MAMAT JASSEH

STUDENT ID 141305227

STATISTICS FOR ANALYTICS

ASSIGNMENT 2

BAN100

QUESTION 1

SAS CODE:

```
PROC IMPORT DATAFILE='/home/u63568328/MY DATA/MyCodes
STATISTICS/File_Proportion_of_Total_Assets_Invested_in_Stocks (1) (1).xlsx'
OUT=Stocks DBMS=XLSX
REPLACE;
GETNAMES=YES;
RUN;
```

Using proc print and proc contents to obtain detailed information about the data provided.

SAS CODE:

```
data set.proc print data=Asset (obs=20);
run;
```

Obs	Young	Early_Middle_Age	Late_Middle_Age	Senior	E	F	G	H
1	24.8	28.9	81.5	66.8				
2	35.5	7.3	0.0	77.4				
3	68.7	61.8	61.3	32.9				
4	42.2	53.6	0.0	74.0				
5	49.5	0.0	45.4	0.0				
6	64.6	49.4	42.3	35.2				
7	58.3	71.4	75.3	21.4				
8	72.0	53.7	54.7	0.0				
9	25.6	46.9	0.0	61.4				
10	39.8	91.6	20.5	61.8				
11	39.3	46.0	76.4	35.6				
12	55.6	41.8	38.0	53.0				
13	0.0	53.2	39.8	38.5				
14	56.5	0.0	78.4	53.7				
15	37.3	43.7	0.0	69.1				
16	50.3	78.1	76.7	55.5				
17	38.0	54.7	72.7	31.6				
18	42.7	45.7	0.0	0.0				
19	48.4	63.1	33.0	57.3				
20	18.3	50.4	11.0	42.7				

SAS CODE:

```
proc contents data=Asset;
run;
```

The CONTENTS Procedure

Data Set Name	WORK.ASSET	Observations	136
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	10/24/2023 12:35:17	Observation Length	40
Last Modified	10/24/2023 12:35:17	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	3265
Obs in First Data Page	136
Number of Data Set Repairs	0
Filename	/saswork/SAS_work01BB0000DEEF_odaws02-usw2.oda.sas.com/SAS_workF3E20000DEEF_odaws02-usw2.oda.sas.com/asset.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	536936281
Access Permission	rw-r--r--
Owner Name	u63568328
File Size	256KB
File Size (bytes)	262144

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
5	E	Char	1	$1.	$1.	E
2	Early_Middle_Age	Num	8	COMMA15.1		Early_Middle_Age
6	F	Char	1	$1.	$1.	F
7	G	Char	1	$1.	$1.	G
8	H	Char	1	$1.	$1.	H
3	Late_Middle_Age	Num	8	COMMA15.1		Late_Middle_Age
4	Senior	Num	8	COMMA15.1		Senior
1	Young	Num	8	COMMA15.1		Young

The Asset data set has eight variables four characters four numeric and 136 observations, according to the proc content results.

In order to combine the variables Early Middle Age, Late Middle Age, Senior, and Young into a single character variable column, we must next establish a new column called age group.

SAS CODES:

```
data young;
set Asset; where young is not missing;
age_group = 'Young'; invest_in_stock=young;
DROP young early_middle_age late_middle_age senior E F G H;
run;

data early_middle_age;
set Asset; where early_middle_age is not missing;
age_group = 'Early_Middle_Age'; invest_in_stock=early_middle_age;
DROP young early_middle_age late_middle_age senior E F G H;
run;

data late_middle_age;
set Asset; where late_middle_age is not missing;
age_group = 'Late_Middle_Age'; invest_in_stock=late_middle_age;
DROP young early_middle_age late_middle_age senior E F G H;
```

```
run;

data senior;
set Asset; where senior is not missing;
age_group = 'Senior';invest_in_stock=senior;
DROP young early_middle_age late_middle_age senior E F G H;
run;

data Combine;
length age_group $25;
set young early_middle_age late_middle_age senior;
run;

proc print data=Combine; run;

There are currently 366 observations in total.
```

Obs	age_group	invest_in_stock
1	Young	24.82
2	Young	35.54
3	Young	68.70
4	Young	42.18
5	Young	49.52
6	Young	64.57
7	Young	58.25
8	Young	72.00
9	Young	25.62
10	Young	39.82
11	Young	39.27
12	Young	55.59
13	Young	0.00
14	Young	56.48
15	Young	37.26
16	Young	50.26
17	Young	38.00
18	Young	42.69
19	Young	48.43
20	Young	18.33
21	Young	50.09
22	Young	77.18
23	Young	42.73
343	Senior	62.23
344	Senior	69.41
345	Senior	48.55
346	Senior	53.41
347	Senior	34.76
348	Senior	72.33
349	Senior	72.58
350	Senior	47.05
351	Senior	62.31
352	Senior	50.78
353	Senior	51.12
354	Senior	67.25
355	Senior	42.48
356	Senior	55.34
357	Senior	60.79
358	Senior	71.33
359	Senior	56.79
360	Senior	52.66
361	Senior	51.16
362	Senior	88.49
363	Senior	81.24
364	Senior	72.44
365	Senior	47.56
366	Senior	65.74

After transforming the dataset, we are now prepared to test our theory.

Null Hypothesis - All age groups have equal average stock ownership.
Alternate Hypothesis - All age groups have unequal average stock ownership.

Using one way ANOVA test to determine if there is a difference in stock ownership between the age groups.

SAS CODE

```
/* Runing one way ANOVA test for combine data set*/
proc anova data=Combine;
class age_group; model invest_in_stock=age_group;
means age_group; run;
```

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
age_group	4	Early_Middle_Age Late_Middle_Age Senior Young

Number of Observations Read	366
Number of Observations Used	366

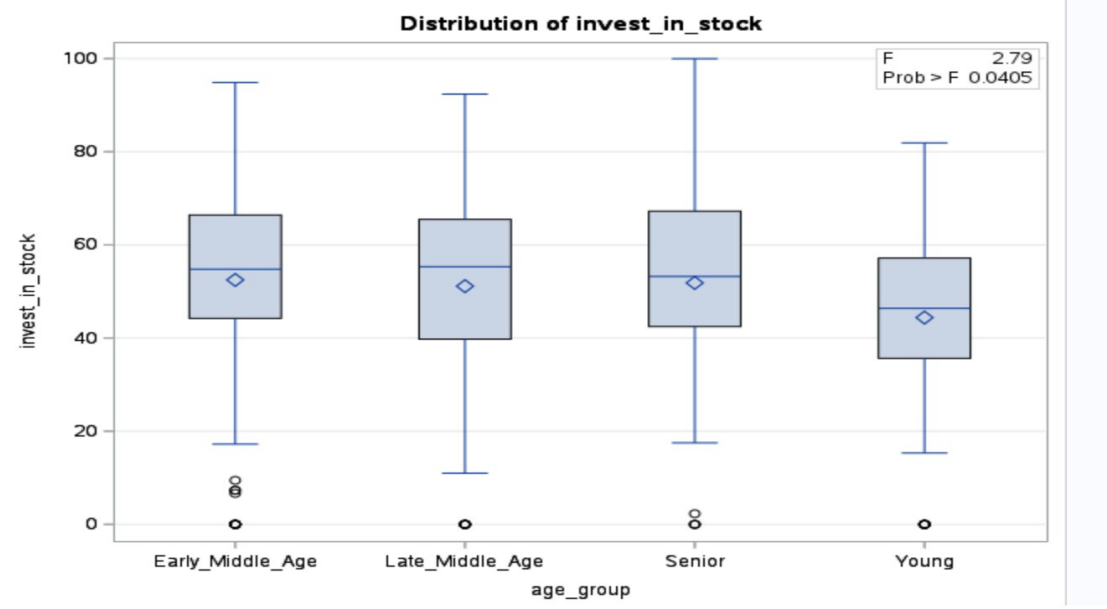
The ANOVA Procedure

Dependent Variable: invest_in_stock

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	invest_in_stock 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

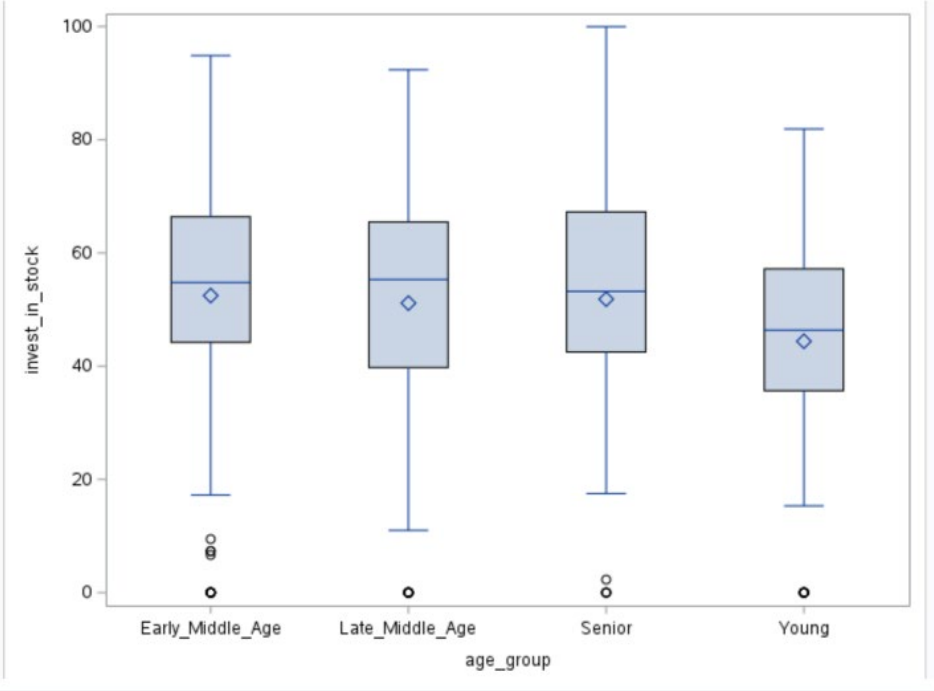


Level of age_group	N	invest_in_stock	
		Mean	Std Dev
Early_Middle_Age	131	52.4724427	21.6664980
Late_Middle_Age	93	51.1390323	21.7215074
Senior	58	51.8381034	21.0900334
Young	84	44.3983333	19.6607843

Running a Box plot for the combined data set

SAS CODE:

```
proc sgplot data=combine;
VBOX invest_in_stock /
category=age_group; run;
```



Runing proc univariate for the combine data set.

SAS CODE

```
proc univariate data=Combine;
PPPLOT invest_in_stock;run;
```

The UNIVARIATE Procedure
Variable: invest_in_stock

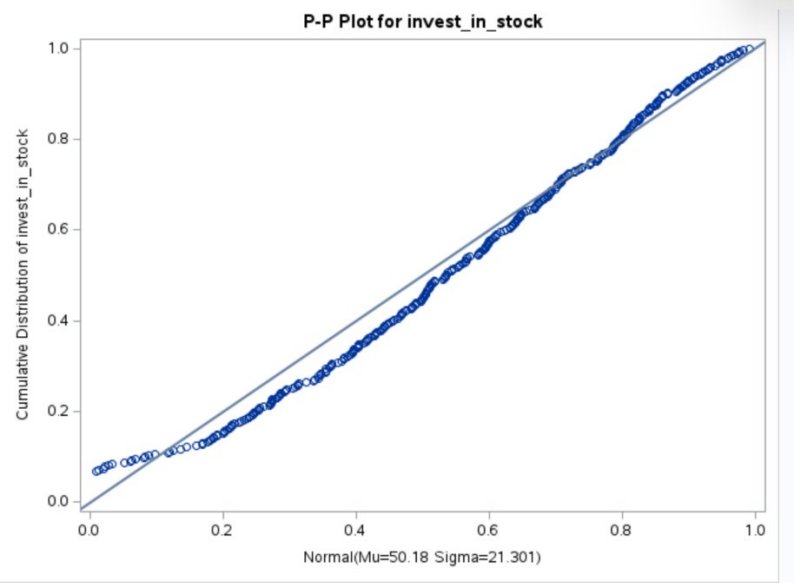
Moments			
N	366	Sum Weights	366
Mean	50.1800273	Sum Observations	18365.89
Std Deviation	21.3009965	Variance	453.732453
Skewness	-0.6627725	Kurtosis	0.35760488
Uncorrected SS	1087213.21	Corrected SS	165612.345
Coeff Variation	42.4491529	Std Error Mean	1.11342092

Basic Statistical Measures			
Location		Variability	
Mean	50.18003	Std Deviation	21.30100
Median	52.04000	Variance	453.73245
Mode	0.00000	Range	99.97000
		Interquartile Range	26.12000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	45.06834	Pr > |t|	<.0001
Sign	M	170.5	Pr >= |M|	<.0001
Signed Rank	S	29155.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	99.97
99%	91.57
95%	80.73
90%	74.01
75% Q3	65.39
50% Median	52.04
25% Q1	39.27
10%	20.62
5%	0.00
1%	0.00
0% Min	0.00

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	326	91.19	145
0	316	91.57	94
0	313	92.37	290
0	293	94.87	117
0	236	99.97	339



To investigate variations in stock ownership among age groups, the ANOVA results were examined. The alternative hypothesis states that average stock ownership varies by age group, contrary to the null hypothesis which states that it does not.

In the ANOVA table, the corresponding probability Pr > F is 0.0405, and the F-value is 2.79. A p-value of less than 0.05 in a hypothesis test is regarded as statistically significant, meaning that there is sufficient data to reject the null hypothesis. Given that the p-value of 0.0405 is less than 0.05, it may be concluded that there are notable variations in stock ownership between the age groups.

However, the R-Square value is 0.022591, indicating that the age groups account for just 2.26% of the variability in the dependent variable (stock investment). This implies that age groups contribute very little to the overall variance, even though they do have a statistically significant impact on stock investment.

This result is further supported by the boxplot visualization, which shows how stock investments are distributed among age groups. The median investment for the Early Middle Age group appears to be slightly larger than the other three age groups' spreads (interquartile ranges). Compared to the other age groups, the young age group has a lower median value and slightly less variability.

In conclusion, there is sufficient evidence to reject the null hypothesis based on the ANOVA results and the supplied visuals. This indicates that the various age groups' stock ownership differs significantly, supporting the alternate hypothesis.

QUESTION 2

```
PROC IMPORT DATAFILE='/home/u63568328/MY DATA/MyCodes
STATISTICS/File_Comparing_the_Lifetime_of_Jobs_by_Educational_Level (1).xlsx'
OUT=Lifetime_jobs
DBMS=XLSX
REPLACE;
GETNAMES=YES;run;
```

We can use the proc print and proc contents commands to retrieve detailed information about the provided data.

SAS CODE

```
proc print data=lifetime_jobs;
run;
```

Obs	Male_E1	Male_E2	Male_E3	Male_E4	Female_E1	Female_E2	Female_E3	Female_E4
1	10	12	15	8	7	7	5	7
2	9	11	8	9	13	12	13	9
3	12	9	7	5	14	6	12	3
4	16	14	7	11	6	15	3	7
5	14	12	7	13	11	10	13	9
6	17	16	9	8	14	13	11	6
7	13	10	14	7	13	9	15	10
8	9	10	15	11	11	15	5	15
9	11	5	11	10	14	12	9	4
10	15	11	13	8	12	13	8	11
11	.	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.	.
14	.	.	.	.	.	.	.	.
15	.	.	.	.	.	.	.	.
16	.	.	.	.	.	.	.	.
17	.	.	.	.	.	.	.	.
18	.	.	.	.	.	.	.	.
19	.	.	.	.	.	.	.	.
20	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.
22	.	.	.	.	.	.	.	.
23	.	.	.	.	.	.	.	.

SAS CODE

```
proc contents data=lifetime_jobs;
run;
```

The CONTENTS Procedure

Data Set Name	WORK.LIFETIME_JOBS	Observations	40
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	10/24/2023 23:42:39	Observation Length	64
Last Modified	10/24/2023 23:42:39	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	2043
Obs in First Data Page	40
Number of Data Set Repairs	0
Filename	/saswork/SAS_workE67A00011C7A_odaws01-usw2.oda.sas.com/SAS_work94E800011C7A_odaws01-usw2.oda.sas.com/lifetime_jobs.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	536872575
Access Permission	rw-r--r--
Owner Name	u63568328
File Size	256KB
File Size (bytes)	262144

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Label
5	Female_E1	Num	8	BEST.	Female_E1
6	Female_E2	Num	8	BEST.	Female_E2
7	Female_E3	Num	8	BEST.	Female_E3
8	Female_E4	Num	8	BEST.	Female_E4
1	Male_E1	Num	8	BEST.	Male_E1
2	Male_E2	Num	8	BEST.	Male_E2
3	Male_E3	Num	8	BEST.	Male_E3
4	Male_E4	Num	8	BEST.	Male_E4

Following the pro content, it was found that the lifetimr_jobs dataset comprises 8 numerical variables and 40 observations.

As a result, a new dataset will be created to explore the factors present in lifetime_jobs and incorporate 3 additional columns, namely gender, education, and the number of jobs.

SAS CODE

```
data Male1;
set lifetime_jobs;
where Male_E1 is not missing; Education = 'E1';
no_of_jobs = Male_E1; Gender = 'Male';
DROP Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3
Female_E4;
run;

data Male2;
set lifetime_jobs;
where Male_E2 is not missing; Education = 'E2';
no_of_jobs = Male_E2; Gender = 'Male';
DROP Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3
Female_E4;
run;

data Male3;
set lifetime_jobs;
where Male_E3 is not missing; Education = 'E3';
no_of_jobs = Male_E3; Gender = 'Male';
DROP Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3
Female_E4;
run;

data Male4;
set lifetime_jobs;
where Male_E4 is not missing; Education = 'E4';
```

```
no_of_jobs = Male_E4; Gender = 'Male';
DROP Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3
Female_E4;
run;

data Female1;
set lifetime_jobs;
where Female_E1 is not missing; Education = 'E1';
no_of_jobs = Female_E1; Gender = 'Female';
DROP Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3
Female_E4;
run;

data Female2;
set lifetime_jobs;
where Female_E2 is not missing; Education = 'E2';
no_of_jobs = Female_E2; Gender = 'Female';
DROP Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3
Female_E4;
run;

data Female3;
set lifetime_jobs;
where Female_E3 is not missing; Education = 'E3';
no_of_jobs = Female_E3; Gender = 'Female';
DROP Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3
Female_E4;
run;

data Female4;
set lifetime_jobs;
where Female_E4 is not missing;Education = 'E4';
no_of_jobs = Female_E4; Gender = 'Female';
DROP Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3
Female_E4;
run;

data education_Combine;
length Gender $10;
set Male1 Male2 Male3 Male4 Female1 Female2 Female3 Female4;
run;

proc print data=education_combine;
run;
```

Obs	Gender	Education	no_of_jobs
1	Male	E1	10
2	Male	E1	9
3	Male	E1	12
4	Male	E1	16
5	Male	E1	14
6	Male	E1	17
7	Male	E1	13
8	Male	E1	9
9	Male	E1	11
10	Male	E1	15
11	Male	E2	12
12	Male	E2	11
13	Male	E2	9
14	Male	E2	14
15	Male	E2	12
16	Male	E2	16
17	Male	E2	10
18	Male	E2	10
19	Male	E2	5
20	Male	E2	11
21	Male	E3	15
22	Male	E3	8
23	Male	E3	7
57	Female	E2	9
58	Female	E2	15
59	Female	E2	12
60	Female	E2	13
61	Female	E3	5
62	Female	E3	13
63	Female	E3	12
64	Female	E3	3
65	Female	E3	13
66	Female	E3	11
67	Female	E3	15
68	Female	E3	5
69	Female	E3	9
70	Female	E3	8
71	Female	E4	7
72	Female	E4	9
73	Female	E4	3
74	Female	E4	7
75	Female	E4	9
76	Female	E4	6
77	Female	E4	10
78	Female	E4	15
79	Female	E4	4
80	Female	E4	11

With the three newly formed columns, we can perform hypothesis testing on a combined dataset.

1. Null Hypothesis (H0): There is no interaction between gender and education in holding jobs
   Alternative Hypothesis (H1): There is an interaction between gender and education in holding jobs.

2. Null hypothesis (H0): There is no gender difference in means.
   Alternative Hypothesis (H1): There are differences in gender means.

3. Null Hypothesis (H0): There are no differences in the means of jobs held across different educational levels.
   Alternative Hypothesis (H1): The means vary according to educational levels.

Performing a two-way ANOVA analysis on the recently merged data set 'Education_combine' to ascertain:

A. whether there is interaction between gender and education in holding jobs.
B. Whether there are differences in holding jobs between men and women.
C. whether there are differences in holding jobs between the educational levels.

SAS CODE

```
proc anova data=education_combine;
class education Gender;
model no_of_jobs = education | gender;
means education gender; run;
```

The ANOVA Procedure

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

Number of Observations Read	80
Number of Observations Used	80

The ANOVA Procedure

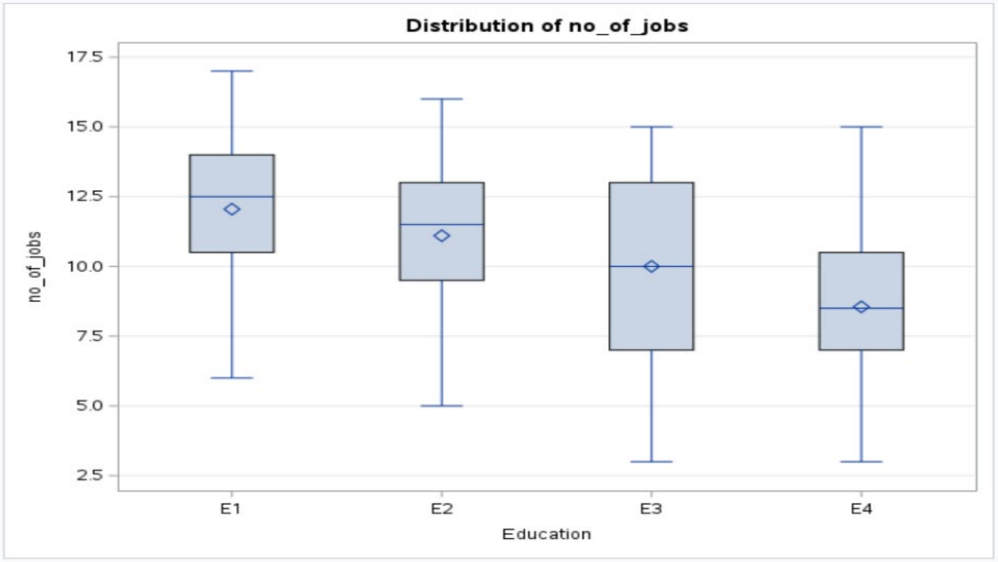
Dependent Variable: no_of_jobs

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	no_of_jobs Mean
0.174351	30.46392	3.175864	10.42500

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

The ANOVA Procedure



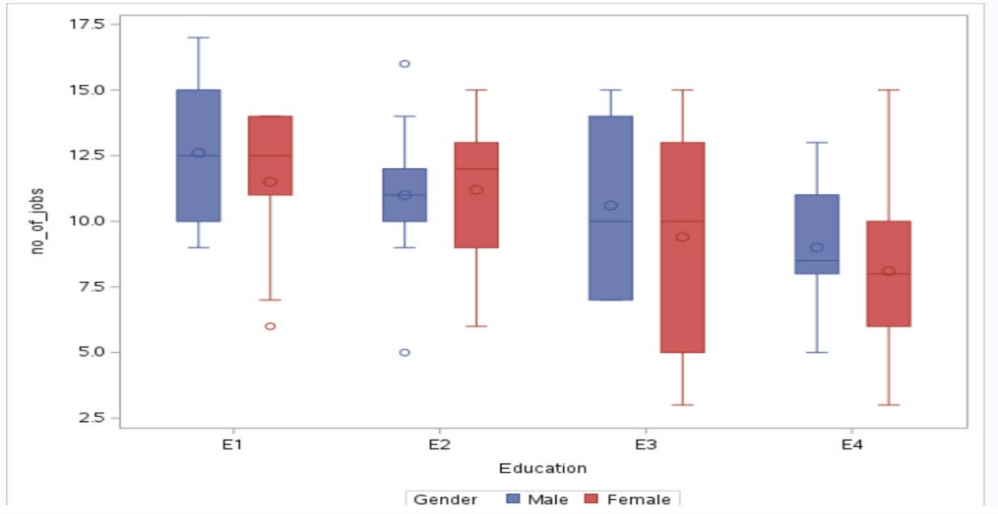
Level of Education	N	no_of_jobs	
		Mean	Std Dev
E1	20	12.0500000	2.85574214
E2	20	11.1000000	2.95403382
E3	20	10.0000000	3.69921756
E4	20	8.5500000	2.92853475

Level of Gender	N	no_of_jobs	
		Mean	Std Dev
Female	40	10.0500000	3.57304725
Male	40	10.8000000	3.08179102

Running a Box plot on the newly created data set ‘education_combine

SAS CODE

```
proc sgplot data=education_combine;
VBOX no_of_jobs / category=Education group=Gender; run;
```



Running a proc univariate on the newly created data set ‘education_combine

SAS CODE

```
proc univariate data=education_combine;
PPPLOT no_of_jobs; run;
```

The UNIVARIATE Procedure
Variable: no_of_jobs

Moments			
N	80	Sum Weights	80
Mean	10.425	Sum Observations	834
Std Deviation	3.33669662	Variance	11.1335443
Skewness	-0.2249118	Kurtosis	-0.6938761
Uncorrected SS	9574	Corrected SS	879.55
Coeff Variation	32.0066822	Std Error Mean	0.37305402

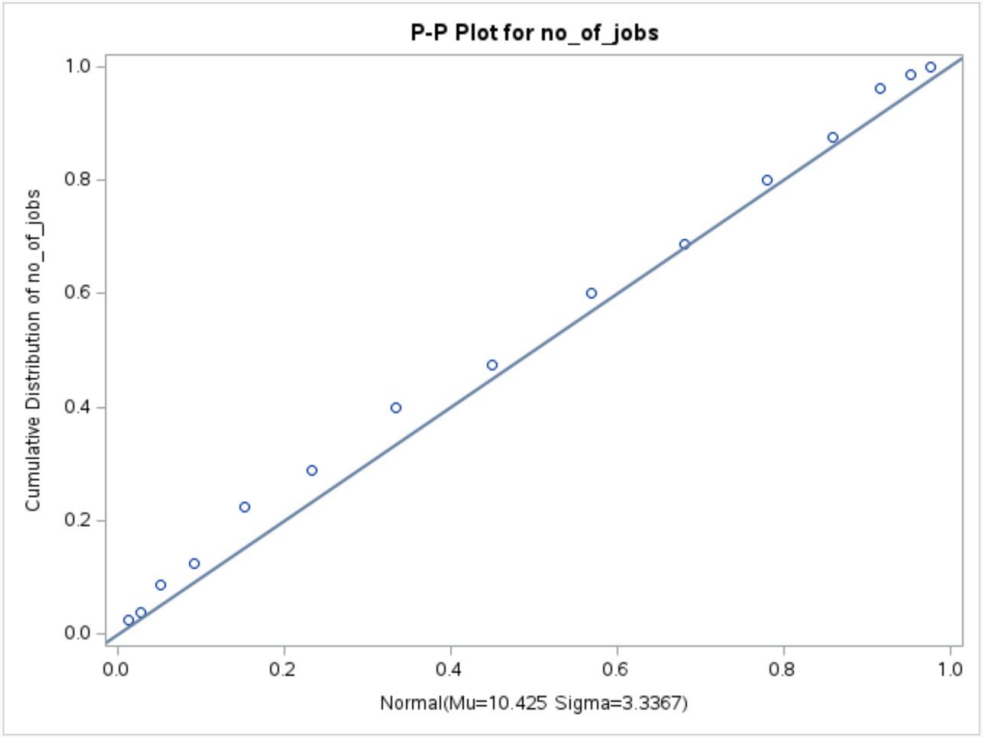
Basic Statistical Measures			
Location		Variability	
Mean	10.42500	Std Deviation	3.33670
Median	11.00000	Variance	11.13354
Mode	11.00000	Range	14.00000
		Interquartile Range	5.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	27.94501	Pr > |t|	<.0001
Sign	M	40	Pr >= |M|	<.0001
Signed Rank	S	1620	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	17
99%	17
95%	15
90%	15
75% Q3	13
50% Median	11
25% Q1	8
10%	6
5%	5
1%	3
0% Min	3

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
3	73	15	67
3	64	15	78
4	79	16	4
5	68	16	16
5	61	17	6

The UNIVARIATE Procedure



A. Is there any interaction between gender and education in holding jobs?

Based on the results, the interaction term (Education*Gender) has a p-value of 0.8915, which is notably higher than the widely accepted significance level of 0.05. As a result, it appears that we cannot reject the null hypothesis.

Furthermore, the main effects for gender and education indicate that there is a discrepancy in the number of jobs across different educational levels (p-value = 0.0060, which is significant). However, the number of occupations doesn't seem to be considerably affected by gender alone (p-value = 0.2944).

In summary, the data implies that there is no significant interactional impact between gender and education, despite the variances in the number of jobs across all educational levels. Since there is no additional interaction effect, the combined impact of gender and education on the number of jobs is merely the sum of their individual effects.

B. Are there any re differences in holding jobs between men and women?

Our goal was to investigate whether there are notable differences in job holdings between men and women, based on the ANOVA results we have obtained. We formulated two hypotheses: the null hypothesis, which suggests that there is no significant difference in the average number of jobs held between genders, and the alternative hypothesis, which proposes the existence of such a difference. The p-value for the "Gender" factor is 0.2944, which is greater than the conventional significance threshold of 0.05.

Therefore, we don’t have sufficient evidence to reject the null hypothesis. Upon examining the descriptive figures, we found that women hold an average of 10.05 jobs, while men have a slightly higher average of 10.8 jobs. However, this minor difference is not statistically significant. Based on the evidence we have gathered; we can conclude that there is no significant disparity in the number of jobs held between men and women.

C. Are there any differences in holding jobs between the educational levels?

After conducting an ANOVA analysis, we developed two hypotheses to determine whether there are differences in the amount of work held at different levels of education. Our null hypothesis states that there is no significant difference in the average number of jobs held based on education level, while the alternative hypothesis states that there is a difference.

When analyzing the ANOVA data, we obtained a p-value of 0.0060 for the variable “Education”, which is significantly lower than the standard significance level of 0.05. Accordingly, we reject the null hypothesis and conclude that there is substantial evidence supporting the claim that the number of jobs held varies significantly at different levels of education.