MIS-64036: Business Analytics

Assignment I

Total Marks: 100
Contribution to the Final Mark: 20%
Submission deadline: 21 October
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).**

   Answer:

   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 is 0.0197

```
> pnorm(700, mean = 494, sd=100,lower.tail = FALSE)
[1] 0.01969927
```

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

   Answer:

   The Probability of getting a score between 350 and 450 on the same GMAT exam is 0.25

```
> pnorm(450, mean = 494, sd=100) - pnorm(350, mean = 494, sd=100)
[1] 0.2550349
```

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)**

Answer:

```
> z = qnorm(.8665)
> z
[1] 1.109998
> # zscore = (x- mean/sd)
> mean = 449 -(z * 36)
> mean
[1] 409.0401
```

The average per diem cost in Buenos Aires is $409.04

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.

Answer:

The pearson correlation coefficient can be calculated using the formula mentioned below.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is the number of samples
- x_i, y_i are the single samples indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

It can be solved as follows:

$$\bar{x} = \frac{59 + 68 + 78 + 60}{4} = 66.25$$

$$\bar{y} = \frac{90 + 82 + 78 + 75}{4} = 81.25$$

- $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) =$
  $(59 - 66.25)(90 - 81.25) + (68 - 66.25)(82 - 81.25) + (78 - 66.25)(78 - 81.25$
  $+ (60 - 66.25)(75 - 81.25) = -61.25$

- $\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})\text{^}2} =$
  $\sqrt{(59 - 66.25)^2 + (68 - 66.25)^2 + (78 - 66.25)^2 + (60 - 66.25)^2} = 15.26$

  $\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})\text{^}2} =$
  $\sqrt{(90 - 81.25)^2 + (82 - 81.25)^2 + (78 - 81.25)^2 + (75 - 81.25)^2} = 11.26$

  Therefore $r = \frac{-61.25}{15.26 * 11.26} = -0.3566$

  Pearson Correlation Coefficient) between the temperatures of the two cities is -0.356

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.

Note: I am considering number of distinct transactions while solving the questions in Part B of this assignment. For Example: One InvoiceNo can have multiple Items. I have read given Online_Retail.csv into mydata.

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)**

   Answer:

```
> library(dplyr)
> mydata <- read.csv("Online_Retail.csv")
> Result4 <- mydata %>% group_by(Country) %>% summarise(Total_Transactions=n_distinct(InvoiceNo))
%>% mutate(Percentage = Total_Transactions/sum(Total_Transactions)*100)
> newresult <- subset(Result4,Percentage > 1)
> newresult
# A tibble: 4 x 3
  Country        Total_Transactions Percentage
  <fct>                       <int>      <dbl>
1 EIRE                          360       1.39
2 France                        461       1.78
3 Germany                       603       2.33
4 United Kingdom              23494      90.7
```

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)**

   Answer:

```
> TransactionValue = mydata$UnitPrice * mydata$Quantity
> mydata <- mutate(mydata,TransactionValue)
> head(mydata)
  InvoiceNo StockCode                          Description Quantity    InvoiceDate UnitPrice CustomerID        Country
1    536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER        6 12/1/2010 8:26      2.55      17850 United Kingdom
2    536365     71053                 WHITE METAL LANTERN        6 12/1/2010 8:26      3.39      17850 United Kingdom
3    536365    84406B      CREAM CUPID HEARTS COAT HANGER        8 12/1/2010 8:26      2.75      17850 United Kingdom
4    536365    84029G KNITTED UNION FLAG HOT WATER BOTTLE        6 12/1/2010 8:26      3.39      17850 United Kingdom
5    536365    84029E      RED WOOLLY HOTTIE WHITE HEART.        6 12/1/2010 8:26      3.39      17850 United Kingdom
6    536365     22752        SET 7 BABUSHKA NESTING BOXES        2 12/1/2010 8:26      7.65      17850 United Kingdom
  TransactionValue
1            15.30
2            20.34
3            22.00
4            20.34
5            20.34
6            15.30
```

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)**

   Answer:

```
> Result6 <- summarise(group_by(mydata,Country), Sum_transaction = sum(TransactionValue))
> newresult1 <- subset(Result6, Sum_transaction > 130000)
> newresult1
# A tibble: 6 x 2
  Country        Sum_transaction
  <fct>                    <dbl>
1 Australia              137077.
2 EIRE                   263277.
3 France                 197404.
4 Germany                221698.
5 Netherlands            284662.
6 United Kingdom        8187806.
```

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.

Answer:

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

Answer:

```
> Result7A <- mydata %>% group_by(Invoice_Day_Week) %>% summarise(N_txn=n_distinct(InvoiceNo))
> Result7A <- mutate(Result7A, Percentage = N_txn/sum(N_txn)*100)
> Result7A
# A tibble: 6 x 3
  Invoice_Day_Week N_txn Percentage
  <chr>            <int>      <dbl>
1 Friday            4184      16.2
2 Monday            4138      16.0
3 Sunday            2381       9.19
4 Thursday          5660      21.9
5 Tuesday           4722      18.2
6 Wednesday         4815      18.6
```

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

Answer:

```
> Result7B <- mydata %>% group_by(Invoice_Day_Week) %>% summarise(Volume=sum(TransactionValue))
> Result7B <- mutate(Result7B, Percentage = Volume/sum(TransactionValue)*100)
> as.data.frame(Result7B)
  Invoice_Day_Week    Volume Percentage
1           Friday 1540610.8  15.804787
2           Monday 1588609.4  16.297194
3           Sunday  805678.9   8.265282
4         Thursday 2112519.0  21.671867
5          Tuesday 1966182.8  20.170636
6        Wednesday 1734147.0  17.790232
```

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

Answer:

```
> Result7C <- mydata %>% group_by(New_Invoice_Month) %>% summarise(Volume=sum(TransactionValue))
> Result7C <- mutate(Result7C, Percentage = Volume/sum(TransactionValue)*100)
> as.data.frame(Result7C)
   New_Invoice_Month    Volume Percentage
1                  1  560000.3   5.744919
2                  2  498062.6   5.109515
3                  3  683267.1   7.009487
4                  4  493207.1   5.059703
5                  5  723333.5   7.420519
6                  6  691123.1   7.090080
7                  7  681300.1   6.989308
8                  8  682680.5   7.003469
9                  9 1019687.6  10.460751
10                10 1070704.7  10.984123
11                11 1461756.2  14.995836
12                12 1182625.0  12.132290
```

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

Answer:
The date with the highest Number of transactions from Australia is 2011-02-07

```
> Result7D <- subset(mydata,Country =="Australia")
> Result7D <- summarise(group_by(Result7D,New_Invoice_Date),No_txn=n_distinct(InvoiceNo))
> Result7D <- arrange(Result7D,desc(No_txn))
> head(Result7D,1)
# A tibble: 1 x 2
  New_Invoice_Date No_txn
  <date>           <int>
1 2011-02-07           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)

Answer:

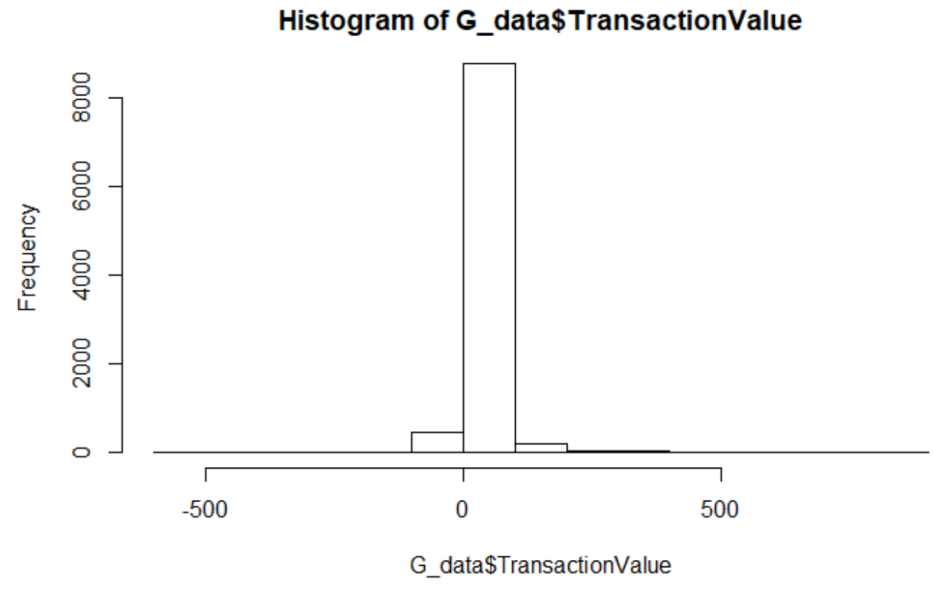
```
> Result7E <- summarise(group_by(mydata,New_Invoice_Hour),No_txn=n_distinct(InvoiceNo))
> Result7E <- filter(Result7E,New_Invoice_Hour >= 7 & New_Invoice_Hour <= 20)
> library(zoo)
> x <- rollapply(Result7E$No_txn,3,sum)
> y <- which.min(x)
> m <- Result7E$New_Invoice_Hour[c(y,y+2)]
> m
[1] 18 20
```

Company can shut down the website from 18:00 to 20:00.

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

Answer:

```
G_data <- subset(mydata,Country=="Germany")
hist(G_data$TransactionValue)
```



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

Answer:

Customer 14911 has Highest number of transactions.

```
> Result9A <- summarise(group_by(mydata,CustomerID),Cust_Transactions=n_distinct(InvoiceNo))
> Result9A <- arrange(Result9A,desc(Cust_Transactions)) %>% na.omit(Result9A)
> head(Result9A,1)
# A tibble: 1 x 2
  CustomerID Cust_Transactions
       <int>             <int>
1      14911               248
```

Customer 14646 is the most valuable customer.

```
> Result9B <- summarise(group_by(mydata,CustomerID), Sum_Cust_Txn = sum(TransactionValue))
> Result9B <- arrange(Result9B,desc(Sum_Cust_Txn)) %>% na.omit(Result9B)
> head(Result9B,1)
# A tibble: 1 x 2
  CustomerID Sum_Cust_Txn
       <int>        <dbl>
1      14646      279489.
```

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

Answer:

```
> Miss_V = colMeans(is.na(mydata))
> Miss_V
      InvoiceNo       StockCode     Description        Quantity     InvoiceDate
      0.0000000       0.0000000       0.0000000       0.0000000       0.0000000
      UnitPrice      CustomerID         Country TransactionValue
      0.0000000       0.2492669       0.0000000       0.0000000
> Percentange_MV = Miss_V * 100
> Percentange_MV
      InvoiceNo       StockCode     Description        Quantity     InvoiceDate
        0.00000         0.00000         0.00000         0.00000         0.00000
      UnitPrice      CustomerID         Country TransactionValue
        0.00000        24.92669         0.00000         0.00000
```

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

Answer:

```
> Result11A <- filter(mydata,is.na(mydata$CustomerID))
> Result11 <- summarise(group_by(Result11A,Country),Cust_Transactions=n_distinct(InvoiceNo))
> Result11
# A tibble: 9 x 2
  Country        Cust_Transactions
  <fct>                      <int>
1 Bahrain                        2
2 EIRE                          41
3 France                         3
4 Hong Kong                     15
5 Israel                         3
6 Portugal                       1
7 Switzerland                    3
8 United Kingdom              3637
9 Unspecified                    5
```

12. On average, how often the customers 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.

Answer:

The average number of days between the consecutive shopping is 67 days.

```
> Result12 <- mydata %>% arrange(CustomerID, New_Invoice_Date) %>% group_by(CustomerID) %>% mutate(diffDate = (New_Invoice_Date - lag(New_Invoice_Date)))
> Result12 <- select(Result12,CustomerID,InvoiceNo,New_Invoice_Date,diffDate)
> Result12 <- filter(Result12,diffDate != 0)
> Result12 <- summarise(group_by(Result12,CustomerID),Avg_days = mean(diffDate))
> round(mean(Result12$Avg_days))
Time difference of 67 days
```

13. 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.

Answer:

The return rate rr for the French Customers is 14.97%

```
> French_Cust <- filter(mydata,mydata$Country== "France")
> c <- subset(French_Cust,French_Cust$Quantity < 0)
> ct <- n_distinct(c$InvoiceNo)
> tt <- n_distinct(French_Cust$InvoiceNo)
> rr <- ct/tt *100
> rr
[1] 14.96746
```

14. 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)

Answer:

The product DOT has generated the highest revenue for the retailer.

```
> Result14A <- summarise(group_by(mydata,StockCode), Sum_Cust_Txn = sum(TransactionValue))
> Result14 <- arrange(Result14A,desc(Sum_Cust_Txn))
> head(Result14,1)
# A tibble: 1 x 2
  StockCode Sum_Cust_Txn
  <fct>           <dbl>
1 DOT           206245.
```

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

Answer:

There are 4373 Unique customers in the database.

```
> Unique_cust <- unique(mydata$CustomerID)
> length(Unique_cust)
[1] 4373
> #OR
> n_distinct(mydata$CustomerID)
[1] 4373
```