

## ▼ Análisis, Visualización y Curación de Datos

### Consigna 3

Elegir (al menos) tres variables, las cuales crean que pueden estar correlacionadas con la satisfacción del estudiante al terminar el diálogo. Para cada una de ellas calcular la probabilidad de que el estudiante dé una evaluación negativa (1 o 2), condicionada a esa variable.

---

## ▼ Importación de librerías

```
1 # Cargo las librerías
2 import os
3 import pandas as pd
4 import numpy as np
5 import matplotlib.pyplot as plt
6 import seaborn
7
8 from columns_to_keep import C2K
9 from type_to_fix import T2F
```

## ▼ Carga de la tabla de metadata

---

```
1 data_dir = os.path.join('.', 'dataset')
2 data_file_name = 'datadump-20150801-20171219.csv'
3 full_data_file_name = os.path.join(data_dir, data_file_name)
4 df = pd.read_csv(full_data_file_name)
5 display(df.shape)
6
```

```
1 df = pd.read_csv('datadump-20150801-20171219.csv')
2 display(df.shape)
```



```
/usr/local/lib/python3.6/dist-packages/IPython/core/interactiveshell.py:2718: DtypeWarning: Col
interactivity=interactivity, compiler=compiler, result=result)
(63265, 111)
```

```
1 # Remuevo espacios al inicio y final del nombre de la columna
2 df.columns = [column.strip() for column in df.columns]
3
4 # Selecciono solo las columnas que vamos a necesitar
5 df = df[C2K]
6
7 # Se corrigen los tipos de los datos
8 df = df.astype(T2F)
9
10 print("El dataframe tiene", df.shape[0], "filas y ", df.shape[1], "columnas.")
```



El dataframe tiene 63265 filas y 39 columnas.

## ▼ Pre-procesamiento de los datos

---

### Pasos:

1. Selección de las variables que se analizarán para determinar si las mismas tienen una correlación positiva con la variable **student\_rating**, con el fin de determinar si dichas variables **wait\_time**, **avg\_tutor\_response\_time** y **feedback\_score** influyen sobre el puntaje asignado al tutor (student\_rating).
2. Eliminación de los valores NA de las variables seleccionadas.
3. Convierto las variables wait\_time y avg\_tutor\_response\_time de segundos a minutos.
4. Se grafica la relación entre student\_rating y las variables elegidas.
5. Cálculo del coeficiente de correlación de Spearman.

### 1. Las variables seleccionadas para analizar la correlación son:

- **wait\_time**: cuánto tiempo esperó el estudiante antes de ser emparejado con un tutor expresado en segundos ((escala entre 1 y 5).
- **avg\_tutor\_response\_time**: tiempo promedio en segundos desde la presentación del estudiante hasta el enunciado del tutor.
- **feedback\_score**: puntaje asignado a la sesión (escala entre 0 y 1).

```
1 # 1 y 2.
2 df_sinNA = df.dropna(axis=0, subset=['student_rating', 'wait_time', 'avg_tutor_response_time',
3
4 df_graficos = df_sinNA.loc[:, ['student_rating', 'wait_time', 'avg_tutor_response_time', 'feedback_score']]
5 display(df_graficos)
6 df_graficos.shape
7
8 print ("Luego de la limpieza, obtengo un dataframe de", df_sinNA.shape[0], "filas y", df_sinNA.shape[1], "columnas")
```



student\_rating wait\_time avg\_tutor\_response\_time feedback\_score

```

1 # 3. convierto de secs a minutos
2 df_graficos['wait_time'] = (df_sinNA['wait_time'] / 60)
3 df_graficos['wait_time'] = round(df_graficos['wait_time'],1)
4
5 df_graficos['avg_tutor_response_time'] = (df_sinNA['avg_tutor_response_time'] / 60)
6 df_graficos['avg_tutor_response_time'] = round(df_graficos['avg_tutor_response_time'],1)
7 display(df_graficos)

```



	student_rating	wait_time	avg_tutor_response_time	feedback_score
0	4	0.8	0.6	0.4
2	1	0.2	0.4	0.0
4	5	0.1	0.3	0.8
5	5	0.1	0.2	1.0
7	5	0.1	0.3	0.9
...	...	...	...	...
63260	5	0.3	0.4	0.7
63261	5	0.2	0.2	1.0
63262	5	0.1	0.2	1.0
63263	5	0.1	0.3	1.0
63264	5	0.2	0.3	1.0

44907 rows × 4 columns

```

1 # 3. convierto de secs a minutos
2 #df_graficos['wait_time'] = (df_sinNA['wait_time'].astype('timedelta64[m]'))
3
4 # convierto la variable avg_tutor_response_time de secs a minutos
5 #df_graficos['avg_tutor_response_time'] = (df_sinNA['avg_tutor_response_time'].astype('timedelta64[m]'))
6 #display(df_graficos)
7

```



	student_rating	wait_time	avg_tutor_response_time	feedback_score
0	4	00:45:00	00:37:00	0.4
2	1	00:10:00	00:24:00	0.0
4	5	00:07:00	00:16:00	0.8
5	5	00:04:00	00:11:00	1.0
7	5	00:05:00	00:20:00	0.9
...	...	...	...	...
63260	5	00:17:00	00:25:00	0.7
63261	5	00:10:00	00:12:00	1.0
63262	5	00:07:00	00:11:00	1.0
63263	5	00:06:00	00:17:00	1.0
63264	5	00:11:00	00:16:00	1.0

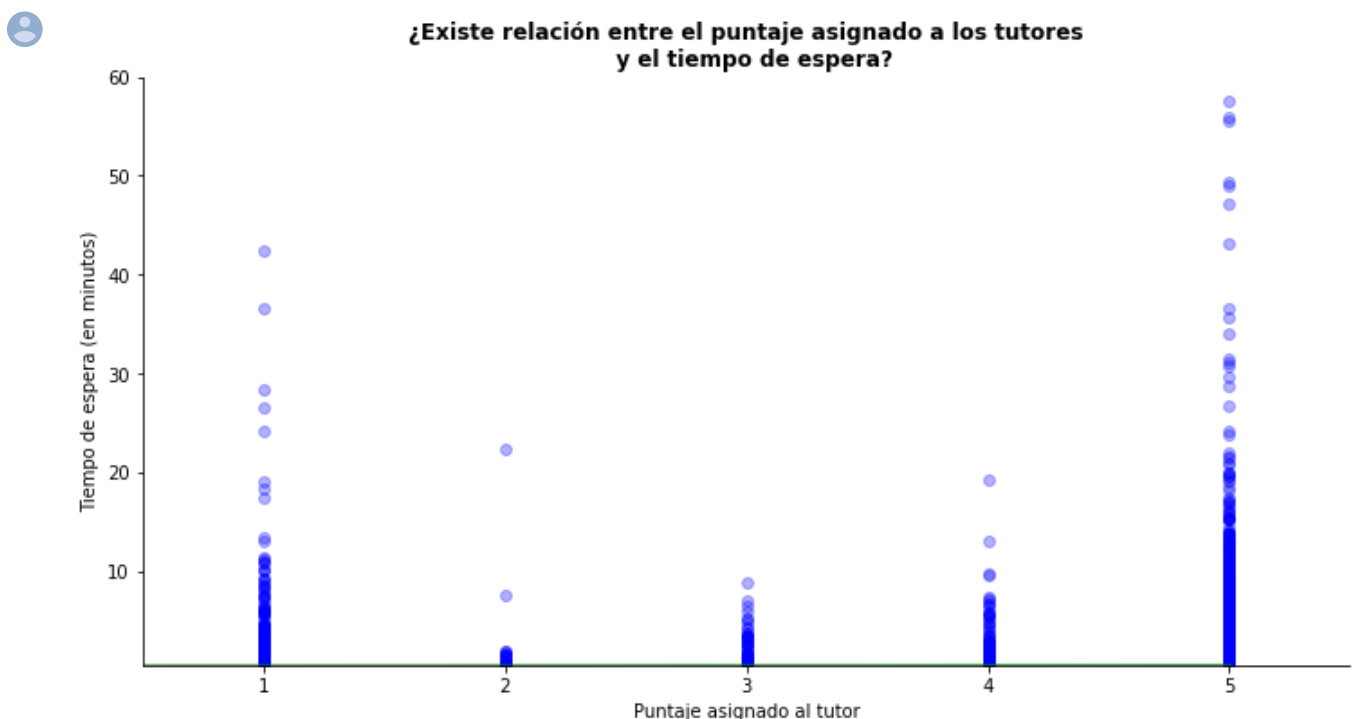
44907 rows × 4 columns

## 4. ¿Existe alguna variable que influya sobre la satisfacción de los estudiantes?

La visualización de datos es una parte fundamental de cualquier análisis de datos, ya que no sólo se utiliza para comunicar los hallazgos o resultados finales sino que también son de suma importancia en la etapa de exploración de datos para confirmar o rechazar hipótesis planteadas acerca de los datos utilizados. Entonces para verificar si existe alguna variable que influya en la satisfacción de los estudiantes por un lado, se utilizarán gráficos y por el otro, se calculará el coeficiente de correlación de Spearman. A continuación utilizaremos tres tipos distintos de gráficos, primero un regplot o gráfico de regresión lineal, luego un scatterplot o gráfico de dispersión y por último un pairplot. Cada uno nos permitirá observar si existe relación entre `student_rating` y las variables elegidas `wait_time`, `avg_tutor_response_time` y `feedback_score`.

### ▼ Gráfico de regresión lineal entre las variables `student_rating` y `wait_time`

```
1 # 4. Gráfico de regresión lineal entre student_rating y wait_time
2 #df_graficos['wait_time'] = (df_sinNA['wait_time'] / 60)
3
4 plt.figure(figsize=(12,6))
5 seaborn.regplot(x=df_graficos.student_rating.astype(float), y=df_graficos.wait_time.astype(float)
6 plt.title('¿Existe relación entre el puntaje asignado a los tutores\n y el tiempo de espera?',
7           fontsize=12, weight="bold")
8 plt.xlabel('Puntaje asignado al tutor')
9 plt.ylabel('Tiempo de espera (en minutos)')
10 plt.ylim([0.5, 60])
11 plt.xlim([0.5, 5.5])
12 seaborn.despine()
13 plt.show()
```

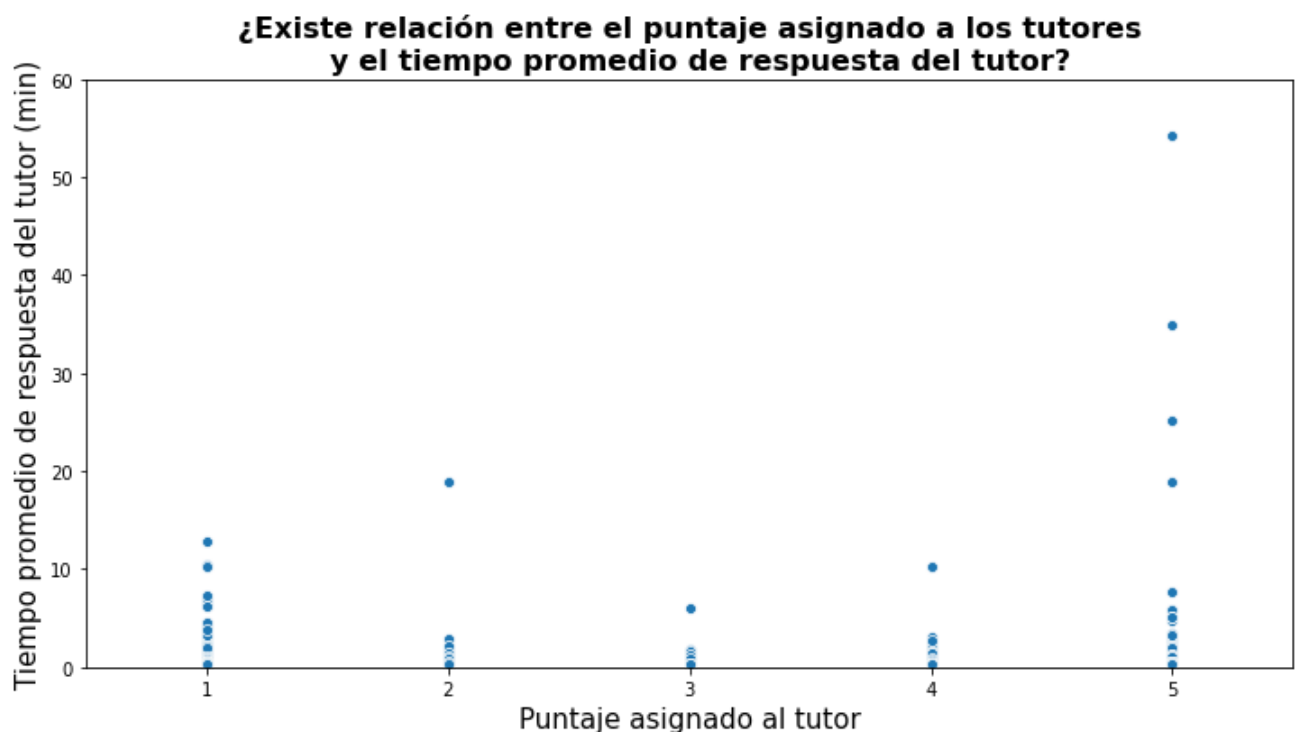


Luego de observar el gráfico obtenido, se observan puntos dispersos y una recta que no aproxima dichos valores, por lo que se evidencia que no existe una relación entre ambas variables, por lo tanto la variable tiempo de espera no influiría sobre el comportamiento de la variable puntaje asignado a los tutores. El cálculo del coeficiente de spearman en el punto 5 permitirá confirmar dicha afirmación.

## ▼ Scatterplot entre las variables student\_rating y avg\_tutor\_response\_time

El gráfico de dispersión o scatterplot permite verificar si existe o no correlación entre las variables student\_rating (puntaje asignado a los tutores) y avg\_tutor\_response\_time (tiempo promedio de respuesta del tutor).

```
1 plt.figure(figsize=(12,6))
2 #df_graficos['avg_tutor_response_time'] = (df_sinNA['avg_tutor_response_time'] / 60)
3
4 seaborn.scatterplot(x=df_graficos.student_rating.astype(float), y=df_graficos.avg_tutor_response_time)
5 plt.title('¿Existe relación entre el puntaje asignado a los tutores\n y el tiempo promedio de respuesta del tutor?')
6 plt.xlabel('Puntaje asignado al tutor', fontsize=15)
7 plt.ylabel('Tiempo promedio de respuesta del tutor (min)', fontsize=15)
8 plt.xlim([0.5, 5.5])
9 plt.ylim([0, 60])
10 plt.show()
```

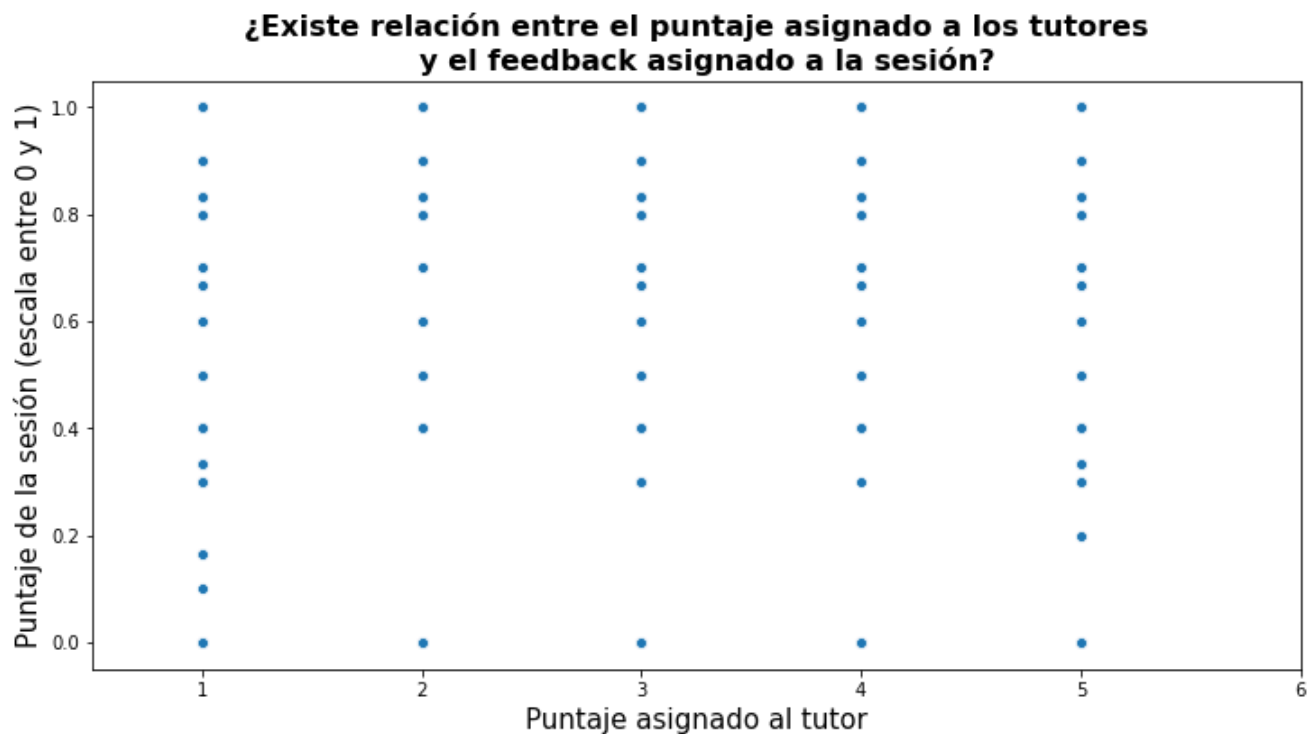


Luego de observar el gráfico obtenido, se aprecian puntos dispersos que no siguen ningún patrón y se evidencia que no existe correlación entre ambas variables, esto además se verificará en el punto 5 con el cálculo del coeficiente de Spearman.

## ▼ Scatterplot entre las variables student\_rating y feedback\_score

El gráfico de dispersión o scatterplot permite verificar si existe o no correlación entre las variables `student_rating` (puntaje asignado a los tutores) y `feedback_score` (feedback de la sesión).

```
1 plt.figure(figsize=(12,6))
2 seaborn.scatterplot(x=df_graficos.student_rating, y=df_graficos.feedback_score)
3 plt.title('¿Existe relación entre el puntaje asignado a los tutores\n y el feedback asignado a
4 plt.xlabel('Puntaje asignado al tutor', fontsize=15)
5 plt.ylabel('Puntaje de la sesión (escala entre 0 y 1)', fontsize=15)
6 plt.xlim([0.5, 6])
7 plt.show()
```

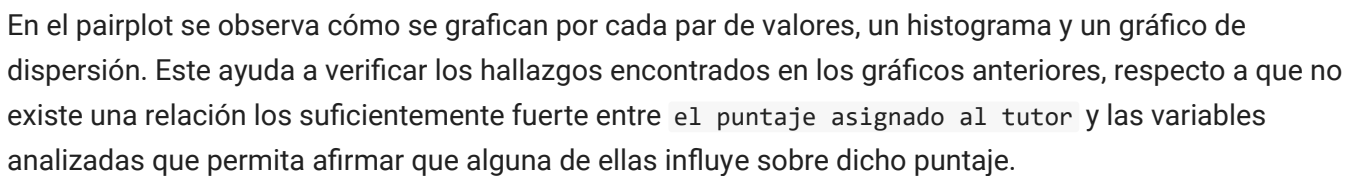
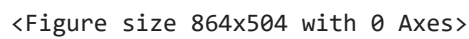


Luego de observar el gráfico obtenido, se evidencia que no existe correlación entre el puntaje asignado al tutor y el feedback de la sesión, esto además se verificará en el punto 5 con el cálculo del coeficiente de Spearman.

## Gráfico pairplot entre `student_rating` y las variables `wait_time`, `avg_tutor_response_time` y `feedback_score`.

Por último, se utilizará un gráfico del tipo pairplot, el cual permitirá observar de manera rápida las relaciones entre las variables del dataframe `df_pairplot`, descrito a continuación.

```
1 df_pairplot = df_graficos.loc[:, ['student_rating', 'wait_time', 'avg_tutor_response_time', 'fe
2 plt.figure(figsize=(12,7))
3 seaborn.pairplot(data= df_pairplot)
4 #plt.title('¿Existe relación entre el puntaje asignado al tutor\n, el tiempo esperado (min), el
5 #plt.xlim([0.5, 1000])
6 plt.ylim([0.5, 180])
7 plt.show()
```



▼

```
1 from scipy.stats import spearmanr
```

```
1 dfspearman = df[["student_rating", "wait_time", "feedback_score", "length_of_session", "avg_tutor_rating"]]
```

## ▼ ¿Existe correlación entre las variables `student_rating` y `wait_time`?


```
1 spearman_p2 = spearmanr(dfspearman['student_rating'],dfspearman['wait_time'])
2 display(spearman_p2)
```

 SpearmanrResult(correlation=-0.005554422719821421, pvalue=0.23955038005708534)

Con este resultado se confirma lo observado en el gráfico de regresión lineal, es decir, que no existe influencia de la variable `tiempo de espera` sobre el comportamiento del `puntaje asignado a los tutores`.

## ▼ ¿Existe correlación entre las variables `student_rating` y `feedback_score`?


```
1 spearman_p = spearmanr(dfspearman['student_rating'],dfspearman['feedback_score'])
2 display(spearman_p)
```

 SpearmanrResult(correlation=0.28743640456627073, pvalue=0.0)

El valor del coeficiente de correlación es bastante bajo, por lo que con este resultado se confirma lo observado en el gráfico de dispersión, es decir, que no existe influencia de la variable `feedback de la sesión` sobre el `puntaje asignado a los tutores`.

## ▼ ¿Existe correlación entre las variables `student_rating` y `avg_tutor_response_time`?

```
1 spearman_p4 = spearmanr(dfspearman['student_rating'],dfspearman['avg_tutor_response_time'])
2 display(spearman_p4)
```

 SpearmanrResult(correlation=-0.08054335545485349, pvalue=2.016065258359058e-65)

Con el anterior resultado se confirma lo observado en el gráfico de dispersión, es decir, que no existe influencia de la variable `tiempo promedio de respuesta del tutor` sobre el `puntaje asignado a los tutores`.

## Conclusión.

Luego de verificar si existía alguna correlación entre `student_rating` y las variables `wait_time`, `avg_tutor_time_response` y `feedback_score`, mediante la visualización de dichas variables con distintos gráficos y el cálculo del coeficiente de Spearman, se puede afirmar que no se encontró evidencia que permita establecer alguna relación lo suficientemente fuerte entre el `puntaje asignado a los tutores` y las variables seleccionadas, que indique influencia de algunas de ellas sobre el comportamiento de la variable `satisfacción del estudiante`.



