A Mini Project Report

on

Health Care Cost Prediction using Linear Regression Model

In Subject: **Probability and Statistics**

by

Karan Lodha	**271028**
Saiprasad Mane	**271033**
Rugwed Nand	**271036**
Saikumar Padmawar	**271048**

Department of Artificial Intelligence and Data Science

VIIT

2021-2022

Contents

Chapter-1 Introduction

1.1 Introduction

Medical expenses are one of the major recurring expenses in a human life. It's a common knowledge that one lifestyle and various physical parameters dictates diseases or ailments one can have and these ailments dictates medical expenses. According various studies, major factors that contribute to higher expenses in personal medical care include smoking, aging, BMI.

In this study, we aim to find a correlation between personal medical expenses and different factors, and compare them. Then we use the prominent attributes as predictors to predict medical expenses by creating linear regression models and comparing them using ANOVA.

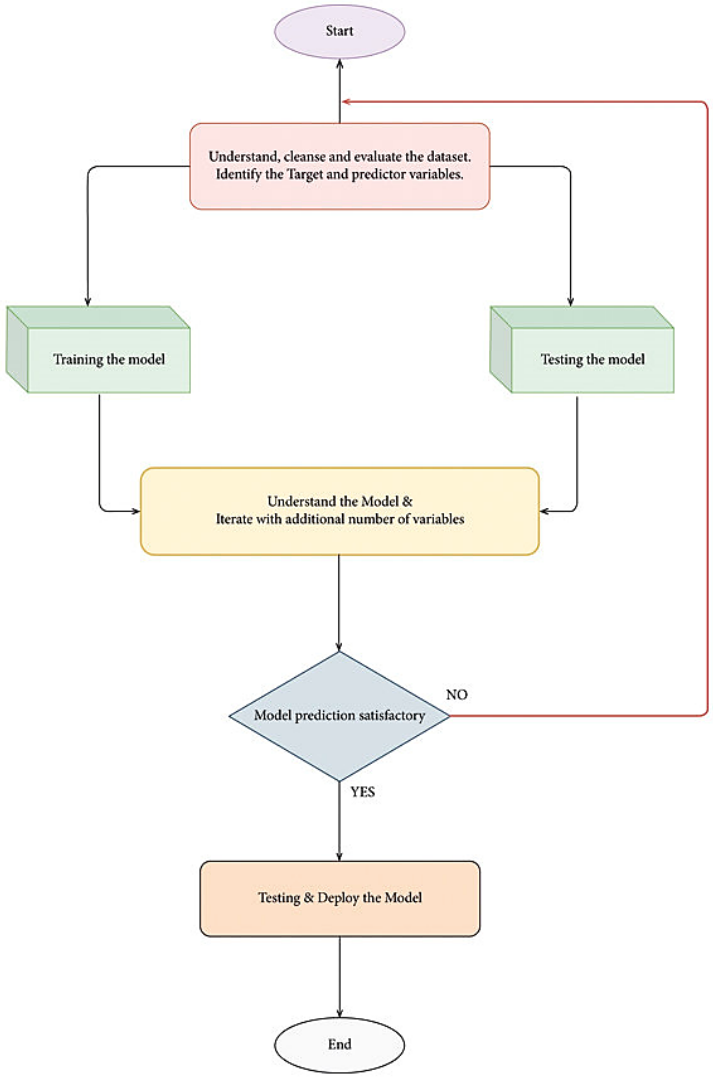
In research, we found that smoking, age and higher BMI have a high correlation with higher medical expenses indicating they are major factors in contributing to the charges and the regression can predict with more than 75% accuracy the charges.

1.2 Requirements

- Software: R studio
- Skills: Knowledge about the Linear Regression Model, R libraries, Statics, Probabilities.

1.3 Design & Problem Statement

Problem statement: To predict Health Care Cost using Linear Regression Model.



1.4 Proposed work

library(ggplot2)

library(dplyr)

library(Hmisc)

library(cowplot)

library(WVPlots)

set.seed(123)

##library(readr)

##insurance <- read_csv("~/SY/PAS/PBL/insurance.csv")

Data <- read.csv("~/SY/PAS/PBL/insurance.csv")

sample_n(Data, 5)

```{r describe, message=FALSE, warning=FALSE, paged.print=TRUE}

describe(Data)

```

No missing values at this point in the dataset.

Exploratory Data Analysis

```{r EDA, message=FALSE, warning=FALSE, paged.print=TRUE}

x <- ggplot(Data, aes(age, charges)) +

geom_jitter(color = "blue", alpha = 0.5) +

theme_light()

y <- ggplot(Data, aes(bmi, charges)) +

geom_jitter(color = "green", alpha = 0.5) +

theme_light()
```

```
p <- plot_grid(x, y)

title <- ggdraw() + draw_label("1. Correlation between Charges and Age / BMI",
fontface='bold')

plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))


x <- ggplot(Data, aes(sex, charges)) +

geom_jitter(aes(color = sex), alpha = 0.7) +

theme_light()


y <- ggplot(Data, aes(children, charges)) +

geom_jitter(aes(color = children), alpha = 0.7) +

theme_light()


p <- plot_grid(x, y)

title <- ggdraw() + draw_label("2. Correlation between Charges and Sex / Children covered
by insurance", fontface='bold')

plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))


x <- ggplot(Data, aes(smoker, charges)) +

geom_jitter(aes(color = smoker), alpha = 0.7) +

theme_light()


y <- ggplot(Data, aes(region, charges)) +

geom_jitter(aes(color = region), alpha = 0.7) +

theme_light()


p <- plot_grid(x, y)
```

title <- ggdraw() + draw_label("3. Correlation between Charges and Smoker / Region", fontface='bold')

plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))

```

* **Plot 1**: As Age and BMI go up Charges for health insurance also trends up.

* **Plot 2**: No obvious connection between Charges and Age. Charges for insurance with 4-5 chilren covered seems to go down (doesn't make sense, does it?)

* **Plot 3**: Charges for Smokers are higher for non-smokers (no surprise here). No obvious connection between Charges and Region.

## Linear Regression Model

### Preparation and splitting the data

```{r prep, message=FALSE, warning=FALSE, paged.print=TRUE}

n_train <- round(0.8 * nrow(Data))

train_indices <- sample(1:nrow(Data), n_train)

Data_train <- Data[train_indices, ]

Data_test <- Data[-train_indices, ]


formula_0 <- as.formula("charges ~ age + sex + bmi + children + smoker + region")

```

Train and Test the Model

```{r model_0, message=FALSE, warning=FALSE, paged.print=TRUE}

model_0 <- lm(formula_0, data = Data_train)

summary(model_0)

#Saving R-squared

r_sq_0 <- summary(model_0)$r.squared

#predict data on test set

prediction_0 <- predict(model_0, newdata = Data_test)

#calculating the residuals

residuals_0 <- Data_test$charges - prediction_0

#calculating Root Mean Squared Error

rmse_0 <- sqrt(mean(residuals_0^2))

### Train and Test New Model

```{r model_1, message=FALSE, warning=FALSE, paged.print=TRUE}

formula_1 <- as.formula("charges ~ age + bmi + children + smoker + region")

model_1 <- lm(formula_1, data = Data_train)

summary(model_1)

r_sq_1 <- summary(model_1)$r.squared

prediction_1 <- predict(model_1, newdata = Data_test)

residuals_1 <- Data_test$charges - prediction_1

rmse_1 <- sqrt(mean(residuals_1^2))

```

Compare the models

```{r comparison, message=FALSE, warning=FALSE, paged.print=TRUE}

print(paste0("R-squared for first model:", round(r_sq_0, 4)))

print(paste0("R-squared for new model: ", round(r_sq_1, 4)))

print(paste0("RMSE for first model: ", round(rmse_0, 2)))

print(paste0("RMSE for new model: ", round(rmse_1, 2)))

### Model Performance
```

```{r performance, message=FALSE, warning=FALSE, paged.print=TRUE}

Data_test$prediction <- predict(model_1, newdata = Data_test)

ggplot(Data_test, aes(x = prediction, y = charges)) +

geom_point(color = "blue", alpha = 0.7) +

geom_abline(color = "red") +

ggtitle("Prediction vs. Real values")

Data_test$residuals <- Data_test$charges - Data_test$prediction

ggplot(data = Data_test, aes(x = prediction, y = residuals)) +

geom_pointrange(aes(ymin = 0, ymax = residuals), color = "blue", alpha = 0.7) +

geom_hline(yintercept = 0, linetype = 3, color = "red") +

ggtitle("Residuals vs. Linear model prediction")

ggplot(Data_test, aes(x = residuals)) +

geom_histogram(bins = 15, fill = "blue") +

ggtitle("Histogram of residuals")

GainCurvePlot(Data_test, "prediction", "charges", "Model")

### Applying on new data

```{r new_test, message=FALSE, warning=FALSE, paged.print=TRUE}

Bob <- data.frame(age = 19,

                  bmi = 27.9,

                  children = 0,

                  smoker = "yes",

                  region = "northwest")
```
```

```
print(paste0("Health care charges for Bob: ", round(predict(model_1, Bob), 2)))

Lisa <- data.frame(age = 40,

                   bmi = 50,

                   children = 2,

                   smoker = "no",

                   region = "southeast")

print(paste0("Health care charges for Lisa: ", round(predict(model_1, Lisa), 2)))


John <- data.frame(age = 30,

                   bmi = 31.2,

                   children = 0,

                   smoker = "no",

                   region = "northeast")

print(paste0("Health care charges for John: ", round(predict(model_1, John), 2)))
```

Chapter-2 Methodology

2.1 Datasets

- In this We Used dataset of Insurance.csv.
- Age: insurance contractor age, years
- Sex: insurance contractor gender, [female, male]
- BMI: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9

- Children: number of children covered by health insurance / Number of dependents

- Smoker: smoking [yes, no]
- Region: the beneficiary's residential area in the US, [northeast, southeast, southwest, northwest]
- Charges: Individual

Screenshot of dataset:

	A	B	C	D	E	F	G	H
1	age	sex	bmi	children	smoker	region	charges	
2	19	female	27.9	0	yes	southwest	16884.92	
3	18	male	33.77	1	no	southeast	1725.552	
4	28	male	33	3	no	southeast	4449.462	
5	33	male	22.705	0	no	northwest	21984.47	
6	32	male	28.88	0	no	northwest	3866.855	
7	31	female	25.74	0	no	southeast	3756.622	
8	46	female	33.44	1	no	southeast	8240.59	
9	37	female	27.74	3	no	northwest	7281.506	
10	37	male	29.83	2	no	northeast	6406.411	
11	60	female	25.84	0	no	northwest	28923.14	
12	25	male	26.22	0	no	northeast	2721.321	
13	62	female	26.29	0	yes	southeast	27808.73	
14	23	male	34.4	0	no	southwest	1826.843	
15	56	female	39.82	0	no	southeast	11090.72	
16	27	male	42.13	0	yes	southeast	39611.76	
17	19	male	24.6	1	no	southwest	1837.237	
18	52	female	30.78	1	no	northeast	10797.34	
19	23	male	23.845	0	no	northeast	2395.172	
20	56	male	40.3	0	no	southwest	10602.39	
21	30	male	35.3	0	yes	southwest	36837.47	
22	60	female	36.005	0	no	northeast	13228.85	
23	30	female	32.4	1	no	southwest	4149.736	
24	18	male	34.1	0	no	southeast	1137.011	
25	34	female	31.92	1	yes	northeast	37701.88	
26	37	male	28.025	2	no	northwest	6203.902	
27	59	female	27.72	3	no	southeast	14001.13	
28	63	female	23.085	0	no	northeast	14451.84	
29	55	female	32.775	2	no	northwest	12268.63	

2.2 Approach

1) Firstly, we are setting the environment and import the data.

```
##       age    sex    bmi children
smoker    region   charges
## 385   44   male 22.135        2
no northeast  8302.536
## 1054  47   male 29.800        3
yes southwest 25309.489
## 547   28   male 35.435        0
no northeast  3268.847
## 1179  23 female 34.865        0
no northeast  2899.489
## 1255  34 female 27.720        0
no southeast  4415.159
```

2) Then read the data from the csv file.

```
## age
##       n  missing distinct     Info     Mean      Gmd      .05      .10
##    1338        0       47    0.999    39.21    16.21       18       19
##     .25      .50      .75      .90      .95
##      27       39       51       59       62
##
## lowest : 18 19 20 21 22, highest: 60 61 62 63 64
## ------------------------------------------------------------------------
## sex
##       n  missing distinct
##    1338        0        2
##
## Value      female   male
## Frequency     662    676
## Proportion  0.495  0.505
## ------------------------------------------------------------------------
## bmi
##       n  missing distinct     Info     Mean      Gmd      .05      .10
##    1338        0      548        1    30.66    6.893    21.26    22.99
##     .25      .50      .75      .90      .95
##   26.30    30.40    34.69    38.62    41.11
##
## lowest : 15.960 16.815 17.195 17.290 17.385, highest: 48.070 49.060 50.380 52.580 53.130
## ------------------------------------------------------------------------
## children
##       n  missing distinct     Info     Mean      Gmd
##    1338        0        6    0.899    1.095    1.275
##
## Value          0     1     2     3     4     5
## Frequency    574   324   240   157    25    18
## Proportion 0.429 0.242 0.179 0.117 0.019 0.013
## ------------------------------------------------------------------------
## smoker
##       n  missing distinct
##    1338        0        2
##
## Value        no   yes
## Frequency   1064   274
## Proportion 0.795 0.205
## ------------------------------------------------------------------------
## region
##       n  missing distinct
##    1338        0        4
##
## Value      northeast northwest southeast southwest
## Frequency        324       325       364       325
## Proportion     0.242     0.243     0.272     0.243
## ------------------------------------------------------------------------
## charges
##       n  missing distinct     Info     Mean      Gmd      .05      .10
```

3) Here, the parameters like age, BMI, Sex, Region, etc. are linearly affecting on dependent parameter charges.

4) We splitted the dataset for training and testing purpose in the ratio 70:30.

```
##
## Call:
## lm(formula = formula_0, data = Data_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10814.9  -3037.9   -978.6   1618.7  29863.0
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -12518.87    1102.55 -11.354  < 2e-16 ***
## age               252.85      13.52  18.707  < 2e-16 ***
## sexmale          -127.93     378.43  -0.338  0.73538
## bmi               369.02      32.14  11.481  < 2e-16 ***
## children          425.64     155.97   2.729  0.00646 **
## smokeryes       23746.57     468.18  50.721  < 2e-16 ***
## regionnorthwest   -348.52     541.19  -0.644  0.51972
## regionsoutheast   -951.40     545.46  -1.744  0.08141 .
## regionsouthwest  -1298.90     536.82  -2.420  0.01570 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6171 on 1061 degrees of freedom
## Multiple R-squared:  0.7467, Adjusted R-squared:  0.7448
## F-statistic: 390.9 on 8 and 1061 DF,  p-value: < 2.2e-16
```

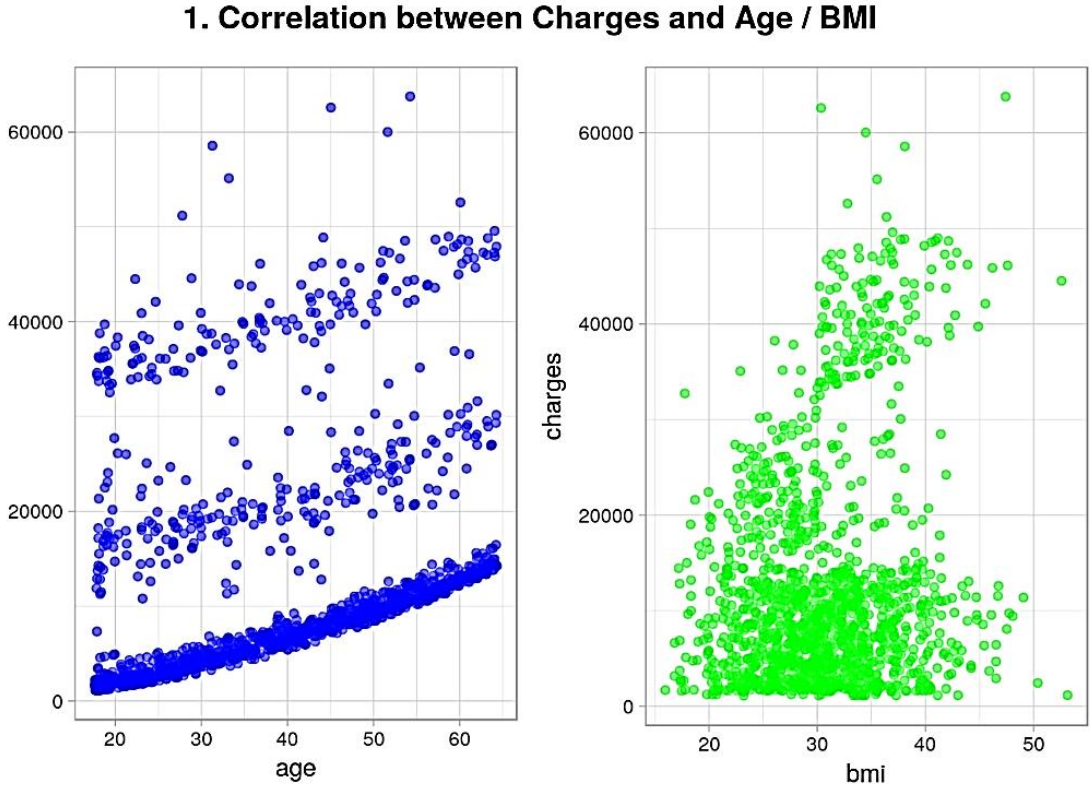
5)Here we trained two machine learning models and tested for most significant variables affecting on the target variable.

```
##
## Call:
## lm(formula = formula_1, data = Data_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10873.7  -3035.9   -977.2   1604.4  29806.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -12573.22    1090.32 -11.532  < 2e-16 ***
## age               252.87      13.51  18.716  < 2e-16 ***
## bmi               368.68      32.11  11.480  < 2e-16 ***
## children          424.85     155.89   2.725  0.00653 **
## smokeryes       23736.72     467.08  50.820  < 2e-16 ***
## regionnorthwest   -347.87     540.96  -0.643  0.52033
## regionsoutheast   -949.67     545.21  -1.742  0.08183 .
## regionsouthwest  -1295.39     536.50  -2.415  0.01592 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6168 on 1062 degrees of freedom
## Multiple R-squared:  0.7466, Adjusted R-squared:  0.745
## F-statistic: 447.1 on 7 and 1062 DF,  p-value: < 2.2e-16
```

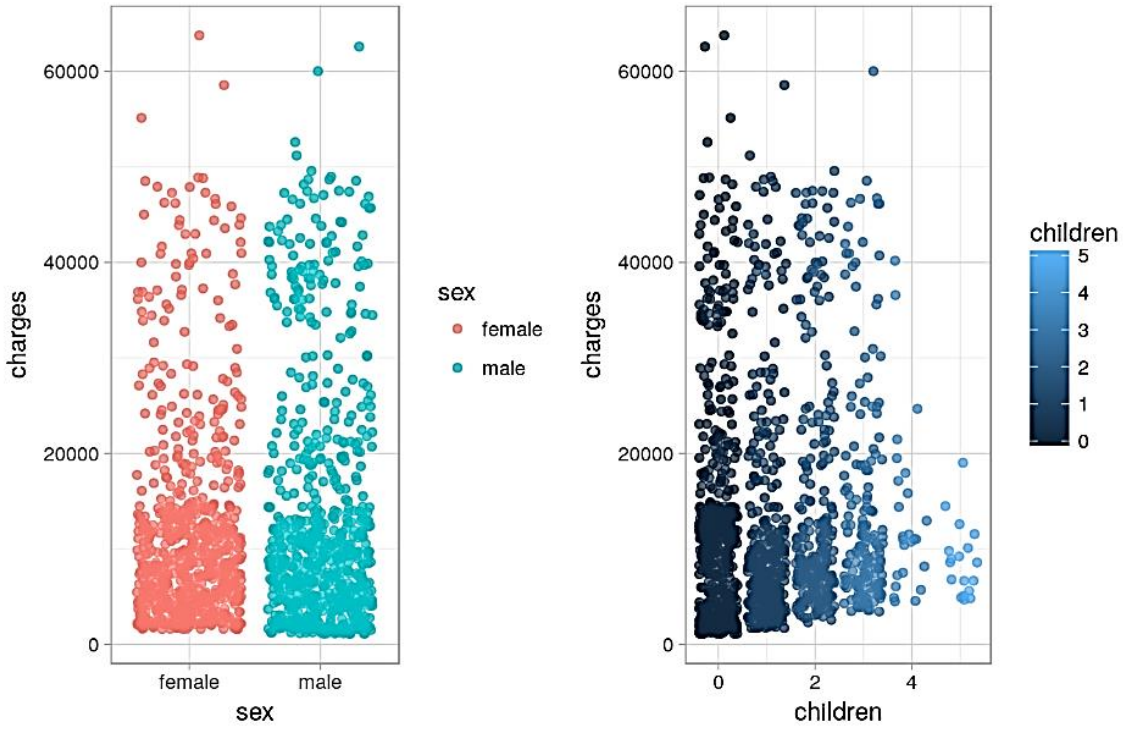
6) In that, firstly we have processed the data in which we found no missing values.

```
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1338        0       47    0.999    39.21    16.21       18       19
##      .25      .50      .75      .90      .95
##       27       39       51       59       62
##
## lowest : 18 19 20 21 22, highest: 60 61 62 63 64
## ------------------------------------------------------------------------------
## sex
##        n  missing distinct
##     1338        0        2
##
## Value      female   male
## Frequency     662    676
## Proportion  0.495  0.505
## ------------------------------------------------------------------------------
## bmi
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1338        0      548        1    30.66    6.893    21.26    22.99
##      .25      .50      .75      .90      .95
##    26.30    30.40    34.69    38.62    41.11
##
## lowest : 15.960 16.815 17.195 17.290 17.385, highest: 48.070 49.060 50.380 52.580 53.130
## ------------------------------------------------------------------------------
## children
##        n  missing distinct     Info     Mean      Gmd
##     1338        0        6    0.899    1.095    1.275
##
## Value          0     1     2     3     4     5
## Frequency    574   324   240   157    25    18
## Proportion 0.429 0.242 0.179 0.117 0.019 0.013
## ------------------------------------------------------------------------------
## smoker
##        n  missing distinct
##     1338        0        2
##
## Value         no   yes
## Frequency   1064   274
## Proportion 0.795 0.205
## ------------------------------------------------------------------------------
## region
##        n  missing distinct
##     1338        0        4
##
## Value      northeast northwest southeast southwest
## Frequency        324       325       364       325
## Proportion     0.242     0.243     0.272     0.243
## ------------------------------------------------------------------------------
## charges
##        n  missing distinct     Info     Mean      Gmd      .05      .10
```

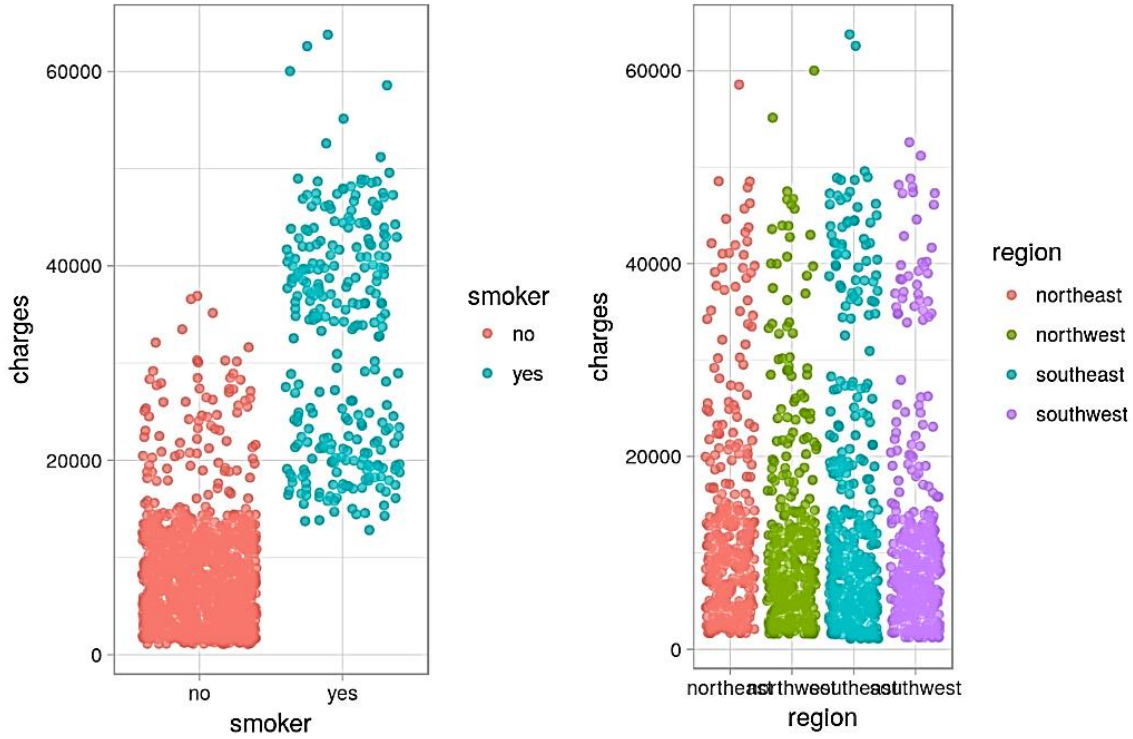
7) After that we plotted the correlation between dependent and target variables for finding the most significant variable among all.



2. Correlation between Charges and Sex / Children covered by insurance



3. Correlation between Charges and Smoker / Region



2.3 Platform and Technology

- R Studio

- GGPLOT
- DPLYR
- COWPLOT
- HMISC
- WVPLOT

2.3 Outcomes & Use cases

1. Comparing the models and finding the best.

```
## [1] "R-squared for first model:0.7467"
```

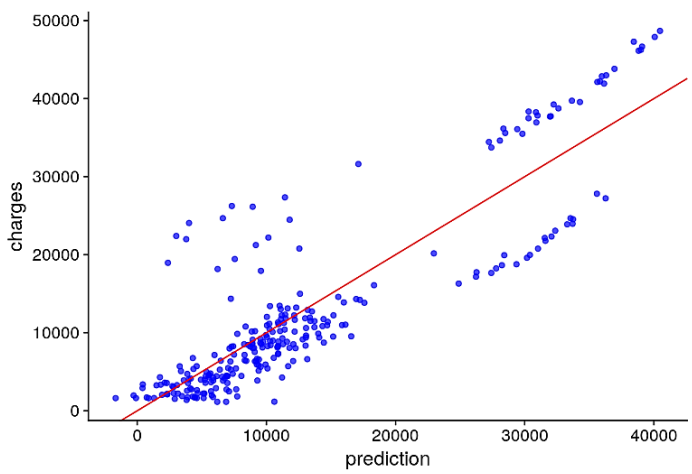
```
## [1] "R-squared for new model: 0.7466"
```

```
## [1] "RMSE for first model: 5641.95"
```

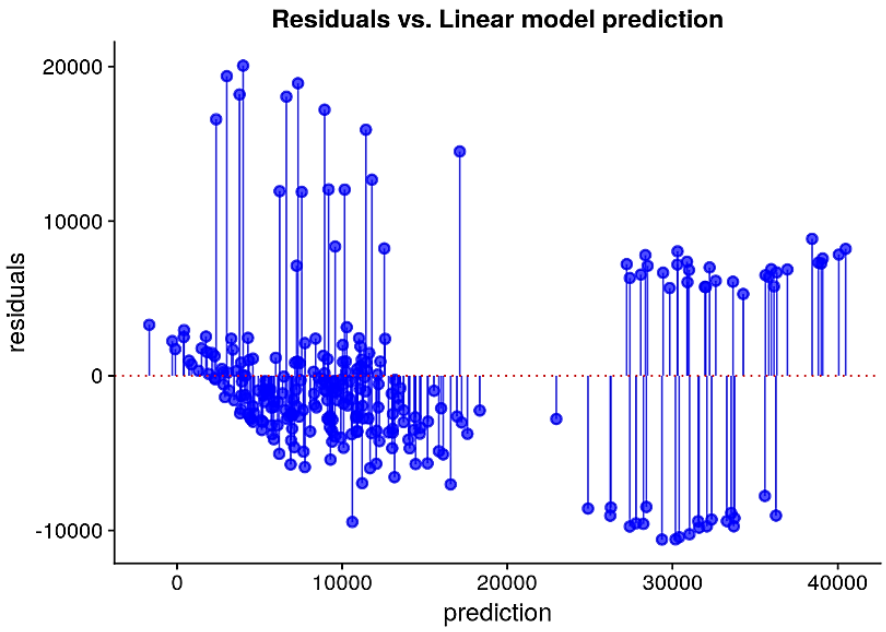
```
## [1] "RMSE for new model: 5642.45"
```

2. Finding the model performances.

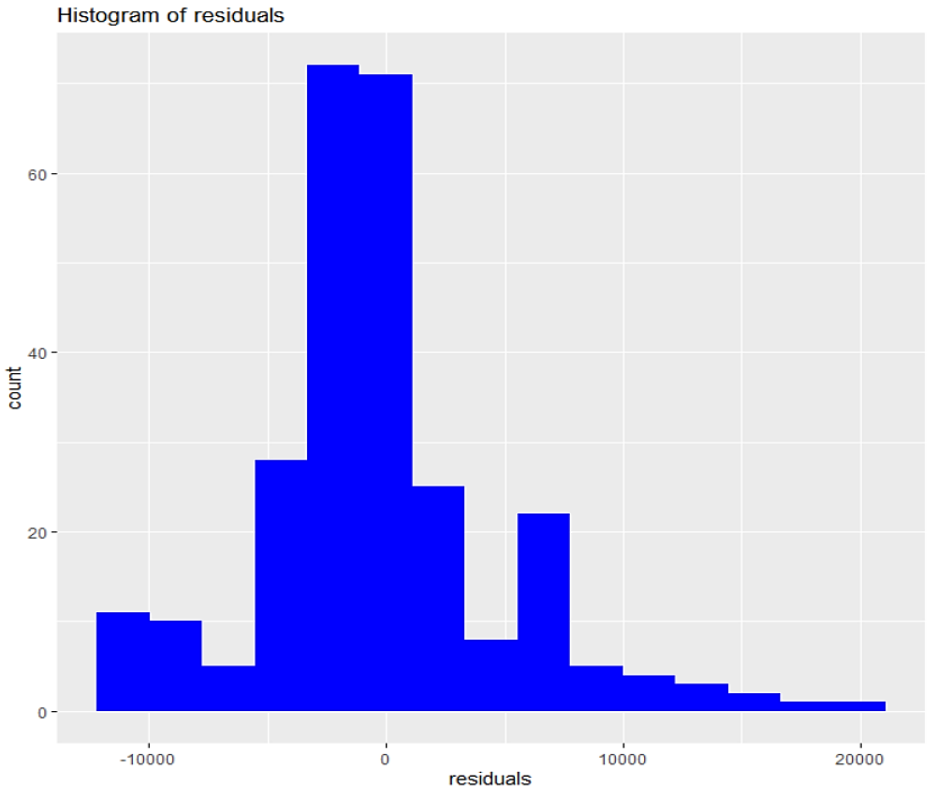
    a. Prediction vs Real values



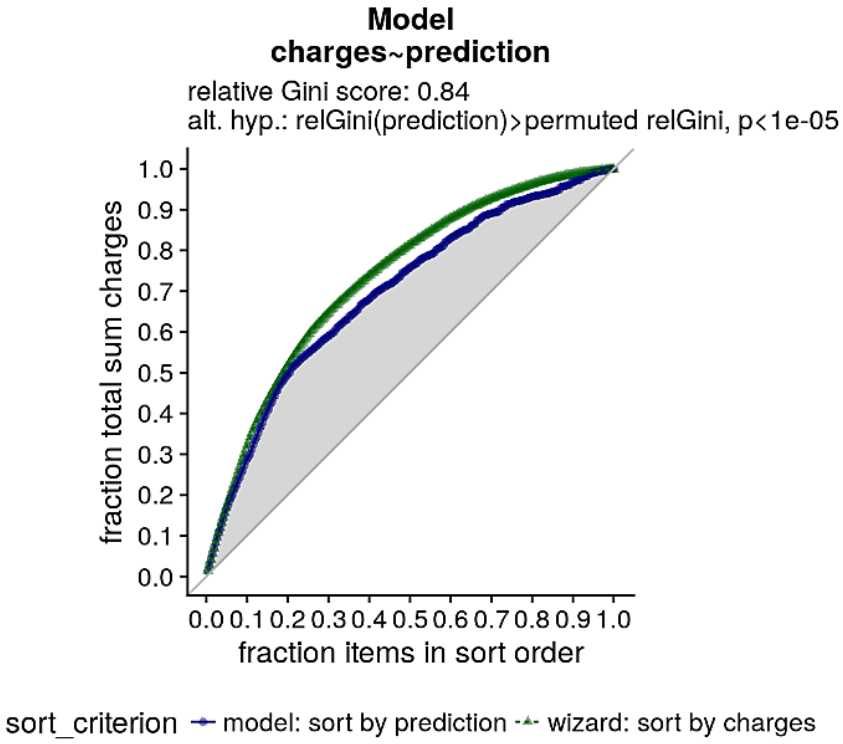
b. Residuals vs linear model prediction



3. By comparing the residual values with predictable and real values. We have found the histogram about the residuals.



4. Lastly, we have compared the model sort by prediction and model sort by charges and finding the model predictor.



5. We can see the errors in the model are close to zero, so model predicts quite well.

6. Predicted health care charges in dollars for Bob, Lisa and John are:

```
[1] "Health care charges for Bob: 25999.5"

[1] "Health care charges for Lisa: 15368.96"

[1] "Health care charges for John: 6355.71"
```

2.4 Challenges

- Finding and Solving errors during project implementation.
- Choosing the appropriate idea to perform project.
- Using few commands for the first time with no prior experience.
- Taking application-based project and implementing it.

Chapter-3 Conclusion

3.1 Conclusion

In this project we have implemented various concepts of machine learning like data processing, data cleaning, predicting data, finding model performances. For that we successfully trained the data and with the help of graphs find the most significant factor i.e., Smokers which affect most on the charges in health care.

3.2 Future work

In this project we have just taken the few parameters. In the future work we will take the symptoms and parameters related in causing health problems more and predict the charge of the disease correctly.

3.3 References

- https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a
- https://machinelearningmastery.com/linear-regression-for-machine-learning/?utm_source=canva&utm_medium=iframely
- https://ieeexplore.ieee.org/abstract/document/8250771
- https://www.pnhp.org/single_payer_resources/health_care_systems_four_basic_models.php

THANK YOU