Slicr and DiceR Project 1

Spencer Abbot, Chi Cheung, Eric Goldman

10/4/2017

#install packages for the project  
install.packages("maps", repos="http://cran.rstudio.com/")

## Installing package into 'C:/Users/ericg/Documents/R/win-library/3.3'  
## (as 'lib' is unspecified)

## package 'maps' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\ericg\AppData\Local\Temp\RtmpkHUMYT\downloaded\_packages

install.packages("ggplot2", repos="http://cran.rstudio.com/")

## Installing package into 'C:/Users/ericg/Documents/R/win-library/3.3'  
## (as 'lib' is unspecified)

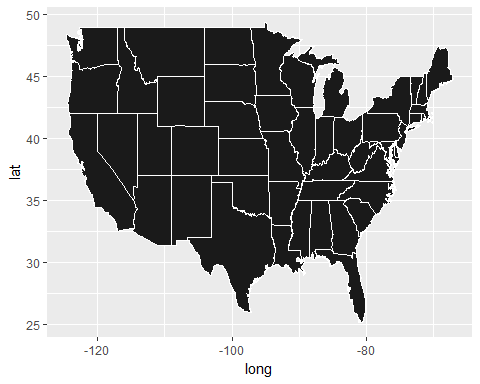
## package 'ggplot2' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\ericg\AppData\Local\Temp\RtmpkHUMYT\downloaded\_packages

install.packages("pastecs", repos="http://cran.rstudio.com/")

## Installing package into 'C:/Users/ericg/Documents/R/win-library/3.3'  
## (as 'lib' is unspecified)

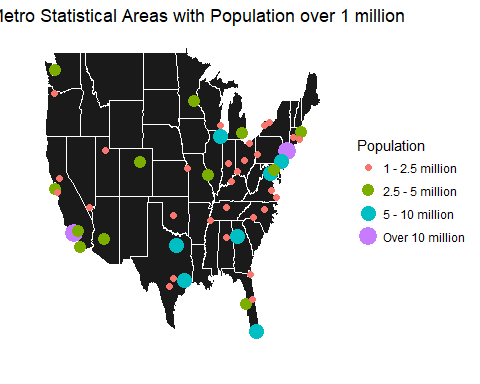
## package 'pastecs' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\ericg\AppData\Local\Temp\RtmpkHUMYT\downloaded\_packages

#Plotting Map of the US and the fifty metro cities analyzed  
#Call library for maps and plotting already installed   
library(ggplot2)  
library(maps)  
  
#US map - data  
all\_states <- map\_data("state")  
  
#Read in Lat Long Data from City  
citylatlong <- data.frame(read.csv("citylatlong.csv"))  
  
#Plot US states using ggplot2  
p <- ggplot()  
p <- p + geom\_polygon( data=all\_states,   
aes(x=long, y=lat, group = group),colour="white", fill="grey10" )  
p;



#Plot data points (lat/long of MSA's) to US map  
p + geom\_point(data=citylatlong,   
aes(y=LAT, x=LONG, color=Population, size=Population)) +  
 ggtitle("US Metro Statistical Areas with Population over 1 million")+  
 theme\_bw() +  
 theme( plot.title=element\_text(hjust=0.5),  
 axis.line = element\_blank(),  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 panel.border = element\_blank(),  
 panel.background = element\_blank(),  
 axis.text.y =element\_blank(),  
 axis.text.x=element\_blank(),  
 axis.ticks.y=element\_blank(),  
 axis.ticks.x=element\_blank()) +  
 xlab("") +   
 ylab("")

## Warning: Using size for a discrete variable is not advised.



#Read in CityData csv and call stats package  
library(stats)  
citydata <- data.frame(read.csv("CityData.csv"))  
  
#CityData Exploration - 50 obs of 9 variables  
head(citydata)

## MSA ZillowHVI ZillowRent Population  
## 1 New York 427300 2619.397 20153634  
## 2 Los Angeles-Long Beach-Anaheim 612400 3065.439 13310447  
## 3 Chicago 213300 1783.704 9512999  
## 4 Dallas-Fort Worth 213000 1795.106 7233323  
## 5 Houston 182100 1834.825 6772470  
## 6 Washington 384500 2108.527 6131977  
## CollegeDegree F500\_HQ CrimeIndex Amazon\_HQIndex MLB\_Team  
## 1 36.0 44 371.4 49.50080 YES  
## 2 31.0 15 368.9 40.91774 YES  
## 3 34.0 28 380.1 41.91189 YES  
## 4 31.1 22 303.0 41.91715 YES  
## 5 28.4 24 567.4 30.38100 YES  
## 6 46.8 12 316.6 49.78833 YES

str(citydata)

## 'data.frame': 50 obs. of 9 variables:  
## $ MSA : Factor w/ 50 levels "Atlanta","Austin",..: 28 21 8 12 16 50 31 24 1 5 ...  
## $ ZillowHVI : int 427300 612400 213300 213000 182100 384500 218800 256100 181700 430200 ...  
## $ ZillowRent : num 2619 3065 1784 1795 1835 ...  
## $ Population : int 20153634 13310447 9512999 7233323 6772470 6131977 6070500 6066387 5789700 4794447 ...  
## $ CollegeDegree : num 36 31 34 31.1 28.4 46.8 33.1 28.1 34.1 43 ...  
## $ F500\_HQ : int 44 15 28 22 24 12 10 5 12 9 ...  
## $ CrimeIndex : num 371 369 380 303 567 ...  
## $ Amazon\_HQIndex: num 49.5 40.9 41.9 41.9 30.4 ...  
## $ MLB\_Team : Factor w/ 2 levels "NO","YES": 2 2 2 2 2 2 2 2 2 2 ...

summary(citydata)

## MSA ZillowHVI ZillowRent Population   
## Atlanta : 1 Min. : 121800 Min. : 730.8 Min. : 1078879   
## Austin : 1 1st Qu.: 161725 1st Qu.:1290.5 1st Qu.: 1642789   
## Baltimore : 1 Median : 217900 Median :1595.6 Median : 2383627   
## Birmingham: 1 Mean : 270392 Mean :1732.0 Mean : 3554510   
## Boston : 1 3rd Qu.: 273550 3rd Qu.:1864.5 3rd Qu.: 4470282   
## Buffalo : 1 Max. :1038900 Max. :4404.2 Max. :20153634   
## (Other) :44   
## CollegeDegree F500\_HQ CrimeIndex Amazon\_HQIndex MLB\_Team  
## Min. :19.50 Min. : 0.00 Min. : 127.6 Min. :10.56 NO :25   
## 1st Qu.:28.10 1st Qu.: 2.00 1st Qu.: 325.7 1st Qu.:33.58 YES:25   
## Median :30.65 Median : 4.00 Median : 393.6 Median :37.68   
## Mean :31.65 Mean : 6.62 Mean : 420.2 Mean :37.95   
## 3rd Qu.:34.08 3rd Qu.: 8.75 3rd Qu.: 491.4 3rd Qu.:44.23   
## Max. :46.80 Max. :44.00 Max. :1033.5 Max. :51.22   
##

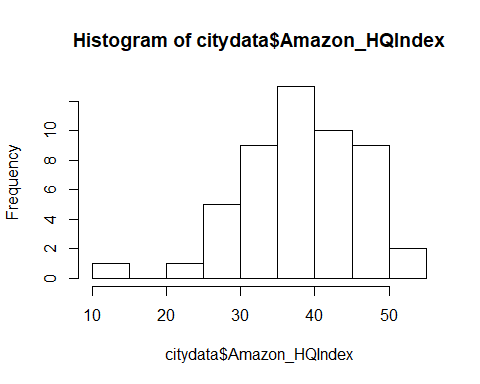
#Gather stats for citydata  
library(pastecs)

## Loading required package: boot

stat.desc(citydata)

## MSA ZillowHVI ZillowRent Population CollegeDegree  
## nbr.val NA 5.000000e+01 5.000000e+01 5.000000e+01 50.0000000  
## nbr.null NA 0.000000e+00 0.000000e+00 0.000000e+00 0.0000000  
## nbr.na NA 0.000000e+00 0.000000e+00 0.000000e+00 0.0000000  
## min NA 1.218000e+05 7.307500e+02 1.078879e+06 19.5000000  
## max NA 1.038900e+06 4.404208e+03 2.015363e+07 46.8000000  
## range NA 9.171000e+05 3.673458e+03 1.907476e+07 27.3000000  
## sum NA 1.351960e+07 8.659797e+04 1.777255e+08 1582.5000000  
## median NA 2.179000e+05 1.595634e+03 2.383627e+06 30.6500000  
## mean NA 2.703920e+05 1.731959e+03 3.554510e+06 31.6500000  
## SE.mean NA 2.539544e+04 1.028234e+02 4.779721e+05 0.8128320  
## CI.mean NA 5.103405e+04 2.066314e+02 9.605210e+05 1.6334471  
## var NA 3.224642e+10 5.286329e+05 1.142287e+13 33.0347959  
## std.dev NA 1.795729e+05 7.270715e+02 3.379773e+06 5.7475904  
## coef.var NA 6.641206e-01 4.197971e-01 9.508408e-01 0.1815984  
## F500\_HQ CrimeIndex Amazon\_HQIndex MLB\_Team  
## nbr.val 50.000000 5.000000e+01 50.0000000 NA  
## nbr.null 4.000000 0.000000e+00 0.0000000 NA  
## nbr.na 0.000000 0.000000e+00 0.0000000 NA  
## min 0.000000 1.276000e+02 10.5556385 NA  
## max 44.000000 1.033500e+03 51.2168606 NA  
## range 44.000000 9.059000e+02 40.6612221 NA  
## sum 331.000000 2.100860e+04 1897.5626728 NA  
## median 4.000000 3.936000e+02 37.6773177 NA  
## mean 6.620000 4.201720e+02 37.9512535 NA  
## SE.mean 1.153129 2.271481e+01 1.1663479 NA  
## CI.mean 2.317299 4.564712e+01 2.3438638 NA  
## var 66.485306 2.579812e+04 68.0183661 NA  
## std.dev 8.153852 1.606179e+02 8.2473248 NA  
## coef.var 1.231700 3.822671e-01 0.2173136 NA

#Testing for normality  
hist(citydata$Amazon\_HQIndex)

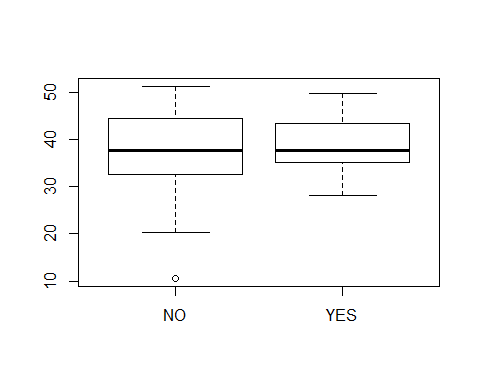


shapiro.test(citydata$Amazon\_HQIndex)

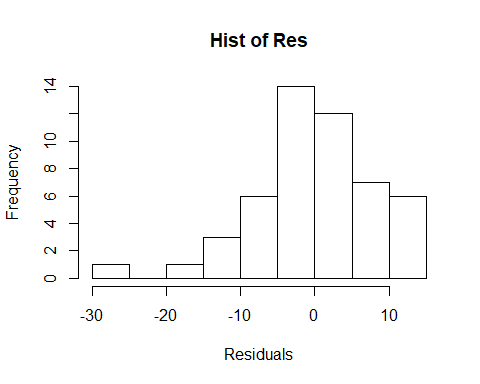
##   
## Shapiro-Wilk normality test  
##   
## data: citydata$Amazon\_HQIndex  
## W = 0.95943, p-value = 0.08412

P-value is 0.084. Fail to reject normality but there appears to be an outlier. Will explore outlier city (Memphis) later in Rmarkdown

#Comparing MLB Team on Amazon HQ Index  
boxplot(citydata$Amazon\_HQIndex~citydata$MLB\_Team)



#Appears to be even, using ANOVA to investigate  
  
model.degree <- aov(Amazon\_HQIndex~MLB\_Team, data = citydata)  
hist(model.degree$residuals, main = "Hist of Res", xlab = "Residuals")



summary(model.degree)

## Df Sum Sq Mean Sq F value Pr(>F)  
## MLB\_Team 1 69 69.36 1.02 0.318  
## Residuals 48 3264 67.99

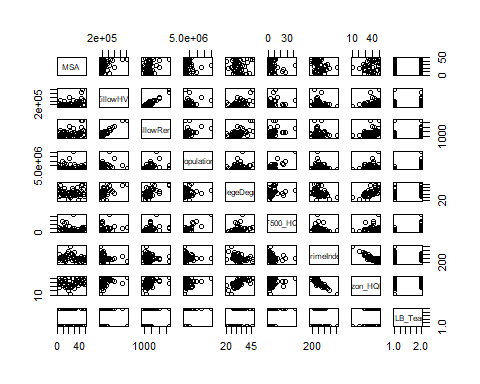
TukeyHSD(model.degree)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Amazon\_HQIndex ~ MLB\_Team, data = citydata)  
##   
## $MLB\_Team  
## diff lwr upr p adj  
## YES-NO 2.355621 -2.333613 7.044855 0.3175427

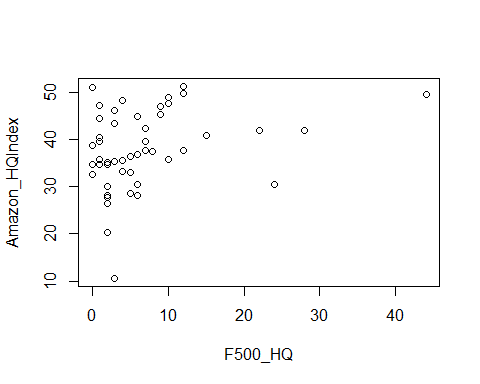
x <- lm(Amazon\_HQIndex~MLB\_Team, data = citydata)  
summary(x)

##   
## Call:  
## lm(formula = Amazon\_HQIndex ~ MLB\_Team, data = citydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.2178 -4.2530 0.2246 5.6994 14.4434   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.773 1.649 22.30 <2e-16 \*\*\*  
## MLB\_TeamYES 2.356 2.332 1.01 0.318   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.246 on 48 degrees of freedom  
## Multiple R-squared: 0.02081, Adjusted R-squared: 0.0004115   
## F-statistic: 1.02 on 1 and 48 DF, p-value: 0.3175

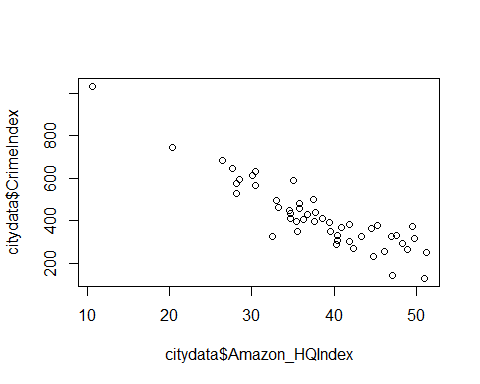
#P-value of .318, fail to reject the means are equal.   
  
  
#CityData Correlation Exploration  
plot(citydata)



plot(Amazon\_HQIndex ~ F500\_HQ, data=citydata)



plot(citydata$Amazon\_HQIndex, citydata$CrimeIndex)



cor(citydata$ZillowHVI,citydata$CollegeDegree, method="pearson")

## [1] 0.5926257

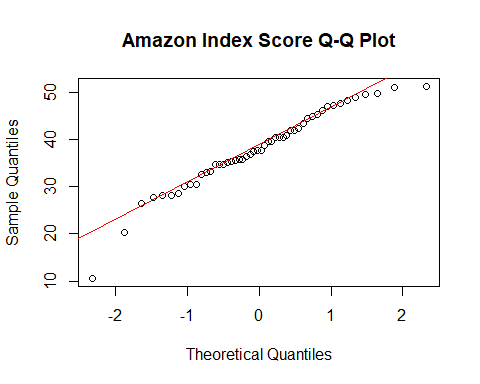
cor(citydata$Amazon\_HQIndex,citydata$Population, method="pearson")

## [1] 0.2112102

cor(citydata[,2:8], method="pearson")

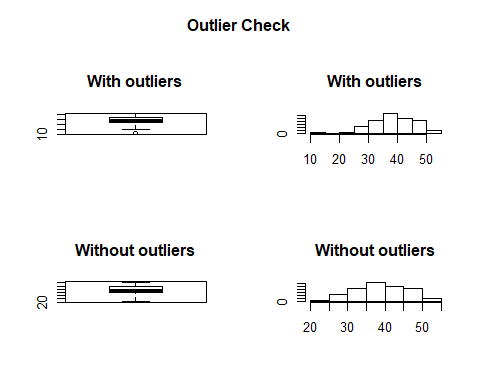
## ZillowHVI ZillowRent Population CollegeDegree F500\_HQ  
## ZillowHVI 1.0000000 0.9432573 0.2849783 0.5926257 0.23289064  
## ZillowRent 0.9432573 1.0000000 0.3836204 0.5799922 0.35392882  
## Population 0.2849783 0.3836204 1.0000000 0.1879175 0.85981750  
## CollegeDegree 0.5926257 0.5799922 0.1879175 1.0000000 0.31691112  
## F500\_HQ 0.2328906 0.3539288 0.8598175 0.3169111 1.00000000  
## CrimeIndex -0.2452589 -0.1768678 -0.0703023 -0.3547689 -0.09402744  
## Amazon\_HQIndex 0.4242073 0.3769843 0.2112102 0.6761967 0.28594755  
## CrimeIndex Amazon\_HQIndex  
## ZillowHVI -0.24525894 0.4242073  
## ZillowRent -0.17686775 0.3769843  
## Population -0.07030230 0.2112102  
## CollegeDegree -0.35476886 0.6761967  
## F500\_HQ -0.09402744 0.2859475  
## CrimeIndex 1.00000000 -0.9021359  
## Amazon\_HQIndex -0.90213592 1.0000000

qqnorm(citydata[,8], main="Amazon Index Score Q-Q Plot")  
qqline (citydata[,8], col=2)



The last city, Memphis is an outlier among the Amazon HQ Index. Will test for the outlier and remove the obs. Then will test the normality again which will yield a higher p-value

#Outlier Function  
attach(citydata)  
citydata.sort <- citydata[order(Amazon\_HQIndex, decreasing = TRUE),]  
detach(citydata)  
outlierKD <- function(dt, var) {  
 var\_name <- eval(substitute(var),eval(dt))  
 na1 <- sum(is.na(var\_name))  
 m1 <- mean(var\_name, na.rm = T)  
 par(mfrow=c(2, 2), oma=c(0,0,3,0))  
 boxplot(var\_name, main="With outliers")  
 hist(var\_name, main="With outliers", xlab=NA, ylab=NA)  
 outlier <- boxplot.stats(var\_name)$out  
 mo <- mean(outlier)  
 var\_name <- ifelse(var\_name %in% outlier, NA, var\_name)  
 boxplot(var\_name, main="Without outliers")  
 hist(var\_name, main="Without outliers", xlab=NA, ylab=NA)  
 title("Outlier Check", outer=TRUE)  
 na2 <- sum(is.na(var\_name))  
 cat("Outliers identified:", na2 - na1, "n")  
 cat("Propotion (%) of outliers:", round((na2 - na1) / sum(!is.na(var\_name))\*100, 1), "n")  
 cat("Mean of the outliers:", round(mo, 2), "n")  
 m2 <- mean(var\_name, na.rm = T)  
 cat("Mean without removing outliers:", round(m1, 2), "n")  
 cat("Mean if we remove outliers:", round(m2, 2), "n")  
 response <- readline(prompt="Do you want to remove outliers and to replace with NA? [yes/no]: ")  
 if(response == "y" | response == "yes"){  
 dt[as.character(substitute(var))] <- invisible(var\_name)  
 assign(as.character(as.list(match.call())$dt), dt, envir = .GlobalEnv)  
 cat("Outliers successfully removed", "n")  
 return(invisible(dt))  
 } else{  
 cat("Nothing changed", "n")  
 return(invisible(var\_name))  
 }  
}  
#test for outlier but will say NO to remove as we will remove it on the next line of code  
outlierKD(citydata.sort, Amazon\_HQIndex)



## Outliers identified: 1 nPropotion (%) of outliers: 2 nMean of the outliers: 10.56 nMean without removing outliers: 37.95 nMean if we remove outliers: 38.51 nDo you want to remove outliers and to replace with NA? [yes/no]:   
## Nothing changed n

#Remove Memphis  
citydata.sort.remove\_Memphis <-citydata.sort[-c(50),]  
hist(citydata.sort.remove\_Memphis$Amazon\_HQIndex)  
shapiro.test(citydata.sort.remove\_Memphis$Amazon\_HQIndex)

##   
## Shapiro-Wilk normality test  
##   
## data: citydata.sort.remove\_Memphis$Amazon\_HQIndex  
## W = 0.97927, p-value = 0.5349

#p value of 0.5349. Much different than the earlier .0841  
  
model.degree.sort <- aov(Amazon\_HQIndex~MLB\_Team, data = citydata.sort.remove\_Memphis)  
hist(model.degree.sort$residuals, main = "Hist of Res", xlab = "Residuals")  
summary(model.degree.sort)

## Df Sum Sq Mean Sq F value Pr(>F)  
## MLB\_Team 1 19.5 19.54 0.36 0.551  
## Residuals 47 2547.5 54.20

TukeyHSD(model.degree.sort)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Amazon\_HQIndex ~ MLB\_Team, data = citydata.sort.remove\_Memphis)  
##   
## $MLB\_Team  
## diff lwr upr p adj  
## YES-NO 1.263213 -2.969362 5.495787 0.551121

qqnorm(citydata.sort.remove\_Memphis[,8], main="Amazon Index Score w/o Outlier Q-Q Plot")  
qqline (citydata.sort.remove\_Memphis[,8], col=2)  
  
  
#Removing Memphis

