Insert heading

KHL

remove the pound sign and insert the file path to the image on your computer in the

Introduction

The research question(s); Background/significance of the research; and relevant highlighted information about the data set. (abbreviated version of part 1)

- Does a higher level of educational attainment generally increase personal earnings income across different states?
- Does personal earnings increase with an individual's health?
- Do older individuals generally earn more money than younger individuals?

Given the rise in inflation and cost of living, exploring the relationships connected to personal earnings across different states is fundamental to understanding how factors like education, health status, and age, influence income disparities at a regional level.This inquiry is grounded in the longstanding debate within economic and social research regarding the return on investment in education, health, and wellness. By examining these relationships across diverse geographical areas, the analysis can uncover nuanced insights into how local economies, policies, and opportunities shape the economic benefits of educational attainment.

The U.S. Census is pivotal in the nation's political and economic framework, ensuring each community receives its fair share based on its specific needs (Bureau, 2021).

In addition, it is crucial to the political sphere with its use in redrawing a multitude of political boundaries to ensure each district contains roughly equal numbers of people (Mather & Scommegna, 2019). Its political importance extends even to the U.S. House of Representatives, which bases its apportionment of House seats on Census population data, safeguarding the equity of voting power within the nation (Farley, 2020).

Originating from multiple reputable sources, the dataset focuses on the United States demographic, economic, and educational landscapes as of 2020. It's based on the US Census Bureau's Annual Social and Economic Supplement (ASEC) survey, including the Current Population Survey (CPS) for employment statistics, and adds questions on poverty and migration. Unemployment data comes from the Bureau of Labor Statistics, and urban population percentages from the Census Bureau's Decennial Census. State sales tax rates, sourced from the Tax Foundation, provide a financial perspective.

The dataset was merged on a state basis, focusing on individuals 18 and older to better represent the adult population. It includes averages of education level, gender, work expenses, and age from the ASEC survey, combined with unemployment rates from the Bureau of Labor Statistics and sales tax rates from the Tax Foundation at the state level.

Methods and Analysis

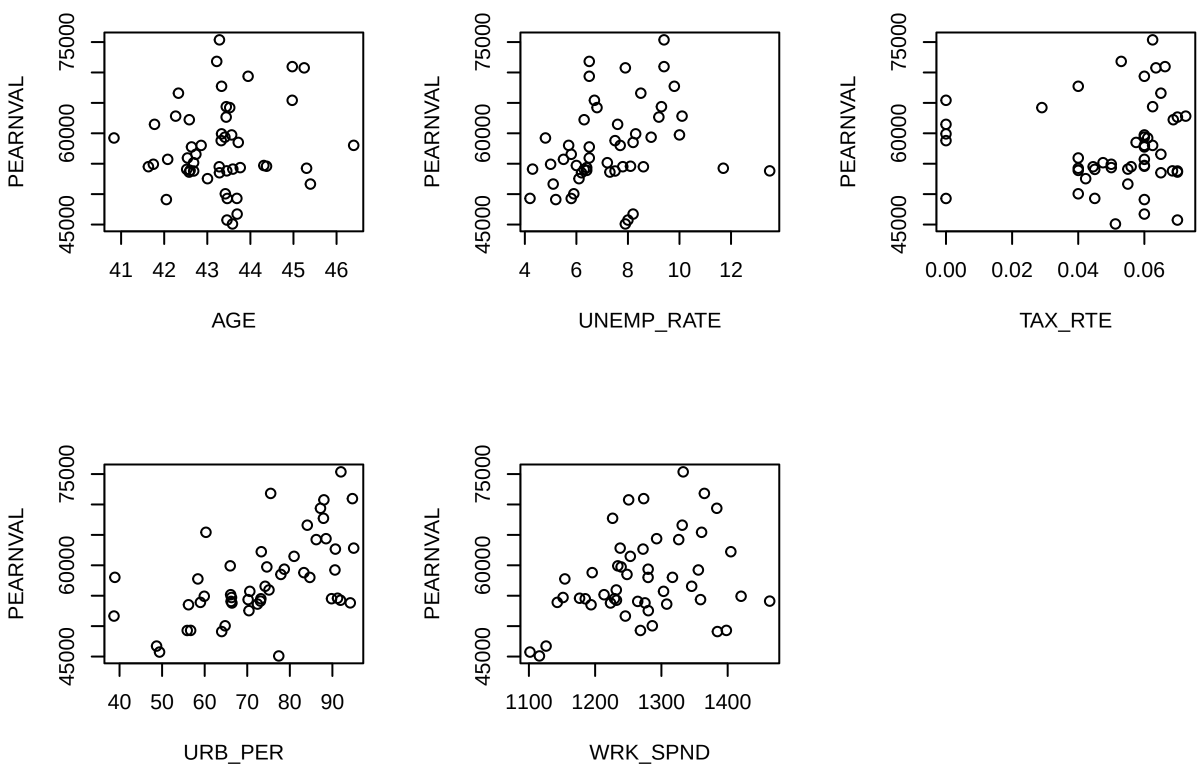
Include EDA from Report 1

ANALYSIS: - Multiple linear regression logistic regression - Include required analysis steps • Include the "added techniques" that you selected • Assessing the model. • Selecting a final "best" model. • NOTE: Your analysis should follow the appropriate order on your poster with a logical flow

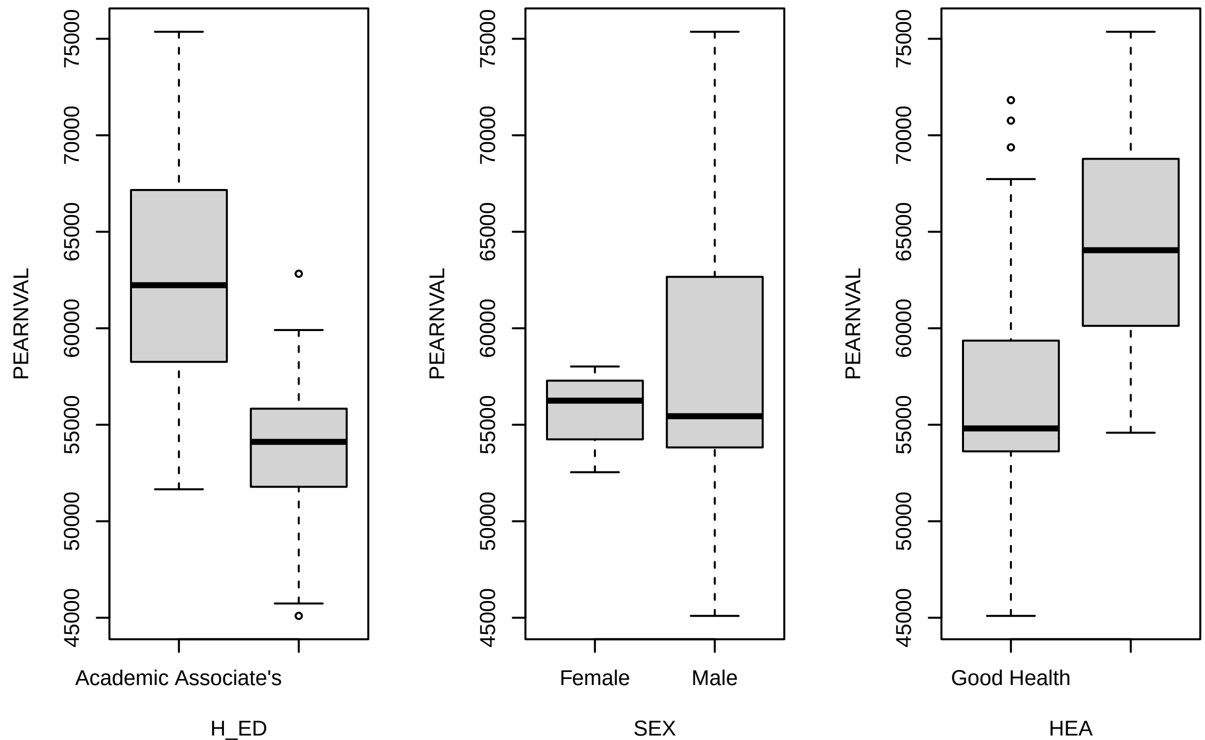
Model building with significance testing (should be supported by EDA and/or variable screening) • Identify and check for multicollinearity • Residual analysis (assumptions + extreme observations) Make necessary adjustments as you see fit. If you are attempting to correct a violation, you only need to try up to three corrections, if the first doesn't work. Include plots or output as needed. • Final model selected should be assessed. It may not be great, but you will explain that in conclusion • Include at least ONE additional techniques from the list to add to your analysis: Weighted least squares (Ch 9.4), External Model validation (Ch 5.11), box -cox transformation (supplemental), or another technique or aspect of MLR you learned on your own. • NOTE: if you have a unique situation with your data-discuss your analysis plan with Prof Varanyak

```
## From the EDA, we saw that the level "Bachelors" for H_ED only had one response. It al
#Displaying unequal responses to levels:
incomeData |>
  group_by(H_ED) |>
  summarise(count=n())
#Subsetting out DC:
incomeData<-incomeData |>
  filter(STATEFIPS != 11)

##EDA to explore variables in relation to the response, PEARNVAL
#Scatter plots for quantitative:
par(mfrow = c(2, 3))
plot(incomeData$AGE,incomeData$PEARNVAL,xlab="AGE",ylab="PEARNVAL")
plot(incomeData$UNEMP_RATE,incomeData$PEARNVAL,xlab="UNEMP_RATE",ylab="PEARNVAL")
```

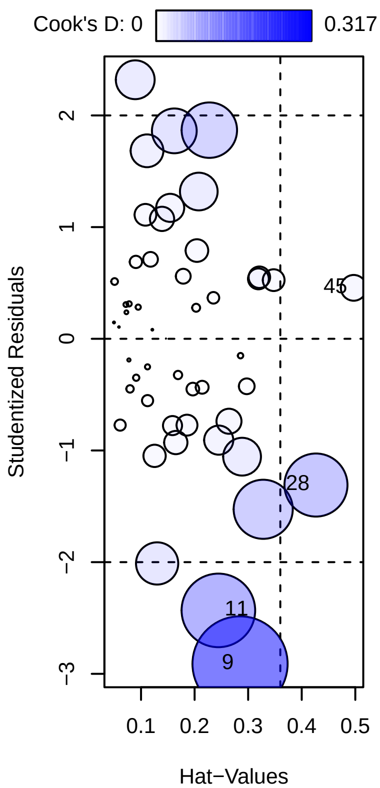
```
plot(incomeData$TAX_RTE,incomeData$PEARNVAL,xlab="TAX_RTE",ylab="PEARNVAL")
plot(incomeData$URB_PER,incomeData$PEARNVAL,xlab="URB_PER",ylab="PEARNVAL")
plot(incomeData$WRK_SPND,incomeData$PEARNVAL,xlab="WRK_SPND",ylab="PEARNVAL")

#Boxplots for qualitative:
par(mfrow=c(1,3))
```



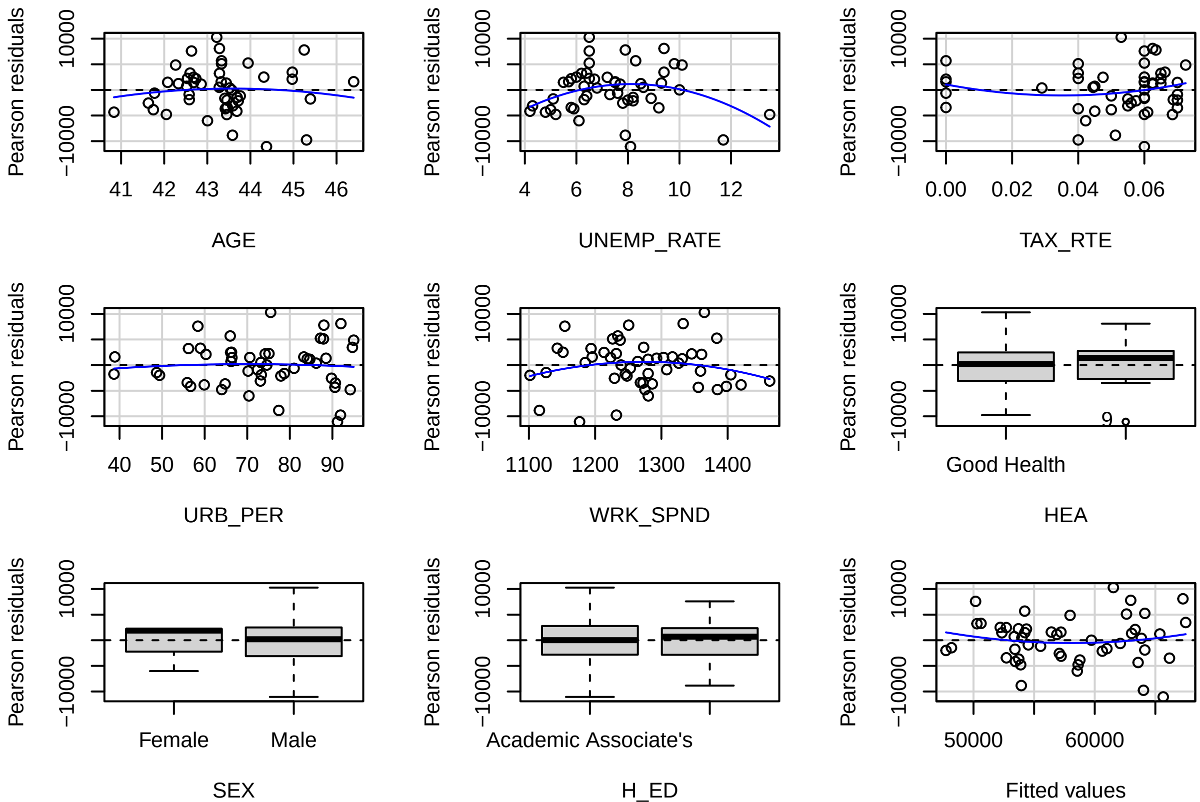

```
boxplot(PEARNVAL~H_ED,data=incomeData)
boxplot(PEARNVAL~SEX,data=incomeData)
boxplot(PEARNVAL~HEA,data=incomeData)
```



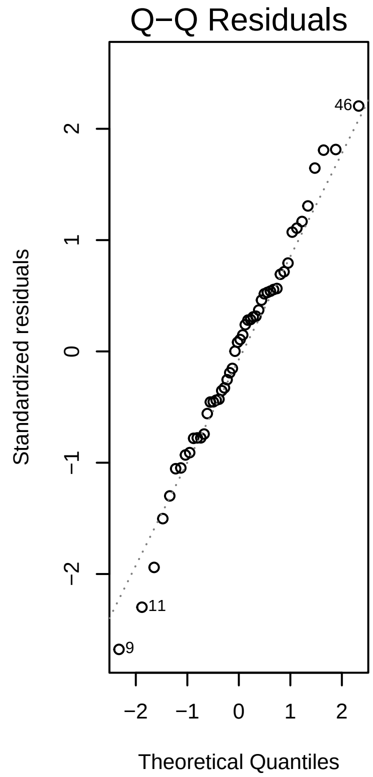

```
##Model with all predictors:
incomemod1<-lm(PEARNVAL~AGE+UNEMP_RATE+TAX_RTE+URB_PER+WRK_SPND+HEA+SEX+H_ED,data=income
summary(incomemod1)
#Results in a globally significant model

##Checking first model for influential observations:
influencePlot(incomemod1)
#Using the Studentized residual threshold of an absolute value of 2, there are 2 outlier

residualPlots(incomemod1,tests=F)
```

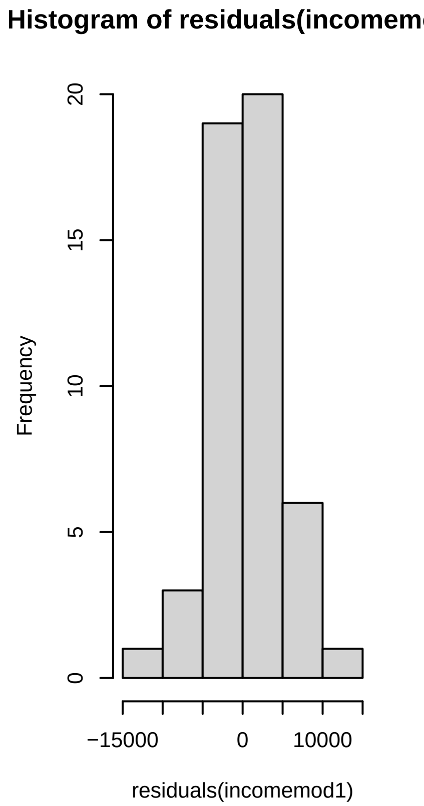


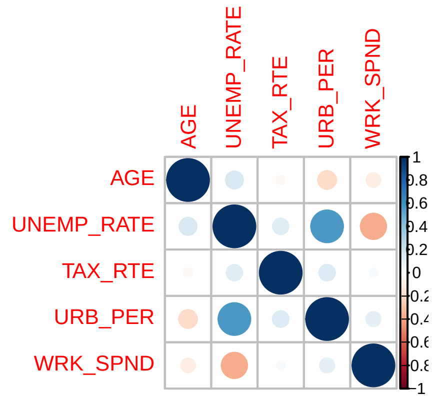
Cook's D: 0
0.317
Studentized Residuals
Hat−Values
2
1
0
−1
−2
−3
0.1
0.2
0.3
0.4
0.5
45
28
11
9



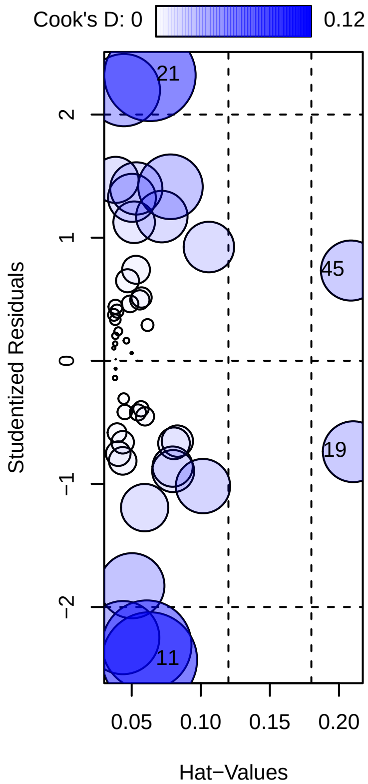

```
#Looking at residual plot, there appears to be a fanning relationship with residuals for

plot(incomemod1,which=2)
#QQplot is approximately linear

hist(residuals(incomemod1))
#residuals are approximately normally distributed

##Checking for multicollinearity
cor(incomeData[7:11])
quantincomeData<-incomeData[7:11]
corrplot(cor(quantincomeData))
```



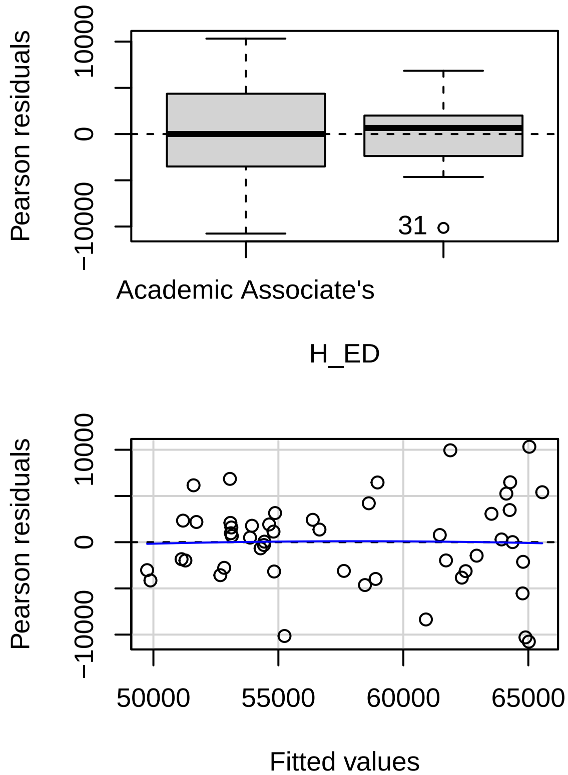




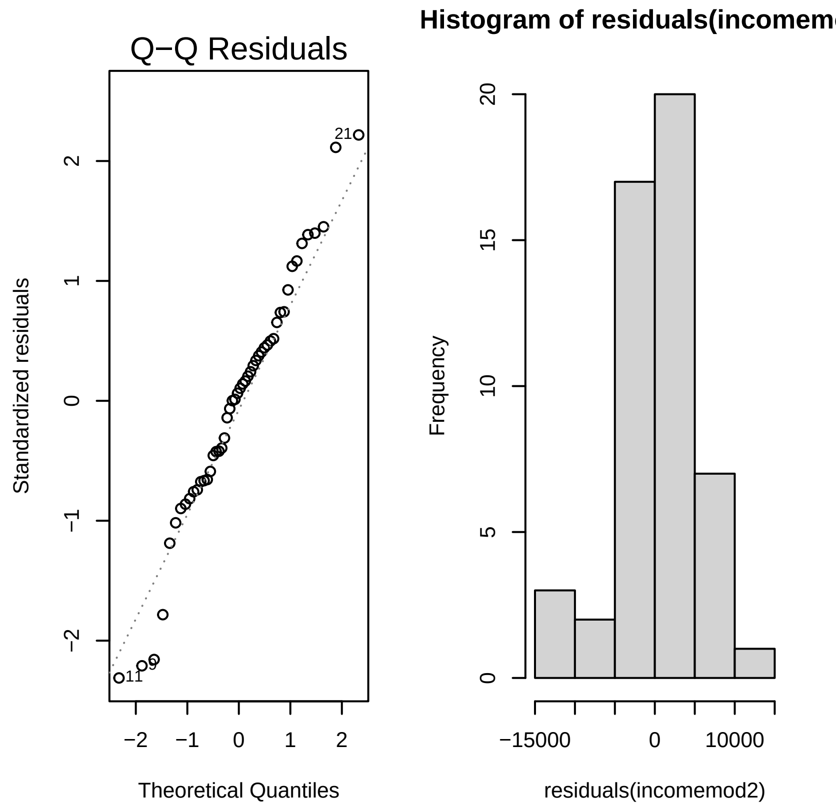

```
#No significant multicollinearity between quantitative variables

##Building a model using step wise regression
ols_step_both_p(incomemod1,pent=0.15,prem=0.15,details=T)
incomemod2<-lm(PEARNVAL~H_ED+URB_PER,data=incomeData)
summary(incomemod2)
#Results in a globally significant model

##Checking second model for influential observations:
influencePlot(incomemod2)
#Again, two influential observations (11 and 21)

residualPlots(incomemod2,tests=F)
```



Cook's D: 0
0.12
21
45
19
11
Studentized Residuals
Hat−Values
2
1
0
−1
−2
0.05
0.10
0.15
0.20




```
#No problems with residuals

plot(incomemod2,which=2)
#Approximate linearity, but worse than first model

hist(residuals(incomemod2))
#Residuals approximately normal, but distribution worse (less normal) than first model

##Building a third model with our 3 important var from part I of the project:
incomemod3<-lm(PEARNVAL~H_ED+HEA+AGE,data=incomeData)
summary(incomemod3)
#Globally significant model, but HEA and AGE are individually insignificant
```



Stage 1 - Quantitative Variables

Initial: $PEARNVAL = \beta_0 + \beta_1 UNEMP\_RATE + \beta_2 TAX\_RTE + \beta_3 URB\_PER + \beta_4 WRK\_SPND$

```
incomemods1<-lm(PEARNVAL~UNEMP_RATE+TAX_RTE+URB_PER+WRK_SPND,data=incomeData)
summary(incomemods1)
```

```
Call:
lm(formula = PEARNVAL ~ UNEMP_RATE + TAX_RTE + URB_PER + WRK_SPND,
    data = incomeData)

Residuals:
     Min       1Q   Median       3Q      Max
-11640.8  -3898.9   -252.2   2976.4  11562.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   2624.53   15416.17   0.170  0.86558
UNEMP_RATE     617.90     630.94   0.979  0.33265
TAX_RTE     -24252.83   41269.02  -0.588  0.55969
URB_PER        205.68      75.08   2.739  0.00879 **
WRK_SPND        28.85      11.60   2.487  0.01667 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5638 on 45 degrees of freedom
Multiple R-squared:  0.3968,    Adjusted R-squared:  0.3431
F-statistic: 7.399 on 4 and 45 DF,  p-value: 0.0001142
```

Final: $PEARNVAL = \beta_0 + \beta_1 URB\_PER + \beta_2 WRK\_SPND$

Stage 2 - Qualitative Variables

Initial: $PEARNVAL = \beta_0 + \beta_1 URB\_PER + \beta_2 WRK\_SPND + \beta_3 H\_EDassociates + \beta_4 HEAverygoodhealth$

```
incomemods2<-lm(PEARNVAL~URB_PER+WRK_SPND+H_ED+HEA,data=incomeData)
summary(incomemods2)
```

```
Call:
lm(formula = PEARNVAL ~ URB_PER + WRK_SPND + H_ED + HEA, data = incomeData)

Residuals:
     Min       1Q   Median       3Q      Max
-10946.1  -2639.6    734.3   2831.9  10274.5

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                32272.59   11702.52   2.758 0.008381 **
URB_PER                      190.77      48.51   3.933 0.000287 ***
WRK_SPND                      10.89       8.57   1.271 0.210227
H_EDVocational Associate's -5385.42    1623.56  -3.317 0.001806 **
HEAVery Good Health         3044.40    2065.74   1.474 0.147508
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4709 on 45 degrees of freedom
Multiple R-squared:  0.5792,    Adjusted R-squared:  0.5418
F-statistic: 15.48 on 4 and 45 DF,  p-value: 4.89e-08
```

Final: $PEARNVAL = \beta_0 + \beta_1 URB\_PER + \beta_2 H\_EDassociates$

```
incomemodfinal<-lm(PEARNVAL~URB_PER+H_ED,data=incomeData)
summary(incomemodfinal)
```

```
Call:
lm(formula = PEARNVAL ~ URB_PER + H_ED, data = incomeData)

Residuals:
     Min       1Q   Median       3Q      Max
-10761.5  -3089.3    395.4   2397.4  10323.6

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                47415.20    4009.21  11.827 1.09e-15 ***
URB_PER                      191.57      49.54   3.867 0.000338 ***
H_EDVocational Associate's -6997.66    1433.19  -4.883 1.25e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4810 on 47 degrees of freedom
Multiple R-squared:  0.5414,    Adjusted R-squared:  0.5219
F-statistic: 27.74 on 2 and 47 DF,  p-value: 1.107e-08
```

Stage 3 - Interactions There are no interactions believed to be influencing the data.

Checking the model with stepwise regression:

```
ols_step_both_p(incomemodfinal,pent=0.15,prem=0.15,details=T)
```

```
Stepwise Selection Method
-------------------------

Candidate Terms:

1. URB_PER
2. H_ED


Step   => 0
Model  => PEARNVAL ~ 1
R2     => 0

Initiating stepwise selection...

Step      => 1
Selected  => H_ED
Model     => PEARNVAL ~ H_ED
R2        => 0.395

Step      => 2
Selected  => URB_PER
Model     => PEARNVAL ~ H_ED + URB_PER
R2        => 0.541


                              Stepwise Summary
----------------------------------------------------------------------------
Step    Variable         AIC         SBC         SBIC         R2       Adj. R2
----------------------------------------------------------------------------
 0      Base Model     1029.621    1033.445    886.179    0.00000    0.00000
 1      H_ED (+)       1006.458    1012.194    863.717    0.39545    0.38286
 2      URB_PER (+)     994.646    1002.294    853.126    0.54137    0.52186
----------------------------------------------------------------------------
```

Final Model Output

Model Summary

R	0.736	RMSE	4663.477
R-Squared	0.541	MSE	23136192.918
Adj. R-Squared	0.522	Coef. Var	8.331
Pred R-Squared	0.479	AIC	994.646
MAE	3633.934	SBC	1002.294

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	1283585421.133	2	641792710.566	27.74	0.0000
Residual	1087401067.151	47	23136192.918		
Total	2370986488.284	49			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig
(Intercept)	47415.205	4009.212		11.827	0.000
H_EDVocational Associate’s	-6997.660	1433.189	-0.506	-4.883	0.000
URB_PER	191.569	49.539	0.401	3.867	0.000

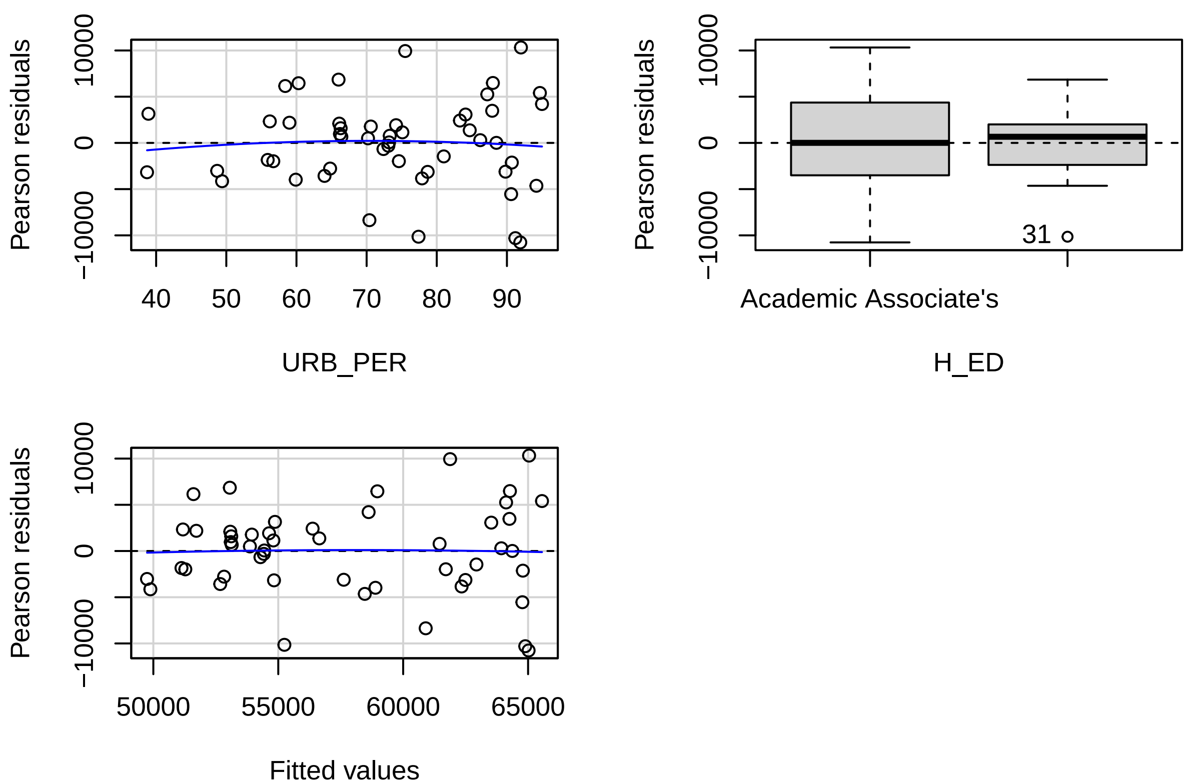
```
#Yipeee!!! the model resulting from this method (consider the EDA) matches the model gen
#(so that means the model building process is pretty sound)
#The model itself is lowkey mid tho
```

Checking assumptions:

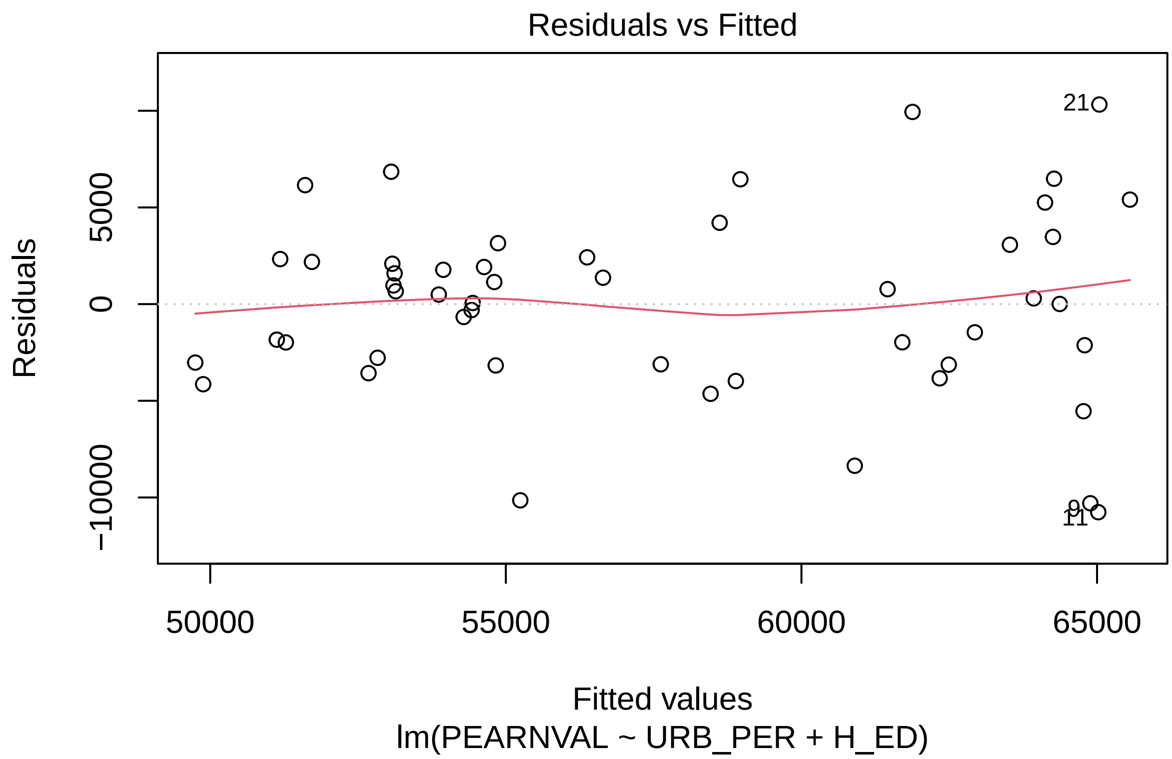
```
#Lack of Fit:
residualPlots(incomemodfinal)
```

```
           Test stat Pr(>|Test stat|)
URB_PER      -0.3779           0.7072
H_ED
Tukey test   -0.1065           0.9152
```

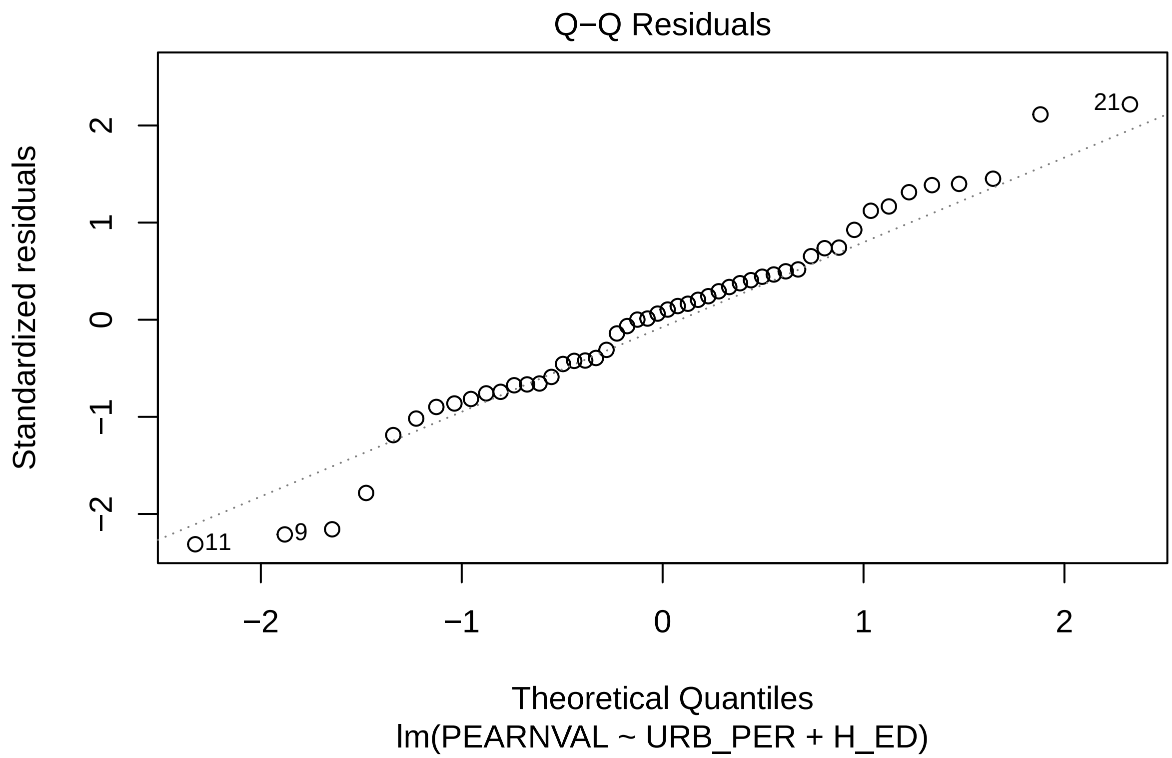
```
#Consant Variance
residualPlots(incomemodfinal,tests=F)
```



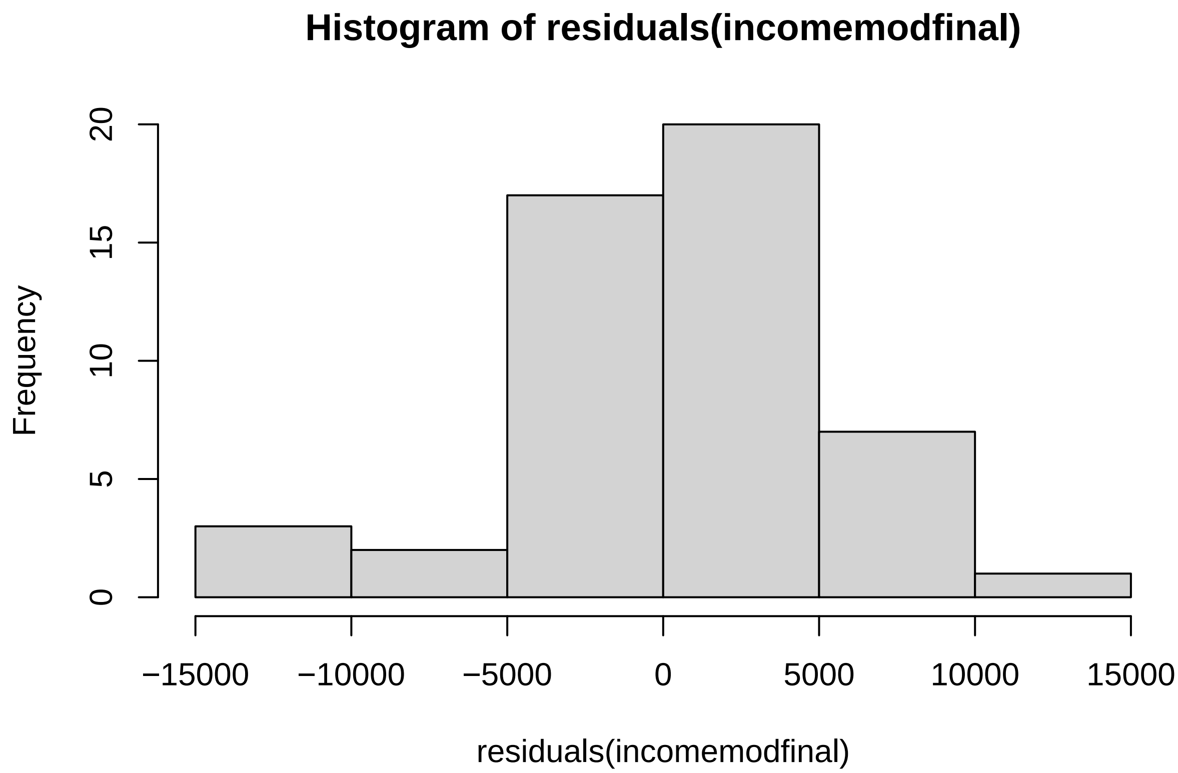
```
plot(incomemodfinal,which=1)
```




```
#Normality:
plot(incomemodfinal,which=2)
```




```
hist(residuals(incomemodfinal))
```

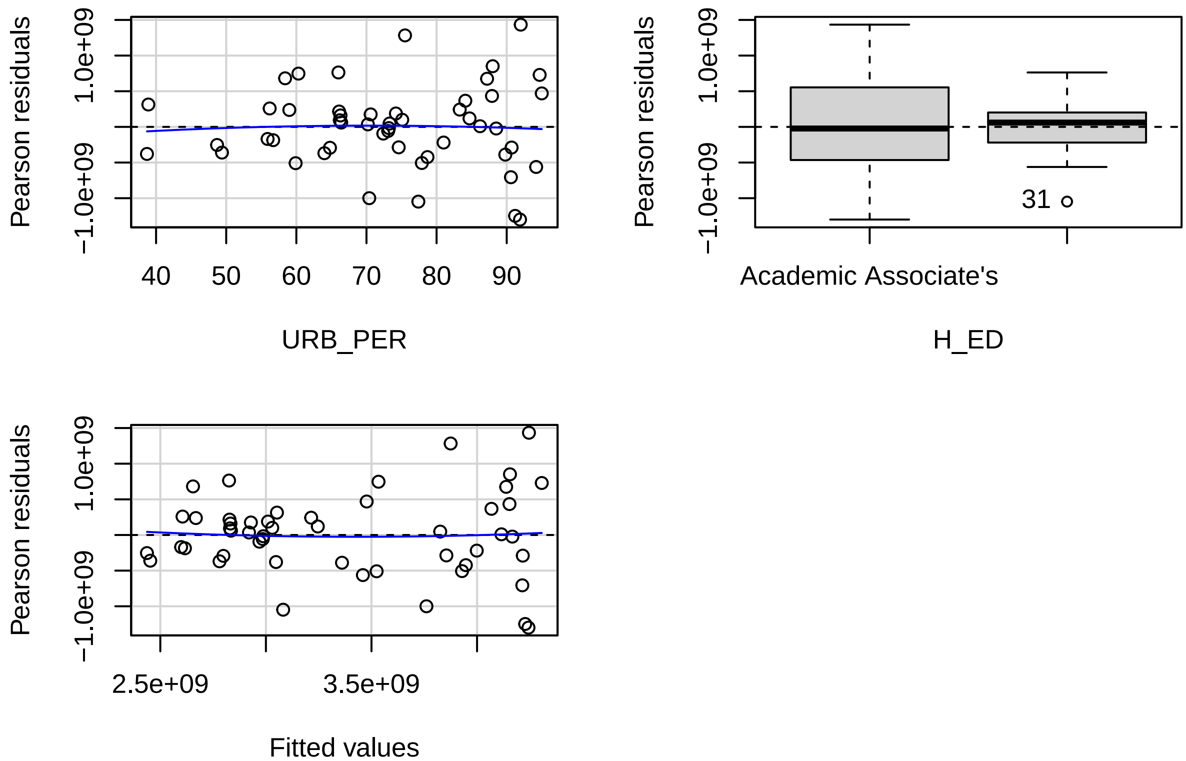



```
#potentially violated:

#NEW TRANSFORMED MODEL
revisedincomemod<-lm(PEARNVAL^2~URB_PER+H_ED,data=incomeData)

##REVISED MODEL
#Lack of Fit:
residualPlots(revisedincomemod)
```

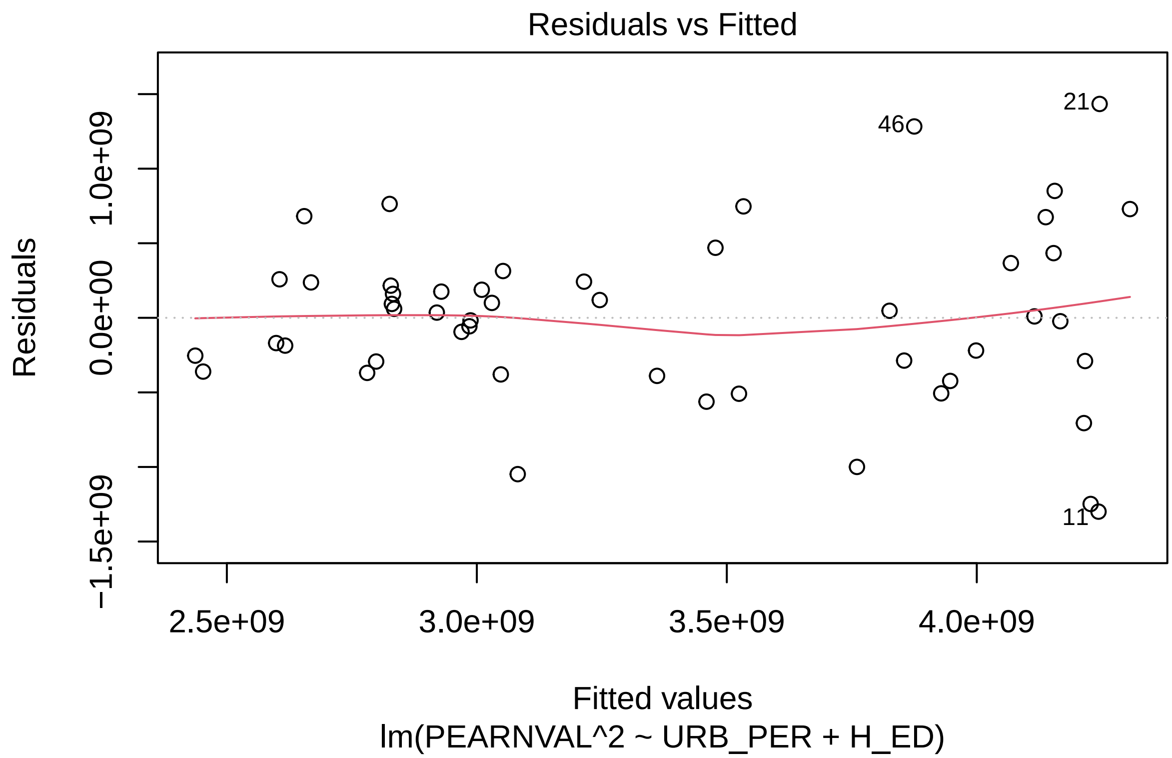
```
           Test stat Pr(>|Test stat|)
URB_PER      -0.2466           0.8063
H_ED
Tukey test    0.2154           0.8294
```

```
#no lack of fit
```

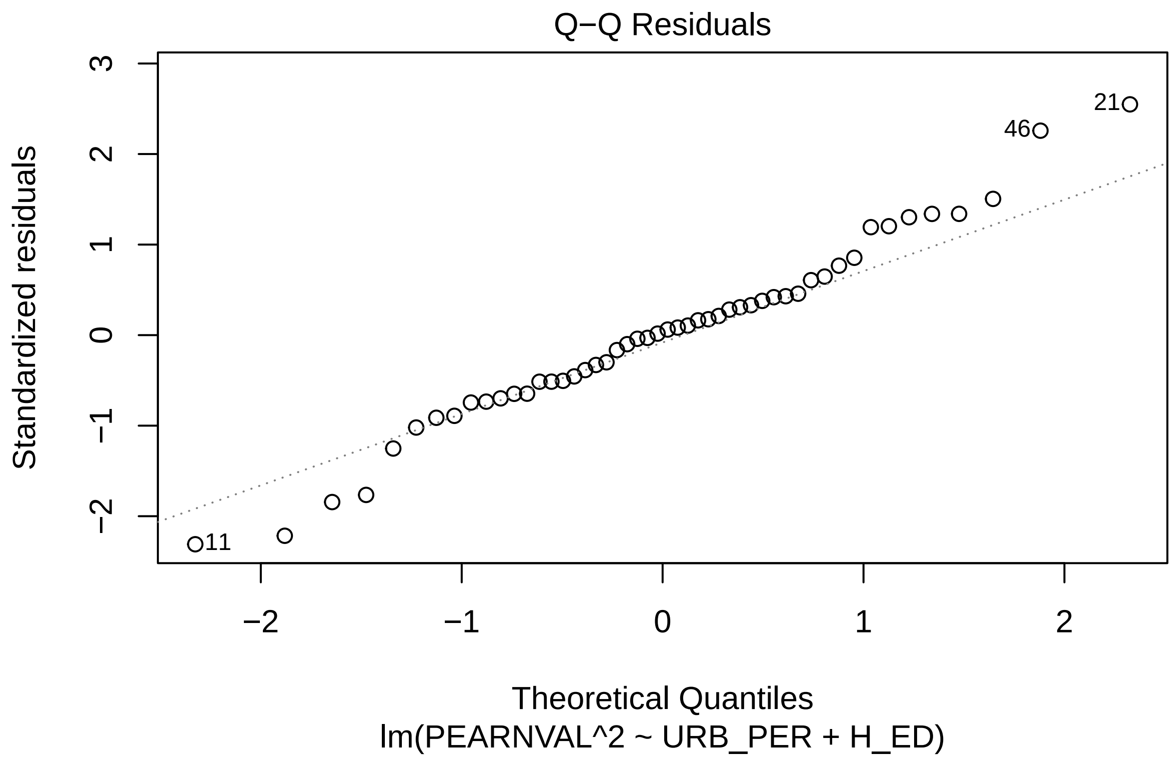
```
#Constant Variance
residualPlots(revisedincomemod,tests=F)
```



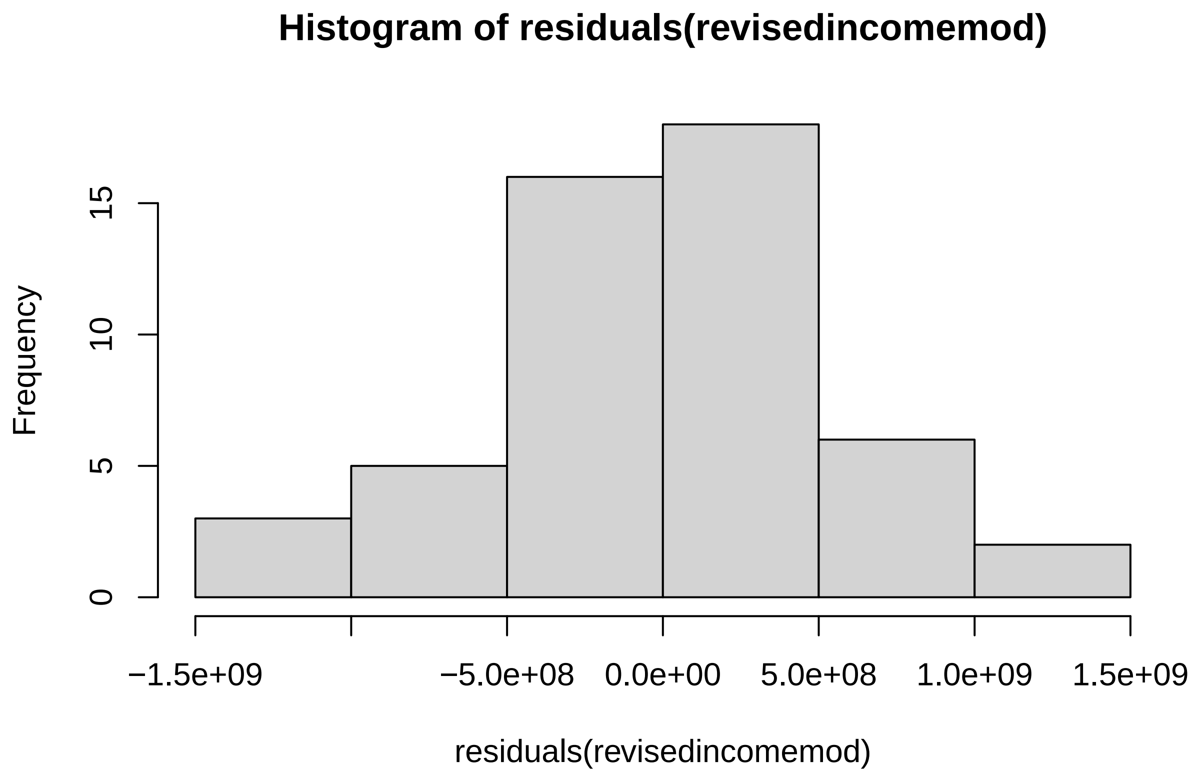

```
plot(revisedincomemod,which=1)
```




```
#Normality:
plot(revisedincomemod,which=2)
```

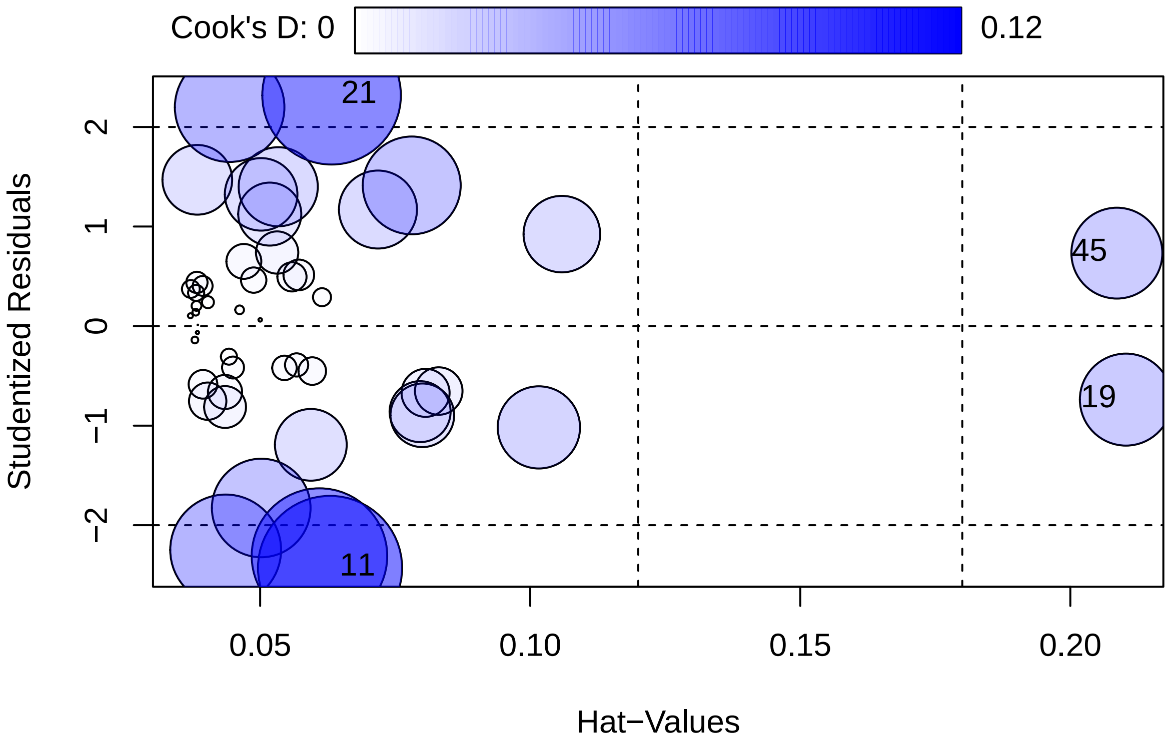



```
hist(residuals(revisedincomemod))
```



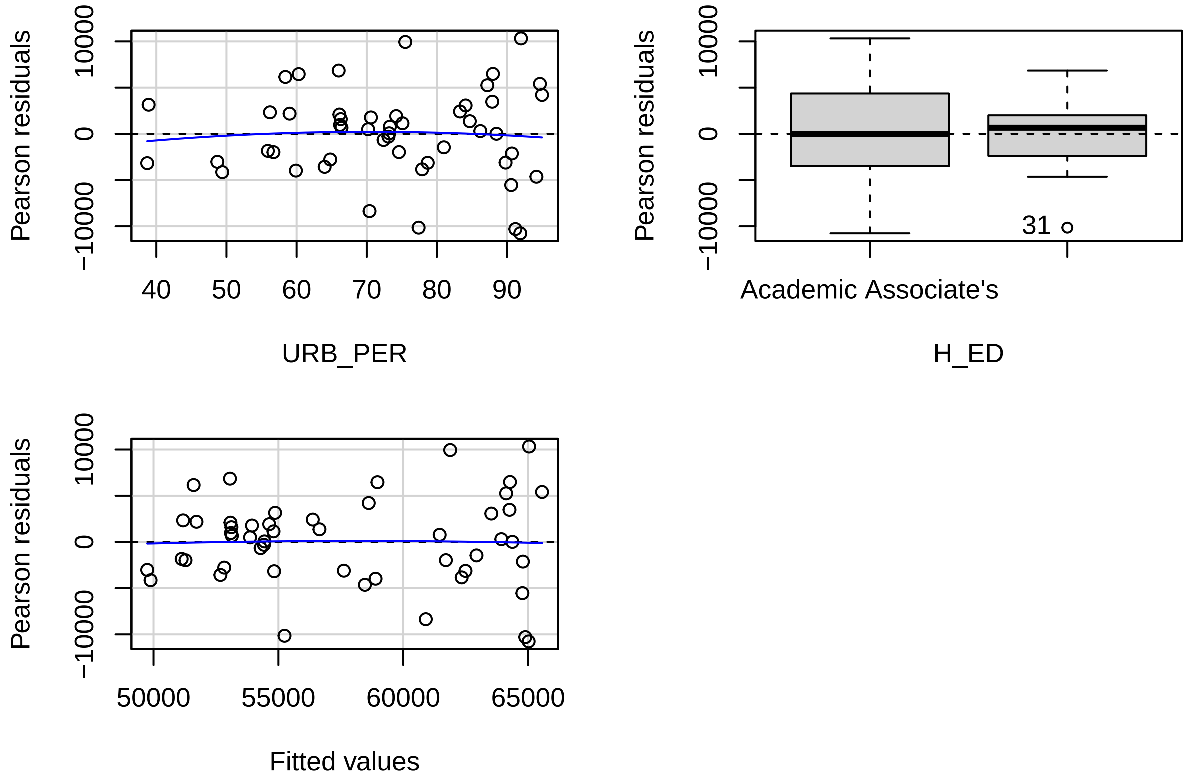
Constant Variance:

```
##Checking final model for influential observations:
influencePlot(incomemodfinal)
```

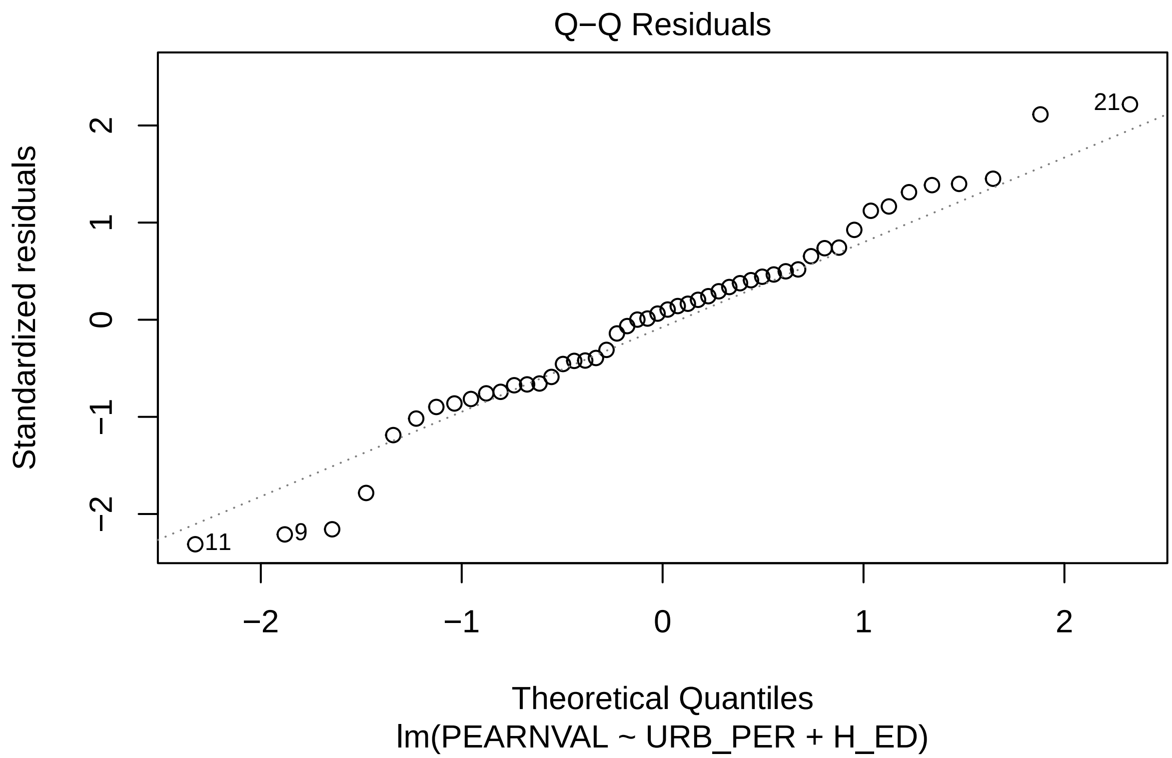



```
      StudRes        Hat      CookD
11 -2.4286854 0.06293499 0.11958773
19 -0.7384973 0.21029459 0.04888333
21  2.3184308 0.06322337 0.11062508
45  0.7328573 0.20861622 0.04766255
```

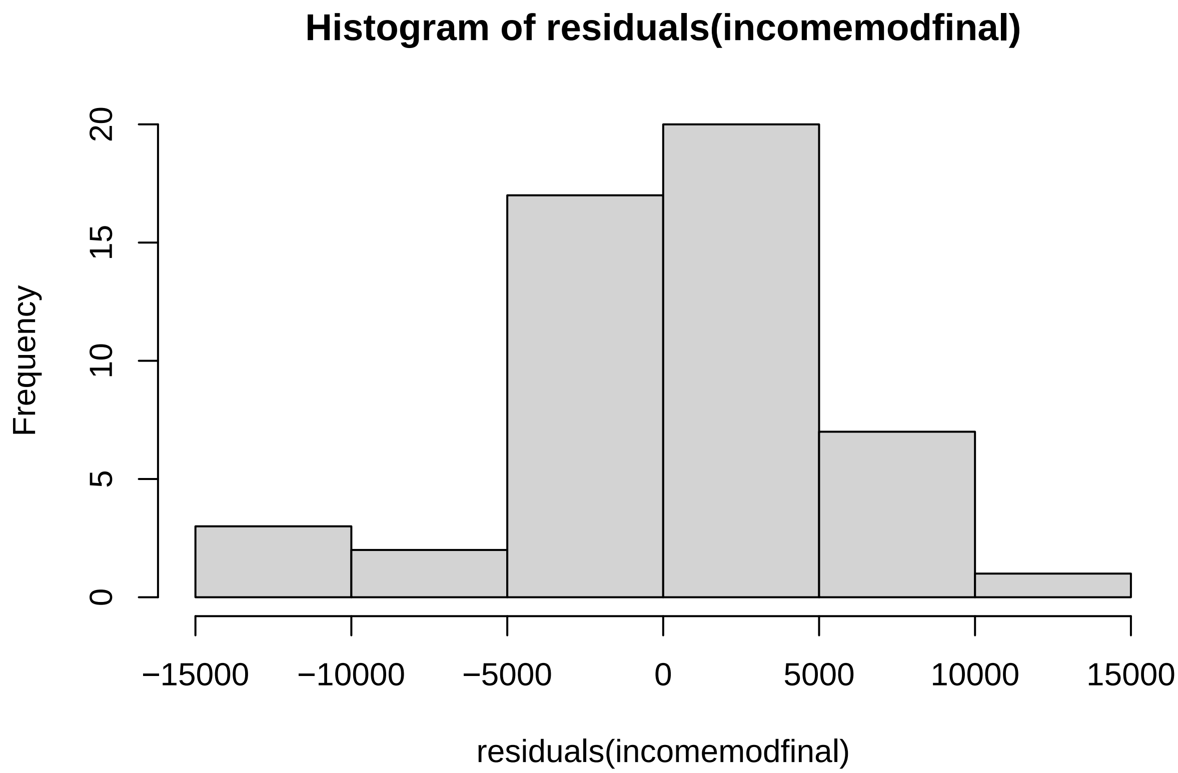
```
#Using the Studentized residual threshold of an absolute value of 2, there are 2 outlier

residualPlots(incomemodfinal,tests=F)
```




```
#Looking at residual plot, there appears to be a fanning relationship with residuals for

plot(incomemodfinal,which=2)
```




```
#QQplot is approximately linear

hist(residuals(incomemodfinal))
```




```
#residuals are approximately normally distributed
```

Remove observations 11, 19, 45, 21

```
subsetincomeData<-incomeData[-c(11,19,45,21),]
subsetincomemod<-lm(PEARNVAL~URB_PER+H_ED,data=subsetincomeData)
summary(subsetincomemod)
```

```
Call:
lm(formula = PEARNVAL ~ URB_PER + H_ED, data = subsetincomeData)

Residuals:
     Min       1Q   Median       3Q      Max
-10328.4  -2955.4    382.2   2298.4   9917.9

Coefficients:
```

```
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  47408.53    4833.78   9.808 1.55e-12 ***
URB_PER                        191.95      58.28   3.294  0.00198 **
H_EDVocational Associate's -7017.40    1488.60  -4.714 2.57e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4433 on 43 degrees of freedom
Multiple R-squared:  0.5795,    Adjusted R-squared:   0.56
F-statistic: 29.63 on 2 and 43 DF,  p-value: 8.141e-09
```

Results

The statistical interpretation of the final model. This should be in statistical terms and overall interpreting and assessing the statistical usefulness of the model with the appropriate metrics. There should be no R output (that will go in the appendix). However, you will include your final model. $PEARNVAL = \beta_0 + \beta_1 URB\_PER + \beta_2 H\_EDassociates$ $PEARNVAL = 47408.52 + 191.95 URB\_PER - 7017.40 H\_EDassociates$

Conclusions

- interpreting your results of the analyses in context of the problem
- commenting on areas of future improvements.

Appendix A: Data Dictionary

Reference Name	Variable Name	Description
State by FIPS Code	`STATEFIPS`	A qualitative measure that identifies the U.S. state (or D.C.) corresponding to the observation by a standardized numeric code. The 51 possible levels are discrete, ranging from 1-56, omitting 3, 7, 14, 43, and 52.
State	`State`	A qualitative measure that identifies the state corresponding to the observation. The 51 possible levels are names of the 50 U.S. states and the District of Columbia.
Educational Attainment	`H_ED`	A qualitative measure that identifies the average of highest education among adult residents of a given state. The three possible levels include a Vocational Associate's Degree, an Academic Associate's Degree, and a Bachelor's Degree.
Majority Sex	`SEX`	A qualitative measure that identifies the predominant sex among a state's adult residents. Two possible levels, male and female, indicate if the adult population of a state is predominately male or female.
Health Status	`HEA`	A qualitative measure that reports the average health status of a state's residents. Two levels, very good health and good health indicate the average health status of a state's residents.
Personal Earnings	`PEARNVAL`	A continuous quantitative measure that reports the average personal earnings of a state's residents, reported in U.S. Dollars. Possible values within the data range from $45096.53 to $95387.40.
Age	`AGE`	A continuous quantitative measure that reports the average age of a state's adult residents in years. Values range from 40.83460 to 46.39759.

Reference Name	Variable Name	Description
Unemployment Rate	`UNEMP_RATE`	A continuous quantitative measure of a state's unemployment rate from 2020. Unemployment rate is reported as a percentage; the range of possible values within the data is from 4.2% to 13.5%.
Sales Tax Rate	`TAX_RTE`	A continuous quantitative measure of a state's sales tax. Sales Tax Rate is reported as a numerical figure; the range of possible values within the data is from 0.0% (0% sales tax) to 7.25% (7.25% sales tax).
Percentage of Urban Residents	`URB_PER`	A continuous quantitative measure of a state's proportion of urban residents to nonurban residents. This variable is reported as a percentage; the range of possible values within the data is from 38.7% to 100.0%.
Work Expenses	`WRK_SPND`	A continuous quantitative measure that identifies the average amount of money spent on work-related expenses among residents of a state, reported in U.S. Dollars. Possible values in the data range from $1101.676 to $1463.411.

Appendix B: Data Rows

	STATEFIPS	State	H_ED	SEX	HEA
1	1	Alabama	Vocational Associate's	Male	Good Health
2	2	Alaska	Vocational Associate's	Male	Good Health
3	4	Arizona	Vocational Associate's	Male	Good Health
4	5	Arkansas	Vocational Associate's	Male	Good Health
5	6	California	Vocational Associate's	Male	Good Health
6	8	Colorado	Academic Associate's	Male	Good Health
7	9	Connecticut	Academic Associate's	Male	Good Health
8	10	Delaware	Vocational Associate's	Male	Good Health
9	12	Florida	Academic Associate's	Male	Very Good Health
10	13	Georgia	Vocational Associate's	Female	Good Health
11	15	Hawaii	Academic Associate's	Male	Good Health
12	16	Idaho	Vocational Associate's	Male	Good Health
13	17	Illinois	Academic Associate's	Male	Good Health
14	18	Indiana	Vocational Associate's	Male	Good Health
15	19	Iowa	Vocational Associate's	Male	Good Health

	PEARNVAL	AGE	UNEMP_RATE	TAX_RTE	URB_PER	WRK_SPND
1	53905.05	42.60043	6.4	0.0400	59.0	1142.836
2	59908.18	43.33103	8.3	0.0000	66.0	1234.210
3	54509.31	41.63326	7.8	0.0560	89.8	1229.184
4	53513.54	43.28300	6.2	0.0650	56.2	1193.809
5	62824.72	42.26563	10.1	0.0725	95.0	1237.779
6	64224.86	43.52050	6.8	0.0290	86.2	1326.110
7	70758.66	45.24870	7.9	0.0635	88.0	1250.562
8	58795.20	43.32292	7.5	0.0000	83.3	1195.540
9	54585.91	44.37481	8.1	0.0600	91.2	1176.506
10	55946.86	42.54033	6.5	0.0400	75.1	1231.916
11	54258.90	45.30385	11.7	0.0400	91.9	1232.113
12	55717.66	42.08074	5.5	0.0600	70.6	1303.237
13	64375.88	43.44074	9.3	0.0625	88.5	1292.794
14	53621.19	42.57899	7.3	0.0700	72.4	1308.013
15	49106.65	42.04989	5.2	0.0600	64.0	1384.532

Appendix C: Tables and Figures

Appendix D: References

Background

- Bureau, U. C. (2021, November 23). Why we conduct the decennial census of Population and Housing. Census.gov. https://tinyurl.com/5fdyh82c
- Mather, M., & Scommegna, P. (2019, March 15). Why is the U.S. Census so important?. Population Reference Bureau https://www.prb.org/resources/importance-of-u-s-census/
- Farley, R. (2020, January 31). The importance of census 2020 and the challenges of getting a complete count. Harvard Data Science Review. https://hdsr.mitpress.mit.edu/pub/rosc6trb/release/3

Data

- 2020 Unemployment Rates: U.S. Bureau of Labor Statistics. (2024). Unemployment rates for states. U.S. Bureau of Labor Statistics. https://www.bls.gov/lau/lastrk20.htm
- Urban percentage of the population for states, historical. Urban Percentage of the Population for States, Historical | Iowa Community Indicators Program. (2024.). https://www.icip.iastate.edu/tables/population/urban-pct-states
- State and local sales tax rates, 2020. Tax Foundation. (2024, February 22). https://taxfoundation.org/data/all/state/2020-sales-taxes/
- Bureau, U. C. (2022, October 27). 2020 annual social and economic supplements. Census.gov. https://www.census.gov/data/datasets/2020/demo/cps/cps-asec-2020.html
- ASEC 2020 Public Use Data Dictionary. (2020). https://tinyurl.com/3h8vexva

Supplemental Code and Analysis Help

1. List your references used to learn more about your techniques and coding here https://rpubs.com/muxicheng/1004550