

Why Excel Is A Bad Format For Storing Data

Andy Grogan-Kaylor

4 Apr 2023 12:43:30

Background

Excel is sometimes used as a program to collect and store data. However, Excel may be problematic as a data storage solution for a number of reasons detailed below. Notably, statistical programs like Stata, SAS, or SPSS all store additional information with each variable such as: a *variable label*, describing the contents of the variable, or the survey question that resulted in the variable; and a *value label*, which attaches qualitative information to each possible value of the response.

Excel does not generally contain this extra information about each variable, or column of data, which may lead to errors in working with quantitative information.

The data below are stored in Stata format, but could as easily be stored in SAS or SPSS format.

Get The Data

```
. use "simulated-happiness-data.dta", clear
```

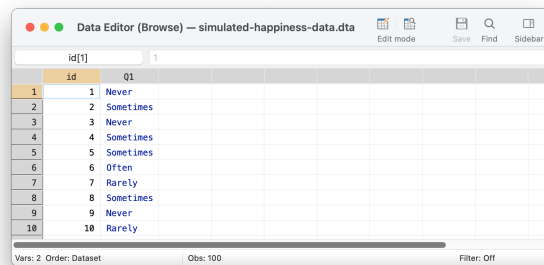


Figure 1: Screenshot of Stata

Describe The Data

```
. describe
```

Contains data from simulated-happiness-data.dta

Observations: 100
Variables: 2 4 Apr 2023 12:43

| Variable name | Storage type | Display format | Value label | Variable label |
|---------------|--------------|----------------|-------------|------------------------------|
| id | float | %9.0g | | id |
| Q1 | float | %9.0g | Likert1 | How often do you feel happy? |

Sorted by:

Descriptive Statistics and Bar Graph

Notice how the descriptive statistics and graph are informative in that they contain information on the *variable label* and *value label*. These help us to get an intuitive sense of the information in the data. We see this information when we list out the data as well.

Descriptive Statistics

```
. tabulate Q1
```

| How often do you feel happy? | Freq. | Percent | Cum. |
|------------------------------------|-------|---------|--------|
| Never | 21 | 21.00 | 21.00 |
| Rarely | 29 | 29.00 | 50.00 |
| Sometimes | 28 | 28.00 | 78.00 |
| Often | 22 | 22.00 | 100.00 |
| Total | 100 | 100.00 | |

Bar Graph

```
. graph bar, over(Q1) scheme(michigan2) asyvars  
.  
. graph export mybar1.png, width(1500) replace  
file /Users/agrogan/Desktop/GitHub/agrogan1.github.io/myposts/mybar1.png saved as PNG  
format
```

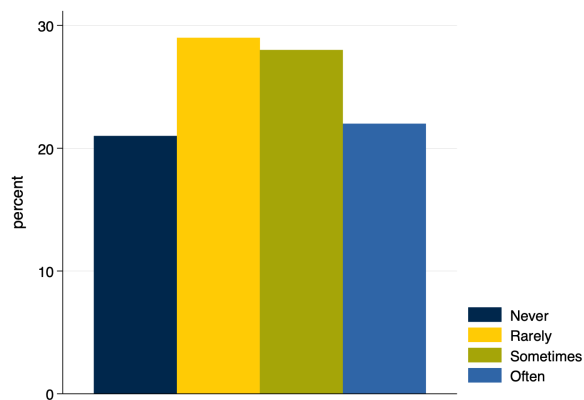


Figure 2: Bar Graph

List Out A Sample Of The Data

```
. list in 1/10
```

| | id | Q1 |
|----|----|-----------|
| 1. | 1 | Never |
| 2. | 2 | Sometimes |
| 3. | 3 | Never |
| 4. | 4 | Sometimes |
| 5. | 5 | Sometimes |
| 6. | 6 | Often |
| 7. | 7 | Rarely |
| 8. | 8 | Sometimes |
| 9. | 9 | Never |

Now Use The Data In Excel Format

We've saved this simulated data in Excel format.

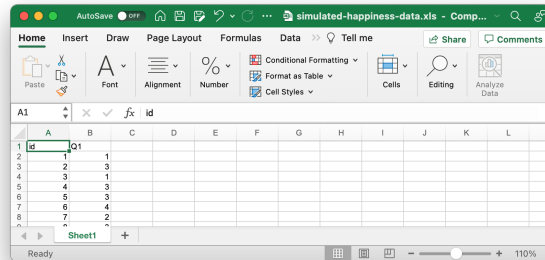


Figure 3: Screenshot of Excel

We now import the Excel data file. We use the first row of data as variable names.

```
. import excel "simulated-happiness-data.xls", sheet("Sheet1") firstrow clear
(2 vars, 100 obs)
```

We see right away—when we list some of the data—that the data are less informative.

```
. list in 1/10
```

| | id | Q1 |
|-----|----|----|
| 1. | 1 | 1 |
| 2. | 2 | 3 |
| 3. | 3 | 1 |
| 4. | 4 | 3 |
| 5. | 5 | 3 |
| 6. | 6 | 4 |
| 7. | 7 | 2 |
| 8. | 8 | 3 |
| 9. | 9 | 1 |
| 10. | 10 | 2 |

Adding this valuable information back into the data set may take a great deal of extra effort.

Descriptive Statistics and Bar Graph

Notice how the descriptive statistics and graph are much less informative. For example, it is now not immediately clear what **Q1** represents.

It is also not clear whether higher values of **Q1** represent higher levels of *happiness*, or higher levels of *unhappiness*, a crucially important substantive distinction. The information on variable label and value label will have to be added back into the data when preparing a report for dissemination.

Descriptive Statistics

```
. tabulate Q1
      Q1 |      Freq.      Percent      Cum.
```

| | | | |
|-------|-----|--------|--------|
| 1 | 21 | 21.00 | 21.00 |
| 2 | 29 | 29.00 | 50.00 |
| 3 | 28 | 28.00 | 78.00 |
| 4 | 22 | 22.00 | 100.00 |
| Total | 100 | 100.00 | |

Bar Graph

```
. graph bar, over(Q1) scheme(michigan2) asyvars
.
. graph export mybar2.png, width(1500) replace
file /Users/agrogan/Desktop/GitHub/agrogan1.github.io/myposts/mybar2.png saved as PNG
format
```

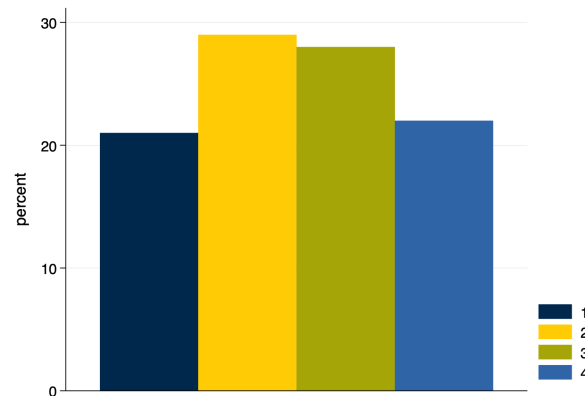


Figure 4: Bar Graph

A Few Final Issues

Notice how Excel doesn't enforce the idea of whether variables are *numeric*, or *string*, and so would allow storage of different types of information in the same column.

Secondly, Excel would allow some of your columns to have the same name, which might make data difficult to work with in other software.

| x | y | verylongvariablename | verylongvariablename |
|----------------|---|----------------------|----------------------|
| 100 | 1 | Smith | 20 |
| 200 | 2 | 30 | NA |
| not applicable | x | yes | 60 |