Introducción a la Ciencia de Datos - 2C 2022

Guía de Trabajos Prácticos 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). Entonces empecemos:

Parte 1. Check-in

1. Abran RStudio desde Ubuntu. Pueden hacerlo desde el menú de aplicaciones (abajo a la derecha) o tocando la tecla "Meta" (Windows) y tipeando "Rstudio". Antes de que terminen va a aparecer el ícono de la aplicación.

2. Por las dudas de que no esté instalado el paquete que tiene el set de vuelos, en la consola ingresen

```
> install.packages('nycflights13')
```

Si el paquete ya está instalado, esto no va a hacer nada.

3. Abran un nuevo script (**Ctrl+Shift+N** o en el menú de arriba a la izquierda) sobre el cual van a trabajar. Recuerden que para ejecutar líneas individuales, hay que seleccionarlas (a menos que sea una sola) y usar **Ctrl+Enter**.

4. En las primeras líneas, escriban los comandos que cargan los paquetes que vamos a usar y ejecuten:

```
library(tidyverse)
library(nycflights13)
```

Ahora deberían tener definida la variable flights, que contiene la tabla de todos los vuelos. Prueben meter en la consola

```
> flights
```
o bien
```
> view(flights)
```

¿Cuántas filas tiene el dataset? ¿Cuántas columnas?
¿Qué hace el comando `length(flights)`? ¿Y `nrow(flights)`?

5. Identifiquen el atributo que indica la aerolínea del vuelo. (Tip: con el comando `colnames()` se imprimen los *headers* del *dataframe*) ¿Entienden qué es cada columna? Si no, pueden usar `?flights` en la consola para leer documentación.

Parte 2. Abordando

6. Averigüen cuáles son las aerolíneas que operaron vuelos desde Nueva York en 2013. Para eso, pueden usar el comando `unique()`, como se hacía en Google Sheets, pero aplicado a una sola columna de la tabla.

Para elegir una columna, hay varias opciones:

a. `flights['nombre_de_la_columna']`
b. `select(flights, nombre_de_la_columna)`
c. `flights$nombre_de_la_columna`

Aplicar el comando `unique()` a estas opciones da resultados diferentes. ¿Cuál de estas opciones les devuelve una tabla?

7. Agrupen `flights` por aerolínea y cuenten cuántos vuelos realizó cada una en el año. Filtren la tabla original para quedarse solo con las aerolíneas que operaron más de 1000 vuelos en el año. Para esto, tienen que hacer una combinación de `group_by`, `mutate`, `ungroup` y `filter`.

8. Elijan una aerolínea de las que operaron más de 1000 vuelos, y filtren la tabla resultante del punto anterior para quedarse solo con los vuelos operados por esta aerolínea.

9. Realicen un histograma de la variable `arr_delay`. Para eso, usen `geom_histogram`. ¿Qué variable hay que mapear a x? ¿Necesitaremos alguna otra variable? Modifiquen la cantidad de bins (argumento `bins`). ¿Qué diferencia hay si usan `geom_freqpoly` en lugar de `geom_histogram`? ¿Y si usan `geom_density`? Noten la escala del eje x; ¿es la adecuada? Prueben cambiar los ejes sumando al plot el comando `xlim`.

10. Calculen la media, mediana, desviación estándar y distancia intercuartil para esta variable. (**Nota**: pueden hacer esto con `summarise`, usando las funciones `mean`, `median`, `sd` y `IQR`.) Usen `geom_vline` y `geom_hline` o `geom_segment` para agregar líneas en el gráfico anterior que señalen los valores de estas métricas.

11. Agreguen una variable a la tabla con los vuelos de esta aerolínea que indique el tiempo que recuperó en el aire. Piensen cómo calcular esto en base a las columnas existentes e impleméntenlo con `mutate`.

12. Visualicen la distribución de esta variable con alguna de las herramientas del punto 9. Calculen además las métricas de resumen (media, etc.). ¿Tiene sentido que haya valores negativos? ¿Pueden dar una explicación para los valores que obtienen?

Parte 3. Despegue

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

13. 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.

14. 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`.

15. Repitan para las distribuciones de tiempo ganado.

16. 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.

17. 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?

18. 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?