Final Project

2024-04-16

Updated:
DateSet1:
2024-04-16 1): Finish descriptive statistics
2): Finish dataset Visulation
3): Finish parts of ANOVA(with Post- hoc test), T-Test
4): (linear/logistic) model- Finish Linear Model of age and charges, Linear Model of bmi and charges
5): Question !!! So, to seperate the column to do ANOVA or combine all columns to do ANOVA?? (3.2 &3.5)

2024-04-22 1) Organzie ANOVA 2) Make a correleation matrix 3) To do a linear regression model in 3D plot

DateSet2:
2024-04-16 1: Finish descriptive statistics

2024-04-22 1) Remove dataset2

2024-05-02 1) Update the graph:Linear (graph:Linear) Model of age and charges 2) Update the graph:Linear (graph:Linear) Model by BMI_Group 3) Update the graph:3D (graph:3D) graph 4) Update Model for linear regression 5) Add the interpretation

2020-05-05 1) bmi chart (change x, y) 2)Table Descriptive Statistics Update
##

```
#update.packages(ask = FALSE)
#install.packages("dplyr")
#install.packages("corrplot")
```

Dataset1

Connect to MySQL and Select the data we want to use as df in R.

```
#install and load Packages.

#install.packages("RMySQL")
library(RMySQL)
```

```
## Loading required package: DBI
```

```
library(DBI)
library(RMySQL)


# Load ggplot library
library(ggplot2)
```

```
con <-dbConnect(MySQL(),
#host='127.0.0.1',
#user='root',
#password='12345678',
#dbname='HDS_R_Final')

host = 'srv1389.hstgr.io',
user = 'u721770684_root',
password ='HDShds1234',
dbname = 'u721770684_HDS_test')
```

```
# Load dataset1
test_data <-dbGetQuery(con,"
SELECT
    A.age,
    A.bmi,
    A.charges,
    A.children,
    A.region,
    A.sex,
    A.smoker,
  CASE
        -- make sure how to define age group? any studies or refeerence?
        -- now: <18, 18-29, 30-44, 45-64
        WHEN A.age < 18 THEN 'Age<18'
        WHEN A.age >= 18 AND A.age < 30 THEN'Age18-29'
        WHEN A.age >= 30 AND A.age <45 THEN 'Age30-44'
        WHEN A.age >= 45 AND A.age <65 THEN 'Age45-64'
        WHEN A.age >65 THEN 'Age>65'

        ELSE 'NA'
            END AS 'Age_Group',

    CASE
            -- bmi groups REFERENCES: https://www.cdc.gov/healthyweight/assessing/bmi/ad
ult_bmi/index.html
            WHEN A.bmi < 18.5 THEN 'Underweight'
            WHEN A.bmi >=18.5 AND A.bmi<24.9 THEN 'Healthy Weight'
            WHEN A.bmi >=24.9 AND A.bmi<29.9 THEN 'Overweight'
            WHEN A.bmi >=29.0 THEN 'Obesity'
            ELSE 'NA'
        END AS 'BMI_Group'

FROM
    insurance A;
")
df <- as.data.frame(test_data) #convert to dataframe
#print(df)

head(df, 10)
```

```
##    age    bmi   charges children    region    sex smoker Age_Group
## 1   19 27.900 16884.924        0 southwest female    yes  Age18-29
## 2   18 33.770  1725.552        1 southeast   male     no  Age18-29
## 3   28 33.000  4449.462        3 southeast   male     no  Age18-29
## 4   33 22.705 21984.471        0 northwest   male     no  Age30-44
## 5   32 28.880  3866.855        0 northwest   male     no  Age30-44
## 6   31 25.740  3756.622        0 southeast female     no  Age30-44
## 7   46 33.440  8240.590        1 southeast female     no  Age45-64
## 8   37 27.740  7281.506        3 northwest female     no  Age30-44
## 9   37 29.830  6406.411        2 northeast   male     no  Age30-44
## 10  60 25.840 28923.137        0 northwest female     no  Age45-64
##         BMI_Group
## 1      Overweight
## 2         Obesity
## 3         Obesity
## 4  Healthy Weight
## 5      Overweight
## 6      Overweight
## 7         Obesity
## 8      Overweight
## 9      Overweight
## 10     Overweight
```

```
#Clean data:Remove NA

df <- na.omit(df)

head(df,10)
```

```
##    age    bmi   charges children    region    sex smoker Age_Group
## 1   19 27.900 16884.924        0 southwest female    yes  Age18-29
## 2   18 33.770  1725.552        1 southeast   male     no  Age18-29
## 3   28 33.000  4449.462        3 southeast   male     no  Age18-29
## 4   33 22.705 21984.471        0 northwest   male     no  Age30-44
## 5   32 28.880  3866.855        0 northwest   male     no  Age30-44
## 6   31 25.740  3756.622        0 southeast female     no  Age30-44
## 7   46 33.440  8240.590        1 southeast female     no  Age45-64
## 8   37 27.740  7281.506        3 northwest female     no  Age30-44
## 9   37 29.830  6406.411        2 northeast   male     no  Age30-44
## 10  60 25.840 28923.137        0 northwest female     no  Age45-64
##         BMI_Group
## 1      Overweight
## 2         Obesity
## 3         Obesity
## 4  Healthy Weight
## 5      Overweight
## 6      Overweight
## 7         Obesity
## 8      Overweight
## 9      Overweight
## 10     Overweight
```

```
#Recode: columns needs to turn to 0.1 2 3 4 format to do anaylize

# Age_c columns
df$age_c <- 0  # Create a new column and initialize all values to 0
df$age_c[df$Age_Group == 'Age18-29'] <- 1
df$age_c[df$Age_Group == 'Age30-44'] <- 2  # Replace with 2 where age is '30-44'
df$age_c[df$Age_Group == 'Age45-64'] <- 3  # Replace with 3 where age is '45-64'
df$age_c[df$Age_Group == 'Age>65'] <- 4 # Replace with 4 where age is lager than 65

#region_c columns
df$region_c[df$region == 'southwest'] <-0
df$region_c[df$region == 'southeast'] <-1
df$region_c[df$region == 'northwest'] <-2
df$region_c[df$region == 'northeast'] <-3

#BMI_group columns
df$bmi_group_c[df$BMI_Group == 'Underweight'] <-0
df$bmi_group_c[df$BMI_Group == 'Healthy Weight'] <-1
df$bmi_group_c[df$BMI_Group == 'Overweight'] <-2
df$bmi_group_c[df$BMI_Group == 'Obesity'] <-3

#sex_c columns
df$sex_c[df$sex == 'female'] <-0
df$sex_c[df$sex == 'male'] <-1

#smoker column

df$smoker_c[df$smoker =='no'] <-0
df$smoker_c[df$smoker =='yes']  <-1

head(df,10)
```

```
##    age    bmi   charges children    region    sex smoker Age_Group
## 1   19 27.900 16884.924        0 southwest female    yes  Age18-29
## 2   18 33.770  1725.552        1 southeast   male     no  Age18-29
## 3   28 33.000  4449.462        3 southeast   male     no  Age18-29
## 4   33 22.705 21984.471        0 northwest   male     no  Age30-44
## 5   32 28.880  3866.855        0 northwest   male     no  Age30-44
## 6   31 25.740  3756.622        0 southeast female     no  Age30-44
## 7   46 33.440  8240.590        1 southeast female     no  Age45-64
## 8   37 27.740  7281.506        3 northwest female     no  Age30-44
## 9   37 29.830  6406.411        2 northeast   male     no  Age30-44
## 10  60 25.840 28923.137        0 northwest female     no  Age45-64
##          BMI_Group age_c region_c bmi_group_c sex_c smoker_c
## 1      Overweight     1        0           2     0        1
## 2         Obesity     1        1           3     1        0
## 3         Obesity     1        1           3     1        0
## 4  Healthy Weight     2        2           1     1        0
## 5      Overweight     2        2           2     1        0
## 6      Overweight     2        1           2     0        0
## 7         Obesity     3        1           3     0        0
## 8      Overweight     2        2           2     0        0
## 9      Overweight     2        3           2     1        0
## 10     Overweight     3        2           2     0        0
```

```
df_clean <- subset(df, select = c("age", "bmi","sex_c","smoker_c", "charges"))

head(df_clean,10)
```

```
##    age    bmi sex_c smoker_c   charges
## 1   19 27.900     0        1 16884.924
## 2   18 33.770     1        0  1725.552
## 3   28 33.000     1        0  4449.462
## 4   33 22.705     1        0 21984.471
## 5   32 28.880     1        0  3866.855
## 6   31 25.740     0        0  3756.622
## 7   46 33.440     0        0  8240.590
## 8   37 27.740     0        0  7281.506
## 9   37 29.830     1        0  6406.411
## 10  60 25.840     0        0 28923.137
```

```
#install.packages("tldr")
#install.packages("knitr")
#install.packages("tableone")
```

#1)Descriptive statistics for research question 1

Phase3 -3a) model 1

```
#summary_stats <- summary(test_data)
#print(test_data)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)

dataset1_summary <- test_data %>%
  select(where(is.numeric)) %>%
  summary()

print(dataset1_summary)
```

```
##       age             bmi           charges         children
##  Min.   :18.00   Min.   :15.96   Min.   : 1122   Min.   :0.000
##  1st Qu.:27.00   1st Qu.:26.30   1st Qu.: 4740   1st Qu.:0.000
##  Median :39.00   Median :30.40   Median : 9382   Median :1.000
##  Mean   :39.21   Mean   :30.66   Mean   :13270   Mean   :1.095
##  3rd Qu.:51.00   3rd Qu.:34.69   3rd Qu.:16640   3rd Qu.:2.000
##  Max.   :64.00   Max.   :53.13   Max.   :63770   Max.   :5.000
```

```
# Calculating summary statistics for BMI and Charges
summary_df <- df %>%
  summarise(
    BMI_min = round(min(bmi, na.rm = TRUE),2),
    BMI_Median = round(median(bmi, na.rm = TRUE),2),
    BMI_Mean = round(mean(bmi, na.rm = TRUE),2),
    BMI_SD = round(sd(bmi, na.rm = TRUE),2),
    BMI_Q1 = round(quantile(bmi, 0.25, na.rm = TRUE),2),
    BMI_Q3 = round(quantile(bmi, 0.75, na.rm = TRUE),2),
    BMI_max = round(max(bmi, na.rm = TRUE),2),

    Age_min = round(min(age, na.rm = TRUE),2),
    Age_Median = round(median(age, na.rm = TRUE),2),
    Age_Mean = round(mean(age, na.rm = TRUE),2),
    Age_SD = round(sd(age, na.rm = TRUE),2),
    Age_Q1 = round(quantile(age, 0.25, na.rm = TRUE),2),
    Age_Q3 = round(quantile(age, 0.75, na.rm = TRUE),2),
    Age_max = round(max(age, na.rm = TRUE),2),

    children_min = round(min(children, na.rm = TRUE),2),
    children_Median = round(median(children, na.rm = TRUE),2),
    children_Mean = round(mean(children, na.rm = TRUE),2),
    children_SD = round(sd(children, na.rm = TRUE),2),
    children_Q1 = round(quantile(children, 0.25, na.rm = TRUE),2),
    children_Q3 = round(quantile(children, 0.75, na.rm = TRUE),2),
    children_max = round(max(children, na.rm = TRUE),2),



    Charges_min = round(min(charges, na.rm = TRUE),2),
    Charges_Median = round(median(charges, na.rm = TRUE),2),
    Charges_Mean = round(mean(charges, na.rm = TRUE),2),
    Charges_SD = round(sd(charges, na.rm = TRUE),2),
    Charges_Q1 = round(quantile(charges, 0.25, na.rm = TRUE),2),
    Charges_Q3 = round(quantile(charges, 0.75, na.rm = TRUE),2),
    Charges_max = round(max(charges, na.rm = TRUE),2),


  ) %>%
  pivot_longer(
    cols = everything(),
    names_to = c("Variable", ".value"),
    names_pattern = "(.*)_(.*)"
  )

# Viewing the transformed summary
print(summary_df)
```

```
## # A tibble: 4 × 8
##   Variable    min Median     Mean       SD     Q1      Q3     max
##   <chr>     <dbl>  <dbl>    <dbl>    <dbl>  <dbl>   <dbl>   <dbl>
## 1 BMI        16.0   30.4    30.7      6.1    26.3    34.7    53.1
## 2 Age        18     39      39.2     14.0    27      51      64
## 3 children    0      1       1.09     1.21    0       2       5
## 4 Charges  1122.  9382.  13270.   12110.   4740.  16640.  63770.
```

```
library(dplyr)
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

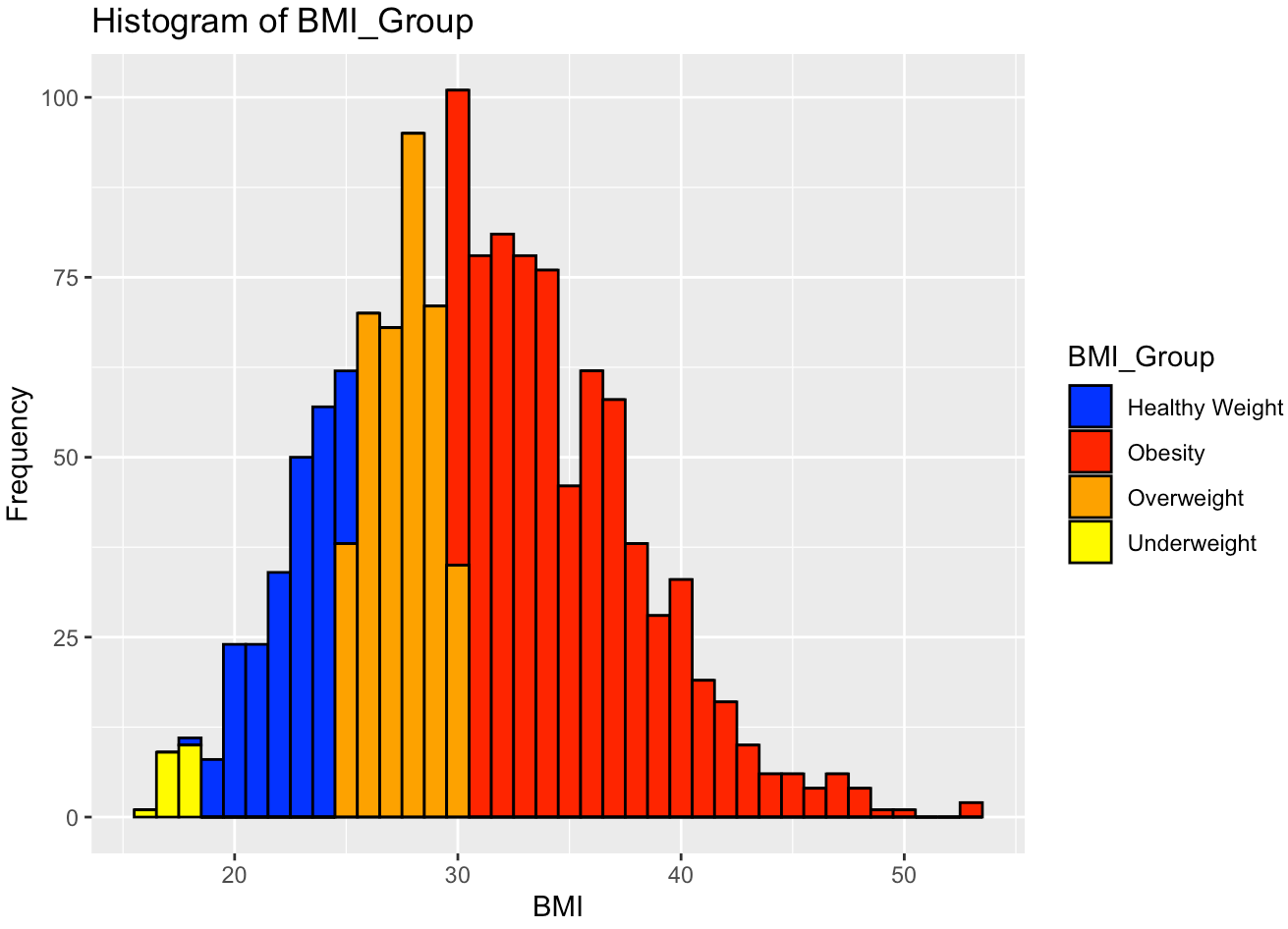
```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
# Assuming test_data is your dataset
#dataset1_summary <- summary_df %>%
#  select(where(is.numeric)) %>%
#  summary()


# Using kable to create a formatted table and make headers bold
kable(summary_df, format = "html", caption = "Descriptive Statistics") %>%
  kable_styling(full_width = FALSE, font_size = 12) %>%
  row_spec(0, bold = TRUE, background = "#D3D3D3") %>%  # Making header row bold and giv
ing it a background color
  column_spec(1, width = "10em") %>%  # Adjust the width for the Variable name column
  column_spec(2:8, width = "5em")  # Adjust the width for each statistics column
```

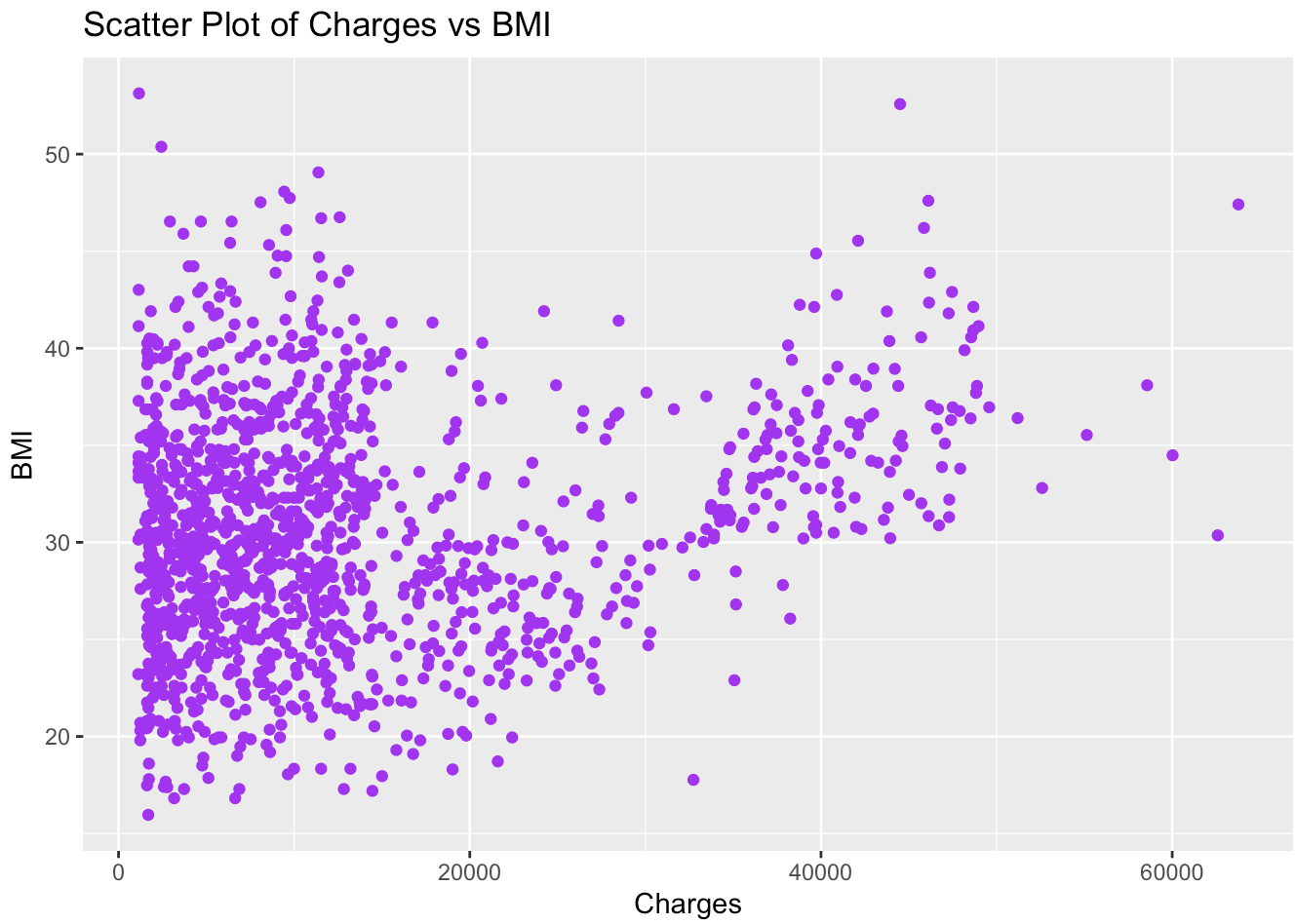
Descriptive Statistics

Variable	min	Median	Mean	SD	Q1	Q3	max
BMI	15.96	30.40	30.66	6.10	26.30	34.69	53.13
Age	18.00	39.00	39.21	14.05	27.00	51.00	64.00
children	0.00	1.00	1.09	1.21	0.00	2.00	5.00
Charges	1121.87	9382.03	13270.42	12110.01	4740.29	16639.91	63770.43

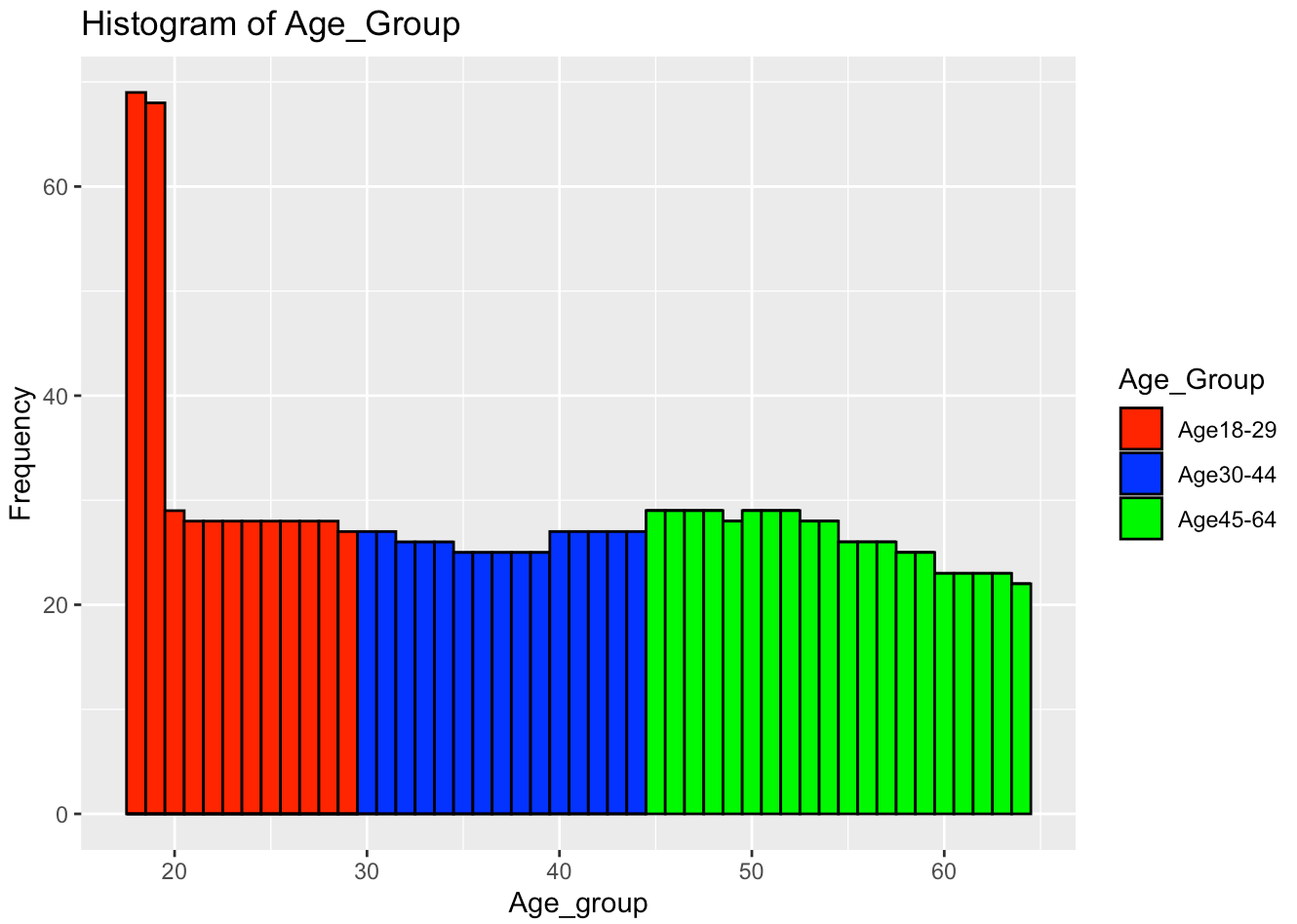
```
# 2.1)histogram (bmi_group) and scatter plot (bmi)

#Create normal normal histogram
# # Histogram for BMI
# ggplot(df, aes(x=bmi)) +
#   geom_histogram(binwidth = 1, fill="blue", color="black") +
#   ggtitle("Histogram of BMI") +
#   xlab("BMI") +
#   ylab("Frequency")


# Create the histogram by BMI_Group
ggplot(df, aes(x = bmi, fill = BMI_Group)) +
  geom_histogram(binwidth = 1, color = "black") +
  scale_fill_manual(values = c("Underweight" = "yellow", "Healthy Weight" = "blue", "Ove
rweight" = "orange", "Obesity" = "red")) +
  ggtitle("Histogram of BMI_Group") +
  xlab("BMI") +
  ylab("Frequency")
```



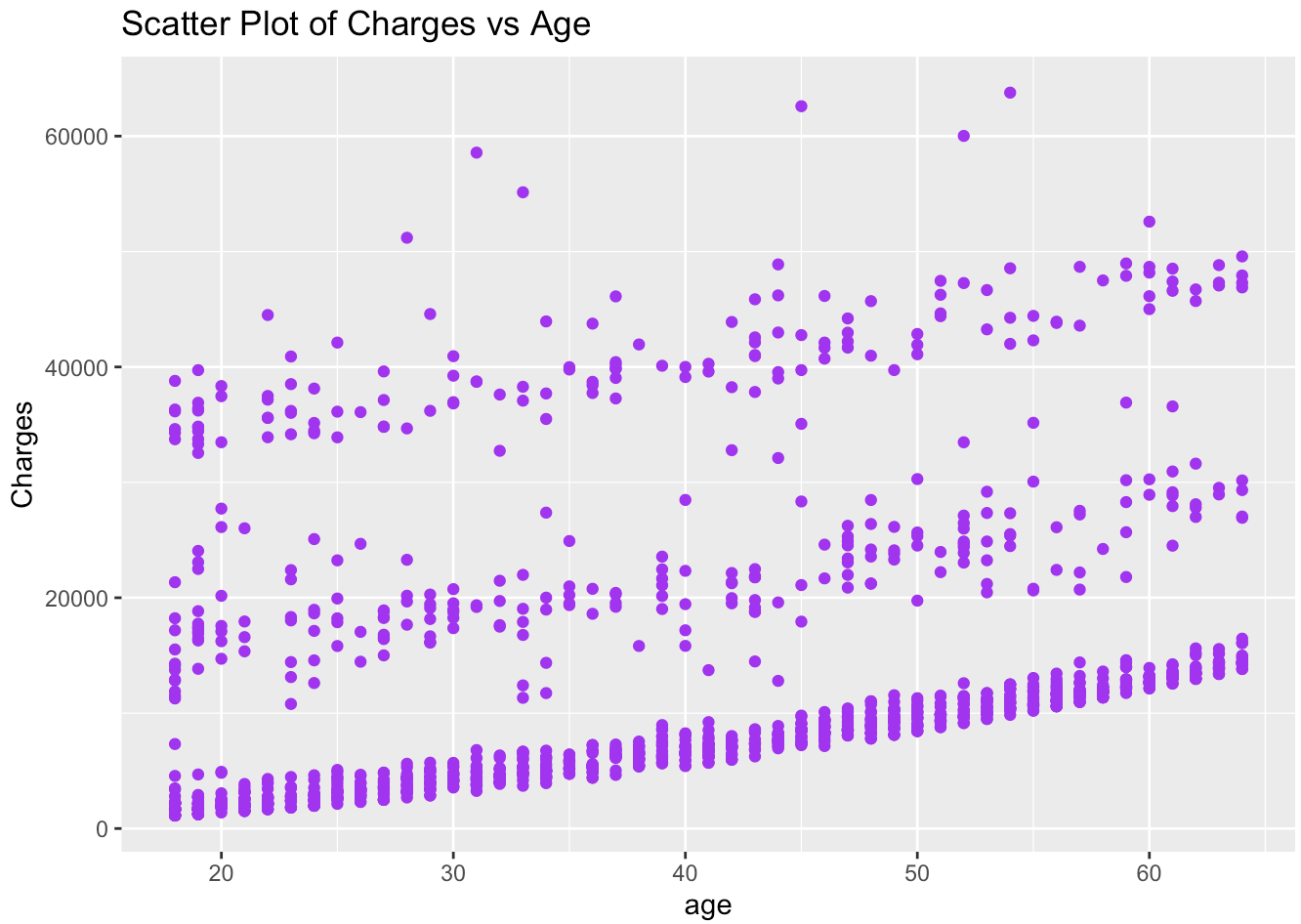
```
#scatter plot
# Scatter plot for BMI vs Charges
ggplot(df, aes(x= charges, y=bmi)) +
  geom_point(color="purple") +
  ggtitle("Scatter Plot of Charges vs BMI") +
  xlab("Charges") +
  ylab("BMI")
```



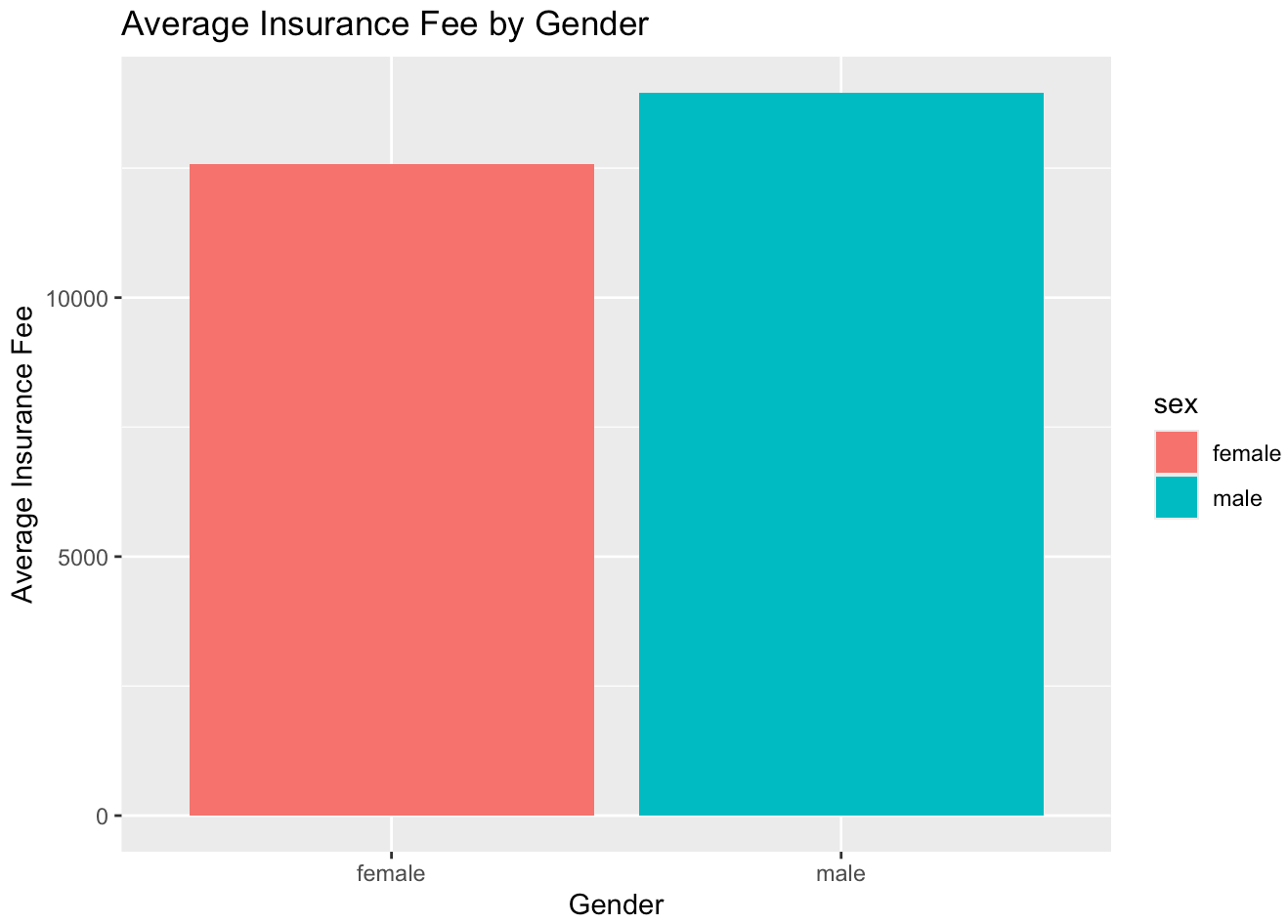

```
# 2.2)histogram (age_group) and scatter plot(age)

ggplot(df, aes(x = age, fill = Age_Group)) +
  geom_histogram(binwidth = 1, color = "black") +
  scale_fill_manual(values = c("Age<18" = "yellow",'Age18-29' = 'Red', "Age30-44" = "blu
e", "Age45-64" = 'Green', "Age >65" = 'Purple')) +
  ggtitle("Histogram of Age_Group") +
  xlab("Age_group") +
  ylab("Frequency")
```



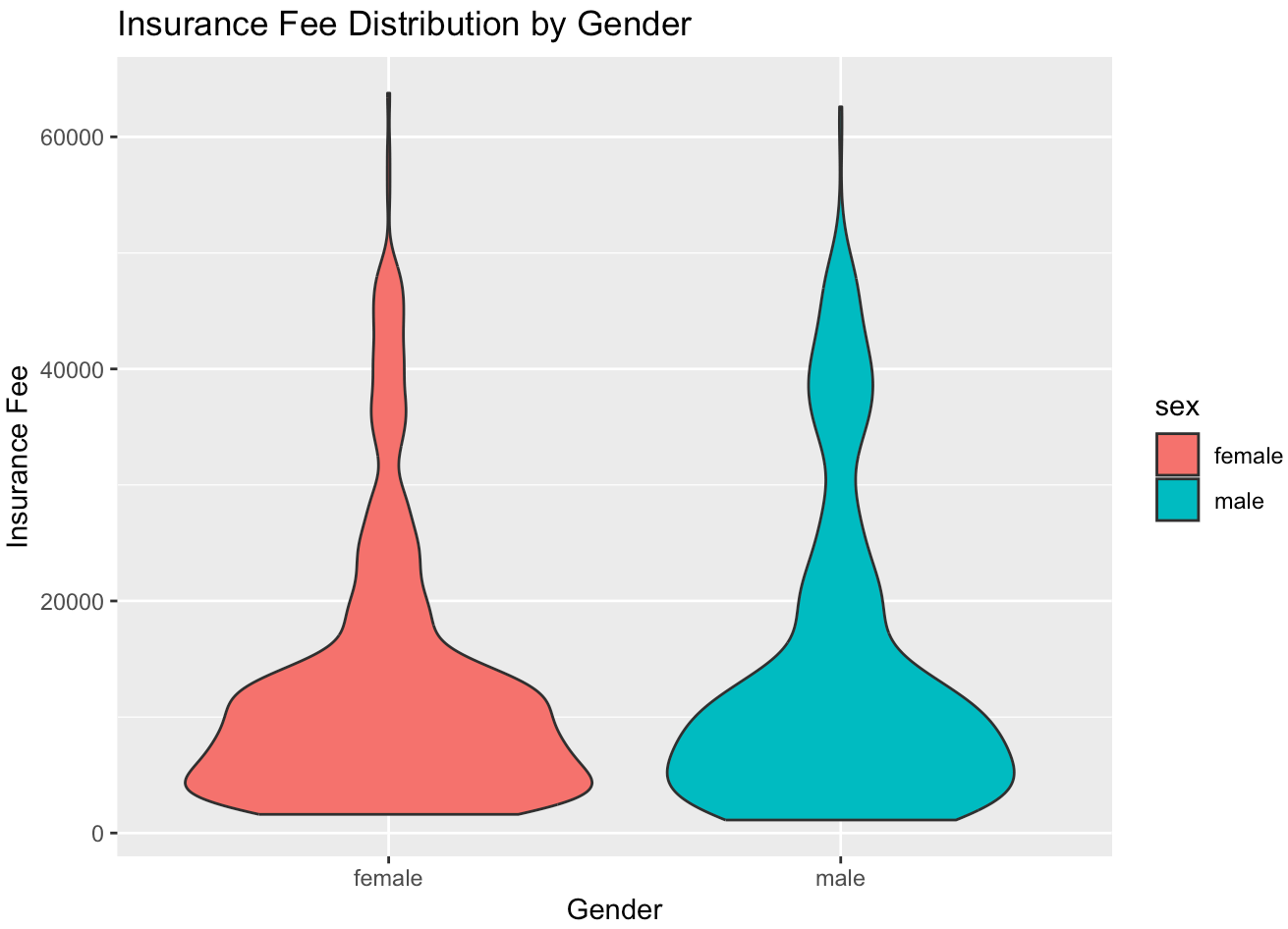

```
#scatter plot
# Scatter plot for age vs Charges
ggplot(df, aes(x=age, y=charges)) +
  geom_point(color="purple") +
  ggtitle("Scatter Plot of Charges vs Age") +
  xlab("age") +
  ylab("Charges")
```



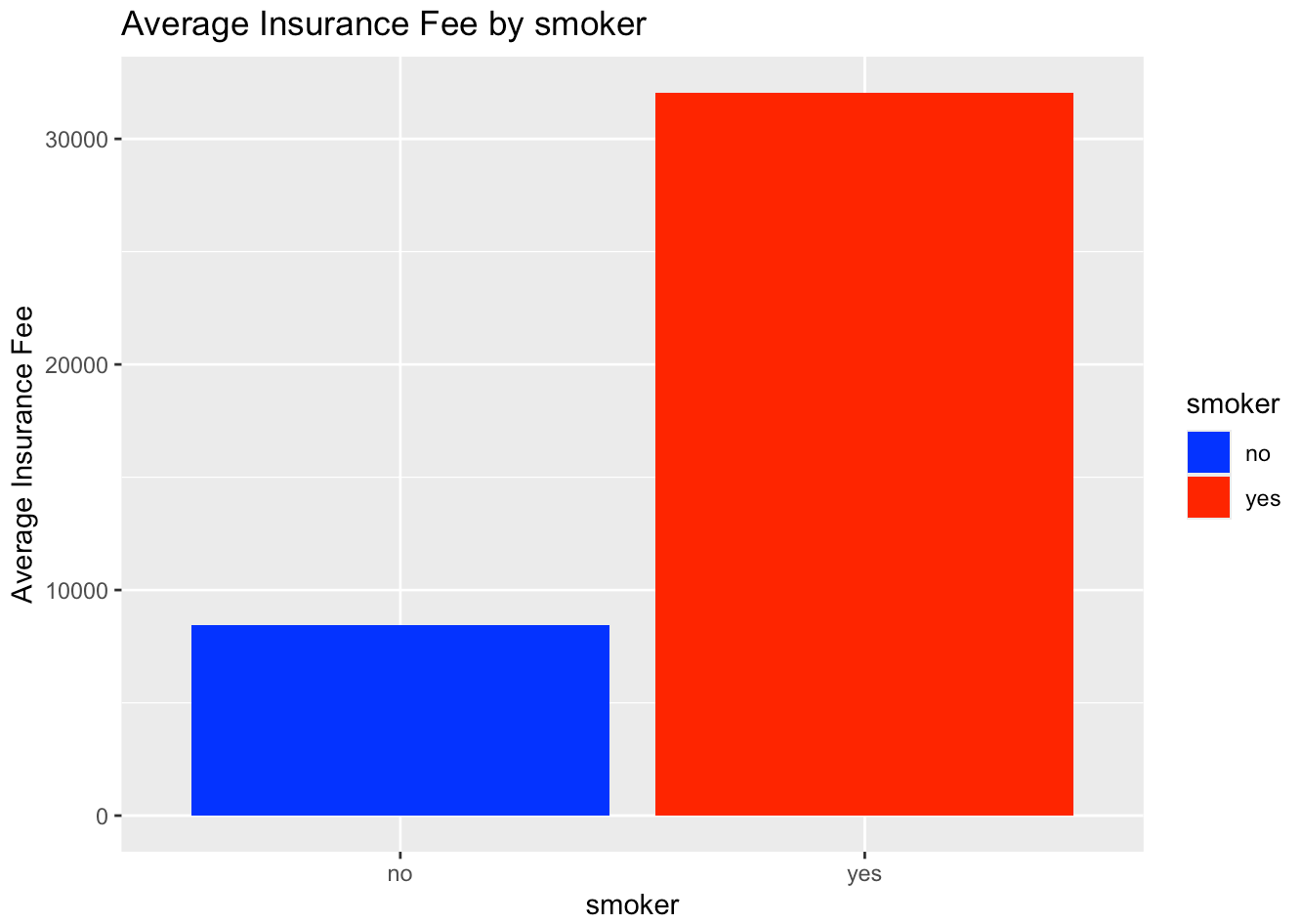

```
# 2.3)bar chart (sex) and violin plot(sex)

# bar chart
ggplot(df, aes(x=sex, y=charges, fill=sex)) +
    geom_bar(stat="summary", fun=mean) +
    labs(title="Average Insurance Fee by Gender", x="Gender", y="Average Insurance Fee")
```



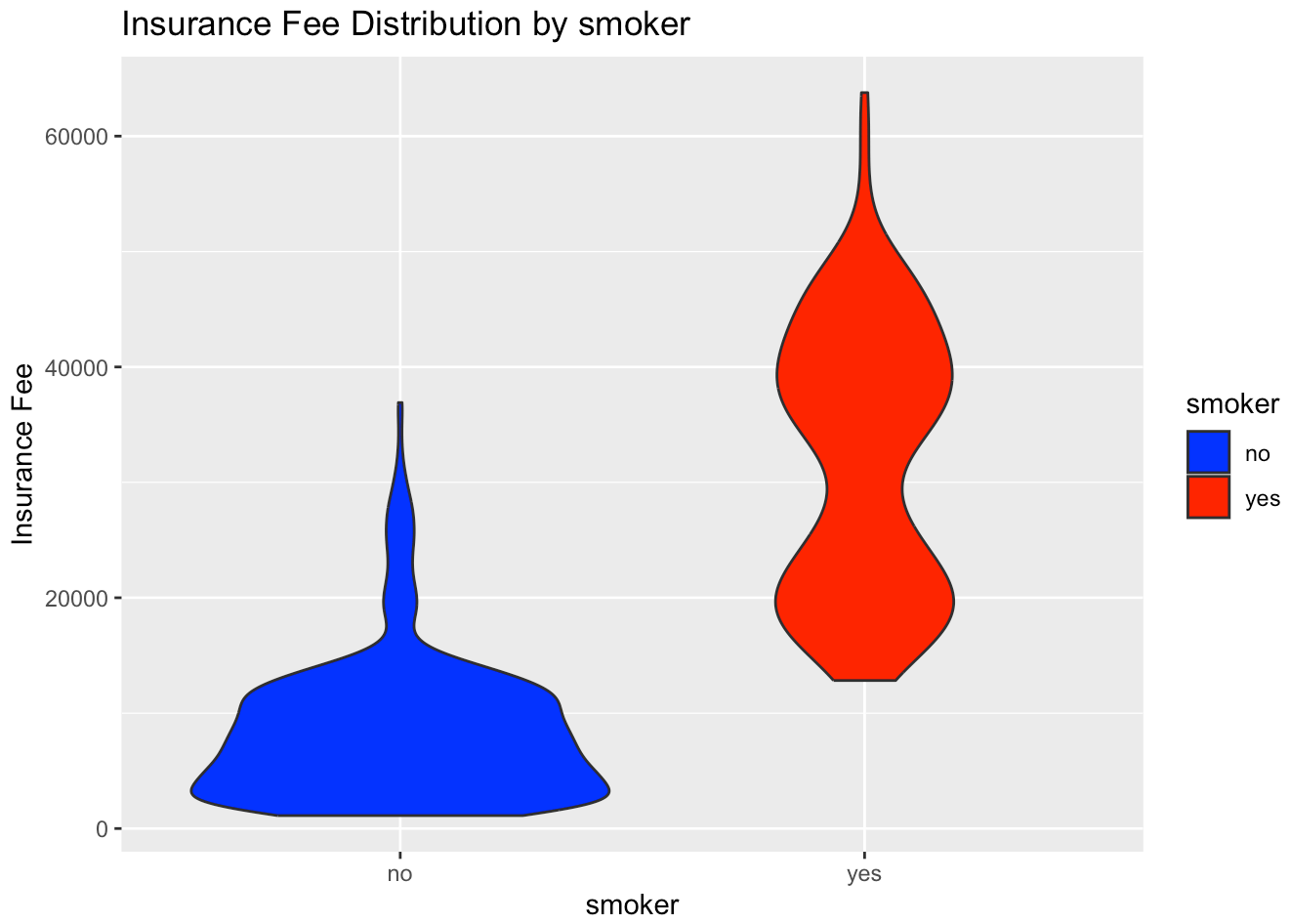

```
#Violin plot
ggplot(df, aes(x=sex, y=charges, fill=sex)) +
    geom_violin() +
    labs(title="Insurance Fee Distribution by Gender", x="Gender", y="Insurance Fee")
```



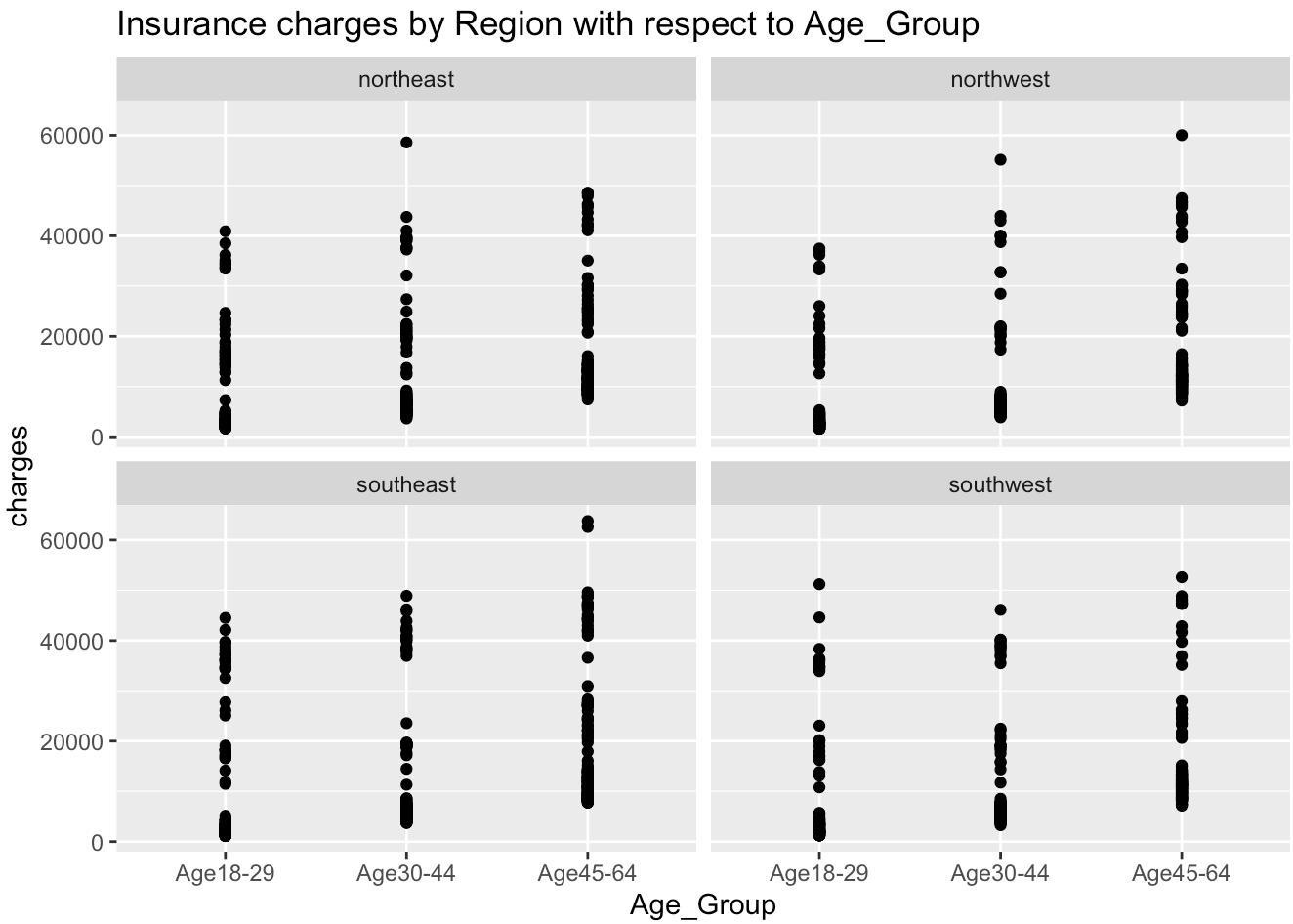

```
# 2.4) smoke plot

# bar chart
ggplot(df, aes(x=smoker, y=charges, fill=smoker)) +
    geom_bar(stat="summary", fun=mean) +
    scale_fill_manual(values=c("no"="blue", "yes"="red")) +   # change colors
    labs(title="Average Insurance Fee by smoker", x="smoker", y="Average Insurance Fee")
```



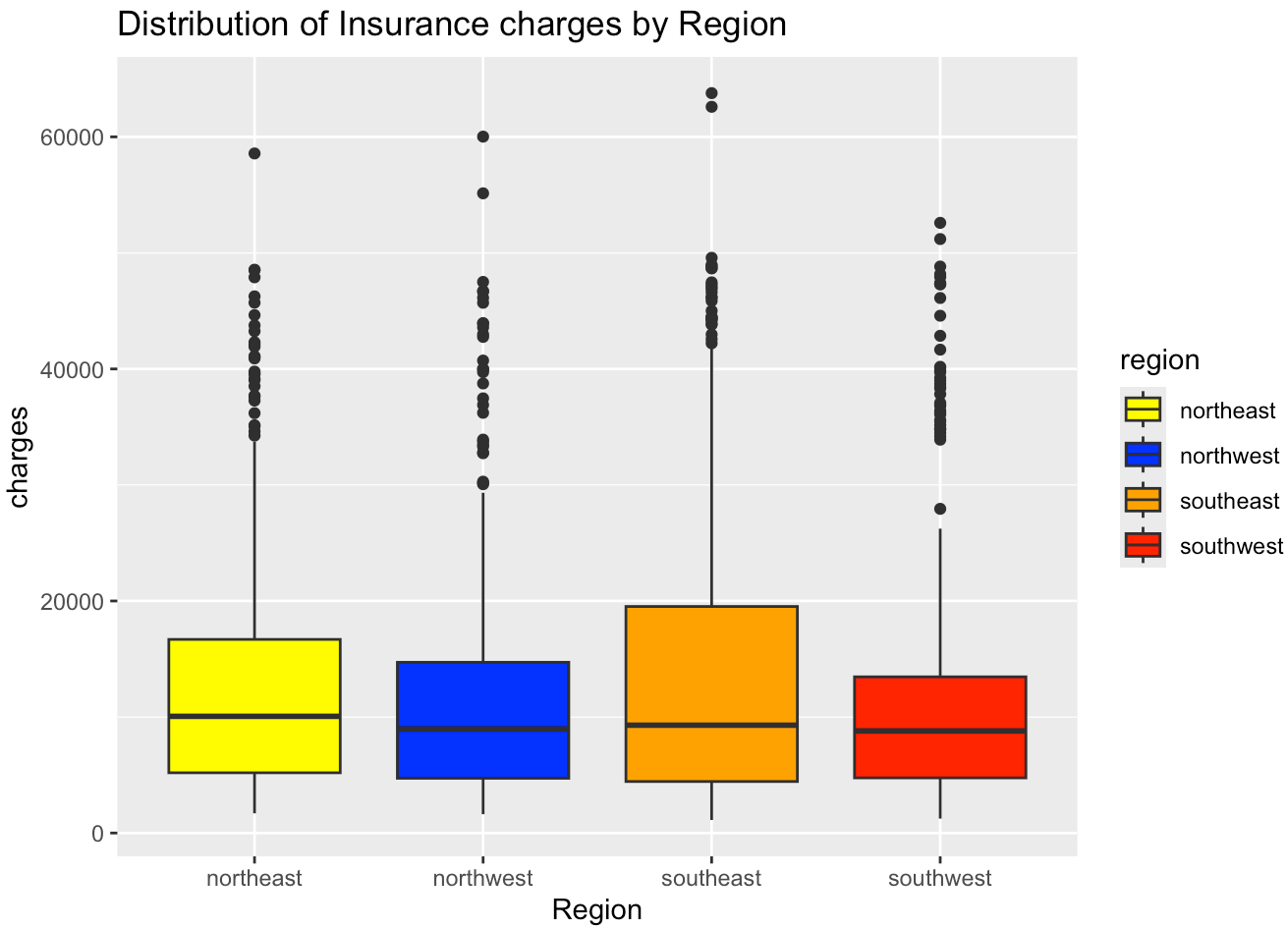

```
#Violin plot
ggplot(df, aes(x=smoker, y=charges, fill=smoker)) +
    geom_violin() +
    scale_fill_manual(values=c("no"="blue", "yes"="red")) +   # change colors
    labs(title="Insurance Fee Distribution by smoker", x="smoker", y="Insurance Fee")
```



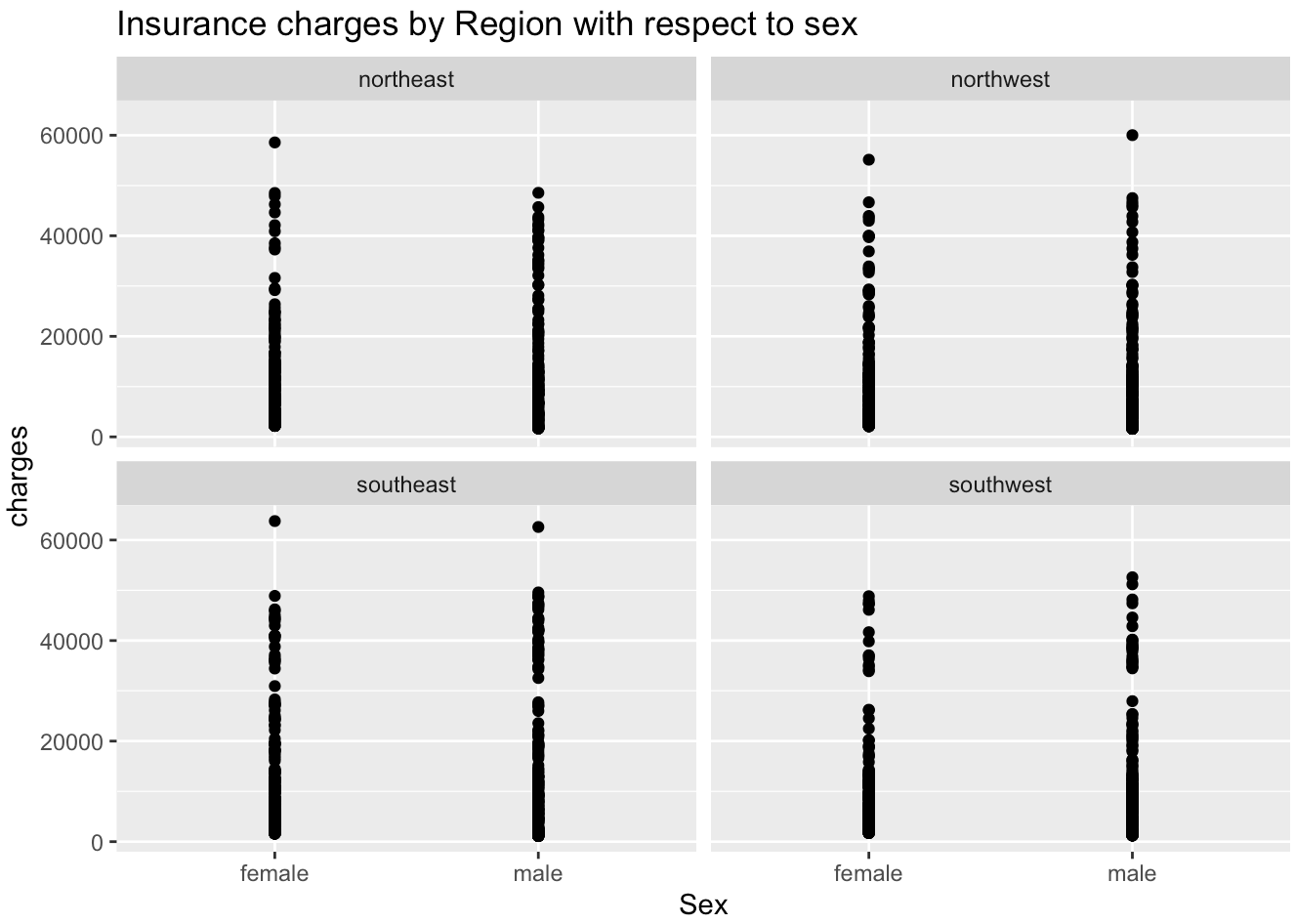

```
# 2.5) region plot

ggplot(df, aes(x=Age_Group, y=charges)) +
  geom_point() +
  facet_wrap(~region) +
  ggtitle("Insurance charges by Region with respect to Age_Group") +
  xlab("Age_Group") +
  ylab("charges")
```



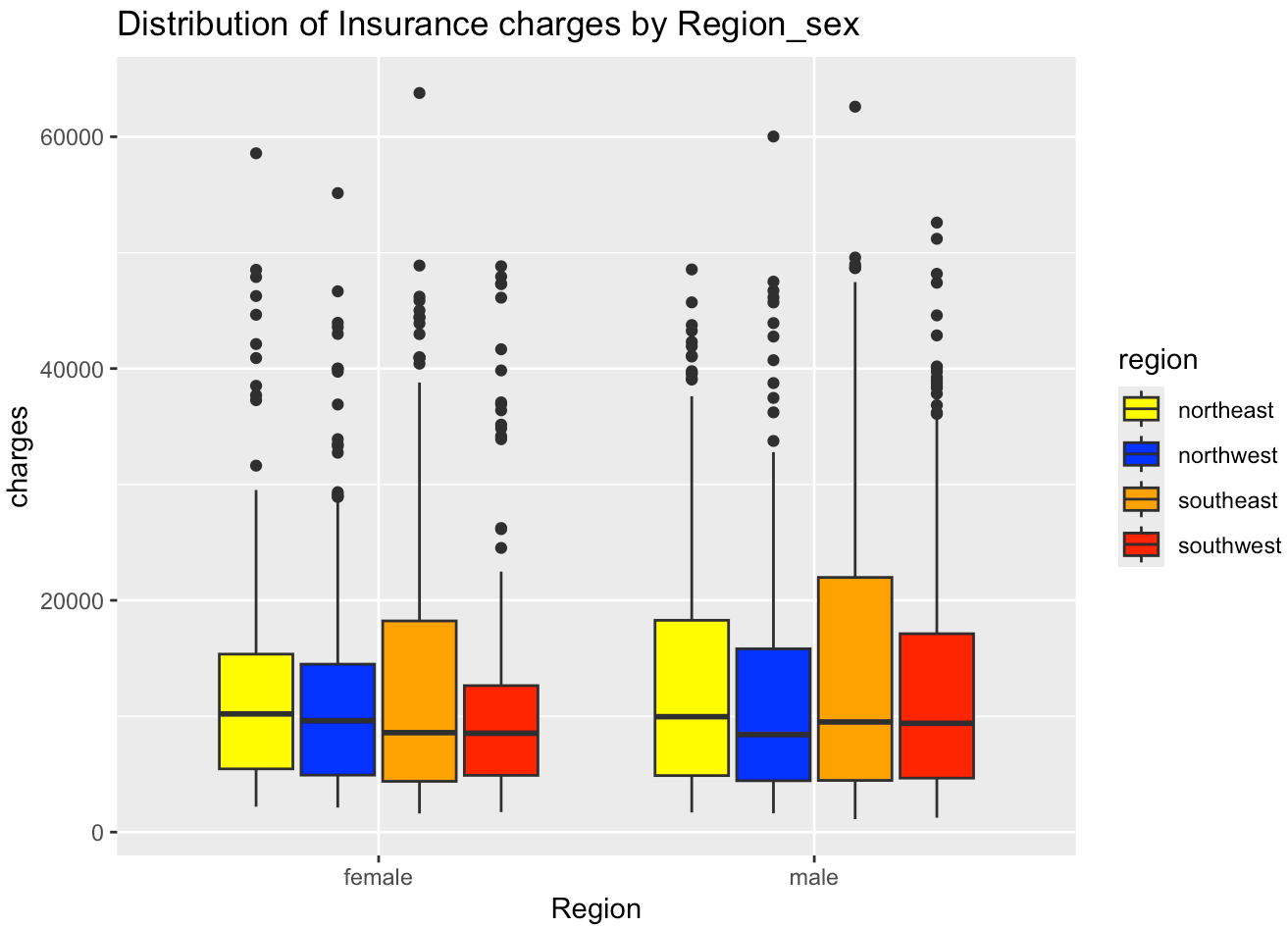

```
ggplot(df, aes(x=region, y=charges,fill=region)) +
  geom_boxplot() +
  ggtitle("Distribution of Insurance charges by Region") +
  scale_fill_manual(values = c("northeast" = "yellow", "northwest" = "blue", "southeast"
= "orange", "southwest" = "red")) +
  xlab("Region") +
  ylab("charges")
```




```
ggplot(df, aes(x=sex, y=charges)) +
  geom_point() +
  facet_wrap(~region) +
  ggtitle("Insurance charges by Region with respect to sex") +
  xlab("Sex") +
  ylab("charges")
```




```
ggplot(df, aes(x=sex, y=charges,fill=region)) +
  geom_boxplot() +
  ggtitle("Distribution of Insurance charges by Region_sex") +
  scale_fill_manual(values = c("northeast" = "yellow", "northwest" = "blue", "southeast"
= "orange", "southwest" = "red")) +
  xlab("Region") +
  ylab("charges")
```




```
# 3.1) ANOVA for all (Age_group, bmi_group,region) --> if significant , to do post- hoc test (3.4 can get more explantion)
anova_result <- aov(charges ~ BMI_Group +Age_Group+region , data = df)
summary(anova_result)
```

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## BMI_Group     3 7.586e+09 2.529e+09  19.252 3.19e-12 ***
## Age_Group     2 1.313e+10 6.563e+09  49.963  < 2e-16 ***
## region        3 8.018e+08 2.673e+08   2.035    0.107
## Residuals  1329 1.746e+11 1.313e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Post- hoc test example code:
#pairwise.t.test(df$charges, df$BMI_Group, p.adjust.method = "bonferroni")
#pairwise.t.test(df$charges, df$Age_Group, p.adjust.method = "bonferroni")
```

In this ANOVA analysis, BMI_Group and Age_Group show statistically significant differences in the variable being analyzed, indicating that the means of these groups differ. However, the region does not show a significant difference (p-value 0.107), suggesting that the means for different regions might not significantly differ.

```
# 3.2) ANVOVA for Age_group and charges.
# When doing ANOVA, it does not need to recode the data to 1.2.3.4.... Thus, just using the original column.

anova_result_age <- aov(charges ~ Age_Group , data = df)
summary(anova_result_age)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## Age_Group      2 1.485e+10 7.423e+09   54.68 <2e-16 ***
## Residuals   1335 1.812e+11 1.358e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#because of significant result, so doing Post- hoc test
pairwise.t.test(df$charges, df$Age_Group, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$charges and df$Age_Group
##
##          Age18-29 Age30-44
## Age30-44 0.00017  -
## Age45-64 < 2e-16  1.4e-08
##
## P value adjustment method: bonferroni
```

In this ANOVA analysis,it shows significant differences given the F value and extremely low p-value ($p < 2e-16$). Every comparison between pairs is statistically significant, suggesting clear variations in charges across various age groups after accounting for multiple comparisons using the Bonferroni technique. These findings indicate that age has a substantial impact on the variable ‘charges’, and each age group exhibits a statistically distinct average charge. This underscores the need of implementing customized financial plans or policies for various age groups.

```
# 3.3) ANVOVA for bmi and charges.


anova_result_bmi <- aov(charges ~ BMI_Group , data = df)
summary(anova_result_bmi)
```

```
##               Df    Sum Sq   Mean Sq F value   Pr(>F)
## BMI_Group      3 7.586e+09 2.529e+09    17.9 2.17e-11 ***
## Residuals   1334 1.885e+11 1.413e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#because of significant result, so doing Post- hoc test
pairwise.t.test(df$charges, df$BMI_Group, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$charges and df$BMI_Group
##
##             Healthy Weight Obesity Overweight
## Obesity     1.9e-07        -       -
## Overweight  1.000          3.5e-08 -
## Underweight 1.000          0.086   1.000
##
## P value adjustment method: bonferroni
```

This ANOVA is to evaluate if there are significant differences in charges across different BMI categories. The result is significant differences, with the F value being 17.9 and the p-value very small (2.17e-11). In addition, given the significant ANOVA results, a post-hoc test (pairwise.t.test) is conducted to determine which specific BMI categories differ in terms of charges. The result shows that “Obesity vs. Healthy Weight” and “Obesity vs. Overweight” are significant different.

```
# 3.4) ANOVA for region.

anova_result_region <- aov(charges ~ region, data = df)
summary(anova_result_region)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## region         3 1.301e+09 433586560    2.97 0.0309 *
## Residuals   1334 1.948e+11 146007093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# p-value =0.0309 (P<.05)

#Because the result is significant, to do post hoc test to get details. ANOVA only can t
est if there is significant between regions. Post-hoc can know more details (such as any
2 regions.)


#because of significant result, so doing Post- hoc test
pairwise.t.test(df$charges, df$region, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$charges and df$region
##
##           northeast northwest southeast
## northwest 1.000     -         -
## southeast 0.901     0.072     -
## southwest 1.000     1.000     0.058
##
## P value adjustment method: bonferroni
```

```
#the result of post-hoc test: there is no p-value < 0.05. It means that after Post hoc t
est, the details actually are not significant.
```

Ihis ANOVA is to check if there are significant differences in charges across different regions. The result indicates that it is significant difference (p-value =0.0309). However, after post-hoc test, there is no single pair shows a statistically significant difference after adjusting for multiple comparisons.

```
# 3.5)  T-test for sex and charges.
# before T-test, the data should convert to binary(0.1) format
# 1 = male, 0 = female

# Perform a t-test to compare charges between males and females
t_test_result_sex <- t.test(charges ~ sex_c, data = df)

# Print the results
print(t_test_result_sex)
```

```
##
##  Welch Two Sample t-test
##
## data:  charges by sex_c
## t = -2.1009, df = 1313.4, p-value = 0.03584
## alternative hypothesis: true difference in means between group 0 and group 1 is not e
qual to 0
## 95 percent confidence interval:
##  -2682.48932   -91.85535
## sample estimates:
## mean in group 0 mean in group 1
##        12569.58        13956.75
```

The t-test has returned a p-value of 0.03584, which is below the conventional alpha level of 0.05, indicating that there is a statistically significant difference in mean charges between the two groups.

The confidence interval for the difference in means does not include zero (91.85535 to 2682.48932), which supports the finding that the difference is significant.

```
# 3.6)  T-test for smoke.
# before T-test, the data should convert to binary(0.1) format
# 0 = yes, 1 = no

t_test_result_smoke <- t.test(charges ~ smoker_c, data = df)

# Print the results
print(t_test_result_smoke)
```

```
##
##  Welch Two Sample t-test
##
## data:  charges by smoker_c
## t = -32.752, df = 311.85, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not e
qual to 0
## 95 percent confidence interval:
##  -25034.71 -22197.21
## sample estimates:
## mean in group 0 mean in group 1
##        8434.268       32050.232
```

T-test is used to test the hypothesis that two populations(smokers vs non-smokers) have equal means. The result is significant different (p< 2.2e-16). It provides strong evidence against the null hypothesis that the means of the two groups are equal.

```
#Correlation Matrix

# Ensure the data frame only are all numeric
numeric_df <- df_clean[sapply(df_clean, is.numeric)]

#correlation matrix
cor_matrix <- cor(numeric_df)

# Print the correlation matrix
print(cor_matrix)
```

```
##                 age         bmi       sex_c     smoker_c    charges
## age       1.00000000 0.109271882 -0.02085587 -0.025018752 0.29900819
## bmi       0.10927188 1.000000000  0.04637115  0.003750426 0.19834097
## sex_c    -0.02085587 0.046371151  1.00000000  0.076184817 0.05729206
## smoker_c -0.02501875 0.003750426  0.07618482  1.000000000 0.78725143
## charges   0.29900819 0.198340969  0.05729206  0.787251430 1.00000000
```

```
# 7) Model for linear regression


# charges vs bmi
model_BMI_Group <- lm(charges ~ bmi , data = df)
summary(model_BMI_Group)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = df)
##
## 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
```

```
# charges vs Region -> does not have sigificant result


# charge vs all
model_all <- lm(charges ~ age+bmi+sex+smoker , data = df)
summary(model_all)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + sex + smoker, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12364.7  -2972.2   -983.2   1475.8  29018.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11633.49     947.27 -12.281   <2e-16 ***
## age            259.45      11.94  21.727   <2e-16 ***
## bmi            323.05      27.53  11.735   <2e-16 ***
## sexmale       -109.04     334.66  -0.326    0.745
## smokeryes    23833.87     414.19  57.544   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6094 on 1333 degrees of freedom
## Multiple R-squared:  0.7475, Adjusted R-squared:  0.7467
## F-statistic: 986.5 on 4 and 1333 DF,  p-value: < 2.2e-16
```

```
# Compare the small models and large model
anova(model_BMI_Group,model_all)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ bmi
## Model 2: charges ~ age + bmi + sex + smoker
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1336 1.8836e+11
## 2   1333 4.9509e+10  3 1.3885e+11 1246.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Interpreation: 'model_all' is best than other two model, because the result of the p-v
alue is sinigicant.
```

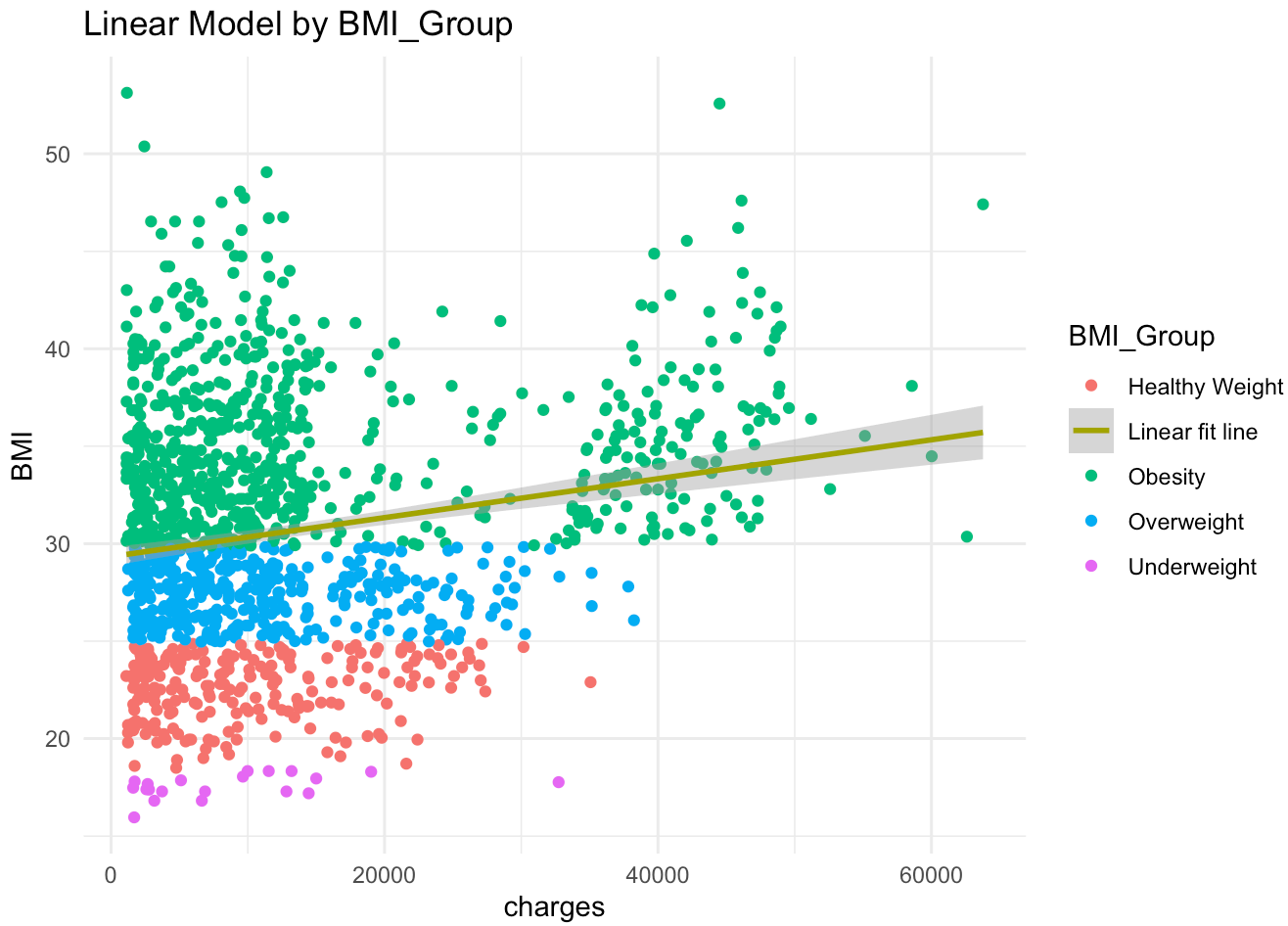
The linear regression of the all variables which have siginificant diffence result: The result shows that “age”, “bmi”, and “smoker_yes” are significant in coefficients. R-squared is around 74%. It indicates that approximately 74.75% of the variance in the charge amounts could be explained by the model’s predictors (age, BMI, sex, and smoking status). The F-statistic is 986.5 on 4 and 1333 DF with a p-value < 2.2e-16, strongly suggesting that the model as a whole is statistically significant.

```
library(ggplot2)

#Linear Model of age and charges
ggplot(df, aes(x = age, y = charges)) +
  geom_point(color ="blue") +
  geom_smooth(method = "lm", se = FALSE, aes(color = "Linear fit line")) +
  theme_minimal()+
  labs(title = "Linear Model Fit", x = "Age", y = "Charges")
```

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


```
#Linear Model of bmi and charges
ggplot(df, aes(x = charges, y = bmi))+
  geom_point(aes(color = BMI_Group))+
  labs(title = "Linear Model by BMI_Group", x = "charges", y = "BMI")+
  theme_minimal()+
  geom_smooth(method = "lm", se = TRUE, aes(color = "Linear fit line"))
```

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




```
library(plotly)
```

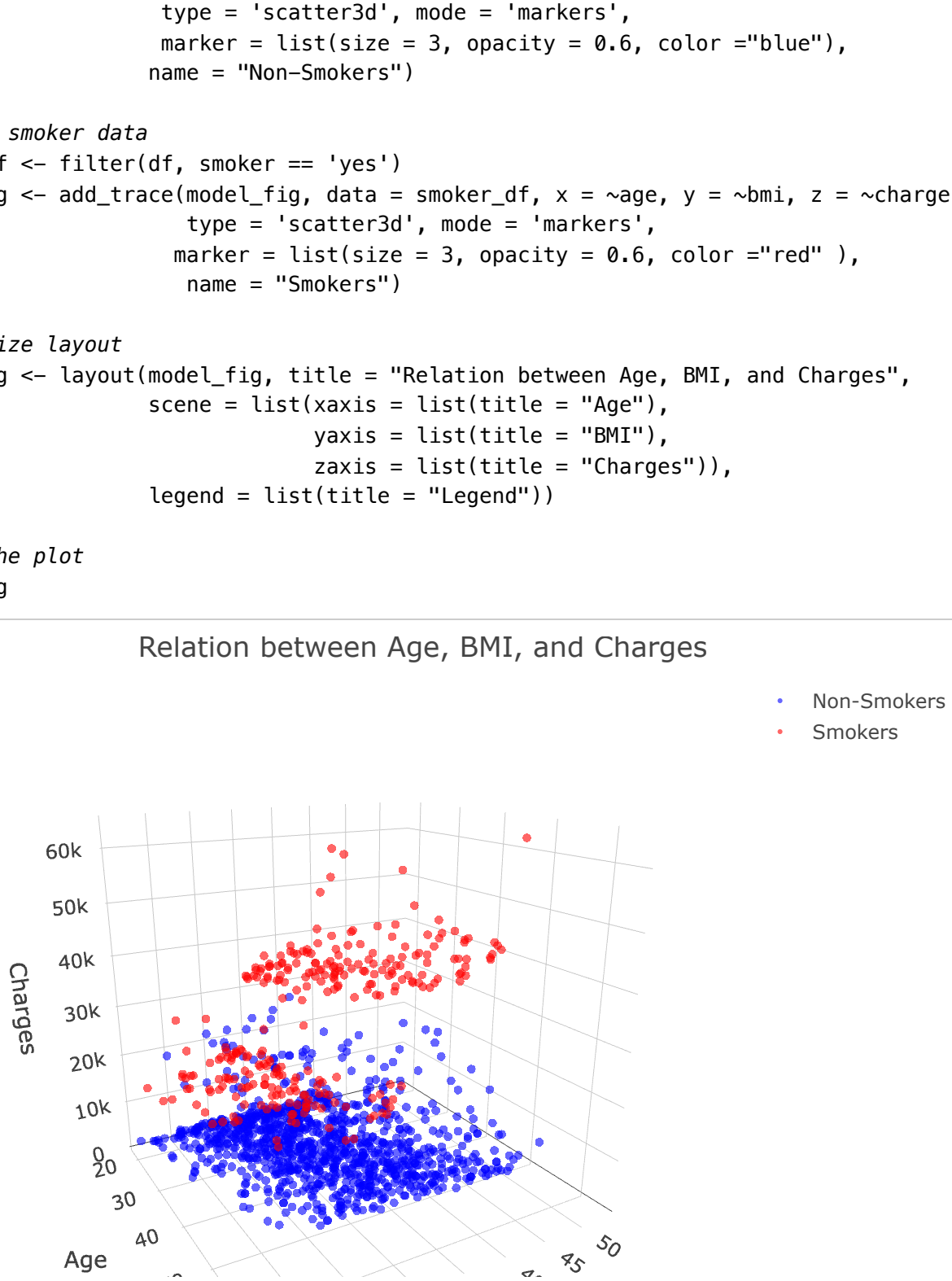
```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
# From the correlation, we can see that 'smoker' has strong correlation with charge. Thus, we can do a model which select 'non-smoker' and find the relation between 'age', bmi' and 'charge'.


# Filter non-smoker data
non_smoker_df <- filter(df, smoker == 'no')

# Create a scatter plot for non-smokers
model_fig <- plot_ly(data = non_smoker_df, x = ~age, y = ~bmi, z = ~charges,
                     type = 'scatter3d', mode = 'markers',
                     marker = list(size = 3, opacity = 0.6, color ="blue"),
                    name = "Non-Smokers")

# Filter smoker data
smoker_df <- filter(df, smoker == 'yes')
model_fig <- add_trace(model_fig, data = smoker_df, x = ~age, y = ~bmi, z = ~charges,
                       type = 'scatter3d', mode = 'markers',
                      marker = list(size = 3, opacity = 0.6, color ="red" ),
                       name = "Smokers")

# Customize layout
model_fig <- layout(model_fig, title = "Relation between Age, BMI, and Charges",
                    scene = list(xaxis = list(title = "Age"),
                                 yaxis = list(title = "BMI"),
                                 zaxis = list(title = "Charges")),
                    legend = list(title = "Legend"))

# Show the plot
model_fig
```




50
60
25
30
35
BMI