Laboratory 2

Data Summarization, Visualization and Pre-Processing

The data provided is based on a fictional Web site—a portal with ecommerce functions (sales of videotapes or DVDs). The data is entirely synthetic and bears no relationship to any actual Web site or company. The data was collected in the period of a week, from 18/02/2001 (Sunday) to 25/02/2001 (Sunday). The films are divided into the following categories: Musical, Western, Comedy, Drama, Sci-Film, Horror, Adventure, Children.

1. Start to load the dataset week_purchases18-25.txt into R session and next:
   - **a.** check its size
   - **b.** get a summary of data
   - **c.** check which columns are numeric and which are categorical
   - **d.** check for missing data by rows and by columns

2. Using the appropriate graphs, analyse the various attributes of the data set and make necessary corrections to the data.:
   - **a.** Title Client
   - **b.** Gender
   - **c.** Country

3. Check:
   - d. Number of films sold by Style
   - e. Number of different Film_Title sold by Style
   - f. Number of units sold for each Film_title / Style
   - g. Number of Customers who bought 1,2,3,4 DVDs
   - h. Create a summary table with Style/Uni_Vnd/Min_Prc/Max_Prk/Avg_Prck/DesvPad_Prd

4. Derive the following new attributes:
   - **a.** Day of the Week Purchase (Sunday, ..., Saturday)
   - **b.** Purchase Period (morning, afternoon, evening)
   - **c.** Price range (low, medium, high)
   - **d.** Age Group (young, middle-aged, elderly)

5. Using the appropriate graphics analyse the novel attributes:
   - **a.** Purchases by day
   - **b.** Purchases by time of day
   - **c.** Price range vs. style
   - **d.** Age Group vs. style
   - **e.** Age Group vs. Price range

6. Review the shopping only to registered customers and make sure that the preferences given by customers confer with products they purchased

   Favourites vs. Style vs. Favourites Buy for