**Institute of Systems Science**
**National University of Singapore**

# MASTER OF TECHNOLOGY IN ENTERPRISE BUSINESS ANALYTICS/ KNOWLEDGE ENGINEERING

## Unit 5 PCA_Factor_Cluster_Assignment Analysis - 2018

## Subject: ***Advanced Analytics / Data Mining Methodology & Methods***

**8 assignments chosen at random for discussion from both KE & EBAC groups**

**EBAC Group EXPLORER**

**Objective:**
Baby low birth weight problem can happen to any parents in the world. If baby born weight is less than 5 pounds 8 ounces, it can be considered as low birth weight. Usually babies with low birthweight look much smaller than other babies of normal birth weight. Their head maybe appear to be bigger than the rest of the body and often looks thin with little body fat.
In this assignment, we wish to analyze information on new born babies and their parents. We have found the dataset from the website that they have collected it from UK hospitals.

**Data Source & Description:**
In this dataset, baby body size and mother pregnancy detail as well as parent lifestyles are recorded as part of experiment. Dataset contains baby body characteristic after birth such as length, weights, mother information (gestation, smother, age and cigarette and maternal weight/height) and father information (age, height, education year and cigarette).

| name | Measure | Label |
|---|---|---|
| id | Ordinal | Baby number |
| head_circumference | Ordinal | Head circumference at birth |
| body_length | Ordinal | Length of baby at birth (**inches**) |
| birth_weight | Ordinal | Weight of baby at birth (lbs) - Predictor Y |
| gestation | Ordinal | Gestational age at birth (**weeks**) |
| mother_smoker | Nominal | Smoker (1= smoker, 0 = non-smoker) |
| mother_age | Ordinal | Age of mother |
| mother_cigarette | Ordinal | Number of cigarettes smoked by mother (per **day**) |
| maternal_height | Ordinal | Maternal height (**inches**) |
| maternal_weight | Ordinal | Mothers pre-pregnancy weight (**lbs**) |
| father_age | Ordinal | Father's age |
| father_edu_yr | Ordinal | Fathers years in education |
| father_cigarette | Ordinal | Number of cigarettes smoked by father (per day) |
| father_height | Ordinal | Height of father (**inches**) |
| low_birth_weight | Category | Low birth weight baby |
| mother_over35 | Nominal | Mother over 35 |
| low_brith_weightt | Nominal | Birth weight (**lb**) |

**Correlation Analysis:**
The result of correlation matrix shows that variables related to baby and mother gestation period are correlated each other. Variables related to father info seem to be low correlated to other variables. Here are some findings with correlation matrix.
- Baby birth body related attributes are highly correlated each other
- The bigger baby is, the mother tend to be taller and heavier
- The parents usually married at the same age
- Father lifestyle and physical attributes as well education has no correlation with baby.

| | head_circumference | body_length | birth_weight | gestation | mother_age | mother_cigarette | maternal_height | maternal_weight | father_age | father_edu_yr | father_cigarette | father_height |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| head_circumference | 1.0000 | 0.5653 | 0.7364 | 0.4440 | 0.1121 | -0.1314 | 0.3813 | 0.3576 | 0.3014 | 0.0834 | -0.0277 | 0.0405 |
| body_length | 0.5653 | 1.0000 | 0.6970 | 0.6514 | -0.0207 | -0.1571 | 0.4147 | 0.3044 | 0.0789 | -0.0507 | 0.0197 | 0.1871 |
| birth_weight | 0.7364 | 0.6970 | 1.0000 | 0.7063 | 0.0010 | -0.1512 | 0.3679 | 0.3896 | 0.1768 | 0.0739 | -0.0889 | 0.0248 |
| gestation | 0.4440 | 0.6514 | 0.7063 | 1.0000 | 0.0108 | 0.0432 | 0.2309 | 0.2505 | 0.1422 | 0.1310 | -0.1138 | 0.1879 |
| mother_age | 0.1121 | -0.0207 | 0.0010 | 0.0108 | 1.0000 | 0.3403 | 0.0468 | 0.2776 | 0.8066 | 0.4417 | 0.0909 | -0.2036 |
| mother_cigarette | -0.1314 | -0.1571 | -0.1512 | 0.0432 | 0.3403 | 1.0000 | 0.1719 | 0.1540 | 0.2484 | 0.1985 | 0.2573 | 0.0084 |
| maternal_height | 0.3813 | 0.4147 | 0.3679 | 0.2309 | 0.0468 | 0.1719 | 1.0000 | 0.6712 | -0.0717 | 0.0162 | 0.0491 | 0.2728 |
| maternal_weight | 0.3576 | 0.3044 | 0.3896 | 0.2505 | 0.2776 | 0.1540 | 0.6712 | 1.0000 | 0.2534 | 0.1877 | 0.0508 | 0.1083 |
| father_age | 0.3014 | 0.0789 | 0.1768 | 0.1422 | 0.8066 | 0.2484 | -0.0717 | 0.2534 | 1.0000 | 0.3005 | 0.1359 | -0.2986 |
| father_edu_yr | 0.0834 | -0.0507 | 0.0739 | 0.1310 | 0.4417 | 0.1985 | 0.0162 | 0.1877 | 0.3005 | 1.0000 | -0.2631 | -0.0053 |
| father_cigarette | -0.0277 | 0.0197 | -0.0889 | -0.1138 | 0.0909 | 0.2573 | 0.0491 | 0.0508 | 0.1359 | -0.2631 | 1.0000 | 0.3255 |
| father_height | 0.0405 | 0.1871 | 0.0248 | 0.1879 | -0.2036 | 0.0084 | 0.2728 | 0.1083 | -0.2986 | -0.0053 | 0.3255 | 1.0000 |

**Naming the chosen PCA-s or Factors:**

1 : PCA (non-rotated component)
The outputs of PCA (formatted loading matrix) are hard to interpret, there is still correlation between orthogonal components. It is confirmed that we need to rotate component to find meaningful components. So we try to use rotated analysis for PCA.
2 : PCA (Rotated analysis)
Even after rotated loading in PCA, the communality estimation has issue with low value (<0.5) for many variables. So next is to try to use Factor analysis with variables to find latent information from the dataset.
3 : Factor analysis (Default setting)
The result output of Factor analysis become clear and meaningful to interpret to compare to the principal component. Variance Explained and Final Communality Estimates after Rotation seem to be better.
4 : Factor analysis (Factoring method/Prior community - Principal component), Varimax
Factor analysis with this setting (Factoring method/Prior community - Principal component and varimax) result in much better result compared to previous factor analysis with default setting. Up to factor 4 (eigenvalue greater than 1) produce 71% of variances. Although factor 5 has less than 1 (0.95), it is applicable to take it after considering the variable combination for factor 5. But we didn't include it for further analysis in cluster analysis.

**Clustering:**

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|
| Average baby head and weight<br>Mother weight are below average<br>Parents are younger than average | Biggest baby size<br>Parent are tend to be older<br>Mother has larger weight during pregnancy | Average baby weight and head size<br>Above average mother weight and baby size | Older parents with average baby weight and size |

**Regression :**
**We can predict low birth weight of baby with orthogonal variables using multiple linear regression. The 4 rotated components from factor analysis are used as independent variables and low birth weight is as dependent variable.**

| | |
|---|---|
| Factor 1 | Parent age with mother smoking cigarette |
| Factor 2 | Baby body size and gestation period |
| Factor 3 | Mother physical condition during pregnancy |
| Factor 4 | Smoking parents with father height |
| low_birth_weight | Low birth weight (lb) - Y variables |

First, use all of 4 latest factors as input variables. The result show as followed.

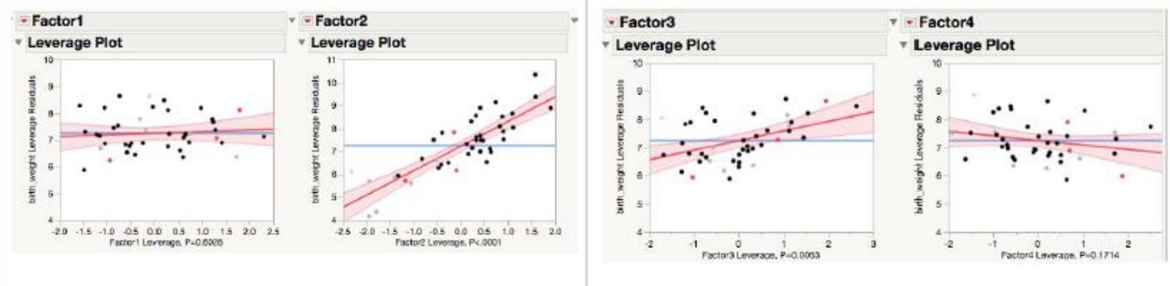
**Summary of Fit**

| | |
|---|---|
| RSquare | 0.726952 |
| RSquare Adj | 0.697433 |
| Root Mean Square Error | 0.731438 |
| Mean of Response | 7.264286 |
| Observations (or Sum Wgts) | 42 |

| AICc | BIC |
|---|---|
| 101.9968 | 110.0228 |

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intercept | 7.2642857 | 0.112863 | 64.36 | <.0001* |
| Factor1 | 0.0599554 | 0.114231 | 0.52 | 0.6028 |
| Factor2 | 1.0685548 | 0.114231 | 9.35 | <.0001* |
| Factor3 | 0.3385498 | 0.114231 | 2.96 | 0.0053* |
| Factor4 | -0.159311 | 0.114231 | -1.39 | 0.1714 |




Looking at the p-values in the Parameter Estimates report, Factor2 and Factor3 appear to be significant predictors of low_birth_weight. The next step might be to reduce the model by removing insignificant predictors.

**Conclusion:**
From result of PCA/Factor analysis and cluster analysis in this assignment, we have found that it is not appropriate to jump to choose either PCA or factor analysis before analysing components. If only result of components is interpretable and practical or making sense, we can choose PCA or Factor analysis. At the same time, it is good to try different setting of dimension reduction because finding latent factor is not straightforward.

In cluster analysis, using interpretable components from factor analysis, it does not guarantee to produce meaningful clusters. Using different cluster size in this experiment, we found that cluster size of 4 produce distinct cluster profile. Moreover, It is important to look at each variable whether it produce more understandable and interpretable clusters.

Finally we have learnt that dependent Y variable (low birth weight) should not be included in PCA and factor analysis because it will be predicted in regression analysis with factors and it's necessary to check which X variables are significant enough with p-value and only then be included in regression equation.

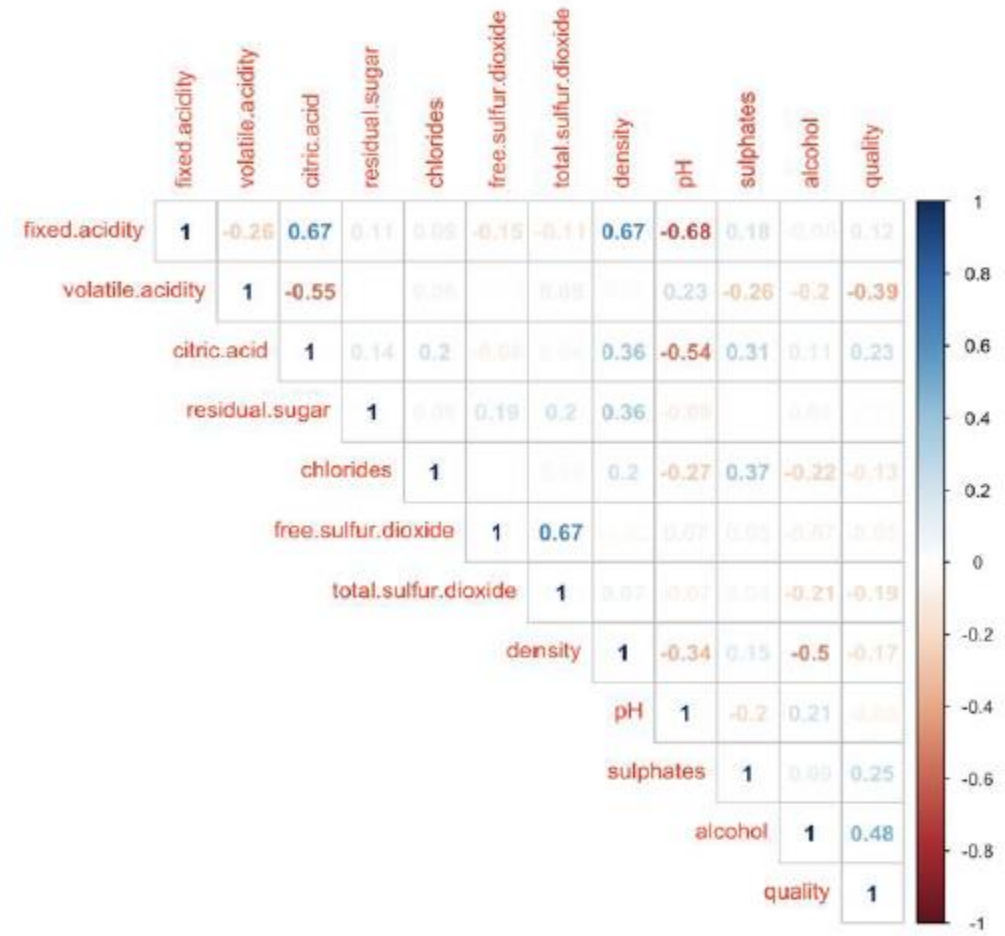
## EBAC Group The R Ninjas

**Objective:**
Wine industry has been experiencing a continuous growth as more and more consumers turn to wine for its unique aroma and taste. Wine certification and quality assessment are the key elements the industry considers when setting price. Laboratory-based physicochemical tests are generally used in the wine certification process to measure wine features such as acidity, pH level, presence of sugar and other chemical properties. Quality is a sensory result evaluated by wine tasters, which is relatively subjective and can vary from person to person.
For wine industry, it would be helpful to know the connection between chemical properties and the sensory quality ranking of wine to provide guidance for wine makers regarding product stratification and the expected product price.
We will use the publicly available red wine quality dataset from the UCI Machine Learning Repository to perform our analysis. The dataset contains 1599 instances with 12 attributes for red variants of the Portuguese "Vinho Verde" wine.

**Data Source & Description:**
The red wine quality dataset contains 1599 instances with 11 physicochemical features (density, pH level, alcohol, etc.) and 1 sensory variable (quality). The physicochemical features are all continuous

and the sensory variable is ordinal, with 0 being the worst and 10 being the best. Each sample of wine was evaluated by a minimum of three sensory assessors through blind taste test, and a score was given. The median of these evaluation scores was recorded as the final sensory quality score. In our dataset, quality scores range from the 3, the lowest, to 8, the highest.

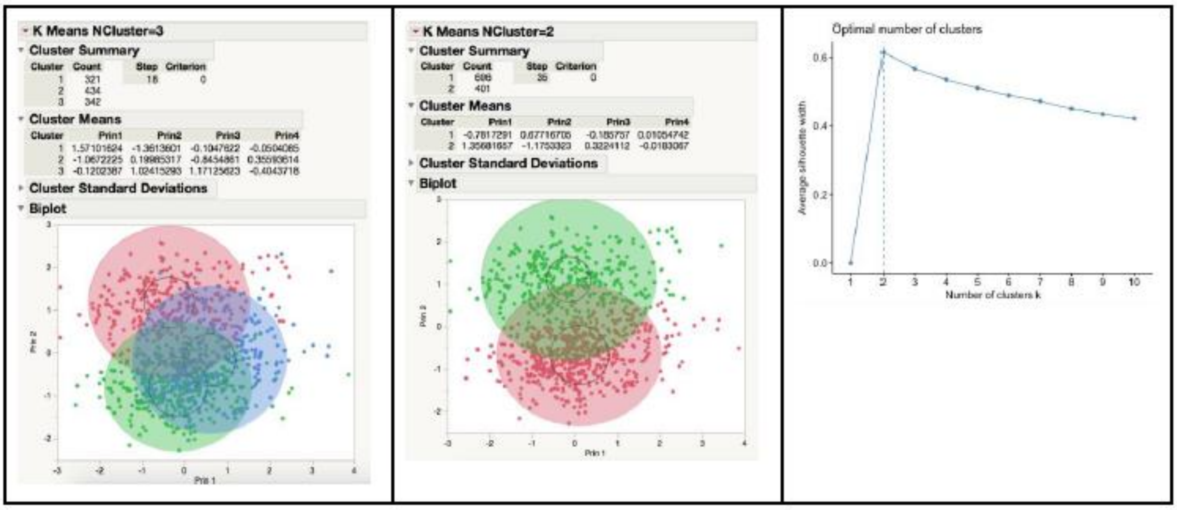
**Correlation Analysis:**
The correlation table below shows that some of the attributes have ±0.5 correlation coefficient with each other, indicating that there may be some underlying linear relationships among these attributes. Hence, PCA is appropriate in this case to identify a set of uncorrelated components to reduce dimension in order to build a better predictive regression model.




**Naming the chosen PCA-s or Factors:**
PC1 - Acidity
Principal component 1 is associated with four variables: citric.acid, fixed.acidity, volatile.acidity and pH value. Citric.acid and fixed.acidity have high positive correlation while volatile.acidity and pH have negative correlation. Citric.acid and fixed.acidity are both measure of acidity so they have high positive correlation. Volatile.acidity is the acid that gives vinegar its aroma and taste, which is less preferred in wines. It should have at least a moderate negative correlation with the preferred type of acidity in wine. Since lower pH value means higher acidity, pH has negative correlation with acidity. According the meanings of the original variables, component 1 represents the acidity of each sample.
PC2 - Fermentation Products
Principal component 2 contains two variables: alcohol and sulphates. Alcohol is the direct product of fermentation and sulphates are derived from the fermentative byproduct sulfites. These two variables co-vary and are combined to form the second component, the fermentation product.
PC3 - Sulfur Dioxide
Principal component 3 also contains two variables: free.sulfur.dioxide and total.sulfur.dioxide, measuring SO2 level of each sample.
PC4 - Flavour

The two variables in principal component 4 are residual.sugar and chlorides. The fourth component denotes “Flavour” since sugar, taste of sweetness, and chlorides, taste of saltiness, contribute to the non-sour taste of the samples.

**Clustering:**
Cluster analysis is used to check for similarity of groupings within all the records. We performed clustering technique on the lower-dimensional dataset using different tools (SPSS, R, JMP), using the K-means method. JMP recommended the three clusters model as the best solution.
As shown above, there is a tremendous overlap between cluster 2 and cluster 3 in the three clusters model, while the two clusters model have similar cluster size and a relatively clearer boundary. Also, R result shows that the two clusters model have the best average silhouette width. Hence, we decide to choose two clusters model and profile the clusters using R.




**Regression :**
After dimension reduction by PCA, a regression model can be built to estimate the relationship between physicochemical features and sensory output (quality). Since quality is a sensory result graded by different people, it is highly likely to be affected by personal preferences and experience. We need to reduce the variation in the quality score due to such subjectivity. In order to do so, the target variable, quality scores, is divided into 3 groups: low, medium, and high, as shown below.

| Quality Score | Quality Level |
|---|---|
| 3-4 | Low |
| 5-6 | Medium |
| 7-8 | High |

**Conclusion:**
By conducting PCA, cluster analysis, and regression model fitting on red wine quality dataset, we have discovered the relationship between physicochemical features of wine and taster’s sensory quality score. Physicochemical features of wine can be summarized into four main aspects: Alcohol content, Acidity, Flavour, and SO2 content. 2 clusters are formed accordingly. The Good Wine cluster is characterized by its high acidity and alcohol concentration level, while the Average Wine cluster contains wines with lower acidity and alcohol. In the final regression model, all of the explanatory variables are significant in making prediction. We are able to predict the quality ranking with 86.5% accuracy and are able to classify the majority of medium ranked wines correctly. In addition, we have discovered some limitations with the current dataset that may affect the applicability of our findings. It is important to note that the tasting of wine is highly correlated with personal likings and the result can vary greatly due to immeasurable factors.

**EBAC Group 4**

**Objective:**
A sample of the Singapore HDB resale flat transaction data from the past 2 years was analysed using the Principle Component Analysis (PCA), K Means Clustering and Linear Regression. The sampled dataset contained 4,800 records with each record having 16 variables. During the data preparation

stage, the dataset was scrubbed for any data integrity issues. From PCA, eight of the variables were processed and two components were extracted. Next, three clusters were then obtained from the two factors identified through K Means Clustering. Lastly, linear regression was applied to predict the transaction price (price per square metre) of the flat based on the two components obtained from PCA and other categorical variables.

**Data Source & Description:**
The price trend of public housing had always been a hot topic of discussion among Singaporeans. Approximately 80% of the Singapore resident population lived in public housing, of which 90% own their home. The "Singapore HDB flat resale transaction" dataset contained the HDB resale transaction information from the past two years (2016 to 2018) with a total of 4,800 records. Each record represented a completed HDB resale transaction. The initial dataset contained 16 variables as shown Table 1, and detailed in Table 2. Data cleaning was done to remove any missing or incomplete record. Through running the Data Audit node on SPSS Modeller, no major data quality issues were observed.

Table 1: List of variables

| No. | Variable | Description |
|---|---|---|
| 1 | Transaction Time | Transaction Settlement Date dd/mm/yyyy |
| 2 | LEASE START | Time when the 99-year HDB lease starts |
| 3 | AGE | Age of the flat |
| 4 | STOREY_No. | Storey number of the unit |
| 5 | FLOOR AREA | Floor Area of the unit |
| 6 | PRICE_Total | Total transaction price |
| 7 | PRICE_PSM | transaction price per square meter |
| 8 | Town | HDB town location |
| 9 | No. of Room | Number of rooms in the flat |
| 10 | BLCK_AVERAGE_SIZE | Average floor area for the block of building |
| 11 | BlCK_LOWEST_PRICE | Lowest sale price for the block of building (in past one year) |
| 12 | BLCK_AVERAGE_PRICE | Average sale price for the block of building (in past one year) |
| 13 | BLCK_HIGHEST_PRICE | Highest sale price for the block of building (in past one year) |
| 14 | BLCK_LOWEST_RENTAL | Lowest rental price for the block of building (in past one year) |
| 15 | BLCK_AVERAGE_RENTAL | Average rental price for the block of building (in past one year) |
| 16 | BLCK_HIGHEST_RENTAL | Highest rental price for the block of building (in past one year) |

**Correlation Analysis: N/A**

**Naming the chosen PCA-s or Factors:**
Input Variable Selection
A scatter-plot matrix (as shown in Fig 1) was plotted to uncover any multi-collinearity trends. From the scatter plot matrix, it was observed that the variable LEASE START and AGE have strong correlation. This was not surprising as the start of the lease was usually just after the flat has been completed. Hence only the variable AGE was selected. Next the three independent variables, BLCK_LOWEST_PRICE, BLCK_AVERAGE_PRICE and BLCK_HIGHEST_PRICE were found to have very strong linear correlation with one another, thus only BLCK_AVERAGE_PRICE was selected while the other two were dropped. Next, we found that the variable STOREY_No. has low communality on the extracted components, thus this variable was also excluded. The remaining eight variables
were inputted into the PCA process.

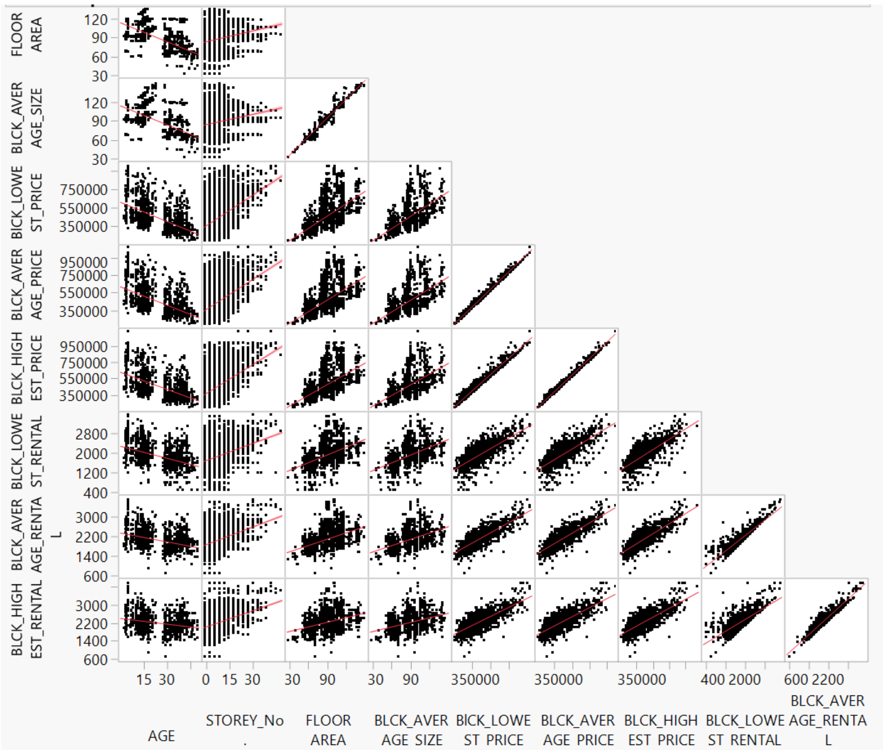
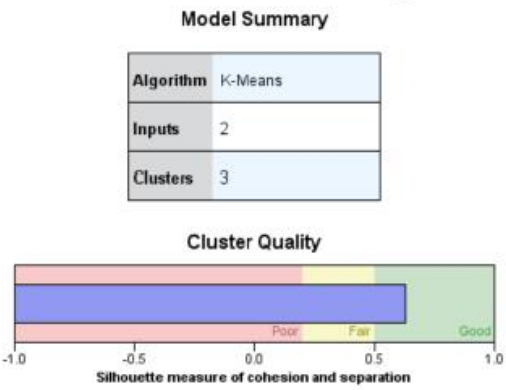


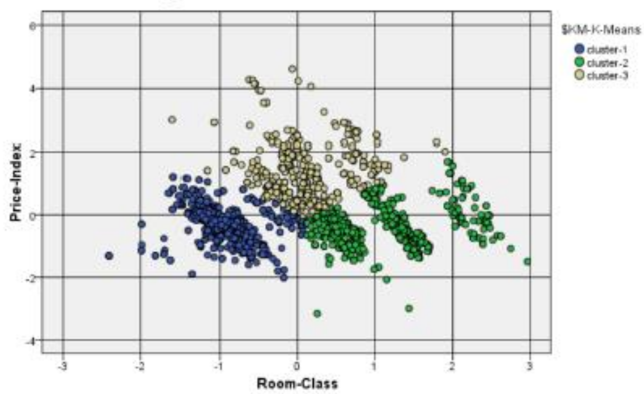

Fig1: Scatter Plot Matrix of Independent variables

**Clustering:**

K-Means clustering is used in this application, as the input variables are continuous. To reduce the complexity and produce meaningful cluster, the clustering step only use the two derived components (Unit-Features and Price-Features) from the PCA process in section 3. The number of clusters are set to 3 after multiple trials to obtain the most meaningful clusters.

The clustering result is summarised in Fig 4 below. Cluster 1 has both low Price-Index and Unit-Features, Cluster 2 has high Unit-Features but medium Price-Index, whereas Cluster 3 has highest Price-Index and medium Unit-Features.




**Regression :**

The objective of the regression analysis is to predict Price_PSM based on the derived factors and other variables. Generalized Linear Model Node was used in SPSS modeler for the regression analysis in order to take in categorical input variables. Since Price_PSM was close to normal distribution, Normal Distribution and Identity Link Function were selected in model parameter.

**Parameter Estimates**

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | |
|---|---|---|---|---|---|---|---|
| | | | Lower | Upper | Wald Chi-Square | df | Sig. |
| (Intercept) | 3858.288 | 59.6896 | 3741.298 | 3975.278 | 4178.219 | 1 | .000 |
| STOREY_No. | 58.294 | 1.5052 | 55.344 | 61.244 | 1499.902 | 1 | .000 |
| [Town=AMK ] | 965.275 | 64.6858 | 838.493 | 1092.057 | 222.682 | 1 | .000 |
| [Town=Bedok ] | 462.625 | 67.5559 | 330.218 | 595.032 | 46.896 | 1 | .000 |
| [Town=Bishan ] | 1586.212 | 86.1484 | 1417.364 | 1755.060 | 339.023 | 1 | .000 |
| [Town=Bukit Merah] | 1563.154 | 75.7505 | 1414.686 | 1711.622 | 425.826 | 1 | .000 |
| [Town=Bukitbatok] | 317.488 | 78.5020 | 163.627 | 471.350 | 16.357 | 1 | .000 |
| [Town=CCK ] | -241.149 | 77.2152 | -392.488 | -89.810 | 9.754 | 1 | .002 |
| [Town=Central ] | 1563.813 | 91.1556 | 1385.152 | 1742.475 | 294.309 | 1 | .000 |
| [Town=Clementi ] | 970.950 | 79.5308 | 815.072 | 1126.827 | 149.047 | 1 | .000 |
| [Town=Hougang ] | 495.789 | 77.3184 | 344.248 | 647.330 | 41.118 | 1 | .000 |
| [Town=Jurongeast] | 382.487 | 73.3777 | 238.670 | 526.305 | 27.171 | 1 | .000 |
| [Town=Jurongwest] | -83.687 | 67.2241 | -215.444 | 48.070 | 1.550 | 1 | .213 |
| [Town=Kallang ] | 792.997 | 69.8854 | 656.024 | 929.970 | 128.757 | 1 | .000 |
| [Town=Parsirris ] | -222.929 | 134.5344 | -486.611 | 40.754 | 2.746 | 1 | .098 |
| [Town=Pungol ] | 9.879 | 66.7133 | -120.876 | 140.635 | .022 | 1 | .882 |
| [Town=Queenstown] | 1595.420 | 73.4846 | 1451.393 | 1739.447 | 471.365 | 1 | .000 |
| [Town=Sembawang ] | -204.635 | 74.3564 | -350.371 | -58.900 | 7.574 | 1 | .006 |
| [Town=Sengkang ] | 500.641 | 66.1472 | 370.995 | 630.287 | 57.283 | 1 | .000 |
| [Town=Serangoon ] | 1143.973 | 81.0936 | 985.033 | 1302.914 | 199.002 | 1 | .000 |
| [Town=Tampines ] | 511.857 | 76.2534 | 362.403 | 661.311 | 45.059 | 1 | .000 |
| [Town=ToaPayoh ] | 961.840 | 70.0994 | 824.447 | 1099.232 | 188.268 | 1 | .000 |
| [Town=Woodlands ] | -102.006 | 69.6213 | -238.461 | 34.449 | 2.147 | 1 | .143 |
| [Town=Yishun ] | 0[a] | . | . | . | . | . | . |
| Unit-Features | 3.049 | 14.6011 | -25.569 | 31.667 | .044 | 1 | .835 |
| Price-Index | 608.525 | 13.1849 | 582.683 | 634.367 | 2130.101 | 1 | .000 |
| (Scale) | 338493.907[b] | | | | | | |

Dependent Variable: PRICE_PSM
Model: (Intercept), STOREY_No., Town, Unit-Features, Price-Index

a. Set to zero because this parameter is redundant.

b. Computed based on the deviance.

**Conclusion:**
The results from the model output was consistent with the general factors associated with the resale value of a HDB flat. Keeping certain features like size, storey number and age constant, the location was a significant variable that determined the price. This made sense as flats located in prime districts generally would be able to fetch higher price when all other factors are kept constant.

**EBAC Group AbracaDATA**

**Objective:**
The main aim of this project is to identify whether the patient is diagnosed with Chronic kidney disease (CKD) or not. There are ways like Kidney Function Test (KFT) that measure various indicators such as Serum Creatinine, helpful in identifying CKD patients. However, taking such large number of predictors makes visualization and modelling complex. Hence, we break down the problem in 3 stages, as below:

Reduce Dimensionality using Principal Component Analysis and select suitable number of obtained orthogonal variables for profiling

Perform suitable number of clusters using K-means completed with their description for better understanding of data

Finally, use logistic regression to classify whether the patient has Chronic Kidney disease

**Data Source & Description:**

The dataset, which is publicly available for research, is related to Chronic Kidney disease is taken from:https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. The dataset consists of 25 (11 numeric,14 nominal) attributes includes one class variable classification. The 24 health related attributes were taken in 2-month period of 402 patients in Tamil Nadu, India.

| Attribute | Name | Type | Description |
|---|---|---|---|
| *age* | Age | Num | Age of the patient at the time of testing |
| *bp* | Blood Pressure | Num | The pressure of circulating blood on the walls of blood vessels |
| *sg* | Specific Gravity | Cat | Measure of the concentration of solutes in the urine. Values [1.005,1.010,1.015,1.020,1.025] |
| *al* | Albumin | Cat | Family of Globular Protein, Values [ 0,1,2,3,4,5] |
| *su* | Sugar | Cat | Blood sugar level range, Values [ 0,1,2,3,4,5] |
| *rbc* | Red Blood Cells | Cat | Type of blood cells carrying Oxygen, Values [normal, abnormal] |
| *ps* | Pus Cells | Cat | Body's immune system to fight infection, Values [normal, abnormal] |
| *pcc* | Pus Cell Clumps | Cat | Pus Cell detected in Urine, Values [present, not present] |
| *ba* | Bacteria | Cat | Whether Bacteria present in Urine |
| *bgr* | Blood Glucose Random | Num | Blood Glucose level tested at random, should be <= 100 mgs/dl normally |
| *bu* | Blood Urea | Num | Urea content in the blood, in mgs/dl |
| *sc* | Serum Creatinine | Num | Measure for how well kidney filters, in mgs/dl |
| *sod* | Sodium | Num | Sodium level in blood, in mEq/dl |
| *pot* | Potassium | Num | Potassium level in blood, in mEq/dl |
| *hemo* | Haemoglobin | Num | Haemoglobin level in blood, in gms |
| *pcv* | Packed Cell Volume | Num | Volume percentage of RBC in blood |
| *wc* | White Blood Cell Count | Num | Count of WBC, in cells/cumm |
| *rc* | Red Blood Cells | Num | Count of RBC, in millions/cmm |
| *htn* | Hypertension | Cat | Suffering from Hypertension, Values [yes, no] |
| *dm* | Diabetes Mellitus | Cat | Metabolic disorder associated with prolonged high blood sugar levels, Values [yes, no] |
| *cad* | Coronary Artery Disease | Cat | Values [yes, no] |
| *appet* | Appetite | Cat | Desire to eat food, Values [good, poor] |
| *pe* | Pedal Edema | Cat | The accumulation of fluid in the feet and lower legs, Values [yes, no] |
| *ane* | Anaemia | Cat | Condition related to Iron Deficiency, Values [yes, no] |
| *classification* | Whether CKD or Not | Cat | identified with CKD, Values (ckd, notckd) |

Cat: Categorical, Num: Numerical

**Correlation Analysis:**

Here is the correlation matrix which shows there is significant correlation between various input variable.

**Correlations**

| | bgr | bu | sc | sod | pot | hemo | pcv | wc | rc | age | bp |
|---|---|---|---|---|---|---|---|---|---|---|---|
| bgr | 1.0000 | 0.0985 | 0.0872 | -0.2738 | 0.0575 | -0.2470 | -0.2167 | -0.0798 | -0.2530 | 0.1945 | 0.1231 |
| bu | 0.0985 | 1.0000 | 0.6896 | -0.2963 | 0.1870 | -0.5537 | -0.5571 | -0.1020 | -0.5080 | 0.1999 | 0.1693 |
| sc | 0.0872 | 0.6896 | 1.0000 | -0.4707 | 0.1621 | -0.5005 | -0.5114 | -0.1415 | -0.4485 | 0.1356 | 0.1870 |
| sod | -0.2738 | -0.2963 | -0.4707 | 1.0000 | 0.0210 | 0.3888 | 0.3861 | 0.1224 | 0.3231 | -0.0963 | -0.0484 |
| pot | 0.0575 | 0.1870 | 0.1621 | 0.0210 | 1.0000 | -0.1667 | -0.1962 | -0.0974 | -0.1876 | 0.1005 | 0.0627 |
| hemo | -0.2470 | -0.5537 | -0.5005 | 0.3888 | -0.1667 | 1.0000 | 0.8978 | 0.3511 | 0.7655 | -0.1603 | -0.2702 |
| pcv | -0.2167 | -0.5571 | -0.5114 | 0.3861 | -0.1962 | 0.8978 | 1.0000 | 0.3298 | 0.7687 | -0.2154 | -0.3018 |
| wc | -0.0798 | -0.1020 | -0.1415 | 0.1224 | -0.0974 | 0.3511 | 0.3298 | 1.0000 | 0.1860 | -0.0644 | -0.0894 |
| rc | -0.2530 | -0.5080 | -0.4485 | 0.3231 | -0.1876 | 0.7655 | 0.7687 | 0.1860 | 1.0000 | -0.2545 | -0.2271 |
| age | 0.1945 | 0.1999 | 0.1356 | -0.0963 | 0.1005 | -0.1603 | -0.2154 | -0.0644 | -0.2545 | 1.0000 | 0.1281 |
| bp | 0.1231 | 0.1693 | 0.1870 | -0.0484 | 0.0627 | -0.2702 | -0.3018 | -0.0894 | -0.2271 | 0.1281 | 1.0000 |

**Naming the chosen PCA-s or Factors:**

| *Principal Component* | Loaded On (refer Rotated Factor Loadings) | Name | Description |
|---|---|---|---|
| *1* | Blood Urea (bu), Serum Creatinine (sc), Sodium (Sod), Hemo, Packed Cell Volume(pcv), Red Blood Cell(rc) | **Urine Indicators** | This Component is highly loaded on various indicators which are present in Urine Test |
| *2* | Haemoglobin (Hemo), Packed Cell Volume (pcv), White Blood Cells (wc) | **Blood Indicators** | This is highly loaded on blood test indicators |
| *3* | Blood Glucose Random (BGR), Age | **Aged Diabetic** | This component will have high value for aged and diabetic patients. Note- both age and diabetic affects CKD as it is progressive disease and degrades with time |
| *4* | Hypertension (bp) | **Hypertension** | It only represents people who have high blood pressure |
| *5* | Potassium (pot) | **High Salt in-take** | It represents patients who have high salt intake |

**Clustering:**

We used K-means clustering with K=2 and K=3 (Avg. Silhouette Method) using first two components and first three components separately. » Though we observed 2 clear clusters using Urine Indicator and Blood Indicator but that did not give us any additional information. However, when we used K=3 clusters using first two components, there was a new class of clusters was observed. We calculated mean for various attributes in these cluster groups and the results are presented in the table above. » The Third cluster consists of CKD patients with highly abnormal mean of indicators like blood urea, serum creatinine which indicates these patients are at high risk.

| Cluster | 1 | 2 | 3 | Normal |
|---|---|---|---|---|
| bgr | 116.42 | 187.83 | 151.08 | <=100 mgs/dl |
| bu | 31.90 | 54.84 | 158.21 | 7-20 mg/dl |
| sc | 1.16 | 2.34 | 9.44 | 0.6-1.2 mgs/dl |
| sod | 141.36 | 136.31 | 131.73 | 135-145 mEq/l |
| pot | 4.29 | 4.49 | 4.77 | 3.6-5.2 mEq/l |
| hemo | 14.67 | 10.69 | 8.22 | 13.5-17.5 g |
| pcv | 45.20 | 32.89 | 24.90 | 31-40% |
| wc | 55.14 | 28.99 | 35.34 | 42-55 |
| rc | 5.33 | 4.03 | 3.19 | 4.2-5.4 |
| age | 44.42 | 57.81 | 55.50 | NA |
| bp | 71.68 | 79.76 | 83.80 | 120-80 |

Table 3: Mean Attribute's Values in Clusters

**Regression :**

Depending on the nature of the dataset and the nature of the output predictor variable i.e. dichotomous (we predicted whether a patient is having a chronic kidney disease or not), Logistic Regression was chosen as the preferred method of classification modelling.

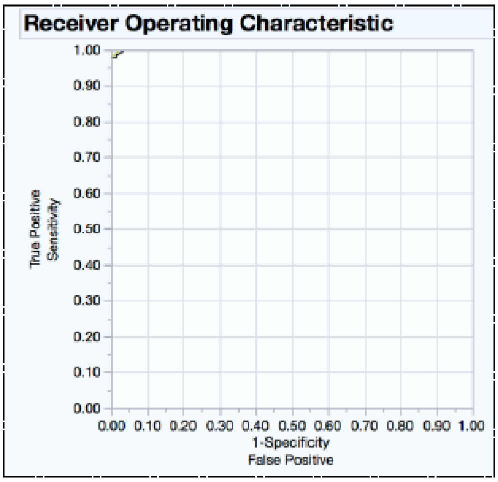
**Effect Summary**

| Source | LogWorth | PValue |
|---|---|---|
| sg | 11.023 | 0.00000 |
| Urine Indicators (PC1) | 8.093 | 0.00000 |
| Blood Indicators (PC2) | 6.648 | 0.00000 |
| al | 6.138 | 0.00000 |
| rbc | 4.855 | 0.00001 |

**Whole Model Test**

| Model | -LogLikelihood | DF | ChiSquare | Prob>ChiSq |
|---|---|---|---|---|
| Difference | 255.85316 | 12 | 511.7063 | <.0001* |
| Full | 9.70915 | | | |
| Reduced | 265.56232 | | | |

| | |
|---|---|
| RSquare (U) | 0.9634 |
| AICc | 46.3564 |
| BIC | 97.3722 |
| Observations (or Sum Wgts) | 402 |




**Conclusion: N/A**

**KE Group DMMM**

**Objective:**

This report has established a predictive model of enterprise bankruptcy to predict the bankruptcy probability. The expectation is to provide signs of bankruptcy to enterprises as soon as possible so that effective measures may be taken to improve enterprises operation as well as protect enterprises from bankruptcy crisis. Since explanation with the prediction model is necessary and all variables are financial attributes involving continuous numerical variables, Factor Analysis is used as the dimensional reduction technique as an intermediate step to investigate the latent factors of defining the financial health of a company. K-means clustering is performed for company profiling and insight discovery. In the end, results of logistic regression were used to generate a bankruptcy prediction model based on the latent factors extracted.

**Data Source & Description:**

To eliminate the influence of industry factors, the data taken in this report is from the typical manufacturing industry of Polish companies collected from Emerging Markets Information Service. The subset of 5th-year data containing 5910 observations with 410 representing companies bankrupted after 1 year and the rest 5500 firms still operating in the period from 2000 to 2013 is chosen for the following model building. There are 64 continuous numerical variables (X-variables) with a binary class (Y-variable). Class 0 indicates companies which did not bankrupt and class 1 indicates companies which bankrupted 1 year later.

**Correlation Analysis: N/A**

**Naming the chosen PCA-s or Factors:**

Factor 1 – Earning and Profit (Profitability)

This factor measures the efficiency of the company in generating earnings and profit. Observing the denominator of the variables with high contribution to this factor, majority of them are ‘total assets’ and ‘sales’. Most numerators involve gross profit, EBITDA, net profit, etc., which are different ways of calculating earnings of the company. Popular profitability indicators like profit margin and Return on Assets (ROA) are included in this factor. Attr58 (total costs/total sales) contributes negatively to this factor and is align with our naming that higher cost reduces profit and therefore correlates negatively with profitability. Hence, this form our basis of naming this factor as ‘Profitability’. A higher value indicates that the company is more efficient in generating profit and thus has higher profitability.

Factor 2 – Ability to Cover Liabilities (Debt Ratio)

The main contributing variables in this factor comprises ratios of asset, equity and liabilities, which are the components making up the accounting equation: Assets = Equity + Liabilities. Looking at the positive contributors, they are all subsets of ratio between equity to assets and assets to liabilities. In fact, the ratio of equity to assets could be written as (1-ratio of liabilities to assets). This aligns to the negative contributors of this factor, which are subsets of the ratio of liabilities to assets. These variables are trying to assess the financing situation of a company. A high debt ratio means most of the assets are financed through debt and the risk is higher with a lowered borrowing capacity.

Factor 3 – Value of Fixed Asset (Fixed Asset Ratio)
The asset of the company could be coarsely divided into those that are liquid (current assets) and illiquid (fixed assets). These fixed assets are long-term tangible assets and are not expected to be liquidated in the short-term. The denominators of the main contributors are fixed assets, while the numerators are attributes that correlates positively to current assets. Hence, we could generalize this factor as the ratio between the current and fixed assets. Higher value means that fixed assets are being utilized efficiently or the company has very few equipment and normally outsource its operations.
Factor 4 – Inefficiency in Collecting Cash
The top 2 contributors are Attr44 (receivables*365)/sales) and Attr61 (sales/receivables), which are the inverse of each other (Attr44 unit in days). Attr61 is essentially the receivables turnover ratio and negatively contributes to this factor. Hence this factor is measuring the inefficiency of the company in collecting cash.

**Clustering:**

| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|
| General Character | Cash-rich Companies | FMCG Companies | Poor Operation Companies | Sustaining Companies | Seasonal Product Companies |
| Description | Medium debt ratio<br>Very solvent<br><br>=> Stable in generating sales and profit<br>=> Good performance to pay long-term debt | High Debt Ratio<br>Low excess inventory<br>Relatively efficient in collection cash<br><br>=>Promising from the bank view<br>=> Bad performance to pay long-term debt<br>=>Large sales volume<br>=>Weak ability to gain profit but collecting cash quickly | Very low debt ratio<br>Very efficient in generating sales<br>Very low profitability<br><br>=> Lack the ability to gain profit<br>=> Poor credibility from bank's perspective | Low debt ratio<br>Very insolvent<br>Very inefficient in generating sales<br>Relatively low fixed asset<br><br>=> Stable in generating sales and profit<br>=> Bad performance to pay long-term debt | High debt ratio<br>Very high excess inventory<br>Relatively low profitability<br>Average efficiency in generating sales<br><br>=> High inventory<br>=> Bad performance to pay long-term debt |

**Regression :**

$$P=1/(1+e^{-z});$$

Where Z = -1.477*F1-1.173*F2+0.545*F3-0.074*F4-0.507*F5-2.020*F6-0.154*F7-2.123

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Lower | Upper |
| Step 1ᵃ | $F-Factor-1 | -1.477 | .099 | 222.155 | 1 | .000 | .228 | .188 | .277 |
| | $F-Factor-2 | -1.173 | .101 | 134.054 | 1 | .000 | .310 | .254 | .378 |
| | $F-Factor-3 | .545 | .077 | 49.601 | 1 | .000 | 1.725 | 1.482 | 2.008 |
| | $F-Factor-4 | -.074 | .082 | .805 | 1 | .370 | .929 | .791 | 1.091 |
| | $F-Factor-5 | -.507 | .084 | 36.165 | 1 | .000 | .603 | .511 | .711 |
| | $F-Factor-6 | -2.020 | .138 | 214.172 | 1 | .000 | .133 | .101 | .174 |
| | $F-Factor-7 | -.154 | .085 | 3.282 | 1 | .070 | .857 | .725 | 1.013 |
| | Constant | -2.123 | .095 | 498.196 | 1 | .000 | .120 | | |

Figure 5-3 Variables in the Equation

**Conclusion:**
Throughout the deep analysis of business sense of the whole dataset of 5910 Polish bankruptcy enterprises and the multiple trials of several different prediction methods, two relatively strong and

precise models have been derived, i.e. a 4-cluster K-Means clustering model with relatively explicit classification of different kinds of companies and a Logistic Regression model with a relatively high accuracy in the acceptable tolerance of the overall error rate. With the clustering model, companies can be classified into different profiles with distinctive characteristics, while in the use of Logistic Regression model sign or warning of bankruptcy could be provided to enterprises.

## KE Group Food

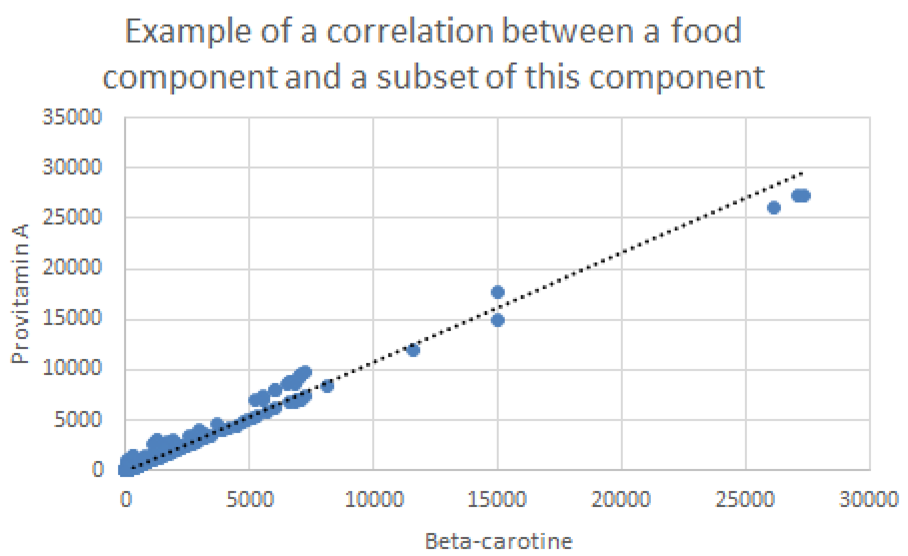
**Objective:**

During the initial data exploration, our team have noticed that there are a huge number of raw ingredients in the dataset itself. Examples of the raw ingredients are baking soda, butter, sugar, Iodine. Since our objective is to classify food into different groups, our team have decided to remove food products that are inedible on its own. Also, we have also removed all forms of alcoholic drinks in the dataset as well as we do not consider alcohol as part of the food group. Last but not the least, we have removed heavily nutrition fortified food as well as cooked food as we consider them to be a mixture of ingredients and condiments. This will result in our analysis to have a huge variance and have an impact in our analytical performance.

**Data Source & Description:**

Our group has chosen the challenging task of analysing the different types of food based on their nutrition values and aim to cluster them into groups that are easily understandable for the masses.
The X variables for PCA will be the different nutrition types. The Y/target variable will be the classification of the food types by the Food Standards Australia New Zealand Authorities (FSANWZA).
This dataset is suitable for this project due to the large number of X variables which are well documented in the appendix they had provided. Domain knowledge on Nutrition Types can easily be retrieved online as well. At the same time, there are numerous ways in which foods can be classified. In addition, there can be a lot of interesting relationships between the different food types of which people were not aware. Speaking for our team, we learned a lot about the nutrition value of the foods we were analysing.
The necessary assumptions, steps followed to reduce the dimension, comment on the variability in the data explained by the selected orthogonal variables.

**Correlation Analysis:**




**Naming the chosen PCA-s or Factors:**

| PC No. | Rotated Factor Loadings | | Foods data represented | Description | Name |
|---|---|---|---|---|---|
| 1 | Protein (g) | 0.825 | Beans, Veal, lean meats | Food that contains High Protein, Vitamin B and inorganic nutrients | Nutritious Super Food (Wide spectrum of nutrients with high values) |
| | Tryptophan (mg) | 0.792 | | | |
| | Niacin (B3) (mg) | 0.778 | | | |
| | Vitamin B6 (mg) | 0.713 | | | |
| | Zinc (Zn) (mg) | 0.652 | | | |
| | Phosphorus (P) (mg) | 0.635 | | | |
| | Potassium (K) (mg) | 0.574 | | | |
| | Iron (Fe) (mg) | 0.533 | | | |
| | Magnesium (Mg) (mg) | 0.491 | | | |
| | Selenium (Se) (µg) | 0.451 | | | |
| 2 | Total fat (g) | 0.852 | Meats (mutton, lamb, beef, chicken, pork), cream | Foods that contain high amounts of fats | High energy fat-based food |
| | Total saturated fat (g) | 0.829 | | | |
| | Total trans fatty acids (mg) | 0.823 | | | |
| | Total monounsaturated fat (g) | 0.694 | | | |
| | Alpha-linolenic acid (g) | 0.421 | | | |
| 3 | Alpha-tocopherol (mg) | 0.853 | Nut, Seeds, Almonds | Foods with high amount of Vitamin E compounds | Nuts and seeds |
| | Vitamin E (mg) | 0.843 | | | |
| | Total polyunsaturated fat (g) | 0.752 | | | |
| | Magnesium (Mg) (mg) | 0.577 | | | |
| | Total monounsaturated fat (g) | 0.470 | | | |
| 4 | Available carbohydrates, without sugar alcohol (g) | 0.872 | Chocolate-based foods, banana chips | Food with high levels of carbohydrates and Low water content | High-sugar energy Carbohydrates |
| | Total sugars (g) | 0.765 | | | |
| | Low moisture | -0.753 | | | |
| | Starch (g) | 0.554 | | | |
| 5 | C22:6w3 Docosahexaenoic (mg) | 0.921 | Fish (Herring, Fish roe, salmon, trout, mackerel, bream, cod) | Food with high level of omega fatty acids | Seafood |
| | C20:5w3 Eicosapentaenoic (mg) | 0.909 | | | |
| | C22:5w3 Docosapentaenoic (mg) | 0.640 | | | |
| | Selenium (Se) (µg) | 0.424 | | | |

| | | | | | |
|---|---|---|---|---|---|
| 6 | Total Folates (µg) | 0.738 | All kinds of bread (wholemeal, white, mixed grain, breadcrumbs), Poppadum | Foods that have high in folates and iodine. | Starchy Carbohydrates Food |
| | Iodine (I) (µg) | 0.637 | | | |
| | Alpha-linolenic acid (g) | 0.552 | | | |
| | Starch (g) | 0.542 | | | |
| | Dietary fibre (g) | 0.418 | | | |
| 7 | Vitamin B12 (µg) | 0.767 | Kidneys, heart, oyster, eggs | Animal organs, animal-based food with high cholesterol | High Cholesterol Food |
| | Riboflavin (B2) (mg) | 0.721 | | | |
| | Cholesterol (mg) | 0.555 | | | |
| | Preformed vitamin A (retinol) (µg) | 0.526 | | | |
| | Selenium (Se) (µg) | 0.482 | | | |
| 8 | Sodium (Na) (mg) | 0.891 | Soup (chicken, pea & ham, French onion, tomato), plum(salted) | Foods that have are high in sodium and inorganic minerals. | Salty Food |
| | Ash (g) | 0.841 | | | |
| | Thiamin (B1) (mg) | 0.596 | | | |
| 9 | Calcium (Ca) (mg) | 0.873 | Cheese (parmesan, mozzarella, cheddar) | Processed dairy product | Processed dairy product |
| | Phosphorus (P) (mg) | 0.458 | | | |
| 10 | Vitamin C (mg) | 0.696 | Lime, Parsley, Tomato, Chilli, sweet potato, juice carrot, cabbage | Vegetables rich in beta-carotene | Vegetables |
| | Beta-carotene (µg) | 0.638 | | | |
| | Potassium (K) (mg) | 0.414 | | | |
| 11 | Caffeine (mg) | 0.900 | Coffee (espresso, cappuccino, latte), Tea | Foods with caffeine contents | Coffee and Tea |

**Clustering:**

| Cluster | Top Most Prominent Factors | Foods (quantity) | Description | Name of Cluster |
|---|---|---|---|---|
| Cluster 1 | Factor 10 (Vegetables)<br>Factor 3 (Nuts and Seeds)<br>Factor 1 (Nutritious Super Food) | Nuts (15)<br>Breakfast Cereal (9)<br>Potato Chips (8)<br>Bean (5)<br>Egg (4)<br>Bread (4) etc. | High Energy & Nutritious food | Energy Food |
| Cluster 2 | Factor 4 (Hi Energy Carbohydrates Food)<br>Factor 10 (Vegetables)<br>Factor 5 (Hi Omega Fatty Acids Food) | Bread (136)<br>Cake (133)<br>Biscuit (114)<br>Muffin (27)<br>Chocolate (21)<br>Pie (18)<br>Confectionery (14)<br>Pastry (14) etc. | Starchy & High Carbohydrates Food | Carbs |
| Cluster 3 | Factor 4 (Hi Energy Carbohydrates Food)<br>Factor 2 (Hi Energy Fat-based Food)<br>Factor 1 (Nutritious Super Food) | Coffee (48)<br>Cabbage (4)<br>Chilli (4)<br>Juice (4) etc. | High Caffeine Food | Food that keeps you alert |
| Cluster 4 | Factor 4 (Hi Energy Carbohydrates Food)<br>Factor 1 (Nutritious Super Food)<br>Factor 6 (Starchy Carbohydrates Food) | Beef (70)<br>Lamb (57)<br>Pork (35)<br>Chicken (23)<br>Veal (20) etc. | Meat | Meat |
| Cluster 5 | Factor 4 (Hi Energy Carbohydrates Food)<br>Factor 2 (Hi Energy Fat-based Food)<br>Factor 7 (Hi Cholesterol Food) | Cheese (58)<br>Cream (12)<br>Beef (10)<br>Lamb (10)<br>Pork (8) etc. | Processed Dairy Products | Dairy Products |
| Cluster 6 | Factor 1 (Nutritious Super Food)<br>Factor 2 (Hi Energy Fat-based Food)<br>Factor 4 (Hi Energy Carbohydrates Food) | Soup (60)<br>Yoghurt (56)<br>Juice (39)<br>Ice Cream (35)<br>Milk (34)<br>Cordial (29)<br>Sushi (18)<br>Apple (17)<br>Pear (10) etc. | High Moisture foods (Soups, Vegetable, & Fruits) | Food with high water content |

**Regression :**

| | P. 1 | P. 2 | P. 3 | P. 5 | P. 6 | P. 8 | P. 9 | P. 10 | P. 11 | P. 12 | P. 13 | P. 15 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Act. 1 | 161 | 11 | 0 | 0 | 11 | 0 | 15 | 0 | 0 | 0 | 9 | 1 | 208 |
| Act. 2 | 2 | 663 | 0 | 2 | 0 | 1 | 7 | 11 | 0 | 3 | 13 | 1 | 703 |
| Act. 3 | 1 | 3 | 9 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 |
| Act. 5 | 0 | 10 | 0 | 64 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 76 |
| Act. 6 | 18 | 1 | 0 | 0 | 92 | 0 | 0 | 1 | 0 | 0 | 5 | 0 | 117 |
| Act. 7 | 2 | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 6 |
| Act. 8 | 0 | 1 | 0 | 0 | 0 | 342 | 0 | 1 | 0 | 0 | 2 | 0 | 346 |
| Act. 9 | 6 | 11 | 0 | 0 | 1 | 0 | 249 | 1 | 0 | 0 | 0 | 0 | 268 |
| Act. 10 | 3 | 12 | 0 | 1 | 6 | 1 | 0 | 66 | 0 | 1 | 27 | 0 | 117 |
| Act. 11 | 2 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 29 | 0 | 3 | 0 | 36 |
| Act. 12 | 3 | 3 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 41 | 0 | 0 | 51 |
| Act. 13 | 18 | 8 | 0 | 0 | 14 | 2 | 1 | 5 | 2 | 0 | 124 | 1 | 175 |
| Act. 15 | 0 | 0 | 0 | 3 | 2 | 2 | 0 | 0 | 0 | 0 | 6 | 2 | 15 |
| Total | 216 | 726 | 9 | 71 | 127 | 349 | 276 | 86 | 33 | 45 | 189 | 5 | 2132 |

**Conclusion:**
Our team have successfully completed the project. We have found and selected appropriate data sets which have a good number of explanatory variables suitable for dimension reduction. The dataset also have suitable Y variable as target for regression as well. In addition, we have also completed dimension reduction successfully with PCA and is able to logically describe and name the individual components. With the new variables created using PCA, we have also completed clustering successfully, generating six logical clusters that we are able to name and describe. Finally we also have completed regression analysis and we were able to predict with high accuracy of 86% in our regression analysis to determine to which food group that a food correctly belongs to.

## KE Group DIMENSION FOUR

**Introduction: N/A**

**Data Source & Description:**
The data set was derived from a Kaggle competition data set here. The data set contains data that describes member details, listening habits, transactions and churn statistics of a Taiwanese music streaming service called KKBox. KKBox offers subscription based music streaming service. The original data set contained 4 CSV files: transactions.csv, user_logs.csv, members.csv and train.csv. The transactions.csv contains monetary transactions performed by KKBox members to subscribe or renew their subscriptions to a service provided by KKBox. The user_logs.csv contains daily user logs describing listening behavior of the users. The members.csv contains the details of the members and the train.csv contains whether a user had churned or not. Due to the volume of the data, Python

scripts were developed to join the records in the different files mentioned above on the user id. Since there were multiple logs per user, average of all continuous values were considered in the combined data for further processing.

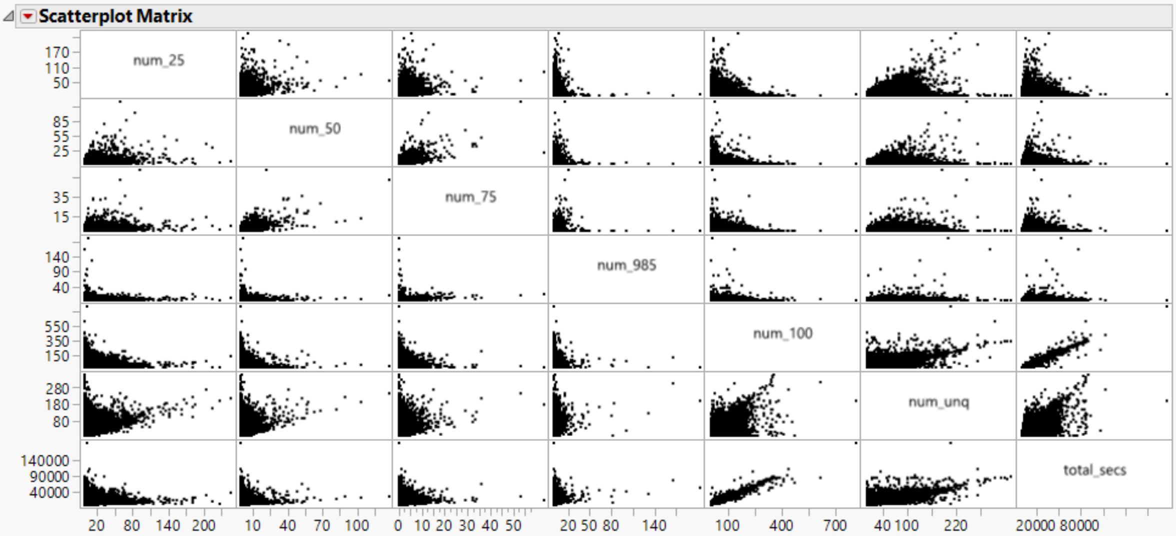
**Correlation Analysis:**

**Multivariate**

**Correlations**

| | num_25 | num_50 | num_75 | num_985 | num_100 | num_unq | total_secs |
|---|---|---|---|---|---|---|---|
| num_25 | 1.0000 | 0.4959 | 0.4542 | 0.2340 | 0.0898 | 0.4248 | 0.1388 |
| num_50 | 0.4959 | 1.0000 | 0.6482 | 0.2309 | 0.0802 | 0.3298 | 0.1454 |
| num_75 | 0.4542 | 0.6482 | 1.0000 | 0.3068 | 0.1266 | 0.3337 | 0.1947 |
| num_985 | 0.2340 | 0.2309 | 0.3068 | 1.0000 | 0.1266 | 0.2678 | 0.2098 |
| num_100 | 0.0898 | 0.0802 | 0.1266 | 0.1266 | 1.0000 | 0.7753 | 0.9832 |
| num_unq | 0.4248 | 0.3298 | 0.3337 | 0.2678 | 0.7753 | 1.0000 | 0.8087 |
| total_secs | 0.1388 | 0.1454 | 0.1947 | 0.2098 | 0.9832 | 0.8087 | 1.0000 |

**Scatterplot Matrix**




**Naming the chosen PCA-s or Factors:**

**Unrotated Factor Loading**

| | Factor 1 | Factor 2 |
|---|---|---|
| num_25 | 0.135952 | 0.637634 |
| num_50 | 0.138693 | 0.772087 |
| num_75 | 0.187061 | 0.740367 |
| num_985 | 0.196721 | 0.359764 |
| num_100 | 0.986576 | -0.078453 |
| num_unq | 0.808271 | 0.308160 |
| total_secs | 0.997367 | 0.010019 |

**Rotated Factor Loading**

| | Factor 1 | Factor 2 |
|---|---|---|
| num_25 | 0.0686014 | 0.6483474 |
| num_50 | 0.0572822 | 0.7823507 |
| num_75 | 0.1086994 | 0.7558565 |
| num_985 | 0.1580639 | 0.3783452 |
| num_100 | 0.9893740 | 0.0250332 |
| num_unq | 0.7716588 | 0.3909056 |
| total_secs | 0.9908644 | 0.1141484 |

**Clustering:**

Cluster 1: Customers who are heavy users of the KKBox platform and need to be provided special attention. They are also less in number.

Cluster 2: The major number of customers of KKBox service are in this cluster. Their overall consumption of songs in the platform is less.

Cluster 3: These are users who have high overall listening summary but low interval summary. This shows they prefer to finish the songs that they have started to listen. These may comprise of businesses that play KKBox songs in the background. So, this cluster can be leveraged for more ads or no ads premium services as per KKBox's plans.

Cluster 4: These are the users who consume high overall content as well as scan through several songs without listening to them completely. These customers can be considered for long term and high value plans.

**Regression :**

**Confusion Matrix**

Training

| Actual | Predicted Count | |
|---|---|---|
| **is_churn** | **0** | **1** |
| 0 | 28236 | 190 |
| 1 | 1252 | 1464 |

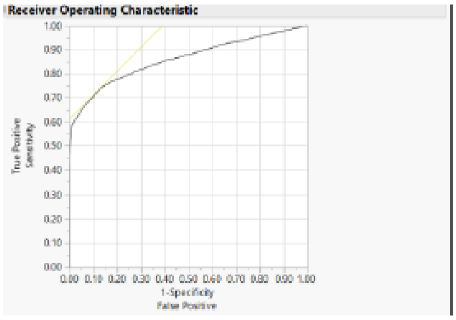



*Figure 14: Confusion matrix and ROC for Logistic regression with unbalanced data*

**Conclusion:**

The below summarizes the limitations of the findings for this exercise:

- Due to the data balancing performed to get a higher specificity, which reduced the amount of training data, the resulting model may be an overfitting model.
- Since only the principal components were considered for the KMeans clustering, clusters that would include the categorical variables were not explored.

**KE Group FMA MUSIC ANALYSIS:**
**GENRE PREDICTION**

**Objective:**

This report serves to showcase the usefulness of PCA by applying it on a real-world dataset - FMA (Free Music Archive). The rotated loading matrix from the technique (varimax rotated PCA) is used to provide domain relevant names to orthogonal variables. The following independent sub-tasks are also considered after generation of the principal components: i. Using a clustering technique with the reduced feature set to profile music tracks, ii. Performing nominal regression analysis using the reduced number of orthogonal variables and analyzing the prediction quality using standard metrics. All related experiments and analysis are conducted using the software - JMP Pro 13.

**Data Source & Description:**

The dataset under consideration is a dump of the Free Music Archive (FMA), an interactive library of high-quality, legal audio downloads [1] The dataset was created by a group of researchers at EPFL, Switzerland. The original dataset contains 781 variables and 13129 observations. Here is the detailed description of the data.

| Index | Column # | Variable Name | Type | Description |
|---|---|---|---|---|
| A | 1 | Track id | Continuous | Unique identifier of the song |
| B | 2-13 | Social & Audio features | Continuous | Scores of songs based on various features such as valence, liveness, danceability, tempo, acoustics, energy, and instrumentals. Scores of artists on basis of discovery, familiarity and hotness. |
| C | 14-237 | Temporal Echo nest features [4] | Continuous | It consists calculations of statistical moments of various segments such as pitch, timbre, loudness, derived using Echo nest API [5] |
| D | 238-241 | Album attributes | Various | Consists of album type, album listens, album tracks and album active days |
| E | 242-244 | Artist info | Various | Consists of artist id, artist location |
| F | 245-246 | Set info | Categorical | Column for dataset split and subset. |
| G | 247-263 | Track info | Continuous | Track details such as genre, duration, days active, number of time it was selected as favorite, number of listens. |
| H | 264-515 | Chroma calculations | Continuous | Chroma features represent audio by projecting the entire spectrum onto 12 bins representing the 12 distinct semitones of the musical octave. Different representations of chroma values using various techniques are derived using Librosa package [3] |
| I | 516-655 | MFCC stats | Continuous | Features representing speech in compact form. [6] |
| J | 656-662 | RMSE stats | Continuous | The RMS level is proportional to the amount of energy over a period of time in the signal. |
| K | 663-774 | TONNETZ stats | Continuous | A planar representation of pitch relations showing harmonic relationships in European classical music. [7] |
| L | 775-781 | ZCR stats | Continuous | Rate at which the signal changes from positive to negative or back. |

**Correlation Analysis: N/A**

**Naming the chosen PCA-s or Factors:**

| Prin | Profile | Description |
|---|---|---|
| **PC-1** | Temporal features | This component is highly loaded with 3 of the temporal features (120,194,210) |
| **PC-2** | High frequency temporal features | This component represents three temporal features (200,202,216) whose eigen vectors lie along one direction and the vector of the mfcc component(mfcc_skew_1) is in the opposite direction |
| **PC-3** | Measure of Low frequency instrumental sounds | This component is loaded with speech recognition components whose eigen vectors lie along one direction (mfcc_median_6 & mfcc_mean_6) and vector of acousticsness component which is a measure of usage natural musical sounds. Lower the value of the acousticsness, it means usage of electronic instruments rather than natural sounds. |
| **PC-4** | Measure of Human voice | This component is highly dependent low frequency speech components (mfcc_mean_2 & mfcc_mean_median 2) and also the acousticsness which is basically natural sounds (like human voice) |

| | | |
|---|---|---|
| **PC-5** | Measure of Low frequency natural sounds | This component is highly dependent on spectral contrast mean 4 which explains the frequency band of the sound and the acousticsness variable. It inversely varies with the energy variable. |
| **PC-6** | Track dance ability | It is highly loaded with two low frequency speech components of mfcc(mfcc_skew_1 & mfcc_std_6), 1 temporal feature(123) & also highly correlated with danceability of the audio feature |
| **PC-7** | Track Popularity | This component is highly loaded with track interest, track listens and artist favorites. On the whole, this explains the popularity of a track |
| **PC-8** | Artist Popularity | This consists of artist hotness and familiarity. |
| **PC-9** | Track energy | This component highly varied by the values of rmse_max_1 and rmse_mean_1 which is a function of energy component of the audio. |
| **PC-10** | Measure of high frequency sounds | This component is inversely dependent on spectral_contrast_1, temporal feature_99 and mfcc_std_6 which makes it a measure of high frequency range and also directly varied by mfcc min 4 |
| **PC-11** | track emotion | This component mostly influenced by n the valence, danceability and the temporal feature_99. Valence explains the emotions of the audio in the scale of sad to happy |
| **PC-12** | Instrumental usage | This is directly proportionated to instrumentalness and temporal feature_99 and inversely proportional to low frequency mfcc component (mfcc_std6) and the temporal feature_100 |
| **PC-13** | track lifespan | It is heavily loaded with variables track days active and album days active |
| **PC-14** | Album Size | It is heavily loaded with variables like album tracks, track number and negatively correlated to album days active |

**Clustering:**

Cluster 1: The mean of the variables like audio_danceability, audio_energy, track interest and track listens are very high in this cluster. Profile: Upbeats

Cluster 2: The mean of audio_acousticness is high which means tracks which doesn't use electronic musical instruments come under this group. Low valence signifies, songs revolve around sadness which will have less energy and danceability. Profile: Underrated Blues

Cluster 3: This cluster has higher means for artist_familiarity, artist_hotness and artist_favorites. The tracks under this cluster are just listened and liked because of the popularity of the artist. Profile: Celebrity

**Regression :**

## Nominal Logistic Fit for track_genre

### Effect Summary

| Source | LogWorth | PValue |
|---|---|---|
| Prin1 | 339.254 | 0.00000 |
| artist_location | 170.035 | 0.00000 |
| Prin11 | 71.585 | 0.00000 |
| Prin6 | 57.652 | 0.00000 |
| Prin2 | 56.077 | 0.00000 |
| Prin5 | 40.738 | 0.00000 |
| Prin12 | 31.458 | 0.00000 |
| Prin14 | 29.721 | 0.00000 |
| Prin13 | 26.479 | 0.00000 |
| Prin4 | 25.081 | 0.00000 |
| Prin3 | 13.324 | 0.00000 |
| Prin7 | 12.161 | 0.00000 |
| Prin8 | 10.475 | 0.00000 |
| Prin9 | 5.811 | 0.00000 |
| Prin10 | 5.108 | 0.00001 |

Remove Add Edit FDR

**Conclusion:**

From PCA Analysis

Since the data had high number of dimensions, PCA helped us to reduce the complexity and facilitated us to understand correlation between various attributes in the data. We can also safely conclude how PCA captures variance across the attributes without much information loss and hence doesn't hamper the performance of the model (tradeoff between performance and reducing the dimensions is very small)

From Cluster Analysis

It enabled us to identify distinct groups in the data and observe their characteristic. For instance, we identified the group with high celebrity value which consisted of tracks which were followed because of the popularity of an artist.

From Modelling

We conclude that for determining genre of a song, artist location and temporal echo nest features(PC1) plays a very important role along with a track's dance ability (PC 6) and its acoustic scores.