**Clase 10. Herramientas de visualización**

**Visualización de datos**

## Introducción

Los datos son un **arma de doble filo**. A la hora de sacarles provecho, pueden utilizarse para hacer el bien, pero también para engañar y desinformar. Es importante recordar en este momento el Principio de Peter Parker: “**Con un gran poder viene una gran responsabilidad**”. Más allá de lo que digan los superhéroes, es cierto que tenemos que ser **fuertemente responsables en el uso de los datos**. Veamos este concepto.

En primer lugar, ¿cómo pensar en responsabilidad si lo único que estoy haciendo es analizar datos? La respuesta está en la **salida del proceso**. Los datos son fundamentalmente **insumos para la toma de decisiones**. En este sentido, pueden **comunicarse en forma directa** o utilizarse como **entrada de algoritmos** y desarrollos de ciencia de datos.

Para el primer caso, y en todos los casos en los que debamos realizar una comunicación de nuestros resultados, debemos echar mano de las prácticas y herramientas de la **visualización de datos.** La visualización es mucho más que seleccionar un tipo de gráfico al azar y ubicarlo sobre una hoja de cálculo. Es, más bien, un área del conocimiento orientada a la **preparación detallada para la presentación efectiva de la información recolectada**. Un gráfico siempre debe **decir mucho con una economía de elementos**, llevando al mundo real la conocida máxima que enuncia que “una imagen vale más que mil palabras”. Ciertamente, un gráfico debe valer por lo menos por la cantidad de datos que pretende mostrar. Y por supuesto, debe ser siempre certero y fiel a estos datos.

Para profundizar sobre el conocimiento del tema visualización de datos, recomendamos los excelentes libros de uno de los estudiosos más importantes de la materia, Edward Tufte, en particular “The visual display of quantitative information”, todos ellos disponibles para su venta en librerías y en su sitio personal <https://www.edwardtufte.com/tufte/>

Para lo que sigue en esta clase, tengamos presente lo visto en la clase anterior, dado que hay una **conexión muy estrecha entre la definición de una variable y su visualización**.

## Elementos de un gráfico

Un buen gráfico debe ser **claro y conciso** para lograr comunicar efectivamente los datos que lo conforman. Siempre debe contar como mínimo con los siguientes elementos:

* Un **título** conciso y representativo, que indique claramente qué información se pretende mostrar con el gráfico
* Una indicación concreta y precisa de las **variables involucradas**, en especial qué nombre tiene cada variable y en qué unidades de medición se expresa.
* Si hubiera distintas **categorías o variables** juntas, una enunciación y representación clara y bien diferenciada de cada una

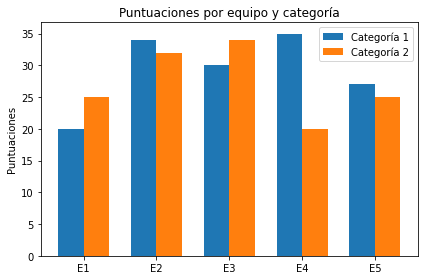
Con estos elementos, tenemos el terreno preparado para colocar los datos propiamente dichos. Si los datos se ubican en un gráfico adecuado, entonces servirán para comunicar efectivamente la información relevada.

Análisis básico de los datos

## Gráficos de barras

La operación más básica que puede realizarse con un conjunto de datos es **contarlos**, responder a la pregunta de “¿cuántos hay?”, considerándolos en conjunto, o bien con la posibilidad de desglosarlos en categorías, si hubiera (“¿cuántos hay de cada tipo?”). Otra operación muy básica es la de simplemente **mostrar valores puntuales** asociados a una categoría, por ejemplo el monto total de ventas por línea de producto.

Para estos fines utilizamos los **gráficos de barras**. Los gráficos de barras se utilizan con **datos categóricos**, y consisten en un eje horizontal donde se ubican las categorías y una barra por cada una de ellas cuya altura en el eje vertical **representa la cantidad o valor** de los elementos de la categoría en cuestión.



En la figura anterior, tenemos las puntuaciones de 5 equipos, del E1 al E5, separados por categorías, nombradas como como Categoría 1 en color azul y Categoría 2 en color anaranjado.

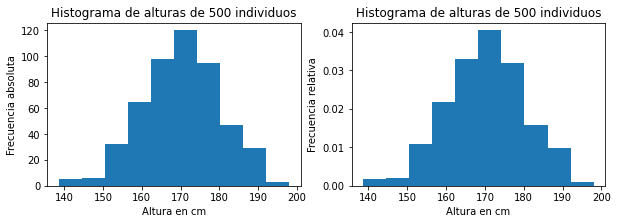
Recomendación: **No usar gráficos de barras para representar datos en el tiempo** porque las relaciones y tendencias son más difíciles de visualizar. Veremos cómo representar datos en el tiempo más adelante en la sección de Series de Tiempo.

## Histogramas

Veremos aquí los fundamentos de uso para los histogramas, a los que introdujimos brevemente como método de la librería matplotlib en la Clase 5. Los histogramas **muestran las frecuencias de aparición para cada intervalo de valores de la variable**, por eso son también denominados **histogramas de frecuencia**. Los histogramas son, en su forma, muy similares a los gráficos de barras, pero muy diferentes en su comportamiento. Un histograma **muestra una distribución de datos**, tal como mencionábamos en la clase anterior.

Consideremos los datos correspondientes a las alturas de 500 personas, medidas en centímetros. Para armar un histograma, organizaremos las mediciones de altura disponibles en intervalos, también llamados bins. Para trabajar con intervalos se procede de la siguiente forma: se toman todos los valores de la variable, desde el mínimo al máximo y se divide el rango correspondiente en intervalos de igual tamaño. Existen distintas maneras de elegir cuántos intervalos utilizar, pero generalmente los lenguajes de programación ya proveen un valor calculado por defecto que sirve adecuadamente para nuestro análisis. En Python, podemos repasar lo visto en la clase 5 para el método hist de matplotlib, donde la cantidad de intervalos se define con el parámetro bins. De esta forma, un histograma muestra **la importancia de cada intervalo de valores con respecto al total de datos**.

Una vez definidos los intervalos, ya estamos en condiciones de presentar un histograma. Vamos a nuestro ejemplo: la figura siguiente muestra los histogramas para las alturas de 500 individuos.



En el eje horizontal, se muestra **la variable a ser analizada**. Cada intervalo de valores de la variable tiene una altura que representa la cantidad de observaciones en dicho intervalo. Esta cantidad puede ser el resultado de un **conteo simple: “¿Cuántas unidades hay para cada intervalo de la variable?”**, o bien pueden expresarse como una **proporción con respecto al total: “¿Cuánto del total representa cada intervalo de la variable?”**. En este caso, **la cantidad de cada intervalo** **se divide por el total de observaciones**, con lo que se obtienen las proporciones mencionadas. En el primer caso hablaremos de **histogramas de frecuencias absolutas**, mostrado en el gráfico de la izquierda. En el segundo caso hablaremos de **histogramas de frecuencias relativas**, mostrado en el gráfico de la derecha. La forma es la misma, lo que cambia es la interpretación de los resultados: en el primer caso hablamos de **cantidad de observaciones para cada intervalo de la variable**, en el segundo caso hablamos de **proporción de veces de cada intervalo de la variable con respecto al total de observaciones de la variable**.

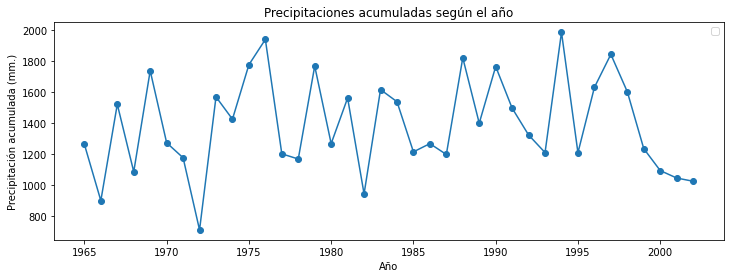
De acuerdo a lo visto en la clase anterior, es claro que esta distribución se aproxima mucho a la normal, con los valores acumulándose en mayor cantidad alrededor de un promedio, y en menor cantidad y más dispersos a medida que nos alejamos del promedio (alrededor de los 170 centímetros). El histograma **brinda una idea de la forma de la distribución**, lo que ayudará a extraer conclusiones útiles sobre los datos. Recomendamos usar el histograma como **una de las primeras herramientas de análisis para las variables cuantitativas**.

Relaciones entre elementos

Muchas veces, una variable tiene relaciones intrínsecas entre sus valores, o bien queremos verificar si una variable tiene relación con otra u otras. En esta sección veremos dos herramientas gráficas básicas para verificar relaciones: la serie de tiempo y los diagramas de dispersión.

## Series de tiempo

Siempre que tengamos una variable que tiene **algún tipo de evolución a través del tiempo**, o **donde organizar sus valores a lo largo del tiempo tiene sentido para el análisis**, podemos graficar una **serie de tiempo**. Para este tipo de gráfico dibujaremos un punto cuya altura representará el valor de la variable, y cuya posición con respecto al eje horizontal representará el momento en el tiempo en el que se mide esa variable. Luego, uniremos los puntos en forma secuencial. Consideremos el siguiente ejemplo, tomado de la Clase 5.

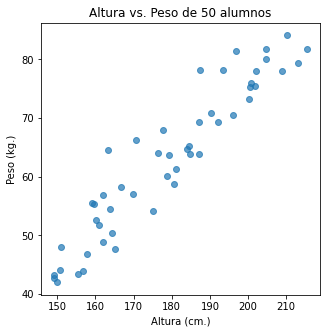


Recordemos que este gráfico mostraba la cantidad de precipitaciones acumuladas, medidas en milímetros, por año calendario. La altura de cada punto representa la cantidad de precipitaciones y su ubicación en el eje horizontal representa su año de medición. Las líneas que conectan los puntos nos ayudan a **percibir rápidamente si existe alguna tendencia, repetición de valores o cualquier otra característica** para empezar a conocer el comportamiento de los datos en el tiempo.

Recomendamos fuertemente **usar este tipo de gráfico para mostrar la evolución de una serie de datos, realizarla siempre con puntos y no con barras, y que el tiempo esté siempre en el eje horizontal, avanzando de izquierda a derecha**. Esto parece muy obvio, pero no siempre se tiene en cuenta y es fundamental para una buena práctica en datos afectados por el tiempo.

## Diagramas de dispersión. Nociones de correlación

Cuando queremos analizar la **relación entre dos variables**, queremos ver qué pasa cuando las observamos en conjunto. Por ejemplo, consideremos el caso de la Clase 5 acerca de los pesos en kilogramos y alturas en centímetros para un conjunto de 50 alumnos.



Cada punto representa un alumno. Para cada alumno podemos ver la intersección del eje horizontal, que indica las alturas, con el eje vertical, que indica los pesos.

Este es un **diagrama de dispersión o scatterplot**, donde se colocan los puntos en dos variables y pueden **analizarse las relaciones entre las mismas**. Para el caso, podemos ver que a mayores alturas corresponden mayores pesos y viceversa.

Hay que tener **mucho cuidado a la hora de expresar estas conclusiones en forma de relaciones simples y no en vínculos de tipo causa-efecto**. No es lo mismo decir que a mayores alturas corresponden mayores pesos (lo que es equivalente a decir que a mayores pesos corresponden mayores alturas) a decir que si un individuo es alto, entonces tendrá mayor peso. Para poder afirmar esta última frase es necesario comprobar que una variable depende de la otra (esto es, afirmar que un individuo con gran altura tendrá necesariamente un mayor peso). Esta comprobación no puede realizarse con el gráfico de dispersión.

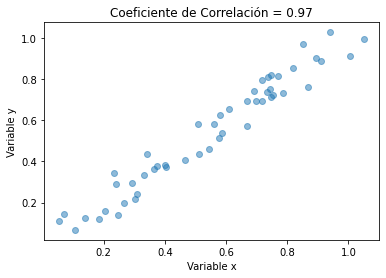
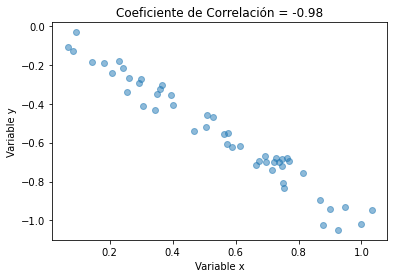
Si bien la diferencia es sutil, no deja de ser importante, y **hablaremos en este caso de relación entre variables**. Presentaremos la noción de relación como el resultado de la ocurrencia conjunta de los valores de las variables en estudio. Es importante tener en cuenta que **cuando hablamos de relación no estamos hablando necesariamente de dependencia de una variable con respecto a la otra**. La dependencia es objeto de estudio del análisis de Regresión, que veremos más adelante.

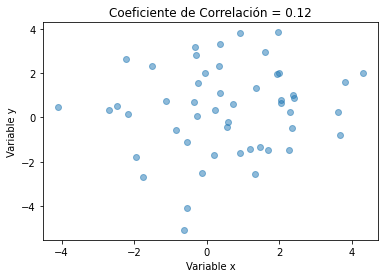
### Correlación

Si la relación **tiene alguna forma definida** será considerada como **correlación entre las variables**. Comúnmente hablaremos de **correlación lineal**, que es la más práctica a efectos del análisis en Data Science. La correlación lineal puede medirse a través de un indicador denominado **coeficiente de correlación**. El coeficiente de correlación puede tener valores entre -1 y 1, y lo interpretamos de la siguiente manera:

* Si el valor de la correlación es **cercano a 1**, significa que cuando los valores de una variable son altos, los de la otra variable también lo son, y análogamente con los valores bajos.
* Si el valor de la correlación es **cercano a -1** significa que cuando los valores de una variable son bajos, los valores de la otra variable son altos, y análogamente con los valores altos.
* Si el valor de la correlación es **cercano a 0**, significa que que **no hay una correlación lineal fuerte** entre las variables.

Se muestran ejemplos de los tres valores de correlación en las figuras siguientes. El coeficiente de correlación de 0.97 es cercano a 1 y **corresponde al primer caso,** **hablamos de correlación positiva**. El coeficiente de correlación de -0.98 es cercano a -1 y **corresponde al segundo caso, hablamos de correlación negativa**. El coeficiente de correlación de 0.12 es cercano a 0 y **corresponde al tercer caso, y hablamos de correlación nula o ausencia de correlación**.





Análisis de distribuciones de datos

## Entender la forma de los datos

Como mencionamos en la sección correspondiente a histogramas, y en las clases anteriores, es importante entender la forma de los datos. Para este fin, utilizaremos principalmente dos herramientas gráficas: por un lado el histograma de frecuencias, y por otro lado el diagrama de caja y bigotes.

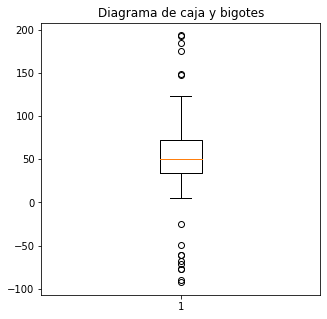
¿Qué buscamos cuando analizamos la forma de los datos? La respuesta es larga, pero podemos resumirla en los puntos siguientes:

* La distribución, ¿**es simétrica o asimétrica**?
* ¿En qué parte de la distribución se ubica **la mayor concentración de datos**?
* Los datos, ¿están más bien **dispersos o concentrados en torno a algún valor**?
* ¿Puede ser que existan datos **demasiado alejados del resto**?

Como vimos anteriormente, el histograma da una buena idea acerca de la **simetría** de la distribución. También da una noción acerca de la acumulación o **concentración** de los datos, o bien de su grado de **dispersión**. Pero a veces **puede que sea confuso, o que la forma que se ve en el gráfico no es demasiado concluyente**. Entonces necesitamos otra herramienta que complemente nuestro análisis.

## Diagramas de caja y bigotes.

Por lo mencionado anteriormente, para poder complementar este análisis e incorporar la mediana y cuartiles como indicadores para medir la concentración de los datos, es muy útil el **diagrama de caja y bigotes**. Este diagrama consiste precisamente en una caja con dos “palitos” o “bigotes”, en un diseño muy simple pero a la vez muy efectivo. Veamos un ejemplo donde desarrollaremos sus elementos.



Mediana

Cuartil 1 (25%)

Cuartil 3 (75%)

Máximo inlier

Mínimo inlier

25% de los datos

25% de los datos

50% “central”   
de los datos

Outliers superiores

Outliers inferiores

En la figura se muestra el diagrama de caja y bigotes, en este caso en forma vertical. Los lenguajes de programación ofrecen además la posibilidad de graficarlo en forma horizontal. En el eje vertical aparece una escala para la variable analizada. A continuación detallamos los elementos del gráfico:

* La línea anaranjada que está dentro de la caja representa **la mediana de la distribución**. Esto significa que el 50% de las observaciones están por debajo del valor de la variable marcado por esta línea, y el 50% restante está por arriba de este valor
* **El borde inferior de la caja marca el valor del cuartil 1 (25%) y el borde superior de la caja marca el valor del cuartil 3 (75%)**. Esto significa que la caja contiene un 75-25 = 50% de los datos. Otra forma de decir esto es que **el 50% de la concentración más “central” de los datos está delimitado por la caja**.
* Esto también significa que **el 25% de los valores está por debajo del límite inferior de la caja, y el 25% de los valores está por encima del límite superior de la caja**

De esta forma tenemos una idea detallada sobre la concentración y distribución de los datos. Hablemos ahora de los valores **extremos u outliers**, que serán cubiertos con más detalle en la clase siguiente. Los outliers son valores **“demasiado” alejados de la masa central de datos**. Si los valores no son outliers, o sea que están **“más cerca” del centro de los datos**, los denominaremos **inliers**. Existen muchos criterios para delimitar el espacio donde residen los inliers, y fuera del cual tendremos los outliers. El diagrama de caja utiliza uno en particular que veremos en detalle la próxima clase.

Supongamos que tenemos esa barrera que separa los inliers de los outliers. **La barrera en cuestión no está representada en el gráfico**, esto es importante porque puede dar lugar a confusión. Veamos entonces el resto de los elementos:

* Los bigotes representan **los valores de los últimos inliers**, esto es, los últimos valores que están alejados del centro de los datos pero que todavía no son outliers.
* Finalmente, **los outliers, si existieran, están representados por los círculos ubicados fuera de los bigotes**.

Como señalábamos al principio, el diagrama de caja y bigotes es muy simple en su diseño pero **muy poderoso por toda la información importante que contiene** para poder conocer la distribución de los datos.

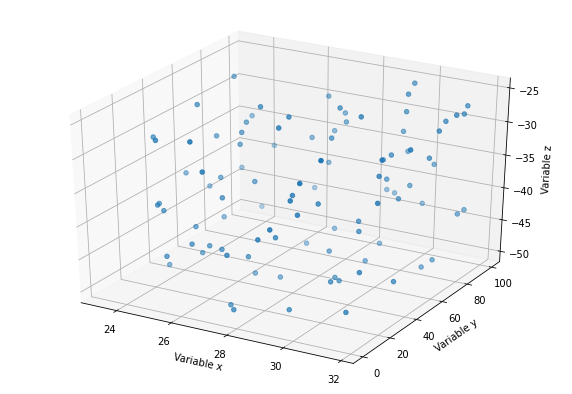
Gráficos para ver más de dos dimensiones

## Más de dos dimensiones

Cuando tenemos que comparar más de dos variables estamos en un problema. Esto se da porque un gráfico sobre el papel o la pantalla no puede mostrar más de dos variables o dimensiones. Para trabajar en más dimensiones podemos echar mano de la perspectiva, o de otras características visuales que permiten analizar los datos. Estas nuevas herramientas, si bien no proporcionan información completa sobre los datos, posibilitan una nueva forma de análisis que muchas veces ayuda a entenderlos mejor. Veremos algunas de ellas a continuación.

## Gráficos en tres dimensiones.

La forma más básica de representar datos, cuando tenemos **tres variables**, es hacer un **gráfico de dispersión en 3 dimensiones**. De esta forma, hacemos uso de la perspectiva y muchas veces podemos ver claramente formas y comportamientos en los datos. Veamos un ejemplo en la figura siguiente.



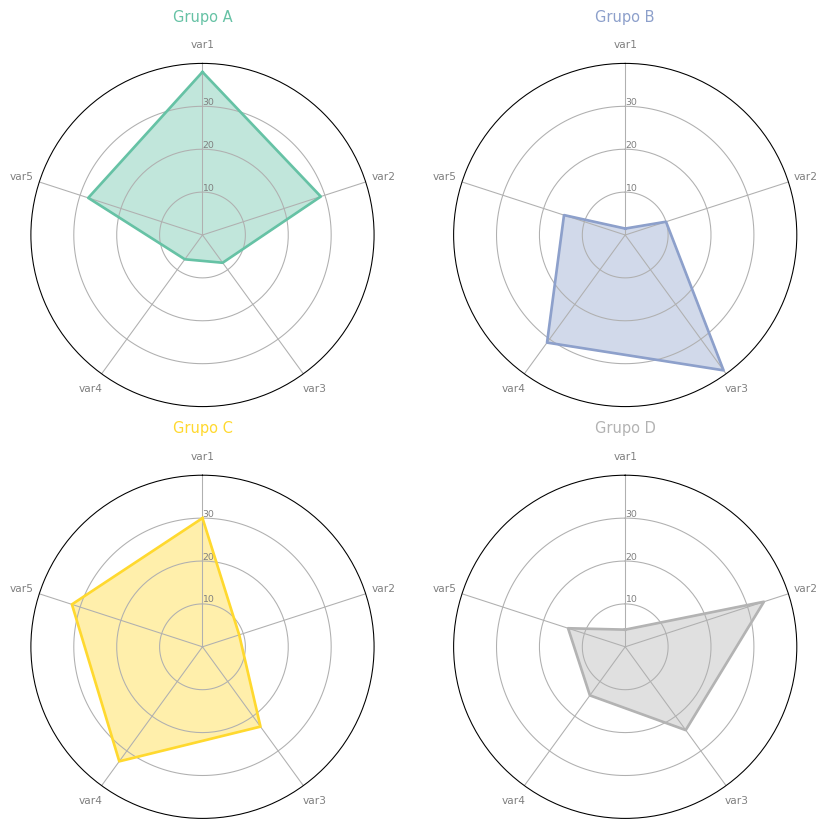
En este caso las variables x, y y z están ubicadas en los bordes de un cubo, y los puntos están ubicados en el espacio interior del cubo. Cada punto representa **la intersección de los valores de las variables x, y y z**.

## Otras formas de mostrar datos en más dimensiones

¿Qué pasa si tenemos datos en más dimensiones? Aquí tenemos que apelar al **ingenio y a las características que permiten que nuestro cerebro “piense” en más dimensiones**. Mostraremos dos herramientas en esta sección

### Gráfico de radar

El gráfico de radar surge para mostrar observaciones **en varias variables**, y tratar de compararlas visualmente. Se dibuja un radar o “telaraña”, donde cada eje representa una variable. Luego se grafica el valor correspondiente a cada variable en el eje, y se traza el área delimitada por los puntos. **Las observaciones similares tendrán áreas parecidas, y análogamente con las observaciones diferentes**. Veamos un ejemplo

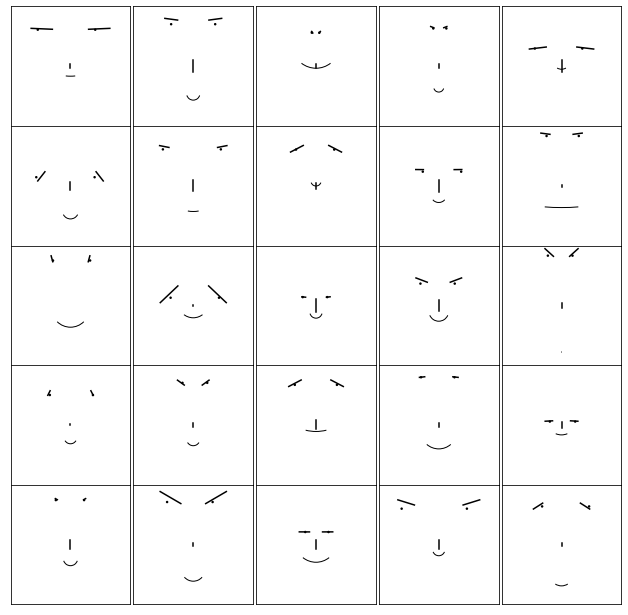


Este gráfico **permite comparar rápidamente múltiples cualidades en pocas unidades de observación**. En este ejemplo, podríamos decir que los grupos A y C tienen las áreas más grandes, por lo que sus variables tienen en general mayores valores. Además, es claro que el grupo A es muy diferente del grupo B, porque sus valores se destacan en distintas variables (1, 2 y 5 para el grupo A; 3 y 4 para el grupo B).

### Caras de Chernoff

Las caras de Chernoff se basan en el mismo principio que los gráficos de radar, pero en este caso **apelan al funcionamiento de nuestro cerebro** y cómo reacciona frente al reconocimiento de rostros. Habitualmente tendemos a agrupar rostros con expresiones faciales similares. En base a este hecho, las caras de Chernoff **asocian cada variable con una característica de la cara**: forma del rostro, forma de la nariz, expresión de las cejas, expresión de la boca, forma de las orejas, etc. También pueden agregarse accesorios, tales como sombreros, o bien colores, aunque en este último caso hay que tener en cuenta posibles destinatarios con problemas de daltonismo, con lo cual hay que ser cuidadoso al elegir la paleta de colores.

El siguiente ejemplo muestra caras en forma simplificada (a la manera de monigotes) para cada observación.



Aquí hay una cuadrícula de 25 observaciones. Cada observación está representada por un conjunto de cejas, ojos, nariz y boca. A pesar de ser muy rudimentarias, **las expresiones pueden inferirse claramente**: hay caras “tristes”, caras “aburridas” y caras “enojadas”, por ejemplo. **Las caras con expresiones similares tienen elementos similares**, por lo tanto los valores de las variables son similares. De forma análoga a los gráficos de radar, las caras de Chernoff son muy útiles, pero **solamente cuando las observaciones no son demasiadas**. Esto es lógico, porque si bien podemos reconocer expresiones de rostros, esta tarea se torna más dificultosa cuando los rostros a analizar son muchos, como por ejemplo cuando estamos frente a una multitud.

Para ver una cara de Chernoff interactiva, donde se muestra cómo la modificación de una variable afecta la expresión de un rostro (dibujado como monigote), puede visitarse el sitio <https://gramener.com/faces/> y jugar moviendo los controles deslizables.

FIN