Statistical Analysis of Medical Cost Dataset

Ata Deniz Arslaner

INTRODUCTION

This paper presents various applications on a given insurance dataset in R. The main purpose of this project is to thoroughly construct the relationships between features and target as well as test a variety of hypotheses to accept or reject them in order to better understand the links between variables in the dataset.

First we imported the dataset ,after that dataset is viewed and checked whether there are duplicated or null values.After the summary of dataset , the visualizations are presented and various Tests are applied and conclusions are made afterwards.

```
#Data Importing
data<-read.csv("C:/Users/Alperitoo/OneDrive/Masaüstü/RPROJE/deneme/insurance.csv")
#Data Analyses
View(data)
#Check duplicated and nulls
sum(duplicated(data))
sum(is.na(data))
summary(data)
```

DATA DESCRIPTION

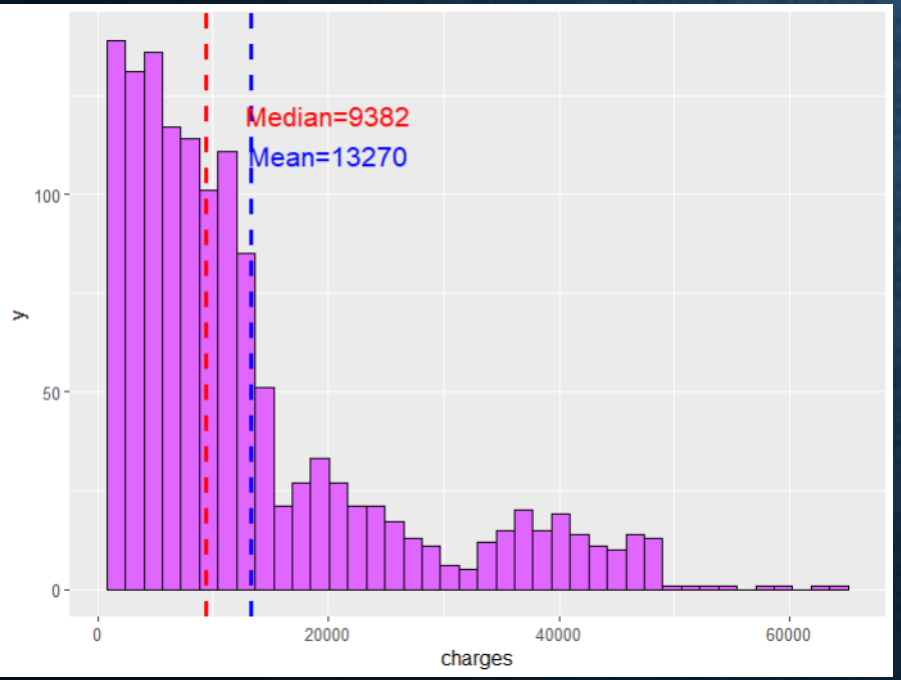
Dataset is been collected from website called "Kaggle" with title "Medical Cost Personal Dataset". This dataset consists of 1338 rows and 7 columns.File name is 'insurance.csv

Summary of the dataset with "summary()" function:

```
> summary(data)
      age            sex                 bmi           children        smoker
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000   Length:1338
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000   Class :character
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character
 Mean   :39.21                      Mean   :30.66   Mean   :1.095
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
 Max.   :64.00                      Max.   :53.13   Max.   :5.000
    region             charges
 Length:1338        Min.   : 1122
 Class :character   1st Qu.: 4740
 Mode  :character   Median : 9382
                    Mean   :13270
                    3rd Qu.:16640
                    Max.   :63770
```

VISUALIZATIONS

Charges Distribution:



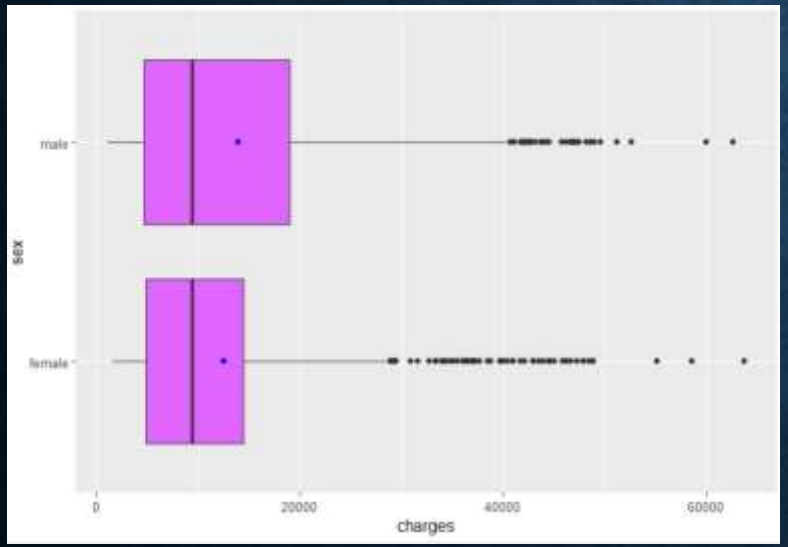
R code:

```
mean(data$charges)
median(data$charges)
cplot<-ggplot(data, aes(x=charges))+
  geom_histogram(color="black", fill="mediumorchid1", bins=40)+
  geom_vline(aes(xintercept= 13270), color="blue", linetype="dashed", size=1)+
  geom_vline(aes(xintercept= 9382), color="red", linetype="dashed", size=1)+
  annotate("text", x= 20000, y=110, size=5, label="Mean=13270", color="blue")+
  annotate("text", x= 20000, y=120, size=5, label="Median=9382", color="red")
```

Mean and median values of “charges” column is calculated and mentioned in the plot. As it can be seen in the figure, distribution is Right skewed and Mean value is greater than the median value.

CHARGES IN TERMS OF SEX

Plot



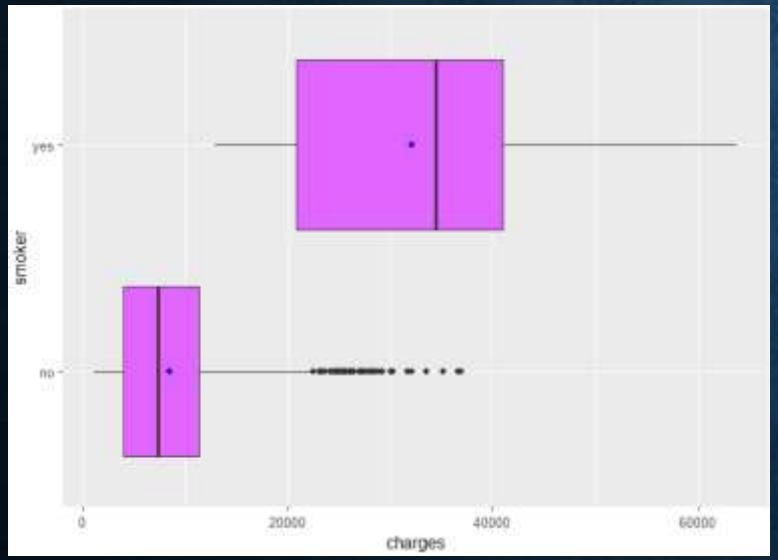
R CODE

```
#Plotting Charges according to Sex
chrsx<-ggplot(data, aes(x=sex, y=charges)) +
  geom_boxplot(fill="mediumorchid1")+
  stat_summary(fun=mean, geom="point", color="blue")+
  coord_flip()
```

Blue points are representing the mean value.

CHARGES IN TERMS OF SMOKING STATUS

Plot

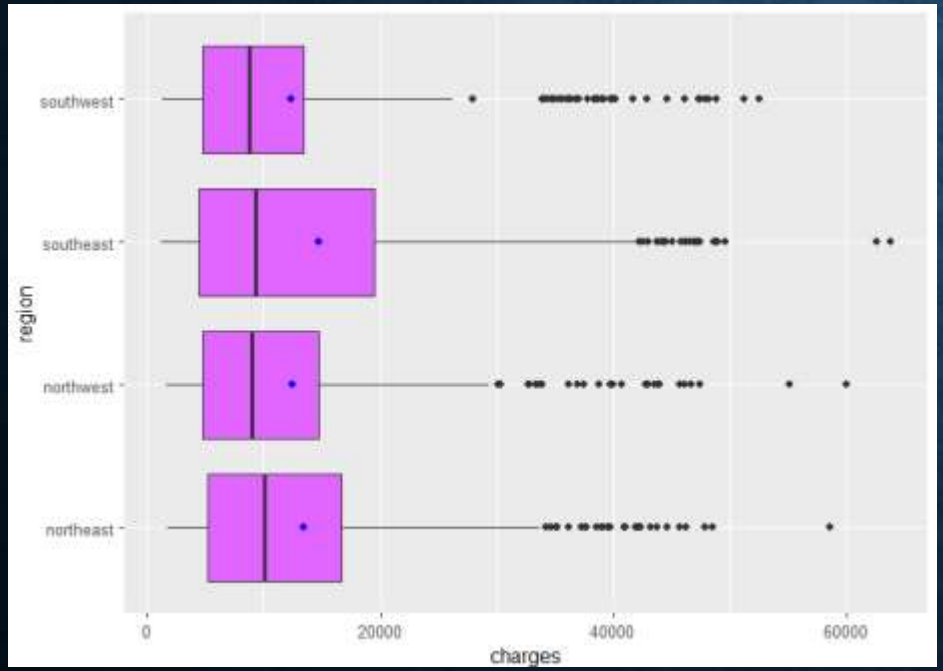


R CODE

```r
#Plotting Charges according to smoking status
chrsmk<-ggplot(data, aes(x=smoker, y=charges)) +
  geom_boxplot(fill="mediumorchid1")+
  stat_summary(fun=mean, geom="point", color="blue")+
  coord_flip()
#Plotting Charges according to regions
```

CHARGES IN TERMS OF REGIONS

Plot

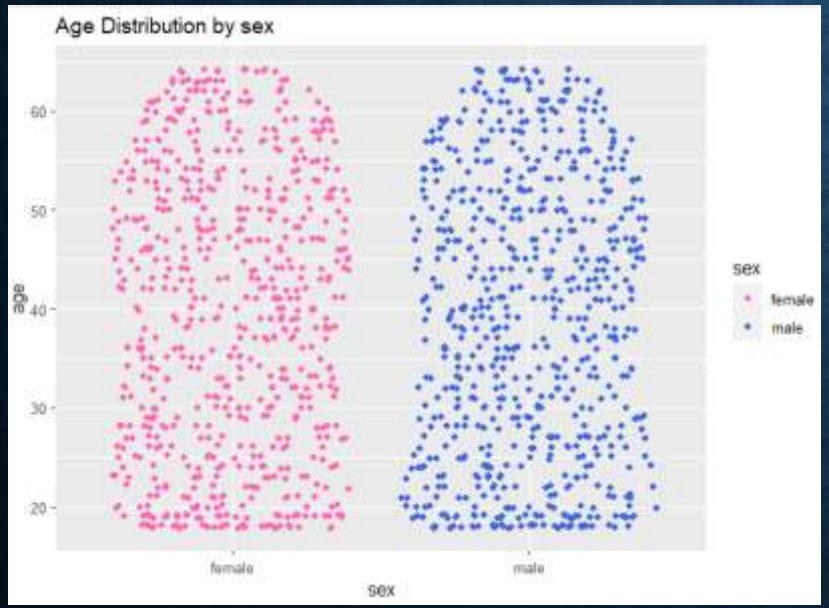


R CODE

```
chrreg<-ggplot(data, aes(x=region, y=charges)) +
  geom_boxplot(fill="mediumorchid1")+
  stat_summary(fun=mean, geom="point", color="blue")+
  coord_flip()
```

AGE DISTRIBUTION BY SEX

Plot

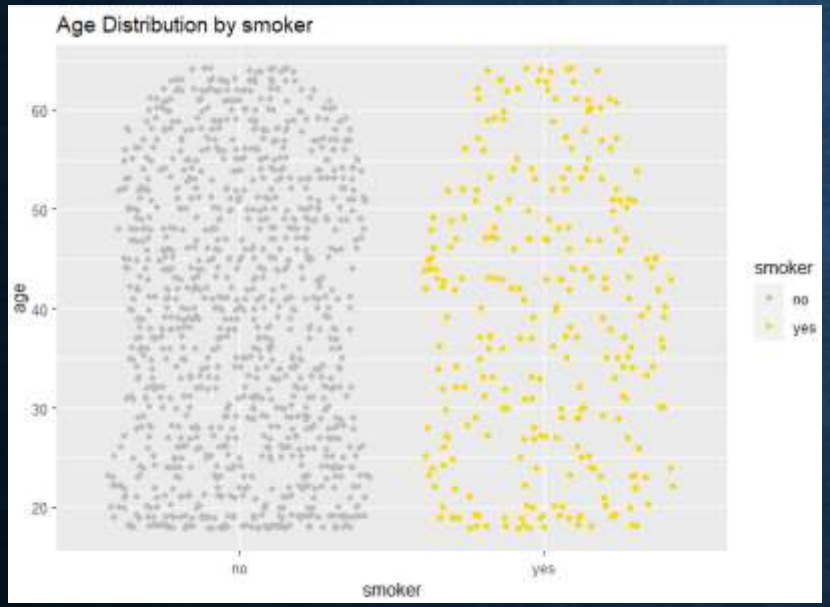


R CODE

```
agesx<-ggplot(data, aes(x=sex, y=age, color=sex)) +
  geom_sina()+
  scale_color_manual(values=c('hotpink', "royalblue"))+
  labs(title="Age Distribution by sex")
```

It can be observed that the age distribution of males and females are similar to each other.

AGE DISTRIBUTION BY SMOKING STATUS

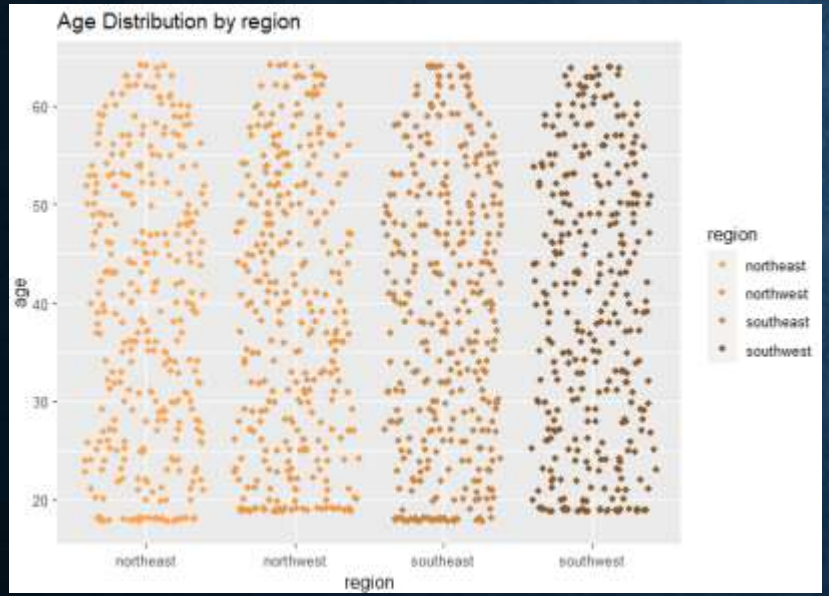


R CODE

```
agesm<-ggplot(data, aes(x=smoker, y=age, color=smoker)) +
  geom_sina()+
  scale_color_manual(values=c('grey', "gold"))+
  labs(title="Age Distribution by smoker")
```

It can be observed that number of nonsmokers are greater than smokers.

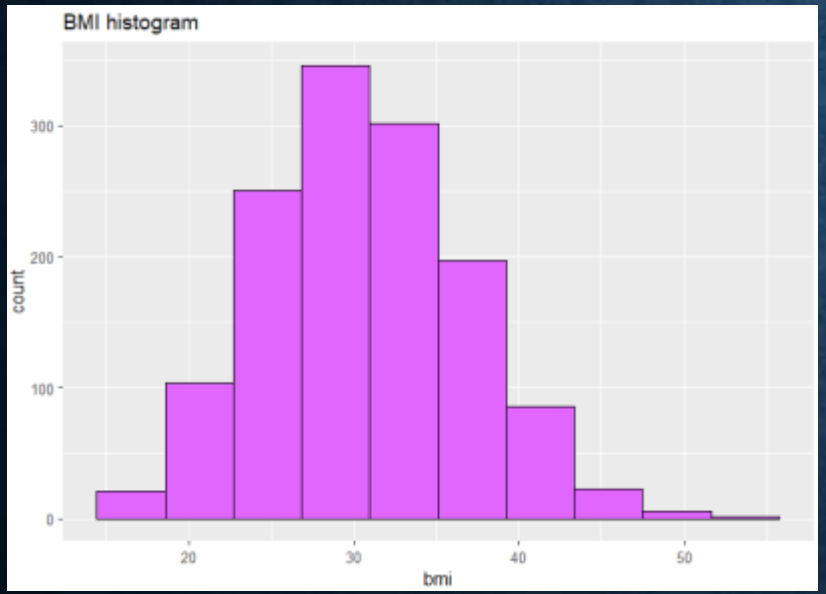
AGE DISTRIBUTION BY REGIONS



R CODE

```r
#Plotting age distribution by smokers
agesm<-ggplot(data, aes(x=smoker, y=age, color=smoker)) +
  geom_sina()+
  scale_color_manual(values=c('grey', "gold"))+
  labs(title="Age Distribution by smoker")
```

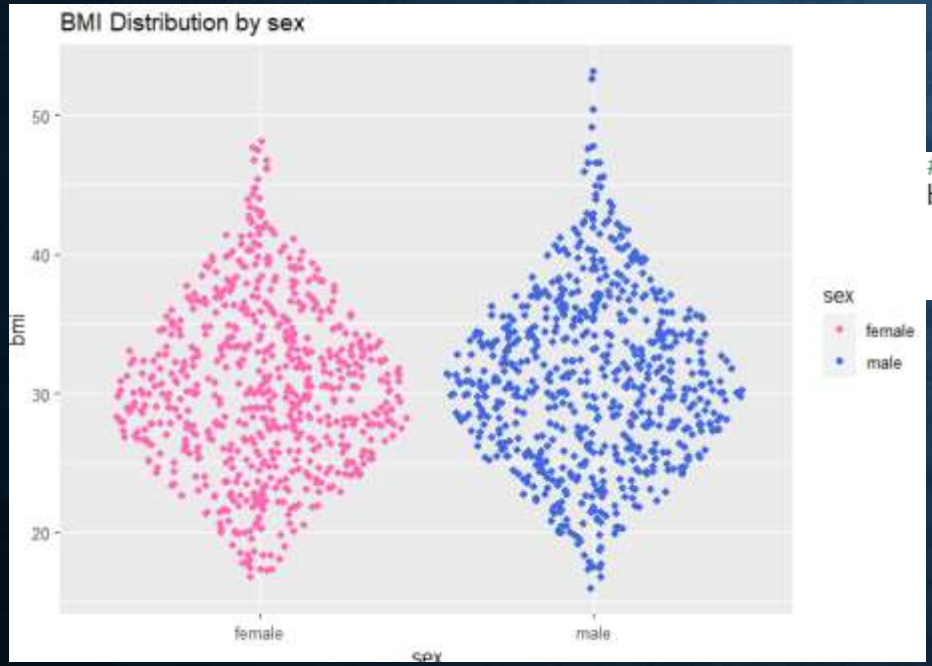
BODY MASS INDEX DISTRIBUTION (BMI)



R CODE

```r
#BMI Histogram
bmihist<-ggplot(data, aes(x=bmi))+
  geom_histogram(color="black", fill="mediumorchid1", bins=10)+
  labs(title="BMI histogram")
```

BMI DISTRIBUTION BY SEX

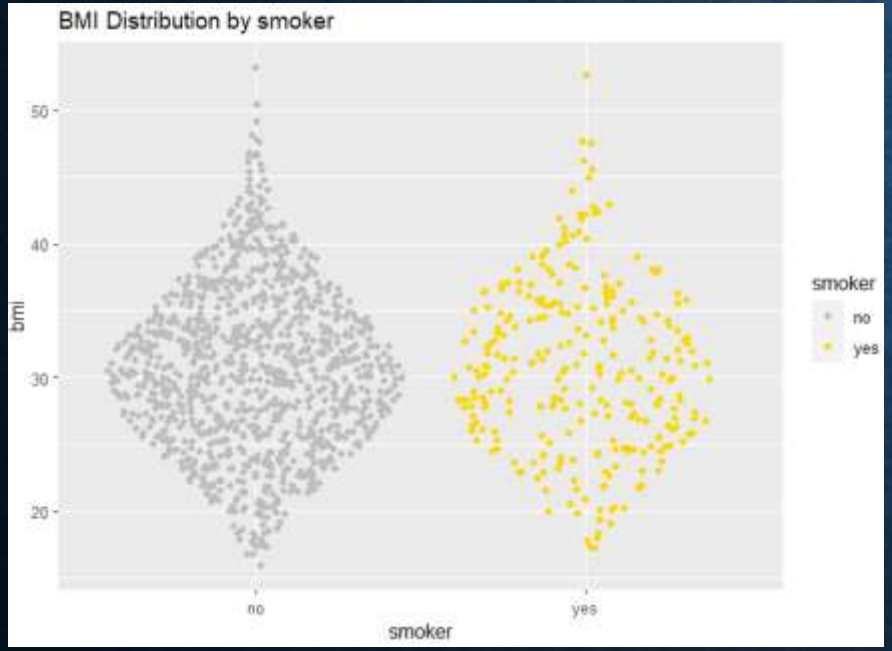


R CODE

```r
#Bmi distribution by sex
bmisx<-ggplot(data, aes(x=sex, y=bmi, color=sex)) +
  geom_sina()+
  scale_color_manual(values=c('hotpink', "royalblue"))+
  labs(title="BMI Distribution by sex")
```

It can be observed that maximum bmi value of females is below 50.Maximum bmi value of males is greater than 50.Also it can be observed that the among people with bmi value less than 20 , females are outnumbered by men.

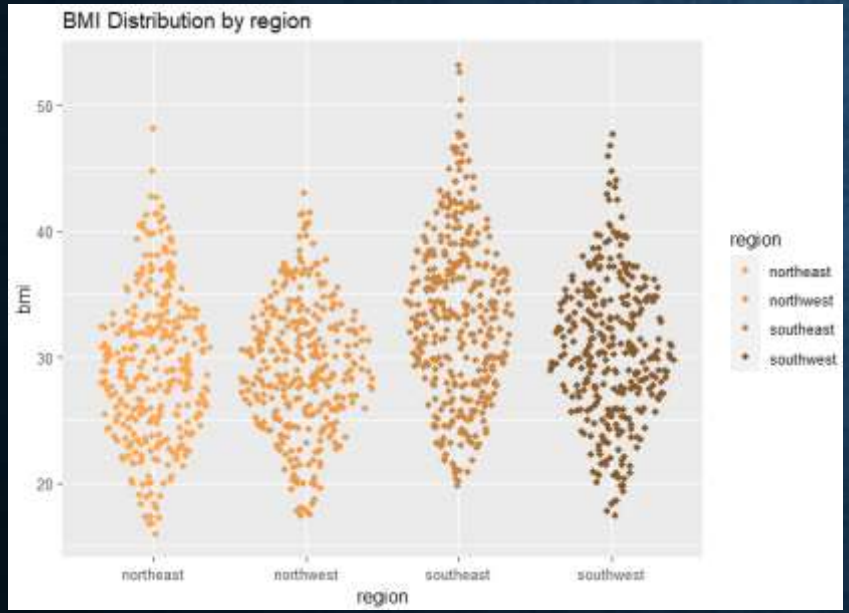
BMI DISTRIBUTION BY SMOKING STATUS



R CODE

```
#Bmi distribution by smoker
bmismk<-ggplot(data, aes(x=smoker, y=bmi, color=smoker)) +
  geom_sina()+
  scale_color_manual(values=c('grey', "gold"))+
  labs(title="BMI Distribution by smoker")
```

BMI DISTRIBUTION BY REGIONS



R CODE

```r
#Bmi distribution by region
bmireg<-ggplot(data, aes(x=region, y=bmi, color=region)) +
  geom_sina()+
  scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4"))+
  labs(title="BMI Distribution by region")
```

T-TEST EXAMPLES

1-) Two sided One Sample T-test

Is the mean of BMI equal to 30 or statistically different from 30?

- H0: mean = 30
- H1: mean != 30
- 95% confidence interval: 30.33635 - 30.99045
- sample estimates:

mean of BMI = 30.6634

- **t** = 3.9792
- **p-value** = 7.284e-05

Conclusion: A one sample t-test was conducted to find if the mean of BMI is equal to 30 or not. We got a p-value that was much lower than alpha (α = 0.05). So we reject the null hypothesis being that the mean is equal to 30 and accept the alternative hypothesis being the mean is not equal to 30.

R CODE AND CONSOLE OUTPUT

```
> mean(data$bmi)
[1] 30.6634
> t.test(data$bmi,mu=30,alternative="two.sided",conf.level = 0.95)

        One Sample t-test

data:  data$bmi
t = 3.9792, df = 1337, p-value = 7.284e-05
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 30.33635 30.99045
sample estimates:
mean of x
  30.6634
```

2-)TWO SAMPLE T-TEST

Do smokers pay more insurance charges?
- **H0:** smokers pay more charges
- **H1:** smokers don't pay more charges
- 95% confidence interval: -25034.71 - -22197.21
- sample estimates:
mean in group no = 8434.268
mean in group yes = 32050.232
- **t** = -32.752
- **p-value** < 2.2e-16

Conclusion: A Welch two sample t-test was conducted to see if individuals who smoke pay more for the insurance. The p-value is very low compared to our defined alpha value (α =0.05). So we reject the null hypothesis and accept the alternative hypothesis and can conclude that smokers pay more insurance charges than nonsmokers.

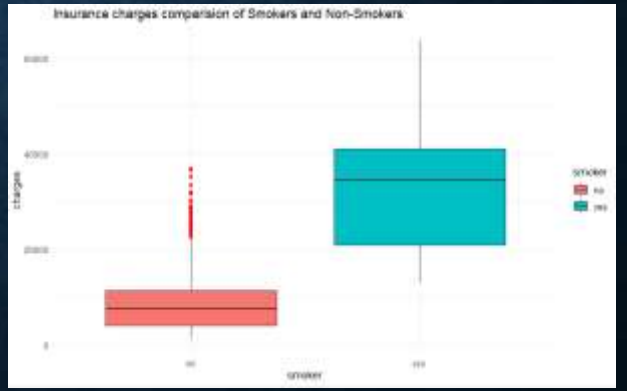
R CODE AND CONSOLE

```
> t.test(charges~smoker,data,conf_level=0.95)

        Welch Two Sample t-test

data:  charges by smoker
t = -32.752, df = 311.85, p-value < 2.2e-16
alternative hypothesis: true difference in means between group no and
 group yes is not equal to 0
95 percent confidence interval:
 -25034.71 -22197.21
sample estimates:
 mean in group no mean in group yes
         8434.268         32050.232
```

PLOT



F-TEST EXAMPLE

Is the male BMI mean equal to the female BMI mean?

- **H0**: male BMI = female BMI
- **H1**: male BMI != female BMI
- **Confidence level**: 95%: 0.8861438 - 1.2004937
- **F** = 1.0315,
- **p-value** = 0.6892
- sample estimates:

ratio of variances = 1.031475

Conclusion: We conducted an F test to compare two variances. Here, the null hypothesis is that they have the same variance. The p-value being greater than the significance level of 0.05 allows us to accept the null hypothesis of two normal samples with same variance.

We also used t-test for this hypothesis….

R CODE AND CONSOLE OUTPUT

```
> var.test(data[which(data$sex=="male"), "bmi"], data[which(data$sex=
="female"), "bmi"],conf.level = 0.95)

        F test to compare two variances

data:  data[which(data$sex == "male"), "bmi"] and data[which(data$sex
 == "female"), "bmi"]
F = 1.0315, num df = 675, denom df = 661, p-value = 0.6892
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8861438 1.2004937
sample estimates:
ratio of variances
          1.031475
```

Two Sample t-test

t=1.6968 , **df**=1336 , **p-value**=0.08998
alternative hypothesis: true difference in means is not equal to 0 with 95 percent **confidence interval:**
-0.08829755 1.21905646
sample estimates:
mean of male bmi **mean of female bmi**
30.94313 30.37775
Conclusion: With the p-value greater than 0.05 we can accept the null hypothesis of equal
average bmi (not significantly different) between men and women.

R CODE AND CONSOLE OUTPUT

```
> t.test(data[which(data$sex=="male"), "bmi"],
+        data[which(data$sex=="female"), "bmi"], alternative="two.sided",
 var.equal= TRUE,conf.level = 0.95)

        Two Sample t-test

data:  data[which(data$sex == "male"), "bmi"] and data[which(data$sex ==
 "female"), "bmi"]
t = 1.6968, df = 1336, p-value = 0.08998
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.08829755  1.21905646
sample estimates:
mean of x mean of y
 30.94313  30.37775
```

ALSO WE TESTED IF BMI VALUES OF MALES AND FEMALES ARE NORMALLY DISTRIBUTED

Test if it is normally distributed or not

Shapiro-Wilk Normality Test

-Male Bmi

W = 0.99305, **p-value** = 0.003133

-Female Bmi

W = 0.99303, **p-value** = 0.00354

Conclusion: The p-value smaller than 0.05 for both cases let us reject the null hipothesis of normally distribution.

R CODE AND CONSOLE OUTPUT

```
> shapiro.test(data[which(data$sex=="male"), "bmi"])

        Shapiro-Wilk normality test

data:  data[which(data$sex == "male"), "bmi"]
W = 0.99305, p-value = 0.003133

> shapiro.test(data[which(data$sex=="female"), "bmi"])

        Shapiro-Wilk normality test

data:  data[which(data$sex == "female"), "bmi"]
W = 0.99303, p-value = 0.003543
```

ANOVA

Is BMI equal among regions?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region	3	4056	1352.0	39.49	<2e-16 ***
Residuals	1334	45664	34.2		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Conclusion: One way ANOVA test was conducted to compare the mean of more than two independent groups. The p-value - Pr(>F) - is less than alpha ($\alpha = 0.05$), so we can reject the null hypothesis of equal means. Some of them are different.

To see what are different we can use Tukey pairwise comparison:

R CODE AND CONSOLE OUTPUT

```
> aovt<-aov(bmi~region,data=data)
> summary(aovt)
              Df Sum Sq Mean Sq F value Pr(>F)
region         3   4056  1352.0   39.49 <2e-16 ***
Residuals   1334  45664    34.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

TUKEY PAIRWISE COMPARISON

region	diff	lwr	upr	p adj
northwest-northeast	0.02628153	-1.1552239	1.207787	0.9999328
southeast-northeast	4.18248592	3.0330135	5.331958	0.0000000
southwest-northeast	1.42311230	0.2416069	2.604618	0.0106965
southeast-northwest	4.15620440	3.0076679	5.304741	0.0000000
southwest-northwest	1.39683077	0.2162360	2.577426	0.0127393
southwest-southeast	-2.75937363	-3.9079101	-1.610837	0.0000000

Conclusion: Using Tukey test with 95% family-wise confidence level, we made multiple comparisons of means. Adjusted p-values lower than alpha indicate that the differences are significant - here, the most part of pairwise comparisons.

R CODE AND CONSOLE OUTPUT

```
> TukeyHSD(aovt)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = bmi ~ region, data = data)

$region
                           diff        lwr       upr     p adj
northwest-northeast   0.02628153 -1.1552239  1.207787 0.9999328
southeast-northeast   4.18248592  3.0330135  5.331958 0.0000000
southwest-northeast   1.42311230  0.2416069  2.604618 0.0106965
southeast-northwest   4.15620440  3.0076679  5.304741 0.0000000
southwest-northwest   1.39683077  0.2162360  2.577426 0.0127393
southwest-southeast -2.75937363 -3.9079101 -1.610837 0.0000000
```

CHI-SQUARE TESTS (GOODNESS OF FIT)

1- INDEPENDENCE TEST
~Is there a relationship between the sex and region values?
- **H0:** There is no relationship between the sex and region values. They are independent of
each other.
- **H1:** There is a relationship between the sex and region values. They are dependent upon each other.
- Critical point is 7.81 since degrees of freedom is 3 and alpha = 0.05
Conclusion: Since the X-squared value is lower than the Critical value , we failed to reject the null hypothesis. In conclusion there is no relationship between smokers and their regions.

R CODE AND CONSOLE OUTPUT

```
> tbl1=table(data$sex,data$region)
> chisq.test(tbl1,correct=FALSE)

        Pearson's Chi-squared test

data:  tbl1
X-squared = 0.43514, df = 3, p-value = 0.9329
```

CHI-SQUARE TESTS (GOODNESS OF FIT)

2- Is there a difference in the proportion of smokers between genders?

- **H0** : There is no difference in the proportion of smokers between genders.
- **H1** : There is a difference in the proportion of smokers between genders.
- Critical point is 3.84 since degrees of freedom is 1.
- X-squared = 7.7659
- p-value = 0.005324

Conclusion: Since the X-squared value is greater than the critical value , we reject the null hypothesis and accept the alternative hypothesis and conclude that there is a difference in proportions of smokers in terms of gender.

R CODE AND CONSOLE OUTPUT

```
> tbl2=table(data$sex,data$smoker)
> chisq.test(tbl2,correct=FALSE,c)

        Pearson's Chi-squared test

data:  tbl2
X-squared = 7.7659, df = 1, p-value = 0.005324
```

CORRELATION

Testing the relationship between variables (age&charges)
First normality test is applied for both columns.

```
> shapiro.test(data$charges)

        Shapiro-Wilk normality test

data:  data$charges
W = 0.81469, p-value < 2.2e-16
```

```
> shapiro.test(data$age)

        Shapiro-Wilk normality test

data:  data$age
W = 0.9447, p-value < 2.2e-16
```

The low p-values let us reject the null hypothesis of normality so, a non-parametric method can be a better choice here.Correlation is used.

R code of correlation

```
> cor.test(data$charges,data$age,method="kendall")

        Kendall's rank correlation tau

data:  data$charges and data$age
z = 25.758, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.4753024
```

The low p-value would let us believe on a significant correlation.Correlation is 0.47

REGRESSION

Train/Test Split and Predict

Train: 75%
Test: 25%
Trained dataframe without region variable:
Residuals:

Min	1Q	Median	3Q	Max
-11437	-2952	-1091	1345	29837

Residual standard error: 6144 on 1000 degrees of freedom
Multiple R-squared: 0.7387,
Adjusted R-squared:0.7374
F-statistic: 565.3 on 5 and 1000 DF, p-value: < 2.2e-16

Conclusion: Since the p- value is very low, we can conclude that at least one predictor is really related to the outcome. Besides the RMSE value is 5,853 and the R squared value is 0.78.If we consider the coefficients (column Estimate) and their significance levels: The intercept is -11979.14, and almost all the predictors (except sex) are significant, according to the p-values

R CODE AND CONSOLE

```
> #TrainTestSplit
> set.seed(703)
> train<-createDataPartition(data$charges,times=1,p=0.75,list=FALSE)
> dtrain<-data[train,]
> dtest<-data[-train,]
> #train dataframe without region variable
> lrtrain<-lm(charges ~ age+sex+bmi+children+smoker, data = dtrain)
> summary(lrtrain)

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

Residuals:
   Min     1Q Median     3Q    Max
-11437  -2952  -1091   1345  29837

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11979.14    1115.72 -10.737  < 2e-16 ***
age             257.12      13.99  18.380  < 2e-16 ***
sexmale          58.07     389.10   0.149  0.88140
bmi             315.78      31.72   9.955  < 2e-16 ***
children        480.55     158.84   3.025  0.00255 **
smokeryes     23519.69     478.62  49.141  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6144 on 1000 degrees of freedom
Multiple R-squared:  0.7387,    Adjusted R-squared:  0.7374
F-statistic: 565.3 on 5 and 1000 DF,  p-value: < 2.2e-16

> #prediction
> pred<-predict(lrtrain,dtest)
> RMSE(pred,dtest$charges)
[1] 5852.901
> R2(pred,dtest$charges)
[1] 0.7822309
```

NONPARAMETRIC TEST EXAMPLES

1- One-sample Sign-Test

Is median age value is statistically equal to 39?
H0: M=39
H1: M!=39
data: data$age
s = 664, p-value = 0.6992
Conclusion: Since the p-value is higher than the alpha (0.05) we can reject the null hypothesis and accept alternative hypothesis and conclude as true median is not equal to 39.
95 percent confidence interval:
38< <41
sample estimates:
median of x 39

R CODE AND CONSOLE OUTPUT

```
> SIGN.test(data$age,md=39)

        One-sample Sign-Test

data:  data$age
s = 664, p-value = 0.6992
alternative hypothesis: true median is not equal to 39
95 percent confidence interval:
 38 41
sample estimates:
median of x
         39

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.9478     38     41
Interpolated CI       0.9500     38     41
Upper Achieved CI     0.9541     38     41
```

2- WILCOXON SIGNED RANK TEST

Is median age value less than 39 ?

H0: MU=39
H1: MU<39
data: data$age
V = 437452, **p-value** = 0.6723
Conclusion: Since the p-value is greater than the alpha(0.05) we can reject the null hypothesis , accept the alternative hypothesis and conclude as true location is less than 39.

R CODE AND CONSOLE OUTPUT

```
> wilcox.test(data$age,mu=39,alternative="less")

        Wilcoxon signed rank test with continuity correction

data:  data$age
V = 437452, p-value = 0.6723
alternative hypothesis: true location is less than 39
```

3- KRUSKAL-WALLIS RANK SUM TEST

data: bmi by region
Kruskal-Wallis chi-squared = 94.689,
df = 3, **p-value** < 2.2e-16
Conclusion:
As we conclude in ANOVA test that bmi is not equal in all regions , with the result of p-value is less than alpha (0.05) we reached the same result here with Kruskal-Wallis Test.

R CODE AND CONSOLE OUTPUT

```
> kruskal.test(bmi~region,data=data)

        Kruskal-Wallis rank sum test

data:  bmi by region
Kruskal-Wallis chi-squared = 94.689, df = 3,
 p-value < 2.2e-16
```

4- SHAPIRO-WILK NORMALITY TEST

The low p-values let us reject the null hypothesis of normality so, a non-parametric method can be a better choice.Shapiro-Wilk test example is below as we used on gender&bmi problem.

Shapiro-Wilk normality test
male and bmi
W = 0.99305, **p-value** = 0.003133
female and bmi
W = 0.99303, **p-value** = 0.003543
Conclusion: The p-value smaller than 0.05 for both cases let us reject the null hypothesis of
normally distribution.

R CODE AND CONSOLE OUTPUT

```
> shapiro.test(data[which(data$sex=="male"), "bmi"])

        Shapiro-Wilk normality test

data:  data[which(data$sex == "male"), "bmi"]
W = 0.99305, p-value = 0.003133

> shapiro.test(data[which(data$sex=="female"), "bmi"])

        Shapiro-Wilk normality test

data:  data[which(data$sex == "female"), "bmi"]
W = 0.99303, p-value = 0.003543
```

CONCLUSION

We analyzed the dataset by conducting various experiments. We applied tests such as T-test, Chi-squared test, F test, etc. Using the results we got from these tests, we made various inferences about the dataset, and made some conclusions. We estimated the charges values with the linear model we built, which works with sufficient accuracy. We examined the relationships between the columns from the tests we made and obtained some results.

References

(n.d.). Retrieved from https://www.kaggle.com/datasets/mirichoi0218/insurance.

THE END