Proyecto Final de Cienciade Datos

David Saldaña Sánchez

Introducción a la Ciencia de Datos / Jaime Alejandro Romero Sierra

Miércoles 27 de noviembredel 2024

2. Introducción.

Objetivo. Evaluar cómo el uso de la realidad virtual en la educación afecta el compromiso y creatividad de los estudiantes, así como la relación entre el acceso a equipos de RV y los resultados de aprendizaje.

Justificación. Este proyecto es importante porque la educación está en un punto de transformación donde la tecnología puede jugar un papel crucial para mejorar los métodos tradicionales de enseñanza. La realidad virtual, en particular, ofrece la posibilidad de crear entornos de aprendizaje altamente interactivos y personalizados que pueden ayudar a los estudiantes a entender conceptos complejos a través de la experiencia directa. Sin embargo, para aprovechar este potencial, es necesario comprender mejor los factores que influyen en la efectividad de la RV en contextos educativos reales. Investigaciones anteriores han demostrado que la RV puede mejorarla retención de información, aumentar la motivación de los estudiantes y fomentar habilidades como la creatividad y la resolución de problemas. No obstante, su aplicación práctica aún enfrenta obstáculos que deben ser superados. Al evaluar cómo la RV impacta el compromiso, la creatividad y los resultados de aprendizaje, este proyecto proporcionará datos valiosos para respaldar la toma de decisiones sobre su integración en el currículo educativo. También permitirá identificar qué factores, como el acceso a equipos o la competencia de los instructores, son críticos para maximizar su efectividad. En última instancia, este proyecto busca no solo proporcionar una visión clara del estado actual del uso de la RV en la educación, sino también ofrecer recomendaciones concretas para su mejora y expansión. Esto puede llevar a una adopción más generalizada de la RV en diversas áreas del conocimiento, lo que, a largo plazo, beneficiará a estudiantes, educadores y a la infraestructura educativa en su conjunto.

Contexto. La realidad virtual (RV) ha sido vista como una herramienta revolucionaria en la educación, capaz de proporcionar experiencias inmersivas que fomentan un aprendizaje más profundo y significativo. Sin embargo, su implementación aún es limitada y enfrenta varios desafíos, especialmente en términos de acceso a la tecnologíay formación adecuada de los instructores. El problema que aborda este proyecto es entender cómo la RV está siendo utilizada actualmente en los entornos educativos y cuál es su impacto real en el aprendizaje de los estudiantes, el compromiso con sus estudios y su creatividad. También se explorará cómo el acceso desigual a los equipos de RV y las habilidades de los instructores en esta tecnología afectan estos resultados. En muchos casos, las instituciones educativas enfrentan dificultades para integrar la RV debido a los costos de los equipos y la falta de recursos formativos para los educadores. Esto ha generado una variabilidad en la experiencia de los estudiantes, con algunos accediendo plenamente a esta tecnología y otros limitados por la falta de equipos o instructores capacitados. Este proyecto investigará si los estudiantes que tienen acceso

a equipos de RV y reciben instrucción de docentes capacitados en el uso de esta tecnología logran mejores resultados académicos y presentan un mayor compromiso y creatividad en comparación con aquellos que no tienen estas ventajas. Además, se analizarán las barreras que puedan estar limitando el uso más amplio de la RV, como las diferencias geográficas en el acceso a la tecnología y el apoyo institucional. La implementación exitosa de RV en la educación depende no solo de la tecnología en sí, sino también de la infraestructura de apoyo que rodea su uso, desde la formación de los instructores hasta las decisiones administrativas sobre su inclusión en el currículo.

Fuente de Datos. Base de datos con las siguientes características: 4939 registros distribuidos en 19 columnas, detalladas de la siguiente manera:

• **Edad:** Valores representando edades (13, 16, 15, 24, 21, 28, 19, 29, 26, 22, 27, 18, 17, 23, 25, 12, 30,14, 20)

• **Género:** Género de los estudiantes ('No-Binario', 'Prefiero no decirlo', 'Mujer', 'Hombre')

• **Nivel_Grado:** Nivel educativo ('Post-Grado', 'Pre-Grado', 'Secundaria')

• **Campo_Estudio:** Campo de estudio ('Ciencias', 'Medicina', 'Ingeniería', 'Artes', 'Negocios', 'Educación', 'Leyes')

• **Uso_RV_Educación:** Uso de realidad virtual en educación ('No', 'Si')

• **Horas_RV_Semana:** Horas de uso de VR por semana (6, 4, 2, 10, 9, 1, 0, 5, 3, 8, 7)

• **Nivel_de_Compromiso:** Nivel de compromiso (1, 5, 4, 3, 2)

• **Mejora_en_Resultados_Aprendizaje:** Mejora en resultados de aprendizaje ('Si', 'No')

• **Materia:** Asignatura ('Ciencias de la computación', 'Matemáticas', 'Arte', 'Economía', 'Historia', 'Física', 'Biología')

• **Nivel_Instructor_RV:** Nivel de competencia en VR del instructor ('Intermedio', 'Principiante', 'Avanzado')

• **Efectividad_Persivida_de_la_RV:** Efectividad percibida del uso de VR (3, 2, 5, 4, 1)

• **Acceso_Equipo_VR:** Acceso al equipo de VR ('Si', 'No')

• **Impacto_en_Creatividad:** Impacto en la creatividad (5, 3, 2, 1, 4)

• **Nivel_Estres_Usando_VR:** Nivel de estrés relacionado con el uso de VR ('Alto', 'Bajo','Medio')

• **Colaboración_con_Compañeros_a_través_RV:** Colaboración con compañeros a través de VR ('No', 'Si')

• **Retroalimentación_de_Educadores_Sobre_RV:** Retroalimentación de los educadores sobre el uso de VR ('Neutral', 'Positiva', 'Negativa')

• **Interés_en_Seguir_con_Aprendizaje_Basado_RV:** Interés en continuar el aprendizaje basado en VR ('No', 'Si')

• **Región:** Región geográfica ('Asía', 'Oceanía', 'Europa', 'America del Norte', 'África', 'América del Sur')

• **Apoyo_Escuela_para_RV_en_el_currículo:** Apoyo escolar para la integración de VRen el currículo ('No', 'Si')

La base de datos no contiene datos nulos (NaN), valores inválidos ni filasduplicadas

Proceso de Limpieza de Datos

1. Eliminación de Duplicados en 'Student_ID':

Comencé centrando la atención en la columna **‘Student_ID’**, que actúa como nuestro índice único. Primero, eliminé los registros duplicados en esta columna, creando una nueva columna que excluye los valores NaN. Luego, eliminé los NaN de la columna original y combiné ambas columnas, lo que limpió los datos duplicados, aunque los valores NaN permanecieron en la columna original.

2. Manejo de Valores Numéricos:

A continuación, me enfoqué en las columnas que contenían datos numéricos. Utilicé **pd.to_numeric** para convertir los valores 'bbb' en NaN. Luego, apliqué boxplots para identificar y descartar los valores atípicos. Finalmente, utilicé **fillna** para rellenar los NaN restantes con el promedio de cada columna.

3. Conversión de Valores No Numéricos:

Para las columnas que contenían valores no numéricos, asigné un número a cada variable a través de un diccionario. Después, apliqué **pd.to_numeric** para convertir estas variables en valores numéricos. Utilicé **fillna** nuevamente para reemplazar los NaN con el promedio correspondiente.

4. Reasignación de 'Student_ID':

Posteriormente, en la columna **‘Student_ID’**, asigné un nuevo valor lineal y reinicié el índice usando **reset_index**. Creé una nueva columna con el índice, precedido por el prefijo ‘STUD’. Reemplacé los valores de la columna original y eliminé la columna temporal.

5. Restauración de Valores Originales:

Después, revertí los diccionarios que había utilizado para convertir las variables de texto a números, para restaurar las columnas a sus valores originales.

6. Renombrado de Columnas:

Utilicé **rename** para cambiar los nombres de las columnas del inglés al español, asegurando que los nombres fueran claros y representativos.

7. Conversión de Valores a Español:

A través de diccionarios, convertí los valores de cada columna al español (por ejemplo, 'male' se cambió a 'hombre').

8. Cambio de Tipo de Datos:

Una vez que la base de datos estuvo limpia y completa, cambié el tipo de dato de las columnas numéricas a entero utilizando **astype.**

9. Eliminación de la Columna 'Student_ID':

Finalmente, eliminé la columna **‘Student_ID’** original usando **.drop.**

Pasos importantes

1. **borrar los datos duplicados sin contar los NaN de la columna 'Student_ID'.**

```
# Los datos duplicados son en toda la fila y no solo en 'Student_ID'
# Vamos a borrar los datos duplicados sin contar los NaN

# Creamos una copia del df1
df2 = df.copy()

# Separar filas con NaN en 'Student_ID'
nan_rows = df2[df2['Student_ID'].isna()]

# Eliminar duplicados solo en los valores no NaN de 'col1'
df2_no_nan = df2.dropna(subset=['Student_ID']).drop_duplicates(subset=['Student_ID'], keep='first')

# Combinar nuevamente las filas con NaN
df3 = pd.concat([df2_no_nan, nan_rows])

df3
```

	Student_ID	Age	Gender	Grade_Level	Field_of_Study	Usage_of_VR_in_Education	Hours_of_VR_Usage_Per_Week	Engagement_Level	Improvement_in_Learning_Outcomes	Subject	In
0	STUD0001	13	Non-binary	Postgraduate	Science	No	6.0	1	Yes	Computer Science	
1	STUD0002	16	Non-binary	Undergraduate	Medicine	No	6.0	1	Yes	Math	
2	STUD0003	15	Prefer not to say	High School	Science	No	4.0	5	Yes	Art	
3	STUD0004	24	Female	Postgraduate	Engineering	Yes	2.0	4	No	Economics	
4	STUD0005	bbb	Non-binary	Undergraduate	Arts	Yes	10.0	3	No	Art	
...	...	...	...	...	...	...	...	...	...	...	
5240	NaN	14	Male	High School	Science	No	9.0	5	Yes	Physics	
5252	NaN	17	Female	Postgraduate	Science	Yes	4.0	4	Yes	Math	
5273	NaN	12	Male	Postgraduate	Business	No	2.0	2	No	Physics	
5293	NaN	21	Male	Postgraduate	Medicine	Yes	8.0	5	No	History	
5341	NaN	28	Prefer not to say	Undergraduate	Education	Yes	0.0	4	No	Physics	

4939 rows × 20 columns

2. Usar **pd.to_numeric** en datos ya numéricos y sacar el promedio.

```
# Ya no hay datos duplicados
# Para solucionar los datos incorrectos de todas las columnas restantes, vamos a remplazar los por el promedio de cada columna
# Primero, lo vamos a hacer con columnas que ya tengan los datos en numéricos

df4 = df3.copy()

# Convertimos los datos a tipo flotante
# Además todo lo que no sea un dato numérico lo vamos a transformar en un NaN
df4['Age'] = pd.to_numeric(df4['Age'], errors='coerce')
df4['Hours_of_VR_Usage_Per_Week'] = pd.to_numeric(df4['Hours_of_VR_Usage_Per_Week'], errors='coerce')
df4['Engagement_Level'] = pd.to_numeric(df4['Engagement_Level'], errors='coerce')
df4['Perceived_Effectiveness_of_VR'] = pd.to_numeric(df4['Perceived_Effectiveness_of_VR'], errors='coerce')
df4['Impact_on_Creativity'] = pd.to_numeric(df4['Impact_on_Creativity'], errors='coerce')
```

```
# Segundo, antes de sacar el promedio eliminaremos los datos atípicos
# Se define el cuartil 75 y se le resta el cuartil 25
iqr = df4['Age'].quantile(0.75) - df4['Age'].quantile(0.25)
# Desarrollamos filtros superior o inferior
filtro_inferior = df4['Age'] > df4['Age'].quantile(0.25) - (iqr * 1.5)
filtro_superior = df4['Age'] < df4['Age'].quantile(0.75) + (iqr * 1.5)

df_filtrado_age = df4[filtro_inferior & filtro_superior]

# Graficando el boxplot
fig = px.box(df_filtrado_age, y='Age', title='Datos atípicos eliminados')
fig.update_layout(width=300,height=500)
fig.show()
```

```
[ ] # Ahora que ya no consideramos los datos atípicos sacamos el promedio
    promedio_age = df_filtrado_age['Age'].mean()
    promedio_age

21.148180055998278
```

```
[ ] # Repetimos el proceso con demás columnas
    # Lista de columnas restantes
    columnas = ['Hours_of_VR_Usage_Per_Week', 'Engagement_Level', 'Perceived_Effectiveness_of_VR', 'Impact_on_Creativity']

    # Diccionario para almacenar los promedios
    promedios = {}

    # Iterar sobre cada columna en la lista
    for columna in columnas:
        # Calcular el IQR
        iqr = df4[columna].quantile(0.75) - df4[columna].quantile(0.25)

        # Filtros superior e inferior
        filtro_inferior = df4[columna] > df4[columna].quantile(0.25) - (iqr * 1.5)
        filtro_superior = df4[columna] < df4[columna].quantile(0.75) + (iqr * 1.5)

        # Filtrar el DataFrame
        df_filtrado = df4[filtro_inferior & filtro_superior]

        # Calcular el promedio de la columna filtrada
        promedio = df_filtrado[columna].mean()

        # Almacenar el promedio en el diccionario
        promedios[columna] = promedio

        # Graficar el boxplot
        fig = px.box(df_filtrado, y=columna, title=f'Datos atípicos eliminados para {columna}')
        fig.update_layout(width=300, height=500)
        fig.show()
```

```
0s  # Imprimimos los promedios
    promedios

{'Hours_of_VR_Usage_Per_Week': 5.022362869198313,
 'Engagement_Level': 3.0174681906405003,
 'Perceived_Effectiveness_of_VR': 2.9436264198569626,
 'Impact_on_Creativity': 3.012205387205387}
```

3. Remplazar los NaN con el promedio que calculamos

```
[ ] # Remplazamos los NaN de cada columna con los datos anteriores

    df4['Age'] = df4['Age'].fillna(21)
    df4['Hours_of_VR_Usage_Per_Week'] = df4['Hours_of_VR_Usage_Per_Week'].fillna(5)
    df4['Engagement_Level'] = df4['Engagement_Level'].fillna(3)
    df4['Perceived_Effectiveness_of_VR'] = df4['Perceived_Effectiveness_of_VR'].fillna(3)
    df4['Impact_on_Creativity'] = df4['Impact_on_Creativity'].fillna(3)
```

4. Asignar un número a cada valor de las columnas no numéricas (ejemplo)

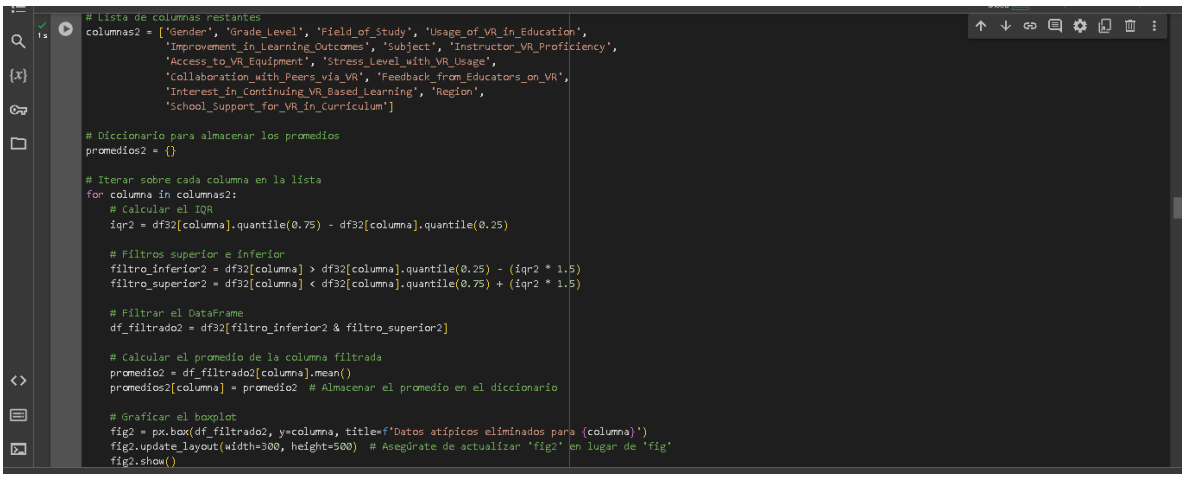
```
[ ] # Definir un diccionario para la asignación
    Gender = {
        'Non-binary': 1,
        'Prefer not to say': 2,
        'Female': 3,
        'Male': 4
    }

    # Reemplazar los valores en la columna 'Género'
    df5 = df4.copy()
    df5['Gender'] = df5['Gender'].replace(Gender)

    df5
```

	Student_ID	Age	Gender	Grade_Level	Field_of_Study	Usage_of_VR_in_Education	Hours_of_VR_Usage_Per_Week	Engagement_Level	Improvement_in_Learning_Outcomes	Subject	In
0	STUD0001	13.0	1	Postgraduate	Science	No	6.0	1.0	Yes	Computer Science	
1	STUD0002	16.0	1	Undergraduate	Medicine	No	6.0	1.0	Yes	Math	
2	STUD0003	15.0	2	High School	Science	No	4.0	5.0	Yes	Art	
3	STUD0004	24.0	3	Postgraduate	Engineering	Yes	2.0	4.0	No	Economics	
4	STUD0005	21.0	1	Undergraduate	Arts	Yes	10.0	3.0	No	Art	
...	...	...	...	...	...	...	...	...	...	...	
5240	NaN	14.0	4	High School	Science	No	9.0	5.0	Yes	Physics	
5252	NaN	17.0	3	Postgraduate	Science	Yes	4.0	4.0	Yes	Math	
5273	NaN	12.0	4	Postgraduate	Business	No	2.0	2.0	No	Physics	
5293	NaN	21.0	4	Postgraduate	Medicine	Yes	8.0	5.0	No	History	
5341	NaN	28.0	2	Undergraduate	Education	Yes	0.0	4.0	No	Physics	

4939 rows × 20 columns

Pasos siguientes: Generar código con df5 | Ver gráficos recomendados | New interactive sheet

5. Calcular el promedio de estas columnas sin considerar los datos atípicos

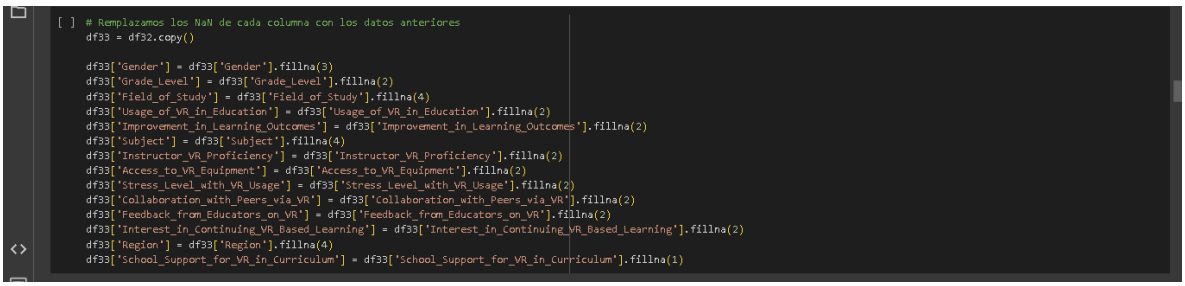


```
# Lista de columnas restantes
columnas2 = ['Gender', 'Grade_Level', 'Field_of_Study', 'Usage_of_VR_in_Education',
             'Improvement_in_Learning_Outcomes', 'Subject', 'Instructor_VR_Proficiency',
             'Access_to_VR_Equipment', 'Stress_Level_with_VR_Usage',
             'Collaboration_with_Peers_via_VR', 'Feedback_from_Educators_on_VR',
             'Interest_in_Continuing_VR_Based_Learning', 'Region',
             'School_Support_for_VR_in_Curriculum']

# Diccionario para almacenar los promedios
promedios2 = {}

# Iterar sobre cada columna en la lista
for columna in columnas2:
    # Calcular el IQR
    iqr2 = df32[columna].quantile(0.75) - df32[columna].quantile(0.25)

    # Filtros superior e inferior
    filtro_inferior2 = df32[columna] > df32[columna].quantile(0.25) - (iqr2 * 1.5)
    filtro_superior2 = df32[columna] < df32[columna].quantile(0.75) + (iqr2 * 1.5)

    # Filtrar el DataFrame
    df_filtrado2 = df32[filtro_inferior2 & filtro_superior2]

    # Calcular el promedio de la columna filtrada
    promedio2 = df_filtrado2[columna].mean()
    promedios2[columna] = promedio2  # Almacenar el promedio en el diccionario

    # Graficar el boxplot
    fig2 = px.box(df_filtrado2, y=columna, title=f'Datos atípicos eliminados para {columna}')
    fig2.update_layout(width=300, height=500)  # Asegúrate de actualizar 'fig2' en lugar de 'fig'
    fig2.show()
```

```
# Imprimimos los promedios
promedios2
```

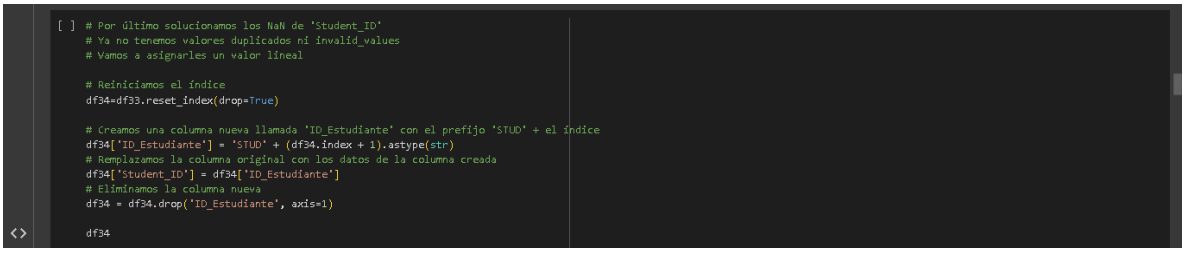
```
{'Gender': 2.503329752953813,
 'Grade_Level': 2.001687051876845,
 'Field_of_Study': 3.965990251390053,
 'Usage_of_VR_in_Education': 1.504323982282219,
 'Improvement_in_Learning_Outcomes': 1.503581963758955,
 'Subject': 4.00633445945946,
 'Instructor_VR_Proficiency': 1.9924114671163575,
 'Access_to_VR_Equipment': 1.505698607007176,
 'Stress_Level_with_VR_Usage': 1.9795302736479208,
 'Collaboration_with_Peers_via_VR': 1.5043047778039174,
 'Feedback_from_Educators_on_VR': 1.9928894634776988,
 'Interest_in_Continuing_VR_Based_Learning': 1.505054850505485,
 'Region': 3.5517687661777395,
 'School_Support_for_VR_in_Curriculum': 1.4864978902953587}
```

6. Remplazar los valores con los datos anteriores

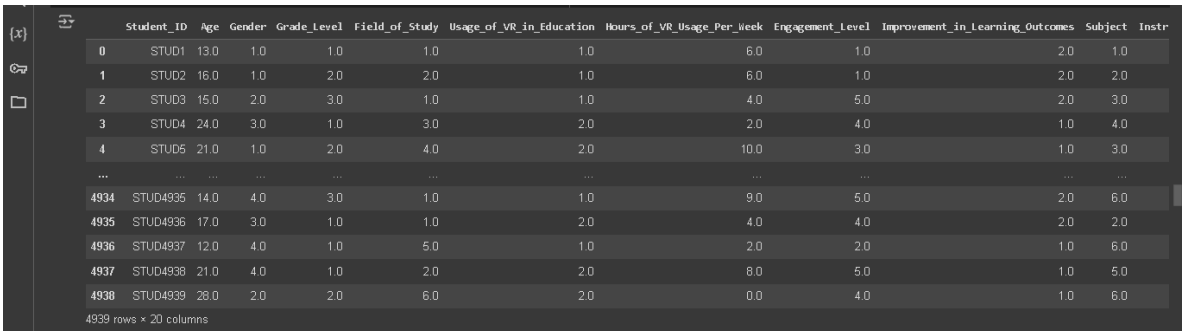


```
# Remplazamos los NaN de cada columna con los datos anteriores
df33 = df32.copy()

df33['Gender'] = df33['Gender'].fillna(3)
df33['Grade_Level'] = df33['Grade_Level'].fillna(2)
df33['Field_of_Study'] = df33['Field_of_Study'].fillna(4)
df33['Usage_of_VR_in_Education'] = df33['Usage_of_VR_in_Education'].fillna(2)
df33['Improvement_in_Learning_Outcomes'] = df33['Improvement_in_Learning_Outcomes'].fillna(2)
df33['Subject'] = df33['Subject'].fillna(4)
df33['Instructor_VR_Proficiency'] = df33['Instructor_VR_Proficiency'].fillna(2)
df33['Access_to_VR_Equipment'] = df33['Access_to_VR_Equipment'].fillna(2)
df33['Stress_Level_with_VR_Usage'] = df33['Stress_Level_with_VR_Usage'].fillna(2)
df33['Collaboration_with_Peers_via_VR'] = df33['Collaboration_with_Peers_via_VR'].fillna(2)
df33['Feedback_from_Educators_on_VR'] = df33['Feedback_from_Educators_on_VR'].fillna(2)
df33['Interest_in_Continuing_VR_Based_Learning'] = df33['Interest_in_Continuing_VR_Based_Learning'].fillna(2)
df33['Region'] = df33['Region'].fillna(4)
df33['School_Support_for_VR_in_Curriculum'] = df33['School_Support_for_VR_in_Curriculum'].fillna(1)
```

7. Arreglar el índice de la columna **‘Student_ID’**



```
# Por último solucionamos los NaN de 'Student_ID'
# Ya no tenemos valores duplicados ni invalid_values
# Vamos a asignarles un valor lineal

# Reiniciamos el índice
df34=df33.reset_index(drop=True)

# Creamos una columna nueva llamada 'ID_Estudiante' con el prefijo 'STUD' + el índice
df34['ID_Estudiante'] = 'STUD' + (df34.index + 1).astype(str)
# Remplazamos la columna original con los datos de la columna creada
df34['Student_ID'] = df34['ID_Estudiante']
# Eliminamos la columna nueva
df34 = df34.drop('ID_Estudiante', axis=1)

df34
```

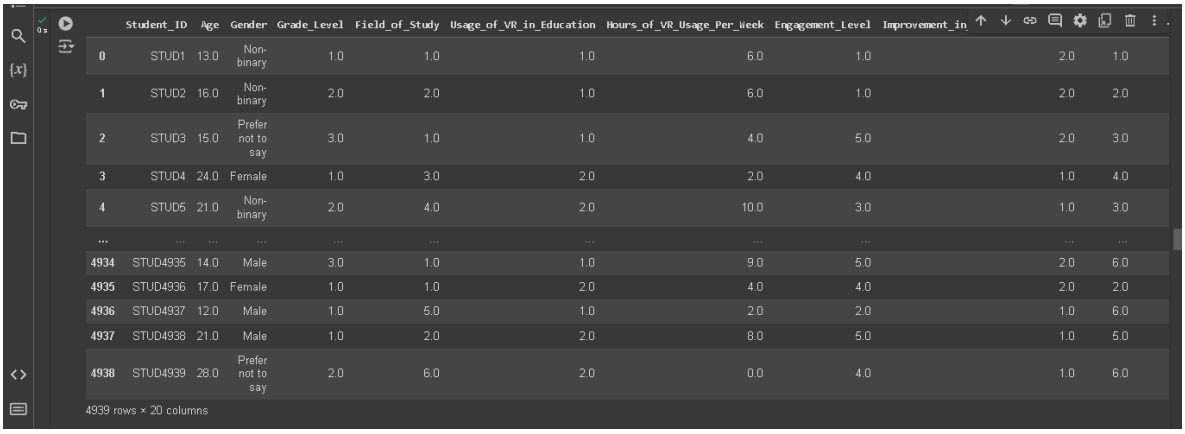


	Student_ID	Age	Gender	Grade_Level	Field_of_Study	Usage_of_VR_in_Education	Hours_of_VR_Usage_Per_Week	Engagement_Level	Improvement_in_Learning_Outcomes	Subject	Instr
0	STUD1	13.0	1.0	1.0	1.0	1.0	6.0	1.0	2.0	1.0	
1	STUD2	16.0	1.0	2.0	2.0	1.0	6.0	1.0	2.0	2.0	
2	STUD3	15.0	2.0	3.0	1.0	1.0	4.0	5.0	2.0	3.0	
3	STUD4	24.0	3.0	1.0	3.0	2.0	2.0	4.0	1.0	4.0	
4	STUD5	21.0	1.0	2.0	4.0	2.0	10.0	3.0	1.0	3.0	
...	...	...	...	...	...	...	...	...	...	...	
4934	STUD4935	14.0	4.0	3.0	1.0	1.0	9.0	5.0	2.0	6.0	
4935	STUD4936	17.0	3.0	1.0	1.0	2.0	4.0	4.0	2.0	2.0	
4936	STUD4937	12.0	4.0	1.0	5.0	1.0	2.0	2.0	1.0	6.0	
4937	STUD4938	21.0	4.0	1.0	2.0	2.0	8.0	5.0	1.0	5.0	
4938	STUD4939	28.0	2.0	2.0	6.0	2.0	0.0	4.0	1.0	6.0	

4939 rows × 20 columns

8. Regresar los valores que convertimos a numéricos a su valor original (ejemplo)

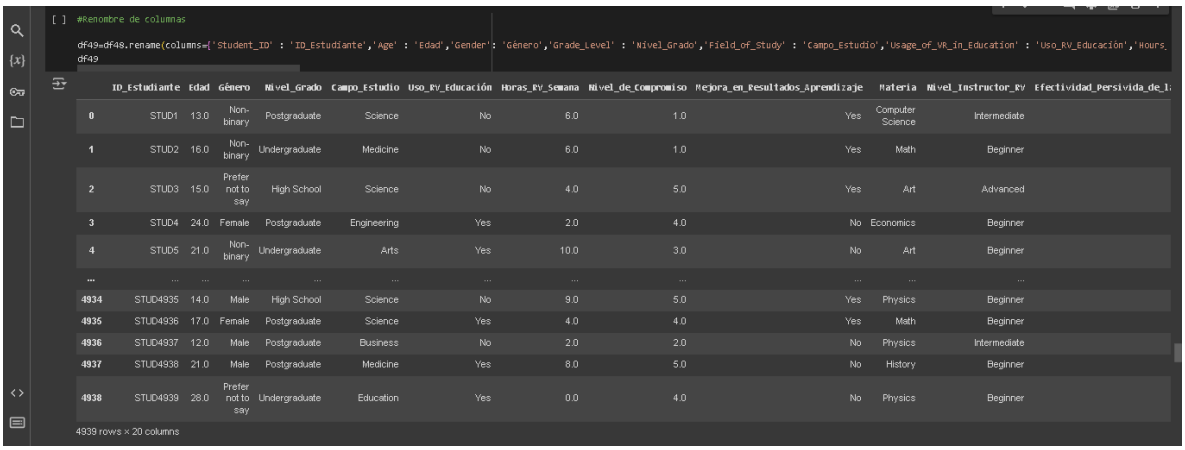
```
# Como ya tenemos nuestra base completamente llena regresaremos los valores a su tipo origina

df35 = df34.copy()

# Convertir los valores de la columna 'Gender' a cadenas
df35['Gender'] = df35['Gender'].astype(str)

# Definir un diccionario para la asignación
GenderOriginal = {
    '1.0': "Non-binary",
    '2.0': "Prefer not to say",
    '3.0': "Female",
    '4.0': "Male"
}

# Reemplazar los valores en la columna 'Género'
df35['Gender'] = df35['Gender'].replace(GenderOriginal)

df35
```



	Student_ID	Age	Gender	Grade_Level	Field_of_Study	Usage_of_VR_in_Education	Hours_of_VR_Usage_Per_Week	Engagement_Level	Improvement_in		
0	STUD1	13.0	Non-binary	1.0	1.0	1.0	6.0	1.0		2.0	1.0
1	STUD2	16.0	Non-binary	2.0	2.0	1.0	6.0	1.0		2.0	2.0
2	STUD3	15.0	Prefer not to say	3.0	1.0	1.0	4.0	5.0		2.0	3.0
3	STUD4	24.0	Female	1.0	3.0	2.0	2.0	4.0		1.0	4.0
4	STUD5	21.0	Non-binary	2.0	4.0	2.0	10.0	3.0		1.0	3.0
...	...	...	...	...	...	...	...	...		...	...
4934	STUD4935	14.0	Male	3.0	1.0	1.0	9.0	5.0		2.0	6.0
4935	STUD4936	17.0	Female	1.0	1.0	2.0	4.0	4.0		2.0	2.0
4936	STUD4937	12.0	Male	1.0	5.0	1.0	2.0	2.0		1.0	6.0
4937	STUD4938	21.0	Male	1.0	2.0	2.0	8.0	5.0		1.0	5.0
4938	STUD4939	28.0	Prefer not to say	2.0	6.0	2.0	0.0	4.0		1.0	6.0

4939 rows × 20 columns

9. Renombre de las columnas a español



```
#Renombre de columnas

df49=df48.rename(columns={'Student_ID' : 'ID_Estudiante','Age' : 'Edad','Gender': 'Género','Grade_Level' : 'Nivel_Grado','Field_of_Study' : 'Campo_Estudio','Usage_of_VR_in_Education' : 'Uso_RV_Educación','Hours_
df49
```

	ID_Estudiante	Edad	Género	Nivel_Grado	Campo_Estudio	Uso_RV_Educación	Horas_RV_Semana	Nivel_de_Compromiso	Mejora_en_Resultados_Aprendizaje	Materia	Nivel_Instructor_RV	Efectividad_Persivida_de_l
0	STUD1	13.0	Non-binary	Postgraduate	Science	No	6.0	1.0	Yes	Computer Science	Intermediate	
1	STUD2	16.0	Non-binary	Undergraduate	Medicine	No	6.0	1.0	Yes	Math	Beginner	
2	STUD3	15.0	Prefer not to say	High School	Science	No	4.0	5.0	Yes	Art	Advanced	
3	STUD4	24.0	Female	Postgraduate	Engineering	Yes	2.0	4.0	No	Economics	Beginner	
4	STUD5	21.0	Non-binary	Undergraduate	Arts	Yes	10.0	3.0	No	Art	Beginner	
...	...	...	...	...	...	...	...	...	...	...	...	
4934	STUD4935	14.0	Male	High School	Science	No	9.0	5.0	Yes	Physics	Beginner	
4935	STUD4936	17.0	Female	Postgraduate	Science	Yes	4.0	4.0	Yes	Math	Beginner	
4936	STUD4937	12.0	Male	Postgraduate	Business	No	2.0	2.0	No	Physics	Intermediate	
4937	STUD4938	21.0	Male	Postgraduate	Medicine	Yes	8.0	5.0	No	History	Beginner	
4938	STUD4939	28.0	Prefer not to say	Undergraduate	Education	Yes	0.0	4.0	No	Physics	Beginner	

4939 rows × 20 columns

10. Traducción de los datos a español (ejemplo)

```
[ ] #Traducción de los datos

    df50 = df49.copy()

    traduccion1 = {
    'Non-binary' : 'No-Binario',
    'Prefer not to say': 'Prefiero no decirlo',
    'Female' : 'Mujer',
    'Male': 'Hombre',
    }

    df50['Género'] = df50['Género'].replace(traduccion1)
    df50
```

	ID_Estudiante	Edad	Género	Nivel_Grado	Campo_Estudio	Uso_RV_Educación	Horas_RV_Semana	Nivel_de_Compromiso	Mejora_en_Resultados_Aprendizaje	Materia	Nivel_Instructo
0	STUD1	13.0	No-Binario	Postgraduate	Science	No	6.0	1.0	Yes	Computer Science	Interme
1	STUD2	16.0	No-Binario	Undergraduate	Medicine	No	6.0	1.0	Yes	Math	Beg
2	STUD3	15.0	Prefiero no decirlo	High School	Science	No	4.0	5.0	Yes	Art	Adva
3	STUD4	24.0	Mujer	Postgraduate	Engineering	Yes	2.0	4.0	No	Economics	Beg
4	STUD5	21.0	No-Binario	Undergraduate	Arts	Yes	10.0	3.0	No	Art	Beg
...	...	...	...	...	...	...	...	...	...	...	
4934	STUD4935	14.0	Hombre	High School	Science	No	9.0	5.0	Yes	Physics	Beg
4935	STUD4936	17.0	Mujer	Postgraduate	Science	Yes	4.0	4.0	Yes	Math	Beg
4936	STUD4937	12.0	Hombre	Postgraduate	Business	No	2.0	2.0	No	Physics	Interme
4937	STUD4938	21.0	Hombre	Postgraduate	Medicine	Yes	8.0	5.0	No	History	Beg
4938	STUD4939	28.0	Prefiero no decirlo	Undergraduate	Education	Yes	0.0	4.0	No	Physics	Beg

4939 rows × 20 columns

11. Cambiar el tipo de datos

```
#Cambiar el tipo de datos

df51 = df50.copy()

df51['Edad']=df51['Edad'].astype(int)
df51['Horas_RV_Semana']=df51['Horas_RV_Semana'].astype(int)
df51['Nivel_de_Compromiso']=df51['Nivel_de_Compromiso'].astype(int)
df51['Efectividad_Persivida_de_la_RV']=df51['Efectividad_Persivida_de_la_RV'].astype(int)
df51['Impacto_en_Creatividad']=df51['Impacto_en_Creatividad'].astype(int)
```

12. Quitar la columna **‘Student_ID’**

```
[ ] #Quita el indice incluido en la base de datos

    df52=df51.drop(columns=['ID_Estudiante'])
    df52
```

	Edad	Género	Nivel_Grado	Campo_Estudio	Uso_RV_Educación	Horas_RV_Semana	Nivel_de_Compromiso	Mejora_en_Resultados_Aprendizaje	Materia	Nivel_Instructor_RV	Efectividad_Persivida_de_la_RV	Acceso_E
0	13	No-Binario	Post-Grado	Ciencias	No	6	1	Si	Ciencias de la computación	Intermedio	3	
1	16	No-Binario	Pre-Grado	Medicina	No	6	1	Si	Matemáticas	Principiante	2	
2	15	Prefiero no decirlo	Secundaria	Ciencias	No	4	5	Si	Arte	Avanzado	5	
3	24	Mujer	Post-Grado	Ingeniería	Si	2	4	No	Economía	Principiante	5	
4	21	No-Binario	Pre-Grado	Artes	Si	10	3	No	Arte	Principiante	4	
...	...	...	...	...	...	...	...	...	...	...	...	
4934	14	Hombre	Secundaria	Ciencias	No	9	5	Si	Física	Principiante	4	
4935	17	Mujer	Post-Grado	Ciencias	Si	4	4	Si	Matemáticas	Principiante	3	
4936	12	Hombre	Post-Grado	Negocios	No	2	2	No	Física	Intermedio	5	
4937	21	Hombre	Post-Grado	Medicina	Si	8	5	No	Historia	Principiante	2	
4938	28	Prefiero no decirlo	Pre-Grado	Educación	Si	0	4	No	Física	Principiante	3	

4939 rows × 19 columns

13. Guardar el archivo

```
[ ] #Guardar el archivo

    df52.to_csv("Proyecto_limpio_1.3.csv",index=False)
```

3. Metodología.

1. Descripción general de los datos.

4939 registros distribuidos en 19 columnas, detalladas de la siguiente manera:

• **Edad:** Valores representando edades (13, 16, 15, 24, 21, 28, 19, 29, 26, 22, 27, 18, 17, 23, 25, 12, 30,14, 20).

- **Variable entera.** Media: 21.139299**, Mediana:** 21.0, **Mínimo:** 12.0**, Máximo:** 30.0

• **Género:** Género de los estudiantes ('No-Binario', 'Prefiero no decirlo', 'Mujer', 'Hombre').

- **Variable tipo objeto.** Frecuencia: Mujer – 1413, Hombre – 1193, No-Binario -1172, Prefiero no decirlo - 1161

• **Nivel_Grado:** Nivel educativo ('Post-Grado', 'Pre-Grado', 'Secundaria').

- **Variable tipo objeto.** Frecuencia: Pre-Grado – 1801, Secundaria – 1573, Post-Grado - 1565

• **Campo_Estudio:** Campo de estudio ('Ciencias', 'Medicina', 'Ingeniería', 'Artes', 'Negocios', 'Educación', 'Leyes').

- **Variable tipo objeto.** Frecuencia: Artes – 953, Ciencias – 696, Negocios – 683, Medicina – 673, Leyes – 648, Educación – 645, Ingeniería - 641

• **Uso_RV_Educación:** Uso de realidad virtual en educación ('No', 'Si').

- **Variable tipo objeto.** Frecuencia: Si – 2589, No - 2350

• **Horas_RV_Semana:** Horas de uso de VR por semana (6, 4, 2, 10, 9, 1, 0, 5, 3, 8, 7).

- **Variable entera. Media:** 5.021462**, Mediana:** 5.0**, Mínimo:** 0**, Máximo:** 10.0

• **Nivel_de_Compromiso:** Nivel de compromiso (1, 5, 4, 3, 2).

- **Variable entera. Media:** 3.016400**, Mediana:** 3.0**, Mínimo:** 1.0**, Máximo:** 5.0

• **Mejora_en_Resultados_Aprendizaje:** Mejora en resultados de aprendizaje ('Si', 'No').

- **Variable tipo objeto.** Frecuencia: Si – 2583, No - 2356

• **Materia:** Asignatura ('Ciencias de la computación', 'Matemáticas', 'Arte', 'Economía', 'Historia', 'Física', 'Biología').

- **Variable tipo objeto.** Frecuencia: Economía - 875, Historia – 696, Física – 687, Arte – 684, Matemáticas – 675, Ciencia de la computación - 662, Biología - 660

• **Nivel_Instructor_RV:** Nivel de competencia en VR del instructor ('Intermedio', 'Principiante', 'Avanzado').

- **Variable tipo objeto.** Frecuencia: Principiante – 1809, Intermedio – 1583, Avanzado - 1547

• **Efectividad_Persivida_de_la_RV:** Efectividad percibida del uso de VR (3, 2, 5, 4, 1).

- **Variable entera. Media:** 2.945738**, Mediana:** 3.0**, Mínimo:** 1.0**, Máximo:** 5.0

• **Acceso_Equipo_VR:** Acceso al equipo de VR ('Si', 'No').

- **Variable tipo objeto.** Frecuencia: Si – 2597, No - 2342

• **Impacto_en_Creatividad:** Impacto en la creatividad (5, 3, 2, 1, 4).

- **Variable entera. Media:** 3.011743**, Mediana:** 3.0**, Mínimo:** 1.0**, Máximo:** 5.0

• **Nivel_Estres_Usando_VR:** Nivel de estrés relacionado con el uso de VR ('Alto', 'Bajo','Medio').

- **Variable tipo objeto.** Frecuencia: Bajo – 1868, Alto: 1583, Medio: 1488

• **Colaboración_con_Compañeros_a_través_RV:** Colaboración con compañeros através de VR ('No', 'Si').

- **Variable tipo objeto.** Frecuencia: Si – 2636, No - 2303

• **Retroalimentación_de_Educadores_Sobre_RV:** Retroalimentación de los educadores sobre el uso de VR ('Neutral', 'Positiva', 'Negativa').

- **Variable tipo objeto.** Frecuencia: Positiva – 1906, Neutral – 1533, Negativa - 1500

• **Interés_en_Seguir_con_Aprendizaje_Basado_RV:** Interés en continuar el aprendizaje basado en VR ('No', 'Si').

- **Variable tipo objeto.** Frecuencia: Si – 2638, No - 2301

• **Región:** Región geográfica ('Asía', 'Oceanía', 'Europa', 'América del Norte', 'África', 'América del Sur').

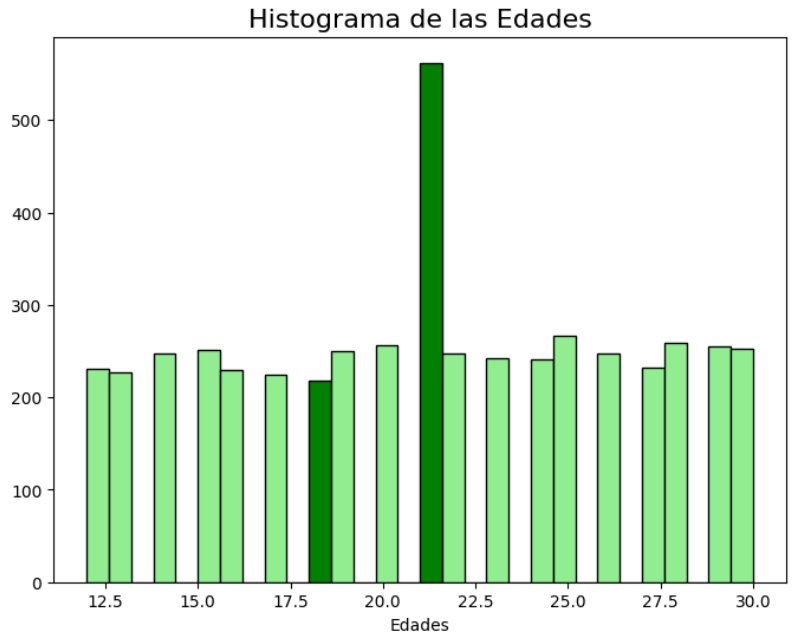
- **Variable tipo objeto.** Frecuencia: Asía – 1087, América del Sur – 813, África – 779, América del Norte – 770, Europa – 757, Oceanía - 733

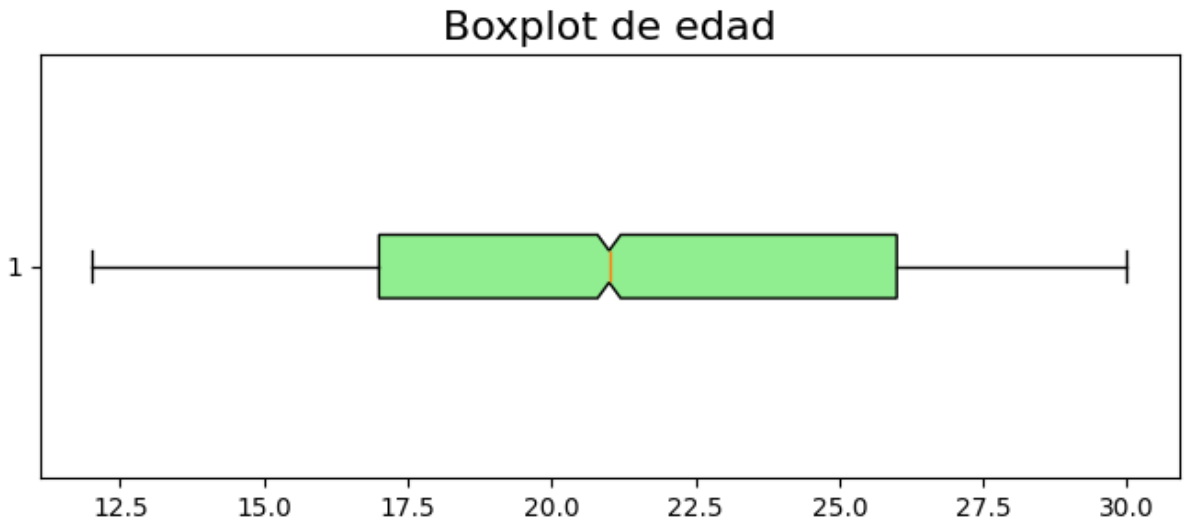
• **Apoyo_Escuela_para_RV_en_el_currículo:** Apoyo escolar para la integración de VR en el currículo ('No', 'Si').

- **Variable tipo objeto.** Frecuencia: Si – 2306, No - 2633

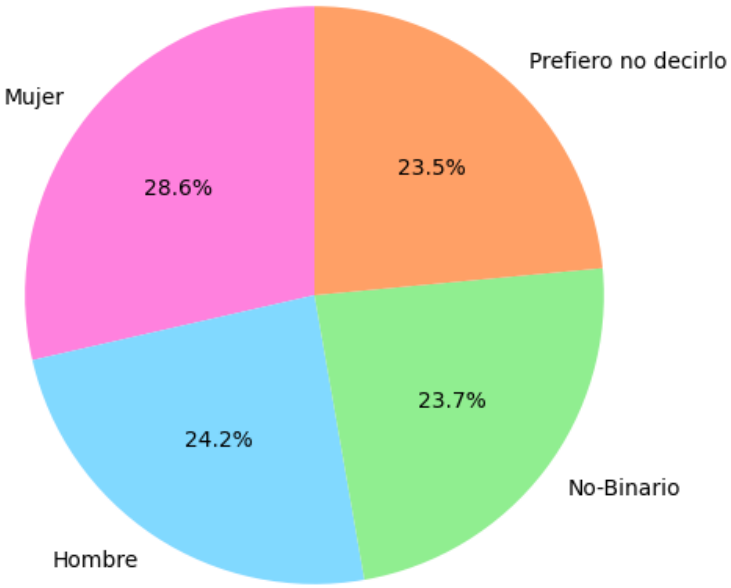
La base de datos no contiene datos nulos (NaN), valores inválidos ni filasduplicadas

2. Visualización y Distribución de Variables Individuales

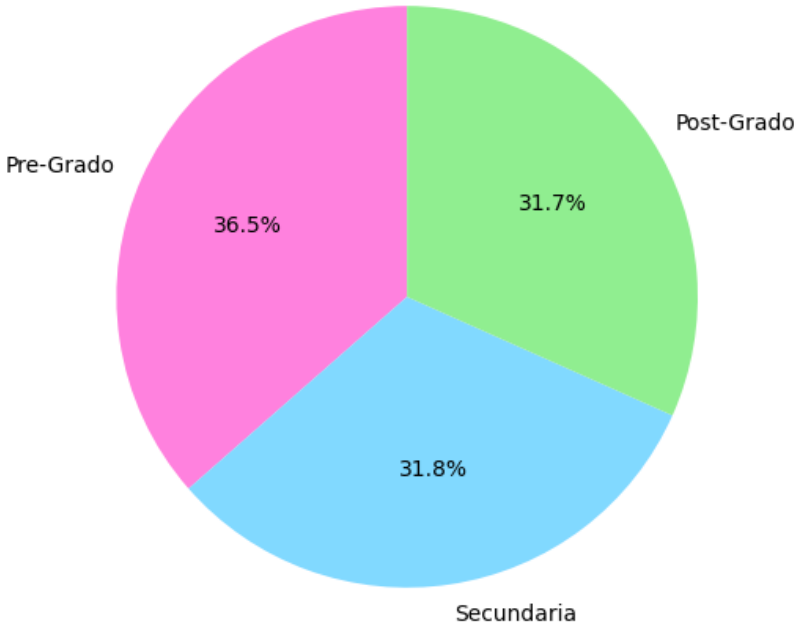


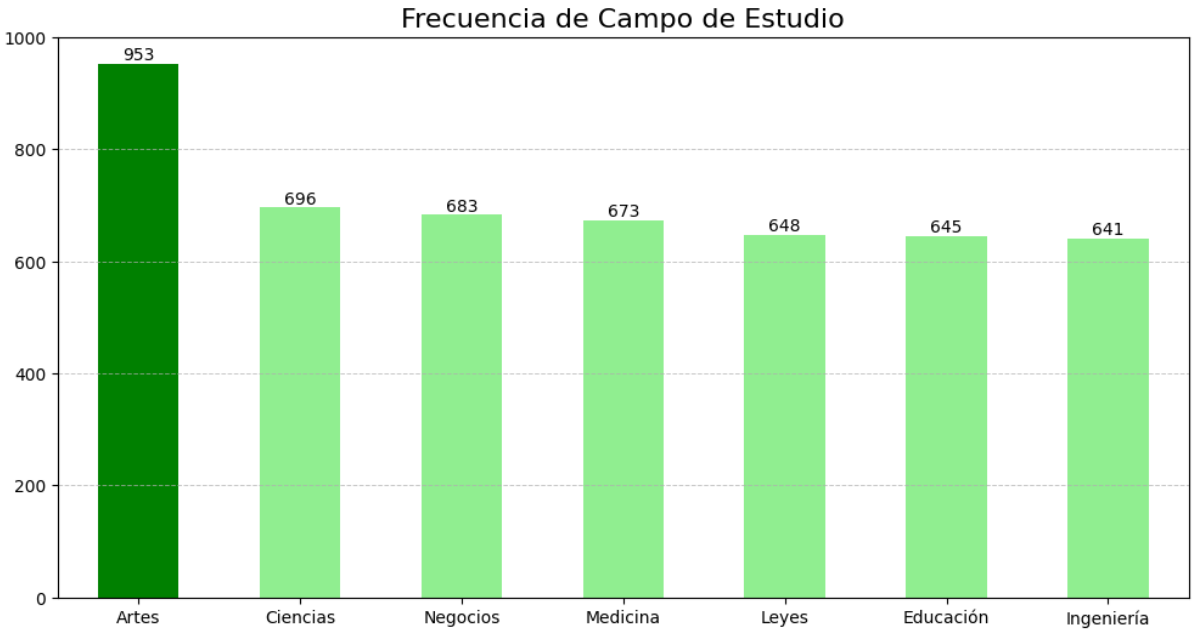


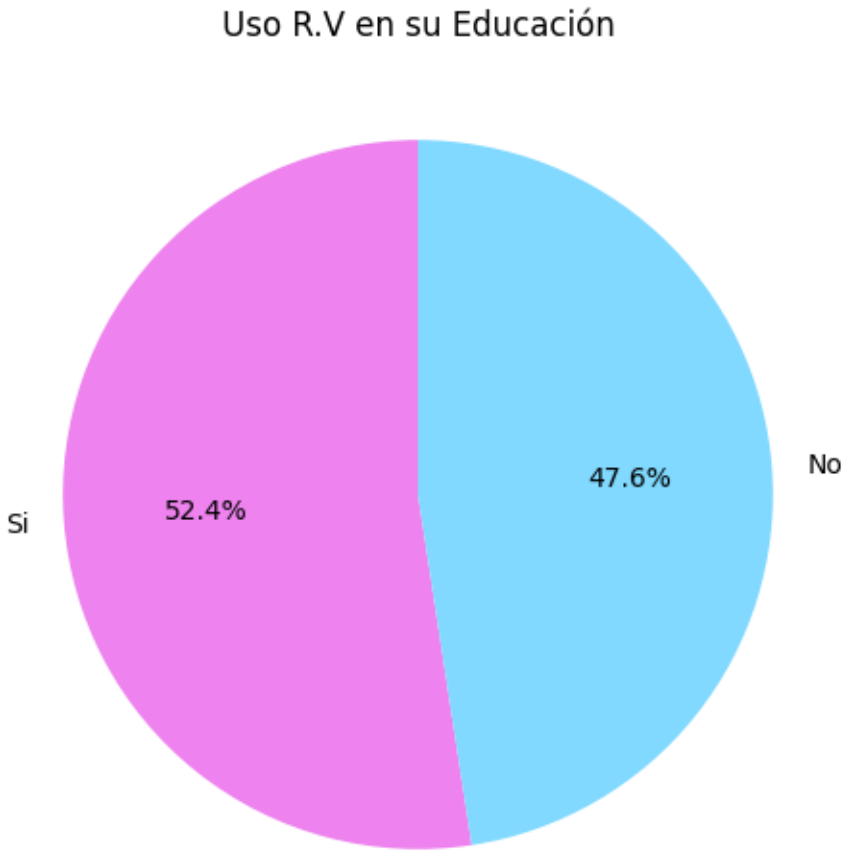
Distribución por Género

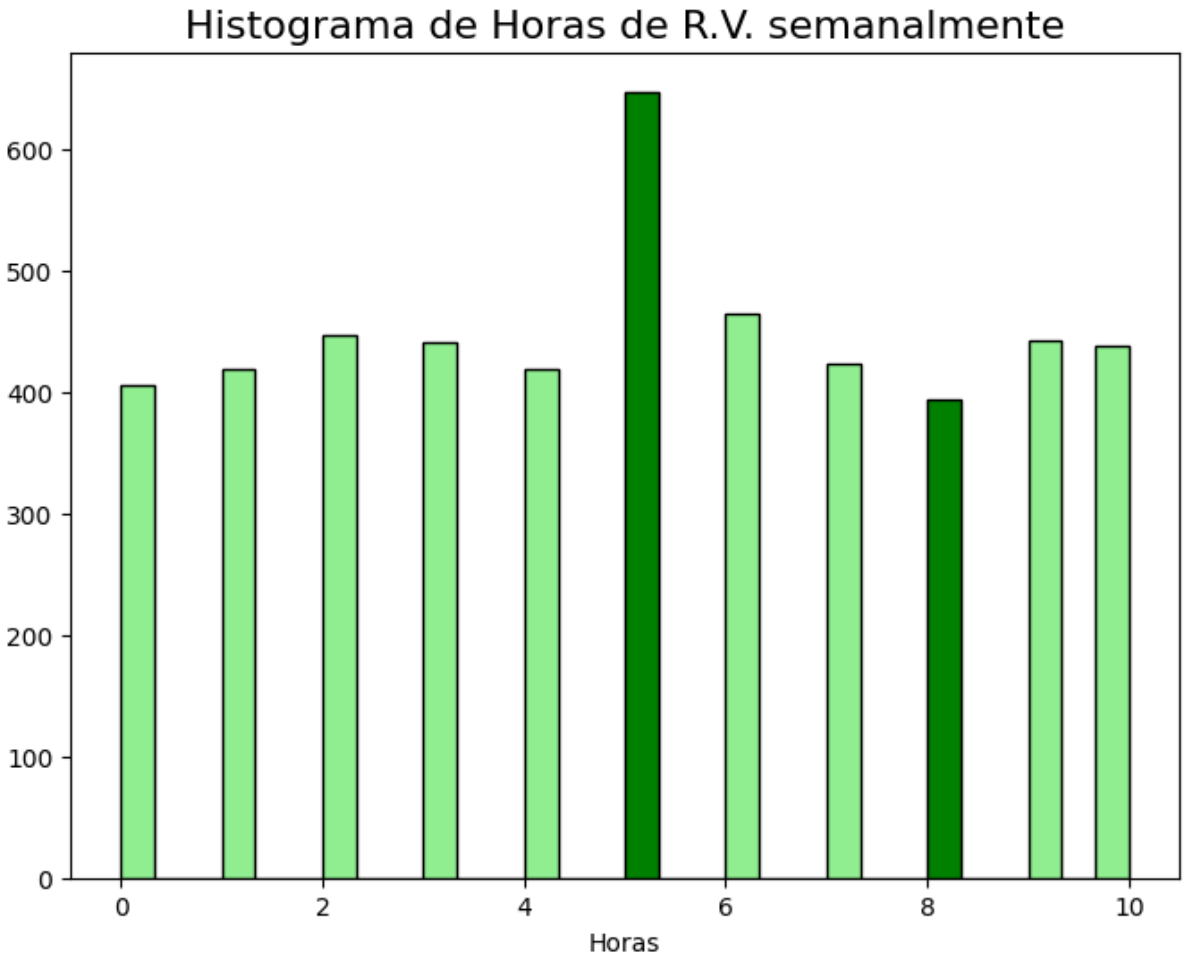


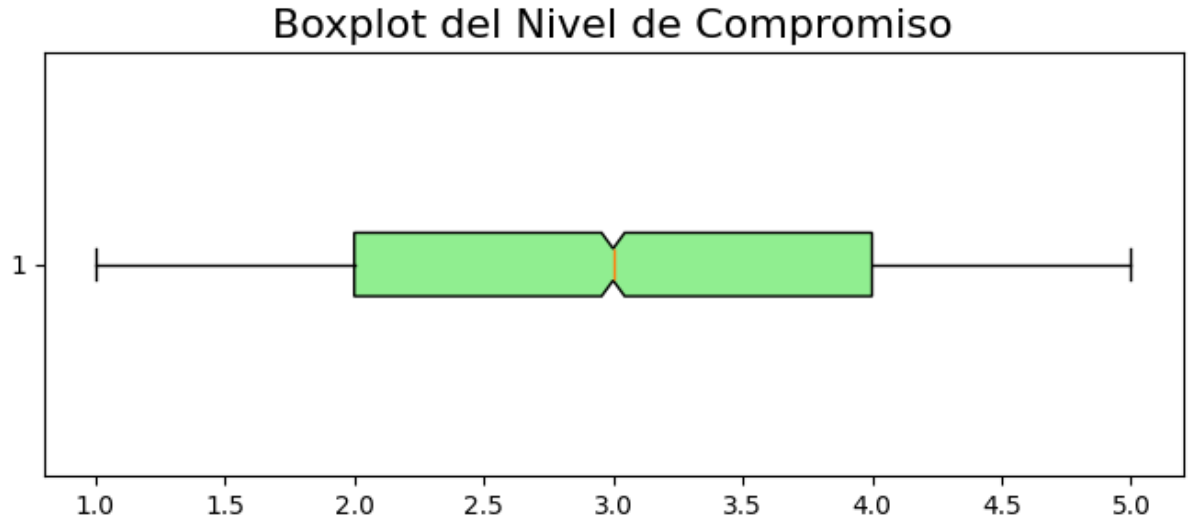
Distribución por Grado

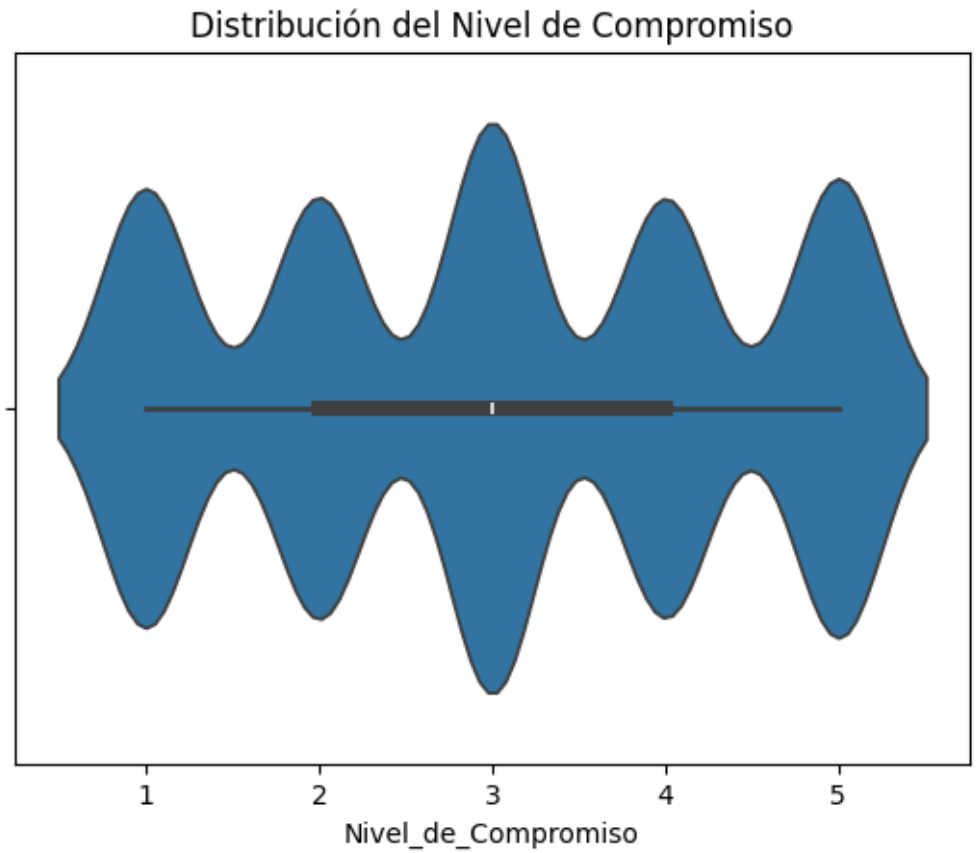


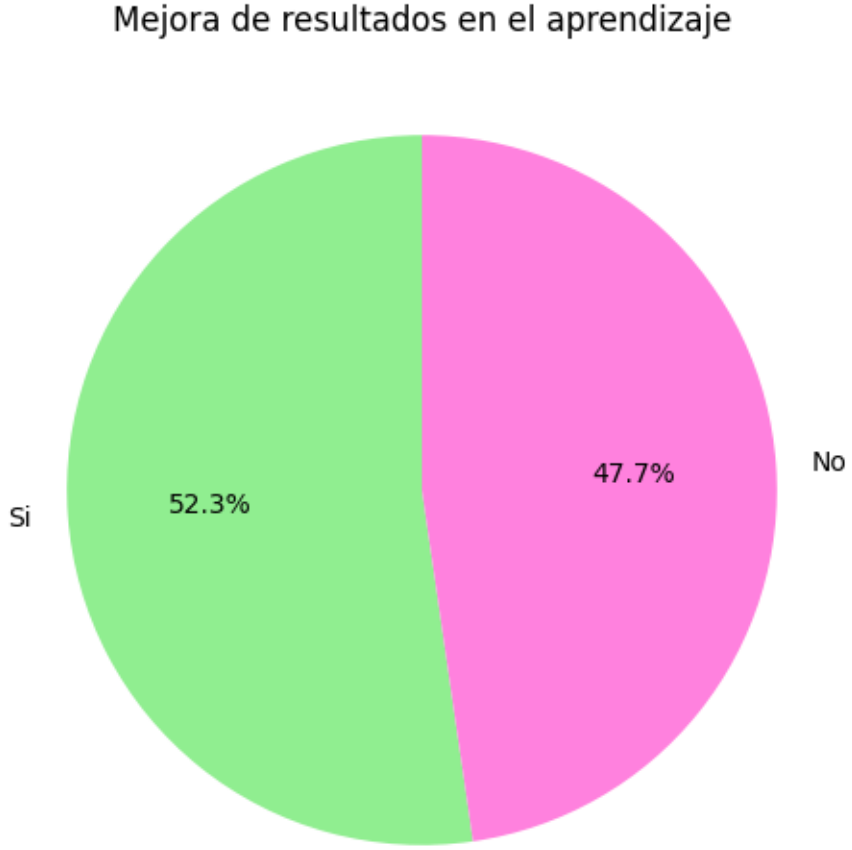

Frecuencia de Campo de Estudio
1000
800
600
400
200
0
953
696
683
673
648
645
641
Artes
Ciencias
Negocios
Medicina
Leyes
Educación
Ingeniería

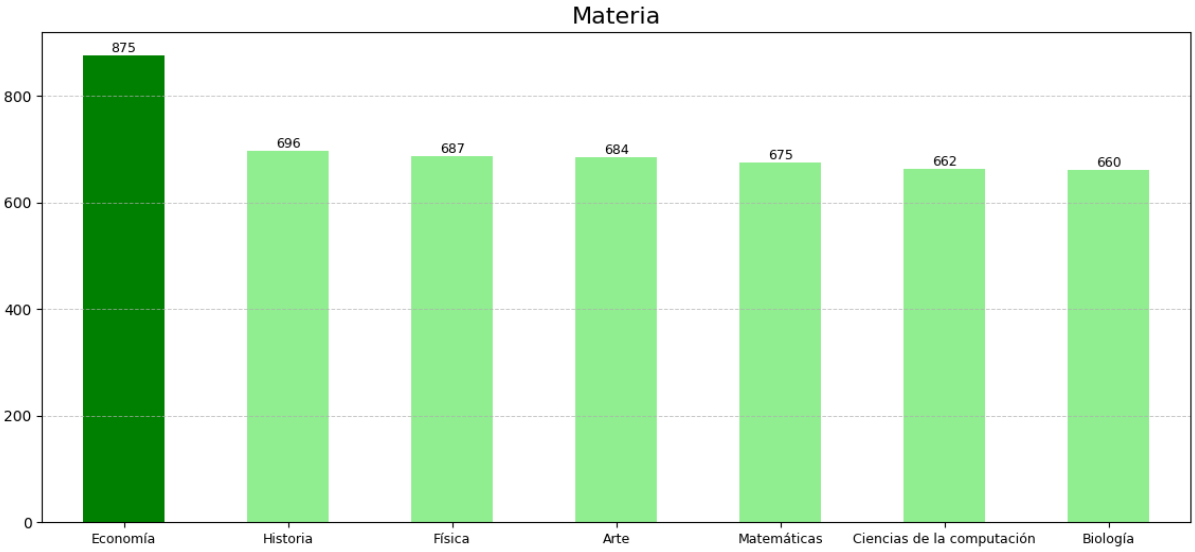

Uso R.V en su Educación
Si
52.4%
47.6%
No

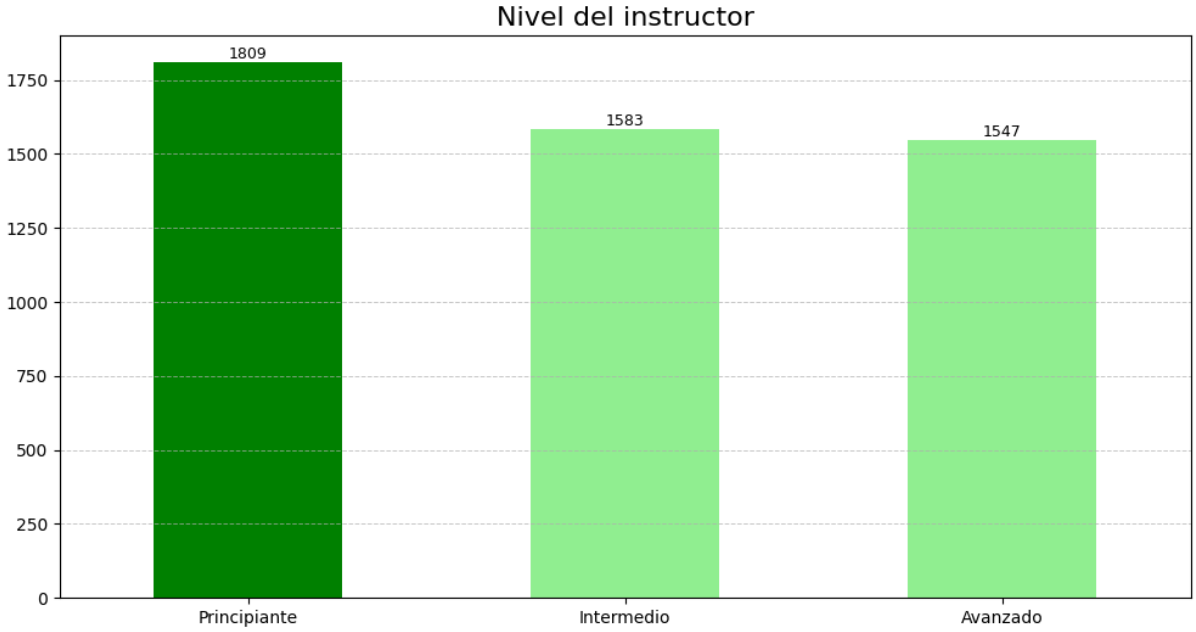

Histograma de Horas de R.V. semanalmente
600
500
400
300
200
100
0
0
2
4
6
8
10
Horas

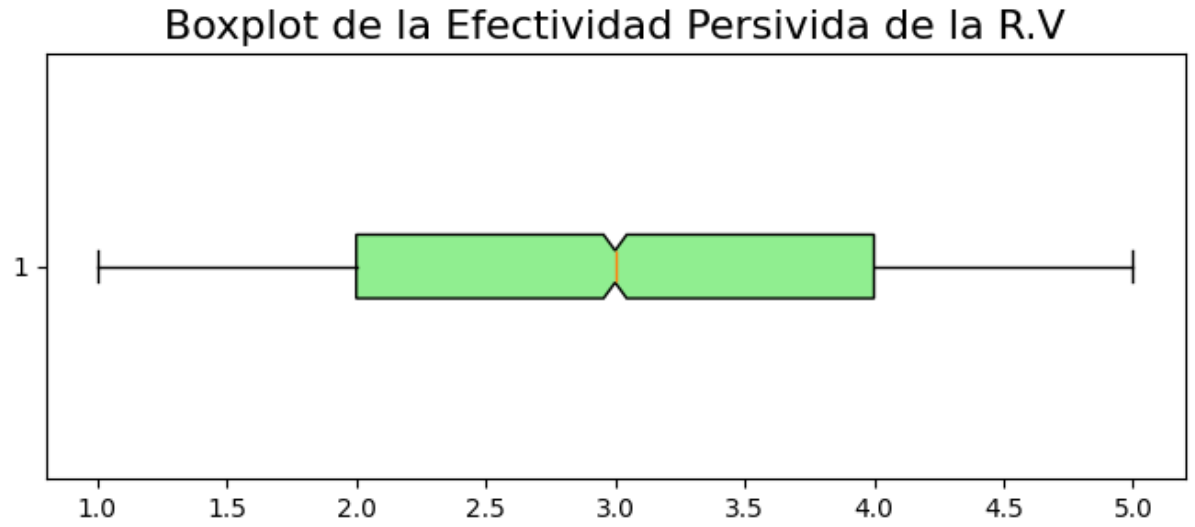

Boxplot del Nivel de Compromiso
1
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

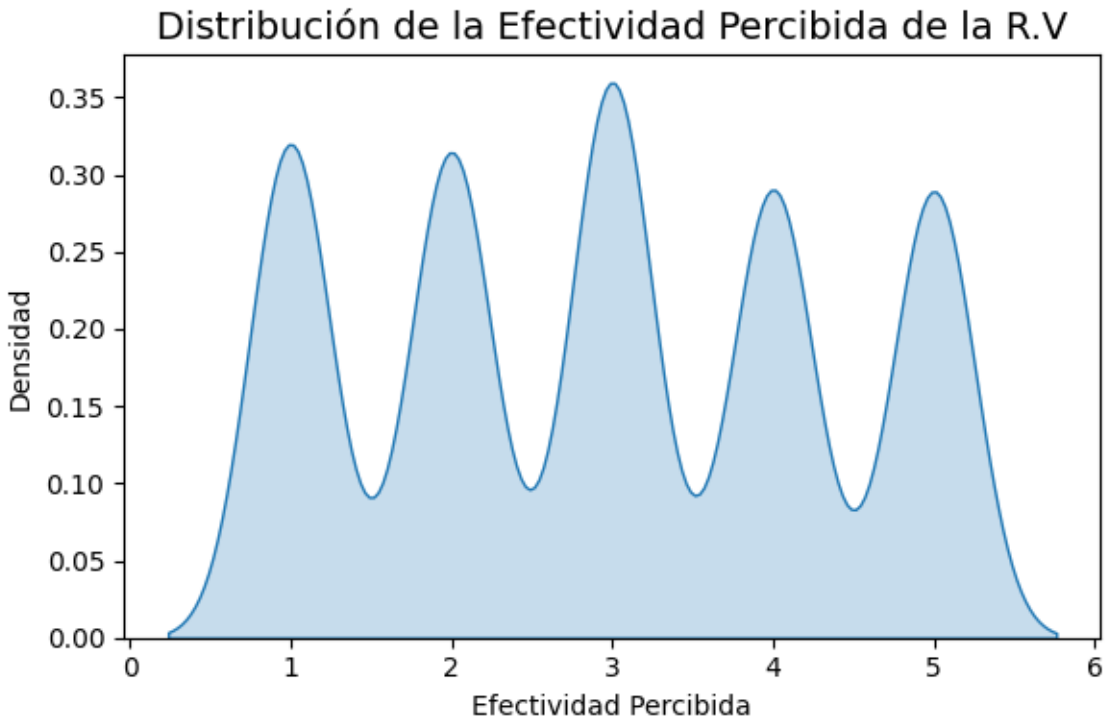

Distribución del Nivel de Compromiso
1
2
3
4
5
Nivel_de_Compromiso

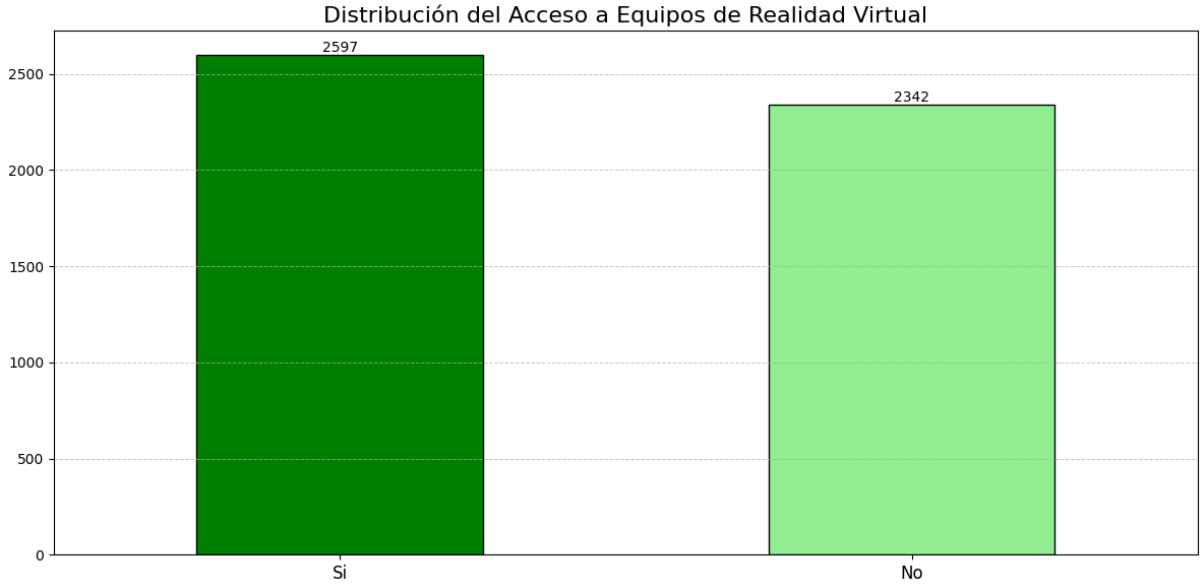

Mejora de resultados en el aprendizaje
Si
52.3%
47.7%
No

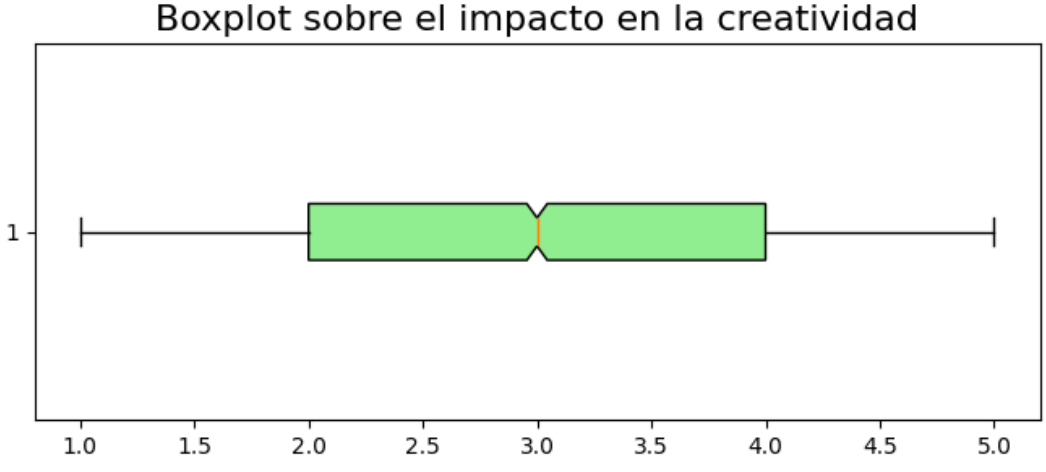

Materia
875
696
687
684
675
662
660
800
600
400
200
0
Economía
Historia
Física
Arte
Matemáticas
Ciencias de la computación
Biología

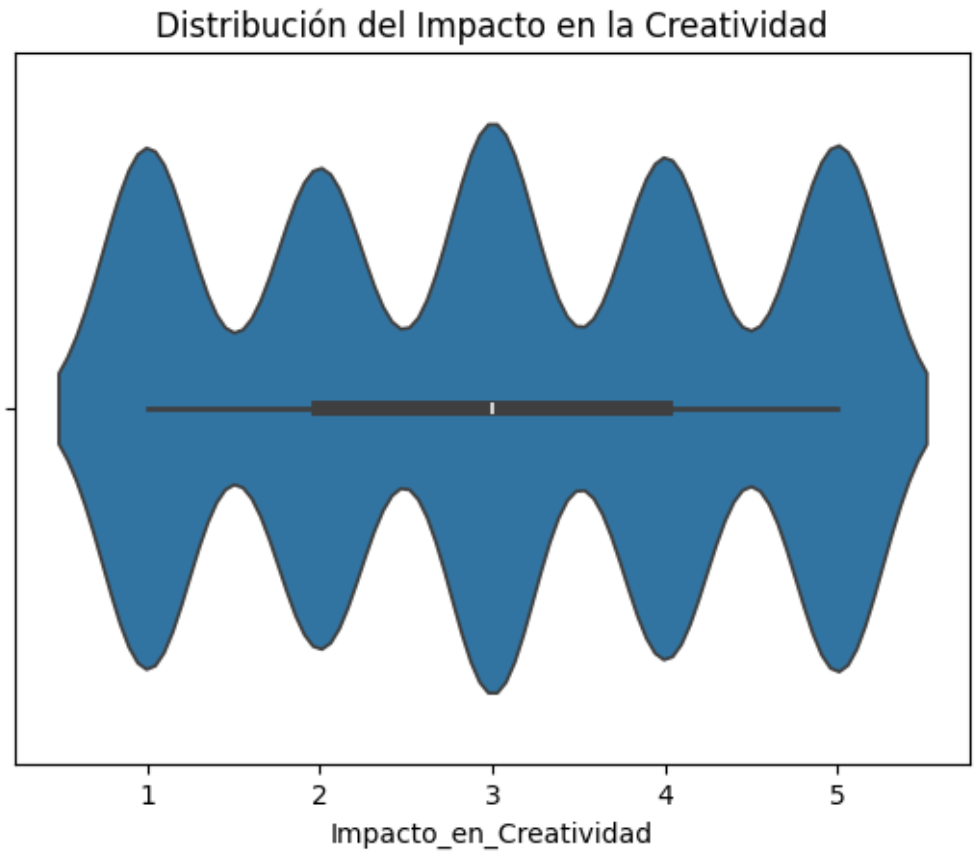

Nivel del instructor
1809
1583
1547
1750
1500
1250
1000
750
500
250
0
Principiante
Intermedio
Avanzado

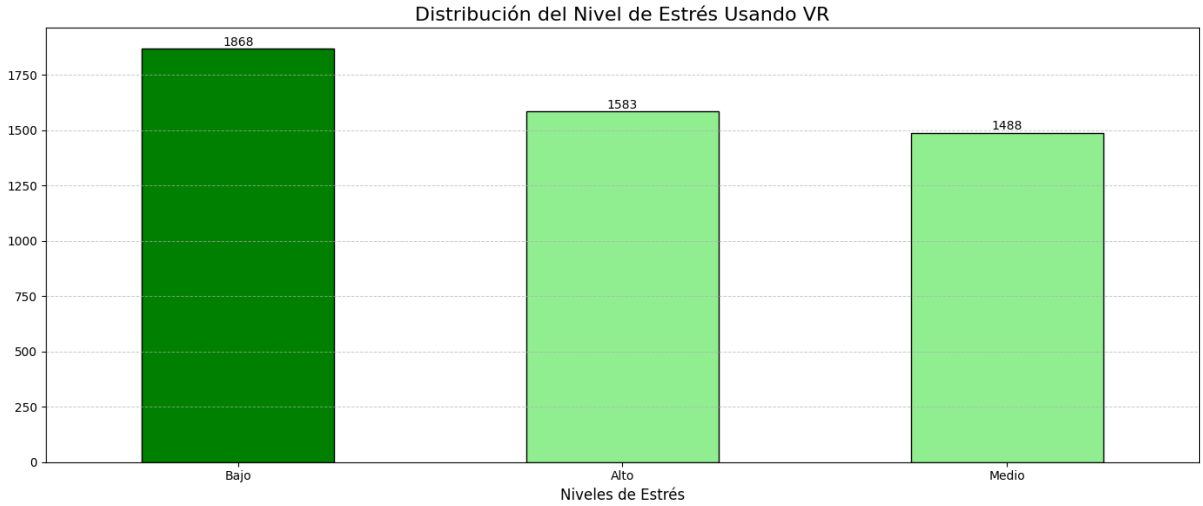

Boxplot de la Efectividad Persivida de la R.V
1
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

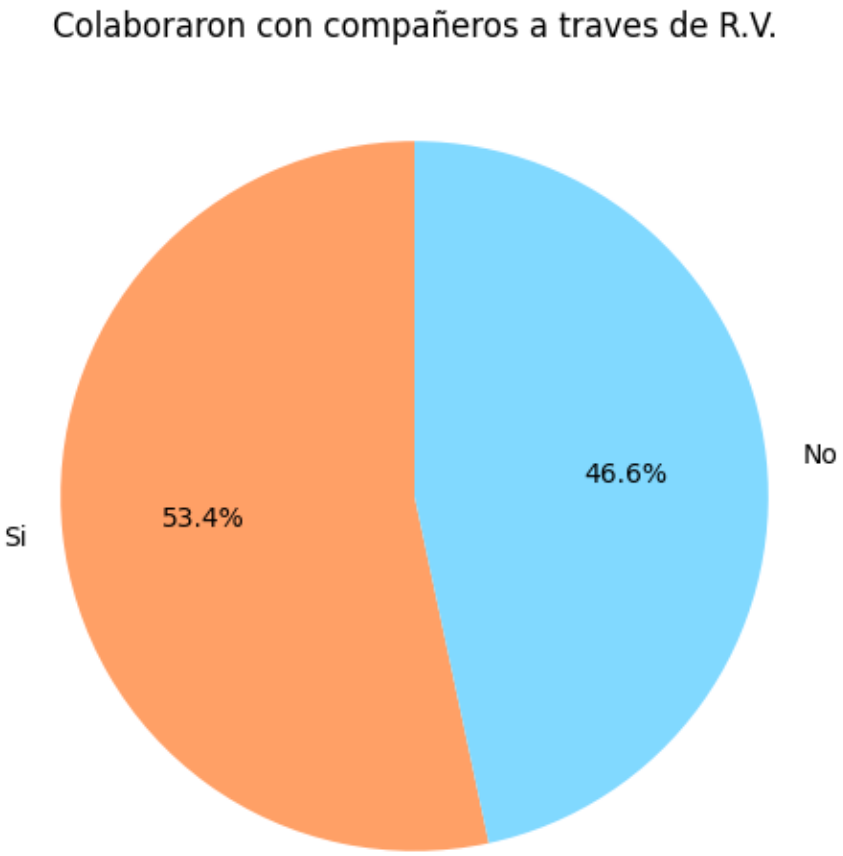

Distribución de la Efectividad Percibida de la R.V
Densidad
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
6
Efectividad Percibida

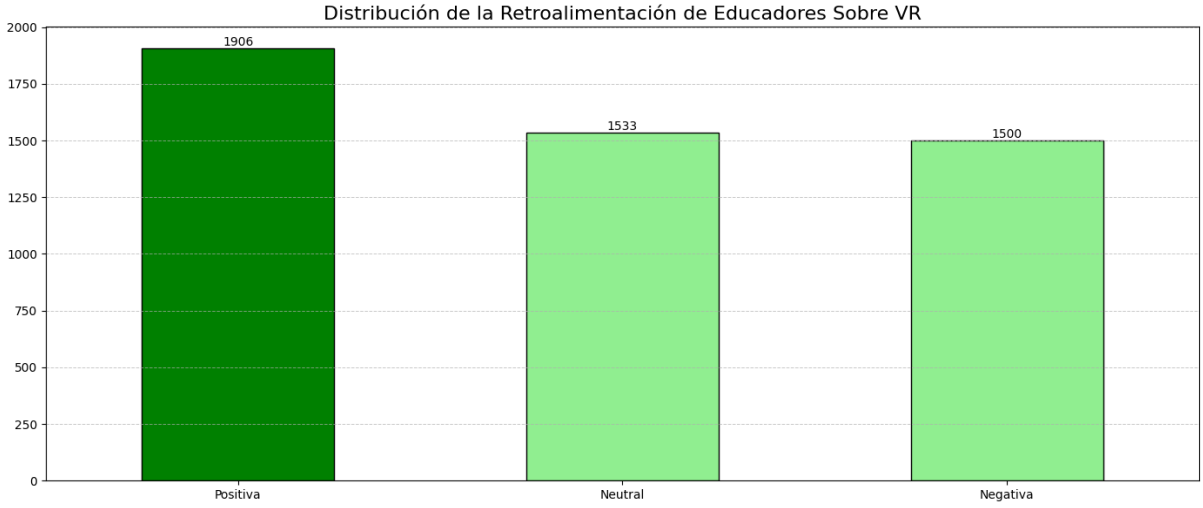

Distribución del Acceso a Equipos de Realidad Virtual
2597
2342
2500
2000
1500
1000
500
0
Si
No

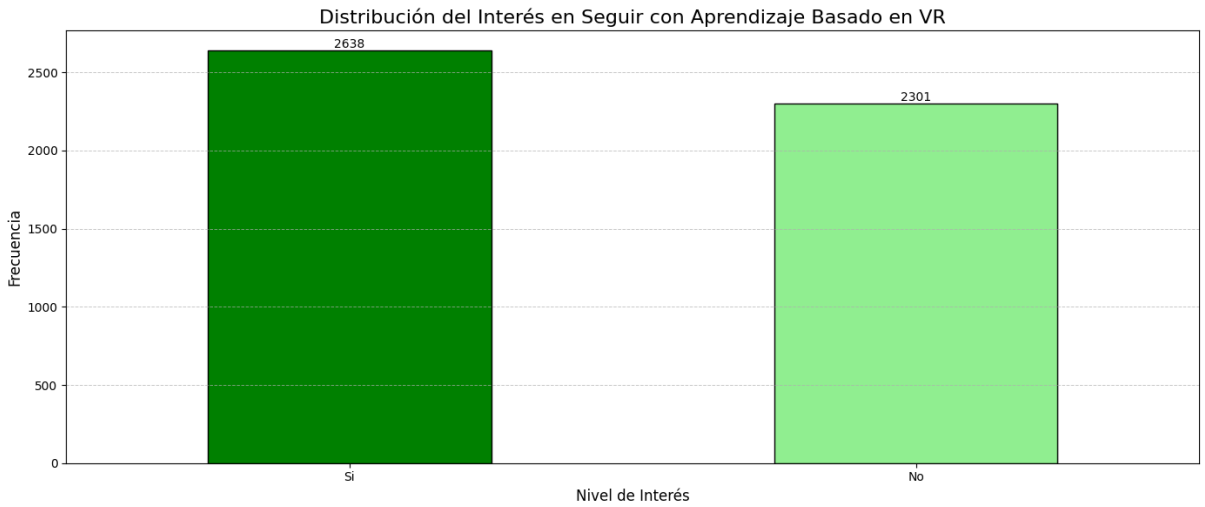

Boxplot sobre el impacto en la creatividad
1
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Distribución del Impacto en la Creatividad
1
2
3
4
5
Impacto_en_Creatividad

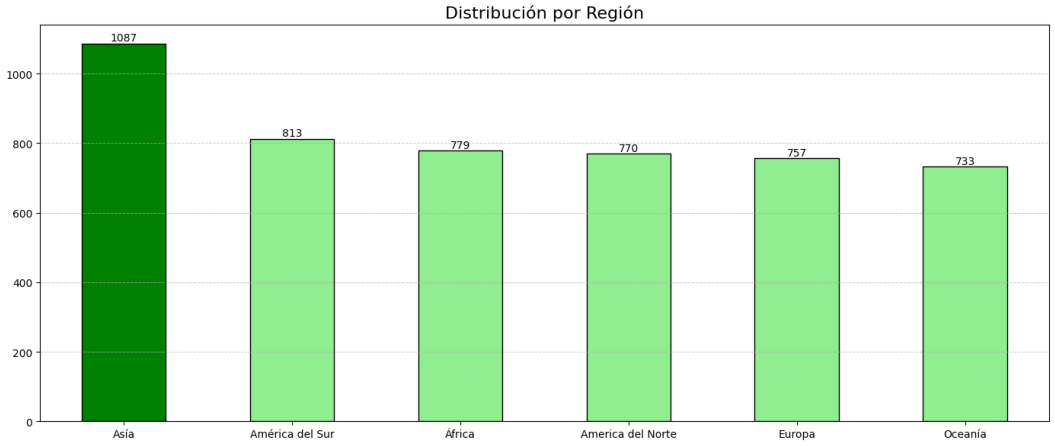

Distribución del Nivel de Estrés Usando VR
1868
1583
1488
1750
1500
1250
1000
750
500
250
0
Bajo
Alto
Medio
Niveles de Estrés

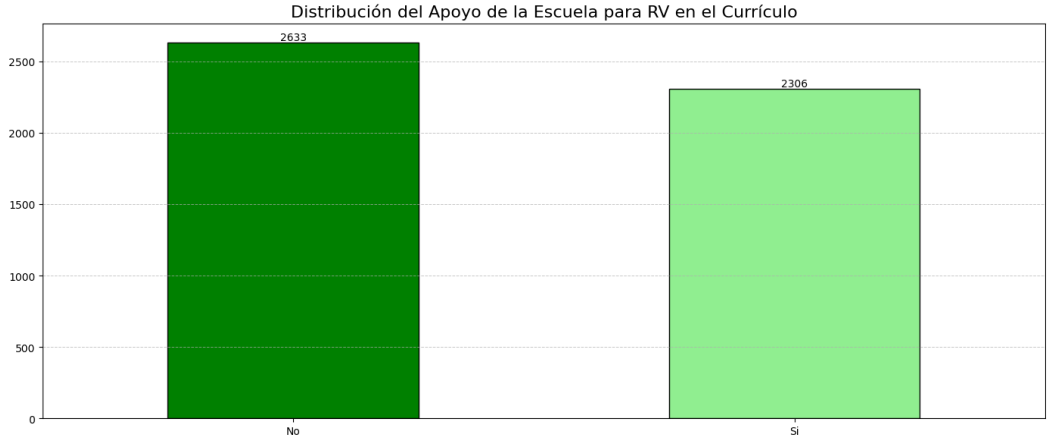

Colaboraron con compañeros a traves de R.V.
Si
53.4%
46.6%
No

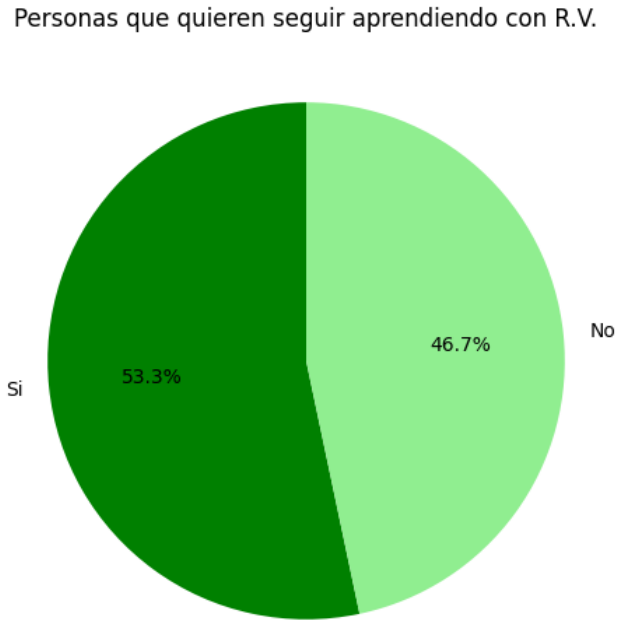

Distribución de la Retroalimentación de Educadores Sobre VR
2000
1750
1500
1250
1000
750
500
250
0
1906
1533
1500
Positiva
Neutral
Negativa


Distribución del Interés en Seguir con Aprendizaje Basado en VR
2638
2301
2500
2000
1500
1000
500
0
Frecuencia
Si
No
Nivel de Interés


Personas que quieren seguir aprendiendo con R.V.
53.4%
46.6%
Si
No


Distribución por Región
1087
813
779
770
757
733
1000
800
600
400
200
0
Asía
América del Sur
África
America del Norte
Europa
Oceanía


Distribución del Apoyo de la Escuela para RV en el Currículo
2633
2306
2500
2000
1500
1000
500
0
No
Si


Personas que quieren seguir aprendiendo con R.V.
46.7%
53.3%
No
Si

- Distribución de Género: Los datos muestran una distribución equitativa entre los cuatro géneros evaluados, asegurando una representación balanceada en el análisis.

- Uso de RV en Educación: Un 52.4% de las personas encuestadas reportaron haber utilizado RV en sus procesos educativos, mientras que el 47.6% no lo hicieron, indicando un uso casi equilibrado.

- Mejora en Resultados de Aprendizaje: De las personas que utilizaron RV, el 52.3% reportó mejoras en sus resultados académicos, evidenciando un impacto positivo en el aprendizaje.

- Nivel del Instructor en RV: Entre los instructores, 1809 tenían un nivel de competencia principiante, 1583 intermedio y 1547 avanzado, lo que refleja una distribución relativamente uniforme en la experiencia con RV.

- Acceso a Equipos de RV: El 52.5% (2597 personas) reportó tener acceso a equipos de RV, mientras que el 47.5% (2342 personas) no contaba con dicho acceso, destacando la necesidad de expandir la disponibilidad de estos recursos.

- Niveles de Estrés: Los niveles de estrés relacionados con el uso de RV también fueron evaluados, con 1583 personas reportando estrés alto, 1488 estrés intermedio, y 1883 estrés bajo, lo que sugiere que la mayoría experimenta un nivel de estrés manejable al usar esta tecnología.

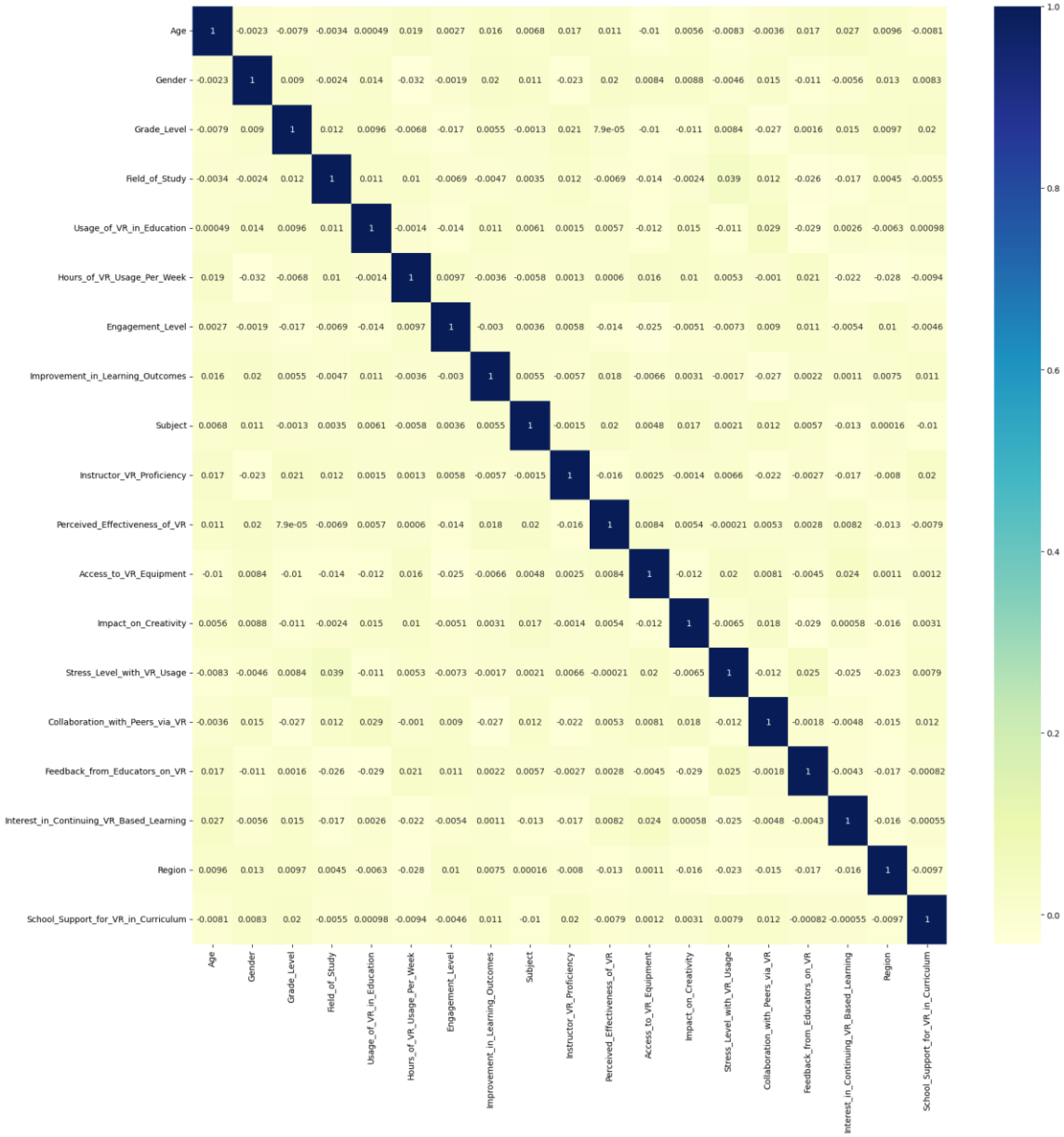
- Interés en Aprendizaje Continúo Basado en RV: El 53.4% de los encuestados mostró interés en continuar aprendiendo con RV, lo que subraya su percepción positiva y potencial para la adopción a largo plazo.

- Región Geográfica: El estudio evaluó al menos a 700 personas en cada una de las regiones consideradas, asegurando una diversidad geográfica significativa en los datos.

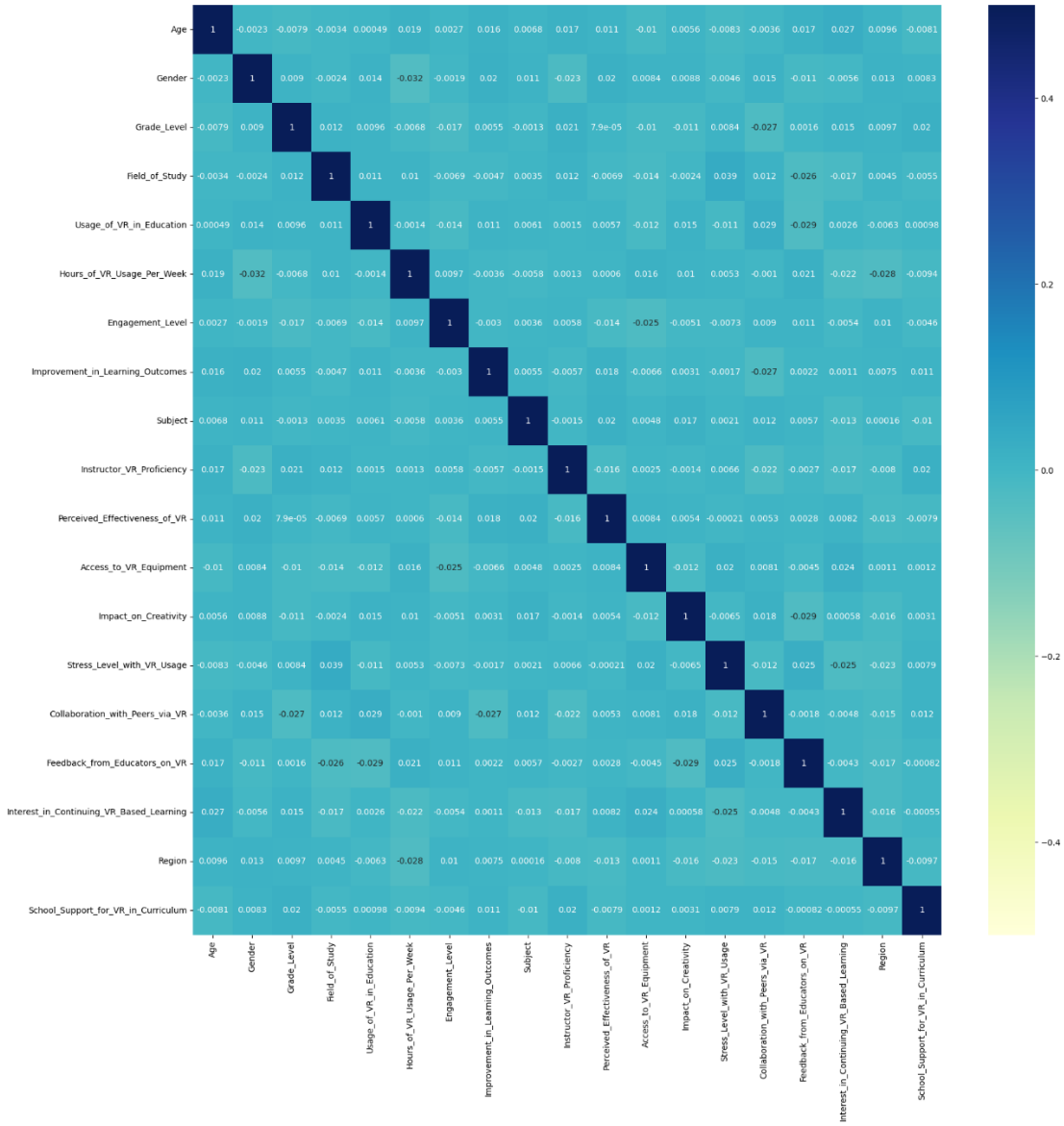
- Apoyo Escolar para la RV en el Currículo: Los resultados indican que el apoyo institucional para el aprendizaje basado en RV ha tenido un impacto favorable en la integración curricular, sugiriendo que este factor es clave para promover el uso de esta tecnología.

3. Correlación entre variables

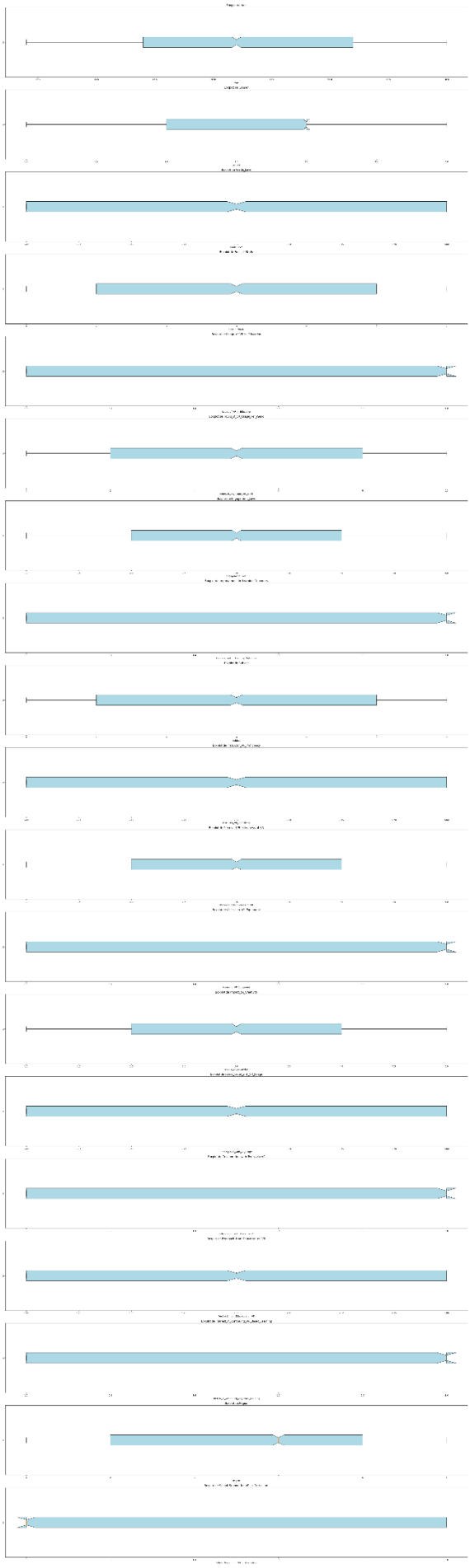
Mapa de calor



Bajamos nuestro rango para ver si podemos observar alguna relación



Sin embargo, no parece haber alguna correlación entre nuestras columnas



Además, como mencionamos anteriormente nuestros datos están distribuidos uniformemente, ya que en nuestros bloxsplots no tenemos puntos que nos indiquen valores atípicos.

Además de que no tenemos datos faltantes, ambos casos se deben al pretratamiento y limpieza de los datos.