Homework-1

**Out date:** June 11, 2021

**Due date:** June 20, 2021 at 11:59PM

Team#: \_\_\_

Team Member-1: Charles\_Colgan\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_

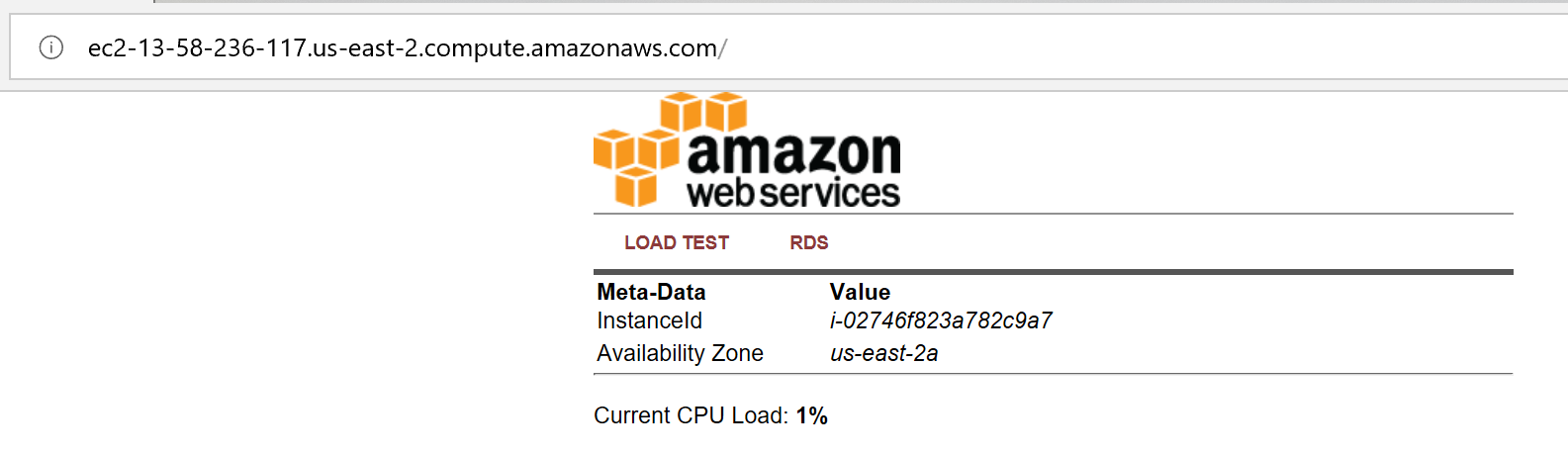
Team Member-2:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_

Team Member-3:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_

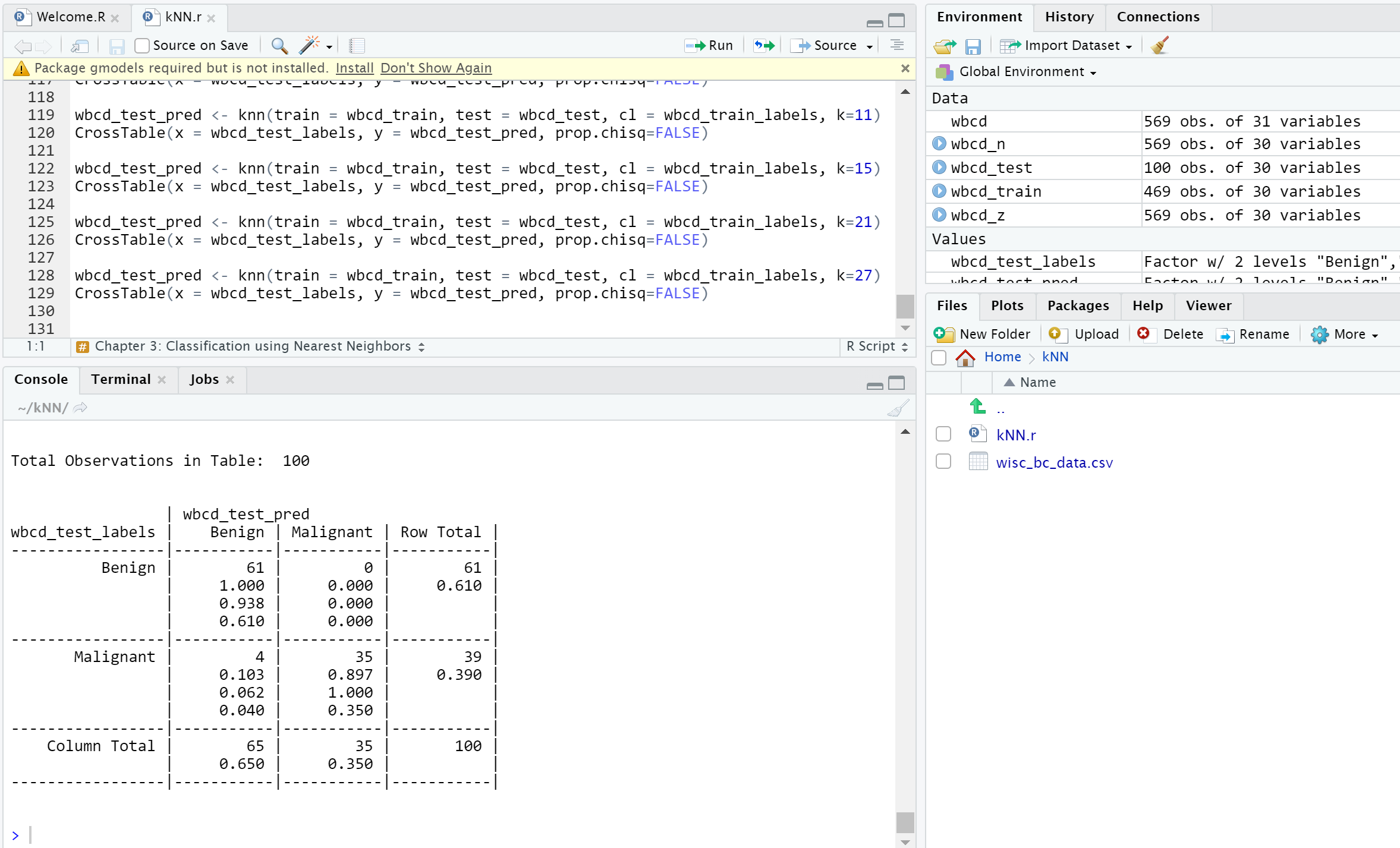
**Submission**

1. Work on the following tasks.
2. Rename this word file to “HW1-YourTeam#” (e.g., HW1-Team0.docx).
3. Upload the file to the blackboard system.

**Question-1:** Create a Linux instance on AWS to run web services. The steps to create the instance are available in the tutorial document (**1-Creating Linux Instance on AWS.pdf**). A video tutorial for this task is available on Blackboard. Once you are done, take a snap of the server instance and paste it here. A sample snap is shown below. **[20 points].**



**Question-2:** Create an AWS Ubuntu instance to run an RStudio server on it. The steps to create the instance are available in the tutorial document (**2-Creating Rstudio Server on AWS.pdf**). A video tutorial for this task is available on Blackboard. The R code and the data are available in the tutorial folder. Modify the code to store the CrossTable outputs in csv files for all k-values (k=1…27). Submit your csv files.Once you are done, take a snap of the server instance and paste it here. A sample snap is shown below. **[20 points].**



**Question-3:** Describe in detail at least four applications of big data. **[20 points]**

**Answer:**

**Question-4:** Describe in detail structured data, unstructured data, semi-structured data, and metadata. **[20 points]**

**Answer:** Structured data is data that follows a rigid system of organization which is never deviated from, thus preserving the relationships of the data and allowing for more ease of storage and use. Unstructured data is data that does not follow any system of organization, thus it is difficult to use and store effectively. Semi-structured data is data that follows a system of organization, but does not directly preserve relationships of the data. Metadata is data that provides information about the characteristics and function about some other set of data.

**Question-5:** Describe in detail the big data characteristics - the five V’s of big data. **[20 points]**

**Answer:** Big Data can be characterized by volume, the overall size of the raw data from a source, velocity, the speed at which the data arrives from that source, variety, the types and formats of the data, veracity, the quality of the data in terms of its signal to noise ratio, and value, the useful the of the data to the user.