**Solution Document**

**Assignment 1**

**Business Analytics with SAS – F16**

**(MIS 6324.502)**

Submitted by

Group 8

|  |  |
| --- | --- |
| Mayank Kothari | mkk160130 |
| Shamsundar Kulkarni | sxk162231 |
| Divya Jayaprakash | dxj160830 |
| Gautami Murugan | gxm161330 |
| Vijay Balaji | vxb161530 |

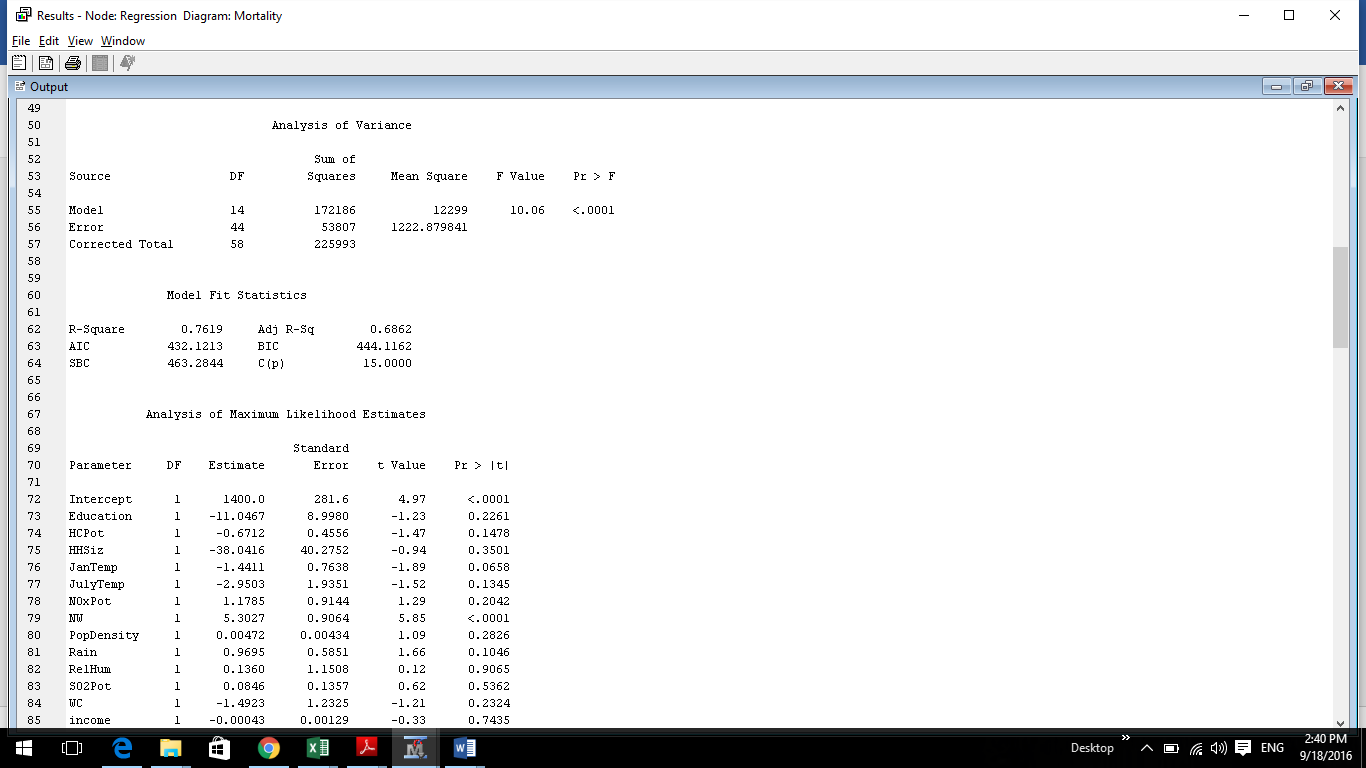
**Exercise 1: Principal Components Analysis**

**Assignment step 1:** New diagram created.

**Assignment step 2:** Dataset file import and variable editing were carried out.

**Assignment Step 3: Run a linear regression using all the original variables**

Result Screenshot:



**a. What are the** R2 **and adjusted** R2 **values?**

**Result:**

R2 Value: 0.7619

Adjusted R2 Value: 0.6862

This can be put as 68.62% of the variations in mortality can be explained by the variation in all other independent variables. Since we have multiple independent variables here, we use adjusted R2.

**b. Which variables (if any) are significant based on the t value statistics and associated probabilities (Pr>|t|)**

**Result:**

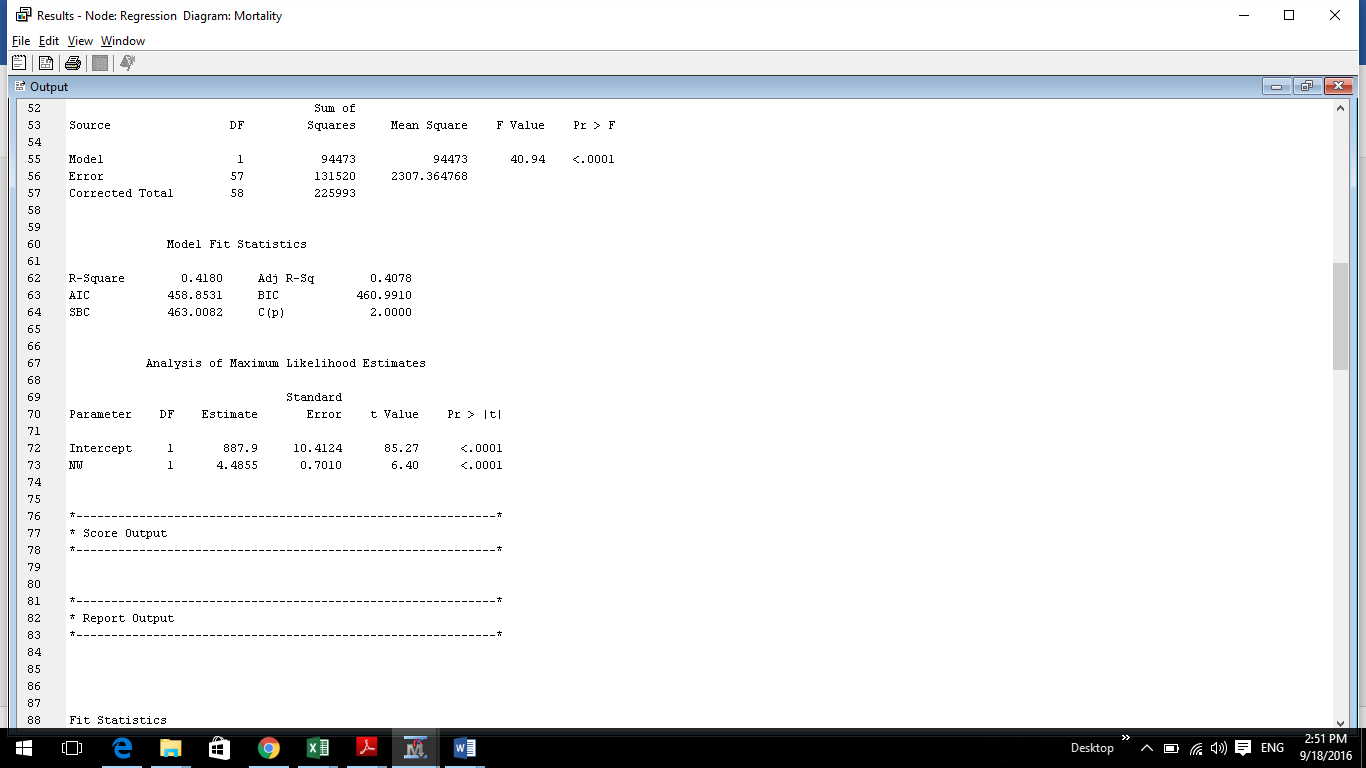
On the Observations of |t| value and associated probability (p-value < 0.05) we can say that the most significant variable is **NW (Percentage of Non-whites)**

Parameter DF Estimate Error t Value Pr > |t|

NW 1 5.3027 0.9064 5.85 <.0001

**Assignment Step 4: Run another linear regression using only the variables that were significant.**

**Result:** We ran the regression model for variable **NW (Percentage of Non-whites)**



**a. Why is this an important step when running regression models?**

**Result:** As NW is the only Significant variable, we reran regression with only this variable to see the result as how this variable is contributing to the variation in mortality.

**b. What are the R2 and adjusted R2 values?**

**Result:**

R2 value: 0.4180

Adjusted R2 value: 0.4078

This can be used to say that almost 41% variation in mortality can be explained by the variation in the percentage of Non-Whites (NW).

**c. Which variables (if any) are significant based on the t value statistics and associated**

**probabilities (Pr>|t|)**

**Result:**

On the Observations of |t| value and associated probability (p-value < 0.05) we can say that the most significant variable is **NW (Percentage of Non-whites)**

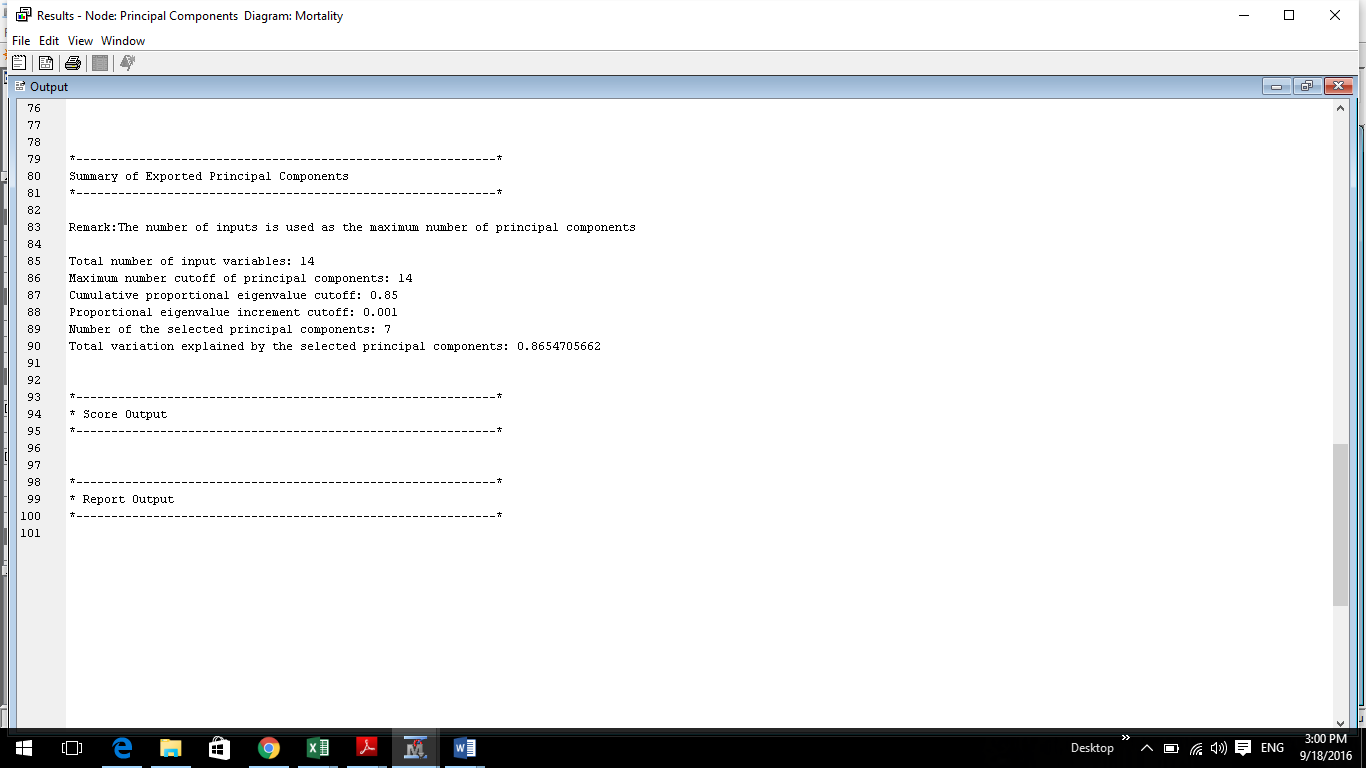
Parameter DF Estimate Error t Value Pr > |t|

NW 1 4.4855 0.7010 6.40 <.0001

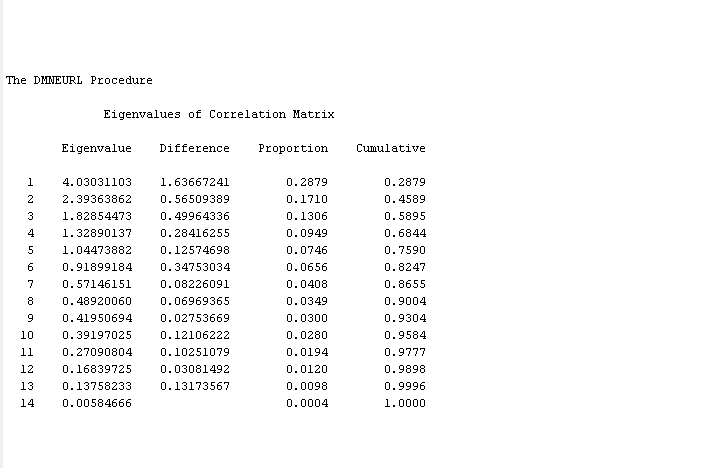
**Assignment Step 5: Apply Principal Component Analysis to this data after changing the Cumulative Eigenvalue Cutoff to 0.85.**

1. **How many principal components are selected?**

**Result:**  Seven components are selected among the 14 input components.

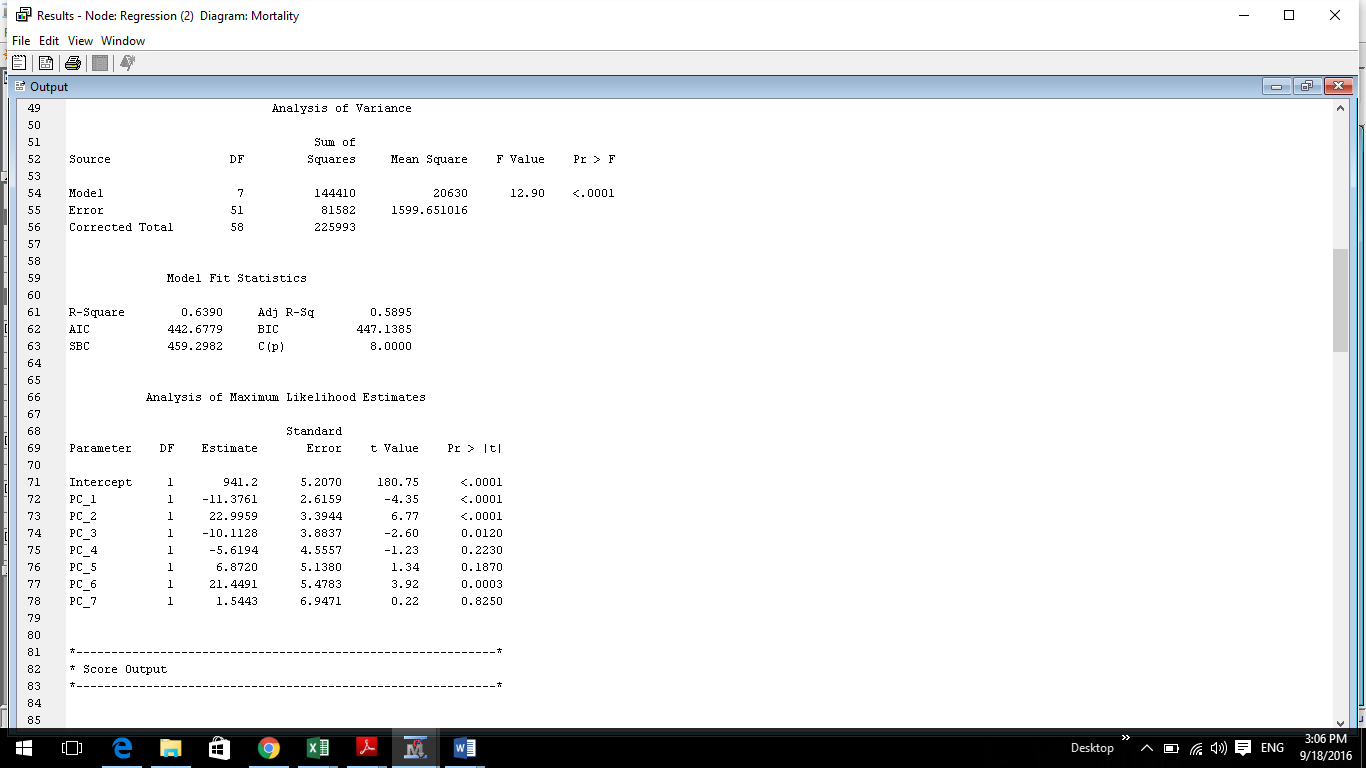


1. **Does this seem like a reasonable number? Explain your answer.**

****

Yes, seven is a reasonable number, but observing from the Eigen Values for all the components we can say that the first six components have a higher value and carry more information than others. Thus we can say that six components are sufficient for our analysis. Also we are almost reducing the number of input components to half, which is a significant dimensionality reduction and this would definitely help in preprocessing of the data.

**Assignment Step 6: Run a linear regression using all the selected principal components**



**a. What are the R2 and adjusted R2 values?**

**Result:**

R2 value: 0.6390

Adjusted R2 value: 0.5895

This can be used to say that almost 58.90% of variation in mortality can be explained by the variation in the selected principal components.

**b. Which variables (if any) are significant based on the t value statistics and associated**

**probabilities (Pr>|t|)**

**Result:**

On the Observations of |t| value and associated probability (p-value < 0.05) we can say that the most significant variables are:

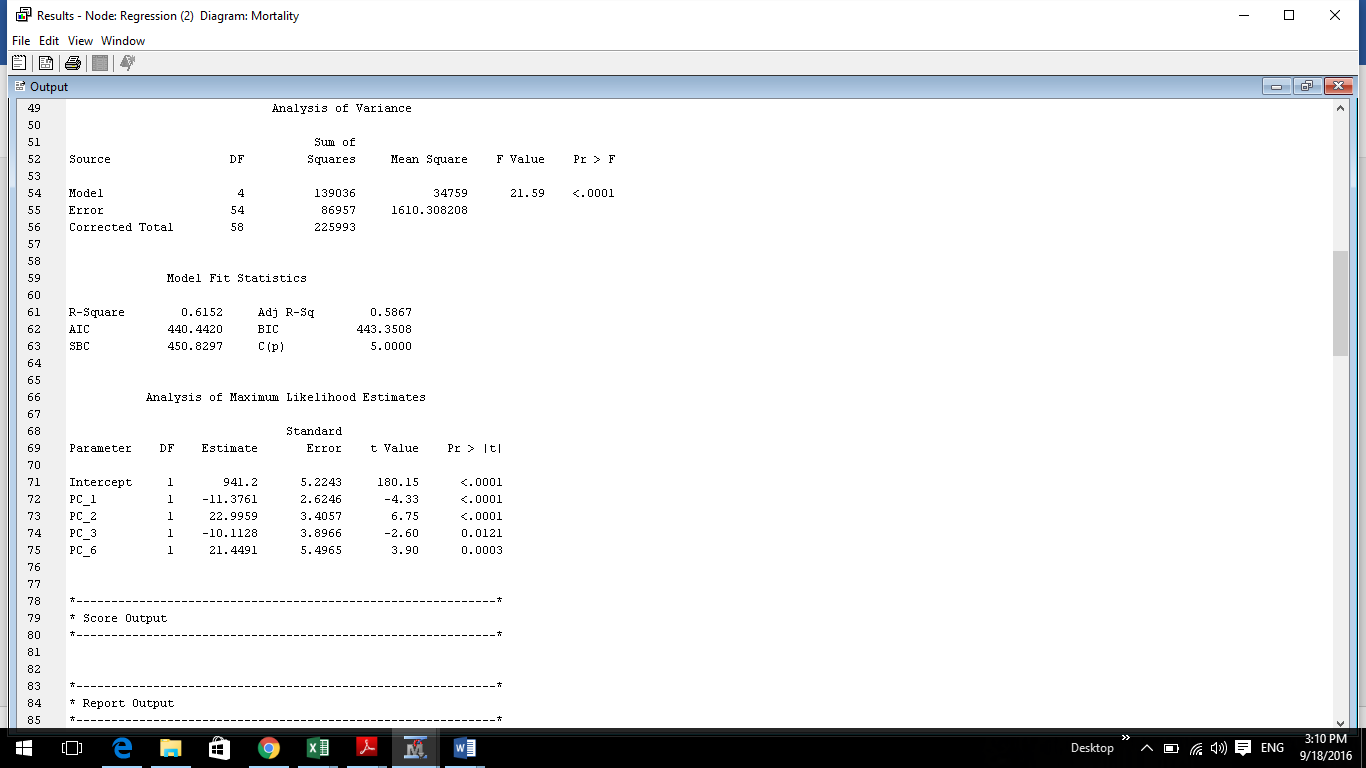
PC\_1

PC\_2

PC\_3

PC\_6

**Assignment Step** **7: Run another linear regression using on the components that were significant.**



**a. What are the R2 and adjusted R2 values?**

**Result:**

R2 value: 0.6152

Adjusted R2 value: 0.5867

This can be used to say that almost 58.67% of variation in mortality can be explained by the variation in the finally selected principal components.

**b. Which variables (if any) are significant based on the t value statistics and associated**

**probabilities (Pr>|t|)**

**Result**: On the Observations of |t| value and associated probability (p-value < 0.05) we can say that all the four selected principle components are significant.

PC\_1

PC\_2

PC\_3

PC\_6

**Assignment Step 8: Comment on the results of the regressions in questions 3‐7. Do you have a preference among them? Explain.**

**Result:**

As we observe from all the above findings, we can say that the step7th regression result is preferred over the others. We see that it has a considerably good adjusted R2 value and also after regression, all the variables, which were given as input to step 7, are observed to be significant.

**Exercise 2: Association Rule Mining**

**Assignment Steps 1- 6:** Creating a new diagram, data source and association node, and editing variables are all done

**Assignment Step 7: Examine the results of the association analysis.**

**a. What is the highest lift value for the resulting rules? Which rule has this value?**

**Result**: Highest Lift value for the resulting rules is 3.6

Perfume ==> Toothbrush

Toothbrush ==> Perfume



**b. Show how this lift value was calculated.**

**Result:**

Lift = Confidence of the rule/Support for the Consequent OR

Lift = Confidence of the rule/ Expected confidence

1. **Perfume ==> Toothbrush**

Confidence of the Rule (Perfume ==> Toothbrush) = 24.26

Support for (Toothbrush) = [Expected confidence value from SAS E Miner] = 6.73

Lift = 24.26/6.73 = 3.6047548291233283803863298662704 ~ 3.60

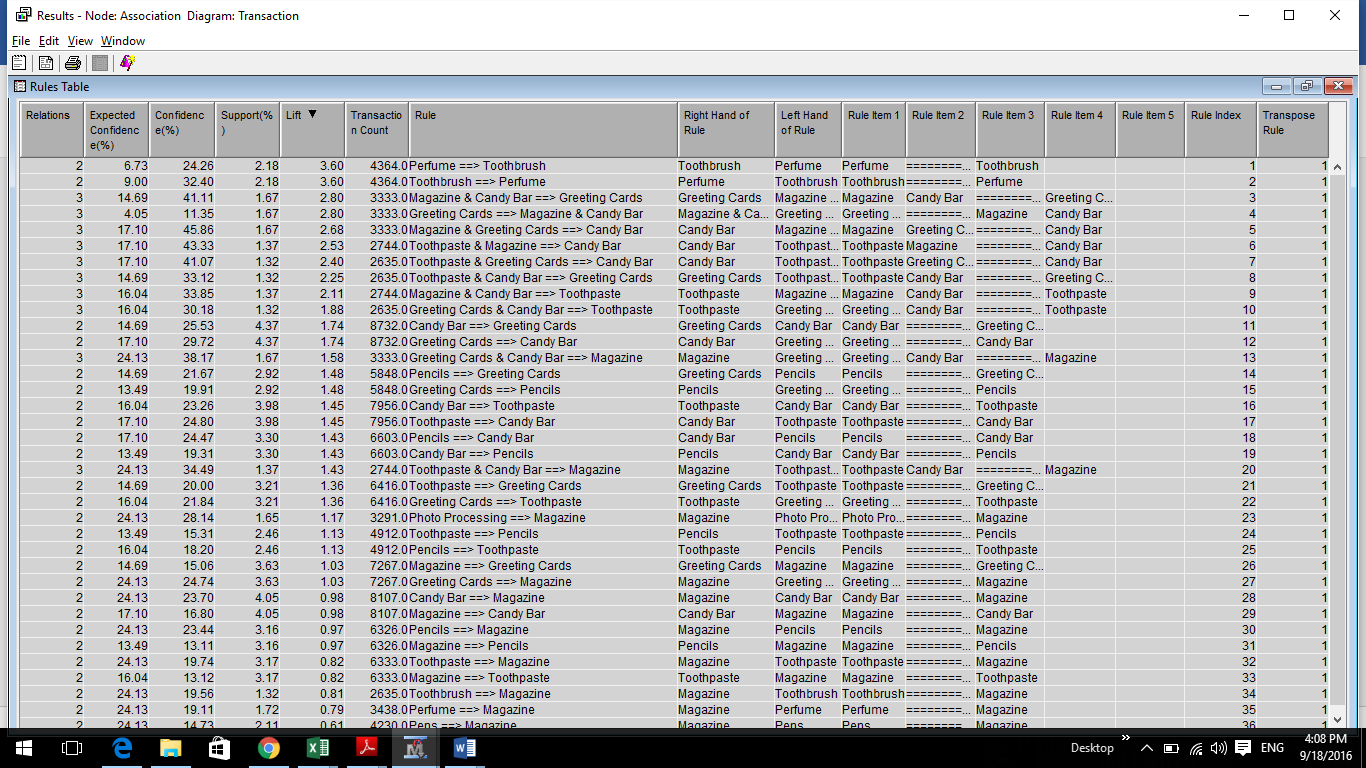
1. **Toothbrush ==> Perfume**

Confidence of the Rule (Perfume ==> Toothbrush) = 32.40

Support for (Toothbrush) = [Expected confidence value from SAS E Miner] =9.00

Lift = 32.40/9.00 = 3.60

**c. Interpret the first five rules in the output in words.**



**1**. **Perfume = => Toothbrush**

2.18 % of transactions(baskets) contains both Perfume and Toothbrush.

24.26 % of transactions(baskets) containing Perfume also contains Toothbrush.

**2. Toothbrush = => Perfume**

2.18 % of transactions(baskets)contains both Toothbrush and perfume.

32.40 % of transactions(baskets)containing Toothbrush also contains perfume.

**3. Magazine & Candy bar = = >Greetings Cards**

1.67 % of transactions(baskets) contain Magazine, Candy bar and Greetings Cards.

41.11 % of transactions(baskets) containing Magazine & Candy bar also contains Greetings Cards.

**4. Greetings Cards = = > Magazine & Candy Bar**

1.67 % of transactions(baskets) contains Greetings Cards, Magazine and Candy Bar.

11.35 % of transactions(baskets) containing Greetings Cards also contains Magazine and Candy Bar.

**5. Magazines and Greetings Cards = = > Candy Bar**

1.67 % of transactions(baskets) contains Magazines, Greetings Cards and Candy Bar.

45.86 % of transactions(baskets) containing Magazines and Greetings Cards also contains Candy Bar.

**d. Reviewing the first 10 rules, comment on their redundancy and how you would assess**

**their utility.**

**Result:**We have highlighted the first 10 rules in the below figure:



There is a lot of redundancy in these rules. First two rules (1. Perfume ==> Toothbrush & 2. Toothbrush ==> Perfume) are redundant. Similarly, we can find redundancy in (3rd ,4th & 5th rules), (6th & 9th rules) and (7th & 10th) rules. So for further analysis we can drop many of these redundant rules and have only the below significant rules:

Perfume == > Toothbrush

Magazine & Candy Bar == > Greetings Cards

Toothpaste & Magazine == > Candy Bar

Toothpaste & Greeting Cards == > Candy Bar

These association rules that we inferred are very useful in drawing marketing strategies

1. **Market Basket:** Arrange the toothpaste, greeting cards and candy bar in adjacent or opposite shelves in the shop [Toothpaste & Greeting Cards == > Candy Bar]
2. **Target marketing**: Market toothpaste to those who buy perfume [Perfume == > Toothbrush]
3. **Analysis of purchasing patterns over time**: Evaluate on which days of the week do people buy magazine and candy bar more. Then provide offers on greeting cards on those days.

[Magazine & Candy Bar == > Greetings Cards]

1. **Collaborative filtering:** Notify “Also buy Candy bar!!!” to those who buy toothpaste and magazine [Toothpaste & Magazine == > Candy Bar]
2. **Customer Profiling:** Analyse what type of customers buy what sort of toothpaste and greeting Cards and then market specific types of candy bars to them