Análisis de datos con R.

Usaremos un archivo csv, proporcionado por google, para realizar análisis de datos don R.
Utilizaremos herramientas básicas para los cálculos, pues no necesitamos más por el momento.
Tenga a mano el programa que escribió para calcular la moda.
Utilizaremos la desviación estándar muestral, pues no existe una diferencia sustancial entre 16999 y 17000 datos.
No olvide que debe presionar el ícono de "play" para que se ejecute el código.
Comenzamos con los preparativos:

```
#Para saber en que directorio estamos
getwd() #que significa obtener directorio de trabajo
```

debería dar como resultado */content*, si es así, se pasa a la siguiente celda de código.
Si obtiene */content/ sample_data* no ejecute la siguiente celda de código.

```
#Lo segundo es ubicarse en el directorio donde están los csv de google
setwd("/content/sample_data") #que significa establecer directorio de trabajo
```

Lo que sigue son dos comandos que instalan un paquete de funciones para la lectura de ficheros o archivos.

```
#Lo tercero es cargar las herramientas necesarias para leer csv
install.packages("readr") #este comando instala el paquete readr
library(readr)
```

Importando los datos.

Para importar los datos se usa un comando presente en la librería readr que se "llamó" en la celda anterior.
Estos datos se deben guardar en una variable, esta variable se llama "data frame", tabla de datos para nosotros.

```
#ahora se importa el csv en una variable, debe ir entre comillas el nombre completo del archivo
datos<-read_csv("california_housing_train.csv")
```

Visualizar los datos.

Podemos ver todos los datos, podemos ver las columnas en caso de que estén definidas, y tambien podemos parte de los datos, nos interesa saber las columnas de esta tabla, pues nos permitirá tratar datos específicos y relacionarlos.
Averigüe, si es que quiere, los comandos que se pueden utilizar para esto.

```
#Esto es solamente para confirmar que los datos se cargaron
#datos
#Y esto es para ver que datos se encuentran en la tabla.
names(datos)
```

```
'longitude' · 'latitude' · 'housing_median_age' · 'total_rooms' · 'total_bedrooms' · 'population' · 'households' · 'median_income' · 'median_house_value'
```

Una vez que se saben los nombres de las columnas, se pueden utilizar y visualizar siguiendo la sigiente forma:
NOMBRE_DEL_DATA_FRAME$NOMBRE_DE_COLUMNA
por ejemplo, para ver los datos de población (columna population) se debe escribir lo siguiente:

```
datos$population #muestra todos los datos de población, los 17 mil
```

Sería bueno que probara con otras columnas, a modo de práctica.

Repasando medidas de tendencia central.

Para calcular la media del total del total de habitaciones:

```
mean(datos$total_rooms)
```

Para la mediana del total de dormitorios:

```
median(datos$total_bedrooms)
```

Para construir un gráfico de caja-bigote de los valores medios de las casas:

```
boxplot(datos$median_house_value)
```

Para que el gráfico se muestre de forma horizontal:

```
boxplot(datos$median_house_value, horizontal=TRUE)
```

Para un histograma:

```
hist(datos$median_house_value)
```

Medidas de dispersión.

El comando range() entrega el valor mínimo y el valor máximo, aplicado a la latitud (posición norte-sur), queda así:

```
range(datos$latitude)
```

```
32.54 · 41.95
```

Si queremos saber la distancia entre el máximo y el mínimo, debemos restarlos directamente:

```
max(datos$housing_median_age)-min(datos$housing_median_age)
```

no debe confundirse, la variable DEBE SER LA MISMA EN EL MÍNIMO Y EN EL MÁXIMO, mantenga su atención en eso, sobre todo al copiar o reutilizar código.

Para la desviación estándar de la antigüedad media de la casa:

```
sd(datos$housing_median_age)
```

Para la varianza de la antigüedad media de la casa:

```
var(datos$housing_median_age)
```

Recuerde que la varianza es igual al cuadrado de la desviación estándar, por lo que se puede obtener de dos formas distintas, acá está la otra forma:

```
(sd(datos$housing_median_age))^2
```

Correlación.

El coeficiente de correlación de Pearson se calcula utilizando la siguiente fórmula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x - \overline{x})^2 \cdot \sum_{i=1}^{n}(y - \overline{y})^2}}$$

traducir esto a una función es algo demasiado complicado y se escapa del objetivo que tenemos en este electivo, afortunadamente en R se puede calcular con un simple comando.
Para calcular la correlación entre la antigüedad de las casas y su precio medio, el comando es el siguiente:

```
cor(datos$housing_median_age,datos$median_house_value)
```

Uso de variables auxiliares.

Para poder manipular la información de forma más fácil, podemos usar "variables auxiliares", por ejemplo, guardaré los datos de población en un vector llamado "pob", obviamente se utiliza un nombre que diga algo o que dé pistas sobre la variable, si utiliza letras, que sean las iniciales, así será más fácil saber que contiene la variable en posteriores usos:

```
pob<-c(datos$population) #se guardan todos los datos de la población en pob
pob #se muestran todos los datos de la columna de poblacion
```

A modo de práctica, realice todos los cálculos anteriores utilizando variables auxiliares

Ejercitando y aplicando

Ahora responda las siguientes preguntas:
¿Qué es lo que más influye sobre el precio medio?, incluya todos los resultados obtenidos.
¿Cuál es la dirección (calle y número) que corresponde a la mediana de las ubicaciones? (Una ubicación se compone de latitud y longitud, en ese orden)
¿Calcule las medidas de tendencia central y las medidas de dispersión para cada columna de datos en que tenga sentido.
Fundamente sus respuestas utilizando las celdas de código y los resultados numéricos.