# 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

```
# Cargo las librerías
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn

from columns_to_keep import C2K
from type_to_fix import T2F
```

## Carga de la tabla de metadata

---

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

```

```
df = pd.read_csv('datadump.csv')
display(df.shape)
```

```
# Remuevo espacios al inicio y final del nombre de la columna
df.columns = [column.strip() for column in df.columns]

# Selecciono solo las columnas que vamos a necesitar
df = df[C2K]

# Se corrigen los tipos de los datos
df = df.astype(T2F)

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 variables: las variables seleccionadas para analizar la correlación con student_rating, son wait_time, length_of_session, avg_tutor_response_time y feedback_score.
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 y 2.
df_sinNA = df.dropna(subset=['student_rating', 'wait_time', 'length_of_session', 'avg_tutor_resp

df_graficos = df_sinNA.loc[:, ['student_rating', 'wait_time', 'length_of_session', 'avg_tutor_re
df_graficos.shape

print ("Luego de la limpieza, obtengo un dataframe de", df_sinNA.shape[0],"filas y", df_sinNA.sh
```

```
Luego de la limpieza, obtengo un dataframe de 44837 filas y 39 col.
```

```
# 3. convierto de secs a minutos
df_graficos['wait_time'] = (df_sinNA['wait_time'].astype('timedelta64[m]'))
#df_sinNA['wait_time'] = (df_sinNA['wait_time'] / 60)
#df_sinNA['wait_time'] = round(df_sinNA['wait_time'],1)

# convierto  la variable length_of_session de secs a minutos
df_graficos['length_of_session'] = (df_sinNA['length_of_session'].astype('timedelta64[m]'))

# convierto  la variable avg_tutor_response_time de secs a minutos
df_graficos['avg_tutor_response_time'] = (df_sinNA['avg_tutor_response_time'].astype('timedelta6
display(df_graficos)

```

| | student_rating | wait_time | length_of_session | avg_tutor_response_time | feedback_score |
|---|---|---|---|---|---|
| **0** | 4 | 00:45:00 | 0 days 08:41:00 | 00:37:00 | 0.4 |
| **2** | 1 | 00:10:00 | 0 days 04:20:00 | 00:24:00 | 0.0 |
| **4** | 5 | 00:07:00 | 0 days 12:50:00 | 00:16:00 | 0.8 |
| **5** | 5 | 00:04:00 | 2 days 09:39:00 | 00:11:00 | 1.0 |
| **7** | 5 | 00:05:00 | 0 days 15:34:00 | 00:20:00 | 0.9 |
| **...** | ... | ... | ... | ... | ... |
| **63260** | 5 | 00:17:00 | 0 days 19:09:00 | 00:25:00 | 0.7 |
| **63261** | 5 | 00:10:00 | 1 days 02:52:00 | 00:12:00 | 1.0 |
| **63262** | 5 | 00:07:00 | 4 days 06:15:00 | 00:11:00 | 1.0 |
| **63263** | 5 | 00:06:00 | 0 days 14:58:00 | 00:17:00 | 1.0 |
| **63264** | 5 | 00:11:00 | 0 days 17:47:00 | 00:16:00 | 1.0 |

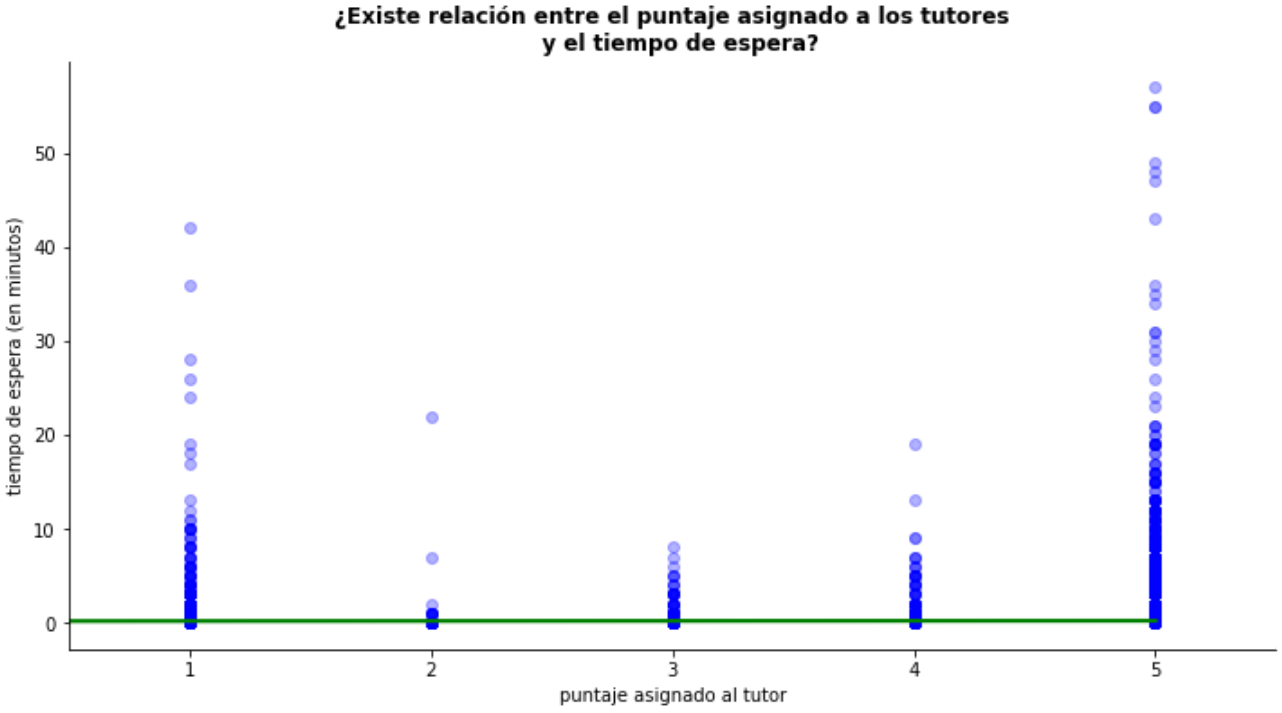
44837 rows × 5 columns

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

Para verificar si existe alguna variable que influya sobre la satisfacción de los estudiantes por un lado, se graficará la relación entre student_rating y las variables elegidas anteriormente y por otro, se calculará el coeficiente de correlación de Spearman.

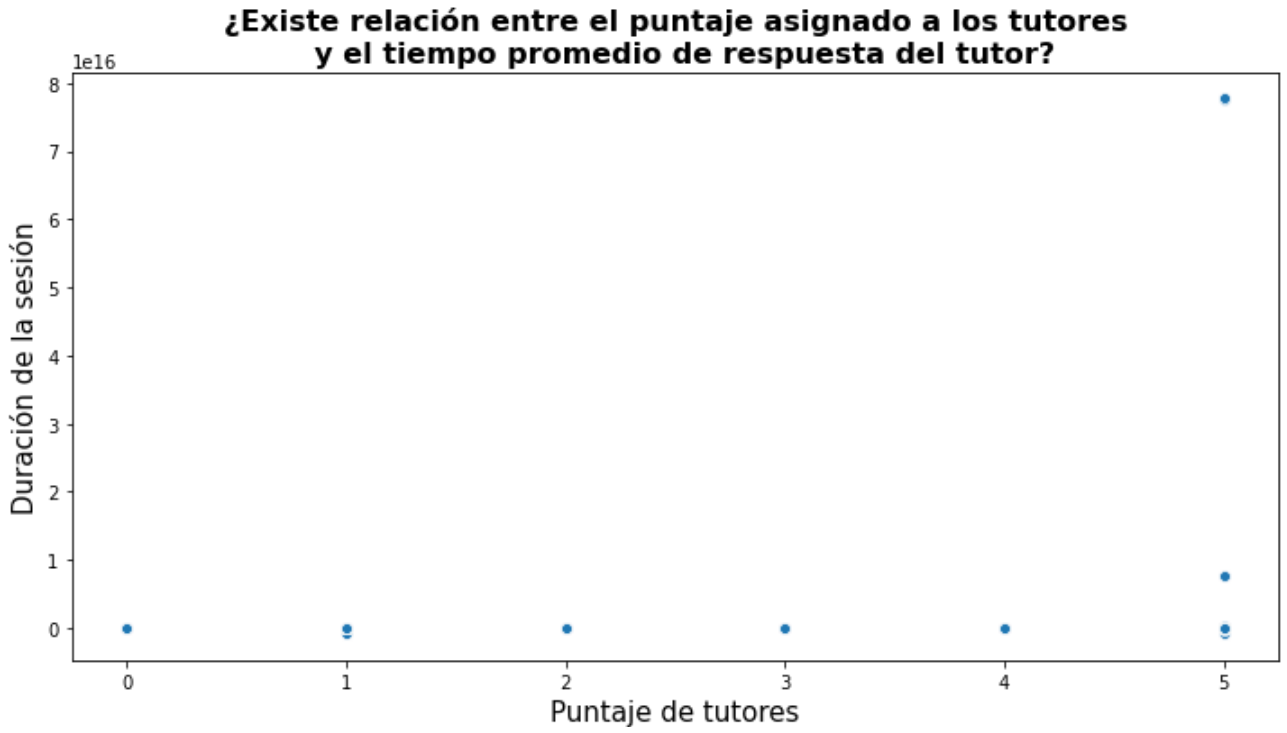
### Gráfico de regresión lineal entre student_rating y wait_time

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



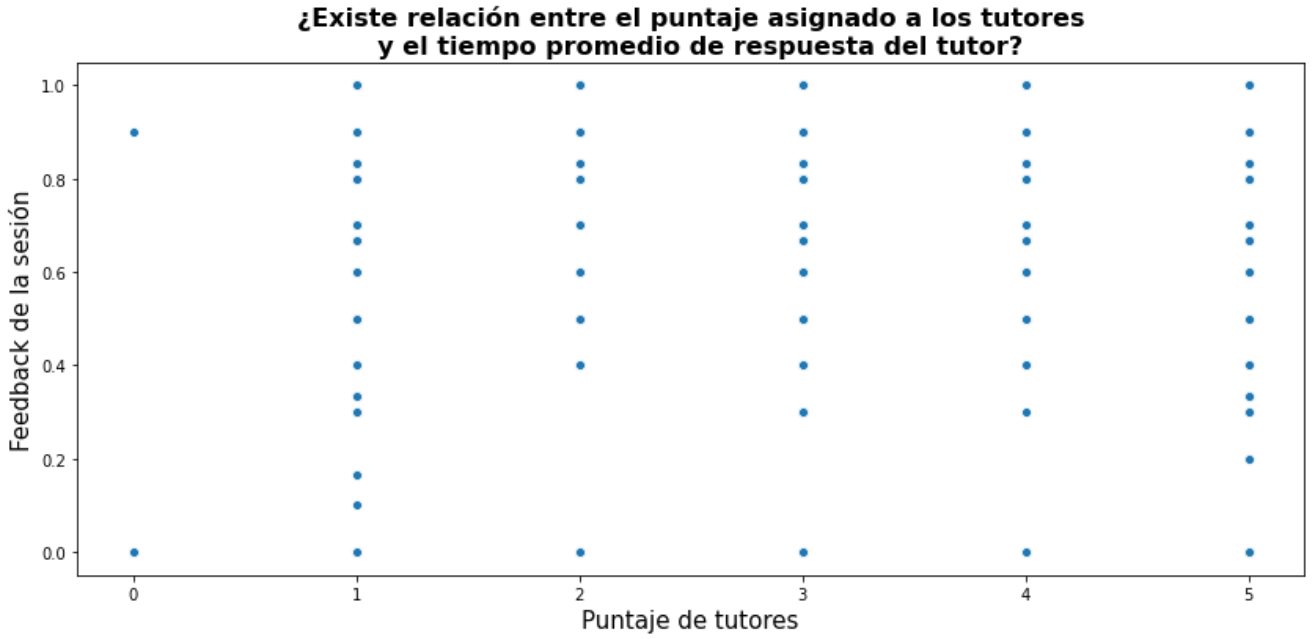

## Scatterplot entre el student_rating y el avg_tutor_response_time

```
plt.figure(figsize=(12,6))  #df_graficos.wait_time.astype(float)
seaborn.scatterplot(x=df_graficos.student_rating, y=df_graficos.avg_tutor_response_time.astype(
plt.title('¿Existe relación entre el puntaje asignado a los tutores\n  y el tiempo promedio de
plt.xlabel('Puntaje de tutores', fontsize=15)
plt.ylabel('Duración de la sesión', fontsize=15)
plt.show()
```




## Scatterplot entre el student_rating y el feedback_score

```
plt.figure(figsize=(14,6))  #df_graficos.wait_time.astype(float)
seaborn.scatterplot(data= df_graficos, x="student_rating", y="feedback_score") #x=df_graficos.st
plt.title('¿Existe relación entre el puntaje asignado a los tutores\n  y el tiempo promedio de r
plt.xlabel('Puntaje de tutores', fontsize=15)
plt.ylabel('Feedback de la sesión', fontsize=15)
plt.show()
```




## Gráfico pairplot entre student_rating y las variables elegidas

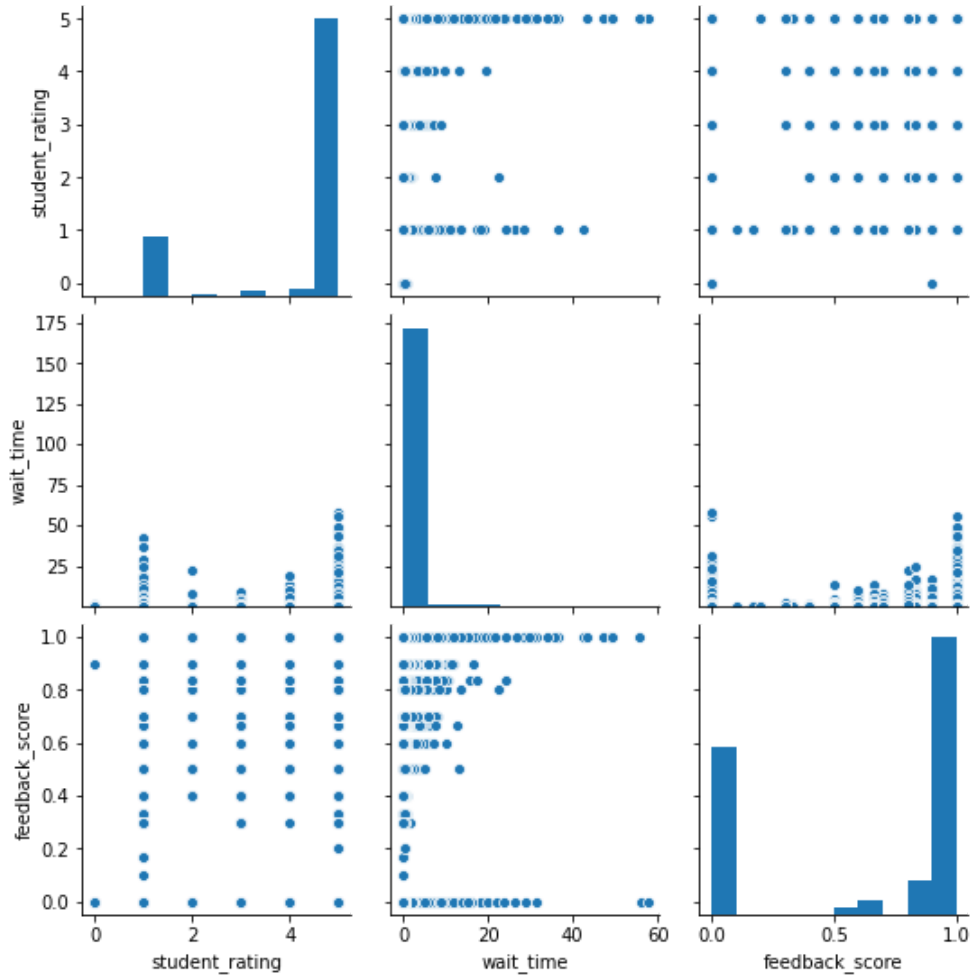
```

```

```
df_pairplot = df_graficos.loc[:, ["student_rating", "wait_time", "feedback_score", "avg_tutor_re
#plt.title('¿Existe relación entre el puntaje asignado a los profesores\n, el tiempo esperado, l
plt.figure(figsize=(12,7))

seaborn.pairplot(data= df_pairplot)
plt.ylim([0.5, 180])
plt.show()
```

```
<Figure size 864x504 with 0 Axes>
```




## 5.Cálculo de coeficientes de correlación

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

```
dfspearman = df[["student_rating", "wait_time", "feedback_score", "length_of_session", "avg_tuto

```

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

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

```
SpearmanrResult(correlation=-0.005554422719821421, pvalue=0.23955038005708534)
```

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

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

```
SpearmanrResult(correlation=0.28743640456627073, pvalue=0.0)
```

## ¿Existe correlación entre las variables student_rating y length_of_session?

```
spearman_p3 = spearmanr(dfspearman['student_rating'],dfspearman['length_of_session'])
display(spearman_p3)
```

```
SpearmanrResult(correlation=0.2570648912611863, pvalue=0.0)
```

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

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

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

### Conclusión:

Luego de verificar si existía alguna correlación entre student_rating y las variables wait_time, lenght_of_session y avg_tutor_time_response, mediante los gráficos anteriores y el coeficiente de Spearman, se llega a la conclusión de que no existe ninguna relación lo suficientemente fuerte entre cada par de variable, por lo tanto no se puede afirmar que alguna de ellas influya de manera significativa sobre la satisfacción del estudiante.