Prediction of fare and changes in airfare with Southwest Airline's entry in a new route.

Sarthak Mohapatra

Loading all the required packages that will be used in the code. In case if the package is not installed, pacman will install it and then load it.

```
pacman::p_load(data.table, forecast, leaps, tidyverse, caret, corrplot, glmnet, mlbench, ggplot2,
               gplots, pivottabler, ggpubr, MASS, knitr, rmarkdown)
```

Reading the I/P file Airfares.CSV from the working directory and generating respective files that will be used in the code.

```
## [1] "Displaying the first 6 records of the I/P file."
```

```
##   S_CODE                S_CITY E_CODE                E_CITY COUPON NEW
## 1      * Dallas/Fort Worth   TX      * Amarillo            TX   1.00   3
## 2      * Atlanta             GA      * Baltimore/Wash Intl MD   1.06   3
## 3      * Boston              MA      * Baltimore/Wash Intl MD   1.06   3
## 4    ORD Chicago             IL      * Baltimore/Wash Intl MD   1.06   3
## 5    MDW Chicago             IL      * Baltimore/Wash Intl MD   1.06   3
## 6      * Cleveland           OH      * Baltimore/Wash Intl MD   1.01   3
##   VACATION  SW   HI S_INCOME E_INCOME   S_POP   E_POP       SLOT GATE DISTANCE
## 1       No Yes 5292    28637    21112 3036732  205711       Free Free      312
## 2       No  No 5419    26993    29838 3532657 7145897       Free Free      576
## 3       No  No 9185    30124    29838 5787293 7145897       Free Free      364
## 4       No Yes 2657    29260    29838 7830332 7145897 Controlled Free      612
## 5       No Yes 2657    29260    29838 7830332 7145897       Free Free      612
## 6       No Yes 3408    26046    29838 2230955 7145897       Free Free      309
##     PAX   FARE
## 1  7864  64.11
## 2  8820 174.47
## 3  6452 207.76
## 4 25144  85.47
## 5 25144  85.47
## 6 13386  56.76
```

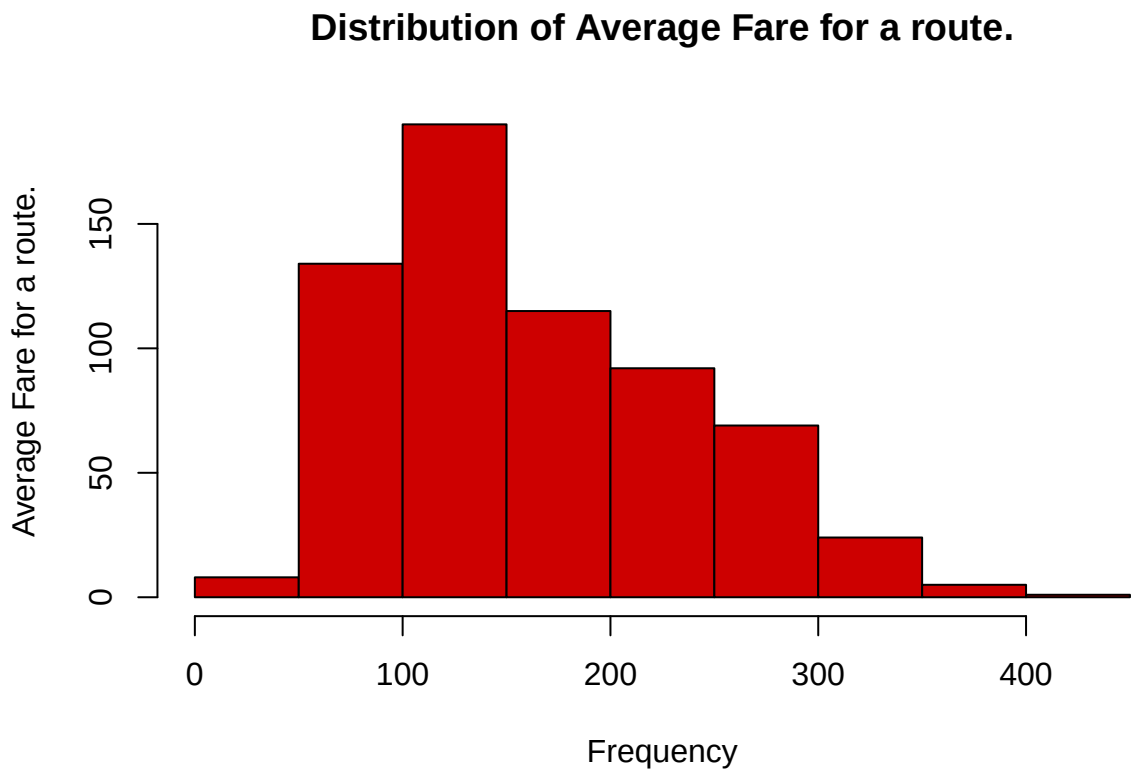
```
## [1] "Displaying the first 6 records of all the numeric variables in the I/P file."
```

```
##   COUPON NEW   HI S_INCOME E_INCOME   S_POP   E_POP DISTANCE   PAX   FARE
## 1   1.00   3 5292    28637    21112 3036732  205711      312  7864  64.11
## 2   1.06   3 5419    26993    29838 3532657 7145897      576  8820 174.47
## 3   1.06   3 9185    30124    29838 5787293 7145897      364  6452 207.76
## 4   1.06   3 2657    29260    29838 7830332 7145897      612 25144  85.47
## 5   1.06   3 2657    29260    29838 7830332 7145897      612 25144  85.47
## 6   1.01   3 3408    26046    29838 2230955 7145897      309 13386  56.76
```

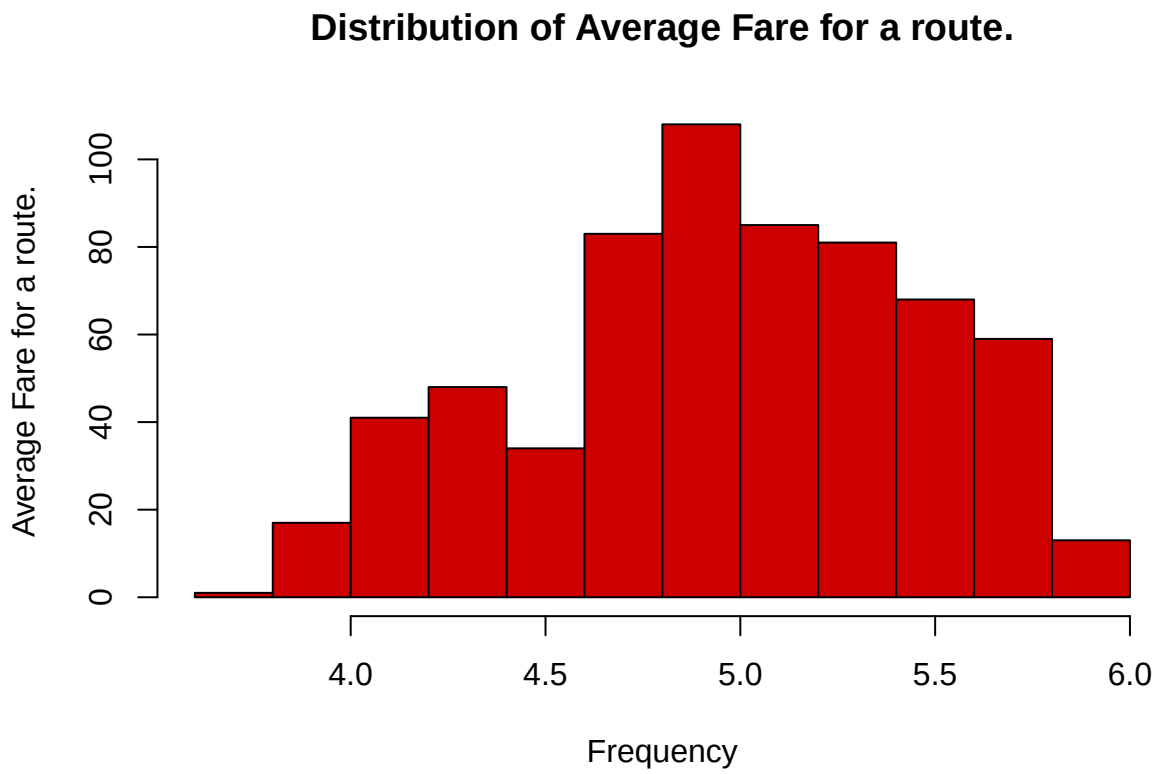
```
## [1] "The Statistical summary of every variable of the data set is mentioned below:"
```

```
##      COUPON          NEW             HI           S_INCOME        E_INCOME
##  Min.   :1.00   Min.   :0.00   Min.   : 1230   Min.   :14600   Min.   :14600
##  1st Qu.:1.04   1st Qu.:3.00   1st Qu.: 3090   1st Qu.:24706   1st Qu.:23903
##  Median :1.15   Median :3.00   Median : 4208   Median :28637   Median :26409
##  Mean   :1.20   Mean   :2.75   Mean   : 4442   Mean   :27760   Mean   :27664
##  3rd Qu.:1.30   3rd Qu.:3.00   3rd Qu.: 5481   3rd Qu.:29694   3rd Qu.:31981
##  Max.   :1.94   Max.   :3.00   Max.   :10000   Max.   :38813   Max.   :38813
##      S_POP            E_POP           DISTANCE         PAX
##  Min.   :  29838   Min.   : 111745   Min.   : 114   Min.   : 1504
##  1st Qu.:1862106   1st Qu.:1228816   1st Qu.: 455   1st Qu.: 5328
##  Median :3532657   Median :2195215   Median : 850   Median : 7792
##  Mean   :4557004   Mean   :3194503   Mean   : 976   Mean   :12782
##  3rd Qu.:7830332   3rd Qu.:4549784   3rd Qu.:1306   3rd Qu.:14090
##  Max.   :9056076   Max.   :9056076   Max.   :2764   Max.   :73892
##       FARE
##  Min.   : 42.5
##  1st Qu.:106.3
##  Median :144.6
##  Mean   :160.9
##  3rd Qu.:209.3
##  Max.   :402.0
```

```
##
## First, let us have a look at the way the FARE data variable is distributed in the input file.
##
hist((inp_file.df$FARE), col='red3', border='black',
     main='Distribution of Average Fare for a route.',
     xlab = 'Frequency', ylab = 'Average Fare for a route.')
```

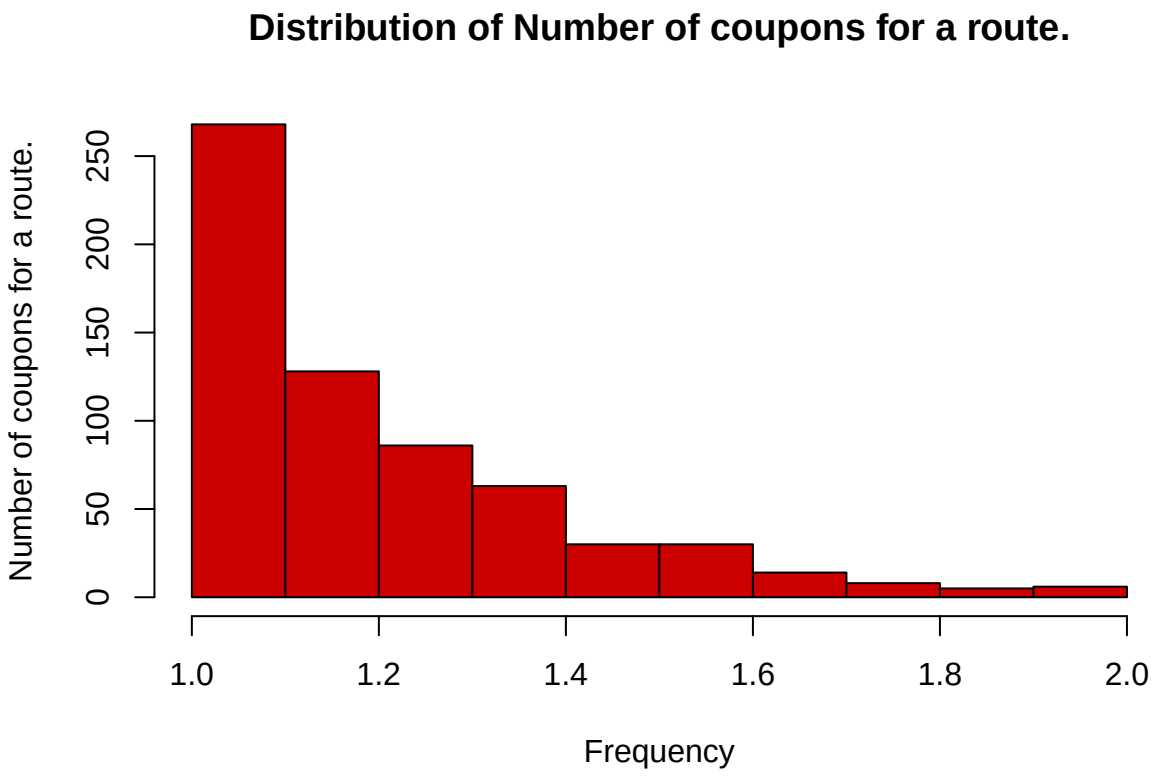



```
hist(log(inp_file.df$FARE), col='red3', border='black',
     main='Distribution of Average Fare for a route.',
     xlab = 'Frequency', ylab = 'Average Fare for a route.')
```



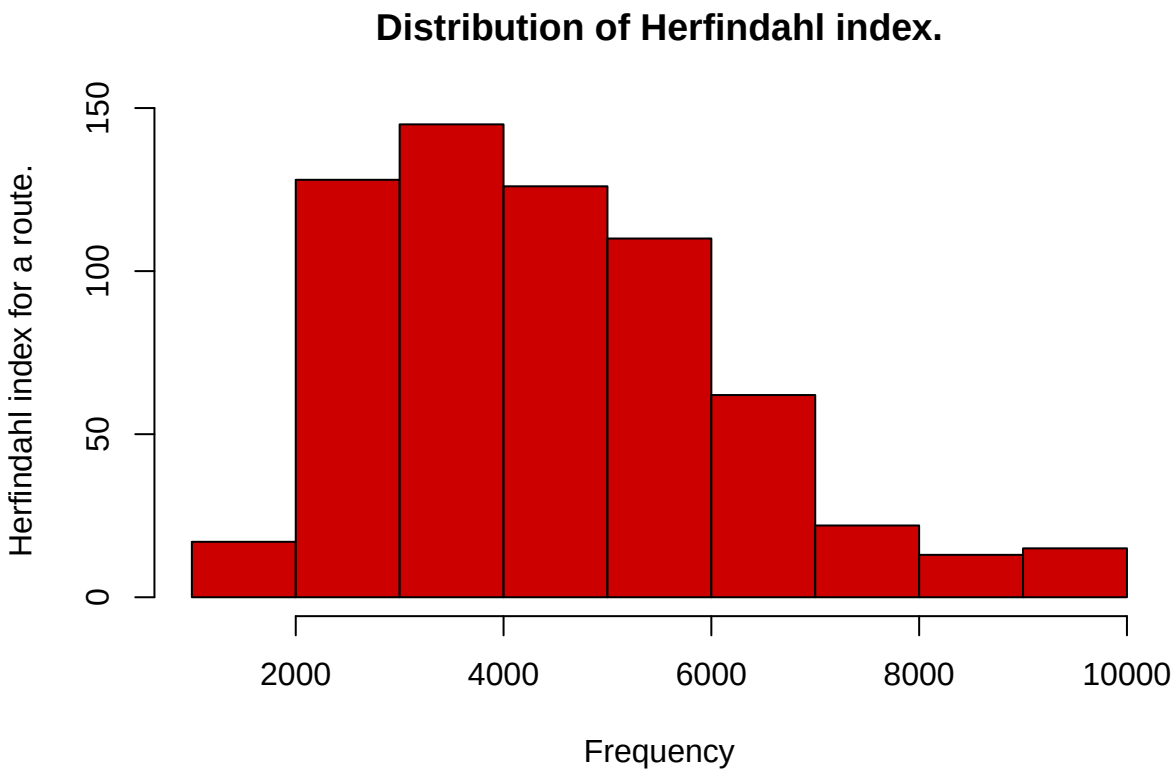
The above histograms shows the distribution of Average Fare for a route. The Fare distribution has been log transformed so that the percentage change in the Air Fares is approximately normally distributed.

```
##
## Now, let us have a look at the way COUPON data variable is distributed in the input file.
##
hist(inp_file.df$COUPON, col='red3', border='black',
     main='Distribution of Number of coupons for a route.',
     xlab = 'Frequency', ylab = 'Number of coupons for a route.')
```



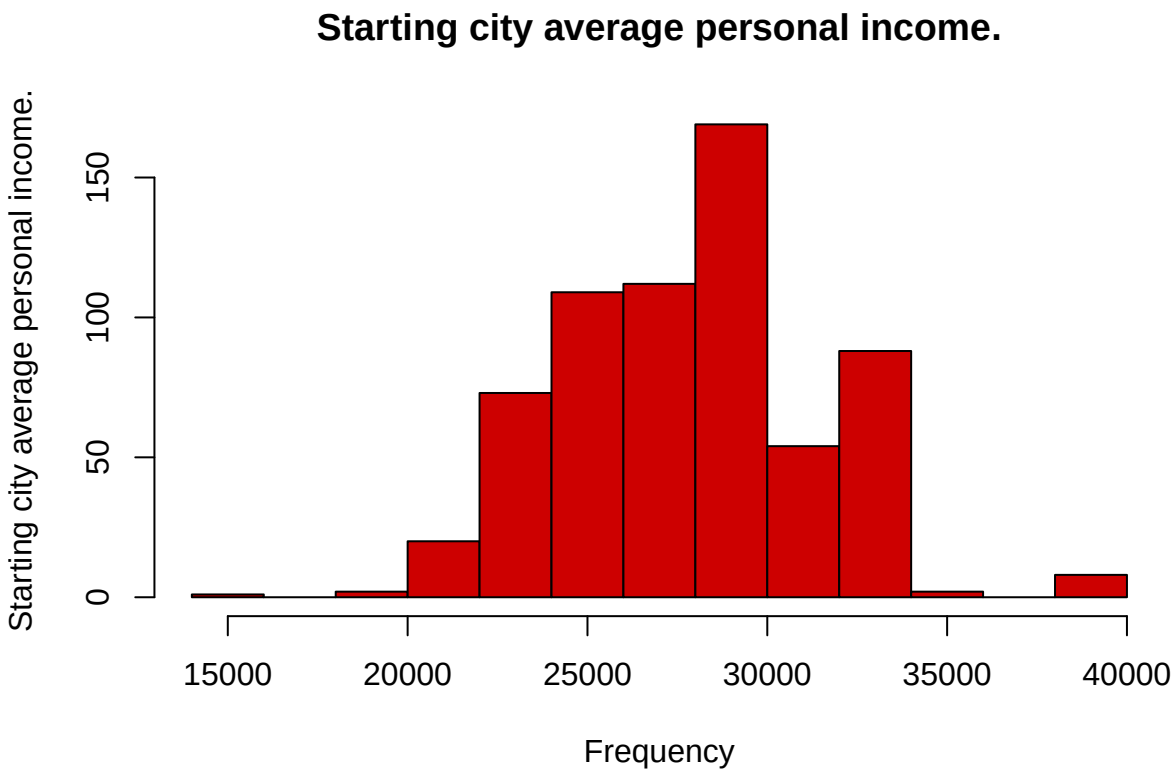
The above histogram shows the sitribution of number of coupons for a route. It can be seen that for majority of the routes, the average number of coupons that are present are close to 1.

```
##
## Now let'slook at the way HI data column is distributed in the input file.
##
hist((inp_file.df$HI), col='red3', border='black',
     main='Distribution of Herfindahl index.',
     xlab = 'Frequency', ylab = 'Herfindahl index for a route.')
```

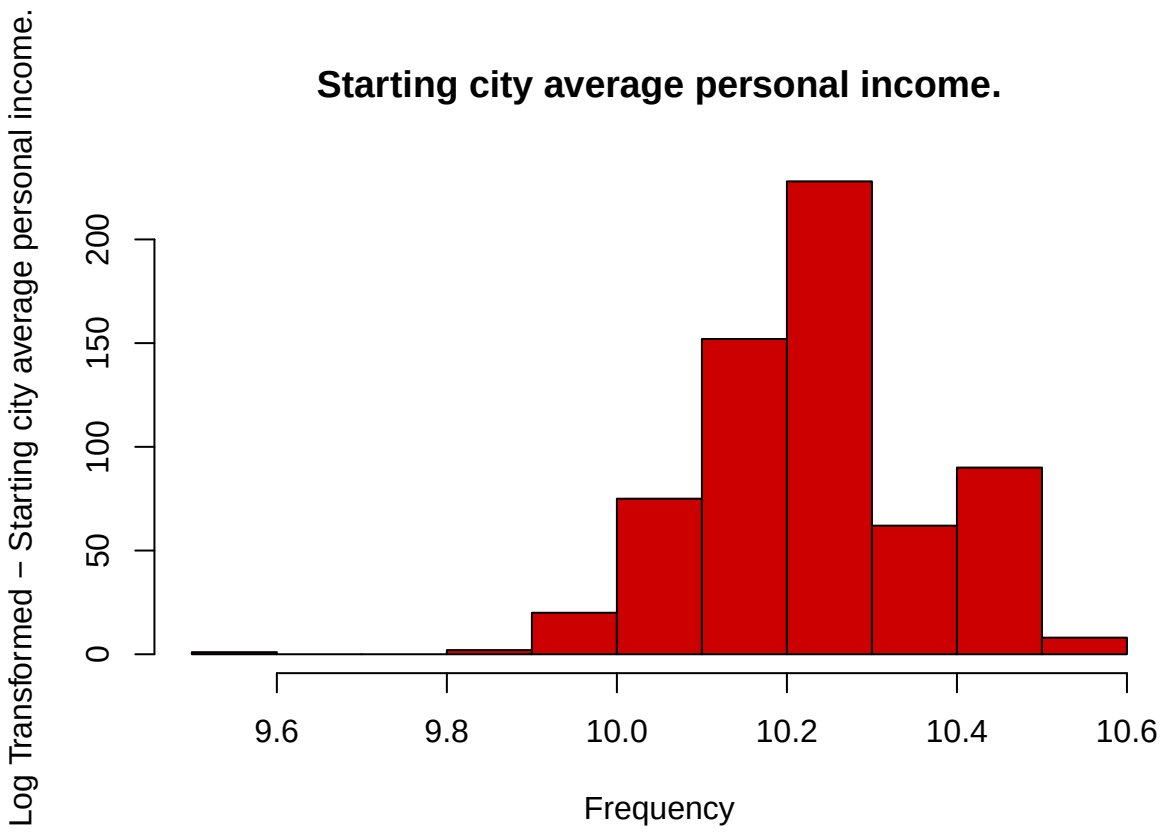


The above histogram shows the distribution of HI Index, a measure of market concentration (higher number means smaller number of available carriers on that route). The distribution seems to be approximately normal for HI index.

```
##
## Now let's look at the way S_INCOME data is distributed in the input file.
##
hist((inp_file.df$S_INCOME), col='red3', border='black',
     main='Starting city average personal income.',
     xlab = 'Frequency', ylab = 'Starting city average personal income.')
```

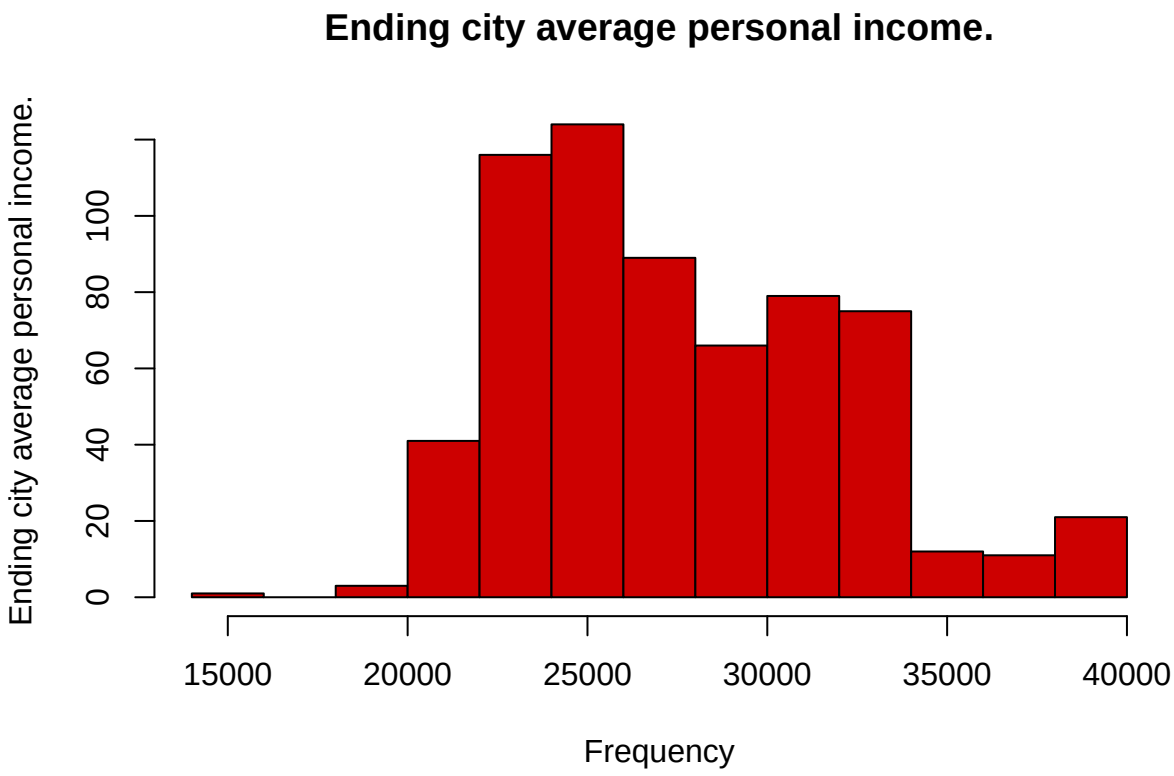



```
hist(log(inp_file.df$S_INCOME), col='red3', border='black',
     main='Starting city average personal income.', xlab = 'Frequency',
     ylab = 'Log Transformed - Starting city average personal income.')
```

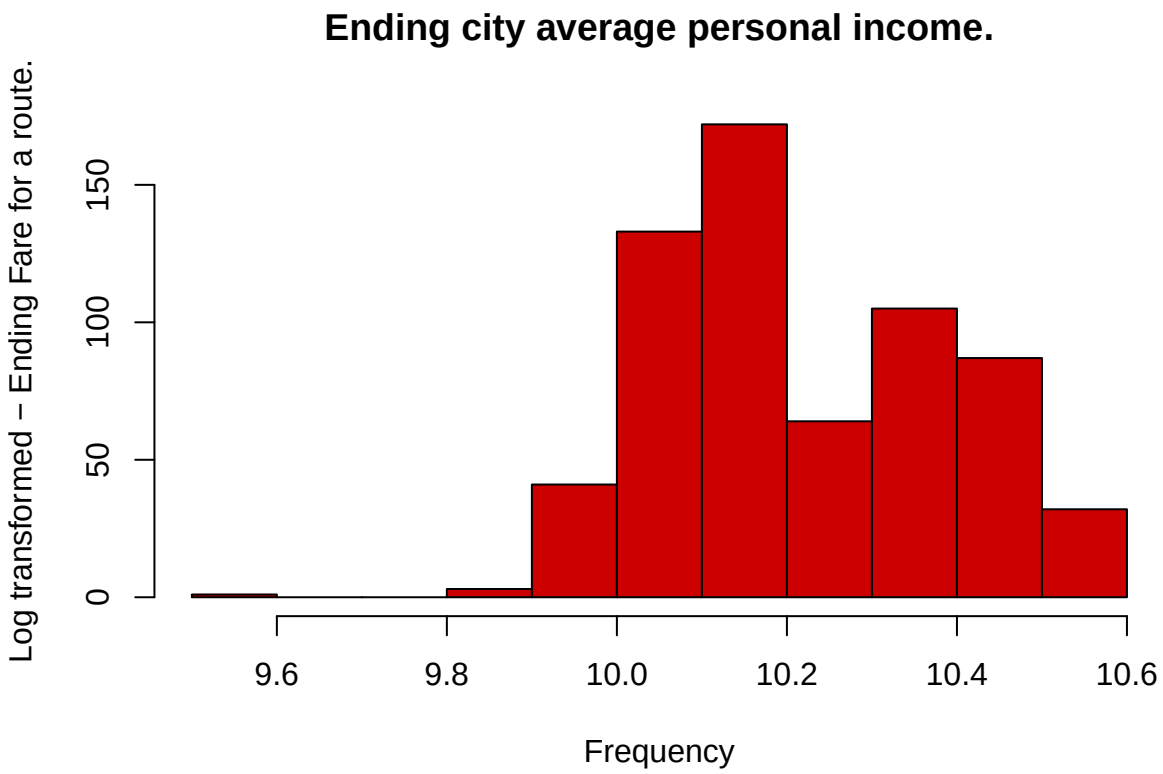


The above histograms shows the distribution of starting city's personal income. The starting point's average annual income is log transformed to capture the percentage change in the income.

```
##
## Let's look at the way end city personal income data is distributed in the input file.
##
hist((inp_file.df$E_INCOME), col='red3', border='black',
     main='Ending city average personal income.',
     xlab = 'Frequency', ylab = 'Ending city average personal income.')
```

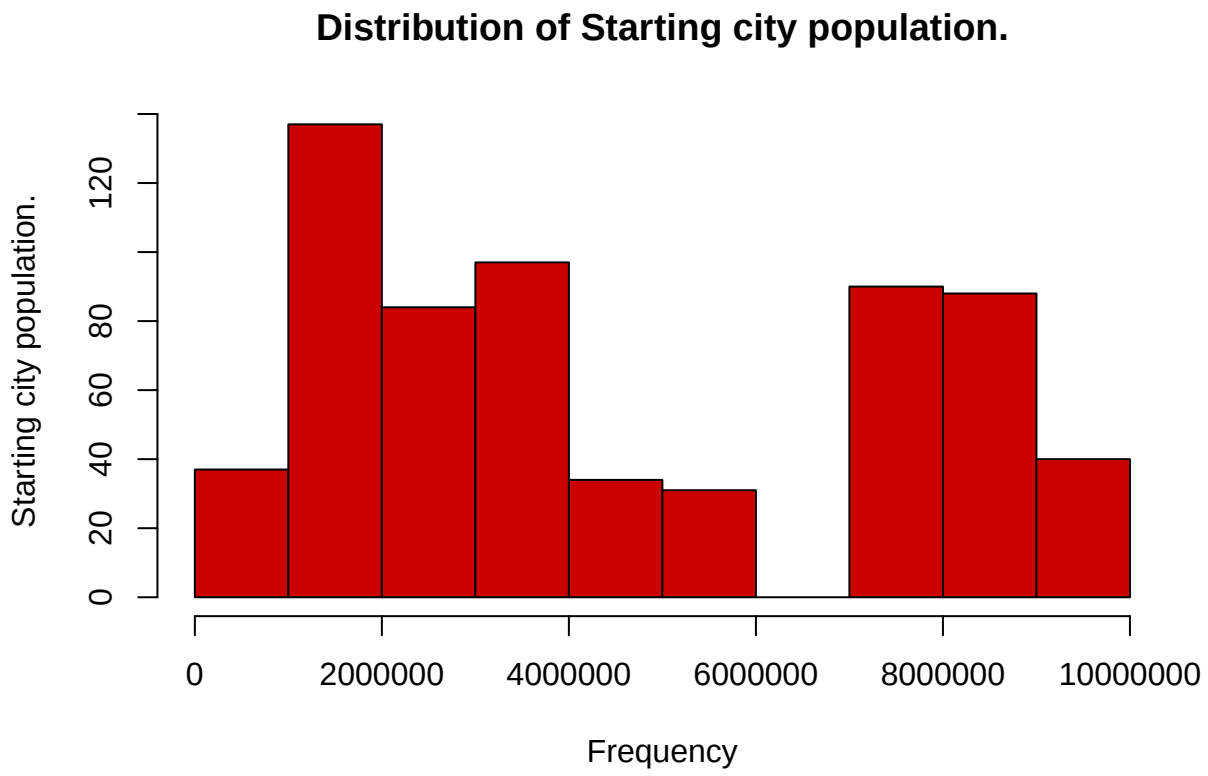



```
hist(log(inp_file.df$E_INCOME), col='red3', border='black',
     main='Ending city average personal income.',
     xlab = 'Frequency', ylab = 'Log transformed - Ending Fare for a route.')
```



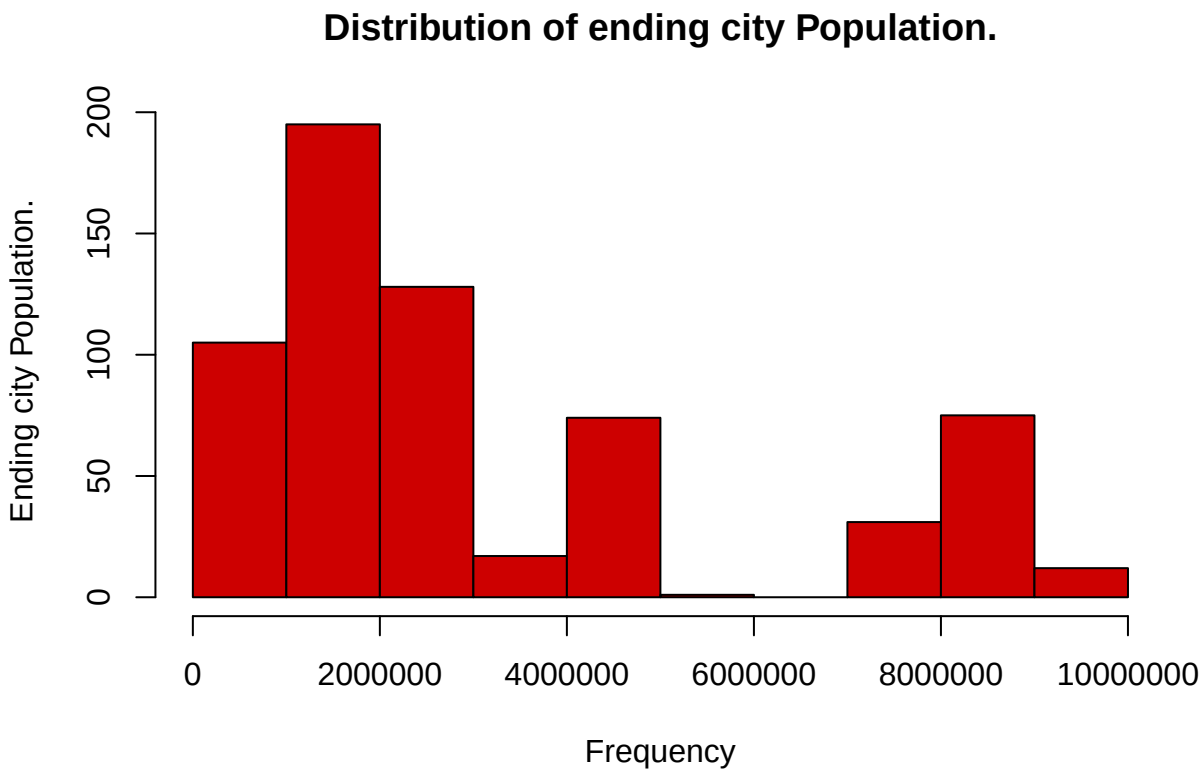
The above histogram shows the log transformed version of the End city/destination city's average personal income. The end point's average annual income is log transformed to capture the percentage change in the income.

```
##
## Let us have a look at the way start city population data is distributed in the input file.
##
hist((inp_file.df$S_POP), col='red3', border='black',
     main='Distribution of Starting city population.',
     xlab = 'Frequency', ylab = 'Starting city population.')
```



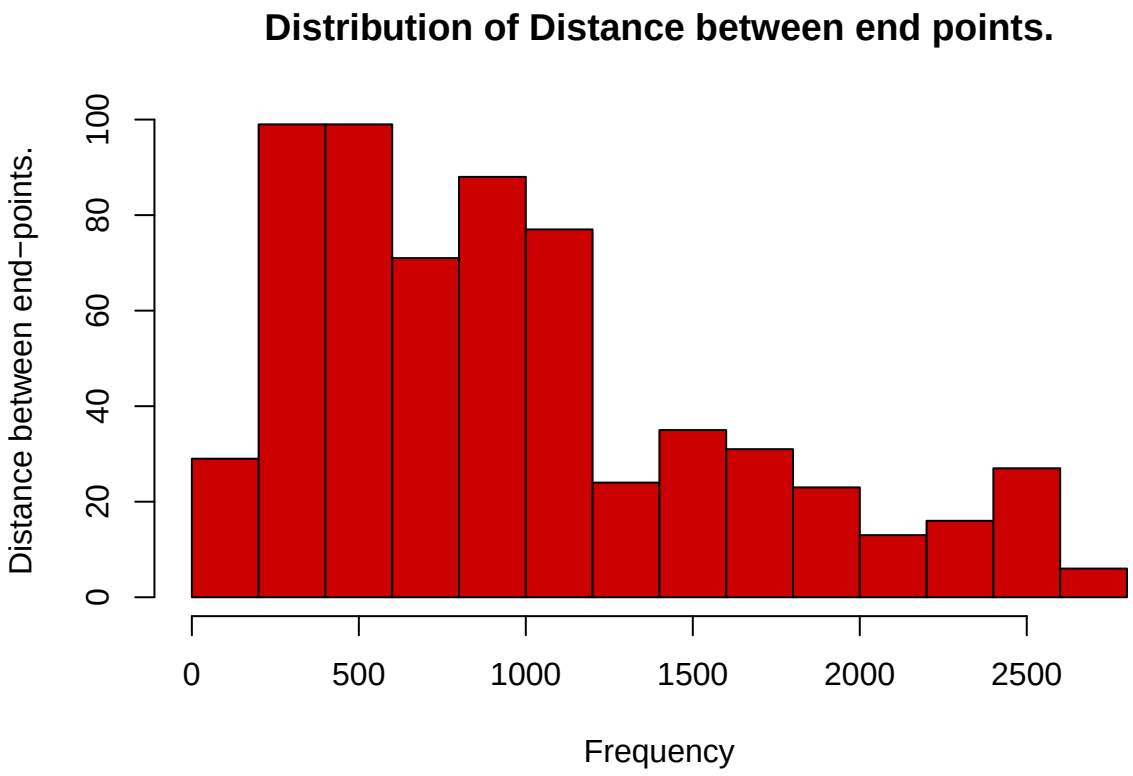
The above histogram shows the distribution of the Starting city's population. It can be seen that for almost 140 observations the starting city was the same.

```
## Let us have a look at the way ending city population data is distributed in the input file.
##
hist((inp_file.df$E_POP), col='red3', border='black',
     main='Distribution of ending city Population.',
     xlab = 'Frequency', ylab = 'Ending city Population.')
```



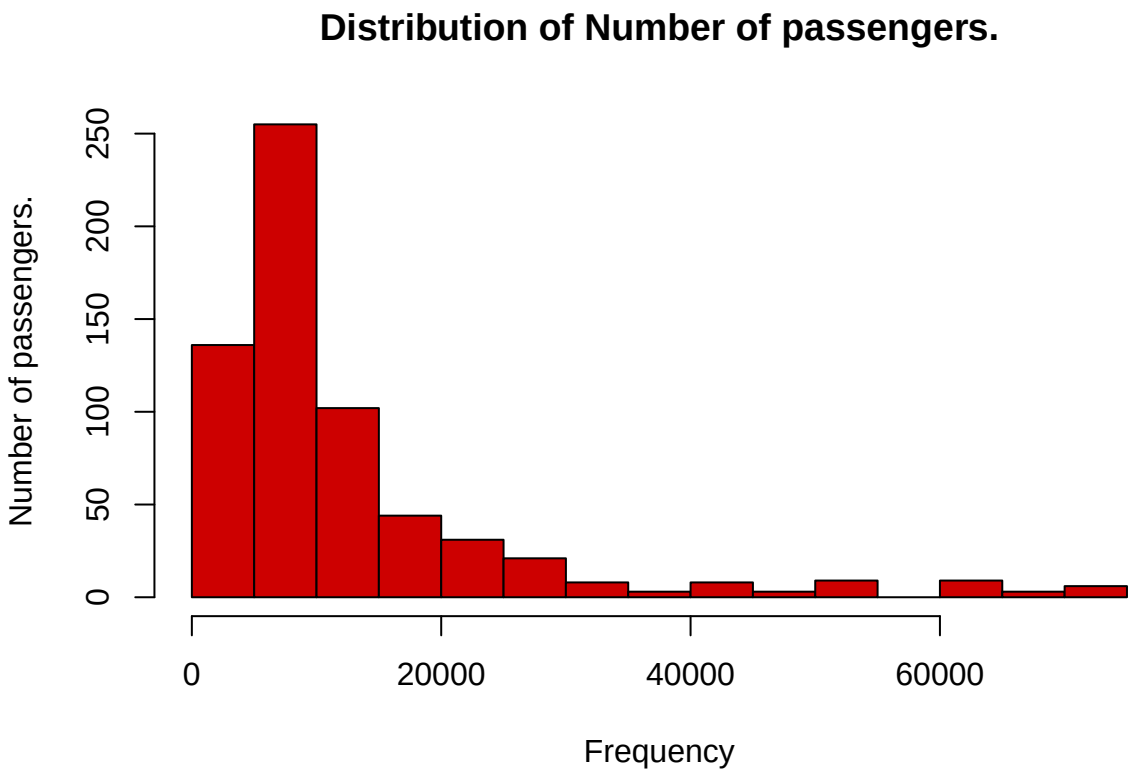
The above histogram shows the population distribution for the ending city. It can be seen that there are almost 200 observations where the ending city was the same.

```
##
## Let us have a look at the way distance between two city data is distributed in the input file.
##
hist(inp_file.df$DISTANCE, col='red3', border='black',
     main='Distribution of Distance between end points.',
     xlab = 'Frequency', ylab = 'Distance between end-points.')
```



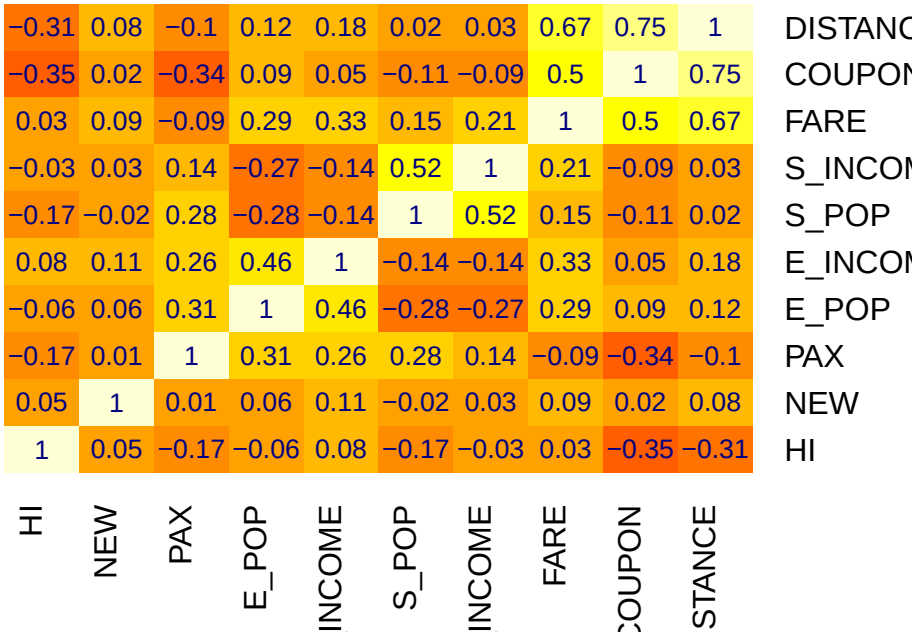
The above histogram shows the distribution of distance between two end points.

```
##
## Let us have a look at the way number of passengers data is distributed in the input file.
##
hist((inp_file.df$PAX), col='red3', border='black',
     main='Distribution of Number of passengers.',
     xlab = 'Frequency', ylab = 'Number of passengers.')
```



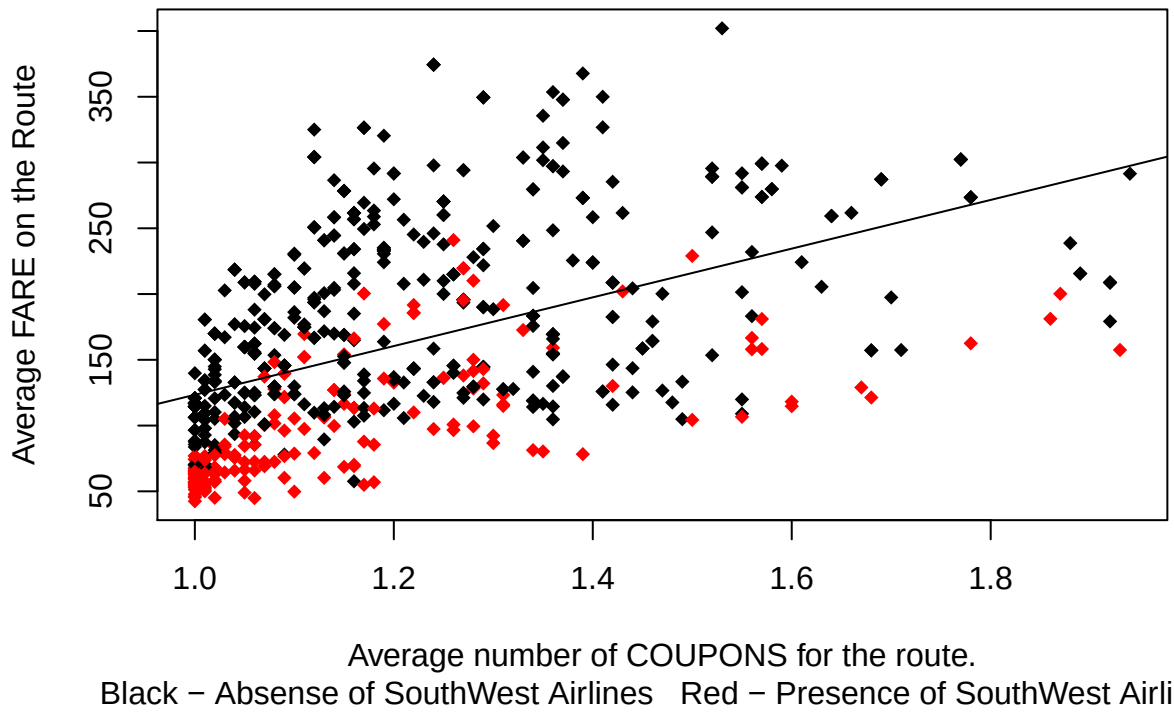
The above histogram explains the distribution of number of passengers in a route.

```
##
## Let's plot the heat map to look at the correlation between variables in the data set.
##
heatmap.2(cor(inp_file_num), dendrogram = "none",
          cellnote = round(cor(inp_file_num),2),
          notecol = "navy", key = FALSE, trace = "none", symm=T)
```



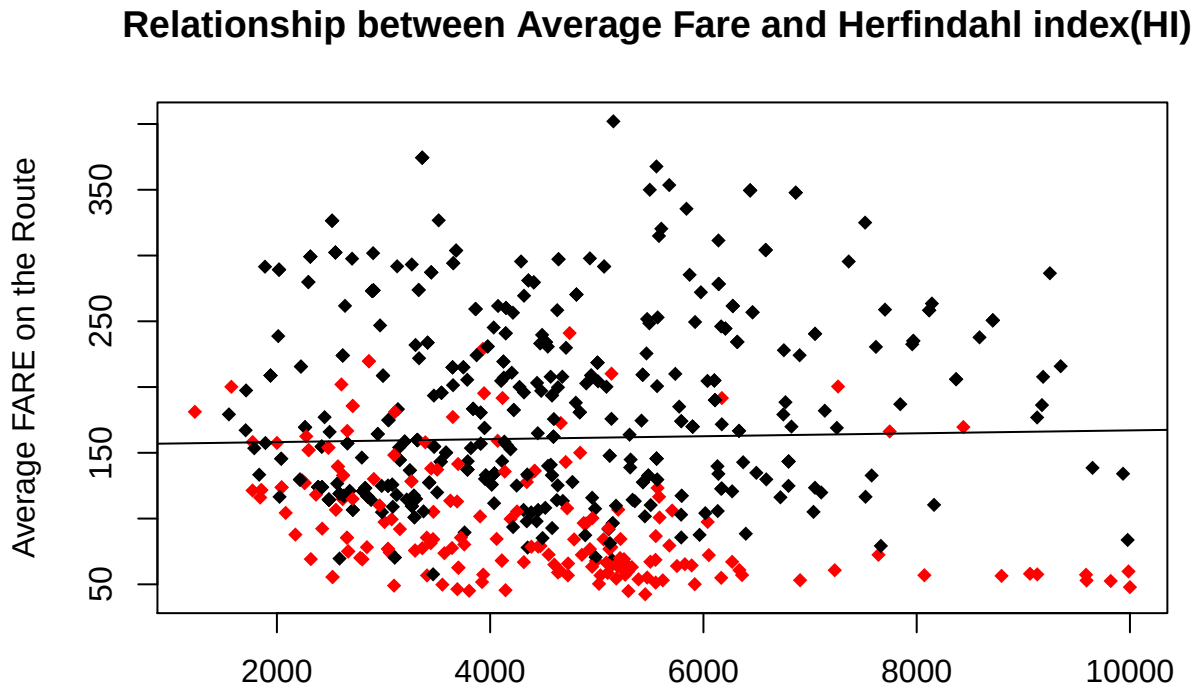
The above heat-map shows the correlation between every variable. 1 - It can be seen that FARE has highest positive correlation with DISTANCE. It would mean that with increase in distance, the FARE is going to increase. 2 - DISTANCE has a strong positive correlation between COUPON. It means that, if distance between two points is more, then it is likely that there will be more coupons for that route. 3 - DISTANCE has the high negative correlation with HI. It can mean that, if distance between two points is less, then there would be lesser flights opertaing and so the HI index would be more. 4 - COUPON has the highest positive correlation with DISTANCE. It would mean that if distance is more, then there is a possibility that there will be more coupons for that route. 5 - COUPON has high negative correlation with HI. It would mean that if a route has lesser flights, then the HI index would be more and coupons for that route will be less.

```
##
##  Relation between Average Fare for a route and the Average number of Coupons for the route.
##
coupon_sct <- plot(inp_file.df$COUPON, inp_file.df$FARE, pch=18, col=factor(inp_file.df$SW),
                   xlab='Average number of COUPONS for the route.', ylab ='Average FARE on the Route',
                   sub='Black - Absense of SouthWest Airlines   Red - Presence of SouthWest Airlines',
                   main='Relationship between Average Fare and Average Number of Coupons.')
abline(lm(inp_file.df$FARE~inp_file.df$COUPON))
```



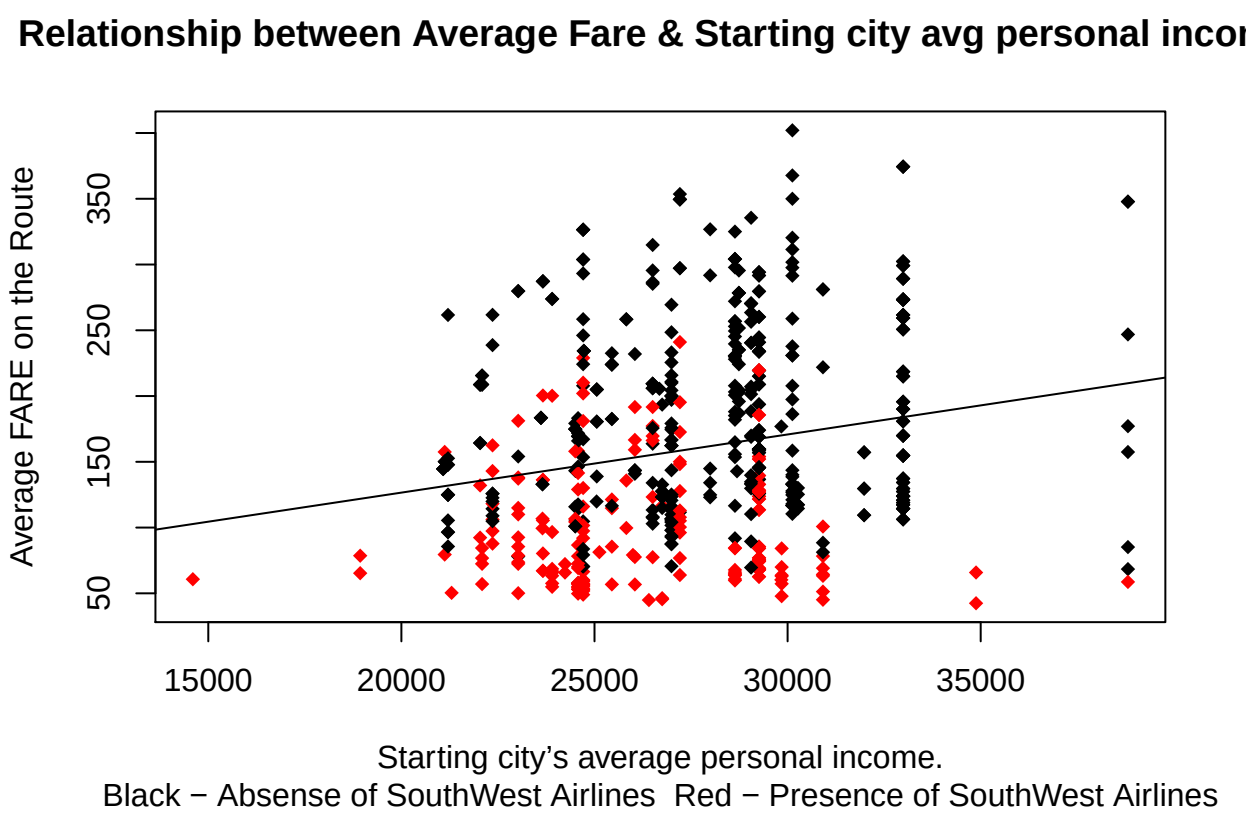
The above scatter plot explains the relationship between Average COUPONS for a route and the FARE for that route. There is a positive linear relationship between both variables. With increase in FARE for a route, there are more number of coupons for that route. It can also be seen that for routes where SouthWest airlines is operating (denatoted by RED colour), the FARE is low in majority of the cases.

```
##
## Let's check the relationship between Average Fare of a route with
## the HI (Herfindahl index, a measure of market concentration)
##
plot(inp_file.df$HI, inp_file.df$FARE, pch=18, col=factor(inp_file.df$SW),
     xlab='Herfindahl index, a measure of market concentration.', ylab ='Average FARE on the Route',
     sub='Black - Absense of SouthWest Airlines  Red - Presence of SouthWest Airlines',
     main='Relationship between Average Fare and Herfindahl index(HI).')
abline(lm(inp_file.df$FARE~inp_file.df$HI))
```



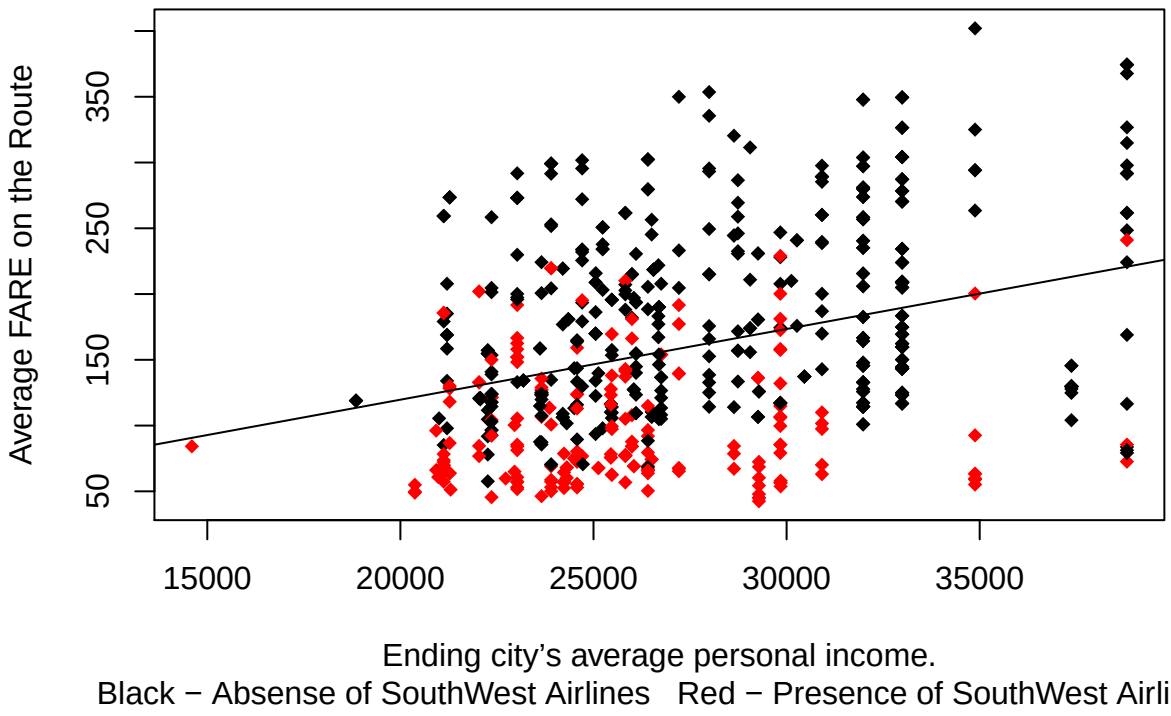
The above scatter plot shows the relationship between FARE and HI index. It can be seen that there is no significant relationship between FARE and HI index. It can also be verifies from the Heat Map generated above where the correlation between HI and FARE is very close to 0.

```
##
## Let's check the relationship between Average Fare for a route(FARE)
## and Starting City's Average Personal Income (S_INCOME)
##
plot(inp_file.df$S_INCOME, inp_file.df$FARE, pch=18, col=factor(inp_file.df$SW),
     xlab='Starting city's average personal income.', ylab ='Average FARE on the Route',
     sub='Black - Absense of SouthWest Airlines  Red - Presence of SouthWest Airlines',
     main='Relationship between Average Fare & Starting city avg personal income.')
abline(lm(inp_file.df$FARE~inp_file.df$S_INCOME))
```



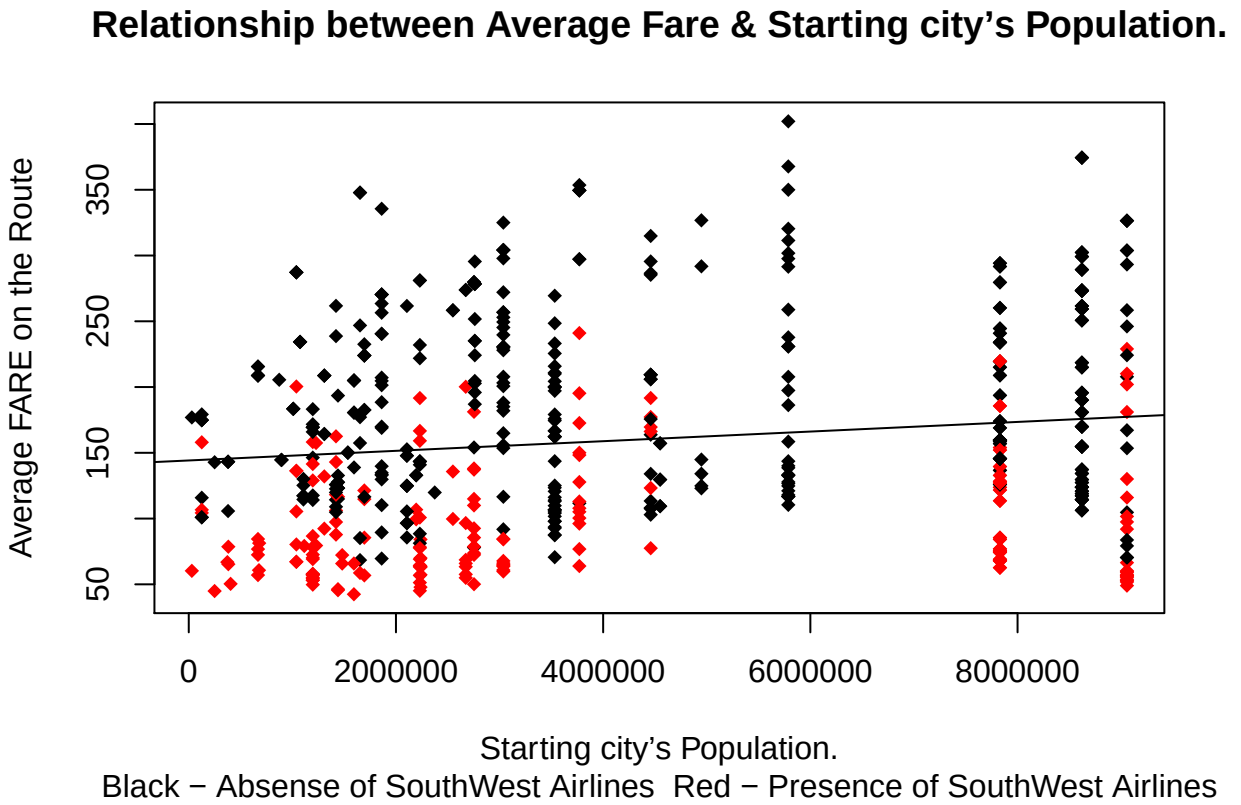
The above scatter plot shows he relationship between FARE and Starting City's Average Income. We can see that it has a linear positive relationship between FARE and S_INCOME. Also, based on the plot we can say that, majority of the SoutWest Airlines customer starting from a city are having a average personal income in between $20000 to $30000. Also, we can infer that for majority of the cases, the FARE for a route operated by SouthWest Airlines is below $200.

```
##
## Now, let's check the relationship between Average Fare for a route(FARE)
## and Ending City's Average Personal Income (S_INCOME)
##
plot(inp_file.df$E_INCOME, inp_file.df$FARE, pch=18, col=factor(inp_file.df$SW),
     xlab='Ending city's average personal income.', ylab ='Average FARE on the Route',
     sub='Black - Absense of SouthWest Airlines   Red - Presence of SouthWest Airlines',
     main='Relationship between Average Fare & Ending city's avg personal income.')
abline(lm(inp_file.df$FARE~inp_file.df$E_INCOME))
```



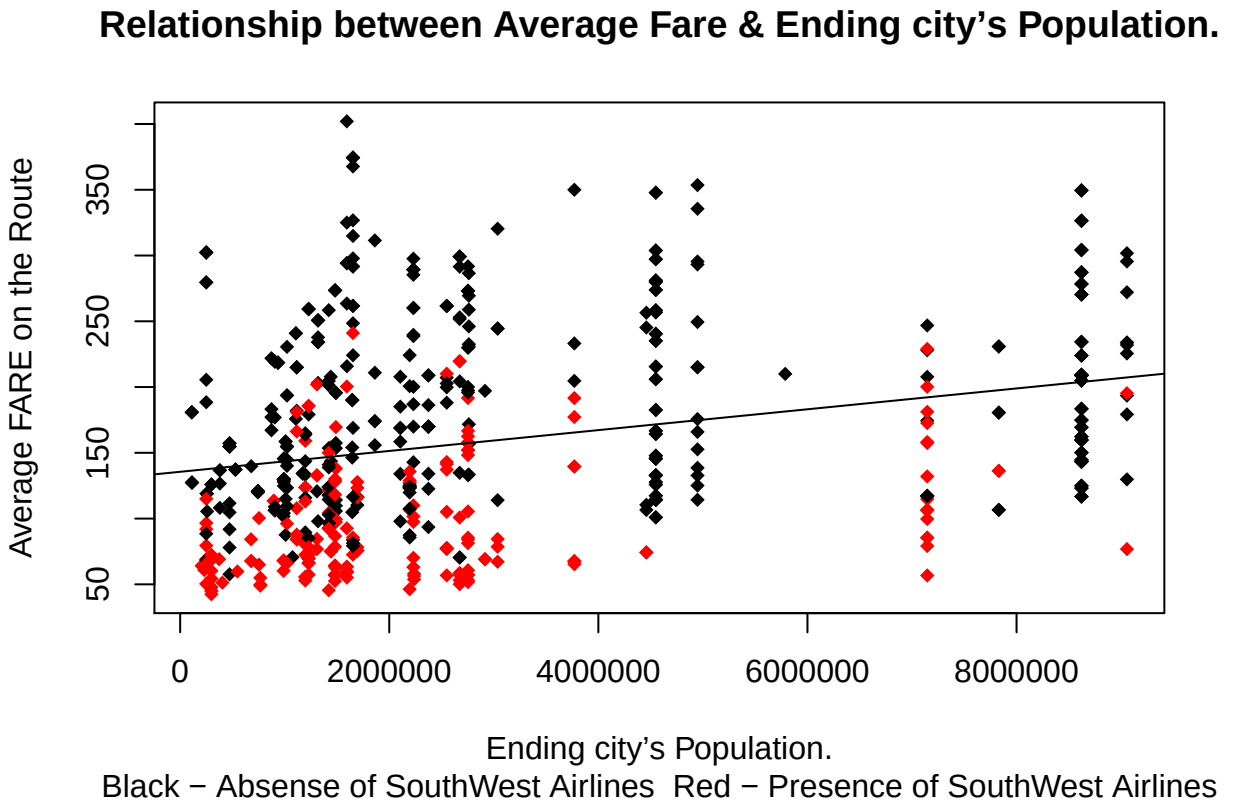
The above scatter plot shows he relationship between FARE and Ending City's Average Income. We can see that it has a linear positive relationship between FARE and E_INCOME. Also, based on the plot we can say that, majority of the SoutWest Airlines customer starting from a city are having a average personal income in between $22000 to $30000. Also, we can infer that for majority of the cases, the FARE for a route operated by SouthWest Airlines is below $200.

```
##
## Let's check the relationship between Average Fare for a route(FARE)
## and Starting City's Population (S_POP)
##
plot(inp_file.df$S_POP, inp_file.df$FARE, pch=18, col=factor(inp_file.df$SW),
     xlab='Starting city's Population.', ylab ='Average FARE on the Route',
     sub='Black - Absense of SouthWest Airlines  Red - Presence of SouthWest Airlines',
     main='Relationship between Average Fare & Starting city's Population.')
abline(lm(inp_file.df$FARE~inp_file.df$S_POP))
```



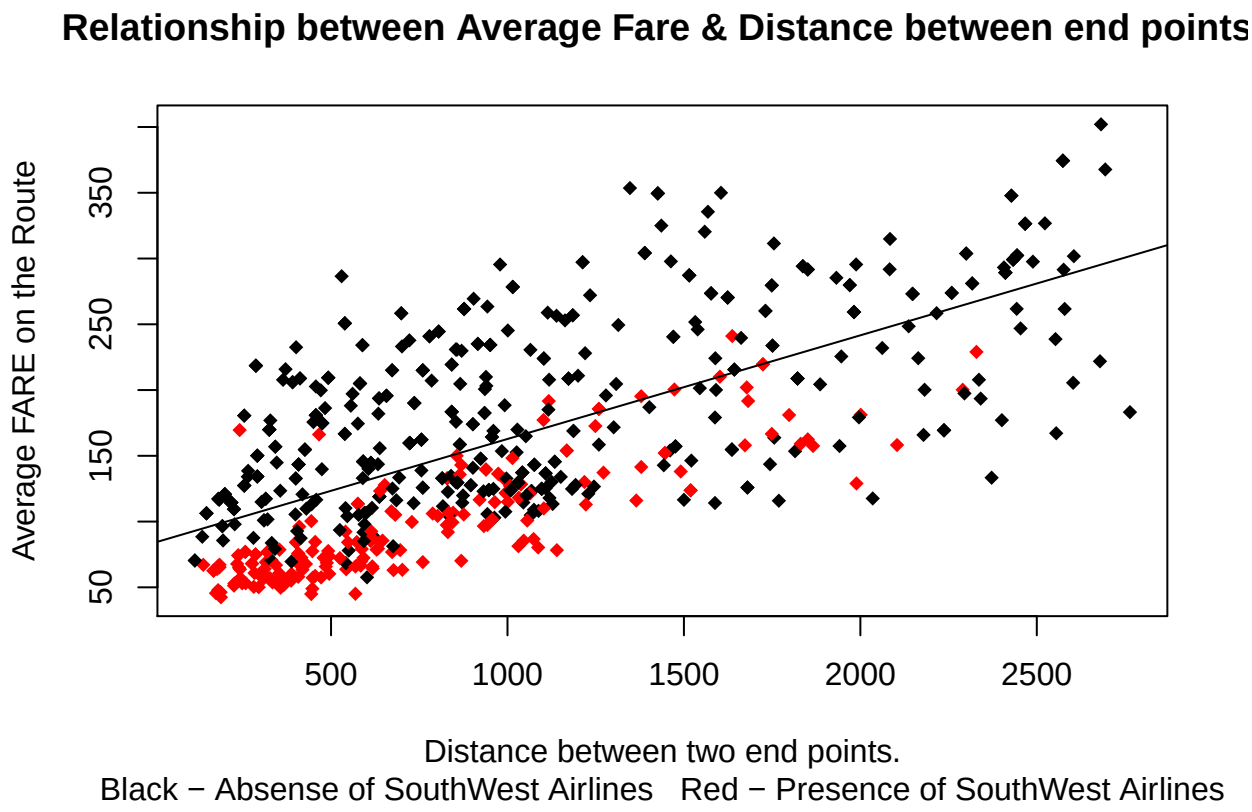
The above scatter plot explains the relationship between Avergae Fare and Starting City's population. There is a small positive correlation between both the variables. It can also be verified from the heat map generated above which shows the correlation between both variables as 0.15

```
##
## Let's check the relationship between Average Fare for a route(FARE)
## and Ending City's Population (S_POP)
##
plot(inp_file.df$E_POP, inp_file.df$FARE, pch=18, col=factor(inp_file.df$SW),
     xlab='Ending city's Population.', ylab ='Average FARE on the Route',
     sub='Black - Absense of SouthWest Airlines  Red - Presence of SouthWest Airlines',
     main='Relationship between Average Fare & Ending city's Population.')
abline(lm(inp_file.df$FARE~inp_file.df$E_POP))
```



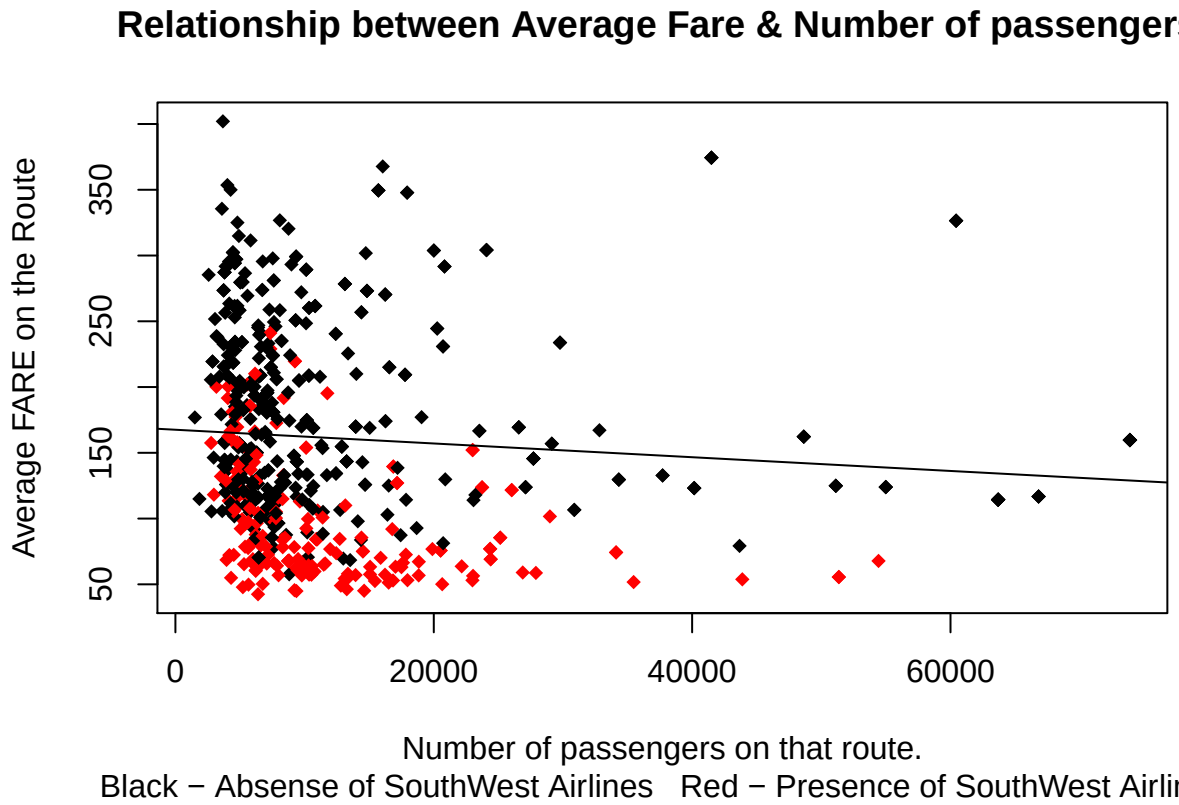
The above scatter plot explains the relationship between Avergae Fare and Starting City's population. There is a small positive correlation between both the variables. It can also be verified from the heat map generated above which shows the correlation between both variables as 0.15

```
##
## Now, let's check the relationship between Average Fare for a route(FARE)
## and the distance between two end points.
##
plot(inp_file.df$DISTANCE, inp_file.df$FARE, pch=18, col=factor(inp_file.df$SW),
     xlab='Distance between two end points.', ylab ='Average FARE on the Route',
     sub='Black - Absense of SouthWest Airlines   Red - Presence of SouthWest Airlines',
     main='Relationship between Average Fare & Distance between end points.')
abline(lm(inp_file.df$FARE~inp_file.df$DISTANCE))
```



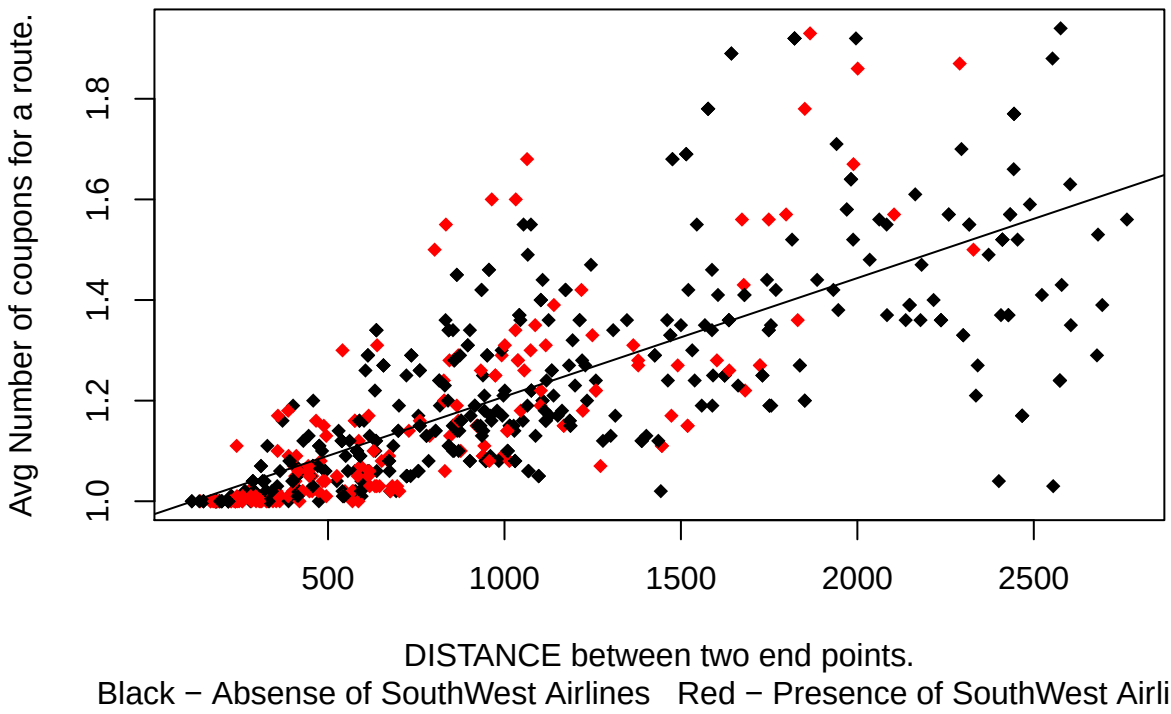
The above scatter plot explains the relationship between FARE and DISTANCE between two points. We can see that there is a strong positive correlation between DISTANCE and FARE. Also, we can infer that for majority of the cases, the FARE for a route operated by SouthWest Airlines is below $200.

```
##
## Lastly, let's check the relationship between Average Fare for a route(FARE)
## and the Number of passsenger on that route.
##
plot(inp_file.df$PAX, inp_file.df$FARE, pch=18, col=factor(inp_file.df$SW),
     xlab='Number of passengers on that route.', ylab ='Average FARE on the Route',
     sub='Black - Absense of SouthWest Airlines   Red - Presence of SouthWest Airlines',
     main='Relationship between Average Fare & Number of passengers.')
abline(lm(inp_file.df$FARE~inp_file.df$PAX))
```



The above scatter plot explains the relationship between Average Fare and Number of Passengers on that route. We can see that there is a negative linear relationship between FARE and Number of passengers. Also, we can infer that for majority of the cases, the FARE for a route operated by SouthWest Airlines is below $200.

```
##
## Let's check the relationship between Average Fare for a route(FARE)
## and the Number of passsenger on that route.
##
plot(inp_file.df$DISTANCE, inp_file.df$COUPON, pch=18, col=factor(inp_file.df$SW),
     xlab='DISTANCE between two end points.', ylab ='Avg Number of coupons for a route.',
     sub='Black - Absense of SouthWest Airlines   Red - Presence of SouthWest Airlines',
     main='Average number of Coupons ~ Distance between two points.')
abline(lm(inp_file.df$COUPON~inp_file.df$DISTANCE))
```



The above scatter plot expalins the relationship between Number of Coupons for a route and the distance between two end points. We can see that both variables have a strong positive correlation. It can also be verified from the heat map / correlaton map generated which shows a postive correlation of 0.75 between both variables.

```
## [1] "Percentage of flights based on Vacation catagory: "

##
##     No    Yes
## 0.7335 0.2665

## [1] "Percentage of flights based on SW catagory: "

##
##     No    Yes
## 0.6959 0.3041

## [1] "Percentage of flights based on Slot catagory:"

##
## Controlled       Free
##     0.2853     0.7147

## [1] "Percentage of flights based on Gate catagory:"
```

```
##
## Constrained        Free
##      0.1944      0.8056
```

1 - We can see that almost 27 percent of the flights are operated on the vacation route and the rest other flights are operated on regular routes. 2 - For almost 30 percent of the routes, SouthWest airlines is providing it's service. 3 - Almost 28.5 percent of the routes are Slot controlled. 4 - Similarly, almost 20 percent of the endpoints are gate controlled.

```
## Creating and displaying a pivot table with average fare in each catagory.
pt <- PivotTable$new()
pt$addData(inp_file.df)
pt$addColumnDataGroups("SW")
pt$addRowDataGroups("VACATION")
pt$addRowDataGroups("SLOT")
pt$addRowDataGroups("GATE")
pt$defineCalculation(calculationName="Mean Fare", summariseExpression="mean(FARE)")
pt$renderPivot()
```

The above pivot table summerizes the Average Fare basedon each categorical variable and it's combination. For Example, the average fare for a route where SouthWest airlines doesn't operates and which is not on a vacation route and is not slot controlled and not Gate constrained is around $196.18

For building the model, below code will partition the data to training and testing dataset and will be used in subsequent code.

From the total data set, 75% of records are allocated to the training data set and rest 25% to the testing data set. We will build and train the model using the training data set and will test it's efficiency using the testing data set.

Now we will build a linear regression model using the Step Wise subset selection process which will help us to get the best model.

```
## [1] "Statistics of stepwise regression analysis with leap package:"
```

```
## Subset selection object
## Call: regsubsets.formula(FARE ~ COUPON + NEW + VACATION + SW + HI +
##     S_INCOME + E_INCOME + S_POP + E_POP + SLOT + GATE + DISTANCE +
##     PAX, data = train.df, nvmax = dim(train.df)[2], method = "seqrep")
## 13 Variables  (and intercept)
##             Forced in Forced out
## COUPON          FALSE      FALSE
## NEW             FALSE      FALSE
## VACATIONYes     FALSE      FALSE
## SWYes           FALSE      FALSE
## HI              FALSE      FALSE
## S_INCOME        FALSE      FALSE
## E_INCOME        FALSE      FALSE
## S_POP           FALSE      FALSE
## E_POP           FALSE      FALSE
## SLOTFree        FALSE      FALSE
## GATEFree        FALSE      FALSE
## DISTANCE        FALSE      FALSE
```

```
## PAX             FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: 'sequential replacement'
##           COUPON NEW VACATIONYes SWYes HI  S_INCOME E_INCOME S_POP E_POP
## 1  ( 1 )  " "    " " " "         " "   " " " "      " "      " "   " "
## 2  ( 1 )  " "    " " " "         "*"   " " " "      " "      " "   " "
## 3  ( 1 )  " "    " " "*"         "*"   " " " "      " "      " "   " "
## 4  ( 1 )  " "    " " "*"         "*"   "*" " "      " "      " "   " "
## 5  ( 1 )  " "    " " "*"         "*"   "*" " "      " "      " "   " "
## 6  ( 1 )  " "    " " "*"         "*"   "*" " "      " "      " "   " "
## 7  ( 1 )  " "    " " "*"         "*"   "*" " "      "*"      " "   " "
## 8  ( 1 )  " "    " " "*"         "*"   "*" " "      "*"      " "   " "
## 9  ( 1 )  " "    " " "*"         "*"   "*" " "      " "      "*"   "*"
## 10  ( 1 ) "*"    "*" "*"         "*"   "*" "*"      "*"      "*"   "*"
## 11  ( 1 ) " "    " " "*"         "*"   "*" "*"      "*"      "*"   "*"
## 12  ( 1 ) " "    "*" "*"         "*"   "*" "*"      "*"      "*"   "*"
## 13  ( 1 ) "*"    "*" "*"         "*"   "*" "*"      "*"      "*"   "*"
##           SLOTFree GATEFree DISTANCE PAX
## 1  ( 1 )  " "      " "      "*"      " "
## 2  ( 1 )  " "      " "      "*"      " "
## 3  ( 1 )  " "      " "      "*"      " "
## 4  ( 1 )  " "      " "      "*"      " "
## 5  ( 1 )  "*"      " "      "*"      " "
## 6  ( 1 )  "*"      "*"      "*"      " "
## 7  ( 1 )  "*"      "*"      "*"      " "
## 8  ( 1 )  "*"      "*"      "*"      "*"
## 9  ( 1 )  "*"      "*"      "*"      "*"
## 10  ( 1 ) "*"      " "      " "      " "
## 11  ( 1 ) "*"      "*"      "*"      "*"
## 12  ( 1 ) "*"      "*"      "*"      "*"
## 13  ( 1 ) "*"      "*"      "*"      "*"

##    (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   TRUE  TRUE        TRUE  TRUE  TRUE     TRUE     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    FALSE    FALSE FALSE
## 11  TRUE     TRUE     TRUE     TRUE  TRUE
## 12  TRUE     TRUE     TRUE     TRUE  TRUE
## 13  TRUE     TRUE     TRUE     TRUE  TRUE
```

```
## [1] "The R-Squared value for each combination is displayed below:"
```

```
##  [1] 0.3721 0.5627 0.6674 0.6933 0.7101 0.7358 0.7420 0.7486 0.7625 0.6389
## [11] 0.7694 0.7714 0.7721
```

```
## [1] "The Adjusted R-Squared value for each combination is displayed below:"
```

```
##  [1] 0.3708 0.5609 0.6653 0.6907 0.7071 0.7324 0.7382 0.7443 0.7579 0.6312
## [11] 0.7640 0.7655 0.7657
```

Based on the step wise subset selection process, the model with best adjusted R squared value is being chosen and all the variables are being considered.

Now, let create the best model based on the subset selection process for our prediction.

```
stepwise.best <- lm(FARE ~ COUPON + NEW + VACATION + SW + HI + S_INCOME + E_INCOME
                    + S_POP + E_POP + SLOT + GATE + DISTANCE + PAX, data = train.df)
summary(stepwise.best)
```

```
##
## Call:
## lm(formula = FARE ~ COUPON + NEW + VACATION + SW + HI + S_INCOME +
##     E_INCOME + S_POP + E_POP + SLOT + GATE + DISTANCE + PAX,
##     data = train.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -118.74  -22.41   -1.29   20.76  132.21
##
## Coefficients:
##                  Estimate    Std. Error t value             Pr(>|t|)
## (Intercept) -21.095776311  33.035039668   -0.64              0.52341
## COUPON       16.425531216  13.756518567    1.19              0.23308
## NEW          -3.926493265   2.027995077   -1.94              0.05346 .
## VACATIONYes -31.332223592   4.388186096   -7.14    0.000000000003623 ***
## SWYes       -39.832153354   4.175682426   -9.54 < 0.0000000000000002 ***
## HI            0.008779634   0.001137533    7.72    0.000000000000073 ***
## S_INCOME      0.001409060   0.000604543    2.33              0.02019 *
## E_INCOME      0.002134283   0.000598323    3.57              0.00040 ***
## S_POP         0.000004017   0.000000753    5.33    0.000000152533957 ***
## E_POP         0.000004050   0.000000904    4.48    0.000009481041492 ***
## SLOTFree    -16.758138891   4.489845557   -3.73              0.00021 ***
## GATEFree    -22.946440281   4.479102358   -5.12    0.000000441841791 ***
## DISTANCE      0.068236039   0.004295485   15.89 < 0.0000000000000002 ***
## PAX          -0.000859264   0.000166110   -5.17    0.000000343632903 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.2 on 464 degrees of freedom
## Multiple R-squared:  0.772,  Adjusted R-squared:  0.766
## F-statistic:  121 on 13 and 464 DF,  p-value: <0.0000000000000002
```

Based on the summary of the stepwise.best model, going by the p-value, we can see that apart from COUPON, all other variables included are statistically significant a 5 percent level of significance. Hence we will now remove COUPON from our model and proceed.

```
linear.best <- lm(FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME
                  + S_POP + E_POP + SLOT + GATE + DISTANCE + PAX, data = train.df)
summary(linear.best)
```

```
##
## Call:
## lm(formula = FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME +
##     S_POP + E_POP + SLOT + GATE + DISTANCE + PAX, data = train.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -117.10  -22.79   -1.52   21.22  133.14
##
## Coefficients:
##                   Estimate    Std. Error t value             Pr(>|t|)
## (Intercept)   1.830374399  26.894294450    0.07              0.94577
## NEW          -4.095324077   2.023985559   -2.02              0.04360 *
## VACATIONYes -31.865465438   4.367399211   -7.30      0.0000000000013 ***
## SWYes       -40.053543681   4.173472770   -9.60 < 0.0000000000000002 ***
## HI            0.008420234   0.001097489    7.67      0.0000000000001 ***
## S_INCOME      0.001343434   0.000602315    2.23              0.02619 *
## E_INCOME      0.002102855   0.000598017    3.52              0.00048 ***
## S_POP         0.000003958   0.000000752    5.26      0.0000002173729 ***
## E_POP         0.000004090   0.000000904    4.52      0.0000076965902 ***
## SLOTFree    -17.273391096   4.471105044   -3.86              0.00013 ***
## GATEFree    -23.238036965   4.474486385   -5.19      0.0000003093146 ***
## DISTANCE      0.071829175   0.003066598   23.42 < 0.0000000000000002 ***
## PAX          -0.000938703   0.000152274   -6.16      0.0000000015347 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.2 on 465 degrees of freedom
## Multiple R-squared:  0.771,  Adjusted R-squared:  0.766
## F-statistic:  131 on 12 and 465 DF,  p-value: <0.0000000000000002
```

Based on the summary of the linear.best model, going by the p-value, we can see that all variables included are statistically significant a 5 percent level of significance. Hence we will now consider this model for our prediction purpose.

To confirm our decision, let's compare the predictive accuracy of both the above models (with and without COUPON) using the testing dataset.

```
## [1] "Displaying the accuracy of models with and without COUPON:"
```

```
## [1] "Accuracy of model with COUPON"
```

```
##                ME  RMSE   MAE    MPE  MAPE
## Test set -0.1119 37.38 29.46 -5.406 21.25
```

```
## [1] "Accuracy of model without COUPON"
```

```
##               ME RMSE   MAE    MPE  MAPE
## Test set 0.0232 37.1 29.21 -5.258 21.13
```

Based on the RMSE value of the models we can conclude that the model without the COUPON variable is a better model as the RMSE value is less.

Now, to further validate our findings from the model, let's consider backward selection process with StepAIC to confirm our findings.

```
## [1] "Statistics of backward selection regression analysis with stepAIC:"
```

```
## Start:  AIC=3418
## 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      1766 576598 3418
## <none>                   574832 3418
## - NEW        1      4644 579476 3420
## - S_INCOME   1      6730 581562 3422
## - E_INCOME   1     15764 590595 3429
## - SLOT       1     17259 592091 3430
## - E_POP      1     24846 599678 3436
## - GATE       1     32514 607346 3442
## - PAX        1     33150 607982 3443
## - S_POP      1     35212 610044 3445
## - VACATION   1     63159 637991 3466
## - HI         1     73798 648630 3474
## - SW         1    112729 687561 3502
## - DISTANCE   1    312627 887458 3624
##
## Step:  AIC=3418
## FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME + S_POP +
##     E_POP + SLOT + GATE + DISTANCE + PAX
##
##             Df Sum of Sq     RSS  AIC
## <none>                   576598 3418
## - NEW        1      5077  581675 3420
## - S_INCOME   1      6169  582767 3421
## - E_INCOME   1     15332  591930 3428
## - SLOT       1     18507  595105 3431
## - E_POP      1     25384  601982 3436
## - GATE       1     33445  610043 3443
## - S_POP      1     34341  610939 3443
## - PAX        1     47122  623720 3453
## - VACATION   1     66011  642609 3467
```

```
## - HI        1     72991  649589 3473
## - SW        1    114211  690809 3502
## - DISTANCE  1    680312 1256910 3788
```

```
## [1] "The acuracy measures of the model is being displayed below:"
```

```
##              ME RMSE   MAE    MPE  MAPE
## Test set 0.0232 37.1 29.21 -5.258 21.13
```

So, based on the final model results obtained from the backward selection process with StepAIC, COUPON is not statistically significant and can be removed from our model. So, our linear.best model can be considered for our prediction purpose.

Let's assume a situation with the following characterstics: 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. Now, let's predict the FARE for the route.

```
## [1] "The avegare fare for a route where Southwest airlines is not serving is:"
```

```
##   NEW
## 242.5
```

With our linear.best model, the value of average Fare for a route where Southwest Airline is not operating in the route (SW = No/0) is $242.5. The value of Fare is high in this case because when there are no low cost airlines operating in that route, the compition is less and the price will be more.

Now, let's predict the FARE for the route when SouthWest Airlines enters the route SW = Yes.

```
## [1] "The avegare fare for a route where Southwest airlines is serving is:"
```

```
##   NEW
## 202.4
```

With our linear.best model, the value of average Fare for a route where Southwest Airline serves the route (SW = Yes/1) is $202.4. The value of Fare is decreasing in this case because Southwest Airlines is a low cost airlines and when it starts a new route, the fare will decrease.

But, from the heatmap / correlation map generated above, we can see that there is a strong positive correlation between COUPON and DISTANCE. Removing it from the model would lead to omitted variable bias and would result in our prediction being underestimated. So, let's include the variable COUPON and it's interaction with DISTANCE in a model and check it's statistical significance and observe it's accuracy measures.

```
linear.reg.interaction <- lm(FARE ~ COUPON + NEW + VACATION + SW + HI + S_INCOME
                             + E_INCOME + S_POP + E_POP + SLOT + GATE
                             + DISTANCE + PAX + (COUPON*DISTANCE) , data = train.df)
summary(linear.reg.interaction)
```

```
##
## Call:
## lm(formula = FARE ~ COUPON + NEW + VACATION + SW + HI + S_INCOME +
##     E_INCOME + S_POP + E_POP + SLOT + GATE + DISTANCE + PAX +
##     (COUPON * DISTANCE), data = train.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -110.59  -22.26    0.61   19.86  124.32
##
## Coefficients:
##                       Estimate     Std. Error t value             Pr(>|t|)
## (Intercept)     -147.058393527   39.316750892   -3.74              0.00021 ***
## COUPON           121.984004920   23.301207149    5.24   0.0000002506151101 ***
## NEW               -4.215170317    1.967090567   -2.14              0.03265 *
## VACATIONYes      -35.697732992    4.327641894   -8.25   0.0000000000000017 ***
## SWYes            -38.644436739    4.054552681   -9.53 < 0.0000000000000002 ***
## HI                 0.008605909    0.001103430    7.80   0.0000000000000416 ***
## S_INCOME           0.001616591    0.000587382    2.75              0.00615 **
## E_INCOME           0.002194761    0.000580253    3.78              0.00018 ***
## S_POP              0.000003966    0.000000731    5.43   0.0000000916901520 ***
## E_POP              0.000003718    0.000000879    4.23   0.0000281264976413 ***
## SLOTFree         -15.202215449    4.362569508   -3.48              0.00054 ***
## GATEFree         -21.924168512    4.346993165   -5.04   0.0000006573225953 ***
## DISTANCE           0.153126939    0.015919412    9.62 < 0.0000000000000002 ***
## PAX               -0.000769146    0.000161888   -4.75   0.0000027035476053 ***
## COUPON:DISTANCE   -0.071318086    0.012908282   -5.52   0.0000000550685952 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.1 on 463 degrees of freedom
## Multiple R-squared:  0.786,  Adjusted R-squared:  0.78
## F-statistic:  122 on 14 and 463 DF,  p-value: <0.0000000000000002
```

```
linear.reg.intr.pred <- predict(linear.reg.interaction, test.df)
print("Accuracy of model with COUPON and it's interaction with DISTANCE")
```

```
## [1] "Accuracy of model with COUPON and it's interaction with DISTANCE"
```

```
accuracy(linear.reg.intr.pred, test.df$FARE)
```

```
##                ME  RMSE   MAE    MPE  MAPE
## Test set -0.01878 38.46 30.87 -4.993 22.36
```

Based on the summary result of the above model, based on the p-value, we can now see that both the variable COUPON and it's interaction with DISTANCE (COUPON:DISTANCE) along with all other variables are statistically significant now at every significance level.

Now, let's predict the airfares based on the variables characterstics mentioned above.

```
## [1] "The avegare fare for a route where Southwest airlines is serving is:"
```

```
## NEW
## 357
```

With our linear model with COUPON and it's interaction with DISTANCE, the value of average Fare for a route where Southwest Airline serves the route (SW = Yes/1) is $357. The value of Fare is decreasing in this case because Southwest Airlines is a low cost airlines and when it starts a new route, the fare will decrease.

```
## [1] "The avegare fare for a route where Southwest airlines is not serving is:"
```

```
##   NEW
## 395.6
```

With our linear model with COUPON and it's interaction with DISTANCE, the value of average Fare for a route where Southwest Airline serves the route (SW = No/0) is $395.6. The value of Fare is increasing in this case because if there are n low cost aitlines for a route, then FARE for that route will be higher because of less compitition.

Based on the above prediction, we can come to a conclusion that whenever a low cost airlines like SouthWest airlines will start it's operation in a new route, the ticket FARE for that route will most likely decrease.