Aspectos tomados en cuenta al realizar el preprocesamiento de los datos del sistema de Becas Crema:

Renombramiento de columnas

Se realizó un cambio en los nombres de las variables debido a que muchas tenían caracteres de más o eran muy ambiguas, además de tener un mejor control sobre ellas mediante el DataFrame con un clave más fácil de recordar y acceder.

Eliminación de instancias:

- Se eliminaron columnas como el “ID” que era la primera columna originalmente sin nombre.
- “ContrajoMatrimonio” debido a que ya existe una variable con el estado civil de una persona, por lo que esa es reduntante.
- “"VecesQueHaRealizadoTesisTrabajoDeGradoOPasantíasDeGrado", no es relevante para el sistema.
- “MontoMensualDeViviendaAlquiladaOResidenciaEstudiantil”, ya en las columnas sobre los egresos existe un gasto de viviendas alquiladas o residencias estudiantiles.

Transformación de tipos de datos en las columnas

Gracias a la utilización del paquete de Python-pandas, muchas de las columnas ya tenían un tipo de dato que les confería, sin embargo, en columnas donde se realizó algún cambio se hizo un parseo al tipo de dato deseado (generalmente entero).

Tales fueron los casos como la fecha o el período a renovar, que luego de ser separado en tres columnas para día, mes y año, dichas columnas fueron pasadas a entero.

Y el semestre actual que cursa el estudiante, el cual tenía una parte en texto que fue eliminada ya que con los datos numéricos era suficiente.

Aplicación de estandarización o normalización

Ciertas columnas tenían valores de texto que podían ser estandarizados o correspondidos por valores numéricos que son más fáciles de computar. Las correspondencias son las siguientes:

- Estado civil = {"Soltero (a)": 0, "Casado (a)": 1, "Viudo (a)": 2, "Unido (a)": 3}
- Sexo = {"Masculino": 0, "Femenino": 1}
- Escuela = {"Enfermería": 0, "Bioanálisis": 1}
- ModalidadIngresoUCV = {"Convenios Interinstitucionales (nacionales e internacionales)": 0, "Prueba Interna y/o propedéutico": 1, "Convenios Internos (Deportistas, artistas, hijos empleados docente y obreros, Samuel Robinson)": 2, "Asignado OPSU": 4 }

Reordenamiento de las columnas

Se realizó un cambio en el orden que tenían las columnas colocando la cédula de identidad como primer valor de la vista minable, antes que período a renovar; ya que es un valor que ayuda a identificar un registro al momento de que una persona lea los datos

Limpieza de datos según algún criterio

La columna edad fue preprocesada para eliminar strings que en algunas ocasiones tenía como la palabra “años” luego de la edad para tener sólo números en esa variable.

La columna semestre actual fue tratada de una manera similar a edad, con el uso de expresiones regulares se extrajeron los números que representaban el semestre como tal del individuo y esos fueron los datos que prevalecieron.

La columna de fecha de nacimiento a renovar fue tratada para ser dividida en tres columnas que son para los días, meses y años. Para separar los valores se usaron expresiones regulares para hallar patrones y separar los datos mediante delimitadores que se especificaron. Una vez obtenidos se hizo una búsqueda en los años para arreglar los que no poseían el año completo sino la abreviación.