Airfare Prediction

Ajinkya

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
pacman::p_load(forecast, tidyverse, gplots, GGally, mosaic,
               scales, mosaic, mapproj, caret, data.table,reshape, reshape2, leaps,tidyverse,MASS)



airfare <- read.csv("Airfares.csv")
air.dt <- setDT(airfare)
names (air.dt)
```

```
##  [1] "S_CODE"   "S_CITY"   "E_CODE"   "E_CITY"   "COUPON"   "NEW"
##  [7] "VACATION" "SW"       "HI"       "S_INCOME" "E_INCOME" "S_POP"
## [13] "E_POP"    "SLOT"     "GATE"     "DISTANCE" "PAX"      "FARE"
```

1. Create a correlation table and scatterplots between FARE and the predictors. What seems to be the best single predictor of FARE? Explain your answer.

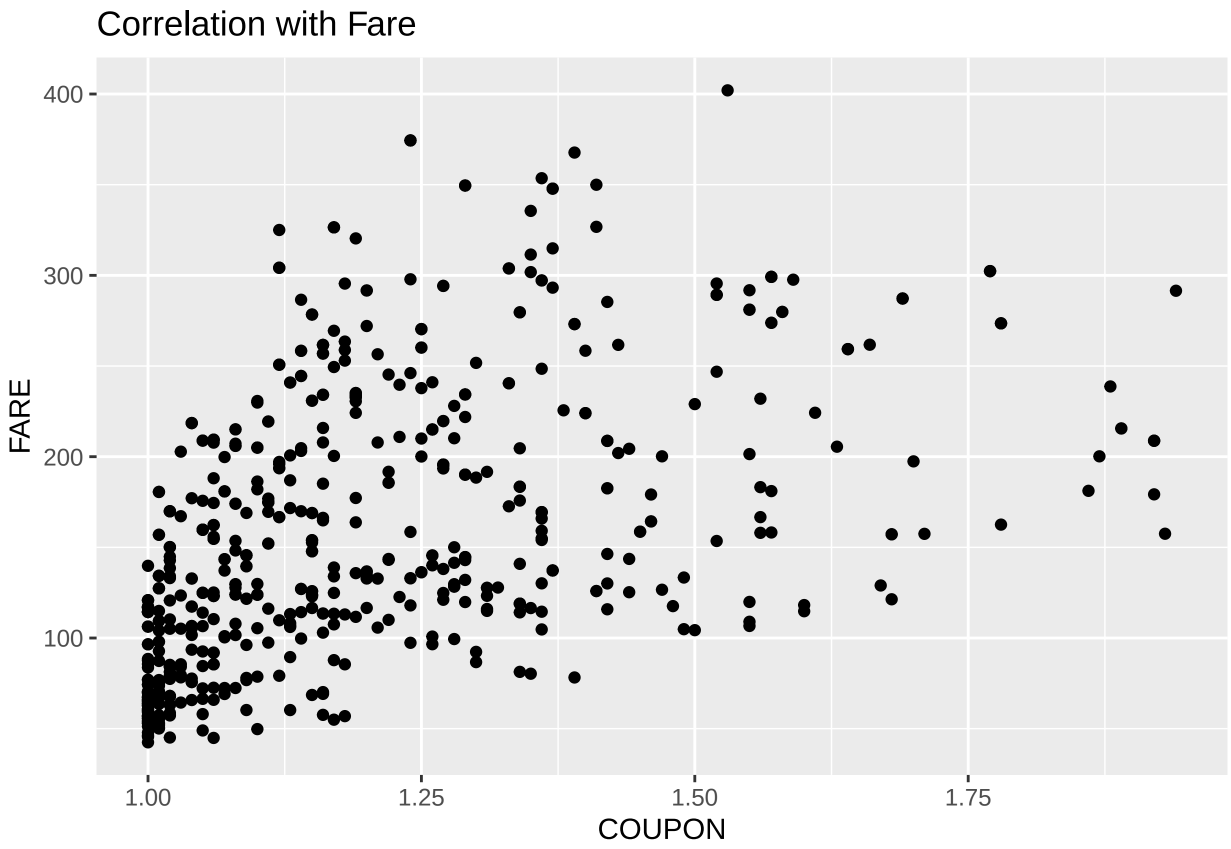
```
#1. Correlation plot



View(air.dt)
air.dt<-air.dt[,-(1:4)]
str(air.dt)
```

```
## Classes 'data.table' and 'data.frame':   638 obs. of  14 variables:
##  $ COUPON  : num  1 1.06 1.06 1.06 1.06 1.01 1.28 1.15 1.33 1.6 ...
##  $ NEW     : int  3 3 3 3 3 3 3 3 3 2 ...
##  $ VACATION: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ SW      : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 1 2 2 2 ...
##  $ HI      : num  5292 5419 9185 2657 2657 ...
##  $ S_INCOME: num  28637 26993 30124 29260 29260 ...
##  $ E_INCOME: num  21112 29838 29838 29838 29838 ...
##  $ S_POP   : int  3036732 3532657 5787293 7830332 7830332 2230955 3036732 1440377 3770125 1694803 ..
##  $ E_POP   : int  205711 7145897 7145897 7145897 7145897 7145897 7145897 7145897 7145897 7145897 ...
##  $ SLOT    : Factor w/ 2 levels "Controlled","Free": 2 2 2 1 2 2 2 2 2 2 ...
##  $ GATE    : Factor w/ 2 levels "Constrained",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ DISTANCE: int  312 576 364 612 612 309 1220 921 1249 964 ...
##  $ PAX     : int  7864 8820 6452 25144 25144 13386 4625 5512 7811 4657 ...
##  $ FARE    : num  64.1 174.5 207.8 85.5 85.5 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

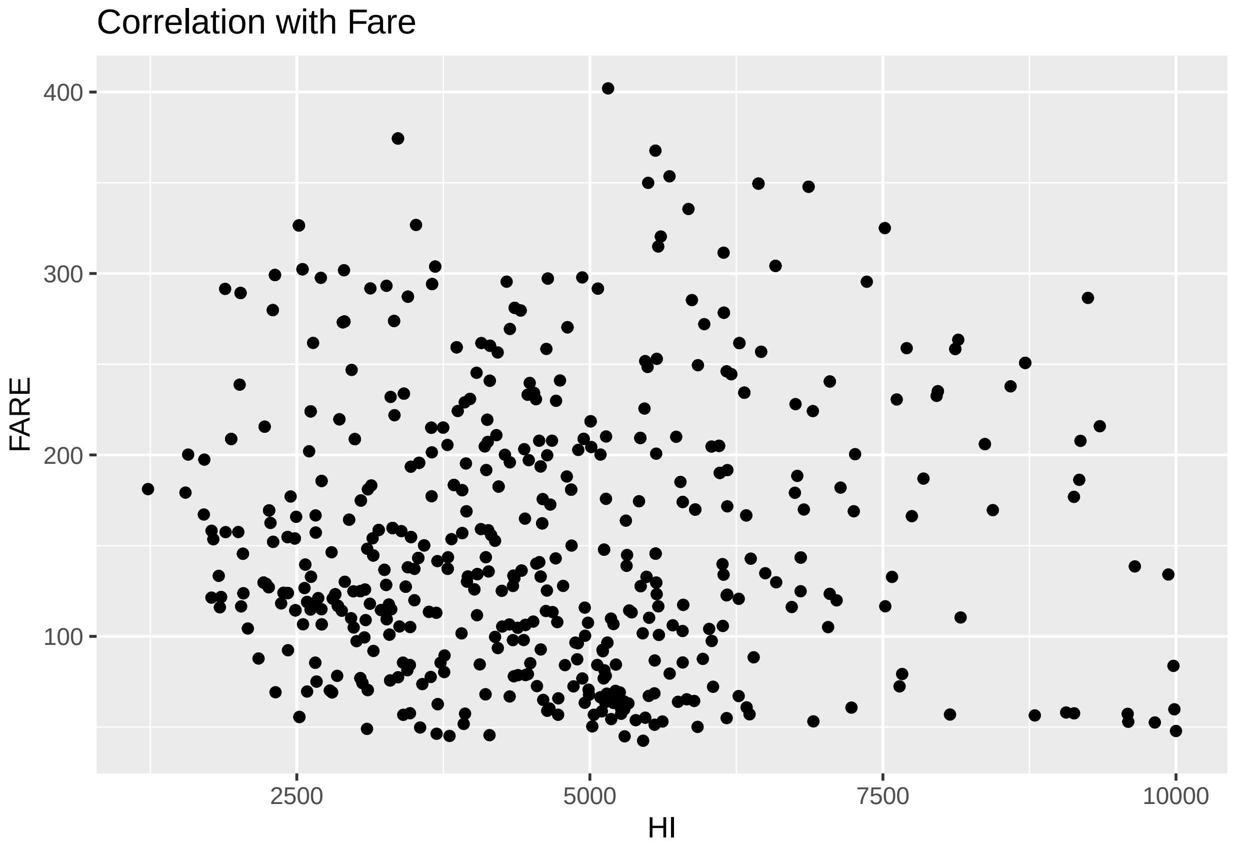
```
names(air.dt)
```

```
##  [1] "COUPON"   "NEW"      "VACATION" "SW"       "HI"       "S_INCOME"
##  [7] "E_INCOME" "S_POP"    "E_POP"    "SLOT"     "GATE"     "DISTANCE"
## [13] "PAX"      "FARE"
```

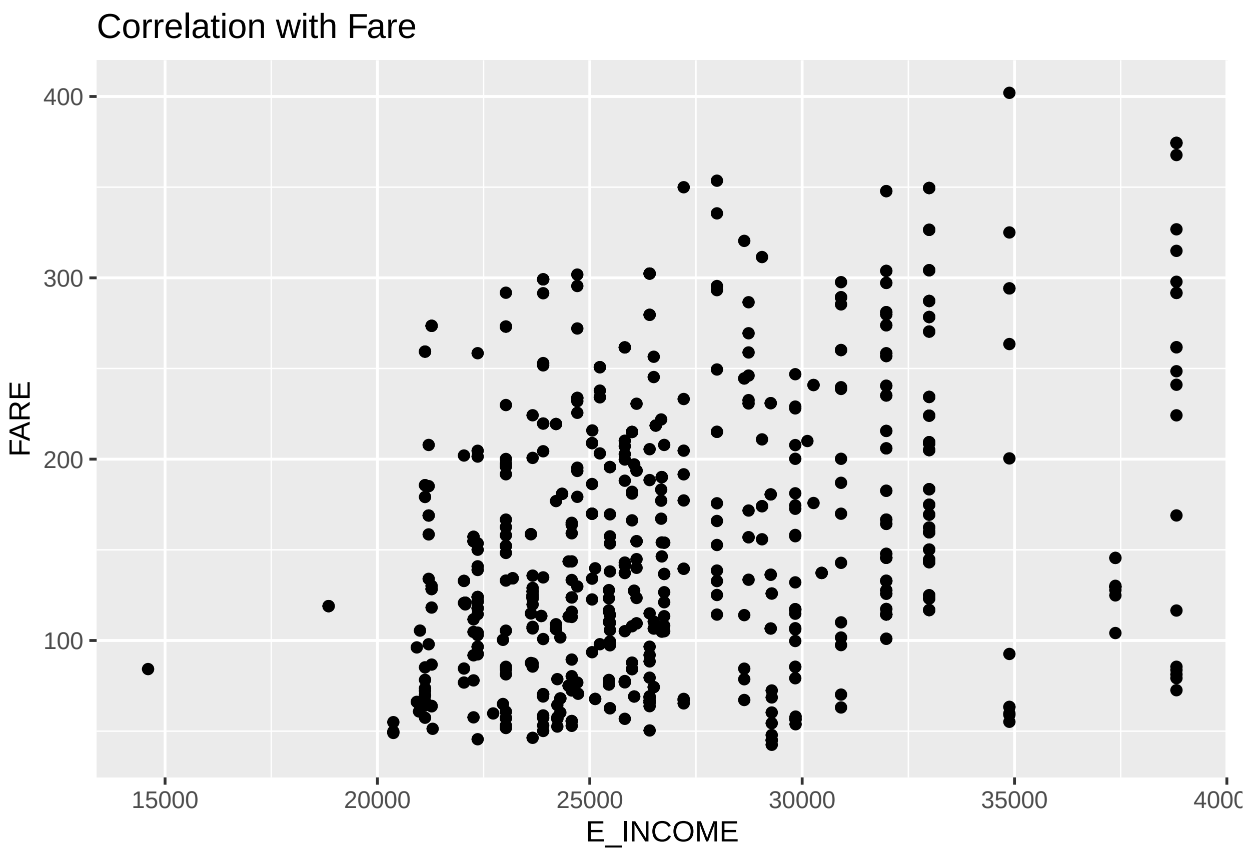
```
View(air.dt)

ggplot(air.dt, aes(x = COUPON, y = FARE)) + geom_point() +
  ggtitle("Correlation with Fare")
```



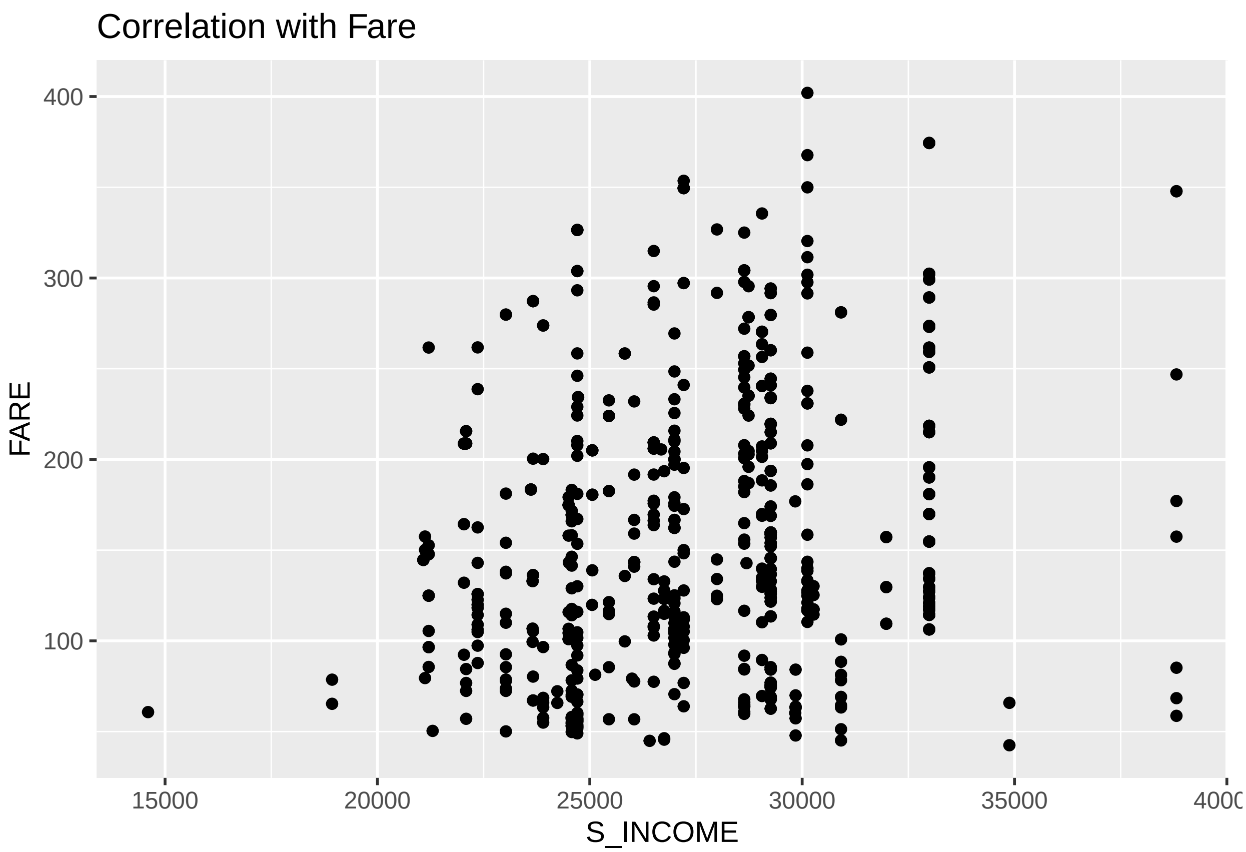
```
ggplot(air.dt, aes(x = HI, y = FARE)) + geom_point() +
  ggtitle("Correlation with Fare")
```



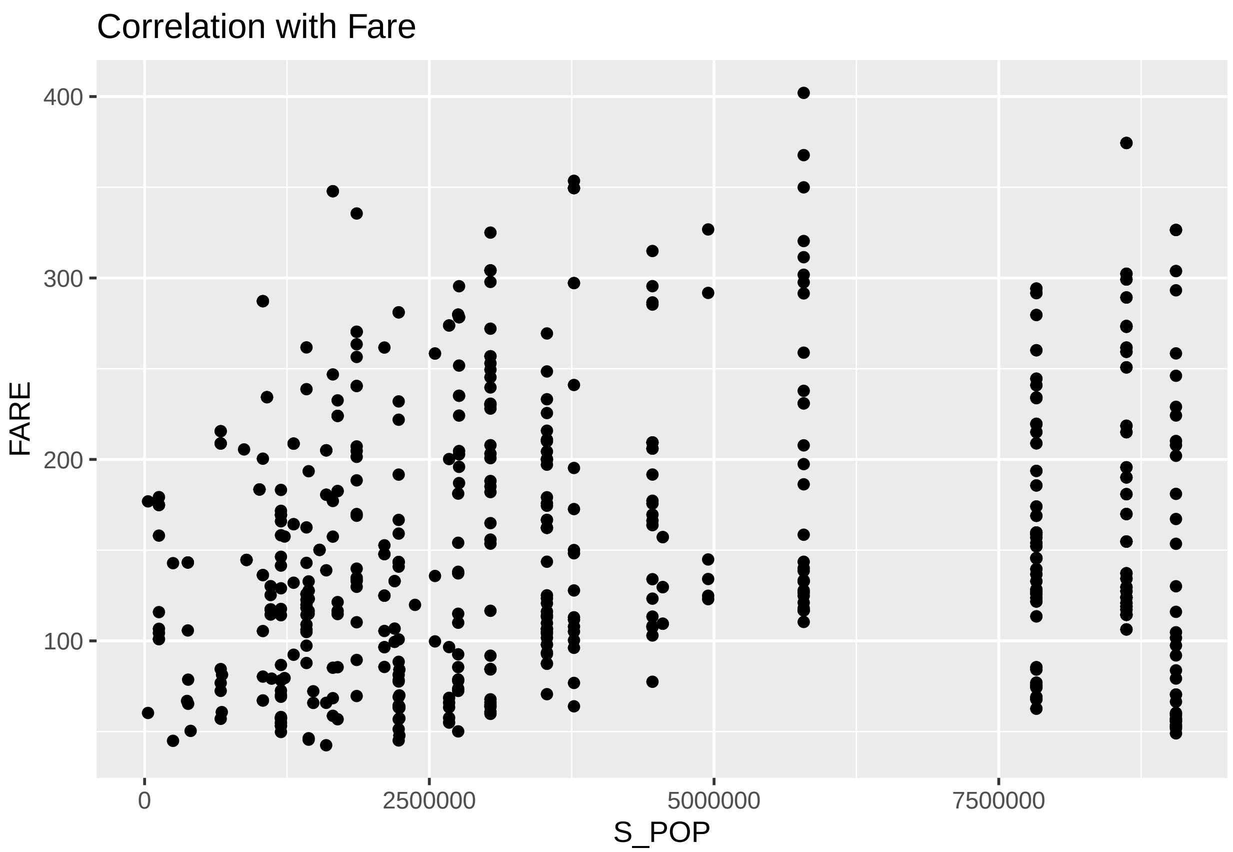
```r
ggplot(air.dt, aes(x = E_INCOME, y = FARE)) + geom_point() +
  ggtitle("Correlation with Fare")
```



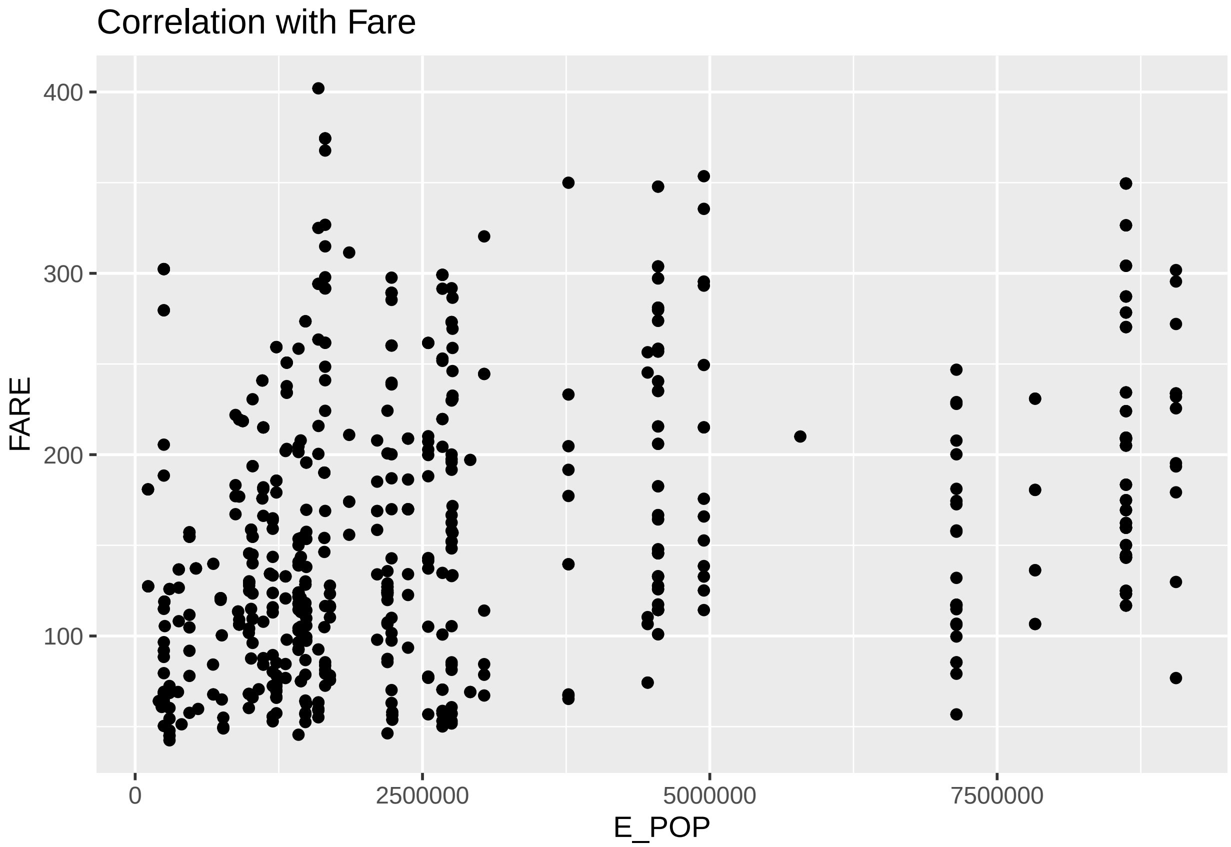
```r
ggplot(air.dt, aes(x = S_INCOME, y = FARE)) + geom_point() +
  ggtitle("Correlation with Fare")
```



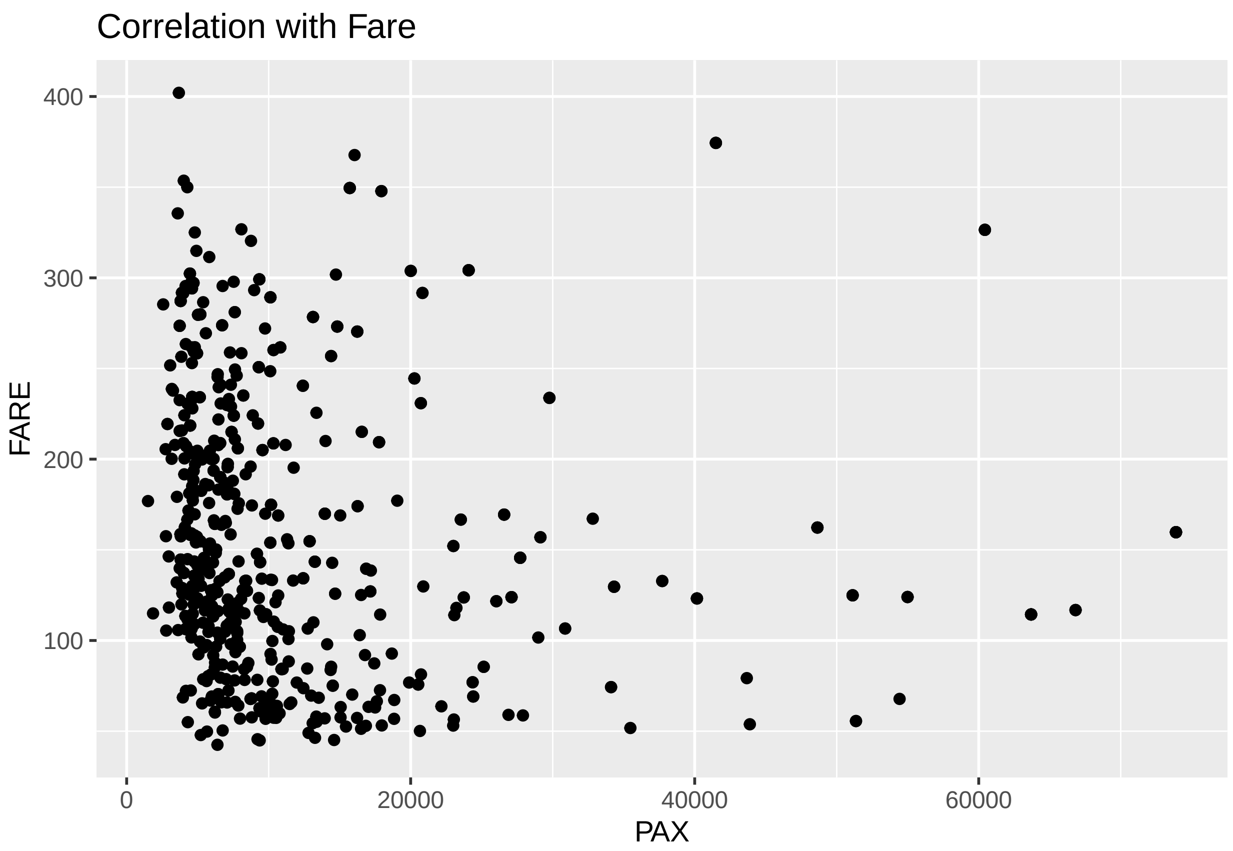
```
ggplot(air.dt, aes(x = S_POP, y = FARE)) + geom_point() +
  ggtitle("Correlation with Fare")
```



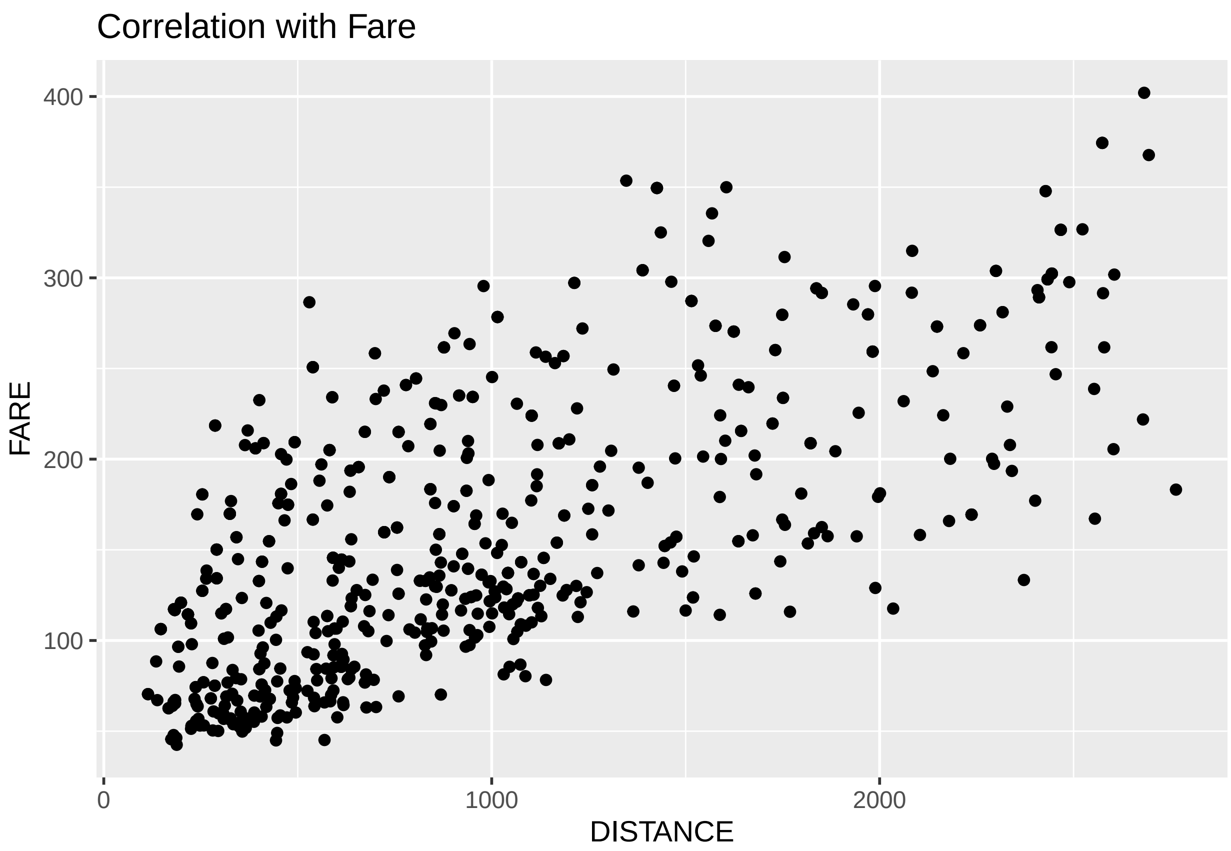
```
ggplot(air.dt, aes(x = E_POP, y = FARE)) + geom_point() +
  ggtitle("Correlation with Fare")
```



```r
ggplot(air.dt, aes(x = PAX, y = FARE)) + geom_point() +
  ggtitle("Correlation with Fare")
```



```r
ggplot(air.dt, aes(x = DISTANCE, y = FARE)) + geom_point() +
  ggtitle("Correlation with Fare")
```



```
air.cor.dt <- round(cor(air.dt[,!c(3,4,10,11)]),2)
air.cor.dt
```

```
##            COUPON   NEW    HI S_INCOME E_INCOME S_POP E_POP DISTANCE   PAX
## COUPON       1.00  0.02 -0.35    -0.09     0.05 -0.11  0.09     0.75 -0.34
## NEW          0.02  1.00  0.05     0.03     0.11 -0.02  0.06     0.08  0.01
## HI          -0.35  0.05  1.00    -0.03     0.08 -0.17 -0.06    -0.31 -0.17
## S_INCOME    -0.09  0.03 -0.03     1.00    -0.14  0.52 -0.27     0.03  0.14
## E_INCOME     0.05  0.11  0.08    -0.14     1.00 -0.14  0.46     0.18  0.26
## S_POP       -0.11 -0.02 -0.17     0.52    -0.14  1.00 -0.28     0.02  0.28
## E_POP        0.09  0.06 -0.06    -0.27     0.46 -0.28  1.00     0.12  0.31
## DISTANCE     0.75  0.08 -0.31     0.03     0.18  0.02  0.12     1.00 -0.10
## PAX         -0.34  0.01 -0.17     0.14     0.26  0.28  0.31    -0.10  1.00
## FARE         0.50  0.09  0.03     0.21     0.33  0.15  0.29     0.67 -0.09
##             FARE
## COUPON      0.50
## NEW         0.09
## HI          0.03
## S_INCOME    0.21
## E_INCOME    0.33
## S_POP       0.15
## E_POP       0.29
## DISTANCE    0.67
## PAX        -0.09
## FARE        1.00
```

```
melt.aircor.dt <- melt(air.cor.dt)
```

ANS – ASSUMPTIONS Correlation coefficient scale:- 0 to 0.29 - Weak 0.30 to 0.60 - Moderate 0.61 to 1.0 - Strong The '+' and '-' signs indicate positive and negative relations respectively.

– CONCLUSION We can CONCLUDE that DISTANCE is the BEST SINGLE PREDICTOR of fare, as it has a linear relationship with the fare and as the distance increases, the fare of flight also increases.

2. Explore the categorical predictors by computing the percentage of flights in each category. Create a pivot table with the average fare in each category. Which categorical predictor seems best for predicting FARE? Explain your answer.

```
#2.
Vacation1 <- table(air.dt$VACATION)

vacation <- round((100*prop.table(Vacation1)),digits = 2)
vacation
```

```
##
##    No   Yes
## 73.35 26.65
```

```
Sw1 <- table(air.dt$SW)
SW <- (round(100*prop.table(Sw1),digits=2))
SW
```

```
##
##    No   Yes
## 69.59 30.41
```

```
slot <- table(air.dt$SLOT)
slot_Percentage <- round(100*prop.table(slot),digits=0)
slot_Percentage
```

```
##
## Controlled       Free
##         29         71
```

```
Gate <- table(air.dt$GATE)
Gate_Percentage <- round(100*prop.table(Gate),digits = 2)
Gate_Percentage
```

```
##
## Constrained        Free
##       19.44       80.56
```

```
Total.Percetage <- list(vacation,SW,slot_Percentage,Gate_Percentage)
Total.Percetage
```

```
## [[1]]
##
##    No   Yes
## 73.35 26.65
##
## [[2]]
##
##    No   Yes
## 69.59 30.41
##
## [[3]]
##
## Controlled       Free
##         29         71
##
## [[4]]
##
## Constrained        Free
##       19.44       80.56
```

```
mlt <- melt(air.dt, id=c("VACATION", "SW", "SLOT", "GATE"), measure=c("FARE"))

cast(mlt, VACATION~.,subset=variable=="FARE", margins = TRUE ,mean)
```

```
##   VACATION    (all)
## 1       No 173.5525
## 2      Yes 125.9809
## 3    (all) 160.8767
```

```
cast(mlt, SW~.,subset=variable=="FARE" , margins = TRUE, mean)
```

```
##      SW     (all)
## 1    No 188.18279
## 2   Yes  98.38227
## 3 (all) 160.87668
```

```
cast(mlt, SLOT~.,subset=variable=="FARE" , margins = TRUE, mean)
```

```
##         SLOT    (all)
## 1 Controlled 186.0594
## 2       Free 150.8257
## 3      (all) 160.8767
```

```
cast(mlt, GATE~.,subset=variable=="FARE" , margins = TRUE, mean)
```

```
##          GATE    (all)
## 1 Constrained 193.1290
## 2        Free 153.0960
## 3       (all) 160.8767
```

ANS We infer that SW is the best categorical predictor of FARE, as there is a significant difference between average fare of the route on which South West airlines operates and the average fare on route where South West Airlines does not operate. The average fare of the route on which South West operates is nearly half the average fare of the route on which it does not.

3. Create data partition by assigning 80% of the records to the training dataset. Use rounding if 80% of the index generates a fraction. Also, set the seed at 42.

ANS

```
#3.
set.seed(42)
splitair <- round(nrow(air.dt) * 0.8)
train.air <- air.dt[1:splitair, ]
test.air <- air.dt[(splitair+1):nrow(air.dt), ]
dim(air.dt)
```

```
## [1] 638  14
```

```
dim(train.air)
```

```
## [1] 510  14
```

```
dim(test.air)
```

```
## [1] 128  14
```

```
train.air
```

```
##      COUPON NEW VACATION  SW      HI S_INCOME E_INCOME   S_POP   E_POP
##   1:   1.00   3       No Yes 5291.99    28637    21112 3036732  205711
##   2:   1.06   3       No  No 5419.16    26993    29838 3532657 7145897
##   3:   1.06   3       No  No 9185.28    30124    29838 5787293 7145897
##   4:   1.06   3       No Yes 2657.35    29260    29838 7830332 7145897
##   5:   1.06   3       No Yes 2657.35    29260    29838 7830332 7145897
##  ---
## 506:   1.02   3       No Yes 5201.65    30916    34880 2230831 1594251
## 507:   1.37   3      Yes  No 3787.29    32991    30460 8621121  528868
## 508:   1.37   1      Yes  No 3787.29    32991    30460 8621121  528868
## 509:   1.37   3      Yes  No 3787.29    32991    30460 8621121  528868
## 510:   1.02   3       No  No 6372.59    28690    30916  249642 2230831
##            SLOT        GATE DISTANCE   PAX   FARE
##   1:       Free        Free      312  7864  64.11
##   2:       Free        Free      576  8820 174.47
##   3:       Free        Free      364  6452 207.76
##   4: Controlled        Free      612 25144  85.47
##   5:       Free        Free      612 25144  85.47
##  ---
## 506:       Free        Free      702 15072  63.30
## 507: Controlled        Free     1042  4028 137.25
## 508: Controlled        Free     1042  4028 137.25
## 509:       Free Constrained     1042  4028 137.25
## 510:       Free        Free     1443 14474 142.83
```

```
test.air
```

```
##      COUPON NEW VACATION  SW      HI S_INCOME E_INCOME   S_POP   E_POP
##   1:   1.47   3       No  No 5090.58    26993    30916 3532657 2230831
##   2:   1.59   3       No  No 2705.03    30124    30916 5787293 2230831
##   3:   1.11   3       No Yes 6039.76    24706    30916 9056076 2230831
##   4:   1.25   1       No  No 4148.56    29260    30916 7830332 2230831
##   5:   1.25   3       No  No 4148.56    29260    30916 7830332 2230831
##  ---
## 124:   1.08   3      Yes  No 2216.70    32991    37375 8621121  991717
## 125:   1.08   0      Yes  No 2216.70    32991    37375 8621121  991717
## 126:   1.17   3      Yes  No 6797.80    27994    37375 4948339  991717
## 127:   1.28   3      Yes  No 5566.43    31981    37375 4549784  991717
## 128:   1.28   3      Yes  No 5566.43    31981    37375 4549784  991717
##            SLOT        GATE DISTANCE   PAX   FARE
##   1:       Free        Free     2182  6124 200.20
##   2:       Free        Free     2489  4560 297.61
##   3:       Free        Free      943  5638  97.46
##   4: Controlled        Free     1731 10343 260.16
##   5:       Free        Free     1731 10343 260.16
##  ---
## 124: Controlled        Free     1030 34324 129.63
## 125:       Free Constrained     1030 34324 129.63
## 126:       Free        Free      960  6016 124.87
## 127:       Free        Free      858  4877 129.62
## 128: Controlled        Free      858  4877 129.62
```

4. Using leaps package, run stepwise regression to reduce the number of predictors. Discuss the results from this model.

```
#4.
#Stepwise regression model
names(train.air)
```

```
##  [1] "COUPON"   "NEW"      "VACATION" "SW"       "HI"       "S_INCOME"
##  [7] "E_INCOME" "S_POP"    "E_POP"    "SLOT"     "GATE"     "DISTANCE"
## [13] "PAX"      "FARE"
```

```
options(scipen = 999)
air.reg <- lm(FARE ~ ., data = train.air)
air.stepwise_reg <- step(air.reg, direction = "both")
```

```
## Start:  AIC=3682.13
## FARE ~ COUPON + NEW + VACATION + SW + HI + S_INCOME + E_INCOME +
##     S_POP + E_POP + SLOT + GATE + DISTANCE + PAX
##
##            Df Sum of Sq    RSS    AIC
## - COUPON    1       231 659824 3680.3
## <none>                  659594 3682.1
## - NEW       1      5319 664913 3684.2
## - S_INCOME  1      7393 666986 3685.8
## - SLOT      1     22055 681648 3696.9
## - E_INCOME  1     27320  686914 3700.8
## - E_POP     1     28928  688522 3702.0
## - S_POP     1     31677  691271 3704.1
## - PAX       1     34804  694398 3706.4
## - GATE      1     36936  696530 3707.9
## - HI        1     76763  736356 3736.3
## - VACATION  1     81514  741107 3739.6
## - SW        1    118653  778247 3764.5
## - DISTANCE  1    435609 1095202 3938.7
##
## Step:  AIC=3680.31
## FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME + S_POP +
##     E_POP + SLOT + GATE + DISTANCE + PAX
##
##            Df Sum of Sq     RSS    AIC
## <none>                   659824 3680.3
## + COUPON    1       231  659594 3682.1
## - NEW       1      5496  665320 3682.5
## - S_INCOME  1      7213  667037 3683.9
## - SLOT      1     22641  682465 3695.5
## - E_INCOME  1     27095  686919 3698.8
## - E_POP     1     29677  689502 3700.7
## - S_POP     1     31552  691377 3702.1
## - GATE      1     37304  697128 3706.4
## - PAX       1     45270  705094 3712.2
## - HI        1     80147  739971 3736.8
## - VACATION  1     82289  742114 3738.3
## - SW        1    119505  779329 3763.2
## - DISTANCE  1    867774 1527599 4106.4
```

```
summary(air.stepwise_reg)
```

```
##
## Call:
## lm(formula = FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME +
##     S_POP + E_POP + SLOT + GATE + DISTANCE + PAX, data = train.air)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -112.462  -23.712   -0.434   21.870  132.369
##
## Coefficients:
##                   Estimate     Std. Error t value             Pr(>|t|)
## (Intercept)   1.4869956055  25.1322694498   0.059               0.9528
## NEW          -4.2298355653   2.0789764672  -2.035               0.0424 *
## VACATIONYes -33.9319056278   4.3099561107  -7.873   0.0000000000000219 ***
## SWYes       -39.8104470405   4.1960448999  -9.488 < 0.0000000000000002 ***
## HI            0.0085198816   0.0010965465   7.770   0.0000000000000454 ***
## S_INCOME      0.0014094582   0.0006047016   2.331               0.0202 *
## E_INCOME      0.0020167534   0.0004464213   4.518   0.0000078224624821 ***
## S_POP         0.0000036400   0.0000007467   4.875   0.0000014659282375 ***
## E_POP         0.0000038580   0.0000008160   4.728   0.0000029576605526 ***
## SLOTFree    -18.6655641718   4.5199233476  -4.130   0.0000426223903430 ***
## GATEFree    -23.9184649522   4.5122398053  -5.301   0.0000001737147895 ***
## DISTANCE      0.0761209176   0.0029773962  25.566 < 0.0000000000000002 ***
## PAX          -0.0008869095   0.0001518839  -5.839   0.0000000094720970 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.44 on 497 degrees of freedom
## Multiple R-squared:  0.779,  Adjusted R-squared:  0.7736
## F-statistic:   146 on 12 and 497 DF,  p-value: < 0.00000000000000022
```

```
# Which variables were dropped/added?
```

ANS If we consider confidence interval of 95%, the p-value of F-statistics is significanly less than 0.05, which indicates overall the model is good. On performing stepwise regression, model has dropped COUPON variable and all other independent variables are significant in predicting the dependent variable, FARE. The adjusted r-square value is 77.36% which indicates that 77.36% of variance of the dependent variable, FARE is explained by the change in predictors.

5. Repeat the process in (4) using exhaustive search instead of stepwise regression. Compare the resulting best model to the one you obtained in (4) in terms of the predictors included in the final model.

```
#5.
#exhaustive search
set.seed(42)

air.exhaustive_reg <- regsubsets(FARE ~ ., data = train.air, nbest = 1,
                                 nvmax = dim(train.air)[2],
                     method = "backward")
sum <- summary(air.exhaustive_reg)
sum$which
```

```
##    (Intercept) COUPON   NEW VACATIONYes SWYes    HI S_INCOME E_INCOME
## 1         TRUE  FALSE FALSE       FALSE FALSE FALSE    FALSE    FALSE
## 2         TRUE  FALSE FALSE       FALSE  TRUE FALSE    FALSE    FALSE
## 3         TRUE  FALSE FALSE        TRUE  TRUE FALSE    FALSE    FALSE
## 4         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 5         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 6         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 7         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 8         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 9         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 10        TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE     TRUE
## 11        TRUE  FALSE FALSE        TRUE  TRUE  TRUE     TRUE     TRUE
## 12        TRUE  FALSE  TRUE        TRUE  TRUE  TRUE     TRUE     TRUE
## 13        TRUE   TRUE  TRUE        TRUE  TRUE  TRUE     TRUE     TRUE
##    S_POP E_POP SLOTFree GATEFree DISTANCE   PAX
## 1  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 2  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 3  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 4  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 5  FALSE FALSE     TRUE    FALSE     TRUE FALSE
## 6  FALSE FALSE     TRUE     TRUE     TRUE FALSE
## 7  FALSE  TRUE     TRUE     TRUE     TRUE FALSE
## 8   TRUE  TRUE     TRUE     TRUE     TRUE FALSE
## 9   TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 10  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 11  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 12  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 13  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
```

```
sum$rsq
```

```
##  [1] 0.4226632 0.5886108 0.6864977 0.7108464 0.7267688 0.7494916 0.7520226
##  [8] 0.7571676 0.7676974 0.7749202 0.7771180 0.7789590 0.7790363
```

```
sum$adjr2
```

```
##  [1] 0.4215267 0.5869880 0.6846390 0.7085560 0.7240582 0.7465034 0.7485648
##  [8] 0.7532901 0.7635160 0.7704096 0.7721949 0.7736220 0.7732449
```

```
sum$cp
```

```
##  [1] 789.95487 419.45025 201.72244 149.06669 115.32529  66.31930  62.63778
##  [8]  53.08878  31.45236  17.23936  14.30602  12.17341  14.00000
```

ANS We are using adjusted r-square for selection criteria in exhaustive search. Adjusted r-square is indicating that we should drop the last variable(COUPON) and consider all other independent variables. From this we can conclude that both stepwise and exhaustive search are indicating to drop the same variable, COUPON while using other variables for predicting the dependent variable, FARE.

6. Compare the predictive accuracy of both models—stepwise regression and exhaustive search—using measures such as RMSE.

```
#6.
#accuracy

names(test.air)
```

```
##  [1] "COUPON"   "NEW"      "VACATION" "SW"       "HI"       "S_INCOME"
##  [7] "E_INCOME" "S_POP"    "E_POP"    "SLOT"     "GATE"     "DISTANCE"
## [13] "PAX"      "FARE"
```

```
#stepwise
air.stepwise_accuracy <- predict(air.stepwise_reg, test.air)

accuracy(air.stepwise_accuracy, test.air$FARE)
```

```
##                 ME     RMSE      MAE       MPE     MAPE
## Test set 0.9419492 32.28078 25.13929 -1.118252 18.07573
```

```
#exhuustive
names(test.air)
```

```
##  [1] "COUPON"   "NEW"      "VACATION" "SW"       "HI"       "S_INCOME"
##  [7] "E_INCOME" "S_POP"    "E_POP"    "SLOT"     "GATE"     "DISTANCE"
## [13] "PAX"      "FARE"
```

```
air.exh.reg <- lm(FARE~.-COUPON, data = train.air )
summary(air.exh.reg)
```

```
##
## Call:
## lm(formula = FARE ~ . - COUPON, data = train.air)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -112.462  -23.712   -0.434   21.870  132.369
##
## Coefficients:
##                    Estimate     Std. Error t value             Pr(>|t|)
## (Intercept)   1.4869956055  25.1322694498   0.059               0.9528
## NEW          -4.2298355653   2.0789764672  -2.035               0.0424 *
## VACATIONYes -33.9319056278   4.3099561107  -7.873   0.0000000000000219 ***
## SWYes       -39.8104470405   4.1960448999  -9.488 < 0.0000000000000002 ***
## HI            0.0085198816   0.0010965465   7.770   0.0000000000000454 ***
## S_INCOME      0.0014094582   0.0006047016   2.331               0.0202 *
## E_INCOME      0.0020167534   0.0004464213   4.518   0.0000078224624821 ***
## S_POP         0.0000036400   0.0000007467   4.875   0.0000014659282375 ***
## E_POP         0.0000038580   0.0000008160   4.728   0.0000029576605526 ***
## SLOTFree    -18.6655641718   4.5199233476  -4.130   0.0000426223903430 ***
## GATEFree    -23.9184649522   4.5122398053  -5.301   0.0000001737147895 ***
## DISTANCE      0.0761209176   0.0029773962  25.566 < 0.0000000000000002 ***
## PAX          -0.0008869095   0.0001518839  -5.839   0.0000000094720970 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.44 on 497 degrees of freedom
## Multiple R-squared:  0.779,  Adjusted R-squared:  0.7736
## F-statistic:   146 on 12 and 497 DF,  p-value: < 0.00000000000000022
```

```
air.exh_accuracy <- predict(air.exh.reg, test.air)

accuracy(air.exh_accuracy, test.air$FARE)
```

```
##                 ME     RMSE      MAE       MPE     MAPE
## Test set 0.9419492 32.28078 25.13929 -1.118252 18.07573
```

ANS On comparing ME, RMSE, MAE, MPE, MAPE results for the two models, we can say that they are producing similar results, meaning both the models are equally efficient in predicting FARE.

7. Using the exhaustive search model, predict the average fare on a route with the following characteristics: COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S_INCOME = $28,760, E_INCOME = $27,664, S_POP = 4,557,004, E_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 miles.

```
#7.
predict.air <- predict(air.exh.reg,data.table(COUPON = 1.202, NEW = 3, VACATION = "No",  SW = "No", HI
                                              E_INCOME =27664, S_POP = 4557004,
                                              E_POP = 3195503, SLOT = "Free",
                                              GATE = "Free",
                                              PAX = 12782, DISTANCE = 1976))

predict.air
```

```
##        1
## 248.3817
```

ANS Average fare where SouthWest is not providing services is $248.38.

8. Predict the reduction in average fare on the route in question (7.), if Southwest decides to cover this route [using the exhaustive search model above].

```
#8.
#southwest=YES
predict.air.SW <- predict(air.exh.reg,data.table(COUPON = 1.202, NEW = 3, VACATION = "No",
                                                 SW = "Yes", HI = 4442.141, S_INCOME = 28760,
                                                 E_INCOME =27664, S_POP = 4557004,
                                                 E_POP = 3195503, SLOT = "Free",
                                                 GATE = "Free", PAX = 12782,
                                                 DISTANCE = 1976))

predict.air.SW
```

```
##        1
## 208.5713
```

```
Difference_Fare <- abs(predict.air.SW-predict.air)
Difference_Fare
```

```
##        1
## 39.81045
```

ANS The reduction in average fare on the route, if SouthWest decides to cover this route is $39.81.

9. Using leaps package, run backward selection regression to reduce the number of predictors. Discuss the results from this model.

```
#9.
#backward selection

air.back_reg <- step(air.reg, direction = "backward")
```

```
## Start:  AIC=3682.13
## FARE ~ COUPON + NEW + VACATION + SW + HI + S_INCOME + E_INCOME +
##     S_POP + E_POP + SLOT + GATE + DISTANCE + PAX
##
##             Df Sum of Sq     RSS    AIC
## - COUPON    1       231   659824 3680.3
## <none>                   659594 3682.1
## - NEW       1      5319  664913 3684.2
## - S_INCOME  1      7393  666986 3685.8
## - SLOT      1     22055  681648 3696.9
## - E_INCOME  1     27320  686914 3700.8
## - E_POP     1     28928  688522 3702.0
## - S_POP     1     31677  691271 3704.1
## - PAX       1     34804  694398 3706.4
## - GATE      1     36936  696530 3707.9
## - HI        1     76763  736356 3736.3
## - VACATION  1     81514  741107 3739.6
## - SW        1    118653  778247 3764.5
## - DISTANCE  1    435609 1095202 3938.7
##
## Step:  AIC=3680.31
## FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME + S_POP +
##     E_POP + SLOT + GATE + DISTANCE + PAX
##
##            Df Sum of Sq     RSS    AIC
## <none>                   659824 3680.3
## - NEW       1      5496  665320 3682.5
## - S_INCOME  1      7213  667037 3683.9
## - SLOT      1     22641  682465 3695.5
## - E_INCOME  1     27095  686919 3698.8
## - E_POP     1     29677  689502 3700.7
## - S_POP     1     31552  691377 3702.1
## - GATE      1     37304  697128 3706.4
## - PAX       1     45270  705094 3712.2
## - HI        1     80147  739971 3736.8
## - VACATION  1     82289  742114 3738.3
## - SW        1    119505  779329 3763.2
## - DISTANCE  1    867774 1527599 4106.4
```

```
summary(air.back_reg)
```

```
##
## Call:
## lm(formula = FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME +
##     S_POP + E_POP + SLOT + GATE + DISTANCE + PAX, data = train.air)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -112.462  -23.712   -0.434   21.870  132.369
##
## Coefficients:
##                   Estimate     Std. Error t value             Pr(>|t|)
## (Intercept)   1.4869956055  25.1322694498   0.059               0.9528
## NEW          -4.2298355653   2.0789764672  -2.035               0.0424 *
## VACATIONYes -33.9319056278   4.3099561107  -7.873   0.0000000000000219 ***
## SWYes       -39.8104470405   4.1960448999  -9.488 < 0.0000000000000002 ***
## HI            0.0085198816   0.0010965465   7.770   0.0000000000000454 ***
## S_INCOME      0.0014094582   0.0006047016   2.331               0.0202 *
## E_INCOME      0.0020167534   0.0004464213   4.518   0.0000078224624821 ***
## S_POP         0.0000036400   0.0000007467   4.875   0.0000014659282375 ***
## E_POP         0.0000038580   0.0000008160   4.728   0.0000029576605526 ***
## SLOTFree    -18.6655641718   4.5199233476  -4.130   0.0000426223903430 ***
## GATEFree    -23.9184649522   4.5122398053  -5.301   0.0000001737147895 ***
## DISTANCE      0.0761209176   0.0029773962  25.566 < 0.0000000000000002 ***
## PAX          -0.0008869095   0.0001518839  -5.839   0.0000000094720970 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.44 on 497 degrees of freedom
## Multiple R-squared:  0.779,  Adjusted R-squared:  0.7736
## F-statistic:   146 on 12 and 497 DF,  p-value: < 0.00000000000000022
```

ANS If we consider confidence interval of 95%, the p-value of F-statistics is significanly less than 0.05, which indicates overall the model is good. On performing backward selection regression, model has dropped COUPON variable and all other independent variables are significant in predicting the dependent variable, FARE. The adjusted r-square value is 77.36% which indicates that 77.36% of variance of the dependent variable, FARE is explained by the change in predictors.

10. Now run a backward selection model using stepAIC() function. Discuss the results from this model, including the role of AIC in this model.

```
#10.
#step AIC()

air.stepAIC_reg <- stepAIC(air.reg,
        direction = "backward",
        trace = 1, keep = NULL, steps = 1000, use.start = FALSE,
        k = 2)
```

```
## Start:  AIC=3682.13
## FARE ~ COUPON + NEW + VACATION + SW + HI + S_INCOME + E_INCOME +
##     S_POP + E_POP + SLOT + GATE + DISTANCE + PAX
##
##             Df Sum of Sq     RSS    AIC
## - COUPON     1       231  659824 3680.3
## <none>                    659594 3682.1
## - NEW        1      5319  664913 3684.2
## - S_INCOME   1      7393  666986 3685.8
## - SLOT       1     22055  681648 3696.9
## - E_INCOME   1     27320  686914 3700.8
## - E_POP      1     28928  688522 3702.0
## - S_POP      1     31677  691271 3704.1
## - PAX        1     34804  694398 3706.4
## - GATE       1     36936  696530 3707.9
## - HI         1     76763  736356 3736.3
## - VACATION   1     81514  741107 3739.6
## - SW         1    118653  778247 3764.5
## - DISTANCE   1    435609 1095202 3938.7
##
## Step:  AIC=3680.31
## FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME + S_POP +
##     E_POP + SLOT + GATE + DISTANCE + PAX
##
```

```
##            Df Sum of Sq     RSS    AIC
## <none>                  659824 3680.3
## - NEW       1      5496  665320 3682.5
## - S_INCOME  1      7213  667037 3683.9
## - SLOT      1     22641  682465 3695.5
## - E_INCOME  1     27095  686919 3698.8
## - E_POP     1     29677  689502 3700.7
## - S_POP     1     31552  691377 3702.1
## - GATE      1     37304  697128 3706.4
## - PAX       1     45270  705094 3712.2
## - HI        1     80147  739971 3736.8
## - VACATION  1     82289  742114 3738.3
## - SW        1    119505  779329 3763.2
## - DISTANCE  1    867774 1527599 4106.4
```

```
summary(air.stepAIC_reg)
```

```
## 
## Call:
## lm(formula = FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME + 
##     S_POP + E_POP + SLOT + GATE + DISTANCE + PAX, data = train.air)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -112.462  -23.712   -0.434   21.870  132.369 
## 
## Coefficients:
##                  Estimate     Std. Error t value             Pr(>|t|)    
## (Intercept)   1.4869956055  25.1322694498   0.059               0.9528    
## NEW          -4.2298355653   2.0789764672  -2.035               0.0424 *  
## VACATIONYes -33.9319056278   4.3099561107  -7.873   0.0000000000000219 ***
## SWYes       -39.8104470405   4.1960448999  -9.488 < 0.0000000000000002 ***
## HI            0.0085198816   0.0010965465   7.770   0.0000000000000454 ***
## S_INCOME      0.0014094582   0.0006047016   2.331               0.0202 *  
## E_INCOME      0.0020167534   0.0004464213   4.518   0.0000078224624821 ***
## S_POP         0.0000036400   0.0000007467   4.875   0.0000014659282375 ***
## E_POP         0.0000038580   0.0000008160   4.728   0.0000029576605526 ***
## SLOTFree    -18.6655641718   4.5199233476  -4.130   0.0000426223903430 ***
## GATEFree    -23.9184649522   4.5122398053  -5.301   0.0000001737147895 ***
## DISTANCE      0.0761209176   0.0029773962  25.566 < 0.0000000000000002 ***
## PAX          -0.0008869095   0.0001518839  -5.839   0.0000000094720970 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 36.44 on 497 degrees of freedom
## Multiple R-squared:  0.779,  Adjusted R-squared:  0.7736 
## F-statistic:   146 on 12 and 497 DF,  p-value: < 0.00000000000000022
```

ANS If we consider confidence interval of 95%, the p-value of F-statistics is significanly less than 0.05, which indicates overall the model is good. On performing backward selection regression using stepAIC, model has dropped COUPON variable and all other independent variables are significant in predicting the dependent variable, FARE.

The adjusted r-square value is 77.36% which indicates that 77.36% of variance of the dependent variable, FARE is explained by the change in predictors.