IntroShell

Tom Bohbot

April 2021

1 Find the number of colleges with the word college in their name.

```
tombohbot@Toms-MacBook-Pro HBDdatasets % grep -i -c "college" unirank.csv
8
```

Figure 1: Number of colleges

2 Find the percentage of colleges with the word college in their name, relative to the size of the overall data- set

```
tombohbot@Toms-MacBook-Pro Desktop % college_count=$(grep -i -c "college" unirank.csv)
count=$(tail -n +2 unirank.csv | grep -i -c "")
echo "scale=4; ($college_count / $count) * 100" | bc
3.4600
```

Figure 2: Percentage of colleges with word college in name, relative to overall dataset

3 Which state has the most "institutions"?

```
tombohbot@Toms-MacBook-Pro Desktop % cut -f 3 -d, unirank.csv | tail -n +2 | sort | uniq -c | sort -r
  22  CA
  21  NY
  15  MA
  12  PA
  11  IL
  10  TX
   9  OH
   8  NC
   7  VA
   7  NJ
   7  MO
   7  FL
   6  MI
   5  TN
   5  IN
   5  DC
   5  CO
   4  GA
   4  AL
   3  WI
   3  WA
   3  UT
   3  OK
   3  MS
   3  MD
   3  LA
   3  CT
   2  SD
   2  SC
   2  RI
   2  OR
   2  NM
   2  NH
   2  ND
   2  MT
   2  MN
   2  KY
   2  KS
   2  IA
   2  AZ
   1  WY
   1  WV
   1  VT
   1  NV
   1  NE
   1  ME
   1  ID
   1  HI
   1  DE
   1  AR
   1  AK
```

Figure 3: California has the most institutions

4 Are tuition and rank correlated

```
gnuplot> set title 'Are rank and tuition correlated'
gnuplot> set xlabel 'rank'
gnuplot> set ylabel 'tuition'
gnuplot> plot 'unirank.csv'
```

Figure 4: code to produce graph

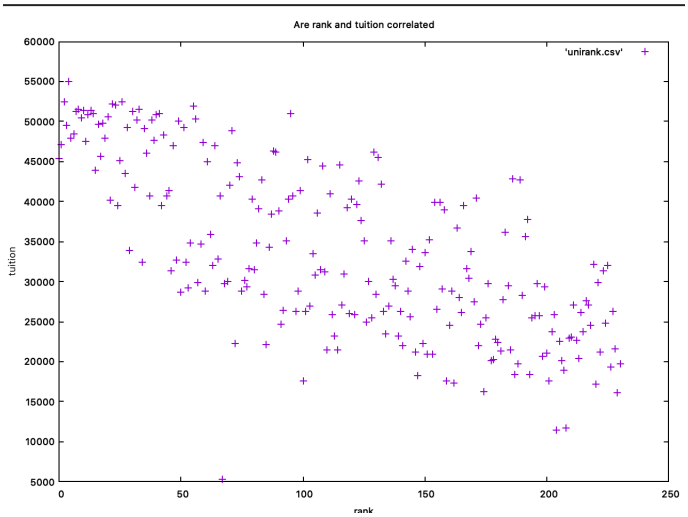

Figure 5: Graph showing correlation

These two columns are clearly negatively correlated as there is a downward slope in which the less money spent on tuition correlates to a lower rank.

Concerning outliers, there is one at around rank 60 which costs very little money. However, an outlier does not disprove the correlation, and if there were to be a trend line inserted into the plot it would clearly show a descending line.

I initially tried to use a cut command to only use the necessary data, but I had trouble applying those commands to gnuplot, so I then tried to resemble the examples you provided through using a "using start:end" command but kept receiving errors. Instead, I transformed the data via the csv file. The way I found the info was through deleting any info that was not tuition or rank in the csv, and then simply plotted all the data in the csv file that remained.