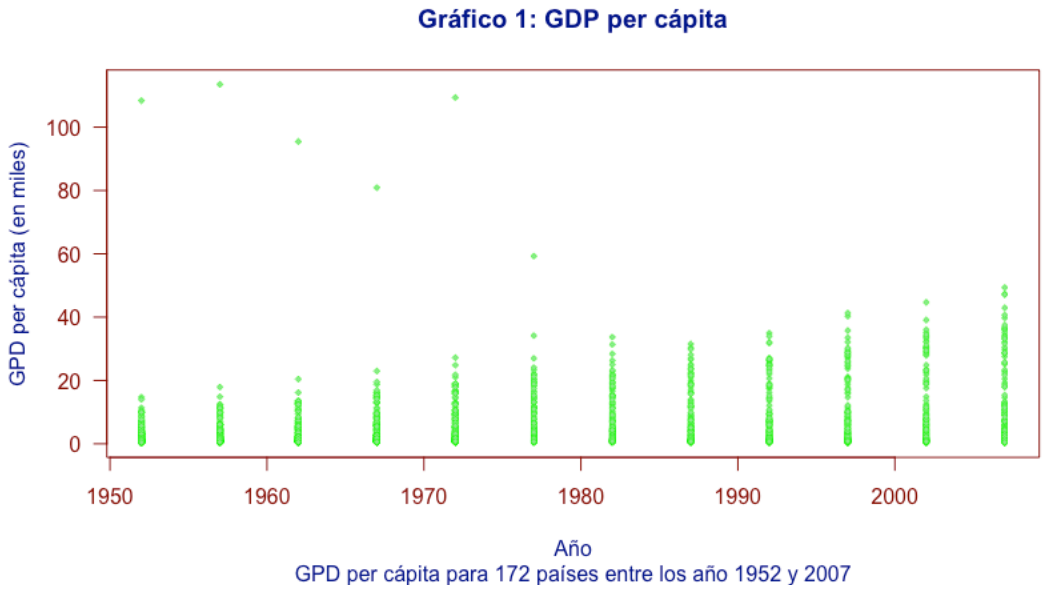
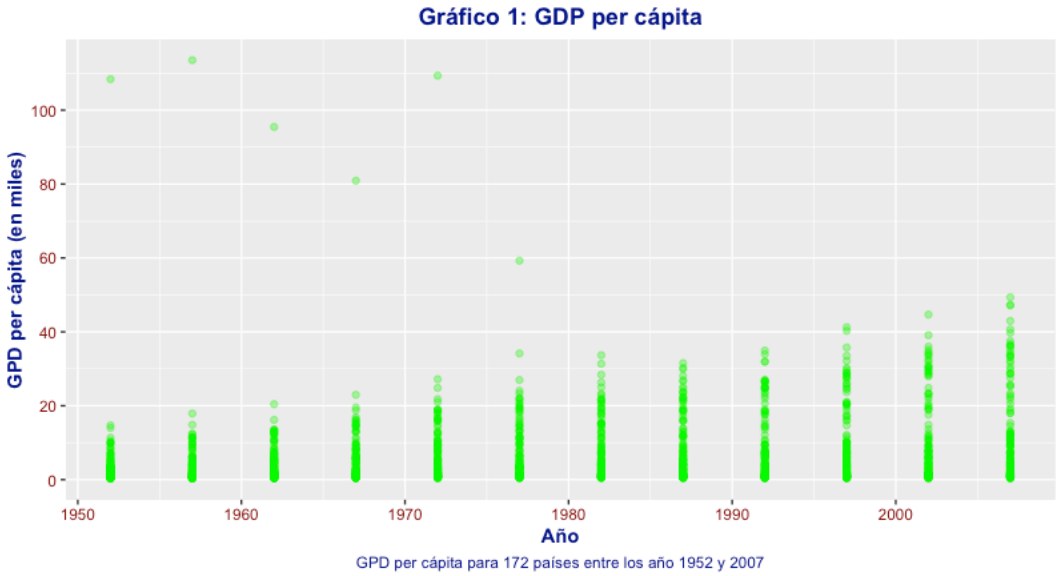
**Tarea 1. EPG3730 - Métodos Exploratorios y Computacionales para Estadística**
**Alumno:** José Reyes
**Fecha:** 12 de abril de 2021

***Nota**: Para la elaboración de los gráficos y el refinanmiento de los datos, el código en R, junto con las bases de datos, se encuentra en el siguiente repositorio github:*
*https://github.com/josereyessaldias/tarea_1_metodos*

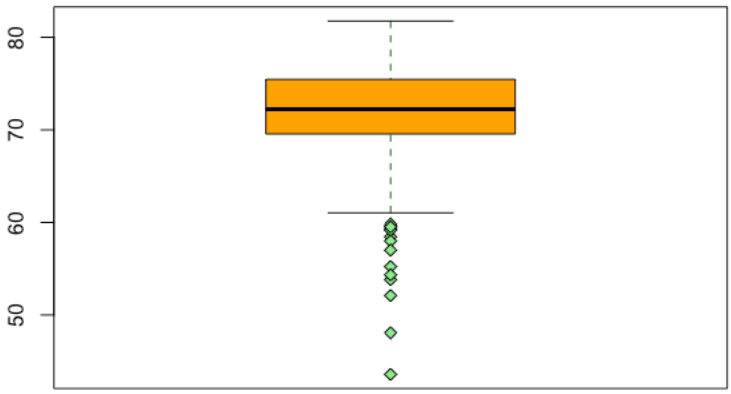
**Problema 1.**
El gráfico generado con el comando *plot()* se muestra a continuación:



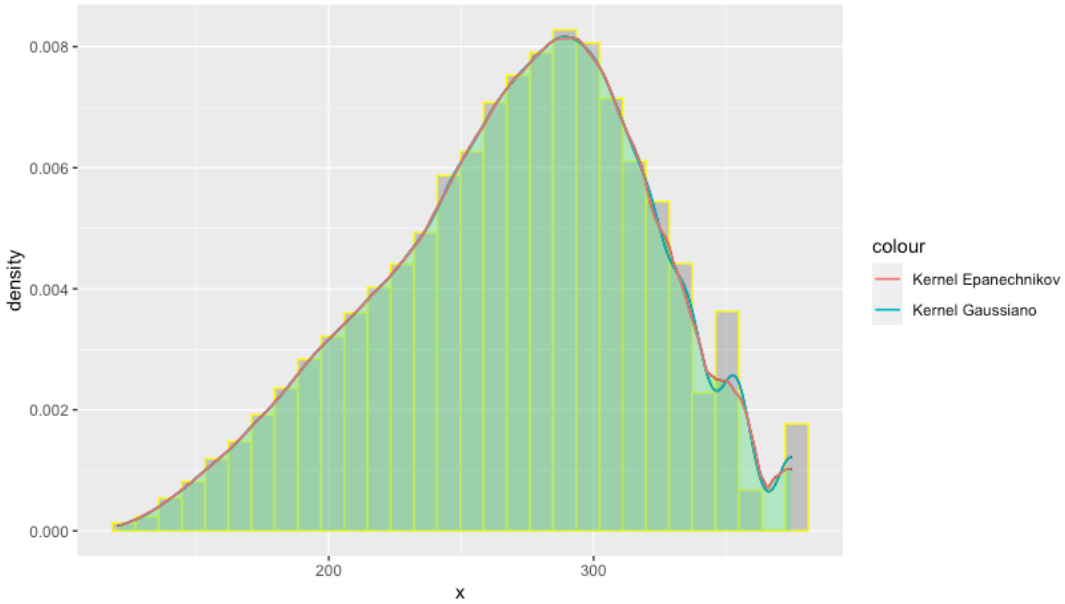
**Problema 2.**
El gráfico generado con la librería gg*plot2()* se muestra a continuación:




**Problema 3.**
El *boxplot* generado a partir de la función elaborada se muestra a continuación:



**Problema 4.**
Se trazó un *density plot* usando el *kernel Gaussiano* y el *kernel Epanechnikov,* el cual se muestra a continuación:




El *density plot* que aparenta tener un mejor ajuste con los datos es el que fue realizado con el *kernel Epanechnikov*. Ello se debe a que dicho *kernel* varía con menor intensidad ante agrupaciones de datos que rompen con la continuidad de la densidad en la muestra.

**Problema 5.**
En primer lugar, sabemos que:

$$\int y\frac{\hat{p}(x,y)}{\hat{p}(x)}dy = \int \frac{y\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}K\left(\frac{X_i-x}{h}\right)K\left(\frac{Y_i-y}{h}\right)}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}K\left(\frac{X_i-x}{h}\right)}dy,$$

$$= \frac{\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)\int yK\left(\frac{Y_i-y}{h}\right)dy}{\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)}$$

$$= \frac{\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)y_i}{\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)}$$

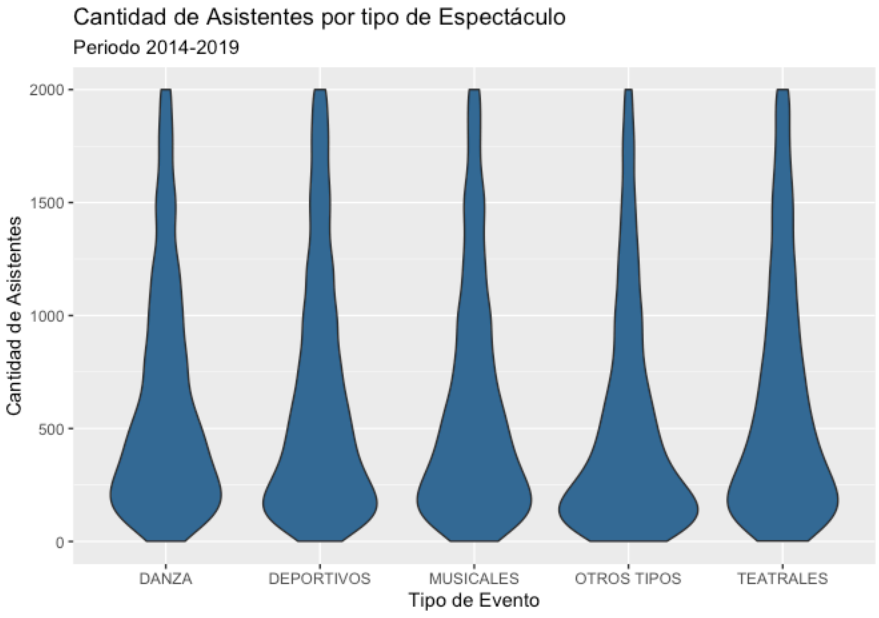
Siendo este último el estimador de kernel de Nadaraya–Watson.

**Problema 6.**

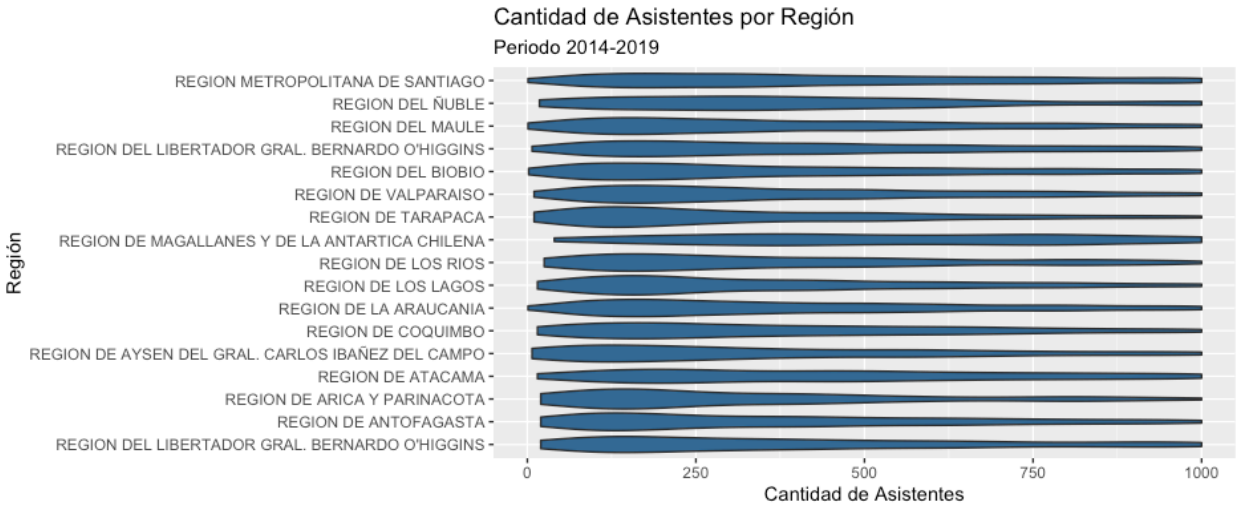
Se recurrió a la base datos denominada "encuesta_espectaculos_publicos_2014-2019", extraída del sitio web del Instituto Nacional de Estadísticas de Chile.[1] Dicha base de datos contiene información sobre la cantidad de funciones y el número de asistentes a espectáculos en Chile entre los años 2014 y 2019. Lo interesante de esta base es que congrega datos categóricos (por, ejemplo, el tipo de espectáculo, región y año) junto con datos continuos (número de asistentes pagados, número de asistentes gratuitos y número de funciones por espectáculo). Además, es una base cuya información puede ser de interés para un contexto post-pandemia en el que se reactiven los eventos en vivo.

En total, la base cuenta con 28.047 observaciones, cada una referida a un espectáculo. Cada espectáculo es categorizado a partir de cinco variables: año, semestre, región, tipo y sub-tipo de espectáculo. Además, las observaciones cuentan 18 variables continuas, las cuales surgen del cruce entre la información del número de funciones, número de asistentes pagados y número de asistentes gratuitos (es decir, 3 variables) para cada uno de los 6 meses considerados en cada uno de los años.

A partir de los datos, entonces, se elaboraron dos *violin plots*. El primero muestra la cantidad de asistentes totales a cada espectáculo en función del tipo de espectáculo. El segundo *violin plot* muestra la cantidad de asistentes totales a cada espectáculo en función de la región de Chile en la cual dicho espectáculo se llevó a cabo. Ambos *violin plots* se muestran a continuación.




[1] Extraída en: https://www.ine.cl/estadisticas/sociales/condiciones-de-vida-y-cultura/cultura


Cantidad de Asistentes por Región
Periodo 2014-2019
Región
REGION METROPOLITANA DE SANTIAGO
REGION DEL ÑUBLE
REGION DEL MAULE
REGION DEL LIBERTADOR GRAL. BERNARDO O'HIGGINS
REGION DEL BIOBIO
REGION DE VALPARAISO
REGION DE TARAPACA
REGION DE MAGALLANES Y DE LA ANTARTICA CHILENA
REGION DE LOS RIOS
REGION DE LOS LAGOS
REGION DE LA ARAUCANIA
REGION DE COQUIMBO
REGION DE AYSEN DEL GRAL. CARLOS IBAÑEZ DEL CAMPO
REGION DE ATACAMA
REGION DE ARICA Y PARINACOTA
REGION DE ANTOFAGASTA
REGION DEL LIBERTADOR GRAL. BERNARDO O'HIGGINS
0
250
500
750
1000
Cantidad de Asistentes