Resumen Capitulo 1

Daniel Villatoro

23/1/2020

#Motivaciones para analizar datos:

#Analisis exploratorio

EDA o analisis exploratorio de datos puede ser definido como el arte de mirar a uno o mas datasets en un esfuerzo para entender la subyacente estructura de datos contenida, esta es una usal descripcion de como podria ir acerca de lo que ofrecespor Diaconis:

Nosotros miramos los numeros o graficos e intentamos encontrar patrones, Nosotros perseguimos pistas sugestivas por un antecedente de informacion, imaginacion patrones percividos t experiencia con otros analisis de datos.

Los datasets pueden contener datos no numericos, nuestro analisis puede estar basado en caracteristicas numericas computadas desde los valores no numericos, las variables pueden tener varios niveles y mas retos problematicosy eso puede venir en dos variedades una es representar las variables como el codigo postal de Estados unidos cual identifica geograficamente las localizaciones , el segundo tipo de varios niveles categoricos describe la inerente structura de la variable que puede ser mostrada para un desarollo especializado en las tecnicas de analisis.

La mension de graficos es particularmente importante desde que los humanos hacen mejor el ver patrines en los graficos, un grafico es una larga coleccion de numeros, es por eso que R soporta una gran cantidad de diferentes graficos , metodos mostrados(scatterplots, barplots,boxplotsm quantile-quantile,plots,histograms,mosaics plots y mucho mas).Primero las tecnicas graficas son muy utiles para el analisis de datos, ya que encontramos importantes estructuras en los dataset que no son necesarios al explicar todas esas ecuentos con otros.Por ejemplo una largo array de dos variables en un grafico de dispersion podria ser una herramienta util para mostrar anomalias o variables en dato subytacente, pero serua extremadamente pobre para presentar resultados de otros por que esencialmente reequiere, que el observador repita el analisis por ellos mismos.

El segundo punto es la utilidad de cualquier grafico mostrado, puede depender fuertemente o exactamente, esta caracteristica depende de dos componentes, la mecanica de como un subset de datos puede ser mostrado y la eleccion de como ira adentro de ese subset de datos, Especificamente eso es importante, el segundo aspecto ya que es importante tener en cuenta de donde podrian venir los datos y quiza no podrian ser muy utiles para el analisis

El analisis exploratorio hace mas extensivo el uso de herramientas graficas

El analisis exoloratorio ofrece caracterizacuibes de naberas diferentes de klas variables y/o las fuentes de los datos y compara esas caracterizaciones

El analisis exploratorio son aspectos extremadamente importantes es la busqueda de lo inusual las anomalias que aparezcan en el dataset.

La cadena del analisis de software es Adquirir —> Analizar —> Explicar

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
```

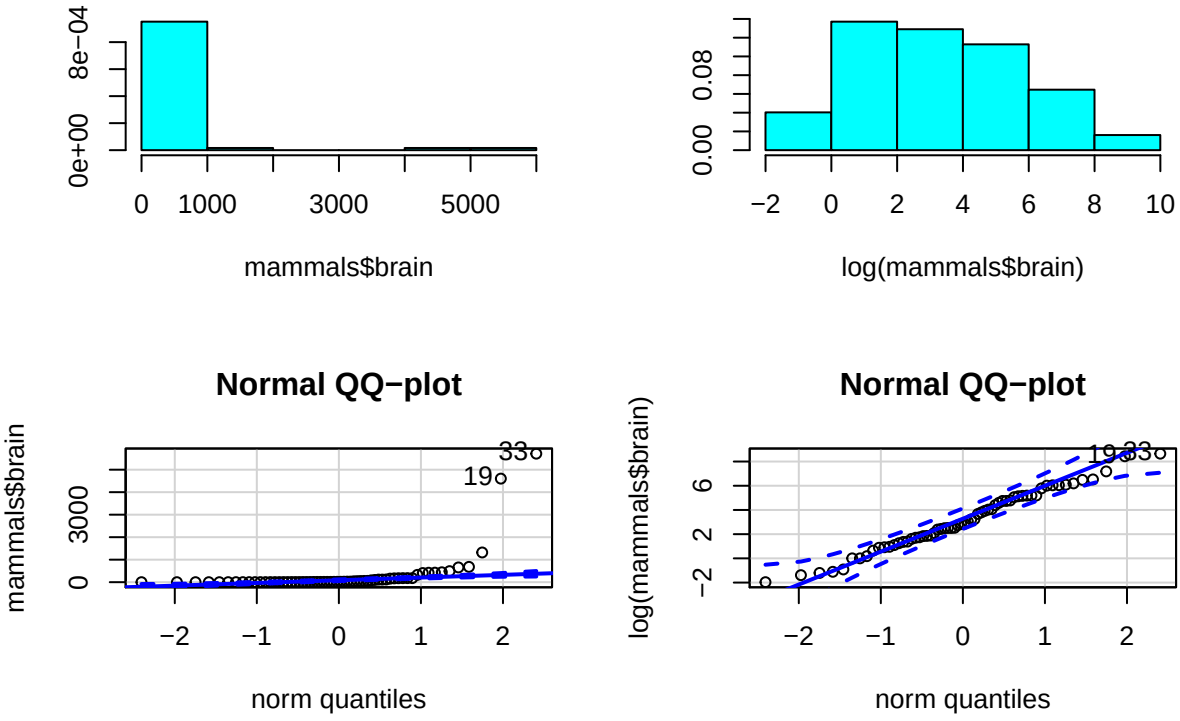
```
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

#Graficos

```
## Loading required package: carData
```

```
## [1] 33 19
```

```
## [1] 33 19
```



#Una representativa sesion en R

En un analisis de una sesion de R se realiza las siguentes preguntas

1.Cuantos records tiene el dataset contenido? 2.Cuantos campos o variables son incluidas en cada record" 3.Que tipos de variables son(numeros reales, integers, variables categoricas como ciudad, tipo o calgo por el estilo) 4.Todas esas variables pueden ser siempre observadas? 5.Todas esas variables incluidas en el dataset las unicas que es lo que se espera? 6.Esas variables en los dataset estan exibiendo el tipo de relacion que nosotros esperamos?

El Sr. Derek Whiteside de la Estación de Investigación de Construcción del Reino Unido registró el consumo semanal de gas y la temperatura externa promedio en su propia casa en el sureste de Inglaterra durante dos temporadas de calefacción, una de 26 semanas antes, y una de las 30 semanas después de que se aisló la pared de la cavidad instalada. El objetivo del ejercicio fue evaluar el efecto de la aislamiento en el consumo de gas.

Apartir de el dataset llamado whiteside que corresponde al consumo de gas se puede realizar una graficacion con la funcionplot, un antes y un despues obteniendose dos referencias lineales.Las dos referencias lineales encajan, en un modela para cada uno de los datos del subset definido por dos valores alternativamente podemos obtener los mismos resultados calzando una regresion lnear de un modelo de dataset, usando temp and insul como variables predictorias.

Cuantos records tiene el dataset contenido? Cuantos campos o variables son incluidas en cada record" Que tipos de variables son(numeros reales, integers, variables categoricas como ciudad, tipo o calgo por el estilo)

Todas esas variables pueden ser siempre observadas? Todas esas variables incluidas en el dataset las unicas que es lo que se espera? Esas variables en los dataset estan exibiendo el tipo de relacion que nosotros esperamos?

```
  head(whiteside)
```

```
##    Insul Temp Gas
## 1 Before -0.8 7.2
## 2 Before -0.7 6.9
## 3 Before  0.4 6.4
## 4 Before  2.5 6.0
## 5 Before  2.9 5.8
## 6 Before  3.2 5.8
```

```
  str(whiteside)
```

```
## 'data.frame':    56 obs. of  3 variables:
##  $ Insul: Factor w/ 2 levels "Before","After": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Temp : num  -0.8 -0.7 0.4 2.5 2.9 3.2 3.6 3.9 4.2 4.3 ...
##  $ Gas  : num  7.2 6.9 6.4 6 5.8 5.8 5.6 4.7 5.8 5.2 ...
```

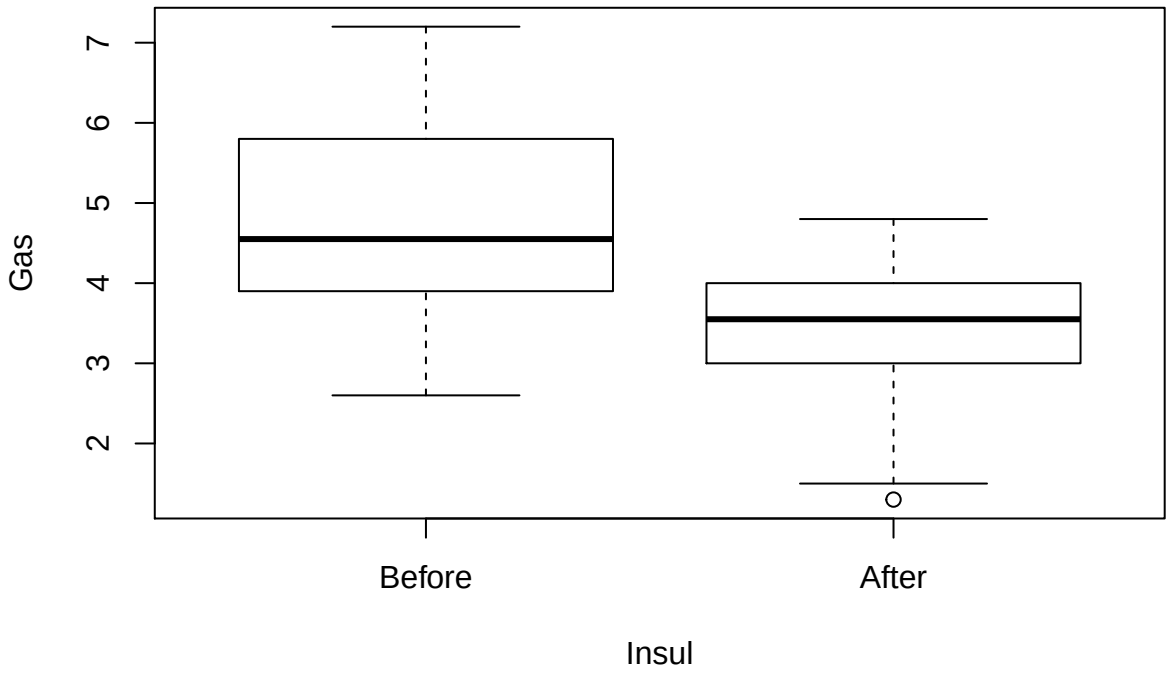
```
 x <- as.character(whiteside$Insul)
 str(x)
```

```
##  chr [1:56] "Before" "Before" "Before" "Before" "Before" "Before" "Before" ...
```

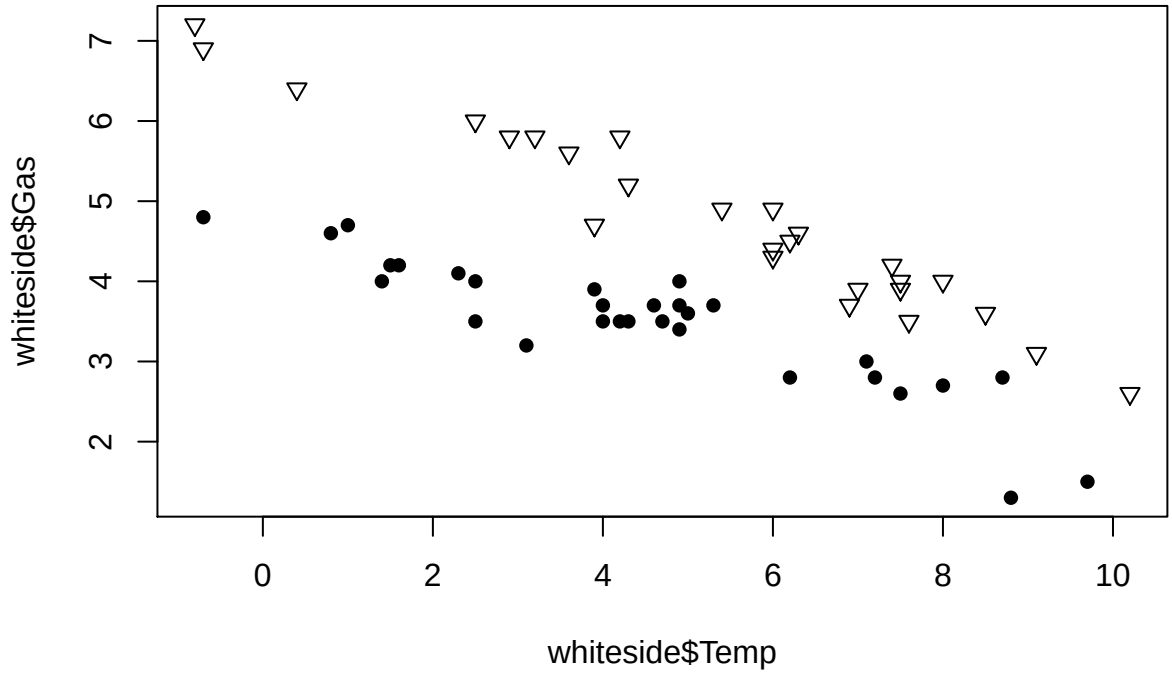
```
summary(whiteside)
```

```
##     Insul         Temp             Gas
##  Before:26   Min.   :-0.800   Min.   :1.300
##  After :30   1st Qu.: 3.050   1st Qu.:3.500
##              Median : 4.900   Median :3.950
##              Mean   : 4.875   Mean   :4.071
##              3rd Qu.: 7.125   3rd Qu.:4.625
##              Max.   :10.200   Max.   :7.200
```

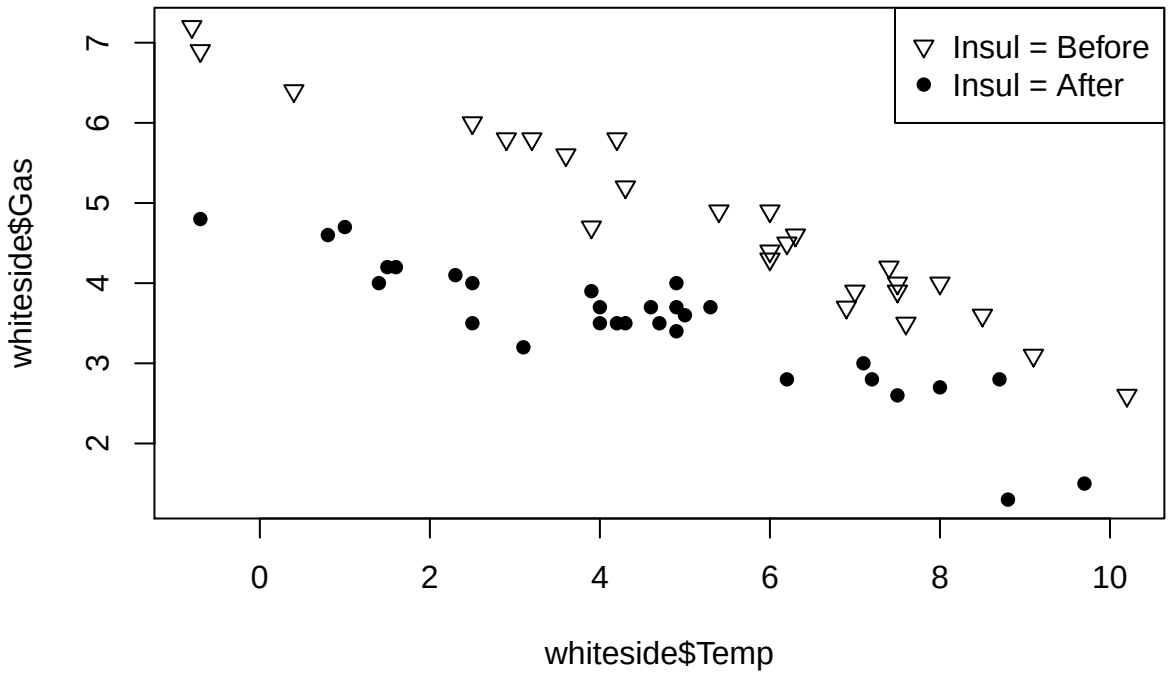
```
boxplot(Gas ~ Insul, data = whiteside)
```




```
plot(whiteside$Temp, whiteside$Gas, pch=c(6,16)[whiteside$Insul])
```




```
plot(whiteside$Temp, whiteside$Gas, pch=c(6,16)[whiteside$Insul])
legend(x="topright",legend=c("Insul = Before","Insul = After"), pch=c(6,16))
```




```
plot(whiteside$Temp, whiteside$Gas, pch=c(6,16)[whiteside$Insul])
legend(x="topright",legend=c("Insul = Before","Insul = After"), pch=c(6,16))
Model1 <- lm(Gas ~ Temp, data = whiteside, subset = which(Insul == "Before"))
Model2 <- lm(Gas ~ Temp, data = whiteside, subset = which(Insul == "After"))
abline(Model1, lty=2)
abline(Model2)
```