Quiz 1

Modern Data Mining/Linda

February 8, 2022

Name: ______________________________

Section (471, 571, 701): ______________________________

This is an open book, 10-minute quiz. Choose the correct answer(s). There might be more than one right answer in some questions. No calculations are needed.

Customer segmentation (clustering) is a process of identifying customers into a few distinctive groups. Within each group people share some common characteristics. The ultimate goal is, for example, to identify high yield segments so that more effort will be devoted to those group.

In this case study we use a dataset coming from Portuguese city regions (Lisbon, Oporto and others) which refers to clients of a wholesale distributor. It includes the annual spending in monetary units on diverse product categories. The dataset is available at UCI machine learning repository at HERE. Unfortunately no detailed information is available in the data.

Our goal is to do some basic EDA to understand the data before running sensible clustering analysis (not done here).

The data contains the following information:

- `Channel`: customers' Channel — **1 (Hotel/Restaurant/Cafe) or 2 (Retail)**
- `Region`: customers' Region — **1 (Lisbon), 2 (Oporto) or 3 (Other)**
- `Fresh`: annual spending on fresh products (continuous)
- `Milk`: annual spending on milk products (continuous)
- `Grocery`: annual spending on grocery products (continuous)
- `Frozen`: annual spending on frozen products (continuous)
- `Detergents_Paper`: annual spending on detergents and paper products (continuous)
- `Delicassen`: annual spending on and delicatessen products (continuous)

1. Let us first read and manipulate the data. The variable `Client` stands for the unique numerical ID of each client.

The data is named as `wholesale` which keeps all the original variables. We list the first three clients in the dataset:

```
url <- 'https://archive.ics.uci.edu/ml/machine-learning-databases/00292/Wholesale%20customers%20
wholesale <- read_csv(url)
wholesale <- wholesale %>%
  mutate(Channel = as.factor(Channel),Region = as.factor(Region)) %>%
  rownames_to_column('Client')
head(wholesale, 3)
```

```
# A tibble: 3 x 9
  Client Channel Region Fresh  Milk Grocery Frozen Detergents_Paper Delicassen
  <chr>  <fct>   <fct>  <dbl> <dbl>   <dbl>  <dbl>            <dbl>      <dbl>
1 1      2       3      12669  9656    7561    214             2674       1338
2 2      2       3       7057  9810    9568   1762             3293       1776
3 3      2       3       6353  8808    7684   2405             3516       7844
```

Log (base 10) is applied to the spending for all the variables. And the data is referred to as `wholesale_log`. The information for the first 3 clients is also listed here:

```
wholesale_log <- wholesale %>%
  mutate(Fresh = log10(Fresh), Milk = log10(Milk), Grocery = log10(Grocery),
         Frozen = log10(Frozen), Detergents_Paper = log10(Detergents_Paper),
         Delicassen = log10(Delicassen))
head(wholesale_log, 3)
```

```
# A tibble: 3 x 9
  Client Channel Region Fresh  Milk Grocery Frozen Detergents_Paper Delicassen
  <chr>  <fct>   <fct>  <dbl> <dbl>   <dbl>  <dbl>            <dbl>      <dbl>
1 1      2       3       4.10  3.98    3.88   2.33             3.43       3.13
2 2      2       3       3.85  3.99    3.98   3.25             3.52       3.25
3 3      2       3       3.80  3.94    3.89   3.38             3.55       3.89
```

Choose the correct description(s) for the two data frames `wholesale` and `wholesale_log`.

(A) The `Client` 2 spends 8.10 monetary units on detergents and paper products.

(B) The `Client` 1 in `wholesale` and `wholesale_log` indicate the same client.

The answer is (B).

2. Here are some summaries of variables in the dataset.

```
summary(wholesale)
```

```
    Client             Channel Region      Fresh             Milk      
 Length:440         1:298   1: 77   Min.   :     3   Min.   :   55  
 Class :character   2:142   2: 47   1st Qu.:  3128   1st Qu.: 1533  
 Mode  :character           3:316   Median :  8504   Median : 3627  
                                    Mean   : 12000   Mean   : 5796  
                                    3rd Qu.: 16934   3rd Qu.: 7190  
                                    Max.   :112151   Max.   :73498  
    Grocery          Frozen      Detergents_Paper   Delicassen    
 Min.   :    3   Min.   :   25   Min.   :    3     Min.   :    3  
 1st Qu.: 2153   1st Qu.:  742   1st Qu.:  257     1st Qu.:  408  
```

```
  Median : 4756   Median : 1526   Median :  816    Median :  966  
  Mean   : 7951   Mean   : 3072   Mean   : 2881    Mean   : 1525  
  3rd Qu.:10656   3rd Qu.: 3554   3rd Qu.: 3922    3rd Qu.: 1820  
  Max.   :92780   Max.   :60869   Max.   :40827    Max.   :47943  
```

True or false? We see the dataset `wholesale` contains missing values.

(A) True

(B) False

(C) Not enough information

The answer is (B).

3. Let us take a look at the dataset more closely.

We next arrange the `wholesale` by sorting `Milk` in an ascending order. The first three and the last three clients are listed below.

Recall:

- `Channel`: customers' Channel — **1 (Hotel/Restaurant/Cafe) or 2 (Retail)**
- `Region`: customers' Region — **1 (Lisbon), 2 (Oporto) or 3 (Other)**

```
wholesale %>% arrange(Milk) %>% head(3)
```

```
  # A tibble: 3 x 9
    Client Channel Region Fresh  Milk Grocery Frozen Detergents_Paper Delicassen
    <chr>  <fct>   <fct>  <dbl> <dbl>   <dbl>  <dbl>            <dbl>      <dbl>
  1 155    1       3        622    55     137     75                7          8
  2 99     1       3        503   112     778    895               56        132
  3 357    1       3      22686   134     218   3157                9        548
```

```
wholesale %>% arrange(Milk) %>% tail(3)
```

```
  # A tibble: 3 x 9
    Client Channel Region Fresh  Milk Grocery Frozen Detergents_Paper Delicassen
    <chr>  <fct>   <fct>  <dbl> <dbl>   <dbl>  <dbl>            <dbl>      <dbl>
  1 86     2       3      16117 46197   92780   1026            40827       2944
  2 48     2       3      44466 54259   55571   7782            24171       6465
  3 87     2       3      22925 73498   32114    987            20070        903
```

Choose the correct description(s).

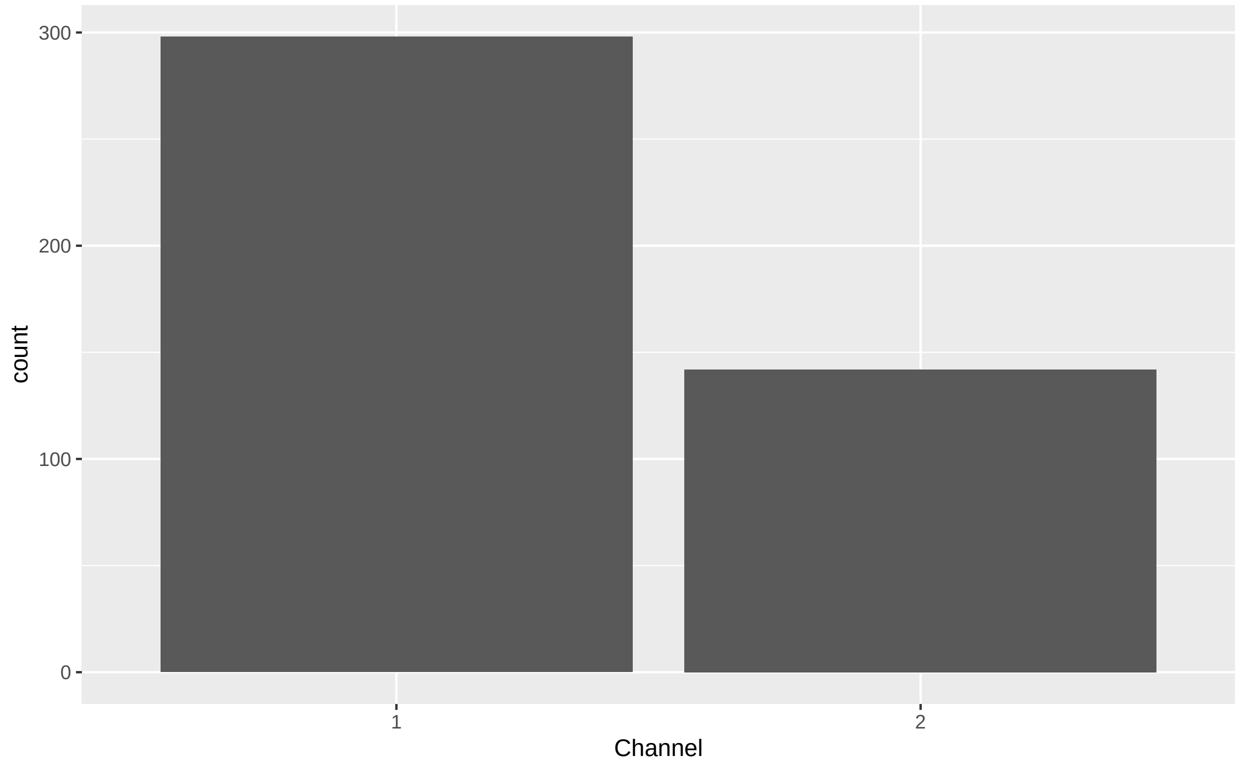
(A) The client who spends the least amount on milk products comes from the Hotel/Restaurant/Cafe channel.

(B) The client who spends the most amount on milk products spends 987 monetary units on frozen products.

The answer is (A) and (B).

4. We next count the number of clients in each channel.

```
ggplot(wholesale) + geom_bar(aes(x=Channel))
```



Which channel do the clients in the dataset come from the most?

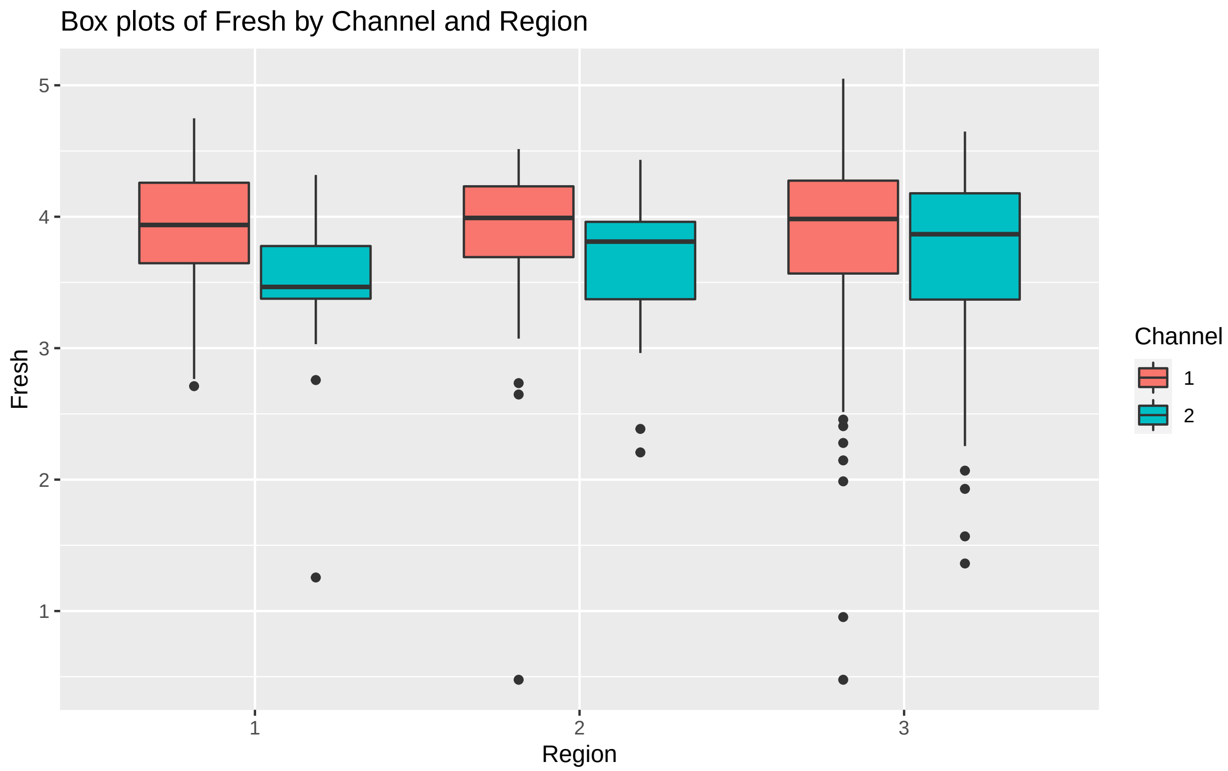
(A) 1:Hotel/Restaurant/Cafe

(B) 2:Retail

The answer is (A).

5. We would further analyze the spending on fresh products with respect to the clients' channel and region.

```
wholesale_log %>% ggplot() +
  geom_boxplot(aes(x = Region, y = Fresh, fill = Channel)) +
  labs( title = "Box plots of Fresh by Channel and Region")
```



Recall:

- `Channel`: customers' Channel — **1 (Hotel/Restaurant/Cafe) or 2 (Retail)**
- `Region`: customers' Region — **1 (Lisbon), 2 (Oporto) or 3 (Other)**

True or false? Based on the boxplot, the median spending of clients from the Hotel/Restaurant/Cafe channel is higher than that of clients from the Retail channel regardless of region.

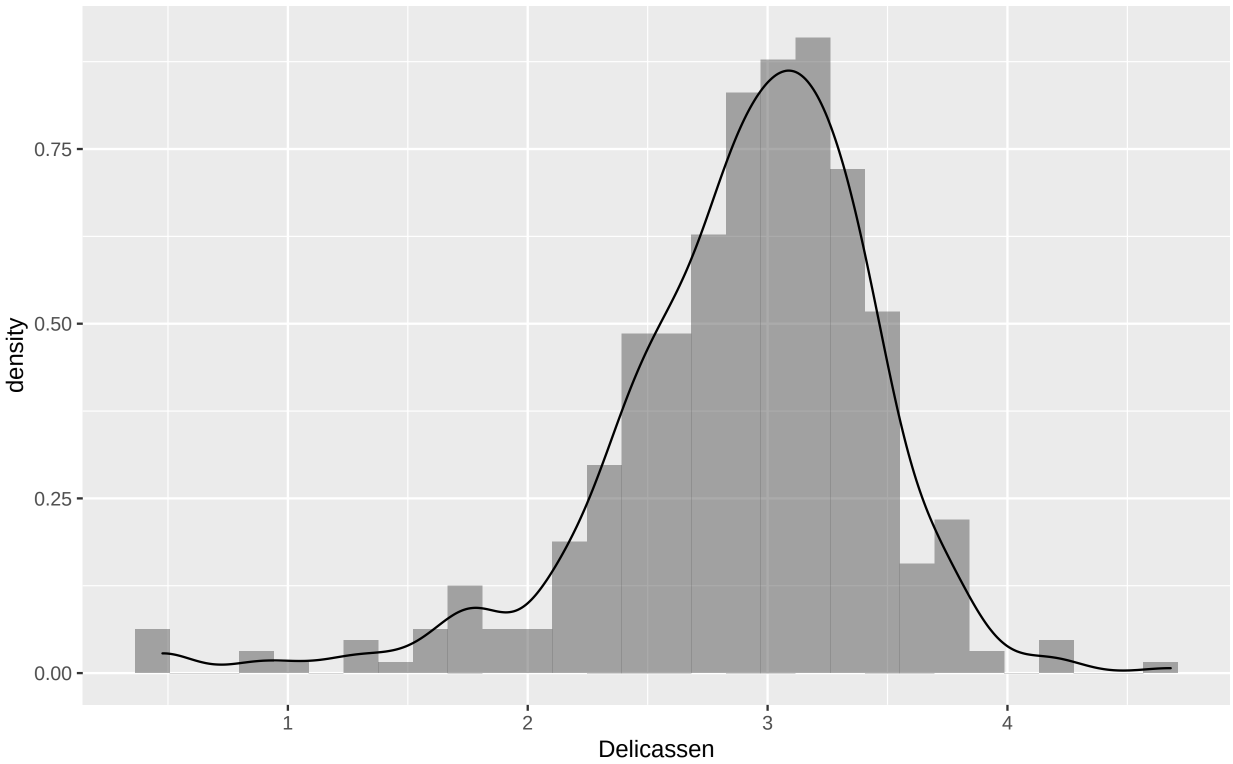
(A) True.

(B) False.

The answer is (A).

6. Assume the log-spending on delicatessen products follows a normal distribution with mean to be 2.89 and the standard deviation to be 0.57. The density plot below supports our assumption.

```
wholesale_log %>% ggplot(aes(x=Delicassen)) +
  geom_histogram(aes(y=..density..), alpha=.5) +
  geom_density()
```



Roughly 95% of the clients' log-spending on delicatessen products falls on the interval:

(A) $[2.89 - 2.89, 2.89 + 2.89]$

(B) $[2.89 - 2 \times 2.89, 2.89 + 2 \times 2.89]$

The answer is (B).

7-9. We perform principal component analysis among all the annual spending on the log-scale.

```
pca <- wholesale_log %>% select(Fresh:Delicassen) %>% prcomp(scale=T)
```

Here is some available information from PCA.

```
pca$rotation
```

```
                  PC1     PC2    PC3    PC4    PC5     PC6
  Fresh          -0.10  0.5905 -0.632  0.489 -0.041 -0.0274
  Milk            0.54  0.1331 -0.076 -0.061  0.762  0.3140
  Grocery         0.57 -0.0063 -0.133 -0.096 -0.098 -0.7978
  Frozen         -0.14  0.5895 -0.034 -0.792 -0.074  0.0059
  Detergents_Paper  0.55 -0.0686 -0.197 -0.077 -0.618  0.5139
  Delicassen      0.21  0.5304  0.733  0.340 -0.144  0.0022
```

```
summary(pca)$importance
```

```
                        PC1  PC2  PC3  PC4   PC5   PC6
  Standard deviation     1.62 1.28 0.80 0.78 0.543 0.429
  Proportion of Variance 0.44 0.27 0.11 0.10 0.049 0.031
  Cumulative Proportion  0.44 0.71 0.82 0.92 0.969 1.000
```

7. True or false? The first principal component is a weighted sum of all 6 variables.

(A) True

(B) False

The answer is (B).

8. The first three principal components explain the following amount of the total variation:

(A) 11%

(B) 54%

(C) 82%

The answer is (C).

9. The first principal component can be written as the following linear combination:

$$\texttt{PC1} = c_1 \times \texttt{Fresh} + c_2 \times \texttt{Milk} + c_3 \times \texttt{Grocery} + c_4 \times \texttt{Frozen} + c_5 \times \texttt{Detergents\_Paper} + c_6 \times \texttt{Delicassen}$$

Which is the correct value of c_5?

(A) 0.55

(B) -0.14

(C) -0.0686

The answer is (A).

Here we list the PC scores for the first 6 clients:

```
pca$x[1:6, ]
```

```
##         PC1   PC2    PC3    PC4   PC5    PC6
## [1,]   1.38 -0.30 -0.216  1.416  0.34  0.265
## [2,]   1.43  0.54  0.084 -0.033  0.11  0.184
## [3,]   1.50  1.22  0.977  0.148 -0.14  0.335
## [4,]  -0.82  1.20  0.245 -0.345 -0.72 -0.406
## [5,]   0.80  1.75  0.311  0.223 -0.25  0.015
## [6,]   0.88  0.13  0.029  0.700  0.34  0.389
```

10. Which client has the largest loading on PC2 in magnitude (i.e., the absolute value) among the `Client` 1-6? (We took this question out of the quiz)

(A) `Client` 1

(B) `Client` 3

(C) `Client` 5

The answer is (C)