Illapani_IS606_Proposal

Srini Illapani

October 16, 2015

```
# Load the necessary packages to be used for this project
library(dplyr)

##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyr)
library(knitr)
library(ggplot2)

# load data - Upload the file to my github repository and use the curl
package to retrieve the dataset in to R.
library(curl)

endorsements <-
read.csv(curl("https://raw.githubusercontent.com/isrini/SI_IS606/master/endor
sements.csv"), header = TRUE)

kable(head(endorsements,10))
```

year	party	candidate	endorsement_points	percentage_endorsement_points	money_raised	percentage_of_money	primary_vote_percentage	won_primary
1980	Republican	George H.W. Bush	5	6.32911	1475332	16.83	23.81	No
1980	Republican	Lowell Weicker	0	0.00000	60000	0.68	0.00	No

1980	Republican	Phil Crane	0	0.00000	2500000	28.53	0.76	No
1980	Republican	John Connally	6	7.59494	2200000	25.10	0.64	No
1980	Republican	Bob Dole	0	0.00000	253286	2.89	0.06	No
1980	Republican	Ronald Reagan	58	73.41770	1400000	15.97	59.79	Yes
1980	Republican	Howard Baker	10	12.65820	643373	7.34	1.41	No
1980	Republican	John Anderson	0	0.00000	201745	2.30	12.19	No
1980	Republican	Larry Pressler	0	0.00000	0	0.00	0.00	No
1980	Republican	Harold Stassen	0	0.00000	30000	0.34	0.20	No

Research question

Do number of endorsements increase the chances of winning the primaries and party nomination?
We would also attempt to see if the money raised by the candidates is any indicator of winning the primaries, if time permits.

Cases

What are the cases, and how many are there?

We will be looking at the democratic and republican party candidates for presidential primaries from the year 1980 to 2012.

Data collection

Describe the method of data collection.

Fundraising data since 1992 is taken from the Federal Election Commission website. Data from 1980 to 1988 is from various news articles at the time of the filing deadline. This data is preliminary, though rarely differs greatly from finalized data.

All the data used excludes self-funding by the candidates.

Type of study

What type of study is this (observational/experiment)?

The presidential candidate endorsements for primary elections is an observational study performed by FiveThirtyEight by collecting the data as mentioned above.

Data Source

The data souce for the puropose of this Data project is FiveThirtyEight (https://github.com/fivethirtyeight/data). Here is the list of variables that will be used for the project:

Value	Description
year	Election year
party	Political party
candidate	Candidate running in primary
endorsement_points	Weighted endorsements through June 30th of the year before the primary
percentage_endorsement_points	Percentage of total weighted endorsement points for the candidate's political party through June 30th of the year before the primary
money_raised	Money raised through June 30th of the year before the primary
percentage_of_money	Percentage of total money raised by the candidate's political party through June 30th of the year before the primary
primary_vote_percentage	Percentage of votes won in the primary
won_primary	Did the candidate win the primary?

Response

What is the response variable, and what type is it (numerical/categorical)?

The response variable for the question we are trying to answer - Do number of endorsements increase the chances of winning the primaries? is 'won_primary'.

It is of type categorial, and can also be treated as numerical by converting to 0 or 1.

Explanatory

What is the explanatory variable, and what type is it (numerical/categorical)?

The explanatory variables in the data set are 'endorsement_points' and 'money_raised'. These are of type numerical.

Relevant summary statistics

Provide summary statistics relevant to your research question. For example, if you're comparing means across groups provide means, SDs, sample sizes of each group. This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
# Sample statistics to be run answer the research question:

# Using group by and summarise to find the mean endorsements by the year and
party
mean = group_by(endorsements,year) %>%
    filter(endorsement_points != 0) %>%
    summarise(mean(endorsement_points))

head(mean)
```

```
## Source: local data frame [6 x 2]
##
##    year mean(endorsement_points)
##   (int)                    (dbl)
## 1  1980                 19.75000
## 2  1984                 23.66667
## 3  1988                 19.72727
## 4  1996                 58.80000
## 5  2000                 69.55556
## 6  2004                 13.60000
```