1er parcial - Análisis de datos científicos en R

- Tienen dos horas+ para completar el parcial.
- Tienen todo el material de clase y online disponible para consulta.
- `NO` tienen a sus compañeros disponibles.
- Tienen que resolver los ejercicios escribiendo las respuestas (tanto texto como código) en un archivo **de texto** (puede ser una notebook en markdown, evitar usar Word, si esto los complica, avísenme por favor).
- **Importantísimo**: enumerar de la misma manera las respuestas (aunque estén vacías) para que coincidan correctamente. Notar que **cada pregunta y subpregunta** están numeradas.
- Algunas preguntas son solo para los alumnos que cursan la materia como **POSGRADO.** Para el resto de los alumnos son opcionales, pero suman ;)
- Al final del parcial me mandan el archivo de texto o notebook por mail o por mensaje privado en Google Classroom (o en un pendrive si no tienen Internet).
- Si no saben o no les sale algo, intenten explicar todo lo que sí saben al respecto, aunque sea parcial o incompleto, porque puede sumarles puntos. Si una línea de código les da algún error, expliquen la lógica de lo que quisieron hacer, ya que si la lógica es correcta, eso también suma. Si identifican la causa del error, aunque no lo sepan corregir, nuevamente eso suma. En general, mejor poner de más que de menos.
- No se estresen!

Problema

1. `DATOS`
    1. Cargar los datos “txhousing” del paquete `ggplot2` en la sesión (o verificar si ya se encuentran disponibles). Son datos de venta de inmuebles en Texas, con datos de ciudad, año, mes, número total de inmuebles vendidos (*sales*), volumen total vendido en dólares (*volume*), mediana del valor del inmueble, número de inmuebles disponibles (*listings*) y meses que se demoraría en vender todo lo disponible al ritmo de venta actual (*inventory*).
    2. Renombrar el dataframe a “tx” (o sea, asignar el objeto txhousing al objeto tx).
2. `PRIMER ANÁLISIS`
    1. ¿Qué clase de estructura de datos es `tx`?
    2. ¿Qué dimensiones tiene `tx` y con que comando lo veo?
    3. ¿Cómo son los *tipos de datos* de cada elemento de `tx`? Poner la instrucción usada para ver esto.
    4. **POSGRADO** Usando *subsetting* o `dplyr`, guardar en una variable llamada “med_houston_10_2010” la mediana del valor de inmuebles vendidos en octubre de 2010 en Houston. La variable debería contener solamente la mediana, es decir, que “med_houston_10_2010” *no debería ser* un dataframe.
3. `PREGUNTAS` - responder usando herramientas de `dplyr` y/o *subsetting*.

1. ¿Cuántas ciudades diferentes hay en el dataframe `tx`?
2. Análisis cuantitativo:
    1. Contar cuántas observaciones (filas) hay por ciudad (o sea, debe haber una columna con las ciudades y otra con el número de observaciones de cada ciudad).
    2. Contar cuántas observaciones hay por ciudad y año (ahora tendremos ciudad, año, y número de observaciones).
3. Análisis de precio:
   Tomando en cuenta todos los años (es decir todo el *dataframe*):
    1. ¿Cuál ciudad tiene los inmuebles más caros en promedio?
    2. ¿Cuál ciudad tiene los más baratos en promedio?
4. Análisis temporal: construir una nuevo *dataframe* llamado "tx_by_yr", con el volumen total de ventas y el número total de inmuebles vendidos, por año (es decir, habrá tres columnas: año, total volumen de ventas y total número de inmuebles vendidos).
    1. ¿En cuál año se recaudó más en total?
    2. ¿En cuál se vendieron menos inmuebles?
5. **POSGRADO** ¿En que mes y año se vendieron más cantidad de inmuebles (en número, o sea, `sales`) en la ciudad de "Amarillo"?
6. **POSGRADO** Agregar al dataframe original una columna adicional llamada *frac*, que tenga el número total de inmuebles vendidos (`sales`) dividido por el de disponibles (`listings`). Hacer un promedio de frac para cada año.

4. `GRAFICAR` con `ggplot2`

   Este punto utiliza el *dataframe* creado en el punto 3.4. Si no salió por alguna razón, pueden crear columnas artificiales usando las herramientas de `R` para generar números aleatorios, como `rnorm()` o `runif()`. Si se dan maña para hacer esto, se considerará el ejercicio (casi) resuelto ;)

    1. Graficar, con puntos, el volumen total de ventas *dividido* el número total de inmuebles vendidos, en función del año.
    2. Mapear el tamaño de los puntos al número total de inmuebles vendidos ese año (puntos más grandes significarán mayor número de inmuebles).
    3. Mapear el color de los puntos al volumen total recaudado.
    4. Ponerle texto explicativo a los ejes x e y, y título al gráfico.
    5. **POSGRADO** Agregar una regresión de tipo *loess* usando una capa `geom_smooth()`.