Introducción a la Ciencia de Datos - 2C 2022

Guía de Trabajos Prácticos Nº 2

Les proponemos comenzar a analizar dos set de datos ubicados en el Campus de la materia a partir de una serie de consignas asociadas a cada uno.

Parte 1. Trabajo en el laboratorio. Análisis dataset adult_census

En el Campus de la materia van a encontrar el dataset adult_census. Estos datos fueron extraídos de la base de datos de la Oficina del Censo de 1994 de los Estados Unidos. A partir de todos los datos recogidos se extrajo un conjunto de registros razonablemente limpios con el objetivo de predecir si una persona gana más de USS 50.000 al año.

El objetivo de los siguientes ejercicios es entender y estudiar la brecha de salario (y si es que existía) entre hombres y mujeres de EEUU en el año 1994. Para eso sigamos los siguientes pasos:

1. Para empezar a conocer el dataset, analícenlo utilizando la herramienta filtros, ordenamiento y la función UNIQUE que les permitirá obtener la siguiente información:
    a. ¿Entre qué edades se encuentra la población de la muestra?
    b. ¿Cuántas categorías de trabajo (*workclass*) hay?
    c. ¿Cuántas categorías de educación hay?
    d. ¿Entre qué valores está la variable de horas trabajadas por semana?
2. Como ya se mencionó, vamos a intentar usar estos datos para buscar y entender una posible brecha salarial entre las dos categorías de *sex*. ¿Tiene sentido incluir en este análisis a los menores de 18 años? ¿Y a las personas que pertenezcan a la categoría *Never-Worked* de la variable *workclass*? Analicen si al eliminar a la categoría *Never-Worked* del dataset se remueven a las personas menores de 18 años. Discutir esto en grupo.
3. Este dataset es una **muestra** de un censo realizado en Estados Unidos. Es decir, se tomó una parte de ese censo en lugar de la información completa. Entonces, unx puede preguntarse si la **muestra** **representa** adecuadamente el total de la **población**. Esto es fundamental para saber si las conclusiones que saquemos del dataset van a ser válidas o no.

Usando tablas dinámicas, calculen la proporción de hombres y mujeres de la muestra y la proporción de personas de diferente etnia, y comparen con las tablas de acá abajo, que describen la composición de la **población** estadounidense. ¿Se mantienen las proporciones de la **población** en la **muestra**? Entonces, ¿es **representativa** la muestra elegida?

Population: sex	
Male	164,385,000
Female	167,509,000
Total	331,894,000

Population: race	
White	73.50%
Black	15.80%
Asian	6.50%
American Indian - Alaska Native	0.80%
Hawaiian	0.20%
Other	3.20%

4. Realizar un gráfico donde se pueda ver la cantidad de hombres y la cantidad de mujeres que ganan más de 50k dólares anuales. Según lo que se observó y concluyó en el ejercicio 1, determinar si es correcto realizar un análisis contando la cantidad de personas en vez de analizar los porcentajes dentro de cada clase (*male-female*).
5. Ahora realizar un gráfico con el porcentaje de hombres que ganan más de 50k dólares al año y el porcentaje de mujeres que ganan más de ese valor. ¿Qué se observa en el gráfico? ¿Qué conclusión (aunque sea provisoria) se les ocurre?
6. En la Hoja 2 del archivo adult_census se encuentra el mismo dataset con el que estamos trabajando, al cual se le agregó una columna donde se segmentó la variable *hours.per.week* de la siguiente forma: **A** : 1 h -20 hs, **B** : 21 hs - 40 hs, **C** : 41 hs - 60 hs, **D** : 61 hs - 80 hs, **E** : 81 hs - 100 hs.

   Realizar un gráfico de barras donde se vea qué porcentaje **del total de hombres** por un lado y qué porcentaje **del total de mujeres** por el otro pertenecen a cada una de estas categorías. Es decir, queremos ver de todos los hombres qué porcentaje pertenece a cada categoría y de todas las mujeres qué porcentaje pertenece a cada categoría. Pensar detenidamente si hay que tomar porcentajes de fila o de columna: para ello tener en cuenta el **total** al cual nos queremos referir. ¿Qué se observa en el gráfico? ¿Qué conclusión podrían proponer?
7. Pensar qué esperan ver si hiciéramos un gráfico que muestre para cada rango de *hours.per.week_segmentado* qué porcentaje del total de cada categoría gana más de 50k dólares al año y qué porcentaje menos de 50k al año. Pensar bien si hay que tomar porcentajes de fila o de columna. Hacer efectivamente el gráfico. ¿Qué se observa en el gráfico? ¿Es lo que esperaban? Si difiere de la expectativa, ¿de qué manera lo hace? ¿Por qué podría ser?
8. Condensen la información de los puntos 6 y 7 en un sólo gráfico utilizando las tablas dinámicas. Para ello elijan como filas de la tabla dinámica las variables *hours.per.week_segmentado* e *income* y como columnas la variable *sex*. Para los valores de la tabla elijan la variable *sex* y que los muestre

como porcentajes de columna para así estudiar todo esto respecto al total de hombres y el total de mujeres. Destildar la opción de "Mostrar Totales" para que la tabla no muestre los totales de cada categoría. Realizar con la tabla resultante un gráfico de barras. ¿Qué se observa en el gráfico? ¿Qué se observa particularmente en la categoría **B** de *hours.per.week_segmentado?*

9. Hacer un gráfico que muestre para cada clase de trabajo (workclass) qué porcentaje gana más de 50k dólares al año y qué porcentaje menos de 50k. Analizar con detenimiento si hay que tomar porcentajes de fila o de columna. ¿Qué tipo de trabajo tiene el mayor porcentaje de personas que ganan >50k anuales?
10. Analizar los distintos trabajos a los que se dedican hombres y mujeres. Para eso, hacer un gráfico de barras que indique el porcentaje de hombres respecto al total de hombres y el porcentaje de mujeres respecto al total de mujeres que se dedican a cada una de las distintas categorías de trabajo. Es decir, queremos ver de todos los hombres qué porcentaje se dedica a cada tipo de trabajo y de todas las mujeres qué porcentaje se dedica a cada tipo de trabajo. En el gráfico realizado se observa que hay una clase de trabajo que supera considerablemente al resto y no permite analizar lo que pasa con el resto de las clases laborales. Hacer otro gráfico sin tener en cuenta ese dato para poder estudiar con más detalle el resto.
11. Estudiar en conjunto los gráficos obtenidos en los incisos c y d. A partir de lo observado, ¿se puede dar una explicación de la conclusión obtenida en el punto 5? Prueben armar un breve argumento que apoye la conclusión basada en esos gráficos. Esto puede ser una lista "*bulleteada*" de los resultados que obtuvieron en los últimos gráficos.
12. Condensen la información de los puntos 9 y 10 como hicieron en el punto 8. Hacer dos gráficos de barras (uno para mujeres y otro para hombres) donde se muestre el porcentaje que se dedica a cada clase de trabajo y si ganan o no más de 50k anuales, respecto al total de mujeres/hombres. ¿Qué se observa en el gráfico?
13. Realicen el mismo análisis que hicieron para las variables *work_class* y *hours.per.week_segmentado* con la variable *education*. Antes de realizar los gráficos, piensen qué esperan que dé. El objetivo es llegar a gráficos como los de los puntos 8 y 12. ¿Qué se observa efectivamente en los gráficos? ¿Es lo que esperaban?
14. Si tuvieran que recomendar alguna medida para equiparar los ingresos entre hombres y mujeres, ¿cuál sería? ¿Cómo intentarían convencer a alguien de que esta medida es útil o justa y que resuelve algún problema? Intenten realmente armar la línea de argumentos en base a los gráficos que hicieron. Puede ser una oración acompañada de un gráfico en cada caso. Prueben. ¡Es difícil! Pero seguro que la primera vez que Messi agarró una pelota no la movía como ahora.