[Date]

Khatwani, Vishal

[company name]

[Document title]

[Document subtitle]

Contents

**Exploring the dataset ...................................................................................................................... 3**

**Part I. Modelling count data ............................................................................................................ 5**

**1. Creation of Count Dataset ........................................................................................................... 5**

**2. NBD Model .................................................................................................................................. 6**

**3. Reach, Average Frequency and GRPs ........................................................................................... 7**

**4. Poisson Regression Model ........................................................................................................... 8**

**5. LL Formula - NBD Regression Model ........................................................................................... 10**

**6. NBD Regression Model using customer characteristics................................................................ 10**

**7. Managerial takeaways between Poisson Regression and NBD Regression……………...................... 11**

**8. Comparing the models - LR Test .................................................................................................. 12**

**Part II. Improving the Model ........................................................................................................... 13**

**9. Variable Selection ....................................................................................................................... 13**

**10. Constructing New Variables ...................................................................................................... 17**

**11. Interaction Effects ..................................................................................................................... 30**

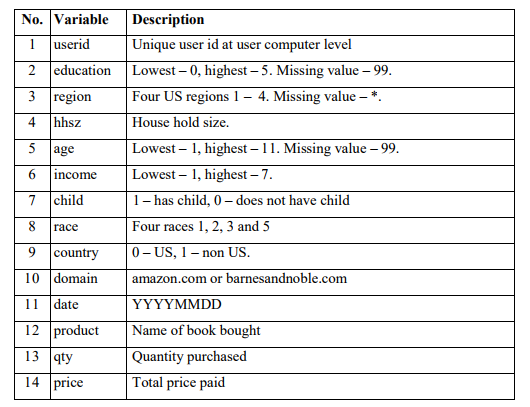
**Part III. Why Certain Customers Prefer Amazon Over BN? ............................................................... 32**

**12. Logistic Regression .................................................................................................................... 32**

**Part IV. Summary ............................................................................................................................ 33**

**13. Insights and Learnings ............................................................................................................... 33**

Data Description

For this project, we focus on analyzing the factors that affect customer purchasing behavior at Barnes and Noble. We are provided with the dataset that records customer purchases at Amazon and Barnes and Noble, as shown in Figure 1. In order to protect consumer privacy, many data fields are transformed, but the order of numbers is preserved. For example, age is shown from 1 to 11 and 8 for age implies a larger true age than 7 for age. With this dataset, we concentrate on building a BA model to make predictions and to figure out what kind of consumer characteristics may cause customers to consume more or less and why certain customers prefer Amazon over Barnes and Noble.

***Figure 1. Dataset “ABA\_PROJECT2\_DATA\_BOOKS”***

The dataset could be classified into three categories – ID variable, demographic variables, and purchase variables. ID variable is userid; Demographic variables include education, region, hhsz, age, income, child, race, and country; Purchase variables include domain, date, product, qty and price.

As we could know from Figure 1, for education and age, missing value are represented by 99, while “\*” represents missing value in region. Therefore, we first replace 99 and “\*” with “.” in SAS.

libname project2 "E:\Fall 17\adv sas\project 2";

**DATA** abap2;

set project2.aba\_project2\_data\_books;

IF education="99" THEN education=".";

IF region="\*" THEN region=".";

IF age="99" THEN age=".";

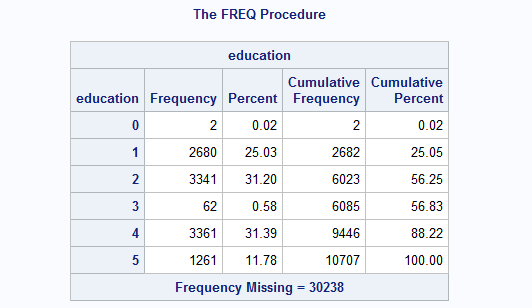
**run**;

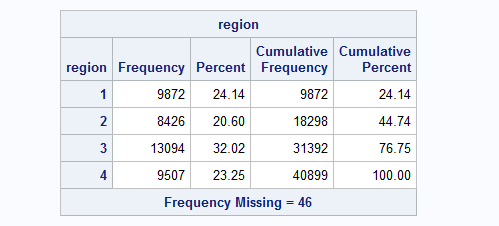
To check missing values of each variable, we use following code to analyze the frequency of variable value.

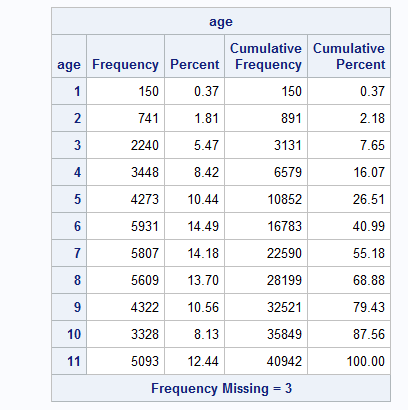
**Proc** **freq** data = abap2;

Tables education region hhsz age income child race country domain;

**RUN**;

From the result, we found only education, region and age have missing values. For education, 30238 records are missing, which takes approximately 74% of the total records. Region and age have relatively less missing value, and the numbers of missing are 46 and 3 respectively.



****

Part 1. Modeling Count Data

1. For each customer, we count the number of books she purchased from BN in 2007, and keep the demographic variables. SAS code and first 10 records of this dataset are shown below:

/\*calculate total books each customer purchased from each book seller\*/

**PROC** **SQL**;

CREATE TABLE project2.table1 as

(SELECT Unique(USERID),EDUCATION,REGION,HHSZ,AGE,INCOME,CHILD,RACE,COUNTRY,DOMAIN,SUM(QTY)AS COUNT

FROM abap2

GROUP BY USERID,DOMAIN);

**QUIT**;

/\*create new variables AMAZON ang BarnesNoble to keep the count\*/

**DATA** project2.table2;

SET project2.table1;

IF domain='amazon.com' THEN DO AMAZON=COUNT;BarnesNoble=**0**;END;

ELSE IF domain='barnesandnoble.com' THEN DO AMAZON=**0**; BarnesNoble=COUNT;END;

DROP COUNT;

**RUN**;

/\*create a new table \*/

**PROC** **SQL**;

CREATE TABLE project2.BarnesNoble AS

(SELECT UNIQUE(USERID),EDUCATION,REGION,HHSZ,AGE,INCOME,CHILD,RACE,COUNTRY,SUM(AMAZON)AS AMAZON,SUM(BarnesNoble)AS BarnesNoble

FROM project2.table2

GROUP BY USERID);

**Quit**;

/\*only keep the number of books purchased in NB and drop the number of Amazon\*/

**DATA** project2.final;

SET project2.BarnesNoble;

DROP AMAZON;

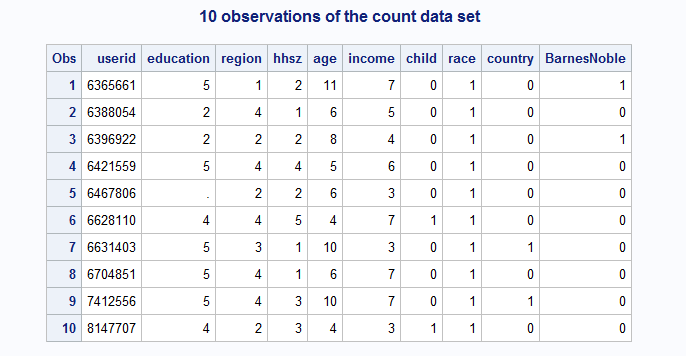
**RUN**;

/\*Print first 10 observations\*/

**proc** **print** data=project2.final(obs=**10**);

TITLE "10 observations of the count data set";

**run**;



1. Run the NBD Model.

/\*create a new table\*/

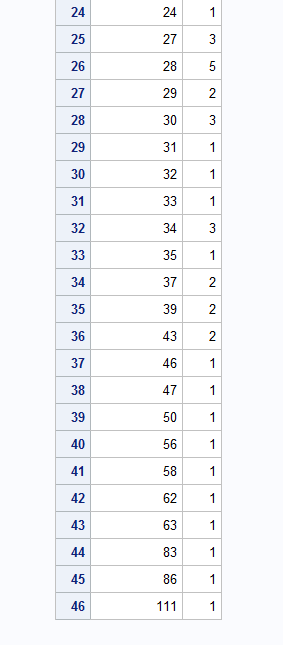
**Proc** **sql**;

Create Table project2.nbd1 as

(select BarnesNoble,count(userid)as freq from project2.final

group by BarnesNoble);

**run**;

*Figure 2. Table project2.nbd1*

/\*run the NBD Model to obtain MlE result\*/

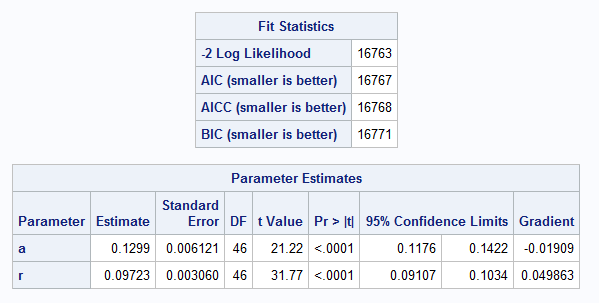
**PROC** **NLMIXED** DATA=project2.nbd1;

PARMS a=**0.5**,r=**0.5**;

ll=freq\*log(gamma(r+BarnesNoble)/(gamma(r)\*fact(BarnesNoble))\*(a/(a+**1**))\*\*r\*(**1**/(a+**1**))\*\*BarnesNoble);

Model freq~general(ll);

**run**;



1. Based on the NBD Model result, we could compute the Reach, Average Frequency and GRPs are 0.1897, 3.946, and 74.85 respectively. Calculations are shown below.

**Reach** = 1 - P(x=0)

= 1 - (0.1299/(0.1299+1))^0.09723

= 1 - 0.8103

= 0.1897

E(x) = r/a= 0.09723/0.1299 = 0.7485

**Average Frequency** = E(x)/(1-p(x=0))

= 0.7485/ 0.1897

= 3.946

**GRPs** = 100 \*Reach\* Average Frequency

= 100 \* E(x)

= 74.85

1. Then we run the Poisson Regression Model considering demographic information.

/\*run a poisson distribution\*/

**Proc** **NLMIXED** DATA=project2.final;

parms m0=**1** b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** b8=**0**;

m=m0\*exp(b1\*EDUCATION+b2\*REGION+b3\*HHSZ+b4\*AGE+b5\*INCOME+b6\*CHILD+b7\*RACE+b8\*COUNTRY);

ll=BarnesNoble\*log(m)-m-log(fact(BarnesNoble));

MODEL BarnesNoble~general(ll);

**run**;

The Maximum LL is -5,200.0991. We can find only RACE(b7) has a P value that is more than 5%. Hence, we can say that EDUCATION, REGION, HHSZ, AGE, INCOME, CHILD and COUNTRY are significant parameters since their P values are less than 5%.

1. LL formula for NBD Regression Model

To capture the differences among individuals, we have:

**P (X =BarnesNoble)** = gamma [(r+ BarnesNoble)/(gamma(r) (BarnesNoble)!)] \*(a/(a+m))^r \*(m/(a+m))^BarnesNoble)

Where

**m** =exp(b1\*education+b2\*region+b3\*hhsz+b4\*age+b5\*income+b6\*child+b7\*race+b8\*country)

Therefore, the LL formula for NBD Regression Model is:

**LL = log (p(x))**

1. Run the NBD Regression Model

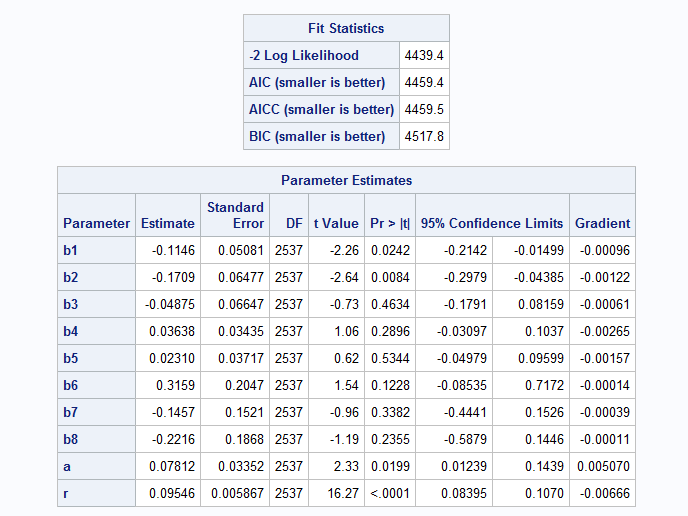
**Proc** **NLMIXED** DATA=project2.final;

parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** a=**0.1** r=**0.1**;

m=exp(b1\*REGION+b2\*HHSZ+b3\*AGE+b4\*INCOME+b5\*CHILD+b6\*RACE+b7\*COUNTRY);

ll=log(gamma(r+BarnesNoble)/(gamma(r)\*fact(BarnesNoble))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*BarnesNoble);

MODEL BarnesNoble~general(ll);

**run**;

The maximum LL is -2219.7086. The results indicate that only EDUCATION (b1) and REGION (b2) are significant. The P values of HHSZ (b3), AGE (b4), INCOME (b5), CHILD (b6), RACE (b7) and COUNTRY (b8) are not significant.

1. Noticeable Difference Regarding the Managerial Takeaways Between Poisson Regression and NBD Regression

In comparing the results of Poisson Regression and NBD Regression, we notice that there’s a significant difference between the two results. First, the log likelihood value of Poisson Regression is -5,200.0991 whereas the log likelihood value of NVD Regression is -2,219. 7086. Second, we also observe that Poisson Regression has more significant parameters than NBD Regression. There are seven variables that are significant in Poisson Regression Model, including Education (b1), Region (b2), Hhsz (b3), age (b4), income (b5), child (b6), and country (b8) while only Education (b1) and Region (b2) are significant in NBD Regression Model.

The reason for these discrepancies probably is that in Poisson Regression Model the mean and variance are equal while NBD Regression allows the variance to differ from the mean. In Poisson Regression Model, we only considered how people differ on a set of available explanatory variables, but in NBD Regression Model, we let λ vary across the population to capture the unobserved component of differences among individuals.

1. Then, we use the Likelihood Ratio Test to determine whether NBD Regression Model fits the data better than Poisson Regression Model.

**Model A**: NBD Regression Model; **Model B**: Poisson Regression Model

We know that LLA is is -2219.7086 and LLB is -5,200.0991

**H0**: Model A is not different from Model B

**H1**: Model A is different from Model B

LR = -2(LLB – LLA)

= -2(-5.200.0991 – (-2,219.7086))

= 5960.781

χ2 (.05,1) = 3.84146

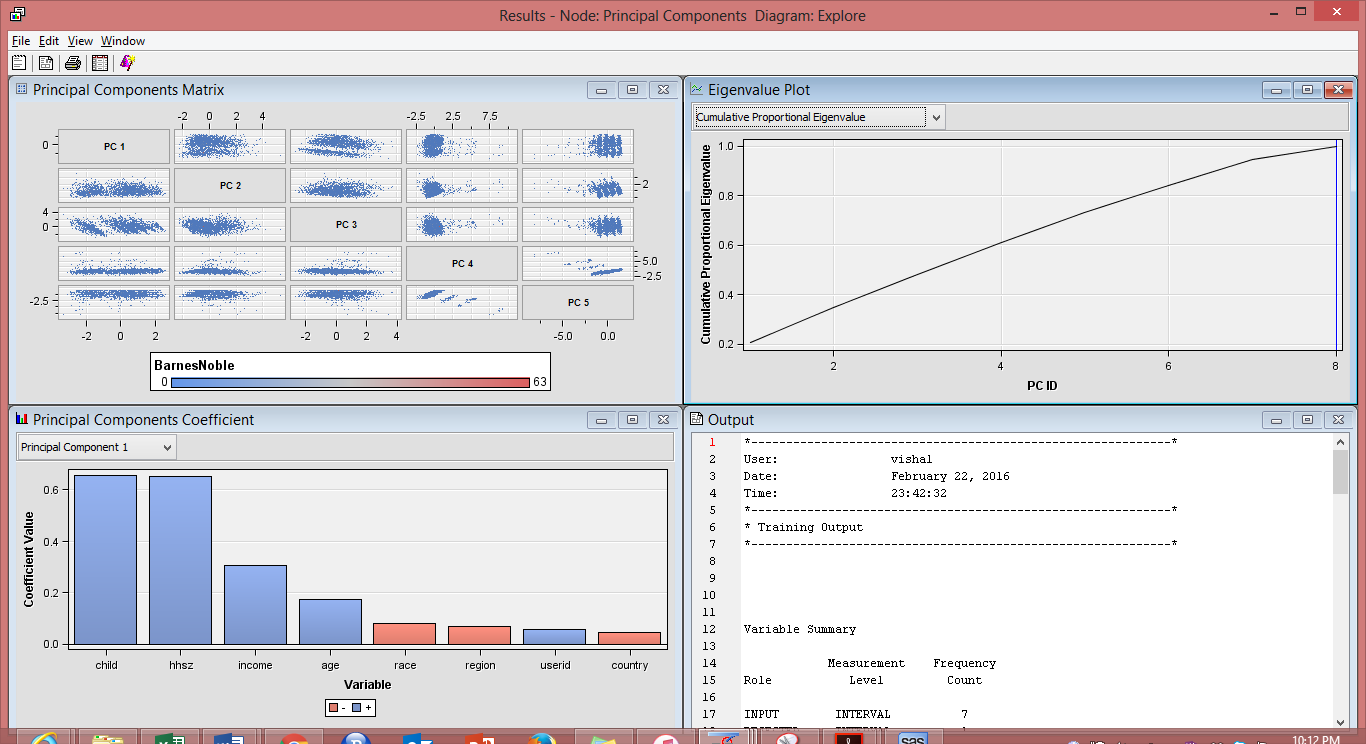
Since LR > χ2 (.05,1), we reject H0. Therefore, we could conclude that NBD Regression Model is better than Poisson Regression Model.

Part II. Improving the Model

9. Variable selection

To start with variable selection we first used the explored the variables in SAS EM and we found out that the variable “Education” has 76% missing values marked as ‘99’ which is 3 times the non-missing values. We did not impute the missing values because that would skew as well as make the data bias hence we rejected the variable.

To further identify the important variables, we skewed ran “Principle component analysis” on SAS EM and below is the result:



Based on the above result we can say that the first 6 variables explain more than 90% about the number of books that a customer’s buys from Barnes and Noble.

Also, the variable “country” is not having much importance with respect to the number of items sold we can explain this by fact the country is the derived variable for regions and thus we decided to remove the country variable and instead using the variable “Region” and since region represent only the USA regions we replaced the missing value with ‘5’ depicting the non-us regions.

**DATA** project2.final;

set project2.final;

IF region="\*" THEN region="5";

**run**;

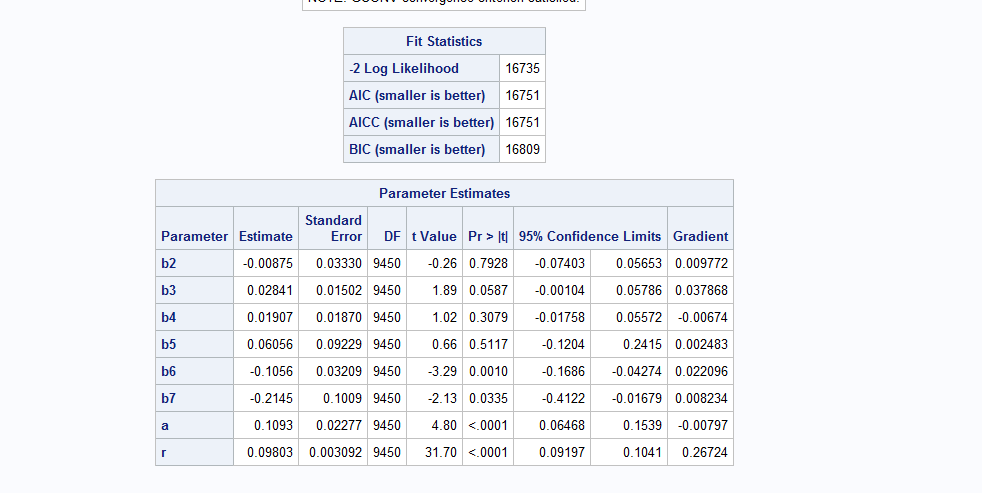
**Proc** **NLMIXED** DATA=project2.final;

parms b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** a=**0.1** r=**0.1**;

m=exp(b2\*HHSZ+b3\*AGE+b4\*INCOME+b5\*CHILD+b6\*REGION+b7\*RACE);

ll=log(gamma(r+BarnesNoble)/(gamma(r)\*fact(BarnesNoble))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*BarnesNoble);

MODEL BarnesNoble~general(ll);

**run**;

We ran the new NBD regression model without the variables “Education” and “Country”, based on the result we can say that the variables Race, Region and Age are significant in explaining the variation in BN sales across all the customers.

But on the other hand, we don’t see much changes in the negative log likelihood value thus we move forward to create some new variables and dig deep into the analysis.

10. Creating new variables

10.1 Modifying the date variable into weekdays and weekends

We extracted the SAS dataset into MS Excel and created two indicators for weekdays and weekends and thus dividing the purchases on weekdays and weekends. Then we used these indicators to draw in further insights from the dataset.

Creation Weekend and Weekday indicators

**PROC** **SQL**;

CREATE TABLE project2.try1 as

(SELECT Unique(USERID),REGION,HHSZ,AGE,INCOME,CHILD,RACE,COUNTRY,WEEKEND,WEEKDAY,DOMAIN,SUM(QTY)AS COUNT

FROM abap24

GROUP BY USERID,DOMAIN);

**QUIT**;

**DATA** project2.try2;

SET project2.try1;

IF domain='amazon.com' THEN DO AMAZON=COUNT;BarnesNoble=**0**;END;

ELSE IF domain='barnesandnoble.com' THEN DO AMAZON=**0**; BarnesNoble=COUNT;END;

if weekend = **1** then do weekendqty=COUNT;end;

if WEEKDAY = **1** then do weekdqty=COUNT; end;

DROP COUNT;

**RUN**;

**PROC** **SQL**;

CREATE TABLE project2.BarnesNobleupd AS

(SELECT UNIQUE(USERID),REGION,HHSZ,AGE,INCOME,CHILD,RACE,COUNTRY,SUM(weekendqty)AS weekend\_qty,SUM(weekdqty)AS weekday\_qty,SUM(BarnesNoble)AS BarnesNoble

FROM project2.try2

GROUP BY USERID);

**Quit**;

**data** project2.BarnesNobleupd1;

set project2.BarnesNobleupd;

if weekend\_qty="." then do weekend\_qty=**0**;end;

if weekday\_qty="." then do weekday\_qty=**0**;end;

**run**;

**DATA** project2.BarnesNobleupd1fn;

SET project2.BarnesNobleupd1;

DROP AMAZON;

**RUN**;

Using the above code, we created a new dataset where we have created two new variables in which we have aggregated the book sales at Barnes and Noble on weekdays and weekends,

We further ran the NBD Regression on this model to analyse the effects of newly created variables.

**Proc** **NLMIXED** DATA=project2.BarnesNobleupd1fn;

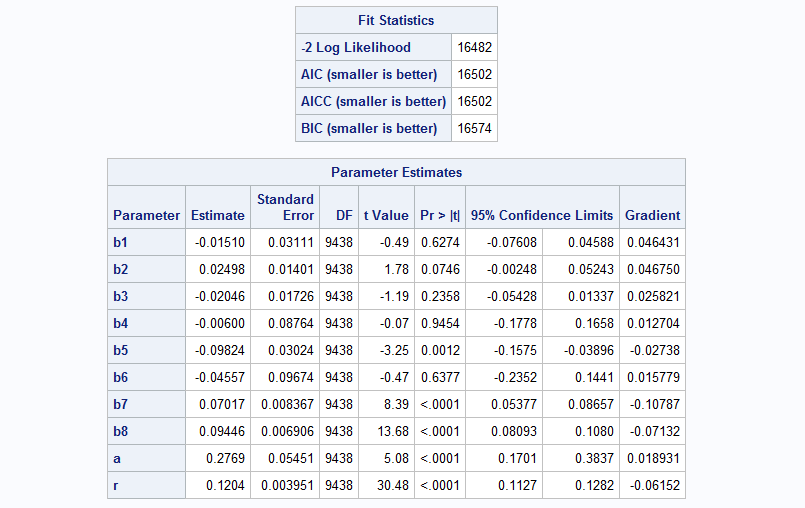
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** b8=**0** a=**1** r=**1**;

m=exp(b1\*HHSZ+b2\*AGE+b3\*INCOME+b4\*CHILD+b5\*REGION+b6\*RACE+b7\*weekend\_qty+b8\*weekday\_qty);

ll=log(gamma(r+BarnesNoble)/(gamma(r)\*fact(BarnesNoble))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*BarnesNoble);

MODEL BarnesNoble~general(ll);

**run**;



Managerial Takeaway

The maximum likelihood further improves with the help of NBD regression model and the significant variables in the above results are count of sales for customers aggregated by Weekdays and weekends, Region, age

Analyzing the Loyalty

Moving forward to analyze the customer loyalty with BN we created two more variables that indicate the customers loyalty towards Amazon and BN. Then we ran the NBD regression model on the new dataset to analyze the customer loyalty towards both the brands we can clearly see from the results that when a customer makes a unit purchase from BN their loyalty increases towards BN by a very high factor of 2.6 whereas in Amazon with every purchase the customer loyalty increases by a factor of 0.81

**Data** project2.loyalty;

Set project2.BarnesNoble;

loyalty\_Barnes=**0**;

if Amazon=**0** AND BarnesNoble > **0** then loyalty\_Barnes=**1**;

**run**;

**Data** project2.loyalty;

Set project2.loyalty;

loyalty\_Amazon=**0**;

if BarnesNoble=**0** AND Amazon > **0** then loyalty\_Amazon=**1**;

**run**;

**proc** **NLMIXED** DATA=project2.loyalty;

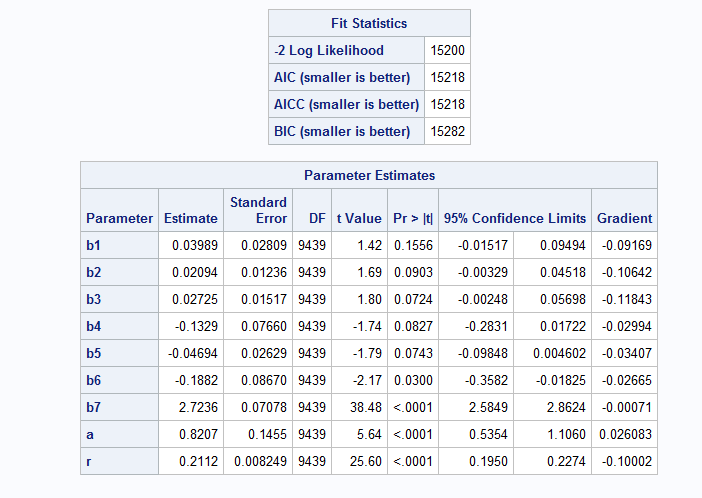
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7= **0** a=**1** r=**1**;

m=exp(b1\*HHSZ+b2\*AGE+b3\*INCOME+b4\*CHILD+b5\*REGION+b6\*RACE+b7\*loyalty\_Barnes);

ll=log(gamma(r+BarnesNoble)/(gamma(r)\*fact(BarnesNoble))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*BarnesNoble);

MODEL BarnesNoble~general(ll);

**run**;



**proc** **NLMIXED** DATA=project2.loyalty;

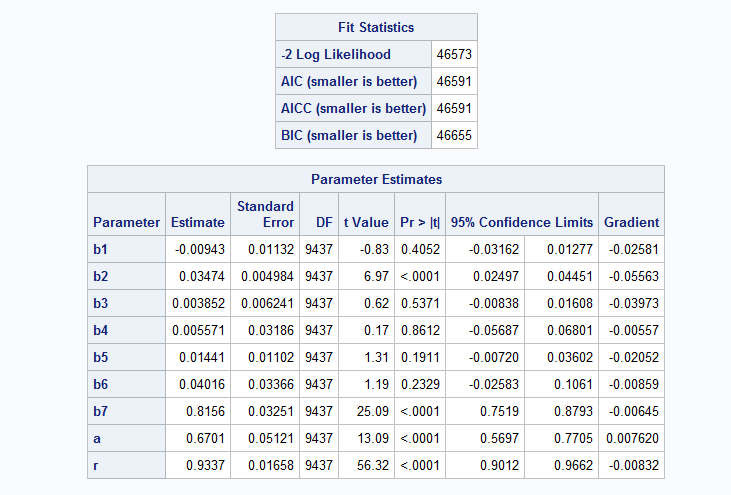
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7= **0** a=**0.1** r=**0.1**;

m=exp(b1\*HHSZ+b2\*AGE+b3\*INCOME+b4\*CHILD+b5\*REGION+b6\*RACE+b7\*loyalty\_Amazon);

IF AMAZON > **160** THEN ll=LOG(exp(-**300**));

else ll=log(gamma(r+Amazon)/(gamma(r)\*fact(Amazon))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*Amazon);

MODEL Amazon~general(ll);

**run**;

Moving ahead we created two more variable for to further analyze the BN customers the variables were the total sales to a customers and average sales to the customer and we ran NBD regression for the same but we didn’t find and significant results that describe the number of books purchased by a customer from BN

**PROC** **SQL**;

CREATE TABLE project2.bn as

(SELECT Unique(USERID),REGION,HHSZ,AGE,INCOME,CHILD,RACE,COUNTRY,WEEKEND,WEEKEND\_1,DOMAIN,SUM(QTY)AS COUNT\_BN, SUM(price) AS BN\_Sales,(SUM(price)/SUM(QTY)) AS avg\_bn

FROM abap24

where domain = 'barnesandnoble.com'

GROUP BY USERID);

**QUIT**;

**Proc** **NLMIXED** DATA=project2.bn;

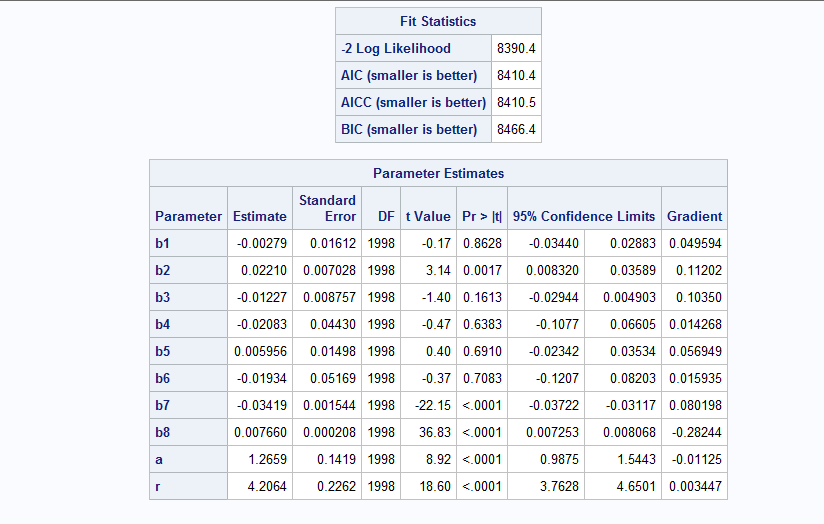
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** b8=**0** a=**1** r=**1**;

m=exp(b1\*HHSZ+b2\*AGE+b3\*INCOME+b4\*CHILD+b5\*REGION+b6\*RACE+b7\*avg\_bn+b8\*BN\_Sales);

ll=log(gamma(r+COUNT\_BN)/(gamma(r)\*fact(COUNT\_BN))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*COUNT\_BN);

MODEL COUNT\_BN~general(ll);

**run**;



11. Interaction Effects

1. Age and Income

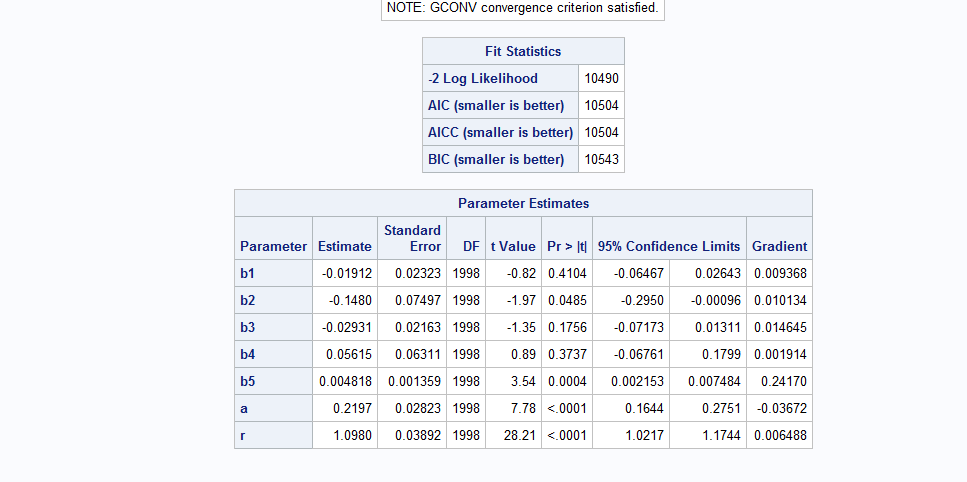
**Proc** **NLMIXED** DATA=project2.bn;

parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** a=**0.1** r=**0.1**;

m=exp(b1\*HHSZ+b2\*RACE+b3\*REGION+b4\*CHILD+b5\*AgeIncome);

ll=log(gamma(r+COUNT\_BN)/(gamma(r)\*fact(COUNT\_BN))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*COUNT\_BN);

MODEL COUNT\_BN~general(ll);

**run**;

From the above results, we can say that the interaction of AGE and INCOME gives us significant value because the p value is < 0.05

1. Age and Child

**Proc** **NLMIXED** DATA=project2.bn;

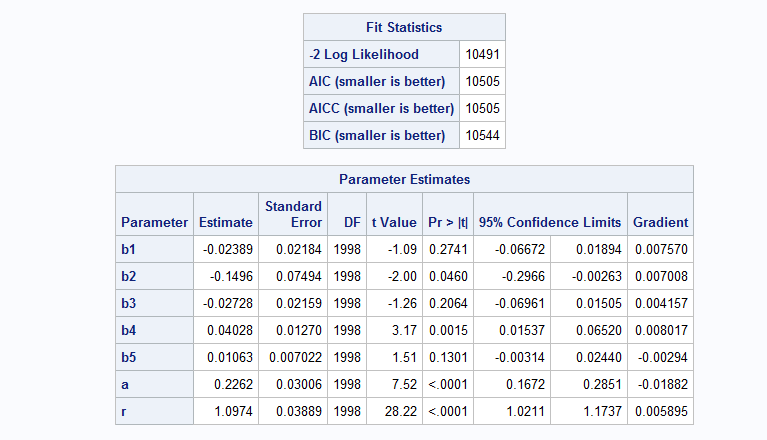
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** a=**0.1** r=**0.1**;

m=exp(b1\*HHSZ+b2\*RACE+b3\*REGION+b4\*Income+b5\*AgeChild);

ll=log(gamma(r+COUNT\_BN)/(gamma(r)\*fact(COUNT\_BN))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*COUNT\_BN);

MODEL COUNT\_BN~general(ll);

**run**;



From the above results, we can say that the interaction of AGE and CHILD does not gives us significant value because the p value is > 0.05

1. Child and Region

**Proc** **NLMIXED** DATA=project2.bn;

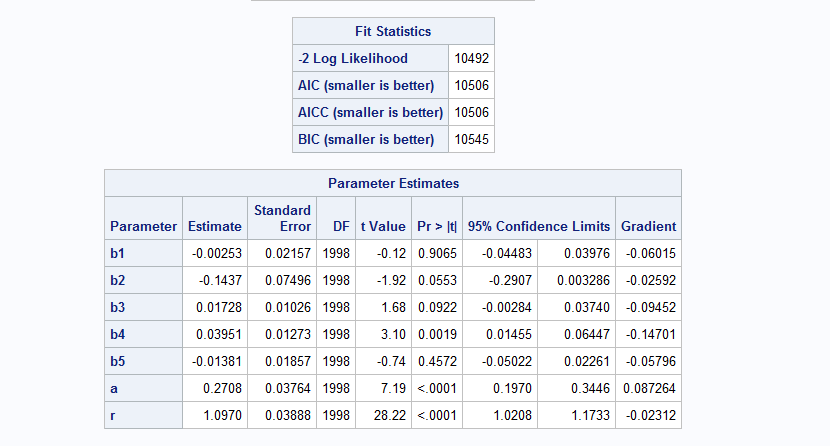
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** a=**0.1** r=**0.1**;

m=exp(b1\*HHSZ+b2\*RACE+b3\*AGE+b4\*Income+b5\*ChildRegion);

ll=log(gamma(r+COUNT\_BN)/(gamma(r)\*fact(COUNT\_BN))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*COUNT\_BN);

MODEL COUNT\_BN~general(ll);

**run**;



From the above results, we can say that the interaction of REGION and CHILD does not gives us significant value because the p value is > 0.05

Region and Income

**Proc** **NLMIXED** DATA=project2.bn;

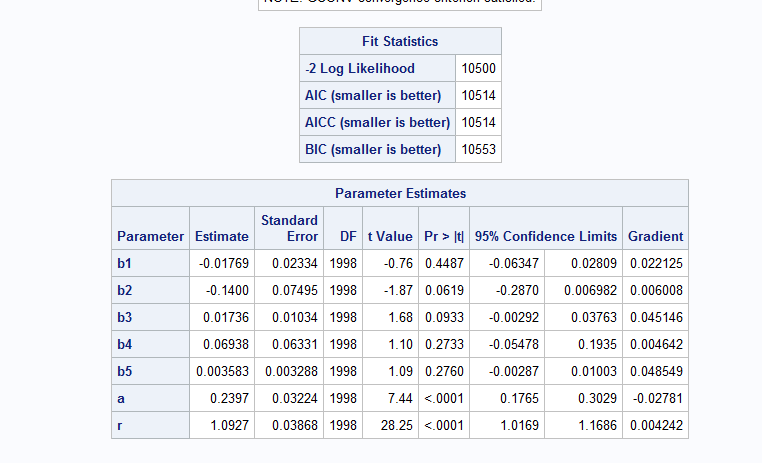
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** a=**0.1** r=**0.1**;

m=exp(b1\*HHSZ+b2\*RACE+b3\*AGE+b4\*Child+b5\*RegionIncome);

ll=log(gamma(r+COUNT\_BN)/(gamma(r)\*fact(COUNT\_BN))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*COUNT\_BN);

MODEL COUNT\_BN~general(ll);

**run**;



From the above results, we can say that the interaction of REGION and INCOME does not gives us significant value because the p value is > 0.05

Part III. Why Certain Customers Prefer Amazon Over BN?

12. Using Logistic regression to compare between BN and Amazon

**Data** project2.logistic;

set project2.BarnesNoble;

if BarnesNoble > **0** then flag\_bn = **1**;

else flag\_bn = **0**;

if AMAZON > **0** then flag\_am = **1**;

else flag\_am = **0**;

**run**;

**proc** **logistic** data = project2.logistic;

model flag\_bn = REGION HHSZ AGE INCOME CHILD RACE COUNTRY/expb;

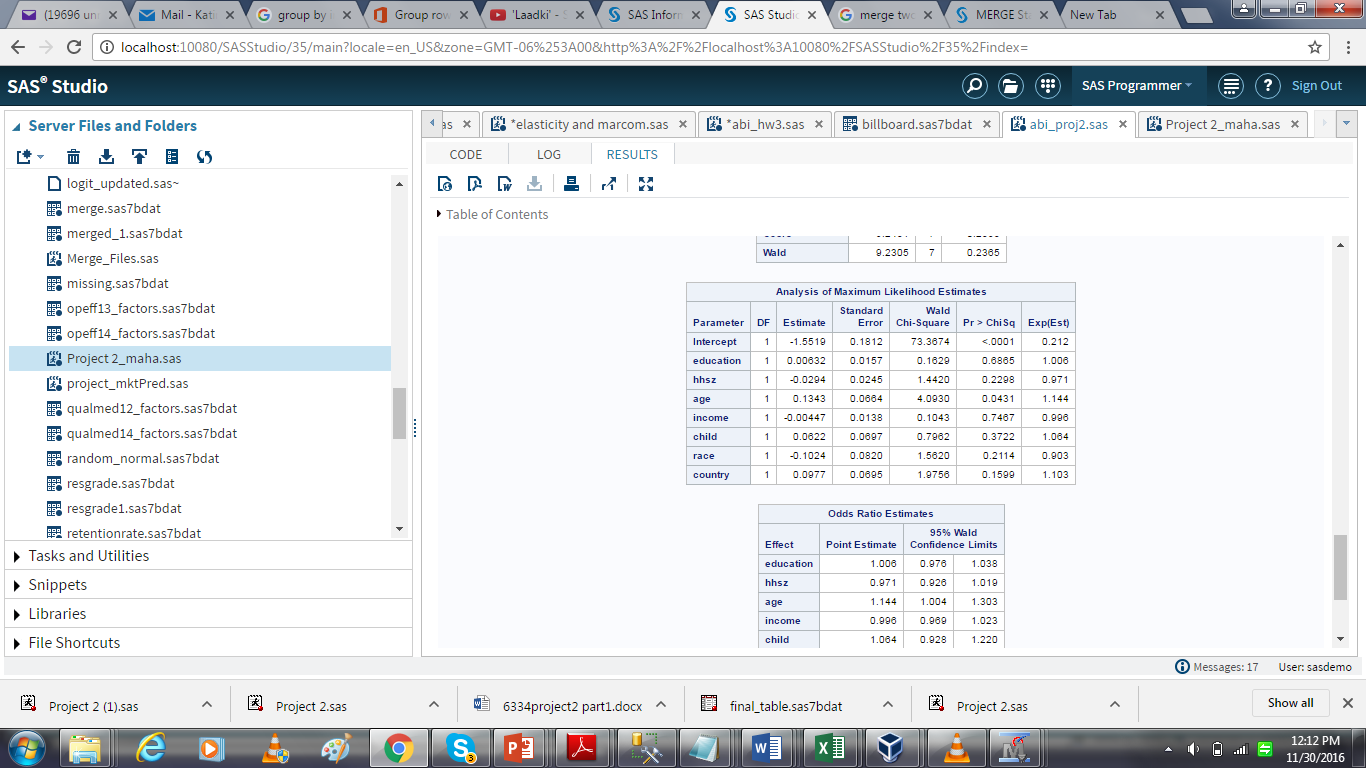
**run**;

**proc** **logistic** data = project2.logistic;

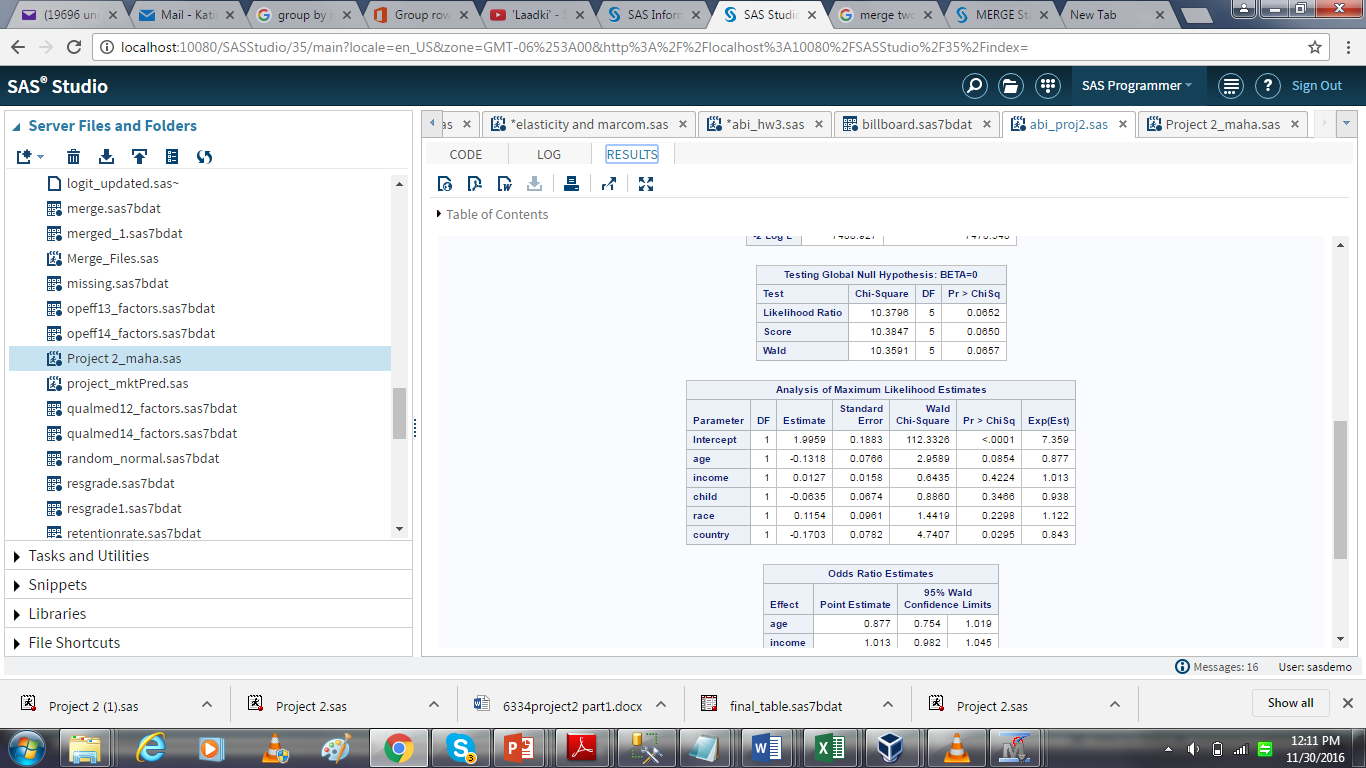
model flag\_am = REGION HHSZ AGE INCOME CHILD RACE COUNTRY/expb;

**run**;

Barnes And Noble



Amazon



Managerial Takeaway:

From the above results, we can easily state that that for Barnes & Nobel “Age” is the only variable that is significant whereas for amazon both the variables “Age” and “Country” are significant

This result is justified because Amazon is a bigger firm than Barnes and Nobel which not only sells book but other articles as well and it I famous across other countries as well whereas a Barnes is an in-store branch which majorly operates in the US.

**Part IV. Summary**

Insights and Learnings

1. Based on our analysis in the above report we can clearly see that reach for Amazon is more than that of Barnes so in this new generation of online retail marketing Barnes must use digital marketing techniques to reach as much people as possible which will help them attract new customers
2. Based on our analysis we saw that Customer loyalty for Barnes is high thus we can say that Barnes is able to retain its customers which make the buy from them again. Thus, they must majorly work on their marketing schemes
3. Age is one of the significant characteristic that Barnes must consider while designing their marketing schemes as age is a significant variable during the sales for Barnes
4. Learned how to build customized analytic model using Base SAS
5. Learned that even though we think that we can explain the dependent variable using all possible independent variables, we are wrong; we always miss on some aspects or variables that can explain the dependent variable
6. Understood the difference between Poisson Regression and NBD Regression Model and how to compare the models using LR test
7. To conclude, working on this project helped us to consider the business insights. It helped everyone individually to attain the Business Intelligence skills which we are sure to help us a lot in the longer run and open our doors to pursue a career in Business Analytics field