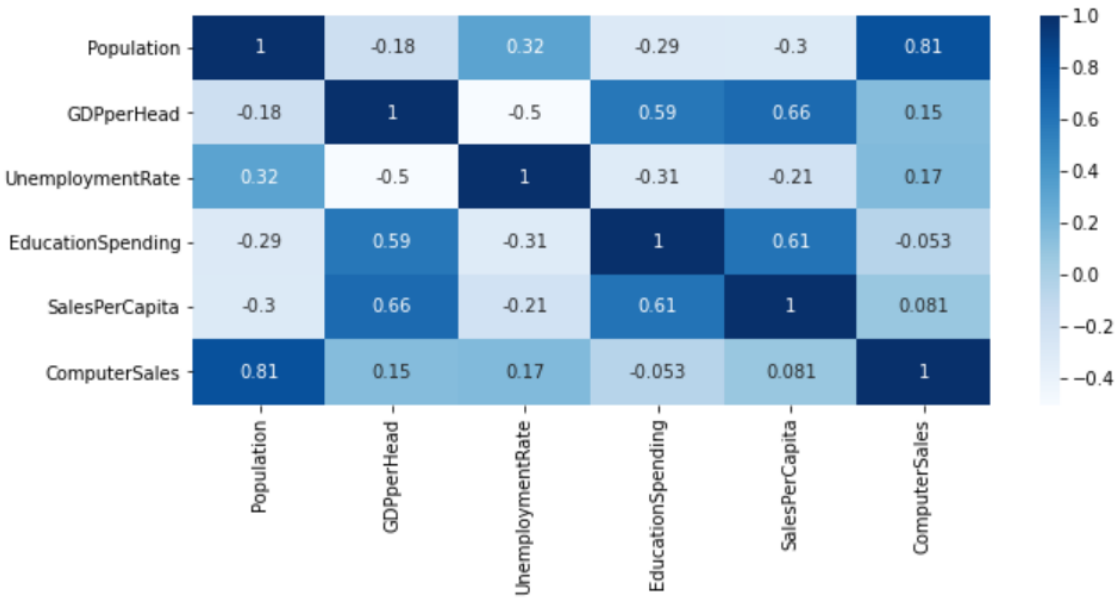
LINEAR REGRESSION WITH EUROPEAN SALES DATASET

Analysis starts by plotting a heatmap to see correlation values of different variables, later linear regression assumptions are checked one by one, train test split is applied, model is built and various metrics like r-squared and percentage error are measured.



Above heatmap shows the correlation values between variables. SalesPerCapita variable seems to be highly correlated with GDPperHead and EducationSpending, on the other hand ComputerSales has its highest correlation with the Population variable.

When distributions are checked, it is possible to see that all four independent variables are normally distributed and therefore confirms the normality assumption. Later, variance inflation factor test is applied, EducationSpending has a high vif value which causes multicollinearity, thereof it must be dropped. Finally, scatter plots are used to clarify if different independent variables have a linear relationship with two dependent variables SalesPerCapita and ComputerSales. Even though it is not a perfect, linearity can be confirmed and tested.

Data is splitted into train and test datasets and trained using the linear regression model. After testing with various variable combinations, GDPperHead seems to be assisting for both models and Population seems to be helpful for predicting ComputerSales with higher r-squared and lower error percentages.

SalesPerCapita

Metrics	Values
R2	0.649993
R2-Adjusted	0.461527
ExplainedVarianceScore	0.737929
MeanAbsoluteError	25.364035
MedianAbsoluteError	26.120741
RootMeanSquaredError	25.742022

TrueValues	Predictions	Error(%)
73	97.922131	34.139906
75	47.680649	36.425802
162	131.388259	18.896136
160	141.397085	11.626822

ComputerSales

Metrics	Values
R2	0.482315
R2-Adjusted	0.203561
ExplainedVarianceScore	0.647333
MeanAbsoluteError	1537.023791
MedianAbsoluteError	438.597219
RootMeanSquaredError	2578.384616

TrueValues	Predictions	Error(%)
813	975.591542	19.998960
2847	2200.706887	22.700847
9887	4778.690818	51.666928
1682	1451.098674	13.727784

For both dependent variables whether r-squared, adjusted r-square or error(%) do not seem satifying. Mean and median absolute errors are not particularly terrible, but insufficient nonetheless.

On the other hand, root mean squared error is lower than standard deviation of the dataset which is considered as a good sign. ExplainedVarianceScore of ComputerSales does not qualify since it is below 0.70 but ExplainedVarianceScore of SalesPerCapita is 0.73 which is satisfying.

In conclusion; even though some metrics seem adequate, most metrics imply that both models should be improved. It could be a lack of data that causes the poor metrics or lack of specific variables which would violate the assumption of endogeneity in linear regression.