Individual Visual Analytics Assignment 5
Anshita Thakkar

First, I read in the data:

**siva = read.csv("/Users/anshitathakkar/Documents/Visual analytics/Week 7 - Hypothesis testing /Assignment 5/siva.csv")**
**siva**

Then, I structured the data:

**str(siva)**

```
'data.frame':   53815 obs. of  29 variables:
 $ xgra_n1clb_nbr          : int  51407 23460 53417 14382 40539 53945 35983 43669 29279 14254 ...
 $ Siva_Rental_Number      : int  67041 56084 70279 15105 49797 71102 43104 54673 33940 14950 ...
 $ 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" "01850 - RICHMOND VA AP" "05743 - PICO CA OAP" ...
 $ loc_nm                  : chr  "DALLAS LOVE FIELD" "VALLEJO HLE" "RICHMOND INTL AP" "PICO HLE" ...
 $ ga_region_desc          : chr  "SOUTHWEST REGION" "WESTERN REGION" "MID ATLANTIC REGION" "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 ...
```

**head(siva)**
**tail(siva)**

```{r}
head(siva)
```

Description: df [6 × 29]

| | xgra_n1clb_nbr <int> | Siva_Rental_Number <int> | rent_area_loc <int> | Date_of_Survey <chr> | Day_of_Week <chr> | Time <chr> | Survey_Type <chr> | Purpose_of_Rental <chr> | Recom_mend_Siva <int> |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 51407 | 67041 | 156 | 5/18/2011 | Wednesday | 7:48:30 | SV Web Sol. | Bus. | 8 |
| 2 | 23460 | 56084 | 204 | 2/5/2011 | Saturday | 22:06:37 | SV Web Sol. | Leis. / Pers. | 8 |
| 3 | 53417 | 70279 | 181 | 6/14/2011 | Tuesday | 5:35:48 | SV Web Sol. | Bus. | 8 |
| 4 | 14382 | 15105 | 1515 | 1/4/2010 | Monday | 23:58:56 | SV Web Sol. | Leis. / Pers. | 7 |
| 5 | 40539 | 49797 | 259 | 12/1/2010 | Wednesday | 8:24:39 | SV Web Sol. | Leis. / Pers. | 9 |
| 6 | 53945 | 71102 | 165 | 6/29/2011 | Wednesday | 5:34:06 | SV Web Sol. | Bus. | 9 |

6 rows | 1-10 of 29 columns

```{r}
tail(siva)
```

Description: df [6 × 29]

| | xgra_n1clb_nbr <int> | Siva_Rental_Number <int> | rent_area_loc <int> | Date_of_Survey <chr> | Day_of_Week <chr> | Time <chr> | Survey_Type <chr> | Purpose_of_Rental <chr> | Recom_mend_Siva <int> |
|---|---|---|---|---|---|---|---|---|---|
| 53810 | 878 | 880 | 138 | 8/13/2009 | Thursday | 6:54:31 | SV Phone | Leis. / Pers. | 9 |
| 53811 | 128 | 128 | 39 | 8/16/2009 | Sunday | 14:10:37 | SV Web Sol. | Leis. / Pers. | 9 |
| 53812 | 9644 | 9867 | 1111 | 11/30/2009 | Monday | 15:17:59 | SV Web Sol. | Leis. / Pers. | 9 |
| 53813 | 18222 | 19641 | 161 | 2/18/2010 | Thursday | 7:38:18 | SV Web Sol. | Bus. | 7 |
| 53814 | 6202 | 78953 | 537 | 10/1/2011 | Saturday | 15:26:14 | SV Web Sol. | Leis. / Pers. | 8 |
| 53815 | 17117 | 45840 | 141 | 11/9/2010 | Tuesday | 5:02:37 | SV Web Sol. | Bus. | 9 |

Then I changed the variables to numeric, and selected columns 9:14

```r
as.numeric(siva$Recom_mend_Siva)
as.numeric(siva$Staff_Courtesy)
as.numeric(siva$Speed_of_Service)
as.numeric(siva$Veh_Equip_Condition)
as.numeric(siva$Trans_Billing_as_Expected)
as.numeric(siva$Value_for_the_Money)

siva_new_numeric <- siva[,c(9:14)]
siva_new_numeric
```

Description: df [53,815 × 6]

| Recom_mend_Siva <int> | Staff_Courtesy <int> | Speed_of_Service <int> | Veh_Equip_Condition <int> | Trans_Billing_as_Expected <int> | Value_for_the_Money <int> |
|---|---|---|---|---|---|
| 8 | 9 | 8 | 9 | 9 | 9 |
| 8 | 8 | 8 | 5 | 8 | 7 |
| 8 | 7 | 8 | 8 | 8 | 8 |
| 7 | 8 | 7 | 8 | 8 | 8 |
| 9 | 9 | 9 | 9 | 9 | 9 |
| 9 | 9 | 9 | 9 | 9 | 9 |
| 9 | 9 | 9 | 9 | 9 | 9 |
| 6 | 9 | 8 | 9 | 9 | 6 |
| 9 | 8 | 9 | 8 | 9 | 7 |
| 5 | 7 | 5 | 8 | 7 | 7 |

1–10 of 53,815 rows

Previous 1 2 3 4 5 6 … 100 Next

Using code, I omitted the missing variables, and checked the change from the siva to the **siva_new** dataset. As you can see below, there are missing variables in the siva dataset, the **siva_new** dataset has no missing variables as they have been omitted.

```r
siva_new <- na.omit(siva_new_numeric)
siva_new
```

```r
colSums(is.na(siva))
```

```
          xgra_n1clb_nbr          Siva_Rental_Number                rent_area_loc 
                       0                           0                           0 
          Date_of_Survey                 Day_of_Week                        Time 
                       0                           0                           0 
             Survey_Type           Purpose_of_Rental           Recom_mend_Siva 
                       0                           0                           0 
          Staff_Courtesy            Speed_of_Service        Veh_Equip_Condition 
                    1453                        1455                        1456 
Trans_Billing_as_Expected        Value_for_the_Money                        Area 
                    1467                        1469                           0 
                  loc_nm                 ga_region_desc                xgra_ckot_ts 
                       0                           0                           0 
            xgra_ckin_ts           xgra_vclass_reserv                xgra_veh_class 
                       0                           0                           0 
           rent_loc_type                cust_tier_code        booking_channel_code 
                       0                         166                           0 
      col34_total_charges               col38_currency                Total_charge_USD 
                       0                           0                           0 
  Survey_checkout_diff        booking_channel_dummy 
                       0                           0 
```

```r
colSums(is.na(siva_new))
```

```
          Recom_mend_Siva           Staff_Courtesy            Speed_of_Service 
                        0                         0                         0 
       Veh_Equip_Condition Trans_Billing_as_Expected        Value_for_the_Money 
                        0                         0                         0 
```

**CORRELATION**

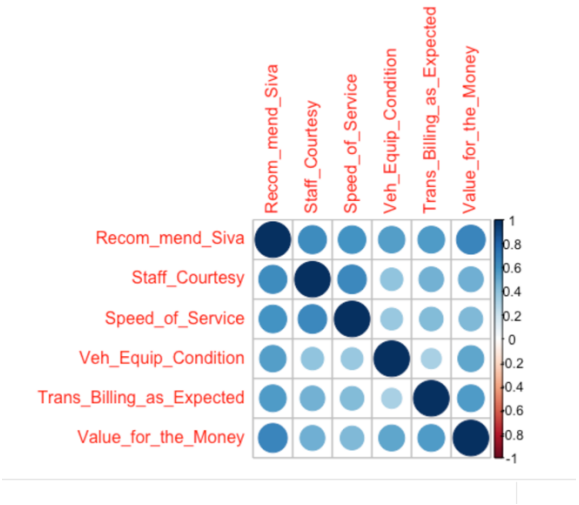
I did a correlation test initially to check the relationship between two variables. The two variables are weakly correlated at r = 0.02988985
cor.test(siva$Recom_mend_Siva, siva$rent_area_loc )

```
        Pearson's product-moment correlation

data:  siva$Recom_mend_Siva and siva$rent_area_loc
t = 6.9368, df = 53813, p-value = 4.055e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.02144641 0.03832904
sample estimates:
      cor 
0.02988985 
```

I thus decided to use variables 9-14 alone as independent variables (Staff Courtesy, speed of the service, vehicle equipment condition, trans billing as expected, and value for the money) and the dependent variable is Recom_mend_Siva.

First, I created a correlation matrix and then a correlation plot of variables 9 to 14 (as specified above)

```
siva_cor <- cor(siva_new)
corrplot(siva_cor)
```




The correlation plot shows how strongly correlated the variables are with each other and with the dependent variable, “Recom_ment_Siva”. The scale on the side indicates the strength of the correlation – all are positively correlated though at varying strengths. To see the strength of the correlation with numbers, I added the co-efficients in the correlation plot below:

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

corrplot(siva_cor,
        method = "color",
        col = col(200),
        type = "lower",
        order = "hclust",
        addCoef.col = "black",
        tl.col = "black",
        tl.srt = 45,
        diag = F)
```


This correlation plot is useful as it shows in numbers the correlation between variables. Value_for_the_money and Recom_mend_siva are two variables that are most strongly correlated at r = 0.66, as opposed to the other variables.

**REGRESSION MODEL**

It is best to use numerical values in a regression module but we can also use categorical variables

siva_model <- lm(Recom_mend_Siva ~ Staff_Courtesy + Speed_of_Service + Veh_Equip_Condition + Total_charge_USD + xgra_vclass_reserv + Day_of_Week, data = siva)

The dependent variables is Recommend Siva, and the other variables are independent.

```
Call:
lm(formula = Recom_mend_Siva ~ Staff_Courtesy + Speed_of_Service +
    Veh_Equip_Condition + Total_charge_USD + xgra_vclass_reserv +
    Day_of_Week, data = siva)

Coefficients:
                (Intercept)               Staff_Courtesy             Speed_of_Service
                  -1.180e-01                    3.931e-01                    2.614e-01
        Veh_Equip_Condition              Total_charge_USD  xgra_vclass_reserv99
                  2.918e-01                   -9.928e-05                    2.012e-01
 xgra_vclass_reservA  xgra_vclass_reservA4  xgra_vclass_reservB  xgra_vclass_reservC
                  3.279e-01                   -4.386e-02                    3.432e-01
xgra_vclass_reservC4  xgra_vclass_reservD  xgra_vclass_reservD4  xgra_vclass_reservE4
                  3.487e-01                    1.812e-01                    4.232e-01
xgra_vclass_reservE6  xgra_vclass_reservF  xgra_vclass_reservF4  xgra_vclass_reservF6
                  3.367e-01                   -9.048e-01                    3.916e-01
 xgra_vclass_reservG  xgra_vclass_reservG4  xgra_vclass_reservH4  xgra_vclass_reservH6
                  3.983e-01                    7.888e-01                    3.350e-01
 xgra_vclass_reservI  xgra_vclass_reservI4  xgra_vclass_reservK  xgra_vclass_reservK4
                  3.366e-01                    4.948e-01                    1.558e-01
xgra_vclass_reservK6  xgra_vclass_reservL  xgra_vclass_reservL4  xgra_vclass_reservM
                  3.997e-01                    3.928e-01                    6.553e-01
xgra_vclass_reservN4  xgra_vclass_reservN6  xgra_vclass_reservP  xgra_vclass_reservP4
                  8.321e-01                   -3.890e-01                    8.040e-01
xgra_vclass_reservP6  xgra_vclass_reservQ  xgra_vclass_reservQ4  xgra_vclass_reservR
                  4.510e-01                    5.074e-01                    6.785e-01
 xgra_vclass_reservS  xgra_vclass_reservS4  xgra_vclass_reservT  xgra_vclass_reservT4
                  1.958e+00                    6.622e-01                    3.144e-01
xgra_vclass_reservT6  xgra_vclass_reservU  xgra_vclass_reservU4  xgra_vclass_reservU6
                  7.051e-01                    4.309e-01                    5.150e-01
xgra_vclass_reservV  xgra_vclass_reservV4  xgra_vclass_reservV6  xgra_vclass_reservW
                  7.081e-01                   -1.794e-01                    4.240e-01
xgra_vclass_reservW4  xgra_vclass_reservW6  xgra_vclass_reservX  xgra_vclass_reservX4
                  4.749e-01                    7.051e-01                    3.056e-01
xgra_vclass_reservX6  xgra_vclass_reservY6  xgra_vclass_reservYF  xgra_vclass_reservYQ
                  2.787e-01                    3.313e-01                    2.877e-01
xgra_vclass_reservZ    Day_of_WeekMonday   Day_of_WeekSaturday     Day_of_WeekSunday
                  2.026e-01                    1.007e+00                   -4.934e-01
  Day_of_WeekThursday    Day_of_WeekTuesday  Day_of_WeekWednesday
                  3.169e-01                   -4.766e-01                    5.008e-01
                 -1.660e-01                    1.225e+00                    1.552e+00
                  2.122e-01                    9.507e-01                    4.720e-01
                  6.238e-01                    1.184e+00                    6.131e-01
                  5.082e-02                    6.089e-02                    1.797e-02
                 -1.780e-02                    3.285e-02                   -8.633e-03
```

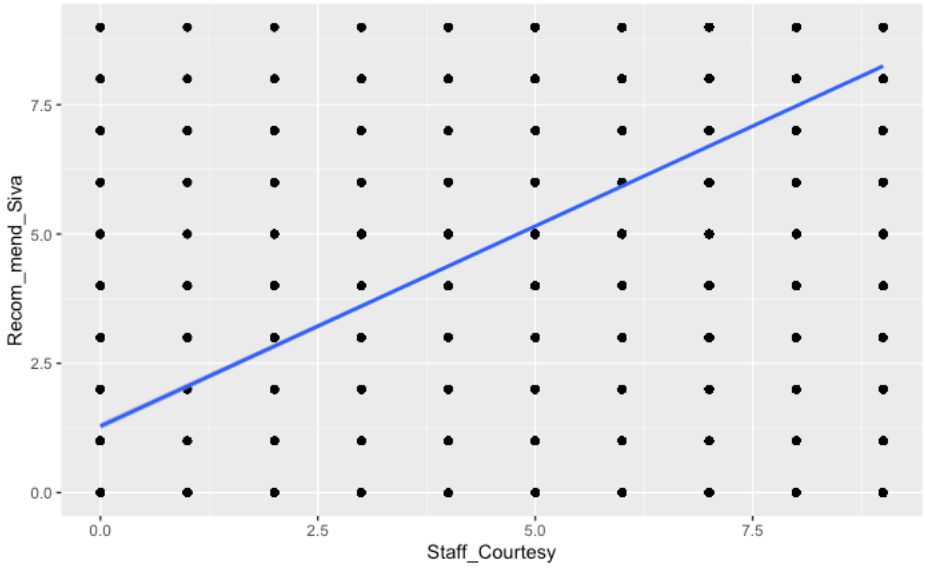
summary(siva_model)

```
Call:
lm(formula = Recom_mend_Siva ~ Staff_Courtesy + Speed_of_Service +
    Veh_Equip_Condition + Total_charge_USD + xgra_vclass_reserv +
    Day_of_Week, data = siva)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8641 -0.5287  0.2312  0.6234  8.6831

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.180e-01  2.253e-01  -0.524  0.60063
Staff_Courtesy           3.931e-01  4.978e-03  78.977  < 2e-16 ***
Speed_of_Service         2.614e-01  3.764e-03  69.451  < 2e-16 ***
Veh_Equip_Condition      2.918e-01  2.994e-03  97.466  < 2e-16 ***
Total_charge_USD        -9.928e-05  2.539e-05  -3.910 9.22e-05 ***
xgra_vclass_reserv99     2.012e-01  2.273e-01   0.885  0.37592
xgra_vclass_reservA      3.279e-01  2.235e-01   1.467  0.14240
xgra_vclass_reservA4    -4.386e-02  8.497e-01  -0.052  0.95884
xgra_vclass_reservB      3.432e-01  2.225e-01   1.542  0.12298
xgra_vclass_reservC      3.487e-01  2.222e-01   1.570  0.11647
xgra_vclass_reservC4     1.812e-01  3.652e-01   0.496  0.61972
xgra_vclass_reservD      4.232e-01  2.233e-01   1.895  0.05809 .
xgra_vclass_reservD4     3.367e-01  3.052e-01   1.103  0.26982
xgra_vclass_reservE4    -9.048e-01  8.497e-01  -1.065  0.28695
xgra_vclass_reservE6     3.916e-01  2.584e-01   1.515  0.12974
xgra_vclass_reservF      3.983e-01  2.222e-01   1.793  0.07305 .
xgra_vclass_reservF4     7.888e-01  2.817e-01   2.800  0.00511 **
xgra_vclass_reservF6     3.350e-01  2.360e-01   1.419  0.15578
xgra_vclass_reservG      3.366e-01  2.310e-01   1.457  0.14500
xgra_vclass_reservG4     4.948e-01  3.944e-01   1.254  0.20967
xgra_vclass_reservH4     1.558e-01  3.606e-01   0.432  0.66568
xgra_vclass_reservH6     3.997e-01  3.813e-01   1.048  0.29452
xgra_vclass_reservI      3.928e-01  2.378e-01   1.652  0.09863 .
xgra_vclass_reservI4     6.553e-01  7.442e-01   0.881  0.37859
xgra_vclass_reservK      8.321e-01  3.324e-01   2.503  0.01231 *
xgra_vclass_reservK4    -3.890e-01  3.524e-01  -1.104  0.26957
xgra_vclass_reservK6     8.040e-01  5.230e-01   1.537  0.12425
xgra_vclass_reservL      4.510e-01  2.255e-01   2.000  0.04551 *
xgra_vclass_reservQ4     4.240e-01  2.249e-01   1.885  0.05937 .
xgra_vclass_reservR      4.749e-01  2.255e-01   2.106  0.03525 *
xgra_vclass_reservS      7.051e-01  3.943e-01   1.788  0.07376 .
xgra_vclass_reservS4     4.309e-01  4.669e-01   0.923  0.35602
xgra_vclass_reservS6     3.056e-01  5.241e-01   0.583  0.55990
xgra_vclass_reservT      2.787e-01  2.345e-01   1.188  0.23466
xgra_vclass_reservT4     3.313e-01  5.492e-01   0.603  0.54639
xgra_vclass_reservT6     2.877e-01  2.552e-01   1.128  0.25948
xgra_vclass_reservU      2.026e-01  2.543e-01   0.797  0.42556
xgra_vclass_reservU4     1.007e+00  5.491e-01   1.834  0.06663 .
xgra_vclass_reservU6    -4.934e-01  5.230e-01  -0.943  0.34553
xgra_vclass_reservV      3.169e-01  4.189e-01   0.757  0.44934
xgra_vclass_reservV4    -4.766e-01  5.012e-01  -0.951  0.34166
xgra_vclass_reservV6     5.008e-01  6.210e-01   0.806  0.42002
xgra_vclass_reservW     -1.660e-01  3.484e-01  -0.476  0.63377
xgra_vclass_reservW4     1.225e+00  6.730e-01   1.820  0.06877 .
xgra_vclass_reservW6     1.552e+00  1.029e+00   1.508  0.13144
xgra_vclass_reservX      2.122e-01  3.351e-01   0.633  0.52667
xgra_vclass_reservX4     9.507e-01  4.825e-01   1.970  0.04883 *
xgra_vclass_reservX6     4.720e-01  8.498e-01   0.555  0.57860
xgra_vclass_reservYF     6.238e-01  1.438e+00   0.434  0.66441
xgra_vclass_reservYQ     1.184e+00  1.438e+00   0.824  0.41021
xgra_vclass_reservZ      6.131e-01  3.323e-01   1.845  0.06503 .
Day_of_WeekMonday        5.082e-02  2.440e-02   2.083  0.03729 *
Day_of_WeekSaturday      6.089e-02  2.780e-02   2.190  0.02851 *
Day_of_WeekSunday        1.797e-02  2.424e-02   0.741  0.45845
Day_of_WeekThursday     -1.780e-02  2.516e-02  -0.707  0.47939
Day_of_WeekTuesday       3.285e-02  2.559e-02   1.284  0.19925
Day_of_WeekWednesday    -8.633e-03  2.572e-02  -0.336  0.73713
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.421 on 52294 degrees of freedom
  (1456 observations deleted due to missingness)
Multiple R-squared:  0.5382,    Adjusted R-squared:  0.5377
F-statistic: 952.5 on 64 and 52294 DF,  p-value: < 2.2e-16
```

In my model, I mainly used numerical variables (as above) but I also used character variables such as vehicle class, day of week. The model only worked with numerical variables.

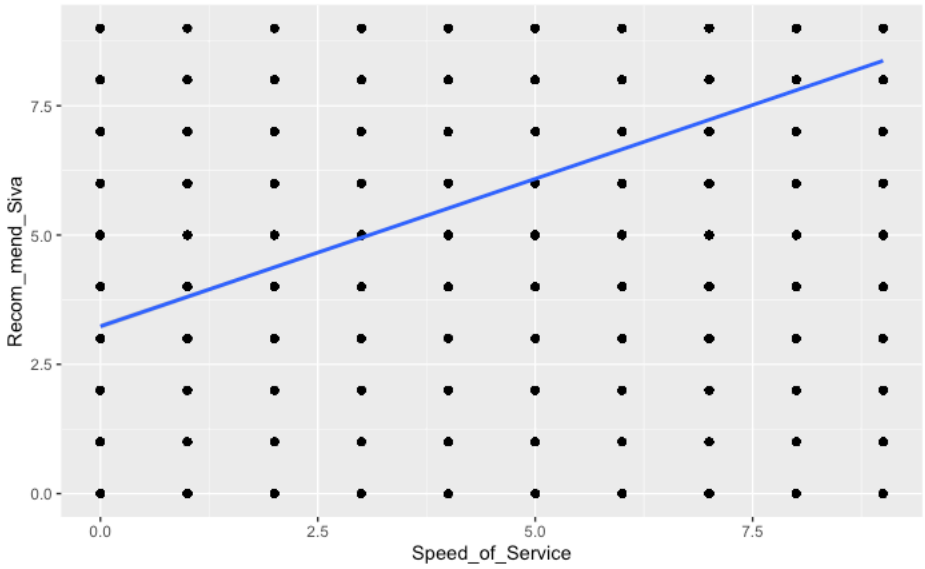
Adding a predicted line; y

```
ggplot(siva, aes(x=Staff_Courtesy, y=Recom_mend_Siva)) +
 geom_point()+
 geom_smooth(method="lm")
```



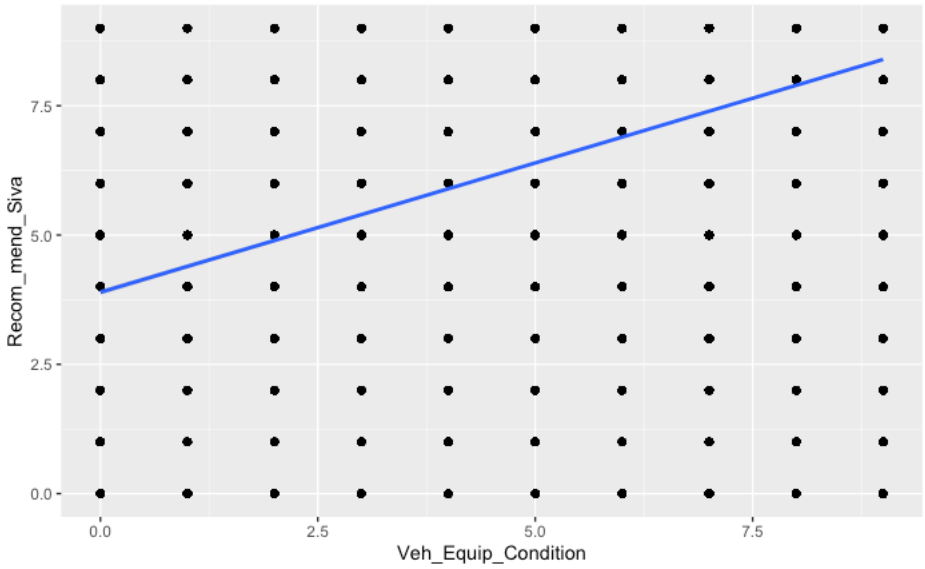

The above graphs shows the relation between the dependent variables: Recom_mend_Siva and the independent variable: Staff_Courtesy. There is a positive trend between the two variables. This means that if a beta coefficient has value→As x (independent variables) increases, y (a dependent variable) increases as well.

```
ggplot(siva, aes(x=Speed_of_Service, y=Recom_mend_Siva)) +
 geom_point()+
 geom_smooth(method="lm")
```



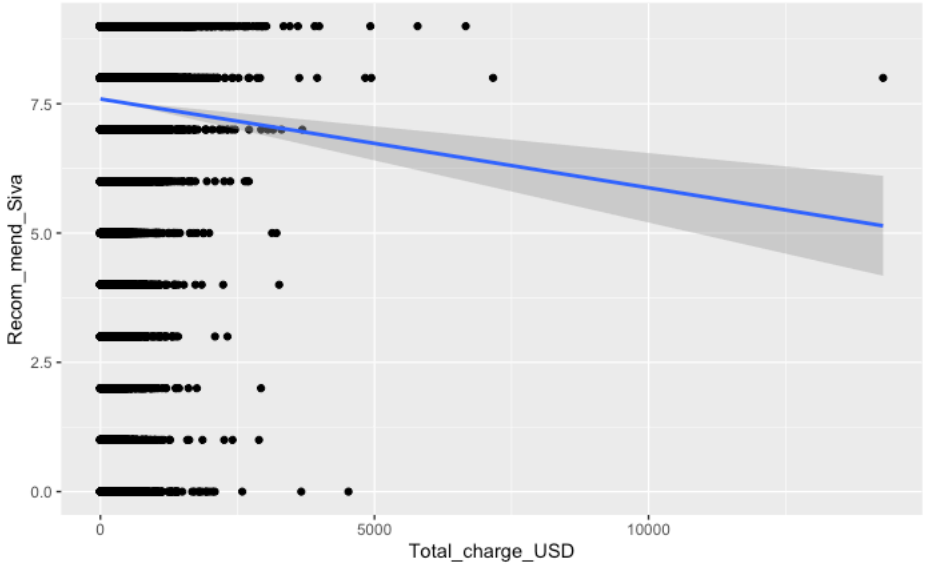

The above graphs shows the relation between the dependent variables: Recom_mend_Siva and the independent variable: Speed_of_Service. There is a positive trend between the two variables. This means that if a beta coefficient has value→As x (independent variables) increases, y (a dependent variable) increases as well.

```
ggplot(siva, aes(x=Veh_Equip_Condition, y=Recom_mend_Siva)) +
 geom_point()+
 geom_smooth(method="lm")
```



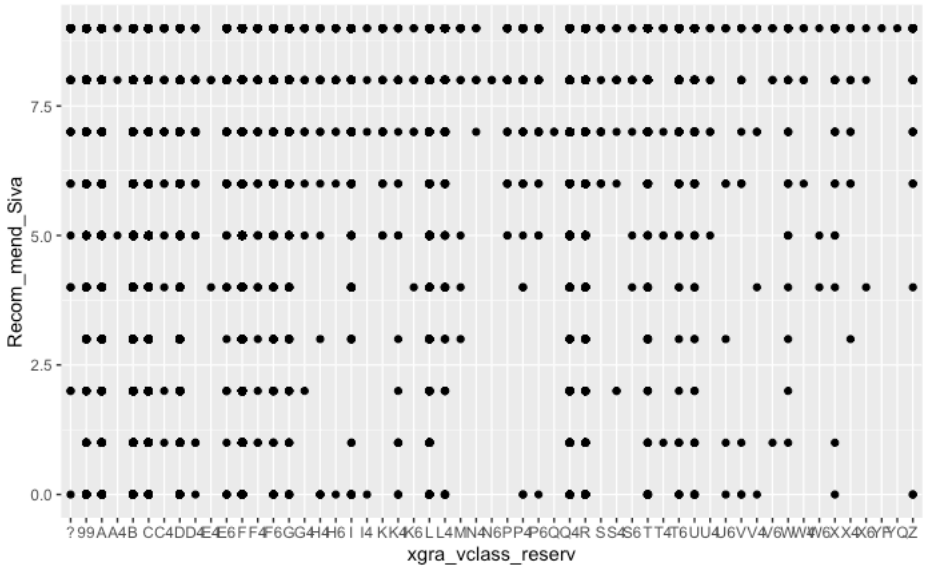

The above graphs shows the relation between the dependent variables: Recom_mend_Siva and the independent variable: Veh_Equip_Condition. There is a positive trend between the two variables. This means that if a beta coefficient has value→As x (independent variables) increases, y (a dependent variable) increases as well.

```
ggplot(siva, aes(x=Total_charge_USD, y=Recom_mend_Siva)) +
 geom_point()+
 geom_smooth(method="lm")
```



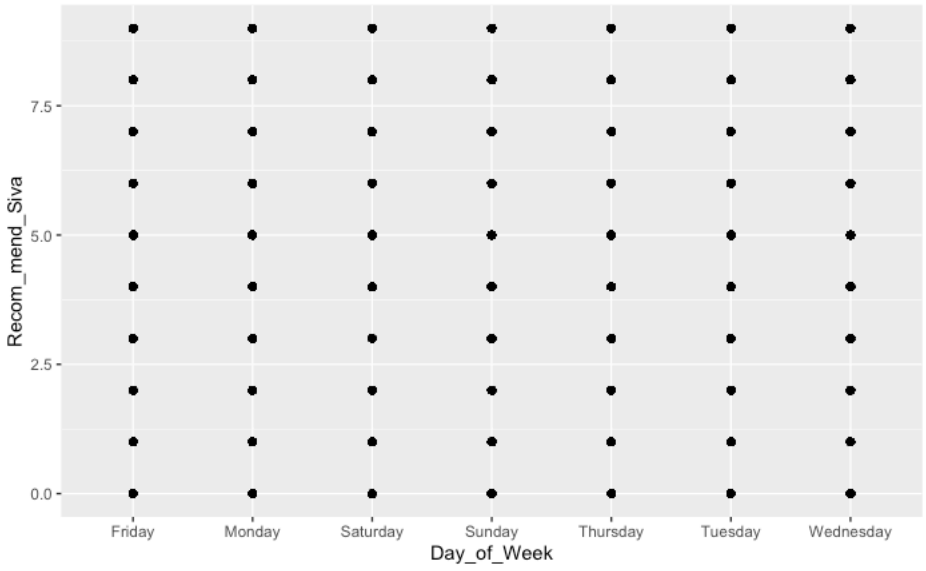
The above graphs shows the relation between the dependent variables: Recom_mend_Siva and the independent variable: Total_charge_USD. There is a negative trend between the two variables. If a beta coefficient has value→As x (independent variables) increases, y (a dependent variable) decreases.

```
ggplot(siva, aes(x=xgra_vclass_reserv, y=Recom_mend_Siva)) +
 geom_point()+
 geom_smooth(method="lm")
```




There is a no predicted line showing a weak correlation between Recom_mend_Siva and xgra_vclass_reserv

```
ggplot(siva, aes(x=Day_of_Week, y=Recom_mend_Siva)) +
 geom_point()+
 geom_smooth(method="lm")
```



There is a no predicted line showing a weak correlation between Recom_mend_Siva and Day_of_week

**CLUSTERING**

Installing and running packages:

```
install.packages("ggpubr")
library(ggpubr)
install.packages("factoextra")
library(factoextra)
```

```
siva_sub1 <- siva[,c(9: 11)]
siva_sub_new2 <- na.omit(siva_sub1)
head(siva_sub_new2)
```

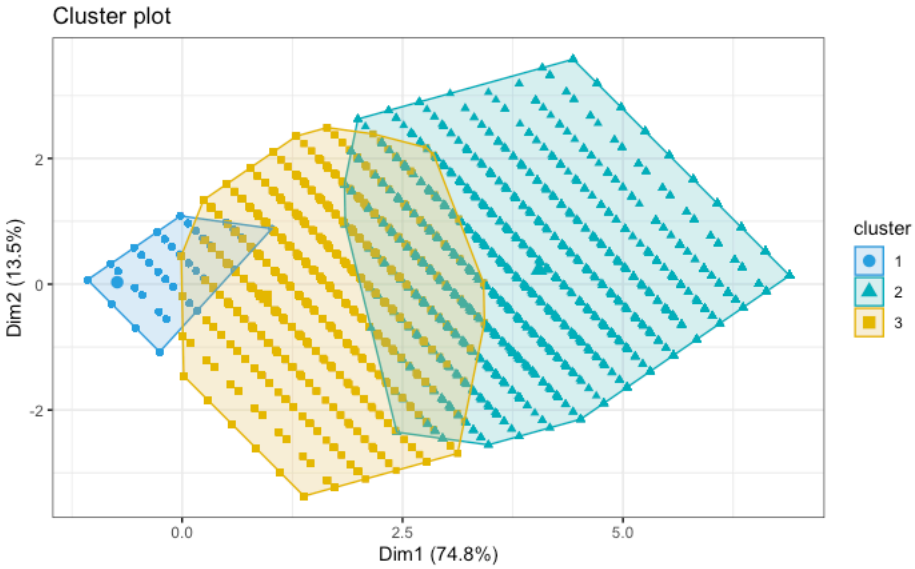
Description: df [6 × 3]

| | Recom_mend_Siva <int> | Staff_Courtesy <int> | Speed_of_Service <int> |
|---|---|---|---|
| 1 | 8 | 9 | 8 |
| 2 | 8 | 8 | 8 |
| 3 | 8 | 7 | 8 |
| 4 | 7 | 8 | 7 |
| 5 | 9 | 9 | 9 |
| 6 | 9 | 9 | 9 |

6 rows

The cluster includes three variables – Recommend Siva, Staff Courtesy and Speed of Service.

```
kmeans.result1 <- kmeans(siva_sub_new2, 3)

fviz_cluster(kmeans.result1, data = siva_sub_new2,
        palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
        geom = "point",
        ellipse.type = "convex",
        ggtheme = theme_bw())
```




Cluster plots are useful for gaining insights into the natural groupings within our data. The three clusters are clearly identified.
The first variable is more closely grouped as opposed to the second and third.