Assignment 6 and 7

Atharv Prashant Tungatkar

2025-10-20

Reading the dataset

```
data<-read.csv("C:/Users/athar/OneDrive/Desktop/MBA Business Analytics/Visual Analytics/Files/siva.csv")
head(data,10)
```

```
##    xgra_n1clb_nbr Siva_Rental_Number rent_area_loc Date_of_Survey Day_of_Week
## 1           51407              67041           156      5/18/2011   Wednesday
## 2           23460              56084           204       2/5/2011    Saturday
## 3           53417              70279           181      6/14/2011     Tuesday
## 4           14382              15105          1515       1/4/2010      Monday
## 5           40539              49797           259      12/1/2010   Wednesday
## 6           53945              71102           165      6/29/2011   Wednesday
## 7           35983              43104           177      9/23/2010    Thursday
## 8           43669              54673           276      1/23/2011      Sunday
## 9           29279              33940          2167       7/6/2010     Tuesday
## 10          14254              14950           953       1/7/2010    Thursday
##        Time Survey_Type Purpose_of_Rental Recom_mend_Siva Staff_Courtesy
## 1   7:48:30 SV Web Sol.              Bus.               8              9
## 2  22:06:37 SV Web Sol.     Leis. / Pers.               8              8
## 3   5:35:48 SV Web Sol.              Bus.               8              7
## 4  23:58:56 SV Web Sol.     Leis. / Pers.               7              8
## 5   8:24:39 SV Web Sol.     Leis. / Pers.               9              9
## 6   5:34:06 SV Web Sol.              Bus.               9              9
## 7   6:24:40 SV Web Sol.              Bus.               9              9
## 8   9:48:08 SV Web Sol.              Bus.               6              9
## 9  14:38:54 SV Web Sol.              Bus.               9              8
## 10  5:49:29 SV Web Sol.              Bus.               5              7
##    Speed_of_Service Veh_Equip_Condition Trans_Billing_as_Expected
## 1                 8                   9                         9
## 2                 8                   5                         8
## 3                 8                   8                         8
## 4                 7                   8                         8
## 5                 9                   9                         9
## 6                 9                   9                         9
## 7                 9                   9                         9
## 8                 8                   9                         9
## 9                 9                   8                         9
## 10                5                   8                         7
##    Value_for_the_Money                                Area
## 1                    9            01602 - LOVE FIELD AP TX
## 2                    7              07286 - VALLEJO CA OAP
## 3                    8              01850 - RICHMOND VA AP
## 4                    8                 05743 - PICO CA OAP
## 5                    9              07787 - NEWPORT RI OAP
## 6                    9               01450 - ATLANTA AP GA
## 7                    9       05426 - ELIZABETHTOWN: KY OAP
## 8                    6        02170 - SALT LAKE CITY UT AP
## 9                    7 07275 - WALSH RD. SANTA CLARA CA OAP
## 10                   7              07781 - BEDFORD MA OAP
##                     loc_nm      ga_region_desc xgra_ckot_ts xgra_ckin_ts
## 1        DALLAS LOVE FIELD    SOUTHWEST REGION    5/15/2011    5/17/2011
## 2              VALLEJO HLE      WESTERN REGION    1/31/2011     2/3/2011
## 3         RICHMOND INTL AP MID ATLANTIC REGION    6/12/2011    6/13/2011
## 4                 PICO HLE      WESTERN REGION   12/23/2009     1/3/2010
## 5               NEWPORT HLE    NORTHEAST REGION   11/29/2010   11/30/2010
## 6  ATLANTA-HARTSFIELD INTL    SOUTHEAST REGION    6/20/2011    6/26/2011
## 7         ELIZABETHTOWN HLE      CENTRAL REGION    9/19/2010    9/22/2010
## 8   SALT LAKE CITY INTL AP WEST CENTRAL REGION    1/17/2011    1/22/2011
## 9     INTEL         (SC12)      WESTERN REGION    6/30/2010     7/1/2010
## 10            JET AVIATION    NORTHEAST REGION     1/3/2010     1/6/2010
```

```
##    xgra_vclass_reserv xgra_veh_class rent_loc_type cust_tier_code
## 1                   C             Q4            AP             FG
## 2                   A              B        OFF AP             N1
## 3                   F             YF            AP             RG
## 4                   D             YF        OFF AP             RG
## 5                   F              A        OFF AP             RG
## 6                   R             YR            AP             RG
## 7                  Q4             Q4        OFF AP             N1
## 8                   F              T            AP             N1
## 9                   B              C        OFF AP             RG
## 10                  D              F        OFF AP             RG
##    booking_channel_code col34_total_charges col38_currency Total_charge_USD
## 1              SIVA.COM               247.29            USD           247.29
## 2              SIVA.COM               128.04            USD           128.04
## 3              SIVA.COM                75.85            USD            75.85
## 4              SIVA.COM               468.51            USD           468.51
## 5              SIVA.COM                42.84            USD            42.84
## 6                  800#               107.92            USD           107.92
## 7             LOCAL RES               224.68            USD           224.68
## 8                  GDSB               387.63            USD           387.63
## 9              SIVA.COM                45.08            USD            45.08
## 10             SIVA.COM               229.10            USD           229.10
##    Survey_checkout_diff booking_channel_dummy
## 1                     2                     1
## 2                     3                     1
## 3                     2                     1
## 4                     2                     1
## 5                     2                     1
## 6                     4                     0
## 7                     2                     0
## 8                     2                     0
## 9                     6                     1
## 10                    2                     1
```

Checking the data types of variables

```
str(data)
```

```
## 'data.frame':    53815 obs. of  29 variables:
##  $ xgra_n1clb_nbr           : int  51407 23460 53417 14382 40539 53945 35983 43669 29279 1
4254 ...
##  $ Siva_Rental_Number       : int  67041 56084 70279 15105 49797 71102 43104 54673 33940 1
4950 ...
##  $ rent_area_loc            : int  156 204 181 1515 259 165 177 276 2167 953 ...
##  $ Date_of_Survey           : chr  "5/18/2011" "2/5/2011" "6/14/2011" "1/4/2010" ...
##  $ Day_of_Week              : chr  "Wednesday" "Saturday" "Tuesday" "Monday" ...
##  $ Time                     : chr  "7:48:30" "22:06:37" "5:35:48" "23:58:56" ...
##  $ Survey_Type              : chr  "SV Web Sol." "SV Web Sol." "SV Web Sol." "SV Web Sol."
...
##  $ Purpose_of_Rental        : chr  "Bus." "Leis. / Pers." "Bus." "Leis. / Pers." ...
##  $ Recom_mend_Siva          : int  8 8 8 7 9 9 9 6 9 5 ...
##  $ Staff_Courtesy           : int  9 8 7 8 9 9 9 9 8 7 ...
##  $ Speed_of_Service         : int  8 8 8 7 9 9 9 8 9 5 ...
##  $ Veh_Equip_Condition      : int  9 5 8 8 9 9 9 9 8 8 ...
##  $ Trans_Billing_as_Expected: int  9 8 8 8 9 9 9 9 9 7 ...
##  $ Value_for_the_Money      : int  9 7 8 8 9 9 9 6 7 7 ...
##  $ Area                     : chr  "01602 - LOVE FIELD AP TX" "07286 - VALLEJO CA OAP" "01
850 - RICHMOND VA AP" "05743 - PICO CA OAP" ...
##  $ loc_nm                   : chr  "DALLAS LOVE FIELD" "VALLEJO HLE" "RICHMOND INTL AP" "P
ICO HLE" ...
##  $ ga_region_desc           : chr  "SOUTHWEST REGION" "WESTERN REGION" "MID ATLANTIC REGIO
N" "WESTERN REGION" ...
##  $ xgra_ckot_ts             : chr  "5/15/2011" "1/31/2011" "6/12/2011" "12/23/2009" ...
##  $ xgra_ckin_ts             : chr  "5/17/2011" "2/3/2011" "6/13/2011" "1/3/2010" ...
##  $ xgra_vclass_reserv       : chr  "C" "A" "F" "D" ...
##  $ xgra_veh_class           : chr  "Q4" "B" "YF" "YF" ...
##  $ rent_loc_type            : chr  "AP" "OFF AP" "AP" "OFF AP" ...
##  $ cust_tier_code           : chr  "FG" "N1" "RG" "RG" ...
##  $ booking_channel_code     : chr  "SIVA.COM" "SIVA.COM" "SIVA.COM" "SIVA.COM" ...
##  $ col34_total_charges      : num  247.3 128 75.8 468.5 42.8 ...
##  $ col38_currency           : chr  "USD" "USD" "USD" "USD" ...
##  $ Total_charge_USD         : num  247.3 128 75.8 468.5 42.8 ...
##  $ Survey_checkout_diff     : int  2 3 2 2 2 4 2 2 6 2 ...
##  $ booking_channel_dummy    : int  1 1 1 1 1 0 0 0 1 1 ...
```

For regression we only use the numerical variables.

Subsetting the data

```
reg_data<-data[,c('Recom_mend_Siva','Staff_Courtesy','Speed_of_Service','Veh_Equip_Conditio
n','Trans_Billing_as_Expected','Value_for_the_Money')]
head(reg_data,10)
```

```
##    Recom_mend_Siva Staff_Courtesy Speed_of_Service Veh_Equip_Condition
## 1                8              9                8                   9
## 2                8              8                8                   5
## 3                8              7                8                   8
## 4                7              8                7                   8
## 5                9              9                9                   9
## 6                9              9                9                   9
## 7                9              9                9                   9
## 8                6              9                8                   9
## 9                9              8                9                   8
## 10               5              7                5                   8
##    Trans_Billing_as_Expected Value_for_the_Money
## 1                          9                   9
## 2                          8                   7
## 3                          8                   8
## 4                          8                   8
## 5                          9                   9
## 6                          9                   9
## 7                          9                   9
## 8                          9                   6
## 9                          9                   7
## 10                         7                   7
```

Check for null values

```
sum(is.na(reg_data))
```

```
## [1] 7300
```

```
sapply(reg_data, function(x) sum(is.na(x)))
```

```
##           Recom_mend_Siva            Staff_Courtesy         Speed_of_Service
##                         0                      1453                      1455
##       Veh_Equip_Condition Trans_Billing_as_Expected       Value_for_the_Money
##                      1456                      1467                      1469
```

There are 7300 null values hence we clear the rows and then get the correlation matrix.

```
df<-na.omit(reg_data)
corr_matrix<-cor(df)
```

Plot the correlaton visualization

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

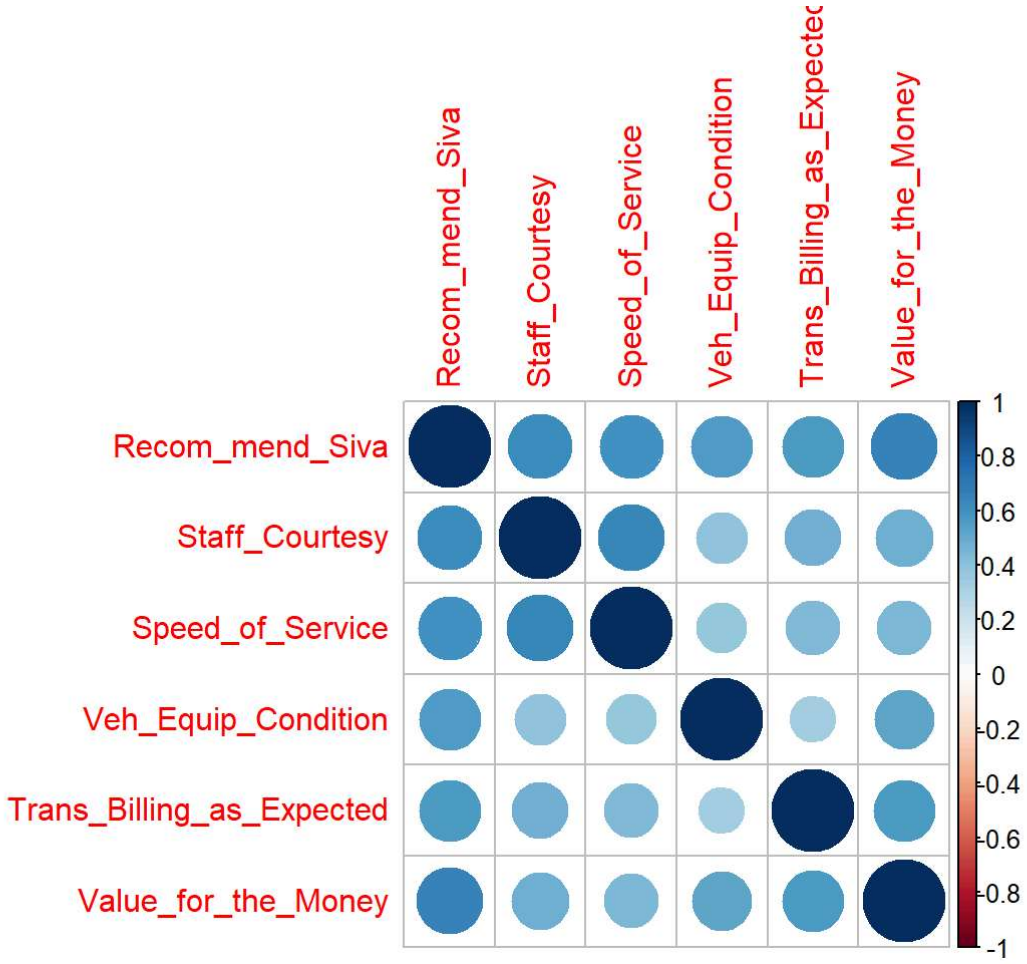
```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.4.3
```

```
## corrplot 0.95 loaded
```

```
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA" ))

corrplot(corr_matrix)
```



Creating the regression model

```
model<-lm(Recom_mend_Siva ~ .,data=df)
summary(model)
```

```
##
## Call:
## lm(formula = Recom_mend_Siva ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0267 -0.5159  0.1299  0.6015  9.1416
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.409053   0.029629  -13.81   <2e-16 ***
## Staff_Courtesy             0.248569   0.004678   53.13   <2e-16 ***
## Speed_of_Service           0.195483   0.003446   56.72   <2e-16 ***
## Veh_Equip_Condition        0.180413   0.002935   61.48   <2e-16 ***
## Trans_Billing_as_Expected  0.156543   0.003327   47.05   <2e-16 ***
## Value_for_the_Money        0.267408   0.003629   73.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.28 on 52340 degrees of freedom
## Multiple R-squared:  0.6247, Adjusted R-squared:  0.6247
## F-statistic: 1.743e+04 on 5 and 52340 DF,  p-value: < 2.2e-16
```

Plotting beta coefficients and their distributions

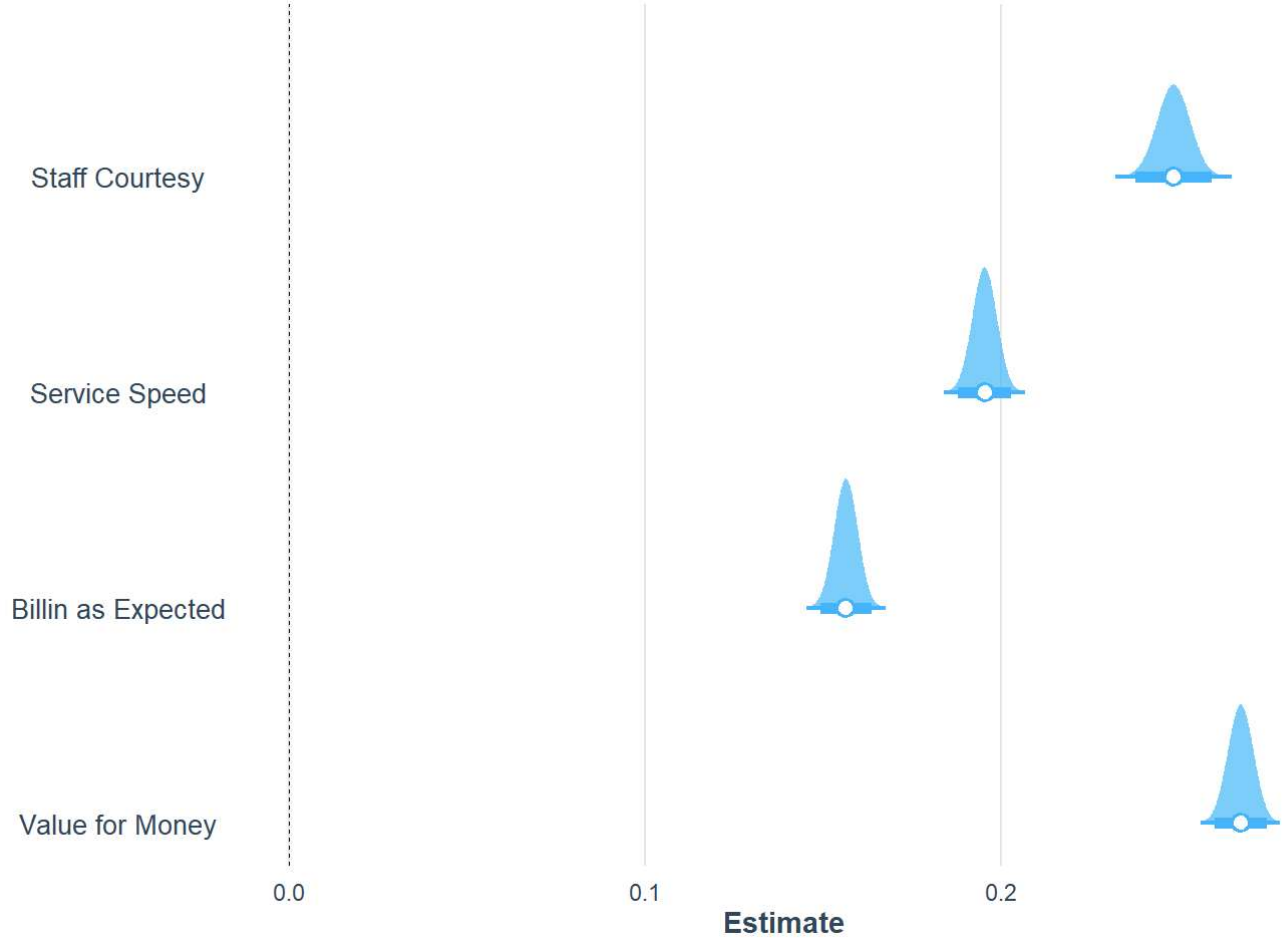
```
library(jtools)
```

```
## Warning: package 'jtools' was built under R version 4.4.3
```

```
library(broom.mixed)
```

```
## Warning: package 'broom.mixed' was built under R version 4.4.3
```

```
plot_summs(model,
           coefs = c("Staff Courtesy" = "Staff_Courtesy", "Service Speed" = "Speed_of_Servic
e","Equipment Condition" = "Veh_Equip_Conditio","Billin as Expected"='Trans_Billing_as_Expect
ed','Value for Money"='Value_for_the_Money' ),
           robust = TRUE, plot.distributions = TRUE, inner_ci_level = .8)
```



Creating the clustering

```
kmeans.result <- kmeans(df, 4)
```

Plotting the clusters

```
library(ggpubr)
```

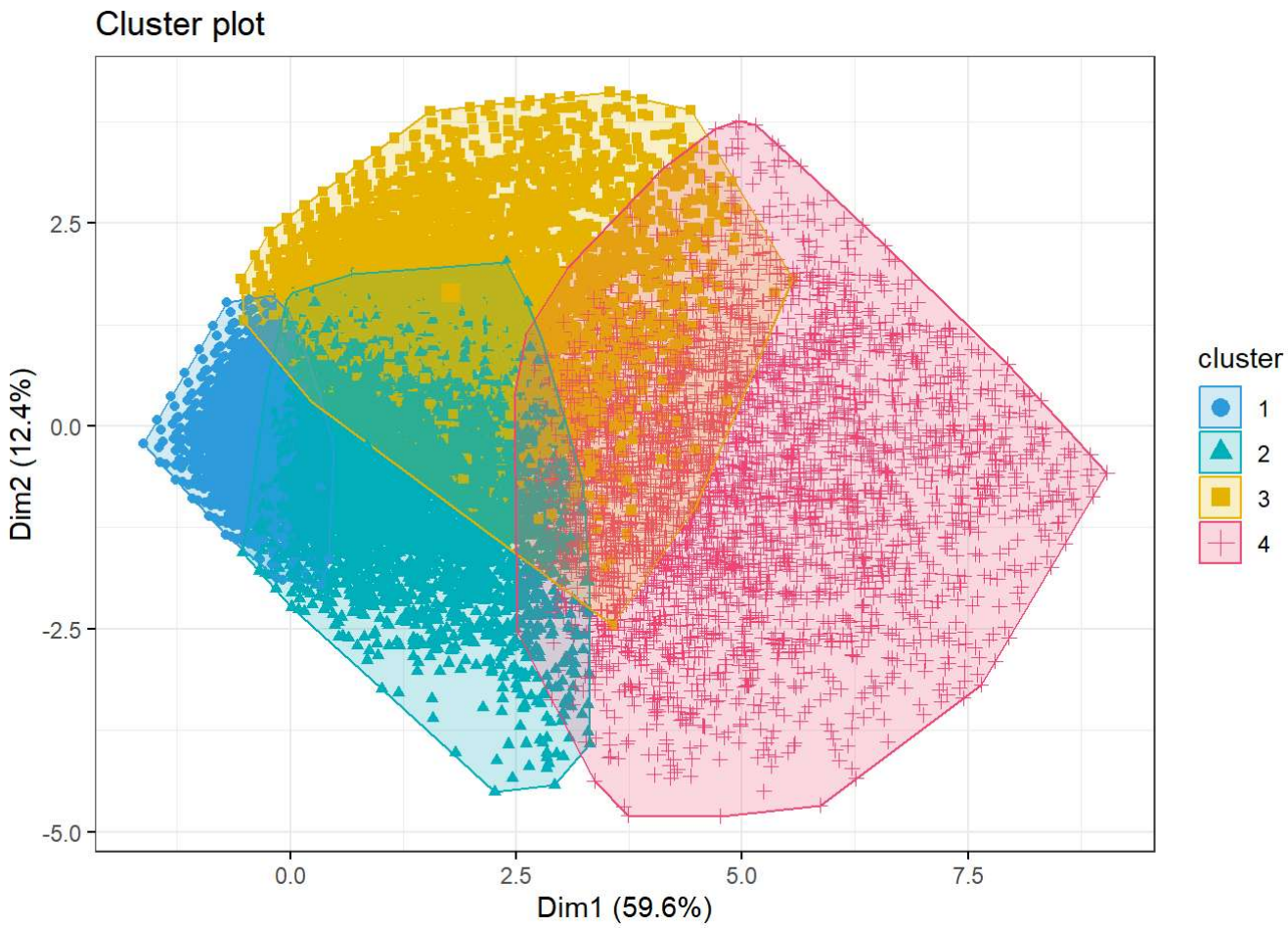
```
## Warning: package 'ggpubr' was built under R version 4.4.3
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.4.3
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_cluster(kmeans.result, data = df,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800","#f14975"),
             geom = "point",
             ellipse.type = "convex",
             ggtheme = theme_bw()
)
```


Cluster plot
2.5
0.0
-2.5
-5.0
Dim2 (12.4%)
0.0
2.5
5.0
7.5
Dim1 (59.6%)
cluster
1
2
3
4