IS507 - Data, Statistical Models and Information - Assignment 3

Dataset Cleaning:

Shrey_Shah_Assignment_3.R

Source on Save

```r
# Load library readr to read the dataset
library(readr)
library(dplyr)
library(tidyr)
library(stringr)

# Set working directory
setwd("~/Desktop/IS507 - Data, Statistical Models and Information/Assignments/Assignment_3")

# Import the dataset using read_csv()
raw_dataset <- read_csv("Starbucks_data-2.csv")
dataset <- raw_dataset

# Shape of dataset
dim(dataset)

# Count of missing values
sum(is.na(dataset))

# Delete rows with missing values
dataset <- na.omit(dataset)

# Shape of new dataset after listwise deletion
dim(dataset)
# There are 2 missing values but both are in same row

library(gmodels)
```

```
> # Set working directory
> setwd("~/Desktop/IS507 - Data, Statistical Models and Information/Assignments/Assignment_3")
>
> # Import the dataset using read_csv()
> raw_dataset <- read_csv("Starbucks_data-2.csv")
Rows: 122 Columns: 23

── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (14): Timestamp, 1. Your Gender, 2. Your Age, 3. Are you currently....?, 4. What is your annual income?, 5. How often d...
dbl  (9): 12. How would you rate the quality of Starbucks compared to other brands (Coffee Bean, Old Town White Coffee..) t...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> dataset <- raw_dataset
>
> # Shape of dataset
> dim(dataset)
[1] 122  23
>
> # Count of missing values
> sum(is.na(dataset))
[1] 20
>
> # Delete rows with missing values
> dataset <- na.omit(dataset)
>
> # Shape of new dataset after listwise deletion
> dim(dataset)
[1] 112  23
```

1) 3. Are you currently....?(question 3) affect How would you rate the ambiance at Starbucks? (lighting, music, etc…) (question 15)?

Solution:

- The variable 'Are you currently….?' Is a categorical variable having 4 different categories of Student, Housewife, Self-employed and Employed.
- The variable 'How would you rate the ambiance at Starbucks? (lighting, music, etc…)?' Is a continuous numeric variable.
- The dependent variable in this case is rating of the ambience at Starbucks and the independent variable is whether is the individual is either a student, housewife, self-employed or employed.
- The data is not paired because we are not not carrying out any pre-test or post-test.
- There are 4 different groups in the independent variable.
- We will perform the Normality tests for the Ambience Rating variable using Shapiro-Wilk test:
  - Since the Housewife category in the variable PersonType has only 2 samples and Shapiro requires atleast 3 samples. So we will discard the rows with Housewife in the PersonRating column
  - The p-value for each group < 0.05, which indicates that the variable is not normal.
- So, in this case we have our conditions satisfies like - each subject is measured just once and there is no pre or post test. The Ambience variable is not normal and hence we will choose the **non-parametric Kruskal-Wallis test**.

Let's formulate the Null and Alternate hypothesis:

Null Hypothesis: There is no effect of the PersonType category on the Ambience Rating at Starbucks.
Alternate Hypothesis: There is an effect of the PersonType category on the Ambience Rating at Starbucks.

Assuming the level of significance $\alpha = 0.05$,
When we perform the Kruskal-Wallis test on the data, we get a p-value of 0.1047 > 0.05 and hence we cannot reject the Null Hypothesis. So, we can conclude that there is no effect of the personType on the Ambience Rating at Starbucks.

```
#=============================================================================================================

# QUESTION 1

# Let's find whether the continuous variable of Ambience rating is normal or not using Shapiro's test.

library(RVAideMemoire)

names(dataset)[4] <- "PersonType"
names(dataset)[16] <- "AmbienceRating"

byf.shapiro(as.matrix(dataset$AmbienceRating)~dataset$PersonType, data=dataset)

# Since the Housewife category in the variable PersonType has only 2 samples and Shapiro requires atleast 3 samples.
# So we will discard the rows with Housewife in the PersonRating column

new_dataset <- dataset[!(dataset$PersonType == 'Housewife'), ]
dim(new_dataset)

byf.shapiro(as.matrix(AmbienceRating)~PersonType, data=new_dataset)

# Using Kruskal-Wallis non-parametric test

kruskal.test(AmbienceRating~PersonType, data=new_dataset)

```

```
> byf.shapiro(as.matrix(dataset$AmbienceRating)~dataset$PersonType, data=dataset)
Error in shapiro.test(resp[as.numeric(fact) == i]) :
  sample size must be between 3 and 5000
> new_dataset <- dataset[!(dataset$PersonType == 'Housewife'), ]
> dim(new_dataset)
[1] 110  23
> byf.shapiro(as.matrix(AmbienceRating)~PersonType, data=new_dataset)

        Shapiro-Wilk normality tests

data:  as.matrix(AmbienceRating) by PersonType

                   W   p-value
Employed      0.8028 2.722e-07 ***
Self-employed 0.7991  0.003581 **
Student       0.9033  0.003163 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> kruskal.test(AmbienceRating~PersonType, data=new_dataset)

        Kruskal-Wallis rank sum test

data:  AmbienceRating by PersonType
Kruskal-Wallis chi-squared = 4.5128, df = 2, p-value = 0.1047
```

2) What is the relationship or association between How would you rate the service at Starbucks? (Promptness, friendliness, etc..)(question 17) and price range (question 13)?

Solution:

- From now on, we will use the new dataset only obtained after deletion of the Housewife type.
- The variable ‘17. How would you rate the service at Starbucks? (Promptness, friendliness, etc..)’ Is a continuous numeric variable.
- The variable ‘13. How would you rate the price range at Starbucks?’ Is a continuous numeric variable.
- The dependent variable in this case is the PriceRating at Starbucks and the independent variable is whether is the ServiceRating.
- The data is not paired because we are not not carrying out any pre-test or post-test.
- We will perform the Normality tests for the ServiceRating and PriceRating variable using Shapiro-Wilk test:
  - The p-value for both variables < 0.05, which indicates that the variables are not normal.
- So, in this case we have our conditions satisfies like - Both the ServiceRating and PriceRating variables are not normal and numeric and we need to find out if any there is any association between them. So we will choose the **Spearman’s correlation test**.

Let's formulate the Null and Alternate hypothesis:

Null Hypothesis: There is no relation/association between ServiceRating and PriceRating at Starbucks.
Alternate Hypothesis: There is some relation/association between ServiceRating and PriceRating at Starbucks.

Assuming the level of significance $\alpha = 0.05$,
When we perform the Spearman's correlation test on the data, we get a p-value of $0.0037 < 0.05$ and hence we reject the Null Hypothesis. So, we can conclude that there is some relationship/ association between the ServiceRating and PriceRating at Starbucks.

```
53 #=============================================================================================================
54
55 # QUESTION 2
56
57 names(new_dataset)[14] <- "PriceRating"
58 names(new_dataset)[18] <- "ServiceRating"
59
60 # Let's find whether the continuous variable Price and Service Rating is normal or not using Shapiro's test.
61
62 shapiro.test(new_dataset$PriceRating)
63 shapiro.test(new_dataset$AmbienceRating)
64
65 # Performing the Spearman correlation test because both the ratings are not normal
66 cor.test(new_dataset$ServiceRating, new_dataset$PriceRating, method = "spearman")
67
68 #=============================================================================================================
```

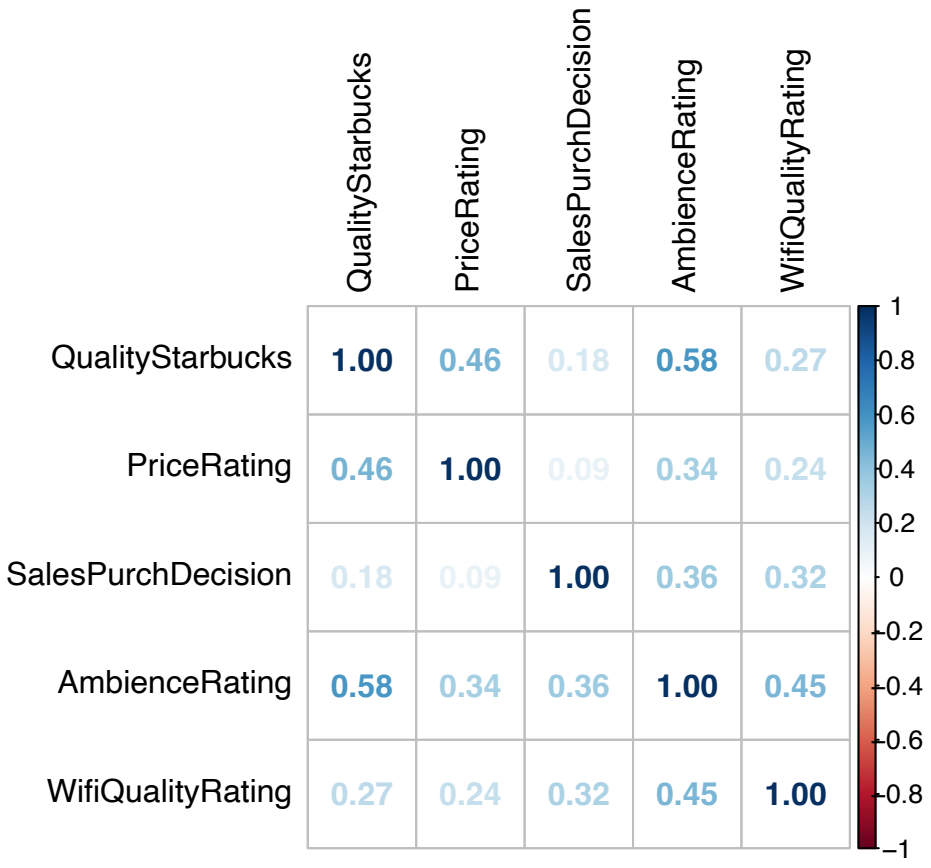
```
> shapiro.test(new_dataset$PriceRating)

        Shapiro-Wilk normality test

data:  new_dataset$PriceRating
W = 0.90691, p-value = 1.147e-06

> shapiro.test(new_dataset$AmbienceRating)

        Shapiro-Wilk normality test

data:  new_dataset$AmbienceRating
W = 0.86297, p-value = 1.149e-08

> # Performing the Spearman correlation test because both the ratings are not normal
> cor.test(new_dataset$ServiceRating, new_dataset$PriceRating, method = "spearman")

        Spearman's rank correlation rho

data:  new_dataset$ServiceRating and new_dataset$PriceRating
S = 160977, p-value = 0.003737
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2742756
```

3) Show and explain a visualization of correlations of questions 12, 13, 14, 15 and 16.
Hint: create one visualization showing all the correlations.

Solution:

- I have created a subset of the dataset using the select function in dplyr package.
- We have renamed the columns with questions 12,13,14,15,16 and used that dataset for developing the corrplot.
- We first check the normality of the all the variables using Shapiro-Wilk test and for all the variables we get p-value (very small, so we assume it as 0.001) ≅ 0.001 < 0.05 and hence conclude that the variables are not normal.
- So we can use the **Spearman's correlation** as method for plotting the correlation plot.

```
#=============================================================================

# QUESTION 3

names(new_dataset)[13] <- "QualityStarbucks"
names(new_dataset)[14] <- "PriceRating"
names(new_dataset)[15] <- "SalesPurchDecision"
names(new_dataset)[16] <- "AmbienceRating"
names(new_dataset)[17] <- "WifiQualityRating"

# Subset the dataset for getting the 5 variables

correlation_dataset <- select(new_dataset, "QualityStarbucks", "PriceRating",
                              "SalesPurchDecision","AmbienceRating",
                              "WifiQualityRating")

install.packages("corrplot")
library(corrplot)

# Plot the correlation
corrplot(cor(correlation_dataset), tl.col = "black", type = "full")

#=============================================================================
```



The above corrplot gives us some important information:

- Firstly, we see that there is a legend on the extreme right side which runs from -1 to +1. Red colour signifies negative relationship whereas blue colour signifies positive relationship.
- In the above corrplot, there is no sign of red colour which means that none of the variables under consideration have any negative correlation with each other.
- We can see that there is a **very weak to weak correlation** among variables:
  - SalesPurchDecision and PriceRating: (0.09)
  - SalesPurchDecision and QualityStarbucks: (0.18)
  - WifiQualityRating and PriceRating: (0.24)
  - WifiQualityRating and QualityStarbucks: (0.27)
- We can see that there is a **weak to moderate correlation** among variables:
  - WifiQualityRating and SalesPurchDecision: (0.32)
  - AmbienceRating and SalesPurchDecision: (0.36)
  - PriceRating and QualityStarbucks: (0.46)
  - AmbienceRating and QualityStarBucks: (0.58)

4) Create a visualization and answer the questions below, which will provide an interesting story or insight within this data.

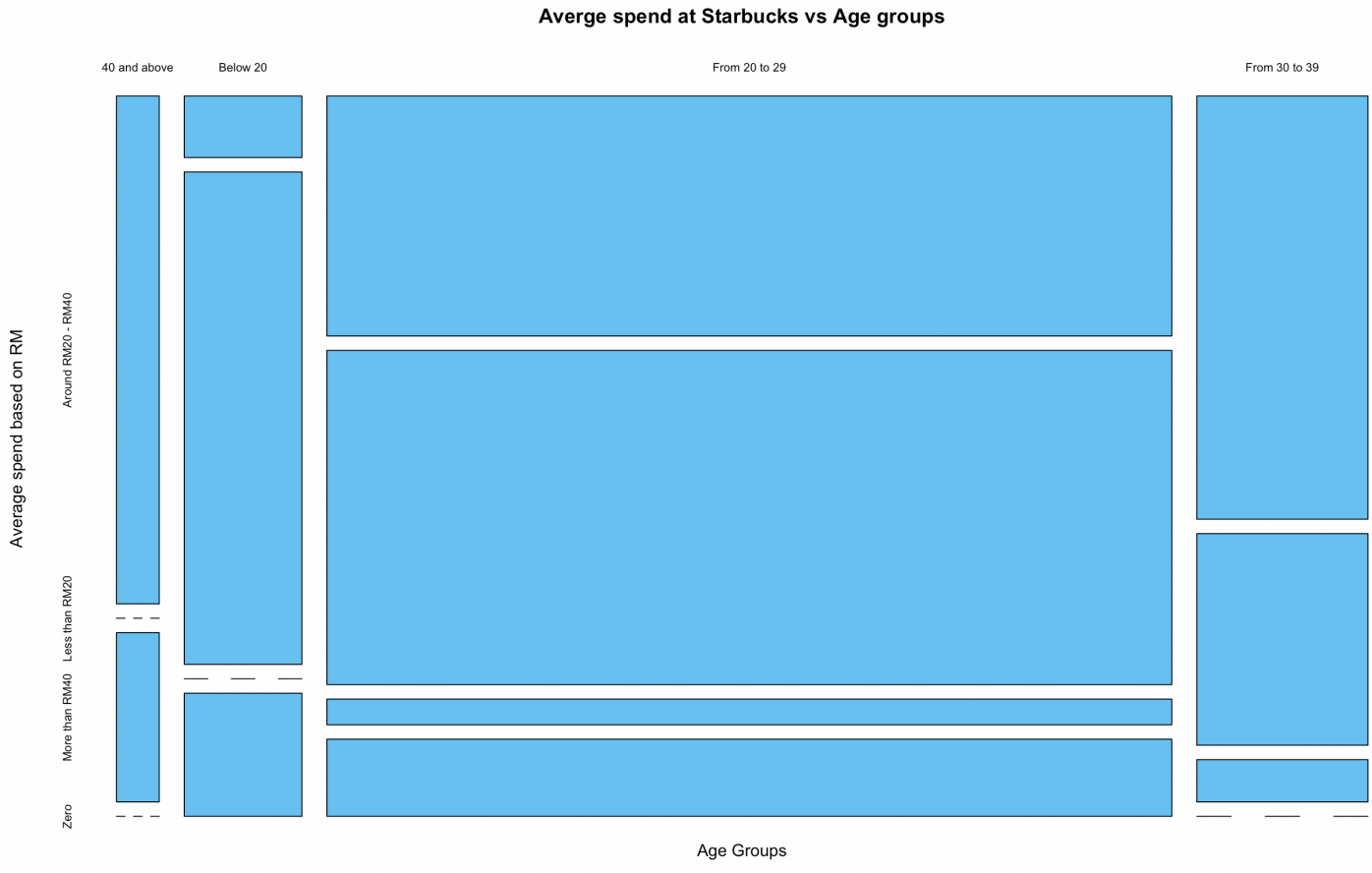
a. Who is your audience?

b. What is the application insight?

c. What does this application insight mean for the audience? Why is it important for the audience to know?

Solution:

From the dataset, we will be plotting a visualization to identify the average spend at Starbucks based on the different age groups.



a. Audience for our visualization is the technical staff including analysts at Starbucks who churn the dataset for improving sales and also non-technical staff like high-level executives who work on business decisions and developing marketing strategies.

b. I have chosen a mosaic plot to get insights about the research questions of 'What is the average spend by the customers categorised according to age groups'.

Based on the plot, we can understand that the age groups from 20 to 39 occupy the majority of the pro and has the highest contribution to the sales of Starbucks followed by approximately equal contribution from the age groups of below 20 and from 30 to 39. The age group 40 and above occupy less portion as they might not be contributing much to the sales indicating that they do not want to purchase at Starbucks.

Also, we can see that the major average spend is between RM20 - RM40 and for all age groups meaning that if they go to Starbucks they do spend definitely.

There are also some people in older age groups 30 - 39 and above who so not spend anything at Starbucks.

Teenagers and adults in range 20 to 39 spend a lot at Starbucks as they may visit and work there along with snacks and coffee.

c. This insight is critical for the executives and decision makers at Starbucks because they want to maximise their sales. So, they need to know the data at the back to get an idea of how the numbers vary accordingly. This application insight is a means for the analysts and executives to understand what are the opportunities and weaknesses in their existing marketing strategies that need to be improvised.

They need to understand how much money is being earned and who contributes the most. Based on the above insights, it is clear that the maximum income is from the people of age groups 20-39 and so they might come up with some strategies to target their main customer segment to earn more.

Alos, they come to know about their areas of weakness and improvements. For eg. in this above case the age groups from 30-39 and above 40 constitute a segment but they may not be spending much as the youth. In fact, there are some people who spend nothing at Starbucks after visiting. So this might be an important area where the executives can devise incentives which can attract this age group crowds as well.

5) There are other types of regression models outside of linear and logistic regression. Using Google Scholar, locate a journal article, which utilizes one of the types of regressions listed below or another regression outside of linear/logistic that interests you. Write a summary of the journal article and how it utilizes the regression model in two to three paragraphs. Cite the paper in APA format.

Solution:

I have chosen the Time Series Regression as I find it interesting and the paper addresses how it is used to for infectious disease and weather predictions.

Time Series Regression (TSR) modelling has been used to explain the non-infectious diseases but the same may not be applicable for infectious ones. The paper describes certain adaptations that can be made to the existing non-infectious TSR models to make them work for the infectious diseases

using the Tokyo influenza and Bangladesh cholera datasets. The common Poisson TSR model is a smooth function of time with x_t which denotes the long term trend which my be time varying variable of importance. The Beta values are the regression coefficients and the measured risk factors are denoted by z.

$$Y_t \sim Poisson(\mu_t)$$
$$\log(\mu_t) = \beta_0 + \beta x_t + \sum_p \beta_p z_{p,t} + f(t)$$

The infectious diseases have a trend that survivors are immune to re-infection and this can be a reason for change in the population vulnerable to infection. The Susceptible-Infectious-Recovery model (SIR) model uses this approach and explains that during the recovery period the vulnerable population is exhausted. But the TSR model assumes that the population at risk is consistent. So the 1st approach is to rely on the smooth function to examine the changes in immune population. Most TSR models for non-infectious disease suggests that the degree of similarity between the time series and its lagged version may be very similar. But for infectious diseases it s found to be different. So the 2nd approach discusses that in the model instead of using the lagged TS version it is better to use the logarithm of the lagged version. The association patterns and lag periods in infectious diseases may be different because its causal pattern is different and can also depend on weather as its effects are delayed. To solve this, the 3rd approach discusses that the broad lag structures and associations for infectious diseases must be selected on the basis of the type of disease. The 4th approach points out that the season and trends can be eliminated in the TSR models because they are already accounted for as they are dependent on weather, human pathogenesis and biological systems. It also says that thee should be provision for including complexity in patterns - like during national holidays can be a reason for population mix and hence infection. The last approach points out the different methods for tackling overdispersion because the variance of infectious diseases is often higher than the variance expected in Poisson TSR models. These include the negative-binomial, transformed Gaussian models etc.

How weather is associated with infectious diseases is described in paper and the approach for the same is explained. ARIMA models can be used as it can allow the addition of complex dependence on past outcomes. The season and trend aspect can be solved by the SARIMA models. Wavelet models is the new approach that can be used to incorporate the trend brought in by the factor of outbreak nature of the diseases. It can be used to understand the degree to which the incidence of infectious diseases is related to the weather changes. Based on the datasets and statistics plotted, there has been a considerable commonality in the TSR models for infectious and non-infectious diseases. The TS-SIR models are very common among the researchers to be used and if the above discussed approaches are incorporated into them, they can be used for the non-infectious diseases as well.

Citation:

Imai, C., Armstrong, B., Chalabi, Z., Mangtani, P., & Hashizume, M. (2015, July 16). *Time series regression model for infectious disease and weather*. Environmental Research. Retrieved October 21, 2022, from https://www.sciencedirect.com/science/article/pii/S0013935115300128