IST 687 Final Project - Team 3

Team Members: Daylin Hernandez, Christopher Murphy,
Chad Alessi, Chrisotpher Cavino

Submission Date: March 25, 2024

Subject of Data Analysis:

Skills Migration Trends of

LinkedIn Member Profiles (2015 – 2019)

Table of Contents

Introduction

Global talent mobility has become key to economic and social development, and understanding the dynamics of skill migration is important for policymakers, businesses, and educational institutions. The migration of skilled professionals across borders, driven by the search for better opportunities, has significant implications for sending and receiving countries.

This project, undertaken by Team 3, delves into the patterns of skills migration over five years from 2015 to 2019. Utilizing LinkedIn member profile data, we aim to uncover the underlying trends of professional mobility on a global scale. The chosen dataset offers a lens through which we can examine the movements of skilled labor, categorized by various attributes, including country code, country name, World Bank income classification, World Bank region, and specific skill groups.

Our analysis aims to identify the key attributes that characterize skill migration trends during the specified period. We also aim to pinpoint the specific skills that have experienced significant changes in migration rates. Additionally, our analysis strives to determine which regions have been most impacted by these migration trends. Lastly, we will highlight the skills that have shown the highest growth rates, thus shedding light on the global demand for emerging areas of expertise.

To achieve these objectives, our analysis involved data retrieval and munging, descriptive statistics and visualizations, and predictive modeling. We leveraged various statistical techniques, including linear modeling, support vector machine prediction modeling, and classification tree model prediction.

Business Questions

1. What are the key attributes?

2. What are the trends?

3. What are the skills experiencing the highest change?

4. Which region was most impacted by migration trends?

5. What are the highest growth rates for the top 10 skills?

Data Set Retrieval & Munging

1) For data retrieval, we read in the excel dataset and saved it in a data frame called skillMigrationDF.
2) We then checked for any missing information or NAs in the dataset using colSums(is.na()). With no missing information, we then perform str() to confirm all our data was correctly imported.
3) We kept all data from the original dataset. No data was replaced or removed.
4) Our next step is to separate out the individual regions.
    a. East Asia & Pacific data is filtered out and save into skillEAP.df
    b. Europe & Central Asia data is filtered out and saved into skillECA.df
    c. Latin America & Caribbean data is filtered out and saved into skillLAC.df
    d. Middle East & North Africa data is filtered out and saved into skillMENA.df
    e. North America data is filtered out and saved into skillNA.df
    f. South Asia data is filtered out and saved into skillSA.df
    g. Sub-Saharan Africa data is filtered out and saved into skillSSA.df
5) We then add a new column to the skillMigrationDF data frame called net_per_10k_cumulative. This new column is the sum of net_per_10k_2015, net_per_10k_2016, net_per_10k_2017, net_per_10k_2018, and net_per_10k_2019.
6) Our next steps were to create a new column called income_class. This column is binary data where a 0 means that the observation is classified under lower income or lower middle income. If it is a 1, the observation is classified as high or upper middle income.
7) Our final step in preparation of the dataset was to create another column which will be binary. In this column a 0 means a negative net_per_10k_2019, and a 1 is a positive net_per_10k_2019.

Skill Migration	
country_code & *country_name*	Country name given by World Bank taxonomy, and 2 letter country code
wb_region & *wb_income*	(World Bank Region & World Bank Income Group) Country categories classified by the latest World Bank region and income group
skill_group_name	Skill groups categorize the 50,000 detailed individual skills into approximately 250 skills groups (skill groups may be excluded based data quality considerations). For example, web development (skill group), may be composed of java, html etc. skills).
net_per_10K_YYYY	Absolute '*netflow_YYYY*' divided by '*total_member_ct_YYYY*', with respect to '*country_name*' and '*skill_group_name*', for specified year. Rate multiplied by 10,000 to simplify interpretation.

```{r}
str(skillMigrationDF)
```

```
'data.frame':   17617 obs. of  12 variables:
 $ country_code        : chr  "af" "af" "af" "af" ...
 $ country_name        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ wb_income           : chr  "Low income" "Low income" "Low income" "Low income" ...
 $ wb_region           : chr  "South Asia" "South Asia" "South Asia" "South Asia" ...
 $ skill_group_id      : num  2549 2608 3806 50321 1606 ...
 $ skill_group_category: chr  "Tech Skills" "Business Skills" "Specialized Industry Skills" "Tech Skills" ...
 $ skill_group_name    : chr  "Information Management" "Operational Efficiency" "National Security" "Software Testing"
...
 $ net_per_10K_2015    : num  -792 -1610 -1731 -958 -1511 ...
 $ net_per_10K_2016    : num  -706 -934 -770 -829 -841 ...
 $ net_per_10K_2017    : num  -550 -776 -757 -965 -842 ...
 $ net_per_10K_2018    : num  -681 -532 -600 -406 -582 ...
 $ net_per_10K_2019    : num  -1209 -790 -768 -740 -719 ...
```

Descriptive Statistics & Visualizations

The descriptive statistics served as the initial step in understanding and summarizing the data. They allow us to organize, visualize, and summarize the raw data to provide a clear and concise overview of the data's key features.

The first step taken for this purpose was calculating the average skill migration per year in each region using the function mean(). The results are summarized in the data frame below:

Average skill migration per year in each region (2015,2016,2017,2018,2019)

averageEAP	averageECA	averageLAC	averageMENA	averageNA	averageSA	averageSSA
43.28502	-15.4445564	-74.59658	6.533663	17.26113	-216.2011	32.35592
25.14751	-0.8243587	-121.26805	-75.543073	33.38045	-206.3829	-44.81603
-15.71553	9.9298714	-146.41812	-118.054603	43.98405	-181.8076	-82.27412
25.14251	33.8352893	-182.72687	-77.502287	56.51573	-151.9598	-75.29864
14.52680	44.9459627	-172.22399	-85.639665	65.59516	-189.5543	-71.27973

We observed that some regions such as LAC presented negative skill migration average for every year while NA showed positive skill migration average for every year.

Additionally, we extracted the row with the highest value of skill migration each year using the function max() and created a summary data frame that gave us an inside on which countries and regions were associated with the highest migration each year as shown below.

country_code	country_name	wb_income	wb_region	skill_group_id	skill_group_category	skill_group_name	net_per_10K_2015	net_per_10K_2016	net_per_10K_2017	net_per_10K_2018	net_per_10K_2019
ml	Mali	Low income	Sub-Saharan Africa	1655	Specialized Industry Skills	Army	2824.97	-1479.05	1906.14	76.53	-732.60
lu	Luxembourg	High income	Europe & Central Asia	921	Soft Skills	Teamwork	1657.96	1796.89	1572.35	1433.05	1345.65
ml	Mali	Low income	Sub-Saharan Africa	1655	Specialized Industry Skills	Army	2824.97	-1479.05	1906.14	76.53	-732.60
lu	Luxembourg	High income	Europe & Central Asia	20581	Specialized Industry Skills	Analytical Reasoning	1125.18	1646.73	1069.69	1515.79	1311.73
ge	Georgia	Lower middle income	Europe & Central Asia	2591	Business Skills	Customer Experience	236.22	-38.81	-205.13	781.72	1901.99

In the same way, we used the function min() to explore the countries with the lowest value of skill migration each year from every region as presented below.

country_code	country_name	wb_income	wb_region	skill_group_id	skill_group_category	skill_group_name	net_per_10K_2015	net_per_10K_2016	net_per_10K_2017	net_per_10K_2018	net_per_10K_2019
cu	Cuba	Upper middle income	Latin America & Caribbean	50304	Tech Skills	Mobile Application Development	-3037.38	-1968.18	-1155.10	-1332.28	-1708.36
ve	Venezuela, RB	Upper middle income	Latin America & Caribbean	31090	Soft Skills	Flexible Approach	-1209.32	-2435.26	-2542.23	-3629.02	-4022.04
dz	Algeria	Upper middle income	Middle East & North Africa	44	Specialized Industry Skills	Recruiting	-55.45	-100.74	-6604.67	-151.61	-260.71
ve	Venezuela, RB	Upper middle income	Latin America & Caribbean	31090	Soft Skills	Flexible Approach	-1209.32	-2435.26	-2542.23	-3629.02	-4022.04
ve	Venezuela, RB	Upper middle income	Latin America & Caribbean	31090	Soft Skills	Flexible Approach	-1209.32	-2435.26	-2542.23	-3629.02	-4022.04

As a result, Europe and central Asia were found to be the region with the highest positive skill migration in three out of the five years. Specifically, Mali was a country that had the maximum skill migration in two of the five years. Specialized Industry skills was the group category listed most frequently (three times) among the skills experiencing the maximum positive migration each year.

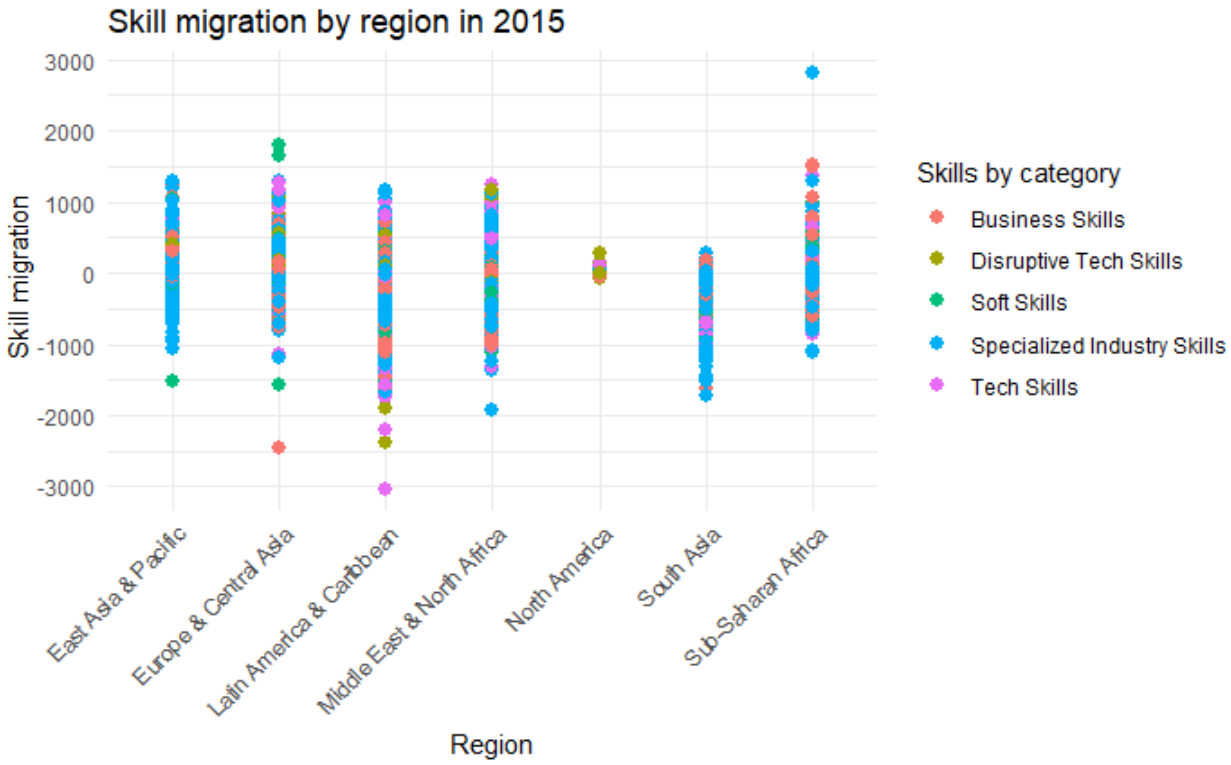
On the other hand, Latin America and Caribbean had the lowest value of skill migration in four of the five years and more specifically, Venezuela was repeated in three of the five years as the country experiencing more skill loss. Soft skills such as a flexible approach were demonstrated to be the most impacted, being listed as the minimum in three out of the five years.

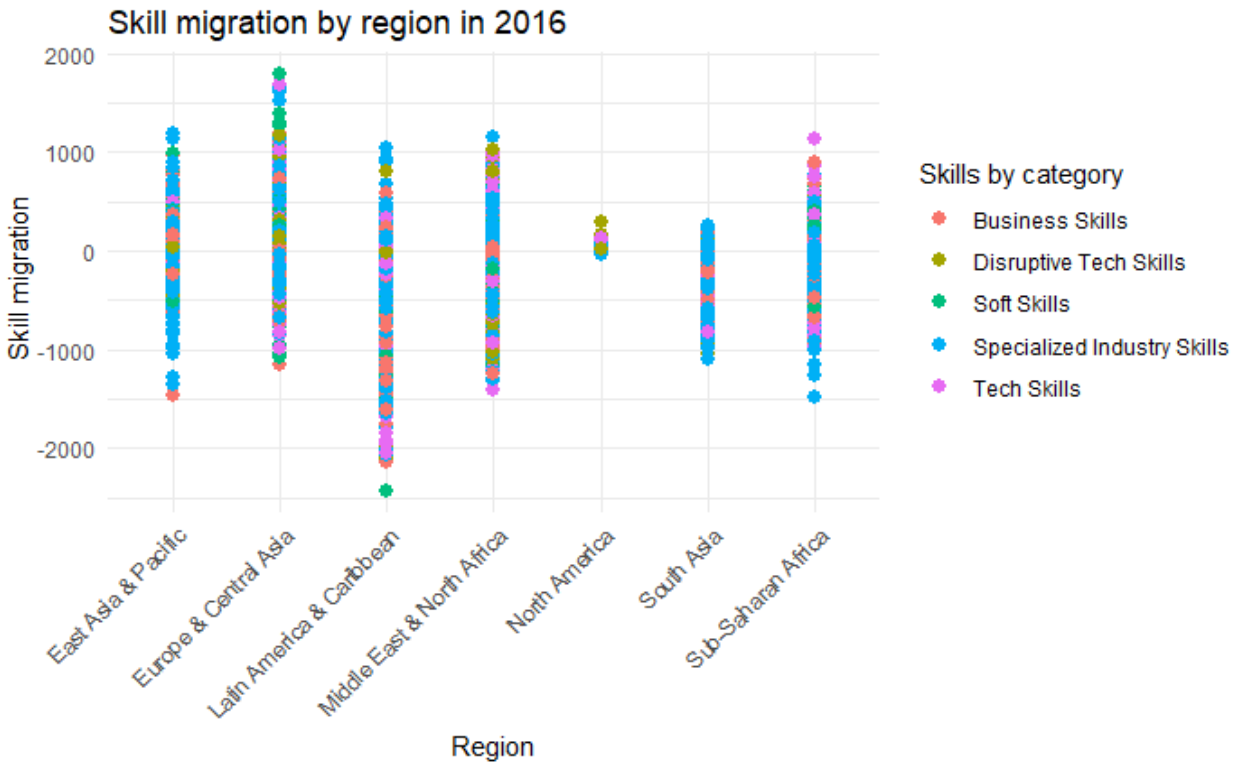
It was also noticed that the income class with minimum values of skill migration was the upper middle income for the five years.

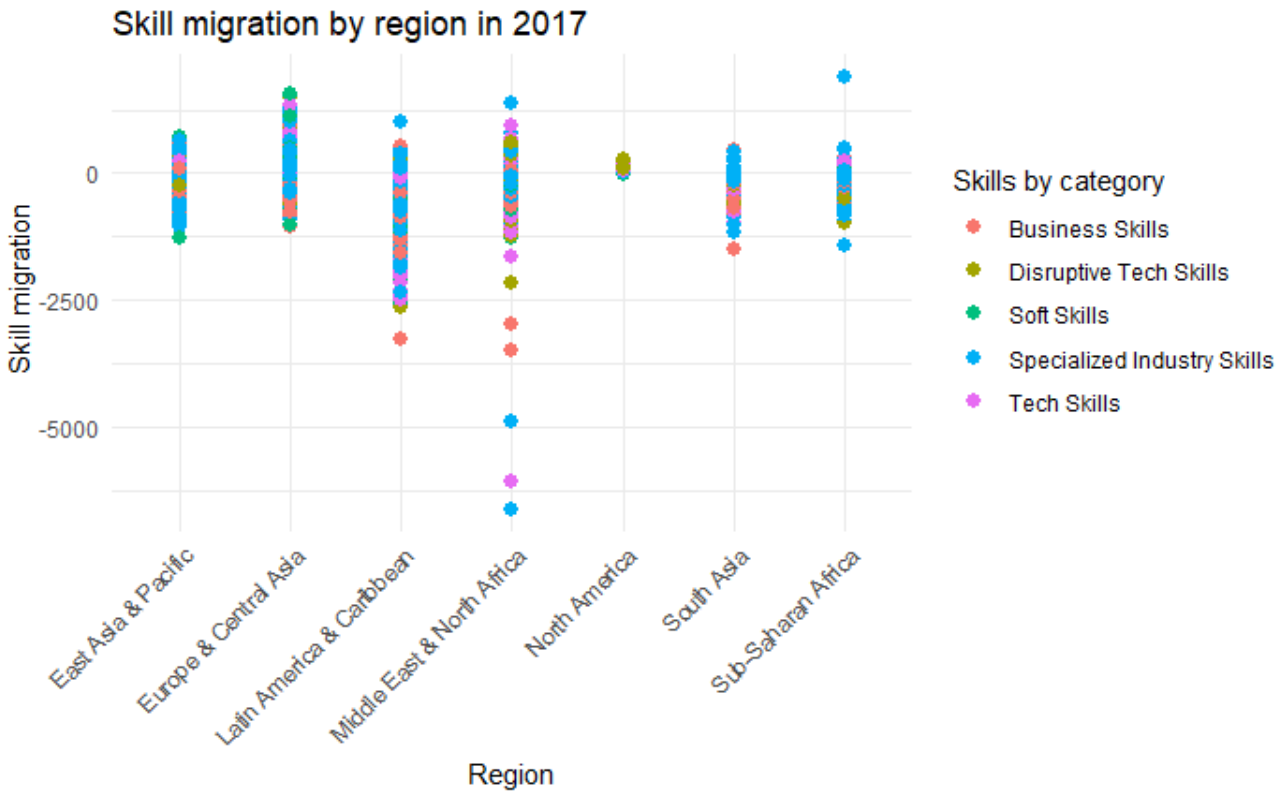
For the data visualization we used the ggplot() function to create scatter plot graphs that represented the skill migration by region each year with the color scale showing the skills classified by category for each region. An example of the code used for building up the graph for the year 2015 is presented below:

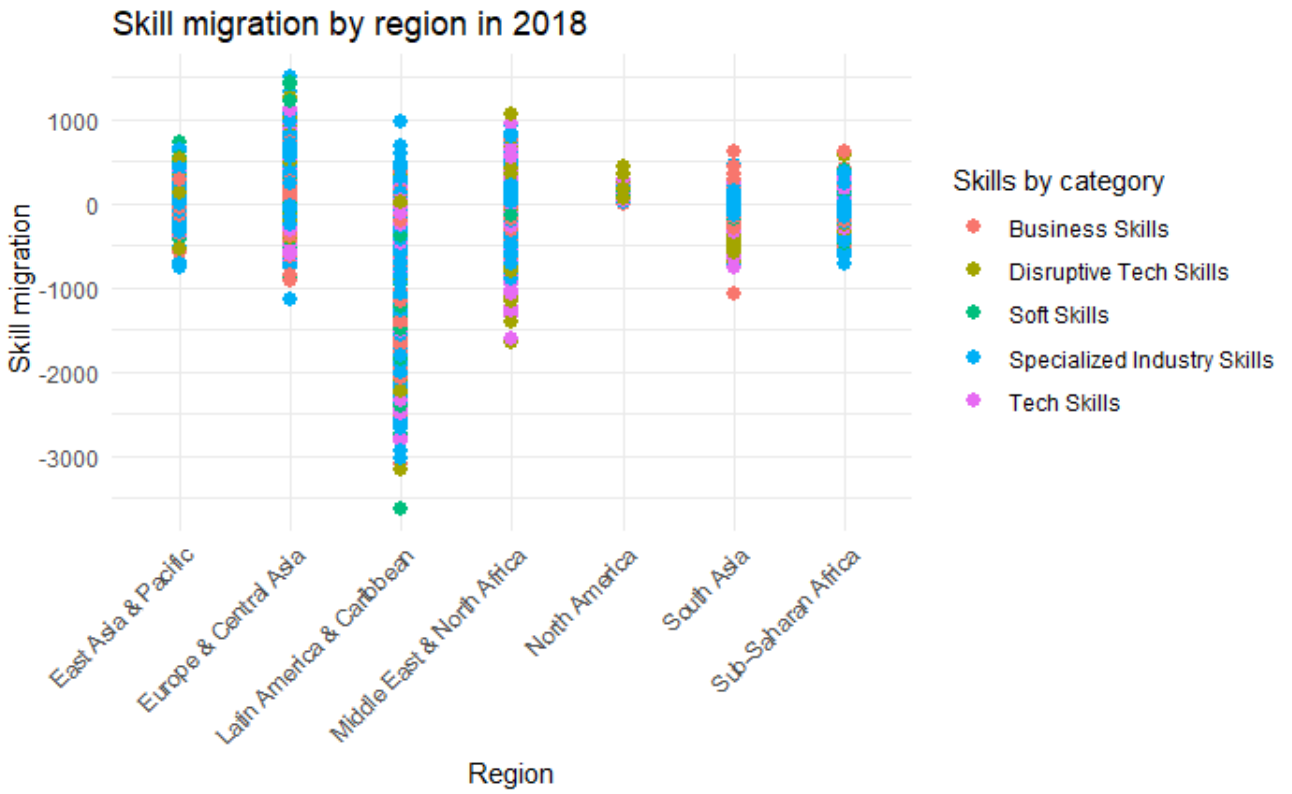
```
#Example of R code to create the graphs
plotskilldf15<-ggplot(skillMigrationDF, aes(x = wb_region, y =net_per_10K_2015, color= skill_group_category)) +
  geom_point(size = 2.5) +
  labs(x = "Region", y = " Skill migration", title = "Skill migration by region in 2015 ", color= "Skills by category") +
  theme_minimal()+
    scale_x_discrete(guide = guide_axis(angle = 45))
plotskilldf15
```

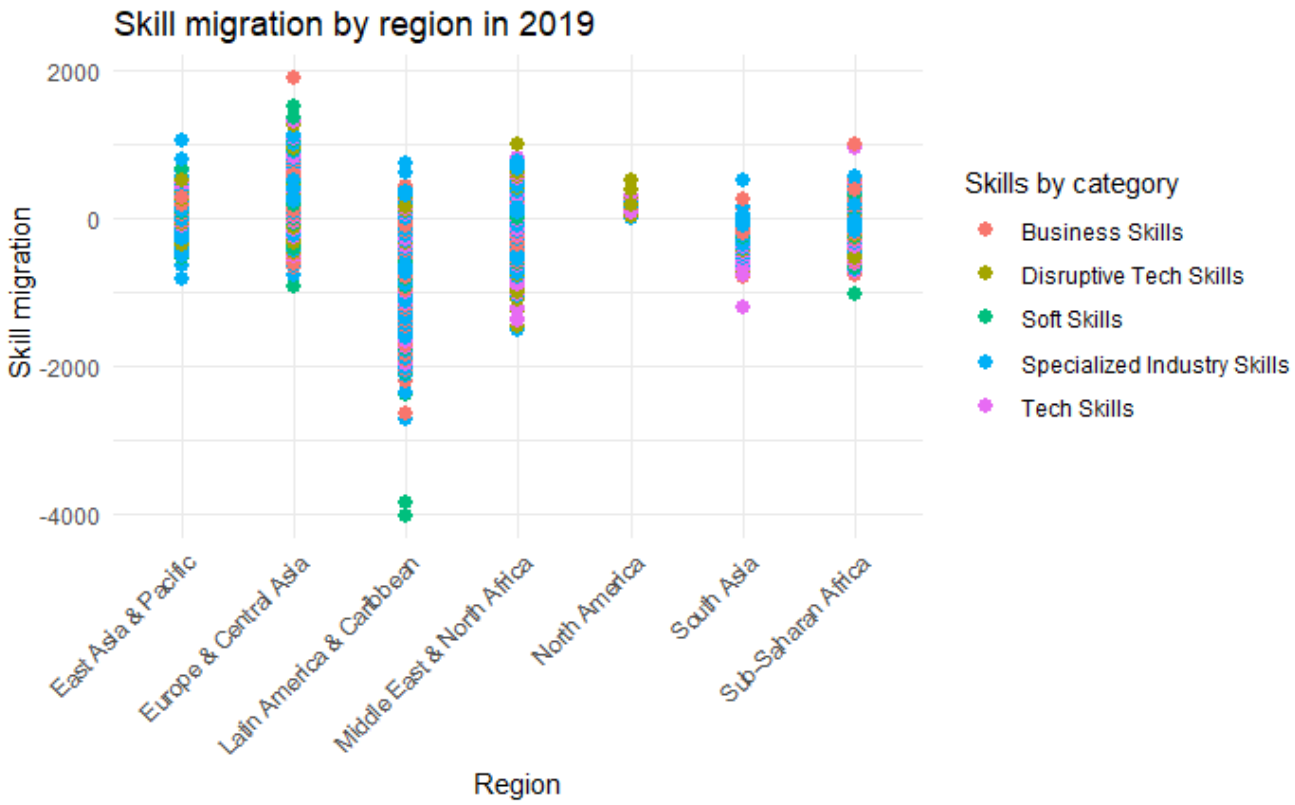
The code provided above was applied to the five years (2015, 2016, 2017,2018 and 2019) to generate the scatter plots. Each region's behavior was found to be similar for five consecutive years with LAC experiencing the biggest loss of skill in several years and NA having mainly positive skill migration over the years. The scatter plots supported the trends previously identified regarding the regions that have been most impacted by these migration trends and the specific skill migration.




Skill migration by region in 2016
Skill migration
2000
1000
0
-1000
-2000
East Asia & Pacific
Europe & Central Asia
Latin America & Caribbean
Middle East & North Africa
North America
South Asia
Sub-Saharan Africa
Region
Skills by category
Business Skills
Disruptive Tech Skills
Soft Skills
Specialized Industry Skills
Tech Skills


Skill migration by region in 2017
Skill migration
0
-2500
-5000
East Asia & Pacific
Europe & Central Asia
Latin America & Caribbean
Middle East & North Africa
North America
South Asia
Sub-Saharan Africa
Region
Skills by category
Business Skills
Disruptive Tech Skills
Soft Skills
Specialized Industry Skills
Tech Skills


Skill migration by region in 2018
Skill migration
1000
0
-1000
-2000
-3000
East Asia & Pacific
Europe & Central Asia
Latin America & Caribbean
Middle East & North Africa
North America
South Asia
Sub-Saharan Africa
Region
Skills by category
Business Skills
Disruptive Tech Skills
Soft Skills
Specialized Industry Skills
Tech Skills


Skill migration by region in 2019
Skill migration
2000
0
-2000
-4000
East Asia & Pacific
Europe & Central Asia
Latin America & Caribbean
Middle East & North Africa
North America
South Asia
Sub-Saharan Africa
Region
Skills by category
Business Skills
Disruptive Tech Skills
Soft Skills
Specialized Industry Skills
Tech Skills

Prediction Modeling & Visualizations

Linear Modeling

The following R code snippet is part of a data analysis process that involves fetching, preparing, and analyzing skill migration data.

```r
#Linear model

# Set URL for the Excel file
migrationInfoURL <- "https://datacatalogfiles.worldbank.org/ddh-published/0038044/DR0046256/public_use-talent-migration.xlsx?versionId=2024-02-13T16:57:39.2869535Z"

# Filter the skill migration data for relevant columns
skillMigrationDF1 <- skillMigrationDF[c("net_per_10K_2015", "net_per_10K_2016", "net_per_10K_2017", "net_per_10K_2018", "net_per_10K_2019")]

# Setup the linear regression model using the migration rates from 2015 to 2018 to predict the rates in 2019
linearModel <- lm(net_per_10K_2019 ~ net_per_10K_2015 + net_per_10K_2016 + net_per_10K_2017 + net_per_10K_2018, data = skillMigrationDF1)

# Summary of the linear model
summary(linearModel)

# Plot the correlation matrix of the net migration rates to visualize their relationships
corrMatrix <- cor(skillMigrationDF1)
corrplot(corrMatrix, method = "circle")

# Optional: Check diagnostics plots to evaluate assumptions (normality, linearity, homoscedasticity, and absence of multicollinearity)
par(mfrow = c(2,2))
plot(linearModel)
```

The analysis focuses on using linear regression to understand the relationship between net migration rates across different years, followed by evaluating the assumptions underlying the regression analysis. Here's a breakdown of what each part of the code is doing:

Fetching Data - Set URL for the Excel file: The code starts by defining a URL migrationInfoURL that points to an Excel file hosted on the World Bank's data catalog. This file contains data on talent migration, presumably including information on net migration rates for various skills or professions across different years.

Data Preparation - Filter the skill migration data for relevant columns: The skillMigrationDF dataframe, which presumably was read from the Excel file in a previous step not shown in the snippet, is filtered to keep only the columns for net migration per 10,000 people for the years 2015 to 2019. This step focuses the analysis on these specific years and prepares the data for regression analysis.

Linear Regression Model - Setup the linear regression model: A linear regression model linearModel is created using the lm() function. The model predicts the net migration rate for the year 2019 (net_per_10K_2019) as a function of the net migration rates from the years 2015 to 2018. This step aims to understand how past migration trends can predict future trends.

Summary of the linear model: The summary() function provides a detailed summary of the linear regression model, including coefficients, statistical significance of predictors, residuals information, and overall model fit (e.g., R-squared value). This summary helps in understanding the relationship between the variables and the predictive power of the model.

Correlation Matrix - Plot the correlation matrix: The cor()function calculates the correlation matrix for the selected columns of skillMigrationDF1, and corrplot() visualizes this matrix. This visualization helps in identifying any potential multicollinearity among predictors, where high correlation between independent variables can undermine the validity of the regression model.

Diagnostic Plots - Check diagnostics plots: The last part of the code uses the plot() function to produce diagnostic plots for the linear model, setting up a 2x2 plotting area with par(mfrow =

$c(2,2)$). These plots are used for evaluating the assumptions of linear regression, including normality of residuals, linearity and homoscedasticity of the relationship between predictors and the dependent variable, and the absence of multicollinearity (to some extent). Common plots include Residuals vs Fitted for checking linearity and homoscedasticity, Normal Q-Q for assessing normality of residuals, Scale-Location for verifying equal variance of residuals, and Residuals vs Leverage to identify influential data points.

Model Summary

The following provides a summary of the results of a linear regression analysis that aimed to predict the net migration per 10,000 people in 2019 (net_per_10K_2019) using the net migration rates from 2015 to 2018 as predictors. Here's a detailed explanation of the key components of this summary:

```
Call:
lm(formula = net_per_10K_2019 ~ net_per_10K_2015 + net_per_10K_2016 +
    net_per_10K_2017 + net_per_10K_2018, data = skillMigrationDF1)

Residuals:
     Min       1Q   Median       3Q      Max
-1677.78   -42.89     2.49    39.72  1382.51

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.204348   0.813821  -5.166 2.42e-07 ***
net_per_10K_2015  0.088106   0.005128  17.183  < 2e-16 ***
net_per_10K_2016  0.069205   0.006868  10.077  < 2e-16 ***
net_per_10K_2017  0.050497   0.006018   8.391  < 2e-16 ***
net_per_10K_2018  0.659981   0.005813 113.539  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 104.7 on 17612 degrees of freedom
Multiple R-squared:  0.8095,    Adjusted R-squared:  0.8094
F-statistic: 1.871e+04 on 4 and 17612 DF,  p-value: < 2.2e-16
```

Coefficients

The coefficients table shows the estimated effects of each predictor variable on the target variable, net migration in 2019, along with their statistical significance:

Estimate: The regression coefficients indicating the expected change in the target variable for a one-unit change in the predictor variable, holding all other variables constant.

- The intercept is -4.204348, which would be the predicted value of net_per_10K_2019 when all predictors are zero.
- The positive coefficients for all years suggest that higher net migration rates in previous years are associated with higher rates in 2019.
- The coefficient for net_per_10K_2018 is notably larger than for the other years, indicating a stronger relationship between the 2018 migration rate and the 2019 rate.

Std. Error refers to the standard error of the estimate, reflecting the variability of the estimate.

t value is the coefficient divided by its standard error, used to determine the statistical significance of the coefficient.

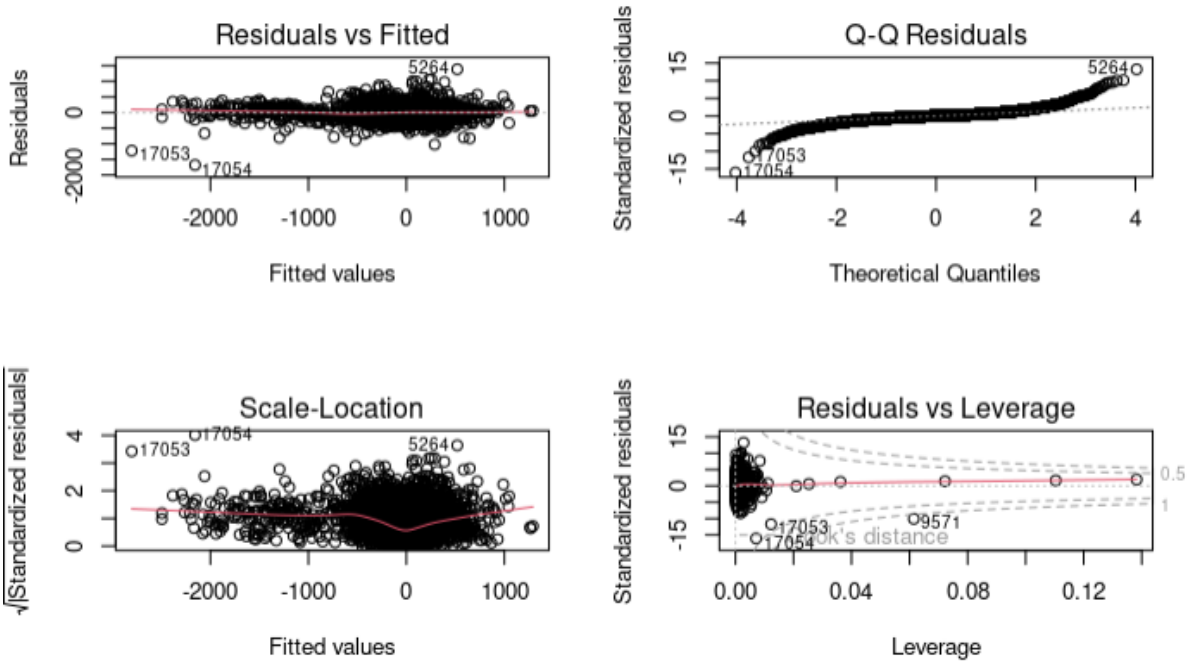
Pr(>|t|) indicates the p-value associated with the t-statistic, testing the null hypothesis that the coefficient is equal to zero (no effect). The significance codes at the bottom help interpret these p-values, with *** indicating a very strong relationship.

Model Fit

- **Residual standard error**: An estimate of the standard deviation of the residuals, which is 104.7. It gives an idea of the typical error in predictions.
- **Multiple R-squared**: 0.8095, representing the proportion of variance in the dependent variable that is predictable from the independent variables. This high value suggests a strong model fit.
- **Adjusted R-squared**: 0.8094, adjusts the R-squared value based on the number of predictors and the sample size, providing a more accurate measure of model fit.
- **F-statistic and its p-value**: The F-statistic tests the null hypothesis that all regression coefficients are equal to zero (i.e., the model has no explanatory power). The extremely low p-value (< 2.2e-16) here rejects this null hypothesis, indicating that the model is statistically significant.

The model suggests a strong relationship between past net migration rates and the net migration rate in 2019, with all predictors being statistically significant. The high R-squared values indicate that the model explains a substantial portion of the variance in the 2019 net migration rate. However, the wide range of residuals suggests that there may be outliers or other factors not captured by the model that could influence the accuracy of predictions for specific observations.

Diagnostics Plot



Residuals vs Fitted:

- **Purpose**: This plot is used to check the assumption of homoscedasticity (constant variance) of residuals. It's also useful for identifying non-linearity in the relationship that the model may not have captured.
- **Interpretation**: Ideally, you want to see a random scatter of points without any discernible pattern. Patterns or a funnel shape would suggest non-constant variance (heteroscedasticity) or non-linear relationships. In this plot, the residuals seem randomly distributed around the zero line without any clear pattern, indicating that the model's variance is constant, and the relationship might be linear.

Q-Q Residuals:

- **Purpose**: The Quantile-Quantile plot of residuals is used to check the normality of residuals. It compares the distribution of residuals to a normal distribution.
- **Interpretation**: A straight line of points suggests that residuals are normally distributed. In this plot, most of the points follow the line closely, but there's a slight deviation at the ends, indicating some departure from normality, especially for extreme values.

Scale-Location (or Spread-Location):

- **Purpose**: This plot, also known as a Scale-Location plot, is used to check if residuals are spread equally along the ranges of predictors (homoscedasticity). It's like the Residuals vs Fitted plot but uses the square root of the absolute values of residuals.
- **Interpretation**: Like the first plot, you're looking for a random scatter without a pattern. The presence of a pattern could indicate heteroscedasticity. This plot shows a relatively even spread, suggesting homoscedasticity, although there's a slight increase in spread for higher fitted values.

Residuals vs Leverage:

- **Purpose**: This plot helps identify influential cases (outliers that have a significant impact on the model's parameters). The Cook’s distance lines (dashed lines) help identify these influential points.
- **Interpretation**: Points that are far from the center horizontally are points with high leverage. Points that are far from zero in the vertical direction are outliers. Points that are outside the Cook's distance lines might be influencing the regression results. In this plot, there are a few points labeled that might be of concern due to their leverage and/or residual values, but none seem to exceed the Cook's distance threshold significantly.

Overall, these plots suggest that the model fits reasonably well.

Support Vector Machine Prediction Modeling

Dataset preparation

To prepare our dataset for prediction models we created a subset of our data keeping only the numeric fields for yearly net skills migration and the region. In addition, we added a binary dummy variable named “blended_higher_income” to classify data from higher income countries.

- If the original record income value was “high income” or “upper middle income”, the blended_higher_income value was set to 1.
- If the original record income value was “low income” or “lower middle income”, the blended_higher_income value was set to 0.

In addition, the following binary dummy variables for each income classification were then created as factors and added to the data set.

- High_income
  - If the original record income value was “High income”, the High_income value was set to 1 (else, 0).
- Upper_middle_income
  - If the original record income value was “upper middle income”, the Upper_middle_income value was set to 1 (else, 0).
- Lower_middle_income
  - If the original record income value was “lower middle income”, the Upper_middle_income value was set to 1 (else, 0).
- *Note*: no dummy variable was created for “Low income”.

(R code) Creation of the streamlined dataset:

```
modelingDataFieldsDF <- data.frame(blended_higher_income=calculatedFieldsDF$blended_higher_income,
                                   high_income=calculatedFieldsDF$high_income,
                                   upper_middle_income=calculatedFieldsDF$upper_middle_income,
                                   lower_middle_income=calculatedFieldsDF$lower_middle_income,
                                   net_per_10K_2015=calculatedFieldsDF$net_per_10K_2015,
                                   net_per_10K_2016=calculatedFieldsDF$net_per_10K_2016,
                                   net_per_10K_2017=calculatedFieldsDF$net_per_10K_2017,
                                   net_per_10K_2018=calculatedFieldsDF$net_per_10K_2018,
                                   net_per_10K_2019=calculatedFieldsDF$net_per_10K_2019,
                                   wb_region=calculatedFieldsDF$wb_region)

modelingDataFieldsDF$wb_region <- as.factor(modelingDataFieldsDF$wb_region)
```

Classification Prediction with a Support Vector Machine

A support vector machine was then used to create a classification model to predict blended_higher_income. The support vector machine was trained on a 70% subset of the blended_higher_income dummy variable field.

(R Code) Creation of the support vector machine:

```
#make the sampling predictable by setting the seed value
set.seed(1000)

#select 70% training data and store indexes
trainList <- createDataPartition(y=modelingDataFieldsDF$blended_higher_income,p=0.70,list=FALSE)

#include all of the elements at the indices in the training data set
trainingDataSet <- modelingDataFieldsDF[trainList,]

#construct the test set from everything that didn't go into training
testingDataSet <- modelingDataFieldsDF[-trainList,]

#train the model
svm.model.blendedHigherIncome <- train(blended_higher_income~.,
                  data=trainingDataSet,method="svmRadial",
                  trControl=trainControl(method="none"),
                  preProc=c("center","scale"))
```

Assessing Fit of the Support Vector Machine Model

We then used the predict () function to assess the accuracy of the support vector machine model using the testing dataset representing a 30% subset of the blended_higher_income dummy variable field. Once the prediction was run, we used the confusionMatrix() function to output the test results and the accuracy of the SVM Model.

(R Code): Assessing fit of the SVM model with test data

```
svmPrediction <- predict(svm.model.blendedHigherIncome, testingDataSet, type="raw")

#build a confusion matrix with caret package function
confusionMatrix <- confusionMatrix(svmPrediction,testingDataSet$blended_higher_income)
```

Results of the Support Vector Machine Model

The support vector machine we trained resulted in a very accurate model for predicting blended_higher_income records in the test data set. The prediction model accuracy was reported at 0.9996, which was far more accurate than the “no information rate” of 0.7515. Also, the confusion matrix showed highly accurate results in plotting the prediction against the blended_higher_income values from the test dataset.

This model would be helpful in determining the likelihood of skills migration data being linked to a particular region or a set of annual net skills migration data.

(R Output): Prediction vs. Reference, and Confusion Matrix Statistics

```
          Reference
Prediction    0    1
         0 1313    2
         1    0 3969
```

```
               Accuracy : 0.9996
                 95% CI : (0.9986, 1)
    No Information Rate : 0.7515
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.999

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.9995
         Pos Pred Value : 0.9985
         Neg Pred Value : 1.0000
             Prevalence : 0.2485
         Detection Rate : 0.2485
   Detection Prevalence : 0.2489
      Balanced Accuracy : 0.9997

       'Positive' Class : 0
```

Interpretation / Insights

- The ability to classify the income level of the country generating migration data may provide insight to how income affects the gain or loss of critical skills.
- Gaining insight into skills migration trends between countries of at the same or different income levels may provide a valuable starting point to policy makers investigating market forces driving migration.
- Migration data for specific skills may then be analyzed to inform country strategy / policy in the face of migration trends for critical skills.

Classification Tree Model Prediction

Team 3 further analyzed the ability to properly predict the blended_higher_income value using a tree model to classify data. For this tree model, we again used 70% of the data set to train the model using the rpart() function, withholding 30% for testing the accuracy of the model. The resulting classification tree is shown below.

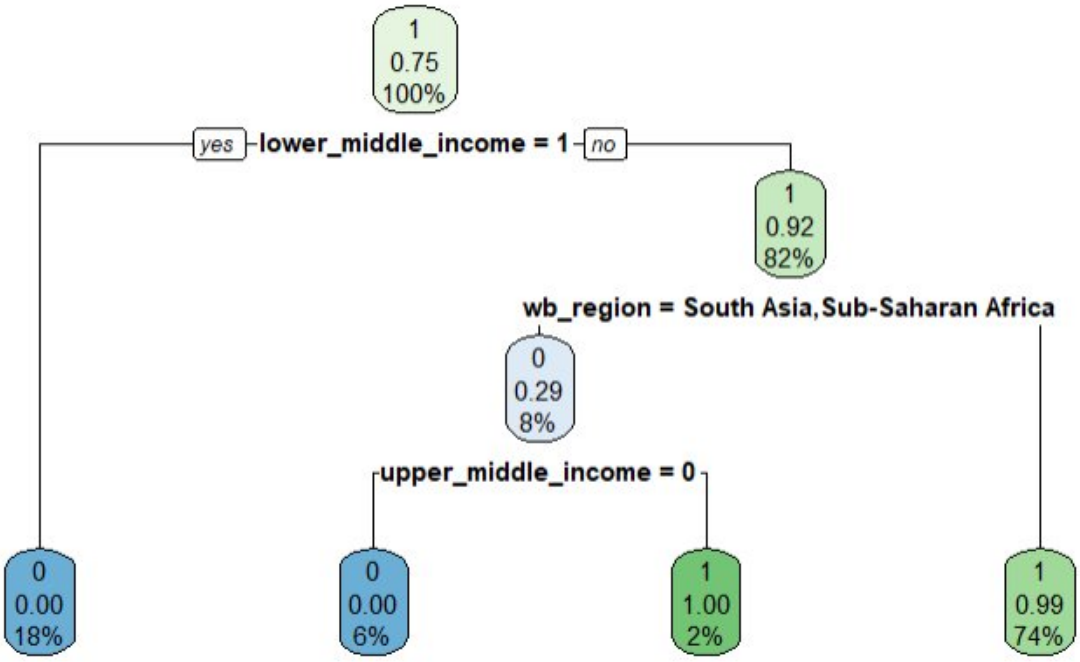
(R Code) Building the classification tree

```
#build the model
model.rpart.blendedHigherIncome <- rpart(blended_higher_income~.,data=trainingDataSet,method="class")

#visualize the results using rpart.plot()
rpart.plot(model.rpart.blendedHigherIncome)
```

Visualizing the Classification Tree

The image below shows the classification and/or decision tree logic of the resultant model. The tree has three layers of checks. First, it checks to see if the lower_middle_income is = 1 and if the answer is =”no”, the second check then validates if the region is either South Asia or Sub-Saharan Africa. If the second check = “no”, the model assumes that the data is from a “High income or “Upper middle income” country (representing 74% of the training data).



Assessing Accuracy of the Classification Tree Model

The team then used the predict() function to assess the accuracy of the classification tree model.

(R Code): Assessing fit of the RPart model with test data

```
#make predictions on the test set
predictions <- predict(model.rpart.blendedHigherIncome, testingDataSet, type="class")

#output results in table format
table(testingDataSet$blended_higher_income,predictions)
```

The following prediction table shows a high accuracy of this model in properly classifying the testing data set. The model has a near-perfect ability to predict when data will have a blended_higher_income = 0 and has a very low error rate for predicting when data will have a blended_higher_income = 1.

```
  predictions
     0    1
0 1266   47
1    0 3971
```

Interpretation / Insights

- The ability to classify the income level of the country generating migration data may provide insight to how income affects the gain or loss of critical skills.
- Gaining insight into skills migration trends between countries of at the same or different income levels may provide a valuable starting point to policy makers investigating market forces driving migration.
- Migration data for specific skills may then be analyzed to inform country strategy / policy in the face of migration trends for critical skills.

Summary

The objective of this project was to explore the global skill migration, leveraging a dataset of LinkedIn member profiles spanning from 2015 to 2019. Through analysis, we were able to uncover some interesting insights into the dynamics of talent mobility.

Our findings showed key attributes that characterize skill migration trends, notably the impact of World Bank income classifications and regional affiliations on migration patterns. We observed distinct trends that underscore the role of economic prosperity and regional opportunities in influencing the flow of skilled professionals.

The analysis revealed that Europe and Central Asia emerged as the regions most impacted by migration trends, showcasing a significant influx of skilled labor, particularly from countries like Mali, which experienced maximum skill migration in two of the five years analyzed. Conversely, Latin America and the Caribbean faced the lowest skill migration rates, with Venezuela marking the least amount of skill migration in three of the five years. Additionally, all the regions with minimum skill migration rates were classified as upper middle income.

We successfully quantified the relationships between various factors and skill migration rates by employing a suite of predictive modeling techniques, including linear modeling, support vector machine prediction, and classification tree model prediction. These models validated our descriptive analyses and provided a robust framework for forecasting future migration trends based on current data.

In conclusion, our investigation into skill migration trends offers valuable insights for stakeholders across multiple sectors. For policymakers, understanding these trends can inform the development of strategies to attract and retain top talent. It highlights the importance of adapting to the changing landscape of skills demand for businesses. And for educational institutions, it underscores the need to align curricula with the evolving needs of the global job market.