Exploring the factors related to average income per person in the United States

Yanlin Li, yanlinli@andrew.cmu.edu

Abstract

In this study, we focused on various factors influencing per-capita income for different counties in the United States, which is an interesting topic for social scientists. We used the county demographic information dataset from Geospatial and Statistical Data Center including several aspects of social well-beings to do our analysis. Methods such as Multiple Linear Regression, Box-Cox Transformation, Analysis of Variance, Variance Inflation Factor, and Baysian Information Criterion were used in our analysis. We discovered several pairwise relationships of the social factors, found the positive relationship between total crimes and per-capita income, developed an optimal linear model for per-capita income, and illustrated the influence of missing values. Finally, all the relationships found were interpreted with real life circumstances, five significant factors in the optimal model (land area, proportion of adults between 18 and 34 in age, number of active physicians, Western region, and percentage of high school graduates in Northeast region) were discussed, and weaknesses were pointed out with specified possible next steps.

Introduction

Personal income is always something that can attract attention from all parts of society. People use their income for basic living, entertainment, and investments. From a social science perspective, scientists are especially interested in how average income per person is related to other variables representing economic, health, and social well-being. In this study, we will use county (a governmental unit in the United States that is larger than a city but smaller than a state) as a basic unit of calculating per-capita income, together with other variables of the county to address various social science questions by data analysis. The questions are as follows:

1. Is there any variables related to each other?
2. Is per-capita income related to crime rate, and is the relationship influenced by region?
3. What is the best model to predict per-capita income?
4. Is there any problem caused by missing states?

We will analyze and try to solve these problems in the following sections.

Data

We will use the CDI dataset from Geospatial and Statistical Data Center, University of Virginia (Kutner et al. 2005). The dataset includes some demographic information (CDI) from the 440 counties with the largest population in the United States in the year 1990. Counties with missing data will be There are 17 columns in this data, including three basic identification of the county, and 14 variables to consider. Here are the definitions of each column:

1. `id`: Identification number (1-400)

2. `county`: County name
3. `state`: State (Two-letter state abbreviation)
4. `land.area`: Land area (square miles)
5. `pop`: Total population (Estimated 1990 population)
6. `pop.18_34`: Percent of 1990 CDI population aged 18–34
7. `pop.65_plus`: Percent of 1990 CDI population aged 65 or old
8. `doctors`: Number of professionally active non-federal physicians during 1990
9. `hosp.beds`: Number of hospital beds (Total number of beds, cribs, and bassinets during 1990)
10. `crimes`: Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11. `pct.hs.grad`: Percent of adult population (persons 25 years old or older) who com- pleted 12 or more years of school
12. `pct.bach.deg`: Percent of adult population (persons 25 years old or older) with bach- elor's degree
13. `pct.below.pov`: Percent of 1990 CDI population with income below poverty level
14. `pct.unemp`: Percent of 1990 CDI population that is unemployed
15. `per.cap.income`: Per-capita income (i.e. average income per person) of 1990 CDI population (in dollars)
16. `tot.income`: Total personal income of 1990 CDI population (in millions of dollars)
17. `region`: Geographic region classification used by the US Bureau of the Census, NE (northeast region of the US), NC (north-central region of the US), S (southern region of the US), and W (Western region of the US)

We first investigate the response variable: per-capita income.

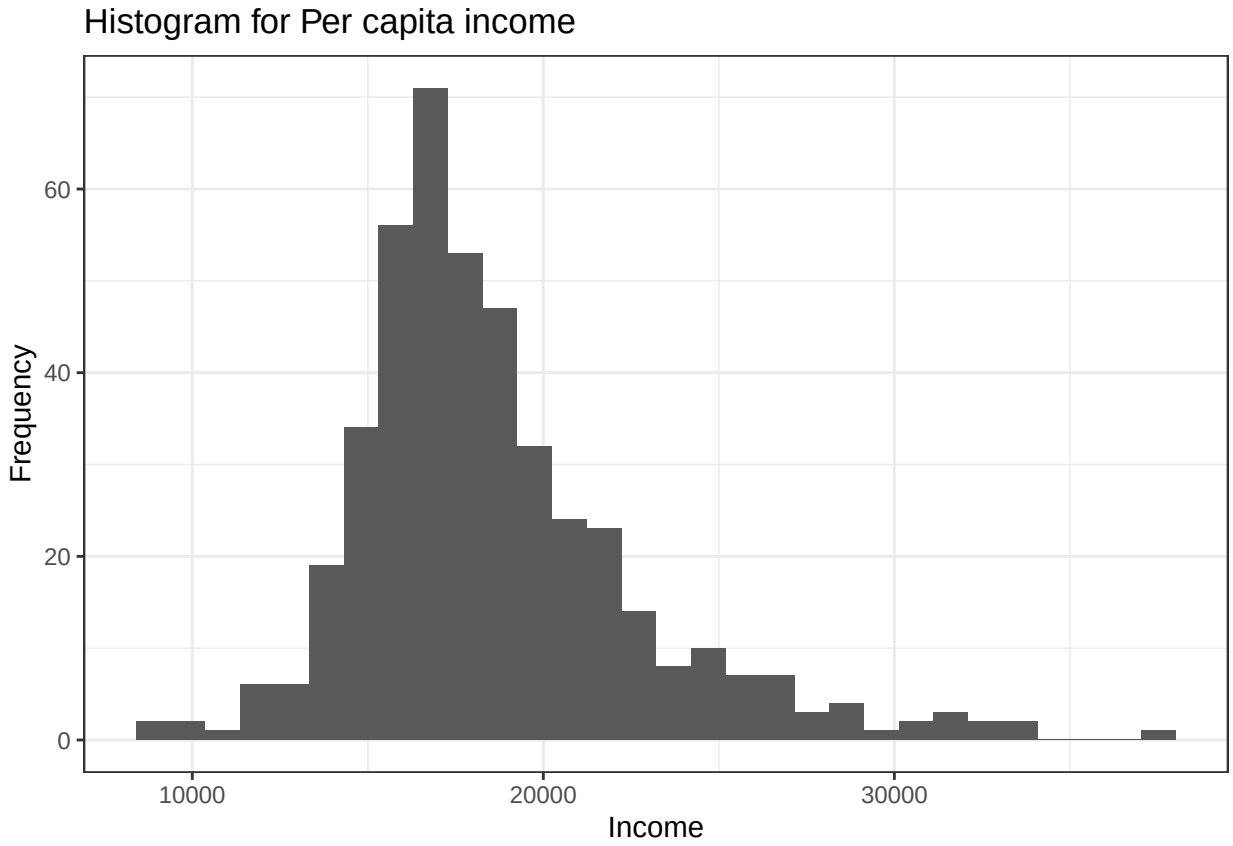

Figure 1: Histogram for Per capita income

As is shown in Figure 1, the distribution of per-capita income in different counties is right skewed and peaking around 17000 dollars. Most counties have per capital income within the range of 14000 and 23000 dollars. There is one county which has a very high per capital income.

Here is a basic summary for the numerical variables. (Table 1)

Table 1: Summary table: Continuous variables

Variable	Min	First.Qu	Median	Mean	SD	Third.Qu	Max
land.area	15.0	451.25	656.50	1041.41	1549.92	946.75	20062.0
pop	100043.0	139027.25	217280.50	393010.92	601987.02	436064.50	8863164.0
pop.18_34	16.4	26.20	28.10	28.57	4.19	30.02	49.7
pop.65_plus	3.0	9.88	11.75	12.17	3.99	13.62	33.8
doctors	39.0	182.75	401.00	988.00	1789.75	1036.00	23677.0
hosp.beds	92.0	390.75	755.00	1458.63	2289.13	1575.75	27700.0
crimes	563.0	6219.50	11820.50	27111.62	58237.51	26279.50	688936.0
pct.hs.grad	46.6	73.88	77.70	77.56	7.02	82.40	92.9
pct.bach.deg	8.1	15.28	19.70	21.08	7.65	25.33	52.3
pct.below.pov	1.4	5.30	7.90	8.72	4.66	10.90	36.3
pct.unemp	2.2	5.10	6.20	6.60	2.34	7.50	21.3
per.cap.income	8899.0	16118.25	17759.00	18561.48	4059.19	20270.00	37541.0
tot.income	1141.0	2311.00	3857.00	7869.27	12884.32	8654.25	184230.0

Per-capita income has a mean of 18561.48 and standard deviation of 4059.19. The minimum is 8899 and the maximum is 37541, which represents a huge difference.

Other interesting findings from the summary table above (Table 1) includes the statistics for crimes. The minimum number of crimes is only 563, and the maximum is 688936, about 1224 times higher than the minimum. The maximum unemployment rate is 21.3%, which means that over one fifth of the population do not have a job. The medical condition of the county also varies a lot. The standard deviation of number of active physicians and number of hospital beds are 1789.75 and 2289.13 respectively.

Here is another summary for the categorical variable: region. (Table 2)

Table 2: Summary table: Geographic region

Region	Count
NC	108
NE	103
S	152
W	77

There are highest number of counties in the southern region, and the number is the lowest in western region.

Methods

To address the questions listed in the introduction section, I did the following analysis of the data.

To answer the first question, I did some graphing using R to analyze the pairwise relationships between some variables. The variable per-capita income was especially emphasized in this part because this is the variable of interest. Confounding factor (factor correlated to two or more variables, making these variables correlated) was detected and analyzed. (Appendix (a))

For the second question, a model was first built for predicting per-capita income using region and crime rate. Another model including the income grouped by regions was also included for comparison. An ANOVA (Analysis of Variance) table was built to assess the necessity for the added interaction. We picked the better model of the two choices and changed the factor crime rate into the total number of crimes. Finally, we compared the changed model to the original one for the best interpretation of the relationship between per-capita income and crimes and region. (Appendix (b))

For the third question, our goal was to find the best model within the given variables to predict the per-capita income of the counties. Here are the steps towards it. First, we explored some transformations for each variable to make them follow a normal distribution using histogram and Box-Cox methods. The reason for that is to satisfy the normality assumption of linear regression. (Appendix (c) Transformations) Second, we did some variable selection. We first deleted the total income variable, which is the one that can calculate per-capita income directly when divided by population. Then, VIF (Variance Inflation Factor) was used to check for multicollinearity conditions. We expect all the variables to be independent of each other to satisfy the multiple linear regression assumptions. Scatterplots were used to detect correlations for factors with high VIF values. BIC was also used after deleting variables for collinearity, to develop a simple model. (Appendix (c) Variable selection) Third, we checked for any possible improvement when grouping one of the variables in the model by region. ANOVA (Analysis of Variance) table was used in this process and we extracted all the p-values of the F-tests. The null hypothesis of the test is that the interaction between the region and the variable does not help improve the model. When the p-value is below the threshold of 0.05, we can reject that null hypothesis and include the interaction term. (Appendix (c) Interaction) Finally, we got the optimal model, and did some model diagnostics. (Appendix (c) Final model discussion)

To answer the fourth question, we investigated the missing counties and states and check for common factors of them. The response variable per-capita income for missing states were also researched to see whether they have extreme values.

Results

Now we can look at the results of the above analysis.

Question 1

We first consider plots related to per-capita income:

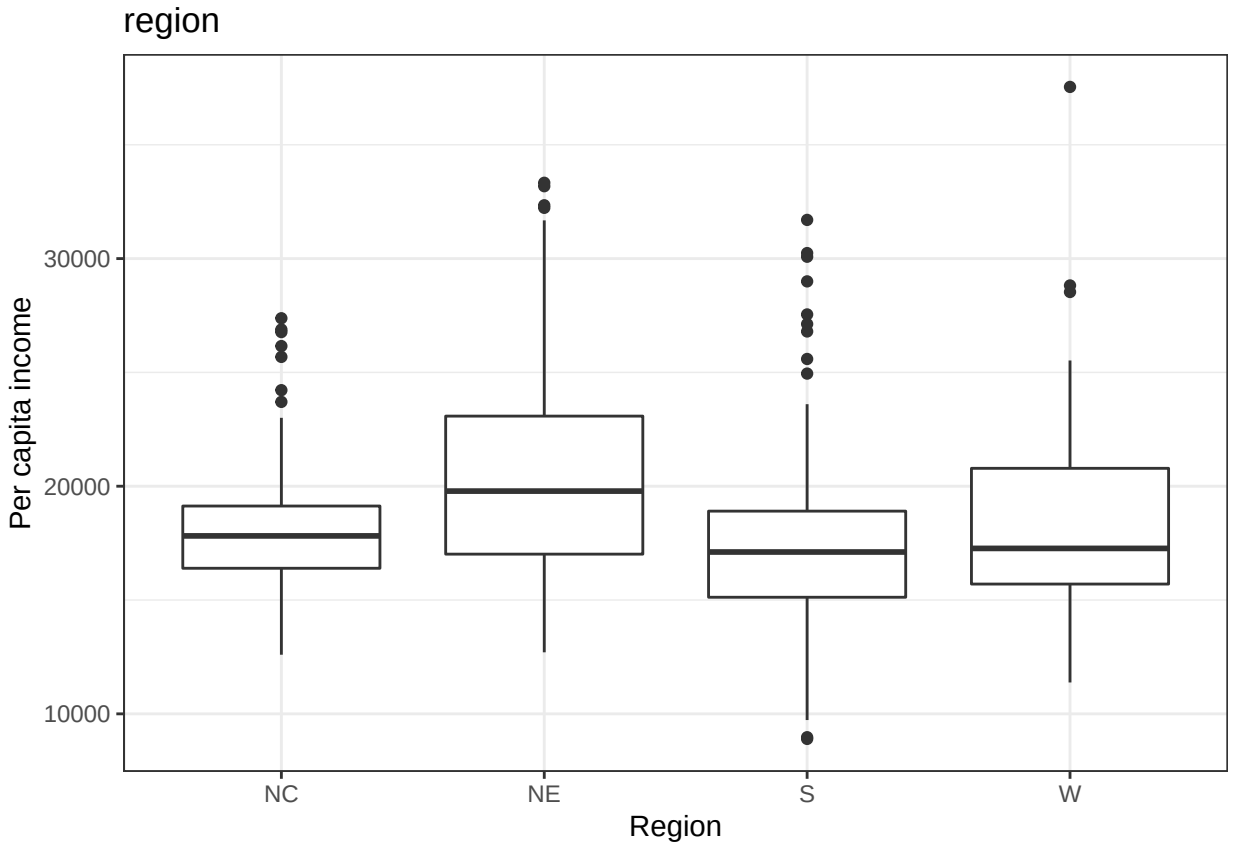

Figure 2: Boxplot for Per capita income vs region

The median per capital is the highest in northeast region, and the lowest in southern region. There are some outliers for the west region with very high values.

There are some scatterplots for numerical factors.

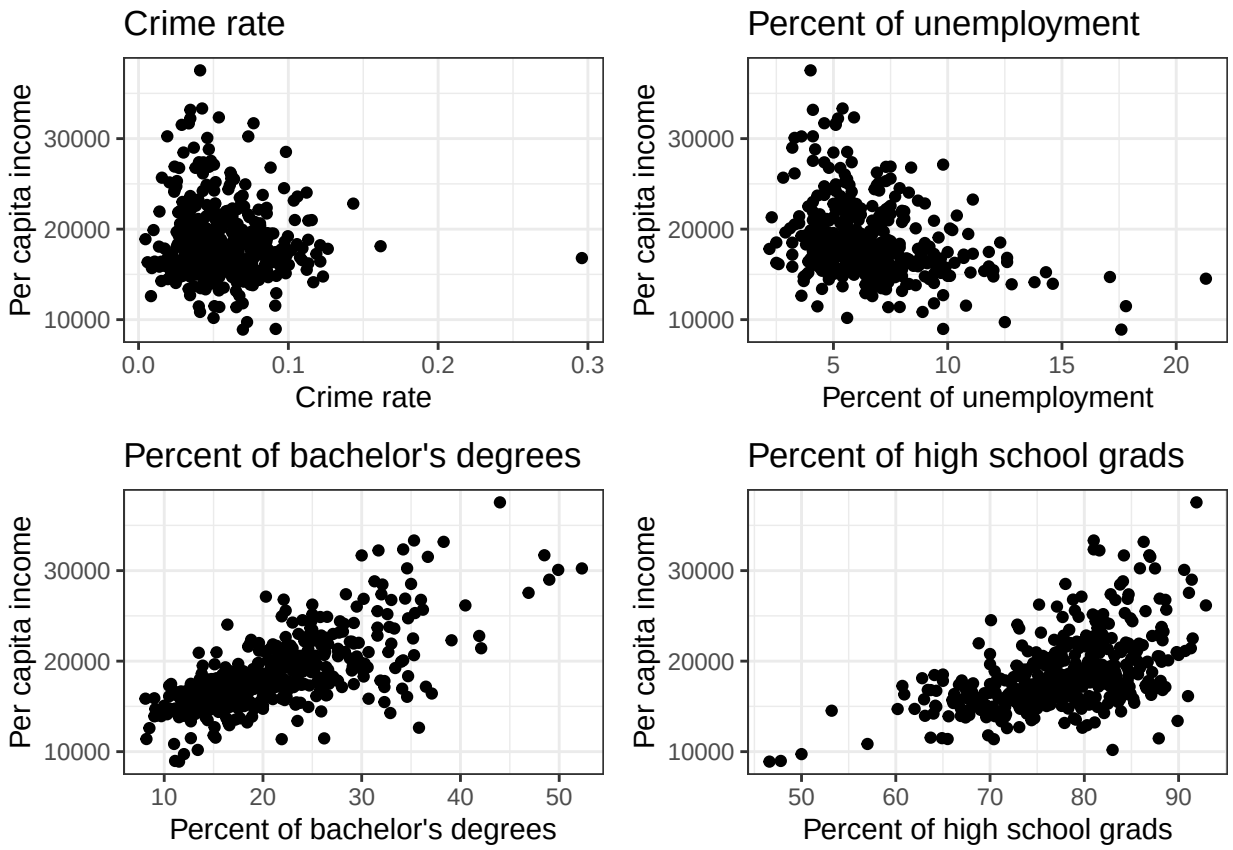

Figure 3: Per capita income vs Crime rate, Percent of unemployment, Percent of bachelor's degrees, Percent of high school grads

By Figure 3, we can see that the counties having high crime rate usually have medium per capital income (around 20000). There is a negative relationship between per capital income and percent of unemployment. Also, we can see increasing trend in the graphs for percent of bachelor's degrees and percent of high school grads. For the latter, there also exists an increasing variance of per-capita income with the growing percent of high school grads.

Then we explore relationships related to number of active physicians.

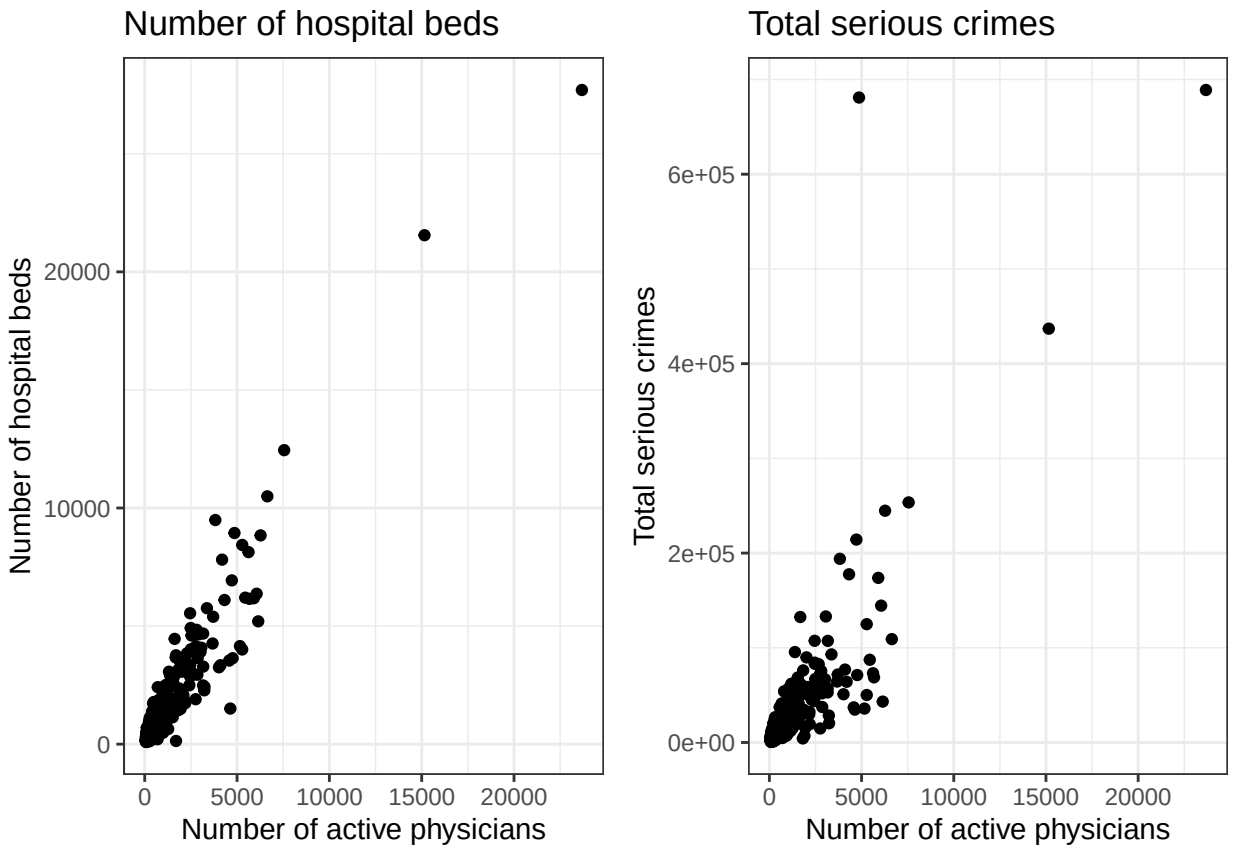

Figure 4: Number of active physicians vs Number of hospital beds, Total serious crimes

We can see a significant positive relationship between the number of active physicians and the number of hospital beds. There is also a positive relationship between number of active physicians and total serious crimes.

Next, we examined the potential confounding factor total population.

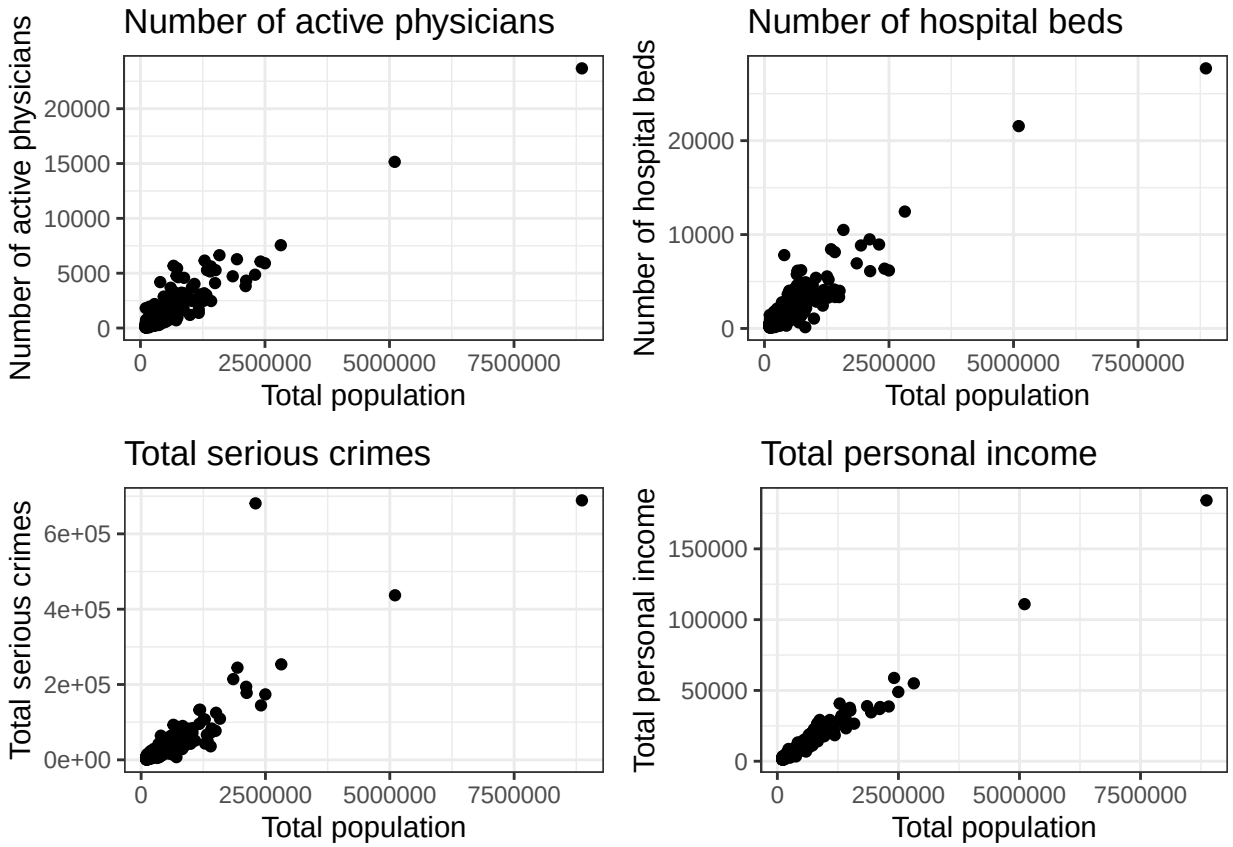

Figure 5: Total population vs Number of active physicians, Number of hospital beds, Total serious crimes, Total personal income

We can see that the factor population is correlated with all the four variables.

Question 2

We first built two models, one predicts per-capita income with crime rate and region (Model 1), and the other added the relationship of the two variables into the model (Model 2). We tested the null hypothesis that we prefer Model 1 to Model 2. The p-value is $0.99 > 0.05$, so Model 1 makes more sense. We should not include the relationship between region and crime rate. Below are the summary tables for both Model 1 and Model 2.

Table 3: Coefficients for the model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18006.04469	537.0395	33.5283439	0.0000000
regionNE	2354.69663	541.9715	4.3446875	0.0000174
regionS	-927.44668	512.3059	-1.8103378	0.0709333
regionW	-34.92294	586.0281	-0.0595926	0.9525075
crime.rate	5773.20230	7520.4126	0.7676710	0.4430992

Table 4: Coefficients for the model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18077.2939	895.2077	20.1934068	0.0000000
regionNE	2329.0367	1101.3942	2.1146260	0.0350340

	Estimate	Std. Error	t value	Pr(>|t|)
regionS	-1010.3526	1323.8024	-0.7632201	0.4457487
regionW	-669.9909	1983.8920	-0.3377154	0.7357417
crime.rate	4379.0699	15893.5069	0.2755257	0.7830441
regionNE:crime.rate	288.3868	20184.6607	0.0142874	0.9886073
regionS:crime.rate	1558.9186	20556.1122	0.0758372	0.9395837
regionW:crime.rate	10655.5422	32322.4079	0.3296642	0.7418135

By Table 3, there is a positive relationship between total serious crimes and per capital income, because the coefficient for crime rate in the model is positive. However, this relationship is not significant, because the p-value for it is $0.44 > 0.05$.

Here is a summary for the model with total crimes, instead of crime rate. (Model 3)

Table 5: Coefficients for the model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18106.9099510	378.4380302	47.8464333	0.0000000
regionNE	2286.0373120	532.4709260	4.2932622	0.0000217
regionS	-860.5567325	486.8305023	-1.7676722	0.0778167
regionW	-142.8266313	579.6196624	-0.2464144	0.8054777
crimes	0.0089153	0.0031877	2.7968322	0.0053895

The coefficient is now strictly above 0, indicating a positive relationship between total crimes and per-capita income. (See Appendix (b) for details)

Question 3

Transformations (See Appendix (c) Transformations for details)

Here are the transformations we used for modeling:

1. Log transform $\log(X)$ for: Per capita income, Land area, Percent of population aged 18–34, Percent of population 65 or older, Number of active physicians, Number of hospital beds, Total serious crimes, Percent bachelor's degrees, Percent below poverty level, Percent unemployment
2. One over square root $X^{-\frac{1}{2}}$ for: Total population, Total personal income
3. Cube square X^3 for: Percent high school graduates

Variable selection (See Appendix (c) Variable selection for details)

We first deleted the variable total income of the county, because this variable divided by population is exactly per-capita income. Including it can be meaningless.

For multicollinearity condition, using VIF table and scatterplots, we decided to delete:

1. Number of hospital beds: It is correlated with number of active physicians
2. Total population: This variable is basically correlated with all the "total" statistics, as a confounding variable between each variable and per-capita income.
3. Percent bachelor's degrees: It is correlated with percent below poverty, unemployment rate and percent of high school graduates.

Using BIC, we finally chose the model with five variables: land area, percent of population aged 18-34, number of active physicians, percent high school graduates, and percent below poverty level. The R-squared value is 0.7822, which is an acceptable value for a model with only 5 parameters.

Interaction (See Appendix (c) Interaction for details)

In our exploration of interaction, we found the relationship between per-capita income and percent of high school graduates can be influenced by region. So the final model will include the interaction term between the two.

Final model (See Appendix (c) Final model discussion for details)

Here is the summary table for our final model.

Table 6: Final Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.6667050	0.1263162	84.444480	0.0000000
land.area	-0.0490628	0.0064826	-7.568407	0.0000000
pop.18_34	-0.2057823	0.0371997	-5.531833	0.0000001
doctors	0.0887292	0.0041972	21.140096	0.0000000
pct.hs.grad	0.0000001	0.0000001	1.019629	0.3084805
regionNE	-0.1135711	0.0631627	-1.798071	0.0728704
regionS	-0.0799485	0.0536723	-1.489565	0.1370751
regionW	0.1441645	0.0664288	2.170212	0.0305401
pct.below.pov	-0.2256078	0.0138861	-16.246969	0.0000000
pct.hs.grad:regionNE	0.0000003	0.0000001	2.194325	0.0287488
pct.hs.grad:regionS	0.0000002	0.0000001	1.713309	0.0873798
pct.hs.grad:regionW	-0.0000002	0.0000001	-1.687523	0.0922314

Finally, here is the model diagnostic plots and the analysis.

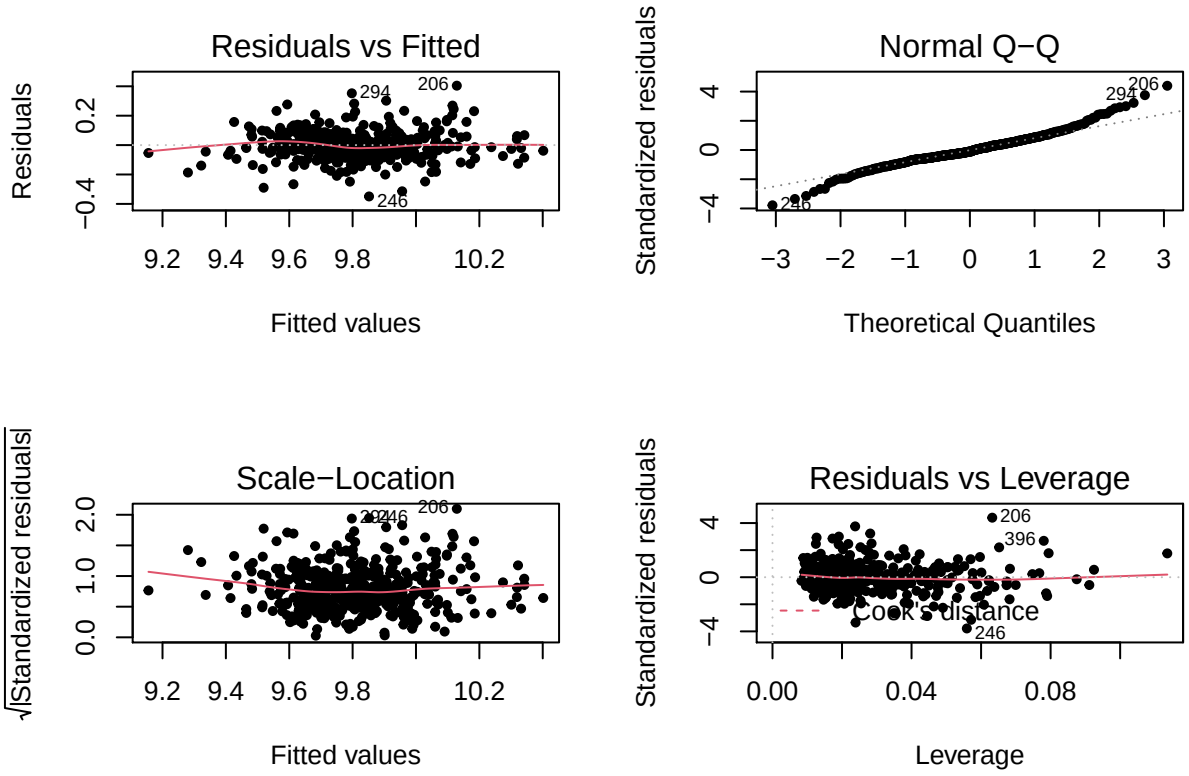

Figure 6: Model diagnostic plots

1. Residuals vs Fitted: The residuals are distributed evenly with a mean 0, with no pattern detectable.
2. Normal Q-Q: The fitted value is slightly skewed and not completely normal.
3. Scale-Location: There is no observable pattern, so the variance looks constant.
4. Residuals vs Leverage: There are no observable leverage points or outliers.

This model can be appropriate for prediction.

Question 4

The missing states are: Alaska, Iowa, Wyoming. The state per-capita income for all the three states are between $31557 and $38915 (Map: Per capita income in past 12 months (in 2019 dollars), 2015-2019, 2019), which is not extremely high or low.

The population for them are 626932, 2926324, and 493782 respectively (The States of the USA on a Map, 2021), which are relatively small, especially for Alaska and Wyoming. The reason of missing them is that they do not have a county with population ranking within 440 among all counties, which is another evidence of small population. Besides, the reason of not including a county is either including missing values, or small in population (majority). So relatively small population can be a common factor in all the missing data. Fortunately, as is shown in the plot below, the population is not significantly correlated with per-capita income.

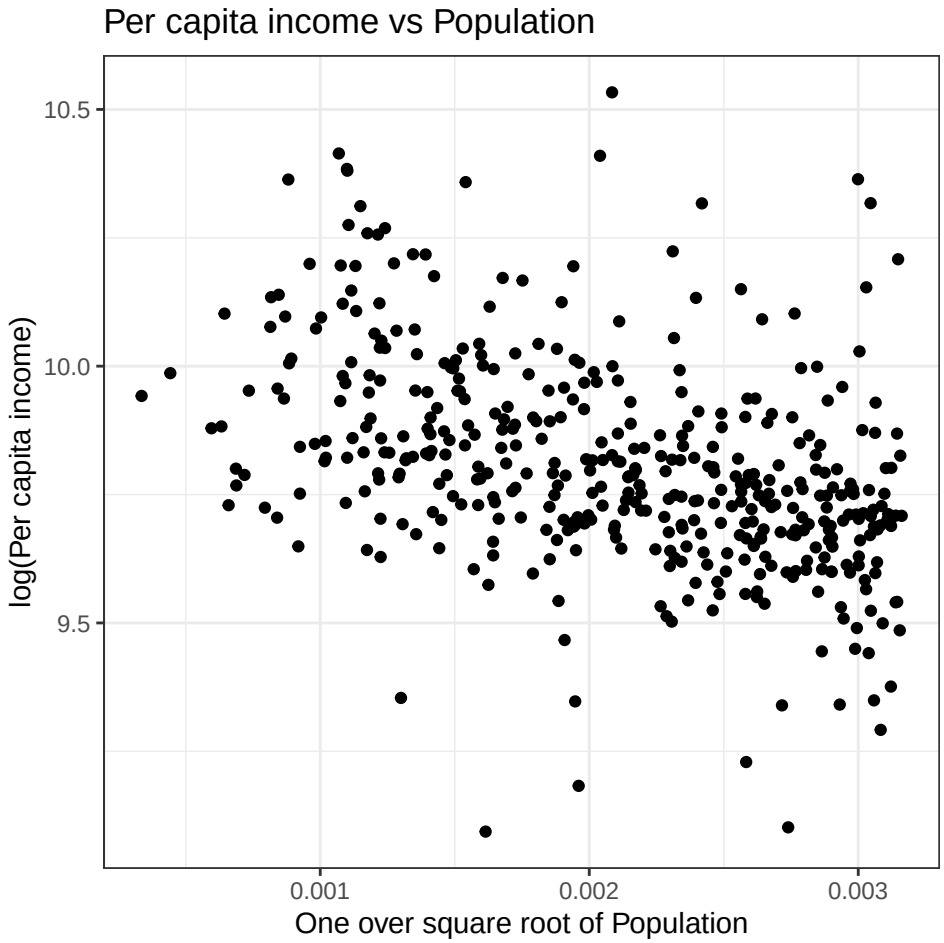

Figure 7: Total population vs per-capita income

Discussion

In this section, we will dig into the results we got from the data and discuss the research questions in the introduction.

Question 1

Is there any variables related to each other?

From the graphs related to per-capita income, there are four main findings:

First, the median per-capita income is the highest in northeast region, and the lowest in southern region. This is consistent to the reality that the density of big cities is the highest in northeast region and the lowest in the southern region. The rich zones in the western region resulted in the outliers.

Second, crime rates are usually higher in areas with medium per-capita income. This can be an interesting result that the crime rate does not significantly decrease with higher income. In reality, almost everyone are having low income in counties with very low per-capita income. One will not get much money from others even if they commit crimes like robbery. In contrast, the income disparity can be larger in middle-income regions, which encourages the criminals to commit crimes.

Third, the negative relationship between per-capita income and unemployment rate is within expectation. Because the unemployed people usually have no or low income, a high percentage of them will result in low per capital income.

Fourth, education is an important factor in determining per-capita income. High school education is a basis of higher per-capita income. It will not definitely result in an increase in income, but a low percent of high school grads will certainly result in a low income. In contrast, percent of adults with a bachelor degree can be a good indicator of per-capita income. The higher the percentage, the higher the per-capita income will be.

From the graphs related to number of active physicians, we detected that the number of active physicians and the number of hospital beds are significantly correlated. It is possible for them to be correlated because both of them can be treated as indicators of a county's medical resource. However, the number of active physicians is surprisingly correlated with total crimes. This cannot be explained by any real life situations, because increasing number of crimes are not likely to make more people become physicians. Thus, we switched to explore some factors that acts as a "medium" to stick these unrelated factors together (i.e. confounding factor). From the graphs related to population, we convinced that population is related to all the variables with aggregation: number of active physicians, number of hospital beds, total serious crimes, and total personal income.

Consequently, the answer of the question is definitely yes, but some of the relationships are influenced by confounding factors.

Question 2

Is per-capita income related to crime rate, and is the relationship influenced by region?

From the result of the model summary, we can see a slightly positive relationship between per-capita income and crime rate. However, we cannot be sure that the relationship is significant and correct. Thus, we can conclude that per-capita income is not necessarily related to crime rate.

When we group crime rate with region, we can see positive relationship between per-capita income and crime rate for all the four regions, indicating that the relationship is not influenced by region. Our test also show that grouping crime rate with regions is unnecessary.

Furthermore, we replaced crime rate with total crimes to see the output. We can infer from the model that total crimes has a significantly positive relationship with per-capita income. Although it can be a result of confounding variable population (mentioned in Question 1), this can still better reflect the social science behind it for its certainty.

Question 3

What is the best model to predict per-capita income?

The best model contains five factors: land area, percent of population aged 18-34, number of active physicians, percent high school graduates influenced by region, and percent below poverty level.

From the optimal model, we can find several significant factors that can influence per-capita income:

1. Land area is negatively correlated with per-capita income. The larger the county, the lower per-capita income will be. It is interesting with the fact that people with highest income are usually in large cities, which are usually small in land area but dense in capital.
2. Proportion of people with age 18 to 34 is also negatively correlated with per-capita income. This can be surprising that the population with many young adults usually have less income. There can be several possible reasons for this fact. The first one is that young adults are usually new in the industry and may still be students without income. The second one is that high income families are less willing to give birth to children, causing aging societies in rich zones. Both of them need further evidence to support.
3. An increase in the number of active physicians is related to an increase in per-capita income. This is understandable, since high-income area usually have more medical resources. Also, physicians usually have high income, which can contribute to the higher per-capita income of the county.
4. Per-capita income in western region is significantly higher than other regions. Many big and high-tech companies have there main sites in western region. Employees and owners of these companies occupy a

large percentage of working population there. They usually have higher income than average, which is largely related to the high per-capita income.

5. Per-capita income is negatively related to percent below poverty. This is a straightforward result, since poverty means low income.

6. Percentage of high school graduates is significantly related to per-capita income only if they are in the northeast region of US. The relationship is positive, which is conceivable. The influence of region can be caused by the density of large cities in the northeast region. There are more working opportunities in these cities than other regions, and graduating from high school can make a difference in their career under this environment.

Question 4

Is there any problem caused by missing states?

We have explored two types of bias that can be caused by missing states. The first one is missing extreme values in our variable of interest, per-capita income. We examined all the three missing states and no extreme value can be found. The second one is the common properties of missing values, which are different from those in the data. Because we included only 440 most populous counties, this common property can be small population. However, from the scatterplot of population and per-capita income, we cannot see any correlation between the two variables.

Thus, we conclude that we cannot detect any problem caused by missing states using the data available.

Weaknesses & Next Step

1. Limited by dataset, we can only consider the data for top 440 counties in population. Although we conclude that there is no problem caused missing values, this is also a result under limited data. We took for granted that the relationship between population and per-capita income is the same for counties with all size of population. This may also cause a bias. In the future, we can include all the counties for analysis.

2. We used only multiple linear regression in this analysis, which cannot reflect all the possible relationships in reality. In the future, we can try other models such as Generalized Additive Model and Linear Mixed Model.

3. In this analysis, we cannot dig into all the social scientific inferences we made. That is, the reason why the relationship is negative or positive, and why the two factors are correlated with each other. We made some guesses in our discussion, but none of them can be proved by real life researches and data. For the next step, we can do more researches to find certain reasons and make recommendations for social well-beings.

References

1. Kutner, M.H., Nachsheim, C.J., Neter, J. & Li, W. 2005. Applied Linear Statistical Models, Fifth Edition. NY: McGraw - Hill/Irwin.

2. Englisch-hilfen.de. 2021. The States of the USA on a Map. [online] Available at: https://www.englisch-hilfen.de/en/texte/states.htm [Accessed 18 October 2021].

3. United States Census Bureau. 2019. Per capita income in past 12 months (in 2019 dollars), 2015-2019. [online] Available at: https://www.census.gov/quickfacts/fact/map/US/INC910219 [Accessed 18 October 2021].

Technical Appendix

(a)

Summary statistics

Categorial variables

Table 7: Summary table: State

State	Count
AL	7
AR	2
AZ	5
CA	34
CO	9
CT	8
DC	1
DE	2
FL	29
GA	9
HI	3
ID	1
IL	17
IN	14
KS	4
KY	3
LA	9
MA	11
MD	10
ME	5
MI	18
MN	7
MO	8
MS	3
MT	1
NC	18
ND	1
NE	3
NH	4
NJ	18
NM	2
NV	2
NY	22
OH	24
OK	4
OR	6
PA	29
RI	3
SC	11
SD	1
TN	8
TX	28
UT	4
VA	9
VT	1
WA	10

State	Count
WI	11
WV	1

Table 8: Summary table: Geographic region

Region	Count
NC	108
NE	103
S	152
W	77

Continuous variables

Table 9: Summary table: Continuous variables

Variable	Min	First.Qu	Median	Mean	Third.Qu	Max
land.area	15.0	451.250	656.50	1.041411e+03	946.750	20062.0
pop	100043.0	139027.250	217280.50	3.930109e+05	436064.500	8863164.0
pop.18_34	16.4	26.200	28.10	2.856841e+01	30.025	49.7
pop.65_plus	3.0	9.875	11.75	1.216977e+01	13.625	33.8
doctors	39.0	182.750	401.00	9.879977e+02	1036.000	23677.0
hosp.beds	92.0	390.750	755.00	1.458627e+03	1575.750	27700.0
crimes	563.0	6219.500	11820.50	2.711162e+04	26279.500	688936.0
pct.hs.grad	46.6	73.875	77.70	7.756068e+01	82.400	92.9
pct.bach.deg	8.1	15.275	19.70	2.108114e+01	25.325	52.3
pct.below.pov	1.4	5.300	7.90	8.720682e+00	10.900	36.3
pct.unemp	2.2	5.100	6.20	6.596591e+00	7.500	21.3
per.cap.income	8899.0	16118.250	17759.00	1.856148e+04	20270.000	37541.0
tot.income	1141.0	2311.000	3857.00	7.869273e+03	8654.250	184230.0

Missing values

There is no missing data in this dataset.

Variable features EDA

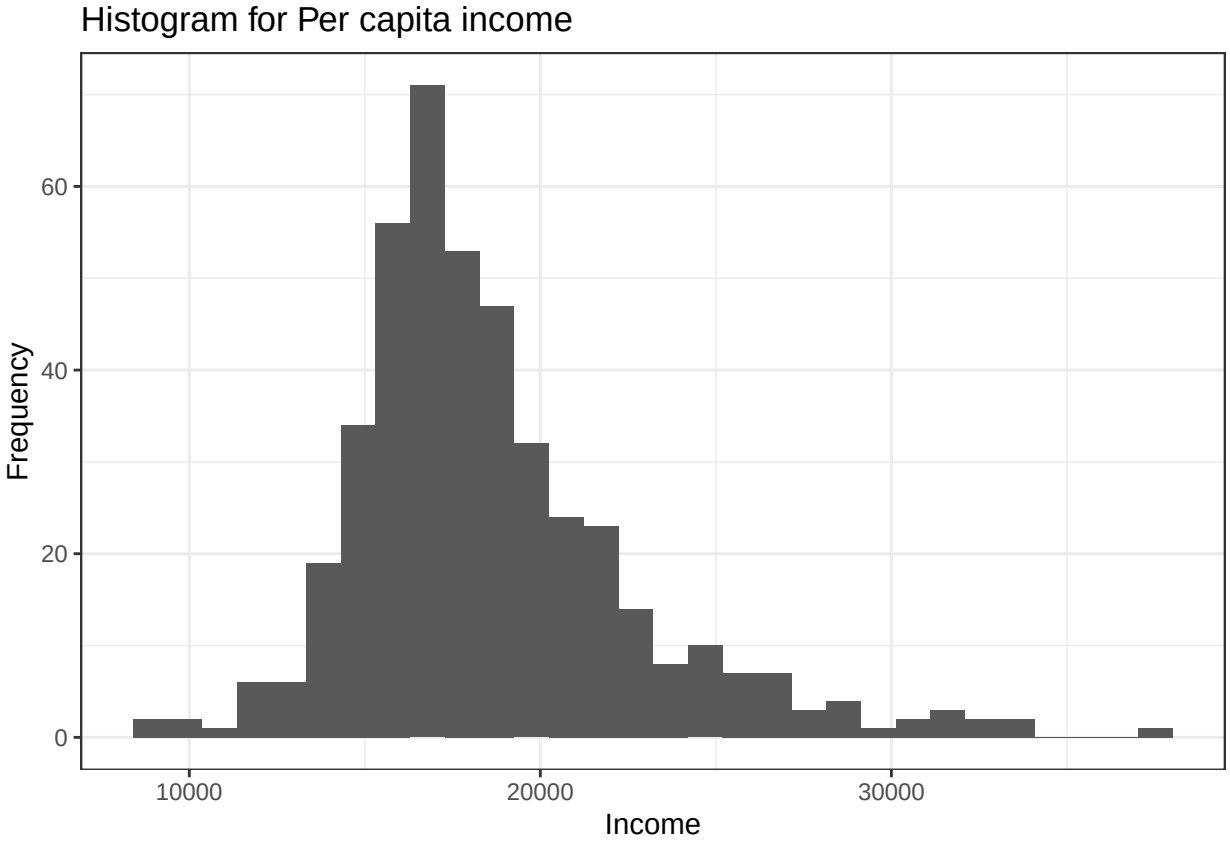

Figure 1: Histogram for Per capita income

As is shown in Figure 1, the distribution of per capital income in different counties is right skewed and peaking around 17000 dollars. Most counties have per capital income within the range of 14000 and 23000 dollars. There is one county which has a very high per capital income.

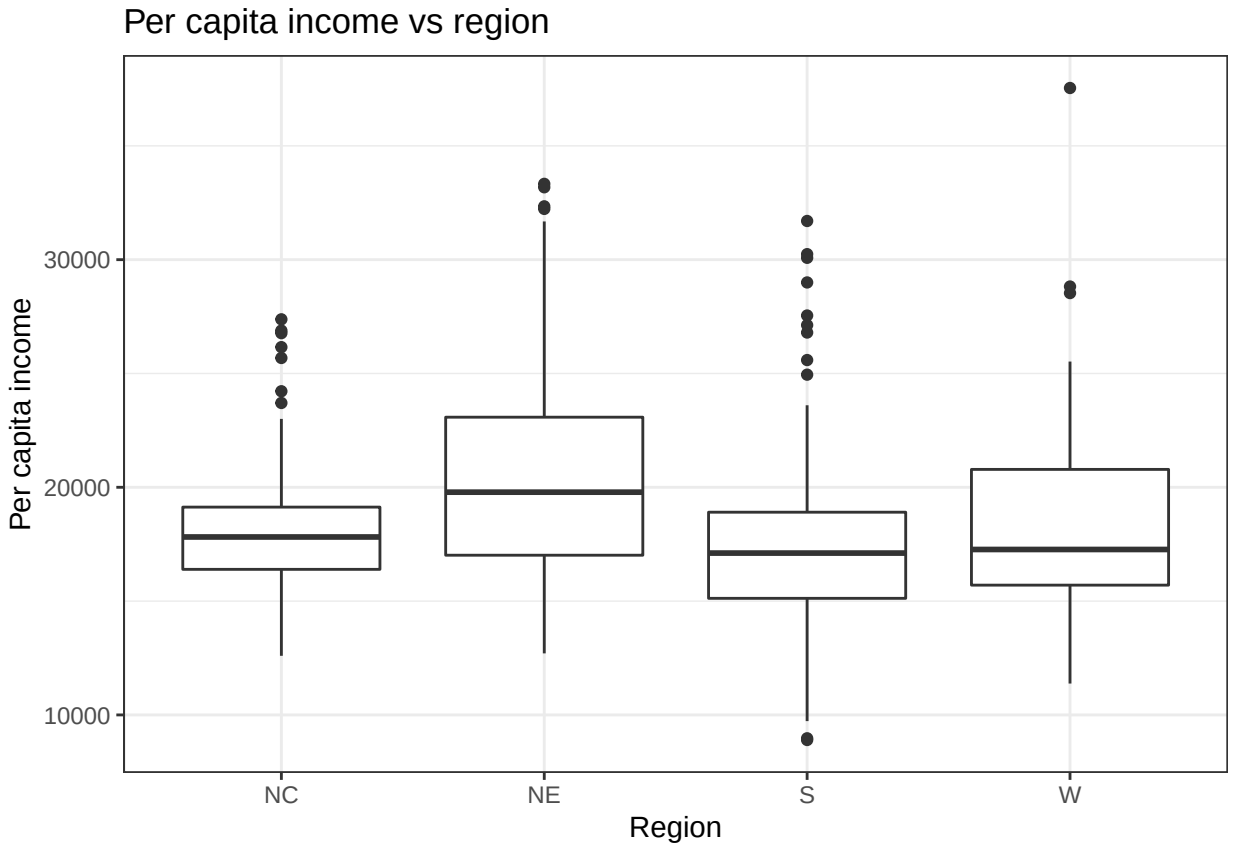

Figure 2: Boxplot for Per capita income vs region

The median per capital is the highest in northeast region, and the lowest in southern region. This is consistent to the reality that the density of big cities is the highest in northeast region and the lowest in the southern region. Per capital income is usually higher in large cities than other areas. Besides, there are some outliers for the west region with very high values. This is because of there is some rich zone in the western region.

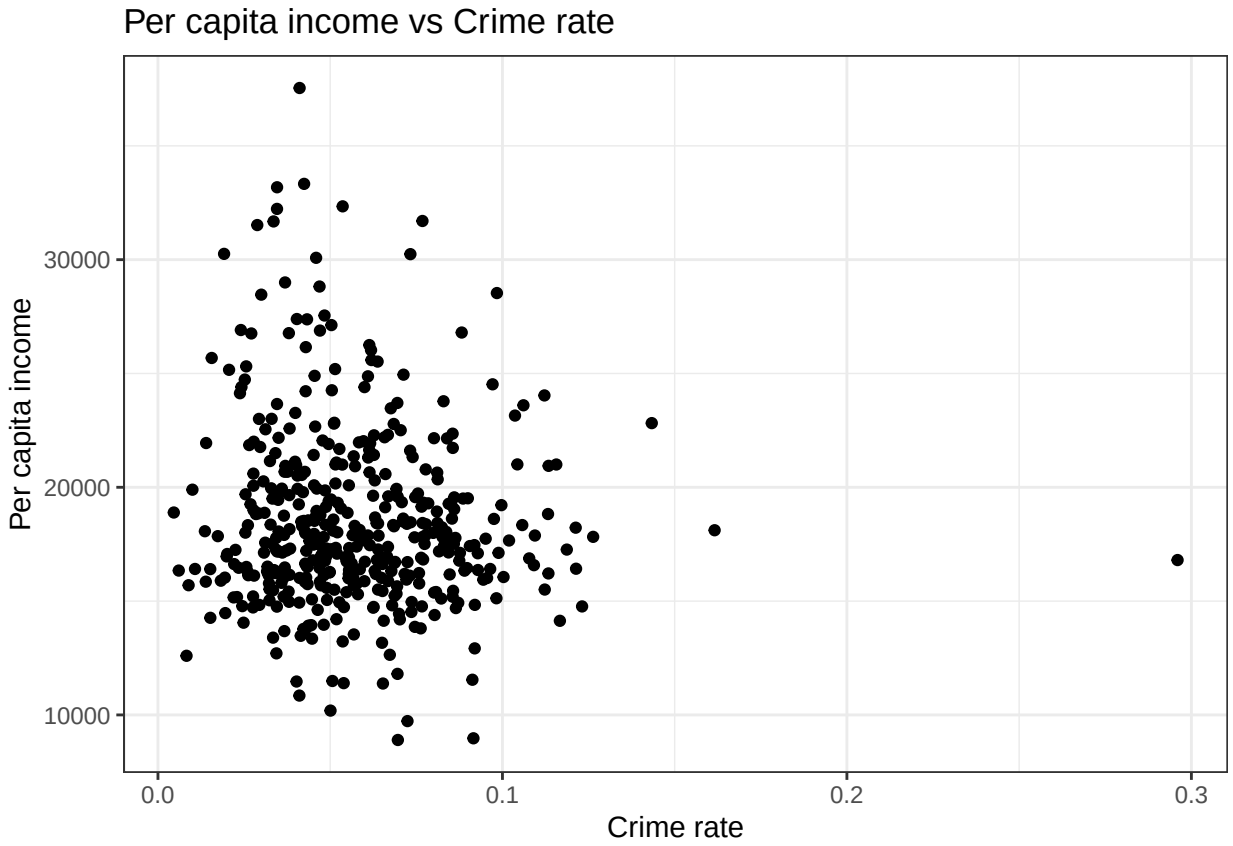

Figure 3: Per capita income vs Crime rate

By Figure 3, we can see that the counties having high crime rate usually have medium per capital income (around 20000).

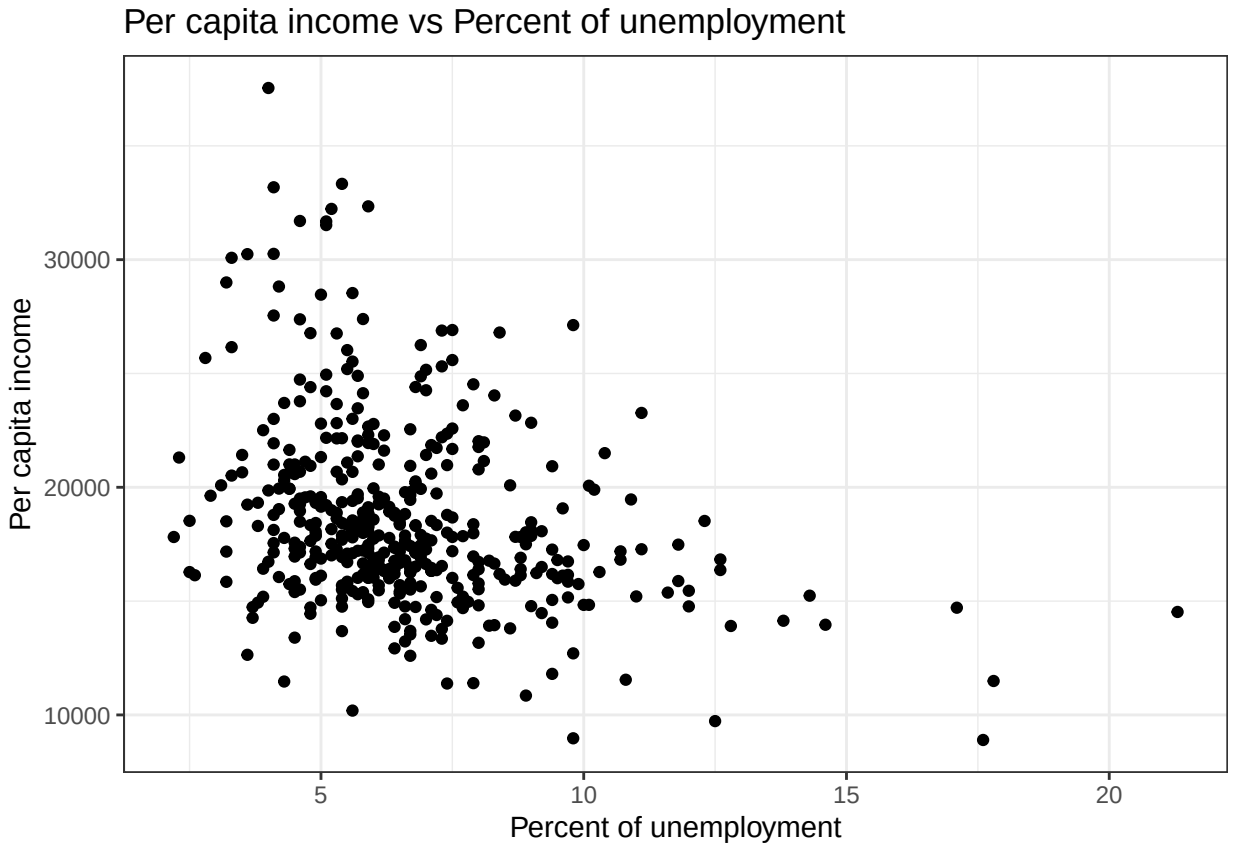

Figure 4: Per capita income vs Percent of unemployment

There is a negative relationship between per capital income and percent of unemployment. Because the unemployed people usually have no or low income, a high percentage of them will result in low per capital income.

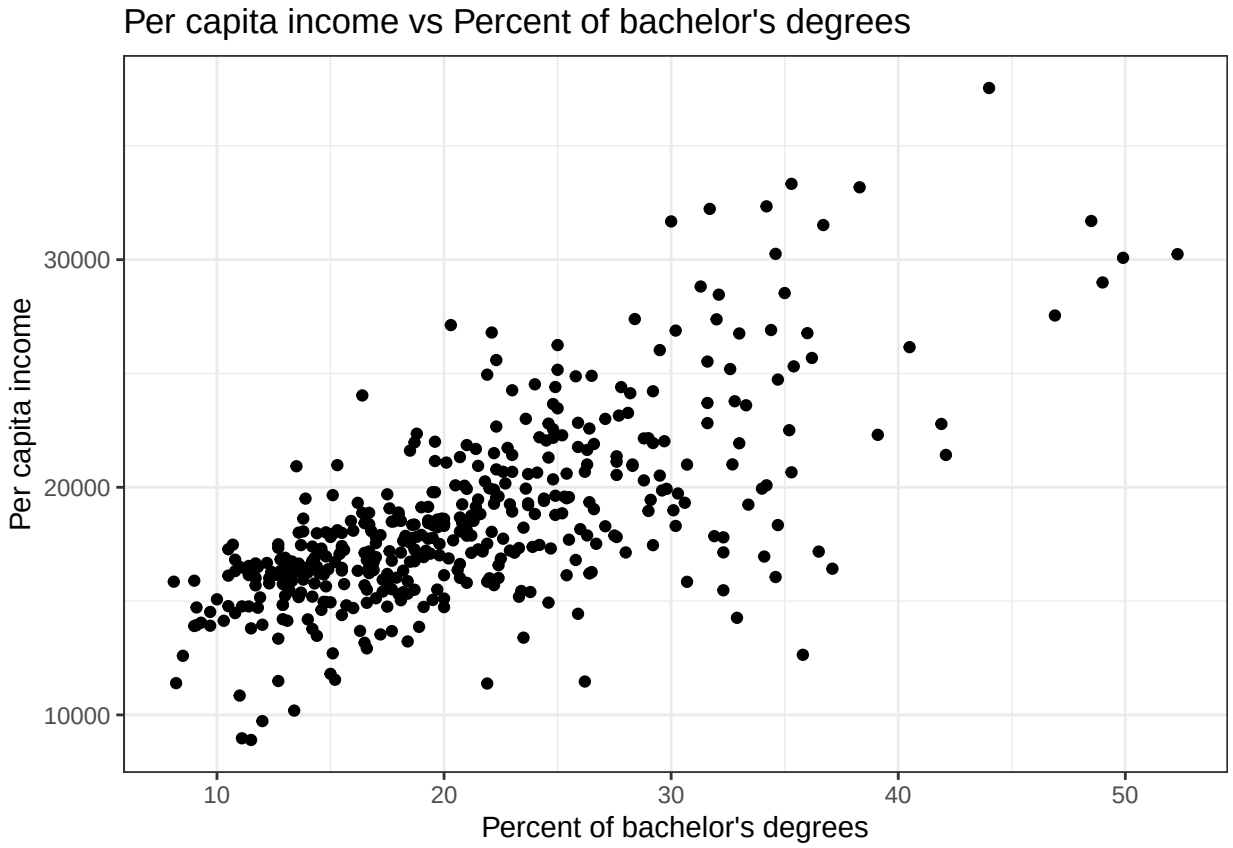

Figure 5: Per capita income vs Percent of bachelor's degrees

We can see a significant increasing trend in Figure 6, meaning that per capital income for a county is significantly correlated to the percent of adult population with a bachelor's degree in that county.

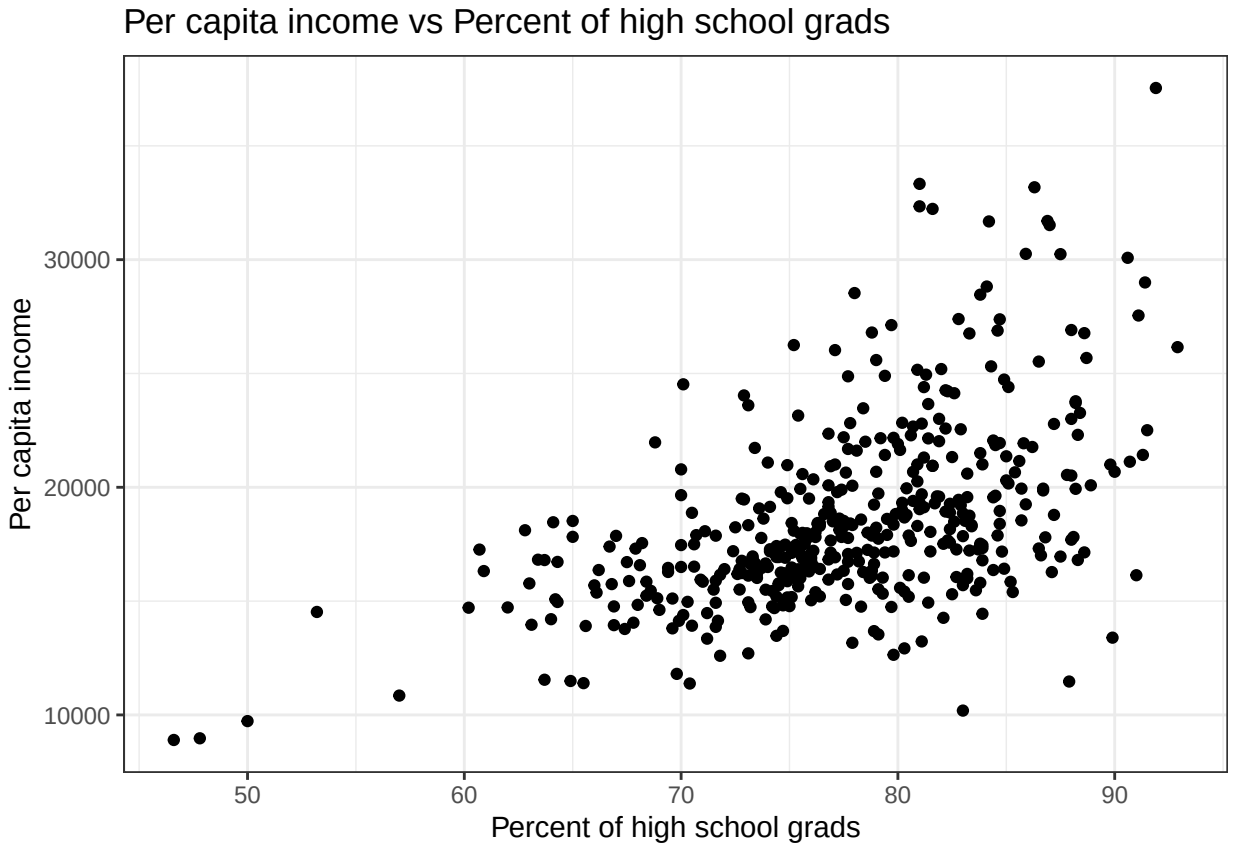

Figure 6: Per capita income vs Percent of high school grads

We can also see an increasing trend in Figure 7, together with an increasing variance with the growing percent of high school grads. We can infer that the higher percent of high school grads will not definitely result in an increase in income, but a low percent of high school grads will certainly result in a low income.

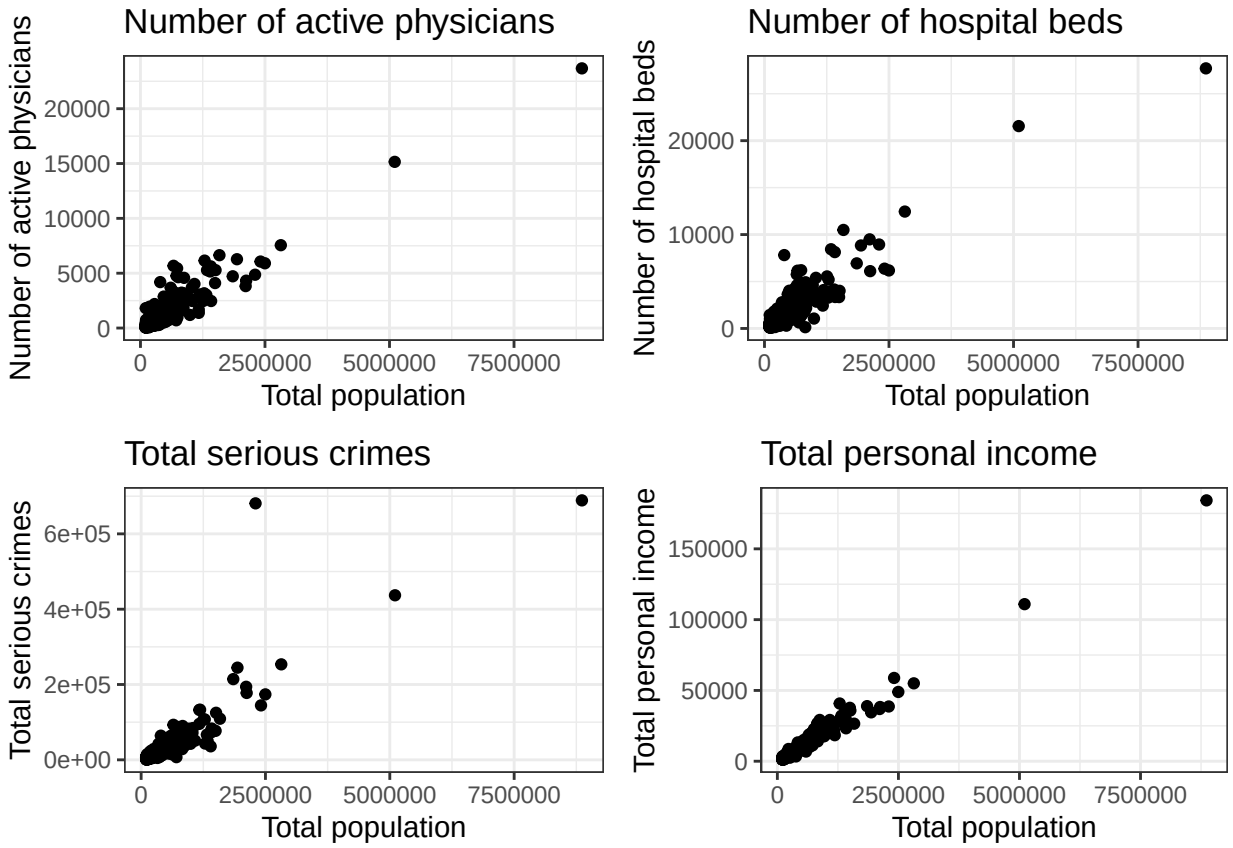

Figure 7: Total population vs Number of active physicians, Number of hospital beds, Total serious crimes, Total personal income

We can see that the factor population is correlated with all the four variables. So it is a counfounding factor to make these variable correlated with each other.

(b)

We will first compare the models with and without the interaction terms. We use a partial F test to compare the models. The null hypothesis is that the coefficient of the interaction terms is zero.

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ region + crime.rate
## Model 2: per.cap.income ~ region * crime.rate
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    435 6609753963
## 2    432 6607856753  3   1897210 0.0413 0.9888
```

By the above ANOVA table, we can see that the p-value for the partial F test is $0.98 > 0.05$. So we fail to reject the null hypothesis and confirm that there should not be any interaction terms in the model.

The final model is:

$$PerCapitalIncome = \beta_0 + \beta_1 Region + \beta_2 Crimes + \epsilon$$

Table 10: Coefficients for the model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18006.04469	537.0395	33.5283439	0.0000000
regionNE	2354.69663	541.9715	4.3446875	0.0000174
regionS	-927.44668	512.3059	-1.8103378	0.0709333
regionW	-34.92294	586.0281	-0.0595926	0.9525075
crime.rate	5773.20230	7520.4126	0.7676710	0.4430992

By Table 4, there is a positive relationship between total serious crimes and per capital income, because the coefficient for crime rate in the model is positive. However, this relationship is not significant, because the p-value for it is $0.44 > 0.05$.

Now we fit the model with the total number of crimes.

Table 11: Coefficients for the model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18106.9099510	378.4380302	47.8464333	0.0000000
regionNE	2286.0373120	532.4709260	4.2932622	0.0000217
regionS	-860.5567325	486.8305023	-1.7676722	0.0778167
regionW	-142.8266313	579.6196624	-0.2464144	0.8054777
crimes	0.0089153	0.0031877	2.7968322	0.0053895

By Table 5, we can see that the coefficient for number of crimes is also positive, meaning that there is a slightly positive relationship between total crimes and per-capita income. The coefficient is significant this time, so this model can best answer the question about the relationship between crimes and per-capita income.

(c)

Transformations

In this part, we will explore transformations for the numeric variables to make them look normal.

For the response variable `per.cap.income`, here is its histogram.

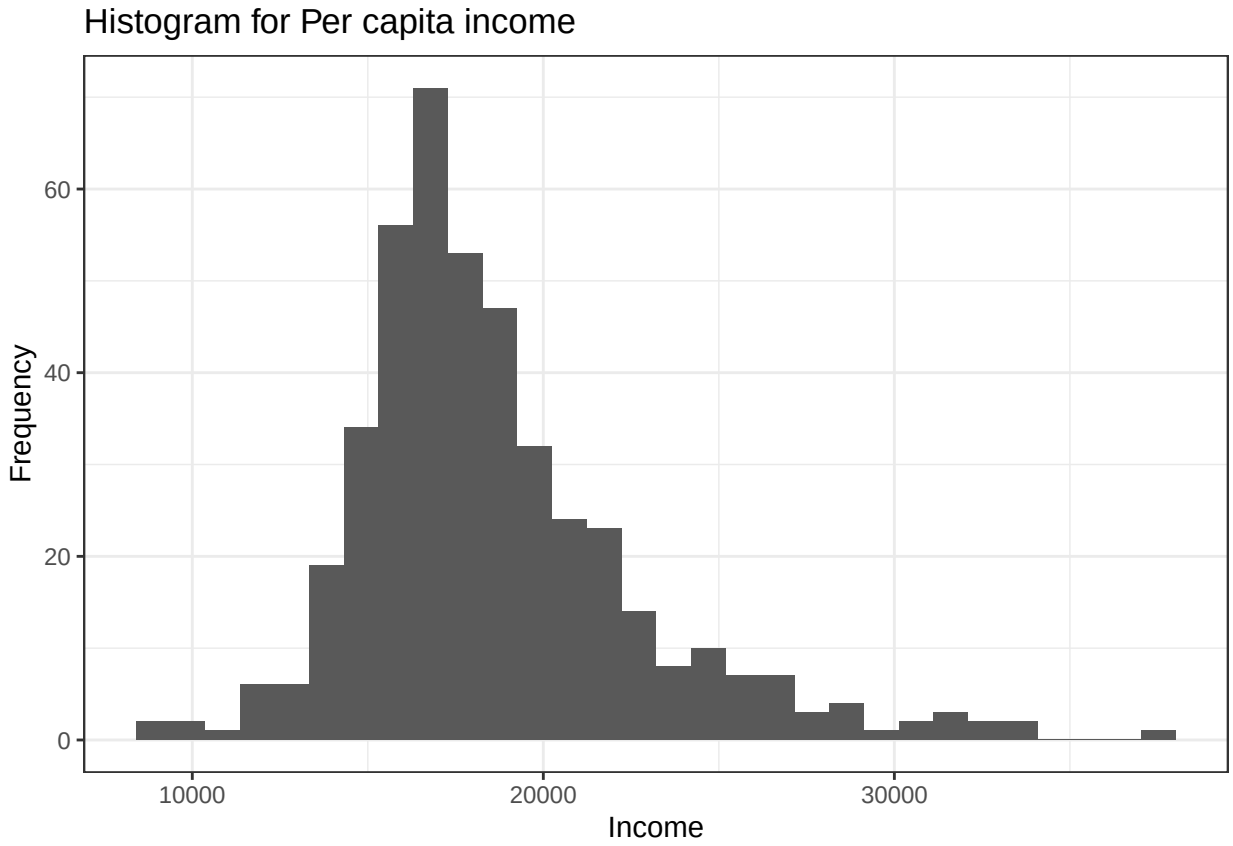

Figure 7: Histogram for Per capital income

We can see that it is right-skewed, so we need a power transform X^λ with $\lambda < 1$. From the Box-Cox method for power transform below, we pick the nearby value $\lambda = 0$. This means we choose the log transform.

```
## [1] "Power Transform Lambda: -0.368336534678286"
```

For the other variables, here is a histogram for all the variables.

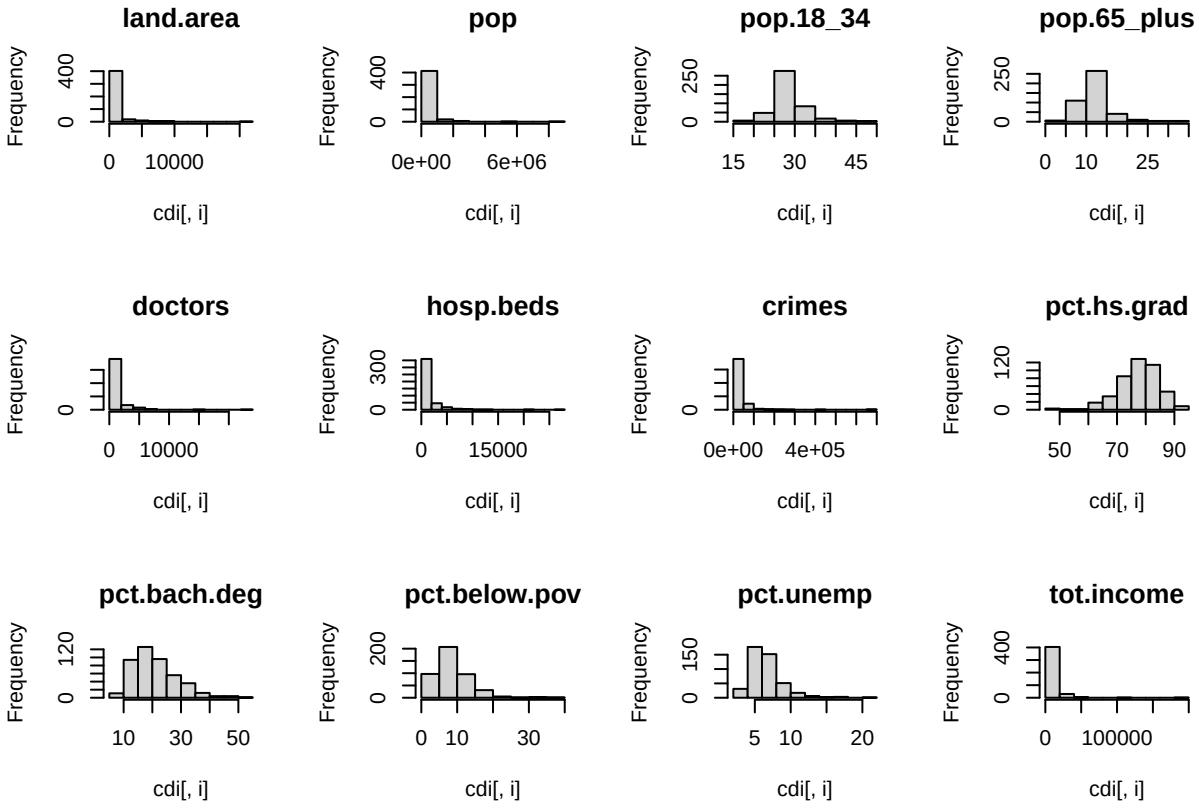

Figure 8: Histograms for variables

We can see from Figure 8 that percent of high school graduates is left-skewed, and all the other variables are right-skewed. We will use Box-Cox method to find proper power transformations.

Table 12: Box-Cox Power transforms

Variable	Lambda
land.area	0.0023048
pop	-0.5795787
pop.18_34	-0.3857010
pop.65_plus	-0.0075542
doctors	-0.2174773
hosp.beds	-0.1541052
crimes	-0.1307109
pct.hs.grad	3.0719249
pct.bach.deg	-0.0317479
pct.below.pov	0.1817562
pct.unemp	-0.1130851
tot.income	-0.4379530

In Table 6, we can see that we can use:

1. Log transform $\log(X)$ for: `land.area`, `pop.18_34`, `pop.65_plus`, `doctors`, `hosp.beds`, `crimes`, `pct.bach.deg`, `pct.below.pov`, `pct.unemp`
2. One over square root $X^{-\frac{1}{2}}$ for: `pop`, `tot.income`
3. Cube square X^3 for: `pct.hs.grad`

Variable selection

Pre-selection We will delete the variable `tot.income`, because total income divided by population (`pop`) is actually the response variable per capital income.

Collinearity In this part, we will check the collinearity condition for all the numeric variables.

Table 13: Coefficients statistics

	Estimate	Std. Error	t value	Pr(>|t|)	VIF
land.area	-0.03	0.00	-6.16	0.00	1.18
pop	68.72	18.11	3.79	0.00	10.08
pop.18_34	-0.29	0.04	-6.61	0.00	2.47
pop.65_plus	0.03	0.02	1.21	0.23	2.69
doctors	0.06	0.01	4.67	0.00	15.45
hosp.beds	0.02	0.01	1.42	0.16	9.83
crimes	0.03	0.01	2.59	0.01	7.47
pct.hs.grad	0.00	0.00	-2.81	0.01	3.84
pct.bach.deg	0.27	0.03	10.66	0.00	5.23
pct.below.pov	-0.25	0.01	-19.15	0.00	2.90
pct.unemp	0.08	0.02	4.99	0.00	1.89

We set 5 as the benchmark VIF value, and there are several correlated variables:

1. Medical conditions: `doctors` and `hosp.bed`. There is a scatterplot showing their linear relationship below (Figure 9).

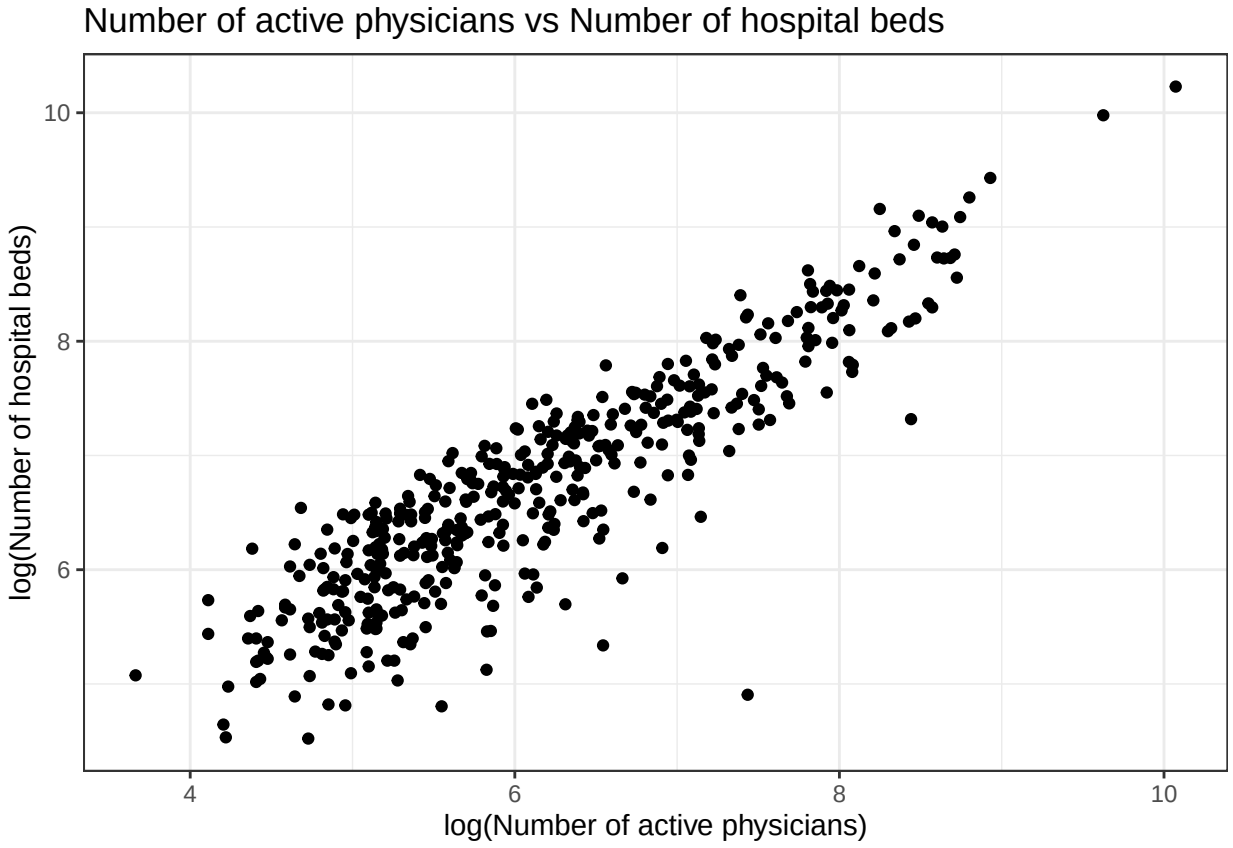

Figure 9: Number of active physicians vs Number of hospital beds

We can see that number of active physicians is highly correlated with number of hospital beds. So we delete the variable `hosp.beds` and keep `doctors`, because this variable is more significant.

2. `crimes`: We will use some scatterplots to check which variable it is correlated with. There are several candidates: `pct.below.pov`, `pct.unemp`, `pct.bach.deg`, `doctors`, `pct.hs.grad`, `pop`

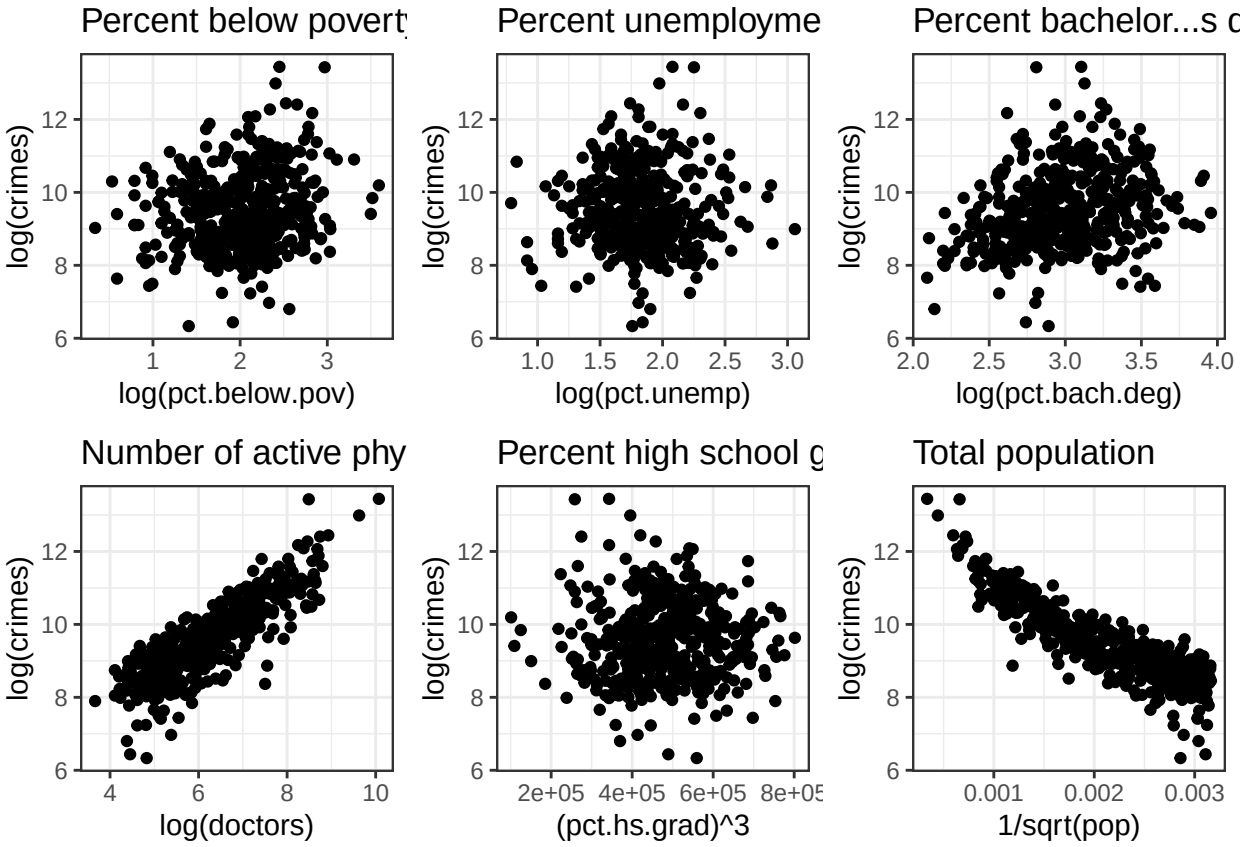

Figure 11: Collinearity check for crimes.

We can see that this variable is highly correlated with number of active physicians and total population. So we choose to delete `pop` and keep both `doctors` and `crimes`, because we are more interested in the relationship between each of them and the per-capita income.

3. `pop`: Already deleted.

4. `pct.bach.deg`: We will use some scatterplots to check which variable it is correlated with. There are several candidates: `pct.below.pov`, `pct.unemp`, `pct.hs.grad`

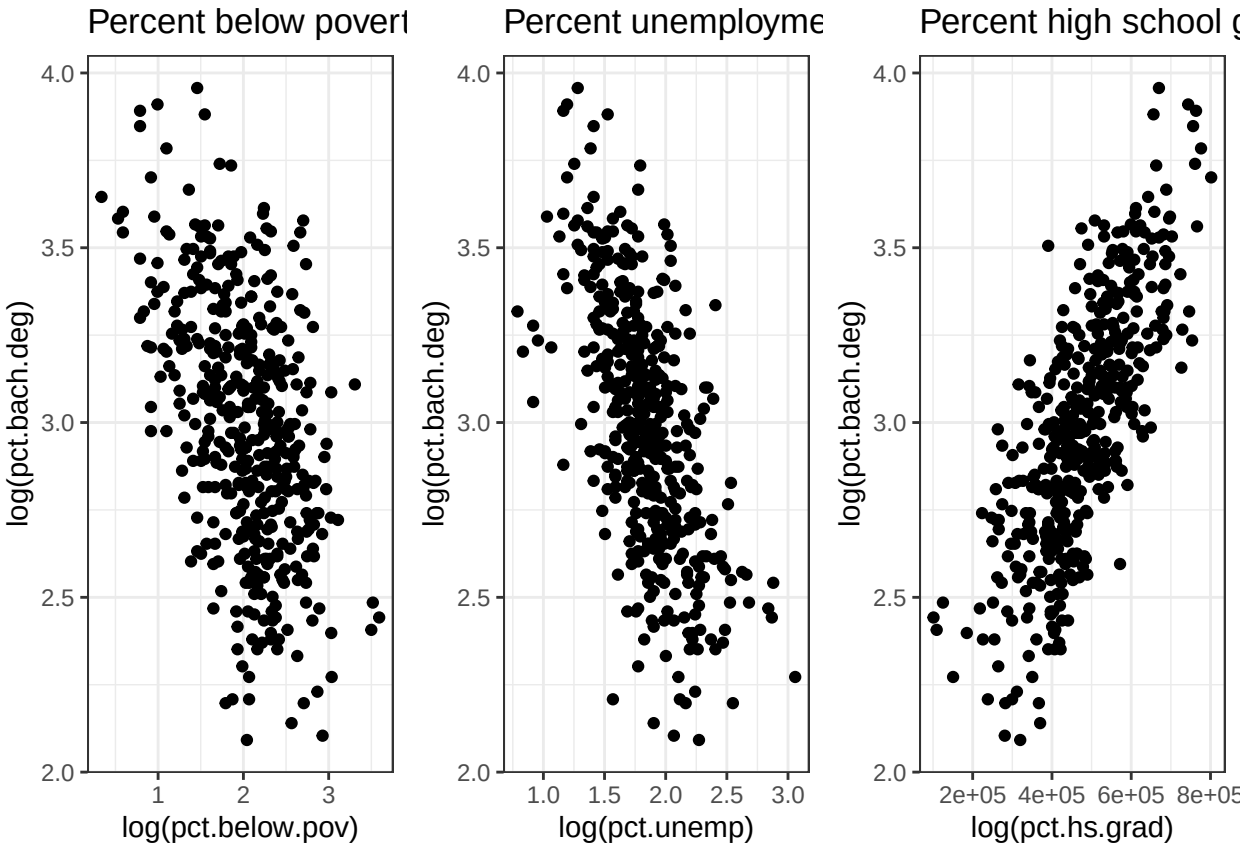

Figure 11: Collinearity check for percent bachelor's degrees.

We can see a correlation with all the three variables, so we choose to delete `pct.bach.deg`

Finally, we deleted: `hosp.beds`, `pop`, `pct.bach.deg`

BIC We can use BIC to do model selection. We chose BIC because we would like a simple model

From the result of BIC, we finally chose the model with five variables: `land.area`, `pop.18_34`, `doctors`, `pct.hs.grad`, `pct.below.pov`. The R-squared value is 0.7822, which is an acceptable value for a model with only 5 parameters.

Interaction

We would like to explore the interaction of `region` with other terms. We test the significance of the interaction term with all the variables using ANOVA table. The null hypothesis is including the interaction term does not make an improvement in the fit. We extract their p-values into a single table below.

Table 14: Interaction analysis

Variables	p_values
land.area	0.2645012
pop.18_34	0.4839704
doctors	0.2802259
pct.hs.grad	0.0004891
pct.below.pov	0.1073062

The p-value is significant for `pct.hs.grad`, meaning that we only reject the null hypithesis for this variable, so we will include that interaction term.

Final model discussion

Final model

Here are the coefficients for the final model.

Table 15: Final Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.6667050	0.1263162	84.444480	0.0000000
land.area	-0.0490628	0.0064826	-7.568407	0.0000000
pop.18_34	-0.2057823	0.0371997	-5.531833	0.0000001
doctors	0.0887292	0.0041972	21.140096	0.0000000
pct.hs.grad	0.0000001	0.0000001	1.019629	0.3084805
regionNE	-0.1135711	0.0631627	-1.798071	0.0728704
regionS	-0.0799485	0.0536723	-1.489565	0.1370751
regionW	0.1441645	0.0664288	2.170212	0.0305401
pct.below.pov	-0.2256078	0.0138861	-16.246969	0.0000000
pct.hs.grad:regionNE	0.0000003	0.0000001	2.194325	0.0287488
pct.hs.grad:regionS	0.0000002	0.0000001	1.713309	0.0873798
pct.hs.grad:regionW	-0.0000002	0.0000001	-1.687523	0.0922314

Model diagnostic check

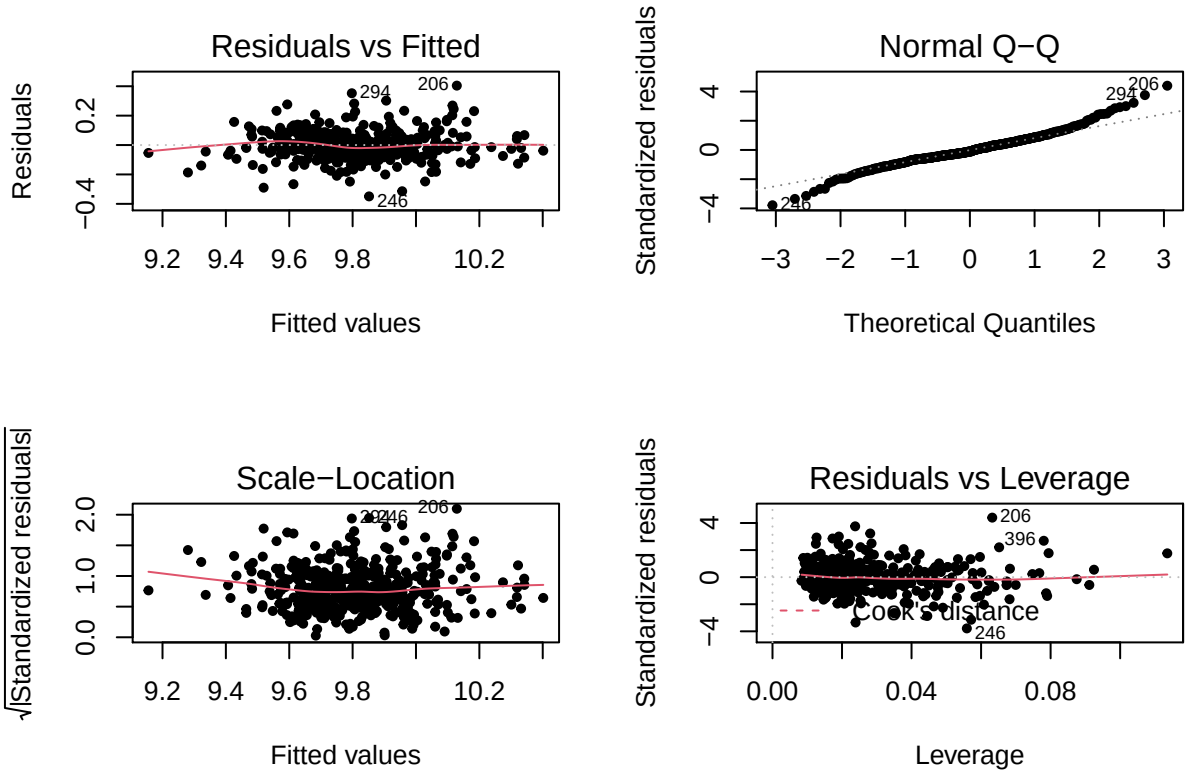

Figure 12: Model diagnostic plots

1. Residuals vs Fitted: The residuals are distributed evenly with a mean 0, with no pattern detectable.

2. Normal Q-Q: The fitted value is slightly skewed and not completely normal.
3. Scale-Location: There is no observable pattern, so the variance looks constant.
4. Residuals vs Leverage: There are no observable leverage points or outliers.

Tradeoff

1. In order to satisfy the normality condition for the variables, we transformed all the variables, which can make the model less interpretable.
2. The transformed data is still not normal, because we chose a nearby value of power for the simplicity of the model.
3. A simple model was chosen using BIC. Some significant factors, such as number of crimes, and unemployment rate were deleted. Other factors, such as population and number of hospital beds were also deleted due to collinearity conditions. This may cause a decrease in R-squared.

Codes

```
# Q1
cdi = read.table("cdi.dat", header = T)

## (a)
smry_state = data.frame(
  State = names(table(cdi$state)),
  Count = as.numeric(table(cdi$state))
)
kable(smry_state, caption = "Summary table: State")

smry_state = data.frame(
  Region = names(table(cdi$region)),
  Count = as.numeric(table(cdi$region))
)
kable(smry_state, caption = "Summary table: Geographic region")

smry_cts = tibble(
  Variable = c(),
  Min = c(),
  First.Qu = c(),
  Median = c(),
  Mean = c(),
  Third.Qu = c(),
  Max = c()
)

for (i in 4:16){
  values = as.numeric(summary(cdi[,i]))
  smry_cts = add_row(smry_cts,
    Variable = names(cdi)[i],
    Min = values[1],
    First.Qu = values[2],
    Median = values[3],
    Mean = values[4],
    Third.Qu = values[5],
    Max = values[6]
  )
```

```
}

kable(smry_cts, caption = "Summary table: Continuous variables")

cdi %>% ggplot()+
  geom_histogram(aes(per.cap.income)) +
  labs(title = "Histogram for Per capita income",
       x = "Income",
       y = "Frequency") +
  theme_bw()

cdi %>% ggplot(aes(x = region, y = per.cap.income)) +
  geom_boxplot()+
  labs(title = "Per capita income vs region",
       x = "Region",
       y = "Per capita income") +
  theme_bw()
cdi %>% ggplot(aes(x = pct.unemp, y = per.cap.income))+
  geom_point() +
  labs(title = "Per capita income vs Percent of unemployment",
       x = "Percent of unemployment",
       y = "Per capita income") +
  theme_bw()

cdi %>% ggplot(aes(x = pct.bach.deg, y = per.cap.income))+
  geom_point() +
  labs(title = "Per capita income vs Percent of bachelor's degrees",
       x = "Percent of bachelor's degrees",
       y = "Per capita income") +
  theme_bw()

cdi %>% ggplot(aes(x = pct.hs.grad, y = per.cap.income))+
  geom_point() +
  labs(title = "Per capita income vs Percent of high school grads",
       x = "Percent of high school grads",
       y = "Per capita income") +
  theme_bw()

## (b)

cdi_more = cdi
cdi_more$crime.rate = cdi$crimes/cdi$pop
lm_1b1 = lm(per.cap.income ~ region + crime.rate, data = cdi_more)
lm_1b2 = lm(per.cap.income ~ region * crime.rate, data = cdi_more)
anova(lm_1b1, lm_1b2)

kable(summary(lm_1b1)$coef, caption = "Coefficients for the model")

lm_1b3 = lm(per.cap.income ~ region + crimes, data = cdi)
kable(summary(lm_1b3)$coef, caption = "Coefficients for the model")


## (c)
```

```r
### Transformations

cdi %>% ggplot()+
  geom_histogram(aes(per.cap.income)) +
  labs(title = "Histogram for Per capita income",
       x = "Income",
       y = "Frequency") +
  theme_bw()

paste("Power Transform Lambda:",powerTransform(cdi$per.cap.income ~ 1)$lambda)

par(mfrow = c(3,4))
for (i in c(4:14,16)){
  hist(cdi[,i], main = names(cdi)[i])
}

box_cdi = tibble(
  Variable = c(),
  Lambda = c()
)

for (i in c(4:14,16)){
  box_cdi = add_row(box_cdi, Variable = names(cdi)[i],
                    Lambda = powerTransform(cdi[,i] ~ 1)$lambda)
}

kable(box_cdi, caption = "Box-Cox Power transforms")

cdi_adj = cdi[,-c(1:2)] %>% mutate(
  per.cap.income = log(per.cap.income),
  land.area = log(land.area),
  pop.18_34 = log(pop.18_34),
  pop.65_plus = log(pop.65_plus),
  doctors = log(doctors),
  hosp.beds = log(hosp.beds),
  crimes = log(crimes),
  pct.bach.deg = log(pct.bach.deg),
  pct.below.pov = log(pct.below.pov),
  pct.unemp = log(pct.unemp),
  pop = 1/sqrt(pop),
  tot.income = 1/sqrt(tot.income),
  pct.hs.grad = (pct.hs.grad)^3
)


### Variable selection

#### Pre-selection

cdi_adj = cdi_adj[,-14]

#### Collinearity
```

```r
lm_cdi_num = lm(per.cap.income ~ ., data = cdi_adj[,c(2:13)])
VIF = vif(lm_cdi_num)
cdi_vif_table = cbind(summary(lm_cdi_num)$coef[-1,], VIF)
kable(round(cdi_vif_table,2), caption = "Coefficients statistics")

cdi_adj %>% ggplot(aes(x = doctors, y = hosp.beds))+
  geom_point() +
  labs(title = "Number of active physicians vs Number of hospital beds",
       x = "log(Number of active physicians)",
       y = "log(Number of hospital beds)") +
  theme_bw()

p1 = cdi_adj %>% ggplot(aes(x = pct.below.pov, y = crimes))+
  geom_point() +
  labs(title = "Percent below poverty level",
       x = "log(pct.below.pov)",
       y = "log(crimes)") +
  theme_bw()

p2 = cdi_adj %>% ggplot(aes(x = pct.unemp, y = crimes))+
  geom_point() +
  labs(title = "Percent unemployment",
       x = "log(pct.unemp)",
       y = "log(crimes)") +
  theme_bw()

p3 = cdi_adj %>% ggplot(aes(x = pct.bach.deg, y = crimes))+
  geom_point() +
  labs(title = "Percent bachelor's degrees",
       x = "log(pct.bach.deg)",
       y = "log(crimes)") +
  theme_bw()

p4 = cdi_adj %>% ggplot(aes(x = doctors, y = crimes))+
  geom_point() +
  labs(title = "Number of active physicians",
       x = "log(doctors)",
       y = "log(crimes)") +
  theme_bw()

p5 = cdi_adj %>% ggplot(aes(x = pct.hs.grad, y = crimes))+
  geom_point() +
  labs(title = "Percent high school graduates",
       x = "(pct.hs.grad)^3",
       y = "log(crimes)") +
  theme_bw()

p6 = cdi_adj %>% ggplot(aes(x = pop, y = crimes))+
  geom_point() +
  labs(title = "Total population",
       x = "1/sqrt(pop)",
       y = "log(crimes)") +
  theme_bw()
```

```r
grid.arrange(p1,p2,p3,p4,p5,p6,nrow = 2, ncol = 3)

p1 = cdi_adj %>% ggplot(aes(x = pct.below.pov, y = pct.bach.deg))+
  geom_point() +
  labs(title = "Percent below poverty level",
       x = "log(pct.below.pov)",
       y = "log(pct.bach.deg)") +
  theme_bw()

p2 = cdi_adj %>% ggplot(aes(x = pct.unemp, y = pct.bach.deg))+
  geom_point() +
  labs(title = "Percent unemployment",
       x = "log(pct.unemp)",
       y = "log(pct.bach.deg)") +
  theme_bw()

p3 = cdi_adj %>% ggplot(aes(x = pct.hs.grad, y = pct.bach.deg))+
  geom_point() +
  labs(title = "Percent high school graduates",
       x = "log(pct.hs.grad)",
       y = "log(pct.bach.deg)") +
  theme_bw()

grid.arrange(p1,p2,p3,nrow = 1, ncol = 3)

#### BIC

cdi_adj = cdi_adj[,-c(3,7,10)]
lm_cdi_full = lm(per.cap.income ~ ., data = cdi_adj)
bic_cdi = stepAIC(lm_cdi_full, direction = 'both', k = log(nrow(cdi_adj)))

### Interaction

lm_cdi_r0 = lm(per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad + pct.below.pov,
            data = cdi_adj)
lm_cdi_r1 = lm(per.cap.income ~ land.area*region + pop.18_34 + doctors + pct.hs.grad + pct.below.pov,
            data = cdi_adj)
lm_cdi_r2 = lm(per.cap.income ~ land.area + pop.18_34*region + doctors + pct.hs.grad + pct.below.pov,
            data = cdi_adj)
lm_cdi_r3 = lm(per.cap.income ~ land.area + pop.18_34 + doctors*region + pct.hs.grad + pct.below.pov,
            data = cdi_adj)
lm_cdi_r4 = lm(per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad*region + pct.below.pov,
            data = cdi_adj)
lm_cdi_r5 = lm(per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad + pct.below.pov*region,
            data = cdi_adj)
inter_table = tibble(
  Variables = c("land.area", "pop.18_34", "doctors", "pct.hs.grad", "pct.below.pov"),
  p_values = c(anova(lm_cdi_r1,lm_cdi_r0)[2,6], anova(lm_cdi_r2,lm_cdi_r0)[2,6],
               anova(lm_cdi_r3,lm_cdi_r0)[2,6], anova(lm_cdi_r4,lm_cdi_r0)[2,6],
               anova(lm_cdi_r5,lm_cdi_r0)[2,6])
)

kable(inter_table, caption = "Interaction analysis")
```

```
### Final model discussion

#### Final model

lm_cdi = lm_cdi_r4
kable(summary(lm_cdi)$coef, caption = "Final Model")

#### Model diagnostic check

par(mfrow = c(2,2))
plot(lm_cdi, pch = 20)
```