FUNDAMENTALS OF DATA SCIENCE

HOMEWORK – 1

NAME – GOUTHAM SELVAKUMAR

ID - 2092286

PROBLEM – 1

INSTALLING PACKAGES:

```
> # PROBLEM 1library(tidyverse)
> library(psych)
Warning message:
package 'psych' was built under R version 4.0.5
> library(matrixStats)
Warning message:
package 'matrixStats' was built under R version 4.0.5
> library(GGally)
Loading required package: ggplot2

Attaching package: 'ggplot2'

The following objects are masked from 'package:psych':

    %+%, alpha

Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
Warning messages:
1: package 'GGally' was built under R version 4.0.5
2: package 'ggplot2' was built under R version 4.0.5
```

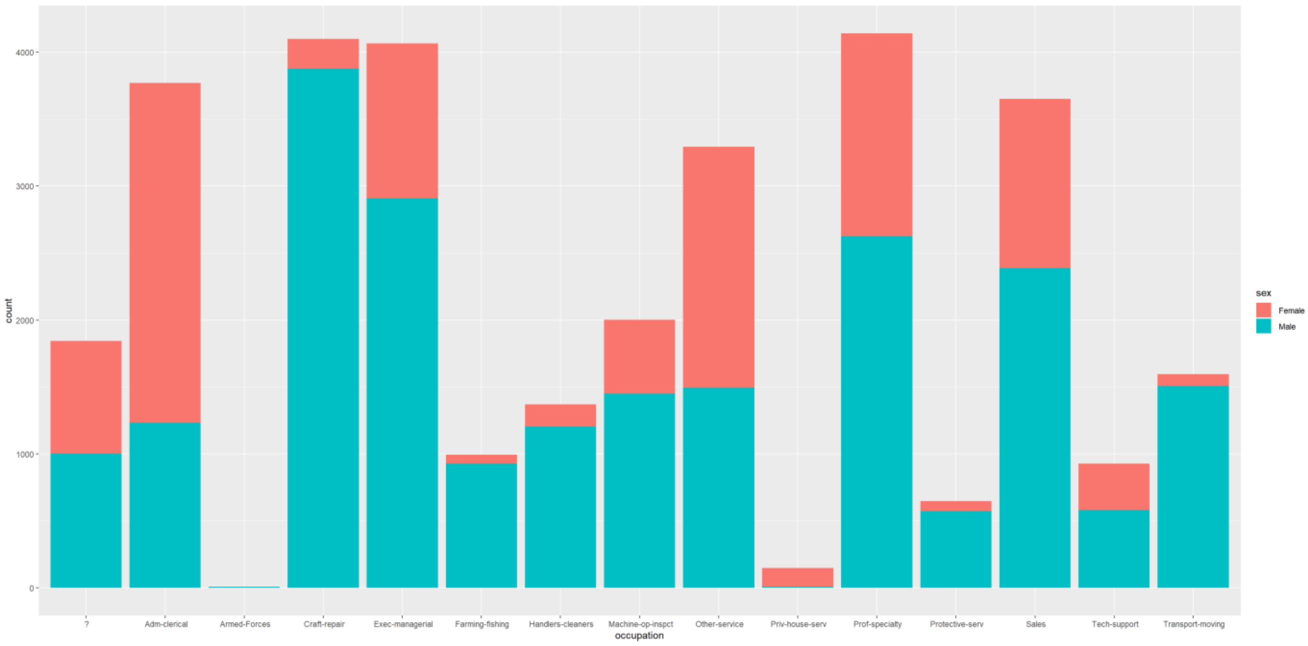
a. First, we look at the summary statistics for all the variables. Based on those metrics, including the quartiles, compare two variables. What can you tell about their shape from these summaries?

```
> setwd("C:/Users/admin/Desktop")
> adult<-read.csv("adult.csv")
> # Describing the summary statistics
> summary(adult)
      age         workclass             fnlwgt          education         education.num
 Min.   :17.00   Length:32561       Min.   :  12285   Length:32561       Min.   : 1.00
 1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character   1st Qu.: 9.00
 Median :37.00   Mode  :character   Median : 178356   Mode  :character   Median :10.00
 Mean   :38.58                      Mean   : 189778                      Mean   :10.08
 3rd Qu.:48.00                      3rd Qu.: 237051                      3rd Qu.:12.00
 Max.   :90.00                      Max.   :1484705                      Max.   :16.00
 marital.status      occupation        relationship          race               sex
 Length:32561       Length:32561       Length:32561       Length:32561       Length:32561
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character



  capital.gain    capital.loss    hours.per.week  native.country     income.bracket
 Min.   :    0   Min.   :   0.0   Min.   : 1.00   Length:32561       Length:32561
 1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00   Class :character   Class :character
 Median :    0   Median :   0.0   Median :40.00   Mode  :character   Mode  :character
 Mean   : 1078   Mean   :  87.3   Mean   :40.44
 3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
 Max.   :99999   Max.   :4356.0   Max.   :99.00
```

Based on this summary, I decided to compare sex to occupation. Based on this summary alone, since the two chosen variables are categoricals, we cannot tell what the relationship is between them.

<u>b.</u> Use a visualization to get a fine-grain comparison (you don't have to use QQ plots, though) of the distributions of those two variables. Why did you choose the type of visualization that you chose? How do your part (a) assumptions compare to what you can see visually?

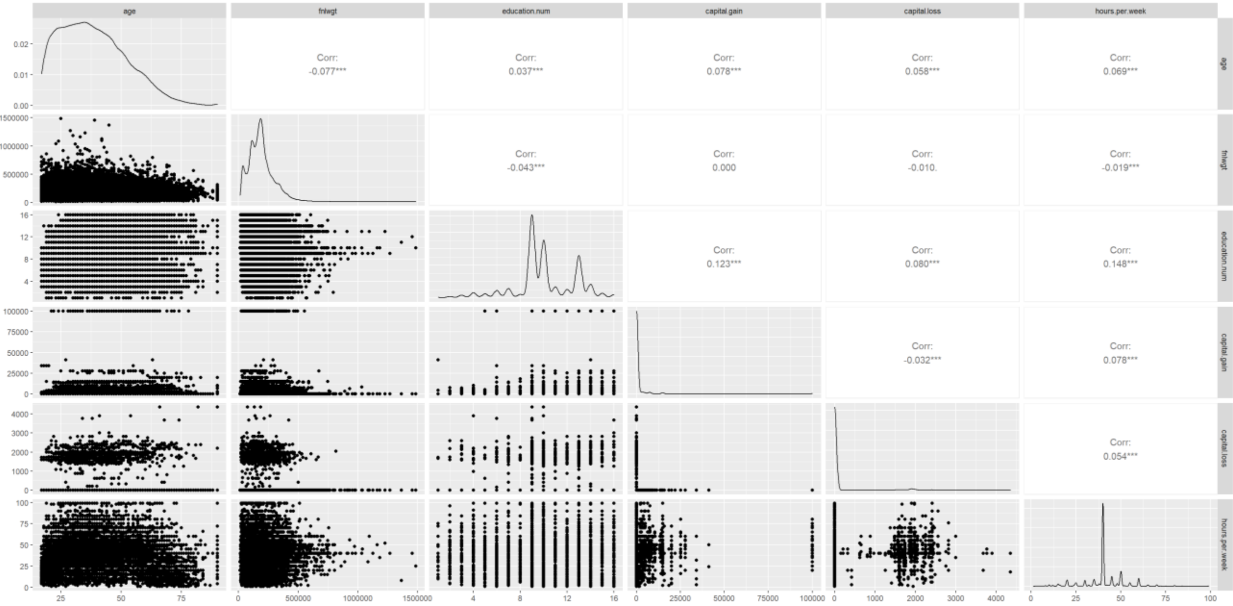
```
# Using ggplot for visualizing
df <- as.data.frame(adult)
p <- ggplot(adult, aes(x=occupation, fill=sex))
p + geom_bar(position="stack")
```



- I chose a Bar Chart to display the number of females and males in each occupation and color-coded it.
- While the initial summary did not provide an accurate description of the data due to the non-numeric nature of the input, the Bar Chart allowed us to visualize it better.
- Per the bar chart above, you can see a clear distribution of how many females and males are represented in each occupation. For example, while not many in this dataset are Priv-house-serv, the females in this dataset dominate this occupation. On the other hand, many males and females are working in the craft-repair field, however, males clearly dominate this field.

c. Now create a scatterplot matrix of the numerical variables. What does this view show you that would be difficult to see looking at distributions?

```
# Creating a scatter plot matrix of the numerical variables
ggpairs(adult[c(1,3,5,11,12,13)])
```



- There are a few things to point out on this Scatter Plot Matrix: 1. The number of adults in this dataset decreases with age, as in the data may be biased towards ages of 17-70, with data points decreasing dramatically after age 75. This could be due to several factors including mortality and employment rates at higher age groups.
- The Pearson-Correlation coefficient is near 0 for most correlations, however, there is a positive correlation between education-num and age. Therefore, this suggests that there is a positive correlation between the years of education and the age of the person.

d. These data are a selection of US adults. It might not be a very balanced sample, though. Take a look at some categorical variables and see if any have a lot more of one category than others. There are many ways to do this, including histograms and following tidyererse group by with count. I recommend you try a few for practice.

```
> adult %>% group_by(race) %>% summarise("count"=n())
# A tibble: 5 x 2
  race                    count
  <chr>                   <int>
1 " Amer-Indian-Eskimo"     311
2 " Asian-Pac-Islander"    1039
3 " Black"                 3124
4 " Other"                  271
5 " White"                27816
```

```
> adult %>% group_by(education) %>% summarise("count"=n())
# A tibble: 16 x 2
   education       count
   <chr>           <int>
 1 " 10th"           933
 2 " 11th"          1175
 3 " 12th"           433
 4 " 1st-4th"        168
 5 " 5th-6th"        333
 6 " 7th-8th"        646
 7 " 9th"            514
 8 " Assoc-acdm"    1067
 9 " Assoc-voc"     1382
10 " Bachelors"     5355
11 " Doctorate"      413
12 " HS-grad"      10501
13 " Masters"       1723
14 " Preschool"       51
15 " Prof-school"    576
16 " Some-college"  7291
```

I chose to view data using group-by to view information on race and education. The white race dominates this dataset, and most within this dataset have a bachelor's degree.

<u>e.</u> Now we'll consider a relationship between two categorical variables. Create a cross tabulation and then a corresponding visualization and explain a relationship between some of the values of the categoricals.

```
> dummy <- dummyVars(sex ~ ., data = adult)
> dummies <- as.data.frame(predict(dummy, newdata = adult))
> head(dummies)
  age workclass ? workclass Federal-gov workclass Local-gov workclass Never-worked workclass Private
1  39           0                     0                   0                      0                 0
2  50           0                     0                   0                      0                 0
3  38           0                     0                   0                      0                 1
4  53           0                     0                   0                      0                 1
5  28           0                     0                   0                      0                 1
6  37           0                     0                   0                      0                 1
  workclass Self-emp-inc workclass Self-emp-not-inc workclass State-gov workclass Without-pay fnlwgt
1                      0                          0                   1                     0  77516
2                      0                          1                   0                     0  83311
3                      0                          0                   0                     0 215646
4                      0                          0                   0                     0 234721
5                      0                          0                   0                     0 338409
6                      0                          0                   0                     0 284582
  education 10th education 11th education 12th education 1st-4th education 5th-6th education 7th-8th
1              0              0              0                 0                 0                 0
2              0              0              0                 0                 0                 0
3              0              0              0                 0                 0                 0
4              0              1              0                 0                 0                 0
5              0              0              0                 0                 0                 0
6              0              0              0                 0                 0                 0
  education 9th education Assoc-acdm education Assoc-voc education Bachelors education Doctorate
1             0                    0                   0                   1                   0
2             0                    0                   0                   1                   0
3             0                    0                   0                   0                   0
4             0                    0                   0                   0                   0
5             0                    0                   0                   1                   0
6             0                    0                   0                   0                   0
  education HS-grad education Masters education Preschool education Prof-school
1                 0                 0                   0                     0
2                 0                 0                   0                     0
3                 1                 0                   0                     0
4                 0                 0                   0                     0
5                 0                 0                   0                     0
6                 0                 1                   0                     0
  education Some-college education.num marital.status Divorced marital.status Married-AF-spouse
1                      0            13                       0                                0
2                      0            13                       0                                0
3                      0             9                       1                                0
4                      0             7                       0                                0
5                      0            13                       0                                0
6                      0            14                       0                                0
  marital.status Married-civ-spouse marital.status Married-spouse-absent
1                                 0                                    0
2                                 1                                    0
3                                 0                                    0
4                                 1                                    0
5                                 1                                    0
6                                 1                                    0
  marital.status Never-married marital.status Separated marital.status Widowed occupation ?
1                            1                        0                      0            0
2                            0                        0                      0            0
3                            0                        0                      0            0
4                            0                        0                      0            0
5                            0                        0                      0            0
6                            0                        0                      0            0
```

```
  occupation Farming-fishing occupation Handlers-cleaners occupation Machine-op-inspct
1                          0                            0                           0
2                          0                            0                           0
3                          0                            1                           0
4                          0                            1                           0
5                          0                            0                           0
6                          0                            0                           0
  occupation Other-service occupation Priv-house-serv occupation Prof-specialty
1                        0                          0                         0
2                        0                          0                         0
3                        0                          0                         0
4                        0                          0                         0
5                        0                          0                         1
6                        0                          0                         0
  occupation Protective-serv occupation Sales occupation Tech-support occupation Transport-moving
1                          0                0                       0                           0
2                          0                0                       0                           0
3                          0                0                       0                           0
4                          0                0                       0                           0
5                          0                0                       0                           0
6                          0                0                       0                           0
  relationship Husband relationship Not-in-family relationship Other-relative relationship Own-child
1                    0                          1                           0                      0
2                    1                          0                           0                      0
3                    0                          1                           0                      0
4                    1                          0                           0                      0
5                    0                          0                           0                      0
6                    0                          0                           0                      0
  relationship Unmarried relationship Wife race Amer-Indian-Eskimo race Asian-Pac-Islander
1                      0                 0                       0                       0
2                      0                 0                       0                       0
3                      0                 0                       0                       0
4                      0                 0                       0                       0
5                      0                 1                       0                       0
6                      0                 1                       0                       0
  race Black race Other race White capital.gain capital.loss hours.per.week native.country ?
1          0          0          1         2174            0             40                0
2          0          0          1            0            0             13                0
3          0          0          1            0            0             40                0
4          1          0          0            0            0             40                0
5          1          0          0            0            0             40                0
6          0          0          1            0            0             40                0
  native.country Cambodia native.country Canada native.country China native.country Columbia
1                       0                     0                    0                       0
2                       0                     0                    0                       0
3                       0                     0                    0                       0
4                       0                     0                    0                       0
5                       0                     0                    0                       0
6                       0                     0                    0                       0
  native.country Cuba native.country Dominican-Republic native.country Ecuador
1                   0                                 0                      0
2                   0                                 0                      0
3                   0                                 0                      0
4                   0                                 0                      0
5                   1                                 0                      0
6                   0                                 0                      0
```

```
  native.country El-Salvador native.country England native.country France native.country Germany
1                          0                      0                     0                      0
2                          0                      0                     0                      0
3                          0                      0                     0                      0
4                          0                      0                     0                      0
5                          0                      0                     0                      0
6                          0                      0                     0                      0
  native.country Greece native.country Guatemala native.country Haiti
1                     0                        0                    0
2                     0                        0                    0
3                     0                        0                    0
4                     0                        0                    0
5                     0                        0                    0
6                     0                        0                    0
  native.country Holand-Netherlands native.country Honduras native.country Hong
1                                 0                       0                   0
2                                 0                       0                   0
3                                 0                       0                   0
4                                 0                       0                   0
5                                 0                       0                   0
6                                 0                       0                   0
  native.country Hungary native.country India native.country Iran native.country Ireland
1                      0                    0                   0                      0
2                      0                    0                   0                      0
3                      0                    0                   0                      0
4                      0                    0                   0                      0
5                      0                    0                   0                      0
6                      0                    0                   0                      0
  native.country Italy native.country Jamaica native.country Japan native.country Laos
1                    0                      0                    0                   0
2                    0                      0                    0                   0
3                    0                      0                    0                   0
4                    0                      0                    0                   0
5                    0                      0                    0                   0
6                    0                      0                    0                   0
  native.country Mexico native.country Nicaragua native.country Outlying-US(Guam-USVI-etc)
1                     0                        0                                         0
2                     0                        0                                         0
3                     0                        0                                         0
4                     0                        0                                         0
5                     0                        0                                         0
6                     0                        0                                         0
  native.country Peru native.country Philippines native.country Poland native.country Portugal
1                   0                          0                     0                       0
2                   0                          0                     0                       0
3                   0                          0                     0                       0
4                   0                          0                     0                       0
5                   0                          0                     0                       0
6                   0                          0                     0                       0
  native.country Puerto-Rico native.country Scotland native.country South native.country Taiwan
1                          0                       0                    0                     0
2                          0                       0                    0                     0
3                          0                       0                    0                     0
4                          0                       0                    0                     0
5                          0                       0                    0                     0
6                          0                       0                    0                     0
```

```
  native.country Thailand native.country Trinadad&Tobago native.country United-States
1                       0                             0                            1
2                       0                             0                            1
3                       0                             0                            1
4                       0                             0                            1
5                       0                             0                            0
6                       0                             0                            1
  native.country Vietnam native.country Yugoslavia income.bracket <=50K income.bracket >50K
1                      0                         0                    1                   0
2                      0                         0                    1                   0
3                      0                         0                    1                   0
4                      0                         0                    1                   0
5                      0                         0                    1                   0
6                      0                         0                    1                   0
```

```
> # Make a table to look at workclass vs sex summary
> table(adult$workclass, adult$sex)

                    Female  Male
  ?                    839   997
  Federal-gov          315   645
  Local-gov            835  1258
  Never-worked           2     5
  Private             7752 14944
  Self-emp-inc         135   981
  Self-emp-not-inc     399  2142
  State-gov            489   809
  Without-pay            5     9
```

```
# Use new data frame to look at more specific data points
table(dummies$'occupationTransport-moving', dummies$educationBachelors)
```

Based on the cross-tabulation above: 25,671 in this survey said they did not work in transport/moving and do not have a bachelors 5293 do not work in transport/moving but have a bachelors 1535 work in transport/moving but do not have a bachelors 62 work in transport/moving and have a bachelors.

PROBLEM – 2

a. **Join the two tables together so that you have one table with each state's population for years 2010- 2019. If you are unsure about what variable to use as the key for the join, consider what variable the two original tables have in common. (Show a head of the resulting table.)**

```
> population_new<-merge(x= population_odd, y= population_even, by="NAME", all.x = TRUE)
> head(population_new)
        NAME ï..STATE.x POPESTIMATE2011 POPESTIMATE2013 POPESTIMATE2015
1    Alabama          1         4799069         4830081         4852347
2     Alaska          2          722128          737068          737498
3    Arizona          4              NA         6632764         6829676
4   Arkansas          5         2940667         2959400         2978048
5 California          6        37638369        38260787        38918045
6   Colorado          8         5121108         5269035         5450623
  POPESTIMATE2017 POPESTIMATE2019 ï..STATE.y POPESTIMATE2010
1         4874486         4903185          1         4785437
2          739700          731545          2          713910
3         7044008         7278717          4         6407172
4         3001345         3017804          5         2921964
5        39358497        39512223          6        37319502
6         5611885         5758736          8         5047349
  POPESTIMATE2012 POPESTIMATE2014 POPESTIMATE2016 POPESTIMATE2018
1         4815588         4841799         4863525         4887681
2          730443          736283          741456          735139
3         6554978         6730413         6941072         7158024
4         2952164         2967392         2989918         3009733
5        37948800        38596972        39167117        39461588
6         5192647         5350101         5539215         5691287
```

b. **Clean this data up a bit (show a head of the data after): a. Remove the duplicate state ID column if your process created one.**

```
# Delete the duplicate Name.y row
population_new$NAME.y <- NULL

# Rename all column names to correct year numbers
colnames(population_new) <- c("State", "Name", "2011", "2013", "2015", "2017","2019", "2010", "2012", "2014", "2016", "2018")

# Reorder Columns
library(tibble)
population_new<-population_new[, c(1, 2, 3, 9, 4, 10, 5, 11, 6, 12, 7)]
head(population_new)
```

```
       State Name     2011     2012     2013     2014     2015     2016     2017     2018     2019
1    Alabama    1  4799069  4785437  4830081  4815588  4852347  4841799  4874486  4863525  4903185
2     Alaska    2   722128   713910   737068   730443   737498   736283   739700   741456   731545
3    Arizona    4       NA  6407172  6632764  6554978  6829676  6730413  7044008  6941072  7278717
4   Arkansas    5  2940667  2921964  2959400  2952164  2978048  2967392  3001345  2989918  3017804
5 California    6 37638369 37319502 38260787 37948800 38918045 38596972 39358497 39167117 39512223
6   Colorado    8  5121108  5047349  5269035  5192647  5450623  5350101  5611885  5539215  5758736
```

c. **Deal with missing values in the data by replacing them with the average of the surrounding years. For example, if you had a missing value for Georgia in 2016, you would replace it with the average of Georgia's 2015 and 2017 numbers. This may require some manual effort.**

```
> # Find missing values; first use the summary function to find all N/A's
> summary(population_new)
    State                 Name             2011         
 Length:52          Min.   : 1.00   Min.   :  567299  
 Class :character   1st Qu.:16.75   1st Qu.: 1712291  
 Mode  :character   Median :29.50   Median : 3872036  
                    Mean   :29.79   Mean   : 6054176  
                    3rd Qu.:42.50   3rd Qu.: 6720105  
                    Max.   :72.00   Max.   :37638369  
                                    NA's   :1         
      2012               2013               2014         
 Min.   :  564487   Min.   :  582122   Min.   :  576305  
 1st Qu.: 1764843   1st Qu.: 1732560   1st Qu.: 1788808  
 Median : 4092836   Median : 3922468   Median : 4142674  
 Mean   : 6020061   Mean   : 6039414   Mean   : 6105105  
 3rd Qu.: 6610438   3rd Qu.: 6673040   3rd Qu.: 6721518  
 Max.   :37319502   Max.   :38260787   Max.   :37948800  
                    NA's   :1                            
      2015               2016               2017         
 Min.   :  585613   Min.   :  582531   Min.   :  578931  
 1st Qu.: 1866664   1st Qu.: 1794895   1st Qu.: 1866476  
 Median : 4425976   Median : 4188796   Median : 4452268  
 Mean   : 6322693   Mean   : 6189152   Mean   : 6416830  
 3rd Qu.: 6996666   3rd Qu.: 6835611   3rd Qu.: 7233685  
 Max.   :38918045   Max.   :38596972   Max.   :39358497  
 NA's   :1                             NA's   :1         
      2018               2019         
 Min.   :  584215   Min.   :  578759  
 1st Qu.: 1793862   1st Qu.: 1789606  
 Median : 4264079   Median : 4217737  
 Mean   : 6275923   Mean   : 6384525  
 3rd Qu.: 7029497   3rd Qu.: 7446805  
 Max.   :39167117   Max.   :39512223  
                    NA's   :1         
```

```
# For 2015
x<-population_new$`2014`[13] + population_new$`2016`[13]
population_new$`2015`<- population_new$`2015` %>% replace_na(mean(x/2, na.rm = TRUE))
# For 2013
x<-population_new$`2012`[36] + population_new$`2014`[36]
population_new$`2013` <- population_new$`2013` %>% replace_na(mean(x/2, na.rm = TRUE))
# For 2011
x<-population_new$`2012`[3]- (population_new$`2013`[3]-population_new$`2012`[3])
population_new$`2011` <- population_new$`2011` %>% replace_na(mean(x, na.rm = TRUE))
# For 2017
x<-population_new$`2016`[27] + population_new$`2018`[27]
population_new$`2017` <- population_new$`2017` %>% replace_na(mean(x, na.rm = TRUE))
# For 2019
x<-population_new$`2018`[50] + (population_new$`2018`[50]-population_new$`2017`[50])
population_new$`2019` <- population_new$`2019` %>% replace_na(mean(x, na.rm = TRUE))
head(population_new)
```

```
       State Name     2011     2012     2013     2014     2015     2016
1    Alabama    1  4799069  4785437  4830081  4815588  4852347  4841799
2     Alaska    2   722128   713910   737068   730443   737498   736283
3    Arizona    4  6181580  6407172  6632764  6554978  6829676  6730413
4   Arkansas    5  2940667  2921964  2959400  2952164  2978048  2967392
5 California    6 37638369 37319502 38260787 37948800 38918045 38596972
6   Colorado    8  5121108  5047349  5269035  5192647  5450623  5350101
      2017     2018     2019
1  4874486  4863525  4903185
2   739700   741456   731545
3  7044008  6941072  7278717
4  3001345  2989918  3017804
5 39358497 39167117 39512223
6  5611885  5539215  5758736
```

d. We can use some tidy verse aggregation to learn about the population.

a. **Get the maximum population for a single year for each state. Note that because you are using an aggregation function (max) across a row, you will need the row wise () command in your tidy verse pipe. If you do not, the max value will not be individual to the row. Of course there are alternative ways.**

```
# Find maximum per row
maxVal <- pmax(adultscopy$`2011`, adultscopy$`2012`, adultscopy$`2013`, adultscopy$`2014`, adultscopy$`2015`, adultscopy$`2016`,
               adultscopy$`2017`, adultscopy$`2018`, adultscopy$`2019`)
head(maxVal)
```

```
> head(maxVal)
[1]  4903185   741456  7278717  3017804 39512223  5758736
```

b. **Now get the total population across all years for each state. This should be possible with a very minor change to the code from (d). Why is that?**

```
> # sum of each row
> rowSums(adultscopy[,-1])
 [1]  38766448   5867903  54418800  23788035 309081943  43219591  28666873
 [8]   7491595   5314929 160798760  81161158  11271424  13208766 102632724
[15]  52823515  24878233  23179021  35337634  37076161  10652283  47614541
[22]  54075351  79459825  43762667  23874611  48520223   9232758  15074693
[29]  22874562  10688112  70872376  16703236 156579533  80042773   5860832
[36]        NA  31034277  32089334 102199631  27882848   8447466  38960588
[43]   6815688  52619925 217446671  23793929   5000599  66543219  57155211
[50]  14696738  42067510   4632963
```

e. **Finally, get the total US population for one single year. Keep in mind that this can be done with a single line of code even without the tidy verse, so keep it simple.**

```
> sum(adultscopy$`2011`)
[1] 314944543
```

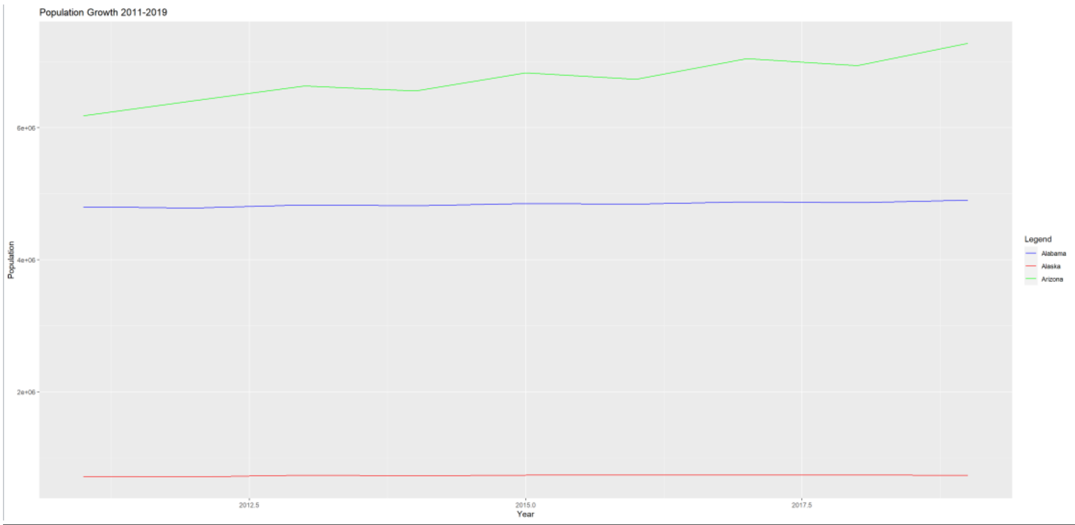
PROBLEM – 3

Continuing with the data from Problem 2, let's create a graph of population over time for a few states (choose at least three yourself). This will require another data transformation, a reshaping. In order to create a line graph, we will need a variable that represents the year, so that it can be mapped to the x axis. Use a transformation to turn all those year columns into one column that holds the year, reducing the 10 year columns down to 2 columns (year and population). Once the data are in the right shape, it will be no harder than any line graph: put the population on the y axis and color by the state. One important point: make sure you have named the columns to have only the year number (i.e.,

without popestimate). That can be done manually or by reading up on string (text) parsing (see the stringr library for a super useful tool). Even after doing that, you have a string version of the year. R is seeing the ‘word’ spelled two-zero-one-five instead of the number two thousand fifteen. It needs to be a number to work on a time axis. There are many ways to fix this. You can look into type_convert or do more string parsing (e.g., stringr). The simplest way is to apply the transformation right as you do the graphing. You can replace the year variable in the ggplot command with as.integer (year).

```
> # Transform Data
> transf <- t(adultscopy)
> head(transf)
        [,1]   [,2]    [,3]    [,4]     [,5]    [,6]    [,7]   [,8]   [,9]
2011 4799069 722128 6181580 2940667 37638369 5121108 3588283 907381 619800
2012 4785437 713910 6407172 2921964 37319502 5047349 3579114 899593 605226
2013 4830081 737068 6632764 2959400 38260787 5269035 3594841 923576 650581
2014 4815588 730443 6554978 2952164 37948800 5192647 3594547 915179 634924
2015 4852347 737498 6829676 2978048 38918045 5450623 3587122 941252 675400
2016 4841799 736283 6730413 2967392 38596972 5350101 3594524 932487 662328
        [,10]    [,11]   [,12]   [,13]    [,14]   [,15]   [,16]   [,17]
2011 19053237  9802431 1379329 1583910 12867454 6516528 3066336 2869225
2012 18845537  9711881 1363963 1570746 12840503 6490432 3050745 2858190
2013 19545621  9972479 1408243 1611206 12895129 6568713 3092997 2893212
2014 19297822  9901430 1394804 1595324 12882510 6537703 3076190 2885257
2015 20209042 10178447 1422052 1613218 12858913 6608422 3120960 2909011
2016 19845911 10067278 1414538 1631112 12884493 6593644 3109350 2900475
       [,18]   [,19]   [,20]   [,21]   [,22]   [,23]   [,24]   [,25]
2011 4369821 4575625 1328284 5839419 6613583 9882412 5346143 2978731
2012 4348181 4544532 1327629 5788645 6566307 9877510 5310828 2970548
2013 4404659 4624527 1328009 5923188 6713315 9913065 5413479 2988711
2014 4386346 4600972 1327729 5886992 6663005 9897145 5376643 2983816
2015 4425976 4664628 1328262 5985562 6794228 9931715 5482032 2988471
2016 4414349 4644013 1330513 5957283 6762596 9929848 5451079 2990468
       [,26]   [,27]   [,28]   [,29]   [,30]   [,31]   [,32]    [,33]
2011 6010275  997316 1840672 2712730 1320202 8828117 2080450 19499241
2012 5995974  990697 1829542 2702405 1316762 8799446 2064552 19399878
2013 6040715 1013569 1865279 2775970 1326622 8856972 2092273 19624447
2014 6024367 1003783 1853303 2743996 1324232 8844942 2087309 19572932
2015 6071732 1030475 1891277 2866939 1336350 8867949 2089291 19654666
2016 6056202 1021869 1879321 2817628 1333341 8864525 2089568 19651049
        [,34]  [,35]    [,36]   [,37]   [,38]    [,39]   [,40]   [,41]
2011  9657592 685225 11544663 3788379 3872036 12745815 3678732 1053649
2012  9574323 674715 11539336 3759944 3837491 12711160 3721525 1053959
2013  9843336 722036       NA 3853214 3922468 12776309 3593077 1055081
2014  9749476 701176 11548923 3818814 3899001 12767118 3634488 1054621
2015 10031646 754066 11617527 3909500 4015792 12784826 3473232 1056065
2016  9932887 737401 11602700 3878187 3963244 12788313 3534874 1055936
       [,42]  [,43]   [,44]    [,45]   [,46]  [,47]   [,48]   [,49]   [,50]
2011 4671994 823579 6399291 25645629 2814384 627049 8101155 6826627 1856301
2012 4635649 816166 6355311 25241971 2775332 625879 8023699 6742830 1854239
2013 4764080 842316 6494340 26480266 2897640 626210 8252427 6963985 1853914
2014 4717354 833566 6453898 26084481 2853375 626090 8185080 6897058 1856872
2015 4891938 853988 6591170 27470056 2981835 625216 8361808 7163657 1842050
2016 4823617 849129 6541223 26964333 2936879 625214 8310993 7054655 1849489
       [,51]  [,52]
2011 5705288 567299
2012 5690475 564487
2013 5736754 582122
2014 5719960 576305
2015 5760940 585613
2016 5751525 582531
```

```r
#use only first 3 columns
transf <- transf[,1:3]
df <- as.data.frame(transf)
# add a column for years 2011 to 2019
df$year <- seq(2011,2019, by=1)
# Let's get plotting
plt = ggplot() +
  geom_line(data = df, aes(x = year, y = V1, group=1), color = "blue")+
  geom_line(data = df, aes(x = year, y = V2, group=1), color = "red")+
  geom_line(data = df, aes(x = year, y = V3, group=1), color = "green")+
  xlab('year')+ ylab('population')
# A graph without a legend is trash, so let's add a legend, manually
colors <- c("Alabama" = "blue", "Alaska" ="red", "Arizona" = "green")
plt = ggplot() +
  geom_line(data = df, aes(x = year, y = V1, group=1, color = "Alabama"))+
  geom_line(data = df, aes(x = year, y= V2, group=1, color = "Alaska"))+
  geom_line(data = df, aes(x = year, y= V3, group=1, color = "Arizona"))+
  labs(x = "Year", y ="Population", color = "Legend") + scale_color_manual(values = colors)
# Add a title
plt <- plt + geom_line() + ggtitle("Population Growth 2011-2019")
print(plt)
```



PROBLEM – 4

a. Describe two ways in which data can be dirty, and for each one, provide a potential solution.

- Data can contain typographical errors, there could be duplicates too. Missing values could be deleted or replaced/corrected by inferring data from known variables, such as using means and percentage to fill gaps.
- Data may have come from different sources/servers/storage mechanisms where the data was handled differently. A potential solution for this would be to add the factors missing in other data (or) simply only relying on same-source data from comparable devices/systems. Otherwise, ignore this data if we have the privilege of doing so.
- Sensor failures; I personally encounter this often and we created failure messages/warnings that trigger a data point in a spreadsheet, which we check every day and use to replace the sensors accordingly. Once a server

fails, the server filters out all data that sensor and completely ignores it. We get an abundant amount of data so ignoring data is not an issue.

b. Explain which data mining functionality you would use to help with each of these data questions.

  a. Suppose we have data where each row is a customer and we have columns that describe their purchases. What are five groups of customers who buy similar things?

  Cluster Analysis – grouping customers based on their similar characteristics.

  b. For the same data: can I predict if a customer will buy milk based on what else they bought?

  Classification and Prediction – Using a historical data to predict future outcomes.

  c. Suppose we have data listing items in individual purchases. What are different sets of products that are often purchased together?

  Association Rule Mining – Events that occur together.

c. Explain if each of the following is a data mining task

  a. Organizing the customers of a company according to education level.

  Not data mining, database query task

  b. Computing the total sales of a company.

  Not data mining, simple mathematical calculation

  c. Sorting a student database according to identification numbers.

  Not data mining, database query task

  d. Predicting the outcomes of tossing a (fair) pair of dice.

  Not data mining, it's a probability calculation

  e. Predicting the future stock price of a company using historical records.

  Data mining, using historical data to predict future outcome.