INFORME

El data set original que fue suministrada (“data.csv”) presentaba un formato poco legible para el usuario, y con el cual podría ser complicado de manipular para el programador en el lenguaje python. Debido a esto se decidió seguir una serie de pasos para que en primer lugar la data fuese fácil de manipular en el lenguaje de programación python , y así tomar una serie de decisiones para la limpieza de datos y finalmente crear el archivo de salida “minable.csv” el cual contendrá presentará de forma legible el data set para el usuario final.

Pasos que se siguieron para obtener la Vista Minable:

I: En primer lugar se eliminó la primera columna que no contenía ninguna información relevante, en principio se pensó que guardaba índices de los individuos de forma desordenada, pero como contenía valores superiores a la cantidad de personas en el formulario, simplemente se decidió descartarla. Se eliminó el atributo ‘Edad’, ya que la supresión de estos datos no conlleva a pérdida de información, la edad puede ser calculada a través del atributo ‘Fecha de Nacimiento’. Se eliminaron los atributos: ‘Se encuentra realizando alguna actividad que le genere ingresos’, y ‘En caso de ser afirmativo, indique el tipo de actividad y frecuencia’. Ya que al obviar estos datos no se pierde información, esta se puede ver reflejada directamente en los campos relacionados con los ingresos económicos del estudiante.

II: Muchos de los nombres de las columnas de la data eran poco legibles para el usuario, además era complicado trabajar con ellos en python ya que por ejemplo muchos contenían varios puntos y esto ocasionaba errores en el lenguaje de programación, debido a esto se decidió renombrar la mayoría de las columnas. De esta forma serian legibles para el usuario, y brindarían facilidad para su manipulación en el lenguaje de programación python

III: Se unificó el formato de columnas como ‘Periodo académico a renovar’ y ‘fecha de nacimiento’ ’ para una mejor lectura. Fueron descartadas aquellas filas que tenían formatos ambiguos, los cuales hacían muy difícil determinar cuál era la fecha de nacimiento de la persona, o su periodo académico a renovar.

IV: Se sustituyeron los campos de las siguientes columnas con valores numéricos:

'Estado Civil': '0' para soltero(a) y '1' para casado(a)

'Sexo': '0' para Masculino, '1' para Femenino

'Escuela':'0' para Enfermería, '1' para Bioanálisis

'Modalidad de Ingreso': '0' para 'Prueba interna y/o propedéutico', '1' para **'Asignado Opsu'**, '2' para 'Convenios Internos'

'Ha cambiado de dirección': '0' para 'No', '1' para 'Si'

'Está realizando tesis, trabajo de grado, o pasantías de grado' : '0' para 'No', '1' para 'Si'

'Contrajo matrimonio': '0' para No, '1' para Si

'Ha solicitado algún beneficio a la universidad u otra institución': '0' para 'No', '1' para Si

IV: Se verificó que el número de materias inscritas del estudiante fuese igual a la suma del número de materias aprobadas en el semestre o año anterior más la suma del número de materias retiradas en el semestre o año anterior más la suma de materias reprobadas en el semestre o año anterior. Serán descartadas aquellas filas que no cumplan esta condición. Antes de esto se llevaron todos los valores a un mismo formato numérico.

V: Se acomodaron las columnas de 'promedio' y 'eficiencia', para que estas tengan un formato valido

VI: En las columnas referentes a los ingresos y gastos del estudiante y/o representante económico, en primer lugar se llevaron los valores a un mismo formato numérico. Luego se tuvo que decidir qué hacer con los valores nan y si tenía sentido que algunos de estos campos tuviesen el valor cero. El criterio a seguir no fue el mismo para todos los casos, y este es descrito a continuación para cada una de estas columnas. Se mantuvieron todos estos atributos, ya que

son relevantes tanto los valores individuales, como los totales, referentes a gastos e ingresos.

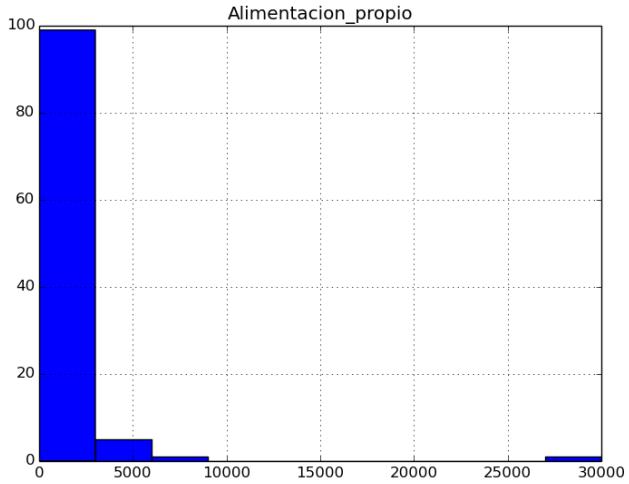
‘Aporte mensual que le brinda su responsable económico’: En este caso no hubo complicaciones, tiene sentido que el aporte mensual recibido por el representante económico sea cero, porque el estudiante no necesariamente recibe esta ayuda. Entonces se procede a imputar ceros a los valores ‘nan’ y no modificar los valores ceros que ya están almacenados en este campo.

Este mismo concepto se aplica para las columnas: ‘Aporte mensual que recibe de familiares y/o amigos ‘ , ‘ingreso mensual que recibe por actividades a destajo o por horas’

‘Alimentación propio’: En este caso se puede interpretar como los gastos en alimentación del estudiante, más no lo que el gasta en alimentación, es decir el estudiante puede gastar en comida cierta cantidad de dinero así él no esté pagando estos gastos, sino otra persona como su responsable económico por ejemplo. Por esto se consideró que los valores ‘nan’ no debían ser sustituidos con ceros, ya que el estudiante siempre tendrá algún gasto en alimentación.

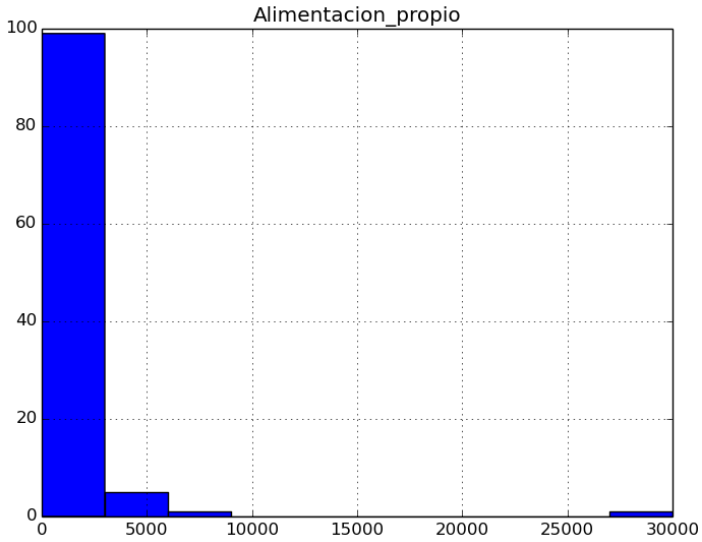
Ahora para saber qué valor imputar para sustituir los valores nan, se realizaron histogramas de frecuencia de la data original para tener una idea de la distribución de los valores, como se muestra en la **figura 1**

Figura 1



Al observar que hay valores atípicos, los cuales se alejan por mucho de los valores más frecuentes, sería poco conveniente usar como estrategia para imputar, a la ‘media’, ya que estos valores atípicos afectarían de gran forma al promedio de los valores. En lugar de esto se optó por imputar la mediana, de esta manera la distribución de los datos no se verá demasiado afectada. Como se puede observar en la **figura 2.**

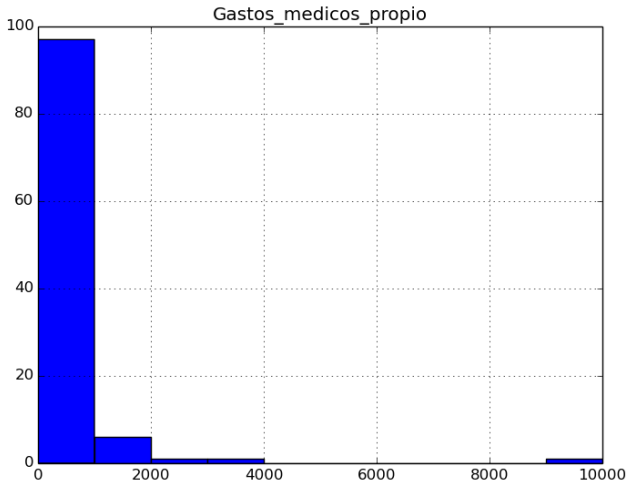
Figura 2



Para la columna ‘Gastos médicos propio ocurre lo mismo. Se consideró que no se deben imputar ceros en los valores ‘nan’, ya que el estudiante en teoría siempre debería tener gastos médicos, por ejemplo, así este asiste a un hospital, las medicinas que necesite es un gasto que debe asumir.

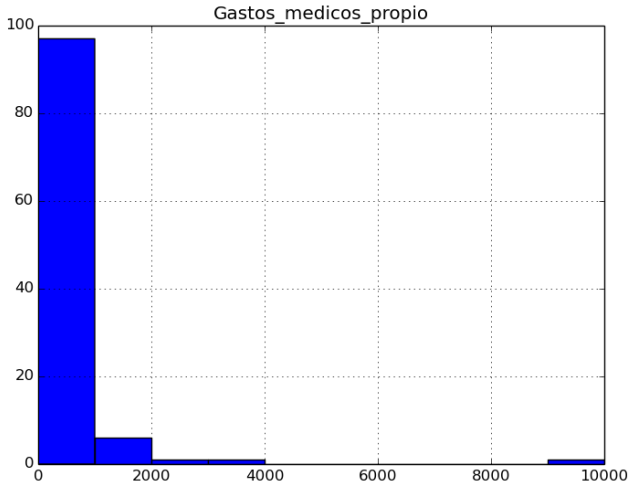
Tenemos valores atípicos que hacen poco conveniente usar la media para imputar los valores ‘nan’, como podemos ver en la **figura 3.**

Figura 3



Una vez que imputamos la mediana a los valores ‘nan’, obtenemos una muestra de datos que varía poco, como se observa en la **figura 4.**

Figura 4



‘Gastos personales propio’: Se consideró que los valores de este campo en teoría no deberían ser 0, ya que las personas por lo general tienen gastos personales que pueden no estar previstos.

Al hacer el histograma de frecuencias se pudo observar que no hay valores atípicos que se alejen de gran manera de la mayoría de los datos de este campo (figura 5).

En este caso no sería mala idea usar la media como estrategia para imputar valores ‘nan’, de hecho la mediana no parece ser la mejor opción ya que la distribución se aleja de la original **(figura 6)**

Figura 5

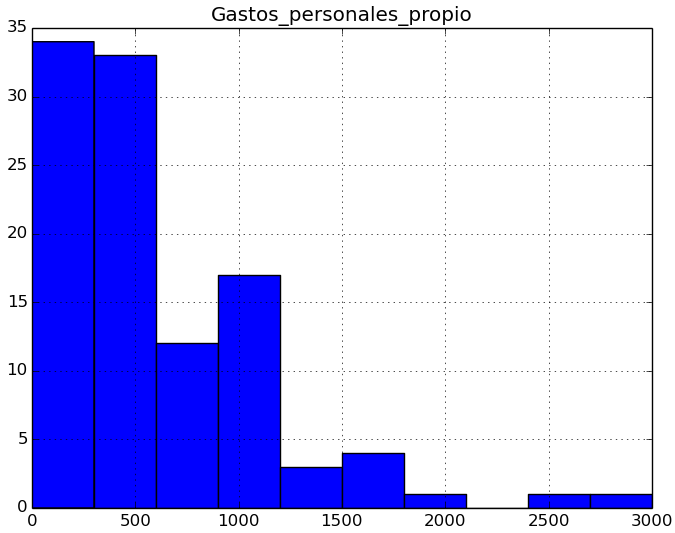
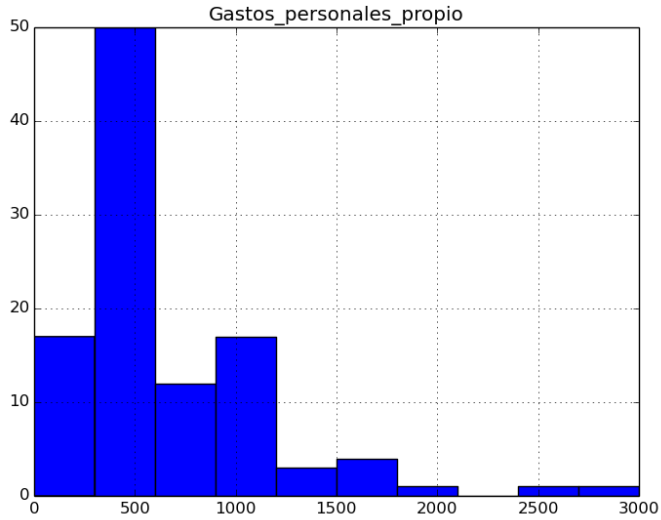
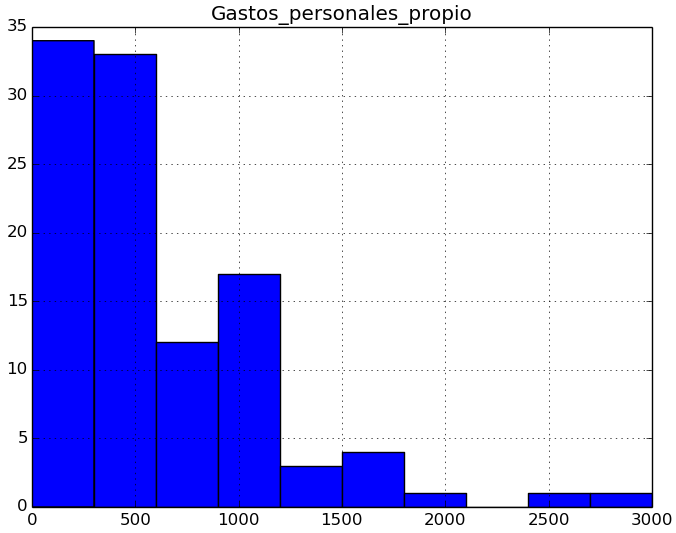


Figura 6



Por este motivo, se eligió como estrategia para imputar la media, de esta forma la distribución de frecuencias quedó bastante parecida a la original **(figura 7)**

Figura 7



VII: En las columnas referentes a los gastos del representante económico se siguieron los criterios anteriores para elegir técnicas para imputar los valores 'nan', dependiendo de si se consideraba que era lógico que algunos valores fuesen ceros, o que otros valores de gastos debían tener algún valor definido

Las columnas referentes a los gastos del representante económico a las que les fueron imputados ceros son:

'Ingreso mensual de su responsable económico': Puede ser cero si el responsable económico se encuentra desempleado, y no tiene ninguna otra entrada de dinero

'Otros ingresos': Simplemente puede no tener otros ingresos

'Vivienda': Si la vivienda es propia, y no paga hipoteca, este valor puede ser cero

'Transporte': Si no tiene que hacer largos desplazamientos o si dispone de vehículo, puede ser cero esta cifra

‘Gastos Odontológicos Representante’: No todas las personas gastan en servicios odontológicos

‘Gastos Educativos’: Este gasto puede ser cero en algunos casos

‘Servicios públicos luz, teléfono, gas’: Algunas personas no pagan estos servicios por alguna razón, por ejemplo las personas que viven en algunos barrios

‘Condominio’: Este valor puede ser cero

A las siguientes columnas referentes a gastos del representante económico, se decidió imputar valores distintos a cero en lugar de los valores ‘nan’, ya que se considera que no tiene sentido que estos valores sean cero, tomando en cuenta el mismo criterio descrito anteriormente en los gastos del estudiante.

‘Alimentación representante’: La persona siempre tendrá gastos en alimentación

‘Gastos Médicos Representante’: Así el representante económico asista a hospitales, y también lleve a estos a las personas de las cuales está encargada, siempre tendrá gastos en medicinas que tendrán que correr por su cuenta

‘Otros Gastos Representante’: Siempre hay gastos no previstos

A estas tres columnas se decidió imputar la media en los valores ‘nan’, ya que al igual que en los gastos del estudiante, se observaron valores atípicos elevados, lo que hacía poco conveniente usar la media como estrategia para imputar valores ‘nan’.