Ian Mulchrone

DSC 441 Homework 1

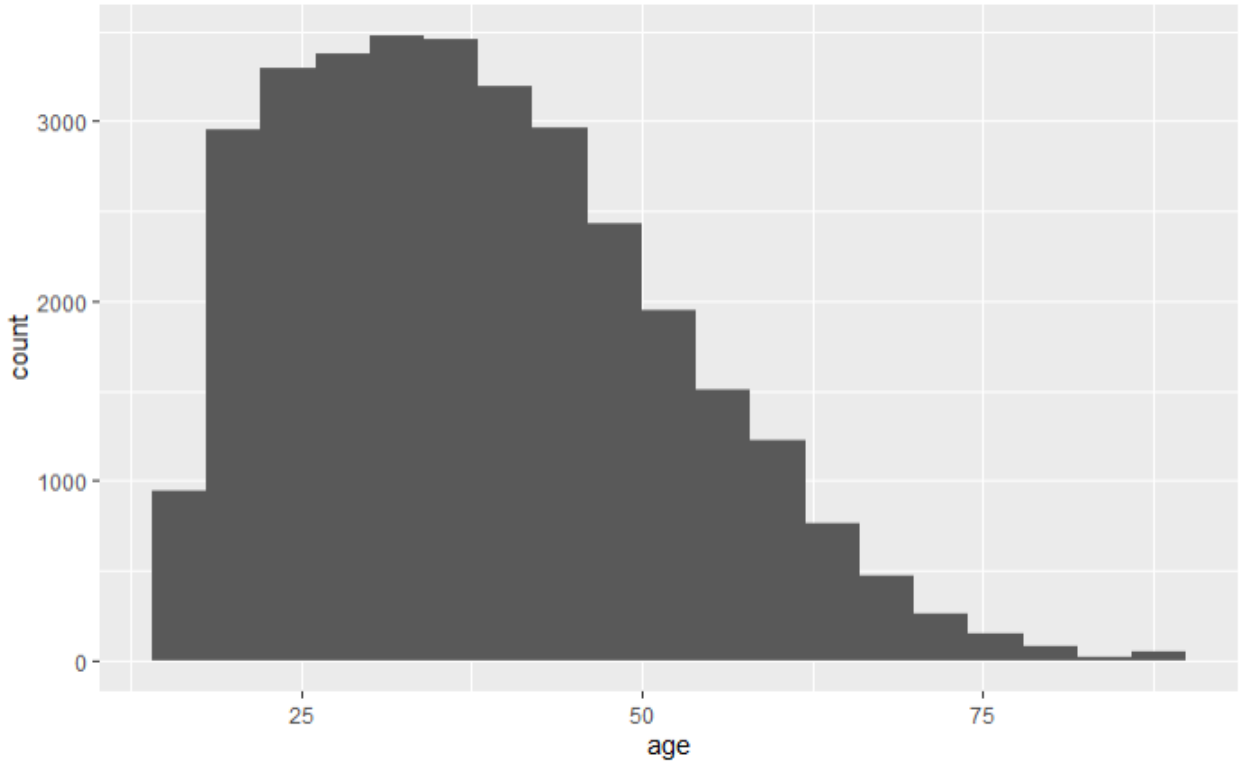
1) a.

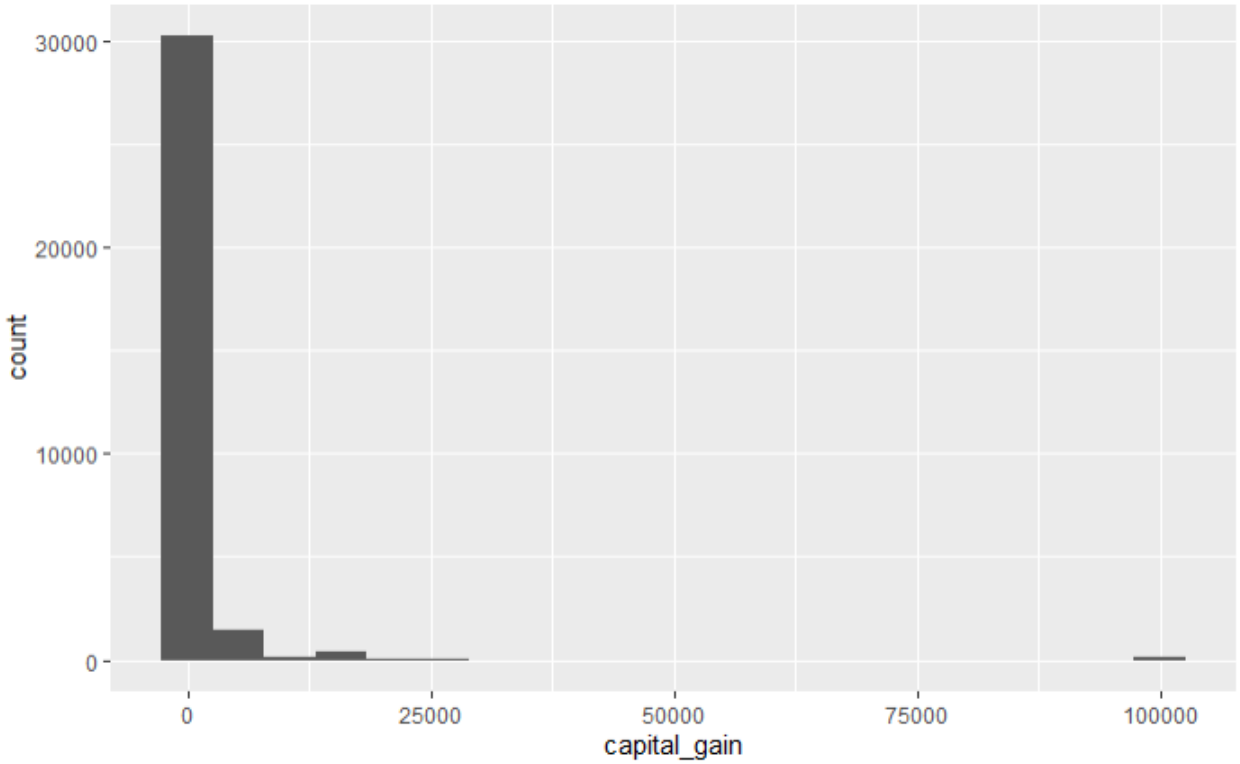
```
      age          workclass             fnlwgt          education          education-num
 Min.   :17.00   Length:32561       Min.   :  12285   Length:32561       Min.   : 1.00
 1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character   1st Qu.: 9.00
 Median :37.00   Mode  :character   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
```

Looking at age, the mean and median are relatively the same and are about equidistant to the 1st and 3rd quartiles. This indicates, based on those metrics, that the distribution is normal. Some questions arise when looking at the max value of 90 being much farther from the mean. Of course, when collecting adult data, you won't have any people under the age of 17 but, in this case, could be as old as 90. So, there may be outliers towards the right tail or may be simply be a case of right skewness.

'Capital-gain' is a much more obvious case of right skewness. All observations are 0 until after the 3rd quartile where after the values shoot up to a maximum of 99999. This causes the mean to be moved to 1078 while the median is 0.

b.

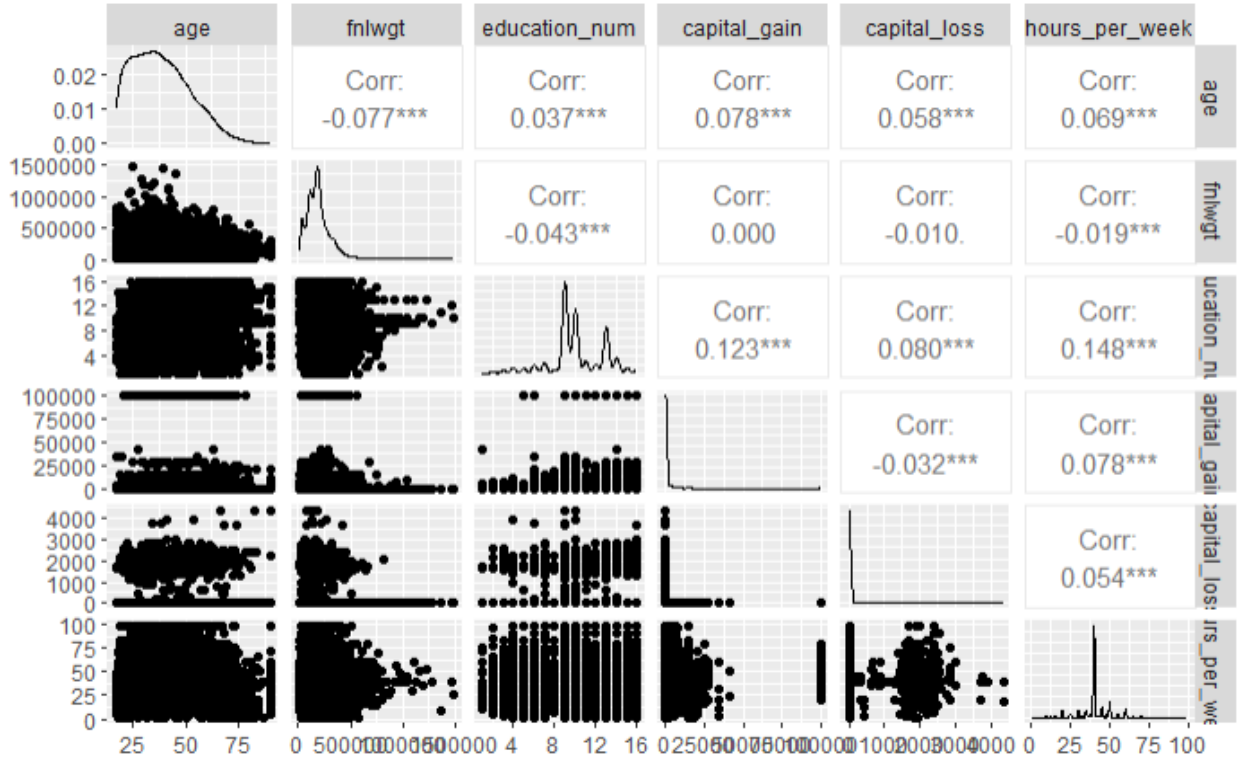




Histograms are the easiest way to display distribution of data and see the shape that is produced from the distribution. Looking at the shape of 'age' it makes sense that there is a right skew because, as mentioned before, the minimum cutoff age for adults is 17. What this visualization shows is that

this is a young population, with a majority of adults being under the age of 40 and very few being over the age of 70. The capital-gain distribution is exactly what we expected. The vast majority of adults reported 0 in capital gain, while a handful of adults reported around a 100000 increase, shifting the mean dramatically to the right.

c.



The scatterplot matrix shows the correlation between two variables and whether they have a pattern in their relationship. This is difficult to see when looking at just distributions because it gives the general shape of the distribution as a whole, as opposed to a more detailed view for each observation and how that can affect other variables.

d.

workclass <chr>	count <int>
?	1836
Federal-gov	960
Local-gov	2093
Never-worked	7
Private	22696
Self-emp-inc	1116
Self-emp-not-inc	2541
State-gov	1298
Without-pay	14

education <chr>	count <int>
10th	933
11th	1175
12th	433
1st-4th	168
5th-6th	333
7th-8th	646
9th	514
Assoc-acdm	1067
Assoc-voc	1382
Bachelors	5355
Doctorate	413
HS-grad	10501
Masters	1723
Preschool	51
Prof-school	576
Some-college	7291

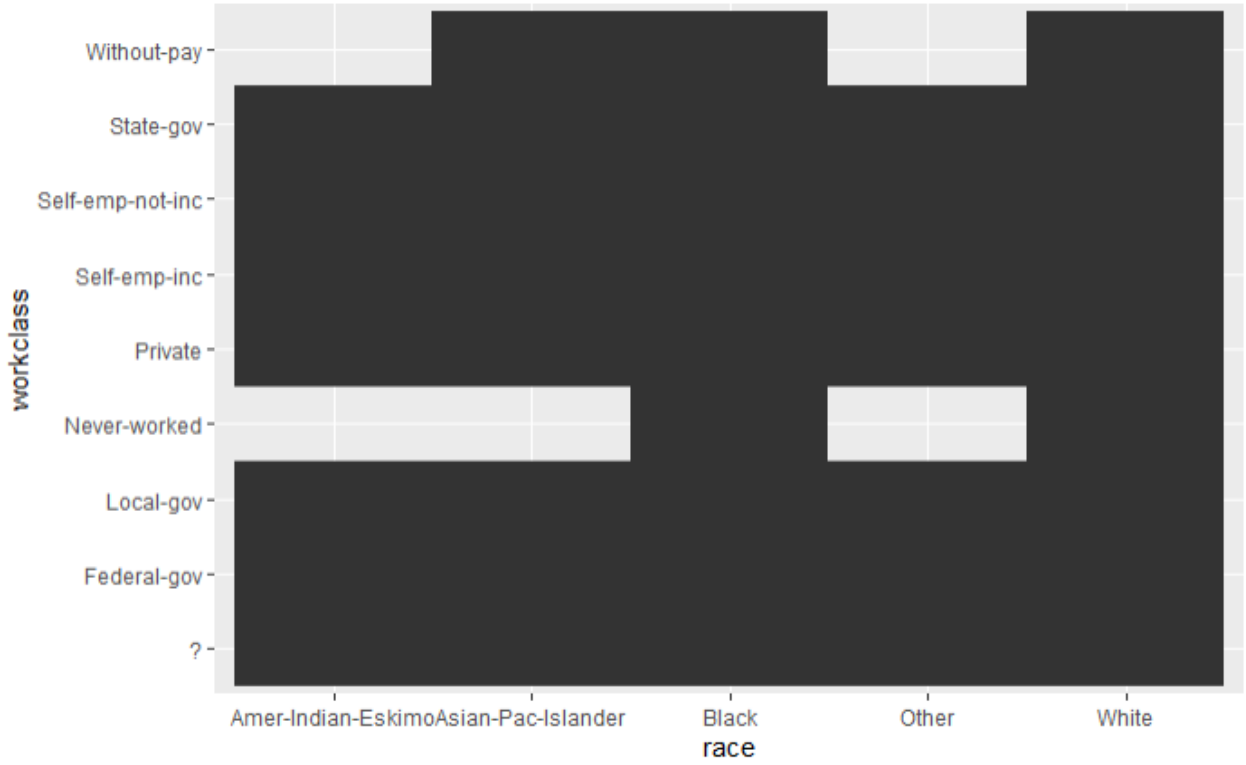
occupation <chr>	count <int>
?	1843
Adm-clerical	3770
Armed-Forces	9
Craft-repair	4099
Exec-managerial	4066
Farming-fishing	994
Handlers-cleaners	1370
Machine-op-inspct	2002
Other-service	3295
Priv-house-serv	149
Prof-specialty	4140
Protective-serv	649
Sales	3650
Tech-support	928
Transport-moving	1597

It is difficult to say if any category is over represented in the sample without knowing the data distribution in the population. For example, there are more than ten times the number of workers in the private sector than there are in the next closest work classification. This may be representative of the population, but it could also be an example of over sampling from this group. For education, high school graduates are the most common group, which is what you would expect, and some college is not too far behind that. For occupation, there are a few groups with around 4000 representatives, so none stand out of being over represented. However, with only 9 adults representing the armed forces category, this may be an example of under representing a group. Again, we'd need to examine the population data to verify this.

e.

workclass <chr>	Amer-Indian-Eskimo <int>	Asian-Pac-Islander <int>	Black <int>	Other <int>	White <int>
?	25	65	213	23	1510
Federal-gov	19	44	169	7	721
Local-gov	36	39	288	10	1720
Never-worked	*NA*	*NA*	2	*NA*	5
Private	190	713	2176	213	19404
Self-emp-inc	2	46	23	5	1040
Self-emp-not-inc	24	73	93	9	2342
State-gov	15	58	159	4	1062
Without-pay	*NA*	1	1	*NA*	12

(should be contingency matrix with gradient colors)



With workers in the private sector being the dominant category, we will choose to focus on non-private sectors. American Indian/Eskimo and Black workers have a high representation in local government positions relative to their private sector counterparts. Asian/Pacific Islanders, on the other hand, have a high percentage of self-employed workers.

2) a.

STATE.x <dbl>	NAME <chr>	POPESTIMATE2010 <dbl>	POPESTIMATE2012 <dbl>	POPESTIMATE2014 <dbl>	POPESTIMATE2016 <dbl>
1	Alabama	4785437	4815588	4841799	4863525
2	Alaska	713910	730443	736283	741456
4	Arizona	6407172	6554978	6730413	6941072
5	Arkansas	2921964	2952164	2967392	2989918
6	California	37319502	37948800	38596972	39167117
8	Colorado	5047349	5192647	5350101	5539215

6 rows | 1-6 of 13 columns

b.

STATE.x <dbl>	NAME <chr>	2010 <dbl>	2011 <dbl>	2012 <dbl>	2013 <dbl>	2014 <dbl>	2015 <dbl>	2016 <dbl>
1	Alabama	4785437	4799069	4815588	4830081	4841799	4852347	4863525
2	Alaska	713910	722128	730443	737068	736283	737498	741456
4	Arizona	6407172	*NA*	6554978	6632764	6730413	6829676	6941072
5	Arkansas	2921964	2940667	2952164	2959400	2967392	2978048	2989918
6	California	37319502	37638369	37948800	38260787	38596972	38918045	39167117
8	Colorado	5047349	5121108	5192647	5269035	5350101	5450623	5539215

6 rows | 1-9 of 12 columns

c.

STATE.x <dbl>	NAME <chr>	2010 <dbl>	2011 <dbl>	2012 <dbl>	2013 <dbl>	2014 <dbl>	2015 <dbl>	2016 <dbl>
1	Alabama	4785437	4799069	4815588	4830081	4841799	4852347	4863525
2	Alaska	713910	722128	730443	737068	736283	737498	741456
4	Arizona	6407172	6481075	6554978	6632764	6730413	6829676	6941072
5	Arkansas	2921964	2940667	2952164	2959400	2967392	2978048	2989918
6	California	37319502	37638369	37948800	38260787	38596972	38918045	39167117
8	Colorado	5047349	5121108	5192647	5269035	5350101	5450623	5539215

6 rows | 1-9 of 12 columns

d.

a.

2011 <dbl>	2012 <dbl>	2013 <dbl>	2014 <dbl>	2015 <dbl>	2016 <dbl>	2017 <dbl>	2018 <dbl>	2019 <dbl>	max <dbl>
4799069	4815588	4830081	4841799	4852347	4863525	4874486	4887681	4903185	4903185
722128	730443	737068	736283	737498	741456	739700	735139	731545	741456
6481075	6554978	6632764	6730413	6829676	6941072	7044008	7158024	7278717	7278717
2940667	2952164	2959400	2967392	2978048	2989918	3001345	3009733	3017804	3017804
37638369	37948800	38260787	38596972	38918045	39167117	39358497	39461588	39512223	39512223
5121108	5192647	5269035	5350101	5450623	5539215	5611885	5691287	5758736	5758736
3588283	3594547	3594841	3594524	3587122	3578141	3573297	3571520	3565287	3594841
907381	915179	923576	932487	941252	948921	956823	965479	973764	973764
619800	634924	650581	662328	675400	685815	694906	701547	705749	705749
19053237	19297822	19545621	19845911	20209042	20613477	20963613	21244317	21477737	21477737

1-10 of 52 rows | 4-13 of 13 columns

Previous 1 2 3 4 5 6 Next

b.

2012 <dbl>	2013 <dbl>	2014 <dbl>	2015 <dbl>	2016 <dbl>	2017 <dbl>	2018 <dbl>	2019 <dbl>	total <dbl>
4815588	4830081	4841799	4852347	4863525	4874486	4887681	4903185	48453198
730443	737068	736283	737498	741456	739700	735139	731545	7325170
6554978	6632764	6730413	6829676	6941072	7044008	7158024	7278717	68057899
2952164	2959400	2967392	2978048	2989918	3001345	3009733	3017804	29738435
37948800	38260787	38596972	38918045	39167117	39358497	39461588	39512223	386181900
5192647	5269035	5350101	5450623	5539215	5611885	5691287	5758736	54031986
3594547	3594841	3594524	3587122	3578141	3573297	3571520	3565287	35826676
915179	923576	932487	941252	948921	956823	965479	973764	9364455
634924	650581	662328	675400	685815	694906	701547	705749	6636276
19297822	19545621	19845911	20209042	20613477	20963613	21244317	21477737	201096314

1-10 of 52 rows | 5-13 of 13 columns Previous 1 2 3 4 5 6 Next

It's a simple change from part a because all you have to do is change the function from max to sum. This way, it will add up all the population numbers instead of simply selecting the largest value for each state.

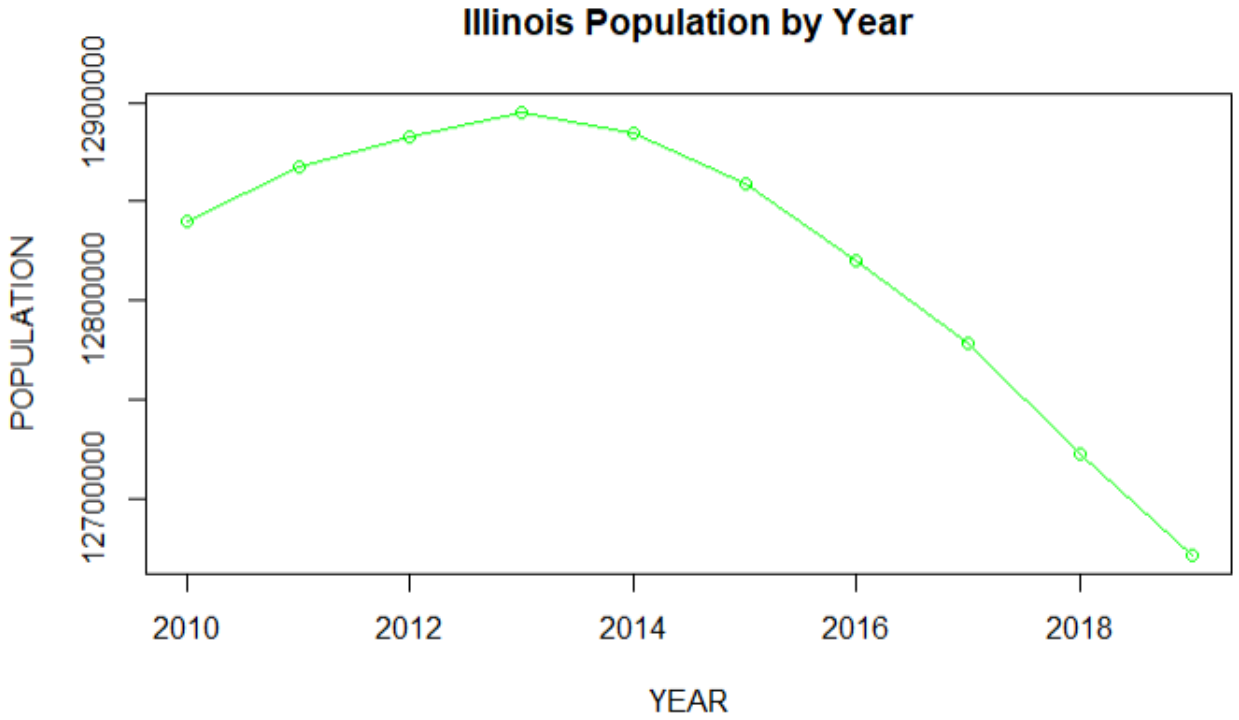
e.

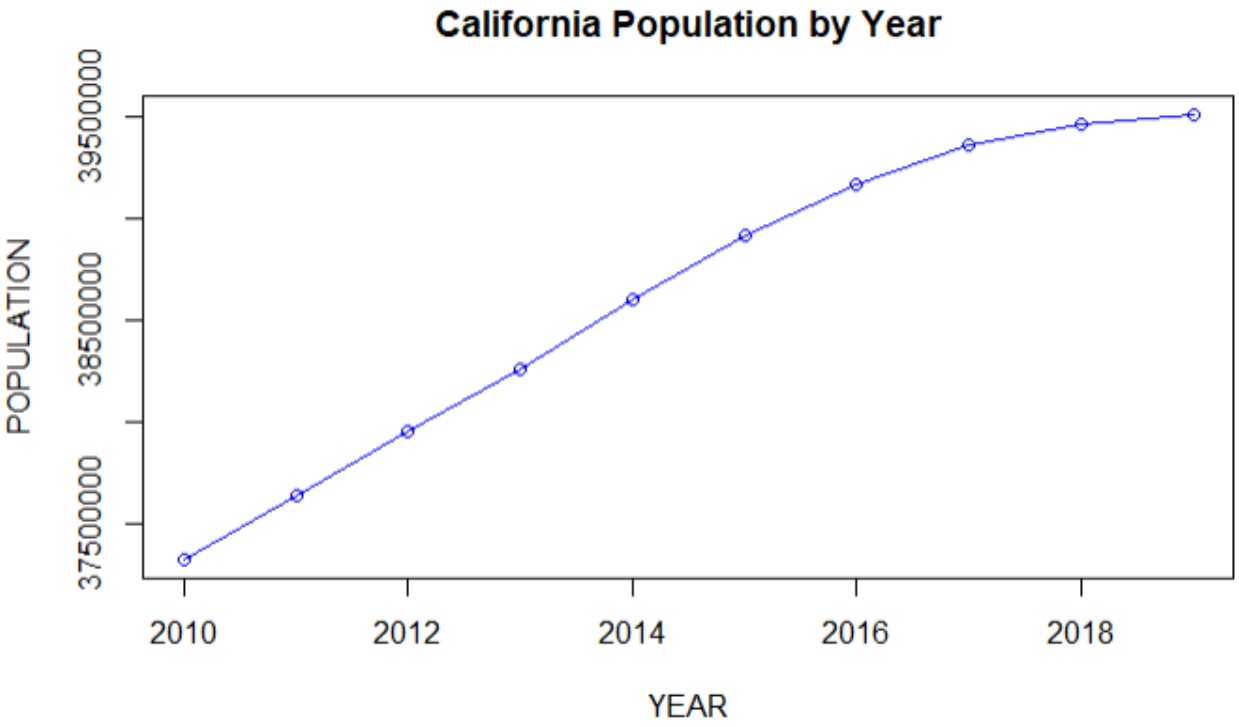
2012 <dbl>	2013 <dbl>	2014 <dbl>	2015 <dbl>	2016 <dbl>	2017 <dbl>	2018 <dbl>	2019 <dbl>	sum2010 <dbl>
4815588	4830081	4841799	4852347	4863525	4874486	4887681	4903185	313043191
730443	737068	736283	737498	741456	739700	735139	731545	313043191
6554978	6632764	6730413	6829676	6941072	7044008	7158024	7278717	313043191
2952164	2959400	2967392	2978048	2989918	3001345	3009733	3017804	313043191
37948800	38260787	38596972	38918045	39167117	39358497	39461588	39512223	313043191
5192647	5269035	5350101	5450623	5539215	5611885	5691287	5758736	313043191
3594547	3594841	3594524	3587122	3578141	3573297	3571520	3565287	313043191
915179	923576	932487	941252	948921	956823	965479	973764	313043191
634924	650581	662328	675400	685815	694906	701547	705749	313043191
19297822	19545621	19845911	20209042	20613477	20963613	21244317	21477737	313043191

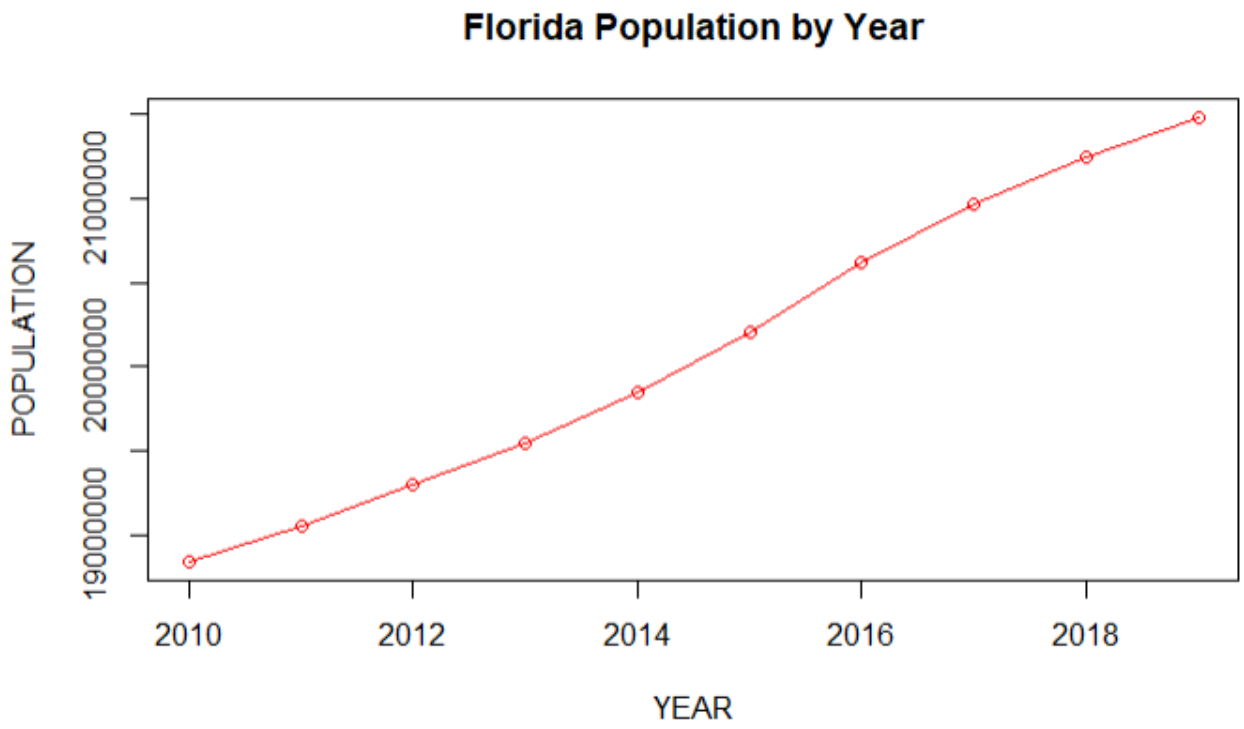
1-10 of 52 rows | 5-13 of 13 columns Previous 1 2 3 4 5 6 Next

Total population in 2010 is 313,043,191.

3. (Incorrectly displayed on three charts instead of one)







4.

a. One way that data can be dirty is when there is a malfunction in equipment used to collect the data. Once this is identified, the equipment can be repaired or replaced and then you can remeasure the data to ensure its accuracy. Another example is when there are inconsistencies in naming conventions or measurement units. If there are inconsistent names, you have to go through your data sets and identify which names are being used and change them to one so that they are all consistent. For different units, you can do conversions to whichever unit makes the most sense.

b.

a. Clustering would allow us to identify which groups of customers buy similar things.

b. Yes, using clustering, you can predict if a customer will buy milk based on if they've bought other items commonly bought with milk.

c. Milk, eggs, and cereal may be often purchased together, along with pasta, marinara sauce, and cheese. So, when trying to determine if milk will be purchased, whether or not the customer also bought cereal may be an attribute to predict that.

c.

a. This is not a data mining task. Simply organizing customers by education level is just grouping people together. If you were to use those groups to find meaningful information about each group's purchasing habits, that would be data mining.

b. This is not data mining; it is just calculating a number.

c. This is not data mining. Sorting students by number gives no meaningful information about each student, it is simply ordering them.

d. This is not data mining either. There is no data necessary to predict the outcome of a dice roll since it is always between 1 and 6. The outcome of the roll, assuming its fair, is random.

e. This is data mining. Using historical information to predict future stock prices requires analysis of identifying attributes that are indicators of stock performance and creating models to calculate predictions.