Profesor: M.Sc. Fabio Fernández Jiménez

Visualización e Interpretación de Datos

Tarea Número 2

Notas:

- Las tareas tienen fecha de entrega una semana después a la clase, y deben ser entregadas **antes del inicio de la clase siguiente**. **Cada día de atraso** en la entrega **tendrá un castigo de 10 puntos.**
- Las tareas son estrictamente de carácter individual
- Debe entregar el resultado como un archivo auto-reproducible (HTML o PDF) con la solución a todas las preguntas. Debe mostrar la respuesta en prosa a lo que se consulte (si aplica), así como el código utilizado y los resultados de ese código (tablas, gráficos, etc.)
- En nombre del archivo debe tener el siguiente formato: *Tarea1_nombre_apellido.pdf(*o *.html)*. Por ejemplo, si el nombre del estudiante es Luis Pérez: ***Tarea1_luis_perez.pdf***. Para la tarea número 2 sería: ***Tarea2_luis_perez.pdf***, y así sucesivamente.
- El puntaje de cada pregunta se indica en su encabezado.

Para todas las siguientes preguntas apóyese en la herramienta visual para dar solución a las mismas. Sea elaborado con respecto a títulos, etiquetas de ejes, así como colores y leyendas en los gráficos. Todas las soluciones deben estar basadas estrictamente a funciones dentro de **ggplot2**.

I. Los datos en WDS2014v2.csv contienen información de 214 países de todos los contienes relacionada con diversos indicadores:
   - P2014: población total en el 2014
   - TMI: Tasa de mortalidad infantil (por cada 1000 nacimientos)
   - TFR: Tasa de fertilidad
   - EVT: Esperanza de Vida al nacer (en años)
   - EVH: Esperanza de Vida al nacer en hombres (en años)

- EVM: Esperanza de Vida al nacer en mujeres (en años)
- REG: Región
- ARE: Área geográfica (continente)

1. ¿Cuál es la relación entre la tasa de fertilidad con la Esperanza de Vida al nacer en hombres? Construya un gráfico de dispersión con **geom_point** y agregue una curva por área (ARE) que indique la relación lineal entre las variables medidas. **(10 pts.)**
2. Adicione una variable booleana al *dataset* original llamada EVT2 que determine si la Esperanza de Vida al nacer (EVT) es menor a 65. A partir de esa nueva variable construya un gráfico de barras **horizontales** que cuente el número de países por región (REG), según esta nueva variable. Para esto use el *geom* **geom_bar**. ¿Qué regiones tienen la menor proporción de países con esperanza de vida mayor a 70? **(10 pts.)**
3. Muestre la relación entre la tasa de fertilidad y la tasa de mortalidad, tanto por área (ARE) como por grupo de EVT2. Para ello apóyese en *facet_grid* para poder mostrar todas las dimensiones solicitadas. **(10 pts.)**
4. Muestre la tasa de mortalidad infantil de cada país por continente (ARE), pero mostrando a la vez una dimensión para tomar en cuenta el tamaño de la población de cada país. Para este caso estamos simulando un *bubble chart* en el cual se muestra una dimensión más por medio de las burbujas.

   Para realizar el gráfico considere: **(15 pts.)**

   a. El área geográfica irá en el eje x
   b. La tasa de mortalidad se mostrará en el eje y
   c. El tamaño de la burbuja deberá ser especificado usando la opción *size*. Muestre únicamente los países con menos de 50 millones de habitantes (variable P2014), edite dicha variable P2014 para ver los datos en millones.
   d. Use el *geom* de *jitter* en vez de *point* para mitigar los *overlaps*. Utilice el *shape* =21 para simular las burbujas.
   e. El eje y debe estar indicado en unidades, para ellos modifique la escala de y (*scale_y_continuous*) con 10 *breaks* de tamaño 10.
   f. Agregue una etiqueta para el país con mayor tasa de mortalidad de cada región (la etiqueta debe tener el nombre del país)

g. Utilice el tema de blanco y negro y agregue etiquetas en los ejes y leyenda según corresponda. La leyenda del eje y debe estar en dirección horizontal (*angle=0*).

5. Modifique el gráfico anterior para mostrar un *boxplot* por área en lugar de las burbujas por país (agregue color según área). Conserve *jitter* para evitar *overlaps* ¿qué puede deducir de los gráficos de caja? Sea detallado. **(10 pts.)**
6. Por medio de un gráfico de densidad, determine cual área (continente) presenta una distribución distinta al resto en cuanto a la expectativa de vida en mujeres. Indique el color (*fill*) para área (ARE) manualmente por medio de *scale_fill_manual*. **(5 pts.)**
7. Detallemos más en el contienen europeo y revisemos qué países tienen una mayor tasa de mortalidad. Para esto utilice un gráfico de puntos en el cual la tasa esté en el eje *x*, y los países en el eje *y*. Ordene los países por tasa TMI para mostrar una mejor visualización (use *y=reorder(factor(Country.Name),TMI)* en el *aes* de *ggplot*). **(10 pts.)**

II. A partir de un set de datos propios, en el cual haya predominantemente variables cuantitativas:

1. Explique en qué consisten los datos. **(5 pts.)**
2. Determine dos problemas o análisis exploratorios que dese realizar. Detalle en qué consisten los análisis, así como las herramientas de ggplot2 que usará para realizar dicho análisis. **(5 pts.)**
3. Muestre el resultado del análisis usando ggplot2 para ambos problemas, sea elaborado en el uso de leyendas, títulos, colores, etc. Interprete los resultados. **(20 pts.)**

Nota: puede usar cualquier tipo de **ggplot2** para los análisis.

(506) 2268.8823 - (506) 8708.9091 info@promidat.com facebook.com/promidat www.promidat.com