

# 2010

Fernandez Hospital Pvt Ltd

Academics Dept & Clinical Research Unit

## [MICROSOFT EXCEL FOR DATA ENTRY]

This document is primarily meant as a guide for students who have to create their own data entry sheets for their dissertations and for those who are scared of spreadsheets...

## Introduction

Data collection is an important part of research that aims to answer a specific research question. The logical step, after data collection, is to analyze the collected data and make meaningful and appropriate interpretations of the data.

Microsoft Excel is a commonly used and rather popular tool for data entry and manipulation. Excel has some statistical capabilities and you can download several add on plug-ins that increase the statistical capabilities of Excel. However, I prefer using statistical software for data analysis as opposed to Excel- call it a personal preference or bias, but somehow, I have not been entirely convinced with the statistical manipulations possible with Excel.

We could use one of several databases for data entry and storage. Excel, is however widely available and used, thanks to the market advantage of Microsoft. We shall look at making use of excel for data entry.

This guide is based on the use of Excel 2007 for data entry.

## The Data Entry-Prior Preparation

The major part of any analysis is not the actual analysis but the preparation for analysis. A properly prepared dataset does not take too much time to analyze. A dataset that is poorly prepared can however lead to frustrating delays with the analysis.

There is a process that can help optimize data entry and give you a clean or almost clean dataset for analysis. This applies to the use of any spreadsheet or database for data entry.

1. Clarity on your research question, your hypothesis, your aims and objectives
2. Clarity on the variables chosen for the study
3. A specific sharp research question, hypothesis, aims and objectives will allow for the use of very specific variables that will allow you to answer your specific question- you will have a very specific data collection format
4. Collecting data on as many variables as possible is not useful unless you will use the data and the data is pertinent to your research question. Remember, you enter your data or pay someone to enter it for you. You do not want to waste your time or money for data you will never use or will not help you answer your question.
5. A proper and thoughtful preparation prior to data collection will ensure that you collect data in a manner that facilitates both data entry as well as data analysis.
6. Prepare your data collection form carefully.
7. Take a piece of paper (or use a word processor as you wish) and write down the specific variables that will help you answer your research question

## The Variables

1. Each variable has a unique name
2. Use short names for variables preferably a maximum of 8 letters
3. Do not use spaces between variable names
4. Use an underscore if you wish to have two names. Example: Gest\_diab instead of gest diab, Sys\_BP instead of systolic BP
5. Each variable identifies a separate piece of data
6. The structure of each variable is predefined
7. Each variable serves a purpose in answering your question
8. You are clear as to how you want to use the variable to answer your question

Example of a chart of variables for a study

Name of Variable	Short Name for the variable	What does it indicate	What is the structure of the variable	Unit of measurement	What is the purpose of the variable	How will you use this variable
Systolic BP	Sys_BP	Systolic Blood Pressure	Numeric, Continuous variable	Mm Hg	Allows me to define hypertension	I will present the distribution of systolic BP (mean, median, range) as well as use this value to categorize people as hypertensive or not
Education	Educ	Education of the person	Categorical variable		Allows me to explore for any potential associations with education levels of the person	I will use this variable to explore for possible differences based on the education of the person

Hypothyroidism	Hypo_thyr	Indicates if the person has hypothyroidism or not	Yes/No (dichotomous response)	0=No 1=Yes	Allows me to categorize people as hypothyroid or not based on predefined criteria	I will use this variable to explore for associations with outcomes of interest
Satisfied with service	Satisfaction	Indicates the level of satisfaction of the person	5 point scale from very dissatisfied, dissatisfied, neither satisfied nor dissatisfied, satisfied, very satisfied	0-4	Allows me to determine the level of satisfaction with services	I will use this variable to explore for associations with outcomes of interest and to see if poor outcomes relate to poor satisfaction

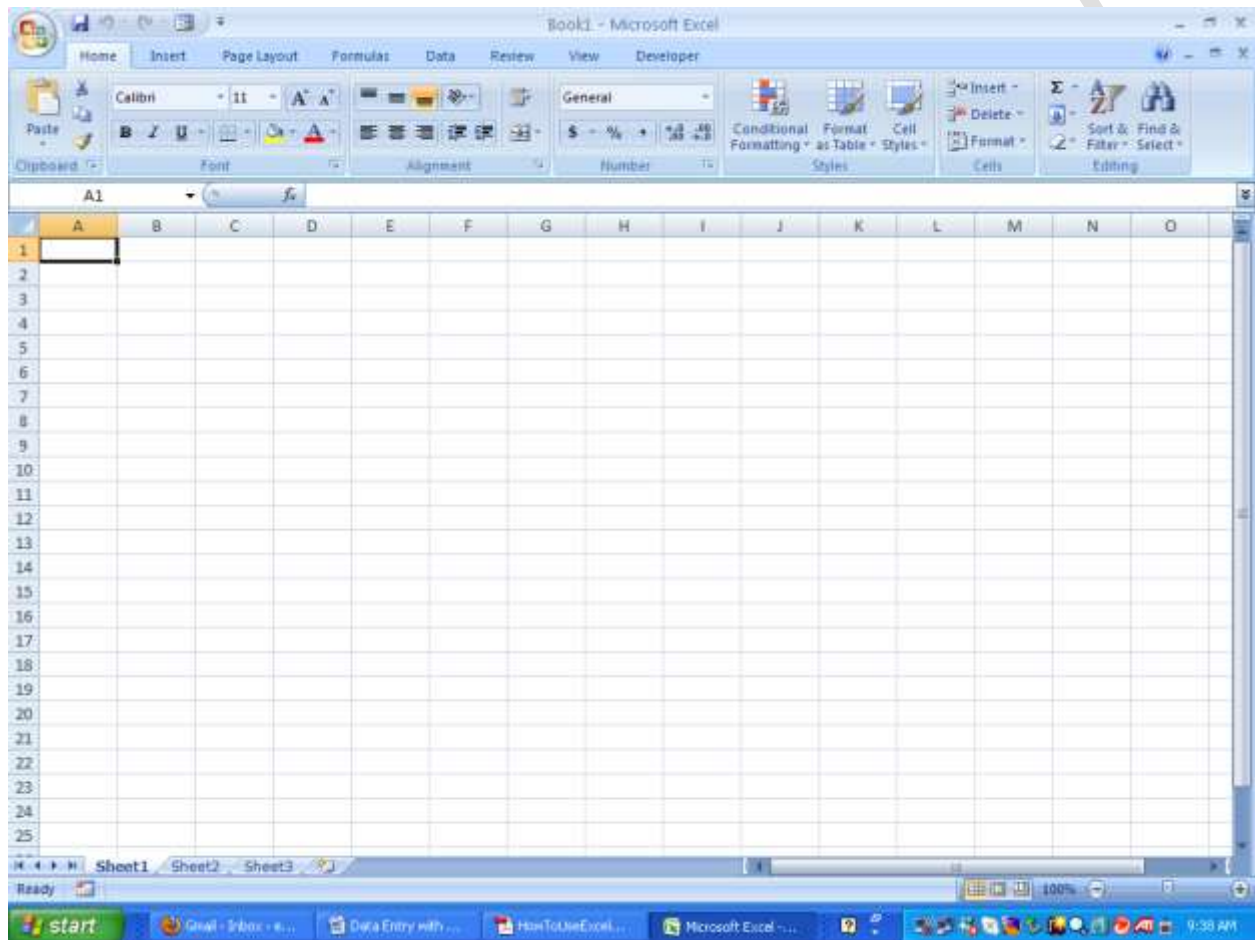
## Data Entry- Data Structure

1. All your data should be in a single spread sheet of a single file
2. Name the sheet appropriately
3. If you use multiple sheets, make sure there are unique identification numbers for each person that will help you to merge the sheets
4. Have the complete data of each person in a single row
5. Have the complete data of each variable in a single column
6. Thus, a row (horizontal direction) should cover the data of a person, while a column (vertical direction) should cover the data of a variable

## Data Entry- The Process

At this stage, we have a research question, we have the variables of interest and we have collected data. We now need to enter the data to analyze in a format that facilitates easy analysis. Remember, the primary aim of data entry is to have the data in a format that facilitates data analysis.

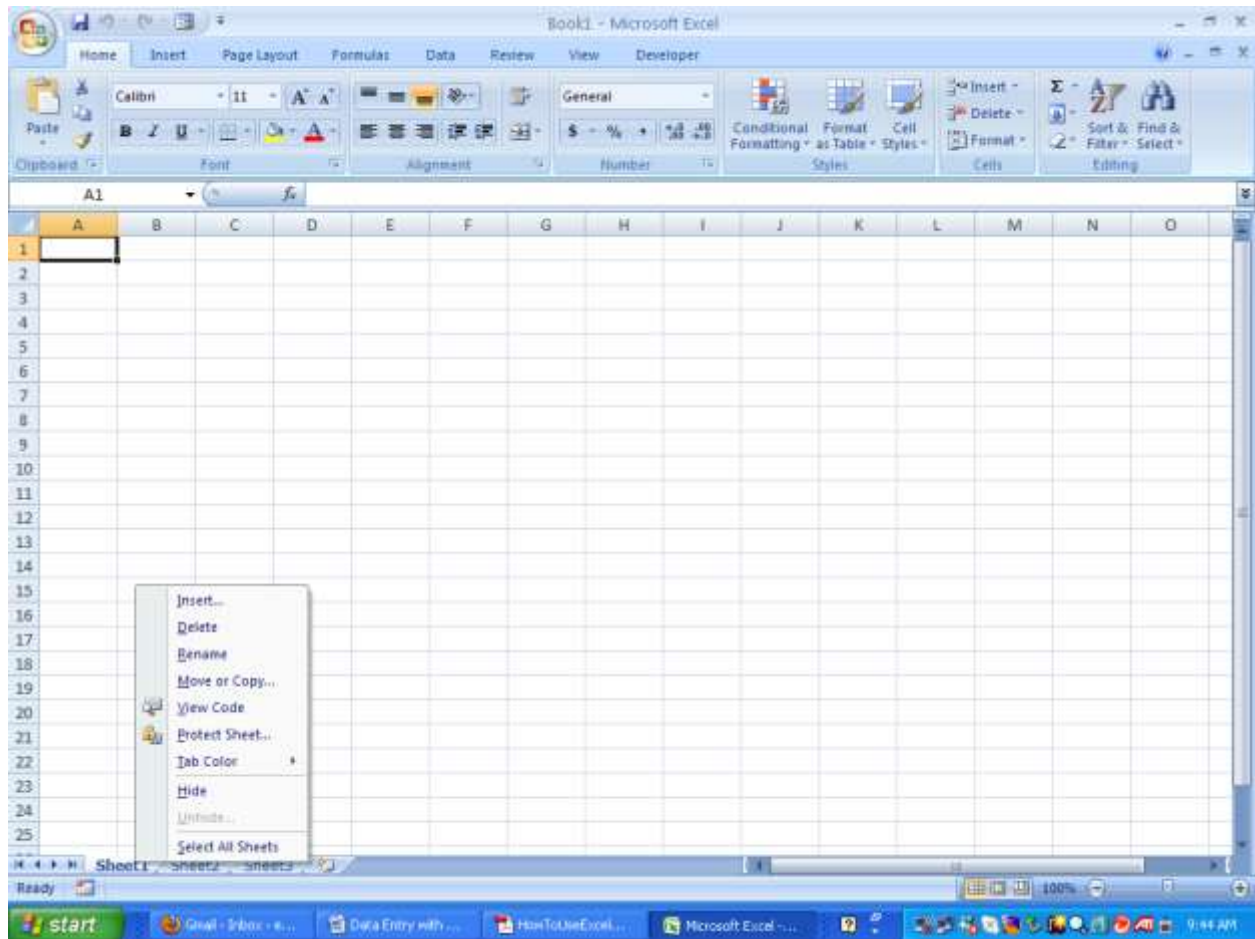
Open MS Excel in your personal computer



This is how a blank excel spreadsheet looks.

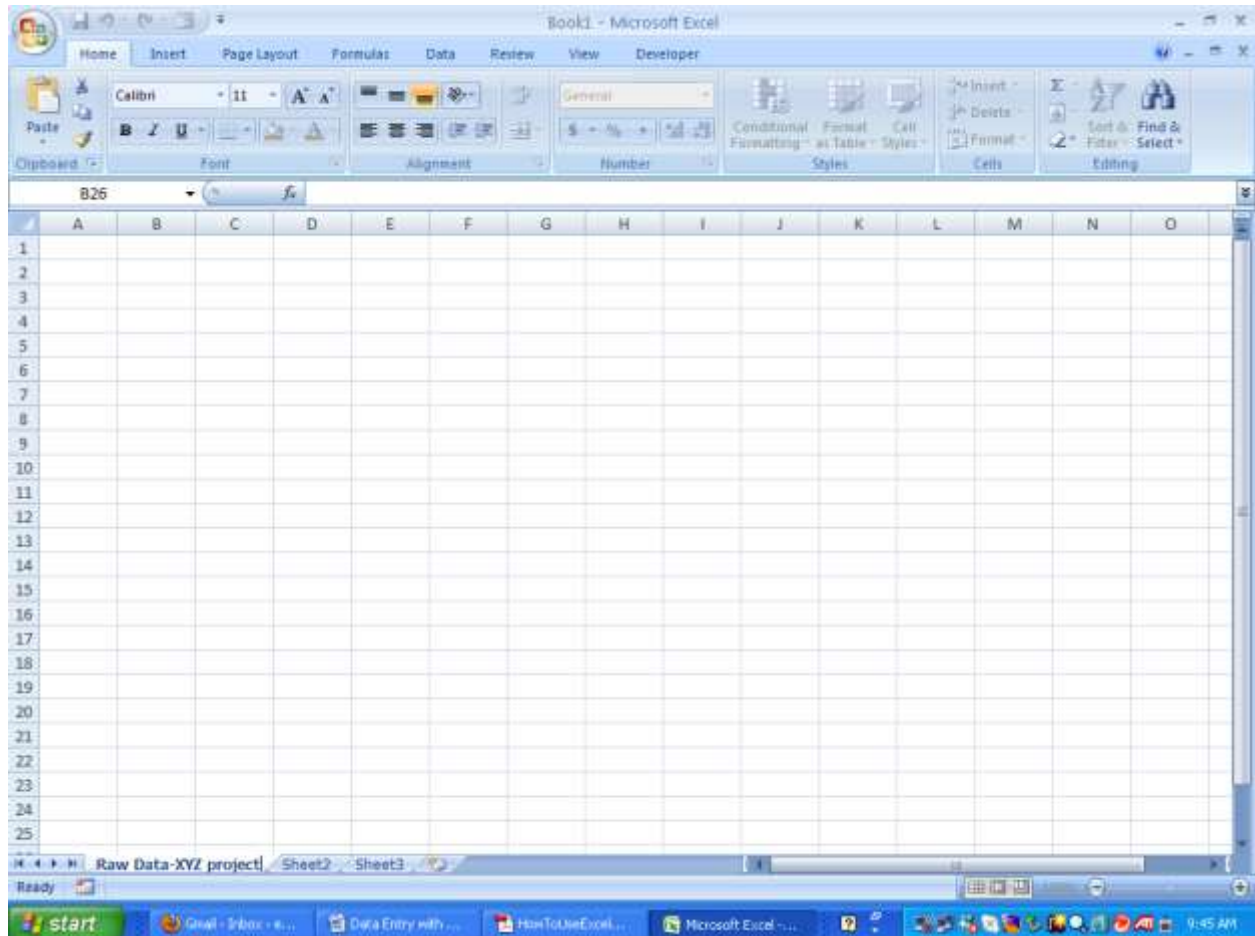
You can see the name of the worksheet at the bottom- Sheet 1, Sheet 2, Sheet 3. We are currently on Sheet 1.

Let us rename the sheet to indicate the project we are working on. Use your mouse to bring the cursor to sheet 1. Right click and you will get a pop up screen with multiple options



Click on Rename. Sheet 1 is now highlighted and you can type in the new name for the sheet. Save your work in the appropriate folder. Save your work frequently. You don't want to lose it!

I have named the sheet Raw Data-XYZ project. Name the sheet appropriately so that you are aware of what the sheet contains. My name now indicates that this sheet pertains to XYZ project and contains the raw data. I do this because I will never use the raw data sheet for analysis-I always copy that sheet and save it under a different name before I work or clean the raw data for analysis.



On the top, you can see the alphabets A, B, C, etc. Each alphabet indicates a column.

On the left hand side you can see the numerics 1, 2, 3 etc. Each number indicates a row.

Let us now get ready to populate the excel spreadsheet with the data we collected.

## STEP 1:

**Important: The top row is always meant for the names of the variables.**

DO NOT use this row for any other purpose including writing the title of the project. We have already named the worksheet based on the title of the project.

Why is this important? When you transfer your data to statistical software, the software will read the first row as the names of the variables and any thing in subsequent rows will be read as data. Do not have any other row of text in the spreadsheet

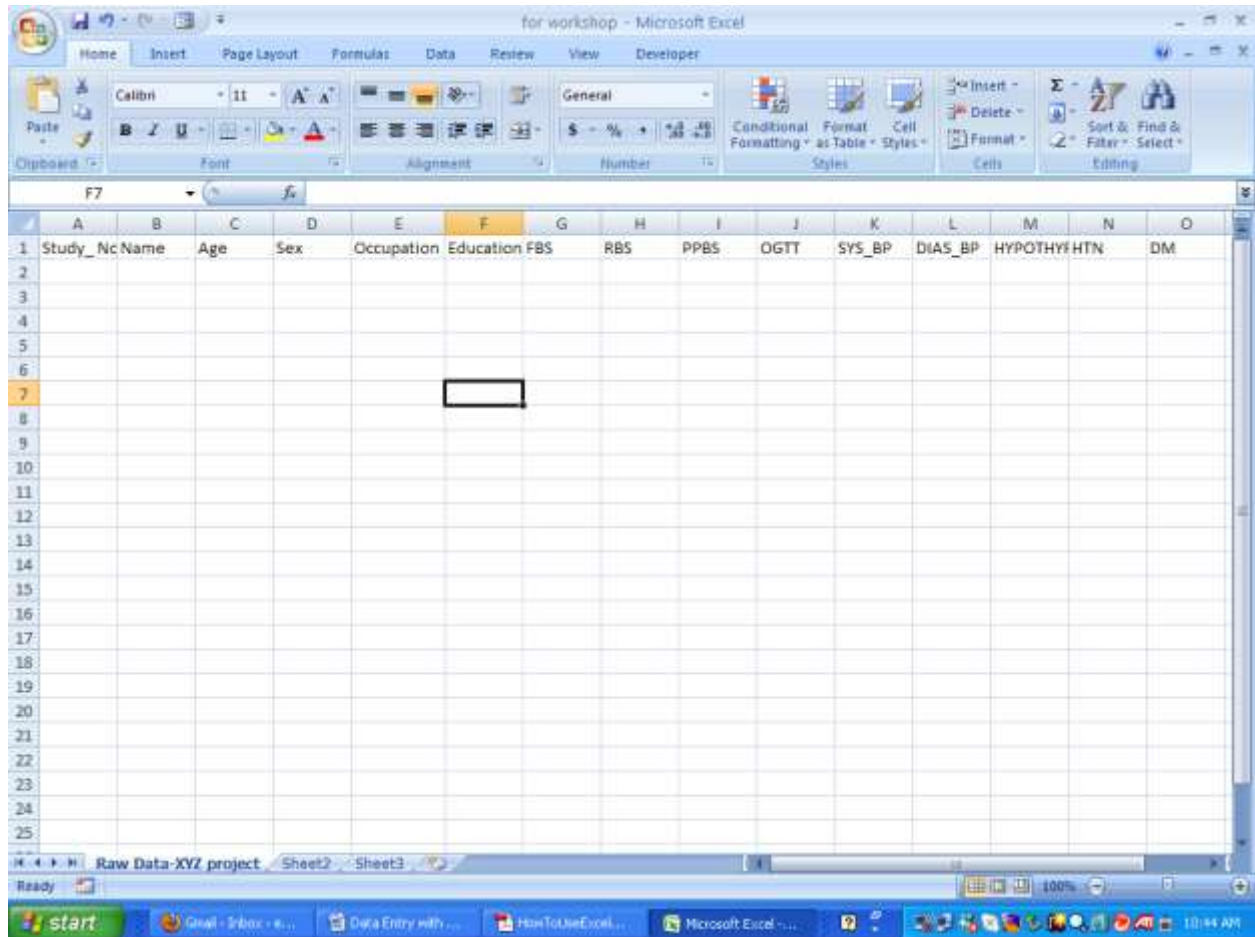
*To put it simply, having anything other than the names of your variables in the first row will only confuse the software and complicate your analysis.*

Let us start entering the name of the variables.

Click on cell A1 and type the first variable.

Tip: Always use a unique identifier number like a study number that will help you to easily identify the person. Do not use the numbers on the left hand side of the excel sheet (the row numbers) for this purpose. You may need to sort the data later, so the row number in Excel would then apply to a different subject or sampling unit.

Once you have clicked and entered the name of first variable in cell A1, repeat the same for each succeeding variable in cell B1, C1, D1 etc.

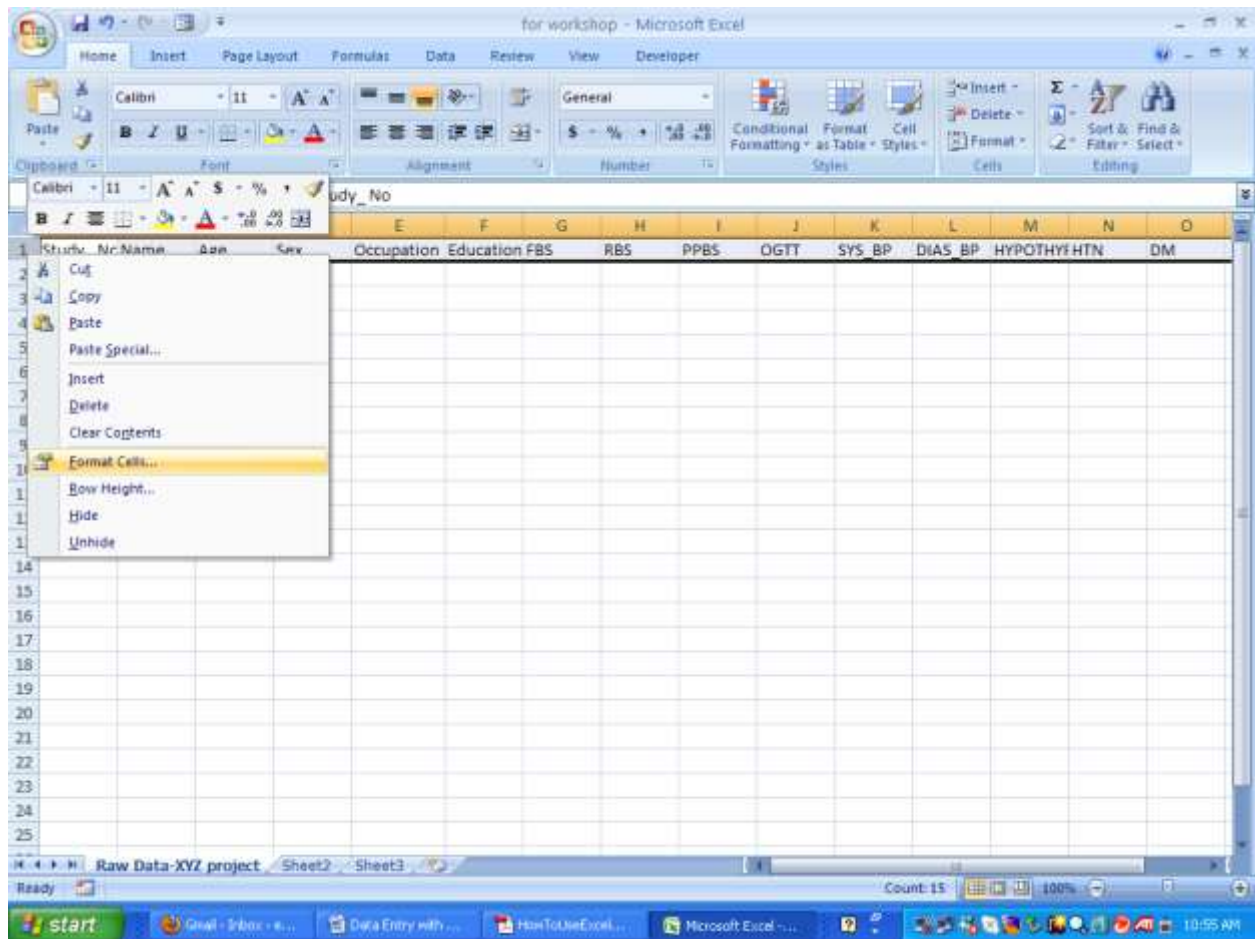


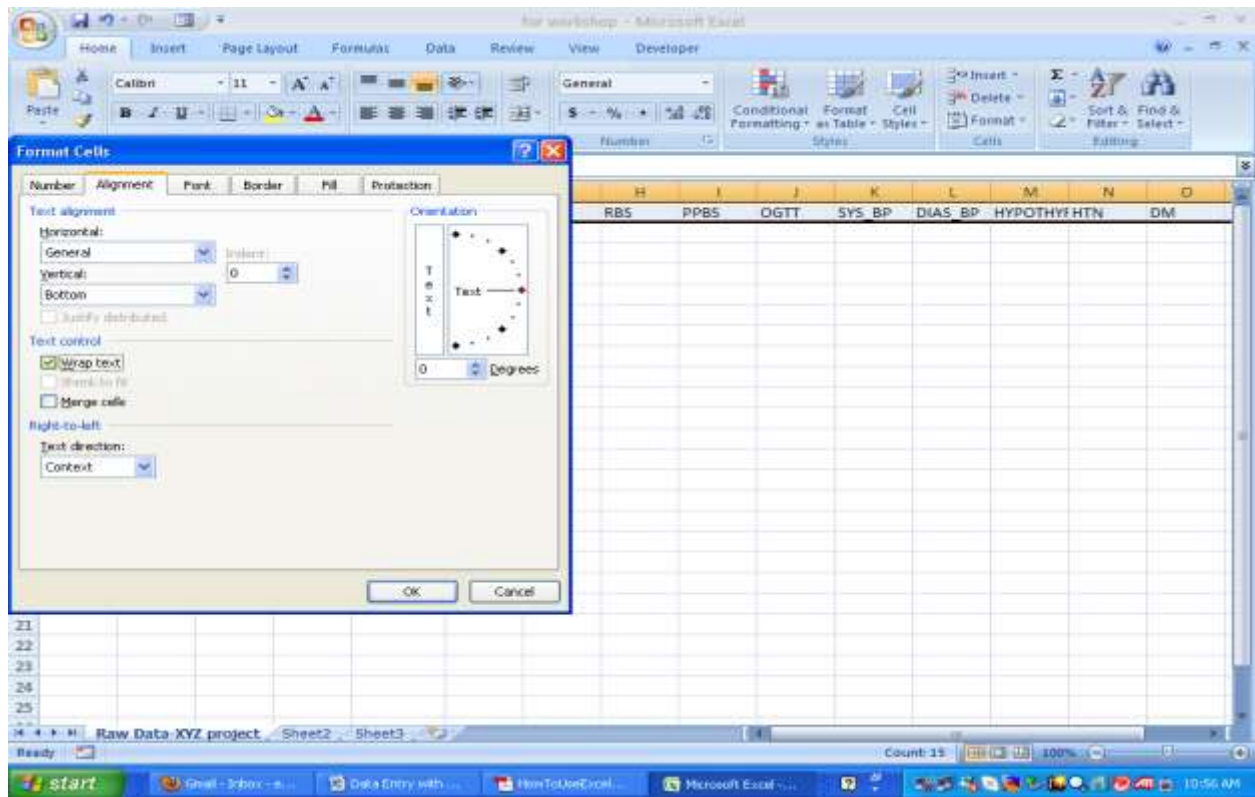
Look at the variable name in column M. This does not read completely, does it?

Select the row 1 by clicking on the row number 1 on the side

Then right click your mouse

A pop up menu will appear- choose format cells and click on that. Another pop up menu will appear (see below). Click on alignment and then select wrap text. Click on ok. The screen with the variable names looks better now (see below for screen shots).





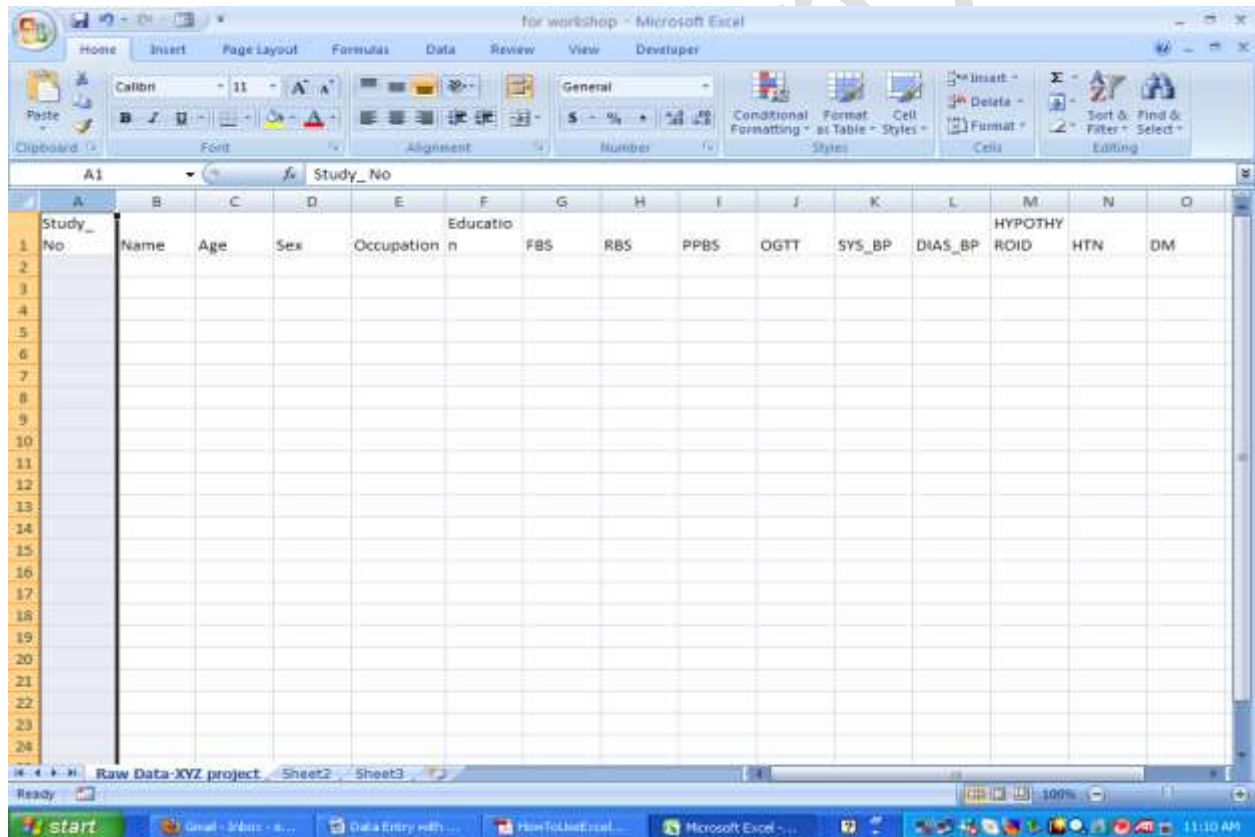


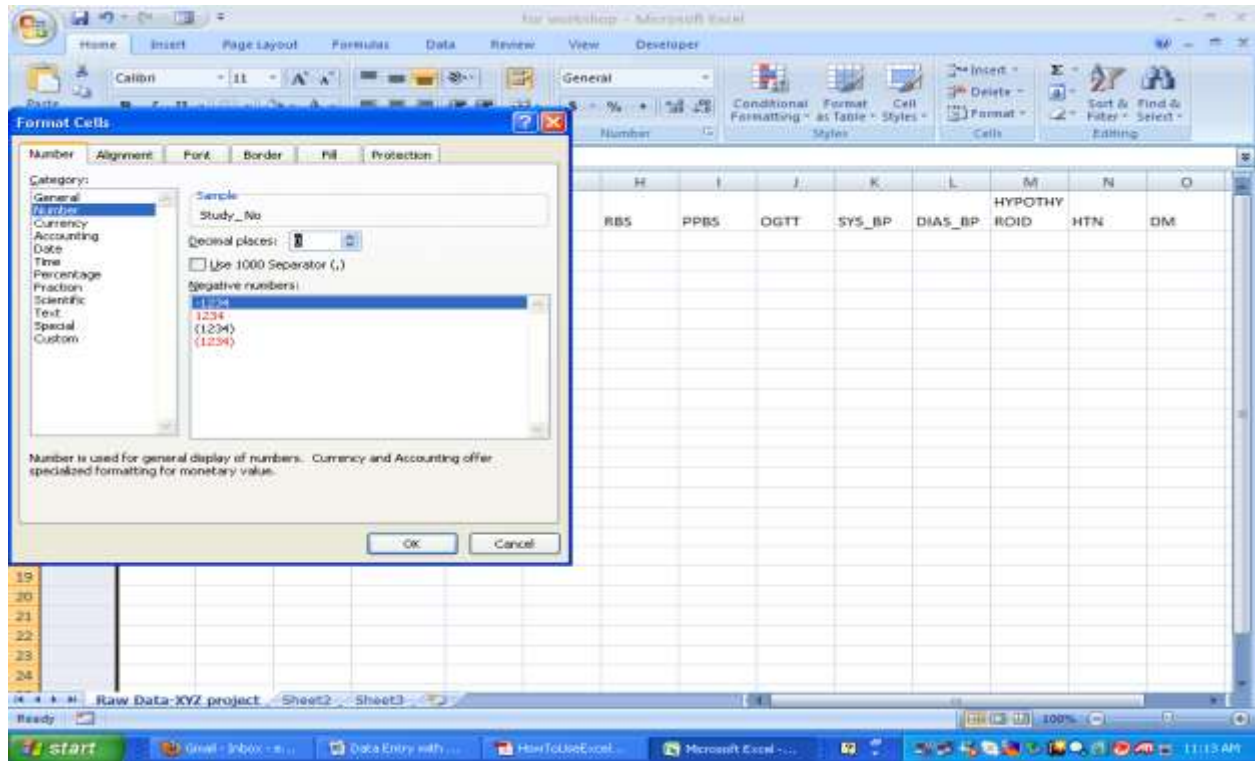
## STEP 2

Let us now look at the structure of the variables.

For this example, I am assigning Study\_No as a numeric variable- 1,2, 3, etc. I will now set the cells under column A to accept only numerical inputs.

Move your mouse cursor onto A and click to highlight column A. Next right click your mouse to get a pop up menu. Choose format cells and click. In the next pop up menu (we have done this step earlier) choose the number tab on the top. See below for the screen shot. Under categories, choose number, set the decimal points to zero, keep the 1000 separator unchecked and then click ok.

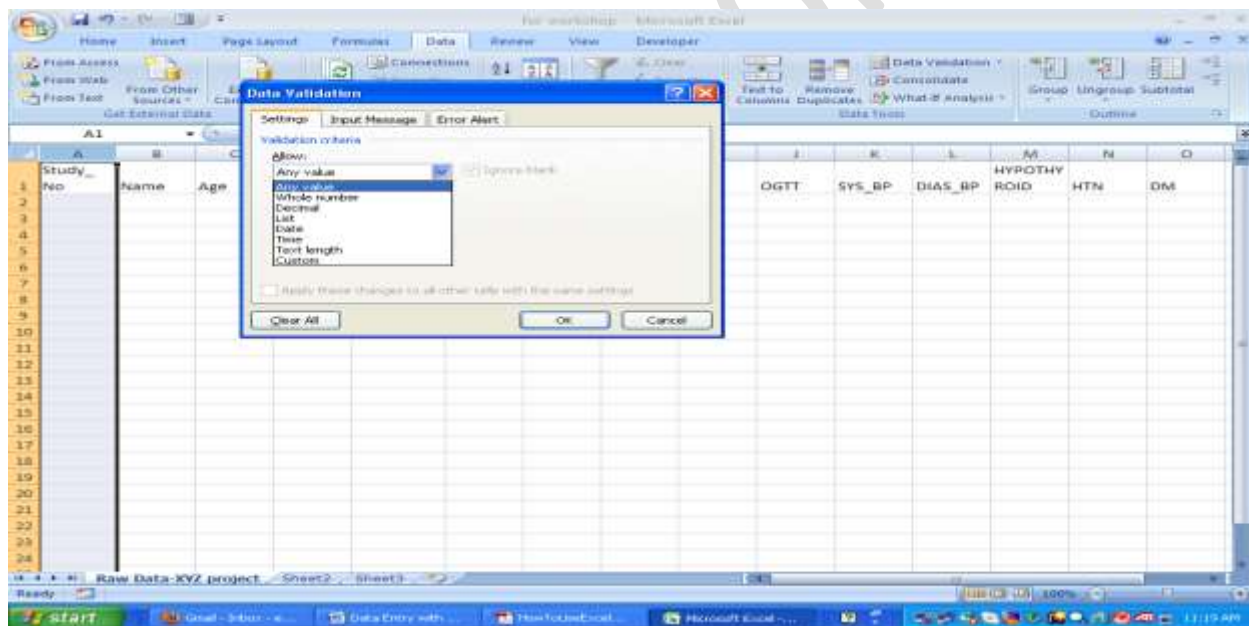
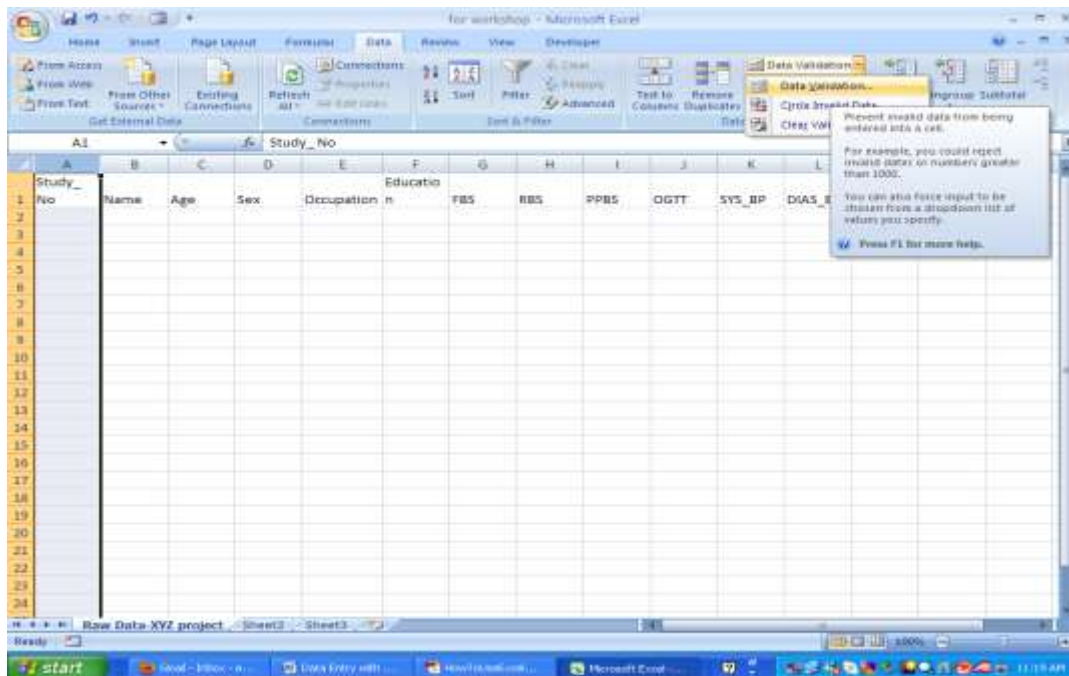




This is one way of formatting the cells to contain numbers. Try typing AA in cell A2 now. Surprise, although we set the format to be a number, this is still allowed. This can lend itself to data entry errors.

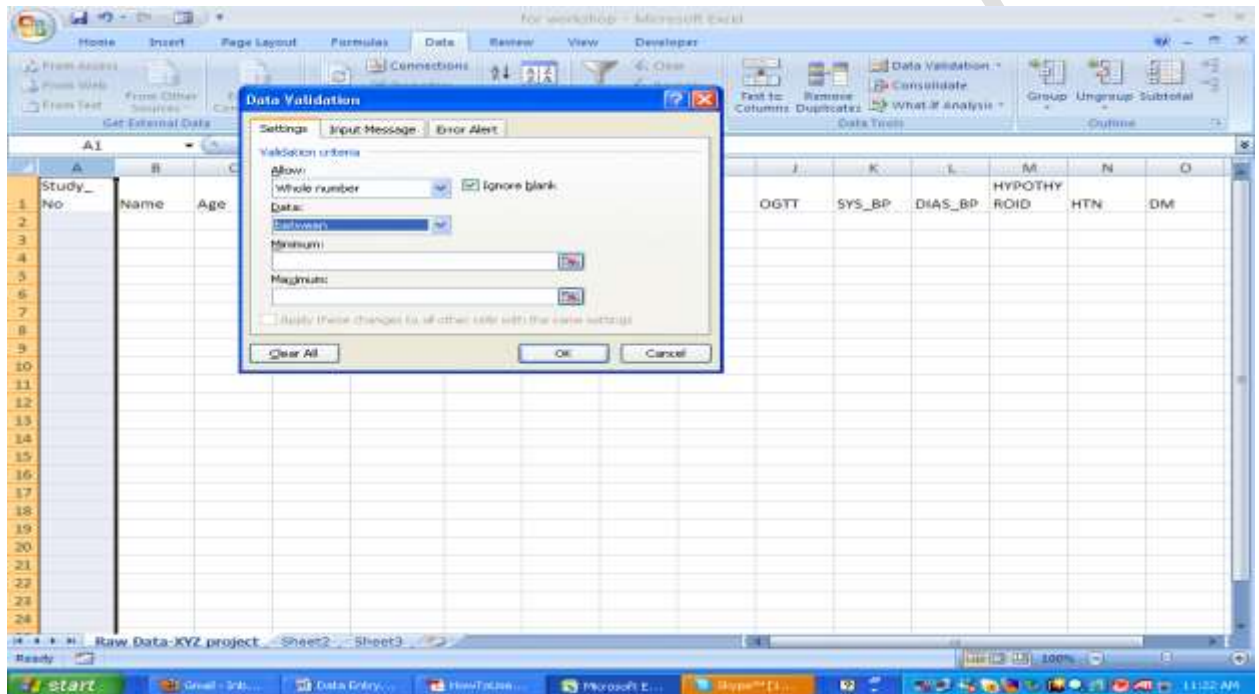
How do we make sure that only whole numbers are allowed?

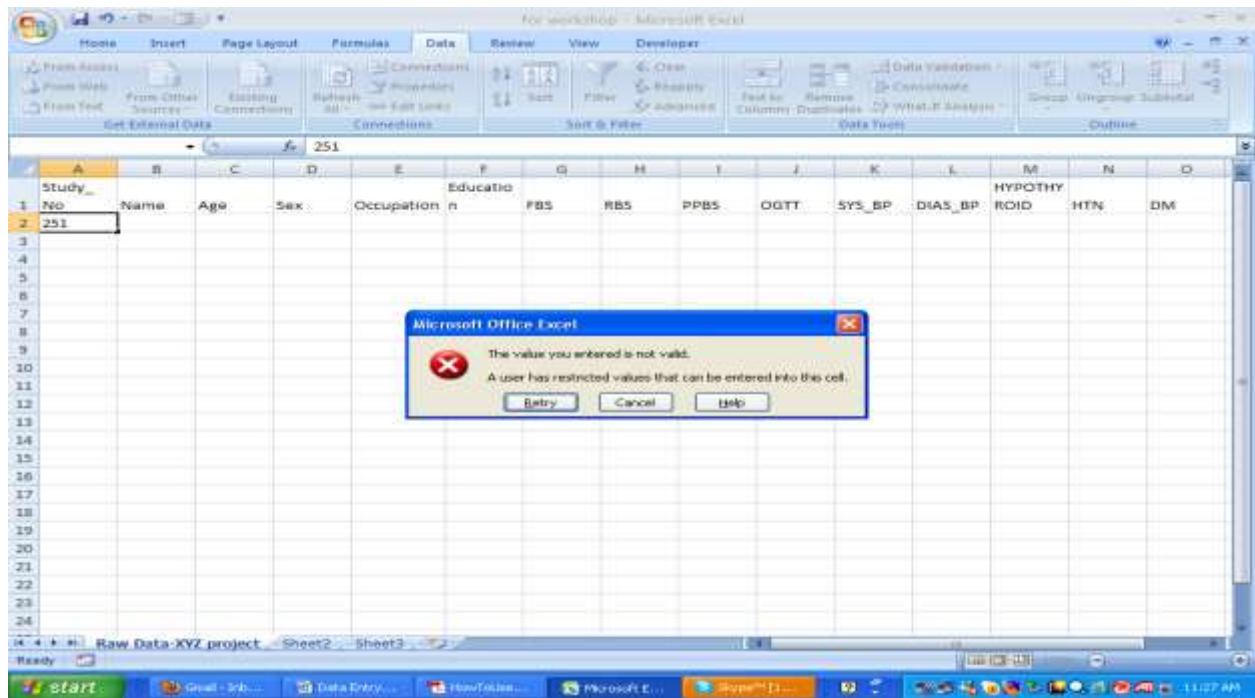
- Highlight Column A.
- Click on Data on the tabs at the top
- Click the arrow next to data validation
- Click on data validation



- In the pop up box, click on the drop down arrow next to any value
- Choose whole numbers
- You get additional choices to make
- You want to have a study number for each person, so uncheck the ignore blank
- You can now set the range of the data-for a study that will have 250 subjects, the minimum can be 1 and the maximum can be 250

- Check the drop down menu under data, there are different ways to do this.
- Click ok
- Try entering 251 in cell A2 (let us see if it works)- we will now get a error screen, select cancel
- Try entering AA in cell A2- we still get the error screen
- *We have now set column A to include only whole numbers.....*





Now check which of your variables are continuous numeric variables and repeat the same process. In the excel sheet that we have illustrated as an example, this can include Age, FBS, RBS, PPBS, OGTT, sys\_BP, Dias\_BP.

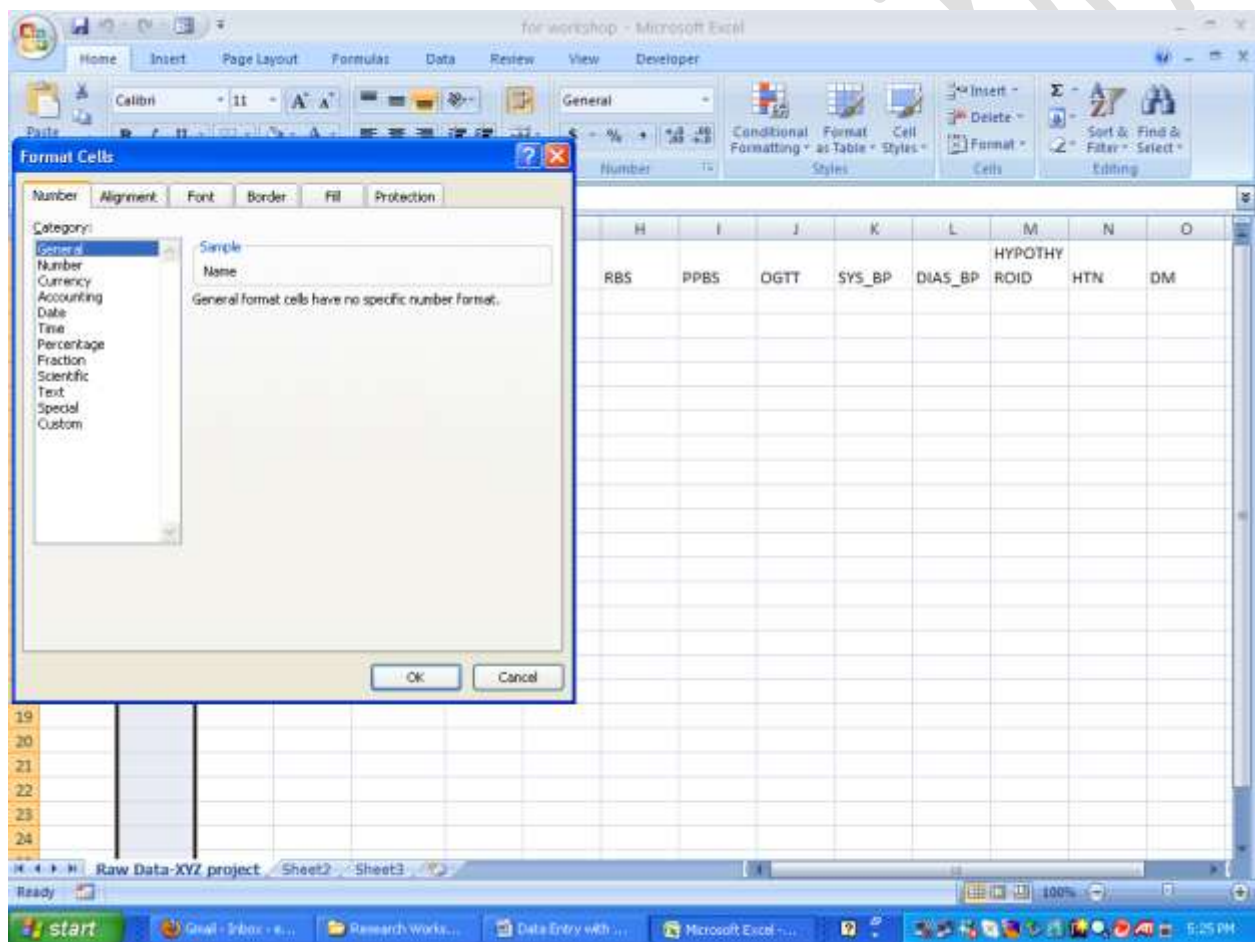
Please note: Do not enter units of measurement in the cell. For example, enter age as 37 instead of 37 years. We enter the unit of measurement in the variable sheet that we have prepared already. If we enter as 37 years in the cell, we will not be able to calculate the mean etc until we remove the years. This can be very painful in large datasets. That is why it is important to define your variable sheet before you start the entry.

ALSO NOTE: WHERE THE DIAGNOSIS CAN BE MADE ON THE BASIS OF EITHER/AND CUTOFFS (EXAMPLE- Hypertension can be diagnosed if the systolic BP is greater than and/or if the diastolic BP is greater than) it is always better to enter the two separately. Having a single column as BP and entering 120/80, 150/90 leads to a lot of cleaning to be done before we determine the mean systolic BP, the mean diastolic BP etc.

### STEP 3

Let us look at another variable structure. Let us look at Name. This will be a text entry. This column is used only as an identifier in the data collection and entry of raw data, and will be deleted during analysis to protect patient privacy.

Select Column B. Right click your mouse. Go to format cells, under the tab , you can choose either general or text, select OK



#### STEP 4

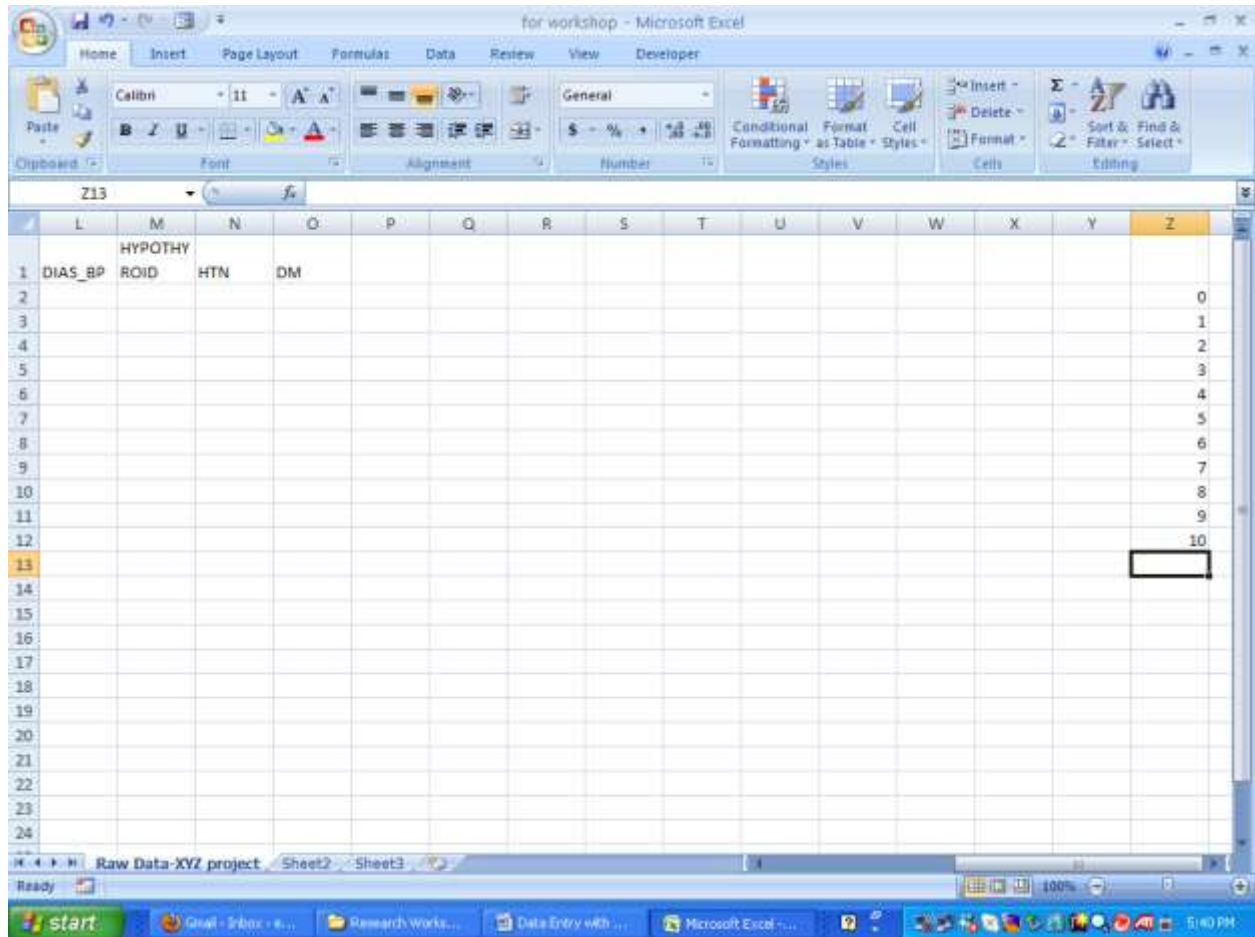
Let us look at another data structure. Let us look at the variable Sex. Data can be entered as Male, Female, or M and F.

However, we recommend that data for all ordinal data is entered as numeric values-For instance 0=Female, 1=Male. We note these labels or codes in our variable sheet (after some time , it is difficult to recollect what the labels indicate). The same applies for education. Instead of entering as primary school, secondary school, etc we enter as 0,1,2, 3 etc and note the codes. Similarly, Hypothyroid becomes 0=No, 1=Yes.

In the example data sheet, this will then include the variables Sex, Occupation, Education, Hypothyroid, HTN, DM.

We could set these values as a drop down menu for ease.

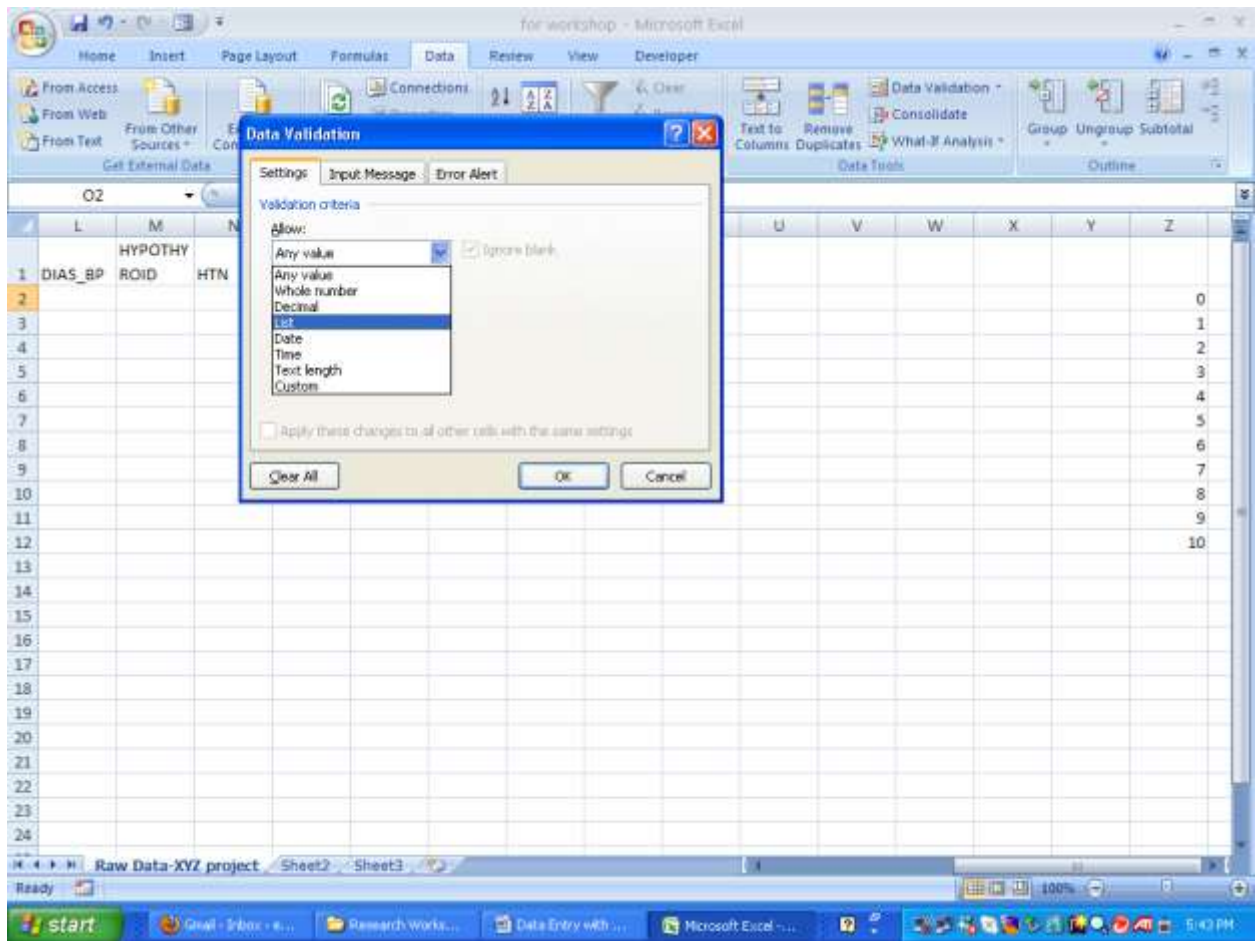
First choose a cell to enter the list of possible values. I will choose this in a cell much after my last variable, so there is enough space to add more variables if necessary.

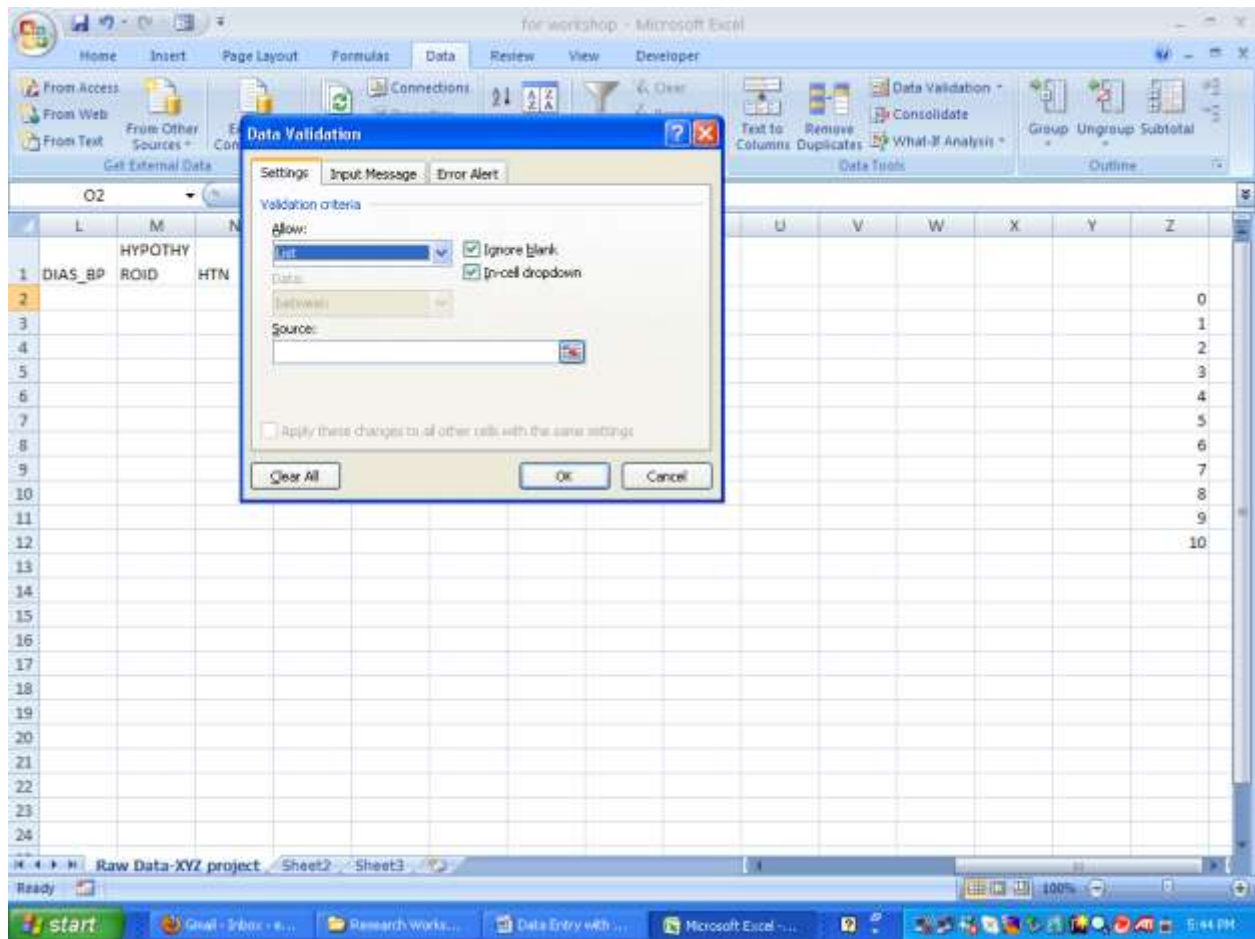


You can see I have left space from P-Y to add more variables if I feel it necessary to do so (this, for example can be categorization variables of data already collected under the variables A to O).

I will use the variable DM as an example.

- Click on cell O2
- Once that cell is selected, click on Data in the tabs on top
- Then click on data validation and choose data validation
- In the drop down menu of allow, select List





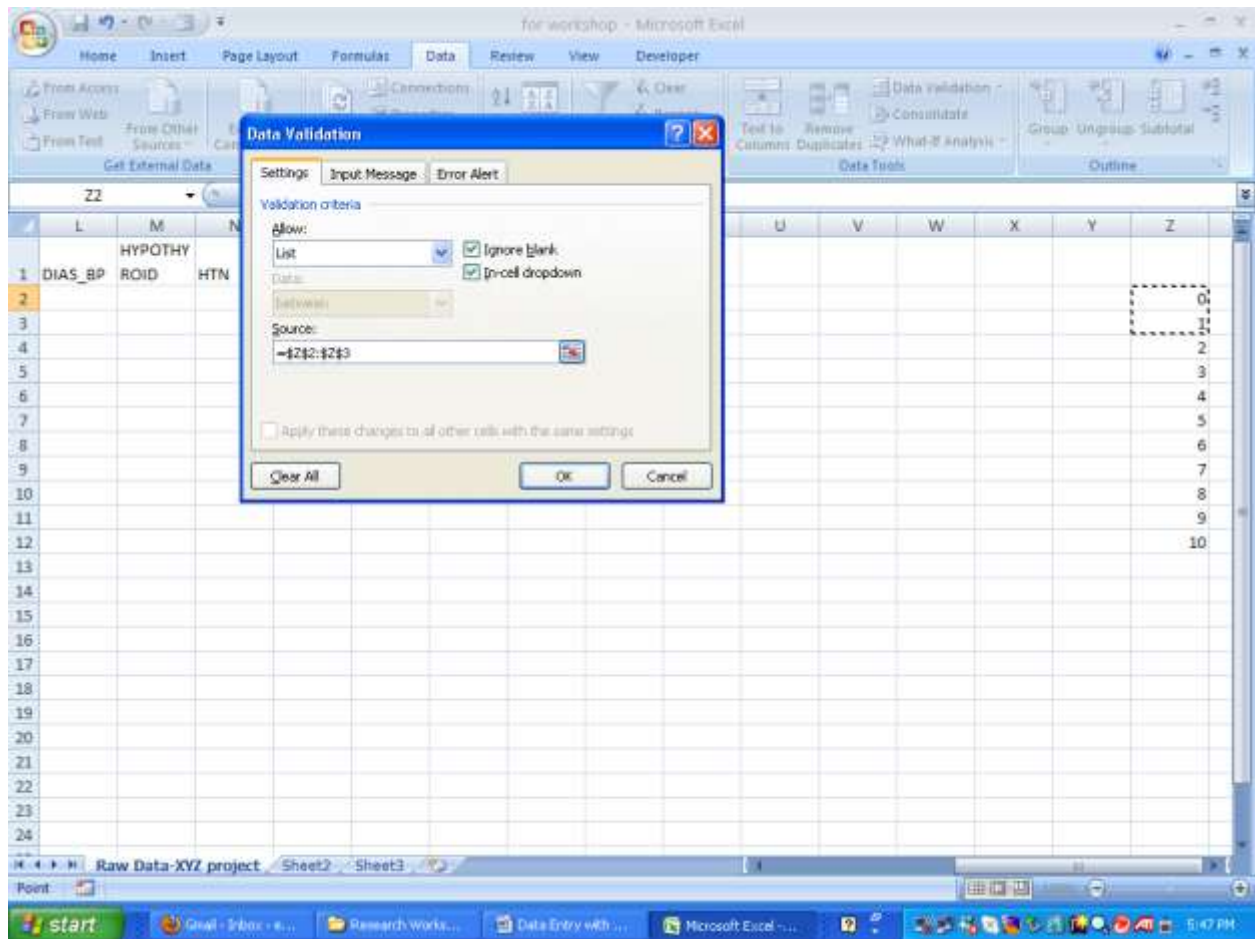
You can check or uncheck the “Ignore Blank” based on your need.

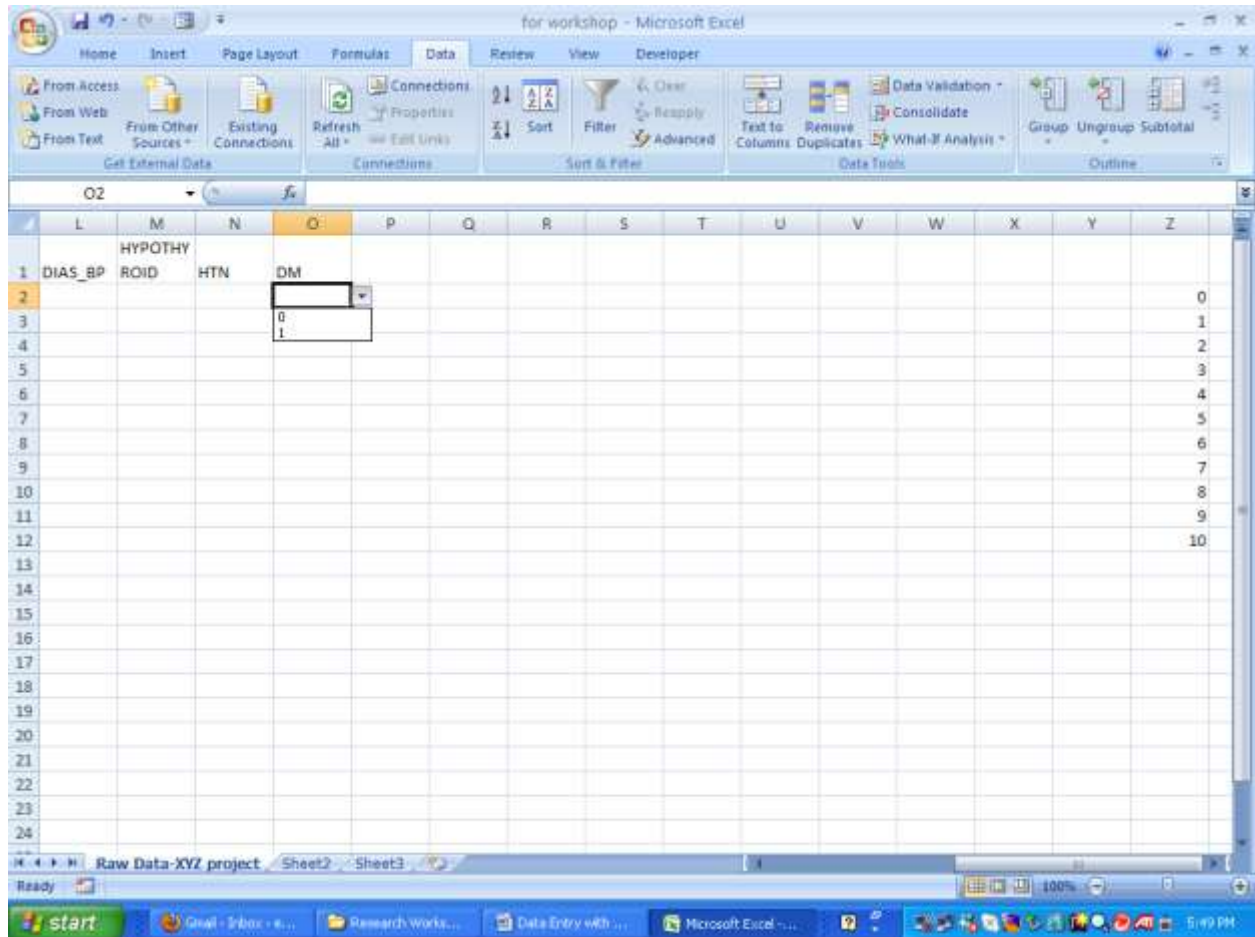
Have the cursor placed in source

The select the data from column Z (in this example) which forms your list; in this instance 0 and 1 will be chosen to represent No and Yes

Click Ok

(Check out the input message and error alert too)





Do the same for other similar variables (sex, education, occupation, HTN, Hypothyroid)

Note, sex, HTN, Hypothyroid can be marked as 0 and 1

Click cell O2

Use control+c to copy

Use control+v to paste under HTN, Hypothyroid and sex

Once you have done the drop down list for all variables, you can choose to hide the list if you wish to.

You can use the drag function of excel to drag the selection for all the 250 cases.

Select the first row.

Click on the right hand side of cell O and drag down.

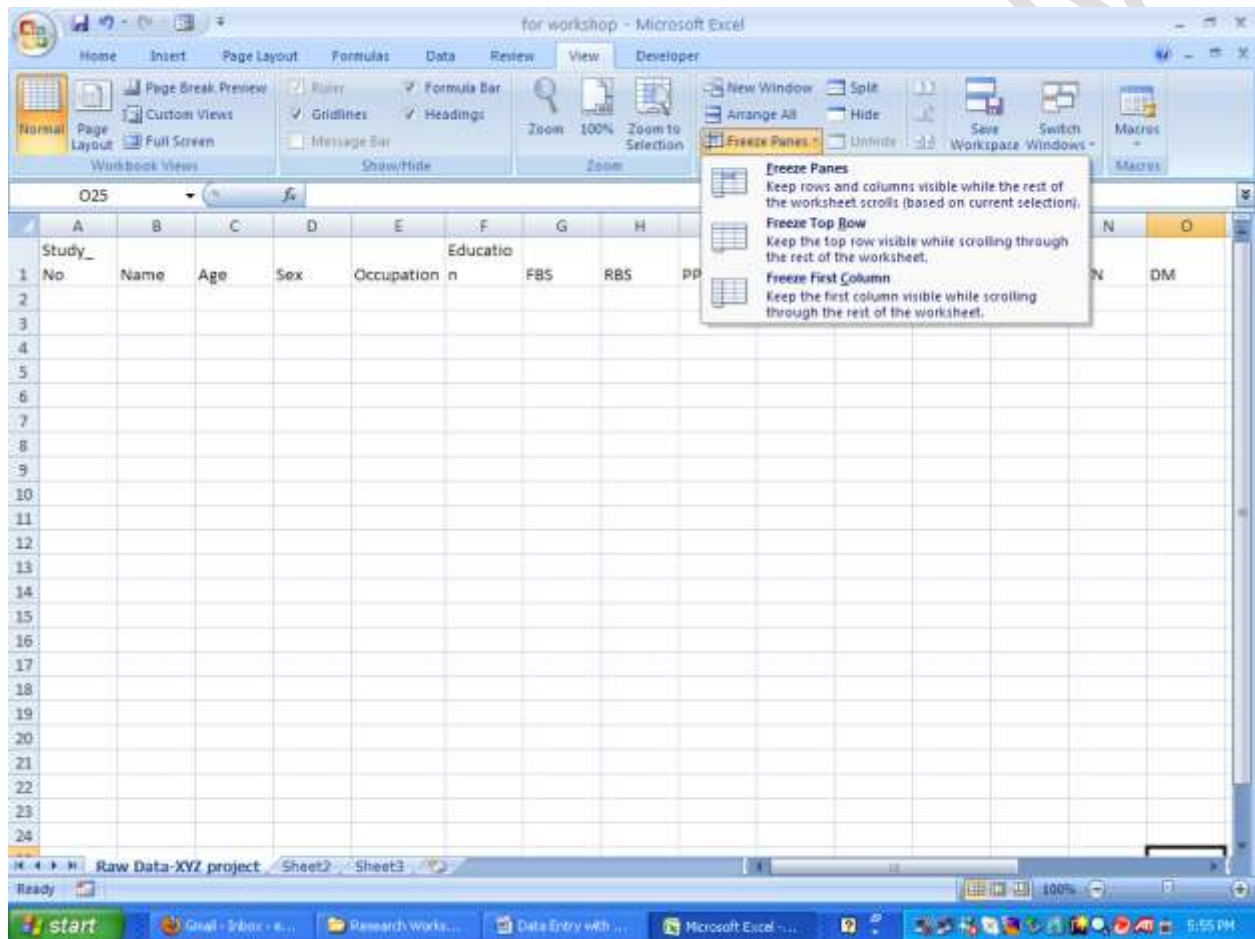
## Step 5

When we reach entry study\_no 24, or row 25, the header with the variables is no longer seen

You can choose to freeze the top row so that the headers are always seen

You can choose to freeze the top row and the first column too if you want to see the study\_number too

Go to View, Freeze Panes.

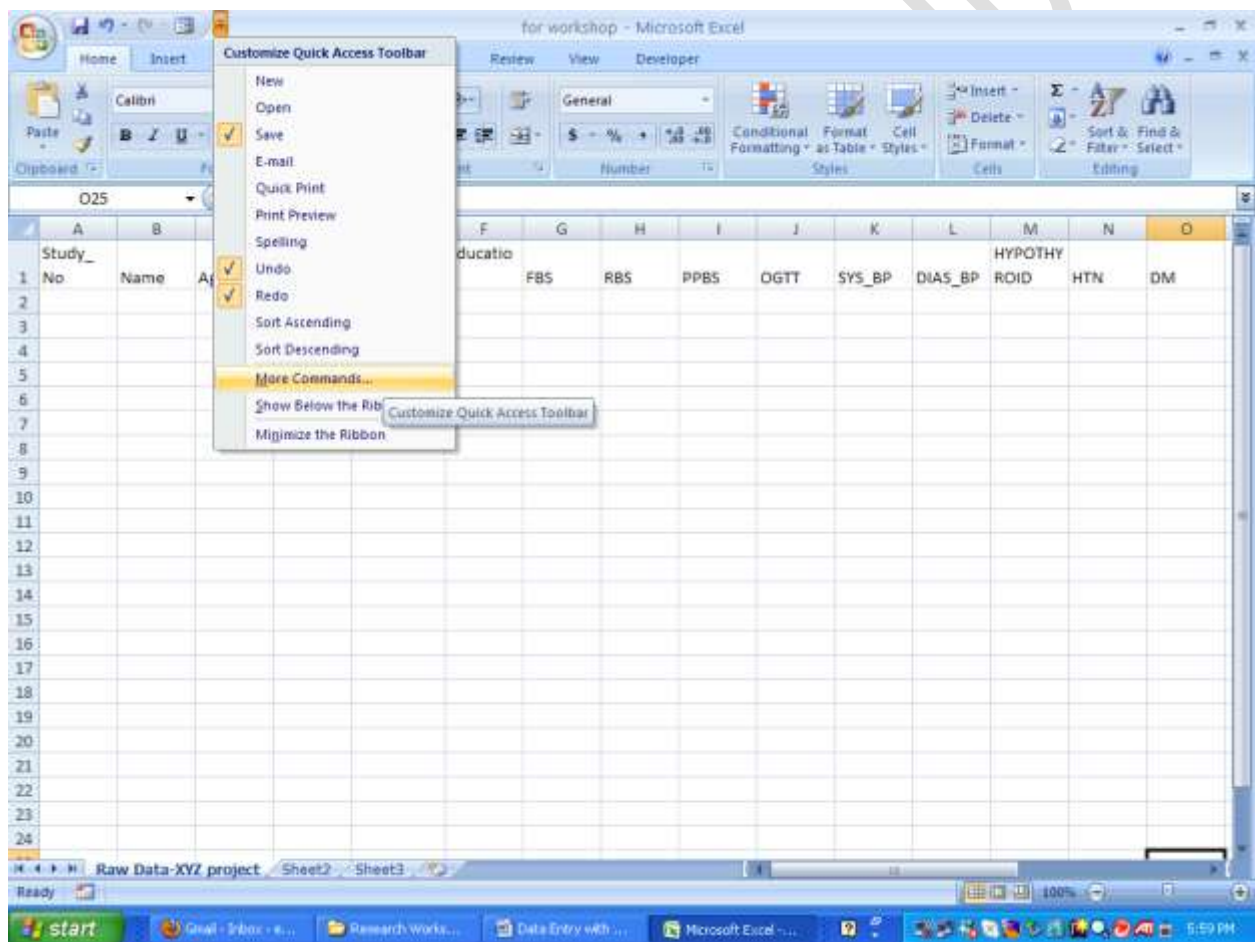


## Step 6

If you would like a form entry as opposed to a spreadsheet entry, there is something you can do instead of freezing the panes. You can create a form in excel, and when you enter the form, data will be entered into the spreadsheet.

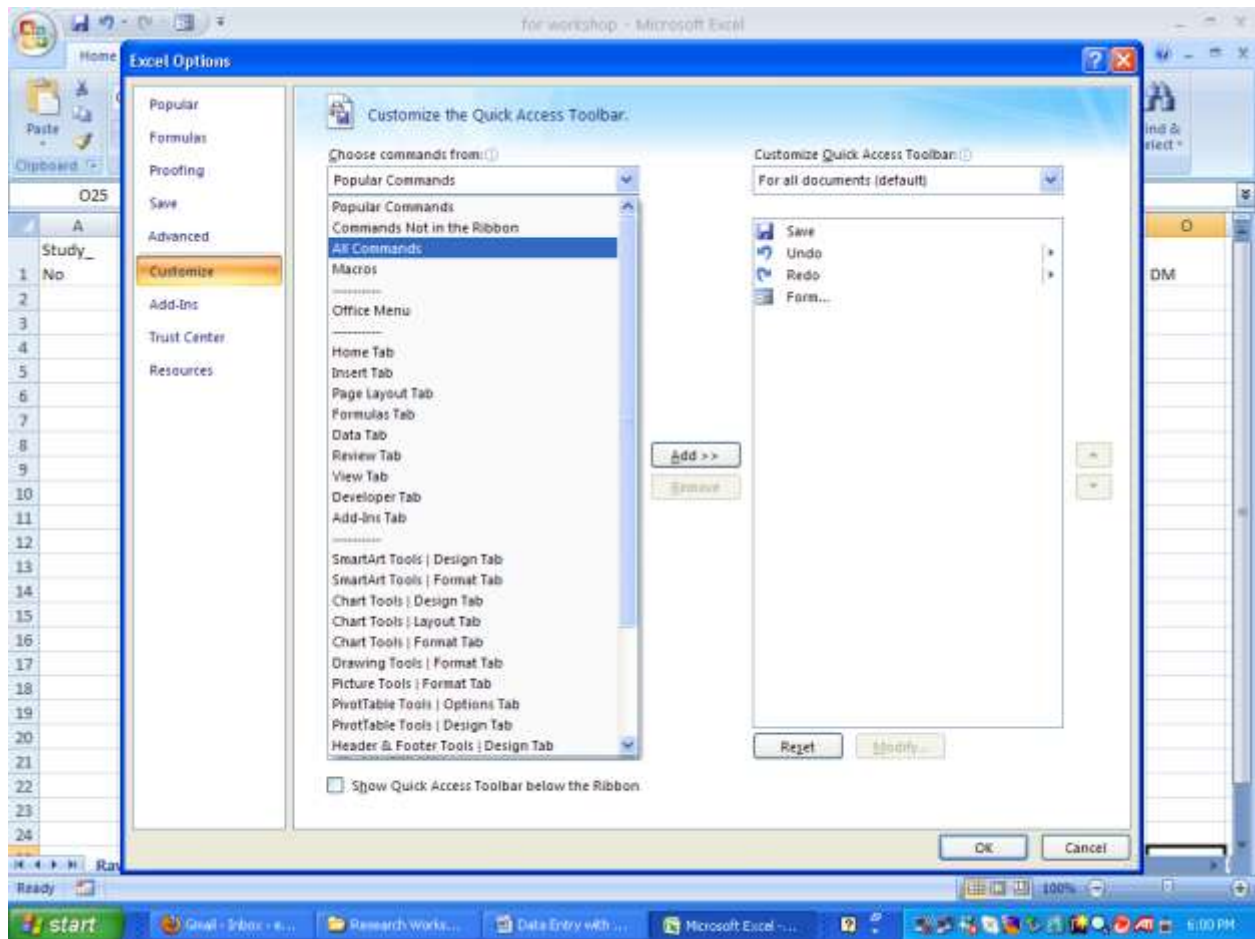
You will see an arrow at the top next to the save redo, undo buttons.

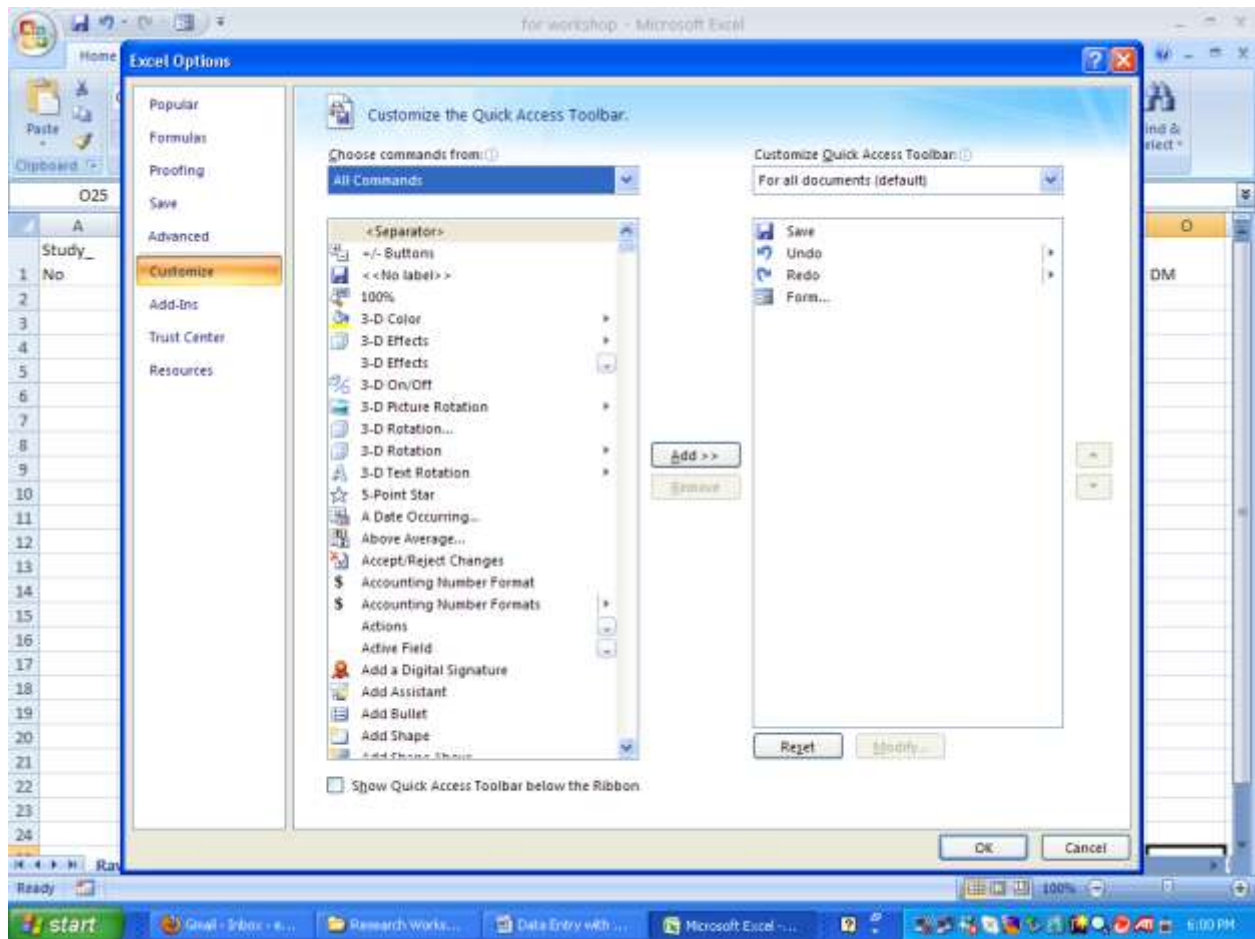
Click on that



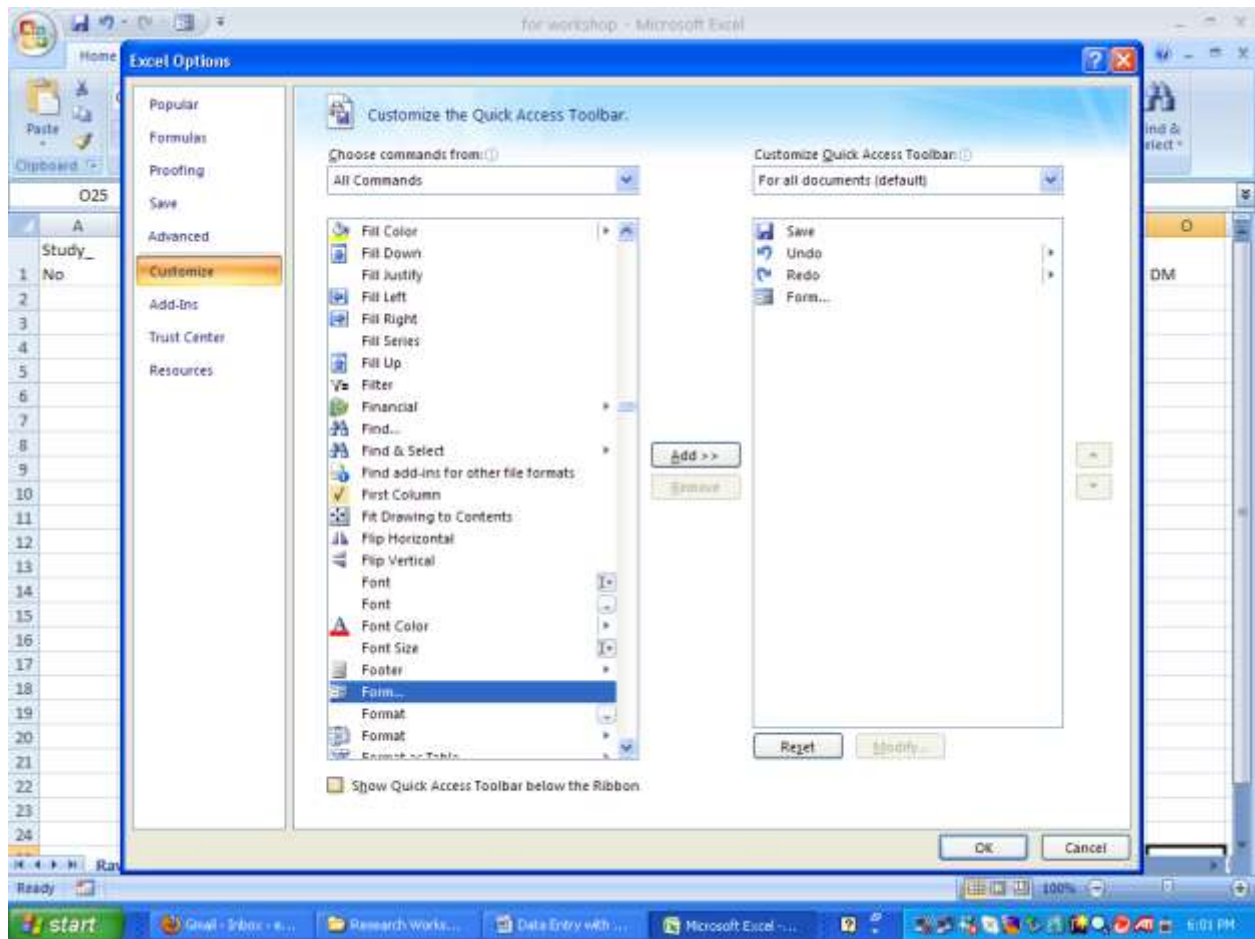
Choose more commands

Under choose commands from, choose all commands





Keep scrolling down till you see Form

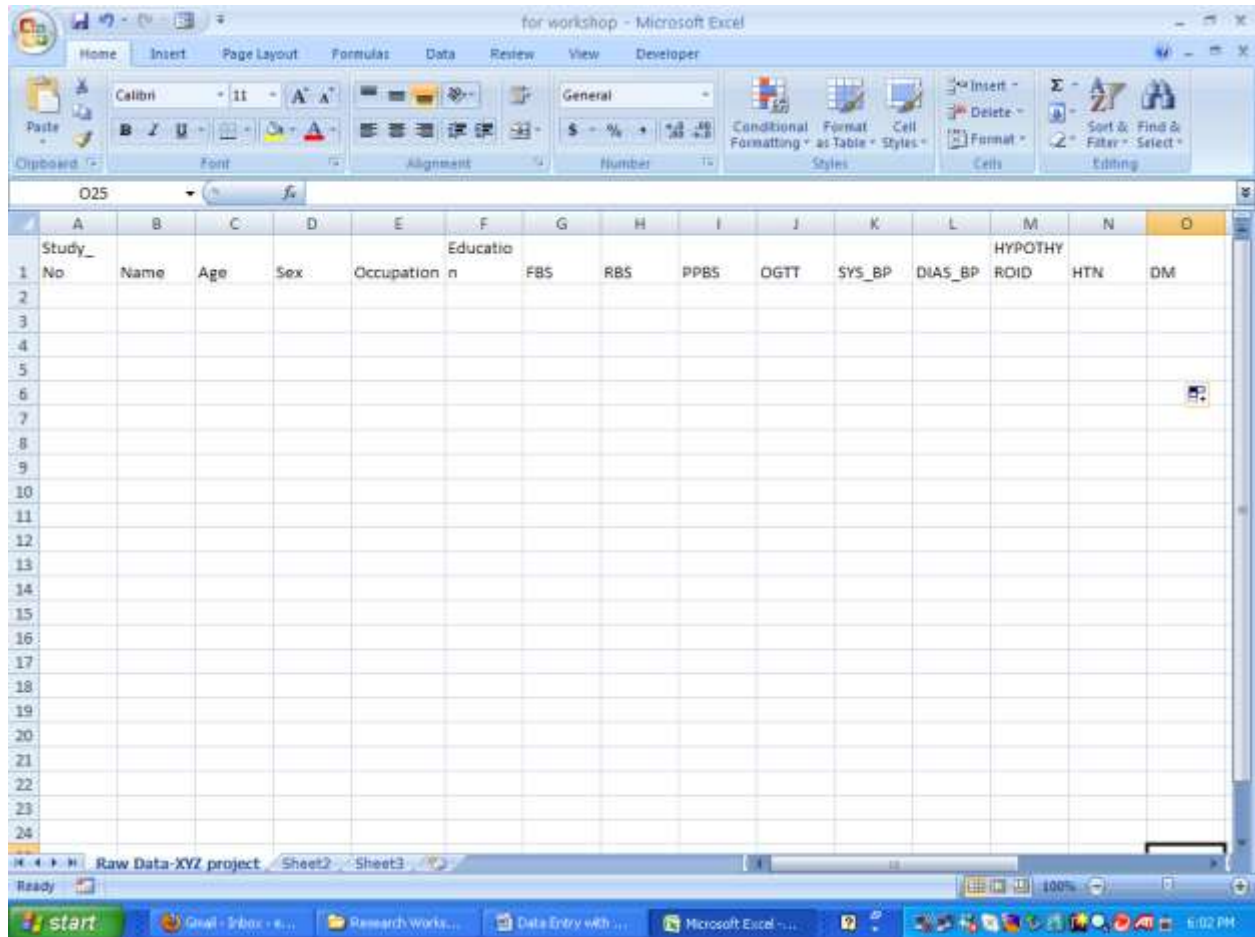


Select it

Click Add

It will move into the box on the right

Click Ok

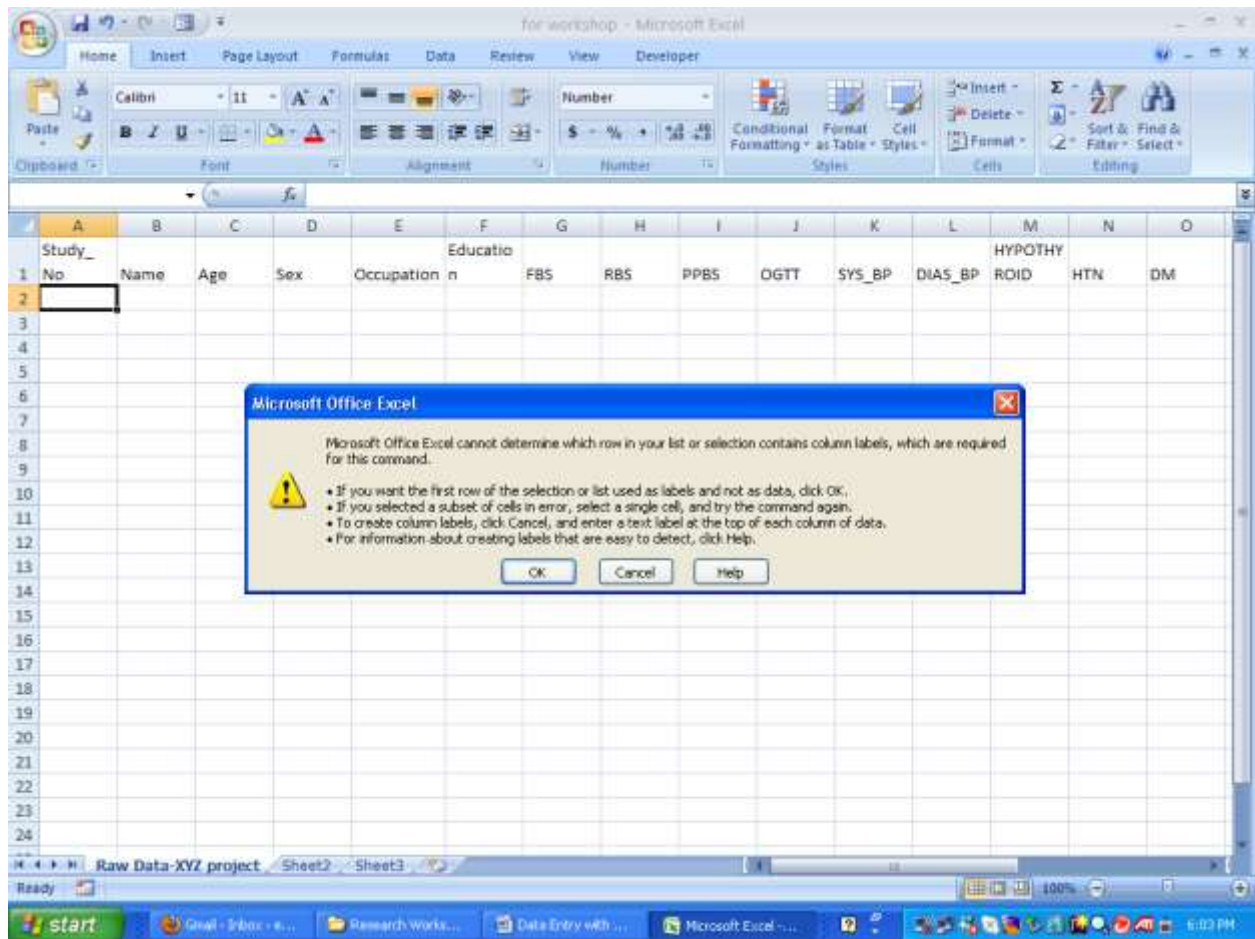


You will now see a form button on the top

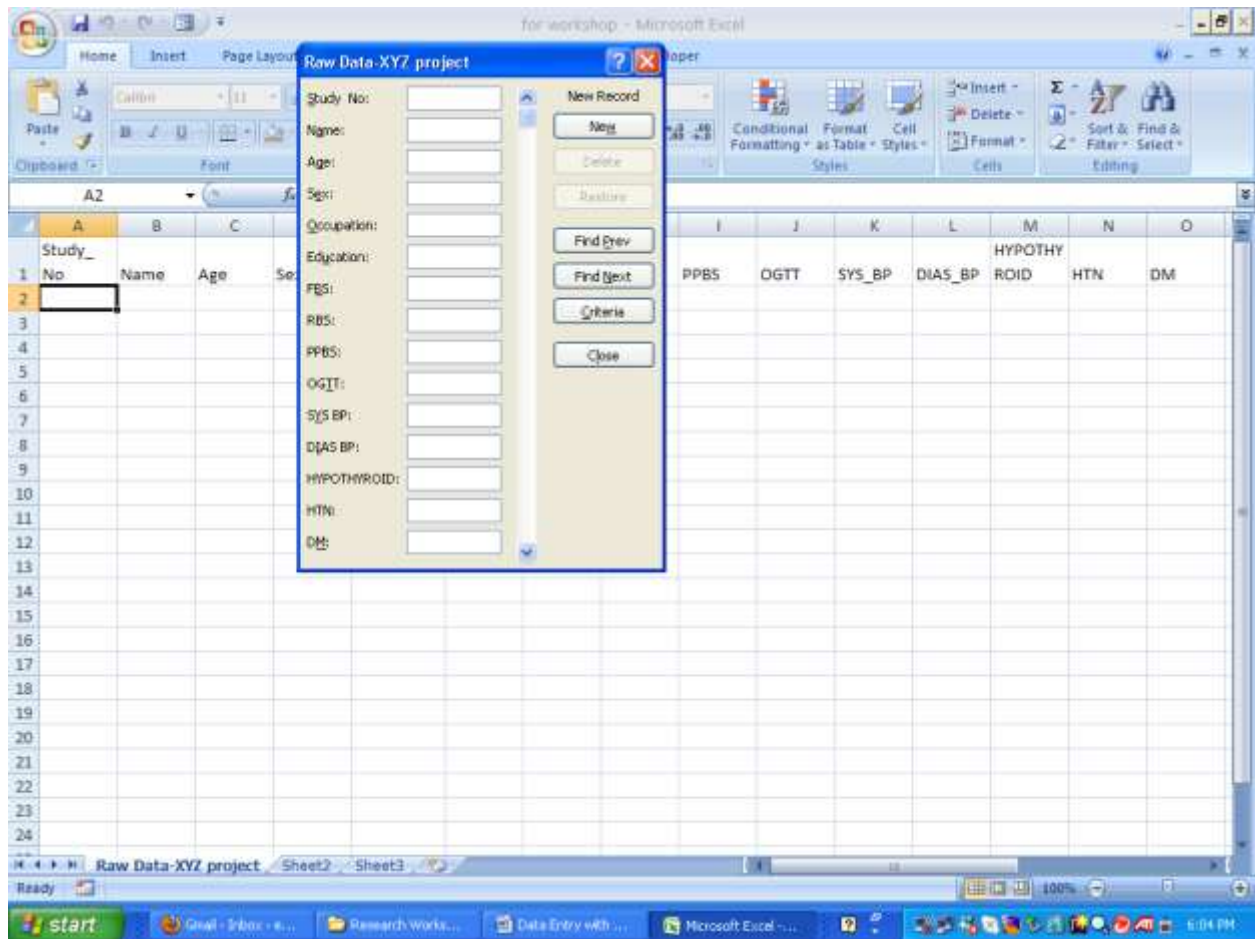
Use the form once you have created and prepared the sheet for data entry

Select cell A2

Click the Form button or icon



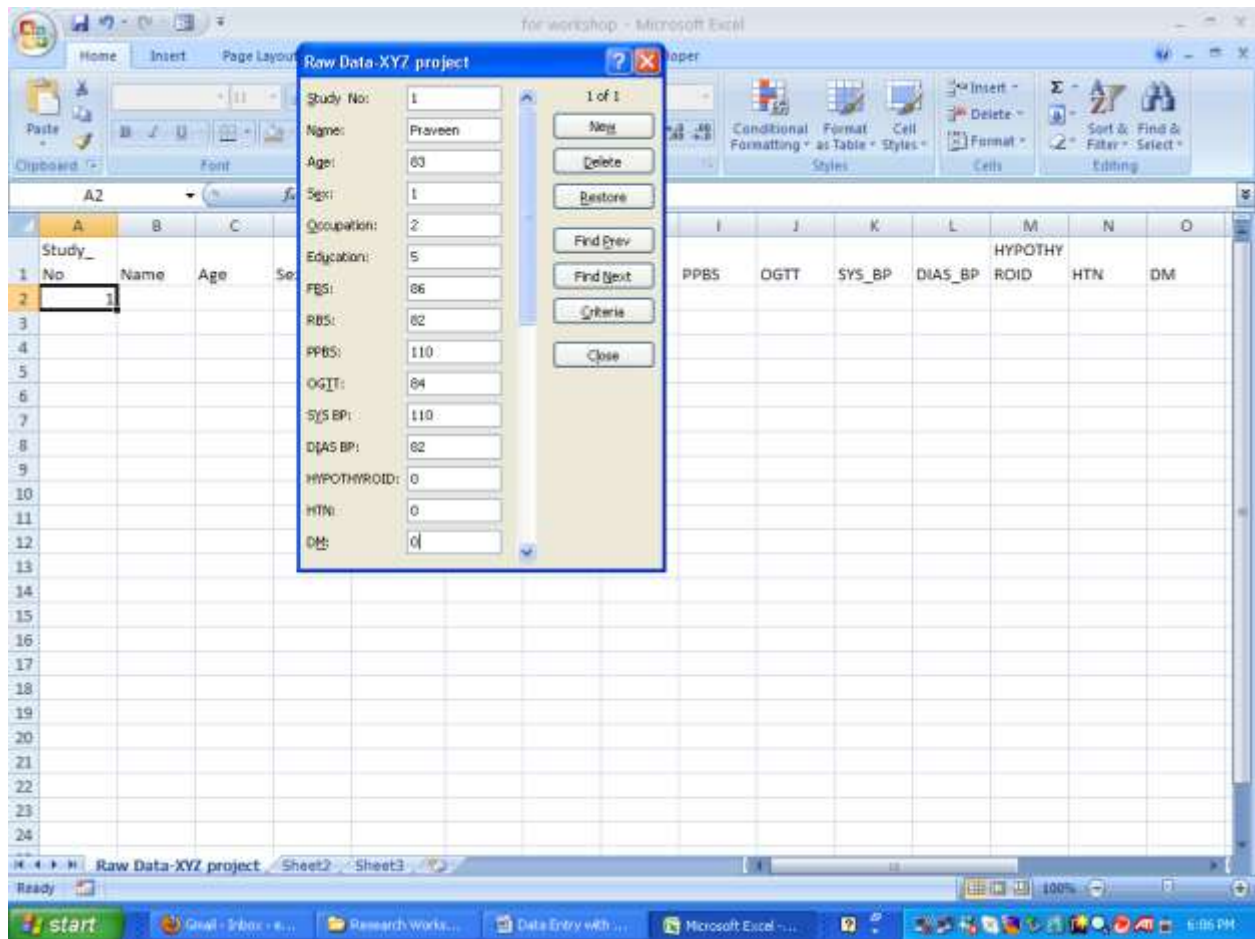
Click Ok



You have a form.....

You can find create a new record, find the previous record, find the next record, search using multiple criteria to find the record you want.....

I have entered data



If I click new , a new record is created

If I click close, the data is there in the spreadsheet

Study_	No	Name	Age	Sex	Occupation	Education	FBS	RBS	PPBS	OGTT	SYS_BP	DIAS_BP	R/OID	HTN	DM
1	1	Praveen	83	1	2	5	86	82	110	84	110	82	0	0	0
3															
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															
16															
17															
18															
19															
20															
21															
22															
23															
24															

THIS SHOULD HELP YOU GET STARTED!

PLEASE FEEL FREE TO ADAPT OR SHARE THIS DOCUMENT-

PLEASE DO ACKNOWLEDGE THE SOURCE AS FERNANDEZ HOSPITAL, HYDERABAD IF YOU SHARE THE DOCUMENT.