Analysis of Medical Cost Dataset

Jianyi Zhang
Chen Wang
Linyu Li

[10 / 01 / 2019]

CONTENTS

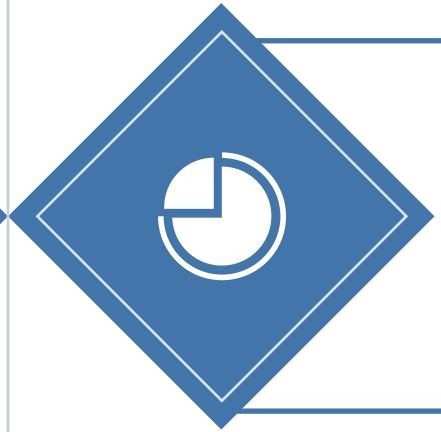
PART ONE

Overview Of Dataset

Overview of medical cost dataset

Source

This dataset is downloaded from kaggle. It was from a book *Machine Learning with R.* It was released from the US Census Bureau.



Content

It contains individual medical costs of 1338 persons and their personal features.

To make their own profits, the insurance company (insurer) must collect more premiums than the amount paid to the insured person. For this, the insurance company invests a lot of time and money in creating a model that accurately predicts health care costs.

Overview of medical cost dataset

Questions inspired

01

What affects the cost?

Is there a relationship between medical expenses and other variables in the dataset?

02

How the cost affected?

How do the key variables affect medical charges?

03

Prediction

To predict results for the medical charges of people, how would their healthcare insurance may vary.

Overview of medical cost dataset

7 Columns

Age

Age of the primary beneficiary

Sex

Gender of the insurance contractor: female, male

BMI

Body mass index, is a measure of body fat based on height and weight that applies to adult. Commonly accepted ranges: underweight: <18.5, normal weight: 18.5-25, overweight: 25-30, obese: >30 *

Children

Number of children covered by health insurance / Number of dependents

Smoker

If the policy holder consider him/herself a smoker

Region

The beneficiary's residential area in the US: northeast, southeast, southwest, northwest

Charges

Individual medical costs billed by health insurance

* "WHO Mean Body Mass Index (BMI)". World Health Organization. Retrieved 5 February 2019.

PART TWO

Exploratory Data Analysis

Exploratory Data Analysis

View dataset

	age	sex	BMI	children	smoker	region	charges
1	19	female	27.9	0	yes	southwest	16884.924
2	18	male	33.77	1	no	southeast	1725.5523
3	28	male	33	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.4706
5	32	male	28.88	0	no	northwest	3866.8552
6	31	female	25.74	0	no	southeast	3756.6216

Dimension

Dimension	
1338	7

Check missing value

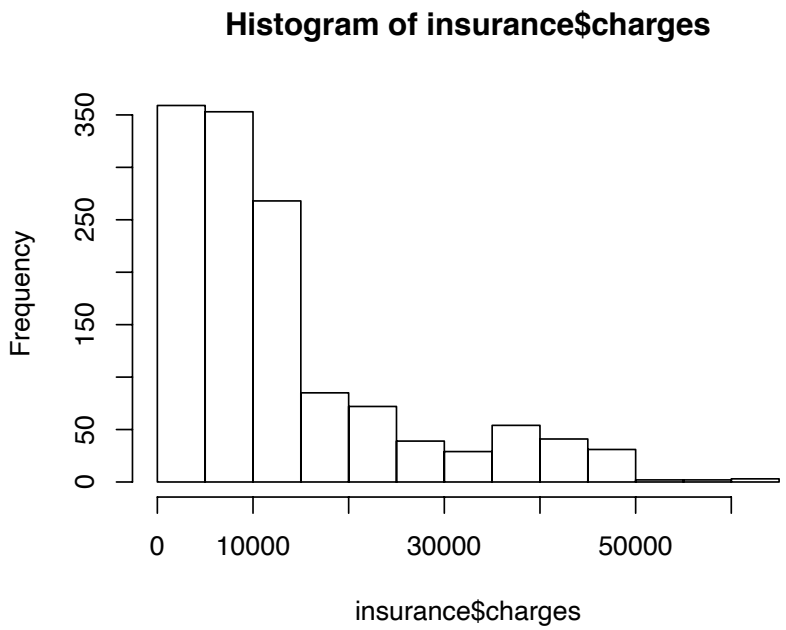
True	False
0	9366

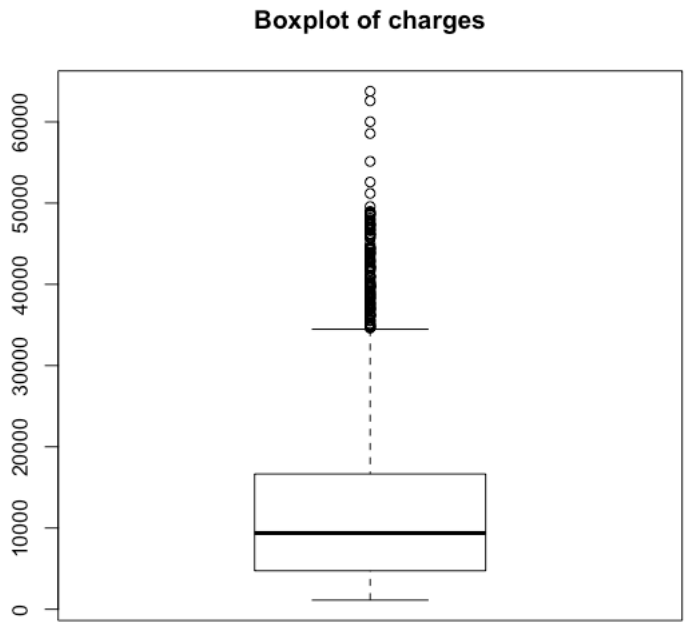
Charges: original values

Summary

Min	1st Quarter	Median	Mean	3rd Quarter	Max
1122	4740	9382	13270	16640	63770

Plots

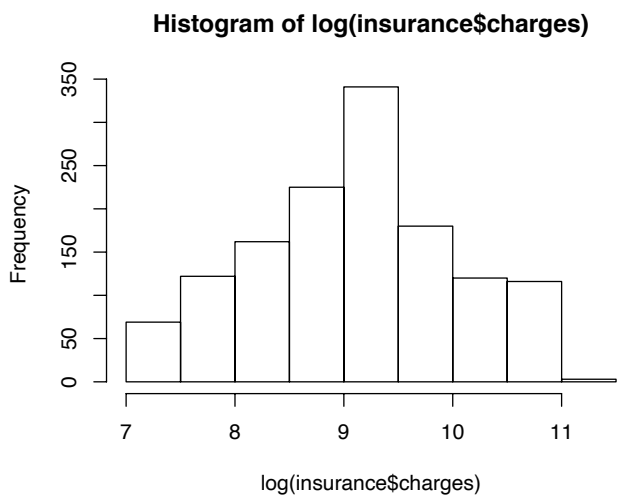


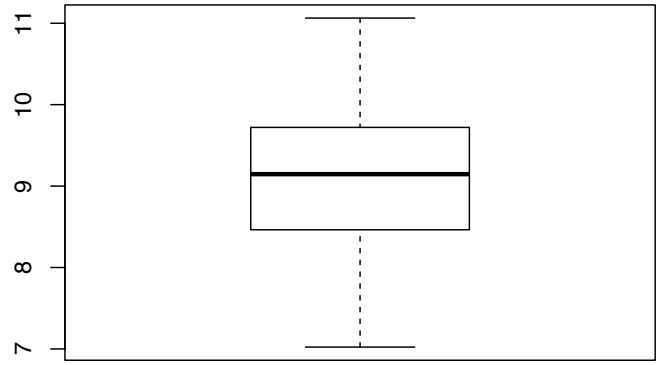


The original values of charges are right-skewed， not following the normal distribution.

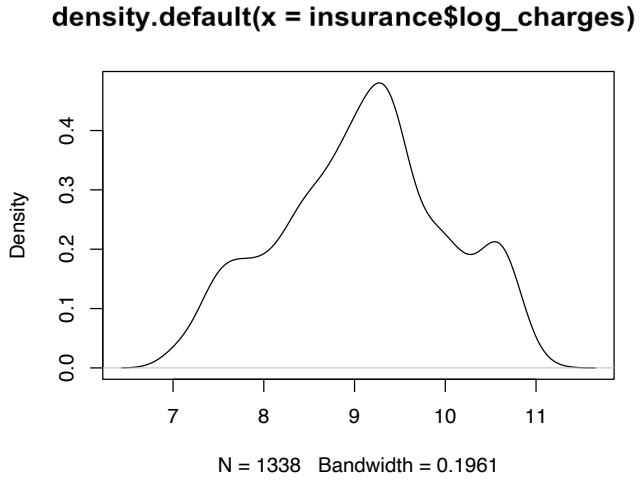
Charges: log transformation

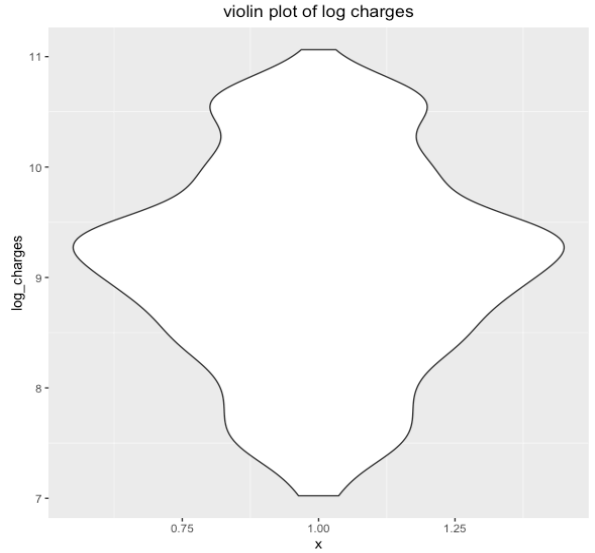
Histogram & Boxplots





Density & Violin plots





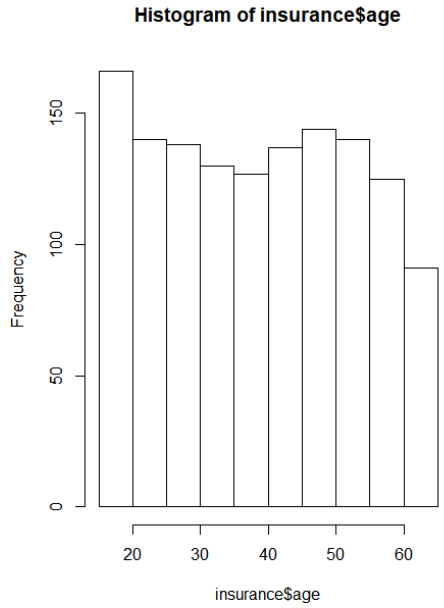
we would consider log transformed data because it looks more normal.

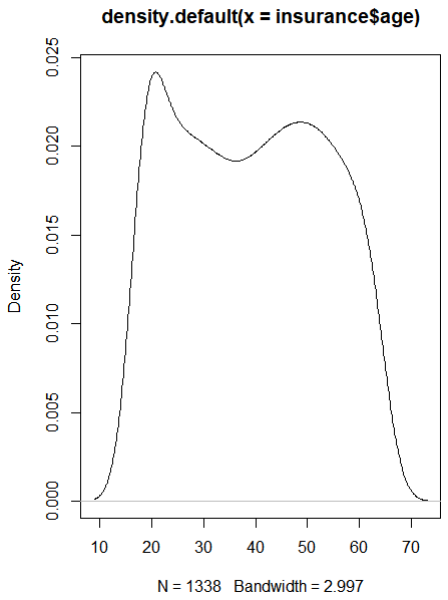
Age

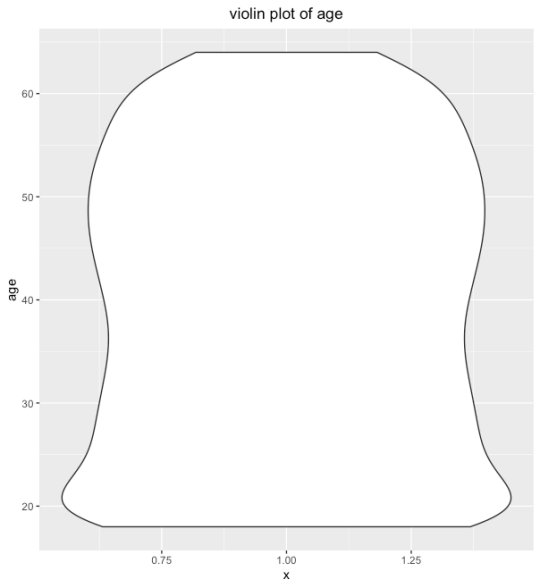
Summary

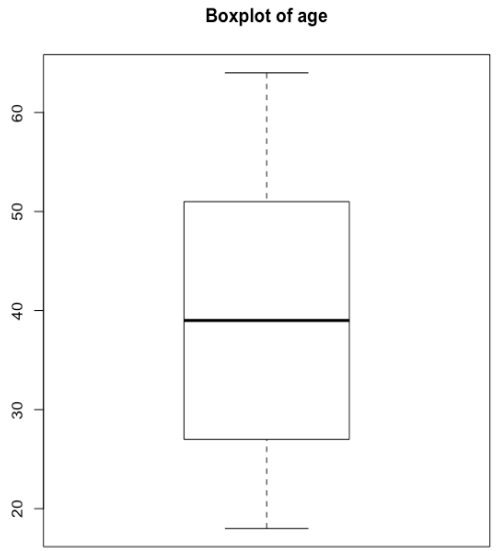
Min	1st Quarter	Median	Mean	3rd Quarter	Max
18.00	27.00	39.00	39.21	51.00	64.00

Plots





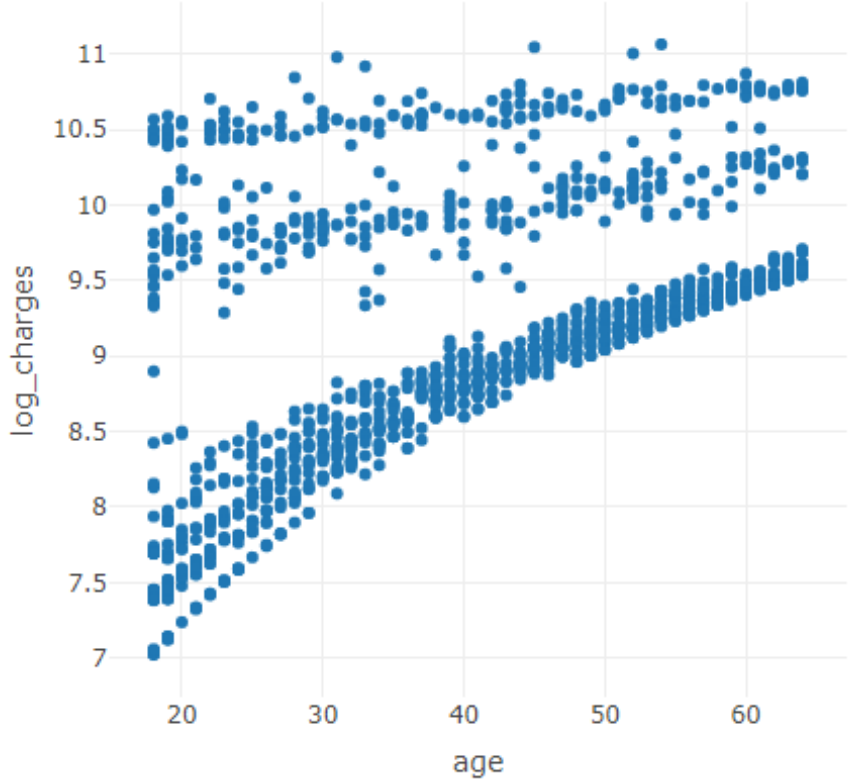




No extremely large or extremely small points. Age doesn‘t have any outlier.

Relationship between age and log charges

Scatter Plot



Correlation

	Age	Log charges
Age	1	0.53
Log charges	0.53	1

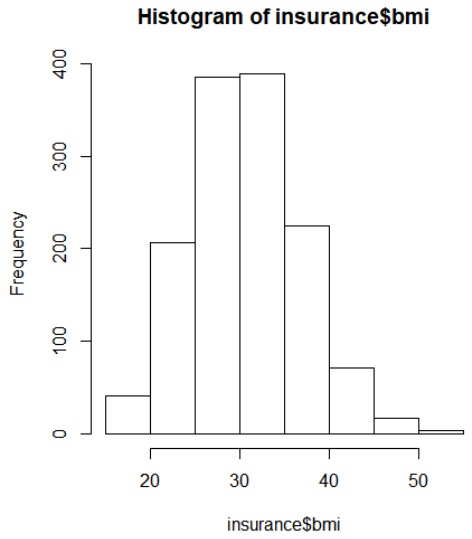
Age and log-charges might have a relationship. Older people spend more.

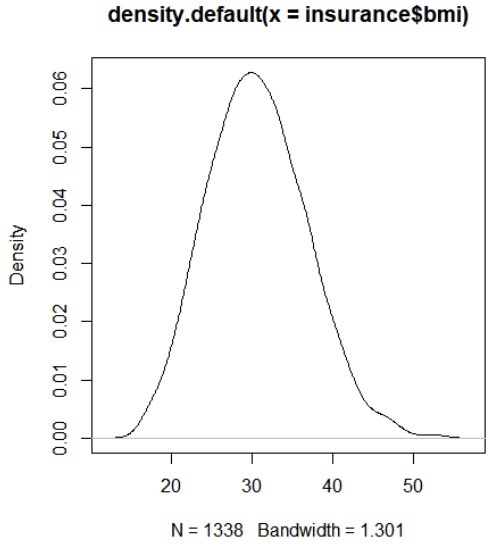
BMI

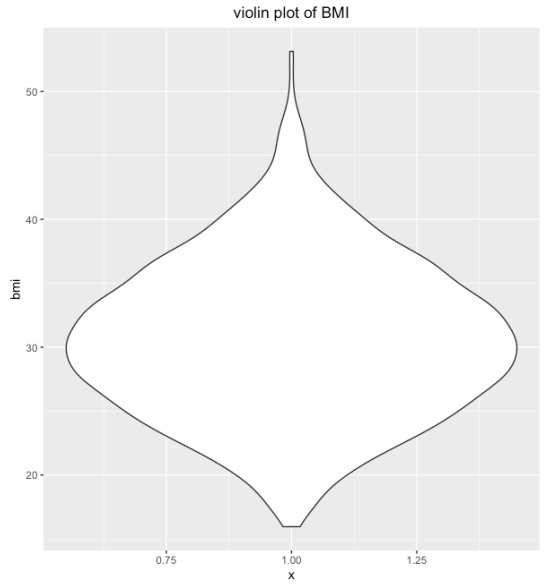
Summary

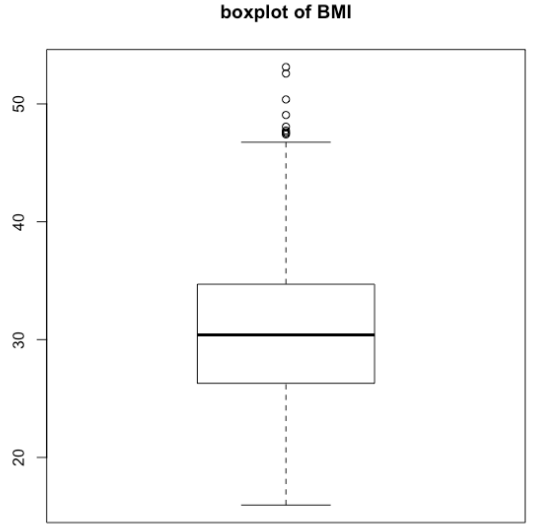
Min	1st Quarter	Median	Mean	3rd Quarter	Max
15.96	26.30	30.40	30.66	34.69	53.13

Plots





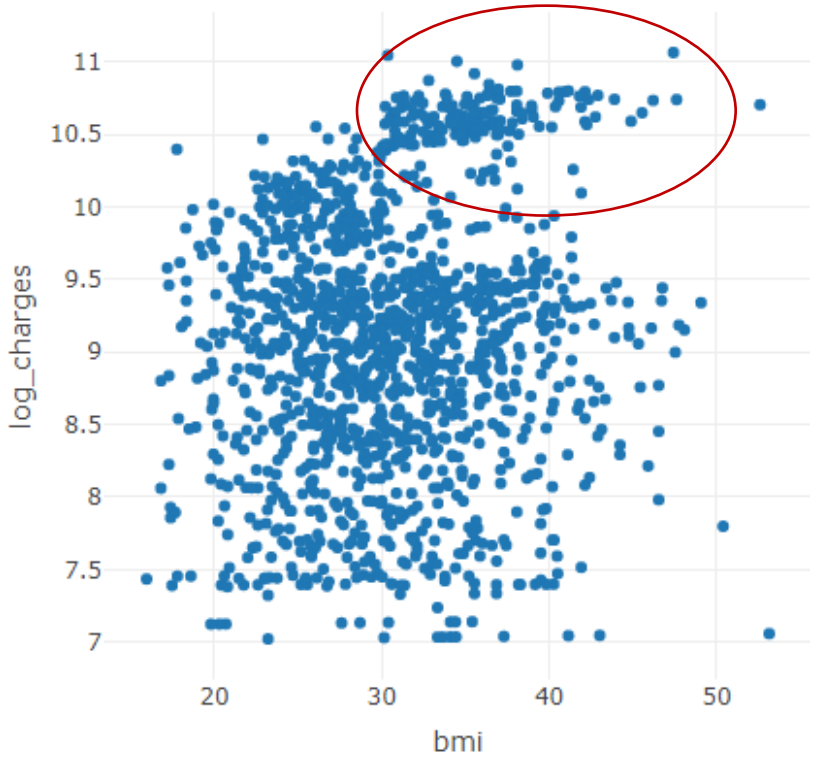




BMI has some extremely large points, but are biological possible, so they could be useful.

Relationship between BMI and log charges

Scatter Plot



Correlation

	BMI	Log charges
BMI	1	0.13
Log charges	0.13	1

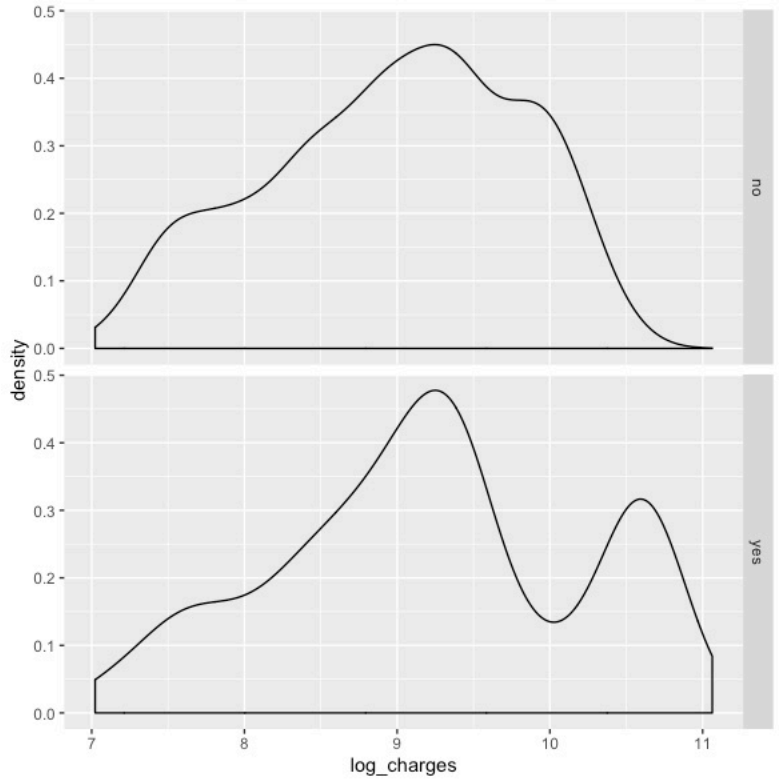
BMI and log-charges don’t seem like to have obvious relationship, but when BMI are over 30 or so, there are some much higher values.

Transformed BMI: Obesity

Here we try to divide BMI into two groups:

- Non-obesity (BMI < 30)
- Obesity (BMI ≥ 30)

Density plots of log charges in two groups



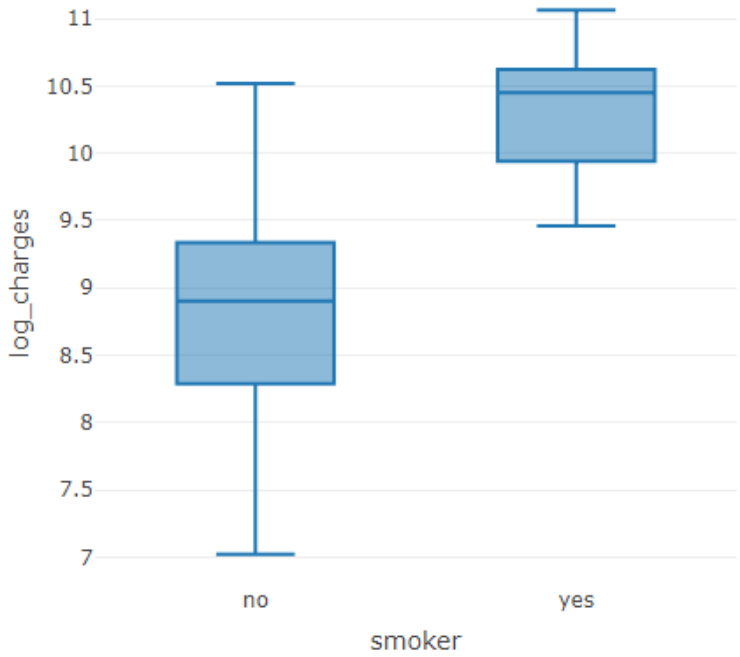
Obesity may influence log-charges

Smokers

Summary

NO	YES
1064	274

Boxplots of log charges in two groups



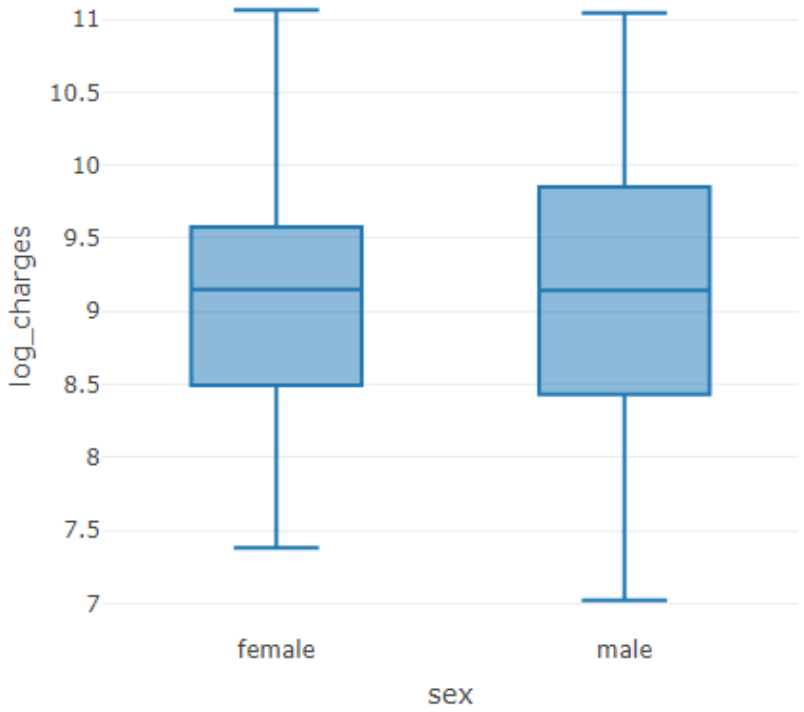
Smoking might has an influence on log-charges. Smokers has higher medical charges than non-smokers.

Sex

Summary

FEMALE	MALE
662	676

Boxplots of log charges in two groups



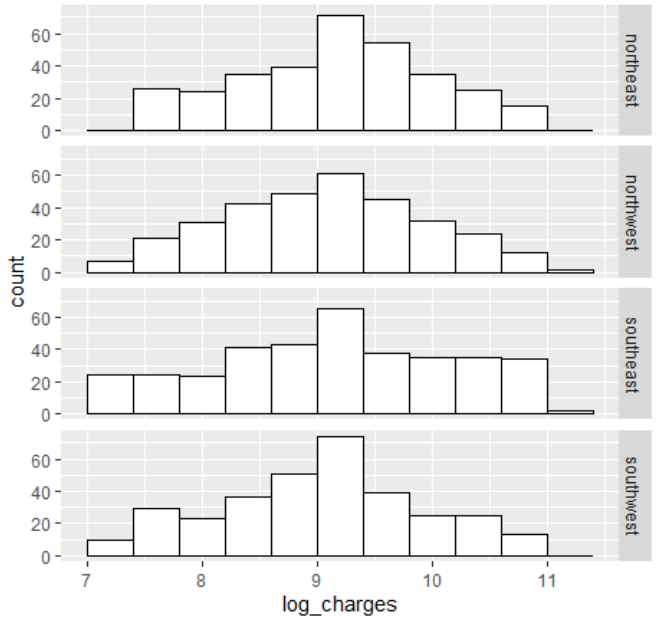
Sex does not seem to have an influence on log-charges.

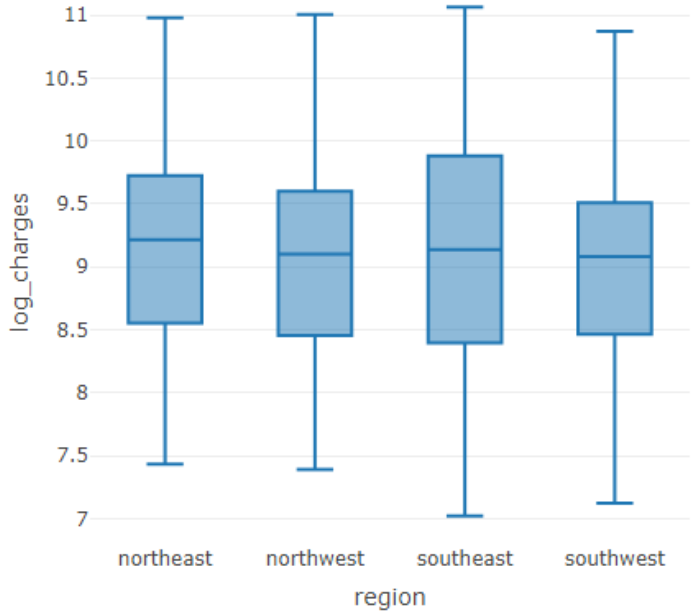
Region

Summary

northeast	northwest	southeast	southwest
324	325	364	325

Histograms & Boxplots of log charges in four groups





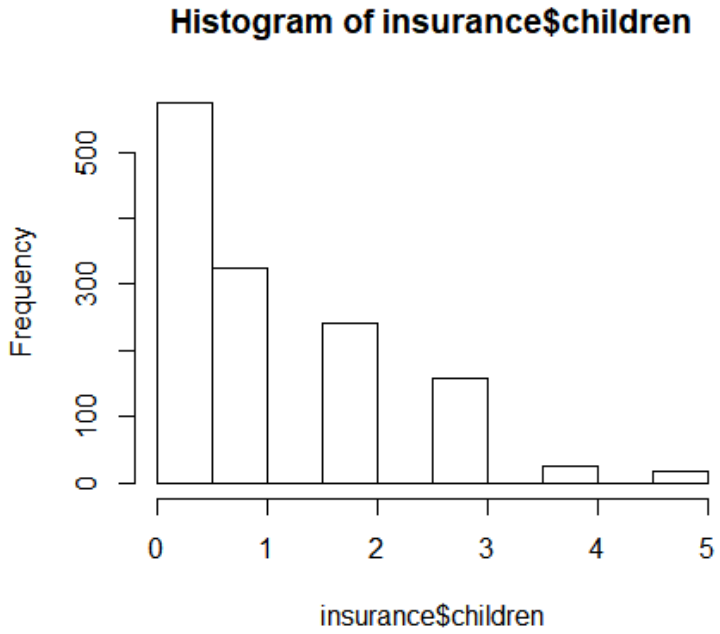
Region does not seem to have an influence on log-charges.

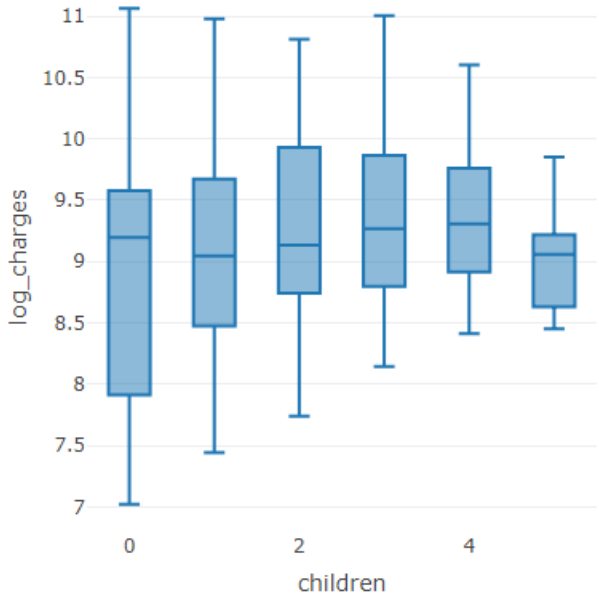
Children

Summary

0	1	2	3	4	5
574	324	240	157	25	18

Plots





The number of people with 4 or 5 children are very small, so it hard to see if there's any relationship between children and log charges; Plus, we didn't find any relationship between.

Conclusion of EDA

Key variables

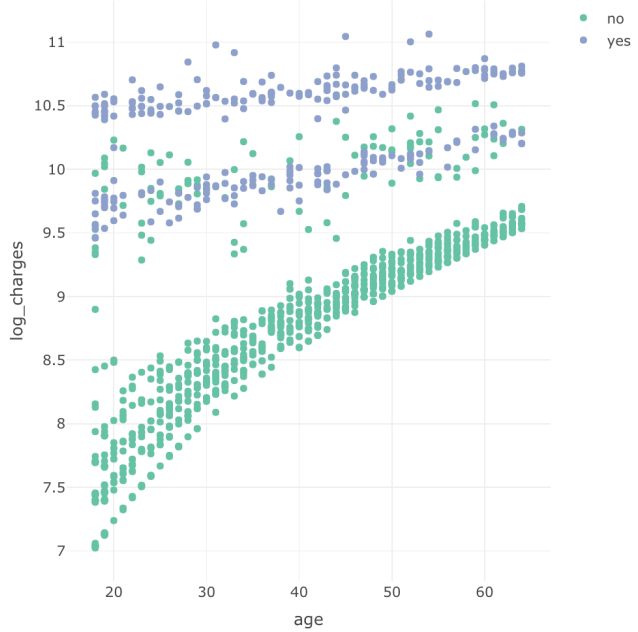
- Age
- Smoking
- Obesity (Transformed BMI)

Reasons

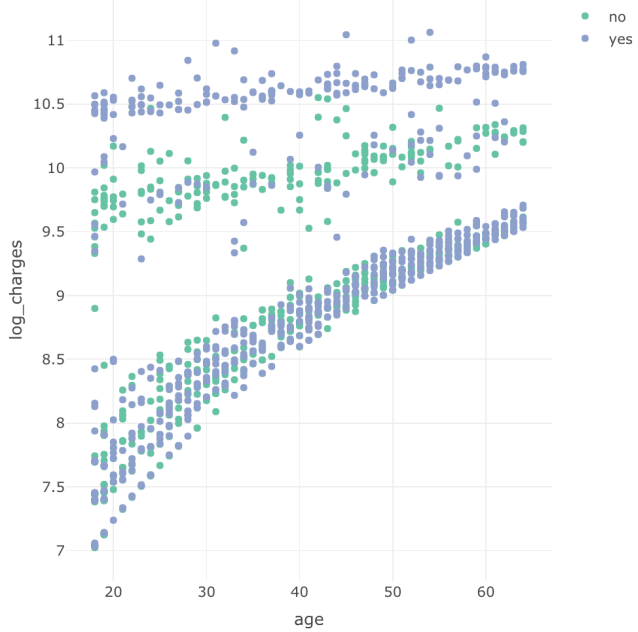
- They obey common sense in life.
- They most likely to have influence on log-charges.

Confounding variables

age vs log-charges (color-coding by smoker)



age vs log-charges (color-coding by obesity)



	Non-obesity	Obesity
Non-Smoker	502	562
Smoker	129	145

Key variables seem don‘t have relationship with each other, so we won’t have to worry about confounding variables.

PART THREE

Hypothesis Testing

Hypothesis Testing

Smokers

H_0: there is no difference in log-charges of smokers and non-smokers		
P-value	$<2.2*10^{-16}$	<0.05, significant at 95% level
95% Confidence interval	(1.45, 1.58)	doesn't contain 0
Mean log-charges of smokers	10.30	Mean log-charges of smokers are lager
Mean log-charges of non-smokers	8.79	

We can conclude the log-charges are significantly different between smokers and non-smokers at 95% level, smokers are more likely to spend more.

Hypothesis Testing

Obesity

H_0: there is no difference in log-charges of Obesity and non-obesity		
P-value	$6.91*10^{-6}$	<0.05, significant at 95% level
95% Confidence interval	(0.13, 0.32)	doesn't contain 0
Mean log-charges of obesity	9.20	Mean log-charges of obesity are lager
Mean log-charges of non-obesity	8.98	

Age

H_0: there is no difference in log-charges of older (age≥median:39) and younger (age<39)		
P-value	$<2.2*10^{-16}$	<0.05, significant at 95% level
95% Confidence interval	(0.75, 0.93)	doesn't contain 0
Mean log-charges of older	9.51	Mean log-charges of smokers are lager
Mean log-charges of younger	8.67	

The t-tests further shows all 3 key variables have influence on the log-charges at 95% level. Smokers, older people, obesity ones tend to have more medical charges.

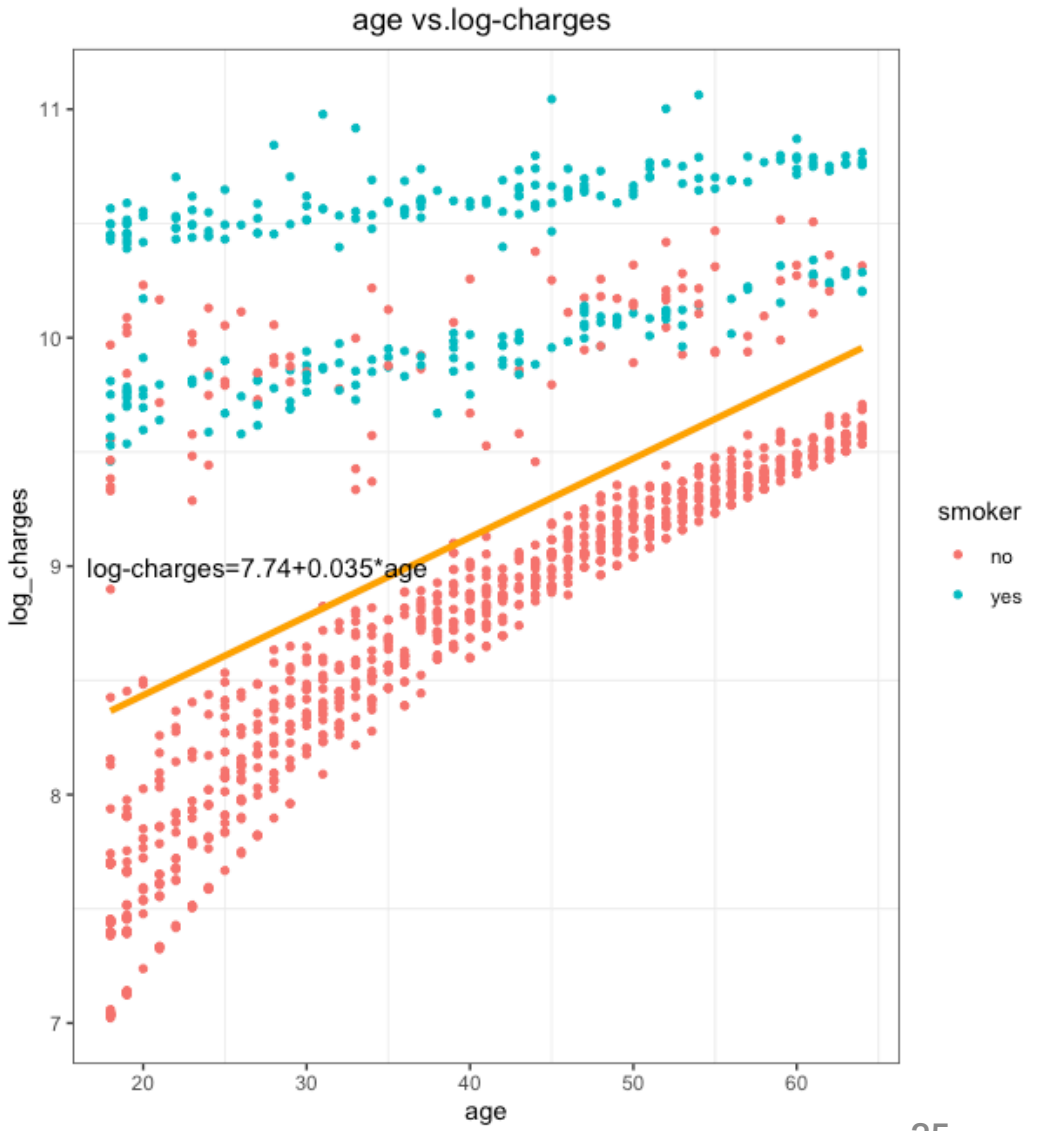
PART FOUR

Linear Regression

Simple Linear Regression - Age

	β_0	β_{age}
Estimate	7.74	0.035
P-value	$<2.2*10^{-16}$	$<2.2*10^{-16}$
Equation	log(charges)=7.74+0.035*age	

- P-value is significant at 95% level, so we can conclude there is association between age and log charges at this level
- β_{age} : Average $\log_e$ charges to go up by 0.04 ($\log_e$ dollars) when increasing age by 1 year
- β_0 : Average $\log_e$ charges to be 7.74 for a person aged 0 year



Multivariable Linear Regression

	β_0(Intercept)	X_{age}	$X_{obesity}$ (baseline is non obesity)	X_{smoker} (baseline is non smoker)
Estimate β_i	7.33	0.04	0.14	1.55
P-value	$<2.2*10^{-16}$	$<2.2*10^{-16}$	$<8.77*10^{-8}$	$<2.2*10^{-16}$
Equation	log(charges)=7.33+0.04*age+0.14*obesity+1.55*smoker			

All p-values are significant and all three independent variables has association with log-charges when adjusting for other variables at 95% level

We estimate the average $\log_e$ charges to:

- β_{age}: go up by 0.04 ($\log_e$ dollars) when increasing age by 1 year and keeping obesity and smoker constant
- $\beta_{obesity}$: go up by 0.14 ($\log_e$ dollars) when comparing obesity to non-obesity and keeping age and smoker constant
- β_{smoker}: go up by 1.55 ($\log_e$ dollars) when comparing smoker to non-smoker and keeping age and obesity constant
- β_0: be 7.33 for a non-obesity, non-smoker person aged 0 year

PART FIVE

Conclusion

Conclusion

01

What affected the cost?

age, obesity, smoking

02

How the cost affected?

- smokers, older people, obesity ones tend to have more medical charges
- the rate of change is the corresponding β in our multivariable regression model

03

Prediction

multivariable regression model
log(charges)=7.33+
0.04*age+0.14*obesity
+1.55*smoker

References

[1] Altman, Naomi, and Martin Krzywinski. "Points of significance: Regression diagnostics." (2016): 385.

[2] Lantz, Brett. "Machine Learning with R. Packt Publishing." Briminghan, UK (2015).

[3] Mendenhall, William, Terry Sincich, and Nancy S. Boudreau. A second course in statistics: regression analysis. Vol. 5. Upper Saddle River, NJ: Prentice Hall, 1996.

[4] Altman, Naomi, and Martin Krzywinski. "Points of significance: simple linear regression." (2015): 999.

Q&A

THANK YOU

[10 / 01 / 2019]