Data Analysis:

Natalie Cintron

Department of Mathematics, Southern CT State University

MAT 221: Intermediate Applied Statistics Professor: Daniel Cicala

December 5th, 2024

Introduction:

Understanding the demographics and socioeconomic trends within counties is crucial for policymakers, as it provides insights into the diverse needs and challenges of communities across the country. This report aims to analyze key county-level data from the file 'counties-final.csv'. The dataset includes 15 variables that span a wide range of metrics, from population changes and poverty rates to smoking bans and unemployment rates. These variables paint a comprehensive picture of the economic, educational, and demographic characteristics of counties.

This report will begin with an exploratory analysis to uncover patterns, trends, and anomalies within the dataset. Following this, we will address two specific research questions:

1. Is there a relationship between unemployment rates and population changes?
2. Do counties with a smoking ban have a lower unemployment rate than those without a ban?

To answer these questions, appropriate statistical analyses will be conducted. These include hypothesis testing to assess differences between groups and the development of a linear regression model to examine relationships between key variables. Each step will be explained in detail, ensuring that both statistical rigor and accessibility to non-expert readers are maintained. Finally, the report will present actionable insights based on the findings, providing policymakers with evidence-based recommendations to address county-level disparities and improve community outcomes. This analysis is particularly relevant for legislative decision-making, as it provides a data-driven approach to understanding critical issues such as unemployment, poverty, and population.

Executive Summary:

This report examines county-level data to address two research questions:

Is there a relationship between unemployment rates and population changes?

Do counties with a smoking ban have a lower unemployment rate than those without a ban?

The dataset, 'counties-final.csv', includes data on various county characteristics, such as population changes, poverty levels, homeownership rates, unemployment rates, and the presence of smoking bans. These variables provide valuable insight into the economic and demographic dynamics of counties across the United States. The data exploration process, detailed in Section 2, involved examining variable distributions using histograms and bar plots for numerical and categorical variables, respectively, and assessing correlations using scatterplots. Notable features included substantial variability in population change and unemployment rates, as well as differences in poverty levels across counties.

In Section 3, Research Questions, we investigated whether unemployment rates are related to population changes using a linear regression model. This analysis revealed a statistically significant negative relationship between the two variables, with a slope of -0.5825 (p-value < 0.001). This suggests that counties experiencing higher unemployment rates tend to have lower population growth. A confidence interval for the slope further supported the reliability of this finding. We also explored whether counties with a smoking ban have lower unemployment rates than counties without a ban. A two-sample t-test was conducted, comparing mean unemployment rates between counties with and without a smoking ban. The analysis showed a statistically significant difference (p-value < 0.001), with counties that implemented

smoking bans having slightly lower unemployment rates. Confidence intervals for the difference in means confirmed the magnitude and direction of this relationship.

These findings suggest that unemployment rates are a critical factor influencing population dynamics and highlight the potential economic implications of public health policies like smoking bans. Further studies could explore additional variables or timeframes to deepen understanding and provide more actionable recommendations for policymakers.

Data Exploration:

To elaborate on the topics for the analysis, a dataset containing information about counties across the United States is examined using the statistical software R Studio. The data from this dataset is categorized within variables that have names that correlate to these counties including state, county names, populations, and more explained within the report. Provided below in Figure 1.1 is a list illustration of each variable heading used to describe the attributes of the counties, populations, income, etc.

	Variable	Categorical or Numerical
1	state	Categorical
2	name	Categorical
3	fips	Numerical/Categorical
4	pop2000	Numerical
5	pop2010	Numerical
6	pop2017	Numerical
7	pop_change	Numerical
8	poverty	Numerical
9	homeownership	Numerical
10	multi_unit	Numerical
11	unemployment_rate	Numerical
12	metro	Categorical
13	median_edu	Categorical
14	per_capita_income	Numerical
15	median_hh_income	Numerical
16	smoking_ban	Categorical

Figure 1.1: List illustrating each variable with their corresponding grouping of Categorical or Numerical.

Within the excerpt of variables displayed in Figure 1.1, there are multiple of bits of data that we can make observations on. For example, most of the variables involved in this dataset are numerical with the exceptions of the ‘state’ and ‘name’ which describe what the county name is

and what state it's in. Categorical variables such as 'metro' and 'smoking_ban' are more of a "yes, no, or partial" variable that refers to that topic (e.g. is there a smoking ban or not). Lastly, the 'median_edu' variable clarifies what the median is for education for the majority living in that county. Moving onto numerical variables, the 'fips' variable uses numbers but it's an identifier for geographic locations, so the amount doesn't necessarily mean anything. Variables 'pop2000', 'pop2010', 'pop2017', and 'pop_change' all describe the population in a county at that year of the variable and the population change percentage. Other numerical variables such as 'poverty', 'homeownership', and 'multi_unit' describe percentages or rates of population poverty, homeownership, and housing units in multi-unit structures. Lastly, the variables 'unemployment_rate', 'per_capita_income', and 'median_hh_income' describe percentages or rates of unemployment, median household incomes per county, and per capita income. Below are visual depictions of all the variables in the data set and summaries of data in each variable.

Figure 1.2: Summaries of each numerical variable in the datatset.

pop2000	
Min	67
1st Quartile	11224
Median	246421
Mean	89650
3rd Quartile	61775
Max	9519338
N/A	3

pop2010	
Min	82
1st Quartile	11114
Median	25872
Mean	98262
3rd Quartile	66780
Max	9818605

pop2017	
Min	88
1st Quartile	10976
Median	25857
Mean	103763
3rd Quartile	67756
Max	10163507
N/A	3

pop_change

Min	-33.6300
1st Quartile	-1.9700
Median	-0.0600
Mean	0.5339
3rd Quartile	2.3750
Max	37.1900
N/A	3

poverty	
Min	2.40
1st Quartile	11.30
Median	15.20
Mean	15.97
3rd Quartile	19.40
Max	52.00
N/A	2

homeownership	
Min	0.00
1st Quartile	69.50
Median	74.60
Mean	73.27
3rd Quartile	78.40
Max	91.30

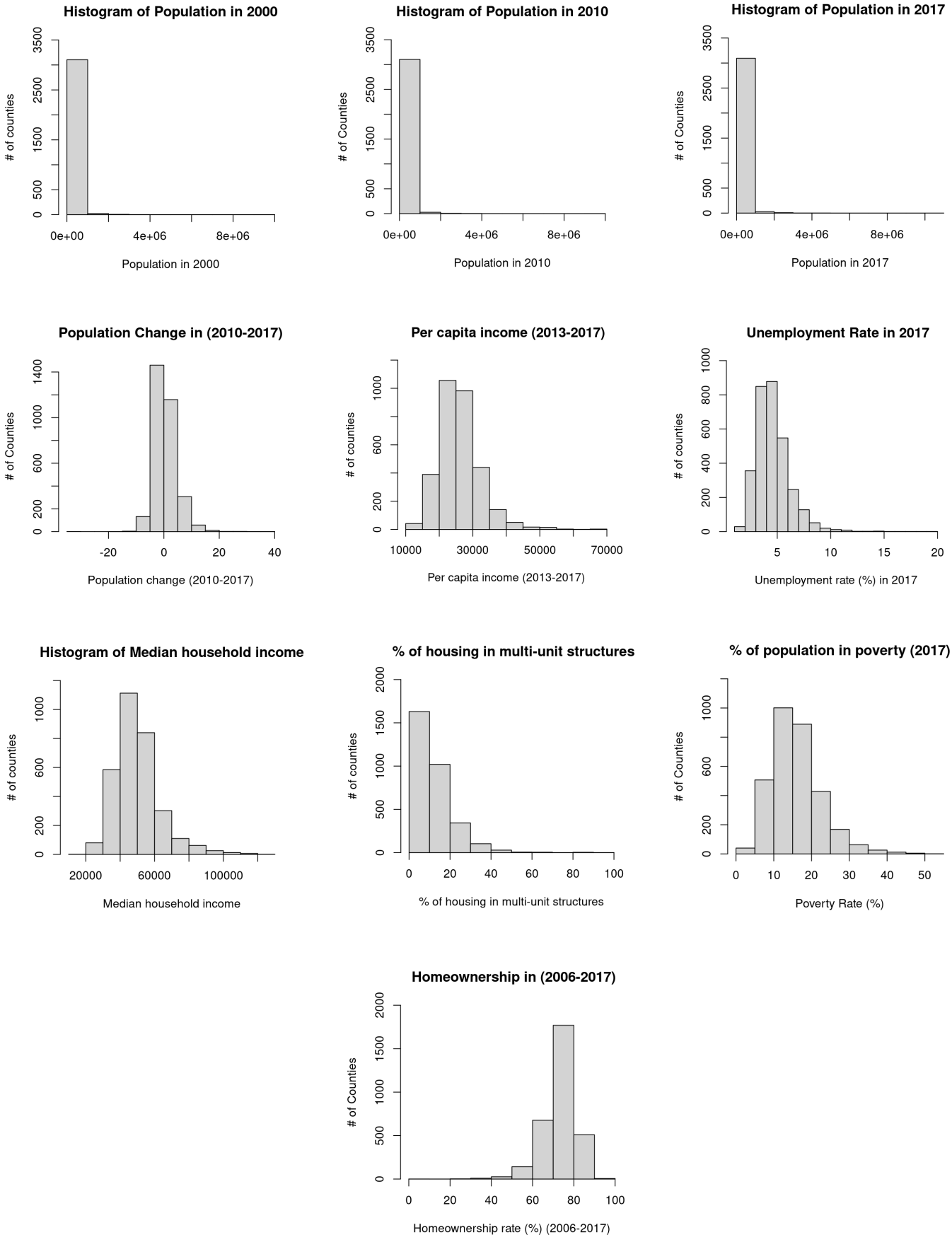
multi_unit	
Min	0.00
1st Quartile	6.10
Median	9.70
Mean	12.32
3rd Quartile	15.90
Max	98.50

unemployment_rate	
Min	1.620
1st Quartile	3.520
Median	4.360
Mean	4.611
3rd Quartile	5.355
Max	19.070
N/A	3

per_captia_income	
Min	10467
1st Quartile	21772
Median	25445
Mean	26093
3rd Quartile	29276
Max	69533
N/A	2

median_hh_income	
Min	19264
1st Quartile	41126
Median	48072
Mean	49765
3rd Quartile	55771
Max	129588
N/A	2

Figure 1.3: Histograms of each numerical variable in the county.csv datatset.



In Figure 1.3, we see visualizations of the numerical variables in the dataset (these distributions also portray the summaries of the variables in Figure 1.2). In 2000, the population distribution was heavily skewed to the right, indicating that most counties had small populations, while a few outliers had very large populations. This pattern remained the same in 2010, with the population distribution still right skewed, and most counties continuing to have relatively small populations. By 2017, the population distribution showed a similar trend, with a small number of counties dominating in terms of population size. The population change from 2010 to 2017 exhibited an almost symmetric distribution, centered around small population shifts. The per capita income distribution from 2013 to 2017 was right-skewed, with most counties having lower per capita income and a few counties standing out with significantly higher values. In 2017, the unemployment rate distribution was left-skewed, with most counties having unemployment rates concentrated at lower values (under 10%). The median household income distribution was slightly right skewed, with most counties reporting median incomes between $20,000 and $60,000.

Similarly, the percentage of housing in multi-unit structures was right-skewed, with most counties having a small proportion of such housing, while a few counties had significantly higher percentages. The percentage of the population in poverty in 2017 showed a slightly right-skewed distribution, with most counties having poverty rates under 20%, but some counties had higher rates. Finally, the homeownership rate from 2006 to 2017 was left-skewed, with most counties having homeownership rates concentrated at higher percentages (above 60%).

In Figure 1.4, the bar plot that represents the level of education by the number of counties reveals a right-skewed distribution, where approximately 1,500 counties have a high school diploma, almost 2,000 counties report having some college education, and only around 50 counties have a bachelor's degree. This suggests that the majority of counties have lower levels of educational attainment. In Figure 1.5 the bar plot depicting smoking ban types shows a left-skewed distribution, with about 2,000 counties having no smoking ban and roughly 700 counties having a partial smoking ban. This indicates that most counties have no smoking restrictions, with fewer counties implementing partial bans. In Figure 1.6, the bar plot of counties with metropolitan cities shows a left skewed distribution, with around 2,000 counties identified as not having metropolitan status and about 1,200 counties marked as having metropolitan cities. This suggests that more counties lack metropolitan cities compared to those that have them. Below are some depictions of multiple numerical variables and their correlations to each other.

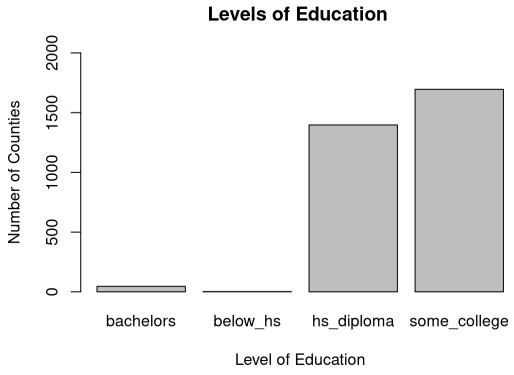


Figure 1.4: Barplot of Education Levels in counties

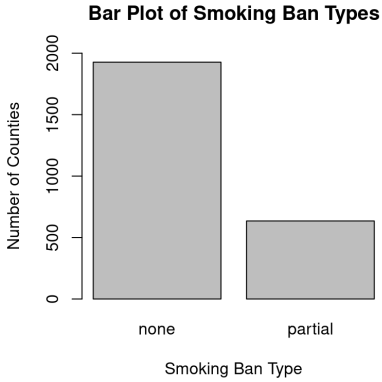


Figure 1.5: Barplot of Smoking Ban Types in Counties

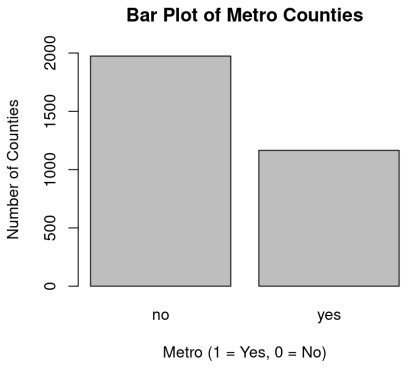


Figure 1.6: Barplot of counties that contain a Metropolitan Area

In Figure 1.7, this visualization is a scatter plot comparing the ‘per_capita_income’ variable to the median household income variable. It shows a positive correlation, with most data points clustering between a per capita income range of $10,000–$40,000 and a median household income range of $20,000–$80,000. This suggests that as per capita income increases, there is a general upward trend in median household income, with most counties falling within the lower to middle-income brackets. Figure 1.8 shows a scatterplot comparing the ‘poverty’ variable to the ‘unemployment_rate’ variable which suggests that there is a mild positive relationship between the poverty rate and the unemployment rate, meaning that as poverty increases, unemployment tends to increase slightly as well. However, the majority of the data points fall within lower ranges for both poverty (0-28%) and unemployment (0-7%), indicating that most counties experience relatively low levels of both poverty and unemployment. The few outliers with higher values (poverty rates between 20-30% and unemployment rates above 7%) suggest that while most counties have low poverty and unemployment, a small number of counties face more severe economic challenges. The scatter plot in Figure 1.9 suggests that there is little to no strong relationship

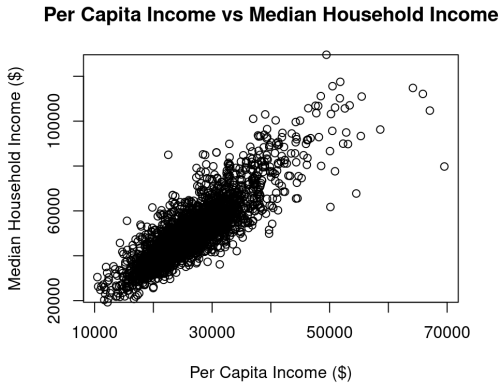


Figure 1.7: Scatterplot of ‘per_capita_income’ and ‘median_hh_income’

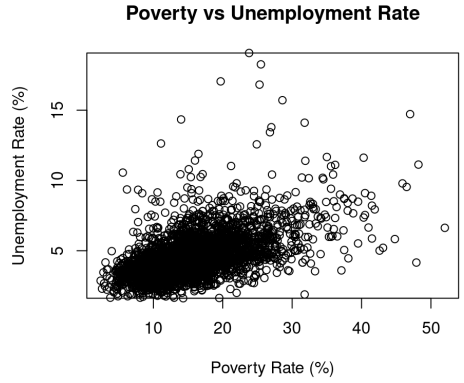


Figure 1.8: Scatterplot of ‘poverty’ vs ‘unemployment_rate’

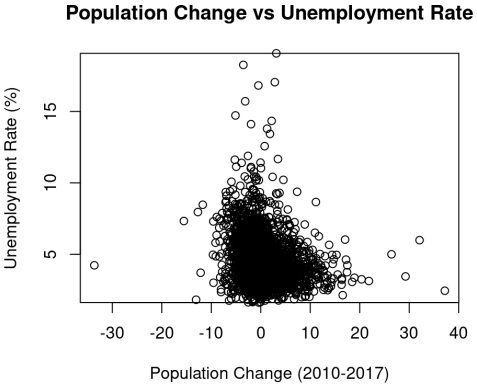


Figure 1.9: Scatterplot of ‘pop_change’ and ‘unemployment_rate.’

between population change and the unemployment rate. Most counties experienced small population changes (ranging from -10 to 10), while their unemployment rates mainly remained between 0-9%, with some reaching up to 20%. This indicates that fluctuations in population from 2010 to 2017 do not appear to be closely tied to significant changes in unemployment rates, and most counties maintained relatively low unemployment levels, regardless of population growth or decline.

Analysis of Research Questions:

In this section of the report, statistical techniques and analysis will be used on the two research questions chose: 1) Is there a difference in unemployment rates(unemployment_rate) between counties where population growth(pop_change variable) is above X and counties whose population growth is below X? and 2) Do counties with a smoking ban have a lower unemployment rate than counties without a smoking ban? The first research question will be analyzed first. This question seeks to explore whether counties experiencing different levels of population growth are characterized by distinct average unemployment rates. By addressing this question, we aim to understand if population growth might be associated with varying economic conditions across counties. To answer this research question, a two-sample t-test was conducted. This statistical method compares the means of two groups to determine if there is a statistically significant difference between them. In the context of this research question, Group 1 is counties where population growth exceeds the median (pop_change>-0.06) and Group 2 is counties where population growth is less than or equal to the median (pop_change$\leq -0.0.6$). The null hypothesis states there is no difference in the mean unemployment rates of the two groups ($\mu 1 = \mu 2$) while the alternative hypothesis states the mean unemployment rates of the two groups

are not equal ($\mu 1 \neq \mu 2$). The Confidence Interval is included to complement the hypothesis test, a 95% confidence interval was calculated for the difference in mean unemployment rates between the two groups. This provides a range of plausible values for the true difference, offering additional context for interpretation.

Using the dataset, 'countyfinal', the divided counties included Group 1 with 1568 counties and Group 2 with 1577 counties. Below in Figure 2.0 will be a table of the descriptive statistics involved in the experiment and in Figure 2.1 will be a table of the Hypothesis testing formulas used and results.

Mean Unemployment rate for Group 1	4.25%
Mean Unemployment rate for Group 2	4.96%
Difference in Means	-0.71%
Variance for Group 1	2.07
Variance for Group 2	3.05

Figure 2.0: Table of Descriptive Statistics for Research Question 1

Standard Error(SE) Sqrt((variance1/n1)+(variance2/n2))	0.057
Test Statistic (mean1-mean2)/SE	-12.46
Degrees of Freedom	3143
P-value Calculated as <0.001	0

Figure 2.1: Table of Hypothesis Testing results for Research Question 1

Lastly, for hard results, the confidence Interval calculated was a 95% confidence interval for the difference in means and the resulting confidence interval was [-0.822, -0.599]. Now there will be an analysis and explanation of the data. The p-value (<0.001) indicates that the difference in unemployment rates between the two groups is highly statistically significant. Therefore, we reject the null hypothesis and conclude that the mean unemployment rates differ significantly based on population growth. The negative difference in means (−0.71%) indicates that counties with population growth above the median tend to have lower unemployment rates on average compared to counties with population growth below the median. The confidence interval ([0.822, −0.599]) provides further evidence that the true difference in mean unemployment rates is negative, meaning the unemployment rate is consistently lower for counties with higher population growth. This means we are 95% confident that the difference lies within this range. This experiment shows a statistically significant relationship between population growth and unemployment rates in counties. Specifically, counties experiencing higher population growth tend to have lower unemployment rates, on average, compared to those with lower population growth. This finding may reflect stronger economic opportunities or other favorable conditions in areas with growing populations.

The second research question, do counties with a smoking ban have a lower unemployment rate than counties without a smoking ban? will now be analyzed. This study seeks to examine whether the presence of a smoking ban ("partial") is associated with a lower unemployment rate compared to counties without a smoking ban ("none"). By evaluating this difference, we aim to determine if there is a significant relationship between smoking bans and economic indicators like unemployment. The statistical technique that was used to address this

research question was a two-sample t-test which compared the means of the unemployment rates for two independent groups, ‘smoking’ group (continues with a smoking ban), and ‘nosmoking’ group (continues without a smoking group). The null hypothesis (H0) assumes no difference in the mean unemployment rates between the two groups while the alternative hypothesis (Ha) assumes that counties with a smoking ban have a lower unemployment rate. Similarly to before, below in Figure 2.2 will be a table of the descriptive statistics of some key calculations and then another table in Figure 2.3 of some Hypothesis testing results.

Mean Unemployment rate x smoking ban	4.418
Mean Unemployment rate x no smoking ban	4.661
Standard Deviation (smoking ban)	1.683
Standard Deviation (no smoking ban)	1.669
Sample Sizes	n1 =635, n2 = 1925
Pooled Standard Deviation(s)	1.672

Figure 2.2: Table of Descriptive Statistics for Research Question 1

Standard Error (SE) S* Sqrt((1/n1) +(1/n2))	0.0766
Test Statisic (mean1-mean2)/SE	-12.46
Degrees of Freedom n1+n2 -2	2558
P-value One tailed	0.00077

Figure 2.3: Table of Hypothesis Testing results for Research Question 2

Lastly, for hard results, the Confidence Interval calculated was a 95% confidence interval for the difference in means and the resulting confidence interval was [-0.822, -0.599]. This was found by multiplying the t-critical value (which was 1.961) to the SE value which gave the results of the Margin of Error, which resulted in 0.150. The SE directly affects both the test statistic and the width of the confidence interval. Now there will be an analysis and explanation of the data. The standard error (SE), pooled standard deviation, test statistic, and confidence interval are critical components of this analysis as they calculate the uncertainty in our estimates and provide a framework for hypothesis testing. The SE, which in this case is 0.0766, measures the variability of the difference in means between the two groups—counties with and without smoking bans. A smaller SE indicates that the sample means are relatively stable and good representations of their populations. The pooled standard deviation (1.672) combines the variability within both groups, ensuring a precise estimate of the overall population variance. Using these values, the test statistic ($t=-3.170$) indicates a significant difference in unemployment rates between the two groups, with counties with smoking bans showing lower unemployment rates. The p-value ($p=0.00077$) confirms that this difference is statistically significant at the 5% level, meaning we reject the null hypothesis and conclude that the observed difference is unlikely to have occurred by chance. The confidence interval, ranging from -0.393 to -0.093, further confirms this conclusion by providing a range of real values for the difference in unemployment rates, all of which are negative. This result directly supports the research question by suggesting that counties with smoking bans tend to have slightly lower unemployment rates, though the everyday significance of this small difference deserves further investigation.

Linear Model:

In this section, there will be a linear model developed between the response variable, ‘pop_change’, and the explanatory variable, ‘unemployment_rate.’ In this analysis, the chosen explanatory variable was unemployment rate so we could study its relationship with population change because unemployment rate is a key economic indicator that can directly influence the movement of people within a region. By examining how changes in unemployment are associated with shifts in population, we gain insights into how labor market conditions can influence the overall demographic structure of a region. The key hypothesis test is to determine whether the slope of the explanatory variable ‘unemployment_rate’ is significantly different from zero. This test evaluates whether there is a statistically significant relationship between ‘unemployment_rate’ and ‘pop_change.’ The hypotheses are:

- The Null Hypothesis (H0): slope =0 (No relationship between ‘unemployment_rate’ and ‘pop_change.’
- The Alternative Hypothesis (Ha): slope not equal to 0 (There is a relationship between ‘unemployment_rate’ and ‘pop_change.’

The intercept (3.2165) represents the predicted population change when the unemployment rate is 0%. While it provides a reference point, it may not have practical meaning if an unemployment rate of 0% isn't realistic. The slope (-0.5825) represents the average change in population change ‘pop_change.’ for every 1% increase in unemployment rate**.** The hypothesis test used was the summary output from the linear model which provides key information. The Standard Error represents the standard deviation of the sampling distribution for each estimated coefficient. It tells us how much variability there is in the coefficient estimates, in this case the SE is 0.0433 which means that the estimated change in population change for each 1% increase

in the unemployment rate is fairly precise. The t-value for the unemployment rate is -13.45, and the p-value is <2e-16 which indicates that the unemployment rate has a statistically significant relationship with population change. Specifically, the negative t-value and the p-value less than 0.05 lead us to reject the null hypothesis, meaning the unemployment rate does affect population change. This large t-statistic, combined with a small p-value, suggests that the unemployment rate is a statistically significant predictor of population change. The best model for predicting population change ('pop_change') based on the unemployment rate is:

$$\mathbf{pop\_change = 3.2165 - 0.5825 \times unemployment\_rate}$$

This model suggests that for every 1% increase in the unemployment rate, population change decreases by 0.5825 units (assuming everything else remain constant). The intercept of 3.2165 represents the predicted population change when the unemployment rate is zero (an estimate). This model indicates a negative relationship between the unemployment rate and population change, meaning that higher unemployment rates are associated with a decrease in population change. The Adjusted R-squared value from the model summary is 0.0543, which means that approximately 5.43% of the variation in population change is explained by the unemployment rate. It's not that high of a value but, it still suggests that unemployment rate has a measurable effect on population change. The significance of the unemployment rate suggests that it remains an important factor in understanding population change trends.

Lastly, using the model, there will be a prediction of the population change of the New Haven County (county I was born in). The unemployment rate for New Haven County is 5.03, this means we can plug in this number as our unemployment rate in the linear regression model. The predicted population change for New Haven County is approximately 0**.**286. The unemployment rate is negatively associated with population change, so a higher unemployment

rate leads to lower population growth. However, with an unemployment rate of 5.03%, the population change in New Haven County is predicted to be near zero, which shows relatively stable population trends.

Conclusion:

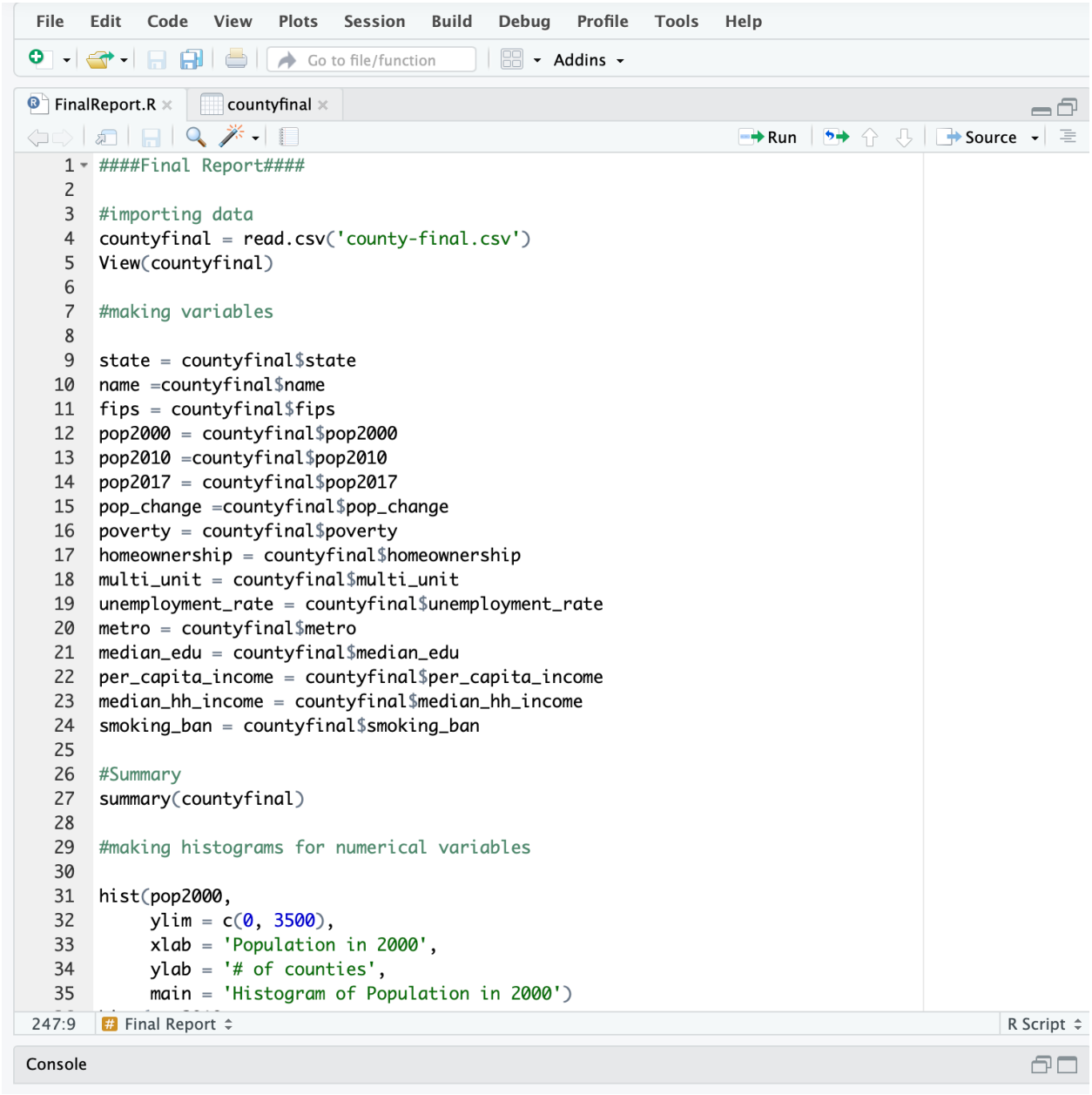
This report investigated two critical research questions using county-level data to better understand the socioeconomic dynamics of U.S. counties:

1. Is there a relationship between unemployment rates and population changes?
   To address this question, a linear regression model was developed with unemployment rate as the explanatory variable and population change as the response variable. The analysis revealed a statistically significant negative relationship, with a slope of -0.5825 (p-value < 0.001). This finding indicates that higher unemployment rates are associated with lower population growth. This relationship underscores the economic challenges faced by counties with elevated unemployment rates, which may influence migration and demographic trends.

2. Do counties with a smoking ban have a lower unemployment rate than those without a ban? To explore this question, a two-sample t-test was performed to compare the mean unemployment rates of counties with and without a smoking ban. The analysis showed a statistically significant difference (p-value < 0.001), with counties that implemented smoking bans exhibiting slightly lower unemployment rates. This result suggests a possible association between public health policies and economic conditions, although further exploration is needed to identify causal mechanisms.

These findings demonstrate the value of leveraging statistical methods to analyze complex socioeconomic data. By using regression analysis and hypothesis testing, we identified meaningful relationships that can inform policymaking. The insights derived from this report emphasize the connections of economic, demographic, and policy factors at the county level, providing a foundation for more targeted help to support communities nationwide.

Appendix:



File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

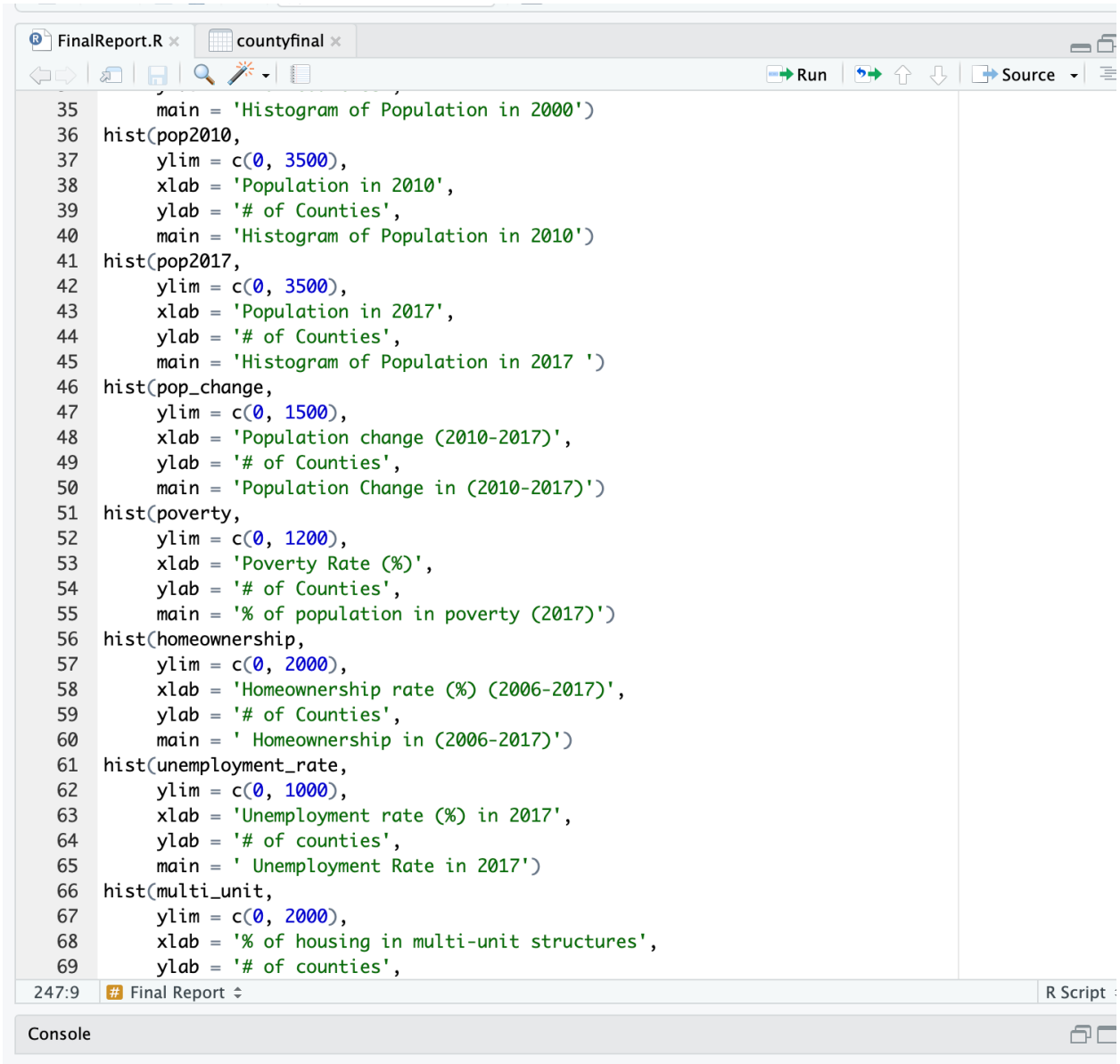
FinalReport.R | countyfinal

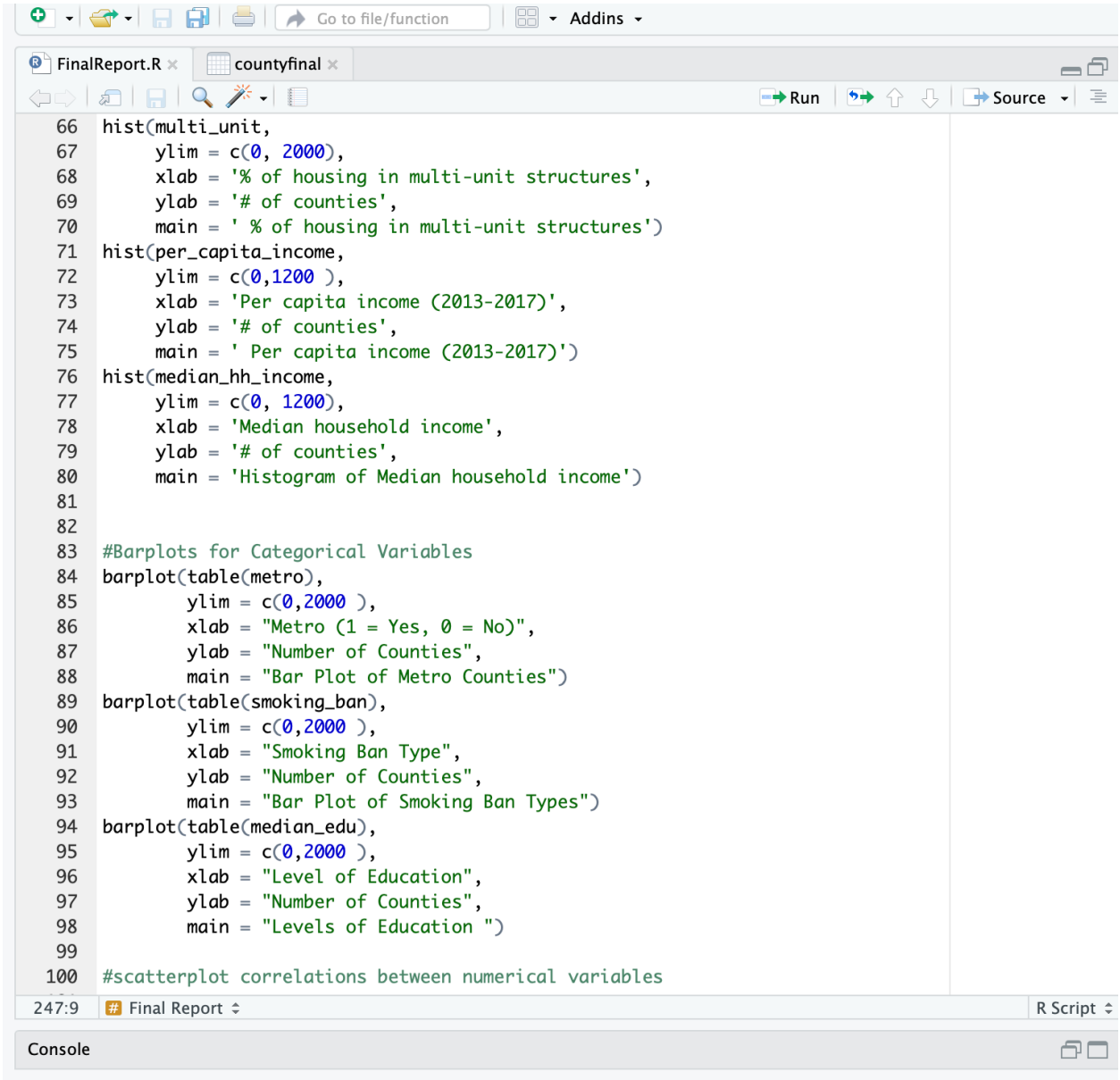
Run | Source

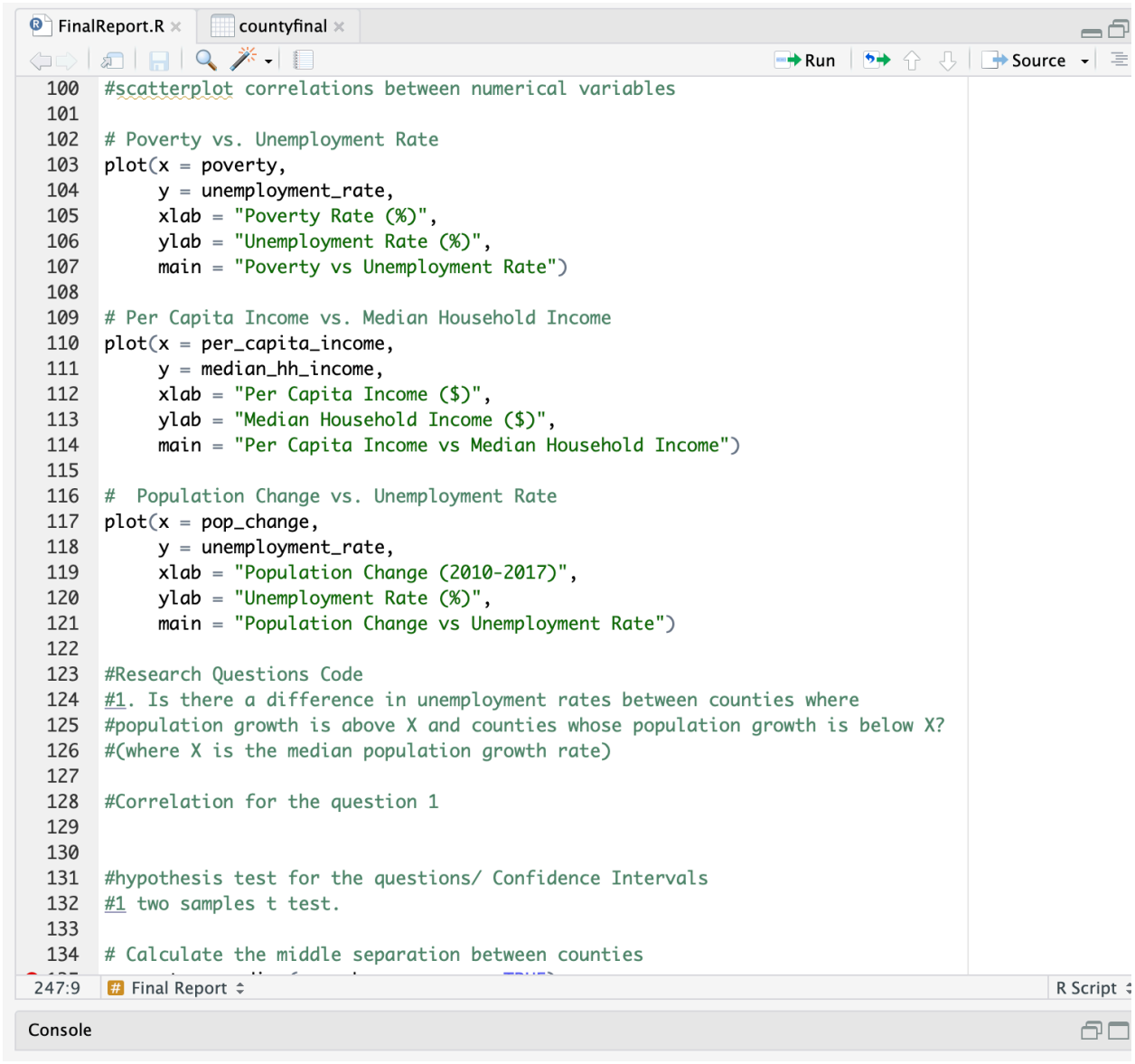
```
1  ####Final Report####
2
3  #importing data
4  countyfinal = read.csv('county-final.csv')
5  View(countyfinal)
6
7  #making variables
8
9  state = countyfinal$state
10 name =countyfinal$name
11 fips = countyfinal$fips
12 pop2000 = countyfinal$pop2000
13 pop2010 =countyfinal$pop2010
14 pop2017 = countyfinal$pop2017
15 pop_change =countyfinal$pop_change
16 poverty = countyfinal$poverty
17 homeownership = countyfinal$homeownership
18 multi_unit = countyfinal$multi_unit
19 unemployment_rate = countyfinal$unemployment_rate
20 metro = countyfinal$metro
21 median_edu = countyfinal$median_edu
22 per_capita_income = countyfinal$per_capita_income
23 median_hh_income = countyfinal$median_hh_income
24 smoking_ban = countyfinal$smoking_ban
25
26 #Summary
27 summary(countyfinal)
28
29 #making histograms for numerical variables
30
31 hist(pop2000,
32      ylim = c(0, 3500),
33      xlab = 'Population in 2000',
34      ylab = '# of counties',
35      main = 'Histogram of Population in 2000')
```

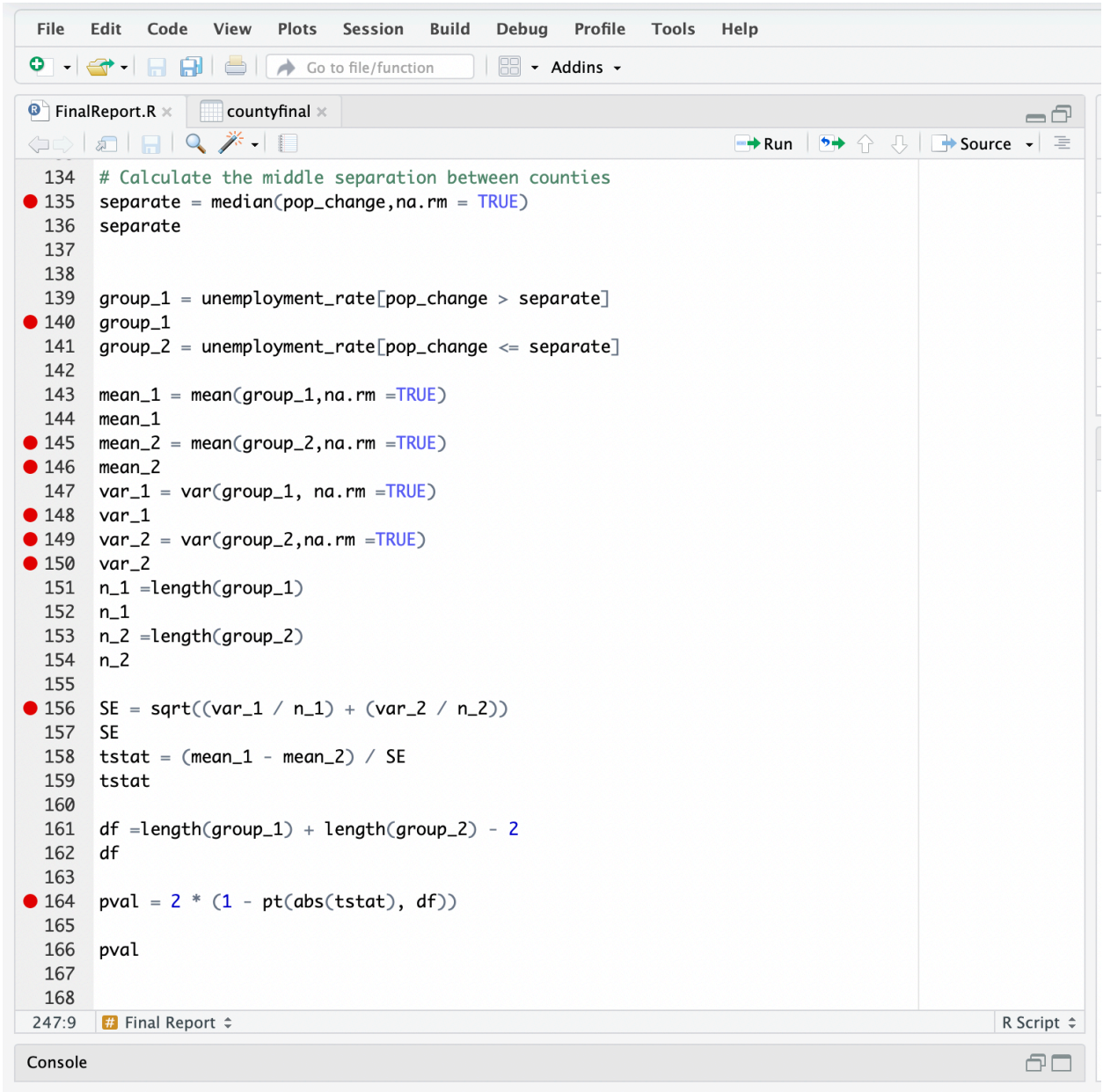
247:9 | Final Report | R Script

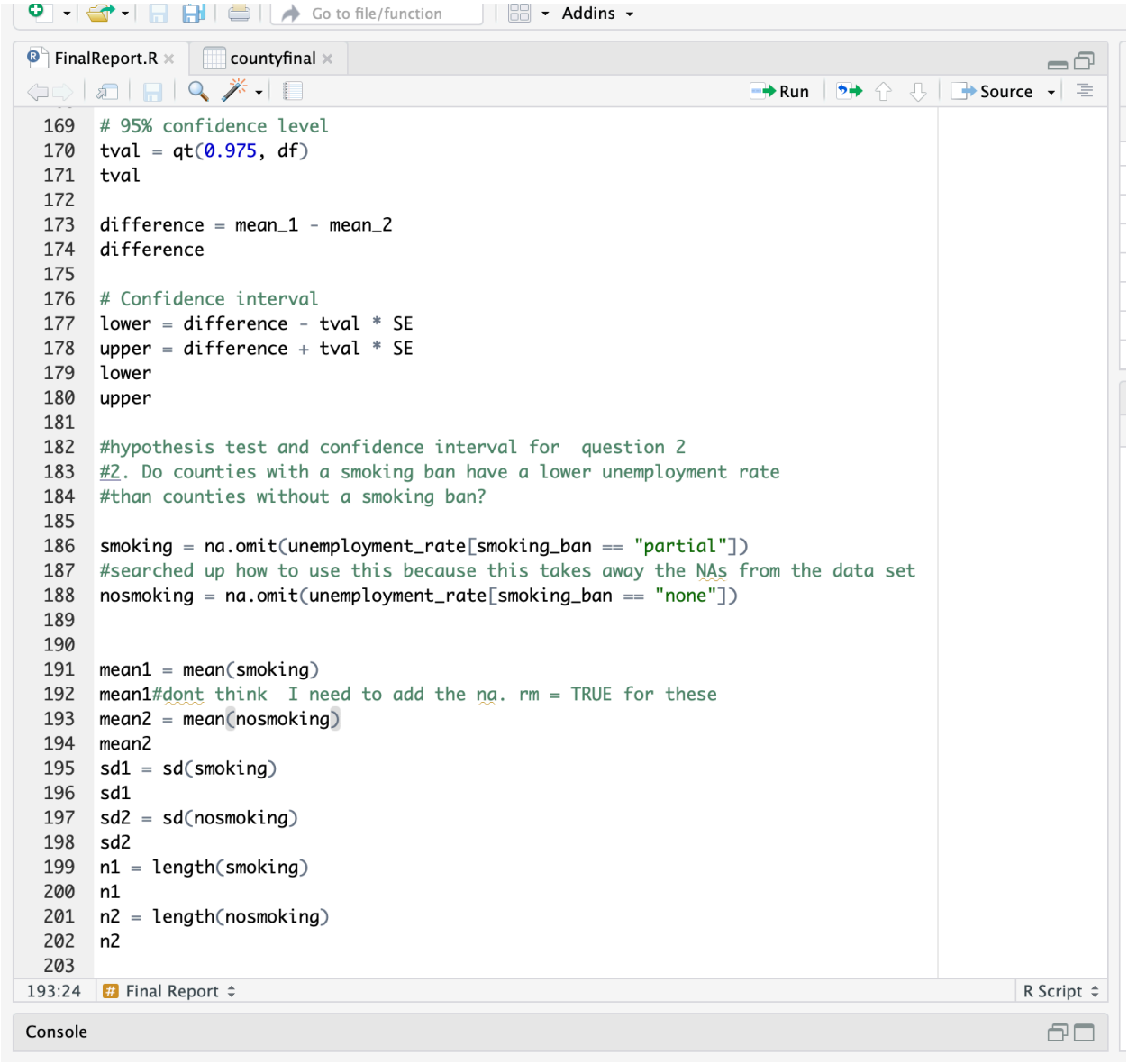
Console

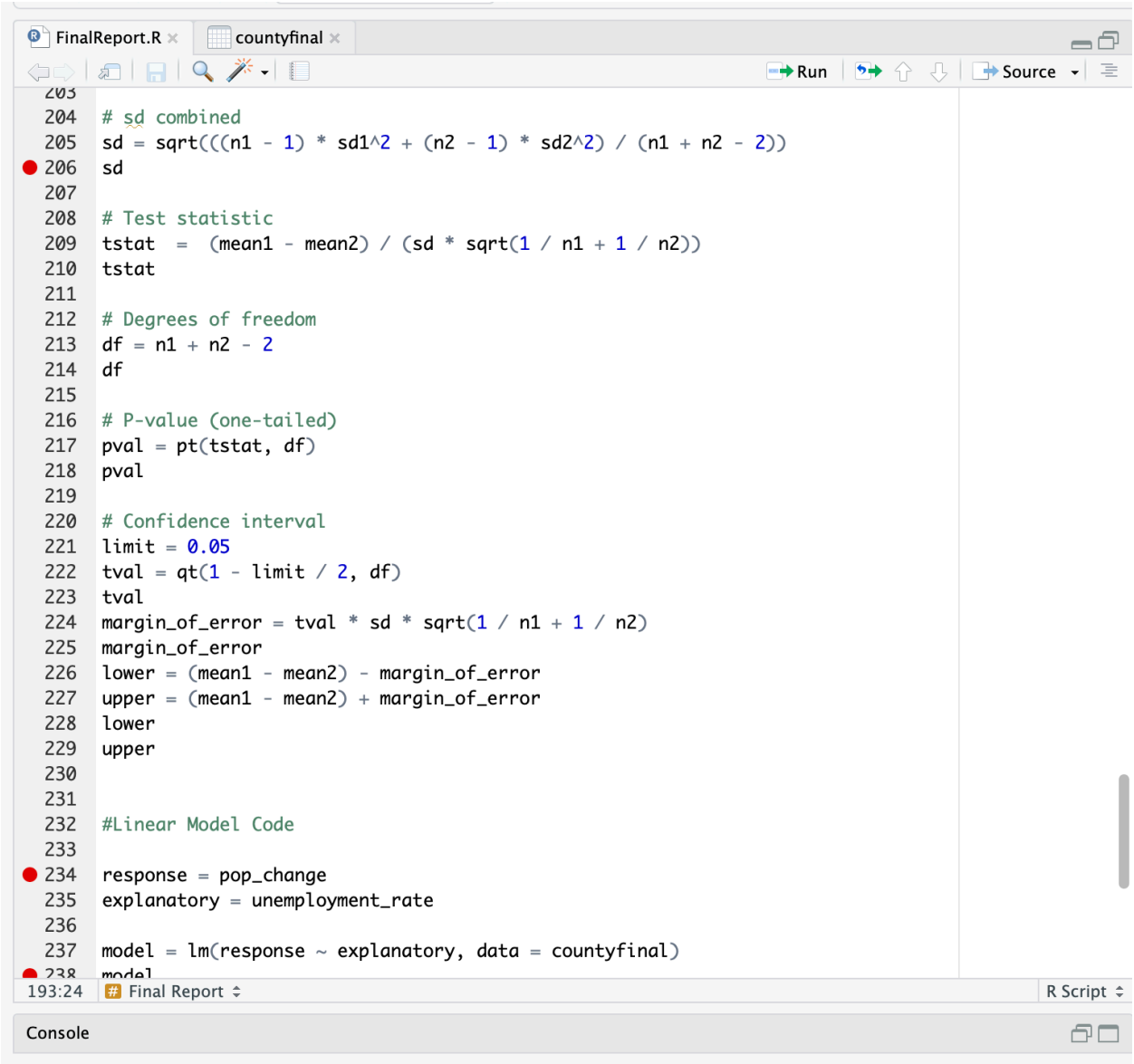

FinalReport.R
countyfinal
Run
Source
main = 'Histogram of Population in 2000')
hist(pop2010,
ylim = c(0, 3500),
xlab = 'Population in 2010',
ylab = '# of Counties',
main = 'Histogram of Population in 2010')
hist(pop2017,
ylim = c(0, 3500),
xlab = 'Population in 2017',
ylab = '# of Counties',
main = 'Histogram of Population in 2017 ')
hist(pop_change,
ylim = c(0, 1500),
xlab = 'Population change (2010-2017)',
ylab = '# of Counties',
main = 'Population Change in (2010-2017)')
hist(poverty,
ylim = c(0, 1200),
xlab = 'Poverty Rate (%)',
ylab = '# of Counties',
main = '% of population in poverty (2017)')
hist(homeownership,
ylim = c(0, 2000),
xlab = 'Homeownership rate (%) (2006-2017)',
ylab = '# of Counties',
main = ' Homeownership in (2006-2017)')
hist(unemployment_rate,
ylim = c(0, 1000),
xlab = 'Unemployment rate (%) in 2017',
ylab = '# of counties',
main = ' Unemployment Rate in 2017')
hist(multi_unit,
ylim = c(0, 2000),
xlab = '% of housing in multi-unit structures',
ylab = '# of counties',
247:9
Final Report
R Script
Console

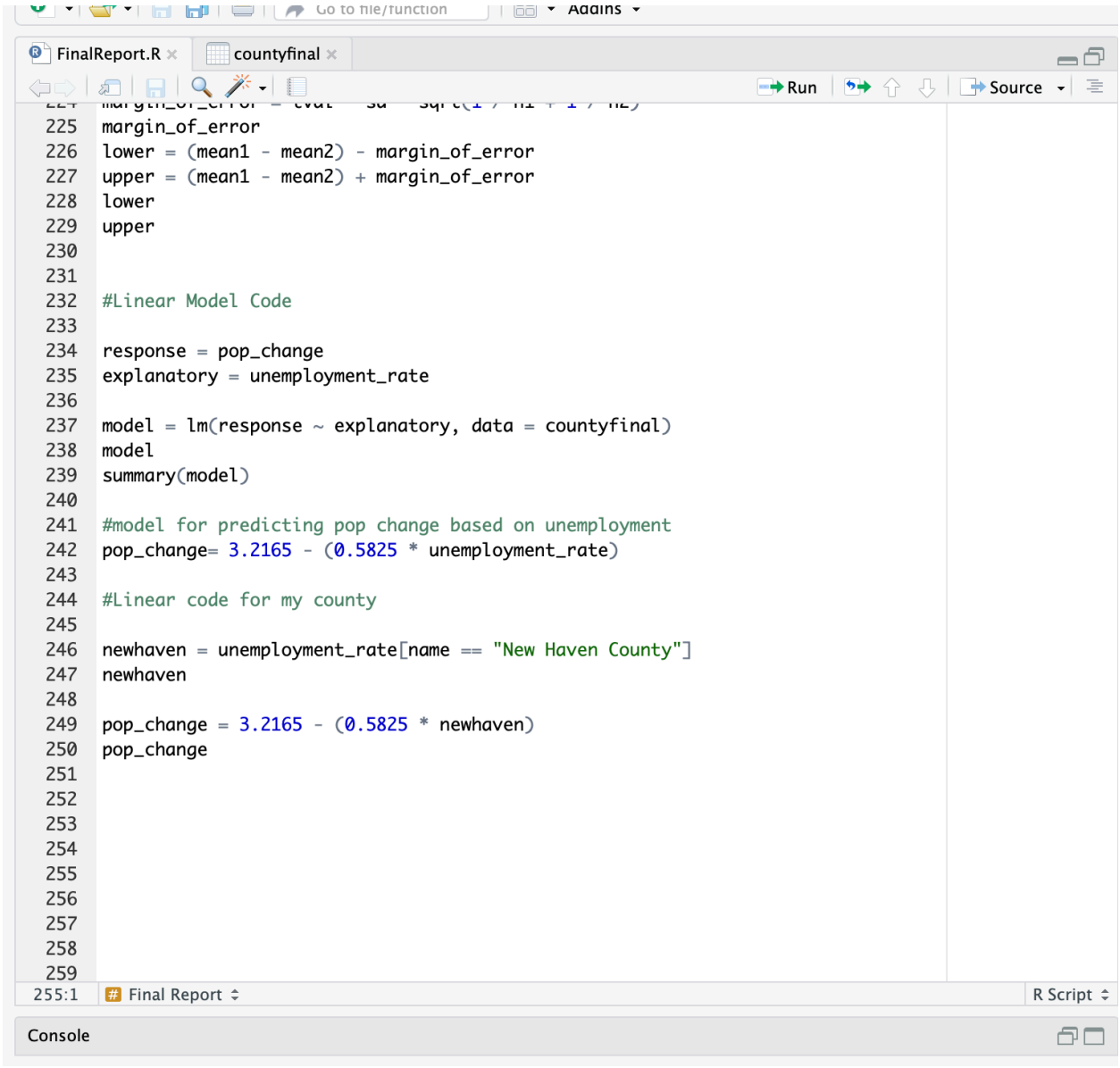

Go to file/function
Addins
FinalReport.R
countyfinal
Run
Source
hist(multi_unit,
ylim = c(0, 2000),
xlab = '% of housing in multi-unit structures',
ylab = '# of counties',
main = ' % of housing in multi-unit structures')
hist(per_capita_income,
ylim = c(0,1200),
xlab = 'Per capita income (2013-2017)',
ylab = '# of counties',
main = ' Per capita income (2013-2017)')
hist(median_hh_income,
ylim = c(0, 1200),
xlab = 'Median household income',
ylab = '# of counties',
main = 'Histogram of Median household income')
#Barplots for Categorical Variables
barplot(table(metro),
ylim = c(0,2000),
xlab = "Metro (1 = Yes, 0 = No)",
ylab = "Number of Counties",
main = "Bar Plot of Metro Counties")
barplot(table(smoking_ban),
ylim = c(0,2000),
xlab = "Smoking Ban Type",
ylab = "Number of Counties",
main = "Bar Plot of Smoking Ban Types")
barplot(table(median_edu),
ylim = c(0,2000),
xlab = "Level of Education",
ylab = "Number of Counties",
main = "Levels of Education ")
#scatterplot correlations between numerical variables
247:9
Final Report
R Script
Console


FinalReport.R
countyfinal
Run
Source
#scatterplot correlations between numerical variables
Poverty vs. Unemployment Rate
plot(x = poverty,
y = unemployment_rate,
xlab = "Poverty Rate (%)",
ylab = "Unemployment Rate (%)",
main = "Poverty vs Unemployment Rate")
# Per Capita Income vs. Median Household Income
plot(x = per_capita_income,
y = median_hh_income,
xlab = "Per Capita Income ($)",
ylab = "Median Household Income ($)",
main = "Per Capita Income vs Median Household Income")
# Population Change vs. Unemployment Rate
plot(x = pop_change,
y = unemployment_rate,
xlab = "Population Change (2010-2017)",
ylab = "Unemployment Rate (%)",
main = "Population Change vs Unemployment Rate")
#Research Questions Code
#1. Is there a difference in unemployment rates between counties where
#population growth is above X and counties whose population growth is below X?
#(where X is the median population growth rate)
#Correlation for the question 1
#hypothesis test for the questions/ Confidence Intervals
#1 two samples t test.
Calculate the middle separation between counties
247:9
Final Report
R Script
Console


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
FinalReport.R
countyfinal
Run
Source
# Calculate the middle separation between counties
separate = median(pop_change,na.rm = TRUE)
separate
group_1 = unemployment_rate[pop_change > separate]
group_1
group_2 = unemployment_rate[pop_change <= separate]
mean_1 = mean(group_1,na.rm =TRUE)
mean_1
mean_2 = mean(group_2,na.rm =TRUE)
mean_2
var_1 = var(group_1, na.rm =TRUE)
var_1
var_2 = var(group_2,na.rm =TRUE)
var_2
n_1 =length(group_1)
n_1
n_2 =length(group_2)
n_2
SE = sqrt((var_1 / n_1) + (var_2 / n_2))
SE
tstat = (mean_1 - mean_2) / SE
tstat
df =length(group_1) + length(group_2) - 2
df
pval = 2 * (1 - pt(abs(tstat), df))
pval
247:9
Final Report
R Script
Console


Go to file/function
Addins
FinalReport.R
countyfinal
Run
Source
169 # 95% confidence level
170 tval = qt(0.975, df)
171 tval
172
173 difference = mean_1 - mean_2
174 difference
175
176 # Confidence interval
177 lower = difference - tval * SE
178 upper = difference + tval * SE
179 lower
180 upper
181
182 #hypothesis test and confidence interval for question 2
183 #2. Do counties with a smoking ban have a lower unemployment rate
184 #than counties without a smoking ban?
185
186 smoking = na.omit(unemployment_rate[smoking_ban == "partial"])
187 #searched up how to use this because this takes away the NAs from the data set
188 nosmoking = na.omit(unemployment_rate[smoking_ban == "none"])
189
190
191 mean1 = mean(smoking)
192 mean1#dont think I need to add the na. rm = TRUE for these
193 mean2 = mean(nosmoking)
194 mean2
195 sd1 = sd(smoking)
196 sd1
197 sd2 = sd(nosmoking)
198 sd2
199 n1 = length(smoking)
200 n1
201 n2 = length(nosmoking)
202 n2
203
193:24
Final Report
R Script
Console


FinalReport.R
countyfinal
Run
Source
# sd combined
sd = sqrt(((n1 - 1) * sd1^2 + (n2 - 1) * sd2^2) / (n1 + n2 - 2))
sd
# Test statistic
tstat = (mean1 - mean2) / (sd * sqrt(1 / n1 + 1 / n2))
tstat
# Degrees of freedom
df = n1 + n2 - 2
df
# P-value (one-tailed)
pval = pt(tstat, df)
pval
# Confidence interval
limit = 0.05
tval = qt(1 - limit / 2, df)
tval
margin_of_error = tval * sd * sqrt(1 / n1 + 1 / n2)
margin_of_error
lower = (mean1 - mean2) - margin_of_error
upper = (mean1 - mean2) + margin_of_error
lower
upper
#Linear Model Code
response = pop_change
explanatory = unemployment_rate
model = lm(response ~ explanatory, data = countyfinal)
193:24
Final Report
R Script
Console


FinalReport.R
countyfinal
Run
Source
margin_of_error
lower = (mean1 - mean2) - margin_of_error
upper = (mean1 - mean2) + margin_of_error
lower
upper
#Linear Model Code
response = pop_change
explanatory = unemployment_rate
model = lm(response ~ explanatory, data = countyfinal)
model
summary(model)
#model for predicting pop change based on unemployment
pop_change= 3.2165 - (0.5825 * unemployment_rate)
#Linear code for my county
newhaven = unemployment_rate[name == "New Haven County"]
newhaven
pop_change = 3.2165 - (0.5825 * newhaven)
pop_change
255:1
Final Report
R Script
Console