MIS-64036: Business Analytics
Assignment I

Total Marks: 100
Contribution to the Final Mark: 20%
Instructions: Please answer all questions. You should use R to solve the questions and include the screen shots in your submission. The Golden questions are optional and carries additional marks. This means that you will not lose marks if you do not answer that question. Please use the link provided on the Blackboard, under the assessment section, to upload your submissions. Late submissions, up to two days, are subject to 30% penalty. Submissions made more than two days after the deadline will not be graded.

*****

Part A) Descriptive Statistics & Normal Distributions

1. a) What is the probability of obtaining a score greater than 700 on a GMAT test that has a mean of 494 and a standard deviation of 100? Assume GMAT scores are normally distributed (5 marks).

Ans. pnorm(700,494,100,lower.tail = FALSE)
0.01969927

b) What is the probability of getting a score between 350 and 450 on the same GMAT exam? (5 marks)

Ans. (pnorm(450,494,100,lower.tail = TRUE) - pnorm(350,494,100,lower.tail = TRUE)) *100
25.50349

2. Runzheimer International publishes business travel costs for various cities throughout the world. In particular, they publish per diem totals, which represent the average costs for the typical business traveler including three meals a day in business-class restaurants and single-rate lodging in business-class hotels and motels. If 86.65% of the per diem costs in Buenos Aires, Argentina, are less than $449 and if the standard deviation of per diem costs is $36, what is the average per diem cost in Buenos Aires? Assume that per diem costs are normally distributed (10 marks)

Ans. qnorm(0.8665,446,36,lower.tail = TRUE)
485.9599

3. Chris is interested in understanding the correlation between temperature in Kent, OH and Los Angeles, CA. He has got the following data for September 2017 from Alpha Knowledgebase. (5 marks)

He has sampled the mid-day temperature for days from Sep 2 to Sep 6 as follows:

Kent=c(59, 68, 78, 60)
Los_Angeles=c(90, 82, 78, 75)

Calculate the correlation (Pearson Correlation Coefficient) between the temperatures of the two cities without using any R commands i.e. calculate step by step.

Ans:
```
kent <- c(59, 68, 78, 60)
los_angeles <- c(90, 82, 78, 75)
N <- sum((kent-mean(kent)) * (los_angeles-mean(los_angeles)) )
N
D <- sqrt(sum(((kent-mean(kent))^2) * sum((los_angeles-mean(los_angeles))^2)))
D
N / D
```
-0.3566049

Part B) Data Wrangling

For the questions in this part, you need to use the 'Online Retail' dataset which can be downloaded in CSV format from the course portal under the assignment folder. This is a transnational data set which contains all the transactions occurring between 01 Dec 2010 and 09 Dec 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. The data contains the following attributes:

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

Download the dataset, and use the read.csv() command to load the file into a R dataframe and answer the following questions.

4. Show the breakdown of the number of transactions by countries i.e. how many transactions are in the dataset for each country (consider all records including cancelled transactions). Show this in total number and also in percentage. Show only countries accounting for more than 1% of the total transactions. (5 marks)

Ans :

```
ort <- read.csv("Online_Retail(2).csv") # Reading the CSV file
> library(plyr) # Calling and using library for Splitting, Applying and Combining Data
> count1 <- count(ort,"Country") # COunting the Country
> per <-prop.table(table(ort$Country))*100 #C calculating Percentage
> rdper <- round(per,digits = 3) # Rounding to 3 Digits
> rdpercent <- paste0(rdper,sep = "%") # Adding % Symbol
> df1<-as.data.frame(per) # Converting
> tbl1 <- cbind(count1,df1[2],rdpercent) # Binding all values in proper Table Format
> names(tbl1)<-c("Country","Count","Percentage","Round %") # Assigning Names
> tbl1 # Final table with required
```

	Country	Count	Percentage	Round %
1	Australia	1259	0.23232683	0.23%
2	Austria	401	0.073997664	0.07%
3	Bahrain	19	0.003506124	0.00%
4	Belgium	2069	0.38179842	0.38%
5	Brazil	32	0.00590505	0.01%
6	Canada	151	0.027864457	0.03%
7	Channel Islands	758	0.139875883	0.14%
8	Cyprus	622	0.114779419	0.12%
9	Czech Republic	30	0.005535985	0.01%
10	Denmark	389	0.07178327	0.07%
11	EIRE	8196	1.512431054	1.51%
12	European Community	61	0.011256502	0.01%
13	Finland	695	0.128250315	0.13%
14	France	8557	1.579047405	1.58%
15	Germany	9495	1.752139197	1.75%
16	Greece	146	0.026941793	0.03%
17	Hong Kong	288	0.053145454	0.05%
18	Iceland	182	0.033584975	0.03%
19	Israel	297	0.05480625	0.06%
20	Italy	803	0.14817986	0.15%
21	Japan	358	0.066062752	0.07%
22	Lebanon	45	0.008303977	0.01%
23	Lithuania	35	0.006458649	0.01%
24	Malta	127	0.023435669	0.02%
25	Netherlands	2371	0.437527334	0.44%
26	Norway	1086	0.200402651	0.20%
27	Poland	341	0.062925694	0.06%
28	Portugal	1519	0.280305365	0.28%
29	RSA	58	0.010702904	0.01%
30	Saudi Arabia	10	0.001845328	0.00%
31	Singapore	229	0.042258017	0.04%
32	Spain	2533	0.467421652	0.47%
33	Sweden	462	0.085254166	0.09%
34	Switzerland	2002	0.369434721	0.37%
35	United Arab Emirates	68	0.012548232	0.01%
36	United Kingdom	495478	91.43195629	91.43%
37	Unspecified	446	0.082301641	0.08%
38	USA	291	0.053699053	0.05%

per[per>1] # countries accounting for more than 1%

EIRE	France	Germany	United Kingdom
1.512431	**1.579047**	**1.752139**	**91.431956**

5. Create a new variable 'TransactionValue' that is the product of the exising 'Quantity' and 'UnitPrice' variables. Add this variable to the dataframe. (5 marks)

Ans. ort$TransactionValue <- ort$Quantity * ort$UnitPrice #New Variable with 'Quantity' and 'UnitPrice'
View(ort)

6. Using the newly created variable, TransactionValue, show the breakdown of transaction values by countries i.e. how much money in total has been spent each country. Show this in total sum of transaction values. Show only countries with total transaction exceeding 130,000 British Pound. (10 marks)

Ans. ctxn <- tapply(ort$TransactionValue,ort$Country,sum) # Summing values by combination
df2 <- as.data.frame(ctxn) # Converting
df2[,]
tbl2 <- cbind(count1,df1[2],rdpercent,df2[,]) # Binding table Format
names(tbl2)<-c("Country","Count","Percentage","Round %","TransactionValue") # Assigning Lables
tbl2 # Final table with required

Country	Count	Percentage	Round %	Transaction Value
Australia	1259	0.23232683	0.23%	137077.27
Austria	401	0.073997664	0.07%	10154.32
Bahrain	19	0.003506124	0.00%	548.4
Belgium	2069	0.38179842	0.38%	40910.96
Brazil	32	0.00590505	0.01%	1143.6
Canada	151	0.027864457	0.03%	3666.38
Channel Islands	758	0.139875883	0.14%	20086.29
Cyprus	622	0.114779419	0.12%	12946.29
Czech Republic	30	0.005535985	0.01%	707.72
Denmark	389	0.07178327	0.07%	18768.14
EIRE	8196	1.512431054	1.51%	263276.82
European Community	61	0.011256502	0.01%	1291.75
Finland	695	0.128250315	0.13%	22326.74
France	8557	1.579047405	1.58%	197403.9
Germany	9495	1.752139197	1.75%	221698.21
Greece	146	0.026941793	0.03%	4710.52
Hong Kong	288	0.053145454	0.05%	10117.04
Iceland	182	0.033584975	0.03%	4310
Israel	297	0.05480625	0.06%	7907.82
Italy	803	0.14817986	0.15%	16890.51
Japan	358	0.066062752	0.07%	35340.62
Lebanon	45	0.008303977	0.01%	1693.88
Lithuania	35	0.006458649	0.01%	1661.06

Malta	127	0.023435669	0.02%	2505.47
Netherlands	2371	0.437527334	0.44%	284661.54
Norway	1086	0.200402651	0.20%	35163.46
Poland	341	0.062925694	0.06%	7213.14
Portugal	1519	0.280305365	0.28%	29367.02
RSA	58	0.010702904	0.01%	1002.31
Saudi Arabia	10	0.001845328	0.00%	131.17
Singapore	229	0.042258017	0.04%	9120.39
Spain	2533	0.467421652	0.47%	54774.58
Sweden	462	0.085254166	0.09%	36595.91
Switzerland	2002	0.369434721	0.37%	56385.35
United Arab Emirates	68	0.012548232	0.01%	1902.28
United Kingdom	495478	91.43195629	91.43%	8187806.36
Unspecified	446	0.082301641	0.08%	4749.79
USA	291	0.053699053	0.05%	1730.92

ctxn[ctxn>130000] # Total transaction exceeding 130,000

Australia	**EIRE**	**France**	**Germany**	**Netherlands**
137077.3	**263276.8**	**197403.9**	**221698.2**	**284661.5**

United Kingdom

8187806.4

7. This is an optional question which carries additional marks (golden questions). In this question, we are dealing with the InvoiceDate variable. The variable is read as a categorical when you read data from the file. Now we need to explicitly instruct R to interpret this as a Date variable. "POSIXlt" and "POSIXct" are two powerful object classes in R to deal with date and time. Click here for more information. First let's convert 'InvoiceDate' into a POSIXlt object:
   Temp=strptime(Online_Retail$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')
   Check the variable using, head(Temp). Now, let's separate date, day of the week and hour components dataframe with names as New_Invoice_Date, Invoice_Day_Week and New_Invoice_Hour:
   Online_Retail$New_Invoice_Date <- as.Date(Temp)
   The Date objects have a lot of flexible functions. For example knowing two date values, the object allows you to know the difference between the two dates in terms of the number days. Try this:
   Online_Retail$New_Invoice_Date[20000]- Online_Retail$New_Invoice_Date[10]
   Also we can convert dates to days of the week. Let's define a new variable for that
   Online_Retail$Invoice_Day_Week= weekdays(Online_Retail$New_Invoice_Date)
   For the Hour, let's just take the hour (ignore the minute) and convert into a normal numerical value:
   Online_Retail$New_Invoice_Hour = as.numeric(format(Temp, "%H"))
   Finally, lets define the month as a separate numeric variable too:
   Online_Retail$New_Invoice_Month = as.numeric(format(Temp, "%m"))

Now answer the flowing questions.

a) Show the percentage of transactions (by numbers) by days of the week (extra 2 marks)

Ans.txnper<tapply(ort$TransactionValue,ort$Invoice_Day_Week,NROW)/NROW(ort$TransactionValue)*100 # percentage of transactions

txnper

Friday	Monday	Sunday	Thursday	Tuesday	Wednesday
15.16731	17.55110	11.87930	19.16503	18.78692	17.45035

b) Show the percentage of transactions (by transaction volume) by days of the week (extra 1 marks)

Friday	Monday	Sunday	Thursday	Tuesday	Wednesday
15.804787	16.297194	8.265282	21.671867	20.170636	17.790232

c) Show the percentage of transactions (by transaction volume) by month of the year (extra 1 marks)

Ans.

1	2	3	4	5	6
5.745	5.109515	7.009487	5.059703	7.420519	7.09008

7	8	9	10	11	12
6.989	7.003469	10.46075	10.98412	14.99584	12.1323

d) What was the date with the highest number of transactions from Australia? (3 marks)

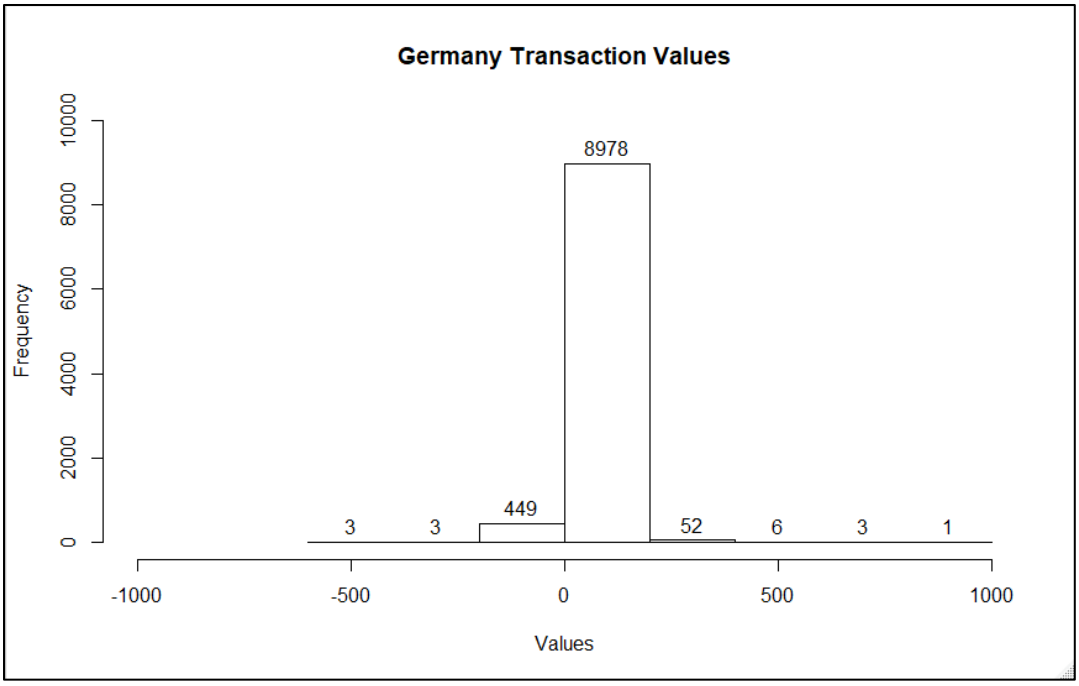
1718.4

e) The company needs to shut down the website for two consecutive hours for maintenance. What would be the hour of the day to start this so that the distribution is at minimum for the customers? The responsible IT team is available from 7:00 to 20:00 every day(3 marks)

Ans. **conse<-table(ort$New_Invoice_Hour)**
abs(diff(conse))

8. Plot the histogram of transaction values from Germany. Use the hist() function to plot. (5 marks)

Ans. cht <- hist(ort$TransactionValue[ort$Country=="Germany"],breaks=10,xlim = c(-1000,1000),ylim = c(0,10000)) # Creating Histogram
cht
text(cht$mids,cht$counts,labels = cht$counts,adj = c(0.5,-0.5)) # Displaying the Text



9. Which customer had the highest number of transactions? Which customer is most valuable (i.e. highest total sum of transactions)? (10 marks)

Ans. htxn<-tapply(ort$TransactionValue,ort$CustomerID,length) # Summing values by combination
which.max(htxn) # Finding Max value with index number of customer
htxn[4043] # 4043 Customer with highest txns as 17841
17841 # Txns
7983 # Customer
Valuable Customer ***
hvl<-tapply(ort$TransactionValue,ort$CustomerID,sum)
which.max(hvl)
hvl[1704]
14646 # Highest Txns
279489 # Valuable Customer

10. Calculate the percentage of missing values for each variable in the dataset (5 marks). Hint colMeans():

Ans. colMeans(is.na(ort))*100

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	24.92669

Country	TransactionValue
0.00000	0.00000

11. What are the number of transactions with missing CustomerID records by countries? (10 marks)

Ans.

```
fn1<-function(x){
k<-sum(is.na(x))
return(k)}
tapply(retail$CustomerID,ort$Country,fn1)
```

Australia	Austria	Bahrain	Belgium	Brazil	
	0	0	2	0	0
Canada	Channel Islands	Cyprus	Czech Republic	Denmark	
0	0	0	0	0	
EIRE	European Community	Finland	France	Germany	
711	0	0	66	0	
Greece	Hong Kong	Iceland	Israel	Italy	
0	288	0	47	0	
Japan	Lebanon	Lithuania	Malta	Netherlands	
0	0	0	0	0	
Norway	Poland	Portugal	RSA	Saudi Arabia	
0	0	39	0	0	
Singapore	Spain	Sweden	Switzerland	United Arab Emirates	
0	0	0	125	0	
United Kingdom	Unspecified	USA			
133600	202	0			

12. On average, how often the costumers comeback to the website for their next shopping? (i.e. what is the average number of days between consecutive shopping) (Optional/Golden question: 18 additional marks!) Hint: 1. A close approximation is also acceptable and you may find diff() function useful.

Ans. Approx. 110 – 116 Days done my manual Excel calculations.

```
library(lubridate)#using Lubriate library
diff.Date()
diff.Date(date(ort$InvoiceDate), date(lag(ort$InvoiceDate, 1)))
DaysSinceLastPurchase = diff.Date(day("2011-12-10 00:00:00"), date(max(ort$InvoiceDate, na.rm = TRUE)))
TenureDays = diff.Date(day("2011-12-10 00:00:00"), date(min(ort$InvoiceDate, na.rm = TRUE)))
```

15. In the retail sector, it is very important to understand the return rate of the goods purchased by customers. In this example, we can define this quantity, simply, as the ratio of the number of transactions cancelled (regardless of the transaction value) over the total number of

transactions. With this definition, what is the return rate for the French customers? (10 marks). Consider the cancelled transactions as those where the 'Quantity' variable has a negative value.

Ans . NROW(ort$Quantity[ort$Quantity<0 & ort$Country=="France"])/NROW(ort)*100

0.02749539

15. What is the product that has generated the highest revenue for the retailer? (i.e. item with the highest total sum of 'TransactionValue')(10 marks)

Ans. Hrev<-tapply(ort$TransactionValue,ort$Description,sum) # Selecting Product with Highest Revenue
which.max(hrev) # Picking the item with value index
hrev[1140] # Dsiplaying the item with the highest total sum of TransactionValue

DOTCOM POSTAGE

206245.5

15. How many unique customers are represented in the dataset? You can use unique() and length() functions. (5 marks)

Ans. Length(unique(ort$CustomerID)) # Unique Customers

4373

****