Homework 2

Group 5 - Phylisha Martinez, Carolina Munoz, Vikrant Nakod, Hareesh Rajendran, Piyusha Kulkarni

2/21/2020

1. Load required packages

```
## Installing package into 'C:/Users/Home Laptop/Documents/R/win-library/3.5'
## (as 'lib' is unspecified)

## Warning: package 'glmnet' is not available (for R version 3.5.3)

## Warning: unable to access index for repository http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contril
##   cannot open URL 'http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5/PACKAGES'

## Warning: 'BiocManager' not available.  Could not check Bioconductor.
##
## Please use `install.packages('BiocManager')` and then retry.

## Warning in p_install(package, character.only = TRUE, ...):

## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'glmnet'

## Installing package into 'C:/Users/Home Laptop/Documents/R/win-library/3.5'
## (as 'lib' is unspecified)

## Warning: unable to access index for repository http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contril
##   cannot open URL 'http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5/PACKAGES'

## package 'goeveg' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Home Laptop\AppData\Local\Temp\Rtmp2hLVVO\downloaded_packages

##
## goeveg installed

## Warning in pacman::p_load(caret, corrplot, glmnet, mlbench, tidyverse, ggplot2, : Failed to install/l
## glmnet, goeveg

##
## Attaching package: 'MASS'
```

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

```
##  [1] ".GlobalEnv"         "package:MASS"       "package:forecast"
##  [4] "package:data.table" "package:leaps"      "package:gridExtra"
##  [7] "package:reshape"    "package:forcats"    "package:stringr"
## [10] "package:dplyr"      "package:purrr"      "package:readr"
## [13] "package:tidyr"      "package:tibble"     "package:tidyverse"
## [16] "package:mlbench"    "package:corrplot"   "package:caret"
## [19] "package:ggplot2"    "package:lattice"    "package:pacman"
## [22] "package:stats"      "package:graphics"   "package:grDevices"
## [25] "package:utils"      "package:datasets"   "package:methods"
## [28] "Autoloads"          "package:base"
```

2. Read the file 'Airfares.csv'

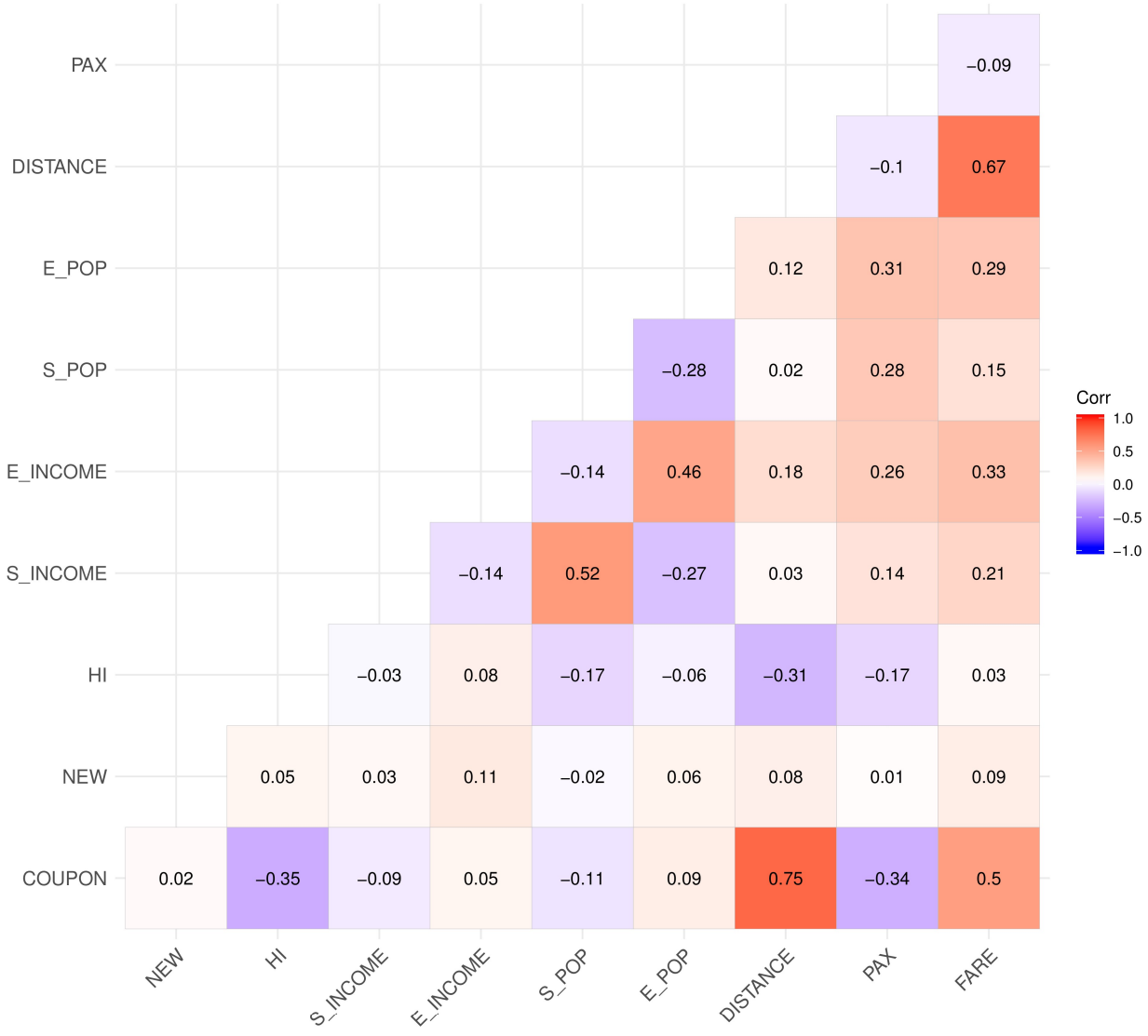
```
## 'data.frame':    638 obs. of  18 variables:
##  $ S_CODE  : Factor w/ 8 levels "*","DCA","EWR",..: 1 1 1 8 7 1 1 1 1 1 ...
##  $ S_CITY  : Factor w/ 51 levels "Albuquerque         NM",..: 14 3 7 9 9 11 14 18 23 25 ...
##  $ E_CODE  : Factor w/ 8 levels "*","DCA","EWR",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ E_CITY  : Factor w/ 68 levels "Amarillo            TX",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ 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 ...
```

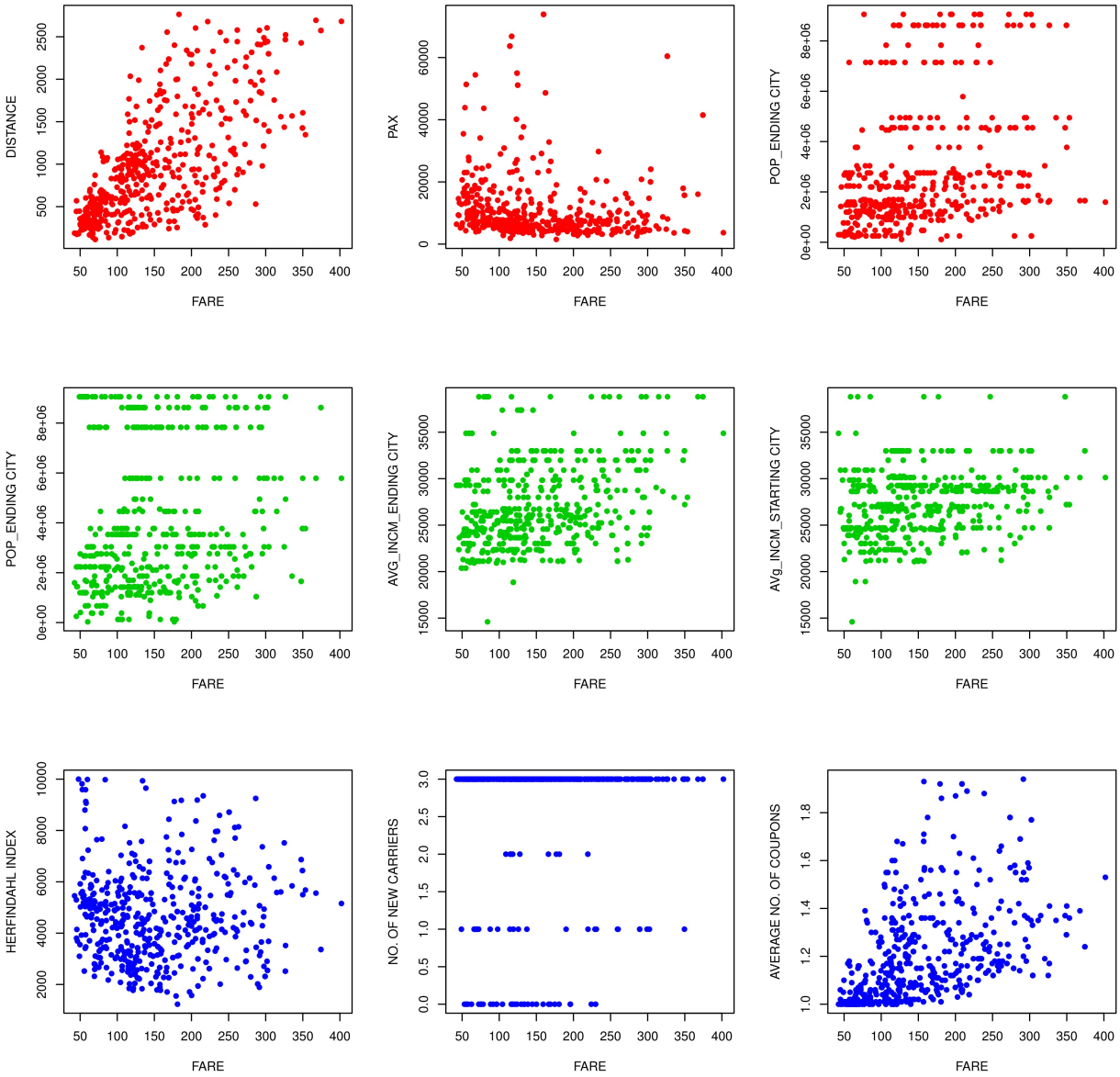
Question 1

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

```
##    COUPON NEW VACATION  SW      HI S_INCOME E_INCOME   S_POP   E_POP       SLOT
## 1:   1.00   3       No Yes 5291.99    28637    21112 3036732  205711       Free
## 2:   1.06   3       No  No 5419.16    26993    29838 3532657 7145897       Free
## 3:   1.06   3       No  No 9185.28    30124    29838 5787293 7145897       Free
## 4:   1.06   3       No Yes 2657.35    29260    29838 7830332 7145897 Controlled
## 5:   1.06   3       No Yes 2657.35    29260    29838 7830332 7145897       Free
```

```
## 6:   1.01   3       No Yes 3408.11    26046    29838 2230955 7145897       Free
##    GATE DISTANCE   PAX   FARE
## 1: Free      312  7864  64.11
## 2: Free      576  8820 174.47
## 3: Free      364  6452 207.76
## 4: Free      612 25144  85.47
## 5: Free      612 25144  85.47
## 6: Free      309 13386  56.76
```




DISTANCE
500 1000 1500 2000 2500
FARE
50 100 150 200 250 300 350 400
PAX
0 20000 40000 60000
POP_ENDING CITY
0e+00 2e+06 4e+06 6e+06 8e+06
AVG_INCM_ENDING CITY
15000 20000 25000 30000 35000
AVg_INCM_STARTING CITY
HERFINDAHL INDEX
2000 4000 6000 8000 10000
NO. OF NEW CARRIERS
0.0 0.5 1.0 1.5 2.0 2.5 3.0
AVERAGE NO. OF COUPONS
1.0 1.2 1.4 1.6 1.8

Explanation [1]

After analyzing the data you can see that distance is the best single predictor of Fare because the correlation between Fare and Distance have the highest positive correlation of .67, which is also apparent by looking at the correlation table/heat map.

Question 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

```
## [1] "Percentage of flights for VACATION"

##   Is Vacation Flight Freq Percentage of Flights
## 1                  No  468             73.35423
## 2                 Yes  170             26.64577

## [1] "Percentage of flights for Southwest Airlines"

##   Southwest Airlines Flight Freq Percentage of Flights
## 1                         No  444              69.59248
## 2                        Yes  194              30.40752

## [1] "Percentage of flights for variable SLOT"

##   Destination Airport SLOT Freq Percentage of Flights
## 1                Controlled  182             28.52665
## 2                      Free  456             71.47335

## [1] "Percentage of flights for variable GATE"

##   Destination Airport GATE Freq Percentage of Flights
## 1               Constrained  124             19.43574
## 2                      Free  514             80.56426

## [1] "Pivot Table with average fare in VACATION categories"

## # A tibble: 2 x 2
##   VACATION AVG_FARE
##   <fct>       <dbl>
## 1 No           174.
## 2 Yes          126.
```

```
## [1] "Pivot Table with average fare in SW categories"
```

```
## # A tibble: 2 x 2
##   SW    AVG_FARE
##   <fct>    <dbl>
## 1 No       188.
## 2 Yes       98.4
```

```
## [1] "Pivot Table with average fare in SLOT categories"
```

```
## # A tibble: 2 x 2
##   SLOT       AVG_FARE
##   <fct>         <dbl>
## 1 Controlled     186.
## 2 Free           151.
```

```
## [1] "Pivot Table with average fare in GATE categories"
```

```
## # A tibble: 2 x 2
##   GATE        AVG_FARE
##   <fct>          <dbl>
## 1 Constrained     193.
## 2 Free            153.
```

```
##    VACATION Average_Fare  SW Average_Fare        GATE Average_Fare       SLOT
## 1:       No     173.5525 Yes     98.38227        Free      153.096       Free
## 2:      Yes     125.9809  No    188.18279 Constrained      193.129 Controlled
##    Average_Fare
## 1:     150.8257
## 2:     186.0594
```

Explanation [2]

After observing the pivot table of average fare with respect to the categorical varaiables, you can see that the Southwest Airlines is the best predictor of FARE. We observe that the average FARE of SW is spread. Flights from Southwest have an average of 98.38 (SW=YES) and flights that are not Southwest have an average of 188.18 (SW=NO) which is much higher, thus SW affects the price of FARE the most.

Question 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.

```
## '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 ...

## [1] 0.799
```

Question 4

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

```
## Start:  AIC=3625.62
## 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      1416  591818 3624.8
## - NEW       1      2019  592422 3625.4
## <none>                   590402 3625.6
## - S_INCOME  1      5682  596085 3628.5
## - E_INCOME  1     19820  610222 3640.5
## - SLOT      1     22230  612633 3642.5
## - S_POP     1     31231  621634 3649.9
## - E_POP     1     31852  622255 3650.4
## - PAX       1     34726  625129 3652.8
## - GATE      1     38560  628962 3655.9
## - HI        1     78939  669342 3687.6
## - VACATION  1     95997  686399 3700.5
## - SW        1    101122  691524 3704.2
## - DISTANCE  1    411223 1001625 3893.2
##
## Step:  AIC=3624.84
## FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME + S_POP +
##     E_POP + SLOT + GATE + DISTANCE + PAX
##
##            Df Sum of Sq     RSS    AIC
## - NEW       1      1960  593779 3624.5
## <none>                   591818 3624.8
## + COUPON    1      1416  590402 3625.6
## - S_INCOME  1      5121  596939 3627.2
## - E_INCOME  1     19209  611027 3639.1
## - SLOT      1     23514  615333 3642.7
## - S_POP     1     30500  622319 3648.5
## - E_POP     1     32819  624637 3650.4
## - GATE      1     38435  630254 3654.9
## - PAX       1     46644  638462 3661.5
## - HI        1     79770  671588 3687.3
## - VACATION  1     98110  689928 3701.1
## - SW        1    105658  697476 3706.6
## - DISTANCE  1    866055 1457874 4082.6
##
## Step:  AIC=3624.52
## FARE ~ VACATION + SW + HI + S_INCOME + E_INCOME + S_POP + E_POP +
##     SLOT + GATE + DISTANCE + PAX
##
##            Df Sum of Sq     RSS    AIC
## <none>                   593779 3624.5
## + NEW       1      1960  591818 3624.8
## + COUPON    1      1357  592422 3625.4
## - S_INCOME  1      4959  598738 3626.8
## - E_INCOME  1     18671  612450 3638.3
## - SLOT      1     23081  616860 3642.0
## - S_POP     1     31054  624833 3648.5
## - E_POP     1     32781  626560 3649.9
## - GATE      1     38744  632523 3654.8
## - PAX       1     46662  640441 3661.1
## - HI        1     79069  672848 3686.3
## - VACATION  1     97524  691303 3700.1
## - SW        1    105525  699304 3705.9
## - DISTANCE  1    864096 1457875 4080.6
```

```
##
## Call:
## lm(formula = FARE ~ VACATION + SW + HI + S_INCOME + E_INCOME +
##     S_POP + E_POP + SLOT + GATE + DISTANCE + PAX, data = train.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.125  -21.857   -2.867   20.563  105.992
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.807e+00  2.328e+01   0.292   0.7701
## VACATIONYes -3.522e+01  3.894e+00  -9.044  < 2e-16 ***
## SWYes       -3.760e+01  3.997e+00  -9.408  < 2e-16 ***
## HI           8.586e-03  1.054e-03   8.143 3.11e-15 ***
## S_INCOME     1.135e-03  5.563e-04   2.039   0.0419 *
## E_INCOME     1.655e-03  4.183e-04   3.957 8.69e-05 ***
## S_POP        3.627e-06  7.107e-07   5.103 4.75e-07 ***
## E_POP        4.232e-06  8.071e-07   5.243 2.33e-07 ***
## SLOTFree    -1.817e+01  4.129e+00  -4.400 1.33e-05 ***
## GATEFree    -2.493e+01  4.374e+00  -5.700 2.05e-08 ***
## DISTANCE     7.527e-02  2.796e-03  26.921  < 2e-16 ***
## PAX         -9.133e-04  1.460e-04  -6.256 8.52e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.53 on 498 degrees of freedom
## Multiple R-squared:  0.7911, Adjusted R-squared:  0.7865
## F-statistic: 171.4 on 11 and 498 DF,  p-value: < 2.2e-16
```

Explantaion [4]

The results of the stepwise regression show that the best model includes 10 predictors, including VACATION + SW +HI + E_INCOME + S_POP + E_POP + SLOT + GATE + DISTANCE + PAX and excluding NEW and COUPON. The Adjusted R-Squared for the model is 0.7759 and the final AIC was 3649.2.

Question 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.

```
##    (Intercept) COUPON   NEW VACATIONYes SWYes    HI S_INCOME E_INCOME S_POP
## 1         TRUE  FALSE FALSE       FALSE FALSE FALSE    FALSE    FALSE FALSE
```

```
## 2         TRUE  FALSE FALSE       FALSE  TRUE FALSE    FALSE    FALSE FALSE
## 3         TRUE  FALSE FALSE        TRUE  TRUE FALSE    FALSE    FALSE FALSE
## 4         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE FALSE
## 5         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE FALSE
## 6         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE FALSE
## 7         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE     TRUE FALSE
## 8         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE     TRUE FALSE
## 9         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE  TRUE
## 10        TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE     TRUE  TRUE
## 11        TRUE  FALSE FALSE        TRUE  TRUE  TRUE     TRUE     TRUE  TRUE
## 12        TRUE  FALSE  TRUE        TRUE  TRUE  TRUE     TRUE     TRUE  TRUE
## 13        TRUE   TRUE  TRUE        TRUE  TRUE  TRUE     TRUE     TRUE  TRUE
##    E_POP SLOTFree GATEFree DISTANCE   PAX
## 1  FALSE    FALSE    FALSE     TRUE FALSE
## 2  FALSE    FALSE    FALSE     TRUE FALSE
## 3  FALSE    FALSE    FALSE     TRUE FALSE
## 4  FALSE    FALSE    FALSE     TRUE FALSE
## 5  FALSE     TRUE    FALSE     TRUE FALSE
## 6  FALSE     TRUE     TRUE     TRUE FALSE
## 7  FALSE     TRUE     TRUE     TRUE FALSE
## 8  FALSE     TRUE     TRUE     TRUE  TRUE
## 9   TRUE     TRUE     TRUE     TRUE  TRUE
## 10  TRUE     TRUE     TRUE     TRUE  TRUE
## 11  TRUE     TRUE     TRUE     TRUE  TRUE
## 12  TRUE     TRUE     TRUE     TRUE  TRUE
## 13  TRUE     TRUE     TRUE     TRUE  TRUE
```

```
##  [1] 0.4511816 0.5891551 0.6990152 0.7262587 0.7408179 0.7644326 0.7681420
##  [8] 0.7734357 0.7837369 0.7893556 0.7911004 0.7917901 0.7922883
```

```
##  [1] 0.4501012 0.5875344 0.6972307 0.7240905 0.7382467 0.7616226 0.7649089
##  [8] 0.7698179 0.7798442 0.7851343 0.7864861 0.7867629 0.7868442
```

```
##  [1] 804.53734 477.06700 216.72926 153.67376 120.90738  66.51745  59.65958
##  [8]  49.01864  26.42010  15.00312  12.83666  13.18967  14.00000
```

Explanation[5]

The results of exhaustive search model were evaluated using the adjusted r-squared metric, for which a higher value is preferred. The results of the exhaustive search show that the best model includes 12 predictors, two more than the model that resulted in question 4. Although the adjusted r-squared begins to stablize around the 9th predictor, the highest adjusted r-square belogs to the model with 12 predictors (0.7760708).

Question 6

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

```
##                ME     RMSE      MAE       MPE     MAPE
## Test set 3.604245 39.26901 30.55444 -2.210884 23.10138
```

```
##                ME     RMSE      MAE       MPE     MAPE
## Test set 3.378458 39.23965 30.61102 -2.537753 22.94292
```

Explanation[6]

The model with the best predictive accuracy is the exhaustive model. While both have similar RMSEs, the exhaustive model has a slightly lower RMSE with 36.41184 than stepwise RMSE which is 36.8617.

Question 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.

```
##       1
## 247.4958
```

Explanation[7]

The average fare with the given test values is $247.4958

Question 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].

```
##        1
## 207.5155
```

Explanation[8]

According to given variable values the exhaustive search model predicts a average fare of $207.1558.SW being the best categorical factor it affects the price and the fair drops from $247.4958 to $207.5155. We can conclude that there is a reduction in average fare when Southwest airlines covers the route as compared to average fare on the route which Southwest airlines doesn't operate on.

Question 9

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

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

```
## 1  FALSE    FALSE    FALSE     TRUE FALSE
## 2  FALSE    FALSE    FALSE     TRUE FALSE
## 3  FALSE    FALSE    FALSE     TRUE FALSE
## 4  FALSE    FALSE    FALSE     TRUE FALSE
## 5  FALSE     TRUE    FALSE     TRUE FALSE
## 6  FALSE     TRUE     TRUE     TRUE FALSE
## 7   TRUE     TRUE     TRUE     TRUE FALSE
## 8   TRUE     TRUE     TRUE     TRUE  TRUE
## 9   TRUE     TRUE     TRUE     TRUE  TRUE
## 10  TRUE     TRUE     TRUE     TRUE  TRUE
## 11  TRUE     TRUE     TRUE     TRUE  TRUE
## 12  TRUE     TRUE     TRUE     TRUE  TRUE
## 13  TRUE     TRUE     TRUE     TRUE  TRUE
```

```
##  [1] 0.4511816 0.5891551 0.6990152 0.7262587 0.7408179 0.7644326 0.7675685
##  [8] 0.7726433 0.7837369 0.7893556 0.7911004 0.7917901 0.7922883
```

```
##  [1] 0.4501012 0.5875344 0.6972307 0.7240905 0.7382467 0.7616226 0.7643274
##  [8] 0.7690128 0.7798442 0.7851343 0.7864861 0.7867629 0.7868442
```

```
##  [1] 804.53734 477.06700 216.72926 153.67376 120.90738  66.51745  61.02905
##  [8]  50.91084  26.42010  15.00312  12.83666  13.18967  14.00000
```

Explanation[9]

From above results, we can interpret this backward search model by taking into consideration the Adjusted R-square. As seen from above Adjusted R-square values there is no significant increase in adjusted r-square after considering 11 variables. However, the highest ajusted r-square belongs to the model with 12 predictors. Therefore according to stepwise search the best variables for predicting FARE are VACATION, NEW, SW, HI, S_INCOME, E_INCOME, S_POP,E_POP, SLOT, GATE, DISTANCE, PAX. However, backward search model in not reccomended as computation cost goes higher with large number of variables.

Question 10

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

```
## Start:  AIC=3625.62
## 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      1416  591818 3624.8
## - NEW       1      2019  592422 3625.4
## <none>                   590402 3625.6
## - S_INCOME  1      5682  596085 3628.5
## - E_INCOME  1     19820  610222 3640.5
## - SLOT      1     22230  612633 3642.5
## - S_POP     1     31231  621634 3649.9
## - E_POP     1     31852  622255 3650.4
## - PAX       1     34726  625129 3652.8
## - GATE      1     38560  628962 3655.9
## - HI        1     78939  669342 3687.6
## - VACATION  1     95997  686399 3700.5
## - SW        1    101122  691524 3704.2
## - DISTANCE  1    411223 1001625 3893.2
##
## Step:  AIC=3624.84
## FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME + S_POP +
##     E_POP + SLOT + GATE + DISTANCE + PAX
##
##            Df Sum of Sq     RSS    AIC
## - NEW       1      1960  593779 3624.5
## <none>                   591818 3624.8
## - S_INCOME  1      5121  596939 3627.2
## - E_INCOME  1     19209  611027 3639.1
## - SLOT      1     23514  615333 3642.7
## - S_POP     1     30500  622319 3648.5
## - E_POP     1     32819  624637 3650.4
## - GATE      1     38435  630254 3654.9
## - PAX       1     46644  638462 3661.5
## - HI        1     79770  671588 3687.3
## - VACATION  1     98110  689928 3701.1
## - SW        1    105658  697476 3706.6
## - DISTANCE  1    866055 1457874 4082.6
##
## Step:  AIC=3624.52
## FARE ~ VACATION + SW + HI + S_INCOME + E_INCOME + S_POP + E_POP +
##     SLOT + GATE + DISTANCE + PAX
##
##            Df Sum of Sq     RSS    AIC
## <none>                   593779 3624.5
## - S_INCOME  1      4959  598738 3626.8
## - E_INCOME  1     18671  612450 3638.3
## - SLOT      1     23081  616860 3642.0
## - S_POP     1     31054  624833 3648.5
## - E_POP     1     32781  626560 3649.9
## - GATE      1     38744  632523 3654.8
## - PAX       1     46662  640441 3661.1
## - HI        1     79069  672848 3686.3
## - VACATION  1     97524  691303 3700.1
## - SW        1    105525  699304 3705.9
## - DISTANCE  1    864096 1457875 4080.6
```

```
##
## Call:
## lm(formula = FARE ~ VACATION + SW + HI + S_INCOME + E_INCOME +
##     S_POP + E_POP + SLOT + GATE + DISTANCE + PAX, data = train.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.125  -21.857   -2.867   20.563  105.992
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.807e+00  2.328e+01   0.292   0.7701
## VACATIONYes -3.522e+01  3.894e+00  -9.044  < 2e-16 ***
## SWYes       -3.760e+01  3.997e+00  -9.408  < 2e-16 ***
## HI           8.586e-03  1.054e-03   8.143 3.11e-15 ***
## S_INCOME     1.135e-03  5.563e-04   2.039   0.0419 *
## E_INCOME     1.655e-03  4.183e-04   3.957 8.69e-05 ***
## S_POP        3.627e-06  7.107e-07   5.103 4.75e-07 ***
## E_POP        4.232e-06  8.071e-07   5.243 2.33e-07 ***
## SLOTFree    -1.817e+01  4.129e+00  -4.400 1.33e-05 ***
## GATEFree    -2.493e+01  4.374e+00  -5.700 2.05e-08 ***
## DISTANCE     7.527e-02  2.796e-03  26.921  < 2e-16 ***
## PAX         -9.133e-04  1.460e-04  -6.256 8.52e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.53 on 498 degrees of freedom
## Multiple R-squared:  0.7911, Adjusted R-squared:  0.7865
## F-statistic: 171.4 on 11 and 498 DF,  p-value: < 2.2e-16

##                      ME     RMSE      MAE      MPE     MAPE
## Test set 1.847608e-13 34.12143 26.84369 -4.57178 20.21843
```

Explanation [10]

Using stepAIC resulted in a model with 10 predictors including VACATION, SW, HI, E_INCOME, S_POP, E_POP, SLOT, GATE, DISTANCE, and PAX and whith a final AIC of 3649.22. AIC quantifies how much information is lost due to simplification and penalizes the model for including too many predictors. Thus, the preferable model will be the one with the lowest AIC. Because it is using backwards selection, in the first runtrhough the model included all 13 predictors. Nonetheless, by the fourth run, the model had already taken out 3 predictors and achieved the lowest AIC with 10 predictors. It is possible stepAIC stopped at 10 predictors because a model with less variables would have a higher AIC, meaning too much information would be lost due to simplification.