SCHOOL OF SCIENCE AND TECHNOLOGY

ASSIGNMENT / PROJECT SUBMISSION FORM

PROGRAMME:
SEMESTER: Jan / Mar / Aug 2020

SUBJECT: IST2024 Applied Statistics

DEADLINE: 3rd July 2020

INSTRUCTIONS TO CANDIDATES

- This is an ~~individual~~ / group project.

IMPORTANT
The University requires students to adhere to submission deadlines for any form of assessment. Penalties are applied in relation to unauthorized late submission of work.

- Coursework submitted after the deadline but within 1 week will be accepted for a maximum mark of 40%.
- Work handed in following the extension of 1 week after the original deadline will be regarded as a non-submission and marked zero.

Lecturer's Remark (Use additional sheet if required)
List down the name of the group members and the student IDs here.

I Jarrod Tham Kuok Yew (Student's Name) 16034753 (Student ID) received the assignment and read the comments.

Jarrod 23/7/2020(Signature/Date)

Academic Honesty Acknowledgement

"I .....(Student's Name) verify that this paper contains entirely my own work. I have not consulted with any outside person or materials other than what was specified (an interviewee, for example) in the assignment or the syllabus requirements. Further, I have not copied or inadvertently copied ideas, sentences, or paragraphs from another student. I realize the penalties *(refer to page 16, 5.5, Appendix 2, page 44 of the student handbook diploma and undergraduate programme)* for any kind of copying or collaboration on any assignment."

…..... (Student's signature / Date)

Table of content

Introduction

The following data is based on a bank. The variables included all relate to the customers from the bank. The bank name and details are unknown.

Table 1

Assignment-Individual-Data Variable Description

Variable	Description	Data Type
ID	Client number	Numeric
GENDER	Gender (M = Male, F = Female)	Character
OWNPROPERTY	Owns a car (Y = Yes, N = No)	Character
CHILDRENCOUNT	Number of children	Numeric
INCOMETOTAL	Annual Income	Numeric
INCOMETYPE	Income category	Character
EDUCATIONLEVEL	Education level	Character
MARITALSTATUS	Marital status	Character
HOUSINGTYPE	Way of living	Character
MOBILE	Owns a mobile phone (1 = Yes, 0 = No)	Numeric
EMAIL	Has an e-mail account (1 = Yes, 0 = No)	Numeric
OCCUPATION	Occupation	Character
FAMSIZE	Family size	Numeric
CREDITSTATUS	Credit loan status	Character

	0: 1-29 days past due 1: 30-59 days past due 2: 60-89 days overdue 3: 90-119 days overdue 4: 120-149 days overdue 5: Bad debts / Write-offs C: paid off for that month X: no loan for the month	

Although the ID variable is a numeric value, it will be excluded from the analysis as it serves no purpose for the modeling stage as the bank does not want to create a model to predict or test the assumptions of a singular customer (ID). Furthermore, the OCCUPATION variable is also discarded because upon data inspection, the variable contains about 30% missing values which could be dangerous when fitting the variable into any model. Although imputation would solve the missing values, in this case, it may cause data outliers which is not necessarily helpful for the bank data.

Descriptive Analysis

Variable	N	N Miss	Minimum	Maximum	Mode	Median	Mean	Std Dev
CHILDRENCOUNT	8000	0	0	14.0000000	0	0	0.4215000	0.7539289
INCOMETOTAL	8000	0	27000.00	1575000.00	135000.00	157500.00	185712.85	100015.41
FAMSIZE	8000	0	1.0000000	15.0000000	2.0000000	2.0000000	2.1883750	0.9200520

Figure 1. Descriptive Statistics for assignment data.

Figure 1 shows descriptive statistics for three numeric variables the assignment data. The descriptive statistics include the number of observations (N), number of missing values (N Miss), minimum, maximum, mode, median, mean, and standard deviation values of the variables. Moreover, the binary variables; EMAIL, MOBILE, GENDER, OWNCAR, and OWNPROPERTY are ignored in this table because their central tendency doesn't mean

anything because all their values just have ‘0’, ‘1’, ‘N’, and ‘Y’ in each of the variables. Likewise, the categorical variables; EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, INCOMETYPE, and CREDITSTATUS are also excluded from *Figure 1* because the values are all character or object format with no numeric data.

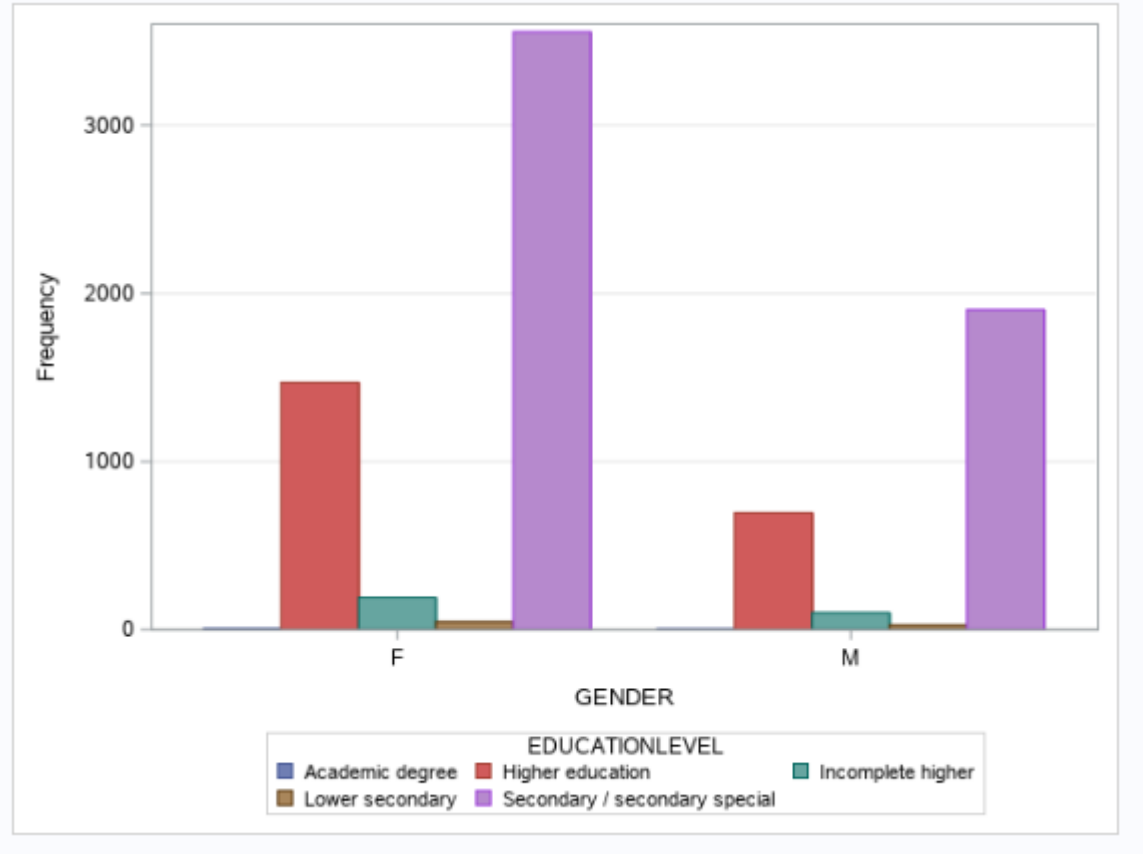


Figure 2. Frequency of customer’s education level by gender.

The vertical bar chart in *Figure 2* shows the customer’s frequency count of each education level by gender. It is clear that female customers surpass the males in each category of education level from holding an academic degree to graduating from secondary or special secondary schools.

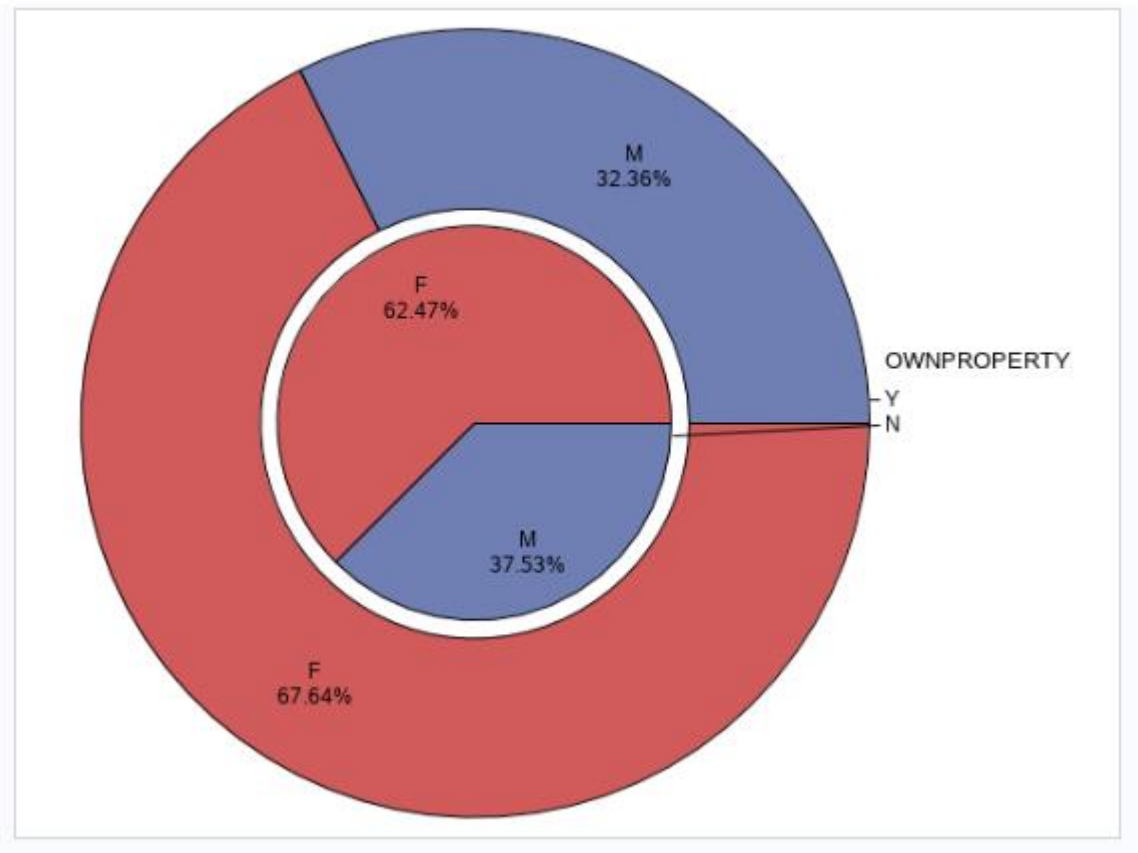


Figure 3. Percentage of customers who own property by gender.

The pie chart in *Figure 3* shows the percentage of customers who own a property by gender. The pie chart indicates over 67% of female customers and 32% of male customers who own a property. Opposingly, over 62% of female customers and 37% of male customers who do not own a property. This can be interpreted as although the majority of female customers own a property, the majority of customers who do not own a single property are also female.

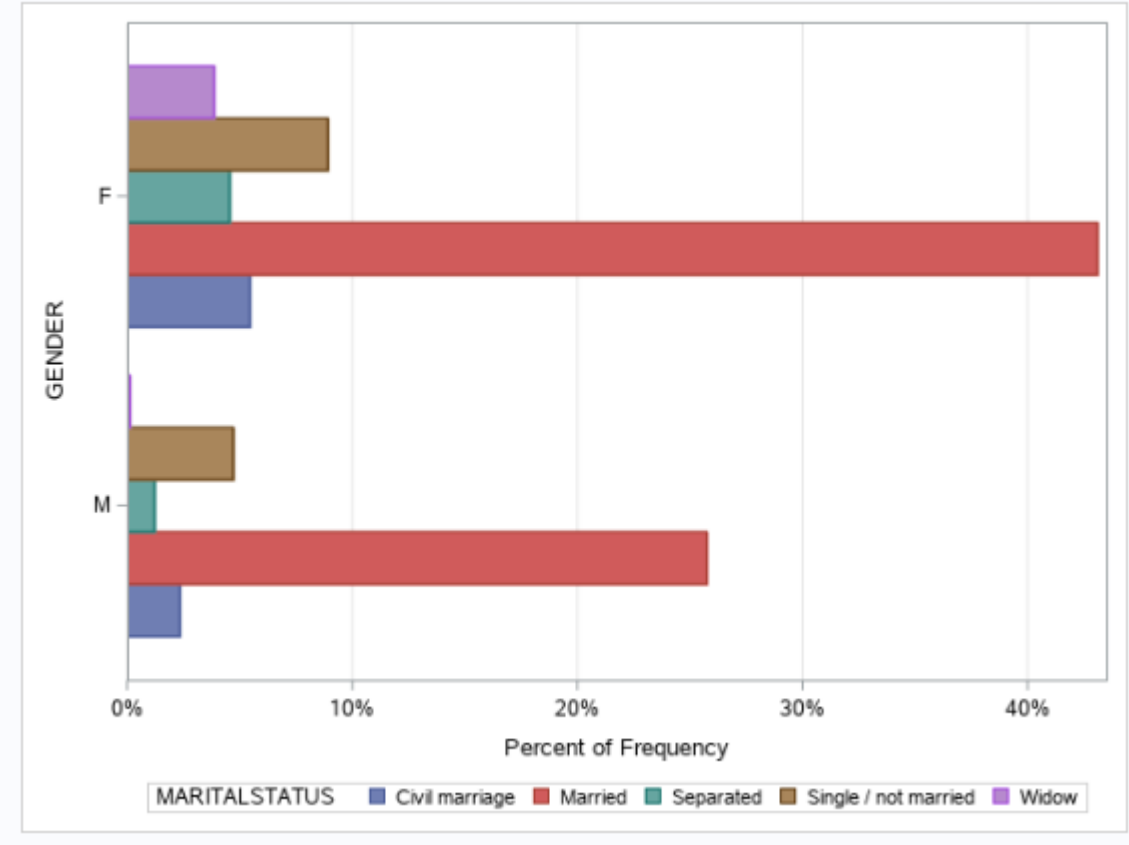


Figure 4. Percentage customer's marital status by gender.

The horizontal bar chart in *Figure 4* indicates the percentage of customer's marital status by gender. The chart shows that over 40% of female customers and 25% of male customers are married. On the other hand, about 5% of females are widows while less than 1% of males are widows.

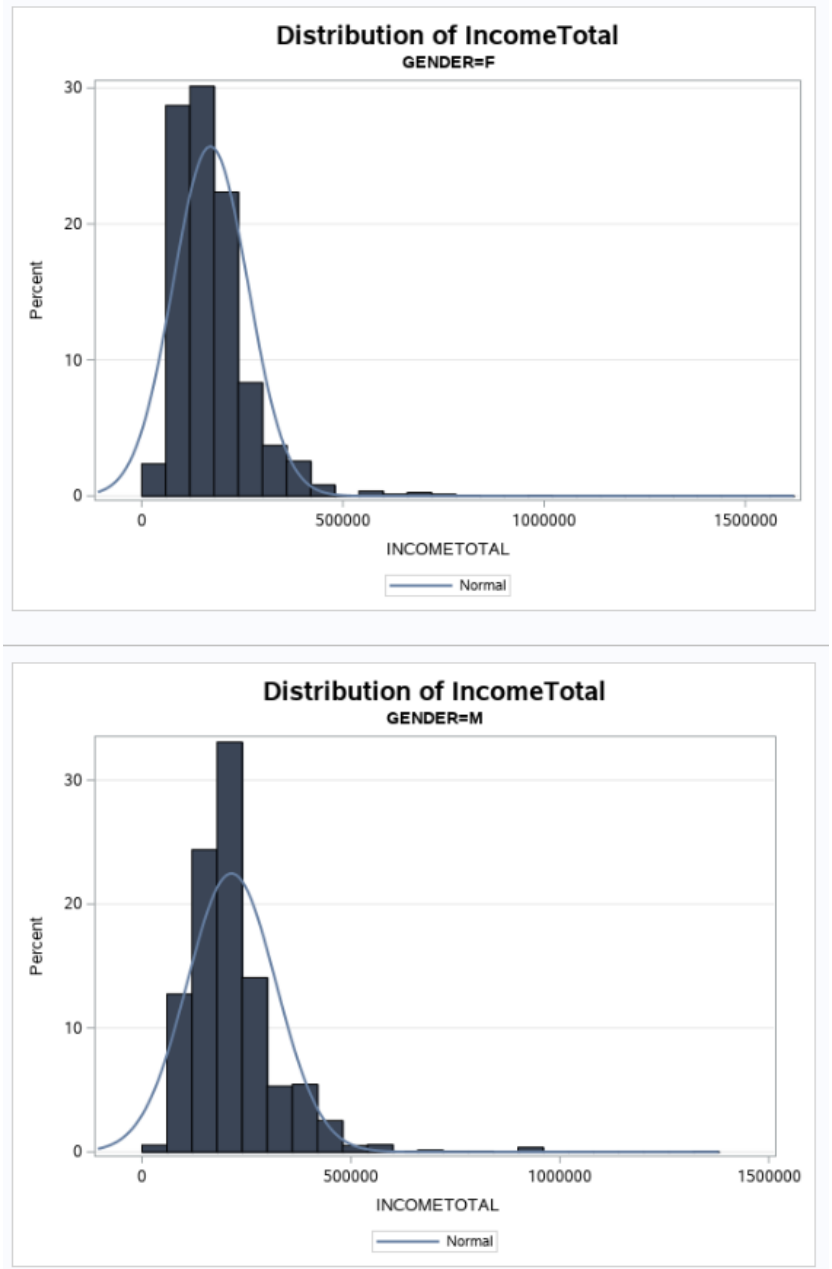


Figure 5. Distribution of annual income by gender.

Figure 5 displays two histograms showing the customer's distribution of annual income by gender. It shows that over 30% of males and 20% females have an annual income of 250000, which indicates that most of the males earned more at its peak compared to females. However, there are 3% of females that earn more during the first quartile of annual income compared to only less than 1% of males.

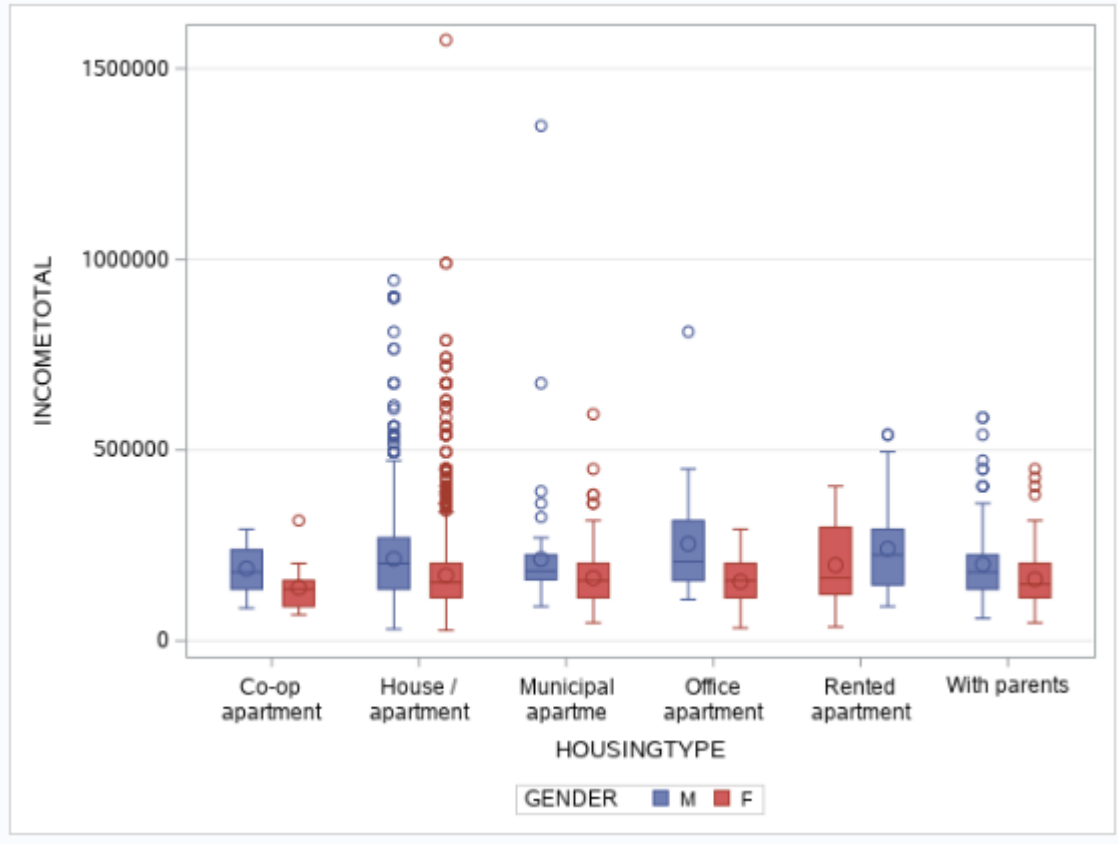


Figure 6. Frequency of customer's annual income by ways of living ordered by gender.

The box plot in *Figure 6* shows the frequency of customer's annual income by ways of living ordered separated by gender. Firstly, the plot shows that the median of annual income for all ways of living seem to be close to each other. Also, the plot shows that customers living in a house or apartment consist of the most outliers that are few times more than the upper quartile or maximum value of annual income. Furthermore, customers living in a municipal apartment and with their parents are also shown to have few outliers few times more than the upper quartile or maximum value of annual income. Lastly, both male and female customers living in a rented apartment show to have the highest maximum annual income across all ways of living.

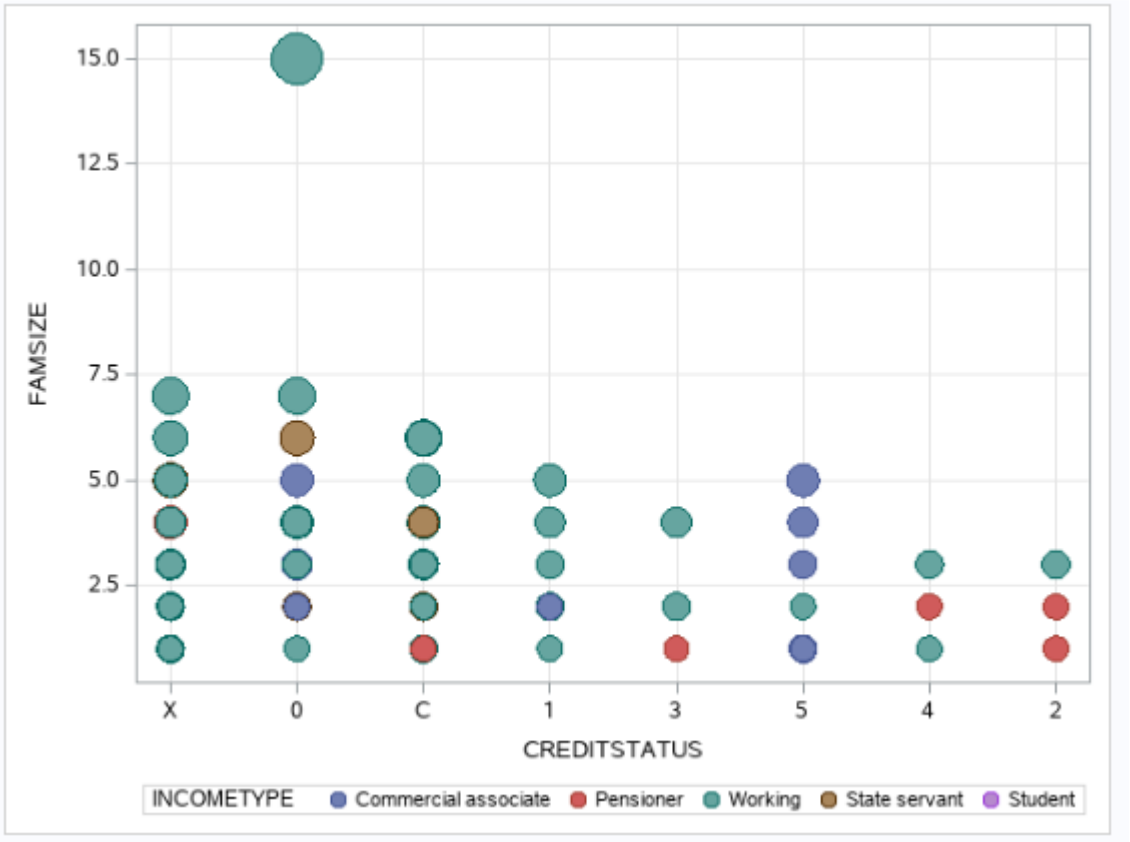


Figure 7. Frequency of customer's family size by credit loan status controlling for income category and number of children.

The bubble plot in *Figure 7* shows the frequency of customer's family size by credit loan status, where the bubble size represents the number of children and bubble color for the income category. Firstly, most customers with no loan for the month (X) for their credit loan status seem to be working class people with a maximum family size of 7. Next, most of the customers with bad debts or write-offs (5) earn their income as commercial associates with a maximum family size of 5. Lastly, the plot shows that as the number of family sizes increases, the number of children (bubble size) also increases.

Analysis

Analysis of Variance (ANOVA)

First Analysis of Variance (ANOVA)

One-way ANOVA

The first ANOVA model is created to study six ways of living (HOUSINGTYPE); Co-op apartment, House / apartment, Municipal apartment, Office apartment, Rented apartment, and With parents. The bank also has information showing the credit loan status (CREDITSTATUS) to each way of living based on the annual income. The One-way ANOVA is performed to see whether the average annual income is significantly different for various ways of living.

First One-Way ANOVA

Class Level Information

Class	Levels	Values
HOUSINGTYPE	6	Co-op apartment House / apartment Municipal apartme Office apartment Rented apartment With parents

Number of Observations Read	8000
Number of Observations Used	8000

First One-Way ANOVA

Dependent Variable: INCOMETOTAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	212405146755	42481029351	4.26	0.0007
Error	7994	7.9802243E13	9982767393		
Corrected Total	7999	8.0014648E13			

R-Square	Coeff Var	Root MSE	INCOMETOTAL Mean
0.002655	53.80016	99913.80	185712.8

Source	DF	Type I SS	Mean Square	F Value	Pr > F
HOUSINGTYPE	5	212405146755	42481029351	4.26	0.0007

Source	DF	Type III SS	Mean Square	F Value	Pr > F
HOUSINGTYPE	5	212405146755	42481029351	4.26	0.0007

Figure 8. First ANOVA model class-level-information table and analysis of variance table.

Based on *Figure 8*, the output specifies the six levels and its values of the class variable (HOUSINGTYPE), and the number of observations read versus the number of observations

used are equal. These values are the same because there are no missing values in any variable in the model. Next, the output contains all of the information needed to test the equality of the group means which is divided into three parts. For the first part, the value of the test statistics, i.e. the *F*-statistic and corresponding *p*-value are reported in the analysis of variance table. Since there are six types of ways of living (HOUSINGTYPE) used, this analysis is to test the hypothesis on whether the means of annual income (INCOMETOTAL) are equal for all types of ways of living. The following shows the details of the null hypothesis (H_0) and the alternative hypothesis (H_1):

H_0: $\mu 1 = \mu 2 = \mu 3 = \mu 4 = \mu 5 = \mu 6$: All the means of annual income are the same

H_1: At least one pair of the annual income (μ_i) means is different

From the analysis of variance table, *F*-value (4.26) with corresponding *p*-value (0.0007) is less than 0.05 level of significance. Thus, the null hypothesis (H_0) is rejected. It is concluded that at least one pair of annual income means is different and indicates that there are significant differences between the means of annual income across ways of living.

Secondly, the coefficient of determination, R^2, denoted in the table as *R*-Square, is a measure of the proportion of variability explained by the independent variables in the analysis. It is interpreted that the way of living (HOUSINGTYPE) explains about 0.3% of the variability of the annual income (INCOMETOTAL) in the model. Likewise, the coefficient of variation (Coeff Var) of 53.8 expresses the root MSE (the estimate of the standard deviation for all treatments) of 99913.8 as a percent of the mean. It is a unitless measure that is useful in comparing the variability of two sets of data with different units of measure. Hence, the INCOMETOTAL Mean is the mean of all of the data values in the variable INCOMETOTAL without regard to HOUSINGTYPE.

Since this is a one-way ANOVA model, the third part shows the information about the class variable (HOUSINGTYPE) in the model is an exact duplicate of the model line of the analysis of variance table.

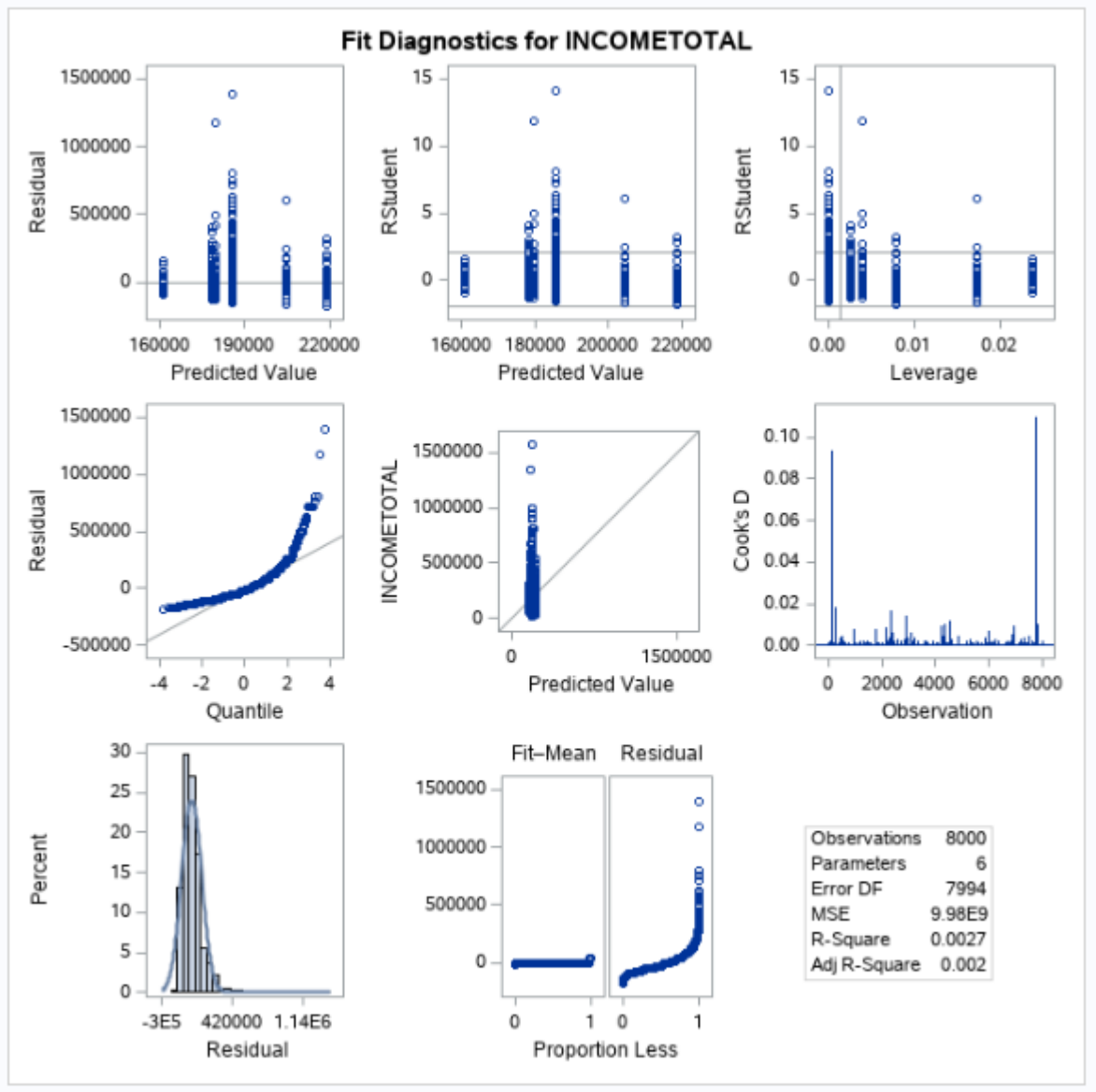


Figure 9. First ANOVA model fit diagnostics plot.

Moving on, the ANOVA model carries three assumptions to be looked into from the three plots in the left pane outputted in *Figure 9*. The plot at the upper-left is a Residuals by Predicted plot, which looks at the random scatter within each group of HOUSINGTYPE. The plot indicates there is a violation in the model assumption due to the data not being scattered

within each group of HOUSINGTYPE. Next, the plot at the center-left is a Quantile-Quantile (Q-Q) plot, which checks for normality assumption in the model. Based on the plot, it is concluded that there appears to have severe departure from normality because the observations tend to lie to the diagonal reference line up till the second quartile where it starts to depart from the line. Lastly, the bottom-left plot shows a positively skewed histogram (tail extending to the right) with a unique peak.

First One-Way ANOVA

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

HOUSINGTYPE	INCOMETOTAL LSMEAN	LSMEAN Number
Co-op apartment	161250.000	1
House / apartment	185725.944	2
Municipal apartme	179769.556	3
Office apartment	204323.276	4
Rented apartment	218636.220	5
With parents	178435.515	6

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

Dependent Variable: INCOMETOTAL

i/j	1	2	3	4	5	6
1		0.6099	0.8771	0.2729	0.0159	0.8974
2	0.6099		0.9408	0.7199	0.0032	0.7231
3	0.8771	0.9408		0.5419	0.0049	1.0000
4	0.2729	0.7199	0.5419		0.9456	0.4388
5	0.0159	0.0032	0.0049	0.9456		0.0012
6	0.8974	0.7231	1.0000	0.4388	0.0012	

Figure 10. First ANOVA model post-hoc analysis: Tukey's Multiple Comparisons.

After that, *Figure 10* shows the post-hoc analysis, Tukey's multiple comparisons, for the first ANOVA model. Based on the output, the way of living with the highest mean annual income is living in a rented apartment and the one with the lowest is living in a co-op apartment. The data is visualized in *Figure 11*.

Furthermore, the *p*-values for comparing each of the group means are displayed as a symmetric matrix which shows all the pairwise comparison *p*-values twice. The matrix is interpreted as the groups having a *p*-value of less than 0.05 are significant to one another. In case this, it is observed that four groups; 5*1, 5*2, 5*3, and 5*6, are significant to one another. Similarly, the data is visualized in *Figure 12*.

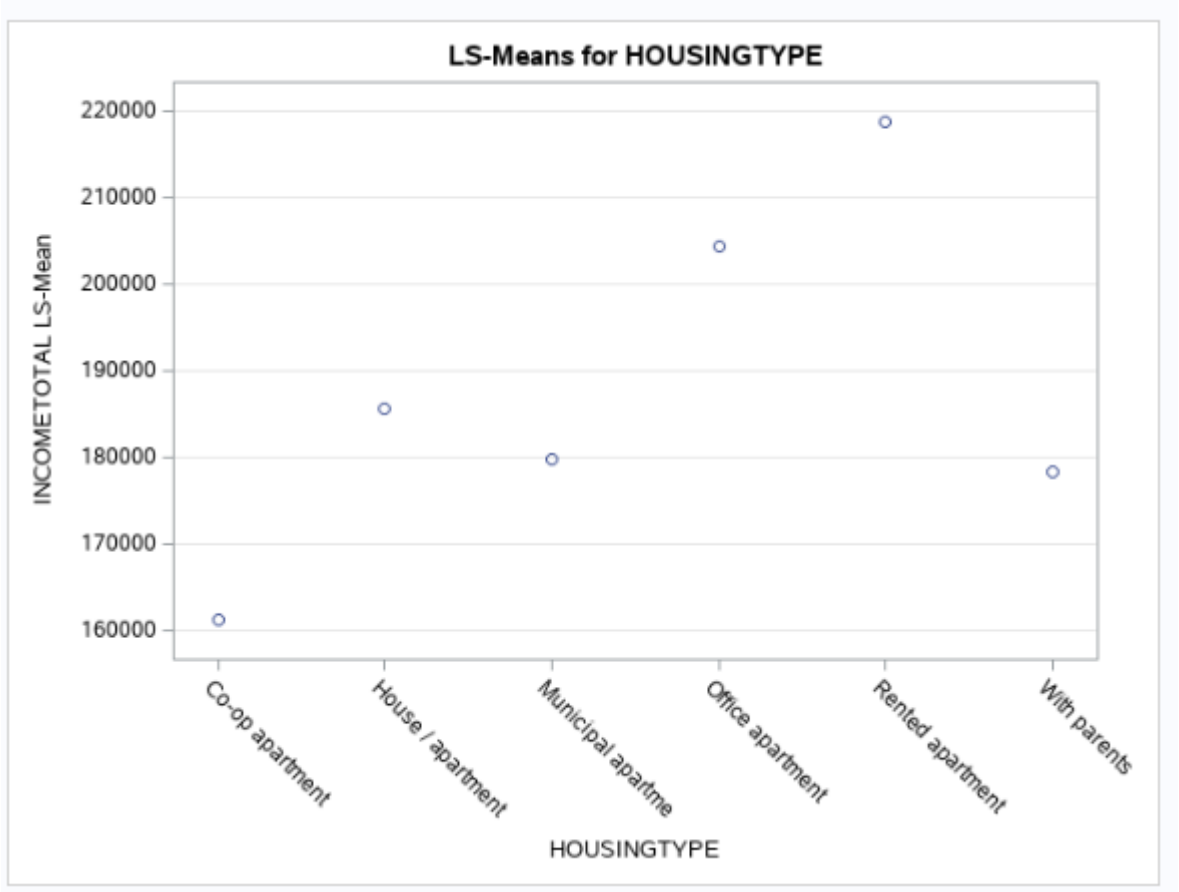


Figure 11. First ANOVA model LS-Means frequency dot plot.

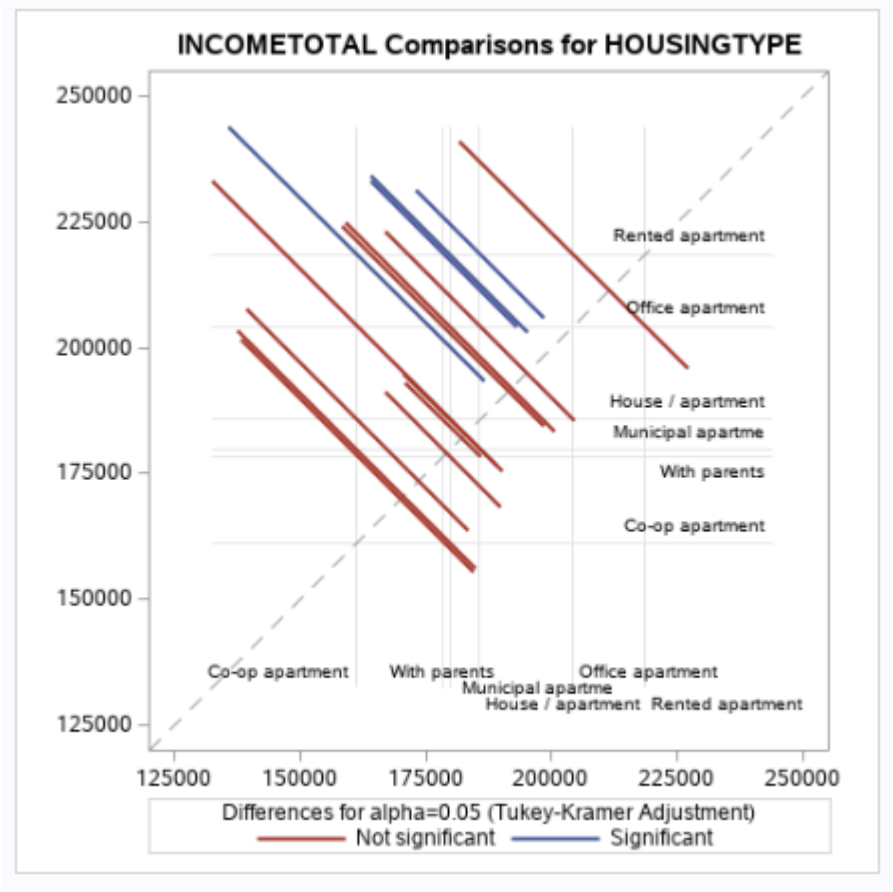


Figure 12. First ANOVA model Tukey's diffogram.

Figure 12 shows a diffogram for the first ANOVA model. A diffogram can be used to quickly tell if two group means are statistically significant. The point estimates for the differences between pairs of group means can be found at the intersections of vertical and horizontal lines. Also, the colored diagonal lines show the confidence intervals for the differences. In this case, the outer numeric data represent the INCOMETOTAL values, and the inner horizontal and vertical data represent the HOUSINGTYPE data. It is observed that there are four significant paired groups based on the blue lines not crossing the diagonal dotted line.

First One-Way ANOVA

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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
HOUSINGTYPE	5	8.995E21	1.799E21	1.10	0.3576
Error	7994	1.306E25	1.634E21		

Figure 13. First ANOVA model Levene's test.

Figure 13 shows the Levene's test to test for the assumption of homogeneity of variance for the first ANOVA model (Kleinman, 2012). Based on the output, *F*-value (1.10) with corresponding *p*-value (0.3532) is greater than 0.05 level of significance. Thus, the null hypothesis (H_0) stands. It is concluded that the assumption of homogeneity of variance is met.

One-way ANOVA with Blocking

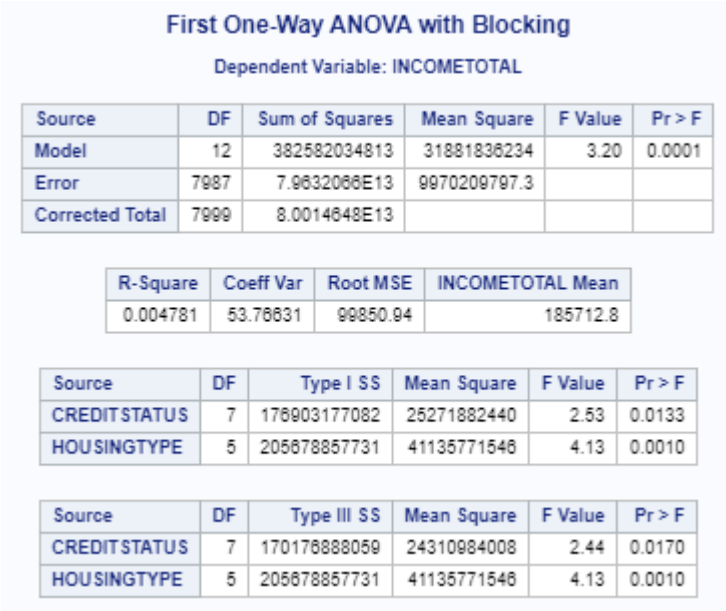
Moving on, blocking is performed to isolate the variability due to the factor of the credit loan status (CREDITSTATUS) from the bank dataset. Therefore, instead of randomizing the way of living across all 48 combinations, it is suggested to only randomize the application of the six ways of living within each of the eight credit loan status. The observations from this randomized block design are outputted below, starting with *Figure 14*.

First One-Way ANOVA with Blocking

Class Level Information

Class	Levels	Values
CREDITSTATUS	8	0 1 2 3 4 5 C X
HOUSINGTYPE	6	Co-op apartment House / apartment Municipal apartme Office apartment Rented apartment With parents

Number of Observations Read	8000
Number of Observations Used	8000



First One-Way ANOVA with Blocking

Dependent Variable: INCOMETOTAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	382582034813	31881836234	3.20	0.0001
Error	7987	7.9632066E13	9970209797.3		
Corrected Total	7999	8.0014648E13			

R-Square	Coeff Var	Root MSE	INCOMETOTAL Mean
0.004781	53.76631	99850.94	185712.8

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CREDITSTATUS	7	176903177082	25271882440	2.53	0.0133
HOUSINGTYPE	5	205678857731	41135771546	4.13	0.0010

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CREDITSTATUS	7	170176888059	24310984008	2.44	0.0170
HOUSINGTYPE	5	205678857731	41135771546	4.13	0.0010

Figure 14. First ANOVA model with blocking class-level-information table and analysis of variance table.

Based on *Figure 14*, the output specifies the eight levels and its values of class (blocking) variable (CREDITSTATUS), six levels and its values of the class variable (HOUSINGTYPE), and the number of observations read versus the number of observations used are equal. Next, the output contains all of the information needed to test the equality of the group means which is divided into four parts. For the first part, the value of the test statistics, i.e. the *F*-statistic and corresponding *p*-value are reported in the analysis of variance table. Since there are six types of ways of living (HOUSINGTYPE) used, this analysis is to test the hypothesis on whether the means of annual income (INCOMETOTAL) are equal for all types of ways of living, controlling for credit loan status (CREDITSTATUS). The following shows the details of the null hypothesis (H_0) and the alternative hypothesis (H_1):

H_0: $\mu 1 = \mu 2 = \mu 3 = \mu 4 = \mu 5 = \mu 6$: All the means of annual income are the same

H_1: At least one pair of the annual income (μ_i) means is different

From the analysis of variance table, the overall *F*-test of *F*-value (3.20) with corresponding *p*-value (0.0001) is less than 0.05 level of significance. Thus, the null hypothesis (H_0) is rejected. It is concluded that at least one pair of annual income means is different, when controlling for credit loan status and indicates that there are significant differences between the means of annual income across ways of living or blocks (credit loan status). By comparing the overall *F*-test between the original one-way ANOVA with the one with the blocking variable, it shows that the overall *F*-test (0.0001) with blocking variable is smaller compared with the model without, with an overall *F*-test value of 0.0007.

Next, by comparing the estimate of the experimental error variance (MSE) between the original one-way ANOVA with the one which blocking variable (CREDITSTATUS) is included, it is noted that the data (99913.80) from the model with blocking performed is smaller compared to the data (99850.94) from the model that included the ways of living (HOUSINGTYPE) only. Depending on the magnitude of the difference, this could affect the comparisons between the treatment means by finding more significant differences than the HOUSINGTYPE-only model, given the same sample sizes.

Also, the R-square for this model (0.4%) is slightly greater than that in the previous model (0.2%). To some degree, this is a function of having more model degrees of freedom, but it is unlikely this is the last reason for this magnitude of difference. Most importantly, it is observed that the effect of ways of living (HOUSINGTYPE) in this model is still significant with *F*-value (4.13) with corresponding *p*-value (0.0010). Likewise, the effect of the blocking variable (CREDITSTATUS) is also significant with *F*-value (2.44) with corresponding *p*-value (0.0170). Since the overall *F*-test from this model is better than the model without blocking performed and the class variable still significant, it is concluded that adding the blocking variable (CREDITSTATUS) into the design and analysis is detrimental to the test of HOUSINGTYPE.

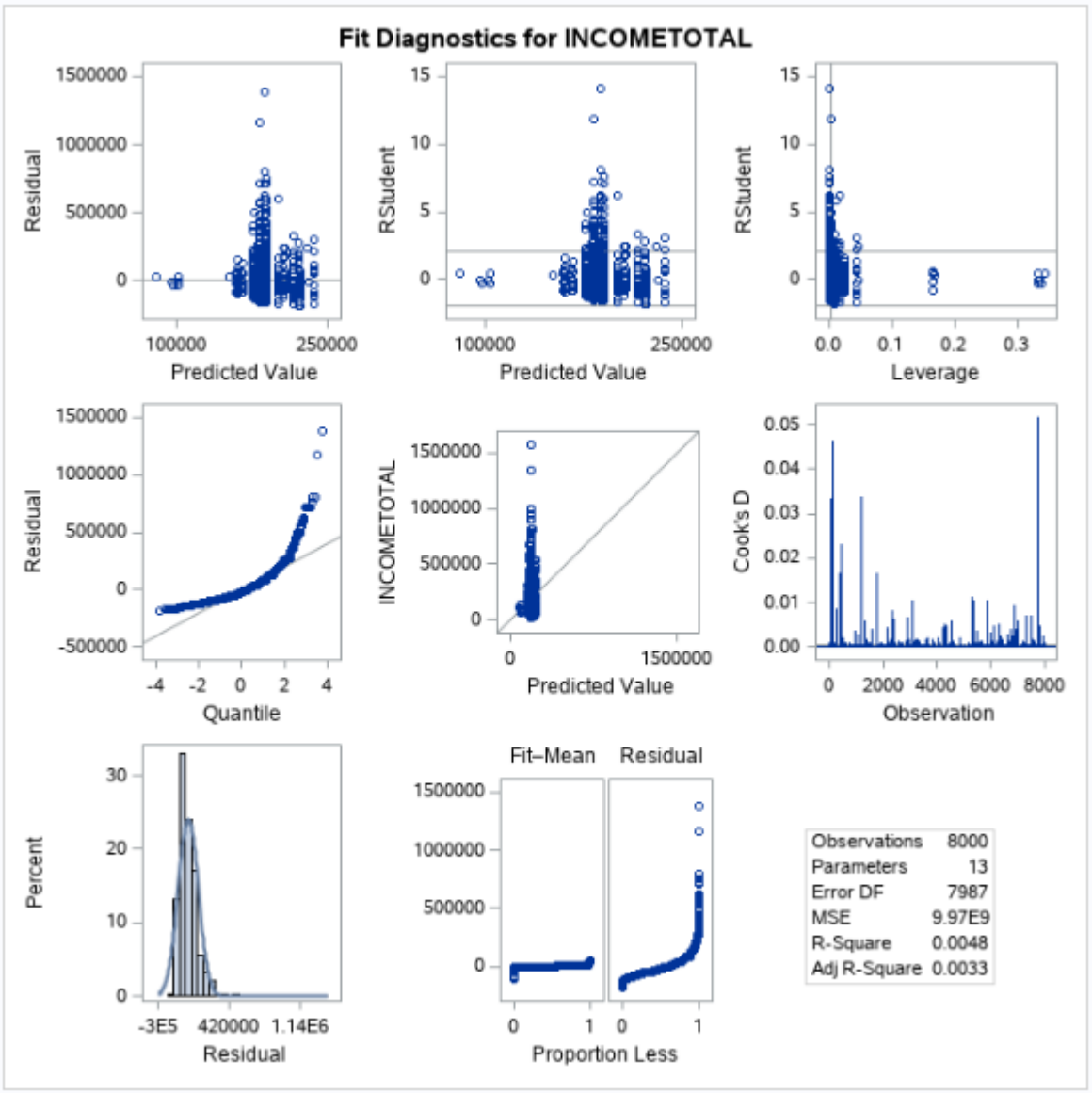


Figure 15. First ANOVA model with blocking fit diagnostics plot.

Figure 15 shows the fit diagnostics plot for the ANOVA model with the blocking variable. In foresight, the plot shows similar results with the model without including the blocking variable. For instance, it shows the data are not scattered throughout the Residuals plot, there is a severe departure from normality from the Q-Q plot, and the histogram is positively skewed with the tail extending to the right.

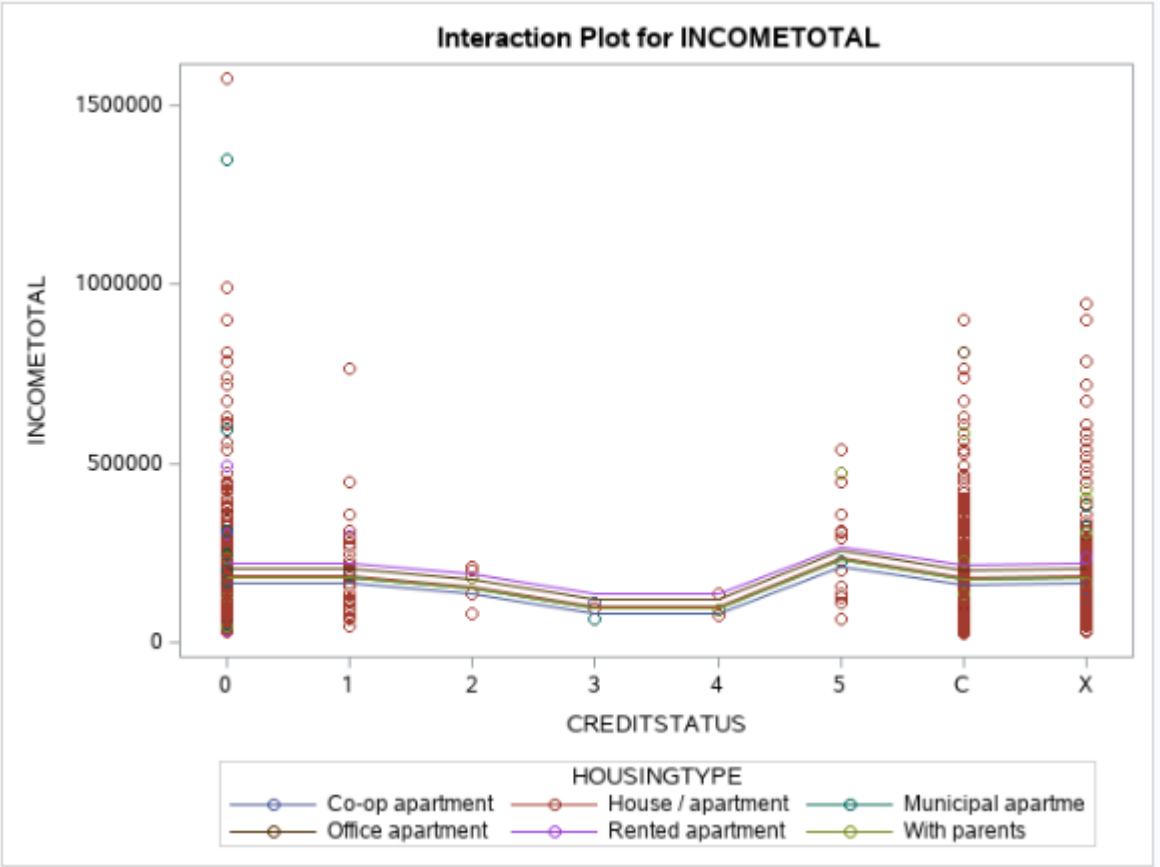


Figure 16. First ANOVA model with blocking interaction plot.

Figure 16 shows the interaction plot with the dependent variable (INCOMETOTAL), categorical variable (HOUSINGTYPE), and blocking variable (CREDITSTATUS). Although the plot is hard to examine due to extensive numbers of dots, it is indicated that there is no interaction between the variables for all 48 combinations. Based on the data given in *Figure 14*, there is a significant difference (p-value=0.0001< 0.05) for the different group, meaning different combinations could have a different annual income. It is concluded that there is a significant difference in the annual income (INCOMETOTAL) due to different credit loan status (CREDITSTATUS) and ways of living (HOUSINGTYPE).

First One-Way ANOVA with Blocking and Post-Hoc Pairwise Comparisons

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

HOUSINGTYPE	INCOMETOTAL LSMEAN	LSMEAN Number
Co-op apartment	145908.608	1
House / apartment	168828.857	2
Municipal apartme	162979.663	3
Office apartment	186920.517	4
Rented apartment	201759.765	5
With parents	161914.317	6

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

Dependent Variable: INCOMETOTAL

i/j	1	2	3	4	5	6
1		0.6790	0.9108	0.3295	0.0215	0.9234
2	0.6790		0.9450	0.7428	0.0032	0.7660
3	0.9108	0.9450		0.5696	0.0050	1.0000
4	0.3295	0.7428	0.5696		0.9368	0.4791
5	0.0215	0.0032	0.0050	0.9368		0.0013
6	0.9234	0.7660	1.0000	0.4791	0.0013	

Figure 17. First ANOVA model with blocking post-hoc analysis: Tukey's Multiple Comparisons.

After that, *Figure 17* shows the post-hoc analysis, Tukey's multiple comparisons, for the first ANOVA model with blocking performed. Based on the output, the way of living with the highest mean annual income is living in a rented apartment and the one with the lowest is living in a co-op apartment, which conclusion of output is similar to the model without blocking performed. The data is visualized in *Figure 18*.

Likewise, the *p*-values for comparing each of the group means are displayed as a symmetric matrix which shows all the pairwise comparison *p*-values twice. The matrix is interpreted as the groups having a *p*-value of less than 0.05 are significant to one another. In case this, it is observed that four groups; 5*1, 5*2, 5*3, and 5*6, are significant to one another. Similarly, the data is visualized in *Figure 19*.

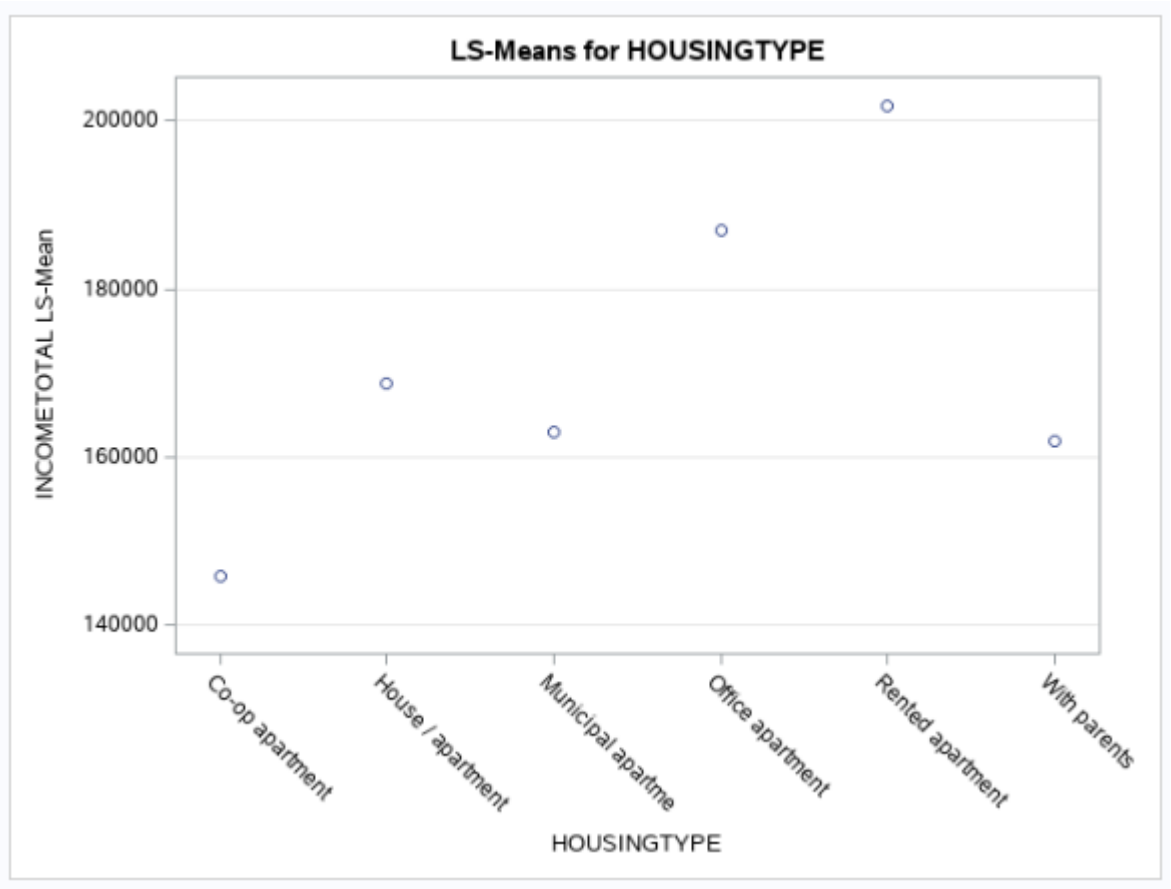


Figure 18. First ANOVA model with blocking LS-Means frequency dot plot.

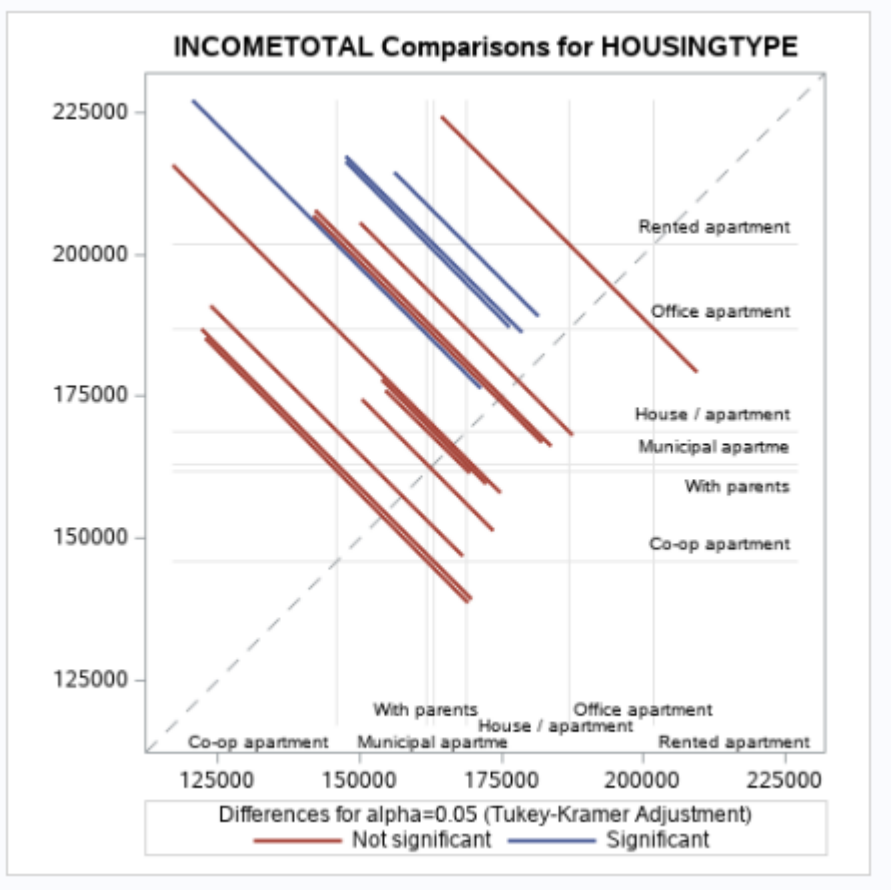


Figure 19. First ANOVA model with blocking Tukey's diffogram.

Figure 19 shows a diffogram for the first ANOVA model with blocking performed. The diffogram provides the same results similar to the ANOVA model without blocking performed, where the outer numeric data represent the INCOMETOTAL values, and the inner horizontal and vertical data represent the HOUSINGTYPE data. Similarly, It is observed that there are four significant paired groups based on the blue lines not crossing the diagonal dotted line.

First One-Way ANOVA with Blocking and Post-Hoc Pairwise Comparisons

Least Squares Means
Adjustment for Multiple Comparisons: Dunnett-Hsu

HOUSINGTYPE	INCOMETOTAL LSMEAN	H0:LSMean=Control Pr > |t|
Co-op apartment	145908.608	0.0072
House / apartment	168828.857	0.0010
Municipal apartme	162979.663	0.0016
Office apartment	186920.517	0.7752
Rented apartment	201759.765	
With parents	161914.317	0.0004

Figure 20. First ANOVA model with blocking post-hoc analysis: Dunnett's Multiple Comparisons.

Subsequently, *Figure 20* shows the post-hoc analysis, Dunnett's multiple comparisons, for the first ANOVA model with blocking performed. Based on the output, the way of living with the highest mean annual income is living in a rented apartment, which is used as a control group for the blocking model, and the one with the lowest is living in a co-op apartment, which conclusion of output is similar to the model without blocking performed. Since the data "rented apartment" is used as a control group for the Dunnett's test, it will not display any value for the *p*-value. The p-value is the two-tailed probability computed using t distribution. It is the probability of observing a greater absolute value of t under the null hypothesis. For a one-tailed test, halve this probability. If p-value is less than the pre-specified alpha level (usually .05 or .01), it is concluded that mean is statistically significantly different from zero (Yankovsky, 2015). In this case, only the "office apartment" data from HOUSINGTYPE is more than the alpha level of 0.05, which is concluded that "Office apartment" is not significantly different from the control group. The data is visualized in *Figure 21*.

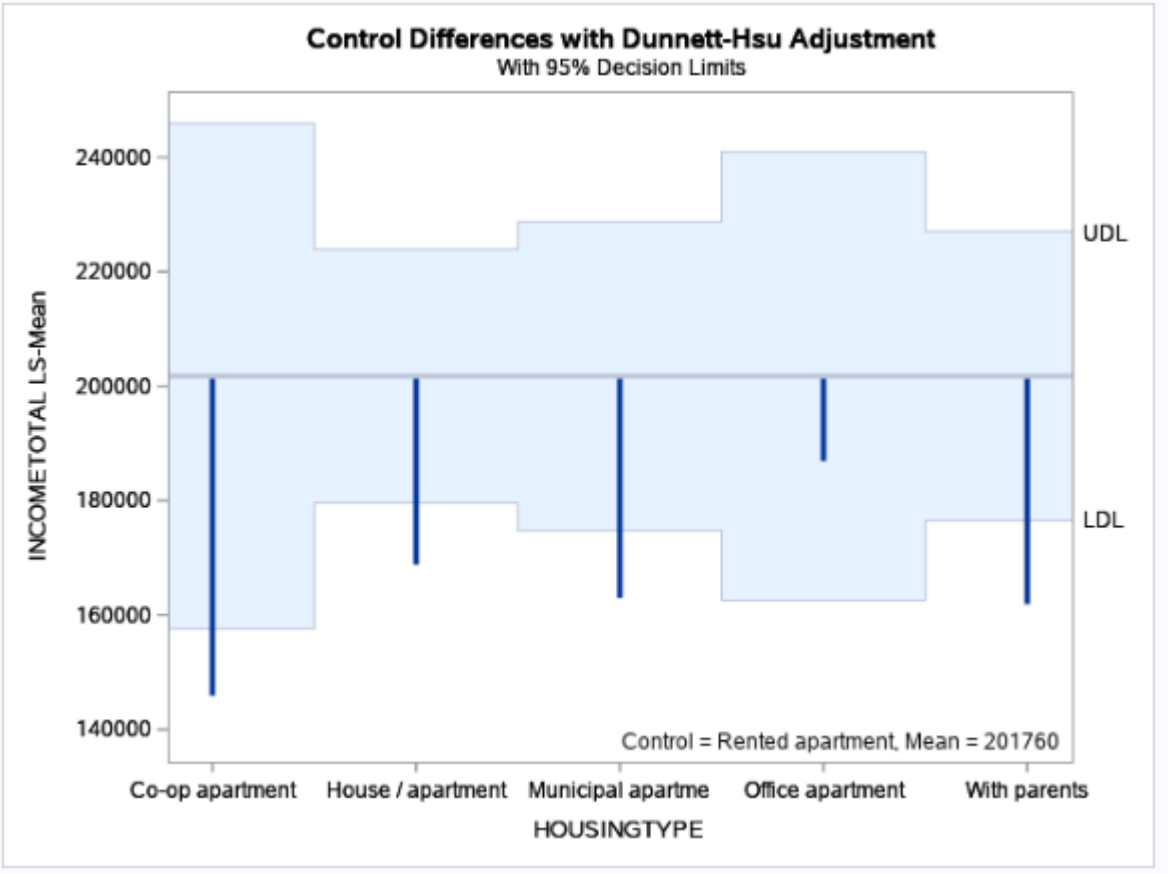


Figure 21. First ANOVA model with blocking control plot.

Figure 21 shows the LS-mean control plot for the first ANOVA model with blocking, given the control group is specified to be “Rented apartment” for the HOUSINGTYPE variable. The value of the control group mean is shown as a horizontal line, which is around the LS-mean of 200000. The shaded light blue area is bounded by the UDL (Upper Decision Limit) and LDL (Lower Decision Limit). Based on the output, the all vertical lines extend past the shaded area except for the “Office apartment” line. Hence, all of the means in the group represented by the line is significantly different from the control group except for “Office apartment”.

First One-Way ANOVA with Blocking and Post-Hoc Pairwise Comparisons

Least Squares Means

HOUSINGTYPE	INCOMETOTAL LSMEAN	LSMEAN Number
Co-op apartment	145908.608	1
House / apartment	168828.857	2
Municipal apartme	162979.663	3
Office apartment	186920.517	4
Rented apartment	201759.765	5
With parents	161914.317	6

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

Dependent Variable: INCOMETOTAL

i/j	1	2	3	4	5	6
1		0.1397	0.3068	0.0432	0.0017	0.3254
2	0.1397		0.3648	0.1695	0.0002	0.1822
3	0.3068	0.3648		0.1003	0.0004	0.8955
4	0.0432	0.1695	0.1003		0.3485	0.0752
5	0.0017	0.0002	0.0004	0.3485		<.0001
6	0.3254	0.1822	0.8955	0.0752	<.0001	

Figure 22. First ANOVA model with blocking post-hoc analysis: Least Squares Means (default)

After that, *Figure 22* shows the default post-hoc analysis, which really signifies no adjustment for multiple comparisons for the first ANOVA model with blocking performed ("The GLM Procedure", 2012). Based on the output, the way of living with the highest mean annual income is living in a rented apartment and the one with the lowest is living in a co-op apartment, which conclusion of output is similar to the model without blocking performed.

Likewise, the *p*-values for comparing each of the group means are displayed as a symmetric matrix which shows all the pairwise comparison *p*-values twice. The matrix is interpreted as the groups having a *p*-value of less than 0.05 are significant to one another. In case this, it is observed that five groups; 4*1, 5*1, 5*2, 5*3, and 5*6, are significant to one another. Similarly, the data is visualized in *Figure 23*.

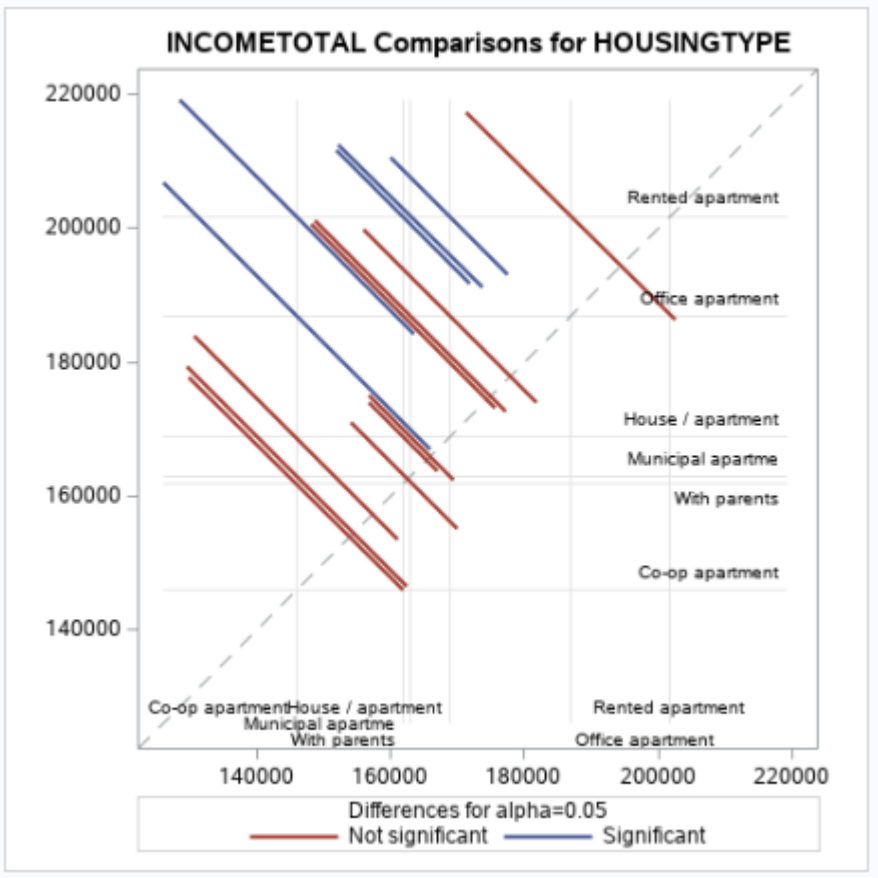


Figure 23. First ANOVA model with blocking diffogram (default).

Figure 23 shows a diffogram for the first ANOVA model with blocking performed. The default diffogram shows the outer numeric data represent the INCOMETOTAL values, and the inner horizontal and vertical data represent the HOUSINGTYPE data. It is observed that there are five significant paired groups based on the blue lines not crossing the diagonal dotted line.

Second Analysis of Variance (ANOVA)

One-way ANOVA

Moving on, the second ANOVA model is created to study five education levels (EDUCATIONLEVEL); Academic degree, Higher education, Incomplete higher, Lower secondary, and Secondary / secondary special. The bank also has information showing the way of living (HOUSINGTYPE) to each education level based on the annual income. The One-way ANOVA is performed to see whether the average annual income is significantly different for various education levels.

Second One-Way ANOVA

Class Level Information		
Class	Levels	Values
EDUCATIONLEVEL	5	Academic degree Higher education Incomplete higher Lower secondary Secondary / secondary special

Number of Observations Read	8000
Number of Observations Used	8000

Second One-Way ANOVA

Dependent Variable: INCOMETOTAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4.585029E12	1.1462572E12	121.50	<.0001
Error	7995	7.5429619E13	9434598963.7		
Corrected Total	7999	8.0014648E13			

R-Square	Coeff Var	Root MSE	INCOMETOTAL Mean
0.057302	52.30218	97131.86	185712.8

Source	DF	Type I SS	Mean Square	F Value	Pr > F
EDUCATIONLEVEL	4	4.585029E12	1.1462572E12	121.50	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EDUCATIONLEVEL	4	4.585029E12	1.1462572E12	121.50	<.0001

Figure 24. Second ANOVA model class-level-information table and analysis of variance table.

Based on *Figure 24*, the output specifies the five levels and its values of the class variable (EDUCATIONLEVEL), and the number of observations read versus the number of observations used are equal. These values are the same because there are no missing values in any variable in the model. Next, the output contains all of the information needed to test the

equality of the group means which is divided into three parts. For the first part, the value of the test statistics, i.e. the *F*-statistic and corresponding *p*-value are reported in the analysis of variance table. Since there are five types of education levels (EDUCATIONLEVEL) used, this analysis is to test the hypothesis on whether the means of annual income (INCOMETOTAL) are equal for all types of education level. The following shows the details of the null hypothesis (H_0) and the alternative hypothesis (H_1):

H_0: $\mu 1 = \mu 2 = \mu 3 = \mu 4 = \mu 5$: All the means of annual income are the same

H_1: At least one pair of the annual income (μ_i) means is different

From the analysis of variance table, *F*-value (121.50) with corresponding *p*-value (<.0001) is less than 0.05 level of significance. Thus, the null hypothesis (H_0) is rejected. It is concluded that at least one pair of annual income means is different and indicates that there are significant differences between the means of annual income across education levels.

Secondly, the coefficient of determination, R^2, denoted in the table as *R*-Square, is a measure of the proportion of variability explained by the independent variables in the analysis. It is interpreted that the education level (EDUCATIONLEVEL) explains about 5.7% of the variability of the annual income (INCOMETOTAL) in the model. Likewise, the coefficient of variation (Coeff Var) of 52.3 expresses the root MSE (the estimate of the standard deviation for all treatments) of 97131.9 as a percent of the mean. Hence, the INCOMETOTAL Mean is the mean of all of the data values in the variable INCOMETOTAL without regard to EDUCATIONLEVEL.

Since this is a one-way ANOVA model, the third part shows the information about the class variable (EDUCATIONLEVEL) in the model is an exact duplicate of the model line of the analysis of variance table.

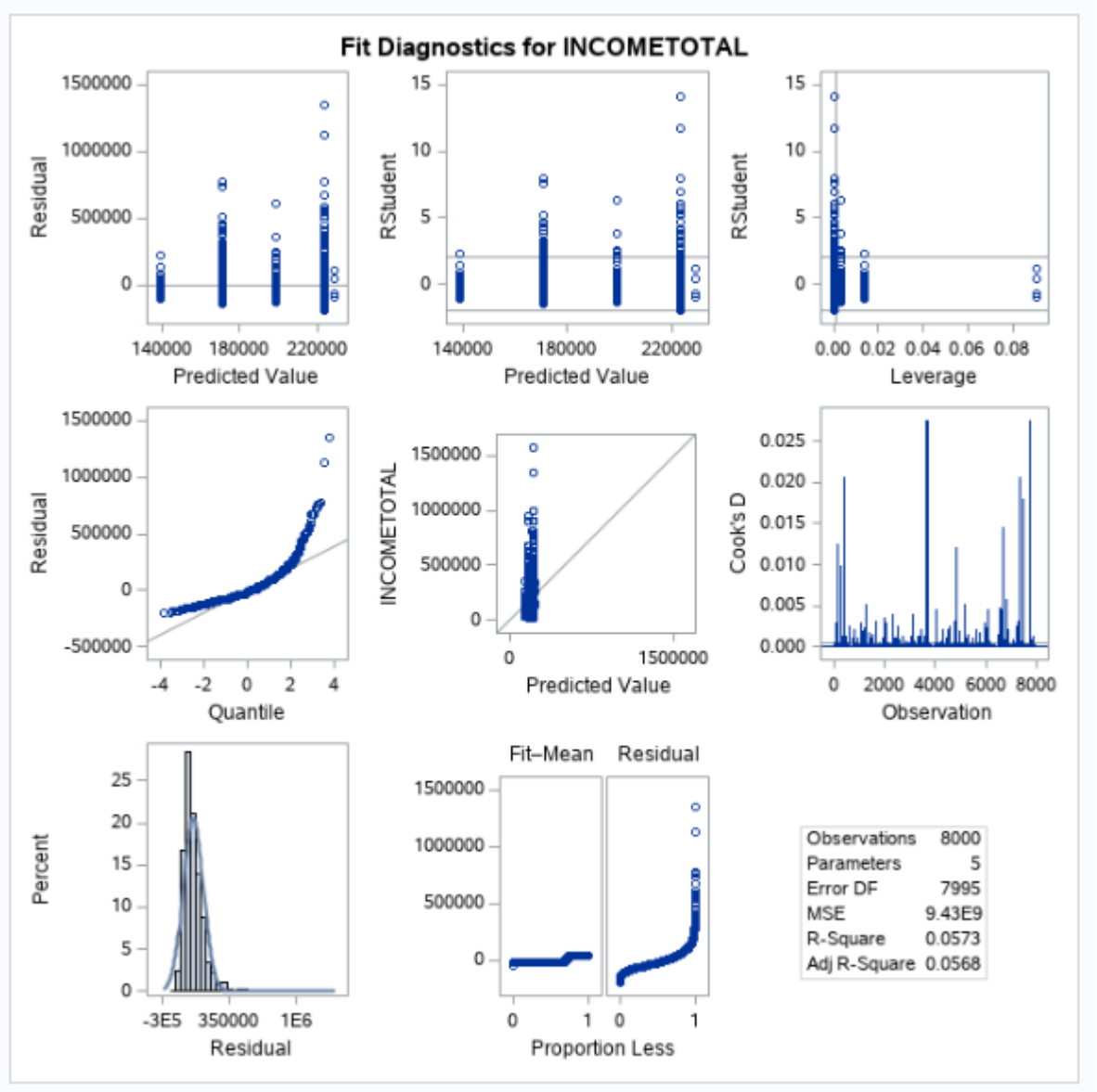


Figure 25. Second ANOVA model fit diagnostics plot.

Moving on, the ANOVA model carries three assumptions to be looked into from the three plots in the left pane outputted in *Figure 25*. The plot at the upper-left is a Residuals by Predicted plot, which looks at the random scatter within each group of EDUCATIONLEVEL. The plot indicates there is a violation in the model assumption due to the data not being scattered within each group of EDUCATIONLEVEL. Next, the plot at the center-left is a Quantile-Quantile (Q-Q) plot, which checks for normality assumption in the model. Based on the plot, it is concluded that there appears to have severe departure from normality because the observations tend to lie to the diagonal reference line up till the second quartile where it

starts to depart from the line. Lastly, the bottom-left plot shows a positively skewed histogram (tail extending to the right) with a unique peak.

Second One-Way ANOVA

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

EDUCATIONLEVEL	INCOMETOTAL LSMEAN	LSMEAN Number
Academic degree	229090.909	1
Higher education	223509.436	2
Incomplete higher	199081.552	3
Lower secondary	138593.919	4
Secondary / secondary special	170557.197	5

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

Dependent Variable: INCOMETOTAL

i/j	1	2	3	4	5
1		0.9997	0.8528	0.0322	0.2676
2	0.9997		0.0006	<.0001	<.0001
3	0.8528	0.0006		<.0001	<.0001
4	0.0322	<.0001	<.0001		0.0396
5	0.2676	<.0001	<.0001	0.0396	

Figure 26. Second ANOVA model post-hoc analysis: Tukey's Multiple Comparisons.

After that, *Figure 26* shows the post-hoc analysis, Tukey's multiple comparisons, for the second ANOVA model. Based on the output, the education level with the highest mean annual income is holding an academic degree and the one with the lowest is considered as lower secondary. The data is visualized in *Figure 27*.

Furthermore, the *p*-values for comparing each of the group means are displayed as a symmetric matrix which shows all the pairwise comparison *p*-values twice. The matrix is interpreted as the groups having a *p*-value of less than 0.05 are significant to one another. In case this, it is observed that seven groups; 1*4, 2*3, 2*4, 2*5, 3*4, 3*5, and 4*5, are significant to one another. Similarly, the data is visualized in *Figure 28.*

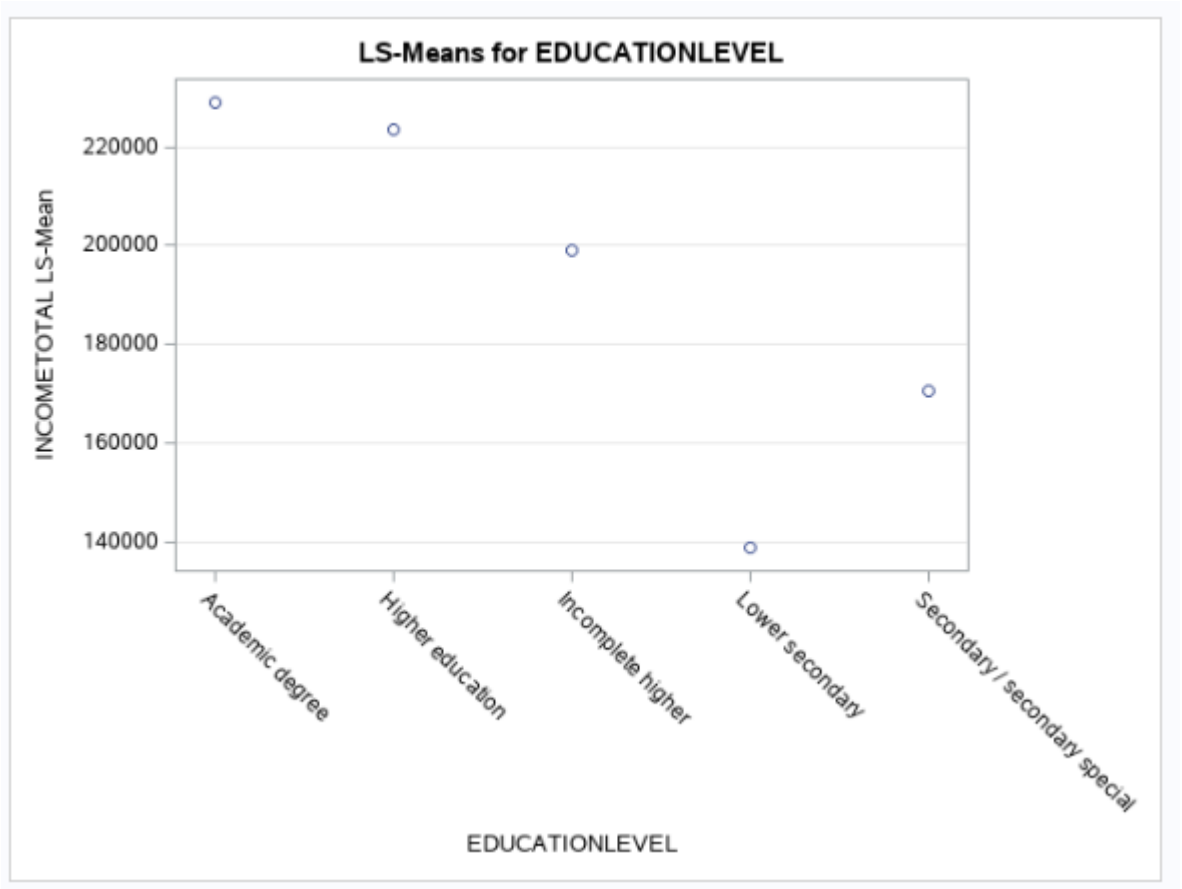


Figure 27. Second ANOVA model LS-Means frequency dot plot.

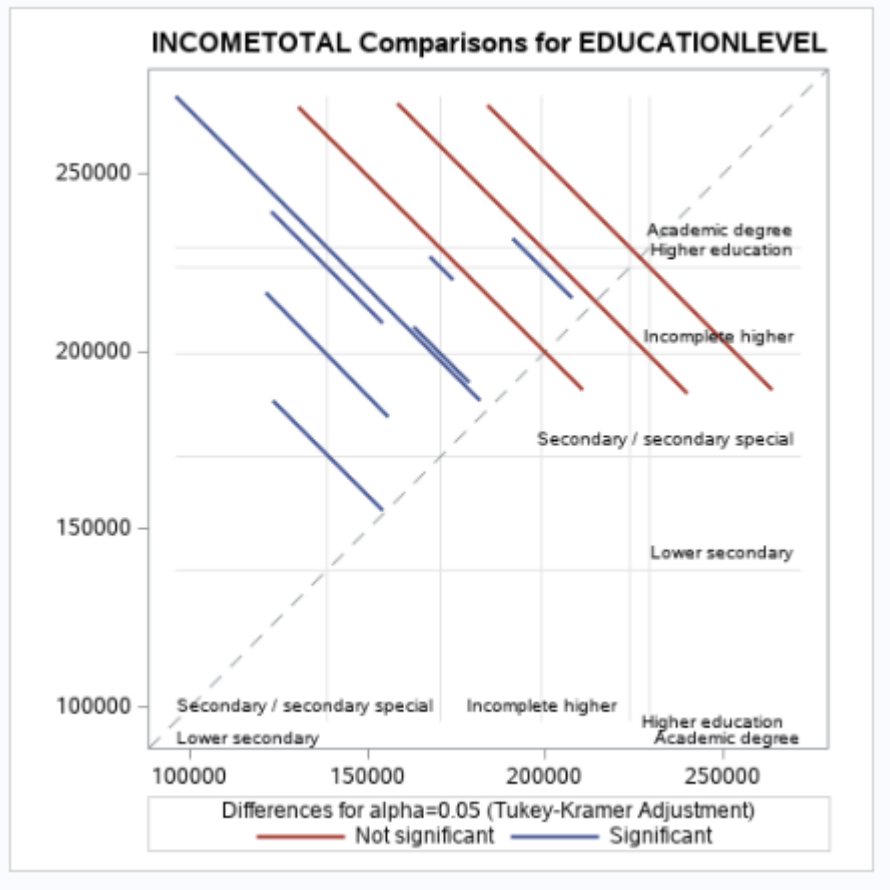


Figure 28. Second ANOVA model Tukey's diffogram.

Figure 28 shows a diffogram for the second ANOVA model. The point estimates for the differences between pairs of group means can be found at the intersections of vertical and horizontal lines. Also, the colored diagonal lines show the confidence intervals for the differences. In this case, the outer numeric data represent the INCOMETOTAL values, and the inner horizontal and vertical data represent the EDUCATIONLEVEL data. It is observed that there are seven significant paired groups based on the blue lines not crossing the diagonal dotted line.

Second One-Way ANOVA

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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
EDUCATIONLEVEL	4	1.095E23	2.738E22	19.26	<.0001
Error	7995	1.136E25	1.421E21		

Figure 29. Second ANOVA model Levene's test.

Figure 29 shows the Levene's test to test for the assumption of homogeneity of variance for the second ANOVA model. Based on the output, *F*-value (19.26) with corresponding *p*-value (<0.0001) is less than 0.05 level of significance. Thus, the null hypothesis (H_0) is rejected. It is concluded that the assumption of homogeneity of variance is not met.

One-way ANOVA with Blocking

Moving on, blocking is performed to isolate the variability due to the factor of the way of living (HOUSINGTYPE) from the bank dataset. Therefore, instead of randomizing the way of living across all 30 combinations, it is suggested to only randomize the application of the five education levels within each of the six ways of living. The observations from this randomized block design are outputted below, starting with *Figure 30*.

Second One-Way ANOVA with Blocking

Class Level Information		
Class	Levels	Values
HOUSINGTYPE	6	Co-op apartment House / apartment Municipal apartme Office apartment Rented apartment With parents
EDUCATIONLEVEL	5	Academic degree Higher education Incomplete higher Lower secondary Secondary / secondary special

Number of Observations Read	8000
Number of Observations Used	8000

Second One-Way ANOVA with Blocking

Dependent Variable: INCOMETOTAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	4.8594515E12	539939054655	57.40	<.0001
Error	7990	7.5155196E13	9406157220.8		
Corrected Total	7999	8.0014648E13			

R-Square	Coeff Var	Root MSE	INCOMETOTAL Mean
0.060732	52.22328	96985.35	185712.8

Source	DF	Type I SS	Mean Square	F Value	Pr > F
HOUSINGTYPE	5	212405146755	42481029351	4.52	0.0004
EDUCATIONLEVEL	4	4.6470463E12	1.1617616E12	123.51	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
HOUSINGTYPE	5	274422520282	54884504056	5.83	<.0001
EDUCATIONLEVEL	4	4.6470463E12	1.1617616E12	123.51	<.0001

Figure 30. Second ANOVA model with blocking class-level-information table and analysis of variance table.

Based on *Figure 30*, the output specifies the six levels and its values of class (blocking) variable (HOUSINGTYPE), five levels and its values of the class variable (EDUCATIONLEVEL), and the number of observations read versus the number of observations used are equal. Next, the output contains all of the information needed to test the equality of the group means which is divided into four parts. For the first part, the value of the test statistics, i.e. the *F*-statistic and corresponding *p*-value are reported in the analysis of variance table. Since there are five types of education levels (EDUCATIONLEVEL) used, this analysis is to test the hypothesis on whether the means of annual income (INCOMETOTAL) are equal for all types of education levels, controlling for ways of living (HOUSINGTYPE). The following shows the details of the null hypothesis (H_0) and the alternative hypothesis (H_1):

H_0: $\mu 1 = \mu 2 = \mu 3 = \mu 4 = \mu 5 = \mu 6$: All the means of annual income are the same

H_1: At least one pair of the annual income (μ_i) means is different

From the analysis of variance table, the overall *F*-test of *F*-value (57.40) with corresponding *p*-value (<.0001) is less than 0.05 level of significance. Thus, the null hypothesis (H_0) is rejected. It is concluded that at least one pair of annual income means is different, when controlling for the ways of living and indicates that there are significant differences between the means of annual income across education levels or blocks (ways of living). By comparing the overall *F*-test between the original one-way ANOVA with the one with the blocking variable, it shows that the overall *F*-test (<.0001) with blocking variable is the same compared with the model without blocking performed.

Next, by comparing the estimate of the experimental error variance (MSE) between the original one-way ANOVA with the one which blocking variable (HOUSINGTYPE) is included, it is noted that the data (96985.35) from the model with blocking performed is smaller compared to the data (97131.86) from the model that included the education levels (EDUCATIONLEVEL) only. Depending on the magnitude of the difference, this could affect the comparisons between the treatment means by finding more significant differences than the EDUCATIONLEVEL-only model, given the same sample sizes.

Also, the R-square for this model (6.1%) is slightly greater than that in the previous model (5.7%). To some degree, this is a function of having more model degrees of freedom, but it is unlikely this is the last reason for this magnitude of difference. Most importantly, it is observed that the effect of education level (EDUCATIONLEVEL) in this model is still significant with *F*-value (123.51) with corresponding *p*-value (<.0001). Likewise, the effect of the blocking variable (HOUSINGTYPE) is also significant with *F*-value (5.83) with corresponding *p*-value (<.0001). Since the overall *F*-test from this model is better than the model without blocking performed and the class variable still significant, it is concluded that adding the blocking variable (HOUSINGTYPE) into the design and analysis is detrimental to the test of EDUCATIONLEVEL.

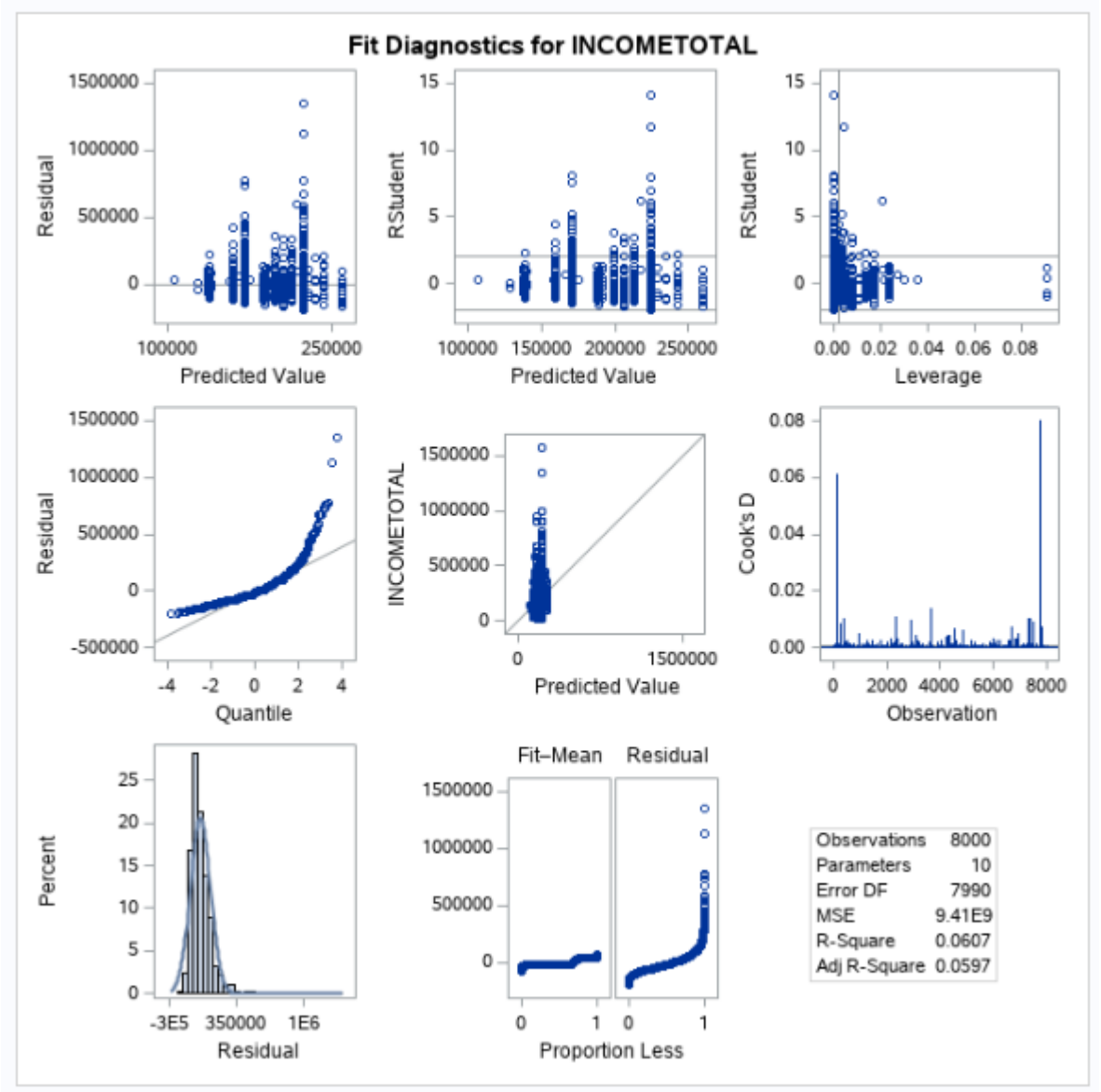


Figure 31. Second ANOVA model with blocking fit diagnostics plot.

Figure 31 shows the fit diagnostics plot for the ANOVA model with the blocking variable. In foresight, the plot shows similar results with the model without including the blocking variable. For instance, it shows the data are not scattered throughout the Residuals plot, there is a severe departure from normality from the Q-Q plot, and the histogram is positively skewed with the tail extending to the right.

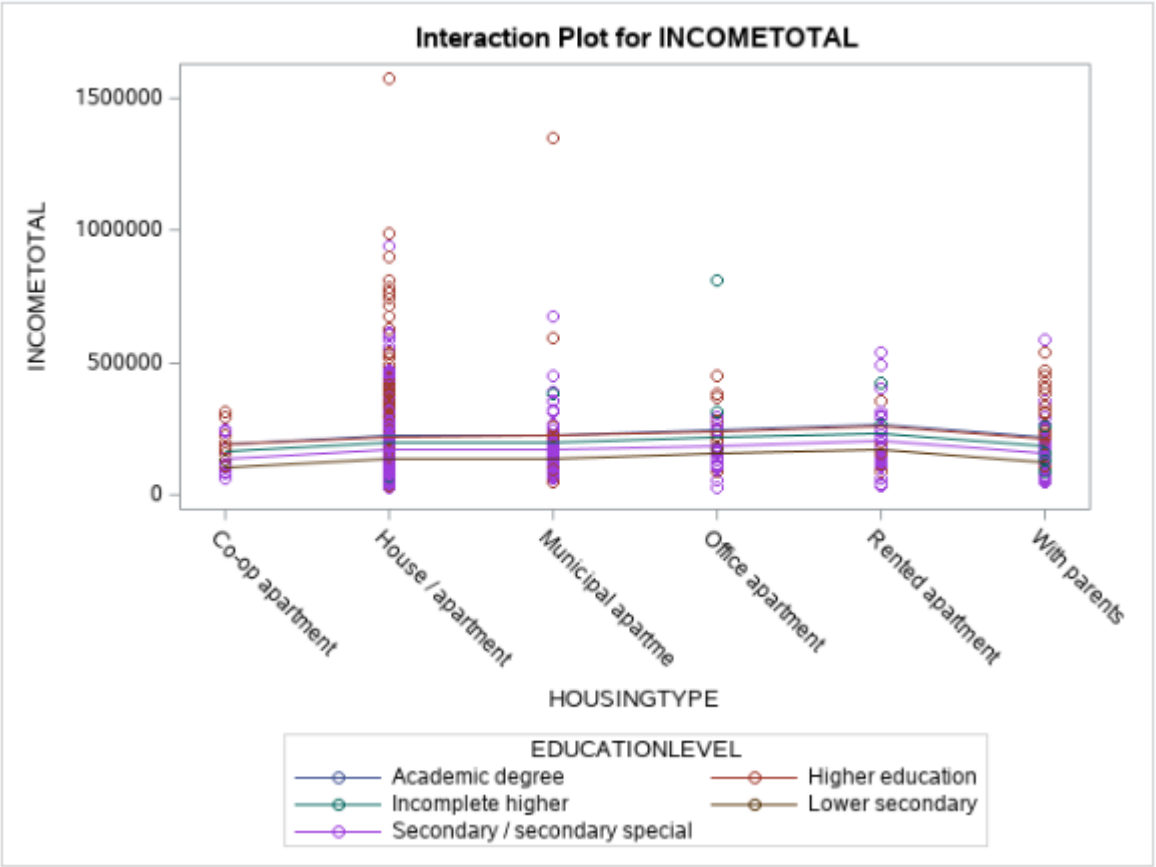
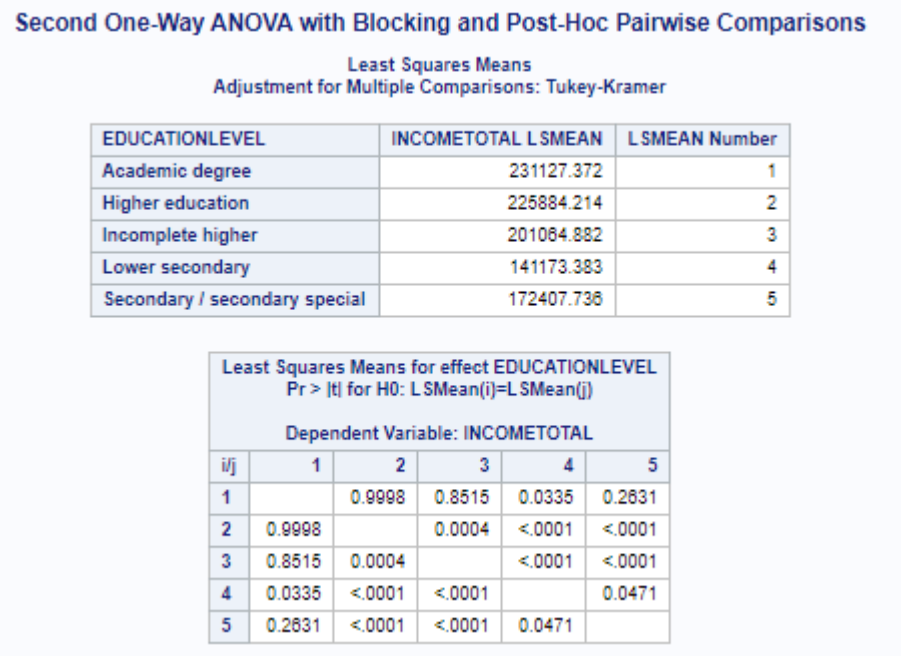


Figure 32. Second ANOVA model with blocking interaction plot.

Figure 32 shows the interaction plot with the dependent variable (INCOMETOTAL), categorical variable (EDUCATIONLEVEL), and blocking variable (HOUSINGTYPE). Although the plot is hard to examine due to extensive numbers of dots, it is indicated that there is no interaction between the variables for all 30 combinations. Based on the data given in *Figure 30*, there is a significant difference (p-value=<.0001< 0.05) for the different group, meaning different combinations could have a different annual income. It is concluded that there is a significant difference in the annual income (INCOMETOTAL) due to different ways of living (HOUSINGTYPE) and education levels (EDUCATIONLEVEL).



Second One-Way ANOVA with Blocking and Post-Hoc Pairwise Comparisons

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

EDUCATIONLEVEL	INCOMETOTAL LSMEAN	LSMEAN Number
Academic degree	231127.372	1
Higher education	225884.214	2
Incomplete higher	201064.882	3
Lower secondary	141173.383	4
Secondary / secondary special	172407.736	5

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

Dependent Variable: INCOMETOTAL

i/j	1	2	3	4	5
1		0.9998	0.8515	0.0335	0.2631
2	0.9998		0.0004	<.0001	<.0001
3	0.8515	0.0004		<.0001	<.0001
4	0.0335	<.0001	<.0001		0.0471
5	0.2631	<.0001	<.0001	0.0471	

Figure 33. Second ANOVA model with blocking post-hoc analysis: Tukey's Multiple Comparisons.

After that, *Figure 33* shows the post-hoc analysis, Tukey's multiple comparisons, for the second ANOVA model with blocking performed. Based on the output, the education level with the highest mean annual income is holding an academic degree and the one with the lowest is considered as lower secondary, which conclusion of output is similar to the model without blocking performed. The data is visualized in *Figure 34*.

Likewise, the *p*-values for comparing each of the group means are displayed as a symmetric matrix which shows all the pairwise comparison *p*-values twice. The matrix is interpreted as the groups having a *p*-value of less than 0.05 are significant to one another. In case this, it is observed that seven groups; 1*4, 2*3, 2*4, 2*5, 3*4, 3*5, and 4*5, are significant to one another. Similarly, the data is visualized in *Figure 35*.

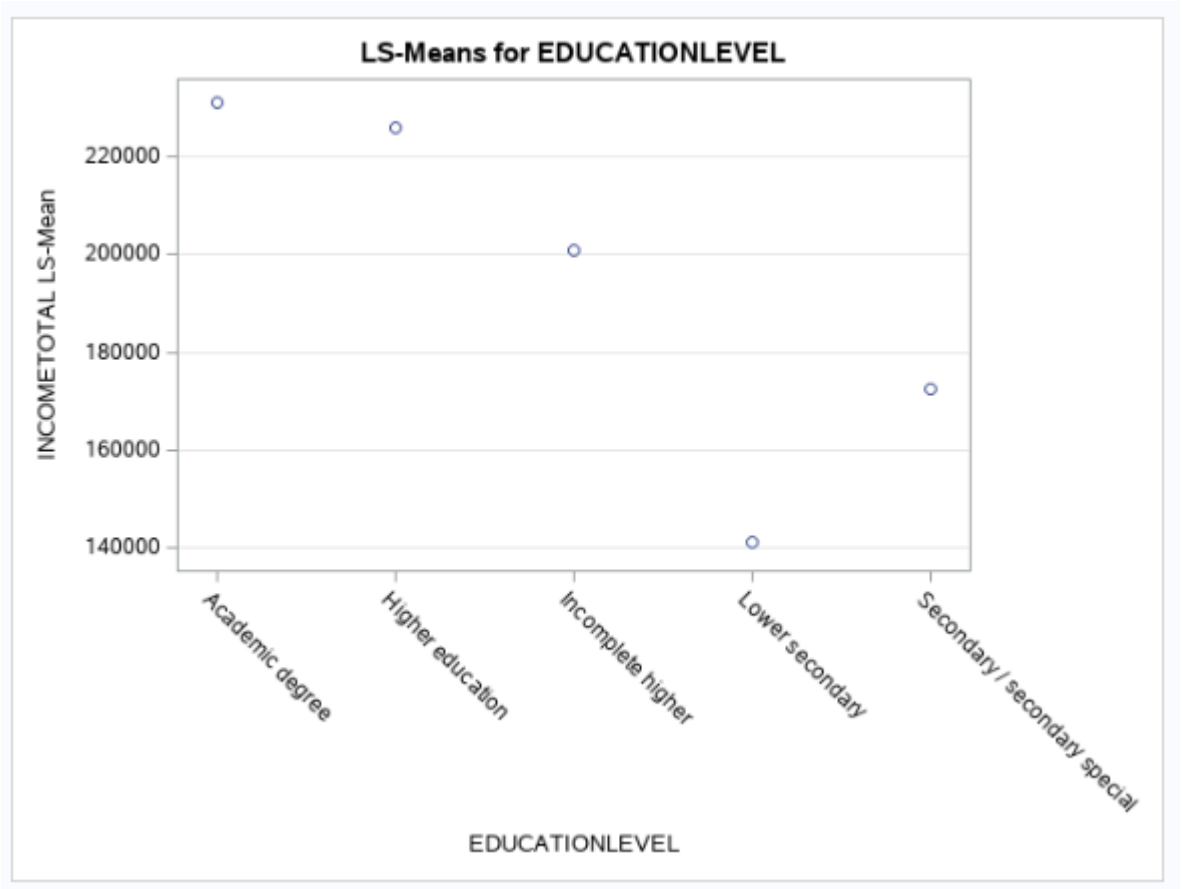


Figure 34. Second ANOVA model with blocking LS-Means frequency dot plot.

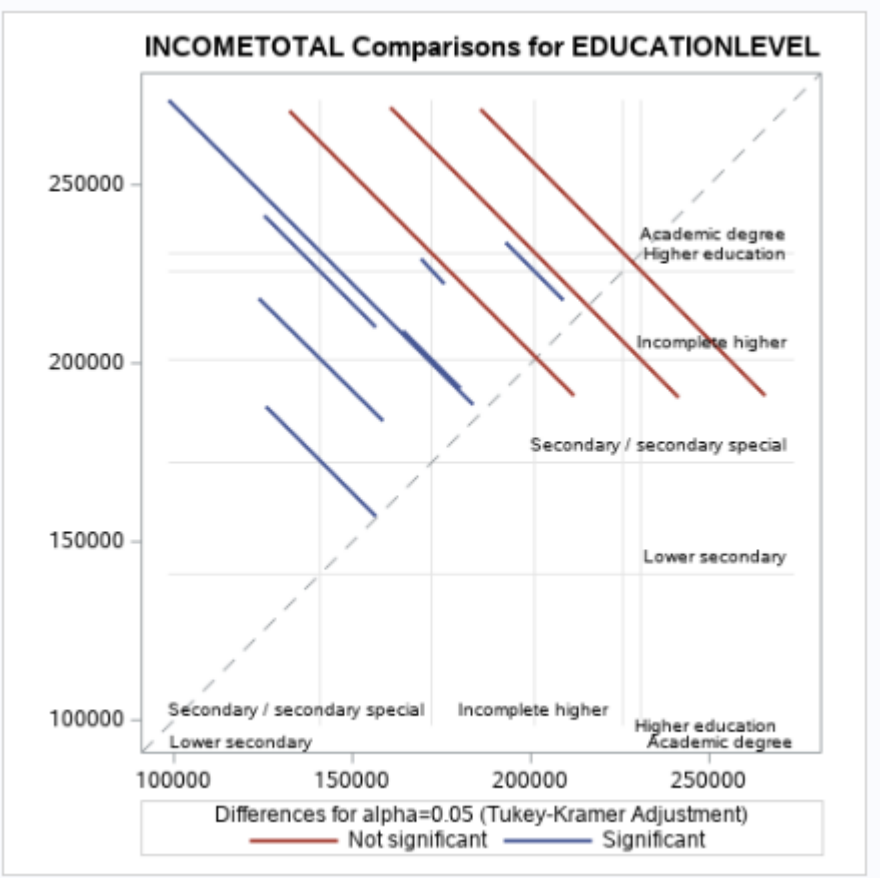


Figure 35. Second ANOVA model with blocking Tukey's diffogram.

Figure 35 shows a diffogram for the second ANOVA model with blocking performed. The diffogram provides the same results similar to the ANOVA model without blocking performed, where the outer numeric data represent the INCOMETOTAL values, and the inner horizontal and vertical data represent the EDUCATIONLEVEL data. Similarly, It is observed that there are seven significant paired groups based on the blue lines not crossing the diagonal dotted line.

Second One-Way ANOVA with Blocking and Post-Hoc Pairwise Comparisons

Least Squares Means
Adjustment for Multiple Comparisons: Dunnett-Hsu

EDUCATIONLEVEL	INCOMETOTAL LSMEAN	H0:LSMean=Control Pr > |t|
Academic degree	231127.372	0.0112
Higher education	225884.214	<.0001
Incomplete higher	201064.882	<.0001
Lower secondary	141173.383	
Secondary / secondary special	172407.736	0.0161

Figure 36. Second ANOVA model with blocking post-hoc analysis: Dunnett's Multiple Comparisons.

Subsequently, *Figure 36* shows the post-hoc analysis, Dunnett's multiple comparisons, for the second ANOVA model with blocking performed. Based on the output, the education level with the lowest mean annual income is classified as "Lower secondary", which is used as a control group for the blocking model, and the one with the highest holds an academic degree, which conclusion of output is similar to the model without blocking performed. Since the data "Lower secondary" is used as a control group for the Dunnett's test, it will not display any value for the *p*-value. In this case, all of the groups are less than the alpha level of 0.05, which is concluded that all groups are significantly different from the control group. The data is visualized in *Figure 37*.

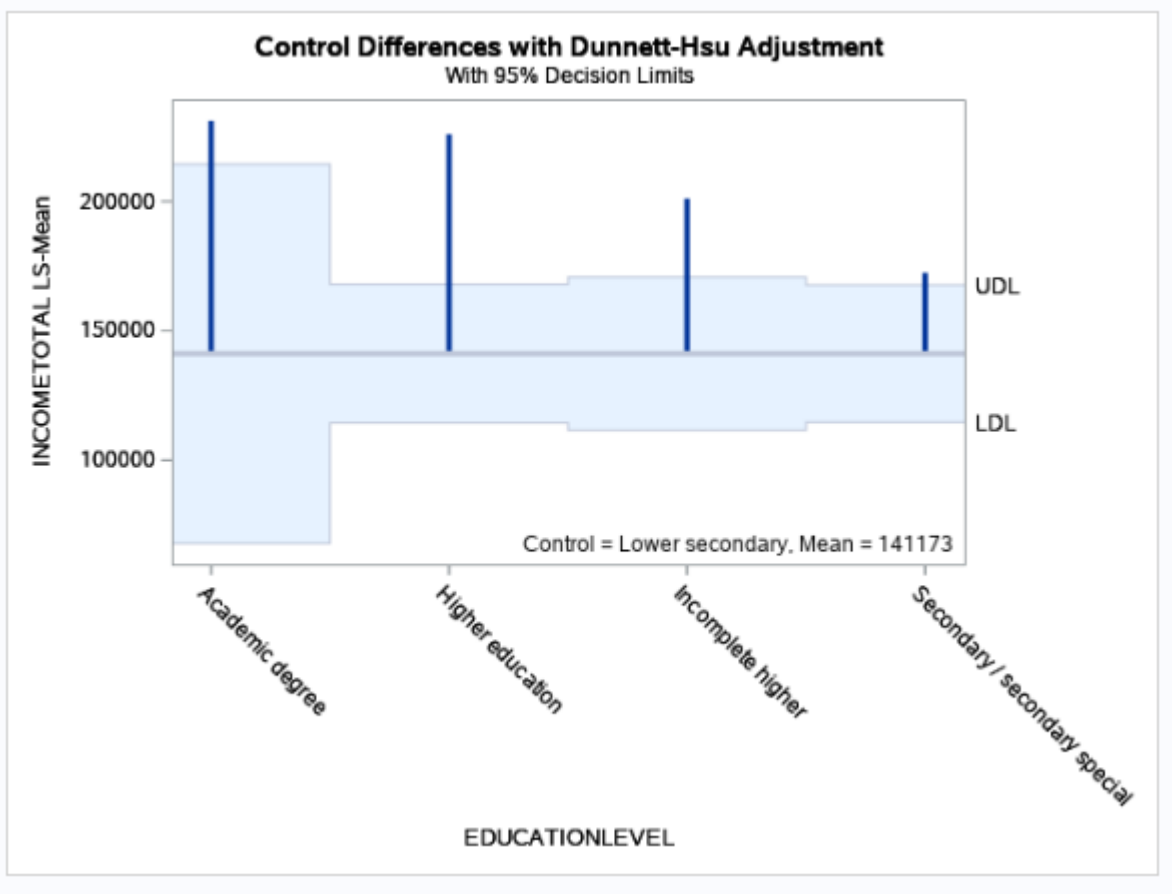


Figure 37. Second ANOVA model with blocking control plot.

Figure 37 shows the LS-mean control plot for the second ANOVA model with blocking, given the control group is specified to be "Lower secondary" for the EDUCATIONLEVEL variable. The value of the control group mean is shown as a horizontal line, which is around the LS-mean of 150000. The shaded light blue area is bounded by the UDL (Upper Decision Limit) and LDL (Lower Decision Limit). Based on the output, the all vertical lines extend past the shaded area. Hence, all of the means in the group represented by the line is significantly different from the control group.

Second One-Way ANOVA with Blocking and Post-Hoc Pairwise Comparisons

Least Squares Means

EDUCATIONLEVEL	INCOMETOTAL LSMEAN	LSMEAN Number
Academic degree	231127.372	1
Higher education	225884.214	2
Incomplete higher	201084.882	3
Lower secondary	141173.383	4
Secondary / secondary special	172407.736	5

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

Dependent Variable: INCOMETOTAL

i/j	1	2	3	4	5
1		0.8581	0.3131	0.0041	0.0449
2	0.8581		<.0001	<.0001	<.0001
3	0.3131	<.0001		<.0001	<.0001
4	0.0041	<.0001	<.0001		0.0060
5	0.0449	<.0001	<.0001	0.0060	

Figure 38. Second ANOVA model with blocking post-hoc analysis: Least Squares Means (default)

After that, *Figure 38* shows the default post-hoc analysis, which really signifies no adjustment for multiple comparisons for the second ANOVA model with blocking performed. Based on the output, the education level with the highest mean annual income is holding an academic degree and the one with the lowest is considered to be lower secondary, which conclusion of output is similar to the model without blocking performed.

Likewise, the *p*-values for comparing each of the group means are displayed as a symmetric matrix which shows all the pairwise comparison *p*-values twice. The matrix is interpreted as the groups having a *p*-value of less than 0.05 are significant to one another. In case this, it is observed that eight groups; 1*4, 1*5, 2*3, 2*4, 2*5, 3*4, 3*5, and 4*5, are significant to one another. Similarly, the data is visualized in *Figure 39*.

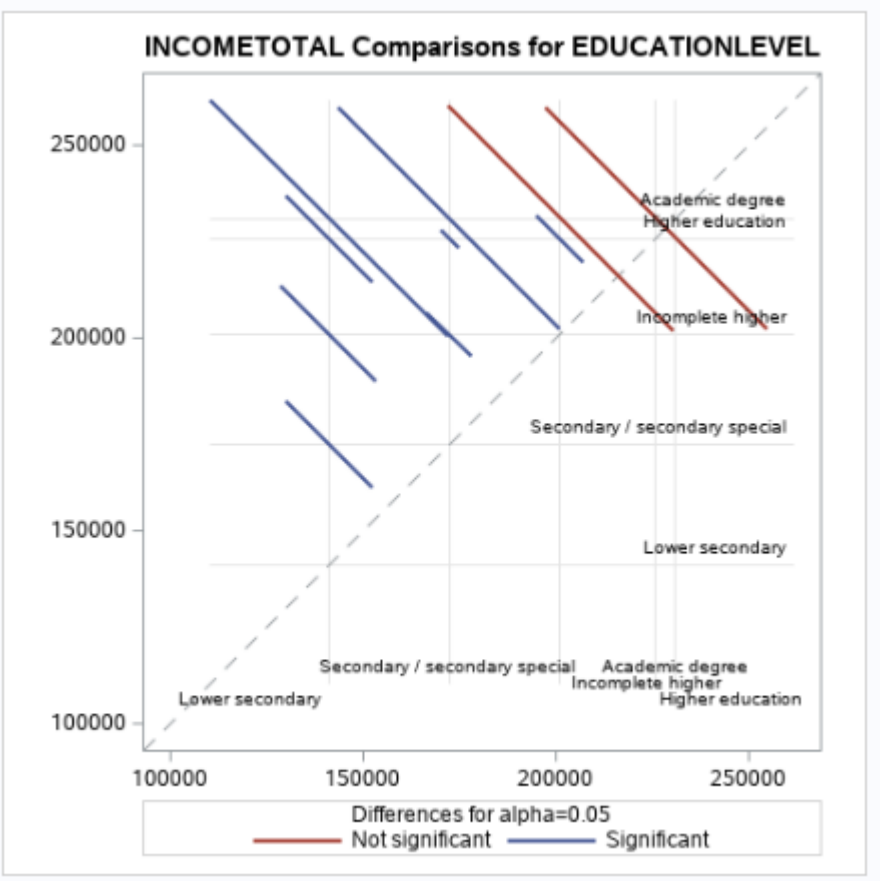


Figure 39. Second ANOVA model with blocking diffogram (default).

Figure 39 shows a diffogram for the second ANOVA model with blocking performed. The default diffogram shows the outer numeric data represent the INCOMETOTAL values, and the inner horizontal and vertical data represent the EDUCATIONLEVEL data. It is observed that there are eight significant paired groups based on the blue lines not crossing the diagonal dotted line.

Nonparametric Test

The ANOVA tests rely on parametric assumptions (requirements about nature of shape of the population involved), especially the assumption that the sample data are normality distributed (assumption of normality). When the sample data does not meet the distributional assumptions, special kinds of statistical procedures known as non-parametric tests (distribution-free tests) are used to help treat the problem.

First Nonparametric One-way ANOVA

Since the sample data from the first AVOVA did not meet all the distribution assumptions of simple random samples and normality, a nonparametric test one-way ANOVA test is performed based on the first ANOVA model to show more appropriate results.

Distribution Examination

Firstly, a distribution data analysis is performed to examine the distribution of data to determine the suitable analyses to be conducted. Based on the data, the plots conducted are histogram and Q-Q plot with the mean, standard deviation, skewness, kurtosis, normal test statistics, and normal test p-value.

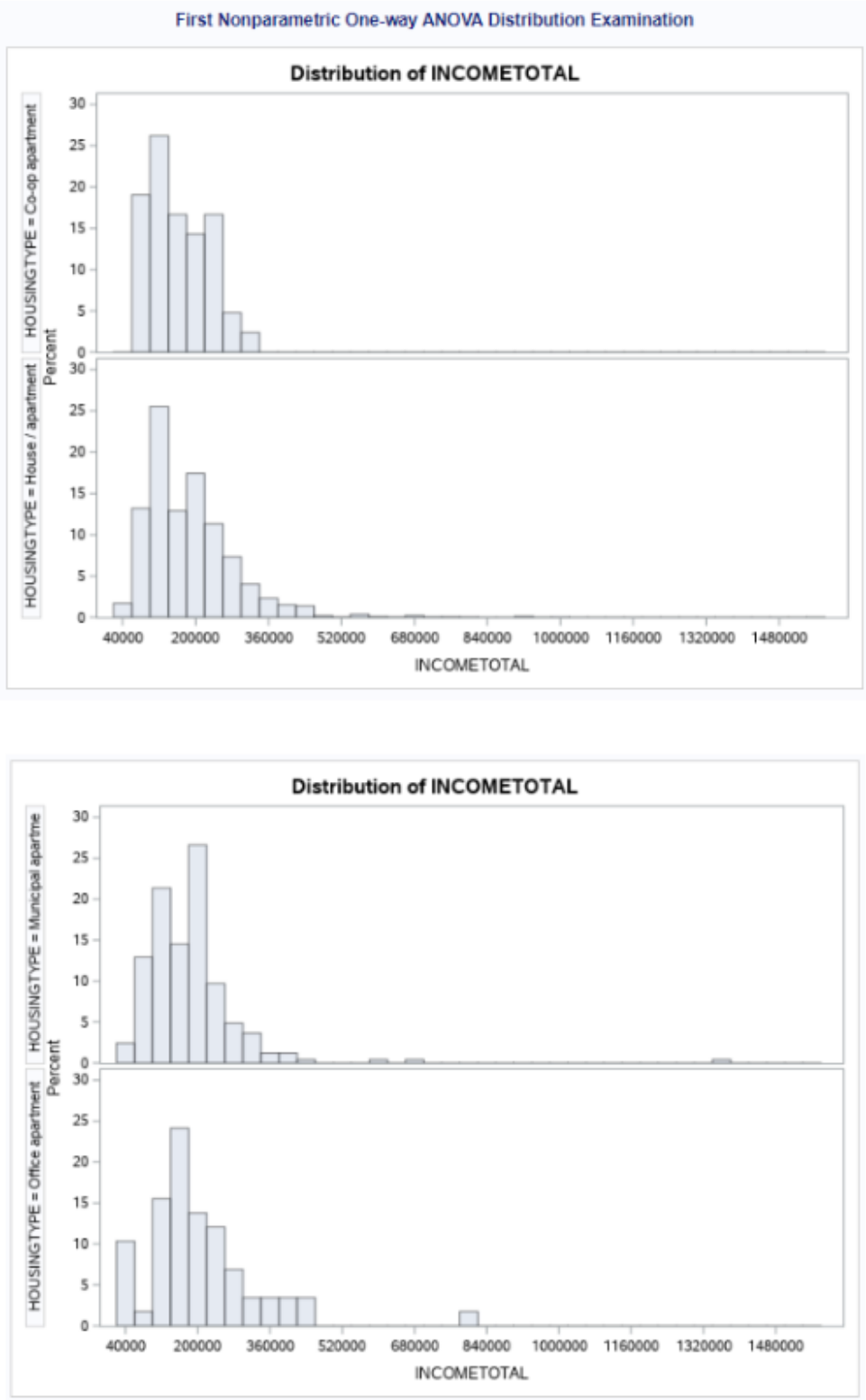


Figure 40. Nonparametric one-way ANOVA histogram distribution based on the first ANOVA model.

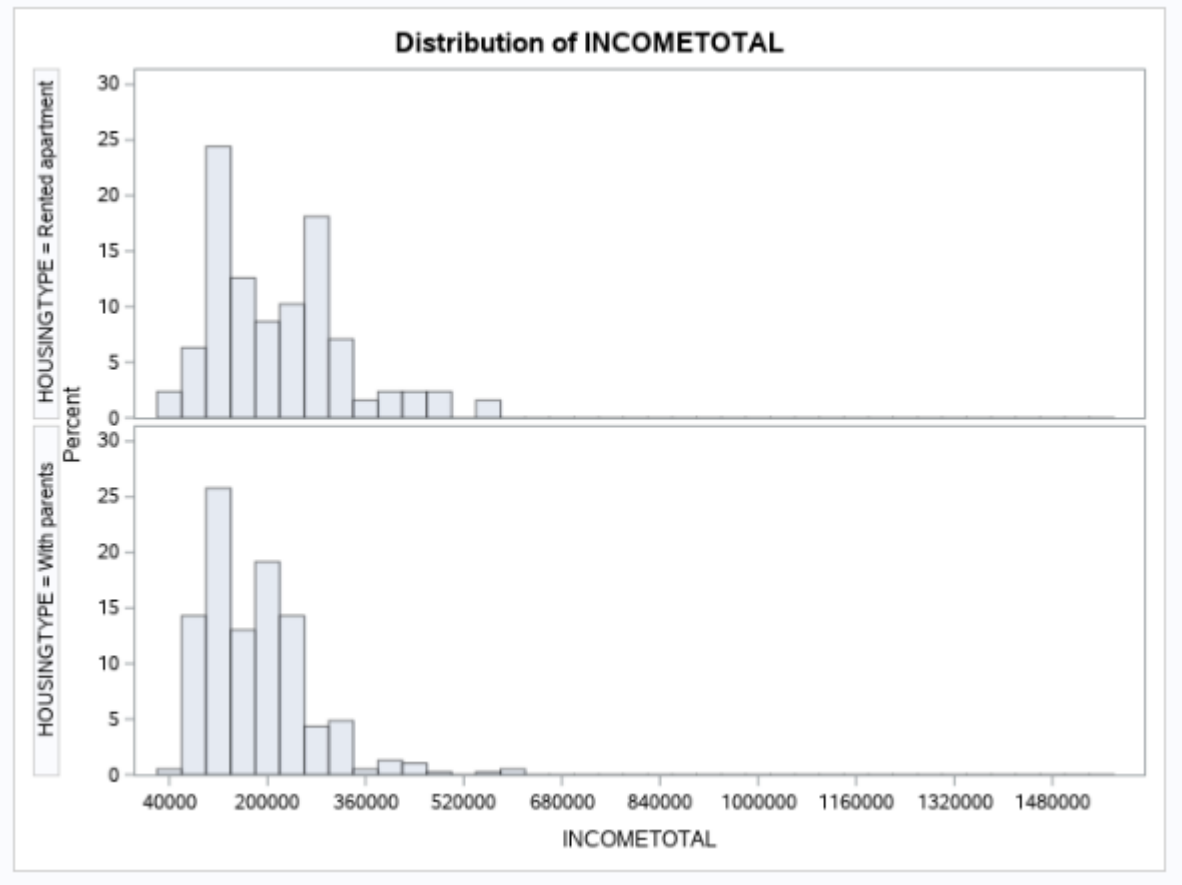


Figure 41. Nonparametric one-way ANOVA histogram distribution based on the first ANOVA model. (continued)

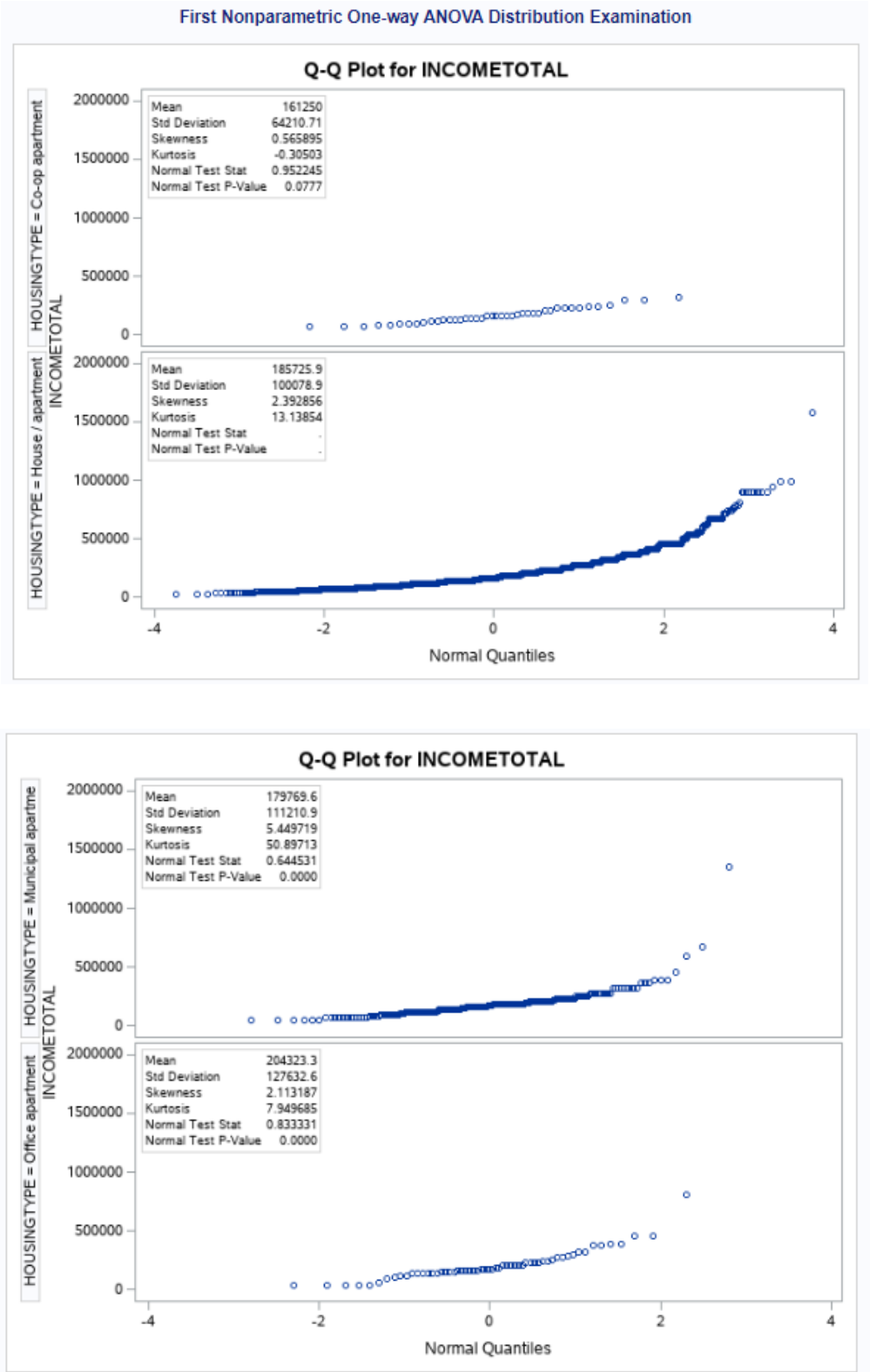


Figure 42. Nonparametric one-way ANOVA Q-Q plot distribution based on the first ANOVA model.

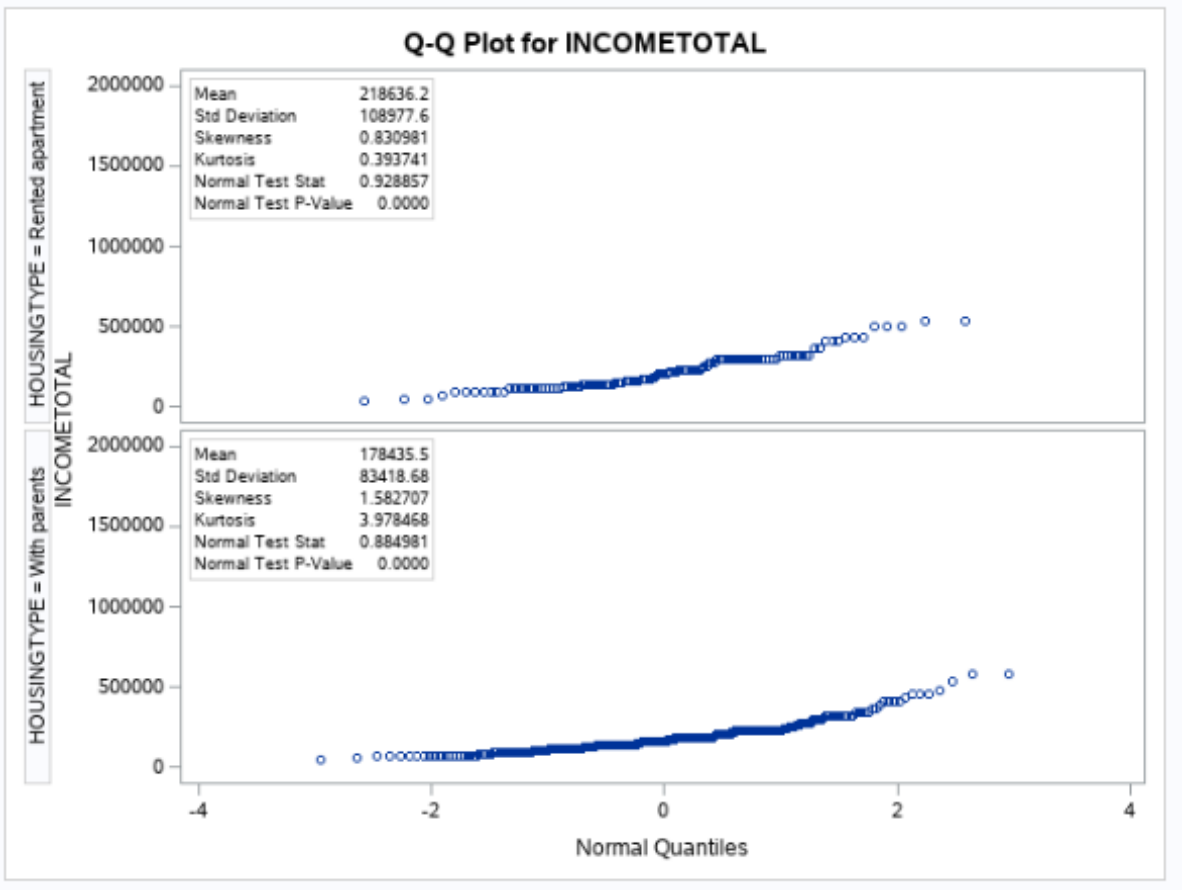


Figure 43. Nonparametric one-way ANOVA histogram distribution based on the first ANOVA model. (continued)

Based on all the plots and values of skewness and kurtosis, the data do not appear to be normally distributed for any speciality code. The goodness-of-fit tests reject the null hypothesis that the data are normally distributed.

Kruskal-Wallis Test

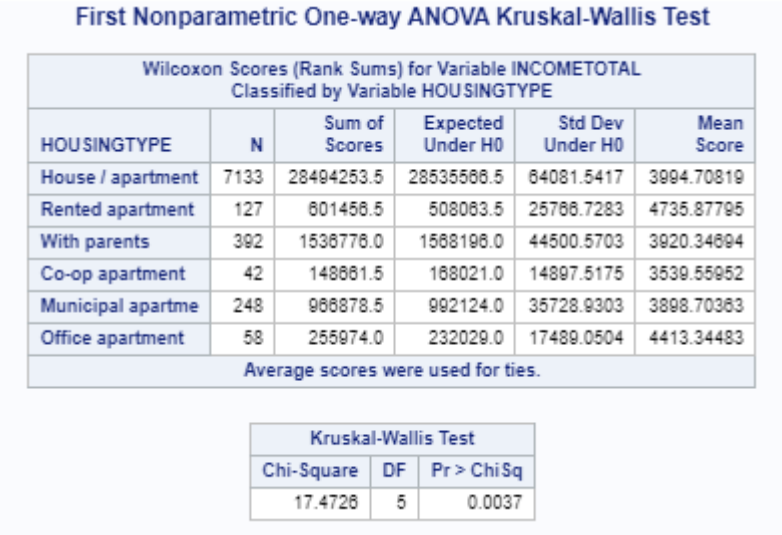
The parametric equivalent of ANOVA is the Kruskal-Wallis test. The purpose of the Kruskal-Wallis test is to test the null hypothesis (H_0) that the independent samples come from

populations, and determine if there are statistically significant differences between two or more groups of an independent variable (HOUSINGTYPE) on a continuous dependent variable (INCOMETOTAL).

H_0: The samples come from populations with equal medians.

H_1: The samples come from populations with medians that are not all equal.

For illustrative purposes, the Wilcoxon option is used to perform a Rank Sum test and the Median option to perform the Median test which uses the ranks of sample data from INCOMETOTAL classified by HOUSINGTYPE. The data is visualized in *Figure 44*.



First Nonparametric One-way ANOVA Kruskal-Wallis Test

Wilcoxon Scores (Rank Sums) for Variable INCOMETOTAL Classified by Variable HOUSINGTYPE

HOUSINGTYPE	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
House / apartment	7133	28494253.5	28535566.5	64081.5417	3994.70819
Rented apartment	127	601456.5	508063.5	25766.7283	4735.87795
With parents	392	1536776.0	1568196.0	44500.5703	3920.34694
Co-op apartment	42	148661.5	168021.0	14897.5175	3539.55952
Municipal apartme	248	966878.5	992124.0	35728.9303	3898.70363
Office apartment	58	255974.0	232029.0	17489.0504	4413.34483

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
17.4726	5	0.0037

Figure 44. Nonparametric one-way ANOVA Wilcoxon-score table and Kruskal-Wallis test table based on the first ANOVA model.

Based on the output in *Figure 44*, the Wilcoxon-score table shows the mean rank score for each of the experimental groups of HOUSINGTYPE. The data with the highest mean score is shown to be “Rented apartment” and the lowest “Co-op apartment”. The data of which are

output in *Figure 45*. The second table represents the analogous ANOVA test or Kruskal-Wallis test, which shows the chi-square approximation and corresponding *p*-value of the model.

The output from the Wilcoxon option shows the actual sums of the rank scores and the expected sums of the rank scores if the null hypothesis is true. From the Kruskal-Wallis test (chi-square approximation), the *p*-value is 0.0037. Therefore, at the 5% level of significance, the null hypothesis (H_0) is rejected. There is enough evidence to conclude that the distributions of change in annual income for the different ways of living are significantly different.

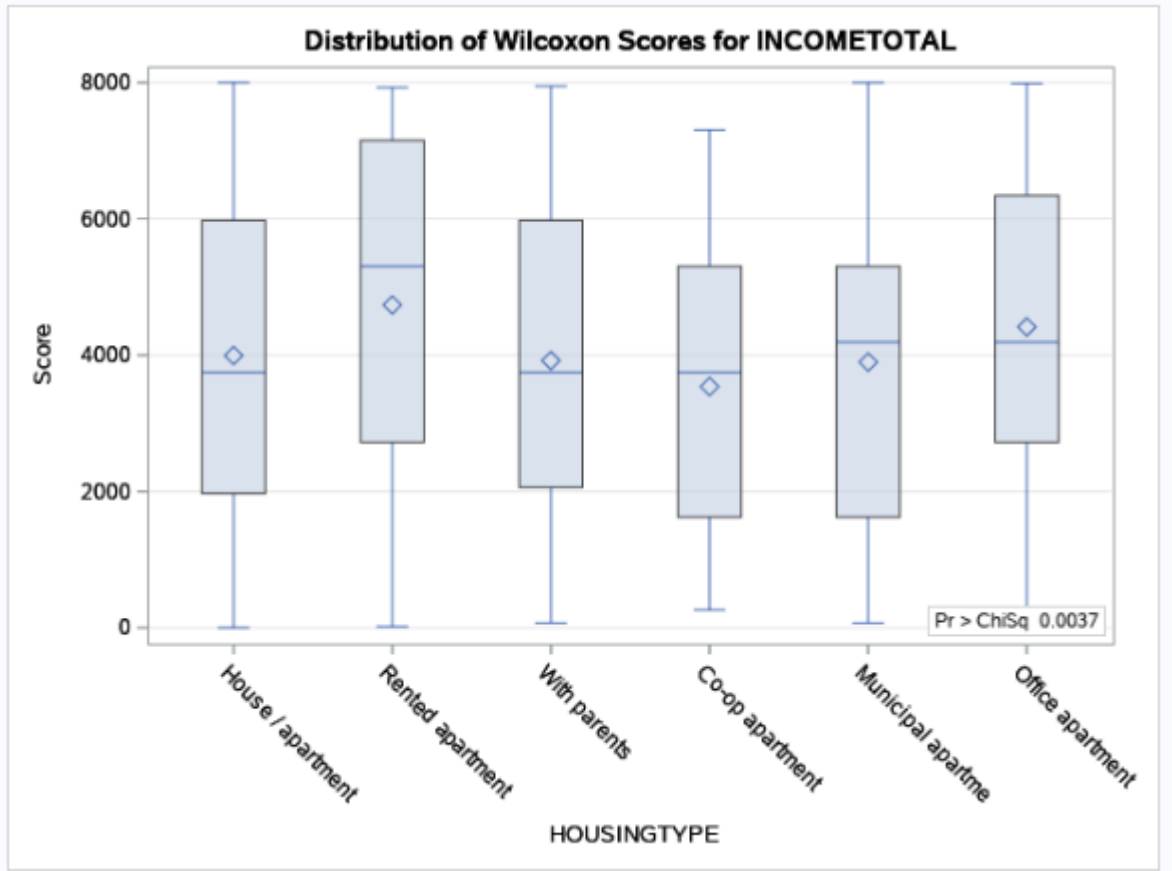


Figure 45. Nonparametric one-way ANOVA box-and-whisker plot based on the first ANOVA model.

Figure 45 shows the box-and-whisker plot of the mean scores graphically illustrates the difference between the groups of HOUSINGTYPE. Although the sample of all the groups have huge differences, the mean score of the groups is observed to revolve around the 4000 mark with maximum and minimum heights of around 8000 and 0 respectively.

First Nonparametric One-way ANOVA Kruskal-Wallis Test

Median Scores (Number of Points Above Median) for Variable INCOMETOTAL Classified by Variable HOUSINGTYPE

HOUSINGTYPE	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
House / apartment	7133	3545.92994	3566.50	13.712744	0.497116
Rented apartment	127	74.66879	63.50	5.513796	0.587943
With parents	392	197.24841	196.00	9.522632	0.503185
Co-op apartment	42	17.57325	21.00	3.187905	0.418411
Municipal apartme	248	132.00637	124.00	7.645598	0.532284
Office apartment	58	32.57325	29.00	3.742464	0.561608

Average scores were used for ties.

Median One-Way Analysis		
Chi-Square	DF	Pr > ChiSq
7.4152	5	0.1915

Figure 46. Nonparametric one-way ANOVA Median-score table and Median one-way-analysis table based on the first ANOVA model.

Next, *Figure 46* shows the score using the Median test and Median one-way analysis. The score in the Median test is an indicator of whether the value of the response variable (HOUSINGTYPE) is above or below the median for all observations. Similar to the Wilcoxon option test, the data with the highest median mean score is shown to be “Rented apartment” and the lowest “Co-op apartment”. The second table represents the median one-

way analysis, which shows the chi-square approximation and corresponding *p*-value of the model. The data of which are output in *Figure 47*.

In this case, performing the analysis using the Median test (chi-square approximation), the *p*-value is 0.1915. Therefore, at the 5% level of significance, the null hypothesis (H_0) is not rejected. There is not enough evidence to conclude that the distributions of change in annual income for the different ways of living are significantly different.

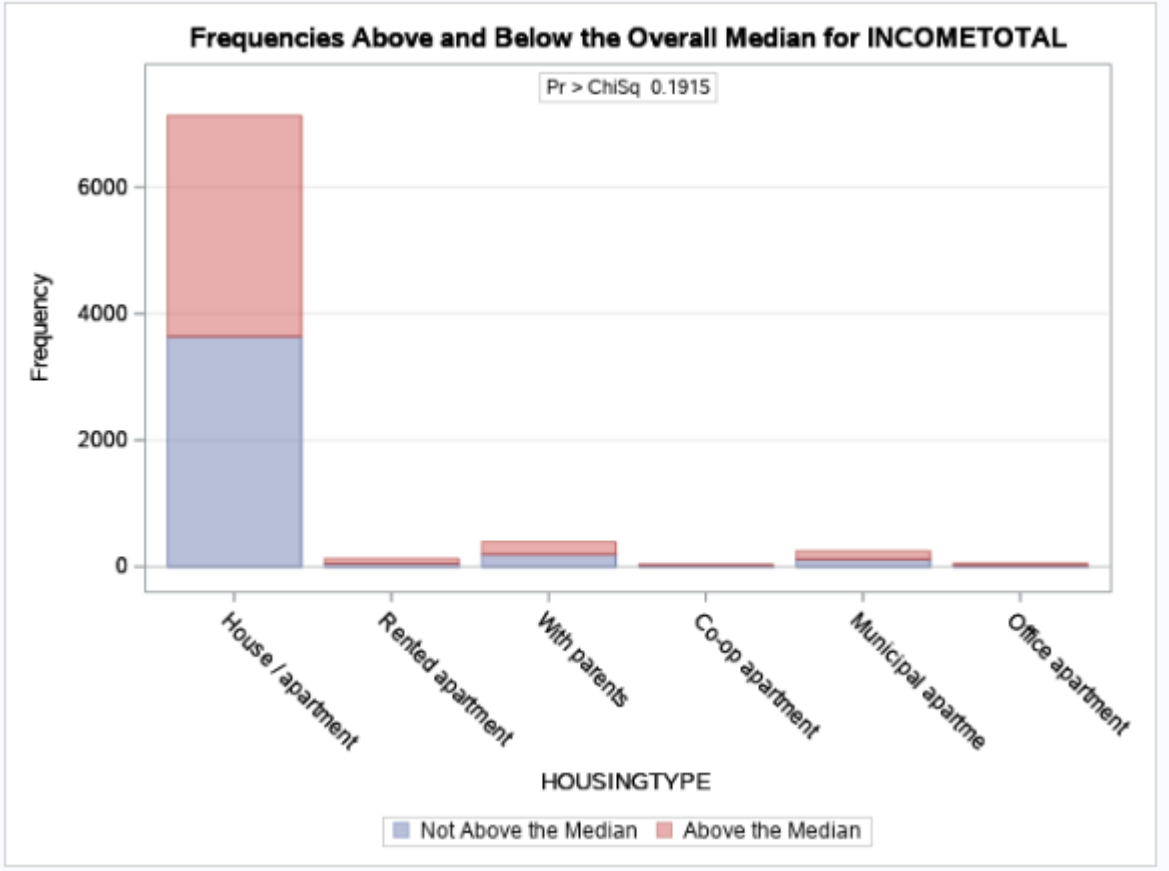


Figure 47. Nonparametric one-way ANOVA frequency plot based on the first ANOVA model.

Figure 47 shows a frequency plot that illustrates the statistics in the Median-score table. The big takeaway from the plot is the “House / apartment” data consisting of median frequency of

close to 4000 and close to 3000 for not above the median and above the median scores respectively.

Second Nonparametric One-way ANOVA

Similarly, the sample data from the second AVOVA did not meet any of the distribution assumptions of simple random samples, normality and homogeneity of variance. Hence, a nonparametric test one-way ANOVA test is performed based on the second ANOVA model to show more appropriate results.

Distribution Examination

Firstly, a distribution data analysis is performed to examine the distribution of data to determine the suitable analyses to be conducted. Based on the data, the plots conducted are histogram and Q-Q plot with the mean, standard deviation, skewness, kurtosis, normal test statistics, and normal test p-value.

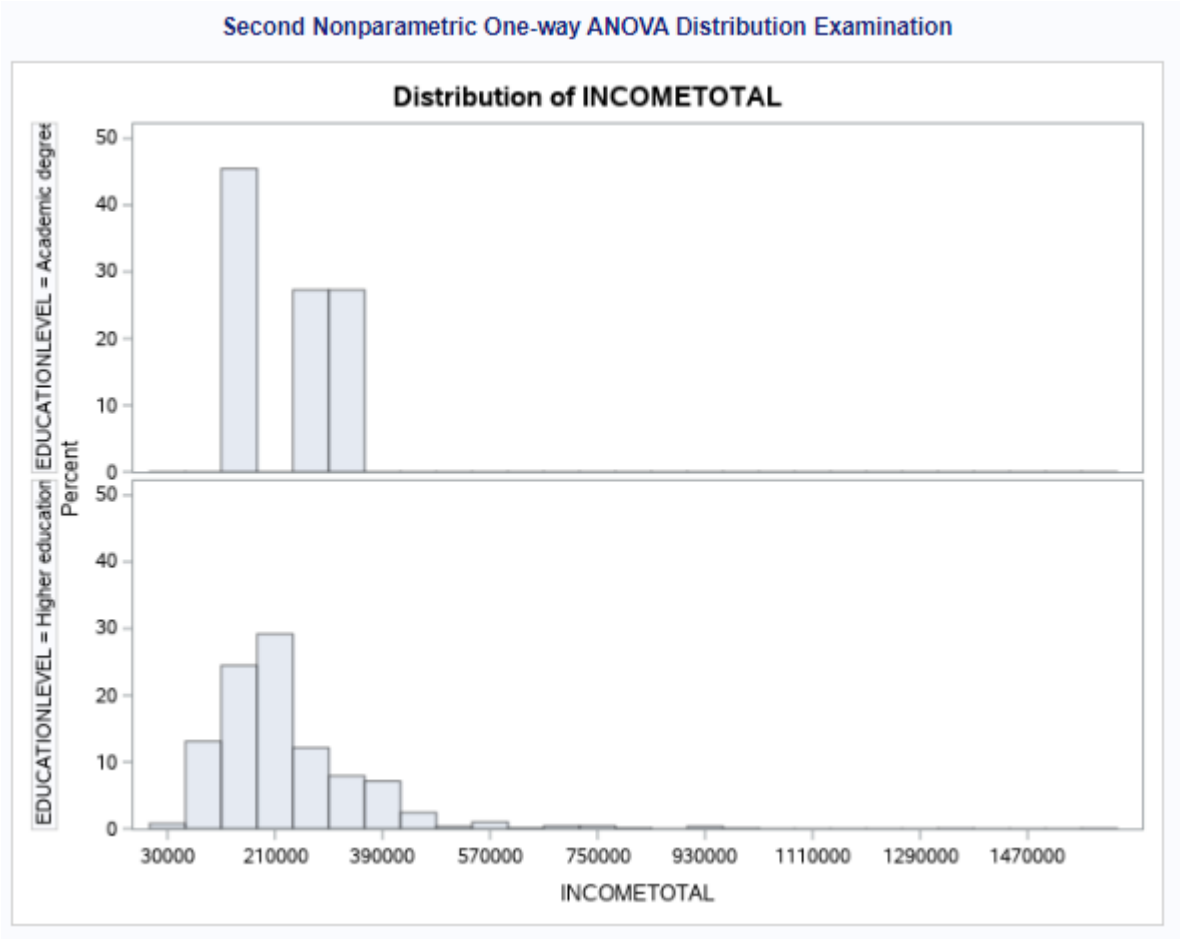
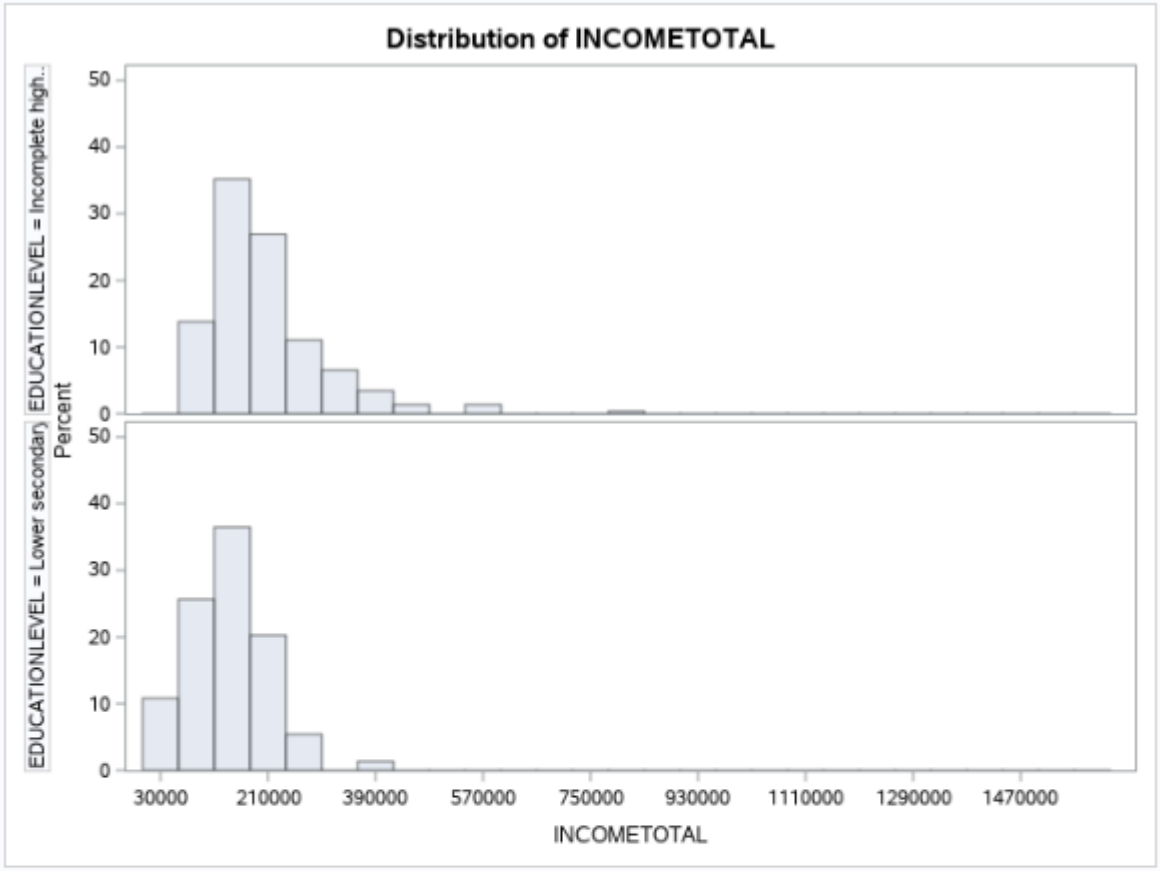


Figure 48. Nonparametric one-way ANOVA histogram distribution based on the second ANOVA model.



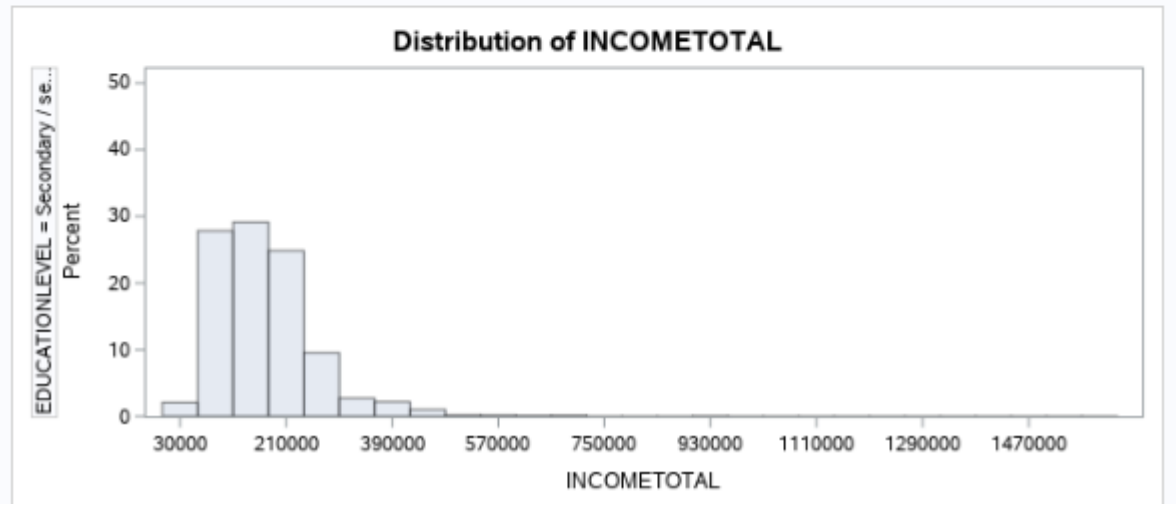


Figure 49. Nonparametric one-way ANOVA histogram distribution based on the second ANOVA model. (continued)

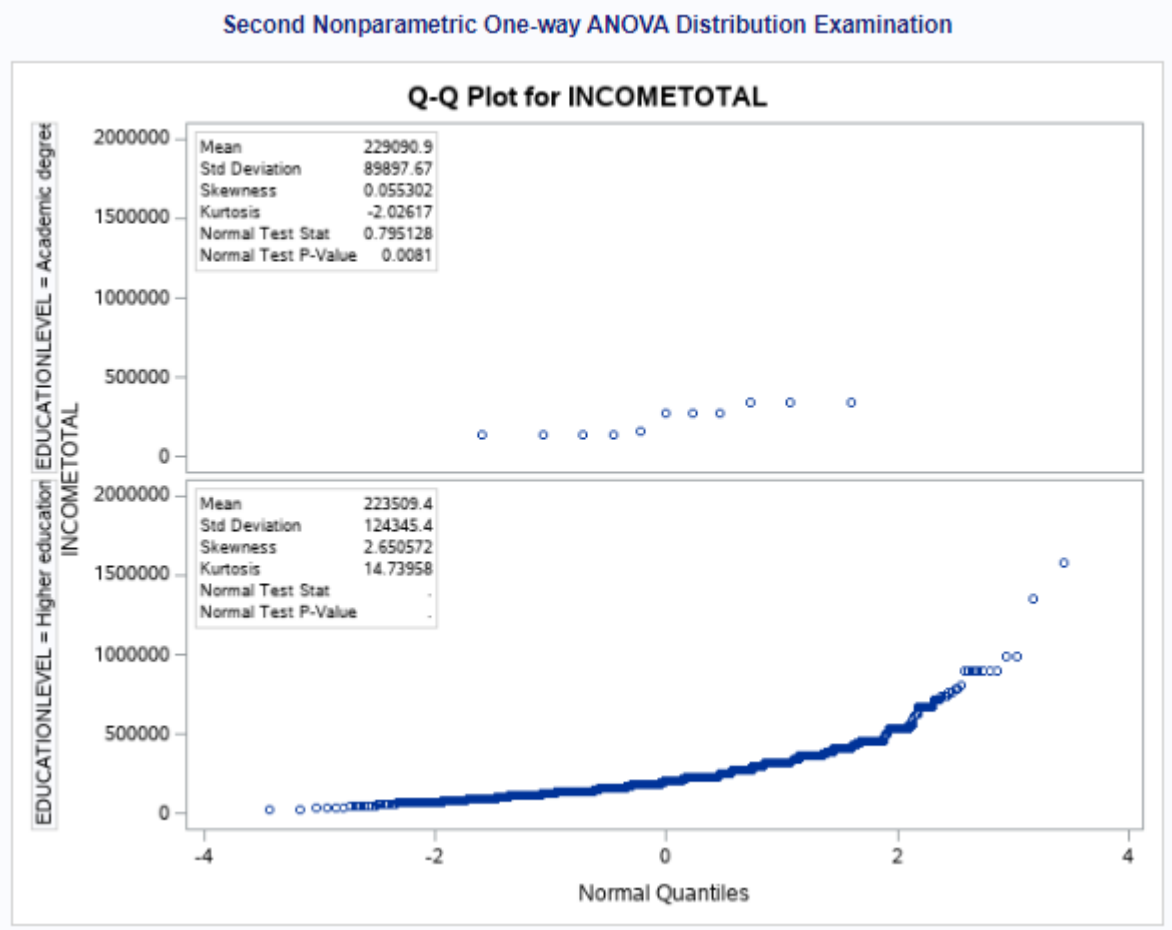


Figure 50. Nonparametric one-way ANOVA Q-Q plot distribution based on the second ANOVA model.

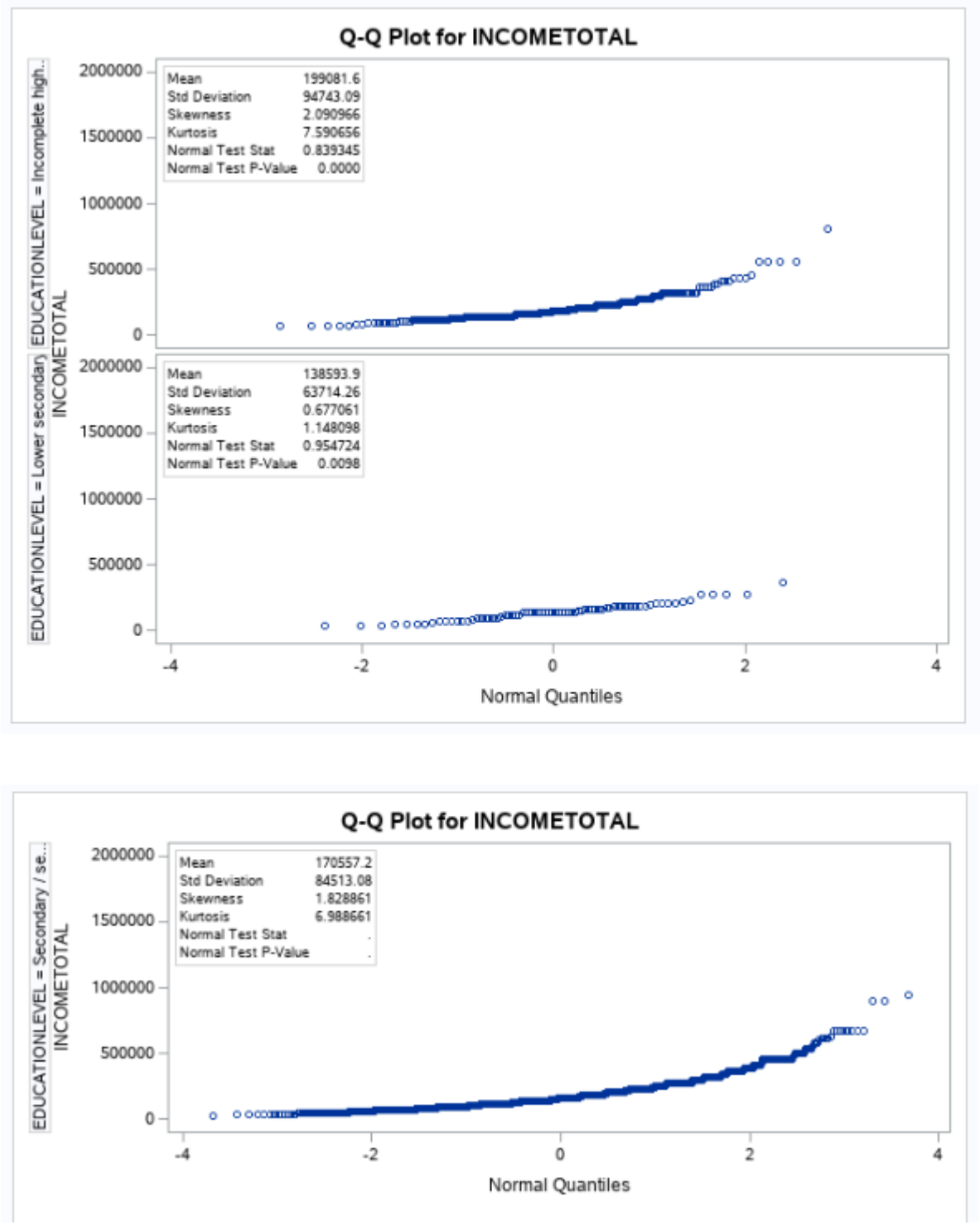


Figure 51. Nonparametric one-way ANOVA Q-Q plot distribution based on the second ANOVA model. (continued)

Based on the plots and values of skewness and kurtosis, the data do not appear to be normally distributed for any speciality code. The goodness-of-fit tests reject the null hypothesis that the data are normally distributed.

Kruskal-Wallis Test

The parametric equivalent of ANOVA is the Kruskal-Wallis test. Similarly, the purpose of the Kruskal-Wallis test is to test the null hypothesis (H_0) that the independent samples come from populations, and determine if there are statistically significant differences between two or more groups of an independent variable (EDUCATIONLEVEL) on a continuous dependent variable (INCOMETOTAL).

H_0: The samples come from populations with equal medians.

H_1: The samples come from populations with medians that are not all equal.

For illustrative purposes, the Wilcoxon option is used to perform a Rank Sum test and the Median option to perform the Median test which uses the ranks of sample data from INCOMETOTAL classified by EDUCATIONLEVEL. The data is visualized in *Figure 52*.

Second Nonparametric One-way ANOVA Kruskal-Wallis Test

Wilcoxon Scores (Rank Sums) for Variable INCOMETOTAL
Classified by Variable EDUCATIONLEVEL

EDUCATIONLEVEL	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Secondary / secondary special	5459	19952530.5	21838729.5	95972.3592	3654.97903
Higher education	2166	10492380.0	8665083.0	91600.8977	4844.12742
Incomplete higher	290	1289817.5	1160145.0	38531.2838	4447.64655
Lower secondary	74	211860.0	296037.0	19734.6844	2862.97297
Academic degree	11	57412.0	44005.5	7638.8828	5219.27273

Average scores were used for ties.

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
444.9020	4	<.0001

Figure 52. Nonparametric one-way ANOVA Wilcoxon-score table and Kruskal-Wallis test table based on the second ANOVA model.

Based on the output in *Figure 52*, the Wilcoxon-score table shows the mean rank score for each of the experimental groups of EDUCATIONLEVEL. The data with the highest mean score is shown to be “Academic degree” and the lowest “Lower secondary”. The second table represents the analogous ANOVA test or Kruskal-Wallis test, which shows the chi-square approximation and corresponding *p*-value of the model. The data of which are output in *Figure 53.*

The output from the Wilcoxon option shows the actual sums of the rank scores and the expected sums of the rank scores if the null hypothesis is true. From the Kruskal-Wallis test (chi-square approximation), the *p*-value is <.0001. Therefore, at the 5% level of significance, the null hypothesis (H_0) is rejected. There is enough evidence to conclude that the distributions of change in annual income for the different education levels are significantly different.

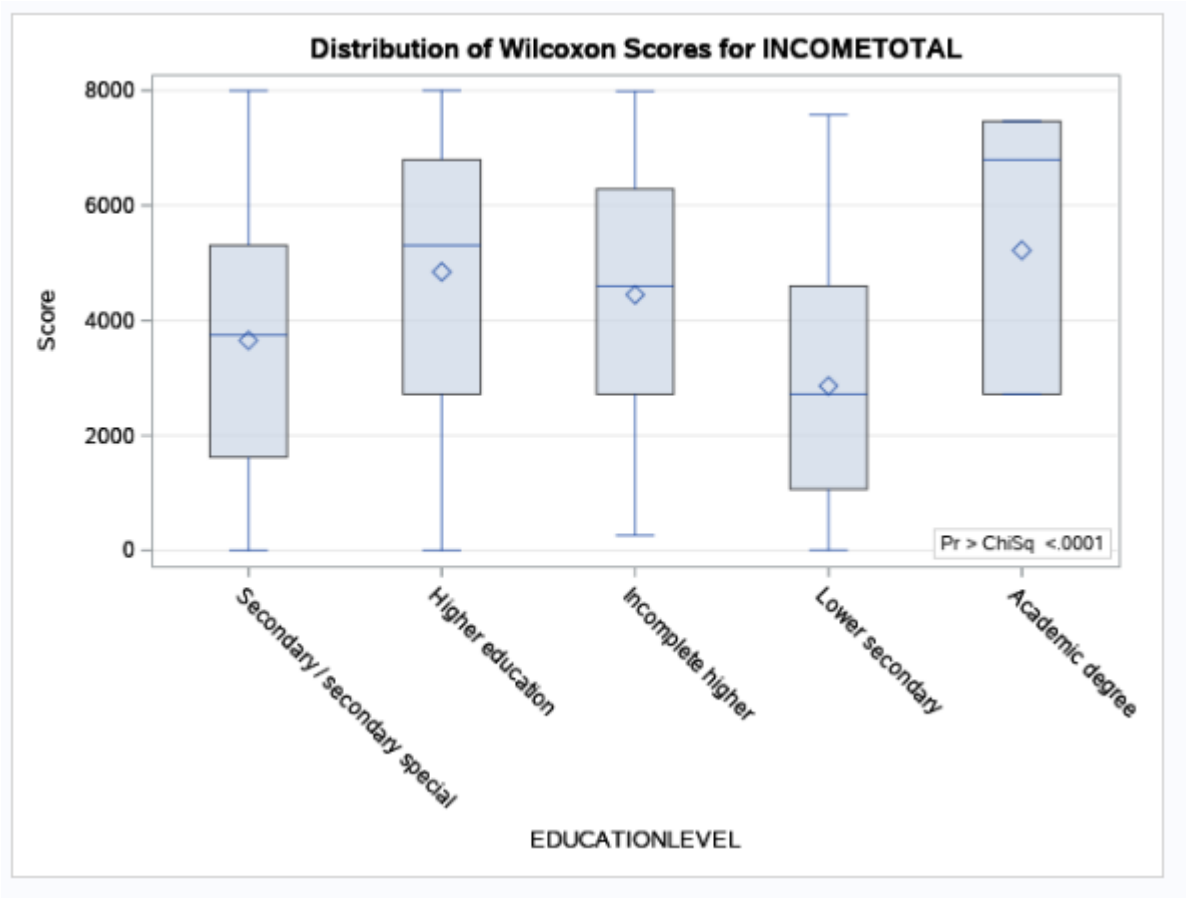


Figure 53. Nonparametric one-way ANOVA box-and-whisker plot based on the second ANOVA model.

Figure 53 shows the box-and-whisker plot of the mean scores graphically illustrates the difference between the groups of EDUCATIONLEVEL. Although the sample of all the groups have huge differences, the mean score of the groups is observed to vary with maximum and minimum heights of around 8000 and 0 respectively except for the “Academic degree” value.

Second Nonparametric One-way ANOVA Kruskal-Wallis Test

Median Scores (Number of Points Above Median) for Variable INCOMETOTAL Classified by Variable EDUCATIONLEVEL

EDUCATIONLEVEL	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Secondary / secondary special	5459	2399.60510	2729.50	20.537028	0.439569
Higher education	2166	1405.24204	1083.00	19.601584	0.648773
Incomplete higher	290	165.57962	145.00	8.245271	0.570964
Lower secondary	74	23.47771	37.00	4.223005	0.317266
Academic degree	11	6.09554	5.50	1.634637	0.554140

Average scores were used for ties.

Median One-Way Analysis

Chi-Square	DF	Pr > ChiSq
295.3398	4	<.0001

Figure 54. Nonparametric one-way ANOVA Median-score table and Median one-way-analysis table based on the second ANOVA model.

Next, *Figure 54* shows the score using the Median test and Median one-way analysis. The score in the Median test is an indicator of whether the value of the response variable (EDUCATIONLEVEL) is above or below the median for all observations. Not similar to the Wilcoxon option test, the data with the highest median mean score is shown to be “Higher education” and the lowest “Lower secondary”. The second table represents the median one-way analysis, which shows the chi-square approximation and corresponding *p*-value of the model. Likewise, the data is visualized in *Figure 55*.

In this case, performing the analysis using the Median test (chi-square approximation), the *p*-value is <.0001. Therefore, at the 5% level of significance, the null hypothesis (H_0) is rejected. There is enough evidence to conclude that the distributions of change in annual income for the different education levels are significantly different.

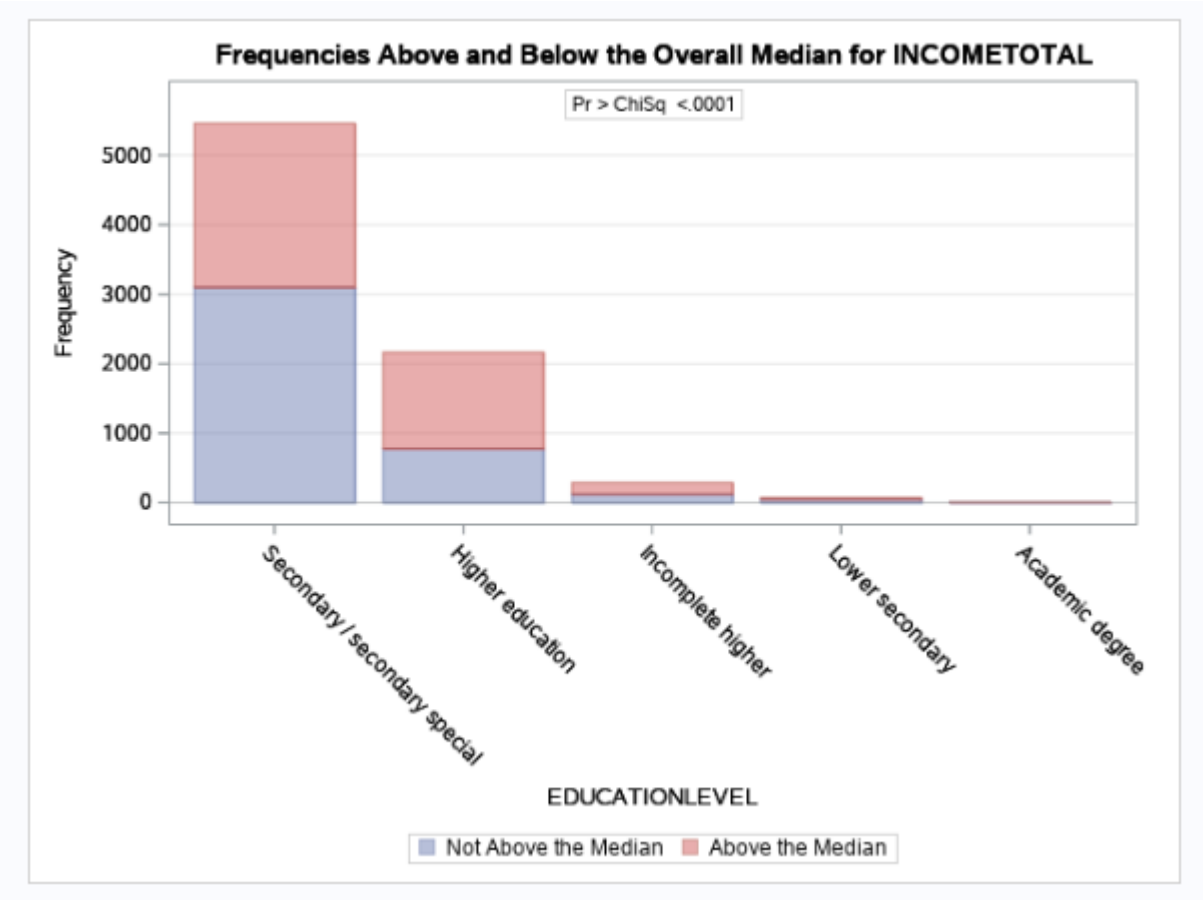


Figure 55. Nonparametric one-way ANOVA frequency plot based on the second ANOVA model.

Figure 55 shows a frequency plot that illustrates the statistics in the Median-score table. The big takeaway from the plot is the "Secondary / secondary special" data consisting of median frequency of close to 3000 and close to 2000 for not above the median and above the median scores respectively. Also, the "Higher education" data stands out with median frequency of close to 1000 and close to 1500 for not above the median and above the median scores respectively.

Conclusion

In this assignment, two one-way ANOVA was performed to determine whether the average annual income is significantly different for various ways of living (HOUSINGTYPE) and education level (EDUCATIONLEVEL). After that, blocking was performed to both of the models to see whether adding a controlled variable improves the model design. Lastly, since the assumptions of both models are not all met, two nonparametric tests were conducted for both of the models to help treat the problem and show better and more appropriate results.

In conclusion, it is determined that different factors are related to the customer's annual income (INCOMETOTAL) where it shows there is enough evidence to conclude that the distribution of change in annual income (INCOMETOTAL) for different education levels (EDUCATIONLEVEL) are significantly different while ways of living (HOUSINGTYPE) are not significantly different. Bankers should make an effort to take initiatives in the education level (EDUCATIONLEVEL) instead focusing on the ways of living (HOUSINGTYPE).

References

Kleinman, K. (2012). Example 9.36: Levene's test for equal variances. Retrieved 23 July 2020, from http://sas-and-r.blogspot.com/2012/06/example-936-levenes-test-for-

equal.html#:~:text=June%2025%2C%202012-
,Example%209.36%3A%20Levene's%20test%20for%20equal%20variances,be%20tested%20via%20Levene's%20test.

The GLM Procedure. (2012). Retrieved 23 July 2020, from https://support.sas.com/documentation//cdl/en/statug/68162/HTML/default/viewer.htm#statug_glm_syntax10.htm

Yankovsky, A. (2015). Let’s explore SAS Proc T-Test. Retrieved 23 July 2020, from https://www.sas.com/content/dam/SAS/en_ca/User%20Group%20Presentations/Calgary-User-Group/Yankovsky-ExploringProcTtest-Apr2015.pdf

Appendix A

SAS Studio Code

```

/* Generated Code (IMPORT) */
/* Source File: assignment-individual-data.csv */
/* Source Path: /home/u42888972/BSBA Sem 6 Applied Statistics/AS Indi Assignment */
/* Code generated on: 21/7/2020 */

/*
I start by uploading the assignment-individual-data.csv file to my folder (AS Group Assignment, stated at the Source Path).
Then I created a permanent library called ASINCW20 and path it to the folder.
*/


libname ASINCW20 '/home/u42888972/BSBA Sem 6 Applied Statistics/AS Indi Assignment';


%web_drop_table(ASINCW20.ASINDT20);

FILENAME REFFILE '/home/u42888972/BSBA Sem 6 Applied Statistics/AS Indi Assignment/assignment-individual-data.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=ASINCW20.ASINDT20;
    GETNAMES=YES;
RUN;

%web_open_table(ASINCW20.ASINDT20);


*Created Macro for easier data manipulation;
%let library = ASINCW20;
%let filename = ASINDT20;
```

```
/** 1. BASIC DATA EXPLORATION & DATA CLEANING/VALIDATION. **/

/*
Using PROC FORMAT AND FREQ to find out which variables have missing data.
(I did not use PROC MEAN because it can only be use for num variables.)
*/

PROC CONTENTS DATA = &library..&filename;
RUN;


PROC FORMAT;
    VALUE $missing_char
        ' ' = 'Missing'
        other = 'Present';

    VALUE missing_num
        . = 'Missing'
        other = 'Present';
RUN;


TITLE 'Listing of Present and Missing Data for Each Variable';
PROC FREQ DATA = &library..&filename;
    TABLES _all_ / missing;
    FORMAT _character_ $missing_char. _numeric_ missing_num.;
RUN;
TITLE;


options nolabel;
PROC MEANS DATA = &library..&filename N NMISS MIN MAX MEAN;
RUN;
```

```
72 *-------------------------------------------------------------------------------------------;
73 /** 2. DESCRIPTIVE ANALYSIS. **/
74
75 *Figure 1. Descriptive Statistics for assignment data.;
76 proc means data=ASINCW20.ASINDT20 chartype n nmiss min max mode median mean std vardef=df;
77     var CHILDRENCOUNT INCOMETOTAL FAMSIZE;
78 run;
79
80
81
82 *Bar Chart1;
83 *Figure 2. Frequency of customer's education level by gender.;
84 ods graphics / reset width=6.4in height=4.8in imagemap;
85
86 proc sgplot data=ASINCW20.ASINDT20;
87     vbar GENDER / group=EDUCATIONLEVEL groupdisplay=cluster;
88     yaxis grid;
89 run;
90
91 ods graphics / reset;
92
93
94
95 *Pie Chart;
96 *Figure 3. Percentage of customers who own property by gender.;
97 proc template;
98     define statgraph SASStudio.Pie;
99         begingraph;
100        layout region;
101        piechart category=GENDER / group=OWNPROPERTY groupgap=2%
102            datalabellocation=inside;
103        endlayout;
104        endgraph;
105    end;
106 run;
107
108 ods graphics / reset width=6.4in height=4.8in imagemap;
109
110 proc sgrender template=SASStudio.Pie data=ASINCW20.ASINDT20;
111 run;
112
113 ods graphics / reset;
```

```
*Bar Chart 2;
*Figure 4. Percentage customer's marital status by gender.;
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=ASINCW20.ASINDT20;
    hbar GENDER / group=MARITALSTATUS groupdisplay=cluster stat=percent;
    xaxis grid;
run;

ods graphics / reset;



*Histogram;
*Figure 5. Distribution of annual income by gender.;
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sort data=ASINCW20.ASINDT20 out=_HistogramTaskData;
    by GENDER;
run;

proc sgplot data=_HistogramTaskData;
    by GENDER;
    title height=14pt "Distribution of IncomeTotal";
    histogram INCOMETOTAL / fillattrs=(color=CX3b4556);
    density INCOMETOTAL;
    yaxis grid;
run;

ods graphics / reset;
title;

proc datasets library=WORK noprint;
    delete _HistogramTaskData;
    run;
```

```
*Box Plot;
*Figure 6. Frequency of customer's annual income by ways of living ordered by gender.;
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=ASINCW20.ASINDT20;
    vbox INCOMETOTAL / category=HOUSINGTYPE group=GENDER;
    yaxis grid;
run;

ods graphics / reset;



*Bubble Plot;
*Figure 7. Frequency of customer's family size by credit loan status controlling for income category and number of children.;
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=ASINCW20.ASINDT20;
    bubble x=CREDITSTATUS y=FAMSIZE size=CHILDRENCOUNT/ group=INCOMETYPE
        bradiusmin=7 bradiusmax=14;
    xaxis grid;
    yaxis grid;
run;

ods graphics / reset;
```

```
186 *-------------------------------------------------------------------------------------------;
187 /** 3. Analysis of Variance (ANOVA) **/
188 *This section look into 2 different ANOVA;
189
190 *First Linear Models Task: One-way ANOVA;
191 /*
192     HOUSINGTYPE = Categorical variable
193     INCOMETOTAL = Dependent variable
194 */
195 TITLE 'First One-Way ANOVA';
196 ods graphics on;
197 proc glm data = &library..&filename plots (maxpoints = none) plots = (residuals diagnostics);
198     class HOUSINGTYPE;
199     model INCOMETOTAL = HOUSINGTYPE;
200     lsmeans HOUSINGTYPE / adjust=tukey pdiff alpha=.05;
201     means HOUSINGTYPE / hovtest=levene;
202 run;
203 ods graphics off;
204 title;
205
206
207 *First ANOVA with Blocking;
208 *First ANOVA with Data from a Randomized Block Design;
209 /*
210     HOUSINGTYPE = Categorical variable
211     INCOMETOTAL = Dependent variable
212     CREDITSTATUS = Blocking variable
213 */
214 TITLE 'First One-Way ANOVA with Blocking';
215 ods graphics on;
216 proc glm data = &library..&filename plots (maxpoints = none) plots = (residuals diagnostics);
217     class CREDITSTATUS HOUSINGTYPE;
218     model INCOMETOTAL = CREDITSTATUS HOUSINGTYPE;
219 run;
220 ods graphics off;
221 title;
```

```
224 *First ANOVA Post Hoc Pairwise Comparisons;
225 /*
226     HOUSINGTYPE = Categorical variable
227     INCOMETOTAL = Dependent variable
228     CREDITSTATUS = Blocking variable
229 */
230 TITLE 'First One-Way ANOVA with Blocking and Post-Hoc Pairwise Comparisons';
231 ods graphics on;
232 proc glm data = &library..&filename plots (maxpoints = none) plots = (residuals diagnostics);
233     class CREDITSTATUS HOUSINGTYPE;
234     model INCOMETOTAL = CREDITSTATUS HOUSINGTYPE;
235     lsmeans HOUSINGTYPE / adjust=tukey pdiff alpha=.05;
236     lsmeans HOUSINGTYPE / pdiff = control ('Rented apartment'); *Dunnett;
237     lsmeans HOUSINGTYPE / adjust = t;
238 run;
239 ods graphics off;
240 title;
```

```
*Second Linear Models Task: One-way ANOVA;
/*
    EDUCATIONLEVEL = Categorical variable
    INCOMETOTAL = Dependent variable
*/
TITLE 'Second One-Way ANOVA';
ods graphics on;
proc glm data = &library..&filename plots (maxpoints = none) plots = (residuals diagnostics);
    class EDUCATIONLEVEL;
    model INCOMETOTAL = EDUCATIONLEVEL;
    lsmeans EDUCATIONLEVEL / adjust=tukey pdiff alpha=.05;
    means EDUCATIONLEVEL / hovtest=levene;
run;
ods graphics off;
title;


*Second ANOVA with Blocking;
*Second ANOVA with Data from a Randomized Block Design;
/*
    EDUCATIONLEVEL = Categorical variable
    INCOMETOTAL = Dependent variable
    HOUSINGTYPE = Blocking variable
*/
TITLE 'Second One-Way ANOVA with Blocking';
ods graphics on;
proc glm data = &library..&filename plots (maxpoints = none) plots = (residuals diagnostics);
    class HOUSINGTYPE EDUCATIONLEVEL;
    model INCOMETOTAL = HOUSINGTYPE EDUCATIONLEVEL;
run;
ods graphics off;
title;

*Second ANOVA Post Hoc Pairwise Comparisons;
/*
    EDUCATIONLEVEL = Categorical variable
    INCOMETOTAL = Dependent variable
    HOUSINGTYPE = Blocking variable
*/
TITLE 'Second One-Way ANOVA with Blocking and Post-Hoc Pairwise Comparisons';
ods graphics on;
proc glm data = &library..&filename plots (maxpoints = none) plots = (residuals diagnostics);
    class HOUSINGTYPE EDUCATIONLEVEL;
    model INCOMETOTAL = HOUSINGTYPE EDUCATIONLEVEL;
    lsmeans EDUCATIONLEVEL / adjust=tukey pdiff alpha=.05;
    lsmeans EDUCATIONLEVEL / pdiff = control ('Lower secondary'); *Dunnett;
    lsmeans EDUCATIONLEVEL / adjust = t;
run;
ods graphics off;
title;
```

```
298 *-------------------------------------------------------------------------------------------;
299 /** 4. Nonparametric One-Way ANOVA **/
300
301 *First Nonparametric One-way ANOVA;
302 /*
303     HOUSINGTYPE = Categorical variable
304     INCOMETOTAL = Dependent variable
305 */
306 *First Distribution Examination;
307 title 'First Nonparametric One-way ANOVA Distribution Examination';
308 ods graphics on;
309 proc univariate data=&library..&filename normal;
310     class HOUSINGTYPE;
311     var INCOMETOTAL;
312     histogram INCOMETOTAL;
313     qqplot INCOMETOTAL;
314     inset mean std skewness kurtosis normaltest probn;
315 run;
316 title;
317
318
319 *First Kruskal-Wallis Test;
320 title 'First Nonparametric One-way ANOVA Kruskal-Wallis Test';
321 ods noproctitle;
322 proc npar1way data=&library..&filename wilcoxon median plots(only)=(wilcoxonboxplot medianplot);
323     class HOUSINGTYPE;
324     var INCOMETOTAL;
325 run;
```

```
331 *Second Nonparametric One-way ANOVA;
332 /*
333     EDUCATIONLEVEL = Categorical variable
334     INCOMETOTAL = Dependent variable
335 */
336 *Second Distribution Examination;
337 title 'Second Nonparametric One-way ANOVA Distribution Examination';
338 ods graphics on;
339 proc univariate data=&library..&filename normal;
340     class EDUCATIONLEVEL;
341     var INCOMETOTAL;
342     histogram INCOMETOTAL;
343     qqplot INCOMETOTAL;
344     inset mean std skewness kurtosis normaltest probn;
345 run;
346
347
348 *Second Kruskal-Wallis Test;
349 title 'Second Nonparametric One-way ANOVA Kruskal-Wallis Test';
350 ods noproctitle;
351 proc npar1way data=&library..&filename wilcoxon median plots(only)=(wilcoxonboxplot medianplot);
352     class EDUCATIONLEVEL;
353     var INCOMETOTAL;
354 run;
```

Appendix B

Marking Rubric

Marking Rubric

Marks (Weightage)	Unsatisfactory	Satisfactory	Good	Excellent	Comments
Results Output (40%) [SLO 2]	Graphs and descriptive summary statistics are not provided, or provided but inappropriate. Dataset is analysed using ONE appropriate statistical technique.	Graphs and descriptive summary statistics are provided but some are inappropriate. Dataset is analysed using ONE appropriate statistical technique.	Appropriate graphs and descriptive summary statistics are provided. Dataset is analysed using TWO appropriate statistical techniques.	Appropriate graphs and descriptive summary statistics are provided. Dataset is analysed using MORE THAN TWO appropriate statistical techniques.	
Interpretation [SLO 2] (30%)	No interpretation on the outputs.	Minimum interpretation (descriptive in nature) on the outputs.	Sufficient interpretation (mostly descriptive, with minimum relevance to the context of the data) on the outputs.	Thorough analysis and interpretation on the outputs with relevance to the context of the data.	
Formatting and Language [SLO 2] (10%)	Formatting is inconsistent with numerous grammatical and spelling errors.	Formatting is consistent with some grammatical and spelling errors.	Formatting is consistent with minimal grammatical and spelling errors.	Formatting is consistent with no grammatical and spelling errors.	
Use of SAS software (20%) [SLO 3]	SAS code/ tasks are not provided.	SAS code / tasks are provided but incomplete.	SAS code / tasks are provided for most of the outputs shown.	SAS code / tasks are provided for all outputs shown.	