SCHOOL OF SCIENCE AND TECHNOLOGY

ASSIGNMENT / PROJECT SUBMISSION FORM

PROGRAMME:

SEMESTER: Mar 2020

SUBJECT: IST2024 Applied Statistics

DEADLINE: 23rd 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....Ng Wei Xiang..... (Student's Name) .....18033167..... (Student ID) received the assignment and read the comments.

.....*Xiang*....16/7/2020..... (Signature/Date)

Academic Honesty Acknowledgement

"I .....Ng Wei Xiang.....(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."

…..... *Xiang*..... (Student's signature / Date)

Data Protection

The protection of personal data is an important concern to Sunway University and any personal data collected on this form will be treated in accordance with the Personal Data Protection Notice of the institution.
http://sunway.edu.my/pdpa/notice_english (English version)
http://sunway.edu.my/pdpa/notice_bm (Malay version)

Table of Contents

1.Introduction..... 2
2.Descriptive Analysis 3
2.1 ID..... 3
2.2 Credit Status..... 3
2.3 Gender 4
2.4 Own Car..... 4
2.5 Own Property 4
2.6 Children Count 5
2.7 Income Total 6
2.8 Income Type 7
2.9 Edu Level..... 7
2.10 Marital Status 8
2.11 Housing Type..... 8
2.12 Mobile..... 9
2.13 Email..... 9
2.14 Occupation 10
2.15 FamSize 11
3.Objectives……..……12
3.1 Objective 1..... 12
3.2 Objective 2..... 15
3.3 Objective 3..... 20
4.Conclusion...……..……24
4.References...……..……26

1. Introduction

A dataset that consists of records on customer from a bank was presented for this assignment. Since it was about customer records from a bank, I decided to handle this dataset from a perspective as a big data analyst working for the bank trying to analyse patterns and relationship hidden in the dataset. According to Hitachi Solutions, big data in banking provides advantages such as “gaining complete view of customer profile”, “tailors customer experience”, “understands customers’ buying pattern”, “identify opportunities”, and “reduce risk of fraudulent behaviour”. Thus, my approach in handling the dataset will be harvesting the dataset and prove the advantages. This leads to the main objectives that is in the area of my interests :

Objective 1 : What variables will contribute and affect the annual income of an individual
Objective 2 : The behaviour of an individual that has a good credit status
Objective 3 : Tailoring product to each customer based on their affordability

The given dataset has 15 variables which were grouped as followed :
Continuous variables were “ChildrenCount” and “IncomeTotal”, ”FamSize”.
Categorical variables were “Gender”, OwnCar”, “OwnProperty”,“IncomeType”, “EducationLevel”, “MaritalStatus ”, “HousingType”, “Mobile”, “email”, “Occupation”, “CreditStatus”.
Unused variable “ID”.

My approach for handling this dataset will be as followed :

1. Perform **descriptive analysis** on the variables to better understand them
2. Perform **ANCOVA** and Perform **ANCOVA** and **ANOVA** to test for relationship between TotalIncome and other variables
3. Perform **logistics regression** to test for relationship between CreditStatus against other variables.
4. Perform **multiple linear regression** on FamSize to provide an idea on how tailoring product to each customer can be carried out.

2. Descriptive Analysis

Descriptive analysis was performed to better understand the behaviour and pattern of the given variables in the dataset. There are 3 continuous variable and 12 categorical variables identified. Microsoft Excel was used to generate a pie chart for analysis for the categorical variables and SAS Studio was used to perform univariate analysis for continuous variables.

1. **ID**

   The variable ID is neither a response or predictor variable in this data set since it does not carry any meaningful value. Thus, it is not analysed and interpreted but rather used as an indicator only.

2. **Credit Status**

   In this dataset, credit loan is a categorical variable. The credit loan status of an individual was represented as CREDITSTATUS in the dataset. The indication was as followed :

   0 : 1 – 29 days past due
   1 : 30 – 59 days past due
   2 : 60 - 89 days past due
   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

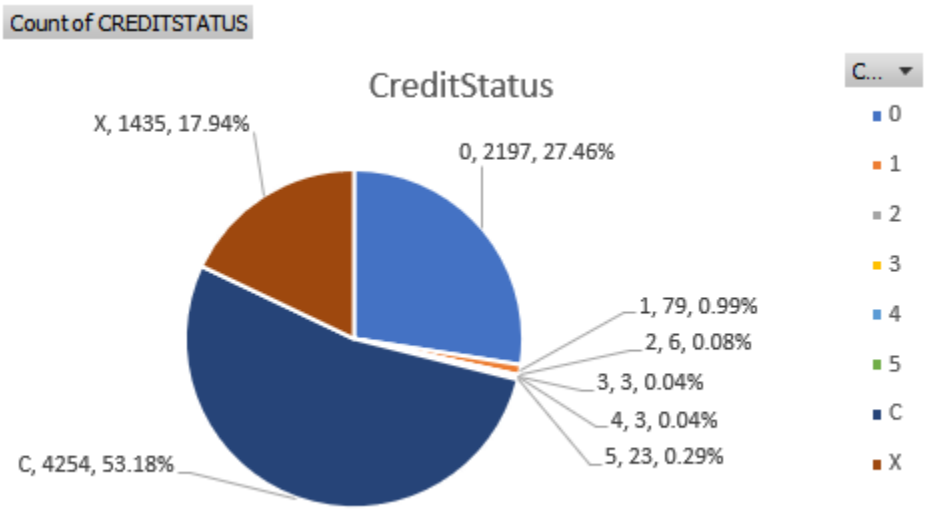


Figure 1 : Pie Chart representation of Credit Status

From the generated pie chart, it is seen that most of the individuals at 53.18% or 4254 of the them had paid off their loans for that month. It is followed by 27.46% or 2197 of them had their loan unpaid 1 to 29 days past due. Next, 17.94% or 1435 of them had no loan for the month. 0.99% or 79 of the individuals had their loan unpaid 30 to 59 days past due. 0.29% or 23 of them had bad debts. 6 of them at 0.08% had their loan unpaid for 60 to 89 days past due. Finally, there are 3 (0.04%) of them each had their loans overdue for 90 to 119 days and 120 to 149 days overdue respectively.

Overall, the credit status of the overall can be considered as great as more than half of them had paid off their loans for the month.

3. **Gender**

In this dataset, gender is a categorical variable. The indicator for Female is **F** and Male is **M.**

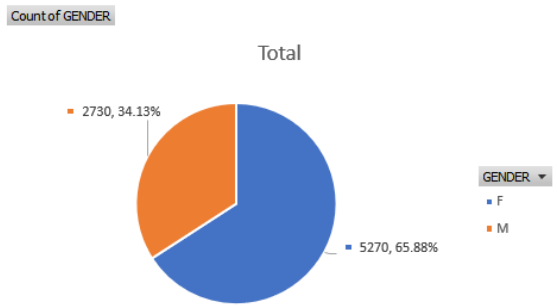


Figure 2 : Pie Chart representation for Gender

It is seen that this dataset consists mainly of Female at 65.88% or 5270 of them. Additionally, 34.13% or 2730 of them are Male.

4. **OwnCar**

In this dataset, OwnCar is a categorical variable. It means whether the individual owns a car with a simple indication of **Y** for Yes and **N** for No.

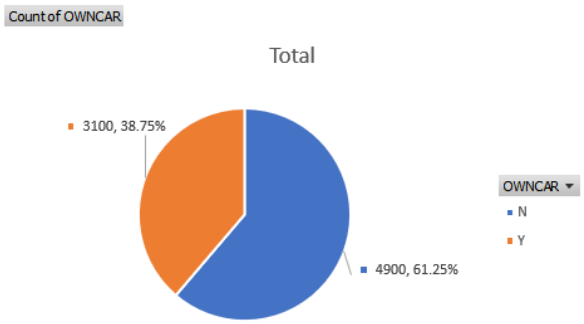


Figure 3 : Pie Chart representation for OwnCar

It seen from this dataset that the majority at 61.25% or 4900 of them do not own a car. Only 38.75% or 3100 of them owns a car.

5. **OwnProperty**

In this dataset, OwnProperty is also a categorical variable. It means that whether the individual owns a property with a simple indication of **Y** for Yes and **N** for No.

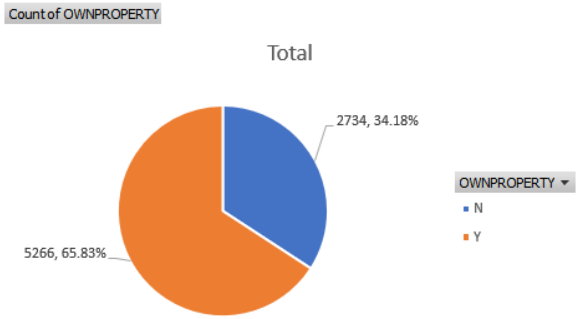


Figure 4 : Pie Chart representation for OwnProperty

It was obvious that the majority of them owns a property at 65.83% or 5266 people. 34.18% or 2734 of them in the dataset do not own a property.

6. ChildrenCount

In this dataset, the number of children was represented as ChildrenCount which is a continuous variable. Thus, a univariate analysis was performed using SAS studio to analyse it.

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

Quantiles (Definition 5)	
Level	Quantile
100% Max	14
99%	3
95%	2
90%	2
75% Q3	1
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

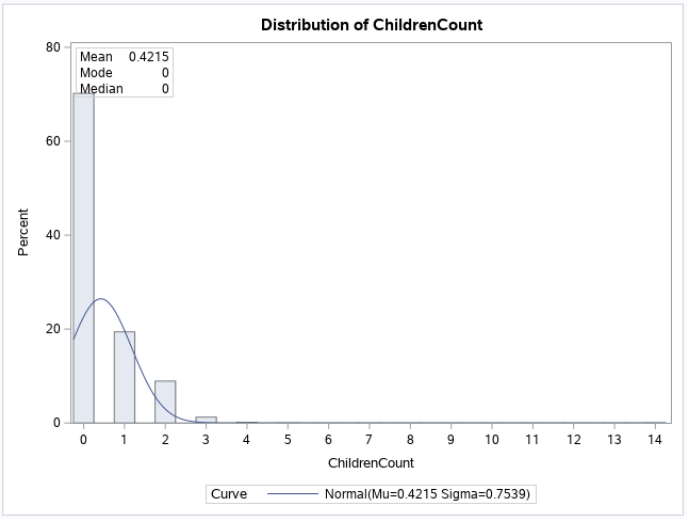


Figure 5 : Univariate analysis output of Children Count

From the generated output, the **Mean** is 0.4215 which means that on average an individual will have 0.4 child. However, number of children cannot be in decimal places thus after rounding up to the closest unit, it will be interpret as the individuals have no children on average. **Median** is at 0 children and **Mode** is also at 0 children. This means that most of the individuals in this dataset has no children. In this dataset, the **Variance** is 0.5684, **Standard Deviation** at 0.75393, with a **Range** of 14 children, and **Interquartile Range** at 1 child.

The value of **Skewness** for children count in this dataset is 2.99. It is a strong positive skewed distribution. This means that its distributed heavily on the right and having a light tail as per shown in the graph in Figure 5.The value for **Kurtosis** is 28.1628. It is a positive kurtosis and can be interpreted that the distribution has heavier tails and sharper peak than the normal distribution as shown in Figure 5.

The most number of children available in this dataset is at 14 children and the least is 0.

7. IncomeTotal

In this dataset, the annual income was represented as IncomeTotal which is also a continuous variable. Thus, a univariate analysis was carried out to analyse on it.

The UNIVARIATE Procedure
Variable: IncomeTotal

Moments			
N	8000	Sum Weights	8000
Mean	185712.847	Sum Observations	1485702779
Std Deviation	100015.406	Variance	1.00031E10
Skewness	2.46936945	Kurtosis	14.3340271
Uncorrected SS	3.55929E14	Corrected SS	8.00146E13
Coeff Variation	53.8548663	Std Error Mean	1118.20623

Basic Statistical Measures			
Location		Variability	
Mean	185712.8	Std Deviation	100015
Median	157500.0	Variance	1.00031E10
Mode	135000.0	Range	1548000
		Interquartile Range	105075

Quantiles (Definition 5)	
Level	Quantile
100% Max	1575000
99%	540000
95%	360000
90%	315000
75% Q3	225000
50% Median	157500
25% Q1	119925
10%	90000
5%	72000
1%	49500
0% Min	27000

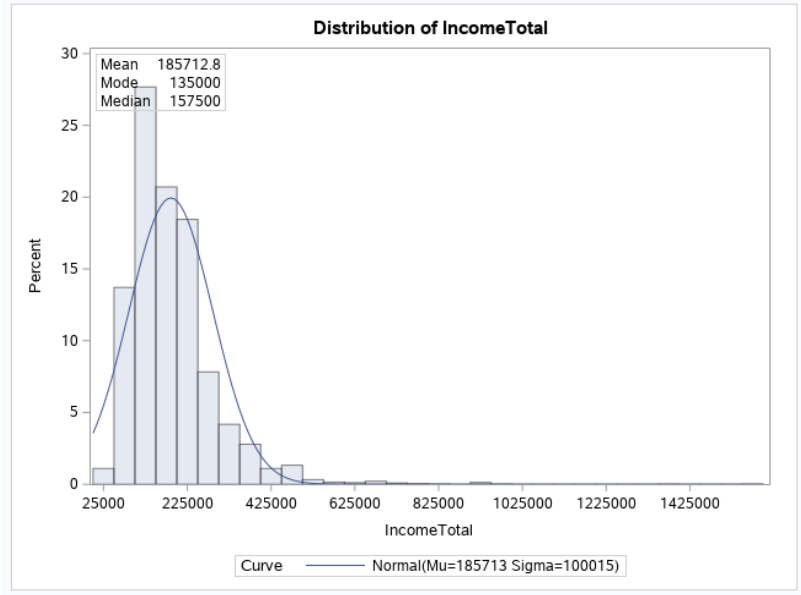


Figure 6 : Univariate analysis output of IncomeTotal

From the generated output, the **Mean** is 185712.847 which means that on average, an individual earns 185712.847 annually. The **Median** salary is 157500. The **Mode** is 135000 which means that most of the individuals in this dataset has an annual income of 135000. The **Variance** of annual income is at 1.00031E10. This is followed by a **Standard Deviation** of 100015.406, a **Range** of 1548000, and **Interquartile Range** of 105075.

The value of **Skewness** for annual income is 2.469. This shows a strong positive skewed distribution. This means that it is distributed heavily on the right and having a light tail as per shown in Figure 6. The value for **Kurtosis** is 14.33. This shows a positive kurtosis that is having a sharp peak and light tails than the normal distribution.

The customer with the highest and lowest annual income is at 1575000 and 27000 respectively.

8. IncomeType

In this dataset, IncomeType is a categorical variable. It means the category of income of that individual. We had 5 categories, which were commercial associate, pensioner, state servant, student, and working.

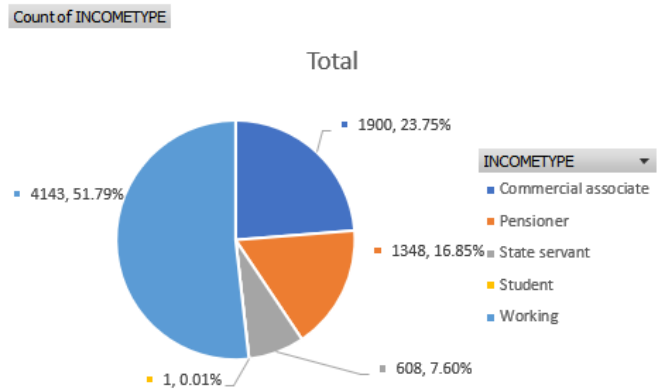


Figure 7 : Pie Chart representation for IncomeType

From the generated pie chart, we can identify that the majority 4143 or 51.79% of the individuals in the dataset has an income through working. Next, 1900 or 23.75% of them has an income of being a commercial associate. 1348 or 16.85% of the individuals is a pensioner and 608 or 7.6% of them is a state servant. Only, 1 of them in the dataset is a student taking 0.01% overall.

9. EducationLevel

In this dataset, EducationLevel is also a categorical variable. It means the level of education of the individual. We had 5 levels which are academic degree, higher education, incomplete higher, lower secondary, and secondary/secondary special.

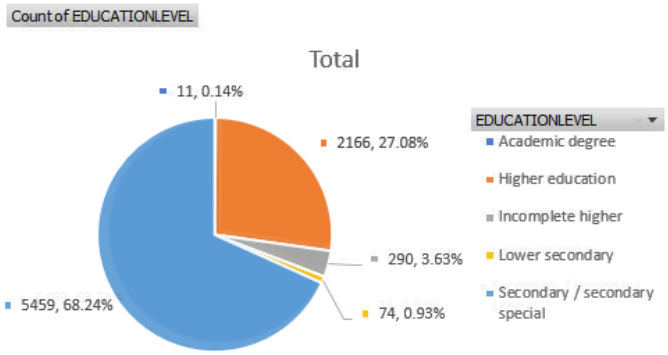


Figure 8 : Pie chart representation for EducationLevel

From the generated pie chart, it is obvious to see that most of the individuals had an education level up till secondary/secondary special at 5459 or 68.24% of them. This is followed by education level that is up to higher education at 2166 or 27.08% of them. Then, 290 or 3.63% of them had incomplete higher education. 74 of them at 0.93% has an education level till lower secondary. Only 11 or 0.14% of the individuals had an education level that is up to an academic degree.

10. MaritalStatus

In this dataset, MaritalStatus is a categorical variable. It is meant by the marital status of that individual in which there are 5 categories : Civil marriage, married, separated, single/not married, and widow.

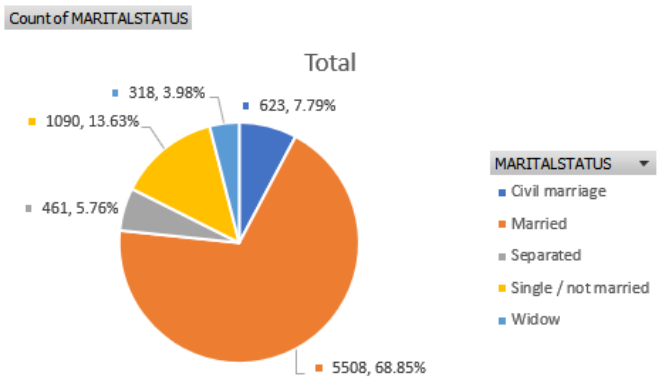


Figure 9 : Pie chart representation of MaritalStatus

Based on the generated pie chart, we are able to identify that there are 68.85% or 5508 of them are married. Next, 13.63% or 1090 of them are still single or not married. 7.79% or 623 of them had a marital status of civil marriage while 5.76% or 461 of them are separated. Lastly, the minority in this dataset had a marital status of being widow which consists of 318 or 3.98% of them.

11. HousingType

In this dataset, HousingType is also a categorical variable. It is meant by the way of living or the type of house they are currently living in. There are 6 categories for HousingType which consists of : co-op apartment, House/apartment, Municipal apartment, Office apartment, Rented apartment, and with parents.

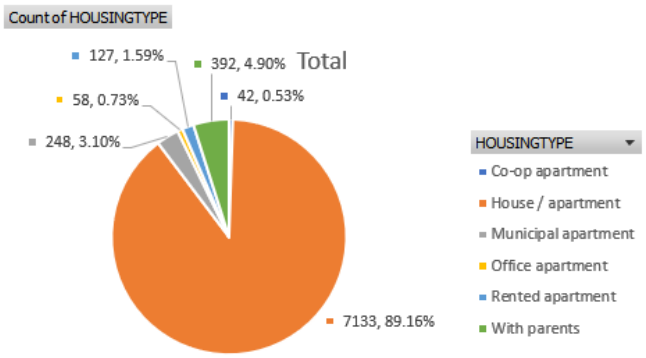


Figure 10 : Pie Chart representation of HousingType

The big majority of the individuals in this dataset are living in a house/apartment which makes up 7133 or 89.16% of them. This is followed by 392 or 4.9% of the individuals are living with their parents. Then, 248 or 3.10% of the individuals live in a municipal apartment. 127 or 1.59% of them live in a rented apartment, 58 or 0.73% of them live in an office apartment. Only 42 persons taking up 0.53% of the whole dataset lives in a co-op apartment.

12. Mobile

In this dataset, Mobile is also one of the categorical variable. It is identified whether the individual owns a mobile phone. The indicator for it is 1 for Yes and 0 for No.

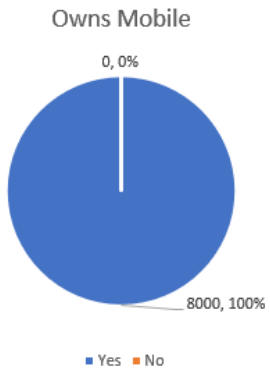


Figure 11 : Pie Chart representation of Mobile

All 8000 of the individuals or 100% of them do own a mobile phone in this dataset.

13. Email

In this dataset, Email is a categorical variable. It meant whether the individual has an email. The indicator in the dataset is 1 for Yes and 0 for No.

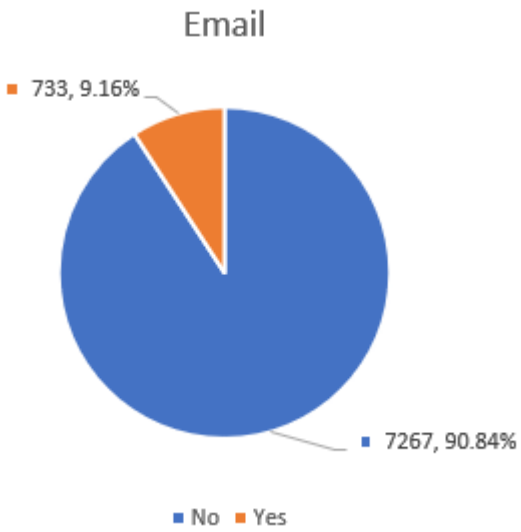


Figure 12 : Pie chart representation for Email

It is obvious to see that 7267 or 90.39% of the individuals do not own an email taking up the majority. Only 733 or 9.61% of them had an email address.

14. Occupation

In this dataset, Occupation is also a categorical variable. The meaning is self-explanatory based on its name, occupation. There are 19 different categories of occupation is found in this dataset which were : Accountants, Cleaning staff, Cooking staff, Core staff, Drivers, High Skill Tech Staff, HR staff, IT staff, Laborers, Low-skill laborers, Managers, Medicine Staff, Private Service Staff, Realty Agents, Sales Staff, Secretaries, Security Staff, Waiters/barmen staff and Others. 1 assumption is made in this dataset which was replacing the emptied occupation by “Others”.

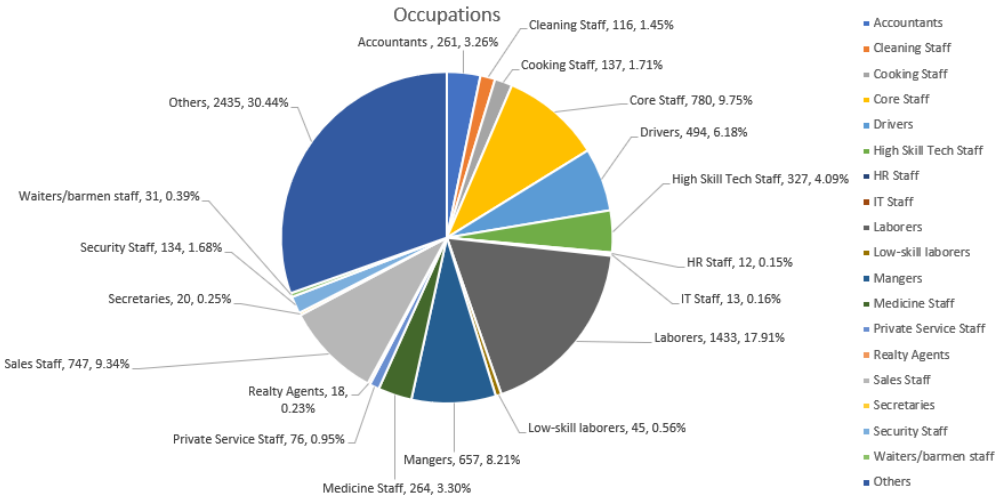


Figure 13 : Pie chart representation of Occupations

Firstly, we can see that the majority in this dataset refuse to disclose their occupation which was then labelled as “Others” taking up 30.44% or 2435 of them overall. There are 1433 or 17.91% of them are laborers, 780 or 9.75% of them are core staff and 747 or 9.34% of them are sales staff. There are 657 managers taking up 8.21%, 494 drivers taking up 6.18%, and 327 High skill tech staff taking up 4.09%. 3.3% or 264 of them are medicine staff, 3.26% or 264 of them are accountants, and 1.71% or 137 of them are cooking staff. In the dataset, there are 1.68% or 134 of them are cooking staff, 1.45% or 116 of them are cleaning staff, 0.95% or 76 of them are 76 are private service staff. Then, 0.56% or 45 of them works as low-skill laborers, 0.39% or 31 of them works as waiter/barmen staff, and 0.25% or 20 of them works as secretaries. Finally, into the least 3 occupations were 18 Realty agents, 13 IT staff, and 12 HR staff which is 0.23%, 0.16% and 0.15% respectively from the dataset.

15. FamSize

In this dataset, FamSize is a continuous variable. It was meant by the size of the family of the individual. Due to its nature of being a continuous variable, an univariate analysis was conducted to better understand it.

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

Quantiles (Definition 5)	
Level	Quantile
100% Max	15
99%	5
95%	4
90%	3
75% Q3	3
50% Median	2
25% Q1	2
10%	1
5%	1
1%	1
0% Min	1

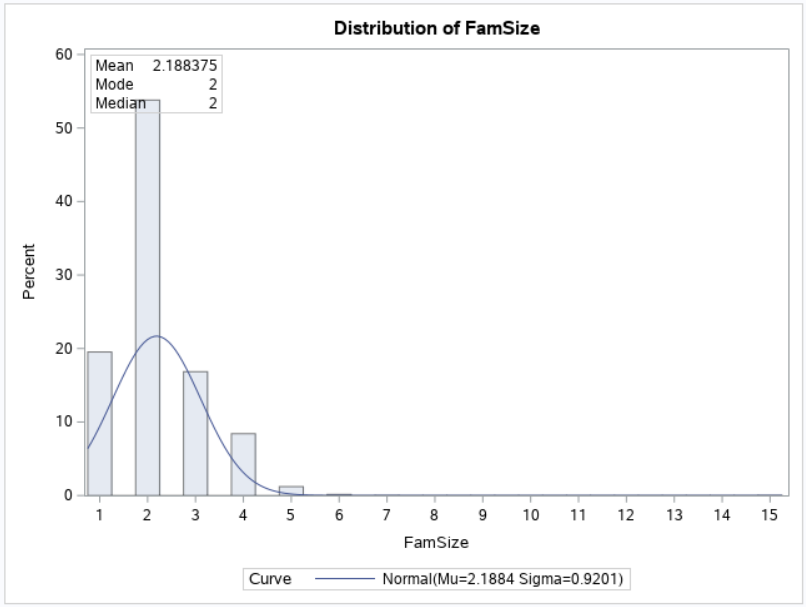


Figure 14 : Univariate analysis for FamSize

Based on the SAS Studio output, we can see that FamSize has a **Mean** of 2.188 which means that an average the family size in this dataset is 2.188 persons. It has a **Median** of 2 and a **Mode** of 2 which means that most of the individuals had a family size of 2.

The Standard Deviation is 0.92, followed with a **Variance** of 0.8464, a **Range** of 14 and **Interquartile range** of 1. It is also seen that the **Minimum** family size is 1 and the **Maximum** is 15 person.

The **Skewness** value is 1.4955 and is a strong positive skewed distribution. This means that it is distributed heavily on the right and having light tails. The **Kurtosis** value is 10 which also reflect a sharp peak and light tail in the distribution chart as shown in Figure 14.

3. Analysis

The analysis of the dataset is conducted from the perspective of meeting the objective listed in introduction. A list of analysis methods will be carried out using SAS Studio to provide the output and justification will be provided to analyse the relationship.

i. Objective 1 - What variables will affect with the annual income of an individual and their relationship in it.

In this objective, I am interested to find out any of the other 13 variables will contribute or affect an individual's annual income and if there is, what is the relationship and strength of them. Thus, the annual income will be my response variable and the remaining will be the predictor variables. The predictor variables to be loaded for analysis consists of both continuous and categorical which is why **Analysis of Covariance(ANCOVA)** is used**.** Before we start our ANCOVA test, it is always a good practice to check for the assumptions to ensure a good model.

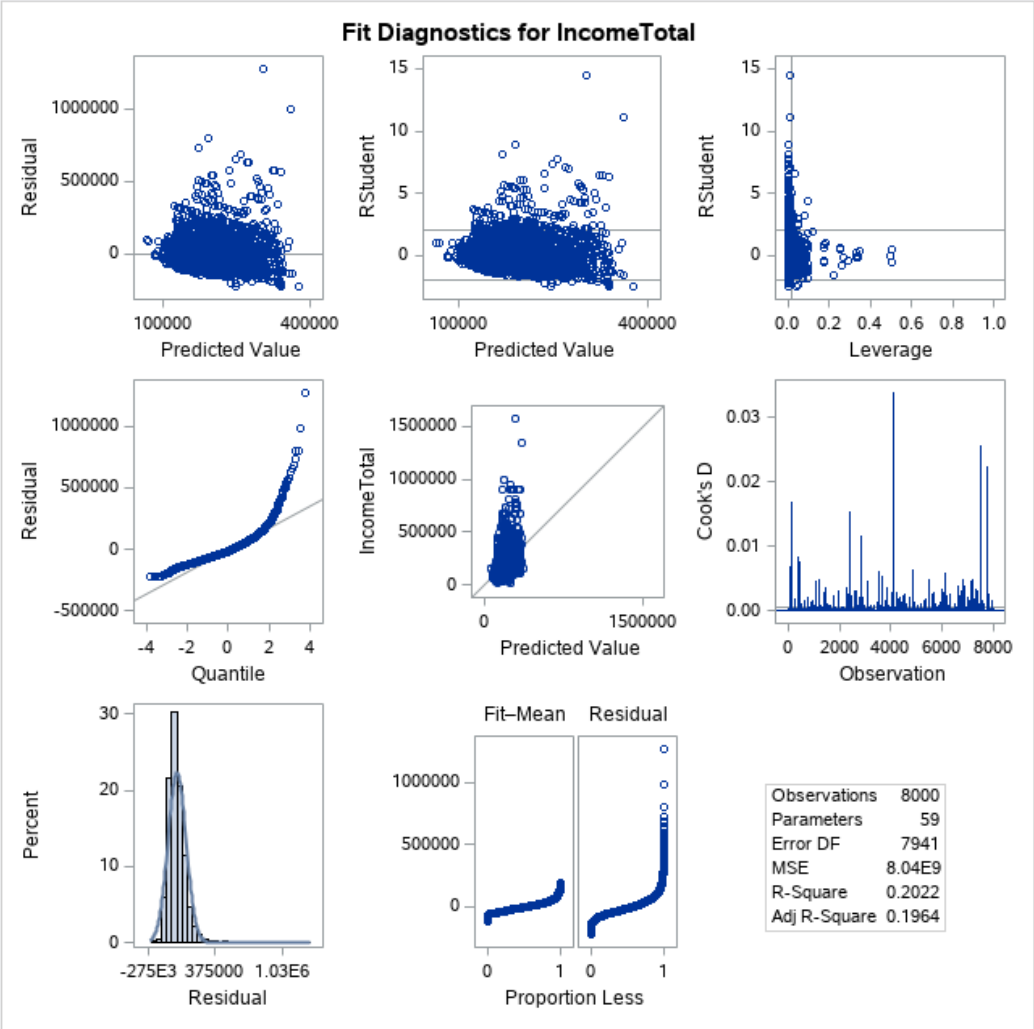


Figure 15 : Diagnostics plot to check for assumptions of ANCOVA

The 3 main assumptions in ANCOVA assumes that all observations are independent, data are normally distributed, and homogeneity of variances. From the residual against predicted value plot, it violates the assumption of observations are independent. Then, The Quantile(QQ) plot do not show a linear line which also fails the assumption of ANCOVA in normally distributed data. The histogram has a strong head to the right which do not show equal variance. Although the model does not seem to meet the assumptions of ANCOVA, but our dataset consists of over 1000 data which should not be a factor that is hindering us from reaching the objective.

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
Email	1	349607275461	349607275461	43.49	<.0001
Occupation	18	3.5514429E12	197302384369	24.54	<.0001
FamSize	6	60544705380	10090784230	1.26	0.2746

Figure 15 : Output of ANCOVA from SAS Studio

Before proceeding for the analysis of the output, a hypothesis is made.
H_0 *: The mean annual income for all the variables are the same*
H_1 *: The mean annual income for at least one of the variables is different*

Based on the output, we obtain a few key values of the analysis of annual income against other variables. Firstly, we obtained a **F Value** of 34.70 followed by a **P-value** of <0.0001. Thus, H_0 is rejected and we conclude that there is a different in the mean annual income for at least one of the variables. Then, we obtained a **mean** value of 185712.8 for IncomeTotal which means that on average a person earns 185712.80 per annum. The **R-Square** value shows 0.2022 which is interpreted as the predictor variables explains about 20.22% of the variability of annual income.

Loading all the variables and considering them do not provide a good model which is why we should looking into the effect of each individual variable against annual income to filter out less influential variables and making a better model. Another set of hypothesis is made in order to proceed for analysis.

H_0 *: There is no significant relationship between the variable against annual income*
H_1 *:There is a significant relationship between the variable against annual income.*

Variables that are above the significance level of 0.05 should be eliminated. Based on the output, we are able to identify that there are 4 variables which is above the significance level. They are CreditStatus, ChildrenCount, MaritalStatus and FamSize. We are able to conclude that the variables CreditStatus, ChildrenCount, MaritalStatus and FamSize had no significant relationship against the annual income at the significance level of 0.05.

Our variables for the equation to analyse against annual income is now left with Gender, OwnCar, OwnProperty, IncomeType, EducationLevel, Email, HousingType, and Occupation. However, it is seen that the remaining variables are all categorical, thus an ANOVA test is conducted to investigate their relationship with incomeTotal.

The GLM Procedure
Least Squares Means

Gender	IncomeTotal LSMEAN
F	170894.910
M	214317.437

The GLM Procedure
Least Squares Means

OwnCar	IncomeTotal LSMEAN
N	168850.605
Y	212366.069

The GLM Procedure
Least Squares Means

OwnProperty	IncomeTotal LSMEAN
N	179888.188
Y	188736.892

The GLM Procedure
Least Squares Means

IncomeType	IncomeTotal LSMEAN
Commerci	215926.032
Pensione	148538.634
State se	198848.931
Student	171000.000
Working	182028.021

The GLM Procedure
Least Squares Means

EducationLevel	IncomeTotal LSMEAN
Academic	229090.909
Higher e	223509.436
Incomple	199081.552
Lower se	138593.919
Secondar	170557.197

The GLM Procedure
Least Squares Means

Occupation	IncomeTotal LSMEAN
Accounta	204008.621
Cleaning	147122.845
Cooking	141684.307
Core sta	188787.179
Drivers	205514.271
HR staff	154875.000
High ski	192576.606
IT staff	227423.077
Laborers	182214.879
Low-skil	125700.000
Managers	279311.918
Medicine	155001.136
Others	167814.606
Private	203092.105
Realty a	232750.000
Sales st	170381.072
Secretar	181575.000
Security	178754.104
Waiters/	164612.903

The GLM Procedure
Least Squares Means

HousingType	IncomeTotal LSMEAN
Co-op ap	161250.000
House /	185725.944
Municipa	179769.556
Office a	204323.276
Rented a	218636.220
With par	178435.515

Figure 17 : least square means of variables having significant relationship

First and foremost, it is seen that Male has a higher mean annual income as compared to female. Then, those individuals with car has a higher annual income than those without. Individuals that owns a property generally had a slightly higher annual income. Those individuals whom having an income type of commercial associate also tend to have a higher annual income. With an academic degree, the individuals in the dataset have a higher mean for annual income. When it comes to housing type, those with a rented apartment also tend to have a higher annual income. Lastly, the individuals that has an occupation of being a manger have the highest mean annual income as compared to others.

As conclusion for the analysis of this objective, 2 main statistical techniques which are ANCOVA and ANOVA were conducted to achieve the objective. After loading in all the variables, 7 variables are successfully narrowed down from the original 13 of them after performing an analysis on ANCOVA. This provides a clearer picture into which few variables exactly have an effect on annual income. They were gender, owncar, ownproperty, incometype, educationlevel, housingtype and occupation. Then, ANOVA is conducted to further investigate the relationship of these variables against the annual income. The relationship of variables in this dataset in having a higher mean annual income are as followed : is a male, having own car, having own property, an income type of commercial associate, completed an academic degree, currently living in a rented apartment, and an occupation as manager.

ii. Objective 2 – The relationship of credit status with other variables

This objective is in the area of interest of mine to understand the relationship of credit status with other variables. This is especially important so that the bank do not approve high amount loan to individual with bad credit status easily. Risk can greatly be reduced. Thus, based on the given dataset, credit status will be the response variable. A logistic regression is executed on SAS Studio to meet the objective.

Model Information	
Data Set	WORK.CREDITSTATUS2
Response Variable	CreditStatus
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	CreditStatus	Total Frequency
1	0	2197
2	1	79
3	2	6
4	3	3
5	4	3
6	5	23
7	C	4254
8	X	1435

Figure 18 : Model Information and Response Profile Table

It is seen that the categorical response variable CreditStatus has 8 levels. Based on the reponse profile table, we can see that creditStatus of “C” which is “paid off for the month” has the highest frequency.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	52.9715	52	0.4364
Score	50.6521	52	0.5270
Wald	44.1900	52	0.7708

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Gender	1	4.1896	0.0407
OwnCar	1	0.0266	0.8705
OwnProperty	1	1.3767	0.2407
ChildrenCount	6	3.3085	0.7692
IncomeType	4	3.6674	0.4529
IncomeTotal	1	0.5696	0.4504
EducationLevel	4	4.2577	0.3723
MaritalStatus	4	8.1889	0.0849
HousingType	5	3.1309	0.6798
Mobile	0	.	.
Email	1	2.3391	0.1262
Occupation	18	9.4274	0.9490
FamSize	6	3.0763	0.7992

Figure 19 : Global Null Hypothesis and Analysis of Effects table

A set of hypothesis is made to test for the collective significance of predictor variables.

H_0 : All the regression coefficients are 0
H_1 : At least one of the regression coefficients is not 0.

Based on the output of the global null hypothesis table, H_0 is not rejected since all the p-values are more than the 0.05 significance level. Thus, we conclude that the predictor variables are not significant collectively and is not suitable to predict the model for this objective. However, we continue the analysis on the model to further identify any details that could help figure out any relationship with credit status even though the model doesn't seem to be useful. Thus, a type 3 analysis of effects table is generated followed by another set of hypothesis for individual variable analysis.

H_0 : The predictor variable has no significant effect to the response variable
H_1 : The predictor variable has a significant effect to the response variable

Based on the type 3 analysis of effects table, only the variable Gender had a p-value of lesser than the significance level of 0.05, at 0.0407. Only gender will be able to reject the null hypothesis. All the other variables failed to meet the significance level which explains and proved the analysis from the global null hypothesis table. We conclude that for this dataset, the only predictor variable that has a significant effect to creditStatus is gender.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	0	1	-0.9874	0.0262	1425.5471	<.0001
Intercept	1	1	-0.9384	0.0259	1313.5125	<.0001
Intercept	2	1	-0.9347	0.0259	1305.1385	<.0001
Intercept	3	1	-0.9329	0.0259	1300.9590	<.0001
Intercept	4	1	-0.9310	0.0259	1296.7845	<.0001
Intercept	5	1	-0.9170	0.0258	1264.9466	<.0001
Intercept	C	1	1.5055	0.0299	2536.0732	<.0001
Gender	F	1	0.0495	0.0225	4.8275	0.0280

Figure 20 : Analysis of maximum likelihood estimates table

A simple logistic regression equation was constructed based on the variables that has a significant effect based on the analysis of maximum likelihood estimates table.

$\text{Logit}(\widehat{\pi}) = \beta_0 + \beta_1 * X_{\text{Female}}$

$\text{Logit}(\widehat{\pi}) = -4.1359 + 0.0495 * X_{\text{Female}}$

Since the only variable that has a relationship with creditStatus is gender also appears to be a categorical variable, it would be appropriate to conduct a contingency table analysis since both response and predictor variables are categorical variables.

Frequency
Expected
Cell Chi-Square
Row Pct

Table of CreditStatus by Gender			
	Gender		
CreditStatus	F	M	Total
0	1486 1447.3 1.0362 67.64	711 749.73 2.0004 32.36	2197
1	52 52.041 327E-7 65.82	27 26.959 0.0001 34.18	79
2	1 3.9525 2.2055 16.67	5 2.0475 4.2575 83.33	6
3	2 1.9763 0.0003 66.67	1 1.0238 0.0006 33.33	3
4	1 1.9763 0.4823 33.33	2 1.0238 0.931 66.67	3
5	11 15.151 1.1374 47.83	12 7.8488 2.1956 52.17	23
C	2801 2802.3 0.0006 65.84	1453 1451.7 0.0012 34.16	4254
X	916 945.31 0.9085 63.83	519 489.69 1.7539 36.17	1435
Total	5270	2730	8000

Figure 21 : Cross tabulation table of gender and creditStatus

Based on the cross tabulation table, with a credit status of "X", female tend to have a higher probability of having that status than male which means than female tend to have lesser loans than male at 63.83% versus 36.17% respectively. When looking at credit status of category "5" which is tendency of having bad debts, male generally have a higher probability of having it at 52.17% against female at 47.83%. Pearson's Chi-Square Test is conducted next to investigate the association of this relationship.

Statistics for Table of CreditStatus by Gender

Statistic	DF	Value	Prob
Chi-Square	7	16.9110	0.0180
Likelihood Ratio Chi-Square	7	16.3445	0.0221
Mantel-Haenszel Chi-Square	1	4.1218	0.0423
Phi Coefficient		0.0460	
Contingency Coefficient		0.0459	
Cramer's V		0.0460	
WARNING: 38% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Figure 20 : Statistics table for association checking

A set of hypothesis is declared for the association testing.

H_0 : There is no association between creditStatus and Gender
H_1 : There is an association between creditStatus and Gender

The x^2 statistic has a value of 16.911 followed by a p-value of 0.018, which is below 0.05. We are able to conclude that, at the 5% significance level, H_0 is rejected and there is an association between creditStatus and Gender. Both the cross tabulation table and Chi-Square test proves that there is a relationship, but it would be better to measure the strength of the association so that we will know whether is gender a strong reference for loan approval or credit status of an individual.

The Cramer's V suggests that the further away its value is from 0, the stronger the association between the variables is. Refer back to Figure 20, the Cramer's V value is 0.0460. This value is very close to the reference value of 0 which says the association between credit status and gender is very weak. The statistics table also included the Mantel-Haenszel Chi-Square test value which checks for the significance of ordinal association between variables. A set of hypothesis then declared.

H_0 : There is no ordinal association between creditStatus and Gender
H_1 : There is an ordinal association between creditStatus and Gender

Statistic	Value	ASE
Gamma	0.0445	0.0202
Kendall's Tau-b	0.0233	0.0106
Stuart's Tau-c	0.0244	0.0111
Somers' D C|R	0.0200	0.0091
Somers' D R|C	0.0271	0.0123
Pearson Correlation	0.0227	0.0111
Spearman Correlation	0.0246	0.0112
Lambda Asymmetric C|R	0.0022	0.0021
Lambda Asymmetric R|C	0.0000	0.0000
Lambda Symmetric	0.0009	0.0009
Uncertainty Coefficient C|R	0.0016	0.0008
Uncertainty Coefficient R|C	0.0010	0.0005
Uncertainty Coefficient Symmetric	0.0012	0.0006

Figure 21 : tabulation to measure strength of ordinal association

Referring to Figure 20, the p-value for the Mantel-Haenszel Chi-Square test is 0.0423, which is also below 0.05. H_0 is rejected and we conclude that there is an ordinal association between credit status and gender. Referring to figure 21, the value of gamma statistics and spearman correlation were 0.0445 and 0.0245 respectively. This proves that the ordinal association is positive, and very weak.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	23.5	Somers' D	0.020
Percent Discordant	21.5	Gamma	0.044
Percent Tied	55.0	Tau-a	0.012
Pairs	19505313	c	0.510

Figure 22 : Concordance Statistics

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	17177.324	17174.492
SC	17226.234	17230.390
-2 Log L	17163.324	17158.492

Figure 23 : Model Fit Statistics

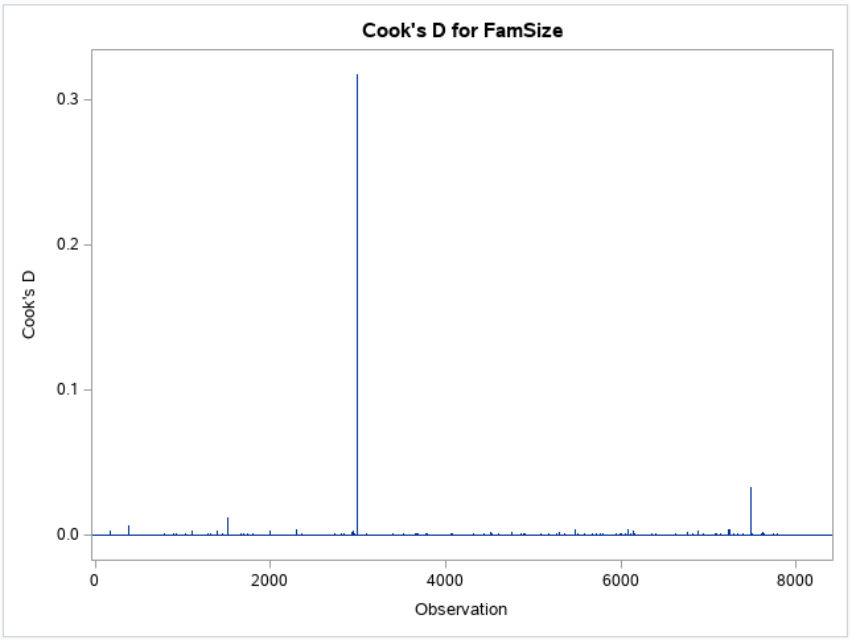
The c statistics from figure 22 shows a value of 0.510. This proves that it is a very weak ability to judge one's credit status based on its gender. The output result of -2 Log L from figure 23 is lesser or equals to 5 when comparing "intercept only" and "intercept and covariates". This further proves that model is not a good fit.

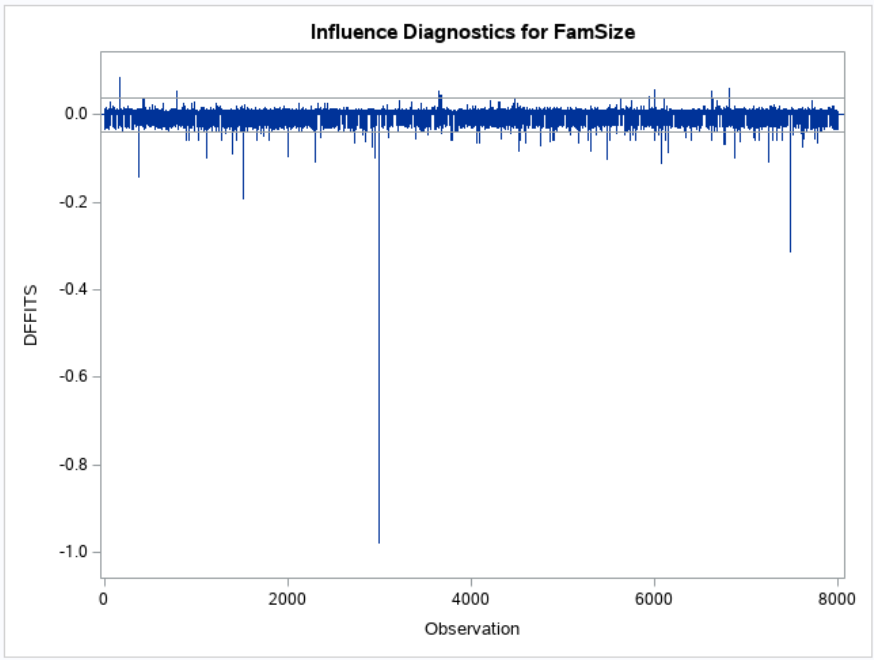
As conclusion for this objective, we first conducted a logistics regression with creditStatus as the response variable and the others as predictor variables. The result shows that only gender has an effect towards credit status of an individual. Then, a cross tabulation was conducted to find the relationship of gender and credit status. The result suggests that there is a pattern found and a statistics table of chi-squares was used as reference to prove the association and strength. The result of Chi-Square proved that there is a weak association and weak ordinal association between the variables. Overall, gender is found to have a relationship with credit status but it is not suggested to use as reference as it is very weak in which is proven in figure 22 and figure 23.

iii. Objective 3 - Tailoring product to each customer based on spending affordability

To figure out the spending affordability of an individual, we will need to find out their predicted expenses so that we understand to what extend they can afford a financial product or service. However, this dataset did not provide information about the spendings of each customer, and thus we can only provide predictions on their affordability. 2 assumptions were made based on this dataset in order to achieve the objective. Firstly, the individuals that owns a car or property is currently under loan for that respective asset. Secondly, that individual is responsible for the expenses of his or her family. With these assumptions made, we proceed to predict expenses of an individual.

A multiple linear regression analysis was conducted with family size as response variable with children count and annual income as predictor variable alongside. The reason behind this is to understand the size of a family on average as it is the main expenses for everyone and relationship with other variables present in this dataset. Before we proceed for the interpretation of the analysis, it is good to check for influential data and violation of regression assumption first.





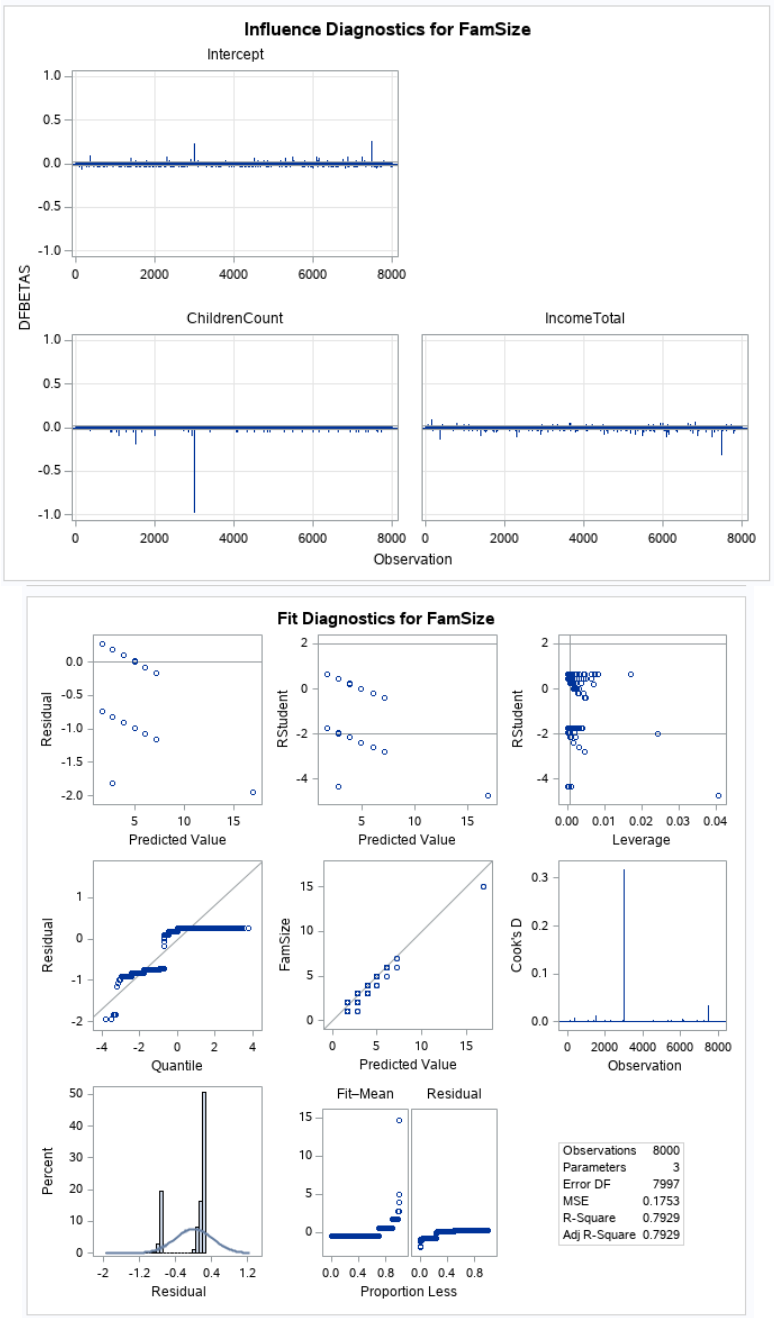


Figure 24 : Cook's D, DFFITS, DFBETAS, and fit diagnostics plot

4 different outputs were generated to check for influential data and assumptions. Cook's D and DFFITS plot shows that there is an influential point for family size. After checking from the dataset, the influential point seems belong to the customers that has a large family size of 15 persons. The result of DFBETAS plot show that all 3 of the variables in this model, FamSize, ChildrenCount, and IncomeTotal had influential points. Based on the plot of residual by FamSize in fit diagnostics, it shows an obvious trend of moving downwards. This violates the independence of residual errors which is an assumption for linear regression. After the completion of diagnostics for regression, we can now finally proceed for the analysis of multiple linear regression.

The REG Procedure
Model: MODEL1
Dependent Variable: FamSize

Number of Observations Read	8000
Number of Observations Used	8000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5369.05679	2684.52839	15311.9	<.0001
Error	7997	1402.06209	0.17532		
Corrected Total	7999	6771.11888			

Root MSE	0.41872	R-Square	0.7929
Dependent Mean	2.18838	Adj R-Sq	0.7929
Coeff Var	19.13367		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.72996	0.01016	170.35	<.0001
ChildrenCount	1	1.08667	0.00621	174.93	<.0001
IncomeTotal	1	2.048533E-9	4.682778E-8	0.04	0.9651

Figure 24 : Multiple linear regression output

The multiple regression equation was constructed based on the output.
$Y = 1.72996 + 1.08667X_1 + 2.048533X_2$

H_0 : There is no significant relationship between the predictor and response variables
H_1 : At least one of the predictor variables has a significant relationship with the response variable

Based on the output, we can see that this model has a **R-Square** value of 0.7929. This is explained as 79.29% of the variation in the family size is explained by the model. The **Root MSE** which is the standard deviation of error term has a value of 0.41872. We can observe that **coefficient of variation** has a value of 19.13, in which we can say that the ratio of the standard deviation to the mean indicates the dispersion from the distribution is 19.13%. The output shows an F value of 15311.9 followed by a **p-value** of <0.0001. At a significance level of 0.05, *H_0* is rejected and we conclude that at least one of the predictor variable in this model has a significant relationship with the response variable.

The parameter estimates table shows the result of the effect of each variables to family size respectively. If the variable has a p-value of lesser than the significance level of 0.05, we shall reject the H_0 and claim that it has a significant effect to family size. Looking at the p-value of childrenCount and IncomeTotal, only ChildrenCount has a p-value of lesser than 0.05. Thus we can conclude that the predictor variable ChildrenCount has a significant effect on FamSize.

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7929 and C(p) = 3.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5369.05679	2684.52839	15311.9	<.0001
Error	7997	1402.06209	0.17532		
Corrected Total	7999	6771.11888			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.72996	0.01016	5087.61355	29018.4	<.0001
ChildrenCount	1.08667	0.00621	5364.82303	30599.6	<.0001
IncomeTotal	2.048533E-9	4.682778E-8	0.00033552	0.00	0.9651

Bounds on condition number: 1.0008, 4.0031

Backward Elimination: Step 1

Variable IncomeTotal Removed: R-Square = 0.7929 and C(p) = 1.0019

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5369.05645	5369.05645	30627.5	<.0001
Error	7998	1402.06243	0.17530		
Corrected Total	7999	6771.11888			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.73034	0.00536	18248	104096	<.0001
ChildrenCount	1.08668	0.00621	5369.05645	30627.5	<.0001

Bounds on condition number: 1, 1

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	IncomeTotal	1	0.0000	0.7929	1.0019	0.00	0.9651

Figure 25 : Backward selection on multiple regression analysis

A model selection method is required to be performed to make the model a better fit and filter out variables with no influence. The selection method chosen was backward and the output was generated on SAS Studio. At step 0, we can see that IncomeTotal has a p-value of 0.9651 which is higher than the default significance level of 0.1. Thus, it is removed from the model as seen in step 1 and ChildrenCount remained to be the variable that has an influence to FamSize while making a better regression model.

Till this point, we can confirm that the number of children closely impacts the family size. Although it is a fact, but a pattern was discovered in this dataset which could help achieve our objective. Based on the descriptive analysis made for children count, it is seen that 29.8% of the individuals in this dataset has at least one child while the remaining 70.2% of them do not seem to have any child. This pattern is useful especially in the banking industry to promote financial products or services. We are able to tailor a suitable product to the individual based on their affordability which is seen through their family size and children count. One with a large family size proves that he/she has many children. We should custom make and promote a lower risk financial product to them because they tend to have higher expenses and

lower affordability. Furthermore, they could not take on greater risk because they had a larger family size and having more children means that they need to be cash-ready for any circumstances. For those that has a smaller family size do mean that they have lesser or no children. If that is the case, a financial product with greater risk but greater return could be a more suitable option for them. Without a large family size and large amount of children, they tend to have greater amount in cash which has no other purpose other than laying around in their savings account.

4. Conclusion

Overall, this dataset doesn't seem normally distributed for all the continuous variables. Thus they do not meet the assumption when carrying out ANCOVA or logistics regression. However, this is not a major issue since the main purpose and objective for this assignment is to implement the statistical techniques learned and come up with interesting analysis of the given dataset. This is why I chose an approach of dealing this dataset from a perspective of a data analyst of the bank so that I can think from new perspective and find out interesting patterns/relationships instead of blindly using statistical techniques.

The result of descriptive analysis on this dataset showed some interesting patterns. Firstly at their loan status, most of them had an excellent performance in managing their credit loan status with around 70% of them had no loan or paid off loans for that month. Then, it is seen that most of them do not own a car but owns a property. This may be due to the phenomenon of both advance and convenience of public transportations and decreasing price of properties. When looking at the children count, little over 70% of them did not have any children which due to high living cost and maintenance of having a child. It is especially interesting to see that more than half of them have an education level of until secondary or secondary special only. Not even over 1% of them had an academic degree. It would be more interesting if we can compare a dataset with the same variables but constructed 10 years later to see whether education level is a point of interest in people in the future. 100% of the individuals owns a mobile phone but 90% of them do not own an email account. Judging by these 2 factors, we can see that mobile phone is now a necessity to everyone but they are not that into technological applications and prefer a more traditional social practice. Lastly, looking at the family size, 53% of the individuals has a family size of 2 person. This could act as another evidence to support the justification made for low children count assuming that these 2 person do not count in a widow and a child.

For the analysis part of the assignment, objectives were identified from the perspective of benefits of using big data analysis in the banking industry. I laid down 3 main objectives to be achieved which will be explained seperately. Firstly, what variables will affect with the annual income of an individual and their relationship in it. An ANCOVA test was executed to find out the variables that has an effect to the total income. The result was gender, owncar, ownproperty, incometype, educationlevel, email, housingtype, and occupation. However, our model seems to violate all 3 of the assumptions for ANCOVA but we can ignore due to our large dataset and investigating of relationship is the top priority. Then, a multiple way ANOVA test was conducted to investigate the relationship. A relationship of an individual with higher income was discovered : male, owns a car, owns a property, has an income type of commercial associate, has an academic degree, living in an office apartment, and has an occupation of managers.

For the 2nd objective, I came up with an objective of investigating the relationship of credit status with other variables. For this, a logistics regression was then conducted to find the variables that has a significant effect to credit status. The result shows that gender was the only variable that has effect which is Females tend to have a better credit status than males. However, as investigation and analysis was conducted further on, it was proven that although gender has a relationship with credit status, it is strongly not recommended to use it as reference because the association is weak with credit status. Thus, as a banking firm, we can only mildly predict the creditability of a customer based on their gender but not believing it firmly.

For the 3rd objective, it is in the area of interest to find out any relationship in this dataset so that we can tailor product to each customer based on their spending affordability using statistical techniques. It is believed that family will be the main source of high expenses and thus making family size our response variable. A multiple linear regression analysis was conducted to find out what affects a family size in and children count was found to be the only one in this dataset. Both family size and children count is a big expenses especially children because they tend to have higher maintenance like education fees, resources, tools & toys etc. So, we are able to conclude that if an individual has a big family size, it has a large children count and we can tailor a financial product with smaller risk but smaller return. If the individual has a smaller family size, another kind of financial product with higher return but higher risk could be recommended because they have lower expenses and higher affordability.

As conclusion, the dataset given was large and consist of many interesting variables that have high analytic value. Although the data did not met the assumptions for the given analytical techniques, but other relationships, association and patterns are still discovered. It is definitely an eye opening experience of how a dataset could be consist of and the power of analytical tool. Without analytical knowledge, no one can even figure out the hidden relationships of data and improve business operation and value. Using a quote from an American consultant named Geoffrey Moore, “Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway”. Big data analytics is proved useful, insightful, and definitely is the future for all industries, especially banks. Without them, businesses are just wandering not on a freeway, but in deep ocean.

References

Gagnon, R. (2020). Your Go-to Guide to Big Data Analytics in Banking. Retrieved 2020, from https://global.hitachi-solutions.com/blog/big-data-banking

Appendix

```
PROC UNIVARIATE DATA=creditStatus;
  VAR ChildrenCount;
  histogram ChildrenCount / normal;
  inset mean mode median;
RUN;
```

Code screenshot to perform univariate analysis for descriptive analysis of ChildrenCount

```
PROC UNIVARIATE DATA=creditStatus;
  VAR IncomeTotal;
  histogram IncomeTotal / normal;
  inset mean mode median;
RUN;
```

Code screenshot to perform univariate analysis for descriptive analysis of IncomeTotal

```
PROC UNIVARIATE DATA=creditStatus;
  VAR FamSize;
  histogram FamSize / normal;
  inset mean mode median;
RUN;
```

Code screenshot to perform univariate analysis for descriptive analysis of FamSize

```
ods graphics on;
proc glm data=creditStatus2 plot(only maxpoints=10000)=(residuals diagnostics);
    class CreditStatus Gender OwnCar OwnProperty ChildrenCount IncomeType EducationLevel MaritalStatus HousingType Mobile Email Occupation FamSize;
    model IncomeTotal = CreditStatus Gender OwnCar OwnProperty ChildrenCount IncomeType EducationLevel MaritalStatus HousingType Email Occupation FamSize;
run;
ods graphics off;
```

Code screenshot to perform ANCOVA

```
proc glm data=creditStatus2;
    class Gender;
    model IncomeTotal = Gender;
    lsmeans Gender;
run;
```

```
proc glm data=creditStatus2;
    class OwnCar;
    model IncomeTotal = OwnCar;
    lsmeans OwnCar;
run;
```

```
proc glm data=creditStatus2;
    class OwnProperty;
    model IncomeTotal = OwnProperty;
    lsmeans OwnProperty;
run;
```

```
proc glm data=creditStatus2;
    class IncomeType;
    model IncomeTotal = IncomeType;
    lsmeans IncomeType;
run;
```

```
proc glm data=creditStatus2;
    class EducationLevel;
    model IncomeTotal = EducationLevel;
    lsmeans EducationLevel;
run;
proc glm data=creditStatus2;
    class HousingType;
    model IncomeTotal = HousingType;
    lsmeans HousingType;
run;
```

Code screenshots to perform ANOVA

```
ods graphics on;
proc logistic data=creditStatus2 plots=oddsratio;
class CreditStatus Gender OwnCar OwnProperty ChildrenCount IncomeType EducationLevel MaritalStatus HousingType Mobile Email Occupation FamSize;
model CreditStatus = Gender OwnCar OwnProperty ChildrenCount IncomeType IncomeTotal EducationLevel MaritalStatus HousingType Mobile Email Occupation FamSize;
units IncomeTotal=100;
effectplot;
run;
```

Code screenshot to perform logistic regression for all variables

```
ods graphics / imagemap=on;
proc reg data=creditStatus2 plot(only maxpoints=10000)=(diagnostics dffits dfbetas cooksd);
    model FamSize = ChildrenCount IncomeTotal;
run;
ods graphics off;
```

Code screenshot to perform multiple linear regression and diagnostics

```
proc reg data=creditStatus2;
    model FamSize = ChildrenCount IncomeTotal / selection=backward;
run;
```

Code screenshot to perform backward selection on linear regression