CS 5402 – Intro to Data Mining
Fall 2021
HW #1

- This assignment is **due by 11:59 p.m. on Monday, Sep. 13, 2021**.
- You are to work on this assignment by yourself. It's ok to discuss general approaches and help one another with technical questions, but your overall work should be your own.
- This assignment is worth **60 points**.

Project Description

For this assignment you are to **preprocess/clean** a dataset. You are only allowed to use **Python and/or Weka methods**; part of the objective of this assignment is to have you practice those methods (as opposed to using Microsoft Excel, R, C++, etc.).

The dataset (**census.csv**), which is posted on Canvas along with this assignment, contains U.S. census data from 1994. It contains **32561 instances** which have the following **16 attributes**:

- **Date**
- **Age** (integer)
- **Workclass** (e.g., Private, Self-emp-not-inc, Federal-gov, etc.)
- **Population-wgt** (integer)
- **Education** (e.g., Bachelors, Some-college, 11th, etc.)
- **Education-num** (integer)
- **Marital-status** (e.g., Divorced, Never-married, etc.)
- **Occupation** (e.g., Tech-support, Sales, etc.)
- **Relationship** (e.g., Wife, Husband, etc.)
- **Race** (e.g., White, Other, etc.)
- **Sex** (e.g., Female, Male)

- **Capital-gain** (integer)
- **Capital-loss** (integer)
- **Hours-per-week** (integer)
- **Native-country** (e.g., United-States, England, etc.)
- **Over-under-50k** (e.g., >50K, <=50K); this is the decision attribute

Specifically, here are the **only preprocessing/cleaning tasks** that you are to perform:

1. **Date**: make the dates have a consistent format (e.g., MM/DD/YYYY); also, if any date has a year other than 1994, change the year to 1994.
2. **Age**: discretize the values into 10 bins using equal width (note: effectively, this now makes Age into a nominal attribute). BTW: 10 was just an arbitrary choice for number of bins here; think about how you would decide on a "good" choice for the number of bins (but specify 10 for this assignment!).
3. **Workclass**: replace missing values (represented as ?) with Other.
4. **Population-wgt**: normalize the values.
5. **Occupation**: replace missing values (represented as ?) with Other.
6. **Sex**: fix typos (valid values are Male and Female).
7. **Hours-per-week**: discretize the values into 30 bins using equal frequency; don't worry if you actually get fewer than 30 bins (note: effectively, this now makes Hours-per-week into a nominal attribute). BTW: 30 was just an arbitrary choice for number of bins here; think about how you would decide on a "good" choice for the number of bins (but specify 30 for this assignment!).
8. **Native-country**: replace missing values (represented as ?) with Unspecified.
9. Perform a **chi-square test** (using 0.05 for significance) between **each pair** of nominal-valued (non-decision) attributes; identify which attributes are not independent of each other by filling in the entries in the table shown below as I=Independent or N=Not independent:

	age	workclass	education	marital-status	occupation	relationship	race	sex	hours-per-week	native-country
age										
workclass										
education										
marital-status										
occupation										
relationship										
race										
sex										
hours-per-week										
native-country										

10. Perform a **Spearman test** between **each pair** of non-nominal (non-decision) attributes; identify which attributes are not independent of each other by filling in the entries in the table shown below as I=Independent or N=Not independent. For the purposes of this assignment, consider the absolute value of correlation coefficient ≥ 0.8 as being "close to 1."

	date	population-wgt	education-num	capital-gain	capital-loss
date					
population-wgt					
education-num					
capital-gain					
capital-loss					

11. In preparation to perform **Principal Components Analysis (PCA)**: (i) change all non-numeric attributes to numeric (one way to do this is with discretization/binning), and standardize each attribute.

12. In **Python**, perform **Principal Components Analysis (PCA)** using **all** of the non-decision attributes. Determine the **10 "most important" attributes** by considering the cumulative contribution that each attribute makes in the **first seven** principal components (as discussed in lecture). Provide results (e.g., eigenvalue and eigenvector values, etc.) that justify your determination. Note: You might ask yourself why we're looking at 7 PCs instead of just 2; consider how much variance each PC is responsible for in the dataset.

13. Using **Python or Weka**, use **all** of the non-decision attributes to determine which pairs of attributes are **most strongly correlated** (positively or negatively - specify which type of correlation for each pair). For this problem, consider a "strong" correlation as one that has absolute value ≥ 0.15 (which was somewhat arbitrarily selected for this assignment). Provide results (e.g., correlation matrix, etc.) that justify your determination.

What To Submit for Grading

You should submit a **zip** file that contains **only two** items:

(1) A single **pdf file** that that **CLEARLY identifies** how you performed **EACH task** (e.g., Python source code, Weka KnowledgeFlow screenshots). Additionally, **provide answers for what you are being asked for in tasks 9, 10, 12, and 13**.

(2) A **csv file** containing your transformed data (this includes transformations done for the PCA).

If your submission contains more than this, we reserve the right to DEDUCT POINTS from your homework score for wasting the grader's time; he has to grade numerous submissions and doesn't have time to wade through extraneous material!

Grading:

Here's how many points each task is worth:

Task	Points Possible
Date: make format consistent and year be 1994	2
Age: discretize into 10 bins using equal width	2
Workclass: replace missing values with Other	2
Population-wgt: normalize values	4
Occupation: replace missing values with Other	2
Sex: fix typos	3
Hours-per-week: discretize into 30 bins using equal frequency	2
Native-country: replace missing values with Unspecified	2
Chi-square test	12
Spearman test	6
Change nominal attributes to numeric for PCA	5
Standardize each attribute for PCA	2
Run PCA and analyze eigenvalue/eigenvector results	10
Correlation analysis	6
Total	**60**