A Data-Driven Exploration of Employment (1).docx

by HONG JING JAY SU.482BSDA2.202401.FT

Submission date: 27-Dec-2024 07:43PM (UTC+0800)
Submission ID: 2558370157
File name: A_Data-Driven_Exploration_of_Employment_1_.docx (1.07M)
Word count: 6379
Character count: 35162

A Data-Driven Exploration of Employment, Productivity, and Well-being

Beatrice Chai Xin Xuan
21060835

Chan Zheng Shao
24020059

Hong Jing Jay
22008338

Ivan Sing Wen Sie
Student ID

Lai Yi Huey
21069620

Vicky Leow Ming Fong
22009591

This study explores the relationship between employment type, productivity, and employee well-being across industries using datasets from LinkedIn Job and Remote Work Productivity sourced from Kaggle. The analysis aims to uncover patterns and correlations that can help optimize work conditions for better productivity and employee satisfaction. Data validation, cleaning, manipulation and merging were performed using SAS OnDemand, ensuring data consistency and reliability before analysis. Key research questions addressed include comprehending the relationship between employment type and well-being score, identifying the impact of employment type on well-being scores and productivity levels, discovering natural groupings among employees based on productivity, well-being and work hours, and understanding the relationship between hours worked per week and productivity per hour. Statistical techniques such as T-tests, ANOVA, and clustering algorithms (KNN and Gaussian Mixture Models) were applied to derive actionable insights. Results indicate that remote employment is associated with higher productivity and well-being scores, while excessive work hours negatively impact productivity per hour. SAS and R programming played a crucial role in data preprocessing, visualization, and hypothesis testing. The findings emphasize the importance of balancing work hours and fostering flexible employment structures to achieve optimal workforce efficiency and well-being.

INTRODUCTION

This research proposes the relationship between employment type, productivity, and well-being across industries. The datasets 'LinkedIn Job' with 5588 observations and 15 variables while 'Remote Work Productivity' consists of 1000 observations and 5 variables, which are discovered from Kaggle. 'LinkedIn Job' dataset contains the job postings in various industries with different employment types on LindkedIn, whereas 'Remote Work Productivity' dataset includes productivity score and well-being score with varying working hours and employment types (remote or on-site). The revealing of the effect of employment type on productivity and well-being metrics establish a deeper understanding of how employment structures influence employee performance and overall well-being. There are 4 problem statements that derived from the datasets. The problem statements aim to identify patterns and correlations that could guide companies in optimising work conditions in order to promote better productivity and enhance employee satisfaction across industries.

Unveiling the Relationship Between Employment Type, Productivity, and Well-being

Problem Statement

In the dynamic work environment nowadays, it is important to understand the impact of employment types and work conditions on employee productivity and well-being for fostering a healthy and efficient workforce. The research aims to investigate deeper understanding on impact of employment structures on employee performance and overall well-being by discovering the following key research questions:

Is There Any Relationship Between Employment Type and Well Being Score?
Does Employment Type have Significant Impact on Productivity Score?
What Natural Groupings Exist Among Employees Based on Productivity, Well-being, and Work Hours?
What is the Relationship Between Hours Worked Per Week and Productivity Per Hour?

Objective

The objective of this research is to analyse the effects of employment type on employee performance with the purpose of optimising work conditions across industries, aligning with the Sustainable Development Goals (SDG) 8: Decent

Work and Economic Growth. We have conducted analysis to fulfil the research questions that we have identified from the datasets.

Toolkits and Software

To address the research questions, SAS OnDemand is the primary tools to be utilized for tasks such as data importing, merging, cleaning, and analyzing according to the research questions. Other than that, R is the additional tools to be employed for data analyzing.

SAS OnDemand

SAS On Demand was primarily utilized for data importing, merging, cleaning, exploratory data analysis, and creating visualizations, including bar charts and box plots. It was also used to conduct regression analysis and correlation analysis such as ANOVA, hypothesis testing, scatter plots.

R Programming

R was mainly used for data visualization, data manipulation and analysis as it offers different library packages such as ggplot2, dplyr, factoextra, cluster for creating cluster plots.

Data Import and Structure

Firstly, the **PROC IMPORT** procedure was used to load the Job Dataset into the SAS environment. The datafile option specifies the file path, while dbms indicates the file format (CSV). The **getnames** option ensures that the first row is read as column headers, and **guessingrows**=max reads the maximum number of rows to determine data types.

```
/* import job dataset*/
PROC IMPORT datafile = '/home/u63870953/AE Assignment/job_cleanData.csv'
dbms = csv replace
out = WORK.JOB;
getnames = yes;
guessingrows = max;
RUN;
```

Fig 1. – Data Import for Job Dataset

Once imported, we verified the structure of the dataset using **PROC CONTENTS**. This command provided an overview of the dataset, including variable names, formats, and data types.

```
/* check the informat and format of the dataset*/
TITLE "Dataset Structure: Job Dataset";
PROC CONTENTS data=work.job;
RUN;
TITLE;
```

Fig 2. – Data Structure of Job Dataset

In Fig 3. we are importing the job_cleanData.csv file into SAS using the **INFILE** statement. The file path is specified, and several options are set, such as **missover**, **dsd**, firstobs=2 and obs=1001 to properly read the data from 2nd row to 1001th row and skip the header row. The **INFORMAT** and **FORMAT** statements are used to specify the data types and formats for the variables in the dataset. For example, job_ID is formatted with best32. to handle larger numbers, while string variables such as job_title and work_type are defined with $ to indicate character data. The **INPUT** statement lists the variables being read from the CSV file. This includes job_ID, job_title, company_id, and other variables relevant to the dataset. This structure ensures that the data is loaded with the correct types and formats for further analysis.

```
DATA work.job ;
INFILE '/home/u63870953/AE Assignment/job_cleanData.csv' dlm=','
missover dsd lrecl=32767 firstobs=2 obs=1001;

informat job_ID best32. ;
informat job_title $31. ;
informat company_id best32. ;
informat comp_name $70. ;
informat work_type $7. ;
informat involvement $10. ;
informat employees_count best32. ;
informat total_applicants best32. ;
informat linkedin_followers best32. ;
informat job_details $12569. ;
informat details_id best32. ;
informat industry $51. ;
informat level $16. ;
informat City $26. ;
informat State $17. ;
format job_ID best12. ;
format job_title $31. ;
format company_id best12. ;
format comp_name $70. ;
format work_type $7. ;
format involvement $10. ;
format employees_count best12. ;
format total_applicants best12. ;
format linkedin_followers best12. ;
format job_details $12569. ;
format details_id best12. ;
format industry $51. ;
format level $16. ;
format City $26. ;
format State $17. ;

INPUT job_ID
      job_title  $
      company_id
      comp_name  $
      work_type  $
      involvement  $
      employees_count
      total_applicants
      linkedin_followers
      job_details  $
      details_id
      industry  $
      level  $
      City  $
      State  $;
```

Fig 3. – Import and Format the Job Dataset

The Productivity Data was imported using a similar process. Here, **PROC IMPORT** was used with DBMS=csv and **replace** to handle the CSV file efficiently. Again, the **guessingrows** option was set to max to analyze all rows for accurate variable type detection.

```
/* import productivity dataset*/
PROC IMPORT datafile = '/home/u63870953/AE Assignment/remote_work_productivity.csv'
dbms = csv replace
out = WORK.productivity;
getnames = yes;
guessingrows = max;
RUN;
```

Fig 4. – Data Import for Productivity Dataset

After importing, **PROC CONTENTS** was used to inspect the dataset structure, providing critical information about variables such as Employment_Type, Hours_Worked_Per_Week, Productivity_Score, and Well_Being_Score. This ensured the data was appropriately formatted and ready for subsequent steps.

```
/* check the informat and format of the dataset*/
TITLE "Dataset Structure: Productivity Dataset";
PROC CONTENTS data=work.productivity;
RUN;
TITLE;
```

Fig 5. - Data Structure of Productivity Dataset

In Fig 6., we are specifying the file path and delimiter with the **INFILE** statement. It uses **missover** to handle missing values, firstobs=2 to start reading from the second row, and limits the dataset to 1001 observations with obs=1001. The **INFORMAT** and **FORMAT** statements define variable formats for accurate reading and display. The **INPUT** statement lists variables to be read, ensuring proper data types are assigned to each variable for further analysis.

```
DATA work.productivity ;
INFILE '/home/u63870953/AE Assignment/remote_work_productivity.csv' dlm=','
missover dsd lrecl=32767 firstobs=2;

informat Employee_ID best32.;
informat Employment_Type $9.;
informat Hours_Worked_Per_Week best32.;
informat Productivity_Score best32.;
informat Well_Being_Score best32.;
format Employee_ID best12.;
format Employment_Type $9.;
format Hours_Worked_Per_Week best12.;
format Productivity_Score best12.;
format Well_Being_Score best12.;

INPUT Employee_ID
      Employment_Type $
      Hours_Worked_Per_Week
      Productivity_Score
      Well_Being_Score;

DROP Employee_ID;
```

Fig 6. - Import and Format the Productivity Dataset

Data Validation and Cleaning

Handling Missing Value

After importing the Job Dataset, we performed several steps for data validation and cleaning. First, we checked for missing values and inconsistencies in the dataset using **arrays** for both numeric and character variables. The **missing_count** variable was used to tally the number of missing values for each record. In addition, we replaced any instances of 'Not Avilable' with a period ('.') to standardize missing data representation.

```
/* check missing values of each variables */
    /* array for numeric variables */
    ARRAY num_vars {*} _NUMERIC_;
    /* array for character variables */
    ARRAY char_vars {*} _CHARACTER_;

    /* count missing values */
    missing_count = 0;

    /* loop through numeric variables */
    DO i = 1 TO DIM(num_vars);
        IF MISSING(num_vars{i}) THEN missing_count + 1;
    END;

    /* loop through character variables */
    DO i = 1 TO DIM(char_vars);
        IF MISSING(char_vars{i}) THEN missing_count + 1;
    END;

    /* replace 'Not Avilable' with '.' */
    DO i = 1 TO DIM(char_vars);
        IF char_vars{i} = 'Not Avilable' THEN char_vars{i} = '.';
    END;
```

Fig 7. – Check and Replace Missing Value

Similarly, for the Productivity Dataset, the missing values were handled using **array** to the Job Dataset, where any missing or inconsistent values were appropriately identified and processed.

```
/* ensure no missing values */
ARRAY num_vars {*} _NUMERIC_;
ARRAY char_vars {*} _CHARACTER_;
missing_count = 0;

DO i = 1 TO DIM(num_vars);
    IF MISSING(num_vars{i}) THEN missing_count + 1;
END;

DO i = 1 TO DIM(char_vars);
    IF MISSING(char_vars{i}) THEN missing_count + 1;
END;
```

Fig 8. – Check Missing Value

Filtering Unnecessary Data

After addressing these issues, we filtered out rows with 'Hybrid' work types to focus only on **On-site** and **Remote** data. The **KEEP** statement ensures that only the variables job_title, comp_name, work_type, total_applicants, industry, and missing_count are retained in the dataset.

```
/* filter out rows with 'Hybrid'*/
IF work_type = 'On-site' OR work_type = 'Remote';

KEEP job_title comp_name work_type total_applicants industry missing_count;
RUN;
```

Fig 9. – Remove unnecessary rows

Standardization of Variable

In Fig 10., we replaced 'In-Office' with 'On-site' using **IF statement** to maintain uniformity across the dataset in order to carry out the next step.

```
/* replace 'In-Office' with 'On-site'*/
IF Employment_Type ='In-Office'
    THEN Employment_Type = 'On-site';
RUN;
```

Fig 10. – Replace Name to Match the Employment Type

To ensure consistency across the datasets, the Employment_Type variable was standardized in both the job and productivity datasets. This involved **renaming** the work_type variable in the job dataset to Employment_Type and explicitly setting the variable **length** to 9 characters in both datasets. Standardizing variable names and lengths is essential to prevent errors during the merging process and to ensure a seamless integration of data.

```
/* standardize variable length*/
DATA work.job;
    LENGTH Employment_Type $9;
    SET work.job (RENAME=(work_type=Employment_Type));
RUN;

DATA work.productivity;
    LENGTH Employment_Type $9;
    SET work.productivity;
RUN;
```

Fig 11. – Change and Set Variable Length of the Employment Type

Data Manipulation and Merging

Adding Row Index for Alignment

The work.job and work.productivity datasets are sorted by Employment_Type to ensure a consistent ordering of the records. Following the sort, we created a new variable, **Row_Index**, to assign a unique identifier to each row in both datasets and temporarily saved in the work.job_with_index and work.producitivity_with_index dataset respectively. This alignment ensures that when we merge the datasets later, each record corresponds correctly.

Fig 12. – Add Row Index for Alignment

Merging Datasets

The following step in Fig 13. merges the work.job_with_index dataset with the work.productivity_with_index dataset using a **LEFT JOIN** in **PROC SQL**. A LEFT JOIN was chosen to retain all records from the job dataset while integrating matching data from the productivity dataset. This ensures no critical job data is lost, unlike an INNER JOIN, which excludes unmatched rows, or a FULL JOIN, which includes unnecessary data. The merge is performed based on two keys: Employment_Type and the newly added Row_Index. This ensures that rows align correctly between the two datasets. The resulting dataset, work.merged_data, contains variables from both datasets, such as job details, total applicants, productivity scores, and well-being scores, enabling a comprehensive analysis across both datasets.

Fig 13. - Merge Job and Productivity Dataset

Derived Variables

After merging the datasets, there are some rows containing missing values. Thus, the missing values in any numeric variables (Hours_Worked_Per_Week, Productivity_Score, Well_Being_Score) or character variables (Employment_Type, job_title, comp_name, industry) were removed to ensure data completeness. The **NMISS** function was used to detect missing values in numeric fields, while the **CMISS** function checked for missing values in character fields. Following this, a new variable, Productivity_Per_Hour, was derived by dividing the Productivity_Score by the Hours_Worked_Per_Week. This variable represents the productivity achieved per hour worked and provides additional insight into employee efficiency. Finally, the Productivity_Per_Hour variable was formatted to display values with two decimal places.

Fig 14. – Remove Rows with Missing Value and Add New Variable in Merged Dataset

In Fig 15, the **PROC SORT** procedure is used with the **NODUPKEY** option, which removes any rows with identical values across the specified variables. Specifically, the dataset work.cleaned_data is sorted and output to a new dataset work.cleaned_data_no_dups, excluding duplicates. The variables considered for duplicate checks include Employment_Type, job_title, comp_name, total_applicants, industry, Hours_Worked_Per_Week, Productivity_Score, Well_Being_Score, and the derived variable Productivity_Per_Hour.

Fig 15. – Check Duplicate Key in Merged Dataset

Data Exploration

The data exploration for the uncleaned merged dataset involved analyzing both numeric and categorical variables. Descriptive statistics were generated using **PROC MEANS** to examine the **central tendencies** (mean) and **dispersion** (standard deviation, minimum, and maximum) for key variables such as Hours_Worked_Per_Week, Productivity_Score, Well_Being_Score, and total_applicants. The decimal place is being set to 2 using **MAXDEC**=2. Additionally, PROC FREQ was used to analyze the **frequency distribution** of Employment_Type and industry, allowing us to check for imbalanced categories or unexpected values.

Fig 16. - Conduct EDA For Uncleaned Merged Dataset

PROC SQL is used to create a summary table, summary_by_employment, by grouping the cleaned dataset based on Employment_Type. Within each group, it calculates key metrics such as the number of records (Count_Records), the average productivity score (Avg_Productivity), the average well-being score (Avg_Well_Being), and the average hours worked per week (Avg_Hours). This aggregation provides an analytical summary of the cleaned data. Next, **PROC PRINT** is used to display the summary table in a clear and organized format. The **FORMAT** statement ensures that all calculated averages are displayed with two decimal places for readability, while the **NOOBS** option removes row numbers for a cleaner presentation.

Fig 17. – Aggregated Summary Statistics by Employment Type on Merged Cleaned Dataset

Based on Fig 18., a stacked bar chart is produced using **PROC SGPLOT** to compare the **average productivity** and **well-being scores** by employment type. The **VBAR** statement creates bars for both metrics (Average Productivity and Average Well-Being) with labels on top of each bar for clarity. The Y-axis is labelled as "Average Score" and the X-axis as "Employment Type," while the legend, positioned at the bottom, clearly distinguishes the two metrics. This chart effectively highlights the differences in productivity and well-being across employment types. The outcome shows that the remote employees have higher Average Productivity and Average Well-Being compared to on-site employees.

Fig 18. – EDA on Merged Clean Dataset by Employment Type

RESULTS AND DISCUSSION

Research Question 1: Is there any relationship between Employment Type and Well Being Score?

One of the research questions to be addressed is if there is any relationship between Employment Type and Well Being Score.

To understand whether it is related or not. We first create an if-else statement and state a new column called well_being_status, this new column will be based on the well_being_score. For instance, wellbeing score with less than 25 will have bad wellbeing score, 26 to 50 will have below average wellbeing score, etc. It should also be known that if there is no value on wellbeing score, then the wellbeing status will be empty as well, indicated by the "." in the dataset.

After the if-else statement a table is created using PROC FREQ with no cumulative. What no cumulative does is that it prevents SAS system from calculating and displaying cumulative frequencies for wellbeing status. [1]

From Fig 19, we can see that out of 451 or 63.52% has an above average wellbeing status. The second most common wellbeing status is below average with 130 employees which took 18.31% of total. Good comes in third with a difference of 1.69% difference from below average while bad contributes the least with only 1.27% or 9 individuals/employees that have bad wellbeing status. Based on this table, there is room for improvement as a combined 20.86% of employees are classified as Below_Average, Bad, or Insufficient_Data, indicating potential concerns regarding employee well-being.

Fig 19. – Table of result based on new column

Moving on, we need to see if there is any difference in wellbeing score. For example, does On-site employment type have a higher wellbeing score or lower. To do this, the best plot would be the box plot.

PROC SGPLOT is implemented to visualize the data given after cleaning. VBOX tells the SAS system to use a box plot based on the column called Well_Being_Score based on the employment types mentioned which are On-site and Remote.

In Fig 20, we can see the boxplot compares the well-being scores of on-site and remote employees, highlighting some significant differences between the two groups. For on-site employees, the median well-being score is around the 60s, with a wider spread in scores, as indicated by the interquartile range (IQR) from approximately 50 to 70. Additionally, on-site workers show more variability, with several low-end outliers below 20, suggesting that some individuals report significantly lower well-being. The overall range extends from about 15 to 100, indicating greater inconsistency in scores.

In contrast, remote employment type has higher well-being scores overall, with a median of around 70 and a narrower IQR from 65 to 75. The data shows fewer outliers and a smaller range, spanning approximately 40 to 95, reflecting more consistent well-being among remote workers. Compared to on-site employees, remote workers tend to have higher and more stable wellbeing scores, suggesting that remote work may offer a more supportive environment for well-being and fewer extreme negative experiences. Having a higher wellbeing score can mean a lot of things. For example, since wellbeing score is related to productivity and health issues, this means that employees with higher wellbeing scores will have increased productivity, decreased absenteeism, improved morale, reduced turnover, and many more advantages. [2]

Fig 20. – Boxplot Result

To further prove that there is a relationship between Employment Type and Well Being Score, a T-Test hypothetical testing is used. T-test is a statistical test that compares means of 2 groups [3]. For this research question we will be using hypothesis testing.

H_0: $\mu 1 = \mu 2$
H_1: $\mu 1 \neq \mu 2$

Where $\mu 1$ is On-site's wellbeing score based on Employment type and $\mu 2$ is remote wellbeing score based on Employment type

In this test we experiment to see if there is any relationship whether there is a significant difference in the mean Well-Being Scores between two Employment Types The above symbols are the null hypothesis (H_0) and alternate hypothesis (H_1). Null hypothesis, or the current situation for the dataset is means that there is no significant difference between On-site and remote employment types. On the other hand, the alternate hypothesis wants to experiment on there is a significance between On-site and Remote employment types.

In Figure 21, the most important thing is the P-value which is labelled as "Pr > F" in Fig 24. The significance level for this test is 0.05. In the table, the P-value is < 0.0001 which is less than the significance level. So, when 0.0001 < 0.05, we reject the null hypothesis. In other words, there is significance difference between On-site and Remote employment type.

Fig 21. – T-test results

For the final test, a Fischer's Exact test is conducted. The reason why Chi-Square test is not involved in this experiment is because SAS system has warned stating "WARNING: 30% of the cells have expected counts less than 5. Chi-Square may not be a valid test" so Fischer's Exact test will be better for this experiment.

H_0: There is no association between Employment Type and wellbeing status.
H_1: There is an association between Employment Type and wellbeing status.

In Fisher's Exact test, using a 0.05 significance level, the probability shown is only 1.4 than 0.0001. Ultimately mean that, the null hypothesis is rejected as 0.0001 < 0.05 significance level. Therefore, there is enough evidence to show that there is an association between employment type and wellbeing status.

In Figure 22, we can see that chi-square is present in the table, but we won't be interpreting it. The phi coefficient is a measure of association between the 2 variables [4]. Contingency coefficient is a measurement to see if the variables are independent or dependent on each other [5]. Cramer's V measures how strongly the 2 variables are dependent on each other. The range of Phi Coefficient is from 0 to 1, 0 being no association and 1 having a high association. Contingency coefficients also have a range from 0 to 1 but measures to see if they are associated. Cramer's V has the same value range but measures to see if they are associated. [6]

From the results, we can see that the Phi Coefficient and Cramer's V value is 0.3098 and Contingency Coefficient is 0.2959. This means to say that all the association between the 2 columns are only moderate, not too much or too little. The same goes for contingency coefficients, both variables are not too dependent on each other. And lastly, the 2 variables are only moderately attracted to each other.

Fig 22. – Result from Fisher's Exact test

Research Question 2: Does Employment_Type have Significant impact on Productivity_Score?

Another question that needs to be clarified would be does employment types influence the productivity scores of employees.
To answer this question, SAS was used to generate graphs and an Anova table to analyze and clarify the relationship between Employment_Type and Productivity_Score.

The Figure 23. shows the average productivity scores of both employment types which is Remote and In-office. The visualisation showed that remote employees had a much higher average score of (73.63) compared to in-office employees with an average score of (63.77). The difference between them is near to 10 points, this highlights a disparity in productivity levels based just on employment type alone, which may imply that a remote work environment promotes a more conducive to higher and better productivity. Many reasons can be contributed to why productivity score is higher if the employment type is remote, remote workers may benefit from reduced commuting time which is one of the factors that wears out an employee the most, fewer workplace distractions. Not only that, but remote workers also usually have flexible working schedules which allows them to work during their peak productive hours, therefore further contributing to their productivity levels.

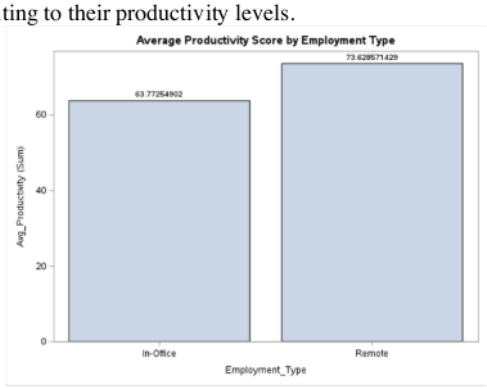


Fig 23. – Bar Chart for employment type on productivity score

Figure 24. is a boxplot that shows the distribution of Productivity_Score based on the Employment_Type. The median is represented as a horizontal line in the middle of the box and the median for in-office employees appears to be closer to the value 60 on the y-axis. This tells us that half of the employee productivity score is below 60 which a significant amount of low performing employees whereas the other half are between 60 and 80, showing a lesser number of employees with higher score.

On the other hand, the remote employee's median seems to be hovering around the value 75. This could mean about half the remote employees have scores above 75 which is an improvement compared to in-office workers with the score of 60. Then, the other half of remote worker's scores is between an estimate of 40 and 75 although it is low but it is still around or higher than those employees who belongs in the in-office median.

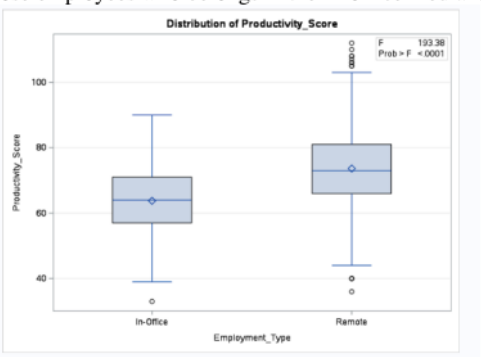


Fig 24. - Box Plot for Distribution of Productivity_Score

While the bar chart provides an indication of a difference in both variables, statistical testing was also conducted to assure whether this difference is significant or not.

Anova also known as Analysis of Variance, allows us to simultaneously compare arithmetic means across groups. We can also determine whether the differences observed are due to random chance or if they reflect genuine, meaningful differences [8].

The output from Anova in Figure 25. shows that there is an independent variable which Employment_Type with 2 levels to it which are "In-Office" and "Remote" These 2 categories represent the groups that will be compared with a total of 1000 observations.

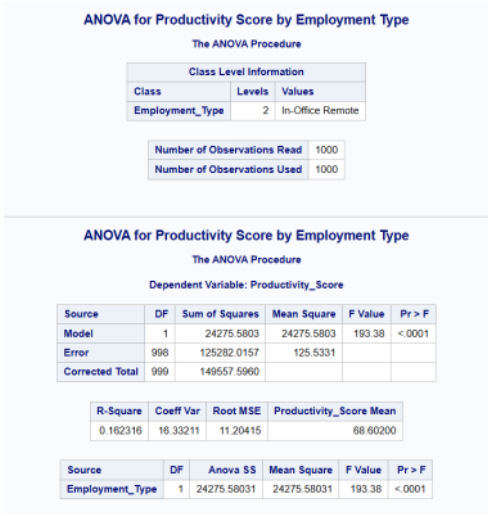
The null hypothesis states that Employment_Type has no significant impact on Productivity_Score and the alternative hypothesis states that Employment_Type has significant impact on Productivity_Score. With this in mind, we can confidently reject the null hypothesis because the corresponding p-value is less than 0.0001, which is below the standardly used significance level of 0.05, indicating the differences in productivity scores between In-Office and Remote employees are unlikely to have occurred by chance but by proper reasons.

On top of that, the F-value shown in the table indicates that the variance between the group means is actually much larger than the variance within the groups. Therefore, a high F-value further strengthens the correlation between productivity_score and employment_type. A large F-statistic also means that the model has also successfully highlighted a significant difference in productivity due to employment types which indicates a finding that employment type has immense influence on productivity.

Additionally, the sum of squares (SS) is a measurement that measures variation. A low sum of squares indicates little variation between data sets while a higher one indicates more variation. Variation refers to the difference of each data set from the mean [9].

In this context the total sum of squares is 149557.5960, which is the combined sum of both model sum of squares and error sum of squares. We can gauge that there is probably a diversity in the Productivity_Score values across the data, which can be due to the different factors that may or may not affect the productivity score that are not yet accounted in the model.

With this we can conclude that Employment_Type significantly influences Productivity_Score based on the evidence above. Because it exists strong evidence that Productivity_Score is significantly influenced by Employment_Type although the small effect sized shown by the R-squared value.



ANOVA for Productivity Score by Employment Type

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
Employment_Type	2	In-Office Remote

Number of Observations Read	1000
Number of Observations Used	1000

ANOVA for Productivity Score by Employment Type

The ANOVA Procedure

Dependent Variable: Productivity_Score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	24275.5803	24275.5803	193.38	<.0001
Error	998	125282.0157	125.5331		
Corrected Total	999	149557.5960			

R-Square	Coeff Var	Root MSE	Productivity_Score Mean
0.162316	16.33211	11.20415	68.60200

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Employment_Type	1	24275.58031	24275.58031	193.38	<.0001

Fig 25. – ANOVA Table

Research Question 3: What Natural Groupings Exist Among Employees Based on Productivity, Well-being, and Work Hours?

The dataset elicits a question regarding the patterns that exist within the natural groups of formed based on their productivity, well-being and work hours to understand different employee profiles within the dataset.

In regard to the question, feature selection is done by picking out the key attributes Productivity_Per_Hour, Well_Being_Score, and Hours_Worked_Per_Week in hopes of understanding the profiles of the employees based on these attributes. Utilizing the cleaned dataset, the data is then imported into R studio to be utilized by conducting a cluster analysis.

K-Nearest Neighbour (KNN) model is chosen as the base clustering model to its wide application, flexibility and simplicity for tackling clustering problems [10]. The approach taken for this particular model begins with the usage of the Elbow Method/Elbow Criterion as it shows a visualization of the optimal clusters expressed by the Sum of Squared Errors (SSE) [11].

$SSE = \sum_{k}{i=1}\sum_{j \in Ci}||{x}_j - {\mu}_i||2 SSE = \sum i=1k \sum j \in Ci||xj - \mu i||2$

The Elbow Method graph seen in figure 26 showcases that the optimal cluster based on silhouette width would be 2 clusters as on average the clusters silhouette width would peak around 0.3 and above, this entails 2 particular details that can be extracted firstly to optimize the quality of the clusters the Elbow Method shows us that 2 clusters would help to maximize the average silhouette over a range of possible values for k and secondly, that the quality of the clusters formed will be of considerate strength as they are above 0.25 (weak) and less than 0.5 (moderate) which are the general ranges of strength established which indicates that the models ability to group data points and differentiate groups from one another are of a low to moderate effectiveness [12].

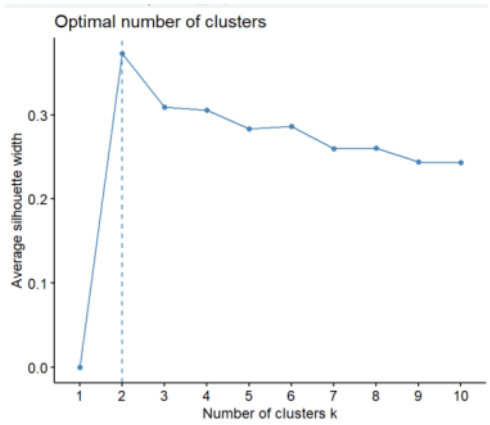


Fig 26. – Elbow Method Graph based on Silhouette Score

When conducting clustering with the KNN model it is observed that the 2 clusters have a lot of data points that are concentrated towards the center as seen in figure 27 which showcases that the clustering features are similar without much distinction in their variability. Although Principal Component Analysis (PCA) had been applied to reduce the dimensions of the plot with a 62.3% and 30.2% dimension for the x and y-axis respectively to explain the variance in the data most points are still gathered in the middle meaning the variability does not spread much across the different components leading to small differences between the measured variables for the 2 clusters making it harder to form highly differentiable clusters with the given variables [13].

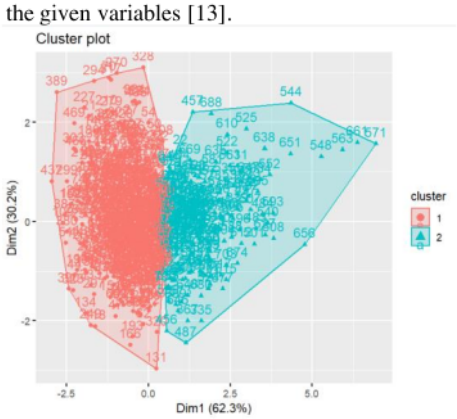


Fig 27. – K-nearest Neighbour Cluster Plot

From the summary table generated above in figure 28 the 2 clusters generated showcases the aggregated statistics for each cluster, for Cluster 1 we can observe that the main defining characteristics of the cluster groups employees by having a low productivity and lower well-being score while having an average high hours worked. This cluster can be used to showcase the overworked employees that experience a low productivity and well-being. Cluster 2 we can observe that the employees are grouped by a high productivity and well-being score while having an average low hours worked. This results in a cluster of employees that are productive in their work due to living a balanced life.

From here it should be noted that the natural grouping formed indicates the difference in productivity scores of the 2 clusters due to the difference in their working hours and well-being. Employees in Cluster 1 which is the overworked cluster should have their working hours reduced to increase well-being scores which in t increases productivity scores. This hypothesis could be tested with a variety of ways such as a basic hypothesis test, Pearson's correlation test or even a Chi-square test for the strength of said correlation. Aside from statistical testing Regression models can also be used to showcase the relationship between well-being score and working hours and make predictions on productivity scores.

As previously mentioned, the KNN cluster model in figure 27 showcased that most of the data points are congregated towards the center of the PCA plot to understand and identify the distinct patterns in the dense center region of the plot, a nonlinear clustering method the Gaussian Mixture Model (GMM) was chosen to help better understand the natural groupings with the given dataset by expanding upon the soft clusters present in the dataset through the usage of a probability model [15].

	Cluster	Avg_Productivity	Avg_Well_Being	Avg_Hours_Worked	Count
	<int>	<dbl>	<dbl>	<dbl>	<int>
1	1	1.44	59.0	44.8	501
2	2	2.30	70.2	32.9	209

Fig 28. – K-nearest neighbour Model Summary Table

The figure 29 showcases the summary of the Gaussian Mixture Model (GMM) after data has been fitted into the model. The Gaussian Mixture Model (GMM) for R using the mclust library runs a Bayesian Information Criterion (BIC) with the data being fitted into the model to identify the optimal amount of clusters by balancing the fit of the model (log-likelihood) and the complexity of the model (parameters and clusters) to discourage overfitting of the model as well; the results suggest that the model has a reasonable fit with the data given while balanced with its complexity at 3 clusters.

Having a value of -2492.347 for the log-likelihood indicates that the data used fitted for the model may not be a perfect fit but aligns with the assumptions made by the Gaussian Mixture Model (GMM) [15]. The results of the Integrated Completed Likelihood (ICL) of the model indicates that by having a lower value compared to the Bayesian Information Criterion (BIC) we can interpret that the clusters identified by the model are not highly distinct and may have overlapping clusters in assigning the data points [16].

```
Mclust EEV (ellipsoidal, equal volume and shape) model with 3
components:

 log-likelihood   n df       BIC       ICL
     -2492.347 710 23 -5135.695 -5527.395

Clustering table:
  1   2   3
331  34 345
```

Fig 29. – Summary of Gaussian Mixture Model (GMM) when fitting data points

The claim regarding interpretations of the Integrated Completed Likelihood (ICL) can be further backed by the figure 30 which shows that clusters do in fact have a lot of overlapping with one another resulting in many soft clusters. From the figure 30 we can also identify that there are a large number of data points that lie outside the elliptical highlighted regions of the clusters which indicate outliers present within the dataset. The spread out data points present in cluster 2 also showcases that there is a high variability between the data points within the Cluster 2.

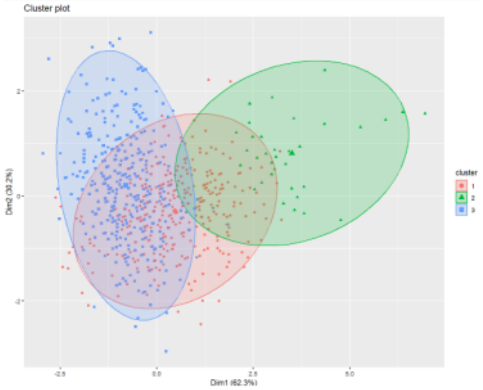


Fig 30. – Cluster Plot of the Gaussian Mixture Model (GMM)

Figure 31 shows the characteristics of the natural clusters formed by the Gaussian Mixture Model (GMM). Once again due to outliers present within the dataset we can observe that Cluster 2 has a low number of data points present within the cluster which may either require the removal of outliers or it could also be caused by having an imbalanced dataset.

From the summary in figure 31 we can see that as mentioned previously with the ICL score that Cluster 1 and Cluster 3 seem to share very similar characteristics while Cluster 2 showcases a very distinct difference from the 2 clusters by having a high productivity and well- being score while having a low working hour.

	Cluster	Avg_Productivity	Avg_Well_Being	Avg_Hours_Worked	Count
	<dbl>	<dbl>	<dbl>	<dbl>	<int>
1	1	1.86	68.4	39.8	331
2	2	3.33	66.9	25.6	34
3	3	1.37	56.0	44.3	345

Fig 31. – Gaussian Mixture Model (GMM) summary table

The model's poor performance in distinguishing between the cluster's 1 and 3 can also be backed up by the silhouette score seen in figure 32 which showcases that the average silhouette width is 0.18 which shows a very weak strength and quality of the clusters formed.

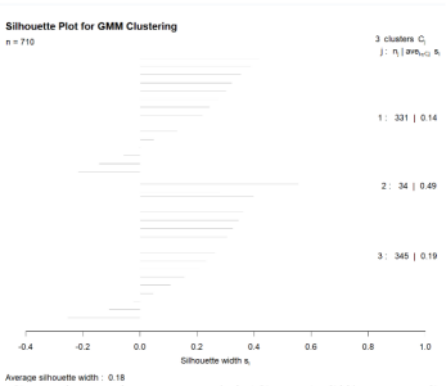


Figure 32. – Gaussian Mixture Model (GMM) Silhouette Score Plot

In conclusion, the KNN model performed better than the GMM model in clustering the natural groupings present in the dataset despite having many overlapping datapoints, which can be attributed to the outliers potentially skewing the results of the GMM.

From the results of the KNN model we can also derive actions that with the insights gained from this research question such as finding out the correlation between productivity, well-being, and work hours and making Regression models to predict productivity scores to ensure high productivity among employees.

Research Question 4: What is the Relationship Between Hours Worked Per Week and Productivity Per Hour?

This research question asks the question of the impact work hours per week has over productivity per hour. The analysis of the relationship between hours worked per week and productivity per hour provides an insight into employee efficiency. SAS was used to generate the scatter plot analysis, descriptive statistics, box plot analysis, linear regression analysis, correlation analysis and Anova test.

Figure 33 shows the relationship between hours worked per week and productivity per hour. Each employee is represented as a datapoint on the graph. The hours worked per week is on the X-axis and productivity per hour is on the Y-axis. The scatter plot shows a trend which indicating that productivity per hour usually decreases as the employee work for more hours [21].

A regression trend line was overlayed on the scatter plot to highlight this relationship. It suggests that while productivity per hour does not increase as there is more hours worked, it usually declines as the hours increases. This means that the productivity per hour is inversely proportional to the hours worked per week. For example, employees working between 30 and 40 hours per week often shows a higher productivity per hour compared to those that exceeds 40 hours. This graph shows some outliers but nothing too extreme.

Fig 33. – Scatter Plot Analysis

Figure 34 shows a box plot that categorizes employees based on their hours worked per week. Here, the work hours are categorized into short hours, medium hours and long hours for easier analysis.

Short hours encompass 1 to 30 hours per week. The median productivity per hour for this group is relatively high compared to the other groups, which indicates that more than half of the employees in this category have productivity scores above the median value. The interquartile range and the presence of outliers at the top end suggests that there's values that pulls the box plot towards the right.

Medium hours encompass 31 to 40 hours per week. This category has the highest median productivity per hour. The interquartile range is narrower compared to the Short Hours group, which indicates a more consistent productivity among employees in this range. This group has better balance, with fewer extreme outliers skewing the box plot.
Long hours encompass 41 and above hours per week. The median productivity per hour in this group is significantly lower compared to the Medium Hours and Short Hours category. The wider interquartile range and the more frequent low-end outliers shows the inefficiencies that is associated with overworking. The means that employees working longer hours tend to exhibit more variability and lower overall productivity.[20]

Fig 34. – Box Plot Analysis

Based on figure 35, an ANOVA test evaluates the differences in productivity per hour across the three categories of hours worked. The null hypothesis states that there is no significant difference in the mean productivity per hour between the three categories of hours worked while the alternative hypothesis states that at least one of the categories has a significantly different mean productivity per hour. Assuming a significance level of 0.05, the null hypothesis is rejected as the p-value of <0.0001 is less than the significance. This reveals statistically significant differences, with the Short Hours group consistently outperforming both the Medium and Long Hours groups in terms of average productivity per hour.

Fig 35. – ANOVA Test

Figure 36 shows the Tukey's post-hoc analysis. The *** indicates that the comparisons are significant at the confidence level of 0.05. Based on the table, every category of the hours is significant. This confirms the box plot analysis of the workers with the group of short working hours workers being significantly more productive compared to the other worker groups, while the group of long working hours workers being the least productive [19].

Fig 36. – Tukey's Studentized Range (HSH) Test

Figure 37 shows a linear regression model that explores the predictive relationship between hours worked per week and productivity per hour. The results indicate a statistically significant relationship, with a positive coefficient up to a certain threshold of hours. Beyond this point, the model shows a diminishing return, where more hours worked contribute less or even negatively impact the productivity per hour. For example, employees working around 35 hours per week achieve the highest predicted productivity, while those exceeding 45 hours show a decline. The R-squared value of 0.6029, though moderate, provides evidence that the hours worked is an important factor but not the exclusive factor in determining productivity [18].

Fig 37. – Linear Regression Model

Figure 38 shows the correlation coefficient between hours worked per week and productivity per hour is moderately positive, highlighting a beneficial relationship up to a point. However, the diminishing returns observed in the regression analysis align with the scatter plot graph, which suggests that the correlation weakens as hours worked increase beyond a certain range.

This analysis reveals that there is an optimal range of hours worked per week that maximizes productivity per hour. Employees working shorter number of hours are the most productive, while overworking leads to diminishing returns. This suggests that extended hours may contribute to fatigue and reduced efficiency [17].

Fig 38. – Correlation Analysis Table

CONCLUSION

In this report, we analyzed the relationship between employment type, employee well-being, and productivity using mainly two datasets sourced from Kaggle which are the 'LinkedIn Job' dataset and the 'Remote Work Productivity' dataset. Our work involved thorough data import, cleaning, and analysis using SAS OnDemand and R, ensuring the datasets were reliable for our research.

Our analysis led to several significant findings based on the four selected research questions. Notably, we discovered that remote workers generally reported higher well-being and productivity scores compared to their on-site counterparts. The analysis illustrated a critical inverse relationship between hours worked and productivity per hour, indicating that longer work hours may lead to diminished productivity. This suggests that flexible work arrangements can enhance employee satisfaction and performance.

Additionally, we found a negative correlation between hours worked per week and productivity per hour, indicating that longer work hours may lead to decreased efficiency. This insight emphasizes the importance of maintaining a balanced workload to optimize productivity. Hence, organizations need to re-evaluate their work-hour policies in favor of more flexible arrangements that can enhance overall performance.

In summary, the potential applications of these findings are significant. Organizations can leverage this knowledge to design work environments that prioritize employee well-being while fostering productivity. In a rapidly evolving workplace landscape, emphasizing flexible work arrangements not only meets the needs of employees but also positions organizations to thrive in a competitive economy. Thus, our findings contribute valuable insights for organizations seeking to enhance their work environments and align with Sustainable Development Goals (SDG) 8: Decent Work and Economic Growth.

A Data-Driven Exploration of Employment (1).docx

ORIGINALITY REPORT

6%	4%	2%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	1%
2	www.investopedia.com Internet Source	<1%
3	Submitted to Adtalem Global Education Student Paper	<1%
4	Submitted to Kean University Student Paper	<1%
5	Submitted to Cranfield University Student Paper	<1%
6	Submitted to Liberty University Student Paper	<1%
7	openknowledge.worldbank.org Internet Source	<1%
8	Pallant, Julie. "SPSS Survival Manual: A Step by Step Guide to Data Analysis using IBM SPSS", SPSS Survival Manual: A Step by Step Guide to Data Analysis using IBM SPSS, 2020 Publication	<1%

9	Submitted to Sunway Education Group Student Paper	<1 %
10	Rosy Oh, Nayoung Woo, Jae Keun Yoo, Jae Youn Ahn. "Predictive analysis in insurance: An application of generalized linear mixed models", Communications for Statistical Applications and Methods, 2023 Publication	<1 %
11	Submitted to American InterContinental University Student Paper	<1 %
12	Submitted to Oklahoma State University Student Paper	<1 %
13	Submitted to UC, Irvine Student Paper	<1 %
14	Submitted to Miami University of Ohio Student Paper	<1 %
15	Submitted to lnu Student Paper	<1 %
16	siliconvalley.basisindependent.com Internet Source	<1 %
17	tutorsonspot.com Internet Source	<1 %
18	Satria Bangsawan, MS Mahrinasari, Ernie Hendrawaty, Rindu Rika Gamayuni et al. "The	<1 %

	Future Opportunities and Challenges of Business in Digital Era 4.0", Routledge, 2020 Publication	
19	Submitted to University of Lancaster Student Paper	<1 %
20	auctoresonline.org Internet Source	<1 %
21	vdocuments.site Internet Source	<1 %
22	epdf.pub Internet Source	<1 %
23	www.journalcra.com Internet Source	<1 %

Exclude quotes	Off	Exclude assignment template	On
Exclude bibliography	Off	Exclude matches	Off