Capitulo3

Daniel Villatoro

23/1/2020

#Analisis de datos exploratorios

La basica nocion del analisis exploratorio fue introducido en el capitulo 1 empezando con la descripcion anteriormente mencionada el analisis exploratorio es el arte de mirar dentro de los datos en una manera cuidadosa y estrycturada, la inicial descripcion a sido seguida por un analisis de alto nivel con tecnicas usadas en analisis exploratorio.

Cuatro conceptos llave en la exploracion de datos

Revelacion

Residual

Re-expresion

Resistencia

Aqui el termino revelacion refiere ala visualizacion de los datos previamentre anotados son importantes en la parte exploratoria del analisis de datos. Una principal diferencia entre los modelos considerados de analisis exploratorios, en modelos predictivos es que esos modelos usados en el analisis exploratorio seguido son muy simples y aveces triviales.

En particular, mirando las diferencias entre los valores de datos individuales y una media o mediana pueden ser útiles en analisis de datos exploratorios, para el uso de comparacion.

La reexpresión, que se refiere a la Aplicación de transformaciones matemáticas a una o más variables. La utilidad de esta idea es una consecuencia del hecho de que los valores de datos que se nos dan analizar no siempre están en su representación más informativa.

La resistencia se refiere a la capacidad de una caracterización de datos para evitar la influencia indebida de valores atípicos u otras anomalías de datos,Un cuidadoso análisis exploratorio antes de intentar utilizar datos en explicar fenómenos, construir modelos predictivos o tomar decisiones es uno de las mejores formas de encontrar estas anomalías antes de que puedan afectar negativamente a nuestra resultados.

#Una estrategia General

Al explorar un nuevo conjunto de datos, la siguiente secuencia de preguntas básica suele ser útil: 1.Evaluar las características generales del conjunto de datos, por ejemplo:

¿Cuántos registros tenemos? Cuantas variables?. ¿Cuáles son los nombres de las variables? ¿Son significativos? ¿De qué tipo es cada variable, por ejemplo, numérica, categórica, lógica? ¿Cuántos valores únicos tiene cada variable? ¿Qué valor ocurre con mayor frecuencia y con qué frecuencia ocurre? ¿Faltan observaciones? Si es así, ¿con qué frecuencia ocurre esto?

2.Examinar estadísticas descriptivas para cada variable.

3.Cuando sea posible, ciertamente para cualquier variable de interés particular examine las visualizaciones exploratorias.

4.Mirar anomalías de datos;

5.Observe las relaciones entre variables clave utilizando las ideas descritas

6.Finalmente, resuma estos resultados en forma de un diccionario de datos, para servir como base para el posterior análisis y explicación de los resultados.

Tipos de variables en la practica

Una de las características clave de los datos incluidas en el resumen preliminar de datos. Como la mayoría de las otras plataformas de análisis de datos, R es compatible con un conjunto de tipos de variables predefinidos, incluidos numérico, de caracteres, lógico, y factores Desafortunadamente, estos tipos de variables básicas no siempre se devuelven por completo,al reflejar el "carácter real" de la variable.

Un ejemplo importante son las variables de fecha, que puede representarse en más de una forma, cada una de las cuales admite diferentes tipos de formatos: las fechas representadas como cadenas de caracteres como "15-nov-2015" son útil para lectores humanos, pero no proporcionan la base para calcular la cantidad de días entre esta fecha y otra.

Eso requiere una respuesta de representacion numericas, y tales representaciones están disponibles, pero no proporcionan la misma utilidad para lectores humanos.

Por lo tanto, a menudo se desarrollan formatos de fecha especiales, que permiten tanto la interpretación humana como los cálculos simples, desafortunadamente, estas representaciones a menudo se pierden en la transferencia de datos de su fuente original a una sesión R, lo que requiere por parte de nosotros reconocer explícitamente las variables de fecha representadas como cadenas de caracteres y volver a codificarlos en un formato de fecha especial.

Numericas contras Ordinal contra variables nominales

Muchos de los problemas de los analisis de datos concierne con los datos numericos el analisis de whiteside dataset es un caso en particular donde las variables primarias de interes donde el calor del consumo semanal de gas y la temperatura de afuera ambos son numeros, una importante caracteristica de los datos numericos es que se puede aplicar muchas operaciones matematicas que pueden ser calculadas, sumas promedioas potencias raices y otras muchas combinaciones o transformacions, esto es posible por las formas basicas de la estadistica descriptiva como la desviacion estandar junto con otras herramientas de datos de caracterizacion.

No todas las variables son numericas sin embargo y en la mayoria de las operaciones matematiocas no son aplicables las variables no numericas, denuevo whiteside dataset prove en la ilustracion the Insul variables es un ejemplo de categoria nominal, o factores de variables que pueden asumir esos dos valores cada uno representados con el caracter string Before y After

En ese caso nominal las variables como Insul ninguna de las operaciones matematicas mencionadas son aplicadas, nosotros no podemos calcular sumas diferencias productos raices potencias en esos valores de datos, Se necesitaria trabajar con variables nominales acerca de lo que podemos hacer es contar y comparar, haciendoce preguntas como las siguientes dadas en una categorica variable C:

Cuantas maneras distintas de valores o niveles hace la variable exibida?

Que tan seguido hace cada de esos niveles en ocurrir en el dataset?

Como es el comportamiento de otra variable x variando sobre los niveles de C?

Otra clase de variables que es, en muchos aspectos, intermedia entre las variables numéricas y nominales es la clase de variables ordinales, se refiere al orden categorico de variables o el orden de los factores. Esas variables asume los valores no numericos en un rango completo de operaciones matematicas y que no estan disponibles para que nosotros podamos trabajar con ellas pero poseen un orden inherente, por lo que podemos decir que un valor de la variable es "menor que" o "precede" a otro valor.

#Texto vs datos

Las variables categoricas tambien conocidas como nominal o ordinal son comunmente representadas como caracteres strings. Los requerimentois principales para esta representacion es que cada distinto nivel de vriables tengan un unico caracter de representacion stringy sean designados seguidamente en casos favorables

sean escogidos para proveer algunos grados para la interpretacion humana. Esto es generico no hay razon de Parse en estos caracteres string en substrings como variables subsecuentes en el analisis de modelados, en contraste los datos de texto consiste en el caracter de strings seguidos por un caracter string de una gran longitud, conllevando una mas compleja informacion con multiple caracteristicas de datos e informacion relacionada entre diferentes entidades de interes. Es es una ventaja para el parse de los caracteres string dentro de componentes, tambien para detectar nuevas caracteristicas en los datos, similitudes o diferencias entre records que no resultan obias de otras varuables, o para el uso de componentes como nuevas covariables en la construccion de modelos predictivos.

#Resumiendo los datos numericos

De hecho, estas dos caracterizaciones son probablemente la estadística descriptiva más común y, en algunos aspectos, la más importante: la media intenta darnos un "valor típico" para una variable numérica, mientras que la desviación estándar intenta transmitir una idea de la "dispersión" o "dispersión" de la observaciones de datos individuales en torno a este valor típico.

En términos más generales, las estadísticas descriptivas suelen ser caracterizaciones independientes del tamaño, que tienen el mismo formato e interpretación para conjuntos de datos pequeños que para conjuntos de datos grandes.

Sin embargo, debido a que son resúmenes simples, las estadísticas descriptivas tienen un poder limitado para describir datos, y en casos desfavorables pueden ser engañosas.

#Anomalias en los datos numericos

Corresponde a los errores que pueden ser encontrado en los metadatos y son potencialmente problemas catastroficos, se debe estar al tanto de los temas de la sistematica de datos extraviados y distinguir los datos extraviados, esa anotacion es representada como un tipo particular de datos.

#Los valores atipicos y su influencias

La definicion adoptada aqui se refiere a la observacion donde podria aparece inconsistencias como recordatorio de ese data set en particular.

#Detectando atipicos univariables

El termino deteccion atipica univariable se refiere al prceso de identificacion de atipicos en una sola variable, en el caso de inconsistencias con recordatorio en los datos es tipicamente interpretado para decir usualmente lejos o cerca de los datos tipicos. Esta idea puede convertirse automaticamente en un procedimiento de deteccion atipica definido de manera calculada como valores tipicos por diferentes medidas para tipico o dispersion de datos y los limites inusuales que obtendremos de diferentos procedimientos automaticos de valores atipicos.

La regla de edicion Tres Sigma

Probablemente la mejor manera de conocer automaticamente la deteccion de procedimientos atipicos es la regla de edicion tres sigma, se conoce en la literatura estadistica de diferentes nombres como la desviacion estudiada identificada, se basa en la observacion por la aproximacion de datos Gausianos se basa en tres puntos.

Los valores tipicos de los datos es x.

Los datos esparcios es la desviacion estandar.

Los limites inusuales es t = 3 desviacones estandar.

#Algunos consejos practios en la deteccion atipicas.

Aplica a todos los tres procedimientos de deteccion atipica en tu secuencia de datos y cuidadosamente examina los reusltados comprando con el numero atipicos detectados por cada procedimieto, los valores de los datos declarados y el rango de los daros de valores no declarados para los atipicos

Si es posible realiza una aplicacion especifica de test rasonables para que ambos limites atipicos y los valores de los rangos no atipicos sean un rango nominal estando razonable y hacer que los valores perifericos parezcan lo extremadamente sospechosos.

Examinar las graficas de datos con los limites atipicos indicados en la grafica, o con los puntos perifericos marcados en diferentes forma de puntos colores o resaltadores los cuales pueden ser determinados como sospechosos en el procedimiento de deteccion de atipicos.

#Inliers y su deteccion

El termino Inliers se refiere a puntos nominales en una secuencia de datos, o los puntos no perifericos, en el libro se se refiere a estos terminos de dos maneras diferentes la primera la define como los valores de los datos dejados en el interior de las distribucion estadistica de distribucion y es el error, una de las fuentes de los puntos nominales es el disfrazado de datos perdidos.

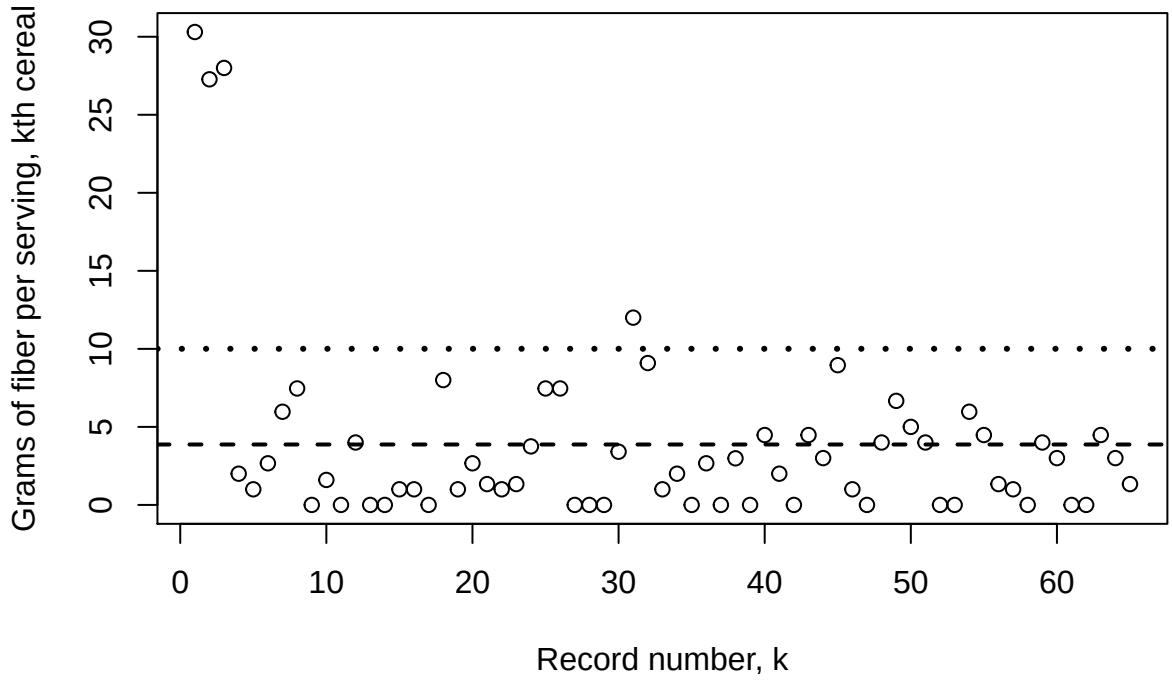
#Erroes en los metadatos

Idealmente los metadatos inclutyen detalladas definiciones de las variables, en rangos de valoresa dmisibles o numeros de observaciones no registradas y las notaciones indicadas para utilizarlas con cualquier otra características o peculiaridades notables. Normalmente, sin embargo, los metadatos son mucho menos completos de lo que nos gustaría, y es difícil mantener el control de calidad porque los metadatos están altamente desestructurados.

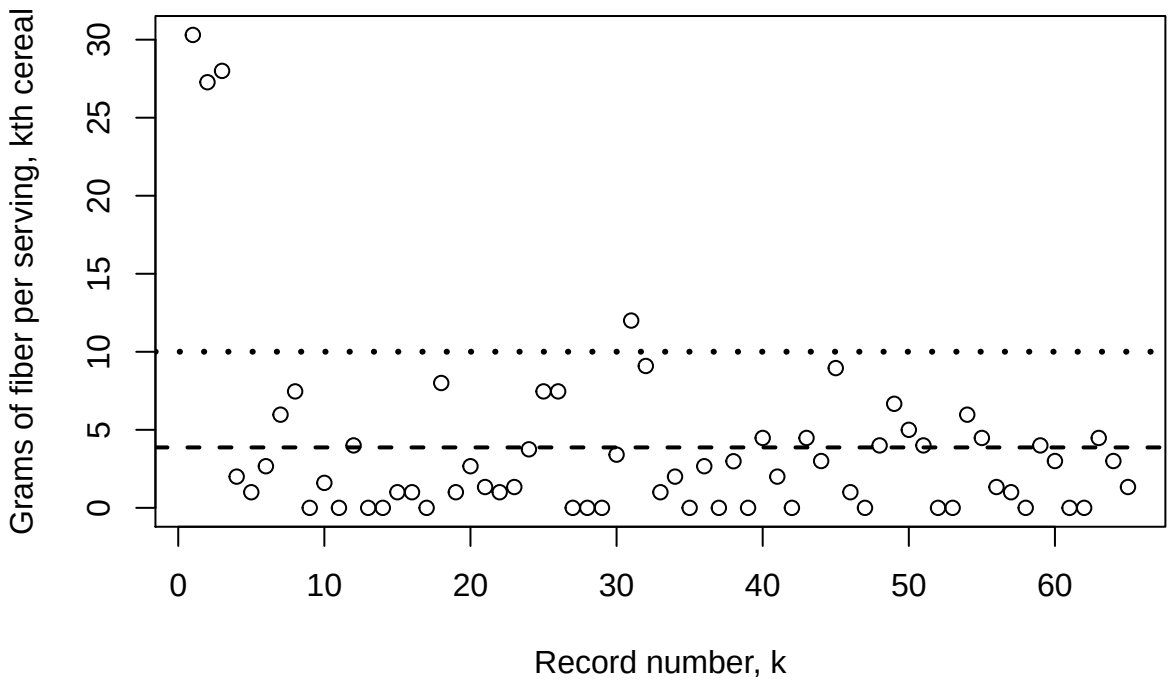
#Datos perdidos posiblemente disfrazados

Varios aspectos de los problemas de los datos perdidos es que no son unicas representaciones universales de ellas, la falta de datos a menudo se supone que es un fenómeno aleatorio. El supuesto de trabajo es que los valores faltantes ocurren completamente al azar.

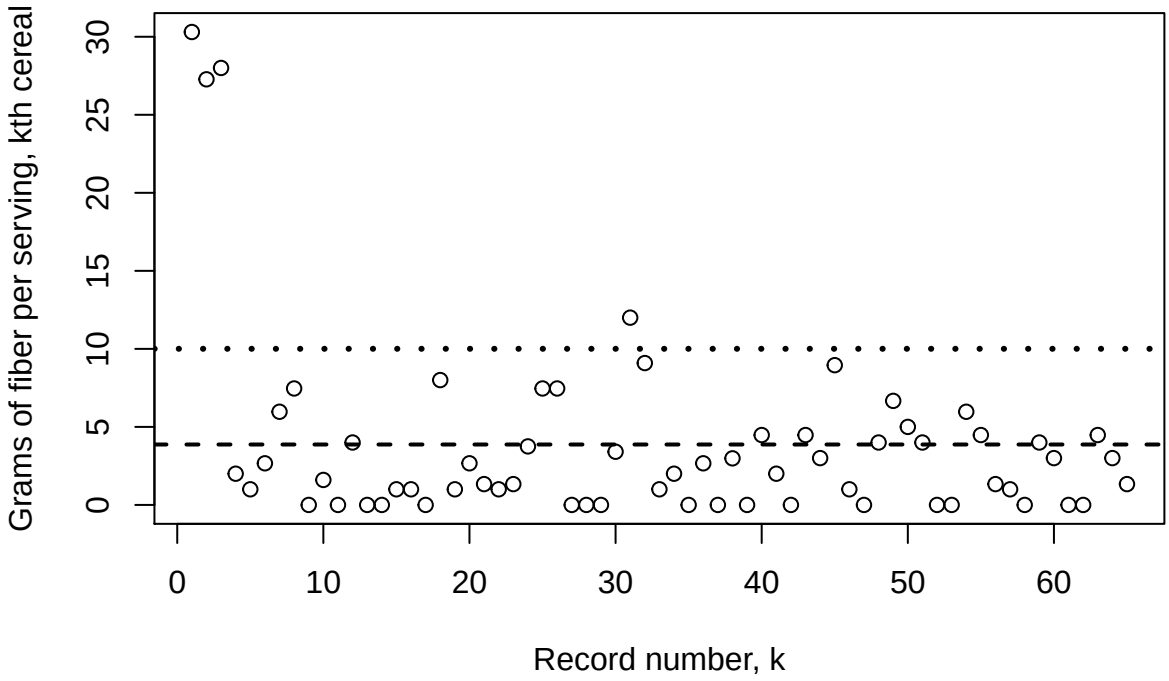
```
library(MASS)
x <- UScereal$fibre
plot(x, xlab="Record number, k",
ylab="Grams of fiber per serving, kth cereal")
abline(h = mean(x), lty=2, lwd=2)
abline(h = mean(x) + sd(x), lty = 3, lwd = 3)
```



```
library(MASS)
x <- UScereal$fibre
plot(x, xlab="Record number, k",
ylab="Grams of fiber per serving, kth cereal")
abline(h = mean(x), lty=2, lwd=2)
abline(h = mean(x) + sd(x), lty = 3, lwd = 3)
```




```
library(MASS)
x <- UScereal$fibre
plot(x, xlab="Record number, k",
ylab="Grams of fiber per serving, kth cereal")
abline(h = mean(x), lty=2, lwd=2)
abline(h = mean(x) + sd(x), lty = 3, lwd = 3)
```




```
outlierIndex <- which(UScereal$fibre > 25)
rownames(UScereal)[outlierIndex]
```

```
## [1] "100% Bran"                "All-Bran"
## [3] "All-Bran with Extra Fiber"
```

```
set.seed(333)
x <- sort(rnorm(200))
mean(x)
```

```
## [1] 0.07512683
```