Assignment # 4: Data Preparation

Consider the data collected by a hypothetical video store for 50 regular customers. This data consists of a table which, for each customer, records the following attributes:

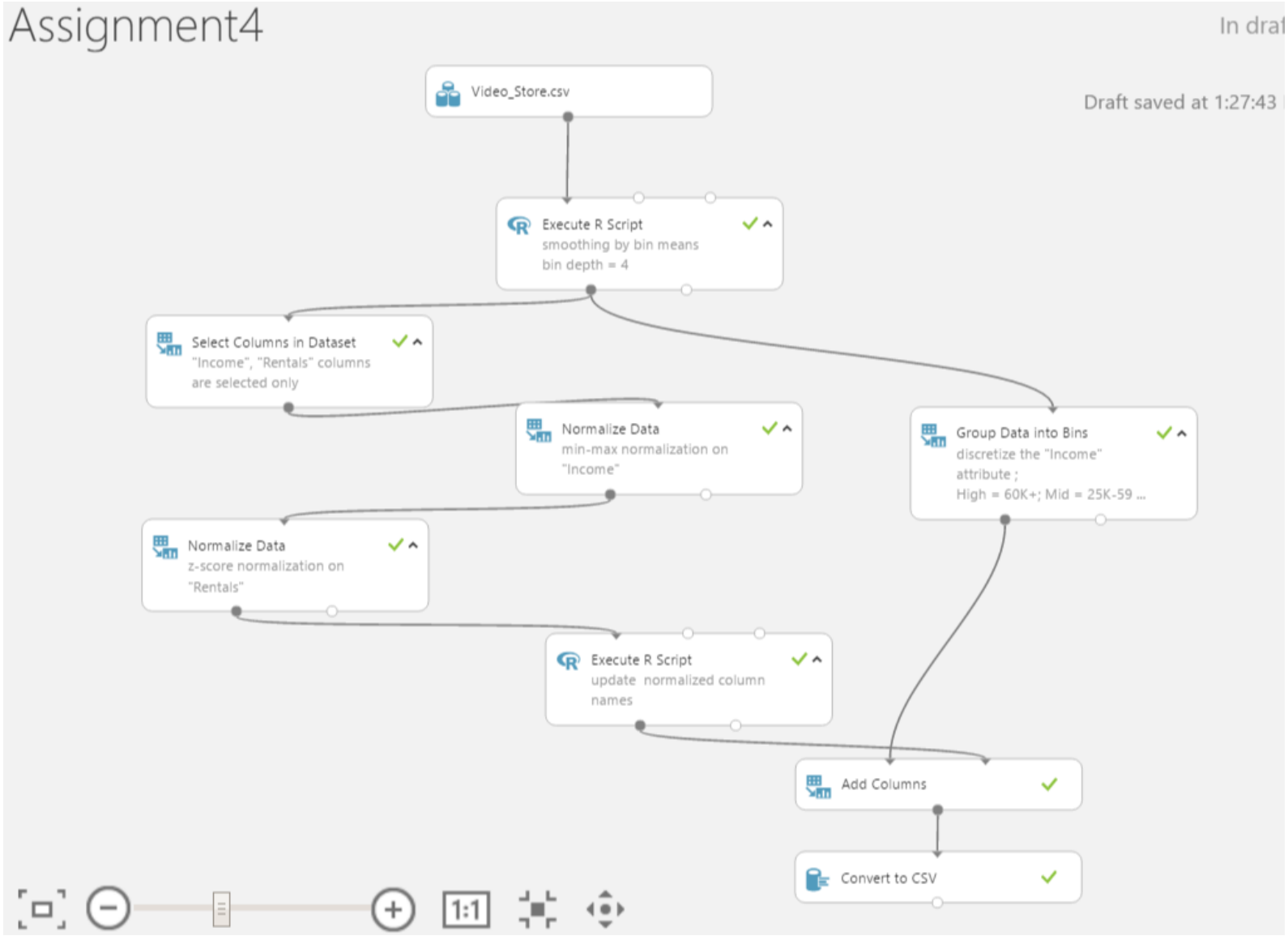
- Gender
- Income
- Age
- Rentals - Total number of video rentals in the past year
- Avg. per visit - Average number of video rentals per visit during the past year
- Incidentals - Whether the customer tends to buy incidental items such as refreshments when renting a video
- Genre - The customer's preferred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

Perform each of the following data preparation tasks:

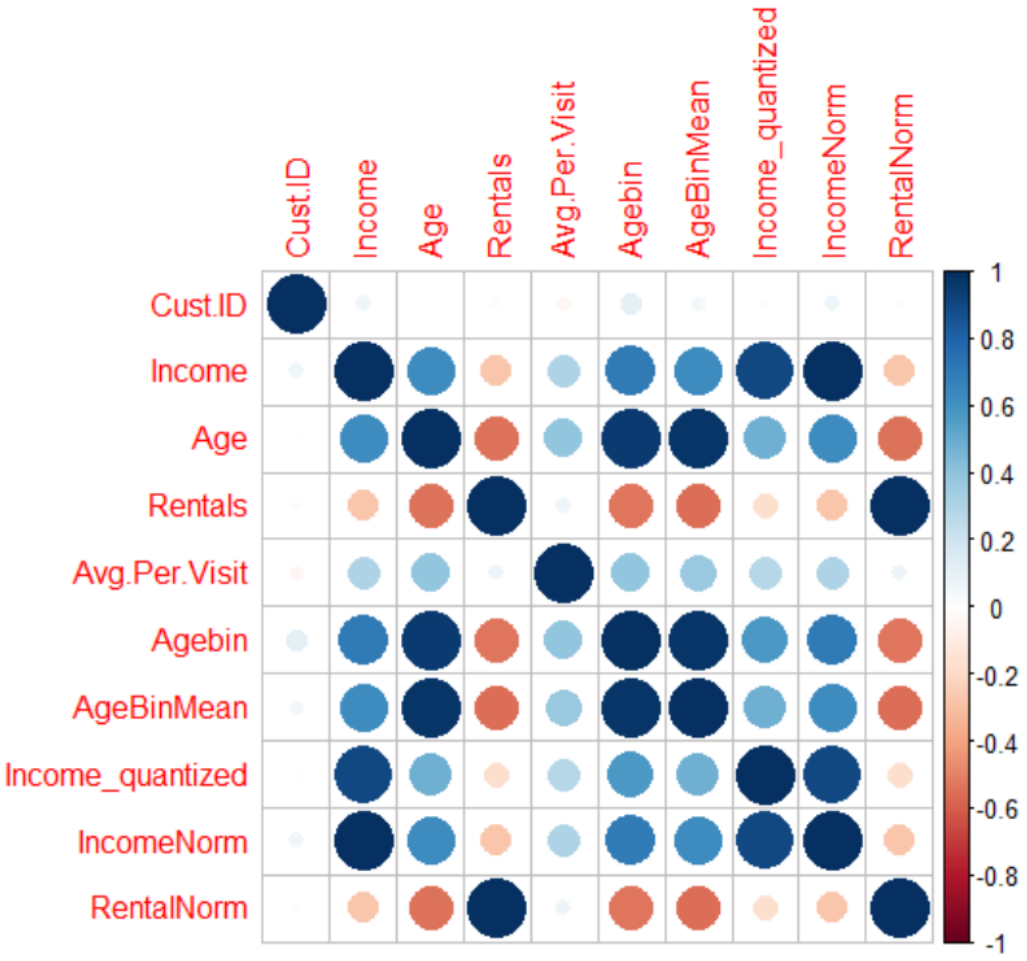
A. Use smoothing by bin means to smooth the values of the Age attribute. Use a bin depth of 4.
B. Use min-max normalization to transform the values of the Income attribute onto the range [0.0-1.0].
C. Use z-score normalization to standardize the values of the Rentals attribute.
D. Discretize the (original)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

Part A to E of this homework is performed using Azure ML (shown in figure below) and resulted dataset is exported as "Assigment4.cvs" (attached).

Part F to I of this homework is performed using R and results are discussed as follows.


Assignment4
In draf
Draft saved at 1:27:43
Video_Store.csv
Execute R Script
smoothing by bin means
bin depth = 4
Select Columns in Dataset
"Income", "Rentals" columns
are selected only
Normalize Data
min-max normalization on
"Income"
Group Data into Bins
discretize the "Income"
attribute ;
High = 60K+; Mid = 25K-59 ...
Normalize Data
z-score normalization on
"Rentals"
Execute R Script
update normalized column
names
Add Columns
Convert to CSV
1:1

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). Can you observe any "significant" patterns among groups of two or more variables? Explain.



- The customer's "Age" and "Income" have the highest positive correlation; i.e. as the customers' income increases as they get older.
- Customer "Age" and "Rentals" have the highest negative correlation; i.e. as the customers gets older, the number of videos they rent decreases.

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. Can you draw any significant conclusions?

```
#   Cell Contents
# |-------------------------|
# |                       N |
# |           N / Row Total |
# |           N / Col Total |
# |-------------------------|
#
# Total Observations in Table:  50
#
#              | data$Genre
#  data$Gender |    Action |    Comedy |     Drama | Row Total |
# -------------|-----------|-----------|-----------|-----------|
#            F |         5 |         6 |        13 |        24 |
#              |     0.208 |     0.250 |     0.542 |     0.480 |
#              |     0.278 |     0.500 |     0.650 |           |
# -------------|-----------|-----------|-----------|-----------|
#            M |        13 |         6 |         7 |        26 |
#              |     0.500 |     0.231 |     0.269 |     0.520 |
#              |     0.722 |     0.500 |     0.350 |           |
# -------------|-----------|-----------|-----------|-----------|
# Column Total |        18 |        12 |        20 |        50 |
#              |     0.360 |     0.240 |     0.400 |           |
# -------------|-----------|-----------|-----------|-----------|
```

Based on the tabulated results (above) following conclusions are made:

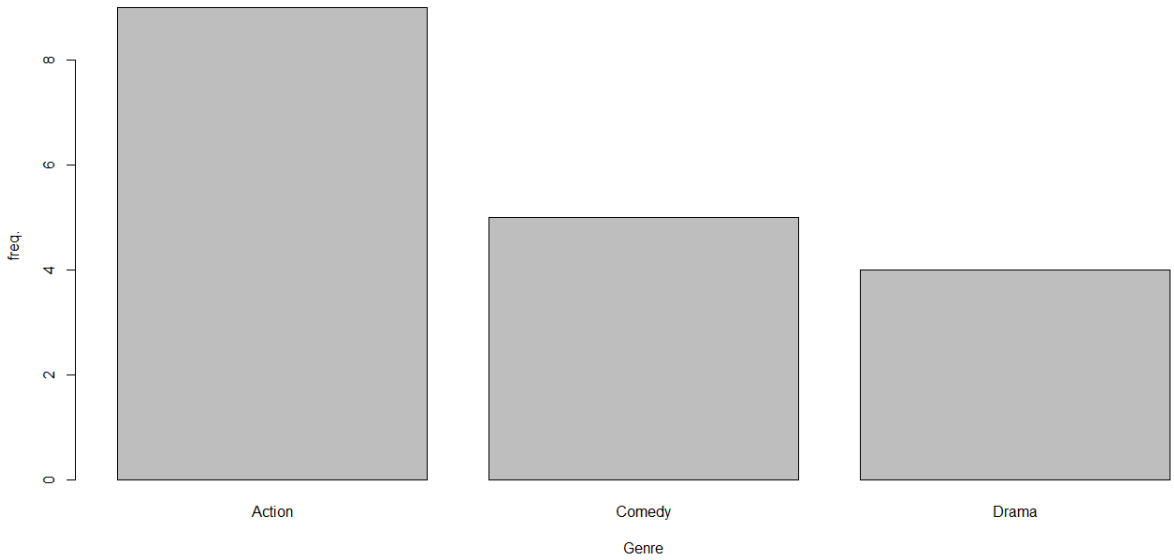
- The total male and female customer numbers are almost equivalent (48 % Female vs 52 % male).
- More than half of the female customers (54.2 %) preferred to rent "Drama" movies, whereas, half of the male customers (50 %) preferred to rent "Action" movies.
- 72.2 % of the "Action" movies were rented by male customers, whereas, 65 % of the "Drama" movies were rented by female customers.

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. 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.

Based on the given criteria, the good customer whose rents more than 30 movies in a year

- whose average salary is around "$30k" or higher
- he/she is in his/her mid-20s
- visits to the store around 2 to 3 times a year or more
- and his/her preference is to rent Action movies
- can be a Female or Male

The clearest conclusion was a good customer rents most likely an action movie (shown below).



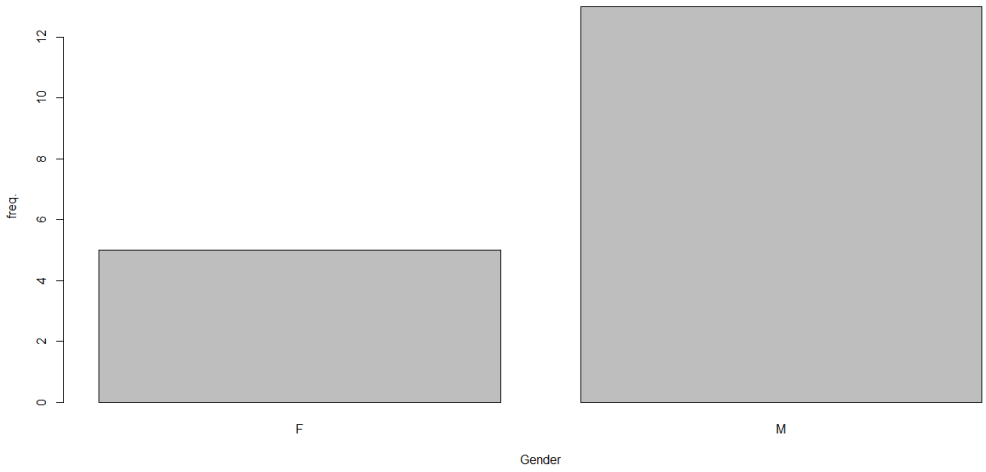
When we compared the summary of the two dataset, "Action" movie renter profile and “good” customer profile, they agreed well. Since an "Action" movie renter customer;

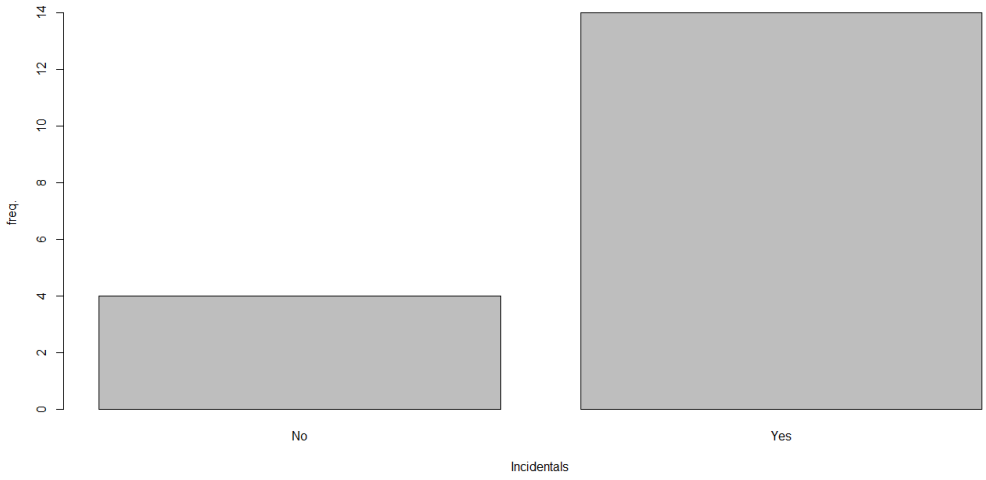
- has an average salary is around "$32k" or higher
- in his mid-20s
- rents more than 30 movies in a year
- visits the store around 2 to 3 times a year or more

One important outcome for this comparison was, a good customer who rents “Action” movies

- most likely is a "Male" customer (shown below)
- and will purchase other "incidentals" (shown below)

Additional graphs are generated in the R-studio and can be reproduced by the attached R-script.





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 (e.g., should customers with specific characteristics be targeted? Should certain types of movies be preferred? etc.). Explain your answer based on your analysis of the data.

Based on the summary of the dataset and the bar plots a customer whose less likely to purchase incidentals

- has an average salary is around "$40k"
- in early or late 20s (bar plot)
- rents more than 20 movies in a year
- visits the store around 2 to 3 times a year or more
- prefers to watch "Comedy" or "Drama" instead of "Action" movies

Another important outcome for this analysis is that a customer who doesn't purchase incidentals is most likely is a "Female" customer (shown below).

Additional graphs are generated in the R-studio and can be reproduced by the attached R-script.