**TITULO PROYECTO:** Depression Risk Analysis

**INTEGRANTES:** Luis Alejandro Londoño Martínez, Simón Correa Marín

1. **ENTENDIMIENTO DEL NEGOCIO**
   1. **DESCRIPCIÓN DEL NEGOCIO**

El **proyecto Depression Risk Analysis** se centra en analizar los factores de riesgo que pueden influir en la probabilidad de desarrollar depresión en adultos. Su objetivo principal es identificar patrones y correlaciones en las respuestas de una encuesta de salud mental para predecir el riesgo de depresión, basándose en datos de estilo de vida y demográficos de personas entre 18 y 60 años. A través de este análisis, se espera aportar a la comprensión de cómo variables como la satisfacción laboral, horas de estudio o trabajo, antecedentes familiares y otros factores de la vida cotidiana se relacionan con el riesgo de sufrir depresión.

Este proyecto se sitúa en el contexto de la **investigación en salud mental,** específicamente en la identificación de factores clave que afectan el bienestar psicológico en un entorno no clínico. La información obtenida se utilizará para desarrollar modelos predictivos de aprendizaje supervisado que podrían emplearse en la formulación de estrategias preventivas y en la toma de decisiones para intervenciones tempranas en el ámbito de la salud pública.

* 1. **DESCRIPCIÓN DEL PROBLEMA**

El problema específico que busca resolver este proyecto es identificar los factores cotidianos y demográficos que se asocian con el riesgo de depresión en adultos. A falta de evaluaciones clínicas y diagnósticos profesionales en este contexto, el desafío radica en emplear un modelo predictivo basado en datos de estilo de vida y satisfacción personal para predecir con precisión el riesgo de depresión en un grupo poblacional diverso. Esta herramienta permitirá entender mejor cómo ciertos factores pueden influir en la salud mental y proporcionar una base para la intervención temprana y la toma de decisiones informadas en materia de bienestar.

El problema que se busca abordar es la **falta de herramientas predictivas para evaluar el riesgo de depresión** en adultos con base en factores no clínicos y de estilo de vida.

* 1. **OBJETIVOS DE LA MINERÍA**
* Identificar los factores clave asociados al riesgo de depresión en adultos mediante un análisis exploratorio y predictivo de los datos recolectados.
* Construir modelos de aprendizaje supervisado para la clasificación del riesgo de depresión. Se emplearán cinco algoritmos de clasificación:
  + Máquinas de soporte vectorial para clasificación (SVM)
  + Red Neuronal para clasificación (ANN)
  + Árboles de decisión
  + K-Nearest Neighbors (KNN)
  + Regresión Logística

Estos modelos se entrenarán sobre un conjunto de datos balanceado al 70%.

* Aplicar cuatro métodos de ensamble para mejorar la robustez y precisión de los modelos predictivos:
  + Random Forest, que combina múltiples árboles de decisión de forma aleatoria.
  + XGBoost y CatBoost, técnicas avanzadas de boosting que ajustan iterativamente los modelos para minimizar los errores residuales.
  + **Voting Hard**, un ensamble que combina múltiples modelos de clasificación mediante un sistema de votación.
* Comparar y evaluar el rendimiento de cada modelo utilizando al menos cuatro métricas de calidad, como precisión, recall, F1-score y curva ROC, además de la matriz de confusión.
* Seleccionar los tres mejores modelos utilizando un análisis estadístico de diferencia significativa (ANOVA y prueba de Tukey) para garantizar que la elección de los modelos tenga una base estadísticamente válida.
* Optimizar los tres modelos seleccionados mediante hiperparametrización usando GridSearch y métodos avanzados como BayesSearchCV (optimización bayesiana) y **GASearchCV** (algoritmos genéticos) para maximizar el desempeño predictivo de cada modelo
* **Desplegar el modelo final** mediante un Pipeline de preparación de datos, integrándolo en una interfaz gráfica desarrollada con **Streamlit.** Esta interfaz permitirá al usuario cargar datos, ejecutar el modelo y visualizar las predicciones de riesgo, proporcionando una herramienta funcional y accesible para la toma de decisiones en salud mental y bienestar psicológico. El despliegue se realizará con **LocalTunnel** para generar un enlace público que facilite el acceso remoto a la aplicación, permitiendo que otros usuarios interactúen con la herramienta de forma sencilla y segura.
  1. **DISEÑO DE SOLUCIÓN**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Problema | Tipo de Análisis | Tipo de Aprendizaje | Tarea Analítica | Requerimientos en los Datos | Métodos | Evaluación | Calidad Esperada |
| Construir modelos de clasificación para predecir el riesgo de depresión e identificar los factores asociados al riesgo. | Predictivo | Supervisado | Análisis exploratorio y clasificación del riesgo | Datos de calidad con variables demográficas, sociales y laborales balanceados al 70% para evitar sesgos | Modelos: SVM, ANN, Árbol de decisión, KNN y Regresión Logística | Matriz de confusión, Precisión, recall, F1-score, curva ROC | Modelos de clasificación con precisión y sensibilidad superiores al 80%. |
| Mejorar la precisión de los modelos utilizando técnicas de ensamble | Predictivo | Supervisado | Ensamble para precisión y estabilidad | Datos limpios y preprocesados; balanceo previo al 70%; partición en conjuntos de entrenamiento y prueba | Ensambles: Random Forest, XGBoost, CatBoost, Voting Hard (con voto mayoritario) | Precisión, recall, F1-score, curva ROC | Modelos ensamblados con desempeño superior a modelos individuales. Aumento en precisión y robustez, con mejora del 5-10% en métricas sobre modelos individuales. |
| Seleccionar los mejores modelos mediante un análisis estadístico de significancia | No aplica | No aplica | Comparación y selección de modelos | Conjunto de datos de entrenamiento y validación, asegurando balanceo previo | ANOVA y prueba de Tukey | Diferencia estadística en rendimiento entre modelos | Selección de modelos basada en significancia estadística con una confiabilidad mínima del 95%. |
| Optimizar el rendimiento de los modelos seleccionados a través de hiperparametrización | No aplica | No aplica | Optimización de hiperparámetros | Conjunto de datos de validación balanceado para GridSearch y optimización avanzada de hiperparámetros | Hiperparametrización con GridSearch, BayesSearchCV y GASearchCV para modelos seleccionados | Matriz de confusión, Precisión, recall, F1-score, curva ROC tras optimización | Mejora en el rendimiento de los modelos seleccionados en al menos un 5% en comparación con los resultados iniciales. |
| Desplegar el modelo final en una interfaz gráfica para predicción del riesgo de depresión | No aplica | No aplica | Despliegue mediante una interfaz gráfica | Pipeline de datos con preprocesamiento completo, normalización, codificación y preparación de datos para despliegue | Implementación en Streamlit, con despliegue público mediante LocalTunnel | No aplica | Interfaz funcional y amigable para el usuario |

* 1. **RECURSOS PARA CREACIÓN DEL MODELO Y PARA DESPLIEGUE**

|  |  |
| --- | --- |
| Aspecto | Detalles |
| Entorno de Desarrollo | La creación del modelo se llevará a cabo de forma **local** utilizando **Jupyter Notebook** en **Visual Studio Code**. |
| Lenguaje de Programación | Se empleará **Python** tanto para la creación y entrenamiento del modelo como para el despliegue en Streamlit. |
| Librerías y Herramientas | - **Limpieza de datos**: pandas, numpy, sklearn  - **Entrenamiento y evaluación de modelos**: scikit-learn - **Optimización y búsqueda de hiperparámetros**: GridSearchCV, BayesSearchCV, GASearchCV  - **Visualización de datos**: matplotlib, seaborn  - **Despliegue**: Streamlit, LocalTunnel |
| Repositorio y Control de Versiones | Los cambios del proyecto se subirán a un repositorio remoto en **GitHub**, permitiendo un control de versiones adecuado y colaboración si es necesario. |
| Despliegue del Modelo | - El modelo se desplegará en una aplicación de **Streamlit**, donde los usuarios podrán cargar datos, ejecutar el modelo y ver los resultados en una interfaz gráfica.  - Se utilizará **LocalTunnel** para generar un enlace público y permitir el acceso remoto a la aplicación, facilitando el uso del modelo en otros dispositivos o ubicaciones sin necesidad de configuración compleja. |
| Interacción Modelo-Interfaz | La interfaz de Streamlit cargará el modelo desde un archivo preentrenado almacenado localmente. Streamlit ejecutará el modelo en tiempo real y mostrará los resultados al usuario, asegurando una comunicación directa entre el modelo y la interfaz. |
| Licencias y Requisitos Legales | - **Licencia de Python y bibliotecas**: Python es de código abierto, y las bibliotecas seleccionadas tienen licencias que permiten el uso y modificación libre. - **Visual Studio Code**: de código abierto, con licencias de uso gratuito para desarrolladores.  - **Streamlit y LocalTunnel**: también de código abierto y gratuitos. |
| IDE | **Visual Studio Code** para la creación y pruebas locales del modelo en Jupyter Notebooks. |

1. **ENTENDIMIENTO DE LOS DATOS**
   1. **CICLO DE LOS DATOS**

**Generación de los datos**  
Los datos se generan a partir de una encuesta anónima que fue diseñada específicamente para evaluar factores demográficos, sociales y de estilo de vida relacionados con el riesgo de depresión en adultos. La encuesta fue distribuida en varias ciudades y capturó información autodeclarada de personas de entre 18 y 60 años sobre factores como la satisfacción laboral, horas de trabajo/estudio, antecedentes familiares, entre otros.

**Almacenamiento de los datos**  
Actualmente, los datos recolectados están almacenados en un archivo CSV que ha sido descargado de Kaggle. Este archivo se gestiona de manera local, pero se almacena en un repositorio remoto en GitHub para facilitar el control de versiones y permitir la colaboración entre los integrantes del equipo.

**Modificación de los datos**  
Dado que se trabaja con un archivo descargado, la manipulación de datos se hace por los miembros del equipo de análisis de datos. Las operaciones de preprocesamiento, limpieza y transformación de datos se realizarán mediante scripts en Python.

**Periodicidad de los datos y reentrenamiento del modelo.**  
Los datos de la encuesta fueron recolectados en un solo periodo (enero a junio de 2023), lo que significa que el conjunto de datos es estático y no recibirá actualizaciones periódicas. En caso de futuras expansiones del proyecto o de la encuesta, se podrían establecer nuevas fases de recolección de datos. De ser así, el modelo podría ser reentrenado cada 2-3 años si se recopilan datos adicionales que reflejen cambios en los factores de riesgo para la depresión.

* 1. DICCIONARIO DE DATOS

|  |  |  |
| --- | --- | --- |
| Variable | Tipo | Descripción |
| Name | Categórica | Nombre del participante (identificador anonimizado). |
| Gender | Categórica | Género del participante. |
| Age | Numérica | Edad del participante, en años. |
| City | Categórica | Ciudad de residencia del participante. |
| Working Professional or Student | Categórica | Estado laboral del participante. |
| Profession | Categórica | Profesión del participante, si aplica. |
| Academic Pressure | Categórica | Nivel de presión académica percibida. |
| Work Pressure | Categórica | Nivel de presión laboral percibida. |
| CGPA | Numérica | Calificación promedio acumulada (solo para estudiantes). |
| Study Satisfaction | Categórica | Nivel de satisfacción con los estudios, solo para estudiantes. |
| Job Satisfaction | Categórica | Nivel de satisfacción laboral, solo para profesionales en activo. |
| Sleep Duration | Categórica | Duración promedio del sueño del participante. |
| Dietary Habits | Categórica | Hábitos alimenticios. |
| Degree | Categórica | Grado académico máximo obtenido por el participante. |
| Have you ever had suicidal thoughts? | Categórica | Si el participante ha tenido pensamientos suicidas. |
| Work/Study Hours | Numérica | Promedio de horas dedicadas al trabajo o estudio diariamente. |
| Financial Stress | Categórica | Nivel de estrés financiero percibido. |
| Family History of Mental Illness | Categórica | Si el participante tiene antecedentes familiares de enfermedades mentales. |
| Depression | Categórica | Variable objetivo: riesgo de depresión, basada en la evaluación de factores demográficos y de vida. |

* 1. **REGLAS DE CALIDAD**

|  |  |
| --- | --- |
| Variable | Regla de calidad |
| Name | ------- |
| Gender | Entre 18 y 60 años. |
| Age | Male, Female. |
| City | Debe coincidir con una ciudad válida en el país. Cualquier otra entrada es un error. |
| Working Professional or Student | Working Professional, Student. |
| Profession | Solo se permiten valores entre las profesiones listadas (e.g., Teacher, HR Manager, Doctor, etc.). |
| Academic Pressure | 1, 2, 3, 4, 5. |
| Work Pressure | 1, 2, 3, 4, 5. |
| CGPA | Entre 5.0 y 10.0. Cualquier otro valor es un error. |
| Study Satisfaction | 1, 2, 3, 4, 5. |
| Job Satisfaction | 1, 2, 3, 4, 5. |
| Sleep Duration | Less than 5 hours, 5-6 hours, 7-8 hours, More than 8 hours. |
| Dietary Habits | Healthy, Moderate, Unhealthy. |
| Degree | Debe coincidir con un título académico válido listado en la base de datos (e.g., B.Tech, B.Sc, M.Tech, PhD, etc.). |
| Have you ever had suicidal thoughts? | Yes, No. |
| Work/Study Hours | Entre 0 y 12 horas diarias. |
| Financial Stress | 1, 2, 3, 4, 5. |
| Family History of Mental Illness | Yes, No. |
| Depression | Yes, No. |

1. **PREPARACIÓN DE DATOS (Estadística)**
   1. **INTEGRACIÓN**

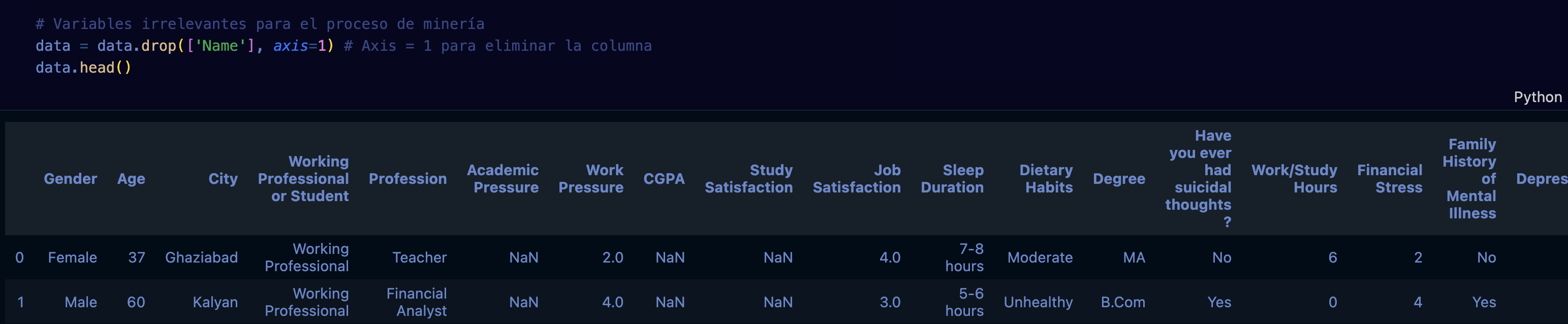
Se consolidará una única tabla (sábana de datos) en la que se integran todas las variables necesarias para el análisis. En este caso, los datos están en un solo archivo CSV, por lo que no se requieren uniones adicionales o joins.

Al inicio, se tienen 19 variables en el dataset.

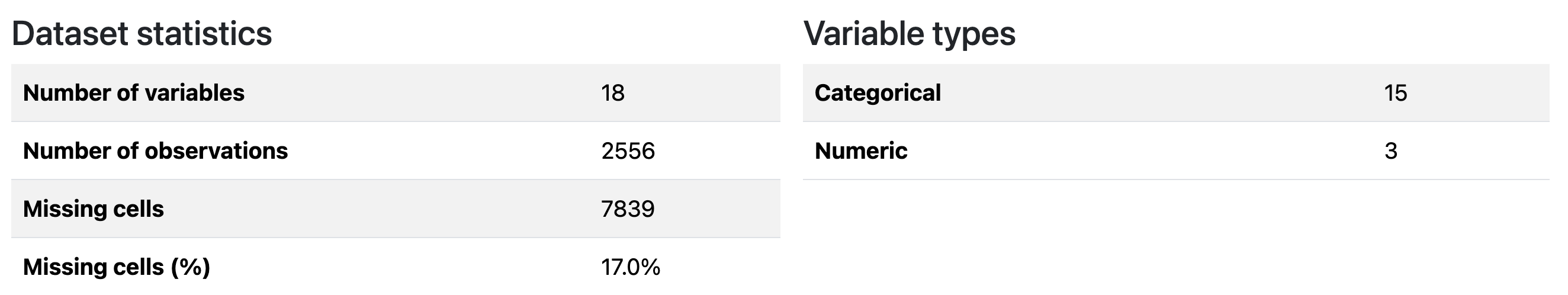
* 1. **SELECCIÓN DE VARIABLES**

Se eliminan las variables irrelevantes o que no aporten al análisis, así como aquellas que puedan comprometer la privacidad de los participantes. En este caso, la variable *Name* no es útil para el modelo y será eliminada.

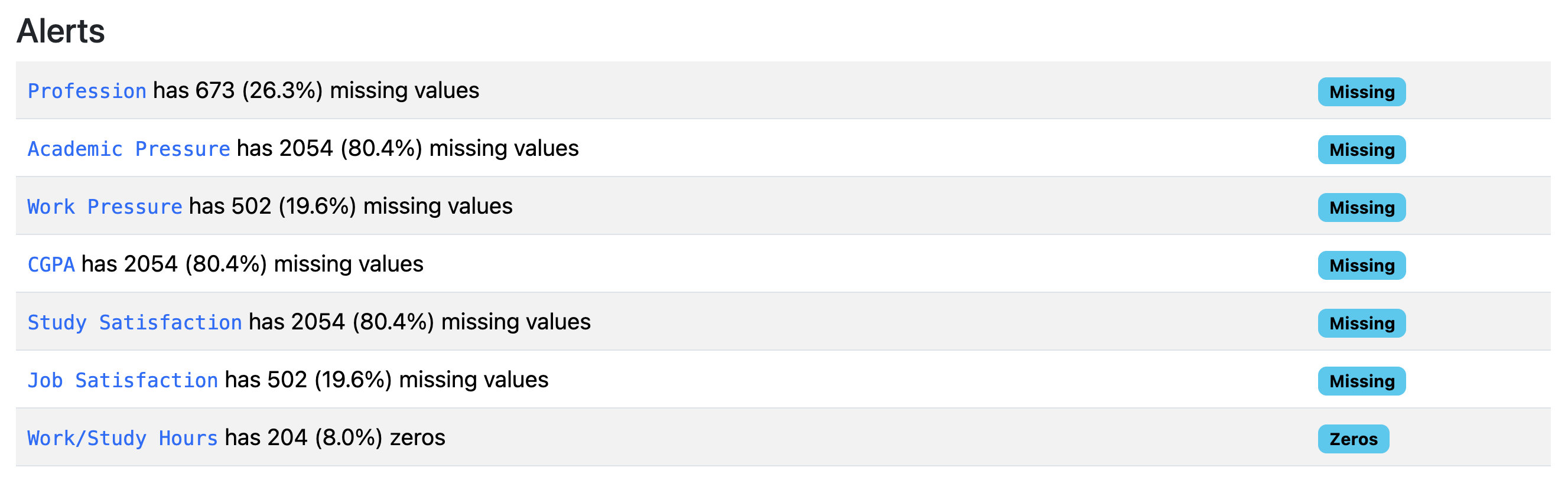


* 1. **DESCRIPCIÓN ESTADÍSTICA**

Se realizará un análisis estadístico descriptivo de las variables numéricas y categóricas.



El dataset contiene 2556 registros y 19 variables, de las cuales 15 son categóricas y 3 numéricas. Hay 7839 celdas nulas que equivalen al 17% del dataset.



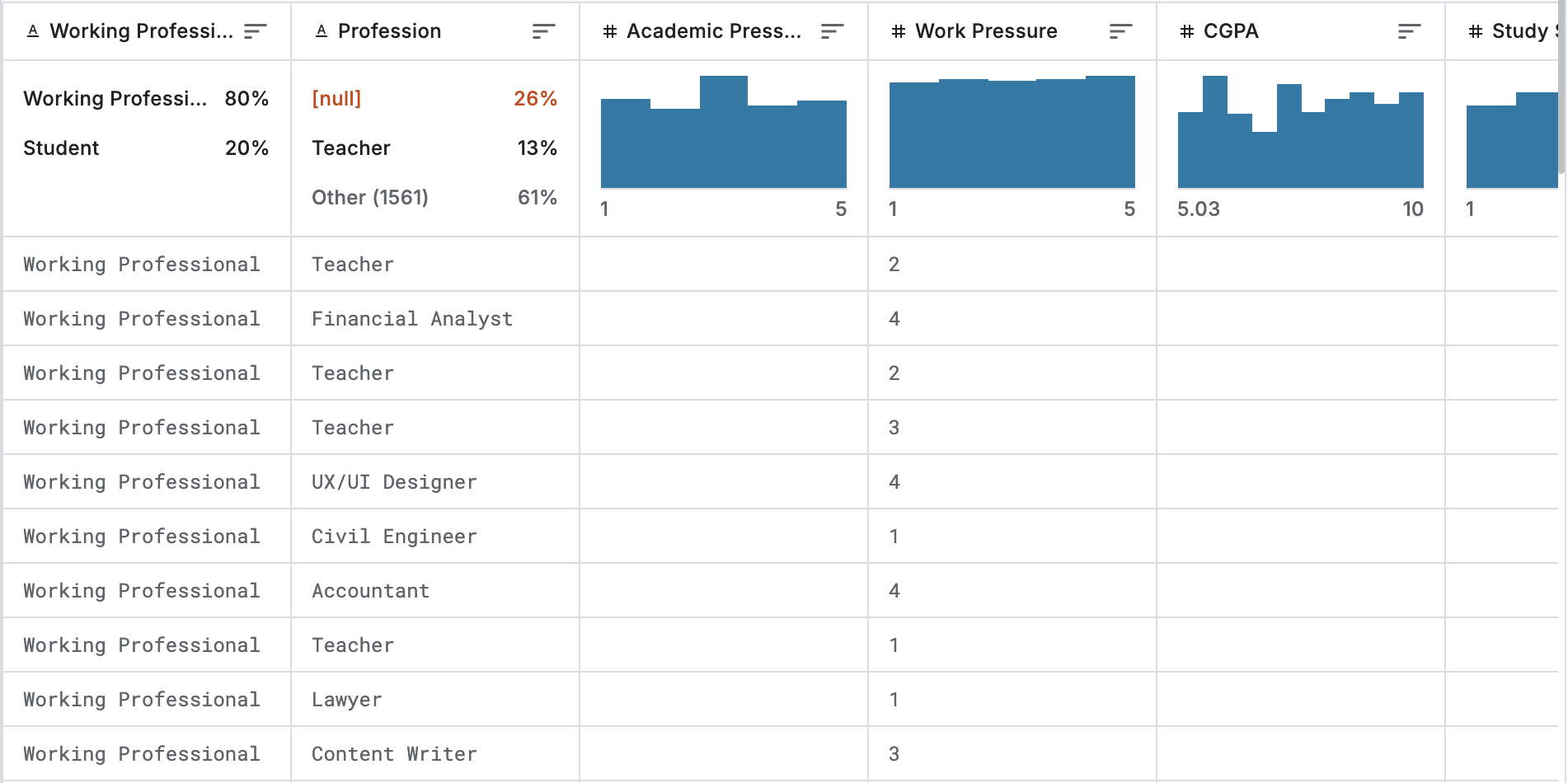
En el conjunto de datos, algunas columnas tienen un porcentaje significativo de valores nulos, particularmente en las variables *Academic Pressure*, *CGPA* y *Study Satisfaction*. Esta situación se explica porque el conjunto de datos incluye tanto a estudiantes como a profesionales (variable *Working Professional or Student*), y estas variables son relevantes solo para los estudiantes.

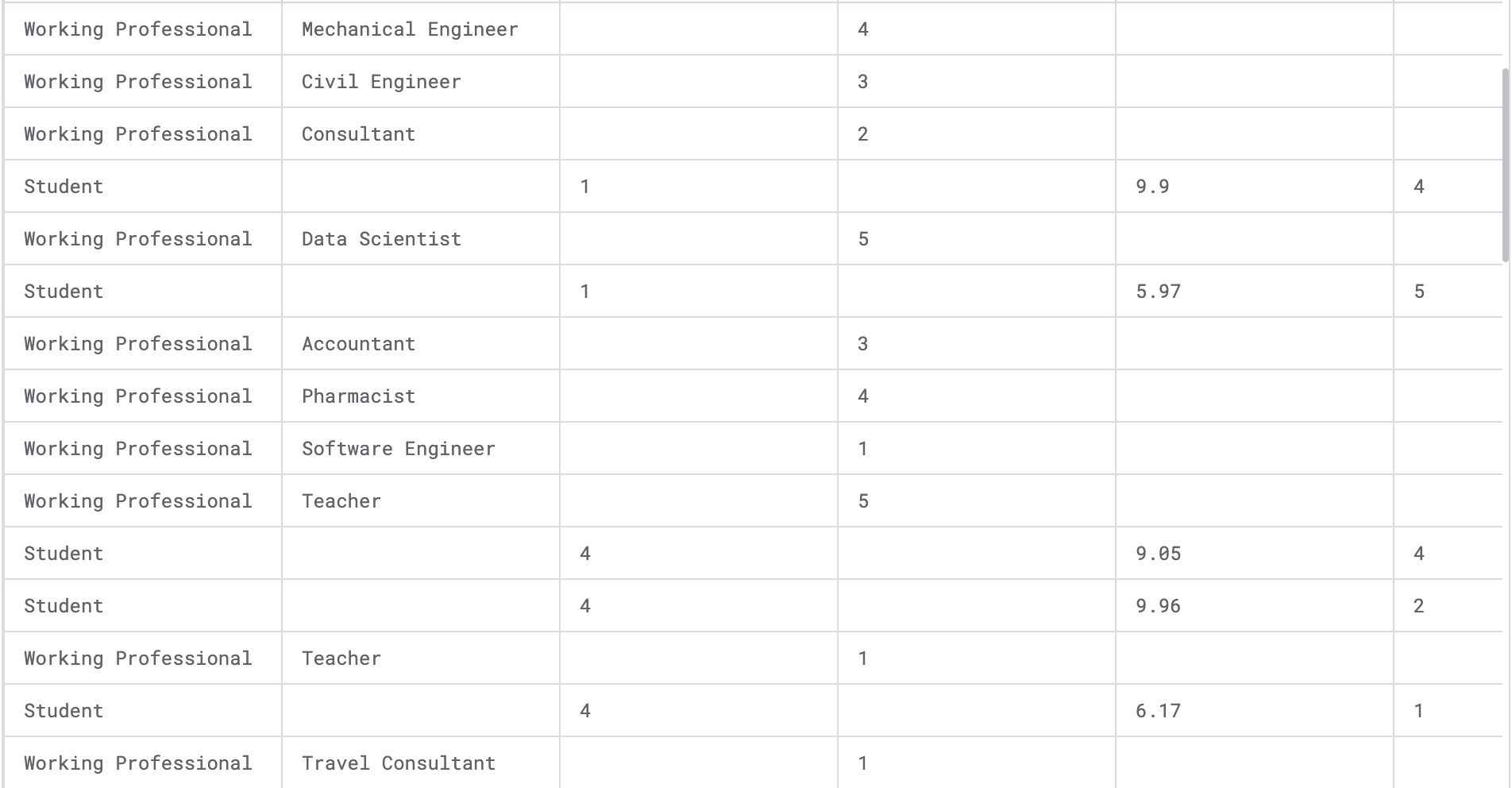
**Academic Pressure:** Mide la presión académica percibida, pero únicamente aplica a los estudiantes. Como los profesionales no están sujetos a presión académica en este contexto, esta columna tiene valores nulos para ellos.

**CGPA:** Representa el promedio acumulado, una medida que se utiliza para estudiantes en contexto académico.

**Study Satisfaction:** Mide el nivel de satisfacción con los estudios, relevante solo para estudiantes.

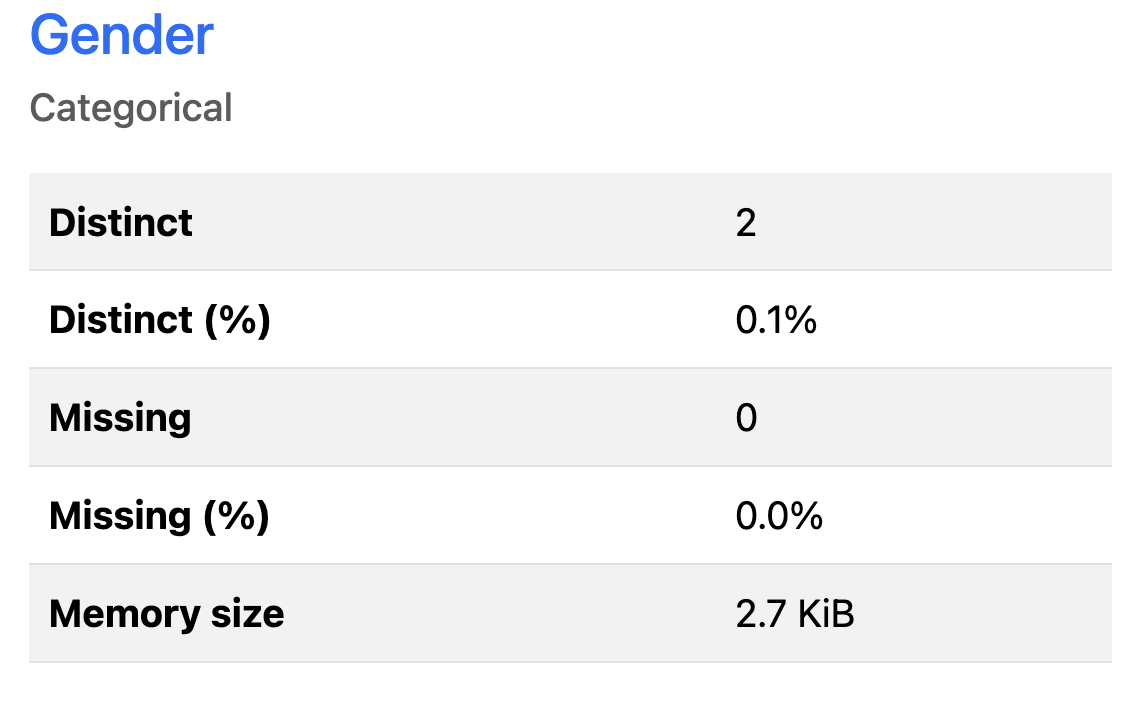
Esto se puede demostrar viendo los datos





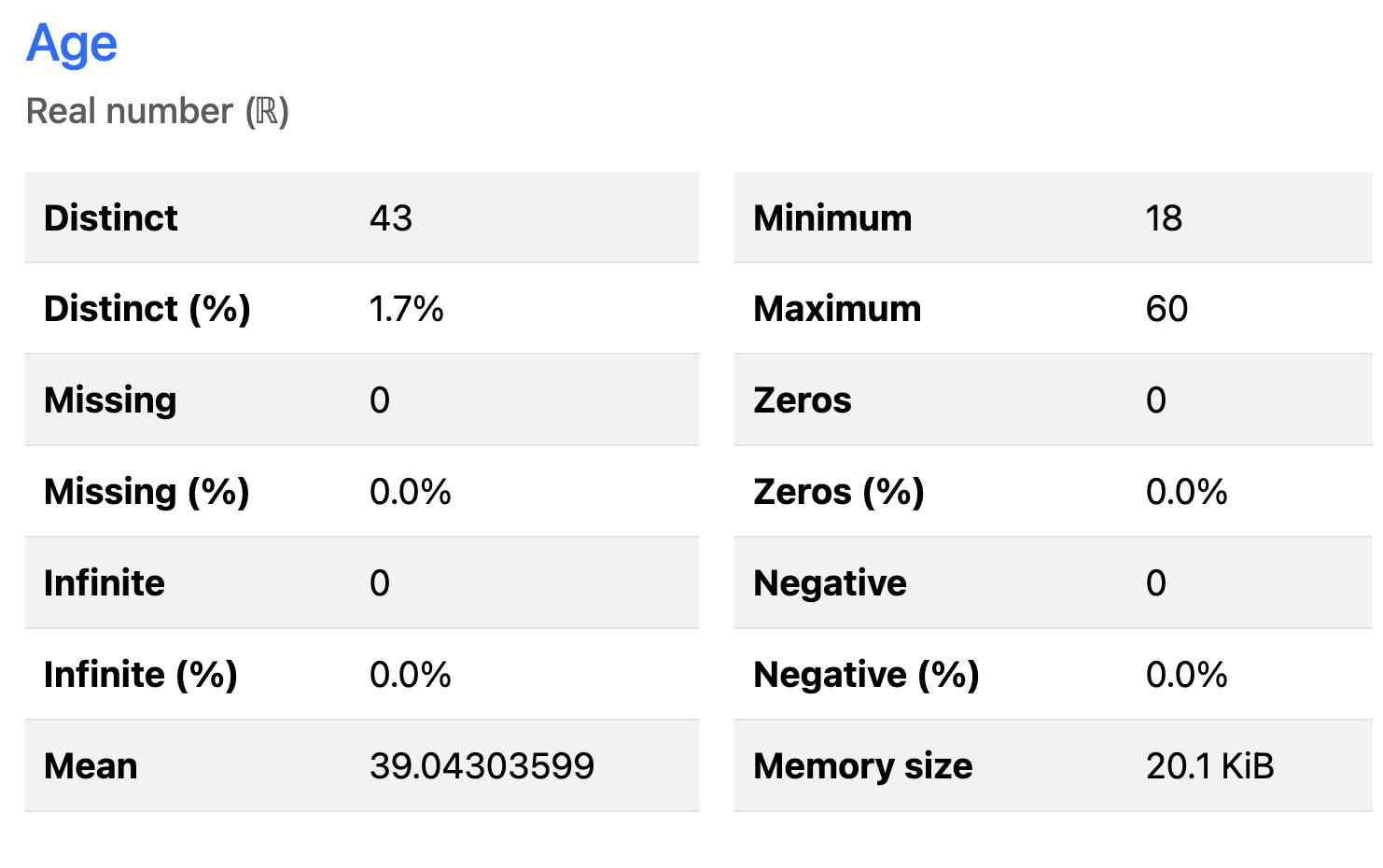
Al observar los datos, se confirma que los registros asociados a ***Working Professional*** tienen valores nulos en estas columnas, mientras que los registros categorizados como ***Student*** contienen valores válidos. Este comportamiento refleja que el conjunto de datos fue diseñado para capturar información específica según el rol del participante (estudiante o profesional), lo cual es la causa de los valores faltantes en ciertas columnas.

**Estudio de variables – Feature Engineering**

 A blue and orange rectangle with white text

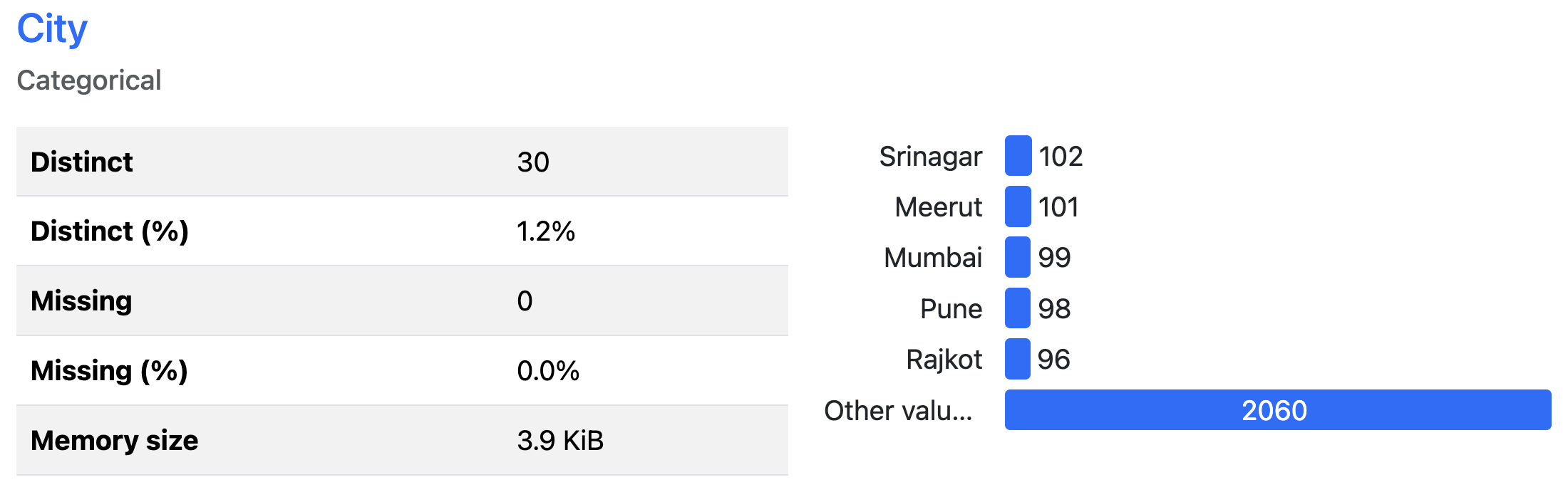
Description automatically generated

Contiene 2 categorías distintas, no tiene nulos, está balanceada.

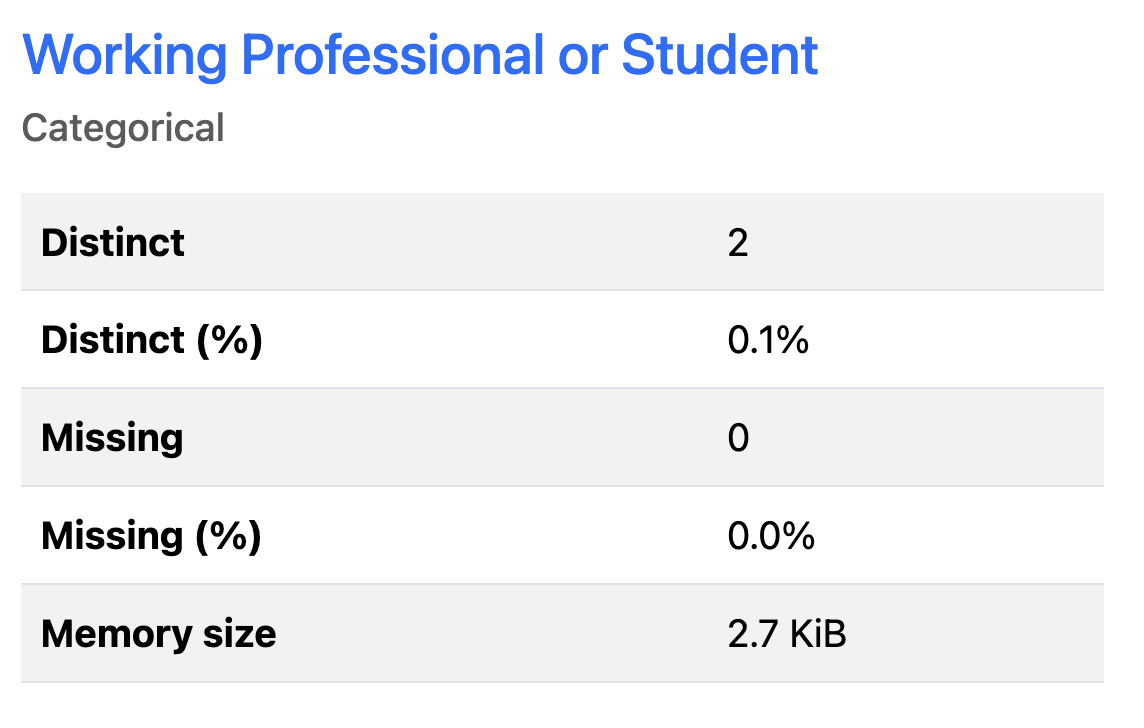
 A blue graph with white text

Description automatically generated

Rango de edad entre 18-60, no tiene nulos, media 39.04, se distribuye balanceadamente.

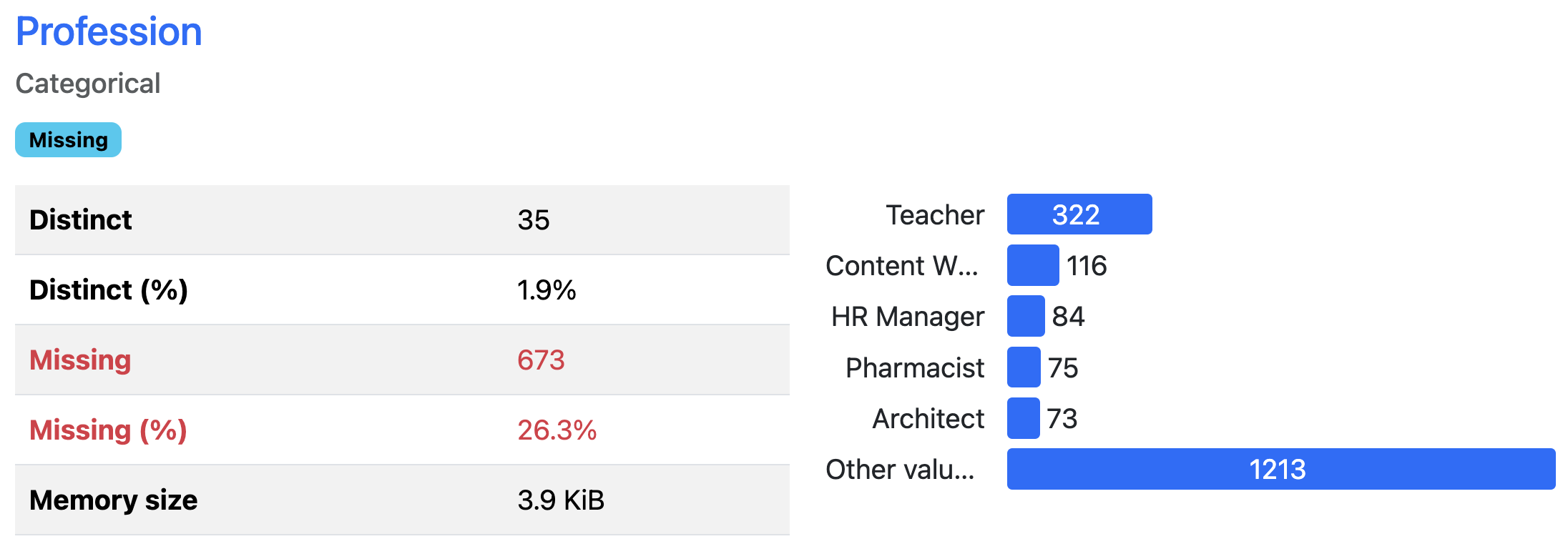


30 cuidades distintas, no contiene nulos.

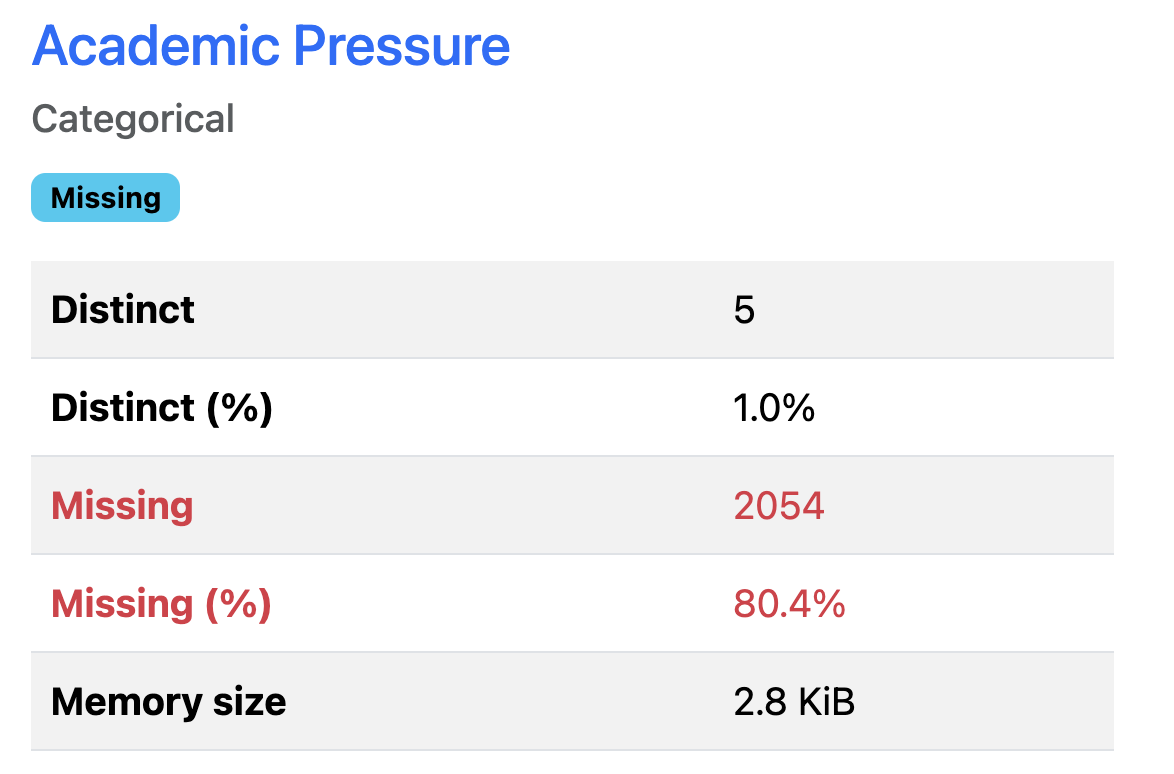
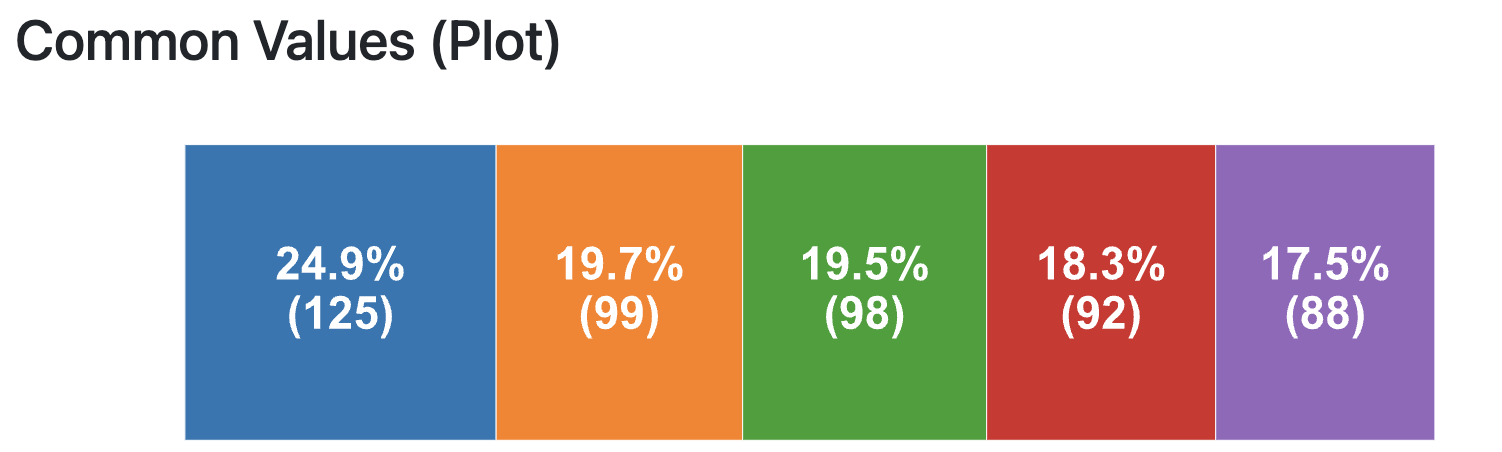
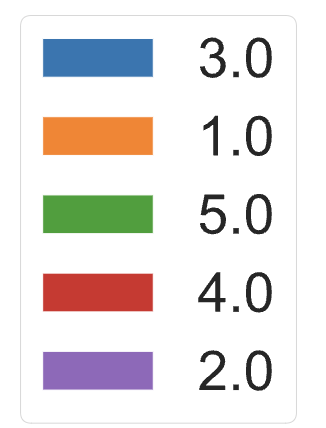
 A blue rectangle with white text

Description automatically generated

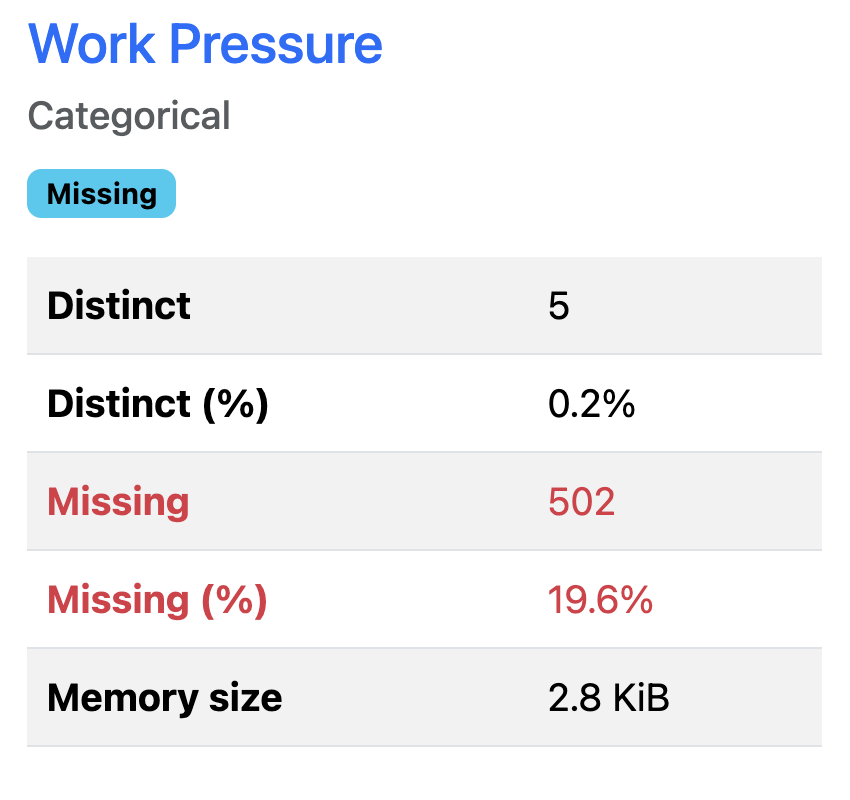
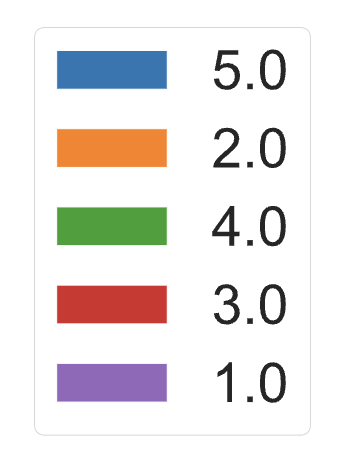
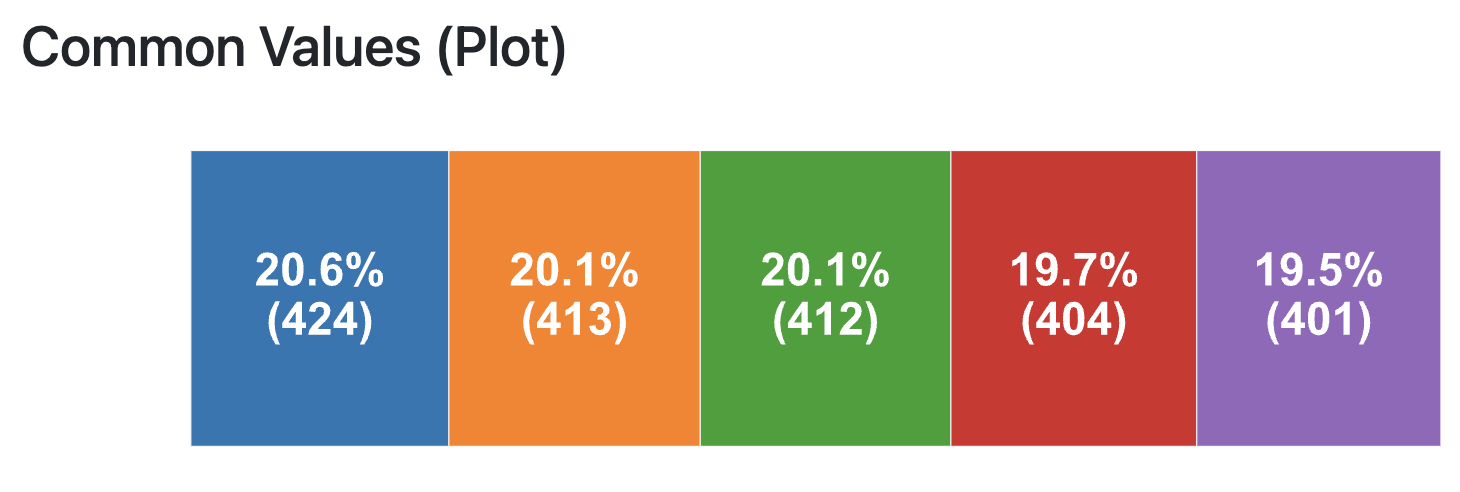
El 80.4% de los encuestados son profesionales, el 19.6% son estudiantes, está desbalanceada y no tiene nulos.



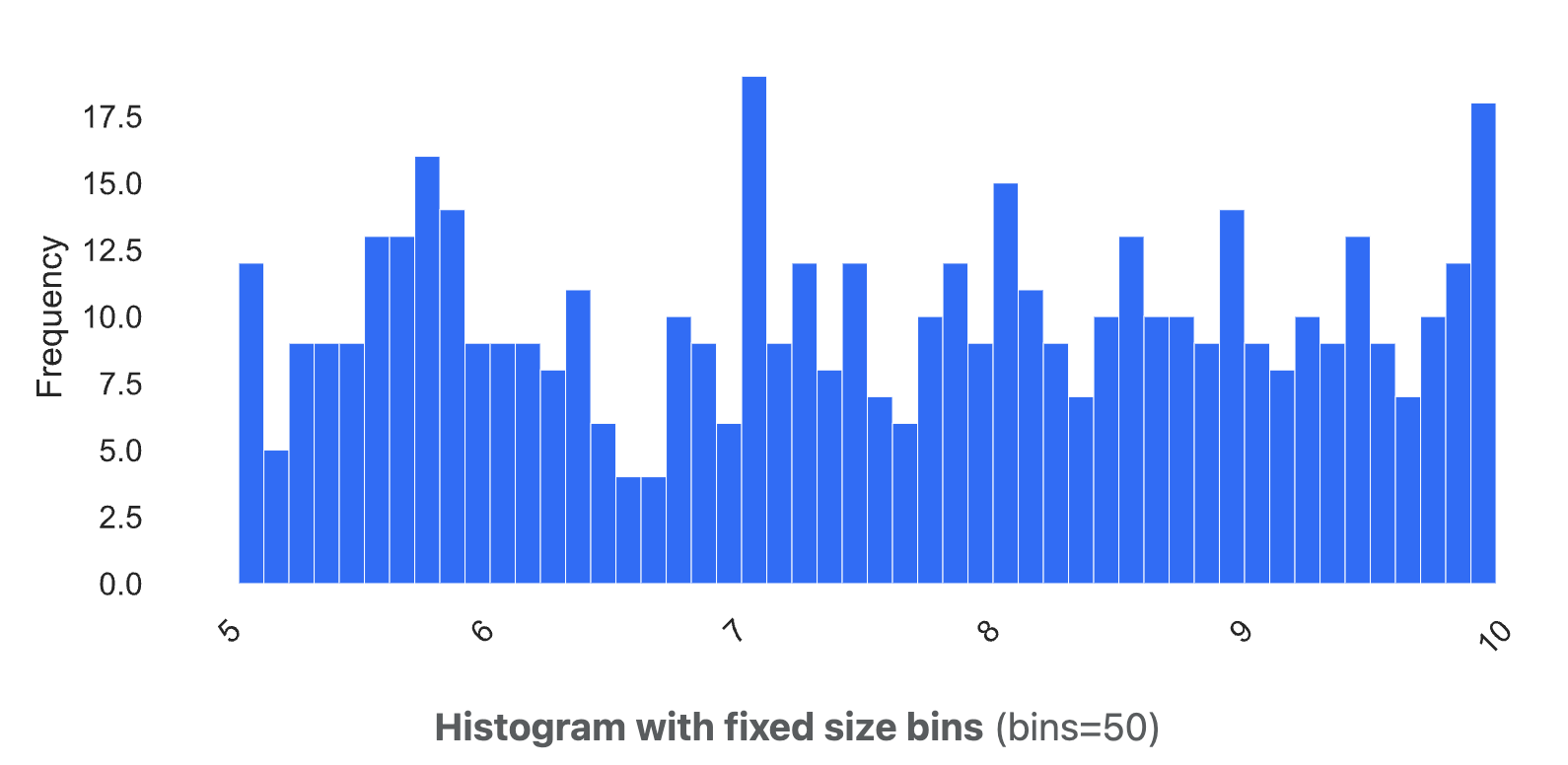
Se tienen 35 profesiones distintas, hay 673 (26.3%) datos nulos que corresponden a los registros de estudiantes que no tienen una profesión aun y a otros datos vacíos.

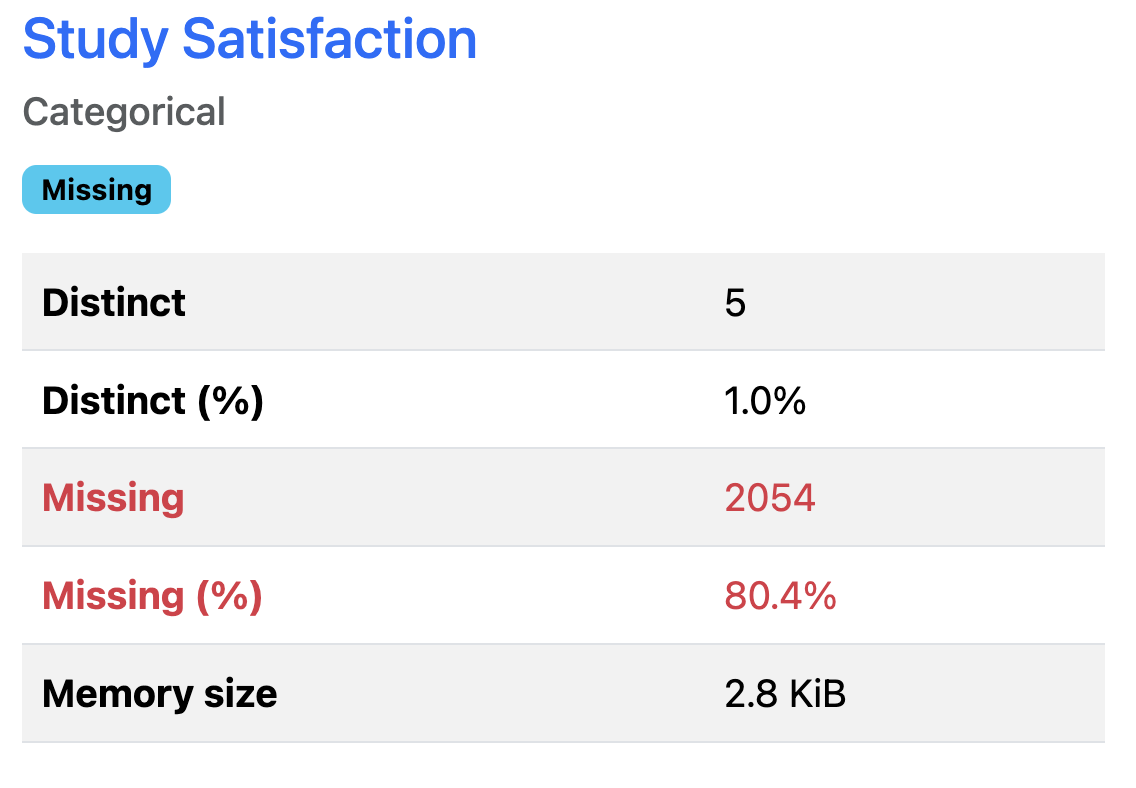
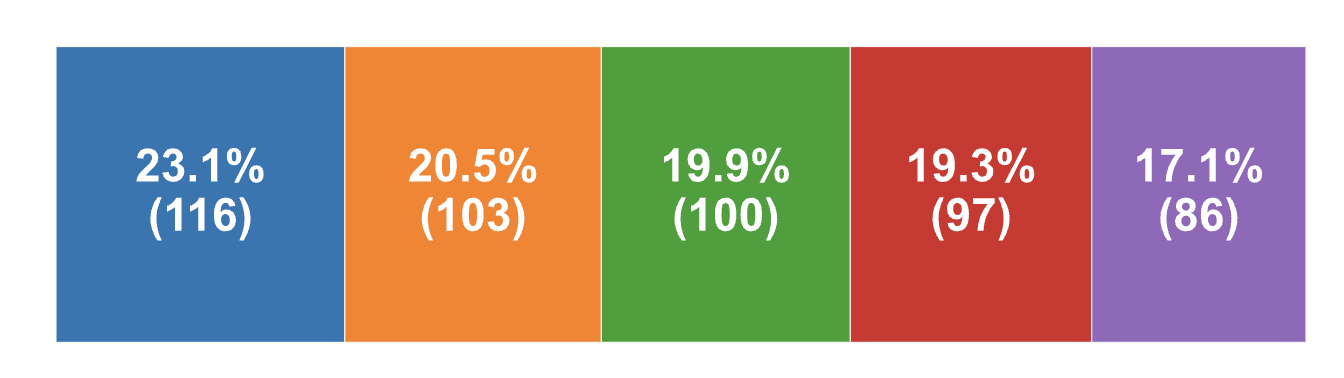
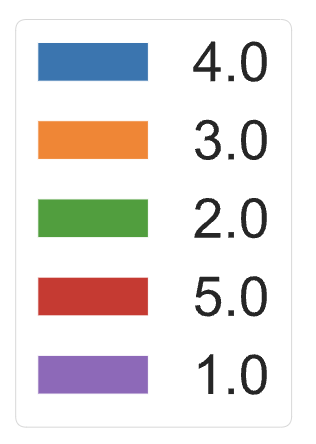
Hay 2054 valores nulos, como se mencionó anteriormente, la mayoría de encuestados son profesionales, razón por la cual esta columna tiene un número significativo de datos nulos.

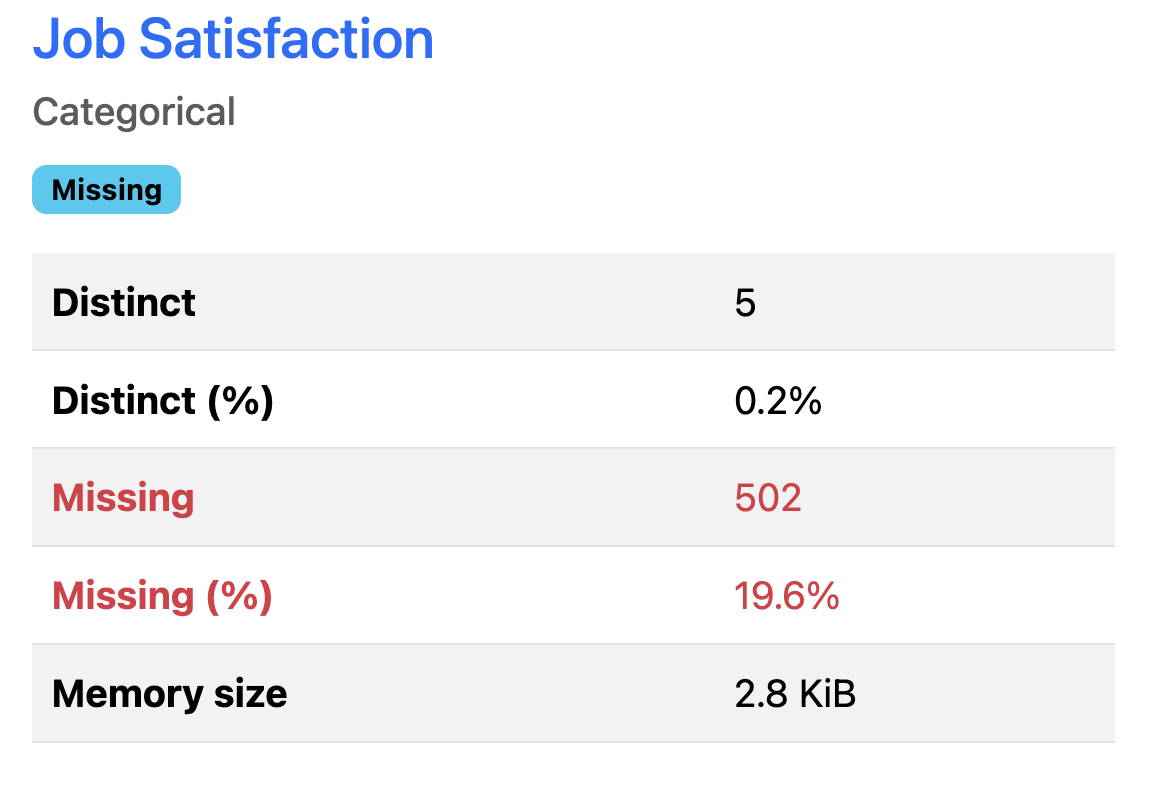
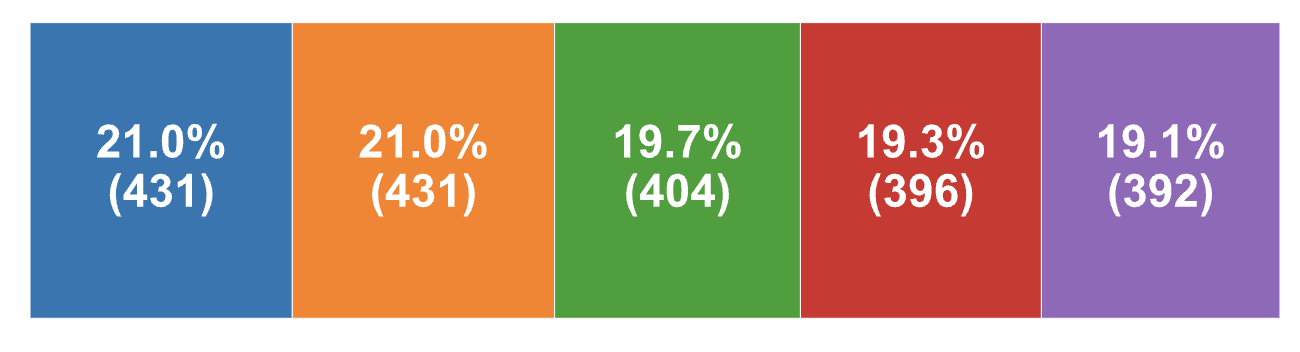
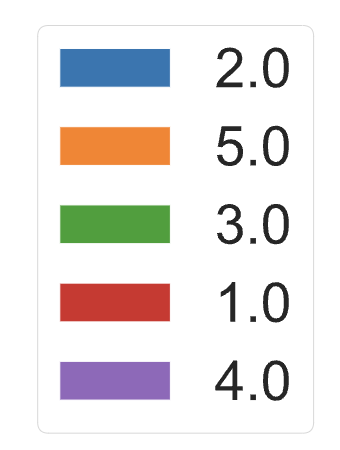
Hay 502 valores nulos que corresponden a los registros de los estudiantes que no presentan presión laboral sino presión académica (variable anterior). Más delante se deciden unificar estas 2 columas en una sola llamada *Work/Academic Pressure.*

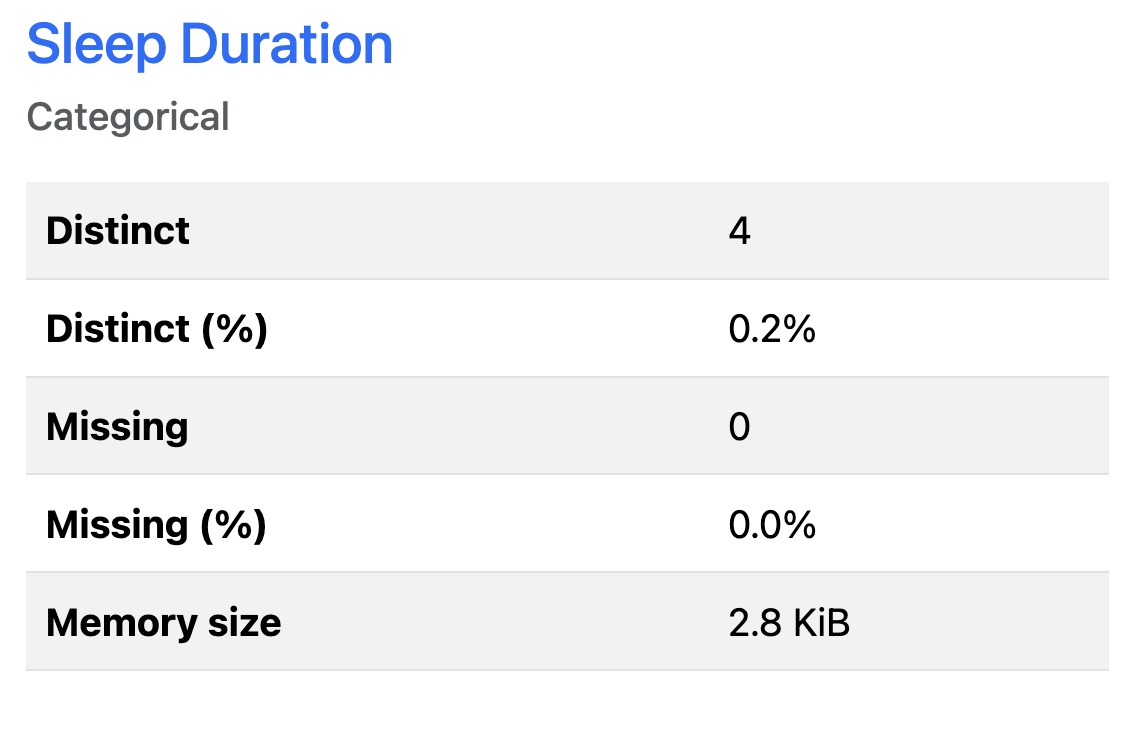
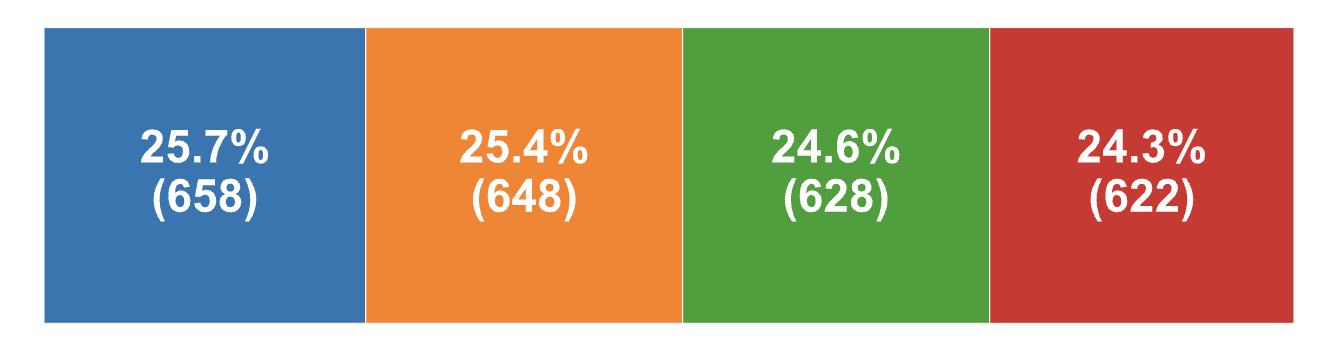
El CGPA es el promedio acumulado de los estudiantes. Se tiene la misma situación, esta variable solo puede ser calculada para estudiantes, por ello tiene tantos nulos.

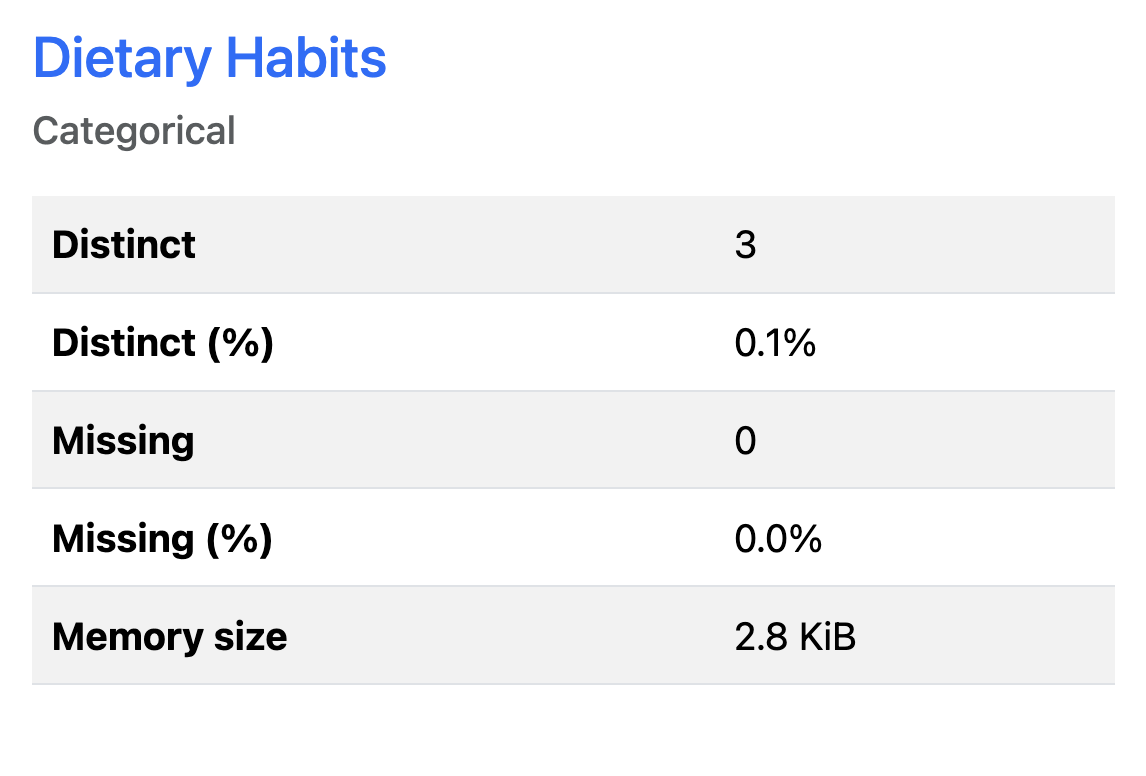
Para la satisfacción con el estudio sucede lo mismo, solo se llena con datos de estudiantes. Por esto tiene tanto nulos.

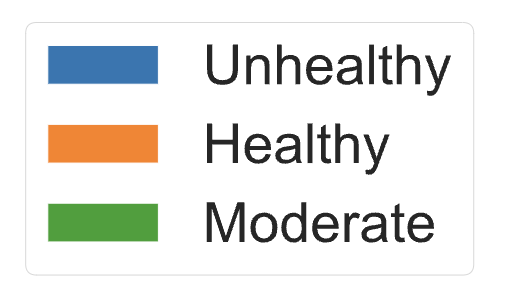
  

Esta columna solo tiene los valores de satisfacción del trabajo (solo profesionales). Tiene 502 valores nulos que corresponden a los registros de los estudiantes. Luego se unifican también estas 2 columnas.

Esta variable no tiene nulos y está distribuida balanceadamente.

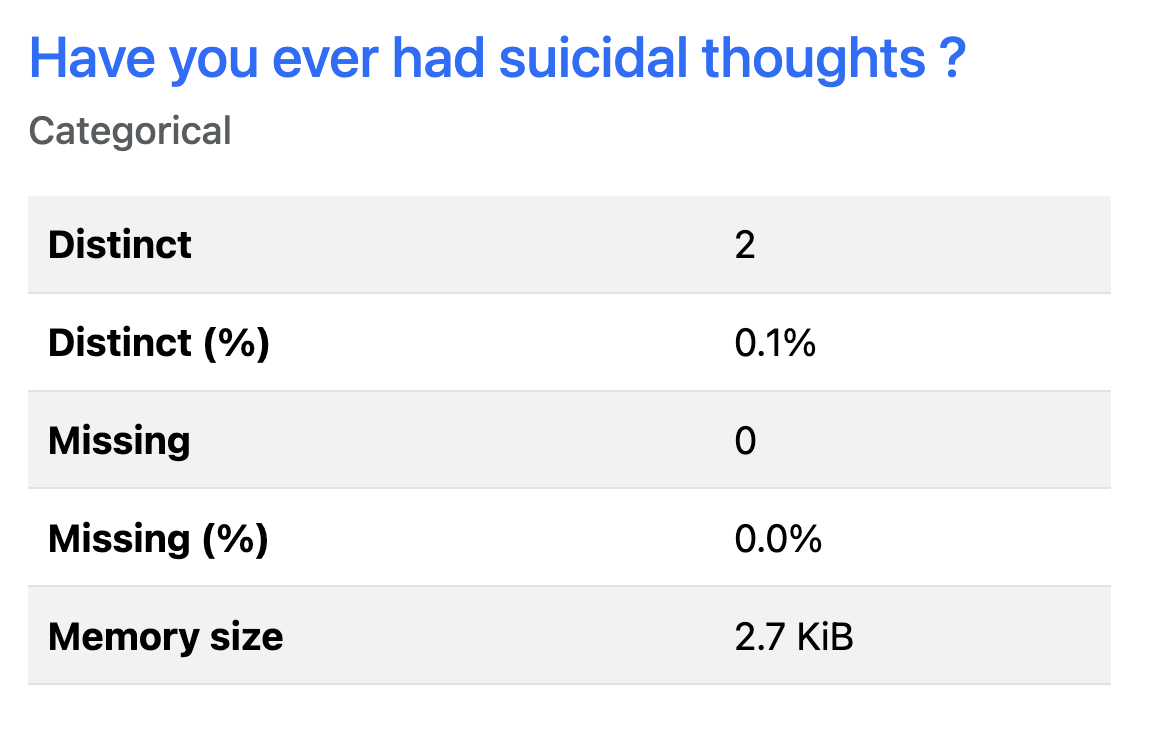
 A orange and blue rectangle with white text

Description automatically generated 

La variable Dietary Habits está balanceda, posee 3 categorías y no tiene nulos.

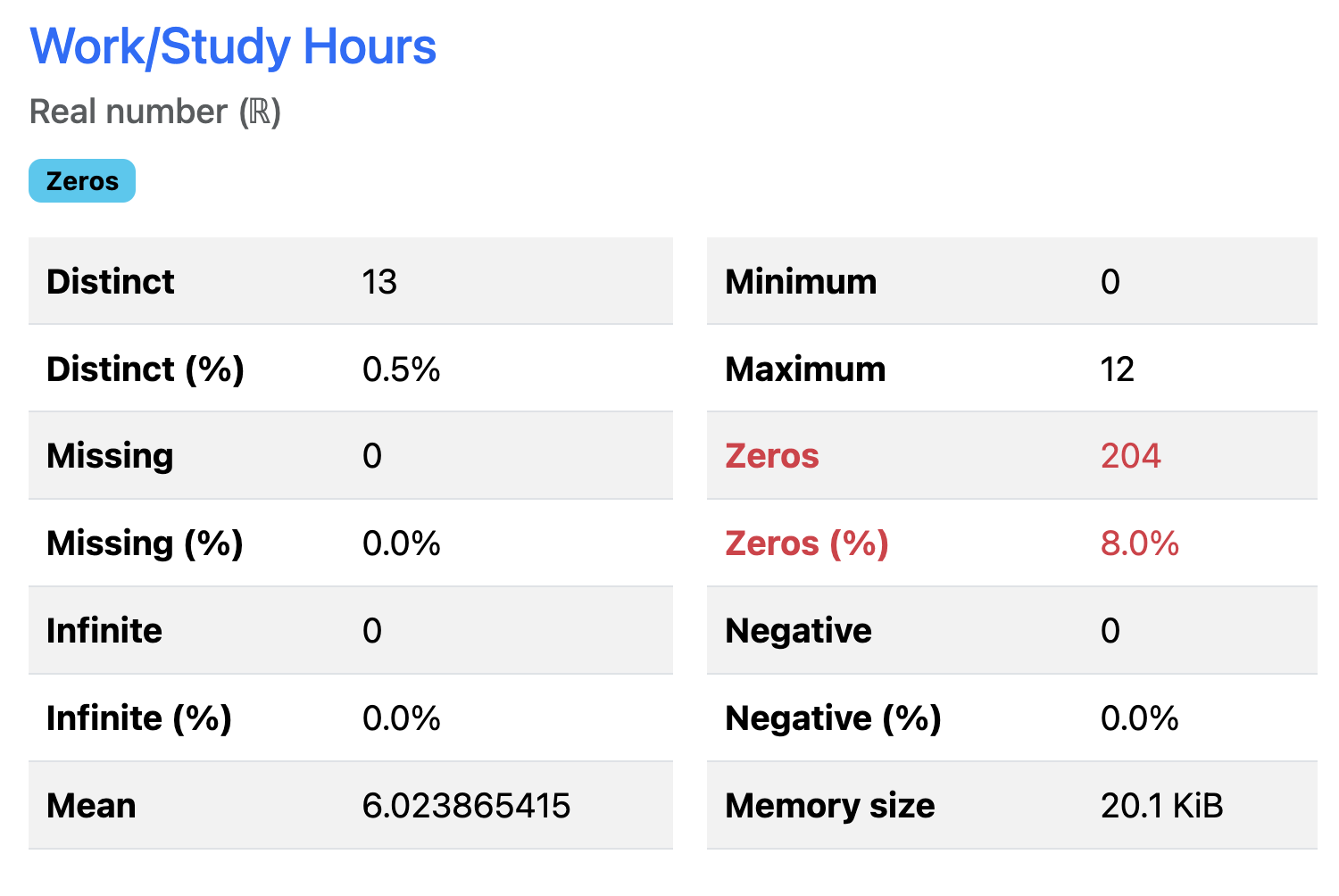
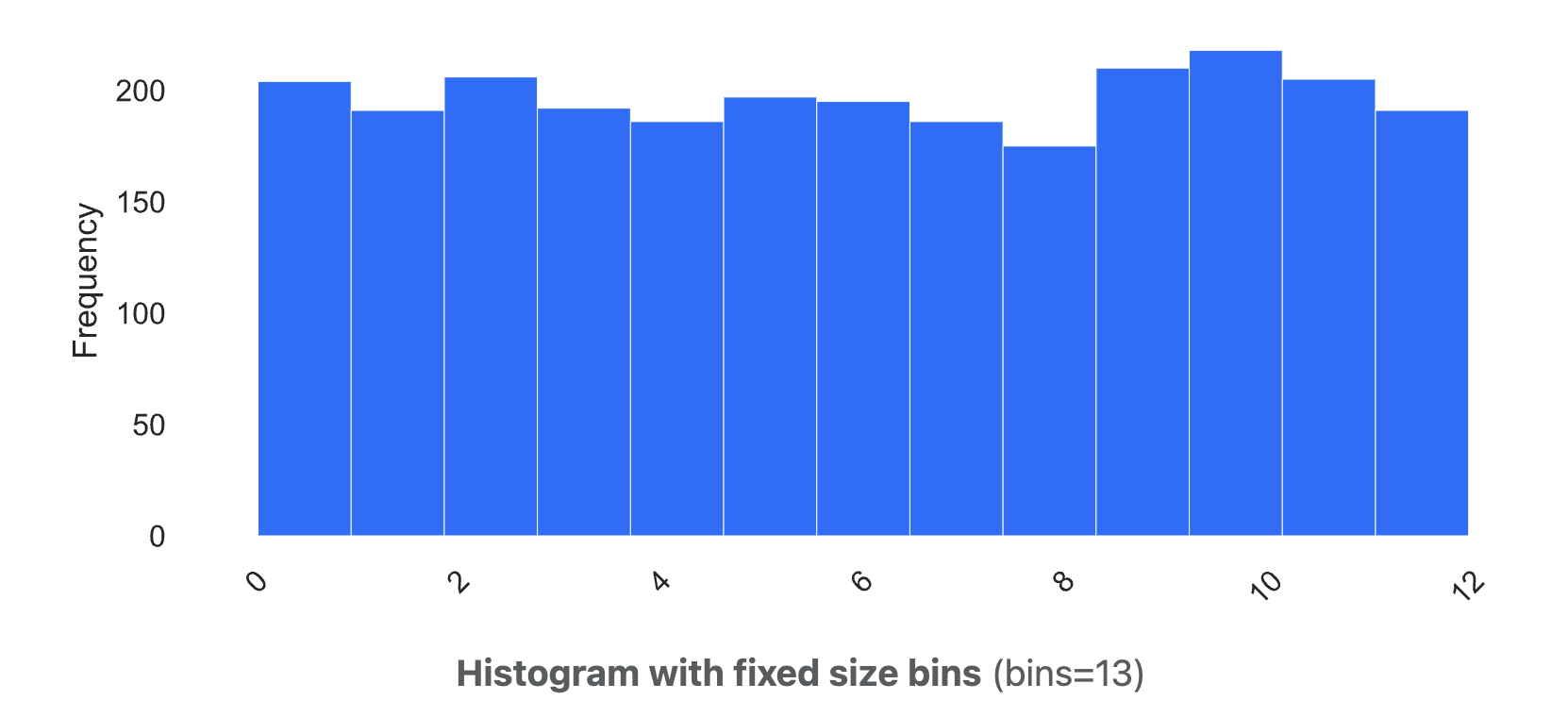


Degree es el nivel educativo del encuestado, esta variable no tiene nulos y tiene 27 categorías.

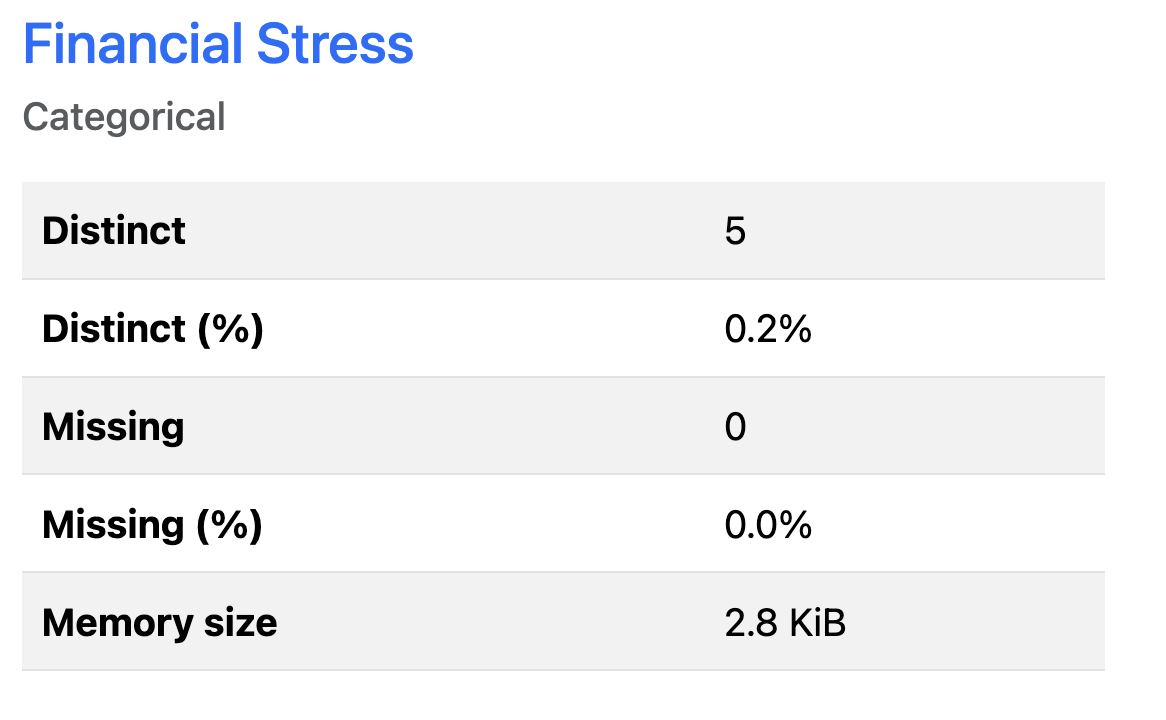
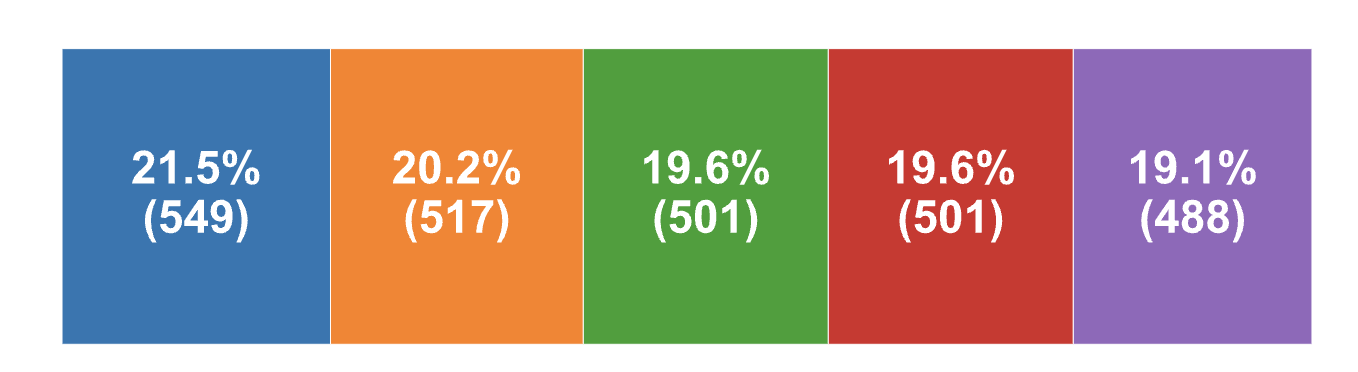
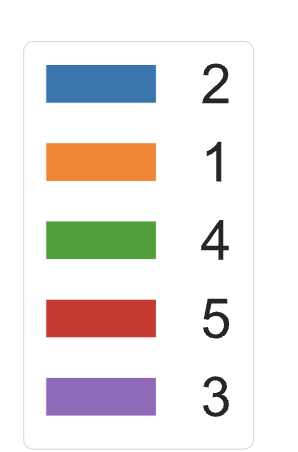
 A blue and orange rectangle with white text

Description automatically generated

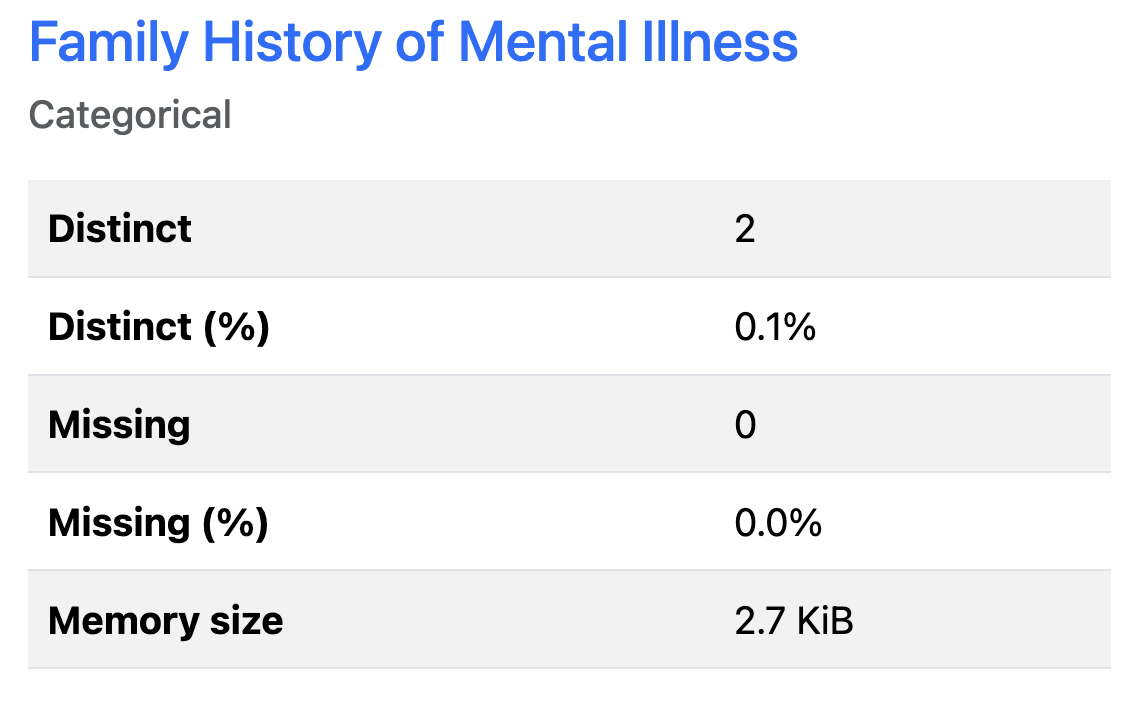
La variable “Has tenido pensamientos suicidas?” no contiene nulos, está distribuida balanceadamente.

Esta columna numérica no tiene valores nulos, el rango de horas de sueño es de 0-12. La media de horas de sueño es de 6.02 horas. Se identifica que hay 204 valores que son 0, esto significa que hay registros de participantes que trabajan o estudian 0 horas, sin embargo, esto puede ser posible.

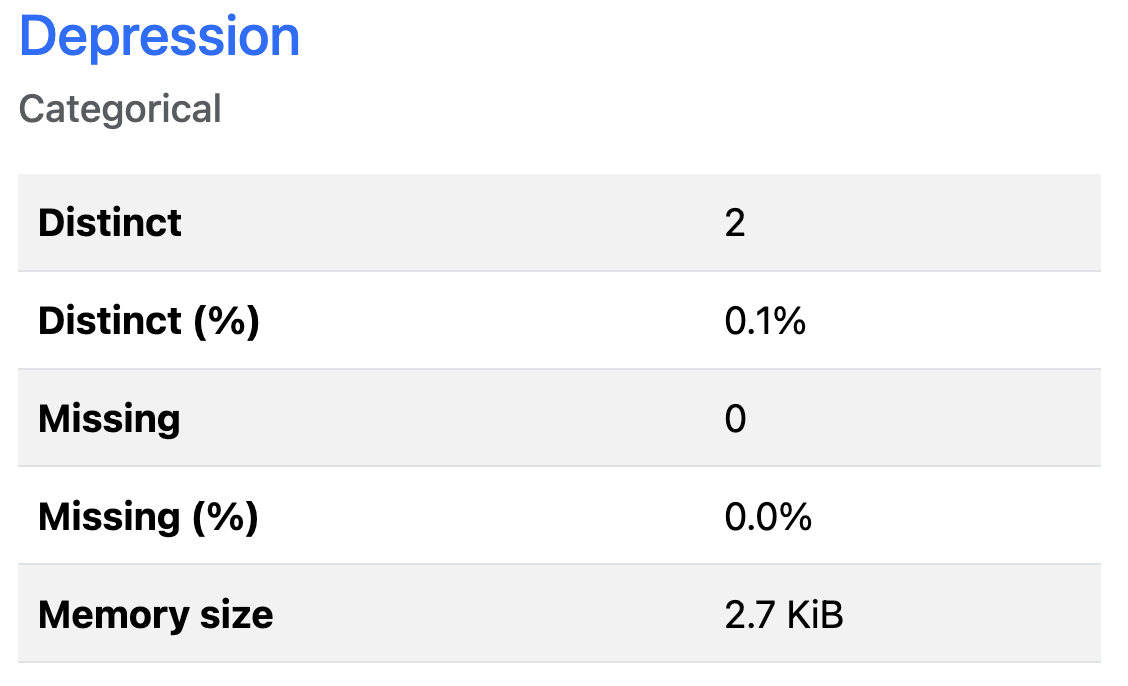
  

La variable estrés financiero está posee una escala de 1-5, esta balanceada y no tiene nulos.

 A blue and orange rectangle with white text

Description automatically generated

La variable indica si se tienen antecedentes familiares con enfermedades mentales. No tiene nulos, está balanceada.

 A blue rectangular box with white text

Description automatically generated

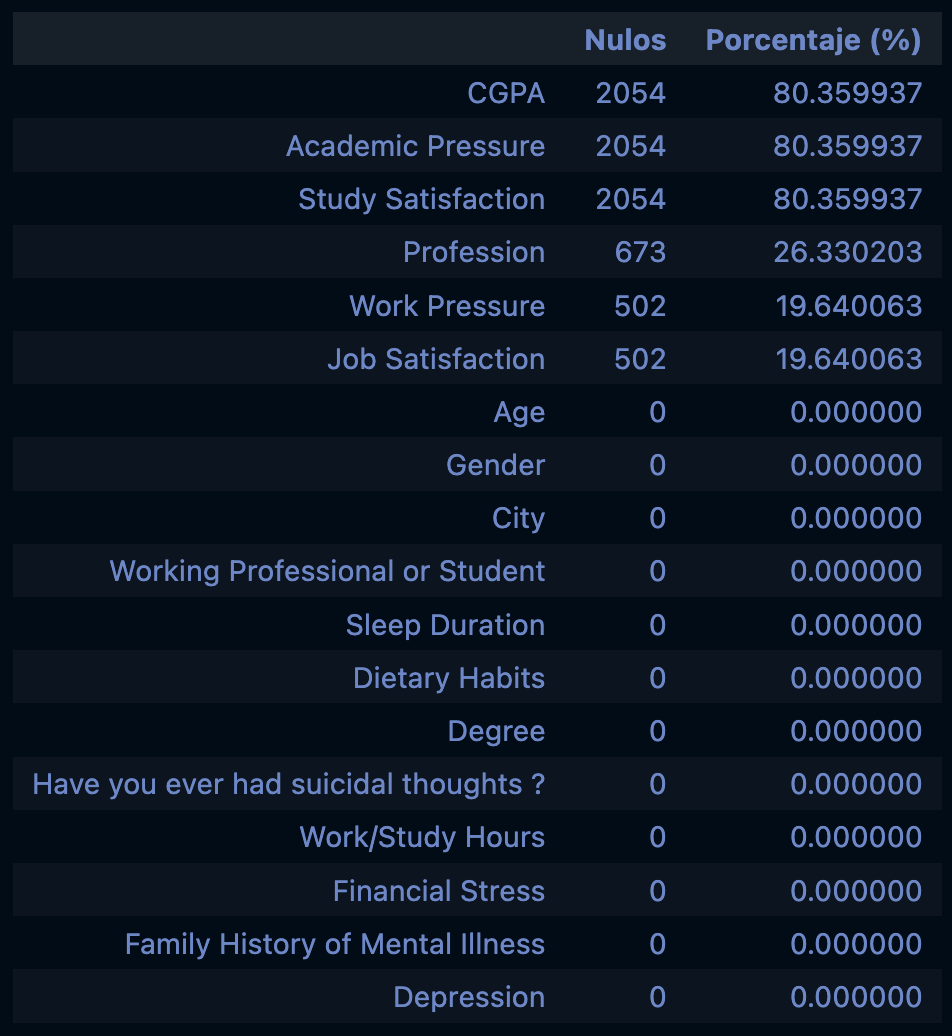
Depresión es la variable objetivo que indica si hay riesgo o no de que un participante sufra depresión. No tiene nulos pero está desbalanceada.

* 1. **LIMPIEZA DE ATÍPICOS**

El dataset no contiene datos atípicos.

* 1. **LIMPIEZA DE NULOS**

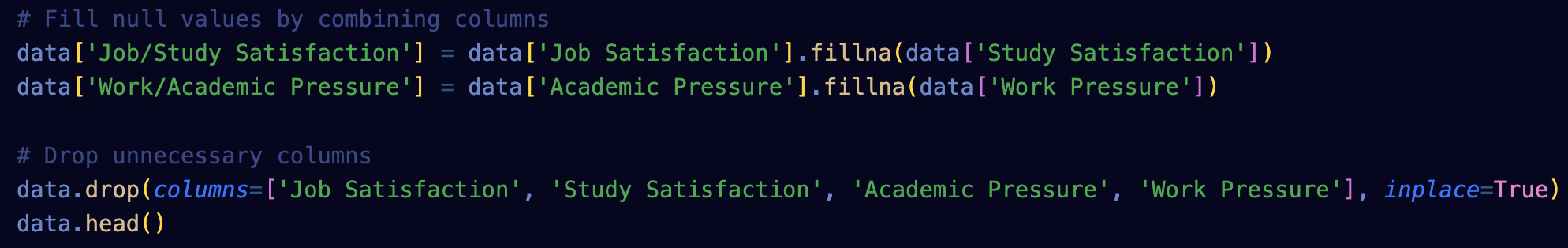
En el análisis y descripción estadística de los datos, se identificó un número significativo de valores nulos en varias columnas, especialmente en aquellas que corresponden exclusivamente a estudiantes.



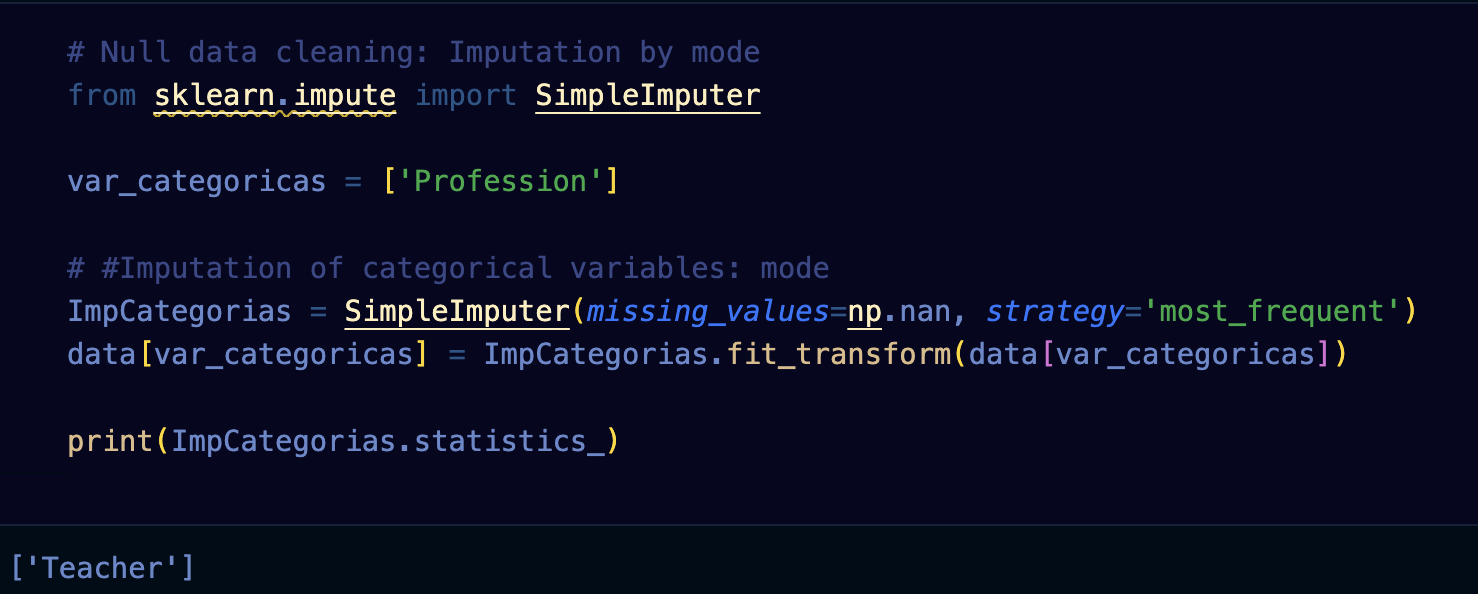
Para abordar este problema, se tomaron las siguientes medidas:

* + - Se decidió eliminar la columna *CGPA* (promedio acumulado de estudiantes) debido a que presenta un porcentaje de valores nulos superior al 30%, lo cual no permite su imputación sin comprometer la calidad de los datos
    - Se creó una nueva columna llamada *Academic/Work Pressure*, que combina los datos de *Academic Pressure* (para estudiantes) y *Work Pressure* (para profesionales). De esta forma, se obtiene una columna unificada sin valores nulos. De manera similar, se creó la columna *Study/Job Satisfaction* unificando *Study Satisfaction* (para estudiantes) y *Job Satisfaction* (para profesionales), logrando así una variable completa y sin nulos.

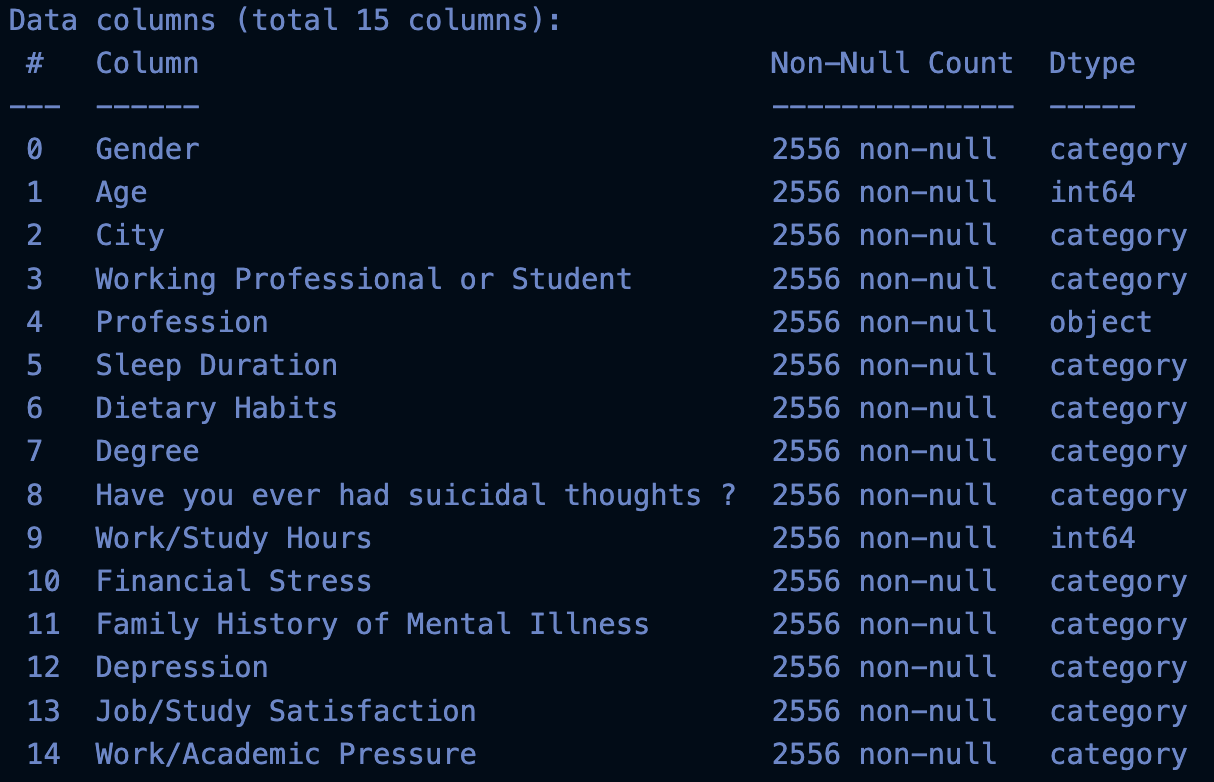
Estas combinaciones fueron posibles gracias a la relación observada entre estudiantes y profesionales durante el análisis de perfilado de datos, permitiendo crear columnas que engloban ambas perspectivas de presión y satisfacción.



* + - Por ultimo, para la columna Profession se rellenaron los valores nulos de la columna *Profession* con la moda, que es *Teacher*. Luego, para los registros donde se detectó que el participante es estudiante (según la columna *Working Professional or Student),* se cambió el valor de *Profession* a *Student*, asegurando así que cada participante tiene una profesión asignada.



Con estas transformaciones, se obtuvo un conjunto de datos más limpio, eliminando los valores nulos en las columnas.



Luego de la limpieza de nulos se obtuvieron **15** columnas sin nulos.

* 1. **CREACIÓN DE NUEVAS VARIABLES**

No se crearon nuevas variables para efectos de este ejercicio de predicción.

* 1. **ANÁLISIS DE CORRELACIONES PARA REDUNDANCIA**

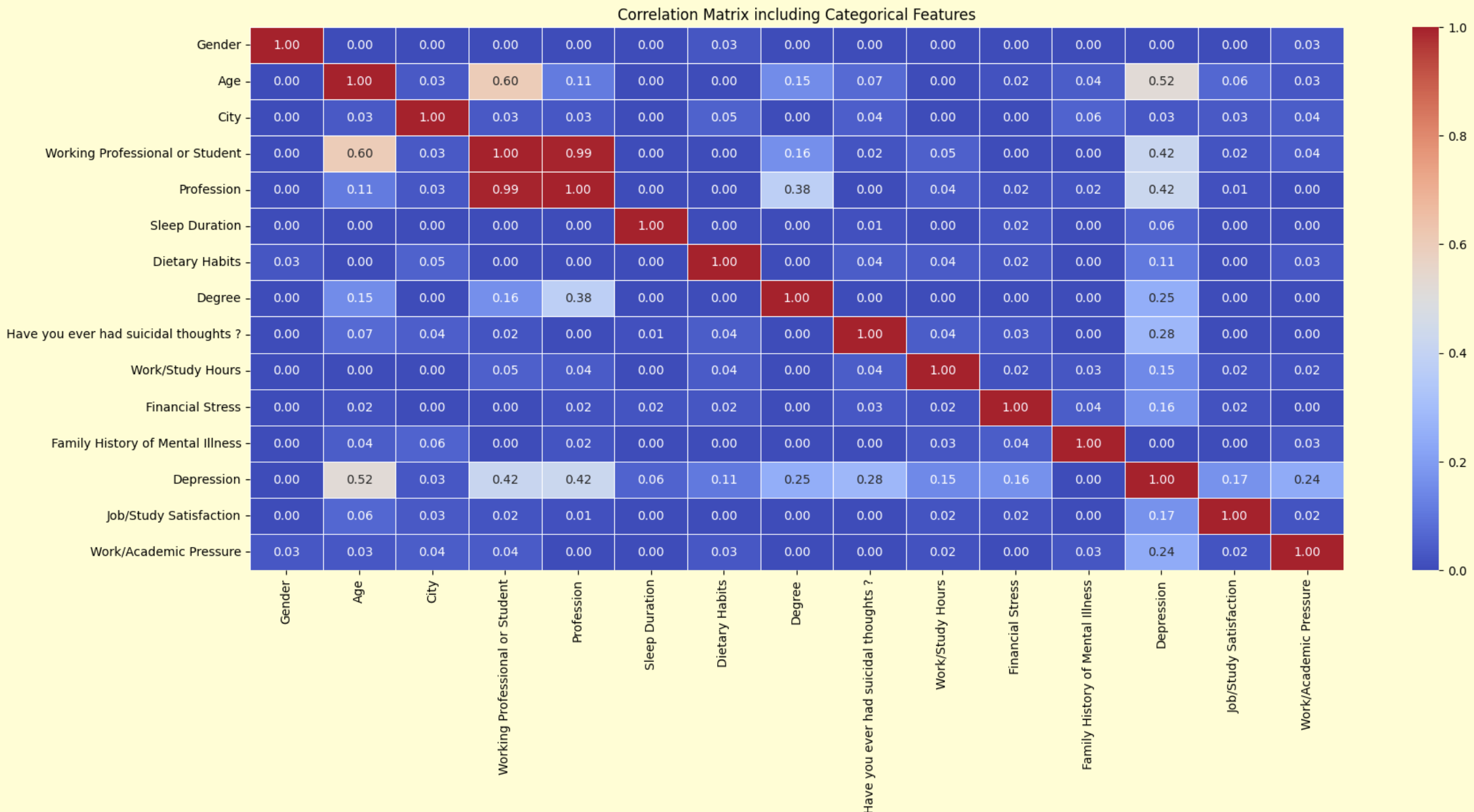
Se analizaron las correlaciones entre las variables para identificar aquellas que sean redundantes. Las variables con una alta correlación entre sí (superiores a 0.8) indicaron redundancia. En tales casos, se eliminaron estas variables.

Estos son los pasos que se siguieron para identificar correlaciones:

Se utilizaron dummies para transformar las variables categóricas en variables numéricas. Luego, la variable objetivo Depression fue transformada en valores numéricos mediante LabelEncoder, asignando 0 para "No" y 1 para "Yes".

**Matriz de Correlación**

Se calculó una matriz de correlación que incluye tanto variables numéricas como categóricas utilizando el paquete dython, lo cual permite analizar asociaciones entre variables categóricas y numéricas y se generó un heatmap para visualizar las correlaciones, identificando correlaciones altas (0.8 - 1.0).



Se observó una correlación alta entre las columnas *Working Professional or Student* y *Profession,* lo que indica que ambas variables representan información similar.

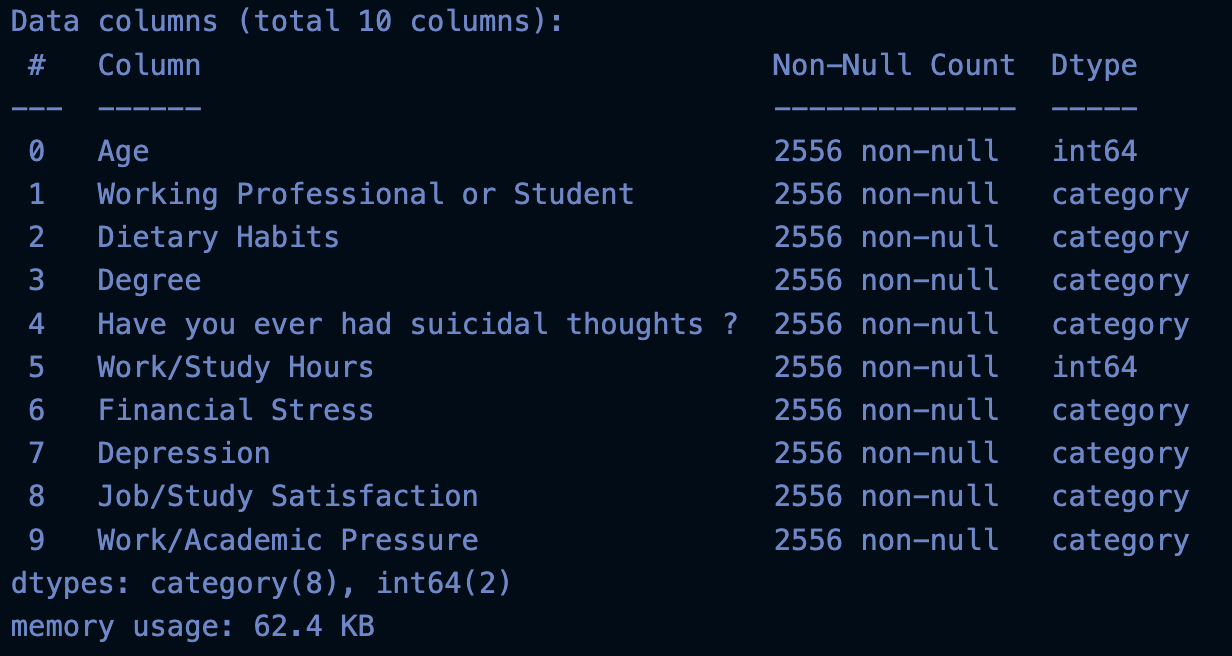
Para reducir la redundancia, se decidió eliminar la columna *Profession*, conservando *Working Professional or Student* como la variable representativa.

* 1. **ANÁLISIS DE CORRELACIONES PARA IRRELEVANCIA**

Se evaluó la relevancia de cada variable para la predicción del riesgo de depresión. Las variables que no tenían ninguna correlación significativa con la variable objetivo ("Depression") fueron consideradas irrelevantes y eliminadas, ya que no aportan al modelo predictivo.

Se identificaron variables con correlaciones muy bajas (en el rango de 0.0 a 0.1) con la variable objetivo Depression. Las variables *City*, *Family History**of Mental Illness, Sleep Duration* y *Gender* presentaron correlaciones muy bajas y, por lo tanto fueron eliminadas del conjunto de datos para simplificar el modelo.

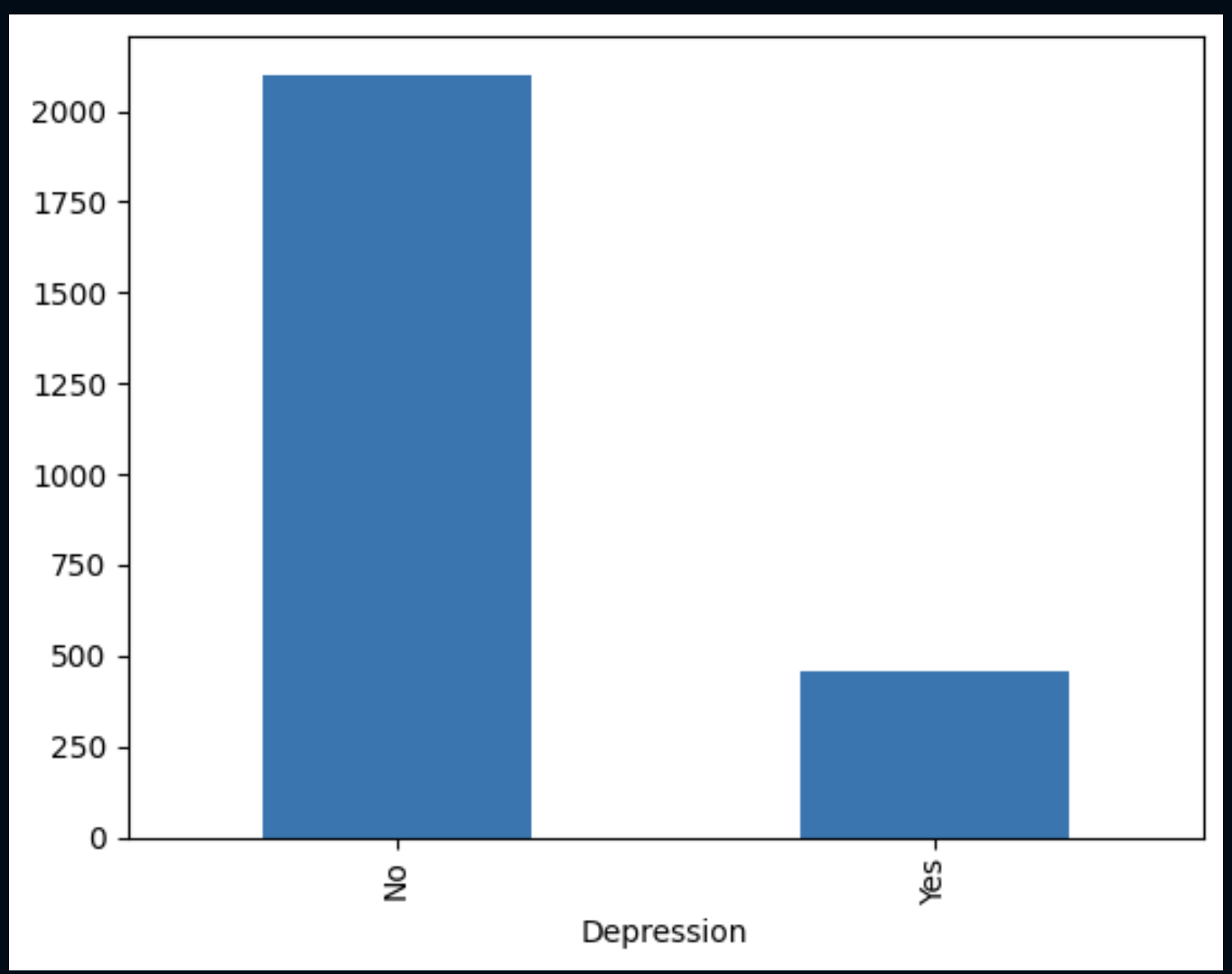
Con estos pasos, se logró reducir la dimensionalidad del conjunto de datos mediante la eliminación de variables redundantes e irrelevantes. Estas son las variables luego del proceso de análisis de correlaciones, quedan 10 variables.



* 1. **REDUCCIÓN DE DIMENSIÓN (OPCIONAL EN PREDICCIONES)**

No se realizó PCA para la reducción de dimensionalidad.

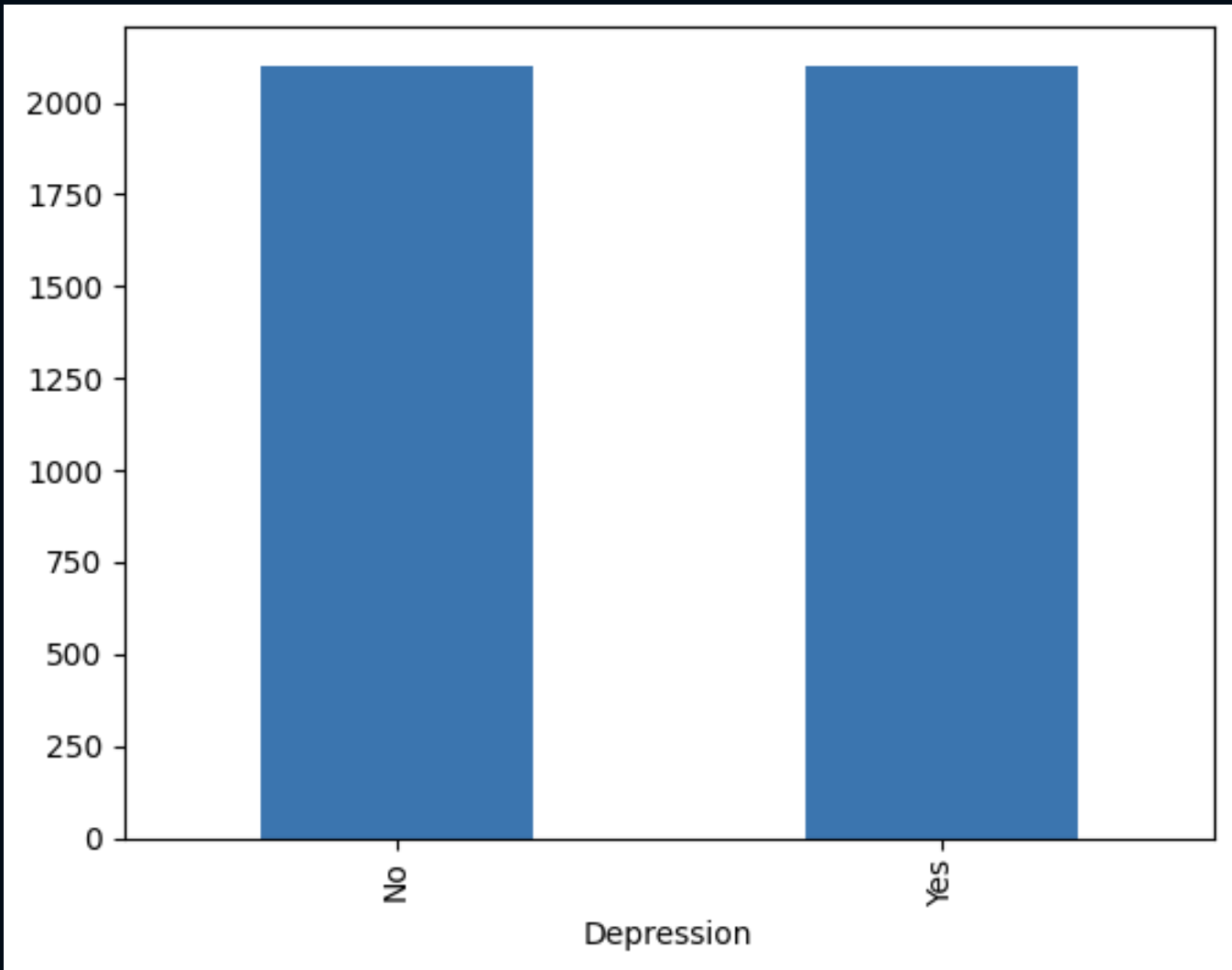
* 1. **BALANCEO**



Dado que la variable objetivo *Depression* está desbalanceada (la mayoría de los registros no presentan riesgo de depresión), es importante aplicar técnicas de balanceo como SMOTE. El objetivo es lograr que el modelo pueda aprender de ambas clases de manera equilibrada y mejorar la precisión de la clasificación para ambas categorías ("Yes" y "No").

Se aplicó el método SMOTE para balancear la variable objetivo, dado que el conjunto de datos incluye tanto variables categóricas como numéricas, se utilizó SMOTENC para el balanceo. Se generaron muestras sintéticas de la clase minoritaria (Yes), logrando un conjunto de datos balanceado.

Tras aplicar SMOTENC, el número total de registros aumentó de 2556 a 4202, logrando una distribución equitativa entre las clases de la variable Depression.



* 1. **TRANSFORMACIONES**

Para adaptarse a los diferentes requerimientos de los modelos de Machine Learning, se generaron dos versiones del conjunto de datos: uno categórico y otro numérico. Esto permite que cada modelo trabaje con el tipo de datos que necesita:

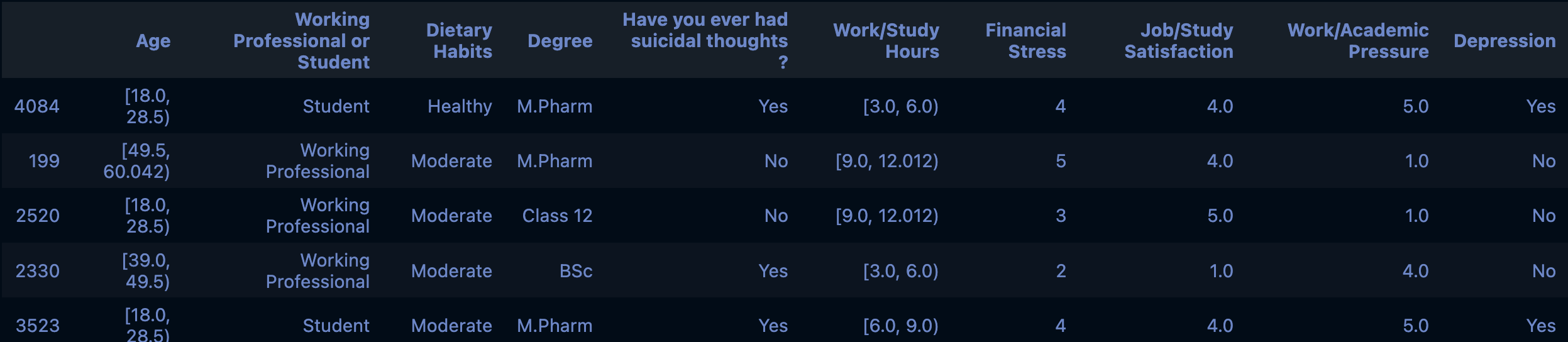
**Dataset categórico:** Diseñado para métodos de ML que requieren variables discretizadas o categóricas (árboles de decisión, Naive Bayes).

* + - **Discretización:** Variables numéricas como *Age* y *Work/Study Hours* se agruparon en intervalos (bins) para modelos que funcionan mejor con datos categóricos, como los árboles de decisión.

A graph of a bar chart

Description automatically generated with medium confidence

Asi se ve el conjunto de datos categórico.



**Conjunto numérico:** Preparado para modelos de ML que requieren datos numéricos y normalizados (Redes neuronales, SVM, regresión logística, KNN).

* + - **Normalización:** Se aplicó la normalización Min-Max a variables numéricas para modelos que requieren datos continuos en un rango de 0 a 1, como redes neuronales y SVM.
    - **Dummies:** Las variables categóricas fueron convertidas en variables numéricas para que todos los modelos pudieran interpretarlas correctamente.
    - **Codificación de la Variable Objetivo:** La variable Depression fue codificada en 0 y 1 con LabelEncoder.



Así se ve el conjunto de datos numérico.

Con estas transformaciones, cada conjunto de datos está listo para ajustarse a los requerimientos específicos de cada modelo de ML, optimizando el rendimiento y la precisión.

1. **MODELAMIENTO, EVALUACIÓN E INTERPRETACIÓN**
   1. CONFIGURACIÓN MÉTODOS DE MACHINE LEARNING

Se aplican los métodos que se especificaron en el diseño de solución, y se configuran.

* 1. ANALISIS DE MEDIDAS DE CALIDAD

Análisis de las medidas de calidad definidas previamente.

* 1. SELECCIÓN DEL MEJOR MODELO

Además de la medida de calidad se deben tener en cuenta otros aspectos como la complejidad computacional. Se usan métodos como la ANOVA o la prueba de Tuckey. Si no hay diferencia significativa en los modelos se parte de la comparación a partir de la complejidad computacional.

1. **DESPLIEGUE**
   1. PREDICCIÓN DE DATOS FUTUROS

Despliegue por interfaz gráfica, app, nube, web, etc.

* 1. CRONOGRAMA DE MANTENIMIENTO

Haya o no errores en el modelo, se hace un mantenimiento cada cierto periodo de tiempo definidos en las primeras etapas.