Principal Component Analysis

DANA-4830 Assignment-2

Hyeri Goh
100349818

1. Data Screening

a. Accuracy Check

Firstly, 2 columns – 'Item Identifier' and 'Outlet Establishment Year'- out of total 11 columns were dropped because it was turned out these 2 columns did not provide important information after all analysis conducted.

In addition to removing unnecessary columns, on this step, there were 2 columns including inaccurate data points detected, which are 'Fat Content' and 'Outlet Size'. In the case of former, 'LF' and 'low fat' were combined to label of 'Low Fat'. (Figure 1) Similarly, 'reg' was added to 'Regular'. On the other hand, outlet size column had 2,410 blank cells. Since they took up over quarter of entire observations, which are 8,523 cases, recoding 2,410 blank cells into NA value will result in serious data missing issue. In order to keep all data points, blank cells were coded as 'Unknown'. (Figure 2)

```
     LF low fat Low Fat     reg Regular
    316     112    5089     117    2889
> data$Item_Fat_Content[data$Item_Fat_Content=='LF']='Low Fat'
> data$Item_Fat_Content[data$Item_Fat_Content=='low fat']='Low Fat'
> data$Item_Fat_Content[data$Item_Fat_Content=='reg']='Regular'
> table(data$Item_Fat_Content)

Low Fat Regular
   5517    3006
```

Figure 1

```
        High Medium  Small
  2410    932   2793   2388
> data$Outlet_Size[data$Outlet_Size='']='Unknown'
> table(data$Outlet_Size)

   High  Medium   Small Unknown
    932    2793    2388    2410
```

Figure 2

b. Missing Data

There were 1,463 blank cells were found in the column named 'Item Weight'. All blank cells were replaced with median of the column which was 12.60 as requested from clients.

c. Outlier

On this step, the categorical variables were not assigned for outlier test since the data of outlet size, outlet type, outlet location type, item type, and fat content cannot be dropped because of its importance. After examining mahalanobis distance score of the numerical variables – item weight, item visibility, item MRP, item outlet sales – the cutoff point was about 22.46 and total 24 outliers were detected and removed.

d. Multicollinearity

In order to execute VIF testing, regression model was created with item outlet sales as Y variable. The result showed outlet size had the most serious multicollinearity issue among all predictors. (Figure 3)

```
> vif(model)
                                      GVIF Df GVIF^(1/(2*Df))
nooutlier$Item_Weight             1.014906  1        1.007425
factor(nooutlier$Item_Fat_Content) 1.216468  1        1.102936
nooutlier$Item_Visibility         1.093816  1        1.045857
factor(nooutlier$Item_Type)       1.262402 15        1.007797
nooutlier$Item_MRP                1.013272  1        1.006614
factor(nooutlier$Outlet_Size)    32.509767  3        1.786497
factor(nooutlier$Outlet_Location_Type) 27.002663  2        2.279563
factor(nooutlier$Outlet_Type)    23.309347  3        1.690136
```

Figure 3

After recalculating the VIF with outlet size excluded, the multicollinearity issue reported from the previous model was solved. (Figure 4). It means outlet size was in the correlation with both outlet location type and outlet type.

```
> vif(model.new)
                                      GVIF Df GVIF^(1/(2*Df))
nooutlier$Item_Weight             1.014813  1        1.007379
factor(nooutlier$Item_Fat_Content) 1.216355  1        1.102885
nooutlier$Item_Visibility         1.093460  1        1.045686
factor(nooutlier$Item_Type)       1.259896 15        1.007731
nooutlier$Item_MRP                1.013071  1        1.006514
factor(nooutlier$Outlet_Type)     2.271088  3        1.146495
factor(nooutlier$Outlet_Location_Type) 2.105862  2        1.204641
```

Figure 4

2. PCA Analysis

In order to select the components, princomp() function from package stats was used. According to the result, a cumulative proportion when 4 components were selected was 62.6%, and the one when 5 components were selected was 73%.

Also, the components with eigen values that are equal to or greater than 1 were 'components 1', 'components 2' and 'components 3', while eigen value of 'components4' and 'components 5' were very closed to 1. (Figure 5)

```
eigen() decomposition
$values
[1] 2.0567511 1.4258083 1.1647796 0.9920822 0.9697477 0.8573287 0.7073079 0.5589651 0.2672294
```

Figure 5

Adding to the eigen values, biplot and screeplot are another indicators to determine how many components should be selected and which variables should be grouped under which components. According to the plots, I could roughly guess 4 components should be selected

since slope of line graph is getting mild. (Figure 7) Furthermore, biplot shows that certain variables are clustered heading to same direction. (Figure 6)

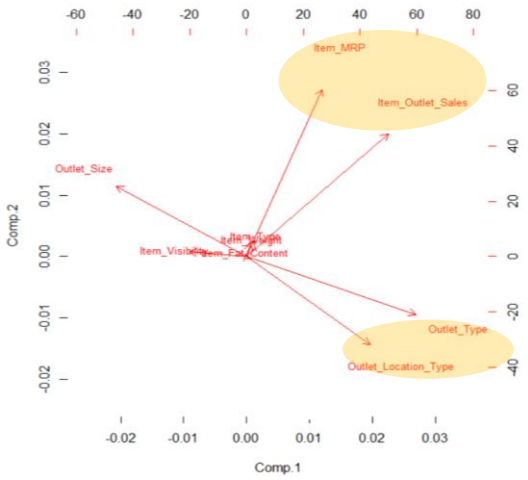

Figure 6

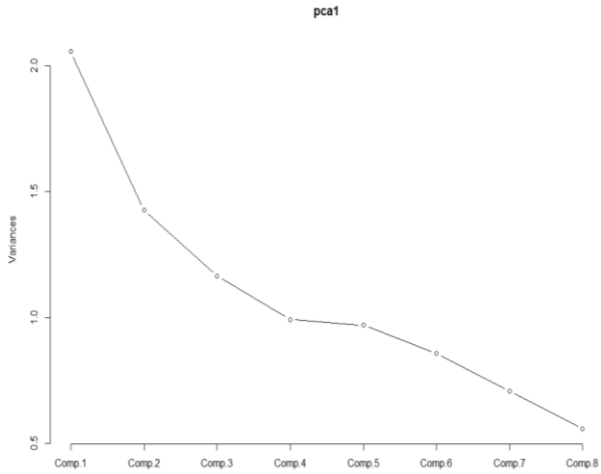

Figure 7

Considering eigen values and the result of plots, it seems to have 4 components should be selected. However, given that 4 components capture only 62.66% of total variance, it is better to include 5th components. Even though capturing 73.4% of total variance are still not enough and 6th component give information over 80% of total, I decided to use only 5 components because 6th component's eigen value is seriously low. (Figure 5, 8) Thus, using 5 components are the best middle point after trade-off between the eigen values and amount of captured variance out of total variance.

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5     Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     1.4341377 1.1940721 1.079250 0.9960332 0.9847577 0.92592048 0.84101601 0.74763969 0.51694235
Proportion of Variance 0.2285279 0.1584231 0.129420 0.1102314 0.1077497 0.09525875 0.07858977 0.06210723 0.02969216
Cumulative Proportion  0.2285279 0.3869510 0.516371 0.6266023 0.7343521 0.82961084 0.90820061 0.97030784 1.00000000
```

Figure 8

After considering all indicators such as loadings, eigen values, plots and cumulative proportion, the result of PCA will be as following (Figure 9):

- 1st component: Outlet_Location_Type, Outlet_Type, Outlet_Size
- 2nd component: Item_MRP, Item_Outlet_Sales
- 3rd component: Item_Fat_Content, Item_Type
- 4th component: Item_Weight
- 5th component: Item_Visibility

The PCA resulted in grouping total 9 variables into 5 components and 2 components - 4th and 5th components – having only one variable, respectively. For 1st component, it can be summarized as outlet feature, for 2nd component as profit, and for 3rd component as item feature.

4th and 5th component do not have grouped variables. The loading shows item weight was loaded to component solely. In the case of item visibility, even though item MRP was assigned in the same component with item visibility, it gives more similar information to item outlet sales. Thus, item visibility was left alone. Since both item weight and item visibility have high loadings, which are 0.970 and -0.885 respectively, deleting these 2 variables are not recommendable. For those reasons, the only way to keep 2 variables was assigning each of them to one component rather than grouping variables forcibly.

```
> print(pca1$loadings, cutoff = 0.3)

Loadings:
                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
Item_Weight                              0.970
Item_Fat_Content          -0.669                -0.712
Item_Visibility                        -0.885
Item_Type                  0.645        -0.305 -0.689
Item_MRP           0.684                                          -0.443  0.492
Outlet_Size       -0.434                                   0.756  0.379
Outlet_Location_Type  0.411 -0.361                        0.631 -0.406
Outlet_Type        0.564                                          0.568  0.544
Item_Outlet_Sales  0.472  0.504                                   0.347 -0.630

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111
Cumulative Var  0.111  0.222  0.333  0.444  0.556  0.667  0.778  0.889  1.000
```

Figure 9

3. Summary

The purpose of PCA is reduction of dimensions. Dimensions can be reduced by finding out variables that are providing similar information and grouping them into one component, which are useful to build smaller model with certain degree of total variables kept and possibly reduce overfitting issue.

In this dataset which originally had 11 variables, by conducting PCA, it ended up being shrunken down to 5 components and still captured 73.4% of total variance. Otherwise, it will cost more of effort and time to find relations among variables within such a huge dimension. Thus, conducting PCA on this dataset suggested the best point where both time/cost efficiency and the quality of analysis settle for.