UNIVERSITI
TEKNOLOGI
MARA

BACHELOR OF SCIENCE (HONS) STATISTICS

FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES

STA610

SAS PROGRAMMING

SALES RECORDED IN SEVERAL REGIONS

PREPARED BY

MIFTAKUL HUDA BIN MUCHAMAD IMRON	2019329007
MUHAMMAD FITRI BIN SALIM	2019314589
NORFAIZATULAH BINTI ABDULLAH	2019314473

PREPARED FOR

MADAM NOOREZATTY BINTI MOHD YUSOP

DATE OF SUBMISSION

29th JUNE 2020

TABLE OF CONTENTS

CHAPTER 1 : INTRODUCTION

1.1 Background of Study

A sale is defined as a transaction that occurs between two or more parties in which the buyer receives tangible or intangible goods, services or assets in exchange for money. In other word, a sale is essentially a contract between the buyer and the seller of a particular good or service in question. Sales can also be completed between businesses such as when one raw materials provider sells available materials to a business that uses the materials to produce consumer goods. This led to a further action into recording a sale that happened as a reference for business purposes in the future. Sales records can be defined as the information that are collected on customers, including but not limited to their contact information, how often they purchase and what they purchase. Recording a sale is important since it is hard to remember key personal and professional details about the customers that have been purchasing goods and services from a company. The information and details gained from each occurring sale may help a company to make better decision in generating more profit in the future.

The secondary dataset in this study is about transactions that have been successfully completed in selling distinct item types to other parties by different countries from different regions. The original dataset consists of fourteen variables that were recorded in a Microsoft Excel sheet, but the researchers only choose five variables that are related to the objectives that are to be achieved which are region, item type, number of items sold, total profit and total revenue. The researchers also decided to use 500 observations out of 1000 observations as the sample for this study. The main criteria of the records are the type of items sold by different countries and regions, the number of items sold, and the total profit and revenue obtained from the transactions. These variables are crucial in helping one country or region in the dataset to take appropriate measures in increasing their revenue and profit while considering on the type of item they are selling and the number of items sold. Therefore, by using these selected variables, a further study on the dataset should be done to determine the impact of selling different type of products and identify the profit standings for each region.

1.2 Research Objectives

1. To compare the differences between Asia region and Europe region towards the total revenue made.
2. To identify whether there is a significant difference in total number of units sold among item types.
3. To determine the association between regions and three total profit levels.

1.3 Data Description

The data used for this study is a secondary dataset that was obtained from http://eforexcel.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/. The dataset is about sales recorded in several regions. The original dataset consists of 14 variables but only 3 quantitative variables (Units Sold, Total Revenue, Total Profit) and 2 qualitative variables (Region, Item Type) were used in this study prior to the objectives of interest as given below.

Table 1.1 : Variable Descriptions

VARIABLES	DESCRIPTION	TYPE
Region	Regions in which sales had occurred	Nominal
Item_Type	The type of items sold	Nominal
Units_Sold	The total number of units sold	Numeric
Total_Revenue	The total income received from the sales	Numeric
Total_Profit	The total profit after deducting the total cost from the total revenue obtained	Numeric

CHAPTER 2 : METHODOLOGY

2.1 Introduction

In this chapter, we will be discussing about all the method that will be applied in order to achieve to the success of this project. Methods discussed in methodology help to achieve the objective of this project.

2.2 Data Management

2.2.1 SAS Data Library

Libname statements is used to assign a libref to SAS library.

General form of the LIBNAME statement:

```
LIBNAME libret ‘SAS-library’;
```

**refer to appendix 1*

2.2.2 Import Data

The data is import from excel document into SAS by using proc import.

General form of the IMPORT statement:

```
PROC IMPORT        OUT= <libref> ‘SAS-data-set’
                   DATAFILE= ‘external-file-name’
                   DBMS=’identifier’;
RUN;
```

**refer to appendix 2*

2.2.3 Creating New Data Set

Data statement will be used to create SAS data set name and the data created can be temporary or permanent data sets.

The general form of the DATA statement:

```
DATA ‘output-SAS-data-set’;

RUN;
```

2.2.4 Reading Data from Existing Data Set

SET statement reads all observations and all variable from an existing SAS data set.

General form of the SET statement:

```
DATA ‘output-SAS-data-set’;

        SET ’input-SAS-data-set’;

RUN;
```

2.2.5 Selecting Variables

KEEP statement can be used to control the variable that want to attain in the new SAS dataset

General form of the KEEP statement:

```
DATA ‘output-SAS-data-set’;

        SET ’input-SAS-data-set’;

        KEEP variables;

RUN;
```

2.2.6 Sub-setting Data

WHERE statement is a sequence of operators and operands. This statement used to select observations that meet condition needed.

General form of WHERE statement:

```
DATA ‘output-SAS-data-set’;
          SET ’input-SAS-data-set’;
          KEEP variables;
          WHERE where-expression
RUN;
```

2.2.7 Formatting Data Values

There are two methods can be use in assigning format to the data values which are formatted values and user-defined formats. For formatted values, FORMAT is used as instruction that SAS uses to write data values while user-defined formats used PROC FORMAT statement to change the attribute name, data format and so on. In this study, user-defined formats will be used.

The general form of PROC FORMAT;

```
PROC FORMAT;
                    VALUE format name range1= ‘label’
                                                  range2= ‘label’;
RUN;
```

2.2.8 The Print Procedure

PRINT PROCEDURE will be applied to display titles, footnotes, description column headings, formatted data values, column totals, column subtotal and page break for each subgroup.

The general form of PROC PRINT:

```
PROC PRINT DATA = SAS-data-set;
RUN;
```

2.2.9 Print Selected Variable

VAR statement will be used to select the variables that will be included in the report and it will arrange follow by its sequence.

The general form of VAR statement:

```
PROC PRINT DATA = SAS-data-set;
                VAR variable(s);
RUN;
```

2.2.10 Defining Title

TITLE statement used to as title of the report when printing.

The general form of TITLE statement:

```
TITLE “title”;
```

2.3 Descriptive Statistics

2.3.1 Summary Statistics

The MEAN PROCEDURE used to display simple descriptive statistic like N, MEAN, STD, MIN and MAX of the variable.

The general form of simple PROC MEANS;

```
PROC MEANS DATA = SAS-data-set;
RUN;
```

2.3.2 Summary Reports

The PROC FREQ statement used to provide output like frequency, percent and cumulative percent.

The General form of PROC FREQ;

```
PROC FREQ DATA = SAS-data-set;
        TABLES variable1*variable2;
RUN;
```

2.3.3 Tabulate Procedure

One and two-dimensional tabular reports can be created by using TABULATE procedure. The report will include control of table construction, specifying statistics and formatting values.

The general form of PROC TABULATE:

```
PROC TABULATE
DATA = SAS-data-set<option>;
        CLASS classvariables;
        VAR analysis-variables;
        TABLES pageexpression;
RUN;
```

2.3.4 Bar and Pie Chart

Chart such as bar chat and pie chart can be display in SAS Programming by using GCHART PROCEDURE. The type of chart needed can be specify by HBAR3D, VBAR or PIE STATEMENT.

General form of the PROC GCHART:

```
PROC GCHART DATA= SAS-DATA-SET;
```

Statements to specify the desired type of chart

```
HBAR3D chart-variable…</options>;
VBAR chart-variable…</options>;
PIE3D chart-variable…</options>;
```

2.4 Inferential Statistics

2.4.1 Test of Normality

Normality test is needed to determine either the data is normally distributed or not normally distributed. By obtaining the result, the variable can be classified into parametric or non-parametric test.

The general form of PROC UNIVARIATE:

```
PROC UNIVARIATE DATA = SAS-data-set;
        VAR variable(s);
        PROBPLOT/NORMAL (MU=EST SIGMA=EST);
RUN;
```

2.4.2 Non-parametric Tests

2.4.2.1 Wilcoxon Rank Sum Test (Mann-Whitney U Test)

The Mann-Whitney Test is a useful nonparametric alternative to the independent T Test. It is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous but not normally distributed.

The general form of PROC NPAR1WAY:

```
PROC NPAR1WAY DATA = SAS-data-set WILCOXON;
        CLASS variable(s);
        VAR variable(s);
        EXACT Wilcoxon;
RUN;
```

2.4.2.2 Kruskal-Wallis Test

Kruskal-Wallis Test used to determine if there are any significant differences between two or more groups of independent variables on a continuous or ordinal dependent variable. It is ANOVA for non-parametric data.

The general form of PROC NPAR1WAY:

```
PROC NPAR1WAY DATA = SAS-data-set WILCOXON;
        CLASS variable(s);
        VAR variable(s);
RUN;
```

2.4.2.3 Chi-square Test

Chi-Square test is used to examine the differences between categorical variables in the data set.

General form of the PROC FREQ:

```
PROC FREQ DATA = SAS-data-set;
        TABLES variable*variable/EXPECTED CHISQ FISHER;
RUN;
```

CHAPTER 3 : RESULTS AND ANALYSIS

3.1 Data Conversion

The original dataset is created in an Excel file that were retrieved from an online website, therefore the dataset is converted into a SAS dataset to be further analyze using the proc import step as given below.

```
proc import   out = project.all
              datafile = "H:\Users\Asus\Documents\Degree UiTM Seremban\PART 5\STA
              610 (SAS)\Project\SalesRecords.xlsx"
              dbms = excel;
run;
```

3.2 Data Management

The dataset which consist of 1000 sales recorded as the population is partitioned into new dataset which only consist of 500 sales recorded since the researchers are interested in selecting 500 observations as the sample of this study. The first 500 observations are selected using the OBS statement. Then, the new sample dataset is partitioned into several datasets named project.AE, project.items and project.regions that are related to each objective of this study that the researchers are interested in achieving. The total profit is classified into three levels which are low, moderate and high for each region to be used on the third objective of this study. All the datasets are coded as below :

1.
   ```
   data project.sales;
   set project.all (obs=500);
   keep Region Item_Type Units_Sold Total_Revenue Total_Profit;
   run;
   ```

2.

```
data project.AE;
set project.sales;
keep Region Total_Revenue;
where Region in ('Asia','Europe');
run;
```

3.

```
data project.items;
set project.sales;
keep item_type units_sold;
run;
```

4.

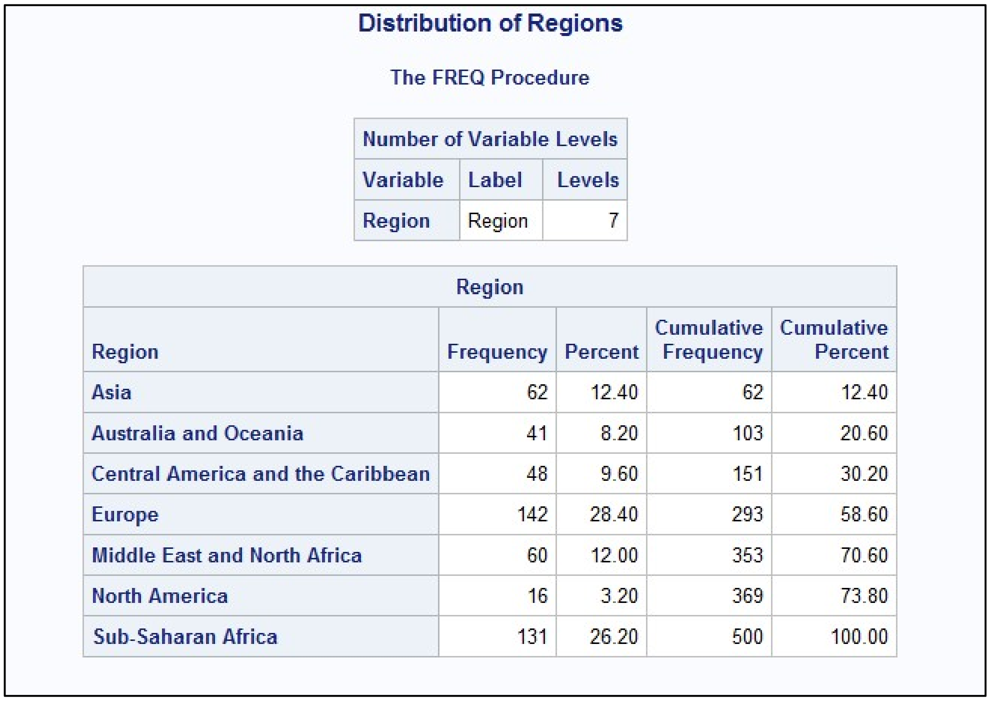
```
proc format;
value profitfmt
  low -<500000 = 'Low'
  500000-<1000000 = 'Moderate'
  1000000-high = 'High';
run;

data project.regions;
set project.sales;
keep region total_profit;
format total_profit profitfmt.;
run;
```

3.3 Descriptive Analysis

3.3.1 Frequency Report

```
title 'Distribution of Regions';
proc freq data=project.sales nlevels;
tables Region;
run;
```



Distribution of Regions

The FREQ Procedure

Number of Variable Levels		
Variable	Label	Levels
Region	Region	7

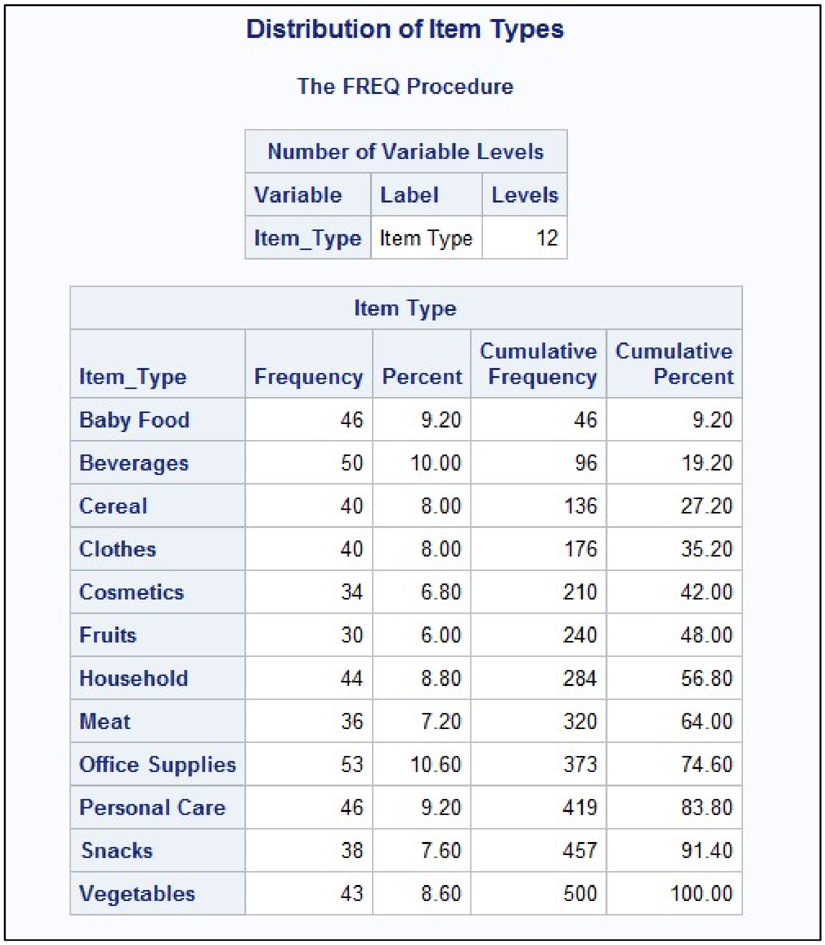
Region				
Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asia	62	12.40	62	12.40
Australia and Oceania	41	8.20	103	20.60
Central America and the Caribbean	48	9.60	151	30.20
Europe	142	28.40	293	58.60
Middle East and North Africa	60	12.00	353	70.60
North America	16	3.20	369	73.80
Sub-Saharan Africa	131	26.20	500	100.00

Figure 3.1 : The frequency report on distribution of regions

Interpretation :

The above frequency report shows that there exist seven different regions that were recorded in the sales dataset. It can be seen that Europe region has the highest number of sales occurred followed by Sub-Saharan Africa region. In contrast, North America region has lowest number of sales recorded.

```
title 'Distribution of Item Types';
proc freq data=project.sales nlevels;
tables Item_Type;
run;
```



Distribution of Item Types

The FREQ Procedure

Number of Variable Levels		
Variable	Label	Levels
Item_Type	Item Type	12

Item Type				
Item_Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Baby Food	46	9.20	46	9.20
Beverages	50	10.00	96	19.20
Cereal	40	8.00	136	27.20
Clothes	40	8.00	176	35.20
Cosmetics	34	6.80	210	42.00
Fruits	30	6.00	240	48.00
Household	44	8.80	284	56.80
Meat	36	7.20	320	64.00
Office Supplies	53	10.60	373	74.60
Personal Care	46	9.20	419	83.80
Snacks	38	7.60	457	91.40
Vegetables	43	8.60	500	100.00

Figure 3.2 : Frequency report on distribution of item types

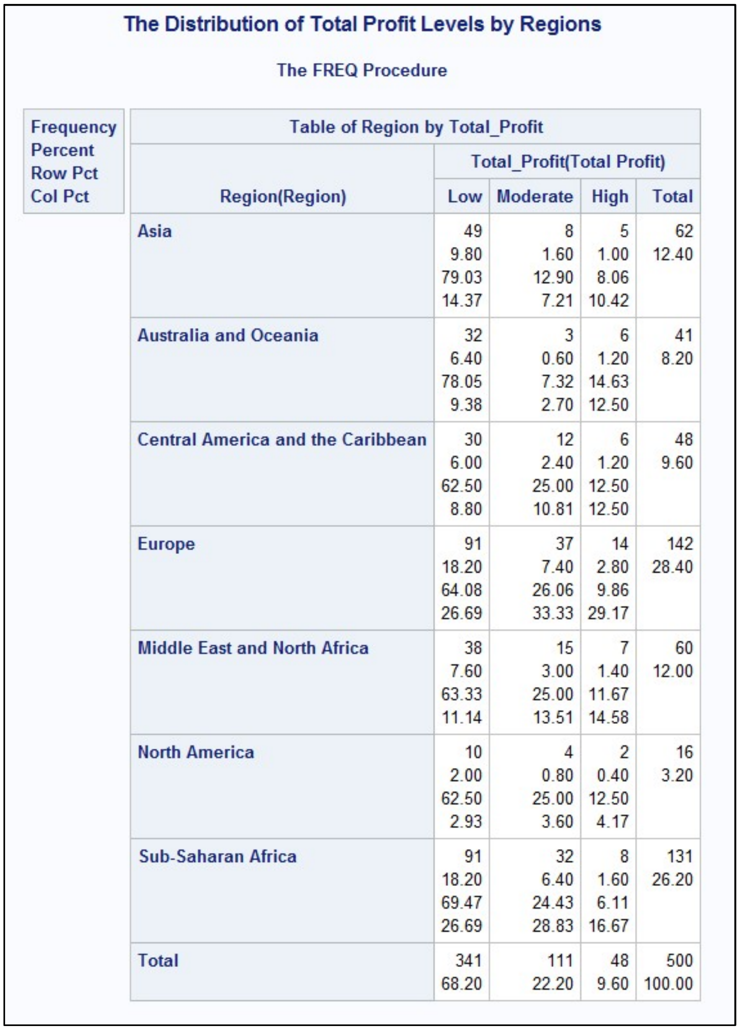
Interpretation :

Figure 3.2 indicates that there are 12 different item types that were managed to be sold by all regions. According to the figure above, office supplies is the most favourable type of item to be sold. In opposition, fruits have the lowest number of sales compared to other item types.

```
proc format;

value profitfmt
        low -<500000 = 'Low'
        500000-<1000000 = 'Moderate'
        1000000-high = 'High';
run;

title 'The Distribution of Total Profit Levels by Regions';
proc freq data=project.sales;
tables Region*Total_Profit;
format Total_Profit profitfmt.;
run;
```



The Distribution of Total Profit Levels by Regions

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Region by Total_Profit				
	Total_Profit(Total Profit)			
Region(Region)	Low	Moderate	High	Total
Asia	49 9.80 79.03 14.37	8 1.60 12.90 7.21	5 1.00 8.06 10.42	62 12.40
Australia and Oceania	32 6.40 78.05 9.38	3 0.60 7.32 2.70	6 1.20 14.63 12.50	41 8.20
Central America and the Caribbean	30 6.00 62.50 8.80	12 2.40 25.00 10.81	6 1.20 12.50 12.50	48 9.60
Europe	91 18.20 64.08 26.69	37 7.40 26.06 33.33	14 2.80 9.86 29.17	142 28.40
Middle East and North Africa	38 7.60 63.33 11.14	15 3.00 25.00 13.51	7 1.40 11.67 14.58	60 12.00
North America	10 2.00 62.50 2.93	4 0.80 25.00 3.60	2 0.40 12.50 4.17	16 3.20
Sub-Saharan Africa	91 18.20 69.47 26.69	32 6.40 24.43 28.83	8 1.60 6.11 16.67	131 26.20
Total	341 68.20	111 22.20	48 9.60	500 100.00

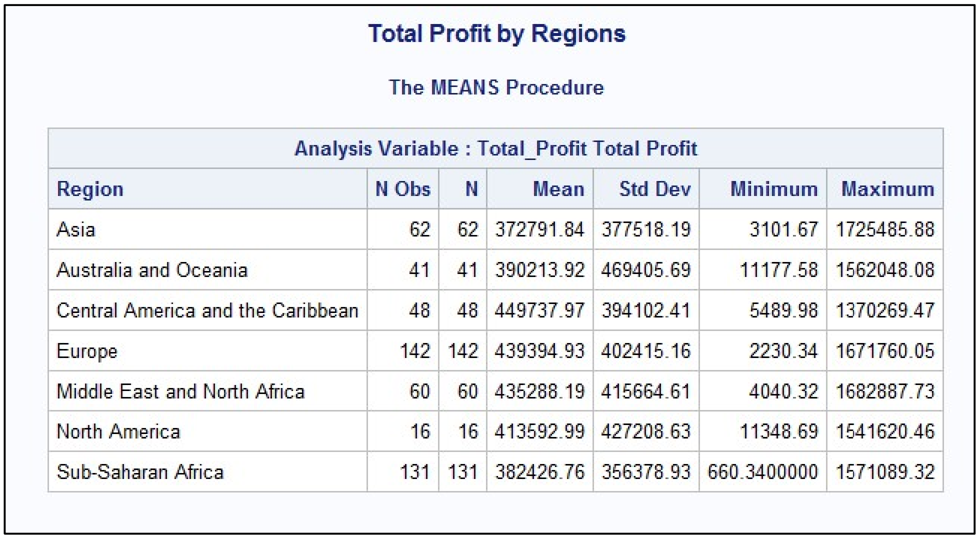
Figure 3.3 : Frequency report on the distribution of total profit levels by regions

Interpretation :

The above figure shows the distribution of regions towards 3 levels of total profit which are low, moderate and high. When a region successfully makes a profit between $0 until $499,000, it will be categorised in the low level. The profit that lies between $500,000 until $999,999 will be categorised in the moderate level and any total profit above that will be categorised in the high level. The result indicates that all the regions have the majority number of sales categorised in the low level.

3.4 Summary Statistic

```
title 'Total Profit by Regions';
proc means data=project.sales;
var Total_Profit;
class Region;
run;
```



Total Profit by Regions

The MEANS Procedure

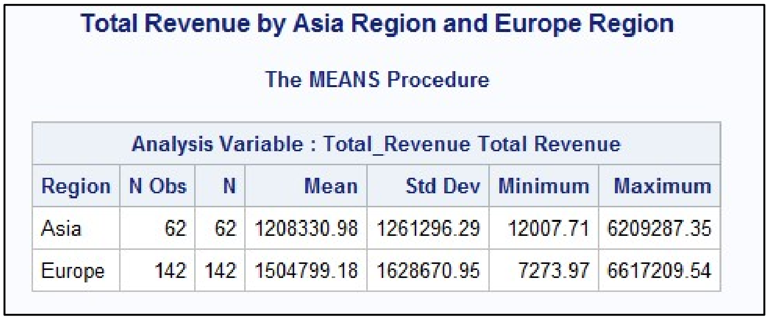
Analysis Variable : Total_Profit Total Profit						
Region	N Obs	N	Mean	Std Dev	Minimum	Maximum
Asia	62	62	372791.84	377518.19	3101.67	1725485.88
Australia and Oceania	41	41	390213.92	469405.69	11177.58	1562048.08
Central America and the Caribbean	48	48	449737.97	394102.41	5489.98	1370269.47
Europe	142	142	439394.93	402415.16	2230.34	1671760.05
Middle East and North Africa	60	60	435288.19	415664.61	4040.32	1682887.73
North America	16	16	413592.99	427208.63	11348.69	1541620.46
Sub-Saharan Africa	131	131	382426.76	356378.93	660.3400000	1571089.32

Figure 3.4 : Summary statistic on total profit by regions

Interpretation :

Figure 3.4 shows that the Central America and the Caribbean region has the highest average total profit of $449,737.97 followed by Europe region of $439,394.93. In contrast, Asia region has the lowest average total profit of $372,791.84.

```
Title ‘Total revenue by Asia Region and Europe Region’;
Proc means data=project.sales maxdec = 2;
Var Total_Revenue;
Class Region;
Where Region in (‘Asia’,’Europe’);
Run;
```



Total Revenue by Asia Region and Europe Region

The MEANS Procedure

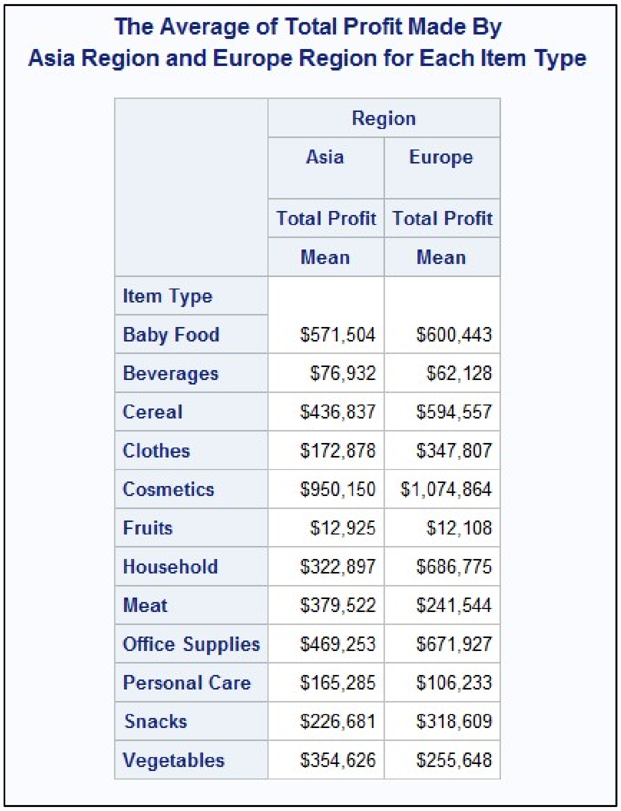
Analysis Variable : Total_Revenue Total Revenue						
Region	N Obs	N	Mean	Std Dev	Minimum	Maximum
Asia	62	62	1208330.98	1261296.29	12007.71	6209287.35
Europe	142	142	1504799.18	1628670.95	7273.97	6617209.54

Figure 3.5 : Summary statistic on total revenue by Asia region and Europe region

Interpretation :

Figure 3.5 shows the comparison on summary statistic of total revenue obtained by Asia region and Europe region. It is revealed that the average revenue for Asia region is $1,208,330.98. Meanwhile, Europe region has a higher average of revenue which is $1504,799.18. The minimum and maximum revenue for Asia region is $12,007.71 and $6,209,287.35 respectively. Meanwhile, the minimum and maximum revenue for Europe region is $7273.97 and $6,617,209.54 respectively.

```
title1 'The Average of Total profit made by Asia Region and Europe Region';
title2 'For Each Item Type';
proc tabulate data=project.sales;
where region in ('Asia','Europe');
class region item_type;
var total_profit;
table item_type, region*total_profit;
run;
```



The Average of Total Profit Made By
Asia Region and Europe Region for Each Item Type

	Region	
	Asia	Europe
	Total Profit	Total Profit
	Mean	Mean
Item Type		
Baby Food	$571,504	$600,443
Beverages	$76,932	$62,128
Cereal	$436,837	$594,557
Clothes	$172,878	$347,807
Cosmetics	$950,150	$1,074,864
Fruits	$12,925	$12,108
Household	$322,897	$686,775
Meat	$379,522	$241,544
Office Supplies	$469,253	$671,927
Personal Care	$165,285	$106,233
Snacks	$226,681	$318,609
Vegetables	$354,626	$255,648

Figure 3.6 : The average of total profit made by Asia and Europe for each item type

Interpretation :

The above figure shows the average of total profit made by Asia and Europe for each item type. It can be indicated that both regions have the highest average of total profit on cosmetics item compared to other item types which are $950,150 and $1,074,864 respectively. The lowest average of total profit made by both regions is still on the same item type which is fruits valued $12,925 and $12,108 respectively.

3.4.1 Pie Chart

```
title h=2 f=broadway 'The Percentage Distribution of Sales between Asia and Europe';
proc gchart data=project.sales;
pie3d region / type=percent noheading;
where region in ('Asia','Europe');
run;
```

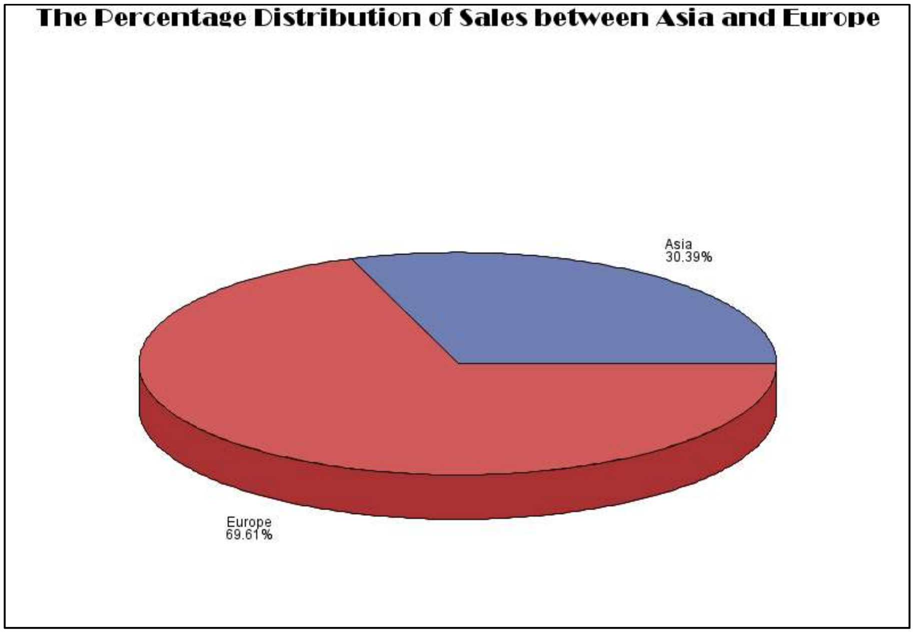


Figure 3.7 : Pie chart on the percentage distribution of sales between Asia and Europe

Interpretation :

The pie chart above describes the percentage distribution of sales made between Asia region and Europe region. By comparing these two regions, it can be indicated that Europe has a higher percentage of 69.61% on sales recorded compared to Asia which is only 30.39%.

3.4.2 Horizontal Bar Chart

```
title h=2 f=broadway 'The Total Number of Units Sold by Each Item Type';
proc gchart data=project.sales;
hbar3d item_type / sumvar=units_sold nostats;
run;
```

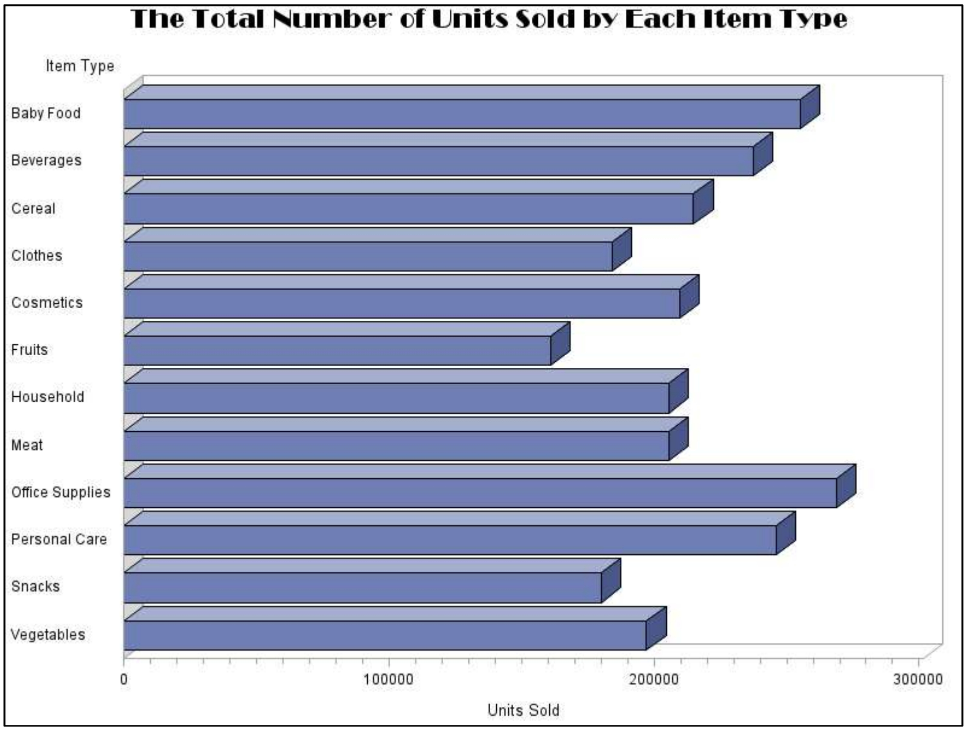


Figure 3.8 : Horizontal bar chart on the total number of units sold by each item type

<u>Interpretation :</u>

The above figure illustrates on the total number of units sold by each item type. The horizontal bar chart above indicates that the highest number of units sold by all regions is office supplies followed by baby food and personal care items. Meanwhile, the lowest item to be sold is fruits.

```
title h=2 f=broadway 'The Total Profit Made by Each Region';
proc gchart data=project.sales;
hbar3d region / sumvar=total_profit nostats;
format total_profit dollar12.;
run;
```

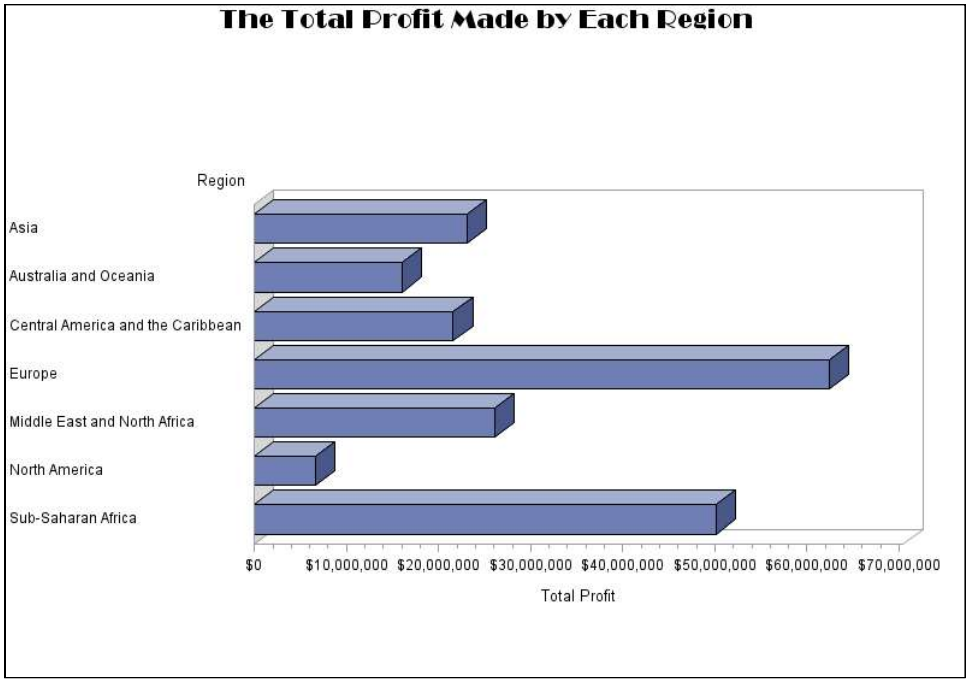


Figure 3.9 : Horizontal bar chart on the total profit made by each region

<u>Intepretation :</u>

The above diagram illustrates on the total profit made by each region. The horizontal bar chart above shows that Europe has the highest total profit made compared to other regions followed by Sub-Saharan African. Meanwhile, North America has the lowest total profit made.

3.4.3 Vertical Bar Chart

```
title h=2 f=broadway c=brown 'The Total Number of Sales Made by Each Region';
axis1 stagger label=none;
axis2 label=(a=90 'Frequency');
proc gchart data=project.sales;
vbar region / patternid=midpoint width=10 maxis=axis1 raxis=axis2;
run;
```

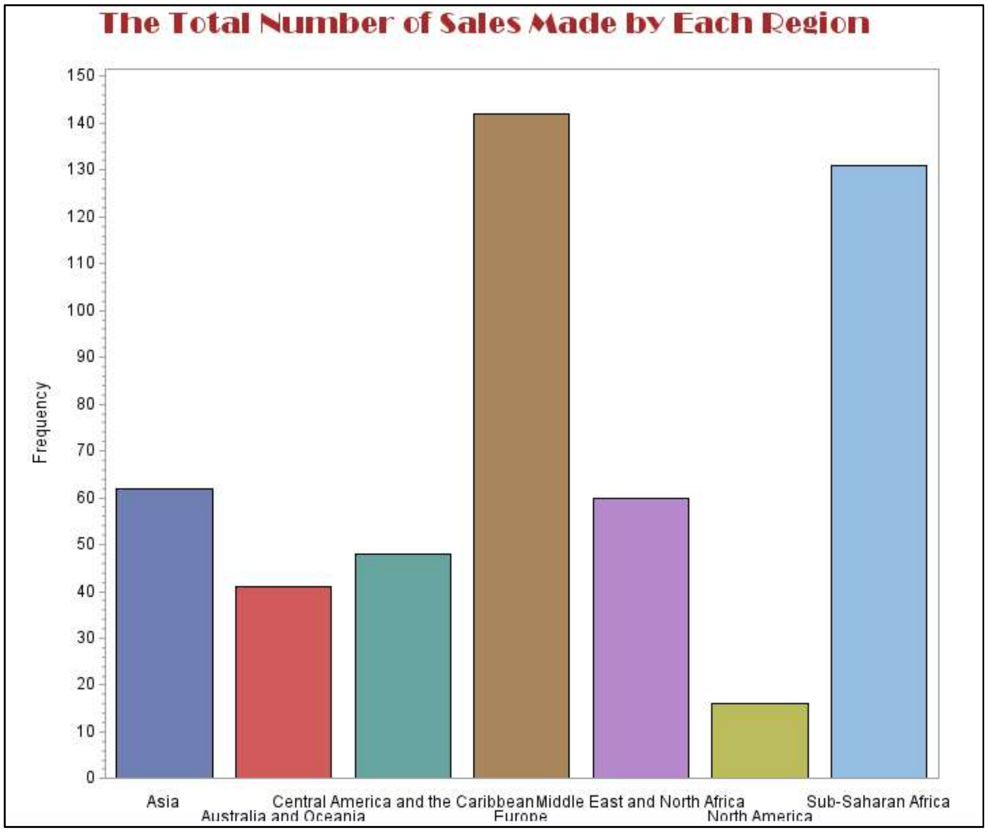


Figure 3.10 : Vertical bar chart on the total number of sales made by each region

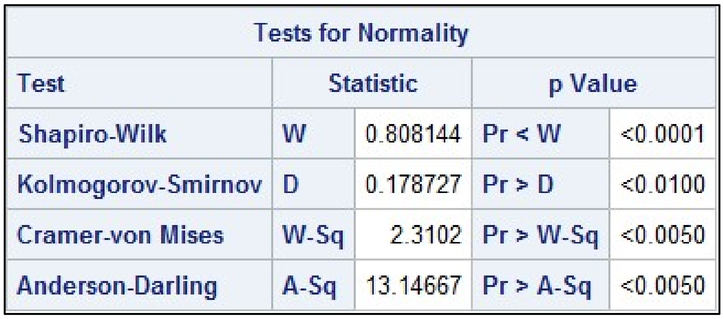
<u>Interpretation :</u>

The above diagram illustrates on the total number of sales recorded for each particular region. From the above vertical bar chart, it can be concluded that Europe is the region with the highest number of sales frequencies followed by Sub-Saharan Africa region. In contrary, North America has the lowest number of sales occurred.

3.5 Inferential Statistic

3.5.1 Normality Test on Total Revenue for Asia Region and Europe Region

```
proc univariate data=project.AE normal;
var total_revenue;
probplot / normal (mu=est sigma=est);
run;
```



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.808144	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.178727	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.3102	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	13.14667	Pr > A-Sq	<0.0050

Figure 3.11 : Test for Normality on Total Revenue

Hypothesis Testing :

H_0 : The data is normally distributed.

H_1 : The data is not normally distributed.

α : 0.05

p-value : 0.01

Decision Rule : Reject H_0 since p-value = 0.01 < α = 0.05

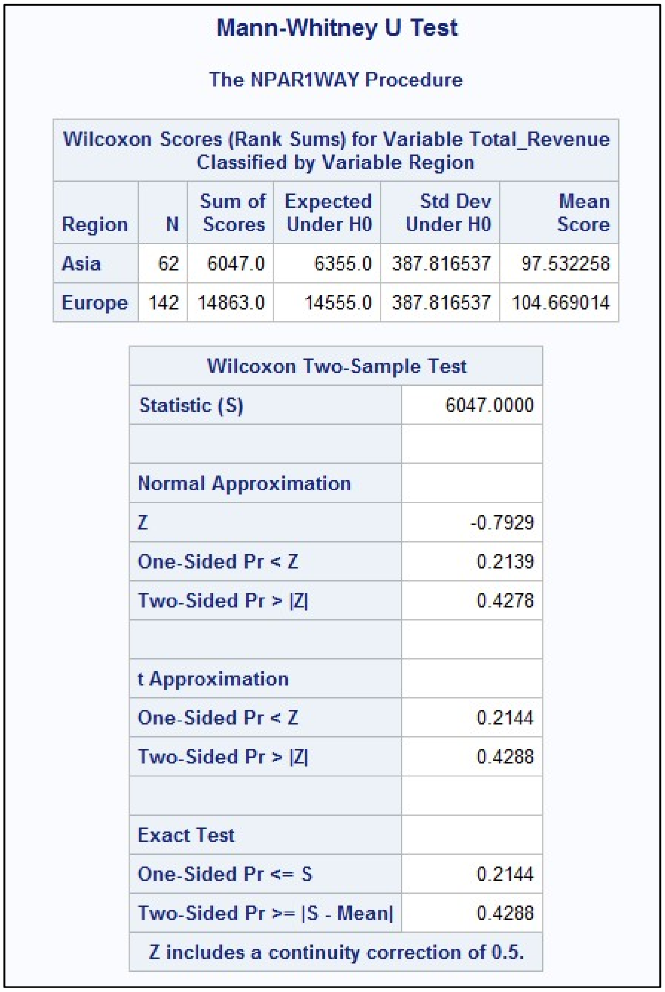
Conclusion : The data is not normally distributed.

Interpretation :

Figure 3.12 shows that there is not enough evidence to conclude that the total revenue for Asia and Europe follows a normal distribution since the p-value of the Kolmogorov-Smirnov of 0.01 is less than the alpha value of 0.05. Therefore, the null hypothesis is rejected and hence, the data is not normally distributed. The above findings can also be proven by Figure C.1 in Appendix C which indicates that the probability plot of total revenue does not follow a normal distribution since the points does not lie approximately in a straight line.

3.5.2 Mann-Whitney U Test to Compare the Differences Between Asia Region and Europe Region on Total Revenue

```
title 'Mann_Whitney U Test';
proc npar1way data=project.AE wilcoxon;
class region;
var total_revenue;
exact Wilcoxon;
run;
```



Mann-Whitney U Test

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Total_Revenue Classified by Variable Region					
Region	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Asia	62	6047.0	6355.0	387.816537	97.532258
Europe	142	14863.0	14555.0	387.816537	104.669014

Wilcoxon Two-Sample Test	
Statistic (S)	6047.0000
Normal Approximation	
Z	-0.7929
One-Sided Pr < Z	0.2139
Two-Sided Pr > |Z|	0.4278
t Approximation	
One-Sided Pr < Z	0.2144
Two-Sided Pr > |Z|	0.4288
Exact Test	
One-Sided Pr <= S	0.2144
Two-Sided Pr >= |S - Mean|	0.4288
Z includes a continuity correction of 0.5.	

Figure 3.12 : Mann-Whitney U Test to compare mean differences in total revenue between Asia and Europe

Hypothesis Testing :

H_0 : $\mu_{Asia} = \mu_{Europe}$

H_1 : $\mu_{Asia} \neq \mu_{Europe}$

α : 0.05

p-value : 0.4288

Decision rule : Failed to reject H_0 since p-value = 0.4288 > α = 0.05

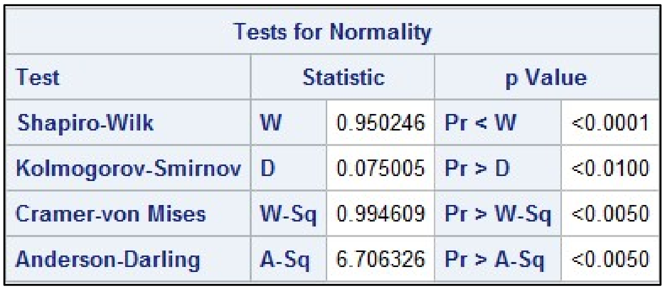
Conclusion : There is no significant mean difference on the total revenue between Asia region and Europe region.

Interpretation :

Figure 3.13 shows the output of a non-parametric Mann-Whitney U test on total revenue among Asia and Europe. Based on the figure, it can be concluded that there is not enough evidence to indicate that there is a significant difference in the mean of total revenue between Asia region and Europe region. This is because the p-value of 0.4288 is greater than 5% significant level and therefore, the null hypothesis failed to be rejected.

3.5.3 Normality Test for Number of Items Sold

```
proc univariate data=project.items normal;
var units_sold;
probplot / normal (mu=est sigma=est);
run;
```



Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.950246	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.075005	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.994609	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	6.706326	Pr > A-Sq	<0.0050

Figure 3.13: Normality Test for number of units sold

Hypothesis Testing :

H_0 : The data is normally distributed.

H_1 : The data is not normally distributed.

α : 0.05

p-value : 0.01

Decision Rule : Reject H_0 since p-value = 0.01 < α = 0.05

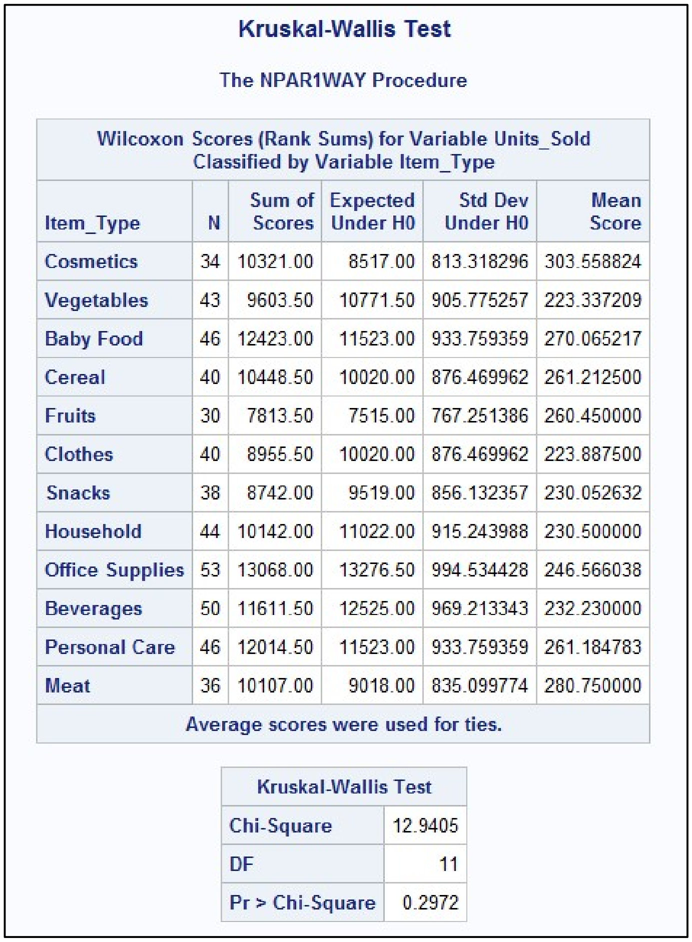
Conclusion : There is not enough evidence to conclude that the data follows a normal distribution.

Interpretation :

Figure 3.14 shows that there is not enough evidence to conclude that the total number of units sold for each item type follows a normal distribution since the p-value of the Kolmogorov-Smirnov of 0.01 is less than the alpha value of 0.05. Therefore, the null hypothesis is rejected and hence, the data is not normally distributed. The above findings can also be proven by Figure D.1 in Appendix D which indicates that the probability plot of number of units sold does not follow a normal distribution since the points does not lie approximately in a straight line.

3.5.4 Kruskal-Wallis Test to Determine Whether There is a Significant Mean Difference in the Number of Items Sold Among Item Types

```
title 'Kruskal-Wallis Test';
proc npar1way data=project.items wilcoxon;
class item_type;
var units_sold;
run;
```



Kruskal-Wallis Test

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Units_Sold Classified by Variable Item_Type

Item_Type	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Cosmetics	34	10321.00	8517.00	813.318296	303.558824
Vegetables	43	9603.50	10771.50	905.775257	223.337209
Baby Food	46	12423.00	11523.00	933.759359	270.065217
Cereal	40	10448.50	10020.00	876.469962	261.212500
Fruits	30	7813.50	7515.00	767.251386	260.450000
Clothes	40	8955.50	10020.00	876.469962	223.887500
Snacks	38	8742.00	9519.00	856.132357	230.052632
Household	44	10142.00	11022.00	915.243988	230.500000
Office Supplies	53	13068.00	13276.50	994.534428	246.566038
Beverages	50	11611.50	12525.00	969.213343	232.230000
Personal Care	46	12014.50	11523.00	933.759359	261.184783
Meat	36	10107.00	9018.00	835.099774	280.750000

Average scores were used for ties.

Kruskal-Wallis Test	
Chi-Square	12.9405
DF	11
Pr > Chi-Square	0.2972

Figure 3.14 : Kruskal-Wallis Test on number of items sold for item types

Hypothesis Testing :

H_0 : $\mu_{cosmetics} = \mu_{vegetables} = \mu_{BabyFood} = \mu_{Cereal} = \mu_{Fruits} = \mu_{Clothes} = \mu_{Snacks} = \mu_{Household} = \mu_{OfficeSupplies} = \mu_{Beverages} = \mu_{PersonalCare} = \mu_{Meat}$

H_1 : $\mu_{cosmetics} \neq \mu_{vegetables} \neq \mu_{BabyFood} \neq \mu_{Cereal} \neq \mu_{Fruits} \neq \mu_{Clothes} \neq \mu_{Snacks} \neq \mu_{Household} \neq \mu_{OfficeSupplies} \neq \mu_{Beverages} \neq \mu_{PersonalCare} \neq \mu_{Meat}$

α : 0.05

p-value : 0.2972

Decision rule : Failed to reject H_0 since p-value = 0.2972 > α = 0.05

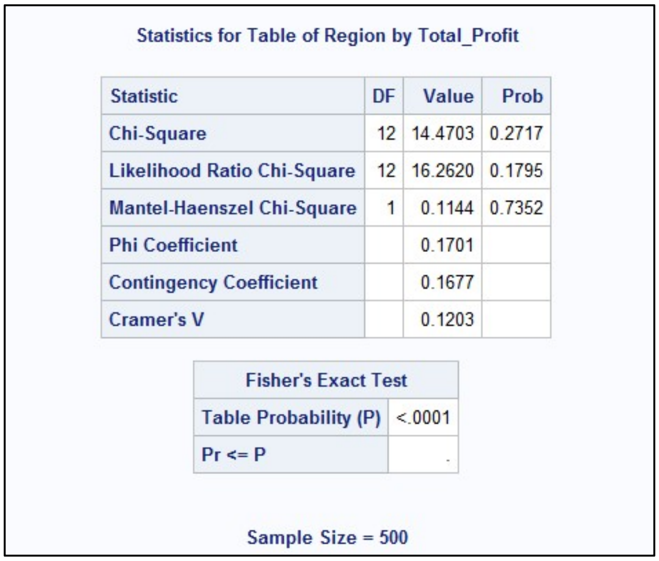
Conclusion : There is no significant mean difference on the number of items sold among the 12 item types.

Interpretation :

A non-parametric Kruskal-Wallis Test is used to compare the differences between the number of items sold and item types since the dependent variable does not follow a normal distribution. Figure 3.13 shows the output of Kruskal-Wallis Test on total number of units sold among 12 different item types. Based on the figure, it can be concluded that there is not enough evidence to indicate that there are statistically significant differences in the mean of number of units sold among the item types. This is because the p-value of the test which is 0.2972 is greater than 5% significant level and therefore, the null hypothesis is to be accepted and therefore the number of items sold are identical for every item type.

3.5.5 Chi-Square Test to Determine the Association Between Regions and Three Total Profit Levels.

```
proc freq data = project.regions;
tables region*total_profit / expected chisq fisher;
run;
```



Statistics for Table of Region by Total_Profit

Statistic	DF	Value	Prob
Chi-Square	12	14.4703	0.2717
Likelihood Ratio Chi-Square	12	16.2620	0.1795
Mantel-Haenszel Chi-Square	1	0.1144	0.7352
Phi Coefficient		0.1701	
Contingency Coefficient		0.1677	
Cramer's V		0.1203	

Fisher's Exact Test	
Table Probability (P)	<.0001
Pr <= P	.

Sample Size = 500

Figure 3.12 : Chi-Square Test on the association between regions and three total profit levels

Hypothesis Testing :

H_0 : There is no association between regions and total profit levels.

H_1 : There is an association between regions and total profit levels.

α : 0.05

p-value : 0.2717

Decision rule : Failed to reject H_0 since p-value = 0.2717 > α = 0.05

Conclusion : There is no association between regions and total profit levels.

Interpretation :

The above figure and hypothesis testing shows that there is not enough evidence to indicate that there is a relationship between regions and total profit levels since the p-value of the Chi-Square test is 0.2717 which is greater than the significant value of 0.05. Therefore, both regions and total profit levels are independent to each other.

CHAPTER 4 : CONCLUSION

As a conclusion for this study regarding the sales recorded by several regions, descriptive statistics procedures such as summary statistic, bar charts and pie charts have been used to produce a better interactive and understanding results. Furthermore, the researchers have successfully applied inferential statistics such as normality test to determine whether the data follows a normal distribution and advanced statistical analysis have also been deployed to answer the objectives of this study.

It was found that the total revenue for both Asia and Europe region does not follow a normal distribution and therefore, a non-parametric Man-Whitney U Test was used. The result indicates that the total revenue is identical for both Asia and Europe region. It can be said that both regions are very competitive in obtaining revenue from the items sold.

A non-parametric was also used to determine whether there is a significant difference in the number of items sold among different item types since there is no sufficient evidence to indicate that the data is normal. As a result, the mean number of items sold by 12 different item types are the same. This shows that all of the items sold recorded in this study are important in determining the profit and revenue obtained from all the regions.

Lastly, a Chi-Square Test was used to observe the association between all regions and three total profit levels. The profit levels are categorised into low, moderate and high. It can be seen that there is no obvious relationship between regions and the total profit obtained and hence, regions and total profit are independent to each other. It may come to a conclusion that regions are not the main factor for the total profit obtained in this study since they did not show any relationship.

As a recommendation for this study, future researchers should consider on using a primary data instead of a secondary data since it will be more accurate for analysis and decision making. Not only that, more observations should be used to fulfil the normality assumption of an analysis so that a parametric test can be conducted.

REFERENCE

Charlot Bennett, M. E. (2016). *SAS Programming 1 : Essentials Course Notes*. North Carolina: SAS Institute Inc.

Charlot Bennett, M. E. (2016). *SAS Programming 2: Data Manipulation Techniques.* North Carolina: SAS Institute Inc.

E for Excel. (n.d.). Retrieved from Sample Data Sets for Testing: http://eforexcel.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/

Law, V. (2004). SAS Programming; The One Day Course. *Journal of Statistical Software*.

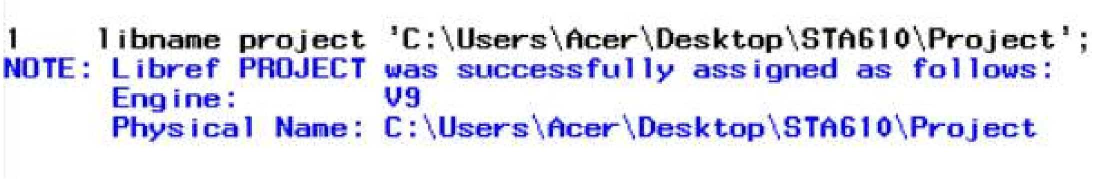
Twin, A. (17 February, 2020). *Sale*. Retrieved from Investopedia: https://www.investopedia.com/terms/s/sale.asp

APPENDICES

Appendix A : Data Management

SAS Data Library

```
libname project 'C:\Users\Acer\Desktop\STA610\Project';
```

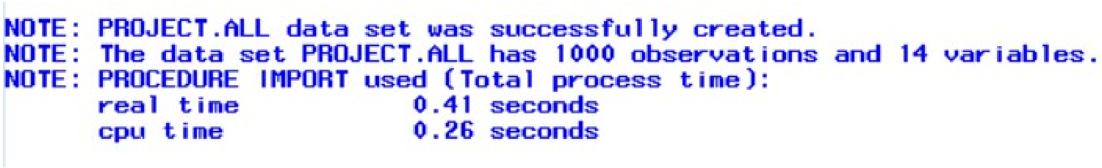


```
1     libname project 'C:\Users\Acer\Desktop\STA610\Project';
NOTE: Libref PROJECT was successfully assigned as follows:
      Engine:        V9
      Physical Name: C:\Users\Acer\Desktop\STA610\Project
```

Appendix 1

Proc Import

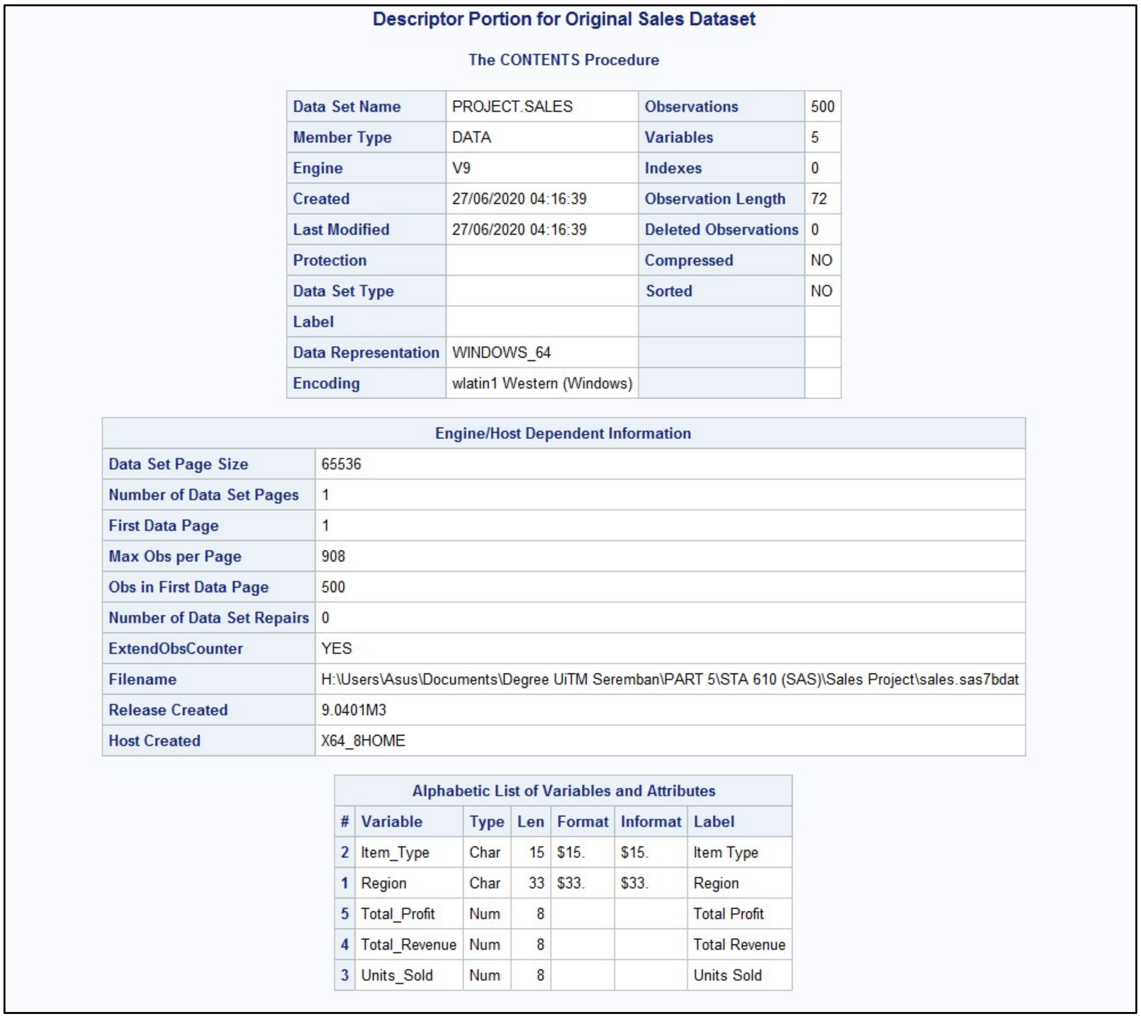
```
proc import       out = project.all
datafile = "C:\Users\Acer\Desktop\STA610\Project\Sales Records.xlsx"
dbms = excel;
run;
```



```
NOTE: PROJECT.ALL data set was successfully created.
NOTE: The data set PROJECT.ALL has 1000 observations and 14 variables.
NOTE: PROCEDURE IMPORT used (Total process time):
      real time           0.41 seconds
      cpu time            0.26 seconds
```

Appendix 2

Appendix B : Descriptor Portion



Descriptor Portion for Original Sales Dataset

The CONTENTS Procedure

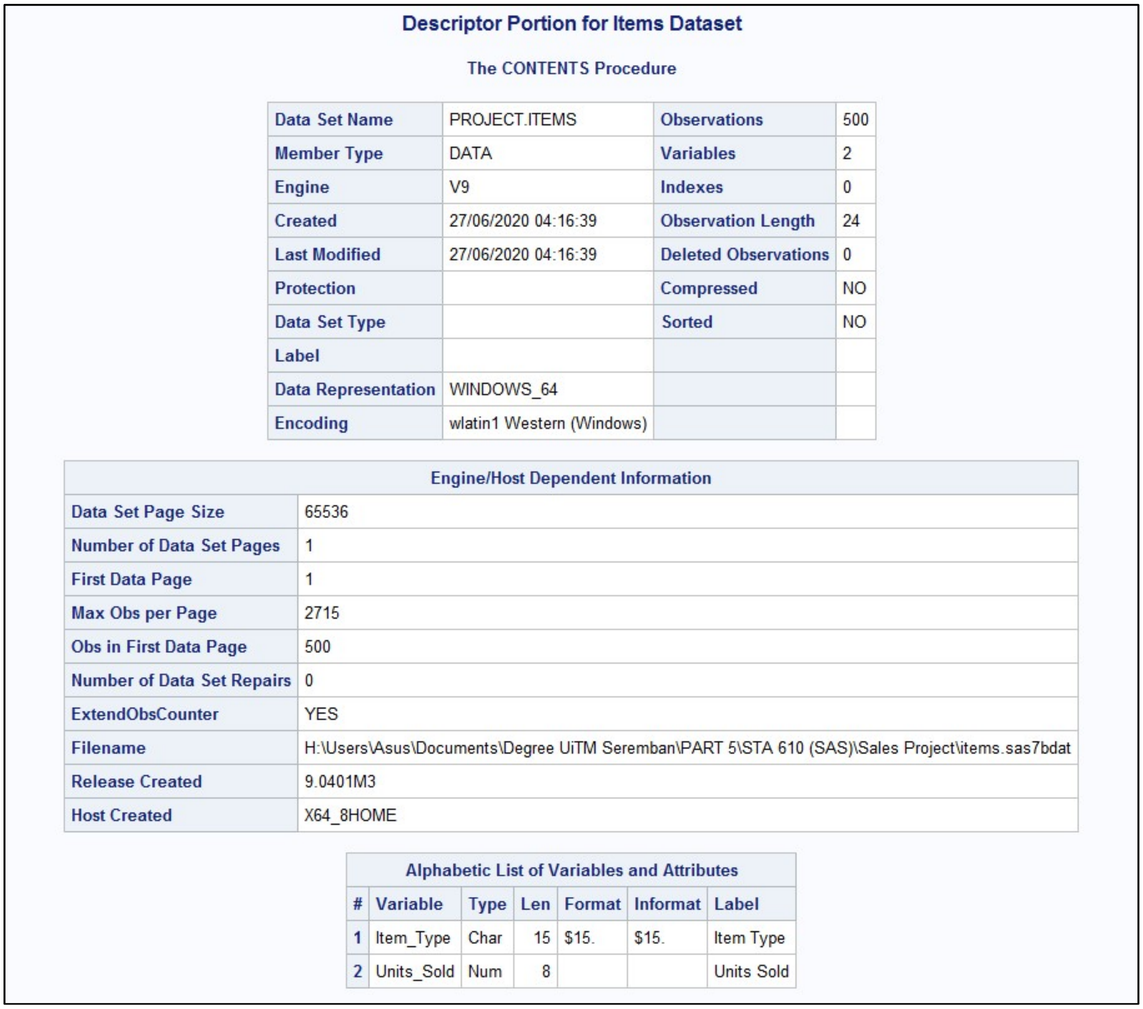
Data Set Name	PROJECT.SALES	Observations	500
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	27/06/2020 04:16:39	Observation Length	72
Last Modified	27/06/2020 04:16:39	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	908
Obs in First Data Page	500
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	H:\Users\Asus\Documents\Degree UiTM Seremban\PART 5\STA 610 (SAS)\Sales Project\sales.sas7bdat
Release Created	9.0401M3
Host Created	X64_8HOME

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
2	Item_Type	Char	15	$15.	$15.	Item Type
1	Region	Char	33	$33.	$33.	Region
5	Total_Profit	Num	8			Total Profit
4	Total_Revenue	Num	8			Total Revenue
3	Units_Sold	Num	8			Units Sold

Figure A.1 : Descriptor Portion for Original Sales Dataset



Descriptor Portion for Items Dataset

The CONTENTS Procedure

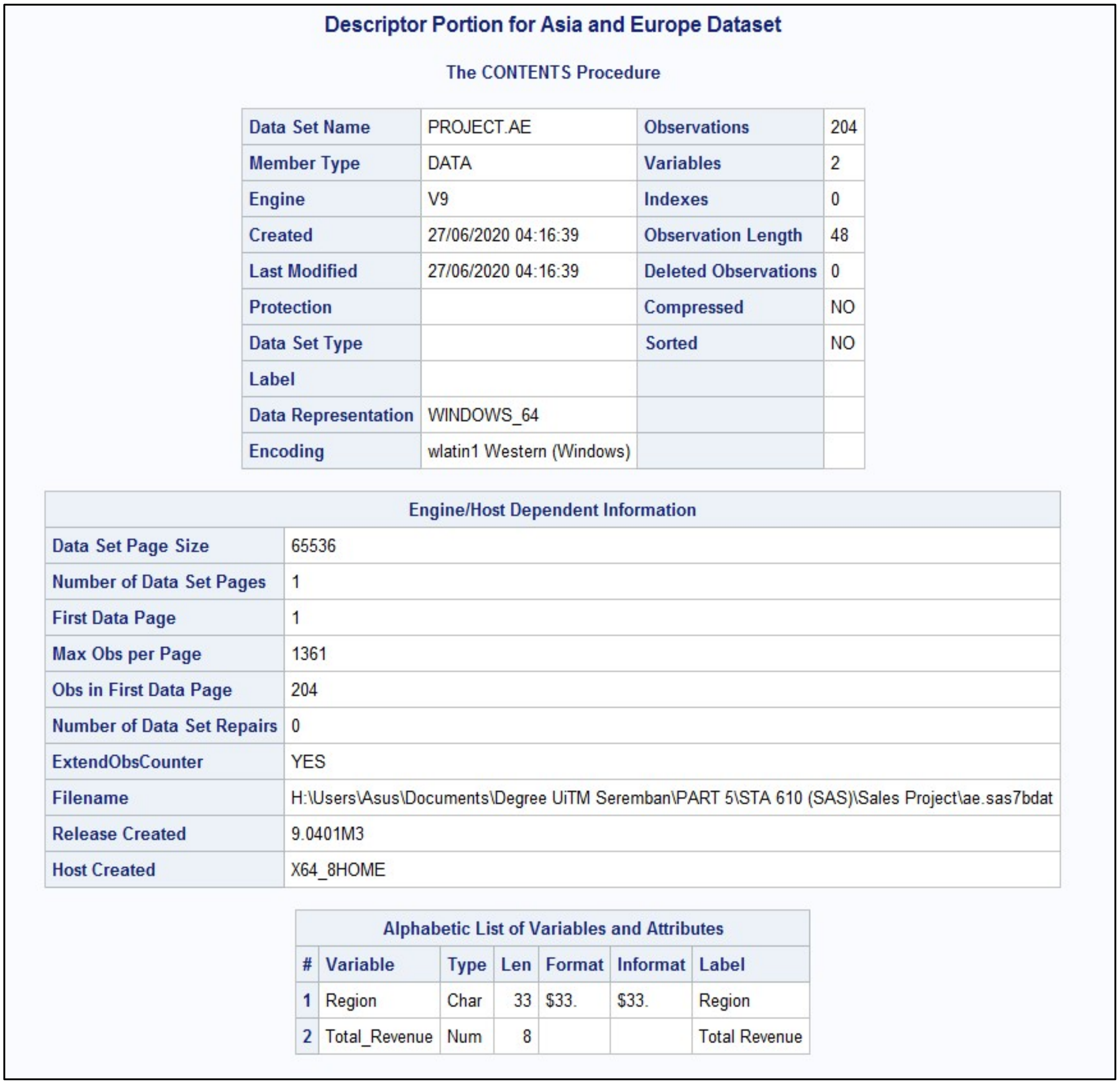
Data Set Name	PROJECT.ITEMS	Observations	500
Member Type	DATA	Variables	2
Engine	V9	Indexes	0
Created	27/06/2020 04:16:39	Observation Length	24
Last Modified	27/06/2020 04:16:39	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	2715
Obs in First Data Page	500
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	H:\Users\Asus\Documents\Degree UiTM Seremban\PART 5\STA 610 (SAS)\Sales Project\items.sas7bdat
Release Created	9.0401M3
Host Created	X64_8HOME

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
1	Item_Type	Char	15	$15.	$15.	Item Type
2	Units_Sold	Num	8			Units Sold

Figure A.2 : Descriptor Portion for Items Dataset



Descriptor Portion for Asia and Europe Dataset

The CONTENTS Procedure

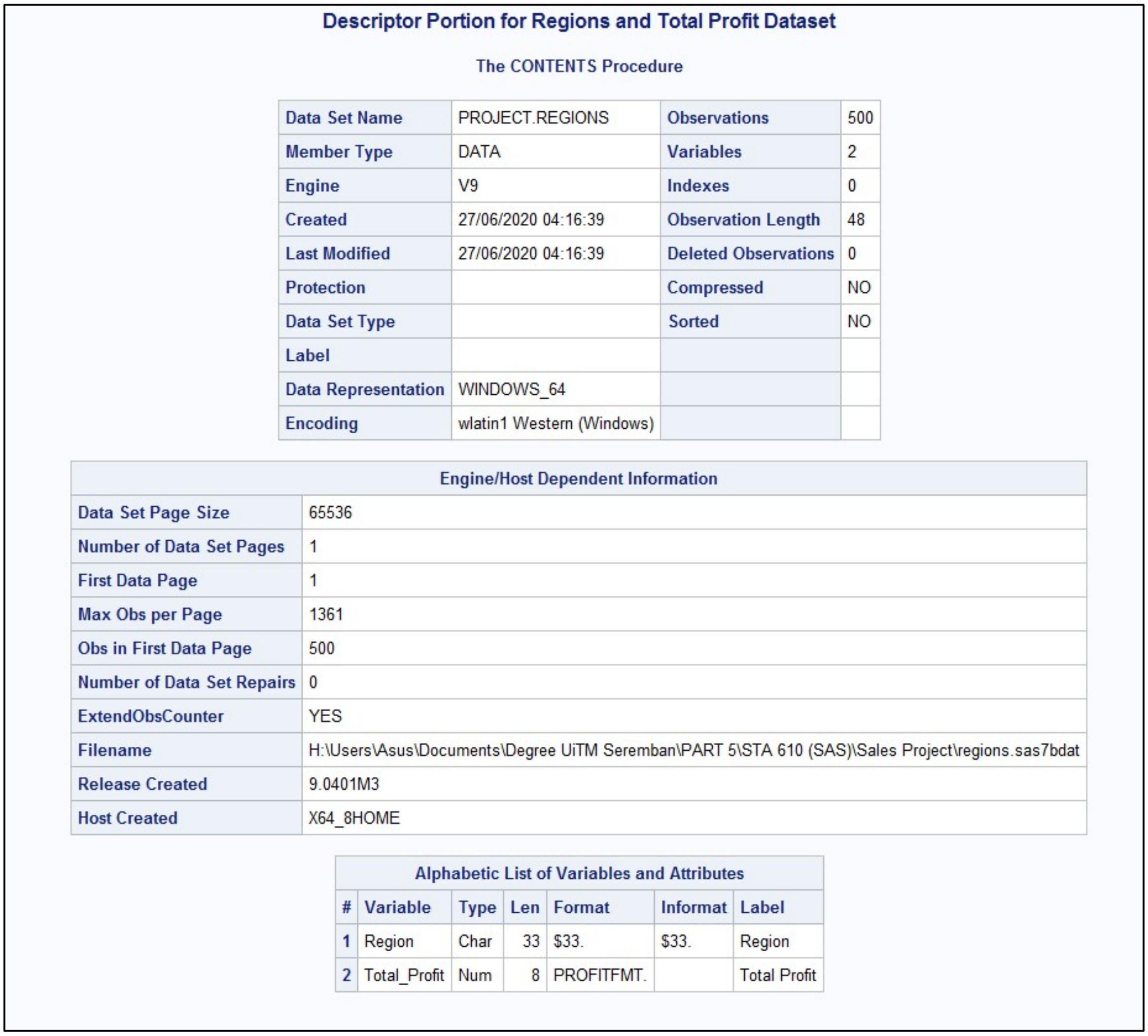
Data Set Name	PROJECT.AE	Observations	204
Member Type	DATA	Variables	2
Engine	V9	Indexes	0
Created	27/06/2020 04:16:39	Observation Length	48
Last Modified	27/06/2020 04:16:39	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	1361
Obs in First Data Page	204
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	H:\Users\Asus\Documents\Degree UiTM Seremban\PART 5\STA 610 (SAS)\Sales Project\ae.sas7bdat
Release Created	9.0401M3
Host Created	X64_8HOME

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
1	Region	Char	33	$33.	$33.	Region
2	Total_Revenue	Num	8			Total Revenue

Figure A.3 : Descriptor Portion for Asia and Europe Dataset



Descriptor Portion for Regions and Total Profit Dataset

The CONTENTS Procedure

Data Set Name	PROJECT.REGIONS	Observations	500
Member Type	DATA	Variables	2
Engine	V9	Indexes	0
Created	27/06/2020 04:16:39	Observation Length	48
Last Modified	27/06/2020 04:16:39	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

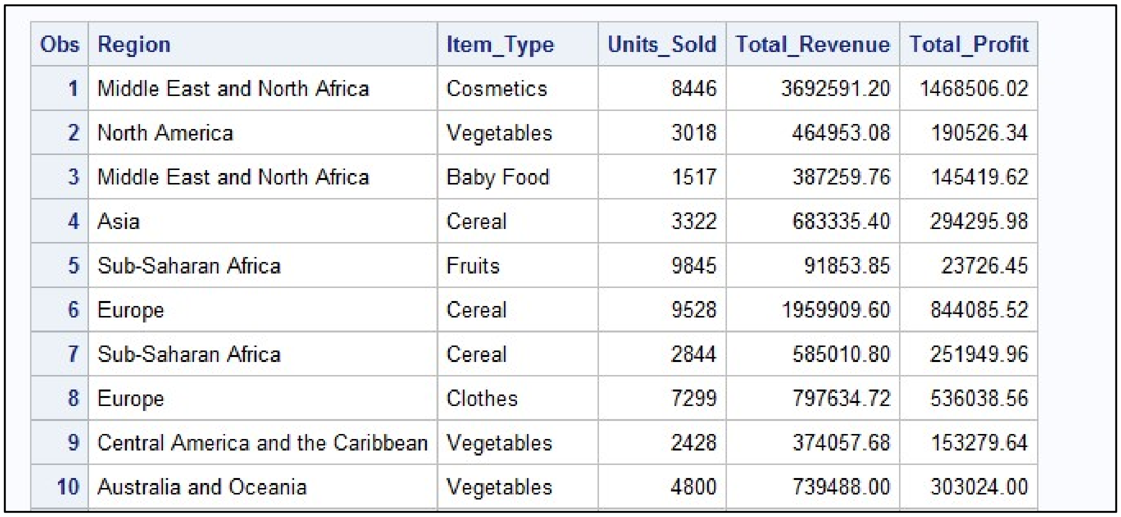
Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	1361
Obs in First Data Page	500
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	H:\Users\Asus\Documents\Degree UiTM Seremban\PART 5\STA 610 (SAS)\Sales Project\regions.sas7bdat
Release Created	9.0401M3
Host Created	X64_8HOME

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
1	Region	Char	33	$33.	$33.	Region
2	Total_Profit	Num	8	PROFITFMT.		Total Profit

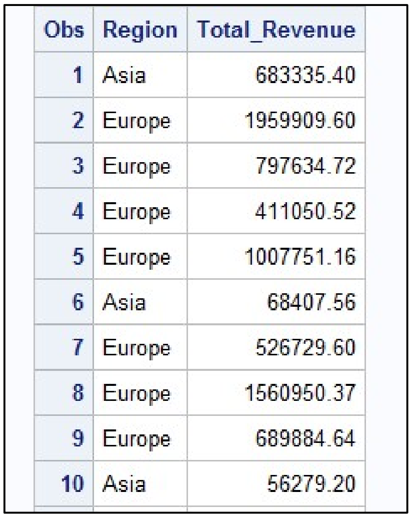
Figure A.4 : Descriptor Portion for Regions and Total Profit Dataset

Appendix C: Printing Partial Dataset



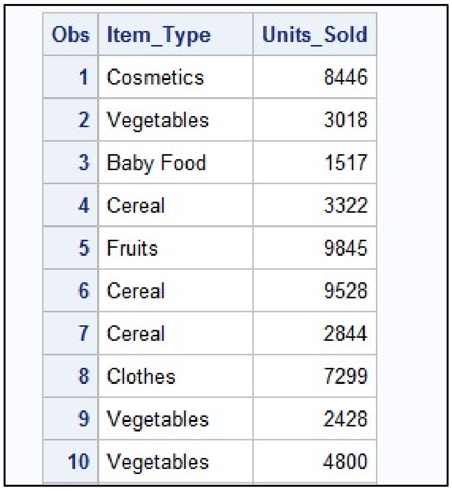
Obs	Region	Item_Type	Units_Sold	Total_Revenue	Total_Profit
1	Middle East and North Africa	Cosmetics	8446	3692591.20	1468506.02
2	North America	Vegetables	3018	464953.08	190526.34
3	Middle East and North Africa	Baby Food	1517	387259.76	145419.62
4	Asia	Cereal	3322	683335.40	294295.98
5	Sub-Saharan Africa	Fruits	9845	91853.85	23726.45
6	Europe	Cereal	9528	1959909.60	844085.52
7	Sub-Saharan Africa	Cereal	2844	585010.80	251949.96
8	Europe	Clothes	7299	797634.72	536038.56
9	Central America and the Caribbean	Vegetables	2428	374057.68	153279.64
10	Australia and Oceania	Vegetables	4800	739488.00	303024.00

Figure B.1 : Partial Data for Original Sales Dataset



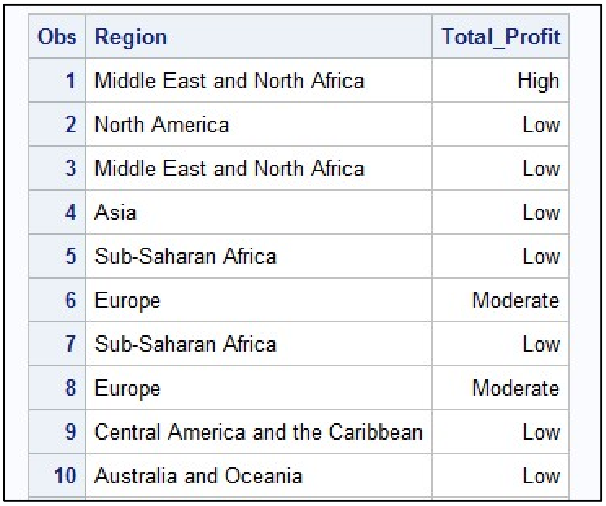
Obs	Region	Total_Revenue
1	Asia	683335.40
2	Europe	1959909.60
3	Europe	797634.72
4	Europe	411050.52
5	Europe	1007751.16
6	Asia	68407.56
7	Europe	526729.60
8	Europe	1560950.37
9	Europe	689884.64
10	Asia	56279.20

Figure B.2 : Partial Data for Asia and Europe Dataset



Obs	Item_Type	Units_Sold
1	Cosmetics	8446
2	Vegetables	3018
3	Baby Food	1517
4	Cereal	3322
5	Fruits	9845
6	Cereal	9528
7	Cereal	2844
8	Clothes	7299
9	Vegetables	2428
10	Vegetables	4800

Figure B.3 : Partial Data for Items Dataset



Obs	Region	Total_Profit
1	Middle East and North Africa	High
2	North America	Low
3	Middle East and North Africa	Low
4	Asia	Low
5	Sub-Saharan Africa	Low
6	Europe	Moderate
7	Sub-Saharan Africa	Low
8	Europe	Moderate
9	Central America and the Caribbean	Low
10	Australia and Oceania	Low

Figure B.4 : Partial Data for Regions and Total Profit Dataset

Appendix D : Normality Test for Total Revenue between Asia and Europe

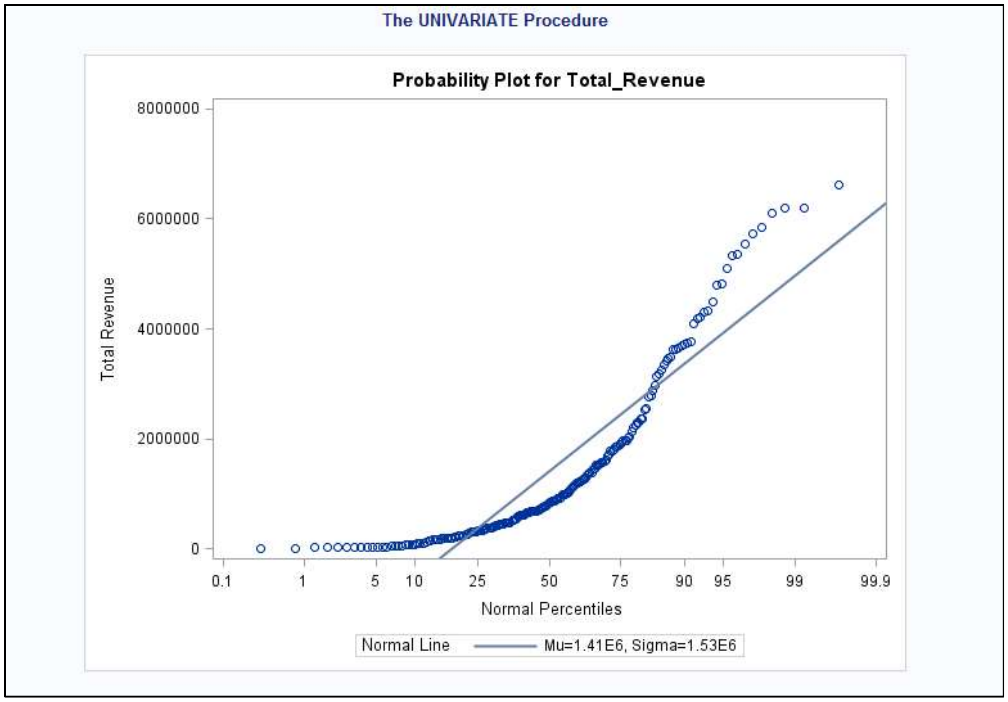


Figure C.1 : Normal P-P Plot for Total Revenue between Asia and Europe Region

Appendix E : Normality Test for Number of Items Sold

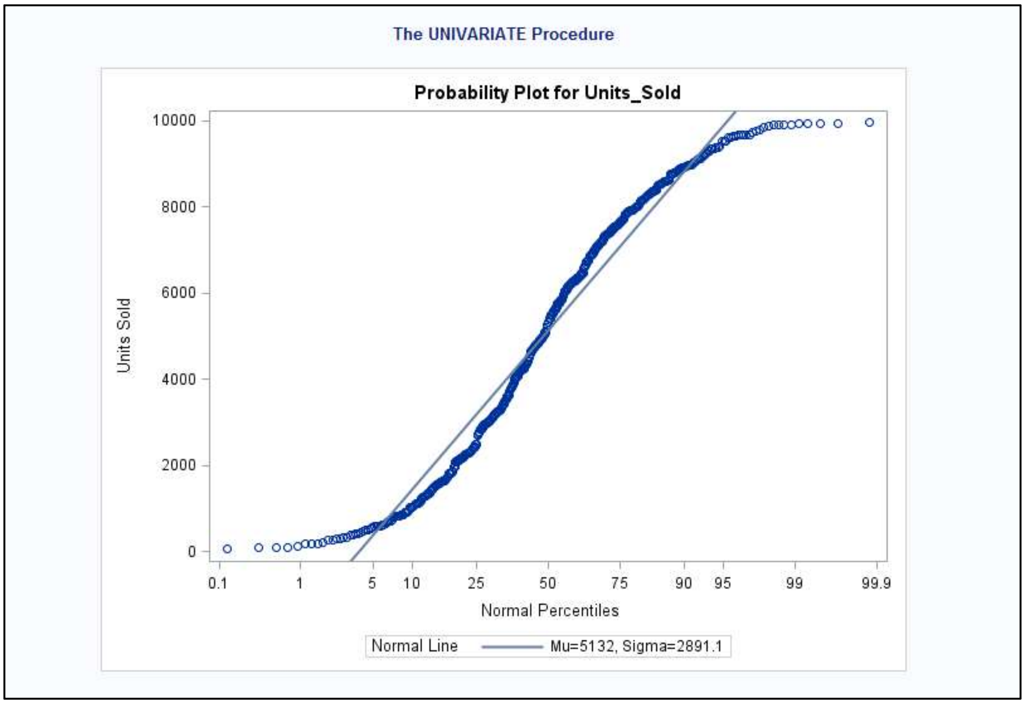


Figure D.1 : Normal P-P Plot for Number of Items Sold

Appendix F : Box Plot for Total Revenue between Asia and Europe

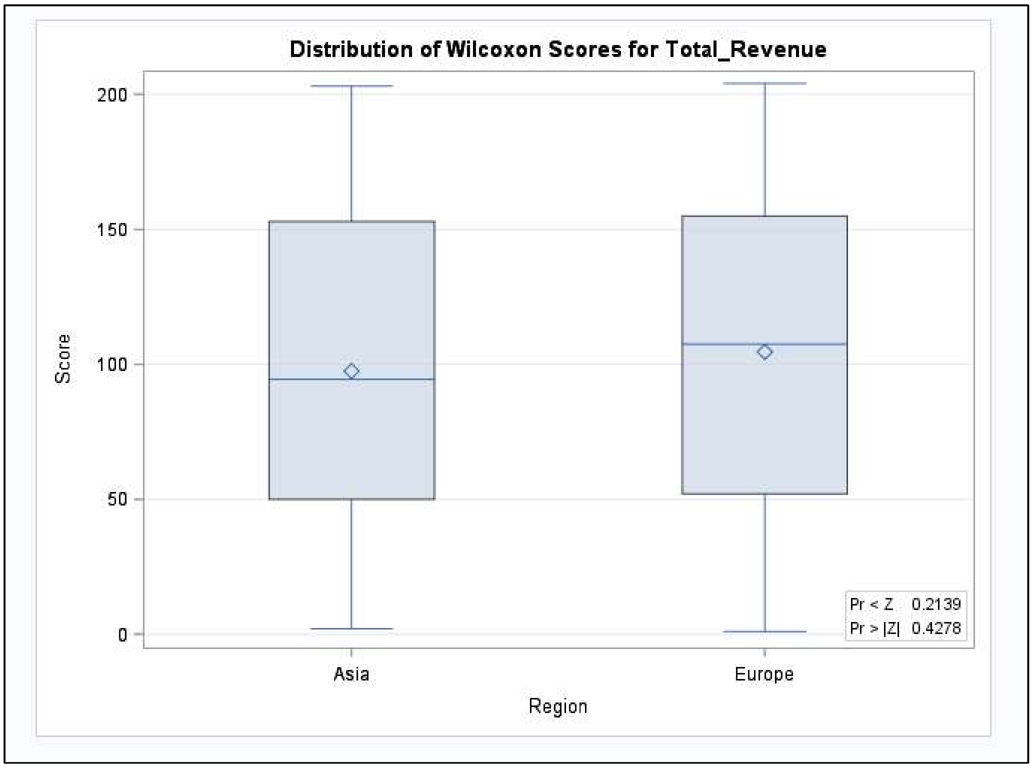


Figure E.1 : Box Plot of Wilcoxon Scores for Total Revenue Among Asia and Europe

Appendix G : Box Plot for Number of Items Sold among Item Types

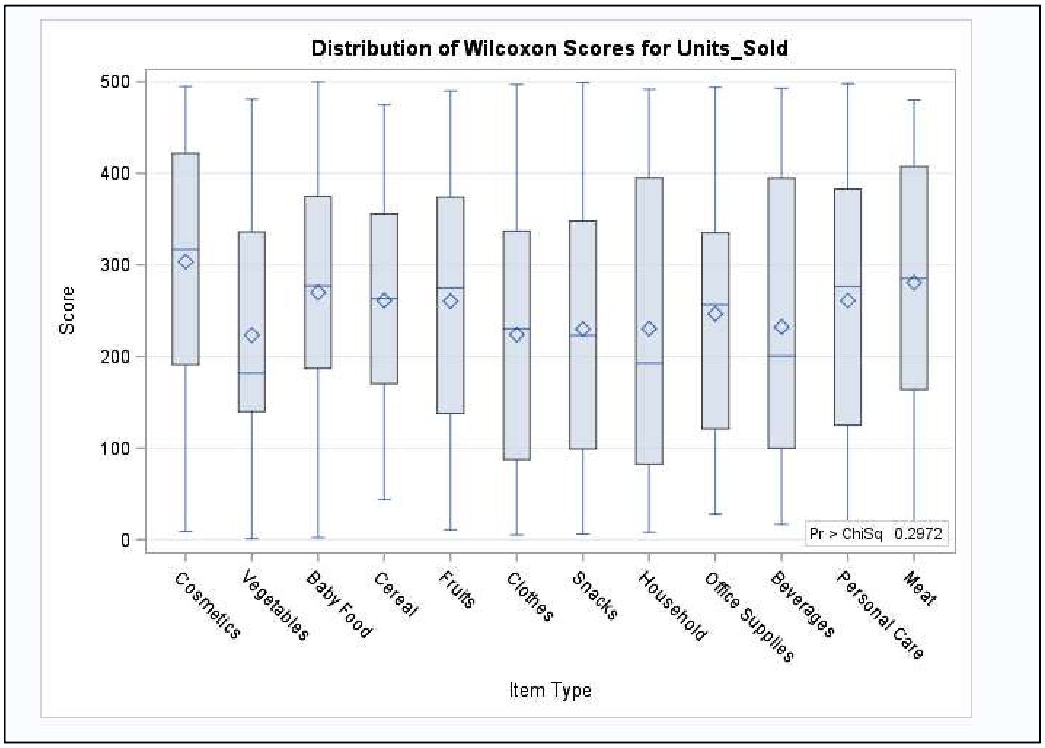


Figure F.1 : Box Plot of Wilcoxon Scores for Number of Units Sold Among Item Types