Layman Description

The data shown consists of countries by region and various population statistics. The aim of the analysis is to determine which factors are significant in influencing life expectancy in 2018.

We started with a model that consists of only quantitative factors and removed factors that were deemed insignificant at predicting our outcome - life expectancy 2018. We found that broadband subscribers, child mortality rate, income per person, and urban population are amongst the factors that heavily influence the life expectancy. We then consider qualitative factors such as the geographical breakdown of data, namely classified as *four_regions* and *world_bank_region* to determine if they are significant enough to be incorporated into our model. As a result, we concluded that only *world_bank_region* was crucial enough to be retained in our model.

After taking measures to account for the potential overlapping association between variables, we were left with a model that is driven by broadband subscribers, child mortality, income per person, urban population and World Bank region classification to predict life expectancy in 2018.

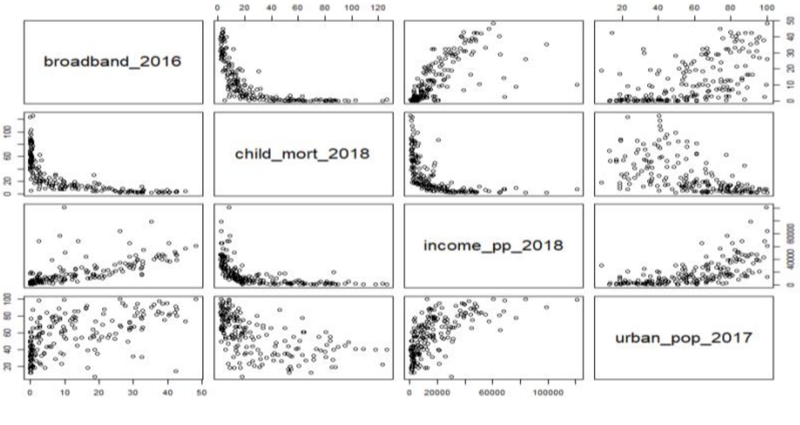
Statistician Description

This report aims to study the drivers of *lifeexp2018* given a set of continuous and categorical covariates from countries across the world. We first cleaned the dataset by adding dummy variables to our categorical covariates (*four_regions* and *world_bank_region*) and removing covariates with insufficient data entries. For clarity, the categorical covariates are set out as *world_bank_region_i,* i ranging from 1 to 6, represents Europe & Central Asia, Middle East & North Africa, Sub-Saharan Africa, Latin America & Caribbean, East Asia & Pacific, and North America respectively. The categorical covariates *four_regions_i,* i ranging from 1 to 3, represents Europe, Africa and Americas respectively. We define sufficiency for columns of covariates where data is available for at least 75% of the total number of countries. Hence, the covariates removed were *agri_2017, oil_per_capita_2016, working_hours_2017,* and *primary_completion_2016.* At the same time, due to the lack of datasets available for North America (only 2 entries), *world_bank_region6* (North America) was merged with *world_bank_region_1* (Europe & Central Asia) as both were regarded as developed regions with similar average life expectancies.

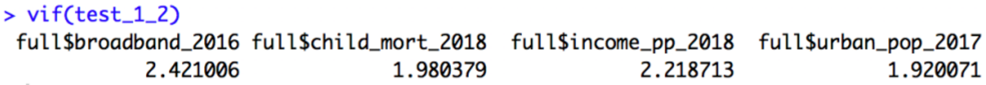
After making alterations to the data, we ran a linear regression model by undertaking an automatic model building approach through backward elimination. We determined that the four covariates which were "significant" (p-value of less than 0.05) in predicting the outcome (*lifeexp2018*) were *broadband_2016*, *child_mort_2018*, *income_pp_2018*, and *urban_pop_2017*. We then used an F-test to determine if we could remove the regression coefficients from our full model, and obtained a p-value of 0.250, implying that there is no evidence against using our nested model (null hypothesis).

We then consider the significance of categorical factors by conducting an F-test on our model without categorical covariates compared to the model that incorporates *world_bank_region*. With Asia and South Asia as our reference points, we obtained a p-value of 0.008, implying that there is sufficient evidence against our null of using the nested model which leads us to retain the *world_bank_regio*n as categorical covariates in our model. Similarly, we derived a p-value of 0.120 after conducting a F-test for *four_regions* and we concluded that there was insufficient evidence against removing *four_regions* as categorical covariates.

Following that, we conducted a test for multicollinearity by producing a matrix plot to assess if the presence of a linear relationship between the covariates *broadband_2016*, *child_mort_2018*, *income_pp_2018*, and *urban_pop_2017* existed, and obtained the following:



We observe that a linear pattern does not exist between the majority of covariates, besides potentially *broadband_2016* and *income_pp_2018*. In order to more accurately determine whether or not this collinearity exists between the two covariates, we then ran a variance inflation test (VIF) and obtained the following:



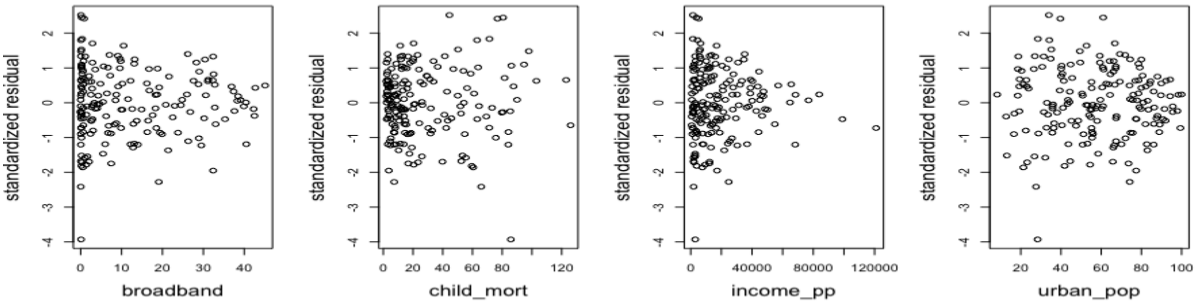
```
> vif(test_1_2)
 full$broadband_2016 full$child_mort_2018  full$income_pp_2018  full$urban_pop_2017
            2.421006             1.980379             2.218713             1.920071
```

As the variance inflation factor coefficients between the four covariates were all relatively low (less than 5), we conclude that there is no strong collinearity between the four variables.

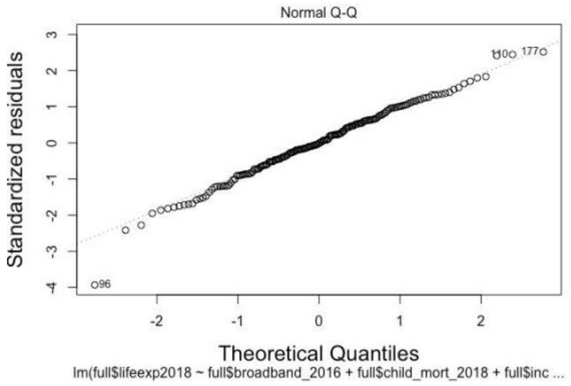
In order to check for dependence between covariates, we have incorporated all interaction terms of categorical variables (*world_bank_region*) numerical covariates into the model. Via ANOVA, we found out that the interaction terms between *world_bank_region* & *broadband_2016* and *world_bank_region* & *child_mort_2018* are significant. We have thus decided to incorporate them into the model.

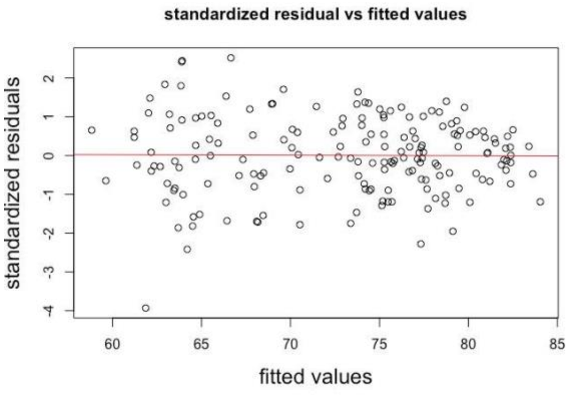
Using an ANOVA test, we tested between Model 1 (without interaction terms) and Model 2 (with interaction terms). We have obtained a p-value of 0.009. This implies that there is sufficient evidence to reject the null hypothesis (H_0: all coefficients of interaction terms = 0). Therefore, we decided to use Model 2 in favour of Model 1.

Finally, we conducted model checking tests to determine if our model exhibited the following required properties: i) linearity ii) homoscedasticity iii) normality. To check for linearity, a matrix plot was used to investigate the presence of a systematic pattern on the scatterplot of standardized residuals against covariates. As a discernible pattern cannot be observed, there is no evidence against linearity.



To check for normality, a QQ plot was used, which shows no evidence against homoscedasticity. To check for homoscedasticity, standardised residuals were plotted against the fitted values. The plots exhibited random scatter centered at zero, which roughly forms a horizontal band. As there are no systematic patterns observed, we conclude that there is no evidence against homoscedasticity.





In conclusion, after running a backward elimination test verified by an F-test and taking the collinearity and interaction between covariates into account, we found that the drivers of *lifeexp2018* are *world_bank_region, broadband_2016*, *child_mort_2018*, *income_pp_2018*, and *urban_pop_2017* (here x_i represents *world_bank_region_i*).

$$Lifeexp2018 = \beta_1 broadband\_2016 + \beta_2 child\_mort\_2018 + \beta_3 income\_pp\_2018 + \beta_4 urban\_pop\_2017 + \sum_{i=1}^{5} \beta_{i+4} x_i + \sum_{i=1}^{5} \beta_{i+9}(x_i)(broadband\_2016) + \sum_{i=1}^{5} \beta_{i+14}(x_i)\,(child\_mort\_2018) + \varepsilon$$