**Project: 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 this 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 |

1. **Variable Selection:**

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

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

**Proc means**

**Proc freq**

**Proc univariate**

**Proc sgplot,…**

1. **SAS Code**

**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: 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. Is it the best time for us to buy a house in this state right now?

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 future market in King County.

**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 |