# Group4 - Ansora Ananth, Corey Parker Roberts, Lokesh Kumar Reddy Desireddy, Utkarsha Deepak Gupte

Project Report - Trends in Skill Migration

2020-11-25

## Executive Summary

Increased scientific research is conducted to understand the relation between migration and development across all countries for various reasons. The increased immigration flows into and out of the country represents an opportunity to better understand the dynamics of labor market and use of the country's government services and resources. With increase or decrease in immigration directly links to other sectors such as education, investment and financial status. The major economic effects of migration for both the sending and receiving countries vary depending specifically on worker skill levels. In this paper we analyze trends in metrics relating country migration, talent migration and development (evaluated using Human development Index) across different countries. We are utilizing the data obtained from the catalog of The World Bank site. This data set is part of the LinkedIn- World Bank Group partnership. This database contains relevant information of migration values with respect to countries and skill group category from 2015 through 2019. The HDI ranking of countries for the same time span is obtained from the site of The United Nations Development programme. We have analyzed the data using Multivariate linear modeling. While observing relation of migration with HDI and income, the results conclude that migration is more towards countries with higher income and is not related to HDI. Observing the skillset of migrants, there is less migration of workers in healthcare sector compared to other sectors.

## Introduction

Human development focuses on improving the quality of lives people lead and the development of the country rather than assuming the economic growth will automatically lead to advancement of human life. The Human Development Index was introduced by United Nations Development Programme (UNDP) for measuring the global ranking of countries with respect to their development. HDI is measured using factors such as life expectancy, education, and per capita income. Human development along with technological growth yield economy for a country, which in turn advances human development. The economic causes and consequences of migration are complex and multi-dimensional, affecting both origin and destination countries. There are 195 countries in the world today. These countries are differentiated by many factors, the main factor of difference being development. The Human Development Index (HDI) is used to assess the development of each country. The HDI is ranked on a scale from 0 to 1. The countries are divided into high developed countries (Score $>$ .8), developing countries (.55 $>$ Score $>$ .8) and least developed countries (Score $<$ .55). The opportunities and challenges presented by the global economy requires the public and private sectors to join forces, share information, share resources, and work towards a common vision to make meaningful, positive and scalable impact. many people migrate from their hometown to other regions for survival or better life opportunities. Within this context, this study aims to explore whether human development or the factors (income, education and health) which determine human development play a role in international migration flows, and whether migrants find an opportunity to increase their human development levels many people migrate from their hometown to other regions for survival or better life opportunities. Within this context, this study aims to explore whether human development or the factors (income, education and health) which determine human development play a role in international migration flows, and whether migrants find

an opportunity to increase their human development levels People migrate to other region for higher standard of living and better opportunities. In the current context, we can evaluate better quality of life with factors life expectancy, education and per-capita income assessed using HDI. Patterns of international migration are usually observed from developing countries to developed countries. The impacts of migration are complex, it has both advantages and limitations for both sending/origin and receiving/destination countries.

- For Origin countries, migration of labor provides economic benefits found as remittances. Further, this decrease in unemployment rate and increases the per capita income. At the same time, these countries can also suffer from loss of educated individuals.
- For Destination countries, immigration provides low cost labor and address skill shortage but in the long run with decrease in demand for labor this can also lead to decrease in domestic wages and increased public welfare burden. Increase in immigrants can also increase productivity of the economy thereby increasing the GDP.

The major economic effects of migration for both the sending and receiving countries vary depending specifically on worker skill levels. The focus here is to determine the relation between development and skill migration and how the business cycle is impacted due to the migration flows. In this paper, we will analyze the growth and fall of skill migration based on skill categories, HDI and income across countries with 3 null hypotheses listed in the next session. Initially we will look at the distribution of data in skill categories and HDI. EDA is administered to further understand the variations of skill migrations with respect to region - income and region – categories. Finally we introduce multivariate linear analysis to investigate the dependency on skill migration with respect to HDI, skill category, income level, and year. This analysis can help in tracking demand of skilled worker population in respective regions and also predict income variations

## Business Relevance

The goal of our project is to identify trends in skill migration.

- Analyzing skill migration trend is central to identifying opportunities that are needed for effective skill development policies to promote a competitive labor force that in turn fosters private sector growth and job creation.
- Monitoring international flows of talent allows policy-makers to shape their talent attraction and retention programs
- As it is believed that skill migration is higher towards higher HDI countries, It enables governments across the globe to work towards achieving a higher human development index.

## Null Hypothesis

The objective of this report is to measure the trends of growth of industries, emerging skills, job displacement across various regions. The three main categories collaborated here are the Migration by Region, Migration by Talent/skill, HDI of country. This report describes relation between metrics country migration, talent migration and HDI for 100+ countries present over all continents on the basis of below Hypotheses.

- Skill migration increases to regions with higher HDI
- Migration of workers with tech skills is predominant over workers with other specialized skills
- High income countries have the highest skill migration across all the income levels included

## Data Description

We have acquired our dataset from below two sites:

- https://datacatalog.worldbank.org/dataset/talent-migration-linkedin-data

- Human Development Index (2015-2018):
  - http://hdr.undp.org/en/content/table-2-human-development-index-trends-1990%E2%80%932018
- Human Development Index (2019):
  - https://worldpopulationreview.com/country-rankings/hdi-by-country

Our interest of data comes from below datasets – * Skill migration and * Human development trends

Hence we are using the linked in matrix along with the dataset of Human data index (HDI) to find answers to the hypotheses we mentioned in the above section.

**1. Skill Migration:**

The original dataset (public_use-talent-migration.xlsx – skill Migration) consists of 17,618 rows and 12 columns. This dataset is intended to communicate below aspects of the country: * If it is low income/high income/Lower middle income/Upper middle income country * Which skill group category is predominant in the country and what are the particular skills under those category. * How much is the net skill migration per population of country for a particular skill group and category.

The dataset also provides information about the net inflow of the talent from/to the country from the year 2015 – 2019 with respect to a country name and skill group name. And this rate is multiplied by a factor of 10,000 for simplification purpose.

Some important variables from the datasets:

| Skill Migration Variable | Description |
|---|---|
| country_code & country_name | Country name given by World Bank taxonomy, and 2 letter |
| wb_region & wb_income | (World Bank Region & World Bank Income Group) Country |
| skill_group_name | Skill groups categorize the 50,000 detailed individual skills |
| net_per_10K_YYYY | Absolute 'netflow_YYYY' divided by 'total_member_ct_YYYY', |

**2. Human Development Index:**

This datasets consists of the HDI values of 189 countries from the year 2015-2019 along with their rankings.

# Preprocessing data:

**Manipulation 1:**

Skill Migration dataset is manipulated in such a way that earlier it had the net flow of skill for every year from 2015 -2019 in separate columns. Our interpreted data now has two columns with all the Years together in a single column along with the net flow.

**Manipulation 2:**

We merged the two datasets of skill migration and Human Development index into the above same table. We updated the country name as specified in the skill migration dataset into the Human development index dataset so as to perform a vlookup and get all the HDI values of the countries for different years into our final dataset.

**Manipulation 3:**

We regrouped some of the skill groups like Emergency medicine, healthcare management, Public health, family medicine etc. under Healthcare sector Industry for better understanding of various sector contribution to a country's skill migration.

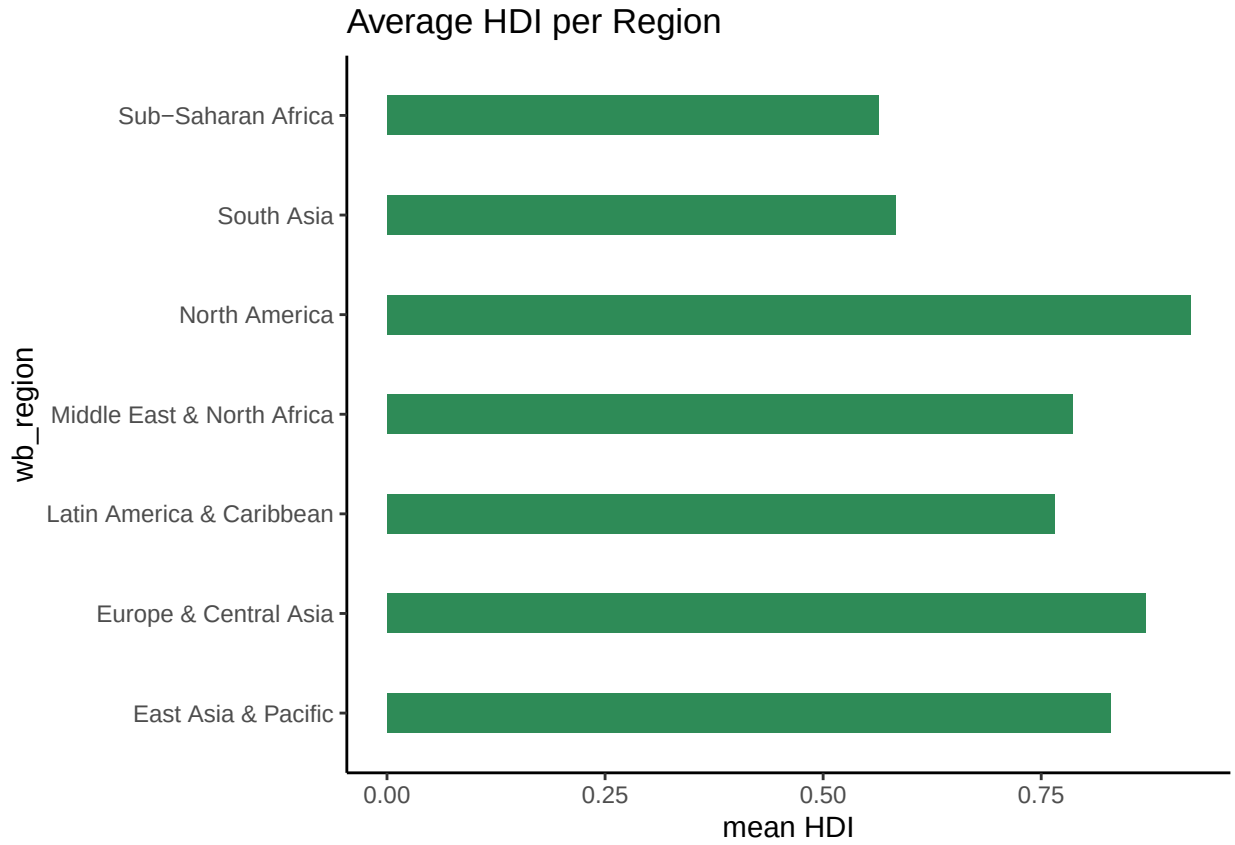
**Manipulation 4:**

In the Skill Migration dataset, we have data for 3 countries ("Taiwan", "Puerto Rico", "West bank and Gaza") which are not part of the HDI data acquired from the "United Nations Development Programme" as they are not officially recognized by the UN. To tackle this issue, We excluded the data of these 3 countries as these 3 countries together gives a data less than 100 rows which when compared to the overall data 14-17 thousand rows is less than 1%. Though it has limitations in drawing conclusions, we should also consider the fact that there is low penetration of LinkedIn membership in many developing countries, especially in the non-tradable, nontechnology, and non-digital sectors.

The resulting dataset contains **86548** observations with **11** variables.

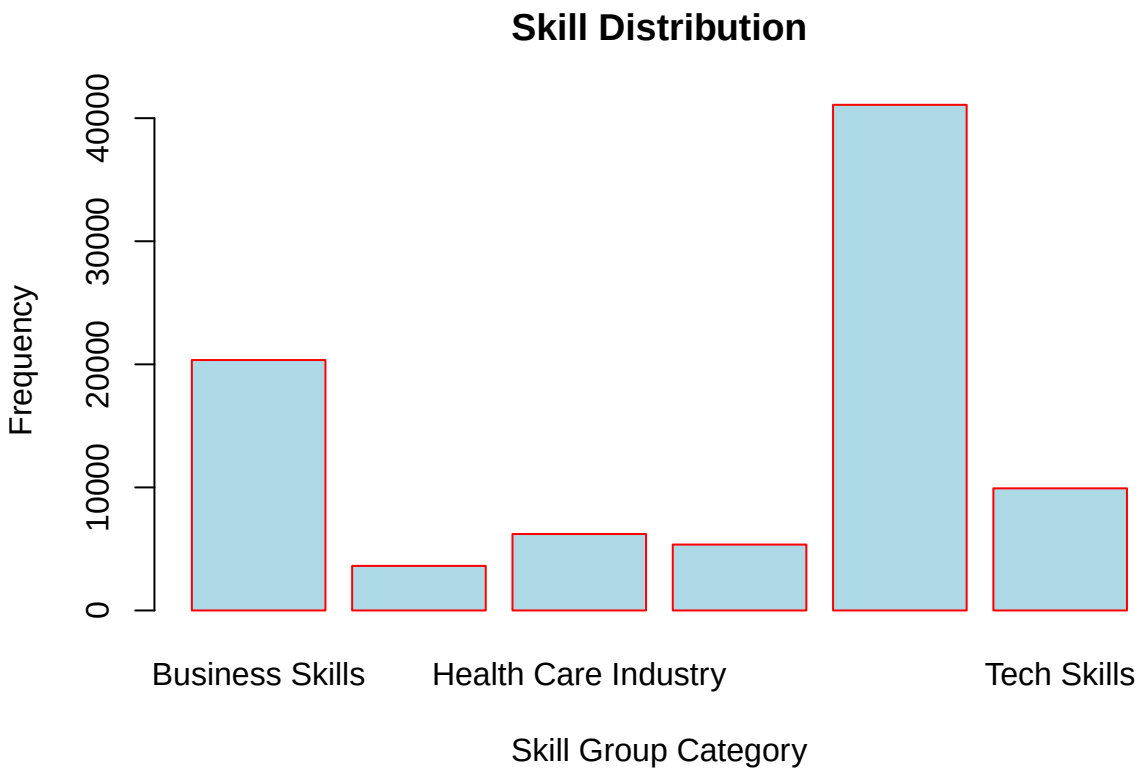
# Exploratory Data Analysis

## Average HDI value of different regions



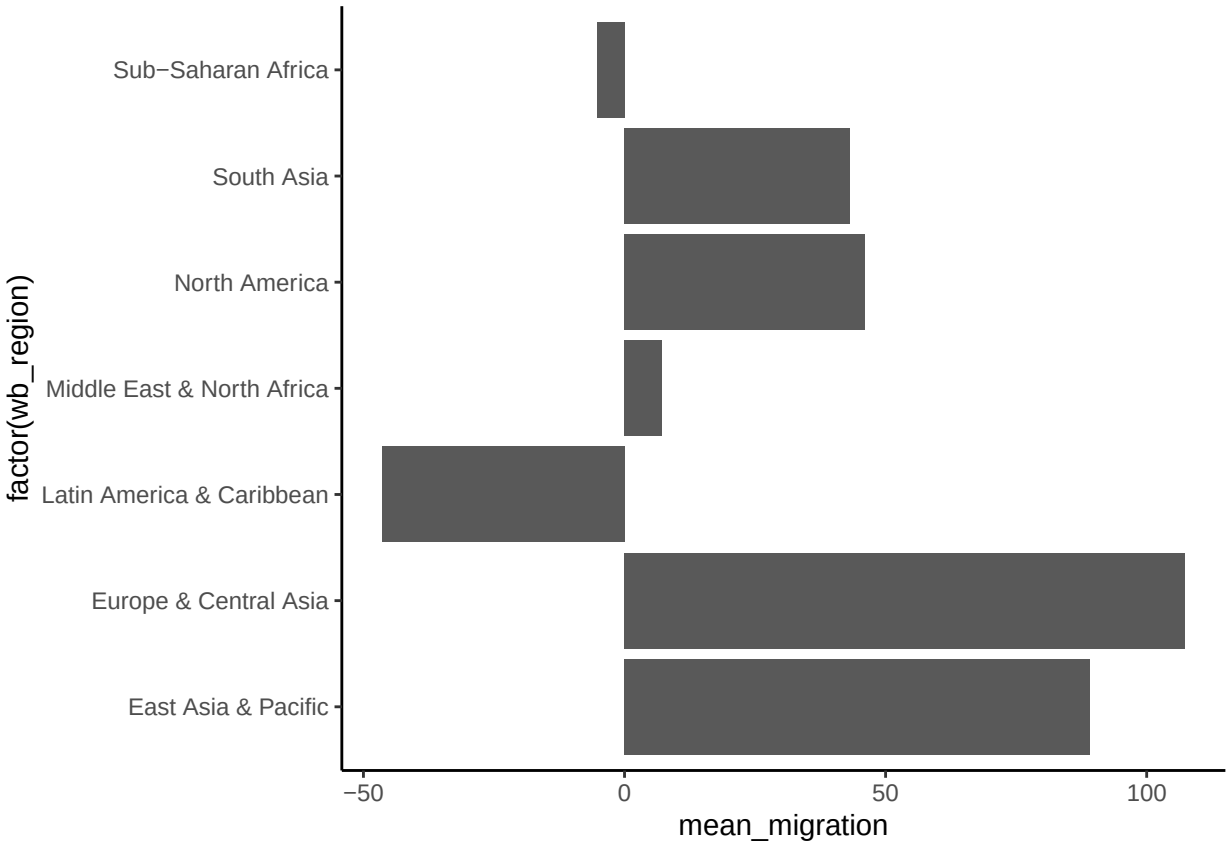

We could observe that the countries in North America and Europe and Central Asia enjoy a high average HDI value. We also tried to figure out which are the countries in Europe and Central Asia which are contributing most to the high value of HDI. And we found out from the below graph that countries like Switzerland, Unite Kingdom, Austria, Belgium etc have the highest contribution

**Various Skill Distribution:**

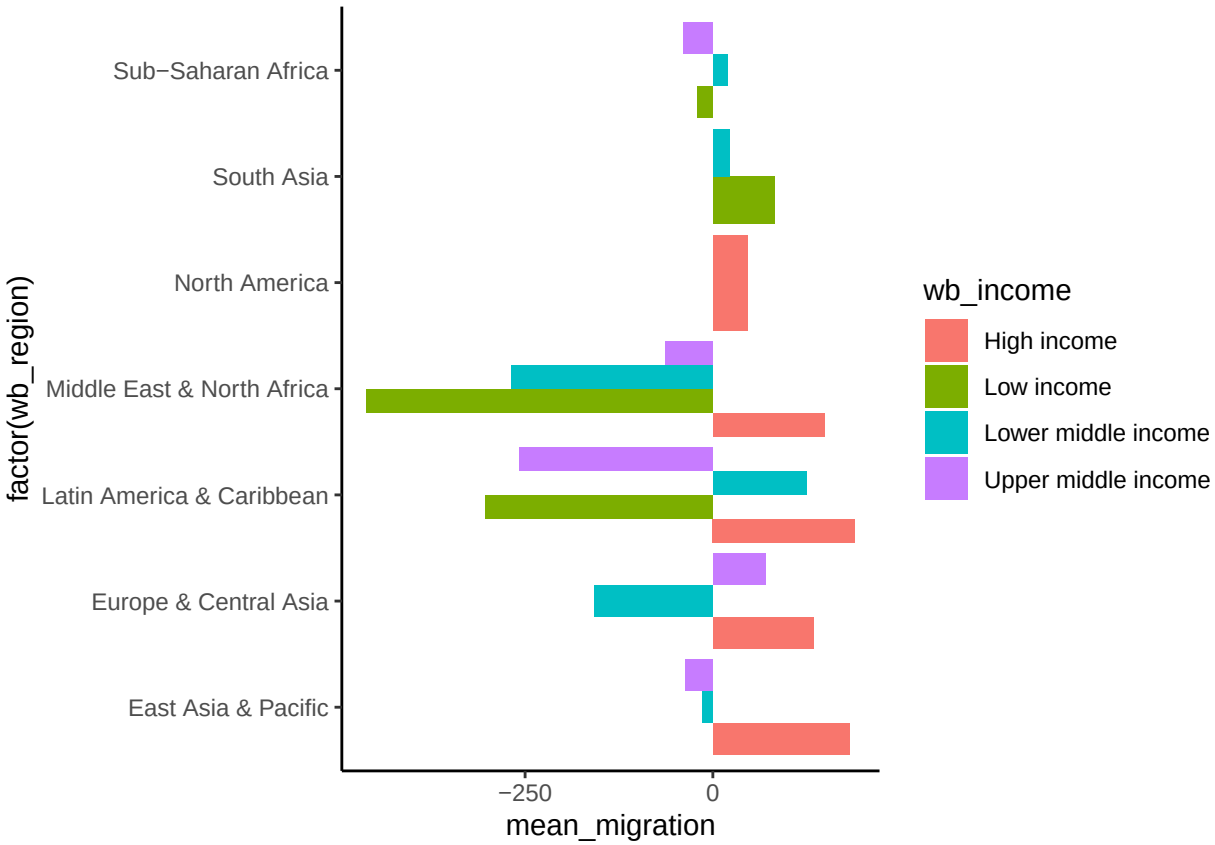



We saw that the most predominant skill in the market is specialized industry skills followed by Business skills and then tech skills. We tried to check the Skills contribution to the different country regions as well as to income regions. But we do not see any major difference in the contribution of one skill to a particular region.

**Skill migration with respect to a region and income**

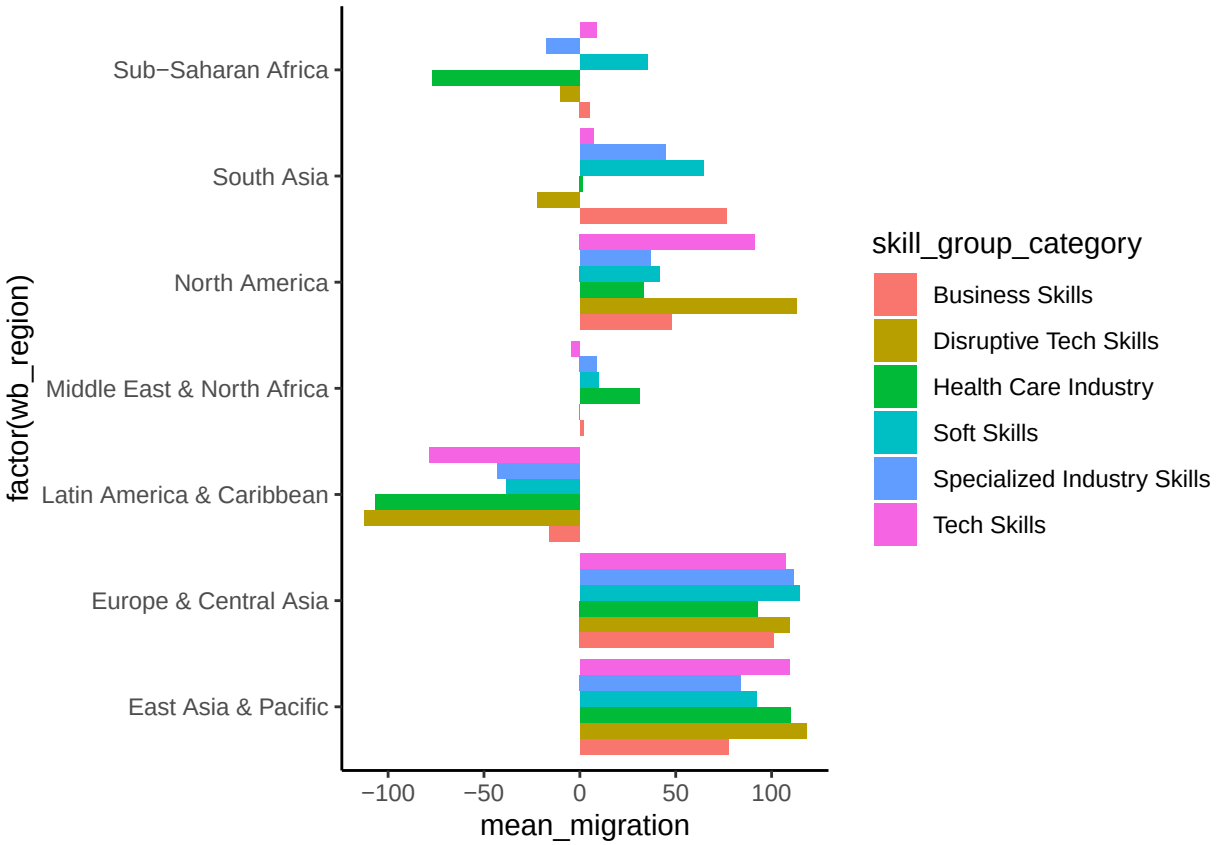



The above graph indicates that Europe and Central Asia saw the most incoming migration followed by East Asia & Pacific and North America and South Asia. On the other hand we saw that the most outgoing migration was observed in Latin America & Caribbean.




The above graph also indicates that the migration is irrespective of the income level.

**Skill migration with respect to a region and skill category**

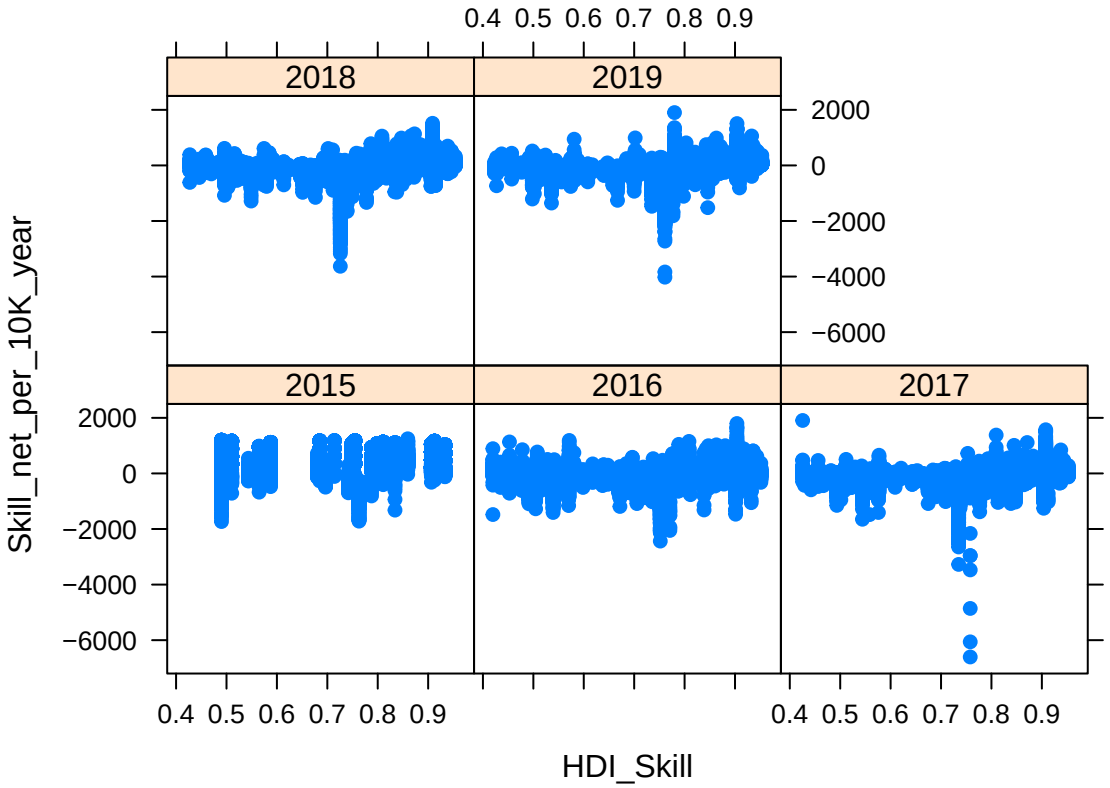



We can see from the graph that disruptive tech skill migration is predominant in regions like North America, Latin America & Caribbean, Europe & Central Asia and East Asia & Pacific.

# Empirical Analysis

## Model Selection

When selecting a model, the primary goal was to be able to assess the null hypothesis that has been proposed. The question being investigated in the first null hypothesis is to understand if there is a positive linear relationship between skill migration and HDI. The model should also be able to distinguish difference between the different income groups and skill groups that are of interest in the next two hypotheses. This can be incorporated into a single fully inclusive model that highlights the changes in HDI while also allowing for differences in skill groups and income levels. To evaluate if a linear analysis would be appropriate for this data, an exploration of a possible linear relationship between skill migration and HDI was explored using a x-y scatterplot which can be seen below.




The scatter plot has been divided to see the relationships occurring within each year. The data seems to follow a linear distribution that is either flat or that increases as the HDI value increases. These graphs give merit to analyzing the linear relationship between skill migration changes and HDI.

### Data Transformation:

The data displayed above shows fairly normal distribution across levels of HDI. The data does seem to be slightly skewed towards the higher side of HDI rankings, so exploratory data transformations were performed. When applying logarithmic changes to either HDI or skill migration, the data became more skewed than the original data without a data transformation. Therefore, the analysis was continued without data transformations for either of these two continuous variables.

### Model Setup

A multivariate linear analysis was conducted to investigate the three null hypotheses being investigated. The regression was set up to have Skill_net_per_10k_year_year as the dependent variable. The independent variables include HDI, skill categories, income levels, and year. Skill_net_per_10k_year was chosen as the dependent variable because the goal of this analysis is to understand the change that occurs to skill migration when there are changes to a countries HDI ranking, changes to the type of skill that is associated with the migration, or changes that occur when the countries have various income levels.

The dependent variable of skill migration and the explanatory variable HDI are both continuous. Skill migration is displayed as a value from -10,000 to 10,000, because it is showing the skill migration per 10,000 individuals. The skill category variable has been transformed into a categorical variable and has the categories of Tech skills, Business skills, Specialized Industry skills, Disruptive Tech skills, Soft skills, and Health Care skills. The model is built to have tech skills as the reference group, so it has been left out of the regression equation. This reference group was chosen to highlight differences between other skill groups the tech skill

group which is the question being analyzed within the second null hypothesis. The income levels used are high income, low income, upper middle income, and lower middle income. The reference group for these variables is high income so that comparisons can be made against this group in order to address the third null hypothesis.

This model includes the data from the years 2015 – 2019, the totality of the data that has been collected. To capture the effects that occur due to the change in the year, a year dummy variable has been added to the regression. The reference year for the year variable is the first year of data that has been collected, 2015. The usage of the year dummy variables is an example of fixed effects modeling for panel data. By using fixed effects for the years, any changes across the years are captured in the coefficient values for the given year that saw a change.

The final form of the multivariate regression is as follows.

**Skill_net_per_10k_year_year ~ HDI + Business skills + specialized industry skills + disruptive tech skills + Soft skills + healthcare industry + low income + upper middle income + lower middle income + Year(2016) + Year(2017) + Year(2018) + Year(2019)**

The following output is for the equation mentioned:

```
##
## Call:
## lm(formula = Skill_net_per_10K_year ~ HDI_Skill + Business_Skills +
##     Specialized_Industry_Skills + Disruptive_Tech_Skills + Soft_Skills +
##     Health_Care_Industry + Lowincome + Upper_middle_income +
##     Lower_middle_income + as.factor(Year), data = train.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6443.6   -84.4     8.7   105.8  2061.7
##
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  437.8007    19.2294   22.767  < 2e-16 ***
## HDI_Skill                     51.6516    21.3513    2.419 0.015560 *
## Business_Skills                7.4833     3.6807    2.033 0.042044 *
## Specialized_Industry_Skills   -0.4237     3.3621   -0.126 0.899713
## Disruptive_Tech_Skills        -0.4793     5.8512   -0.082 0.934711
## Soft_Skills                   19.6884     5.1128    3.851 0.000118 ***
## Health_Care_Industry         -15.6673     4.8814   -3.210 0.001330 **
## Lowincome                   -168.8327     9.2758  -18.201  < 2e-16 ***
## Upper_middle_income         -191.4981     3.7684  -50.817  < 2e-16 ***
## Lower_middle_income         -142.0314     6.1238  -23.193  < 2e-16 ***
## as.factor(Year)2016         -431.7368     3.2301 -133.661  < 2e-16 ***
## as.factor(Year)2017         -446.0863     3.2319 -138.027  < 2e-16 ***
## as.factor(Year)2018         -427.2444     3.2352 -132.062  < 2e-16 ***
## as.factor(Year)2019         -427.2391     3.2368 -131.995  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 268.3 on 68720 degrees of freedom
##   (66 observations deleted due to missingness)
## Multiple R-squared:  0.357,  Adjusted R-squared:  0.3569
## F-statistic:  2935 on 13 and 68720 DF,  p-value: < 2.2e-16
```

```
##
## Call:
```

```
## lm(formula = Skill_net_per_10K_year ~ HDI_Skill + Business_Skills +
##     Specialized_Industry_Skills + Disruptive_Tech_Skills + Soft_Skills +
##     Health_Care_Industry + Lowincome + Upper_middle_income +
##     Lower_middle_income + as.factor(Year), data = train.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6443.6   -84.4     8.7   105.8  2061.7
##
## Coefficients:
##                               Estimate Std. Error  t value             Pr(>|t|)
## (Intercept)                   437.8007    19.2294   22.767 < 0.0000000000000002
## HDI_Skill                      51.6516    21.3513    2.419             0.015560
## Business_Skills                 7.4833     3.6807    2.033             0.042044
## Specialized_Industry_Skills    -0.4237     3.3621   -0.126             0.899713
## Disruptive_Tech_Skills         -0.4793     5.8512   -0.082             0.934711
## Soft_Skills                    19.6884     5.1128    3.851             0.000118
## Health_Care_Industry          -15.6673     4.8814   -3.210             0.001330
## Lowincome                    -168.8327     9.2758  -18.201 < 0.0000000000000002
## Upper_middle_income          -191.4981     3.7684  -50.817 < 0.0000000000000002
## Lower_middle_income          -142.0314     6.1238  -23.193 < 0.0000000000000002
## as.factor(Year)2016          -431.7368     3.2301 -133.661 < 0.0000000000000002
## as.factor(Year)2017          -446.0863     3.2319 -138.027 < 0.0000000000000002
## as.factor(Year)2018          -427.2444     3.2352 -132.062 < 0.0000000000000002
## as.factor(Year)2019          -427.2391     3.2368 -131.995 < 0.0000000000000002
##
## (Intercept)                  ***
## HDI_Skill                    *
## Business_Skills              *
## Specialized_Industry_Skills
## Disruptive_Tech_Skills
## Soft_Skills                  ***
## Health_Care_Industry         **
## Lowincome                    ***
## Upper_middle_income          ***
## Lower_middle_income          ***
## as.factor(Year)2016          ***
## as.factor(Year)2017          ***
## as.factor(Year)2018          ***
## as.factor(Year)2019          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 268.3 on 68720 degrees of freedom
##   (66 observations deleted due to missingness)
## Multiple R-squared:  0.357,  Adjusted R-squared:  0.3569
## F-statistic:  2935 on 13 and 68720 DF,  p-value: < 0.00000000000000022
```

```
## [1] "lm"
```

```
##  [1] add1           alias          anova          case.names     coerce
##  [6] confint        cooks.distance cull_for_do    deviance       dfbeta
## [11] dfbetas        drop1          dummy.coef     effects        extractAIC
## [16] family         forecast       formula        fortify        getResponse
## [21] hatvalues      influence      initialize     kappa          labels
```

```
## [26] logLik         makeFun        model.frame    model.matrix   mplot
## [31] msummary       nobs           plot           predict        print
## [36] proj           qqnorm         qr             residuals      rstandard
## [41] rstudent       sample         show           simulate       slotsFromS3
## [46] summary        TukeyHSD       variable.names vcov
## see '?methods' for accessing help and source code
```

### Confidence Interval

```
##                                    2.5 %      97.5 %
## (Intercept)                  400.1112143  475.490222
## HDI_Skill                      9.8031550   93.500055
## Business_Skills                0.2691506   14.697362
## Specialized_Industry_Skills   -7.0134432    6.166024
## Disruptive_Tech_Skills       -11.9476332   10.988986
## Soft_Skills                    9.6673932   29.709435
## Health_Care_Industry         -25.2347894   -6.099743
## Lowincome                   -187.0131979 -150.652143
## Upper_middle_income         -198.8841683 -184.112058
## Lower_middle_income         -154.0340662 -130.028810
## as.factor(Year)2016         -438.0677819 -425.405857
## as.factor(Year)2017         -452.4208128 -439.751839
## as.factor(Year)2018         -433.5853166 -420.903397
## as.factor(Year)2019         -433.5832333 -420.895043
```

The confidence interval for the coefficients has also been provided.

### Accuracy

```
##                   ME     RMSE      MAE MPE MAPE
## Test set -0.5450084 274.8403 165.3647 NaN  Inf
```

A test for the models' accuracy was run to look at the RSME value for the model by using a data partition of 80% included in the training dataset and 20% of the observations being in the test dataset. All observations containing null values for skill migration and/or HDI index were removed prior to regression analysis. Preliminary models were produced prior to this model. They followed the same process but only were focused on a singular year. These models will not be thoroughly discussed, but comparisons of the results found in those models will be mentioned in this analysis to give evidence towards the validity of the relationships that are identified in this model.

## Regression Interpretation

### HDI

From the output it can be seen that the variable HDI was found to be significant at the $P < .05$ level. The p-value was found to be .016 with a t-value of 2.149. Since the p-value is below .05 it can be concluded that there is sufficient evidence to claim that the value of the coefficient for HDI is different from zero. This is also confirmed when looking at the confidence interval for HDI [9.8, 93.5]. Since zero is not included within the confidence interval then their evidence to reject the null hypothesis within the model on HDI which states that HDI is not different than zero.

The coefficient associated with HDI was found to be 51.65, meaning that a unit increase in HDI (a change from 0 to 1) will lead to an increase in skill migration of approximately 52 individuals per 10,000.

# Null Hypothesis 1 Evaluation

The HDI variable is of high interest to the first null hypothesis that was proposed which stated that skill migration increases to regions with higher HDI irrespective of income. Since HDI was found to be significant, there would be evidence to accept this null hypothesis as HDI was found to be significantly proven to be different than zero. Therefore, when HDI increases, there is sufficient evidence to say that skill migration would also increase.

It is interesting to note that when the preliminary models were run looking into a single year's data rather than all five years' worth of data, HDI was found to be significant in all the models. The coefficient associated with HDI in all these models was also positive. This provides supporting evidence for the finding that skill worker migration increases to regions with higher HDI, as a single unit increase in HDI would lead to a positive increase in the skill migration for a given country.

## Skill Group Categories

As previously mentioned all the skill categories for skill migration are being compared to the base case of tech skills. Of the five categories that were included in the output, three were found to be significant at the $p < .05$ level. Those three categories are business skills, soft skills and health care industry related skills.

Business skills was found to have a p-value of .04 with a t-value of 2.033. This finding leads to the conclusion that business skills are significantly different than tech skills in their relation to changes in skill migration per 10,000 migrants. The coefficient for business skills was found to be 7.48, indicating that business skills leads to an increase in skill migration by approximately 7-8 individuals per 10,000 compared to that of the base case tech skills. The confidence interval for this coefficient provides evidence that the value is different from zero at the 95% confidence level.

Soft skills were found to have a p-value of .0001 with a t-value of 3.851. This information leads to the finding that soft skills are significantly different than tech skills in their relation to skill migration. The coefficient associated with soft skills was found to be 19.69. This would indicate that soft skills are prevalent for skilled migrants by approximately twenty individuals per 10,000 more than that of tech skills. The confidence interval for soft skills covers the values from [9.67, 29.71]. This confidence interval confirms the fact that the coefficient for soft skills is a non-zero value since zero is not included in the confidence interval at a 95% confidence level. These values also give evidence that the soft skills coefficient is a positive value since the entirety of the interval is constructed by positive values.

Healthcare industry related skills were found be significantly different than that of tech skills with a p-value of .001 and a t-value of -3.21. With a p-value that is less than p=.05, it can be concluded that the coefficient associated with the variable is significantly different than zero. The coefficient associated with this variable was found to be -15.66. When comparing tech skills and healthcare skills, healthcare skills are found in approximately sixteen less individuals per 10,000 than that of individuals with tech skills. When looking to the confidence interval for health care industry [-25.23, -6.09], there is no inclusion of the value zero, and all the values are negative supporting this analysis.

Disruptive tech skills and specialized industry skills were all found to not be significantly different to the refence group of tech skills because all of the p-values associated with these variables was found to be greater than .05. The significance results for each of the five skill group categories included in this output was the same for all of the preliminary models that looked at individual year data, finding business skills, soft skills and healthcare skills to be significantly different to that of tech skills while the other categories were not found to be statistically significant.

# Null Hypothesis 2 Evaluation

The second null hypothesis that migration of workers with tech skills are predominant over workers with other specialized skills can be assessed with this information. The migration per 10,000 individuals with, specialized industry and disruptive tech skills are both found to not be different to that of tech skills. This

finding would reject the null hypothesis that tech skills are predominant skills over that of other skill groups for migrating individuals.

The results for soft skills would also support a rejection of the null hypothesis because the coefficient was both positive and significant, meaning that the model shows soft skills are more predominant to that of tech skills when looking at the migration of workers at a global scale across the five year span. This statement is also true for business skills because the value for the coefficient is also significant and greater than zero. Healthcare industry is the only skill category variable that gives evidence to support the second null hypothesis being explored. This is because the coefficient value was found to be negative and significant, meaning that the model indicates that healthcare related skills are found in less migrants per 10,000 than that of the category tech skills.

Overall, it can be concluded that there is more evidence to reject the second null hypothesis that was proposed. Two of the five categories were found to not be significantly different to tech skills in explaining the variation in the number of skilled migrants, two categories were found to show more migration than that of tech skilled migrants, and one category was found to show less migration than that of tech skilled migrants.

### Income Levels

The income related categorical variables are all being compared to the base case of high income. Low income, upper middle income, and high middle income were all found to be statistically significant at the $p < .05$ level. All the p-values for these variables were found to be approximately zero with t-values that had a negative value greater than 18. These p and t values give strong supporting evidence that all the income level outputs have coefficients that are statistically different than zero.

The coefficient value for upper middle income was found to be -191.5. This means that when the observation is found to be within the upper middle-income group, the expected skill migration per 10,000 individuals is approximately 192 people less than that of high-income countries or regions. The value found for lower middle-income group was found to be -142.03. This coefficient has a similar interpretation to that of upper middle income. When a country or region is found to have lower middle income, the expected skill migration per 10,000 individuals is approximately 142 people less than the skill migration that was found to occur in the reference group high income. Lastly, the low-income group was found to have a coefficient value of -168.83. Therefore, when a country or region has low income, the model shows a decrease in skilled migration of approximately 169 people per 10,000 compared to that of high-income areas.

## Null Hypothesis 3 Evaluation

The results for the various income groupings support the null hypothesis that high income countries or regions have increased skill migration to that area. Interestingly, the output shows that this trend is not consistent throughout the other income levels going from high to low. Lower middle-income groups have the second highest expected skill migration rather than the group of upper middle income. This finding suggests that high income and lower middle income have the highest levels of skill migration rather than high income and upper middle income. The low-income group also has a higher level of skill migration than that of the upper middle-income group. This finding suggests that high income group has the highest expected migration followed by lower middle income, followed by low income, followed by high middle income.

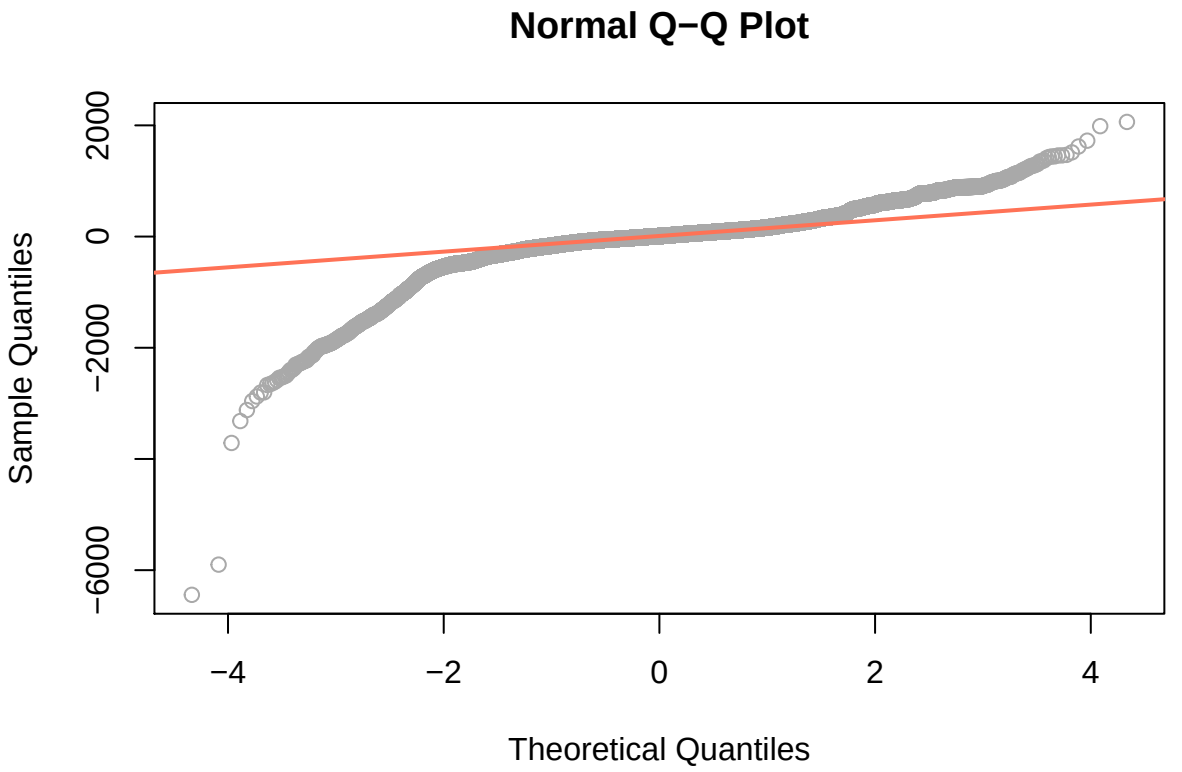
While these findings do support the null hypothesis that high income has the most skill migration, it also provided insights into the migration levels of the other income groups. The most important of which is that higher middle-income countries or regions experience less migration than that of lower middle income and low-income areas.

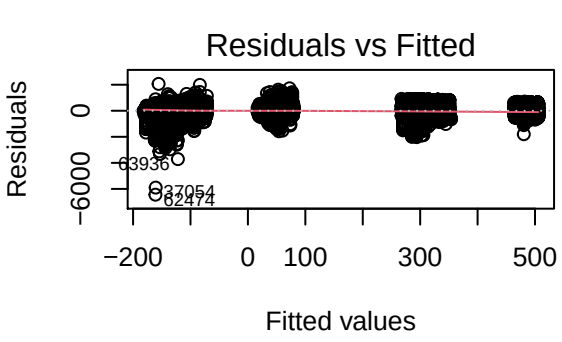
### Evaluation of Years

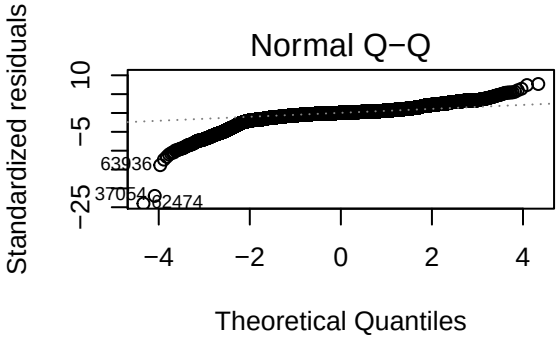
The year variable that was included in the model was not relevant to any of the null hypothesis, but they were still all found to be significant. These variables were put in place in order to give more reliable coefficient values for the other variables included in the model. The year variables can capture the changes that occur

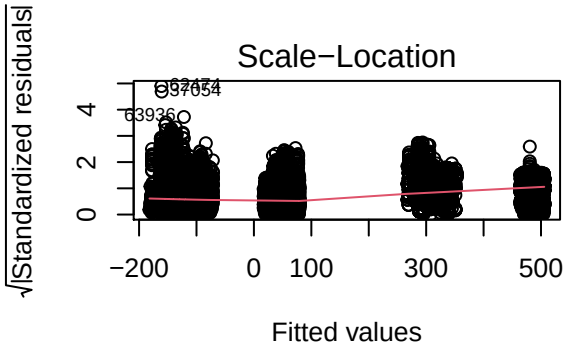
across the different years of the model that are not captured by the other variables that are included. All the year variables were found to have coefficients less than -400. This finding suggests that all years following the reference year of 2015 saw less skill migration for all the countries and regions when holding all other variables constant.

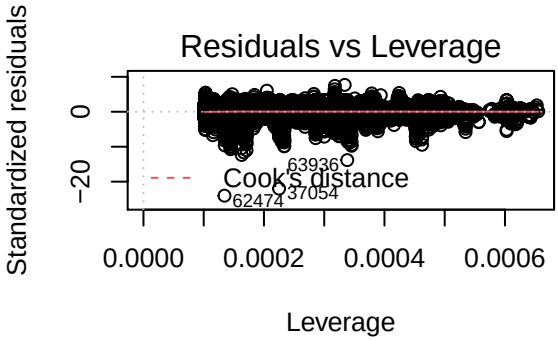
**Testing Assumptions**



The residuals vs. fitted plot show a concentration of residuals around zero. The groupings of the data are expected due to the inclusion of the year, income, and skill category indicator variables within the model that has grouped the data into different subsets, therefore the distribution is not continuous. The distribution of the data looks to be randomly distributed across all levels of the fitted values and does not contain any patterns throughout. The variance also seems to be similar across the distribution with small differences being observed within the first grouping. Overall, the residuals follow a linear pattern around zero signified by the red line within the graph.

The Normal Q-Q plot does show a slight deviation from a perfectly linear graph. The curved tails are of some concern, but with such a large number of observations falling within the expected linear pattern the plot has been considered acceptable.

The residuals vs. leverage plot does not contain any observations that extended outside the metric of Cook's distance, therefor the outliers were not removed from the model. It can be seen that three observations were distinguished as being outliers and they are all located on the lower side of the distribution

### Models Explanatory Power

The model was found to have a R2 value of .3569, meaning that the independent variables explained approximately 36% of the variation that is found in the data for skill migration. The model was found to have a p-value of approximately zero giving significance to the model, stating that at least one of the variables included had a linear relationship with skill migration that was different than zero. When tested against the test data set, the RMSE was found to be 274.84. This accuracy measure is consistent with the preliminary models focused on a single year's data. This provides evidence that the model including all year's data has similar predictive accuracy to the single year models.

## Conclusion

Using the skill migration and HDI data sourced from the World Bank Data and United Nations Development sites, we explored the factors that are influential to skill migration across the globe. This analysis provides information on the relationship between skill migration and HDI ranking, while also taking into consideration income levels of countries/regions, and the types of skills that migrants are bringing with them to their new target country.

The multi-variate regression model showed evidence to accept the first null hypothesis that skill migration and HDI have a positive linear relationship. HDI was found to have a significant linear relationship with skill migration levels when holding all of the other variables constant. The model also provided evidence to reject the second null hypothesis that tech skills led to more skill migration than all other skill groups being explored. Tech skills were not found to have higher levels of skill migration compared to the other skill groups except for when it was compared to that of health care industry migration. All other skill categories were found to not be statistically different or were found to have higher skill worker migration like that of the business skills and soft skill group. The regression model supported the third null hypothesis that high income countries have the highest skill migration across all the income levels included. The analysis also provided evidence that upper middle-income countries and regions have the lowest levels of migration compared to all other income levels.

This information can be used to understand migration trends around the world so that countries are able to understand what migrants are looking for in target countries, and what type of skills migrants are bringing with them. This knowledge can also allow nations to make better decisions regarding their labor force and the future migrants that their country might attract/lose in the future.

## Sources


- https://www.creighton.edu/fileadmin/user/CCAS/departments/PoliticalScience/MVJ/docs/ludlow.pdf

- World Bank Document - https://development-data-hub-s3-public.s3.amazonaws.com/ddhfiles/144635/wbg-linkedin-methodology-report_1.pdf