Medical Cost Personal Datasets

###Step 1

Read your dataset. Done

###Step 2

Write a description of your dataset: • What is the topic?

Medical Cost Personal Datasets

- Where you get it from?

https://www.kaggle.com/mirichoi0218/insurance Yes, it is open dataset.

- What kind of variables you have (numerical/categorical etc)? Describe the variables (the meaning of those).

Short Descriptions:

The independent variable is manipulated by the experimenter and its effects on the dependent variable are measured.

Qualitative variables (categorical variables) are those that express a qualitative attribute such as hair colour, eye colour, religion, favourite movie, gender, and so on.

Quantitative variables are those variables that are measured in terms of numbers (height, weight, shoe size).

Discrete variables can take only certain values (For example, a household could have three children or six children, but not 4.53 children)

Continuous variables can take any value within the range of the scale (For example, "time to respond to a question" are continuous variables since the scale is continuous,say, the response time could be 1.64 seconds).

AGE - dependent continuous variable, quantitative

SEX - categorical

BMI - dependent continuous variable, quantitative

Nr of CHILDREN - dependent discrete variable, quantitative

SMOKER - dependent variable, categorical

REGION - dependent variable, categorical

CHARGES -dependent continuous variable, quantitative

###Step 3

Look for missing values, or errors (NA etc) in the dataset.

Dataset looks complete

Step 4

Do some data visualization (distribution graphs).

```
{echo=FALSE}
knitr::opts_chunk$set(error = TRUE)
```

```
hist(insurance$age, xlab = ’age’, main = ’Age’, col = ’blue’, include = TRUE)
```

```
## Error in hist(insurance$age, xlab = "age", main = "Age", col = "blue", : object ’insurance’ not foun
```

```
table(insurance$sex)
```

```
## Error in table(insurance$sex): object ’insurance’ not found
```

```
hist(insurance$bmi, xlab = ’bmi’, main = ’BMI’, col = ’red’)
```

```
## Error in hist(insurance$bmi, xlab = "bmi", main = "BMI", col = "red"): object ’insurance’ not found
```

```
hist(insurance$children, xlab = ’children’, main = ’Children’, col = ’green’)
```

```
## Error in hist(insurance$children, xlab = "children", main = "Children", : object ’insurance’ not fou
```

```
table(insurance$region)
```

```
## Error in table(insurance$region): object ’insurance’ not found
```

```
hist(insurance$charges, xlab = ’charges’, main = ’Charges’, col = ’yellow’)
```

```
## Error in hist(insurance$charges, xlab = "charges", main = "Charges", col = "yellow"): object ’insura
```

###Step 5

Describe what can you see from those graphs.

Average age 39

Male require insurance more than female

Average BMI is 30.6634, so slightly overweight

Mostly the people doesn’t have any children

So basically, if you have many children, you smoke and you have extra weight your charges will be the highest:-)