Seneca
POLYTECHNIC

Statistics for Analytics

BAN100NAA - PROFESSOR: SAMANEH GHOLAMI

Maaz Hussain
STUDENT ID: 173714221

ASSIGNMENT 2:

ANOVA

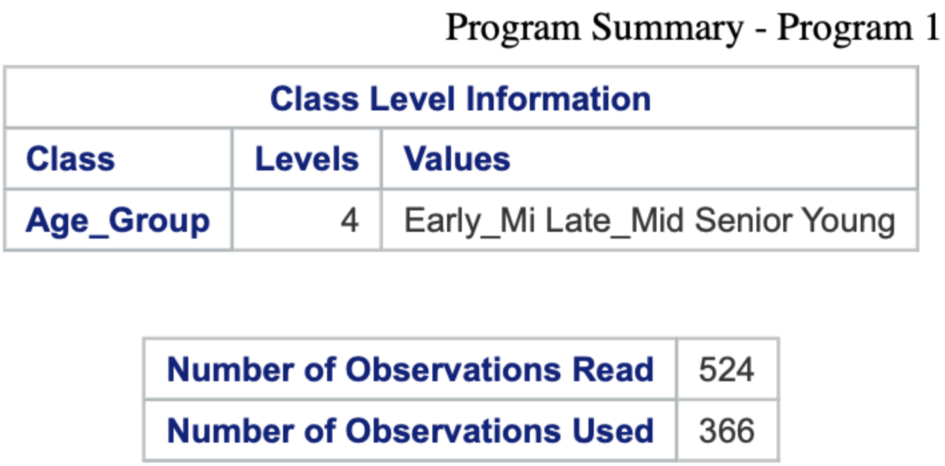
Problem 1:

ANOVA on Stock Ownership by Age Group

Introduction to ANOVA
Analysis of Variance (ANOVA) is a statistical method used to test whether there are any statistically significant differences between the means of three or more independent groups. In this case, we want to determine if stock ownership proportion varies significantly across four age groups: Young, Early Middle Age, Late Middle Age, and Senior.

Hypothesis

- Null Hypothesis (H0): There is no difference in the mean stock ownership among the age groups.
- Alternative Hypothesis (H1): At least one age group has a different mean stock ownership compared to others.



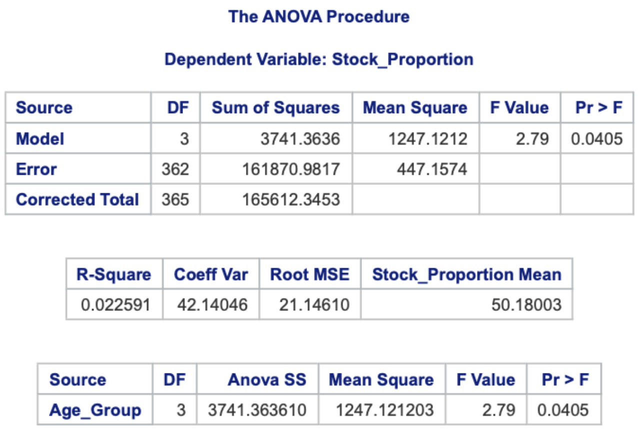
Program Summary - Program 1

Class Level Information		
Class	**Levels**	**Values**
Age_Group	4	Early_Mi Late_Mid Senior Young

Number of Observations Read	524
Number of Observations Used	366

Displays the four levels of Age Group (Young, Early Middle Age, Late Middle Age, Senior) and confirms the number of observations used in the analysis.

I also restructured the data set in order to make it work in the SAS code, so I grouped all the age classifications under one column named Age_Group to make the SAS code work properly.



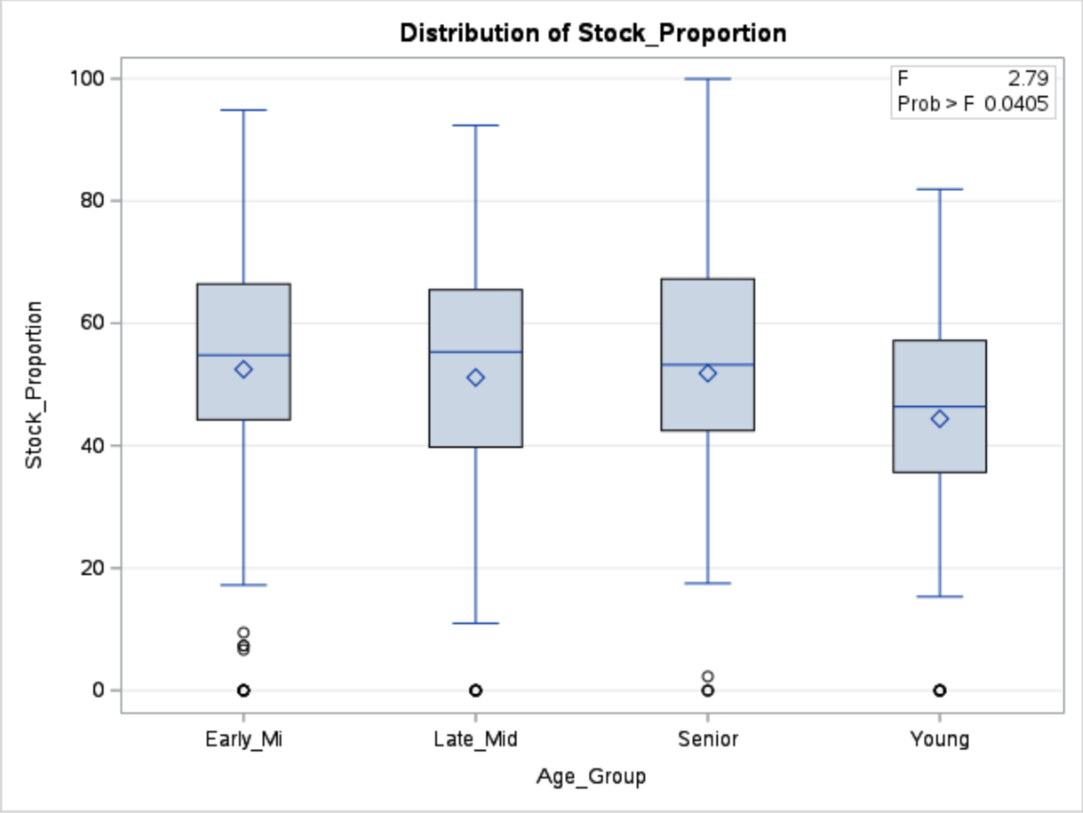
The ANOVA Procedure

Dependent Variable: Stock_Proportion

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3741.3636	1247.1212	2.79	0.0405
Error	362	161870.9817	447.1574		
Corrected Total	365	165612.3453			

R-Square	Coeff Var	Root MSE	Stock_Proportion Mean
0.022591	42.14046	21.14610	50.18003

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Age_Group	3	3741.363610	1247.121203	2.79	0.0405

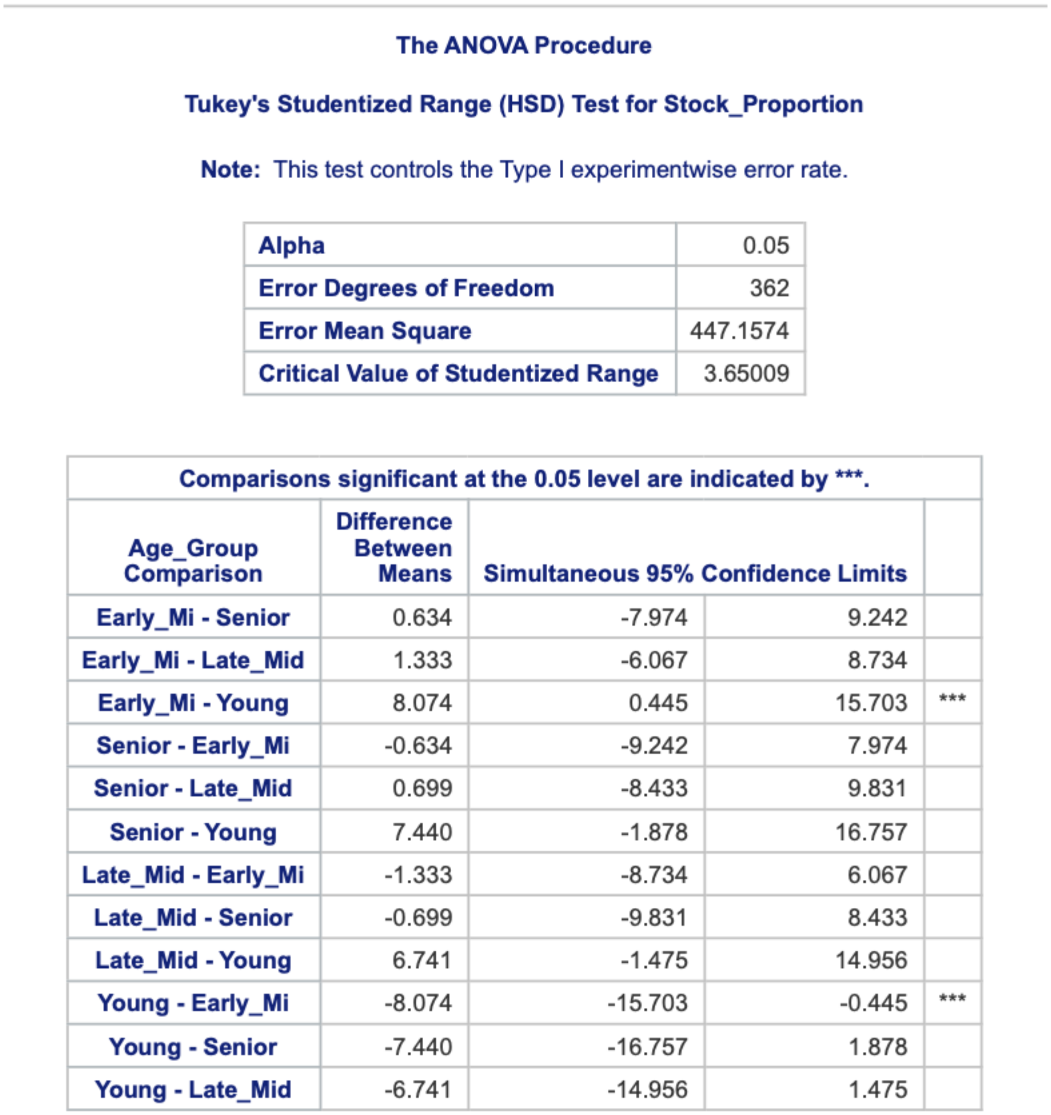


ANOVA Table:
Displays the degrees of freedom, sum of squares, mean square, F-value (2.79), and p-value (0.0405) for the factor Age_Group.

ANOVA Table Results
The ANOVA table shows the following key values:

- **F-Value**: 2.79
- **p-Value**: 0.0405

Since the p-value (0.0405) is less than 0.05, we reject the null hypothesis, indicating that there are statistically significant differences in stock ownership proportions across the age groups. This result suggests that at least one age group has a different average stock ownership level compared to the others.



The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for Stock_Proportion

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	362
Error Mean Square	447.1574
Critical Value of Studentized Range	3.65009

Comparisons significant at the 0.05 level are indicated by *.**

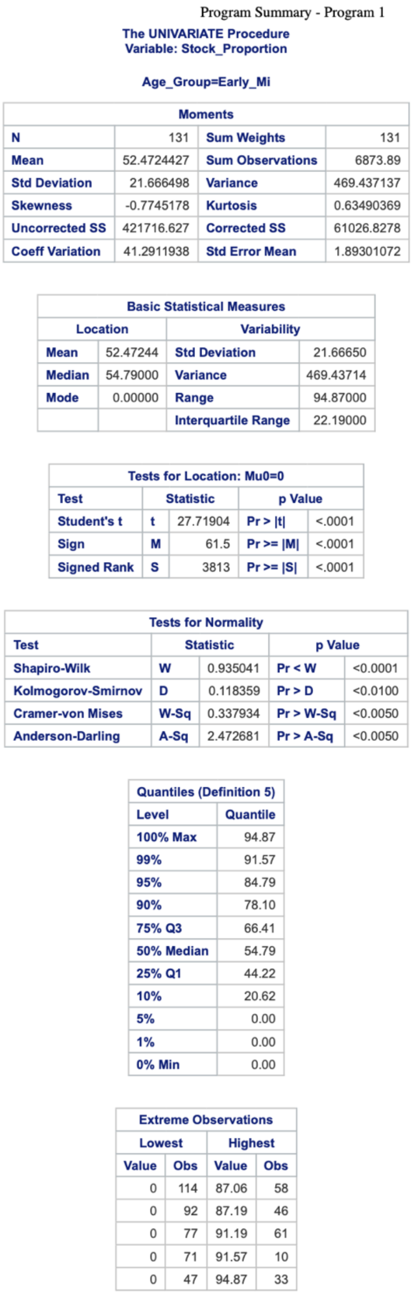
Age_Group Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Early_Mi - Senior	0.634	-7.974	9.242	
Early_Mi - Late_Mid	1.333	-6.067	8.734	
Early_Mi - Young	8.074	0.445	15.703	***
Senior - Early_Mi	-0.634	-9.242	7.974	
Senior - Late_Mid	0.699	-8.433	9.831	
Senior - Young	7.440	-1.878	16.757	
Late_Mid - Early_Mi	-1.333	-8.734	6.067	
Late_Mid - Senior	-0.699	-9.831	8.433	
Late_Mid - Young	6.741	-1.475	14.956	
Young - Early_Mi	-8.074	-15.703	-0.445	***
Young - Senior	-7.440	-16.757	1.878	
Young - Late_Mid	-6.741	-14.956	1.475	

Tukey's Post Hoc Test

Tukey's post hoc test is used to identify specific pairs of groups with significant differences in means. Here are the results for the comparisons:

- The "Young" and "Early Middle Age" groups show a statistically significant difference in stock ownership (mean difference = 8.074, $p < 0.05$).
- No other pairs show significant differences at the 0.05 level.

This indicates that the significant difference in stock ownership proportions is primarily between the "Young" and "Early Middle Age" groups, with "Young" individuals showing a different level of stock ownership than those in early middle age.



Program Summary - Program 1

The UNIVARIATE Procedure
Variable: Stock_Proportion

Age_Group=Early_Mi

Moments			
N	131	Sum Weights	131
Mean	52.4724427	Sum Observations	6873.89
Std Deviation	21.666498	Variance	469.437137
Skewness	-0.7745178	Kurtosis	0.63490369
Uncorrected SS	421716.627	Corrected SS	61026.8278
Coeff Variation	41.2911938	Std Error Mean	1.89301072

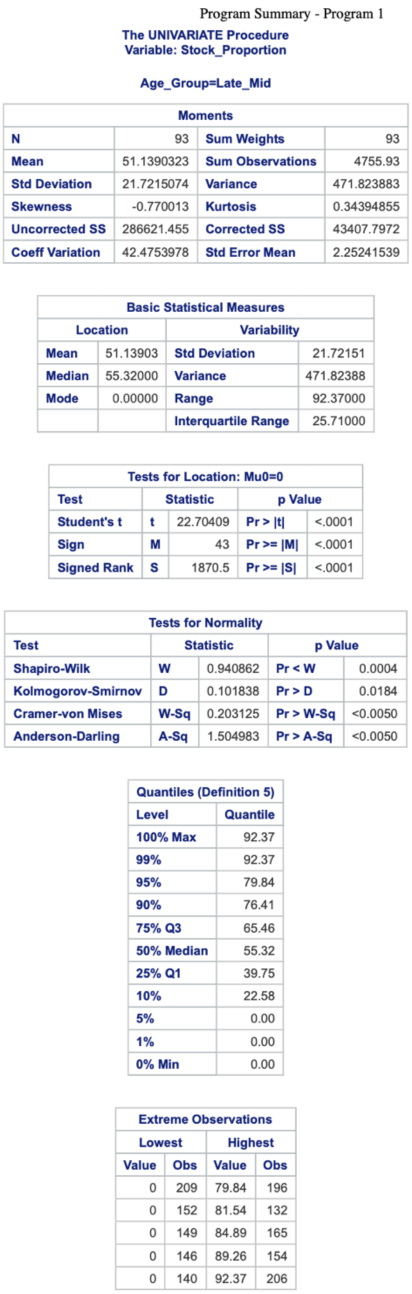
Basic Statistical Measures			
Location		Variability	
Mean	52.47244	Std Deviation	21.66650
Median	54.79000	Variance	469.43714
Mode	0.00000	Range	94.87000
		Interquartile Range	22.19000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	27.71904	Pr > |t|	<.0001
Sign	M	61.5	Pr >= |M|	<.0001
Signed Rank	S	3813	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.935041	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.118359	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.337934	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.472681	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Level	Quantile
100% Max	94.87
99%	91.57
95%	84.79
90%	78.10
75% Q3	66.41
50% Median	54.79
25% Q1	44.22
10%	20.62
5%	0.00
1%	0.00
0% Min	0.00

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	114	87.06	58
0	92	87.19	46
0	77	91.19	61
0	71	91.57	10
0	47	94.87	33



Program Summary - Program 1

The UNIVARIATE Procedure
Variable: Stock_Proportion

Age_Group=Late_Mid

Moments			
N	93	Sum Weights	93
Mean	51.1390323	Sum Observations	4755.93
Std Deviation	21.7215074	Variance	471.823883
Skewness	-0.770013	Kurtosis	0.34394855
Uncorrected SS	286621.455	Corrected SS	43407.7972
Coeff Variation	42.4753978	Std Error Mean	2.25241539

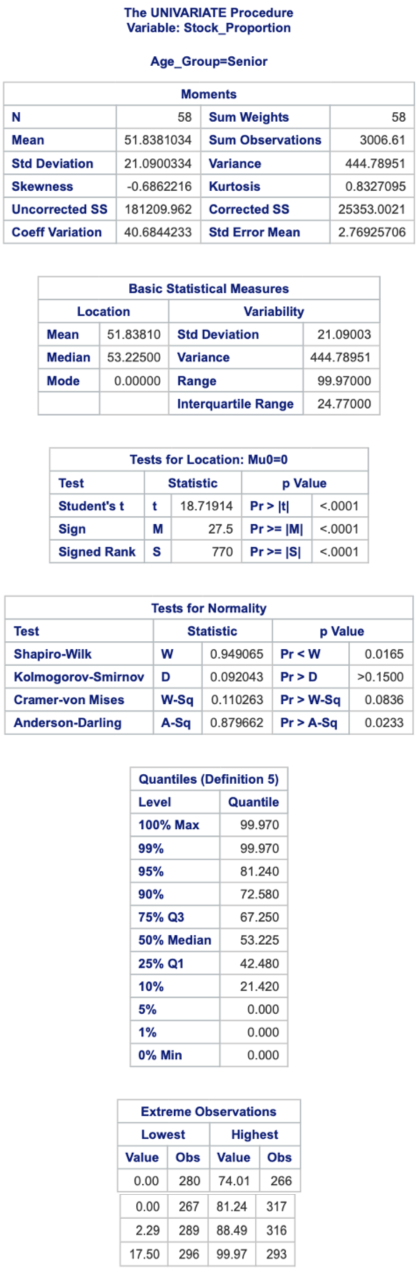
Basic Statistical Measures			
Location		Variability	
Mean	51.13903	Std Deviation	21.72151
Median	55.32000	Variance	471.82388
Mode	0.00000	Range	92.37000
		Interquartile Range	25.71000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	22.70409	Pr > |t|	<.0001
Sign	M	43	Pr >= |M|	<.0001
Signed Rank	S	1870.5	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.940862	Pr < W	0.0004
Kolmogorov-Smirnov	D	0.101838	Pr > D	0.0184
Cramer-von Mises	W-Sq	0.203125	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.504983	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Level	Quantile
100% Max	92.37
99%	92.37
95%	79.84
90%	76.41
75% Q3	65.46
50% Median	55.32
25% Q1	39.75
10%	22.58
5%	0.00
1%	0.00
0% Min	0.00

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	209	79.84	196
0	152	81.54	132
0	149	84.89	165
0	146	89.26	154
0	140	92.37	206



The UNIVARIATE Procedure
Variable: Stock_Proportion

Age_Group=Senior

Moments			
N	58	Sum Weights	58
Mean	51.8381034	Sum Observations	3006.61
Std Deviation	21.0900334	Variance	444.78951
Skewness	-0.6862216	Kurtosis	0.8327095
Uncorrected SS	181209.962	Corrected SS	25353.0021
Coeff Variation	40.6844233	Std Error Mean	2.76925706

Basic Statistical Measures			
Location		Variability	
Mean	51.83810	Std Deviation	21.09003
Median	53.22500	Variance	444.78951
Mode	0.00000	Range	99.97000
		Interquartile Range	24.77000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	18.71914	Pr > |t|	<.0001
Sign	M	27.5	Pr >= |M|	<.0001
Signed Rank	S	770	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.949065	Pr < W	0.0165
Kolmogorov-Smirnov	D	0.092043	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.110263	Pr > W-Sq	0.0836
Anderson-Darling	A-Sq	0.879662	Pr > A-Sq	0.0233

Quantiles (Definition 5)	
Level	Quantile
100% Max	99.970
99%	99.970
95%	81.240
90%	72.580
75% Q3	67.250
50% Median	53.225
25% Q1	42.480
10%	21.420
5%	0.000
1%	0.000
0% Min	0.000

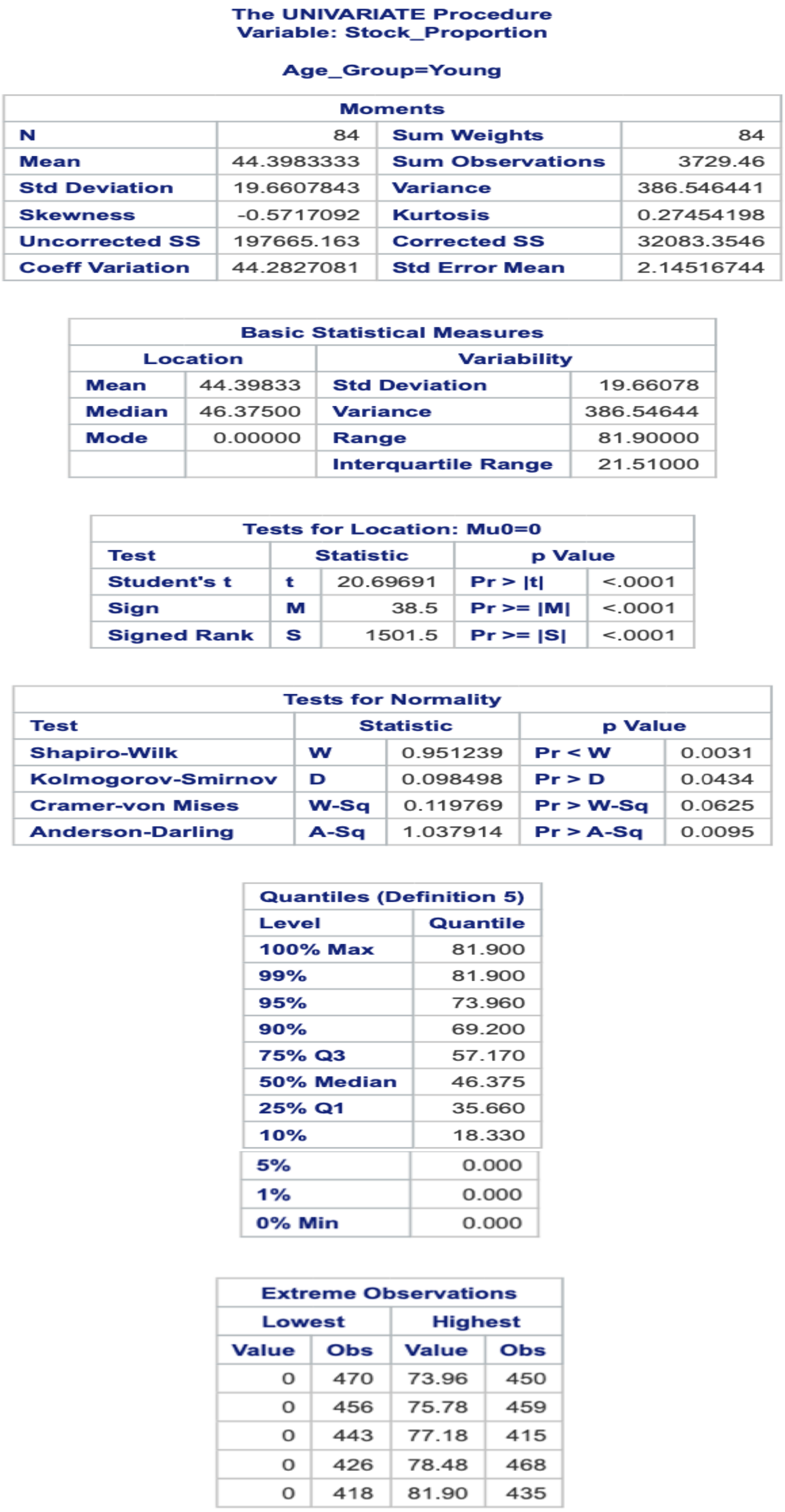
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.00	280	74.01	266
0.00	267	81.24	317
2.29	289	88.49	316
17.50	296	99.97	293

Descriptive Statistics for Each Age Group (UNIVARIATE Procedure):

- Explanation: Summarizes statistical measures (mean, standard deviation, skewness, kurtosis) for stock ownership within each age group.
- Interpretation: Provides a clearer understanding of each age group's distribution, confirming differences in stock ownership tendencies.

Normality Test Results:

- Explanation: Provides p-values for normality tests (e.g., Shapiro-Wilk) for each age group's stock ownership data.
- Interpretation: Low p-values (<0.05) indicate slight deviations from normality, although ANOVA is still reliable in this case.



The UNIVARIATE Procedure
Variable: Stock_Proportion

Age_Group=Young

Moments			
N	84	Sum Weights	84
Mean	44.3983333	Sum Observations	3729.46
Std Deviation	19.6607843	Variance	386.546441
Skewness	-0.5717092	Kurtosis	0.27454198
Uncorrected SS	197665.163	Corrected SS	32083.3546
Coeff Variation	44.2827081	Std Error Mean	2.14516744

Basic Statistical Measures			
Location		Variability	
Mean	44.39833	Std Deviation	19.66078
Median	46.37500	Variance	386.54644
Mode	0.00000	Range	81.90000
		Interquartile Range	21.51000

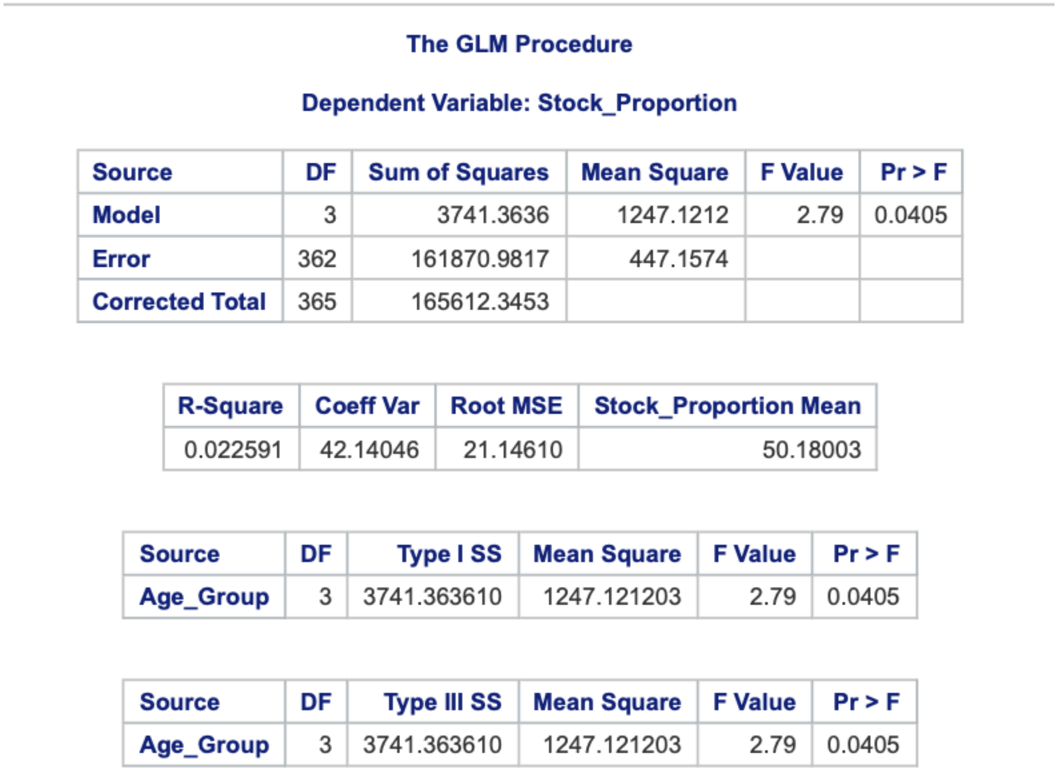
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	20.69691	Pr > |t|	<.0001
Sign	M	38.5	Pr >= |M|	<.0001
Signed Rank	S	1501.5	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.951239	Pr < W	0.0031
Kolmogorov-Smirnov	D	0.098498	Pr > D	0.0434
Cramer-von Mises	W-Sq	0.119769	Pr > W-Sq	0.0625
Anderson-Darling	A-Sq	1.037914	Pr > A-Sq	0.0095

Quantiles (Definition 5)	
Level	Quantile
100% Max	81.900
99%	81.900
95%	73.960
90%	69.200
75% Q3	57.170
50% Median	46.375
25% Q1	35.660
10%	18.330
5%	0.000
1%	0.000
0% Min	0.000

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	470	73.96	450
0	456	75.78	459
0	443	77.18	415
0	426	78.48	468
0	418	81.90	435

Each age group's normality was assessed using the Shapiro-Wilk test. All age groups have p-values below 0.05, indicating deviations from normality. However, ANOVA is generally robust to minor deviations from normality, so this result is not critical.



The GLM Procedure

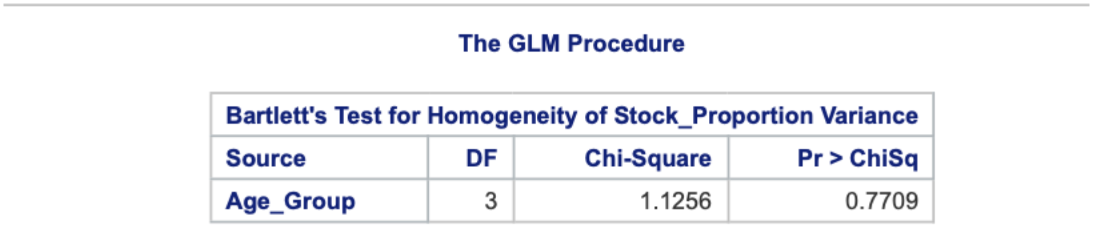
Dependent Variable: Stock_Proportion

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3741.3636	1247.1212	2.79	0.0405
Error	362	161870.9817	447.1574		
Corrected Total	365	165612.3453			

R-Square	Coeff Var	Root MSE	Stock_Proportion Mean
0.022591	42.14046	21.14610	50.18003

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Age_Group	3	3741.363610	1247.121203	2.79	0.0405

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Age_Group	3	3741.363610	1247.121203	2.79	0.0405



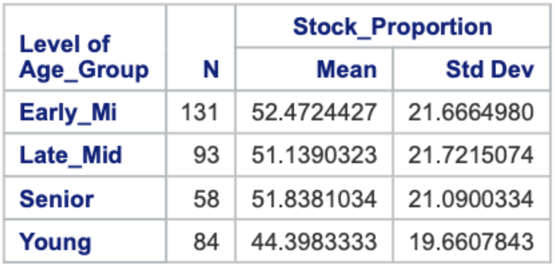
The GLM Procedure

Bartlett's Test for Homogeneity of Stock_Proportion Variance			
Source	DF	Chi-Square	Pr > ChiSq
Age_Group	3	1.1256	0.7709

Bartlett's Test for Homogeneity of Variances:

- Explanation: This test shows a Chi-square statistic with a p-value (0.7709) to assess if variances are equal across age groups.
- Interpretation: The high p-value (above 0.05) indicates that the assumption of equal variances is met, validating the ANOVA results.

Homogeneity of Variances: Bartlett's test provides a p-value of 0.7709, which is greater than 0.05, indicating that the variances across the age groups are similar. This meets the homogeneity of variances assumption required for ANOVA.



Level of Age_Group	N	Stock_Proportion	
		Mean	Std Dev
Early_Mi	131	52.4724427	21.6664980
Late_Mid	93	51.1390323	21.7215074
Senior	58	51.8381034	21.0900334
Young	84	44.3983333	19.6607843

Means and Standard Deviations by Age Group:

- Explanation: Lists the mean stock ownership and standard deviation for each age group.
- Interpretation: This highlights the differences in average stock ownership across groups, supporting the significant ANOVA findings.

Conclusion

In conclusion, there is a statistically significant difference in stock ownership proportions across the age groups, specifically between the "Young" and "Early Middle Age" groups. While there is some deviation from normality, the homogeneity of variance assumption is met, supporting the reliability of these ANOVA results.

Problem 2

Effect of Gender and Education on Number of Jobs Held

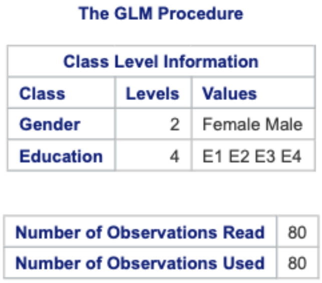
Introduction to Two-Way ANOVA
Two-way ANOVA is used to determine the effects of two factors (in this case, **Gender** and **Education Level**) on a dependent variable (number of jobs held). This analysis allows us to explore both main effects of each factor individually and any interaction effect between them.

Hypothesis

1. **Interaction Effect**:
   - Null Hypothesis (H0): There is no interaction between gender and education level in terms of job holding.
   - Alternative Hypothesis (H1): There is an interaction between gender and education level in terms of job holding.

2. **Main Effect of Gender**:
   - Null Hypothesis (H0): There is no difference in job holding between men and women.
   - Alternative Hypothesis (H1): There is a difference in job holding between men and women.

3. **Main Effect of Education**:
   - Null Hypothesis (H0): There is no difference in job holding across educational levels.
   - Alternative Hypothesis (H1): There is a difference in job holding across educational levels.



The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
Gender	2	Female Male
Education	4	E1 E2 E3 E4

Number of Observations Read	80
Number of Observations Used	80

Class Level Information Table:
- Explanation: Displays the categories for Gender (Male, Female) and Education (E1, E2, E3, E4) along with the total number of observations.
- Interpretation: Confirms the categorical breakdown used in the two-way ANOVA for understanding job-holding patterns.

The GLM Procedure

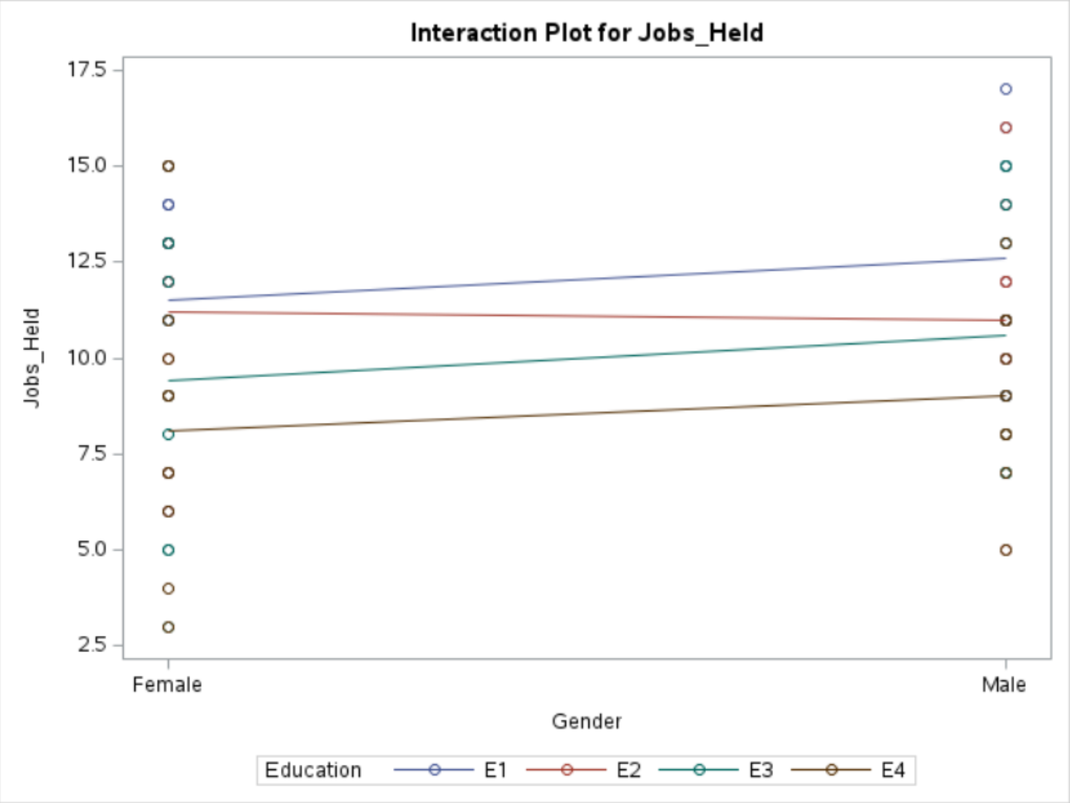
Dependent Variable: Jobs_Held

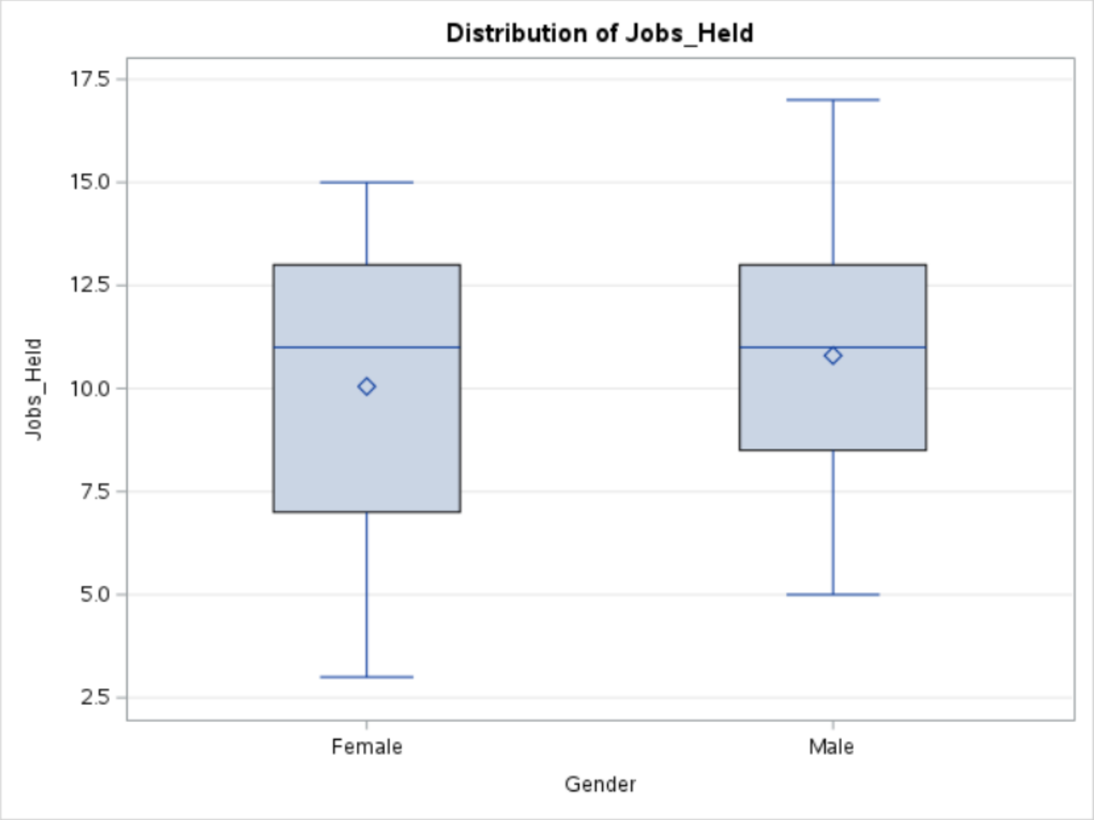
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	153.3500000	21.9071429	2.17	0.0467
Error	72	726.2000000	10.0861111		
Corrected Total	79	879.5500000			

R-Square	Coeff Var	Root MSE	Jobs_Held Mean
0.174351	30.46392	3.175864	10.42500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gender	1	11.2500000	11.2500000	1.12	0.2944
Education	3	135.8500000	45.2833333	4.49	0.0060
Gender*Education	3	6.2500000	2.0833333	0.21	0.8915

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gender	1	11.2500000	11.2500000	1.12	0.2944
Education	3	135.8500000	45.2833333	4.49	0.0060
Gender*Education	3	6.2500000	2.0833333	0.21	0.8915



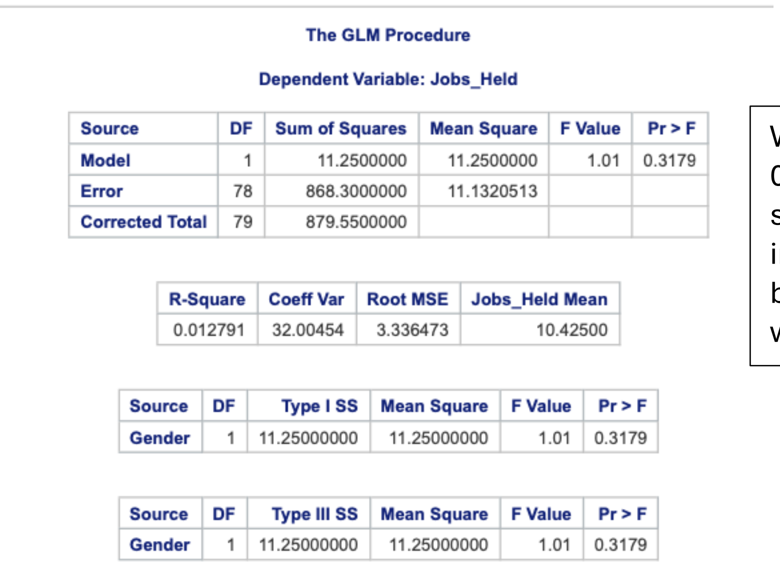


Interaction Effect Results

In the output for the interaction effect between gender and education:

- **F-Value**: 0.21
- **p-Value**: 0.8915

Since the p-value is much higher than 0.05, we fail to reject the null hypothesis for the interaction effect. This indicates that there is no significant interaction between gender and education level in terms of job holding. In other words, the effect of education on job holding does not vary by gender.



The GLM Procedure

Dependent Variable: Jobs_Held

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11.2500000	11.2500000	1.01	0.3179
Error	78	868.3000000	11.1320513		
Corrected Total	79	879.5500000			

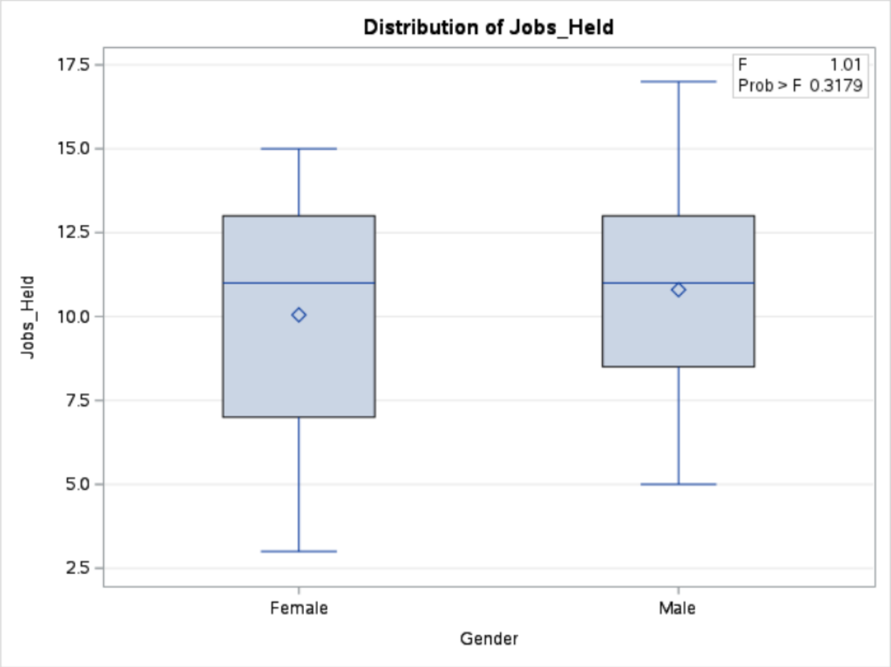
R-Square	Coeff Var	Root MSE	Jobs_Held Mean
0.012791	32.00454	3.336473	10.42500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gender	1	11.25000000	11.25000000	1.01	0.3179

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gender	1	11.25000000	11.25000000	1.01	0.3179

With a p-value above 0.05, there is no significant difference in job holding between men and women.

Program Summary - Program 2

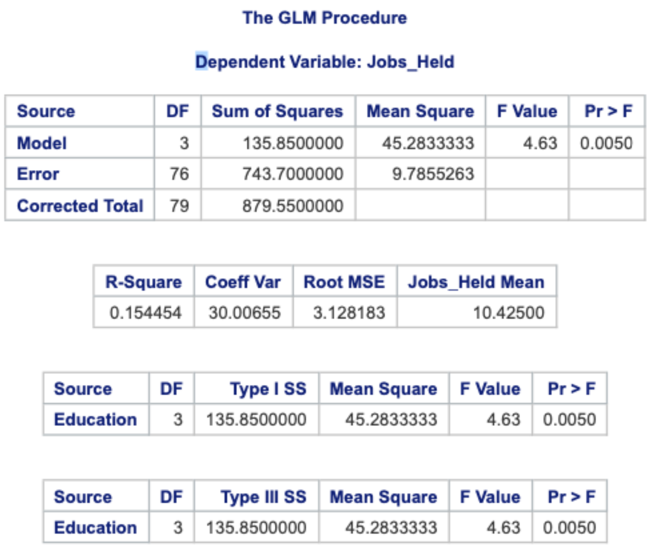


Main Effect of Gender

The output for the main effect of gender provides the following values:

- **F-Value:** 1.01
- **p-Value:** 0.3179

The p-value (0.3179) is above 0.05, indicating that there is no statistically significant difference in the number of jobs held between men and women. We fail to reject the null hypothesis, meaning gender alone does not have a significant effect on job holding.



The GLM Procedure

Dependent Variable: Jobs_Held

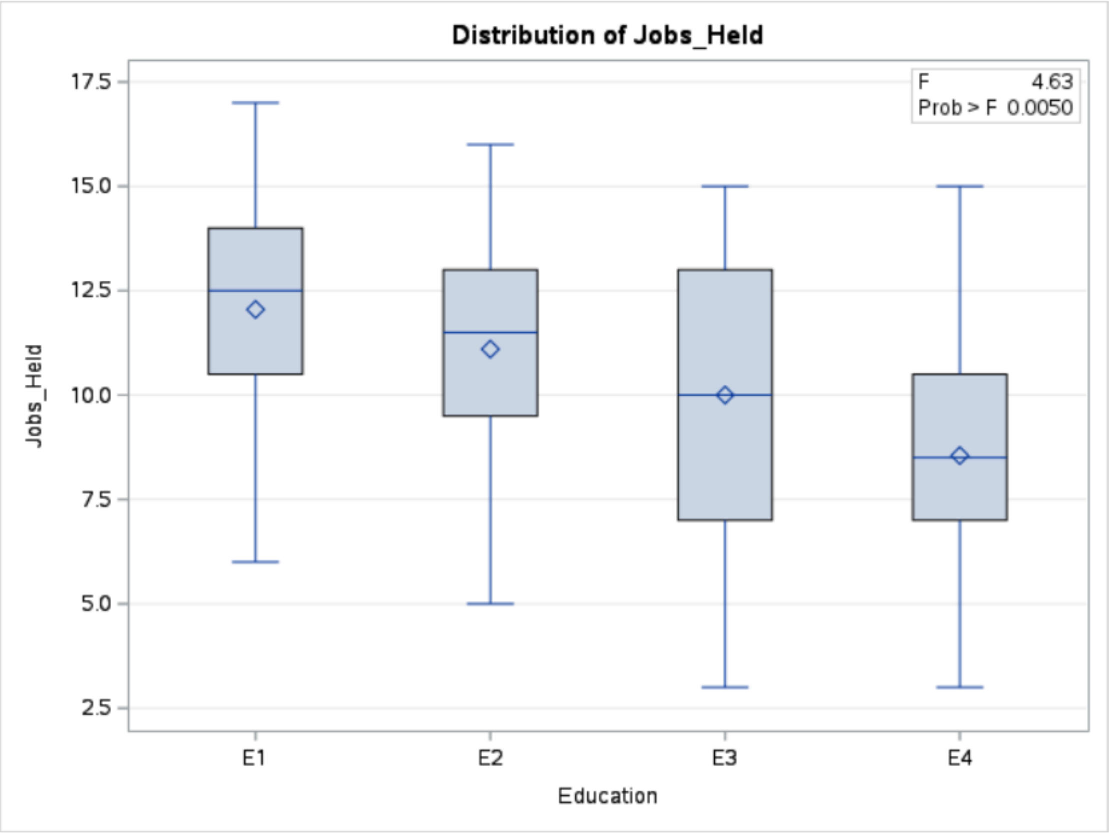
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	135.8500000	45.2833333	4.63	0.0050
Error	76	743.7000000	9.7855263		
Corrected Total	79	879.5500000			

R-Square	Coeff Var	Root MSE	Jobs_Held Mean
0.154454	30.00655	3.128183	10.42500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Education	3	135.8500000	45.2833333	4.63	0.0050

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Education	3	135.8500000	45.2833333	4.63	0.0050

The low p-value indicates significant differences in job holding across educational levels, warranting further exploration via Tukey's test.

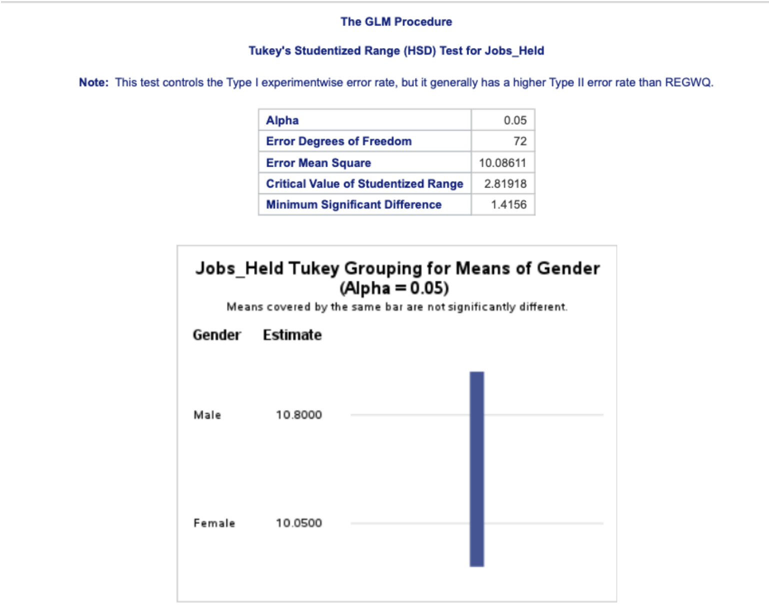


Main Effect of Education

The output for the main effect of education shows:

- **F-Value:** 4.63
- **p-Value:** 0.0050

With a p-value of 0.0050, which is less than 0.05, we reject the null hypothesis for education level. This result indicates that there are significant differences in job holding across different educational levels.



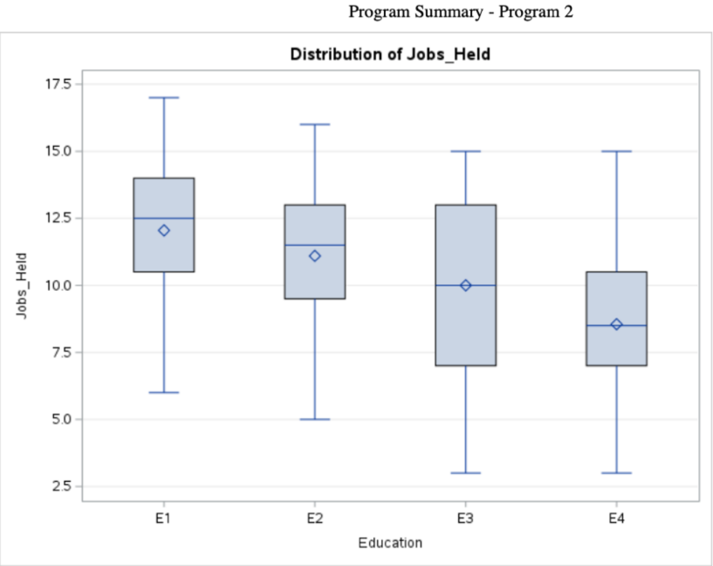
The GLM Procedure

Tukey's Studentized Range (HSD) Test for Jobs_Held

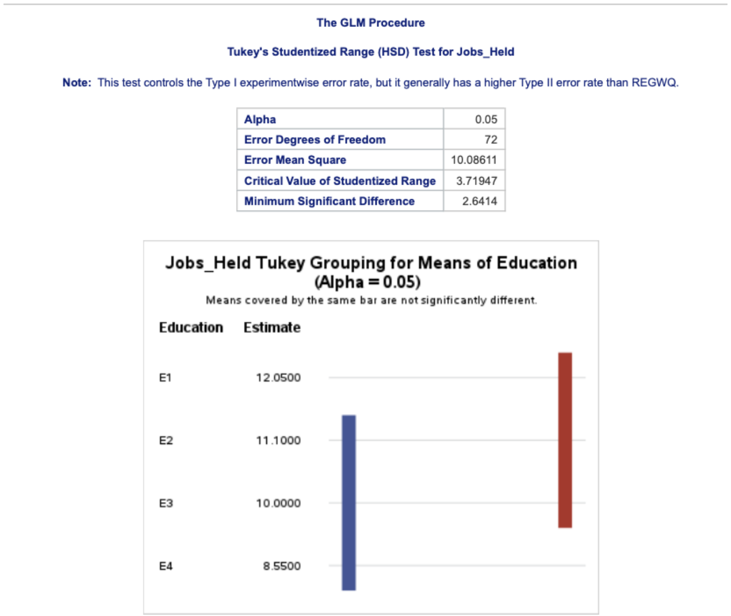
Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	72
Error Mean Square	10.08611
Critical Value of Studentized Range	2.81918
Minimum Significant Difference	1.4156

Jobs_Held Tukey Grouping for Means of Gender
(Alpha = 0.05)
Means covered by the same bar are not significantly different.
Gender
Estimate
Male
10.8000
Female
10.0500



Program Summary - Program 2



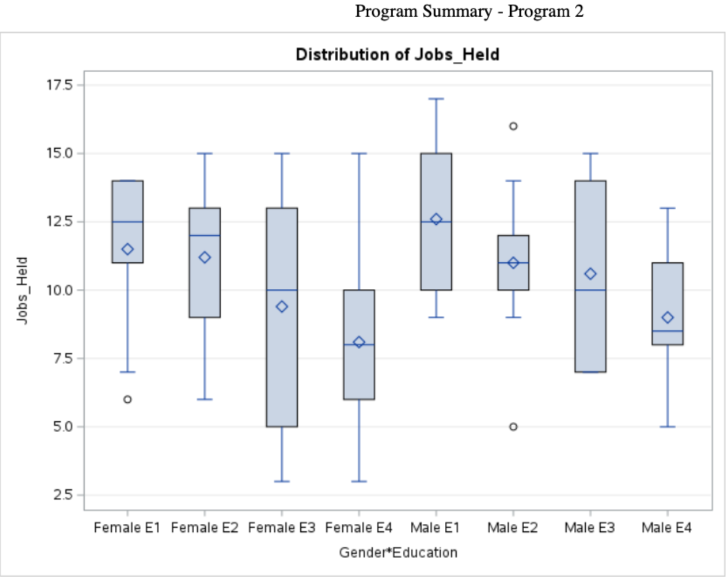
The GLM Procedure

Tukey's Studentized Range (HSD) Test for Jobs_Held

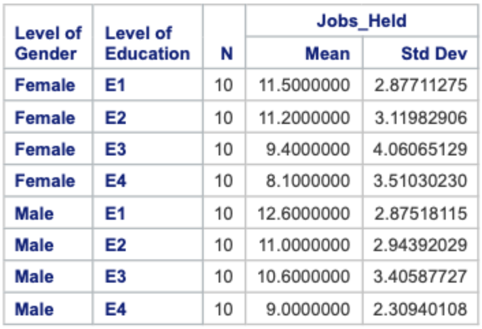
Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	72
Error Mean Square	10.08611
Critical Value of Studentized Range	3.71947
Minimum Significant Difference	2.6414

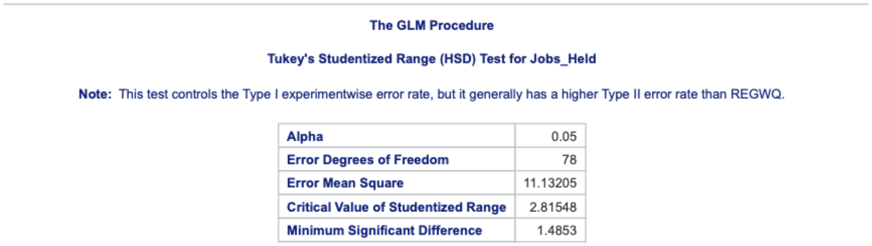
Jobs_Held Tukey Grouping for Means of Education
(Alpha = 0.05)
Means covered by the same bar are not significantly different.
Education
Estimate
E1
12.0500
E2
11.1000
E3
10.0000
E4
8.5500



Program Summary - Program 2



Level of Gender	Level of Education	N	Jobs_Held	
			Mean	Std Dev
Female	E1	10	11.5000000	2.87711275
Female	E2	10	11.2000000	3.11982906
Female	E3	10	9.4000000	4.06065129
Female	E4	10	8.1000000	3.51030230
Male	E1	10	12.6000000	2.87518115
Male	E2	10	11.0000000	2.94392029
Male	E3	10	10.6000000	3.40587727
Male	E4	10	9.0000000	2.30940108



The GLM Procedure

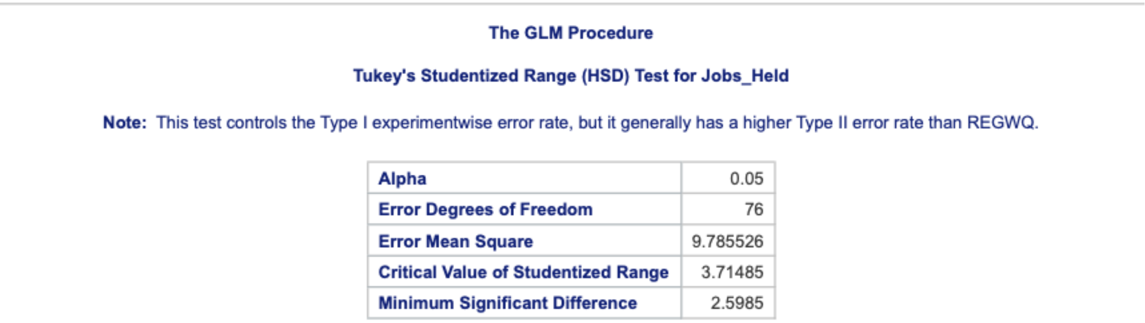
Tukey's Studentized Range (HSD) Test for Jobs_Held

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	78
Error Mean Square	11.13205
Critical Value of Studentized Range	2.81548
Minimum Significant Difference	1.4853

Descriptive Statistics by Gender and Education:

- Explanation: Provides mean job counts and standard deviations for each gender-education combination.
- Interpretation: Helps visualize patterns and differences in job holding across different educational levels and genders,



The GLM Procedure

Tukey's Studentized Range (HSD) Test for Jobs_Held

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	76
Error Mean Square	9.785526
Critical Value of Studentized Range	3.71485
Minimum Significant Difference	2.5985

Tukey's Post Hoc Test for Education
Tukey's test shows specific educational levels that differ significantly in terms of job holding. This result suggests that certain educational groups have statistically different job-holding patterns, which could imply that individuals with certain levels of education may tend to hold more or fewer jobs over time.

Conclusion

For Problem 2:

- There is no interaction between gender and education, meaning their effects on job holding are independent.
- Gender does not significantly affect the number of jobs held.
- Education level significantly affects job holding, with certain educational groups differing in their job-holding patterns.

These results highlight that educational attainment is a more critical factor than gender in determining the number of jobs held by individuals.