MEDICAL COST PERSONAL DATASETS

CONTENTS

01 INTRODUCTION

Data Overview

02 DATA VISUALIZATION

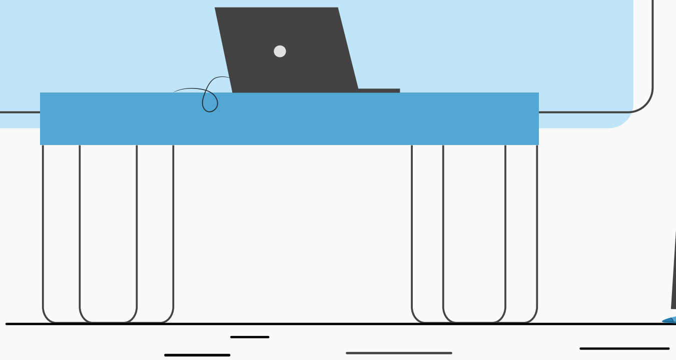
03 MODEL

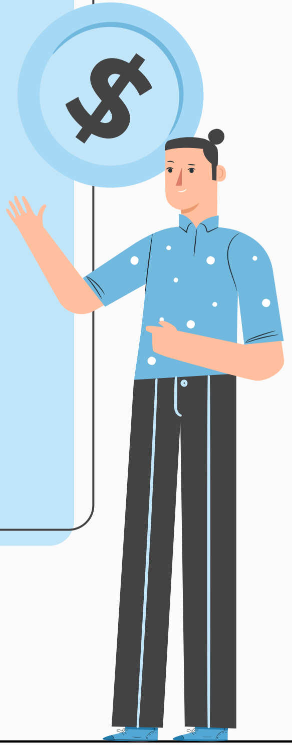
Multivariate linear regression
K-nearest neighbors (k-NN)

04 HYPOTHESIS TESTING

INTRODUCTION

The dataset named Medical Cost Personal Dataset from **Kaggle**. This dataset has a large number of clients from insurance companies of the USA and multiple personal information for each client such as age, sex, body mass index(BMI), number of children, smoker/non-smoker, residential area. We are going to see the relations between the charges (The charges are an important point of estimation for any insurance company.) and some interesting attributes as well as some criteria these companies have.



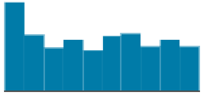
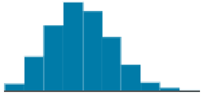
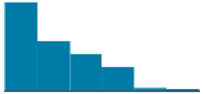
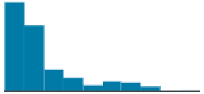


kaggle

Data Overview: insurance.csv

About this file

This dataset consists of 1338 rows.

# age	A sex	# bmi	# children	✓ smoker	A region	# charges
Edad del asegurado	Género	Indice de masa corporal	Número de hijos	Indicador si fuma	Región donde vive el asegurado	Prima del seguro
 18 64	male 51% female 49%	 16 53.1	 0 5	true 0 0% false 0 0%	southeast 27% southwest 24% Other (649) 49%	 1.12k 63.8k
19	female	27.9	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.5523
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47061
32	male	28.88	0	no	northwest	3866.8552
31	female	25.74	0	no	southeast	3756.6216
46	female	33.44	1	no	southeast	8240.5896
37	female	27.74	3	no	northwest	7281.5056
37	male	29.83	2	no	northeast	6406.4107
60	female	25.84	0	no	northwest	28923.13692

https://www.kaggle.com/mirichoi0218/insurance

02.

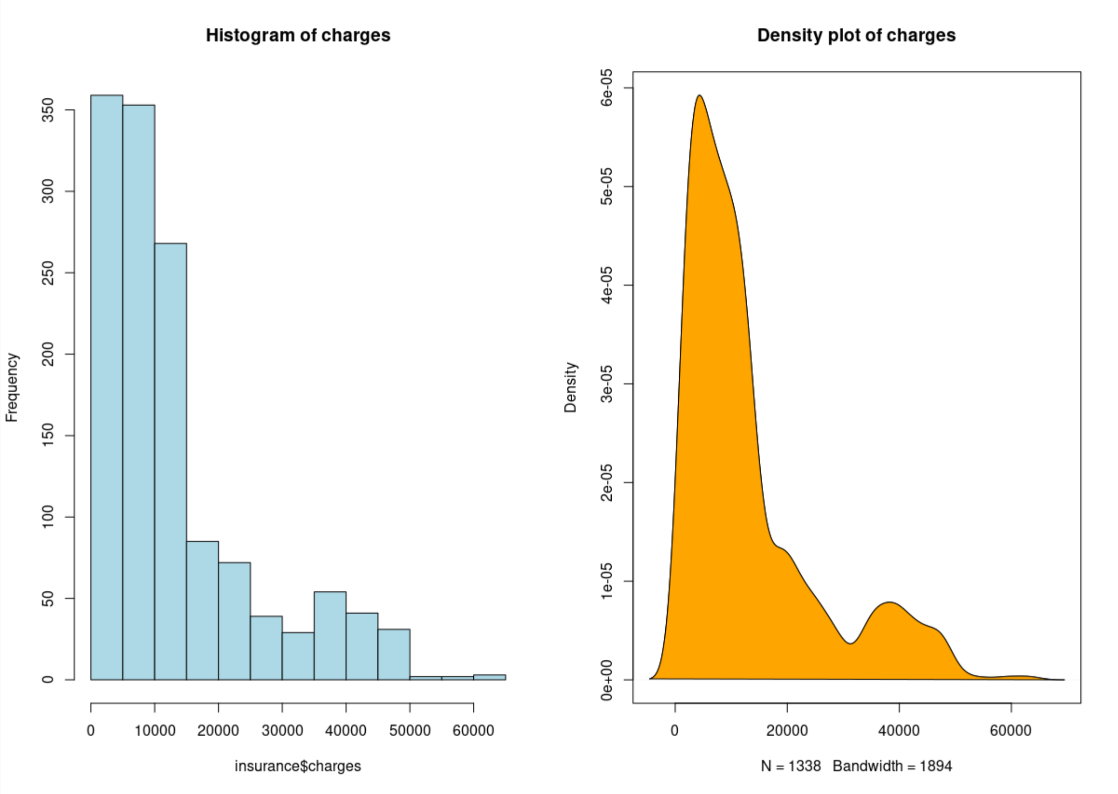
DATA VISUALIZATION

Data Visualization

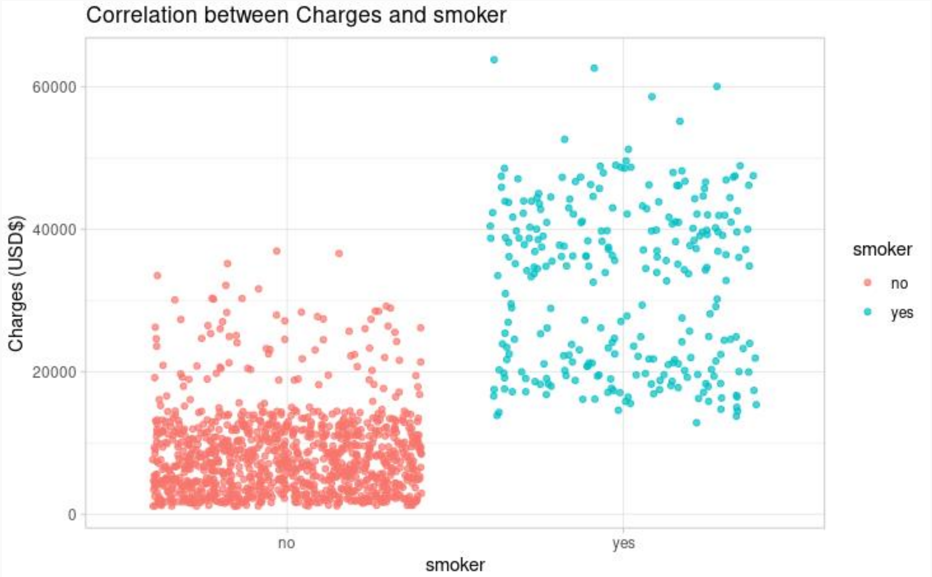
Response (dependent) variable: charges

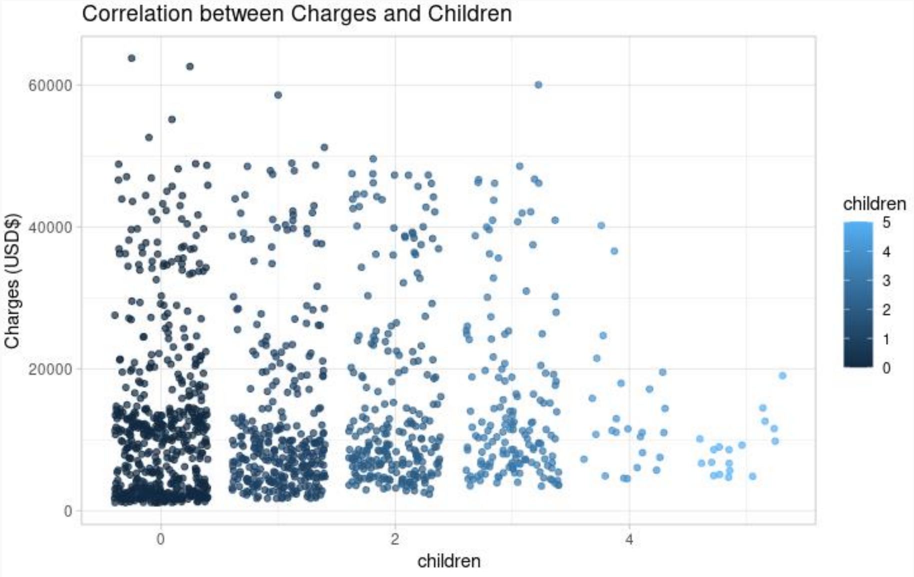
```
summary(insurance$charges)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 1122    4740    9382   13270   16640   63770 
```

Because the mean value is greater than the median, this implies that the distribution of insurance expenses is right-skewed

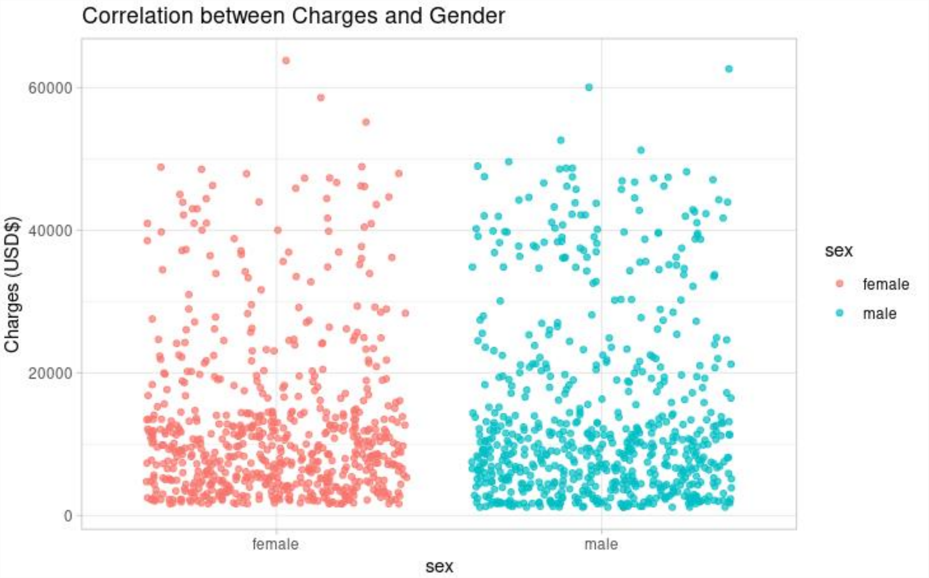


Data Visualization

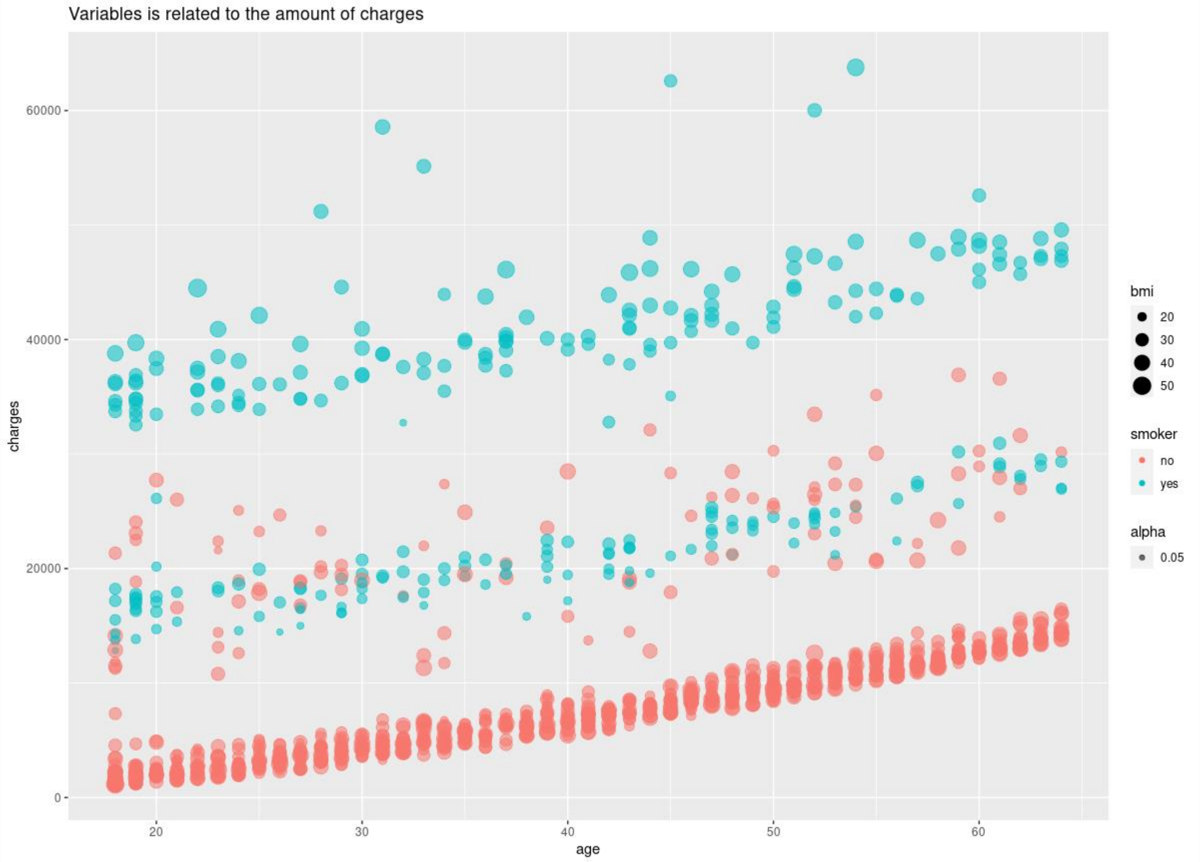




Data Visualization



Data Visualization



03.

MODEL

Multivariate linear regression

K-nearest neighbors (k-NN)

MULTIVARIATE LINEAR REGRESSION

charges ~ age + sex + bmi + children + smoker + region

```
> # lm all column
> mul_model <- lm(charges ~ age + sex + bmi + children + smoker + region, data = insuran
ce)
> # mul_model_ <- lm(charges ~ ., data = insurance_new)
> summary(mul_model)
```

```
Call:
lm(formula = charges ~ age + sex + bmi + children + smoker + 
    region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-11304.9  -2848.1   -982.1   1393.9  29992.8 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348    
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smoker           23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769    
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *  
regionsouthwest   -960.0      477.9  -2.009 0.044765 *  
---
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
```

MULTIVARIATE LINEAR REGRESSION

charges ~ age + bmi + children + smoker

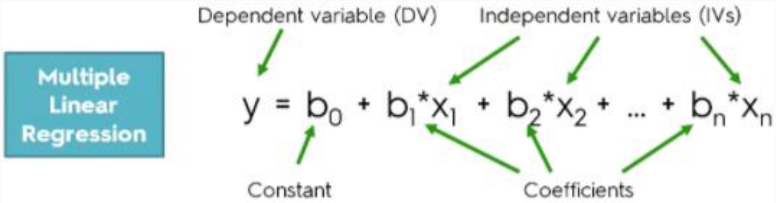
```
> # column that have significance
> mul_model <- lm(charges ~ age + bmi + children + smoker, data = insurance)
> # mul_model_ <- lm(charges ~ ., data = insurance_new)
> summary(mul_model)
```

```
Call:
lm(formula = charges ~ age + bmi + children + smoker, data
= insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-11897.9  -2920.8   -986.6   1392.2  29509.6 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
age            257.85      11.90  21.675  < 2e-16 ***
bmi            321.85      27.38  11.756  < 2e-16 ***
children       473.50     137.79   3.436 0.000608 ***
smoker       23811.40     411.22  57.904  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1
```

```
Residual standard error: 6068 on 1333 degrees of freedom
Multiple R-squared:  0.7497,	Adjusted R-squared:  0.7489 
F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16

> sqrt(0.7489)
[1] 0.8653901
```

MULTIVARIATE LINEAR REGRESSION

The effect of changing one predictor variable while controlling the values of the other predictor variables.




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
age            257.85      11.90  21.675  < 2e-16 ***
bmi            321.85      27.38  11.756  < 2e-16 ***
children       473.50     137.79   3.436 0.000608 ***
smoker       23811.40     411.22  57.904  < 2e-16 ***
```

$$Y = -12102.77 + 257.85(\text{age}) + 321.85(\text{bmi}) + 473.50(\text{Children}) + 23811.40(\text{smoker})$$

```
insurance$Multi <- -12102.77 + 257.85*insurance$age + 321.85*insurance$bmi +
  473.50*insurance$children + 23811.40*insurance$smoker

insurance$prediction <- predict(mul_model, newdata = insurance)
```

	age	bmi	children	smoker	charges	Multi	prediction
1	19	27.900	0	1	16884.924	25587.3950	25587.4252
2	18	33.770	1	0	1725.552	3880.9045	3880.9459
3	28	33.000	3	0	4449.462	7158.5800	7158.6201
4	33	22.705	0	0	21984.471	3713.8843	3713.9005
5	32	28.880	0	0	3866.855	5443.4580	5443.4834

MULTIVARIATE LINEAR REGRESSION

K-NEAREST NEIGHBORS (K-NN)

sampling approach 2 variables: Smoker (n=274) and Non-Smoker (n=1064)

smoker	count	min	median	max	IQR
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
yes	274	12829.	34456.	63770.	20193.
no	1064	1122.	7345.	36911.	7376.

Population (n=1338): set test 20% (n sample = 270) of smoker and non-smoker.
set train: 80% (n sample = 1068).

```
set.seed(612)
test_no <- sample_n(smoker_no, 135, fac = "ID")$ID
test_yes <- sample_n(smoker_yes, 135, fac = "ID")$ID
test <- c(test_no,test_yes)
```

```
# keep just the test data points/rows
all_test <- df[test,-1]
all_train <- df[-(test), -1]
```

K-NEAREST NEIGHBORS (K-NN)

Check the data structure

```
> str(all_train)
'data.frame':	1198 obs. of  7 variables:
 $ age     : int  18 28 32 31 46 37 37 60 25 62 ...
 $ sex     : chr  "male" "male" "male" "female" ...
 $ bmi     : num  33.8 33 28.9 25.7 33.4 ...
 $ children: int  1 3 0 0 1 3 2 0 0 0 ...
 $ smoker  : chr  "no" "no" "no" "no" ...
 $ region  : chr  "southeast" "southeast" "northwest"
 $ charges : num  1726 4449 3867 3757 8241 ...
```

```
> str(all_test)
'data.frame':	140 obs. of  7 variables:
 $ age     : int  19 43 23 31 49 18 20 41 34 53 ...
 $ sex     : chr  "male" "male" "female" "male" ...
 $ bmi     : num  20.6 26 28.1 31.1 34.8 ...
 $ children: int  2 0 0 3 1 0 1 1 1 1 ...
 $ smoker  : chr  "no" "no" "no" "no" ...
 $ region  : chr  "northwest" "northeast" "northwest"
 $ charges : num  2804 6837 2690 5425 9584 ...
```

change the "smoker" variable to be of a factor type

```
# Change the "smoker" variable to be of a factor type
all_train$smoker <- as.factor(all_train$smoker)
all_test$smoker <- as.factor((all_test$smoker))
```

```
> str(all_train)
'data.frame':	1198 obs. of  7 variables:
 $ age     : int  18 28 32 31 46 37 37 60 25 62 ...
 $ sex     : chr  "male" "male" "male" "female" ...
 $ bmi     : num  33.8 33 28.9 25.7 33.4 ...
 $ children: int  1 3 0 0 1 3 2 0 0 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 1 1 1 1 1
 $ region  : chr  "southeast" "southeast" "northwest"
 $ charges : num  1726 4449 3867 3757 8241 ...
```

```
> str(all_test)
'data.frame':	140 obs. of  7 variables:
 $ age     : int  19 43 23 31 49 18 20 41 34 53 ...
 $ sex     : chr  "male" "male" "female" "male" ...
 $ bmi     : num  20.6 26 28.1 31.1 34.8 ...
 $ children: int  2 0 0 3 1 0 1 1 1 1 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 1 1 1 1 1
 $ region  : chr  "northwest" "northeast" "northwest"
 $ charges : num  2804 6837 2690 5425 9584 ...
```

```
> # see the model's details
> model_knn

Call:
train.kknn(formula = smoker ~ ., data = all_train, kmax = 9)

Type of response variable: nominal
Minimal misclassification: 0.04307116
Best kernel: optimal
Best k: 1
```

```
> # Do a prediction on the test data
> prediction <- predict(model_knn, all_test[, -5])
> prediction
  [1] no  no  no  no  no  no  no  no  no  no  no  no  no  no  no  no  no  no  no
 [20] no  no  no  no  no  no  no  yes no  no  no  yes no  no  no  no  no  no  no
 [39] no  no  no  no  yes no  no  no  no  no  no  no  no  yes no  no  no  no  no
 [58] no  no  no  no  no  no  no  no  no  no  no  no  no  yes no  no  no  no  no
 [77] no  no  no  no  no  no  no  no  no  no  no  no  no  no  no  yes no  no  no
 [96] no  no  no  no  no  no  no  no  no  no  no  no  no  no  no  no  no  no  yes
[115] no  no  no  no  no  yes no  no  no  no  no  no  no  no  no  no  no  no  no
[134] no  no  yes yes no  yes yes yes yes yes no  yes yes yes yes yes yes yes yes
[153] yes yes yes yes yes yes yes no  yes yes yes yes no  yes no  no  yes yes yes
[172] yes no  no  yes yes yes yes yes no  no  yes no  yes yes yes yes no  yes yes
[191] yes no  yes yes yes yes yes yes yes yes yes yes no  yes no  no  yes no  yes
[210] yes no  yes yes yes yes yes yes yes yes yes yes yes no  yes yes yes yes no
[229] yes yes yes yes yes yes yes yes yes yes yes no  no  yes yes yes no  yes yes
[248] yes yes yes yes yes no  no  no  yes yes yes yes yes yes no  yes yes yes yes
[267] yes yes yes no
Levels: no yes
```

K-NEAREST NEIGHBORS (K-NN)

```
> #Display results
> solution
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  127  28
       yes   8 107

               Accuracy : 0.8667
                 95% CI : (0.8202, 0.9048)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7333

 Mcnemar's Test P-Value : 0.001542

            Sensitivity : 0.9407
            Specificity : 0.7926
         Pos Pred Value : 0.8194
         Neg Pred Value : 0.9304
             Prevalence : 0.5000
         Detection Rate : 0.4704
   Detection Prevalence : 0.5741
      Balanced Accuracy : 0.8667

       'Positive' Class : no
```

	NO	YES
NO	127	8
YES	28	107

Accuracy = 0.867

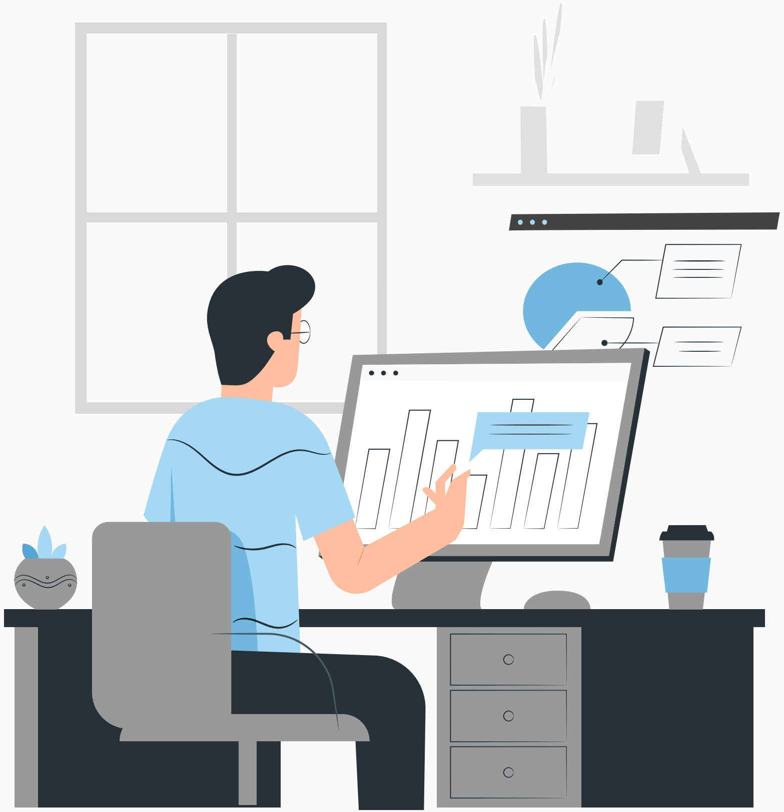
Precision = 0.819

Sensitivity = 0.941

Specificity = 0.793

04.

HYPOTHESIS TESTING



HYPOTHESIS TESTING

Exploratory data analysis has indicated that smoking has an effect on charges.

```
> # Read in our dataset
> df <- read.csv("insurance.csv")
> df %>%
+   group_by(smoker) %>%
+   summarise(
+     count = n(),
+     median = median(charges),
+     mean = mean(charges),
+     SD = sd(charges),
+     Var = var(charges)
+   ) %>%
+   arrange(desc(median))
# A tibble: 2 × 6
  smoker count median   mean     SD        Var
  <chr>  <int>  <dbl>  <dbl>  <dbl>      <dbl>
1 yes      274 34456. 32050. 11542. 133207311.
2 no      1064  7345.  8434.  5994.  35925420.
```

HYPOTHESIS TESTING

Step 1: Define null and alternative hypothesis

H0: $\mu 1 - \mu 2 = 0$ The average charges of smokers is equal to non-smokers

Ha: $\mu 1 - \mu 2 > 0$ The average charges of smokers is greater than non-smokers

Test at the 5% level of significance: $\alpha = 0.05$

Smoker: yes
n: 274
Mean: 32050
SD: 11541.55
Variance: 133207311

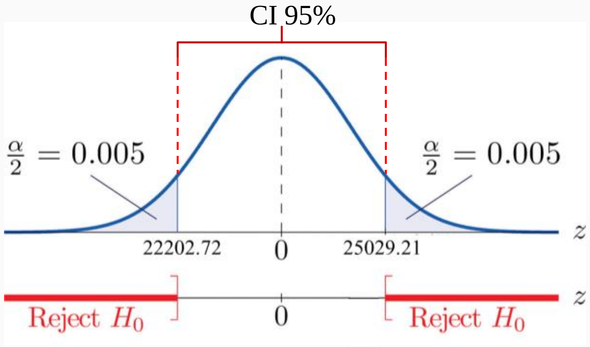
Smoker: no
n: 1064
Mean: 8434
SD: 5993.782
Variance: 35925420

HYPOTHESIS TESTING

Step 2: Confidence Intervals Two-sample hypothesis

```
z.test(df$charges[df$smoker== "yes"], df$charges[df$smoker== "no"],
       mu = 0, sigma.x = sd(df$charges[df$smoker== "yes"]),
       sigma.y = sd(df$charges[df$smoker== "no"]),conf.level = 0.95)
```

```
        Two-sample z-Test

data:  df$charges[df$smoker == "yes"] and df$charges[df$smoker == "no"]
z = 32.752, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 22202.72 25029.21
sample estimates:
mean of x mean of y
32050.232  8434.268
```



We are 95% confident that the difference in the population means lies in the interval [22202.72 , 25029.21]

Step 3: Since the samples are independent and both are large the test statistic is

Where D_0 = hypothesized difference between the means

Test Statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sigma_{(\bar{x}_1 - \bar{x}_2)}} \quad \text{where} \quad \sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Rejection region: $z < -z_\alpha$
[or $z > z_\alpha$ when H_a: $(\mu_1 - \mu_2) > D_0$]

Rejection region: $|z| > z_{\alpha/2}$

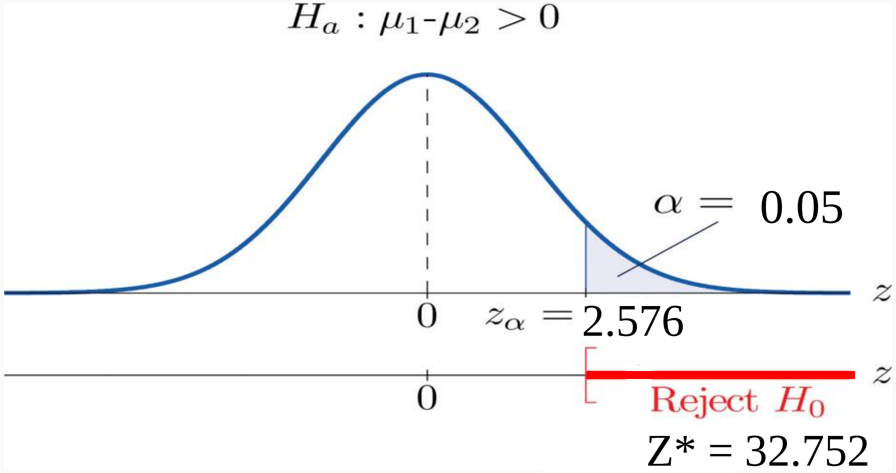
Step 4: Inserting the data into the formula for the test statistic gives

Test Statistic:

$$Z^* = \frac{(32050 - 8434) - 0}{\sqrt{\frac{133207311}{274} + \frac{35925420}{1064}}}$$

$$= \frac{23616}{\sqrt{486158.1 + 33764.49}} = 32.752$$

Step 5: Since the symbol in Ha is “>” this is a right-tailed test, so there is a single critical value, zα=0.05, which from the last line in Figure we read off as 2.576. The rejection region is [2.576,∞)



Step 6: As shown in Figure "Rejection Region and Test Statistic for " the test statistic falls in the rejection region. We reject the null hypothesis and can conclude that people who smoke have on an average larger medical claim compared to people who don't smoke.

Step 7: Find P-Value

Test Statistic: $$Z = \frac{(32050 - 8434) - 0}{\sqrt{\frac{133207311}{274} + \frac{35925420}{1064}}}$$

$$= \frac{23616}{\sqrt{486158.1 + 33764.49}} = 32.752$$

Step 8: The observed significance or *p*-value of the test is the area of the right tail of the standard normal distribution that is cut off by the test statistic $Z = 32.752$. The number 5.684 is too large to appear in Z-table the area of the *right* tail, is therefore 1-1.0000 = 0.0000 (The actual value is approximately 0.00000000000000022 or 2.2e^-16)

```
        Two-sample z-Test

data:  smoker_yes$charges and smoker_no$charges
z = 32.752, p-value < 2.2e-16
alternative hypothesis: true difference in means is n
ot equal to 0
95 percent confidence interval:
 22202.72 25029.21
sample estimates:
mean of x mean of y
32050.232  8434.268
```

Step 9: Since 2.2e^-16 < 0.05(2.576), p -value < α so the decision is to reject the null hypothesis : The data provide sufficient evidence, at the 5% level of significance, to conclude that the mean charges satisfaction for smoker is higher that for non-smoker.