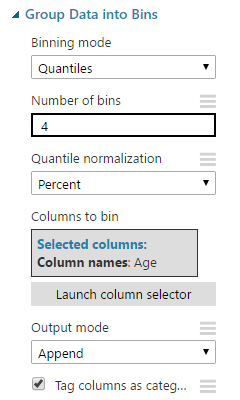
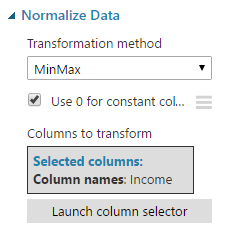
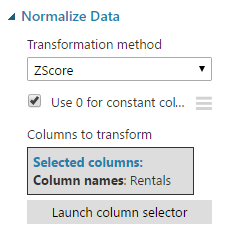
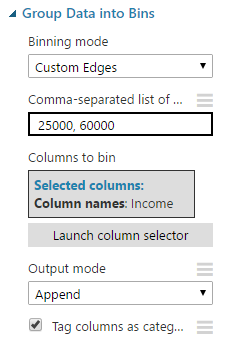
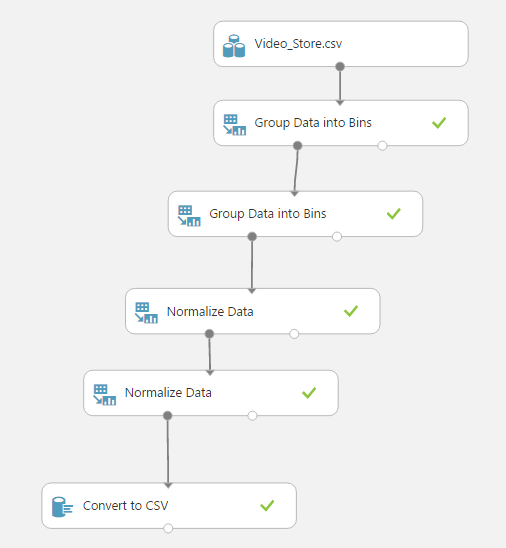
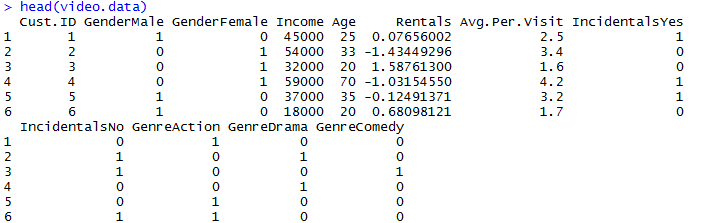
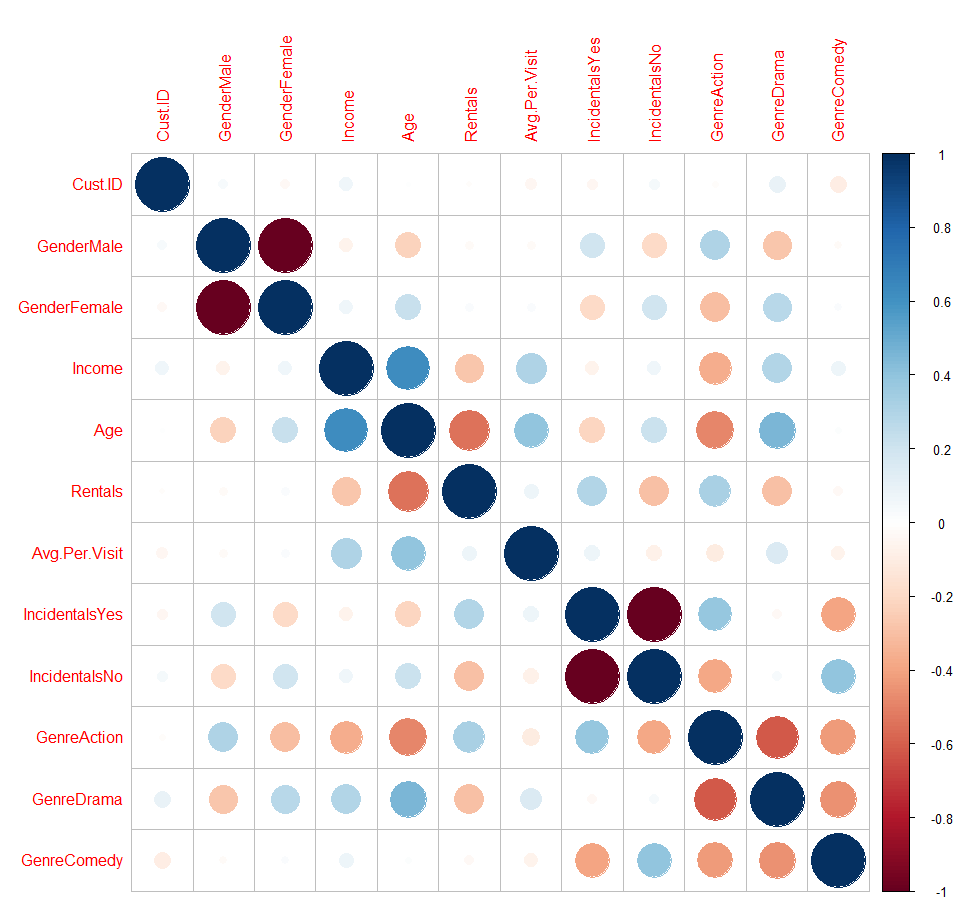
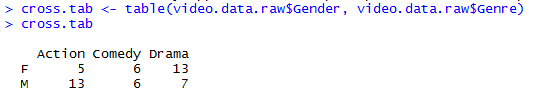
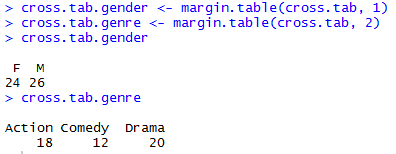
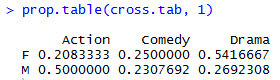
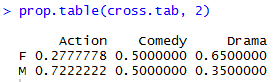
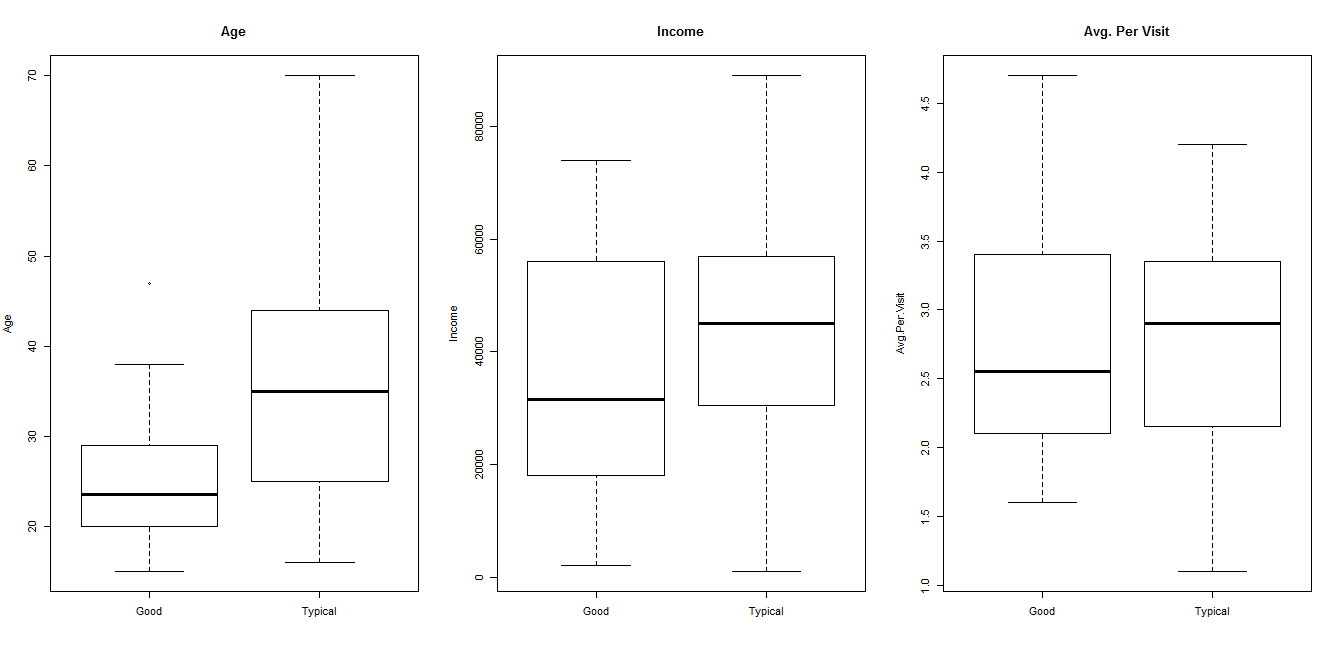
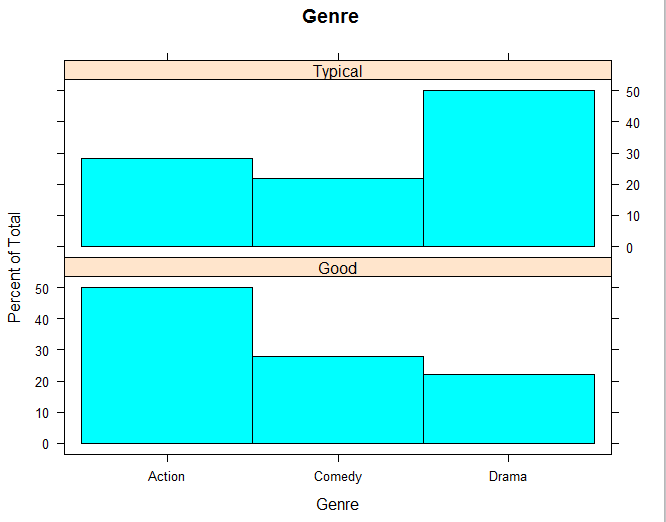
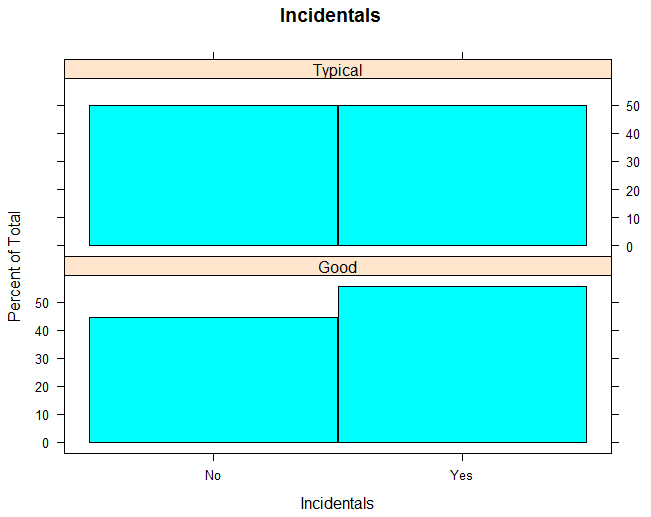
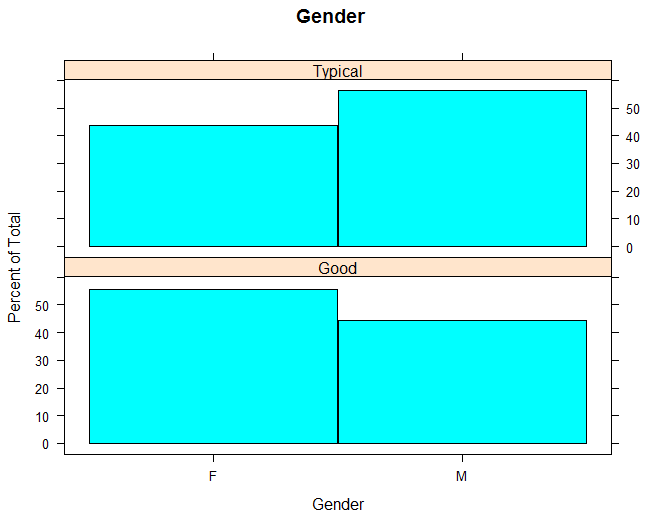
# Homework 5- Data Processing

Prepared By: Danny Godbout

Date: 11/14/2016

1. Use smoothing by bin means to smooth Age (bin depth = 4)  
   
2. Apply min-max normalization to Income  
   
3. Apply z-score normalization to Rentals  
   
4. Discretize Income by <$25k, $25k - <$60k, >=$60k  
   

These data processing steps were combined into the following AzureML flow with the resulting CSV attached to this submission as *AzureML\_PartsAtoD.csv*.  
 

1. Convert the dataset into standard spreadsheet format, expanding categorical features into unique columns.  
   
2. Perform a correlation analysis between all features of the matrix generated in step e. Note significant correlations.  
     
   * The correlation matrix shows a positive correlation between Age and Income, yet shows a negative correlation between Age/Income and Rentals.
   * While higher Age/Income customers appear to rent less, there is a positive correlation with their Average Per Visit expenditure.
   * With Age in particular, we see a positive correlation with Drama rentals, and negative correlation with Action.
   * Action rentals also see a positive correlation with Incidentals, while Comedies show a negative relationship to Incidentals.
   * Males show a slight positive correlation with Action and negative with Drama, while females reverse that pattern.
3. Perform cross-tabulation of gender and genre:  
     
   We see fairly even distribution between males and females, but more Action and Drama rentals than Comedy:  
     
   Looking at the percentages by-gender, we see that females bias towards Drama while males bias towards Action:  
     
   Tabulating percentages by genre, we similarly see that most Action renters are male, most Drama renters are female, but Comedy is split evenly between genders:  
   
4. Select good customers with a high Rental attribute (>= 30) and examine which attributes differ from the general population.  
     
   Among numerical variables, the following boxplots show a tendency towards younger Age for “good” customers. Mean Income is lower for good customers, but the 1st-3rd quartiles overlap significantly. Average Per Visit also overlaps significantly for both groups.  
     
   Among categorical variables, we see that good customers bias more towards Action movies than the general bias towards Drama. Good customers also have a very slight bias towards buying Incidentals and slightly bias towards female Gender.  
     
     
   
5. How could sales of incidentals be increased?  
   Incidentals tend to be purchased by customers that are male, younger, rent often (i.e. “good customers”), and rent action movies.   
   Marketing campaigns could be tailored towards these demographics. In particular, good customers could be targeted with a buy-10-get-1-free style or loyalty program.