Universidad Central de Venezuela
Facultad de Ciencias
Escuela de Computación
Minería de Datos

Documento de explicación de la vista minable

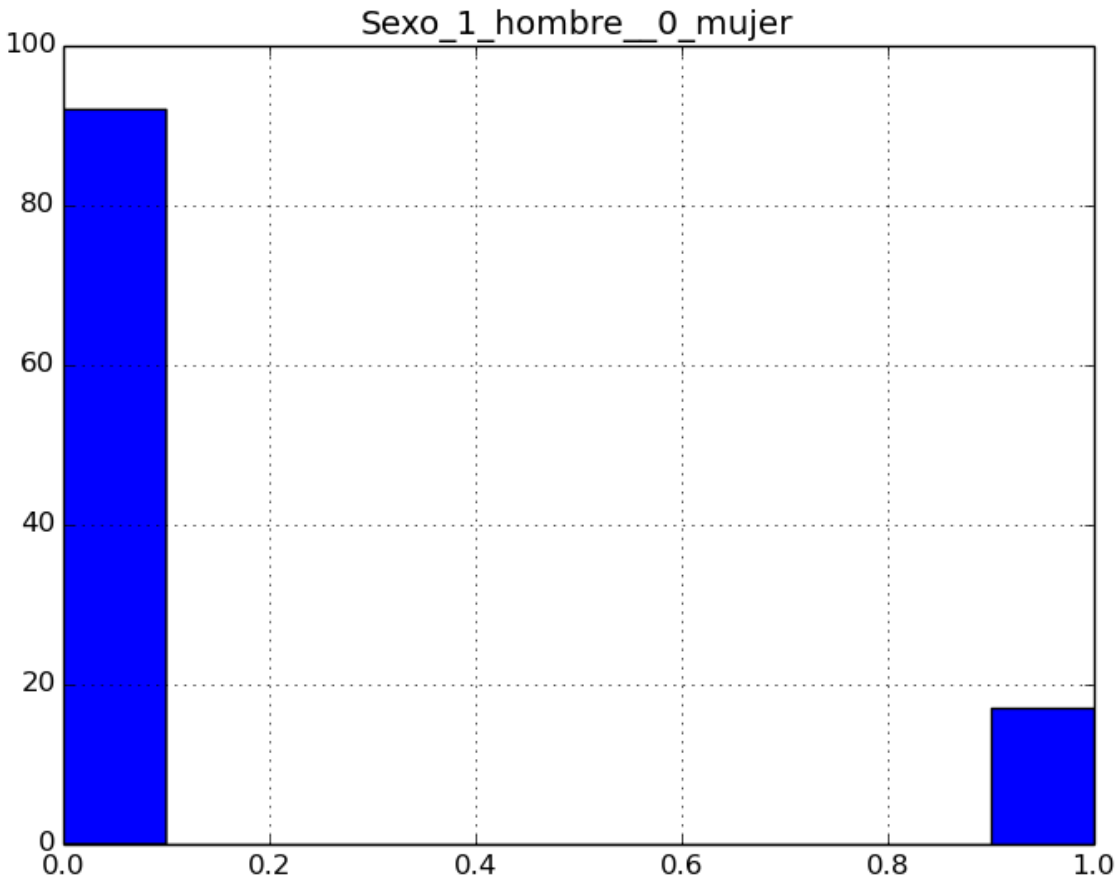
En el presente informe se explicará parte del proceso de limpieza y pre-procesamiento del set de datos inicial, herramientas utilizadas, reducción de la dimensionalidad del set de datos , exclusión de columnas redundantes, acompañado de algunas gráficas que nos mostrarán el comportamiento de los valores de acuerdo a cada columna.

Un grupo de columnas fueron eliminadas por ser duplicados (monto mensual alquiler o residencia y residencia o alquiler) o por ser elementos que pueden ser calculados a través de operaciones aritméticas (Edad, ingresos o egresos totales, entre otras), otras columnas fueron modificadas por su nombre (la mayoría, ya que el lenguaje utilizado requería el
cambio de nombre para poder accesar sin problemas) y otras columnas fueron modificadas o “binarizadas” (Sexo, entre otras).

A continuación, a través de gráficas se explicará algunas columnas resultantes luego del proceso de mineria de datos aplicado a la data inicial.

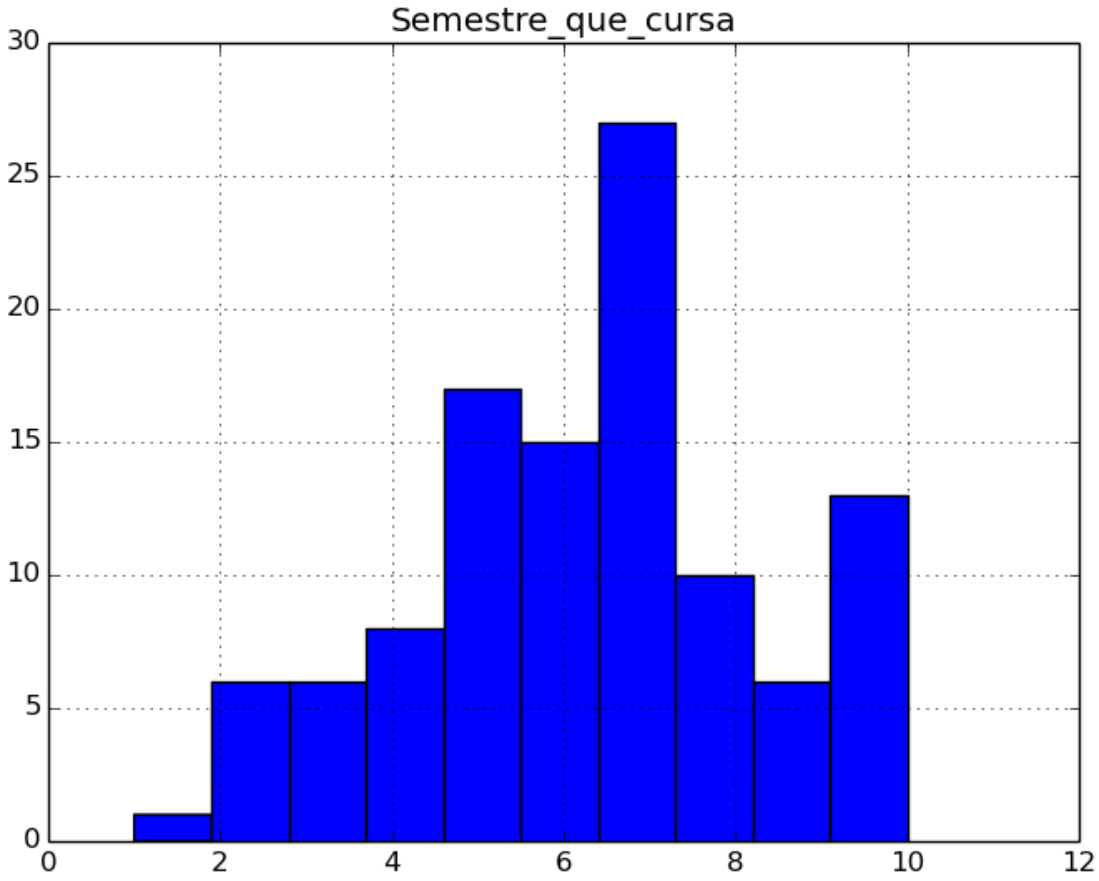
Análisis de Gráficas

csv_file1[['Sexo_1_hombre__0_mujer']].hist()



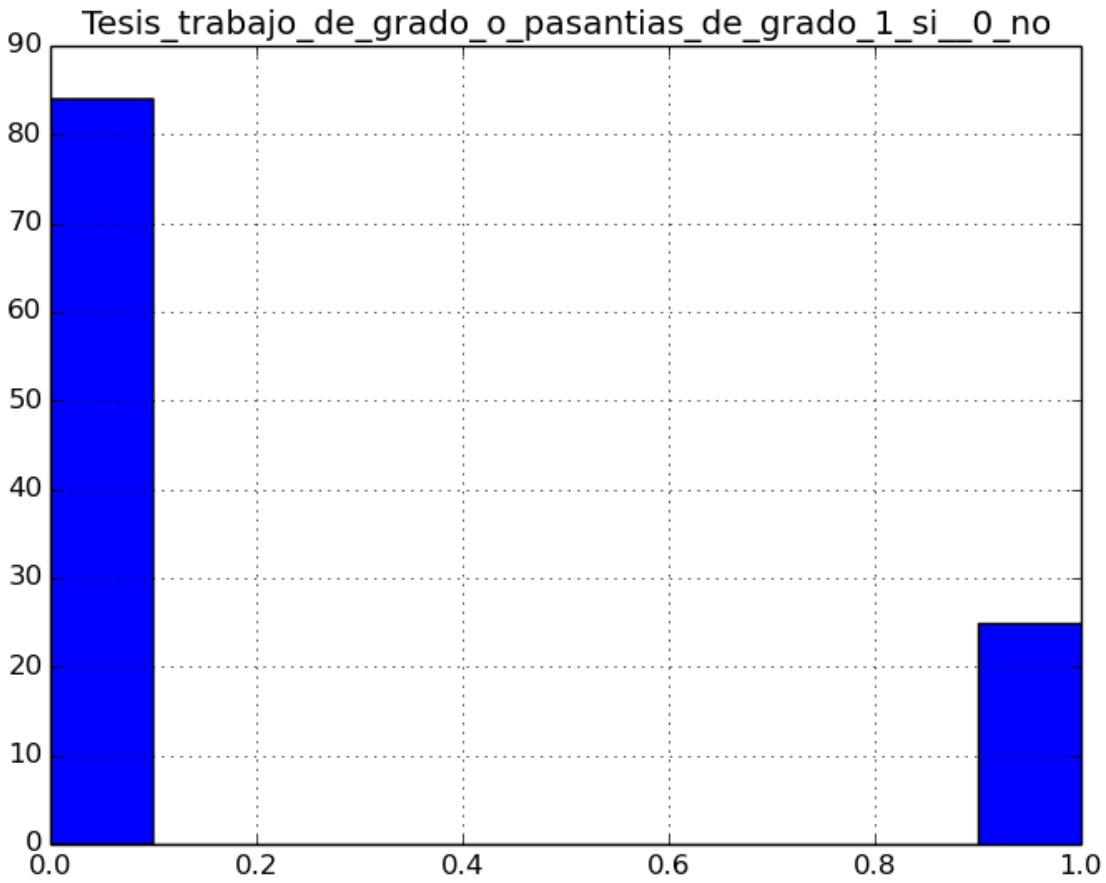
La gráfica muestra la distribución del género sexual de cada uno de los aspirantes a becas, mostrando una cantidad superior de estudiantes de género femenino, por sobre el grupo de estudiantes de género masculino.

csv_file1[['Semestre_que_cursa']].hist()



La gráfica anterior muestra como se distribuyen los aspirantes en sus respectivos semestres, siendo claro el máximo entre los semestres 6to y 8vo y un mínimo entre el 1ro y 2do semestre. Lo que indica que los estudiantes correspondientes a los semestres intermedios (del 5to semestre al 7mo semestre aproximadamente) son los que más han solicitado becas.

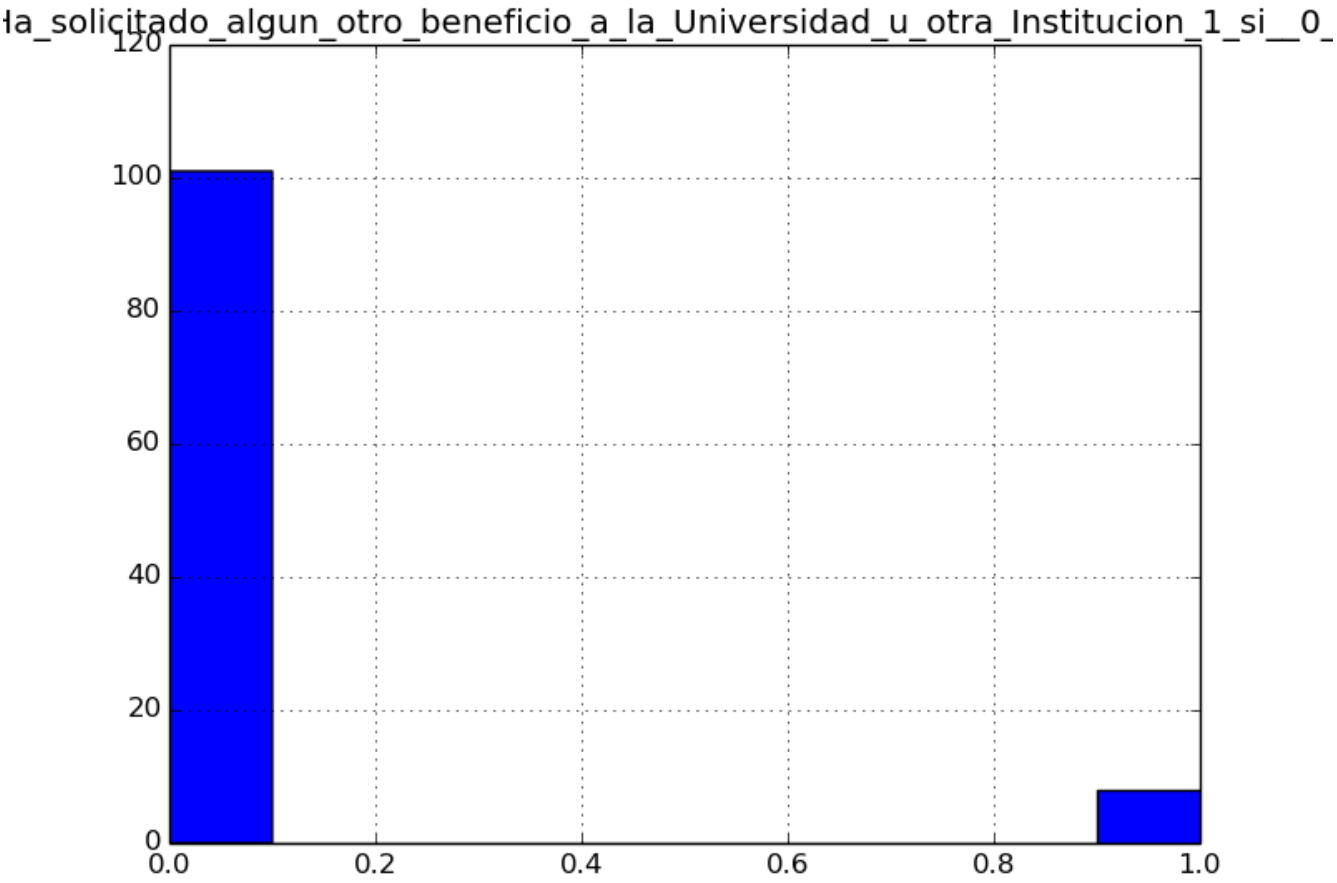
csv_file1[['Tesis_trabajo_de_grado_o_pasantias_de_grado_1_si__0_no']].hist()



La gráfica superior indica que la menor parte de los becarios aspirantes se encuentra realizando el trabajo especial de grado (Aproximadamente 23); difiriendo claramente de aquellos estudiantes que no están en el mismo, el cual

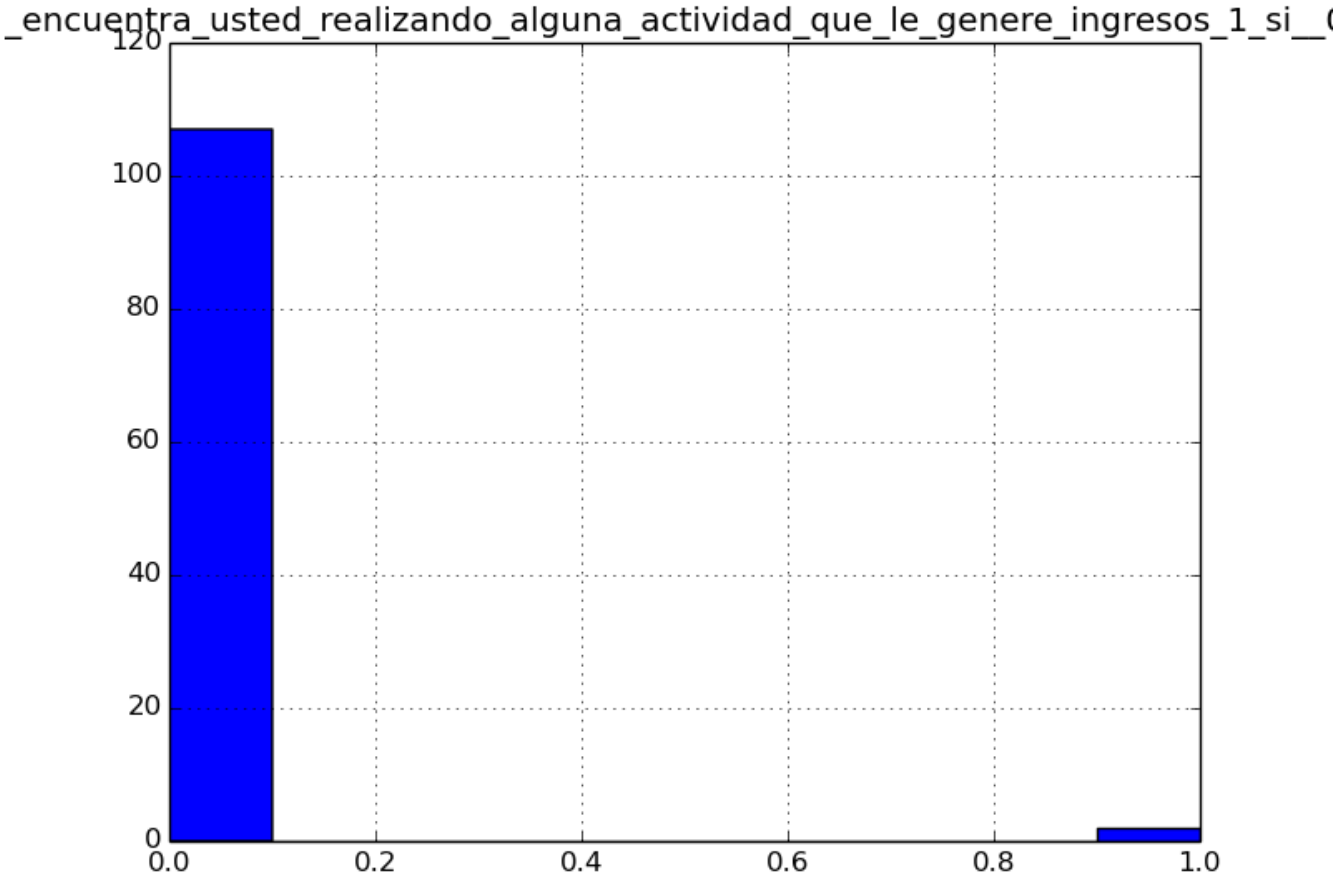
representa una clara mayoría (Aproximadamente 85).

csv_file1[['Ha_solicitado_algun_otro_beneficio_a_la_Universidad_u_otra_Institucion_1_si__0_no']].hist()



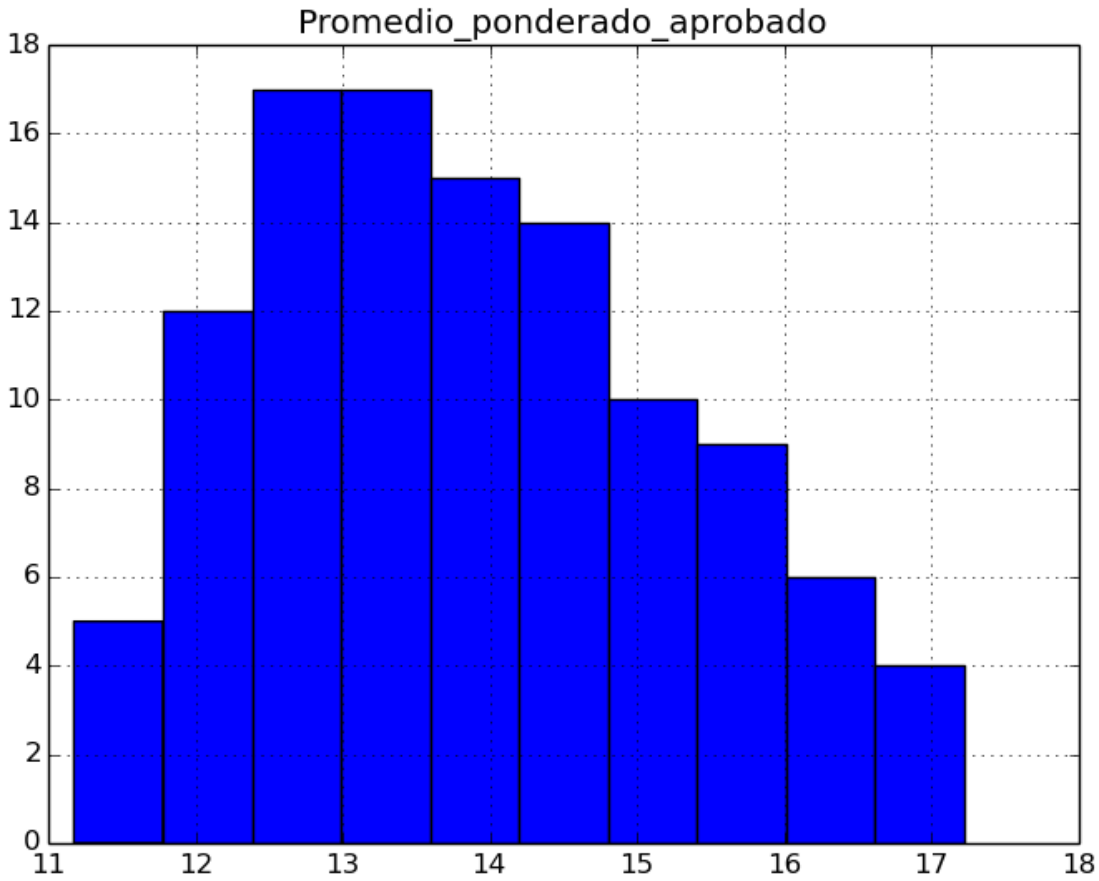
Este gráfico expresa como se distribuyen los estudiantes que han solicitado otro beneficio / beca con respecto a los que no. Se puede observar una gran mayoría que no ha solicitado otro beneficio aparte del que solicitan actualmente (Aproximadamente 100)

csv_file1[['Se_encuentra_usted_realizando_alguna_actividad_que_le_genere_ingresos_1_si__0_no']].hist()



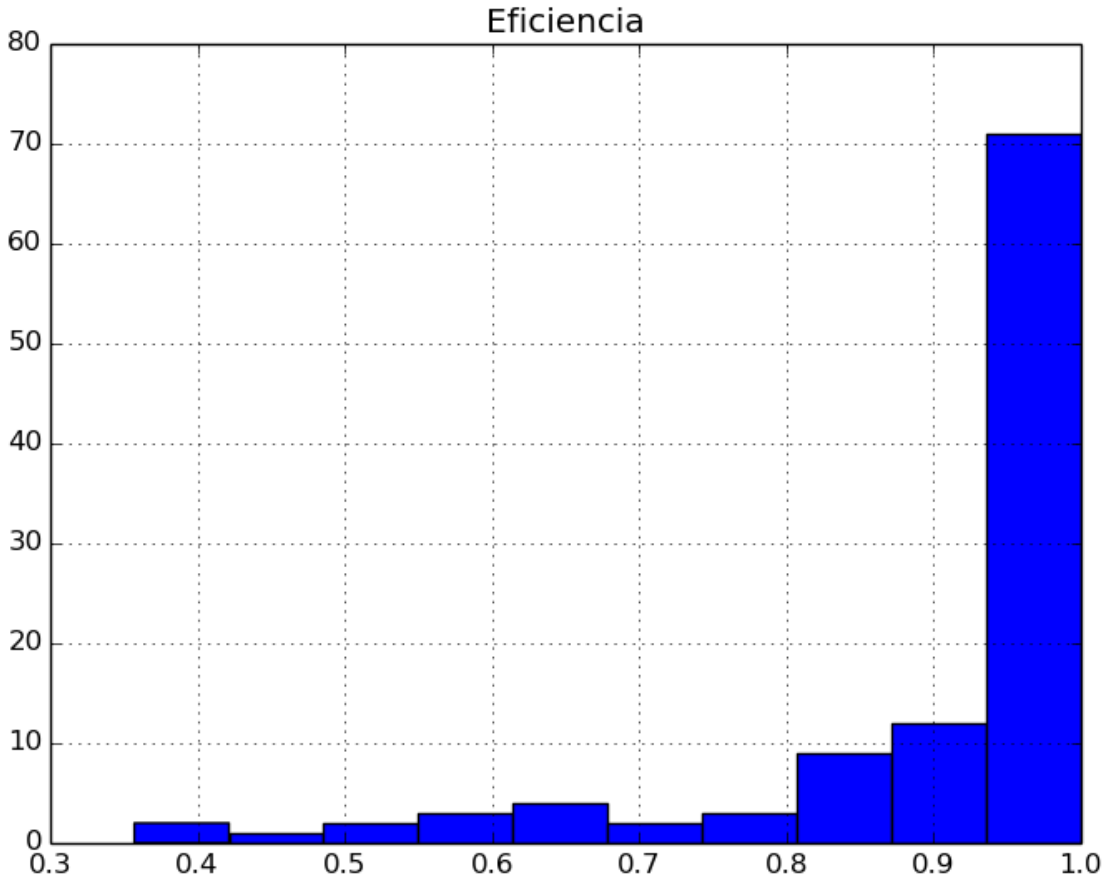
Quizás sea esta la gráfica con mas polarización dentro del informe, debido a que muestra la clara diferencia entre aquellos estudiantes que realizan alguna otra actividad que les genera ingresos (aproximadamente 1 – 5 estudiantes) contra la mayoría que no realiza una actividad de este estilo (aproximadamente 102-105).

csv_file1[['Promedio_ponderado_aprobado']].hist()



El Promedio ponderado es una de las gráficas mas plurales dentro del presente informe debido a que con promedios desde aproximadamente 11.5 puntos hasta promedios que alcanzan el 17.5 puntos aproximadamente, se nota un máximo de notas dentro del rango entre 12 y 14 puntos, que son las notas intermedias del rango total del eje (11-18), permitiendo catalogar el como un comportamiento de función normal respecto a esta columna.

csv_file1[['Eficiencia']].hist()



La eficiencia posee un máximo en 1 (con aproximadamente 70 estudiantes) y el resto de las eficiencias entre 0.35 y 0.99 aproximadamente, se distribuye entre el resto de los aproximadamente 40 estudiantes. Lo que indica que la mayoría de los estudiantes no ha reprobado materias hasta el semestre en que se realizaba el sondeo que dió como resultado el set de datos sobre que el que se trabaja.

Explicación del Código

La lectura del archivo fue realizada con el paquete Pandas y se realizaron cambios a los nombres de algunas de las columnas para que fueran reconocidos como Data Frame por el lenguaje Python, que fue el seleccionado para la realización del trabajo de minería de datos, específicamente limpieza y pre-procesamiento.
Además de esto, se realizó una estandarización de las fechas y los períodos a través de expresiones regulares pertenecientes al paquete 're' (Regular Expression) y a través del paquete Datetime de Python.

Se hicieron recorridos sobre aquellas columnas que requerían estandarizaciones o cambios para lograr así, verificar cuales filas cumplen o no con determinadas condiciones (dependiendo de cada columna) y de esta manera utilizar criterios de filtrado para la escritura en el nuevo archivo con la vista minable. Cabe acotar que con tanto al momento de lectura como con la eliminación de filas, se tuvo que reiniciar el contador de índices para que no hubiese inconsistencia de datos al momento de recorrer las posteriores columnas, y porque en el set de datos inicial los índices estaban desordenados.

Para la estandarización del promedio y la eficiencia se usaron algoritmos condicionales que para determinados rangos lograban establecer valores legibles y acordes a cada columna

Algunos de los valores atípicos fueron estudiados y posteriormente ajustados (por ejemplo, en la columna “eficiencia”, valores mucho mayores a 1) ó eliminados (fila completa, por ejemplo, columna “promedio ponderado” poseía un valor con 0.6)

Conclusión

El set de datos fue estudiado, analizado y posteriormente modificado para lograr una limpieza de datos en la cual se lograron estandarizar valores de una misma columna, lograron eliminarse columnas con valores redundantes y filas con datos anómalos o ilegibles, logrando así la reducción del set de datos, acción que posteriormente lograría una mas efectiva lectura de los datos y análisis a través de gráficas y algoritmos.

Esta primera etapa de limpieza y pre-procesamiento permite así preparar la data, dejando una vista “minable”, para el proceso de minería en el cual se realizarán actividades como el análisis de componentes principales, algoritmos para reconocimiento de patrones o algoritmos para reconocimiento de grupos, entre otros.