Informe

Edad y Semestre
Con la columna edad y semestre extraje el número ya que resulta mucho más sencillo procesar números que string

Escuela:
Al haber solo 2 escuelas les di un valor numérico el cuál es 1 = Bioquímica y 0 = Enfermera ya que es mucho más fácil hacer el procesamiento con números

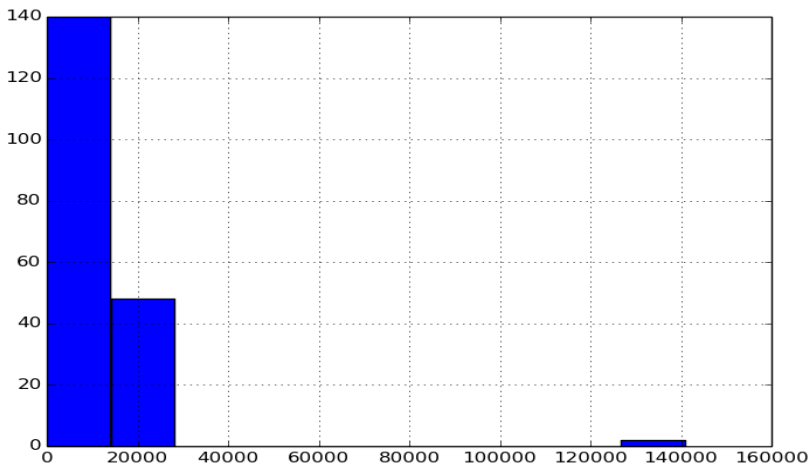
Inscritas anterior – Aprobadas – Reprobadas y Retiradas:
Extraje número de aprobadas y luego sume las aprobadas,reprobadas y retiradas luego elimine aquellas filas que no coincidieran ya que resulta imposible saber en cuál de las columnas existe el error

Período:
En un principio pensé en colocar los periodos en un formato específico de aaaa-I(ejemplo 2014-I) pero es de mucho mayor utilidad usar números por lo tanto al terminar el procesamiento por lo tanto al final queda:
2014-I = 1
2014 – II = 2
2015 -I = 3
2015-II = 4

Fecha de Nacimiento:
Fue todo llevado al estándar dd-mm-aaaa para mayor facilidad en vez de tener distintos estándares

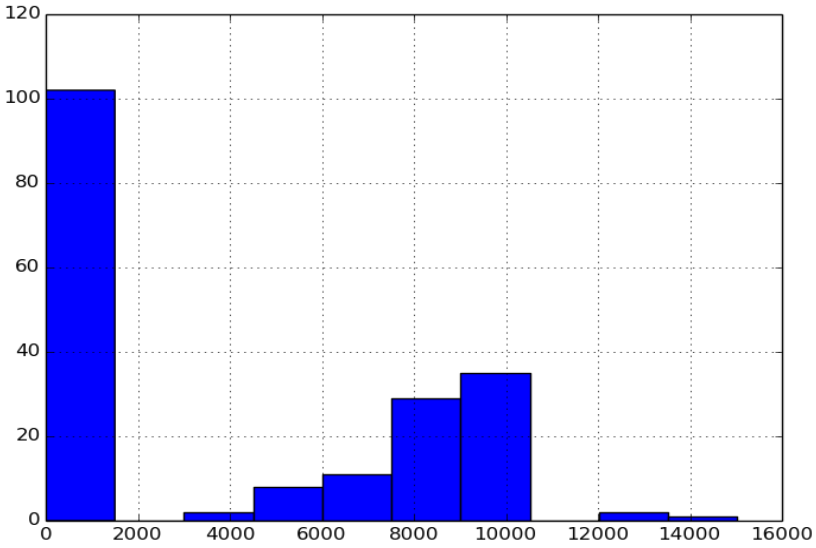
Estado Civil:
Existen 4 estados civiles los cuales represento con un entero, cualquier otro estado civil lo considero como soltero
Soltero = 1, Casado = 2, Viudo = 3 , Divorciado = 4 (no está en la data)

Promedio Aprobado :



Como se puede apreciar en el histograma , los promedios estaban muy por encima del 20 , por lo tanto dividí entre mil y 10 respectivamente para colocar los datos entre 10 y 20

Eficiencia:



Se puede ver en el histograma que hay una gran cantidad de datos que son mucho mayor que 1 , cuando la eficiencia está entre 0 y 1 por lo tanto divido entre múltiplos de 10 para lograr que la eficiencia esté en ese rango

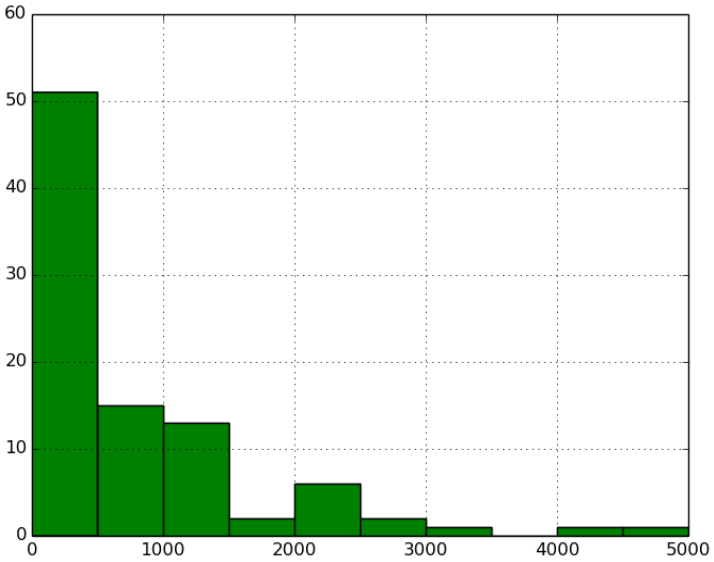
Modalidad de Ingreso
Transforme la modalidad de ingreso en valores numéricos.

asignado por OPSU = 1
prueba interna o propedéutico = 2
convenio interno = 3
convenios interinstitucionales = 4

Tesis Pasantías:
Elimine la columna “X.Estás.realizado.tesis...trabajo.de.grado.o.pasantías.de.grado.” ya que de ser positivo tiene un valor en la columna “Veces_Inscritas_Tesis_Pasantias” y si es negativo tiene el valor 0 por lo tanto esa columna es innecesaria

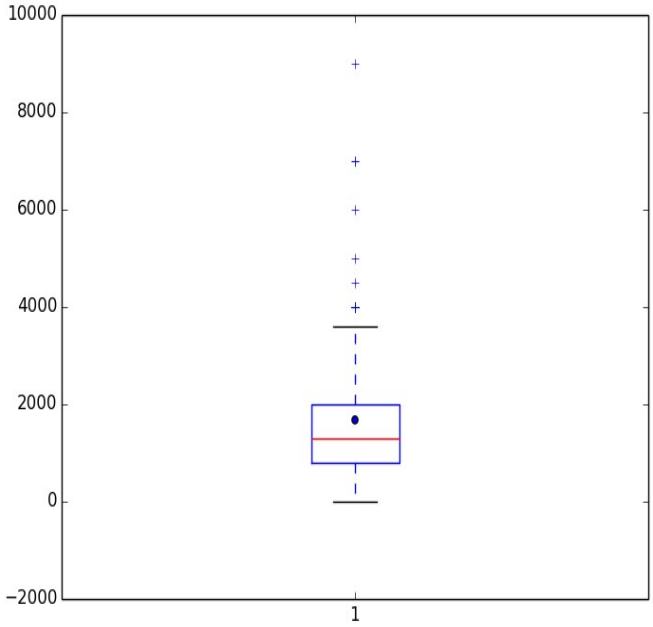
Ingresos Estudiante:
Primero transforme la columna X_Actividad_Ingreso' en 'SI' = 1 y 'No' = 0.

Decidí usar la columna 'X_Actividad_Ingreso' como la verdad sobre si tenía un ingreso adicional o no , si la respuesta es 'No' entonces el ingreso por actividad es 0
1- Aporte de Amigos Estudiante:



Existen muchos valores 'NaN' por eso coloco un histograma en vez de un boxplot , además se puede ver que los valores “outliers” no son demasiado altos por lo tanto no afectan mucho la media , por eso imputo con la media

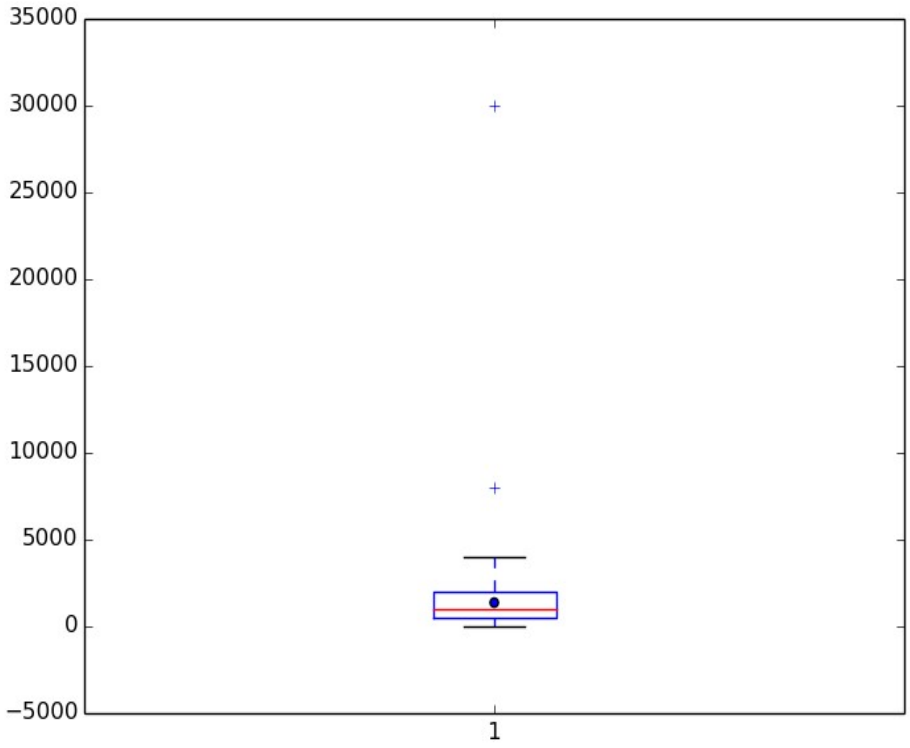
2- Aporte del Responsable al Estudiante:



Se puede observar que existen varios outliers pero ninguno es demasiado significativo por lo tanto es preferible imputar con la media

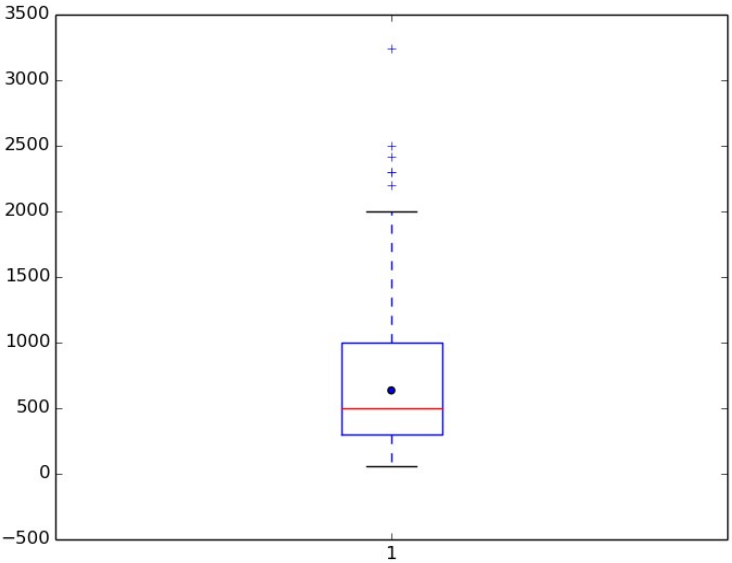
Egresos Estudiante:

1- Alimentación Estudiante:



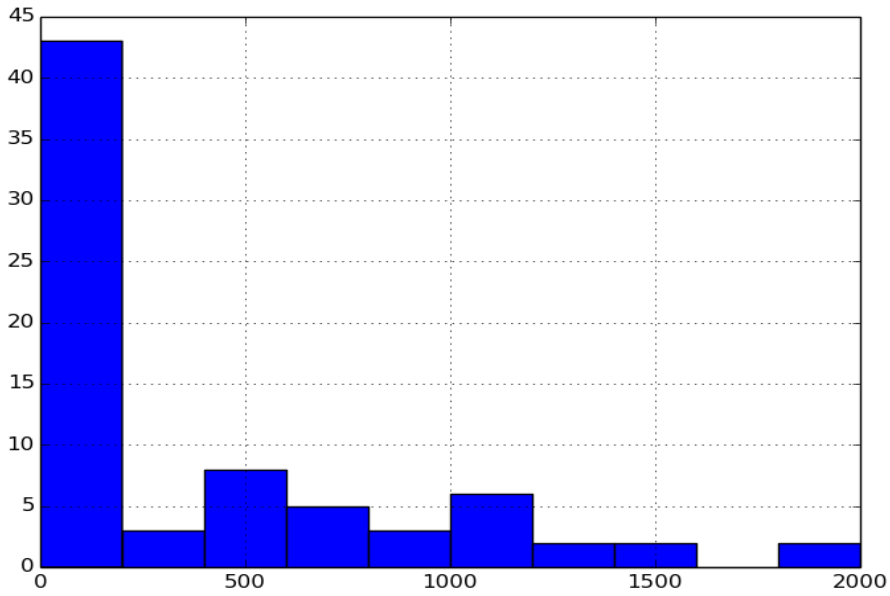
Imputo con la mediana porque existen 2 outliers especialmente el que esta por 30000 – 32000 bs como gasto mensual.

2- Transporte Estudiante:



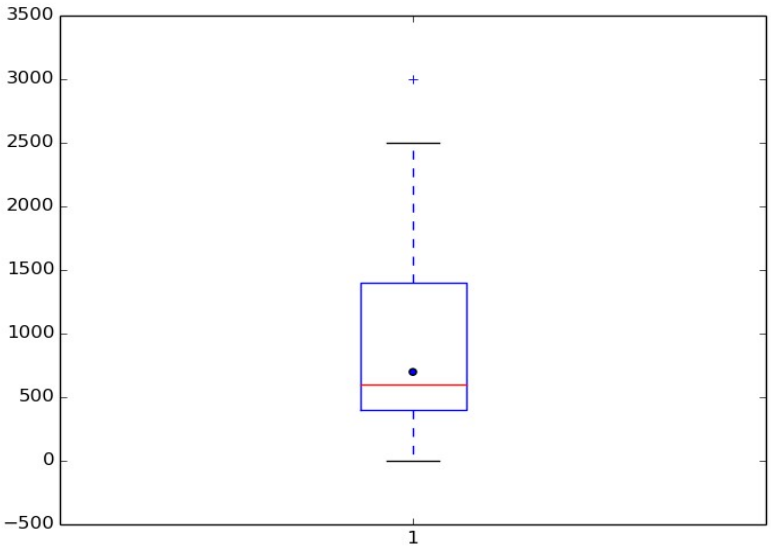
Imputo con media porque los outliers están bastante cerca de los demás valores

3-Gastos Odontológico Estudiante:



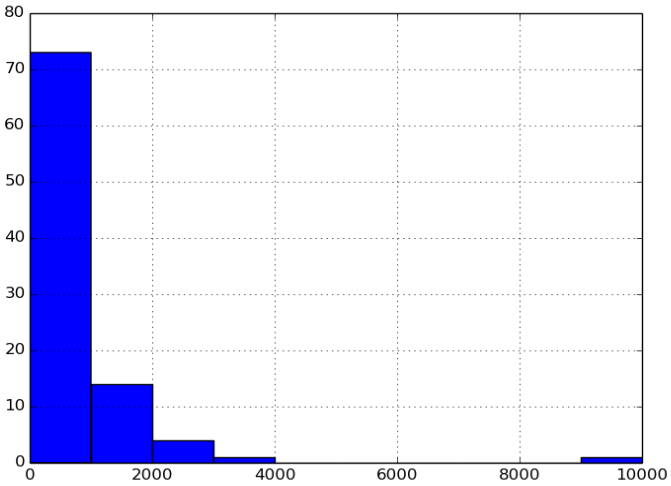
Imputo con la media ya que todos los valores están en un rango relativamente aceptable

4- Gastos Personales Estudiante:



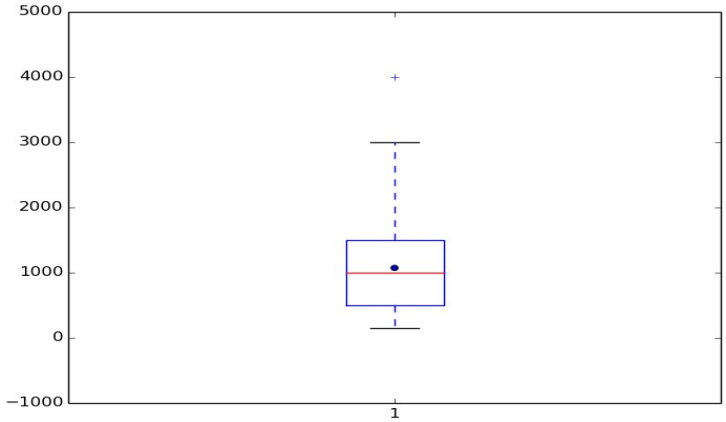
Como sólo existe un outlier y realmente no es TAN diferente a los demás me parece mejor imputar con la media

4-Salud Estudiante:



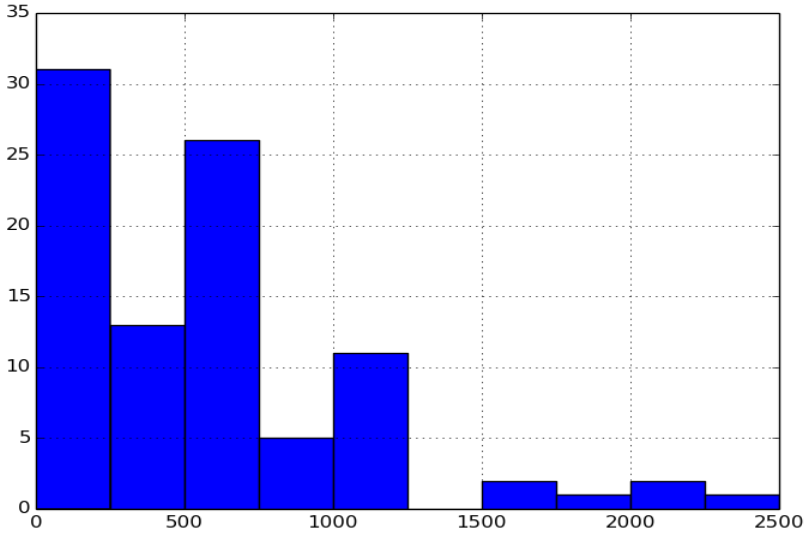
Como existen bastantes valores “bajos” y un solo outlier que tampoco es tán alto me parece que es mejor usar la media para imputar

5-Materiales:



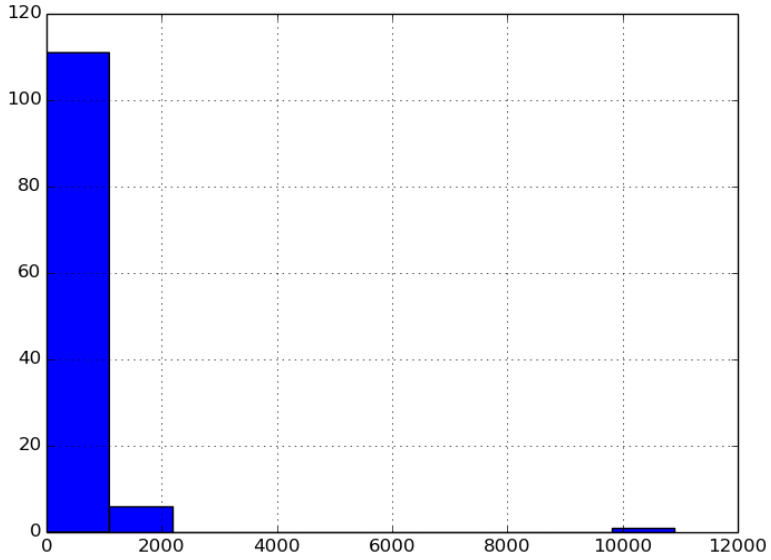
Imputo con la media porque no existen outliers que afecten la media de manera significativa

6- Recreación



Coloco un histograma en vez de un boxplot porque la mediana es 'NaN' y no se ve tan bien con el boxplot en esos casos , de todas maneras imputo con la media porque no existe ningún outlier significativo

7- Otros Gastos Estudiante:



Existen muchísimos 'NaN' (aprox 1/3 de los datos) y no existe un outlier tan significativo por eso prefiero usar la media

8-Monto Mensual Residencia / Gastos Residencia Egresos:

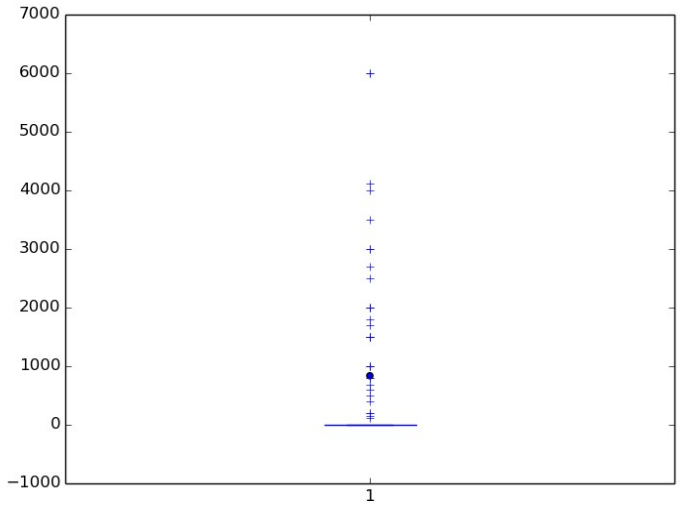
Existen 2 columnas que se refieren a lo mismo que son los gastos por residencia , compare ambos y me quede con el mayor , luego de comparar elimine la columna “Monto Mensual de Residencia”

Ingresos del Responsable:

Primero uso las columnas ingresos y otros ingresos para verificar el total de ingreso y crear consistencia entre estas columnas, luego elimino las columnas de ingreso y otros ingresos porque considero que son irrelevantes, ya que poco importa a un estudio de beca si los ingresos totales son parte de sus ingresos principales o de otros ingresos

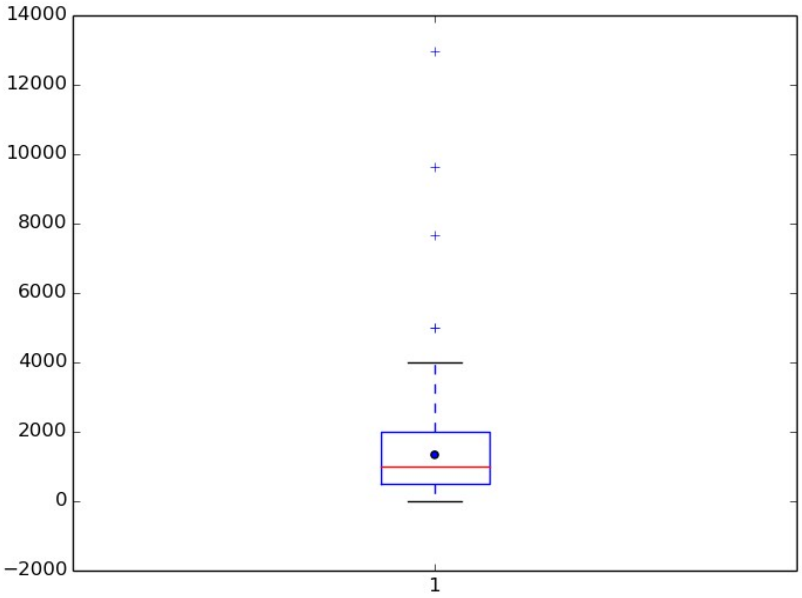
Egresos del Responsable:

1-Vivienda Responsable:



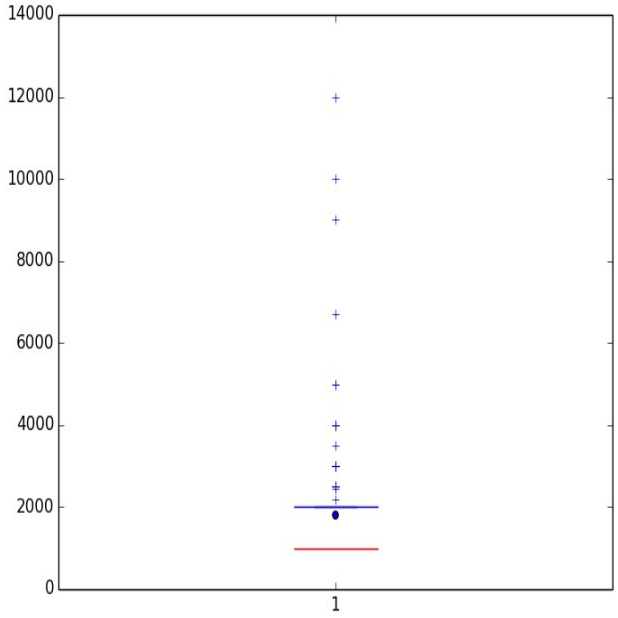
el boxplot queda así porque la mediana es 'NaN', por lo tanto prefiero imputar con la media (aunque el imputer de sklearn ignora los 'NaN')

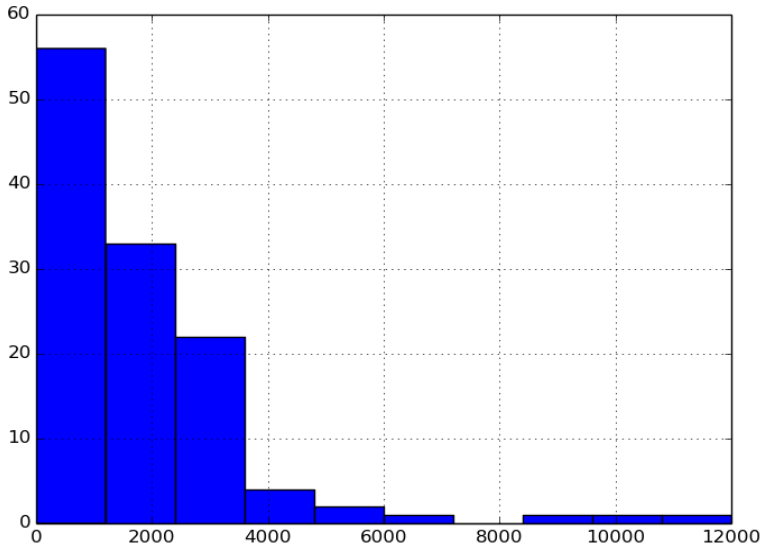
Transporte Responsable:



Como se puede observar en la imagen existen los valores “anomálos” no tienen una diferencia tan alta , pero aún así es una diferencia que afecta por lo tanto prefiero imputar con la mediana

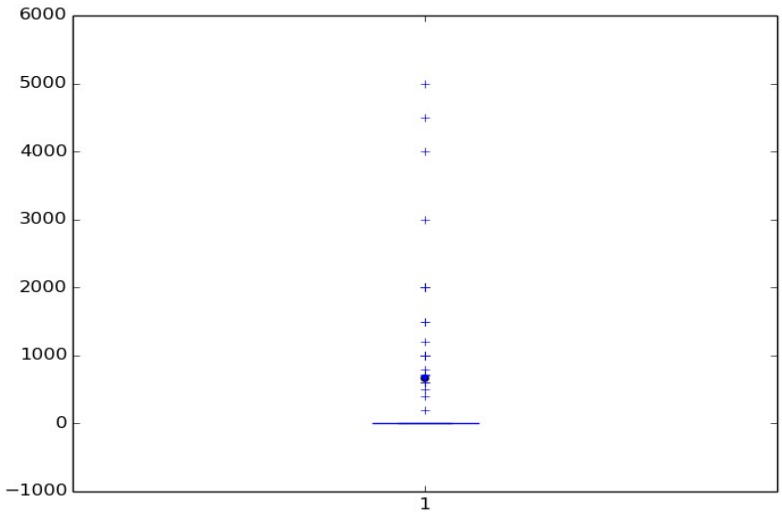
Salud Responsable:





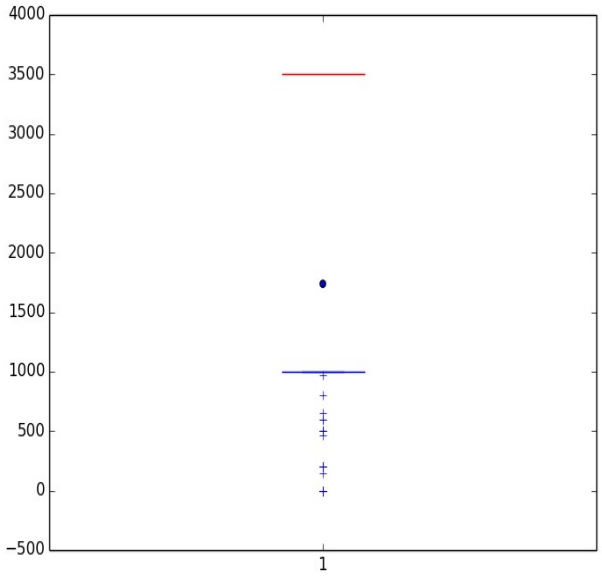
En el boxplot se observar que existen valores atípicos a igual que en el histograma por lo tanto es preferible imputar los 'NaN' con mediana.

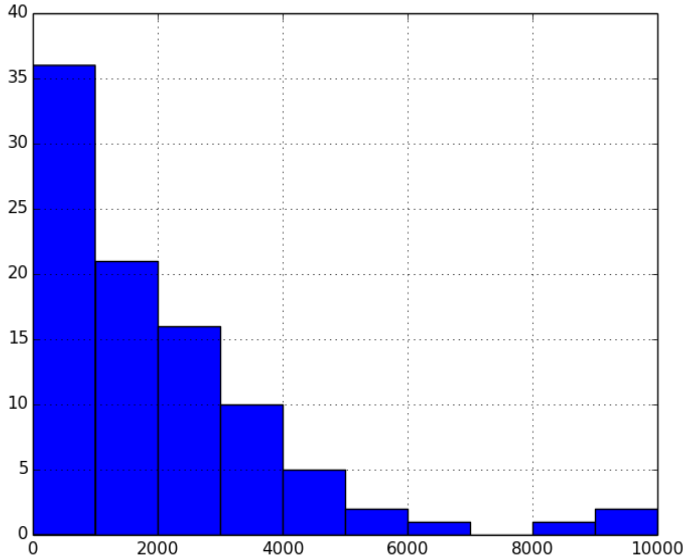
Gastos Odontológico Responsable:



Se puede observar que la mediana es 'NaN' y que los valores no están tan dispersos entre sí por lo tanto imputo con la media

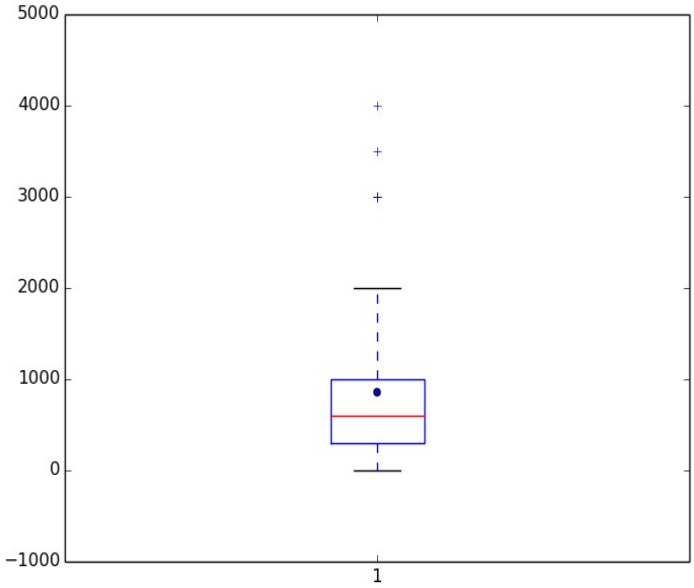
Educativos Responsable:





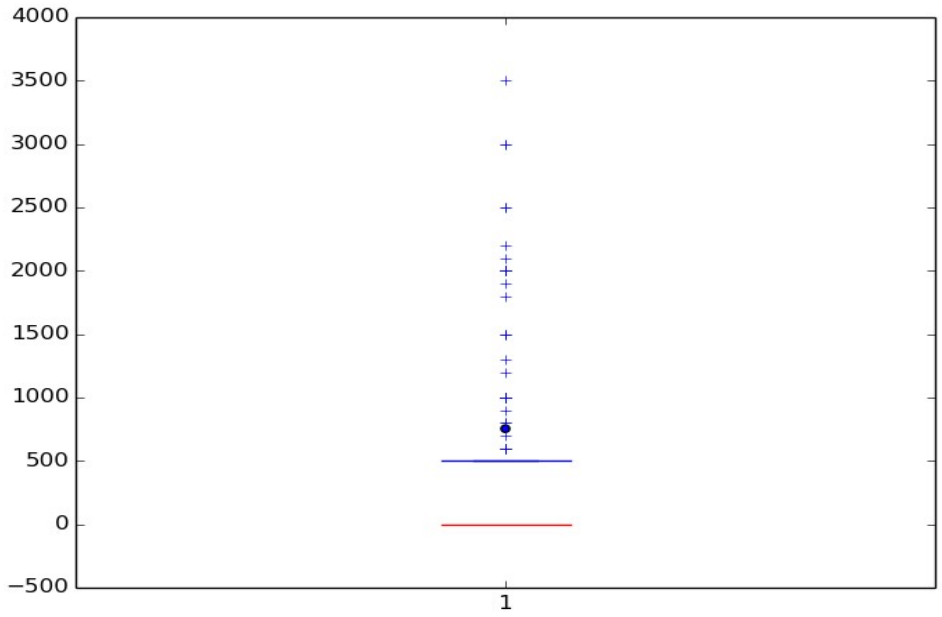
Se puede observar que existen bastantes 'NaN' y que además existen bastantes valores por debajo de la media , por lo tanto prefiero imputar con la media

Gastos Servicios Responsable:



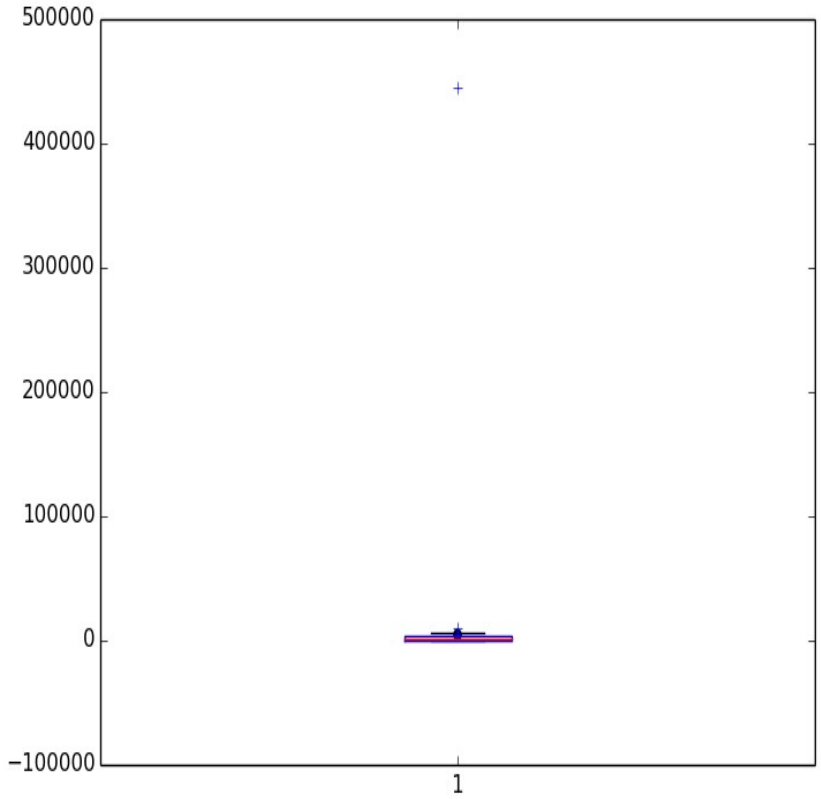
Como los valores atípicos no están muy alejados de los demás la media y la mediana no son tan distantes entre ellos , por ello prefiero usar la media para imputar

Condominio Responsable:



La mediana es cero pero hablando estrictamente de valores realmente no están tan alejados por eso prefiero imputar los 'NaN' con la media

Otros Gastos Responsable:



Se observa que existe un outlier con un valor muy superior al resto por lo tanto imputo con la mediana para que este outlier no afecte el valor de los demás , además no elimino este outlier porque no existe forma de saber si es el valor correcto o si fue un error de escritura en los datos

Total Egresos del Responsable:

Para los totales , compare el total con la suma de los demás egresos (aux) y en caso de que aux fuese menor que el total , le sumo esa diferencia a “Otros Gastos” en caso contrario simplemente el total es igual a aux , además elimino la columna “Total Egresos” porque considero que es una columna calculable a partir de las otras columnas de egresos del Responsable

Otras columnas eliminadas

1- data['Tipo_Actividad_y_Frecuencia']:

Es una columna de la cual su gran mayoría son 'NaN' y los pocos valores que tiene , no son suficientes para realizar un proceso de minería , además son valores totalmente distintos.

2- del data['col0']:

No se sabe qué significa esta columna por lo tanto la elimino

3- del data['X_Matrimonio']:

Es una columna redundante porque ya existe la columna de estado civil

4- Razón de Cambio de Dirección:

Son pocos las filas que tienen un valor en esta columna y son respuestas diferentes que no se pueden categorizar.