Homework 2

1. After unzipping the Kaggle CSV files, make a new directory for the original zip files, and move the files there. In case you accidentally mess up one of the CSV files, you'll be able unzip the data again.

- step one: create a new directory to put all the zip files in using.

```
mkdir directory_name
```

- step two: move all the zip files to that directory.

```
mv *.zip zipfiles/
```

2. The "diabetes_prediction_dataset.csv" file has a lot of entries. Create 3 new CSV files, each with about 1/3 of the data.

```
nano filename.csv #to open and see what the whole data looks like head
filename.csv #to see a couple of the datas
```

```
awk -F',' '{print NF; exit}' filename.csv #to see how many colunms
```

```
wc -l filename.csv #to see how many rows there are.
```

```
head -n 1 existing.csv > file1.csv #to create 3 files with just the first
line by redirecting output of head into a file using >.
```

```
head -n 3335 diabetes_prediction_dataset.csv | tail -n +2 >> diabetes1.csv
head -n 6668 diabetes_prediction_dataset.csv | tail -n +3336 >> diabetes2.csv
head -n 10001 diabetes_prediction_dataset.csv | tail -n +6669 >>
diabetes3.csv #Chain/pipe head and tail to select specific lines, redirecting
output to append to the 3 files you created using >>.
```

3. Create 2 new CSV files from `Heart_Disease_Prediction.csv`, one containing rows with "Presence" label and another with "Absence" label. Make sure that the first line of each file contains the field names.

```
head Heart_Disease_Prediction.csv # to see the data
```

```
head -n 1 Heart_Disease_Prediction.csv >heartabsence.csv head -n 1
Heart_Disease_Prediction.csv >heartpresence.csv #First create 2 files with
just the first line by redirecting output of head into a file using >.grep
"Absence" Heart_Disease_Prediction.csv >> heartabsence.csv
```

```
grep "Absence" Heart_Disease_Prediction.csv >> heartabsence.csv grep
"Presence" Heart_Disease_Prediction.csv >> heartpresence.csv #Use grep to
select lines that contain "Absence" or "Presence" and append the output to
the appropriate file created in the previous step.
```

4. What fraction of cars in `car_web_scraped_dataset.csv` have had no accidents? 2223

```
grep "No accidents reported" car_web_scraped_dataset.csv | wc -l #Use grep to
select the appropriate lines. #Pipe the output of grep into wc (using |) to
count the lines.
```

5. Make the following replacements in `Housing.csv`, output the result into a new CSV:

- yes → 1
- no → 0
- unfurnished → 0
- furnished → 1
- semi-furnished → 2

```
head -n 1 Housing.csv >Housing.csv
```

```
unknown option to `s' bettynega@Bettynega:~$ sed -e 's/semi-furnished/2/g' \ -e 's/unfurnished/0/g' \ -e 's/furnished/1/g' \ -e 's/yes/1/g' \ -e 's/no/0/g' \ Housing.csv > new_Housing.csv #this is to make chnages in the Housing.csv file and transfer it to Newhousing.csv
```

```
rm filename.csv #to delete a csv file
```

6. Create a new CSV file from `Mall_Customers`, removing "CustomerID" column.

```
cut -d ',' -f 2- Mall_Customers.csv >NoIDmall_customers.csv #use 'cut' to cut out the first column and start from second column on.
```

7. Create a new file that contains the sum of the following fields for each row:

- Research Quality Score
- Industry Score

- International Outlook
- Research Environment Score

8. Sort the "cancer patient data sets.csv" file by age. Make sure the output is a readable CSV file.

```
sort -t ',' -k 3n "cancer patient data sets.csv" > "sorted cancer patient
data sets.csv"
```

make a homework folder on Data4380 directory:

```
mkdir Homework cp /mnt/c/Users/betty/Downloads/lecture.5.ipynb ~/ #this moves
lecture.5.ipynb in to lunix mv Lecture.5.ipynb Data4380/Homework #this moves
it from
```