**MKT 6337**

**MARKETING PREDTIVE ANALYTICS PROJECT REPORT**

**SALTY SNACKS**

Group 9

Ayush Singhal

Madurranjani

Rajesh Rajipalem

Siddharth Vasudevan

Yuvanesh Vedaraju

**OBJECTIVES**

- To understanding the market share by sales and volume for brands by descriptive analysis

- To compare Brand preference, based on promotional features across top brands using multi-nominal logit

- To perform clustering using RFM and describe the demographic characteristics of each clusters

- To model ordinal logit regression with rank as a predictor variable to analyze how demographic variable influences

each customer’s rank

-To compare demographics based on preference of product packing type

**DATA CLEANING**

Data Analysis and cleaning on the demographics file leads to understand the distribution of the customer in the given data-set. In this process, we have analyzed and removed missing values and attributes with high null values like MALE\_SMOKE FEM\_SMOKE, Number\_of\_cats, Language, Number\_Of\_TVs\_Hooked\_to\_cable, HISP\_FLAG, HISP\_CAT HH\_Head\_Race\_Race2\_Microwave\_Owned\_by\_HH, ZIPCODE FIPSCODE market\_based\_upon\_zipcode, EXT\_FACT

**FILE MERGING**

In order to merge the store and product data, we used UPC\_Code. For merging store, and product data with the delivery store, we used IRI\_Week. For merging Panel and Demographics data, we used PANID. Finally, to merge the panel data with the store-level data we used UPC\_CODE, IRI\_KEY, IRI\_WEEK.

**DESCRIPTIVE STATISTICS**

**Market Share**

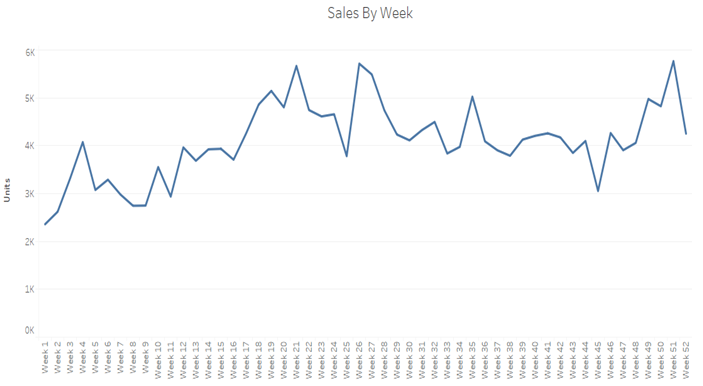
|  |  |
| --- | --- |
|  |  |

**Market share and Package Type**

|  |  |
| --- | --- |
| **image001** | **image003** |

We used Proc Freq on different sizes to find the optimal packaging sizes, after an initial analysis we binned the sizes as follows (Box -> 4.9oz to 6.7oz, party size -> 12oz to 15oz, regular size -> 9oz to 11oz and small size -> .6oz to 1.2oz)

Based on the above sizes (in diagram 1) we can infer that the majority of the salty snack products sold were in the size range of 12oz to 15oz (party size), after this we even made further analysis on the party sizes and found that the majority of the brands sold in party size were from one parent company PepsiCo Inc.



After this we analyzed the sales of units by week and found that during the month of July, October and December there was a huge spike in the overall unit sales, possibility due to holidays

Healthy vs Non-healthy:

We tried to compare, based on people’s preference for heathy on comparison to non-healthy snacks, we included a field, and called it healthy, if the brand had oven or baked in its title and made the others as non-healthy. We conducted the analysis by age and income and Family Size, but we were not able to find any significant difference in any of the segments that made them buy more of healthy products in the salty snacks category.

**MNL Analysis**

We have used MNL Analysis to analyze customer preference among the brands. In order to make it simple, initially we check the top brands purchased by the customer using the following code:

**proc** **sql**;

create table saltsnck.PanelBrand as

select Brand, sum(dollars) as Dollars

from saltsnck.panel\_prod

group by Brand

Order by Dollars DESC;

This has given us the Top brands on which most of the money is spent. From this we have taken the top four brands (i.e) Lays, Tostitos, Wise and Pringles and removed the other brands which are not in study. We have not considered Doritos, Fritos and Ruffles because, they came under Frito lays brand.



We have selected Lays, Tostitos, Wise (which has one of the top sales), and Pringles.

In the next step, we have taken the weekly mean price of all the four brands and merged it with our original data sets. Similarly, for the display, Price reduction flag, features, customers with house and Fat content.

After merging the complete data set, our next task was to clean all the missing values and the dirty data. Once the cleaning is performed, A new data set (panel\_prod\_demo\_cleaned\_MNL) is created since PROC MDC requires that each individual decision maker has one case for each alternative in his choice set.

The MDC procedure produces a summary of model estimation displayed. Since there are multiple observations for each individual, the "Number of Cases" (275456)—that is, the total number of choices faced by all individuals—is larger than the number of individuals, "Number of Observations" (68864). Below shows the frequency distribution of the three choice alternatives.

|  |
| --- |
| Tree Diagram using METHOD = WARD |

The MDC Procedure

Conditional Logit Estimates

|  |
| --- |
| Algorithm converged. |

|  |  |
| --- | --- |
| Model Fit Summary | |
| Dependent Variable | decision |
| Number of Observations | 68864 |
| Number of Cases | 275456 |
| Log Likelihood | -72134 |
| Log Likelihood Null (LogL(0)) | -95466 |
| Maximum Absolute Gradient | 0.0002958 |
| Number of Iterations | 6 |
| Optimization Method | Newton-Raphson |
| AIC | 144346 |
| Schwarz Criterion | 144702 |

The parameter estimates can be used to forecast the choice probability of individuals that are not in the input data set. As shown below, are the variables which significantly impacting the decision variables.

|  |
| --- |
| Tree Diagram using METHOD = WARD |

The MDC Procedure

Conditional Logit Estimates

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | |
| **Parameter** | **DF** | **Estimate** | **Standard** | **t Value** | **Approx** |
| **Error** | **Pr > |t|** |
| **dBR2** | 1 | -1.7795 | 0.0766 | -23.24 | 0.0001 |
| **dBR3** | 1 | -0.3538 | 0.0757 | -4.67 | 0.0001 |
| **dBR4** | 1 | -1.0448 | 0.0848 | -12.33 | 0.0001 |
| **Price** | 1 | -1.1935 | 0.4807 | -2.48 | 0.013 |
| **Display** | 1 | 0.3377 | 0.00873 | 38.67 | 0.0001 |
| **Feature** | 1 | 0.6979 | 0.0121 | 57.51 | 0.0001 |
| **LOW\_FAT2** | 1 | 0.9347 | 0.0378 | 24.76 | 0.0001 |
| **LOW\_FAT3** | 1 | -2.0724 | 0.129 | -16.06 | 0.0001 |
| **LOW\_FAT4** | 1 | 1.7568 | 0.0363 | 48.39 | 0.0001 |
| **inc2** | 1 | 0.0639 | 0.00552 | 11.59 | 0.0001 |
| **inc3** | 1 | -0.0147 | 0.00557 | -2.64 | 0.0083 |
| **inc4** | 1 | 0.00011 | 0.00619 | 0.02 | 0.9858 |
| **HH\_OWN2** | 1 | 0.1815 | 0.0355 | 5.11 | 0.0001 |
| **HH\_OWN3** | 1 | 0.0892 | 0.0329 | 2.71 | 0.0067 |
| **HH\_OWN4** | 1 | -0.0201 | 0.0375 | -0.54 | 0.5918 |
| **Age\_Male\_HH2** | 1 | -0.0698 | 0.0104 | -6.71 | 0.0001 |
| **Age\_Male\_HH3** | 1 | 0.0156 | 0.00987 | 1.58 | 0.1141 |
| **Age\_Male\_HH4** | 1 | -0.0713 | 0.0115 | -6.18 | 0.0001 |
| **HH\_Male\_EDU2** | 1 | 0.0374 | 0.00908 | 4.12 | 0.0001 |
| **HH\_Male\_EDU3** | 1 | -0.012 | 0.00957 | -1.25 | 0.2096 |
| **HH\_Male\_EDU4** | 1 | 0.0521 | 0.0105 | 4.97 | 0.0001 |
| **HH\_Male\_OCC2** | 1 | -0.0122 | 0.00379 | -3.21 | 0.0013 |
| **HH\_Male\_OCC3** | 1 | 0.00599 | 0.00402 | 1.49 | 0.1358 |
| **HH\_Male\_OCC4** | 1 | -0.0106 | 0.00437 | -2.43 | 0.015 |
| **Age\_Female\_HH2** | 1 | -0.1023 | 0.0106 | -9.61 | 0.0001 |
| **Age\_Female\_HH3** | 1 | 0.0234 | 0.01 | 2.34 | 0.0193 |
| **Age\_Female\_HH4** | 1 | -0.0245 | 0.0115 | -2.12 | 0.0337 |
| **HH\_Female\_EDU2** | 1 | 0.119 | 0.00829 | 14.35 | 0.0001 |
| **HH\_Female\_EDU3** | 1 | -0.0795 | 0.0089 | -8.93 | 0.0001 |
| **HH\_Female\_EDU4** | 1 | 0.0365 | 0.00958 | 3.81 | 0.0001 |
| **HH\_Female\_OCC2** | 1 | 0.0078 | 0.00283 | 2.76 | 0.0058 |
| **HH\_Female\_OCC3** | 1 | -0.0124 | 0.0031 | -4.02 | 0.0001 |
| **HH\_Female\_OCC4** | 1 | 0.0102 | 0.00325 | 3.15 | 0.0016 |
| **nmemb2** | 1 | 0.00108 | 0.0121 | 0.09 | 0.9287 |
| **nmemb3** | 1 | -0.0439 | 0.013 | -3.38 | 0.0007 |
| **nmemb4** | 1 | -0.0745 | 0.0144 | -5.17 | 0.0001 |
| **kid2** | 1 | -0.1474 | 0.0333 | -4.43 | 0.0001 |
| **kid3** | 1 | -0.0556 | 0.0366 | -1.52 | 0.1283 |
| **kid4** | 1 | 0.2306 | 0.0392 | 5.88 | 0.0001 |

Observing the above table, we have come up with following interpretation:

1. Lays is preferred over all the other brands
2. Decision is significantly dependent on price. As price decrease the customer tendency to purchase increases.
3. If display increase the customer will purchase brand
4. Low fat lovers highly prefer brand 4 over other brands
5. High income group customers prefer brand 1, 2 & 4
6. House hold owner prefer brand 2 over other brands
7. The head of household male with age of 45-54 prefer brand 1
8. Mostly male laborers prefer brand 1
9. College graduated male prefer brand 2, 4 and 1
10. Age group of female between 35-44 prefer brand 1
11. Female machine operators prefer brand 2 & 4
12. Family size of 2 or 3 people prefer brand 1
13. Kids prefer brand 4

**Brand 1- Lays**

High income group customers

Age group of 45-55 males who are head of household

Male Laborers

College going males

Family size of 2- 3 people

**Brand 2- Tostitos**

Preferred by household customers

High income group and low fat eating customers

Female machine operators

College graduated males

**Brand 3 – Wise**

No major insight from MNL analysis

**Brand 4 – Pringles**

Preferred by kids

High income group customers

Low fat customers

Female machine operators

College going males

**RFM Analysis:**

All the 3 panel data are merged with product details file using UPC code.

Using the merged file, records only with brand “pringles” are written in new dataset.

Using PROC SQL, new file is created with PANID, max(week), sum(units) and sum(dollars) for each customer.

PROC RANK is used to rank the customers based on Recency, Frequency and Monetary and grouped as 5 clusters.

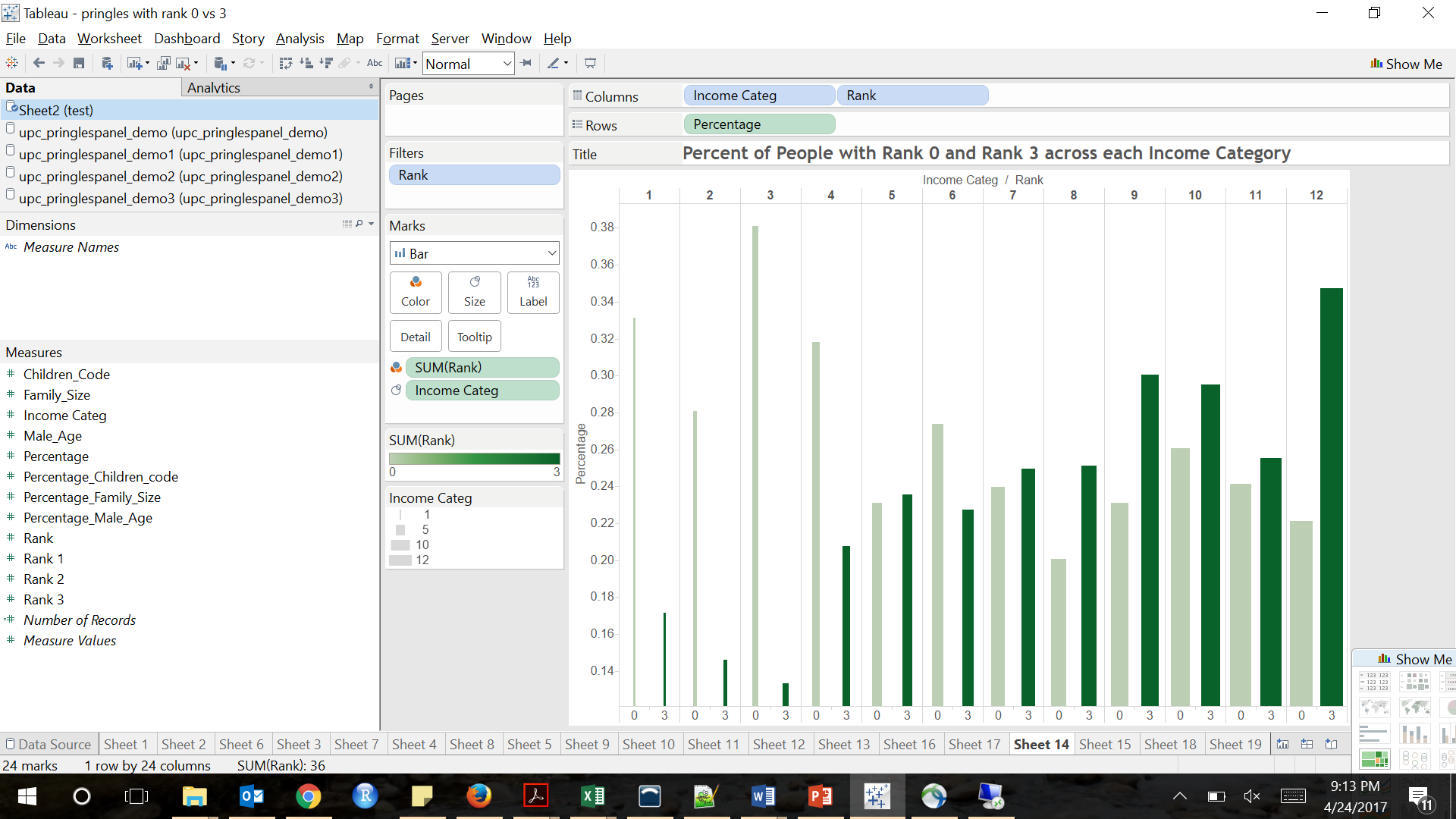
We assigned weights as R=2,F=1,M=3. We did not assign new weight for Frequency, because a customer can buy less value chips many times and another customer can buy high value chips only once. Hence monetary value is more important than frequency. We noticed the Recency Rank 4 was given only for Dec last week of the year and Recency rank3 was given from Oct 22 to Dec 23rd. For first 6 months of the year, system assigned rank 0 and 1, and for next 6 months system assigned rank 2, 3, 4.

Hence we wanted to give weight for R. High weightage is given for M since that decides the profit of the company. (Assumption: Since M is the most important deciding factor in my segmentation of clusters/customers. People who bring more money are profitable.)

Using PROC RANK again on this summed weighted score, we ranked again and clustered as group of 4 (rank 0,1,2 and 3).

We exported this ranked cluster file to Tableau and compared the demographic profiles of these clusters. We wanted to differentiate Rank0 and Rank3 customers, so that we can understand who are Rank3 customers.

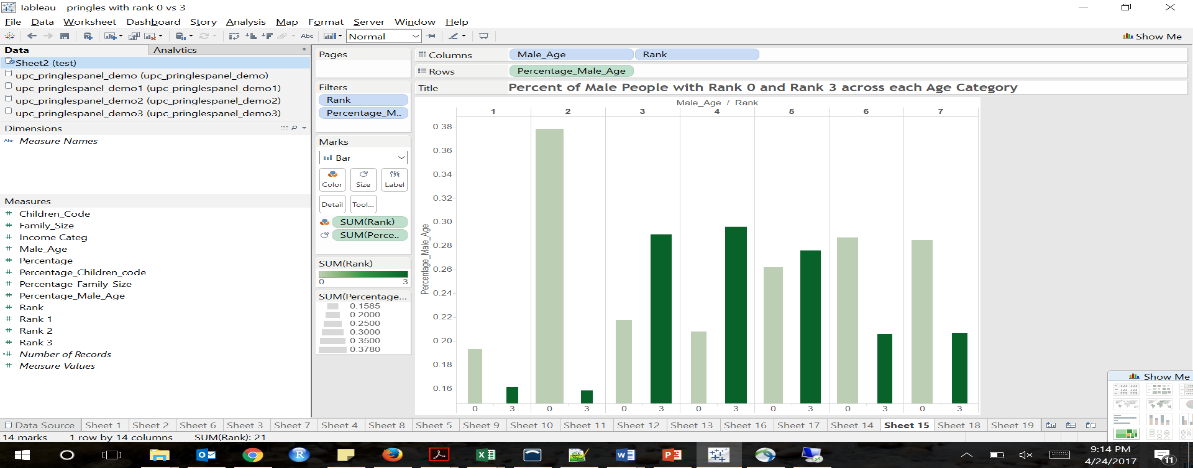
**Comparison of Income category of Cluster1 and Cluster4 (Rank0 and Rank3)**

From the below graph, we can infer that Rank3 customers lie in the high income category. Below chart is drawn in a way to understand what is the percentage of Rank0,1,2,3 customers across each income category. Income category 7 to 12 has more percentage of Rank3 customers (Especially category 9 and 12). 30% of the category 9 are Rank3 and 35% of the category 12 are Rank3.

Income Group 7 to 12

**Comparison of Age group with Cluster1 and Cluster4(Rank0 and Rank3)**

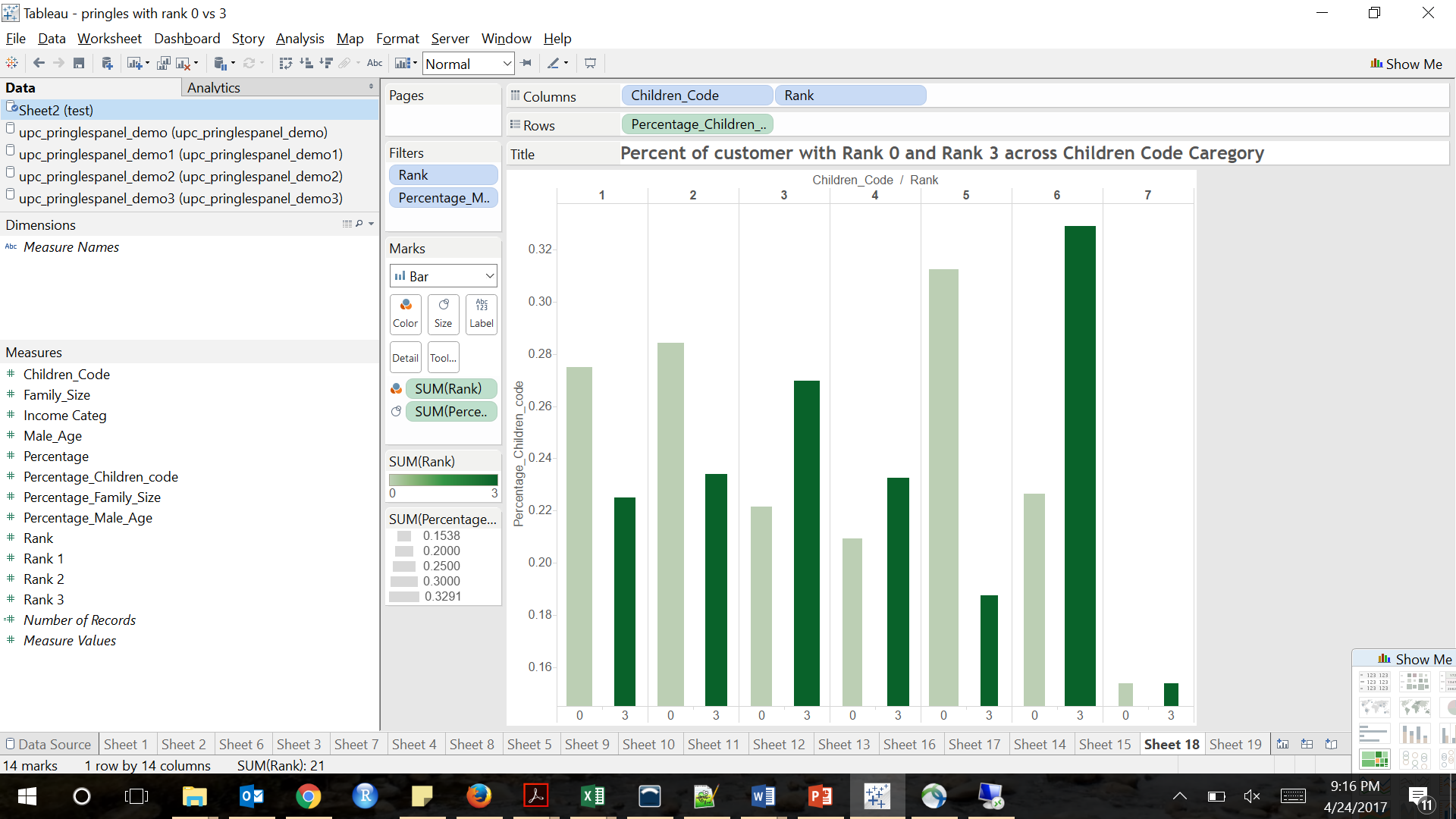
From the below graph, we can infer that Rank3 customers lie in the category 3 and 4. Given all age category, approximately 30% of the Rank3 customers present in category 3 and 4.



Male Age Group 3 & 4

**Comparison of Children group code with Cluster1 and Cluster4(Rank0 and Rank3)**

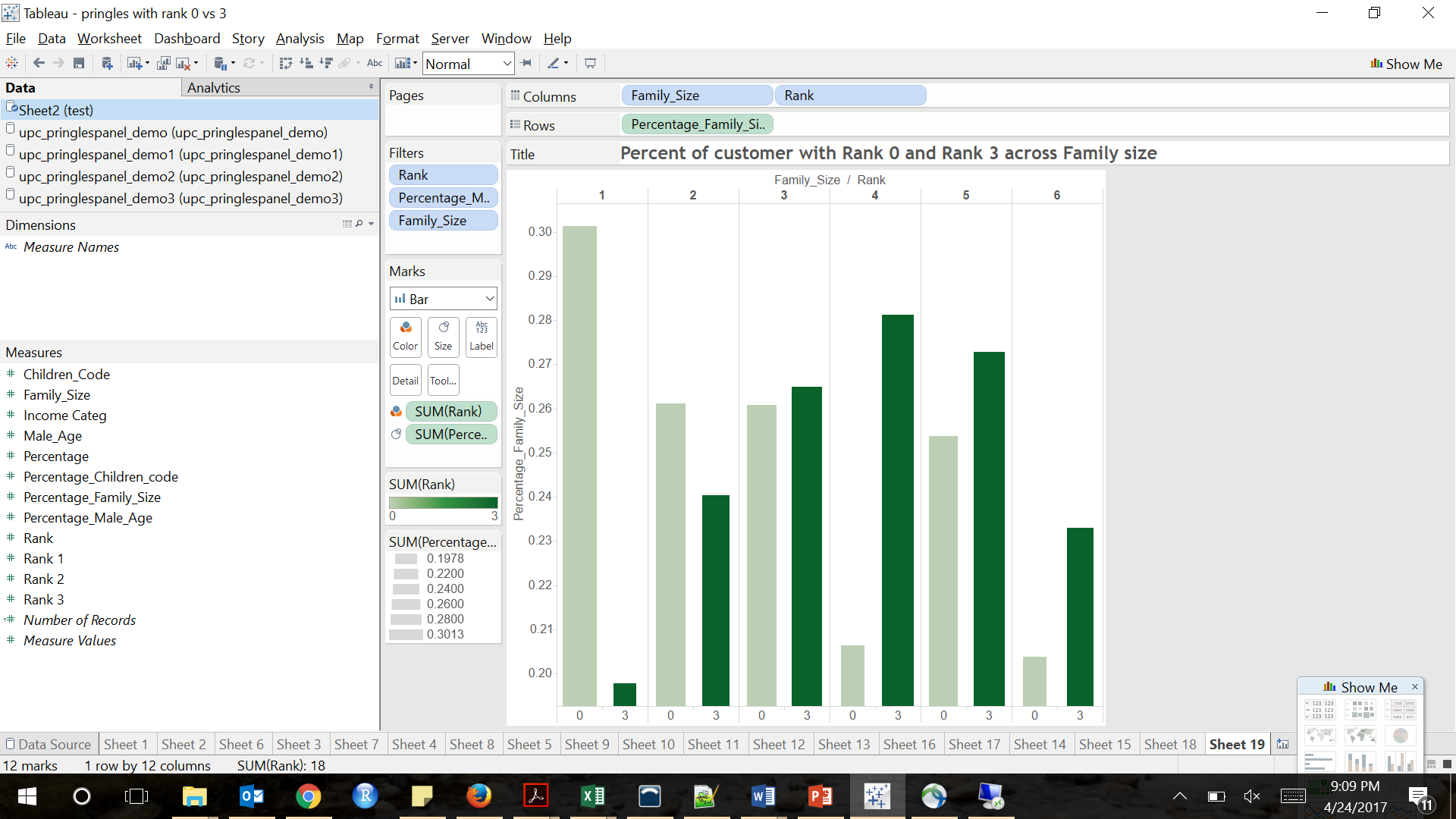
From the below graph, we can infer that Rank3 customers lie in the category 3 and 6. Given all age children category, approximately 33% of the Rank3 customers present in category 6 and 27% of the Rank3 customers present in category3.



Children Group 3 & 6

**Comparison of Family size with Cluster1 and Cluster4(Rank0 and Rank3)**

From the below graph, we can infer that Rank3 customers predominantly lie in the category 4. Given all family size category, approximately 28% of the Rank3 customers present in category 4 whereas only 20% of the Rank0 customers present in this category.



Family size 4 & 5

**Ordinal Logit:**

Next we wanted to see whether these demographics really help us to understand the ranking of customers. So with y variable as rank(0 to 3) and x variables as demographics, we modelled ordinal logistic regression to understand the effect of these demographics in the ranking of customers.

**proc** **logistic** data =saltproducts\_pringles\_demo\_trim DESC;

class Combined\_Pre\_Tax\_Income\_of\_HH (REF="1")

Family\_Size (REF="1")

HH\_RACE (REF="1")

Type\_of\_Residential\_Possession Age\_Group\_Applied\_to\_Male\_HH (REF="1") Education\_Level\_Reached\_by\_Male (REF="1")

Occupation\_Code\_of\_Male\_HH (REF="0") Male\_Working\_Hour\_Code (REF="1")

Age\_Group\_Applied\_to\_Female\_HH (REF="1") Education\_Level\_Reached\_by\_Femal (REF="1")

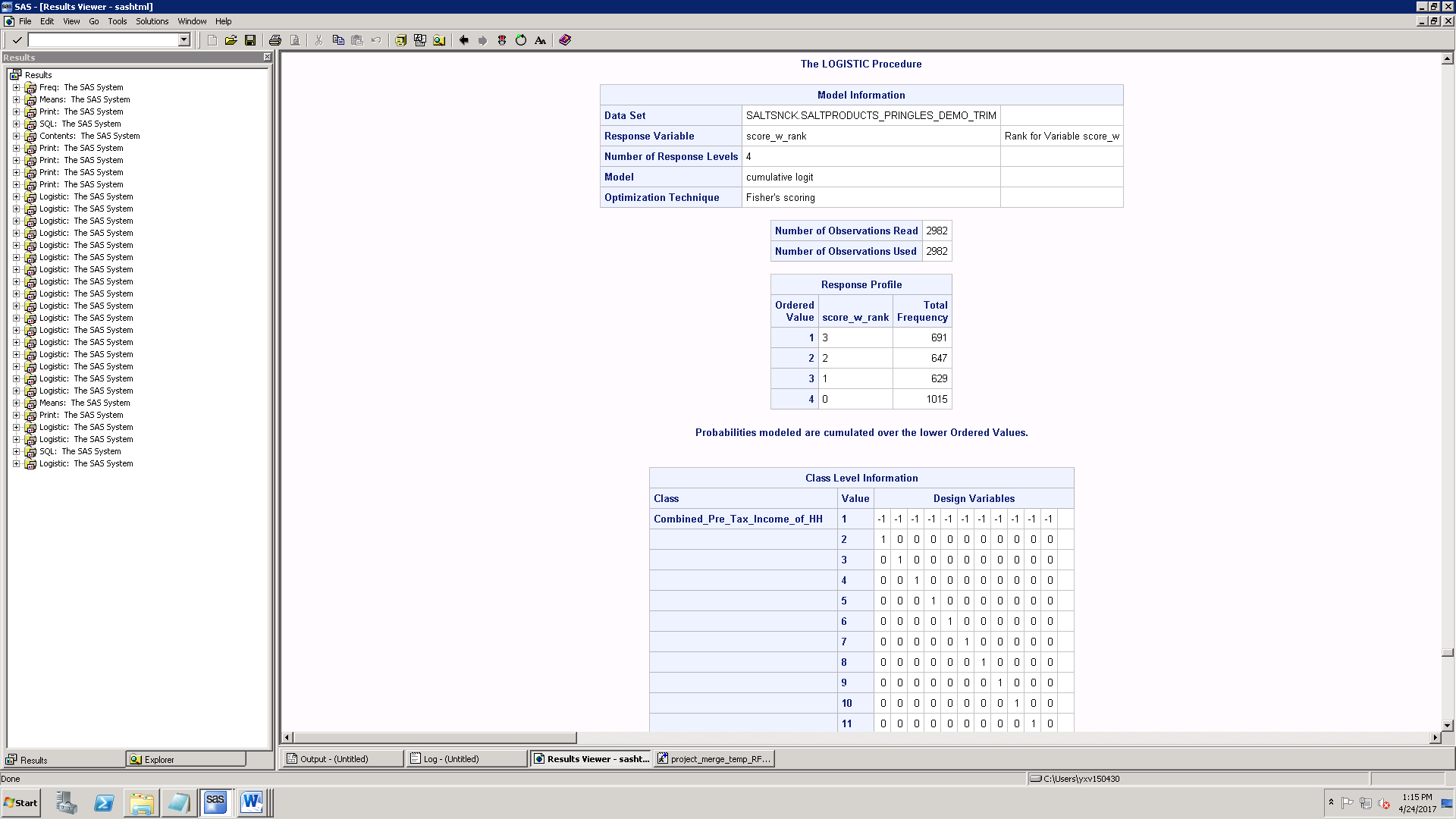
Occupation\_Code\_of\_Female\_HH (REF="0") Female\_Working\_Hour\_Code (REF="1") Children\_Group\_Code Marital\_Status (REF="1");

id PANID;

model score\_w\_rank = Combined\_Pre\_Tax\_Income\_of\_HH Family\_Size HH\_RACE Type\_of\_Residential\_Possession Age\_Group\_Applied\_to\_Male\_HH Education\_Level\_Reached\_by\_Male Occupation\_Code\_of\_Male\_HH Male\_Working\_Hour\_Code Age\_Group\_Applied\_to\_Female\_HH Education\_Level\_Reached\_by\_Femal Occupation\_Code\_of\_Female\_HH Female\_Working\_Hour\_Code Children\_Group\_Code Marital\_Status;

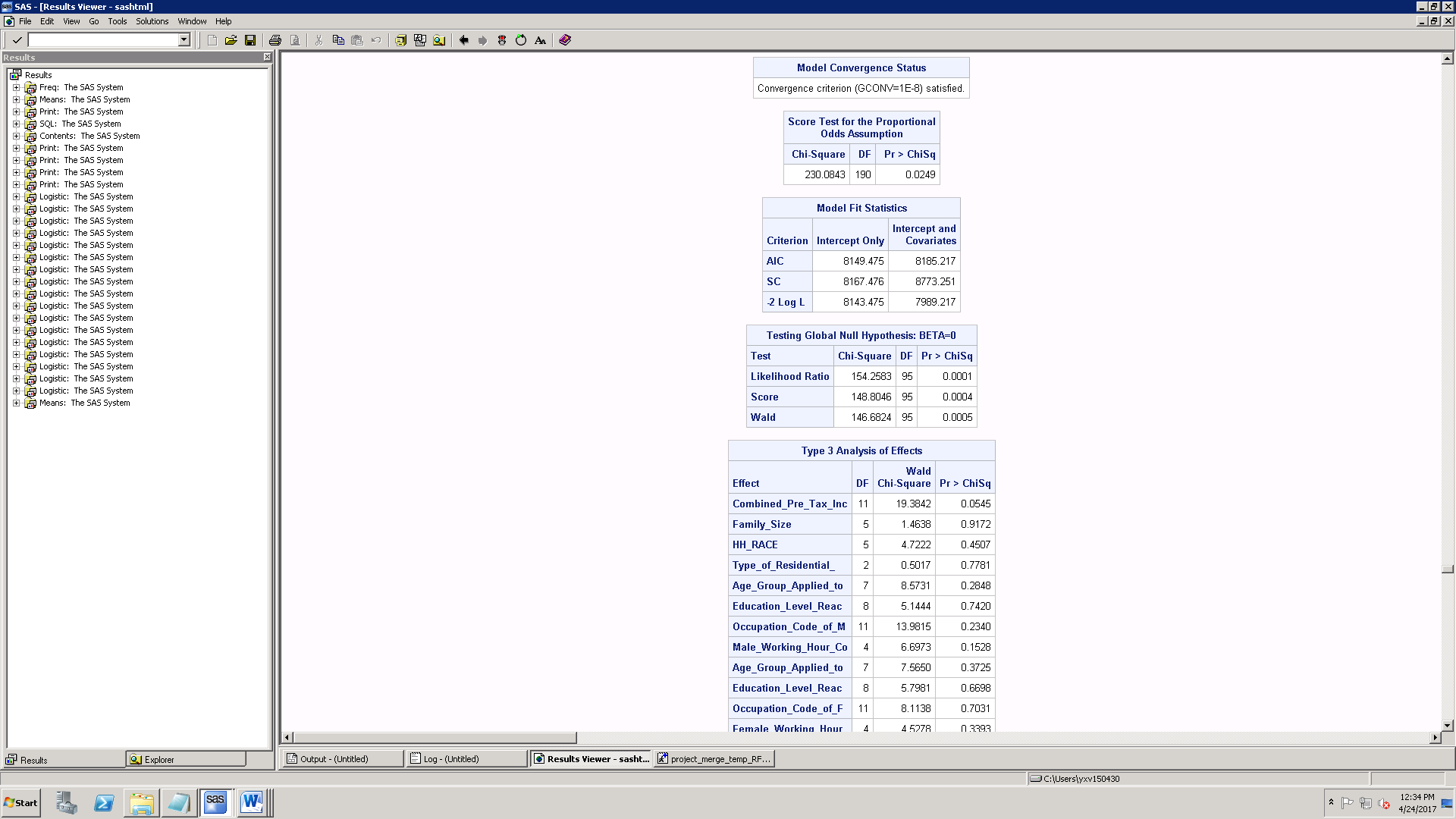
**run**;

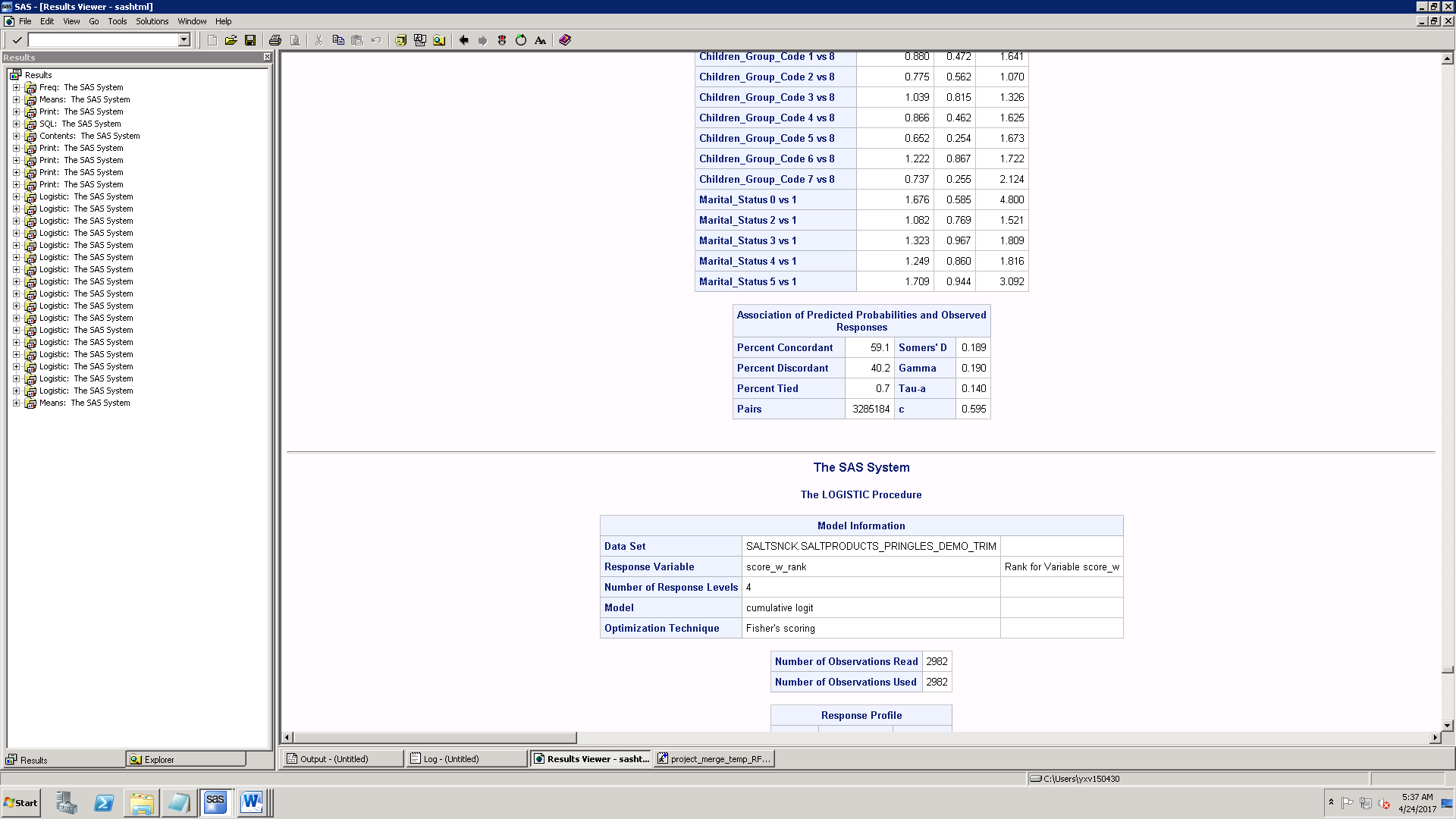
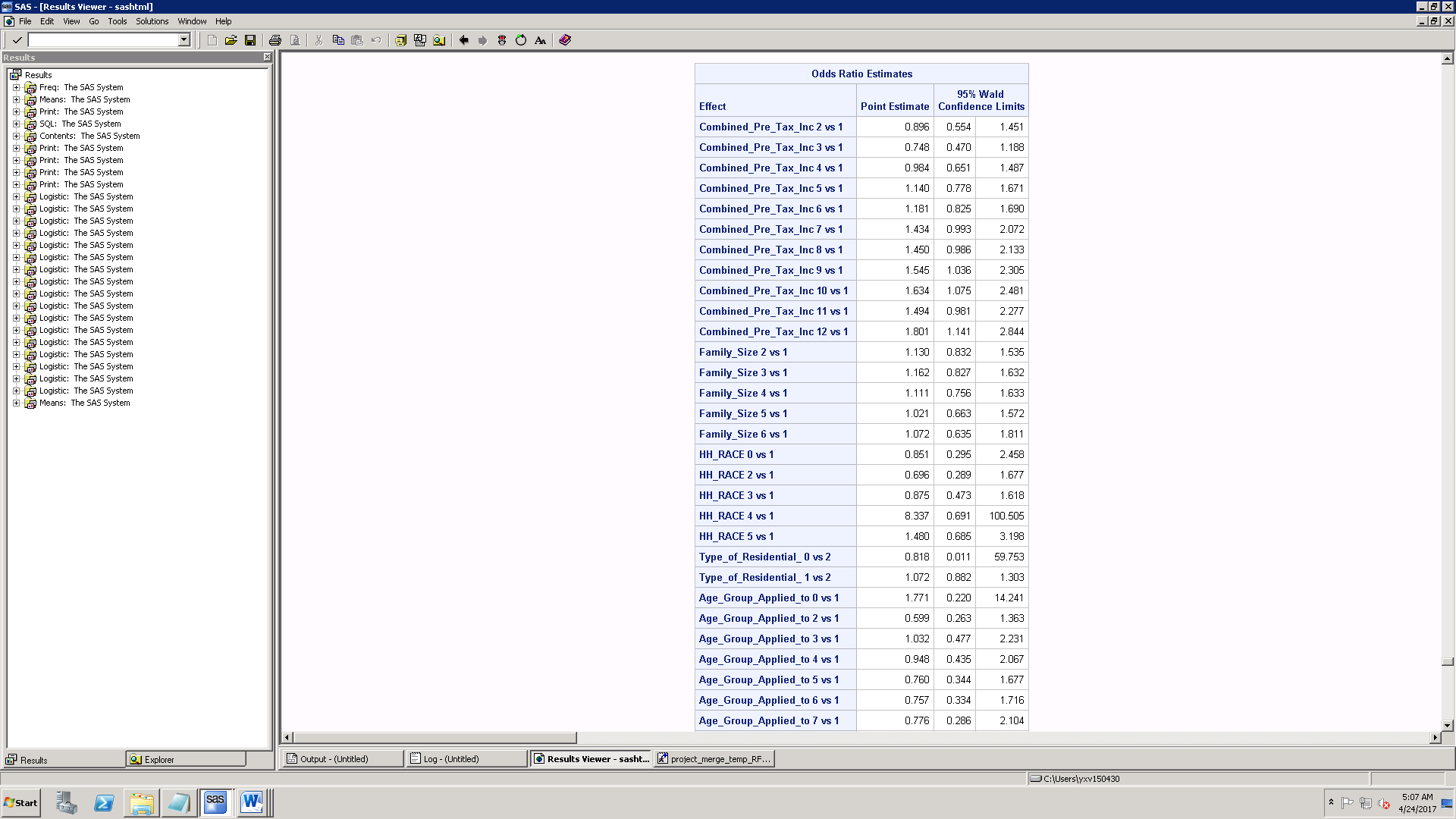
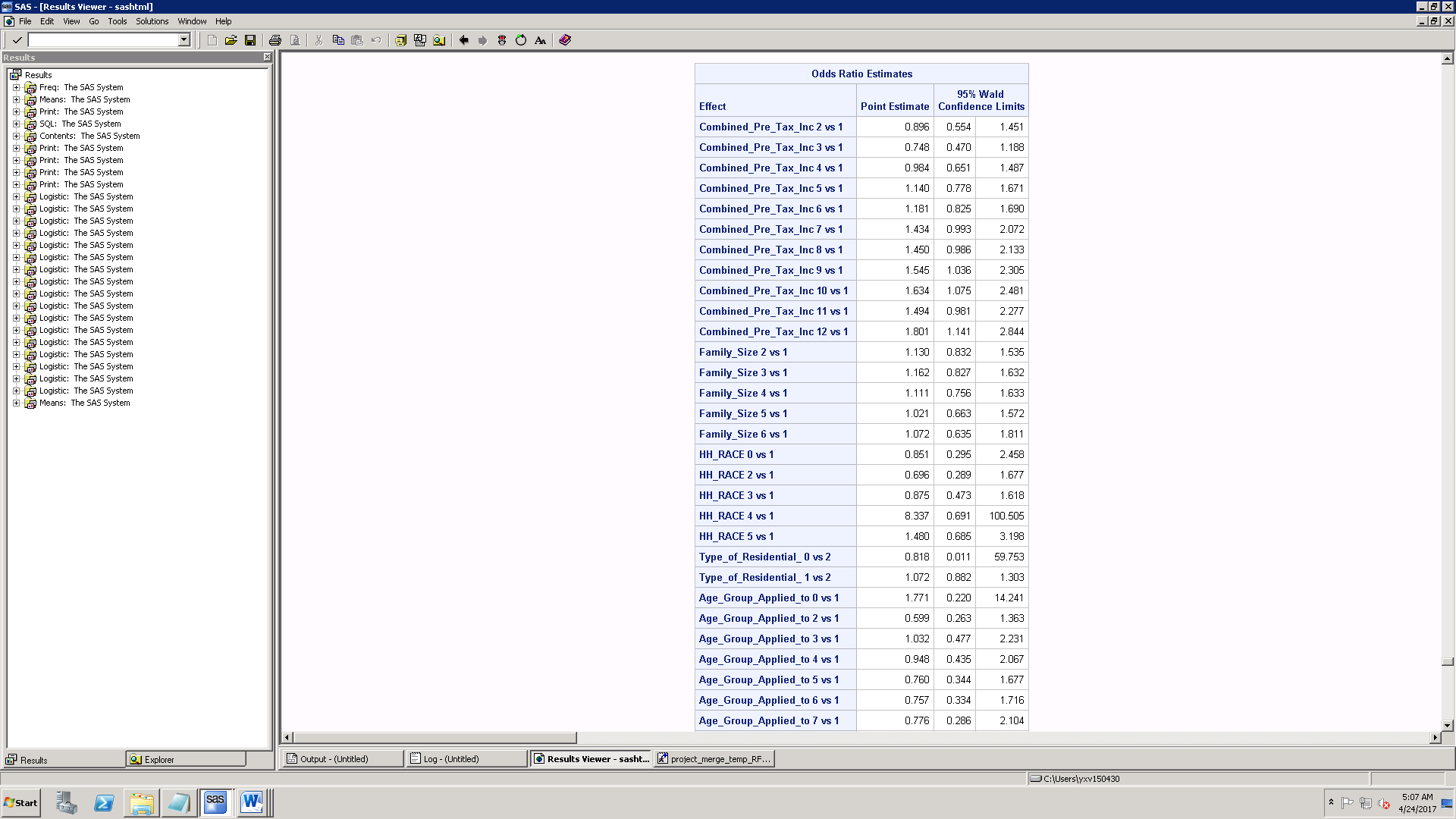
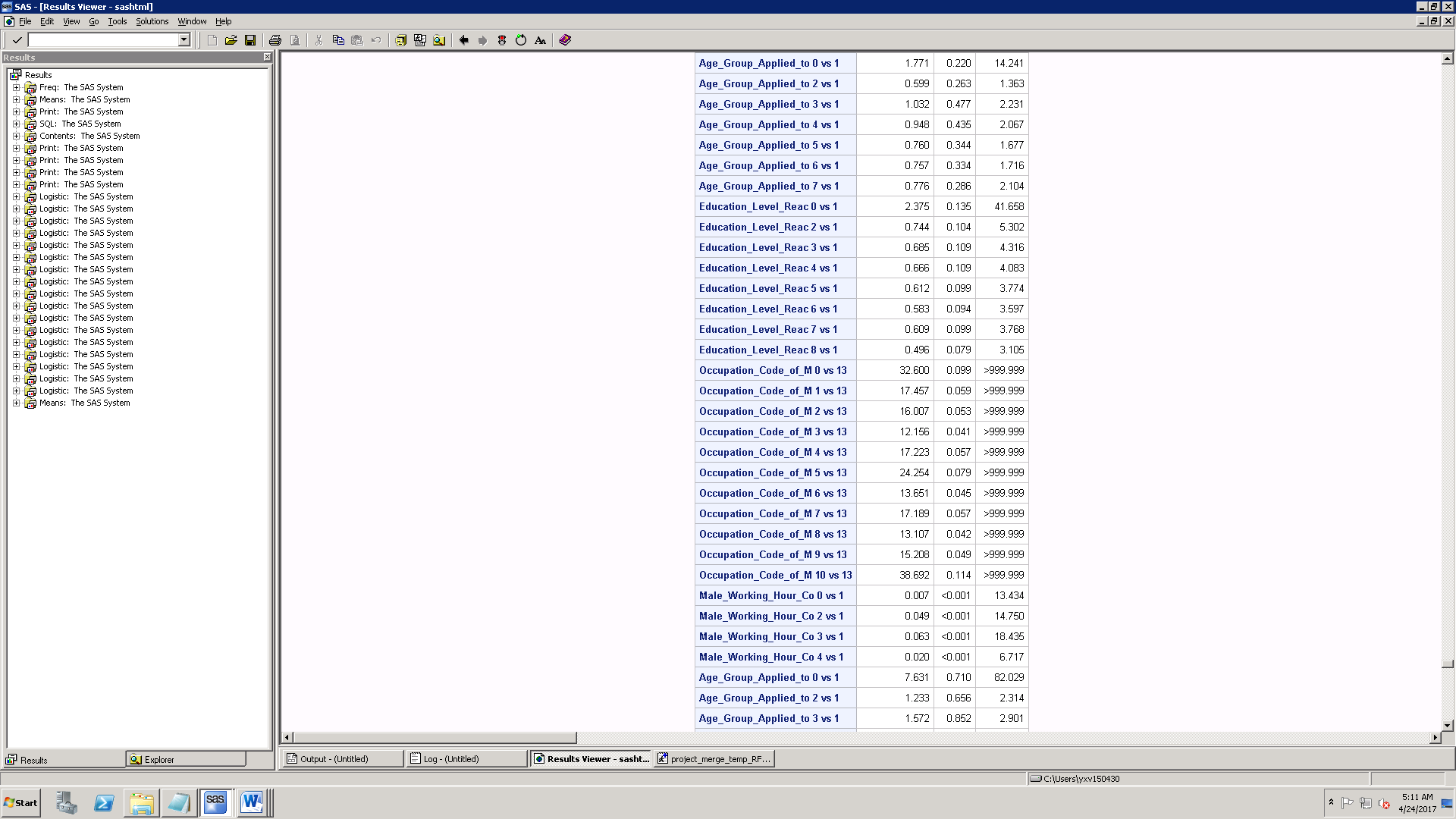
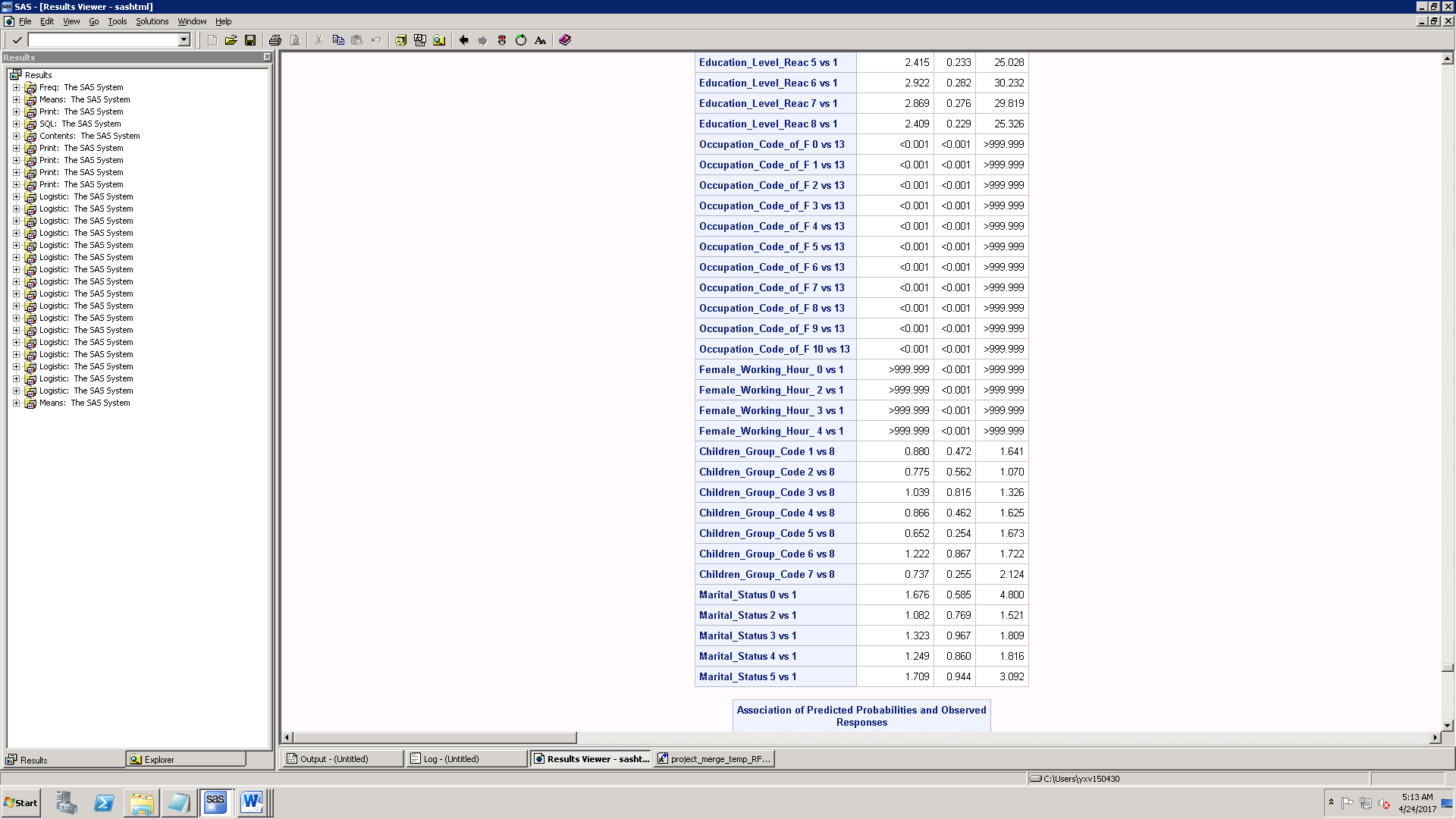
Model Output:

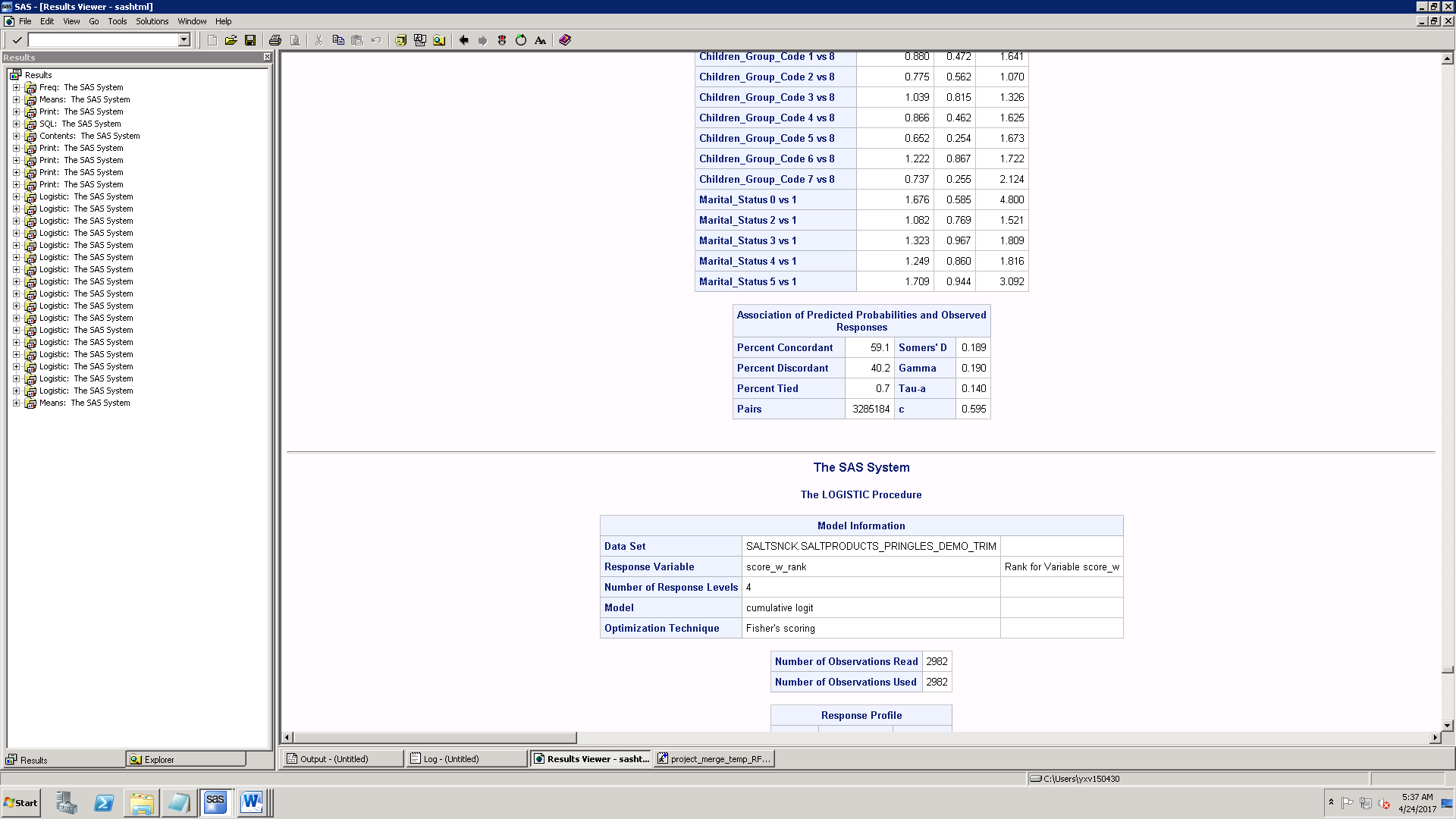
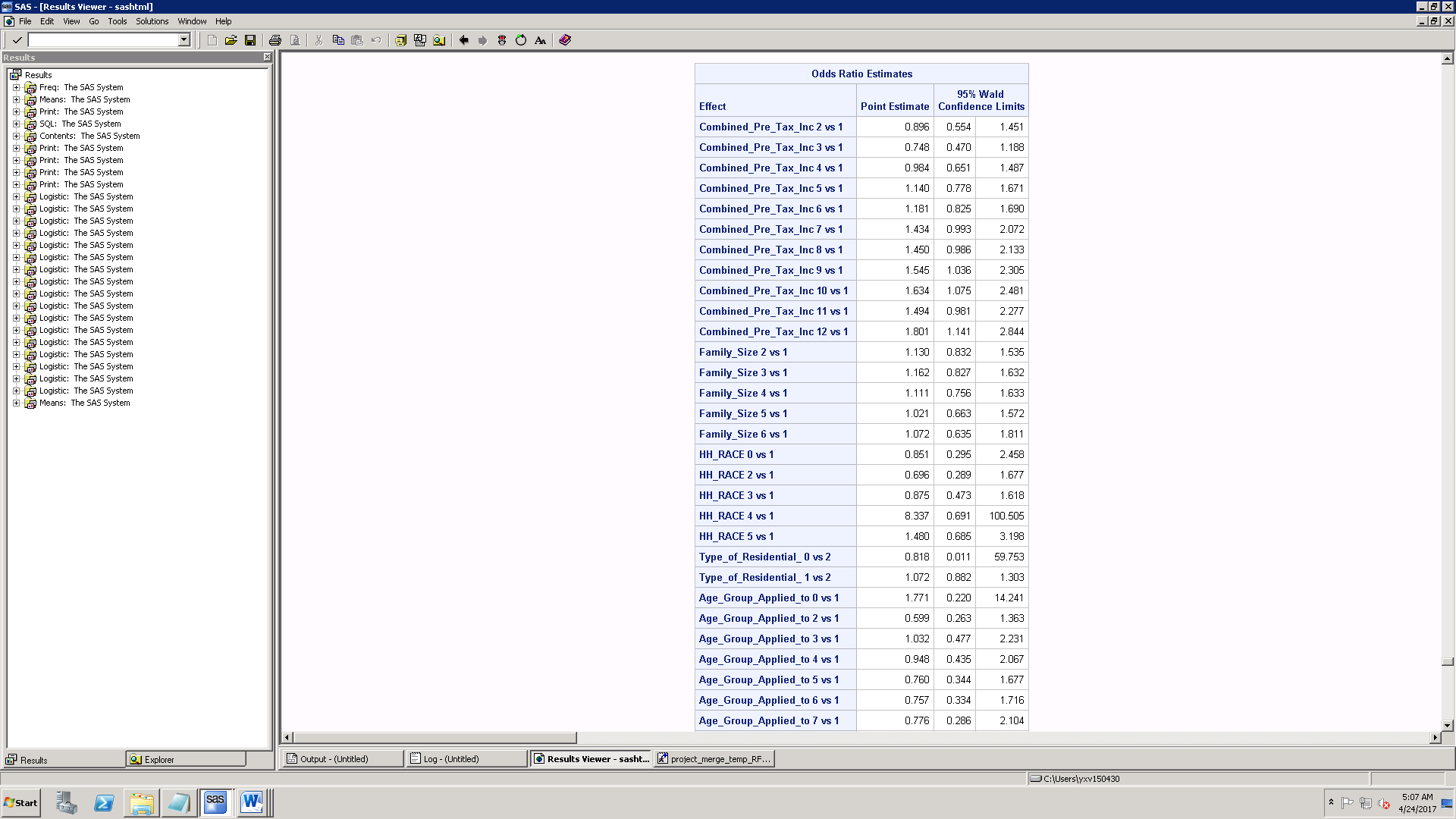
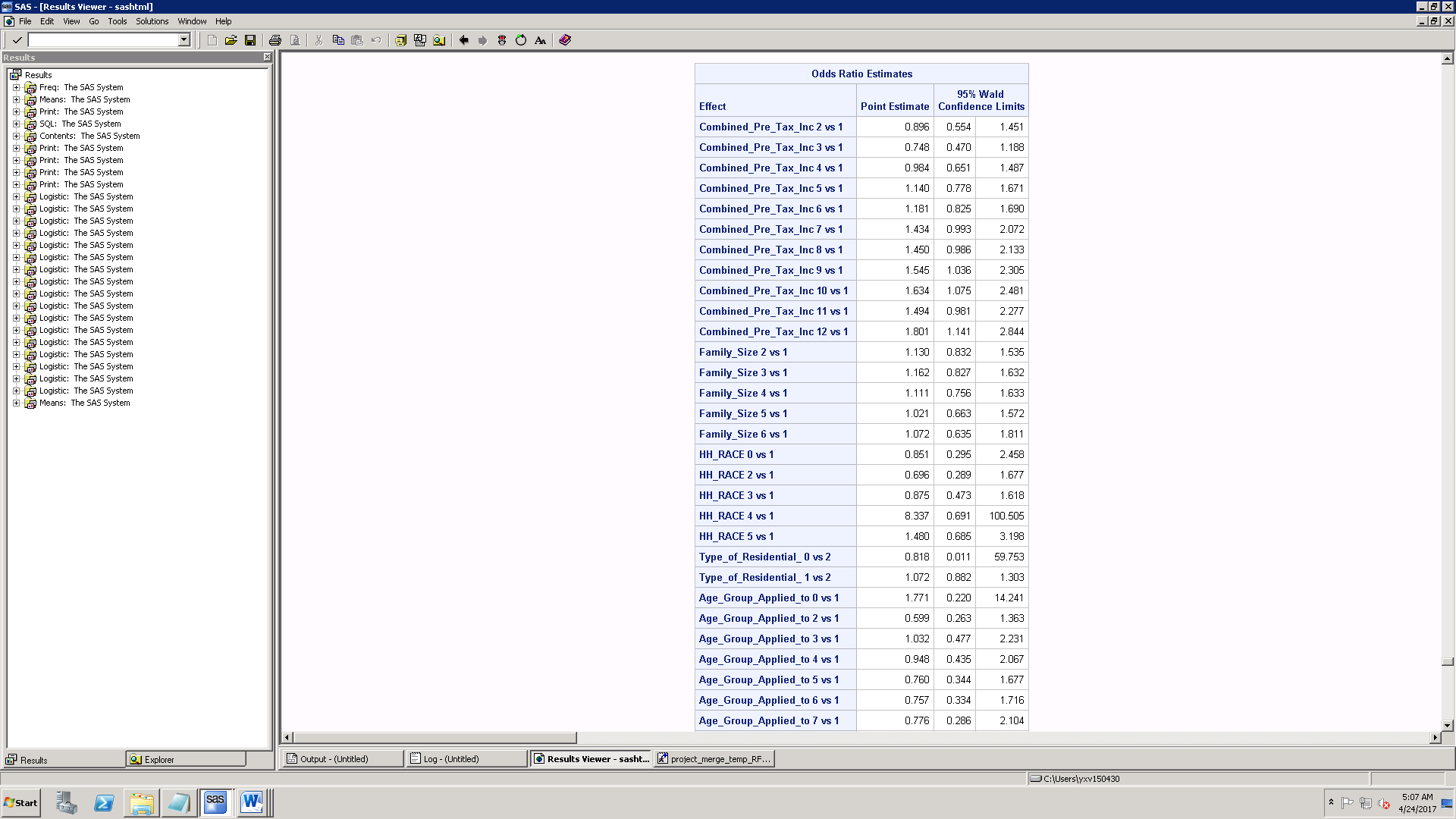
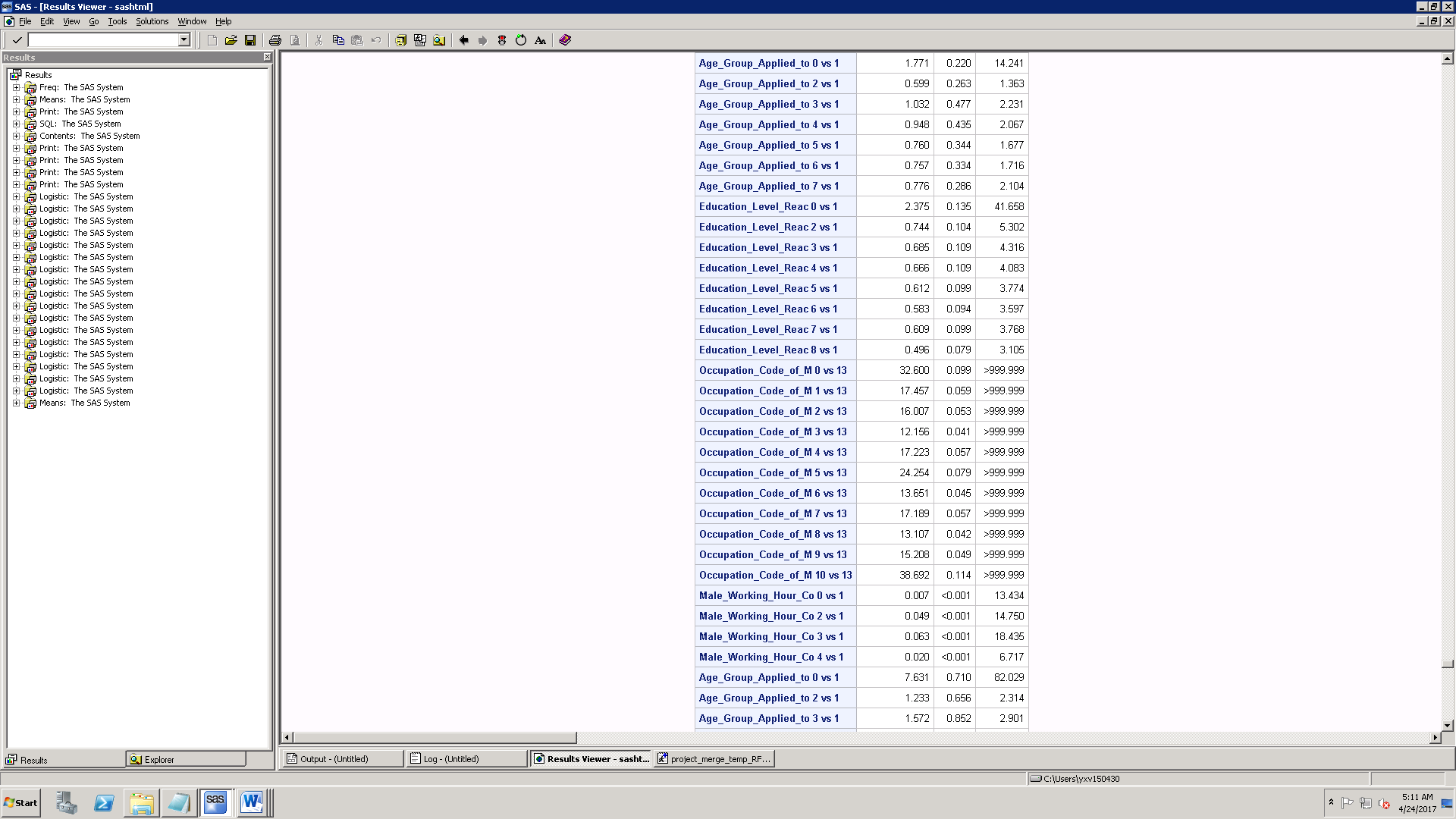
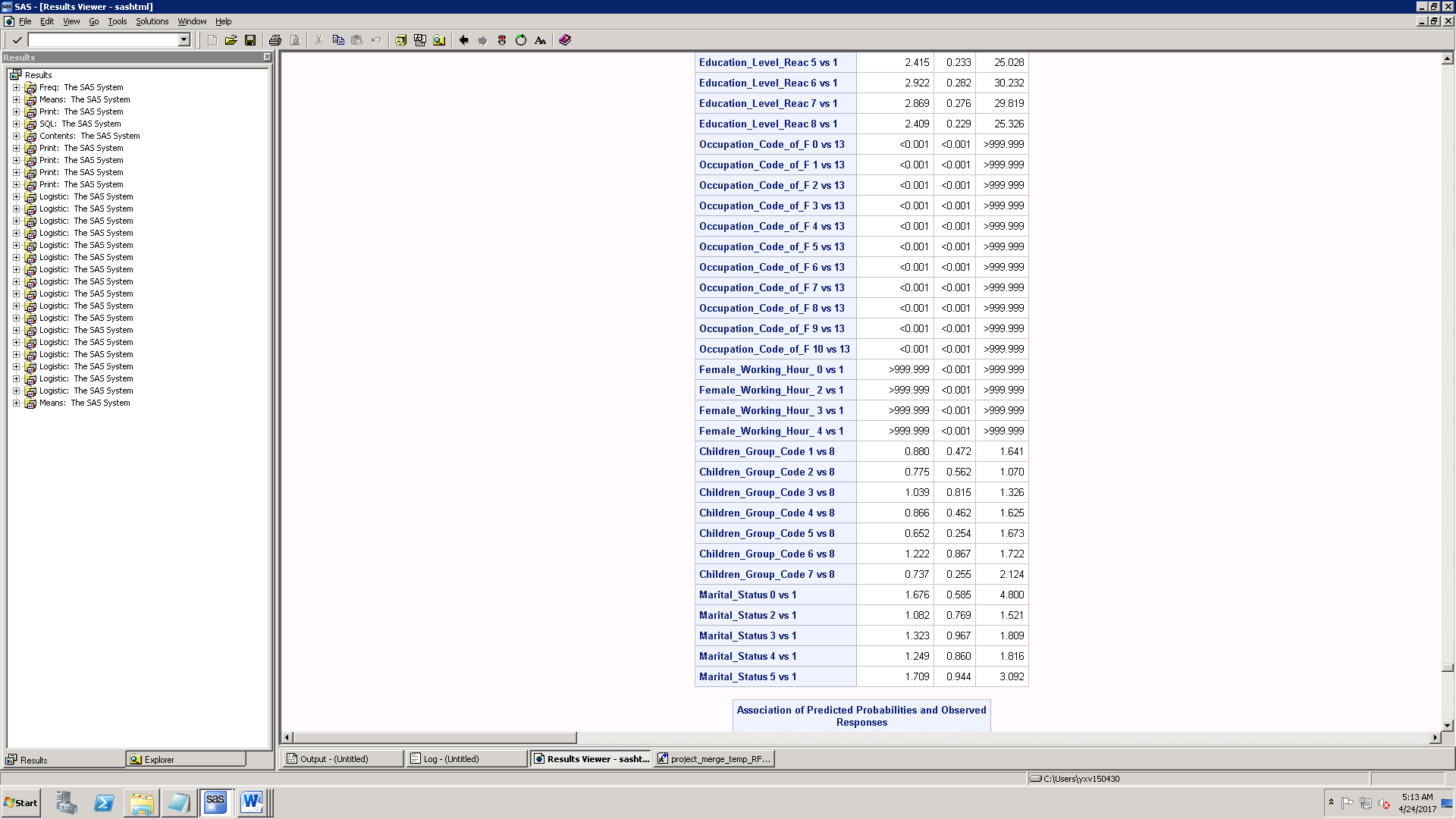


From -2logl value, we can interpret that model behaves well with covariates.

Overall fitness of the model can be identified by checking p value of testing global null hypothesis beta=0. Since chi square p is <0.05, overall model is valid and significant.



****

****

**Interpretation of Odds Ratio:**

From the below Odds ratio, as we explained in the previous Tableau analysis, we can see that income category 7,8,9,10,11 & 12 has higher odds ratio.

When Income category changes from 1 to 9, percentage change in odds of customers becoming high rank will increase by 54%.

Detailed interpretation:

Odds(Rank3)|Odds(Rank2) = 1.54

Odds(Rank2)|Odds(Rank1) = 1.54

Odds(Rank1)|Odds(Rank0) = 1.54

Similarly when Family size increase from 1 to 2,3,4, percentage change in odds of customer being highly ranked will increase by 13%,16.2%, 11.1% respectively.

Compared to the base reference group of Male Age1, when Age category becomes 3, percentage change in odds of customer being highly ranked will increase by 3.2%

Compared to the base reference group of child group code 8, when child group code changes to 3 and 6, percentage change in odds of customer being highly ranked will increase by 3.9% and 22.2%

Accuracy of Naïve classification: Classifying all records as high ranked customers 691/2982=23.17%

Percent concordant is 59.1

So this model is better compared to naïve model in predicting the customer rank.

**Understanding demographics based on the preference of packing type:**

We summed how many Party type, Regular, Small and Box packing type snacks were bought by each customer.

Using PROC Rank, we ranked customers whether he will buy Party type, Regular, Small and Box with group = 2 (he will buy, not buy).

Then we ran logistic regression 4 times (each for Party\_rank, Regular\_Rank, Small\_rank and Box\_rank) with rank variable as output and demographics as input.

Using odds ratio, we took the significant category for each categorical variable(demographics).

Then we compared the demographics of these 4 categories to understand what kind of customers will buy Party, Regular, Small and Box packing type.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Demographics** | **Party** | **Small** | **Regular** | **Box** |
| Income | 15k-19k$ | 10k-12k | 100k$ | 45k$-55k$ |
| Family size | 6 | 3 | 5 | 3 |
| resident | Owner | Renter | Owner | Renter |
| Male Age Group | 18-24 yrs old | 25-34 yrs old | 55-64 yrs old | 18-24 yrs old |
| Edu Level of Male | Grade School &  College - Not graduated | Grade School &  College - Graduated | Grade School &  high School - Not graduated | Completed High School and  not graduated from Grade School |
| Occ code Male | Not Employed,sales, retired | Private household worker | Not Employed,Retired | Cleaning, food,  health service, sales |
| Male work hr | Student,part-time | Not employed, full time | Student, part time | Not employed , full time |
| Female Age Group | 18-24 yrs old |  | 25-43 yrs old | 18-24 yrs old, 65+ |
| Edu Level of Female | College |  | Technical School | Grade school not completed |
| Occ code Female | Private house hold worker |  | Not employed, retired | Managers, clerical |
| Female work hr | Full time | Not employed, | Student | PART TIME |
| Number\_of\_Dogs | 2,3 | 2,1 | 5,1 | 4,0 |
| Number\_of\_Cats | 0,2 | 3,1 | 4,3 | 4,1 |
| Children\_Group\_Code | No children | child 6-11 yrs | No children, child 0-5 yrs | child 0-5 yrs old,  6-11 yrs old, 12-17 yrs old |
| Marital\_Status | Divorced, Single | Divorced, Single | Married | Married |

From the above result, one can identify the demographic properties of each packing type and market accordingly.

**Recommendations**

The following recommendations are made based on the above analysis.

* By rightly identifying the target customers, Pringles can increase its market share and customer base.
* Brand Pringles has been preferred more by the High income, health conscious group.

To increase the sales of Pringles, we recommend them to concentrate on displays and promotions which could target the customer groups like high income low family size with teenage kids.

* Package preference based on demographics, we found teenage students group prefer portable box pack which has been a unique feature in the package of Pringles that creates a unique brand identity, which attracts the youth, so Pringles can introduce youth specific flavors to increase their market share.