**Cost of living: A strong indicator of economic well-being with respect to GDP?**

Final Project Report

Submitted By: Porur E. Kumanan

Student ID: **500993016**

Date: April 6, 2020

Development of nations is measured by economic factors such as income per capita, GDP per capita, life expectancy, literacy rates and access to necessary healthcare. However, for our project we are considering whether a nation’s GDP shows strength of correlation between cost of living and the economic well being of a nation. Cost of living is influenced by changes in cost of basic amenities like fuel, food, clothing, housing and most importantly; healthcare. The factors stated above support that standard of living which is closely linked to cost of living is a strong factor to consider when differentiating between a developing vs. developed nation.

Thus far, we haven’t been able to identify whether cost of living shows positive correlation to a nation’s GDP and whether it is also a strong indicator of a nation’s economic health. For the current study we are focused on utilizing predictive analytical models and pattern mining to identify the key indicators of a nation’s economic health based on factors of standard of living.

Juho Kiuru et al. employed predictive modeling techniques to measure economic growth based on human capital such as level of education and employment skills. Key findings of this study states that there is a strong correlation between intra metropolitan clustering of innovations and availability of a highly skilled workforce.1

Manuel Wolff et al. studied the change in urban residential density in Europe since 1990. Their studies reveal that there is a strong correlation between increasing per capita land consumption, household income growth and national GDP.2

Rafaela Costa Martins de Mello Dourado et al. studied the cost of living in the best livable cities in the world. They used cluster analysis and logistic regression to predict the best city to move to in order to decrease cost of living. The results show a strong association between livable index and cost of living.3

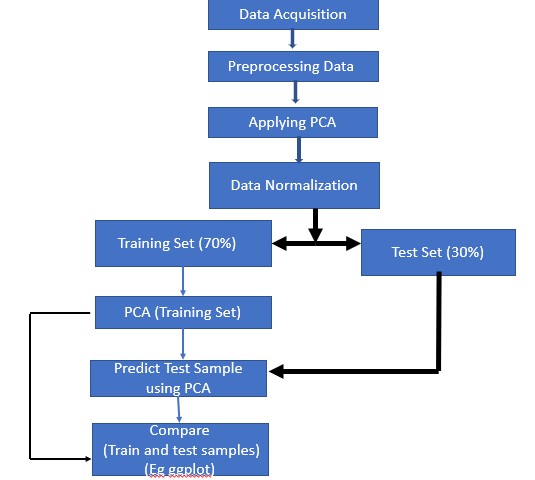
L.F. Masson et al. studies the Scottish diet in correlation with their household tobacco use. Results show that smokers have a poorer diet resulting in poor health thereby significantly reducing standard of living.4

Mingchen Feng et al. forecasted crime data in major American cities using pattern mining and trend prediction to identify the relationship between crime levels and increasing standard of living. They also utilized exploratory data analysis to aid with visualization and trend prediction.5

**Description of the Dataset**

GDP is the total value of all goods and services produced in a country and provides insight into a country’s economic wellbeing. The dataset used contains the cost of essential goods i.e. cost of meals, milk, water, fuel, vegetables, cost of public transportation, vehicle maintenance and other consumables etc. in various major cities around the globe. The data was gathered from the various countries and permission has been provided to freely use the data for analysis (<https://www.kaggle.com/stephenofarrell/cost-of-living>). In order to effectively conduct our analysis and arrive at a comprehensive result, GDP must also be considered for comparison purposes. GDP data was obtained from the World Bank and contains the GDP of all world nations listed in US$ for the 2018 fiscal year (<https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?view=map>).

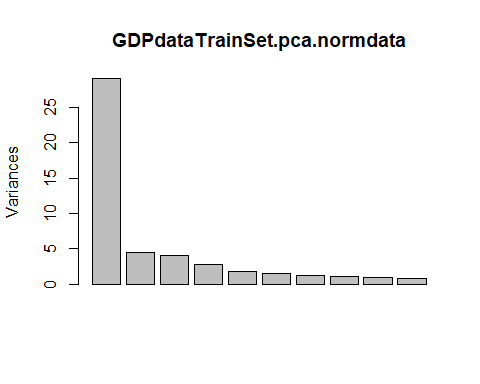
**Approach**



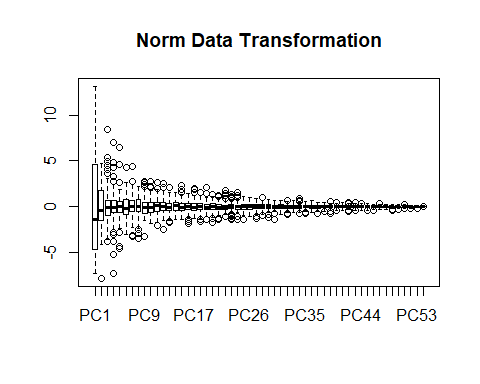
The following steps have been taken in order to accurately employ the data to conduct effective analysis:

* **Data Processing** – Removal of missing values and addition of GDP data along with the obtained cost of living data.
* **Principle Component Analysis**:
  + Dividing the data into a test (30%) and training set (70%) to conduct dimensional reduction due to the data having over 160 attributes.
  + Identifying principle components with an Eigen value greater than 1. Considering variance close to 90% in order to reduce dimensionality from 160 to lower values.
* **Data Normalization**- Applying normalization for scaling and centering the data.
* **Predictive Modeling-** Using test set (30%) for comparing training set and test set results using GG plot to identify any deviations is results.

**Final Results**

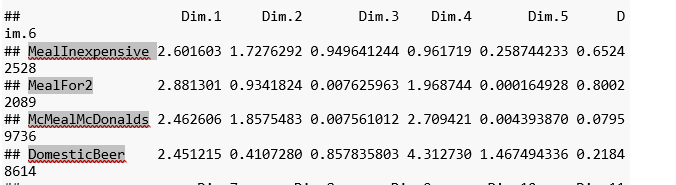


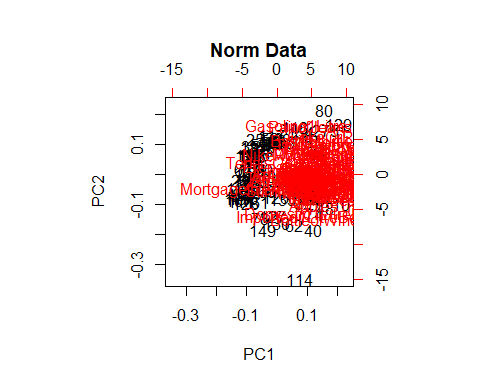
As stated previously, we have employed principal component analysis on 58 variables from GDP Data to identify individual variables that influence a nation’s GDP. Based on our principal component analysis, we obtained 55 principal components and 19 of those components account for variance above 95% of total variance in the data. While the chart above displays the first ten principal components of the training set, the first principal component displays the most variance.

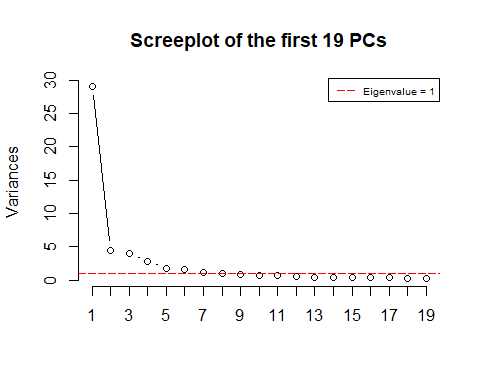


The norm data transformation plot shown above is a biplot between the first two principal components due to their increased dominance in variance. Visualizing the first two components of our principal component analysis, we have identified the following variables contribute to the PCA 1 and PCA: (Kindly refer to dim 1 and dim 2 and the table below). These attributes are not readily visible in our comparison graph between PC 1 and PC 2 listed below due to the number of variables listed in the acquired dataset.

* Meal Inexpensive
* Meal for 2
* McDonalds Meal
* Domestic Beer

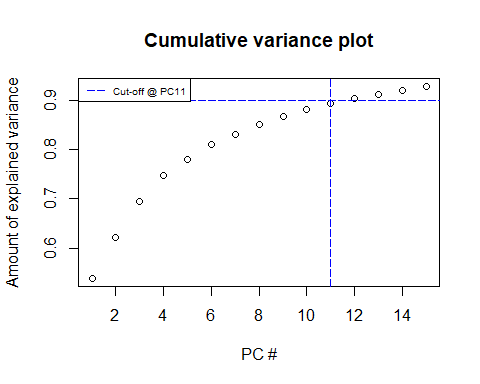


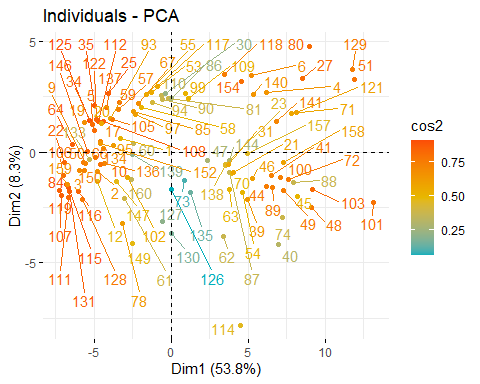




Based on the chart above, scree plot was utilized on the first 19 principal components to identify the Eigen value of the principal components. The first 11 principal components with an Eigen value greater than or equal to 1 explains almost 90% of the variance, thereby; effectively reducing dimensionality from 160 to 11 should we consider 90% of the stated variance.

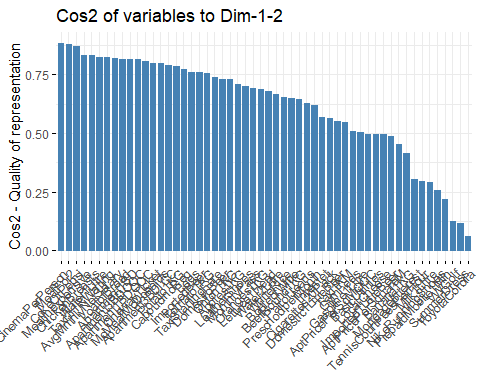
The cumulative variance plot listed below further confirms that 11 principal contributes 90% of variance within our principal component analysis.



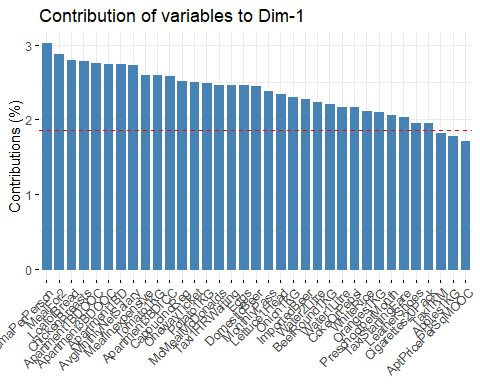


The above chart clearly states that the distance between variables and the origin measures the quality of the variables on the factor map. The quality of representation of the variables listed on the factor map is identified as cos2.

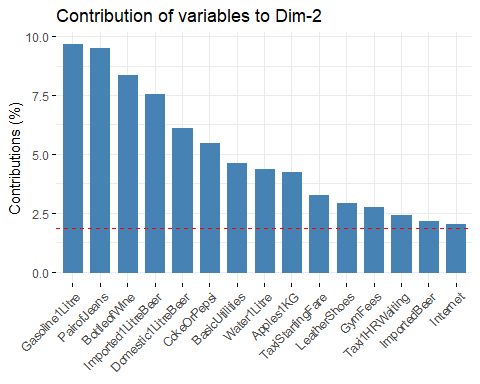
**Total Cos2 of variables on Dimensions 1-2**



The above chart displays the quality of representation of the variables of the principal component. A high cos2 value portrays a good representation of the variables in principal components which shows a farthest value in the circle of correlations (refer to Individual PCA Cos2 Plot for further confirmation). A low cos2 value indicates that the variables are not perfectly represented by the principal components. In this case, the variable is close to the center of the circle of correlation.

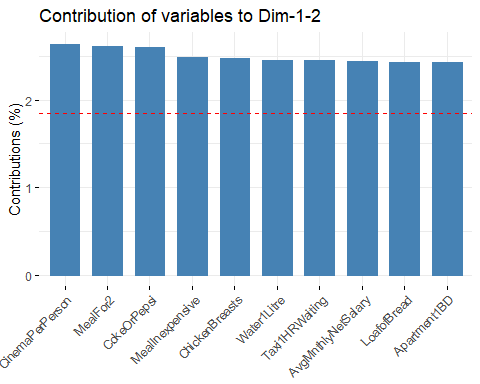


The above chart portrays the most useful variables that contribute to PC1. The red dotted line in the figure indicates the expected average contribution. The variables whose values are greater than the threshold are important contributors. The next step is in identifying the contribution of variables to PC2 displayed below:

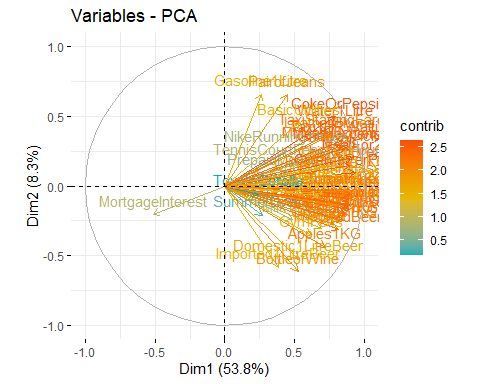


Comparing PC1 and PC2 states that Coke or Pepsi, Imported Beer, Water (1L), Leather Shoes, Taxi Sitting Fare, Taxi (1 HR Waiting), Oranges (1KG), Apples (1KG) contribute the most to PC1 and PC2.

**Total Contribution to PC1 and PC2:**



The above chart shows the contribution of variables towards PC1 and PC2.



The above graph highlights the most important variables that contribute to the Principal Component Analysis.

**Regression Analysis:**

Regression analysis helps to show the relationship between a response variable with respect to the predicted variables to build a model for estimating the GDP of a country based on the monetary value of all the goods and services produced in that nation.

In the present study, actual GDP data and principal components are used to analyze the strength of correlation between GDP and Cost of Living. Further analysis of F-statistics (p-value: 2.2e-16) shows that at least one predicted variable is significantly related to the outcome variable.

Analysis of t-values indicates that local cheese, domestic 1L beer, imported 1L beer, one-way ticket, chicken breast, monthly pass, prepaid mobile phone, average monthly net salary, lettuce (1 Head) and white rice are strongly associated to changes in GDP.

Regression Analysis Using PCA:

Standard linear models are potentially more powerful when employing PCA rather than using predicted variables alone. This analysis proves that only the first principal component turned out to be significant

**Model Accuracy Assessment**

In multiple linear regression, R2 represents the correlation coefficient between the observed variables of the outcome variables (y) and the fitted (i.e., predicted) values of y. R2 value close to 1 indicates that the model explains a large portion of variance in the outcome variable but R2 will always increase when additional variables are incorporated into the model. The primary solution is to adjust R2 by taking into account, the number of predictor variables.

By comparing multiple regression analysis using actual GDP data, all the principal components and the first principal component indicates that PCA shows the maximum R2 value (0.9241) than using only the first PCA (0.265) and predictor variables (GDP Data- 0.9004). Our study suggests that the first 9 principal components are sufficient for analysis. However, PCA is not a feature selection method because each of the calculated principal components is a linear combination of the original variables.

**Conclusion**

Based on our PCA analysis, all the variables in PCA 1 are negative, hence; they all contribute to the GDP. PCA 2 includes domestic beer, imported beer, Milk (1L), Loaf of Bread, Water (2L), Bottle of Wine, domestic beer (1L), imported beer (1L) and 20 Pack of Cigarettes all portray positive values, whereas inexpensive meal, Meal for 2 , McDonalds Meal, Coke or Pepsi, Water (1L), Eggs, Local Cheese, One Way ticket and chicken breasts portray negative values. In the case of PCA 3 variables such as Meal for 2, McDonalds Meal, domestic beer, imported beer, Milk (1L), eggs, local cheese, bottle of wine and domestic beer (1L) display negative values whereas inexpensive meal, Coke or Pepsi, Water (1L), Water (2L), 20 Pack of Cigarettes, One Way ticket and chicken breasts have positive values thereby positively influencing GDP.

However, considering principal component analysis, negative values are more important than positive values; i.e. GDP is significantly influenced by cost of living if only negative valued variables are considered. Similar results have been reported when conducting multiple regression analysis with PCA. However, multiple regression using actual data shows contribution from different variables due to PCA not being a feature selection method as each of the calculated principal components are a linear combination of the original variables. In conclusion, GDP shows strong positive correlation when only the variables of the first principal component are analyzed.

**References**

1. Juho Kiuru and Tommi Inkinen, ‘Predicting innovative growth and demand with proximate human capital: A case study of the Helsinki metropolitan area’, *Cities*, 64 9–17, (2017).
2. Manuel Wolff, Dagmar Haase and Annegret Haase, ‘Compact or spread? A quantitative spatial model of urban areas in Europe since 1990’, *PLOS ONE*, https://doi.org/10.1371/journal.pone.0192326 February 28, (2018).
3. Rafaela Costa Martins de Mello Dourado and Alessandra de Ávila Montini, ‘The cost of living in the best livable cities in the world: a brief predictive quantitative analysis’, *Int. J. Multivariate Data Analysis,* Vol. 1, No. 1, (2016).
4. L.F. Masson, K.L. Barton and W.L. Wrieden, ‘The Scottish diet is poorer in households purchasing tobacco products: analysis of Living Costs and Food Survey data from 2001–2012’, *Proceedings of the Nutrition Society*, 75 (OCE3), E145, (2016).
5. Mingchen Feng, Jiangbin Zheng, Jinchang Ren, Amir Hussain ,Yue Xi and Qiaoyuan Liu, ‘Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data’, *IEEE Access*, Volume 7, pp. 106111-106123, (2019).
6. Cost of Living Data- <https://www.kaggle.com/stephenofarrell/cost-of-living>
7. World Bank GDP Data- <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?view=map>