (1 item)

smoking_status

● Discrete

○ Continuous

*Y axis: (1 item)

work_type

● Discrete

○ Continuous

Color: (1 item)

age

Statistic: Mean

Figure 9 SAS Studio Screen Shot showing the making of a Heat Map visual

The figure above shows the variables involved in making a SAS Studio Heat Map. On the left side of the Heat Map we can see the roles on each axis.

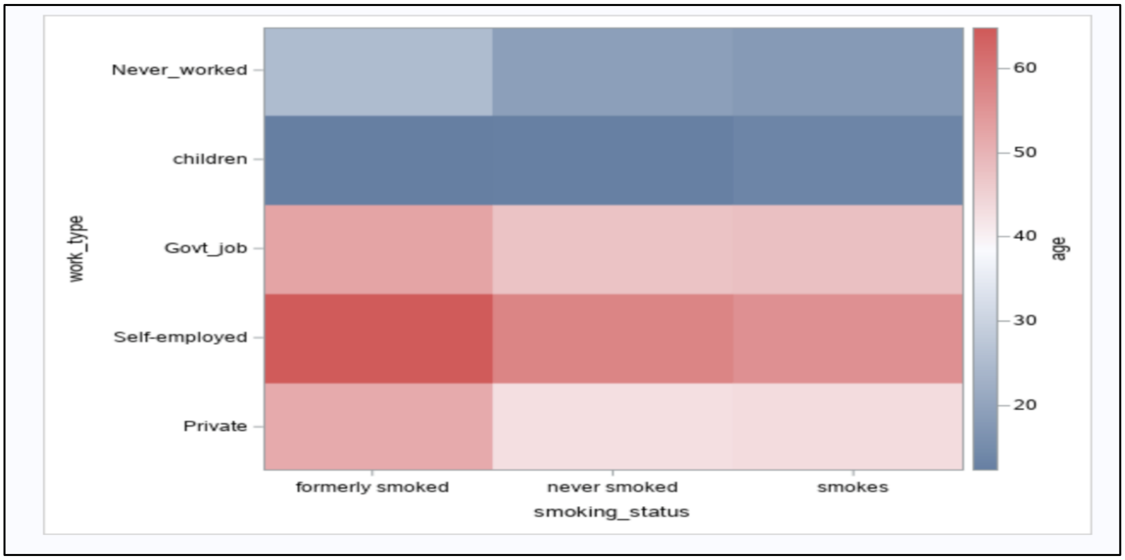
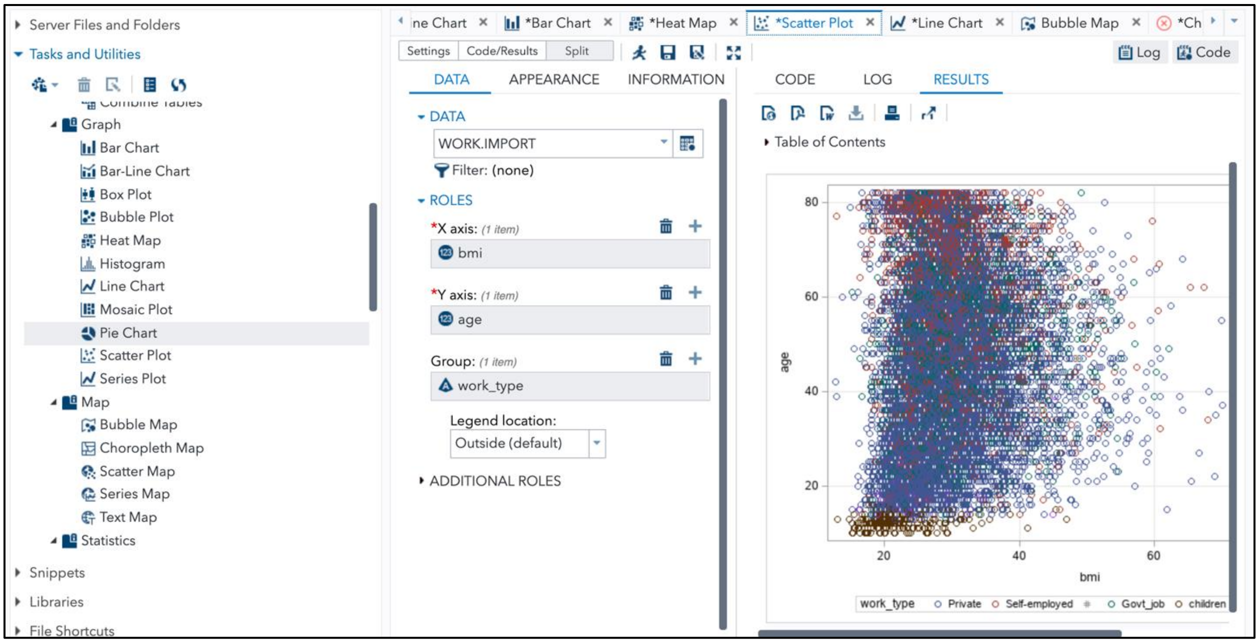


Figure 10 Heat Map showing the Nexus between Age, Work Type and Smoking Status.

The Heat Map above has a dual y axis, which explains work type and age. On the X axis we can see the smoking status. The colors change by age. The higher the age the reddish the color and the lower the age the more bluish in the color. Hence, we can see that people who are above 60 years and are self employed have formerly smoked in their lives more than anyone else. Also, we can see that children under the age of 20 have formerly smoked too. In all age categories we can see that people who have formerly smoked are the largest numbers, except those who have never worked.

What is the distribution of BMI among different work types and ages?



DATA

WORK.IMPORT

Filter: (none)

ROLES

*X axis: (1 item)

bmi

*Y axis: (1 item)

age

Group: (1 item)

work_type

Legend location:

Outside (default)

ADDITIONAL ROLES

Figure 11 Scatter Plot showing different work types at BMI levels and different age groups.

The figure above shows a scatter plot where BMI and Age constitute the Y axis on both sides and they are grouped based on work type.

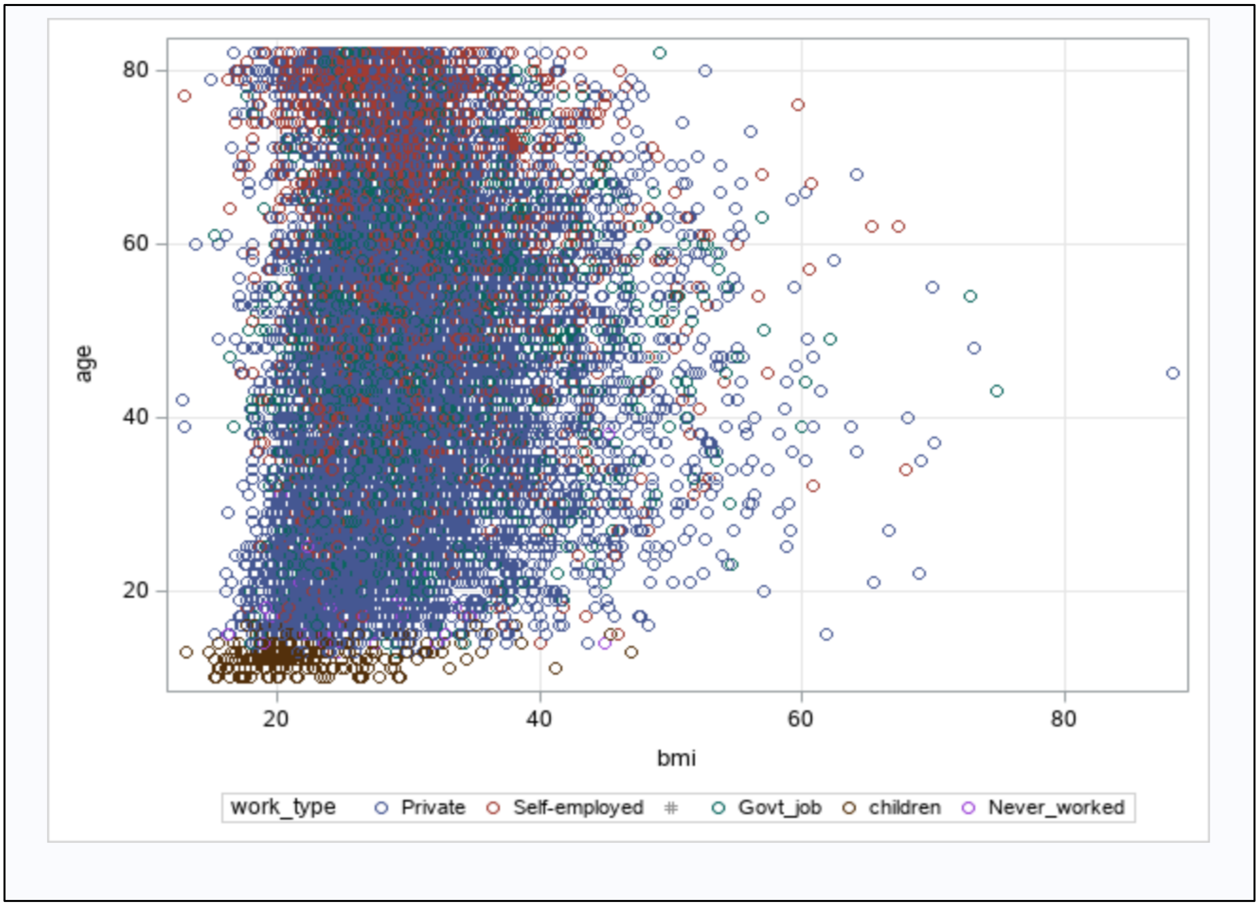
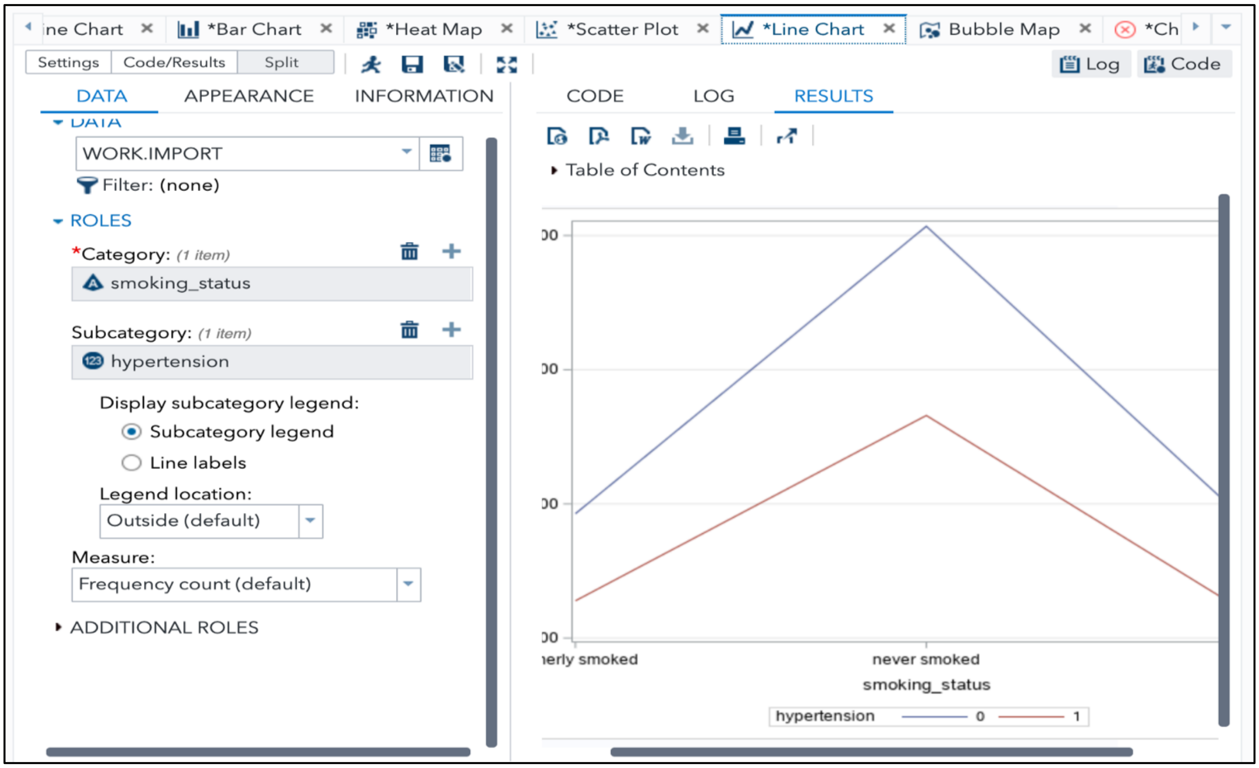


Figure 12 Scatter Plot showing BMI and Age based on Work Type

The Chart above shows the distribution of different work types across different BMI levels. We can see that most BMI levels are between 20-40 the color description of work type shows that most people have private jobs and then followed by government jobs. The maximum sample size lies between the age categories of 40-60 and BMI Level of 30. We can also derive from the plot that more elderly people are self-employed. This chart also can refute any notion that people who have never worked are more likely to have higher levels of BMI.

What is the frequency of hypertension among different categories of smokers and non-smokers?



DATA

WORK.IMPORT

Filter: (none)

ROLES

*Category: *(1 item)*

smoking_status

Subcategory: *(1 item)*

hypertension

Display subcategory legend:

● Subcategory legend

○ Line labels

Legend location:

Outside (default)

Measure:

Frequency count (default)

ADDITIONAL ROLES

Figure 13 SAS Studio Screenshot showing a Line Chart

The Figure above is showing a Line Chart with smoking status as category and hypertension as sub-category. Smoking Statues are; formerly smoked, never smoked, and smoking.

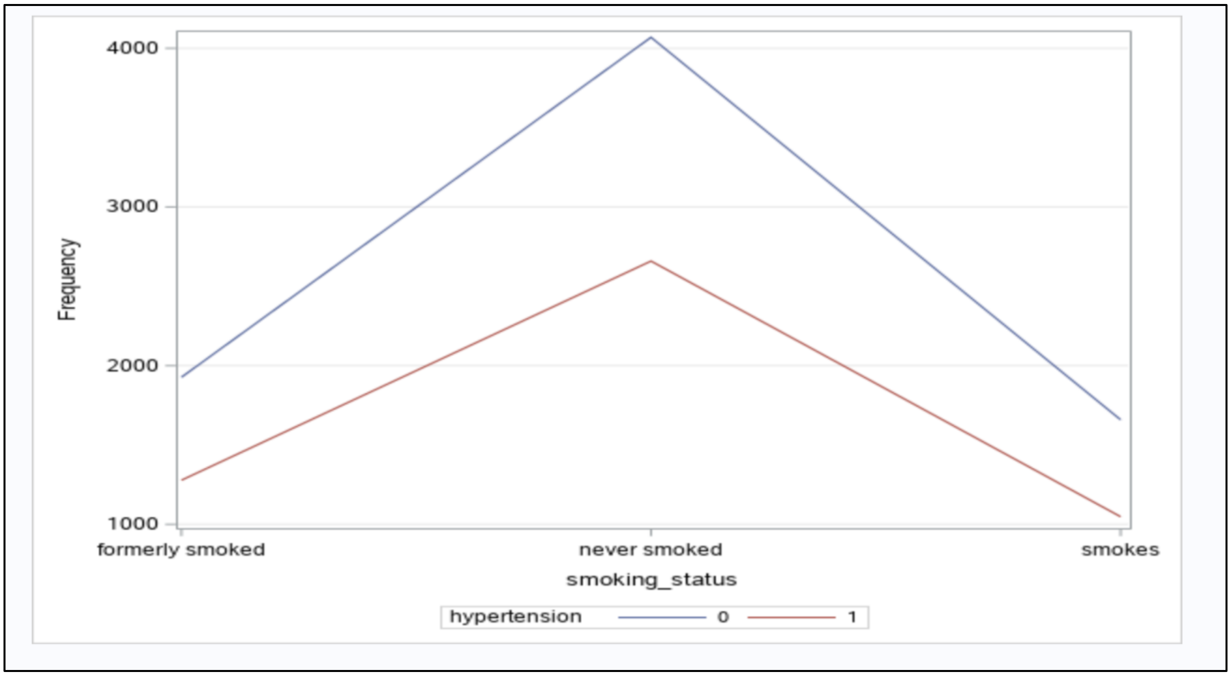
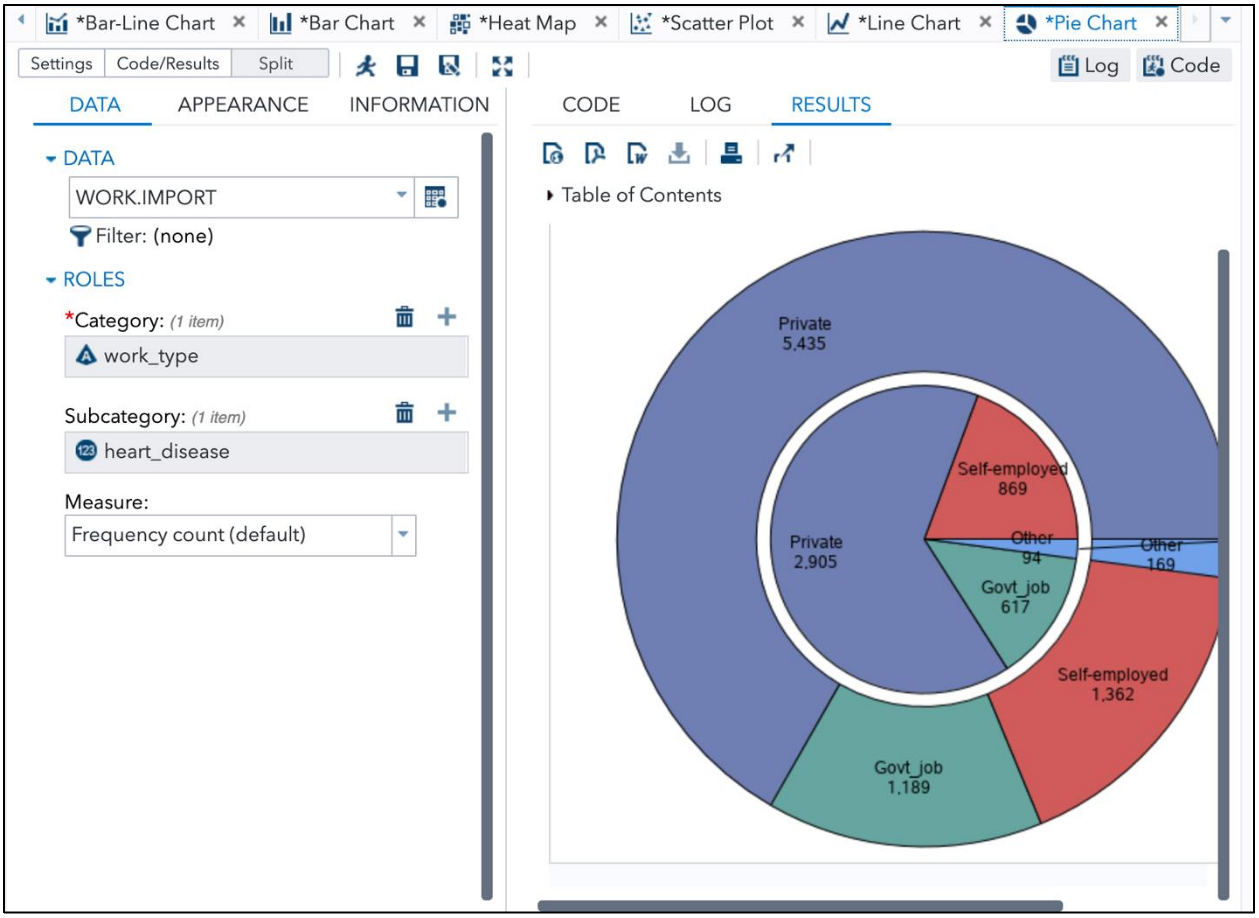


Figure 14 Line Chart showing hypertension frequency and smoking status in relation to hypertension

The chart above is a line chart which shows hypertension frequency among smokers and non-smokers. We can derive that there are higher number of people who do not experience hypertension and have never smoked compared to the people who do not experience hypertension and are currently smoking. The dataset also shows that people who have never smoked are more in the samples of our data set. We can also derive that people who smoke experience the least hypertension among different categories. Hence among smokers we can see that those who smoke are less prone to hypertension.

How is heart disease spread among different work categories?



DATA

WORK.IMPORT

Filter: (none)

ROLES

*Category: *(1 item)*

work_type

Subcategory: *(1 item)*

heart_disease

Measure:

Frequency count (default)

Figure 15 SAS Studio Screenshot showing a Pie Chart

The figure above shows a pie chart from SAS Studio where work type is a category and heart disease is a subcategory on which the pie chart is divided up on.

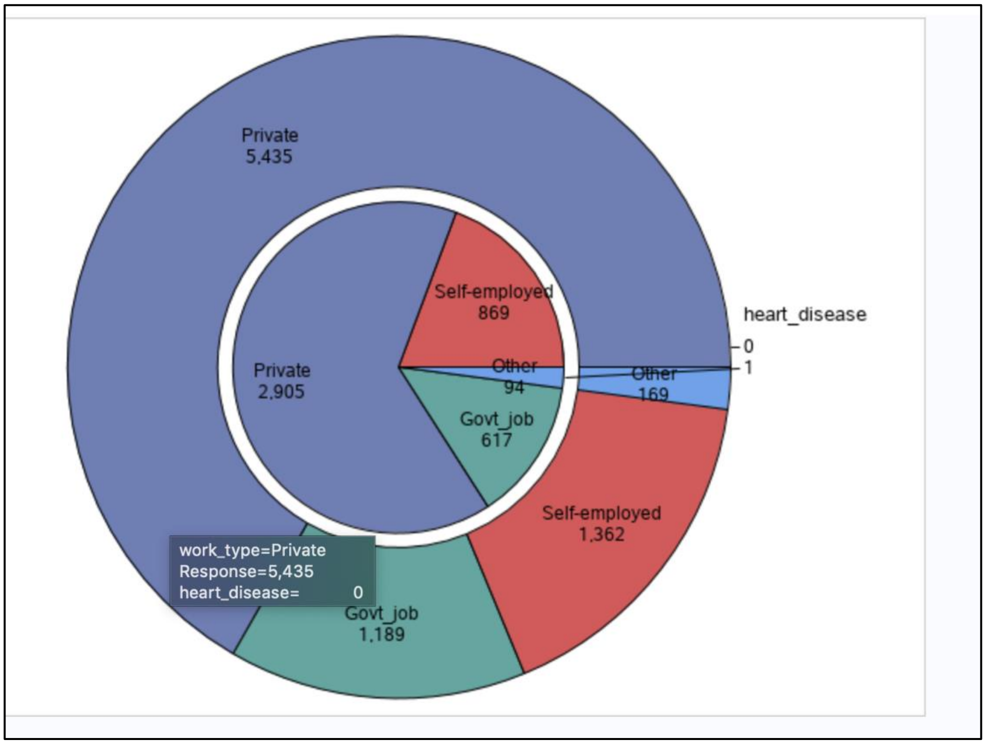
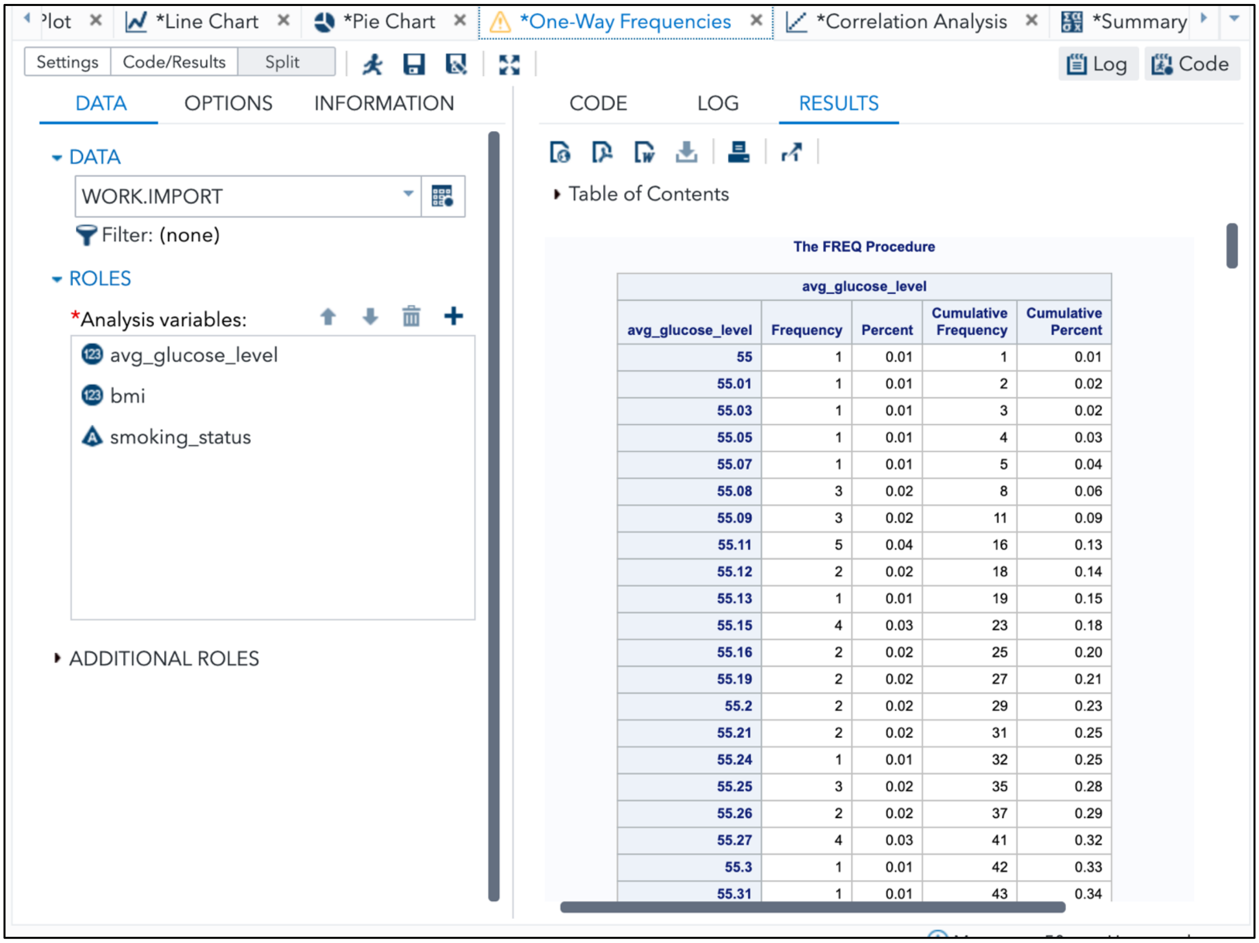


Figure 16 Pie Chart showing hear disease based on work type

The pie chart above shows different categories of work type. Private Category is the largest, while self employed is second most and government jobs the third most common category. Overall, we see that there are more people without heart disease. The outer circle identified with '0' Zero on the right side of the chart shows people without heart disease and the inner circle identified with '1' One shows people with heart disease. Hence, we see that among those with heart disease people working in the private sector have a count of 2,095 and people working in Government jobs have a count of 617. It is clear that the proportion of self-employed people without heart disease to those with heart disease is the most.

Statistics

One Way Frequency Test



DATA

WORK.IMPORT

Filter: (none)

ROLES

*Analysis variables:

- avg_glucose_level
- bmi
- smoking_status

The FREQ Procedure

avg_glucose_level				
avg_glucose_level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
55	1	0.01	1	0.01
55.01	1	0.01	2	0.02
55.03	1	0.01	3	0.02
55.05	1	0.01	4	0.03
55.07	1	0.01	5	0.04
55.08	3	0.02	8	0.06
55.09	3	0.02	11	0.09
55.11	5	0.04	16	0.13
55.12	2	0.02	18	0.14
55.13	1	0.01	19	0.15
55.15	4	0.03	23	0.18
55.16	2	0.02	25	0.20
55.19	2	0.02	27	0.21
55.2	2	0.02	29	0.23
55.21	2	0.02	31	0.25
55.24	1	0.01	32	0.25
55.25	3	0.02	35	0.28
55.26	2	0.02	37	0.29
55.27	4	0.03	41	0.32
55.3	1	0.01	42	0.33
55.31	1	0.01	43	0.34

Figure 17 One Way Frequency Test from SAS Studio

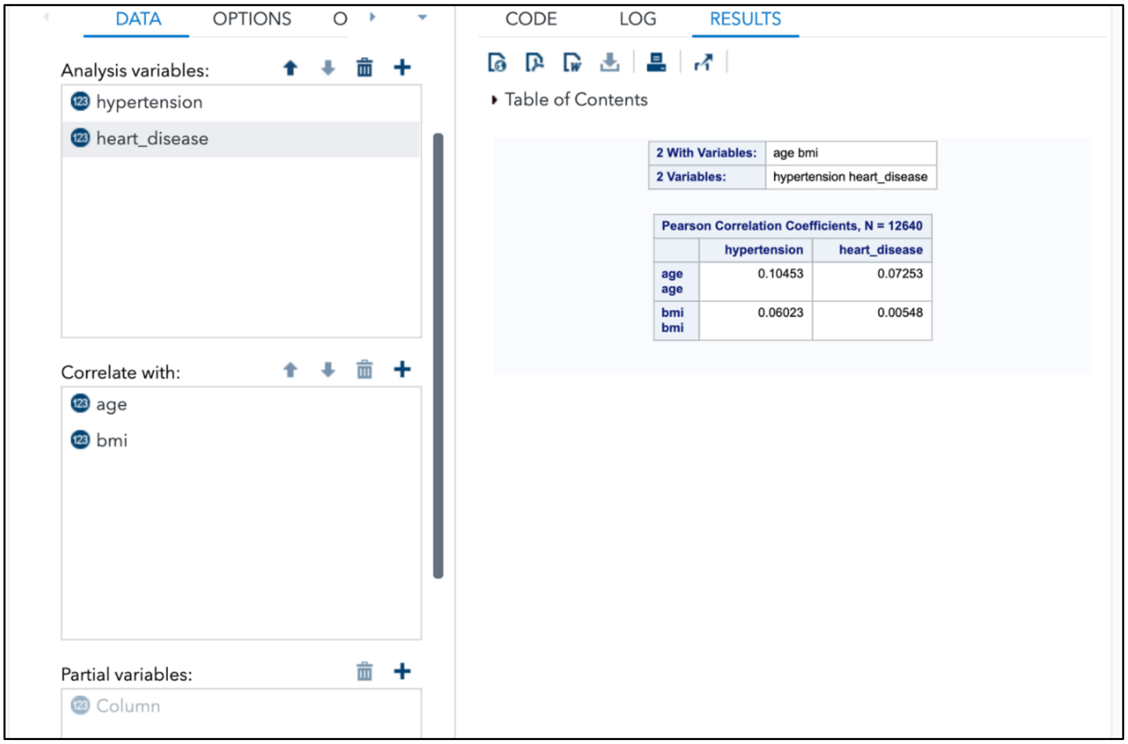
The figure above shows a one-way frequency test that takes into account average glucose levels, BMI and smoking status. We can see that average glucose level of 55.11 has the highest frequency seconded by 55.15. The highest cumulative frequency is the average glucose level of 55.31 with that of cumulative .34 %. It can a When a one-way table shows frequency counts for a particular category of a categorical variable, it is called a frequency table. When a one-way table shows relative frequencies (i.e., percentages or proportions) for particular categories of a categorical variable, it is called a relative frequency table.

From the above given scenario, we are trying to figure out the frequency counts for a category in this instance avg_glucose_level and its relative frequencies.

From our table, as the glucose level increases the count of people diagnosed with it also increases. For instance, the avg_glucose level of 55.27 has frequency count of 4 and relative frequency percent of 0.03.

Also it can be derived that Average glucose levels are above 55 in all cases.

Correlation Analysis



Analysis variables:
hypertension
heart_disease

Correlate with:
age
bmi

Partial variables:
Column

2 With Variables:	age bmi
2 Variables:	hypertension heart_disease

Pearson Correlation Coefficients, N = 12640		
	hypertension	heart_disease
age age	0.10453	0.07253
bmi bmi	0.06023	0.00548

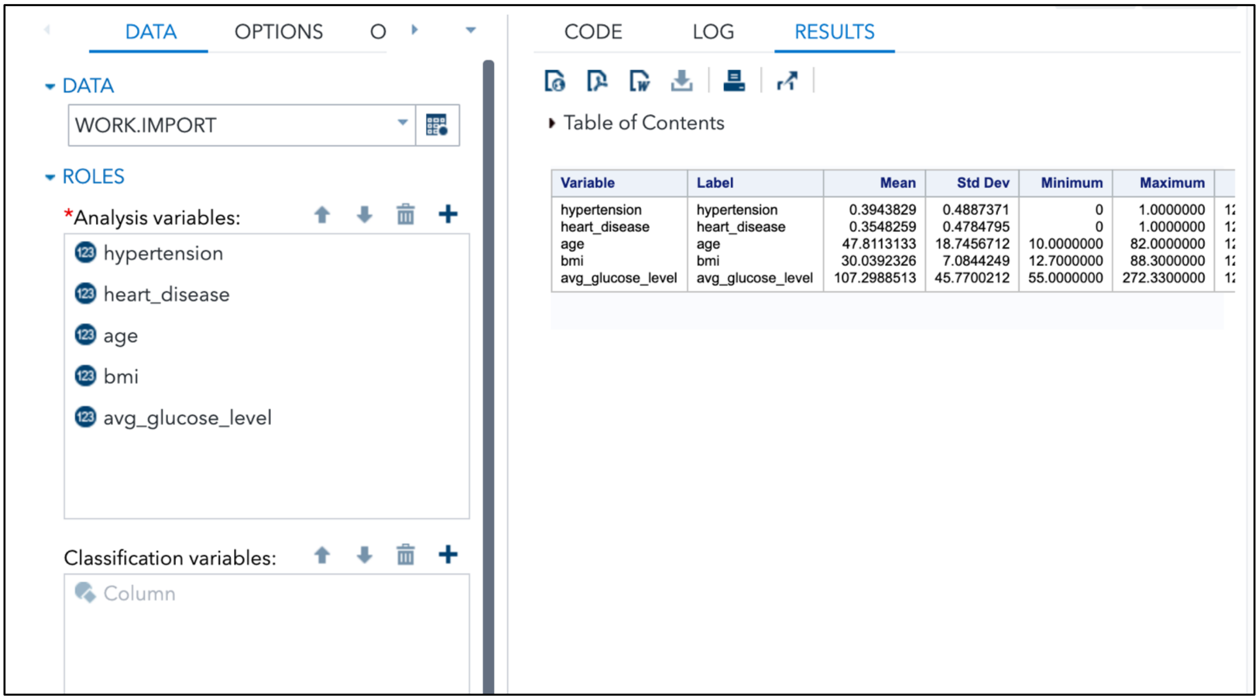
Figure 18 SAS Studio Correlation Analysis

The figure above shows correlation between hypertension and heart disease with age and BMI. We can see that age and hypertension have the highest correlation while hypertension and heart disease have the least correlation. It can also be derived that heart disease and age have the highest correlation while heart disease and BMI have the lease t correlation. This visualization is very helpful in understanding different variables and their relations with each other in order to conduct further analysis or statistical regressions. First, we find a correlation then we conduct an analysis.

Correlation analysis deals with relationships among variables. Correlation analysis in SAS is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables (e.g. height and weight). SAS Correlation analysis is a particular type of analysis, useful when a researcher wants to establish if there are possible connections between variables. In other words, it’s a measure of how things are related.

In the above visual, we have tried to establish correlation between different variables like age, BMI, Hypertension and heart disease. From our understanding, there is a strong relation between age - hypertension and age – heart disease. From the given correlation, we can interpret that with people of different age brackets are more prone to hypertension and stroke. In today’s stressful life, it more pre-dominant that young students are most likelihood of stroke and hypertension as compared to their older counterparts.

Statistical Summary



DATA

WORK.IMPORT

ROLES

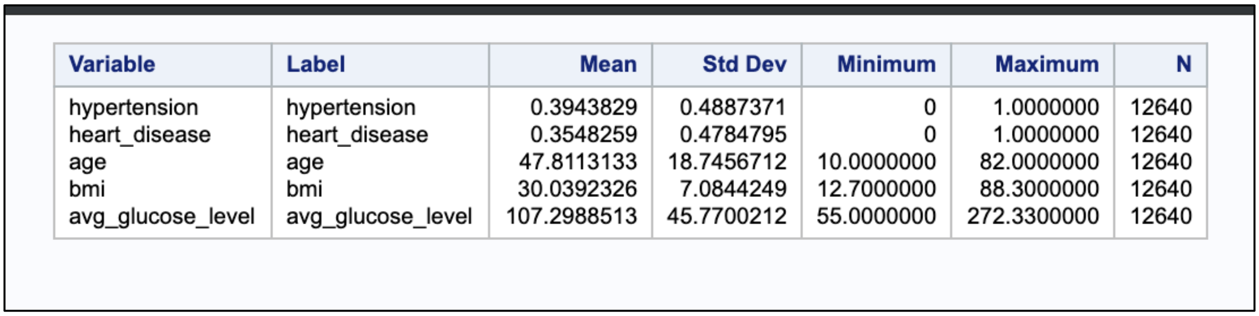
*Analysis variables:

- hypertension
- heart_disease
- age
- bmi
- avg_glucose_level

Classification variables:

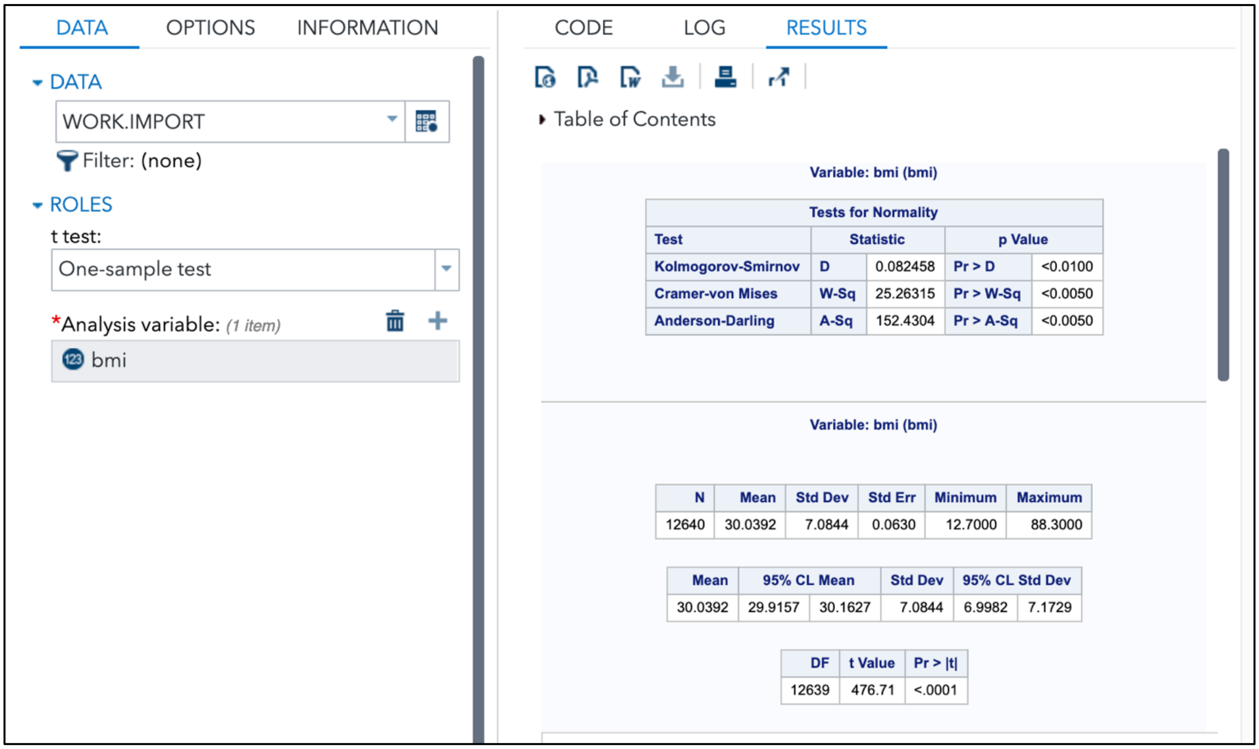
- Column

Variable	Label	Mean	Std Dev	Minimum	Maximum
hypertension	hypertension	0.3943829	0.4887371	0	1.0000000
heart_disease	heart_disease	0.3548259	0.4784795	0	1.0000000
age	age	47.8113133	18.7456712	10.0000000	82.0000000
bmi	bmi	30.0392326	7.0844249	12.7000000	88.3000000
avg_glucose_level	avg_glucose_level	107.2988513	45.7700212	55.0000000	272.3300000



Variable	Label	Mean	Std Dev	Minimum	Maximum	N
hypertension	hypertension	0.3943829	0.4887371	0	1.0000000	12640
heart_disease	heart_disease	0.3548259	0.4784795	0	1.0000000	12640
age	age	47.8113133	18.7456712	10.0000000	82.0000000	12640
bmi	bmi	30.0392326	7.0844249	12.7000000	88.3000000	12640
avg_glucose_level	avg_glucose_level	107.2988513	45.7700212	55.0000000	272.3300000	12640

The Statistical Summary above shows us the different variables and their mean, minimum and standard deviations. The smaller box with clear labels clearly shows the details of the statistical analysis in respect to Mean and Standard Deviation.



DATA

WORK.IMPORT

Filter: (none)

ROLES

t test: One-sample test

*Analysis variable: (1 item)

bmi

Variable: bmi (bmi)

Tests for Normality				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	D	0.082458	Pr > D	<0.0100
Cramer-von Mises	W-Sq	25.26315	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	152.4304	Pr > A-Sq	<0.0050

Variable: bmi (bmi)

N	Mean	Std Dev	Std Err	Minimum	Maximum
12640	30.0392	7.0844	0.0630	12.7000	88.3000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
30.0392	29.9157	30.1627	7.0844	6.9982	7.1729

DF	t Value	Pr > |t|
12639	476.71	<.0001

Figure 19 T Test table from SAS Studio

The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the **null hypothesis (H_0)** of a study question is true – the definition of ‘extreme’ depends on how the hypothesis is being tested. P is also described in terms of rejecting **H_0** when it is actually true, however, it is not a direct probability of this state. The term **significance level (alpha)** is used to refer to a pre-chosen probability and the term "P value" is used to indicate a probability that you calculate after a given study.

The **alternative hypothesis (H$_1$)** is the opposite of the null hypothesis; in plain language terms this is usually the hypothesis you set out to investigate

In our t-test, the p-value is less than 0.0001 then we reject the null hypothesis i.e. accept that our sample (BMI) gives reasonable evidence to support the alternative hypothesis. The variable i.e. BMI whose mean is compared to the hypothesized population mean (i.e., Test Value).

Based on the results, we can state the following:

- There is a significant difference in mean BMI value between the sample and the overall adult population ($p < .001$).
- The average BMI of the sample is about 0.12 kg/m^2 than the overall adult population average

References

(2014, 05 15). Retrieved from National Library of Medicine : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4048595/

(2017, 03 14). Retrieved from British Heart Foundation: https://www.bhf.org.uk/informationsupport/risk-factors/smoking

(2018, 09 25). Retrieved from www.geneticsandfertility.com