HEALTH INSURANCE IN US

MANOJ KUMAR – 2048015

CAC2

Domain – Finance, Healthcare, and Insurance
Method – Regression Analysis

PROJECT DESCRIPTION

Health insurance is a means for financing a person's or persons' health care expenses. Most people have private health insurance in the US, usually obtained through a current employer, and the minority are covered through government-sponsored programs.

The insurer calculates premiums for their insurance policies relying on two primary factors the cost the insurer predicts to pay under their policies and the cost of operating particular policies or plans. The cost of medical expenses is calculated in many ways - the policyholder's health status, region of residence, employment status, and wages can all be included in the estimate.

Aims,

- To determine if there is a relationship between attributes and medical costs.
- To determine if there a significant difference in medical costs between different groups.
- To fit a multiple linear regression to predict costs.

REGRESSION ANALYSIS

In terms of statistical methods, regression analysis is often used to estimate healthcare costs and calculate insurance premiums. A policyholder's medical expenses can be influenced by many factors, such as their habits, chronic illnesses, age, economic factors, occupational hazards, place of residence, and so on. Regression analysis can be used to identify the factors that are significant in their influence on medical costs. Regression analysis can also predict the actual cost of an insurance policy, allowing for insurance companies to set competitive prices. Selecting the same price for all policyholders is not a competitive strategy as those with low expenses would overpay and possibly leave the service. Those with high costs would remain using the service and make a loss of the insurance company. Regression models are tools that can use to establish proper classification systems that offer a fair price to customers and maximize the company's profits.

DATASET

The dataset for this report comes from the Kaggle data science community and is in the public domain. The dataset includes information about the insurance policyholder, their dependents, and their medical expenses throughout a year.

- **Age**: Age of primary policyholder.
- **Sex**: Sex of the policy policyholder.
- **BMI**: Body Mass Index of policyholder, defined as the body mass divided by the square of the body height (kg/m2).
- **Smoker status**: Whether the policyholder is a smoker or a non-smoker.
- **Children**: Number of children/dependents covered in the policy.
- **Region of residence**: Residential areas of the policyholder (in the US) - North East, South East, South West, North West.
- **Charges**: Yearly medical expenses billed by the medical insurance provider ($).

Import and pre-processing

```
suppressMessages(library(tidyverse))        # Data manipulation and plots
suppressMessages(library(funModeling))      # Overview stats

library(magrittr)                           # To use pipes
library(skimr)             # To get a quick summary table
library(caret)             # To create the partition for train/test datasets

options(scipen = 999)      # Turn off scientific notation for numbers
options(repr.plot.width=12,
        repr.plot.height=8)# Set universal plot size
```

```
# Reading the dataset
df          <- read.csv('insurance.csv')

# Denote factor variables
df$sex      <- factor(df$sex)
df$smoker   <- factor(df$smoker)
df$region   <- factor(df$region)
df$children <- factor(df$children)
```

```
# check for missing values
df %>%
    is.na() %>%
    sum()

## [1] 0
```

```
# check data types
df %>%
    str()
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: Factor w/ 6 levels "0","1","2","3",..: 1 2 4 1 1 1 2 4 3 1 ..
.
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3
2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
skim(df)
```

Data summary

Name	df
Number of rows	1338
Number of columns	7
_______________________	
Column type frequency:	
factor	4
numeric	3
________________________	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
sex	0	1	FALSE	2	mal: 676, fem: 662
children	0	1	FALSE	6	0: 574, 1: 324, 2: 240, 3: 157
smoker	0	1	FALSE	2	no: 1064, yes: 274
region	0	1	FALSE	4	sou: 364, nor: 325, sou: 325, nor: 324

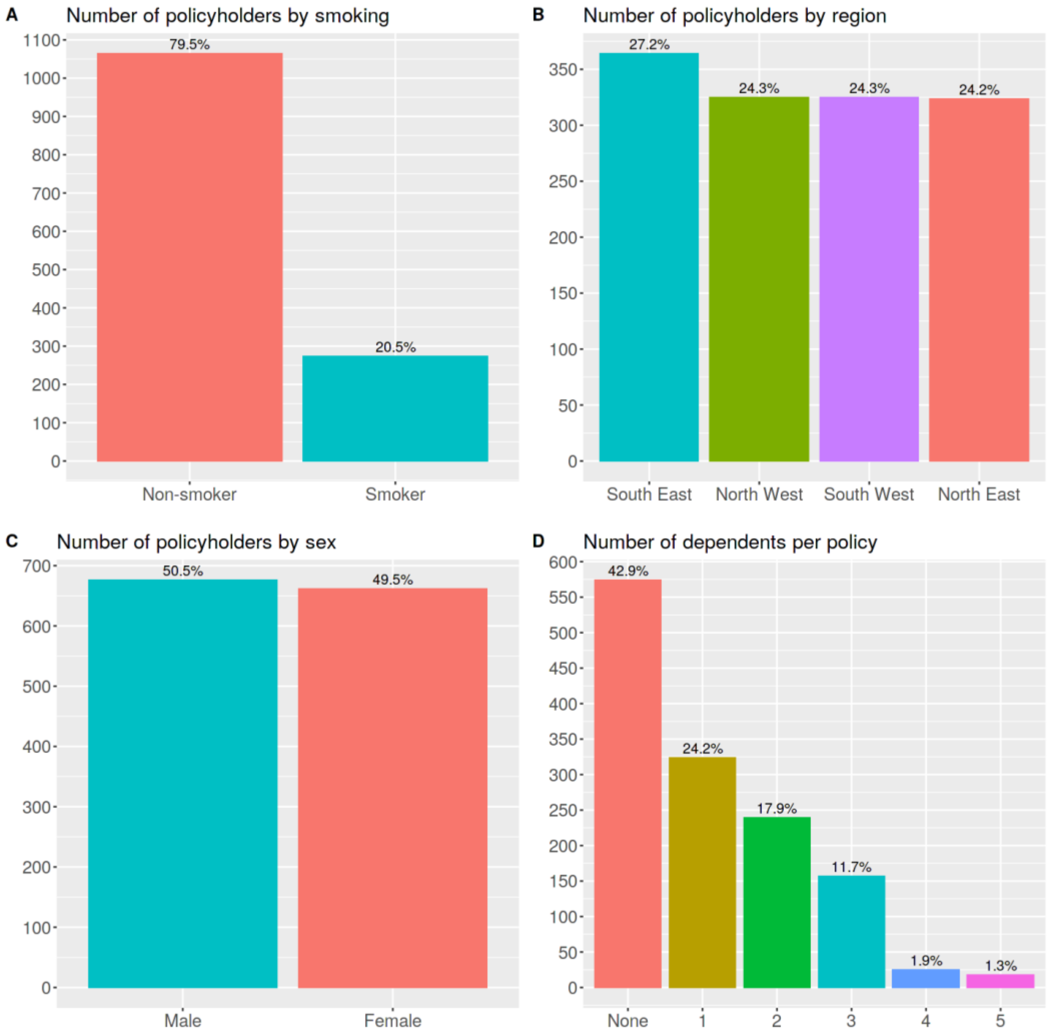
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	39.21	14.05	18.00	27.00	39.00	51.00	64.00	▇▅▅▆▆
bmi	0	1	30.66	6.10	15.96	26.30	30.40	34.69	53.13	▂▇▇▂▁
charges	0	1	13270.42	12110.01	1121.87	4740.29	9382.03	16639.91	63770.43	▇▂▁▁▁

EXPLORATORY DATA ANALYSIS

1. Distributions of Categorical variables

```
# Plot grid
cowplot::plot_grid( smoker, region, sex, children, labels="AUTO", ncol = 2, n
row = 2 )
```

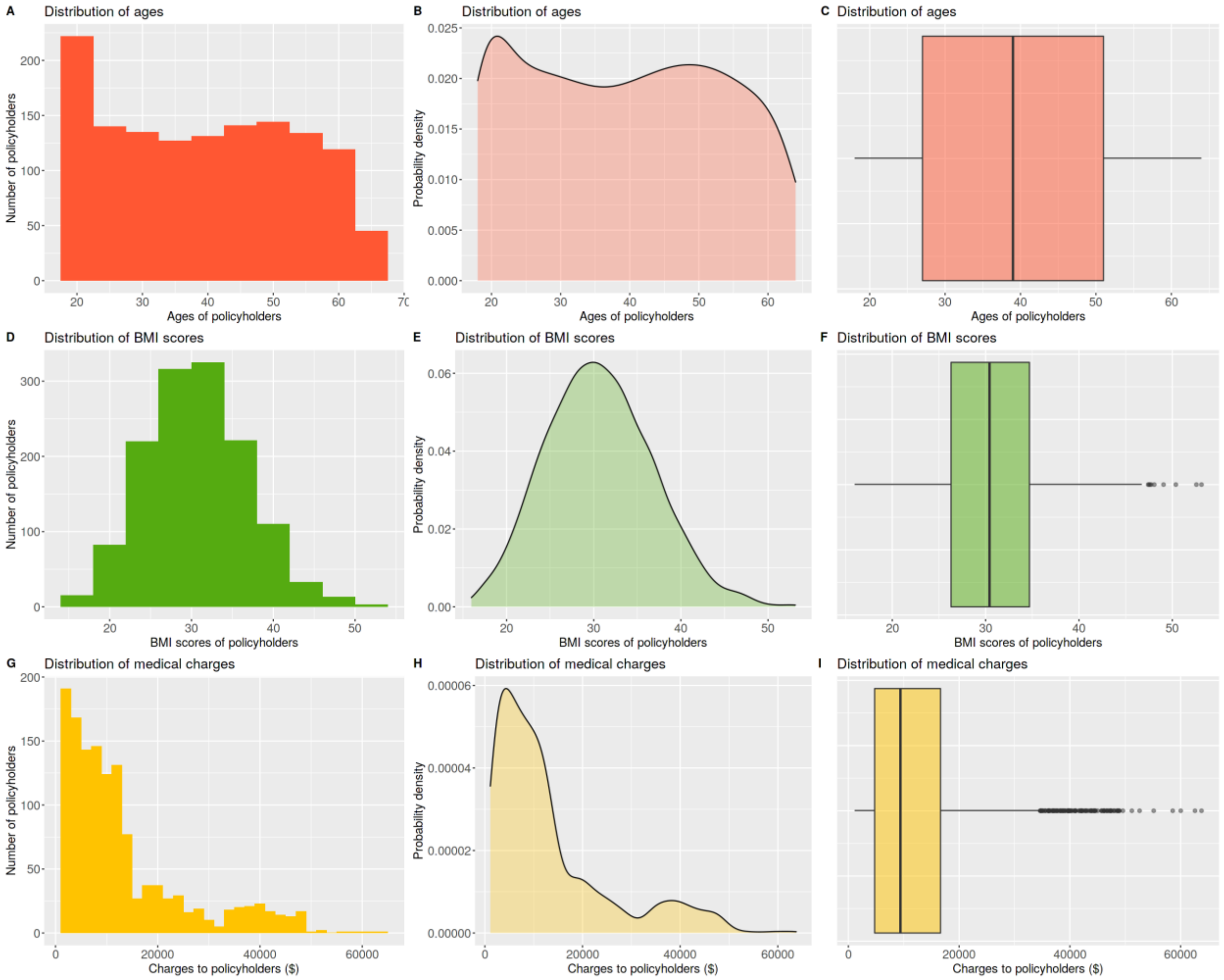


Insights

- **Smoking status**: There are many more non-smokers (80%) than smokers (20%).
- **Region of residence**: Policyholders are evenly distributed across areas, with South East being the most populous one (27%), with the rest of the regions containing around 24% of policyholders each.
- **Sex**: There are slightly more men (51%) than women (49%) in the sample.
- **Dependents**: Most policyholders (43%) do not have dependents covered in their policy. For those who have dependents covered in their policy, most have one dependent (24%). The maximum number of dependents covered is five (1%).

2. Distributions of Numerical variables

```
figsize <- options(repr.plot.width=20, repr.plot.height=16)
# Age, BMI, Charges distribution
cowplot::plot_grid( age_hist, age_dens, age_box, bmi_hist, bmi_dens, bmi_box,
    charges_hist, charges_dens, charges_box,
labels="AUTO", ncol = 3, nrow = 3 )
```

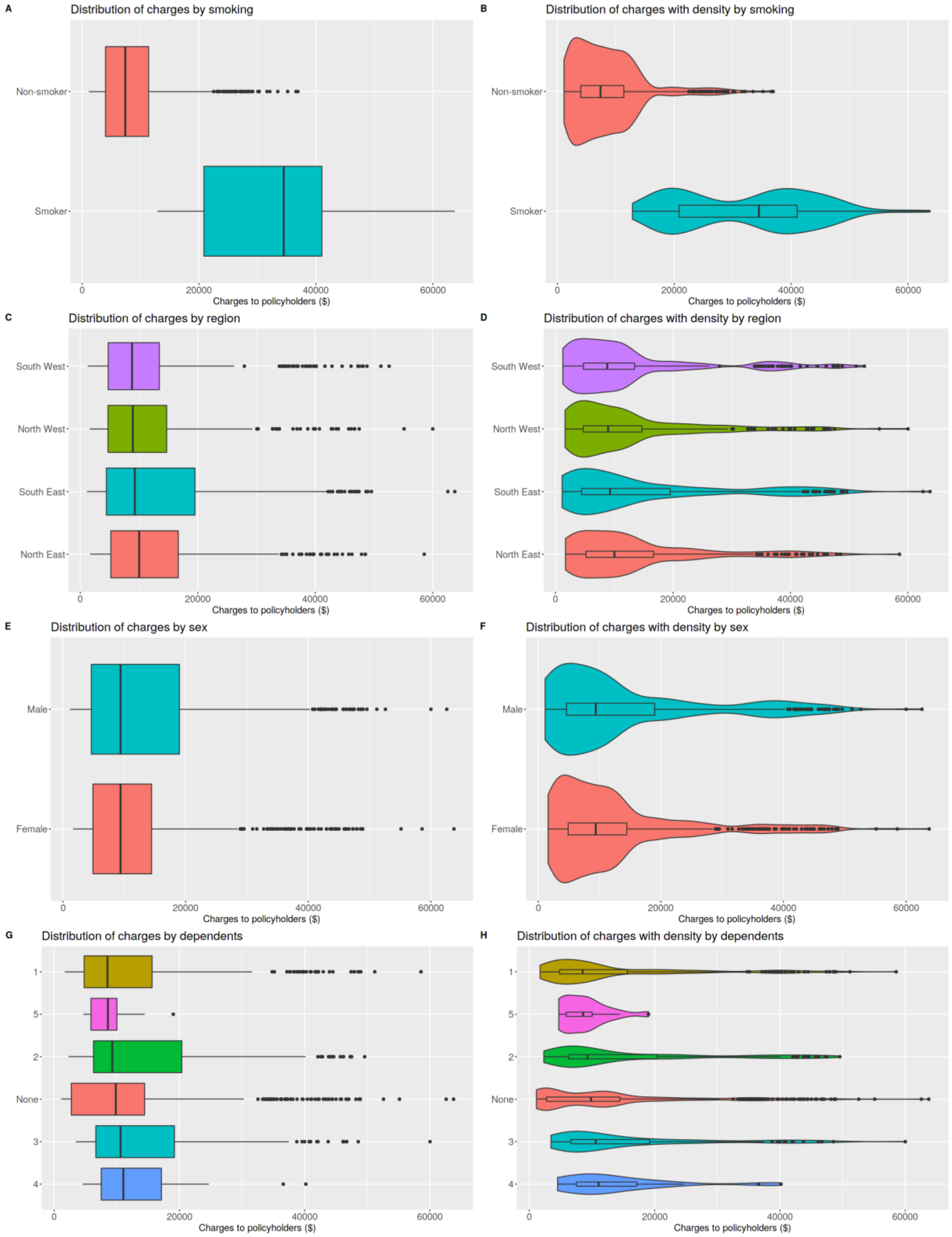


Insights

- **Age**: Youngest policyholder is 18, and the eldest is 64. All ages in the range are represented reasonably equally apart from the youngest and eldest policyholders. 18-23-year-olds are the most populous (among all 5-year segments), and 60-64-year-olds are the least represented 5-year age group. There are no outliers.
- **BMI**: BMI is normally distributed, with the smallest and the most significant values being the least common and median and almost identical. There are a few outliers on the larger side. The minimum recorded BMI score is 16, and the maximum is 53.1.
- **Charges**: Charges are heavily right-skewed, with many outliers on the larger side. This means most charges are fairly low, with a few particularly high charges. Smallest charge is $1,122 and largest charge is $63,770.

3. Charges vs all other factors

```
# make a grid
```

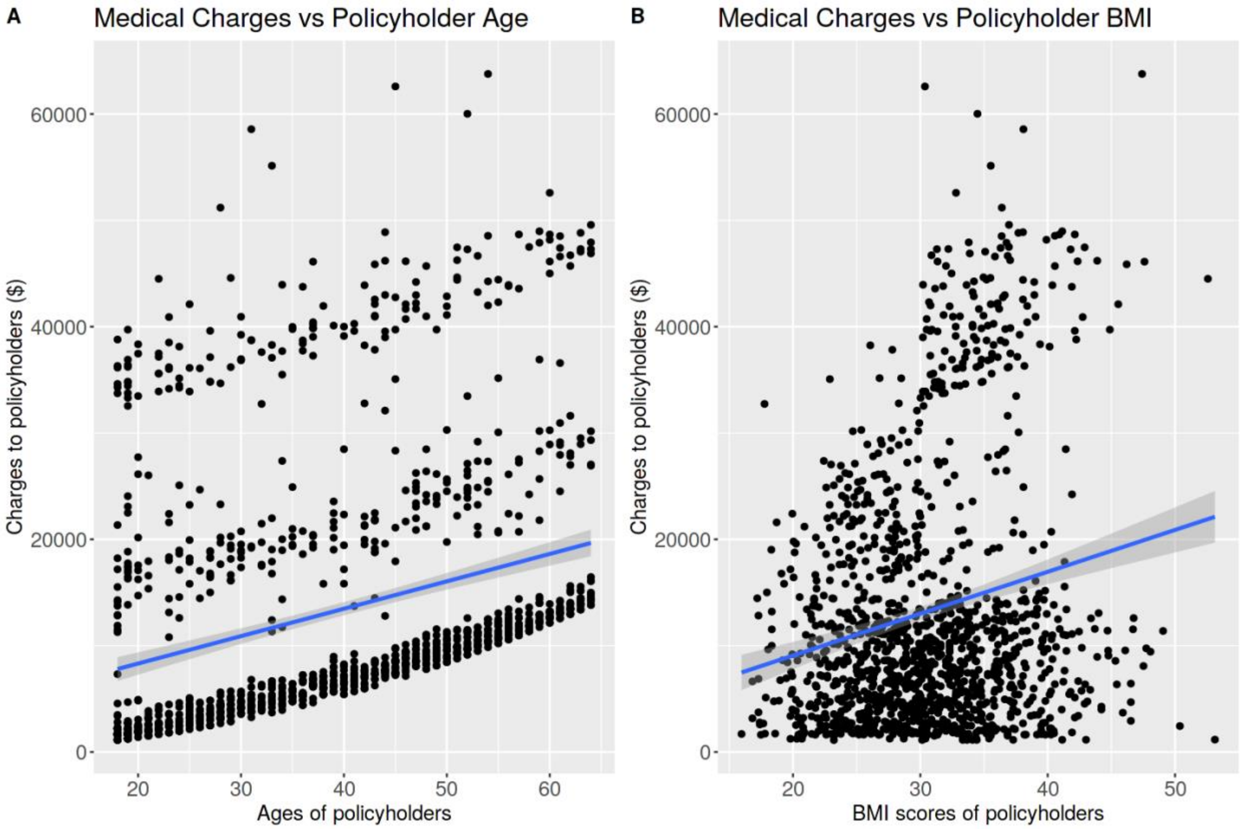


Insights

- **Smoking**: There is a big difference in medians between smokers ($34,456) and non-smokers (\$7,345). Non-smokers show many outliers on the larger side, while the vast majority of charges are on the smaller side. Smokers show the bimodal distribution and no outliers.
- **Region of residence**: There are slight differences in medians between all groups. All groups have outliers. The spread of values is similar for all groups apart from South East which has a more extensive interquartile range (IQR).
- **Sex**: Males have a marginally more significant median ($9,413) than females (\$9,370), a difference of just $43. Both groups show outliers on the larger side. The spread of values is reasonably similar.
- **Dependents**: There are some differences in medians between the groups, but they are not drastic.

4. Medical Charges vs Policyholder Age/BMI

```
# make a grid
```



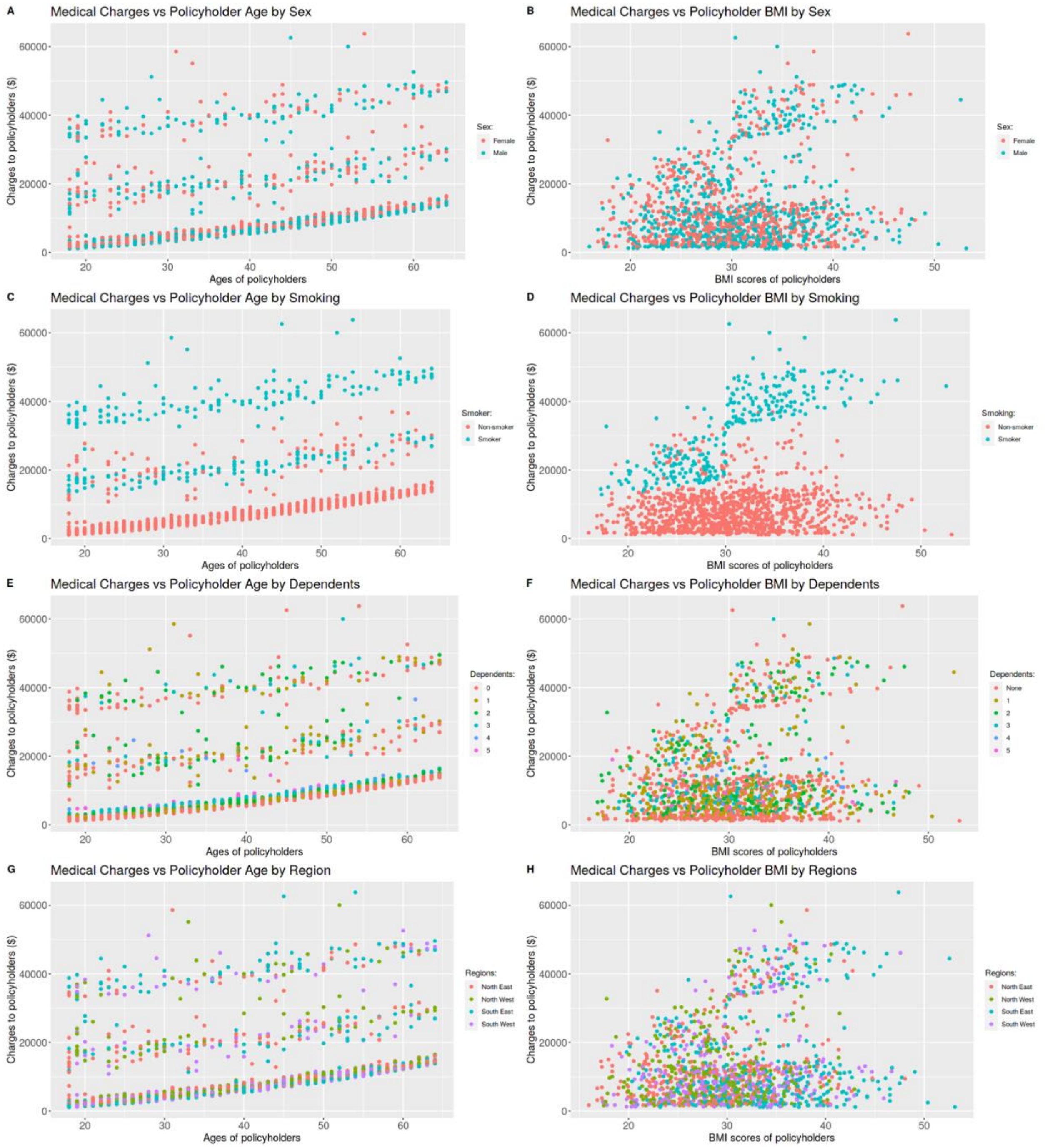
Insights

- **Medical Charges vs Policyholder Age**: There are interesting patterns here, showing three groups. The lower band shows a strong relationship between medical charges and age, with two other bands showing a less intense relationship. The general trend is a positive correlation, meaning as age increases, so do medical expenses.

- **Medical charges vs Policyholder BMI**: There is a positive relationship between BMI and medical expenses, meaning people with higher BMI scores have higher medical bills. The relationship is not, however, decisive. There are possibly two groups to the scatter plot, judging by the spread of points above and below the regression line.

5. Medical Charges vs Different factors

```
# make a grid
```



Insights

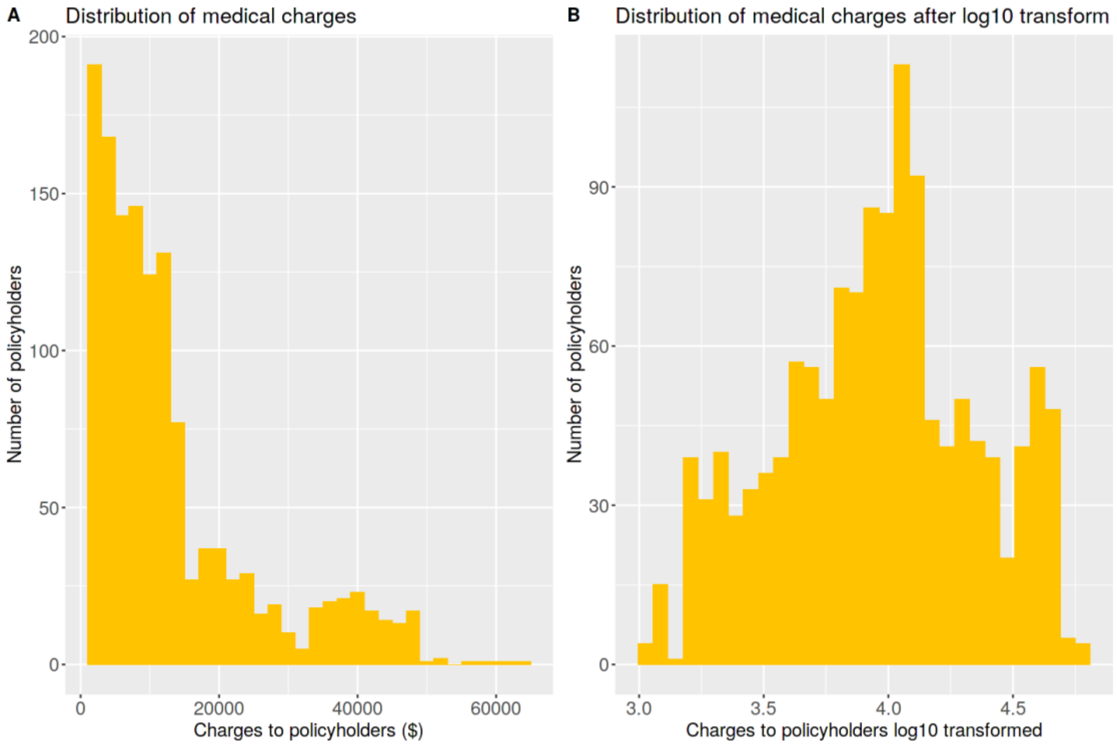
- A very clear pattern emerges looking at plots C and D. There are clusters in these two scatter plots, according to the smoking status of the policyholder. Other factors don't show such clear groupings in this visual analysis.

MODELLING AND ACCURACY

Multiple linear regression

- Multiple linear regression (MLR) models allow for effective summarization of multivariate datasets. It is an extension of the single linear regression in which instead of one independent variable, multiple independent variables are used to predict the value of the response variable.
- The response variable (charges) is to be transformed to reduce skewness and meet normality for the MLR model.
- The dataset is split into a training dataset (80% of all data) and a testing dataset (20% of all data).

```
cowplot::plot_grid( charges_hist, charges_hist_log10, labels="AUTO", ncol = 2
, nrow = 1 )
```



The hypotheses for this model are such:

- **H0:** there will be no significant prediction of medical expenses by the policyholder's smoking status, BMI score, age, region of residence, sex, and number of dependents covered by the policy.
- **H1:** there will be significant prediction based on the above mentioned factors.

Split the dataset and train the model

```
set.seed(122)            # Set the seed to make the partition reproducible
training.samples <- df$logCharges %>%
  createDataPartition(p = 0.8, list = FALSE)

train  <- df[training.samples, ]
test <- df[-training.samples, ]
```

Importing required packages

```
library(ridge)
library(e1071)
library(ggplot2)
library(rpart)
library(rpart.plot)
```

```
# Fitting Multiple Linear Regression to the Training set
formula <- as.formula("logCharges ~ smoker + bmi + age + children + sex + reg
ion")
model <- lm(formula, data = train)

summary(model)
## Call:
## lm(formula = formula, data = train)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40628 -0.09013 -0.02321  0.03314  0.93626
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)      3.0308795  0.0342260  88.555 < 0.0000000000000002 ***
## smokeryes        0.6760329  0.0144515  46.779 < 0.0000000000000002 ***
## bmi              0.0058070  0.0009931   5.848  0.00000000663898698 ***
## age              0.0153611  0.0004142  37.090 < 0.0000000000000002 ***
## children1        0.0538452  0.0146927   3.665             0.000260 ***
## children2        0.1286999  0.0161328   7.978  0.00000000000000385 ***
## children3        0.1086741  0.0189414   5.737  0.00000001254227630 ***
## children4        0.2109837  0.0411729   5.124  0.00000035470555674 ***
## children5        0.1835554  0.0552900   3.320             0.000931 ***
## sexmale         -0.0304837  0.0115905  -2.630             0.008661 **
## regionnorthwest -0.0305449  0.0164321  -1.859             0.063325 .
## regionsoutheast -0.0599089  0.0168307  -3.559             0.000388 ***
## regionsouthwest -0.0562769  0.0165515  -3.400             0.000699 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1884 on 1059 degrees of freedom
## Multiple R-squared:  0.7789, Adjusted R-squared:  0.7764
## F-statistic: 310.9 on 12 and 1059 DF,  p-value: < 0.00000000000000022
```

Interpretation

- A significant regression equation was found (F(12,1057) = 303.9, $p < 0.001$), with an adjusted R-squared of 0.7789. In other words, the model explains 77.9% of total variance in the sample. Null hypothesis is rejected.

EVALUATING THE MODEL

```
# Make predictions on the training dataset
predictions <- model %>% predict(train)
# Calculating the residuals
residuals <- train$logCharges - predictions
# Calculating Root Mean Squared Error
rmse <- sqrt(mean(residuals^2))

rmse %>%
    round(digits=3)
## [1] 0.187
```

```
predictions <- model %>% predict(test)
residuals <- test$logCharges - predictions
rmse <- sqrt(mean(residuals^2))

rmse %>%
    round(digits=3)
## [1] 0.208
```

```
# Calculating RMSE for training data with backtransformed data
predictions <- model %>% predict(train)
residuals <- 10^train$logCharges - 10^predictions # backtransform measured an
d predicted values
rmse <- sqrt(mean(residuals^2))
round(rmse)
## [1] 8334
```

```
# Calculating RMSE for testing data with backtransformed data
predictions <- model %>% predict(test)
residuals <- 10^test$logCharges - 10^predictions # backtransform measured and
predicted values
rmse <- sqrt(mean(residuals^2))
round(rmse)

## [1] 9000
```

To measure robustness of the model, an absolute measure of fit - RMSE was calculated, RMSE (test set) = 0.208, RMSE (training set) = 0.187. This is an indicator that the model is not overfitting. However, further investigation is needed to confirm this. After back transforming the residuals, the RMSE for the test set was $9000, meaning the model's predictions are usually off by this amount.

CONCLUSIONS

Smoking having the strongest effect on medical expenses is reasonably expected. Increases in the BMI score lead to relatively small expense increases. However, it is worth pointing out that standard BMI scores are not indicative of ill health. Only people in the underweight (BMI < 18.5), overweight (BMI 25.0 to 29.9), and obese (BMI ≥ 30) ranges would be expected to have poorer health outcomes.

The same should be said of the effect of aging - 22-year-olds would be expected to enjoy the same level of health as 18-year-olds despite being 4 years older. However, middle-aged and elderly people will most likely see a rapid decline in health year by year.

Medical expenses increasing with an increased number of dependents is to be expected. However, having three dependents covered by insurance seems cheaper than having two dependents, and five dependents see a lesser increase in charges than four. This may be explained by the uneven number of observations in each group. For example, no dependents group has 574 observations when five dependents group only has 18.

It is also interesting to note that even though the median difference of medical charges between men and women is only $43, the relationship between sex and medical charges was significant in the multiple linear regression model.

Lastly, whether the model is robust can only determine the acceptable cost of error. Being able to explain 77.9% of the total variance with an RMSE of $9000 may well be enough if the company can deal with the potential mispredictions.

REFERENCES

[1] Fulton, B. D. (2017). Health care market concentration trends in the United States: evidence and policy responses. Health Affairs, 36(9), 1530-1538.

[2] Ho, K. (2009). Insurer-provider networks in the medical care market. American Economic Review, 99(1), 393-430.

[3] Frees, E. W. (2009). Regression modeling with actuarial and financial applications. Cambridge University Press.

[4] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.