Maestría en Ciencia de Datos

Visualización de Datos

Curso 2023

Obligatorio 1

25/07/2023

Camila Rojí

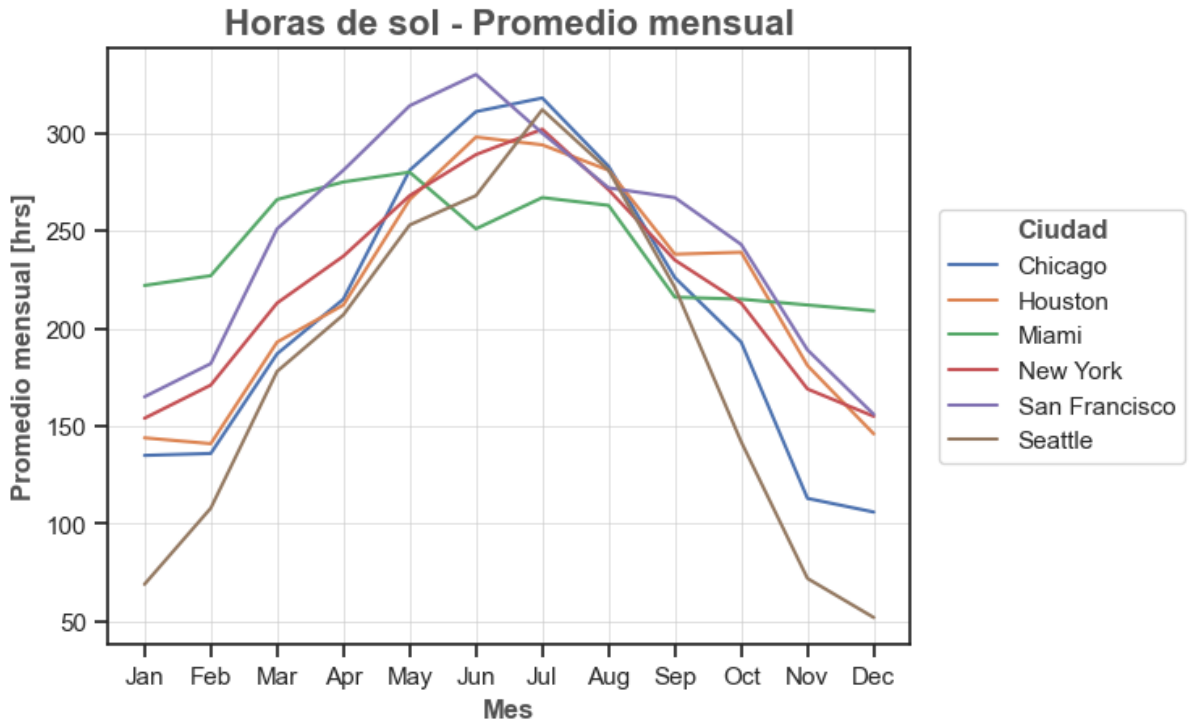
Diego Velasco

Introducción

En este trabajo vamos a presentar distintas visualizaciones que representen la información de los datos de Clima otorgados para analizar en el obligatorio.
Para elaborar las visualizaciones vamos a plantear preguntas de los datos que se desean obtener de forma de generar una visualización que la responda.

Preguntas:

1. ¿Para cada ciudad cuántas horas de sol se tienen a lo largo del año?



En la gráfica anterior representamos la cantidad de horas de sol que tiene cada ciudad en los distintos meses a lo largo del año.

Para diferenciar la información de cada ciudad utilizamos líneas con distintos colores, donde en el eje de las *x* se representan los distintos meses y en el eje de las *y* se representan las horas de sol. De esta forma obtenemos como resultado las horas de sol a lo largo del año para cada una de las ciudades encontradas en los datos. Se utiliza una línea de manera de facilitar la visualización de la tendencia anual para cada ciudad. Asimismo, se opta por una representación en orden cronológico, de manera de facilitar la identificación e interpretación del patrón subyacente. También se coloca una grilla horizontal y vertical de manera de facilitar la estimación de valores de interés.

La gráfica anterior fue realizada en python utilizando las librerías pandas, matplotlib y seaborn. El código puede verse a continuación.

```
fig,ax = plt.subplots()

for city in df['city'].unique():
    df_city = df[df['city'] == city]
    sns.lineplot(data = df_city, x = df_city.month, y =
df_city.sunshine, label = city)

ax.grid(lw = 0.4)
legend = ax.legend(bbox_to_anchor=(1.02, 0.75), title = 'Ciudad',
markerscale = 0)
legend.get_title().set_fontsize('12')
legend.get_title().set_fontweight('bold')
legend.get_title().set_alpha(0.8)

ax.set_xlabel('Mes', fontsize = 12, fontweight = 'bold', alpha = 0.8)
ax.set_ylabel('Promedio mensual [hrs]', fontsize = 12, fontweight =
'bold', alpha = 0.8)
ax.set_title ('Horas de sol - Promedio mensual', fontsize = 16,
fontweight = 'bold', alpha = 0.8)
```

Es posible apreciar como todas las ciudades presentan la misma tendencia, teniendo más horas de sol en los meses de Junio y Julio y menos hacia los meses de Enero y Diciembre. Esto era esperable debido a que todas las ciudades están ubicadas en el hemisferio norte, hemisferio para el cual el verano se da a mitad de año. También es posible ver que en los meses de verano, las ciudades en cuestión tienen una cantidad promedio de horas de sol más cercana entre sí, mientras que en los meses de invierno estos valores tienden a alejarse.

2. ¿Dónde se encuentran ubicadas las ciudades antes analizadas en el mapa de Estados Unidos?

En la visualización anterior representamos las ubicaciones de las ciudades en el mapa. Esto lo hicimos para utilizar los datos de latitud y longitud que no se habían representado en la gráfica anterior.

Para realizar esta gráfica, primero filtramos los datos del dataset de forma de tener una única fila por ciudad.

```
library(tidyverse)
data_filtered <- Clima %>% distinct(city, .keep_all = TRUE)
```

Luego para representar el mapa utilizamos la librería maps, y para etiquetar los nombres la librería ggrepel, con el comando map_data("state") obtuvimos el mapa a mostrar.

La ejecución de comandos sería la siguiente:

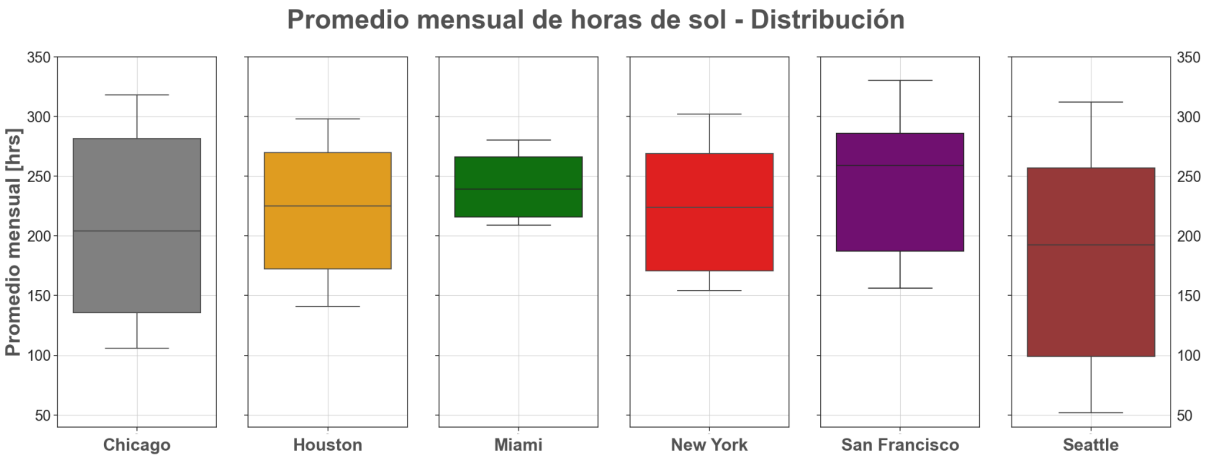
```
library(ggplot2)
library(maps)
library(ggrepel)
library(tidyverse)

us_map <- map_data("state")
data_filtered <- Clima %>% distinct(city, .keep_all = TRUE)
ggplot(data = us_map, aes(x = long, y = lat, group = group)) +
  geom_polygon(fill = "lightgray", color = "white") +
  geom_point(data = data_filtered, aes(x = lon, y = lat, group = city), color='grey' size = 3) +
  geom_text_repel(data = data_filtered, aes(x = lon, y = lat, label = city, group = city),
```

box.padding = 0.5, point.padding = 0.5, max.overlaps = Inf) +
labs(title = "Ubicación de las ciudades en Estados Unidos", x = "Longitud", y = "Latitud") +
theme_minimal()

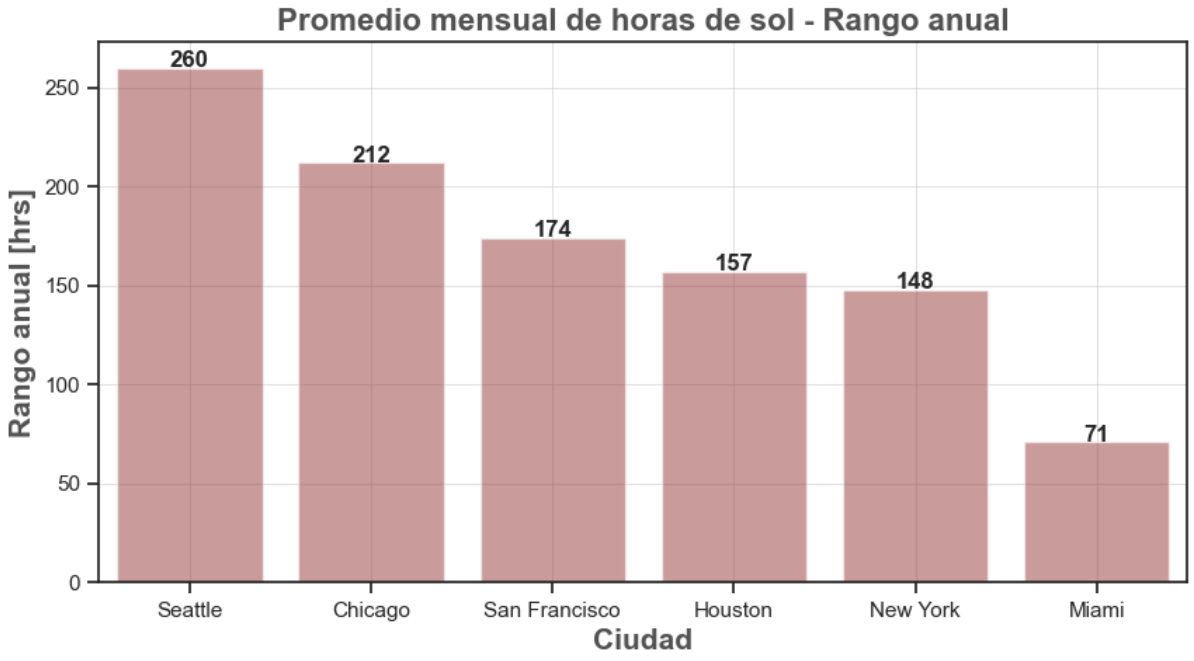
3. ¿Qué rango anual de cantidad de horas de sol presenta cada ciudad?

En primer lugar se realizan boxplots de los valores de horas de sol para cada ciudad. Se grafican todos en la misma fila y se utiliza la misma escala en el eje vertical de manera de facilitar la comparación visual entre las distintas ciudades.



Es posible ver como el rango anual de horas de sol de Miami, por ejemplo, es mucho más acotado y con un promedio mayor que el de Chicago y Seattle. Asimismo, estás últimas dos ciudades tienen el rango más amplio y el menor promedio anual de horas de sol de todas las ciudades analizadas.

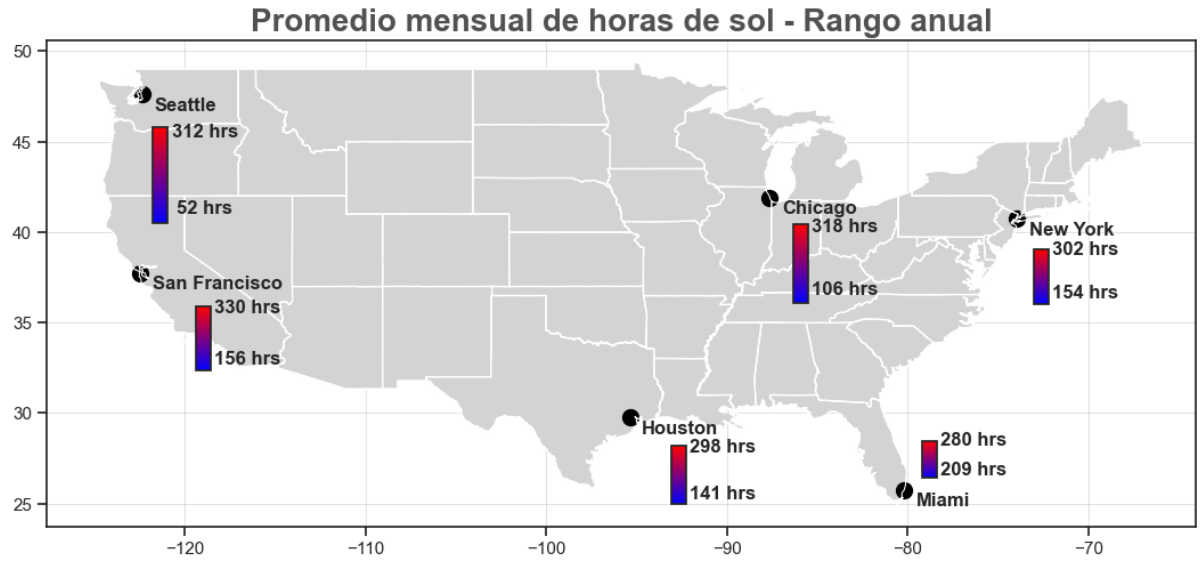
Asimismo, de manera de poder cuantificar estos rangos, se realiza el siguiente gráfico.



Se opta por un gráfico de barras debido a que la variable en el eje horizontal es de tipo nominal. Asimismo, se utiliza el mismo color para todas las barras de manera de uniformizar el gráfico, ya que se busca facilitar la comparación entre las distintas ciudades. Las barras se representan con cierta transparencia para poder visualizar la grilla, y se coloca el valor real de la barra como texto por encima de la misma de manera de evitar la necesidad de estimación.

De esta manera, puede verse claramente qué ciudades tienen un mayor y menor rango anual, confirmando lo que sugiere el gráfico anterior. Resulta de interés, a su vez, incluir la ubicación de estas ciudades en el análisis, de manera de evaluar si surge alguna tendencia.

Para ello se toma como base el mapa realizado anteriormente y se realiza otro gráfico similar, en el cual se muestra nuevamente el mapa de Estados Unidos, destacando las ciudades en cuestión, y se indica los rangos anuales de promedio mensual de horas de sol para cada una de ellas.



En este gráfico, el rango se muestra al lado de cada ciudad, con un gradiente de colores que va desde el azul para el valor mínimo del rango hasta el rojo para el valor máximo. Se utilizan estos colores para mostrar que, cómo se vio más arriba, el valor máximo del rango se da en los meses de verano y el mínimo en los meses de invierno. Asimismo, el tamaño de la barra que indica el rango es proporcional al rango en sí mismo, de manera de visualizar rápidamente qué ciudades tienen un rango más amplio. Es posible ver cómo, exceptuando New York, las ciudades ubicadas más al norte tienen rangos más amplios, presentando, en general, máximos mayores y mínimos menores, mientras que las ciudades más cercanas al Ecuador presentan rangos más acotados, y por tanto menores variaciones anuales.

Estos gráficos fueron realizados en python, utilizando las librerías numpy, pandas, matplotlib, seaborn y geopandas. El código para cada uno de los gráficos es presentado a continuación.

Boxplots

```
path = 'Clima.csv'
```

```
df = pd.read_csv(path, sep = ',')

fig,ax = plt.subplots(ncols=6, nrows=1, figsize = (30,10), sharey =
True)
colors = ['gray', 'orange', 'green', 'red', 'purple', 'brown']

i = 0
for city in df['city'].unique():
    df_city = df[df['city'] == city]
    ax1 = ax.flat[i]
    sns.boxplot(data = df_city[['sunshine']], ax = ax1,
color=colors[i])
    ax1.set_ylim(40, 350)
    ax1.set_title(city, fontweight = 'bold', fontsize = 26, alpha =
0.8, y = -0.075)
    ax1.grid(lw = 0.8)
    ax1.tick_params (axis='y', labelsize = 0)
    ax1.tick_params (axis='x', labelsize = 0)

    i += 1

ax[0].tick_params (axis='y', labelsize = 22)
ax[5].tick_params (axis='y', labelsize = 22, labelright = True,
labelleft = False)
ax[0].set_ylabel ('Promedio mensual [hrs]', fontweight = 'bold',
fontsize = 30, alpha = 0.8)
fig.suptitle ('Promedio mensual de horas de sol - Distribución',
fontsize = 40, fontweight = 'bold', alpha = 0.8)
```

Gráfico de barras

```
rangos = {}
for city in df['city'].unique():
    df_city = df[df['city'] == city]
    rangos[city] = df_city['sunshine'].max() -
df_city['sunshine'].min()

df_rangos = pd.DataFrame.from_dict(rangos, orient='index',
columns=['Values']).reset_index()
df_rangos = df_rangos.rename(columns={'index': 'city', 'Values':
'rango'})

fig,ax = plt.subplots(figsize = (10,5))
```

```
sns.barplot(data = df_rangos.sort_values(by = 'rango', inplace = True,
ascending = False), x = df_rangos.city, y = df_rangos.rango, color =
'brown', alpha = 0.5)
ax.grid(lw = 0.4)

for i in range(6):
    plt.text(x=i, y=df_rangos.iloc[i]['rango'] + 0.40,
s=str(df_rangos.iloc[i]['rango']), ha='center', fontweight = 'bold')

ax.set_xlabel ('Ciudad', fontsize = 0, fontweight = 'bold', alpha =
0.8)
ax.set_ylabel ('Rango anual [hrs]', fontsize = 15, fontweight = 'bold',
alpha = 0.8)
ax.set_title ('Promedio mensual de horas de sol - Rango anual',
fontsize = 16, fontweight = 'bold', alpha = 0.8)
```

Mapa con rangos

```
# Carga de mapa de USA
usa = gpd.read_file("States_shapefile.shx")

gdf = gpd.GeoDataFrame(df_rangos,
geometry=gpd.points_from_xy(df_rangos.lon, df_rangos.lat))

# Gradiente de colores
gradient_colors = ['#0000FF','#FF0000']
gradient_cmap =
plt.cm.colors.LinearSegmentedColormap.from_list('gradient_cmap',
gradient_colors)

# Gráfico de mapa y ploteo de ciudades
fig, ax = plt.subplots(figsize=(13, 6.5))
usa.boundary.plot(ax=ax, linewidth=0.75, color='white')
usa.plot(ax=ax, facecolor='lightgray')
gdf.plot(ax=ax, color='black', markersize=100, label='Cities', alpha =
1)
ax.grid(lw = 0.4)

for idx, row in gdf.iterrows():
    ax.annotate(row['city'], (row['lon'], row['lat']), xytext=(7.5,
-10), textcoords="offset points", fontweight = 'bold')
```

```
    sm = ScalarMappable(cmap=gradient_cmap,
norm=plt.Normalize(vmin=row['min'], vmax=row['max']))
    sm.set_array([])

    #Agrego barra de colores
    cax = fig.add_axes([u[row['city']][0], u[row['city']][1], 0.01,
max(0.05, row['rango']/2000)])
    cb = plt.colorbar(sm, cax=cax, orientation='vertical', ticks = [])
    cax.yaxis.set_ticks_position('right')

    cax.text(3.5, 0.1, str(row['min']).replace('.0','')+' hrs',
ha='center', va='baseline', fontsize=12, transform=cax.transAxes,
fontweight = 'bold')
    cax.text(3.5, 0.9, str(row['max']).replace('.0','')+' hrs',
ha='center', va='baseline', fontsize=12, transform=cax.transAxes,
fontweight = 'bold')

# Agrego título
ax.set_title ('Promedio mensual de horas de sol - Rango anual',
fontsize = 20, fontweight = 'bold', alpha = 0.8)
```