BA assignment 1

#### the file contains Descriptive statistics solutions and Data wrangling results of a dataset of Transaction data of Online Retail company and its countries in multiple countries.

### Part A) Descriptive Statistics & Normal Distributions

1.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 ?

\*We will do this in three steps: + Step 1: Lets calculate the probabiliy of getting score greater than 50. The Z-score for 700 is (700-494)/100=2.06 + Step 2:Calculate probabibily of z score

1-pnorm(2.06)

## [1] 0.01969927

### or

1-pnorm(700,mean=494,sd=100)

## [1] 0.01969927

So 1.97% of total would exceed score of 700.

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.

\*We will do this in two steps: + The Z-score for 86.65% is Calculated by q norm

qnorm(0.8665)

## [1] 1.109998

if z is 1.11 value of mean = $449 - $36(1.11)= $409.04

3.Calculate the correlation (Pearson Correlation Coefficient) between the temperatures of the two cities without using any R commands i.e. calculate step by step. Kent=c(59, 68, 78, 60) Los\_Angeles=c(90, 82, 78, 75)

library(ISLR)  
kent=c(59,68,78,60)  
Los\_Angeles=c(90,82,78,75)  
cor(kent,Los\_Angeles,method = 'pearson')

## [1] -0.3566049

## Part B) Data Wrangling

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

invoice <- read.csv(file = "Online\_Retail.csv")  
colnames(invoice)

## [1] "InvoiceNo" "StockCode" "Description" "Quantity" "InvoiceDate"  
## [6] "UnitPrice" "CustomerID" "Country"

4.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). Showing total number and also in percentage. only countries accounting for more than 1% of the total transactions.

invoice %>% group\_by(Country) %>%   
 summarise(Total\_Transactions=n(), Percentage\_Transactions = (n()/nrow(invoice))\*100) %>%   
 filter(Percentage\_Transactions > 1)

## # A tibble: 3 x 3  
## Country Total\_Transactions Percentage\_Transactions  
## <fct> <int> <dbl>  
## 1 France 967 1.48  
## 2 Germany 982 1.50  
## 3 United Kingdom 61186 93.4

1. a new variable ‘TransactionValue’ that is the product of the exising ‘Quantity’ and ‘UnitPrice’ variables. Add this variable to the dataframe.

invoice$TransactionValue = invoice$Quantity \* invoice$UnitPrice  
head(invoice)

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

6.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

invoice %>% group\_by(Country) %>% summarise(TransValue=sum(TransactionValue)) %>% filter(TransValue>130000)

## # A tibble: 1 x 2  
## Country TransValue  
## <fct> <dbl>  
## 1 United Kingdom 965043.

7 .for the set of next fpur questions we , first covert and split the invoice date field into day of the week , month and hour

Temp=strptime(invoice$InvoiceDate,format='%m/%d/%Y %H:%M',tz='GMT')  
invoice$New\_Invoice\_Date <- as.Date(Temp)  
invoice$Invoice\_Day\_Week= weekdays(invoice$New\_Invoice\_Date)  
invoice$New\_Invoice\_Hour = as.numeric(format(Temp, "%H"))  
invoice$New\_Invoice\_Month = as.numeric(format(Temp, "%m"))

7.a)percentage of transactions (by numbers) by days of the week

invoice %>% group\_by(Invoice\_Day\_Week) %>% summarise(trans\_perday = n()/nrow(invoice) \* 100)

## # A tibble: 6 x 2  
## Invoice\_Day\_Week trans\_perday  
## <chr> <dbl>  
## 1 Friday 17.4   
## 2 Monday 19.0   
## 3 Sunday 9.86  
## 4 Thursday 18.5   
## 5 Tuesday 16.4   
## 6 Wednesday 18.9

7.b)percentage of transactions (by transaction volume) by days of the week

Total\_TransactionVal = sum(invoice$TransactionValue)  
Total\_TransactionVal

## [1] 1122291

7.c)percentage of transactions (by transaction volume) by month of the year

invoice %>% group\_by(New\_Invoice\_Month) %>% summarise(trans\_perday = sum(TransactionValue)/Total\_TransactionVal \* 100)

## # A tibble: 2 x 2  
## New\_Invoice\_Month trans\_perday  
## <dbl> <dbl>  
## 1 1 33.3  
## 2 12 66.7

7.d)the date with the highest number of transactions from Australia

Aus\_Tran = invoice %>% filter(Country == 'Australia') %>% group\_by(New\_Invoice\_Date) %>% summarise(max\_tran\_amt = sum(TransactionValue))  
Aus\_Tran %>% filter(max\_tran\_amt == max(max\_tran\_amt))

## # A tibble: 1 x 2  
## New\_Invoice\_Date max\_tran\_amt  
## <date> <dbl>  
## 1 2011-01-06 7154.

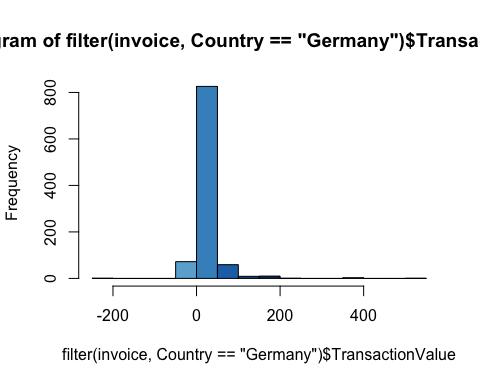
7.e)the hour of the day to start this so that the distribution is at minimum for the customers.The responsible IT team is available from 7:00 to 20:00 every day

t = invoice %>% group\_by(New\_Invoice\_Hour) %>% summarise(tran\_count = n())

# Result :7 to 8 would be the best hours

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

hist(filter(invoice,Country=='Germany')$TransactionValue,col = blues9)

 9.the highest number of transactions and the most valuable customer (i.e. highest total sum of transactions) +customer with highest transaction

customer\_grp <- invoice %>% group\_by(CustomerID) %>% summarise(customer\_transaction = n())  
filter(customer\_grp,customer\_transaction==max(customer\_grp$customer\_transaction))

## # A tibble: 1 x 2  
## CustomerID customer\_transaction  
## <int> <int>  
## 1 NA 25281

+customer with highest value

highestval\_cust <- invoice %>% group\_by(CustomerID) %>% summarise(highestvalue = sum(TransactionValue))  
highestval\_cust = highestval\_cust[complete.cases(highestval\_cust),]  
filter(highestval\_cust,highestvalue==max(highestval\_cust$highestvalue))

## # A tibble: 1 x 2  
## CustomerID highestvalue  
## <int> <dbl>  
## 1 18102 27835.

10.the percentage of missing values for each variable in the dataset

colMeans(is.na(invoice))

## InvoiceNo StockCode Description Quantity   
## 0.0000000 0.0000000 0.0000000 0.0000000   
## InvoiceDate UnitPrice CustomerID Country   
## 0.0000000 0.0000000 0.3859754 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

11.the number of transactions with missing CustomerID records by countries

invoice %>% group\_by(Country) %>% summarise(missing\_customerid = sum(is.na(CustomerID)))

## # A tibble: 24 x 2  
## Country missing\_customerid  
## <fct> <int>  
## 1 Australia 0  
## 2 Austria 0  
## 3 Bahrain 2  
## 4 Belgium 0  
## 5 Channel Islands 0  
## 6 Cyprus 0  
## 7 Denmark 0  
## 8 EIRE 42  
## 9 Finland 0  
## 10 France 0  
## # ... with 14 more rows

12.how often the costumers comeback to the website for their next shopping

dt <- invoice %>% group\_by(New\_Invoice\_Date) %>% summarise(cnt=n())  
avg\_days = max(diff(dt$New\_Invoice\_Date,1))  
print(avg\_days)

## Time difference of 12 days

13.return rate for the French customers.Considering the cancelled transactions as those where the ‘Quantity’ variable has a negative value.

frech\_cust <- filter(invoice,Country == 'France')   
  
cancell\_tran <- filter(frech\_cust,Quantity < 0)   
return\_rate = nrow(cancell\_tran) / nrow(frech\_cust)  
print(return\_rate)

## [1] 0.01034126

14.the product that has generated the highest revenue for the retailer

prod = invoice %>% group\_by(Description) %>% summarise(highestprodval = sum(TransactionValue))  
filter(prod, highestprodval == max(highestprodval))

## # A tibble: 1 x 2  
## Description highestprodval  
## <fct> <dbl>  
## 1 REGENCY CAKESTAND 3 TIER 35120.

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

a = invoice   
a[complete.cases(a),] %>% summarise(uniquer\_customer = n())

## uniquer\_customer  
## 1 40218