INTL 450: Advanced Data Analysis in Python

Dora Çelik 64324 - Homework #2

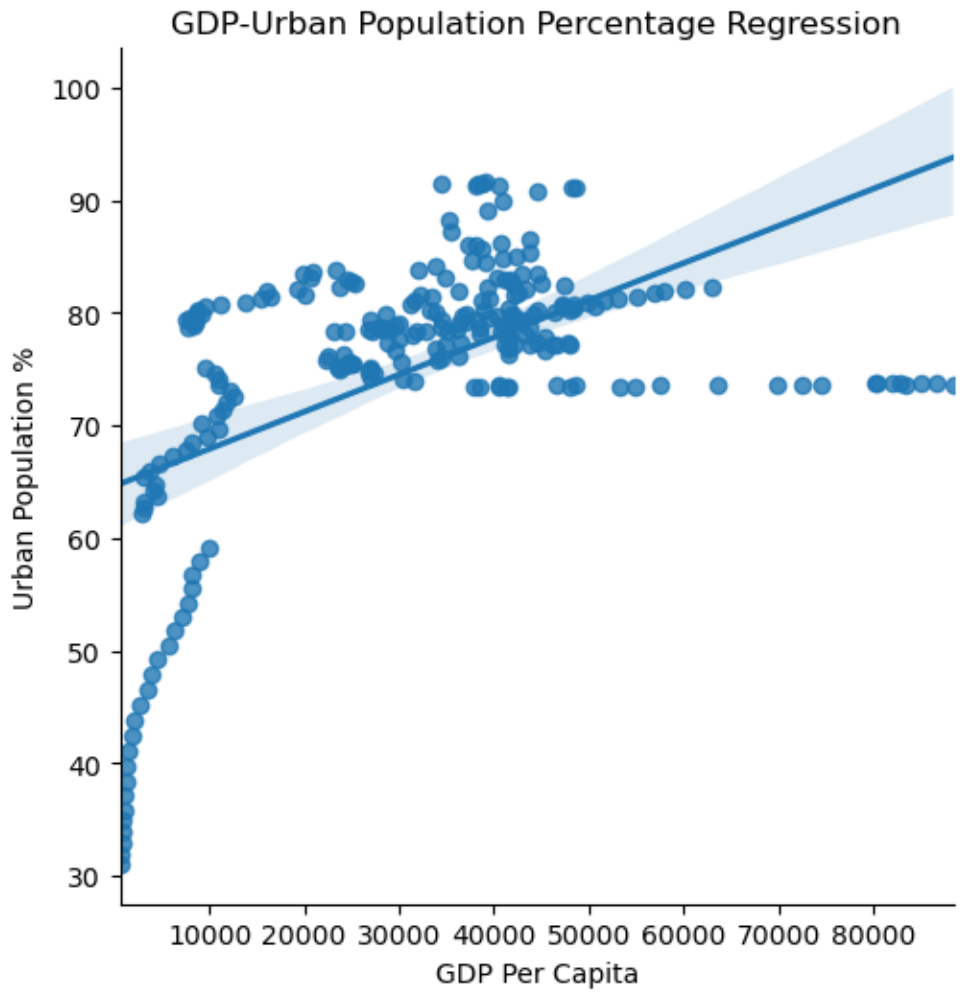
Introduction

Urbanization is a metric which is getting more important every day for measurement of a country's development. Therefore, this homework is aiming to observe the relationship between a country's urbanization percentage in its population and its gross domestic product (GDP) in USD$ terms. Data will be obtained from World Bank's API by using Python; linear regression model will be used as a quantitative method to examine the relationship between these two metrics.

Analysis

Here are the results of my regression and its plot.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.53155597							
R Square	0.28255174							
Adjusted R Square	0.27953726							
Standard Error	16130.3016							
Observations	240							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	2.4388E+10	2.4388E+10	93.7312406	6.6412E-19			
Residual	238	6.1924E+10	260186631					
Total	239	8.6312E+10						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-32569.378	6724.71522	-4.8432353	2.3017E-06	-45816.943	-19321.813	-45816.943	-19321.813
Urban population (%	856.650024	88.4832873	9.68148959	6.6412E-19	682.339583	1030.96047	682.339583	1030.96047



As it can be seen in the plot, two countries grab attention: China's GDP increase and increase in its population clusters away from other countries. Switzerland's stable urban population percentage and GDP growth also clusters away from other countries in the plot.

In our regression model, we end up with an intercept of -32.569 and a coefficient of 856.65 for our independent variable (urban population percentage). Therefore, we end up with an equation of our GDP predictor as **GDP = -32.569 + 856.65*(Urban Population %).** This equation indicates that with a 1% increase in urban population, GDP of a country will also increase approximately $856.

P-values for our variable is almost 0, indicating that it is statistically significant. If we would be conducting a multivariate regression model with multiple variables, which I intended to at the first place, we would have to remove variables which have scored higher than 0.05 P-value result in order for our model to be statistically more accurate and significant.