**Project Option - 1: 2016 US Election**

**Student: Aishwarya Ravi, Hong Zhang**

1. **Business Scenario**:

The presidential election season is upon us. On November 8, 2016, Americans will head to the polls and choose their new president. Hillary Clinton is the presidential nomination from the Democratic Party who is the first female presidential nominee of a major party in U.S. history. Will she win this election and become the first female president?

1. **Objectives**:

The presidential primary elections and caucuses were held between February 1 and June 14, 2016, staggered among the 50 states, Washington, D.C., and U.S. territories. They determined the Republican and Democratic presidential nominations. Winning the early primaries is a major key for electoral victory in November. In the first model, we will use primary result and choose the variable, fraction\_votes to predict which party will win this election.

1. **Data Sources**:

|  |  |  |  |
| --- | --- | --- | --- |
| Data Source | Files | total number of rows | time period |
| <https://www.kaggle.com/benhamner/2016-us-election> | primary\_results | 24,611 | 2016 |
| <http://www.presidency.ucsb.edu/elections.php> | history\_final\_results | 450 | 1980~2012 |
| Still missing | history\_primary\_results | ? | ? |

1. **Variable Selection:**

**Total: 65 variables**

|  |  |  |
| --- | --- | --- |
| **Predictor(X)** | **Variable** | **Type** |
| **fraction\_votes** | **continuous** |
| **female** | **continuous** |
| **Over65years** | **continuous** |
| **Response(Y)** | **party** | **categorical** |

1. **SAS Code**

**The SAS procedures we are planning to use are**

**Proc means**

**Proc freq**

**Proc univariate**

**Proc sgplot,…**

**Sample Data selection from the dataset:**

/\*Select sample data from dataset\*/

/\*Sample data\*/

data election.data\_set;

set election.data;

keep state party candidate votes fraction\_votes county;

where votes>=10000;

run;

**Exploratory Data Analysis**

/\*Frequency Plot to understand data distribution\*/

/\*Exploration of all variables that are available for analysis.\*/

/\*%let statements define macro variables containing lists of continuous variables\*/

%let tfilename=election.data\_set;

%let interval= votes fraction\_votes;

/\*UNIVARIATE proc is used to plot histogram, probability graph and to display basic statistics\*/

proc univariate data=&tfilename;

var &interval;

histogram &interval / normal kernel;

inset n mean median std skewness kurtosis / position=ne;

probplot &interval / normal (mu=est sigma=est);

inset skewness kurtosis;

title 'Descriptive Statistics Using PROC UNIVARIATE';

run;

**Exploration of all categorical variables that are available for analysis**

/\*Exploration of all categorical variables that are available for analysis.\*/

%let tfilename=election.data\_set;

%let categorical= party candidate state county;

proc means data=&tfilename maxdec=2 fw=10 printalltypes n mean median std var

q1 q3;

class &categorical;

var fraction\_votes;/\* U can check fr votes also\*/

output out=means mean=votes;

title 'Selected Descriptive Statistics for number of votes';

run;

title;

/\*proc FREQ displays frequency graph of categorical variables\*/

proc freq data=&tfilename;

tables &categorical / plots=freqplot;

title "Categorical Variable Frequency Analysis";

run;

title;

**Association between categorical response and continuous predictores using SGPLOT (VBOX**

/\* Association between categorical response and continuous predictors using SGPLOT (VBOX) \*/

proc sgplot data=election.data\_set;

vbox fraction\_votes/ category=party connect=mean;

run;

proc sgplot data=election.data\_set;

vbox votes/ category=party connect=mean;

run;

**Project Option-2: House Sales in King County, USA**

**Student: Aishwarya Ravi, Hong Zhang**

1. **Business Scenario**:

Across Washington, home prices are rising faster than in any state in the country — the first time that’s happened in a quarter-century. The housing prices hit new highs and inventory is at new lows. For example, King County’s housing market hit two milestones in December, 2015: the median single-family home price set a new high, $508,000, topping the $481,000 peak reached in July 2007 before prices began their long slide. At the same time, the supply of available homes has been at less than two months for nearly two years. For most of 2016, inventory has hovered around one month's supply. Buyers think it is a best time to buy a house in this state. However, what is a good dealer?

1. **Objectives**:

King County is the most populous county in Washington. The county seat is Seattle, which is the state's largest city. We will choose the house sale prices from 2014 May to 2015 May in this county to do analysis. There are many factors describing the condition of a house, and they do not weigh equally in determining the home value. We will try to find these factors to influent the housing price and build a model to predict the house price.

1. **Data Sources**:

This dataset contains the data of homes sold between May 2014 and May 2015 in King County.

|  |  |  |  |
| --- | --- | --- | --- |
| Data Source | Files | total number of rows | time period |
| <https://www.kaggle.com/harlfoxem/housesalesprediction> | kc\_house\_data | 21,613 | 2014/5/2 ~2015/5/27 |

1. **Variable Selection:**

**Total: 21 variables**

|  |  |  |
| --- | --- | --- |
| **Predictor(X)** | **Variable** | **Type** |
| **bedrooms** | **continuous** |
| **Sqft-living** | **continuous** |
|  | **bathrooms** | **continuous** |
| **Response(Y)** | **pirce** | **continuous** |

1. **SAS Code**

data house.h\_data;

       set house.house\_data;

      run;

  /\*Exploration of all variables that are available for analysis.\*/

  /\*%let statements define macro variables containing lists of \*/ /\*dataset variables\*/

%let interval= view bedrooms sqft\_living sqft\_lot bathrooms view;

proc sgscatter

data=house.h\_data;

plot price\*(&interval) / reg;

title "Associations of Interval

 Variables with House Sale Price";

run;

%let tfilename=house.h\_data;

%let interval= bedrooms sqft\_living sqft\_lot bathrooms view

;

/\*SGSCATTER displays scatter plot with regression line for continuous predictors and response variable\*/

proc sgscatter data=&tfilename;

    plot price \*(&interval) / reg;

    title "Associations of Price with predictors";

run;

/\*Correlation between predictors and response \*/

ods graphics / reset=all imagemap;

proc corr data=&tfilename rank

          plots(only)=scatter(nvar=all ellipse=none);

   var &interval;

   with price;

   title "Correlations and Scatter Plots with sales price";

run;

title

**Project Option-3: Income in Illinois, USA**

**Student: Aishwarya Ravi, Hong Zhang**

1. **Business Scenario**:

Everybody must care about the same question what I can earn more money. "Success is a learnable skill," emphasizes T. Harv Eker in his book "Secrets of the Millionaire Mind." "If you want to be a great golfer, you can learn how to do it. If you want to be a great piano player, you can learn how to do it ... If you want to be rich, you can learn how to do it." If you want to learn how to get rich — how to grow and master your money — consider these variable which can affect your income.

1. **Objectives**:

American Community Survey Office is the premier source for detailed information about the American people and workforce. We will choose the dataset of its population survey from 2010 to 2014 and try to do analysis. The study location will be our living state, Illinois. There are many factors describing a person, and they do not weigh equally in determining this person’s income. We will try to find them, cross compare and build a multiple regression model to predict the income.

1. **Data Sources**:

This dataset contains the data of American Community Survey 2010-2014 ACS 5-year PUMS.

Prepared by:

American Community Survey Office

U.S. Census Bureau

Create: ‎October ‎28, ‎2015

|  |  |  |  |
| --- | --- | --- | --- |
| Data Source | Files | total number of rows | time period |
| <http://www.census.gov/programs-surveys/acs/data/pums.html> | psam\_p17.sas7bdat | 635532 | 2000~2014 |

1. **Variable Selection:**

**Total: 293 variables**

|  |  |  |
| --- | --- | --- |
| **Predictor(X)** | **Variable** | **Type** |
|  |  |
|  |  |
|  |  |  |
| **Response(Y)** |  |  |

1. **SAS Code**

%let path=/home/hzhan1210/census;

libname census "&path";

/\*Select particular variables from data set\*/

data census.psam\_p17\_subset;

set census.psam\_p17;

if AGEP < 16 then delete;

if PINCP < 100 then delete;

if WKHP <=0 then delete;

if FOD1P <0 then delete;

if SCHL <0 then delete;

PINCP = PINCP\*(ADJINC/1000000);

keep SCH SEMP AGEP PINCP ADJINC WKH WKHP COW SCHL FOD1P MAR OCCP10 RELP RAC1P SEX SCHG ANC1P05 ANC2P05;

label

SCH ='School Enrolment'

SEMP ='Self employment Income'

AGEP = 'Age'

PINCP = 'Total person income (signed)'

ADJINC = 'Adjustment factor for income and earnings dollar amounts (6 implied decimal places)'

WKHP = 'Usual hours worked per week past 12 months'

COW = 'Class of worker'

SCHL = 'Educational attainment'

FOD1P = 'Recoded field of degree - first entry'

MAR = 'Marital status'

RELP = 'Relationship'

RAC1P = 'Recoded detailed race code - 0'

SEX = 'Sex'

OCCP10 = 'Occupation recode'

SCHG='Grade Level Attending'

ANC1P05 = 'Recoded Detailed Ancestry - first entry'

ANC2P05 = 'Recoded Detailed Ancestry - second entry';

run;

/\*Get the sample dataset of 100 obs\*/

proc surveyselect data=census.psam\_p17\_subset

method=srs n=101 out=census.psam\_p17\_subset1;

run;

/\*Exploration of all variables that are available for analysis.\*/

/\*%let statements define macro variables containing lists of continuous variables\*/

%let tfilename=census.psam\_p17\_subset1;

%let interval= AGEP WKHP PINCP SEMP;

/\*UNIVARIATE proc is used to plot histogram, probability graph and to display basic statistics\*/

proc univariate data=&tfilename;

var &interval;

histogram &interval / normal kernel;

inset n mean median std skewness kurtosis / position=ne;

probplot &interval / normal (mu=est sigma=est);

inset skewness kurtosis;

title 'Descriptive Statistics Using PROC UNIVARIATE';

run;

/\*Exploration of all categorical variables that are available for analysis.\*/

%let categorical= COW SCHL SCH SCHG FOD1P MAR RELP SEX ANC1P05 ANC2P05 ;

%let tfilename=census.psam\_p17\_subset1;

proc means data=&tfilename maxdec=2 fw=10 printalltypes

n mean median std var q1 q3;

class &categorical;

var PINCP;

output out=means mean=PINCP;

title 'Selected Descriptive Statistics for Person Income';

run;

title;

/\*proc FREQ displays frequency graph of categorical variables\*/

proc freq data=&tfilename;

tables &categorical / plots=freqplot ;

title "Categorical Variable Frequency Analysis";

run;

title;

/\*Survey select only 100 obs\*/

proc surveyselect data=census.psam\_p17\_subset

method=srs n=5000 out=census.data\_corrl;

run;

proc print data=census.data\_corrl(obs=5000);

run;

proc contents data=census.data\_corrl;

run;

/\*PROC SGPLOT is used here with the VBOX statement to produce vertical bar charts\*/

/\*PROC SGPLOT can only produce one plot at a time and so the macro is written to\*/

/\*produce one plot for each member in the list of the &categorical macro variable.\*/

%let categorical= SCH SCHL SCHG MAR RELP SEX COW FOD1P AGEP ANC1P05 ANC2P05 ;

%let tfilename=census.data\_corrl;

ods graphics on/width= 800;

%macro box(dsn = ,

response = ,

Charvar = );

%let i = 1 ;

%do %while(%scan(&charvar,&i,%str( )) ^= %str()) ;

%let var = %scan(&charvar,&i,%str( ));

proc sgplot data=&dsn;

\*panelby season;

vbox &response / category=&var

grouporder=ascending

connect=mean;

title "&response across Levels of &var";

run;

%let i = %eval(&i + 1 ) ;

%end ;

%mend box;

%box(dsn = &tfilename,

response = PINCP,

charvar = &categorical);

title;

options label;

%let tfilename=census.data\_corrl;

%let interval= SEMP AGEP WKHP ;

/\*SGSCATTER displays scatter plot with regression line for continuous predictors and response variable\*/

proc sgscatter data=&tfilename;

plot PINCP\*(&interval) / reg;

title "Associations of Income with predictors";

run;

/\*Correlation between predictors and response \*/

ods graphics / reset=all imagemap;

proc corr data=&tfilename rank

plots(only)=scatter(nvar=all ellipse=none);

var &interval;

with PINCP;

title "Correlations and Scatter Plots with Income";

run;

title;