**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.**

**\*\*\*\*\***

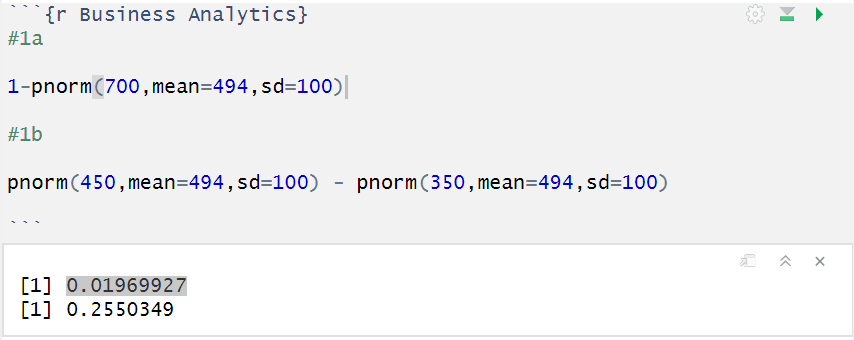
**Part A) Descriptive Statistics & Normal Distributions**

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).**

**1.96%**

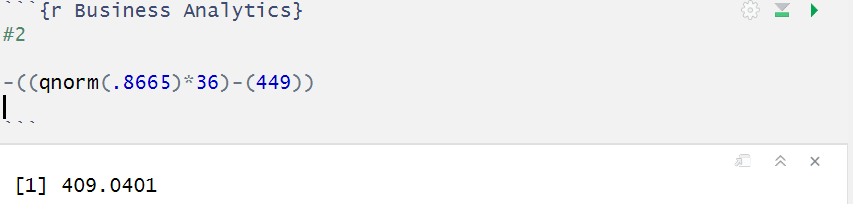
**b)** **What is the probability of getting a score between 350 and 450 on the same GMAT exam?(5 marks)**

**25.5%**



1. **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)**

**409.0401**



1. **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)**

****

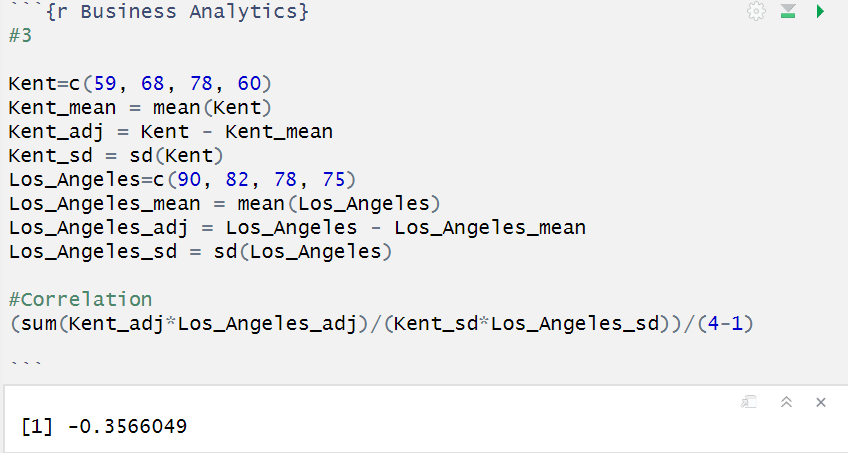
**He has sampled the mid-day temperature for days from Sep 2 to Sep 6 as follows:**

**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.**

**-0.3566049**



**Part B) Data Wrangling**

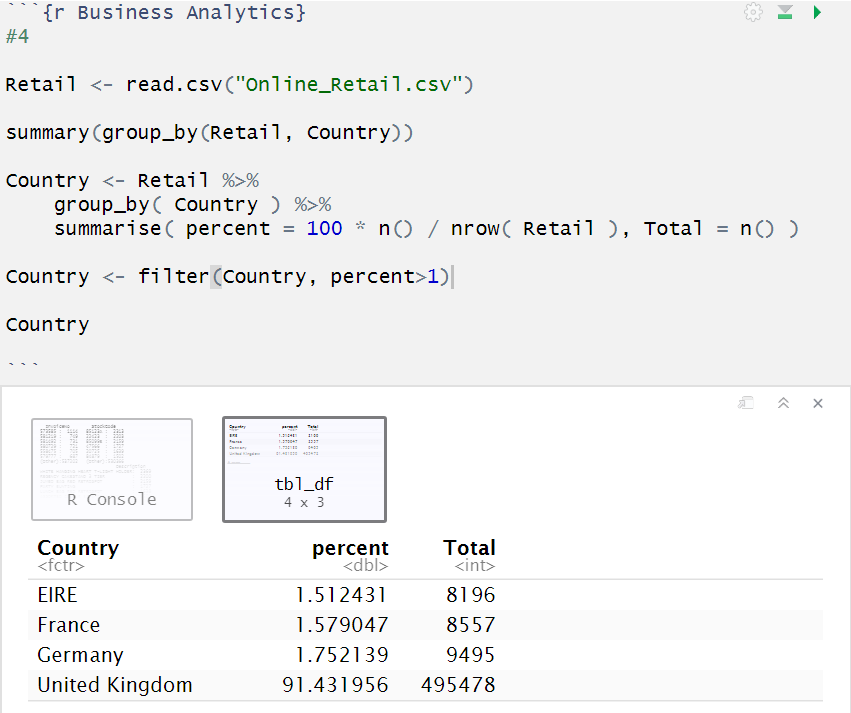
**For the questions in this part, you need to use the ‘Online Retail’ dataset which can be downloaded in CSV format from the course portal under the assignment folder. This is a transnational data set which contains all the transactions occurring between 01 Dec 2010 and 09 Dec 2011 for a UK-based and 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.**

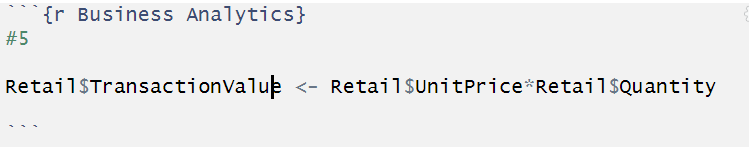
1. **Show the breakdown of the number of transactions by countries i.e. how many transactions are in the dataset for each country (consider all records including cancelled transactions). Show this in total number and also in percentage. Show only countries accounting for more than 1% of the total transactions.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

**See the below result.**

**See** 

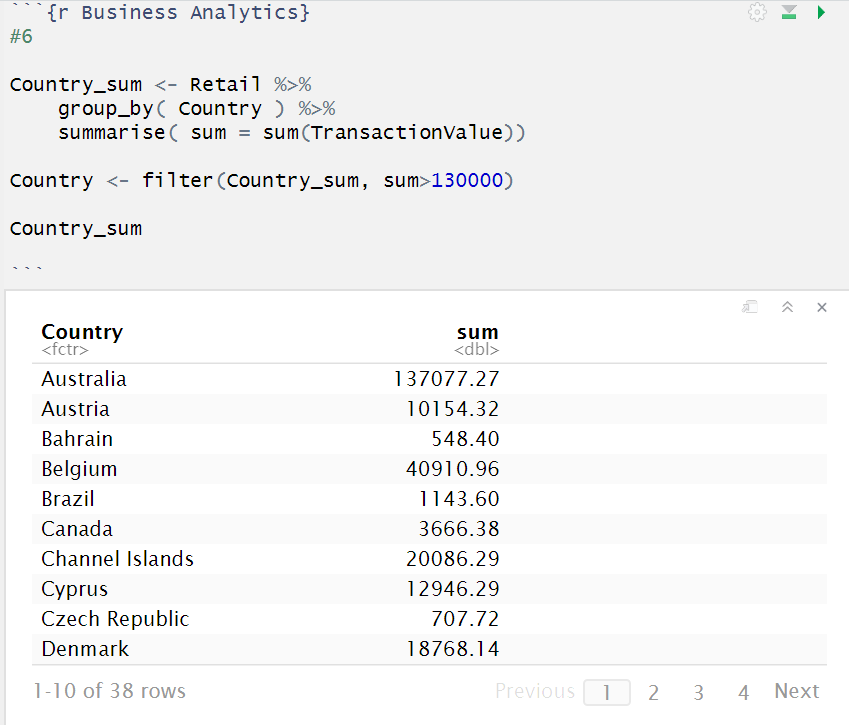
1. **Create a new variable ‘TransactionValue’ that is the product of the exising ‘Quantity’ and ‘UnitPrice’ variables. Add this variable to the dataframe. (5 marks)**

**See the below code:**



1. **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)**

**See the below results**



1. **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**](https://stat.ethz.ch/R-manual/R-devel/library/base/html/as.POSIXlt.html) **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"))**

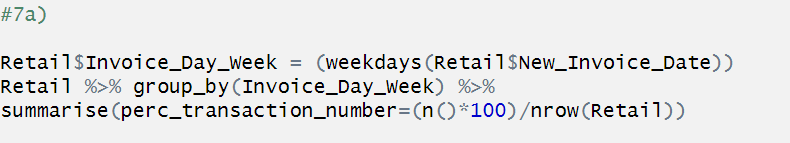
**Finally, lets define the month as a separate numeric variable too:**

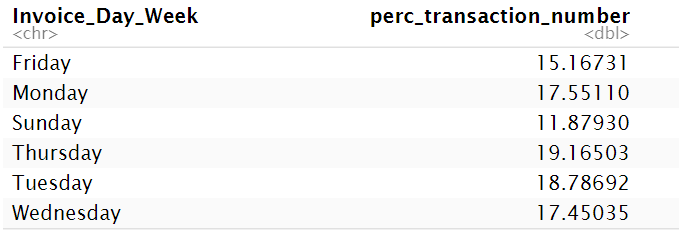
**Online\_Retail$New\_Invoice\_Month = as.numeric(format(Temp, "%m"))**

**Now answer the flowing questions.**

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

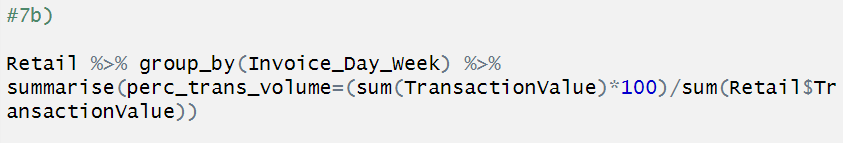
**See the below tables**

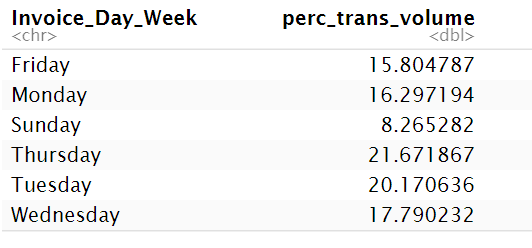




1. **Show the percentage of transactions (by transaction volume) by days of the week (extra 1 marks)**

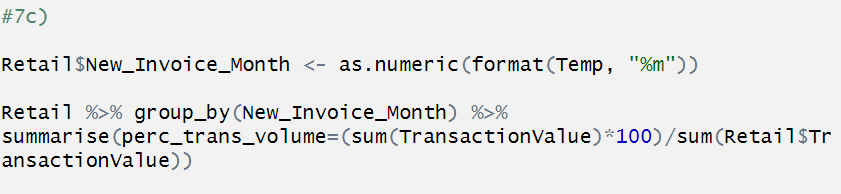
**See the below chart**

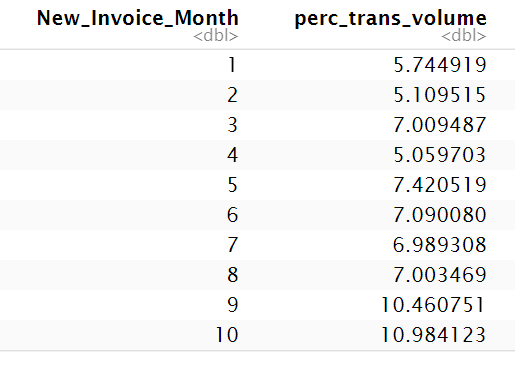




1. **Show the percentage of transactions (by transaction volume) by month of the year (extra 1 marks)**

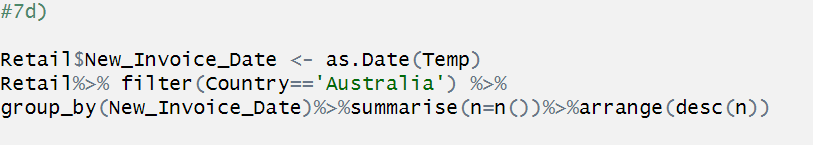
**See the below chart**

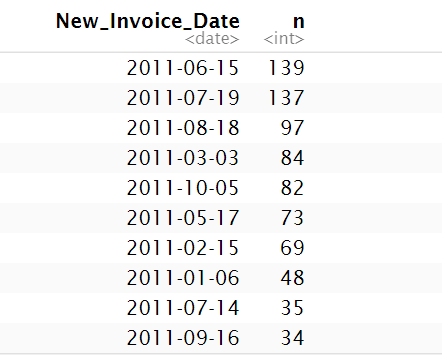




1. **What was the date with the highest number of transactions from Australia? (3 marks)**

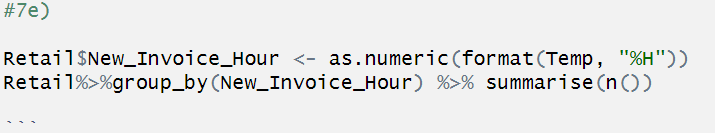
**See the below chart**

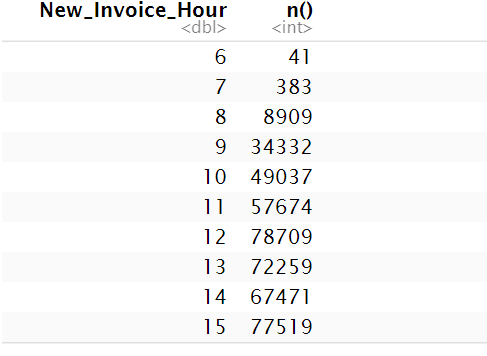




1. **The company needs to shut down the website for two consecutive hours for maintenance. What would be the hour of the day to start this so that the distribution is at minimum for the customers? The responsible IT team is available from 7:00 to 20:00 every day(3 marks)**

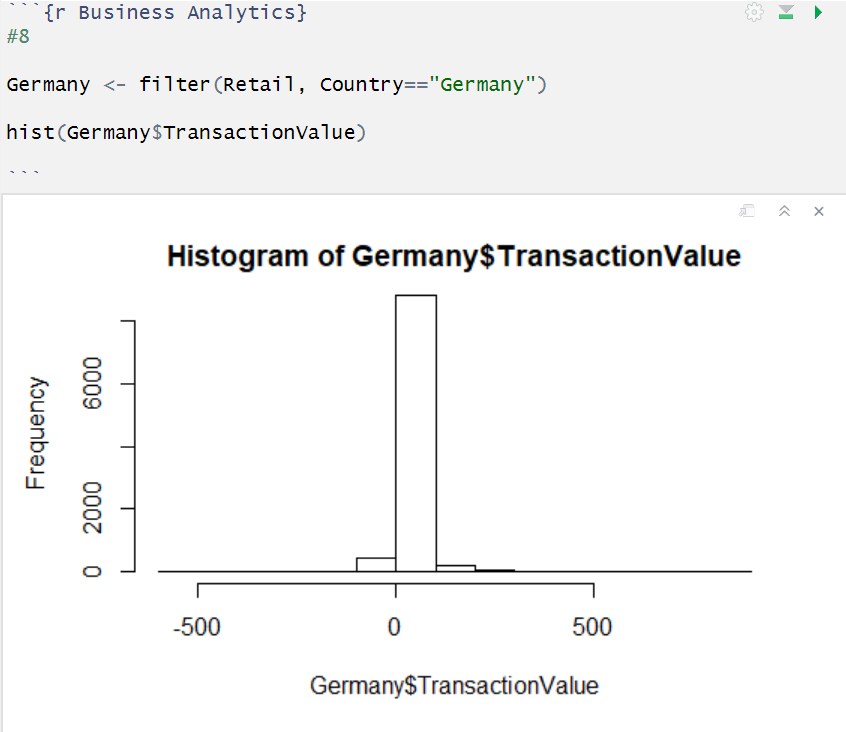
**See the below chart**





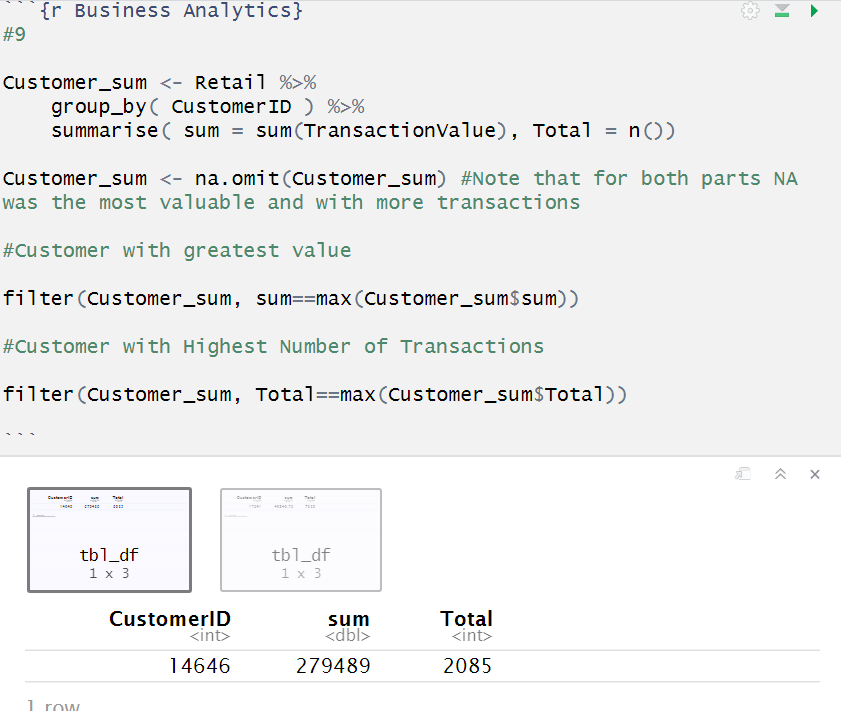
**8. Plot the histogram of transaction values from Germany. Use the hist() function to plot. (5 marks)**

**See the below chart**

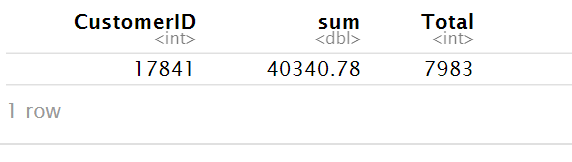


1. **Which customer had the highest number of transactions? Which customer is most valuable (i.e. highest total sum of transactions)? (10 marks)**

**See the code for all and the chart for the highest value**

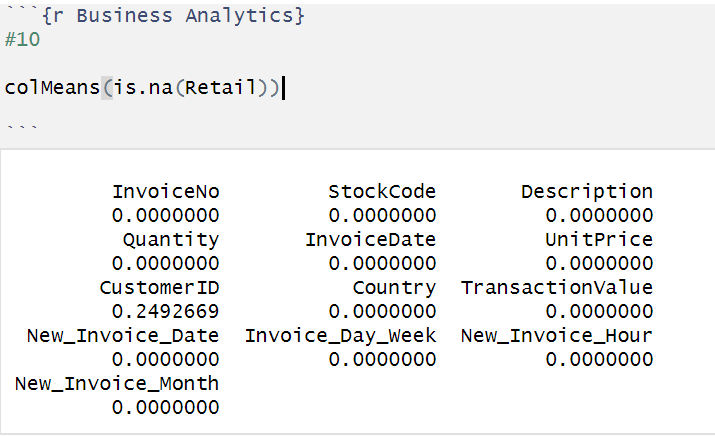


**Highest Number of Transactions**



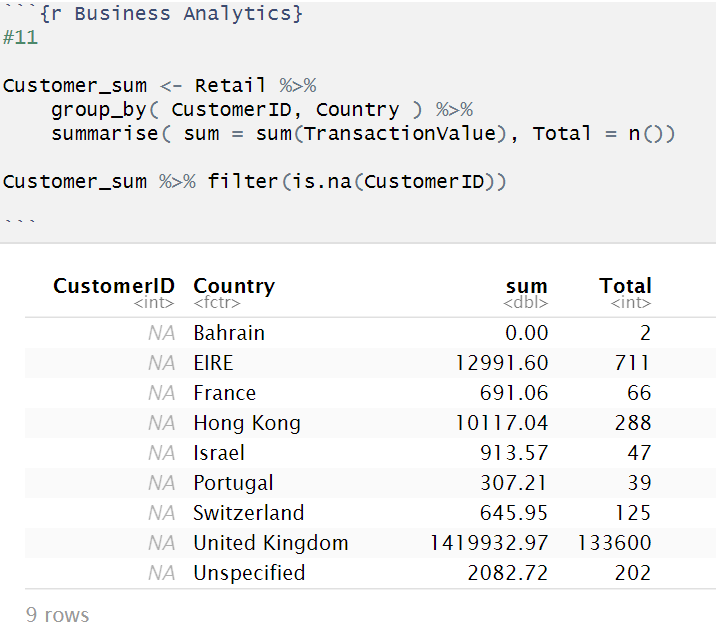
1. **Calculate the percentage of missing values for each variable in the dataset (5 marks). Hint colMeans():**

**See the below code and results**



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

**See the below code and results**

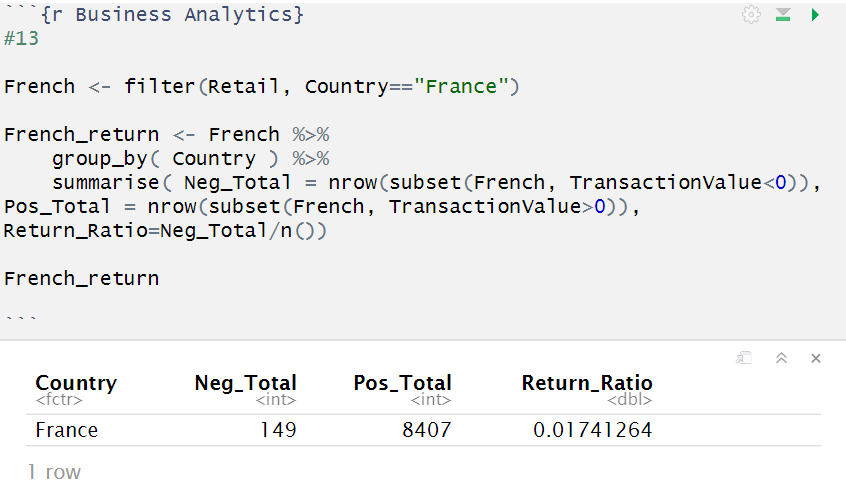


**12. On average, how often the costumers comeback to the website for their next shopping? (i.e. what is the average number of days between consecutive shopping) (Optional/Golden question: 18 additional marks!) Hint: 1. A close approximation**

**you may find** [**diff() function**](https://stat.ethz.ch/R-manual/R-devel/library/base/html/diff.html) **useful.**

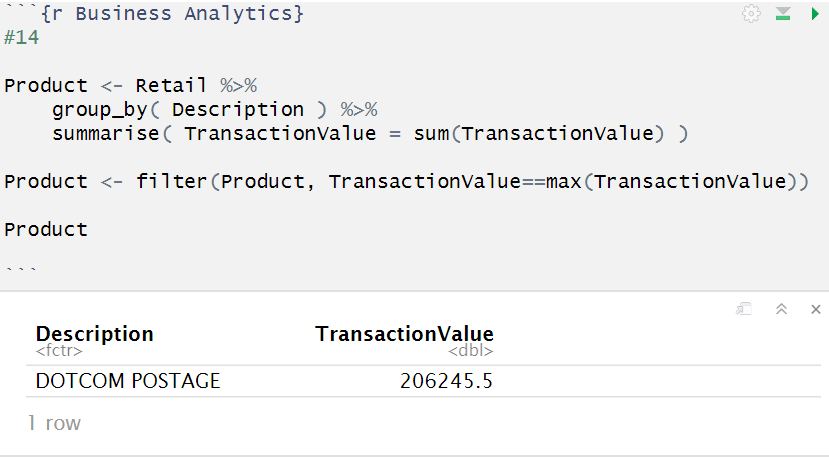
**13. In the retail sector, it is very important to understand the return rate of the goods purchased by customers. In this example, we can define this quantity, simply, as the ratio of the number of transactions cancelled (regardless of the transaction value) over the total number of transactions. With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

**See the below code**



**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)**

**See the below code**



**15. How many unique customers are represented in the dataset? You can use** [**unique()**](http://stat.ethz.ch/R-manual/R-devel/library/base/html/unique.html) **and** [**length()**](https://stat.ethz.ch/R-manual/R-devel/library/base/html/length.html) **functions. (5 marks)**

**See the below code**

