**Georgetown CSSA Data Workshop #1**

Date: 2023/2/10

Shiyao Wang

Software:

1. \*GitHub Desktop
2. \*Microsoft Excel (WPS may as a substitute, but Excel is preferred)
3. RStudio: <https://posit.co/download/rstudio-desktop/>

You need to follow a two-step process installing R language first and then RStudio IDE

1. Anaconda (Python for data analysis): <https://www.anaconda.com/>

This will give you: python + two IDES: Jupyter Notebook and Spyder

1. Stata: can be downloaded from Georgetown Webstore

*\* means required for this session*

Topic:

1. Any application is just a tool, the idea is the key;
   1. import data: <https://grade-data.oss-cn-hangzhou.aliyuncs.com/galaviz_rawdata/auto.csv>
   2. browse data
   3. summarize the mean for a column
   4. draw a histogram

Some questions to think about:

1. What is a CSV file, what’s the difference between csv and .xlsx file?
2. Why in R programming language, we need to quote the data frame’s name before we quote a column (i.e. a variable in Stata)?
3. How important it is to learn the terminologies - any idea of what is an application, what is a program?
4. What is a “function”, what is a syntax/sentence? Can these definitions be universally used?
5. Excel group work - World Bank’s mock exam practice;

Some questions to think about:

1. What are the biggest difficulties in conducting/learning data analysis?
2. Why firms are still using Excels, and for what purposes?
3. (Just to give a sense) Excel dynamic chart;
4. (Just to give a sense) Excel and VBA.