Ryan Timbrook

Data Science 450, Spring 2017

Date: 05/21/2017

Assignment 4

Shared Azure Notebook URL:

https://notebooks.azure.com/n/VSAv262I9VE/notebooks/ds_450_ass4_v1.0.ipynb

Description: Data Preparation

Video store data set of 50 regular customers

This data consists of a table which, for each customer, records the following attributes: Gender •

- **Income** •
- **Age** •
- **Rentals** - Total number of video rentals in the past year •
- **Avg. per visit** - Average number of video rentals per visit during the past year
- **Incidentals** - Whether the customer tends to buy incidental items such as refreshments when renting a video
- **Genre** - The customer's preferred movie genre

Perform each of the following data preparation tasks:

- a) Use **smoothing** by bin means to smooth the values of the Age attribute. Use a bin depth of 4.
- b) Use **min-max** normalization to transform the values of the Income attribute onto the range [0.0-1.0].
- c) Use **z-score normalization** to standardize the values of the Rentals attribute.
- d) **Discretize** the (original) Income attribute based on the following categories: High = 60K+; Mid = 25K-59K; Low = less than $25K

Output Data Set:

ds_ass4_aTod.csv

- e) **Convert the original data** (not the results of parts a-d) into the standard spreadsheet format (note that this requires that you create, for every categorical attribute, additional attributes corresponding to values of that categorical attribute; numerical attributes in the original data remain unchanged).

Output Data Set:

ds_ass4_task_E.csv

- f) Using the standardized data set (from part e), perform **basic correlation analysis among the attributes**.
  Discuss your results by indicating any strong correlations (positive or negative) among pairs of attributes. You need to **construct a complete Correlation Matrix** (Please read the brief document Basic Correlation Analysis (see course website) for more detail).
  Question: Can you observe any "significant" patterns among groups of two or more variables? Explain.¶

  Answer: Income, Age and Gender play a significant role in the genre of moves being selected.

Output Results: Table 1

Strong Correlations: range |.4:1|

Positive:

-Age-Income

-Age-Genre_Drama

-Genre_Comedy-Incidentals_No

Negative:

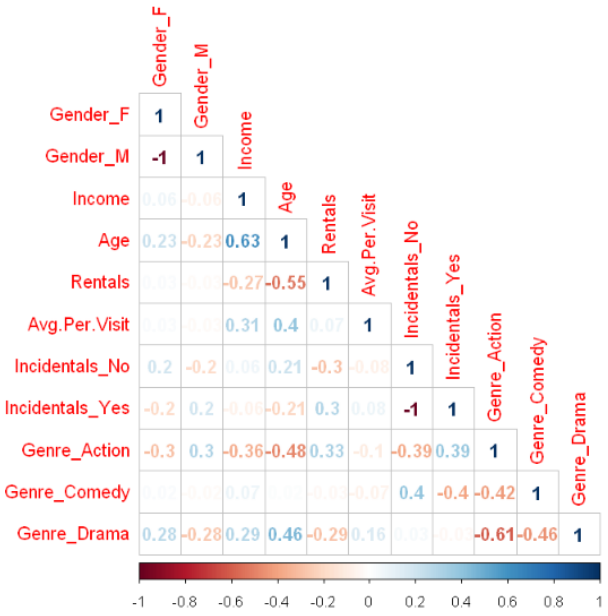
-Age-Rentals

-Age-Genre_Action

-Genre_Action-Genre_Drame

-Genre_Drama-Genre_Comedy

-Genre_Comedy-Incidentals_Yes



- g) Perform a **cross-tabulation** of the two "gender" variables versus the three "genre" variables. #### Show this as a 2 x 3 table with entries representing the total counts.

  Then, use a graph or chart that provides the best visualization of the relationships between these sets of variables.

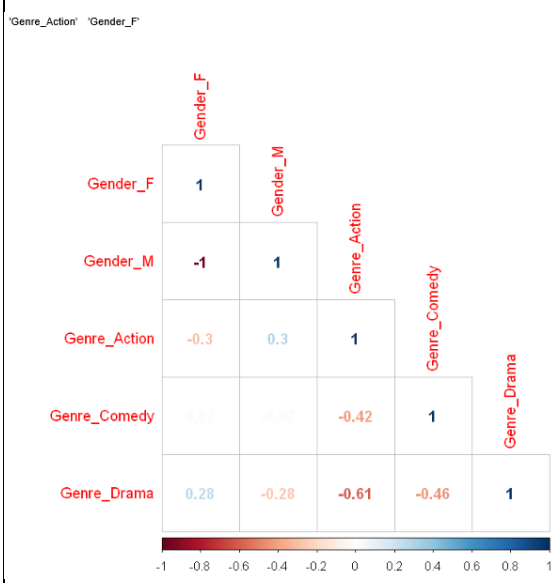
  Question: Can you draw any significant conclusions?¶

  Answer: Women prefer Dramas over Comedy and Action equally as Men prefer Actions over Drama and Comedy's

Output Results:

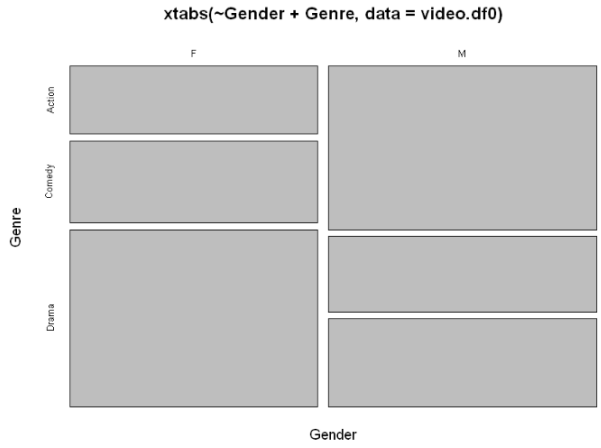
```
       Genre
Gender Action Comedy Drama
     F      5      6    13
     M     13      6     7

Call: xtabs(formula = ~Gender + Genre, data = video.df0)
Number of cases in table: 50
Number of factors: 2
Test for independence of all factors:
        Chisq = 5.284, df = 2, p-value = 0.07122
```




```
       Genre
Gender Action Comedy Drama
     F      5      6    13
     M     13      6     7

Call: xtabs(formula = ~Gender + Genre, data = video.df0)
Number of cases in table: 50
Number of factors: 2
Test for independence of all factors:
        Chisq = 5.284, df = 2, p-value = 0.07122
```

- h) Select all "good" customers with a high value for the Rentals attribute (a "good customer is defined as one with a Rentals value of greater than or equal to 30). Then, create a summary (e.g., using means, medians, and/or other statistics) of the selected data with respect to all other attributes.

Question: Can you observe any significant patterns that characterize this segment of customers? Explain. ¶

Note: To know whether your observed patterns in the target group are significant, you need to compare them with the general population using the same metrics.

[1] "#### Good Customers Head data and Summary Statistics ####"

Gender_F	Gender_M	Income	Age	Rentals	Avg.Per.Visit	Incidentals_No	Incidentals_Yes	Genre_Action	Genre_Comedy	Genre_Drama
1	0	2000	15	30	2.5	1	0	0	1	0
1	0	6000	16	39	1.8	0	1	1	0	0
1	0	15000	18	37	2.1	0	1	1	0	0
0	1	17000	19	32	1.8	1	0	1	0	0
1	0	32000	20	42	1.6	1	0	0	1	0
0	1	18000	20	33	1.7	1	0	1	0	0

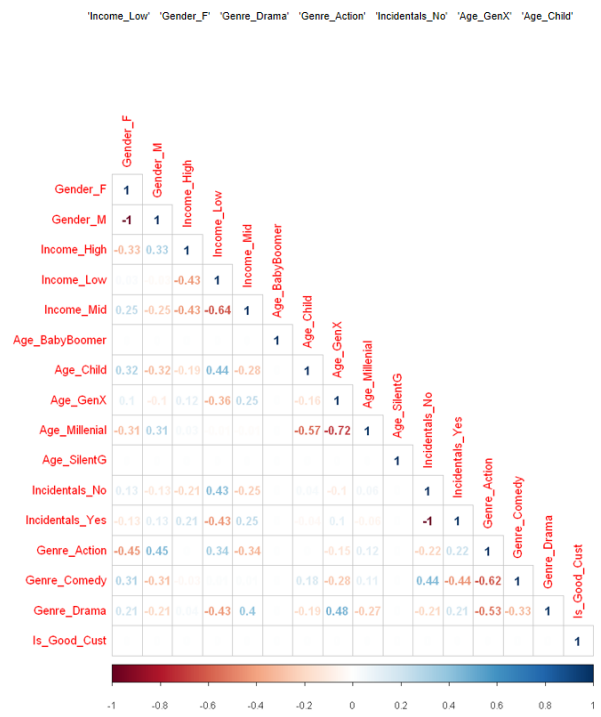
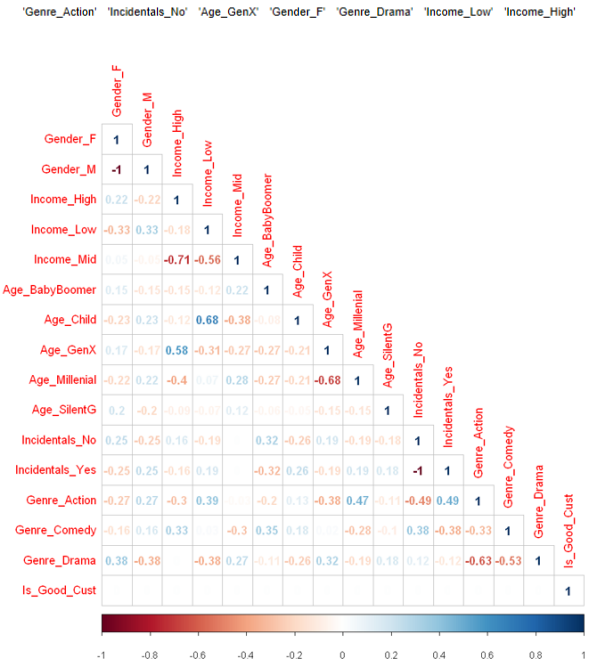
```
    Gender_F         Gender_M          Income           Age       
 Min.   :0.0000   Min.   :0.0000   Min.   : 2000   Min.   :15.00  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:19500   1st Qu.:20.00  
 Median :1.0000   Median :0.0000   Median :31500   Median :23.50  
 Mean   :0.5556   Mean   :0.4444   Mean   :37667   Mean   :25.17  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:56000   3rd Qu.:28.75  
 Max.   :1.0000   Max.   :1.0000   Max.   :74000   Max.   :47.00  
    Rentals      Avg.Per.Visit   Incidentals_No   Incidentals_Yes 
 Min.   :30.00   Min.   :1.600   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:32.25   1st Qu.:2.150   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :36.50   Median :2.550   Median :0.0000   Median :1.0000  
 Mean   :37.28   Mean   :2.778   Mean   :0.4444   Mean   :0.5556  
 3rd Qu.:41.75   3rd Qu.:3.375   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :48.00   Max.   :4.700   Max.   :1.0000   Max.   :1.0000  
  Genre_Action  Genre_Comedy     Genre_Drama    
 Min.   :0.0   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.5   Median :0.0000   Median :0.0000  
 Mean   :0.5   Mean   :0.2778   Mean   :0.2222  
 3rd Qu.:1.0   3rd Qu.:0.7500   3rd Qu.:0.0000  
 Max.   :1.0   Max.   :1.0000   Max.   :1.0000  
```

[1] "#### Bad Customers Head data and Summary Statistics ####"

Gender_F	Gender_M	Income	Age	Rentals	Avg.Per.Visit	Incidentals_No	Incidentals_Yes	Genre_Action	Genre_Comedy	Genre_Drama
0	1	12000	16	23	2.2	0	1	1	0	0
0	1	1000	16	25	1.4	0	1	0	1	0
0	1	17000	19	26	2.2	0	1	1	0	0
0	1	38000	21	18	2.1	1	0	0	1	0
1	0	26000	22	29	2.9	0	1	1	0	0
0	1	35000	24	24	1.7	1	0	0	0	1

```
    Gender_F         Gender_M          Income           Age       
 Min.   :0.0000   Min.   :0.0000   Min.   : 1000   Min.   :16.00  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:31250   1st Qu.:25.00  
 Median :0.0000   Median :1.0000   Median :45000   Median :35.00  
 Mean   :0.4375   Mean   :0.5625   Mean   :44906   Mean   :35.16  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:57000   3rd Qu.:43.50  
 Max.   :1.0000   Max.   :1.0000   Max.   :89000   Max.   :70.00  
    Rentals      Avg.Per.Visit   Incidentals_No Incidentals_Yes
 Min.   : 9.00   Min.   :1.100   Min.   :0.0    Min.   :0.0    
 1st Qu.:16.00   1st Qu.:2.175   1st Qu.:0.0    1st Qu.:0.0    
 Median :20.50   Median :2.900   Median :0.5    Median :0.5    
 Mean   :20.03   Mean   :2.731   Mean   :0.5    Mean   :0.5    
 3rd Qu.:24.25   3rd Qu.:3.325   3rd Qu.:1.0    3rd Qu.:1.0    
 Max.   :29.00   Max.   :4.200   Max.   :1.0    Max.   :1.0    
  Genre_Action     Genre_Comedy     Genre_Drama    
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0    
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0    
 Median :0.0000   Median :0.0000   Median :0.5    
 Mean   :0.2812   Mean   :0.2188   Mean   :0.5    
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0    
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0    
```

Table: Good Customer Correlation Matrix	Table: Bad Customer Correlation Matrix

Significant Attributes:	Significant Attributes:
	

- i) Suppose that because of the high profit margin, the store would like to increase the sales of incidentals. Based on your observations in previous parts discuss how this could be accomplished.

  Answer: Target marketing toward blue collar, mid-income level, millennial males who are either married or in a relationship. Focus on Action movies that are suspenseful and cross into the Drama domain.

Explain your answer based on your analysis of the data.

Question: Should customers with specific characteristics be targeted?¶

Cluster 3 was composed 2/3 of Millenials and 1/3 GenX Cluster 3 was mostly composed of Mid Income level observations Cluster 3 has twice as many Male observations then Female

Question: Should certain types of movies be preferred?¶

Cluster 3 had a 100% Incidentals_Yes of it's observations, where Action(10) and Drama(8) Movies were preferred

ds_ass4_cluster_analysis.csv

Table: Matrix of Cluster Summary Values

count	Good_Customers	Bad_Customers	Males	Females	High_Incomes	Mid_Incomes	Low_Incomes	BabyBoomers	Children	GenXs	Millenials	Rentals_Mean	Avg_Per_Visit	Incidentals_Yes	Incidentals_No	Genre_Actions	Genre_Comedys	Genre_Dram
20	3	17	6	14	7	13	0	3	0	11	6	-0.607321649	2.715	2	18	0	8	
11	7	4	7	4	0	0	11	0	4	0	7	0.529083318	2.381818182	5	6	8	3	
19	8	11	13	6	3	16	0	0	0	5	13	0.332974552	2.994736842	19	0	10	1	

Notes: Use your favorite machine learning tool, Excel or scripting to perform the following tasks on the original data set. - Review basic statistics for different attributes by clicking on the name of each one in "attribute" panel. - Consider discretizing the Age attribute. - Convert all of the remaining numerical attribute into [0…1] scale.

Save the resulting data set into an ARFF formatted or CSV file and submit with your answers for the above questions.

Attached is the output of this data exploration:

ds_ass4_all_numeric.csv

You can give the final results of parts (a) through (d) as a single table which includes the original data and has an added column for each of the parts (a) through (d).

Attached is the output for this data conversion:

The results of part (e) should be a separate table.

For the correlation analysis (part f) give your correlation matrix (rows and columns of the matrix are the attributes, and entries would represent correlation value for a pair of attributes (e.g., "Income" versus "Age").