veera mukesh aripaka -m6 - U97302307

2022-10-18

#veera mukesh aripaka -m6 - U97302307  
#Preprocessing:  
  
#⦁ Load the file “6304 Module 6 Assignment Data.xlsx” data set into R. This data shows airfares and passengers for certain U.S. Domestic Routes for the 4th quarter of 2002. This is not an exhaustive list of all flights.   
  
library(readxl)  
M6 <- read\_excel("C:/Users/mukes/Downloads/6304 Module 6 Assignment Data (1).xlsx")  
  
#⦁ Create a random selection of flights of n=50. Be certain to include in your sample only the origin airports of LAS, LAX, BWI, LGA, MCI, MCO, ATL, and BNA. Make sure to convert any character (chr) variables to factor variables. This will be your primary data set for analysis.  
M6$Origin = as.factor(M6$Origin)  
M6$Destination = as.factor(M6$Destination)  
M6$`market leading airline`= as.factor(M6$`market leading airline`)  
M6$`Low price airline`=as.factor(M6$`Low price airline`)  
str(M6)

## tibble [1,000 × 9] (S3: tbl\_df/tbl/data.frame)  
## $ Origin : Factor w/ 90 levels "ABQ","ACY","ALB",..: 17 17 3 3 3 3 3 3 3 3 ...  
## $ Destination : Factor w/ 85 levels "ATL","AUS","BDL",..: 1 41 1 8 53 19 33 34 41 83 ...  
## $ Average Fare : num [1:1000] 114.5 122.5 214.4 69.4 158.1 ...  
## $ Distance : num [1:1000] 528 860 852 288 723 ...  
## $ Avg weekly passengers : num [1:1000] 425 277 216 607 313 ...  
## $ market leading airline: Factor w/ 16 levels "AA","AS","B6",..: 7 7 5 15 13 15 15 5 15 14 ...  
## $ route market share : num [1:1000] 70.2 75.1 78.9 97 39.8 ...  
## $ Low price airline : Factor w/ 19 levels "AA","AQ","AS",..: 9 7 6 18 18 7 18 17 18 7 ...  
## $ price : num [1:1000] 111 118.9 167.1 68.9 145.4 ...

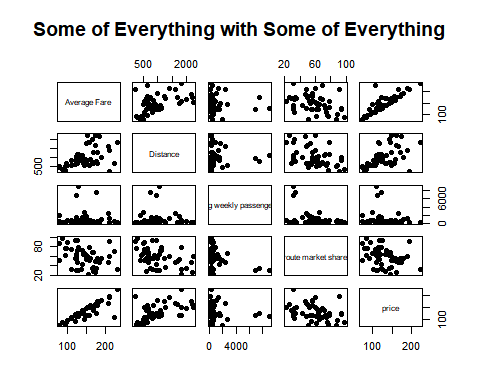
dss = subset(M6,M6$Destination=='LAS'|M6$Destination=='LAX'|M6$Destination=='BWI'|M6$Destination=='LGA'|M6$Destination=='MCI'|M6$Destination=='MCO'|M6$Destination=='ATL'|M6$Destination=='BNA')  
  
set.seed(97302307)  
sam=dss[sample(1:nrow(dss),50),replace=FALSE]  
  
#Analysis:  
# Show the results of an str() command.  
  
str(sam)

## tibble [50 × 9] (S3: tbl\_df/tbl/data.frame)  
## $ Origin : Factor w/ 90 levels "ABQ","ACY","ALB",..: 52 29 40 67 6 64 19 36 27 26 ...  
## $ Destination : Factor w/ 85 levels "ATL","AUS","BDL",..: 41 40 34 36 34 36 40 41 41 36 ...  
## $ Average Fare : num [1:50] 164.9 168.2 212.4 85.7 178.6 ...  
## $ Distance : num [1:50] 683 629 1397 288 1238 ...  
## $ Avg weekly passengers : num [1:50] 281 412 1292 271 508 ...  
## $ market leading airline: Factor w/ 16 levels "AA","AS","B6",..: 11 11 4 7 1 1 4 5 1 13 ...  
## $ route market share : num [1:50] 68.8 67.1 60.1 98.2 56 ...  
## $ Low price airline : Factor w/ 19 levels "AA","AQ","AS",..: 9 18 18 9 18 15 18 7 7 8 ...  
## $ price : num [1:50] 145.4 135.6 188.3 85.6 172 ...

# Show the results of a table() command on the origin variable  
  
table(sam$Origin)

##   
## ABQ ACY ALB AMA ATL AUS BDL BHM BNA BOI BOS BTV BUF BUR BWI CAE CAK CHS CLE CLT   
## 2 0 1 0 1 2 1 0 0 1 1 1 1 0 1 0 0 1 3 0   
## CMH COS CRP CVG DAY DEN DFW DSM DTW ELP EUG FLL FNT GPT GRR GSO GSP HRL IAD IAH   
## 2 1 0 0 0 2 2 0 4 0 0 1 1 0 1 1 0 0 0 3   
## IND ISP JAX LAS LAX LGA LGB LIT MCI MCO MDT MEM MHT MIA MKE MSN MSP MSY MYR OAK   
## 2 2 1 0 0 1 0 0 2 0 0 1 1 0 0 1 0 1 0 0   
## OKC OMA ONT ORD ORF PDX PHF PHL PHX PIT PSP PVD RDU RNO ROC RSW SAN SAT SDF SEA   
## 0 0 0 2 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0   
## SFO SJC SLC SMF SNA STL SYR TLH TPA TYS   
## 0 0 0 0 0 0 0 0 0 0

#⦁ Show a scatterplot matrix of the continuous variables only. From this matrix which pair of variables do you believe would have the strongest linear relationship? How did you arrive at this conclusion?   
  
csm = subset(sam,select=c(3,4,5,7,9))  
  
plot(csm,pch=19,main="Some of Everything with Some of Everything")



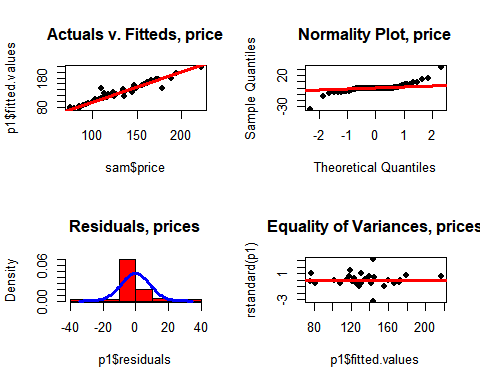
cor(csm)

## Average Fare Distance Avg weekly passengers  
## Average Fare 1.00000000 0.44827677 -0.02301651  
## Distance 0.44827677 1.00000000 -0.01957975  
## Avg weekly passengers -0.02301651 -0.01957975 1.00000000  
## route market share -0.32670046 -0.42663169 -0.34341110  
## price 0.85906686 0.51849990 -0.12172625  
## route market share price  
## Average Fare -0.3267005 0.8590669  
## Distance -0.4266317 0.5184999  
## Avg weekly passengers -0.3434111 -0.1217263  
## route market share 1.0000000 -0.3137368  
## price -0.3137368 1.0000000

#The linear relationship between price variable and avergae fare is the strongest with a straight line graph between them and also the cor command shows the 0.8590 correlation between them.  
  
#⦁ Parameterize a full regression model with y=price. Include all other continuous variables as well as the origin variable. Show the R summary of this model.   
  
p1 =lm(price~`Average Fare`+Distance+`Avg weekly passengers`+`route market share`+Origin,data=sam)  
  
summary(p1)

##   
## Call:  
## lm(formula = price ~ `Average Fare` + Distance + `Avg weekly passengers` +   
## `route market share` + Origin, data = sam)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -34.640 -1.383 0.000 0.778 34.640   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 139.074001 46.058042 3.020 0.0107 \*  
## `Average Fare` 0.476392 0.196622 2.423 0.0321 \*  
## Distance 0.003161 0.010575 0.299 0.7701   
## `Avg weekly passengers` -0.003794 0.004725 -0.803 0.4377   
## `route market share` -1.204358 0.463921 -2.596 0.0234 \*  
## OriginALB 2.717840 26.475619 0.103 0.9199   
## OriginATL 65.233953 34.267177 1.904 0.0812 .  
## OriginAUS 14.745306 20.862288 0.707 0.4932   
## OriginBDL 2.214831 27.470772 0.081 0.9371   
## OriginBOI -32.090887 25.049096 -1.281 0.2244   
## OriginBOS -29.685624 22.821094 -1.301 0.2178   
## OriginBTV -11.455100 29.548279 -0.388 0.7050   
## OriginBUF -35.773480 24.505293 -1.460 0.1700   
## OriginBWI -10.675039 27.301595 -0.391 0.7026   
## OriginCHS -8.187669 23.752897 -0.345 0.7363   
## OriginCLE -7.031633 20.500825 -0.343 0.7375   
## OriginCMH -43.780670 22.124566 -1.979 0.0712 .  
## OriginCOS -38.916998 25.412355 -1.531 0.1516   
## OriginDEN -6.337660 19.938836 -0.318 0.7561   
## OriginDFW -18.536413 22.178374 -0.836 0.4196   
## OriginDTW -19.776820 20.390109 -0.970 0.3512   
## OriginFLL -20.228616 53.186399 -0.380 0.7103   
## OriginFNT -17.138461 25.339948 -0.676 0.5116   
## OriginGRR -51.448462 25.348816 -2.030 0.0652 .  
## OriginGSO 4.119861 27.883329 0.148 0.8850   
## OriginIAH 11.708355 19.158864 0.611 0.5525   
## OriginIND -34.069635 22.262357 -1.530 0.1519   
## OriginISP 35.490548 29.393428 1.207 0.2505   
## OriginJAX -3.602647 22.802289 -0.158 0.8771   
## OriginLGA -23.653601 45.132096 -0.524 0.6098   
## OriginMCI 10.565017 27.881055 0.379 0.7114   
## OriginMEM 9.512763 25.313756 0.376 0.7136   
## OriginMHT 3.363728 27.503117 0.122 0.9047   
## OriginMSN -41.885199 25.658399 -1.632 0.1285   
## OriginMSY 15.742624 31.918053 0.493 0.6308   
## OriginORD -32.538931 34.143065 -0.953 0.3594   
## OriginORF -5.207398 29.294113 -0.178 0.8619   
## OriginPHF 24.091028 34.372060 0.701 0.4967   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.64 on 12 degrees of freedom  
## Multiple R-squared: 0.9265, Adjusted R-squared: 0.6999   
## F-statistic: 4.088 on 37 and 12 DF, p-value: 0.005983

#⦁ Drawing on Step 4, give a verbal interpretation of the impact of the levels of the origin variable on price in your model.   
  
#As seen in the summary output the pvalues of the different origins there are only two origins ATL,CMH,GRR are only significant because the p values are nearly equal to 0.05 and the beta coefficients of these origins also a significant impact on deciding the price. ATL beta = 65.233 which means if the the origin is ATL then the price increase by 65.233, similarly if CMH then price decreases by 43.78 and if its GRR then decreases by 51.  
  
#⦁ Drawing on Step 4, determine whether your model meets the LINE assumptions of regression.  
  
par(mfrow=c(2,2))  
# Linearity  
plot(sam$price,p1$fitted.values,  
 pch=19,main="Actuals v. Fitteds, price")  
abline(0,1,col="red",lwd=3)  
  
#most of the points lie on the abline which means the fitted values and prices are the same so linearity can be assumed.  
  
# Normality  
qqnorm(p1$residuals,pch=19,  
 main="Normality Plot, price")  
qqline(p1$residuals,lwd=3,col="red")  
hist(p1$residuals,col="red",  
 main="Residuals, prices",  
 probability=TRUE)  
curve(dnorm(x,mean(p1$residuals),  
 sd(p1$residuals)),  
 from=min(p1$residuals),  
 to=max(p1$residuals),  
 lwd=3,col="blue",add=TRUE)  
  
# the histogram shows a normal distribution curve which is fine but the QQ plot shows a pattern with tails not aligning with the abline but most points are on the abline so this can be assumed.  
# Equality of Variances  
plot(p1$fitted.values,rstandard(p1),  
 pch=19,main="Equality of Variances, prices")  
abline(0,0,lwd=3,col="red")



#this is not satisfied as there are leverage points apart from this other data points also almost lie in the same range of sd's  
  
#⦁ Drawing on Step 4, report in a single vector the origin and destination airports and original price of the flight for which actual price deviates most from your model’s regression line.  
  
  
boxplot(sam$price,col="red",ylim=c(0,300),pch=19)  
  
leverages=hat(model.matrix(p1))  
  
sam[leverages>(3\*mean(leverages)),1]

## # A tibble: 0 × 1  
## # … with 1 variable: Origin <fct>

sam[which.max(sam$price),c(1,2,9)]

## # A tibble: 1 × 3  
## Origin Destination price  
## <fct> <fct> <dbl>  
## 1 ABQ LGA 222.

#there are outliers  
  
#⦁ Drawing on Steps 4 and 7, report in a single vector the origin and destination airports and original price of the flight for which actual price deviates least from your model’s regression line.  
  
sam[which.leverages=(mean(leverages)),c(1,2,9)]

## # A tibble: 3 × 9  
## Origin Destination Average Far…¹ Dista…² Avg w…³ marke…⁴ route…⁵ Low p…⁶ price  
## <fct> <fct> <dbl> <dbl> <dbl> <fct> <dbl> <fct> <dbl>  
## 1 MEM MCO 165. 683 281. NW 68.8 FL 145.  
## 2 DTW MCI 168. 629 412. NW 67.1 WN 136.  
## 3 DFW MCO 190. 984 1010. AA 55.4 DL 178.  
## # … with abbreviated variable names ¹​`Average Fare`, ²​Distance,  
## # ³​`Avg weekly passengers`, ⁴​`market leading airline`, ⁵​`route market share`,  
## # ⁶​`Low price airline`

