Introduction

The objective of the research is to thoroughly examine a dataset from 2020 that contains information on employment and salaries across different industries and locations. The dataset was refined using stratified sampling to ensure a balanced representation of various job roles, industries, and geographical areas (Thomas, 2022). This dataset was chosen because it includes a diverse range of data types, such as categorical, nominal, and continuous variables, which allows for a comprehensive multidimensional analysis.

The report is structured as follows: it begins with a detailed data dictionary that explains each variable. Next, data exploration is conducted using descriptive statistics and visualisations to identify important trends and patterns. The analysis then includes t-tests to identify differences among specific groups, factor analysis to uncover underlying variables, and cluster analysis to effectively group employees. The results of this analysis aim to provide insights into the factors that influence employee satisfaction and job performance.

1. Creating a data dictionary of what the variables (.i.e., columns) represent.

Table 1: Data Dictionary Overview

Number	Variable Name	Description	Data Type	Scale/Encoding
1	**Work Year**	The year in which the salary data was recorded	Numeric	YYYY
2	**Age**	The age of the employee	Numeric	Years
3	**Experience Level**	The level of experience of the employee	Ordinal	Entry, Intermediate, Senior, Expert
4	**Employment Type**	The type of employment contract	Categorical	Full-time, Part-time, Contract, Freelance
5	**Job Title**	The title of the job held by the employee	Categorical	Various job titles
6	**Salary**	The annual salary of the employee in local currency	Numeric	Currency units
7	**Salary Currency**	The currency in which the salary is paid	Categorical	Currency codes (e.g., USD, EUR)
8	**Salary in USD**	The annual salary of the employee converted to US dollars	Numeric	USD
9	**Employee Residence**	The country of residence of the employee	Categorical	Country codes or names
10	**Company Location**	The location of the company	Categorical	Country codes or names
11	**Remote Ratio**	The percentage of work conducted remotely	Numeric	0-100 (%)
12	**Company Size**	The size of the company where the employee works	Ordinal	Small, Medium, Large
13	**Performance Rating**	The performance rating of the employee	Ordinal- Likert Scale	1-5 scale
14	**Work Life Balance**	The work-life balance rating given by the employee	Ordinal-Likert Scale	1-5 scale
15	**Remote Work Effectiveness**	Rating of effectiveness of working remotely	Ordinal-Likert Scale	1-5 scale (1-Very Ineffective to 5-Very Effective)
16	**Job Satisfaction**	Employee's overall job satisfaction	Ordinal-Likert Scale	1-5 scale (1-Very Dissatisfied to 5-Very Satisfied)
17	**Professional Development**	Rating of professional development opportunities	Ordinal-Likert Scale	1-5 scale (1-Very Poor to 5-Excellent)
18	**Engagement Score**	Level of employee's engagement with their work and the organisation	Ordinal-Likert Scale	1-5 scale (1-Very Low to 5-Very High)

The data dictionary presents a comprehensive overview of the dataset (Atlan, 2022), encompassing intricate details regarding employment and salary information across various industries as of the year 2020. This dataset comprises a combination of categorical and numerical data types. Categorical data, such as employment type, company location, and employee residence, serve as valuable tools for investigating the potential influence of these factors on job satisfaction and career opportunities. On the other hand, ordinal data includes scales for performance ratings and job satisfaction, enabling the assessment of employee engagement levels and overall contentment with work-life balance (Bhandari, 2022b). Furthermore, numerical data, including age, salary, and remote work ratio, offer quantitative insights into demographic trends, compensation levels, and remote working practices. By analysing this data, organisations can gain a deeper understanding of the factors that

impact employee satisfaction and performance, ultimately facilitating the development of more effective workplace policies.

2. Presenting the measures of central tendency and variation in tables and providing interpretations for each of tables.

Table 2: Descriptive Statistics for Numerical Data

Descriptive Measures	Age	Salary in local currency	Salary in USD
Mean	37.23	548221.08	112404.65
Standard Error	0.65	168462.24	6223.67
Median	35.00	115000.00	89487.00
Mode	30.00	100000.00	100000.00
Standard Deviation	9.65	2487315.12	91891.45
Sample Variance	93.07	6186736499688.81	8444038962.85
Kurtosis	-0.46	99.22	5.00
Skewness	0.46	9.10	1.82
Range	40.00	30396000.00	597141.00
Minimum	19.00	4000.00	2859.00
Maximum	59.00	30400000.00	600000.00
Sum	8116.00	119512196.00	24504213.00
Count	218.00	218.00	218.00

Age: The data indicates that the average age of individuals in the dataset is 37.23 years, with a standard deviation of 9.65 years. This suggests that the age distribution is somewhat close to the mean. The age range of 40 years, spanning from 19 to 59, implies that there is a mix of younger and older employees in the workforce (Bhandari, 2022a).

Salary in local currency: The average salary in local currency is 548,221.08, which is a significant figure. However, the standard deviation of 2,487,315.12 and a wide range of 30,396,000 indicate substantial disparities in income levels. This is further evidenced by the sample variance of around 6.19 trillion, highlighting a notably uneven distribution of salaries (What Is Variance in Statistics? Definition, Formula, and Example, n.d.). The average salary is greatly influenced by a high skewness of 9.10, which suggests a long tail of high earners in the salary distribution.

Salary in USD: The mean salary of 112,404.65 USD and a standard deviation of 91,891.45 USD indicate that there is less variation in salaries compared to those in the local currency. However, the range of 597,141 USD still shows a significant difference in earnings when using a standardized monetary unit. The skewness value of 1.82 indicates that most employees earn less than the average salary, highlighting a preference for lower salary brackets in USD amounts.

The salary data shows a wide range of variability, with some individuals earning much higher salaries than the majority. This is evident from the high kurtosis values (99.22 for local currency and 5.00 for USD) and the fact that mean salaries are significantly higher than the medians in both local currency and USD (115,000 for local currency and 89,487 for USD). However, when salaries are converted to USD, the disparity in earnings is less pronounced, possibly due to exchange rate effects. In contrast, the age distribution appears to be more evenly spread, with lower kurtosis (-0.46) and a near-zero skewness (0.46), suggesting that age may not have a significant impact on salary differences. These findings have implications for pay scale strategies and highlight the importance of addressing potential salary disparities within the industries being studied.

3. Conducting data visualisation on variables and providing interpretations for each of the plots.

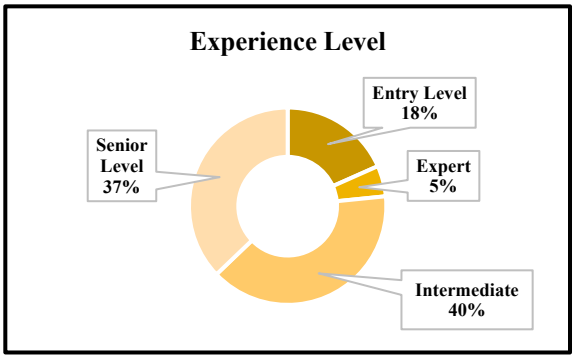

Figure 1: Distribution of Workforce by Experience Level

In Figure 1, the donut chart shows that 40% of the workforce is at Intermediate level, with Senior level close behind at 37%. This indicates a significant number of experienced employees. Entry-level workers make up 18%, while Experts represent 5%, creating a pyramid-like distribution of expertise.

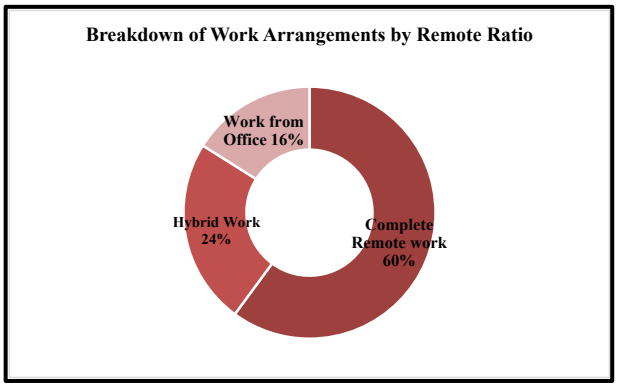

Figure 2: Distribution of Work Arrangements Among Employees

The donut chart in Figure 2 shows that the vast majority, 60%, are working entirely remotely, 24% in a hybrid setting, and just 16% are working from the office, indicating a strong trend towards remote work practices.

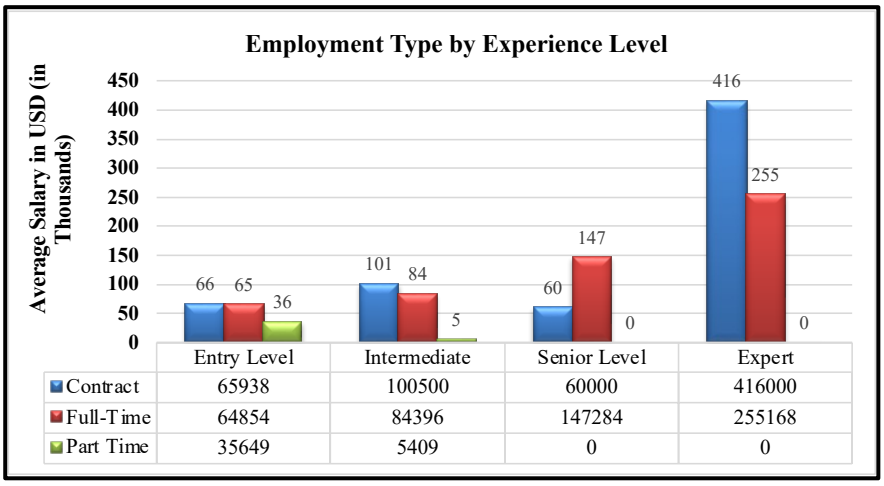

Figure 3: Average Salary by Employment Type Across Experience Levels

The bar graph in Figure 3 shows that contract employees at the Senior level have surprisingly lower average salaries at 60,000 USD compared to their Full-Time counterparts who earn 147,284 USD. In contrast, Expert

level contract employees top the salary scale at 416,000 USD, far exceeding the Full-Time experts' average of 255,168 USD.

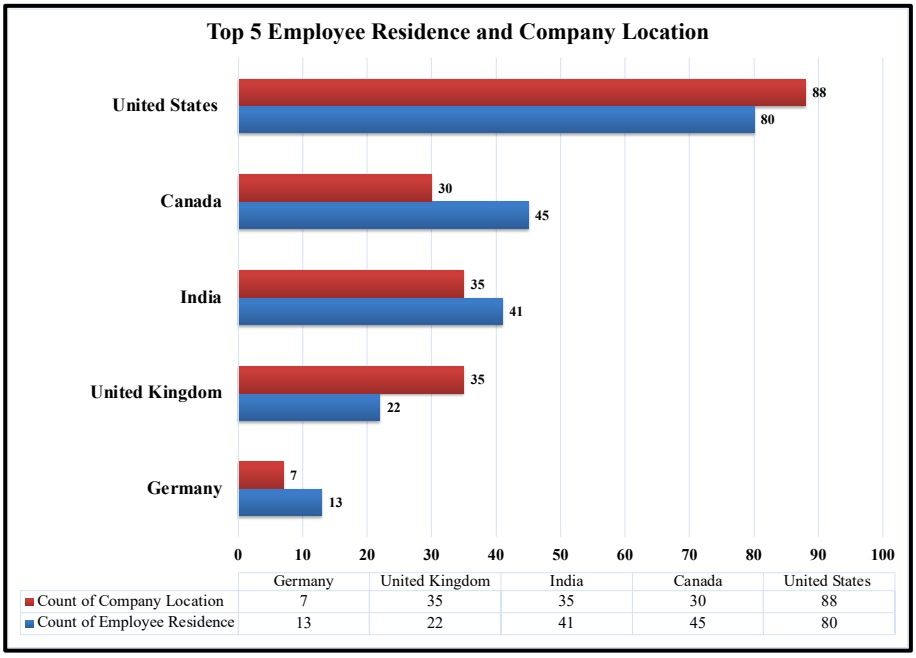

Figure 4: Comparative Analysis of Top 5 Employee Residences and Company Locations

The bar chart presents a comparison of employee residence versus company location for the top five countries. The United States leads with the highest number of companies (88) and employee residences (80), indicating it as a major hub for both business operations and workforce. India follows, with more employees (41) than company locations (35), potentially pointing to a trend of outsourcing or remote employment. Germany shows a reverse trend, with more company locations (13) than residences (7), suggesting it could be a focal point for corporate presence over employee base.

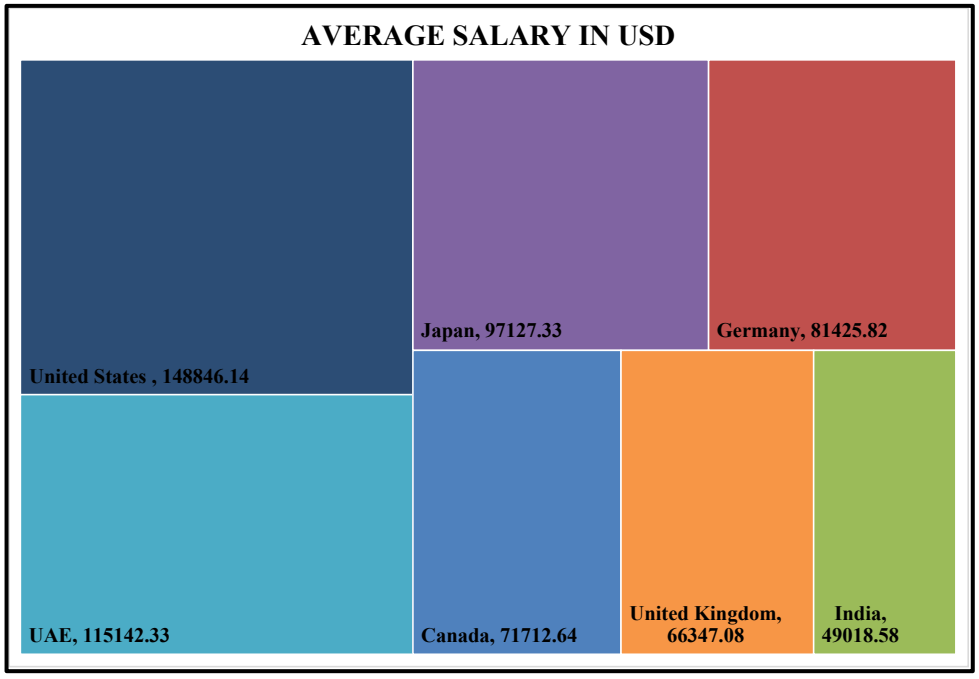

Figure 5: Tree Map of Average Salary in USD by Company Location

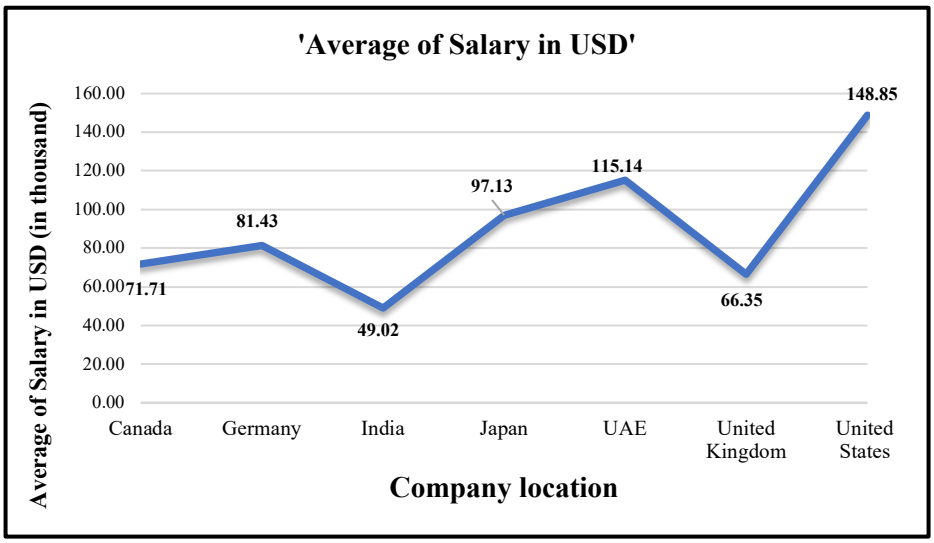

Figure 6: Line Chart Average Salary in USD by Company Location

Figure 5 and Figure 6 shows the average salary in USD across various company locations, with the United States topping the list at 148,846.14 USD, indicating a higher wage scale relative to other countries. The UAE follows with an average salary of 115,142.33 USD, while salaries in India are the lowest among the listed countries at 49,018.58 USD, reflecting the vast economic disparities across these regions.

Job Title by Experience Level	Average of Salary in USD
Applied Data Scientist	
Entry Level	70956.00
Intermediate	149527.60
Senior Level	278500.00
Data Analytics Lead	
Senior Level	**405000.00**
Data Science Manager	
Intermediate	200000.00
Senior Level	149994.20
Data Specialist	
Senior Level	165000.00
Director of Data Science	
Expert	238948.67
Financial Data Analyst	
Entry Level	100000.00
Intermediate	**450000.00**
Head of Data	
Expert	232500.00
Intermediate	32974.00
Senior Level	151419.50
Machine Learning Scientist	
Entry Level	225000.00
Intermediate	109325.00
Senior Level	201666.67
Principal Data Engineer	
Expert	**600000.00**
Senior Level	192500.00
Principal Data Scientist	
Expert	**416000.00**
Intermediate	151000.00
Senior Level	187939.40

Figure 7: Average Salary in USD by Company Location

Figure 7, highlighted with conditional formatting, indicates the average salaries in USD for various data-centric job titles by experience level. Principal Data Scientists at the Expert level earn the highest salaries, averaging at 416,000 USD, signifying the premium on top-tier expertise. In contrast, Entry Level roles like

Applied Data Scientist start at a lower end with 70,956 USD, reflecting the salary progression tied to experience and job title.

4. Test for differences between the variables using t-tests and providing interpretations for each of the tests.

1) **T-Test Hypotheses for Comparing Average Salaries Between Entry Level and Senior Level Employees:**

- **Null Hypothesis (H0):** There is no significant difference in the average salaries between Entry Level and Senior Level employees.

- **Alternative Hypothesis (H1):** There is a significant difference in the average salaries between Entry Level and Senior Level employees.

	Entry Level	Senior level
Mean	**61257.88**	**146206.56**
Variance	2244537637	6171592197
Observations	40	81
Hypothesized Mean	0	
df	114	
t Stat	-7.39	
P(T<=t) one-tail	**0.0000000000135**	
t Critical one-tail	1.66	
P(T<=t) two-tail	**0.000000000027**	
t Critical two-tail	1.98	

Figure 8: t-Test: Two-Sample Assuming Unequal Variances

Findings: The comparison of average salaries between those just starting their careers and those at a senior level shows a really big gap. People at the start are earning an average of $61,257.88, while those who've climbed the ladder to a senior level are making around $146,206.56 on average. The P(T<=t) two-tail value is extremely small (**0.0000000000135**), which is well below the common alpha threshold of 0.05. This indicates a statistically significant difference in the average salaries between Entry Level and Senior Level employees. With such a low p-value, we can reject the null hypothesis that there is no difference in means, confirming that experience level does have a significant effect on salary.

2) **T-Test Hypotheses for Comparing Work-Life Balance Ratings between 100% Remote work ratio and 0% Remote work ratio:**

- **Null Hypothesis (H0):** There is no significant difference in the average work-life balance ratings between employees with a 100% remote ratio and those with a 0% remote ratio.

- **Alternative Hypothesis (H1):** There is a significant difference in the average work-life balance ratings between employees with a 100% remote ratio and those with a 0% remote ratio.

	WorkLifeBalance(0% remote ratio)	WorkLifeBalance (100% remote ratio)
Mean	**2.94**	**2.82**
Variance	0.53	0.72
Observations	35.00	131.00
Hypothesized Mean Difference	0.00	
df	61.00	
t Stat	0.88	
P(T<=t) one-tail	**0.19**	
t Critical one-tail	1.67	
P(T<=t) two-tail	**0.38**	
t Critical two-tail	2.00	

Figure 9: t-Test: Two-Sample Assuming Unequal Variances

Findings: The average work-life balance rating for employees with a 0% remote ratio is slightly lower at approximately 2.94 compared to those with a 100% remote ratio at about 2.82. However, the t-statistic of 0.88 indicates that this difference is not statistically significant, as the p-value of 0.38 is much higher than the conventional threshold of 0.05. Hence, we cannot reject the null hypothesis; thus, it suggests that remote work ratio does not have a significant effect on employees' perception of their work-life balance within this sample.

3) T-Test Hypotheses for Comparing Engagement Scores between Entry-Level and Intermediate-Level Employees

- **Null Hypothesis (H0):** There is no significant difference in the average engagement scores between entry-level and intermediate-level employees.

- **Alternative Hypothesis (H1):** There is a significant difference in the average engagement scores between entry-level and intermediate-level employees.

	EngagementScore of Entry level	EngagementScore for Intermediate level
Mean	**3.20**	**3.13**
Variance	0.83	1.34
Observations	40.00	86.00
Hypothesized Mean Difference	0.00	
df	95.00	
t Stat	0.38	
P(T<=t) one-tail	**0.35**	
t Critical one-tail	1.66	
P(T<=t) two-tail	**0.71**	
t Critical two-tail	1.99	

Figure 10: t-Test: Two-Sample Assuming Unequal Variances

Findings: The average engagement score for entry-level employees is 3.20, while for intermediate-level employees, it's slightly lower at 3.13. The t-statistic of 0.38 indicates the difference in means is not statistically significant, as the p-value of 0.71 is much higher than the conventional threshold of 0.05. With 95 degrees of freedom, we cannot reject the null hypothesis, suggesting that there is no significant difference in the average engagement scores between entry-level and intermediate-level employees.

5. Conduct a factor analysis on the variables. Keep the number of factors to a maximum of 5. Provide interpretations for the factors.

Variable Selection: For the factor analysis, we selected six variables that capture different aspects of employee experience and performance: Performance Rating, Work Life Balance, Remote Work Effectiveness, Job Satisfaction, Professional Development, and Engagement Score. These variables were chosen as they are expected to reflect underlying factors related to workplace satisfaction and effectiveness, especially in the context of professional development and remote work dynamics.

Data Suitability for Factor Analysis:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.497
Bartlett's Test of Sphericity	Approx. Chi-Square	435.883
	df	15
	Sig.	<.001

Figure 11: Kaiser- Meyer-Olkin(KMO) and Bartlett's Test

Correlation Matrix

		PerformanceRating	WorkLifeBalance	RemoteWorkEffectiveness	JobSatisfaction	ProfessionalDevelopment	EngagementScore
Sig. (1-tailed)	PerformanceRating		.215	.392	.426	.143	.343
	WorkLifeBalance	.215		.197	.304	.113	.195
	RemoteWorkEffectiveness	.392	.197		.000	.449	.300
	JobSatisfaction	.426	.304	.000		.492	.185
	ProfessionalDevelopment	.143	.113	.449	.492		.439
	EngagementScore	.343	.195	.300	.185	.439	

Figure 12: Correlation Matrix

The Kaiser-Meyer-Olkin (KMO) Figure 13, measure of sampling adequacy is 0.497, which falls below the commonly accepted threshold of 0.6, suggesting that the dataset may not be suitable for factor analysis (Li et al., 2020). However, Bartlett's Test of Sphericity has a significant result ($p < .001$), indicating that the variables are sufficiently interrelated for the analysis.

The correlation matrix (Figure 14) shows moderate positive relationships between Work Life Balance and Professional Development (0.449) and between Remote Work Effectiveness and Job Satisfaction (0.492), suggesting that employees who experience better work-life balance may perceive greater professional development opportunities and those more effective at remote work tend to be more satisfied with their jobs. The significance of these correlations is not specified, and as such, caution should be exercised in their interpretation without further statistical confirmation.

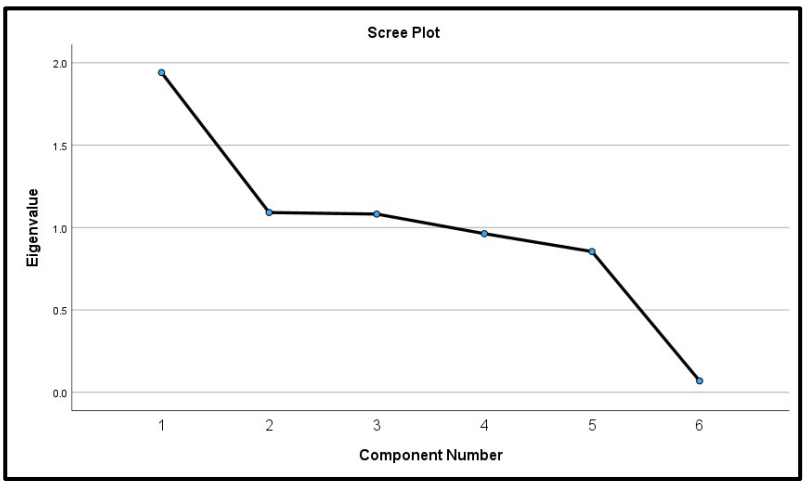

Figure 13: Scree Plot

Scree Plot Interpretation: The scree plot visually suggests a 'break' or 'elbow' after the third component, indicating that the first three components are likely to be the most significant. This pattern is a common criterion for deciding the number of components to retain, suggesting that beyond the third component, additional components contribute minimally to explaining the variance.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.941	32.343	32.343	1.941	32.343	32.343
2	1.091	18.185	50.528	1.091	18.185	50.528
3	1.082	18.029	68.557	1.082	18.029	68.557
4	.963	16.048	84.605			
5	.854	14.239	98.844			
6	.069	1.156	100.000			

Extraction Method: Principal Component Analysis.

Figure 14: Total Variance Explained

The total variance explained figure 16 shows that the first component explains 32.343% of the variance, the second 18.185%, and the third 18.029%, cumulatively accounting for 68.557% of the variance in the dataset. This indicates that a substantial portion of the total variance can be attributed to these three components, justifying their retention in the analysis.

Component Matrix[a]

	Component		
	1	2	3
PerformanceRating	-.025	.311	-.687
WorkLifeBalance	.103	.777	.077
RemoteWorkEffectiveness	.979	-.062	.006
JobSatisfaction	.979	-.070	-.014
ProfessionalDevelopment	.017	.394	.725
EngagementScore	-.106	-.476	.278

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Figure 15: Component Matrix

Component Matrix: In the component matrix (Figure 17) from the factor analysis using Principal Component Analysis, three components were extracted:

- **Component 1:** Strong positive loadings on Remote Work Effectiveness (0.979) and Job Satisfaction (0.979), indicating this component is strongly associated with how effective employees feel while working remotely and their overall job satisfaction.

- **Component 2:** The most notable loading is on Work Life Balance (0.777), suggesting this component reflects the balance between professional and personal life.

- **Component 3:** Performance Rating has a notable negative loading (-0.687), and Professional Development has a strong positive loading (0.725). This component seems to capture the contrast between employees' development opportunities and their current performance ratings, which might suggest that individuals with higher opportunities for professional growth may perceive their current performance rating differently.

These components help conceptualise broad themes in the data: Component 1 can be seen as capturing the 'Remote Work Dynamics', Component 2 as 'Work-Personal Life Equilibrium', and Component 3 may represent 'Growth and Performance Perception'.

6. Conducting Cluster Analysis

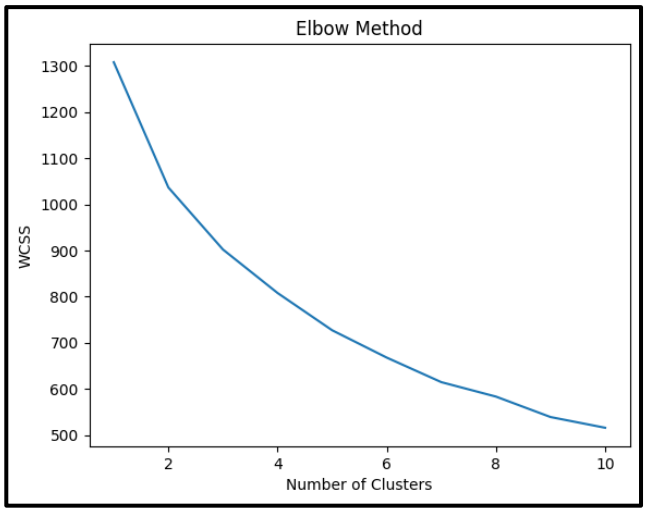

Figure 16: Elbow method for optimal number of clusters

Applying the Elbow Method to determine the best number of clusters (Gupta, 2019), the analysis demonstrated that three clusters represent the optimal solution, as evidenced by the distinct change in the rate of reduction in the Within-Cluster Sum of Squares (WCSS). This point, where the benefits of additional clusters begin to diminish, aligns with the findings from the Silhouette Score Analysis, reinforcing that three clusters ensure an optimal mix of quantity and coherence within the data. Consequently, the K-means algorithm was implemented to fit the dataset into these three identified clusters.

```
Cluster
2    122
0     55
1     41
Name: count, dtype: int64
   PerformanceRating  WorkLifeBalance  RemoteWorkEffectiveness  \
0           3.018182         2.727273                 1.545455
1           4.000000         3.048780                 3.121951
2           2.942623         2.918033                 3.467213

   JobSatisfaction  ProfessionalDevelopment  EngagementScore
0         1.727273                 3.090909         3.290909
1         3.317073                 2.853659         3.073171
2         3.696721                 3.032787         3.098361
```

Figure 17: Cluster Matrix

The cluster centroid table from the K-means analysis presents three distinct employee segments based on the variables Performance Rating, Work Life Balance, Remote Work Effectiveness, Job Satisfaction, Professional Development, and Engagement Score.

Cluster 0 are encountering challenges with the lowest average Performance Rating of approximately 3.02 and the lowest scores for Remote Work Effectiveness (around 1.55) and Job Satisfaction (about 1.73), which might suggest challenges in adapting to remote working and overall job contentment. These difficulties suggest potential mismatches between employees' needs and the support provided, calling for a strategic review of the existing remote work policies.

Cluster 1 is characterised by the highest Performance Rating with an average of 4.00, indicating a group that excels in their roles. They also have a good Work Life Balance score of 3.05, indicating a healthier equilibrium between work and personal life. With Remote Work Effectiveness at 3.12 and Job Satisfaction at 3.32, they appear to be effective and content in their work, although their Professional Development score is relatively lower at 2.85. This group seems engaged, as shown by an Engagement Score of 3.07, and might represent well-rounded employees who are effective in their roles and satisfied, yet still see room for professional growth.

Cluster 2 has the highest scores in Remote Work Effectiveness (3.47) and Job Satisfaction (3.70), suggesting a segment that is thriving in the remote work environment and showing high levels of job contentment. Their Performance Rating is near the lower end at 2.94, with Work Life Balance at 2.92, which indicates room for improvement. They are doing well in Professional Development (3.03) and Engagement Score (3.10), which could reflect employees who are generally content and have the resources they need for professional growth.

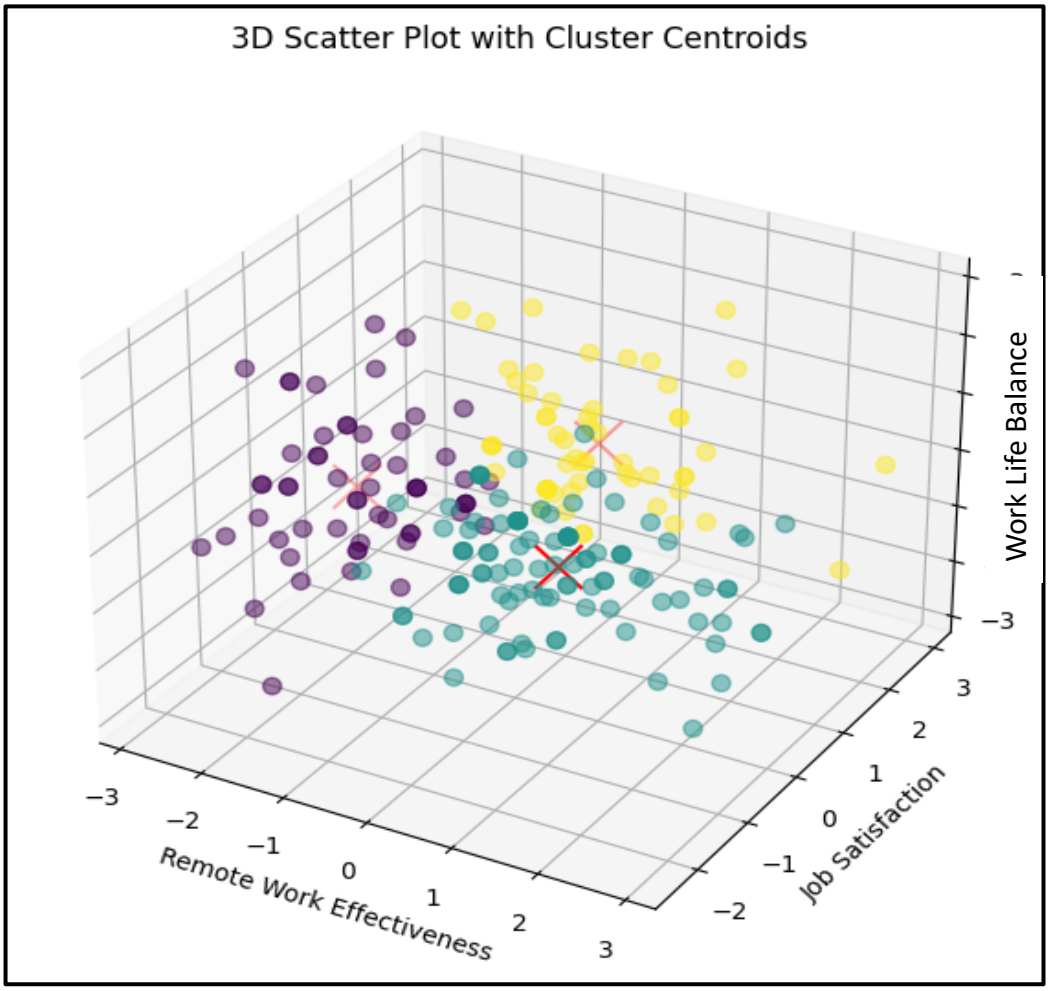

Figure 18: 3D Scatter Plots with Cluster Centroids

This 3D scatter plot depicts clusters of employees based on Job Satisfaction, Remote Work Effectiveness, and Work Life Balance. Cluster 0 (purple) may reflect employees with varied satisfaction and effectiveness but generally lower work-life balance. Cluster 1 (yellow) likely represents employees with high job satisfaction and remote work effectiveness, potentially the most well-adjusted group. Cluster 2 (teal) seems to consist of employees with moderate job satisfaction and effectiveness and possibly better work-life balance. The clusters' spread and the red centroid crosses indicate each group's average position in this multidimensional space. The position of the red 'X' marks shows the average score of each cluster along the three principal components, providing a reference point for the core attributes defining each cluster.

Strategic Enhancements for Diverse Employee Clusters:

The clustering analysis reveals actionable insights for enhancing workforce strategies. For Cluster 0, with its lower performance and satisfaction metrics, the organisation should bolster remote work support and enact targeted interventions to elevate performance and satisfaction. Cluster 1 exhibits high performance and job satisfaction, meriting recognition programs and enriched professional development to maintain and capitalize on this group's achievements. Meanwhile, Cluster 2 shows high satisfaction but comparatively lower performance, suggesting that aligning roles more closely with individual capabilities and providing appropriate professional growth opportunities could translate their contentment into improved performance. By implementing these tailored strategies, the organisation stands to significantly boost overall engagement and productivity.

Conclusion

This comprehensive analysis of the 2020 employment and salary dataset highlighted critical dynamics within different employee segments, using descriptive statistics, factor analysis, and cluster analysis. Three principal components emerged—'Remote Work Dynamics', 'Work-Personal Life Equilibrium', and 'Growth and Performance Perception', identifying distinct groups with unique needs: Cluster 0 requires enhanced remote work support, Cluster 1 benefits from continued recognition and growth opportunities, and Cluster 2 needs alignment of job satisfaction with performance.

The study faced limitations due to the suitability of the data for factor analysis and its cross-sectional nature, which might affect the generalisability of the findings, particularly given the unique context of 2020. Future research should extend to longitudinal studies to verify the persistence of these trends and explore causal relationships.

These insights advocate for tailored HR strategies that address the specific challenges and potentials within each cluster. Implementing such differentiated approaches can significantly enhance workplace satisfaction, engagement, and overall organisational productivity, paving the way for a supportive and thriving work environment.