TASK A

PURPOSE OF THE ANALYSIS

The analysis aims to explore the connections among key demographic factors while demonstrating that the value for money of properties can be effectively modelled through linear relationships using specific components from the survey questionnaire administered to first-time house buyers.

METHOD AND ASSUMPTIONS USED FOR ANALYSIS

Two main statistical techniques, Correlation and Regression, were employed to explore the connection between demographic factors or variables and the linear model for 'value.'

Demographic variables (Age, Income, Advertised, Sex, and Gender) were briefly examined pre-analysis. Outliers and irrelevant data were identified and removed using box plots, and invalid values such as negative ages and values over 100 were excluded.

Assumptions relied on central tendency measures. Variables mostly deviated from normal distribution; 'Advertised' was treated as ranked. Hence, a 5% significance level non-parametric Spearman's correlation was chosen for analysis.

Regarding Regression, the variables 'Location,' 'Travel Expense,' and 'Cost of Living' were chosen to model the Value for the money for property. Assumptions included independent responses for predictors, and the residuals were assumed to be normal. To achieve this, a square root transformation was applied to the skewed dependent variable 'Value' resulting in stabilised variance.

RESULTS

CORRELATION ANALYSIS

To determine the non-parametric correlation between the continuous variables of interest under Demographics (Income, Age and Advertised) the following Null and Alternate Hypothesis were stated at 5% level of significance.

H_0: There exists no relationship between variables.
H_1: There exists a relationship between variables.

Spearman Correlation Coefficients, N = 786 Prob > |r| under H0: Rho=0			
	Age	**Income**	**Advertised**
Age **Age**	1.00000	0.02030 0.5699	-0.05260 0.1407
Income **Income**	0.02030 0.5699	1.00000	-0.01946 0.5859
Advertised **Advertised**	-0.05260 0.1407	-0.01946 0.5859	1.00000

Table.1a - Correlation matrix between variables.

The output for the correlation analysis is detailed in Figure 1a above, it was observed that the correlation coefficient at a 5% level of significance between Age and Income is 0.02030 with p-value of 0.5699. That of Age and Advertised resulted in a negative coefficient of correlation of -0.05260 with p-value 0.1407. Also, Income and advertised had a negative correlation coefficient of 0.01946 with p-value 0.5859.

REGRESSION ANALYSIS

The following Hypotheses were stated for the 'Value' to be modelled linearly.

$H_0 : \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$

$H_1 :$ *At least one of the* β *parameters in* H_0 *is non-zero.*

Number of Observations Read	762
Number of Observations Used	762

Table. 2a

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	875.34852	875.34852	2977.97	<.0001
Error	760	223.39546	0.29394		
Corrected Total	761	1098.74398			

Table. 3b

Root MSE	0.54216	**R-Square**	0.7967
Dependent Mean	6.39761	**Adj R-Sq**	0.7964
Coeff Var	8.47447		

Table. 4c

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	3.10364	0.06348	48.89	<.0001
Cost Of Living	Cost Of Living	1	0.05940	0.00109	54.57	<.0001

Table. 5d

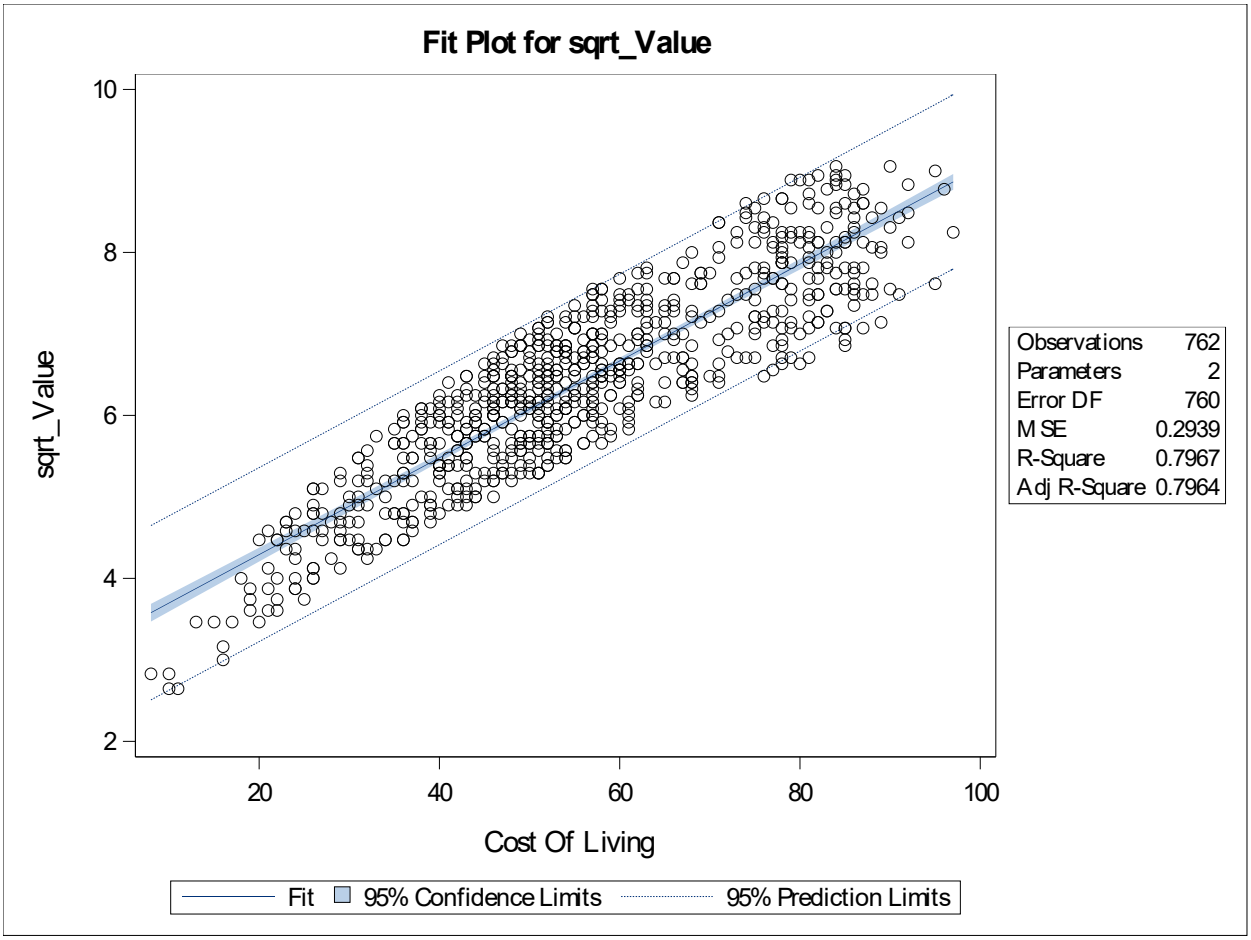


Figure. 1a

Durbin-Watson D	1.972
Number of Observations	762
1st Order Autocorrelation	0.014

Figure. 2a

To assess whether the independent variables in our model effectively explain the dependent variable, the analysis revealed a p-value of 0.0001(from Table 3b), indicating strong evidence that the model is fit for further analysis. The degree of freedom was 1(from Table 3b), representing the number of independent variables considered in the analysis.

To check the importance of the predictors, the p-values for the intercept and cost of living were both less than 0.0001. To evaluate how well the model fits, we use R^2 and adjusted R^2, which indicate how much the independent variable explains changes in the dependent variable. For our model, we recorded values of 0.7967 and 0.7964 (See Table 4c) as the coefficient of determinations.

To check if errors are independent, we used the Durbin-Watson test for correlated residuals, and it gave a value of 1.972(See Fig.2a).

The linear model for value estimate was obtained as follows using outputs from Table 5d.

Value = 3.10364 + 0.05940(Cost of living).

CONCLUSION

From the results of our analysis above, it is evident that the correlation coefficients show weak relationships between variables. Additionally, the correlation analysis results revealed non-significant p-values (p-value > 0.05) for any two variables. Consequently, we fail to reject the null hypothesis, indicating that there is no significant relationship between the variables.

In conclusion, the regression model, supported by a highly significant p-value of 0.0001, rejects the null hypothesis in favour of the alternative, which asserts the presence of at least one non-zero β coefficient. The impactful regressors are the intercept and the cost of living, each exhibiting a noteworthy p-value of 0.0001. The model's goodness of fit is reflected in a 79.67% variance in Value explained by the cost of living. The Durbin-Watson statistics suggest reasonably independent errors within the range of 1-3. With only the cost of living and intercept as factors, the parameter estimates table yields the following interpretation: a cost of living coefficient of 0.05940 (See Table 5d) signifies that a £1 increase in living costs corresponds to a 0.05940 increase in value for money of properties. Additionally, the linear model expresses the value of money for property with no cost of living as £3.10364 (See Table 5d). Therefore, the model, while unable to

detect an influence from location and travel expenses on Value, indicates a relatively low but discernible impact of Cost of living on Value, as captured by the formula: Value = 3.10364 + 0.05940(Cost of living).

TASK B

PURPOSE OF THE ANALYSIS

To create a new set of variables from a complex multivariate dataset to uncover patterns and relationships through transformation and explore differences between regions using the newly constructed dataset.

METHOD AND ASSUMPTIONS USED FOR ANALYSIS

After cleaning 798 initial questionnaire observations to 762, Principal Component Analysis (PCA) reduced 14 variables, yielding interpretable factors. The Scree Plot guided factor selection.

Employing Varimax rotation with 4 components addressed cross-loading issues. This orthogonal rotation enhanced result interpretability by maximizing high correlations and minimizing low ones.

Utilizing the new factors, a non-parametric one-way ANOVA that is the Kruskal-Wallis test was conducted to compare newly constructed factor scores across regions.

RESULTS

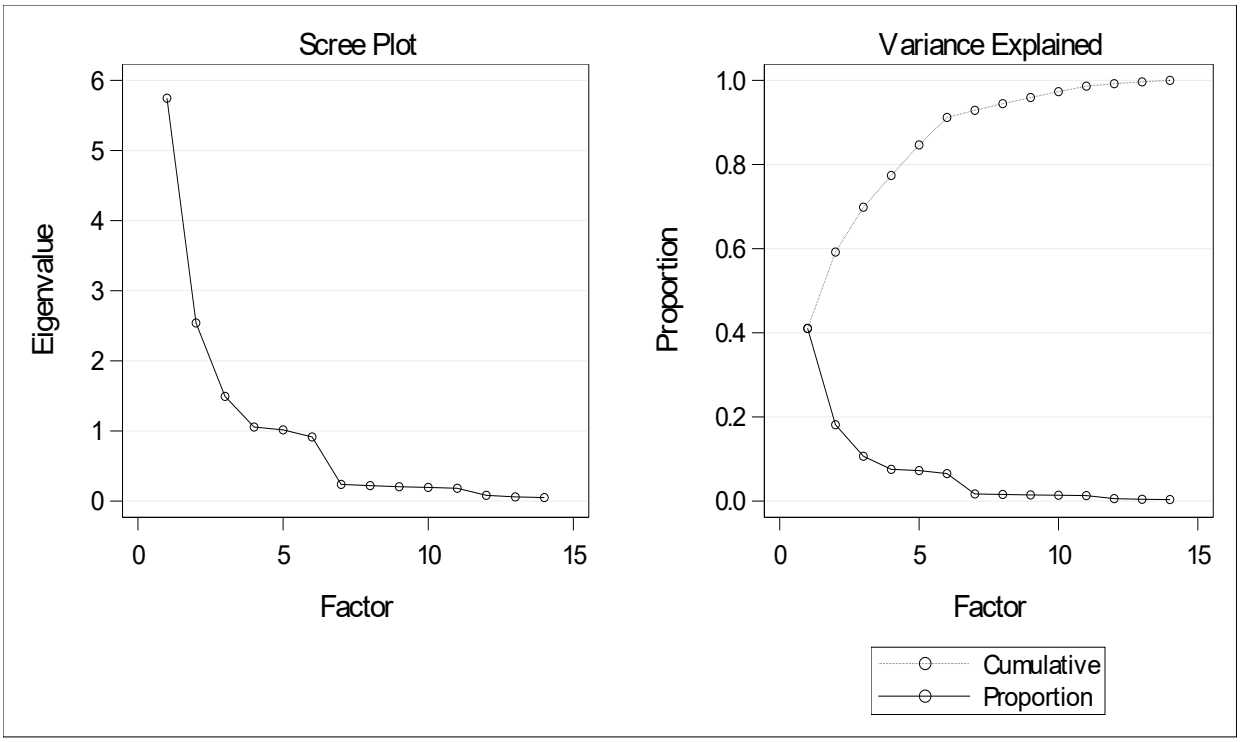
PCA OUTPUT GENERATED

Input Data Type	Raw Data
Number of Records Read	762
Number of Records Used	762
N for Significance Tests	762

Table 1b

Eigenvalues of the Correlation Matrix: Total = 14 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	5.74598172	3.20553707	0.4104	0.4104
2	2.54044465	1.04736166	0.1815	0.5919
3	1.49308300	0.43655803	0.1066	0.6985
4	1.05652496	0.04070276	0.0755	0.7740
5	1.01582220	0.10110242	0.0726	0.8466
6	0.91471978	0.67685514	0.0653	0.9119
7	0.23786464	0.01749710	0.0170	0.9289
8	0.22036755	0.01564685	0.0157	0.9446
9	0.20472070	0.00971621	0.0146	0.9593
10	0.19500448	0.01214055	0.0139	0.9732
11	0.18286393	0.10066185	0.0131	0.9862
12	0.08220208	0.02235744	0.0059	0.9921
13	0.05984464	0.00928896	0.0043	0.9964
14	0.05055568		0.0036	1.0000

Table 2b.


Scree Plot
Eigenvalue
Factor
Variance Explained
Proportion
Factor
Cumulative
Proportion
Figure 1b.

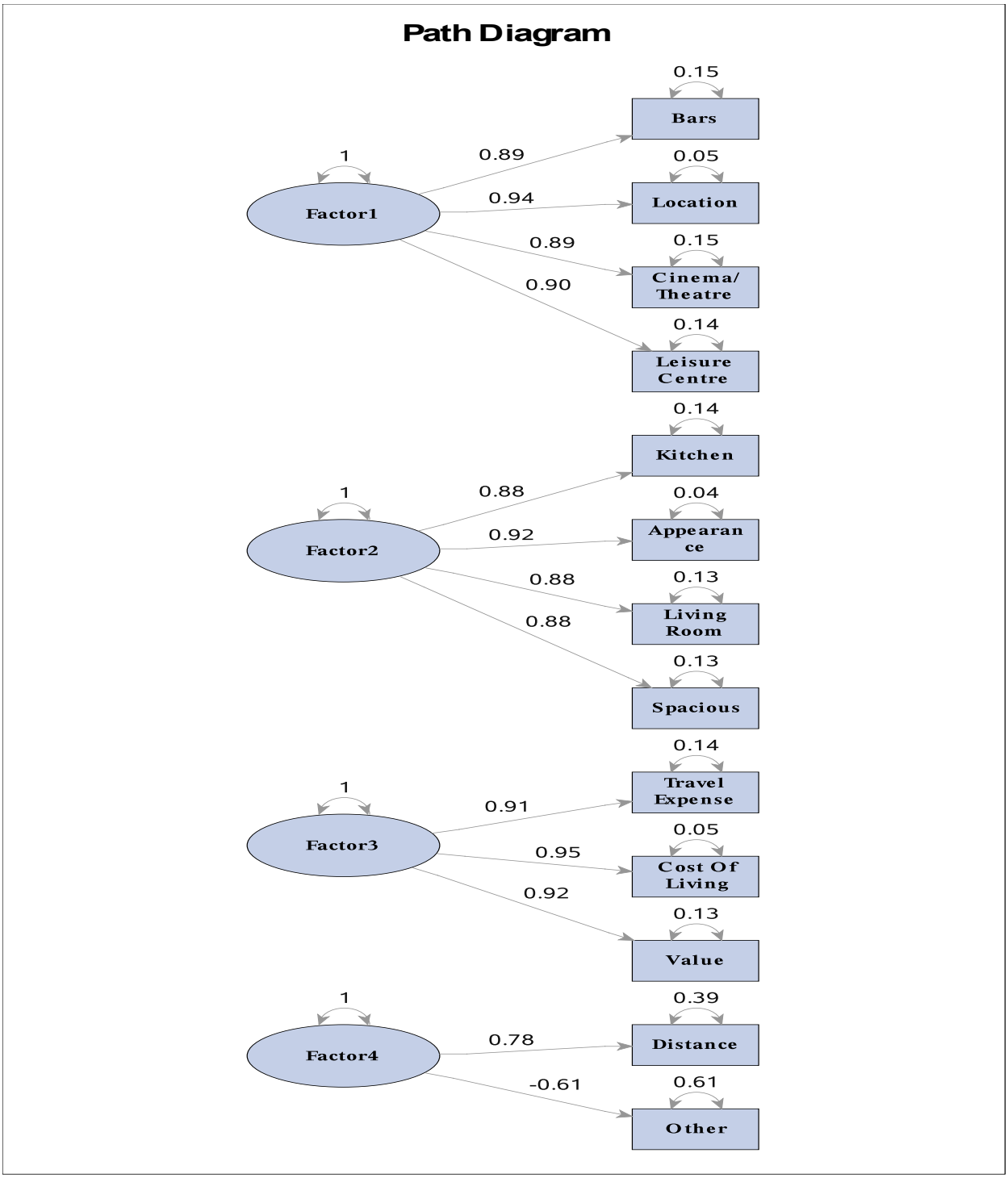

Path Diagram
Factor1
Factor2
Factor3
Factor4
Bars
Location
Cinema/ Theatre
Leisure Centre
Kitchen
Appearance
Living Room
Spacious
Travel Expense
Cost Of Living
Value
Distance
Other
1
0.89
0.94
0.89
0.90
0.88
0.92
0.88
0.88
0.91
0.95
0.92
0.78
-0.61
0.15
0.05
0.15
0.14
0.14
0.04
0.13
0.13
0.14
0.05
0.13
0.39
0.61

Figure 2b.

Table 2b eigenvalues and Fig 1b, Scree Plot were used to determine the number of components. Factors 1-5 in the table had eigenvalues exceeding one, and the scree plot indicated 4 or 5 components. Choosing 4 components based on eigenvalues, scree plot, and proportion of variance retained 77.40%. The explained variance for each of the 4 factors is detailed below, contributing to a cumulative variance of 77.40%.

In exploring latent variables influencing common variance and understanding variable interrelationships, factor analysis was employed. After PCA determined the number of factors, iterated principal factor analysis sought the minimum factors explaining common variance measured by original variable communalities.

However, the test revealed an Ultra-Heywood case, where final communalities exceeded 1 for the 4 factors from PCA. This suggests potential negative variance or excessive common factors, invalidating the solution. Therefore, the common factor model is inappropriate for the data.

ANOVA OUTPUT

Considering a significance level of 5% the hypothesis for the ANOVA test can be stated as:

Let μ_i be the mean score of each factor from region.
Let k= Kenston, n= Northton, f= Fernham and s= Southware.

$$H_0: \mu_k = \mu_n = \mu_f = \mu_s$$
$$H_1: at\ least\ one\ \mu_i\ not\ equal$$

Kruskal-Wallis Test		
Chi-Square	**DF**	**Pr > ChiSq**
2.4736	3	0.4801

Table 2c. (Factor 4)

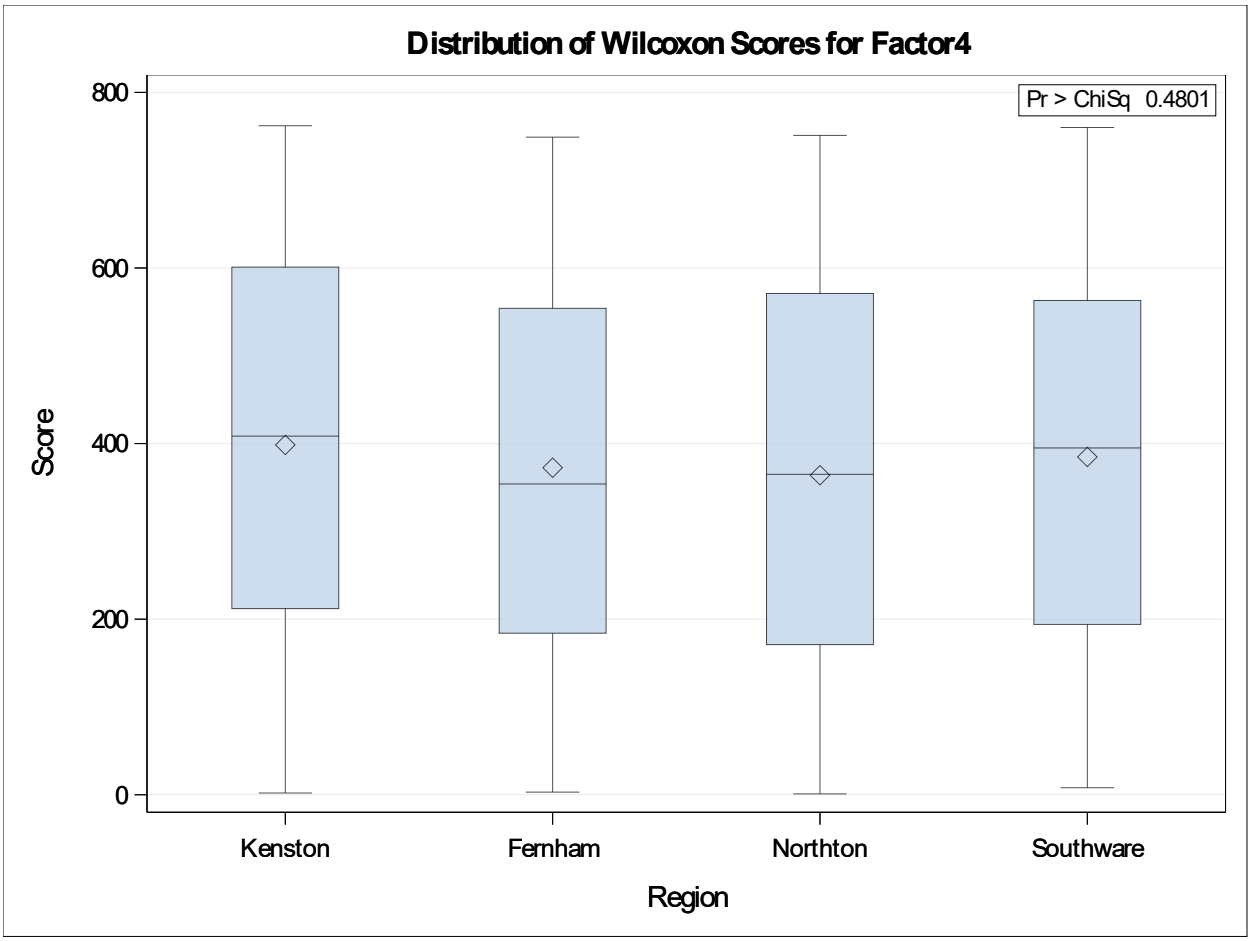


Figure 2c.

The Kruskal-Wallis test gave p-values on the analysis for factor 1, 2 and 3 as 0.001.
But for factor 4, the p-value is 0.4801(See Table 2e).
Each test is accompanied by a box plot to confirm if differences exist between the different regions.(see Fig 2c).

CONCLUSION

With p-values for factors 1-3 being less than 0.05, the null hypothesis is rejected which indicates that at least one of μ_i is not equal, hence there exist differences across regions for factors 1-3.

For factor 4, the p-value of 0.4801 which is greater than 0.05, the null hypothesis is failed to be rejected, confirming that no difference between the regions regarding factor 4.

TASK C

PURPOSE OF THE ANALYSIS

Identifying and categorizing employees with similar motivation based on key factors such as job, pay, hours, and teamwork. Subsequently, examining whether variations exist in these key features among the newly established groups.

METHOD AND ASSUMPTIONS USED FOR ANALYSIS

Clustering analysis was conducted to form distinct groups with similar characteristics, shedding light on the pharmaceutical company's workforce. Utilizing an agglomerative clustering approach, specifically the ward minimum-variance method in SAS, variables such as job satisfaction, pay scale, working hours, and collaboration in teams were measured at an ordinal level. The dissimilarity measure employed was the Euclidean measure.

To explore potential differences among these various factors grouped by clusters, a non-parametric one-way ANOVA was applied, recognizing the ranked/scoring nature of the data. A significance level of 0.05 was used to identify statistically significant differences between the groups.

RESULTS

To determine the appropriate number of clusters for the study, we examined the Cubic Clustering Criterion (CCC), Pseudo t-squared, Pseudo F statistic, and dendrogram.

Cluster History										
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Tie
29	CL99	CL96	6	0.0015	.963	.	.	92.6	12.3	
28	CL72	OB75	5	0.0015	.961	.	.	93.0	8.9	
27	CL36	CL75	10	0.0015	.960	.	.	93.6	4.9	
26	CL44	CL62	7	0.0017	.958	.	.	94.3	4.8	
25	CL80	CL32	8	0.0017	.956	.936	6.75	95.2	3.4	
24	CL81	CL66	7	0.0017	.955	.933	6.78	96.3	11.7	
23	CL60	CL37	7	0.0018	.953	.930	6.85	97.6	4.3	
22	CL28	CL40	10	0.0019	.951	.928	6.92	99.1	4.2	
21	CL29	CL41	13	0.0024	.949	.924	6.84	99.9	6.4	
20	CL35	CL47	9	0.0027	.946	.921	6.71	100	7.4	
19	CL30	CL77	7	0.0028	.943	.917	6.65	102	6.1	
18	CL38	CL50	10	0.0029	.940	.914	6.59	103	8.8	
17	CL51	CL24	12	0.0030	.937	.910	6.60	105	8.2	
16	CL31	CL33	12	0.0033	.934	.905	6.58	107	6.8	
15	CL39	CL26	13	0.0039	.930	.900	6.50	108	8.2	
14	CL23	CL49	10	0.0054	.925	.895	6.14	109	9.9	
13	CL27	CL20	19	0.0056	.919	.889	5.87	110	10.2	
12	CL15	CL42	18	0.0056	.914	.882	5.74	112	8.5	
11	CL12	CL45	23	0.0073	.906	.875	5.43	114	8.6	
10	CL18	CL34	15	0.0077	.899	.867	5.23	117	13.3	
9	CL25	CL17	20	0.0088	.890	.857	5.04	121	14.1	
8	CL11	CL19	30	0.0106	.879	.846	4.80	126	9.9	
7	CL16	CL22	22	0.0119	.867	.832	4.69	133	17.6	
6	CL8	CL14	40	0.0140	.853	.816	4.04	143	10.6	
5	CL7	CL21	35	0.0144	.839	.794	4.48	161	14.8	
4	CL6	CL13	59	0.0156	.823	.764	4.51	194	11.2	
3	CL9	CL10	35	0.0169	.806	.716	6.49	262	15.8	
2	CL3	CL5	70	0.0600	.746	.601	6.67	374	41.3	
1	CL4	CL2	129	0.7463	.000	.000	0.00	.	374	

Table 3a.

From Table 3a. The pseudo t-squared recorded 41.3 for cluster 2 but recorded 15.8 for cluster 3 indicating a high increment from cluster 3 to 2. The pseudo-F statistic recorded a high value of 262 for cluster 3.

The Cubic Clustering Criterion recorded a value of 6.49 which is above the threshold of 2.

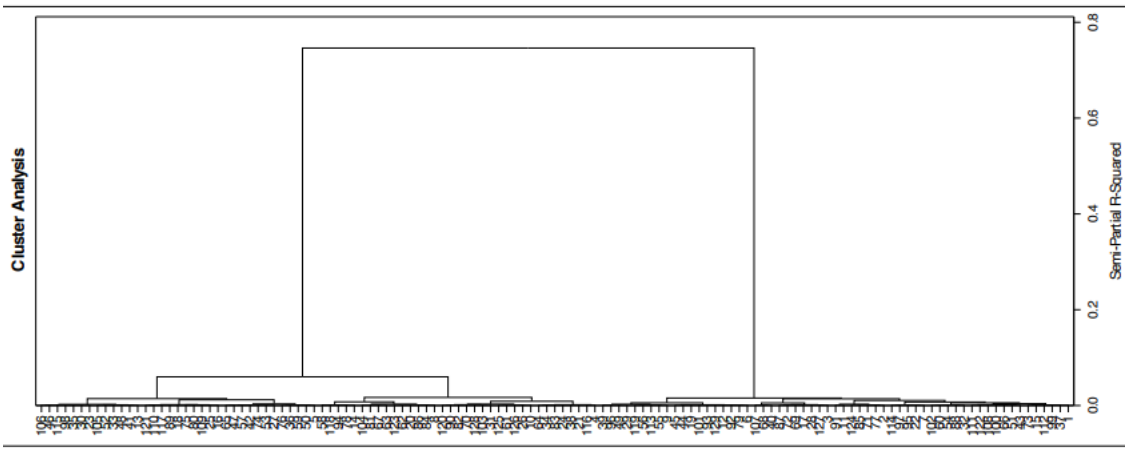


Figure 3a.

The dendrogram (Figure 3a) generated from the cluster observation analysis suggests that 3 clusters are optimal for grouping employees. This conclusion is affirmed by the initial choice of 3 clusters in k-means clustering, yielding a pseudo F statistic of 1212.36 and a CCC of 81.89. Conversely, opting for 4 clusters resulted in values of 1033.39 for Cubic Clustering Criterion and a pseudo-F statistic of 78.62, indicating that the ideal number of clusters is 3.

The Kruskal-Wallis test, a non-parametric one-way ANOVA, was performed to assess whether differences existed in the variables among the newly formed clusters.

The hypothesis for each of the variables for the study was:
H_0: There exists no difference between a variable by cluster.
H_1: There exists a difference between a variable by cluster.

Examining each box plot and conducting the Kruskal-Wallis test for the four variables, it is observed that the p-values for each variable are below 0.05. This suggests rejecting the null hypothesis in favour of the alternative hypothesis stating there exist differences in the scores of variables for clusters. (See Figures. 3b).

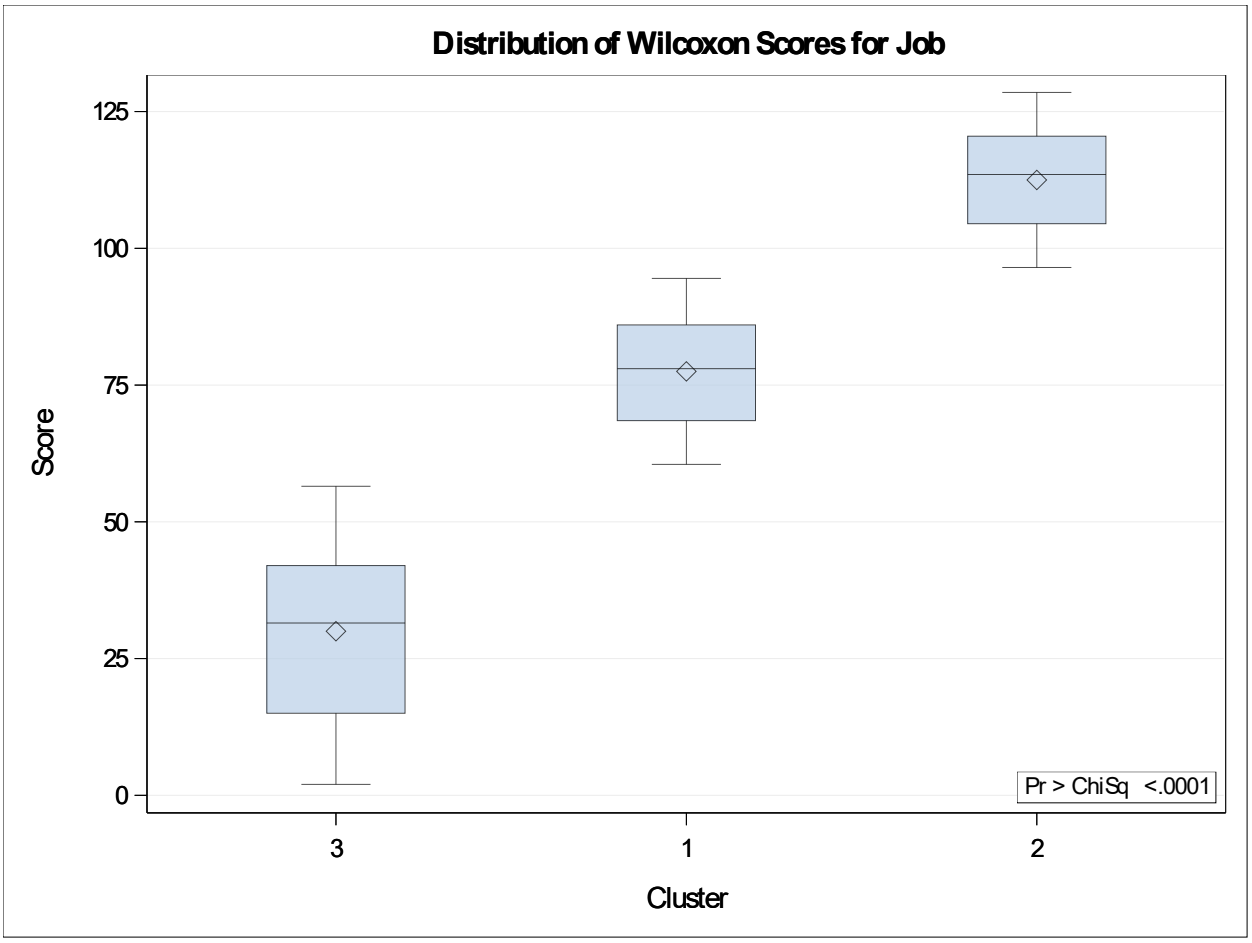


Figure 3b (Box plot showing mean scores across clusters for variable, Job)

CONCLUSION

In summary, we can conclude that the number of clusters chosen to group employees of similar motivation is 3, and there exists a difference across clusters under the variables of study.