Business Analytics Assignment 1

SrushtiP

10/14/2019

Question 1 A)

```
pnorm(700, mean = 494, sd = 100, lower.tail = FALSE)
```

```
## [1] 0.01969927
```

Probabbility of obtaining score greater than 700 is 0.01969927

Question 1 B)

```
pnorm(450, mean = 494, sd = 100) - pnorm(350, mean = 494, sd = 100)
```

```
## [1] 0.2550349
```

Probability of getting a score between 350 and 450 on the same GMAT exam is 0.2550349.

Question 2)

```
Avg_per_dim_cost <- 449 - (qnorm(0.8665) * 36)
Avg_per_dim_cost
```

```
## [1] 409.0401
```

The average per diem cost in Buenos Aires is 409.0401

Question 3)

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

km <- mean(Kent)
lm <- mean(Los_Angeles)

numerator = sum((Kent - km)*(Los_Angeles - lm))
denomenator = sqrt(sum((Kent - km)^2)) * sqrt(sum((Los_Angeles- lm)^2))
Correlation <- numerator/denomenator

Correlation
```

```
## [1] -0.3566049
```

```
cor(Kent, Los_Angeles)
```

```
## [1] -0.3566049
```

Correlation is negative by 0.3566049

Read the data file.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)

OR_object <- read_csv("Online_Retail.csv")
```

```
## Parsed with column specification:
## cols(
##   InvoiceNo = col_character(),
##   StockCode = col_character(),
##   Description = col_character(),
##   Quantity = col_double(),
##   InvoiceDate = col_character(),
##   UnitPrice = col_double(),
##   CustomerID = col_double(),
##   Country = col_character()
## )
```

```
head(OR_object)
```

```
## # A tibble: 6 x 8
##   InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice CustomerID
##   <chr>     <chr>     <chr>          <dbl> <chr>           <dbl>      <dbl>
## 1 536365    85123A    WHITE HANG~        6 12/1/2010 ~      2.55      17850
## 2 536365    71053     WHITE META~        6 12/1/2010 ~      3.39      17850
## 3 536365    84406B    CREAM CUPI~        8 12/1/2010 ~      2.75      17850
## 4 536365    84029G    KNITTED UN~        6 12/1/2010 ~      3.39      17850
## 5 536365    84029E    RED WOOLLY~        6 12/1/2010 ~      3.39      17850
## 6 536365    22752     SET 7 BABU~        2 12/1/2010 ~      7.65      17850
## # ... with 1 more variable: Country <chr>
```

```
summary(OR_object)
```

```
##   InvoiceNo          StockCode         Description
##  Length:541909      Length:541909      Length:541909
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Quantity         InvoiceDate          UnitPrice
##  Min.   :-80995.00   Length:541909      Min.   :-11062.06
##  1st Qu.:     1.00   Class :character   1st Qu.:     1.25
##  Median :     3.00   Mode  :character   Median :     2.08
##  Mean   :     9.55                      Mean   :     4.61
##  3rd Qu.:    10.00                      3rd Qu.:     4.13
##  Max.   : 80995.00                      Max.   : 38970.00
##
##    CustomerID       Country
##  Min.   :12346    Length:541909
##  1st Qu.:13953    Class :character
##  Median :15152    Mode  :character
##  Mean   :15288
##  3rd Qu.:16791
##  Max.   :18287
##  NA's   :135080
```

Question 4)

```
Total_Country_Transaction <- tapply(OR_object$InvoiceNo, OR_object$Country, NRO
W) / NROW(OR_object$InvoiceNo) * 100
subset(Total_Country_Transaction ,as.data.frame(Total_Country_Transaction) >1)
```

```
##           EIRE         France        Germany United Kingdom
##       1.512431       1.579047       1.752139      91.431956
```

Countries accounting for more than 1% of the total transactions are EIRE, France, Germany, United Kingdom

Question 5)

```
OR_object$TransactionValue <- OR_object$Quantity * OR_object$UnitPrice
summary(OR_object)
```

```
##   InvoiceNo          StockCode         Description
##  Length:541909      Length:541909      Length:541909
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Quantity         InvoiceDate          UnitPrice
##  Min.   :-80995.00   Length:541909      Min.   :-11062.06
##  1st Qu.:     1.00   Class :character   1st Qu.:     1.25
##  Median :     3.00   Mode  :character   Median :     2.08
##  Mean   :     9.55                      Mean   :     4.61
##  3rd Qu.:    10.00                      3rd Qu.:     4.13
##  Max.   : 80995.00                      Max.   : 38970.00
##
##    CustomerID       Country          TransactionValue
##  Min.   :12346    Length:541909      Min.   :-168469.60
##  1st Qu.:13953    Class :character   1st Qu.:      3.40
##  Median :15152    Mode  :character   Median :      9.75
##  Mean   :15288                       Mean   :     17.99
##  3rd Qu.:16791                       3rd Qu.:     17.40
##  Max.   :18287                       Max.   : 168469.60
##  NA's   :135080
```

Question 6)

```
Money_Spend <- aggregate(OR_object$TransactionValue, by = list(OR_object$Countr
y), FUN = sum)
colnames(Money_Spend)<- c("Country", "Transaction_Value_Spend")
subset(Money_Spend, Money_Spend[2] > 130000)
```

```
##           Country Transaction_Value_Spend
## 1       Australia                137077.3
## 11           EIRE                263276.8
## 14         France                197403.9
## 15        Germany                221698.2
## 25    Netherlands                284661.5
## 36 United Kingdom               8187806.4
```

Countries with total transaction exceeding 130,000 British Pound are Australia, EIRE, France, Germany, Netherlands, United Kingdom.

Question 7)

```
Temp=strptime(OR_object$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')
OR_object$New_Invoice_Date <- as.Date(Temp)
OR_object$New_Invoice_Date[20000]- OR_object$New_Invoice_Date[10]
```

```
## Time difference of 8 days
```

```
OR_object$Invoice_Day_Week= weekdays(OR_object$New_Invoice_Date)
OR_object$New_Invoice_Hour = as.numeric(format(Temp, "%H"))
OR_object$New_Invoice_Month = as.numeric(format(Temp, "%m"))

head(OR_object)
```

```
## # A tibble: 6 x 13
##   InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice CustomerID
##   <chr>     <chr>     <chr>          <dbl> <chr>           <dbl>      <dbl>
## 1 536365    85123A    WHITE HANG~        6 12/1/2010 ~      2.55      17850
## 2 536365    71053     WHITE META~        6 12/1/2010 ~      3.39      17850
## 3 536365    84406B    CREAM CUPI~        8 12/1/2010 ~      2.75      17850
## 4 536365    84029G    KNITTED UN~        6 12/1/2010 ~      3.39      17850
## 5 536365    84029E    RED WOOLLY~        6 12/1/2010 ~      3.39      17850
## 6 536365    22752     SET 7 BABU~        2 12/1/2010 ~      7.65      17850
## # ... with 6 more variables: Country <chr>, TransactionValue <dbl>,
## #   New_Invoice_Date <date>, Invoice_Day_Week <chr>,
## #   New_Invoice_Hour <dbl>, New_Invoice_Month <dbl>
```

```
summary(OR_object)
```

```
##   InvoiceNo          StockCode         Description
##  Length:541909      Length:541909      Length:541909
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Quantity         InvoiceDate          UnitPrice
##  Min.   :-80995.00   Length:541909      Min.   :-11062.06
##  1st Qu.:     1.00   Class :character   1st Qu.:     1.25
##  Median :     3.00   Mode  :character   Median :     2.08
##  Mean   :     9.55                      Mean   :     4.61
##  3rd Qu.:    10.00                      3rd Qu.:     4.13
##  Max.   : 80995.00                      Max.   : 38970.00
##
##    CustomerID       Country          TransactionValue
##  Min.   :12346    Length:541909      Min.   :-168469.60
##  1st Qu.:13953    Class :character   1st Qu.:      3.40
##  Median :15152    Mode  :character   Median :      9.75
##  Mean   :15288                       Mean   :     17.99
##  3rd Qu.:16791                       3rd Qu.:     17.40
##  Max.   :18287                       Max.   : 168469.60
##  NA's   :135080
##  New_Invoice_Date     Invoice_Day_Week   New_Invoice_Hour
##  Min.   :2010-12-01   Length:541909      Min.   : 6.00
##  1st Qu.:2011-03-28   Class :character   1st Qu.:11.00
##  Median :2011-07-19   Mode  :character   Median :13.00
##  Mean   :2011-07-04                      Mean   :13.08
##  3rd Qu.:2011-10-19                      3rd Qu.:15.00
##  Max.   :2011-12-09                      Max.   :20.00
##
##  New_Invoice_Month
##  Min.   : 1.000
##  1st Qu.: 5.000
##  Median : 8.000
##  Mean   : 7.553
##  3rd Qu.:11.000
##  Max.   :12.000
##
```

Question 7 A)

```
tapply(OR_object$InvoiceNo , OR_object$Invoice_Day_Week, NROW) / NROW(OR_object
$InvoiceNo) * 100
```

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

Question 7 B)

```
tapply(OR_object$TransactionValue , OR_object$Invoice_Day_Week, sum) / sum(OR_o
bject$TransactionValue) * 100
```

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

Question 7 C)

```
tapply(OR_object$TransactionValue , OR_object$New_Invoice_Month , sum) / sum(OR
_object$TransactionValue) * 100
```

```
##         1         2         3         4         5         6         7
##  5.744919  5.109515  7.009487  5.059703  7.420519  7.090080  6.989308
##         8         9        10        11        12
##  7.003469 10.460751 10.984123 14.995836 12.132290
```

Question 7 D)

```
OR_object$New_Invoice_Date[max(OR_object$TransactionValue[OR_object$Country ==
"Australia"])]
```

```
## [1] "2010-12-01"
```

Question 7 E)

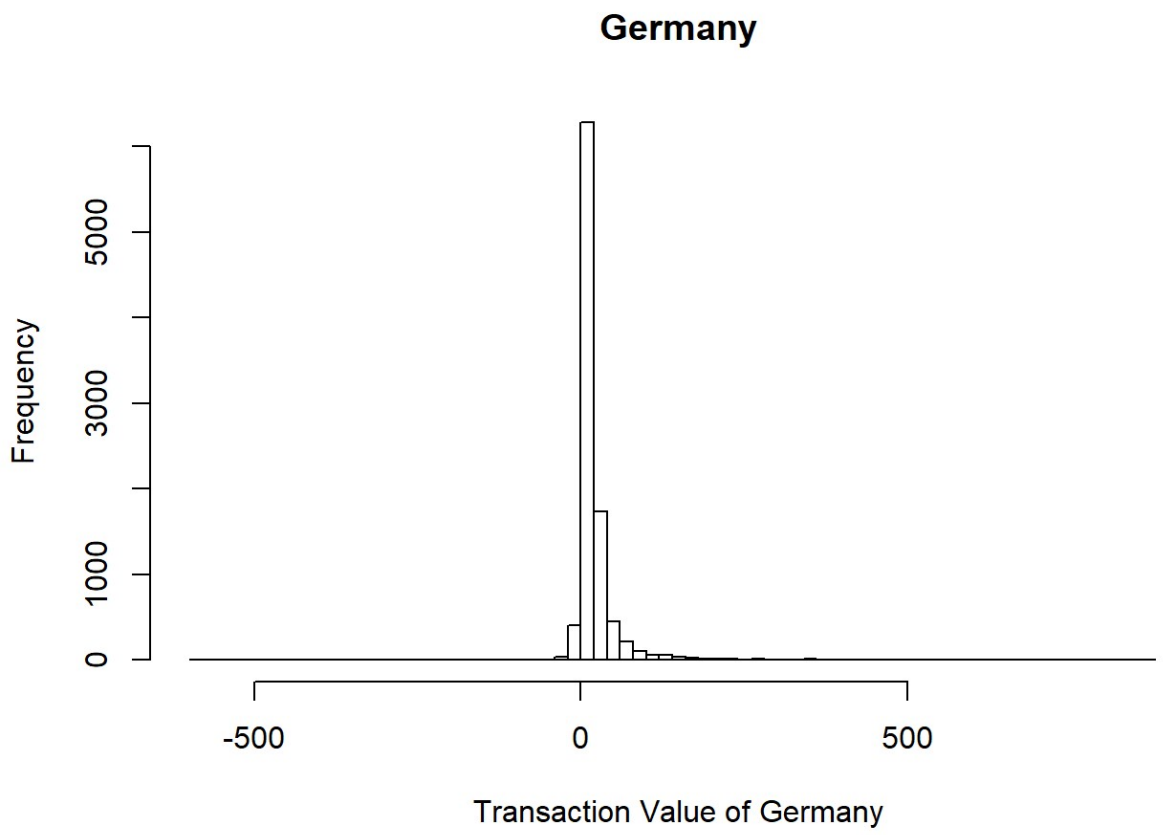
```
which.min(abs(diff(table(OR_object$New_Invoice_Hour), lag = 1, differences =
2)))
```

```
## 20
## 13
```

As the minimum value is for the 20th hour hence the store can be closed for maintenence between 6 PM to 8 PM

Question 8)

```
Germany_Transaction <- subset(OR_object$TransactionValue, OR_object$Country ==
"Germany")
hist(Germany_Transaction, xlim = c (-600, 900), breaks = 100 , xlab = "Transact
ion Value of Germany", main = "Germany")
```



Question 9)

```
NumberOfTransaction <- tapply(OR_object$TransactionValue,OR_object$CustomerID,
length)
NumberOfTransaction[which.max(NumberOfTransaction)]
```

```
## 17841
##  7983
```

```
ValuableTransaction <- tapply(OR_object$TransactionValue,OR_object$CustomerID,
sum)
ValuableTransaction[which.max(ValuableTransaction)]
```

```
##  14646
## 279489
```

Customer had the highest number of transactions is with CustomerID as 17841. Most Valuable customer is with customerID 14646.

Question 10)

```
colMeans(is.na(OR_object)) * 100
```

```
##         InvoiceNo         StockCode       Description          Quantity
##         0.0000000         0.0000000         0.2683107         0.0000000
##       InvoiceDate         UnitPrice        CustomerID           Country
##         0.0000000         0.0000000        24.9266943         0.0000000
##  TransactionValue  New_Invoice_Date  Invoice_Day_Week  New_Invoice_Hour
##         0.0000000         0.0000000         0.0000000         0.0000000
## New_Invoice_Month
##         0.0000000
```

Question 11)

```
NA_Sum <- function(input){
  Total_NA <- sum(is.na(input))
  return(Total_NA)
}

tapply(OR_object$CustomerID , OR_object$Country, NA_Sum)
```

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

Question 12)

```
Customer_Visit_Count<-as.data.frame(table(OR_object$CustomerID))
colnames(Customer_Visit_Count)<-c("CustomerID","NumberOfVisits")
round(mean(abs(diff(Customer_Visit_Count$NumberOfVisits))))
```

```
## [1] 117
```

Avgerage times a customer visits a store is 117 times.

```
func <- function(x){
  y <- abs(diff.Date(x))
  z <- mean.difftime(x)
  return(z)
}

temp <- OR_object[order(OR_object$CustomerID),]
xyz <- aggregate(temp$New_Invoice_Date, by = list(temp$CustomerID), FUN = func)
View(xyz)

xyz <- unlist(xyz$x)
xyz<- xyz[xyz != 0]

round(mean(xyz)/(24*60))
```

```
## Time difference of 11
```

On an average a customer returns to the online store after 11 days.

Question 13)

```
NROW(OR_object$Quantity [OR_object$Quantity < 0 & OR_object$Country == "Franc
e"] ) / NROW(OR_object) * 100
```

```
## [1] 0.02749539
```

Return rate for the French customers 0.02749539%

Question 14)

```
Revenue<- aggregate(OR_object$TransactionValue, by = list(OR_object$Descriptio
n), FUN = sum)
colnames(Revenue) <- c("Customer", "Revenue")
Revenue[which.max(Revenue$Revenue),]
```

```
##            Customer  Revenue
## 1128 DOTCOM POSTAGE 206245.5
```

The Dotcom Postage customer is most valuable customer with highest revenue.

Question 15)

```
length(unique(OR_object$CustomerID))
```

```
## [1] 4373
```

There are 4373 unique customers in dataset.