1. Introduction

The objectives that has been set for this statistical analysis are as follows:

- To determine the whether the variables GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE, CREDITSTATUS that have a significant effect on the response variable INCOMETOTAL.
- To determine the whether the variables GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, INCOMETOTAL, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION and FAMSIZE that have a significant effect on the response variable CREDITSTATUS.
- To determine how variable GENDER is associated to variable OWNCAR and OWNPROPERTY.

As part of the statistical analysis, I have employed Analysis of Covariance, Logistic Regression and Categorical Data Analysis – Table Analysis for the objectives respectively.

For the **Analysis of Covariance**, variable INCOMETOTAL has been identified as the Response Variable while variables GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE, CREDITSTATUS has been identified as the Predictor variables.

Since, the Response Variable Identified is a Continuous Variable and the Predictor Variables identified are a mix of Categorical and Continuous variables, the Analysis of Covariance has been employed.

For the **Logistic Regression**, variable CREDITSCORE has been identified as the Response Variable while variables GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, INCOMETOTAL, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE, CREDITSTATUS has been identified as the Predictor Variable.

Since, the Response Variable Identified is a Categorical Variable and the Predictor Variables identified are a mix of Categorical and Continuous variables, the Logistic Regression has been employed.

As for the final statistical analysis, **Categorical Data Analysis – Table Analysis**, the identified Response Variable is GENDER while the Predictor Variables are OWNCAR and OWNPROPERTY.

The Table Analysis method is used as the statistical set contains two categorical variables. Classification of both the variables are also have 2 possible values.

In all the statistical analysis, variable ID has been removed as it does not carry any meaningful value. The variable ID is a unique number that is assigned to the customers of the bank. Hence, it cannot be computed, and the trends cannot be analysed properly.

2. Descriptive Analysis

As part of my descriptive analysis for each variable, the categorical variables are analysed with a Pie Chart to best represent the data while the continuous variables are analysed using the UNIVARIATE Procedure.

The **measure of central tendency** refers to a summary statistic that is used to represent the centre point of a dataset (Narkhede, 2018). Both the mean and median indicates the centre of the data. However, the median is less affected by the outlier as compared to the mean. (Minitab, 2019)

As for the **measure of dispersion**, this measures the variability within the data (Narkhede, 2018). The range is basically the difference between the largest values and the smallest values. The quartile range is calculated by subtracting the 3rd quartile value by the first quartile value.

Skewness measures the asymmetry of the probability distribution about its mean. (Narkhede, 2018). Kurtosis is the measure of whether the data contains an abundance or lack of outliers relative to a normal distribution (Narkhede, 2018).

In order to optimise the statistical analysis process, I have substituted several values within the dataset to be represented by a numerical value. The substitutions are explained in the next section. The pie charts are the representation of the organic data before any changes was made.

VARIABLES

GENDER

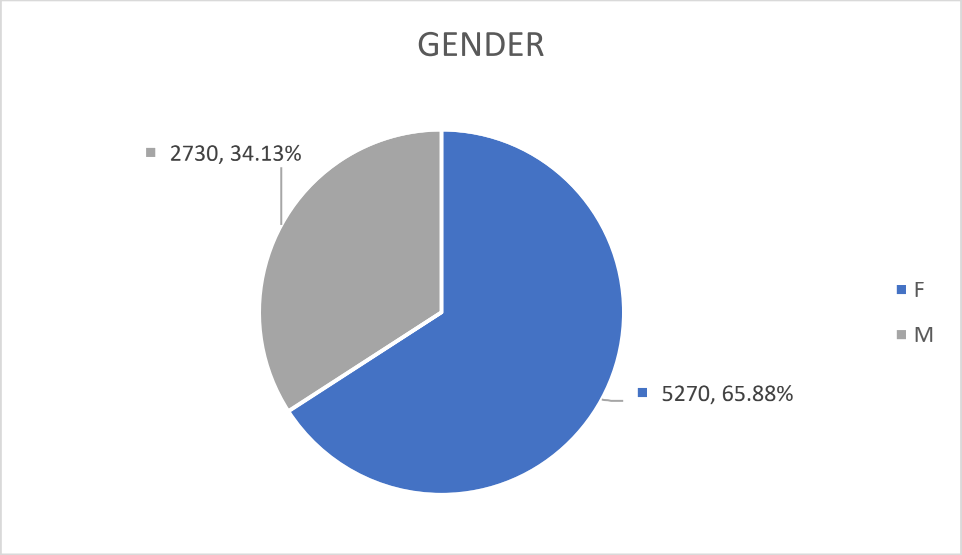


FIGURE 1. Pie chart representation of variable GENDER

The values that have been replaced are as follows:

- M = 0
- F = 1

A pie chart was generated for variable 'GENDER'.
It is seen that a total of 2730 (34.13%) of the customers are males and 5270 (65.88%) are females.
It is shown that females represent more than 50% of the dataset.

OWNCAR

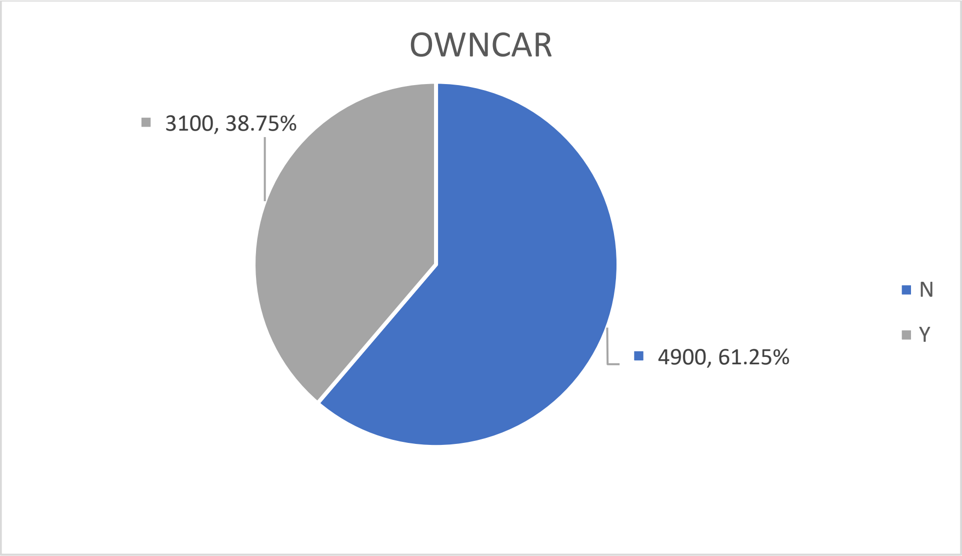


FIGURE 2. Pie chart representation of variable OWNCAR

The values that have been replaced are as follows:

- N = 0
- Y = 1

A pie chart was generated for variable ‘OWNCAR.
It is seen that a total of 3100 (38.75%) of the customers own a car while 4900 (61.25%) do not.
It is shown that more than half of the customers in the dataset do not own a car.

OWNPROPERTY

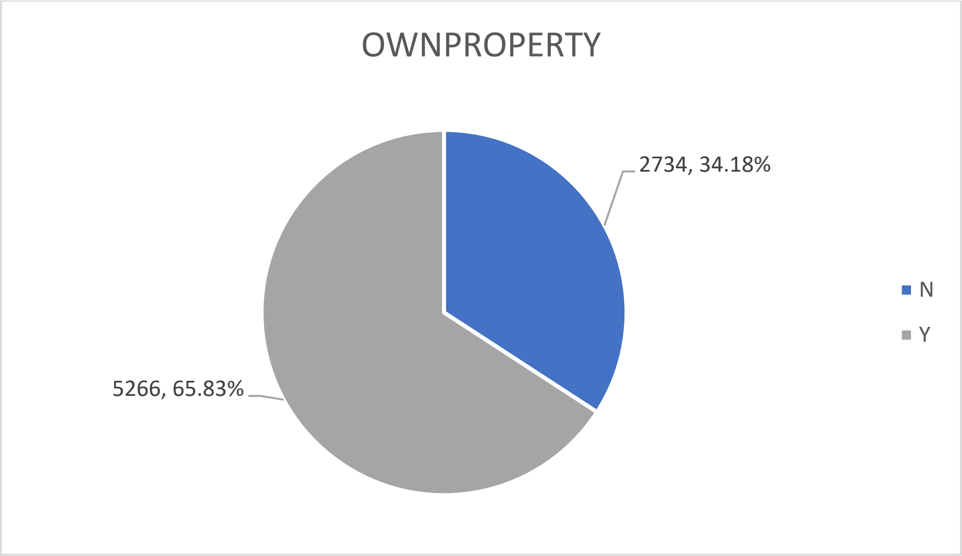


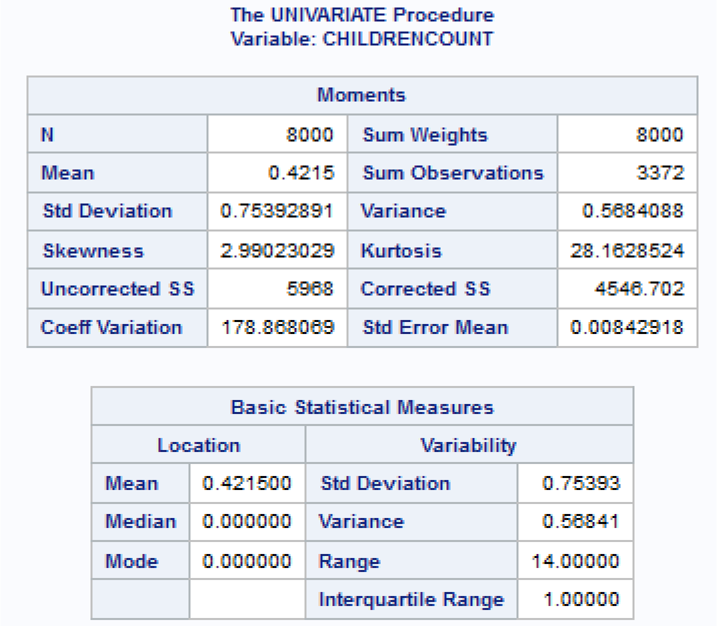
FIGURE 3. Pie chart representation of variable OWNPROPERTY

The values that have been replaced are as follows:

- N = 0
- Y = 1

A pie chart was generated for variable ‘OWNPROPERTY.
It is seen that a total of 5266 (65.83%) of the customers own a property while 2734 (34.18%) do not.
It is shown that more than half of the customers in the dataset do not own a property.

CHILDRENCOUNT



The UNIVARIATE Procedure
Variable: CHILDRENCOUNT

Moments			
N	8000	Sum Weights	8000
Mean	0.4215	Sum Observations	3372
Std Deviation	0.75392891	Variance	0.5684088
Skewness	2.99023029	Kurtosis	28.1628524
Uncorrected SS	5968	Corrected SS	4546.702
Coeff Variation	178.868069	Std Error Mean	0.00842918

Basic Statistical Measures			
Location		Variability	
Mean	0.421500	Std Deviation	0.75393
Median	0.000000	Variance	0.56841
Mode	0.000000	Range	14.00000
		Interquartile Range	1.00000

FIGURE 4. Univariate analysis representation of variable CHILDRENCOUNT

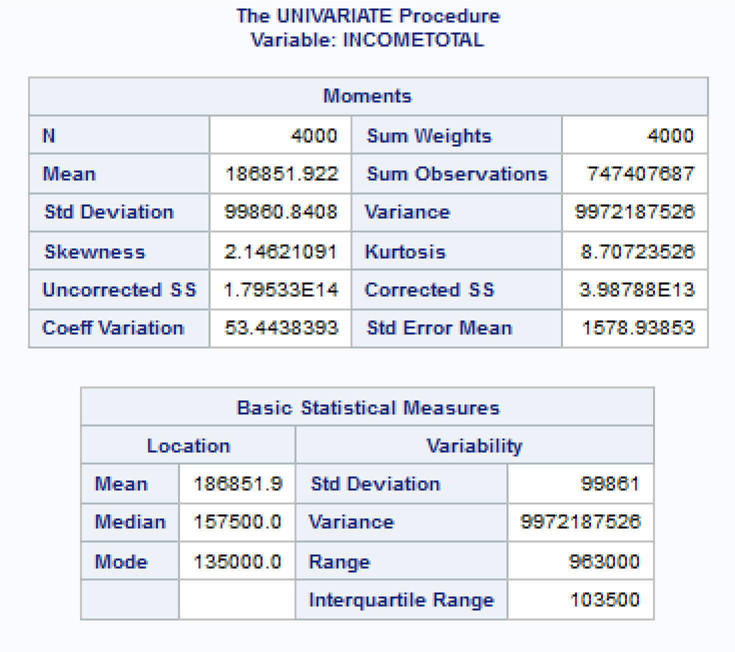
Measure of Central Tendency

- The **mean** is 0.4215
    - On average, a customer has 1 (rounded up from 0.415) children.
- The **median** mark is 0
- The smallest **mode** is 0
    - The value 0 has the highest frequency in the CHILDRENCOUNT variable.

Measure of Dispersion

- The **standard deviation** is 0.7539
- The **variance** is 0.5684
- The **range** is 14
- The **interquartile range** is 1
- The **skewness** value is 2.9902
    - The positive values indicate that it is positively skewed.
    - The skewness value can be said to be moderate, hence its difference from the normal distribution is moderate.
- The **Kurtosis** value is 28.1629
    - The positive value indicates that the distribution has thick and heavy tails and a sharper peak than the normal distribution.
    - This is called Leptokurtic.

INCOMETABLE



The UNIVARIATE Procedure
Variable: INCOMETOTAL

Moments			
N	4000	Sum Weights	4000
Mean	186851.922	Sum Observations	747407687
Std Deviation	99860.8408	Variance	9972187526
Skewness	2.14621091	Kurtosis	8.70723526
Uncorrected SS	1.79533E14	Corrected SS	3.98788E13
Coeff Variation	53.4438393	Std Error Mean	1578.93853

Basic Statistical Measures			
Location		Variability	
Mean	186851.9	Std Deviation	99861
Median	157500.0	Variance	9972187526
Mode	135000.0	Range	963000
		Interquartile Range	103500

FIGURE 5. Univariate analysis representation of variable INCOMETOTAL

Measure of Central Tendency

- The **mean** is 186851.922
  - On average, a customer has a total income of 186851.922.
- The **median** mark is 157500
- The smallest **mode** is 135000
  - The value 135000 has the highest frequency in the INCOMETOTAL variable.

Measure of Dispersion

- The **standard deviation** is 99861
- The **variance** is 9972187526
- The **range** is 963000
- The **interquartile range** is 103500
- The **skewness** value is 9972187526
  - The positive values indicate that it is positively skewed.
  - The skewness value can be said to be large, hence its difference from the normal distribution is large.
- The **Kurtosis** value is
  - The positive value indicates that the distribution has thick and heavy tails and a sharper peak than the normal distribution.
  - This is called Leptokurtic.

EDUCATIONLEVEL

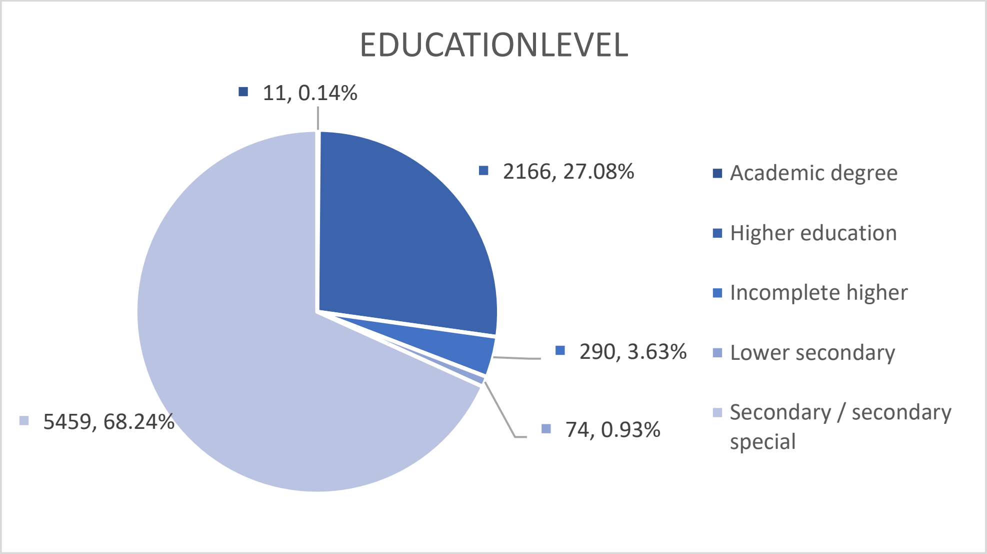


FIGURE 6. Pie chart representation of variable EDUCATIONLEVEL

The values that have been replaced are as follows:

• Secondary/secondary/special	- 1
• Higher education	- 2
• Incomplete higher	- 3
• Lower secondary	- 4
• Academic degree	- 5

A pie chart was generated for variable ‘EDUCATIONLEVEL’.
It is seen that a total of 5459 (68.24%) of the customers are of Secondary / Secondary Special Level while 2166 (27.08%) are of Academic degree.
The 2 EDUCATIONLEVEL with the lowest representation are Higher Education at 290 (3.63%) and Lower Secondary at 74 (0.93%).
It is shown that more than half of the customers in the dataset has obtained an education level of secondary.

MARITALSTATUS

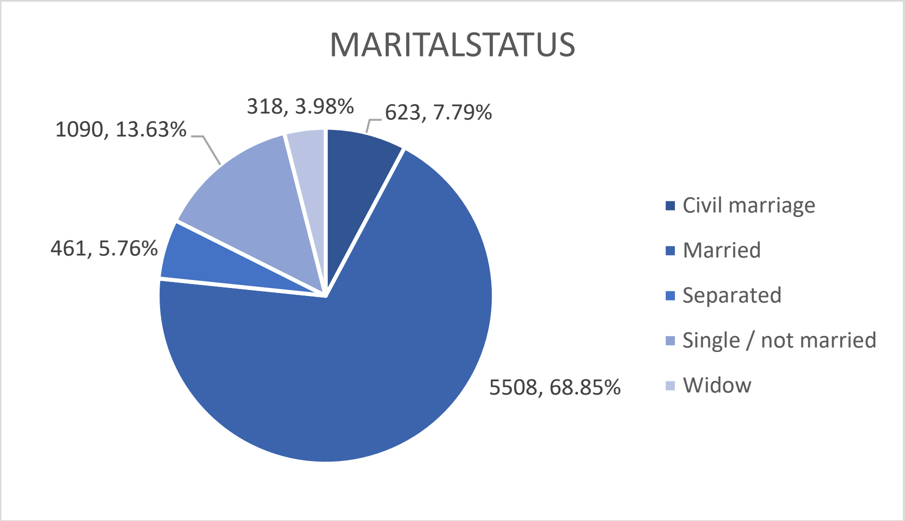


FIGURE 7. Pie chart representation of variable MARITALSTATUS

The values that has been replaced are as follows:

- Married - 1
- Single / not married - 2
- Civil marriage - 3
- Separated - 4
- Widow - 5

A pie chart was generated for variable ‘MARITALSTATUS’.
It is seen that a total of 5508 (68.85%) of the customers are Married while 1090 (13.63%) are Single / Not married.
The 3 MARITALSTATUS with the lowest representation are Civil Marriage at 623 (7.79%), Separated at 461 (5.76%) and finally Widow at 318 (3.98%).
It is shown that more than half of the customers in the dataset are married.

HOUSINGTYPE

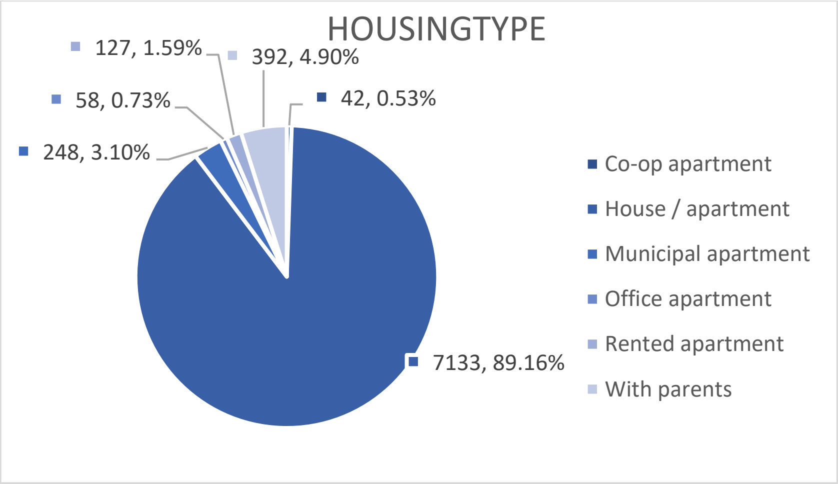


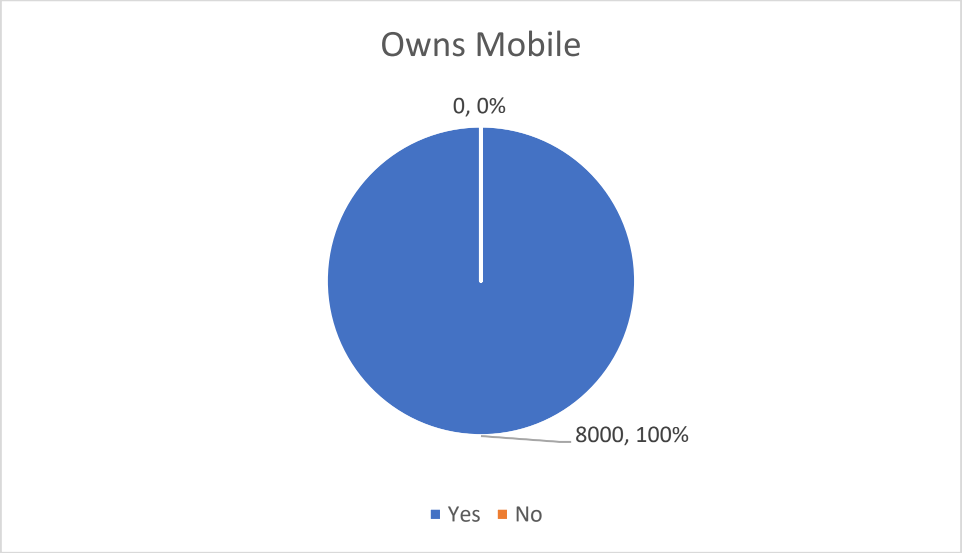
FIGURE 8. Pie chart representation of variable HOUSINGTYPE

The values that has been replaced are as follows:

- House / apartment - 1
- Rented apartment - 2
- With parents - 3
- Co-op apartment - 4
- Municipal apartment - 5
- Office apartment - 6

A pie chart was generated for variable ‘HOUSINGTYPE.
It is seen that a total of 7133 (89.16%) of the customers stays in a house/apartment and a total of 392 (4.90%) stay with their family.
The 4 HOUSINGTYPE with the lowest representation are Municipal Apartment at 248 (3.10%), Rented Apartment at 127 (1.59%), Office Apartments at 58 (0.73%) and finally Co-op Apartment at 42 (0.53%).
It is shown that more than half of the customers in the dataset stay in a house.

MOBILE



1 = Owns a mobile phone; 0 = Does not own a mobile phone
FIGURE 9. Pie chart representation of variable OWNSMOBILE

A pie chart was generated for variable ‘MOBILE.
It is seen 8000 (100%) of the customers owns a mobile phone.
It is shown that all of the customers in the dataset owns a mobile phone.

EMAIL

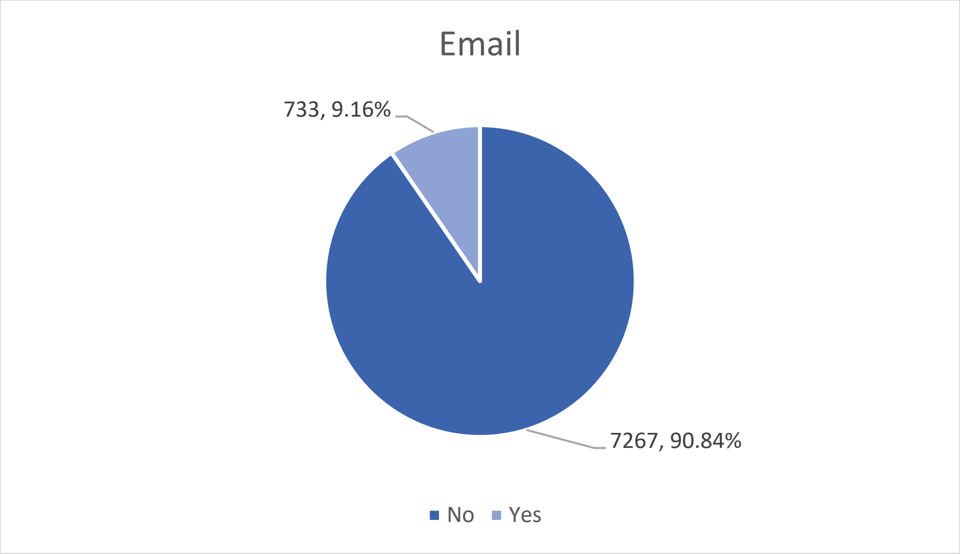


FIGURE 10. Pie chart representation of variable EMAIL

A pie chart was generated for variable 'EMAIL.
It is seen that a total of 7267 (90.39%) of the customers does not have an email while 773 (9.61%) has an email.
It is shown that more than half of the customers in the dataset do not use an email.

OCCUPATION *FIGURE 6. Pie chart representation of variable EDUCATIONLEVEL*

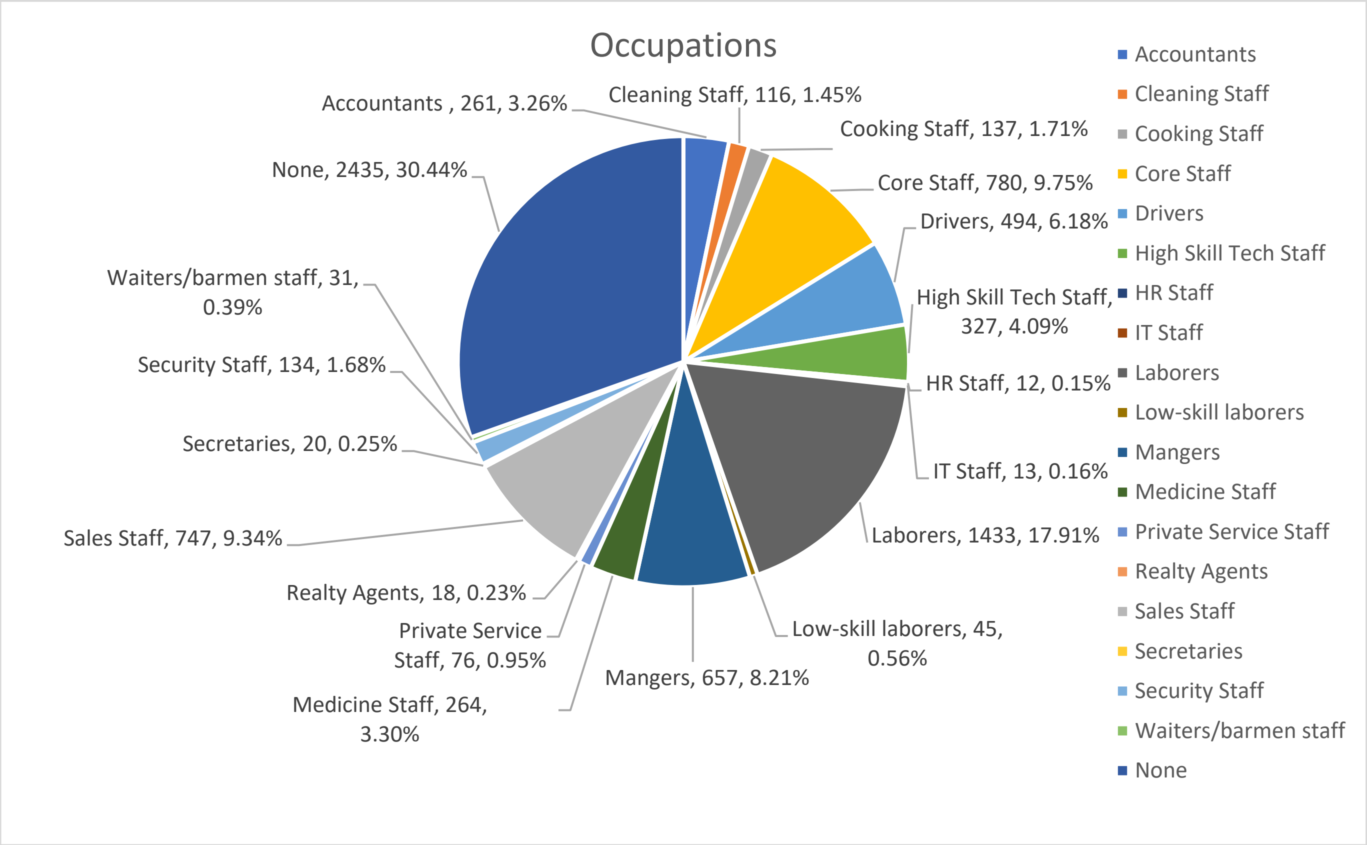


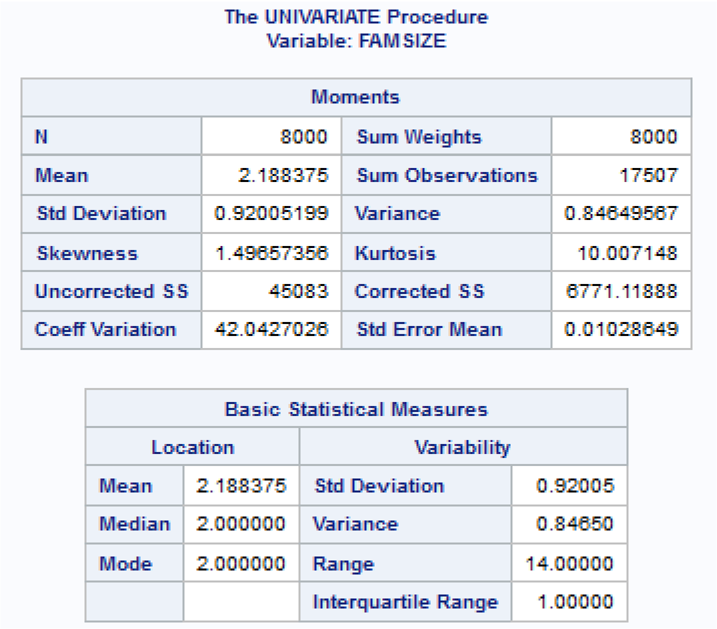
FIGURE 11. Pie chart representation of variable OCCUPATION

The values that have been replaced are as follows:

- None - 0
- Laborers - 1
- Managers - 2
- Drivers - 3
- Security staff - 4
- Accountants - 5
- Core staff - 6
- High skill tech staff - 7
- Sales staff - 8
- Cleaning staff - 9
- Cooking staff - 10
- Medicine staff - 11
- Waiters/barmen staff - 12
- Low-skill Laborers - 13
- Private service staff - 14
- Realty agents - 15
- Secretaries - 16
- HR staff - 17
- IT staff - 18

A pie chart was generated for variable ‘OCCUPATION.
There are numerous occupations listed in the dataset.
No occupation has the highest number of representations at 2435 (30.44%).
The lowest occupation is shown to be IT Staff at 13 (0.16%).

FAMSIZE



The UNIVARIATE Procedure
Variable: FAMSIZE

Moments			
N	8000	Sum Weights	8000
Mean	2.188375	Sum Observations	17507
Std Deviation	0.92005199	Variance	0.84649567
Skewness	1.49657356	Kurtosis	10.007148
Uncorrected SS	45083	Corrected SS	6771.11888
Coeff Variation	42.0427026	Std Error Mean	0.01028649

Basic Statistical Measures			
Location		Variability	
Mean	2.188375	Std Deviation	0.92005
Median	2.000000	Variance	0.84650
Mode	2.000000	Range	14.00000
		Interquartile Range	1.00000

FIGURE 12. Univariate analysis representation of variable FAMSIZE

Measure of Central Tendency

- The **mean** is 2.1884
    - On average, a customer has a family size of 3 (rounded up 2.1884).
- The **median** mark is 2.000
- The smallest **mode** is 2.000
    - The value 2.000 has the highest frequency in the FAMSIZE variable.

Measure of Dispersion

- The **standard deviation** is 0.92005
- The **variance** is 0.8465
- The **range** is 14.000
- The **interquartile range** is 1.000
- The **skewness** value is 1.4966
    - The positive values indicate that it is positively skewed.
    - The skewness value can be said to be moderate, hence its difference from the normal distribution is moderate.
- The **Kurtosis** value is 10.0071
    - The positive value indicates that the distribution has thick and heavy tails and a sharper peak than the normal distribution.
    - This is called Leptokurtic.

CREDITSTATUS

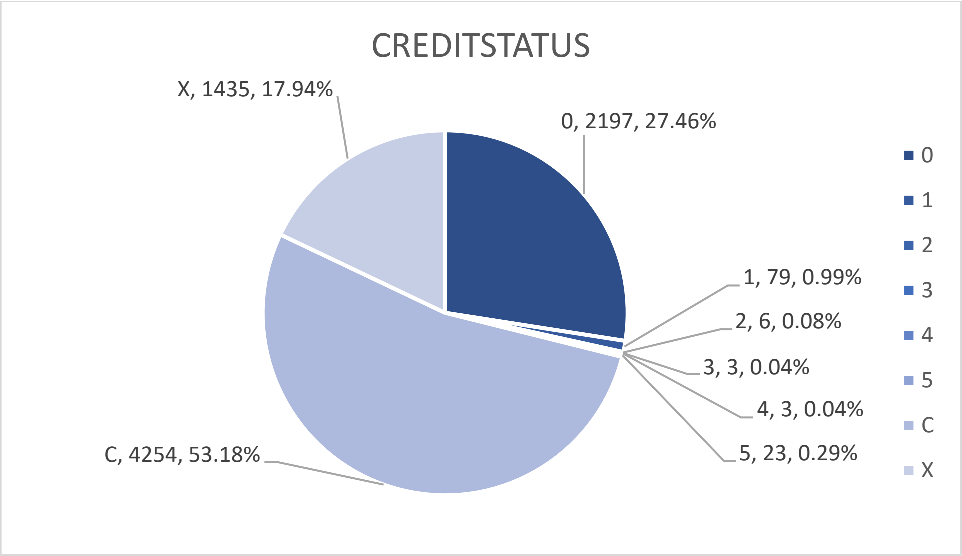


FIGURE 13. Pie chart representation of variable CREDITSTATUS

The values that have been replaced are as follows:

- 0 - 0: 1-29 days past due
- 1 - 1: 30-59 days past due
- 2 - 2: 60-89 days overdue
- 3 - 3: 90-119 days overdue
- 4 - 4: 120-149 days overdue
- 5 - 5: Bad debts / Write-offs
- C - 6: paid off for that month
- X - 7: no loan for the month

A pie chart was generated for the variable ‘CREDITSTATUS.
It is seen that a total of 4254 (53.18%) of the customers has a credit score of C and a total of 2197 (27.46%) on a credit score of 0.
Next up, 1435 (17.94%) of the customers had a credit score of X.
The 5 CREDITSTATUS with the lowest representation are credit score at 79 (0.99%), credit score of 5 at 23 (0.29%), credit score 2 at 6 (0.08%), credit score 3 at 3 (0.04%), credit score 4 at 3 (0.04%) and finally credit score 5 at 23 (0.29%).
It is shown that more than half of the customers in the dataset has a credit score of C.

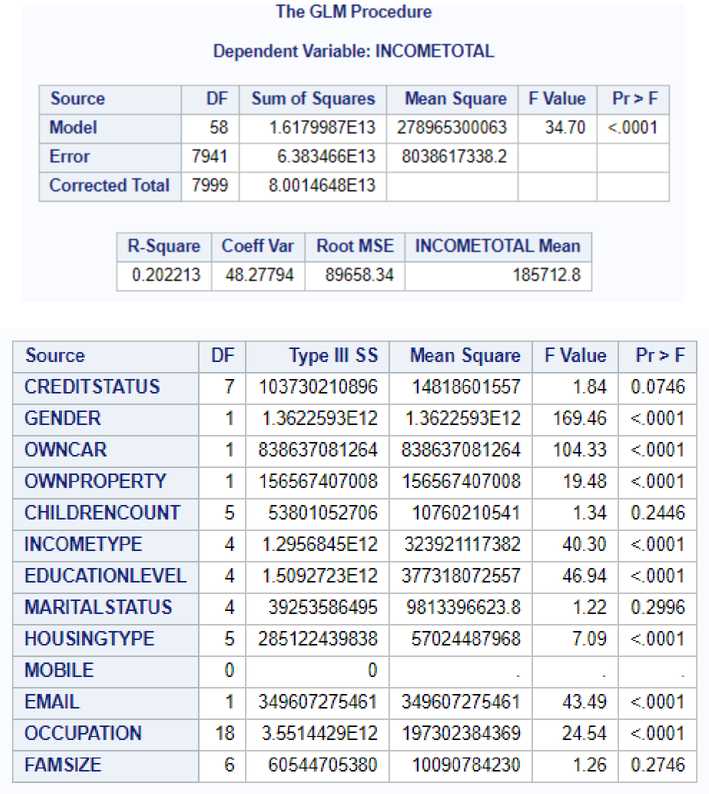
3. Analysis

Objective 1:

To determine the whether the variables GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE, CREDITSTATUS that have a significant effect on the response variable INCOMETOTAL.

The purpose of this analysis is to identify the variables that has a significant relationship with customer's annual income sum along with the relationship and significance. I have identified INCOMETOTAL as my Response Variable and GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE, CREDITSTATUS as my Predictor variables.

The statistical analysis technique being used is Analysis of Covariance (ANCOVA) as the test consist of a continuous variable as a Response Variable and continuous variables and categorical variables as Predictor Variables.



The GLM Procedure

Dependent Variable: INCOMETOTAL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	58	1.6179987E13	278965300063	34.70	<.0001
Error	7941	6.383466E13	8038617338.2		
Corrected Total	7999	8.0014648E13			

R-Square	Coeff Var	Root MSE	INCOMETOTAL Mean
0.202213	48.27794	89658.34	185712.8

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CREDITSTATUS	7	103730210896	14818601557	1.84	0.0746
GENDER	1	1.3622593E12	1.3622593E12	169.46	<.0001
OWNCAR	1	838637081264	838637081264	104.33	<.0001
OWNPROPERTY	1	156567407008	156567407008	19.48	<.0001
CHILDRENCOUNT	5	53801052706	10760210541	1.34	0.2446
INCOMETYPE	4	1.2956845E12	323921117382	40.30	<.0001
EDUCATIONLEVEL	4	1.5092723E12	377318072557	46.94	<.0001
MARITALSTATUS	4	39253586495	9813396623.8	1.22	0.2996
HOUSINGTYPE	5	285122439838	57024487968	7.09	<.0001
MOBILE	0	0	.	.	.
EMAIL	1	349607275461	349607275461	43.49	<.0001
OCCUPATION	18	3.5514429E12	197302384369	24.54	<.0001
FAMSIZE	6	60544705380	10090784230	1.26	0.2746

FIGURE 14. SAS Studio Output

Based on the SAS Studio output, the response variable INCOMETOTAL has a mean of 185712.8. It can be said that on average a customer earns 185712.80 per annum.

The R-Square value is 0.2022. This means that 20.22% of the variation in INCOMETOTAL is explained by the variation in the predictor variables.

The model does not provide a good statistical analysis as the predictor variables only account to 20.22% of the changes in the response variable INCOMETOTAL. However, further analysis is to be done in order to identify if the predictor variables have a significant effect on the response variable.

F-test for the overall model

H_0 : There is no significant relationship between the predictor variables, GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE, CREDITSTATUS and the response variable INCOMETOTAL.

H_1 : At least one of the predictor variables, GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE, CREDITSTATUS has a significant relationship with the response variable INCOMETOTAL.

Based on the SAS Studio output, the F value is shown to be 34.70 along with a p-value of <0.0001.

H_0 is rejected at the 0.05 level of significance (α =0.05).

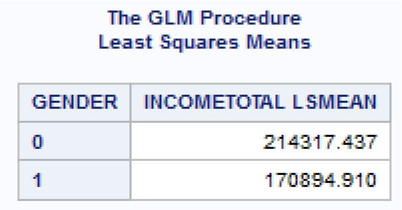
At least one of the predictor variables (GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE, CREDITSTATUS) have a significant effect on the predictor variable, INCOMETOTAL.

Individual Test for Significance

Based on the SAS Studio output, H_0 is rejected for the variables with a p-value below the significant level of 0.05.

There are 8 variables with a p-value less than 0.05. Hence, variables GENDER, OWNCAR, OWNPROPERTY, INCOMETYPE, EDUCATIONLEVEL, HOUSINGTYPE, EMAIL AND OCCUPATION has a significant effect on the response variable INCOMETOTAL.

Least Squares Means Analysis on the Significant Predictor Variables

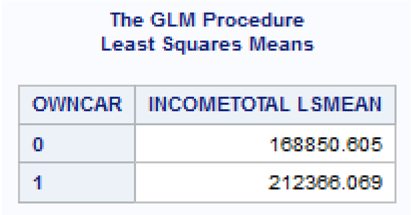


The GLM Procedure
Least Squares Means

GENDER	INCOMETOTAL LSMEAN
0	214317.437
1	170894.910

FIGURE 15. GENDER Least Square Means

Based on the SAS Studio result, it is shown that the GENDER 1 (Male) has a higher mean annual income compared to a GENDER 0 (Female).

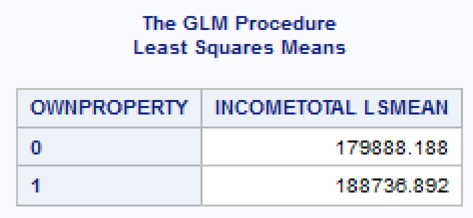


The GLM Procedure
Least Squares Means

OWNCAR	INCOMETOTAL LSMEAN
0	168850.605
1	212366.069

FIGURE 16. OWNCAR Least Square Means

Based on the SAS Studio result, it is shown that those with Own a Car value 1 (Owns a car) has a higher mean annual income compared to those who don’t.

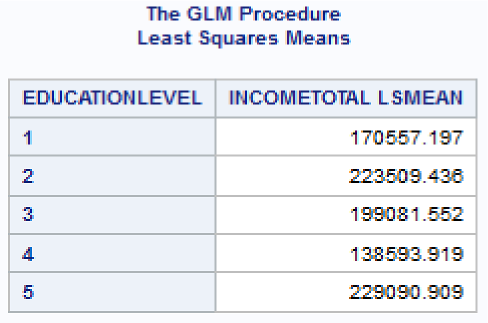


The GLM Procedure
Least Squares Means

OWNPROPERTY	INCOMETOTAL LSMEAN
0	179888.188
1	188736.892

FIGURE 17. OWNPROPERTY Least Square Means

Based on the SAS Studio result, it is shown that those who has a OwnPropertyvalue of 1 (Owns a Property) has a higher mean annual income compared to those who don’t.

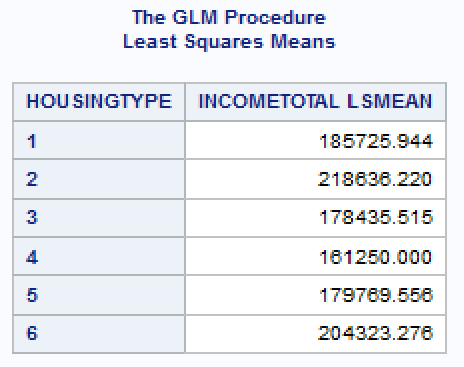


The GLM Procedure
Least Squares Means

EDUCATIONLEVEL	INCOMETOTAL LSMEAN
1	170557.197
2	223509.436
3	199081.552
4	138593.919
5	229090.909

FIGURE 18. EDUCATIONLEVEL Least Square Means

Based on the SAS Studio result, it is shown that those with EDUCATIONLEVEL 5 (Academic) has a higher mean annual income compared to the other EDUCATIONLEVEL.

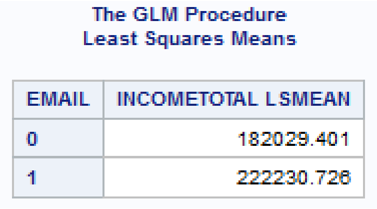


The GLM Procedure
Least Squares Means

HOUSINGTYPE	INCOMETOTAL LSMEAN
1	185725.944
2	218636.220
3	178435.515
4	161250.000
5	179769.556
6	204323.276

FIGURE 19. HOUSINGTYPE Least Square Means

Based on the SAS Studio result, it is shown that those with HOUSINGTYPE 2 (Rented Apartment) has a higher mean annual income compared to the other HOUSINGTYPE.

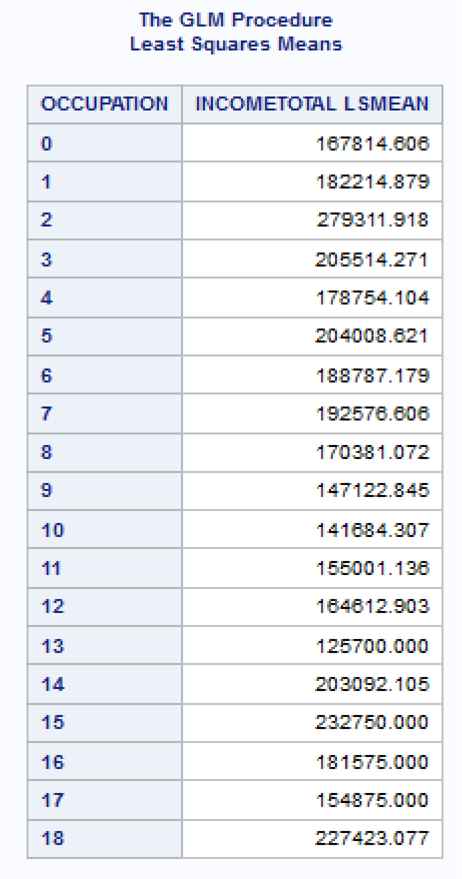


The GLM Procedure
Least Squares Means

EMAIL	INCOMETOTAL LSMEAN
0	182029.401
1	222230.726

FIGURE 20. EMAIL Least Square Means

Based on the SAS Studio result, it is shown that those with EMAIL LEVEL 1 (YES) has a higher mean annual income compared to those who do not have an email.



The GLM Procedure
Least Squares Means

OCCUPATION	INCOMETOTAL LSMEAN
0	167814.606
1	182214.879
2	279311.918
3	205514.271
4	178754.104
5	204008.621
6	188787.179
7	192576.606
8	170381.072
9	147122.845
10	141684.307
11	155001.136
12	164612.903
13	125700.000
14	203092.105
15	232750.000
16	181575.000
17	154875.000
18	227423.077

FIGURE 21. OCCUPATION Least Square Means

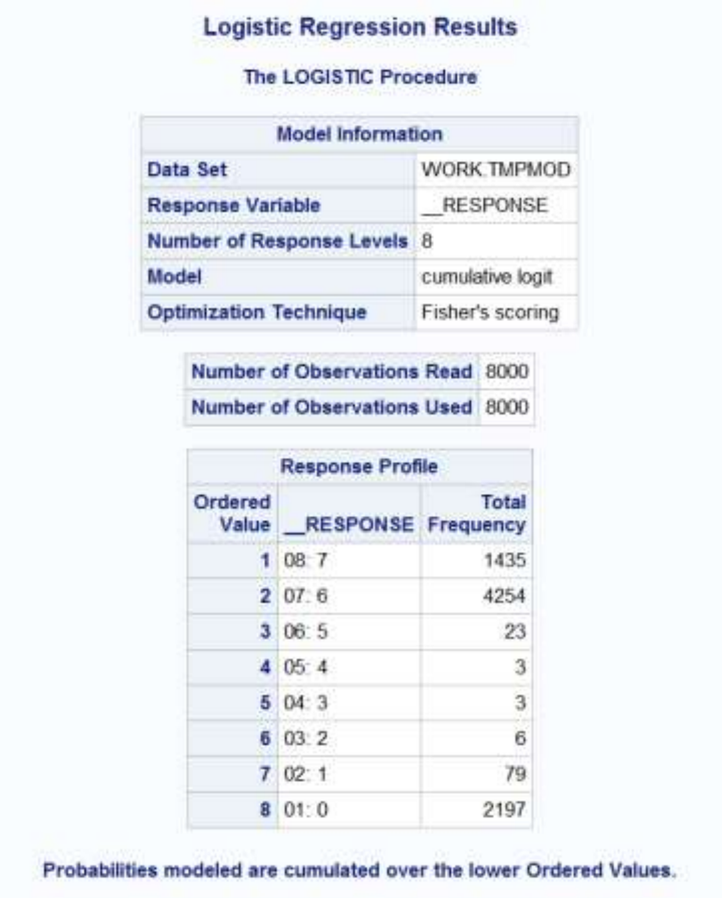
Based on the SAS Studio result, it is shown that those with OCCUPATION 15 (Manager) has a higher mean annual income compared to the other OCCUPATION.

Objective 2:

To determine the whether the variables GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, INCOMETOTAL, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION and FAMSIZE that have a significant effect on the response variable CREDITSTATUS.

The purpose of this analysis is to identify the variables that has a significant relationship with customer's annual income sum along with the relationship and significance. I have identified CREDITSCORE as my Response Variable and GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, INCOMETOTAL, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE, CREDITSTATUS as the Predictor variables.

The statistical analysis technique being used is **Logistics Regression** as the test consist of a categorical variable as a Response Variable and continuous variables and categorical variables as Predictor Variables.



Logistic Regression Results

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TMPMOD
Response Variable	__RESPONSE
Number of Response Levels	8
Model	cumulative logit
Optimization Technique	Fisher's scoring

Number of Observations Read	8000
Number of Observations Used	8000

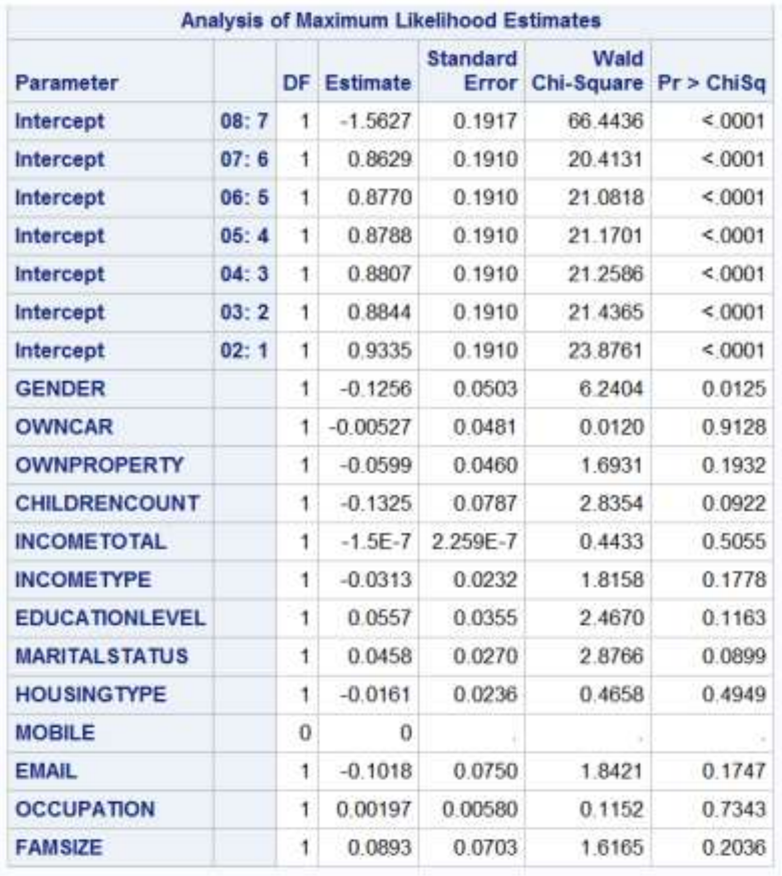
Response Profile		
Ordered Value	__RESPONSE	Total Frequency
1	08: 7	1435
2	07: 6	4254
3	06: 5	23
4	05: 4	3
5	04: 3	3
6	03: 2	6
7	02: 1	79
8	01: 0	2197

Probabilities modeled are cumulated over the lower Ordered Values.

FIGURE 22. SAS Studio Output

The Model Information Table describes the Logistic Regression process. Its description also includes the number of response level. In this case, variable _RESPONSE represents variable CREDITSCORE which is set as the Response Variable. The variable has 8 levels of values that are 0= 1-29 days past due, 1=30-59 days past due, 2= 60-89 days due, 3= 90-119 days overdue, 4= 120-149 days overdue, 5= Bad debts / Write-offs, 6= paid off for that month and 7= no loan for the month.

The Response Profile Table indicates the total frequency of the 8 values of variable CREDITSCORE.

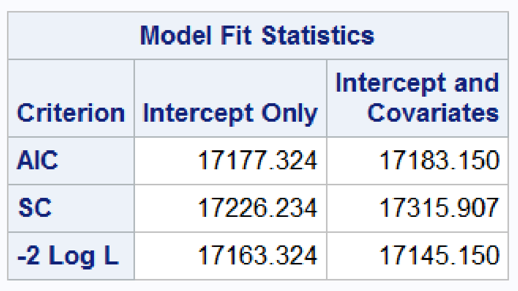


Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	08: 7	1	-1.5627	0.1917	66.4436	<.0001
Intercept	07: 6	1	0.8629	0.1910	20.4131	<.0001
Intercept	06: 5	1	0.8770	0.1910	21.0818	<.0001
Intercept	05: 4	1	0.8788	0.1910	21.1701	<.0001
Intercept	04: 3	1	0.8807	0.1910	21.2586	<.0001
Intercept	03: 2	1	0.8844	0.1910	21.4365	<.0001
Intercept	02: 1	1	0.9335	0.1910	23.8761	<.0001
GENDER		1	-0.1256	0.0503	6.2404	0.0125
OWNCAR		1	-0.00527	0.0481	0.0120	0.9128
OWNPROPERTY		1	-0.0599	0.0460	1.6931	0.1932
CHILDRENCOUNT		1	-0.1325	0.0787	2.8354	0.0922
INCOMETOTAL		1	-1.5E-7	2.259E-7	0.4433	0.5055
INCOMETYPE		1	-0.0313	0.0232	1.8158	0.1778
EDUCATIONLEVEL		1	0.0557	0.0355	2.4670	0.1163
MARITALSTATUS		1	0.0458	0.0270	2.8766	0.0899
HOUSINGTYPE		1	-0.0161	0.0236	0.4658	0.4949
MOBILE		0	0	.	.	.
EMAIL		1	-0.1018	0.0750	1.8421	0.1747
OCCUPATION		1	0.00197	0.00580	0.1152	0.7343
FAMSIZE		1	0.0893	0.0703	1.6165	0.2036

FIGURE 23. Analysis of Maximum Likelihood Estimates

Model Fit Statistic



Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	17177.324	17183.150
SC	17226.234	17315.907
-2 Log L	17163.324	17145.150

FIGURE 24. Model Fit Statistics

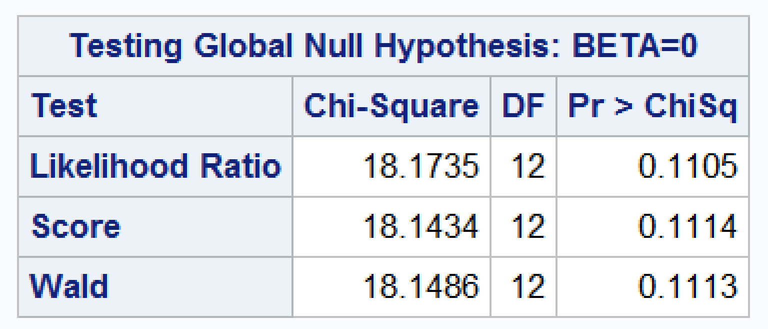
The goodness of fit measures is measured by comparing the difference between Intercept Only and Intercept and Covariates.

Based on the output result, criterion -2 Log L has an Intercept only value of 17163.324 whereas the intercept and covariates have a value of 17145.150.

The model has a good fit as there is a difference greater than 5 between the intercept only and intercept and covariates.

Thus, the model fit statistics indicate that **GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETYPE, INCOMETOTAL, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION and FAMSIZE** as predictor variables gives a better fit than an empty model.

Test For Collective Significance



Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.1735	12	0.1105
Score	18.1434	12	0.1114
Wald	18.1486	12	0.1113

FIGURE 25. Testing Global Null Hypothesis Table

This table is used to identify the collective significance of the predictor variables in the model.

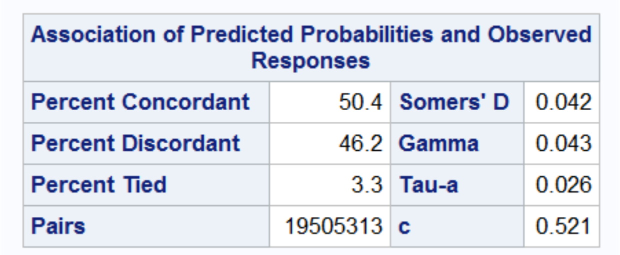
H_0 : All the regression coefficients are 0.

H_1 : At least one of the regression coefficients is not 0.

Based on the output, all 3 tests - Likelihood Ratio, Score and Wald has a p-value of >0.05.

At the 0.05 significance level, H_0 is not rejected. Hence, we can conclude that the predictor variables in this logistic regression model are not collectively significant.

Concordance Statistic Value

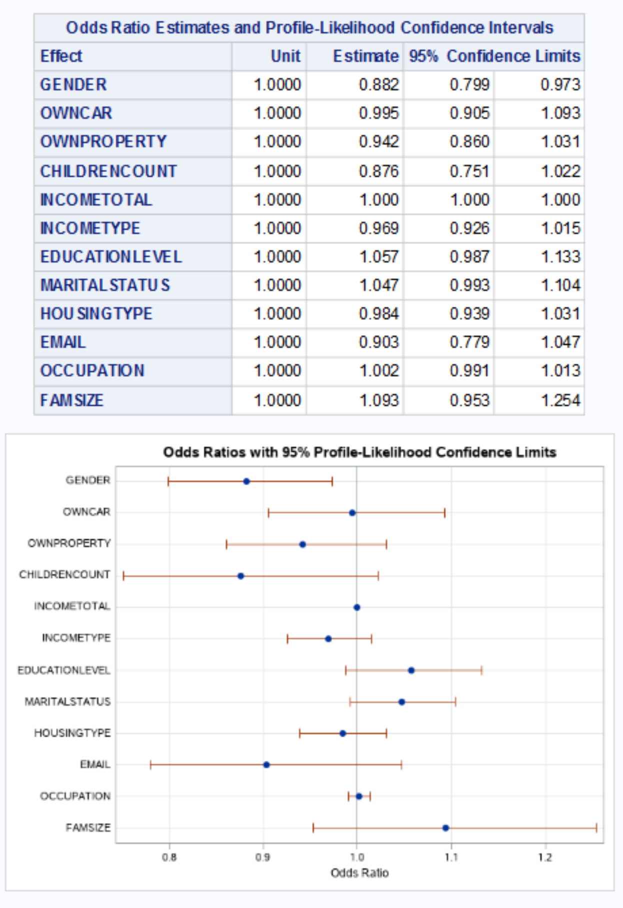


Association of Predicted Probabilities and Observed Responses			
Percent Concordant	50.4	Somers' D	0.042
Percent Discordant	46.2	Gamma	0.043
Percent Tied	3.3	Tau-a	0.026
Pairs	19505313	c	0.521

FIGURE 26. Concordance Table

The c (concordance) statistic value is 0.521 for this model, indicating that the model can correctly classify the outcome at a percentage of 52.10%.

Odds Ratio Plot



Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
GENDER	1.0000	0.882	0.799	0.973
OWNCAR	1.0000	0.995	0.905	1.093
OWNPROPERTY	1.0000	0.942	0.860	1.031
CHILDRENCOUNT	1.0000	0.876	0.751	1.022
INCOMETOTAL	1.0000	1.000	1.000	1.000
INCOMETYPE	1.0000	0.969	0.926	1.015
EDUCATIONLEVEL	1.0000	1.057	0.987	1.133
MARITALSTATUS	1.0000	1.047	0.993	1.104
HOUSINGTYPE	1.0000	0.984	0.939	1.031
EMAIL	1.0000	0.903	0.779	1.047
OCCUPATION	1.0000	1.002	0.991	1.013
FAMSIZE	1.0000	1.093	0.953	1.254

FIGURE 27. Odds Ratio Estimates Table

At a 95% confidence level, Gender is significant as it does not cross the reference line.

We are 95% confident that the effect of the odds ratio of Gender is between 0.799 and 0.973.

Effects Plot

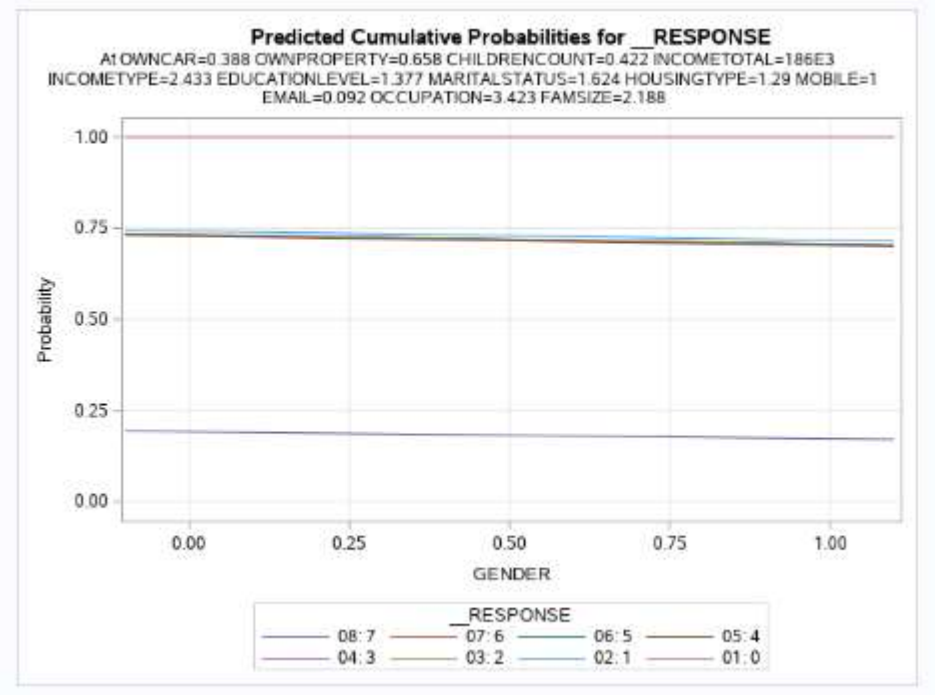


FIGURE 28. Effects Plot

The effects plot above shows the probability of **CREDITSCORE** = 7 (No loan for the month) across all the different combinations of categories and levels of predictor variables.

The plot shows that the probability of GENDER male and female in achieving CREDITSCORE 7 (No loan for the month).It is shown that the females (1) has a lower possibility of having no loan for the month compared to males (0).

Objective 3 :

To determine how variable GENDER is associated to variable OWNCAR and OWNPROPERTY.

The purpose of this analysis is to identify the relationship between variable GENDER and variable OWNCAR and OWNPROPERTY.

The statistical analysis technique being used is **Categorical Data Analysis – Table Analysis** as the test consist of a categorical variable as a Response Variable and continuous variables and categorical variables as Predictor Variables.

Test between variable OWNCAR and GENDER

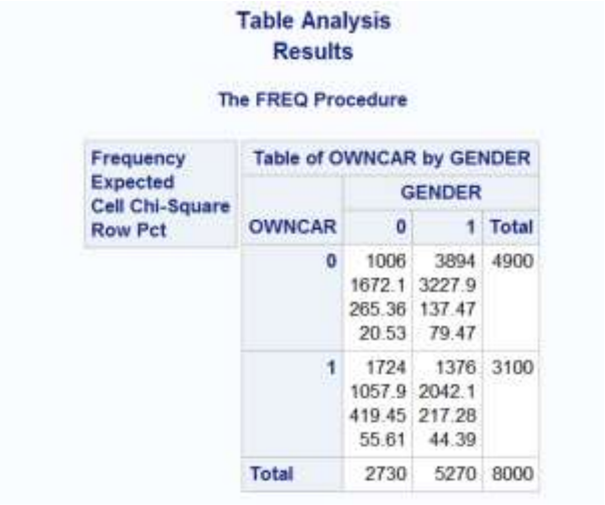


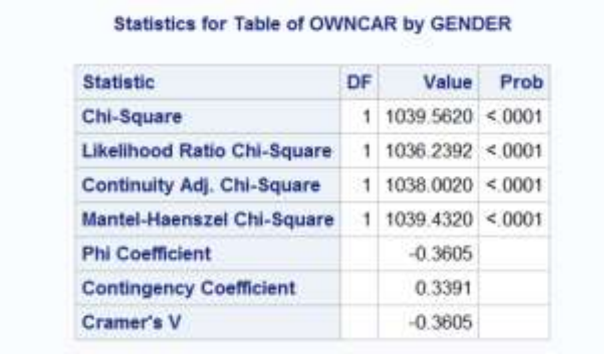
Table Analysis
Results

The FREQ Procedure

Frequency Expected Cell Chi-Square Row Pct	Table of OWNCAR by GENDER			
		GENDER		
	OWNCAR	0	1	Total
	0	1006 1672.1 265.36 20.53	3894 3227.9 137.47 79.47	4900
	1	1724 1057.9 419.45 55.61	1376 2042.1 217.28 44.39	3100
	Total	2730	5270	8000

FIGURE 29. Table Analysis

Based on the results from SAS EG, we can observe a relationship between variable GENDER and variable OWNCAR.



Statistics for Table of OWNCAR by GENDER

Statistic	DF	Value	Prob
Chi-Square	1	1039.5620	<.0001
Likelihood Ratio Chi-Square	1	1036.2392	<.0001
Continuity Adj. Chi-Square	1	1038.0020	<.0001
Mantel-Haenszel Chi-Square	1	1039.4320	<.0001
Phi Coefficient		-0.3605	
Contingency Coefficient		0.3391	
Cramer's V		-0.3605	

FIGURE 30. Statistics for Table of OWNCAR by Gender

Test for Association

H_0 : There is no association between variable OWNCAR and variable GENDER.

H_1 : There is an association between variable OWNCAR and variable GENDER.

Based on the SAS EG Output, X^2 has a value of 1039.5620 , with a p-value of <0.0001.

At a 5% significance level, H_0 is rejected.

There is evidence of an association between variable OWNCAR and variable GENDER.

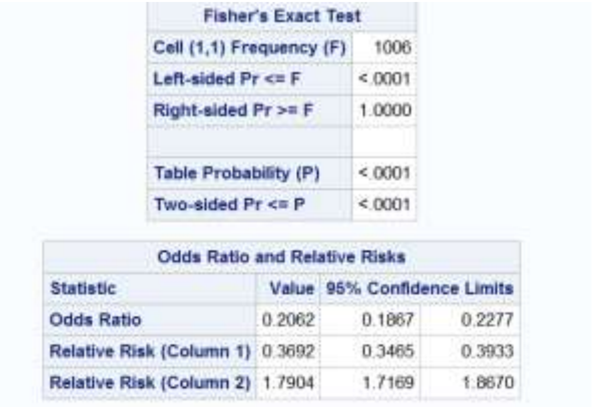
Test for Strength of Association

Phi Coefficient : -0.3605

Contingency Coefficient : 0.3391

Cramer’s V : -0.3605

Cramer’s V of -0.3605 indicates that the association detected between variable OWNCAR and variable GENDER with the chi-square test is relatively moderate.



Fisher's Exact Test	
Cell (1,1) Frequency (F)	1006
Left-sided Pr <= F	<.0001
Right-sided Pr >= F	1.0000
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.2062	0.1867	0.2277
Relative Risk (Column 1)	0.3692	0.3465	0.3933
Relative Risk (Column 2)	1.7904	1.7169	1.8670

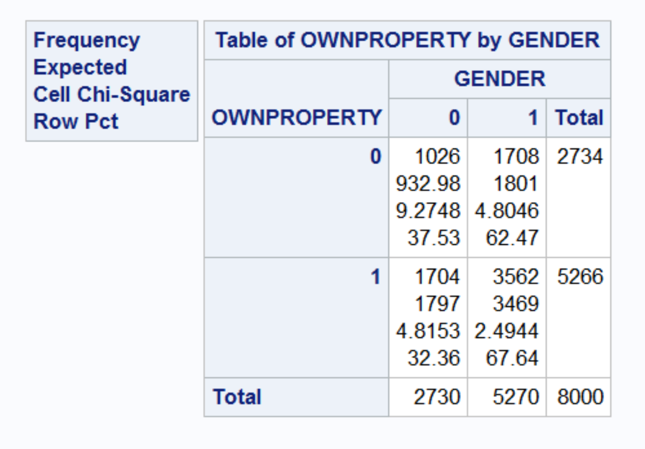
FIGURE 30. Fisher’s Exact Test

Test for independence

Based on the Fisher’s Exact Test, the two-sided p-value of <0.0001 is very small.

At the 5% significance level, there is evidence of an association between variable GENDER and variable OWNCAR.

Test between variable OWNPROPERTY and GENDER

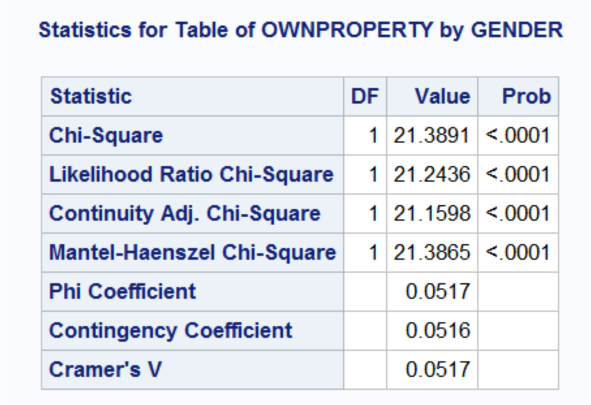


Frequency
Expected
Cell Chi-Square
Row Pct

Table of OWNPROPERTY by GENDER			
	GENDER		
OWNPROPERTY	0	1	Total
0	1026 932.98 9.2748 37.53	1708 1801 4.8046 62.47	2734
1	1704 1797 4.8153 32.36	3562 3469 2.4944 67.64	5266
Total	2730	5270	8000

FIGURE 30. Table of OWNPROPERTY by GENDER

Based on the results from SAS EG, we can observe a relationship between variable GENDER and variable OWNPROPERTY.



Statistics for Table of OWNPROPERTY by GENDER

Statistic	DF	Value	Prob
Chi-Square	1	21.3891	<.0001
Likelihood Ratio Chi-Square	1	21.2436	<.0001
Continuity Adj. Chi-Square	1	21.1598	<.0001
Mantel-Haenszel Chi-Square	1	21.3865	<.0001
Phi Coefficient		0.0517	
Contingency Coefficient		0.0516	
Cramer's V		0.0517	

FIGURE 31. Statistics for Table of OWNPROPERTY by Gender

Test for Association

H_0 : There is no association between variable OWNPROPERTY and variable GENDER.

H_1 : There is an association between variable OWNPROPERTY and variable GENDER.

Based on the SAS EG Output, X^2 has a value of 21.3891, with a p-value of <0.0001.

At a 5% significance level, H_0 is rejected.

There is evidence of an association between variable OWNPROPERTY and variable GENDER.

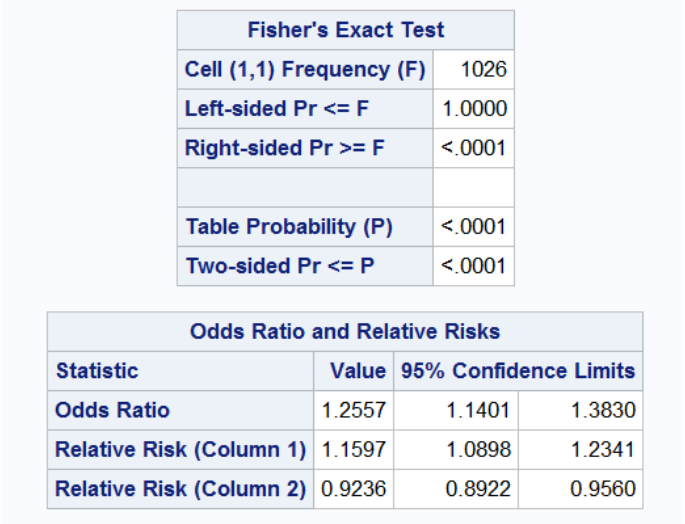
Test for Strength of Association

Phi Coefficient : 0.0517

Contingency Coefficient : 0.0516

Cramer’s V : 0.0517

Cramer’s V of 0.0517 indicates that the association detected between variable OWNPROPERTY and variable GENDER with the chi-square test is low.



Fisher's Exact Test	
Cell (1,1) Frequency (F)	1026
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	1.2557	1.1401	1.3830
Relative Risk (Column 1)	1.1597	1.0898	1.2341
Relative Risk (Column 2)	0.9236	0.8922	0.9560

FIGURE 30. Fisher’s Exact Test

Test for independence

Based on the Fisher’s Exact Test, the two-sided p-value of <0.0001 is very small.

At the 5% significance level, there is evidence of an association between variable GENDER and variable OWNPROPERTY.

4. Conclusion

In conclusion, I have carried out 3 Statistical Techniques, Analysis of Covariance, Logistic Regression, and Categorical Data Analysis – Table Analysis.

I have been able to identify whether the specific predictor variable(s) have a significant effect on the response variable.

In the Analysis of Covariance, I have examined which predictor variables has a significant effect on the response variable INCOMETOTAL.

I have concluded that variables GENDER, OWNCAR, OWNPROPERTY, INCOMETYPE, EDUCATIONLEVEL, HOUSINGTYPE, EMAIL AND OCCUPATION has a significant effect on the response variable INCOMETOTAL. In addition to that, I have also identified which level within the specific variables which has a higher mean annual income.

In the Logistic Regression, I have examined which predictor variable has a significant effect on the response variable, CREDITSCORE. This allows me to understand which variable has a positive effect on the CREDITSCORE. As part of our analysis I have included model fit statistics, odds ratio plot, concordance statistic value and effects plot.

I have concluded that variable GENDER is statistically significant in predicting the CREDITSCORE of the customer while the rest of the predictor variables are not.

Finally, I have also employed the Categorical Data Analysis – Table Analysis. The purpose of this analysis was to find out whether variable GENDER is associated to variables OWNCAR and OWNPROPERTY. This analysis included the test for Association, test for Strength of Association and Test for Independence. With the output from the analysis, I are able to identify and compare whether variable GENDER is associated to variables OWNCAR and OWNPROPERTY.

I have concluded that variable GENDER is statistically associated with variables OWNCAR and OWNPROPERTY.

References

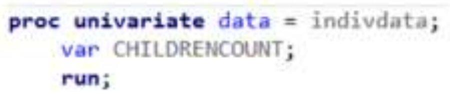
(2019, June 15). Retrieved from Minitab: https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/basic-statistics/summary-statistics/descriptive-statistics/interpret-the-results/all-statistics-and-graphs/

Naird, W. (2019, May 5). Retrieved from Statalogy: https://www.statology.org/how-to-interpret-the-c-statistic-of-a-logistic-regression-model/#:~:text=The%20c%2Dstatistic%2C%20also%20known,0.5%20indicates%20a%20poor%20model.&text=The%20closer%20the%20value%20is,is%20at%20correctly%20classifying%20outcomes.

Narkhede, S. (2018, June 6). *medium.com*. Retrieved from TowardsDataScience: https://towardsdatascience.com/understanding-descriptive-statistics-c9c2b0641291

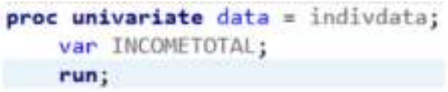
Appendix

Descriptive Analysis



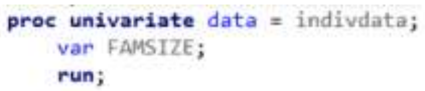
```
proc univariate data = indivdata;
    var CHILDRENCOUNT;
    run;
```

Code Screenshot of Univariate Analysis for Variable CHILDRENCOUNT



```
proc univariate data = indivdata;
    var INCOMETOTAL;
    run;
```

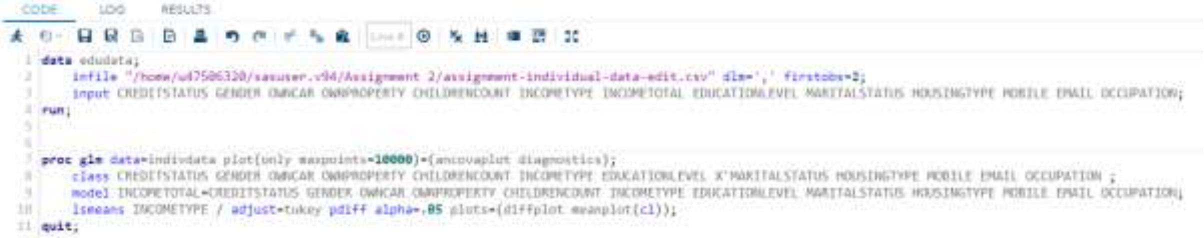
Code Screenshot of Univariate Analysis for Variable INCOMETOTAL



```
proc univariate data = indivdata;
    var FAMSIZE;
    run;
```

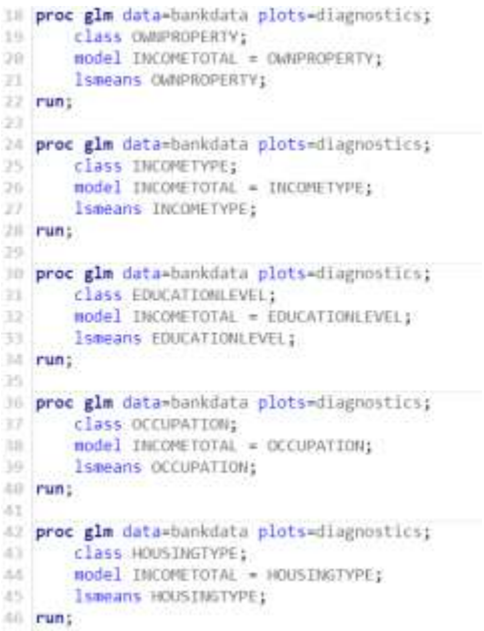
Code Screenshot of Univariate Analysis for Variable FAMSIZE

Analysis Of Covariance:



```
data edudata;
    infile "/home/u47506320/sasuser.v94/Assignment 2/assignment-individual-data-edit.csv" dlm=',' firstobs=2;
    input CREDITSTATUS GENDER OWNCAR OWNPROPERTY CHILDRENCOUNT INCOMETYPE INCOMETOTAL EDUCATIONLEVEL MARITALSTATUS HOUSINGTYPE MOBILE EMAIL OCCUPATION;
run;


proc glm data=indivdata plot(only maxpoints=10000)=(ancovaplot diagnostics);
    class CREDITSTATUS GENDER OWNCAR OWNPROPERTY CHILDRENCOUNT INCOMETYPE EDUCATIONLEVEL X'MARITALSTATUS HOUSINGTYPE MOBILE EMAIL OCCUPATION ;
    model INCOMETOTAL=CREDITSTATUS GENDER OWNCAR OWNPROPERTY CHILDRENCOUNT INCOMETYPE EDUCATIONLEVEL MARITALSTATUS HOUSINGTYPE MOBILE EMAIL OCCUPATION;
    lsmeans INCOMETYPE / adjust=tukey pdiff alpha=.05 plots=(diffplot meanplot(cl));
quit;
```

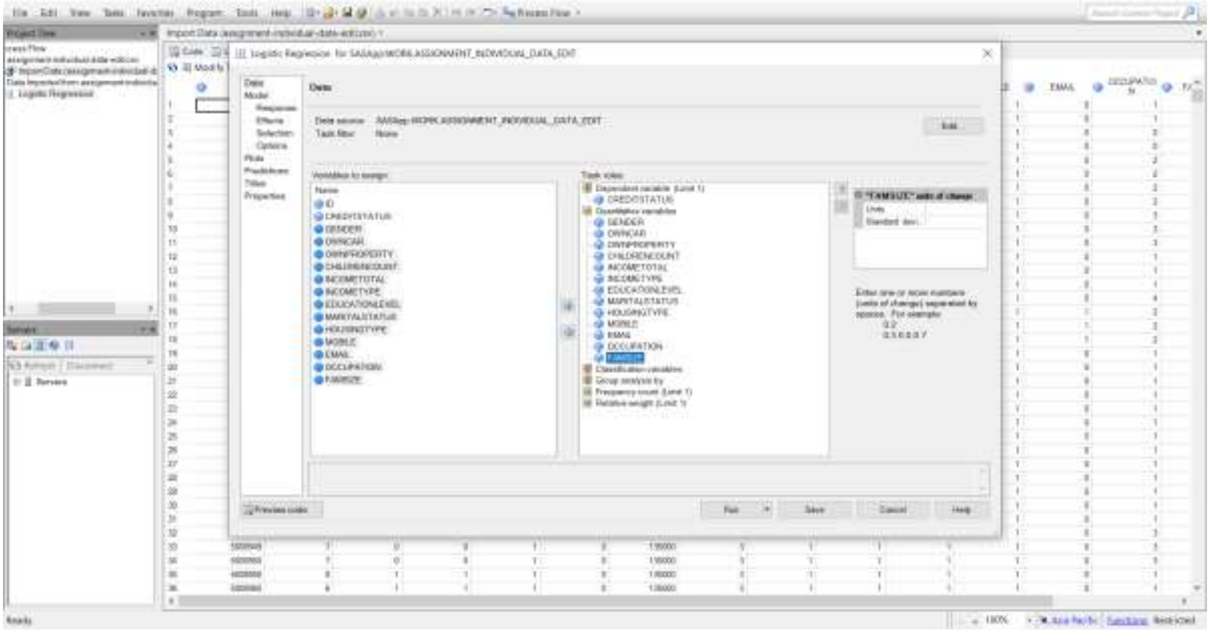
Code screenshot to perform ANCOVA



```
proc glm data=bankdata plots=diagnostics;
    class OWNPROPERTY;
    model INCOMETOTAL = OWNPROPERTY;
    lsmeans OWNPROPERTY;
run;

proc glm data=bankdata plots=diagnostics;
    class INCOMETYPE;
    model INCOMETOTAL = INCOMETYPE;
    lsmeans INCOMETYPE;
run;

proc glm data=bankdata plots=diagnostics;
    class EDUCATIONLEVEL;
    model INCOMETOTAL = EDUCATIONLEVEL;
    lsmeans EDUCATIONLEVEL;
run;

proc glm data=bankdata plots=diagnostics;
    class OCCUPATION;
    model INCOMETOTAL = OCCUPATION;
    lsmeans OCCUPATION;
run;

proc glm data=bankdata plots=diagnostics;
    class HOUSINGTYPE;
    model INCOMETOTAL = HOUSINGTYPE;
    lsmeans HOUSINGTYPE;
run;
```

Code screenshot to calculate LSMEANS

Logistic regression:



Logistic Regression for SASApp:WORK.ASSIGNMENT_INDIVIDUAL_DATA_EDIT

Data source: SASApp:WORK.ASSIGNMENT_INDIVIDUAL_DATA_EDIT
Task filter: None

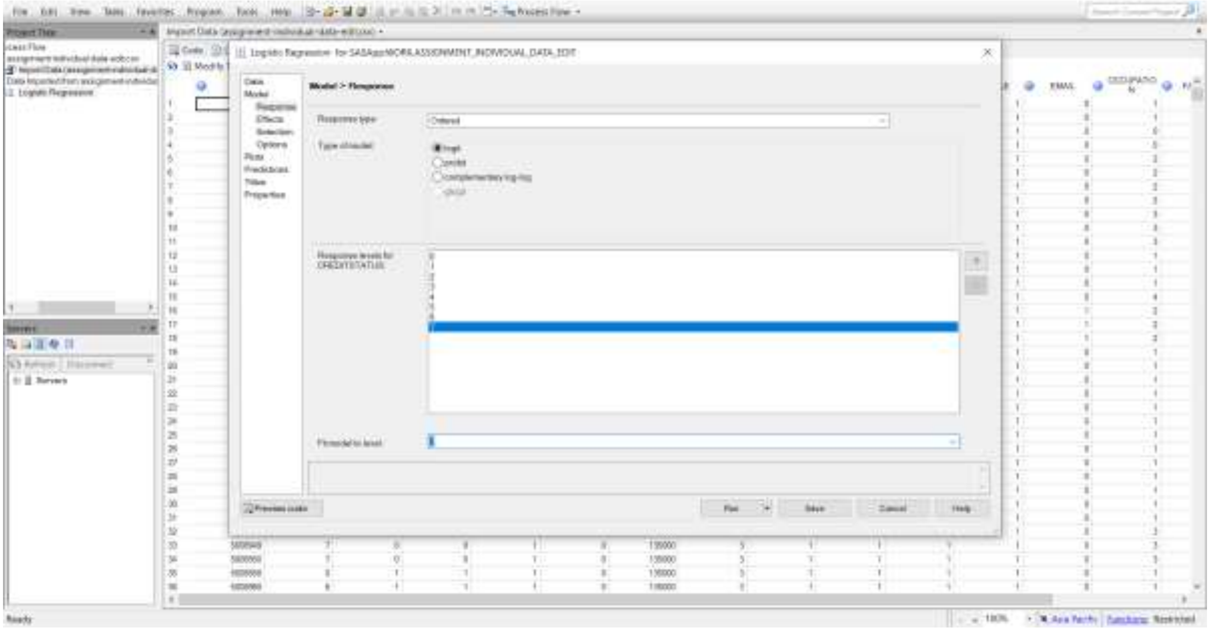
Variables to assign: ID, CREDITSTATUS, GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETOTAL, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE

Task roles:
- Dependent variable (Limit 1): CREDITSTATUS
- Quantitative variables: GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETOTAL, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE
- Classification variables
- Group analysis by
- Frequency count (Limit 1)
- Relative weight (Limit 1)

"FAMSIZE" units of change
Units
Standard dev...

Enter one or more numbers (units of change) separated by spaces. For example:
0.2
0.5 6 0.8 7

Assigning Variables on SAS EG



Logistic Regression for SASApp:WORK.ASSIGNMENT_INDIVIDUAL_DATA_EDIT

Model > Response

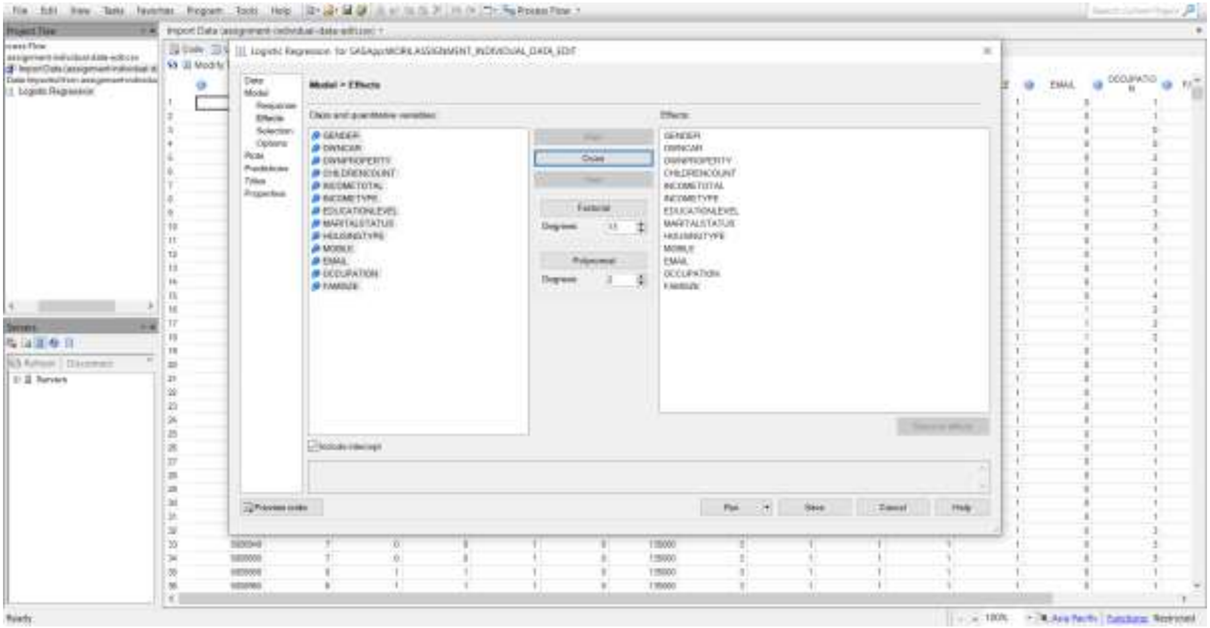
Response type: Ordered

Type of model: ● logit ○ probit ○ complementary log-log

Response levels for CREDITSTATUS: 0 1 2 3 4 5 6 7

Fit model to level: 7

Setting the FIT MODEL LEVEL as well as RESPONSE LEVEL to 7



Logistic Regression for SASApp:WORK.ASSIGNMENT_INDIVIDUAL_DATA_EDIT

Model > Effects

Class and quantitative variables: GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETOTAL, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE

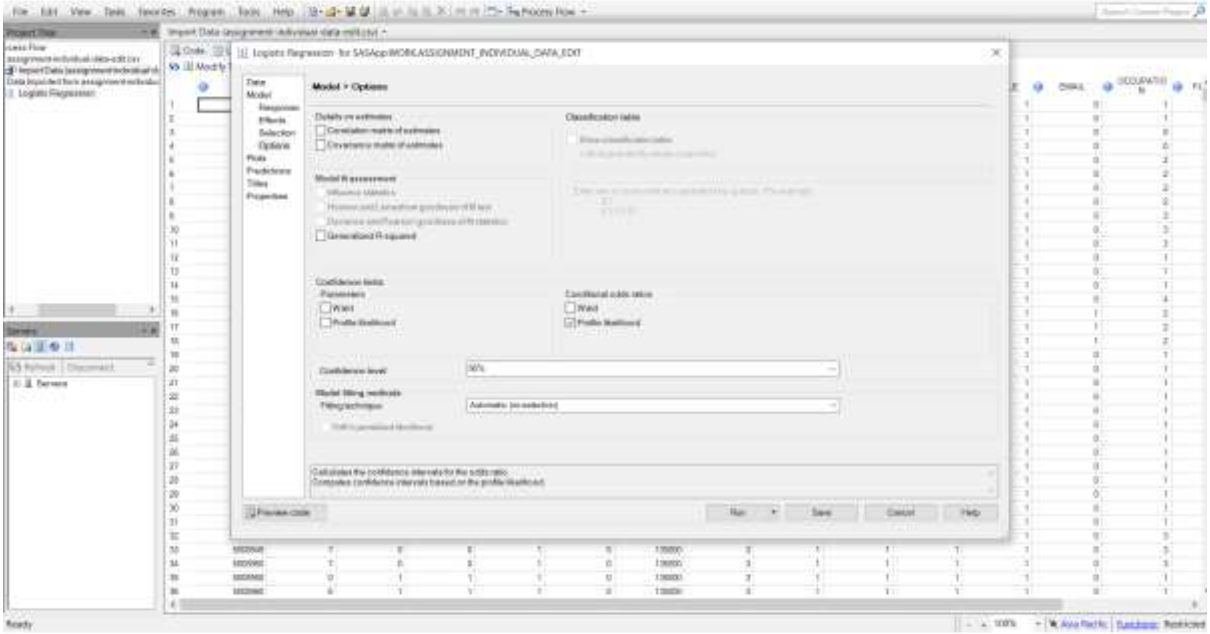
Effects: GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETOTAL, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE

Factorial Degrees: 13

Polynomial Degrees: 2

☑ Include intercept

Assigning the variables as Main Effects in the model.


Logistic Regression for SASApp:WORK.ASSIGNMENT_INDIVIDUAL_DATA_EDIT

Model > Options

Details on estimates

□ Correlation matrix of estimates

□ Covariance matrix of estimates

Classification table

Model fit assessment

□ Generalized R-squared

Confidence limits

Parameters

□ Wald

□ Profile likelihood

Conditional odds ratios

□ Wald

☑ Profile likelihood

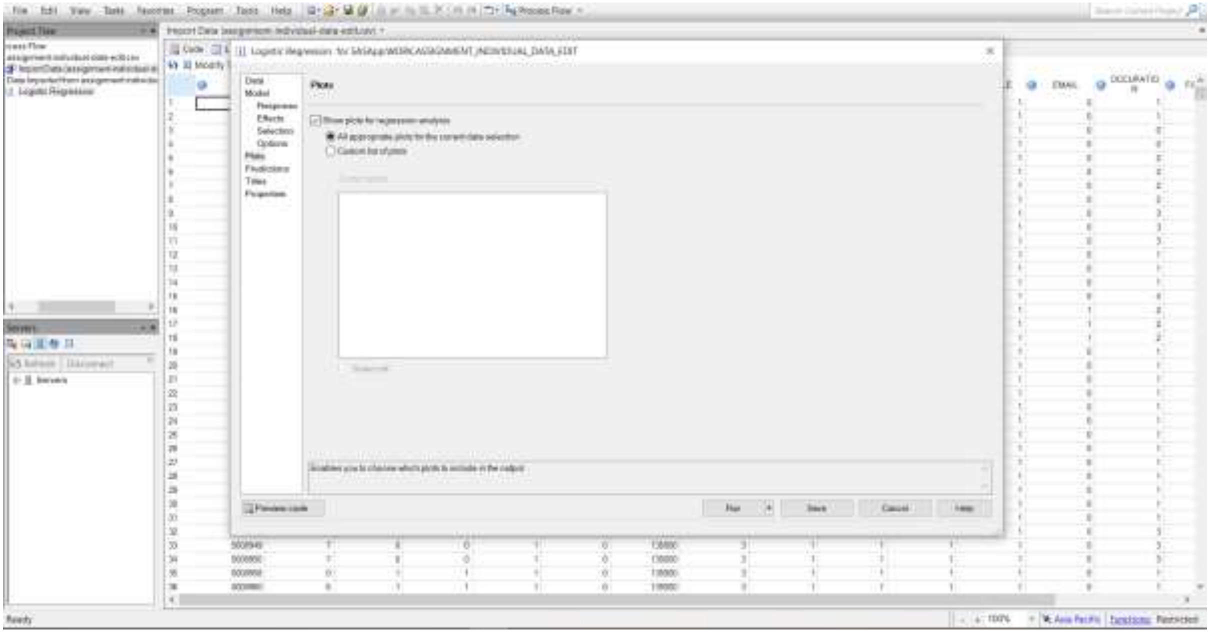
Confidence level: 95%

Model fitting methods

Fitting technique: Automatic (no selection)

Calculates the confidence intervals for the odds ratio.
Computes confidence intervals based on the profile likelihood.

Checking Profile Likelihood under Conditional Odds Ratio


Logistic Regression for SASApp:WORK.ASSIGNMENT_INDIVIDUAL_DATA_EDIT

Plots

☑ Show plots for regression analysis

● All appropriate plots for the current data selection

○ Custom list of plots

Enables you to choose which plots to include in the output

Ensuring the plots are generated

Categorical Data Analysis – Table Analysis

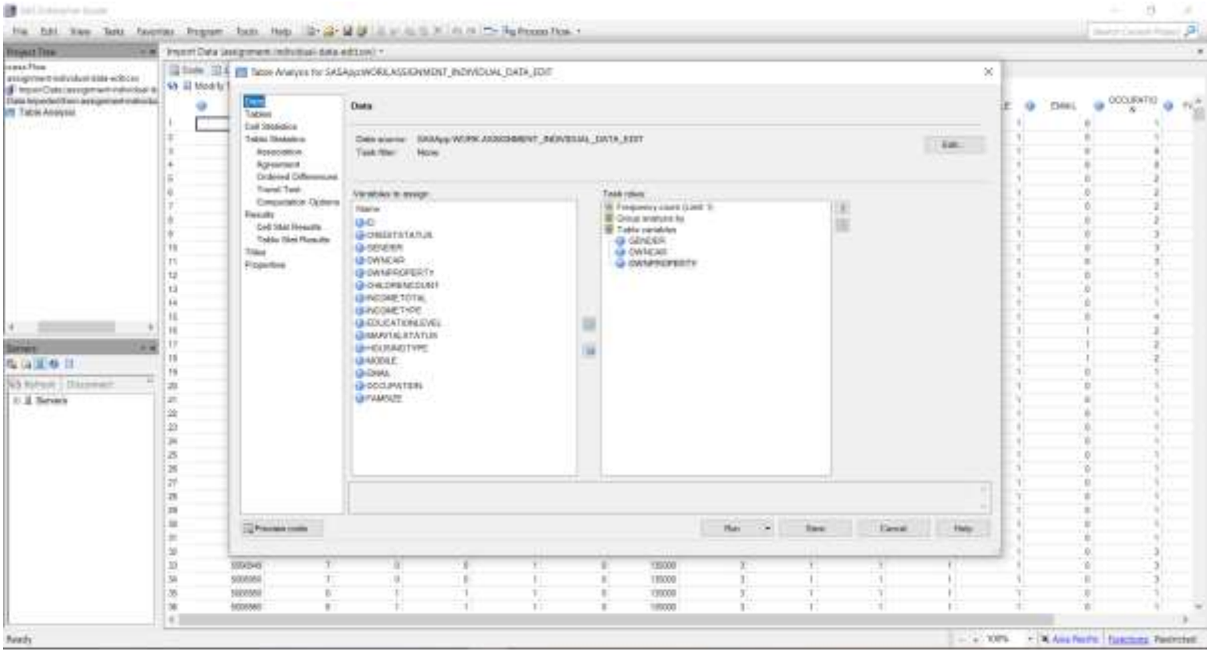


Table Analysis for SASApp:WORK.ASSIGNMENT_INDIVIDUAL_DATA_EDIT

Data

Data source: SASApp:WORK.ASSIGNMENT_INDIVIDUAL_DATA_EDIT
Task filter: None

Variables to assign: ID, CREDITSTATUS, GENDER, OWNCAR, OWNPROPERTY, CHILDRENCOUNT, INCOMETOTAL, INCOMETYPE, EDUCATIONLEVEL, MARITALSTATUS, HOUSINGTYPE, MOBILE, EMAIL, OCCUPATION, FAMSIZE

Task roles:
- Frequency count (Limit: 1)
- Group analysis by
- Table variables
  - GENDER
  - OWNCAR
  - OWNPROPERTY

Assigning the variables to Table Variables

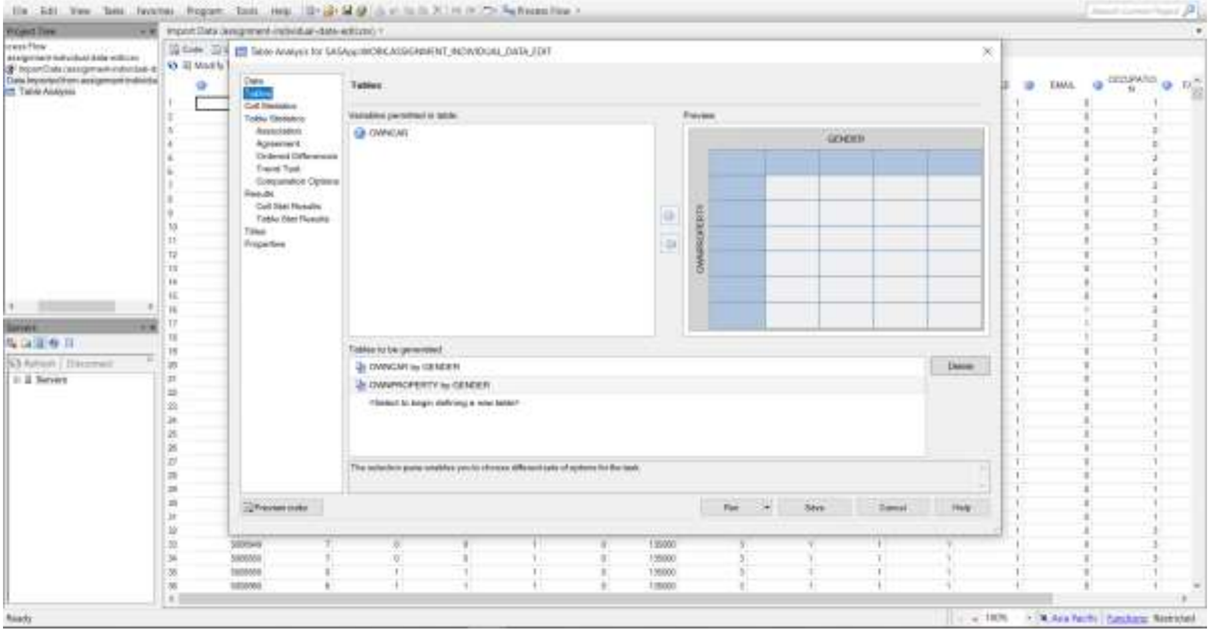


Table Analysis for SASApp:WORK.ASSIGNMENT_INDIVIDUAL_DATA_EDIT

Tables

Variables permitted in table: OWNCAR

Preview: GENDER by OWNPROPERTY

Tables to be generated:
- OWNCAR by GENDER
- OWNPROPERTY by GENDER
- <Select to begin defining a new table>

The selection pane enables you to choose different sets of options for the task.

Assigning the variables according to their roles

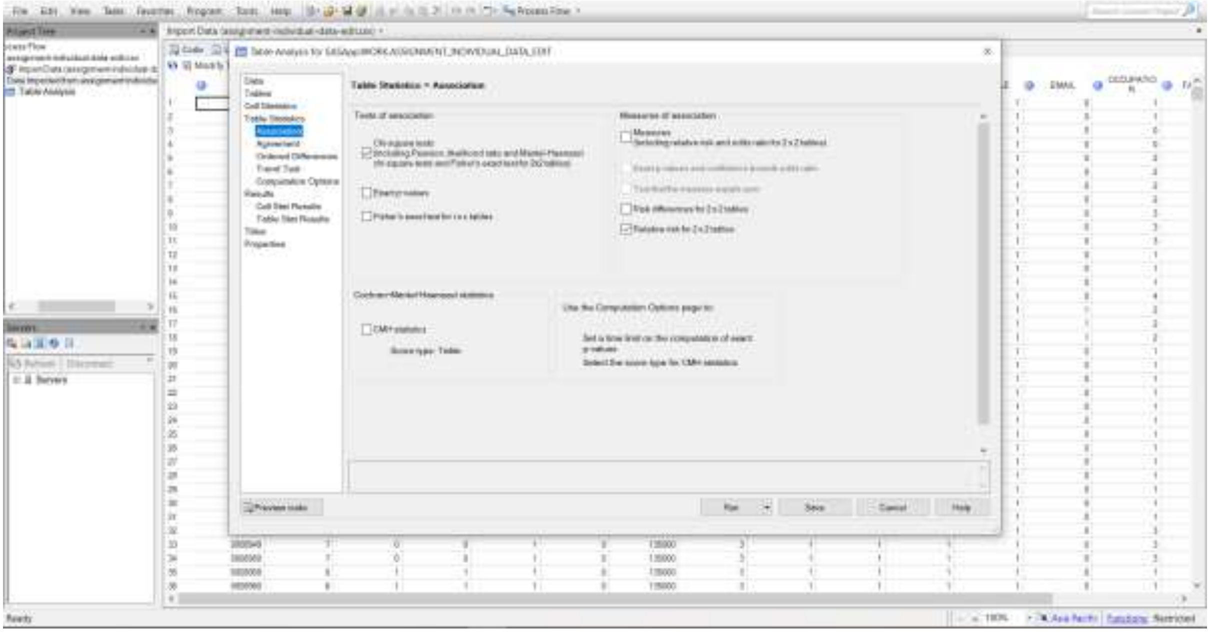


Table Analysis for SASApp:WORK.ASSIGNMENT_INDIVIDUAL_DATA_EDIT

Table Statistics > Association

Tests of association

☑ Chi-square tests (including Pearson, likelihood ratio and Mantel-Haenszel chi-square tests and Fisher's exact test for 2x2 tables)

□ Exact p-values

□ Fisher's exact test for r x c tables

Measures of association

□ Measures (including relative risk and odds ratio for 2 x 2 tables)

Exact p-values and confidence bounds odds ratio

Test that the measures equals zero

□ Risk differences for 2 x 2 tables

☑ Relative risk for 2 x 2 tables

Cochran-Mantel-Haenszel statistics

□ CMH statistics

Score type: Table

Use the Computation Options page to:

Set a time limit on the computation of exact p-values

Select the score type for CMH statistics

Checking on Row Percentage and unchecking Column Percentage