stat-364-codes-final-project

2021-06-18

# STAT 364 CODES FINAL PROJECT (18 June 2021)  
  
# Alper Tunahan OZTURK 2290856  
# Dayanch AKMYRADOV 2347292  
# Hamza Anil YILDIRIM 2218386  
# Kubilay TASYUREK 2218329  
  
library(AID)

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

library(ggplot2)  
library(sjPlot)

## Install package "strengejacke" from GitHub (`devtools::install\_github("strengejacke/strengejacke")`) to load all sj-packages at once!

library(MASS)  
  
#data handling  
data <- read.csv("data.csv", sep = ";")  
data$APR\_offered = as.factor(data$APR\_offered)  
  
# removing unwanted parts  
data <- data[,c(-1, -2, -9, -11)]  
  
str(data)

## 'data.frame': 1001 obs. of 13 variables:  
## $ Age\_group : chr "55+" "18-25" "25-35" "18-25" ...  
## $ APR\_offered : Factor w/ 4 levels "10.99","12.99",..: 2 1 2 3 2 1 3 3 1 4 ...  
## $ Grocery\_spend : num 166 1200 708 1872 1965 ...  
## $ Travel\_spend : num 1445.5 2360 47.4 2861 1471.6 ...  
## $ Apparels\_spend : num 911 1099 1403 773 793 ...  
## $ DE\_spend : num 816 188 538 391 176 ...  
## $ Income\_Group : chr "6000-9000" "12000+" "6000-9000" "3000-6000" ...  
## $ Avg\_monthly\_spend : num 2674 4944 2336 5766 4448 ...  
## $ Avg\_Monthly\_balance: num 6783 12368 4756 11958 7722 ...  
## $ Avg\_monthyl\_payrate: num 0.54 0.47 0.12 0.71 0.58 0.57 0.27 0.15 0.24 0.55 ...  
## $ student\_loan\_ind : int 1 1 0 1 0 0 0 1 0 0 ...  
## $ Times\_Mailed : int 10 9 7 3 9 10 9 7 8 8 ...  
## $ No\_of\_tradelines : int 9 9 6 2 9 10 7 4 5 8 ...

summary(data)

## Age\_group APR\_offered Grocery\_spend Travel\_spend   
## Length:1001 10.99:267 Min. : 0.9 Min. : 0.7586   
## Class :character 12.99:248 1st Qu.: 509.3 1st Qu.: 387.2445   
## Mode :character 14.99:253 Median : 972.0 Median : 826.8961   
## 16.99:233 Mean :1011.1 Mean : 938.3039   
## 3rd Qu.:1538.6 3rd Qu.:1267.2932   
## Max. :1999.2 Max. :2999.1945   
## Apparels\_spend DE\_spend Income\_Group Avg\_monthly\_spend  
## Min. : 0.55 Min. : 0.9 Length:1001 Min. : 377.1   
## 1st Qu.: 392.77 1st Qu.:259.1 Class :character 1st Qu.:2169.0   
## Median : 731.72 Median :503.7 Mode :character Median :2865.0   
## Mean : 749.48 Mean :507.3 Mean :2913.4   
## 3rd Qu.:1117.13 3rd Qu.:751.6 3rd Qu.:3559.6   
## Max. :1498.39 Max. :999.6 Max. :6246.6   
## Avg\_Monthly\_balance Avg\_monthyl\_payrate student\_loan\_ind Times\_Mailed   
## Min. : 641.8 Min. :0.1000 Min. :0.0000 Min. : 3.000   
## 1st Qu.: 3764.9 1st Qu.:0.3000 1st Qu.:0.0000 1st Qu.: 7.000   
## Median : 5438.0 Median :0.5000 Median :0.0000 Median : 8.000   
## Mean : 5847.0 Mean :0.4828 Mean :0.3786 Mean : 7.974   
## 3rd Qu.: 7473.0 3rd Qu.:0.6500 3rd Qu.:1.0000 3rd Qu.: 9.000   
## Max. :17775.7 Max. :1.0000 Max. :1.0000 Max. :10.000   
## No\_of\_tradelines  
## Min. : 0.000   
## 1st Qu.: 5.000   
## Median : 7.000   
## Mean : 6.485   
## 3rd Qu.: 8.000   
## Max. :10.000

attach(data)  
  
####################################################################  
### Research Question 1  
  
# fit the model  
fit1 <- lm(Avg\_monthly\_spend ~ ., data = data)  
summary(fit1)

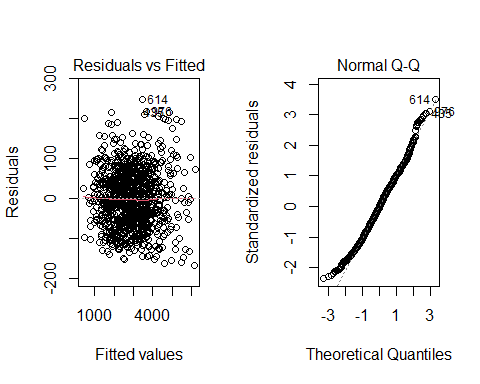
##   
## Call:  
## lm(formula = Avg\_monthly\_spend ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -166.106 -52.788 -2.724 47.879 247.772   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 268.241244 19.460186 13.784 < 2e-16 \*\*\*  
## Age\_group25-35 -1.114489 6.498700 -0.171 0.86387   
## Age\_group35-55 -10.625524 6.570400 -1.617 0.10616   
## Age\_group55+ -9.554624 6.450542 -1.481 0.13887   
## APR\_offered12.99 -3.689805 6.377983 -0.579 0.56304   
## APR\_offered14.99 -0.551200 6.367753 -0.087 0.93104   
## APR\_offered16.99 -12.014341 6.532624 -1.839 0.06620 .   
## Grocery\_spend 0.990671 0.004640 213.514 < 2e-16 \*\*\*  
## Travel\_spend 0.997979 0.004575 218.127 < 2e-16 \*\*\*  
## Apparels\_spend 0.988788 0.006019 164.286 < 2e-16 \*\*\*  
## DE\_spend 0.007527 0.007965 0.945 0.34491   
## Income\_Group12000+ 6.237707 7.061197 0.883 0.37725   
## Income\_Group3000-6000 12.823942 7.277986 1.762 0.07838 .   
## Income\_Group6000-9000 -4.515018 7.448629 -0.606 0.54455   
## Income\_Group9000-12000 10.291725 7.269477 1.416 0.15717   
## Avg\_Monthly\_balance 0.004306 0.001322 3.257 0.00116 \*\*   
## Avg\_monthyl\_payrate 63.059085 11.877546 5.309 1.36e-07 \*\*\*  
## student\_loan\_ind -3.677673 4.702670 -0.782 0.43438   
## Times\_Mailed -10.271695 2.498588 -4.111 4.27e-05 \*\*\*  
## No\_of\_tradelines -0.942304 2.029242 -0.464 0.64249   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 71.81 on 981 degrees of freedom  
## Multiple R-squared: 0.9952, Adjusted R-squared: 0.9951   
## F-statistic: 1.072e+04 on 19 and 981 DF, p-value: < 2.2e-16

# check for multicollinearity  
car::vif(fit1)

## Registered S3 methods overwritten by 'car':  
## method from  
## influence.merMod lme4  
## cooks.distance.influence.merMod lme4  
## dfbeta.influence.merMod lme4  
## dfbetas.influence.merMod lme4

## GVIF Df GVIF^(1/(2\*Df))  
## Age\_group 1.043168 3 1.007069  
## APR\_offered 1.063134 3 1.010256  
## Grocery\_spend 1.442809 1 1.201170  
## Travel\_spend 1.991527 1 1.411215  
## Apparels\_spend 1.311760 1 1.145321  
## DE\_spend 1.025376 1 1.012608  
## Income\_Group 1.065226 4 1.007930  
## Avg\_Monthly\_balance 2.527185 1 1.589712  
## Avg\_monthyl\_payrate 1.320844 1 1.149280  
## student\_loan\_ind 1.010065 1 1.005020  
## Times\_Mailed 3.859051 1 1.964447  
## No\_of\_tradelines 3.467554 1 1.862137

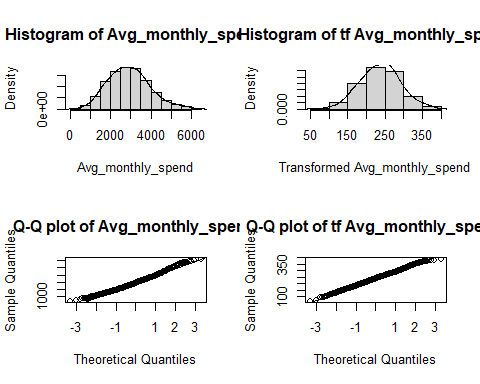
# plot residual vs fitted and normal q-q  
par(mfrow=c(1,2))  
plot(fit1, c(1,2))



# normality test for avg\_monthly\_spend (response)  
shapiro.test(Avg\_monthly\_spend)

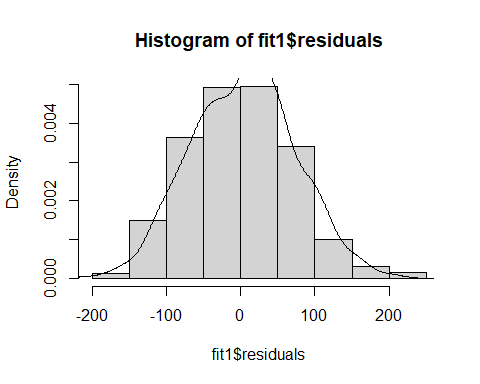
##   
## Shapiro-Wilk normality test  
##   
## data: Avg\_monthly\_spend  
## W = 0.98993, p-value = 2.322e-06

# apply transformation  
boxcoxnc(Avg\_monthly\_spend)

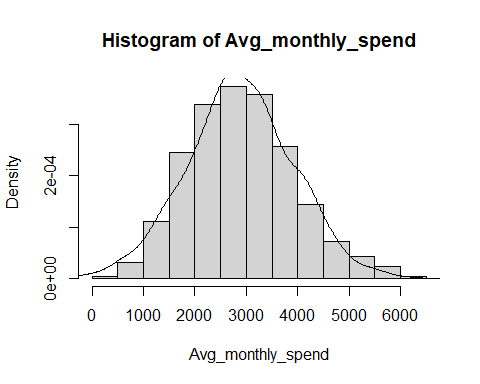


##   
## Box-Cox power transformation   
## -------------------------------------------------------------------   
## data : Avg\_monthly\_spend   
##   
## lambda.hat : 0.63   
##   
##   
## Shapiro-Wilk normality test for transformed data (alpha = 0.05)  
## -------------------------------------------------------------------   
##   
## statistic : 0.998778   
## p.value : 0.7397786   
##   
## Result : Transformed data are normal.   
## -------------------------------------------------------------------

# Normality check for residuals  
par(mfrow=c(1,1))  
hist(fit1$residuals, probability = T)  
lines(density(rnorm(1001, mean(fit1$residuals), sd(fit1$residuals))))



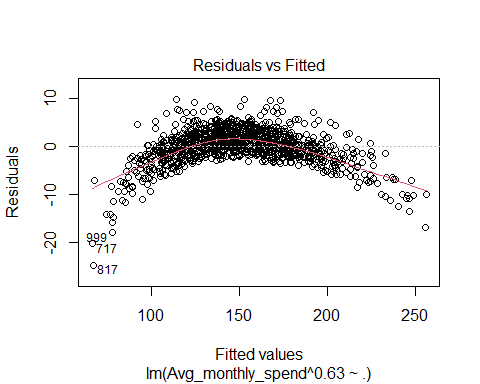
# Normality check for response  
hist(Avg\_monthly\_spend, probability = T)  
lines(density(rnorm(1001, mean(Avg\_monthly\_spend), sd(Avg\_monthly\_spend))))



# fit the model for transformed data  
fit1 <- lm(Avg\_monthly\_spend^0.63 ~ ., data = data)  
summary(fit1)

##   
## Call:  
## lm(formula = Avg\_monthly\_spend^0.63 ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.8311 -1.7873 0.2651 2.4890 9.8618   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.088e+01 1.042e+00 58.410 < 2e-16 \*\*\*  
## Age\_group25-35 -2.849e-01 3.481e-01 -0.819 0.41326   
## Age\_group35-55 -6.542e-01 3.519e-01 -1.859 0.06334 .   
## Age\_group55+ -3.161e-01 3.455e-01 -0.915 0.36044   
## APR\_offered12.99 -2.678e-01 3.416e-01 -0.784 0.43328   
## APR\_offered14.99 6.791e-02 3.411e-01 0.199 0.84221   
## APR\_offered16.99 -1.842e-01 3.499e-01 -0.526 0.59872   
## Grocery\_spend 3.360e-02 2.485e-04 135.186 < 2e-16 \*\*\*  
## Travel\_spend 3.263e-02 2.451e-04 133.151 < 2e-16 \*\*\*  
## Apparels\_spend 3.411e-02 3.224e-04 105.796 < 2e-16 \*\*\*  
## DE\_spend 4.092e-04 4.266e-04 0.959 0.33765   
## Income\_Group12000+ 2.475e-02 3.782e-01 0.065 0.94783   
## Income\_Group3000-6000 7.716e-01 3.898e-01 1.979 0.04805 \*   
## Income\_Group6000-9000 -7.979e-02 3.990e-01 -0.200 0.84153   
## Income\_Group9000-12000 9.442e-01 3.894e-01 2.425 0.01549 \*   
## Avg\_Monthly\_balance 7.783e-05 7.081e-05 1.099 0.27196   
## Avg\_monthyl\_payrate 9.809e-01 6.362e-01 1.542 0.12341   
## student\_loan\_ind -3.484e-01 2.519e-01 -1.383 0.16697   
## Times\_Mailed -4.023e-01 1.338e-01 -3.006 0.00271 \*\*   
## No\_of\_tradelines 1.864e-01 1.087e-01 1.715 0.08667 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.846 on 981 degrees of freedom  
## Multiple R-squared: 0.9876, Adjusted R-squared: 0.9874   
## F-statistic: 4112 on 19 and 981 DF, p-value: < 2.2e-16

# Check for residual vs fitted   
plot(fit1, 1) # variance is not constant



# model output  
tab\_model(fit1)

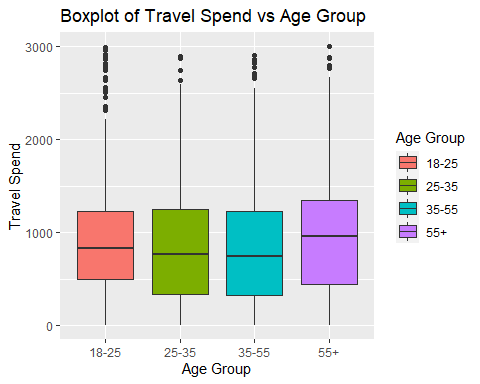
# standardized estimates  
effectsize::standardize\_parameters(fit1)

## Warning in summary.lm(x): essentially perfect fit: summary may be unreliable

## Warning in summary.lm(model): essentially perfect fit: summary may be unreliable  
  
## Warning in summary.lm(model): essentially perfect fit: summary may be unreliable  
  
## Warning in summary.lm(model): essentially perfect fit: summary may be unreliable

## # Standardization method: refit  
##   
## Parameter | Coefficient (std.) | 95% CI  
## -----------------------------------------------------------  
## (Intercept) | 0.00 | [ 0.00, 0.00]  
## Avg\_monthly\_spend^0.63 | 1.00 | [ 1.00, 1.00]  
## Age\_group25-35 | 1.25e-16 | [ 0.00, 0.00]  
## Age\_group35-55 | 2.77e-16 | [ 0.00, 0.00]  
## Age\_group55+ | 2.30e-16 | [ 0.00, 0.00]  
## APR\_offered12.99 | -1.71e-17 | [ 0.00, 0.00]  
## APR\_offered14.99 | 1.54e-17 | [ 0.00, 0.00]  
## APR\_offered16.99 | 1.73e-16 | [ 0.00, 0.00]  
## Grocery\_spend | 6.02e-16 | [ 0.00, 0.00]  
## Travel\_spend | 6.25e-16 | [ 0.00, 0.00]  
## Apparels\_spend | 3.70e-16 | [ 0.00, 0.00]  
## DE\_spend | 3.85e-17 | [ 0.00, 0.00]  
## Income\_Group12000+ | -2.79e-16 | [ 0.00, 0.00]  
## Income\_Group3000-6000 | -2.78e-16 | [ 0.00, 0.00]  
## Income\_Group6000-9000 | -3.18e-16 | [ 0.00, 0.00]  
## Income\_Group9000-12000 | -3.48e-16 | [ 0.00, 0.00]  
## Avg\_Monthly\_balance | -6.09e-17 | [ 0.00, 0.00]  
## Avg\_monthyl\_payrate | 2.99e-17 | [ 0.00, 0.00]  
## student\_loan\_ind | -3.71e-17 | [ 0.00, 0.00]  
## Times\_Mailed | 9.00e-17 | [ 0.00, 0.00]  
## No\_of\_tradelines | -2.17e-16 | [ 0.00, 0.00]

####################################################################  
### Research Question 2  
  
# boxplot for travel spend of age groups  
ggplot(data, aes(x=Age\_group, y=Travel\_spend,fill=Age\_group)) +  
 geom\_boxplot() + xlab("Age Group") + ylab("Travel Spend") +  
 labs(fill = "Age Group") +  
 ggtitle("Boxplot of Travel Spend vs Age Group")

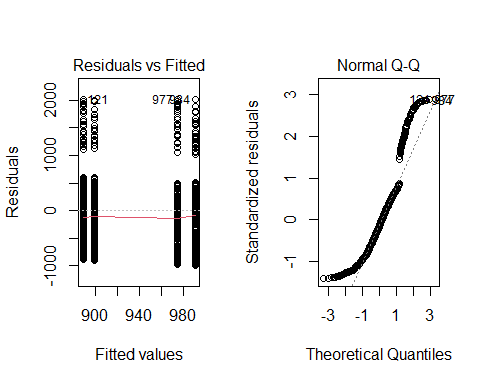


# fit the model  
fit2 = lm(Travel\_spend ~ Age\_group, data)  
summary(fit2) # estimates are not significant

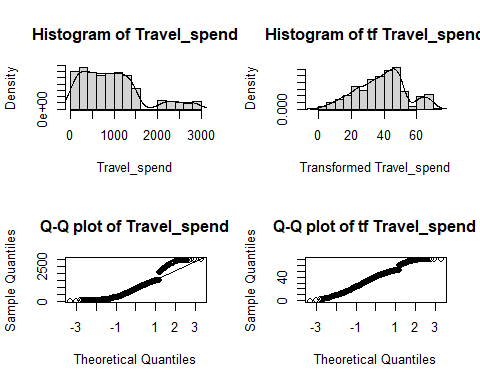
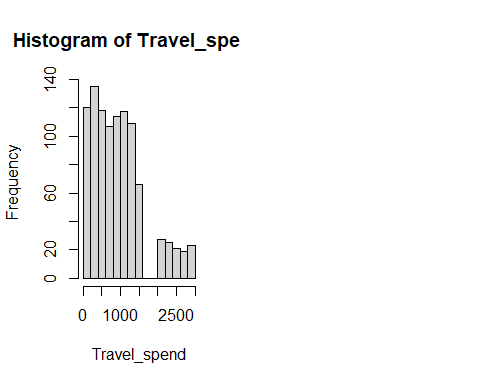
##   
## Call:  
## lm(formula = Travel\_spend ~ Age\_group, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -985.5 -537.5 -115.0 324.2 2012.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 974.50 45.18 21.567 <2e-16 \*\*\*  
## Age\_group25-35 -76.19 63.07 -1.208 0.227   
## Age\_group35-55 -86.07 63.57 -1.354 0.176   
## Age\_group55+ 15.70 62.49 0.251 0.802   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 700 on 997 degrees of freedom  
## Multiple R-squared: 0.004152, Adjusted R-squared: 0.001155   
## F-statistic: 1.386 on 3 and 997 DF, p-value: 0.2457

# model output  
tab\_model(fit2)

# plot residual vs fitted and normal q-q  
par(mfrow=c(1,2))  
plot(fit2, c(1,2))



hist(Travel\_spend)  
  
# transformation on data  
boxcoxnc(Travel\_spend)



##   
## Box-Cox power transformation   
## -------------------------------------------------------------------   
## data : Travel\_spend   
##   
## lambda.hat : 0.43   
##   
##   
## Shapiro-Wilk normality test for transformed data (alpha = 0.05)  
## -------------------------------------------------------------------   
##   
## statistic : 0.9854625   
## p.value : 2.013102e-08   
##   
## Result : Transformed data are not normal.   
## -------------------------------------------------------------------

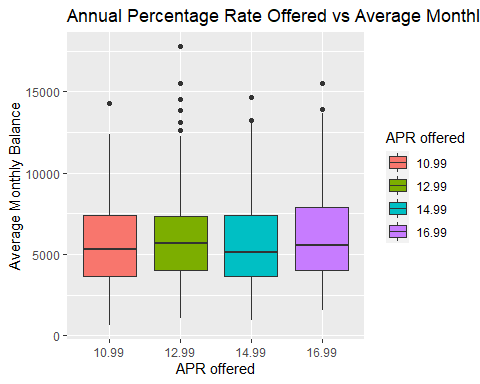
# robust regression  
huber <- rlm(Travel\_spend ~ Age\_group, data)  
summary(huber) # estimates are not significant

##   
## Call: rlm(formula = Travel\_spend ~ Age\_group, data = data)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -908.18 -457.22 -31.17 412.18 2113.49   
##   
## Coefficients:  
## Value Std. Error t value   
## (Intercept) 872.1478 38.1476 22.8625  
## Age\_group25-35 -57.7472 53.2513 -1.0844  
## Age\_group35-55 -73.4277 53.6729 -1.3681  
## Age\_group55+ 40.6823 52.7562 0.7711  
##   
## Residual standard error: 641.2 on 997 degrees of freedom

bisquare <- rlm(Travel\_spend ~ Age\_group, data, psi = psi.bisquare)  
summary(bisquare) #estimates are not significant

##   
## Call: rlm(formula = Travel\_spend ~ Age\_group, data = data, psi = psi.bisquare)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -918.01 -456.31 -31.18 413.32 2117.24   
##   
## Coefficients:  
## Value Std. Error t value   
## (Intercept) 868.3942 41.4673 20.9417  
## Age\_group25-35 -56.8912 57.8854 -0.9828  
## Age\_group35-55 -72.8024 58.3437 -1.2478  
## Age\_group55+ 54.2641 57.3472 0.9462  
##   
## Residual standard error: 638.6 on 997 degrees of freedom

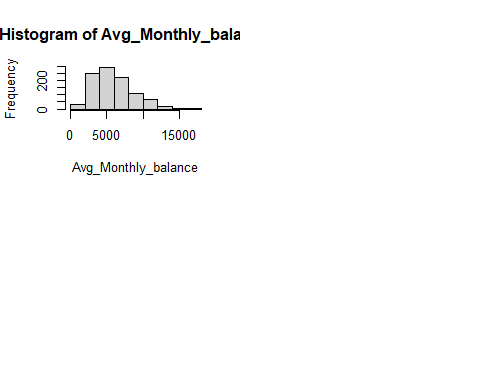
####################################################################  
### Research Question 3   
# boxplot for average monthly balance of apr offered  
ggplot(data, aes(x=APR\_offered, y=Avg\_Monthly\_balance,fill=APR\_offered)) +  
 geom\_boxplot() + xlab("APR offered") + ylab("Average Monthly Balance") +  
 labs(fill = "APR offered") +  
 ggtitle("Annual Percentage Rate Offered vs Average Monthly Balance")



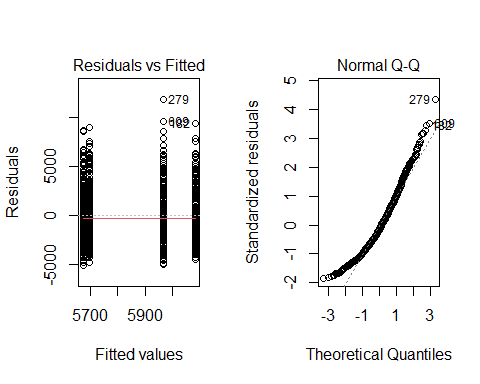
# fit the model  
fit3 = lm(Avg\_Monthly\_balance ~ APR\_offered, data)  
summary(fit3)

##   
## Call:  
## lm(formula = Avg\_Monthly\_balance ~ APR\_offered, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5031.5 -2010.4 -469.4 1698.4 11809.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5673.21 167.02 33.967 <2e-16 \*\*\*  
## APR\_offered12.99 293.11 240.69 1.218 0.2236   
## APR\_offered14.99 24.05 239.45 0.100 0.9200   
## APR\_offered16.99 408.46 244.67 1.669 0.0953 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2729 on 997 degrees of freedom  
## Multiple R-squared: 0.004037, Adjusted R-squared: 0.00104   
## F-statistic: 1.347 on 3 and 997 DF, p-value: 0.2577

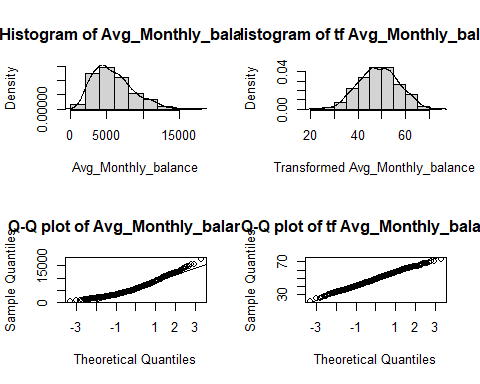
hist(Avg\_Monthly\_balance)  
par(mfrow=c(1,2))



plot(fit3, c(1,2))



# transformation on data  
boxcoxnc(Avg\_Monthly\_balance)



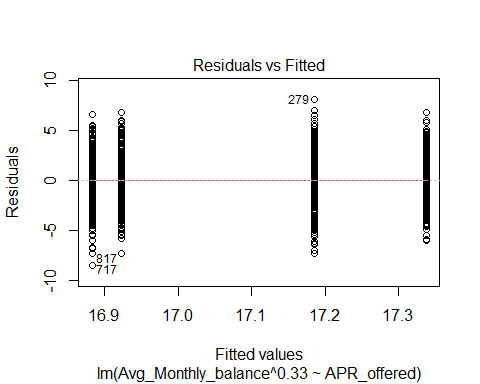
##   
## Box-Cox power transformation   
## -------------------------------------------------------------------   
## data : Avg\_Monthly\_balance   
##   
## lambda.hat : 0.33   
##   
##   
## Shapiro-Wilk normality test for transformed data (alpha = 0.05)  
## -------------------------------------------------------------------   
##   
## statistic : 0.9978442   
## p.value : 0.2226993   
##   
## Result : Transformed data are normal.   
## -------------------------------------------------------------------

# fit the model for transformed data  
fit3 = lm(Avg\_Monthly\_balance ^ 0.33 ~ APR\_offered, data)  
summary(fit3)

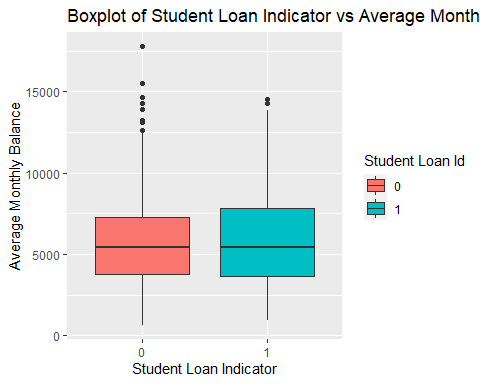
##   
## Call:  
## lm(formula = Avg\_Monthly\_balance^0.33 ~ APR\_offered, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.441 -1.877 -0.041 1.952 8.075   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.88267 0.16637 101.474 <2e-16 \*\*\*  
## APR\_offered12.99 0.30270 0.23975 1.263 0.2070   
## APR\_offered14.99 0.03954 0.23852 0.166 0.8684   
## APR\_offered16.99 0.45472 0.24372 1.866 0.0624 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.719 on 997 degrees of freedom  
## Multiple R-squared: 0.004706, Adjusted R-squared: 0.001712   
## F-statistic: 1.572 on 3 and 997 DF, p-value: 0.1947

# model output  
tab\_model(fit3)

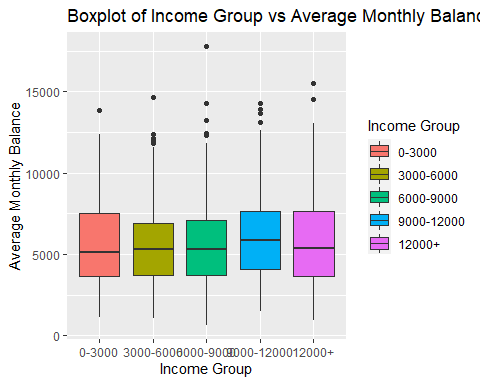
# residuals vs fitted (for constant variance)  
par(mfrow=c(1,1))  
plot(fit3, 1) #seems fine



####################################################################  
### Research Question 4  
  
# boxplot for balance of student loan indicator  
ggplot(data, aes(x=as.factor(student\_loan\_ind), y=Avg\_Monthly\_balance,fill=as.factor(student\_loan\_ind))) +  
 geom\_boxplot() +xlab("Student Loan Indicator")+ylab("Average Monthly Balance")+labs(fill="Student Loan Id")+  
 ggtitle("Boxplot of Student Loan Indicator vs Average Monthly Balance ")



# boxplot for balance of income groups  
data$Income\_Group <- factor(data$Income\_Group,levels=c("0-3000","3000-6000","6000-9000","9000-12000","12000+"))  
ggplot(data, aes(x = Income\_Group, y = Avg\_Monthly\_balance, fill = Income\_Group)) +xlab("Income Group")+  
 ylab("Average Monthly Balance")+ggtitle("Boxplot of Income Group vs Average Monthly Balance")+geom\_boxplot() +  
 labs(fill = "Income Group")



# no need for checking normality, see research question 3  
  
# fit the model for transformed data  
fit4 <- lm(Avg\_Monthly\_balance ^ 0.33 ~ student\_loan\_ind + Income\_Group, data)  
summary(fit4)

##   
## Call:  
## lm(formula = Avg\_Monthly\_balance^0.33 ~ student\_loan\_ind + Income\_Group,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.5948 -1.9565 -0.0214 1.8881 8.2240   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.81946 0.20645 81.470 <2e-16 \*\*\*  
## student\_loan\_ind 0.04801 0.17769 0.270 0.7871   
## Income\_Group3000-6000 0.09066 0.27437 0.330 0.7411   
## Income\_Group6000-9000 0.21706 0.28111 0.772 0.4402   
## Income\_Group9000-12000 0.60968 0.27234 2.239 0.0254 \*   
## Income\_Group12000+ 0.24814 0.26501 0.936 0.3493   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.72 on 995 degrees of freedom  
## Multiple R-squared: 0.005959, Adjusted R-squared: 0.0009641   
## F-statistic: 1.193 on 5 and 995 DF, p-value: 0.3105

# model output  
tab\_model(fit4