Workshop - Week 5 - 2021SEM1

1. Download, open and study the file smoking data us 1995 2010-fixed.csv, showing United States population smoking data from 1995 to 2010. In the first twenty rows, there are seven errors that all fall into one of the following categories:

   a. Semantic

   b. Range errors

   c. Format errors

   Identify the errors and what category they fall into. Where possible fix the errors manually and save the new spreadsheet as *smoking-info-corrected.csv* Suggest how you would write a program to detect them.

2. Write a python code for the following tasks.

   a. Import your file *smoking-info-corrected.csv* into a pandas data frame

   b. Remove the percentage symbols from the data. For removing/replacing characters, see

   c. http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.replace.html and after the removals you will also need to convert all the strings to numeric values (*XXX.apply(pd.to numeric)*)