Analysis- Regression

Bomi So

Part 1. Principal Component Analysis

1. Introduction

Immigration is a significant phenomenon in contemporary society, and the UK is a popular destination for many migrants, particularly in Europe. According to the UK Parliament [4], the number of immigrants in the UK has been steadily increasing over time, and there was a significant increase in 2020 after Covid-19 restrictions were eased.

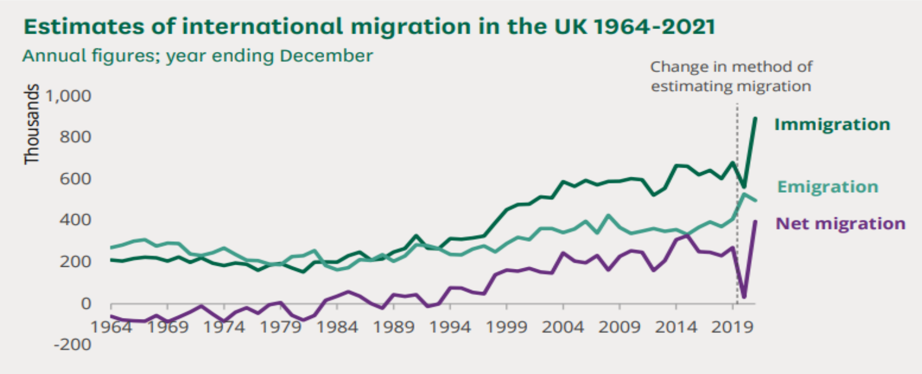


Figure 1 Migration annual figure, source: UK parliament

Given this phenomenon, this paper focuses on the impact of different family types on immigration to the UK and asks: 'How do different family types impact immigration in the UK?'. More specifically, it focusses on capital city of UK, London. Therefore, research question is set as below:

Research question	How do different family types of impact on number of immigrations in London?

2.Reseach method

For answering the research question – How different types of family impacts on migration number in London, Principal Component Analysis (PCA) is conducted via SPSS. PCA is a method for reducing the number of variables in a dataset that has many interrelated variables [5]. The goal of PCA is to transform the original variables into a new set of variables called "principal components" (PCs) that are uncorrelated and ordered in a way that the first few PCs retain as much of the variation present in the original variables as possible [5]. This allows for the dimensionality of the data to be reduced while still preserving as much information as possible [5]. This analysis concept is firstly introduced by Karl Pearson in 1901 [6]. Principal component analysis (PCA) is a popular method for data analysis because it is a non-parametric technique that can be used to identify important variables in a complex dataset with multiple variables [7].

In this research, PCA is used to identify important independent variables that have a statistically significant impact on the dependent variable. By doing so, PCA helps to improve the accuracy of the test.

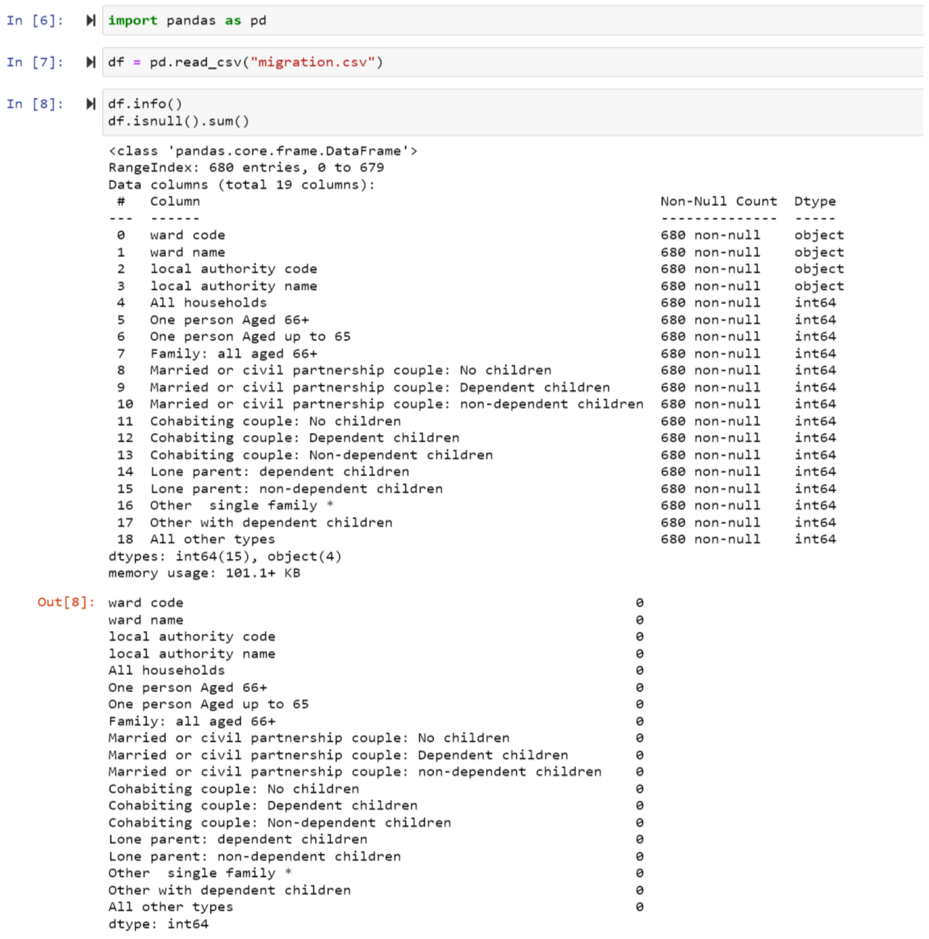
3. Data Description

a. Dataset

The dataset was obtained from the London Data Store, which is operated by the UK parliament. The data was collected during the 2021 London census and is copyrighted by the Greater London Authority [8]. It includes information on the total number of migrants in London in 2021, specifically detailing each migrant's family type.

b. Data pre-processing

Data pre-processing was completed using Python. The characteristics of the data were identified, and appropriate pre-processing techniques were chosen. The dataset includes 19 variables and a total of 680 data points for each variable, based on the census survey of all migrants in London in 2021. No missing data was identified. To improve the accuracy of the analysis, five columns (variables) were removed. More specially, four categorical variables including local authority and district codes and names (such as ward code, ward name, local authority code, and local authority name) were excluded from the analysis. Also, one variable, "all other types," was removed due to its ambiguity. This variable, which indicates "others" in the dataset, is ambiguous in determining the exact factor of the house type, therefore, it is removed from the research and factor analysis.



```
In [6]:  import pandas as pd

In [7]:  df = pd.read_csv("migration.csv")

In [8]:  df.info()
         df.isnull().sum()

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 680 entries, 0 to 679
         Data columns (total 19 columns):
          #   Column                                                  Non-Null Count  Dtype
         ---  ------                                                  --------------  -----
          0   ward code                                               680 non-null    object
          1   ward name                                               680 non-null    object
          2   local authority code                                    680 non-null    object
          3   local authority name                                    680 non-null    object
          4   All households                                          680 non-null    int64
          5   One person Aged 66+                                     680 non-null    int64
          6   One person Aged up to 65                                680 non-null    int64
          7   Family: all aged 66+                                    680 non-null    int64
          8   Married or civil partnership couple: No children        680 non-null    int64
          9   Married or civil partnership couple: Dependent children 680 non-null    int64
          10  Married or civil partnership couple: non-dependent children 680 non-null    int64
          11  Cohabiting couple: No children                          680 non-null    int64
          12  Cohabiting couple: Dependent children                   680 non-null    int64
          13  Cohabiting couple: Non-dependent children               680 non-null    int64
          14  Lone parent: dependent children                         680 non-null    int64
          15  Lone parent: non-dependent children                     680 non-null    int64
          16  Other  single family *                                  680 non-null    int64
          17  Other with dependent children                           680 non-null    int64
          18  All other types                                         680 non-null    int64
         dtypes: int64(15), object(4)
         memory usage: 101.1+ KB

Out[8]:  ward code                                                   0
         ward name                                                   0
         local authority code                                        0
         local authority name                                        0
         All households                                              0
         One person Aged 66+                                         0
         One person Aged up to 65                                    0
         Family: all aged 66+                                        0
         Married or civil partnership couple: No children            0
         Married or civil partnership couple: Dependent children     0
         Married or civil partnership couple: non-dependent children 0
         Cohabiting couple: No children                              0
         Cohabiting couple: Dependent children                       0
         Cohabiting couple: Non-dependent children                   0
         Lone parent: dependent children                             0
         Lone parent: non-dependent children                         0
         Other  single family *                                      0
         Other with dependent children                               0
         All other types                                             0
         dtype: int64
```

4. Result

a. Correlation test

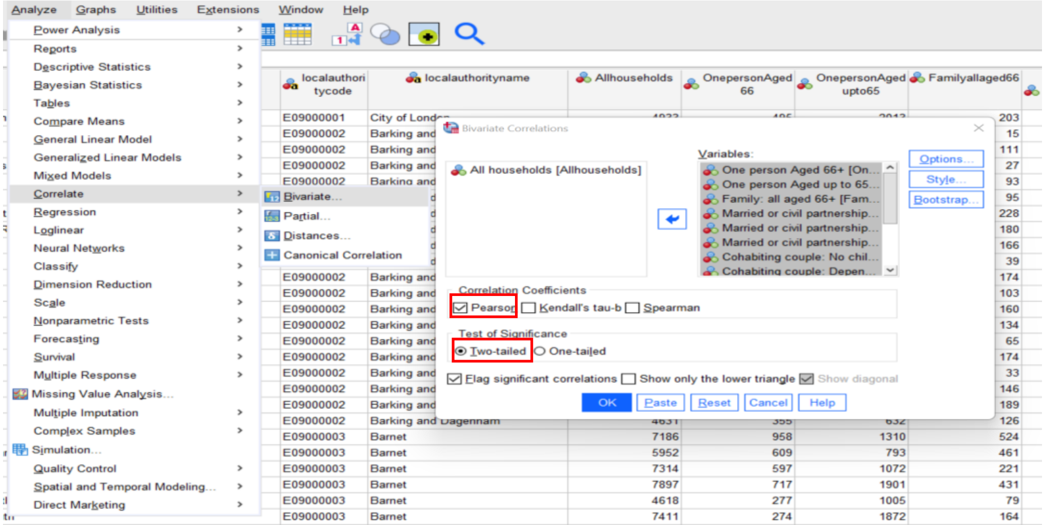
Firstly, bivariate correlation analysis is conducted via SPSS in order to analyse correlation between variables. For visualization, R program is used. The detailed steps are outlined as below.

Also, the hypothesis is set.

H0 (Null Hypothesis)	There is no statistically significant linear relationship between variables.
HA (Alternative Hypothesis)	There is a statistically significant linear relationship between variables.

Lastly, it assumes independent observation of the values of two variables. Test is conducted in significant level 0.05.

1) Test steps



Bivariate Correlations

All households [Allhouseholds]

Variables:
One person Aged 66+ [On...
One person Aged up to 65...
Family: all aged 66+ [Fam...
Married or civil partnership...
Married or civil partnership...
Married or civil partnership...
Cohabiting couple: No chil...
Cohabiting couple: Depen...

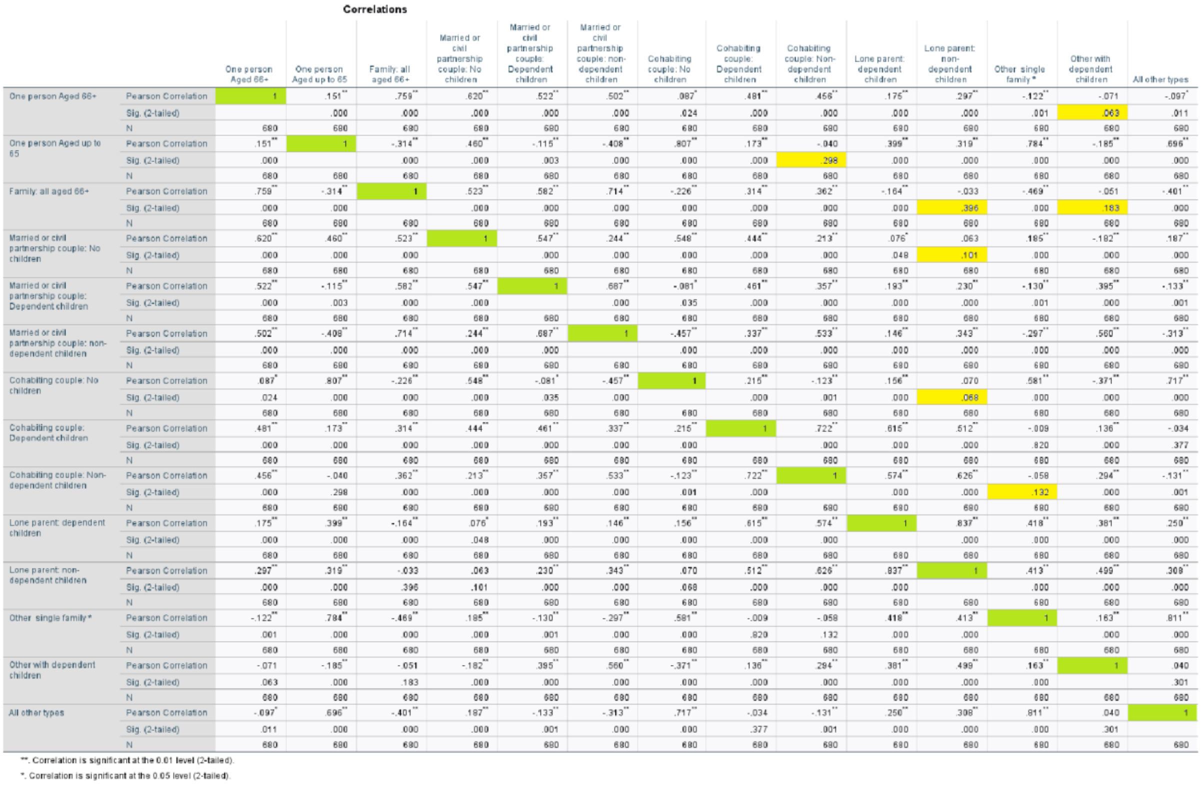
Correlation Coefficients
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations ☐ Show only the lower triangle ☑ Show diagonal

2) Test result

According to test results, the majority of the variables' correlation tests have significant evidence to reject null hypothesis as the significant level (P-values) is less than 0.05 and close to 0.000. It indicates that the null hypothesis is rejected in favour of the alternative hypothesis. This confirms a statistically significant linear relationship between these variables. On the other hand, cases where the P-value is greater than 0.05 do not have statistically significant evidence for rejecting the null hypothesis. These cases are highlighted in yellow in the table below and indicate that there is not enough evidence to decide on a linear relationship between variables.



Correlations

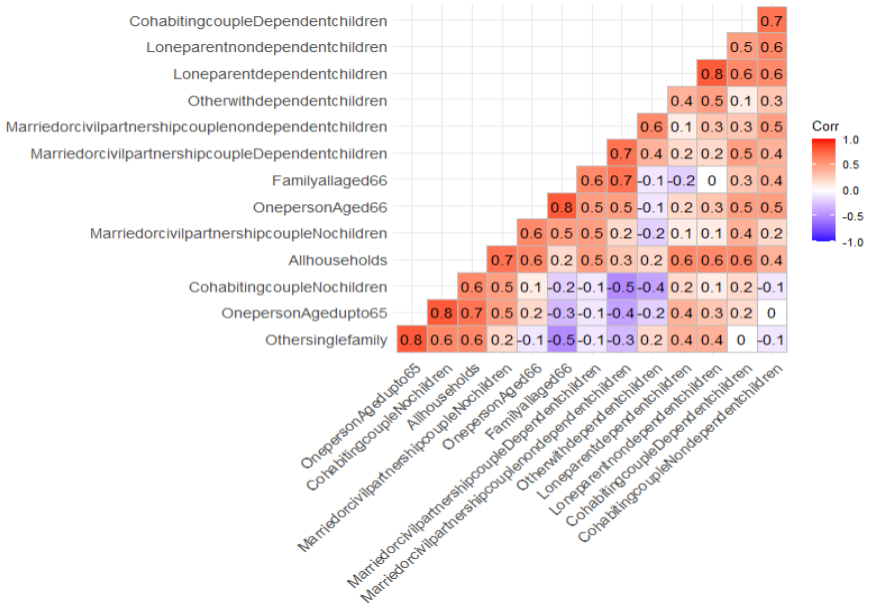
		One person Aged 66+	One person Aged up to 65	Family: all aged 66+	Married or civil partnership couple: No children	Married or civil partnership couple: Dependent children	Married or civil partnership couple: non-dependent children	Cohabiting couple: No children	Cohabiting couple: Dependent children	Cohabiting couple: Non-dependent children	Lone parent: dependent children	Lone parent: non-dependent children	Other single family*	Other with dependent children	All other types
One person Aged 66+	Pearson Correlation	1	.151**	.759**	.620**	.522**	.502**	.087*	.481**	.456**	.175**	.297**	-.122**	-.071	-.097*
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.024	.000	.000	.000	.000	.001	.063	.011
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
One person Aged up to 65	Pearson Correlation	.151**	1	-.314**	.460**	-.115**	-.408**	.807**	.173**	-.040	.399**	.319**	.784**	-.185**	.696**
	Sig. (2-tailed)	.000		.000	.000	.003	.000	.000	.000	.298	.000	.000	.000	.000	.000
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
Family: all aged 66+	Pearson Correlation	.759**	-.314**	1	.523**	.582**	.714**	-.226**	.314**	.362**	-.164**	-.033	-.469**	-.051	-.401**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000	.000	.000	.396	.000	.183	.000
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
Married or civil partnership couple: No children	Pearson Correlation	.620**	.460**	.523**	1	.547**	.244**	.548**	.444**	.213**	.076*	.063	.185**	-.182**	.187**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000	.000	.048	.101	.000	.000	.000
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
Married or civil partnership couple: Dependent children	Pearson Correlation	.522**	-.115**	.582**	.547**	1	.687**	-.081*	.461**	.357**	.193**	.230**	-.130**	.395**	-.133**
	Sig. (2-tailed)	.000	.003	.000	.000		.000	.035	.000	.000	.000	.000	.001	.000	.001
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
Married or civil partnership couple: non-dependent children	Pearson Correlation	.502**	-.408**	.714**	.244**	.687**	1	-.457**	.337**	.533**	.146**	.343**	-.297**	.560**	-.313**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
Cohabiting couple: No children	Pearson Correlation	.087*	.807**	-.226**	.548**	-.081*	-.457**	1	.215**	-.123**	.156**	.070	.581**	-.371**	.717**
	Sig. (2-tailed)	.024	.000	.000	.000	.035	.000		.000	.001	.000	.068	.000	.000	.000
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
Cohabiting couple: Dependent children	Pearson Correlation	.481**	.173**	.314**	.444**	.461**	.337**	.215**	1	.722**	.615**	.512**	-.009	.136**	-.034
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.820	.000	.377
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
Cohabiting couple: Non-dependent children	Pearson Correlation	.456**	-.040	.362**	.213**	.357**	.533**	-.123**	.722**	1	.574**	.626**	-.058	.294**	-.131**
	Sig. (2-tailed)	.000	.298	.000	.000	.000	.000	.001	.000		.000	.000	.132	.000	.001
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
Lone parent: dependent children	Pearson Correlation	.175**	.399**	-.164**	.076*	.193**	.146**	.156**	.615**	.574**	1	.837**	.418**	.381**	.250**
	Sig. (2-tailed)	.000	.000	.000	.048	.000	.000	.000	.000	.000		.000	.000	.000	.000
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
Lone parent: non-dependent children	Pearson Correlation	.297**	.319**	-.033	.063	.230**	.343**	.070	.512**	.626**	.837**	1	.413**	.499**	.308**
	Sig. (2-tailed)	.000	.000	.396	.101	.000	.000	.068	.000	.000	.000		.000	.000	.000
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
Other single family*	Pearson Correlation	-.122**	.784**	-.469**	.185**	-.130**	-.297**	.581**	-.009	-.058	.418**	.413**	1	.163**	.811**
	Sig. (2-tailed)	.001	.000	.000	.000	.001	.000	.000	.820	.132	.000	.000		.000	.000
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
Other with dependent children	Pearson Correlation	-.071	-.185**	-.051	-.182**	.395**	.560**	-.371**	.136**	.294**	.381**	.499**	.163**	1	.040
	Sig. (2-tailed)	.063	.000	.183	.000	.000	.000	.000	.000	.000	.000	.000	.000		.301
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680
All other types	Pearson Correlation	-.097*	.696**	-.401**	.187**	-.133**	-.313**	.717**	-.034	-.131**	.250**	.308**	.811**	.040	1
	Sig. (2-tailed)	.011	.000	.000	.000	.001	.000	.000	.377	.001	.000	.000	.000	.301	
	N	680	680	680	680	680	680	680	680	680	680	680	680	680	680

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Furthermore, in order to measure correlation accurately, it assesses the correlation value. If the correlation value is greater than 0.8, it indicates that the variables have a statistically strong relationship. According to the results, it can be concluded that a total of 4 relationships are approximately over 0.8 (round to one decimal place of the value), indicating that the variables have a statistically significant relationship between them.

Variables
Others single family – One person aged up to 65
Cohabiting couple No children - One person Aged up to 65
One person aged 66+ - Family all aged 66 +
Lone parent non dependent children – Lone parent dependent children



b. Factor analysis

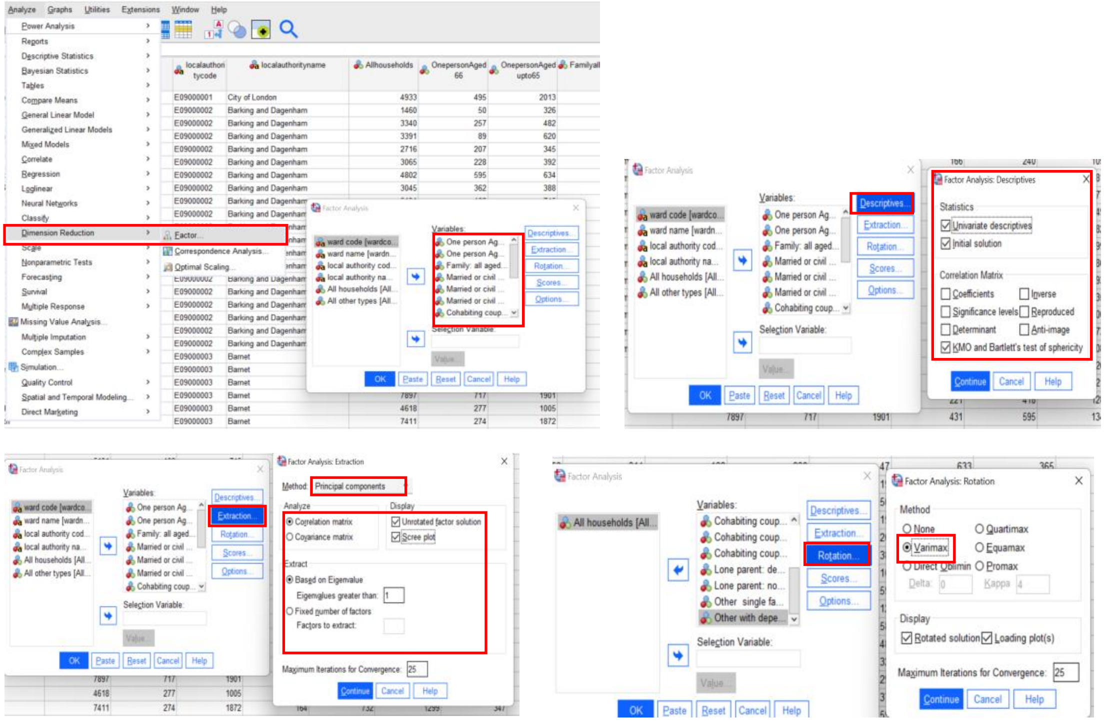
Based on correlation analysis, factor analysis is conducted. It utilized PCA analysis via SPSS.

1) Test steps

PCA test is conducted via SPSS and the testing steps are highlighted as below.

In total, 13 independent variables were added as variables in Factor Analysis. Four categorical variables and one variable "all other types" were excluded, as discussed in the pre-processing stage. Additionally, the variable "All households" was also excluded from PCA analysis as it was set as the dependent variable for the research.

<Step by step process for PCA analysis via SPSS>



Analyze > Dimension Reduction > Factor...

Factor Analysis: Descriptives

Statistics

☑ Univariate descriptives

☑ Initial solution

Correlation Matrix

□ Coefficients □ Inverse

□ Significance levels □ Reproduced

□ Determinant □ Anti-image

☑ KMO and Bartlett's test of sphericity

Factor Analysis: Extraction

Method: Principal components

Analyze: ● Correlation matrix ○ Covariance matrix

Display: ☑ Unrotated factor solution ☑ Scree plot

Extract: ● Based on Eigenvalue — Eigenvalues greater than: 1 ○ Fixed number of factors — Factors to extract:

Maximum Iterations for Convergence: 25

Factor Analysis: Rotation

Method: ○ None ○ Quartimax ● Varimax ○ Equamax ○ Direct Oblimin ○ Promax — Delta: 0 Kappa 4

Display: ☑ Rotated solution ☑ Loading plot(s)

Maximum Iterations for Convergence: 25

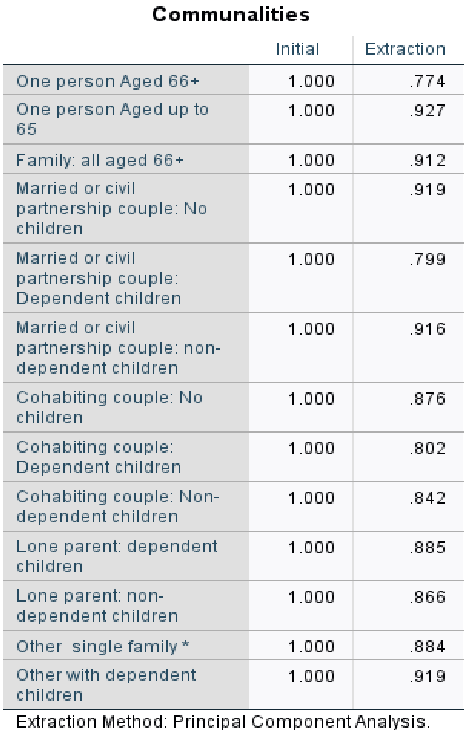
2) Test result

Firstly, according to KMO and Bartlett’s Test, it indicates that the data is reasonable for factor analysis as the KMO sampling adequacy is greater than 0.5. Also, Bartlett’s test set the Null hypothesis as ‘the test variances (independent variables) of two or more groups are equal’ in significant level of 0.05. The test result shows that P- value is lower than 0.05. Therefore, it rejects null hypothesis in favour of alternative hypothesis. As a result, it concludes that the dataset is statistically reasonable for using factor analysis for data reduction technique.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.726
Bartlett's Test of Sphericity	Approx. Chi-Square	9522.421
	df	78
	Sig.	.000

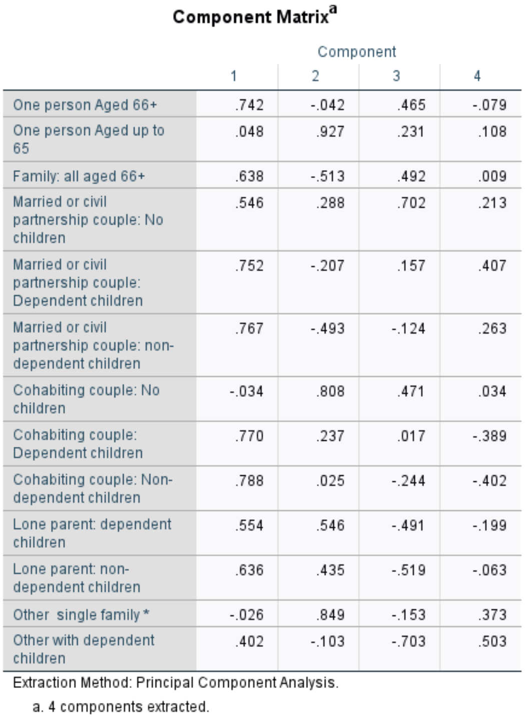
Secondly, following ‘Communalities’ analysis result, there is no extraction value which is lower than 0.4. The threshold for communalities of factor analysis is advised as 0.4 (MacCallum et al, 1999)[9]. Therefore, it concludes that all independent variables are statistically adequate to test factor analysis.



Communalities

	Initial	Extraction
One person Aged 66+	1.000	.774
One person Aged up to 65	1.000	.927
Family: all aged 66+	1.000	.912
Married or civil partnership couple: No children	1.000	.919
Married or civil partnership couple: Dependent children	1.000	.799
Married or civil partnership couple: non-dependent children	1.000	.916
Cohabiting couple: No children	1.000	.876
Cohabiting couple: Dependent children	1.000	.802
Cohabiting couple: Non-dependent children	1.000	.842
Lone parent: dependent children	1.000	.885
Lone parent: non-dependent children	1.000	.866
Other single family *	1.000	.884
Other with dependent children	1.000	.919

Extraction Method: Principal Component Analysis.



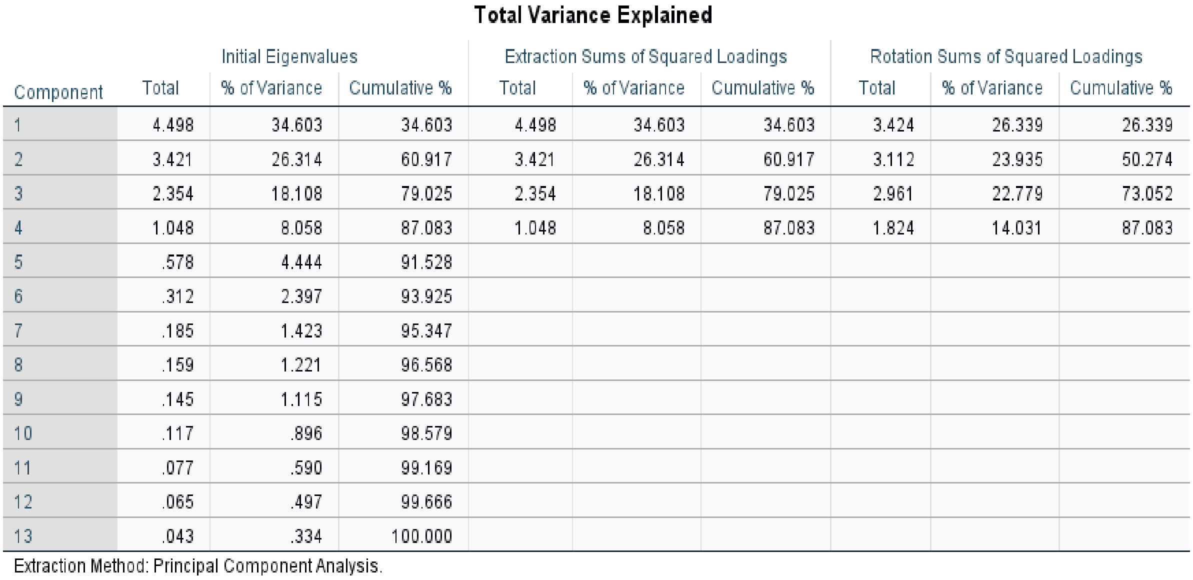
Component Matrix[a]

	Component			
	1	2	3	4
One person Aged 66+	.742	-.042	.465	-.079
One person Aged up to 65	.048	.927	.231	.108
Family: all aged 66+	.638	-.513	.492	.009
Married or civil partnership couple: No children	.546	.288	.702	.213
Married or civil partnership couple: Dependent children	.752	-.207	.157	.407
Married or civil partnership couple: non-dependent children	.767	-.493	-.124	.263
Cohabiting couple: No children	-.034	.808	.471	.034
Cohabiting couple: Dependent children	.770	.237	.017	-.389
Cohabiting couple: Non-dependent children	.788	.025	-.244	-.402
Lone parent: dependent children	.554	.546	-.491	-.199
Lone parent: non-dependent children	.636	.435	-.519	-.063
Other single family *	-.026	.849	-.153	.373
Other with dependent children	.402	-.103	-.703	.503

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Thirdly, the first four principal components have eigenvalues greater than 1. In test result, below four components explain 87% of the variation in the data, according to the "Total Variance Explained" table. These four components (independent variables) are: 1) One person aged 66+ 2) One person aged up to 65 3) Family: all aged 66+ 4) Married or civil partnership couple with no dependent children.

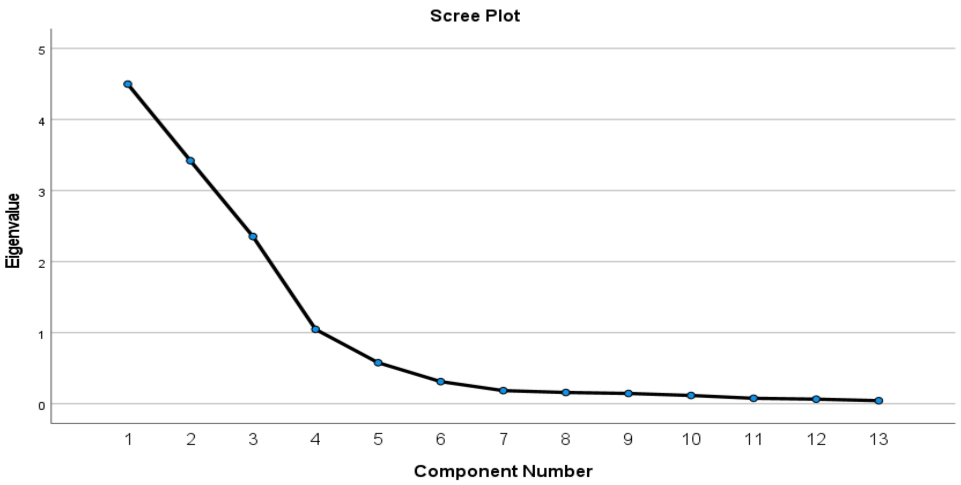


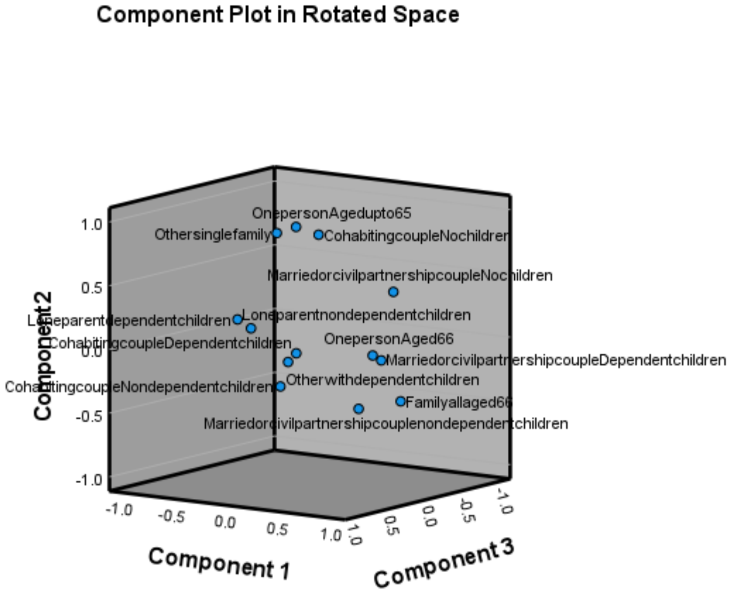
Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.498	34.603	34.603	4.498	34.603	34.603	3.424	26.339	26.339
2	3.421	26.314	60.917	3.421	26.314	60.917	3.112	23.935	50.274
3	2.354	18.108	79.025	2.354	18.108	79.025	2.961	22.779	73.052
4	1.048	8.058	87.083	1.048	8.058	87.083	1.824	14.031	87.083
5	.578	4.444	91.528						
6	.312	2.397	93.925						
7	.185	1.423	95.347						
8	.159	1.221	96.568						
9	.145	1.115	97.683						
10	.117	.896	98.579						
11	.077	.590	99.169						
12	.065	.497	99.666						
13	.043	.334	100.000						

Extraction Method: Principal Component Analysis.

The scree plot also shows that the eigenvalues start to form a straight line after the fourth principal component.





5. Discussion and Analysis of Result

In this research, the question is "How do different types of families impact migration numbers in London?". As a first step, PCA is conducted via SPSS as a linear dimensionality reduction technique to identify meaningful independent variables among the total 13 independent variables. According to the analysis results, four components with eigenvalues greater than 1 are identified. These four components explain 87% of the variation. Secondly, KMO and Bartlett's Test is conducted. The results show that the analysis is statistically significant and justifiable as a factor analysis. The KMO sampling adequacy is greater than 0.5, and the Bartlett's test P-value is less than 0.05 (significant level of 0.05). Therefore, the null hypothesis of Bartlett's test is rejected in favour of the alternative hypothesis and it is concluded that the variances of two or more groups are not equal. In summary, PCA concludes that there are a total of 4 components - 1) One-person aged 66+, 2) One person aged up to 65, 3) Family: all aged 66+, and 4) Married or civil partnership couple: No children - which explain 87% of the total tested independent variables for further investigation. Additionally, this test is statistically reasonable as a factor analysis based on KMO and Bartlett's test.

Further analysis is also suggested on how each family type impacts the number of migrations based on each independent variable's subcategories. For example, the gender differences in single parent or lone parent variables are suggested. Also, analysing subcategories of sexual preference of couples (such as same-sex couples or heterosexual couples) among married or partnership and cohabiting couples.

Part 2. Multiple Linear Regression

1. Introduction and background

In this research, it asks what family type impacts on number of migrations. In order to test this question, Multiple Linear Regression (MLR) test is designed. Multiple linear regression is generally utilized for prediction when examines multiple predictors and its relationship with dependent variables [10]. Additionally, MLR assesses the model fit to identify the amount of error in the model via residuals [11]. Generally, the test results' adjusted R square shows how much the dependent variable can be explained by the independent variables based on this regression model [11]. This model is commonly used for identifying factors that impact the research topic. For example, one research used the MLR model to identify the best fit for air-conditioned building design among 12 key designs [12]. By doing so, the MLR model helps to analyse the best design for managing various climates (from hot summer to cold winter) for maximizing energy efficiency [12].

In this part, it conducts MLR for the dataset which completed PCA in part 1. After PCA test, it identified total four types of family are selected as independent variables as these four variables are available to explain total independent variables (13 independent variables in dataset). Based on previous PCA result, in this analysis, two independent variables are merge into one. The variable 1) One-person aged 66+ and 2) One person aged up to 65 are having relationship. Both are the sub-category of One person who migrated and currently living in London (UK). In order to improve the accuracy of the statistical analysis, these two variables are replaced to one alternative variable: One person (total).

Independent variables
One person (Note. Calculate sum of 'One person aged up to 65' variable and 'One person aged 66+')
Family: all aged 66+
Married or civil partnership couple- no child.

→

Dependent variable
All households (Note. All households indicate total number of migrants in London 2021)

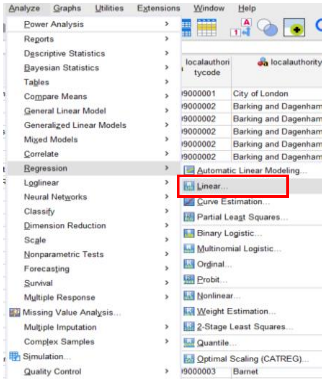
-Research question

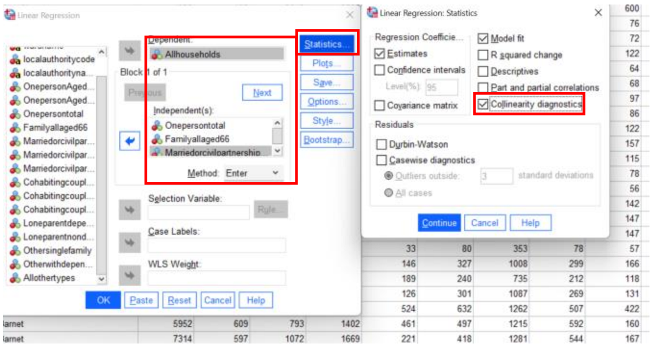
Research question	How three types of family impacts on number of migrations in London, UK. (Three family types are as below.) 1. Consisted with one person, 2. Consisted with family who are all over 66yr 3. Consisted with couple who has no child and has a legal partnership – married or civil partnership

2. Multicollinearity test

a. Test steps

In order to test the collinearity among three independent variables before testing Multiple linear regression model, multicollinearity test is conducted via SPSS. The testing steps are as below.





Dependent: Allhouseholds

Independent(s): Onepersontotal, Familyallaged66, Marriedorcivilpartnership...

Method: Enter

Regression Coefficie...

☑ Estimates

□ Confidence intervals

Level(%): 95

□ Covariance matrix

☑ Model fit

□ R squared change

□ Descriptives

□ Part and partial correlations

☑ Collinearity diagnostics

Residuals

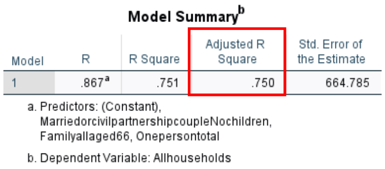
□ Durbin-Watson

□ Casewise diagnostics

● Outliers outside: 3 standard deviations

○ All cases

b. Test result

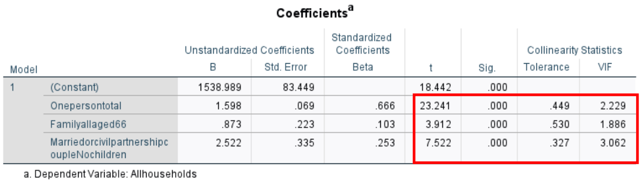


Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.867[a]	.751	.750	664.785

a. Predictors: (Constant), MarriedorcivilpartnershipcoupleNochildren, Familyallaged66, Onepersontotal

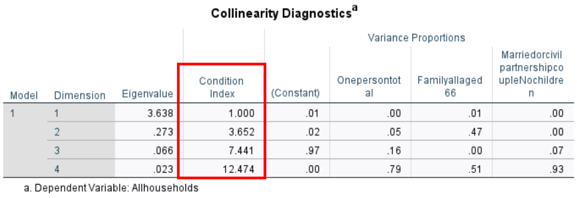
b. Dependent Variable: Allhouseholds



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	1538.989	83.449		18.442	.000		
	Onepersontotal	1.598	.069	.666	23.241	.000	.449	2.229
	Familyallaged66	.873	.223	.103	3.912	.000	.530	1.886
	MarriedorcivilpartnershipcoupleNochildren	2.522	.335	.253	7.522	.000	.327	3.062

a. Dependent Variable: Allhouseholds



Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	Onepersontotal	Familyallaged66	MarriedorcivilpartnershipcoupleNochildren
1	1	3.638	1.000	.01	.00	.01	.00
	2	.273	3.652	.02	.05	.47	.00
	3	.066	7.441	.97	.16	.00	.07
	4	.023	12.474	.00	.79	.51	.93

a. Dependent Variable: Allhouseholds

Firstly, according to model summary, adjusted R square is lower than 0.9 as it is 0.750. Secondly, in ‘Coefficients’ table, significant value is 0.00 for three variables- 1) One person total, 2) family all aged 66 and 3) Married or civil partnership couple no children. Thirdly, collinearity statistics tolerance is all greater than 0.1 and VIF is lower than 5. Lastly, collinearity diagnostics table shows the ‘condition index’ are below 15 for all three variables. Therefore, it can conclude that all three independent variables are sufficient to conclude that they are not having multicollinearity.

3. Multiple linear regression test result

Next, in order to analysis the relationship between independent variables and dependent variables, Multiple Linear Regression (MLR) is tested via SPSS.

a. Hypothesis

Null hypothesis	There is no relationship between independent variables and the dependent variable (all household -migration).	H0 : p =0
Alternative hypothesis	There is relationship between independent variables and the dependent variable (all household -migration).	H1: p≠ 0

The test is also conducted in significant level of 0.05.

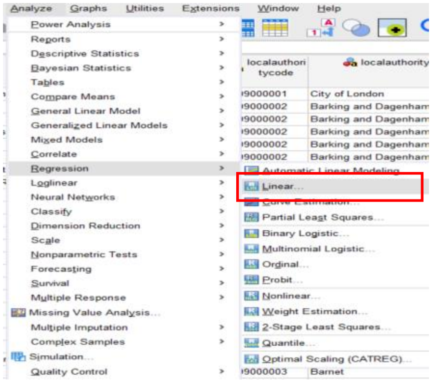
b. Assumption

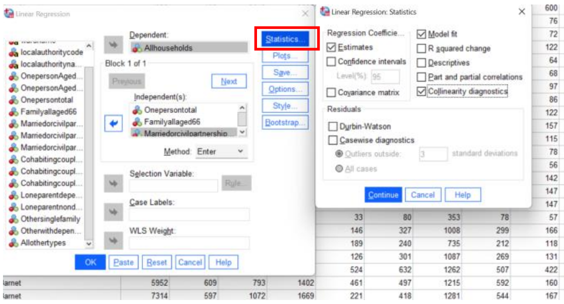
In the multiple linear regression test, it assumed below three conditions.

- Independence of observations
- Normality of errors, meaning that the distribution of residuals is approximately bell-shaped and symmetric.
- No multicollinearity among independent variables
  - This assumption is confirmed in an earlier stage of the paper.

c. Test steps

The multiple linear regression test is conducted via SPSS and details of the testing steps are outlined below.




Linear Regression

Dependent: Allhouseholds

Block 1 of 1

Independent(s): Onepersontotal, Familyallaged66, Marriedorcivilpartnershio...

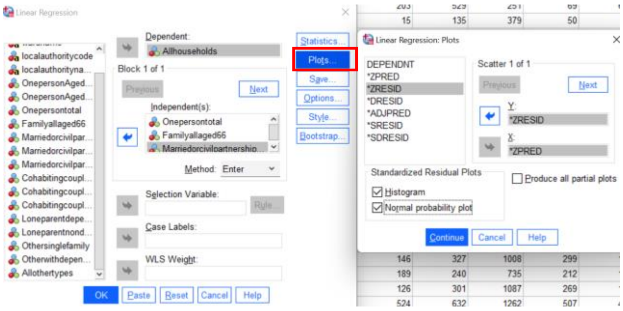
Method: Enter

Linear Regression: Statistics

Regression Coefficients: ☑ Estimates, □ Confidence intervals (Level(%): 95), □ Covariance matrix

☑ Model fit, □ R squared change, □ Descriptives, □ Part and partial correlations, ☑ Collinearity diagnostics

Residuals: □ Durbin-Watson, □ Casewise diagnostics, ● Outliers outside: 3 standard deviations, ○ All cases


Linear Regression

Dependent: Allhouseholds

Block 1 of 1

Independent(s): Onepersontotal, Familyallaged66, Marriedorcivilpartnershio...

Method: Enter

Linear Regression: Plots

DEPENDNT, *ZPRED, *ZRESID, *DRESID, *ADJPRED, *SRESID, *SDRESID

Scatter 1 of 1

Y: *ZRESID

X: *ZPRED

Standardized Residual Plots: ☑ Histogram, ☑ Normal probability plot

□ Produce all partial plots

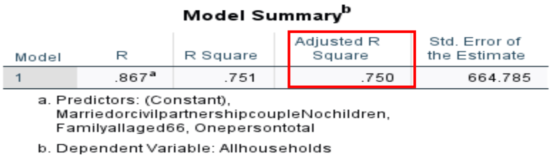
d. Test result

According to the test results, the following confirmations are made:

- The adjusted R-square indicates that the MLR model accounts for 75% of total variance.

- The ANOVA table and coefficients table indicate that the test P-value is 0.000, which is less than 0.05. Therefore, the null hypothesis is rejected in favor of the alternative hypothesis. This conclude that the three independent variables and dependent variable have a statistically significant linear relationship.

- The three graphs- histogram, Normal P-P plot, and scatterplot- indicate that the residuals are normally distributed.

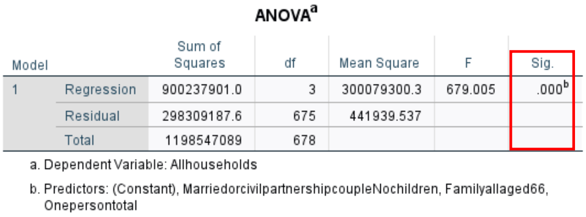


Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.867[a]	.751	.750	664.785

a. Predictors: (Constant), MarriedorcivilpartnershipcoupleNochildren, Familyallaged66, Onepersontotal

b. Dependent Variable: Allhouseholds

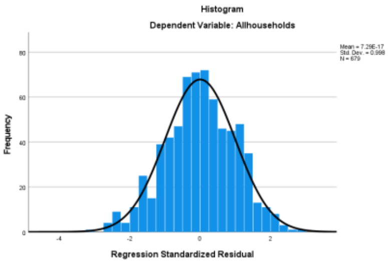


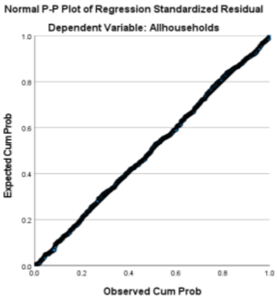
ANOVA[a]

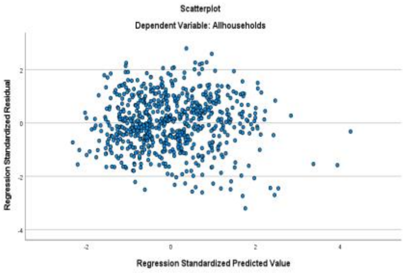
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	900237901.0	3	300079300.3	679.005	.000[b]
	Residual	298309187.6	675	441939.537		
	Total	1198547089	678			

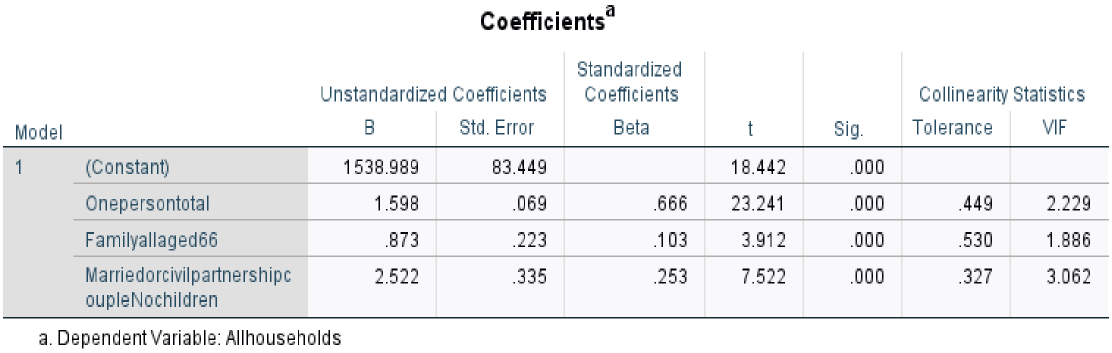
a. Dependent Variable: Allhouseholds

b. Predictors: (Constant), MarriedorcivilpartnershipcoupleNochildren, Familyallaged66, Onepersontotal









Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	1538.989	83.449		18.442	.000		
	Onepersontotal	1.598	.069	.666	23.241	.000	.449	2.229
	Familyallaged66	.873	.223	.103	3.912	.000	.530	1.886
	MarriedorcivilpartnershipcoupleNochildren	2.522	.335	.253	7.522	.000	.327	3.062

a. Dependent Variable: Allhouseholds

Lastly, the prediction is set as below based on test result.

Prediction (Formula)
All household = 1538.989 + 1.598 (One person total) + 0.873 (Family all aged 66+) + 2.522 (Married or civil partnership couple no children)

4. Discussion and Analysis of Result

The Research asks: how do different family types of impact on number of immigrations in London. To test this research question, three selected independent variables (family type: one person, family all aged 66+, married or civil partnership couple with no child) based on PCA were analysed using the MLR model via SPSS. The null hypothesis is set as "There being no statistically significant linear relationship between independent variables and the dependent variable (number of migrations)". The test is conducted in significant level 0.05. According to the test results, the R-square is 75%, indicating that it is a good fit for the MLR model, and the linear regression is statistically significant (coefficients table P-value and ANOVA test table P-value show less than 0.05). Lastly, the PP-plot and scatter plot show that the residuals are normally distributed. Therefore, it is concluded that there is a statistically significant relationship between the three mentioned family types and migration number, and the prediction formula is as follows:

Number of migration (all household type migration) = 1538.989 + 1.598 (family type: one person) + 0.873 (Family type: family members are all aged 66+) + 2.522 (family type: Married or civil partnership couple who has no children).

References

[1] R. J. Hyndman and G. Shmueli, "Forecasting: principles and practice," [Online]. Available: http://www.forecastingbook.com. [Accessed: 01-Jan-2023].

[2] S. Glen, "ARIMA (Box-Jenkins Models): Autoregressive Integrated Moving Average," [Online]. Available: https://www.statisticshowto.com/arima/. [Accessed: 01-Jan-2023].

[3] S.Tyagi, "Introduction to Time Series Forecasting Part 1: Average and Smoothing Models,", Towards Data Science, [Online] Available: https://towardsdatascience.com/introduction-to-time-series-forecasting-part-1-average-and-smoothing-models-a739d832315 [Accessed: 02-Jan-2023]

[4] UK Parliament. (2021). Migration statistics, [Online]. Available: https://researchbriefings.files.parliament.uk/documents/SN06077/SN06077.pdf

[5] I.T. Jolliffe, "Principal component analysis for special types of data," in Principal component analysis, Springer New York, 2002, pp. 338-372

[6] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," The London, Edinburgh, and Dublin philosophical magazine and journal of science, vol. 2, no. 11, pp. 559-572, 1901

[7] J. Shlens, "A Tutorial on Principal Component Analysis," arXiv:1404.1100 [cs], 2014.

[8] "2021 census - wards - demography and migration," data.london.gov.uk, [Online]. Available: https://data.london.gov.uk/dataset/2021-census-wards-demography-and-migration. [Accessed: 30-Dec-2022].

[9] MacCallum RC, Widaman KF, Zhang S and Hong S. (1999) Sample size in factor analysis. Psychological Methods 4(1) 84-99.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with application R, New York: Springer, 2013.

[11] A. Field, Discovering Statistics Using IBM SPSS Statistics, London: Sage Publications, 2013.

[12] M.C. Lu and C.F. Kao, "Multiple linear regression analysis of factors affecting the energy consumption of air conditioning systems," Energy and Buildings, vol. 42, pp. 1845-1852, 2010.

Data source

https://data.london.gov.uk/dataset/2021-census-wards-demography-and-migration