Abir Pattnaik Airline Code.R

COLLEGE:MAHARAJA AGRASEN INSTITUTE OF TECHNOLOGY

PHONE NO.8586896169

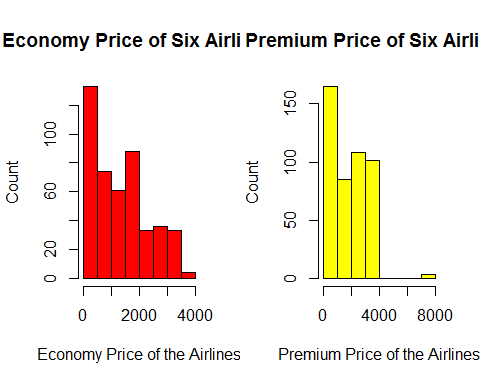
# Analysis of Airline Ticket Pricing  
# NAME: ABIR PATTNAIK  
# EMAIL: abir.pattanaik@gmail.com  
# COLLEGE : Maharaja Agrasen Institute of Technology  
  
# NOTE:The dataset that has been used is SixAirlinesData.csv updated on W4D1  
# PriceRelative=(PricePremium-PriceEconomy)/ (PriceEconomy)  
  
# a)Load the dataset into RStudio and Viewing the dataframe for checking whether the   
# data set uploaded is correct or not.  
#  
# Choose your own location of where you have saved the dataset SixAirlinesData.csv  
  
AirlineAnalysis.df<-read.csv("C:/Users/DRDO HQ/Desktop/DATA ANALYTICS INTERNSHIP/AIRLINE PROJECT/SixAirlinesData.csv")  
View(AirlineAnalysis.df)  
  
# b)To summarize the data with the help of Summary function.  
# A detailed version of the spread can be understood by describe() function  
# 'psych' package was used for accessing the function  
  
summary(AirlineAnalysis.df)

## Airline Aircraft FlightDuration TravelMonth  
## AirFrance: 74 AirBus:151 Min. : 1.250 Aug:128   
## British :175 Boeing:311 1st Qu.: 4.250 Jul: 76   
## Delta : 46 Median : 7.750 Oct:128   
## Jet : 65 Mean : 7.549 Sep:130   
## Singapore: 40 3rd Qu.:10.500   
## Virgin : 62 Max. :14.660   
## IsInternational SeatsEconomy SeatsPremium PitchEconomy   
## Domestic : 40 Min. : 17.0 Min. : 8.00 Min. :30.00   
## International:422 1st Qu.:127.0 1st Qu.:21.00 1st Qu.:31.00   
## Median :185.0 Median :36.00 Median :31.00   
## Mean :200.7 Mean :33.54 Mean :31.21   
## 3rd Qu.:243.0 3rd Qu.:40.00 3rd Qu.:32.00   
## Max. :389.0 Max. :66.00 Max. :33.00   
## PitchPremium WidthEconomy WidthPremium PriceEconomy   
## Min. :34.00 Min. :17.00 Min. :17.00 Min. : 65.0   
## 1st Qu.:38.00 1st Qu.:17.00 1st Qu.:19.00 1st Qu.: 404.8   
## Median :38.00 Median :18.00 Median :19.00 Median :1224.0   
## Mean :37.92 Mean :17.83 Mean :19.48 Mean :1317.1   
## 3rd Qu.:38.00 3rd Qu.:18.00 3rd Qu.:21.00 3rd Qu.:1903.0   
## Max. :40.00 Max. :19.00 Max. :21.00 Max. :3593.0   
## PricePremium PriceRelative SeatsTotal FractionPremiumSeats  
## Min. : 86 Min. :0.0200 Min. : 38.0 Min. :0.0500   
## 1st Qu.: 524 1st Qu.:0.1000 1st Qu.:162.0 1st Qu.:0.1200   
## Median :1710 Median :0.3800 Median :227.0 Median :0.1300   
## Mean :1832 Mean :0.4926 Mean :234.2 Mean :0.1503   
## 3rd Qu.:2989 3rd Qu.:0.7475 3rd Qu.:279.0 3rd Qu.:0.1500   
## Max. :7414 Max. :1.8900 Max. :441.0 Max. :0.5500   
## PitchDifference WidthDifference  
## Min. : 2.000 Min. :0.000   
## 1st Qu.: 6.000 1st Qu.:1.000   
## Median : 7.000 Median :1.000   
## Mean : 6.716 Mean :1.654   
## 3rd Qu.: 7.000 3rd Qu.:3.000   
## Max. :10.000 Max. :4.000

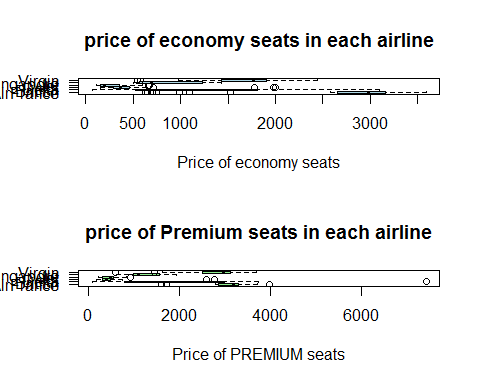
library(psych)  
describe(AirlineAnalysis.df)

## vars n mean sd median trimmed mad  
## Airline\* 1 462 3.02 1.65 2.00 2.90 1.48  
## Aircraft\* 2 462 1.67 0.47 2.00 1.72 0.00  
## FlightDuration 3 462 7.55 3.54 7.75 7.54 4.82  
## TravelMonth\* 4 462 2.56 1.17 3.00 2.58 1.48  
## IsInternational\* 5 462 1.91 0.28 2.00 2.00 0.00  
## SeatsEconomy 6 462 200.71 77.96 185.00 193.76 85.99  
## SeatsPremium 7 462 33.54 13.26 36.00 33.20 11.86  
## PitchEconomy 8 462 31.21 0.66 31.00 31.25 0.00  
## PitchPremium 9 462 37.92 1.32 38.00 38.06 0.00  
## WidthEconomy 10 462 17.83 0.56 18.00 17.81 0.00  
## WidthPremium 11 462 19.48 1.10 19.00 19.54 0.00  
## PriceEconomy 12 462 1317.06 989.81 1224.00 1231.30 1163.84  
## PricePremium 13 462 1832.35 1289.97 1710.00 1782.94 1852.51  
## PriceRelative 14 462 0.49 0.45 0.38 0.43 0.42  
## SeatsTotal 15 462 234.25 86.88 227.00 227.69 90.44  
## FractionPremiumSeats 16 462 0.15 0.06 0.13 0.14 0.03  
## PitchDifference 17 462 6.72 1.78 7.00 6.79 0.00  
## WidthDifference 18 462 1.65 1.20 1.00 1.55 0.00  
## min max range skew kurtosis se  
## Airline\* 1.00 6.00 5.00 0.59 -0.95 0.08  
## Aircraft\* 1.00 2.00 1.00 -0.74 -1.46 0.02  
## FlightDuration 1.25 14.66 13.41 -0.05 -1.12 0.16  
## TravelMonth\* 1.00 4.00 3.00 -0.14 -1.46 0.05  
## IsInternational\* 1.00 2.00 1.00 -2.93 6.60 0.01  
## SeatsEconomy 17.00 389.00 372.00 0.61 -0.26 3.63  
## SeatsPremium 8.00 66.00 58.00 0.25 -0.46 0.62  
## PitchEconomy 30.00 33.00 3.00 -0.03 -0.38 0.03  
## PitchPremium 34.00 40.00 6.00 -1.48 3.43 0.06  
## WidthEconomy 17.00 19.00 2.00 -0.03 -0.12 0.03  
## WidthPremium 17.00 21.00 4.00 -0.09 -0.34 0.05  
## PriceEconomy 65.00 3593.00 3528.00 0.52 -0.88 46.05  
## PricePremium 86.00 7414.00 7328.00 0.51 0.41 60.01  
## PriceRelative 0.02 1.89 1.87 1.14 0.61 0.02  
## SeatsTotal 38.00 441.00 403.00 0.61 -0.44 4.04  
## FractionPremiumSeats 0.05 0.55 0.50 2.70 14.02 0.00  
## PitchDifference 2.00 10.00 8.00 -0.51 1.67 0.08  
## WidthDifference 0.00 4.00 4.00 0.82 -0.60 0.06

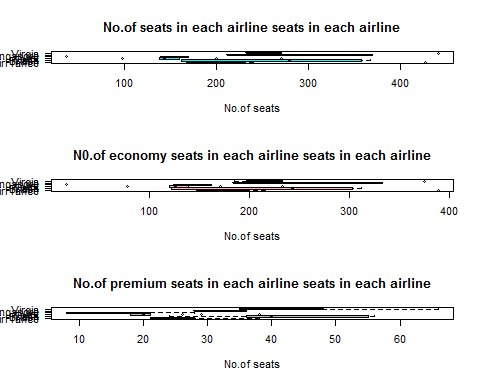
# c)Visualising each variable independently and understanding the spread of each data  
# Histogram was used to depict the distribution of PriceEconomy and PricePremium  
  
attach(AirlineAnalysis.df)  
par(mfrow=c(1,2))  
hist(PriceEconomy,xlab="Economy Price of the Airlines",ylab="Count",main="Economy Price of Six Airlines",col="red")  
hist(PricePremium,xlab="Premium Price of the Airlines",ylab="Count",main="Premium Price of Six Airlines",col="yellow")



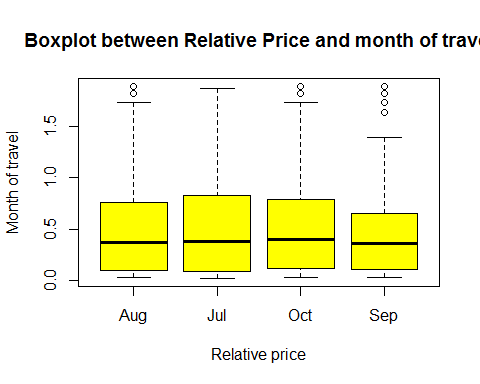
# c1)Boxplots to see the price of economy seats and premium seats in each airline  
  
par(mfrow=c(2,1))  
boxplot(PriceEconomy~Airline,data=AirlineAnalysis.df,horizontal=TRUE,xlab="Price of economy seats",ytab="Airline",main="price of economy seats in each airline",las=1,col="lightblue")  
boxplot(PricePremium~Airline,data=AirlineAnalysis.df,horizontal=TRUE,xlab="Price of PREMIUM seats",ytab="Airline",main="price of Premium seats in each airline",las=1,col="lightgreen")



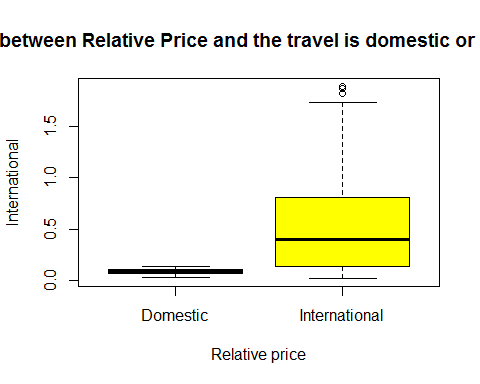
# c2)Boxplots to see the no of total,economy and premium seats in each airline  
  
par(mfrow=c(3,1))  
boxplot(SeatsTotal~Airline,data=AirlineAnalysis.df,horizontal=TRUE,xlab="No.of seats",ytab="Airline",main="No.of seats in each airline seats in each airline",las=1,col="cyan")  
boxplot(SeatsEconomy~Airline,data=AirlineAnalysis.df,horizontal=TRUE,xlab="No.of seats",ytab="Airline",main="N0.of economy seats in each airline seats in each airline",las=1,col="lightpink")  
boxplot(SeatsPremium~Airline,data=AirlineAnalysis.df,horizontal=TRUE,xlab="No.of seats",ytab="Airline",main="No.of premium seats in each airline seats in each airline",las=1,col="grey")



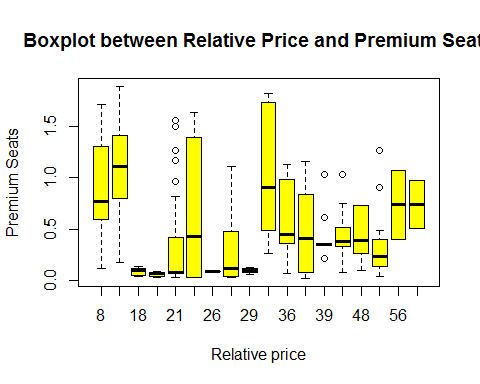
# c3)Visualising other variables to the PriceRelative column  
  
par(mfrow=c(1,1))  
boxplot(PriceRelative~TravelMonth,xlab="Relative price",ylab="Month of travel",main="Boxplot between Relative Price and month of travel",col="yellow")



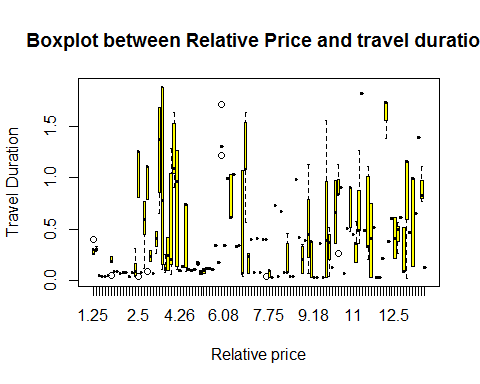
boxplot(PriceRelative~IsInternational,xlab="Relative price",ylab="International",main="Boxplot between Relative Price and the travel is domestic or international",col="yellow")



boxplot(PriceRelative~SeatsPremium,xlab="Relative price",ylab="Premium Seats",main="Boxplot between Relative Price and Premium Seats",col="yellow")



boxplot(PriceRelative~FlightDuration,xlab="Relative price",ylab="Travel Duration",main="Boxplot between Relative Price and travel duration",col="yellow")



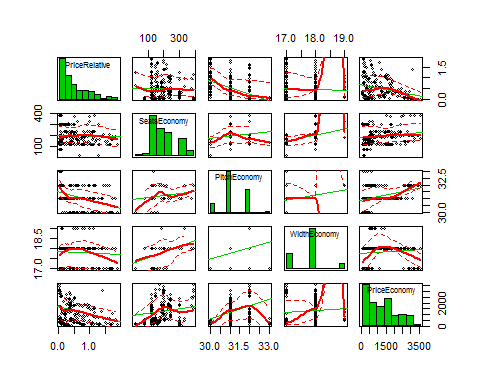
# d)Finding relations between the variables through the scatterplotMatrix.It was suggested that the PriceRelative depended on some variables.The high predictor variables were chosen in this  
# 'car' package was called  
  
library(car)

##   
## Attaching package: 'car'

## The following object is masked from 'package:psych':  
##   
## logit

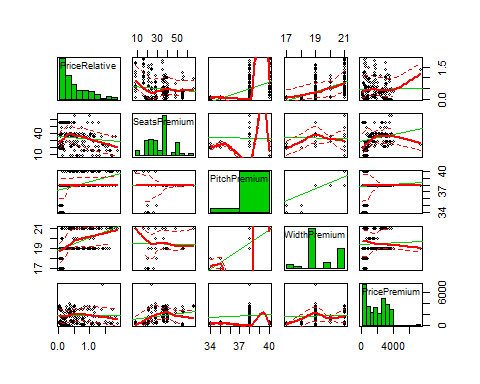
scatterplotMatrix(formula = ~PriceRelative+SeatsEconomy+PitchEconomy+WidthEconomy+PriceEconomy, cex=0.6,data=AirlineAnalysis.df, diagonal="histogram")

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,  
## spread = spread, : could not fit smooth



scatterplotMatrix(formula = ~PriceRelative+SeatsPremium+PitchPremium+WidthPremium+PricePremium, cex=0.6,data=AirlineAnalysis.df, diagonal="histogram")

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,  
## spread = spread, : could not fit smooth



# e) Attempt to to find correlation between the each Premium and Economy  
# Calculating correlations between Prices of Economy and Premium in correlation to other factors  
cor.test(PriceEconomy, PitchEconomy)

##   
## Pearson's product-moment correlation  
##   
## data: PriceEconomy and PitchEconomy  
## t = 8.8003, df = 460, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2987210 0.4550742  
## sample estimates:  
## cor   
## 0.379605

cor.test(PriceEconomy, WidthEconomy)

##   
## Pearson's product-moment correlation  
##   
## data: PriceEconomy and WidthEconomy  
## t = 1.764, df = 460, p-value = 0.0784  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.009330795 0.171911298  
## sample estimates:  
## cor   
## 0.0819679

cor.test(PricePremium, PitchPremium)

##   
## Pearson's product-moment correlation  
##   
## data: PricePremium and PitchPremium  
## t = 1.5338, df = 460, p-value = 0.1258  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.02002801 0.16150915  
## sample estimates:  
## cor   
## 0.07133125

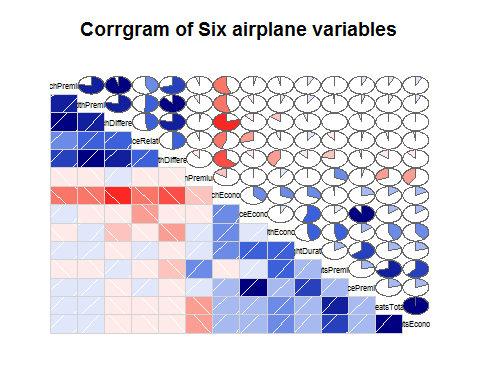
cor.test(PricePremium, WidthPremium)

##   
## Pearson's product-moment correlation  
##   
## data: PricePremium and WidthPremium  
## t = 1.0592, df = 460, p-value = 0.2901  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04209336 0.13992426  
## sample estimates:  
## cor   
## 0.04932498

library(corrgram)

## Warning: package 'corrgram' was built under R version 3.4.1

corrgram(AirlineAnalysis.df, main = "Corrgram of Six airplane variables", lower.panel = panel.shade, upper.panel = panel.pie, text.panel = panel.txt,order=TRUE)



# f)Creating a regression model -In this following section 2 models were created one that is a full linear model and second the model created is leaving out of variables  
# that do not create significant difference in the model.  
# 'QuantPsyc' package contains the lm.beta that involves in the making standardized coefficients   
# lm.fit is the full model containing all the variables that affect the PriceRelative  
# lm.fit5 is the reduced model containing the variables that create or give significant effect on the variable   
# anova() involves the calculating the variance from the actual model  
# 'visreg' package was used to create regression lines on given linear model  
  
lm.fit=lm(PriceRelative~., data=AirlineAnalysis.df)  
summary(lm.fit)

##   
## Call:  
## lm(formula = PriceRelative ~ ., data = AirlineAnalysis.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.77045 -0.08569 0.00114 0.08348 0.84939   
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.023e-01 2.956e+00 -0.102 0.918582   
## AirlineBritish -4.033e-01 1.109e-01 -3.637 0.000308 \*\*\*  
## AirlineDelta -4.182e-01 2.196e-01 -1.904 0.057539 .   
## AirlineJet -2.665e-01 9.587e-02 -2.780 0.005675 \*\*   
## AirlineSingapore -3.754e-01 1.282e-01 -2.929 0.003573 \*\*   
## AirlineVirgin -3.917e-01 2.017e-01 -1.942 0.052781 .   
## AircraftBoeing 4.495e-02 2.923e-02 1.538 0.124737   
## FlightDuration 2.629e-02 4.737e-03 5.552 4.9e-08 \*\*\*  
## TravelMonthJul 1.355e-02 3.136e-02 0.432 0.665938   
## TravelMonthOct 2.797e-02 2.666e-02 1.049 0.294678   
## TravelMonthSep -6.260e-03 2.660e-02 -0.235 0.814079   
## IsInternationalInternational 1.139e-02 2.507e-01 0.045 0.963770   
## SeatsEconomy 3.988e-04 3.175e-04 1.256 0.209799   
## SeatsPremium -4.732e-03 2.290e-03 -2.066 0.039421 \*   
## PitchEconomy -1.755e-02 8.018e-02 -0.219 0.826798   
## PitchPremium 4.932e-02 9.168e-02 0.538 0.590849   
## WidthEconomy -8.244e-02 5.129e-02 -1.607 0.108680   
## WidthPremium 6.097e-02 1.367e-01 0.446 0.655894   
## PriceEconomy -9.311e-04 3.323e-05 -28.022 < 2e-16 \*\*\*  
## PricePremium 5.777e-04 2.301e-05 25.111 < 2e-16 \*\*\*  
## SeatsTotal NA NA NA NA   
## FractionPremiumSeats 4.682e-01 3.031e-01 1.544 0.123185   
## PitchDifference NA NA NA NA   
## WidthDifference NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2129 on 441 degrees of freedom  
## Multiple R-squared: 0.7889, Adjusted R-squared: 0.7793   
## F-statistic: 82.39 on 20 and 441 DF, p-value: < 2.2e-16

library(QuantPsyc)

## Warning: package 'QuantPsyc' was built under R version 3.4.1

## Loading required package: boot

##   
## Attaching package: 'boot'

## The following object is masked from 'package:car':  
##   
## logit

## The following object is masked from 'package:psych':  
##   
## logit

## Loading required package: MASS

##   
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':  
##   
## norm

lm.beta(lm.fit)

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x) on a factor x is deprecated and will become an error.  
## Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x) on a factor x is deprecated and will become an error.  
## Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.  
  
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x) on a factor x is deprecated and will become an error.  
## Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.  
  
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling var(x) on a factor x is deprecated and will become an error.  
## Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.

## Warning in b \* sx: longer object length is not a multiple of shorter object  
## length

## AirlineBritish AirlineDelta   
## -1.468053e+00 -4.333722e-01   
## AirlineJet AirlineSingapore   
## -2.082164e+00 -9.684371e-01   
## AirlineVirgin AircraftBoeing   
## -2.433554e-01 7.734595e+00   
## FlightDuration TravelMonthJul   
## 7.693098e-01 1.979537e-02   
## TravelMonthOct TravelMonthSep   
## 8.165672e-02 -7.744472e-03   
## IsInternationalInternational SeatsEconomy   
## 2.769639e-02 8.711879e-01   
## SeatsPremium PitchEconomy   
## -1.347210e+01 -3.366211e+00   
## PitchPremium WidthEconomy   
## 6.658400e-03 -3.239986e-01   
## WidthPremium PriceEconomy   
## 1.620451e-01 -3.389298e-03   
## PricePremium FractionPremiumSeats   
## 5.987167e-04 3.657921e+00

lm.fit5<-lm(PriceRelative~.-Airline-TravelMonth-Aircraft-IsInternational-SeatsTotal-PitchDifference-WidthDifference-FractionPremiumSeats-SeatsEconomy,data=AirlineAnalysis.df)  
summary(lm.fit5)

##   
## Call:  
## lm(formula = PriceRelative ~ . - Airline - TravelMonth - Aircraft -   
## IsInternational - SeatsTotal - PitchDifference - WidthDifference -   
## FractionPremiumSeats - SeatsEconomy, data = AirlineAnalysis.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.77218 -0.10792 0.00009 0.07971 0.83920   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.366e+00 1.033e+00 -5.197 3.07e-07 \*\*\*  
## FlightDuration 2.252e-02 4.543e-03 4.957 1.01e-06 \*\*\*  
## SeatsPremium -5.245e-03 9.698e-04 -5.409 1.03e-07 \*\*\*  
## PitchEconomy 1.141e-01 2.639e-02 4.324 1.88e-05 \*\*\*  
## PitchPremium 1.192e-01 1.334e-02 8.935 < 2e-16 \*\*\*  
## WidthEconomy -1.775e-01 2.726e-02 -6.512 1.97e-10 \*\*\*  
## WidthPremium 5.402e-02 1.576e-02 3.428 0.000663 \*\*\*  
## PriceEconomy -8.369e-04 2.928e-05 -28.584 < 2e-16 \*\*\*  
## PricePremium 5.451e-04 2.275e-05 23.961 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2217 on 453 degrees of freedom  
## Multiple R-squared: 0.7647, Adjusted R-squared: 0.7606   
## F-statistic: 184.1 on 8 and 453 DF, p-value: < 2.2e-16

anova(lm.fit5,lm.fit)

## Analysis of Variance Table  
##   
## Model 1: PriceRelative ~ (Airline + Aircraft + FlightDuration + TravelMonth +   
## IsInternational + SeatsEconomy + SeatsPremium + PitchEconomy +   
## PitchPremium + WidthEconomy + WidthPremium + PriceEconomy +   
## PricePremium + SeatsTotal + FractionPremiumSeats + PitchDifference +   
## WidthDifference) - Airline - TravelMonth - Aircraft - IsInternational -   
## SeatsTotal - PitchDifference - WidthDifference - FractionPremiumSeats -   
## SeatsEconomy  
## Model 2: PriceRelative ~ Airline + Aircraft + FlightDuration + TravelMonth +   
## IsInternational + SeatsEconomy + SeatsPremium + PitchEconomy +   
## PitchPremium + WidthEconomy + WidthPremium + PriceEconomy +   
## PricePremium + SeatsTotal + FractionPremiumSeats + PitchDifference +   
## WidthDifference  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 453 22.264   
## 2 441 19.981 12 2.2831 4.1992 2.829e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Result:Keeping all the variables it is shown that the lm.fit model has R^2 of 0.78 which is quite good but it does  
# contain multilinearity and other overlapping variables.Hence,for calculation of the most significant oncs  
# lm.fit5 model was created.This model had R^2 of 0.76 and Adjusted R^2 of 0.7606.  
# On applying the anova analysis in these 2 models the F-statistic came out to be 4.1992 and P<0.05  
# which implies that the variables that were removed were less significant and had no impact on tha analysis  
# visreg function plots function of each variable.

library(visreg)

par(mfrow=c(4,5))

visreg(lm.fit5)

