Student Mental Health

Appendix

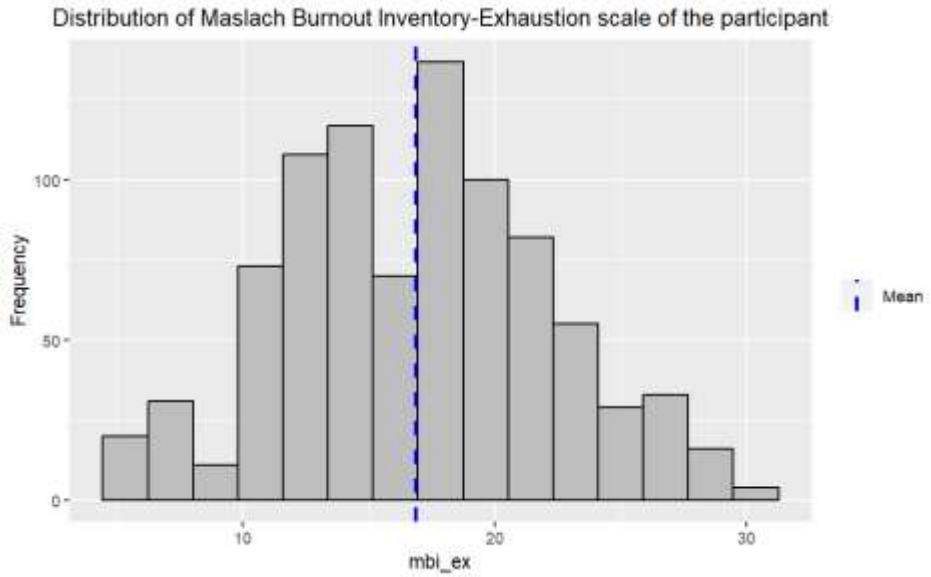
Appendix -1

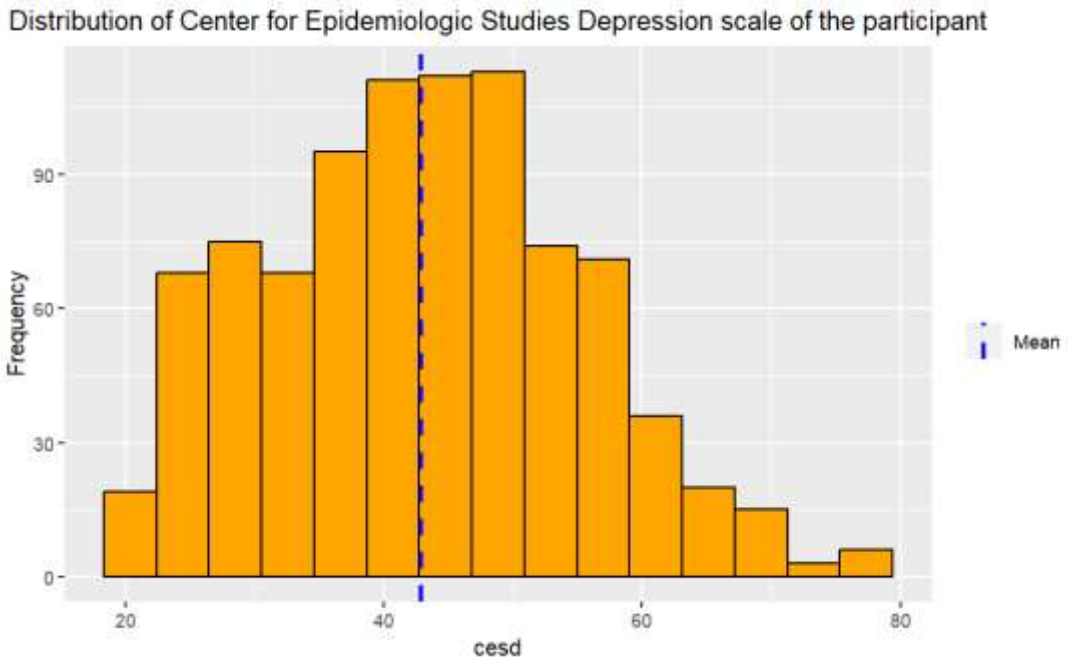
Although the data did not require cleaning, I found that several integer/numeric variables lacked ordinal meaning and should be considered categorical variables. Therefore, I categorized these variables based on the guidance provided in the documentation available at https://www.kaggle.com/datasets/thedevastator/medical-student-mental-health, as well as insights gained from the Q&A session. The resulting data structure is presented below.

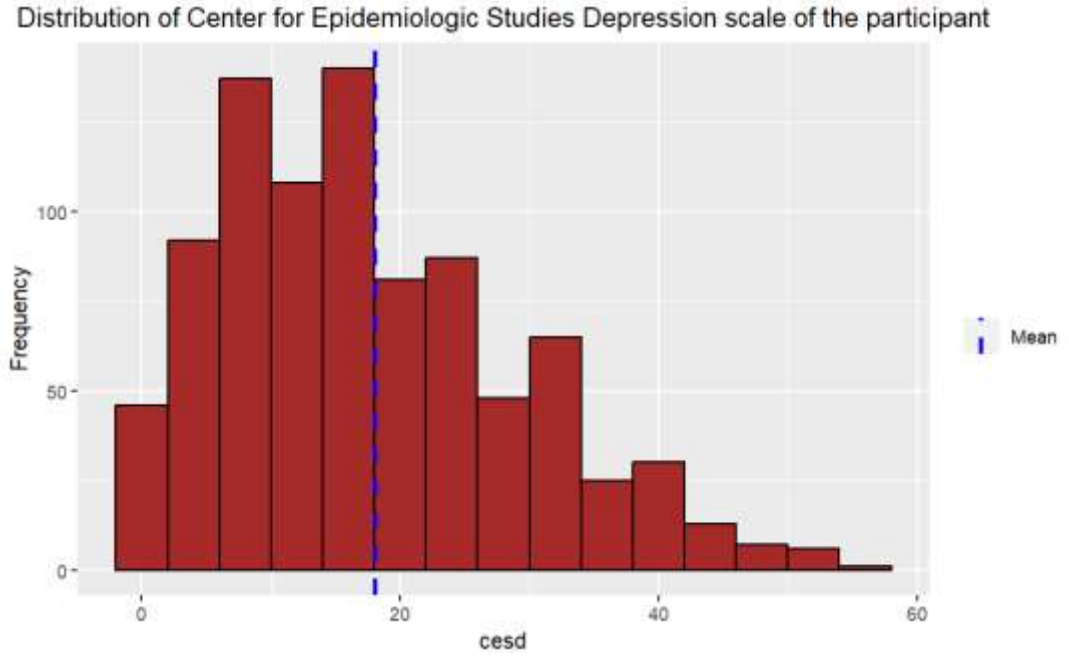
```
'data.frame':  886 obs. of  20 variables:
 $ id       : int  2 4 9 10 13 14 17 21 23 24 ...
 $ age      : int  18 26 21 21 21 26 23 23 23 22 ...
 $ year     : Factor w/ 6 levels "1","2","3","4",..: 1 4 3 2 3 5 5 4 4 2 ...
 $ sex      : Factor w/ 3 levels "1","2","3": 1 1 2 2 1 2 2 1 2 2 ...
 $ glang    : Factor w/ 19 levels "1","15","20",..: 18 1 1 1 1 1 1 1 1 1 ...
 $ part     : Factor w/ 2 levels "0","1": 2 2 1 1 2 2 2 2 2 2 ...
 $ job      : int  0 0 0 1 0 1 0 1 1 0 ...
 $ stud_h   : int  56 20 36 51 22 10 15 8 20 20 ...
 $ health   : Factor w/ 5 levels "1","2","3","4",..: 3 4 3 5 4 2 3 4 2 5 ...
 $ psyt     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ jspe     : int  88 109 106 101 102 102 117 118 118 108 ...
 $ qcae_cog : int  62 55 64 52 58 48 58 65 69 56 ...
 $ qcae_aff : int  27 37 39 33 28 37 38 40 46 36 ...
 $ amsp     : int  17 22 17 18 21 17 23 32 23 22 ...
 $ erec_mean: num  0.738 0.69 0.69 0.833 0.69 ...
 $ cesd     : int  34 7 25 17 14 14 45 6 43 11 ...
 $ stai_t   : int  61 33 73 48 46 56 56 36 43 43 ...
 $ mbi_ex   : int  17 14 24 16 22 18 28 11 26 18 ...
 $ mbi_cy   : int  13 11 7 10 14 15 17 10 21 6 ...
 $ mbi_ea   : int  20 26 23 21 23 18 16 27 22 23 ...
```

Appendix -2

Distribution of the 3 relevant variant variables -





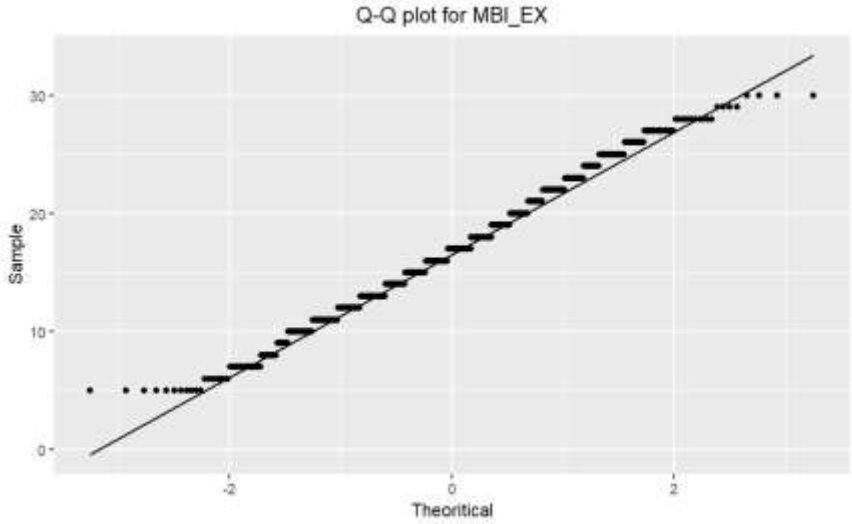


Normality Tests and Q-Q plots -

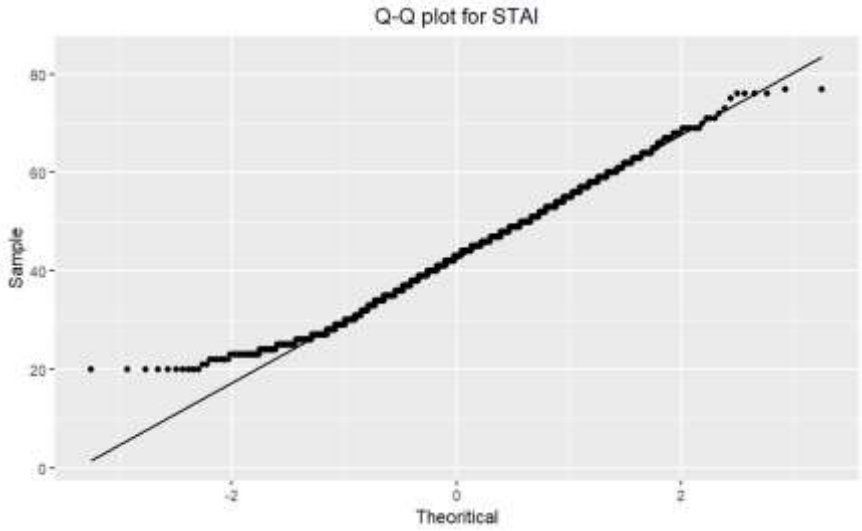
We assume the means of the samples to be normally distributed by CLT due to a large sample size but the normality was checked using the tests and visual inspection and wasn't found to be normal. Based on the Q&A and discussions was assumed to be normal for this assignment.

Shapiro-Wilk normality test

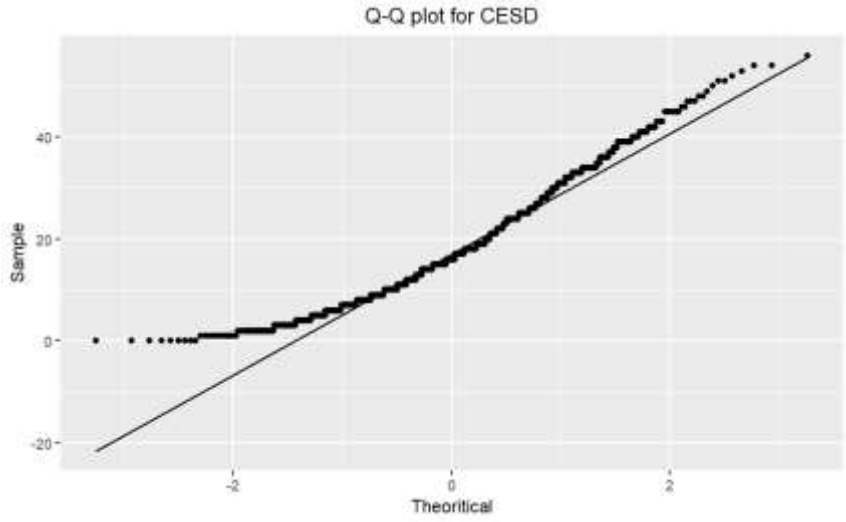
data:
df_filtered_t$mbi_ex_t
W = 0.95864, p-value =
4.085e-15



data:
df_filtered_t$stai_t_t
W = 0.97987, p-value =
1.05e-09



data:
df_filtered_t$cesd_t
W = 0.94305, p-value <
2.2e-16

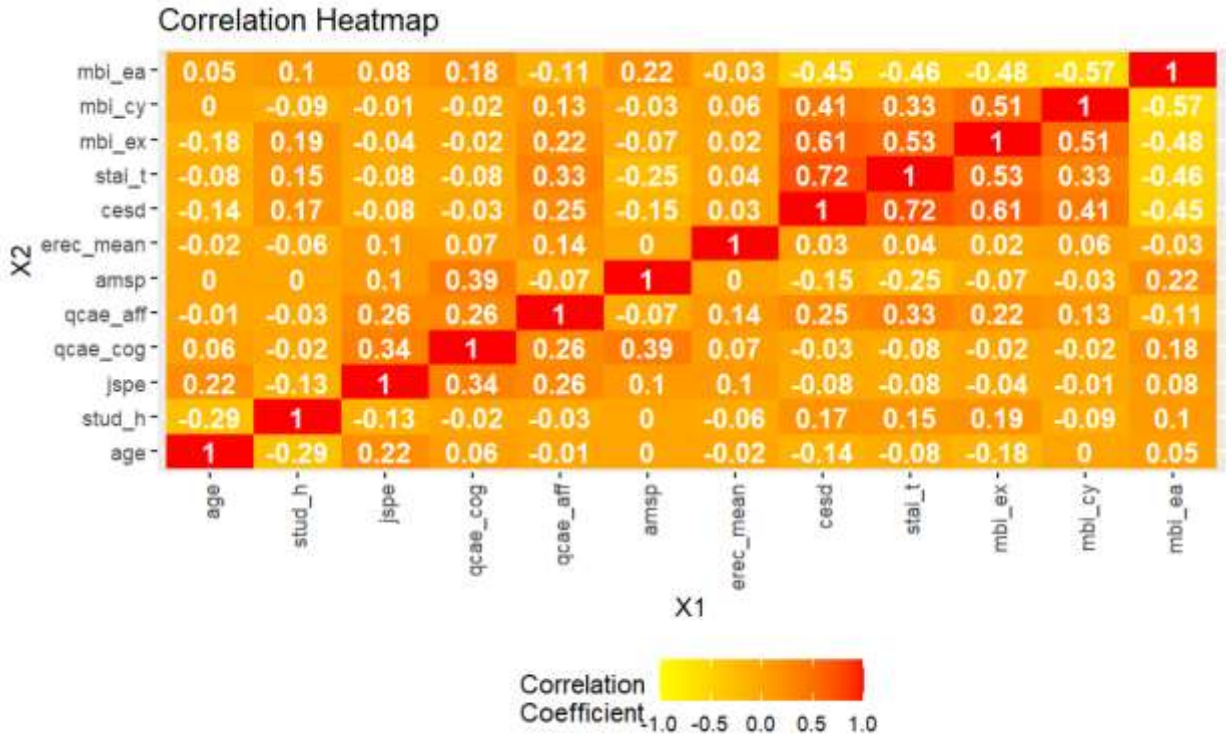


Normality tests after log transformations to the data –

Variable	Transformation	W-value	p-value
df_filtered_t$mbi_ex_t	Log(x+1)	0.95864	4.085e-15
df_filtered_t$stai_t_t	Log(x+1)	0.97987	1.05e-09
df_filtered_t$cesd_t	Log(x+1)	0.94305	< 2.2e-16
df_filtered_t$mbi_ex_t	Square Root	0.98512	7.747e-08
df_filtered_t$stai_t_t	Square Root	0.9893	4.662e-06
df_filtered_t$cesd_t	Square Root	0.99342	0.0006077

Appendix -3

The correlation between all the numeric variables in the dataset. I decided to choose my 3 variables looking at the correlation heatmap backed by a intuition that depression , anxiety and emotional exhaustion are likely to be correlated.



Observations –

High +ve Correlation -	***High/Moderate -ve Correlation -***
cesd x stai_t = 0.72 cesd X mbi_ex = 0.61	mbi_ea X mbi_cy = -0.57 mbi_ea X mbi_ex = -0.47

References –

1. https://www.r-bloggers.com/2013/05/how-to-calculate-a-partial-correlation-coefficient-in-r-an-example-with-oxidizing-ammonia-to-make-nitric-acid/
2. https://www.analyticsvidhya.com/blog/2021/01/correlation-analysis-using-r/
3. https://www.kaggle.com/datasets/thedevastator/medical-student-mental-health