1 Section A - Medical Insurance

1.1 Is the required ML supervised, unsupervised, or semi-supervised learning and why? Which ML task (classification, clustering, regression analysis or any other) is the best in this case and why?

Algorithms are trained on input data that has been labelled for a specific output in supervised machine learning (Jiang, Gradus and Rosellini, 2020). Without labelled outputs, unsupervised machine learning algorithms are employed to find patterns in a data set, labelling each observed input on its own (Gentleman and Carey, 2008). On the other hand semi-supervised machine learning incorporates both supervised and unsupervised machine learning and as such is used to train a combination of labelled and unlabeled data (Zhu and Goldberg, 2009).

Since the predicted or output variable, `medicalCost` is labelled, supervised machine learning is required for this task.

Regression, a category of supervised machine learning is the required task for the stated problem. This is because regression is suited for tasks where the predicted variable is on a continuous scale(Freund, Wilson and Sa, 2006), and for this study, the `medicalCost` predicted variable is expressed in continuous monetary terms.

1.2 Explore your data and document your observation.

1.2.1 Data structure and missingness

The insurance data set was comprised of 1338 observations and 7 variables. These variables form a mixture of numbers, integers and characters. The `age` and `children` variables are integers and the `bmi` and `medicalCost` variables are numeric. The `smoker` and `sex` character variables are binary with levels being `yes and no` and `male and female` respectively. The `region` character variable is multinomial with four levels `southwest, southeast, northwest and northeast`.

```
# Data importation
library(tidyverse)
dat_ins <- read.csv("./insurance.csv")
# Structure of data
str(dat_ins)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age        : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex        : chr  "female" "male" "male" "male" ...
##  $ bmi        : num  27.9 33.8 33 22.7 28.9 ...
##  $ children   : int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker     : chr  "yes" "no" "no" "no" ...
##  $ region     : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ medicalCost: num  16885 1726 4449 21984 3867 ...
```

```
# Overview of first 6 rows
head(dat_ins)
```

```
##   age    sex    bmi children smoker    region medicalCost
## 1  19 female 27.900        0    yes southwest   16884.924
## 2  18   male 33.770        1     no southeast    1725.552
## 3  28   male 33.000        3     no southeast    4449.462
## 4  33   male 22.705        0     no northwest   21984.471
## 5  32   male 28.880        0     no northwest    3866.855
## 6  31 female 25.740        0     no southeast    3756.622
```

There were also no missing values in the insurance data set.

```
# Sum of missing data values per column
colSums(is.na(dat_ins))
```

```
##         age         sex         bmi    children      smoker      region 
##           0           0           0           0           0           0 
## medicalCost 
##           0
```

1.2.2 Distribution of medical cost per smoker status

From figure 1, Individuals who do not smoke have a right-skewed distribution of medical costs. Non-smokers pay relatively lower medical costs with a mode of about USD 8000. Smoker on the other hand had a bi-modal distribution of medical costs. Smokers seemed to pay a higher cost with two modes of about USD 20,000 and USD 41,000. This disparity between smokers and non-smokers may be expected since smokers may be more prone to health issues such as lung cancer, strokes and heart attacks(Hays, et al., 1998), necessitating a higher medical insurance cost.

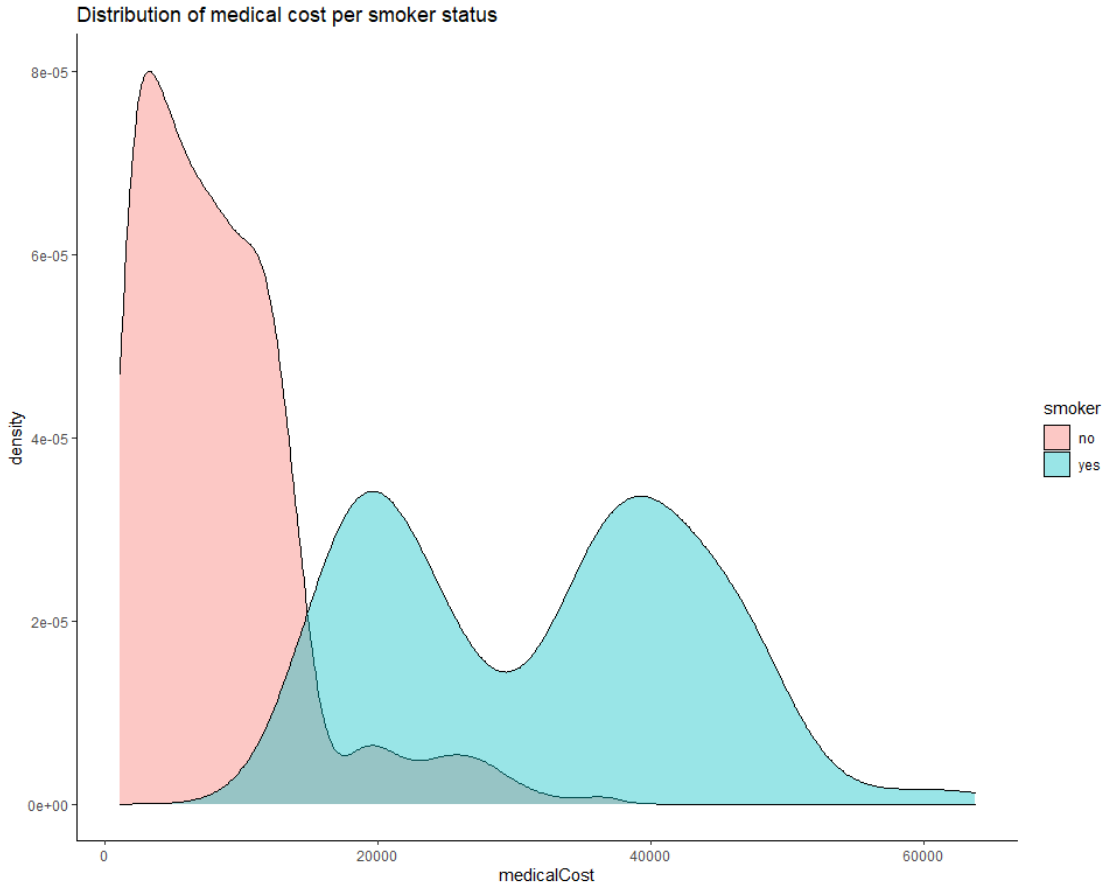


Figure 1-Distribution of medical cost per smoker status

1.2.3 Average medical costs per age group

Figure 2 shows the average medical cost per age group. There is an incremental relationship between the age groups and medical costs with the 60 - 70 age group paying the highest average medical cost of about USD 22,000. This result seems rational because as people age their immune system becomes weaker increasing their chances of getting sick (Martins, et al., 2005). This elevated risk of sickness corresponds with more medical attention and hence a higher medical insurance cost to cover such expenses.

```
 # create age bin variables with breaks of 10 to 70.
dat_ins %>%
  mutate(age_bin = cut(age, breaks=c(10, 20, 30, 40, 50, 60, 70))) %>%
  group_by(age_bin) %>% # group by age
  summarise(avg_medical_cost = mean(medicalCost)) %>% # find mean of medical cost
  ggplot(aes(x=age_bin, y=avg_medical_cost)) +
  geom_segment( aes(x=age_bin,
                    xend=age_bin,
                    y=0, yend=avg_medical_cost), color="skyblue") + # plot segments
  geom_point( color="blue", size=4, alpha=0.6) + # plot points
  theme_light() + # set theme to light
  labs(title = "Average medical costs per age group",
       x = "Age group",
       y = "Average medical cost")+ # title , x and y labels
  coord_flip() + # flip diagram
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank()
  ) # remove panel and axis ticks
```

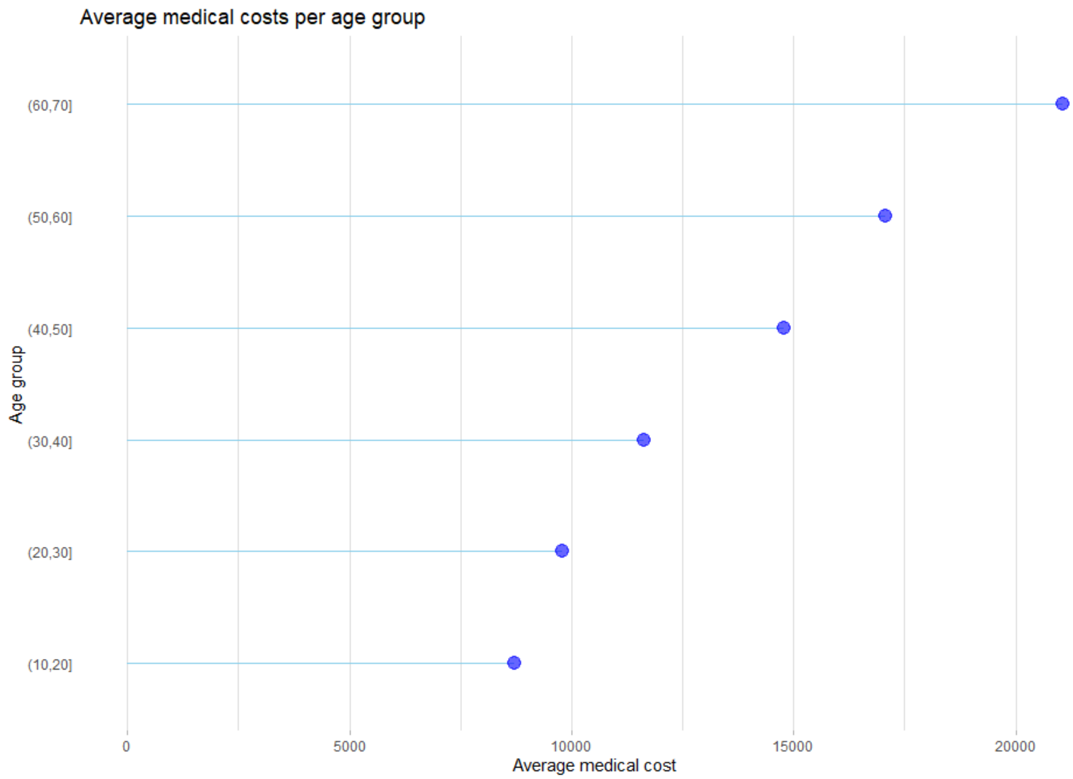


Figure 2-Average medical costs per age group

1.2.4 Relationship between Bmi and MedicalCost

Figure 3 depicts the relationship between BMI (Body Mass Index) and medical cost. The figure shows a weak but positive relationship between these two variables. Higher BMI has been associated with high blood pressure, inflammation and high levels of blood sugar. These may lead to an elevated risk for heart disease, stroke and other serious diseases(Hall, et al., 2014). As such, individuals with higher BMIs may have higher medical insurance costs for potential diseases.

```
dat_ins %>%
  ggplot(aes(x=bmi, y=medicalCost)) +
  geom_point(shape=2) +
  geom_smooth(method = "lm") +
  theme_bw() +
  labs(title = "Relationship between Bmi and MedicalCost")

## `geom_smooth()` using formula 'y ~ x'
```

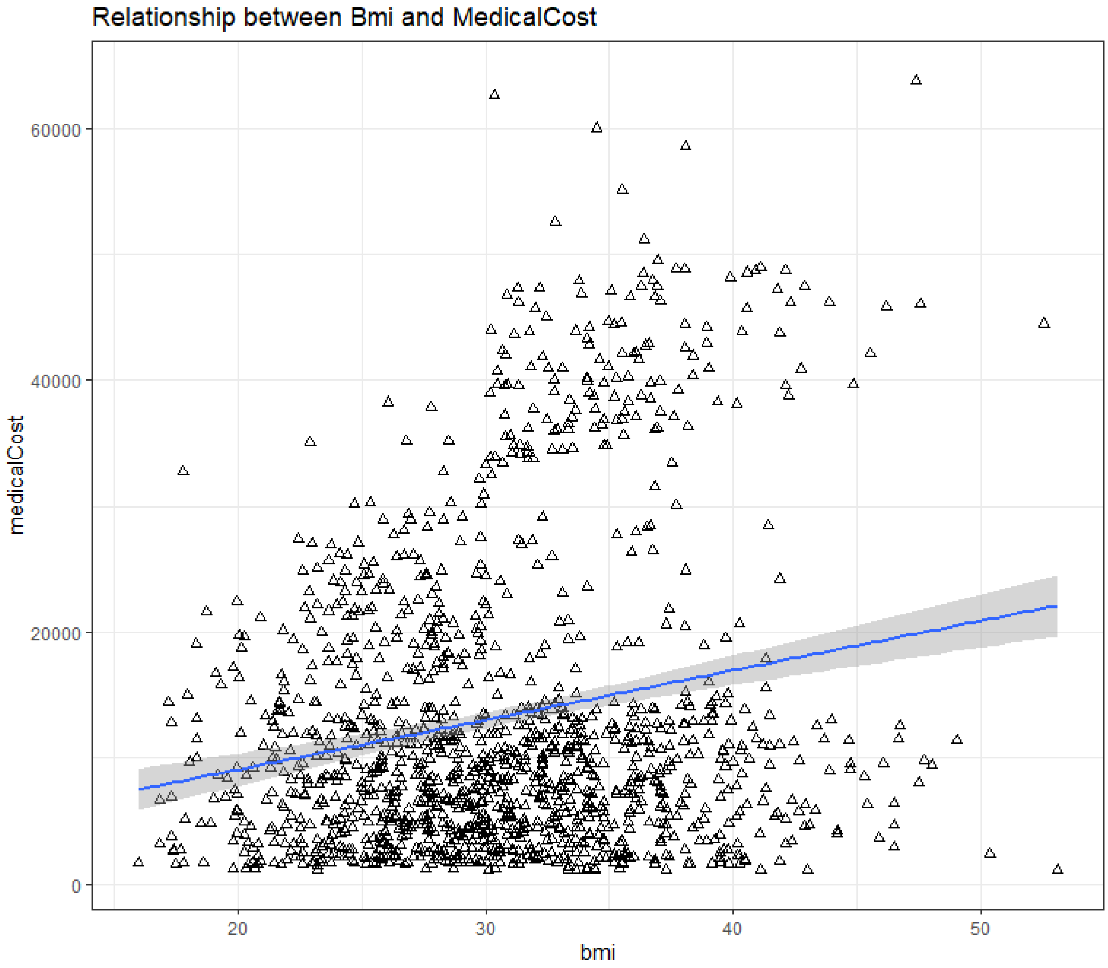


Figure 3-Relationship between BMI and Medical cost

1.2.5 Distribution of medical cost per sex

Figure 4 shows the distribution of medical costs in males and females. The median cost for both sexes is approximately the same, however, the boxplot for males shows a higher upper quartile indicating that 75% of males pay medical costs below USD 19,000. On the other hand, the upper quartile for the female box plot indicates that 75% of females pay medical costs of less than 15,000.

```
# assign sex on the sex on the x and medical cost on the y
dat_ins %>%
  ggplot(aes(x=sex, y=medicalCost)) +
  geom_boxplot(aes(col=sex)) + # boxplot with color based on sex
  theme_minimal() # Minimal theme
```

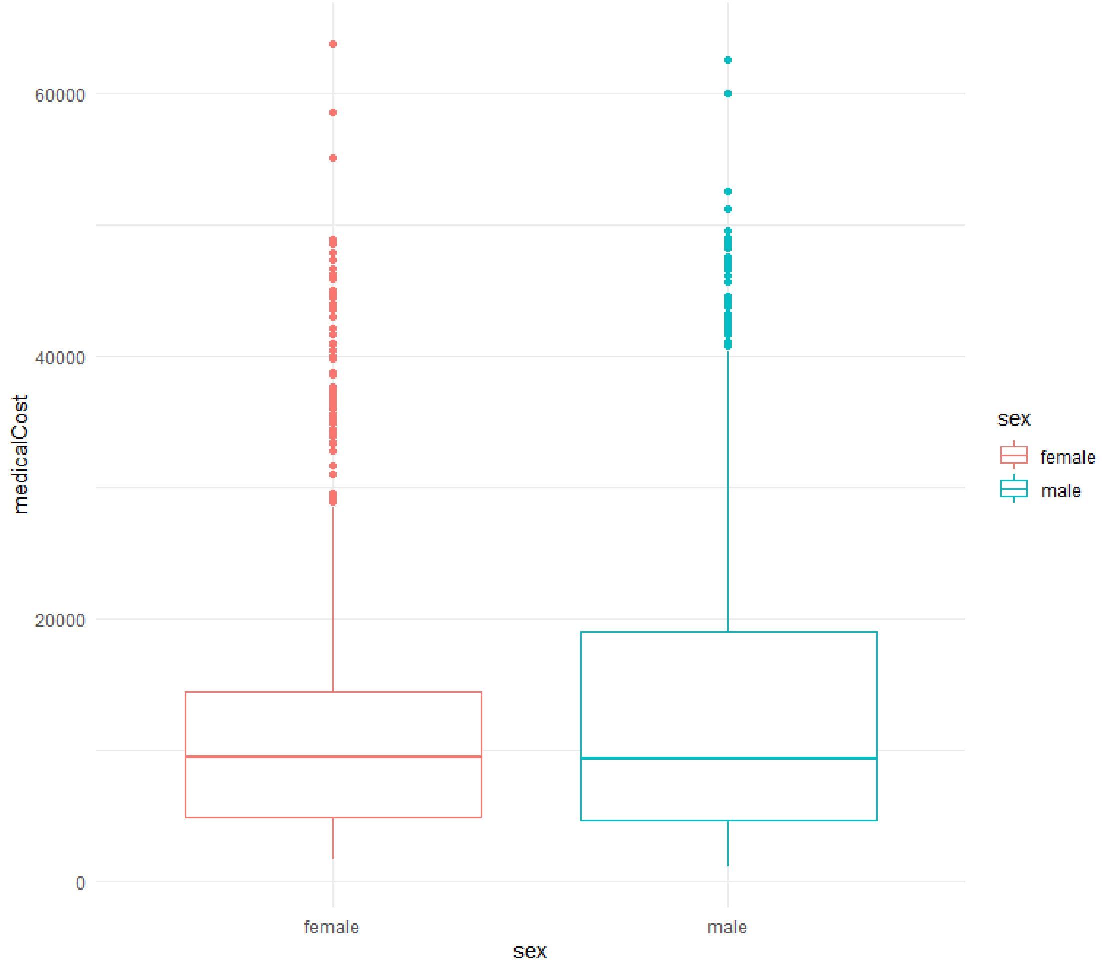


Figure 4-Distribution of medical cost per sex

A view of the average medical costs from both sex table 1 shows males pay a slightly higher premium of USD 13,956.75 compared to females with USD 12,569.58.

```
if (!require(knitr)) install.packages('knitr') # Install knitr if required
library(knitr) # for table generation
dat_ins %>%
  group_by(sex) %>%
  summarise(average.medical.cost = mean(medicalCost)) %>%
  kable()
```

Table 1-Average medical cost per sex

sex	average.medical.cost
female	12569.58
male	13956.75

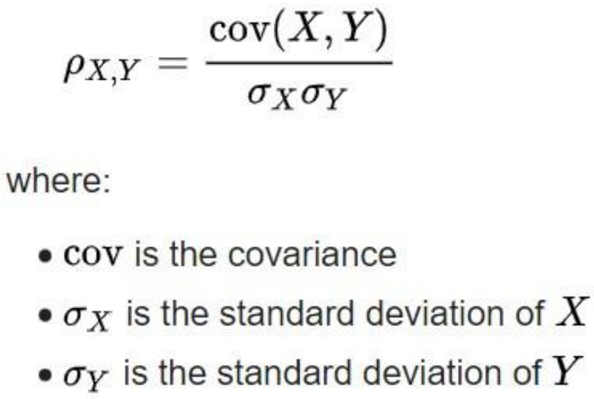
These differences in medical cost may have occurred because men are more susceptible to certain chronic illnesses that can be fatal, such as coronary heart disease, cancer, cerebrovascular disease, emphysema, cirrhosis of the liver, and kidney disease, and atherosclerosis (Vlassoff, 2007). Additionally, men around the world have a shorter life expectancy than women do (Barford, 2006). The greater insurance costs for men may be influenced by these risk factors.

1.3 Study the correlation between each predictor and the medicalCost. What is your conclusion?

1.3.1 Correlation for numerical variables

Correlation is a measure of the strength of a relationship between two variables. To quantify this relationship in numerical variables the Person's correlation coefficient may be suitable.

The Person's correlation coefficient (ρ) or the Pearson Product Moment Correlation expresses the correlation between two variables say X and Y, by dividing their co-variance over the product of their standard deviations.



$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

Person's correlation

This results in a value between -1 and 1, with 1 indicating a strong positive relationship, -1 indicating a strong negative relationship and 0 indicating the absence of any relationship.

The numerical variables were selected from the insurance data set and correlated with each other. To visualize how the numeric independent variables correlated with the dependent variable, the correlation was sorted in a decreasing manner based on `medicalcost`. This was showcased in the results section.

```
num_dat <- dat_ins %>%
  select(-c(sex,smoker,starts_with("region"))) # select only numerical variables

#correlations of all numeric variables with person method
cor_numVar <- cor(num_dat, method = "pearson")
# sort on decreasing correlations with medicalCost
cor_num_matrix <- as.matrix(sort(cor_numVar[,'medicalCost'], decreasing = T))
```

1.3.2 Correlation for categorical variables

Finding the correlation between the categorical variables and the dependent variable may require a different approach. To achieve this the point-biserial correlation was considered. The point biserial correlation is a special case of Pearson's correlation coefficient where the degree of the relationship between one continuous variable and one dichotomous variable is determined.

To compute the point-biserial correlation coefficient it is assumed that a dichotomous variable says `b` has two values 0 and 1. The dat set may then be seperated into two groups based on these values with the first group receiving the value of 1 and the second 0. The calculation will therefore be:

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}},$$

Point-biserial correlation

where: *sn* is the standard deviation for the entire population

M1 is the mean value for the continuous variable `p` for all data points in the first group.

M0 is the mean value of the continuous variable `p` for all points in the second group.

n1 is the number of data points in the first group

n0 is the number of data points in the second group

n is the total sample size.

To apply the point-biserial correlation certain assumptions are required. These are:

1. Dichotomy of categorical variables
2. Normal distribution of the continuous variable
3. No outlines in the continuous variable

1.3.2.1 The categorical variables have to be dichotomous

The categorical variables `sex` and `smoke` are already dichotomous or binary with two levels being `male and female` and `yes and no` respectively. These were made numerical with `male and yes` assigned to 1 and `female and no` assigned to 0.

```
# Encode sex and smoke variable
dat_ins$smoker <- ifelse(dat_ins$smoker == 'yes',1,0)
dat_ins$sex <- ifelse(dat_ins$sex == 'male',1,0)

# head to view changes
head(dat_ins)

##   age sex    bmi children smoker    region medicalCost
## 1  19   0 27.900        0      1 southwest   16884.924
## 2  18   1 33.770        1      0 southeast    1725.552
## 3  28   1 33.000        3      0 southeast    4449.462
## 4  33   1 22.705        0      0 northwest   21984.471
## 5  32   1 28.880        0      0 northwest    3866.855
## 6  31   0 25.740        0      0 southeast    3756.622
```

region on the other hand has four levels southwest, southeast, northwest and northeast. This makes the region variable polychotomous, that is having more than two values. To rectify this and comply with the assumptions, one-hot encoding was employed. Through one-hot encoding, each region within the region variable will be made into its variable and assigned a dichotomous value of 0 and 1. 1 will indicate when that particular region or data point occurred within the data set and 0 will indicate when it was absent. Making these variables binary will allow for compliance with the assumption of dichotomy.

```
if (!require(caret)) install.packages('caret')
library(caret)
# one hot encode categorical variables
# Encode region
dummy <- dummyVars(" ~ region", data=dat_ins) # creating dummies for region
newdata <- data.frame(predict(dummy, newdata = dat_ins)) # predict the new on
e hot encoded data with the dummies
dat_ins <- cbind(newdata,dat_ins) # bind the new one hot encoded regions to t
he insurance dataset
dat_ins <- dat_ins %>% select(-region) # Remove the region variable as it is
no longer needed
```

```
head(dat_ins) # over view of changes

##   regionnortheast regionnorthwest regionsoutheast regionsouthwest age sex
## 1               0               0               0               1  19   0
## 2               0               0               1               0  18   1
## 3               0               0               1               0  28   1
## 4               0               1               0               0  33   1
## 5               0               1               0               0  32   1
## 6               0               0               1               0  31   0
##      bmi children smoker medicalCost
## 1 27.900        0      1   16884.924
## 2 33.770        1      0    1725.552
## 3 33.000        3      0    4449.462
## 4 22.705        0      0   21984.471
## 5 28.880        0      0    3866.855
## 6 25.740        0      0    3756.622
```

1.3.2.2 Normally distributed continuous variable

Another assumption is that the continuous variable must be normally distributed. This implies that the `medicalCost` variable must be symmetrically distributed around its mean with most values near the central peak.

However the histogram of `medicalCost` in figure 5, shows a right-skewed distribution flouting the normality assumption.

```
hist(dat_ins$medicalCost)
```

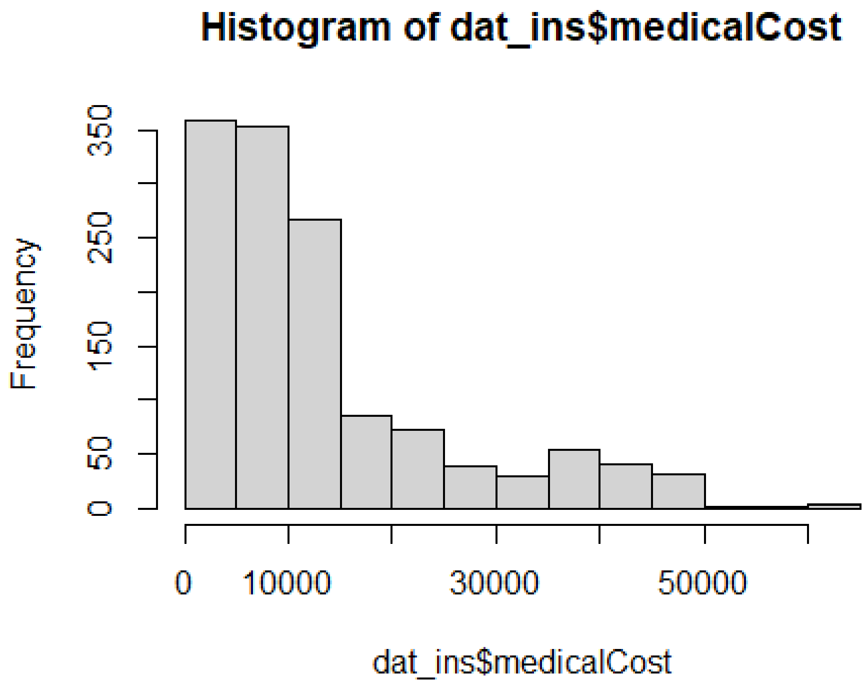


Figure 5-Distribution of medical cost

To rectify this the `medicalCost` variable was logged. Doing so allowed for a symmetrical distribution, implying normality and meeting the assumption as seen in figure 6.

```r
# Make continuous normally distributed
dat_ins$medicalCost <- log(dat_ins$medicalCost)
hist(dat_ins$medicalCost)
```

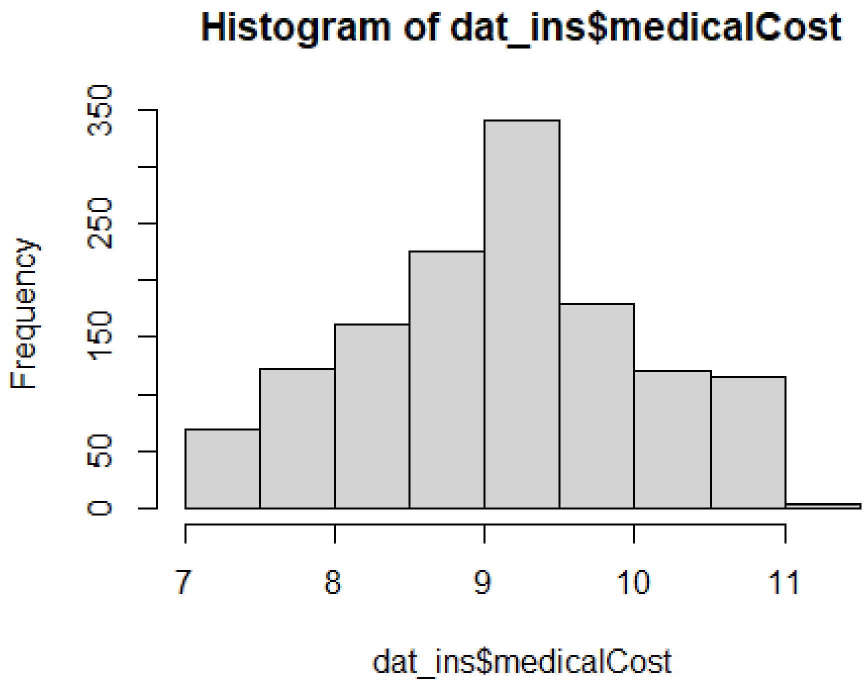


Figure 6-Logged distribution of medical cost

1.3.2.3 No outliers in the continuous variable

From figure 7, it may be concluded that after the log transformation of the dependent variable no outliers were present.

```
boxplot(dat_ins$medicalCost)
```

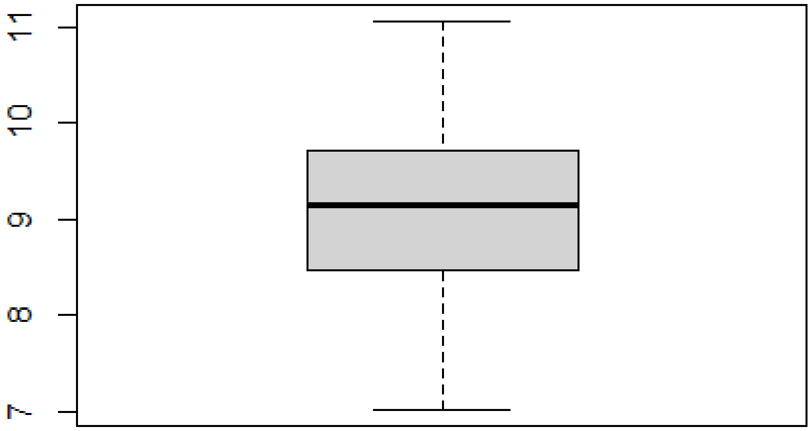


Figure 7-Boxplot of medical cost for outlier detection

After meeting the assumptions, the categorical variables and the dependent variables were selected and correlated. The categorical variables were sorted in a decreasing order based on `medicalCost`. These results were discussed in the next section. The dependent variable was converted back to its original state after results were obtained by computing the exponential of the logged values.

```r
# select the categorical variables and medical cost

# sex, smoker, medical cost and variables starting with region.
cat_dat <- dat_ins %>%
  select(sex,smoker,medicalCost,starts_with("region"))

cor_cat <- cor(cat_dat, method = "pearson") # Find correaltion with pearson

#sort on decreasing correlations with medicalCost
cor_cat_matrix <- as.matrix(sort(cor_cat[,'medicalCost'], decreasing = T))
```

```
# Reverse medical cost back to normal by finding the exponential
dat_ins$medicalCost <- exp(dat_ins$medicalCost)
```

1.3.3 Results of correlation analysis

from figure 8, the correlogram of numerical variables shows age as the most correlated variable to `medicalCost`. This is followed by `bmi` and `children`.

```
if (!require(corrplot)) install.packages('corrplot') # install corrplot if re
quired
library(corrplot) # Library for corrplot
CorHigh <- rownames(cor_num_matrix) # names of numerical variables

# index row and column names form correlation matrix
cor_numVar <- cor_numVar[CorHigh, CorHigh]

# plot correlation with corrplot, show only lower part of correlation plot.
corrplot(cor_numVar,type = "lower")
```

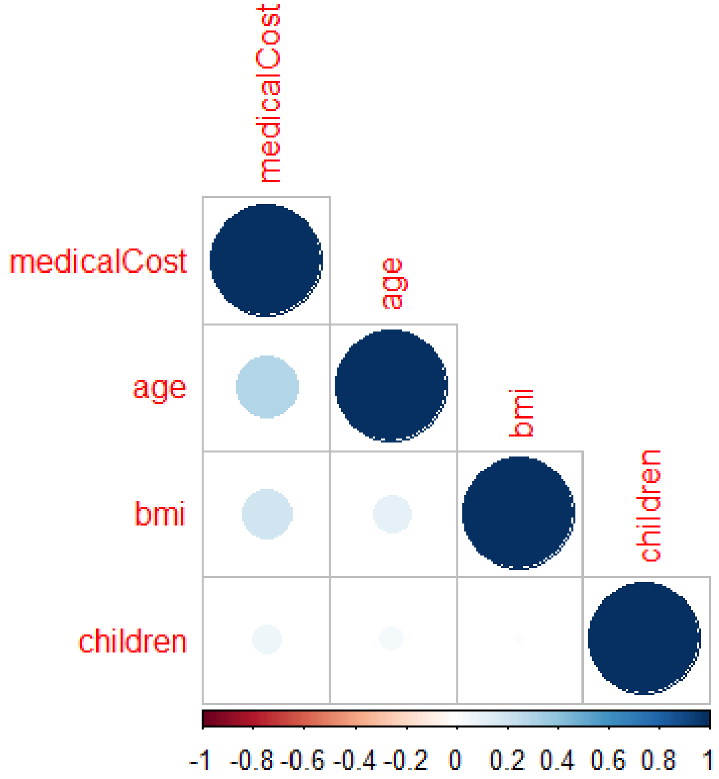


Figure 8-Correlogram of numerical variables

From figure 9, the correlogram of categorical variables shows smoker as the most correlated variable to medicalCost. This is followed by regionnortheast, regionsoutheast, sex, regionnorthwestand regionsouthwest.

```
CorHigh <- rownames(cor_cat_matrix) # names of categorical variables
cor_cat <- cor_cat[CorHigh, CorHigh] # index row and column names form correl
ation matrix
# plot correlation with corrplot, show only lower part of correlation
corrplot(cor_cat,type = "lower")
```

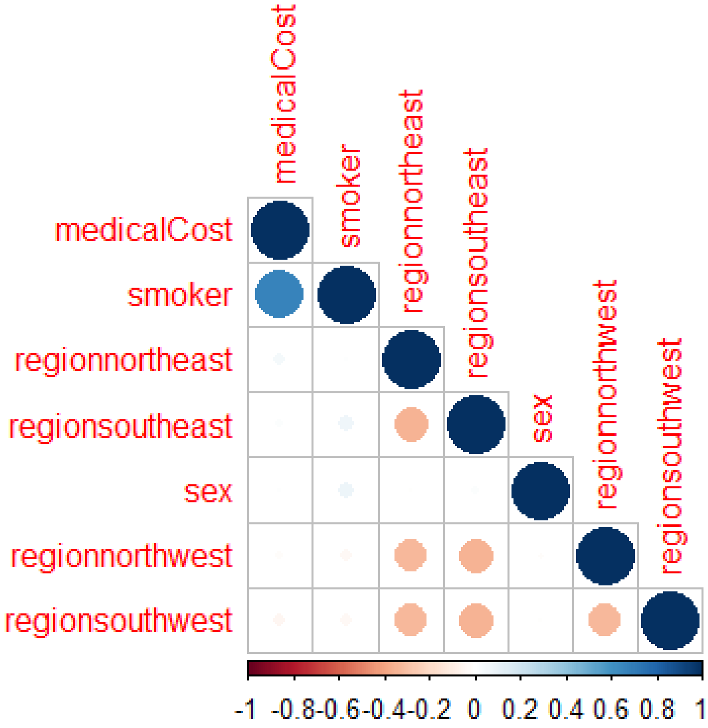


Figure 9-Correlogram of categorical variables

Table 2 shows the combined correlation of both the numerical and categorical variables. The smoker variable was the most correlated variable with a coefficient of about 0.67. This shows a strong positive relationship between smoking and medical costs. The age and bmi variables were the next most correlated variables with coefficients of approximately 0.3 and 0.2. These variables had a relatively weak but positive correlation with the dependent variable. The sex variable exhibited the weakest correlation with a coefficient of about

0.006. The regionnorthwest and regionsouthwest variables were negatively correlated to the medicalCost variable with coefficients of about -0.018 and -0.042 respectively.

```r
cor_dat <- rbind(cor_cat_matrix,cor_num_matrix) # bind numerical and categorical correlation matrix
cor_dat1 <- cor_dat[-1,] # remove duplicated medical Cost variable
cor_sorted <- as.matrix(sort(cor_dat1,decreasing = T)) # sort them in decreasing order and convert to matrix form
cor_sorted_df <- as.data.frame(cor_sorted) # convert to data frame
names(cor_sorted_df)[1]<-paste("Correlation") # Name column and correlation

cor_sorted_df %>% kable() # present in a table format
```

Table 2-Correlation coefficient of all predictors

	Correlation
medicalCost	1.0000000
smoker	0.6655057
age	0.2990082
bmi	0.1983410
children	0.0679982
regionnortheast	0.0431151
regionsoutheast	0.0157907
sex	0.0056319
regionnorthwest	-0.0178243
regionsouthwest	-0.0416321

1.4 Use the correlation analysis to select the three best predictors and build a simple linear regression model based on each of the predictors.

As discussed in section 1.3.3, the most correlated variables to medicalCost are smoker, age and bmi.

```
corr_var <- cor_sorted_df %>% rownames()
corr_var[2:4] # best three correlated variables

## [1] "smoker" "age"    "bmi"
```

These variables will be used for a simple linear regression with their results discussed in a later section.

```
# linear model with medical cost and smoker
fit_smoker <- lm(medicalCost~smoker,data = dat_ins)

# linear model with medical cost and age
fit_age <- lm(medicalCost~age,data = dat_ins)

# linear model with the medical cost and bmi
fit_bmi <- lm(medicalCost~bmi,data = dat_ins)
```

1.5 Evaluate the performance with the statistical performance measures to evaluate the statistical significance of your results.

To measure the statistical performance of the regression models the adjusted R-squared (R^2) and the residual standard error would be evaluated and compared.

The adjusted R^2 is a corrected goodness-of-fit measure for linear models. It is used to determine the percentage of the variations in the dependent variable that is explained by the independent variable, in other words, it shows how well the data at hand fits the regression model. The adjusted R^2 is calculated by dividing the residual mean squared error by the total mean squared error, with the result deducted from 1(Miles, 2005).

$$R^2 = 1 - \frac{Residual\ sum\ of\ squares}{Total\ sum\ of\ squares}$$

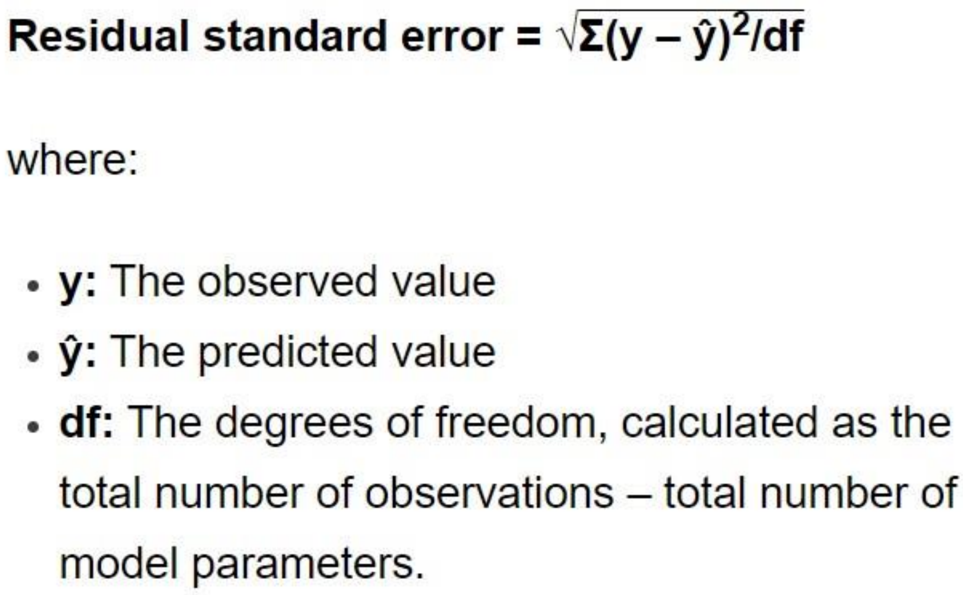
Adjusted R-squared

The total prediction error of our regression model is represented by the residual sum of squares. On the other hand, the total sum of squares represents the variation that the mean model could not account for, and as a result, it is proportional to the variance of the data(Morgan and Tatar, 1972).

A model that completely predicts the values of the dependent variables has an adjusted R^2 value of 1. A result that is less than or equal to 0 indicates that the model cannot predict anything. Simply put the model gets better as the adjusted R^2 gets to 1.

How well the model fits the data is also assessed using the residual standard error. It is the residuals' regression model's standard deviation.

It is computed as:



Residual standard error = $\sqrt{\Sigma(y - \hat{y})^2/df}$

where:

- **y:** The observed value
- **ŷ:** The predicted value
- **df:** The degrees of freedom, calculated as the total number of observations – total number of model parameters.

The smaller the residual standard error the better the regression model fits the dataset.

1.5.1 Statistical performance of the smoker variable

The regression equation using the `smoker` variable is $medicalCost = 8434.3 + 23616smoker$

This implies that holding all other explanatory variables constant an increase in the smoker variable by 1 unit will lead medical costs to increase by USD 23,616. The p-value is less than 0.001, indicating a statistically significant relationship.

The adjusted R^2 for this model is 0.6195 and the residual squared error also stands at 7470. The R^2 value implies that about 62% of the variations in `medicalCost` is explained by the `smoker` variable. The adjusted R^2 of 0.6195 may also indicate that the model fits the data set quite well due to its proximity to 1.

```
summary(fit_smoker)

## 
## Call:
## lm(formula = medicalCost ~ smoker, data = dat_ins)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -19221  -5042   -919   3705  31720 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   8434.3      229.0   36.83   <2e-16 ***
## smoker       23616.0      506.1   46.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7470 on 1336 degrees of freedom
## Multiple R-squared:  0.6198, Adjusted R-squared:  0.6195 
## F-statistic:  2178 on 1 and 1336 DF,  p-value: < 2.2e-16
```

1.5.2 Statistical performance of the age variable

The regression equation using the age variable is $medicalCost = 3165.9 + 257.7age$

This implies that holding all other explanatory variables constant an increase in age by 1 year will lead medical costs to increase by USD 257.7. The p-value is less than 0.001, indicating a statistically significant relationship.

The adjusted R^2 for this model is 0.08872 and the residual squared error stands at a high of 11560. The R^2 value implies that about 9% of the variations in `medicalCost` is explained by the age variable. The adjusted R^2 of 0.08872 may also indicate that the model has a weak fit on the data set due to its proximity to 0 rather than 1.

```
summary(fit_age)

##
## Call:
## lm(formula = medicalCost ~ age, data = dat_ins)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8059  -6671  -5939   5440  47829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3165.9      937.1   3.378 0.000751 ***
## age            257.7       22.5  11.453  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11560 on 1336 degrees of freedom
## Multiple R-squared:  0.08941,    Adjusted R-squared:  0.08872
## F-statistic: 131.2 on 1 and 1336 DF,  p-value: < 2.2e-16
```

1.5.3 Statistical performance of age variable

The regression equation using the `bmi` variable is $medicalCost = 1192.94 + 393.87bmi$

This implies that holding all other explanatory variables constant an increase in bmi by 1 unit will lead medical costs to increase by USD 393.87. The p-value is less than 0.001, indicating a statistically significant relationship.

The adjusted R^2 for this model is 0.03862 and the residual squared error stand at its highest with a value of 11870. The R^2 value implies that about 4% of the variations in `medicalCost` is explained by the `bmi` variable. The adjusted R^2 of 0.03862 may also indicate that the model has the weakest fit on the data set in comparison to `smoker` and `age` due to its proximity to 0 rather than 1.

```
summary(fit_bmi)

## 
## Call:
## lm(formula = medicalCost ~ bmi, data = dat_ins)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -20956  -8118  -3757   4722  49442 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1192.94    1664.80   0.717    0.474    
## bmi           393.87      53.25   7.397 2.46e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 11870 on 1336 degrees of freedom
## Multiple R-squared:  0.03934,    Adjusted R-squared:  0.03862 
## F-statistic: 54.71 on 1 and 1336 DF,  p-value: 2.459e-13
```

1.6 Build two multivariate regression models with the three predictors above and with all the predictors in the dataset. Evaluate and compare the two models.

1.6.1 Multivariate regression with three best predictors

The regression equation using the three predictors is: $medicalCost = -11676.83 + 259.55age + 23823.68smoker + 393.87bmi$

This implies that holding all other explanatory variables constant an increase in age, `smoker` and `bmi` by 1 unit will lead medical costs to increase by USD 259.55, 23,823.68 and 393.87 respectively. The p-value is less than 0.001 for each predictor, indicating a statistically significant relationship.

```
# Multivariate regression with the three most correlated variables
fit_three_pred <- lm(medicalCost~age+smoker+bmi,data = dat_ins)
summary(fit_three_pred)

##
## Call:
## lm(formula = medicalCost ~ age + smoker + bmi, data = dat_ins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12415.4  -2970.9   -980.5   1480.0  28971.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11676.83     937.57  -12.45   <2e-16 ***
## age            259.55      11.93   21.75   <2e-16 ***
## smoker       23823.68     412.87   57.70   <2e-16 ***
## bmi            322.62      27.49   11.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 6092 on 1334 degrees of freedom
## Multiple R-squared:  0.7475, Adjusted R-squared:  0.7469
## F-statistic:  1316 on 3 and 1334 DF,  p-value: < 2.2e-16
```

1.6.2 Multivariate regression with all predictors

The regression equation using all predictors is: $medicalCost = -12898.59 + 960.05 regionnortheast + 607.09 regionnorthwest + 256.86 age + 339.19 bmi + 475.50 children + 23848.53 smoker - 74.97 regionsoutheast - 131.31 sex$

This implies that holding all other explanatory variables constant an increase in regionnortheast, regionnorthwest, age, bmi, children, smoker, regionsoutheast and sex by 1 unit will lead medical costs to increase by USD 960.05, 607.09, 256.86, 393.87, 475.50, 23848.53 and fall by 74.97 and 131.31 respectively. The p-value is less than 0.001 for age, bmi, children and smokers indicating a strong statistically significant relationship. The regionnortheast variable had a p-value of 0.044765 which is lower than 0.05. This is statistically significant but not as strong as the former variables. Sex, regionsoutheast and regionnorthwest had p values above 0.05 as such are not statistically significant. The coefficient of the regionsouthwest variable was not estimable so no statistical significance could be inferred. The absence of estimation may occur due to co-linearity with other variables and so including it in the model may not provide additional information.

```
# Multivariate regression with all the models
fit_all <- lm(medicalCost~.,data = dat_ins)
summary(fit_all)

##
## Call:
## lm(formula = medicalCost ~ ., data = dat_ins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
```

```
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -12898.59    1020.96 -12.634  < 2e-16 ***
## regionnortheast    960.05     477.93   2.009 0.044765 *
## regionnorthwest    607.09     477.20   1.272 0.203533
## regionsoutheast    -74.97     470.64  -0.159 0.873460
## regionsouthwest        NA         NA      NA       NA
## age                256.86      11.90  21.587  < 2e-16 ***
## sex               -131.31     332.95  -0.394 0.693348
## bmi                339.19      28.60  11.860  < 2e-16 ***
## children           475.50     137.80   3.451 0.000577 ***
## smoker           23848.53     413.15  57.723  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

1.6.3 Comparison of statistical results of models

```
data.frame(Model=c("Best three predictors","All predictors"),
           Adjusted_R_squred = c(0.7469 ,0.7494),
           Residual_standard_error = c(6092,6062 )) %>% kable()
```

Table 3-Comparison of results for multiple regression model.

Model	Adjusted_R_squred	Residual_standard_error
Best three predictors	0.7469	6092
All predictors	0.7494	6062

A comparison of both models produces similar statistical metrics as indicated in table 3. With regards to the model with the best three predictors, the adjusted R^2 was 0.7469 and the residual standard error was 6062. The R^2 value implies that about 75% of the variations in `medicalCost` is explained by the three predictor variables namely `smoker`, `age` and `bmi`. The adjusted R^2 of 0.7469 may also indicate that the model has a strong fit on the data set due to its proximity to 1.

With regards to the model which utilized all the predictors, significant differences in statistical results were not observed. The adjusted R^2 obtained was 0.7494 and the residual standard error was 6092. The R^2 value implies that about 75% of the variations in `medicalCost` was explained by all predictors. The adjusted R^2 of 0.7494 may also indicate that the model has a strong fit on the data set due to its proximity to 1.

Although there was a slight improvement in statistical performance metrics by using all the predictors, the adjusted R^2 obtained only showed an improved explanation in the variations within the dependent variable by a mere 0.3%.

```
# Improvement in Adjusted R_squared
imp_r2 <- (0.7494-0.7469)/0.7469 * 100
paste0(round(imp_r2,1),"%")

## [1] "0.3%"
```

The residual standard error also indicated a reduction in error by only 0.5%

```
rse <- (6062-6092)/6092 * 100
paste0(round(rse,1),"%")

## [1] "-0.5%"
```

1.7 State your overall conclusions for this task.

In linear regression, the degree of correlation of independent variables to the dependent variables is significant for predictive ability. This study discovered a direct relationship between the correlation of an independent and its ability to explain the variations in the given dependent variable. From table 4, it can be inferred that as the degree of correlation fell from 0.67 in the smoker variable to 0.20 in the age variable, the adjusted R^2 also fell from 62% down to 4%. The residual standard error also worsened as the correlation weakened, increasing from 7470 to 11870. Table 3 also illustrated that using the most correlated variables led to results that were not statistically different from using all independent variables in the model.

```
# make row names part of data frame
cor_df <- rownames_to_column(cor_sorted_df,'Predictor')

# select top three most correlated
cor_df_three <-cor_df[c(2,3,4),]

# create a data frame for adjusted r squared and residual standard error
r2_rse <- data.frame(Adjusted_R_squared=c(0.6195,0.08872,0.03862),
                     Residual_standard_error =c(7470,11560 ,11870))

cor_df_three_stats <- cbind(cor_df_three,r2_rse) # bind data frames
# round correlations to 2 decimals
cor_df_three_stats$Correlation <- round(cor_df_three_stats$Correlation,2)

# round adjusted r squared to 2 decimals, convert to per cent and add the % s
ign
cor_df_three_stats$Adjusted_R_squared<- paste0(round(cor_df_three_stats$Adjus
ted_R_squared,2)*100,'%')

# show in table form
cor_df_three_stats %>% kable()
```

Table 4-Correlation, Adjusted R-Squared and Residual standard error

	Predictor	Correlation	Adjusted_R_squared	Residual_standard_error
2	smoker	0.67	62%	7470
3	age	0.30	9%	11560
4	bmi	0.20	4%	11870

As such in the absence of multicollinearity, the most correlated independent variables to a dependent variable may provide significantly sufficient information for the regression modelling process. Independent variables with weaker correlation may be removed to save computational resources and time.