Question-2.R

kess

2022-10-07

examdata = read.csv('data.csv', header = T)  
library(ggplot2)  
library(afex)

## Loading required package: lme4

## Loading required package: Matrix

## \*\*\*\*\*\*\*\*\*\*\*\*  
## Welcome to afex. For support visit: http://afex.singmann.science/

## - Functions for ANOVAs: aov\_car(), aov\_ez(), and aov\_4()  
## - Methods for calculating p-values with mixed(): 'S', 'KR', 'LRT', and 'PB'  
## - 'afex\_aov' and 'mixed' objects can be passed to emmeans() for follow-up tests  
## - NEWS: emmeans() for ANOVA models now uses model = 'multivariate' as default.  
## - Get and set global package options with: afex\_options()  
## - Set orthogonal sum-to-zero contrasts globally: set\_sum\_contrasts()  
## - For example analyses see: browseVignettes("afex")  
## \*\*\*\*\*\*\*\*\*\*\*\*

##   
## Attaching package: 'afex'

## The following object is masked from 'package:lme4':  
##   
## lmer

library(emmeans)  
library(dbplyr)  
library(car)

## Loading required package: carData

library(corrplot)

## corrplot 0.92 loaded

summary(examdata)

## cust.id in\_relationship health\_conscious nut\_free   
## Min. : 1.0 Length:500 Length:500 Length:500   
## 1st Qu.:125.8 Class :character Class :character Class :character   
## Median :250.5 Mode :character Mode :character Mode :character   
## Mean :250.5   
## 3rd Qu.:375.2   
## Max. :500.0   
## fruit\_free product\_satisfaction age annual\_income   
## Length:500 Min. : 3.00 Min. :20.00 Min. : 62.00   
## Class :character 1st Qu.:12.00 1st Qu.:40.00 1st Qu.: 83.00   
## Mode :character Median :12.00 Median :50.00 Median : 90.50   
## Mean :12.79 Mean :49.82 Mean : 90.13   
## 3rd Qu.:15.00 3rd Qu.:60.00 3rd Qu.: 98.00   
## Max. :21.00 Max. :90.00 Max. :123.00   
## conscientiousness\_personality log\_relationship\_length  
## Min. :18.00 Min. :0.0000   
## 1st Qu.:19.00 1st Qu.:0.0000   
## Median :20.00 Median :0.0000   
## Mean :20.06 Mean :0.2829   
## 3rd Qu.:21.00 3rd Qu.:0.9031   
## Max. :23.00 Max. :1.3010

str(examdata)

## 'data.frame': 500 obs. of 10 variables:  
## $ cust.id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ in\_relationship : chr "no" "no" "yes" "no" ...  
## $ health\_conscious : chr "medium" "medium" "medium" "medium" ...  
## $ nut\_free : chr "no" "no" "no" "yes" ...  
## $ fruit\_free : chr "no" "yes" "no" "no" ...  
## $ product\_satisfaction : int 15 21 15 9 15 15 12 9 12 15 ...  
## $ age : int 40 60 40 50 40 40 50 40 50 50 ...  
## $ annual\_income : int 81 77 81 83 108 100 100 85 91 95 ...  
## $ conscientiousness\_personality: int 20 21 20 20 21 21 21 18 20 20 ...  
## $ log\_relationship\_length : num 0 0 1.04 0 0 ...

table(examdata$in\_relationship)

##   
## no yes   
## 358 142

table(examdata$health\_conscious)

##   
## high low medium   
## 63 40 397

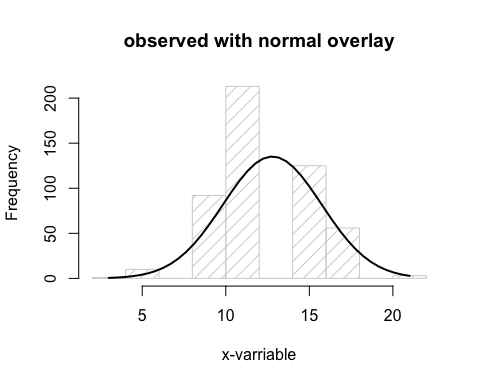
table(examdata$nut\_free)

##   
## no yes   
## 347 153

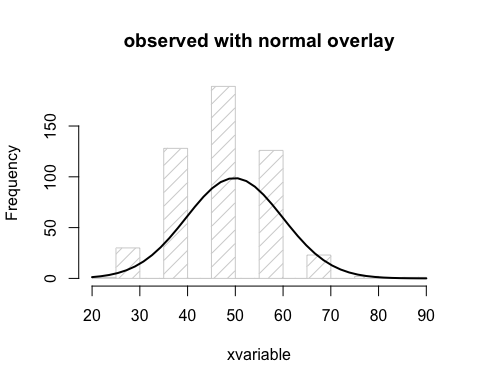
table(examdata$fruit\_free)

##   
## no yes   
## 186 314

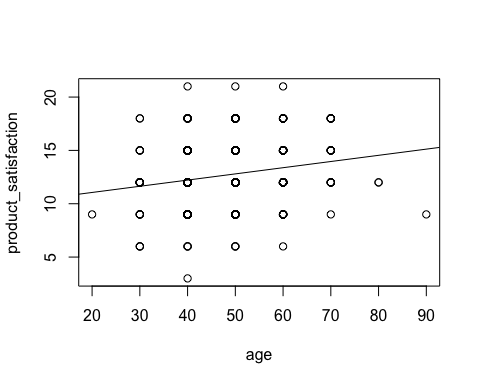
# Numeric variable of each categorical variable  
examdata$in\_relationship = factor(examdata$in\_relationship, levels=c("yes", "no"), labels=c("yes", "no"))  
examdata$num\_in\_relationship = factor(examdata$in\_relationship, levels=c("yes", "no"), labels=c(1,0))  
examdata$num\_in\_relationship = as.numeric(as.character(examdata$num\_in\_relationship))  
examdata$nut\_free =factor(examdata$nut\_free, levels=c("yes", "no"), labels=c("yes","no"))  
examdata$num\_nut\_free= factor(examdata$nut\_free, levels=c("yes", "no"), labels=c(1,0))  
examdata$num\_nut\_free <- as.numeric(as.character(examdata$num\_nut\_free))  
examdata$fruit\_free =factor(examdata$fruit\_free, levels=c("yes", "no"), labels=c("yes","no"))  
examdata$num\_fruit\_free= factor(examdata$fruit\_free, levels=c("yes", "no"), labels=c(1,0))  
examdata$num\_fruit\_free <- as.numeric(as.character(examdata$num\_fruit\_free))  
examdata$health\_conscious = factor(examdata$health\_conscious, levels=c("high", "low", "medium"), labels=c("high", "low", "medium"))  
examdata$in\_relationship = factor(examdata$in\_relationship, levels=c("yes", "no"), labels=c("yes", "no"))  
# The client wants to understand what variables are related to product satisfaction. Which variables are related to product satisfaction and which are not?  
# The outcome variable is product satisfaction. I examine the distribution of satisfaction scores to see if there are any outliers:  
g = examdata$product\_satisfaction  
h = hist(g, breaks = 10, density =10,  
 col = "lightgray", xlab = "x-varriable",main = "observed with normal overlay")  
xfit = seq(min(g), max (g), length = 40)  
yfit = dnorm(xfit, mean = mean(g), sd = sd(g))  
yfit = yfit \* diff(h$mids[1:2]) \* length(g)  
  
lines(xfit, yfit, col = "black", lwd =2)



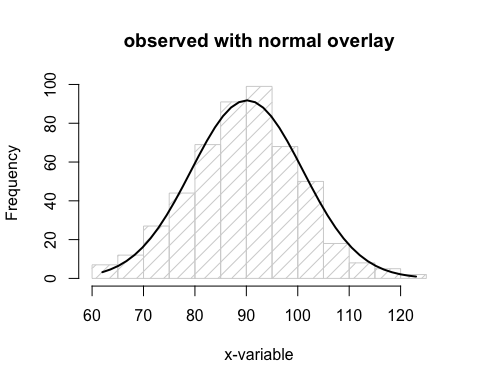
# The distribution for product\_satisfaction looks fairly symmetric and there are no unusual values in either tail.  
# Visulaize distributions of the numeric variables that we will test to see if they are related to product\_satisfaction.   
# Display of each numeric variable with a histogram and then in a scatterplot with product\_satisfaction.  
  
# age  
g = examdata$age  
h = hist(g, breaks =10, density = 10,  
 col = "lightgray", xlab = "xvariable", main = "observed with normal overlay")  
xfit = seq(min(g), max(g), length = 40)  
yfit = dnorm(xfit, mean = mean(g), sd = sd(g))  
yfit = yfit \* diff(h$mids [1 :2]) \* length(g)  
  
lines(xfit, yfit, col = "black", lwd = 2)



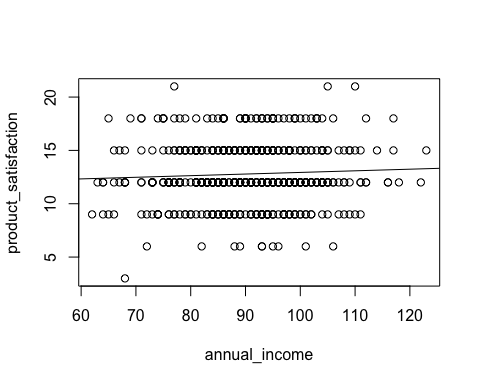
age\_model = lm(product\_satisfaction ~ age, data = examdata)  
plot(product\_satisfaction ~ age, data = examdata)  
abline(age\_model)



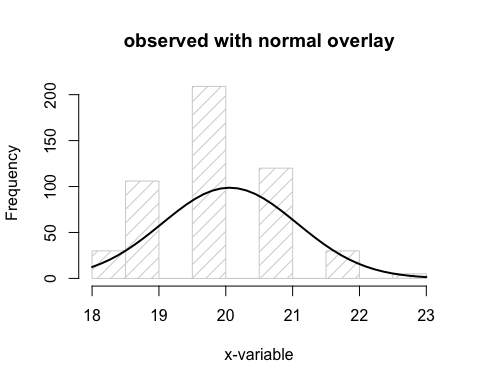
# annual\_income  
g = examdata$annual\_income  
h = hist(g, breaks = 10, density =10,  
 col = "lightgray", xlab = "x-variable", main = "observed with normal overlay")  
xfit = seq(min(g), max(g), length = 40)  
yfit = dnorm(xfit, mean = mean(g), sd = sd(g))  
yfit = yfit \* diff(h$mids[1:2]) \* length(g)  
  
lines(xfit, yfit, col = "black", lwd = 2)



annual\_income\_model = lm(product\_satisfaction ~ annual\_income, data = examdata)  
plot(product\_satisfaction ~ annual\_income, data = examdata)  
abline(annual\_income\_model)



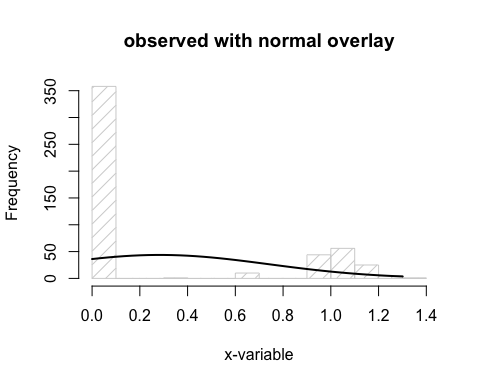
#conscientiousness\_personality  
g = examdata$conscientiousness\_personality  
h = hist(g, breaks = 10, density =10,  
 col = "lightgray", xlab = "x-variable", main = "observed with normal overlay")  
xfit = seq(min(g), max(g), length = 40)  
yfit = dnorm(xfit, mean = mean(g), sd = sd(g))  
yfit = yfit \* diff(h$mids[1:2]) \* length(g)  
  
lines(xfit, yfit, col = "black", lwd = 2)



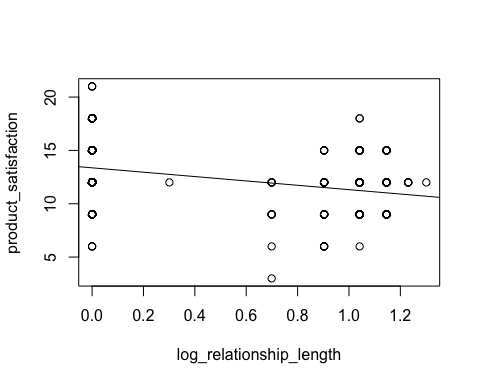
conscientiousness\_personality\_model = lm(product\_satisfaction ~ conscientiousness\_personality, data = examdata)  
plot(product\_satisfaction ~ conscientiousness\_personality, data = examdata)  
abline(conscientiousness\_personality\_model)



# log\_relationship\_length  
g = examdata$log\_relationship\_length  
h = hist(g, breaks = 10, density =10,  
 col = "lightgray", xlab = "x-variable", main = "observed with normal overlay")  
xfit = seq(min(g), max(g), length = 40)  
yfit = dnorm(xfit, mean = mean(g), sd = sd(g))  
yfit = yfit \* diff(h$mids[1:2]) \* length(g)  
  
lines(xfit, yfit, col = "black", lwd = 2)



log\_relationship\_length\_model = lm(product\_satisfaction ~ log\_relationship\_length, data = examdata)  
plot(product\_satisfaction ~ log\_relationship\_length, data = examdata)  
abline(log\_relationship\_length\_model)



# Test which variables are related to satisfaction with the product using multiple regression.   
# Null hypothesis : The variable is not related to satisfaction.  
# Alternative hypothesis : The variable is related to satisfaction.  
str(examdata)

## 'data.frame': 500 obs. of 13 variables:  
## $ cust.id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ in\_relationship : Factor w/ 2 levels "yes","no": 2 2 1 2 2 2 2 2 2 2 ...  
## $ health\_conscious : Factor w/ 3 levels "high","low","medium": 3 3 3 3 3 3 3 2 3 3 ...  
## $ nut\_free : Factor w/ 2 levels "yes","no": 2 2 2 1 2 2 2 1 2 2 ...  
## $ fruit\_free : Factor w/ 2 levels "yes","no": 2 1 2 2 2 2 2 1 2 2 ...  
## $ product\_satisfaction : int 15 21 15 9 15 15 12 9 12 15 ...  
## $ age : int 40 60 40 50 40 40 50 40 50 50 ...  
## $ annual\_income : int 81 77 81 83 108 100 100 85 91 95 ...  
## $ conscientiousness\_personality: int 20 21 20 20 21 21 21 18 20 20 ...  
## $ log\_relationship\_length : num 0 0 1.04 0 0 ...  
## $ num\_in\_relationship : num 0 0 1 0 0 0 0 0 0 0 ...  
## $ num\_nut\_free : num 0 0 0 1 0 0 0 1 0 0 ...  
## $ num\_fruit\_free : num 0 1 0 0 0 0 0 1 0 0 ...

corrplot(cor(examdata[, c(6,7:13)]), method = "ellipse", type = "upper")



model3 <- lm(product\_satisfaction ~ age + annual\_income + conscientiousness\_personality + log\_relationship\_length + num\_in\_relationship + num\_nut\_free+ num\_fruit\_free + health\_conscious, data=examdata)  
summary(model3)

##   
## Call:  
## lm(formula = product\_satisfaction ~ age + annual\_income + conscientiousness\_personality +   
## log\_relationship\_length + num\_in\_relationship + num\_nut\_free +   
## num\_fruit\_free + health\_conscious, data = examdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.7814 -1.8843 -0.3067 1.6961 7.8517   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.30160 2.69446 1.225 0.2210   
## age 0.02351 0.01246 1.887 0.0598 .   
## annual\_income -0.02649 0.01250 -2.119 0.0346 \*   
## conscientiousness\_personality 0.55972 0.13776 4.063 5.64e-05 \*\*\*  
## log\_relationship\_length 2.30802 1.71769 1.344 0.1797   
## num\_in\_relationship -3.94684 1.75326 -2.251 0.0248 \*   
## num\_nut\_free -0.67371 0.31140 -2.164 0.0310 \*   
## num\_fruit\_free -0.45362 0.27594 -1.644 0.1008   
## health\_consciouslow 0.51366 0.54304 0.946 0.3447   
## health\_consciousmedium 0.49287 0.39487 1.248 0.2126   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.659 on 490 degrees of freedom  
## Multiple R-squared: 0.2024, Adjusted R-squared: 0.1877   
## F-statistic: 13.81 on 9 and 490 DF, p-value: < 2.2e-16

# The p-values for the t statistics for annual\_income,conscientiousness\_personality,num\_in\_relationship and num\_nut\_free, are less than a significance level of .05.  
# For these 4 explanatory variables, we reject the null hypothesis and conclude that each relates to satisfaction (taking into account the associations that each variable in the model has with satisfactions)  
# The p-values for the t statistics for age, num\_fruit\_log\_relationship\_length health\_consciouslow and health\_consciousmedium are greater than a significance level of .05.  
# For these 4 explanatory variables, we accept the null hypothesis and conclude that none of these variables relate to satisfaction (taking into account the associations that each variable in the model has with satisfactions).  
# We see that as a set, the explanatory variables account for about 18% of the variation in satisfaction. This indicates that the client may want to invest more effort into finding other variables that are driving satisfaction.  
# Model 4 contains variables that were not statiscally significant  
model4 = lm(product\_satisfaction ~ age + annual\_income + conscientiousness\_personality + num\_in\_relationship + num\_nut\_free, data=examdata)  
summary(model4)

##   
## Call:  
## lm(formula = product\_satisfaction ~ age + annual\_income + conscientiousness\_personality +   
## num\_in\_relationship + num\_nut\_free, data = examdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.4541 -1.8377 -0.1747 1.6358 8.0741   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.23158 2.56305 0.871 0.384354   
## age 0.02745 0.01238 2.217 0.027045 \*   
## annual\_income -0.02130 0.01221 -1.745 0.081654 .   
## conscientiousness\_personality 0.59167 0.13615 4.346 1.69e-05 \*\*\*  
## num\_in\_relationship -1.66833 0.28064 -5.945 5.24e-09 \*\*\*  
## num\_nut\_free -0.93678 0.27134 -3.452 0.000603 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.669 on 494 degrees of freedom  
## Multiple R-squared: 0.1901, Adjusted R-squared: 0.1819   
## F-statistic: 23.19 on 5 and 494 DF, p-value: < 2.2e-16

anova(model4, model3)

## Analysis of Variance Table  
##   
## Model 1: product\_satisfaction ~ age + annual\_income + conscientiousness\_personality +   
## num\_in\_relationship + num\_nut\_free  
## Model 2: product\_satisfaction ~ age + annual\_income + conscientiousness\_personality +   
## log\_relationship\_length + num\_in\_relationship + num\_nut\_free +   
## num\_fruit\_free + health\_conscious  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 494 3518.3   
## 2 490 3464.9 4 53.436 1.8892 0.1111

# From the anova test used to compare the fit of these two models, we see that the test statistic has a p value that is greater than .05, so the fit of the full model is statistically different from the fit of the reduced model. In fact, the adjusted R squared value is still at about 18%.  
# Satisfaction with the product is related to whether the customer prefers fruit-free chocolate, whether the customer health consciousness is high or low or if the customer is has been in a reelatonship for long.