Se hacen una cuenta de cloud, Nuevo proyecto y se importa la base de la eph, ver si los datos para demography hace falta que yo les de los datos o si los bajan del sitio de Vladimir

Esta reunida una intro con ggplot, falta agregarle demography para ggplot, inspirarse en el módulo 4 para ejemplificar con EPH

Y http://www.cookbook-r.com/Graphs/ Está muy organizado por temas

Estimador de Theil-sen

p1<-ggplot(asalariados.con.ingreso.y.horas.cordoba, aes(PP3E\_TOT, PP08D1))+

geom\_point()

p1+geom\_smooth()

p1+ geom\_smooth(method = "lm")

# Un mejor modelo podría ser usar el estimador de Theil-Sen, que es la mediana de todas las pendientes entre dos puntos distintos está en el paquete mblm.

library(mblm)

TheilSen <- function(..., weights = NULL) {

mblm::mblm(...)

}

p1+geom\_smooth(method = "TheilSen")

Sobre R

https://nadaesgratis.es/fernandez-villaverde/r

RStudio

http://www.unavarra.es/personal/tgoicoa/ESTADISTICA\_RMarkdown\_tomas/basicRmarkdown/index.html

Ggplot:

Markdown

http://www.unavarra.es/personal/tgoicoa/ESTADISTICA\_RMarkdown\_tomas/basicRmarkdown/index.html

Open science

https://ropensci.org/

verificar que en cloud ande markdown

puede ser ultimo día que trabajen sobre sus datos

a veces gace falta especificar como leer una base: por ejemplo en ectss el CON es numérico, pero no debe leerse así

un solo documento Rmd:

1. operaciones básicas, incluye recodificación
2. matriz de datos y medidas resumen
3. gráficos
4. demography
5. Markdown

|  |  |  |
| --- | --- | --- |
| Tema | Material | aplicación |
| El Entorno de Desarrollo Integrado (IDE) *RStudio*. Operaciones elementales del paquete *base*. Los objetos R. | Curso filmado sobe la base de "operaciones", del libro, cambiando por datos eph: fusionar "primer script" con ""materiales y herramientas" | EPH |
| Lectura desde diferentes formatos: paquetes *foreign* y *readxl*. |  |
| Exploración y limpieza de una base. Recodificación de variables |  |
| Análisis descriptivo: tablas, medidas resumen. |  |
| Visualización: el paquete *ggplot2*. | Hay algo en el curso visualización |  |
| Datos demográficos: el paquete *demography*. | Esta avanzado |  |
| Confección de informes: *RMarkdown*. | Básico, es corto |  |

## Modalidad de cursado

Se trata de un seminario taller, en el que los participantes trabajarán sobre sus propias computadoras. En cada encuentro se presentarán los contenidos y se aplicarán los procedimientos sobre los mismos conjuntos de datos. Luego, los cursantes reproducirán los análisis con otros datos, que pueden ser propios o provistos en el curso.

## Evaluación

El curso se aprueba con un documento de análisis de datos, presentado en formato RMarkdown que incluya: objetivos, fuente(s) de datos, metodología de análisis, conclusiones.

Se hacen una cuenta de cloud, Nuevo proyecto y se importa la base de la eph, ver si los datos para demography hace falta que yo les de los datos o si los bajan del sitio de Vladimir

Creating plots in R using ggplot2. http://t-redactyl.io/archives.html

part 1: line plots

part 2: area plots

part 3: bar plots

part 4: stacked bar plots

part 5: scatterplots

part 6: weighted scatterplots

part 7: histograms

part 8: density plots

part 9: function plots

part 10: boxplots

part 11: linear regression plots

Agregar la ggplot cheatsheet

probando ggplot con eph 2-2015

file.choose()##ver donde está

eph215=read.spss("C:\\Documents and Settings\\Usuario\\Escritorio\\t215\_sav\\Individual\_t215.sav")

eph215=as.data.frame(eph215)##llevarlo a df

summary(eph215.1$p47t)

eph215.1=subset(eph215,eph215$p47t>0 & p47t<20000)##recorto p47

uno=ggplot(eph215.1, aes(p47t))

uno+geom\_histogram()

dos=uno+geom\_histogram(aes(fill=ch04))

dos+coord\_flip()

dos+scale\_x\_continuous(limits = c(200,20000))

tres=ggplot(eph215.1, aes(x=ch04, y=p47t))

tres+geom\_boxplot()

tres+geom\_boxplot(aes(fill=ch04))

eph215.1$sexo=eph215.1$ch04

levels(eph215.1$sexo)=c("varones", "mujeres")

eph215.1$ingreso=eph215.1$p47t

tres=ggplot(eph215.1, aes(x=sexo, y=ingreso))

tres+geom\_boxplot()

tres+geom\_boxplot(aes(fill=sexo))##muy bien, lo mismo con mejores nombres

###

cuatro=ggplot(eph215.1, aes(nivel\_ed,ingreso))

cuatro+stat\_summary(fun.y = mean,

fun.ymin=function(x) mean(x)-sd(x),

fun.ymax=function(x) mean(x)+sd(x),

geom="pointrange")##bien, salvo el orden de las categorías

cuatro+stat\_summary(fun.y = mean,

fun.ymin=function(x) mean(x)-sd(x)/sqrt(length(eph215.1$ingreso)),

fun.ymax=function(x) mean(x)+sd(x)/sqrt(length(eph215.1$ingreso)),

geom="pointrange")##se achicaron los intervalos y quedan solo puntos

##falla el orden de las categorías

eph215.1$educacion=as.character(eph215.1$nivel\_ed)##la defino como carácter

table(eph215.1$educacion)##ahora el orden es alfabético

eph215.1$educacion=factor(eph215.1$educacion, levels=c("Sin instrucción",

"Primaria Incompleta (incluye educación especial)",

"Primaria Completa","Secundaria Incompleta",

"Secundaria Completa","Superior Universitaria Incompleta",

"Superior Universitaria Completa"))

##con esto volvimos a factor, en el orden que quiero, a ver...

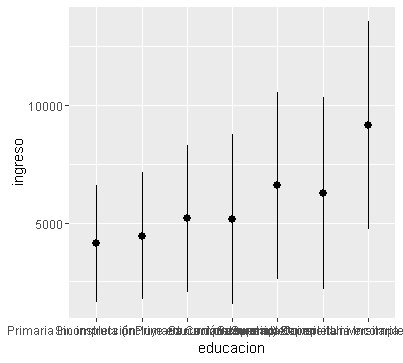
cinco=ggplot(eph215.1, aes(educacion,ingreso))

cinco+stat\_summary(fun.y = mean,

fun.ymin=function(x) mean(x)-sd(x),

fun.ymax=function(x) mean(x)+sd(x),

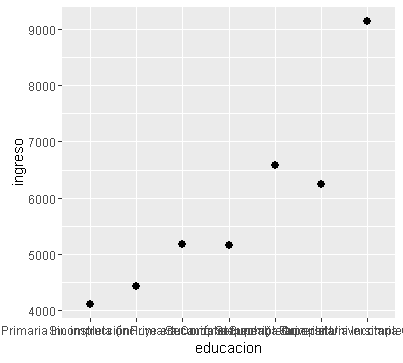
geom="pointrange")##great



cinco+stat\_summary(fun.y = mean,

fun.ymin=function(x) mean(x),

fun.ymax=function(x) mean(x),

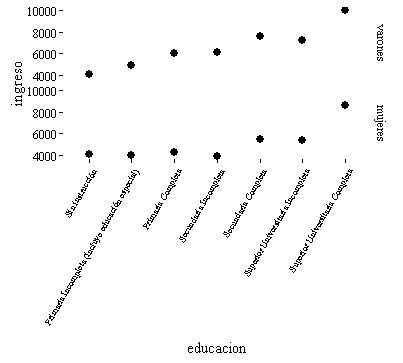
 geom="pointrange")##requiere los limites de y, así se logra quitar los errores

Faceteado y formateado:

siete=seis+facet\_grid(sexo~.)

siete+ theme\_tufte()+ theme(axis.text.x=element\_text(angle = 60, vjust=1, size = 7,

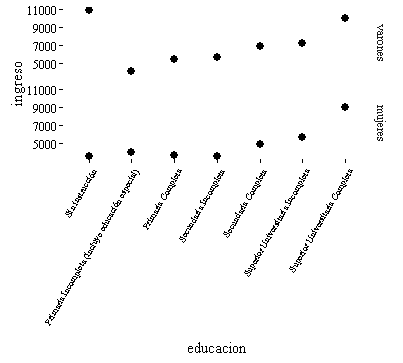
hjust =1 ))



En vez de facetear, hay que probar usar color para sexo, ver redactil como cambia la posición de la leyenda y los colores...

repetimos con

cba=subset(eph215.1,aglomerado=="Gran Córdoba")



Muy raro, la media de los varones sin educación es de 11000, lo muestra:

aggregate(cba$ingreso,list(cba$educacion,cba$sexo),mean)

Verificar con infostat: es 2-2015, cordoba

Ver esto, que viene de Rcommander

with(eph214.trim, plotMeans(p47t, educacion, error.bars="se"))

plotMeans??? que paquete?

La construcción de los gráficos en ggplot se hace por medio de capas que se van agregando. Las capas tienen cinco componentes:

* Los datos, que es la base de donde provienen la variables que se van a graficar. Si más tarde se grafica lo mismo para otra base, solo se debe cambiar ese origen, lo mismo si la base se modifica.
* Un conjunto de mapeos estéticos (“aes”), que describen el modo en que las variables de la base van a ser representadas en las propiedades estéticas de la capa.
* El “geom”, que describe la figura geométrica que se va a usar para dibujar la capa.
* La transformación estadística (“stat”) que opera sobre los datos originales para sintetizarlos de modo que se los pueda representar.
* Los  ajustes de posición

**Creación de un gráfico**

La primera instrucción para crear un gráfico es ggplot(). Esta instrucción va a tener dos argumentos: los datos y el mapeo estético (esta información puede ser omitida aquí, si se la indica en alguna de las capas que se agreguen).

Para iniciar un gráfico de la variable sexo, esa instrucción tiene la forma

ggplot(asalariados, aes(sexo))

El segundo argumento es la “estética”, que aquí solo contiene la variable que nos interesa. Pero si se ejecuta esta instrucción (Ctrl+r), no hay efecto. Hasta que no se agregue una capa, no habrá nada para mostrar, salvo los ejes coordenados con sus nombres.

Una capa mínima, puede especificar solamente el “geom”, es decir, la forma que pedimos a la representación. Por ejemplo, si queremos un gráfico de barras, se agrega (como una capa) un geom:

ggplot(eph.215, aes(sexo))+geom\_bar()

En el documento geoms, se exploran diferentes manera de graficar la relación entre los ingresos salariales y el nivel de educación

La relación entre geom y stat viene, en general dada por defecto. Según el tipo de objeto gráfico que se haya elegido, hay una transformación  establecida, por ejemplo, para el gráfico de barras, se cuenta la frecuencia absoluta. Vea el documento geoms y stats para un ejemplo.

 A continuación se ve la generación de un gráfico de barras simples

Reproducción Vídeo

**Mapear y fijar**

Una diferencia muy importante en la relación entre características gráficas y variables es la que hay entre mapear y establecer (o fijar).

**Mapear** es vincular valores de una variable a atributos estéticos del gráfico, como el color, la forma, o el tamaño, según qué gráfico sea. En geom\_point(aes(size="edad")) se dibujan puntos cuyo tamaño es proporcional a los valores de la variable edad.

**Fijar** es establecer una atributo en un valor predeterminado para todo el gráfico. Las expresiones size=3 o color="red" fijan el tamaño en el valor 3 o el color en rojo, sin tener en cuenta alguna variable.

Para mapear, la instrucción debe ir dentro de la estética (aes), mientras que para fijar, va fuera. Los dos ejemplos siguientes muestran la diferencia.

### Mapear la edad al tamaño y el sexo al color de los puntos:

ggplot(df)+geom\_point(aes(sexo, edad, size=edad, color=sexo))

### Establecer el tamaño y el color de los puntos:

ggplot(df)+geom\_point(aes(sexo,edad), size=5, color="green")

Si se fija un valor para el atributo, pero se lo ubica dentro de la estética, el resultado no es comprensible. Para los mismos ejemplos de arriba, la instrucción

ggplot(df)+geom\_point(aes(sexo,edad, color="green"))

es contradictoria, porque mapea el color (dado que está dentro de la estética) a un valor fijo. El resultado es:

Se ha mapeado (no fijado) el color al valor "green", con lo que se creó una nueva variable que solo tiene la categoría "green" y el color es mapeado a esa variable, como tiene una sola categoría, la leyenda es incomprensible.

El conjunto de colores disponibles en ggplot y sus nombres  son los siguientes:

**Avanzar en la estética del gráfico**

Las capas dan opciones para mejorar el aspecto de las representaciones. El paquete ggthemes tiene varias opciones de tema predeterminado, algunos que imitan los que usan publicaciones conocidas, como The Economist o el Wall Street Journal. En [este artículo](https://cran.r-project.org/web/packages/ggthemes/vignettes/ggthemes.html) de Jeffrey Arnold se ven los temas y las escalas de color que provee este paquete.

A continuación mostramos la construcción de un gráfico que ilustra las diferencias salariales por sexo, según nivel de educación. Está hecho a partir de datos de 2004, lo interesante es que una vez que se logra un gráfico satisfactorio, la misma sintaxis puede aplicarse a las bases de los diferentes años para obtener la serie.

[Diferencias salariales](https://youtu.be/WTrE62ULh_I)

Las mismas representaciones puede hacerse con los ingresos por hora, a fin de mejorar la comparación, porque así se elimina el efecto de la eventualmente distinta cantidad de horas que podrían trabajar varones y mujeres.

[Comparación por salarios-hora](https://youtu.be/UCaLTZmXwv4)

Estos ejemplos son para el primer trimestre de 2004, a fin de lograr los mismos resultados para otros períodos, solo debe cambiarse la base que se lee al comienzo, porque las variables tienen el mismo nombre.