Homework 3 DATA3421

LeMaur Kydd

2023-02-22

```
a = matrix(1:36,3,12, byrow = T)
a
```

1.) Using R, create a 3*12 matrix (3 rows and 12 columns); then calculate the means for each column of the matrix.

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
## [1,]    1    2    3    4    5    6    7    8    9    10    11    12
## [2,]   13   14   15   16   17   18   19   20   21    22    23    24
## [3,]   25   26   27   28   29   30   31   32   33    34    35    36
```

```
a_mean = apply(a,2,mean)
a_mean
```

```
##  [1] 13 14 15 16 17 18 19 20 21 22 23 24
```

```
job = read.csv('job.csv')
balance = job$Balance

sum(balance)
```

2.Calculate sum, mean, median, standard deviation, skewness, quantile, kurtosis, and variance for variable "Balance" in the job data set.

```
## [1] 159622523
```

```
mean(balance)
```

```
## [1] 39766.45
```

```
median(balance)
```

```
## [1] 33567.33
```

```
sd(balance)
```

```
## [1] 29859.49
```

```
skewness(balance)
```

```
## [1] 0.9755534
```

```
quantile(balance)
```

```
##        0%       25%       50%       75%      100%
##     11.52  16115.37  33567.33  57533.93 183467.70
```

```
kurtosis(balance)
```

```
## [1] 0.7675059
```

```
var(balance)
```

```
## [1] 891589095
```

```
age = job$Age
quantile(age, probs = c(0.30,0.60,0.80))
```

3.Recode the "Age" variable to three categories of "Young Adult","Middle-aged Adult", and "Old Adult", based on quantiles of 30%, 60%,and 80%.

```
## 30% 60% 80%
##  33  40  47
```

```
age_categorized = cut(age,breaks = c(-Inf,33,40,47,Inf), labels = c('YoungAdult','Middle-agedAdult','Ol
head(age_categorized, n = 10)
```

```
##  [1] YoungAdult       Middle-agedAdult OldAdult         YoungAdult
##  [5] Middle-agedAdult YoungAdult       Middle-agedAdult Other
##  [9] YoungAdult       OldAdult
## Levels: YoungAdult Middle-agedAdult OldAdult Other
```

```
bal_sorted = job[order(job$Balance, decreasing = T), ]
bal_jobclass_sorted = bal_sorted[c('Balance','Job.Classification')]
head(bal_jobclass_sorted, n= 10)
```

4.Sort the "Balance" variable from high to low in the job data set and create a new subset of the data set with just Balance and Job Classification variables.

```
##       Balance Job.Classification
## 1307 183467.7       White Collar
## 3450 181681.0       White Collar
## 2010 172085.5              Other
## 3832 161517.8        Blue Collar
## 914  149698.1        Blue Collar
## 1823 149684.4       White Collar
## 3108 146892.4       White Collar
## 639  146569.8       White Collar
## 524  145996.0       White Collar
## 3185 144607.4       White Collar
```

```
job_subset = subset(job, Region == 'England' & Balance > 100000)
head(job_subset, n=10)
```

5.Subset the job data set with just the England region and with a balance of higher than 100,000.

```
##     Gender Age  Region Job.Classification  Balance
## 1     Male  21 England       White Collar 113810.1
## 3     Male  46 England       White Collar 101536.8
## 120 Female  31 England       White Collar 136370.4
## 124   Male  39 England              Other 111149.5
## 183   Male  31 England       White Collar 115312.9
## 206 Female  36 England       White Collar 109026.8
## 227   Male  29 England              Other 106172.1
## 243   Male  47 England        Blue Collar 116144.9
## 274 Female  34 England              Other 111818.3
## 281   Male  37 England        Blue Collar 103875.8
```

```
job_subset = job_subset[c('Region','Balance')]
head(job_subset, n=10)
```

```
##      Region  Balance
## 1   England 113810.1
## 3   England 101536.8
## 120 England 136370.4
## 124 England 111149.5
## 183 England 115312.9
## 206 England 109026.8
## 227 England 106172.1
## 243 England 116144.9
## 274 England 111818.3
## 281 England 103875.8
```

```
delay_time = mutate(flights, delay_time=arr_time-sched_arr_time)
head(delay_time$delay_time, n=10)
```

6.Using the flights data set, create a new variable using existing variables(arr_time and sched_arr_time) and call it delay time.

```
##  [1]  11  20  73 -18 -25  12  59 -14  -8   8
```