Para responder estas preguntas se ha utilizado EDA →notebooks→questions_help.ipynb y EDA → reports → images

Apartado C

9.Responde a las preguntas:

¿Ha sido posible demostrar la hipótesis? ¿Por qué?

Si, ha sido posible demostrar la hipótesis porque los datos mostraban tendencias lógicas a las preguntas que se formulaban en cada caso para poder verificar esa hipótesis. Se ha demostrado que las principales causas de los casos de alcoholismo vienen por pautas educacionales intencionadas de los tutores y no por otros parámetros no controlados por el tutor. Además, se ha estudiado el caso singular del sexo, que presentaba una fuerte oposición a la afirmación de la hipótesis. Pero se ha podido analizar que esta disparidad que presenta el sexo deriva de la fuerte diferencia estratégica que tiene la sociedad para tratar de diferente manera a alumnos y alumnas, y no deriva de un canon sexual que diferencia a ambos sexos por rasgos biológicos.

¿Qué has podido concluir del estudio de tus datos?

Por contenido, que existen motivos significativos en las pautas educativas de los tutores para definir la tasa de consumo de alcohol del sujeto. Mayormente, por definición de los tiempos de estudio y tiempo libre y la varianza de soporte educacional según el sexo. En contraposición, se encuentra poca correlación con el tipo de población (urbana o rural), la edad del alumno, su situación sentimental, la situación de la relación familiar o la educación de los padres, por ejemplo. Para estos últimos casos, no se puede definir que haya pautas educativas clasificadas, pero si que las hay al permitir que los alumnos gocen de largos tiempos de libertad y no tengan la obligatoriedad o necesidad de llevar a cabo largos tiempos de estudio o solicitar ayuda académica.

En referencia a la segunda parte de la hipótesis, la correlación entre el alcohol y las notas no es muy significativa (-0.14), por lo que sí, existe una pequeña correlación y se puede justificar que si el alumno aumenta su ingesta de alcohol, hay riesgo de obtener peores notas, pero no tanta como para afirmar que los alumnos que beben obtienen peores notas. No obstante, "higher", el parámetro que estudia la motivación de los alumnos para seguir estudiando al finalizar, indica que a medida que aumenta las notas de los alumnos en el dataset, también lo hacen sus motivaciones académicas.

¿Qué cambiarías si hicieras otro proyecto EDA?

Primero definiría mejor los parámetros de escritura que tienen que ver con la persona y el lenguaje, ya que al querer formalizarlo, el cuerpo del proyecto era muy extenso y es algo que podría haber clasificado desde un principio.

También, cogí una tabla muy extensa (en columnas) desde un principio pensando que eso me daría mucho juego para formar un buen proyecto *EDA*. No obstante, ha actuado como un arma de doble filo ya que, después de trabajar las gráficas y elaborar todos los pasos que pide el proyecto, me he dado cuenta que hay alguna columna que no le he acabado dando uso; aunque por el tema estudiado, podían haberse utilizado, pero para el caso de estudio ya se han tocado suficientes parámetros. Por ello, considero que tengo que plantear mejor el trayecto que quiero hacer para evitar tener columnas no utilizadas en mi *DataFrame* procesado.

Y finalmente, tener la idea de inicio un poco más clara. Si bien el proyecto *EDA* es un proyecto que avanza pautadamente, no tener la idea de inicio muy clara o no ver directamente en el *dataset* claras oportunidades con las que trabajar, puede ralentizar mucho el proceso e incluso confundir al programador en algunas situaciones de visualización y recogida de conclusiones.

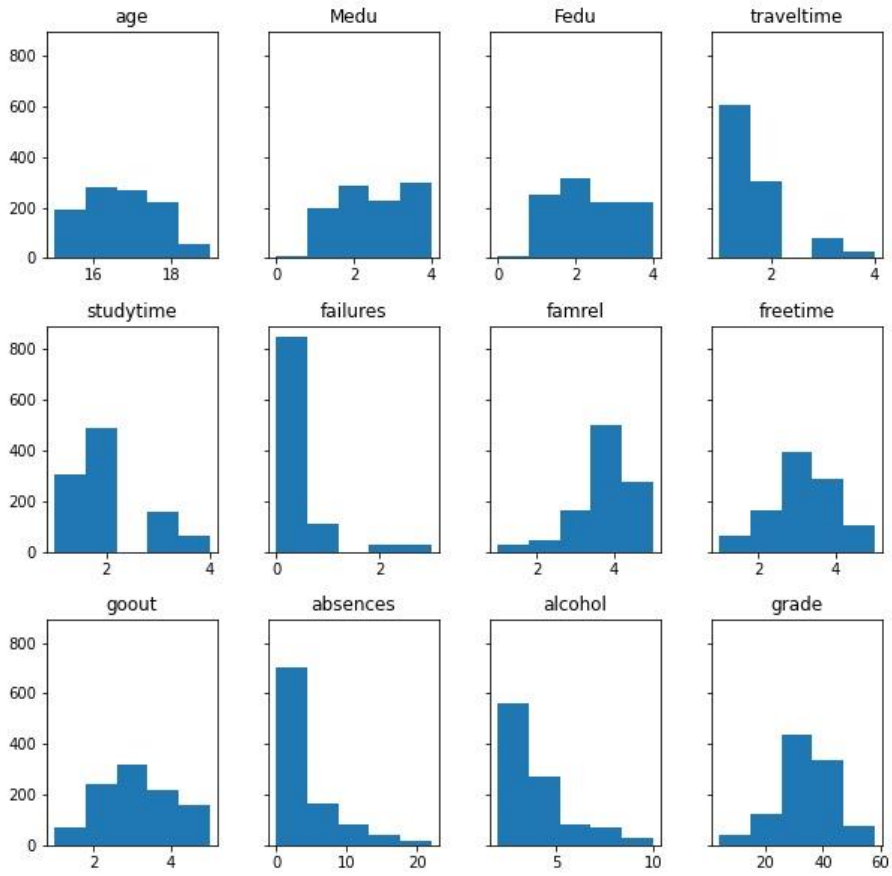
¿Qué has aprendido al hacer este proyecto?

Que lo más importante es tener claro los pasos y saber clasificarlos, ya que es muy grande, a primera vista siempre cuesta de entender, y tengo que centrar mi tiempo en tares específicas, y, a medida que las vaya completando, confiar en que si están bien clasificadas, saldrá un buen proyecto.

Apartado B

1.Muestra el histograma de cada columna en tu dataset con bins=5. ¿Como están los rangos pintados?

Los rangos están pintados a 5 intervalos que separan los datos de cada columna, inesperadamente bien adecuados a la mayoría columnas teniendo en cuenta que muchas de las columnas cuenta con un bajo número de valores únicos. Como es el caso de edad (“age”), que toma valores desde 15 hasta 19.



2.¿Cuáles son las columnas con mayor correlación? Dibuja una matriz de correlación.

Las columnas con mayor correlación positiva, como muestra la gráfica abajo, son:

-Fedu : Medu

-goout : alcohol

-freetime : goout

-higher : grade

-studytime : sex

Los columnas con mayor correlación negativa, como muestra la gráfica abajo, son:

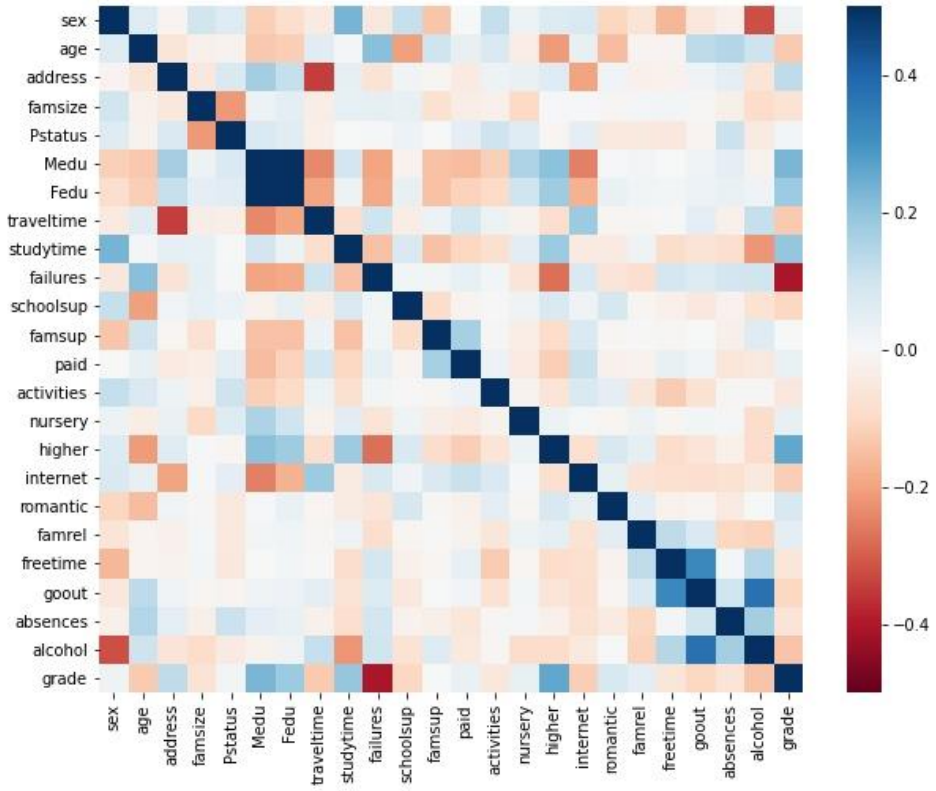
-failures : grade

-traveltime : address

-sex : alcohol

-failures: higher

-internet : Medu



Apartado A

4.Responde a las preguntas:

¿Hay outliers o data irregular?

Sí, un segmento de la data no se ha considerado en el estudio porque mostraba unos valores demasiado irregulares y se ha demostrado, para ello, en el proceso de data cleaning en el archivo main.py, que dichos valores se separaban en gran medida de la desviación estándar con el estudio de las medias y no presentaban atributos de normalización ya que mostraban valores muy poco frecuentes.

¿Qué columnas tienen más valores repetidos?

Como se puede ver en la imagen, las columnas “grade” y “absences” tienen una variedad de variables únicas significativamente superior a las demás. Por lo que, considerando que todas las columnas tienen el mismo número de valores, los parámetros que en la gráfica se muestran con una barra más corta, son los que tienen más repeticiones. Por conocimiento del dataset, son columnas con valores binarios (“sex”, “address”, “famsize”, “Pstatus”, “schoolsup”, “famsup”, “paid”, “activities”, “nursery”, “higher”, “internet”,” romantic”).