Spring 2018

Windows User

[company name]

Mini Project 1

Group 2

Hanson Wu

Medha Tiwary

Animesh Kansal

Jaahnavi Tiruthani

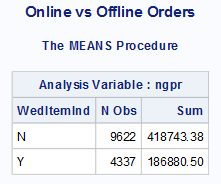
**Changes and Assumptions**

For the Catalogue dataset, we determined that the Gross Product Revenue Amount (GPR) was the most important variable. However, we needed to adjust this variable depending on the date. We created a new variable NGPR or Net Gross Product Revenue amount by subtracting Gift Certificate Amount, Coupon Amount, and Refund Amount from GPR if the Date of order was before 1/25/2007 and subtracting Gift Certificate Amount, Coupon Amount, Shipping and Handling, Sales Tax, Refund Amount, Cancel Amount, and Return Amount if the Date of order was on 1/25/2007 or after.

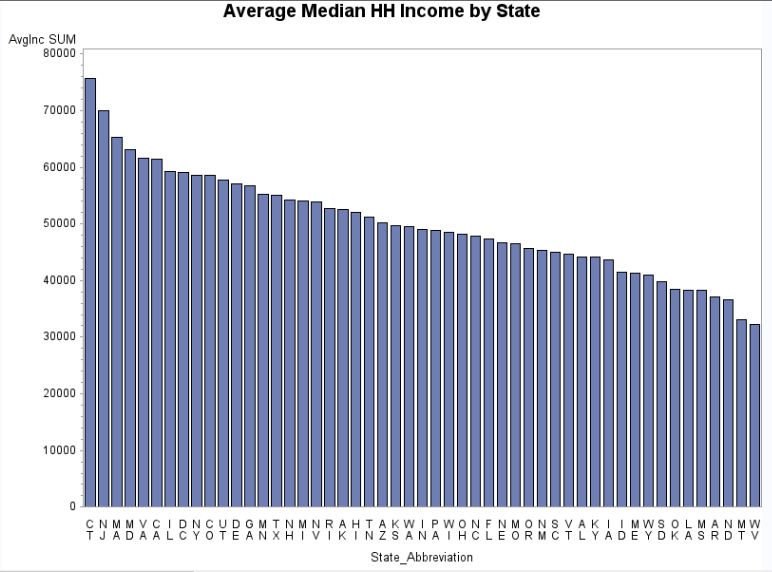
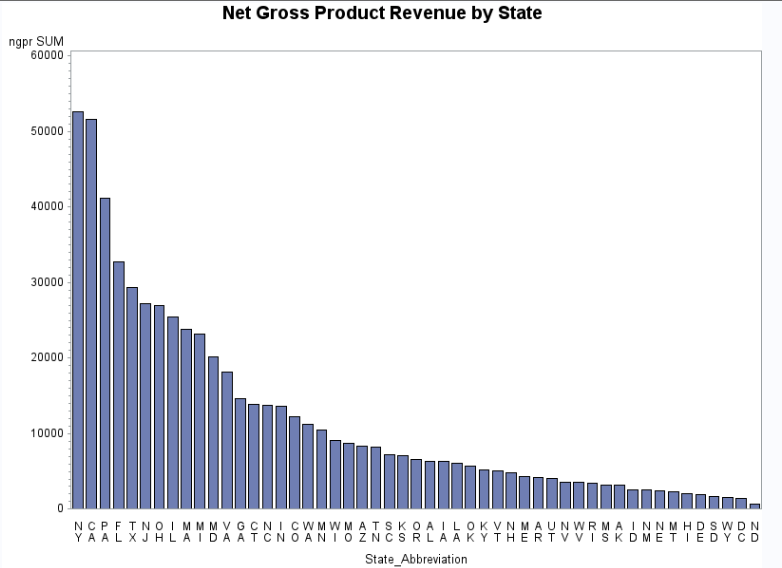
The initial data import was relatively straight forward using the DATA and INPUT commands. We formatted the dates where applicable using FORMAT DATE MMDDYY10..We analyzed the data and dropped quite a few variables. Shipping and Handling Amount, Sales Tax , Cancel Amount, Returned Amount, Refund Amount, Gift Certificate Amount, and Coupon Amount were all now represented in the newly created variable NGPR so they were no longer needed. Refund Status Code and Refund Reason were dropped since Refund Amount is already considered. Any refund is generally considered bad for the company and that is reflected by the refund amount being subtracted from the GPR. Offer Drop Code was dropped because the variable itself was no longer valid after September 2008. Variables related to Gift Certificates and Coupons were dropped because less than 1% of the orders utilized them. Additional Charges Amount and Code were dropped since these variables are already reflected in NGPR. Web Item Indicator was dropped since that data is already reflected in Web Item Quantity. Write Off amount was dropped because all values are $0.

**Marketing Ideas and Exploratory Analysis**

*What are the main methods people are spending money on the catalogue?* - We should focus advertising on more successful methods and see if they are changing over time.

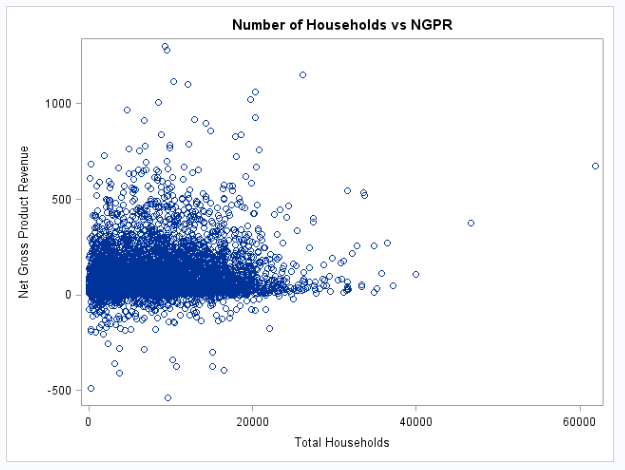


*Which zip codes or states spend the most money? Which make the most money?* – Focus advertising on states that spend more money or less money depending on the situation.

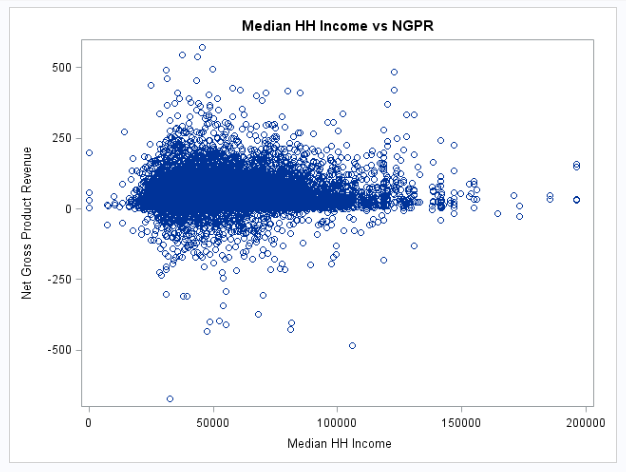


*Are there any locations that have a higher rate of returns, refunds, etc than others?* – spend less resources in these ‘problem’ areas.

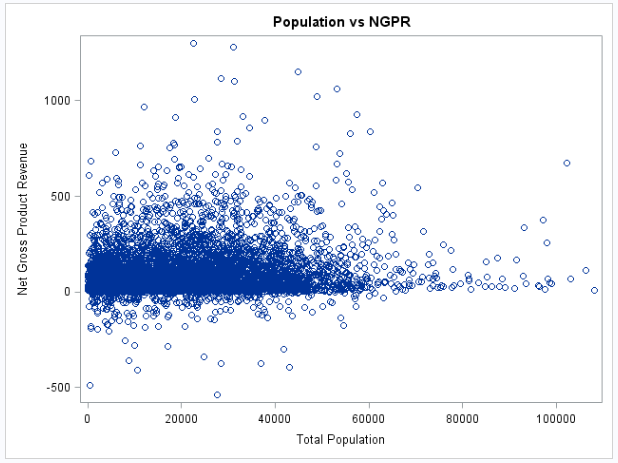
*Is there a correlation between number of households and spending?* – We can target areas with many households but low amount of total spending. More households in an area means our advertisements will reach more people.



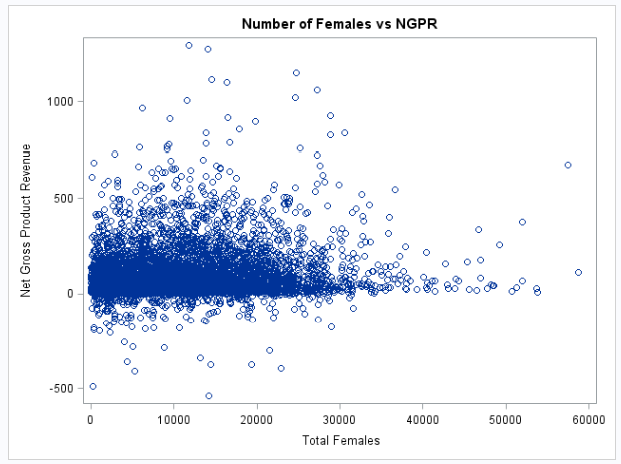
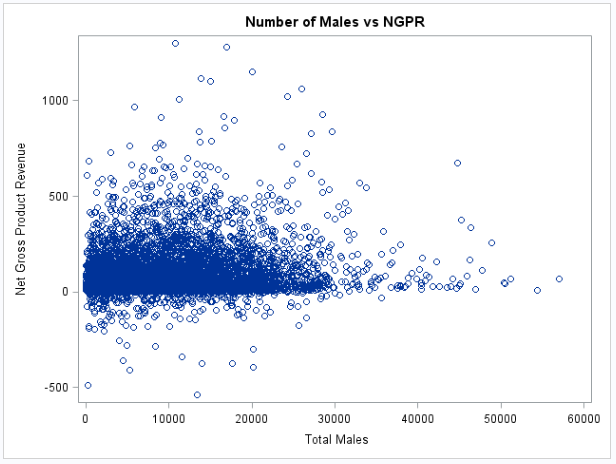
*Is there a correlation between median household income and spending?* – do peoples’ incomes affect how much they spend? There may be a direct or inverse relationship that needs to be explored.



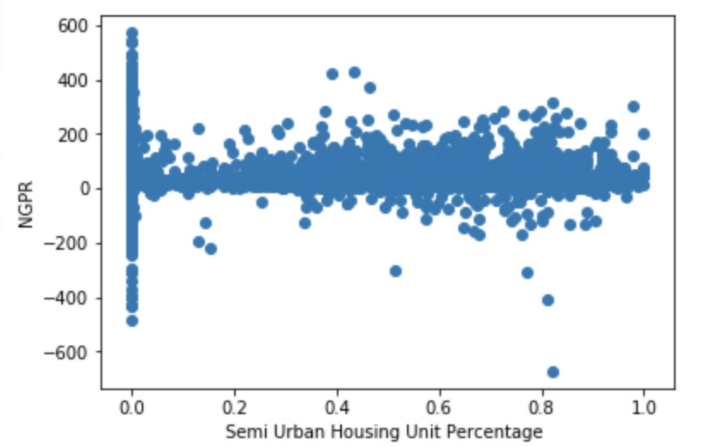
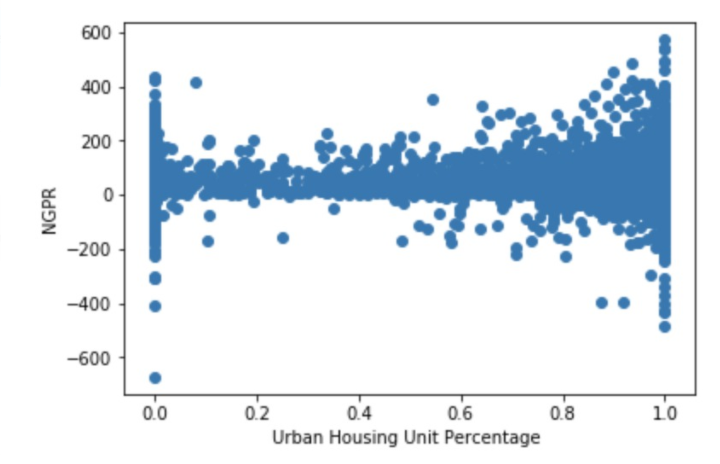
*Is there a correlation between total population and spending?* If we can determine how much an area is spending per person, we can determine who to target.

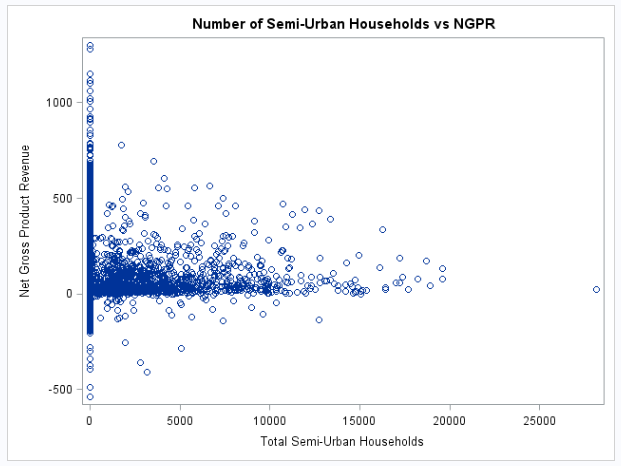
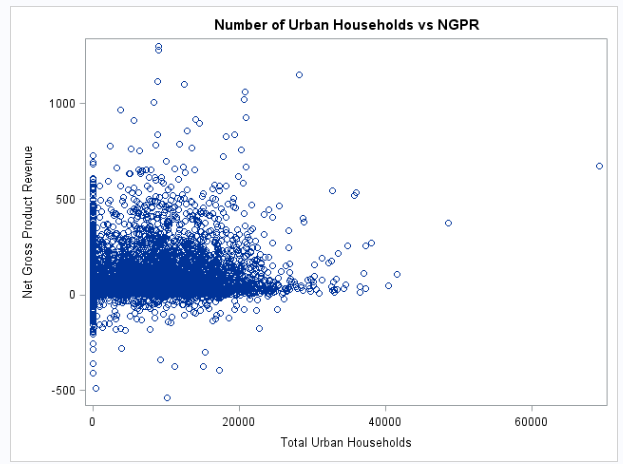


*How do the numbers of Males and Females in a population affect spending?* If we can determine whether gender plays a role, we can target a specific gender.

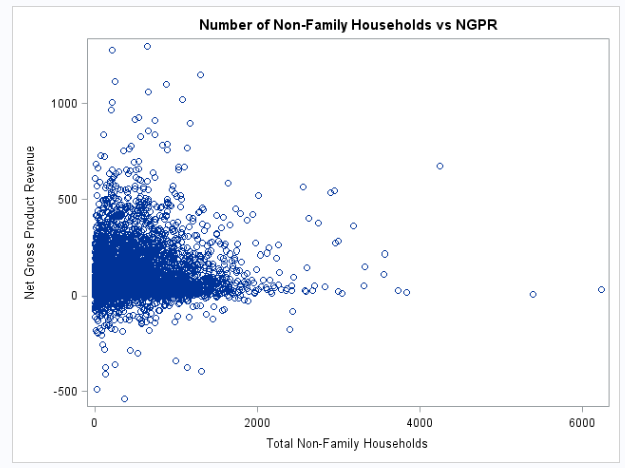
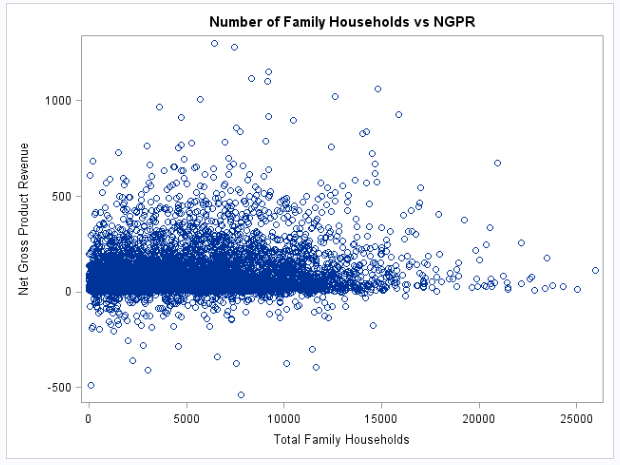


*How do the Percentages and numbers of Urban Households and semi-urban households affect spending?* – We used Urban and Semi-urban because rural households tend to be farther away and more spread apart so costs of shipping will often be more expensive.



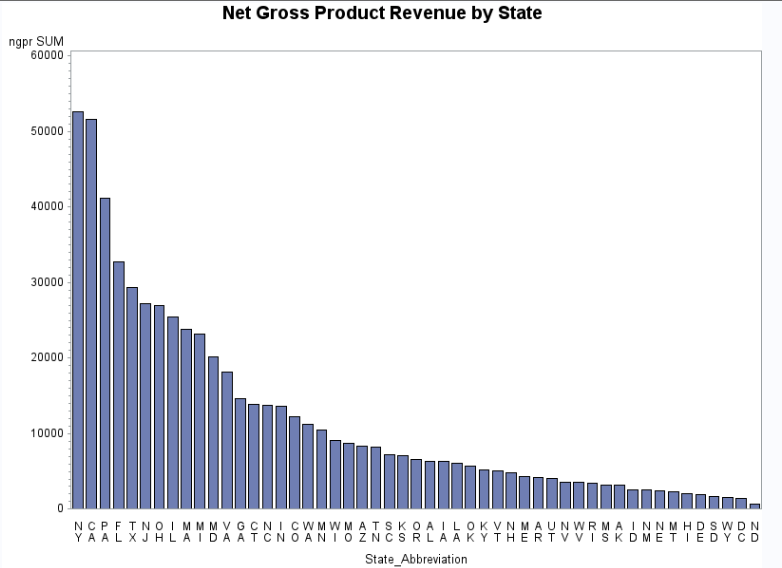


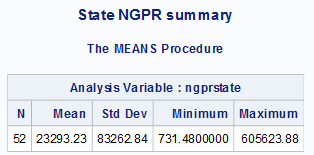
*Do areas with more family households spend more than non-family households?* – If family households spend more, we can make advertisements more family-friendly.



**Analysis, Results, and Conclusions**

1. *We need to target states with the highest Net Gross Product Revenues*





The States with the highest NGPRs are New York, California, Pennsylvania, Florida, Texas, New Jersey, Ohio, Illinois, Massachusetts, and Michigan should be the target for more advertisements or incentives to buy more things. They all have higher NGPRs than average and also happen to have the largest populations so advertisements will reach more people. One might try to target the states with lower NGPRs but that would allocate resources away from the high earning states. It is also less efficient to ship items to these farther away states since oftentimes the infrastructure is not as good due to low population. Having to send a truck or plane to a remote area for only a few customers is not economical. Another strategy would be to target states with relatively low NGPRs but high populations.

1. *Are people spending more online?*

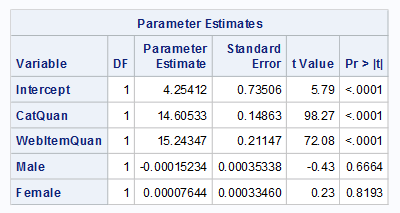
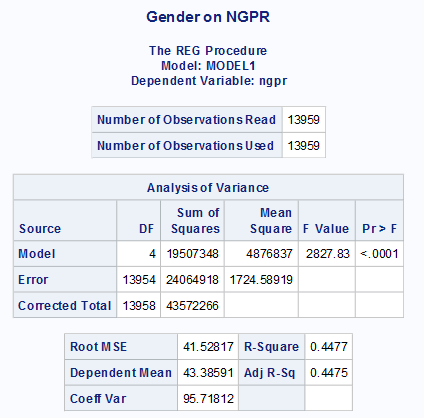
H0: There is no significant difference in how much people spend online vs offline.

HA: There is a difference between how much people spend online vs offline.



As seen in the t-test performed above, people tend to spend more when ordering from sources other than online. The difference is only about 43 cents but $0.43 can make a large difference in thousands of orders. According to the results of this test, the company should spend more resources on offline orders in order to maximize NGPR. BUT our suggestion is to encourage more online orders. The reason is because offline order NGPR per order is only greater because a customer needs to order everything they may need all at once; whereas, in online orders they can order something with just a few clicks whenever they need something and don’t have to wait for a catalog to be mailed or go to a kiosk. As the world is becoming more modern, online shopping is becoming more prevalent.

1. *Encourage package and bulk shopping*



According to the regression, more females have a positive effect on NGPR while more males have a slight negative effect. This leads us to believe that Females tend to spend a bit more when ordering from the Catalogue. However, the results were not significant. The Catalogue Item Quantity and Web Item Quantity are both positive which makes sense since more items ordered equals more money for the company. A discount for more items purchased could be implemented to increase sales and encourage the purchase of more items. The R-Squared value is 0.4477 which means that 44.77% of the variation in Net Gross Product Revenue can be explained by the variables Catalogue Item Quantity, Web Item Quantity, Male, and Female.

**Code Appendix**

/\*Mini Project 1\*/

/\*Importing the dataset\*/

**DATA** catalogue;

infile "C:\Users\Hanson\Documents\Hanson\College\UTD Spring 2018\Predicive Analaysis using SAS - BUAN 6337\Project\Catalog Data\Catalog-Data.TXT";

input ID **1**-**10** ORDNUM **11**-**24** DATE MMDDYY10. SubInd $ **35** SubQuan **36**-**42** CatInd $ **43** CatQuan **44**-**50** NumRecip **51**-**64** GPR **65**-**72** SaH **73**-**80**

SalesTax **81**-**88** Cancel **89**-**96** Return **97**-**104** AccRecAm **105**-**112** RefundDate MMDDYY10. RefundAm **123**-**130** RefStatus $ **131** RefType $ **132**

RefReason $ **133** OfferType $ **134** OfferDrop **135** Offer $ **136**-**138** GCRed $ **139** GCAm **140**-**147** CoupRedInd $ **148** CoupAm **149**-**156** GiftInd $ **157**

PayType $ **158**-**161** PayCat $ **162** PayStatus $ **163** OrdType $ **164** AncItemInd $ **165** AncItemQuan **166**-**179** AddChargeAm **180**-**187** AddChargCode $ **188**-**189**

WebItemInd $ **190** WebItemQuan **191**-**197** WOAm **198**-**205** DivCode **206**-**207** RankInDate **235**-**241** Zip **242**-**246**;

IF DATE < **1**/**25**/**2007** THEN ngpr = gpr - GCAm - CoupAm - RefundAm;

IF DATE > **1**/**24**/**2007** THEN ngpr = gpr - GCAm - CoupAm - SaH - SalesTax - RefundAm - Cancel - Return;

FORMAT DATE MMDDYY10. RefundDate MMDDYY10.;

**RUN**;

**PROC** **PRINT** data=catalogue(obs=**20**);

**RUN**;

/\*Drop uselss variables\*/

**DATA** catalogueN;

SET catalogue;

DROP Return AccRecAm RefStatus RefType RefReason OfferDrop GCRed GCAm CoupRedInd CoupAm GiftIndc AddChargeAm AddChargeCode WebItemInd

WOAm RankInDate;

**RUN**;

**PROC** **PRINT** data=catalogueN(obs=**20**);

**RUN**;

**PROC** **SORT** data = catalogueN;

BY Zip;

**RUN**;

/\*import excel file\*/

**PROC** **IMPORT** DATAFILE = "C:\Users\Hanson\Documents\Hanson\College\UTD Spring 2018\Predicive Analaysis using SAS - BUAN 6337\Project\us\_postal\_codes.csv"

OUT = ZipCode (rename=(Postal\_Code=Zip));

**RUN**;

**PROC** **PRINT** data=ZipCode(obs=**20**);

**RUN**;

**PROC** **IMPORT** DATAFILE = "C:\Users\Hanson\Documents\Hanson\College\UTD Spring 2018\Predicive Analaysis using SAS - BUAN 6337\Project\US 2000 Census Data\Household Income Distribution.csv"

OUT = Income;

**RUN**;

**PROC** **PRINT** data=Income(obs=**20**);

**RUN**;

**PROC** **IMPORT** DATAFILE = "C:\Users\Hanson\Documents\Hanson\College\UTD Spring 2018\Predicive Analaysis using SAS - BUAN 6337\Project\US 2000 Census Data\Household Type by Householder Age and Family Type.csv"

OUT = AgeFamily;

**RUN**;

**PROC** **PRINT** data=AgeFamily(obs=**20**);

**RUN**;

**PROC** **IMPORT** DATAFILE = "C:\Users\Hanson\Documents\Hanson\College\UTD Spring 2018\Predicive Analaysis using SAS - BUAN 6337\Project\US 2000 Census Data\Population by Age and Gender.csv"

OUT = AgeGender;

**RUN**;

**PROC** **PRINT** data=AgeGender(obs=**20**);

**RUN**;

**PROC** **IMPORT** DATAFILE = "C:\Users\Hanson\Documents\Hanson\College\UTD Spring 2018\Predicive Analaysis using SAS - BUAN 6337\Project\US 2000 Census Data\Urban Rural Housing Units.csv"

OUT = UrbanRural;

**RUN**;

**PROC** **PRINT** data=UrbanRural(obs=**20**);

**RUN**;

/\*Merge Code\*/

**DATA** Merged;

MERGE CatalogueN(in=inCatalogueN) ZipCode(DROP = State Latitude Longitude VAR8 in=inZipCode);

BY Zip;

IF inCatalogueN and inZipCode;

**RUN**;

**PROC** **PRINT** DATA=Merged(obs=**20**);

**RUN**;

/\* Merging with the Household Income distribution\*/

**DATA** MergedInc;

MERGE Merged(in=inMerged) Income(KEEP = Zip Total\_Households Median\_HH\_Income in=inIncome);

BY Zip;

IF inMerged and inIncome;

**RUN**;

**PROC** **PRINT** DATA=MergedInc(obs=**20**);

**RUN**;

/\*Merging with the Household Type\*/

**DATA** MergedFamtype;

MERGE MergedInc(in=inMergedInc) AgeFamily(KEEP = Zip Family\_households\_ Nonfamily\_households\_\_\_\_Househol in=inAgeFamily);

BY Zip;

IF inMergedInc and inAgeFamily;

**RUN**;

**PROC** **PRINT** DATA=MergedFamtype(obs=**20**);

**RUN**;

/\*Merging with Population Type\*/

**DATA** MergedAgeGen;

MERGE MergedFamtype(in=inMergedFamtype) AgeGender(KEEP = Zip Total\_population Male Female in=inAgeGender);

BY Zip;

IF inMergedFamtype and inAgeGender;

**RUN**;

**PROC** **PRINT** DATA=MergedAgeGen(obs=**20**);

**RUN**;

/\*Merging with Housingunits Type\*/

**DATA** MergedUrbRur;

MERGE MergedAgeGen(in=inMergedAgeGen) UrbanRural(in=inUrbanRural);

BY Zip;

IF inMergedAgeGen and inUrbanRural;

**RUN**;

**PROC** **PRINT** DATA=MergedUrbRur(obs=**20**);

**RUN**;

/\* This is the dataset with all files merged\*/

**DATA** MergedTotal;

SET MergedUrbRur;

**RUN**;

**PROC** **PRINT** data = MergedTotal(obs=**20**);

**RUN**;

/\*creating SumNGPR\*/

**PROC** **MEANS** data=MergedTotal SUM;

CLASS Zip;

VAR ngpr;

OUTPUT OUT = SumNGPR SUM = Sumngpr;

**RUN**;

**DATA** Final;

MERGE MergedTotal (in=inMergedTotal) SumNGPR (in=inSumNGPR);

BY Zip;

IF inMergedTotal and inSumNGPR;

**RUN**;

**PROC** **PRINT** DATA=Final(obs=**20**);

**RUN**;

/\*Refund Analysis

TITLE1 "Areas with highest amount of Refunds";

PROC MEANS data=Final SUM;

CLASS Zip;

VAR RefundAm;

OUTPUT OUT = RefundTotal (drop=\_TYPE\_ \_FREQ\_) SUM=RefTotal;

RUN;

DATA Final1;

MERGE Final (in=inFinal) RefundTotal (in=inRefundTotal);

BY Zip;

IF inFinal and inRefundTotal;

RUN;

PROC PRINT DATA=Final1(obs=20);

RUN;

PROC SORT data=Final1 OUT=REFUND1 (Keep=Place\_Name State\_Abbrevation County RefTotal);

BY RefTotal DESCENDING;

RUN;

PROC PRINT data=REFUND1(obs=20);

RUN;\*/

/\*charts, tables, and graphs\*/

**PROC** **MEANS** data=MergedTotal MEAN;

CLASS State\_Abbreviation;

VAR Median\_HH\_Income;

OUTPUT OUT=AvgMedInc MEAN=AvgInc;

**RUN**;

**PROC** **PRINT** data=AvgMedInc(OBS=**20**);

**RUN**;

/\*State NGPR\*/

**PROC** **MEANS** data=Final SUM;

CLASS State\_Abbreviation;

VAR ngpr;

OUTPUT OUT=statengpr SUM=ngprstate;

**RUN**;

TITLE1 "State NGPR summary";

**PROC** **MEANS** data=statengpr;

VAR ngprstate;

**RUN**;

TITLE1 "Average Median HH Income by State";

**PROC** **GCHART** data=AvgMedInc;

vbar State\_Abbreviation /TYPE=SUM sumvar=AvgInc descending;

**RUN**;

TITLE1 "Net Gross Product Revenue by State";

**PROC** **GCHART** data=Final;

vbar State\_Abbreviation/TYPE=SUM sumvar=ngpr descending;

**RUN**;

TITLE1 "Median HH Revenue by State";

TITLE1 "Online vs Offline Orders";

**PROC** **MEANS** data=Final SUM;

CLASS WedItemInd;

VAR ngpr;

**RUN**;

**PROC** **SGPLOT** DATA=Final;

SCATTER X=Total\_Households Y=sumngpr;

TITLE 'Number of Households vs NGPR';

LABEL Total\_Households = 'Total Households' sumngpr='Net Gross Product Revenue';

**RUN**;

**PROC** **SGPLOT** DATA=Final;

SCATTER X=Median\_HH\_Income Y=sumngpr;

TITLE 'Median HH Income vs NGPR';

LABEL Median\_HH\_Income = 'Median HH Income' sumngpr='Net Gross Product Revenue';

**RUN**;

**PROC** **SGPLOT** DATA=Final;

SCATTER X=Total\_Population Y=sumngpr;

TITLE 'Population vs NGPR';

LABEL Total\_Population = 'Total Population' sumngpr='Net Gross Product Revenue';

**RUN**;

**PROC** **SGPLOT** DATA=Final;

SCATTER X=Male Y=sumngpr;

TITLE 'Number of Males vs NGPR';

LABEL Male = 'Total Males' sumngpr='Net Gross Product Revenue';

**RUN**;

**PROC** **SGPLOT** DATA=Final;

SCATTER X=Female Y=sumngpr;

TITLE 'Number of Females vs NGPR';

LABEL Female = 'Total Females' sumngpr='Net Gross Product Revenue';

**RUN**;

**PROC** **SGPLOT** DATA=Final;

SCATTER X=Urban Y=sumngpr;

TITLE 'Number of Urban Households vs NGPR';

LABEL Urban = 'Total Urban Households' sumngpr='Net Gross Product Revenue';

**RUN**;

**PROC** **SGPLOT** DATA=Final;

SCATTER X=Semi\_Urban Y=sumngpr;

TITLE 'Number of Semi-Urban Households vs NGPR';

LABEL Semi\_Urban = 'Total Semi-Urban Households' sumngpr='Net Gross Product Revenue';

**RUN**;

**PROC** **SGPLOT** DATA=Final;

SCATTER X=Family\_households\_ Y=sumngpr;

TITLE 'Number of Family Households vs NGPR';

LABEL Family\_households\_ = 'Total Family Households' sumngpr='Net Gross Product Revenue';

**RUN**;

**PROC** **SGPLOT** DATA=Final;

SCATTER X=Nonfamily\_households\_\_\_\_Househol Y=sumngpr;

TITLE 'Number of Non-Family Households vs NGPR';

LABEL Nonfamily\_households\_\_\_\_Househol = 'Total Non-Family Households' sumngpr='Net Gross Product Revenue';

**RUN**;

/\*Tests and Regressions\*/

/\*Web items vs catalogue items\*/

TITLE1 "Web orders vs Catalogue orders";

**proc** **ttest** data=Final;

class WedItemInd;

var ngpr;

**run**;

/\*proc reg data=Final;

model ngpr = CatQuan WebItemQuan Urban Semi\_Urban Rural\_\_\_\_Nonfarm Rural\_\_\_\_Farm;

RUN;\*/

/\*Regression using gender as a variable\*/

TITLE1 "Gender on NGPR";

**proc** **reg** data=Final;

model ngpr = CatQuan WebItemQuan Male Female;

**RUN**;

TITLE1 "Gender on NGPR";

**proc** **reg** data=Final;

model ngpr = CatQuan WebItemQuan Female;

**RUN**;

**PROC** **CORR** data=Final;

var ngpr CatQuan WebItemQuan Male Female;

**RUN**;

/\*Family vs non-family

proc reg data=Final;

model ngpr = CatQuan WebItemQuan Family\_households\_ Nonfamily\_households\_\_\_\_Househol;

RUN;\*/