```
df <- read.csv("/content/IBM - IBM_CAO_data_challenge_DS_2021_.csv")
```

```
head(df)
```

A data.frame: 6 × 9

	ORDER_ID	PROD_ID	PROD_CAT	PRICE_ORIG	PRICE_DISC	CLIENT_ID	INDUSTRY	S
	<chr>	<int>	<chr>	<int>	<dbl>	<int>	<chr>	<
1	DS100049976	77563	SW	6882	3303.36	9228913	IT	2
2	DS100049976	19692	SW	4744	2277.12	9228913	IT	2
3	DS100049976	51090	SW	7725	3708.00	9228913	IT	2
4	DS100049976	94654	SAAS	138	66.24	9228913	IT	2
5	DS100049976	77969	SAAS	96	46.08	9228913	IT	2
6	DS100049976	25795	SW	5598	2687.04	9228913	IT	2

```
df['Discount_Percentage'] = ((df['PRICE_ORIG'] - df['PRICE_DISC'])/ df['PRICE_ORIG']) * 100
```

```
install.packages("superml")

library("superml")
```

```
label <- LabelEncoder$new()
df$PROD_CAT <- label$fit_transform(df$PROD_CAT)
```

```
install.packages("jtools")
library("jtools")
```

```
df$INDUSTRY <- label$fit_transform(df$INDUSTRY)
df$SIZE <- label$fit_transform(df$SIZE)
df$STATE <- label$fit_transform(df$STATE)
```

```
head(df)
```

```
lmOut = lm( Discount_Percentage ~ PROD_ID + PROD_CAT + CLIENT_ID + INDUSTRY + SIZE + STATE, d

summary(lmOut)
```

```
Call:
lm(formula = Discount_Percentage ~ PROD_ID + PROD_CAT + CLIENT_ID + 
    INDUSTRY + SIZE + STATE, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-41.67 -18.66  -3.65  15.83  55.45 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.000e+01  3.499e-01 114.304  < 2e-16 ***
PROD_ID      1.288e-05  3.212e-06   4.009 6.10e-05 ***
PROD_CAT     2.787e-01  1.048e-01   2.659  0.00784 ** 
CLIENT_ID   -1.777e-07  3.484e-08  -5.100 3.41e-07 ***
INDUSTRY    -4.393e+00  6.278e-02 -69.982  < 2e-16 ***
SIZE        -1.924e+00  1.056e-01 -18.217  < 2e-16 ***
STATE        3.080e-02  2.014e-02   1.529  0.12616    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.09 on 65746 degrees of freedom
Multiple R-squared:  0.07833,   Adjusted R-squared:  0.07824 
F-statistic: 931.2 on 6 and 65746 DF,  p-value: < 2.2e-16
```

The above results clearly explain that, Prod_id, Prod_Cat, Client_id, Industry, Size and state are the key drivers for discount percentage. The adjusted R2 is 0.8 and P value is < 0.005 which means that the results obtained are statistically significant and not by chance

Now let's check individual drivers for discount

```
lmOut = lm(Discount_Percentage ~ PROD_ID, data = df)
summary(lmOut)
```

```
Call:
lm(formula = Discount_Percentage ~ PROD_ID, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-33.007 -20.069  -4.806  15.510  53.122

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

R squared value is very low and P value is also 0.0001 This means that Prod_id and Discount_percentage are slighly correlated

```
lmOut = lm(Discount_Percentage ~ PROD_CAT, data = df)
summary(lmOut)
```

```
Call:
lm(formula = Discount_Percentage ~ PROD_CAT, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-32.894 -20.286  -4.692  15.511  52.714

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.2863     0.1461 221.031   <2e-16 ***
PROD_CAT      0.2027     0.1089   1.862   0.0626 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.05 on 65751 degrees of freedom
Multiple R-squared:  5.273e-05, Adjusted R-squared:  3.753e-05
F-statistic: 3.468 on 1 and 65751 DF,  p-value: 0.06259
```

R squared value is very low and P value is not less than 0.005 This means that Prod_Cat and Discount_percentage are not correlated

```
lmOut = lm(Discount_Percentage ~ CLIENT_ID, data = df)
summary(lmOut)
```

```
Call:
lm(formula = Discount_Percentage ~ CLIENT_ID, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-33.239 -20.115  -4.858  15.556  53.241

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

R squared value is very low and P value is also 0.0001 This means that Client_id and Discount_percentage are slighly correlated

```
lmOut = lm(Discount_Percentage ~ INDUSTRY, data = df)
summary(lmOut)
```

```
Call:
lm(formula = Discount_Percentage ~ INDUSTRY, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-38.379 -18.379  -3.379  15.621  53.638

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.37890    0.12182  315.04   <2e-16 ***
INDUSTRY    -4.50415    0.06261  -71.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.15 on 65751 degrees of freedom
Multiple R-squared:  0.07298,   Adjusted R-squared:  0.07296
F-statistic:  5176 on 1 and 65751 DF,  p-value: < 2.2e-16
```

R squared value is high and P value is also 0.0001 This means that INDUSTRY and Discount_percentage are highly correlated

```
lmOut = lm(Discount_Percentage ~ SIZE, data = df)
summary(lmOut)
```

```
Call:
lm(formula = Discount_Percentage ~ SIZE, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-35.303 -19.967  -4.635  15.365  57.700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.3030     0.1479  238.76   <2e-16 ***
```

R squared value is very low and P value is less than 0.05 This means that SIZE and Discount_percentage are slighly correlated

```
Residual standard error: 23.94 on 65751 degrees of freedom
```

```
lmOut = lm(Discount_Percentage ~ STATE, data = df)
summary(lmOut)
```

```
Call:
lm(formula = Discount_Percentage ~ STATE, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-32.540 -20.467  -4.504  15.511  52.569

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.540133   0.160523 202.714   <2e-16 ***
STATE       -0.007284   0.020966  -0.347    0.728
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.05 on 65751 degrees of freedom
Multiple R-squared:  1.836e-06, Adjusted R-squared:  -1.337e-05
F-statistic: 0.1207 on 1 and 65751 DF,  p-value: 0.7283
```

R squared value is very low and P value not less than 0.05 This means that Prod_id and Discount_percentage are not correlated

✓ 0s completed at 7:41 AM