SUMMARY REPORT

SIVATEJA PATHIPATI

DATE:- 29-APR-2024

**9) Describing Segment (step 7)**

**9.1) Developing a Complete Picture of Market Segments:**

* It is similar to profiling step. The only difference is that the variable being inspected have not been used to extract market segment.
* Market segments are described using addition information about segment members (instead of variables used in extraction step).
* Segments should be further described (profiling) and typified by crossing them with all other variables (i.e. to avoid nasty surprises).
* Profiling means investigating difference between segments with respect to motive variables (used for extraction step) themselves.
* But the segment description step uses additional information such as segment members variables (like Age, gender and other behaviours) that are typically not used for target segmentation extraction. These additional variables are referred to as **Descriptor Variables.**
* Good description of market segments is critical to gaining detailed insight into nature of segments and segments descriptions are essential for the development of a customised marketing mix
* Two ways to study the difference between market segments with respect to descriptor variables
  + Description statistics including visualization
  + We can analyse data using infernal statistics(statistical testing and tabular presentation of differences in descriptor variables)

**9.2) Using Visualisation to Describe Market Segments:**

* Graphical representations serve to transmit the very essence of marketing research result
* The descriptor variables are divided into
  + Nominal and Ordinal descriptor variables (ex: gender, level of education, etc.)
  + Metric descriptor variables (ex: age, money spent, etc.)

**9.2.1) Nominal and Descriptor Variables:**

* For visualisation we need a cross tabulation of segment membership with descriptor variable
* This is done by adding segment membership as a categorical variable to the data frame of descriptor variables
* Instead of using a stacked bar chart better to use **Mosaic plot**
* In mosaic plot
  + Width of the bar indicates the absolute segment size
  + Each bar has same height but it is divided into multiple slots (rectangles)
  + Height of the rectangle represent portion of each descriptor variable
  + To note the difference, colouring is used based on Pearson residuals(p-value)
  + White dashed border implies lower than expected but not much difference
  + White solid border implies higher than expected but not much difference
  + Red implies lower than expected but with significant difference
  + Blue implies higher than expected but with significant difference

**9.2.2) Metric Descriptor Variables**

* Here conditional plots are used because they are well suited for visualizing difference between market segments using metric descriptor variables
* The visualisation includes Histograms (total n plots for n segments), Boxplots. Whisker plots etc.
* For histograms x-axis is descriptor variable and y- axis segment percentage and total n plots for each segment of n target segments
* Box and whisker plots have y axis as descriptor variable and x – axis as target segments and total of 1 plot
* Segment Level Stability across Solution (SLSa) plot can also be plotted based on descriptor variables

**9.3) Testing for segment difference in Descriptor Variables:**

* The outcome of segment extraction step (step 5) is segment membership, i.e. The assignment of each consumer to one market segment
* The segment membership can be treated like any other nominal variable. It represent a nominal summary statistic of the segmentation variables
* The association between the nominal segment membership variable and another nominal/ordinal variable (like gender, level of education, country of origin, etc) is visualized using cross-tabulation of both variables as basis for the mosaic plot
* – test is one of the test used to test for significant difference between columns and rows of a table
* P-value is also calculated using the parameters of the distribution
* The p- value indicates how likely the observed frequency occurs if there is no association between two variables
* For metric variables **Mean or Median**  is used to check the difference
* The most popular method for testing for significant difference in the **means** of more than two groups is **Analysis of Variance (ANOVA)**.
* ANOVA performs an F-test where the F- value compares the weighted variance between market segment. It means variance within the market segments
* Pairwise comparison between segments provide the information about differing segments
* Method proposed by Holm, whenever a series of tests is computed using the same dataset to assess a single hypothesis, p-values need to be adjusted for multiple testing
* The simplest way to correct p-values for multiple testing is **Bonferroni Correction**. It multiplies all p-values by number of tests computed and represents a very conservative approach
* Other alternative includes False discovery rate procedure proposed by Benjamini et.al, and alternate to pairwise p-tests, we can plot Tukey’s honest significant difference

**9.4) Predicting Segments from Descriptor Variables:**

* Another way of learning about market segments is to try to predict segment membership from descriptor variable
* To achieve this, we use a regression model where our segment memberships are dependent variable (target y) and descriptor variables as independent variables (X). Then we use machine learning models for supervised learning to train our model
* The prediction performance indicates how well members of market segments can be identified given the descriptor variables.
* Linear regression model used for continuous distribution so not mostly used for segment membership (categorical column). So, we use generalized linear models (linear models accompanied by link function), like binary and multinomial logistics.
* In these models the dependent variable follows either a binary or multinomial distribution and the link function is the logit function.

**Binary- Logistic Regression:**

* Where θ is linear function and the function result is between 0 and 1 so it is to be applied on each segment individually

**Multinomial- Logistic Regression:**

* It is applied simultaneously to predict each segments, so it is preferred in prediction of target segments

**Tree Based Methods:**

* The tree approach uses a step wise procedure to fit the model, at each step consumers are split into groups based on the independent variable. The aim of the split is for the resulting group to be as pure as possible with respect to dependent variables
* This approach is also referred as **recursive partitioning.**
* The start node where all customers are present is called root node and the end node where no further partition can be done is called terminal nodes.