Draft

Nima Jamshidi & Diana Lin

3/6/2020

Contents

Introduction

The dataset we have chosen to work with is the "Medical Expenses" dataset used in the book Machine Learning with R, by Brett Lantz. This dataset was extracted from Kaggle by Github user @meperezcuello. The information about this dataset has been extracted from their GitHub Gist.

This dataset is very interesting as the USA does not have universal healthcare, and is known for bankrupting its citizens with hospital visits despite having insurance. It will be interesting to see the relationship between characteristics of a beneficiary, such as `BMI` and `Smoking` status, and the `charges` incurred.

Research Question

In this study, we are analyzing the data to find a relationship between the features and the amount of insurance cost.

Does having an increased BMI increase your insurance costs? What about age? Number of dependents? Smoking status? Are certain areas of the USA associated with higher insurance costs?

In order to answer the questions above we're planning to perform a linear regression analysis and plot the regression line and relevant variables. The variables need to be normalized before performing the regression analysis.

Data Description

This dataset explains the medical insurance costs of a small sample of the USA population. Each row corresponds to a beneficiary. Various metadata was recorded as well.

The columns (except the last one) in this dataset correspond to metadata, where the last column is the monetary charges of medical insurance. Here are the possible values for each of the columns:

Variable	Type	Description
Age	integer	the primary beneficiary's age in years
Sex	factor	the beneficiary's sex: `female` or `male`
BMI	double	the beneficiary's Body Mass Index, a measure of their body fat based on height and weight (measured in kg/m2), an ideal range of 18.5 to 24.9
Children	integer	the number of dependents on the primary beneficiary's insurance policy
Smoker	factor	whether or not the beneficiary is a smoker: `yes` or `no`
Region	factor	the beneficiary's residential area in the USA: `southwest`, `southeast`, `northwest`, or `northeast`
Charges	double	the monetary charges the beneficiary was billed by health insurance

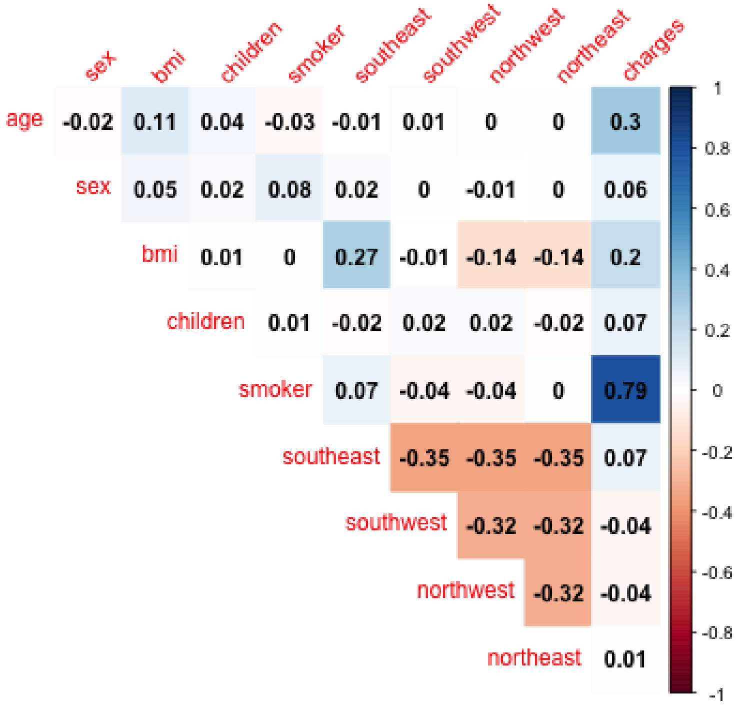
Exploring the Dataset

Here is a summary of the dataset, and the values of each variable:

age	sex	bmi	children	smoker	region	charges
Min. :18.00	female:662	Min. :15.96	Min. :0.000	yes: 274	southwest:325	Min. : 1122
1st Qu.:27.00	male :676	1st Qu.:26.30	1st Qu.:0.000	no :1064	southeast:364	1st Qu.: 4740
Median :39.00		Median :30.40	Median :1.000		northwest:325	Median : 9382
Mean :39.21		Mean :30.66	Mean :1.095		northeast:324	Mean :13270
3rd Qu.:51.00		3rd Qu.:34.69	3rd Qu.:2.000			3rd Qu.:16640
Max. :64.00		Max. :53.13	Max. :5.000			Max. :63770

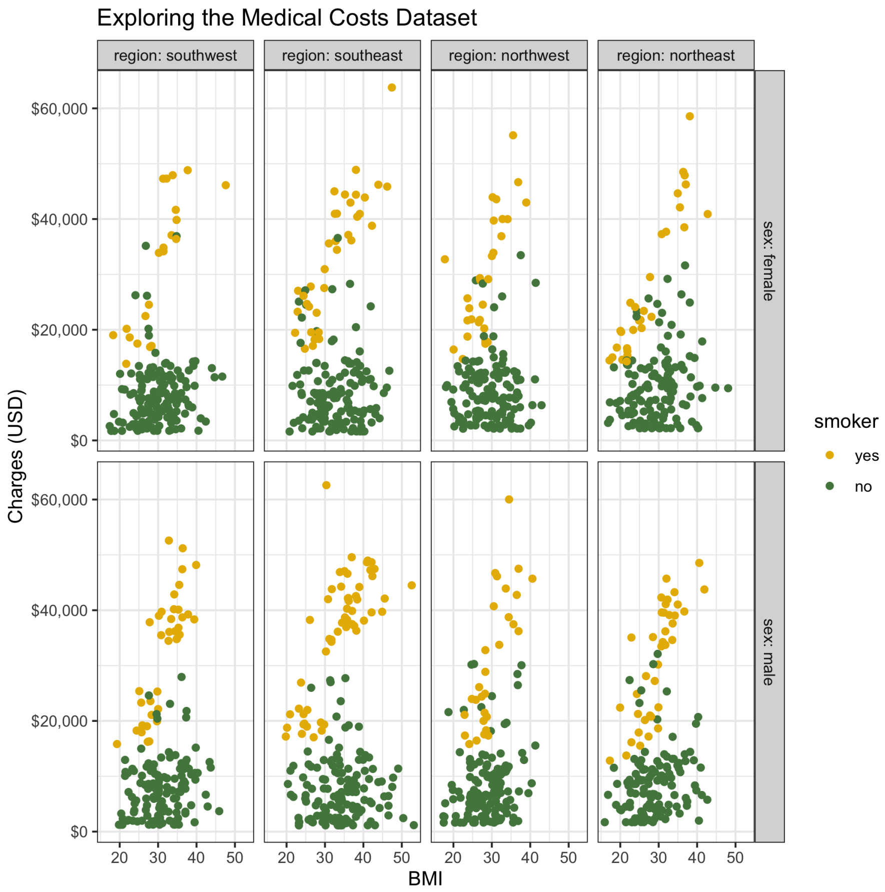
Next, we want to inspect the data set to see if there is any correlation between the variables. From now on we want to consider charges as our dependent variable. In order to analyze correlation between variables, the ones that are categorical with two categories, are translated into binery vectors. The only categorical variable with more than two categories, is region. We split this variable into four different binery vectors, each indicating if the sample data has category (1) or not (0).

After using dummy variables for sex, smoker, and region, according to the correlogram below, smoker and charges has the strongest correlation of 0.79. No high collinearity between independent variables is observed.

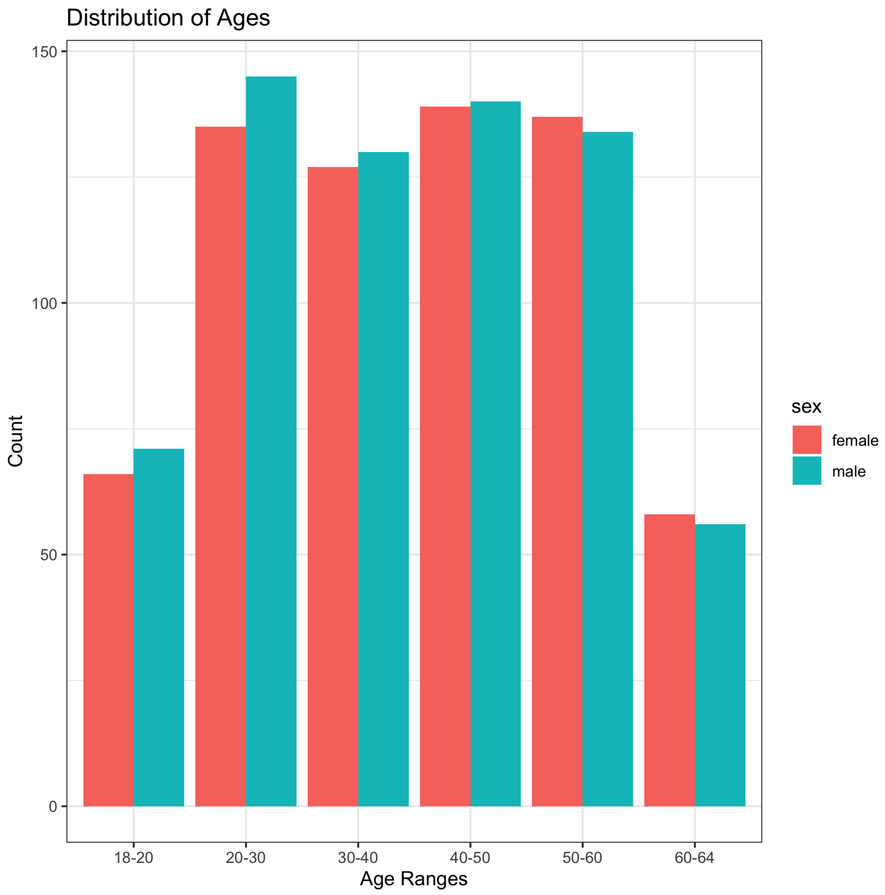


	sex	bmi	children	smoker	southeast	southwest	northwest	northeast	charges
age	-0.02	0.11	0.04	-0.03	-0.01	0.01	0	0	0.3
sex		0.05	0.02	0.08	0.02	0	-0.01	0	0.06
bmi			0.01	0	0.27	-0.01	-0.14	-0.14	0.2
children				0.01	-0.02	0.02	0.02	-0.02	0.07
smoker					0.07	-0.04	-0.04	0	0.79
southeast						-0.35	-0.35	-0.35	0.07
southwest							-0.32	-0.32	-0.04
northwest								-0.32	-0.04
northeast									0.01

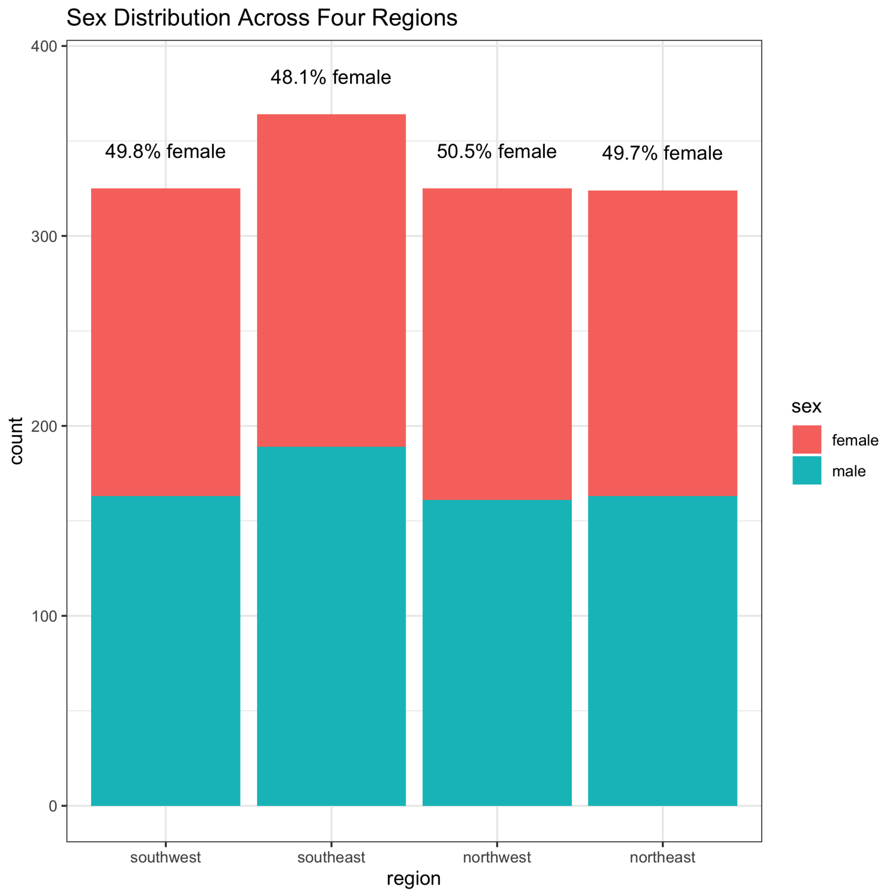
In order to to check if there is any cluster of data points, we use faceted plot. While the data between regions and sex does not appear to vary much, the smokers vs nonsmokers of each facet appear to cluster together, with the non-smokers having an overall lower medical cost.



How is the distribution of sex among different age groups? Looking at the dataset, there appears to be more beneficiaries in the 20-60 age range. The biggest difference in the number of beneficiaries from different sex is seen in the 20-30 bracket.



How about the distribution of sex among the regions? This plot shows the distribution of sex in each of the four regions. At a glance, the dataset looks very even when it comes to sex, but there are slightly more beneficiaries in the southeast.



Methods

```
# PLACE HOLDER FOR LINEAR REGRESSION
```

Results

```
# PLACE HOLDER FOR LINEAR REGRESSION
```

Discussion

```
# PLACE HOLDER FOR LINEAR REGRESSION
```

Conclusion

```
# PLACE HOLDER FOR LINEAR REGRESSION
```

References

1. Medical Costs Dataset - https://gist.github.com/meperezcuello/82a9f1c1c473d6585e750ad2e3c05a41
2. BMI - https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmi-m.htm