Jason Larosiliere

Professor Lamere

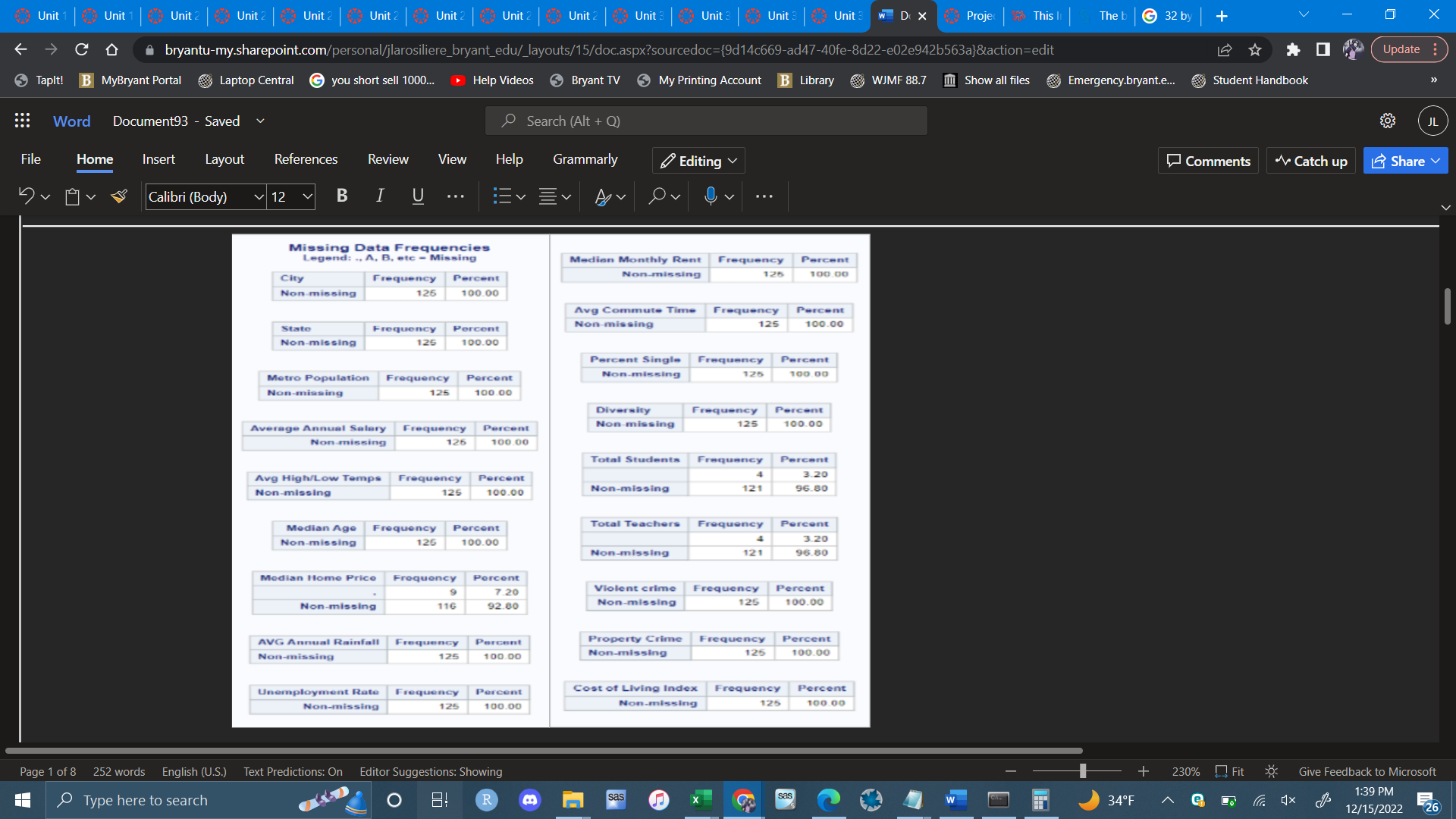
Applied Multivariate Stat

17 December 2022

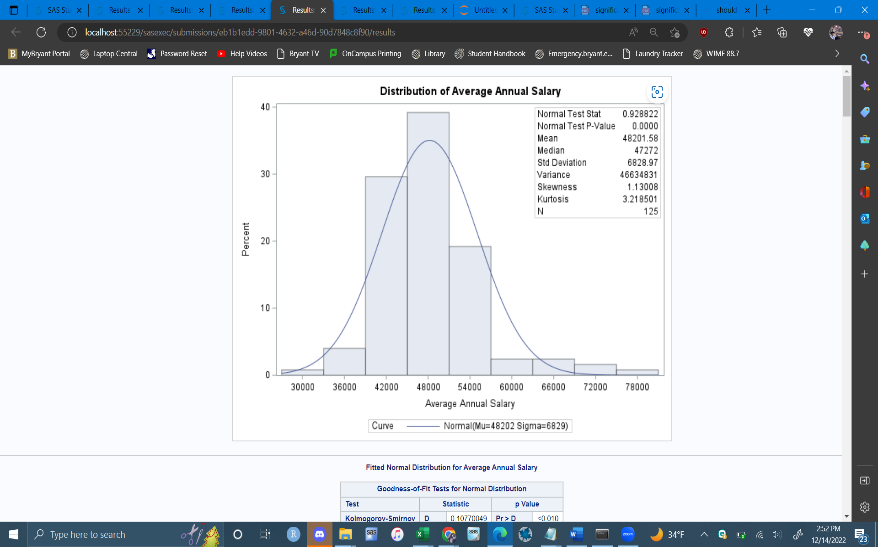
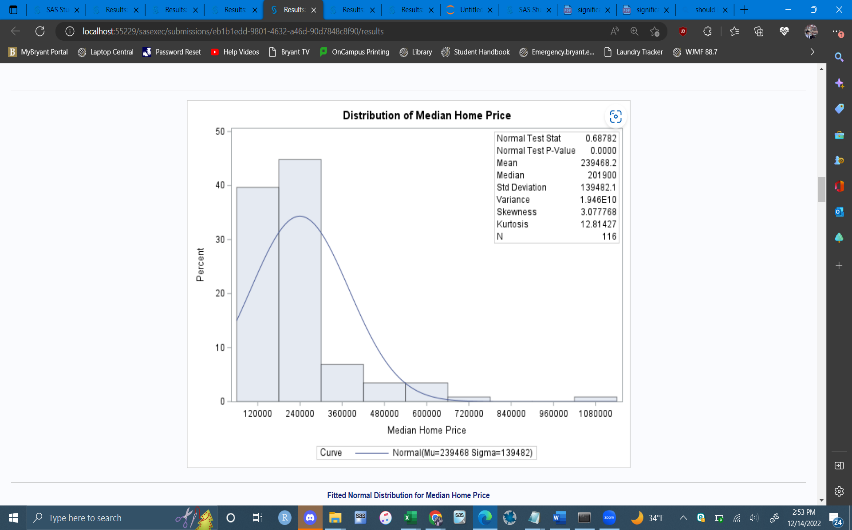
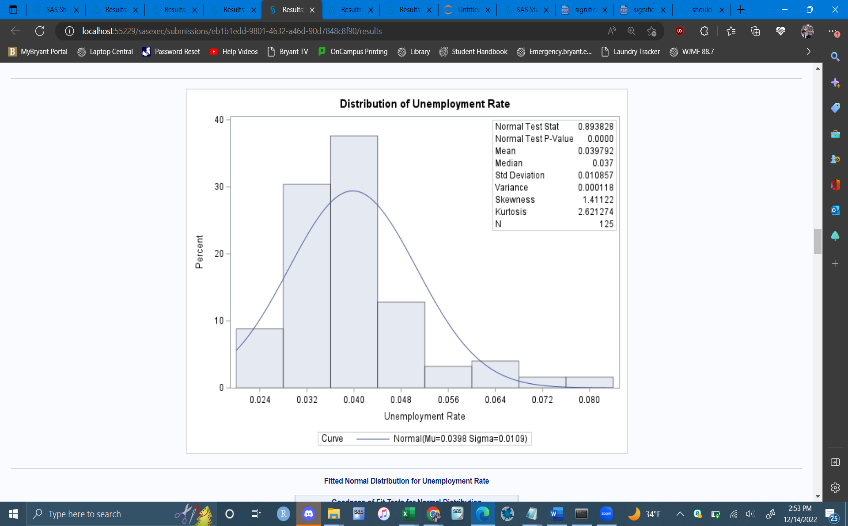
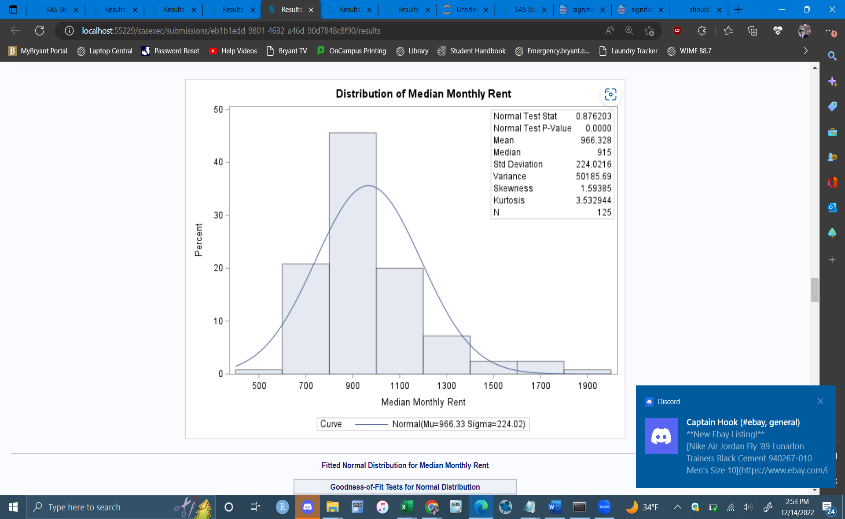
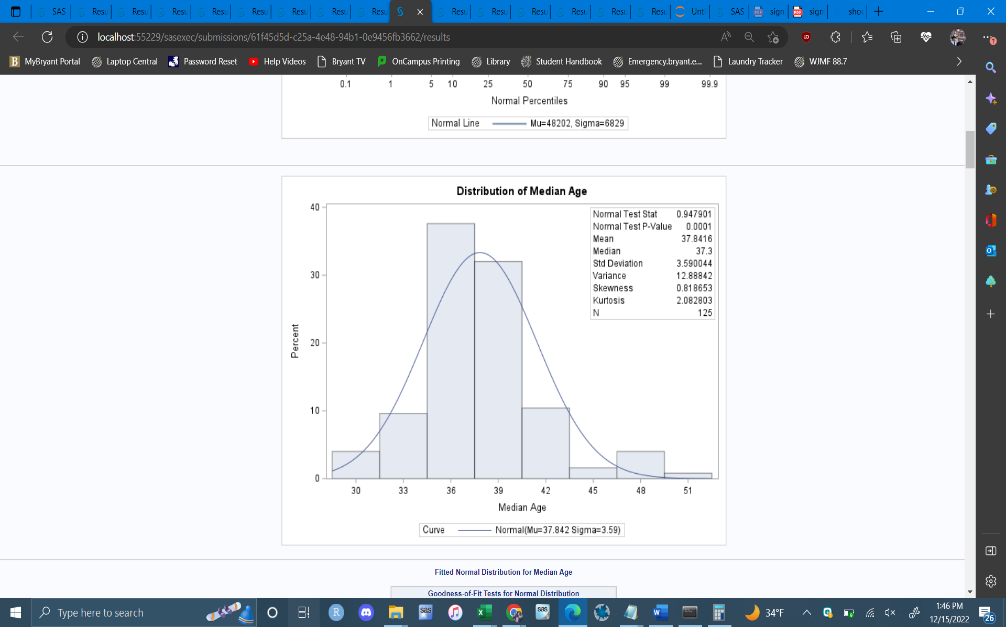
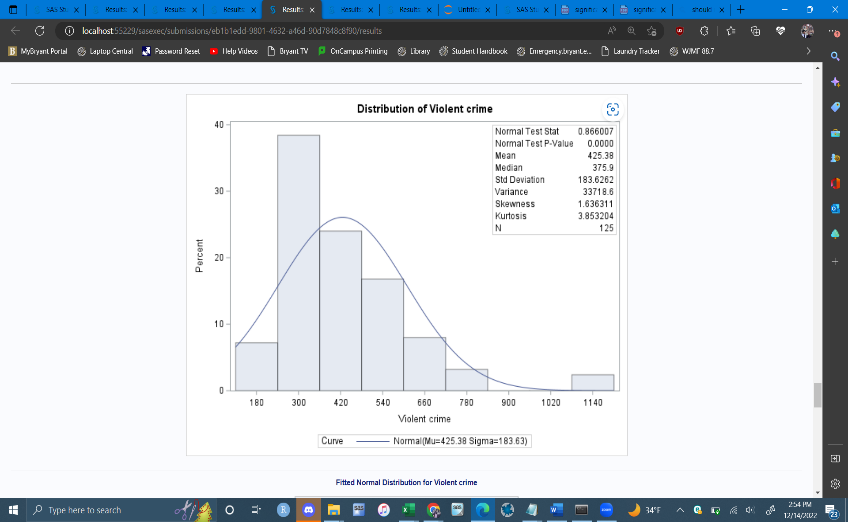
Final Project: Cost of Living in the U.S.

The dataset I have includes 125 random cities and 16 variables. This topic has always been something in the back of my mind. Eventually in a few years, where to start my career is something I must decide, so how do I choose what is the best city to settle down in? There are a lot of factors that go into making this move but it's important to first consider the cost of living. My main purpose here is to create a model that first analyzes the factors to find underlying dimensions, then find the correlation with them, and my response variable. In the end this will give you a clearer insight into the cost of living and what affects this.

This dataset was put together on Excel using several other datasets found on Kaggle, Data.World and AdvisorSmith. It has 125 observations and 16 variables with cost-of-living index as the response. I did have some issues with my data. I originally had 250 observations but did not realize half of them were just duplicated over. Fortunately, the ratio of observations to variables is still above 5 to 1.

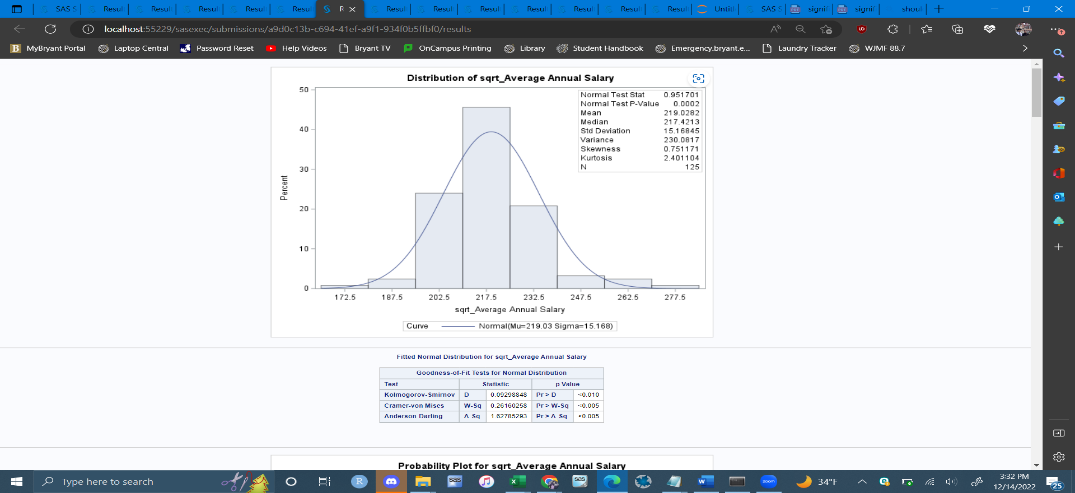
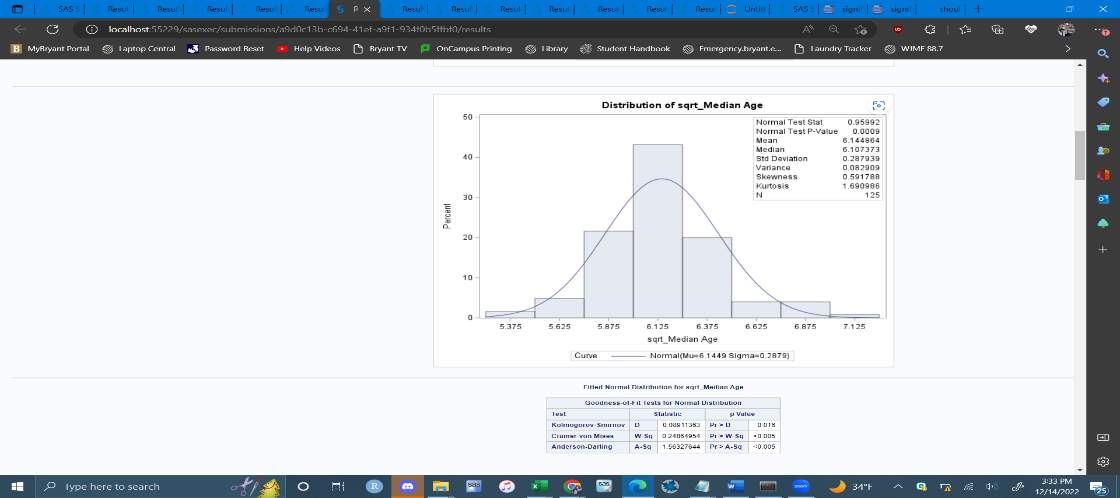
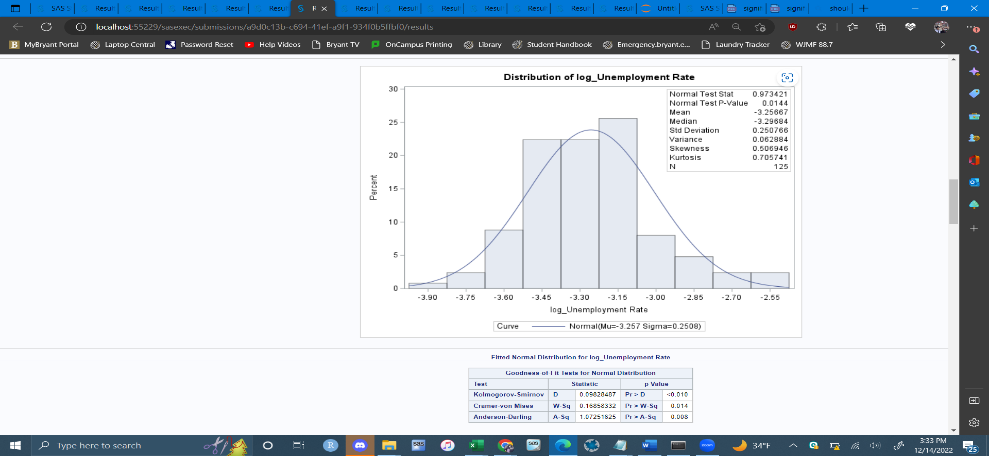
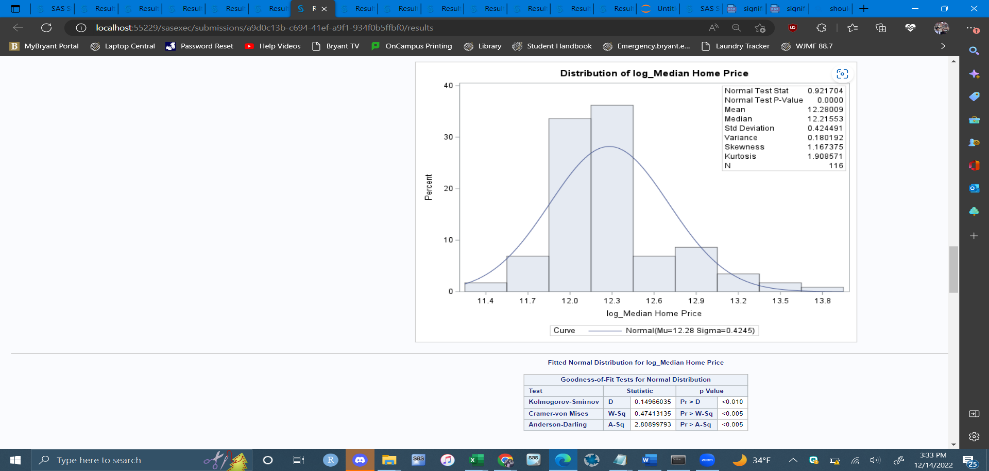
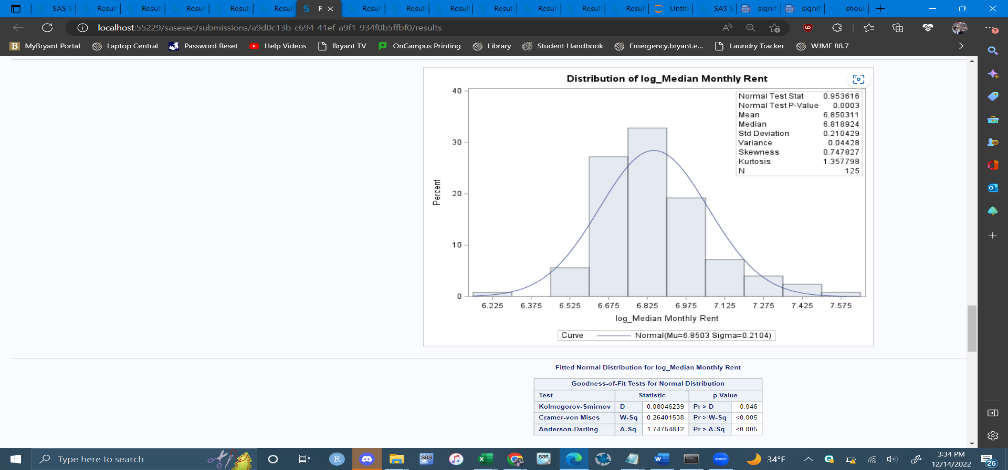
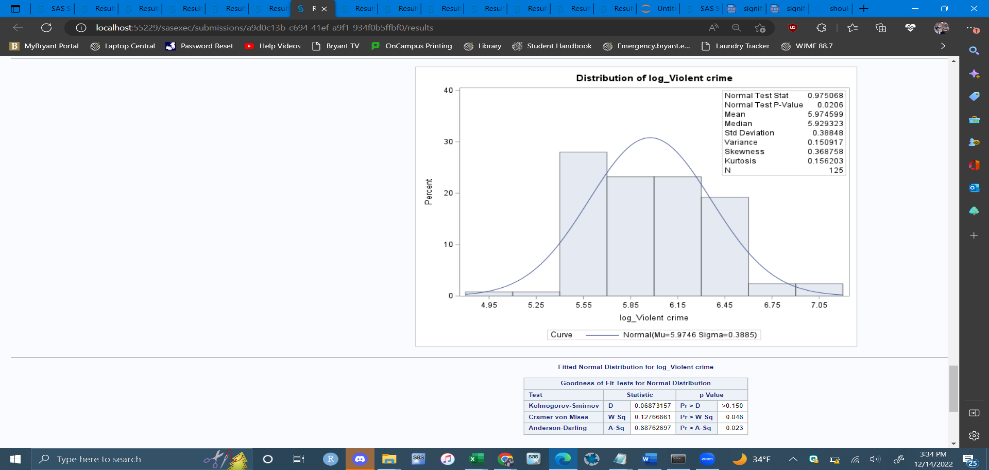
After checking for missing values, I found that all observations contained nearly all their data and only 3 variables had missing data. Since it is less than 10%, the amount is not high enough to affect our results. Therefore, we will ignore it.

Before Transformations:

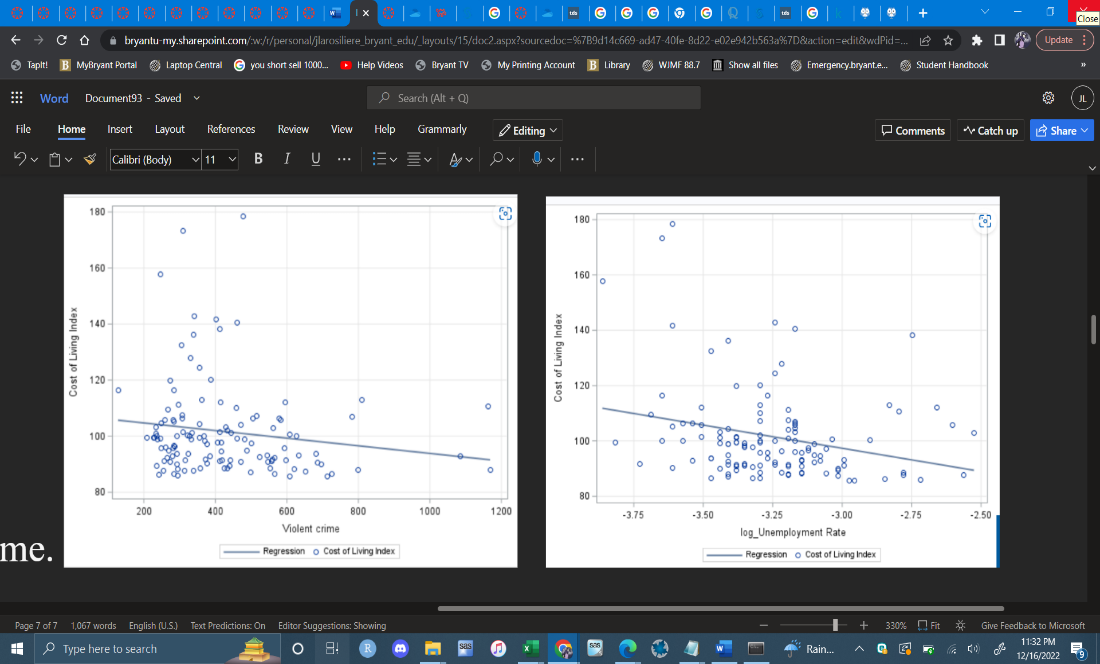
After running a Distribution Analysis, I found that the independent variables, median home price, average annual salary, median age, unemployment rate, median monthly rent, total students, total teachers, property and violent crime all violate assumptions of normality. As you can see all the histograms are skewed to the right and the mean-variance ratio of all variables allows for transformation.

After Transformation:

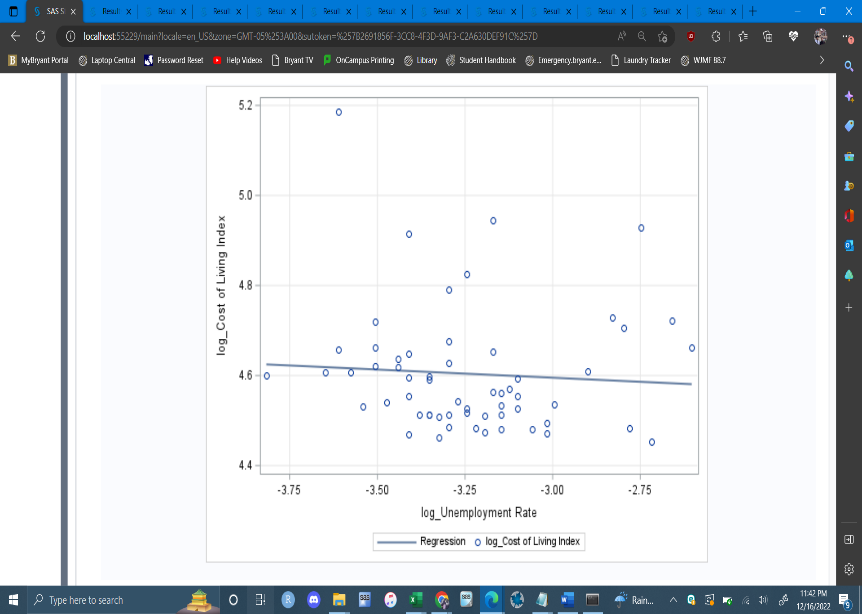
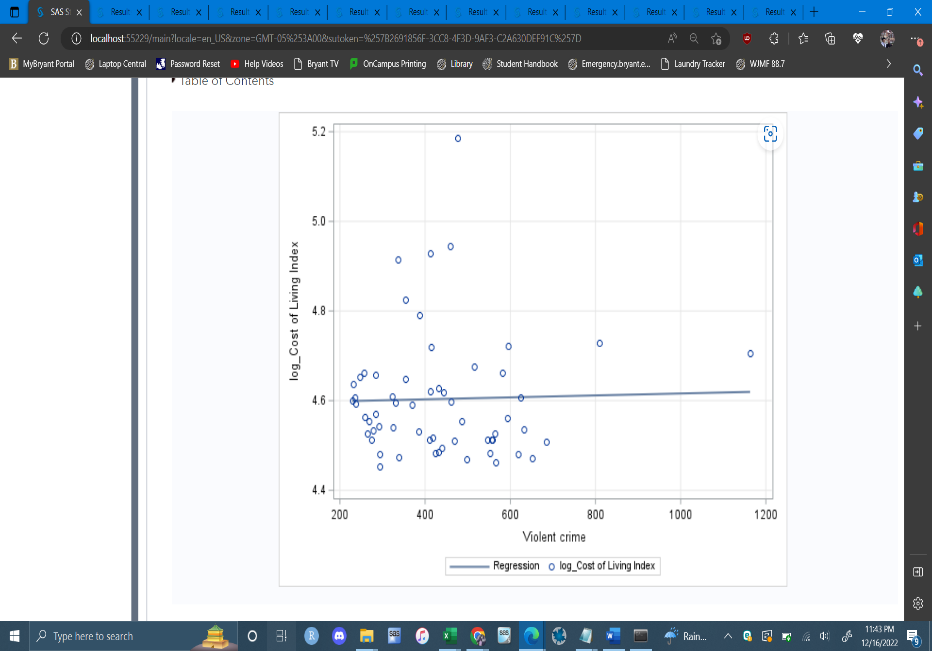
     

As you can see the transformations brought skewness of all the variables at a normal range if not close. The histograms also appear to be no longer strongly skewed in any direction.

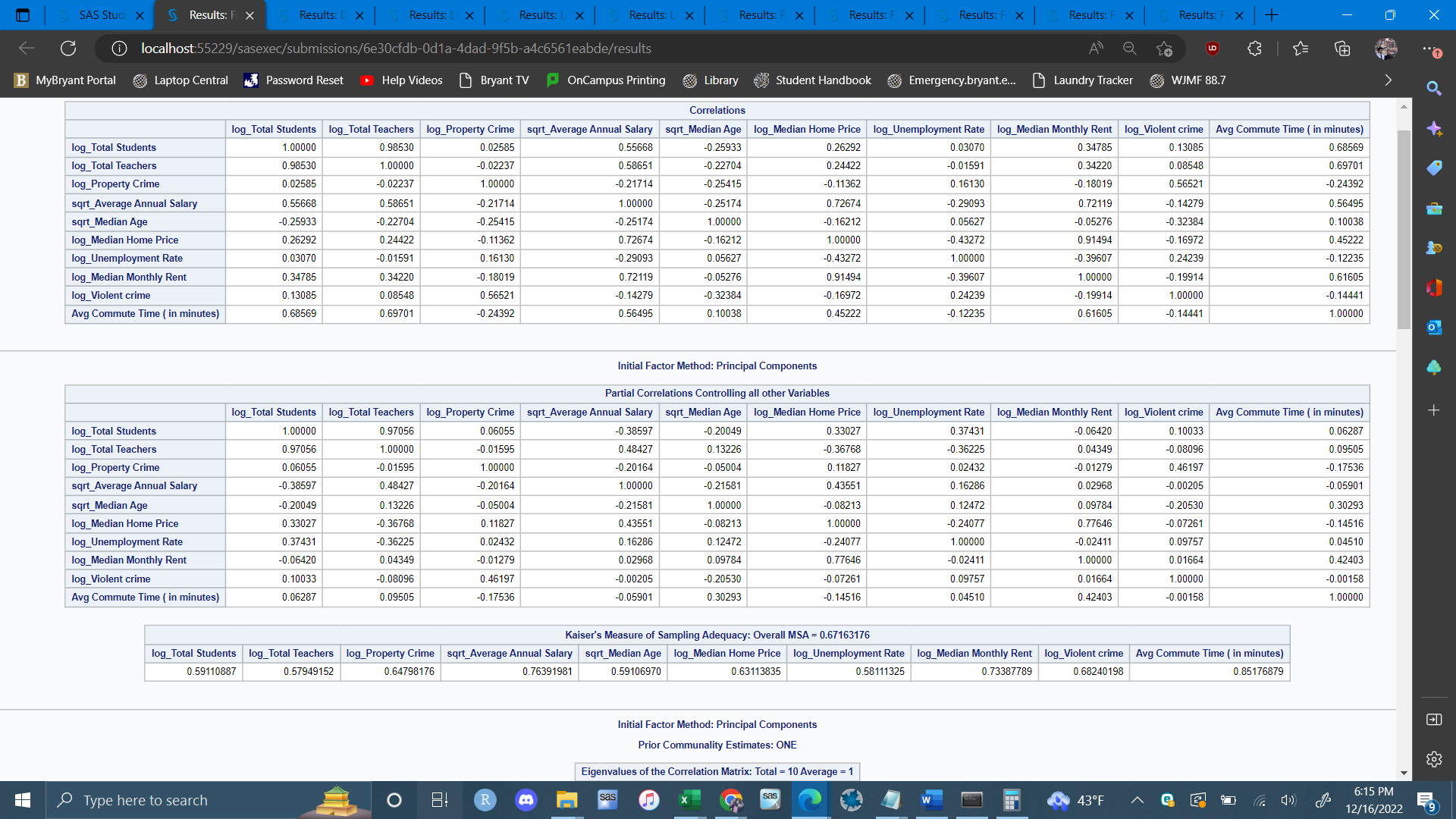
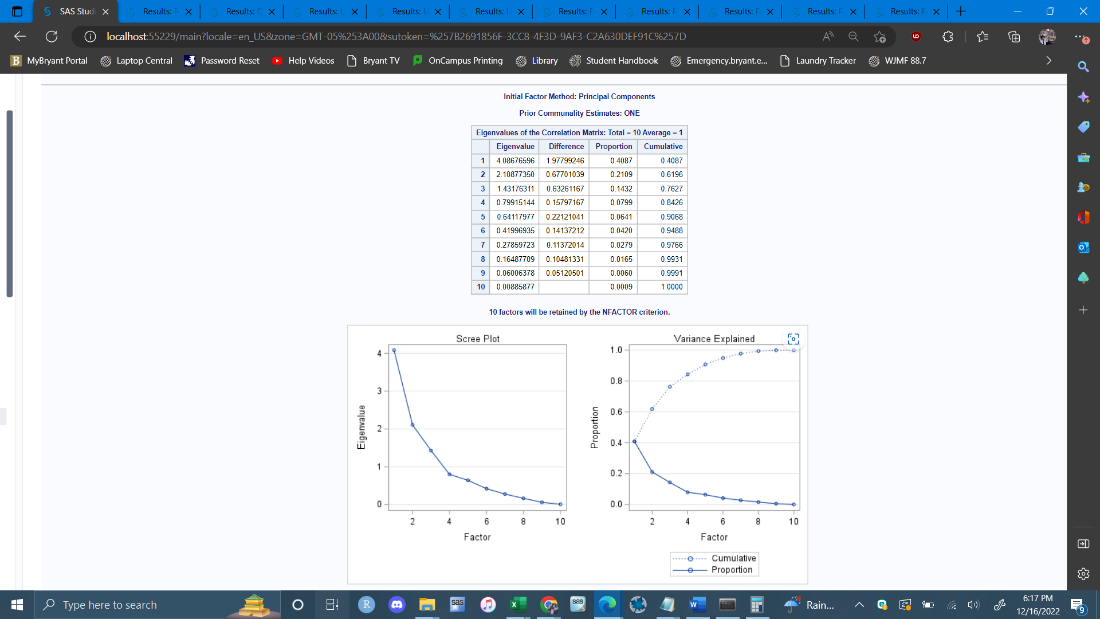
We could also check for normality through the ks-statistic and p-value.

As for homoscedasticity, the variables that violate this are median age, average rainfall, unemployment rate, % single and violent crime. To deal with the violations I will transform the response variable cost-of-living by using the log. 

After Transformation:

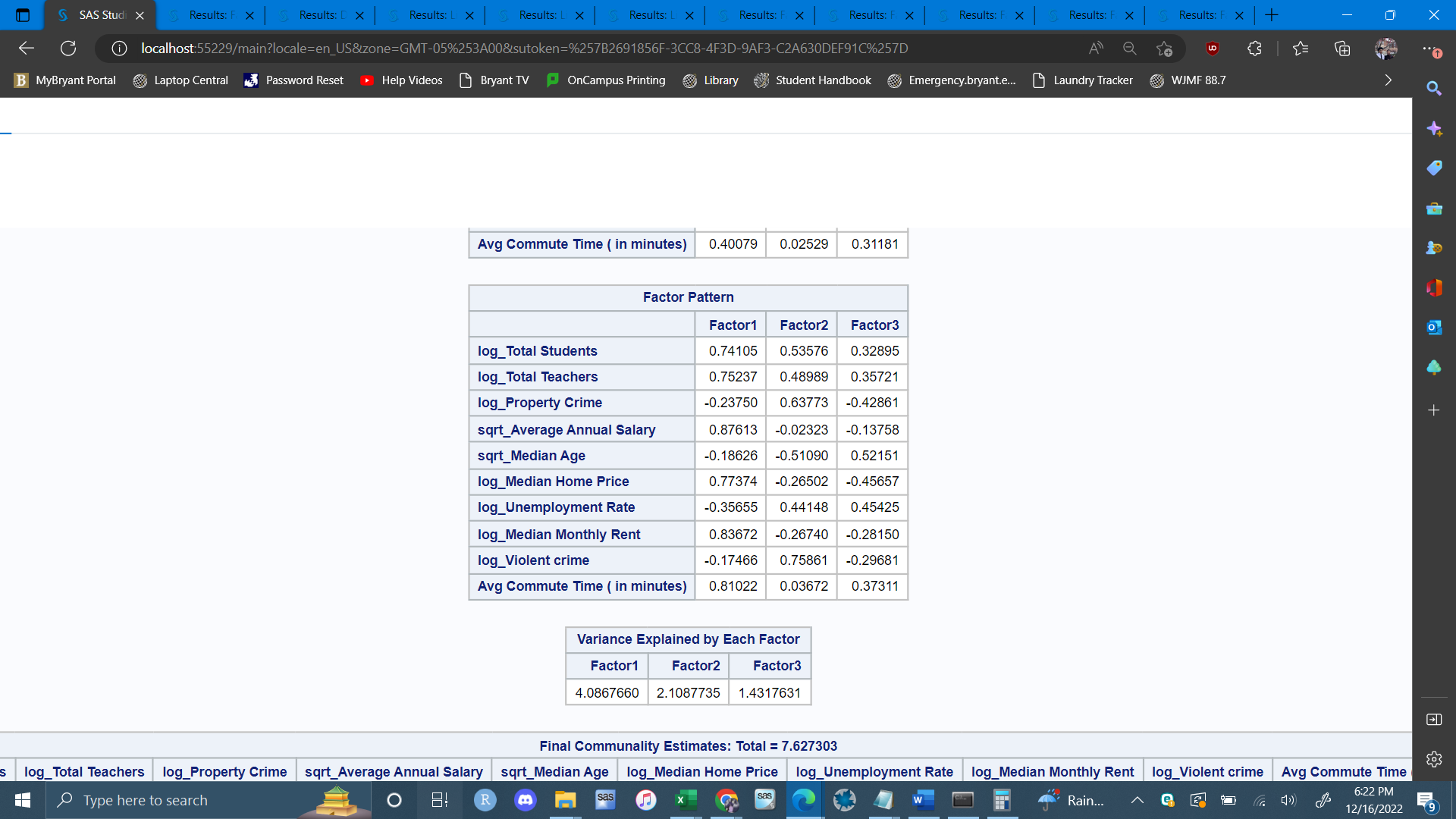
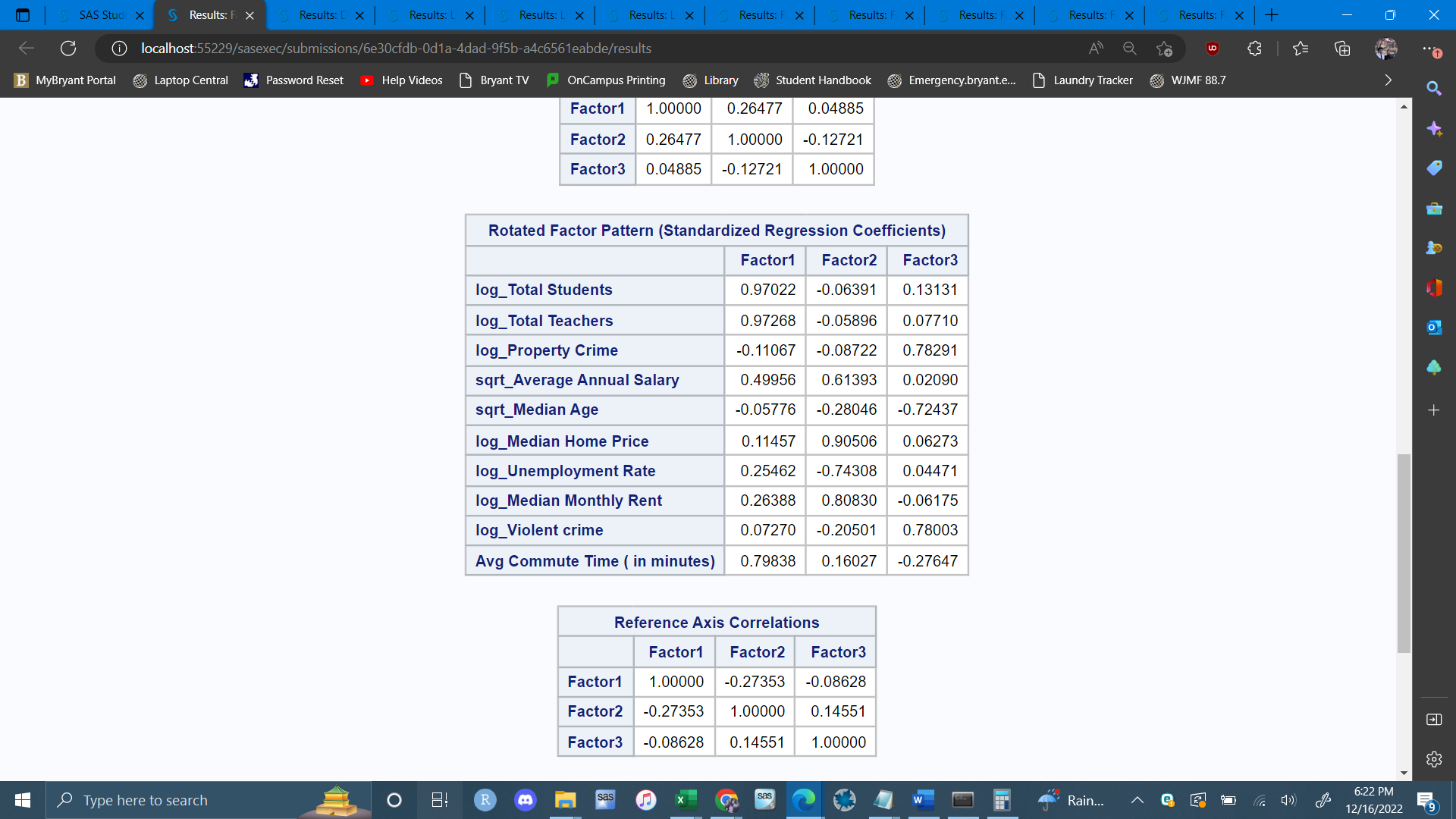
 

For Dimension Reduction, I decided to go with factor analysis. I feel like there's definitely an underlying structure in the variables. After running the Factor Analysis the first time, the overall MSA was .64 and the variable annual rainfall had a kaisure measure of below .5. I got rid of this variable and ran it a second time and the overall MSA went up to .67 and there were no longer any variables under 0.5. As for the number of factors to use, I decided to go with 3 factors. These three factors explain about 75% of the variance and are the only factors with eigenvalues greater than 1.

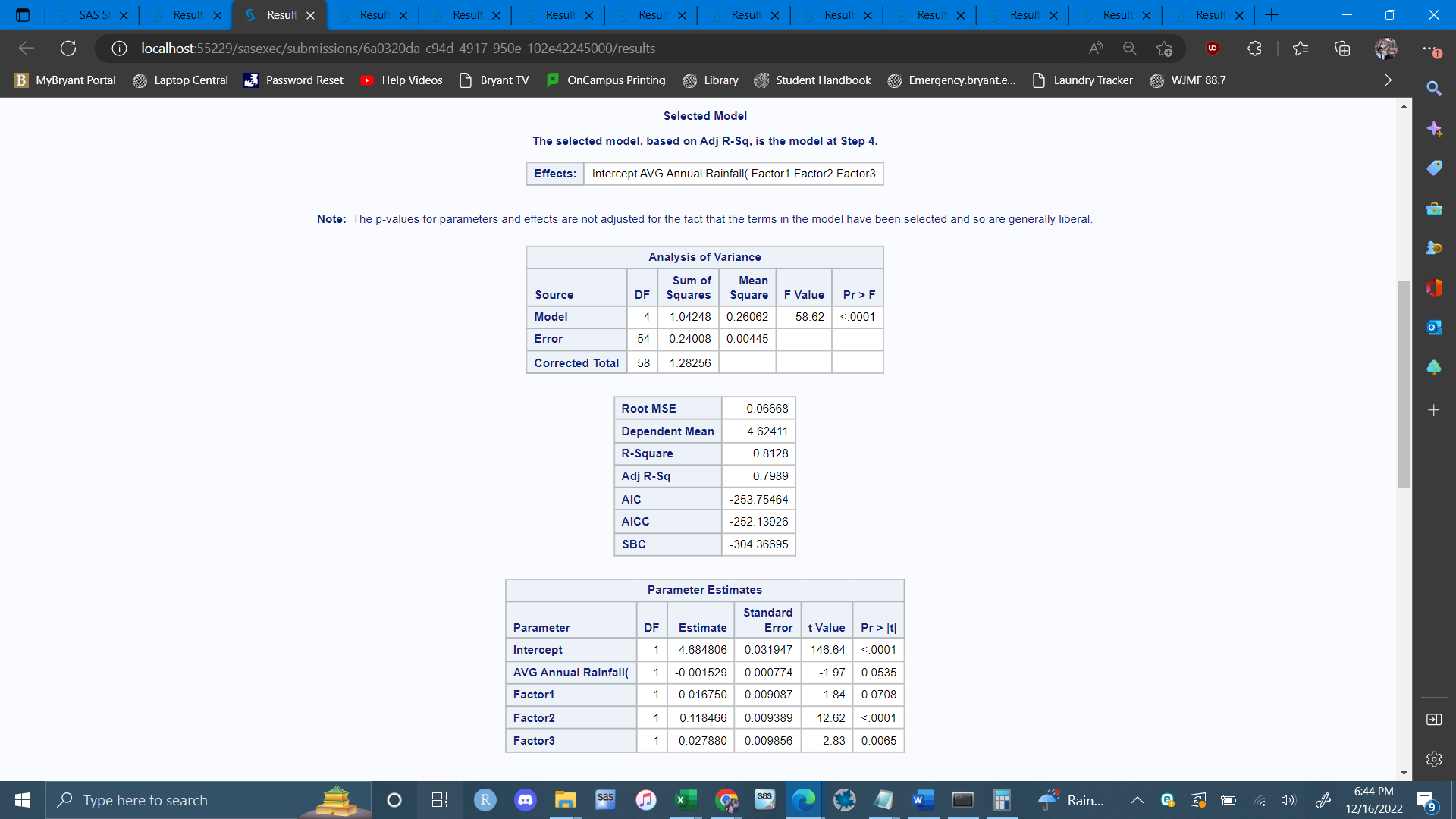
 

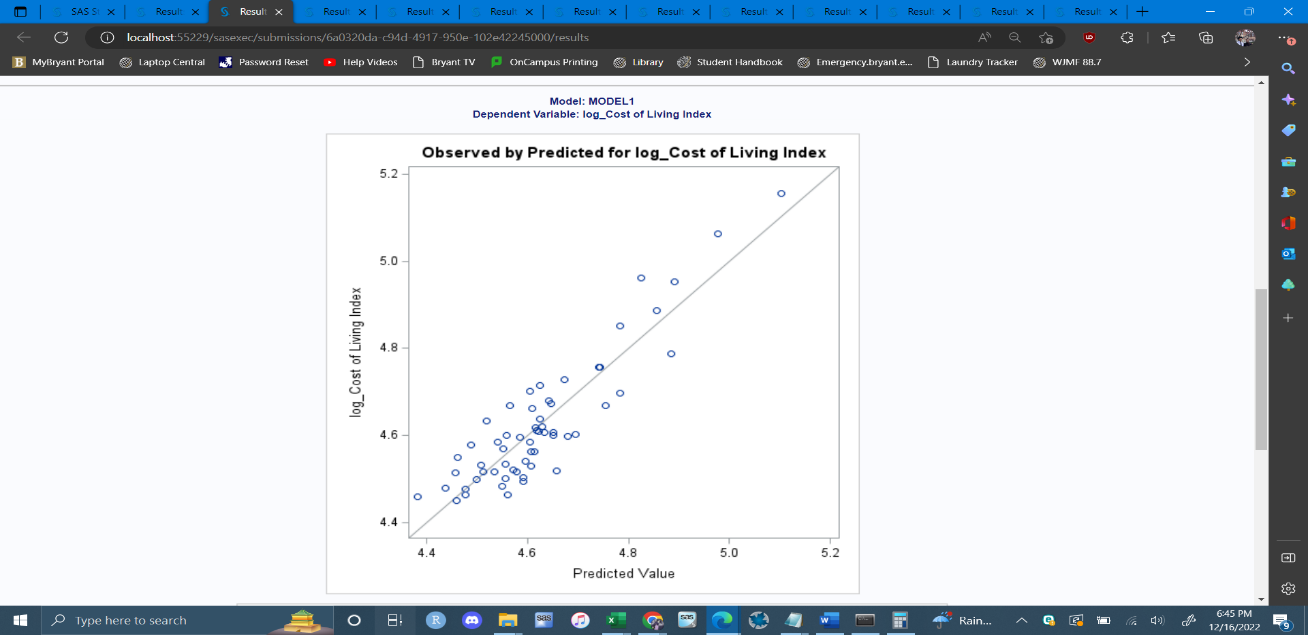
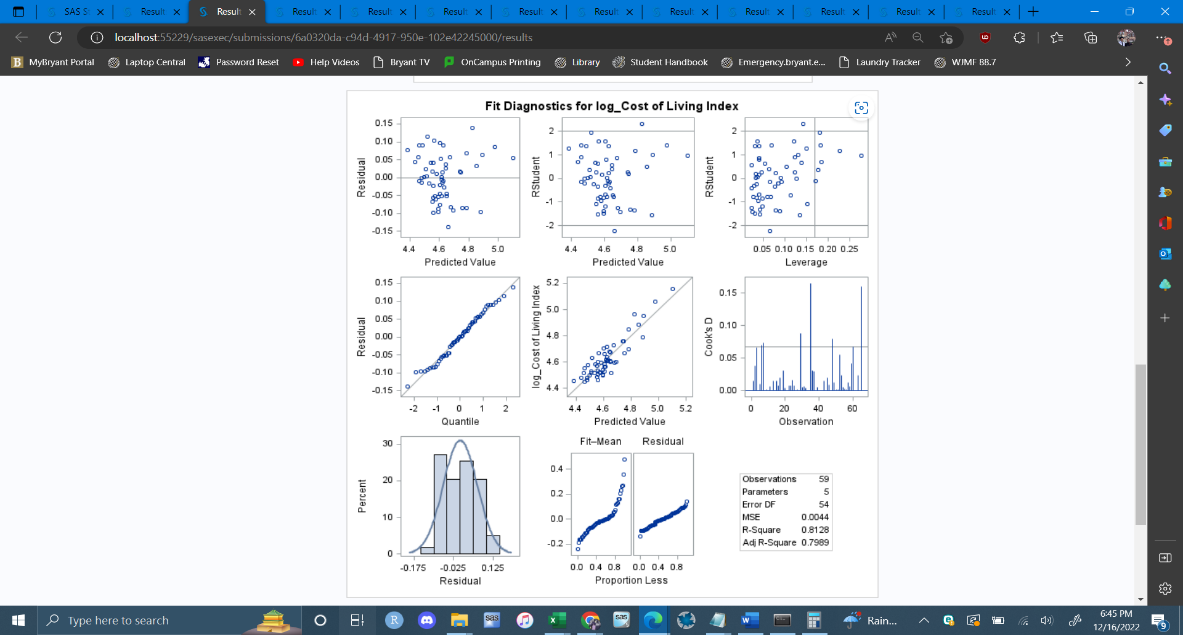
After running the Factor Analysis with 3 factors, I feel like there is definitely a need for rotation. I tried different types of rotation and found that oblimin rotation gave the best results. As you can see there’s no more signs of cross loading and the variables that belong to each factor show strong correlation.

Before Rotation: After Rotation:

The variables average commute time, total teachers and total students were assigned to Factor 1. The variables median monthly rent, unemployment rate, median home price and average annual salary were assigned to Factor 2 and violent crime, median age and property crime were assigned to Factor 3. Factor 1 seems to represent the school system and how long teachers take to get to schools. Factor 2 seems to represent the cost of a home and how to relates to work and pay. In a city, jobs will probably pay people in relation to how expensive it is to own a place and vice versa. Factor 3 appears to represent people in the city and the environment. It includes how many of them are not in a relationship, whether they are committing acts of violence or property crimes and their age. In conclusion, I feel like this method worked great with the dataset.



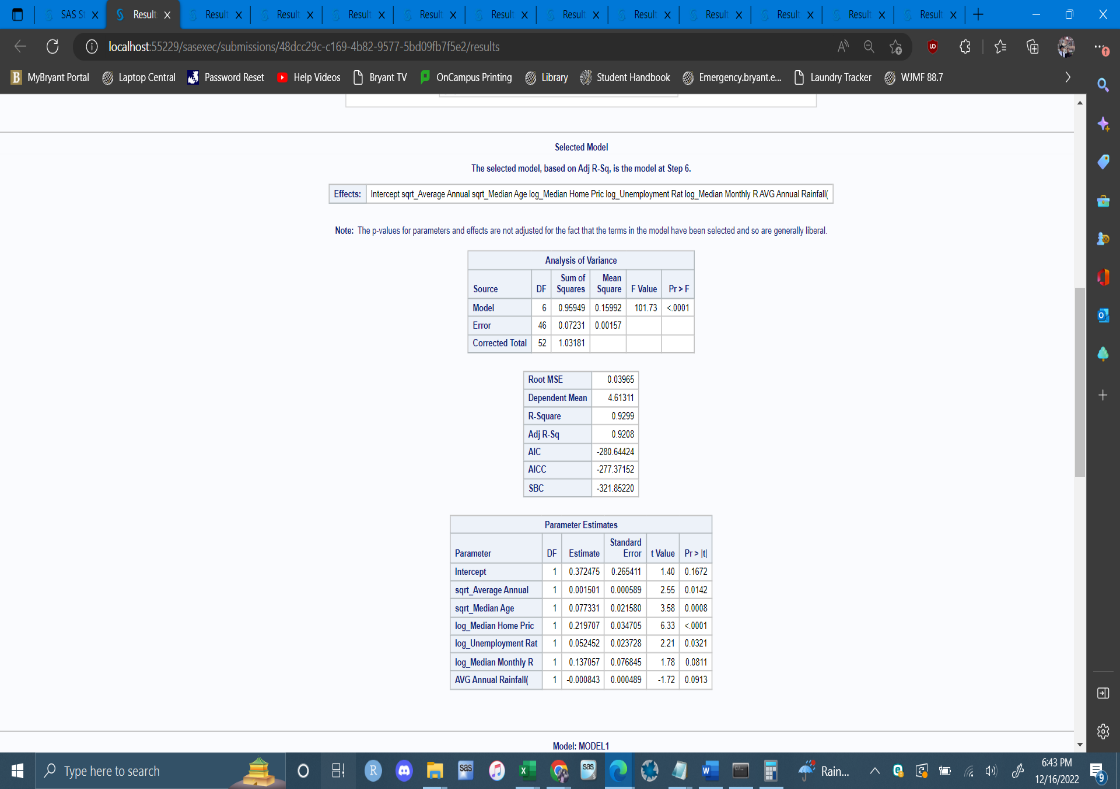
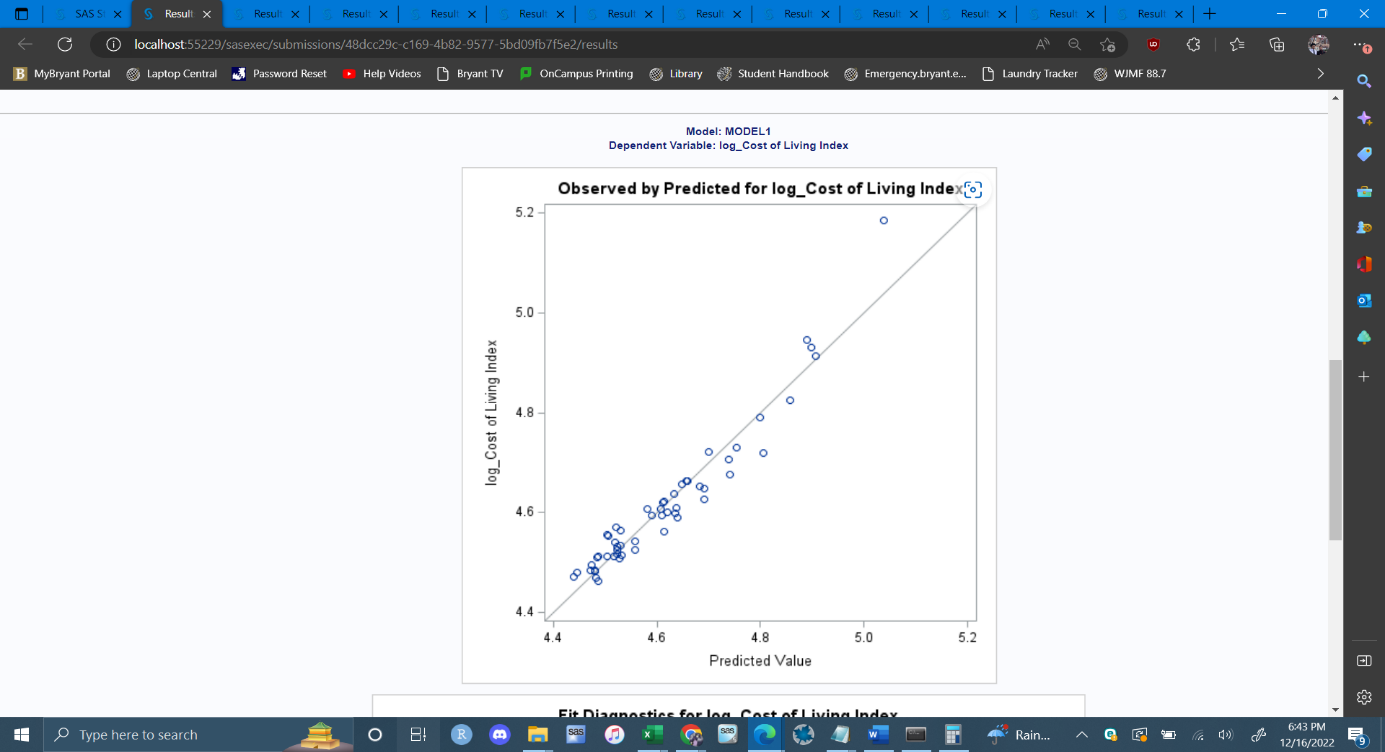
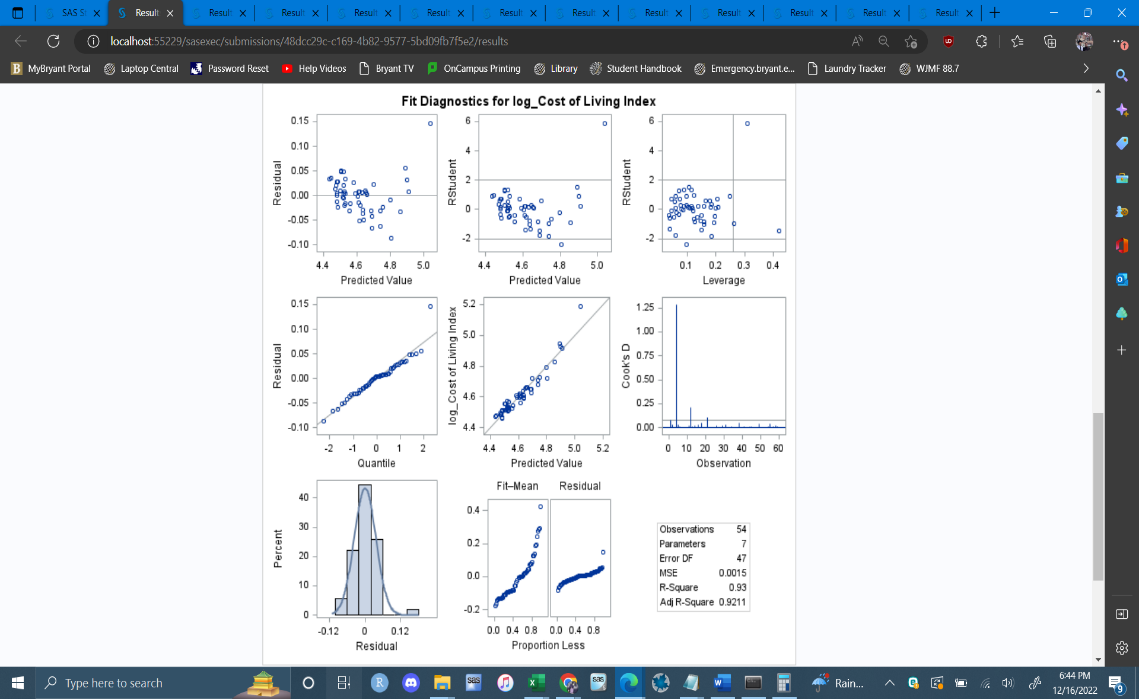
 

I decided to use stepwise selection for the regression. After running this with the dimension reduced data the selected model included all 4 variables to achieve the “best” fit.

**Cost of Living(index) = 4.68 - .0015(avg rainfall) + 0.017(factor 1) + 0.12(factor 2) - 0.028(factor 3)**

**One increase in Factor 1 would lead to a 0.017 increase in the cost of living index**

The normal probability plot looks good and there does not seem to be any influential observations in the residuals. The adjusted R-Squared came out to be .7989 which indicates a strong linear relationship. The MSE, which is how well the model fits the cost of living, is .00445. The F Statistic is below 0.05 so that proves the regression model is statistically significant.

After running the stepwise regression with the original partitioned from before I found that the results were slightly better. This dataset does not include any dimension reduction. The selected model stopped at 6 variables and left out violent crime, total students and teachers, commute time and property crime to achieve the “best fit.”

**Cost of Living(index) = 0.372 + .0015(avg rainfall) + 0.077(median age) + 0.22(median home price) - 0.052(unemployment rate) + 0.137(median monthly rent) - 0.0008(avg annual rainfall)**

The normal probability plot and residual plots look normal. The adjusted R squared is .921 which is higher than the first regression model. As for the MSE, it was 0.00157 which proves this model fits the cost of living better.

In conclusion, what I can interpret from the model is that the cost of living is strongly correlated with median home price, median monthly rent, median age. I also found that rent/home price is correlated with annual salary. I can infer that cost of living is also affected by other things which make it complex to calculate. One thing I would change is to gather more observations and factors. The addition of factors would help the factor analysis find a better underlying structure. Lastly if I continued this study, I would try region as a covariate to see how it plays into cost of living.