# Taller 3 - Análisis Descriptivo para Proyecto Final

Programación en R

Profesor Santiago Tellez Cañas

Fecha de entrega: Domingo, mayo 9 de 2021 a las 11:59 pm

## Instrucciones Generales:

- **Les sugiero leer el taller de manera completa antes de empezar.**
- Antes de comenzar deben crear un repositorio en Github con un nombre que identifique al grupo. El repositorio puede ser público y deben estar todos los integrantes del grupo como colaboradores. Antes de hacer la entrega, me degen agregar a mi como colaborador, también. Después de crearlo, creen un proyecto en RStudio y vincúlenlo al repositorio que creó. **En este repositorio deben incluir este taller, el análisis preliminar, la presentación y la entrega final.**
- El taller debe ser resuelto en los grupos en los que están trabajando el proyecto final.
- Todo el taller lo deben realizar en un archivo de RMarkdown. Este archivo debe producir un documento de Word o pdf que incluya el código y las respuestas, incluyendo las gráficas, cuando sea el caso. El RMarkdown debe comenzar con la activación del conjunto de paquetes `tidyverse` y otros paquetes que requieran para desarrollar el taller.
- El RMarkdown y el archivo de Word o pdf deben ser cargados el aula virtual antes de la fecha arriba mencionada.
- Una vez creado el proyecto, abran el archivo de RMarkdown y guárdenlo en esa carpeta. En la carpeta del proyecto también deben incluir las base de datos que usarán en el taller.
- Solamente deben usar funciones que hagan parte de los paquetes que hemos trabajando en clase.
- El taller se va a responder con la(s) base(s) de datos que están trabajando para el proyecto final. **Toda la manipulación de datos, incluyendo la importación y unión de las bases de datos (en caso de ser necesario), la transformación de la base, la creación y modificación de variables y, en general la limpieza de**

**la(s) base(s) la deben realizar en R, usando las funciones que hemos visto hasta el momento. No deben hacer ningún tipo de manipulación de la base de datos original por fuera de R (P.Ej. En Excel).**

- Cualquier pregunta me pueden contactar por correo electrónico, o pueden asistir a Sala Pitágoras los siguientes viernes de 10 a 12. También pueden acudir el jueves a la monitoría.

# Puntos del Taller

1. Carguen la(s) base(s) de datos usando la función que corresponda de los paquetes *readr*, *readxl* o *haven*.

2. De ser necesario, asegúrense que la base de datos esté lista para el análisis, usando la(s) función(es) del paquete *tidyr* que correspondan. Sean cuidadosos con la eliminación de los valores faltantes. No los eliminen salvo que lo consideren esencial de manera justificada.

3. Si están usando datos provenientes de más de una base de datos, unan las bases de datos usando la función que corresponda del paquete *dplyr*.

4. De ser necesario modifiquen las variables de interés **usando el paquete `dplyr`** para asegurar lo siguiente:

    - Las variables deben ser adecuadamente identificadas por R según su tipo. Por ejemplo, todas las variables numéricas -continuas y discretas-, deben identificarse coomo tal y no como carácteres, y todas las variables categóricas deben identificarse por R como factores ordenados o no ordenados, según el caso.
    - Las variables deben tener nombres que sean fáciles de entender. Por ejemplo, si la variable de género se llama P6020, deben modificar el nombre para que sea fácil entender cuál es el contenido de la variable.
    - Las variables binarias deben tener valores de 0 o 1. El valor de 1 debe corresponder con el nombre de la variable. Por ejemplo, si la variable se llama mujer, 1 corresponde a las mujeres y 0 a los hombres, o si la variable se llama bachiller, 1 debe corresponder a quienes son bachilleres y 0 a quienes no.

5. Usen alguna de las siguientes funciones para realizar una tabla de estadísticas descriptivas para las variables que van a incluir en su estudio, según sus necesidades: `summarize`, `statdesc()`, `summary()`, `stargazer()`, `datasummary()` o `datasummary_balance`. Si su variable dependiente o la principal variable independiente es una variable categórica, realice una distribución de frecuencias que incluya las frecuencias absoluta y relativa, y de ser el caso, las frecuencias absoluta y relativa acumuladas. Incluya esta tabla en el RMarkdown usando la función `kable` del paquete knitr, o `stargazer` directamente, según sea el caso.

6. Realicen un histograma y/o gráfico de barras, según se trate de una variable cuantitativa o categórica, para la variable dependiente y para la principal variable independiente. Las gráficas las deben realizar usando el paquete *ggplot*, y deben modificarlas en cuanto a número de bins, títulos, subtítulos, etiquetas de los ejes, escalas numéricas de los ejes y colores de una manera que permitan visualizar con claridad la información.

7. Para las variables cuantitativas (continuas y discretas), calculen la correlación entre cada pareja de variables y presenten estas correlaciones en una matriz de correlaciones. Para las variables categóricas hagan una tabla cruzada para cada par de variables. En ambos casos, comente sobre los resultados más relevantes.

8. Realicen una gráfica que les permita relacionar la variable dependiente y la principal variable independiente de interés. Si tienen varias variables independientes de interés, realicen una gráfica para mostrar la relación entre cada una de estas variables y la variable dependiente. Realicen las transformaciones en las variables que consideren necesarias para mostrar de mejor manera la información en estas gráficas. Las gráficas las deben realizar usando el paquete *ggplot*, y deben modificar los títulos, etiquetas de los ejes, escalas numéricas de los ejes y colores de una manera que permitan visualizar con claridad la información. Siga las siguientes recomendaciones:

    - Sí las dos variables son cuantitativas el gráfico debe ser un diagrama de dispersión.
    - Sí las dos variables son categóricas el gráfico debe ser un gráfico de barras que permita visualizar las dos variables.
    - Sí una variable es cuantitativa y la otra categórica, en un mismo gráfico realice un polígono de frecuencias, una densidad o un diagrama de cajas de la variable cuantitativa para cada una de las categorías de la variable categórica. Por ejemplo, si fueran a analizar la relación entre el tipo de colegio y el puntaje, deberían hacer un polígono de frecuencias, una densidad o un diagrama de cajas para el puntaje para los colegios públicos y otro para los colegios privados. Ambos polígonos o densidades deberían estar en la misma gráfica (i.e. mismo plano cartesiano) para poder apreciar las diferencias en la distribución de los puntajes de los estudiantes de ambos tipos de colegios. Si la variable categórica tuviera más de tres categorías, la gráfica podría hacerse con una función `facet_`.

9. **[Esta pregunta vale el 30% de esta parte del taller]** Usando la información presentada en las preguntas 6 a 9, ¿qué concluyen sobre la distribución de sus variables de interés y sobre la relación entre estas variables? Responda esta pregunta en al menos tres párrafos.