hw1

2024-09-29

HW 1 - DSC 441

Problem 1

- 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?

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
adult <- read_csv("/Users/danielkim/Downloads/FUNDAMENTALS OF DATA SCIENCE - 9232024 - 133 PM/adult.csv")
```

```
## Rows: 32561 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): workclass, education, marital-status, occupation, relationship, rac...
## dbl (6): age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(adult)
```

```
## Rows: 32,561
## Columns: 15
## $ age              <dbl> 39, 50, 38, 53, 28, 37, 49, 52, 31, 42, 37, 30, 23, 3~
## $ workclass        <chr> "State-gov", "Self-emp-not-inc", "Private", "Private"~
## $ fnlwgt           <dbl> 77516, 83311, 215646, 234721, 338409, 284582, 160187,~
## $ education        <chr> "Bachelors", "Bachelors", "HS-grad", "11th", "Bachelo~
## $ `education-num`  <dbl> 13, 13, 9, 7, 13, 14, 5, 9, 14, 13, 10, 13, 13, 12, 1~
## $ `marital-status` <chr> "Never-married", "Married-civ-spouse", "Divorced", "M~
## $ occupation       <chr> "Adm-clerical", "Exec-managerial", "Handlers-cleaners~
## $ relationship     <chr> "Not-in-family", "Husband", "Not-in-family", "Husband~
## $ race             <chr> "White", "White", "White", "Black", "Black", "White",~
## $ sex              <chr> "Male", "Male", "Male", "Male", "Female", "Female", "~
## $ `capital-gain`   <dbl> 2174, 0, 0, 0, 0, 0, 0, 0, 14084, 5178, 0, 0, 0, 0, 0~
## $ `capital-loss`   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

```
## $ `hours-per-week` <dbl> 40, 13, 40, 40, 40, 40, 16, 45, 50, 40, 80, 40, 30, 5~
## $ `native-country` <chr> "United-States", "United-States", "United-States", "U~
## $ `income-bracket` <chr> "<=50K", "<=50K", "<=50K", "<=50K", "<=50K", "<=50K",~
```

```
dim(adult)
```

```
## [1] 32561    15
```

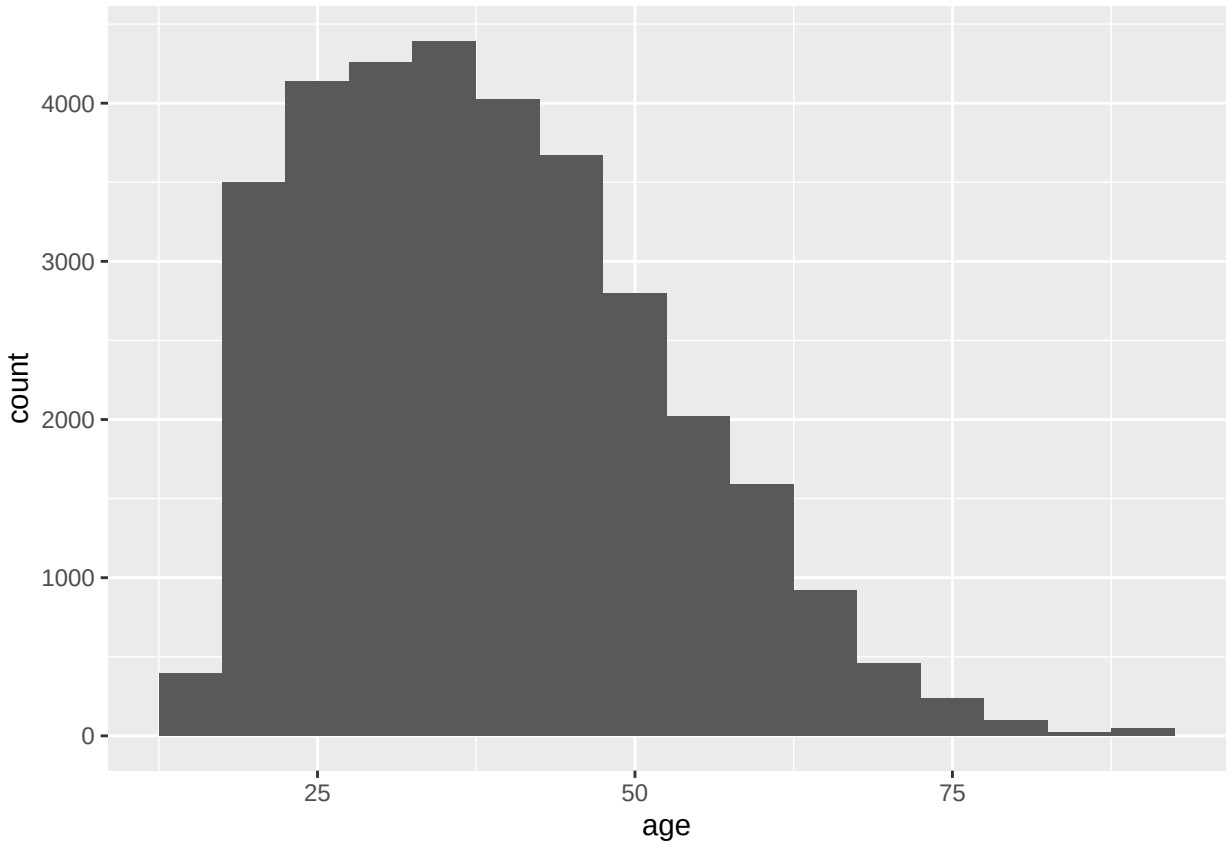
```
summary(adult)
```

```
##       age          workclass             fnlwgt          education
##  Min.   :17.00   Length:32561       Min.   :  12285   Length:32561
##  1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character
##  Median :37.00   Mode  :character   Median : 178356   Mode  :character
##  Mean   :38.58                      Mean   : 189778
##  3rd Qu.:48.00                      3rd Qu.: 237051
##  Max.   :90.00                      Max.   :1484705
##  education-num   marital-status      occupation        relationship
##  Min.   : 1.00   Length:32561       Length:32561       Length:32561
##  1st Qu.: 9.00   Class :character   Class :character   Class :character
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :10.08
##  3rd Qu.:12.00
##  Max.   :16.00
##      race               sex             capital-gain    capital-loss
##  Length:32561       Length:32561       Min.   :    0   Min.   :   0.0
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0
##  Mode  :character   Mode  :character   Median :    0   Median :   0.0
##                                        Mean   : 1078   Mean   :  87.3
##                                        3rd Qu.:    0   3rd Qu.:   0.0
##                                        Max.   :99999   Max.   :4356.0
##  hours-per-week  native-country     income-bracket
##  Min.   : 1.00   Length:32561       Length:32561
##  1st Qu.:40.00   Class :character   Class :character
##  Median :40.00   Mode  :character   Mode  :character
##  Mean   :40.44
##  3rd Qu.:45.00
##  Max.   :99.00
```

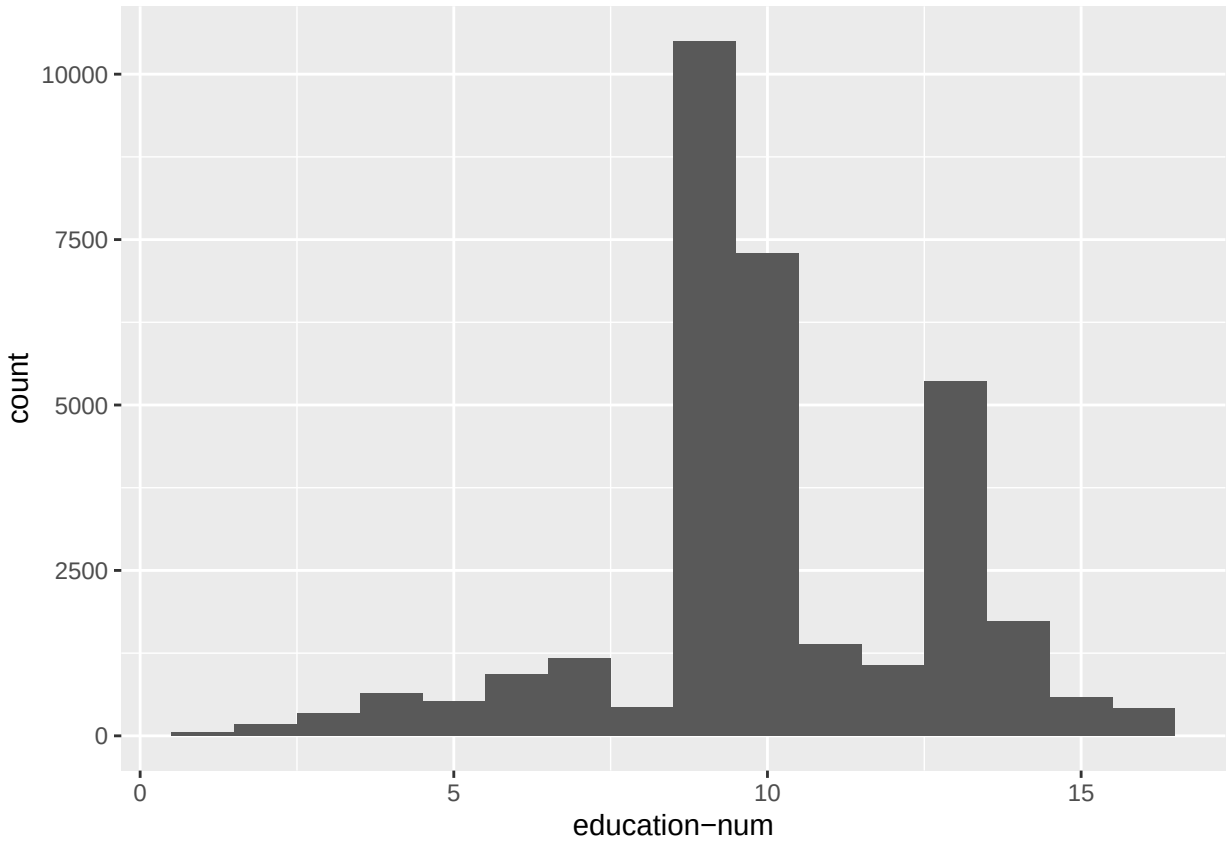
Based on the results of the summary statistics for all the variables, the age of the US population had an unimodal, positively-skewed, or right skewed, distribution in which the difference between maximum age and 3rd quartile was roughly 40 years while the difference between the minimum age and 1st quartile was roughly 10 years. The number of years in education for the US population appears unimodal as well, but with symmetrical distribution with little-to-no skewness given the median and mode of variable are similar. Additionally, the difference between the 1st quartile and 3rd quartile to the minimum and maximum, respectively, are similar in value.

- b: 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?

```
ggplot(adult, aes(age)) + geom_histogram(binwidth = 5)
```



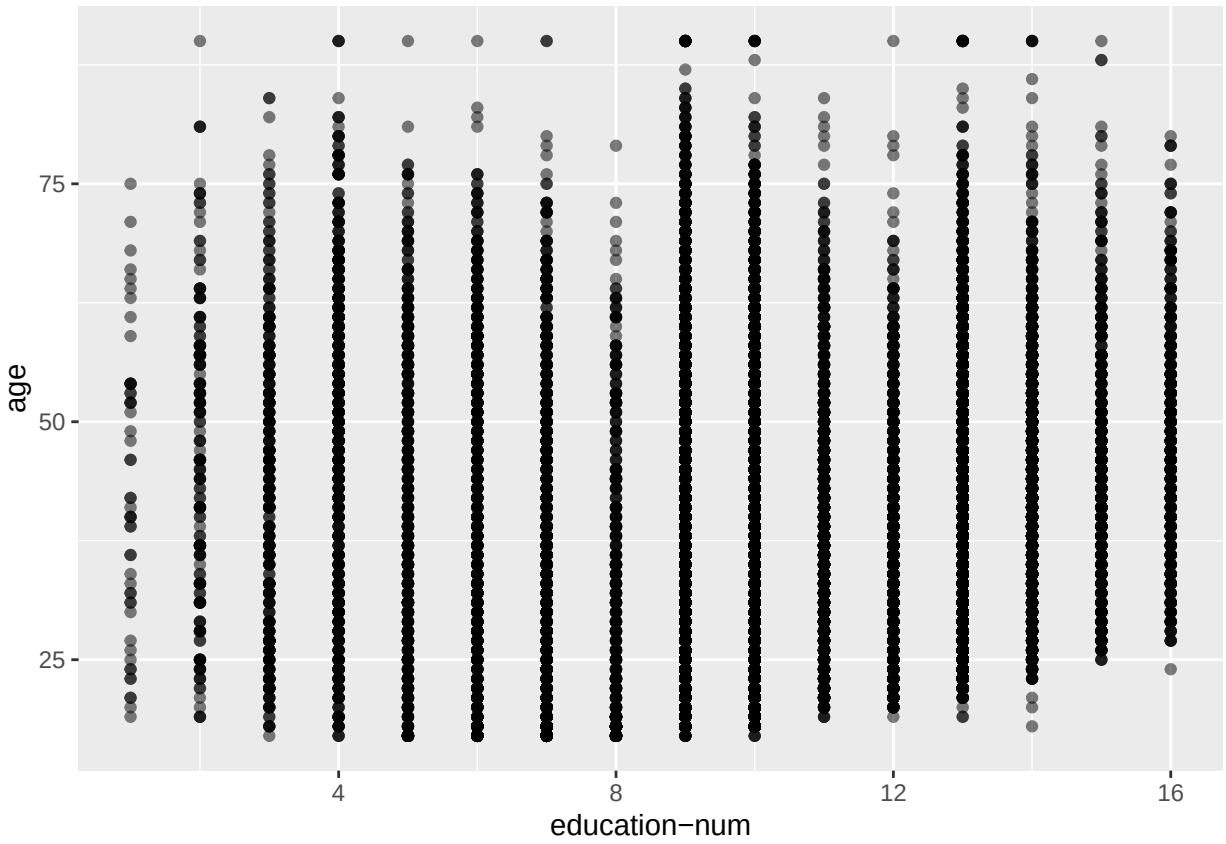

```
ggplot(adult, aes(`education-num`)) + geom_histogram(binwidth = 1)
```



I decided to take advantage of the histogram visualizations to illustrate the distribution of these two numeric variables. The advantage of histograms over other distribution visualizations such as boxplots is the level of detail that can be added to the distribution plot with the parameters binwidth or bins. With these parameters, the histogram can take into account varying amounts of details for the distribution. When compared to my responses in part a) of the assignment, my assumptions were accurate in drawing the distribution for the age variable. However, the extent of this distribution was more severe than initially thought. In this instance, the distribution for the variable had a severe right-sided skewness with higher-valued outliers. The distribution for years of education was bimodal distribution with a mild left-sided skewness, most likely as a result of most students completing highschool and/or college.

- 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?

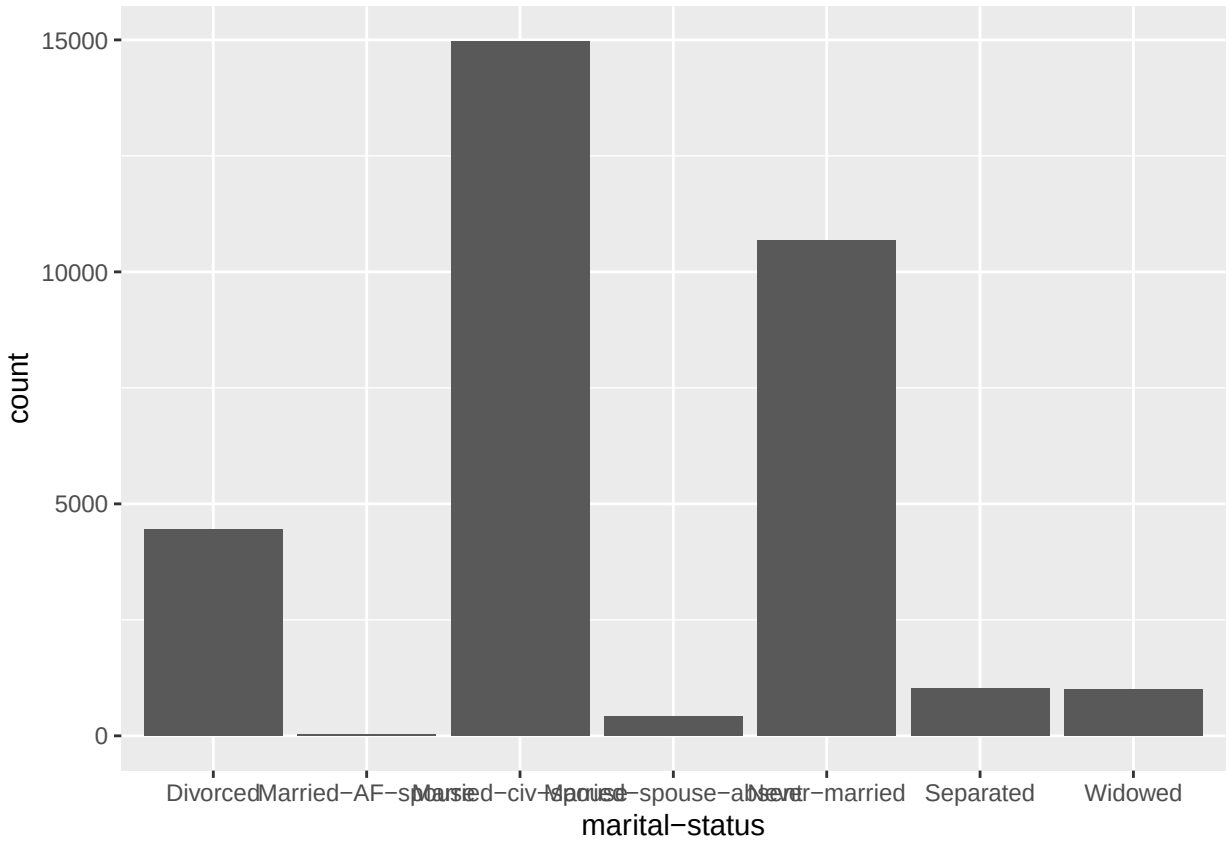
```
ggplot(adult, aes(`education-num`, age)) + geom_point(alpha=0.5)
```



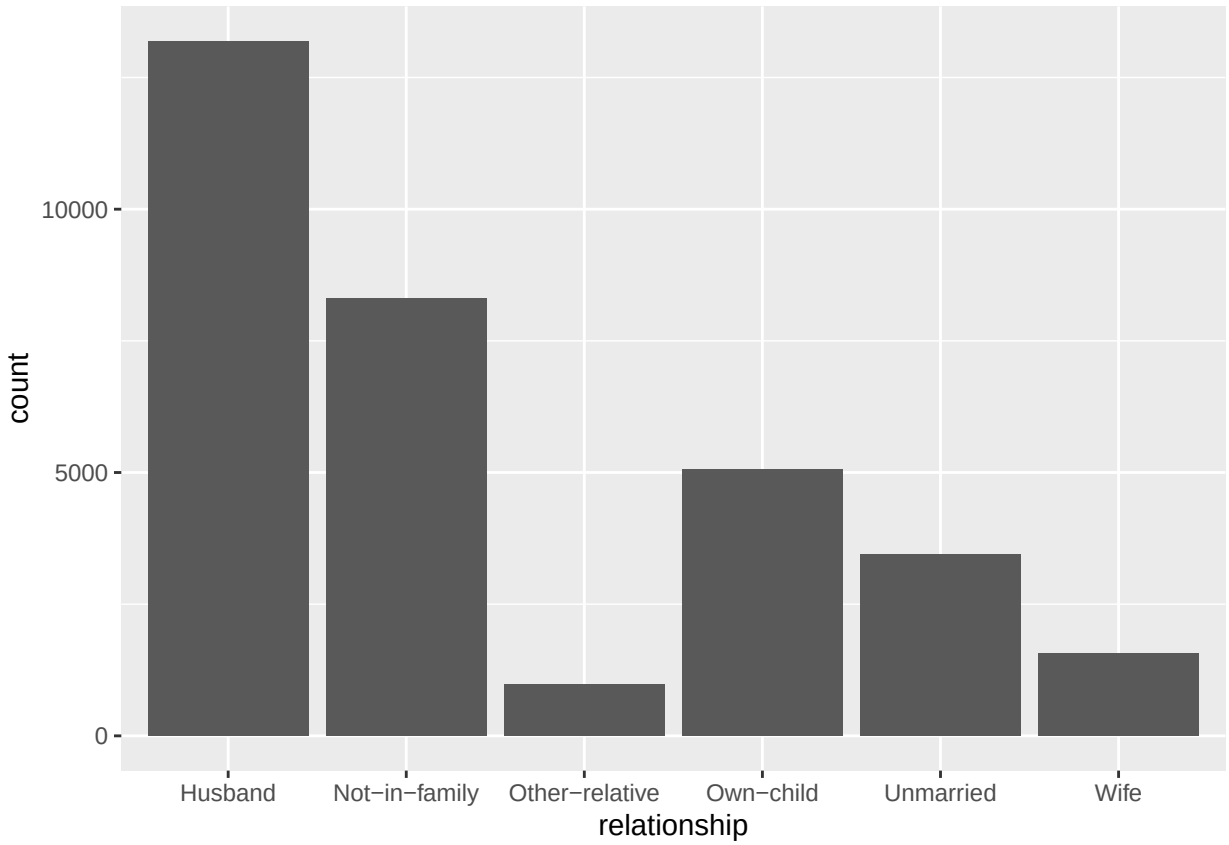
The scatter plots are generally useful to highlight any relationship between two numerical variables. Each point represents a sample or observation through which a discernible pattern can be detected, if any, when looking at data as a whole. In the instance above, the scatter plot illustrates no relationship between age and the number of years in school.

- 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 tidyverse group_by with count. I recommend you try a few for practice.

```
ggplot(adult, aes(`marital-status`)) + geom_bar()
```




```
ggplot(adult, aes(relationship)) + geom_bar()
```




```
adult %>% group_by(sex) %>% select(sex, `marital-status`) %>% table()
```

```
##         marital-status
## sex      Divorced Married-AF-spouse Married-civ-spouse Married-spouse-absent
##   Female     2672                14               1657                   205
##   Male       1771                 9              13319                   213
##         marital-status
## sex      Never-married Separated Widowed
##   Female          4767       631     825
##   Male            5916       394     168
```

```
adult %>% group_by(`native-country`) %>% summarise(count = n()) %>% head()
```

```
## # A tibble: 6 x 2
##   `native-country` count
##   <chr>            <int>
## 1 ?                  583
## 2 Cambodia            19
## 3 Canada             121
## 4 China               75
## 5 Columbia            59
## 6 Cuba                95
```

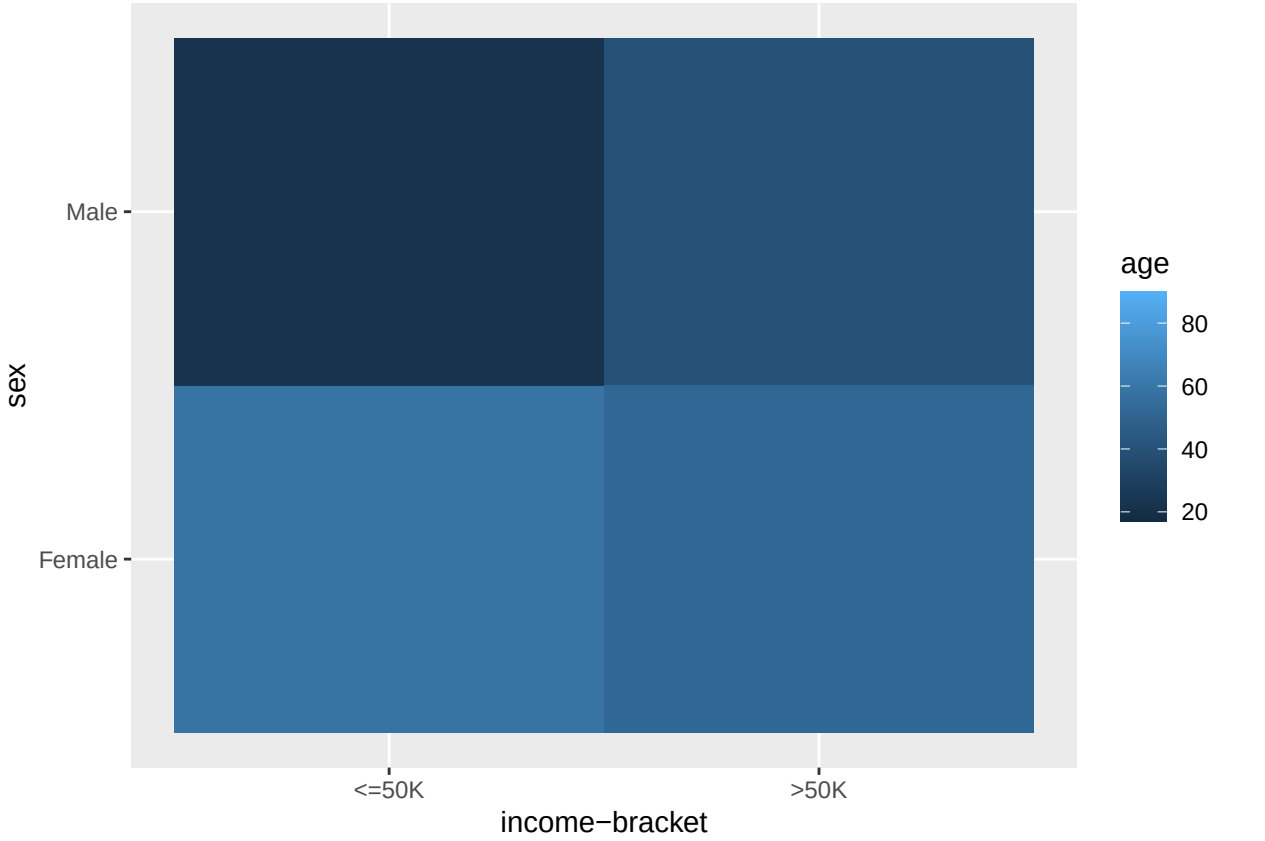
```
adult %>% group_by(`income-bracket`) %>% summarise(count = n())
```

```
## # A tibble: 2 x 2
##   `income-bracket` count
##   <chr>            <int>
## 1 <=50K            24720
```

```
## 2 >50K                7841
```

- e: 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 categorical.

```
ggplot(adult, aes(x=`income-bracket`, y=sex, fill = age)) + geom_tile()
```




```
adult %>% select(sex, `income-bracket`) %>% table()
```

```
##         income-bracket
## sex      <=50K  >50K
##   Female  9592  1179
##   Male   15128  6662
```

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.)

```
library(dplyr)

evens <- read_csv("/Users/danielkim/Downloads/FUNDAMENTALS OF DATA SCIENCE - 9232024 - 133 PM/populatio
```

```
## Rows: 52 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): NAME
## dbl (6): STATE, POPESTIMATE2010, POPESTIMATE2012, POPESTIMATE2014, POPESTIMA...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(evens)
```

```
## # A tibble: 6 x 7
##   STATE NAME     POPESTIMATE2010 POPESTIMATE2012 POPESTIMATE2014 POPESTIMATE2016
##   <dbl> <chr>              <dbl>           <dbl>           <dbl>           <dbl>
## 1     1 Alabama          4785437         4815588         4841799         4863525
## 2     2 Alaska            713910          730443          736283          741456
## 3     4 Arizona          6407172         6554978         6730413         6941072
## 4     5 Arkansas         2921964         2952164         2967392         2989918
## 5     6 Califor~        37319502        37948800        38596972        39167117
## 6     8 Colorado         5047349         5192647         5350101         5539215
## # i 1 more variable: POPESTIMATE2018 <dbl>
```

```
odds <- read_csv("/Users/danielkim/Downloads/FUNDAMENTALS OF DATA SCIENCE - 9232024 - 133 PM/population_
```

```
## Rows: 52 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): NAME
## dbl (6): STATE, POPESTIMATE2011, POPESTIMATE2013, POPESTIMATE2015, POPESTIMA...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(odds)
```

```
## # A tibble: 6 x 7
##   STATE NAME     POPESTIMATE2011 POPESTIMATE2013 POPESTIMATE2015 POPESTIMATE2017
##   <dbl> <chr>              <dbl>           <dbl>           <dbl>           <dbl>
## 1     1 Alabama          4799069         4830081         4852347         4874486
## 2     2 Alaska            722128          737068          737498          739700
## 3     4 Arizona               NA         6632764         6829676         7044008
## 4     5 Arkansas         2940667         2959400         2978048         3001345
## 5     6 Califor~        37638369        38260787        38918045        39358497
## 6     8 Colorado         5121108         5269035         5450623         5611885
## # i 1 more variable: POPESTIMATE2019 <dbl>
```

```
pop <- evens %>% inner_join(odds, by='STATE')
head(pop)
```

```
## # A tibble: 6 x 13
##   STATE NAME.x   POPESTIMATE2010 POPESTIMATE2012 POPESTIMATE2014 POPESTIMATE2016
##   <dbl> <chr>              <dbl>           <dbl>           <dbl>           <dbl>
## 1     1 Alabama          4785437         4815588         4841799         4863525
## 2     2 Alaska            713910          730443          736283          741456
## 3     4 Arizona          6407172         6554978         6730413         6941072
## 4     5 Arkansas         2921964         2952164         2967392         2989918
## 5     6 Califor~        37319502        37948800        38596972        39167117
## 6     8 Colorado         5047349         5192647         5350101         5539215
## # i 7 more variables: POPESTIMATE2018 <dbl>, NAME.y <chr>,
## #   POPESTIMATE2011 <dbl>, POPESTIMATE2013 <dbl>, POPESTIMATE2015 <dbl>,
## #   POPESTIMATE2017 <dbl>, POPESTIMATE2019 <dbl>
```

- 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.

```
pop <- pop %>% select(-c("NAME.y"))
head(pop)
```

```
## # A tibble: 6 x 12
##   STATE NAME.x   POPESTIMATE2010 POPESTIMATE2012 POPESTIMATE2014 POPESTIMATE2016
##   <dbl> <chr>              <dbl>           <dbl>           <dbl>           <dbl>
## 1     1 Alabama          4785437         4815588         4841799         4863525
## 2     2 Alaska            713910          730443          736283          741456
## 3     4 Arizona          6407172         6554978         6730413         6941072
## 4     5 Arkansas         2921964         2952164         2967392         2989918
## 5     6 Califor~        37319502        37948800        38596972        39167117
## 6     8 Colorado         5047349         5192647         5350101         5539215
## # i 6 more variables: POPESTIMATE2018 <dbl>, POPESTIMATE2011 <dbl>,
## #   POPESTIMATE2013 <dbl>, POPESTIMATE2015 <dbl>, POPESTIMATE2017 <dbl>,
## #   POPESTIMATE2019 <dbl>
```

  - b: Rename columns to be just the year number.

```
pop <- pop %>% setNames(c("STATE", "NAME", 2010, 2012, 2014, 2016, 2018, 2011, 2013, 2015, 2017, 20
head(pop)
```

```
## # A tibble: 6 x 12
##   STATE NAME     `2010` `2012` `2014` `2016` `2018`  `2011` `2013` `2015` `2017`
##   <dbl> <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>  <dbl>  <dbl>  <dbl>
## 1     1 Alabama  4.79e6 4.82e6 4.84e6 4.86e6 4.89e6  4.80e6 4.83e6 4.85e6 4.87e6
## 2     2 Alaska   7.14e5 7.30e5 7.36e5 7.41e5 7.35e5  7.22e5 7.37e5 7.37e5 7.40e5
## 3     4 Arizona  6.41e6 6.55e6 6.73e6 6.94e6 7.16e6 NA      6.63e6 6.83e6 7.04e6
## 4     5 Arkansas 2.92e6 2.95e6 2.97e6 2.99e6 3.01e6  2.94e6 2.96e6 2.98e6 3.00e6
## 5     6 Califor~ 3.73e7 3.79e7 3.86e7 3.92e7 3.95e7  3.76e7 3.83e7 3.89e7 3.94e7
## 6     8 Colorado 5.05e6 5.19e6 5.35e6 5.54e6 5.69e6  5.12e6 5.27e6 5.45e6 5.61e6
## # i 1 more variable: `2019` <dbl>
```

  - c: Reorder the columns to be in year order.

```
pop <- pop %>% select(c("STATE", "NAME", "2010", "2011", "2012", "2013", "2014", "2015", "2016", "2
head(pop)
```

```
## # A tibble: 6 x 12
##   STATE NAME     `2010`  `2011` `2012` `2013` `2014` `2015` `2016` `2017` `2018`
##   <dbl> <chr>     <dbl>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1     1 Alabama  4.79e6  4.80e6 4.82e6 4.83e6 4.84e6 4.85e6 4.86e6 4.87e6 4.89e6
## 2     2 Alaska   7.14e5  7.22e5 7.30e5 7.37e5 7.36e5 7.37e5 7.41e5 7.40e5 7.35e5
## 3     4 Arizona  6.41e6 NA      6.55e6 6.63e6 6.73e6 6.83e6 6.94e6 7.04e6 7.16e6
## 4     5 Arkansas 2.92e6  2.94e6 2.95e6 2.96e6 2.97e6 2.98e6 2.99e6 3.00e6 3.01e6
## 5     6 Califor~ 3.73e7  3.76e7 3.79e7 3.83e7 3.86e7 3.89e7 3.92e7 3.94e7 3.95e7
## 6     8 Colorado 5.05e6  5.12e6 5.19e6 5.27e6 5.35e6 5.45e6 5.54e6 5.61e6 5.69e6
## # i 1 more variable: `2019` <dbl>
```

- 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.

```
pop2 <- pop %>% select(-c("STATE")) %>% column_to_rownames(var = "NAME")
head(pop2)
```

```
##                2010      2011      2012      2013      2014      2015      2016
```

```
## Alabama     4785437  4799069  4815588  4830081  4841799  4852347  4863525
## Alaska       713910   722128   730443   737068   736283   737498   741456
## Arizona     6407172       NA  6554978  6632764  6730413  6829676  6941072
## Arkansas    2921964  2940667  2952164  2959400  2967392  2978048  2989918
## California 37319502 37638369 37948800 38260787 38596972 38918045 39167117
## Colorado    5047349  5121108  5192647  5269035  5350101  5450623  5539215
##                2017     2018     2019
## Alabama     4874486  4887681  4903185
## Alaska       739700   735139   731545
## Arizona     7044008  7158024  7278717
## Arkansas    3001345  3009733  3017804
## California 39358497 39461588 39512223
## Colorado    5611885  5691287  5758736
```

```
transpose_pop2 <- as.data.frame(t(pop2))
transpose_pop2
```

```
##      Alabama Alaska Arizona Arkansas California Colorado Connecticut Delaware
## 2010 4785437 713910 6407172  2921964   37319502  5047349     3579114   899593
## 2011 4799069 722128      NA  2940667   37638369  5121108     3588283   907381
## 2012 4815588 730443 6554978  2952164   37948800  5192647     3594547   915179
## 2013 4830081 737068 6632764  2959400   38260787  5269035     3594841   923576
## 2014 4841799 736283 6730413  2967392   38596972  5350101     3594524   932487
## 2015 4852347 737498 6829676  2978048   38918045  5450623     3587122   941252
## 2016 4863525 741456 6941072  2989918   39167117  5539215     3578141   948921
## 2017 4874486 739700 7044008  3001345   39358497  5611885     3573297   956823
## 2018 4887681 735139 7158024  3009733   39461588  5691287     3571520   965479
## 2019 4903185 731545 7278717  3017804   39512223  5758736     3565287   973764
##      District of Columbia  Florida  Georgia  Hawaii   Idaho Illinois Indiana
## 2010               605226 18845537  9711881 1363963 1570746 12840503 6490432
## 2011               619800 19053237  9802431 1379329 1583910 12867454 6516528
## 2012               634924 19297822  9901430 1394804 1595324 12882510 6537703
## 2013               650581 19545621  9972479 1408243 1611206 12895129 6568713
## 2014               662328 19845911 10067278 1414538 1631112 12884493 6593644
## 2015               675400 20209042 10178447 1422052      NA 12858913 6608422
## 2016               685815 20613477 10301890 1427559 1682380 12820527 6634304
## 2017               694906 20963613 10410330 1424393 1717715 12778828 6658078
## 2018               701547 21244317 10511131 1420593 1750536 12723071 6695497
## 2019               705749 21477737 10617423 1415872 1787065 12671821 6732219
##         Iowa  Kansas Kentucky Louisiana   Maine Maryland Massachusetts Michigan
## 2010 3050745 2858190  4348181   4544532 1327629  5788645       6566307  9877510
## 2011 3066336 2869225  4369821   4575625 1328284  5839419       6613583  9882412
## 2012 3076190 2885257  4386346   4600972 1327729  5886992       6663005  9897145
## 2013 3092997 2893212  4404659   4624527 1328009  5923188       6713315  9913065
## 2014 3109350 2900475  4414349   4644013 1330513  5957283       6762596  9929848
## 2015 3120960 2909011  4425976   4664628 1328262  5985562       6794228  9931715
## 2016 3131371 2910844  4438182   4678135 1331317  6003323       6823608  9950571
## 2017 3141550 2908718  4452268   4670560 1334612  6023868       6859789  9973114
## 2018 3148618 2911359  4461153   4659690 1339057  6035802       6882635  9984072
## 2019 3155070 2913314  4467673   4648794 1344212  6045680       6892503  9986857
##      Minnesota Mississippi Missouri Montana Nebraska  Nevada New Hampshire
## 2010   5310828     2970548  5995974  990697  1829542 2702405       1316762
## 2011   5346143     2978731  6010275  997316  1840672 2712730       1320202
## 2012   5376643     2983816  6024367 1003783  1853303 2743996       1324232
## 2013   5413479     2988711  6040715 1013569  1865279 2775970       1326622
```

```
## 2014   5451079     2990468  6056202 1021869  1879321 2817628       1333341
## 2015   5482032     2988471  6071732 1030475  1891277 2866939       1336350
## 2016   5522744     2987938  6087135 1040859  1905616 2917563       1342307
## 2017   5566230     2988510  6106670      NA  1915947 2969905       1348787
## 2018   5606249     2981020  6121623 1060665  1925614 3027341       1353465
## 2019   5639632     2976149  6137428 1068778  1934408 3080156       1359711
##      New Jersey New Mexico New York North Carolina North Dakota     Ohio
## 2010    8799446    2064552 19399878        9574323       674715 11539336
## 2011    8828117    2080450 19499241        9657592       685225 11544663
## 2012    8844942    2087309 19572932        9749476       701176 11548923
## 2013    8856972    2092273 19624447        9843336       722036       NA
## 2014    8864525    2089568 19651049        9932887       737401 11602700
## 2015    8867949    2089291 19654666       10031646       754066 11617527
## 2016    8870827    2091630 19633428       10154788       754434 11634370
## 2017    8885525    2091784 19589572       10268233       754942 11659650
## 2018    8886025    2092741 19530351       10381615       758080 11676341
## 2019    8882190    2096829 19453561       10488084       762062 11689100
##      Oklahoma  Oregon Pennsylvania Rhode Island South Carolina South Dakota
## 2010  3759944 3837491     12711160      1053959        4635649       816166
## 2011  3788379 3872036     12745815      1053649        4671994       823579
## 2012  3818814 3899001     12767118      1054621        4717354       833566
## 2013  3853214 3922468     12776309      1055081        4764080       842316
## 2014  3878187 3963244     12788313      1055936        4823617       849129
## 2015  3909500 4015792     12784826      1056065        4891938       853988
## 2016  3926331 4089976     12782275      1056770        4957968       862996
## 2017  3931316 4143625     12787641      1055673        5021268       872868
## 2018  3940235 4181886     12800922      1058287        5084156       878698
## 2019  3956971 4217737     12801989      1059361        5148714       884659
##      Tennessee    Texas    Utah Vermont Virginia Washington West Virginia
## 2010   6355311 25241971 2775332  625879  8023699    6742830       1854239
## 2011   6399291 25645629 2814384  627049  8101155    6826627       1856301
## 2012   6453898 26084481 2853375  626090  8185080    6897058       1856872
## 2013   6494340 26480266 2897640  626210  8252427    6963985       1853914
## 2014   6541223 26964333 2936879  625214  8310993    7054655       1849489
## 2015   6591170 27470056 2981835  625216  8361808    7163657       1842050
## 2016   6646010 27914410 3041868  623657  8410106    7294771       1831023
## 2017   6708799 28295273 3101042  624344  8463587    7423362       1817004
## 2018   6771631 28628666 3153550  624358  8501286    7523869       1804291
## 2019   6829174 28995881 3205958  623989  8535519    7614893       1792147
##      Wisconsin Wyoming Puerto Rico
## 2010   5690475  564487     3721525
## 2011   5705288  567299     3678732
## 2012   5719960  576305     3634488
## 2013   5736754  582122     3593077
## 2014   5751525  582531     3534874
## 2015   5760940  585613     3473232
## 2016   5772628  584215     3406672
## 2017   5790186  578931     3325286
## 2018   5807406  577601     3193354
## 2019        NA  578759     3193694
```

```
summary(transpose_pop2) #used to id columns with null values
```

```
##     Alabama            Alaska           Arizona           Arkansas
##  Min.   :4785437   Min.   :713910   Min.   :6407172   Min.   :2921964
```

```
##  1st Qu.:4819211   1st Qu.:730718   1st Qu.:6632764   1st Qu.:2953973
##  Median :4847073   Median :735711   Median :6829676   Median :2972720
##  Mean   :4845320   Mean   :732517   Mean   :6841869   Mean   :2973844
##  3rd Qu.:4871746   3rd Qu.:737390   3rd Qu.:7044008   3rd Qu.:2998488
##  Max.   :4903185   Max.   :741456   Max.   :7278717   Max.   :3017804
##                                     NA's   :1
##    California          Colorado        Connecticut         Delaware
##  Min.   :37319502   Min.   :5047349   Min.   :3565287   Min.   :899593
##  1st Qu.:38026797   1st Qu.:5211744   1st Qu.:3574508   1st Qu.:917278
##  Median :38757508   Median :5400362   Median :3583118   Median :936870
##  Mean   :38618190   Mean   :5403199   Mean   :3582668   Mean   :936446
##  3rd Qu.:39310652   3rd Qu.:5593718   3rd Qu.:3592964   3rd Qu.:954848
##  Max.   :39512223   Max.   :5758736   Max.   :3594841   Max.   :973764
##
##  District of Columbia    Florida            Georgia             Hawaii
##  Min.   :605226       Min.   :18845537   Min.   : 9711881   Min.   :1363963
##  1st Qu.:638838       1st Qu.:19359772   1st Qu.: 9919192   1st Qu.:1398164
##  Median :668864       Median :20027476   Median :10122862   Median :1415205
##  Mean   :663628       Mean   :20109631   Mean   :10147472   Mean   :1407135
##  3rd Qu.:692633       3rd Qu.:20876079   3rd Qu.:10383220   3rd Qu.:1421687
##  Max.   :705749       Max.   :21477737   Max.   :10617423   Max.   :1427559
##
##      Idaho            Illinois           Indiana             Iowa
##  Min.   :1570746   Min.   :12671821   Min.   :6490432   Min.   :3050745
##  1st Qu.:1595324   1st Qu.:12789253   1st Qu.:6545456   1st Qu.:3080392
##  Median :1631112   Median :12849708   Median :6601033   Median :3115155
##  Mean   :1658888   Mean   :12822325   Mean   :6603554   Mean   :3109319
##  3rd Qu.:1717715   3rd Qu.:12878746   3rd Qu.:6652134   3rd Qu.:3139005
##  Max.   :1787065   Max.   :12895129   Max.   :6732219   Max.   :3155070
##  NA's   :1
##      Kansas           Kentucky         Louisiana           Maine
##  Min.   :2858190   Min.   :4348181   Min.   :4544532   Min.   :1327629
##  1st Qu.:2887246   1st Qu.:4390924   1st Qu.:4606861   1st Qu.:1328072
##  Median :2904596   Median :4420162   Median :4646404   Median :1329398
##  Mean   :2895960   Mean   :4416861   Mean   :4631148   Mean   :1331962
##  3rd Qu.:2910386   3rd Qu.:4448746   3rd Qu.:4663394   3rd Qu.:1333788
##  Max.   :2913314   Max.   :4467673   Max.   :4678135   Max.   :1344212
##
##     Maryland       Massachusetts        Michigan         Minnesota
##  Min.   :5788645   Min.   :6566307   Min.   :9877510   Min.   :5310828
##  1st Qu.:5896041   1st Qu.:6675582   1st Qu.:9901125   1st Qu.:5385852
##  Median :5971422   Median :6778412   Median :9930782   Median :5466556
##  Mean   :5948976   Mean   :6757157   Mean   :9932631   Mean   :5471506
##  3rd Qu.:6018732   3rd Qu.:6850744   3rd Qu.:9967478   3rd Qu.:5555358
##  Max.   :6045680   Max.   :6892503   Max.   :9986857   Max.   :5639632
##
##   Mississippi         Missouri          Montana           Nebraska
##  Min.   :2970548   Min.   :5995974   Min.   : 990697   Min.   :1829542
##  1st Qu.:2979303   1st Qu.:6028454   1st Qu.:1003783   1st Qu.:1856297
##  Median :2985877   Median :6063967   Median :1021869   Median :1885299
##  Mean   :2983436   Mean   :6065212   Mean   :1025335   Mean   :1884098
##  3rd Qu.:2988500   3rd Qu.:6101786   3rd Qu.:1040859   3rd Qu.:1913364
##  Max.   :2990468   Max.   :6137428   Max.   :1068778   Max.   :1934408
##                                      NA's   :1
```

```
##      Nevada        New Hampshire       New Jersey        New Mexico
##  Min.   :2702405   Min.   :1316762   Min.   :8799446   Min.   :2064552
##  1st Qu.:2751990   1st Qu.:1324830   1st Qu.:8847950   1st Qu.:2087804
##  Median :2842284   Median :1334846   Median :8866237   Median :2090599
##  Mean   :2861463   Mean   :1336178   Mean   :8858652   Mean   :2087643
##  3rd Qu.:2956820   3rd Qu.:1347167   3rd Qu.:8879349   3rd Qu.:2092151
##  Max.   :3080156   Max.   :1359711   Max.   :8886025   Max.   :2096829
##
##     New York        North Carolina      North Dakota         Ohio
##  Min.   :19399878   Min.   : 9574323   Min.   :674715   Min.   :11539336
##  1st Qu.:19507018   1st Qu.: 9772941   1st Qu.:706391   1st Qu.:11548923
##  Median :19581252   Median : 9982266   Median :745734   Median :11617527
##  Mean   :19560912   Mean   :10008198   Mean   :730414   Mean   :11612512
##  3rd Qu.:19631183   3rd Qu.:10239872   3rd Qu.:754815   3rd Qu.:11659650
##  Max.   :19654666   Max.   :10488084   Max.   :762062   Max.   :11689100
##                                                         NA's   :1
##     Oklahoma           Oregon         Pennsylvania       Rhode Island
##  Min.   :3759944   Min.   :3837491   Min.   :12711160   Min.   :1053649
##  1st Qu.:3827414   1st Qu.:3904868   1st Qu.:12769416   1st Qu.:1054736
##  Median :3893844   Median :3989518   Median :12783550   Median :1055804
##  Mean   :3876289   Mean   :4014326   Mean   :12774637   Mean   :1055940
##  3rd Qu.:3930070   3rd Qu.:4130213   3rd Qu.:12788145   3rd Qu.:1056594
##  Max.   :3956971   Max.   :4217737   Max.   :12801989   Max.   :1059361
##
##  South Carolina     South Dakota      Tennessee           Texas
##  Min.   :4635649   Min.   :816166   Min.   :6355311   Min.   :25241971
##  1st Qu.:4729036   1st Qu.:835754   1st Qu.:6464008   1st Qu.:26183427
##  Median :4857778   Median :851558   Median :6566196   Median :27217194
##  Mean   :4871674   Mean   :851796   Mean   :6579085   Mean   :27172097
##  3rd Qu.:5005443   3rd Qu.:870400   3rd Qu.:6693102   3rd Qu.:28200057
##  Max.   :5148714   Max.   :884659   Max.   :6829174   Max.   :28995881
##
##       Utah            Vermont          Virginia         Washington
##  Min.   :2775332   Min.   :623657   Min.   :8023699   Min.   :6742830
##  1st Qu.:2864441   1st Qu.:624348   1st Qu.:8201917   1st Qu.:6913790
##  Median :2959357   Median :625215   Median :8336400   Median :7109156
##  Mean   :2976186   Mean   :625201   Mean   :8314566   Mean   :7150571
##  3rd Qu.:3086248   3rd Qu.:626037   3rd Qu.:8450217   3rd Qu.:7391214
##  Max.   :3205958   Max.   :627049   Max.   :8535519   Max.   :7614893
##
##  West Virginia       Wisconsin          Wyoming        Puerto Rico
##  Min.   :1792147   Min.   :5690475   Min.   :564487   Min.   :3193354
##  1st Qu.:1820509   1st Qu.:5719960   1st Qu.:576629   1st Qu.:3345632
##  Median :1845770   Median :5751525   Median :578845   Median :3504053
##  Mean   :1835733   Mean   :5748351   Mean   :577786   Mean   :3475493
##  3rd Qu.:1854158   3rd Qu.:5772628   3rd Qu.:582429   3rd Qu.:3624135
##  Max.   :1856872   Max.   :5807406   Max.   :585613   Max.   :3721525
##                    NA's   :1
```

```
which(is.na(transpose_pop2)) #id location of null values
```

```
## [1]  22 126 268 354 500
```

```
#fix null in Arizona
az_na <- transpose_pop2 %>% select(Arizona)
```

```
az_na
```

```
##      Arizona
## 2010 6407172
## 2011      NA
## 2012 6554978
## 2013 6632764
## 2014 6730413
## 2015 6829676
## 2016 6941072
## 2017 7044008
## 2018 7158024
## 2019 7278717
```

```
az_val <- sum(az_na[c("2010", "2012"),])/2
az_val
```

```
## [1] 6481075
```

```
transpose_pop2$Arizona <- transpose_pop2$Arizona %>% replace_na(az_val)

#fix null in Idaho
id_na <- transpose_pop2 %>% select(Idaho)
id_na
```

```
##        Idaho
## 2010 1570746
## 2011 1583910
## 2012 1595324
## 2013 1611206
## 2014 1631112
## 2015      NA
## 2016 1682380
## 2017 1717715
## 2018 1750536
## 2019 1787065
```

```
id_val <- sum(id_na[c("2014", "2016"),])/2
id_val
```

```
## [1] 1656746
```

```
transpose_pop2$Idaho <- transpose_pop2$Idaho %>% replace_na(id_val)

#fix null in Montana
mt_na <- transpose_pop2 %>% select(Montana)
mt_na
```

```
##      Montana
## 2010  990697
## 2011  997316
## 2012 1003783
## 2013 1013569
## 2014 1021869
## 2015 1030475
## 2016 1040859
## 2017      NA
```

```
## 2018 1060665
## 2019 1068778
```

```
mt_val <- sum(mt_na[c("2016", "2018"),])/2
transpose_pop2$Montana <- transpose_pop2$Montana %>% replace_na(mt_val)

#fix null in Ohio
oh_na <- transpose_pop2 %>% select(Ohio)
oh_na
```

```
##          Ohio
## 2010 11539336
## 2011 11544663
## 2012 11548923
## 2013       NA
## 2014 11602700
## 2015 11617527
## 2016 11634370
## 2017 11659650
## 2018 11676341
## 2019 11689100
```

```
oh_val <- sum(oh_na[c("2012", "2014"),])/2
oh_val
```

```
## [1] 11575812
```

```
transpose_pop2$Ohio <- transpose_pop2$Ohio %>% replace_na(oh_val)

#fix null in Wisconsin, replace null with mean of Wisconsin pop since no 2020 data available
wi_na <- transpose_pop2 %>% select(Wisconsin)
wi_na
```

```
##      Wisconsin
## 2010   5690475
## 2011   5705288
## 2012   5719960
## 2013   5736754
## 2014   5751525
## 2015   5760940
## 2016   5772628
## 2017   5790186
## 2018   5807406
## 2019        NA
```

```
transpose_pop2$Wisconsin <- transpose_pop2$Wisconsin %>% replace_na(mean(transpose_pop2$Wisconsin, na.r

summary(transpose_pop2)
```

```
##     Alabama            Alaska           Arizona           Arkansas
##  Min.   :4785437   Min.   :713910   Min.   :6407172   Min.   :2921964
##  1st Qu.:4819211   1st Qu.:730718   1st Qu.:6574424   1st Qu.:2953973
##  Median :4847073   Median :735711   Median :6780044   Median :2972720
##  Mean   :4845320   Mean   :732517   Mean   :6805790   Mean   :2973844
##  3rd Qu.:4871746   3rd Qu.:737390   3rd Qu.:7018274   3rd Qu.:2998488
##  Max.   :4903185   Max.   :741456   Max.   :7278717   Max.   :3017804
##    California          Colorado        Connecticut         Delaware
```

```
##  Min.   :37319502   Min.   :5047349   Min.   :3565287   Min.   :899593
##  1st Qu.:38026797   1st Qu.:5211744   1st Qu.:3574508   1st Qu.:917278
##  Median :38757508   Median :5400362   Median :3583118   Median :936870
##  Mean   :38618190   Mean   :5403199   Mean   :3582668   Mean   :936446
##  3rd Qu.:39310652   3rd Qu.:5593718   3rd Qu.:3592964   3rd Qu.:954848
##  Max.   :39512223   Max.   :5758736   Max.   :3594841   Max.   :973764
##  District of Columbia    Florida            Georgia             Hawaii
##  Min.   :605226       Min.   :18845537   Min.   : 9711881   Min.   :1363963
##  1st Qu.:638838       1st Qu.:19359772   1st Qu.: 9919192   1st Qu.:1398164
##  Median :668864       Median :20027476   Median :10122862   Median :1415205
##  Mean   :663628       Mean   :20109631   Mean   :10147472   Mean   :1407135
##  3rd Qu.:692633       3rd Qu.:20876079   3rd Qu.:10383220   3rd Qu.:1421687
##  Max.   :705749       Max.   :21477737   Max.   :10617423   Max.   :1427559
##      Idaho            Illinois           Indiana             Iowa
##  Min.   :1570746   Min.   :12671821   Min.   :6490432   Min.   :3050745
##  1st Qu.:1599294   1st Qu.:12789253   1st Qu.:6545456   1st Qu.:3080392
##  Median :1643929   Median :12849708   Median :6601033   Median :3115155
##  Mean   :1658674   Mean   :12822325   Mean   :6603554   Mean   :3109319
##  3rd Qu.:1708881   3rd Qu.:12878746   3rd Qu.:6652134   3rd Qu.:3139005
##  Max.   :1787065   Max.   :12895129   Max.   :6732219   Max.   :3155070
##      Kansas           Kentucky         Louisiana           Maine
##  Min.   :2858190   Min.   :4348181   Min.   :4544532   Min.   :1327629
##  1st Qu.:2887246   1st Qu.:4390924   1st Qu.:4606861   1st Qu.:1328072
##  Median :2904596   Median :4420162   Median :4646404   Median :1329398
##  Mean   :2895960   Mean   :4416861   Mean   :4631148   Mean   :1331962
##  3rd Qu.:2910386   3rd Qu.:4448746   3rd Qu.:4663394   3rd Qu.:1333788
##  Max.   :2913314   Max.   :4467673   Max.   :4678135   Max.   :1344212
##     Maryland       Massachusetts        Michigan         Minnesota
##  Min.   :5788645   Min.   :6566307   Min.   :9877510   Min.   :5310828
##  1st Qu.:5896041   1st Qu.:6675582   1st Qu.:9901125   1st Qu.:5385852
##  Median :5971422   Median :6778412   Median :9930782   Median :5466556
##  Mean   :5948976   Mean   :6757157   Mean   :9932631   Mean   :5471506
##  3rd Qu.:6018732   3rd Qu.:6850744   3rd Qu.:9967478   3rd Qu.:5555358
##  Max.   :6045680   Max.   :6892503   Max.   :9986857   Max.   :5639632
##   Mississippi         Missouri          Montana           Nebraska
##  Min.   :2970548   Min.   :5995974   Min.   : 990697   Min.   :1829542
##  1st Qu.:2979303   1st Qu.:6028454   1st Qu.:1006230   1st Qu.:1856297
##  Median :2985877   Median :6063967   Median :1026172   Median :1885299
##  Mean   :2983436   Mean   :6065212   Mean   :1027877   Mean   :1884098
##  3rd Qu.:2988500   3rd Qu.:6101786   3rd Qu.:1048286   3rd Qu.:1913364
##  Max.   :2990468   Max.   :6137428   Max.   :1068778   Max.   :1934408
##      Nevada        New Hampshire       New Jersey        New Mexico
##  Min.   :2702405   Min.   :1316762   Min.   :8799446   Min.   :2064552
##  1st Qu.:2751990   1st Qu.:1324830   1st Qu.:8847950   1st Qu.:2087804
##  Median :2842284   Median :1334846   Median :8866237   Median :2090599
##  Mean   :2861463   Mean   :1336178   Mean   :8858652   Mean   :2087643
##  3rd Qu.:2956820   3rd Qu.:1347167   3rd Qu.:8879349   3rd Qu.:2092151
##  Max.   :3080156   Max.   :1359711   Max.   :8886025   Max.   :2096829
##     New York        North Carolina      North Dakota         Ohio
##  Min.   :19399878   Min.   : 9574323   Min.   :674715   Min.   :11539336
##  1st Qu.:19507018   1st Qu.: 9772941   1st Qu.:706391   1st Qu.:11555645
##  Median :19581252   Median : 9982266   Median :745734   Median :11610114
##  Mean   :19560912   Mean   :10008198   Mean   :730414   Mean   :11608842
##  3rd Qu.:19631183   3rd Qu.:10239872   3rd Qu.:754815   3rd Qu.:11653330
```

```
##  Max.   :19654666   Max.   :10488084   Max.   :762062   Max.   :11689100
##     Oklahoma           Oregon         Pennsylvania       Rhode Island
##  Min.   :3759944   Min.   :3837491   Min.   :12711160   Min.   :1053649
##  1st Qu.:3827414   1st Qu.:3904868   1st Qu.:12769416   1st Qu.:1054736
##  Median :3893844   Median :3989518   Median :12783550   Median :1055804
##  Mean   :3876289   Mean   :4014326   Mean   :12774637   Mean   :1055940
##  3rd Qu.:3930070   3rd Qu.:4130213   3rd Qu.:12788145   3rd Qu.:1056594
##  Max.   :3956971   Max.   :4217737   Max.   :12801989   Max.   :1059361
##  South Carolina     South Dakota      Tennessee           Texas
##  Min.   :4635649   Min.   :816166   Min.   :6355311   Min.   :25241971
##  1st Qu.:4729036   1st Qu.:835754   1st Qu.:6464008   1st Qu.:26183427
##  Median :4857778   Median :851558   Median :6566196   Median :27217194
##  Mean   :4871674   Mean   :851796   Mean   :6579085   Mean   :27172097
##  3rd Qu.:5005443   3rd Qu.:870400   3rd Qu.:6693102   3rd Qu.:28200057
##  Max.   :5148714   Max.   :884659   Max.   :6829174   Max.   :28995881
##       Utah            Vermont          Virginia         Washington
##  Min.   :2775332   Min.   :623657   Min.   :8023699   Min.   :6742830
##  1st Qu.:2864441   1st Qu.:624348   1st Qu.:8201917   1st Qu.:6913790
##  Median :2959357   Median :625215   Median :8336400   Median :7109156
##  Mean   :2976186   Mean   :625201   Mean   :8314566   Mean   :7150571
##  3rd Qu.:3086248   3rd Qu.:626037   3rd Qu.:8450217   3rd Qu.:7391214
##  Max.   :3205958   Max.   :627049   Max.   :8535519   Max.   :7614893
##  West Virginia       Wisconsin          Wyoming        Puerto Rico
##  Min.   :1792147   Min.   :5690475   Min.   :564487   Min.   :3193354
##  1st Qu.:1820509   1st Qu.:5724158   1st Qu.:576629   1st Qu.:3345632
##  Median :1845770   Median :5749938   Median :578845   Median :3504053
##  Mean   :1835733   Mean   :5748351   Mean   :577786   Mean   :3475493
##  3rd Qu.:1854158   3rd Qu.:5769706   3rd Qu.:582429   3rd Qu.:3624135
##  Max.   :1856872   Max.   :5807406   Max.   :585613   Max.   :3721525
```

```
sum(is.na(transpose_pop2))
```

```
## [1] 0
```

- d: We can use some tidyverse 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 rowwise() command in your tidyverse pipe. If you do not, the max value will not be individual to the row. Of course there are alternative ways.

```
transpose_pop <- as.data.frame(t(transpose_pop2))
transpose_pop
```

```
##                             2010     2011     2012     2013     2014     2015
## Alabama                  4785437  4799069  4815588  4830081  4841799  4852347
## Alaska                    713910   722128   730443   737068   736283   737498
## Arizona                  6407172  6481075  6554978  6632764  6730413  6829676
## Arkansas                 2921964  2940667  2952164  2959400  2967392  2978048
## California              37319502 37638369 37948800 38260787 38596972 38918045
## Colorado                 5047349  5121108  5192647  5269035  5350101  5450623
## Connecticut              3579114  3588283  3594547  3594841  3594524  3587122
## Delaware                  899593   907381   915179   923576   932487   941252
## District of Columbia      605226   619800   634924   650581   662328   675400
## Florida                 18845537 19053237 19297822 19545621 19845911 20209042
## Georgia                  9711881  9802431  9901430  9972479 10067278 10178447
```

```
## Hawaii                1363963  1379329  1394804  1408243  1414538  1422052
## Idaho                 1570746  1583910  1595324  1611206  1631112  1656746
## Illinois             12840503 12867454 12882510 12895129 12884493 12858913
## Indiana               6490432  6516528  6537703  6568713  6593644  6608422
## Iowa                  3050745  3066336  3076190  3092997  3109350  3120960
## Kansas                2858190  2869225  2885257  2893212  2900475  2909011
## Kentucky              4348181  4369821  4386346  4404659  4414349  4425976
## Louisiana             4544532  4575625  4600972  4624527  4644013  4664628
## Maine                 1327629  1328284  1327729  1328009  1330513  1328262
## Maryland              5788645  5839419  5886992  5923188  5957283  5985562
## Massachusetts         6566307  6613583  6663005  6713315  6762596  6794228
## Michigan              9877510  9882412  9897145  9913065  9929848  9931715
## Minnesota             5310828  5346143  5376643  5413479  5451079  5482032
## Mississippi           2970548  2978731  2983816  2988711  2990468  2988471
## Missouri              5995974  6010275  6024367  6040715  6056202  6071732
## Montana                990697   997316  1003783  1013569  1021869  1030475
## Nebraska              1829542  1840672  1853303  1865279  1879321  1891277
## Nevada                2702405  2712730  2743996  2775970  2817628  2866939
## New Hampshire         1316762  1320202  1324232  1326622  1333341  1336350
## New Jersey            8799446  8828117  8844942  8856972  8864525  8867949
## New Mexico            2064552  2080450  2087309  2092273  2089568  2089291
## New York             19399878 19499241 19572932 19624447 19651049 19654666
## North Carolina        9574323  9657592  9749476  9843336  9932887 10031646
## North Dakota           674715   685225   701176   722036   737401   754066
## Ohio                 11539336 11544663 11548923 11575812 11602700 11617527
## Oklahoma              3759944  3788379  3818814  3853214  3878187  3909500
## Oregon                3837491  3872036  3899001  3922468  3963244  4015792
## Pennsylvania         12711160 12745815 12767118 12776309 12788313 12784826
## Rhode Island          1053959  1053649  1054621  1055081  1055936  1056065
## South Carolina        4635649  4671994  4717354  4764080  4823617  4891938
## South Dakota           816166   823579   833566   842316   849129   853988
## Tennessee             6355311  6399291  6453898  6494340  6541223  6591170
## Texas                25241971 25645629 26084481 26480266 26964333 27470056
## Utah                  2775332  2814384  2853375  2897640  2936879  2981835
## Vermont                625879   627049   626090   626210   625214   625216
## Virginia              8023699  8101155  8185080  8252427  8310993  8361808
## Washington            6742830  6826627  6897058  6963985  7054655  7163657
## West Virginia         1854239  1856301  1856872  1853914  1849489  1842050
## Wisconsin             5690475  5705288  5719960  5736754  5751525  5760940
## Wyoming                564487   567299   576305   582122   582531   585613
## Puerto Rico           3721525  3678732  3634488  3593077  3534874  3473232
##                          2016     2017     2018     2019
## Alabama               4863525  4874486  4887681  4903185
## Alaska                 741456   739700   735139   731545
## Arizona               6941072  7044008  7158024  7278717
## Arkansas              2989918  3001345  3009733  3017804
## California           39167117 39358497 39461588 39512223
## Colorado              5539215  5611885  5691287  5758736
## Connecticut           3578141  3573297  3571520  3565287
## Delaware               948921   956823   965479   973764
## District of Columbia   685815   694906   701547   705749
## Florida              20613477 20963613 21244317 21477737
## Georgia              10301890 10410330 10511131 10617423
## Hawaii                1427559  1424393  1420593  1415872
```

```
## Idaho                1682380  1717715  1750536  1787065
## Illinois            12820527 12778828 12723071 12671821
## Indiana              6634304  6658078  6695497  6732219
## Iowa                 3131371  3141550  3148618  3155070
## Kansas               2910844  2908718  2911359  2913314
## Kentucky             4438182  4452268  4461153  4467673
## Louisiana            4678135  4670560  4659690  4648794
## Maine                1331317  1334612  1339057  1344212
## Maryland             6003323  6023868  6035802  6045680
## Massachusetts        6823608  6859789  6882635  6892503
## Michigan             9950571  9973114  9984072  9986857
## Minnesota            5522744  5566230  5606249  5639632
## Mississippi          2987938  2988510  2981020  2976149
## Missouri             6087135  6106670  6121623  6137428
## Montana              1040859  1050762  1060665  1068778
## Nebraska             1905616  1915947  1925614  1934408
## Nevada               2917563  2969905  3027341  3080156
## New Hampshire        1342307  1348787  1353465  1359711
## New Jersey           8870827  8885525  8886025  8882190
## New Mexico           2091630  2091784  2092741  2096829
## New York            19633428 19589572 19530351 19453561
## North Carolina      10154788 10268233 10381615 10488084
## North Dakota          754434   754942   758080   762062
## Ohio                11634370 11659650 11676341 11689100
## Oklahoma             3926331  3931316  3940235  3956971
## Oregon               4089976  4143625  4181886  4217737
## Pennsylvania        12782275 12787641 12800922 12801989
## Rhode Island         1056770  1055673  1058287  1059361
## South Carolina       4957968  5021268  5084156  5148714
## South Dakota          862996   872868   878698   884659
## Tennessee            6646010  6708799  6771631  6829174
## Texas               27914410 28295273 28628666 28995881
## Utah                 3041868  3101042  3153550  3205958
## Vermont               623657   624344   624358   623989
## Virginia             8410106  8463587  8501286  8535519
## Washington           7294771  7423362  7523869  7614893
## West Virginia        1831023  1817004  1804291  1792147
## Wisconsin            5772628  5790186  5807406  5748351
## Wyoming               584215   578931   577601   578759
## Puerto Rico          3406672  3325286  3193354  3193694
```

```
max_pop <- transpose_pop %>% apply(1, max)
max_pop
```

```
##              Alabama               Alaska              Arizona
##              4903185               741456              7278717
##             Arkansas           California             Colorado
##              3017804             39512223              5758736
##          Connecticut             Delaware District of Columbia
##              3594841              973764               705749
##              Florida              Georgia               Hawaii
##             21477737             10617423              1427559
##                Idaho             Illinois              Indiana
##              1787065             12895129              6732219
##                 Iowa               Kansas             Kentucky
```

```
##              3155070              2913314              4467673
##            Louisiana                Maine             Maryland
##              4678135              1344212              6045680
##        Massachusetts             Michigan            Minnesota
##              6892503              9986857              5639632
##          Mississippi             Missouri              Montana
##              2990468              6137428              1068778
##             Nebraska               Nevada        New Hampshire
##              1934408              3080156              1359711
##           New Jersey           New Mexico             New York
##              8886025              2096829             19654666
##       North Carolina         North Dakota                 Ohio
##             10488084               762062             11689100
##             Oklahoma               Oregon         Pennsylvania
##              3956971              4217737             12801989
##         Rhode Island       South Carolina         South Dakota
##              1059361              5148714               884659
##            Tennessee                Texas                 Utah
##              6829174             28995881              3205958
##              Vermont             Virginia           Washington
##               627049              8535519              7614893
##        West Virginia            Wisconsin              Wyoming
##              1856872              5807406               585613
##          Puerto Rico
##              3721525
```

– 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?

```
total_pop <- apply(transpose_pop, 1, sum)
total_pop
```

```
##              Alabama               Alaska              Arizona
##             48453198              7325170             68057899
##             Arkansas           California             Colorado
##             29738435            386181900             54031986
##          Connecticut             Delaware District of Columbia
##             35826676              9364455              6636276
##              Florida              Georgia               Hawaii
##            201096314            101474720             14071346
##                Idaho              Illinois              Indiana
##             16586740            128223249             66035540
##                 Iowa                Kansas              Kentucky
##             31093187             28959605             44168608
##            Louisiana                Maine             Maryland
##             46311476             13319624             59489762
##        Massachusetts             Michigan            Minnesota
##             67571569             99326309             54715059
##          Mississippi             Missouri              Montana
##             29834362             60652121             10278773
##             Nebraska               Nevada        New Hampshire
##             18840979             28614633             13361779
##           New Jersey           New Mexico             New York
##             88586518             20876427            195609125
##       North Carolina         North Dakota                 Ohio
```

```
##           100081980             7304137           116088422
##            Oklahoma              Oregon        Pennsylvania
##            38762891            40143256           127746368
##        Rhode Island      South Carolina        South Dakota
##            10559402            48716738             8517965
##           Tennessee               Texas                Utah
##            65790847           271720966            29761863
##             Vermont            Virginia          Washington
##             6252006            83145660            71505707
##       West Virginia           Wisconsin             Wyoming
##            18357330            57483513             5777863
##         Puerto Rico
##            34754934
```

The total population across all years for each state has similar code from (d) because the minor change between these two codes is a change in aggregation function. Rather than using the max function in (d), the code for the total population across all years for each state uses the sum aggregation funtion.

- 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 tidyverse, so keep it simple.

```
max_US_pop <- apply(transpose_pop, 2, sum)
max_US_pop
```

```
##      2010      2011      2012      2013      2014      2015      2016      2017
## 313043191 315244038 317465478 319585920 321835882 324114082 326347983 328309105
##      2018      2019
## 329880855 331359134
```

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)*.

```
transpose_pop
```

```
##                       2010     2011     2012     2013     2014     2015
## Alabama            4785437  4799069  4815588  4830081  4841799  4852347
## Alaska              713910   722128   730443   737068   736283   737498
## Arizona            6407172  6481075  6554978  6632764  6730413  6829676
## Arkansas           2921964  2940667  2952164  2959400  2967392  2978048
## California        37319502 37638369 37948800 38260787 38596972 38918045
## Colorado           5047349  5121108  5192647  5269035  5350101  5450623
## Connecticut        3579114  3588283  3594547  3594841  3594524  3587122
## Delaware            899593   907381   915179   923576   932487   941252
```

```
## District of Columbia   605226   619800   634924   650581   662328   675400
## Florida              18845537 19053237 19297822 19545621 19845911 20209042
## Georgia               9711881  9802431  9901430  9972479 10067278 10178447
## Hawaii                1363963  1379329  1394804  1408243  1414538  1422052
## Idaho                 1570746  1583910  1595324  1611206  1631112  1656746
## Illinois             12840503 12867454 12882510 12895129 12884493 12858913
## Indiana               6490432  6516528  6537703  6568713  6593644  6608422
## Iowa                  3050745  3066336  3076190  3092997  3109350  3120960
## Kansas                2858190  2869225  2885257  2893212  2900475  2909011
## Kentucky              4348181  4369821  4386346  4404659  4414349  4425976
## Louisiana             4544532  4575625  4600972  4624527  4644013  4664628
## Maine                 1327629  1328284  1327729  1328009  1330513  1328262
## Maryland              5788645  5839419  5886992  5923188  5957283  5985562
## Massachusetts         6566307  6613583  6663005  6713315  6762596  6794228
## Michigan              9877510  9882412  9897145  9913065  9929848  9931715
## Minnesota             5310828  5346143  5376643  5413479  5451079  5482032
## Mississippi           2970548  2978731  2983816  2988711  2990468  2988471
## Missouri              5995974  6010275  6024367  6040715  6056202  6071732
## Montana                990697   997316  1003783  1013569  1021869  1030475
## Nebraska              1829542  1840672  1853303  1865279  1879321  1891277
## Nevada                2702405  2712730  2743996  2775970  2817628  2866939
## New Hampshire         1316762  1320202  1324232  1326622  1333341  1336350
## New Jersey            8799446  8828117  8844942  8856972  8864525  8867949
## New Mexico            2064552  2080450  2087309  2092273  2089568  2089291
## New York             19399878 19499241 19572932 19624447 19651049 19654666
## North Carolina        9574323  9657592  9749476  9843336  9932887 10031646
## North Dakota           674715   685225   701176   722036   737401   754066
## Ohio                 11539336 11544663 11548923 11575812 11602700 11617527
## Oklahoma              3759944  3788379  3818814  3853214  3878187  3909500
## Oregon                3837491  3872036  3899001  3922468  3963244  4015792
## Pennsylvania         12711160 12745815 12767118 12776309 12788313 12784826
## Rhode Island          1053959  1053649  1054621  1055081  1055936  1056065
## South Carolina        4635649  4671994  4717354  4764080  4823617  4891938
## South Dakota           816166   823579   833566   842316   849129   853988
## Tennessee             6355311  6399291  6453898  6494340  6541223  6591170
## Texas                25241971 25645629 26084481 26480266 26964333 27470056
## Utah                  2775332  2814384  2853375  2897640  2936879  2981835
## Vermont                625879   627049   626090   626210   625214   625216
## Virginia              8023699  8101155  8185080  8252427  8310993  8361808
## Washington            6742830  6826627  6897058  6963985  7054655  7163657
## West Virginia         1854239  1856301  1856872  1853914  1849489  1842050
## Wisconsin             5690475  5705288  5719960  5736754  5751525  5760940
## Wyoming                564487   567299   576305   582122   582531   585613
## Puerto Rico           3721525  3678732  3634488  3593077  3534874  3473232
##                          2016     2017     2018     2019
## Alabama               4863525  4874486  4887681  4903185
## Alaska                 741456   739700   735139   731545
## Arizona               6941072  7044008  7158024  7278717
## Arkansas              2989918  3001345  3009733  3017804
## California           39167117 39358497 39461588 39512223
## Colorado              5539215  5611885  5691287  5758736
## Connecticut           3578141  3573297  3571520  3565287
## Delaware               948921   956823   965479   973764
## District of Columbia   685815   694906   701547   705749
```

```
## Florida              20613477 20963613 21244317 21477737
## Georgia              10301890 10410330 10511131 10617423
## Hawaii                1427559  1424393  1420593  1415872
## Idaho                 1682380  1717715  1750536  1787065
## Illinois             12820527 12778828 12723071 12671821
## Indiana               6634304  6658078  6695497  6732219
## Iowa                  3131371  3141550  3148618  3155070
## Kansas                2910844  2908718  2911359  2913314
## Kentucky              4438182  4452268  4461153  4467673
## Louisiana             4678135  4670560  4659690  4648794
## Maine                 1331317  1334612  1339057  1344212
## Maryland              6003323  6023868  6035802  6045680
## Massachusetts         6823608  6859789  6882635  6892503
## Michigan              9950571  9973114  9984072  9986857
## Minnesota             5522744  5566230  5606249  5639632
## Mississippi           2987938  2988510  2981020  2976149
## Missouri              6087135  6106670  6121623  6137428
## Montana               1040859  1050762  1060665  1068778
## Nebraska              1905616  1915947  1925614  1934408
## Nevada                2917563  2969905  3027341  3080156
## New Hampshire         1342307  1348787  1353465  1359711
## New Jersey            8870827  8885525  8886025  8882190
## New Mexico            2091630  2091784  2092741  2096829
## New York             19633428 19589572 19530351 19453561
## North Carolina       10154788 10268233 10381615 10488084
## North Dakota           754434   754942   758080   762062
## Ohio                 11634370 11659650 11676341 11689100
## Oklahoma              3926331  3931316  3940235  3956971
## Oregon                4089976  4143625  4181886  4217737
## Pennsylvania         12782275 12787641 12800922 12801989
## Rhode Island          1056770  1055673  1058287  1059361
## South Carolina        4957968  5021268  5084156  5148714
## South Dakota           862996   872868   878698   884659
## Tennessee             6646010  6708799  6771631  6829174
## Texas                27914410 28295273 28628666 28995881
## Utah                  3041868  3101042  3153550  3205958
## Vermont                623657   624344   624358   623989
## Virginia              8410106  8463587  8501286  8535519
## Washington            7294771  7423362  7523869  7614893
## West Virginia         1831023  1817004  1804291  1792147
## Wisconsin             5772628  5790186  5807406  5748351
## Wyoming                584215   578931   577601   578759
## Puerto Rico           3406672  3325286  3193354  3193694
```

```
pivot_long <- transpose_pop %>% rownames_to_column(var="State") %>% pivot_longer(cols = c("2010", "2011
pivot_long$Year <- as.factor(as.integer(pivot_long$Year))

pivot_long
```

```
## # A tibble: 520 x 3
##    State   Year  Population
##    <chr>   <fct>      <dbl>
##  1 Alabama 2010     4785437
##  2 Alabama 2011     4799069
##  3 Alabama 2012     4815588
```

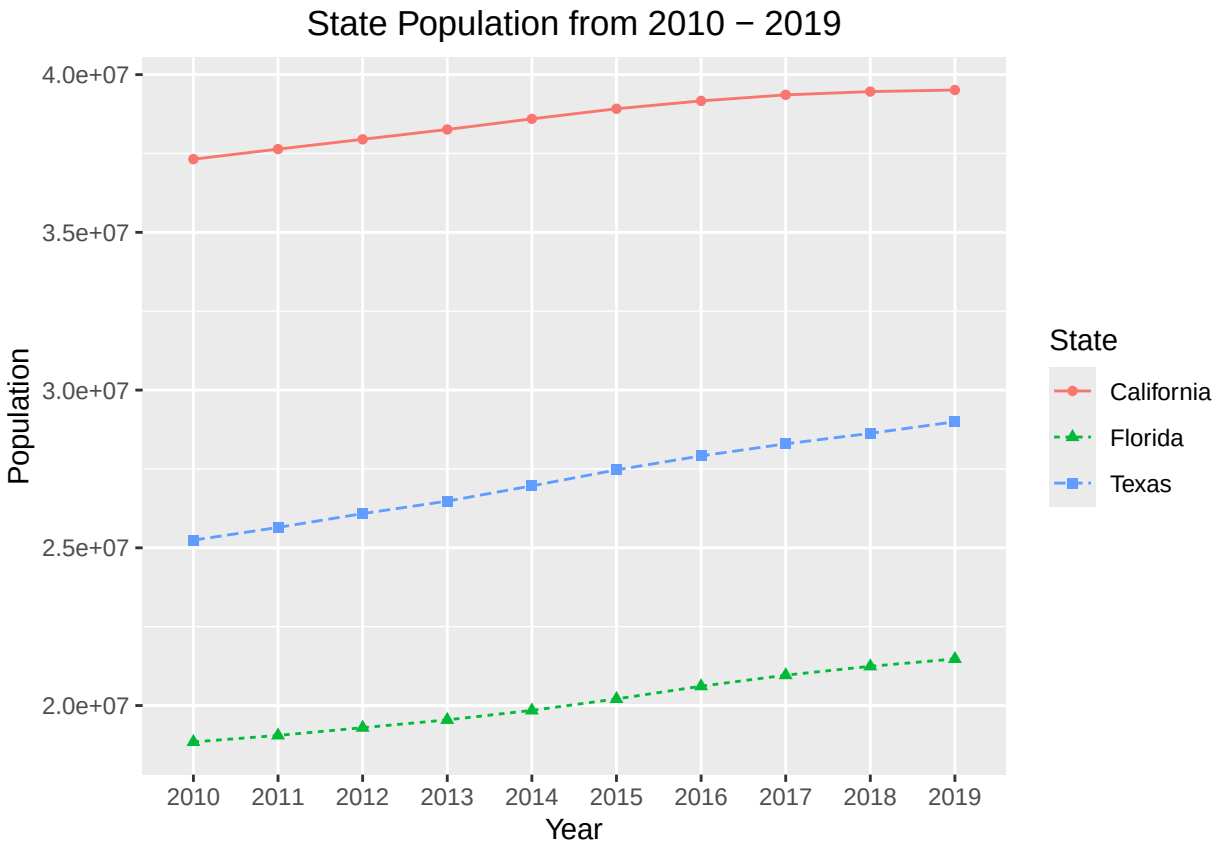
```
##  4 Alabama 2013     4830081
##  5 Alabama 2014     4841799
##  6 Alabama 2015     4852347
##  7 Alabama 2016     4863525
##  8 Alabama 2017     4874486
##  9 Alabama 2018     4887681
## 10 Alabama 2019     4903185
## # i 510 more rows
```

```
pivot_long_for_line <- filter(pivot_long, State == "California" | State == "Texas" | State == "Florida"

pivot_long_for_line
```

```
## # A tibble: 30 x 3
##    State      Year  Population
##    <chr>      <fct>      <dbl>
##  1 California 2010    37319502
##  2 California 2011    37638369
##  3 California 2012    37948800
##  4 California 2013    38260787
##  5 California 2014    38596972
##  6 California 2015    38918045
##  7 California 2016    39167117
##  8 California 2017    39358497
##  9 California 2018    39461588
## 10 California 2019    39512223
## # i 20 more rows
```

```
plt <- ggplot(pivot_long_for_line, aes(x=Year, y=Population, group = State)) + geom_line(aes(linetype =
plt
```



Problem 4

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

  Missing data and inconsistent formatting of data are two ways in which data can be dirty. Missing data occurs when data points are absent or missing. A potential solution to address missing data is to impute a statistical value for the missing value. These descriptive statistical values can vary, however, common statistics utilized are the mean or median. If the sample size is enough, the missing data, or null value, can be dropped from the data set. Inconsistent formatting of data occurs when data entries are not standardized. A common formatting inconsistency deals with text. For example, an entry can have "USA" vs. "U.S.A." Although these entries signify the same value, statistical analysis would produce different conclusions for "USA" and "U.S.A." The solution for this inconsistency is to normalize the data and standardize the formatting. Data normalization can also be applied to data with varying scale so that one variable with larger scale of measurement does not have more influence on the model or analysis than another variable with a smaller scale.

- 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?

  As the question attempts to group customers who buy similar things into separate entities without pre-established classifications, the data mining functionality for cluster analysis should be used to help with this data question. The cluster analysis would help form groups that are not yet established based on their purchase history.

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

  As the question attempts to gain insight into the purchase habits of customers and predicting the likelihood of buying milk based on other purchases made, the data mining funcationality for classification

should be used to help with this data question. Classification requires the prediction of an outcome into a category, in this case would be to determine whether a customer will purchase milk based on the their other purchases.

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

  As the question aims to uncover products that are often purchased together, the data mining functionality for association rule mining would be used to help with this data question. Association rule mining finds the events that occur together, in this case would be the sets of products that are purchased together.

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

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

  This is not a data mining task as the task is to simply reorganize the dataset according to a variable, the education level of customers. It does not provide any new actionable insight based on pattern discovery.

  - b: Computing the total sales of a company.

  This is not a data mining task as computing the total sales of a company does not provide any valuable pattern that would help develop an actionable plan.

  - c: Sorting a student database according to identification numbers.

  This is not a data mining task as the task is to sort, or organize, the dataset according to a variable, which is similar to the task asked in part (a).

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

  This is a data mining task, specifically a classification task as a model would need to be created to predict the likely outcomes of tossing a pair of dice. The model can provide general rules of thumb in games where winning is dependent on rolling a pair of dice.

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

  This is a data mining task, specifically a time-series analysis as this model would be needed to predict the future stock price of a company. Based on historic prices, the time-series analysis would enable investors, shareholders, to understand the optimal time a company stock should be purchased or sold.