Assignment 1: Data Exploration (100 points)

INFO3237: Business Analytics II

Instructor: Dr. Xue Guo

**Purpose:** To load, explore, and manipulate a data set

**Note:** These instructions are provided for those who use RStudio. However, the students can use any other software to complete the assignment.

**What Should You Do?** You need to follow these steps:

1. Login to [canvas.uncc.edu](https://citrix.wpcarey.asu.edu) and download the assignment file(s).
2. Download and save the data set (census\_2018.csv) from Assignments/Assignment 1
3. Launch RStudio
4. Load and read the data using RStudio: Click on File -> Import Dataset -> From Text (base) or From CSV. You can also use function “read.csv()” to load and read the data set.
5. Please look at the data dictionary at the end of this document to learn about the variables.
6. Answer the following questions:
   1. **(4 points)** How many attributes this data set has?

There are **25** attributes in the dataset

* 1. **(4 points)** How many records this data set has?

There are **164610** records in this dataset

* 1. **(12 points)** What is the range of values for each of these attributes:

*educ\_num:* **1 to 11**

*citypop:* **965 to 83987**

*incwage:* **0 to not reported or 0 to 718000 if not counting NA’s**

* 1. **(12 points)** The missing values in this data set are marked with “not reported”. Which attributes have missing values?

**empstat, looking, available, incwage**

1. Use functions “typeof()”, “mean()”, “sd()”, and “unique()” to answer the following questions:
   1. **(16 points)** What is the type of each of these attributes:

*met2013:* **Character**

*educ*: **Character**

*classwkr*: **Character**

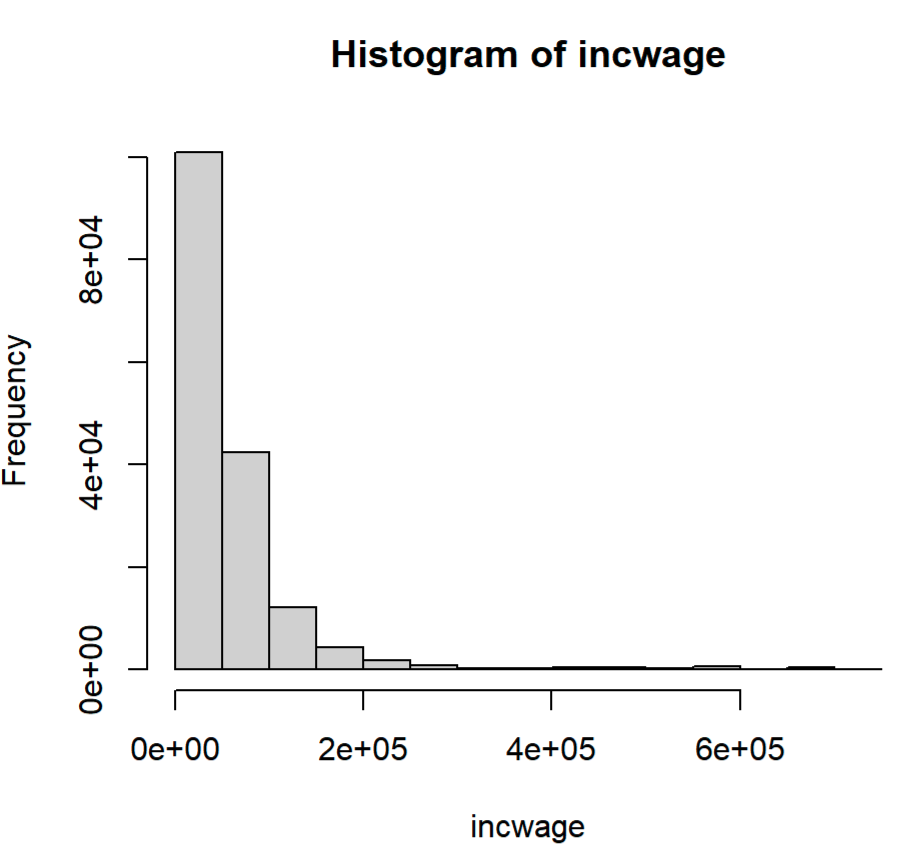
inctot: **Integer**

* 1. **(4 points)** What are the categories in variable *classwkrd*?
* **“Federal govt employee”**
* **“State govt employee”**
* **“wage/salary, private”**
* **“wage/salary at non-profit”**
* **“Self - employed, not incorporated”**
* **“Self-employed, incorporated”**
* **“Local govt employee”**
* **“unpaid family worker”**

* 1. **(8 points)** What is the mean and standard deviation for *inctot?*

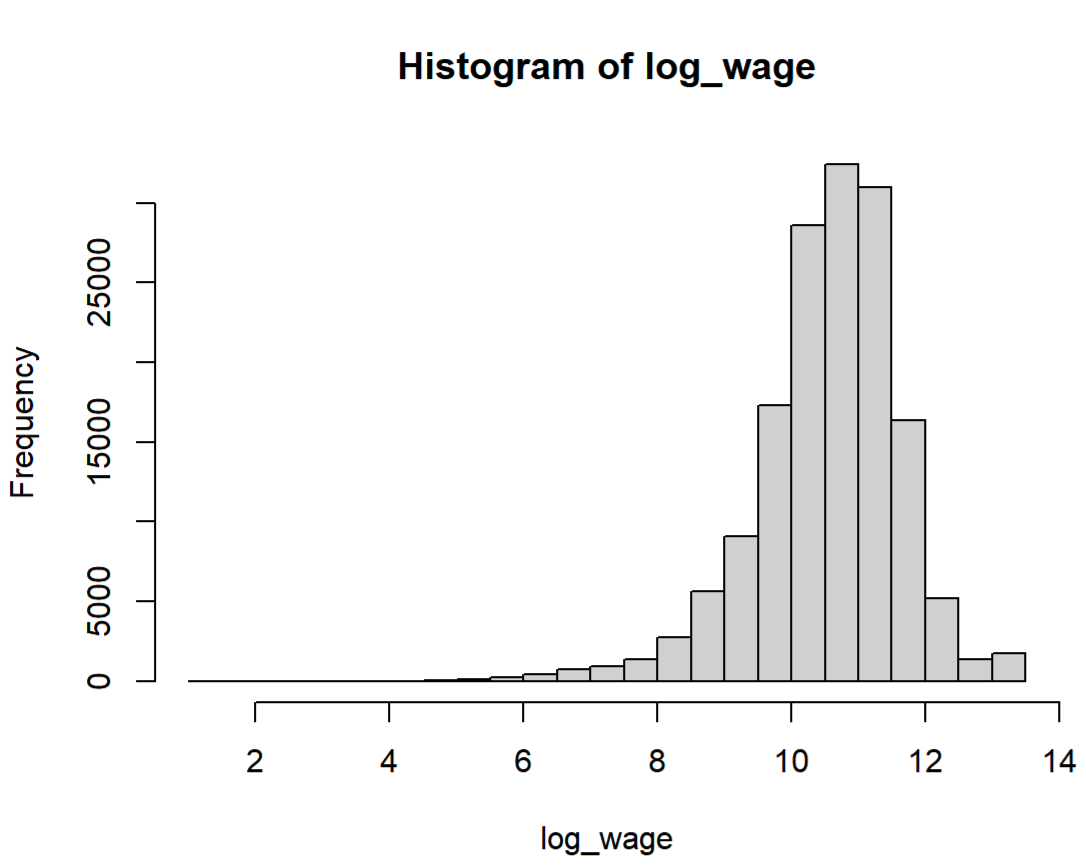
The mean is **65805.12** and the standard deviation is **83775.61**

1. Incwage will be one very important variable in your future analysis. Explore this variable and answer the following questions.
   1. **(10 points)** Create a Histogram graph for attribute *incwage.* Insert the graph in the space below and answer how this attribute is distributed in the data set (4 points).

l

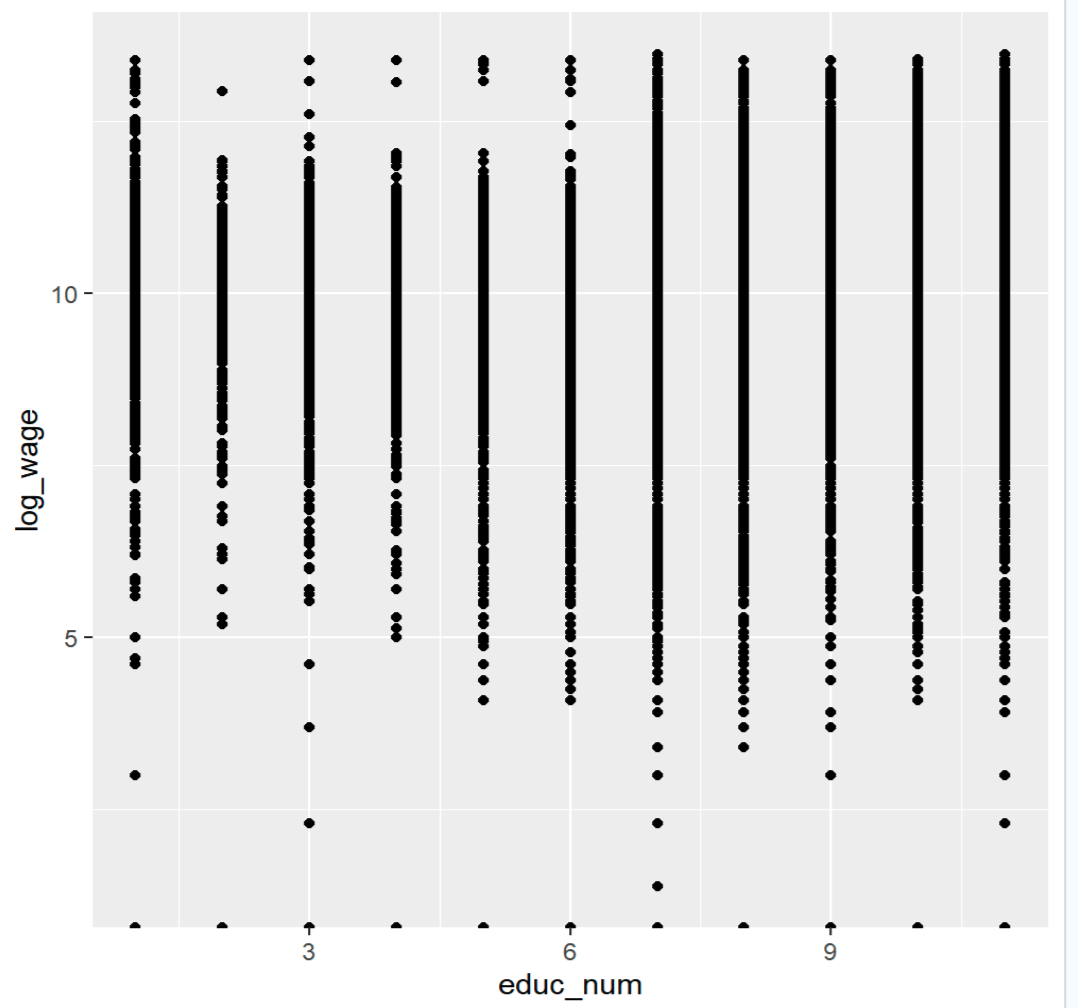
There are more records with a lower wage and salary income than there are with a higher wage and salary income. The dataset also seem to have only high outliers and no low value outliers.

* 1. **(10 points)** Generate a new attribute *log\_wage* = log(wage). Create a Histogram graph for attribute *log\_wage.* Insert the graph in the space below and explain what the differences between this histogram and the histogram are from 8.1.



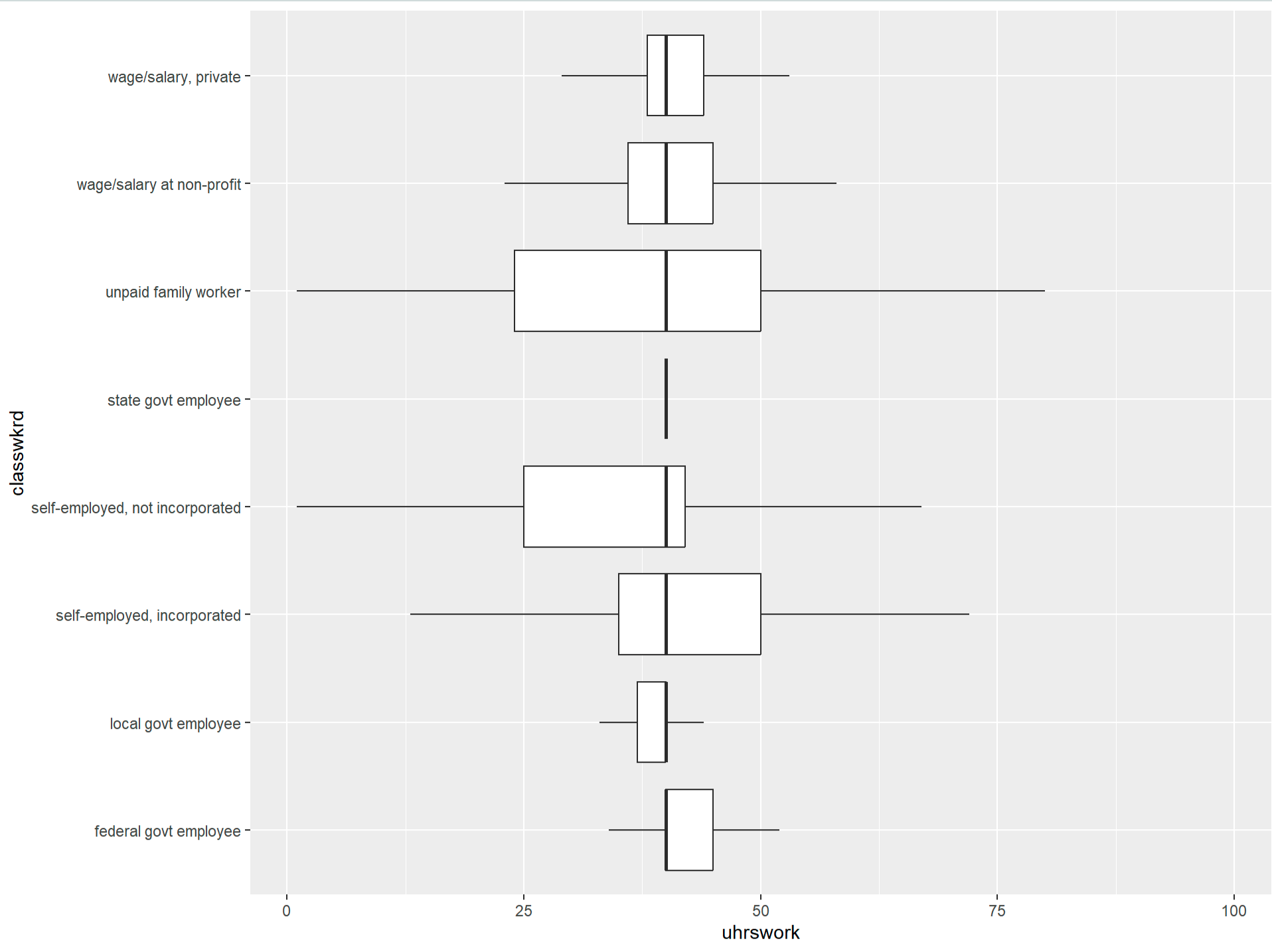
The log attribute allows us to see the distribution of the incwage attribute in a more detailed view. It also allows us to see the bell curve in a more centralized spot, so we can see that there are more lower income records that can be broken down further than in the 8.1 graph.

1. **(10 points)** You want to explore the relationship between *education\_num* and *log\_wage*. Choose one appropriate graph type and create a graph using ggplot2. By observing the graphs you created, please interpret the relationship between education and wage. Insert the graph in the space below:



These are two continuous variables, therefore the scatterplot would be the best graph to use to look at the relationship. It can be seen that how many years of education someone has does affect the log\_wage of a person. We can see that the larger clusters that are for higher salaries correlated with a higher education rate. You can also see that there are more outliers with the lower education records versus the higher education records.

1. **(10 points)** Create a graph shows the number of working hours (*uhrswork*) by detailed worker’s class (*classwkrd*). Choose one appropriate graph type and create a graph using ggplot2. Explain why this graph fit the data. Insert the graph in the space below.



This graph best fits this relationship as the attribute on the x axis is a discrete variable (Classwkrd) and the y axis variable is continuous (uhrswork). Of course, I used the flip command that switched the axes for the picture, so it is the opposite.

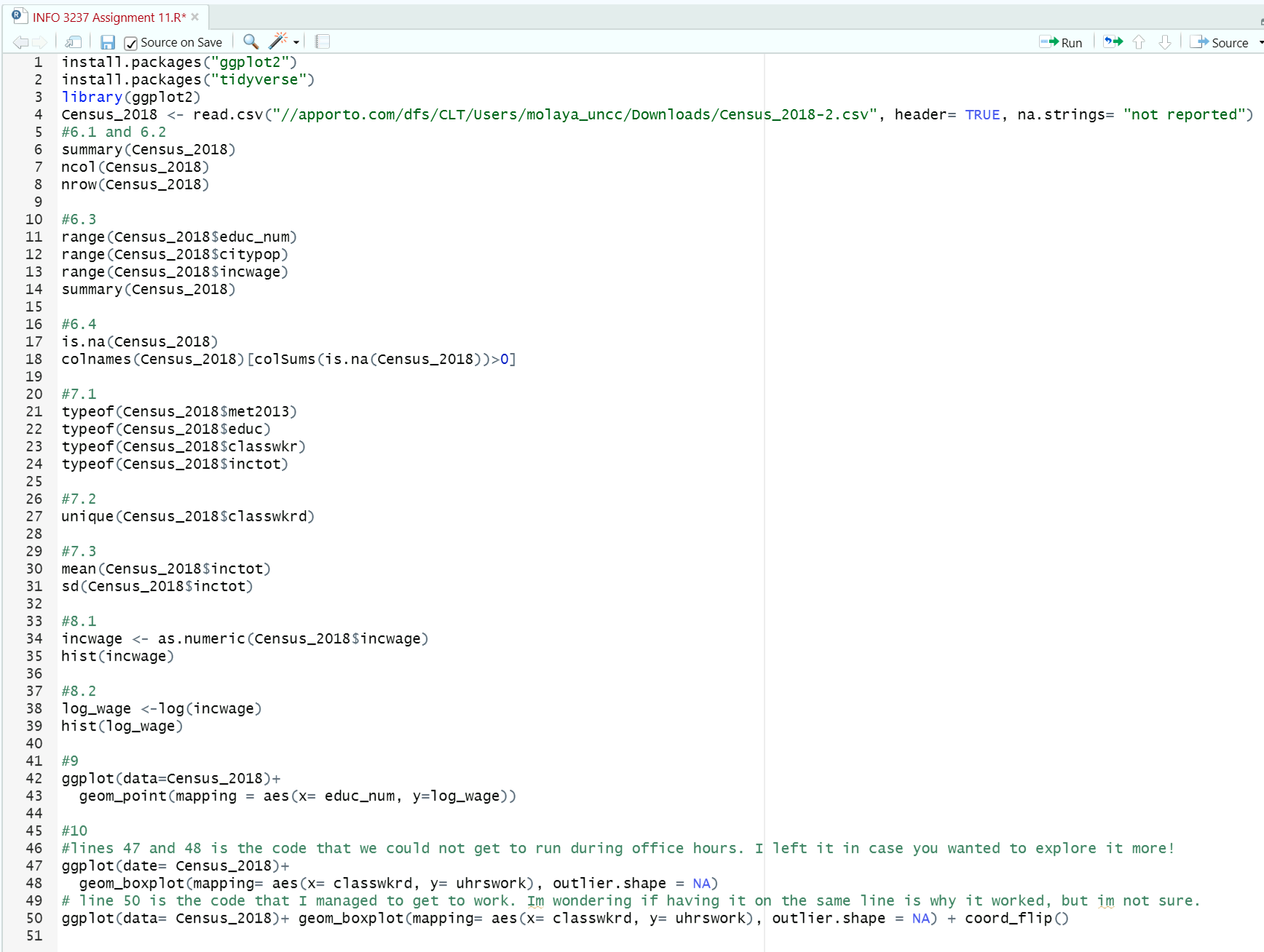


Table 1. Data Dictionary

| **Variables** | **Definition** |
| --- | --- |
| ID | Individual identifier (one person one row) |
| Year | Census year |
| State | State |
| County | County |
| Met2013 | Metropolitan areas |
| City | City |
| Citypop | City population |
| Pernum | Person number in sample unit |
| Perwt | Person weight |
| Bpl | Birthplace [general version] |
| Bpld | Birthplace [detail version] |
| Educ | Educational attainment [general version] |
| Educ\_num | Numeric representation of educational attainment |
| Educd | Educational attainment [detail version] |
| Empstat | Employment status |
| Classwkr | Class of worker [general version] |
| Classwkrd | Class of worker [detail version]] |
| Wksword2 | Weeks worked last year |
| Uhrswork | Usual hours worked per week |
| Looking | Looking for work |
| Available | Available for work |
| Workedyr | Worked last year |
| Inctot | Total personal income |
| Ftotinc | Total family income |
| Incwage | Wage and salary income |