VIT
BANGALORE

ONLINE RETAIL
REGRESSION ANALYSIS USING MINITAB

STUDENT
NAME:DHINESH.S.P
REGISTER NUMBER:
23MSP3032

Table of Contents

1. Introduction

Determining and measuring the link between one or more independent variables and a dependent variable is the aim of this endeavor. Understanding the ANOVA, model validation, and Chi-square test is another project goal.

Online Retail Analysis dataset :

Online Retail - UCI Machine Learning Repository

This project is important because it provides insight into the correlations between variables, enabling data-driven decision-making. Identifying important variables, forming predictions, and testing hypotheses are helpful.

2. Dataset Description

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Source of dataset::

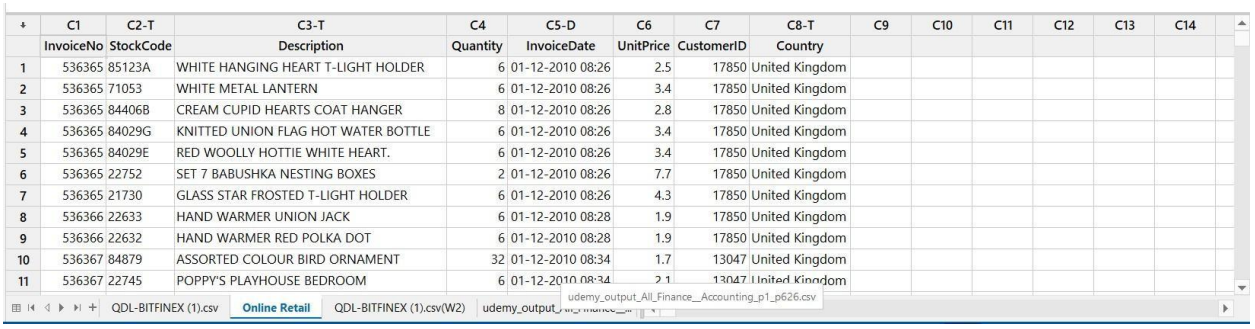
Online Retail - UCI Machine Learning Repository

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
InvoiceNo	ID	Categorical		a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation		no
StockCode	ID	Categorical		a 5-digit integral number uniquely assigned to each distinct product		no
Description	Feature	Categorical		product name		no
Quantity	Feature	Integer		the quantities of each product (item) per transaction		no
InvoiceDate	Feature	Date		the day and time when each transaction was generated		no
UnitPrice	Feature	Continuous		product price per unit	sterling	no
CustomerID	Feature	Categorical		a 5-digit integral number uniquely assigned to each customer		no
Country	Feature	Categorical		the name of the country where each customer resides		no

3.y Data Analysis and Visualization

a) Data Analysis:

- Launch Minitab and open a new project.
- **Import your dataset** by going to "File" > "Open Worksheet" or "File" > "Import Data." Ensure your data is in a compatible format (e.g., CSV, Excel).



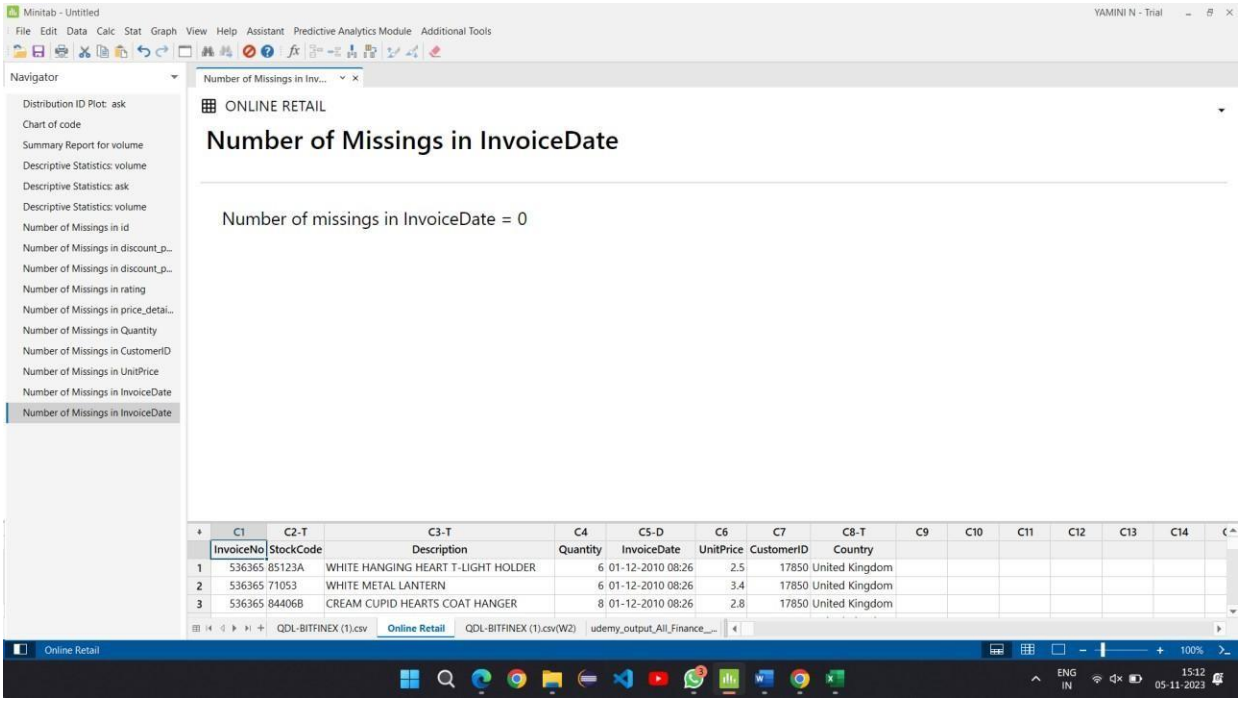
	C1	C2-T	C3-T	C4	C5-D	C6	C7	C8-T	C9	C10	C11	C12	C13	C14
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country						
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.5	17850	United Kingdom						
2	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.4	17850	United Kingdom						
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.8	17850	United Kingdom						
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.4	17850	United Kingdom						
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.4	17850	United Kingdom						
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010 08:26	7.7	17850	United Kingdom						
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01-12-2010 08:26	4.3	17850	United Kingdom						
8	536366	22633	HAND WARMER UNION JACK	6	01-12-2010 08:28	1.9	17850	United Kingdom						
9	536366	22632	HAND WARMER RED POLKA DOT	6	01-12-2010 08:28	1.9	17850	United Kingdom						
10	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	01-12-2010 08:34	1.7	13047	United Kingdom						
11	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	01-12-2010 08:34	2.1	13047	United Kingdom						

udemy_output_All_Finance__Accounting_p1_p626.csv

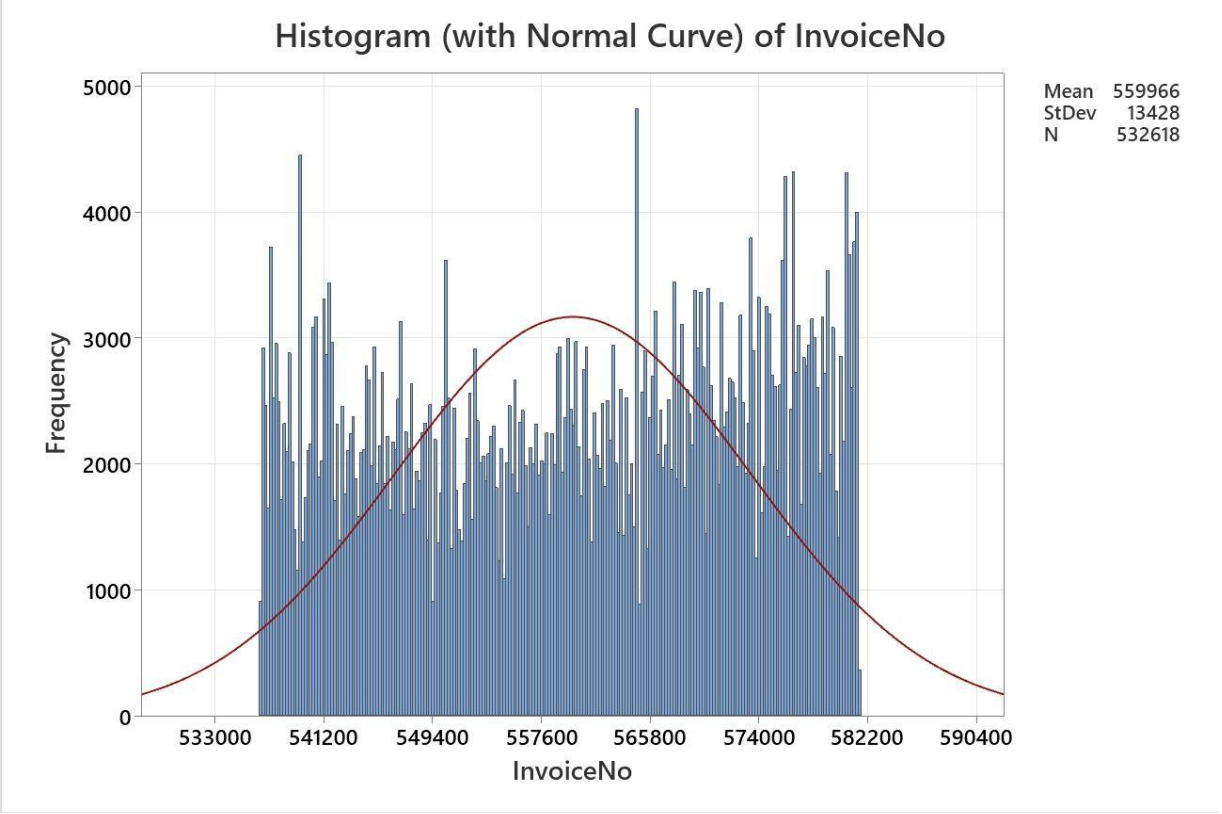
QDL-BITFINEX (1).csv | Online Retail | QDL-BITFINEX (1).csv(W2) | udemy_output_...

TO CHECK MISSING VALUES :

Calc > Column Statistics. Select N missing option. Provide the column name in the input variable and click on ok.



To fill the missing values, check the datatype of the column first.
If it is categorical, fill it with the mode value.
If it is continuous, check the skewness.
To check skewness, go to stat > Basic Statistics > Display Descriptive Statistics.
Provide the column name and in the statistic option, select skewness and press ok and in the graphs option, select histogram of data, with normal curve and press ok. Again press ok.
If the skewness is 0 i.e., the data is uniform normally distributed, fill it with the mean value. If not, fill it with the median value.

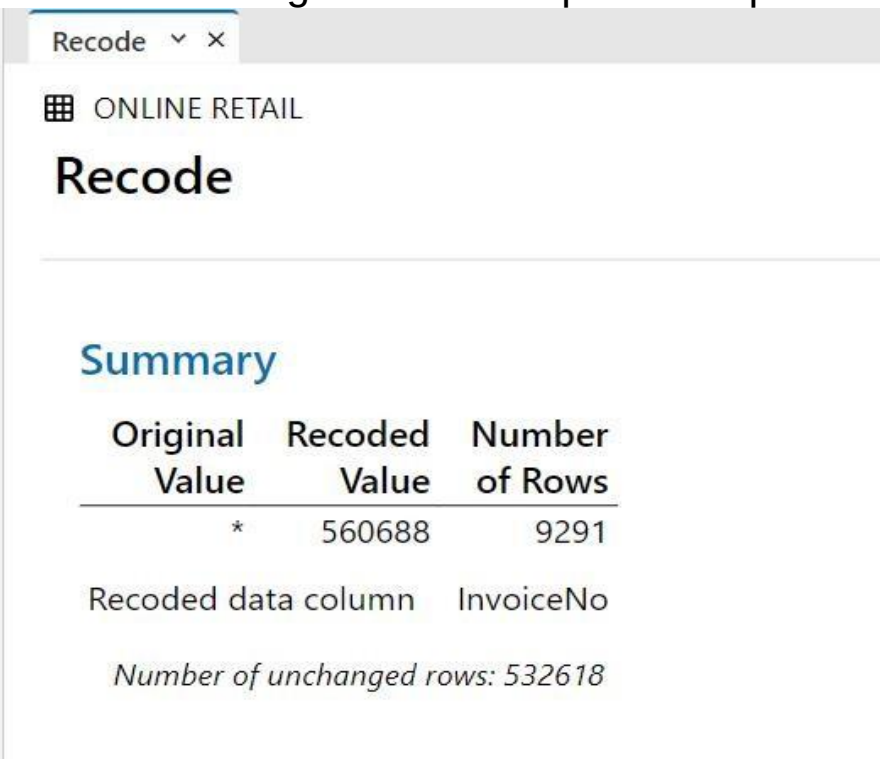


Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
InvoiceNo	532618	9291	559966	18.4	13428	536365	547906	560688	571841	581587

Variable	Skewness
InvoiceNo	-0.11

In this case, the data is negatively skewed. So, fill it with the median value.
Go to Data > Recode > To Numeric. Select the column name, then in method, select recode a single value option. Provide the current value as * and provide the recorded value with the median value. In the storage location for the recorded columns, select in the original column option and press ok.



Anomalies(Outliers):
For detecting outliers, go to stat > Basic Statistics > Outlier Test. Provide the column names and press ok.

ONLINE RETAIL

Outlier Test: UnitPrice

Method

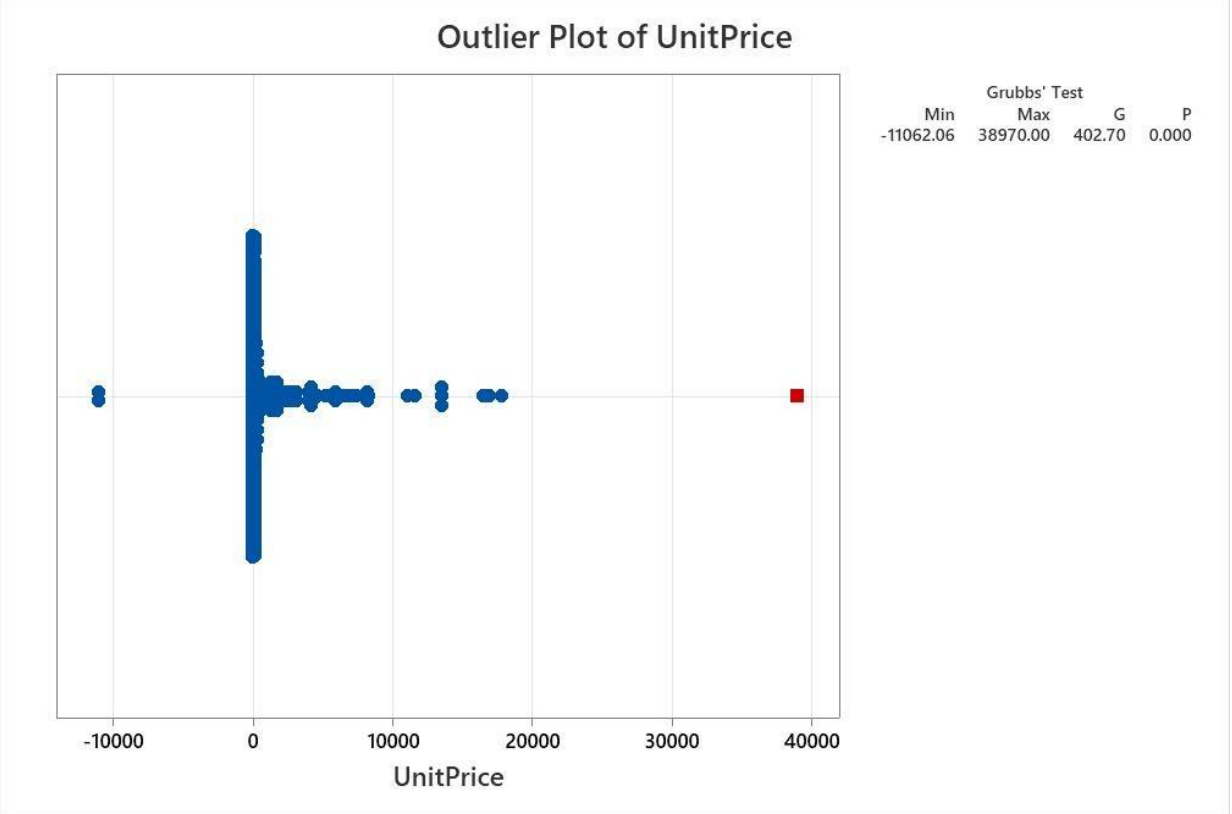
Null hypothesis	All data values come from the same normal population
Alternative hypothesis	Smallest or largest data value is an outlier
Significance level	α = 0.05

Grubbs' Test

Variable	N	Mean	StDev	Min	Max	G	P
UnitPrice	541909	4.61	96.8	-11062.1	38970.0	402.70	0.000

Outlier

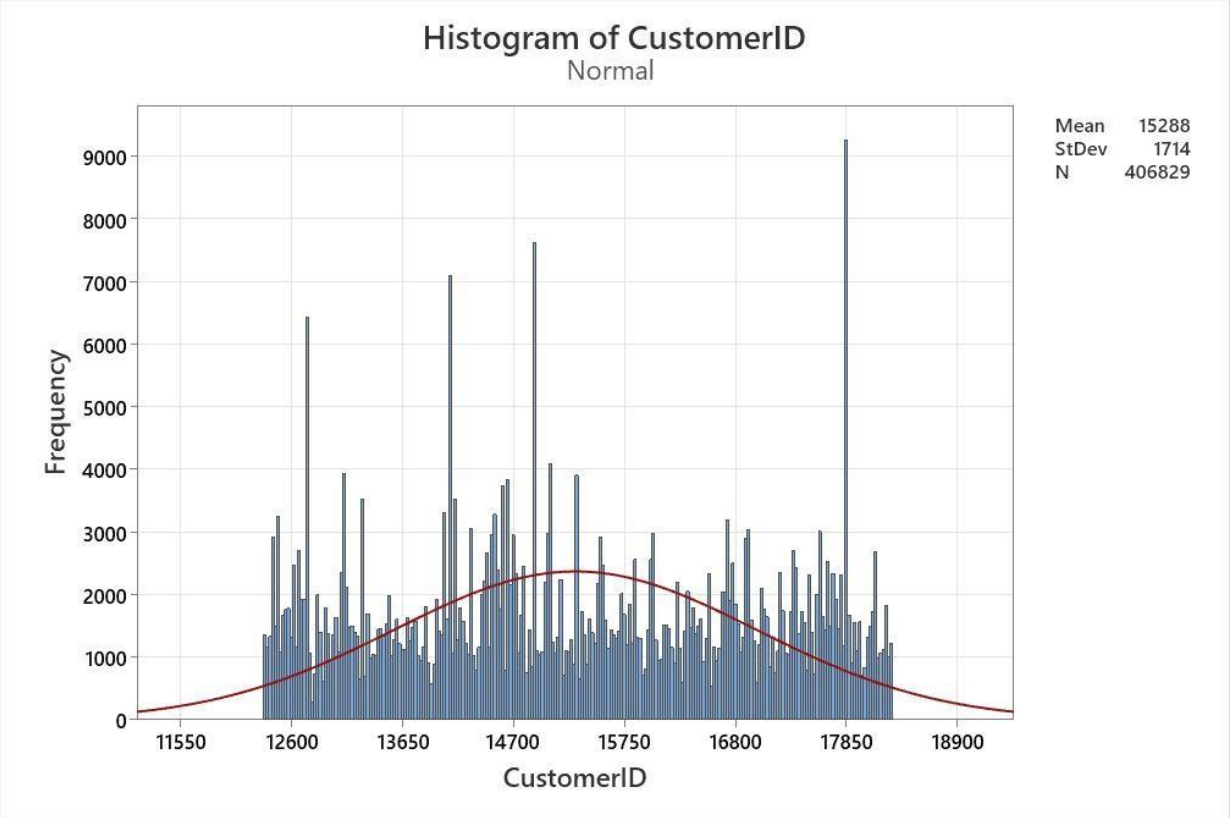
Variable	Row	Outlier
UnitPrice	222682	38970



b) VISUALISATIONS:

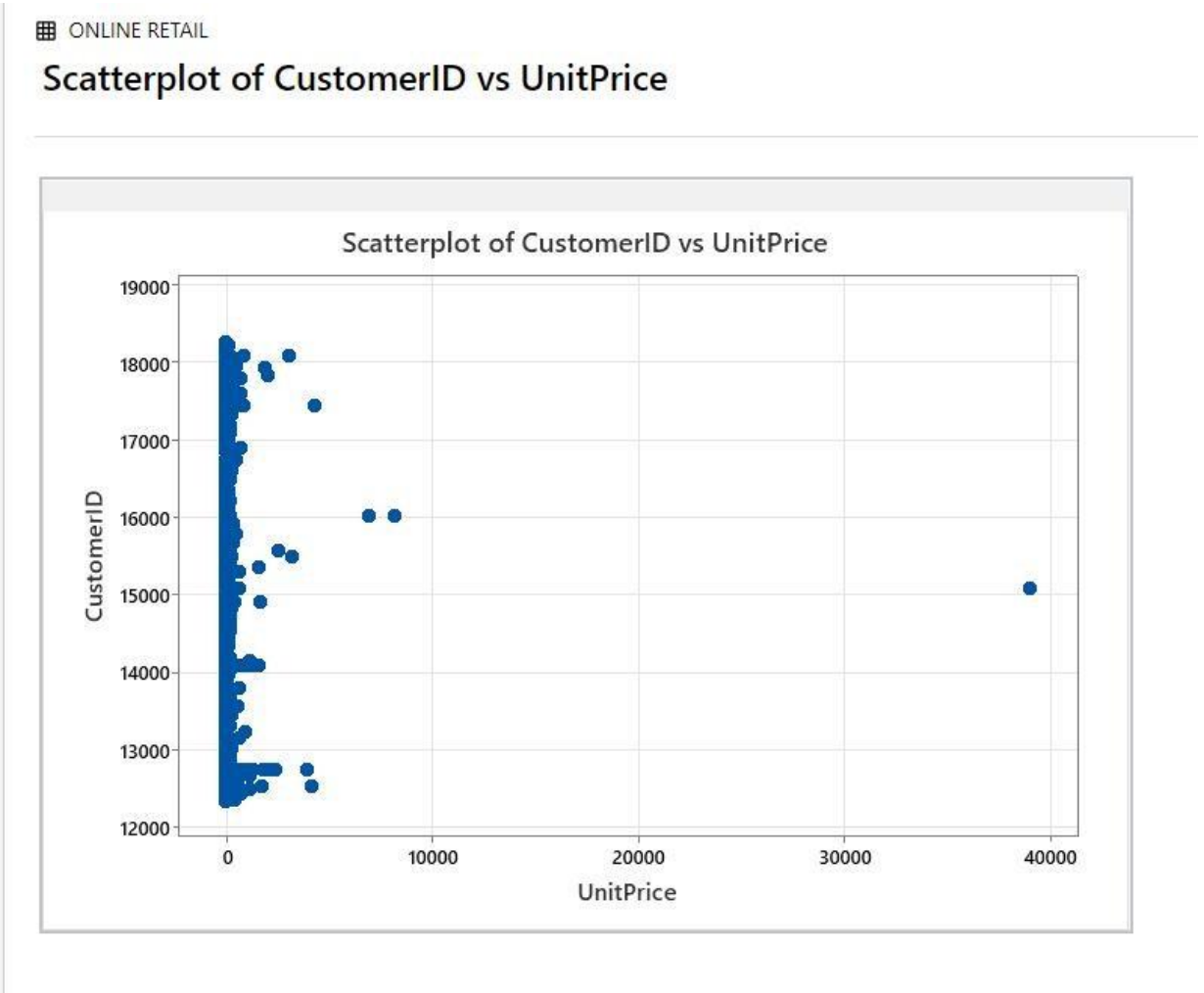
Histogram/Density Plot:

Go to Graph > Histogram. Select with fit option. Provide the column name and press ok.



The mean is 15288, standard deviation is 1714. The data is positively skewed.

Scatter Plot:
Go to Graph > Scatterplot. Select a simple option. Provide the column names for x and y and press ok

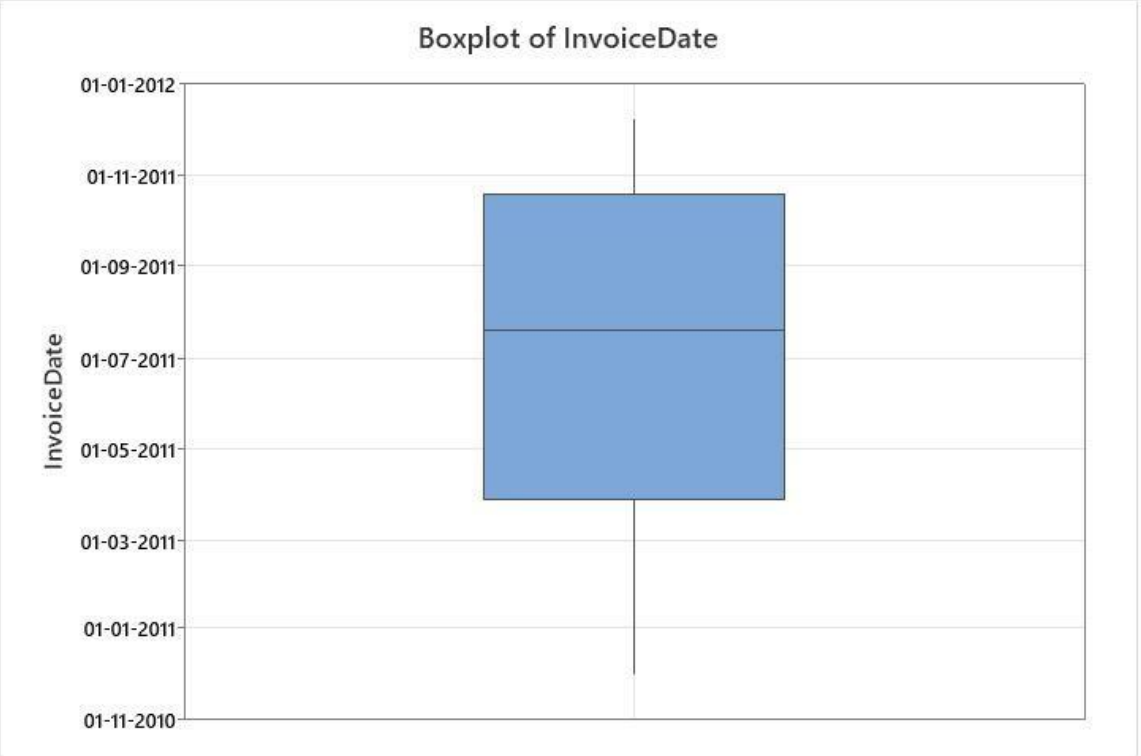


Box Plot:

Go to Graph > Boxplot. Select simple option. Provide the column name and press ok.

ONLINE RETAIL

Boxplot of InvoiceDate



Probability Distribution Analysis:

Go to Calc > Probability Distribution > Uniform.
Select the Probability Density Option. Provide the Minimum value of
descriptive statistics in the lower endpoint and the Minimum value of descriptive statistics in the upper endpoint. Provide the target
column in the input column and press OK.

ONLINE RETAIL

Probability Density Function

Continuous uniform on 0 to 1

x	f(x)
2.5	0
3.4	0
2.8	0
3.4	0
3.4	0
7.7	0
4.3	0
1.9	0
1.9	0
1.7	0
2.1	0
2.1	0
3.8	0
1.6	0

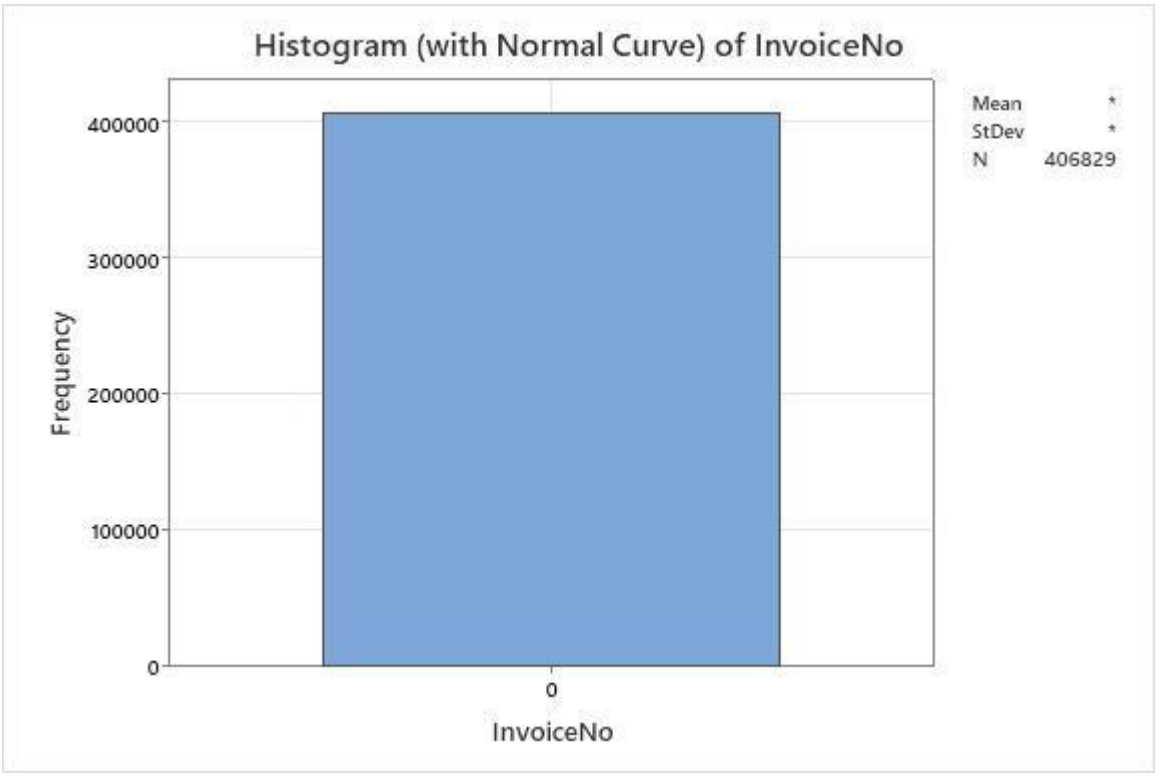
4. Descriptive Statistics

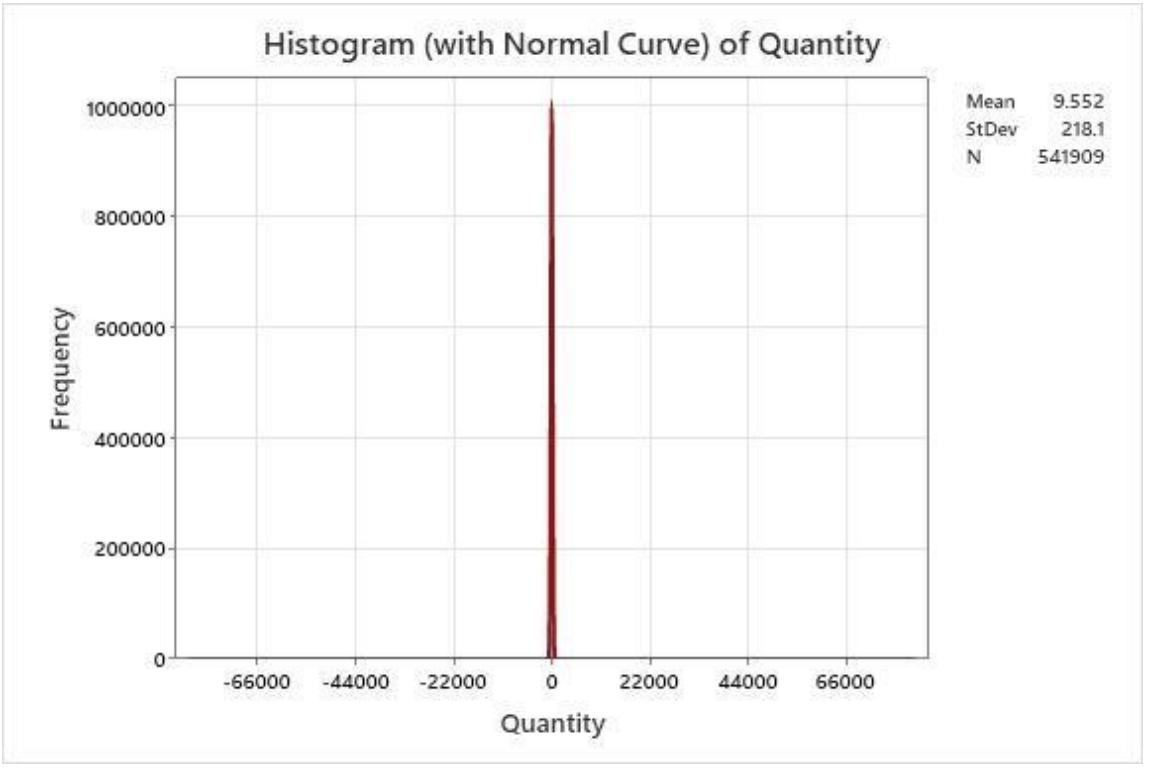
Go to Stat > Basic Statistics > Display Descriptive Statistics. Provide the required column names for variables. In the Statistics option, select the options mean, median, mode, range, variance, standard deviation, skewness and kurtosis. In the Graphs option, select histogram of data, with normal curve option and press ok.

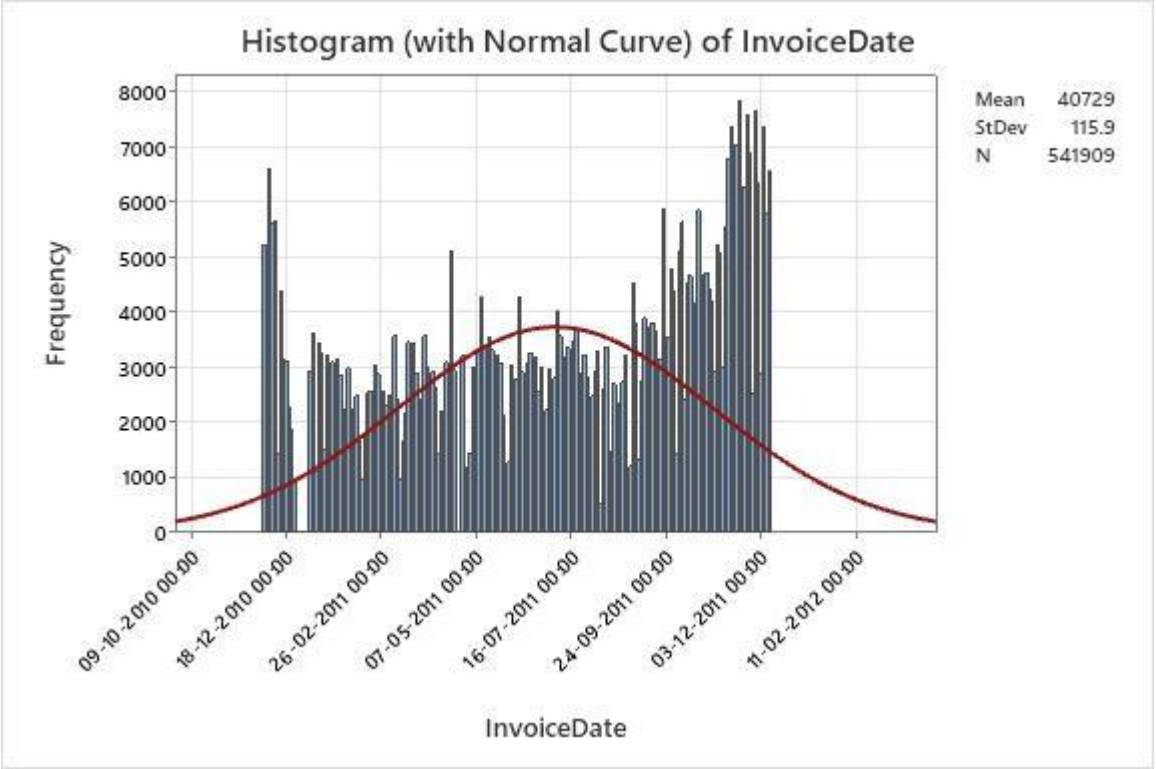
Statistics

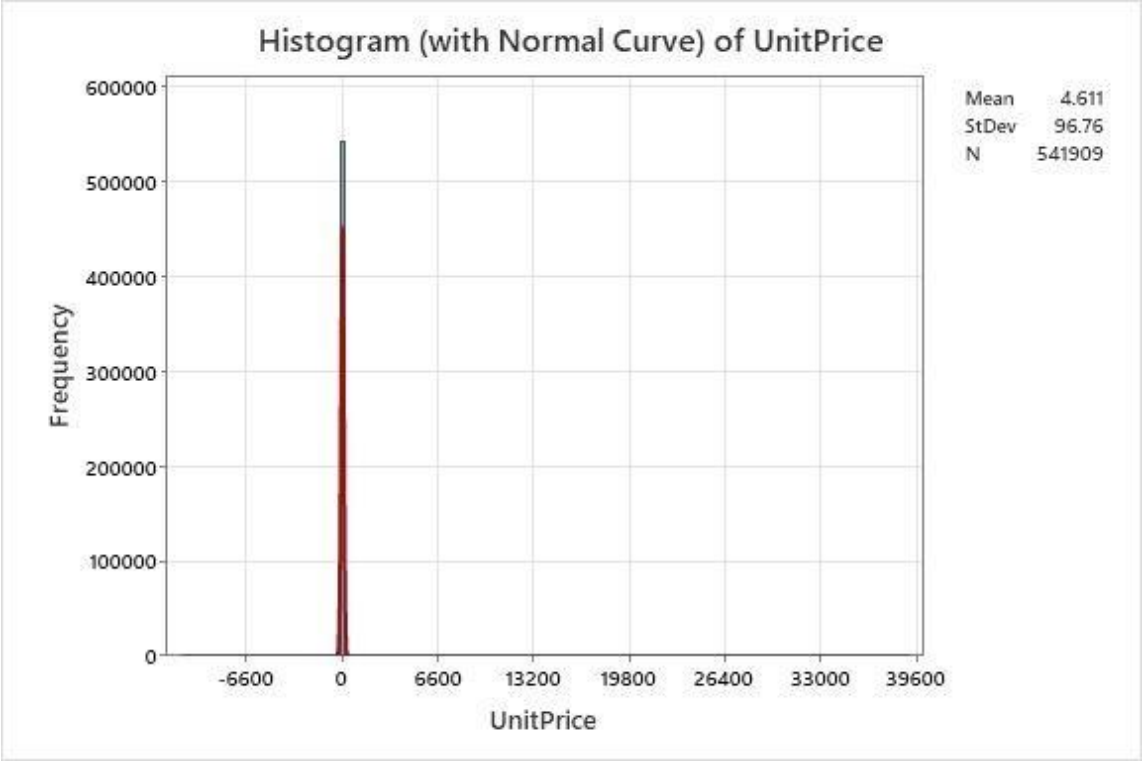
Variable	N	N*	Mean	SE Mean	StDev	Variance	CoefVar	Q1
InvoiceNo	406829	135080	0.000000	0.000000	0.000000	0.000000	*	0.000000
Quantity	541909	0	9.55	0.296	218.1	47559.4	2283.03	1.00
InvoiceDate	541909	0	40729	0.157	116	13428	0.28	40630
UnitPrice	541909	0	4.61	0.131	96.8	9362.5	2098.41	1.25
CustomerID	406829	135080	15288	2.69	1714	2936426	11.21	13953

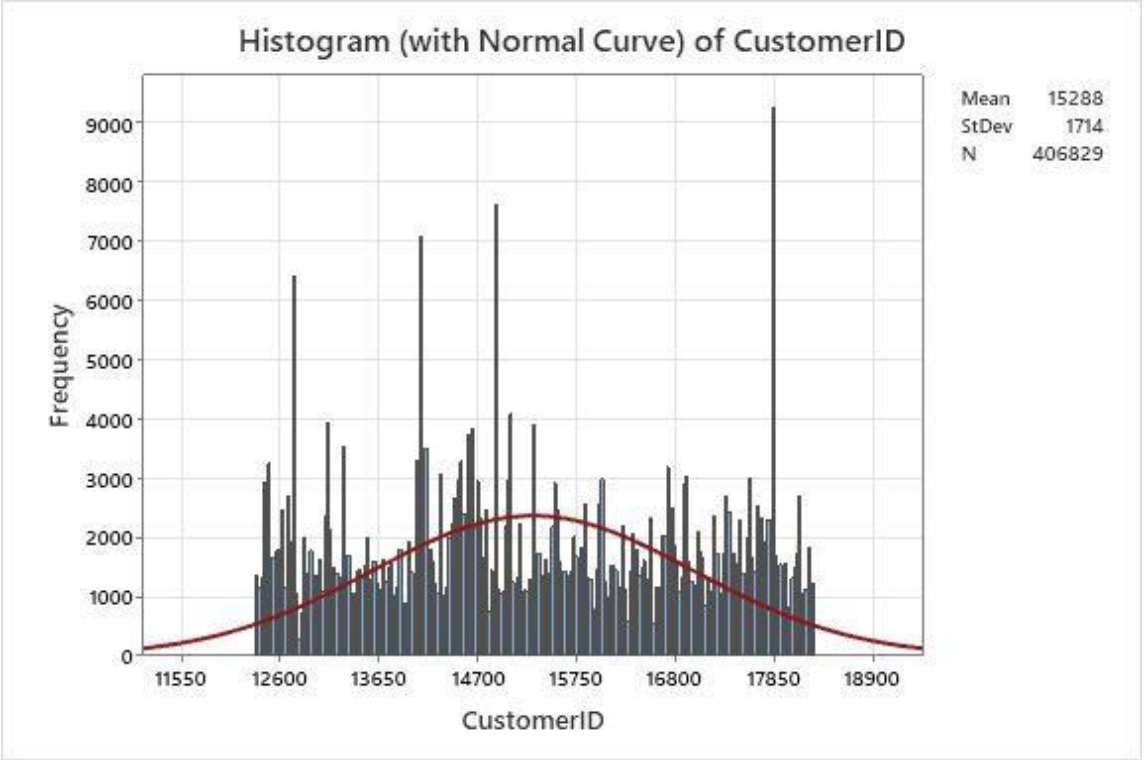
Variable	Median	Q3	Range	Mode	N for Mode	Skewness	Kurtosis
InvoiceNo	0.000000	0.000000	0.000000	0	406829	*	*
Quantity	3.00	10.0	161990.0	1	148227	-0.26	119769.16
InvoiceDate	40744	40835	373	40847.6	1114	-0.34	-1.20
UnitPrice	2.08	4.13	50032.1	1.25	50496	186.51	59005.72
CustomerID	15152	16791	5941	17841	7983	0.03	-1.18


Histogram (with Normal Curve) of InvoiceNo
Mean *
StDev *
N 406829
Frequency
400000
300000
200000
100000
0
0
InvoiceNo


Histogram (with Normal Curve) of Quantity
Mean 9.552
StDev 218.1
N 541909
Frequency
1000000
800000
600000
400000
200000
0
-66000
-44000
-22000
0
22000
44000
66000
Quantity


Histogram (with Normal Curve) of InvoiceDate
Mean 40729
StDev 115.9
N 541909
Frequency
8000
7000
6000
5000
4000
3000
2000
1000
0
09-10-2010 00:00
18-12-2010 00:00
26-02-2011 00:00
07-05-2011 00:00
16-07-2011 00:00
24-09-2011 00:00
03-12-2011 00:00
11-02-2012 00:00
InvoiceDate


Histogram (with Normal Curve) of UnitPrice
Mean 4.611
StDev 96.76
N 541909
Frequency
600000
500000
400000
300000
200000
100000
0
-6600
0
6600
13200
19800
26400
33000
39600
UnitPrice


Histogram (with Normal Curve) of CustomerID
Mean 15288
StDev 1714
N 406829
Frequency
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
11550
12600
13650
14700
15750
16800
17850
18900
CustomerID

5. Regression Analysis

a) Simple Linear Regression:

The predictor chosen is open.
Go to stat > Regression > Regression > Fit Regression Model. In responses, give the column name. In continuous predictors, give a predictor column name. In the graphs option, select four in one graph and press ok and again press ok.

ONLINE RETAIL

Regression Analysis: Quantity versus UnitPrice, Country

Method

Categorical predictor coding (1, 0)

Regression Equation

Country			
Australia	Quantity	=	66.45 - 0.00279 UnitPrice
Austria	Quantity	=	12.0 - 0.00279 UnitPrice
Bahrain	Quantity	=	13.7 - 0.00279 UnitPrice
Belgium	Quantity	=	11.20 - 0.00279 UnitPrice
Brazil	Quantity	=	11.1 - 0.00279 UnitPrice
Canada	Quantity	=	18.3 - 0.00279 UnitPrice
Channel Islands	Quantity	=	12.52 - 0.00279 UnitPrice

ONLINE RETAIL

Regression Analysis: Quantity versus UnitPrice, Country

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	66.45	6.14	10.82	0.000	
UnitPrice	-0.00279	0.00306	-0.91	0.363	1.00
Country					
Austria	-54.4	12.5	-4.35	0.000	1.32
Bahrain	-52.8	50.4	-1.05	0.295	1.02
Belgium	-55.25	7.79	-7.09	0.000	2.63
Brazil	-55.3	39.0	-1.42	0.156	1.03
Canada	-48.1	18.8	-2.56	0.010	1.12
Channel Islands	-53.9	10.0	-5.38	0.000	1.60
Cyprus	-56.3	10.7	-5.27	0.000	1.49
Czech Republic	-46.7	40.3	-1.16	0.246	1.02
Denmark	-45.4	12.6	-3.59	0.000	1.31
EIRE	-49.03	6.60	-7.43	0.000	7.40
European Community	-58.3	28.6	-2.04	0.041	1.05
Finland	-51.1	10.3	-4.96	0.000	1.55
France	-53.53	6.58	-8.13	0.000	7.67
Germany	-54.07	6.54	-8.27	0.000	8.39
Greece	-55.8	19.1	-2.93	0.003	1.12
Hong Kong	-49.8	14.2	-3.50	0.000	1.23
Iceland	-52.9	17.3	-3.06	0.002	1.14

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
217.994	0.09%	0.08%	0.09%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	38	22287170	586504	12.34	0.000
UnitPrice	1	39394	39394	0.83	0.363
Country	37	22247865	601294	12.65	0.000
Error	541870	25750527510	47522		
Lack-of-Fit	3332	231429921	69457	1.47	0.000
Pure Error	538538	25519097589	47386		
Total	541908	25772814680			

Regression Analysis: Quantity versus UnitPrice, Country

Fits and Diagnostics for Unusual Observations

Obs	Quantity	Fit	Resid	Std Resid	
198	6.0	66.4	-60.4	-0.28	X
199	8.0	66.4	-58.4	-0.27	X
200	12.0	66.4	-54.4	-0.25	X
201	6.0	66.4	-60.4	-0.28	X
202	4.0	66.4	-62.4	-0.29	X
203	6.0	66.4	-60.4	-0.28	X
204	3.0	66.4	-63.4	-0.29	X
205	2.0	66.4	-64.4	-0.30	X
206	4.0	66.4	-62.4	-0.29	X
207	4.0	66.4	-62.4	-0.29	X
208	2.0	66.4	-64.4	-0.30	X
209	2.0	66.4	-64.4	-0.30	X
210	24.0	66.4	-42.4	-0.19	X
211	24.0	66.5	-42.5	-0.19	X
386	96.0	84.4	11.6	0.05	X
387	1.0	84.4	-83.4	-0.38	X
731	600.0	8.6	591.4	2.71	R
871	480.0	8.6	471.4	2.16	R
1237	50.0	17.7	32.3	0.15	X
1238	96.0	17.7	78.3	0.36	X
1239	8.0	17.7	-9.7	-0.04	X

Regression Analysis: Quantity versus UnitPrice, Country

Fits and Diagnostics for Unusual Observations

Obs	Quantity	Fit	Resid	Std Resid	
198	6.0	66.4	-60.4	-0.28	X
199	8.0	66.4	-58.4	-0.27	X
200	12.0	66.4	-54.4	-0.25	X
201	6.0	66.4	-60.4	-0.28	X
202	4.0	66.4	-62.4	-0.29	X
203	6.0	66.4	-60.4	-0.28	X
204	3.0	66.4	-63.4	-0.29	X
205	2.0	66.4	-64.4	-0.30	X
206	4.0	66.4	-62.4	-0.29	X
207	4.0	66.4	-62.4	-0.29	X
208	2.0	66.4	-64.4	-0.30	X
209	2.0	66.4	-64.4	-0.30	X
210	24.0	66.4	-42.4	-0.19	X
211	24.0	66.5	-42.5	-0.19	X
386	96.0	84.4	11.6	0.05	X
387	1.0	84.4	-83.4	-0.38	X
731	600.0	8.6	591.4	2.71	R
871	480.0	8.6	471.4	2.16	R
1237	50.0	17.7	32.3	0.15	X
1238	96.0	17.7	78.3	0.36	X
1239	8.0	17.7	-9.7	-0.04	X

The predictor chosen is Weight.

<u>Regression Equation</u>

b) Perform simple MonteCarlo Simulation for ttest

Perform simple MonteCarlo Simulation for ttest: For MonteCarlo Simulation in minitab. Go to Calc > Random Data > t. Select Number of rows of data to be generated, provide the column name in store in column(s) option and provide the degrees of freedom.

	C1	C2-T	C3-T	C4	C5-D	C6	C7	C8-T
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	1.950	01-12-2010 08:26	2.5	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	-0.736	01-12-2010 08:26	3.4	17850	United Kingdom
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	-1.464	01-12-2010 08:26	2.8	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	-1.433	01-12-2010 08:26	3.4	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	1.969	01-12-2010 08:26	3.4	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	1.241	01-12-2010 08:26	7.7	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	-0.807	01-12-2010 08:26	4.3	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	-148.156	01-12-2010 08:28	1.9	17850	United Kingdom
9	536366	22632	HAND WARMER RED POLKA DOT	0.284	01-12-2010 08:28	1.9	17850	United Kingdom
10	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	-0.214	01-12-2010 08:34	1.7	13047	United Kingdom

Online Retail | Online Retail(W1) | Online Retail(W2)

The Monte Carlo Simulation values are generated and stored in tradecount.

Code:

```
library (brms)
library(rstan)
data<-read.csv("C:\\Users\\Dhinesh\\Downloads\\online+retail\\onlinrretail-red.csv")
head (data)
fit<-brm(sulphates~density+pH,data = data,family = gaussian())
summary(fit)
pred<-data.frame(density=1,pH=3.5)
predict (fit, pred)

#bayesfactor
library(BayesFactor)
g1=data$density[data$pH==3.51]
g2=data$density[data$pH==3.2]
res<-ttestBF(x=g1,y=g2)
summary(res)


library(MonteCarlo)
set.seed (9)
ttest<-function (n, loc, scale) {
  sample<-rnorm(n,loc, scale)
  stat<-sqrt(n)*mean(sample)/sd(sample)
  decision<-abs(stat) >1.96
  return(list("decision"=decision))
}
n_grid<-c(50,100,250,500)
loc_grid<-seq(0,1,0.2)
```

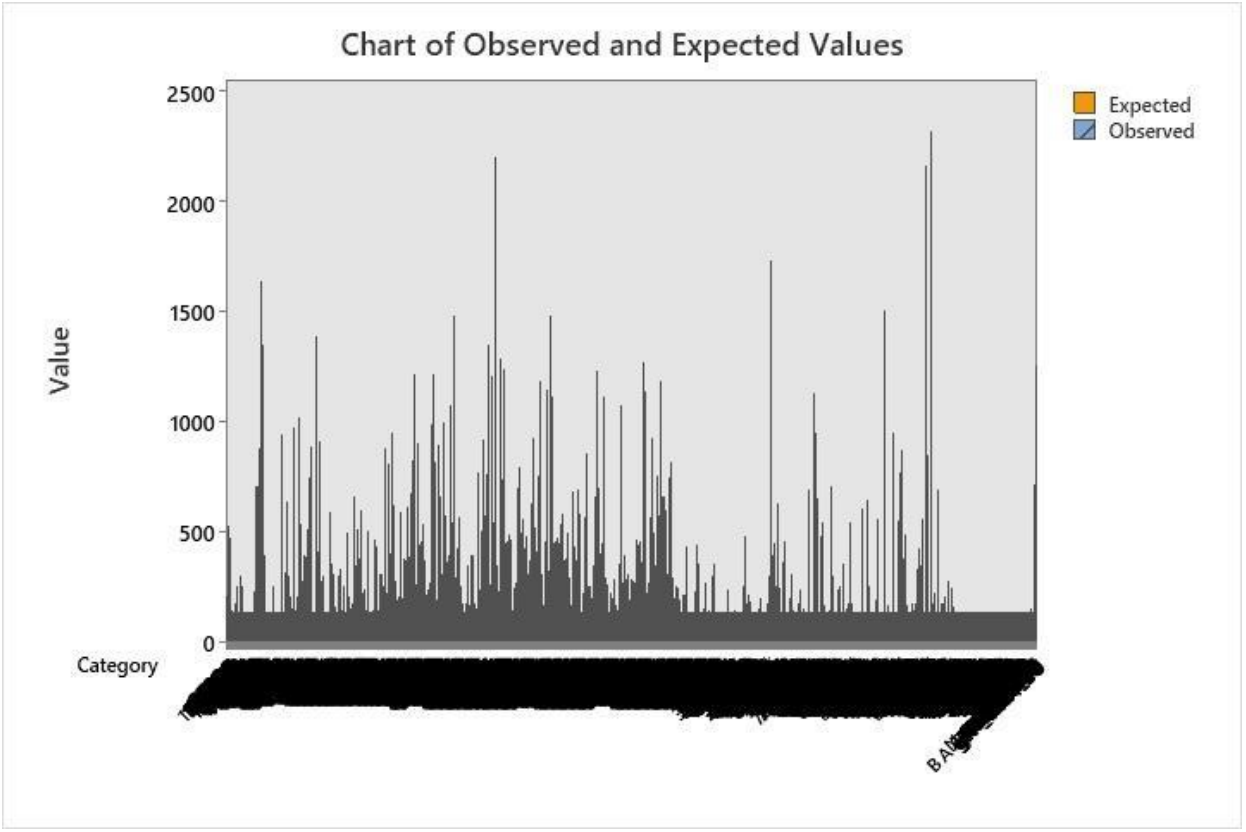
```
scale_grid<-c(1,2)
param_list<-list("n"=n_grid,"loc"=loc_grid,"scale"=scale_grid)
res<-MonteCarlo(func=ttest,nrep=1000,param_list =param_list)
summary(res)
rows<-c ("n")
cols<-c("loc", "scale")
MakeTable(output = res,rows=rows,cols=cols,digit=2)
```

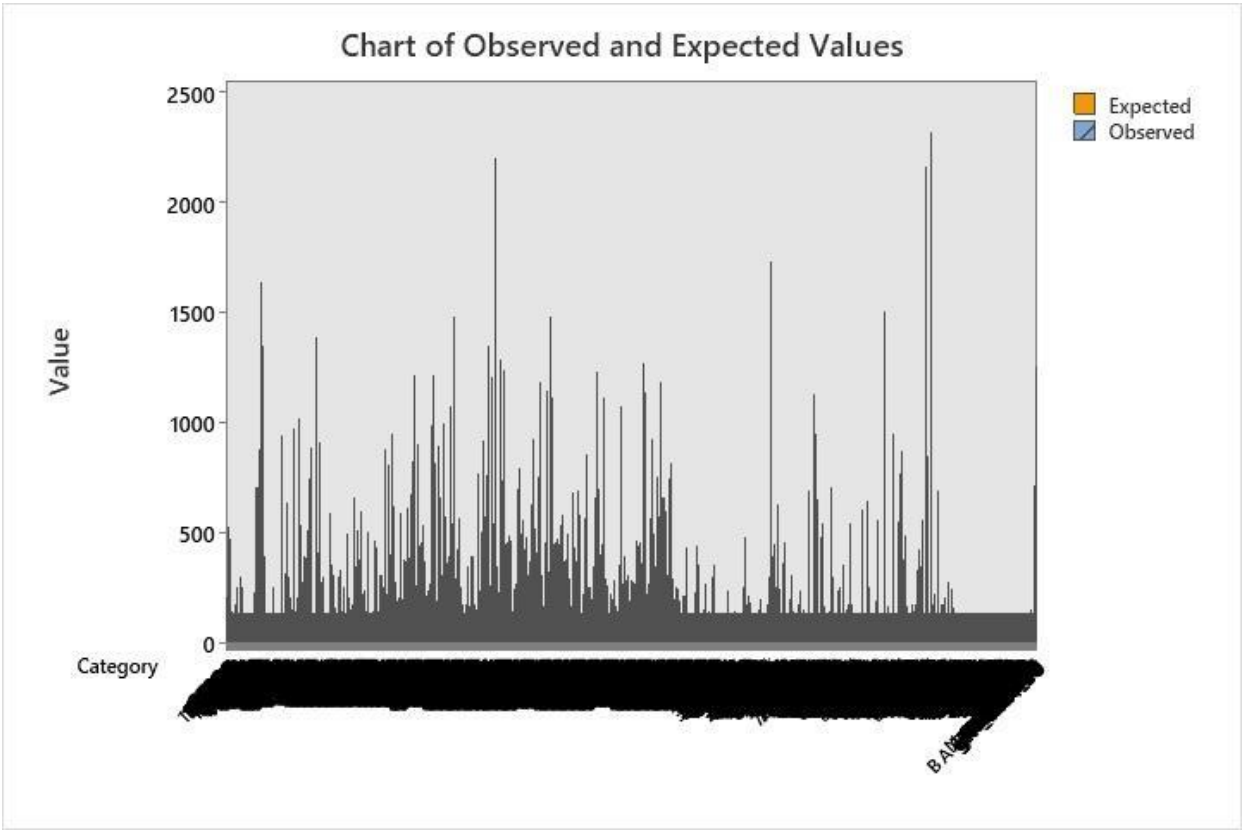
6. Chi-square Test

a) Goodness-of-fit:

Chi-Square Test

N	N*	DF	Chi-Sq	P-Value
541909	0	4069	1221225	0.000





The p-value from the output is 0.000

P < 0.05 therefore we reject the null hypothesis.

The null hypothesis is that the data is in equal proportions.
The Alternative hypothesis is that data is not in equal proportions.
Based on the p-value, the value is less than 0.05. So, the null hypothesis is rejected and the alternative hypothesis is accepted.
Therefore, the Category_High values 1 and 2 are in unequal proportions.

b) Test of Association:

Let us consider that we need to check whether the column category_high and category_low.

Go to Stat > Tables > Chi-Square Test for Association. Select Raw data(categorical variables). Provide the column names in rows and columns. In the Statistics option, select each cell's contribution to chi-square and press ok and press ok.

ONLINE RETAIL

Chi-Square Test for Association: Quantity, Country

Rows: Quantity Columns: Country

	Australia	Austria	Bahrain	Belgium	Brazil	Canada	Channel Islands
-80995	0	0	0	0	0	0	0
	0.002323	0.000740	0.000035	0.003818	0.000059	0.000279	0.001399
	0.002323	0.000740	0.000035	0.003818	0.000059	0.000279	0.001399
-74215	0	0	0	0	0	0	0
	0.002323	0.000740	0.000035	0.003818	0.000059	0.000279	0.001399
	0.002323	0.000740	0.000035	0.003818	0.000059	0.000279	0.001399
-9600	0	0	0	0	0	0	0
	0.004647	0.001480	0.000070	0.007636	0.000118	0.000557	0.002798
	0.004647	0.001480	0.000070	0.007636	0.000118	0.000557	0.002798
-9360	0	0	0	0	0	0	0
	0.002323	0.000740	0.000035	0.003818	0.000059	0.000279	0.001399
	0.002323	0.000740	0.000035	0.003818	0.000059	0.000279	0.001399
-9058	0	0	0	0	0	0	0
	0.002323	0.000740	0.000035	0.003818	0.000059	0.000279	0.001399

Chi-Square Test for Association: Quantity, Country

	0.001148	0.000055	0.000718	0.015124	0.000113	0.001283
80995	0	0	0	0	0	0
	0.001148	0.000055	0.000718	0.015124	0.000113	0.001283
	0.001148	0.000055	0.000718	0.015124	0.000113	0.001283
All	622	30	389	8196	61	695

Cell Contents
Count
Expected count
Contribution to Chi-square
The entire table cannot be displayed.

Chi-Square Test

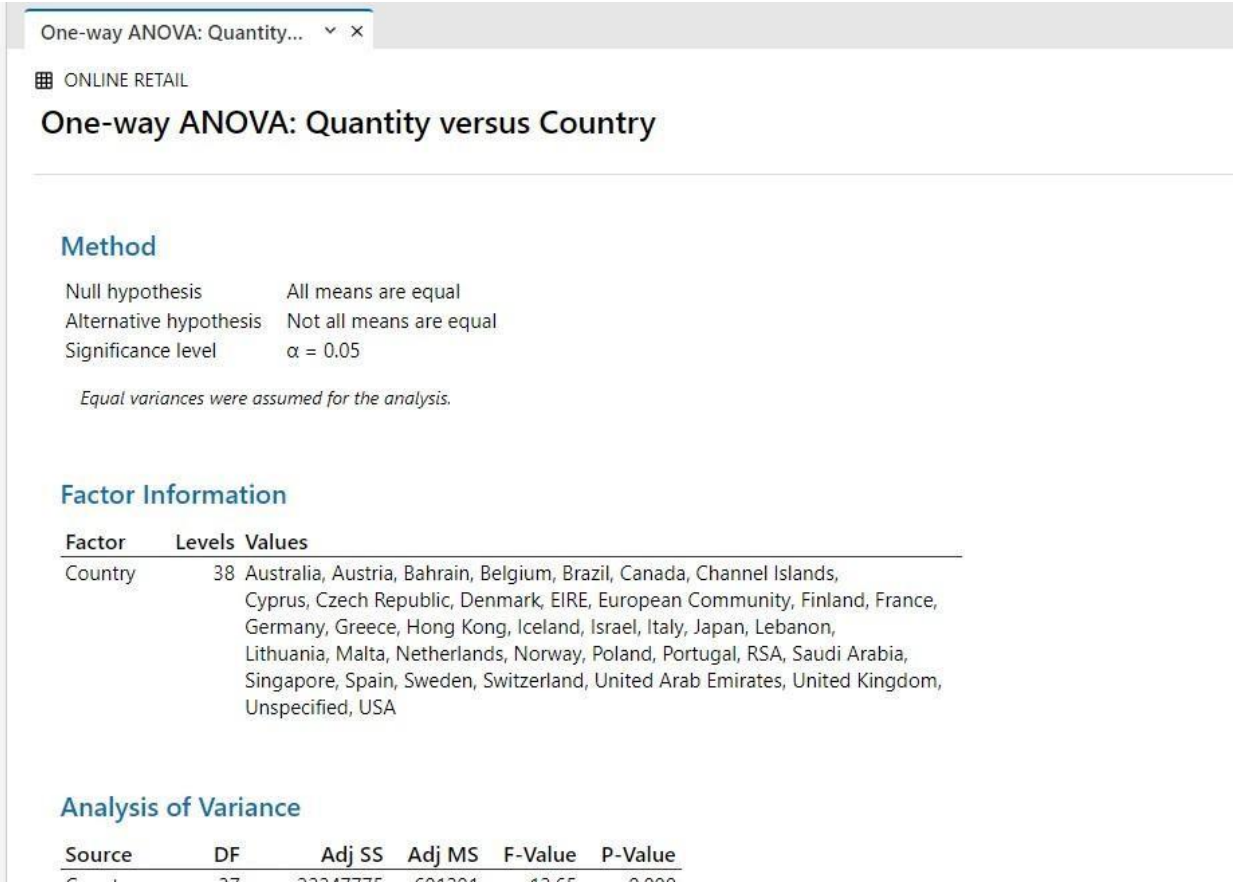
	Chi-Square	DF
Pearson	126303.436	26677
Likelihood Ratio	51528.084	26677

26043 cell(s) with expected counts less than 1.
Chi-Square approximation probably invalid.
26733 cell(s) with expected counts less than 5.

The null hypothesis is that the columns are independent.
The Alternative hypothesis is that columns are dependent.
Based on the p-value, the value is less than 0.05. So, the null hypothesis is rejected and the alternative hypothesis is accepted.
Therefore, the columns Category_High and Category_Low are dependent.

7. ANOVA

Let us consider the problem to check whether two categories of Category_High have equal means of tradecount.
Go to Stat > ANOVA > One-Way. Select Response data are in one column for all factor levels. Provide tradecount as the response variable and Category_High as the factor and press ok.



One-way ANOVA: Quantity...

ONLINE RETAIL

One-way ANOVA: Quantity versus Country

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0.05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
Country	38	Australia, Austria, Bahrain, Belgium, Brazil, Canada, Channel Islands, Cyprus, Czech Republic, Denmark, EIRE, European Community, Finland, France, Germany, Greece, Hong Kong, Iceland, Israel, Italy, Japan, Lebanon, Lithuania, Malta, Netherlands, Norway, Poland, Portugal, RSA, Saudi Arabia, Singapore, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom, Unspecified, USA

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value

One-way ANOVA: Quantity versus Country

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Country	37	22247775	601291	12.65	0.000
Error	541871	25750566904	47522		
Total	541908	25772814679			

Model Summary

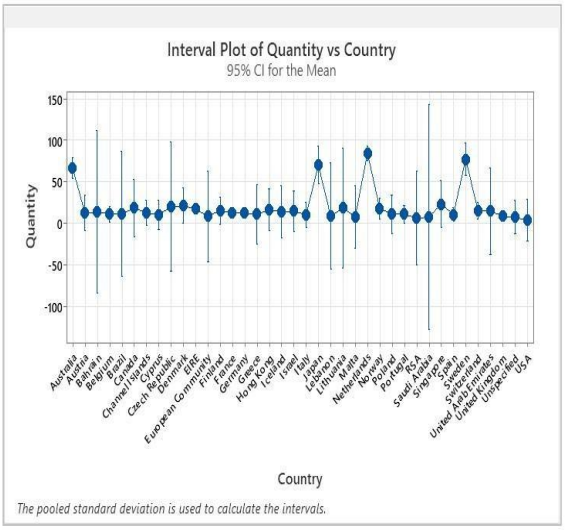
S	R-sq	R-sq(adj)	R-sq(pred)
217.994	0.09%	0.08%	0.09%

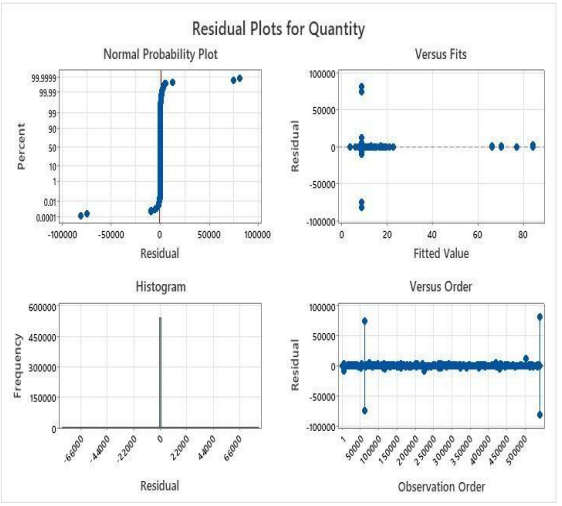
Means

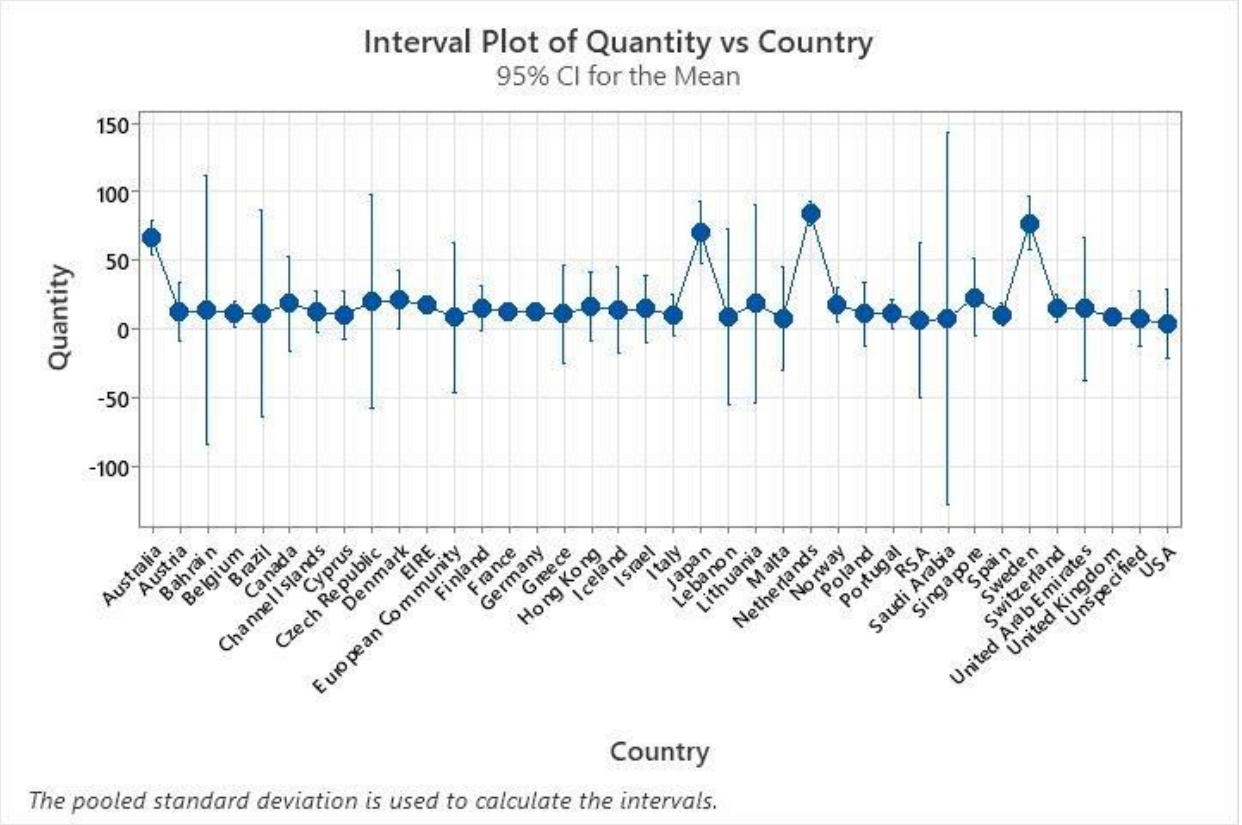
Country	N	Mean	StDev	95% CI
Australia	1259	66.44	97.69	(54.40, 78.49)
Austria	401	12.04	21.75	(-9.30, 33.37)
Bahrain	19	13.68	30.02	(-84.34, 111.70)
Belgium	2069	11.190	13.601	(1.797, 20.583)
Brazil	32	11.13	8.48	(-64.41, 86.66)
Canada	151	18.30	46.68	(-16.47, 53.07)

ONLINE RETAIL

One-way ANOVA: Quantity versus Country



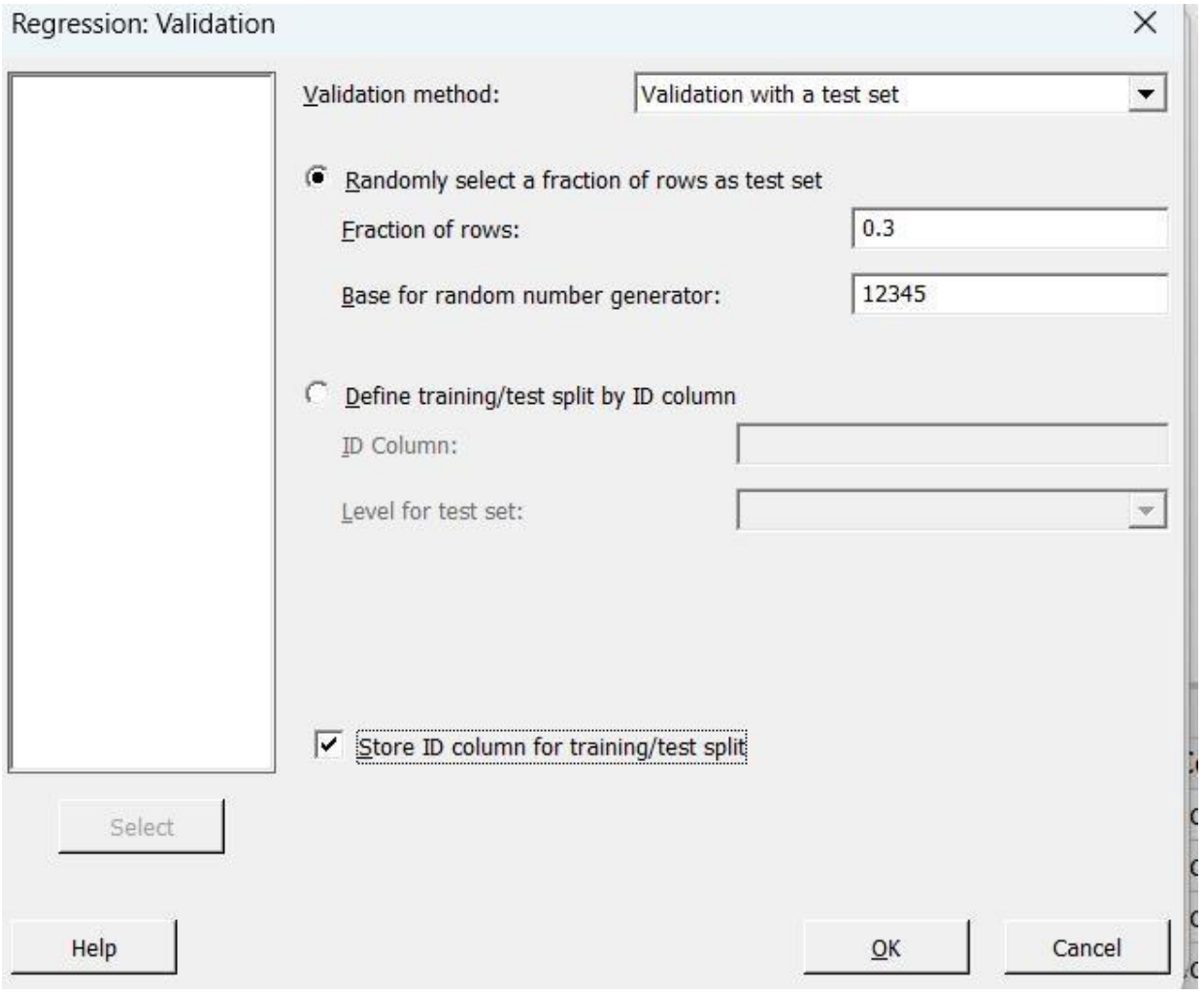




The null hypothesis is that the groups have equal means.
The Alternative hypothesis is that the groups have unequal means.
Based on the p-value, the value is greater than 0.05. So, the null hypothesis is accepted and the alternative hypothesis is rejected.
Therefore, the columns Category_High have categories with equal means of tradecount.

8. Model Validation, Diagnostic, and Prediction

Go to stat > Regression > Regression > Fit Regression Model.
In responses, give the column name. In continuous predictors, give a predictor column name. In the graphs option, select four in one graph and press ok. In the validation option, select the validation with a test set and press ok and again press ok.
The size of the training and testing sets is 70%*50 and 30%*50 respectively as the fraction provided was 0.3 which means 30% of the data will be taken into testing.



ONLINE RETAIL(W3)

Regression Analysis: Quantity versus UnitPrice, Country

Method

Categorical predictor coding	(1, 0)
Test set fraction	30.0%

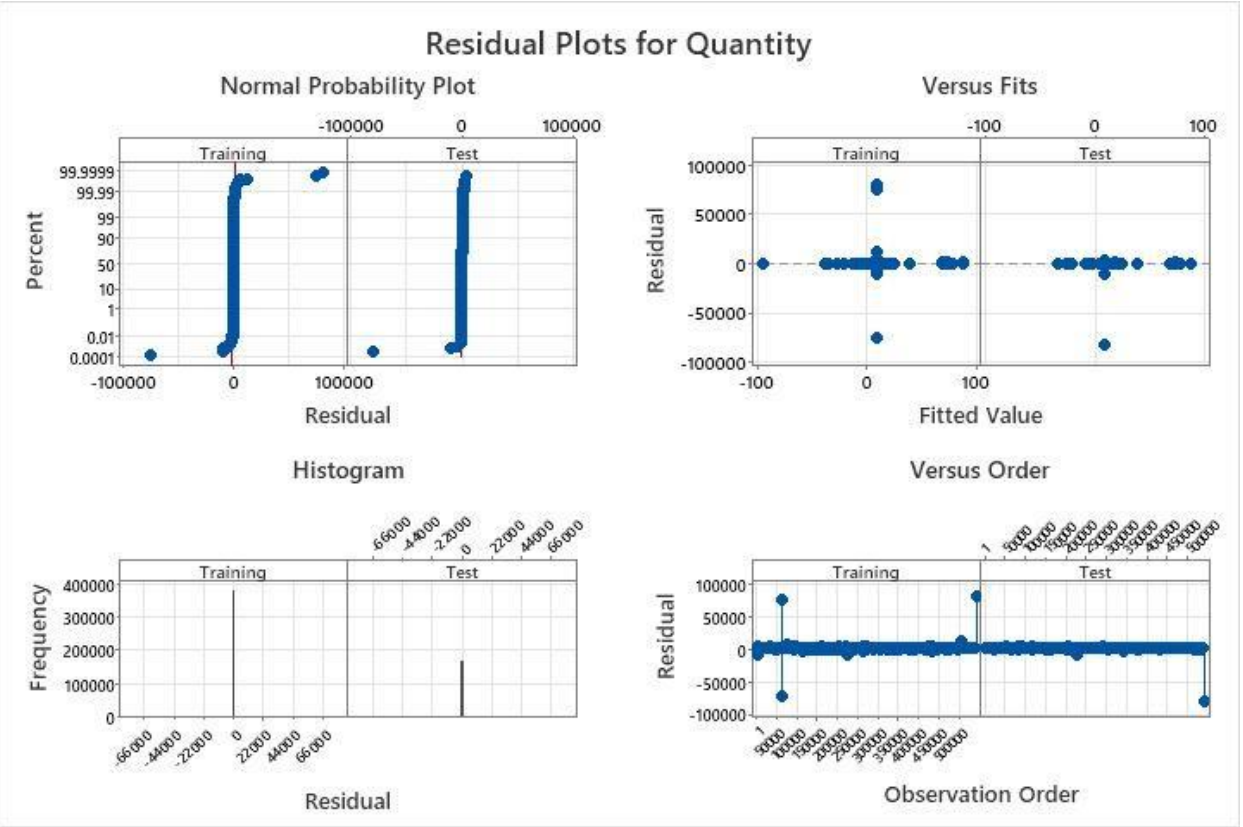
Regression Equation

Country			
Australia	Quantity	=	67.66 - 0.00263 UnitPrice
Austria	Quantity	=	12.2 - 0.00263 UnitPrice
Bahrain	Quantity	=	20.6 - 0.00263 UnitPrice
Belgium	Quantity	=	11.38 - 0.00263 UnitPrice

Fits and Diagnostics for Unusual Observations

Training Set

Obs	Quantity	Fit	Resid	Std Resid	
198	6.0	67.6	-61.6	-0.28	X
199	8.0	67.6	-59.6	-0.27	X
200	12.0	67.7	-55.7	-0.25	X
201	6.0	67.6	-61.6	-0.28	X
202	4.0	67.6	-63.6	-0.29	X
204	3.0	67.6	-64.6	-0.29	X
205	2.0	67.6	-65.6	-0.29	X
206	4.0	67.6	-63.6	-0.29	X
207	4.0	67.6	-63.6	-0.29	X
209	2.0	67.6	-65.6	-0.29	X
210	24.0	67.7	-43.7	-0.20	X
211	24.0	67.7	-43.7	-0.20	X



Based on the R-square value and Test R-square value, it can be decided that the model is an underfitting model but the difference is not much high so the model can be considered as a good model. Based on the four in one chart, the model can be considered as a statistically significant model.

➔ prediction

ONLINE RETAIL(W3)

Prediction for Quantity

Regression Equation

Quantity = 67.66 - 0.00263 UnitPrice + 0.000000 Country_Australia - 55.5 Country_Austria
- 47.1 Country_Bahrain - 56.28 Country_Belgium - 56.6 Country_Brazil
- 47.0 Country_Canada - 55.1 Country_Channel Islands - 58.0 Country_Cyprus
- 49.7 Country_Czech Republic - 46.2 Country_Denmark - 50.05 Country_EIRE
- 61.7 Country_European Community - 51.7 Country_Finland - 54.65 Country_France
- 55.32 Country_Germany - 56.8 Country_Greece - 51.3 Country_Hong Kong
- 53.3 Country_Iceland - 52.4 Country_Israel - 58.1 Country_Italy
+ 4.7 Country_Japan - 58.5 Country_Lebanon - 49.7 Country_Lithuania
- 60.0 Country_Malta + 19.09 Country_Netherlands - 50.4 Country_Norway
- 57.5 Country_Poland - 57.0 Country_Portugal - 61.3 Country_RSA
- 61.2 Country_Saudi Arabia - 43.2 Country_Singapore - 57.14 Country_Spain
+ 9.8 Country_Sweden - 52.69 Country_Switzerland - 53.7 Country_United Arab Emirates - 58.81 Country_United Kingdom - 60.1 Country_Unspecified
- 64.1 Country_USA

Prediction for Quantity

- 64.1 Country_USA

Settings

Variable	Setting
UnitPrice	2.5
Country	Australia

Prediction

Fit	SE Fit	95% CI	95% PI	
67.6529	7.47518	(53.0018, 82.3040)	(-368.941, 504.246)	XX

XX denotes an extremely unusual point relative to predictor levels used to fit the model.

8. Conclusion

Based on the analysis performed using Minitab for the online retail dataset, the following conclusions can be drawn:

1. Data Preparation: Missing values were identified and filled appropriately based on the data type, and outliers were detected using outlier tests.
2. Descriptive Statistics: Descriptive statistics, including mean, median, mode, range, variance, standard deviation, skewness, and kurtosis, were calculated to understand the data distribution.
3. Regression Analysis: Simple linear regression models were constructed, and Monte Carlo simulations were performed for t-tests. The results indicated that the model is statistically significant.
4. Chi-square Test: The goodness-of-fit chi-square test revealed that the data is not in equal proportions, and the test of association indicated a dependency between certain categorical variables.
5. ANOVA: The ANOVA test showed that different categories within the "Category_High" variable have equal means of "tradecount."
6. Model Validation: Model validation was conducted, and the model was found to be slightly underfitting but still statistically significant.

In conclusion, the analysis provides valuable insights into the online retail dataset. It identifies key relationships, dependencies, and statistical significance within the data. These findings can be used to make data-driven decisions and predictions, contributing to a better understanding of the online retail business.