HW5-Rohit Thakur

Rohit Thakur

3/19/2020

Problem 1: Author: Maanasa Kaza Source: https://www.kaggle.com/carrie1/ecommerce-data/data#

```
library(tidyr)
library(readr)
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(ggplot2)
library(forcats)
```

```
## Warning: package 'forcats' was built under R version 3.6.3
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```
df<-read.csv("D:/Spring 20 Sem 2/DMP/data.csv")
head(df,n=10)
```

```
##    InvoiceNo StockCode                         Description Quantity
## 1     536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER        6
## 2     536365     71053                 WHITE METAL LANTERN        6
## 3     536365    84406B      CREAM CUPID HEARTS COAT HANGER        8
## 4     536365    84029G KNITTED UNION FLAG HOT WATER BOTTLE        6
## 5     536365    84029E      RED WOOLLY HOTTIE WHITE HEART.        6
## 6     536365     22752        SET 7 BABUSHKA NESTING BOXES        2
## 7     536365     21730   GLASS STAR FROSTED T-LIGHT HOLDER        6
## 8     536366     22633              HAND WARMER UNION JACK        6
```

```
## 9     536366     22632           HAND WARMER RED POLKA DOT        6
## 10    536367     84879       ASSORTED COLOUR BIRD ORNAMENT       32
##       InvoiceDate UnitPrice CustomerID        Country
## 1  12/1/2010 8:26      2.55      17850 United Kingdom
## 2  12/1/2010 8:26      3.39      17850 United Kingdom
## 3  12/1/2010 8:26      2.75      17850 United Kingdom
## 4  12/1/2010 8:26      3.39      17850 United Kingdom
## 5  12/1/2010 8:26      3.39      17850 United Kingdom
## 6  12/1/2010 8:26      7.65      17850 United Kingdom
## 7  12/1/2010 8:26      4.25      17850 United Kingdom
## 8  12/1/2010 8:28      1.85      17850 United Kingdom
## 9  12/1/2010 8:28      1.85      17850 United Kingdom
## 10 12/1/2010 8:34      1.69      13047 United Kingdom
```

Problem 2

```
#Putting into tidy format
df<-separate(df,col="InvoiceDate",into=c("Date","Time"),sep=" ")
head(df)
```

```
##   InvoiceNo StockCode                         Description Quantity      Date
## 1    536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER        6 12/1/2010
## 2    536365     71053                 WHITE METAL LANTERN        6 12/1/2010
## 3    536365    84406B      CREAM CUPID HEARTS COAT HANGER        8 12/1/2010
## 4    536365    84029G KNITTED UNION FLAG HOT WATER BOTTLE        6 12/1/2010
## 5    536365    84029E      RED WOOLLY HOTTIE WHITE HEART.        6 12/1/2010
## 6    536365     22752        SET 7 BABUSHKA NESTING BOXES        2 12/1/2010
##   Time UnitPrice CustomerID        Country
## 1 8:26      2.55      17850 United Kingdom
## 2 8:26      3.39      17850 United Kingdom
## 3 8:26      2.75      17850 United Kingdom
## 4 8:26      3.39      17850 United Kingdom
## 5 8:26      3.39      17850 United Kingdom
## 6 8:26      7.65      17850 United Kingdom
```

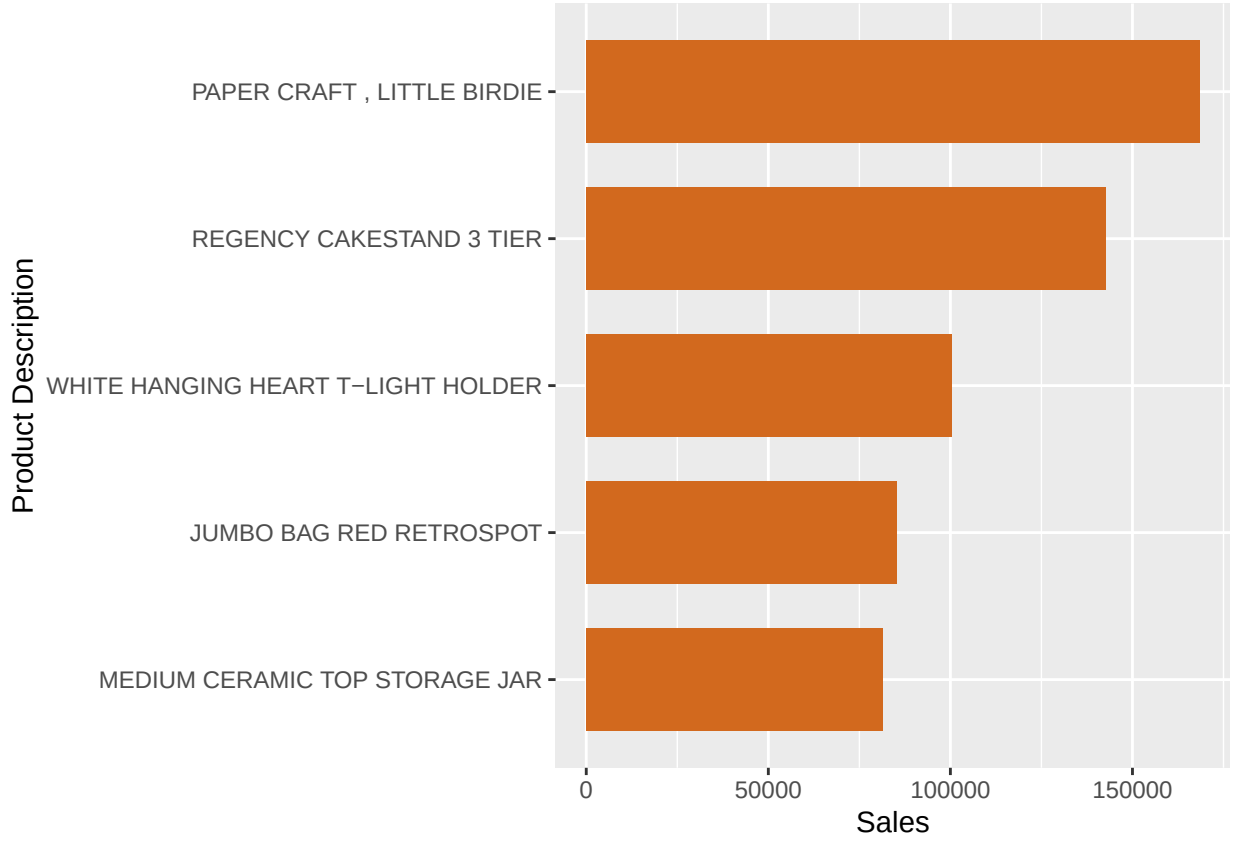
```
df<-filter(df,Description!="")
df<-na.omit(df)
df<-filter(df,UnitPrice > 0)
df<-filter(df,Quantity>0)
head(df, n=10)
```

```
##    InvoiceNo StockCode                         Description Quantity      Date
## 1     536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER        6 12/1/2010
## 2     536365     71053                 WHITE METAL LANTERN        6 12/1/2010
## 3     536365    84406B      CREAM CUPID HEARTS COAT HANGER        8 12/1/2010
## 4     536365    84029G KNITTED UNION FLAG HOT WATER BOTTLE        6 12/1/2010
## 5     536365    84029E      RED WOOLLY HOTTIE WHITE HEART.        6 12/1/2010
## 6     536365     22752        SET 7 BABUSHKA NESTING BOXES        2 12/1/2010
## 7     536365     21730   GLASS STAR FROSTED T-LIGHT HOLDER        6 12/1/2010
## 8     536366     22633              HAND WARMER UNION JACK        6 12/1/2010
## 9     536366     22632           HAND WARMER RED POLKA DOT        6 12/1/2010
## 10    536367     84879       ASSORTED COLOUR BIRD ORNAMENT       32 12/1/2010
##    Time UnitPrice CustomerID        Country
```

```
## 1  8:26      2.55      17850 United Kingdom
## 2  8:26      3.39      17850 United Kingdom
## 3  8:26      2.75      17850 United Kingdom
## 4  8:26      3.39      17850 United Kingdom
## 5  8:26      3.39      17850 United Kingdom
## 6  8:26      7.65      17850 United Kingdom
## 7  8:26      4.25      17850 United Kingdom
## 8  8:28      1.85      17850 United Kingdom
## 9  8:28      1.85      17850 United Kingdom
## 10 8:34      1.69      13047 United Kingdom
```

```
df<-mutate(df,Time=hm(Time))
df1<-df
df1<-df1%>%mutate(Expenditure=Quantity*UnitPrice)
```

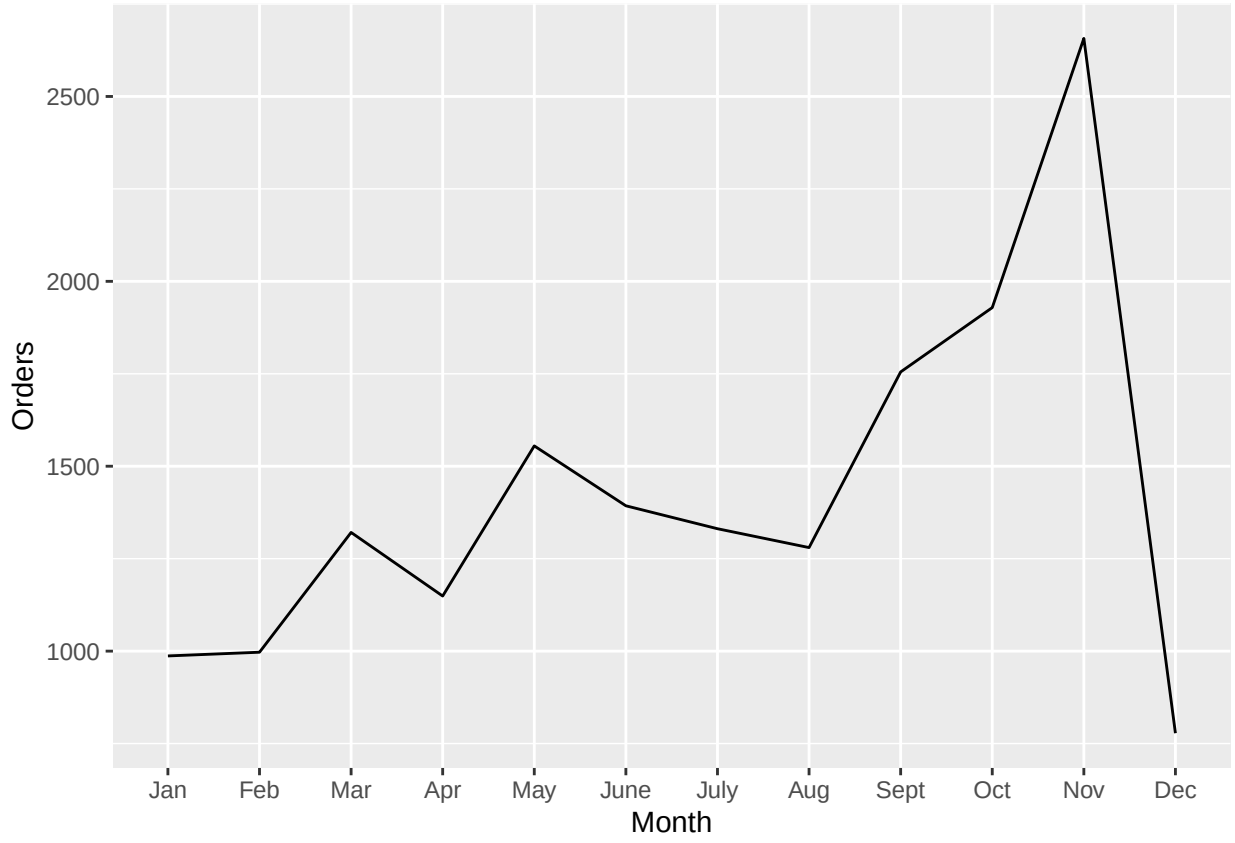
```
df2 <- summarise(group_by(df1, Description), Sales=sum(Expenditure))
df2<-df2%>% arrange(desc(Sales))
df3<-df2[1:5,]
ggplot()+geom_col(aes(x=reorder(df3$Description, df3$Sales),y=df3$Sales)
                  ,fill='chocolate',width=0.7)+
  coord_flip()+xlab("Product Description")+ylab("Sales")
```




```
df4<-df1%>%distinct(InvoiceNo,.keep_all=TRUE)
df4<-separate(df4,col="Date",into=c("mm","dd","yyyy"),sep="/")
df4<-df4%>%filter(yyyy=="2011")
```

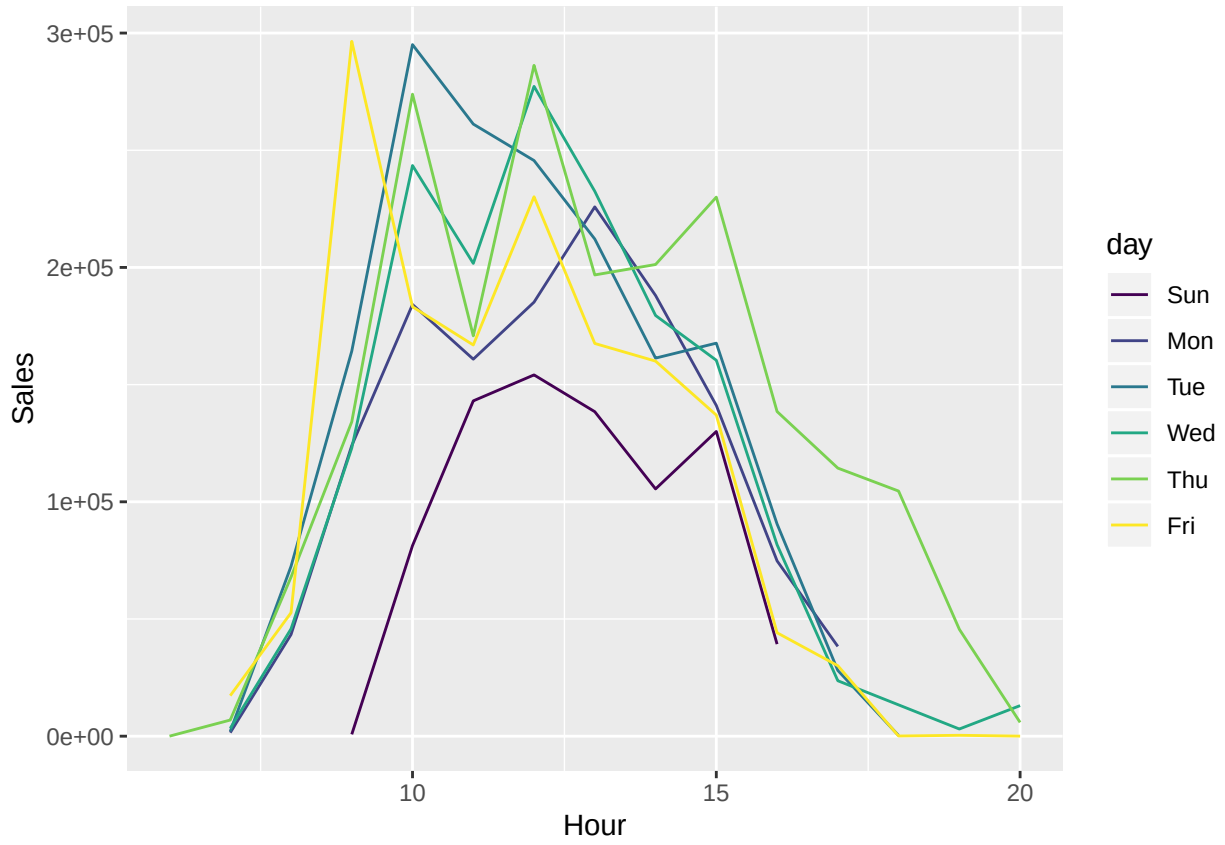
```
df4<-df4%>%group_by(mm)%>%summarise(Orders=n())

df4$mm <- df4$mm%>%factor(levels = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12))
df4<- df4%>%
  mutate(mm=recode(mm, "1"="Jan","2"="Feb","3"="Mar","4"="Apr","5"="May","6"="June",
                       "7"="July","8"="Aug","9"="Sept","10"="Oct","11"="Nov","12"="Dec"))
df4%>%ggplot()+geom_line(aes(x=mm,y=Orders,group=1))+xlab("Month")
```




```
df5<-select(df1,Date,Time,Expenditure)
df5$day <- wday(as.Date(df5$Date,'%m/%d/%Y'), label = TRUE)


df5<-group_by(df5,hour(Time),day)%>%summarise(total=sum(Expenditure))
plot<- ggplot(data = df5 ,
                      mapping = aes(x = `hour(Time)`, y = total, group=day)) +
geom_line(stat = 'identity', aes(color = day)) +
labs(x = 'Hour', y = 'Sales')
print(plot)
```



Problem 3

```
library(mlbench)
library(modelr)
library(purrr)
data(PimaIndiansDiabetes2)
cv<-function(formula,data,folds){
  set.seed(1)
  data_cv <- crossv_kfold(data,folds)
  data_cv <- data_cv%>%
              mutate(fit = map(train,
                 ~ lm(formula,data=data)))
  data_cv <- data_cv %>%
            mutate(rmse_test = map2_dbl(fit, test, ~ rmse(.x,.y)))
  mean_error<-mean(data_cv$rmse_test)
  return(mean_error)
}

cv(mass~pressure+triceps+diabetes+pregnant,PimaIndiansDiabetes2,5)
```

```
## [1] 4.964229
```

Problem 4

First fit for each predictor variable

```
rmse_triceps<-cv(mass~triceps,PimaIndiansDiabetes2,5)
print(rmse_triceps)
```

```
## [1] 5.222564
```

```
rmse_diabetes<-cv(mass~diabetes,PimaIndiansDiabetes2,5)
print(rmse_diabetes)
```

```
## [1] 6.555314
```

```
rmse_pressure<-cv(mass~pressure,PimaIndiansDiabetes2,5)
print(rmse_pressure)
```

```
## [1] 6.57702
```

```
rmse_insulin<-cv(mass~insulin,PimaIndiansDiabetes2,5)
print(rmse_insulin)
```

```
## [1] 6.782915
```

```
rmse_glucose<-cv(mass~glucose,PimaIndiansDiabetes2,5)
print(rmse_glucose)
```

```
## [1] 6.71721
```

triceps is the best predictor with rmse 5.222

```
rmse_diabetes2<-cv(mass~triceps+diabetes,PimaIndiansDiabetes2,5)
print(rmse_diabetes2)
```

```
## [1] 5.124542
```

```
rmse_pressure2<-cv(mass~triceps+pressure,PimaIndiansDiabetes2,5)
print(rmse_pressure2)
```

```
## [1] 5.096432
```

```
rmse_insulin2<-cv(mass~triceps+insulin,PimaIndiansDiabetes2,5)
print(rmse_insulin2)
```

```
## [1] 5.167705
```

```
rmse_glucose2<-cv(mass~triceps+glucose,PimaIndiansDiabetes2,5)
print(rmse_glucose2)
```

```
## [1] 5.180415
```

Pressure is best predictor with rmse 5.09 we will add it in our model

```
rmse_diabetes3<-cv(mass~triceps+pressure+diabetes,PimaIndiansDiabetes2,5)
print(rmse_diabetes3)
```

```
## [1] 5.026427
```

```
rmse_insulin3<-cv(mass~triceps+pressure+insulin,PimaIndiansDiabetes2,5)
print(rmse_insulin3)
```

```
## [1] 5.070169
```

```
rmse_glucose3<-cv(mass~triceps+pressure+glucose,PimaIndiansDiabetes2,5)
print(rmse_glucose3)
```

```
## [1] 5.080728
```

we will further add diabetes in our model because it reduces rmse to 5.02.

```
rmse_insulin4<-cv(mass~triceps+pressure+diabetes+insulin,PimaIndiansDiabetes2,5)
print(rmse_insulin4)
```

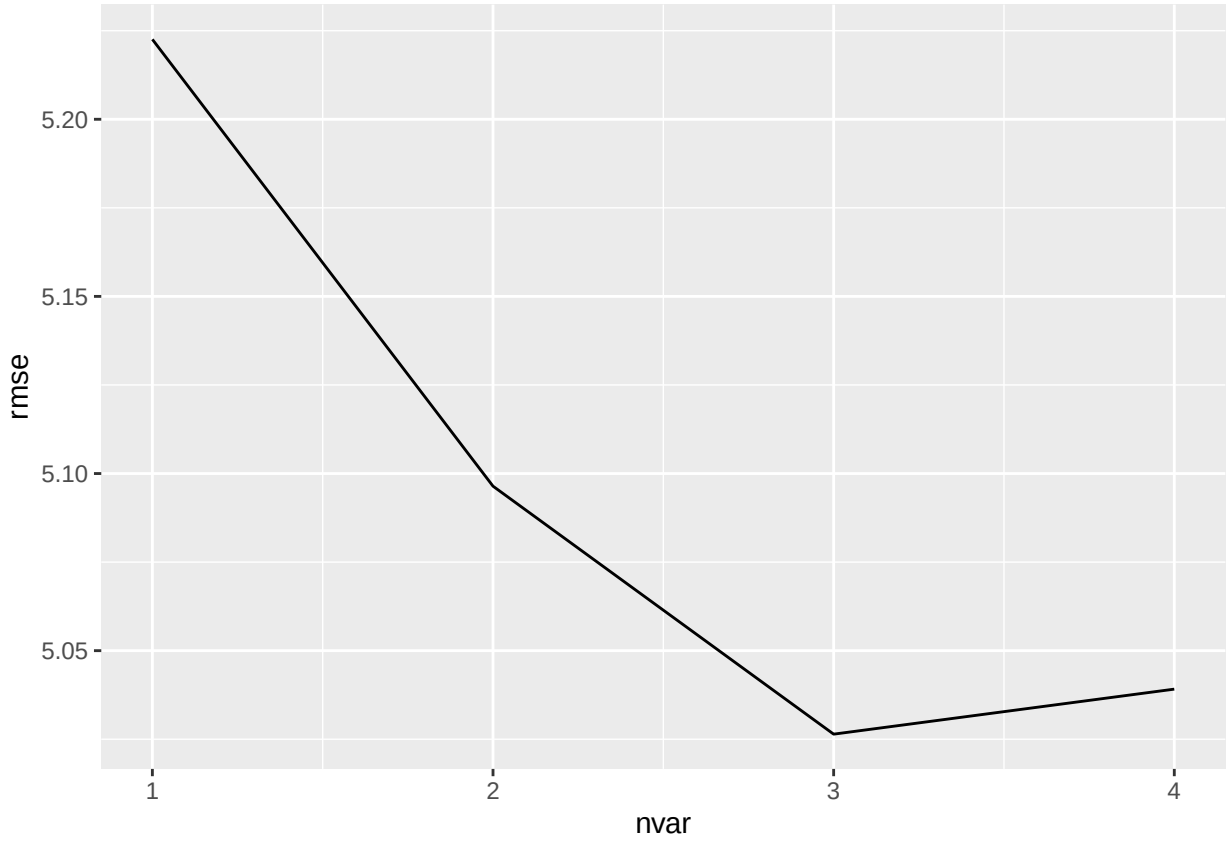
```
## [1] 5.053876
```

```
rmse_glucose4<-cv(mass~triceps+pressure+diabetes+glucose,PimaIndiansDiabetes2,5)
print(rmse_glucose4)
```

```
## [1] 5.039135
```

Addition of predictor variables further does not improve model. Thus final model will be mass~triceps+pressure+diabetes with rmse 5.02

```
fits_rmse <- tibble(nvar = 1:4,
                    rmse = c(rmse_triceps,rmse_pressure2,rmse_diabetes3,rmse_glucose4))
ggplot(fits_rmse) + geom_line(aes(x=nvar, y=rmse))
```



In homework 4 proposed model for response variable mass was using 4 predictor variables(pressure,triceps,diabetes,pregnant) it had rmse 4.97 while current model with predictor variables(triceps,pressure,diabetes) gives rmse 5.02 which is slightly worse than previous model.

PROBLEM 5

Final cross validated rmse for best model we found in problem 4 CANNOT be reported as good measure of rmse on new data. One of the reason is that best model for this data may not provide good estimate for new set of data. For RMSE to be a good estimate , we should calculate it on set of test data so as to make sure that it was not used to select the best model.