

Práctica Aproximación a la calidad de datos

La página <https://data.buenosaires.gob.ar/dataset/presupuesto-ejecutado> contiene información referida a la ejecución presupuestaria de la ciudad de Buenos Aires. Elijan el archivo correspondiente al trimestre del mes del cumpleaños del año 2018 del integrante más joven del equipo y resuelvan los siguientes ítems

- a) Lean la documentación sobre que representan las variables que funciona en la misma página. Analicen el dataset, como puede observar hay varias variables que funcionan de a pares, por ejemplo (car y car_desc). Analicen si existe correspondencia entre esos pares. Tengan en cuenta que en algunos casos existe una jerarquía
- b) Efectuar una apreciación sobre el nivel de cada una de las cualidades intrínsecas de calidad para este conjunto de datos. ¿Qué puede decir particularmente de la consistencia?
- c) Efectuar un análisis descriptivo de cada una de las variables, incluyendo el % de datos faltantes
- d) Efectuar un análisis bivariado de dos variables categóricas, dos numéricas y una numérica y una categórica
- e) ¿Cuáles son los elementos que más le llaman la atención?

Software

Este trabajo puede realizarse con diferentes softwares, un software muy utilizado para los temas de análisis de datos es R (<https://cran.r-project.org/>), que se puede utilizar por línea de comando o a través de una interfase grafica (<https://www.rstudio.com/>)

En internet hay muchísimos ejemplos de como usar R, pero acá hay dos links que hacen especial hincapié en las funciones estadísticas

<http://www.etsii.upm.es/ingor/estadistica/Grado/rESTDESC.pdf> (es un catalogo de funciones)

https://cran.r-project.org/doc/contrib/Chicana-Introduccion_al_uso_de_R.pdf (es un documento mas extenso donde las funciones están definidas en mas detalle)

En el siguiente link pueden encontrar un tutorial de cómo usar ggplot que es una librería de R que tiene gráficos más sofisticados:

<http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html>