final_project

Student 1 & Student 2

2024-12-01

0. Contribution Statement

Student 1

Student 1 mostly worked on questions ...

Student 2

Student 2 mostly worked on questions ...

Introduction

Cities provide a glimpse into the local population of an area and are reflective of the culture and lifestyles in the region in which they are located. However, not all cities are created equal, as different socieoeconomic factors within the population and the surrounding region impact the quality of life of residents.

Data

Our primary dataset consists of a collection of various studies done by the CDC in 2018 that detail various characteristics of a particular city and its demographics, ranging from the percentage of people that are unemployed to the percentage of civilian noninstitutionalized population with a disability. There are 72836 entries.

We include a secondary dataset in our **Advanced Analysis** section of our report that consists of the climate action surveys of major international corporations and if they responded to the survey or not. There are two surveys collected (Water and Climate Change) between 2018 and 2020, inclusive.

Objective

Our objective is to analyze and document any causal relationships between features in the dataset, with an emphasis on per capita income (`EP_PCI`) and unemployment rate (`EP_UNEMP`).

Basic Analysis

Question 1: SIMPLE REGRESSION - Fit a regression line of the data predicting a city's EP_PCI (estimated per capita income)* based on the estimated proportion of unemployment EP_UNEMP. How well does the regression line fit this relationship?

Methods

Data Cleaning & Preparation

0%	25%	50%	75%	100%
-999	3.3	5.2	8	100

```
## [1] 546
```

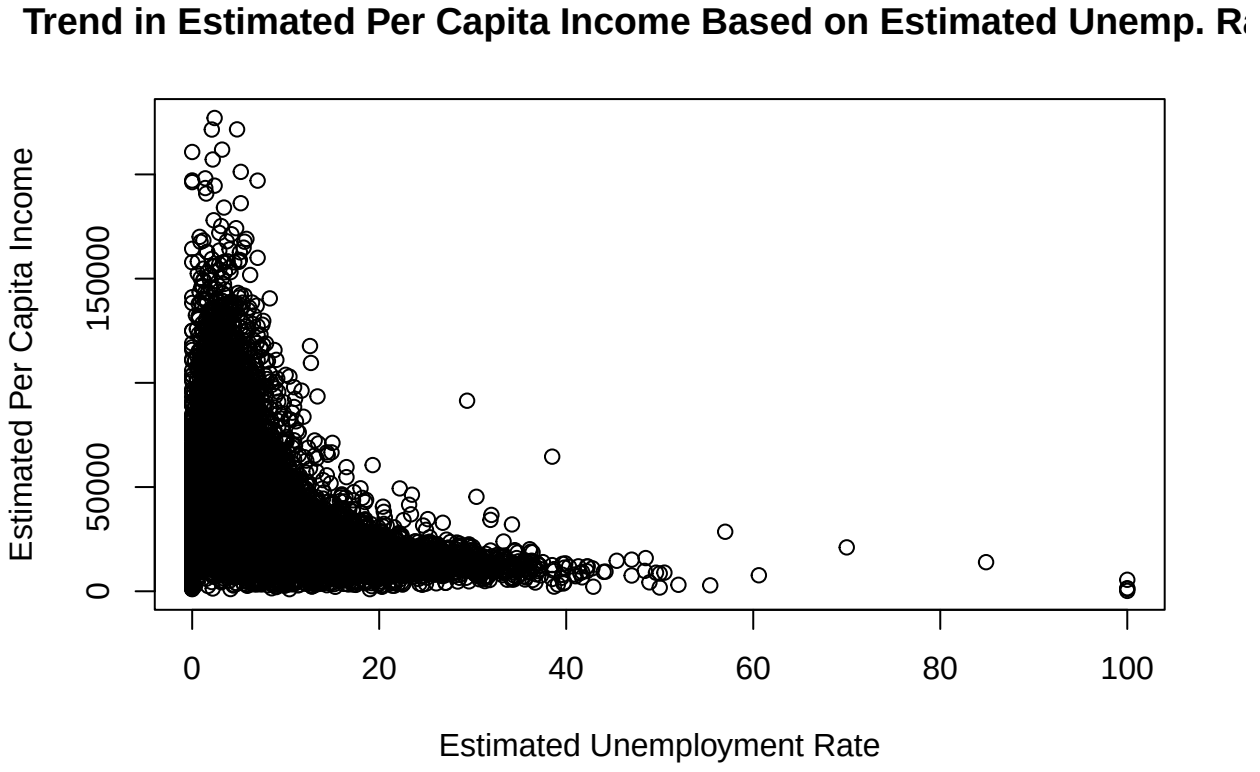
0%	25%	50%	75%	100%
-999	21571	28460	38120	227064

```
## [1] 481
```

We want to create a regression model for EP_PCI, the estimated per capita income, based on the estimated proportion of unemployment, EP_UNEMP. We are predicting EP_PCI because it describes the *rate* of unemployment and is normalized against population, unlike E_PCI, which describes the estimate *count* the unemployed.

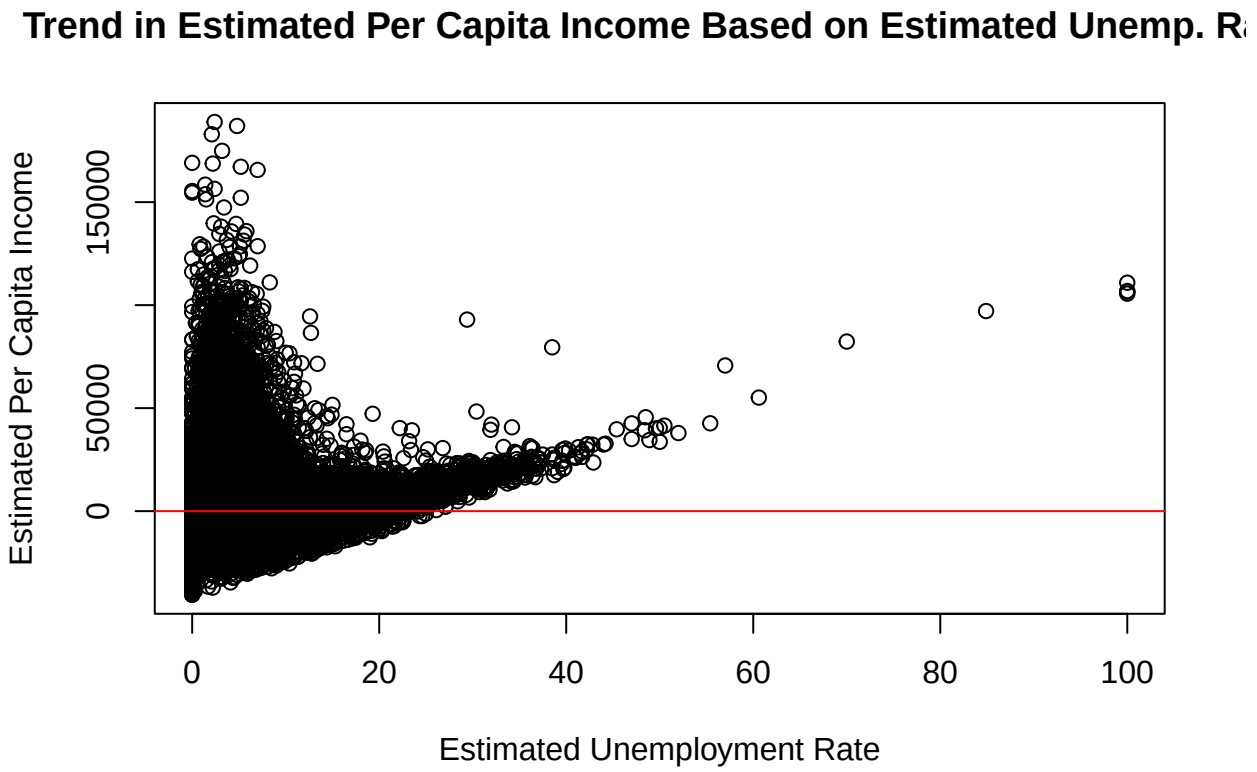
We will check the conditions for fitting a simple linear regression: linearity, homoscedasticity, independence, and normality.

1. **Linearity** - Is there a linear relationship between our explanatory variable, EP_UNEMP and our response variable, EP_PCI?



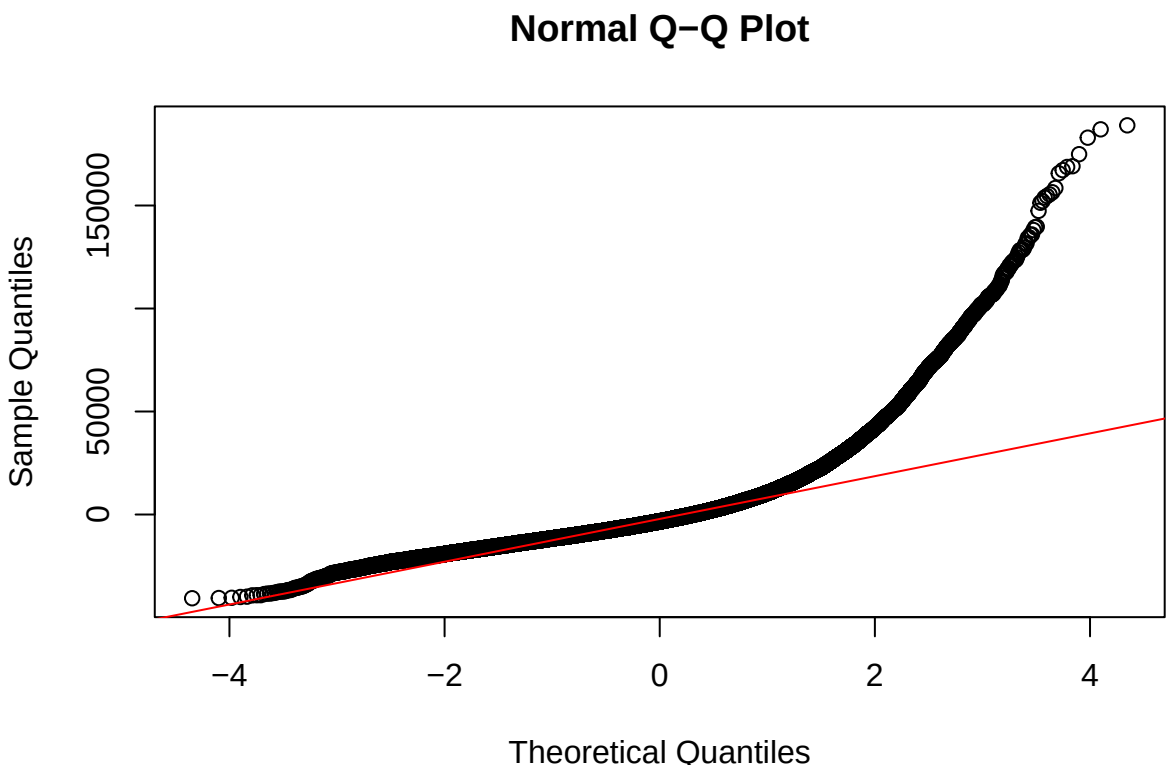
From our scatterplot, there appears to be a nonlinear negative trend between unemployment rate and estimated per capita income. As the estimated unemployment rate increases, the estimated per capita income exponentially decreases.

2. **Homoscedasticity**: Would the variance for the estimated unemployment rate be constant if we were to use a linear model?



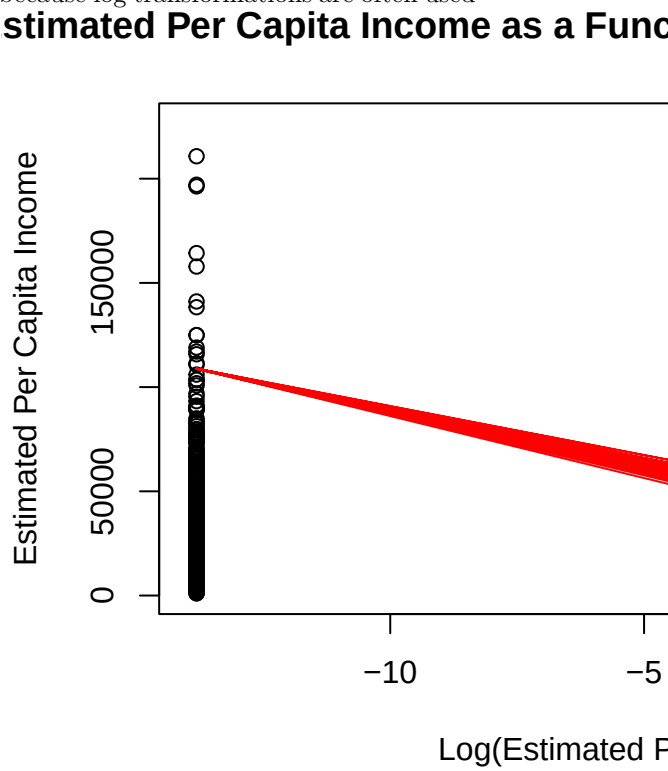
The residuals do *not* have constant variance and so, the data is not homoscedastic.

3. **Independence**: Because the residuals exhibit a clear pattern, independence may be violated.
4. **Normality**: A QQ plot of our residuals shows our data deviates from the y = x line. Its slight curve upward suggests the distribution of our residuals are right-skewed.



Analysis

To transform the data, we will take the log of `EP_UNEMP` and `EP_PCI` because log transformations are often used



to stabilize variance and help make data more suitable for regression.

With our log-transformed model, we have an MSE of about 0.201, which represents the averages squared difference between the actual values of the estimated per capita income and the predicted values. This

number is pretty low, which indicates our model performed well. However, the MSE of our original model is 301,671,723, which means our model performed very poorly without the log transformation.

Conclusion

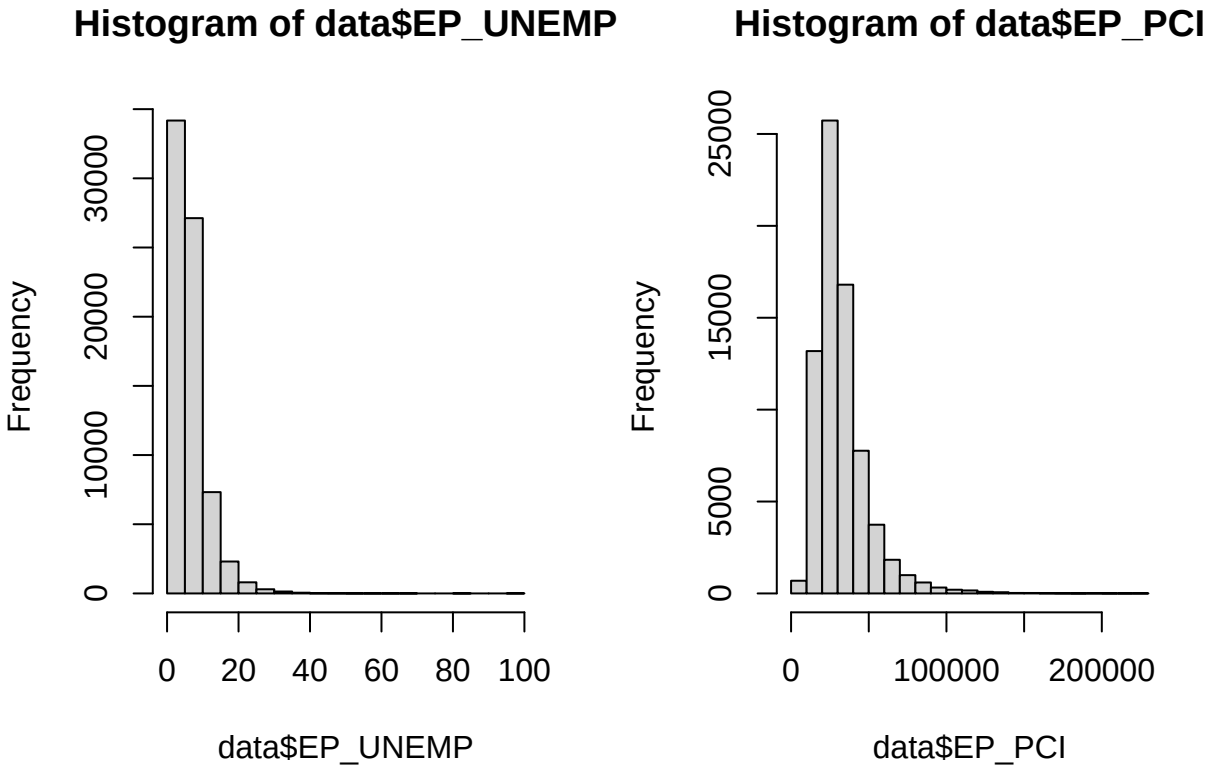
Based on these findings, the relationship between the estimated unemployment rate (EP_UNEMP) and the estimated per capita income (EP_PCI) do not meet the assumptions needed to fit a simple linear regression model without a transformation:

1. **Linearity:** The scatterplot shows a nonlinear, exponential relationship between EP_UNEMP and EP_PCI, which suggests a simple linear model is not the most effective.
2. **Homoscedasticity:** The residuals from the untransformed model show clear patterns and non-constant variance, which reaffirms the idea that our model does not perform well without a log transformation.
3. **Independence:** Similar to the point addressed in #2, the residual plots show clear patterns, which indicate a basic linear model is not suitable for this distribution.
4. **Normality**: The QQ Plot demonstrates a slight right skew, which means that the residuals from a simple linear model are not normal, and so a transformation is necessary.

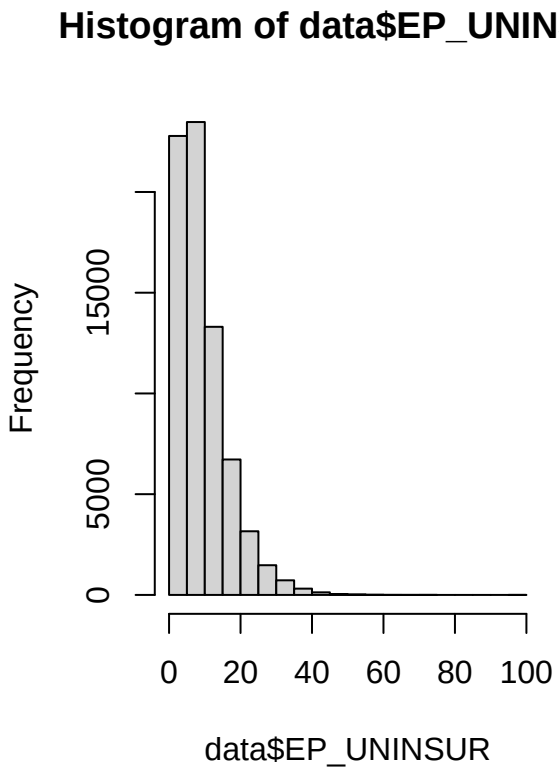
Question 2: Distribution Analysis - How do the distributions of `EP_PCI` and `EP_UNEMP` compare to one another and among other features?

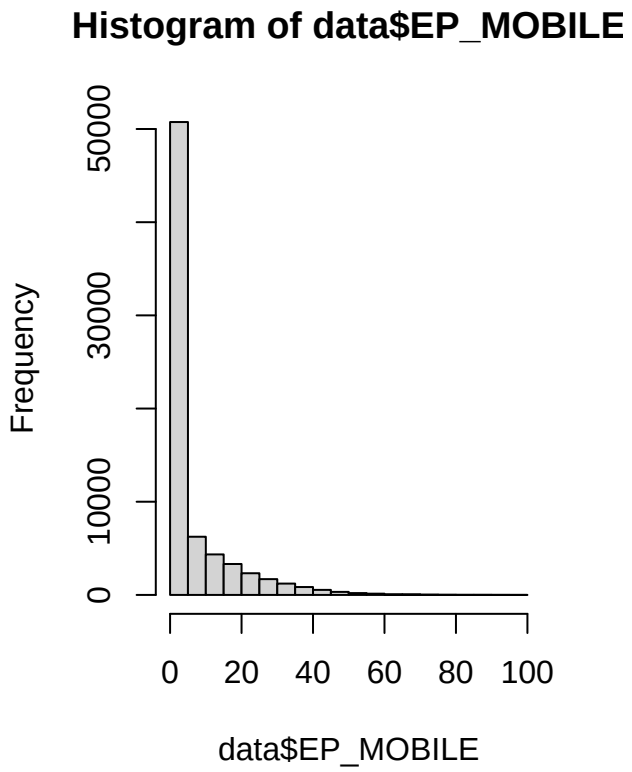
Methods

To see the distributions of the `EP_UNEMP` and `EP_PCI`, we graphed each feature as a histogram.



Additionally, we compared these distributions to other features that we believe would be related to this linear model, such as the percentage of people who are uninsured (`EP_UNINSUR`) and the percentage of mobile homes in the city (`EP_MOBILE`).





Analysis

We will conduct a **KS Analysis** to test whether the distribution of these features in our dataset are statistically similar.

- **H0:** The distribution of both features are the same.
- **H1:** The distribution of the features differ.

We meet the conditions to perform this test because these features are independent from one another and the data is continuous.

Table 3: Exact two-sample Kolmogorov-Smirnov test: `data$EP_UNEMP` and `data$EP_PCI`

Test statistic	P value	Alternative hypothesis
1	NA NA	two-sided

Table 4: Exact two-sample Kolmogorov-Smirnov test: `data$EP_UNEMP` and `data$EP_UNINSUR`

Test statistic	P value	Alternative hypothesis
0.2257	0 * * *	two-sided

Table 5: Exact two-sample Kolmogorov-Smirnov test: `data$EP_UNEMP` and `data$EP_MOBILE`

Test statistic	P value	Alternative hypothesis
0.5283	0 * * *	two-sided

Table 6: Exact two-sample Kolmogorov-Smirnov test: `data$EP_PCI` and `data$EP_UNINSUR`

Test statistic	P value	Alternative hypothesis
1	NA NA	two-sided

Table 7: Exact two-sample Kolmogorov-Smirnov test: `data$EP_PCI` and `data$EP_MOBILE`

Test statistic	P value	Alternative hypothesis
1	NA NA	two-sided

Conclusion

When looking at our visualizations for the features, it appears that are is no significant difference between their distributions, as they are all right-skewed. However, our results from the KS test reflect the opposite conclusion, as the test statistics for each test are less than the test statistic at alpha = 0.05. As a result, we cannot conclude if the distributions of these features are statistically similar, and therefore we cannot conclude that they affect the value of `EP_UNEMP`.

Question 3: Hypothesis Testing - Do cities in states with higher percentages of `EP_PCI` (estimated per capita income) have significantly fewer climate action responses compared to those with lower percentages?

Methods

H0: There is no significant difference between the number of climate action responses in states with higher percentages of EP_PCI compared to states with lower percentages. H1: States with higher percentages of EP_PCI have significantly *fewer* climate action responses compared to states with lower percentages of EP_PCI.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

We have just created a DataFrame that shows the state, whether they took a climate or water action, and the EP_PCI (estimated per capita income) for each city recorded. NOTE: Cities are not included in our merged DataFrame since we are just interested in the cities' states.

Analysis

The summary_df keeps track of the number of climate change-related actions taken for various cities grouped by state and PCI status (either higher or lower than the median value).

```
##
##  Shapiro-Wilk normality test
##
## data:  higher_counts
## W = 0.41224, p-value = 1.8e-11
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lower_counts
## W = 0.45169, p-value = 4.79e-11
```

Neither of these groups are normally distributed, so we need to use a Mann-Whitney U Test for comparing the mean counts of climate action for the groups with EP_PCI lower vs. higher than the median EP_PCI.

```
##
##  Wilcoxon rank sum exact test
##
## data:  count by pci_status
## W = 769, p-value = 0.3852
## alternative hypothesis: true location shift is less than 0
```

Because (p = 0.3852) > (alpha = 0.05), we fail to reject the null hypothesis. Cities in states with higher percentages of EP_PCI *do not* have significantly fewer climate action responses compared to those with lower percentages.

Conclusion

We merged the SVI Data with the city data and performed data cleaning to keep track of the number of cities with an EP_PCI status higher or lower than the median EP_PCI status. Since Shapiro-Wilk tests showed that our data for each group - Higher and Lower - were both not normally distributed, we had to use a Mann Whitney U Test (aka Wilcoxon Rank Sum Test) to compare the two groups since it is a non-parametric statistical test. Since we ended with a p-value of $0.3852 > 0.05$, we fail to reject the null hypothesis. This means that cities in states with higher percentages of EP_PCI *do not* have significantly fewer climate action responses compared to those with lower percentages.

Question 4: Correlation Analysis - Is there a significant correlation between the `EP_PCI` and `EP_UNEMP`? How strong is the relationship, and what does it suggest about the role of unemployment in estimated per capita income?

Methods

To understand the correlation between these two features, we calculate the covariance for the features.

```
## [1] -0.4062332
```

Analysis

Next, we test the significance of their correlation with the following hypotheses:

- **H0:** The correlation between `EP_PCI` and `EP_UNEMP` is equal to zero.
- **H1:** The correlation between `EP_PCI` and `EP_UNEMP` is not equal to zero.

Table 8: Pearson's product-moment correlation: `data$EP_PCI` and `data$EP_UNEMP`

Test statistic	df	P value	Alternative hypothesis	cor
-119.4	72171	0 * * *	two.sided	-0.4062

The correlation coefficient between `EP_PCI` and `EP_UNEMP` is **-0.4062332**, which signifies a somewhat strong negative correlation between the two features. After performing a significance test on this coefficient, it is clear that it is statistically significant, as the test statistic is **-119.4** and the p-value is **0**. The negative sign in the test statistic shows that the correlation between these two features is negative, with its large magnitude reinforcing this relationship. Additionally, the p-value tells us to reject the null hypothesis, providing evidence that the correlation between the two feature is not equal to zero.

Conclusion

We can conclude that there is a statistically significant negative correlation between `EP_PCI` and `EP_UNEMP`, which is supported by the significance test we performed on the calculated correlation coefficient. This shows us that as unemployment increases within a city's population, the city's per capita income is expected to decrease. Additionally, the relationship between these two features is rather strong, as the magnitude of the correlation coefficient is a signiificant magnitude below zero.

Advanced Analysis: Multiple Regression - Fit a regression line of the data predicting a city's `EP_PCI` (estimated per capita income) based on the estimated proportion of unemployment `EP_UNEMP`, city population, and city density. How well does the regression line fit this relationship?

Methods

To fit a regression line as specified above, we will need the following variables from the `data` dataset:

- Dependent Variable: EP_PCI - estimated per capita income
- Independent Variables:
  - `EP_UNEMP` - estimated unemployment rate
  - `E_TOTPOT` - estimated population
  - city density - need to create new column based on `E_TOTPOT` / `AREA_SQMI` (city population / city area in square miles)

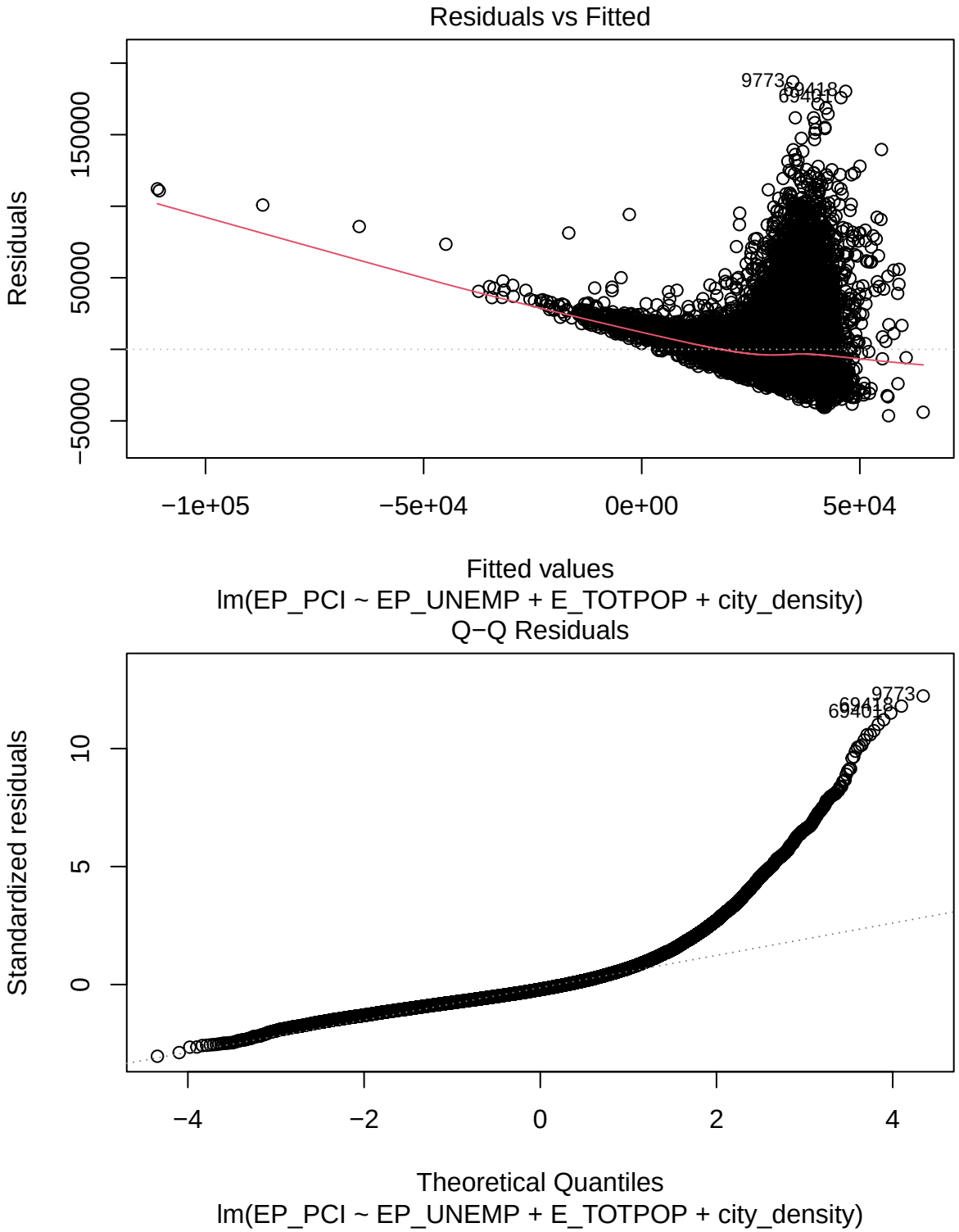
We will make a new DataFrame to keep track of our variables, which we will call `advanced_df`.

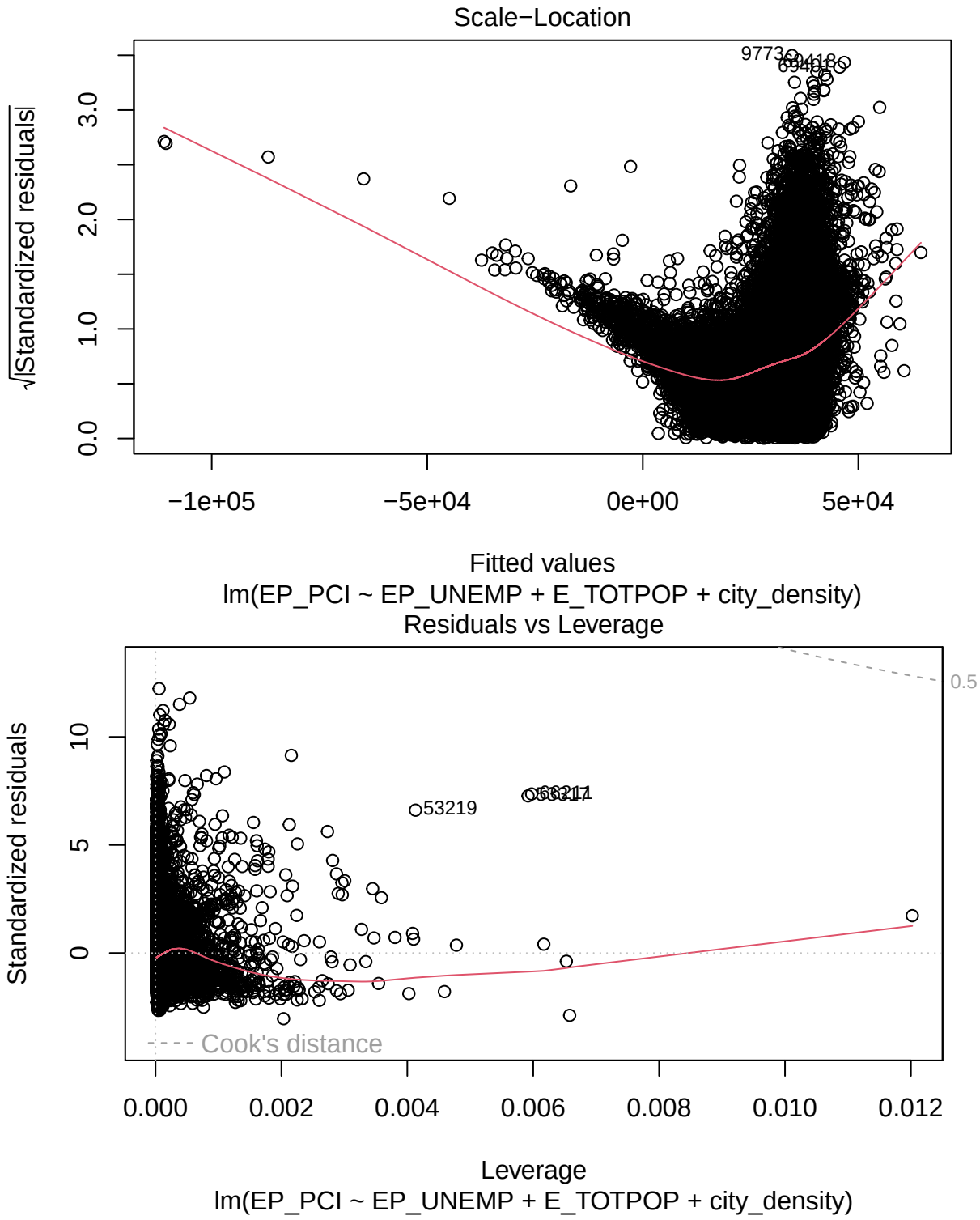
```
##
## Call:
## lm(formula = EP_PCI ~ EP_UNEMP + E_TOTPOP + city_density, data = clean_advanced_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46403  -9089  -3159   5024 186969
##
## Coefficients:
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   4.200e+04  1.585e+02  264.927  < 2e-16 ***
## EP_UNEMP     -1.525e+03  1.259e+01 -121.110  < 2e-16 ***
## E_TOTPOP     -1.309e-01  2.548e-02   -5.138 2.79e-07 ***
```

```
## city_density  1.150e-01  4.764e-03   24.137  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15290 on 72168 degrees of freedom
## Multiple R-squared:  0.1719, Adjusted R-squared:  0.1719
## F-statistic:  4995 on 3 and 72168 DF,  p-value: < 2.2e-16
```

Our multiple linear regression model fits the line EP_PCI-hat = 42000 - 1525 * EP_UNEMP -0.1309 * E_TOTPOP + 0.115 * city_density. Here is how we would interpret our model: * **Slope of EP_UNEMP**: For every 1% increase in unemployed people, we would expect income to decrease by $1,525 per capita. * **Slope of E_TOTPOP**: For every additional person in the total population, we would expect income to decrease by about $0.1309 per capita. * **Slope of city_density**: For every additional person per square mile, we would expect income to increase by about $0.115 per capita. * **Intercept**: All factors held constant, we would expect the income to be $4,200 per capita if the unemployment rate is 0, total population is 0, and city density is 0. However, it is important to note this involves extrapolation and is inappropriate to infer in this context.

Analysis


Residuals vs Fitted
Residuals
150000
50000
−50000
9773
694418
694019
−1e+05
−5e+04
0e+00
5e+04
Fitted values
lm(EP_PCI ~ EP_UNEMP + E_TOTPOP + city_density)
Q−Q Residuals
Standardized residuals
10
5
0
9773
694418
694019
−4
−2
0
2
4
Theoretical Quantiles
lm(EP_PCI ~ EP_UNEMP + E_TOTPOP + city_density)



A multiple linear regression model is probably also not the best choice for the model because the residuals vs. fitted scatterplot show a clear pattern trending upward as the fitted values increase. Additionally, in the QQ Plot, the residuals do not always fall approximately along the diagonal and indicate normality; instead, after the theoretical quantile of 2,the the residuals curve upward, which demonstrate a distribution of the residuals is right-skewed. In the scale-location graph, the residuals do not appear to be evenly spaced, and in the residuals vs leverage graph, it is evident that some residuals are associated with a very high leverage.

Conclusion

```
## [1] 15286.38
```

With an RMSE of 15,289.38, our multiple linear regression model is not very predictive of our data. It is better to use a different type of model.

Conclusion & Discussion

In conclusion, our analysis of the causal relationship between per capita income (`EP_PCI`) and unemployment rate (`EP_UNEMP`) in the socioeconomic vulnerability index dataset reveal that there is a strong negative correlation between them, though a linear regression may not be the best model to capture the nuances of their relationshiip. Additionally, the distributions of these features appear to be similar, but there is no statistical evidence to support this. In our advanced analysis, we attempt to fit a multiple linear regression with the addition of new features that describe the population distribution and density of a city.

One limititation that we noticed with our data was missing data in the socieoeconomic vulnerability index dataset. Although we removed missing data, the resulting data could underestimate or overestimate certain features of the dataset, which inhibits our ability to make accurate calculations.