Lifestyle and Wellbeing Analysis

1 Introduction

We intend to understand what various factors affect the lifestyle of a person. For example, we intend to understand how our sleep schedules or stress levels are affected by factors like Meditation, eating of fruits and vegetables as well as the daily steps we walk. To understand this in depth we formulate a set of hypotheses as well as general correlations to understand and analyze the lifestyle of an individual properly.

We chose the dataset “Lifestyle_and_Wellbeing” (*Dataset Link*, 2021) on Kaggle which contains about 15,977 responses with 24 attributes. It contains values in ranges like 0-5,0-10 for attributes like fruits and veggies, daily steps, etc., and also demographic information like Age, gender, etc.

2 Literature review

We did not refer to any paper for our analysis. Dataset consisted of 24 attributes(individual questions with answers like stress levels, weekly meditation,etc.) With about 16000 responses. We referred to different articles and links, some of which I have attached in the references section (*Correlation Analyses in R - Easy Guides - Wiki*, n.d.) (Frost, n.d.) (Soetewey, 2022). We also referred to a book (*Stats Book*, n.d., #) to grasp some basics properly especially for multiple regression.

3 Problem statement

- Data exploration and visualizations
- Studying correlation between factors which we think should be interrelated.
- Formation of hypotheses on the basis of the various visualizations and correlations that we have looked at and then testing them using statistical techniques taught in class.

4 Methods

4.1 Analysis done by me

Correlation Analysis

All three tests namely spearman,pearson and kendall were used for these.

- BMI vs Daily steps vs Fruits and veggies - pairwise testing was done for the three variables. After this partial correlation was done for Daily steps vs fruits and veggies keeping BMI as the controlling variable.

Simple one to one testing here using the three tests above

- To do-list vs Live vision
- Sleep hours vs weekly meditation
- Time for passion vs achievement
- Lost vacation vs Sleep hours

Finally a chi-squared test for gender vs stress

Hypothesis Analysis

- People with a healthy lifestyle and regular weekly goals are high achievers.
- People who have a regular mindfulness practice have better stress management skills than those who do not practice mindfulness.
- People who give more time for their passion tend to be more energetic and are more immersed in their work.
- People who have high income travel a lot and are generous donors.
- People who are middle aged are suffering from more daily stress.

First, for each of these hypothesis I check data for normality using KS-test and histogram plots. If the data is normal I further check for homogeneity of variances using Bartlett's test. Once this is done I Use the following tests.

- Normal data - Anova(one-way,two-way or three way depending on the data) followed by pairwise t-testing(BH correction) and TukeyHSD post hoc tests)
- Non-normal data - Kruskal wallis test followed by dunn's test and pairwise wilcoxin tests(BH correction).

4.2 Analysis done by teammates:

- Multiple Regression analysis
- Pre-analysis(including boxplots)
- Attempt at Clustering Analysis

5 Results

5.1 Analysis done by me

Correlation Results

1) BMI vs Daily steps vs Fruits and veggies

Significant p-value was obtained for all correlations(< 2* 1e-16). Since this is not the best measure due to large sample size I also consider the test statistic value

BMI vs Daily steps

Spearman - -0.1315824
Pearson - -0.131451
Kendall - -0.113119

Small negative correlations observed here(0.1 - small)

BMI vs Fruits and veggies

Spearman - -0.0918827
Pearson - -0.09296747
Kendall - -0.08186301

Small negative correlations observed here(0.1 - small)

Fruits and veggies vs Daily steps

Spearman - 0.2503675
Pearson - 0.251213
Kendall - 0.1947633

Medium positive correlations observed here(0.3 - medium)

Partial correlation with BMI as a controlling variable showed a decrease for this correlation.

Spearman - 0.2413884
Pearson - 0.242133
Kendall - 0.1873302

This tells you that BMI can partly explain the correlation between the other two But otherwise is not significantly correlated with the other two. Fruits and veggies And Daily steps are significantly correlated.

2) To do-list vs Live vision

P-value significant(<2*1e-16)
Spearman - 0.2854402
Pearson - 0.2698743
Kendall - 0.2169894

Medium positive correlation which tells that the completion of weekly tasks are Significantly correlated with your future planning

3) Sleep hours vs Weekly Meditation

P-value significant(<2*1e-16)
Spearman - 0.1560521
Pearson - 0.1632097
Kendall - 0.1247929

Sleep hours shows small positive correlations with weekly meditation

4) Gender vs Daily Stress

Chi-squared test was done for these as both variables are categorical.
P-value significant(<2*1e-16) and chi-squared statistic - `282.1` implies a fairly significant correlation.

5) Time for passion vs achievement

P-value significant(<2*1e-16)
Spearman - `0.3916002`
Pearson - `0.3689393`
Kendall - `0.2997548`

This was the strongest positive correlation we found among all of our tests.

6) Lost Vacation vs Sleep hours

P-value significant(<2*1e-16)
Spearman - `-0.100191`
Pearson - `-0.09167296`
Kendall - `-0.08282467`

Small negative correlations here. This is expected as well as more the Sleep hours you are getting, chances are you are getting more vacation And hence lesser lost vacation.

Some Sample output screenshots for this analysis

```
Warning: Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  data$GENDER and data$DAILY_STRESS
X-squared = 282.1, df = 6, p-value <
2.2e-16

Warning: Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  data$AGE and data$DAILY_STRESS
X-squared = 139.85, df = 18, p-value
< 2.2e-16
```

```
Warning: Cannot compute exact p-value with ties
        Spearman's rank correlation rho

data:  grp1 and grp2
S = 4.1316e+11, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3916002


        Pearson's product-moment correlation

data:  grp1 and grp2
t = 50.163, df = 15970, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3554645 0.3822608
sample estimates:
      cor
0.3689393


        Kendall's rank correlation tau

data:  grp1 and grp2
z = 50.887, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.2997548
```

Hypothesis Results

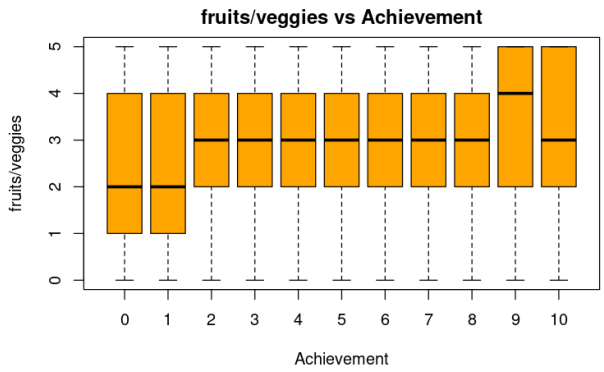
Common points:

- For all of these I use KS-test/histograms for normality testing and Bartlett's test for checking the homogeneity of variances and then proceed accordingly.
- I have applied BH correction after all types of pairwise testing like pairwise t-testing or pairwise wilcoxin testing

Note - Boxplots were drawn by teammates I have only attached one for each for explanation

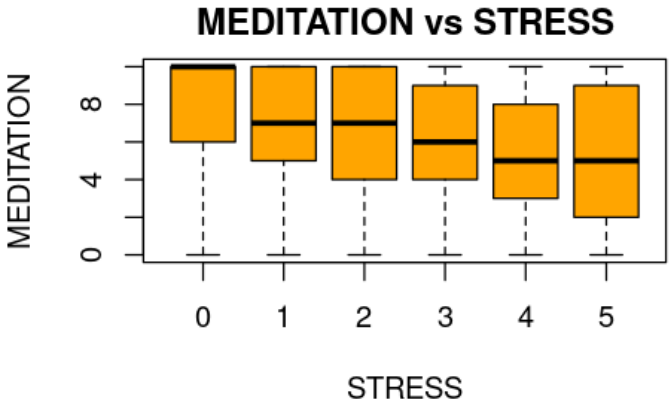
1) People with a healthy lifestyle and regular with weekly goals are high achievers.

We intend to study three way interaction between fruits/veggies,daily steps, and Weekly task completion on achievement.



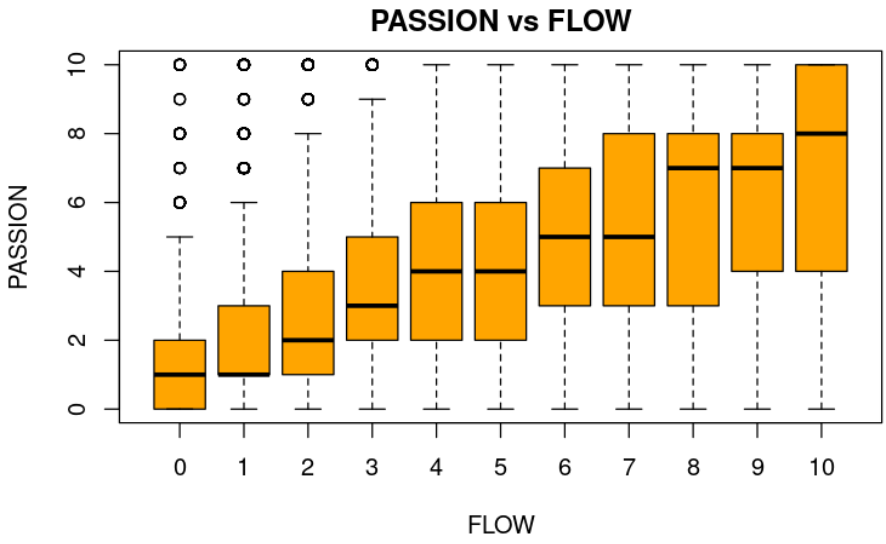
All the individual factors seemed to increase along with increasing achievement according to boxplots. Anova conditions were satisfied. Three way anova tells us that pairwise the three factors had no significant effect in terms of p-values(>0.05) but combined the three factors gave a p-value of 0.02. Individual p-values were all significant as well(2*1e-16). Tukey's HSD posthoc test seemed to give mixed results. Some classes have significant inter p-value while some didn't. We can partially consider this hypothesis.

2) People who have a regular mindfulness practice have better stress management skills than those who do not practice mindfulness.



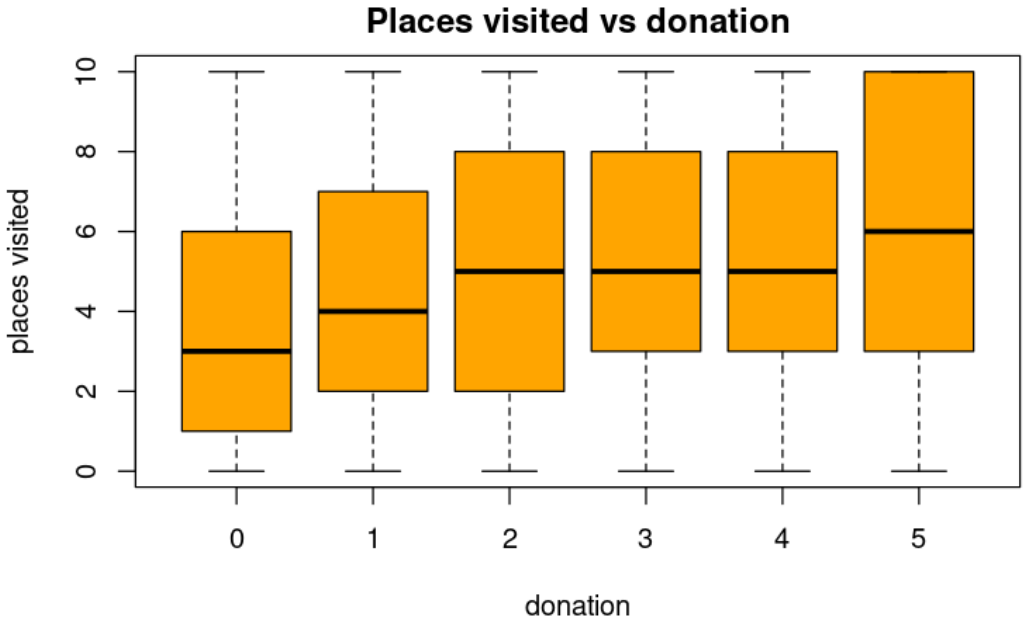
Boxplot suggests a decrease in meditation with increasing stress.
Data here was found to be not normal. We do kruskall wallis test first which gives us significant p-value(<2*1e-16) and a fairly high Kruskal wallis chi-squared Value - 771.85. This was followed by Posthoc-tests namely dunn's and pairwise wilcoxin tests which gave significant correlations between all classes(p-value < 2*1e-16). This hypothesis can be accepted.

3) People who give more time for their passion tend to be more energetic and are more immersed in their work.



Boxplot suggests a very strong positive correlation. Again data was found to be not normal here So we first start with kruskal wallis test which gives us a significant p-value(<2*1e-16) and a very strong Kruskal Wallis chi-squared - 3901.3. Again posthoc tests(Dunn's and pairwise wilcoxin tests) seemed to confirm the same thing with significant p-values between classes. This is our strongest hypothesis.

4) People who have high income travel a lot and are generous donors.
We intend to study a two-way interaction of sufficient_income,places visited on donation.



Boxplots suggest a decent increasing trend on donation and travelling. Data was normal so a two-way anova was done which suggested no significant interaction(p-value > 0.07) individually the factors were quite significant though. Pairwise t-testing suggested same thing with many p-values above 0.05. Hypothesis rejected.

5) People who are middle aged are suffering from more daily stress.

Data is first divided into two sets 36-50(middle aged) and the rest in a separate group.

Boxplots didnot suggest significant difference. On doing a t-test between the two We found a slightly higher mean shift towards non-middle aged group which is the reverse of what is expected. Pairwise t-testing Suggested a decent p-value(<0.05) but not that significant(~0.003). We consider this hypothesis rejected.

5.2 Analysis done by teammates

Pre-Analysis - Boxplots were drawn for each hypothesis by teammates I have attached some sample ones along with the hypothesis. The data was cleaned by removing rows with N/A information, and also replacing redundancies wherever found. **Clustering Analysis** was attempted but it was not continued because it failed our testing for homogeneity of variances through bartlett's test,scree plot,etc.

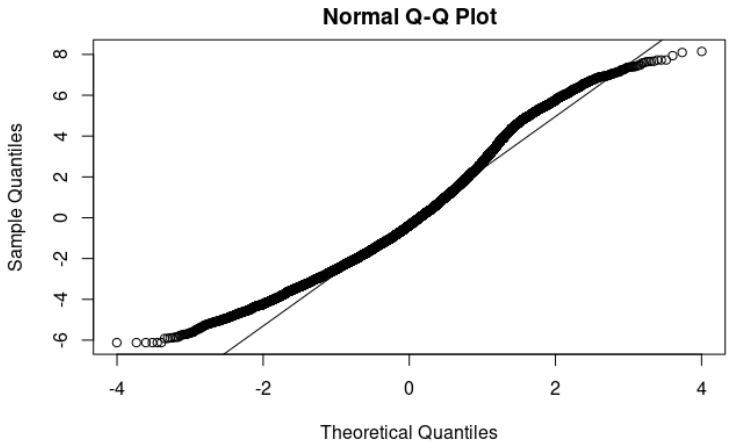
Multiple Regression analysis was done for each hypothesis. Some of the correlations had models drawn as well. Normality of residuals was checked using Q-Q plots and they were checked for heteroscedasticity. Hypothesis gave good models but some of the correlations gave bad results like the models predicting stress and donation. I am attaching some sample models that were made.

```
Call:
lm(formula = ACHIEVEMENT ~ FRUITS_VEGGIES + DAILY_STEPS + WEEKLY_MEDITATION +
    TODO_COMPLETED, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1219 -1.9042 -0.3739  1.5579  8.1463

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.037172   0.069870  14.844  < 2e-16 ***
FRUITS_VEGGIES    0.116290   0.015046   7.729 1.15e-14 ***
DAILY_STEPS       0.101476   0.007416  13.684  < 2e-16 ***
WEEKLY_MEDITATION 0.082889   0.007007  11.829  < 2e-16 ***
TODO_COMPLETED    0.265968   0.008161  32.589  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.583 on 15967 degrees of freedom
Multiple R-squared:  0.1219,    Adjusted R-squared:  0.1217
F-statistic: 554.1 on 4 and 15967 DF,  p-value: < 2.2e-16
```

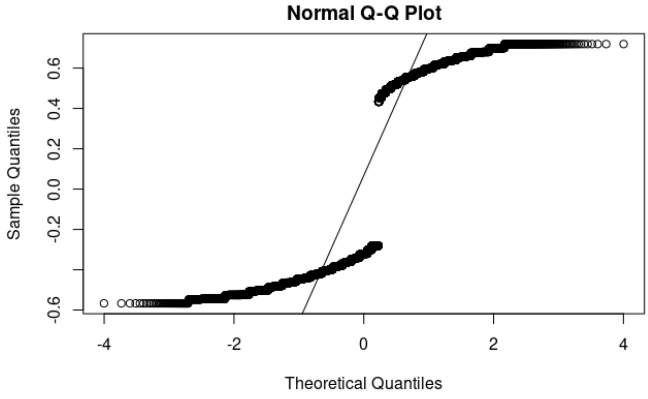



```
Call:
lm(formula = BMI_RANGE ~ DAILY_STEPS + FRUITS_VEGGIES, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5668 -0.4206 -0.3246  0.5554  0.7190

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.586410   0.010472  151.49  < 2e-16 ***
DAILY_STEPS    -0.019634   0.001376  -14.26  < 2e-16 ***
FRUITS_VEGGIES -0.021819   0.002758   -7.91 2.74e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4868 on 15969 degrees of freedom
Multiple R-squared:  0.02111,   Adjusted R-squared:  0.02099
F-statistic: 172.2 on 2 and 15969 DF,  p-value: < 2.2e-16
```



6 Conclusion and Discussion

We study many correlations and hypotheses in depth. We find that factors like weight and sleep hours tend to not have very strong correlations with other factors. While time for passion has a very strong correlation. For our hypothesis we see that

middle aged people may not be the most stressed,People with a rich lifestyle may not be the highest donors and that People with a good diet and routine may sometimes be the highest achievers(partial acceptance). However, we find that giving time for passion always leads to work focus and meditation greatly helps for stress.

Regression supports our hypothesis But some models predicting stress and donation did end up giving us bad results.

We understand that our analysis is far from perfect and in future work we can dive deeper into various aspects of the analysis. We can attempt and see if its possible to do any sort of factor analysis on the data. We can try and build better multiple regression models as much as possible as well that the current models do not produce the best results. We can further try to understand if our testing results(such as anova and post hoc) have been interpreted perfectly and attempt improvement in analysis if required.

7 References

- *Dataset Link*. (2021, March 14). Kaggle. Retrieved May 4, 2023, from https://www.kaggle.com/datasets/ydalat/lifestyle-and-wellbeing-data
- *Correlation Analyses in R - Easy Guides - Wiki*. (n.d.). STHDA. Retrieved May 4, 2023, from http://www.sthda.com/english/wiki/correlation-analyses-in-r
- Frost, J. (n.d.). *Using Post Hoc Tests with ANOVA - Statistics By Jim*. Statistics by Jim. Retrieved May 4, 2023, from https://statisticsbyjim.com/anova/post-hoc-tests-anova/
- *R: Documentation*. (n.d.). The R Project for Statistical Computing. Retrieved May 4, 2023, from https://www.r-project.org/other-docs.html
- Soetewey, A. (2022, March 24). *Kruskal-Wallis test, or the nonparametric version of the ANOVA*. Stats and R. Retrieved May 4, 2023, from https://statsandr.com/blog/kruskal-wallis-test-nonparametric-version-anova/
- *Stats Book*. (n.d.). https://open.umn.edu/opentextbooks/textbooks/559

8 Code

Our entire code is available on the drive link

https://drive.google.com/drive/folders/1JoJhsFFarGMXobA6kKEn8KkcIABC3Bp6?usp=sharing