**Impacts of Leisure Time on Daily Routines**

A PROJECT REPORT

*Submitted by*

**Ms. Shruti Premsagar Ushire**

*in partial fulfillment for the award of the degree*

*of*

**M.Sc. Statistics Part 2**

ACADEMIC YEAR 2023 – 2024

*Submitted to*



**DEPARTMENT OF STATISTICS**

**RAMNIRANJAN JHUNJHUNWALA COLLEGE OF ART’S,**

**SCIENCE & COMMERCE (AUTONOMOUS),**

**GHATKOPAR (W)**

**RAMNIRANJAN JHUNJHUNWALA COLLEGE OF ART’S, SCIENCE**

**& COMMERCE (AUTONOMOUS), GHATKOPAR (W)**



***(Affiliated to University of Mumbai)***

**CERTIFICATE**

*This is to certify that the project entitled* ***Impacts of Leisure Time on Daily Routines*** *bonafide work of* ***Ms. Shruti Premsagar Ushire*** *bearing seat no.* ***906*** *during the year 2023-2024 in partial fulfillment of the requirements for the award of Degree Master of Science in Statistics.*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Signature of Internal Guide**   **Signature of Co-ordinator**

(Mr. Vikas Dubey) (Mr. Jaishankar Singh)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Seal of the College**   **Signature of Examiner**

**ACKNOWLEDGEMENT**

A lot of efforts have been taken in this project; however, this work would not have reached its completion and would not have been possible without the kind support and help of many individuals, so we would like to extend our sincere thanks to all of them. We have learned a lot from working on this research project with different participants and gained valuable skills and insights that will help us in our future careers.

First and foremost, we would like to thank the Department of Statistics, Ramniranjan Jhunjhunwala College, Mumbai for providing us with the necessary resources and facilities to conduct this research. The department's commitment to excellence in teaching and research has been a constant source of inspiration throughout our academic journey.

We would like to extend my heartfelt appreciation to our project mentor Mr. Vikas Dubey, whose unwavering support and guidance have been indispensable in shaping this project. He has provided us with invaluable insights and feedback and has been a constant source of encouragement throughout the process.

We would also like to express our gratitude to Prof. Jaishankar Singh whose expertise and guidance have been invaluable throughout our academic career. He has been a constant source of inspiration, motivation, and encouragement, and his insights and feedback have been instrumental in shaping the direction of this report.

Finally, I would like to express my gratitude to our family and friends for their unwavering support and encouragement. Their love and encouragement have been a constant source of strength throughout my academic journey.

Thank you all for your support, guidance, and encouragement. Without you, this project would not have been possible.

**INDEX**

| **Sr. No** | **Topic** | **Page No** |
| --- | --- | --- |
| 1 | Introduction |  |
| 2 | Overview |  |
| 3 | Objectives & Methodology |  |
| 4 | Multiple Linear Regression |  |
| 5 | Binary Logistic Regression |  |
| 6 | Factor Analysis |  |
| 7 | Association Rules |  |
| 8 | Decision Tree |  |
| 9 | Sentiment & Text Analysis |  |
| 10 | Conclusion / Recommendations |  |
| 11 | Questionnaire |  |
| 12 | Coding of the data |  |
| 13 | Scope |  |
| 14 | Bibliography |  |

**ABSTRACT**

In today's dynamic and competitive world, many individuals are exploring additional sources of income alongside their full-time employment. This research project delves into the ramifications of engaging in side income pursuits on both personal and professional spheres. Through a comprehensive investigation, it seeks to elucidate the potential advantages and challenges associated with such endeavors, aiming to offer a nuanced understanding of their impact on individuals' lives.

The objectives of the study encompass a wide array of dimensions:

* Motivations and Decision Factors: Exploring the driving forces behind individuals' pursuit of side income sources and the factors shaping their decision-making process.
* Financial Well-being Impact: Assessing how side income activities influence financial stability, encompassing personal income, family earnings, and strategies for financial management.
* Work-Life Balance: Investigating the effects of side income pursuits on individuals' allocation of time, stress levels, and overall satisfaction with their work and personal lives.
* Career Development: Analyzing the relationship between side income ventures and professional growth, including opportunities for skill enhancement, networking, and potential transitions into full-time entrepreneurship.
* Challenges and Risks: Evaluating the potential drawbacks associated with managing side income alongside primary employment, encompassing impacts on physical and mental health, conflicts of interest, and work performance.
* Required Skills and Competencies: Identifying the specific proficiencies necessary for successful engagement in side income activities, and exploring the role of education, training, and support systems in fostering individuals' capabilities.

By scrutinizing the influence of side income sources on both personal and professional realms, this research project aims to contribute to existing literature on work-life balance, career advancement, and personal finance. The insights garnered are expected to offer valuable guidance for individuals contemplating or engaged in side income pursuits, as well as for organizations seeking to support their employees in navigating the evolving landscape of work and income generation.

In summary, this study endeavors to deepen our understanding of the motivations and impacts of side income sources on individuals' lives, offering actionable insights for individuals and organizations alike as they navigate the intricacies of modern work and income dynamics.

**Introduction: Impacts of Leisure Time on Daily Routine**

In today's fast-paced and demanding world, where individuals juggle multiple responsibilities and commitments, the significance of leisure time cannot be overstated. Leisure time, defined as the period outside of obligatory work or other commitments, represents a precious opportunity for relaxation, rejuvenation, and personal fulfillment. How individuals choose to allocate and utilize their leisure time can profoundly influence their daily routines, shaping various aspects of their lives.

Moreover, the impact of leisure time on daily routines extends beyond mere relaxation or entertainment—it plays a pivotal role in determining overall well-being, productivity, and life satisfaction. As individuals navigate the complexities of modern life, effectively managing leisure time becomes increasingly crucial for maintaining a healthy work-life balance, managing stress levels, and fostering personal growth.

This study seeks to explore the intricate relationship between leisure time activities and daily routines, aiming to uncover the multifaceted impacts and implications for individuals in today's society, by examining how leisure time is spent and its effects on time management, stress levels, and overall satisfaction with life, we can gain valuable insights into optimizing daily routines for improved well-being and productivity. Furthermore, in the context of evolving work structures and lifestyle preferences, understanding the dynamics of leisure time becomes even more pertinent. With the rise of remote work, flexible schedules, and the gig economy, individuals have greater autonomy over how they allocate their leisure time, presenting both opportunities and challenges in managing daily routines effectively. Through a comprehensive analysis utilizing both qualitative and quantitative methodologies, this study aims to elucidate the complexities of leisure time management and its impact on daily routines. By uncovering patterns, trends, and associations, we seek to provide actionable insights for individuals, organizations, and policymakers alike, facilitating informed decision-making and fostering a culture of well-being and productivity in today's society.

**Introduction:**

Time is often overlooked but holds immense value in our lives, shaping our daily experiences and emotions.

Each person's daily routine varies based on factors like age, location, and occupation, making time management a challenge.

Effective time allocation is crucial, ensuring that important tasks are tackled at the right moments.

While time is intangible and beyond our control, we have the power to make conscious choices in the present.

The saying "Time is Money" reflects our collective understanding of the importance of time, yet we often squander it unknowingly.

This analysis aims to delve into how people spend their time each day, shedding light on various influencing factors.

By examining demographics such as age, occupation, and location, we can understand what impacts individual satisfaction with their daily routines.

Data was collected using a method called 'Simple Random Sampling without replacement,' ensuring diverse representation.

The statistical analyses conducted are interconnected, providing a cohesive understanding of time management patterns.

Our goal is not only to identify areas for improvement but also to empower individuals to make better use of their time.

Through mindful reflection and conscious effort, we can maximize the value of our time and strive for temporal mastery.

**Objectives and Methodology:**

| To understand how the daily routine affects the leisure time of the individuals daily. | Multiple Linear Regression |
| --- | --- |
| To understand which of the demographics are accountable or not for improvement in daily or weekend routine | Binary Logistic Regression |
| To extract the factors considered important by the people for their satisfaction or dissatisfaction with a regular day. | Factor Analysis |
| To identify what different types of activities people mostly go through in a day and how they are associated. | Association Rules |
| To understand which characteristics of the demographics result in improvement in a daily routine is needed. | Decision Tree |
| To study the sentiments of the people on their daily routine and how they want to make an improvement which will assist them. | Sentiment and Text Analysis |

**Testing of Hypothesis:**

Hypothesis testing is a statistical technique used to evaluate the validity of assumptions or claims about a population based on a sample. It allows us to make inferences about the characteristics of a larger population based on a smaller sample of data. In many cases, the data we work with may not follow a normal distribution, which means that we cannot use traditional parametric tests such as t-tests or ANOVA. In these cases, non-parametric tests such as the Wilcoxon test, Kruskal-Wallis test, and Dunn test are used to analyze the data.

The Wilcoxon test is a non-parametric test used to compare two independent samples. It is a powerful alternative to the t-test when the normality assumption is violated. The Kruskal Wallis test is another non-parametric test used to compare three or more independent samples. It is similar to the ANOVA test, but does not require the assumption of normality. The Dunn test is a post-hoc test used to compare multiple groups following a Kruskal-Wallis test.

In this scenario, since the data does not follow a normal distribution, we will perform non parametric tests such as the Wilcoxon test, Kruskal-Wallis test, and Dunn test to analyze the data. These tests are robust and powerful alternatives to traditional parametric tests and can provide valuable insights into the characteristics of the population being studied.

**t-test:**

A t-test is a statistical hypothesis test that is used to compare the means of two populations or samples, when the standard deviation of the population is not known and the sample size is relatively small. The t-test is a parametric test that assumes that the data is normally distributed. It is a widely used statistical test in many fields, including medicine, social sciences, and engineering.

The t-test is based on the t-distribution, which is a probability distribution that is used to estimate the population mean when the sample size is small and the standard deviation is not known. The t-distribution is similar to the normal distribution, but it has more spread and heavier tails.

There are two types of t-tests: the independent samples t-test and the paired samples t-test.

* The independent samples t-test is used when we want to compare the means of two independent samples. This could be, for example, comparing the average scores of two different groups of students on a standardized test.
* The paired samples t-test is used when we want to compare the means of two related samples, such as the performance of the same group of students before and after receiving a particular intervention.

**Assumptions of t-test:**

1. **Normality:** The t-test assumes that the data follows a normal distribution. This means that the data has a symmetric distribution around the mean and the majority of the data is within one standard deviation of the mean.
2. **Independence:** The samples being compared should be independent of each other, meaning that the observations in one sample should not be related to the observations in the other sample.
3. **Equal variances:** The t-test assumes that the variance of the two populations being compared is equal. This can be tested using the F-test or Levene's test for equality of variances.

Violation of these assumptions may affect the accuracy of the t-test and lead to incorrect conclusions. In cases where the data does not meet these assumptions, non-parametric tests such as the Wilcoxon rank-sum test may be more appropriate.

Hypothesis:

The hypothesis of a t-test is a statement about the population or populations being studied. Specifically, the null hypothesis (H0) states that there is no significant difference between the means of the two populations or samples, while the alternative hypothesis (Ha) states that there is a significant difference between the means.

For an independent samples t-test, the null and alternative hypotheses are:

H0: 𝜇1=𝜇2 & H1: 𝜇1≠𝜇2

where μ1 and μ2 are the means of the two populations being compared.

For a paired samples t-test, the null and alternative hypotheses are:

H0: 𝜇𝑑=0 & H1: 𝜇𝑑≠0

where μd is the mean difference between the paired samples.

In both cases, the null hypothesis assumes that there is no difference between the means of the populations being compared, while the alternative hypothesis suggests that there is a significant difference.

The t-test calculates a test statistic, t, which measures the difference between the sample means relative to the variation within the samples. If the calculated t-value is greater than the critical value (determined by the significance level and degrees of freedom), then the null hypothesis is rejected in favor of the alternative hypothesis.

**Test Statistics**

The t-statistic is a test statistic used in the t-test to determine whether the difference between two sample means is statistically significant. It is calculated by dividing the difference between the two sample means by the standard error of the difference. For an independent samples t-test, the t-statistic is calculated as:

t = (x1 - x2) / (s^2pooled \* (1/n1 + 1/n2))^(1/2)

where x1 and x2 are the sample means of the two populations being compared, s^2pooled is the pooled variance of the two samples, n1 and n2 are the sample sizes of the two populations being compared.

For a paired samples t-test, the t-statistic is calculated as:

t = (d - μd) / (sd / √n)

where xd is the sample mean of the differences between the paired samples, μd is the hypothesized mean difference (usually 0), sd is the standard deviation of the differences, and n is the sample size.

The t-statistic follows a t-distribution with degrees of freedom (df) equal to the sum of the sample sizes minus 2 for an independent samples t-test and n-1 for a paired samples t-test. The critical value of the t-distribution is determined by the significance level (α) and the degrees of freedom.

If the calculated t-statistic is greater than the critical value, the null hypothesis is rejected in favor of the alternative hypothesis, indicating that there is a significant difference between the means of the populations being compared. If the calculated t-statistic is less than the critical value, the null hypothesis cannot be rejected, indicating that there is no significant difference between the means of the populations being compared.

**Gender**

Hypothesis:

H0: There is no significant difference in the mean Leisure Time in Female and Male.

H1: There is a significant difference in the mean Leisure Time in Female and Male.

**Preliminary test to check independent t-test assumptions**

**Assumption 1:** Are the two samples independents?

Yes, since the samples from men and women are not related.

**Assumption 2:** Do the data from each of the 2 groups follow a normal distribution?

Hypothesis:

H0: Data follow normal distribution

H1: Data doesn’t follow normal distribution

Output:

| Shapiro-Wilk Normality test for leisure time in female | |
| --- | --- |
| W = 0.93623 | p-value = 3.221e-09 |

| Shapiro-Wilk Normality test for leisure time in male | |
| --- | --- |
| W = 0.90029 | p-value = 2.072e-11 |

**Conclusion:** From the output, the p-values are less than the significance level 0.05. Which implies that the distributions of the data are significantly different from the normal distribution. Hence, we can conclude that there is no normality.

In cases where the normality assumption is not satisfied, the Wilcoxon rank-sum test (also known as the Mann-Whitney U test) can be used as an alternative to the independent t-test. The Wilcoxon test is a non-parametric test that does not require the assumption of normality, making it a useful tool for analyzing non-normally distributed data.

**Wilcoxon Sign Rank Test:**

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e., it is a paired difference test). It can be used as an alternative to the paired Student's t-test (also known as "t-test for matched pairs" or "t-test for dependent samples") when the distribution of the difference between two samples' means cannot be assumed to be normally distributed. A Wilcoxon signed-rank test is a nonparametric test that can be used to determine whether two dependent samples were selected from populations having the same distribution.

The Wilcoxon test works by comparing the ranked values of the two groups being compared. Specifically, the test calculates the sum of the ranks for one of the groups, and then compares this value to the sum of the ranks for the other group. The test statistic used in the Wilcoxon test is the U-statistic, which is calculated based on these rank sums. If the calculated U-statistic is less than the critical value at a chosen significance level (usually α = 0.05), the null hypothesis is rejected in favor of the alternative hypothesis, indicating that there is a significant difference between the medians of the two groups being compared.

If the calculated U-statistic is greater than the critical value, the null hypothesis cannot be rejected, indicating that there is no significant difference between the medians of the two groups being compared.

Assumptions:

1. Data are paired and come from the same population.
2. Each pair is chosen randomly and independently.
3. The data are measured on at least an interval scale when, as is usual, within-pair differences are calculated to perform the test (though it does suffice that within-pair comparisons are on an ordinal scale)

Hypothesis:

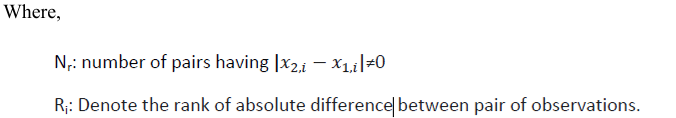
H0: The difference between the pairs follows a symmetric distribution around zero (Mx=MY)

H1: The difference between the pairs does not follow a symmetric distribution around zero. (Mx≠MY)

Test Statistics is,



Where,



Hypothesis:

H0: There is no significant difference in the median leisure time in Female and Male. (Mx=My)

H1: There is a significant difference in the median leisure time in Female and Male. (Mx≠My)

Output:

| Wilcoxon rank sum test with continuity correction | |
| --- | --- |
| W = 33107 | p-value = 0.1719 |

**Conclusion:** The p-value of the test is 0.1719, which is greater than the significance level alpha = 0.05. We can conclude that the median leisure time in females is not significantly different from the median leisure time in Males.

**Similarly for other binary groups**

Hypothesis:

H0: There is no significant difference in the median leisure time and categories of different binary variables. (Mx=My)

H1: There is a significant difference in the median leisure time and categories of different binary variables. (Mx≠My)

| **Wilcoxon rank sum test with continuity correction** | | |
| --- | --- | --- |
| **Binary Group Variables** | **Test Statistics** | **P - Value** |
| Marital Status | 26406 | 0.1135 |
| Type of family | 27599 | 0.6651 |

**Conclusion:**  The p-value of each binary group variable is greater than the significance level alpha=0.05. Hence, there is no significant difference in categories of each binary group variable.

**One Way ANOVA:**

One-way analysis of variance (ANOVA) is a statistical test used to determine whether there is a significant difference between the means of three or more independent groups. The test compares the variance between groups to the variance within groups to determine if the differences in means are significant.

In one-way ANOVA, the dependent variable is continuous and the independent variable has three or more levels or categories. The test involves calculating the F-ratio by dividing the variance between groups by the variance within groups. If the calculated F-value is greater than the critical value at a chosen significance level (usually α = 0.05), the null hypothesis is rejected in favor of the alternative hypothesis, indicating that there is a significant difference between the means of the groups. It's important to note that the test assumes that the groups are independent and normally distributed with equal variances, which may not always be the case in real-world scenarios. When these assumptions are not met, alternative tests such as non parametric tests may be more appropriate.

**Assumptions:**

The one-way ANOVA has several assumptions that need to be met for the results to be reliable. The main assumptions are:

* The variances of the dependent variable are equal across the groups.
* The dependent variable is continuous and normally distributed in each group.
* The observations are independent within and between groups.

**Hypothesis:**

The null hypothesis of the one-way ANOVA is that there is no significant difference between the means of the groups, while the alternative hypothesis is that there is at least one group that differs significantly from the others.

H0: The means of all groups are equal.

Ha: At least one of the means of the groups is different from the others.

**Test Statistics:**

The test statistic used in one-way ANOVA is the F-ratio, which is calculated by dividing the variance between groups by the variance within groups. The formula for the F-ratio is:

F = MS\_between / MS\_within

where MS\_between is the mean sum of squares between groups and MS\_within is the mean sum of squares within groups.

The mean sum of squares between groups is calculated by dividing the sum of squares between groups (SS\_between) by the degrees of freedom between groups (df\_between). The sum of squares between groups measures the variation between the sample means of the groups and is calculated as:

SS\_between = Σni (Yi - YG)²

where ni is the sample size of the ith group, Yi is the mean of the ith group, YG is the grand mean, and Σni is the total sample size.

The degrees of freedom between groups are equal to the number of groups minus one:

df\_between = k – 1

where k is the number of groups.

The mean sum of squares within groups is calculated by dividing the sum of squares within groups (SS\_within) by the degrees of freedom within groups (df\_within). The sum of squares within groups measures the variation within each group and is calculated as:

SS\_within = Σ (Yij - Yi)²

where Yij is the jth observation in the ith group, Yi is the mean of the ith group, and Σ(Yij - Yi)² is the sum of squares within each group.

The degrees of freedom within groups are equal to the total sample size minus the number of groups:

df\_within = N - k

where N is the total sample size.

The F-ratio is then compared to the F-distribution with df\_between and df\_within degrees of freedom to determine whether the null hypothesis should be rejected or not. If the calculated F-value is greater than the critical value at a chosen significance level (usually α = 0.05), the null hypothesis is rejected in favor of the alternative hypothesis, indicating that there is a significant difference between the means of the groups.

**Education Level**

Hypothesis:

H0: The means of all groups are equal.

H1: At least one of the means of the groups is different from the others.

**Preliminary test to check one way ANOVA assumptions:**

The ANOVA test assumes that the data are normally distributed and the variance across groups are homogeneous.

**Assumption 1**: The variances of the dependent variable are equal across the groups.

**Check the homogeneity of variance assumption**

The homogeneity of variance assumption in one-way ANOVA states that the variance of the dependent variable is equal across all groups. This assumption can be checked using the Levene's test or Bartlett's test.

Levene's test is a statistical test used to assess the equality of variances across groups. The null hypothesis of Levene's test is that the variances are equal across all groups. If the p-value of Levene's test is less than the chosen significance level (usually α = 0.05), the null hypothesis is rejected, indicating that the variances are not equal.

Bartlett's test is another statistical test used to assess the equality of variances across groups. The null hypothesis of Bartlett's test is that the variances are equal across all groups. If the p value of Bartlett's test is less than the chosen significance level (usually α = 0.05), the null hypothesis is rejected, indicating that the variances are not equal.

Output-

| **Levene's Test for Homogeneity of Variance** | | | |
| --- | --- | --- | --- |
|  | Df | F value | P-value (>F) |
| group | 5 | 1.0916 | 0.3641 |

**Conclusion:** The p-value is not less than the significance level of 0.05. This means that there is no evidence that the variance across groups is statistically significantly different. Therefore, we can assume the homogeneity of variances in the different groups.

**Assumption 2**: Are the data from each of the 2 groups follow a normal distribution?

Hypothesis:

H0: Data follow normal distribution

H1: Data doesn’t follow normal distribution

Output-

| Shapiro-Wilk Normality test | |
| --- | --- |
| W = 0.92197 | p-value = 2.17e-15 |

**Conclusion:**

The p-value obtained from the Shapiro-Wilk test to check the normality of the data is 2.17e-15, which is less than the significance level of 0.05. This indicates strong evidence against the null hypothesis of normality, suggesting that the data is not normally distributed. Therefore, the assumption of normality for the data is violated, and any statistical tests that require normality assumption may not be appropriate to use. In such a case, non-parametric tests such as Kruskal-Wallis test, or Friedman test could be used as alternatives to the parametric tests.

**Kruskal-Wallis test:**

The Kruskal-Wallis test is a non-parametric statistical test used to compare three or more independent groups that are not normally distributed. It is an alternative to the one-way ANOVA test, which assumes that the data is normally distributed.

The Kruskal-Wallis test is based on the rank sums of the data in each group. It ranks all the data in the combined sample from lowest to highest, then sums the ranks for each group separately. The test compares the sum of ranks between groups to determine if there is a significant difference in the distribution of the data among the groups.

**Assumptions:**

* Independent samples: The groups being compared are independent of each other.
* Random samples: The data is sampled randomly from the population.
* The dependent variable is measured on an ordinal or continuous scale.
* The shape of the distributions of the groups being compared is not assumed to be normal.
* Homogeneity of variance: The variance of the dependent variable is equal across groups.

**The hypothesis of the Kruskal-Wallis test is**

Hypothesis:

H0: The populations from which the groups are drawn have the same distribution.

H1: At least one population has a different distribution than the others.

Output-

| **Kruskal-Wallis rank sum test** | |
| --- | --- |
| Kruskal-Wallis chi-squared | 5.7486 |
| df | 5 |
| p-value | 0.3315 |

**Conclusion:** The p-value obtained from the Kruskal-Wallis test is 0.3315, it indicates that there is significant difference between the groups being compared at the commonly chosen significance level of 0.05. Therefore, we fail to reject the null hypothesis, which states that the populations from which the groups are drawn have the same distribution.

The Kruskal-Wallis test has determined that there are significant differences among the groups being compared, performing a Dunn test (or Dunn's post hoc test) is a suitable next step to identify which specific group comparisons are statistically significant.

**Dunn Test:**

The Dunn test, also known as the Dunn's post hoc test or Dunn's multiple comparison test, is a non-parametric statistical test used to perform pairwise comparisons between multiple groups after conducting a Kruskal-Wallis test or a one-way analysis of variance (ANOVA) on ranked data. It is specifically designed to identify which groups are significantly different from each other.

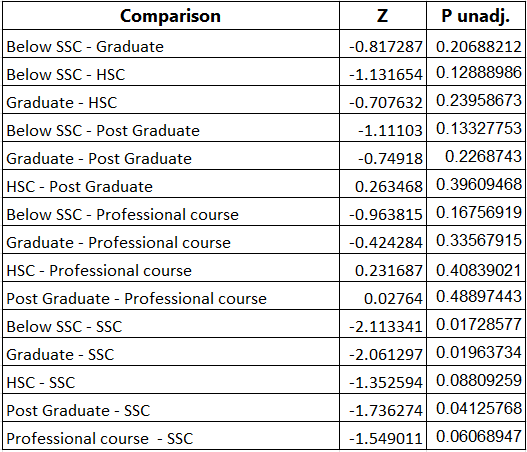
The Dunn test is particularly useful when the assumptions of parametric tests, such as normality or equal variances, are not met. It allows for non-parametric analysis of multiple group comparisons, providing insights into the differences between groups without making distributional assumptions.

**The hypothesis for the Dunn test can be stated as follows:**

**Null Hypothesis (H₀):** There is no significant difference between the two groups being compared.

**Alternative Hypothesis (H₁):** There is a significant difference between the two groups being compared.

Output:



**Similarly, for other variables which contain more than two categories.**

The hypothesis of the Kruskal-Wallis test is that Hypothesis:

H0: The populations from which the groups are drawn have the same distribution.

H1: At least one population has a different distribution than the others.

Output-

| **Kruskal-Wallis rank sum test** | | | |
| --- | --- | --- | --- |
|  | Kruskal-Wallis chi-squared | Df | p-value |
| Occupation | 40.451 | 12 | 6.051e-05 |
| Family annual income | 4.9964 | 5 | 0.4163 |

**Conclusion:**

The p-value obtained from the Kruskal-Wallis test is greater than the level of significance for Family annual income; it indicates that there is no significant difference between the groups being compared at the commonly chosen significance level of 0.05. Therefore, we fail to reject the null hypothesis, which states that the populations from which the groups are drawn have the same distribution.

The p-value obtained from the Kruskal-Wallis test is less than the level of significance for Occupation; it indicates that there is significant difference between the groups being compared at the commonly chosen significance level of 0.05. Therefore, we reject the null hypothesis, which states that at least one population has a different distribution than the others.

**Chi - Square Test**

A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration. As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables.

A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable. Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal. They cannot have a normal distribution since they can only have a few particular values.

For example, a meal delivery firm in India wants to investigate the link between gender, geography, and people's food preferences.

It is used to calculate the difference between two categorical variables, which are:

* As a result of chance or
* Because of the relationship

### **Assumptions:**

1. **Independence of Observations:** The observations used to calculate the chi-square statistic should be independent of each other.
2. **Random Sampling:** The data should be obtained through random sampling, or the sampling method should be such that the observations are representative of the population.
3. **Expected Cell Frequencies:** The expected frequency count for each cell in the contingency table should be at least 5. If this assumption is violated, alternative tests or adjustments may be needed.

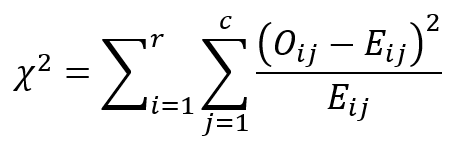
### **Hypotheses:**

* Null Hypothesis (H0): There is no association between the categorical variables.
* Alternative Hypothesis (H1): There is an association between the categorical variables.

### **Test Statistic:**

The test statistic for the chi-square test is calculated based on the differences between the observed and expected frequencies in a contingency table.

For a contingency table with *r* rows and *c* columns, the chi-square test statistic () is calculated as:



Where:

* Oij =Observed frequency in cell (*i*,*j*)
* Eij = Expected frequency in cell (*i*,*j*)

Under the null hypothesis, the chi-square statistic follows a chi-square distribution with (*r*−1)×(*c*−1) degrees of freedom.

Once the chi-square statistic is calculated, it is compared to the critical value from the chi-square distribution (with appropriate degrees of freedom) or directly compared to the p-value. If the p-value is less than the chosen significance level (e.g., 0.05), then the null hypothesis is rejected, and it can be concluded that there is a significant association between the categorical variables.

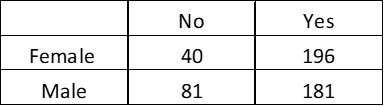
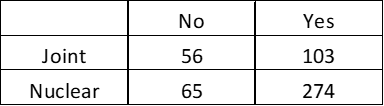
Using the chi square test we will check the association between the demographics and daily routine improvement decisions of the individuals.

| **Variable 1** | **Variable 2** | **P value** |
| --- | --- | --- |
| Gender | Daily routine improvement | p-value = 0.0004249 |
| Marital Status | Daily routine improvement | p-value = 2.488e-05 |
| Type of family | Daily routine improvement | p-value = 0.0001566 |
| Education level | Daily routine improvement | p-value = 3.114e-05 |
| Occupation | Daily routine improvement | p-value = 0.001747 |

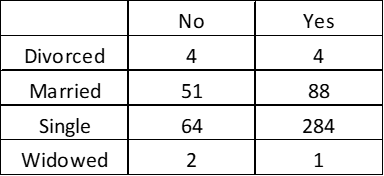
**Conclusion:** In the results of the chi square test we can see that the decision of the daily routine improvement of the individuals has a significant relationship with all the demographics of the individuals.

**Contingency tables:**

Gender - Daily routine improvement Type of Family - Daily routine improvement

Marital Status - Daily routine improvement Education level - Daily routine improvement



Occupation - Daily routine improvement



**Multiple Linear Regression**

The simplest of algorithms in statistics which attempts to predict the relationship between one continuous dependent variable and multiple independent variables.

The Multiple Linear Regression Model:

yi = β0 + β1 xi1 + β1 xi2 +…………. + β1 xin + ε; i= 1,2,3,4……………….n

Where:

* yi​ is the dependent or predicted variable
* β0 is the y-intercept, i.e., the value of y when both xi and x2 are 0.
* β1 and β2 are the regression coefficients representing the change in y relative to a one-unit change in xi1 and xi2, respectively.
* βp is the slope coefficient for each independent variable
* ϵ is the model’s random error (residual) term.

The model has certain assumptions that must hold.

These assumptions are stated as follows

1. Homoscedasticity

2. No Multicollinearity

3. No Outliers

4. Independent observations

5. Residuals must follow Normal Distribution

6. No Autocorrelation

To check that these assumptions hold, the following techniques were used.

| **Assumption** | **Technique** |
| --- | --- |
| Homoscedasticity | A plot of residuals vs fitted values |
| No Multicollinearity | Variance Inflation Factor |
| No Outliers | Cook’s Distance |
| Residuals ~ N (0, σ 2 ) | QQ plot of residuals |
| No Autocorrelation | A plot of ACF values vs LAG |

The hypothesis to test:

Ho: βi = 0 ; i = 1,2,3 …. 10

H1: βi ≠ 0 ; i = 1,2,3 …. 10

The Null hypothesis should be rejected because if all the βi’s are 0 then the model just contains the slope and there is no use in doing the regression analysis.

**Explanation of the terms:**

1. **Homoscedasticity:** It means equal variance in a layman’s language. It is important because we assume the model gives equal weights to all the observations and hence the variance around the regression line for all the predictor variables must be the same.
2. **Multicollinearity**: It means very high intercorrelation among the predictor variables. Therefore, it is termed as a disturbance. High multicollinearity makes the inferences made about the data less reliable.

1. **Outliers:** These are some specific observation(s) that lie at a far distance from the usual observations and may cause errors. Not all the outliers are a problem and it’s up to our choice how to deal with the outliers.

1. **Autocorrelation:** It is the correlation between elements of one series and another. When this is present the errors follow a certain pattern showing a trend.

Checking the assumptions for the Multiple Linear Regression Model.

**Factor Analysis**

Factor analysis is a data reduction technique that attempts to represent a set of observed variables X1, X2……,Xn in terms of a number of 'common' factors plus a factor that is unique to each variable. A smaller set of uncorrelated variables is obtained from the original set of correlated variables.

**Procedure for factor analysis**

Principal component analysis is a method for removing correlation among independent variables that are considered in multivariate regression analysis such as linear or logistic regression. The principal component method of extraction begins by finding a linear combination of variables that accounts for as much variation in the original variables as possible. It then finds another component that accounts for as much of the remaining variation as possible and is uncorrelated with the previous component. This continues until there are as many components as the original variables. Usually, a few components will account for most of the variation and these components can be used to replace the original variables.

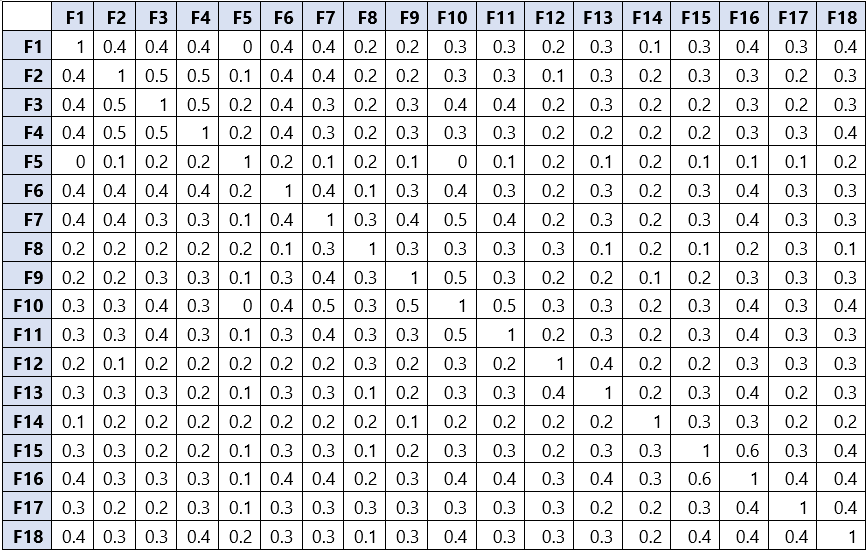
Variables used:

| F1) Your Attitude Toward Life | F10) Weekend satisfaction |
| --- | --- |
| F2) Mental Stability | F11) Promptness (delaying things) |
| F3) Time management | F12) Transport facilities |
| F4) Passion for Living | F13) Environment in your Locality / Area |
| F5) Difficult Interactions | F14) Unorthodox family |
| F6) Healthy Life | F15) Sufficient Resources |
| F7) Fulfill your enjoyments | F16) Fulfillment of desires |
| F8) Daily Struggle | F17) High Expectations toward life |
| F9) Dull / Social Neighborhood | F18) Work / Study Environment |

**Correlation analysis**

Correlation analysis was carried out to check linear relationship between various variables such as Strained Relationship, Orthodox family pressure etc.

**Correlation Matrix**

****

From the above table, we can see that there is a linear relationship among variables and hence

assumption of correlation between different variables considered for factor analysis is satisfied.

**Kaiser-Meyer-Olkin (KMO) measures and Bartlett’s test of Sphericity**

To determine the factor ability of an intercorrelation matrix we use the KMO and Bartlett's test of Sphericity.

**KMO** test measures the sampling adequacy which should be greater than 0.5. Then, we can say that the sample size is adequate for factor analysis.

| Kaiser-Meyer-Olkin Measure of Sampling Adequacy | 0.8968 |
| --- | --- |

**Bartlett’s test of Sphericity**

Under this test null hypothesis is given by

: Population correlation matrix is an identity matrix. (i.e. the variables are uncorrelated)

: Population correlation matrix is not an identity matrix.

| Bartlett’s test of Sphericity | Approx. Chi-Square | 2587.17 |
| --- | --- | --- |
| P - value | 0.00 |

Here p value is less than 0.05, therefore we reject and conclude that variables are correlated.

| Component | Eigen Value | % of Variance | Cumulative % |
| --- | --- | --- | --- |
| **F1** | 5.82865634 | 30.23 | 30.23 |
| **F2** | 1.33970978 | 6.95 | 37.18 |
| **F3** | 1.25710238 | 6.52 | 43.7 |
| **F4** | 1.13161542 | 5.87 | 49.57 |
| **F5** | 1.07335646 | 5.57 | 55.14 |
| **F6** | 1.00345544 | 5.2 | 60.34 |
| **F7** | 0.96473248 | 5 |  |
| **F8** | 0.88893396 | 4.61 |  |
| **F9** | 0.75909035 | 3.94 |  |
| **F10** | 0.72875263 | 3.78 |  |
| **F11** | 0.67906058 | 3.52 |  |
| **F12** | 0.62976319 | 3.27 |  |
| **F13** | 0.59895149 | 3.11 |  |
| **F14** | 0.57630473 | 2.99 |  |
| **F15** | 0.53238736 | 2.76 |  |
| **F16** | 0.46146554 | 2.39 |  |
| **F17** | 0.43355119 | 2.25 |  |
| **F18** | 0.39329569 | 2.04 |  |

**Eigenvalues**

Eigenvalues are the variances of the principal components. Because we conducted our principal component analysis on the correlation matrix, the variables are standardized, which means that each variable has a variance of 1, and the total variance is equal to the number of variables used in the analysis, in this case, 20.

**Percentage Of Variance**

This column contains the % of variance accounted for by each principal component.

**Cumulative Percentage**

This column contains the cumulative % of variance accounted by the current and all preceding principal components. E.g.- The 6th row shows a value of 60.34. This means the first 6 components together account for 54.13%.

**Scree plot**

A scree plot displays Eigenvalues associated with the component of factors in descending order versus the number of components or factors.

This scree plot shows that the 6 factors explain most of the variability because the line starts to be parallel to the x-axis after factor 6. The remaining factors explain a very small proportion of the variability and are likely not important.

**Rotated Component Matrix:**

|  | **Factor1** | **Factor 2** | **Factor3** | **Factor 4** | **Factor 5** | **Factor 6** |
| --- | --- | --- | --- | --- | --- | --- |
| **F2** | 0.6661 | 0.140755 | 0.16507 | 0.0497 | 0.008658 | 0.066337 |
| **F4** | 0.6513 | 0.09658 | 0.17063 | 0.05949 | 0.126286 | 0.193859 |
| **F3** | 0.58805 | 0.09745 | 0.32038 | 0.15577 | 0.238766 | -0.11047 |
| **F1** | 0.49349 | 0.307129 | 0.11005 | 0.11825 | -0.06503 | 0.225244 |
| **F6** | 0.47692 | 0.257748 | 0.23986 | 0.08362 | 0.119675 | 0.011861 |
| **F7** | 0.36344 | 0.305053 | 0.45194 | 0.02838 | 0.011465 | 0.136258 |
| **F18** | 0.32169 | 0.368252 | 0.17494 | 0.25925 | 0.073758 | 0.201937 |
| **F11** | 0.29892 | 0.245384 | 0.42092 | 0.13163 | 0.086064 | 0.056643 |
| **F13** | 0.25471 | 0.303933 | 0.14958 | 0.43219 | 0.087825 | -0.06309 |
| **F10** | 0.23854 | 0.255268 | 0.66523 | 0.18802 | -0.00124 | 0.006616 |
| **F16** | 0.2371 | 0.663705 | 0.22619 | 0.16035 | 0.057367 | 0.142035 |
| **F9** | 0.20621 | 0.06088 | 0.58965 | 0.10069 | 0.035372 | 0.155673 |
| **F17** | 0.19645 | 0.288579 | 0.18532 | 0.11808 | 0.091583 | 0.474363 |
| **F15** | 0.17609 | 0.729 | 0.09255 | 0.12893 | 0.066651 | 0.059338 |
| **F5** | 0.09116 | 0.013373 | -0.0044 | 0.07706 | 0.6695 | 0.065939 |
| **F8** | 0.08346 | -0.0057 | 0.40273 | 0.07531 | 0.252034 | 0.356106 |
| **F14** | 0.06426 | 0.278234 | 0.15653 | 0.11182 | 0.336485 | 0.047777 |
| **F12** | 0.05458 | 0.150347 | 0.16409 | 0.75365 | 0.140284 | 0.160181 |

**FACTOR SCORES**

The factor scores tell us an individual’s score on this subset of measures. Therefore, any further analysis can be done using factor scores rather than the original data. Secondly, factor scores may be appropriate to use for Regression analysis because they are produced from uncorrelated factors. Thus, these scores reduce or eliminate multicollinearity that can cause problems with regression analysis.

**Binary Logistic Regression**

1. Logistic regression is an extension of simple linear regression.
2. In the case where the dependent variable is dichotomous or binary, we cannot use simple linear regression. Logistic regression is the statistical technique used to predict the relationship between predictors (our independent variables) and a predicted variable (the dependent variable) where the dependent variable is binary (e.g., Sex [Male vs. Female], Response [Yes vs. No], Score [High vs. Low]).
3. There must be two or more independent variables (IVs), or predictors, for a logistic regression. The IVs, or predictors, can be continuous (interval/ratio) or categorical (ordinal/nominal).
4. All predictor variables are tested in one block to assess their predictive ability while controlling for the effects of other predictors in the model.

**Logit** models are used for discrete outcome modeling. This can be for binary outcomes (0 and 1) or three or more outcomes (multinomial logit). The logit model operates under the logit distribution (i.e., Gumbel distribution) and is preferred for large sample sizes.

**Probit** models are mostly the same, especially in binary form (0 and 1). However, for three or more outcomes (in this context, it's typically ranking or ordering) it operates much differently. It uses a single regression equation, in which inferences from marginal effects can only be made on the "extreme" (upper and lower rankings) with any certainty. I can elaborate if more information is needed.

**Tobit** models are entirely different. It has nothing to do with binary or discrete outcomes. Tobit models are a form of linear regression. Specifically, if a CONTINUOUS dependent variable needs to be regressed, but is skewed to one direction, the Tobit model is used. The Tobit model allows regression of such a variable while censoring it so that regression of a continuous dependent variable can happen. It allows the analyst to specify a lower (or upper) threshold to censor the regression at while maintaining the linear assumptions needed for linear regression.

Here we use logit model to perform logistic regression which forces the prediction equation to predict values between 0 and 1.

**Binary Logistic Regression Model:**

Where,

п(X): Conditional probability that the outcome is present i.e.(Y=1|X)

Y: Response Variable

X: Vector of Independent Variables

Y=1; People who want to improve their daily routine.

Y=0; People who don’t want to improve their daily routine.

**Independent Variables**

1. Well Being
2. Quality of life
3. Life circumstances
4. Local Infrastructure
5. Social frictions
6. Ambitions

Extracted Factors-

Quality of Life:

1. Time management

2. Sufficient Resources

3. Transport facilities

4. Fulfillment of desires

5. Work / Study Environment

6. Environment in your Locality / Area

Well Being

1. High Expectations towards life

2. Healthy Life

3. Mental Stability

4. Passion for Living

5. Your Attitude Toward Life

Life Circumstances:

1. Weekend Satisfaction

2. Daily Struggle

3. Social Neighbourhood

4. Promptness

5. Fulfill your enjoyment

Social Frictions

1. Difficult Interactions

2. Unorthodox family

Factor Scores-

**ASSOCIATION RULE (MBA) ( power bi )**

Objective:-

● To identify what different types of activities people mostly go through in a day and how they are associated.

The idea has been used in various fields like Sales, Banking, and Marketing.

This is used by retail firms like Walmart, D-Mart, Amazon, etc. to analyze customers buying behavior.

● We used this to uncover the pattern of activities for people who need an improvement in

their weekend/daily routine as compared to those who are satisfied with their ongoing pattern.

● Association rules are techniques to uncover how items are associated with each other. This is

IF-THEN statements help to show the probability of relationships between data items

within large data sets in various types of databases.

● Association rule mining has several applications and is widely used to help discover

correlations in transactional data.

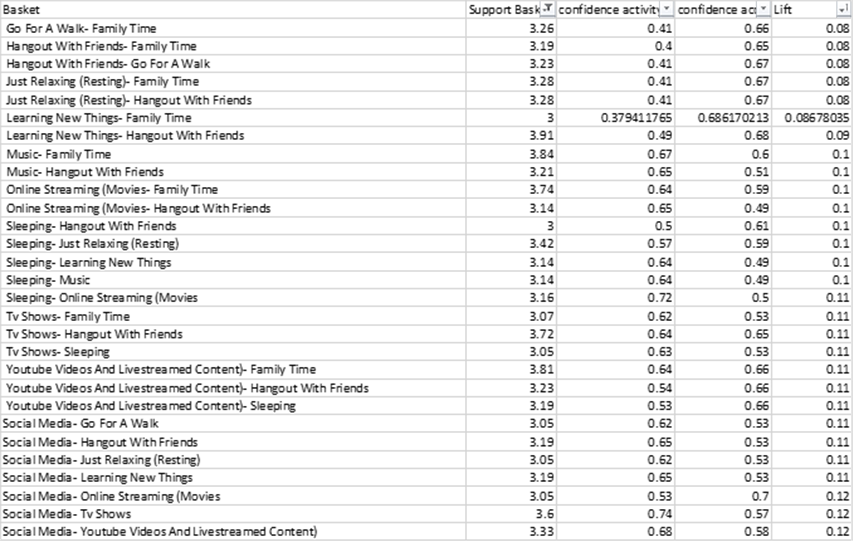
● Three major concepts used in association rules are:

Support

Confidence

Lift

**What are the activities are done in your Leisure time.** (support – 0.3 & confidence between 50% to 70%)



It has been observed that the rules generated for both cases have been found significant.

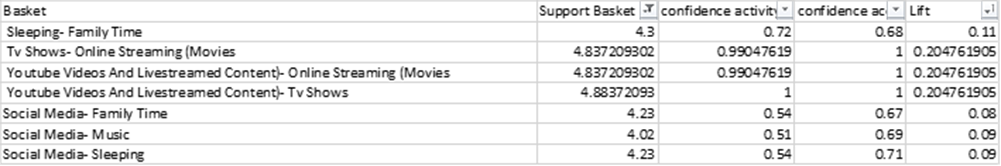
The lift values are higher for the individuals who want to improve their daily schedule with higher lift indicating the greater extent to which the Activities on LHS are dependent on Activities on RHS.

1) People have suggested tasks like going for a walk, Family time, Listen Music, etc. to have an improvement.

2) The possible causes for this may be long working hours or external stress or some other that affect the chance of having more leisure time for these suggested tasks.

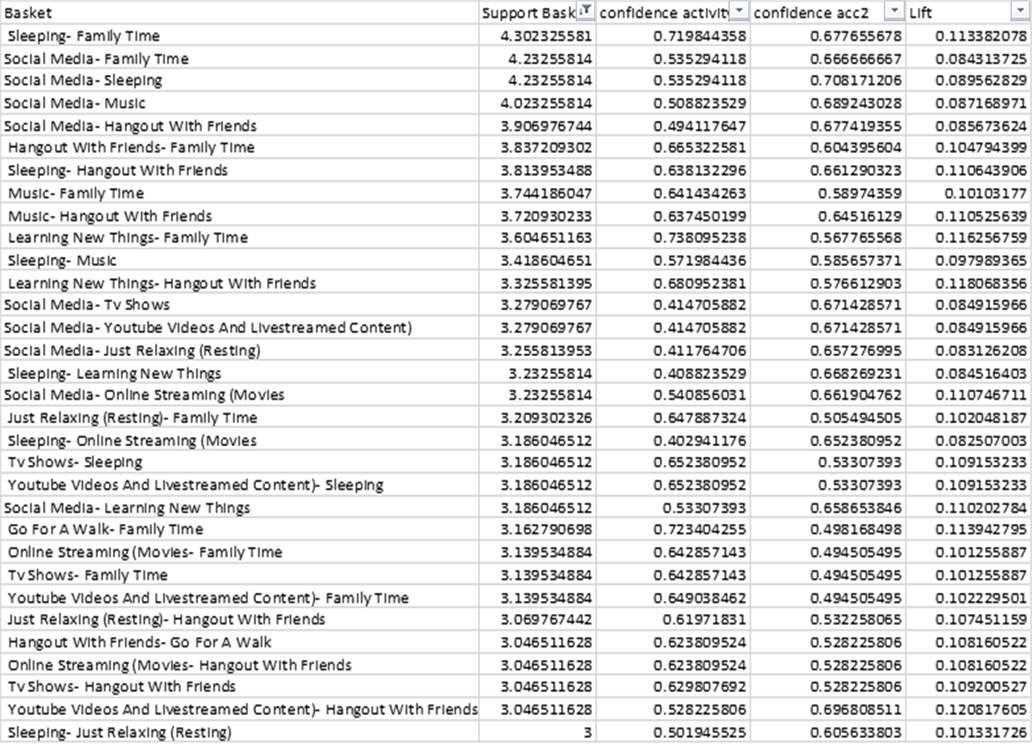
3) More time for these tasks may help the individuals to feel better in their daily routines.

**What activities are done in Leisure time?** (support – 0.4 & confidence between 50% to 70%)



**What activities are done in regular time routines**

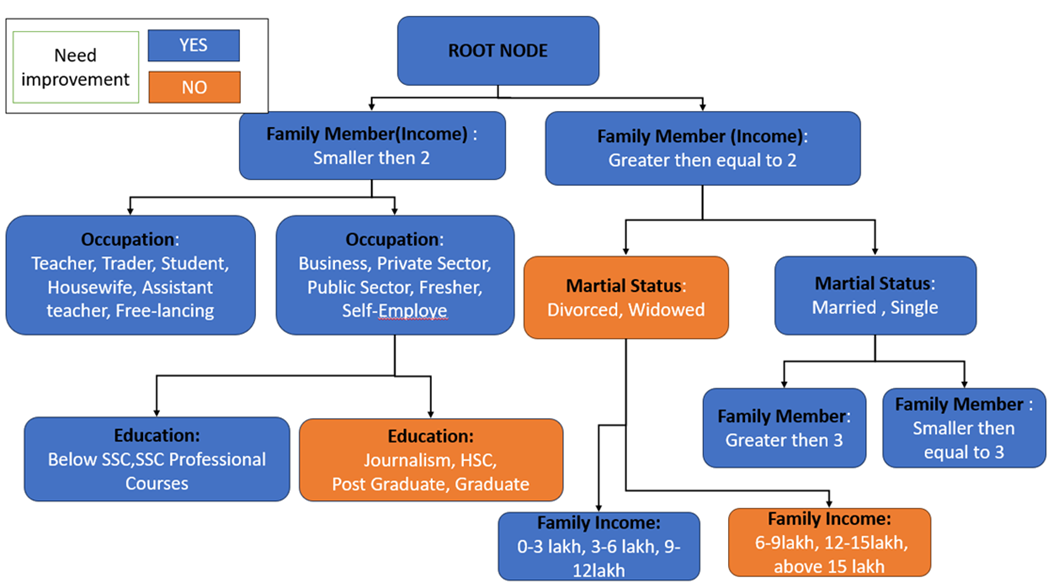
(support – (0.3 to 0.4) & confidence between 50% to 70%)



**DECISION TREE (SAS)**

Objective:- To understand which characteristics of the demographics is the improvement in a daily

routine is needed.



•A person with family size 3,4,6,7,8,9,10, 11,12,13 need an improvement

•A person with a family size of 2,5 doesn’t need an improvement

•A person with occupations private sector, public sector, Business, and housewife needs an improvement

•A person with occupation student, self-employed, teacher, assistant teacher, free-lancing need an improvement

•A person with a Family annual income of 6-9 lakh, above 15 lakhs doesn’t need an improvement

•A person with a Family annual income of 0-3,3-6,9-12,12-15 need an improvement

•A person who spends leisure time with a partner, family, or friends needs improvement.

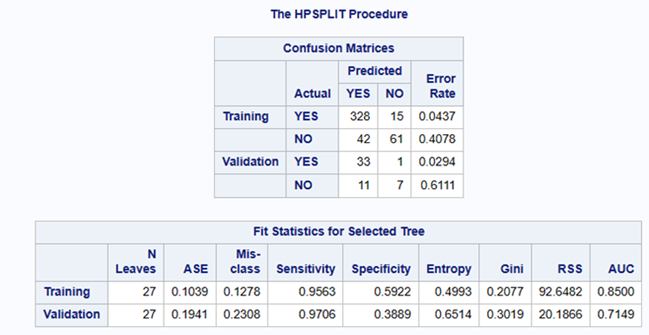
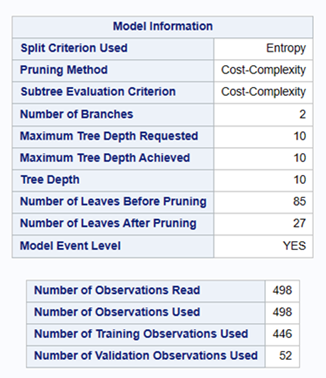
•A person who spends leisure time with ALONE doesn’t need an improvement

•A person with earning members 1,3,2,4,5,7 needs improvement.

•A person with a relationship status divorced and windowed doesn’t need an improvement.

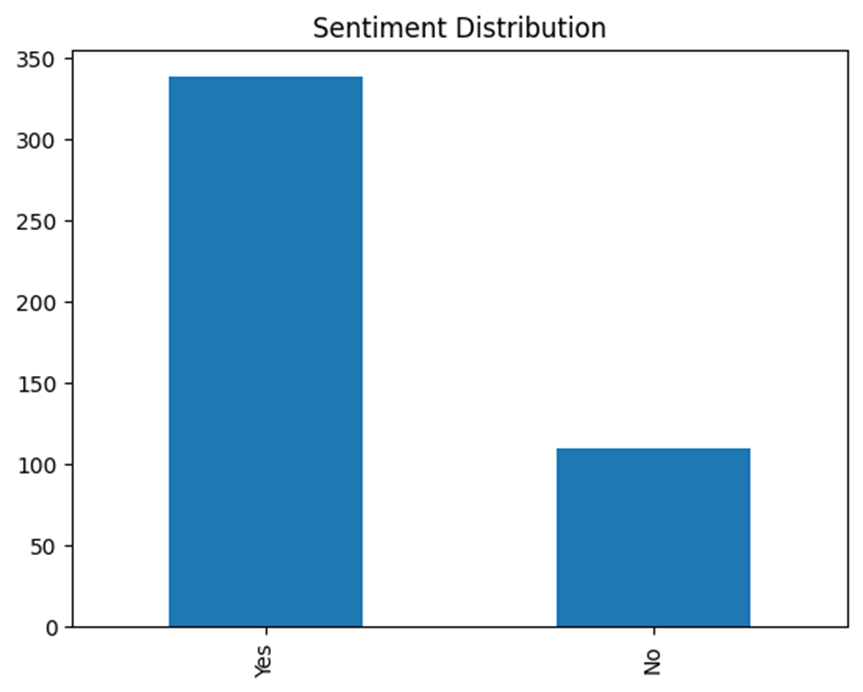
•A person with a relationship status married, Single need an improvement

**Validation check**



**Text Analysis**

This graph shows that of the different emotions expressed, still most of them have been expressed in a positive way and very few in a negative way as to form a response and submit their choice.



**Positive words**



The above word cloud highlights the words like IMPROVE, daily routine,better life ,time etc. the most as compared to other words.

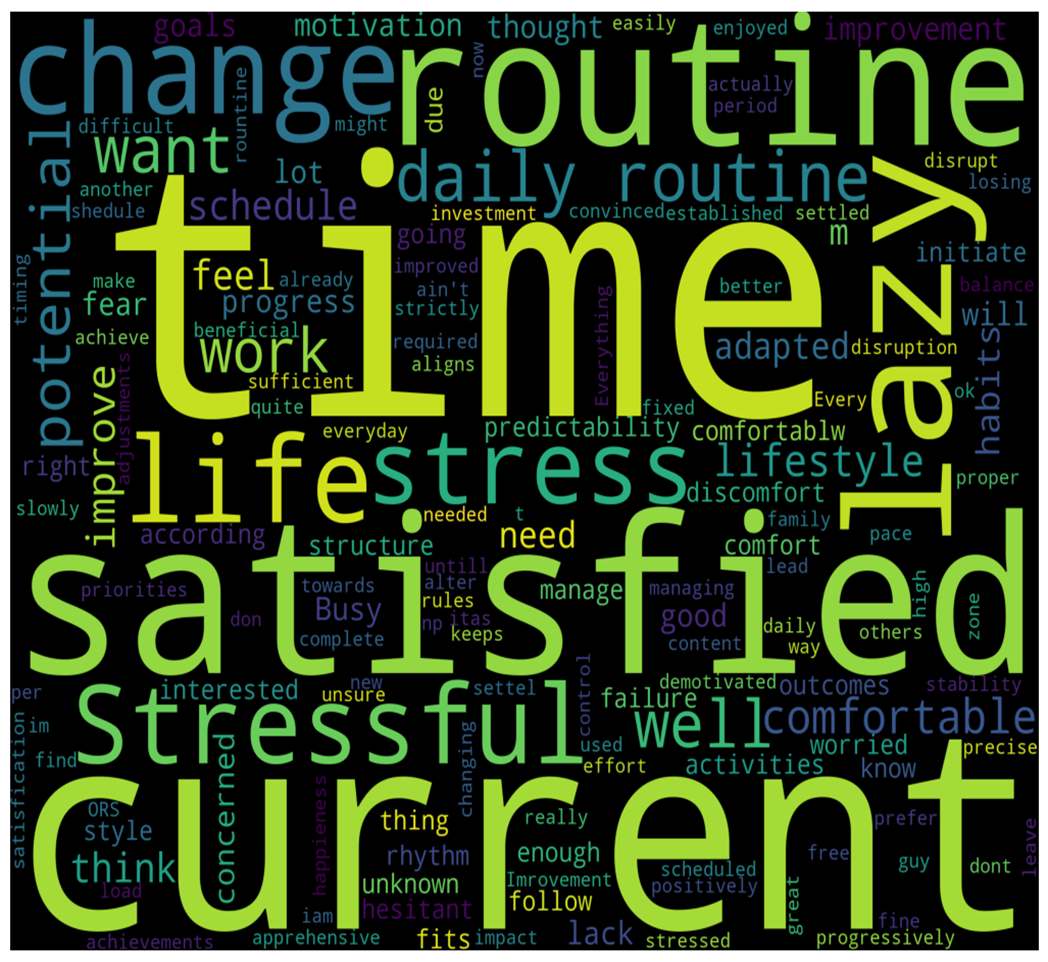
1. This shows that these large sized words have been expressed the most by the people in their responses.

2. Some sort of improvement is required by them in their routines which may be due to any reasons.

3. They are stressed by something and the reason for that may be productivity lacking in their tasks.

4. People want to enjoy a healthy lifestyle and have fun. They be doing something in a proper fashion etc. This gives a clear idea about the responses.

**Negative Words**



The above word cloud highlights the words like change rountine, stressful,change rountine ,lazy etc. the most as compared to other words.

**t test**

Gender

Hypothesis:

H0: There is no significant difference in the mean leisure time in Female and Male.

H1: There is significant difference in the mean leisure time in Female and Male.

Assumptions:

Normality-

Hypothesis

H0: Data follow normal distribution.

H1: Data does not follow normal distribution.

Output-

Shapiro-Wilk Normality test for leisure time in female

W = 0.93623 p-value = 3.221e-09

Shapiro-Wilk Normality test for leisure time in male

W = 0.90029 p-value = 2.072e-11

Conclusion – From the output, the p-values are less than the significance level 0.05. Which implying that the distributions of the data are significantly different from the normal distribution. Hence, we can conclude the no normality.

**Wilcoxon Rank Sum Test**

Assumptions:

1. Data are paired and come from the same population.

2. Each pair is chosen randomly and independently.

3. The data are measured on at least an interval scale when, as is usual, within-pair differences are calculated to perform the test (though it does suffice that within-pair comparisons are on an ordinal scale).

Hypothesis:

H0: There is no significant difference in the median leisure time in females and Male. (Mx=My)

H1: There is a a significant difference in the median leisure time in Female and Male. (Mx≠My)

Output:

Wilcoxon rank sum test with continuity correction

W = 33107 p-value = 0.1719

Conclusion: The p-value of the test is 0.1719, which is greater than the significance level alpha = 0.05. We can conclude that the median leisure time in Female is not significantly different from the median leisure time in Male.

Similarly for other binary groups

Wilcoxon rank sum test with continuity correction

Binary Group Variables Test Statistics P - Value

Marital Status 26406 0.1135

Type of family 27599 0.6651

Conclusion – The p-value of each binary group variables are greater than the significance level alpha=0.05. Hence, there is no significant difference in the categories of each binary group variables.

**One-Way ANOVA**

Assumptions:

The one-way ANOVA has several assumptions that need to be met for the results to be reliable. The main assumptions are:

· The variances of the dependent variable are equal across the groups.

· The dependent variable is continuous and normally distributed in each group.

· The observations are independent within and between groups.

Education Level

Hypothesis:

H0: The means of all groups are equal.

H1: At least one of the means of the groups is different from the others.

Preliminary test to check one-way ANOVA assumptions:

The ANOVA test assumes that, the data are normally distributed and the variance across groups are homogeneous.

Assumption 1: The variances of the dependent variable are equal across the groups.

Output-

Levene's Test for Homogeneity of Variance

Df F value P-value (>F)

group 5 1.0916 0.3641

Assumption 2: Are the data from each of the 2 groups follow a normal distribution?

Hypothesis:

H0: Data follow normal distribution

H1: Data doesn’t follow normal distribution

Output-

**Shapiro-Wilk Normality test**

W = 0.92197 p-value = 2.17e-15

Conclusion:

The p-value obtained from the Shapiro-Wilk test to check the normality of the data is 2.17e-15, which is less than the significance level of 0.05. This indicates strong evidence against the null hypothesis of normality, suggesting that the data is not normally distributed. Therefore, the assumption of normality for the data is violated, and any statistical tests that require normality assumption may not be appropriate to use. In such a case, non-parametric tests such as Kruskal-Wallis test, or Friedman test could be used as alternatives to the parametric tests.

**Kruskal-Wallis test**

Assumptions:

· Independent samples: The groups being compared are independent of each other.

· Random samples: The data is sampled randomly from the population.

· The dependent variable is measured on an ordinal or continuous scale.

· The shape of the distributions of the groups being compared is not assumed to be normal.

· Homogeneity of variance: The variance of the dependent variable is equal across groups.

Output-

Kruskal-Wallis rank sum test

Kruskal-Wallis chi-squared 40.451

df 12

p-values 6.051e-05

Similarly, for other variables which contain more than two categories.

The hypothesis of the Kruskal-Wallis test is that Hypothesis:

H0: The populations from which the groups are drawn have the same distribution.

H1: At least one population has a different distribution than the others.

Output-

Kruskal-Wallis rank sum test

Kruskal-Wallis chi-squared Df p-value

Occupation 40.451 12 6.051e-05

Family annual income 4.9964 5 0.4163

Conclusion:

The p-value obtained from the Kruskal-Wallis test are greater than level of significance for EDU\_LEVEL, Family annual income it indicates that there is no significant difference between the groups being compared at the commonly chosen significance level of 0.05. Therefore, we fail to reject the null hypothesis, which states that the populations from which the groups are drawn have the same distribution.

Similarly, for other variables which contain more than two categories.

The hypothesis of the Kruskal-Wallis test is that Hypothesis:

H0: The populations from which the groups are drawn have the same distribution.

H1: At least one population has a different distribution than the others.

Output-

Kruskal-Wallis rank sum test

Kruskal-Wallis chi-squared Df p-value

Occupation 40.451 12 6.051e-05

Family annual income 4.9964 5 0.4163

Conclusion:

The p-value obtained from the Kruskal-Wallis test are greater than level of significance for EDU\_LEVEL, Family annual income it indicates that there is no significant difference between the groups being compared at the commonly chosen significance level of 0.05. Therefore, we fail to reject the null hypothesis, which states that the populations from which the groups are drawn have the same distribution.

**Pearson Chi square test**

H0 =there is no relationship between the two variables.

H1 =there is a significant relationship between the two variables.

p-value=0.05

If the p-value is less than the significance level, then the null hypothesis is rejected.

Varibale 1 Variable 2 P value

Gender Maritial status p-value = 6.051e-05

Gender Nuclear family p-value = 0.3505

Gender Education level p-value = 0.2154

Gender Occupation p-value = 0.001779

Gender Family annual income p-value = 0.5798

Maritial Status Nuclear family p-value = 6.932e-09

Maritial Status Education level p-value = 0.002723

Maritial Status Occupation p-value < 2.2e-16

Maritial Status Family annual income p-value = 0.000368

Nuclear family Occupation p-value = 0.006692

Nuclear family

Education level p-value = 0.2664

Nuclear family Family annual income p-value = 0.1657

Education level Occupation p-value = 0.002601

Education level Family annual income p-value = 0.0124

Occupation Family annual income p-value = 0.0004672

Chi-Square test of independence is used on discrete data that can be

nominal or ordinal. It can be used to test the null hypothesis that the two

categorical variables under consideration are independent of each other.

The procedure involves comparing the observed cell frequencies with

expected cell frequencies. Expected frequencies are no. of cases that

should fall in each cell if there is no relationship between the 2 categorical

variables.

**Test Statistic:**

**cal** ∑ **Oi – Ei) 2 /Ei)**

Where

= Chi-Square test of Independence

Oi= Observed value of two nominal variables

Ei = Expected value of two nominal variables

**Decision Criteria:** We reject H0 if p-value < 0.05

If the tabulated chi-square value is less than the calculated chi-square value, then

the null hypothesis is rejected and we conclude that the two variables under

consideration are dependent.

21

**The Relationship between Gender and Education.**

H0: There is no Association between Gender and Education.

H1: There is an Association between Gender and Education.

**Decision Criterion:**

Reject H0 if p value < α

**The Relationship between Gender and Education**

