PROJECT
DATA MINING

Data Mining Extended Project

Part 1: PCA:

Problem Statement:

The ‘Hair Salon.csv’ dataset contains various variables used for the context of Market Segmentation. This particular case study is based on various parameters of a salon chain of hair products. You are expected to do Principal Component Analysis for this case study according to the instructions given in the following rubric.

Note: This particular dataset contains the target variable satisfaction as well. Please do drop this variable before doing Principal Component Analysis.

Answer:

	ID	ProdQual	Ecom	TechSup	CompRes
count	100.000000	100.000000	100.000000	100.000000	100.000000
mean	50.500000	7.810000	3.672000	5.365000	5.442000
std	29.011492	1.396279	0.700516	1.530457	1.208403
min	1.000000	5.000000	2.200000	1.300000	2.600000
25%	25.750000	6.575000	3.275000	4.250000	4.600000
50%	50.500000	8.000000	3.600000	5.400000	5.450000
75%	75.250000	9.100000	3.925000	6.625000	6.325000
max	100.000000	10.000000	5.700000	8.500000	7.800000

	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim
count	100.000000	100.000000	100.00000	100.000000	100.000000
mean	4.010000	5.805000	5.12300	6.974000	6.043000
std	1.126943	1.315285	1.07232	1.545055	0.819738
min	1.900000	2.300000	2.90000	3.700000	4.100000
25%	3.175000	4.700000	4.50000	5.875000	5.400000
50%	4.000000	5.750000	4.90000	7.100000	6.100000
75%	4.800000	6.800000	5.80000	8.400000	6.600000
max	6.500000	8.400000	8.20000	9.900000	8.100000

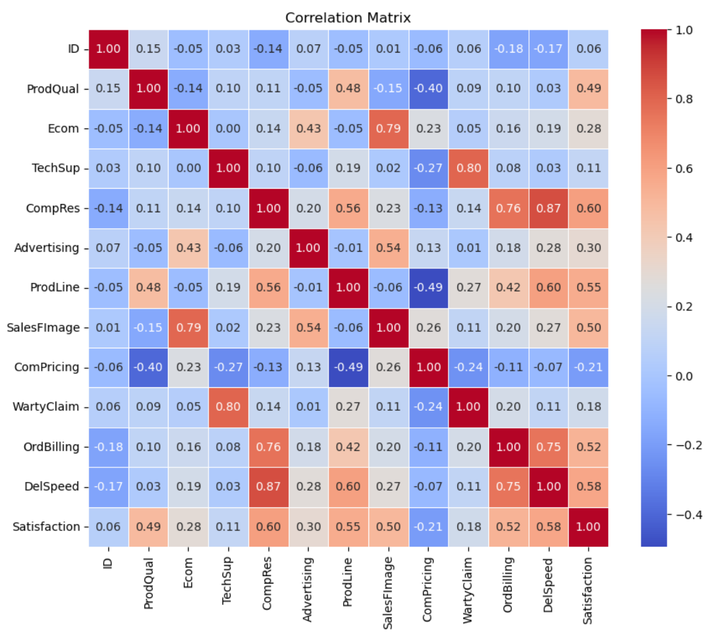
	OrdBilling	DelSpeed	Satisfaction
count	100.00000	100.000000	100.000000
mean	4.27800	3.886000	6.918000
std	0.92884	0.734437	1.191839
min	2.00000	1.600000	4.700000
25%	3.70000	3.400000	6.000000
50%	4.40000	3.900000	7.050000
75%	4.80000	4.425000	7.625000
max	6.70000	5.500000	9.900000

Correlations:

	ID	ProdQual	Ecom	TechSup	CompRes	Advertising \
ID	1.000000	0.145774	-0.046173	0.031838	-0.144322	0.073129
ProdQual	0.145774	1.000000	-0.137163	0.095600	0.106370	-0.053473
Ecom	-0.046173	-0.137163	1.000000	0.000867	0.140179	0.429891
TechSup	0.031838	0.095600	0.000867	1.000000	0.096657	-0.062870
CompRes	-0.144322	0.106370	0.140179	0.096657	1.000000	0.196917
Advertising	0.073129	-0.053473	0.429891	-0.062870	0.196917	1.000000
ProdLine	-0.048641	0.477493	-0.052688	0.192625	0.561417	-0.011551
SalesFImage	0.013848	-0.151813	0.791544	0.016991	0.229752	0.542204
ComPricing	-0.063007	-0.401282	0.229462	-0.270787	-0.127954	0.134217
WartyClaim	0.058592	0.088312	0.051898	0.797168	0.140408	0.010792
OrdBilling	-0.178352	0.104303	0.156147	0.080102	0.756869	0.184236
DelSpeed	-0.172134	0.027718	0.191636	0.025441	0.865092	0.275863
Satisfaction	0.061143	0.486325	0.282745	0.112597	0.603263	0.304669

	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling \
ID	-0.048641	0.013848	-0.063007	0.058592	-0.178352
ProdQual	0.477493	-0.151813	-0.401282	0.088312	0.104303
Ecom	-0.052688	0.791544	0.229462	0.051898	0.156147
TechSup	0.192625	0.016991	-0.270787	0.797168	0.080102
CompRes	0.561417	0.229752	-0.127954	0.140408	0.756869
Advertising	-0.011551	0.542204	0.134217	0.010792	0.184236
ProdLine	1.000000	-0.061316	-0.494948	0.273078	0.424408
SalesFImage	-0.061316	1.000000	0.264597	0.107455	0.195127
ComPricing	-0.494948	0.264597	1.000000	-0.244986	-0.114567
WartyClaim	0.273078	0.107455	-0.244986	1.000000	0.197065
OrdBilling	0.424408	0.195127	-0.114567	0.197065	1.000000
DelSpeed	0.601850	0.271551	-0.072872	0.109395	0.751003
Satisfaction	0.550546	0.500205	-0.208296	0.177545	0.521732

	DelSpeed	Satisfaction
ID	-0.172134	0.061143
ProdQual	0.027718	0.486325
Ecom	0.191636	0.282745



Simple Linear Models :

```
==================================================================
=========
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------
---------
const          3.6759      0.598      6.151      0.000       2.490
4.862
ProdQual       0.4151      0.075      5.510      0.000       0.266
0.565
==================================================================
```

Satisfaction = 3.6759 + 0.4151 * ProdQual

1.beta-naught or intercept coefficient is equal to 3.6759

2.beta-slope or the variable coefficient Product quality = 0.4151

3.for any one unit change in product quality Satisfaction rating would impr ove by 0.4151 keeping other things constant as explained by model

```
==================================================================
=========
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------
---------
const          5.1516      0.616      8.361      0.000       3.929
6.374
Ecom           0.4811      0.165      2.918      0.004       0.154
0.808
==================================================================
```

Satisfaction = 5.1516 + 0.4811 * Ecom

```
=========
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------
---------
const          6.4476      0.436     14.791      0.000       5.583
7.313
TechSup        0.0877      0.078      1.122      0.265      -0.067
0.243
==================================================================
```

Satisfaction = 6.44757 + 0.08768 * TechSup

```
==================================================================
=========
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------
---------
const          3.6800      0.443      8.310      0.000       2.801
4.559
CompRes        0.5950      0.079      7.488      0.000       0.437
0.753
==================================================================
```

Satisfaction = 3.680 + 0.595 * CompRes

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.6259	0.424	13.279	0.000	4.785	6.467
Advertising	0.3222	0.102	3.167	0.002	0.120	0.524

Satisfaction = 5.6259 + 0.3222 * Advertising

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0220	0.455	8.845	0.000	3.120	4.924
ProdLine	0.4989	0.076	6.529	0.000	0.347	0.651

Satisfaction = 4.0220 + 0.4989 * ProdLine

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0698	0.509	8.000	0.000	3.060	5.079
SalesFImage	0.5560	0.097	5.719	0.000	0.363	0.749

Satisfaction = 4.070 + 0.556 * SalesFImage

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.0386	0.544	14.769	0.000	6.958	9.119
ComPricing	-0.1607	0.076	-2.108	0.038	-0.312	-0.009

Satisfaction = 8.0386 + (-0.1607) * ComPricing

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.3581	0.881	6.079	0.000	3.609	7.107
WartyClaim	0.2581	0.145	1.786	0.077	-0.029	0.545

Satisfaction = 5.3581 + 0.2581 * WartyClaim

```
==================================================================
=========
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------
---------
const          4.0541      0.484      8.377      0.000       3.094
5.014
OrdBilling     0.6695      0.111      6.054      0.000       0.450
0.889
==================================================================
```

Satisfaction = 4.0541 + 0.6695 * OrdBilling

```
=========
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------
---------
const          3.2791      0.529      6.194      0.000       2.229
4.330
DelSpeed       0.9364      0.134      6.994      0.000       0.671
1.202
==================================================================
```

Satisfaction = 3.2791 + 0.9364 * DelSpeed

Principal Component Analysis:

Conducting a bartlett sphericity test to check whether Principal Component Analysis can be done on the predictor variables of the dataset:

```
Chi-Square Value: 619.2725577964159
P-value: 1.793370009363552e-96
```

Since the p value for the test is quite less signficance level of alpha = 0.001 so we reject the null hypothesis Ho (that PCA cannot be conducted implying that there is no correlation amongst the predictor variables)

PCA workout

Using the rotation type of varimax we conduct the PCA analysis with 4 factors Dataset hair.corr has all 11 predictor variables (minus the ID column and dependent variable Satisfaction ratings)

```
Factor Loadings:
                     0         1         2         3
ProdQual      0.023986 -0.070194  0.015714  0.646756
Ecom          0.068920  0.781470  0.028048 -0.114545
TechSup       0.019547 -0.025660  0.889679  0.115366
CompRes       0.897429  0.129730  0.053820  0.131827
Advertising   0.166362  0.528760 -0.042875 -0.062563
ProdLine      0.525424 -0.035276  0.127176  0.712145
SalesFImage   0.113605  0.980071  0.063652 -0.132610
ComPricing   -0.075566  0.212761 -0.208944 -0.590359
WartyClaim    0.102623  0.056708  0.878694  0.129163
OrdBilling    0.768271  0.126614  0.088106  0.088788
DelSpeed      0.948841  0.185127 -0.004712  0.087337
```

PCA Explained

The 4 RCs explain explain about 80 % of cumulative variation in the dataset which is good number After studying the PCA results on hair dataset an arbitrary number was choosen as cutoff (0.6) to check whether the variablity of the predictors can be explained by single components. It worked and we can see that every input variable can be explained by the single set of Components (RCs)


```
Factor Loadings:
                     0         1         2         3
ProdQual      0.023986 -0.070194  0.015714  0.646756
Ecom          0.068920  0.781470  0.028048 -0.114545
TechSup       0.019547 -0.025660  0.889679  0.115366
CompRes       0.897429  0.129730  0.053820  0.131827
Advertising   0.166362  0.528760 -0.042875 -0.062563
ProdLine      0.525424 -0.035276  0.127176  0.712145
SalesFImage   0.113605  0.980071  0.063652 -0.132610
ComPricing   -0.075566  0.212761 -0.208944 -0.590359
WartyClaim    0.102623  0.056708  0.878694  0.129163
OrdBilling    0.768271  0.126614  0.088106  0.088788
DelSpeed      0.948841  0.185127 -0.004712  0.087337
```


Scores for individual IDs (rows of observation) was extracted from the PCA analysis and rounded off to two decimal places for ease of computation :

Table for Meaningful names of Principal Components

Components	**Meaningful Names**	**Column Name**
RC1	Purchasing Experience	Pchexp
RC2	Brand Recognition	Bdrecog
RC3	After Sales Service	Aftsvc
RC4	Product	Prodt

Explanation

1. RC1 - Purchasing Experience explains about variables affecting Complaint resolution, Order and Billing and delivery speed to customers

2. RC2 - Brand recognition handles Ecommerce, image of Sales force , Advertising which is face of the product

3. RC3 - After Sales Service gives information about Technical support, and Warranty and claims if there is any problem to customer after he has bought the item

4. RC4 – Product talks about the qualities of product like its varieties and types, prices its quality i.e all tangible aspects about the very existence of company.

Score matrix was converted into a data frame and its variables which are nothing but PCA components were given meaningful names for further analysis We achieved a dimensionality reduction where just 4 factors can explain the complete 11 predictor variables of the hair dataset through PCA analysis.

Score head

```
Variance Inflation Factor (VIF):
       Variable       VIF
0      ProdQual  0.465400
1          Ecom  0.648470
2       TechSup  0.658320
3       CompRes  0.834310
4   Advertising  0.302278
5      ProdLine  0.720493
6   SalesFImage  0.720370
7    ComPricing  0.393829
8    WartyClaim  0.690173
9    OrdBilling  0.668513
10     DelSpeed  0.861591
```

Score data frame was combined with a smaller subset (extracted data frame - hair_new) having ID and Satisfaction ratings as columns to form a meaningful dataset devoid of multicollinearity and manageable predictor variables (just 4) for further Regression model building.

```
Breusch-Pagan Test for Heteroscedasticity:
Test Statistic: 13.192917327906535
P-value: 0.2130847550267204
```

Multiple Linear Regression Model Validity:

```
Regression Model Summary:
                            OLS Regression Results

==============================================================================
=======
```

```
Dep. Variable:           Satisfaction   R-squared:
  0.813
Model:                            OLS   Adj. R-squared:
  0.783
Method:                 Least Squares   F-statistic:
  26.92
Date:                Thu, 04 Jan 2024   Prob (F-statistic):           1
.43e-20
Time:                        12:58:51   Log-Likelihood:
-63.822
No. Observations:                  80   AIC:
  151.6
Df Residuals:                      68   BIC:
  180.2
Df Model:                          11

Covariance Type:            nonrobust

=======================================================================
========
                  coef    std err          t      P>|t|      [0.025
  0.975]
-----------------------------------------------------------------------
--------
const          -0.4363      1.066     -0.409      0.684      -2.564
   1.692
ProdQual        0.3846      0.064      6.039      0.000       0.258
   0.512
Ecom           -0.4739      0.159     -2.984      0.004      -0.791
  -0.157
TechSup         0.0587      0.075      0.778      0.439      -0.092
   0.209
CompRes         0.1652      0.128      1.292      0.201      -0.090
   0.420
Advertising    -0.0172      0.071     -0.242      0.810      -0.159
   0.125
ProdLine        0.1323      0.092      1.437      0.155      -0.051
   0.316
SalesFImage     0.8333      0.113      7.353      0.000       0.607
   1.059
ComPricing     -0.0639      0.054     -1.187      0.239      -0.171
   0.044
WartyClaim     -0.1717      0.155     -1.105      0.273      -0.482
   0.138
OrdBilling      0.1224      0.120      1.020      0.311      -0.117
   0.362
DelSpeed        0.2243      0.239      0.940      0.351      -0.252
   0.701
=======================================================================
=======
Omnibus:                        5.813   Durbin-Watson:
  1.898
Prob(Omnibus):                  0.055   Jarque-Bera (JB):
  5.907
Skew:                          -0.657   Prob(JB):
 0.0522
Kurtosis:                       2.786   Cond. No.
   298.
```

```
==============================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
Regression Model Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:           Satisfaction   R-squared:                       0.813
Model:                            OLS   Adj. R-squared:                  0.783
Method:                 Least Squares   F-statistic:                     26.92
Date:                Thu, 04 Jan 2024   Prob (F-statistic):           1.43e-20
Time:                        12:58:51   Log-Likelihood:                -63.822
No. Observations:                  80   AIC:                             151.6
Df Residuals:                      68   BIC:                             180.2
Df Model:                          11
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const          -0.4363      1.066     -0.409      0.684      -2.564       1.692
ProdQual        0.3846      0.064      6.039      0.000       0.258       0.512
Ecom           -0.4739      0.159     -2.984      0.004      -0.791      -0.157
TechSup         0.0587      0.075      0.778      0.439      -0.092       0.209
CompRes         0.1652      0.128      1.292      0.201      -0.090       0.420
Advertising    -0.0172      0.071     -0.242      0.810      -0.159       0.125
ProdLine        0.1323      0.092      1.437      0.155      -0.051       0.316
SalesFImage     0.8333      0.113      7.353      0.000       0.607       1.059
ComPricing     -0.0639      0.054     -1.187      0.239      -0.171       0.044
WartyClaim     -0.1717      0.155     -1.105      0.273      -0.482       0.138
OrdBilling      0.1224      0.120      1.020      0.311      -0.117       0.362
DelSpeed        0.2243      0.239      0.940      0.351      -0.252       0.701
==============================================================================
Omnibus:                        5.813   Durbin-Watson:                   1.898
Prob(Omnibus):                  0.055   Jarque-Bera (JB):                5.907
Skew:                          -0.657   Prob(JB):                       0.0522
Kurtosis:                       2.786   Cond. No.                         298.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Summary Explained

1. Looking at the Pr(t) values of Coefficients like Intercept (constant beta-naught) we see that it is significant even at 0.001 level. so it definitely not zero and contributes to regression model

2. Similarly predictor variables like Purchase experience, Brand Recognition and Product have significant betas implying that Response variable Satisfaction is linearly associated with them

3. After sales service is the only variable which has some high p-value implying that its beta coefficient may not be contributing that significantly to the model or may be zero

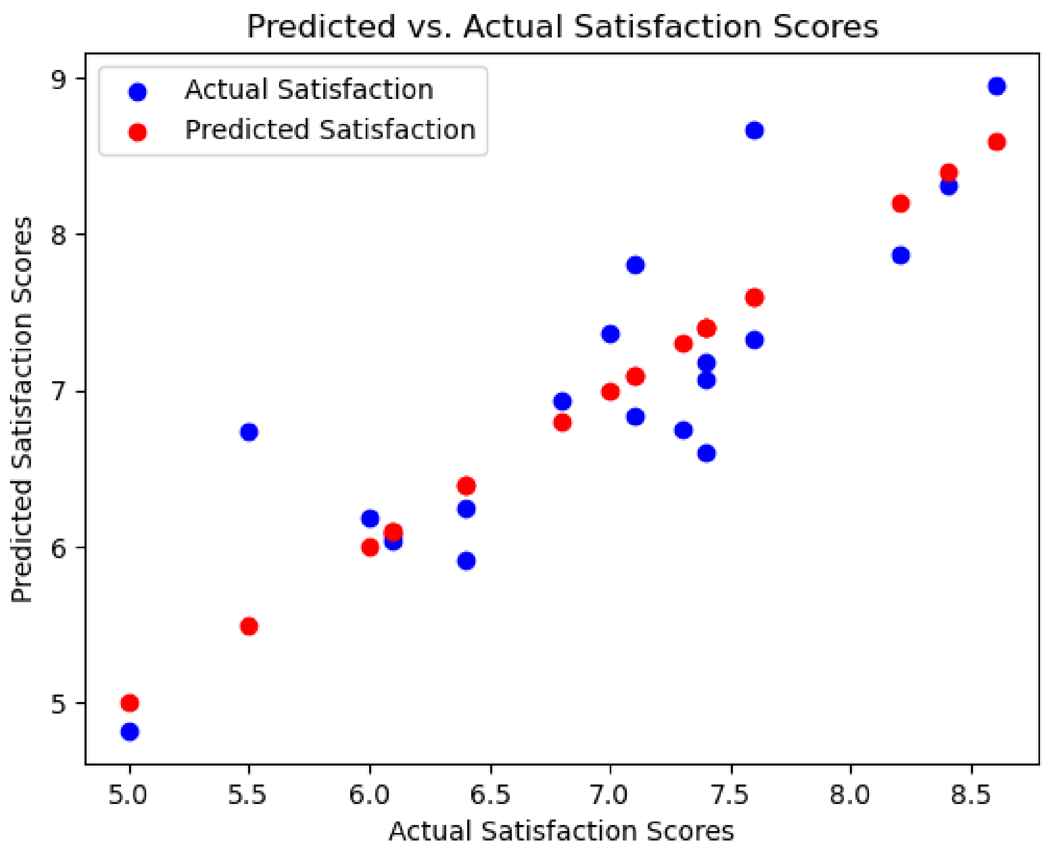
4. All together Adj-R^2 explains that these predictors explains the 64.6 % of the variability in the dataset which is still good enough (may not fall in excellent category)

5. Overall p-value (extremely less e raise to minus 16) of Model given by F-statistic gives evidence against the null-hypothesis. Model is significantly valid at this point

Using the newly built multiple regression model new Satisfaction scores were predicted (pred.Satisfn) to check the validity of the model New dataframe hair_new was formed to have columns as 1. IDs, 2. Satisfaction ratings 3. Purchase Experience 4. Brand Recognition 5. After Sales service 6. predicted satisfaction (from multiple linear model)

Predicted v/s Actual Satisfactions

Plot analysis revealed that our new MLR Regression model is quite good and close to actual Satisfaction scores Blue dots represent Actual Satisfaction ratings Red dots represent Predicted satisfaction scores derived from multiple linear regression model



Conclusion

Based on the consumer goods product – Hair – market segmentation data set, we can conclude that, due to multicollinearity within independent variables, we cannot apply regression model directly on the date set.

So, we created new data set – New hair – based on Principal Component Analysis. We have also recommended subjective new variable names as ServDesk, MktDesk, SuppDesk and RechDesk to the components. And then, based on Factor Analysis study we performed multi linear regressing.

Based on the regression model we have concluded that Sales Service Desk plays – the most significant role in customer satisfaction. That means company should be extra cautious in Complain Resolution, Order & Billing, and Delivery Speed fronts. If Delivery is late or complaint is not resolved in time may leads to decline in company's revenue. However, Brand Marketing Desk and Strategic

Research Desk also plays important role with 0.509 and 0.540 weighted respectively in the regression model.

From the study, we have also concluded that due to consumer goods product type customer do not give significance to Technical Support and Warranty & Claims, And hence SuppDesk variable does not play significance role in customer satisfaction index.

In overall study, we removed multicollinearity from the data, we built regression model, we tested regression model and based on BackTrack data we also predicted Actual vs. Predicted customer satisfaction score in line chart.

In product or service based companies, if customer/prospect is satisfied with product, he will make purchase again and again for that particular product, and that works as revenue multiplier for the company. High customer satisfaction can also leads to cross selling of products.

Hence, we suggest management to conduct customer survey on regular bases to identify trends and relationship for higher customer satisfaction experience.