Introducción a la Ciencia de Datos - 2C 2022

Guía para la entrega Nº 3

Análisis dataset vuelos

Retomemos la tabla de vuelos sobre la cual charlamos en la exposición de la primera clase. Les recordamos que este dataset contiene los vuelos que salieron de la ciudad de Nueva York en el año 2013, información procedente de la Oficina de Estadísticas de Transporte de Estados Unidos. Queremos investigar si los retrasos dependen del mes del año. Para esto, vamos a necesitar el conjunto de datos completo, y no la selección que tiene solo los vuelos del mes de septiembre con la que trabajamos en la Cápsula 3 (para no colgar al Google Sheets), en la guia 2 están los detalles para bajar el dataset completo en Rstudio:

Despegue

Vamos a estudiar cómo cambia el comportamiento de algunas de las variables que venimos estudiando con el mes del año.

1. Agrupen la tabla de los vuelos de su aerolínea por los meses del año; usen `summarise` para calcular resúmenes sobre los retrasos en la salida y tiempo ganado en el aire.

2. Hagan gráficos que permitan ver las distribuciones de retrasos para cada mes. Para esto, experimenten con los parámetros fill y `color` de aes con las herramientas que usaron en los puntos 9 y 12 ¿Con todas las herramientas logramos visualizar las distribuciones? **Nota**: para que `month` funcione como variable categórica, tienen que convertirla usando la función `factor`.

3. Repitan para las distribuciones de tiempo ganado.

4. Ahora presentamos una nueva herramienta, que vive en el paquete `ggridges`. Tal vez tengan que instalarlo:

```
> install.packages('ggridges')
```

Para ver algunos ejemplos de uso, pueden explorar https://r-graph-gallery.com/ridgeline-plot.html. La función que les proponemos usar ahora es `geom_density_ridges`. Esto se usa como cualquiera de los otros geoms que venimos viendo. Experimenten con esta función y den rienda suelta a su artista interior, con los argumentos `alpha`, `fill` y `color`. Si sienten vientito en la cara está bien; estamos volando.

5. Calculen métricas de resumen (también conocidos como estadísticos de resumen) como las que vimos arriba (media, mediana, desvío standard, etc.) para cada mes del año. Usen estos cálculos para hacer gráficos de las métricas en función del mes del año con `geom_point` y `geom_line`. ¿Todos los meses son iguales? Si no, ¿en qué meses los vuelos tienen más retraso?

6. Comparen con la aerolínea de otro grupo. ¿Los gráficos muestran patrones similares? Si es así, ¿será una tendencia global? ¿Cómo la pueden explicar? Si no, ¿en qué difieren?

7. Por último, al igual que la semana pasada entendamos si nuestra producción está lista:

    a. ¿Qué sería necesario agregar o cambiar para que los gráficos se puedan presentar a alguien que no conozca el dataset? ¿Podrían, por ejemplo, presentárselos a estudiantes de otras materias que estén estudiando en el pasillo, o a algún familiar o amigx que esté con ustedes mientras terminan el trabajo, sin explicarles previamente nada del dataset? Si no es así, ¿qué les cambiarían? Realicen esos cambios.

    b. ¡Pongamos a prueba el punto anterior! Muéstrenle los gráficos a alguna persona, puede ser física o virtualmente. Registren la experiencia: ¿qué preguntas y devoluciones recibieron? ¿Hay alguna forma de resolver esas inquietudes en los mismos gráficos? Si es así, ¡háganlo!

Subir la producción realizada al campus de la materia en la solapa de "Entrega Nº 3" junto con un breve párrafo explicando lo que se ve en los gráficos y la conclusión que obtienen a partir de este análisis.