

**Klasifikasi dengan Analisis Diskriminan**

**Modul 4 Praktikum**

**Statistika Sains Data**

**Program Studi Sains Data**

**Fakultas Sains**

**Institut Teknologi Sumatera**

**2024**

**Eksplorasi Data**

**Modul Praktikum 1**

**Statistika Sains Data**

**Eksplorasi Data**

**Modul Praktikum 1**

**Statistika Sains Data**

1. **Tujuan Praktikum**
2. Mahasiswa mampu mengklasifikasikan objek baru ke dalam kelompok berdasarkan fungsi diskriminan yang terbentuk.
3. Mahasiswa mampu mengidentifikasi variabel prediktor yang berkontribusi terhadap pemisahan kelompok data.
4. Mahasiswa mampu membuat fungsi diskriminan yang terdiri atas kombinasi linear berbagai variabel prediktor yang dapat memisahkan objek ke dalam kelompok data.
5. **Teori Dasar**

Analisis Diskriminan merupakan salah satu analisis multivariat yang bertujuan untuk memisahkan beberapa objek ke dalam beberapa kelompok atau kategori dengan cara membentuk sebuah fungsi yang memaksimalkan pemisahan antar kelompok tersebut yang biasa disebut sebagai fungsi diskriminan. Fungsi diskrimininan merupakan fungsi yang terdiri atas kombinasi liniear berbagai variabel prediktor. Pada nantinya akan terbentuk fungsi diskrimininan sebanyak jumlah kelas atau kategori dikurangi satu yang dapat memisahkan data. Lebih lanjut lagi, analisis diskriminan merupakan teknik interdependensi dimana informasi kategori atau kelas data sudah diketahui dan ingin dilihat hubungan kategori atau kelas tersebut dengan variabel prediktor. Analisis diskriminan digunakan pada kasus dimana variabel respons berupa data kategorik dan variabel prediktor berupa data numerik.

Secara umum metode analisis diskriminan serupa dengan metode PCA namun tak sama. Persamaannya adalah kedua metode merupakan metode yang mereduksi dimensi data menjadi dimensi yang lebih kecil dengan cara membentuk sebuah persamaan yang terdiri atas kombinasi linear dari berbagai variabel. Perbedaanya adalah pada 1] analisis diskriminan berfokus untuk membentuk persamaan yang dapat memaksimalkan pemisahan antar kelompok dan analisis ini membutuhkan informasi variabel respons berupa data ketegorik guna membentuk persamaannya. Persamaan ini biasa dinyatakan dalam sebuah fungsi yang dinamakan sebagai **fungsi diskriminan**. Di sisi lain, pada 2] PCA berfokus untuk membentuk persamaan yang dapat memaksimalkan & menjelaskan keragaman data dan analisis ini tidak membutuhkan informasi variabel respons saat membentuk persamaannya. Persamaan ini biasa dinyatakan dalam sebuah fungsi yang dinamakan sebagai **principal component**.

**Asumsi pada Analisis Diskriminan**

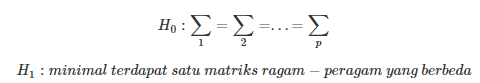
Berbagai asumsi yang diterapkan ketika melakukan analisis diskriminan adalah :

* 1. **Variabel Prediktor berdistribusi Multivariate Normal.**

Uji ini dapat dilakukan dengan mengamati grafik Chi-Square QQ Plot. Jika pada grafik terbentuk garis linear X = Y, maka dapat dikatakan variabel prediktor berdistribusi multivariate normal.

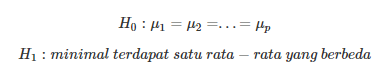
* 1. **Matriks Ragam-peragam Variabel Prediktor Antar Kelompok Sama.**

Dalam prakteknya terdapat kelonggaran pada asumsi ini. Asumsi ini berlaku jika ingin membuat fungsi diskriminan yang linear, namun jika ia tidak terpenuhi atau matriks ragam-peragamnya tidak sama, maka solusi yang dapat dilakukan adalah dengan model diskriminan yang kuadratik. Hipotesis pada asumsi ini adalah :



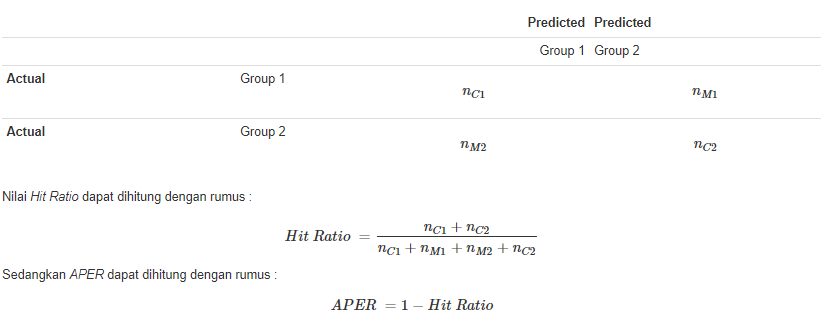
* 1. **Terdapat Perbedaan Rata-rata Antar Kelompok Data.**

Untuk menguji apakah terdapat perbedaan antar kelompok data, dapat menggunakan Uji Manova atau statistik uji Wilk’s Lambda. Hipotesis pada asumsi ini adalah



**Ukuran Performa Model**

Digunakan untuk mengetahui seberapa besar keakurasian model dalam mengklasifikasi suatu objek. Ukuran yang digunakan adalah Hit Ratio atau Apparent Error Rate (APER). &Hit Ratio merupakan proporsi objek yang diklasifikasikan benar oleh model, sedangkan APER kebalikannya, yaitu proporsi objek yang diklasifikan salah oleh model. Untuk memudahkan dalam pengitungannya, perlu dibuat sebuah Confusion Matrix, yaitu matriks tabulasi silang antara kategori sebenarnya dengan kategori yang diprediksi oleh model.



# Contoh Kasus : Pengelompokkan Species Bunga pada Dataset Iris.

Data yang digunakan adalah data Iris yang sudah termuat di dalam R. Data Iris terdiri atas 150 pengamatan bunga Iris dengan variabel : 1] Sepal Length, 2] Sepal Width, 3] Petal Length, 4] Petal Width, dan 5] Species dari bunga Iris tersebut. Tujuan dilakukan analisis adalah untuk mengelompokkan spesies bunga Iris berdasarkan informasi yang tertera dengan metode Analisis Diskriminan.

## **Load Library dan Dataset**

**instal.package("ISLR")**

**instal.package("MASS")**

**instal.package("repr")**

**instal.package("ggplot2")**

**library**(DT) *#Menampilkan tabel agar mudah dilihat di browser*

**library**(MVN) *#Uji multivariate normal*

**library**(MASS) *#Fungsi diskriminan analisis*

**library**(biotools) *#Melakukan uji Box-M*

## ---

## biotools version 3.1

data("iris")

str(iris)

datatable(iris)

Show  entries

Search:

|  | **Sepal.Length** | **Sepal.Width** | **Petal.Length** | **Petal.Width** | **Species** |
| --- | --- | --- | --- | --- | --- |
| 14 | 4.3 | 3 | 1.1 | 0.1 | setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 39 | 4.4 | 3 | 1.3 | 0.2 | setosa |
| 43 | 4.4 | 3.2 | 1.3 | 0.2 | setosa |
| 42 | 4.5 | 2.3 | 1.3 | 0.3 | setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 23 | 4.6 | 3.6 | 1 | 0.2 | setosa |
| 48 | 4.6 | 3.2 | 1.4 | 0.2 | setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |

Showing 1 to 10 of 150 entries

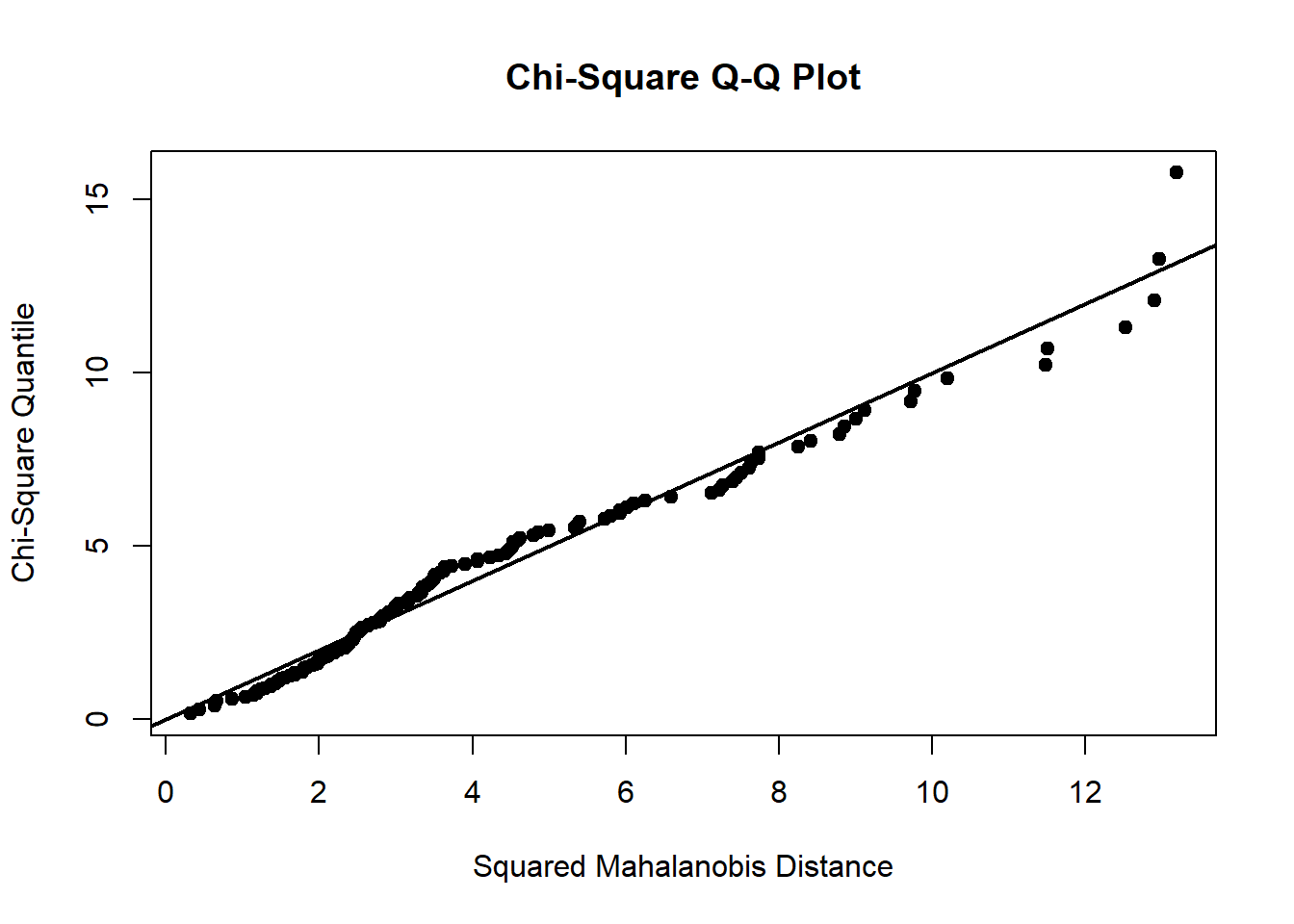
Previous12345…15Next

## **Pengujian Asumsi**

**Multivariate Normal**

Ketika menguji apakah variabel prediktor berdistribusi multivariate normal, di R dapat menggunakan fungsi mvn. Pengujian dilakukan hanya pada variabel prediktor (berskala numerik).

mvn(data = iris[, c(1:4)], multivariatePlot = 'qq') *#hanya mengambil kolom variabel prediktor*



Dari grafik Chi-Square QQ Plot diatas, dapat dilihat bahwasanya secara umum terbentuk garis linear X = Y, maka dapat dikatakan bahwa data berdistribusi multivariate normal.

**Matriks Ragam-peragam antar Kategori Spesies Sama**

Untuk menguji apakah matriks ragam-peragam antar kategori spesies sama, digunakan statistik uji Box’s M. untuk melakukan uji statistik Box’s M di R dapat menggunakan fungsi boxM.

boxM(data = iris[, c(1:4)], grouping = iris[,5])

##

## Box's M-test for Homogeneity of Covariance Matrices

##

## data: iris[, c(1:4)]

## Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16

Output diatas menunjukkan bahwa dengan tingkat signifikansi 5%, didapat keputusan untuk menolak hipotesis nol atau dengan kata lain terdapat perbedaan matriks ragam-peragam antar kategori spesies. Solusinya adalah menggunakan model diskriminan kuadratik, namun pada contoh ini mengabaikan asumsi ini sehingga tetap menggunakan model diskriminan linear.

**Terdapat perbedaan rata-rata antar kategori spesies**

Untuk menguji apakah terdapat perbedaan rata-rata (nilai variabel prediktor) antar kategori spesies, digunakan Uji Manova dengan statistik uji Wilk’s Lambda . Untuk melakukan uji tersebut di R dapat menggunakan fungsi manova dan mengisikan Wilks pada parameter test.

m <- manova(formula = cbind(iris$Sepal.Length, iris$Sepal.Width, iris$Petal.Length,

iris$Petal.Width) ~ iris$Species)

summary(object = m, test = 'Wilks')

## Df Wilks approx F num Df den Df Pr(>F)

## iris$Species 2 0.023439 199.15 8 288 < 2.2e-16 \*\*\*

## Residuals 147

## ---

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

Output diatas menunjukkan bahwa dengan tingkat signifikansi 5%, didapat keputusan untuk menolak hipotesis nol atau dengan kata lain terdapat perbedaan rata-rata (nilai variabel prediktor) antar kategori spesies.

## **Memulai Analisis Diskriminan**

### **Membagi dataset ke dalam Training dan Test.**

Training Data digunakan untuk membuat model diskriminan sedangkan Testing Data digunakan untuk mengevaluasi performa model diskriminan yang terbentuk. Pada contoh ini, dataset Iris akan dibagi menjadi 75% sebagai Training Data dan 25% sebagai Test Data.

set.seed(123)

train\_index <- sample(seq(nrow(iris)), size = floor(0.75 \* nrow(iris)), replace = F)

training\_data <- iris[train\_index, ]

test\_data <- iris[-train\_index, ]

### **Membentuk fungsi diskriminan**

Di dalam R, untuk melakukan analisis diskriminan dapat menggunakan fungsi lda yang terdapat pada library MASS. Model yang dibentuk berdasarkan data yang terdapat pada Training Data.

linearDA <- lda(formula = Species ~., data = training\_data)

linearDA

## Call:

## lda(Species ~ ., data = training\_data)

##

## Prior probabilities of groups:

## setosa versicolor virginica

## 0.3482143 0.3303571 0.3214286

##

## Group means:

## Sepal.Length Sepal.Width Petal.Length Petal.Width

## setosa 4.997436 3.482051 1.448718 0.2487179

## versicolor 5.956757 2.770270 4.308108 1.3405405

## virginica 6.600000 2.997222 5.541667 2.0027778

##

## Coefficients of linear discriminants:

## LD1 LD2

## Sepal.Length 0.5867651 0.004753014

## Sepal.Width 1.6320591 2.388948706

## Petal.Length -1.9853968 -0.666265458

## Petal.Width -2.7922397 2.419828272

##

## Proportion of trace:

## LD1 LD2

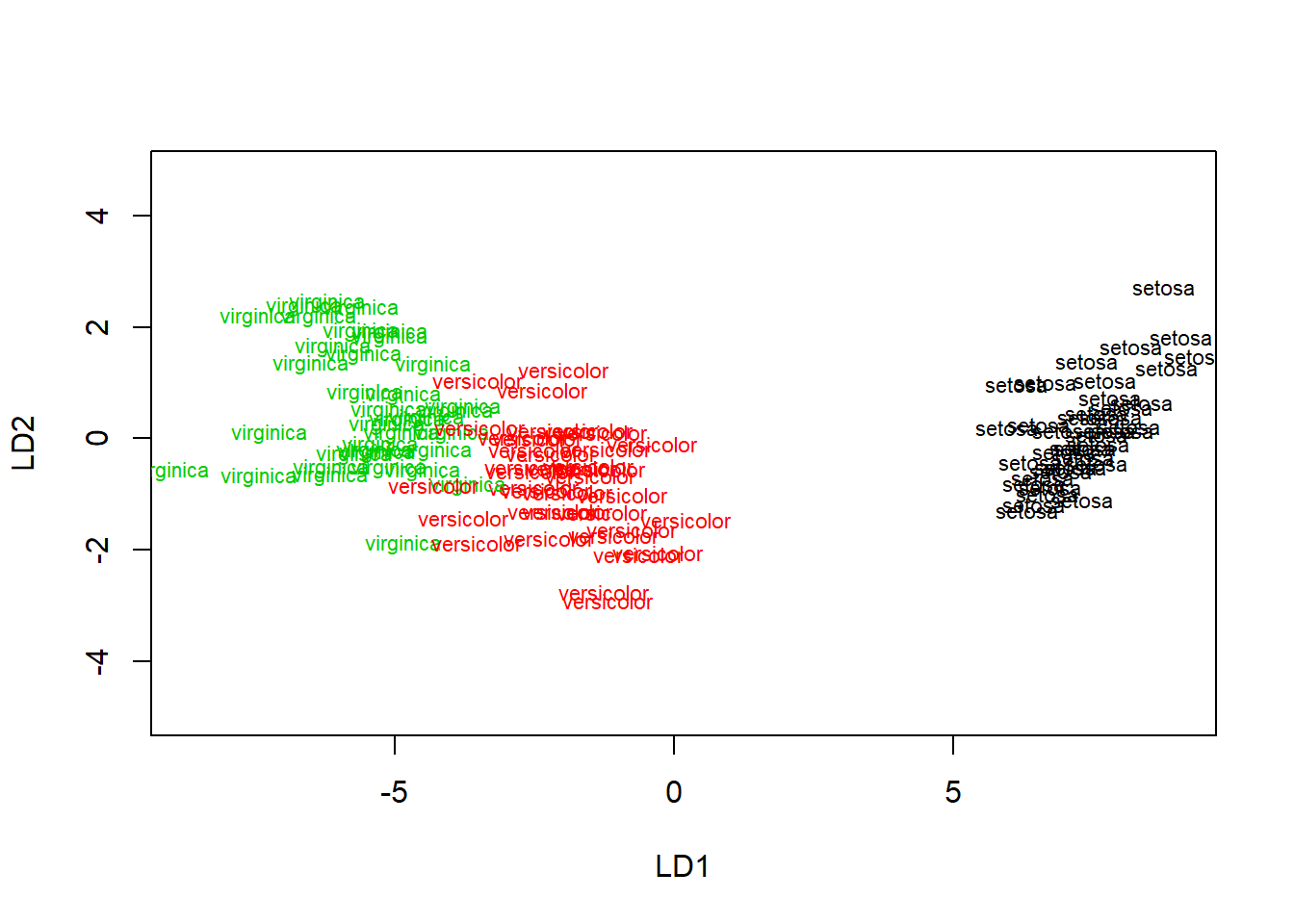
## 0.9898 0.0102

Beberapa output dari fungsi lda adalah sebagai berikut =

* means = rata-rata nilai variabel prediktor pada tiap grup
* priors = peluang yang digunakan (jika tidak disebutkan, maka menggunakan proposi tiap grup)
* scalling = matriks yang berisikan fungsi diskriminan yang dinormalkan

Untuk mengetahui variabel mana yang berpengaruh terhadap perbedaan spesies bunga, salah satu caranya adalah dengan melihat plot antara fungsi diskriminan.

plot(linearDA, col = as.integer(training\_data$Species))



Dilihat dari plot diatas, dapat dikatakan secara umum model mampu mengelompokkan data dengan baik walaupun terdapat sedikit overlap pada kategori Versicolor dan Virginica. Dapat dilihat pula fungsi diskriminan LD1 berperan besar dalam membedakan antara kategori bunga, sedangkan fungsi diskriminan LD2 tidak berperan besar dalam membedakan kategori bunga.

### **Melakukan prediksi di Test Data dan Menguji Performa Model yang dibuat**

Untuk melakukan prediksi menggunakan fungsi predict dari model yang diterapkan kepada Test Data.

predicted <- predict(object = linearDA, newdata = test\_data)

table(actual = test\_data$Species, predicted = predicted$class)

## predicted

## actual setosa versicolor virginica

## setosa 11 0 0

## versicolor 0 13 0

## virginica 0 0 14

Secara keseluruhan model cocok diterapkan, karena model dapat mengklasifikasikan dengan benar seluruh objek yang berada pada Test Data. Nilai Hit Ratio yang diperoleh adalah 1.

1. **Latihan Praktikum**

Pada praktikum kali ini akan dilakukan klasifikasi pada dataset kriminal di Kota Boston, dengan algoritma Linear Discriminant Analysis (LDA).



## **Package**

Silahkan install jika belum ada

instal.package("ISLR")

instal.package("MASS")

instal.package("repr")

instal.package("ggplot2")

Memanggil Package

**library**("ISLR")

**library**("MASS")

**library**("repr")

**library**("ggplot2")

**library**("ROCR")

**library**(class)

head(Boston)

## crim zn indus chas nox rm age dis rad tax ptratio black lstat

## 1 0.00632 18 2.31 0 0.538 6.575 65.2 4.0900 1 296 15.3 396.90 4.98

## 2 0.02731 0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90 9.14

## 3 0.02729 0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83 4.03

## 4 0.03237 0 2.18 0 0.458 6.998 45.8 6.0622 3 222 18.7 394.63 2.94

## 5 0.06905 0 2.18 0 0.458 7.147 54.2 6.0622 3 222 18.7 396.90 5.33

## 6 0.02985 0 2.18 0 0.458 6.430 58.7 6.0622 3 222 18.7 394.12 5.21

## medv

## 1 24.0

## 2 21.6

## 3 34.7

## 4 33.4

## 5 36.2

## 6 28.7

summary(Boston)

## crim zn indus chas

## Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000

## 1st Qu.: 0.08205 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000

## Median : 0.25651 Median : 0.00 Median : 9.69 Median :0.00000

## Mean : 3.61352 Mean : 11.36 Mean :11.14 Mean :0.06917

## 3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10 3rd Qu.:0.00000

## Max. :88.97620 Max. :100.00 Max. :27.74 Max. :1.00000

## nox rm age dis

## Min. :0.3850 Min. :3.561 Min. : 2.90 Min. : 1.130

## 1st Qu.:0.4490 1st Qu.:5.886 1st Qu.: 45.02 1st Qu.: 2.100

## Median :0.5380 Median :6.208 Median : 77.50 Median : 3.207

## Mean :0.5547 Mean :6.285 Mean : 68.57 Mean : 3.795

## 3rd Qu.:0.6240 3rd Qu.:6.623 3rd Qu.: 94.08 3rd Qu.: 5.188

## Max. :0.8710 Max. :8.780 Max. :100.00 Max. :12.127

## rad tax ptratio black

## Min. : 1.000 Min. :187.0 Min. :12.60 Min. : 0.32

## 1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40 1st Qu.:375.38

## Median : 5.000 Median :330.0 Median :19.05 Median :391.44

## Mean : 9.549 Mean :408.2 Mean :18.46 Mean :356.67

## 3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:396.23

## Max. :24.000 Max. :711.0 Max. :22.00 Max. :396.90

## lstat medv

## Min. : 1.73 Min. : 5.00

## 1st Qu.: 6.95 1st Qu.:17.02

## Median :11.36 Median :21.20

## Mean :12.65 Mean :22.53

## 3rd Qu.:16.95 3rd Qu.:25.00

## Max. :37.97 Max. :50.00

## **Generate the response variable**

crime\_rate <- rep(0,506)

crime\_rate[Boston$crim >median(Boston$crim)]=1

df <- Boston[,-1]

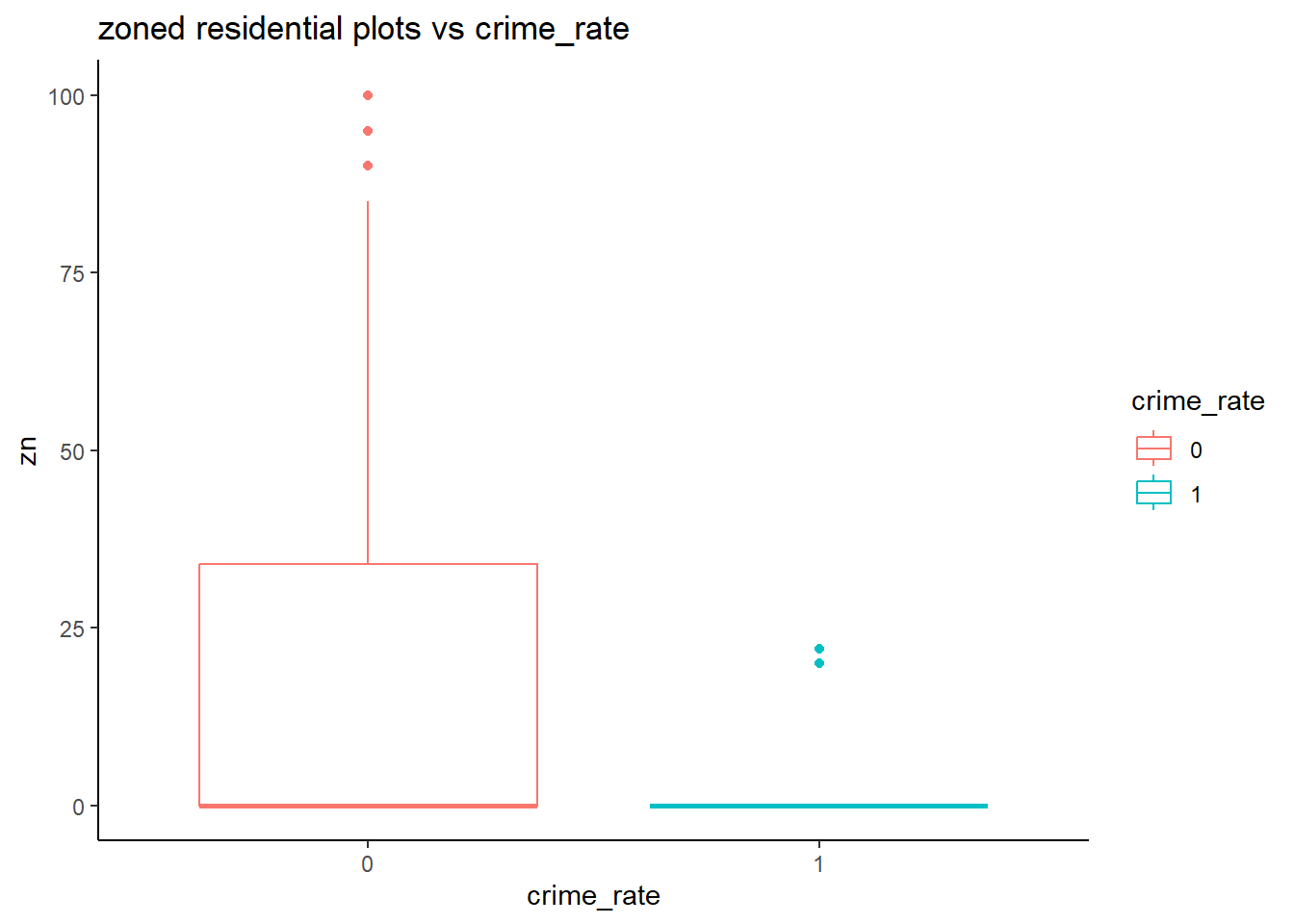
df <- data.frame(df, crime\_rate)

df$crime\_rate <- as.factor(df$crime\_rate)

## **Exploratory data analysis**

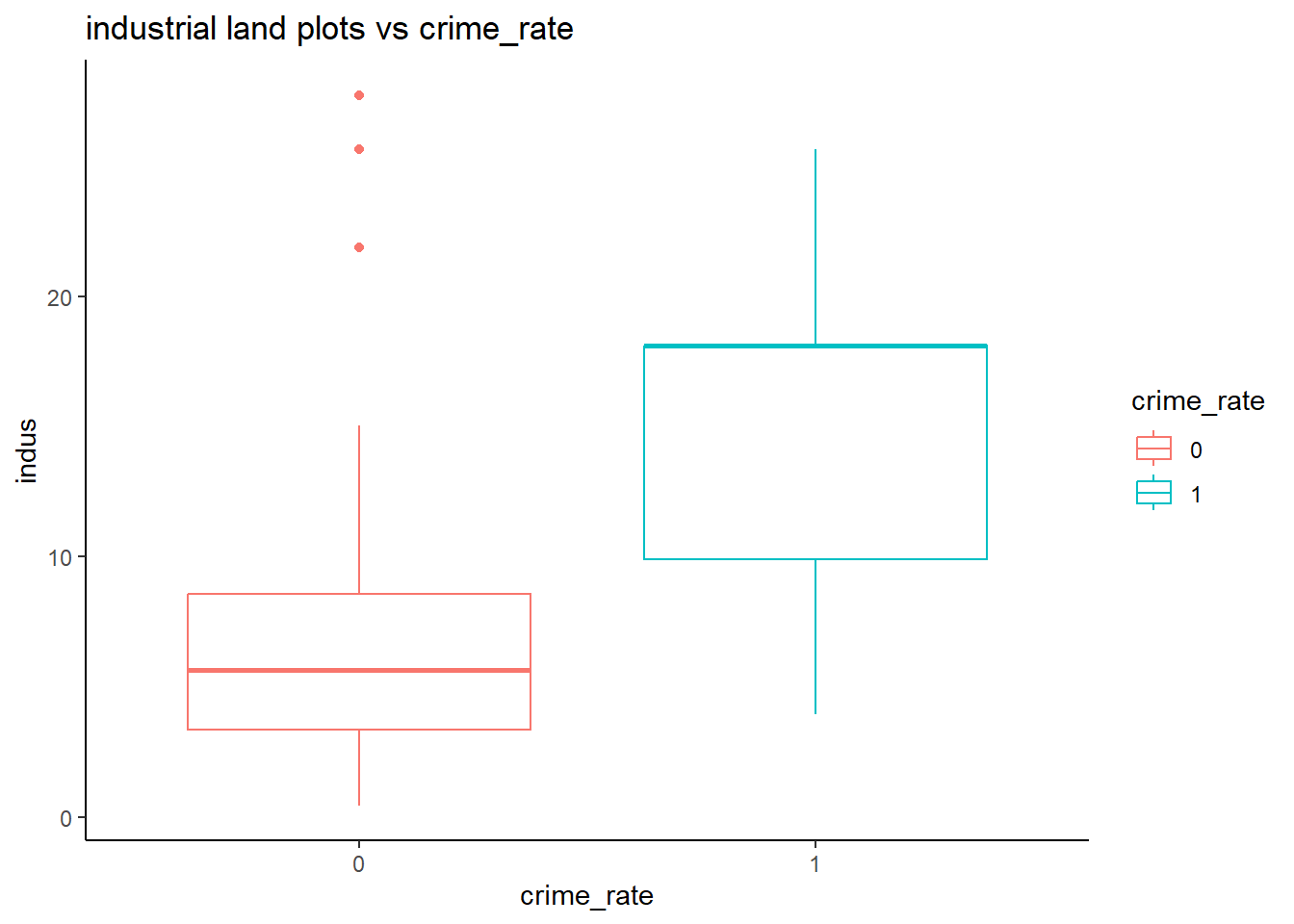
ggplot(df,aes(x=crime\_rate,y=zn, color = crime\_rate))+geom\_boxplot()+theme\_classic()+

labs(title="zoned residential plots vs crime\_rate")



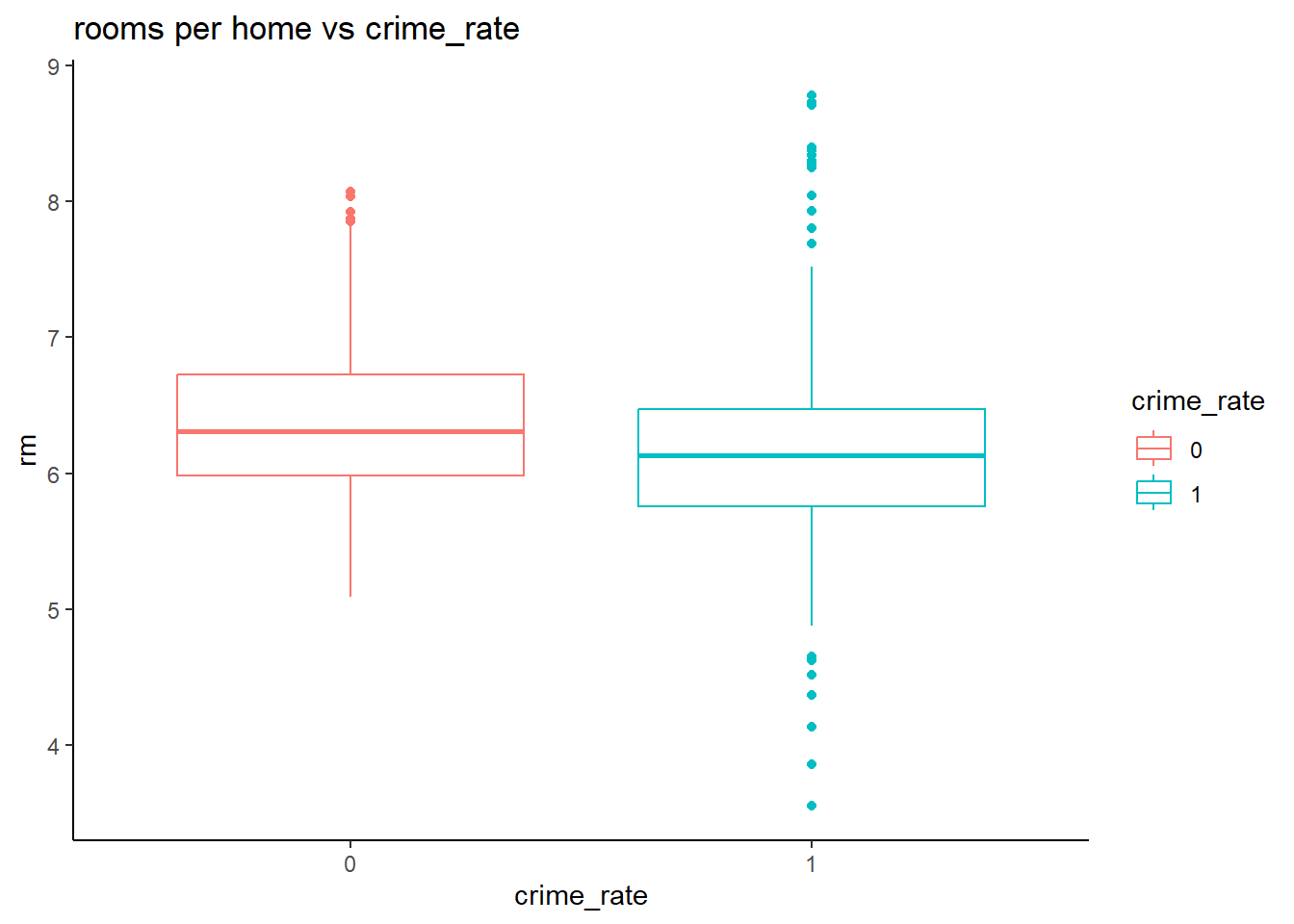
ggplot(df,aes(x=crime\_rate,y=indus, color = crime\_rate))+geom\_boxplot()+theme\_classic()+

labs(title="industrial land plots vs crime\_rate")



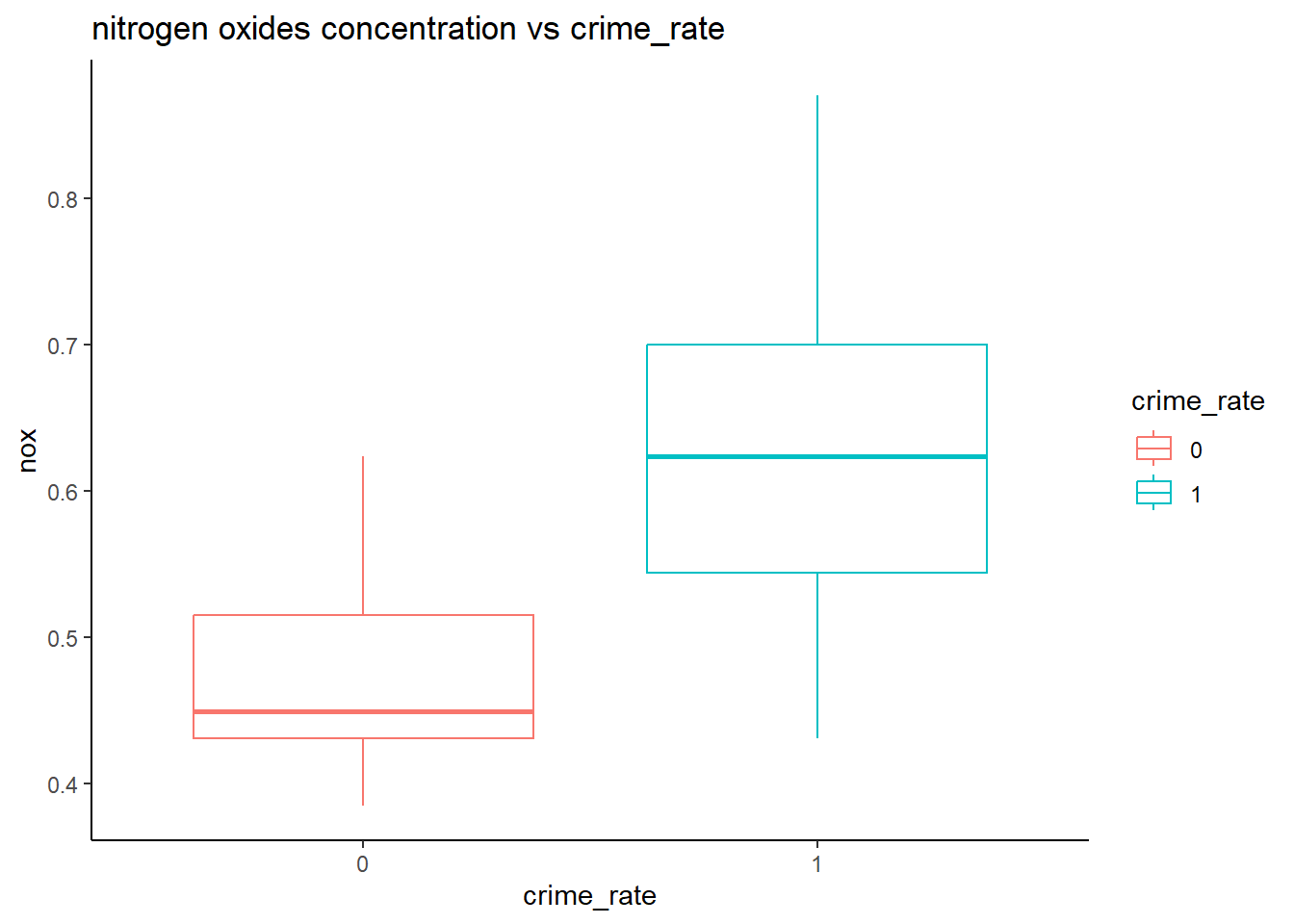
ggplot(df,aes(x=crime\_rate,y=rm, color = crime\_rate))+geom\_boxplot()+theme\_classic()+

labs(title="rooms per home vs crime\_rate")



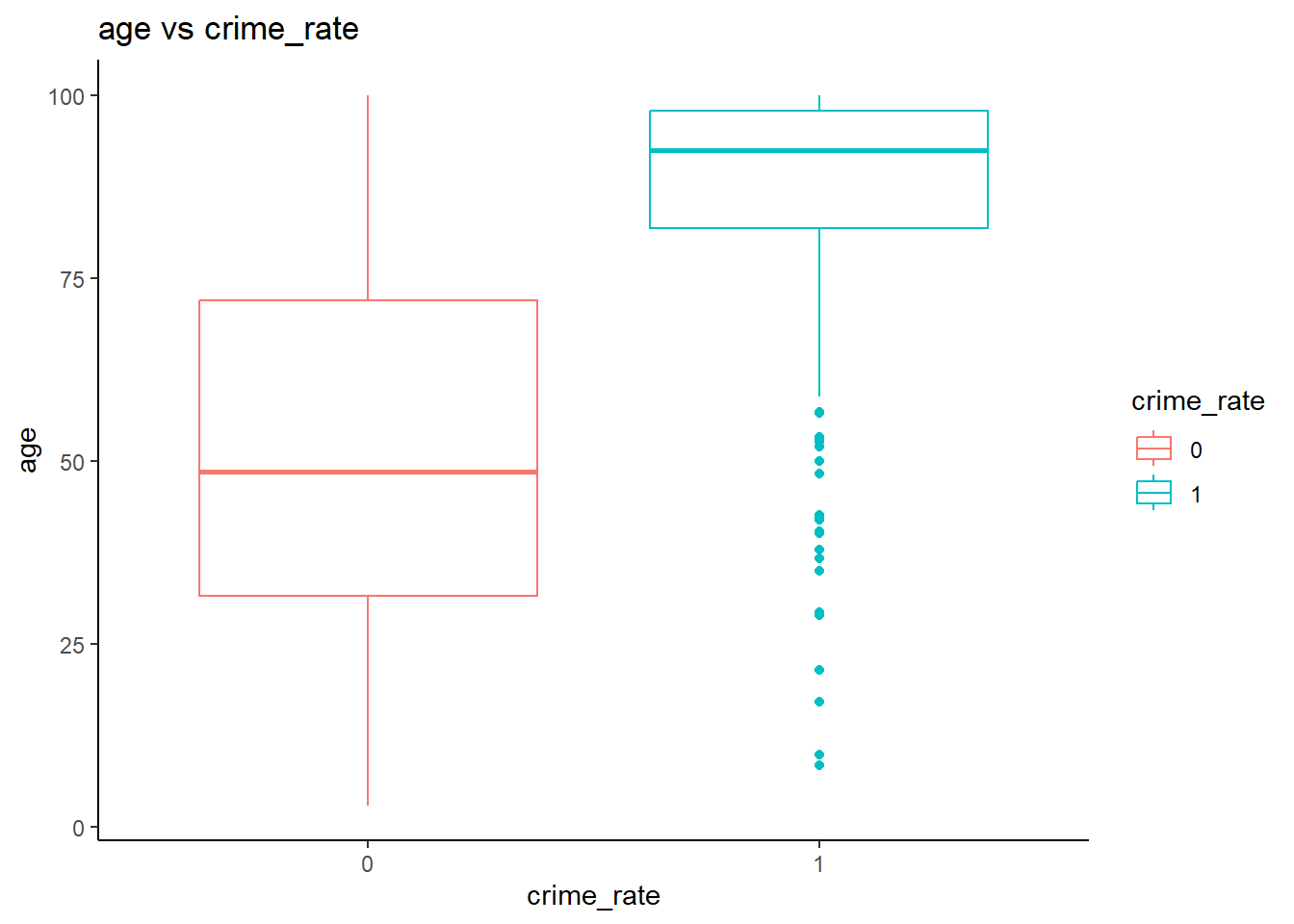
ggplot(df,aes(x=crime\_rate,y=nox, color = crime\_rate))+geom\_boxplot()+theme\_classic()+

labs(title = "nitrogen oxides concentration vs crime\_rate")



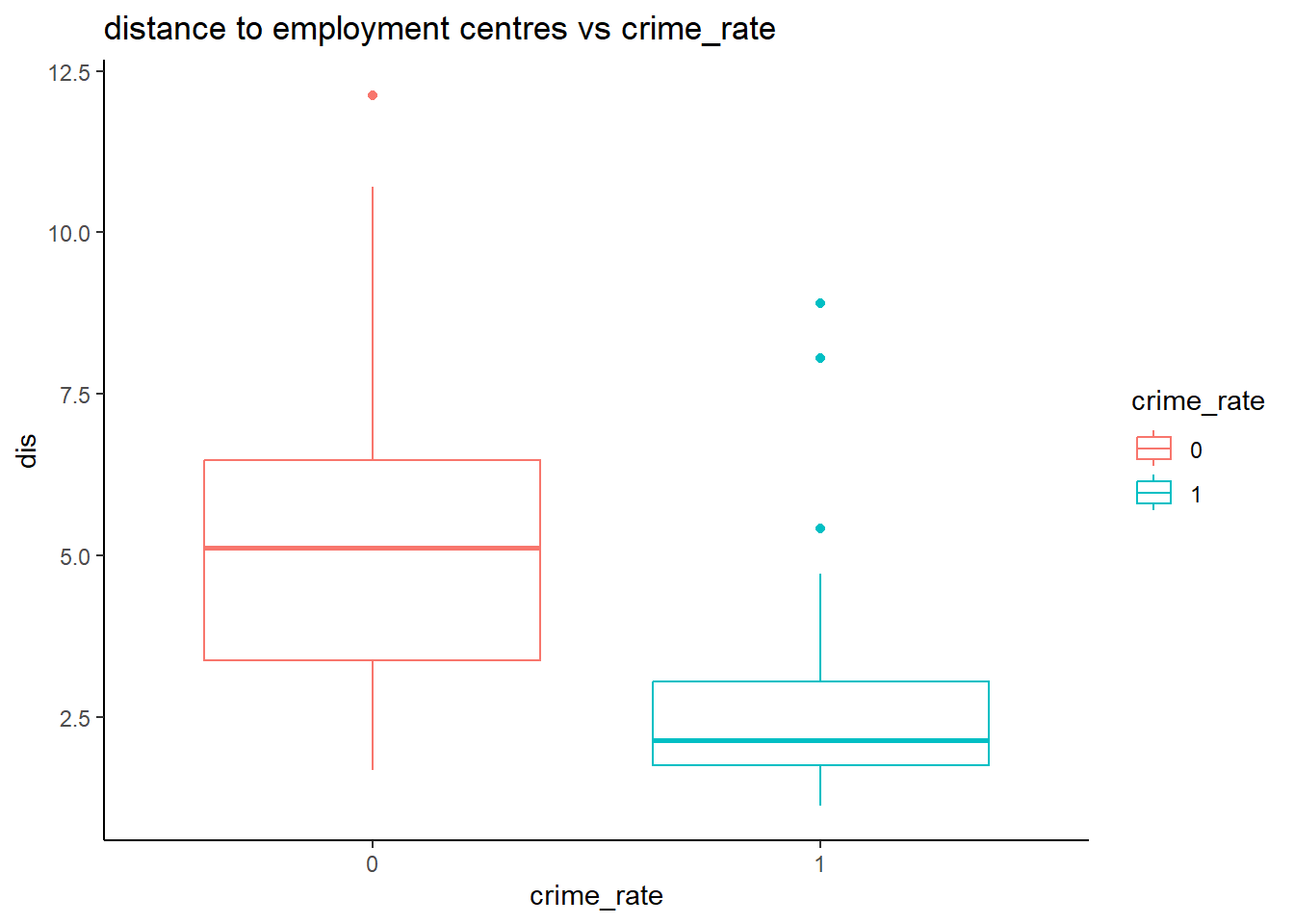
ggplot(df,aes(x=crime\_rate,y=age, color = crime\_rate))+geom\_boxplot()+theme\_classic()+

labs(title = "age vs crime\_rate")



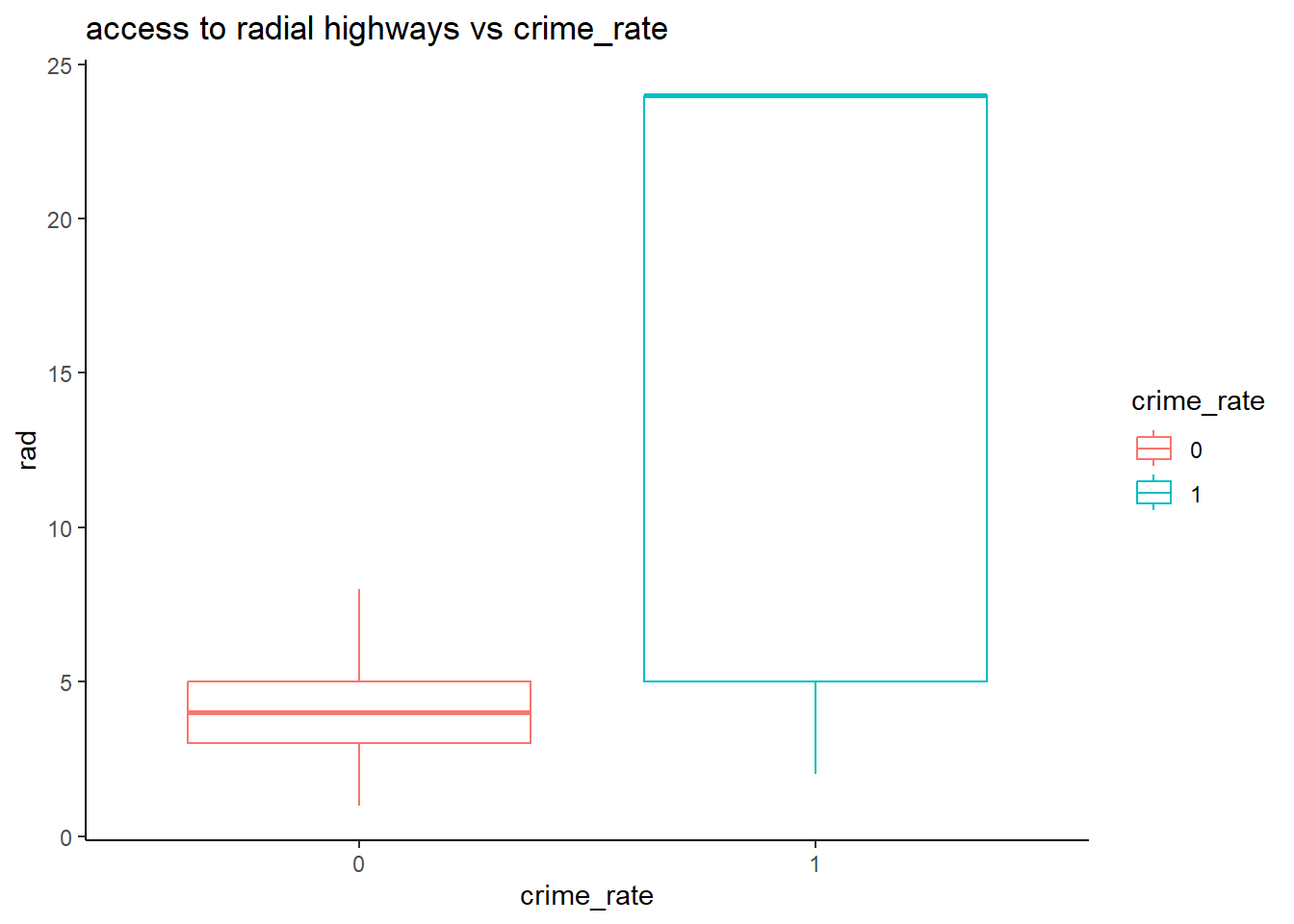
ggplot(df,aes(x=crime\_rate,y=dis, color = crime\_rate))+geom\_boxplot()+theme\_classic()+

labs(title = "distance to employment centres vs crime\_rate")



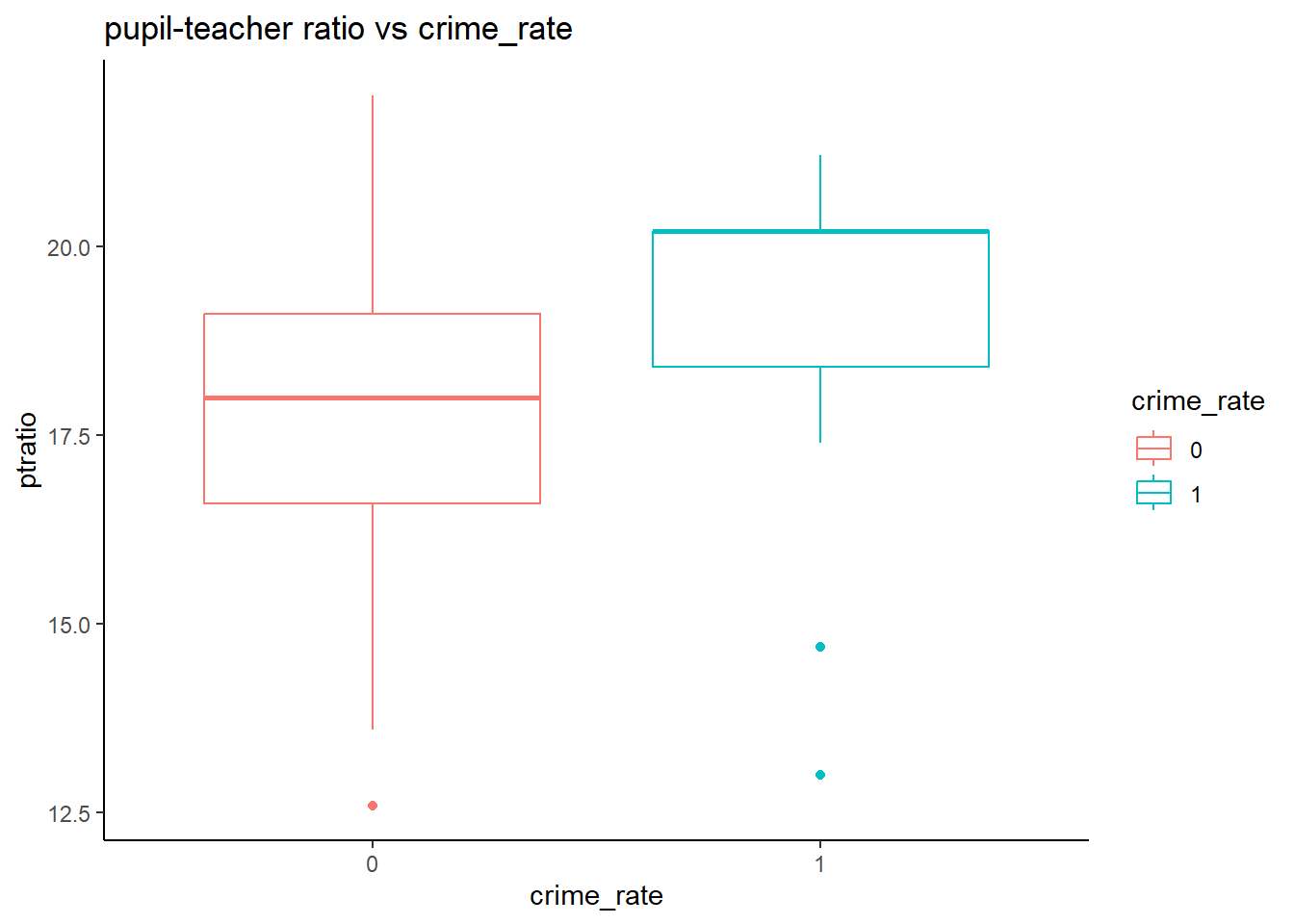
ggplot(df,aes(x=crime\_rate,y=rad,color = crime\_rate))+geom\_boxplot()+theme\_classic() +

labs(title = "access to radial highways vs crime\_rate")



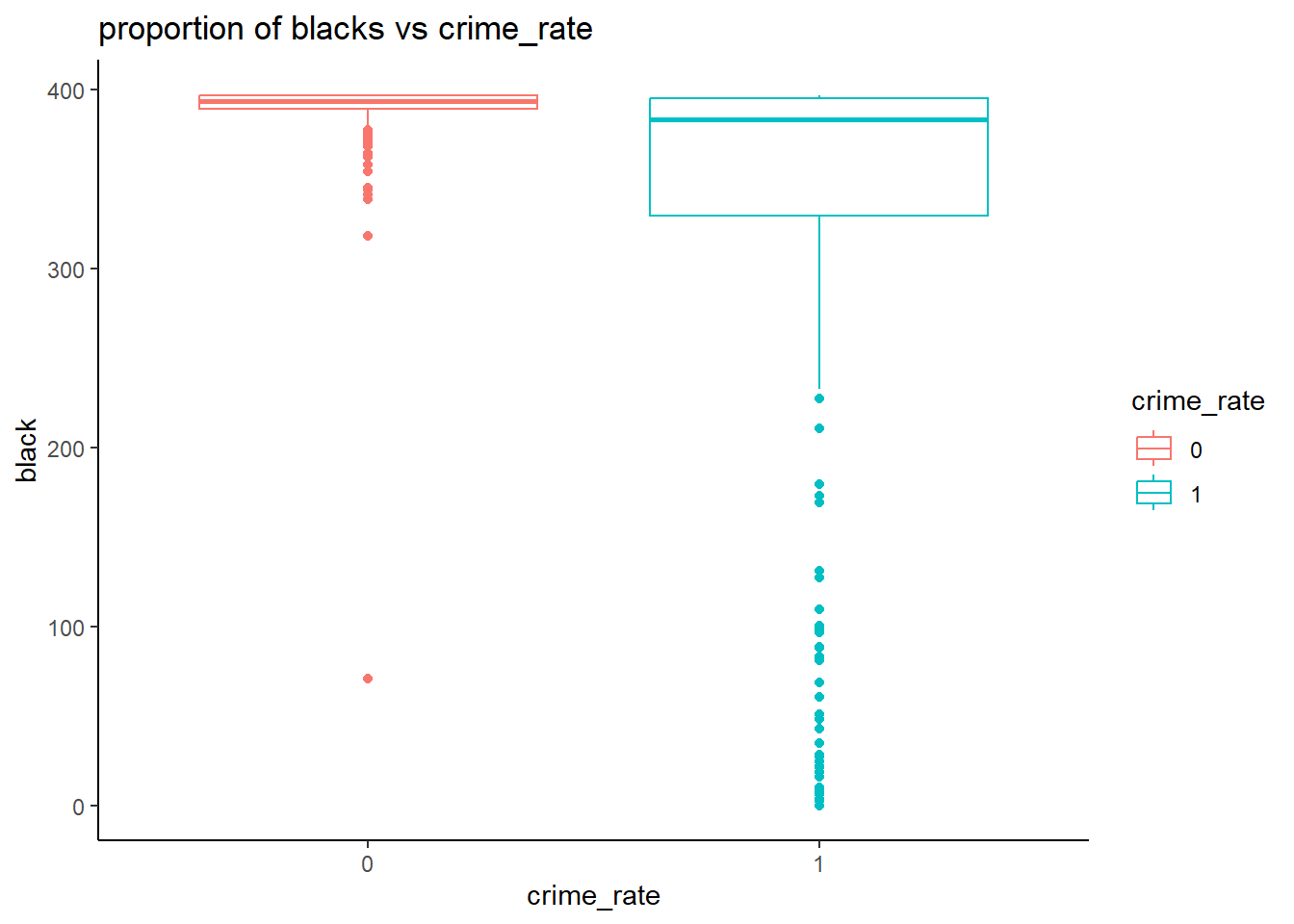
ggplot(df,aes(x=crime\_rate,y=ptratio, color = crime\_rate))+geom\_boxplot()+theme\_classic()+

labs(title = "pupil-teacher ratio vs crime\_rate")



ggplot(df,aes(x=crime\_rate,y=black, color = crime\_rate))+geom\_boxplot()+theme\_classic() +

labs(title = "proportion of blacks vs crime\_rate")



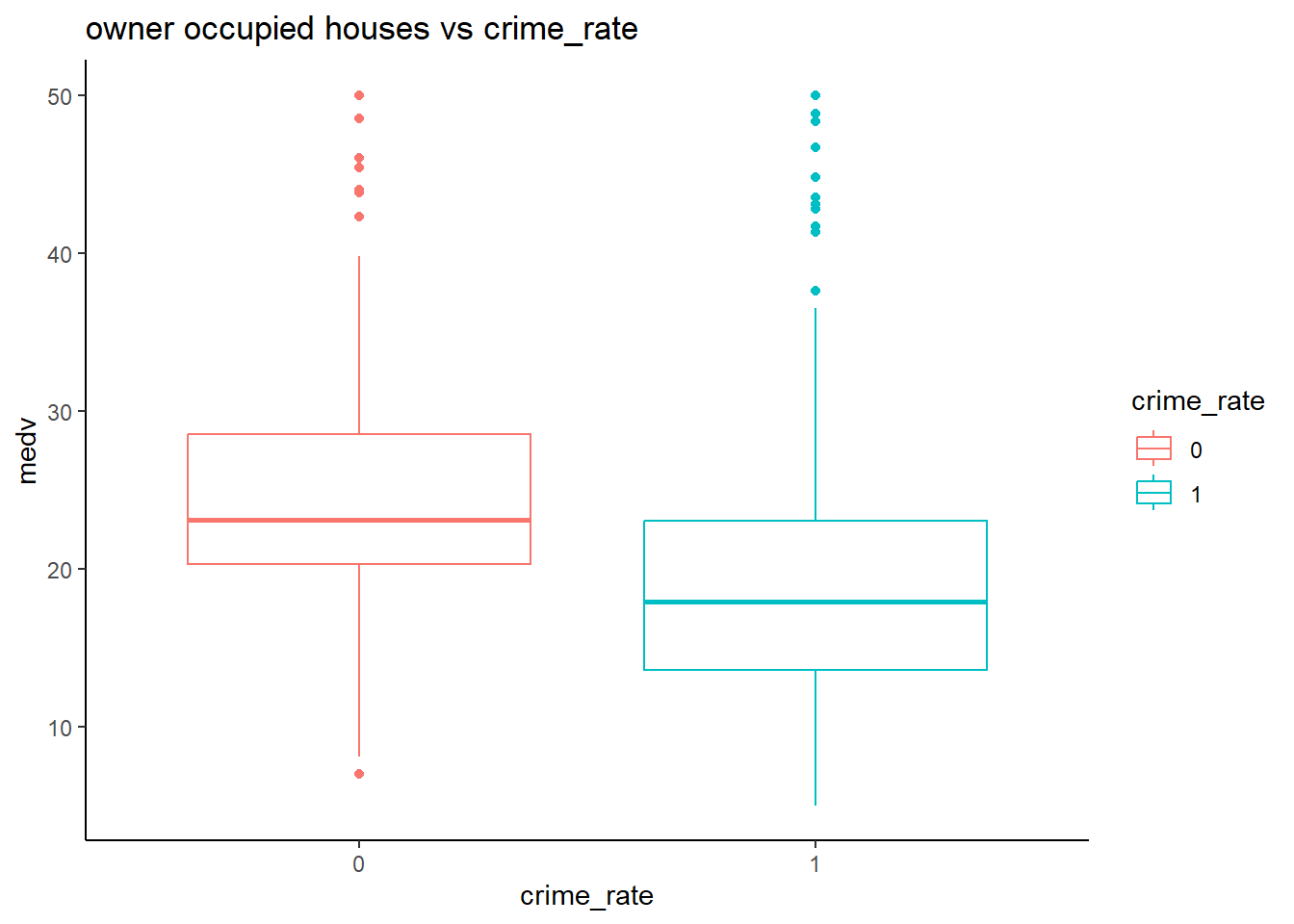
ggplot(df,aes(x=crime\_rate,y=lstat, color = crime\_rate))+geom\_boxplot()+theme\_classic()+

labs(title = "percentage of low income earners vs crime\_rate")



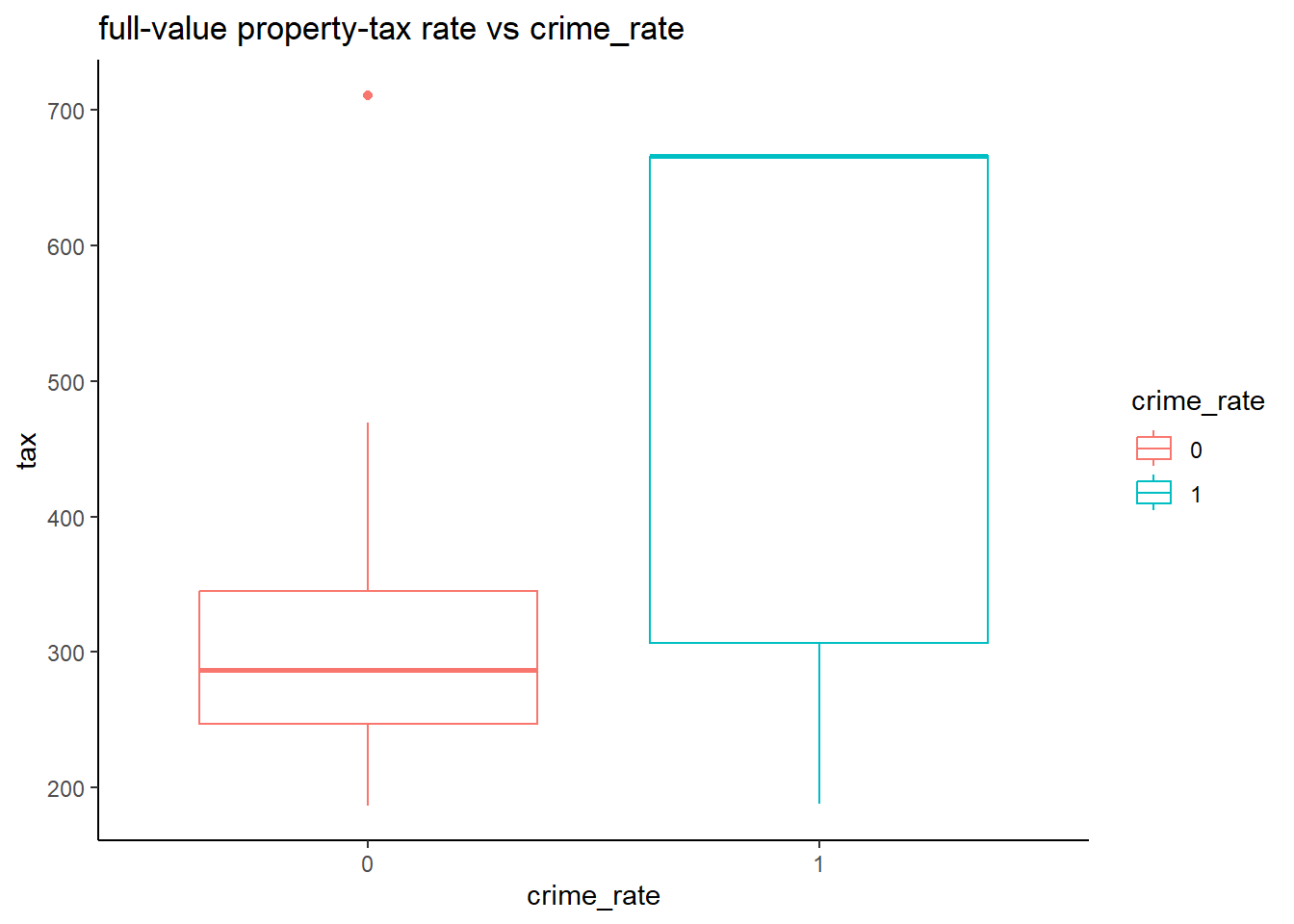
ggplot(df,aes(x=crime\_rate,y=medv, color = crime\_rate))+geom\_boxplot()+theme\_classic()+

labs(title = "owner occupied houses vs crime\_rate")



ggplot(df,aes(x=crime\_rate,y=tax, color = crime\_rate))+geom\_boxplot()+theme\_classic()+

labs(title = "full-value property-tax rate vs crime\_rate")



## **Split the dataset into test and training sets**

set.seed(1110)

df\_split = sort(sample(nrow(df), nrow(df)\*0.8)) *## 80% of the dataset randomly selected*

train<-df[df\_split,]

test<-df[-df\_split,]

## **Linear discriminant analysis (LDA)**

lda.fit=lda(crime\_rate~., data = train)

lda.fit

## Call:

## lda(crime\_rate ~ ., data = train)

##

## Prior probabilities of groups:

## 0 1

## 0.4851485 0.5148515

##

## Group means:

## zn indus chas nox rm age dis rad

## 0 20.596939 6.985561 0.06122449 0.4720398 6.397255 52.23776 5.073745 4.234694

## 1 1.173077 15.388173 0.08653846 0.6407115 6.172274 86.05337 2.500149 15.658654

## tax ptratio black lstat medv

## 0 307.5918 17.89592 389.5360 9.471327 24.96378

## 1 522.7163 19.01202 317.7923 15.791202 19.86827

##

## Coefficients of linear discriminants:

## LD1

## zn -0.006174668

## indus 0.018537416

## chas -0.194227131

## nox 7.434305549

## rm 0.057830566

## age 0.014213350

## dis 0.040855575

## rad 0.091546526

## tax -0.001468892

## ptratio 0.013551006

## black -0.001298476

## lstat -0.004459225

## medv 0.031586482

lda.pred=predict(lda.fit,test)

lda.class =lda.pred$class

table(lda.class, test$crime\_rate)

##

## lda.class 0 1

## 0 54 15

## 1 3 30

### **Predictive accuracy of LDA model**

accuracy.lda <- round(mean(lda.class == test$crime\_rate), digits =2)\*100

print(paste('Accuracy is ',accuracy.lda,"%"))

## [1] "Accuracy is 82 %"

print(paste('Test error is ',100-accuracy.lda,"%"))

## [1] "Test error is 18 %"