# SPSS ANALYSIS ON A GIVEN DATASET

# Basic data analysis and descriptive statistics

1. The variables included in the given HR Dataset and the type of measurement scales to which they belong are,

# Variable Name Variable Type Domain of Values

satisfaction\_level Continuous (Interval) 0.0 to 1.0

last\_evaluation Continuous (Interval) 0.0 to 1.0

number\_project Ordinal Like a number 2, 3, 4, 5….

average\_monthly\_hours Continuous (Ratio) 96 to 310

time\_spend\_company Continuous (Ratio) Measured in years like 2, 3, 4….

Work\_accident Nominal 0 = had no work accident; 1 = had a

work accident

left Nominal 0 = stayed with the company; 1 = left the company

promotion\_last\_5years Nominal 0 = not promoted; 1 = promoted department Nominal Various departments like sales, IT, hr,

etc.,

salary Ordinal 1 (low), 2 (medium), 3 (high)

According to Hatcher (2013), “Nominal scale variables classify participants into groups without measuring quantity. Ordinal scale variables categorize subjects into ordered categories based on quantity. Interval scale variables have equal intervals but lack a true zero point. Ratio scale variables possess all characteristics of interval scales and have a true zero point, making them the most sophisticated measurement scale”.

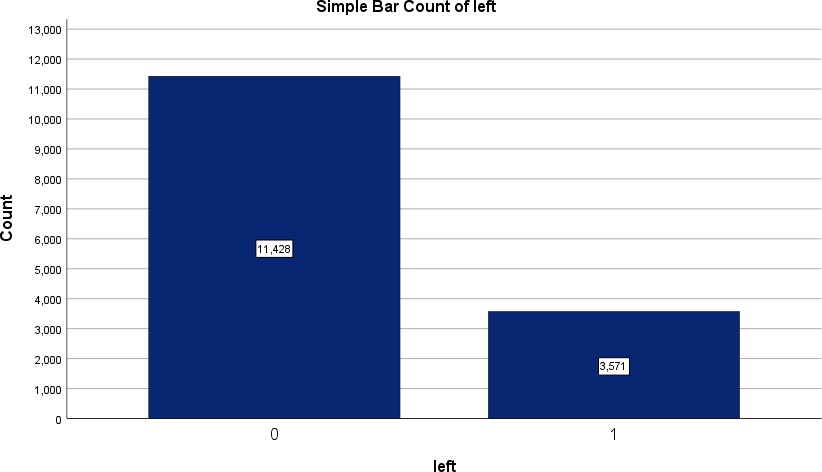
The variable "satisfaction\_level" is an interval scale variable. It is quantitative and exhibits equal intervals between values (e.g., the difference between 0.2 and 0.3 is the same as between 0.7 and 0.8). However, it lacks a true zero point, as a satisfaction level 0 does not represent the absence of satisfaction but rather a point on the scale. Similarly, "last\_evaluation" is also an interval scale variable as it shares the same characteristics with satisfaction\_level. It is quantitative, displaying equal intervals, but not a true zero point. The values in the range from 0.0 to 1.0 represent different levels of evaluation but do not imply the absence of evaluation at 0.0.

The variable "average\_monthly\_hours," a continuous variable with a range of values from 96 to 310, falls under the ratio scale category. This is because it meets the criteria of an interval scale (equal intervals between consecutive numbers) and has a true zero point, meaning that an absence of the variable (zero hours) is meaningful and indicates a complete lack of the measured construct. Similarly, the variable "time\_spend\_company," is a continuous variable measured in years (2, 3, 4...), and falls under the ratio scale category for the same reasons mentioned above.

The variable "number\_project" is considered an ordinal scale because it categorizes participants based on the number of projects undertaken, which can be ordered but lacks equal intervals between values. Similarly, the variable "salary" is considered ordinal because it categorizes salary levels in a defined order (low, medium, high) without implying equal intervals.

The variables Work\_accident, left, promotion\_last\_5years, and department are measured under a nominal scale as they only indicate whether the participants under them belong to two or more groups, but not any quantifiable parameter.

2.



The above is the bar chart demonstrating the number of employees who left the company (indicated by 1) and that of those who stayed with the company (indicated by 0). The number of people who stayed with the company is 11428, which is greater than those who left the company which is 3571.

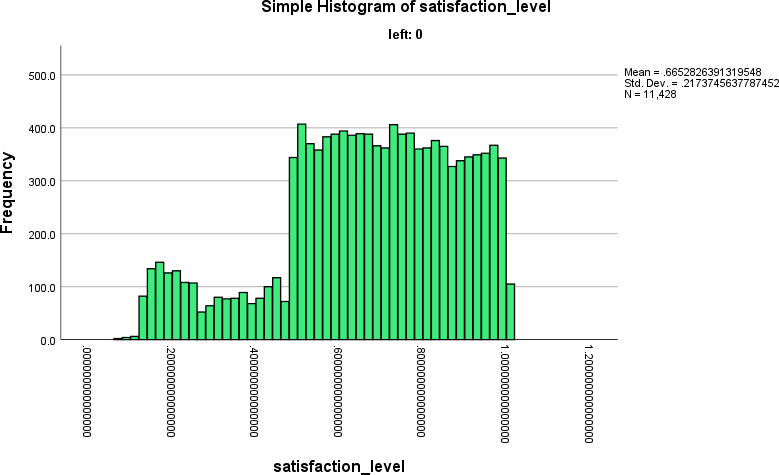
According to Hatcher (2013), “Proportion also known as relative frequency, is determined by dividing the simple frequency for a given value (here the value is 1, indicating the employees who left the company) by the total size of the sample (here it indicates the total number of employees which is equal to 14999), where the simple frequency is the raw number of observations displaying a given value (here it is the number of employees who left the company which is equal to 3571)”.

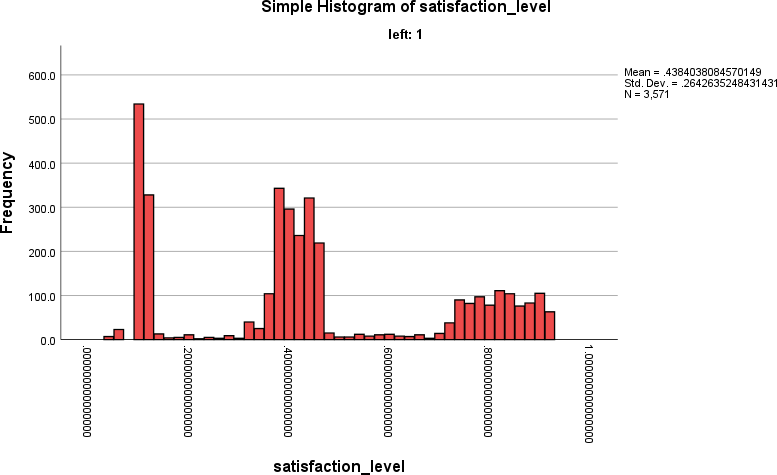
Therefore, the observed proportion of those who left the company from the HR dataset is 3571/14999, which is 0.238.

1. Below are the descriptive statistics for the satisfaction scores of the two groups i.e., those who left the company (as indicated by 1), and those who stayed (as indicated by 0).



1. Below are the histograms for the satisfaction scores of the two groups i.e., those who left the company (as indicated by 1), and those who stayed (as indicated by 0).





For those who stayed with the company, the distribution has a negative skewness of -0.605, indicating that more employees have higher satisfaction scores, with a tail towards lower scores. The kurtosis of -0.211 suggests a platykurtic distribution, indicating a slightly flatter peak than a normal distribution (Dave Longhorn, 2023c). In contrast, for employees who left the company the distribution has a positive skewness of 0.293, indicating more employees with lower satisfaction scores, with a tail towards higher scores. The kurtosis of -1.030 suggests a leptokurtic distribution, indicating a more peaked and narrower shape than a normal distribution (Dave Longhorn, 2023c).

1. The descriptive statistics reveal notable differences in satisfaction scores between employees who stayed with the company and those who left. For those who stayed (N = 11,428), the mean satisfaction level is higher (0.665) compared to those who left (N = 3,571) with a mean satisfaction level of 0.438. The median and mode for those who stayed are also higher, suggesting a relatively more central distribution of satisfaction scores. The standard deviation for those who stayed is lower (0.217) compared to those who left (0.264), indicating less variability in satisfaction scores among employees who remained with the

company. Skewness values imply a slightly leftward skew for employees who stayed (negative skewness of -0.605) and a slight rightward skew for those who left (positive skewness of 0.293). Kurtosis indicates thinner tails for stayers (near-zero kurtosis of -0.211) and relatively lighter tails for leavers (negative kurtosis of -1.030).

In summary, the descriptive statistics highlight that employees who stayed generally have higher and less variable satisfaction scores, with a distribution slightly skewed to the left. Those who left show lower, more variable scores with a distribution slightly skewed to the right and heavier tails.

# Task 2: Z score interpretations

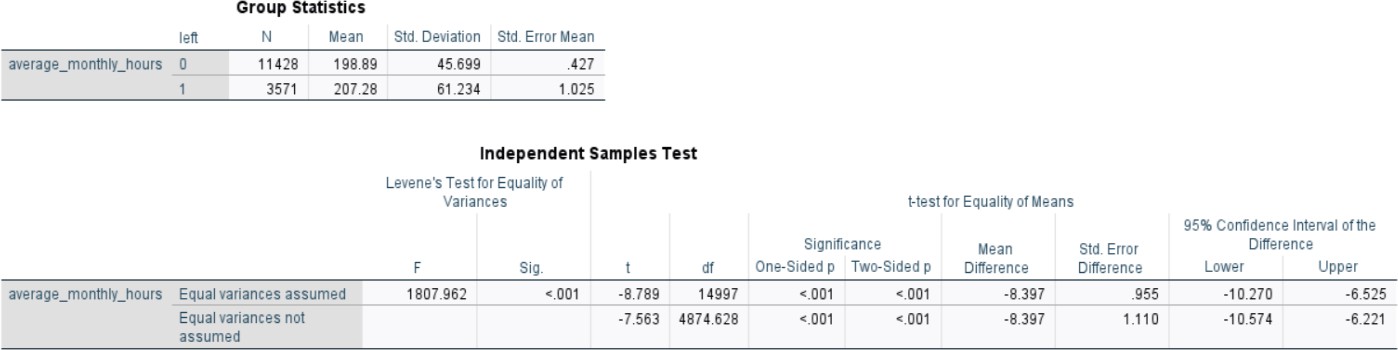
1. The descriptive statistics for the z scores of satisfaction level of all employees are



1. The Z-score of approximately -0.92907355547285 suggests that the first employee's satisfaction level is about 0.92907355547285 standard deviations below the mean satisfaction level of all employees in the dataset. This indicates that the first employee is less satisfied relative to the overall dataset. With a negative Z-score, this employee's satisfaction level falls? below the average, implying a lower level of satisfaction compared to the dataset's average satisfaction level.

**Tests of differences and associations.**

In the present case, to statistically know whether the employees who have left the company differ from those who stayed, according to their average monthly work hours, independent samples t-test could be helpful. Independent samples t-test is applicable only for two groups or samples (here the two groups are the employees who have left the company are one group, and the other group is those who stayed with the company), this distribution must be normal, moreover, the outcome variable must be continuous, i.e., it should be either in an interval or ratio scale (Dave Longhorn, 2023h). Here the continuous outcome variable is the average monthly work hours which is also in the ratio scale. Following are the test results,

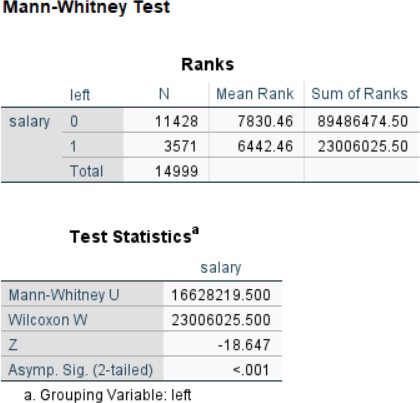


The obtained p-value is p<0.001, which is less than the assumed statistical significance threshold (alpha level) of 0.05.

According to Hatcher (2013), “In the new school approach, if the obtained p-value is less than the alpha level (threshold), it indicates to reject the null hypothesis and conclude that there was a statistically significant difference between the two conditions.”

Hence, as the p-value is less than the alpha level of 0.05, there is a statistically significant difference between the employees who have left the company differ from those who stayed in terms of average monthly work hours.

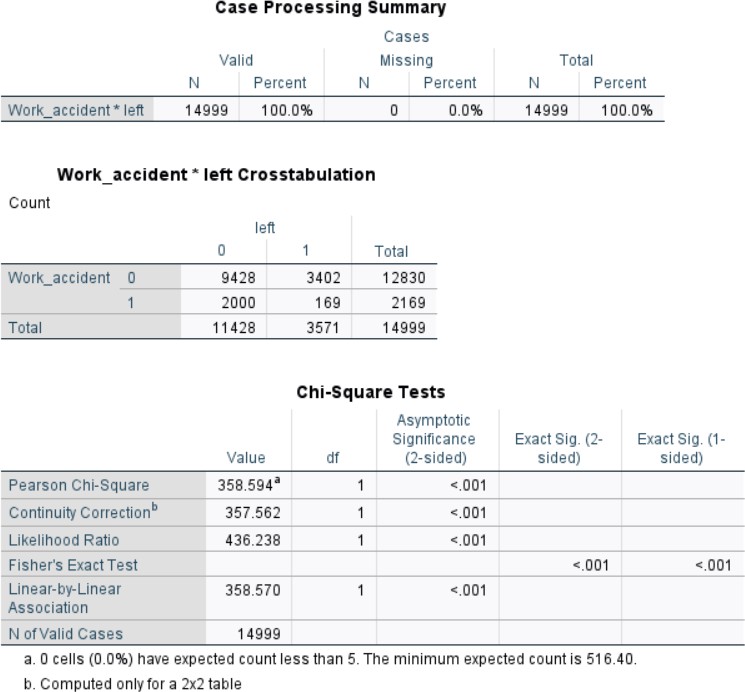
1. In the present case to statistically know whether the employees who have left the company differ from those who stayed, according to their salary level, the Mann-Whitney U test is performed which is a non-parametric statistical test of differences between two groups. This test applies to two groups (here the two groups are the employees who have left the company are one group, and the other group is those who stayed with the company), where distribution does not have to be normal, and the outcome variable can be continuous or ordinal (Dave Longhorn, 2023g). Here the ordinal outcome variable is the employees' salary level. Following are the test results,



Here the obtained p-value is p<0.001, which is less than the assumed statistical significance threshold (alpha level) of 0.05.

Hence, as the p-value is less than the alpha level of 0.05, there is a statistically significant difference between the employees who have left the company differ from those who stayed in terms of their salary level.

1. To check whether there is a statistical association between employees who met with a work accident and whether an employee has left the company, the Chi-square test can be helpful, which is a non-parametric statistical test of association. This test is applicable for two or more groups that have categorical or nominal variables (Dave Longhorn, 2023g). Here as the variables Work\_accident and left are nominal variables; this test is good to be applied. Following are the test results,



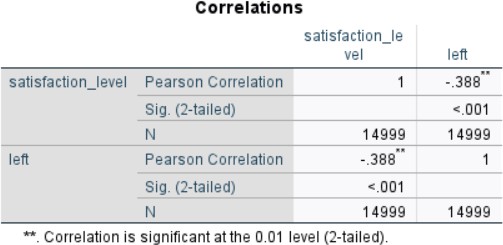
Here the obtained p-value is p<0.001, which is much less than the assumed statistical significance threshold (alpha level) of 0.05.

According to Hatcher (2013), “In the new school approach, if the obtained p-value is less than the alpha level (threshold), it indicates to reject the null hypothesis and conclude that there was a statistically significant association between the two conditions.”

Hence, as the p-value is less than the alpha level, there is a statistical association between employees who met with a work accident and employees leaving the company.

# Task 4: Correlations and regression analysis

1. Correlation between turnover (using the nominal variable "left") and satisfaction level



# Reporting the correlation results

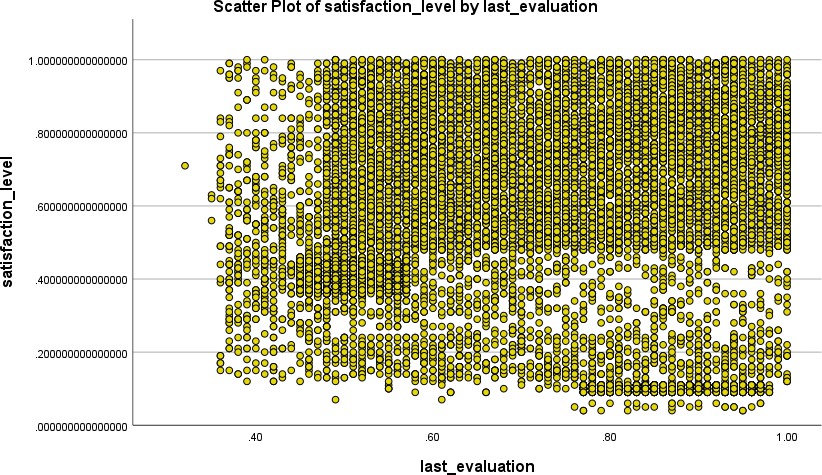
* + Statistical significance: The p-value is p<0.001.
  + Effect size: r = -0.388, this is the Pearson Correlation Coefficient computed for turnover (using the nominal variable “left”) and satisfaction level. Pearson Correlation Coefficient is computed when continuous or nominal variables are analyzed (Dave Longhorn, 2023f).

# Interpreting the correlation results

The correlation between turnover (represented by the nominal variable "left") and satisfaction level is statistically significant with a p<0.001 which is much less than the assumed statistical

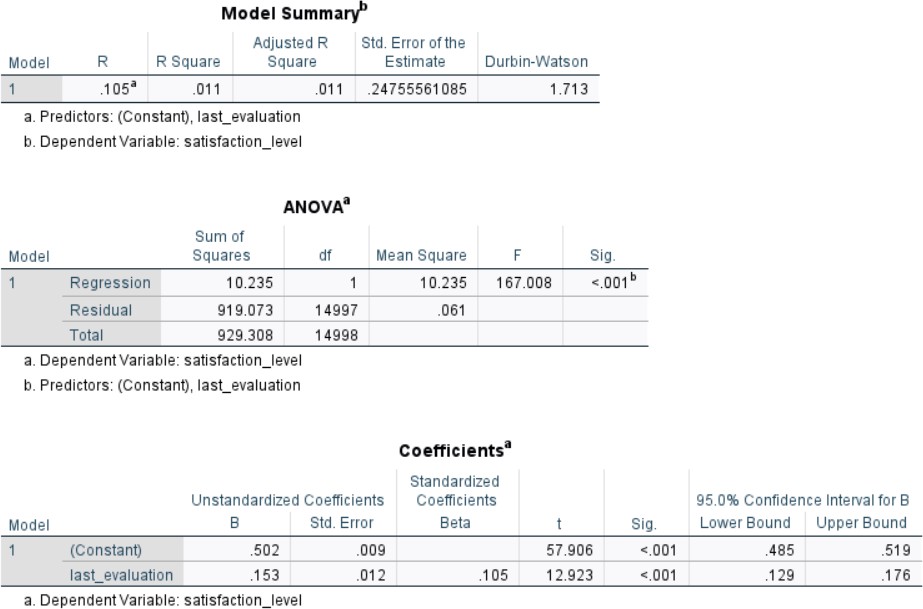
significance threshold (alpha level) of 0.05. The Pearson Correlation Coefficient (r) of -0.388 indicates a moderate negative correlation between these variables. This suggests that as the satisfaction level decreases, the likelihood of turnover increases. The effect size is meaningful, indicating a notable association between employee satisfaction and turnover within the dataset.

1. Scatterplot of employee’s last evaluation score and employee’s satisfaction level



The thick and dispersed distribution of points in the scatterplot of employee satisfaction level and last evaluation scores indicates that there is no clear relationship. It is possible that changes in one variable do not always translate into changes in the other if there is no apparent pattern. The bivariate regression output, which most likely displays a low or insignificant correlation coefficient, supports this observation. In conclusion, the finding that there is little to no significant association between employees' satisfaction levels in the dataset and their last evaluation scores is supported by both visual inspection and statistical analysis.

1. Bivariate linear regression between last evaluation (predictor) and satisfaction level (outcome)



The ANOVA table shows that the bivariate linear regression between the satisfaction level (outcome) and the last evaluation (predictor) points to a statistically significant model. With an F-statistic of 167.008 and a p<0.001, the model appears to be well-fitting and unlikely to be the result of chance. The last evaluation's standardized coefficient (β) is 0.105, suggesting a positive correlation with satisfaction level. The significance of this predictor is supported by the t-statistic for the most recent evaluation, which is 12.923, with a p-value < 0.001. The coefficient's significance is further shown by the fact that zero is excluded from the 95% confidence interval (0.129, 0.176).

With the R2 value of 0.011, the last evaluation may account for just 1.1% of the variation in satisfaction level. The low R2 suggests that additional factors not in the model contribute to the variation in satisfaction levels, even though the difference is statistically significant it shows little

autocorrelation.

In conclusion, the last evaluation and satisfaction level have a positive correlation that is confirmed by the statistical significance of the model. The low R2, however, indicates a poor overall model fit. This suggests that there may be other factors influencing satisfaction levels that are not considered by the current model, even though the most recent evaluation has a statistically significant impact on satisfaction level. Nevertheless, it does not account for a significant portion of the variance in employee satisfaction.

# Summarize the learnings about employee retention and satisfaction from the analysis, as well as the other factors that might be related to retention or satisfaction

1. The analysis suggests that the component most closely associated with retention is satisfaction level. A moderately negative correlation between turnover and satisfaction levels is revealed by the correlation study, meaning that when employee satisfaction declines, there is a greater chance of their leaving the organization. Furthermore, the descriptive statistics demonstrate significant disparities in satisfaction ratings between departing and remaining staff members, with the latter group exhibiting higher satisfaction levels. Consequently, the dataset shows that employee satisfaction level is a major factor impacting retention, indicating that attempts to raise employee satisfaction may result in higher retention rates.
2. The variables that seem to be most associated with satisfaction are turnover (shown by the nominal variable "left") and the results of the most recent evaluation. A moderately negative correlation between turnover and satisfaction level is revealed by the correlation analysis, suggesting that less satisfied employees are more likely to leave the organization. Furthermore, a statistically significant positive association between satisfaction level and last evaluation scores is suggested by the bivariate linear regression, suggesting a relationship between higher evaluation scores and greater satisfaction levels. These results emphasize that

within the dataset, employee satisfaction is significantly influenced by turnover and last evaluation scores.

1. The analysis reveals that to increase employee satisfaction and retention, several important areas need more research. First and foremost, the HR Director ought to investigate the variables affecting satisfaction levels in greater detail than is possible with the current dataset. To investigate additional factors like work-life balance, professional growth prospects, and job role satisfaction, a new survey may need to be conducted. Furthermore, it is recommended that the HR Director look into the precise causes of employee turnover, with a special emphasis on areas where leaving employees expressed lower levels of satisfaction. Comprehending these reasons can facilitate the efficient customization of retention strategies. Thirdly, gathering qualitative information through focus groups or interviews with both current and former workers can reveal important details about their viewpoints, experiences, and recommendations for development. Through more research and analysis, the HR Director can address these areas in-depth and create focused interventions that will improve employee satisfaction generally and retention rates inside the company.

References

Dave Longhorn. (2023c, December 24). *AA5221 - Week 3 - Descriptive Statistics* [Video].

YouTube. <https://www.youtube.com/watch?v=ziCIn4f_9ys>

Dave Longhorn. (2023f, December 24). *AA5221 - Week 5 - Correlation* [Video]. YouTube. <https://www.youtube.com/watch?v=0xFS_667PGU>

Dave Longhorn. (2023g, December 24). *AA5221 - Week 5 - Tests of Association (Chi- Squared test)* [Video]. YouTube. <https://www.youtube.com/watch?v=Tq_Ft9-dvdg>

Dave Longhorn. (2023g, December 24). *AA5221 - Week 5 - Tests of Differences (Mann- Whitney U test)* [Video]. YouTube. <https://www.youtube.com/watch?v=bz8rkFnmOVU>

Dave Longhorn. (2023h, December 24). *AA5221 - Week 5 - Tests of Differences (t-test)*

[Video]. YouTube. <https://www.youtube.com/watch?v=HC6BdrJnzz8>

Hatcher, L. (2013). *Advanced statistics in research: Reading, Understanding, and Writing Up Data Analysis Results*. Shadow Finch Media LLC.