Understanding what controls education spending in Europe

Summary

It's often assumed that money spent on educational services in European cities depends primarily on some local revenue or country wide grants and the cost of other services or even demographics and city type does not affect its own cost. We attempt to identify which variables involved in revenue and expenditures most effectively predict the amount of money spent on education services in European cities using multi regression. We end up seeing that direct total expenditures and utility expenditures are the biggest predictors of how cities end up or will end up spending on education. The first is directly related to spending and encompasses all services while utilities and other government expenditures can make the government re-evaluate how much they will then spend on education. Demographics and social status take a back seat in Europe and do not play a very significant role in dictating money spent toward education

Introduction

Revenue is part of a government's budget and provides options on what a European government can spend on (Kutner). With this in mind most people would think revenue is the key indicator on telling how much the city governments can spend on education. Despite this, there may be other areas that might have a stronger influence on this area of spending. This brings up the question of "What attributes involving expenditures and revenue are strongest in determining how educational services are spent in the cities?" To analyze, our report has all the areas of local government spending and revenue as well as money held and owed along with CPI . For greater context we include the city type, id, and their population levels. Our central hypothesis will be that European cities' educational spending will react primarily on how much money they have or owe. Based on this we expect revenue variables, debt, and cash holdings to be the main predictors of educational spending while the other areas of spending and its demographic related variables are less significant. Our analysis will be done by using multiple linear regression (MLR) since this is well suited toward evaluating a quantitative or numerically continuous response variable that is predicted by both quantitative and categorical variables.

Exploratory Data Analysis

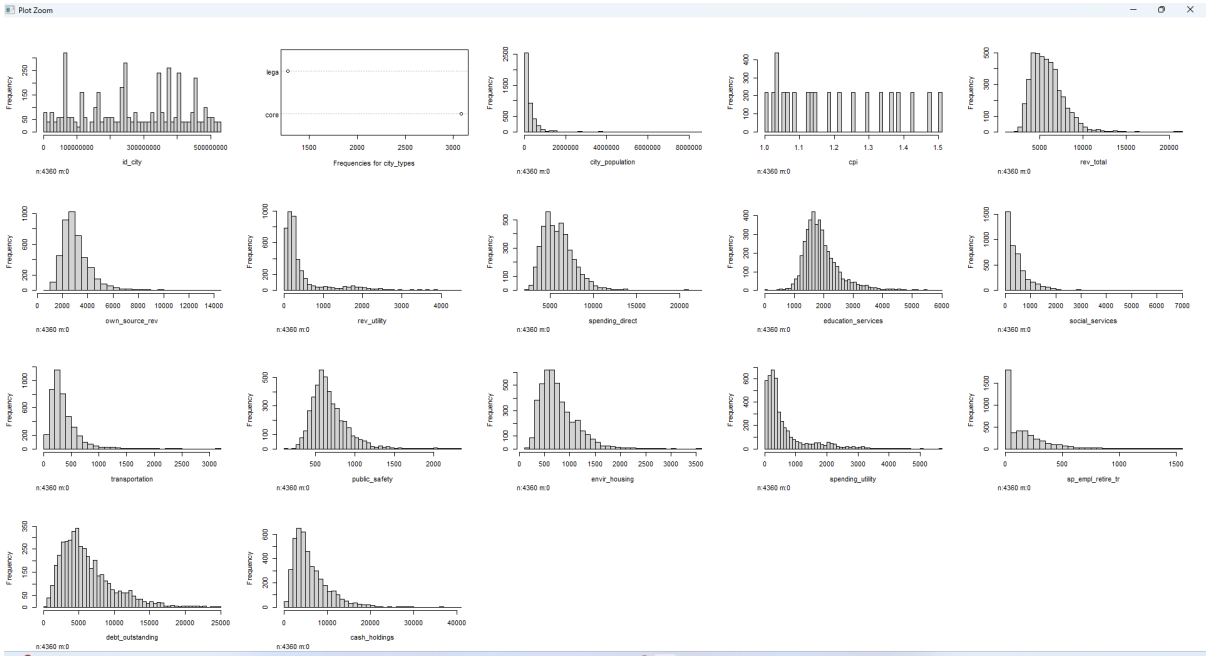
A. Summary Table and Histograms

Our data consists of 4360 observations and the majority of our cities are classified as core cities. In addition the data is already standardized given how all the values for expenditure and revenue are per capita dollars and there are no missing or NA values. For this dataset we do not use the variable years. The majority of our variables are partially right skewed although some of our variables such as total or city revenue are closer to a normal distribution along with target dependent variable education services that are only slightly skewed. We do see sp_empl and cpi with eye catching outliers as the highest frequency for cpi is around 1.03 with a frequency of 1857..

Summary Statistics

Variable	N	Mean	Median	SD	Min	Max	Pct.25	Pct.75
city_types	4360							
… core	3080	70.6%						
… lega	1280	29.4%						
city_population	4360	318068.8	172023	671344.408	15531	8475976	65695	334065.5
cpi	4360	1.207	1.162	0.158	1	1.504	1.064	1.34
rev_total	4360	5976.747	5671.39	2034.719	1389.3	21381.92	4542.212	6938.078
own_source_rev	4360	3107.034	2835.495	1318.151	978.66	14110.41	2315.102	3552.09
rev_utility	4360	497.318	230.395	699.519	0	4451.16	141.872	462.967
spending_direct	4360	6008.107	5704.815	2029.21	2291.44	22390.79	4603.58	6977.19
education_services	4360	1933.636	1810.03	649.378	58.35	5979.53	1525.963	2182.868
social_services	4360	538.017	339.64	732.28	0	6822.65	119.9	645.412
transportation	4360	360.384	296.375	259.424	23.35	3181.02	201.07	440.133
public_safety	4360	681.252	631.125	245.211	123.22	2315.8	529.335	783.35
envir_housing	4360	774.181	689.385	370.262	148.14	3521.94	515.622	951.69
spending_utility	4360	628.679	339.425	750.881	0	5614.71	182.735	720.87
sp_empl_retire_tr	4360	169.828	101.45	225.492	0	1531.81	0	244.158
debt_outstanding	4360	6123.774	5228.36	3722.01	471.93	24889.89	3460.045	7877.29
cash_holdings	4360	6173.956	4981.32	4244.739	270.23	40153.78	3306.27	7950.18

Figure 2 and 3 below: Histograms including the 3 transformed variables and ggpairs plot



B. Predictors/Response correlation, distributions, and transformations

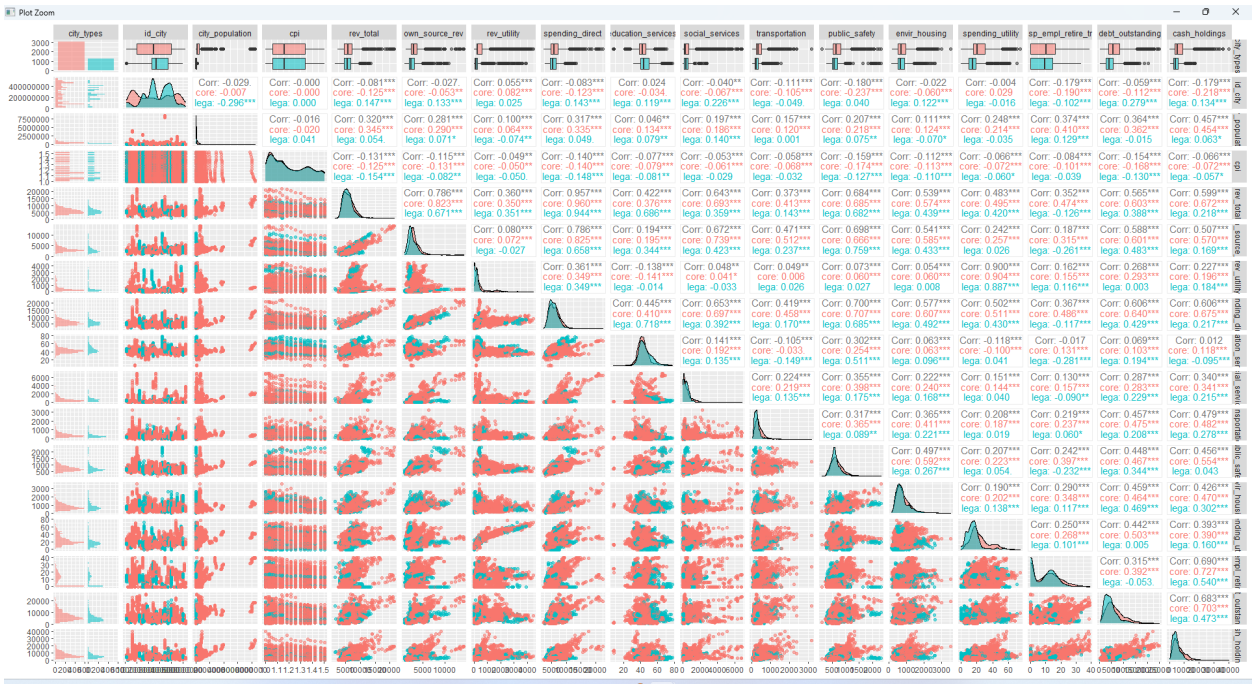
An important thing to note is there are a few variables with strong outliers most notably the zeros from *sp_empl_retire_tr* or even city_population which has a large number of very small cities. To prevent our data from being over dispersed or skewed relative to the normal distribution we will apply square root transformation to our variables *sp_empl_retire_tr, spending_utlity* and our dependent variable *education services*

Looking at the distribution of observations from our ggpairs plot that are divided between those who live in a core city (orange) versus those that live in a legacy city (teal), legacy city

observations are in the minority to begin with because many people have left these cities for core cities. Furthermore we can see how much legacy cities are distant from the centers of distributions for our variables almost often on the edges of the center; city population for example, has very high levels of education services when the city population is low which is a little surprising. In the case of CPI, there is a very polarized distribution for legacy city spendings toward educational services.

Education services seem to have the highest correlation with *spending_direct* and just trailing closely is *rev_total*. This is to say *rev_total* had much greater correlation compared to the *own_source_revenue* in regards to education_services. One important thing to note is *spending_direct* and *rev_total* along with *rev_utility* and *spending_utlity* had a correlation greater than 0.9 which would indicate multicollinearity among predictors. In addition, *revenue_total* and *own_source_revenue*, *spending_direct* and *own_source_rev*, *sp_empl_retire_tr* and *cash_holdings* had correlations greater than 0.7 between each other.

It should be noted that *CPI, spending utility, transportation, and rev_utility* had significant negative correlation with *education_services showing* there may be that these areas may be prioritized over education. The variable *cash_holdings* was not significant and had little correlation with *education_services* most likely because money held has little to do with the government or is involved in educational investment by consumers. Note, *sp_empl_retire_tr, id_city* and *city_population* had little correlation and significance with *education_services* so we won't be talking more about them in this section.

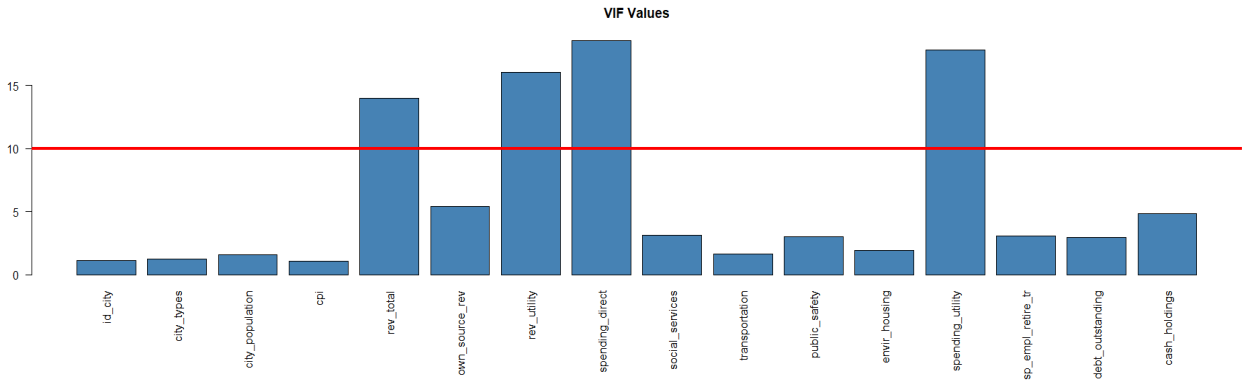


Statistical Analyses

A. Removing variables with multicollinearity and other variable pruning.

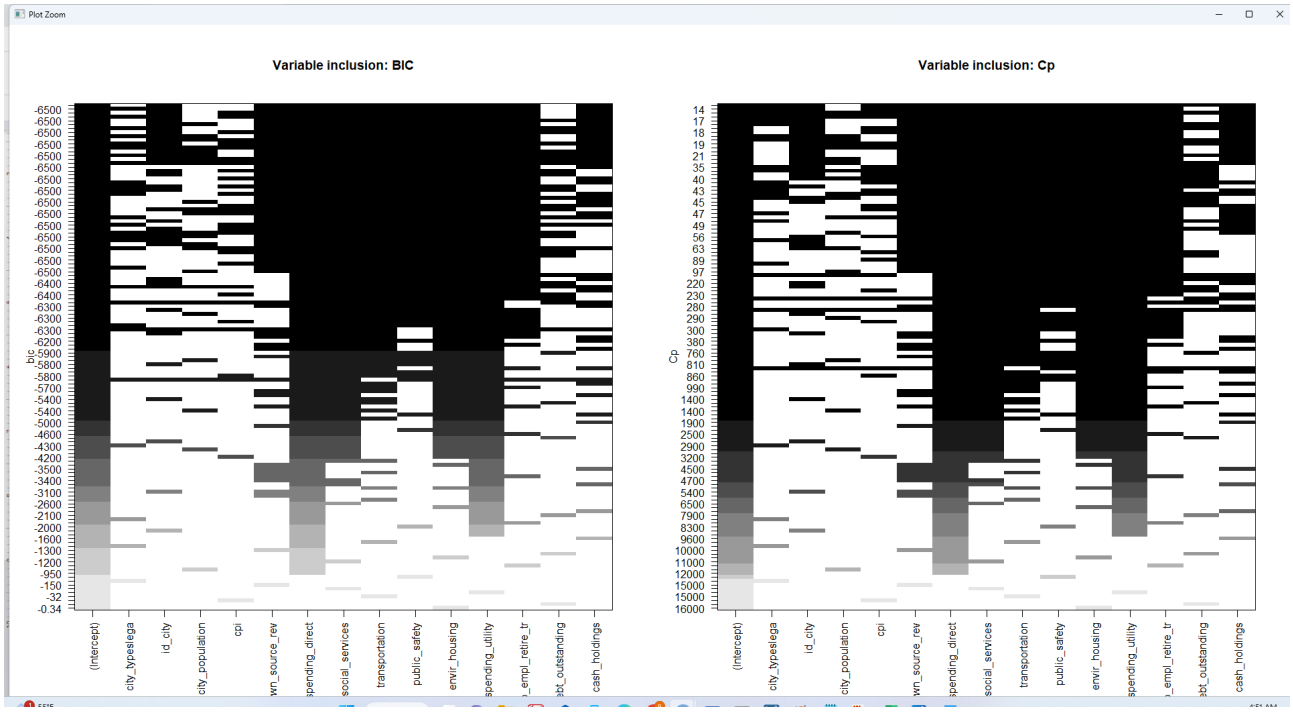
Earlier we saw the extremely high levels of correlation *spending_direct* and *rev_total* along with *rev_utility* and *spending_utlity* had. Strong multicollinearity increases variance which increases standard error that ends up overestimating our coefficients for our model. We can check the Variance inflation factor for these values where a high R squared value indicates a target independent variable being explained by the other target independent variable. As shown

the four variables have VIF > 10 which requires them to be addressed by dropping at least one or two of the variables.. When we remove rev_utility and rev_total we see multiple coefficients see a drop in their standard error; most noticeably spending_utility goes from .018 to .006. Meanwhile spending_direct now goes from .007 to .004. It makes sense that these two revenue related variables be removed especially since they can over explain other variables.



B. Variable selection

With a parameter space within reasonable space, we can perform forward and backward subset analysis. It is best that we use BIC since it is somewhat faster and selects simpler models compared to Mallow's Cp. Different models are compared in each row of the plots and black lines show variable inclusion in the models. The best models overall included *spending_direct, social_services, transportation, public_saftey, envir_housing, spending_utility, and sp_empl_retire_tr.* This is about the same for both metrics.



C. Diagnostic plots

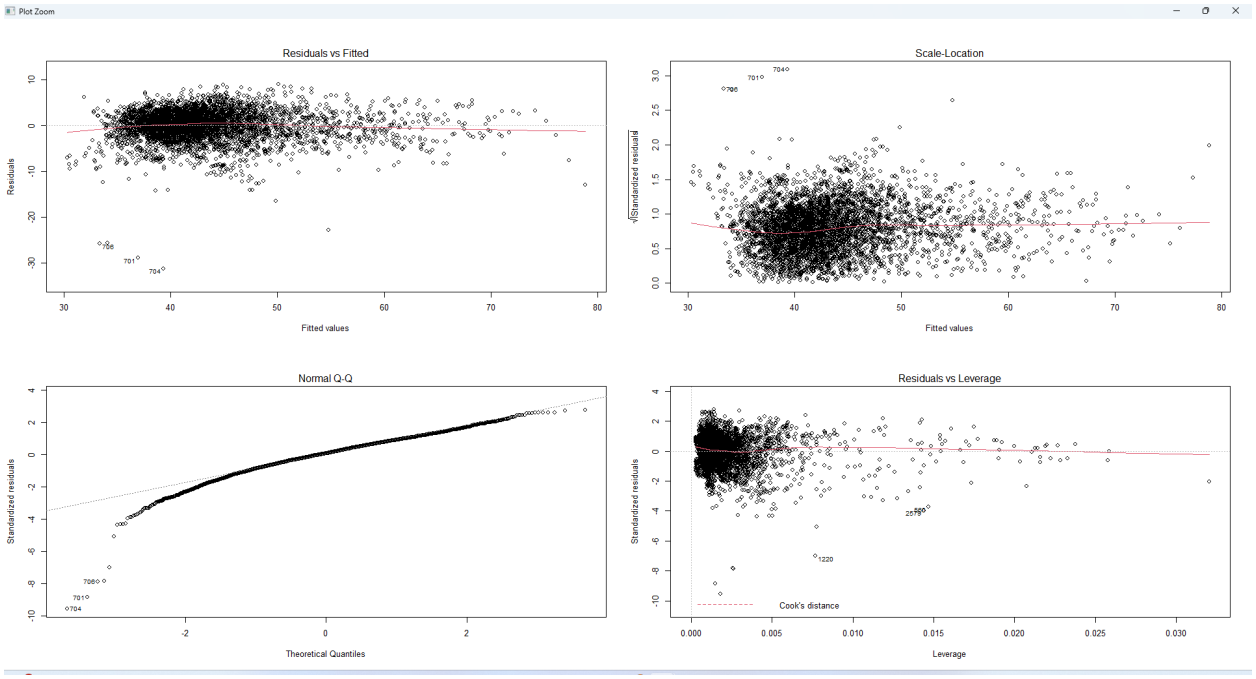
After selecting our best model, we can see that overall our model is significant since our p-value is close to zero while our R squared value is 0.7779. Our predictors are all significant with the p values being less than **α** = .05 level. Given how all of our predictors have significance with ***, we see that the ones with the largest standardized coefficients and model selection from earlier *spending_utlity and sp_empl_retire_tr* have the most importance as predictors followed by social_services and env_housing. This is also in part because they have the same standard units. *Public_saftey, own_source_reve, and transportation* are the lesser of the group.

```
Call:
lm(formula = education_services ~ own_source_rev + spending_direct +
    social_services + transportation + public_safety + envir_housing +
    spending_utility + sp_empl_retire_tr, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-31.1948  -1.7369   0.2667   2.1660   9.0493

Coefficients:
                   Estimate  Std. Error t value             Pr(>|t|)
(Intercept)       32.67078646  0.17809108  183.45 <0.0000000000000002 ***
own_source_rev    -0.00112507  0.00007872  -14.29 <0.0000000000000002 ***
spending_direct    0.00675832  0.00006192  109.15 <0.0000000000000002 ***
social_services   -0.00608460  0.00011322  -53.74 <0.0000000000000002 ***
transportation    -0.00730136  0.00022436  -32.54 <0.0000000000000002 ***
public_safety     -0.00595495  0.00033118  -17.98 <0.0000000000000002 ***
envir_housing     -0.00780001  0.00018086  -43.13 <0.0000000000000002 ***
spending_utility  -0.40390505  0.00508006  -79.51 <0.0000000000000002 ***
sp_empl_retire_tr -0.16339781  0.00636834  -25.66 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.271 on 4351 degrees of freedom
Multiple R-squared:  0.7779,    Adjusted R-squared:  0.7775
F-statistic:  1905 on 8 and 4351 DF,  p-value: < 0.00000000000000022
```



For our diagnostic plots shown above we can safely say that our regression model follows assumptions without any serious violations. The trend line fits fairly well for most of the residuals vs fitted and the scale location plot suggests homoscedasticity. The QQ-plot is good as most of it falls flat on the line showing linearity and while it does have a tail end most likely

due to the 0 values from *sp_empl_retire_tr,* we do not have huge leverage overall so most of the slope is not summarized by a few data points

D. Description of model and inference

In this model we have X_1 as own_source_rev, X_2 as spending_direct X_3 as social services, X_4 as transportation, X_5 as public_safety, X_6 as envir_housing, X_7 as square rooted spending_utlity, X_8 as square rooted sp_empl_retire_tr, and Y is square rooted education_services. Final model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_8 X_8 + \epsilon$$

For H_0 being all the coefficients of β_i being 0.
For H_A we have at least one of them not being equal to 0

Since we have an R squared value not equal to zero and our model is significant as proven from the p value earlier, we can reject the null hypothesis with at least one coefficient not being zero . Our final regression line is this:
Y=$32.67 - .00112507 X_1$ + .00675832 X_2 - .006084 X_3 - 0.00730136 X_4 - 0.00595495 X_5 - 0.00780001 X_6 - 0.40390505 X_7 -0.16339781 X_8

We will now perform some confidence interval tests with $\alpha = .05$ and if the interval for βi contains 0 in it we can know that there is no particular evidence of a linear relationship between the coefficient and our dependent variable education_services. Not surprisingly all our variables had intervals that did not contain 0 especially given how they all had *** for their p values earlier so they are all statistically significant and we reject the null hypothesis that any of the coefficients were 0 and that we are 95% our population mean is.

It should be remembered that we transformed the values for education_services, spending_utility, and sp_empl_retire_tr by square rooting so it is difficult to interpret. For example, for each additional per capita dollar that was spent on utilities there is a decrease between .41 and .39 per capita dollars of education services. In the case of our positive coefficient for spending direct, for every dollar spent on overall expenditures, there was an between .0066 and .0068 dollar spent toward education services

```
> confint(fmqp)
                      2.5 %        97.5 %
(Intercept)       32.321637234 33.0199356768
own_source_rev    -0.001279406 -0.0009707249
spending_direct    0.006636937  0.0068797078
social_services   -0.006306573 -0.0058626221
transportation    -0.007741217 -0.0068615005
public_safety     -0.006604238 -0.0053056625
envir_housing     -0.008154577 -0.0074454353
spending_utility  -0.413864554 -0.3939455507
sp_empl_retire_tr -0.175883010 -0.1509126163
>
```

Conclusions/Discussion

Looking at our results what I did find predictable was spending_direct having a positive correlation and prominence as a predictor of education_services. It is only natural afterall to see that when the government records its overall spendings, education services will naturally be part of it as a government of the city can allocate a certain percentage of money toward the schools or other educational facilities every year, month, day, etc. So when one shows that money was spent, the other will show it as well in a way that can be compared quantitatively.

We should point out that sp_empl_retire_tr is not exactly a government service or project but it is an expense. An expense being trust funds could involve not just government employees in sectors such as transportation but many other common people as well. It is after all negatively correlated so these funds aren't necessarily going to school institutions or services. We had many strong concentrations of near or exactly zero showing some lack of money that the government may have although there is no strong variable correlation between this and our other predictors.

When we look at services we can understand that spending_utilities is more important of a negative predictor than the other government expenditure services Utilities is a necessary expense (eg, electricity, gas) that people require on a daily basis, and it's even more critical than something like transportation or social_services which people can live without. We come to understand that education services can see cuts in their spending when these other public services increase or require attention toward spending.

Debt_outstanding and cash_holdings were two variables I expected to see in the model but didn't appear. In a situation for the government they often hold cash because it can help with transaction costs and help finance certain projects in case some other government projects are costing too much money. It could be the case that reserved cash is not applicable to school institutions or projects. In the case of debt_outstanding it may seem similar to a trust fund because the government almost owes some form of money but debt is current and is not factored toward allocating money spent in the future.

What I did find surprising was how own_source_revenue did not appear as a more prominent variable among the models. Even though it was significant it had less impact as a predictor than the other expenditure predictors. It may be clear that governments in Europe do not entertain spending extra money if they see they have greater levels of revenue. Certain metrics regarding the population and their consumption of goods also did not have enough sway to be well contributing variables in our model, maybe because city governments factor in other services first (utility,transportation, etc..) or they prioritize education regardless of the city type or population. Europe after all is not so similar to America where money is funneled to private schools and smaller wealthier areas and legacy or old American cities with low populations may no long prioritize education services

We did not include years in our variables because as I tested our model selection earlier, years did not really show up among the models that appeared that could have affected the R squared or residuals of our model. Additionally, years is not relevant to spending, as factors that truly influence money spent on education service was almost predictably the services spent by the government and the years in this dataset had repeating values and could not be inferred from that.

Limitations of this study amounted to the fact that we do not know the actual identification behind the city id's so we cannot directly draw connections to variables such as CPI . We also do not know the global situation each year regarding these cities such as war, bankruptcy, etc. In the future we could elucidate who these cities were and what policies they're governments set in place to influence areas of retirement funding as we had many outliers and were not sure why they were there and could not remove them.

References

pubs, R. “Multiple Linear Regression R Guide.” *RPubs*, 18 Apr. 2018, https://rpubs.com/bensonsyd/385183.

Kutner, Michael H., et al. *Applied Linear Regression Models*. McGraw-Hill Education/Asia, 2018.

Maciag, Mike. “What Are Cities Spending Big on? Increasingly, It's Debt.” *Governing*, Governing, 21 Apr. 2021, https://www.governing.com/archive/gov-legacy-cities-bills-debt.html.

files, eurostat. “File: General Government Expenditure by Function and Transaction, EU, 2020 , % of TE.” *Eurostat Statistics Explained*, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File%3AGeneral_government_expenditure_by_function_and_transaction%2C_EU%2C_2020_%2C_%25_of_TE.png.

Appendix

```
#Libraries and imported data that was saved as an excel workbook
#Make sure to to have setwd() set to where your excel workbook is
library(readxl)
library(modelsummary)
library(dplyr)
library(vtable)
library(tidyverse)
library(data.table)
library(GGally)
library(ggplot2)
library(corrplot)
library(Hmisc)
library(car)
library(flexmix)
library(leaps)
library(olsrr)
library(MASS)
library(lmtest)
options(scipen = 999)
df = read_excel("dae_23.xlsx")

#Summary statistics
sumtable(df[,3:18], summ =
c('notNA(x)','mean(x)','median(x)','sd(x)','min(x)','max(x)','pctile(x)[25]', 'pctile(x)[75]'),
summ.names =
        c('N','Mean','Median','SD','Min','Max',"Pct.25","Pct.75"))


#Histograms
hist.data.frame(df[,2:18])

#Transformed variables
df[,10] <- sqrt(df[,10])
df[,15:16] <- sqrt(df[,15:16]


#GGpairs plot
ggpairs(df[,2:18],aes(colour = city_types , alpha = 0.2))


#VIF model
par(mfrow = c(2, 1))
model <- lm(education_services ~id_city + city_types+ city_population + cpi+ rev_total +
own_source_rev +
          rev_utility+ spending_direct +social_services + transportation+
public_safety+envir_housing+  spending_utility
        + sp_empl_retire_tr + debt_outstanding + cash_holdings , data=df )
vif_values <- vif(model)
barplot(vif_values, main = "VIF Values",  col = "steelblue",las=2)
```

```
vif(model)
abline( h=10, col="red",lwd=4)
```

#Removing variables/columns from current data set renaming as df1 and checking column names

```
df1 <- subset (df, select = -c(rev_total,rev_utility,year))
colnames(df)
```

#Subset algorithm for BIC and Cp

```
b_subset <- regsubsets(education_services~., data = df1, nbest=10, nvmax = 16,really.big = T)

par(mfrow = c(1, 2))
plot(b_subset, scale = "bic", main = "Variable inclusion: BIC")
plot(b_subset, scale = "Cp", main = "Variable inclusion: Cp")
par(mfrow = c(1, 1))
```

#Summary of best model from earlier criterion

```
fmqp <- lm(education_services ~ own_source_rev + spending_direct + social_services
        + transportation + public_safety + envir_housing + spending_utility
                + sp_empl_retire_tr, data=df1)
summary(fmqp)
```

#Diagnostic plots

```
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fmqp)
```

#95% Confidence interval

```
confint(fmqp)
```