University of Texas at Arlington

Applied Statistics and Data Science

5301 STAT THEORY AND APPLICATION – Final Project

Title: Healthcare Outcomes: Analyzing Treatment and Age Influences on Scores using Two-way ANOVA

Team 09

Lakshman Kumar Reddy Peddireddy – 1002149745

Vineesha Mallu - 1002149747

Table of contents

Problem Statement:

In the field of healthcare and wellness, understanding the impact of different treatments and age groups on individuals' outcomes is crucial. The dataset at hand encompasses three distinct treatments—mental, physical, and medical—administered to individuals categorized into three age groups: young, mid, and old. We seek to explore whether there are statistically significant variations in the scores based on the type of treatment, age group, or the interaction between these two factors.

Objective:

The goal is to provide a detailed knowledge of how treatment, age, and their interaction affect the observed outcomes in addition to determining whether there are significant changes in scores based on these variables using comprehensive analysis. The findings will contribute valuable insights to healthcare professionals, informing potential adjustments in treatment approaches tailored to specific age groups, ultimately enhancing the effectiveness of healthcare interventions.

Dataset Overview:

Link: https://houssein-assaad.shinyapps.io/TwoWayANOVA/

There are **27** observations in the dataset, each corresponding to a unique individual subjected to a specific treatment and falling into a particular age group.

Features:

- **Treatment:** It represents the type of intervention or therapy administered to individuals.
- **Age:** It categorizes individuals into different age groups.
- **Scores (Main Focus)**: It is a continuous variable which represents the effectiveness of treatments among different age groups.
- **StressReduction**: It represents a level or degree of stress reduction.

Treatment Levels: 'mental,' 'physical,' 'medical'.

Age Levels: 'young,' 'mid,' 'old'

Code:

Output:

```
15  physical  mid        4   18
```

Fig.1: Dataset Overview

Hypothesis:

Main Effects Hypotheses:

Treatment Main Effect:

- **Null Hypothesis (H0):** There is no significant difference in scores among different levels of treatments.
- **Alternative Hypothesis (H1):** There is a significant difference in scores among different levels of treatments.

Age Main Effect:

- **Null Hypothesis (H0):** There is no significant difference in scores among different age groups.
- **Alternative Hypothesis (H1):** There is a significant difference in scores among different age groups.

Interaction Effect Hypothesis:

Treatment by Age Interaction Effect:

- **Null Hypothesis (H0):** The interaction between treatment and age does not significantly influence the scores of individuals.
- **Alternative Hypothesis (H1):** There is a significant interaction effect between treatment and age, leading to variations in scores.

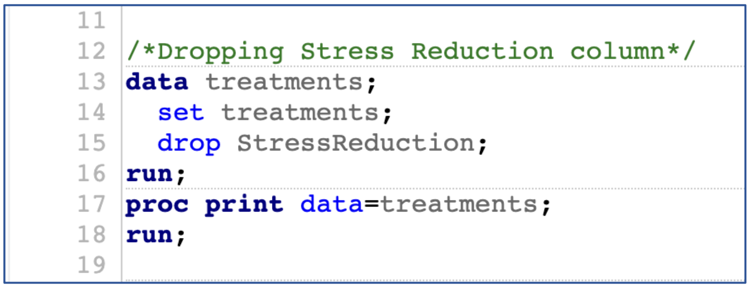
Data Preprocessing:

Data preprocessing is a crucial step in data analysis. It involves cleaning and transforming raw data into a format that is suitable for analysis. The following steps were undertaken during data preprocessing to ensure the dataset's quality and suitability for analysis:

Dropped StressReduction Column:

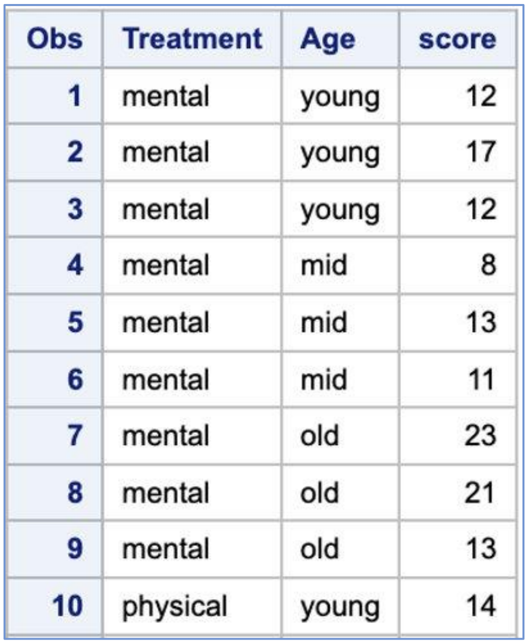
Why: The reason to drop the column is we focus more on score column (Which is responsive variable in our case).

Code:



```

/*Dropping Stress Reduction column*/
data treatments;
  set treatments;
  drop StressReduction;
run;
proc print data=treatments;
run;

```

Output:



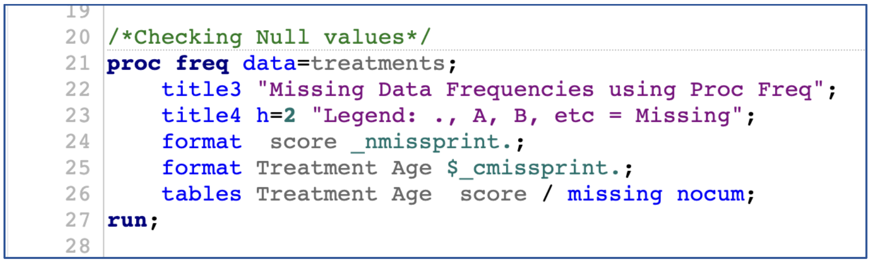
Obs	Treatment	Age	score
1	mental	young	12
2	mental	young	17
3	mental	young	12
4	mental	mid	8
5	mental	mid	13
6	mental	mid	11
7	mental	old	23
8	mental	old	21
9	mental	old	13
10	physical	young	14

Fig.2: Data set overview after dropping a column

Verifying and Handling Missing values:

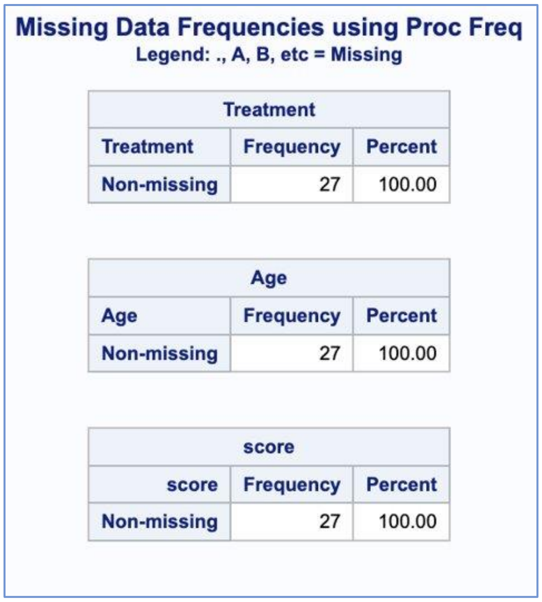
Here, we are checking for any null or missing values in our dataset. Fortunately, our dataset does not have any missing or null values.

Code:



```
/*Checking Null values*/
proc freq data=treatments;
    title3 "Missing Data Frequencies using Proc Freq";
    title4 h=2 "Legend: ., A, B, etc = Missing";
    format  score _nmissprint.;
    format Treatment Age $_cmissprint.;
    tables Treatment Age  score / missing nocum;
run;
```

Output:



Missing Data Frequencies using Proc Freq

Legend: ., A, B, etc = Missing

Treatment		
Treatment	Frequency	Percent
Non-missing	27	100.00

Age		
Age	Frequency	Percent
Non-missing	27	100.00

score		
score	Frequency	Percent
Non-missing	27	100.00

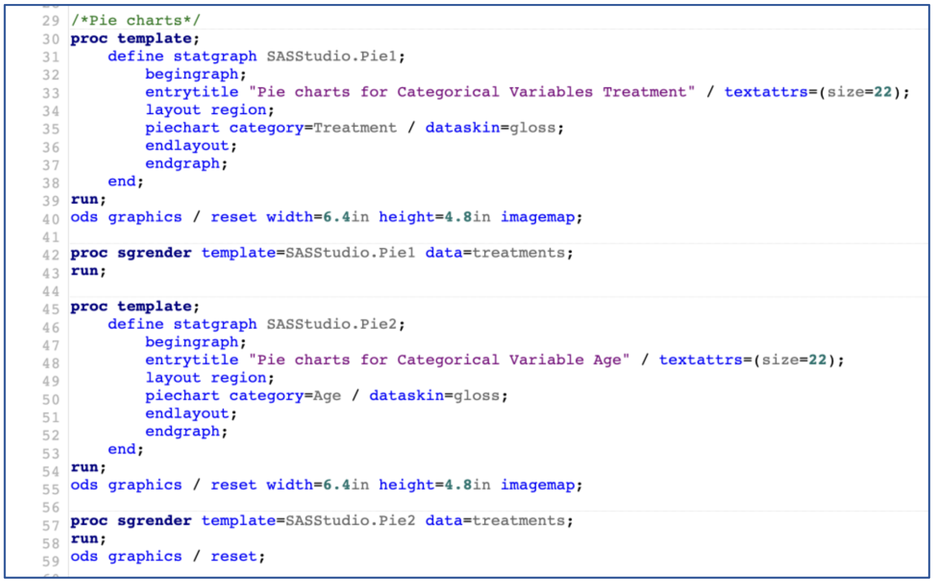
Fig.3: Output of Handling missing and null values

Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is a crucial initial step in data analysis where the main objective is to summarize the main characteristics, often with the help of graphical representations, and to uncover patterns, relationships, or anomalies in the data. EDA involves examining and visualizing the data to gain insights that can guide subsequent analysis and modeling.

Visualization of categorical variables:

Code:



```
/*Pie charts*/
proc template;
    define statgraph SASStudio.Pie1;
        begingraph;
        entrytitle "Pie charts for Categorical Variables Treatment" / textattrs=(size=22);
        layout region;
        piechart category=Treatment / dataskin=gloss;
        endlayout;
        endgraph;
    end;
run;
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgrender template=SASStudio.Pie1 data=treatments;
run;

proc template;
    define statgraph SASStudio.Pie2;
        begingraph;
        entrytitle "Pie charts for Categorical Variable Age" / textattrs=(size=22);
        layout region;
        piechart category=Age / dataskin=gloss;
        endlayout;
        endgraph;
    end;
run;
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgrender template=SASStudio.Pie2 data=treatments;
run;
ods graphics / reset;
```

Output:

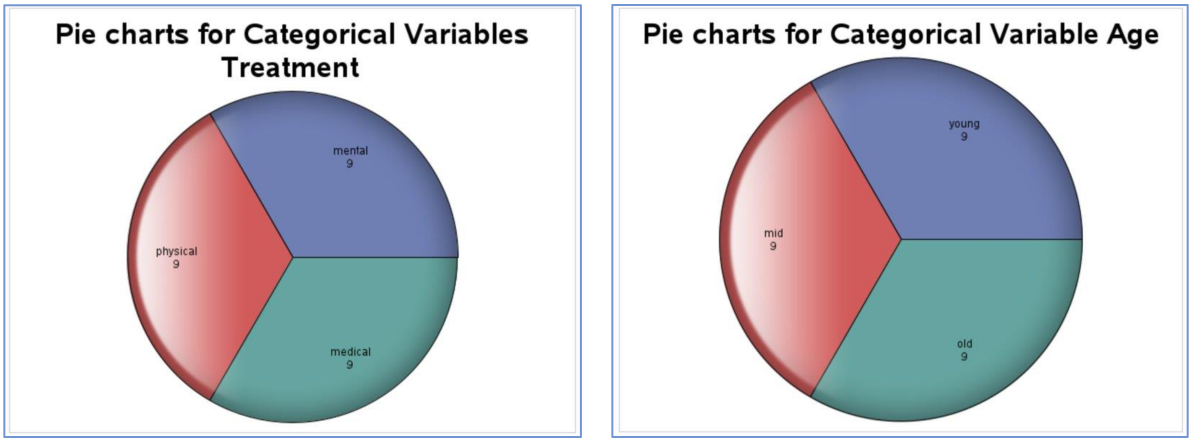
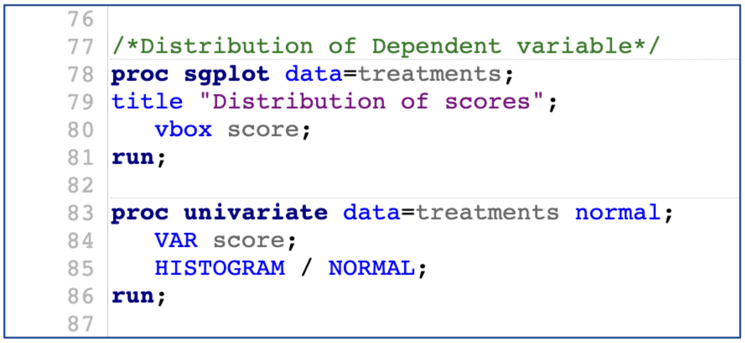


Fig.4: Visualization of categorical variables

From Fig.4, we can see those levels (medical, mental and physical, young, old, mid) in treatment and age variables are equally distributed with 9 observations in each level.

Visual representation of Dependent variable:

Code:



```

/*Distribution of Dependent variable*/
proc sgplot data=treatments;
title "Distribution of scores";
   vbox score;
run;

proc univariate data=treatments normal;
   VAR score;
   HISTOGRAM / NORMAL;
run;

```

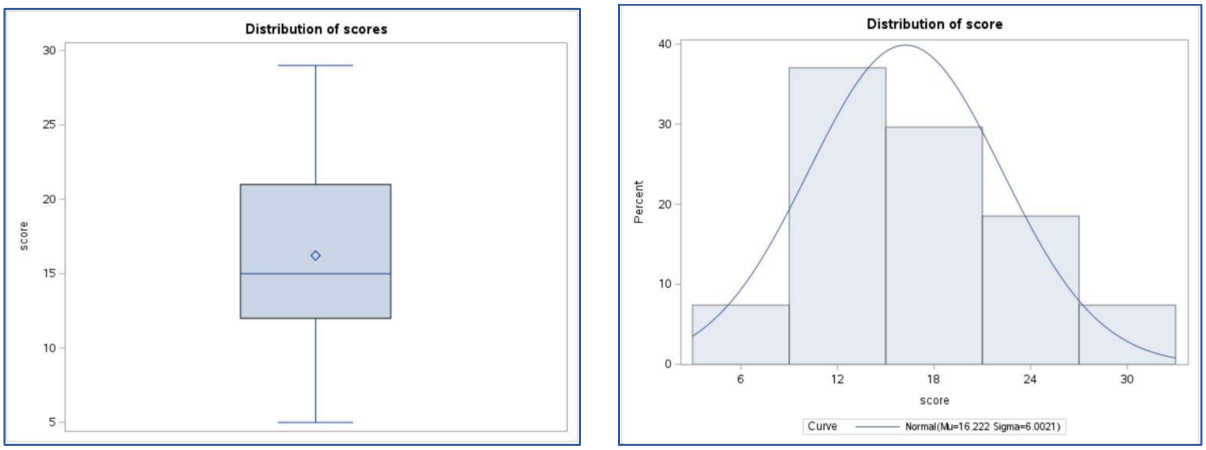
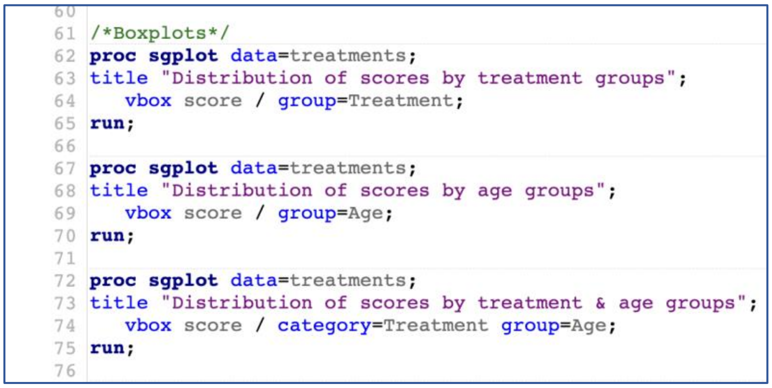


Fig.5: Visualization of dependent variable

From Fig.5, we can see that the distribution of dependent variable is normal, and it doesn’t have any outliers.

Visual representation of levels in categorical variables:

Code:



```
/*Boxplots*/
proc sgplot data=treatments;
title "Distribution of scores by treatment groups";
   vbox score / group=Treatment;
run;

proc sgplot data=treatments;
title "Distribution of scores by age groups";
   vbox score / group=Age;
run;

proc sgplot data=treatments;
title "Distribution of scores by treatment & age groups";
   vbox score / category=Treatment group=Age;
run;

```

Output:

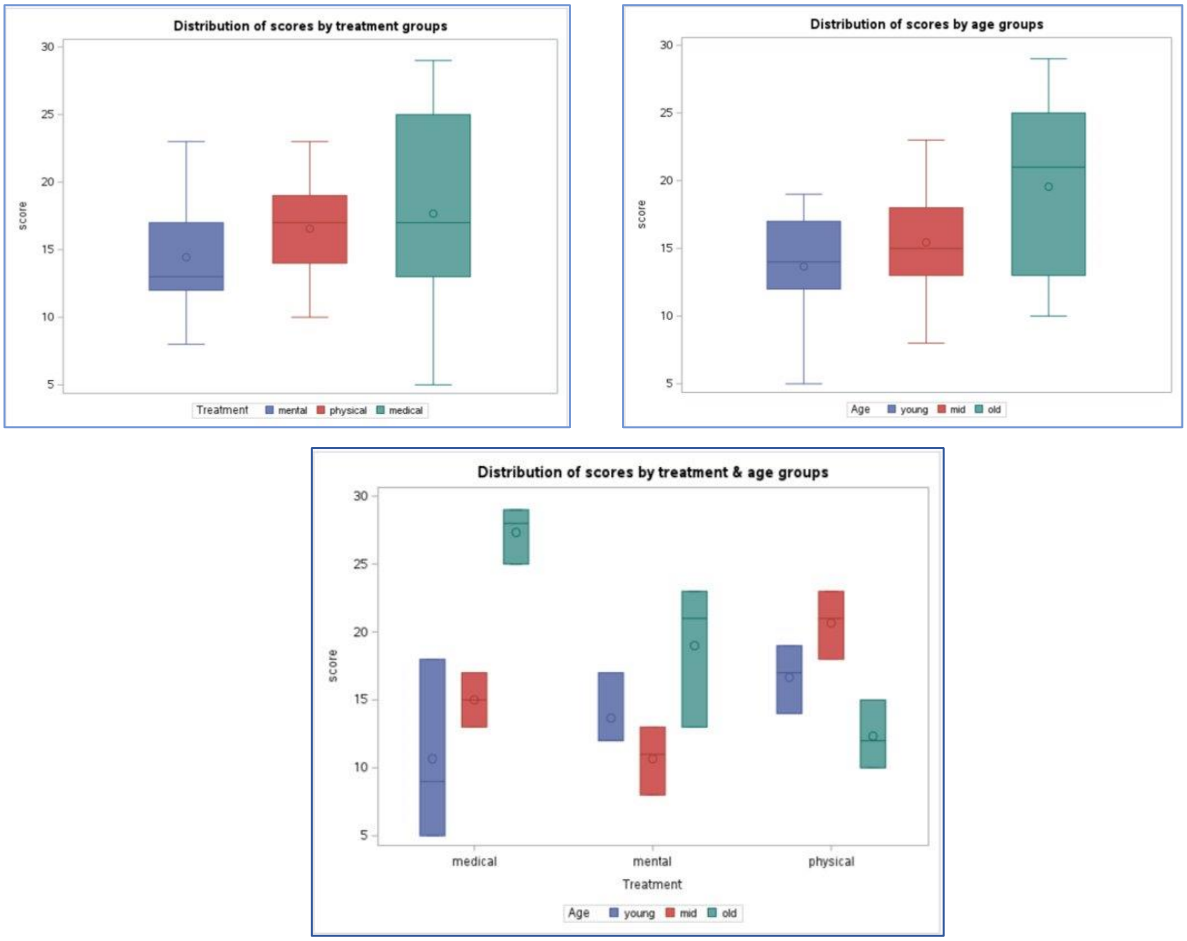


Fig.6: Boxplots of distributions of scores among different groups.

From the above boxplots, we can see that there are no outliers in any of the boxplots of distributions of scores among different groups.

Two Way Analysis of Variance:

Two-way Analysis of Variance (ANOVA) is a statistical method used to assess how two independent categorical variables (factors) simultaneously impact a continuous dependent variable. The primary goal is to determine whether there are significant differences in the means of the dependent variable across different levels of each factor and whether there is an interaction effect between the factors.

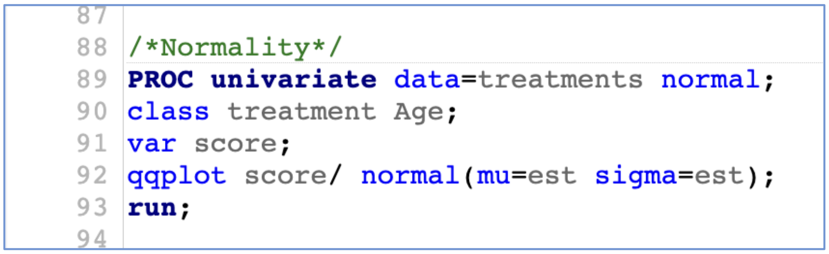
Assumptions of 2-way Anova (at alpha=0.01):

1. Normality
2. Homogeneity of Variances
3. Independence

Normality Assumption:

- **Null Hypothesis (H0):** The scores within each combination of factor levels are normally distributed.
- **Alternative Hypothesis (Ha):** The scores within each combination of factor levels are not normally distributed.

Code:

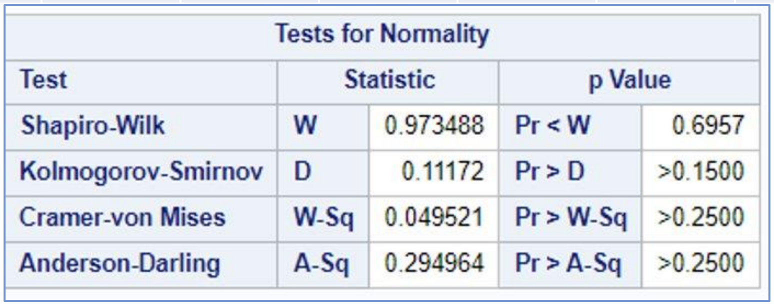


```

/*Normality*/
PROC univariate data=treatments normal;
class treatment Age;
var score;
qqplot score/ normal(mu=est sigma=est);
run;

```

Output:

These are shapiro-wilk p_values of all dependent variable across different levels of each factor.

Medical			Mental			Physical		
Mid	Old	Young	Mid	Old	Young	Mid	Old	Young
1.000	0.463	0.582	0.780	0.363	0.184	0.780	0.780	0.780



Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.973488	Pr < W	0.6957
Kolmogorov-Smirnov	D	0.11172	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.049521	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.294964	Pr > A-Sq	>0.2500

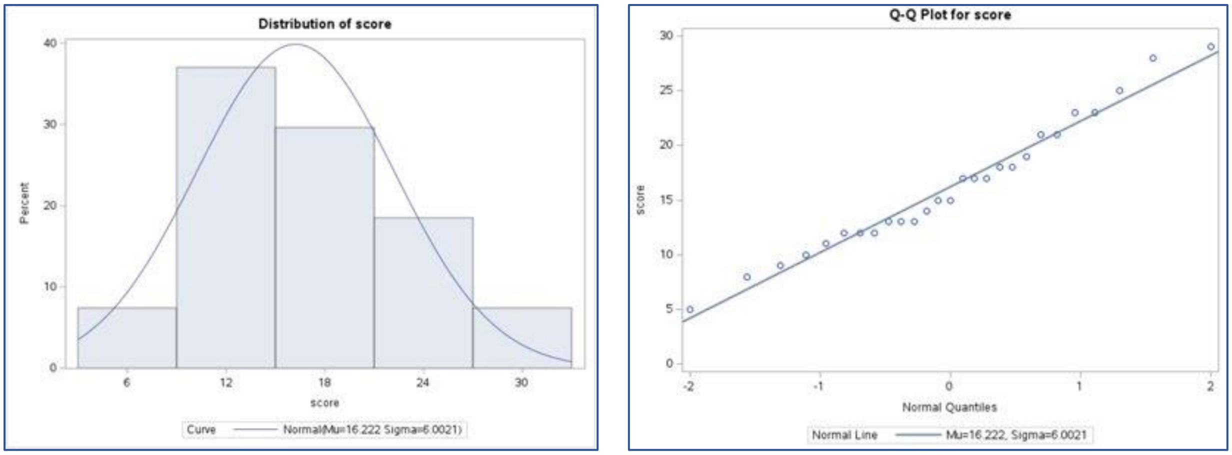


Fig.7: Normality assumption of response variable (scores)

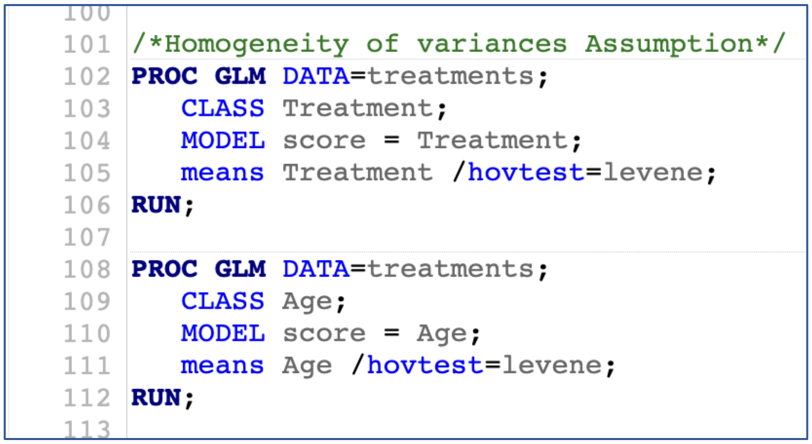
Interpretation:

Above all p-values are greater than alpha 0.01, so there is no sufficient evidence to reject the null hypothesis that means the data is normally distributed.

Homogeneity of Variances Assumption:

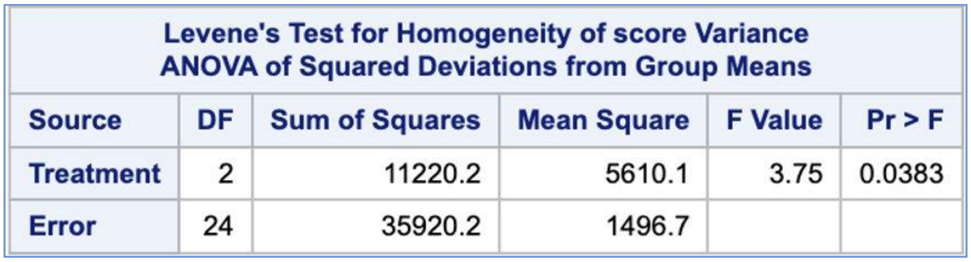
- **Null Hypothesis (H0):** The variances of the scores are equal across all combinations of factor levels.
- **Alternative Hypothesis (Ha):** The variances of the scores differ across at least one combination of factor levels.

Code:



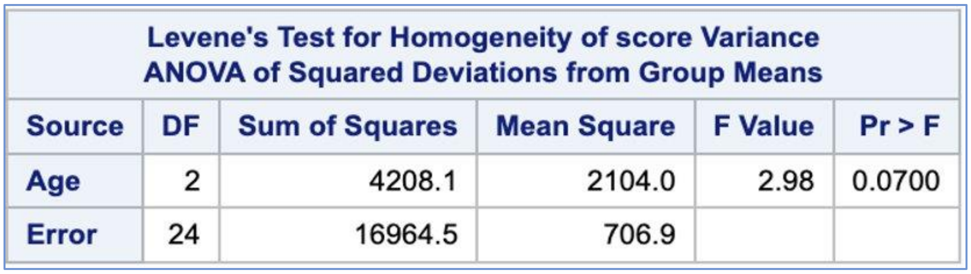
```
/*Homogeneity of variances Assumption*/
PROC GLM DATA=treatments;
   CLASS Treatment;
   MODEL score = Treatment;
   means Treatment /hovtest=levene;
RUN;

PROC GLM DATA=treatments;
   CLASS Age;
   MODEL score = Age;
   means Age /hovtest=levene;
RUN;
```

Output:



Levene's Test for Homogeneity of score Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Treatment	2	11220.2	5610.1	3.75	0.0383
Error	24	35920.2	1496.7		



Levene's Test for Homogeneity of score Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Age	2	4208.1	2104.0	2.98	0.0700
Error	24	16964.5	706.9		

Fig.8: Levene's test for Homogeneity of variances

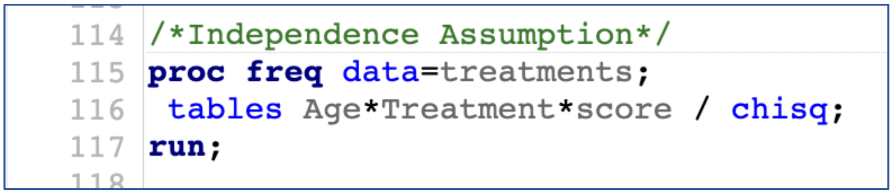
Interpretation:

From the above levene's test, we have p-values (0.0383, 0.070) greater than alpha 0.01, so there is no sufficient evidence to reject the null hypothesis which means that the variances of the scores are equal

Independence Assumption:

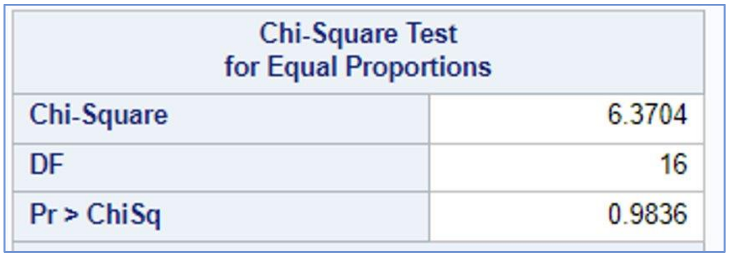
- **Null Hypothesis (H0):** Observations within each combination of factor levels are independent of each other.
- **Alternative Hypothesis (Ha):** There is a relationship between observations within at least one combination of factor levels.

Code:



```
/*Independence Assumption*/
proc freq data=treatments;
 tables Age*Treatment*score / chisq;
run;
```

Output:



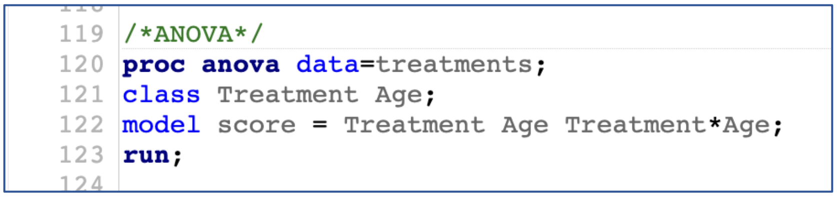
Chi-Square Test for Equal Proportions	
Chi-Square	6.3704
DF	16
Pr > ChiSq	0.9836

Fig.9: Chi-square test for Independence assumption

Interpretation:

From the above Chi-square test, the p-value (0.9836) is greater than alpha 0.01, so there is no sufficient evidence to reject the null hypothesis which means that the observations are independent of each other.

Two-way ANOVA Results:



```
/*ANOVA*/
proc anova data=treatments;
class Treatment Age;
model score = Treatment Age Treatment*Age;
run;
```

Output:

Fig.10: Two-way ANOVA results

Interpretation:

The overall model is statistically significant (F(8, 18) = 6.97, p = 0.0003), indicating that at least one of the factors (Treatment, Age, or their interaction) significantly influences the **scores.**

1. Main Effects:

Treatment Effect:

- The Treatment factor is not statistically significant (F(2, 18) = 1.90, p = 0.1787). There is no sufficient evidence to reject the null hypothesis that there is no significant difference in scores among individuals receiving different treatments**.**

Age Effect:

- The Age factor is statistically significant (F(2, 18) = 6.46, p = 0.0077). There is enough evidence to reject the null hypothesis, indicating that there is a significant difference in scores among individuals in different age groups.

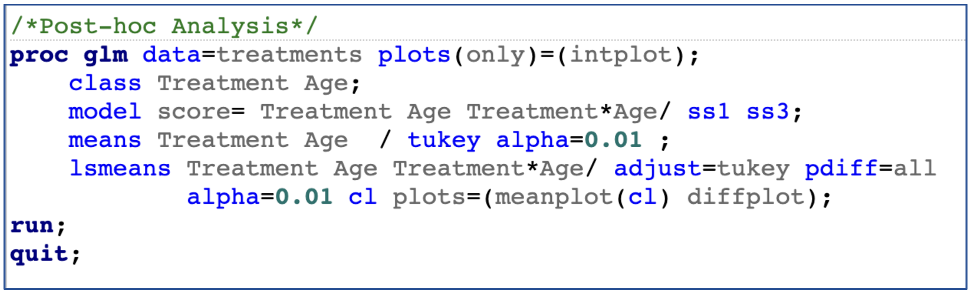
2. Interaction Effect:

- The interaction between Treatment and Age is statistically significant (F(4, 18) = 9.75, p = 0.0002). This suggests that the combined effect of Treatment and Age significantly influences the scores. The interaction effect indicates that the relationship between Treatment and scores varies across different Age groups, and vice versa.

Post-HOC Analysis:

Post-hoc analysis, short for post hoc tests, refers to a set of statistical tests that are conducted after an initial analysis (such as an analysis of variance - ANOVA) has been performed. The primary purpose of post-hoc tests is to explore and identify specific group differences when the overall comparison suggests there is a statistically significant effect**.**

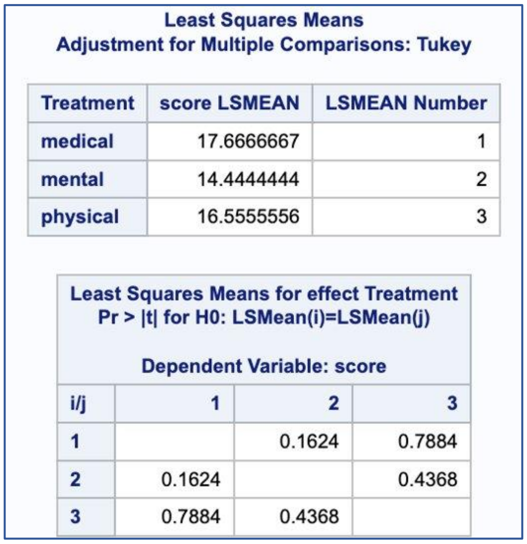
Code:



```
/*Post-hoc Analysis*/
proc glm data=treatments plots(only)=(intplot);
    class Treatment Age;
    model score= Treatment Age Treatment*Age/ ss1 ss3;
    means Treatment Age  / tukey alpha=0.01 ;
    lsmeans Treatment Age Treatment*Age/ adjust=tukey pdiff=all
            alpha=0.01 cl plots=(meanplot(cl) diffplot);
run;
quit;
```

Treatment:

Based on the Tukey's post-hoc analysis, there is no evidence of a significant difference in scores between any pair of treatment groups.



Least Squares Means
Adjustment for Multiple Comparisons: Tukey

Treatment	score LSMEAN	LSMEAN Number
medical	17.6666667	1
mental	14.4444444	2
physical	16.5555556	3

Least Squares Means for effect Treatment
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: score

i/j	1	2	3
1		0.1624	0.7884
2	0.1624		0.4368
3	0.7884	0.4368	

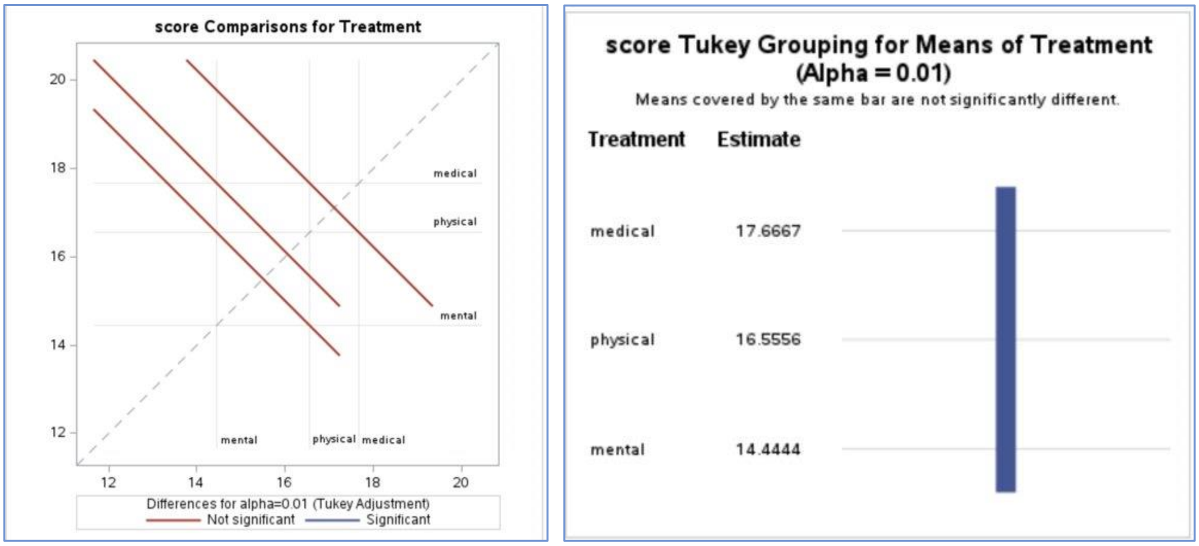
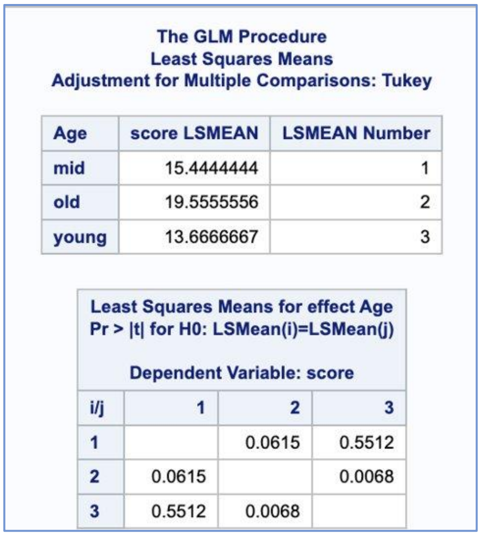


Fig.11: Post-hoc analysis of treatment group

Age:

Based on the Tukey's post-hoc analysis, there is evidence of a significant difference in scores between the "old" and "young" age groups. However, there is no significant difference between the "mid" age group and either the "old" or "young" age groups.



The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

Age	score LSMEAN	LSMEAN Number
mid	15.4444444	1
old	19.5555556	2
young	13.6666667	3

Least Squares Means for effect Age
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: score

i/j	1	2	3
1		0.0615	0.5512
2	0.0615		0.0068
3	0.5512	0.0068	

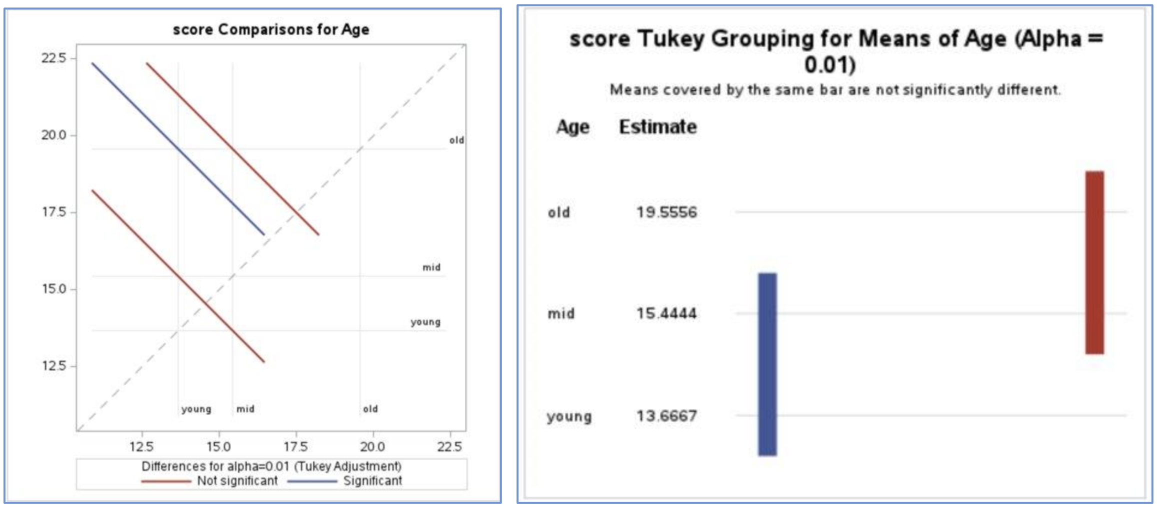
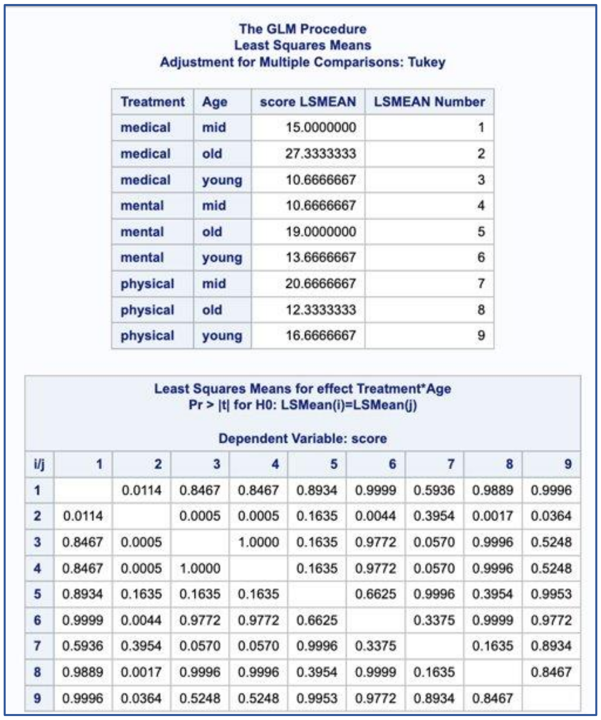


Fig.12: Post-hoc analysis of age group

Interaction:

The results suggest that the interaction between Treatment and Age has a significant impact on the scores. Different combinations of Treatment and Age groups exhibit varying scores, as evidenced by the significant p-values in the post-hoc analysis.



The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

Treatment	Age	score LSMEAN	LSMEAN Number
medical	mid	15.0000000	1
medical	old	27.3333333	2
medical	young	10.6666667	3
mental	mid	10.6666667	4
mental	old	19.0000000	5
mental	young	13.6666667	6
physical	mid	20.6666667	7
physical	old	12.3333333	8
physical	young	16.6666667	9

Least Squares Means for effect Treatment*Age
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: score

i/j	1	2	3	4	5	6	7	8	9
1		0.0114	0.8467	0.8467	0.8934	0.9999	0.5936	0.9889	0.9996
2	0.0114		0.0005	0.0005	0.1635	0.0044	0.3954	0.0017	0.0364
3	0.8467	0.0005		1.0000	0.1635	0.9772	0.0570	0.9996	0.5248
4	0.8467	0.0005	1.0000		0.1635	0.9772	0.0570	0.9996	0.5248
5	0.8934	0.1635	0.1635	0.1635		0.6625	0.9996	0.3954	0.9953
6	0.9999	0.0044	0.9772	0.9772	0.6625		0.3375	0.9999	0.9772
7	0.5936	0.3954	0.0570	0.0570	0.9996	0.3375		0.1635	0.8934
8	0.9889	0.0017	0.9996	0.9996	0.3954	0.9999	0.1635		0.8467
9	0.9996	0.0364	0.5248	0.5248	0.9953	0.9772	0.8934	0.8467	

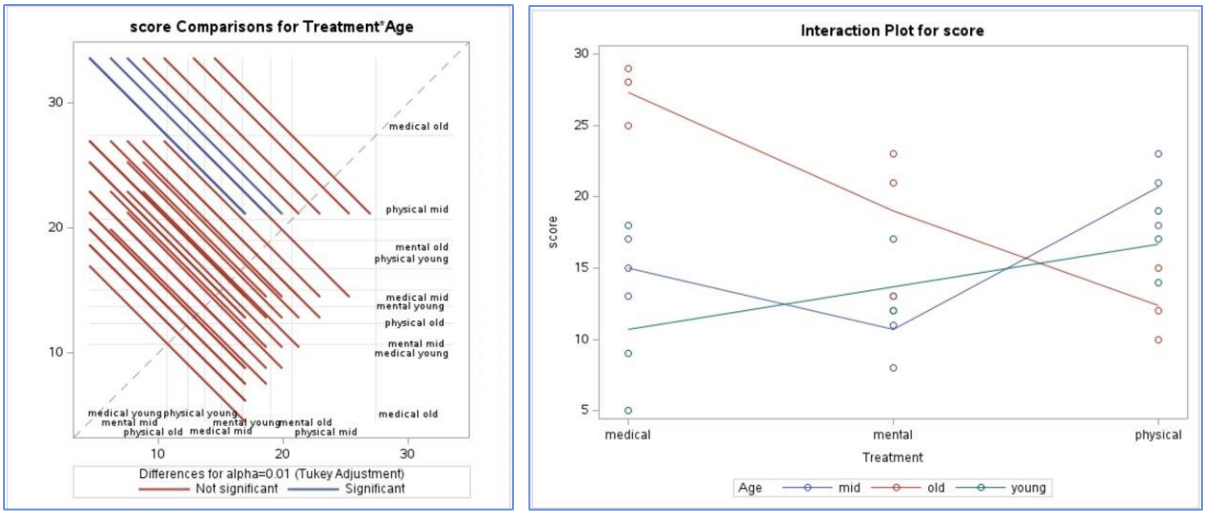
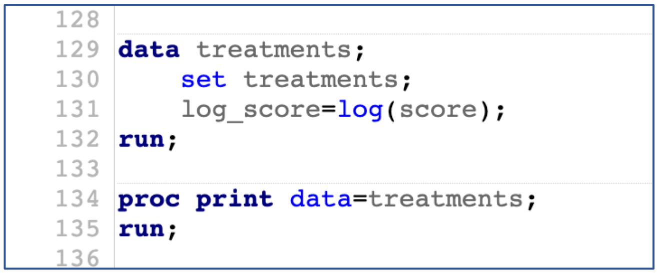


Fig.13: Post-hoc analysis of interaction between treatment and age

Further Analysis (For Future Purpose):

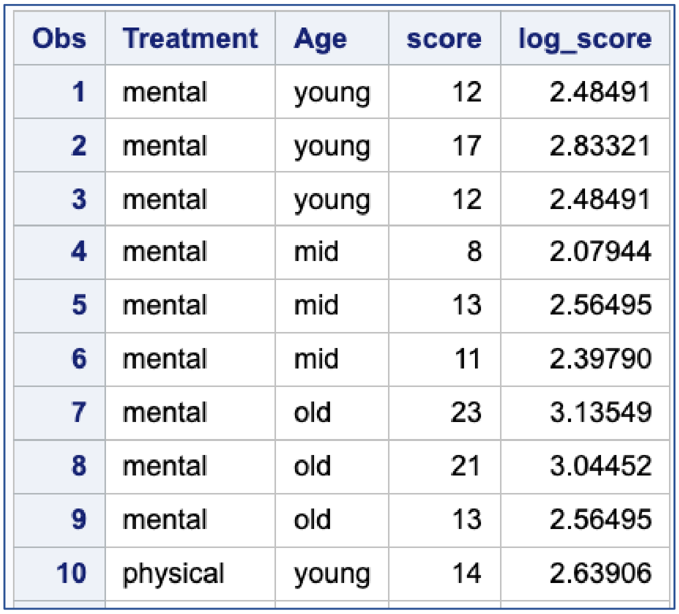
Currently, we did a two-way analysis for significance level 0.01. We can also do the same for significance level 0.05 but if we check the previous results the assumption of homogeneity of variances alone fails, apart from all other observations will pass for alpha=0.05. So, for that we can do log transformation for scores then we can perform homogeneity of variances and two-way Anova on log_scores as a response variable.

Code:



```
data treatments;
    set treatments;
    log_score=log(score);
run;

proc print data=treatments;
run;
```

Output:

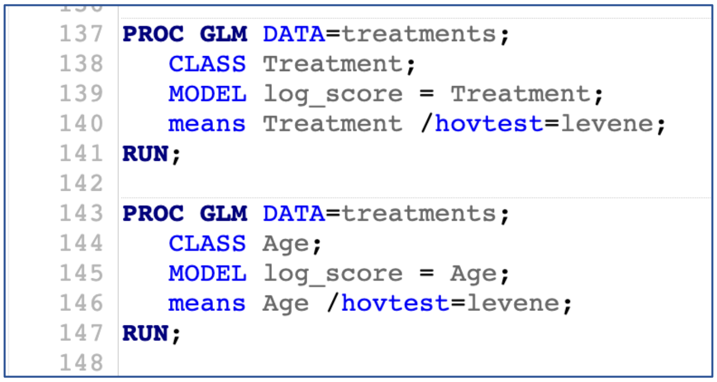


Obs	Treatment	Age	score	log_score
1	mental	young	12	2.48491
2	mental	young	17	2.83321
3	mental	young	12	2.48491
4	mental	mid	8	2.07944
5	mental	mid	13	2.56495
6	mental	mid	11	2.39790
7	mental	old	23	3.13549
8	mental	old	21	3.04452
9	mental	old	13	2.56495
10	physical	young	14	2.63906

Fig.14: Dataset overview after log transformation

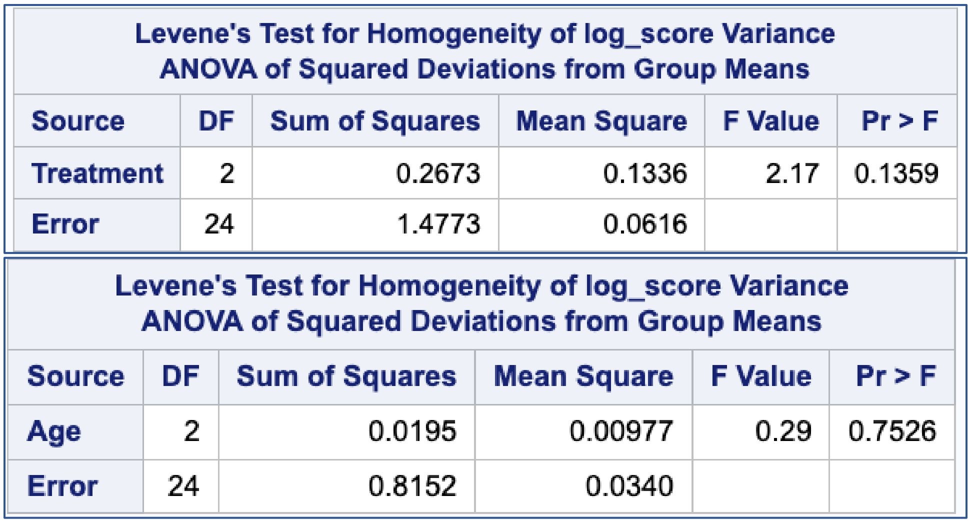
Homogeneity of variances:

Code:



```
PROC GLM DATA=treatments;
   CLASS Treatment;
   MODEL log_score = Treatment;
   means Treatment /hovtest=levene;
RUN;

PROC GLM DATA=treatments;
   CLASS Age;
   MODEL log_score = Age;
   means Age /hovtest=levene;
RUN;
```

Output:



Levene's Test for Homogeneity of log_score Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Treatment	2	0.2673	0.1336	2.17	0.1359
Error	24	1.4773	0.0616		

Levene's Test for Homogeneity of log_score Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Age	2	0.0195	0.00977	0.29	0.7526
Error	24	0.8152	0.0340		

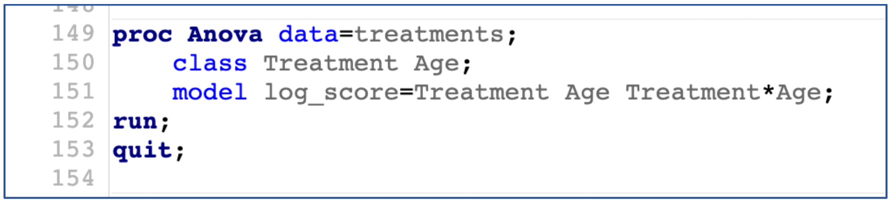
Fig.15: Variance test at significance level 0.05

Interpretation:

From the above levene's test, we have p-values (0.1359, 0.7526) greater than alpha 0.05, so there is no sufficient evidence to reject the null hypothesis which means that the variances of the scores are equal.

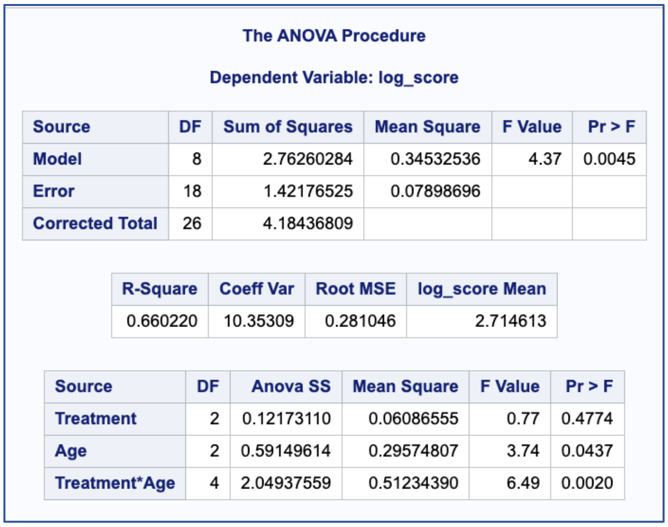
Two-way Anova:

Code:



```
proc Anova data=treatments;
    class Treatment Age;
    model log_score=Treatment Age Treatment*Age;
run;
quit;
```

Output:



The ANOVA Procedure

Dependent Variable: log_score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2.76260284	0.34532536	4.37	0.0045
Error	18	1.42176525	0.07898696		
Corrected Total	26	4.18436809			

R-Square	Coeff Var	Root MSE	log_score Mean
0.660220	10.35309	0.281046	2.714613

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Treatment	2	0.12173110	0.06086555	0.77	0.4774
Age	2	0.59149614	0.29574807	3.74	0.0437
Treatment*Age	4	2.04937559	0.51234390	6.49	0.0020

Fig.16: Two-way ANOVA results at significance level 0.05 after log Transform.

Interpretation:

The overall model is statistically significant (p = 0.0045), indicating that at least one of the factors (Treatment, Age, or their interaction) significantly influences the **scores.**

1. Main Effects:

Treatment Effect:

- The Treatment factor is not statistically significant (p = 0.4774). There is no sufficient evidence to reject the null hypothesis that there is no significant difference in scores among individuals receiving different treatments**.**

Age Effect:

- The Age factor is statistically significant (p = 0.0437). There is enough evidence to reject the null hypothesis, indicating that there is a significant difference in scores among individuals in different age groups.

2. Interaction Effect:

- The interaction between Treatment and Age is statistically significant (p = 0.0020). This suggests that the combined effect of Treatment and Age significantly influences the scores. The interaction effect indicates that the relationship between Treatment and scores varies across different Age groups, and vice versa.

Even at significance level 0.05 we have drawn similar conclusions as significance level 0.01. Here we are not performing post-hoc analysis now, because we did this as a first step for future analysis.

Conclusion:

The project's analysis using a two-way ANOVA with post-hoc testing underscores that there is no significant difference between treatment groups, but specific combinations of age exhibit notable differences, emphasizing the need for treatments; others show no significant variance. While individual treatment types may not significantly affect scores, the age of individuals and the interaction between treatment and age play significant roles in determining the outcomes. These findings underscore the importance of considering both age and the interaction between treatment and age when analyzing and interpreting the scores in this study.

References:

http://staff.pubhealth.ku.dk/~jufo/courses/vr/oldlectures/anova2010-2x2.pdf

DataSet: https://houssein-assaad.shinyapps.io/TwoWayANOVA/