**1.**Describe the two main methods of representing market basket data. What are the benefits and drawbacks of each?

There are two principal methods of representing this type of market basket data: using either the **transactional data** **format** or the **tabular data format**. The transactional data format requires only two fields, an ID field, and a content field, with each record representing a single item only. In the tabular data format, each record represents a separate transaction, with as many 0/1 flag fields as there are items.

One of the drawbacks of each model is: Models built using tabular data can be used to score only tabular data; models built using transactional data can score only transactional data.

One of the benefits of the tabular data format is that it is better for being coded and stored. So, you do not need much memory to store the table. In contrast, in the transaction model, you need much memory. For example, if many items had been purchased in any transaction, we have to save for each of these items a different row in the table and store all of them. So, it needs more memory. However, in tabular data format, you can simply code with using the sequence of 0 and 1, which was repeated for that specific transaction. Also, you can simply find the items in the transaction and check if they are sold or not. Because you also have 0 for the items which had not being sold. But in the other table format, you just have the items which were sold, and you cannot keep track of the items which were not sold easily.

One of the benefits of transaction data format is used for the transactional data, and we can easily check which items were sold in a particular transaction. So we do not have to search all of the items in one transaction to find 1 and 0. Also, it does not have as many columns as the tabular data format for each product. So, if we did not sell one particular item in a transaction, we do not have to occupy a space in the memory for that.

**2.**Describe support and confidence. Express the formula for confidence using support.

We know that an association rule takes the form if A, then B (A ⇒ B), where the antecedent A and the consequent B are proper subsets of I, and A and B are mutually exclusive.

So, for defining the term supports for a particular association rule A ⇒ B we can say that the support is the proportion of these transactions that contain both A and B. So we can say support is the percentage of transactions that contain all of the items in an itemset. The higher the support, the more frequently the itemset occurs. Rules with high support are preferred since they are likely to be applicable to a large number of future transactions.

Now for confidence we can say the term confidence of the association rule A ⇒ B is a measure of the accuracy of the rule, as determined by the percentage of transactions in D containing A that also contains B. In other words, the probability that a transaction that contains the items on the left-hand side of the rule also contains the item on the right-hand side. The higher the confidence, the greater the likelihood that the item on the right-hand side will be purchased or, in other words, the higher the return rate you can expect for a given rule.

**3.**Restate the a priori property in your own words.

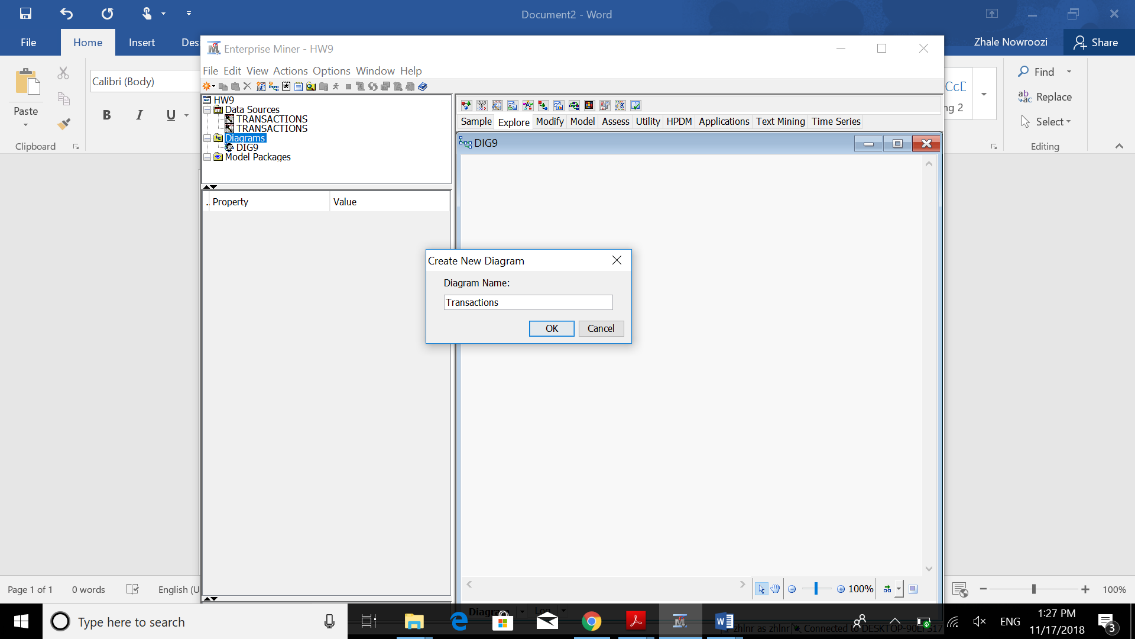
Affinity analysis is the study of attributes or characteristics that “go together.” Methods for affinity analysis, also known as market basket analysis, seek to uncover associations among these attributes; that is, it seeks to uncover rules for quantifying the relationship between two or more attributes.

What types of algorithms can we apply to mine association rules from a particular data set? The daunting problem that awaits any such algorithm is the curse of dimensionality. The number of possible association rules grows exponentially in the number of attributes. Specifically, if there are k attributes, we limit ourselves to binary attributes, we account only for the positive cases (e.g., buy diapers = yes), there are on the order of (k)2k–1 possible association rules.

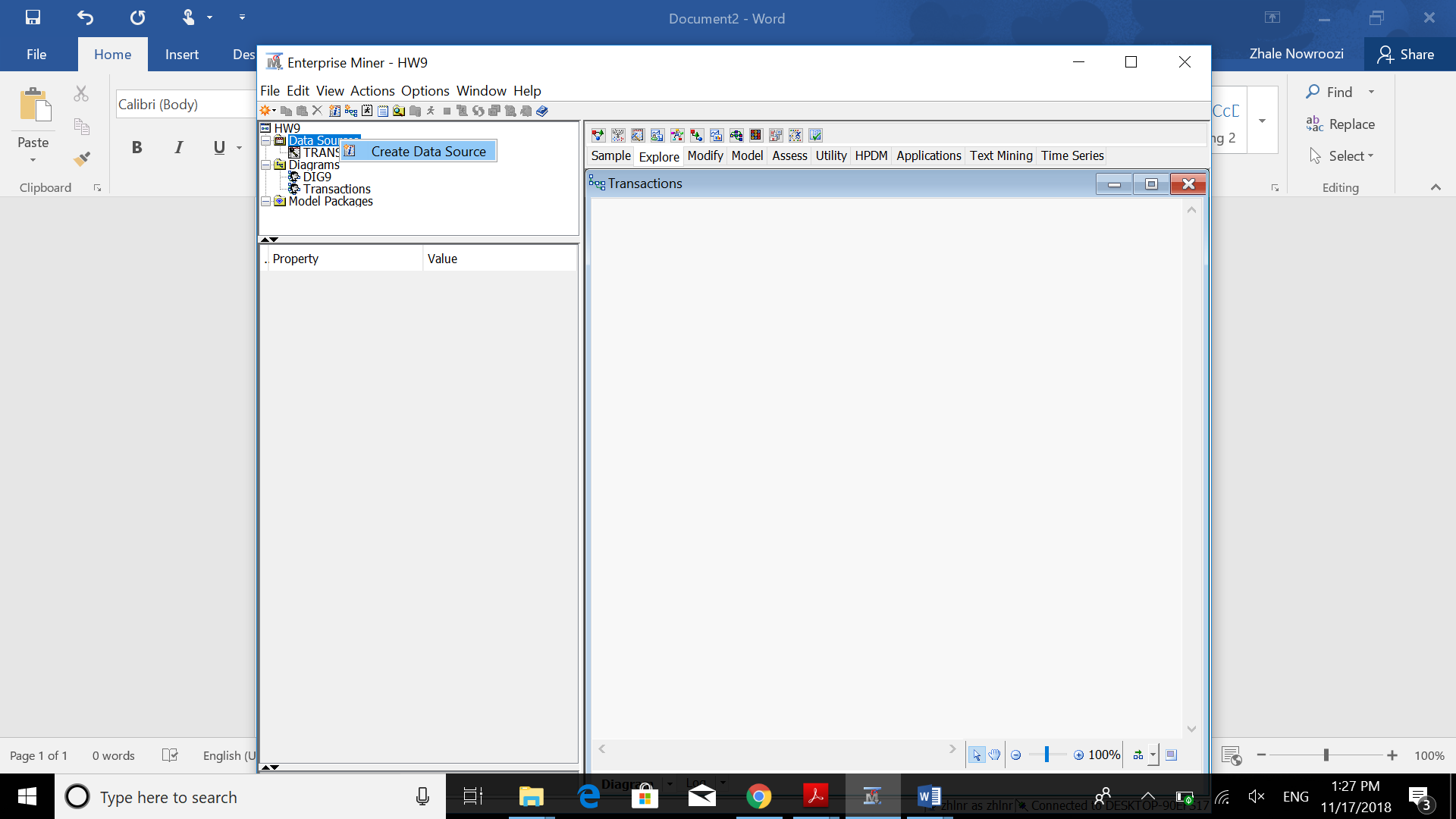
So, we have to use an algorithm to reduce this huge size of association rule in Affinity Analysis. An algorithm we can use is the priori algorithm. This algorithm finds out the most frequent items in the data set. Then it expands these items to a larger item set until all of these items in item sets appear in data set by a reasonably large amount of frequency. So, it will use the priori property to make the search space smaller. In other words, A priori uses an approach, where frequent subsets are extended one item at a time, and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. This helpful property reduces the search space significantly for the a priori algorithm.

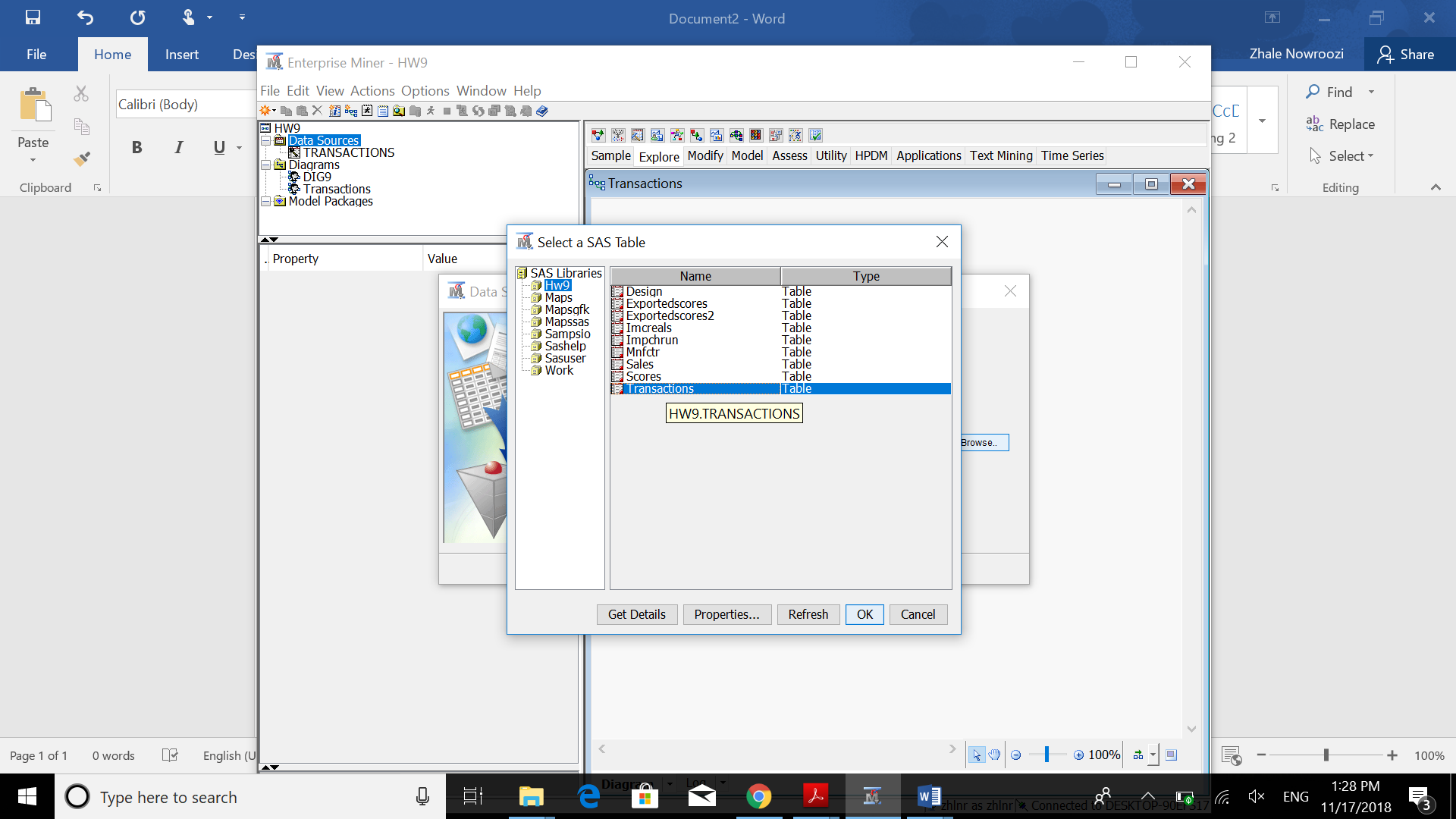
4. Conducting an Association Analysis A store is interested in determining the associations between items purchased from the Health and Beauty Aids Department and the Stationery Department. The store chose to conduct a market basket analysis of specific items purchased from these two departments. The TRANSACTIONS data set contains information about more than 400,000 transactions made over the past three months.

1. Create a new diagram. Name the diagram Transactions.

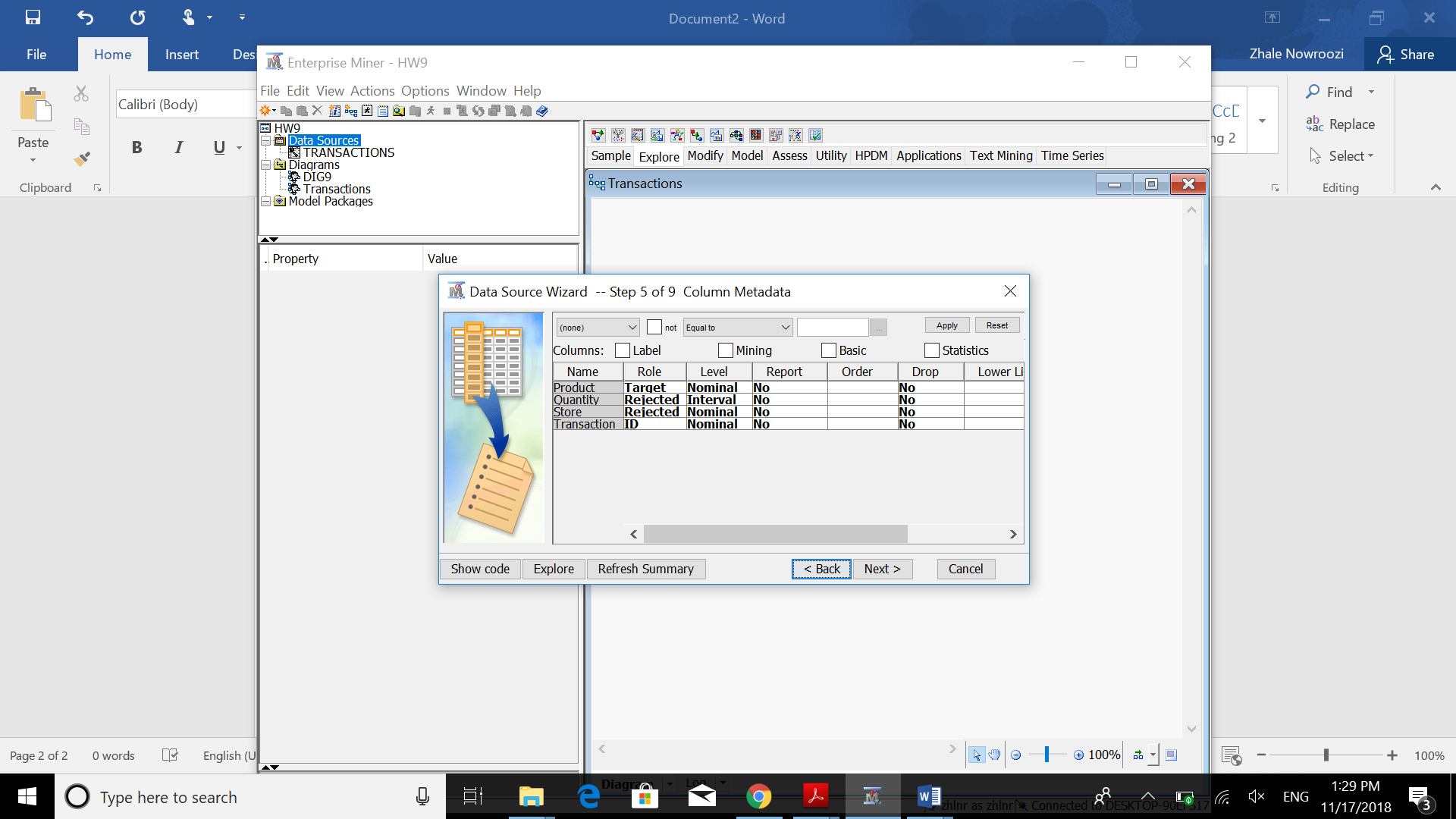


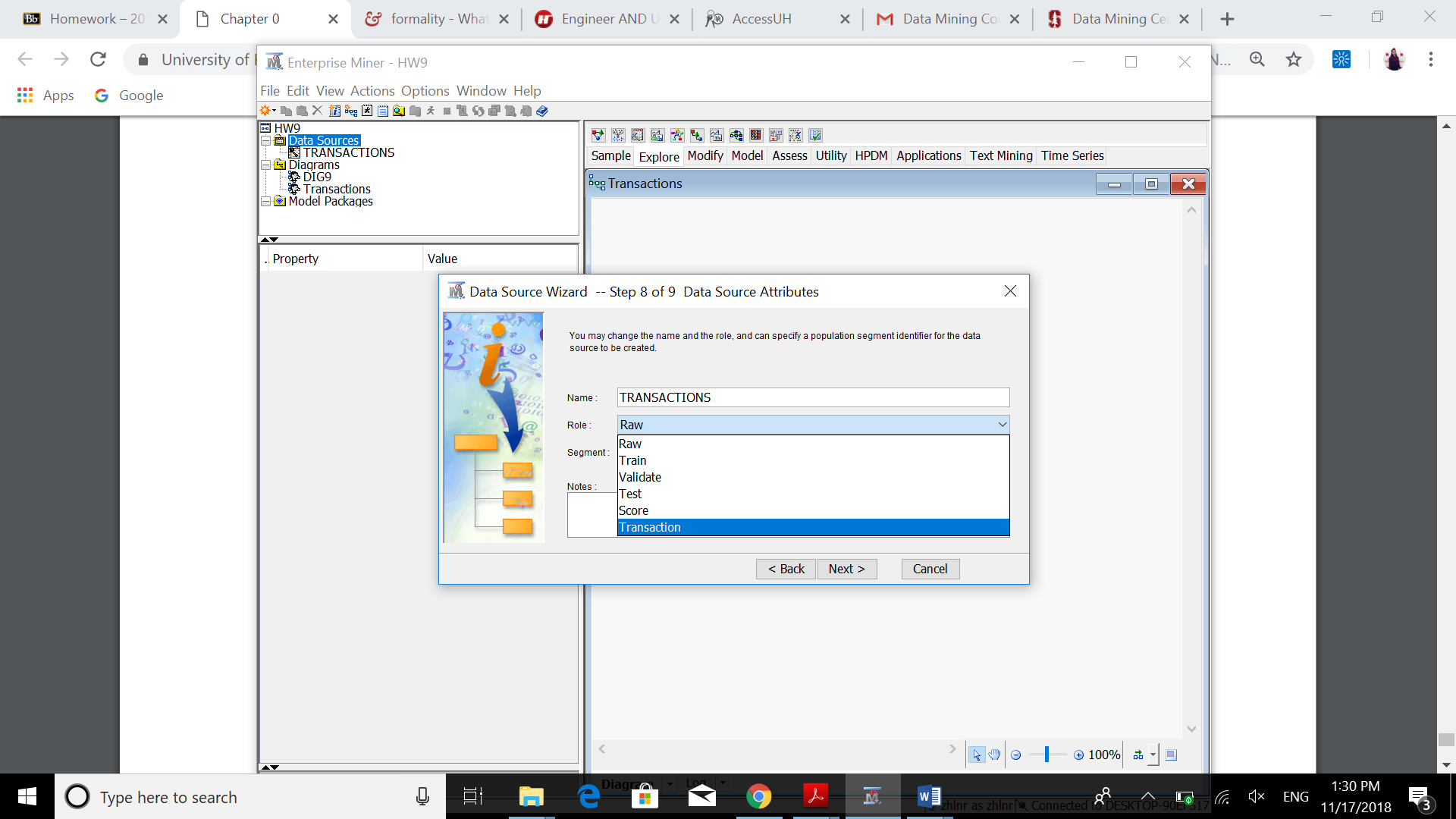
1. Create a new data source using the data set AAEM.TRANSACTIONS.

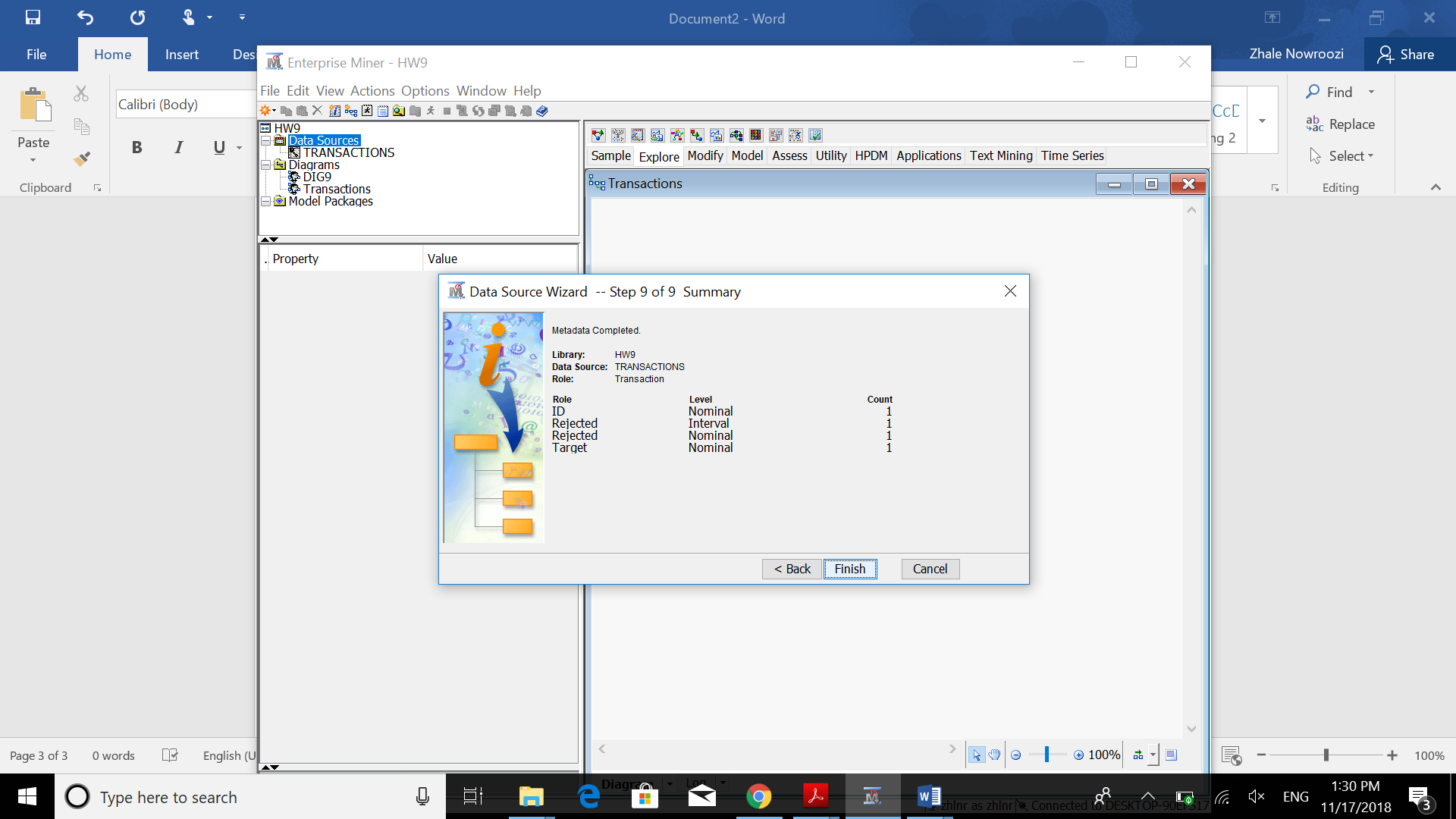




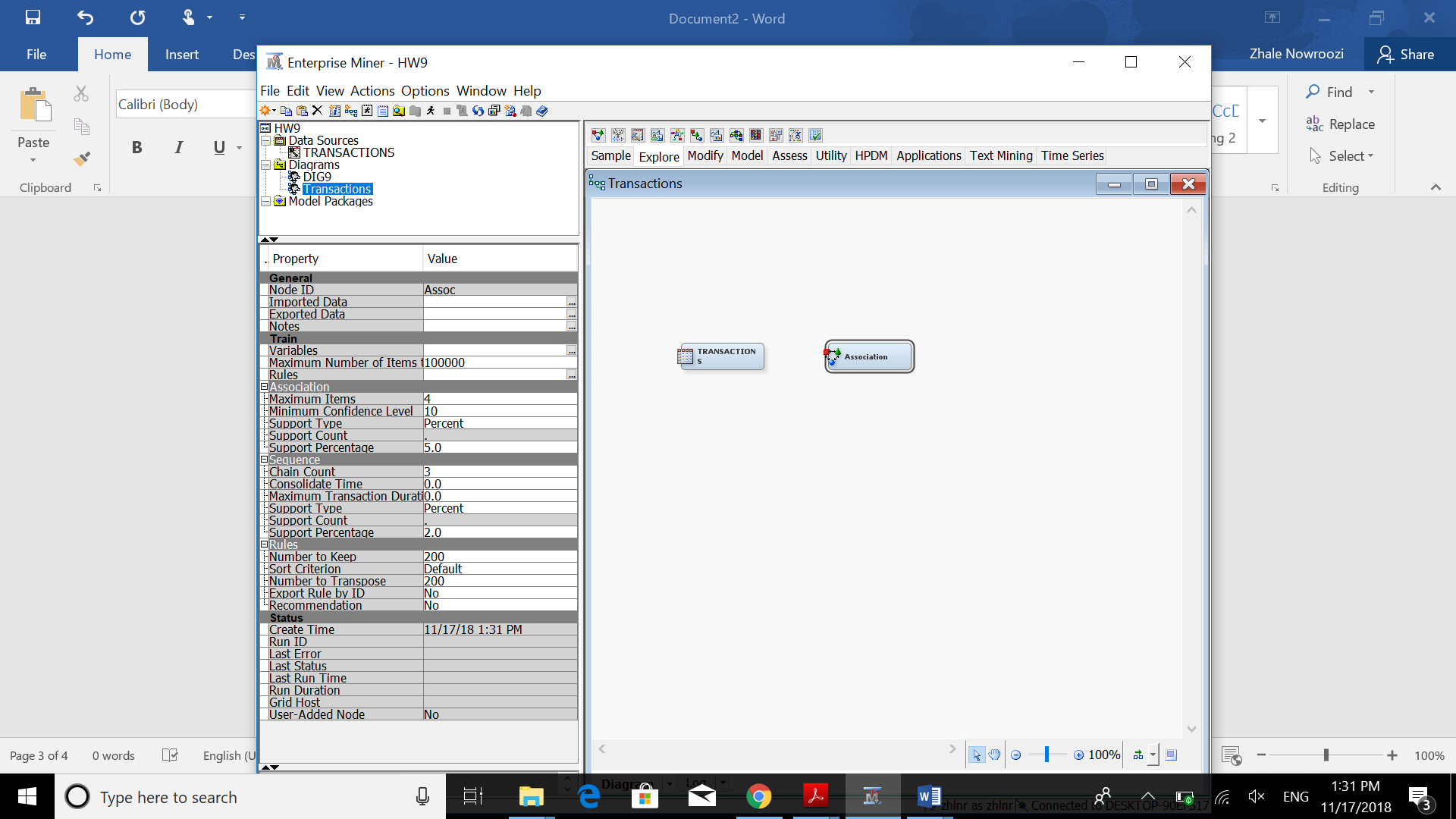
1. Assign the variables STORE and QUANTITY the model role Rejected. These variables will not be used in this analysis. Assign the ID model role to the variable TRANSACTION and the Target model role to the variable PRODUCT.



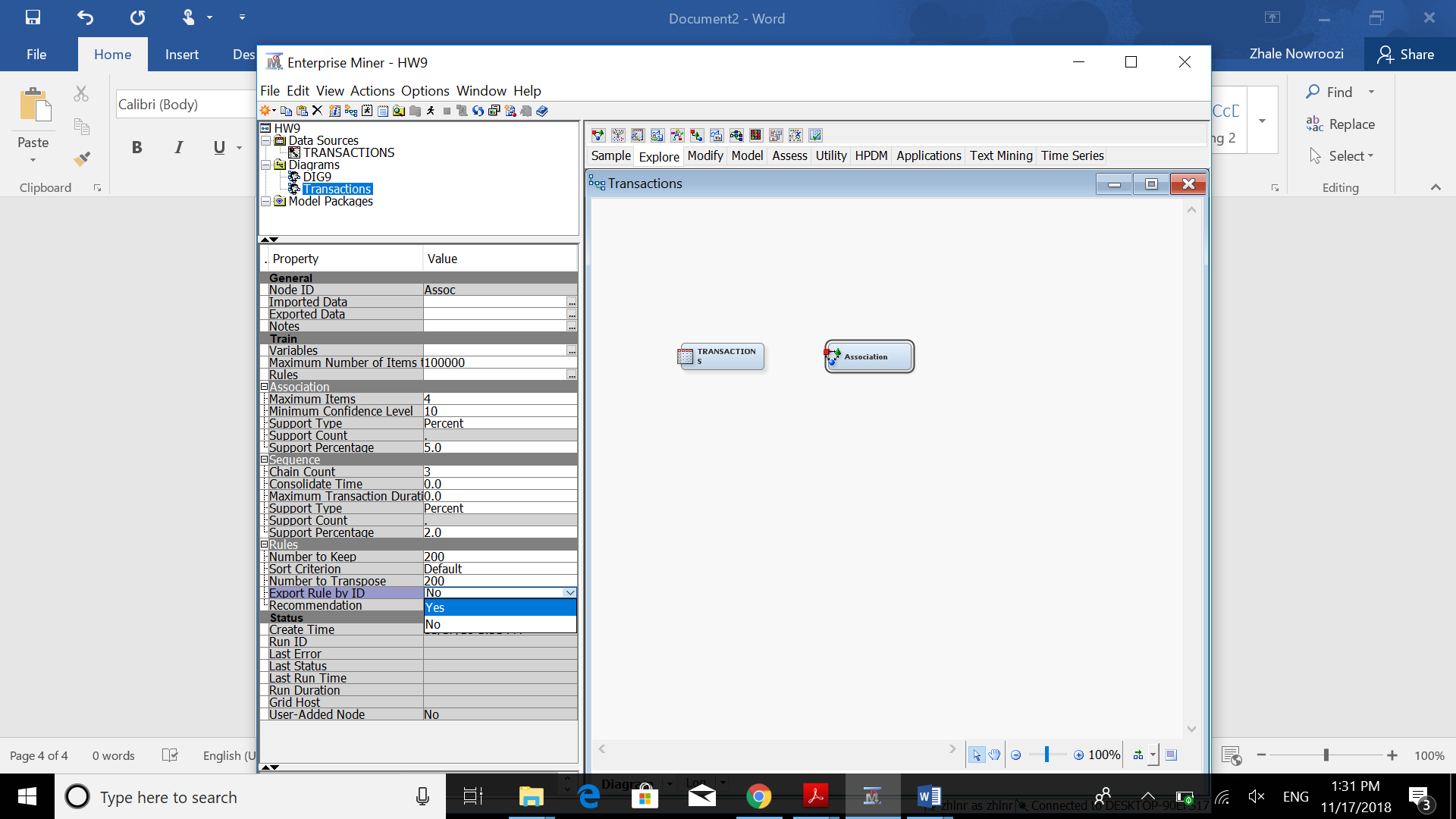




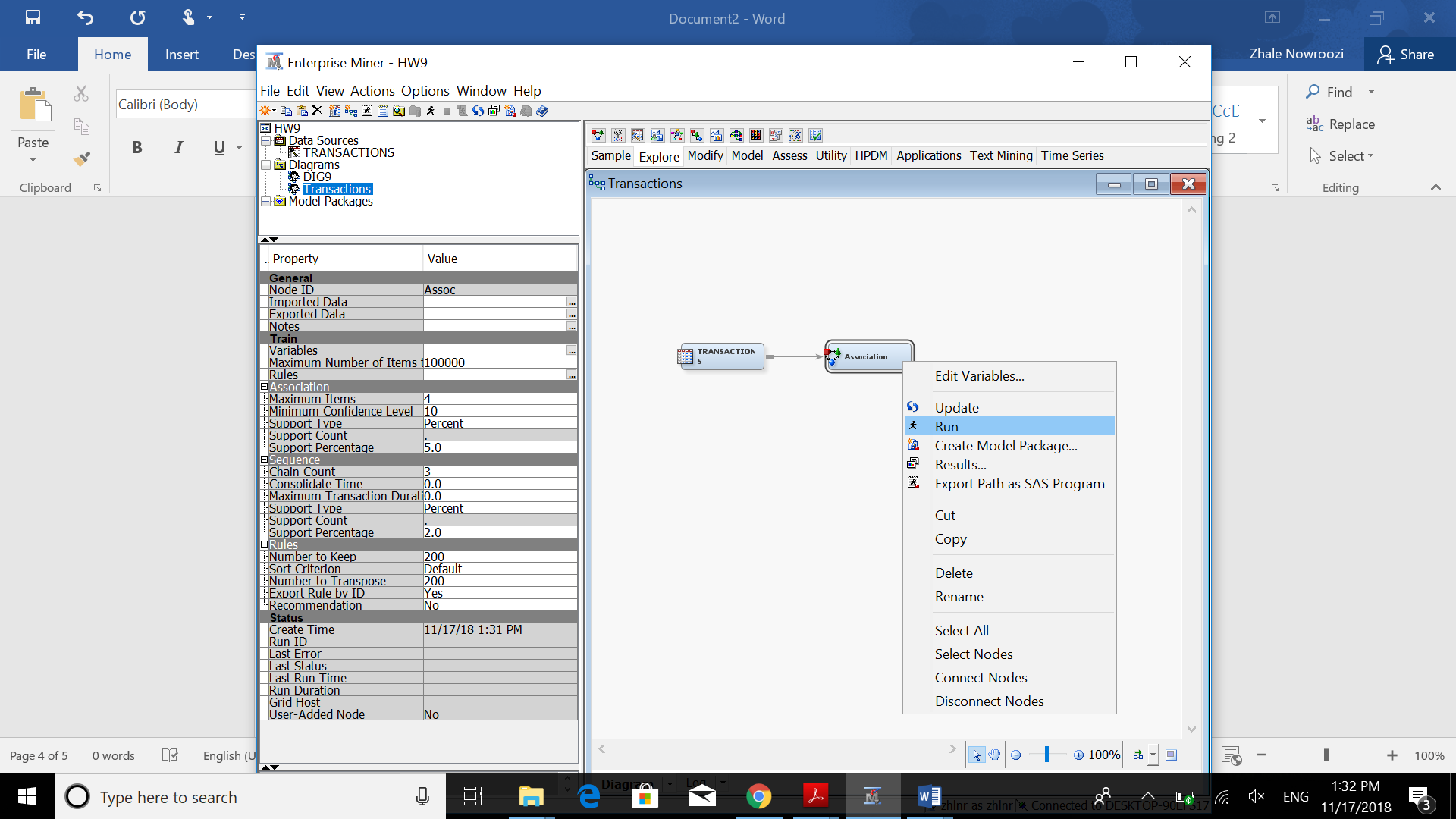
1. Add the node for the TRANSACTIONS data set and an Association node to the diagram.



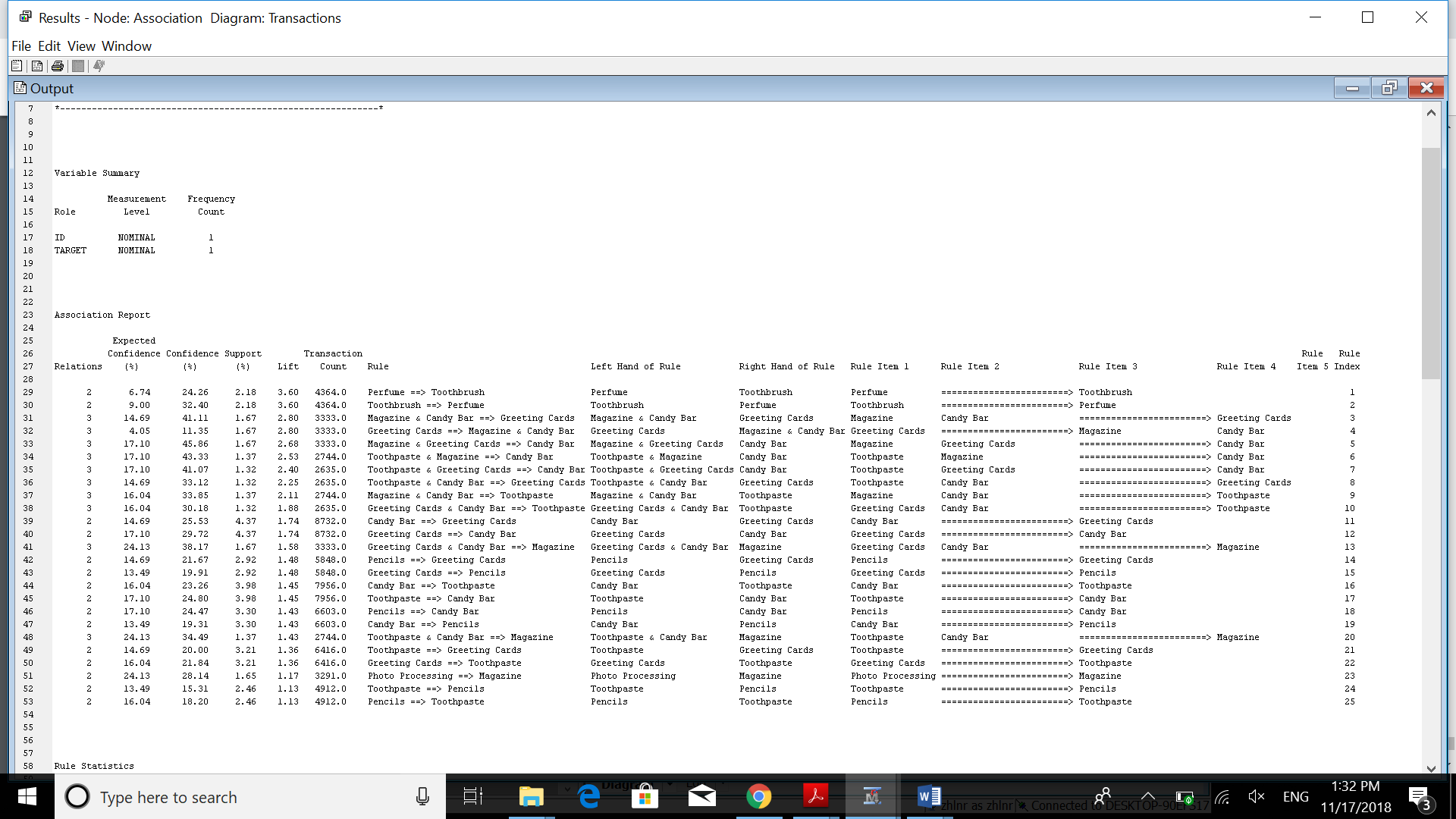
1. Change the setting for Export Rule by ID to Yes.



1. Leave the remaining default settings for the Association node and run the analysis.



1. Examine the results of the association analysis.



What is the highest lift value for the resulting rules? 3.60

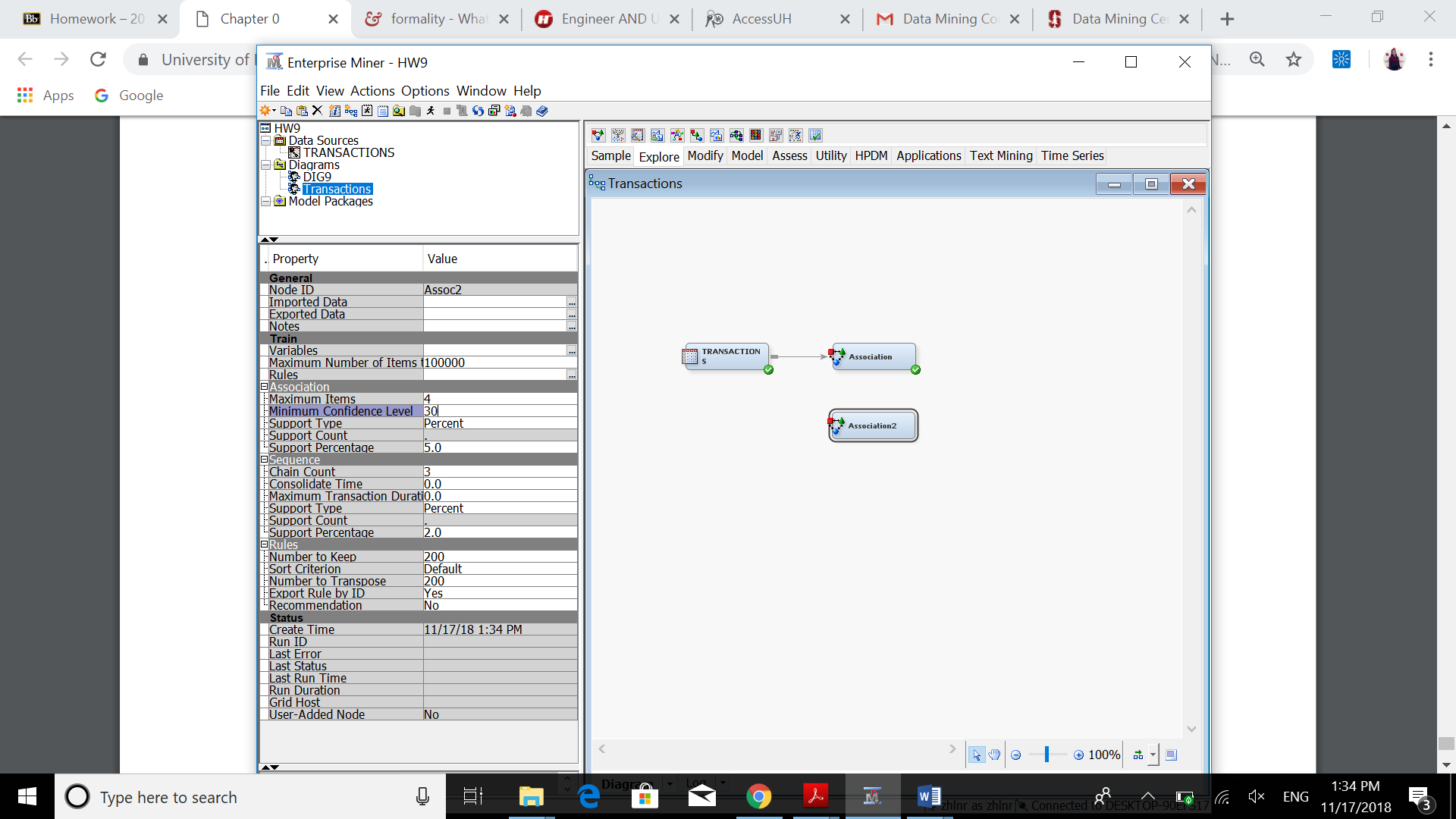
Which rule has this value?

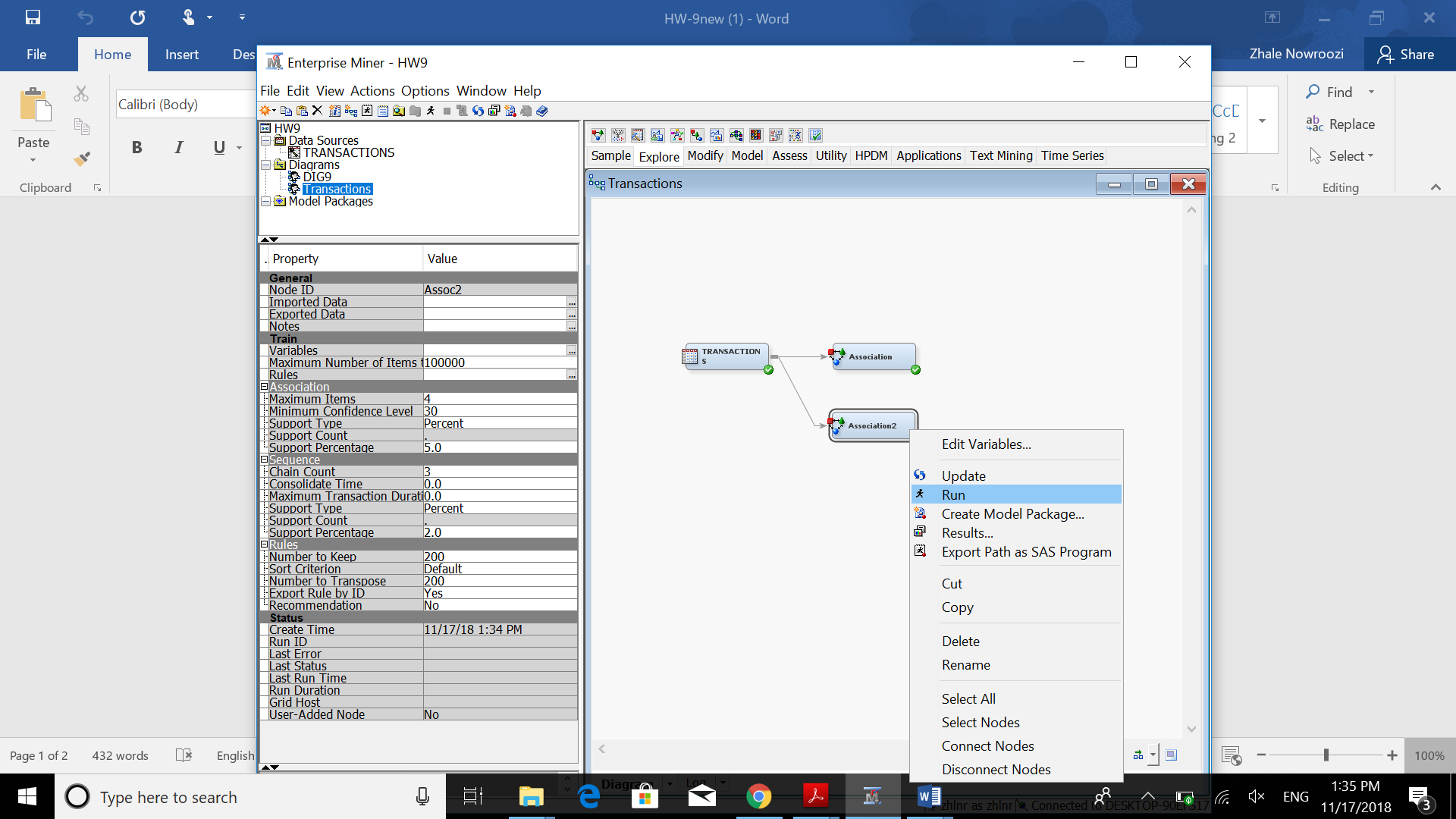
If Perfume 🡺 Toothbrush

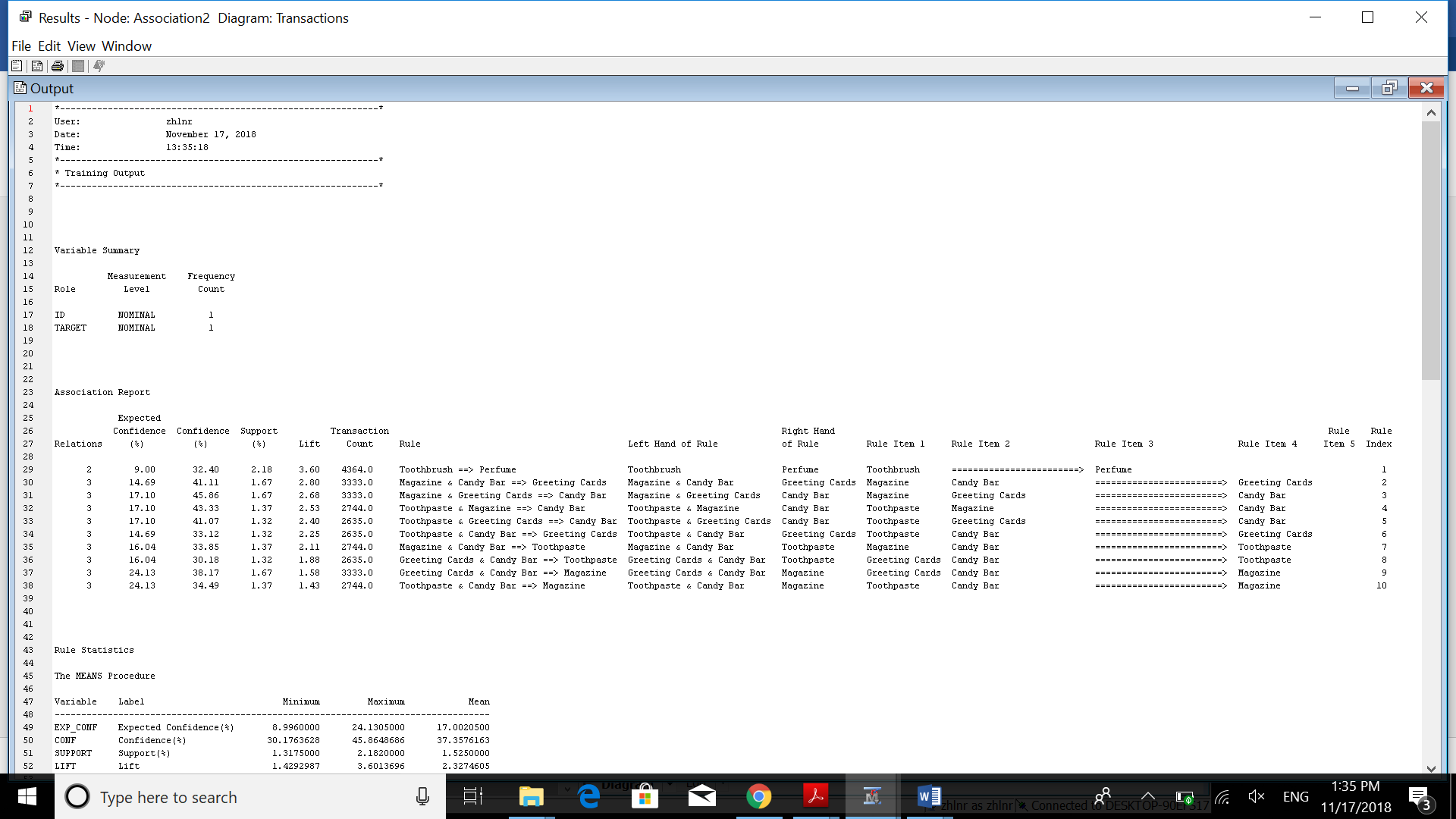
If Toothbrush 🡺 Perfume

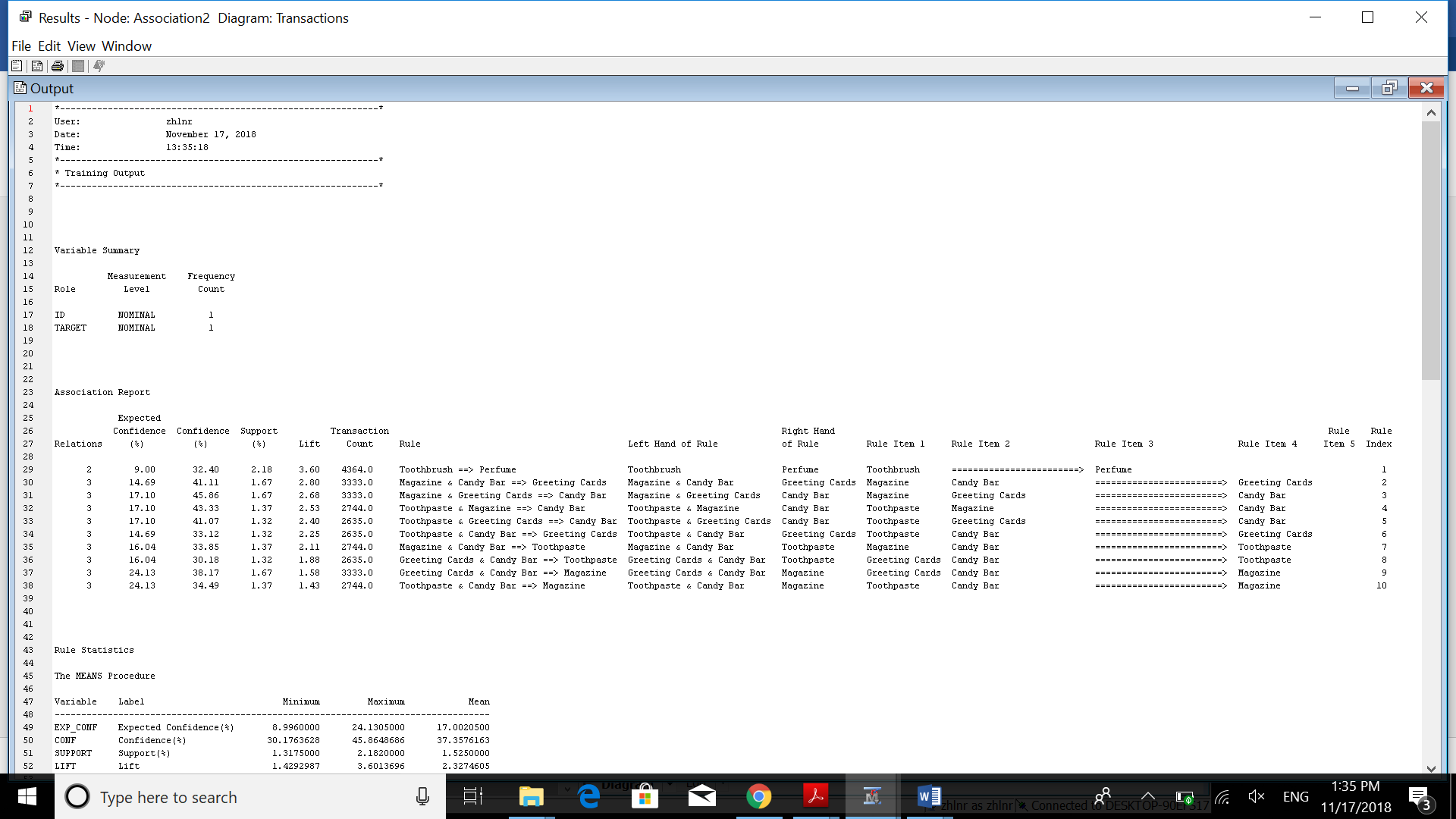
This symmetry is just for these two rules and we do not see this in all of the rules. So we should not generalize this result.

1. Run another Association Analysis by using a minimum confidence of 30 percent.





  
• What is the rule that has the highest lift? 3.60 (shown in blue) If Toothbrush 🡺 Perfume  
• What is the rule that has the highest confident % AND with a lift larger than 1.5? (shown in red)



If Magazine and Greeting Cards 🡺 Candy Bar. Confidence = 45.86% and Lift = 2.68