



Trabajo Final Integrador:
Análisis Descriptivo de Bases de Datos

Materia:
Analítica Descriptiva 82.04

Docentes:
Gómez Aguirre, Mauricio Maximiliano
Enrich, Lucas Ariel

Integrantes del grupo:
Paula González (60784),
Matias González Virgili (61091),
Justo Soleño (61572),
Candela Palomeque (60827)

Índice

1. Introducción	3
2. Objetivos	3
3. Relación entre desempeño y cantidad de materias	4
4. Relación entre estrés y cantidad de tareas	12
5. Relación entre estrés y desempeño	26
6. Nivel de depresión en relación a otras variables	27
a. Depresión y desempeño	29
b. Depresión y el uso del Piazza	31
c. Depresión y cantidad de horas de sueño	36
d. Depresión e interacciones sociales	40
e. Depresión y el nivel de soledad	44
7. Relación entre desempeño y cantidad de horas de sueño	48
8. Relación entre desempeño con interacciones sociales	49
9. Relación entre soledad e interacciones sociales	49
10. Relación entre soledad y desempeño	50
11. Outliers de peor desempeño	51
a. Características del estudiante con ID = 1	51
b. Características del estudiante con ID = 46	52
c. Características del estudiante con ID = 52	53
12. Análisis de correspondencias múltiples (MCA)	54
a. Análisis de la primera dimensión	57
b. Análisis de la segunda dimensión	59
13. Regresión Lineal Múltiple	60
14. Conclusión	62

1. Introducción

La base de datos que fue seleccionada para realizar el trabajo se llama “StudentLife Dataset”¹ y contiene datos de un estudio que fue realizado por Dartmouth College, una universidad ubicada en el estado de New Hampshire, Estados Unidos. Los datos fueron recolectados de una app de Android que contiene sensores y encuestas que evalúan el día a día de 60 alumnos durante un cuatrimestre. El estudio fue innovador dado que fue el primero en recopilar este tipo de data en la vida de los estudiantes lo que permitió, por primera vez, poder entender la vida de los estudiantes y cómo impactan sus patrones de comportamiento en su desempeño académico.

El objetivo principal del trabajo es averiguar **cómo se podría mejorar el desempeño de los estudiantes de Dartmouth College**. Para esto se intentará comprobar que el desempeño se relaciona con la calidad de vida del alumno y se evaluará qué hace que algunos alumnos tengan mejor rendimiento que otros. Con esta información, se propondrán modelos para ayudar a la universidad a aplicar políticas o programas que asista al alumnado para aumentar su rendimiento académico.

A pesar de que este estudio se enfocará exclusivamente en los datos recolectados de los estudiantes en Dartmouth College, parecería ser sumamente enriquecedor replicar una metodología similar de recolección de datos e implementación de modelos en nuestra universidad.

2. Objetivos

Observando las distintas variables numéricas y categóricas, se definieron objetivos secundarios para poder guiar el análisis descriptivo de los datos, enlistados a continuación:

- Analizar la existencia de una relación entre el estrés que sufren los participantes pre y post estudio con el desempeño académico de los mismos.

¹ Fuente: <https://studentlife.cs.dartmouth.edu/dataset.html#sec:ema>

- Estudiar cómo se alteran los niveles de estrés de los estudiantes a lo largo del cuatrimestre y evaluar si tiene relación con la cantidad de tareas que les dan para hacer.
- Evaluar si el desempeño académico está relacionado inversamente con la cantidad de materias que hace el alumno, fijarse cuántas materias cursan los alumnos con mejores notas, y según eso recomendar esa cantidad de materias.
- Evaluar cómo el índice de depresión influye en la utilización de Piazza (la plataforma de campus virtual), en el sueño (la calidad, horas, problemas para mantenerse despierto, etc). Ver si también está relacionado con el nivel de interacciones sociales del estudiante y con el nivel de soledad que siente.

El lenguaje de programación que se estará utilizando para realizar el análisis es R en R Studio.

3. Relación entre desempeño y cantidad de materias

En esta sección se evaluará si el desempeño académico está relacionado inversamente con la cantidad de materias que hace el alumno. Se verá cuántas materias cursan los alumnos con mejores notas, y según eso recomendar dicha cantidad de materias.

Para poder realizar este análisis, fue necesario realizar un análisis exploratorio de los datos de las tablas de la base de datos de “grades” y “Class”.

Se comienza analizando cómo se comportan las notas de los estudiantes. Antes de mostrar los resultados, parece menester destacar que de los 60 estudiantes que son parte de la muestra del estudio, sólo aparecen las notas de 31 de ellos, lo que significa que no aparecen las notas de 29 estudiantes que realizaron el estudio. En la *Figura 1*, se puede observar un breve resumen estadístico de las notas:

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50
1	gpa_all	0	1	3.42	0.398	2.4	3.26	3.49
2	gpa_13s	0	1	3.33	0.798	1	3.33	3.53
3	cs_65	0	1	3.62	0.796	0	3.67	4
4	gpa_allvs13s	0	1	-0.0910	0.667	-2.53	-0.163	0.0497
			p75 p100 hist					
1	3.70	3.95						
2	3.86	4						
3	4	4						
4	0.229	0.627						

Figura 1 - Summary de notas

Columna	Especificación
gpa_all	Promedio general del alumno en la carrera (siendo 4 el valor más alto).
gpa_13s	Promedio del alumno en el semestre.
cs_65	Promedio del alumno en una materia en esa materia en particular (viene esta columna en el archivo original).
pga_allvs13s	Cuantos puntos bajó o subió el alumno en ese semestre en comparación al promedio que venía llevando en la carrera.

Tabla 1 - Descripción variables

Para ayudar en la comprensión de la *Figura 1*, se armó la *Tabla 1* que explica que representa cada variable del resumen estadístico.

Por lo general, lo que se puede ver en estos resultados, es que el desempeño de los alumnos empeoró ligeramente durante el semestre que se realizó el análisis en comparación con su desempeño promedio de su carrera en total.

A pesar de que el promedio de las diferencias es negativo, es decir que en promedio los alumnos empeoraron su desempeño en el semestre del estudio, se observa que la mediana es positiva. Esto nos da a entender que la mayoría mejoró su desempeño con respecto a su promedio histórico. Se puede confirmar esto con la *Figura 2*, donde se observa que son 17 observaciones están a la derecha de la línea roja, o en otras palabras, que mejoraron sus notas, mientras que son solo 13 los que lo empeoraron.

Sin embargo, hay algunos alumnos que están muy lejos de la línea roja, lo que significa que empeoraron considerablemente. Por lo que se podría considerar como los causantes de que el promedio de las diferencias de los promedios sea negativo. Estos valores son posibles *outliers*.

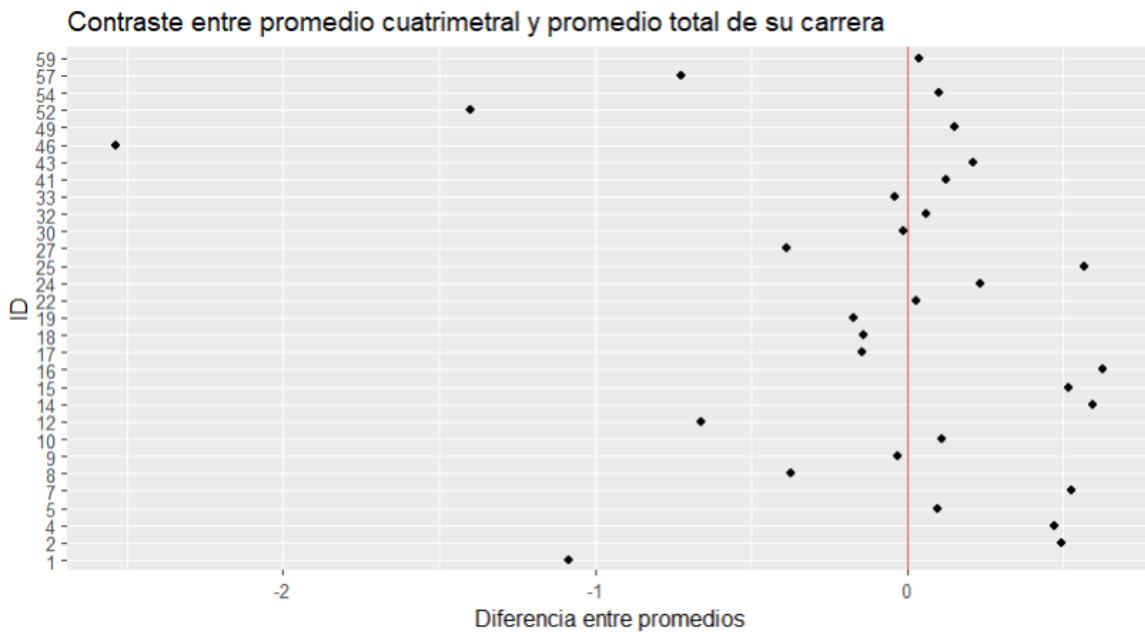


Figura 2 - Diagrama de dispersión de la relación entre variables

Para corroborar esto, se realizó un boxplot con estas diferencias (véase Figura 3). Efectivamente, se encontraron 3 outliers correspondientes a los id 1, 46 y 52. Estos alumnos nos pueden otorgar insights de porque su desempeño se vio tan afectado en este cuatrimestre en particular. Además, se podría analizar qué características tienen en común los alumnos que mejor les fue para poder tomarlos como “modelos a seguir” en cuánto a sus hábitos analizando los patrones de comportamiento y buscando similitudes.

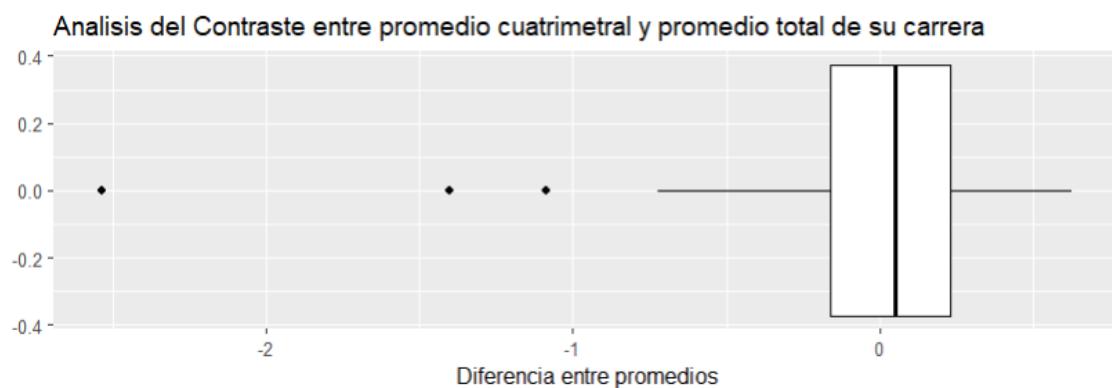


Figura 3 - Boxplot de la relación entre variables

El gráfico presenta asimetría negativa, lo que indica que hay más valores a izquierda de la media que es, a su vez, donde se encuentran los outliers.

Luego, se quiso visualizar cómo se distribuyen los distintos promedios de los alumnos en cada caso. Para eso se realizó el gráfico de dispersión que se puede ver en la Figura 4. En color negro se pueden vizualizar los promedios (el GPA) de cada alumno en toda su carrera,

en color azul se visualizan los promedios de las notas de los alumnos en el semestre del estudio, y en color verde se visualizan el promedio de los alumnos en la materia CS_65 (que es la materia que hace el estudio y por eso todos los estudiantes del estudio la cursan). Hay algunos alumnos cuya nota del promedio del semestre coincide con la nota del promedio de su carrera total por ende esos puntos en la *Figura 4* se superponen.

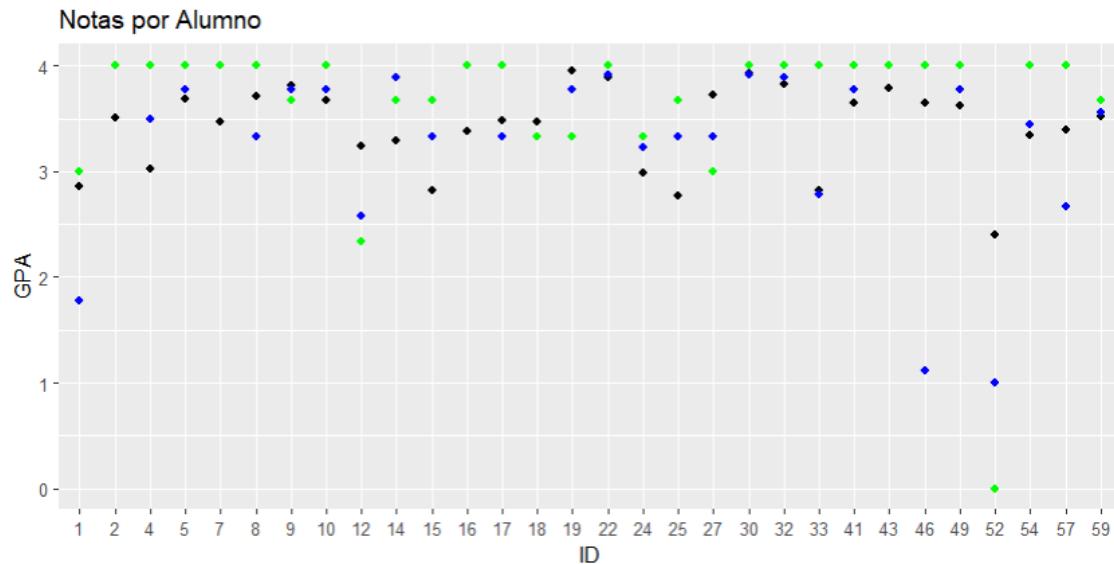


Figura 4 - Diagrama de dispersión de los promedios

Se comprendió que para poder analizar de forma más óptima cómo se comportan los resultados, como la mediana y los cuartiles, era óptimo realizar un boxplot para cada uno de los promedios que se puede observar en la *Figura 5* a continuación.

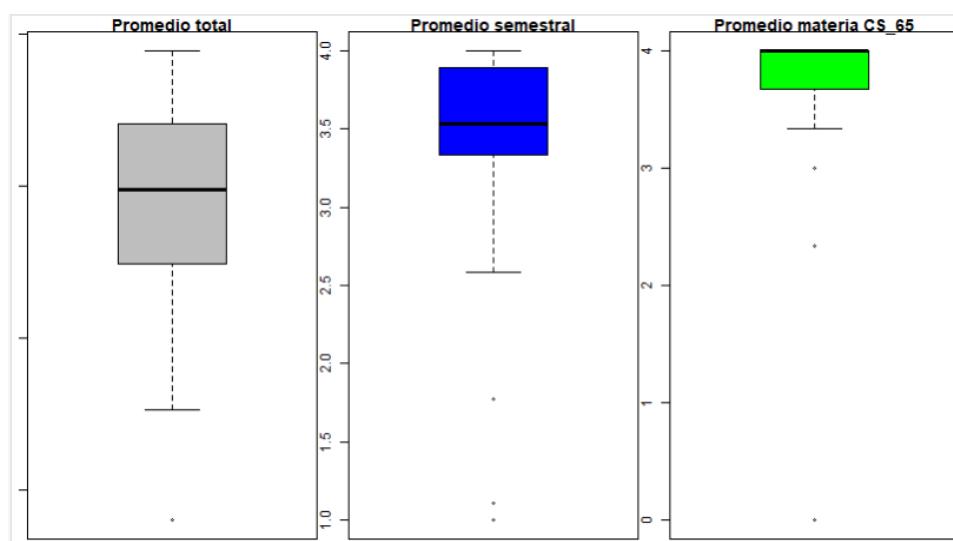


Figura 5 - Múltiples Boxplots de los promedios

Como se puede observar en la *Figura 5*, la mediana del promedio de las notas de la materia CS_65 es muy elevada, que es el puntaje máximo de las notas. Si bien hay algunos casos atípicos, lo de la mediana da lugar a pensar que quizás es una materia muy fácil por ende no es un indicador de si el alumno le va bien en la universidad o no.

Luego de analizar las notas, se continuó analizando las materias que cursa cada alumno. En primer lugar, en esta tabla se reconoce que hay 11 alumnos faltantes, por lo tanto el análisis sobre esta tabla estará basado en los 49 alumnos que sí aparecen. En total, se destaca que hay 43 materias distintas que cursan los estudiantes. En el siguiente resumen estadístico (véase *Figura 6*), se puede destacar que, como la media es aproximadamente 2 alumnos por cada materia², hay una materia que es cursada por 43 alumnos (la de COSC_065).

subject	cantidad	porcentaje
Length:44	Min. : 1.000	Min. : 0.8333
Class :character	1st Qu.: 1.000	1st Qu.: 0.8333
Mode :character	Median : 1.000	Median : 0.8333
	Mean : 2.727	Mean : 2.2727
	3rd Qu.: 2.000	3rd Qu.: 1.6667
	Max. :43.000	Max. :35.8333

Figura 6- Estadísticos básicos de tabla class

Otra observación es que el máximo de materias que cursan los alumnos es de 4, por lo que se supone que la universidad sólo los deja anotarse en esa cantidad de materias.

Luego, se quiso contar la cantidad de materias cursadas por los alumnos. Para eso, se comenzó realizando un resumen estadístico que se puede ver a continuación en la *Figura 7*.

uid	suma_mat
Length:49	Min. :0.000
Class :character	1st Qu.:2.000
Mode :character	Median :3.000
	Mean :2.449
	3rd Qu.:3.000
	Max. :4.000

Figura 7- Estadísticos básicos de nueva tabla de class

Como se puede observar en la *Figura 7*, el promedio de cantidad de materias cursadas por los alumnos es aproximadamente 2.5.

² Cada materia que aparece en la base de datos del estudio, suponemos que en la realidad no hay 2 alumnos por materia en serio.

Cantidad de materias cursadas por alumno

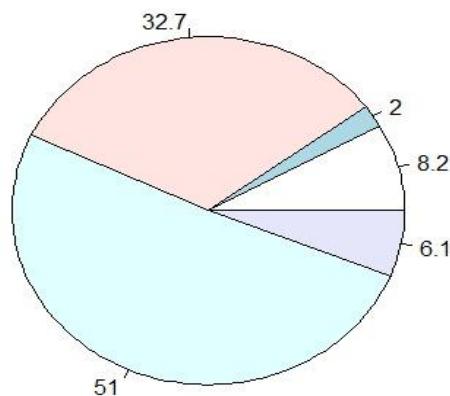


Figura 8 - Gráfico de torta de cantidad de materias

En la *Figura 8*, se puede observar un gráfico de tortas que muestra el porcentaje de alumnos que cursa cierta cantidad de materias. Los colores de los porcentajes representan lo siguiente:

- Blanco: 0 materias cursadas
- Azul: 1 materia cursada
- Rosa: 2 materias cursadas
- Celeste/Verde: 3 materias cursadas
- Violeta: 4 materias cursadas

Como se puede observar, casi la mitad (el 51%) de los estudiantes cursan 3 materias, mientras que casi un tercio de los mismos (el 32.7%) cursan 2 materias, el 6.1% cursan 4, el 8.2% cursan 3 y el 2% cursan 1 materia. Resulta curioso pensar por qué aparecen alumnos en la tabla de las materias si cursan 0, y hay otros que directamente no aparecen, ¿cuál sería la diferencia entre ambos grupos?

Una vez hecho este análisis exploratorio de las tablas, vuelve a aparecer la pregunta: ¿hay relación entre la cantidad de materias que hacen los alumnos con su desempeño? Para poder hacer las correlaciones entre los datos de las tablas, se debe completar los datos faltantes. Como se mencionó previamente, en la tabla de las materias que hace cada alumno hay 11 alumnos faltantes, mientras que la tabla de las notas tiene 29 alumnos faltantes. Observando los faltantes en ambas tablas, se deja ver que hay alumnos que faltan en ambas tablas, por ende se decidió no contar con ellos para este análisis. Luego, se observa que hay algunos datos que aparecen que están cursando materias pero no figuran con notas (en la tabla de notas), por ende se decidió imputar los faltantes con el promedio

de las notas para poder generar un valor “artificial” de las notas, que tenga sentido con el conjunto de datos”, para no tener faltantes y poder realizar una óptima correlación.

Una vez imputados los datos faltantes, se realizaron los test de correlación entre las materias hechas en el semestre (véase *Figura 9*) y, primero el promedio de notas total, y luego el promedio de notas en el cuatrimestre.

```
Pearson's product-moment correlation

data: df_gpa_imp$gpaall and df_gpa_imp$total
t = -0.43683, df = 43, p-value = 0.6644
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.3531124 0.2315843
sample estimates:
cor
-0.06646794
```

Figura 9 - Correlación de Pearson entre variables

La primera prueba devolvió un R^2 de -0,06, lo cual es bajo ya que para variables sociológicas necesitamos por lo menos un R^2 de mod (0,5) y el test de correlación devolvió un valor alto (0,6644), lo que indica que la relación no es significativa.

Esto es evidencia suficiente como para concluir que **no hay relación entre las materias hechas ese cuatrimestre y el promedio de notas general**.

```
Pearson's product-moment correlation

data: df_gpa_imp$gpa13s and df_gpa_imp$total
t = -2.5784, df = 43, p-value = 0.01343
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.59549625 -0.08110299
sample estimates:
cor
-0.3659255
```

Figura 10 - Correlación de Pearson entre variables

En el segundo caso (véase *Figura 10*), hay un coeficiente de correlación (R^2) de -0,3659, lo cual sigue siendo bajo para los estándares establecidos para esta clase de variables. Sin embargo, el p value del test de correlación es de 0,01343, es decir, significativo. Lo que indica esto es que hay relación entre las notas del cuatrimestre y la cantidad de materias cursadas, pero debido al R^2 bajo, la relación no es lineal, por lo que hay que probar si la distribución de datos se ajusta mejor a otro modelo.

Spearman's rank correlation rho

```
data: df_gpa_imp$gpa13s and df_gpa_imp$total
S = 22950, p-value = 0.0003256
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
-0.5118696
```

Figura 11 - Correlación de Spearman entre variables

Luego de probar varios modelos de distribución de datos, tales como chi cuadrado o el que otorga la función dcor, se concluye que la única distribución que ajusta bien al requisito mínimo para realizar una regresión es la logarítmica (véase Figura 11). El p value devuelto del test de correlación es de 0,003256 y el R²es de -0,51186, que si bien no es muy alto, es lo suficiente para realizar la regresión.

```
Call:
lm(formula = gpaall ~ total, data = df_gpa_imp)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.04513 -0.10213  0.06936  0.20087  0.50187 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.54812   0.21539 16.473 <2e-16 ***
total       -0.03433   0.07859 -0.437   0.664    

```

Figura 12

Considerando Y la el promedio de notas del alumno en el cuatrimestre y X la cantidad de materias, la regresión que mejor ajusta es:

```
call:
lm(formula = log10(df_gpa_imp$total) ~ df_gpa_imp$gpaall)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.42957 -0.10535  0.04594  0.07142  0.19893 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.54564   0.17814  3.063  0.00377 ** 
df_gpa_imp$gpaall -0.03832   0.05093 -0.752  0.45590
```

Figura 13

Esto se interpreta como:

$$\ln(y) = 0.54564 - 0.03832x$$

Es decir que **por cada 1% que sube la nota, la cantidad de materias baja en 0,03832 unidades**. O para simplificar aún más, **mientras menos materias uno hace, mejor rendimiento tiene**, por lo que la mejor cantidad de materias que uno puede hacer para mejorar su rendimiento es una materia.

4. Relación entre estrés y cantidad de tareas

En esta sección se estudiará cómo se alteran los niveles de estrés de los estudiantes a lo largo del cuatrimestre y evaluar si tiene relación con la cantidad de tareas que les dan para hacer.

En primer lugar, se hará un análisis exploratorio de la cantidad de tareas que recibió cada estudiante por día desde el día 27/3/2013 hasta el 5/6/2013. Para hacer esto, se crea un dataframe en donde se almacena la cantidad total de tareas por alumno en todo el semestre. Se pudo reconocer que habían 15 alumnos que faltaban, por ende se hizo el análisis de las tareas sobre 44 alumnos. Se realizó un resumen estadístico de esta tabla y nos dieron los valores que se pueden ver en la *Figura 14*.

```
          uid      total
Length:59   Min.   : 7.00
Class :character  1st Qu.:11.00
Mode  :character  Median :15.00
                  Mean   :18.23
                  3rd Qu.:20.00
                  Max.   :84.00
                  NA's    :15
                  

> sd(tareas$total, na.rm = TRUE) #desvio estandar
[1] 13.04886
> sd(tareas$total, na.rm = TRUE)/mean(tareas$total, na.rm = TRUE) #coeficiente de variacion
[1] 0.7158975
```

Figura 14 - Estadísticos básicos de la tabla Deadlines

A simple vista se puede ver que en promedio los alumnos tienen aproximadamente 18 entregas, sin embargo, también se observa que el valor máximo llega a 84 (casi 5 veces por encima de la media) y que el mínimo es de 7 (apenas un tercio de la media). Esto podría estar indicando un alto grado de variación en la cantidad de entregas de los alumnos, hecho que es confirmado por el coeficiente de variación que es de un 71.59%.

Lo que indica esto es que no hay un número determinado de entregas que tiene cada estudiante, hay algunos que tienen muchos más que otros.

Al tener esta gran variedad de resultados, más adelante se analiza que tanto afecta la cantidad de entregas a sus notas, ya que se puede ver si es mejor tener un número elevado de entregas o uno más reducido. Además, se plantea un análisis para entender si existe una correlación entre las materias que cursan los alumnos que se presentan como outliers con la cantidad de tareas que realizan.

Se probará eliminando los outliers para ver cuánto modifican la media, el desvío estándar y el coeficiente de variación. Al hacer eso, devuelve los datos que se pueden ver a continuación en la *Figura 15*:

```
uid      total
Length:42   Min.    : 7.00
Class :character  1st Qu.:11.00
Mode  :character  Median  :15.00
                  Mean   :15.86
                  3rd Qu.:18.75
                  Max.   :31.00
> sd(tareas_sinoutliers$total, na.rm = TRUE) #desvio estandar
[1] 6.28421
> sd(tareas_sinoutliers$total, na.rm = TRUE)/mean(tareas_sinoutliers$total, na.rm = TRUE)
#coeficiente de variacion
[1] 0.3963015
> |
```

Figura 15- Estadísticos básicos sacando outliers de la tabla Deadlines

Como se puede observar, si se eliminan los valores atípicos de la muestra, el promedio da 15.86, que es un valor muy parecido a la mediana, por lo que se puede decir que la distribución es bastante simétrica (mucho más que en el caso anterior con los outliers). También, como era de esperarse, el desvío estándar y el coeficiente de variación se reducen a la mitad.

A continuación en la *Figura 16* se muestra un gráfico de dispersión, donde se puede observar más claramente lo mencionado anteriormente acerca de la *Figura 15*. Se puede ver cómo, si bien la media es de aproximadamente 18 tareas por alumno, la mayoría tiene menos tareas que el promedio lo cual se da porque hay valores outliers que elevan el valor de la media de forma pronunciada, como por ejemplo el alumno que tuvo 84 tareas (se comprueba que es outlier en la *Figura 17*). Tras el análisis de los resultados, se pudo observar que solo fueron 13 alumnos quienes recibieron más tareas que la media mientras que 31 alumnos de 44 recibieron menos tareas que la media.

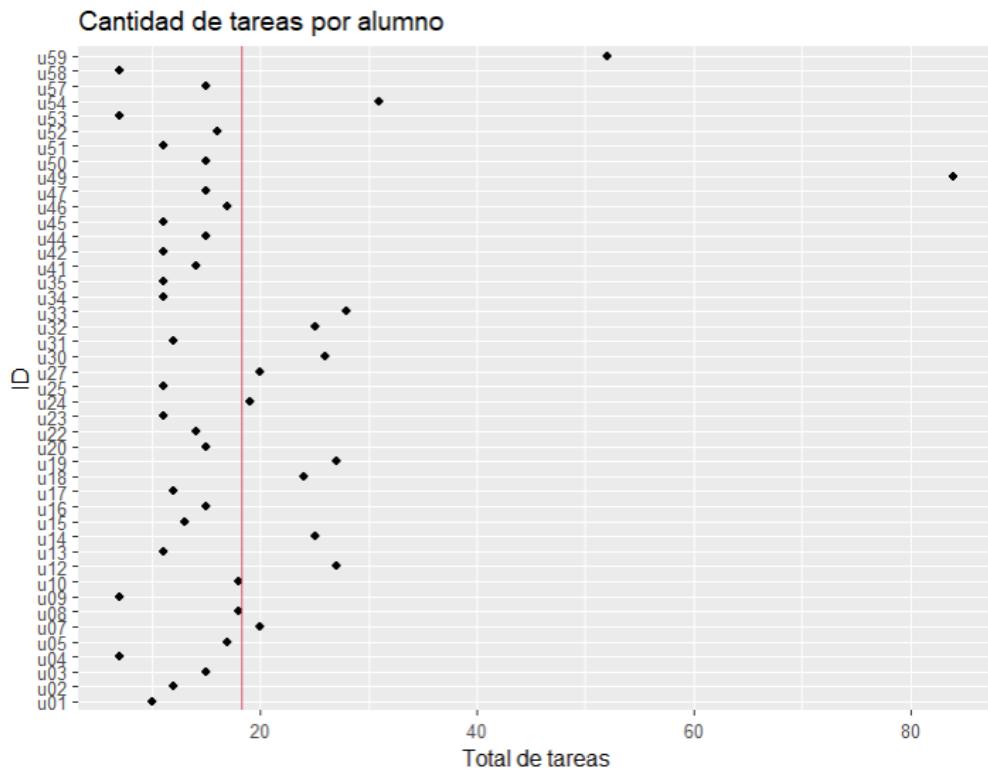


Figura 16- Gráfico de dispersión de Deadlines

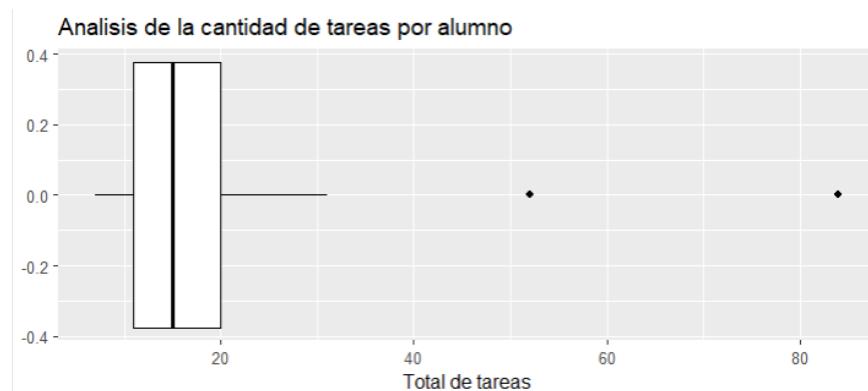


Figura 17- Boxplot de Deadlines

En la Figura 17 se presenta un boxplot con asimetría positiva de la media de tareas por alumno, donde se puede ver que hay 2 outliers significativos (uno con 52 tareas con id = 59 que cursó las materias: GERM 001, COSC 065 y COSC 007, y otro con 84 tareas con id = 49 que cursó las materias: MATH 013, LAT 003 y COSC 065), mientras que la mayoría se encuentra alrededor de 15 tareas por alumno. El participante 59 en promedio tiene un nivel de estrés de 2 puntos de un total de 5 y el 49 tiene un nivel de estrés promedio de 3 puntos de un total de 5. Al contrario de lo que se podría haber pensado en un principio, si bien presentaron una cantidad de tareas superior en comparación con el resto de los participantes del estudio, estos 2 en particular presentaron un nivel de estrés promedio

normal para la cantidad de tareas. Se podría suponer que la dificultad de las mismas no fue elevada o que se realizaron con tiempo lo que pudo haber ayudado a minimizar el estrés.

Ahora, se realizará un análisis del estrés de los estudiantes. En la base de datos, se encontró dos tipos de fuentes de información que nos hablaban del estrés. Uno fue una encuesta que se le hizo a los estudiantes al empezar y finalizar el estudio, y el otro fueron las encuestas que fueron respondiendo frecuentemente a lo largo de todo el período del estudio.

Por el lado de la encuesta respondida al inicio y final del estudio, la misma se llama Perceived stress scale³; es una serie de 10 preguntas que se responden con:

- Never (0)
- Almost never (1)
- Sometimes (2)
- Fairly often (3)
- Very Often (4)

El análisis de esta encuesta permitirá tener una idea general acerca de los niveles de estrés que manejaron los alumnos durante el estudio sin adentrarnos en tanto detalle, y como se puede ver, a cada una de las respuestas se les asigna un número, que luego se usan para sumarlos y realizar un puntaje que determina el nivel de estrés del encuestado. El score que devuelve este cálculo se divide en tres categorías de niveles de estrés:

- Low stress (0 - 13)
- Moderate stress (14 - 26)
- High percibes stress (27 - 40)

Al calcular todos los scores de los estudiantes, se observa que se cuenta con valores faltantes que fueron imputados utilizando el algoritmo miss Forest el cual arrojó un error de estimación total de 0.898%. A partir de eso, se realizó un boxplot para poder visualizar si existían valores atípicos de los niveles de estrés, el cual resultó en el boxplot a continuación.

³ <https://novopsych.com.au/assessments/health/perceived-stress-scale-pss-10/>

Boxplot de los Scores de estrés

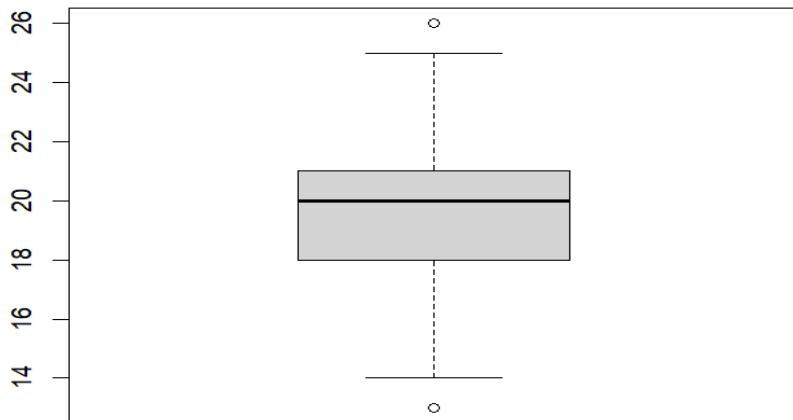


Figura 18 - Boxplot sobre los scores de la encuesta de los alumnos

En el boxplot de la *Figura 18* se observan 2 outliers. El outlier del bigote inferior pertenece a un alumno con bajos niveles de estrés y el otro outlier pertenece a un alumno que se encuentra en el valor más alto de las categoría moderate stress. A ambos outliers se los considera válidos dentro de la muestra por lo que no se los quitará ni modificará en el posterior análisis. También se presenta una asimetría positiva la cual indica que los alumnos tienden a estar menos estresados que la mediana. El 50% de los alumnos en la muestra tienen un score de estrés entre 18 - 21 lo que lleva a concluir que la mitad de los alumnos tuvieron un estrés moderado, no fuera de lo común.

Por otro lado, a continuación se evalúa con más en profundidad los valores del estrés utilizando las encuestas que fueron respondiendo frecuentemente a lo largo de todo el período del estudio. Las mismas estaban en formato .json completamente desestructurado, por lo tanto, para poder manejar los datos se reestructuraron los datos para poder trabajarlos en R.

Lo primero que se hizo visible fue que había un 3% de datos faltantes de los niveles de estrés registrados de los alumnos (que son valores que van del 0 al 5), y para poder imputarlos se decidió reemplazarlos por el promedio del nivel de estrés que se respondió (2.223). Además, habían alumnos que respondieron la encuesta más veces que otros, y esto se hace visible en el siguiente gráfico (*Figura 19*):

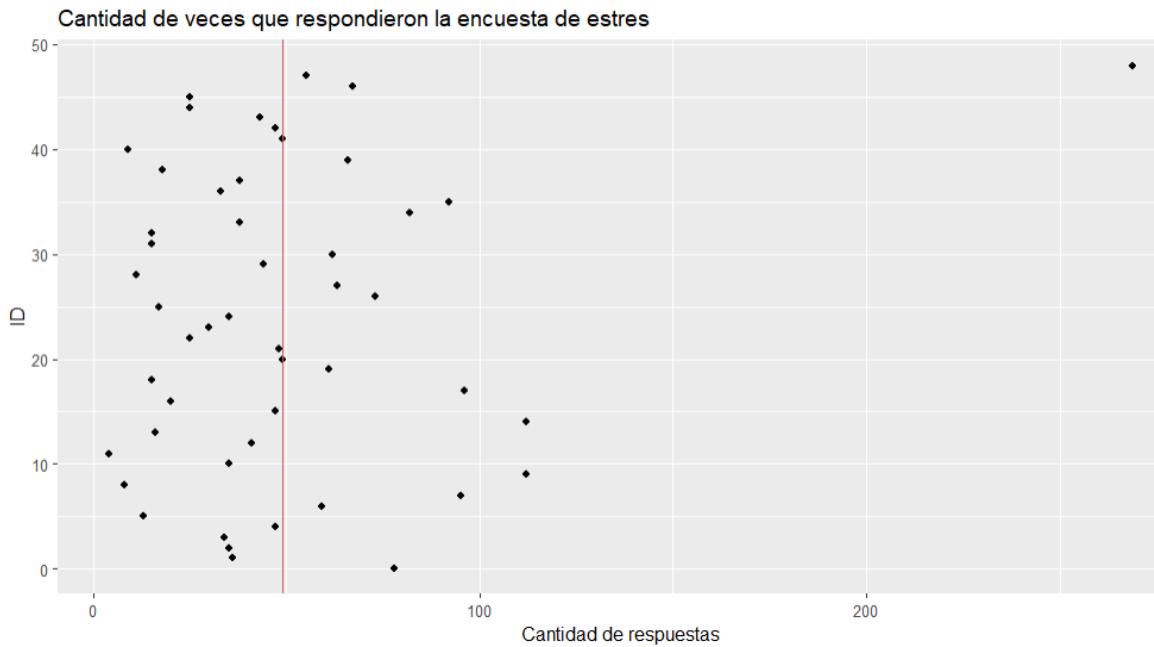


Figura 19 - Dispersión de respuestas a encuesta de estrés

Resulta interesante analizar la dispersión de los promedios de estrés por estudiante (véase Figura 20), y para eso se generó el gráfico que se puede observar a continuación:

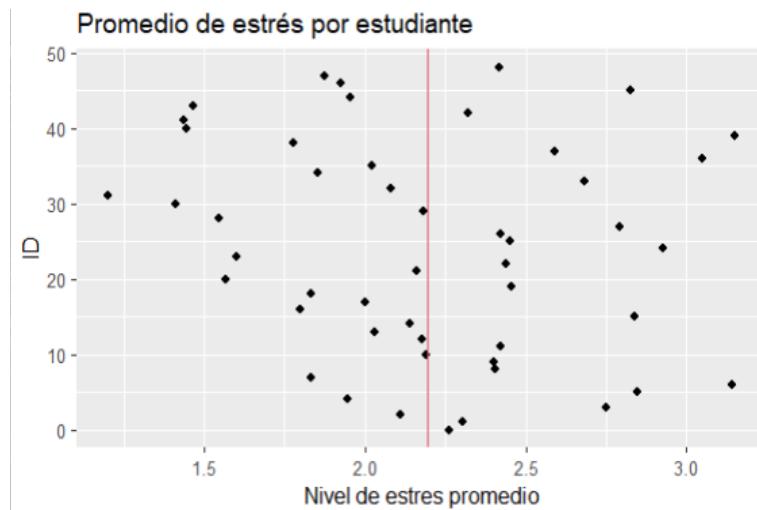


Figura 20

Como se puede ver también en el boxplot de la Figura 21 , no hay outliers (como era de esperarse porque los niveles de estrés se puntuaba del 0 al 5), y los puntos de estrés se distribuyen bastante simétricamente, no hay una tendencia hacia un valor en específico.

Boxplot de los promedios de estrés de los estudiantes

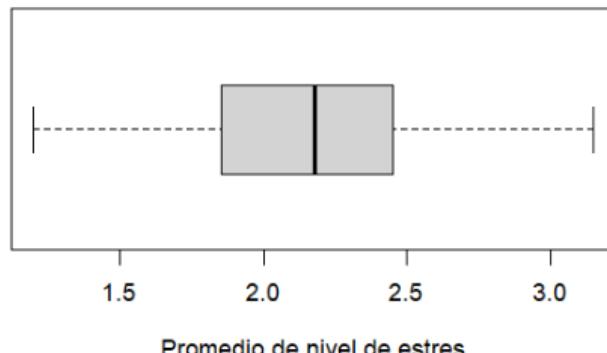


Figura 21 - Boxplot promedio nivel de estrés

A su vez, resulta interesante analizar cómo influye el mes del año en el estrés de los alumnos, por lo que se averiguó el promedio de nivel de estrés de los alumnos según el mes del año en que se hizo el cuestionario (véase Figura 22).

mes	promedio
1	2.18
2	2.26
3	2.13
4	3.81
5	1.75

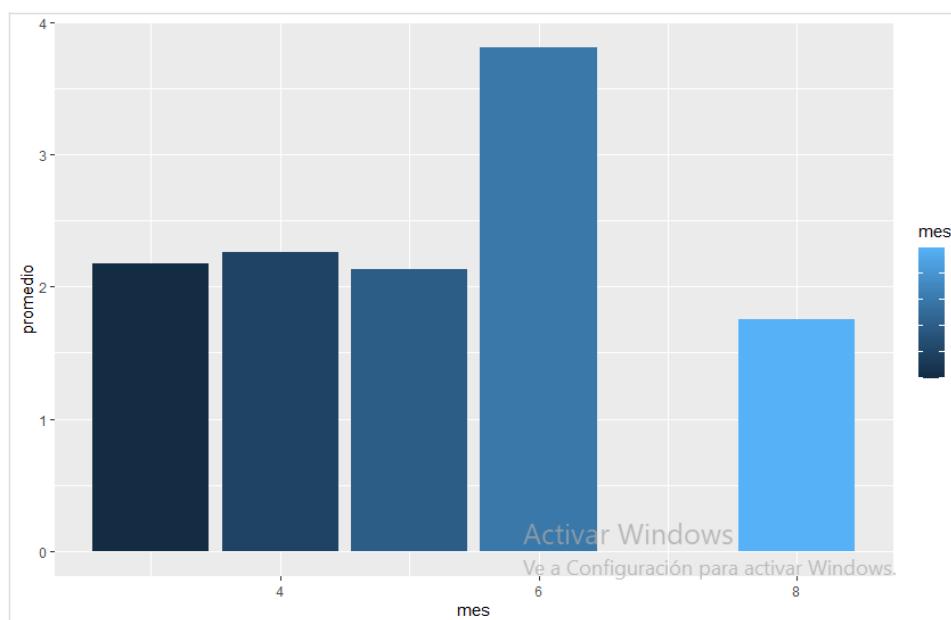


Figura 22

La muestra indica que en junio el nivel de estrés aumenta considerablemente, mientras que en agosto disminuye un poco. Sin embargo, también se realizará un test de hipótesis ANOVA para corroborar que los alumnos se estresan más en junio, o simplemente es un error muestral. Es importante aclarar que para este test, se supone la independencia de las observaciones y la homogeneidad de las varianzas.

```
> anova
Call:
aov(formula = mes_caracter$level ~ mes_caracter$`as.character(df$month)`)

Terms:
            mes_caracter$`as.character(df$month)` Residuals
Sum of Squares           60.783   4326.762
Deg. of Freedom                  4          2402

Residual standard error: 1.342131
Estimated effects may be unbalanced
> summary(anova)

             DF Sum Sq Mean Sq F value Pr(>F)
mes_caracter$`as.character(df$month)`  4     61   15.196  8.436 9.33e-07 ***
Residuals                           2402  4327   1.801
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 23 - Test de hipótesis ANOVA

Al obtener un valor p tan bajo (casi de 0, véase *Figura 23*), indica que por lo menos una media es distinta al resto. Esto lleva al siguiente paso que es ver qué medias son mayores realmente y cuáles no. Por lo que se hará un test de hipótesis de diferencia de media por cada mes (véanse *Figura 24* y *Figura 25*).

```
Welch Two Sample t-test

data: mes_caracter$level[mes_caracter$`as.character(df$month)` == "3"]
nth) == "4"]
t = -1.1904, df = 966.86, p-value = 0.2342
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.2221525  0.0543966
sample estimates:
mean of x mean of y
2.176242  2.260120
```

Figura 24

```
welch Two Sample t-test

data: mes_caracter$level[mes_caracter$`as.character(df$month)` == "3"] and mes_caracter$level
nth)` == "5"]
t = 0.51503, df = 1045, p-value = 0.6066
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.1193189 0.2042449
sample estimates:
mean of x mean of y
2.176242 2.133779
```

Figura 25

```
welch Two Sample t-test

data: mes_caracter$level[mes_caracter$`as.character(df$month)` == "3"] and mes_caracter$level
nth)` == "6"]
t = -7.0179, df = 22.942, p-value = 3.812e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.114789 -1.151774
sample estimates:
mean of x mean of y
2.176242 3.809524
```

Figura 26

```
welch Two Sample t-test

data: mes_caracter$level[mes_caracter$`as.character(df$month)` == "3"] and mes_caracter$level
nth)` == "8"]
t = 0.88348, df = 3.0949, p-value = 0.4402
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.082852 1.935336
sample estimates:
mean of x mean of y
2.176242 1.750000
```

Figura 27

Las imágenes de arriba corresponden a las pruebas de hipótesis de media de los niveles de estrés de marzo con el resto de los meses. En la *Figura 26*, correspondiente al contraste con el mes de mayo, y la *Figura 27*, correspondiente al contraste con el mes de agosto, el valor de p devuelto fue de 0,6066 y de 0,4402 respectivamente. Al ser valores mayores a 0,05, no hay evidencia suficiente como para decir que estas medias son diferentes a las de marzo. En el test correspondiente al mes de marzo y abril, si bien el valor p es casi 0, el intervalo de confianza contiene al 0, esto indica que si bien se puede decir que las medias son distintas, no se puede concluir que una sea mayor a la otra. Sin embargo, en el test realizado con el mes de junio, se dan ambas condiciones deseadas, (valor p casi nulo y un intervalo de confianza que excluye al 0) lo que es prueba suficiente para confirmar que los alumnos en junio están más estresados que en agosto.

Siguiendo con el mes de abril (véanse *Figura 28*, *Figura 29* y *Figura 30*), la diferencia de medias con mayo y agosto no son lo suficientemente grandes como para que sean significativas, ya que el valor p que devolvió los test fueron de 0,06191 y 0,3651 respectivamente.

```
welch Two Sample t-test
data: mes_caracter$level[mes_caracter$`as.character(df$month)` == "4"] and mes_caracter$level[mes_caracter$`as.character(df$month)` == "5"]
t = 1.8691, df = 969.45, p-value = 0.06191
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.006306945 0.258988893
sample estimates:
mean of x mean of y
2.260120 2.133779
> |
```

Figura 28

```
data: mes_caracter$level[mes_caracter$`as.character(df$month)` == "4"] and mes_caracter$level[mes_caracter$`as.character(df$month)` == "6"]
t = -6.7988, df = 21.097, p-value = 9.827e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.023200 -1.075608
sample estimates:
mean of x mean of y
2.260120 3.809524
> |
```

Figura 29

```
welch Two Sample t-test
data: mes_caracter$level[mes_caracter$`as.character(df$month)` == "4"] and mes_caracter$level[mes_caracter$`as.character(df$month)` == "8"]
t = 1.0624, df = 3.0359, p-value = 0.3651
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.007722 2.027962
sample estimates:
mean of x mean of y
2.26012 1.75000
|
```

Figura 30

Sin embargo nuevamente junio muestra mayor nivel de estrés que en abril (véanse *Figura 31* y *Figura 32*), ya que el valor p de la correspondiente prueba es casi 0 y su intervalo de confianza no contiene el 0, por lo que prueba que siempre es mayor.

```
welch Two Sample t-test

data: mes_caracter$level[mes_caracter$`as.character(df$month)` == "5"] and mes_caracter$level[mes_caracter$`as.character(df$month)` == "6"]
t = -7.2268, df = 22.609, p-value = 2.597e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.155883 -1.195607
sample estimates:
mean of x mean of y
2.133779 3.809524

> |
```

Figura 31

```
data: mes_caracter$level[mes_caracter$`as.character(df$month)` == "5"] and mes_caracter$level[mes_caracter$`as.character(df$month)` == "8"]
t = 0.79615, df = 3.0844, p-value = 0.4827
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.126831 1.894389
sample estimates:
mean of x mean of y
2.133779 1.750000

> |
```

Figura 32

En el caso de mayo, el valor devuelto por el test de diferencia de medias con el mes de junio dió significativamente alto, y el intervalo de confianza no contiene al 0, por lo que en junio los estudiantes tienen un nivel más alto de estrés que en mayo. Mientras que la diferencia de su media de niveles de estrés con el mes de agosto no es significativa, debido a su alto valor de p (0,4827)

Para finalizar el análisis del nivel de estrés según el mes, el mes de junio (véase Figura 33) muestra tener niveles más altos que agosto. El valor p es de 0,01446 y no contiene 0 en su IC. Lo que concluye estas pruebas es que en junio los alumnos tienen mucho más estrés que en el resto de los meses, mientras que el resto de los meses no hay mucha diferencia.

```
welch Two Sample t-test

data: mes_caracter$level[mes_caracter$`as.character(df$month)` == "6"] and mes_caracter$level[mes_caracter$`as.character(df$month)` == "8"]
t = 3.894, df = 4.4376, p-value = 0.01446
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.6464488 3.4725989
sample estimates:
mean of x mean of y
3.809524 1.750000

> |
```

Figura 33

Para continuar el análisis exploratorio, se realizarán los mismos pasos que se hicieron con los meses del año, pero con el nivel de estrés según el día de semana (véase Figura 34).

<i><chr></i>	<i><dbl></i>
1 domingo	2.18
2 jueves	2.31
3 lunes	2.22
4 martes	2.15
5 miércoles	2.19
6 sábado	2.50
7 viernes	2.12

Figura 34

A simple vista, no se ve que haya mucha diferencia entre las medias, sin embargo se hará nuevamente un test de ANOVA (véase Figura 35) para ver si alguna media tiene una diferencia significativa, con los mismos supuestos que establecimos en el apartado anterior.

```
call:
  aov(formula = mes_caracter$level ~ mes_caracter`as.character(df$month)`)

Terms:
  mes_caracter`as.character(df$month)` Residuals
  Sum of Squares           60.783   4326.762
  Deg. of Freedom          4          2402

Residual standard error: 1.342131
Estimated effects may be unbalanced
> summary(anova)
      Df  Sum Sq Mean Sq F value    Pr(>F)
mes_caracter`as.character(df$month)`  4     61  15.196  8.436 9.33e-07 ***
Residuals                          2402  4327   1.801
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

Figura 35 - Test de ANOVA

El valor otorgado por el test es casi 0, esto significa que por lo menos una media difiere del resto.

Luego de contrastar el nivel de estrés en los domingos con el resto de los días de la semana (véase Figura 36), el único valor de p que retorno un valor mayor a 0,05 fue con el día sábado. Además, en la imagen se puede ver que el IC no contiene al 0, esto significa que los alumnos están más estresados los sábados que los domingos

```
data: df$level[df$d_week == "domingo"] and df$level[df$d_week == "sábado"]
t = -2.816, df = 484.61, p-value = 0.005061
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.53225737 -0.09475564
sample estimates:
mean of x mean of y
2.183208 2.496714
```

Figura 36

Siguiendo con el día lunes (véase *Figura 37*), nuevamente se encontró que los alumnos reportaron un nivel de estrés significativamente más alto en los sábados, ya que el p es de 0,01 y el IC no contiene 0. Este fenómeno no se repitió con el resto de los días.

```
welch Two Sample t-test

data: df$level[df$d_week == "lunes"] and df$level[df$d_week == "sábado"]
t = -2.3819, df = 489.01, p-value = 0.01761
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.50089027 -0.04806064
sample estimates:
mean of x mean of y
2.222239 2.496714
```

Figura 37

La *Figura 38* muestra la prueba de diferencia de medias entre las medias de los niveles de estrés en los martes y los sábados. Ya que nuevamente sábado es el único día que tiene una diferencia significativa con martes, con un valor p=0,003382 y un IC sin 0. La prueba con el resto de los días dieron todos un valor p mayor a 0,05.

```
welch Two Sample t-test

data: df$level[df$d_week == "martes"] and df$level[df$d_week == "sábado"]
t = -2.9446, df = 508.13, p-value = 0.003382
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.5739737 -0.1145765
sample estimates:
mean of x mean of y
2.152439 2.496714
```

Figura 38

Siguiendo con el día miércoles (véase *Figura 39*), nuevamente el día sábado muestra niveles de estrés significativamente más altos. Mientras que el resto de las pruebas dieron un p-value alto.

```
welch Two Sample t-test

data: df$level[df$d_week == "miércoles"] and df$level[df$d_week == "sábado"]
t = -2.685, df = 507.91, p-value = 0.007491
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.53351489 -0.08265239
sample estimates:
mean of x mean of y
2.188630 2.496714
```

Figura 39

Sin embargo, cuando se sigue con los contrastes de diferencias de medias del jueves (véanse *Figura 40* y *Figura 41*). En ningún día el valor p es lo suficientemente alto como para hacer una conclusión con una justificación estadística sólida, ya que con viernes devuelve 0.06844 y con sábado devuelve 0,1075.

```
data: df$level[df$d_week == "jueves"] and df$level[df$d_week == "viernes"]
t = 1.825, df = 670.54, p-value = 0.06844
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.01444717 0.39520678
sample estimates:
mean of x mean of y
2.30876 2.11838
```

Figura 40

```
data: df$level[df$d_week == "jueves"] and df$level[df$d_week == "sábado"]
t = -1.6122, df = 515.49, p-value = 0.1075
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.41699320 0.04108468
sample estimates:
mean of x mean of y
2.308760 2.496714
```

Figura 41

Para finalizar, el último test muestra que los sábados los alumnos están más estresados que en los viernes (véase *Figura 42*), debido a su p=0,001468 y su IC sin 0. Esto quiere decir que, exceptuando los jueves donde no hay evidencia suficiente, los alumnos están más estresados en los sábados que en cualquier otro día de la semana.

```
welch Two sample t-test

data: df$level[df$d_week == "viernes"] and df$level[df$d_week == "sábado"]
t = -3.1983, df = 514.33, p-value = 0.001468
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.6107294 -0.1459388
sample estimates:
mean of x mean of y
2.118380 2.496714
```

Figura 42

Por último, se analizará la relación entre los niveles de estrés del alumnado y la cantidad de tarea que tienen para hacer (véase *Figura 43*). Antes de empezar con los test, se imputaron los datos faltantes (tanto de las tareas como del estrés) con un algoritmo de miss forest.

Pearson's product-moment correlation

```
data: mat_final$promedio and mat_final$x
t = 0.49328, df = 58, p-value = 0.6237
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.1924477 0.3134163
sample estimates:
cor
0.06463555
```

Figura 43

Al hacer el test de coeficiente, se concluye que **no hay relación entre las variables de estrés y cantidad de tareas**, ya que el p-value devuelto es muy alto (0,6237), lo que significa que cualquier relación que tengan no tienen la significancia estadística necesaria como para proseguir el análisis.

5. Relación entre estrés y desempeño

En esta sección se pondrá en prueba si existe la relación entre las calificaciones de los alumnos y sus niveles de estrés, es importante aclarar que se realizó el estudio sobre los alumnos quienes figuraban en la tabla de notas, ya que si se incluían a todos la muestra estaría muy afectado por los valores generados con el algoritmo de miss Forest. Además, se imputaron los datos de estrés faltantes.

Pearson's product-moment correlation

```
data: as.numeric(df_gr_st_imp$gpaall) and as.numeric(df_gr_st_imp$promedio_stress)
t = -1.4949, df = 28, p-value = 0.1461
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.57574622 0.09799875
sample estimates:
cor
-0.2718691
```

Figura 44

Pearson's product-moment correlation

```
data: as.numeric(df_gr_st_imp$gpa13s) and as.numeric(df_gr_st_imp$promedio_stress)
t = -1.1791, df = 28, p-value = 0.2483
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.5357843  0.1549097
sample estimates:
cor
-0.2174979
```

Figura 45

Se realizó el test de coeficientes para verificar si había una relación entre el estrés y las calificaciones (véase *Figura 44* y *Figura 45*), tanto general como semestral. Sin embargo los resultados indican que tal relación no existe, ya que los p-value dieron demasiados altos (0,1461 y 0,2483, respectivamente), por lo que, sorprendentemente, **no hay suficiente evidencia para establecer una relación entre estrés y desempeño de los estudiantes.**

6. Nivel de depresión en relación a otras variables

En esta sección se buscarán indicadores de la depresión de un alumno. ¿Cómo es posible detectarlo? Para eso, se tratará de encontrar correlaciones entre depresión y uso de piazza, el sueño, el nivel de interacciones sociales que tienen, y por último el nivel de soledad.

En primer lugar, se estará realizando un análisis exploratorio de datos general a la variable de depresión. Para eso, se utilizarán unos resultados a una encuesta llamada PHQ-9⁴, realizada al inicio y final del estudio. La encuesta consta de una serie de 9 preguntas que se responden con las siguientes opciones:

- Not at all (0)
- Several days (1)
- More than half the days (2)
- Nearly every day (3)

Cada una de las respuestas tienen asignado un valor numérico que se utiliza posteriormente para calcular un score que va a indicar un aproximado del nivel de estrés que tiene el encuestado. Las categorías de depresión dependen del score y son las siguientes:

⁴ <https://www.mdcalc.com/calc/1725/phq9-patient-health-questionnaire9>

- Minimal (1 - 4)
- Minor (5 - 9)
- Moderate (10 - 14)
- Moderately severe (15 - 19)
- Severe (20 - 27)

A continuación, la cantidad de estudiantes en cada uno de estos niveles de depresión previo y posterior al estudio:

depression severity	minimal	minor	moderate	moderately severe	severe
score	1-4	5-9	10-14	15-19	20-27
number of students (pre-survey)	17	15	6	1	1
number of students (post-survey)	19	12	3	2	2

Tabla 2 - Resultados encuesta PHQ-9

Con estos valores de la tabla de arriba, se genera un boxplot para poder visualizar si se encuentran valores atípicos, y éste se puede visualizar en la Figura 46.

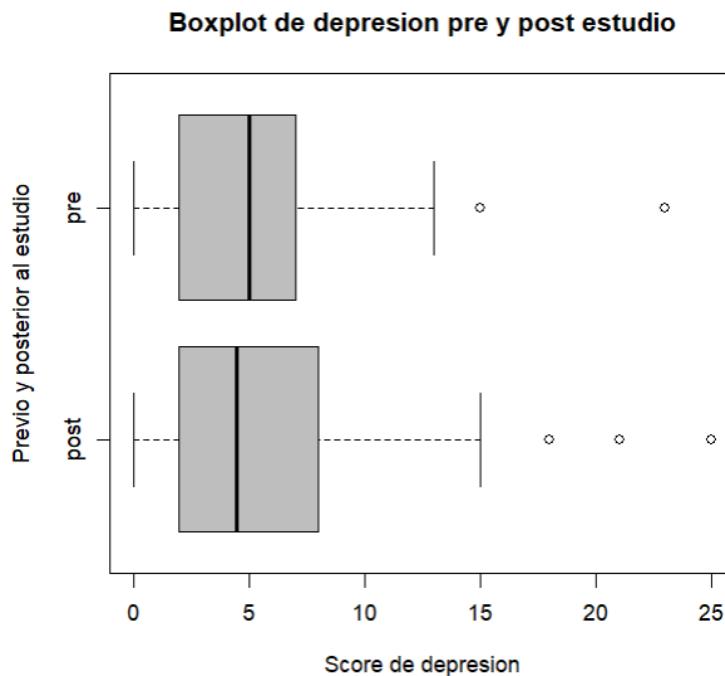


Figura 46 - Boxplot de los scores de la encuesta PHQ-9 pre y post estudio

En la *Figura 46* se representa 2 boxplot, el primero corresponde a la depresión previa al estudio, y el segundo corresponde a los resultados de la depresión al finalizar el estudio. Contrastando los boxplots, hay mayor variabilidad de resultados del segundo boxplot que en el primero, si bien la mediana del boxplot de abajo es menor que la de arriba, es decir, que en el boxplot de abajo los datos tienen aún más marcada una asimetría positiva en los datos, el percentil 75 es mayor al de arriba y hay mayor cantidad de outliers. Esto indica que a pesar que hubo una mejoría en la mayoría de los alumnos, los alumnos que previamente estaban con moderados niveles de depresión ahora están peor. Los valores atípicos con mayor depresión previo al estudio son los estudiantes con ids 17 y 27, mientras que los de mayor depresión posterior al estudio son los que tienen ids 23, 17 y 13.

Sin embargo, también hubo algunos alumnos que se dieron de baja durante el estudio, porque se encontraron 46 encuestas respondidas en la previa al estudio, y 38 respuestas a la encuesta al finalizar el período.

a. Depresión y desempeño

En esta sección se pondrá en análisis la relación que hay entre la variable de depresión y sus notas. Por lo que se realizó el siguiente test de Pearson que se observa en la *Figura 47*.

```
Pearson's product-moment correlation
data: desem_vs_depr$gpaall and desem_vs_depr$depre_score_post
t = -2.6359, df = 25, p-value = 0.01421
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.7189519 -0.1049325
sample estimates:
cor
-0.4663503
```

Figura 47 - Test de Pearson

Se pone en prueba la relación entre el promedio general y el score de depresión post estudio, el cual dio un p significativo (0,01421), lo que significa que hay una relación entre las variables, sin embargo, el índice de correlación dió bajo (-0,4663), por lo que hay que probar si existe otro modelo que ajuste mejor, se procede a realizar el test de Spearman (véase *Figura 48*).

Spearman's rank correlation rho

```
data: desem_vs_depr$gpaall and desem_vs_depr$depre_score_post
S = 4941.4, p-value = 0.006781
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
-0.5083495
```

Figura 48 - Test de Spearman

Al hacer la prueba, el p-valor sigue siendo bajo (0,0067), y el R² dió lo suficientemente alto (-0,5083) como para hacer una regresión logarítmica que podemos visualizar en la *Figura 49*.

```
Call:
lm(formula = log10(desem_vs_depr$gpaall) ~ desem_vs_depr$depre_score_post)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.11941 -0.01415  0.01559  0.03071  0.05427 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.563648  0.014933 37.746 <2e-16 ***
desem_vs_depr$depre_score_post -0.004269  0.001654 -2.581  0.0161 *  

```

Figura 49

Esto se lee como:

$$\ln(y) = 0,5636 - 0,004269 x$$

Es decir que **por cada 1% que sube su score de depresión, los promedios totales de las carreras de los estudiantes bajan en 0,004 unidades.**

Ahora, se repite el análisis pero esta vez con las notas de los estudiantes del semestre de estudio. Realizamos el test de correlación de Pearson (véase *Figura 50*) para evaluar si hay relación entre las notas del semestre con el nivel de depresión de los estudiantes.

Pearson's product-moment correlation

```
data: desem_vs_depr$gpa13s and desem_vs_depr$depre_score_post
t = -2.4016, df = 25, p-value = 0.02408
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.69811657 -0.06338001
sample estimates:
cor
-0.4329659
```

Figura 50 - Test de Pearson

El p-value que devolvió la prueba es de 0,02408, es decir, significativo. Sin embargo, el R² es bajo para los requisitos establecidos. Por lo que hay que ver si se puede ajustar el modelo a una regresión logarítmica. Para eso se utiliza la fórmula que se ve a continuación en la *Figura 51*.

```

Spearman's rank correlation rho

data: desem_vs_depr$gpa13s and desem_vs_depr$depre_score_post
S = 4641.2, p-value = 0.03059
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
-0.4167357

```

Figura 51 - Test de Spearman

Al observar el p-valor de la *Figura 51*, podemos ver que ninguno de los modelos que hay disponibles se ajusta lo suficientemente bien para poder hacer una regresión, sin embargo **se afirma la relación entre las notas de los estudiantes en el semestre del estudio con la depresión.**

Al confirmar que existe una relación entre la depresión de los alumnos con su desempeño, en las próximas secciones se verá en qué otros aspectos afecta el nivel de depresión en la vida de los estudiantes.

b. Depresión y el uso del Piazza

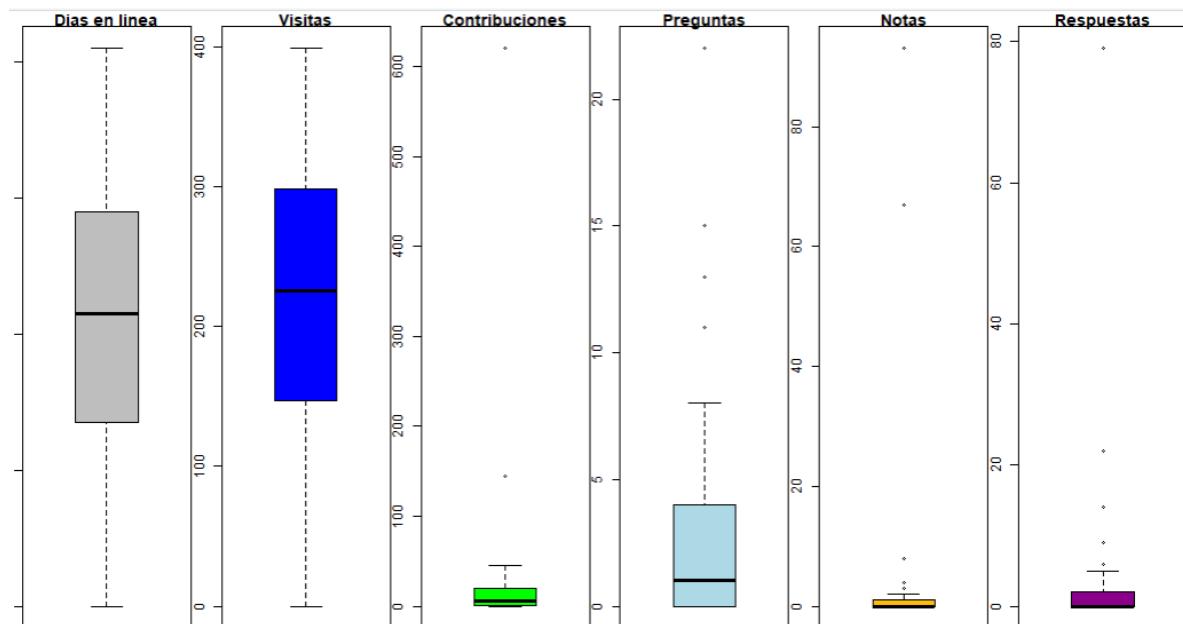
Como el objetivo de esta sección era poder ver si hay alguna correlación entre la depresión y otras variables, por lo que se tendrá que ver en detalle la actividad que tienen los alumnos en el Piazza. En caso que hubiese correlación entre las variables, se podría llegar a inferir la depresión de un alumno con ver su actividad en la plataforma.

Antes que nada, se reconoce que habían 10 estudiantes faltantes de los 60, por lo tanto, el análisis de la actividad en el Piazza se basará en los datos de los 50 alumnos con los que se cuenta. En el resumen estadístico de los datos son los que se pueden ver en la *Figura 52* a continuación.

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
days_online	0	1	42.5	20.2	0	27	43	58	82	
views	0	1	210.	97.8	0	147	226	299	399	
contributions	0	1	25.6	89.5	0	1	6	20	620	
questions	0	1	3.24	5.08	0	0	1	4	22	
notes	0	1	3.82	16.2	0	0	0	1	93	
answers	0	1	3.41	11.7	0	0	0	2	79	

Figura 52- Estadísticos básicos del Piazza

Luego, se realizaron boxplots (véase *Figura 53*) para poder analizar si hay casos atípicos en cada tipo de interacción dentro del Piazza. Lo que se puede observar es que en las variables de “dia en línea” (número de días que el estudiante se conectó a la página de clase de CS65 Piazza) y la de “visitas” (número de publicaciones que el estudiante ha visto), los datos se distribuyen de una forma simétrica y de forma muy similar. Esto hace pensar en la hipótesis de que hay posibilidad de que haya una correlación positiva entre ambas variables. Para verificar esto se realizó un test de hipótesis para la correlación que se puede ver en la *Figura 54*.



5

Figura 53- Boxplots de las variables del piazza

```
> cor(piazza$days_online, piazza$views)
[1] 0.8026388
> cor.test(piazza$days_online, piazza$views)

Pearson's product-moment correlation

data: piazza$days_online and piazza$views
t = 9.2254, df = 47, p-value = 4.044e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.6734363 0.8842588
sample estimates:
cor
0.8026388
```

Figura 54 - Test de hipótesis de relación entre variables

⁵ Notas siendo anotaciones que hacen en el Piazza, no refleja ningún resultado al desempeño del alumno.

Lo que se puede concluir observando la *Figura 54* es que presentan una correlación positiva fuerte entre ambas variables (la misma es de 0.802) y que dado el valor del p-value la probabilidad de cometer un error de tipo I son bajas, no se tienen pruebas suficientes para rechazar la hipótesis nula, es decir de que hay correlación entre ambas variables. También, para visualizar esta fuerte correlación realizamos otro gráfico (véase *Figura 55*) que muestra en azul los datos de las visitas por día de cada alumno y en gris los días en línea en el Piazza. Las líneas horizontales corresponden a los promedios de ambas variables. Por lo tanto, se puede observar que en la mayoría de los casos, los alumnos que están por encima de la media en una variable también lo están en la otra y de igual manera se da con quienes se encuentran por debajo de la media.

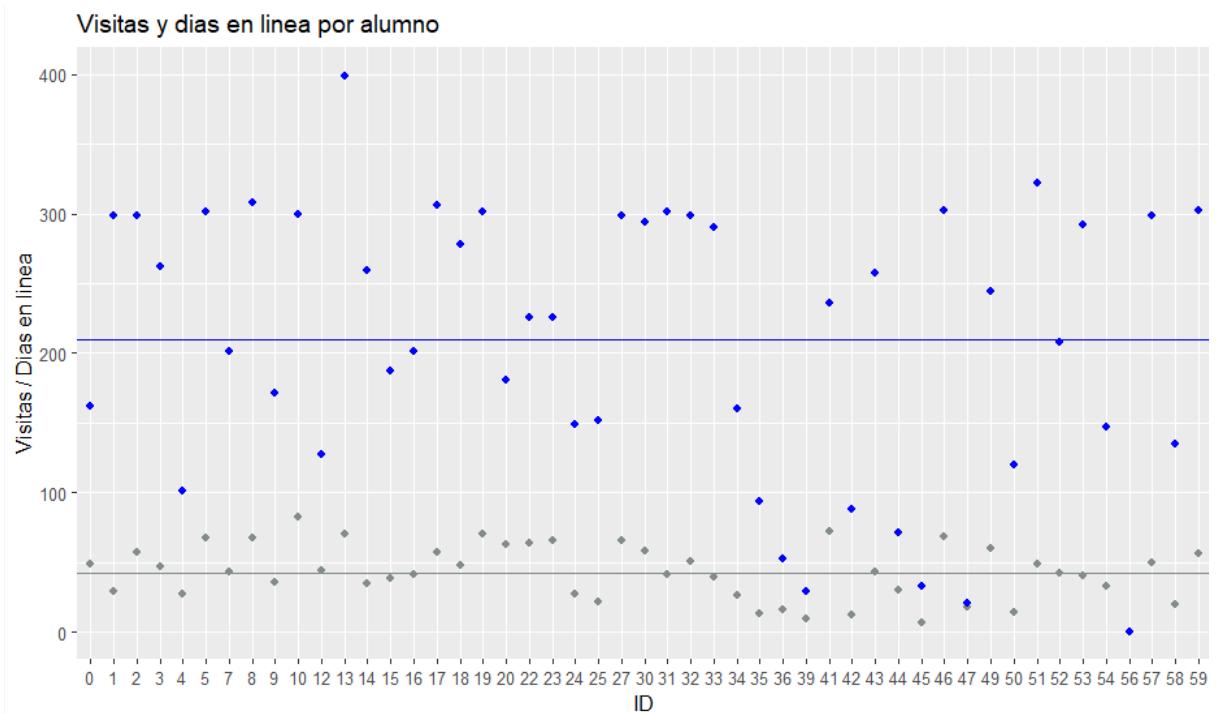


Figura 55- Diagrama de dispersión de visitas y días al Piazza

Volviendo a analizar la *Figura 53*, sería interesante evaluar si los valores atípicos en las variables de contribuciones, preguntas, notas y respuestas son los mismos alumnos y si son los que tienen las mejores notas. Para esto, primero se averigua quienes son los outliers (véase *Figura 56*).

```
> head(arrange(piazza, desc(contributions)), 2) #outliers contribuciones
# A tibble: 2 × 7
  uid  days_online views contributions questions notes answers
  <fct>     <int> <int>          <int>      <int> <int>   <int>
1 13        70    399           620         1    93    79
2 0         49    162           144         0    67    22
> head(arrange(piazza, desc(questions)), 4) #outliers preguntas
# A tibble: 4 × 7
  uid  days_online views contributions questions notes answers
  <fct>     <int> <int>          <int>      <int> <int>   <int>
1 8          67    308           45         22     0      5
2 3          47    262           30         15     2      6
3 41         72    236           27         15     0      2
4 14         35    259           16         13     1      0
> head(arrange(piazza, desc(notes)), 5) #outliers notes
# A tibble: 5 × 7
  uid  days_online views contributions questions notes answers
  <fct>     <int> <int>          <int>      <int> <int>   <int>
1 13        70    399           620         1    93    79
2 0         49    162           144         0    67    22
3 51         49    322           39         0     8    14
4 33         39    290           28         2     4     1
5 9          36    171           20         4     3     4
> head(arrange(piazza, desc(answers)), 5) #outliers respuestas
# A tibble: 5 × 7
  uid  days_online views contributions questions notes answers
  <fct>     <int> <int>          <int>      <int> <int>   <int>
1 13        70    399           620         1    93    79
2 0         49    162           144         0    67    22
3 51         49    322           39         0     8    14
4 50         14    120           10         0     0     9
5 3          47    262           30         15     2     6
```

Figura 56- Outliers de cada variable del Piazza

En la *Figura 56* se busco identificar a quienes se consideran outliers ordenando según las distintas columnas, acá se puede observar que en casi todas los outliers son el estudiante 13 y 0 en ese orden, pero cuando se los ordena para ver los outliers en base a la columna question se deja ver que los mismos son los estudiantes 8 y 3.

Lo siguiente fue ver la relación entre los scores de depresión y la utilización de piazza.

Para realizar este análisis se utilizaron las variables de depresión pre y post estudio, los días online, las veces que entro a la página (view) y una nueva variable llamada “participación”, el cual consta de la suma entre las variables “contributions”, “questions”, “notes” y “answers”.

Luego, se eliminaron los alumnos que no tenían respuesta alguna en ninguna de las variables e imputaron los que tenían por lo menos una variable sin valor.

Para hacer el análisis se analizará las correlaciones de la depresión pre estudio con las variables que corresponden al uso del piazza, las cuales si dan altas, hacer la regresión apropiada. Por último, repetir este proceso pero con la depresión pre estudio.

```
Pearson's product-moment correlation

data: matriz$`Score depresion previo estudio` and matriz$`Días online`
t = 0.48467, df = 47, p-value = 0.6302
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.2149383 0.3448781
sample estimates:
cor
0.0705205
```

Figura 57

```
Pearson's product-moment correlation

data: matriz$`Score depresion previo estudio` and matriz$`vistas`
t = 1.553, df = 47, p-value = 0.1271
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.06425775 0.47275738
sample estimates:
cor
0.2209308
```

Figura 58

```
Pearson's product-moment correlation

data: matriz$`Score depression previo estudio` and matriz$`Participacion`
t = -0.55344, df = 47, p-value = 0.5826
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.3536596 0.2053785
sample estimates:
cor
-0.08046507
```

Figura 59

Al realizar el test, podemos ver que los scores de depresión pre estudio no son influenciados ni por los días que estuvo online ($p=0,6302$, véase Figura 57), ni por las vistas ($p=0,1271$, véase Figura 58), ni por la participación ($p=0,5826$, véase Figura 59) de forma relevante, por lo que no se realizará la regresión.

```
Pearson's product-moment correlation

data: matriz$`Score depresion post estudio` and matriz$`Días online`
t = 1.2964, df = 47, p-value = 0.2012
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.1006466 0.4438164
sample estimates:
cor
0.1858087
```

Figura 60

Pearson's product-moment correlation

```
data: matriz$`Score depresion post estudio` and matriz$Vistas
t = 1.5324, df = 47, p-value = 0.1321
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.06718062 0.47047477
sample estimates:
cor
0.2181368
```

Figura 61

Pearson's product-moment correlation

```
data: matriz$`Score depresion post estudio` and matriz$Participacion
t = -0.72859, df = 47, p-value = 0.4699
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.3757126 0.1808910
sample estimates:
cor
-0.1056811
```

Figura 62

Lo mismo se repite con los scores de depresión post estudio, ya que las relaciones con las otras variables no son estadísticamente relevantes. Esto se puede ver en su test de correlación con los días online ($p=0.2012$, véase *Figura 60*), las vistas ($p=0.1312$, véase *Figura 61*) y la participación ($p=0.4699$, véase *Figura 63*). Es por esto, que se puede concluir que el **nivel de depresión que tiene el estudiante no influye en su actividad en el Piazza**.

c. Depresión y cantidad de horas del sueño

En esta sección se analizará la relación que hay entre la depresión y la cantidad de horas de sueño de los alumnos. Para eso, en primer lugar se hará un análisis exploratorio de los datos de sueño. Se pudo ver que la columna de horas presentaba alrededor de 15.5% datos faltantes por lo que se decidió imputarlos utilizando el algoritmo de missForest.

Haciendo un resumen estadístico de los datos de horas de sueño, se puede notar que la media es un poco mayor a la mediana, pero es casi igual (véase *Figura 63*).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.000	6.000	7.000	7.061	8.000	19.000

Figura 63

Se realizó un boxplot para ver si hay valores atípicos en las horas de sueño que se ve a continuación (véase *Figura 64*) . Se observa que hubo 107 valores atípicos de horas de sueño.

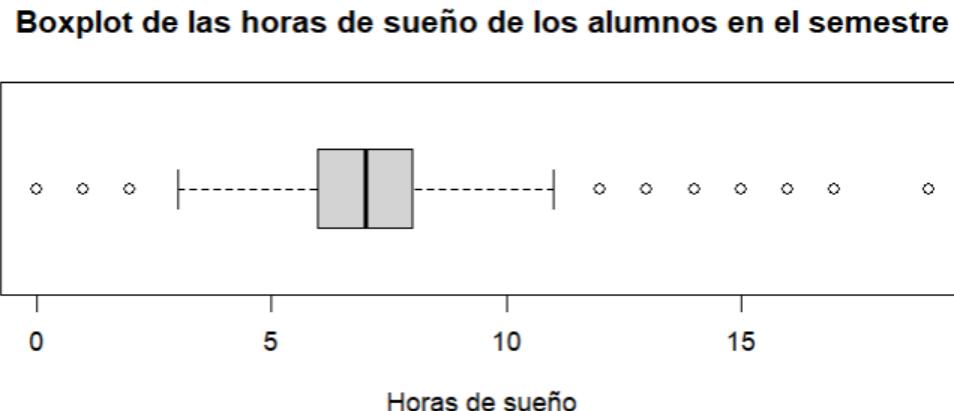


Figura 64

Luego, se quiso realizar un boxplot por día de semana para poder ver cómo van cambiando las horas de sueño según los días (véase *Figura 65*).

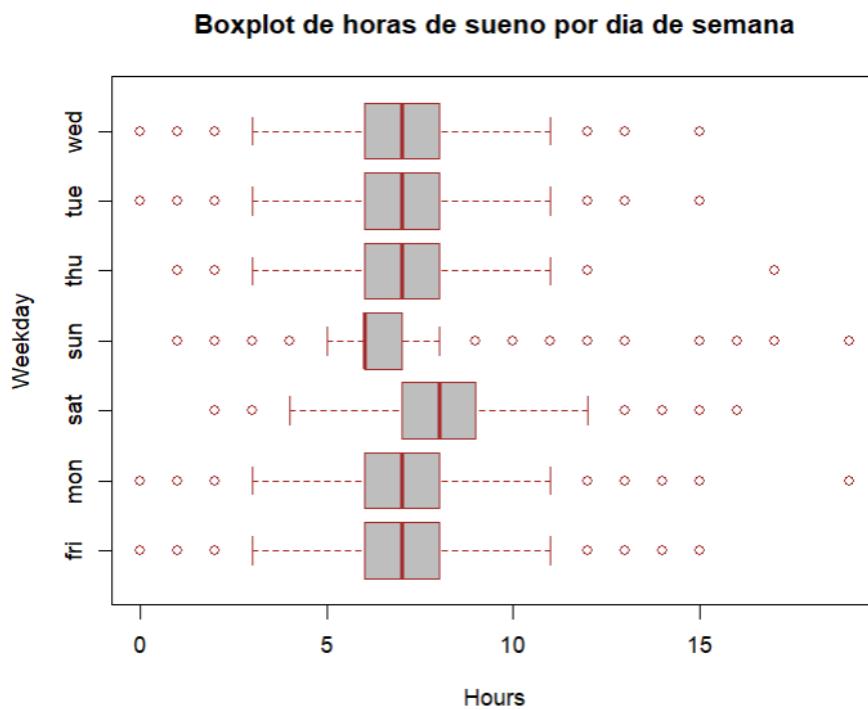


Figura 65

Como se puede ver en el gráfico de arriba, los datos se distribuyen de una forma uniforme para los casos de los días de semana (de lunes a viernes), esto se puede afirmar ya que el rango intercuartílico y los extremos son los mismos para todos esos días. Lo que sí cambia para estos días son los outliers. Por otra parte, en los días sábado y domingo se observa que cambia la mediana, siendo el sábado el día que más alumnos duermen más horas lo cual tiene sentido dado que es el fin de semana y, probablemente, no tengan ninguna obligación.

En todos los casos, hay varios outliers en nuestro estudio pero consideramos que se deben a los patrones de sueño inusuales que presentan los estudiantes universitarios. Debido a esto se los considera importantes para el estudio por lo que no se los va a descartar ni modificar.

Para poder hacer la correlación de las horas de sueño con la depresión, se decidió calcular los promedios de las horas de sueño de cada alumno en el semestre. Estos promedios se pueden visualizar en el siguiente boxplot (véase *Figura 66*), en donde se observan 2 outliers. Se busca a quienes pertenecen los promedios atípicos y pertenecen al de los alumnos con uid 17 (con un promedio de horas de sueño de 5.043478) y el alumno con uid 44 (con un promedio de horas de sueño de 8.841270).

Promedios de hora de sueño de los alumnos

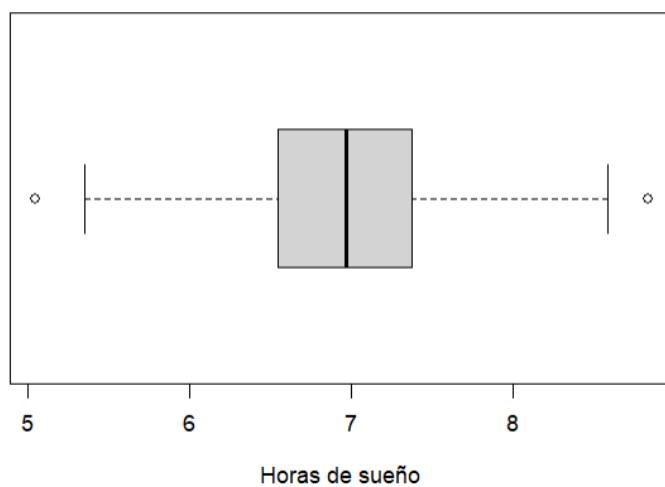


Figura 66

Por una parte, se nota que el alumno con uid 17 comienza el semestre con un score de depresión moderado de 13, y al final del semestre, pasa a tener un score de depresión de

18, es decir, asciende a la categoría de moderado severo de depresión. Por otro lado, analizando los niveles de depresión del alumno con uid 44, el mismo comienza el período de estudio con un score de depresión 1, y finaliza el período con un score de 2, lo que significa que mantuvo un nivel óptimo de salud mental al mantenerse en la categoría de depresión mínima. Es decir, se puede suponer que hay una relación entre horas de sueño y depresión, ya que el alumno con menos horas promedio de sueño tiene un nivel elevado de depresión y el alumno con mayor horas de sueño promedio tiene un nivel reducido de depresión. Sin embargo, esta correlación se va a evaluar con un test de correlación entre el promedio de sueño de los estudiantes y el score de depresión al finalizar el estudio (véase *Figura 67*).

```
Pearson's product-moment correlation

data: provisorio$promedio and provisorio$score_post
t = 2.0417, df = 47, p-value = 0.04681
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.004602076 0.524526468
sample estimates:
cor
0.2854293
```

Figura 67 - Test de correlación de Pearson

Analizando el test de hipótesis de Pearson, como el p-valor es menor a 0.05 se puede suponer que hay relación entre las variables de estudio. Luego, observando el coeficiente de correlación R^2 , como éste es menor a 0.5 se puede asumir que la relación entre las variables es no lineal, entonces se procede a proponer un modelo de regresión logarítmica. Para ver si la regresión se puede ajustar a una logarítmica, se hizo un test de correlación usando el método spearman (véase *Figura 68*), y como se puede ver el p-valor es mayor a 0.05, no se puede asumir que se puede ajustar la regresión a una logarítmica.

```
Spearman's rank correlation rho

data: provisorio$promedio and provisorio$score_post
S = 15701, p-value = 0.1706
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.1989255
```

Figura 68 - Test de correlación de Spearman

Teniendo ya estos datos, si bien no se puede concluir que dicha relación se ajusta a una regresión lineal ni a una regresión logarítmica, si se puede aseverar que **existe una relación entre el nivel de depresión y el promedio de las horas de sueño de los estudiantes.**

d. Depresión e interacciones sociales

En esta sección se evaluará si existe una relación entre la depresión y las interacciones sociales de los estudiantes. Para eso, se estará utilizando una tabla dentro de la base de datos que registró las conversaciones de los estudiantes durante todo el período de estudio.

Para poder trabajar con las conversaciones de los estudiantes, se tuvo que procesar los datos para obtener así una tabla que contenga los uid, luego el promedio de los minutos conversados por día por los estudiantes y el promedio de la cantidad de conversaciones que tuvo por día. Con toda la información recopilada se pudo graficar los boxplots que se pueden observar en las *Figura 69* y *Figura 70*.

Promedios de minutos conversados por estudiante por día

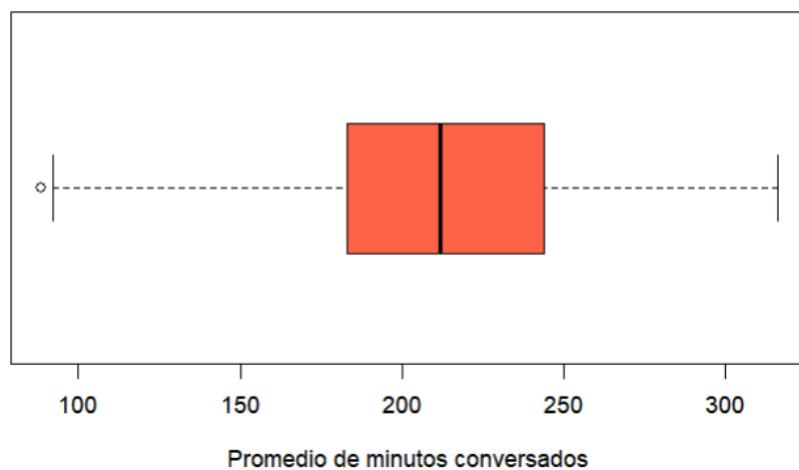


Figura 69

Como se puede observar en la *Figura 69*, solo hay un valor atípico que es el alumno con uid 39. Más adelante en esta sección, se evaluará el nivel de depresión de este alumno.

Promedios de la frecuencia de conversaciones por estudiante por día

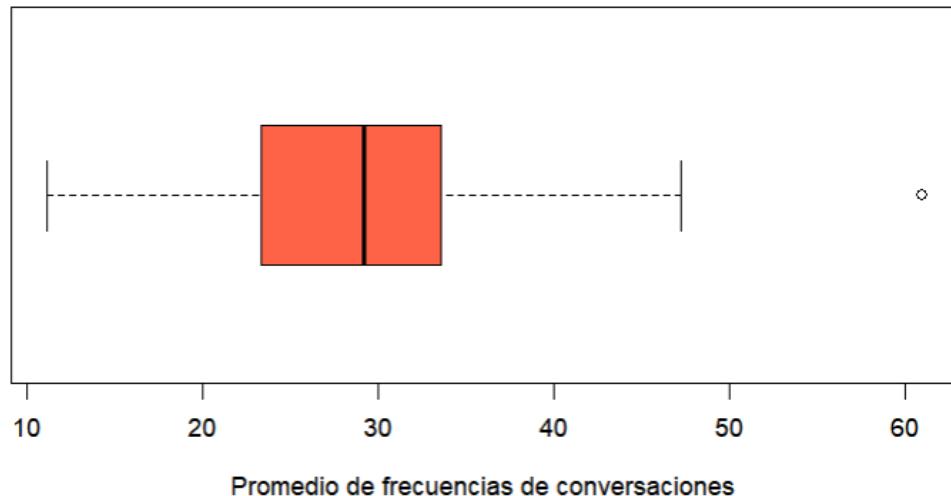


Figura 70

Por otro lado, se puede observar en la *Figura 70* que solo hay un valor atípico que es un estudiante que por día tiene un promedio de 60,97 cantidad de conversaciones por día durante todo el semestre. El uid que le pertenece a este alumno es el 59. Más adelante en esta sección, se evaluará el nivel de depresión de este alumno.

Otro enfoque que se le quiso dar al estudio, fue sumar el tiempo de conversaciones que tiene cada alumno por día durante el período de estudio. El resumen estadístico inicial de los datos es el siguiente que se puede observar en la *Figura 71*.

```
tiempo_conv_en_el_dia
Min. : 0.8333
1st Qu.: 155.1333
Median : 286.6000
Mean   : 299.5492
3rd Qu.: 423.3000
Max.   : 1128.2500
```

Figura 71 - Estadísticas básicas de la variable tiempo conversado por dia

En base a la *Figura 71* se puede observar que la muestra presenta una leve asimetría negativa y que, probablemente, se deba a algunos outliers en el bigote superior. En el resumen falta la desviación estándar que es igual a 181.8184. El coeficiente de variación es $181.8184/299.54 = 0.6069$, que implica una variabilidad moderada de los datos.

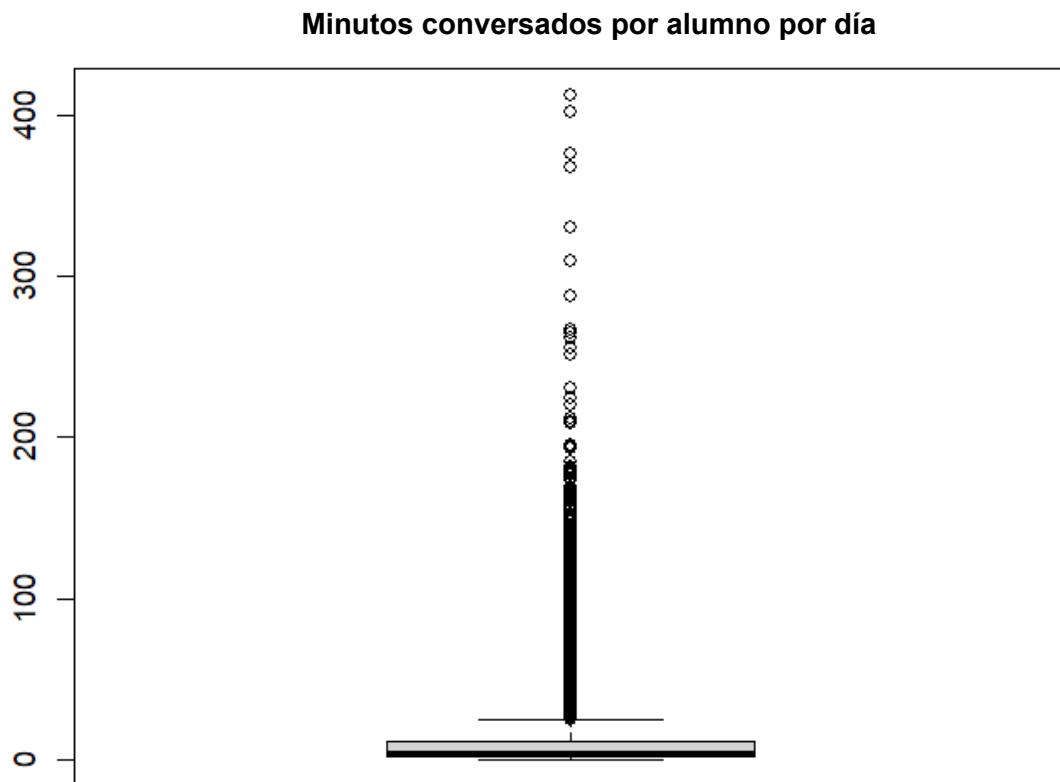


Figura 72 - Boxplot de minutos conversados por dia por alumno

Como se puede ver en la *Figura 72*, el boxplot presenta una gran cantidad de outliers en la parte superior de nuestro gráfico. Se considera que las observaciones que superen las 16 horas habladas en un día son outliers sistemáticos en los que el software que registra las conversaciones falló por lo que se decidió quitarlos de la muestra. Dado que la muestra es grande (alrededor de 2700 observaciones) no se cree que afecte al posterior análisis.

Diagrama de densidad / Histograma de minutos conversados en un día por estudiante

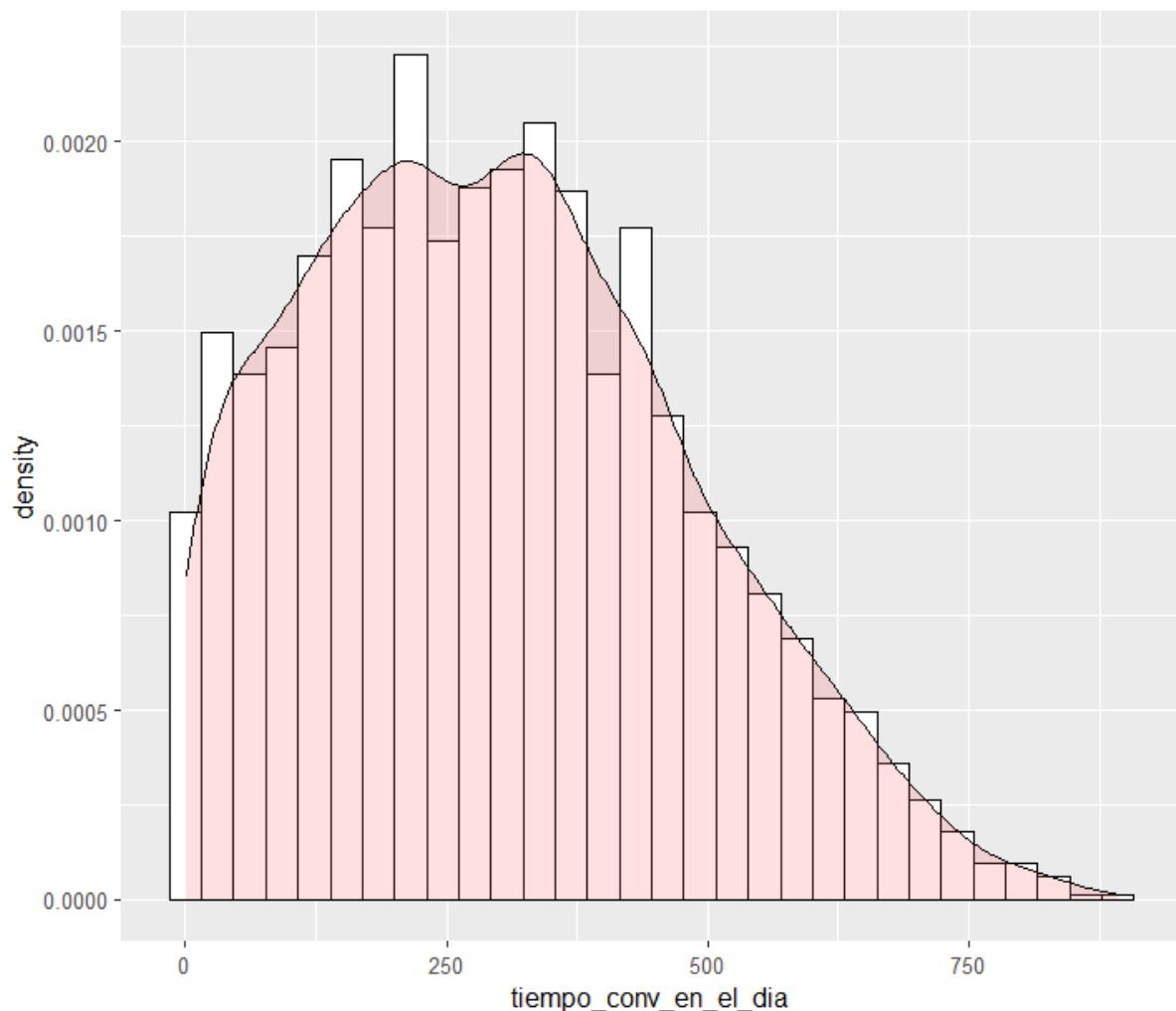


Figura 73 - Histograma del tiempo conv por dia

En la *Figura 73* se puede observar un histograma con las frecuencias de los datos de los minutos conversados por día por alumno. Hay varias conclusiones que se pueden rescatar al observar este gráfico. En primer lugar, se observa que la mayoría de las observaciones se encuentran entre 0 min - 500 min, y en segundo lugar, los datos no se distribuyen normalmente, tienen una asimetría positiva.

Una vez realizado el análisis exploratorio de datos de las conversaciones que tienen los alumnos, se analiza si existe alguna correlación fuerte entre la depresión y el promedio de los minutos conversados por día por los estudiantes y la depresión y el promedio de la cantidad de conversaciones que tuvo por día.

```
> cor.test(conv_vs_depr$score_post, conv_vs_depr$Tiempo_conversacion_prom)
Pearson's product-moment correlation

data: conv_vs_depr$score_post and conv_vs_depr$Tiempo_conversacion_prom
t = -0.91765, df = 47, p-value = 0.3635
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.3989820 0.1542823
sample estimates:
cor
-0.13267

> cor.test(conv_vs_depr$score_post, conv_vs_depr$Freq_prom_conversacion)
Pearson's product-moment correlation

data: conv_vs_depr$score_post and conv_vs_depr$Freq_prom_conversacion
t = -1.498, df = 47, p-value = 0.1408
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.4666583 0.0720470
sample estimates:
cor
-0.2134744
```

Figura 74

Como se puede observar en la *Figura 74*, en ambos test de correlación de Pearson los p-valores dan mayores a 0.05, por lo tanto se asume que no hay relación entre la depresión y el promedio de los minutos conversados por día por los estudiantes y la depresión y el promedio de la cantidad de conversaciones que tuvo por día. Es decir, **no hay relación entre la depresión y el nivel de interacciones sociales de los estudiantes.**

Con respecto al análisis del outlier que se pudo observar anteriormente, el alumno con uid 39 que fue el alumno con menor interacción social, pasó de tener un score de depresión de 3 a no tener respuesta en la encuesta de depresión al finalizar el período. Por lo que no es posible asumir que hay relación entre la poca interacción y la depresión, porque faltan datos. Por otra parte, el otro valor atípico que se puede visualizar, el alumno con uid 59 que fue el que en promedio tuvo más conversaciones por día en el período, presentó un score de depresión de 5 al iniciar el período y de 7 al finalizar. Aunque su nivel de depresión aumenta durante el semestre, sigue siendo un valor mínimo.

e. Depresión y el nivel de soledad

En esta sección se evaluará la relación entre depresión y nivel de soledad de los alumnos de la muestra. Pareció importante evaluar esta relación ya que si es que existe relación entonces se podría plantear algún tipo de política dentro de la universidad para que los estudiantes se sientan acompañados y asistidos en todo momento.

Para realizar esto, se tratará de encontrar correlaciones entre los niveles de depresión de cada participante y sus niveles de soledad obtenidos en la encuesta que aparece en la base de datos.

Teniendo en cuenta que ya se cuenta con el análisis exploratorio de datos de la variable depresión, se procede a realizar un análisis exploratorio de datos general a la variable de soledad. Para eso, se utilizarán los resultados de una encuesta llamada Loneliness Scale⁶, realizada al inicio y final del estudio. La encuesta consta de una serie de 20 preguntas que se responden con las siguientes opciones:

- Never
- Rarely
- Often
- Sometimes

Las preguntas que se realizan en la encuesta se pueden visualizar en la *Figura 75* a continuación.

Scale:

INSTRUCTIONS: Indicate how often each of the statements below is descriptive of you.

C indicates "I often feel this way"

S indicates "I sometimes feel this way"

R indicates "I rarely feel this way"

N indicates "I never feel this way"

1. I am unhappy doing so many things alone	O S R N
2. I have nobody to talk to	O S R N
3. I cannot tolerate being so alone	O S R N
4. I lack companionship	O S R N
5. I feel as if nobody really understands me	O S R N
6. I find myself waiting for people to call or write	O S R N
7. There is no one I can turn to	O S R N
8. I am no longer close to anyone	O S R N
9. My interests and ideas are not shared by those around me	O S R N
10. I feel left out	O S R N
11. I feel completely alone	O S R N
12. I am unable to reach out and communicate with those around me	O S R N
13. My social relationships are superficial	O S R N
14. I feel starved for company	O S R N
15. No one really knows me well	O S R N
16. I feel isolated from others	O S R N
17. I am unhappy being so withdrawn	O S R N
18. It is difficult for me to make friends	O S R N
19. I feel shut out and excluded by others	O S R N
20. People are around me but not with me	O S R N

Scoring:

Make all O's =3, all S's =2, all R's =1, and all N's =0. Keep scoring continuous.

Self Report Measures for Love and Compassion Research: *Loneliness and Interpersonal Problems*  Fetzer Institute

Figura 75 - Composición de la encuesta y valor de las respuestas

Cada una de las respuestas tienen asignado un valor numérico que se utiliza posteriormente para calcular un score que va a indicar un aproximado del nivel de soledad que tiene el encuestado. Las categorías de soledad dependen del score y son las siguientes:

- Leve (25 - 34)
- Moderada (35 - 46)

A continuación, en la *Tabla 3* se mostrará la cantidad de estudiantes en cada uno de estos niveles de soledad previo y posterior al estudio.

Loneliness Scale	Leve	Moderada
Score	25 - 34	35 - 46
Número de estudiantes pre-survey	38	3
Número de estudiantes post-survey	34	7

Tabla 3

Es importante aclarar que la misma cantidad de participantes contestaron la encuesta pre y post estudio, en total fueron 41 alumnos de 60. Se ha utilizado la *Tabla 3* para crear el siguiente boxplot (véase *Figura 76*) que permite visualizar las diferencias entre las medianas y las distribuciones de los datos de los niveles de soledad de los estudiantes previo y posterior al período de análisis.

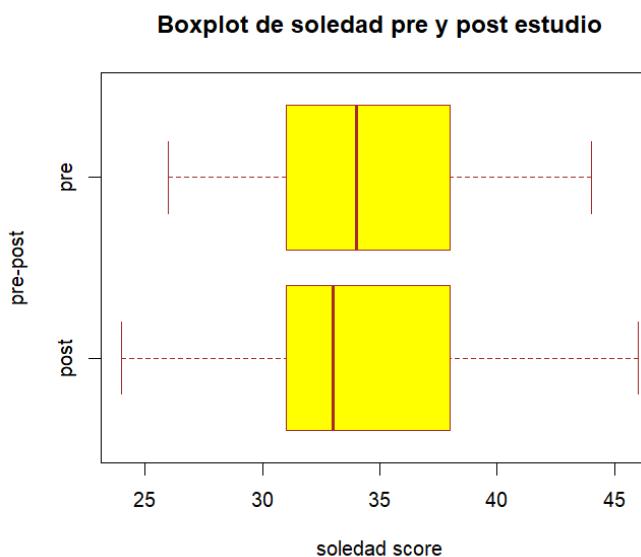


Figura 76 - Boxplot del score de soledad pre y post estudio

En la *Figura 76* se puede ver que ambas muestras son bastante similares con la diferencia de que post estudio la mediana en el score de soledad es menor y tiene una mayor dispersión en sus datos, esto quiere decir que se presentaron un rango más variado en las respuestas y respectivos puntajes de los alumnos. Asimismo, no se puede observar ningún valor atípico en nuestros boxplots, por lo tanto, no existe un estudiante que se sienta extremadamente sólo. Se supone que esta disminución de la mediana de los valores de soledad que obtenemos en el post-estudio, es una cuestión que se da con naturalidad entre los alumnos, ya que al finalizar el semestre ya se conocen más entre sí.

Una vez realizado el breve análisis exploratorio de los datos de soledad, se quiso averiguar si existía una correlación entre la depresión y la soledad de los participantes de la muestra, por lo que para analizar dicha correlación, se realiza una correlación de Pearson (*Figura 77*):

```
Pearson's product-moment correlation

data: as.numeric(df_de_sol_imp$soledad_pre) and as.numeric(df_de_sol_imp$score_pre)
t = 3.6466, df = 58, p-value = 0.0005699
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.1998648 0.6180223
sample estimates:
cor
0.4318717
```

Figura 77

Como se hizo anteriormente, si se analiza el test de hipótesis de Pearson, como el p-valor es menor a 0.05 se puede suponer que hay relación entre las variables de estudio. Cuando se observa el coeficiente de correlación R^2 , se nota que éste es menor a 0.5 por lo que se puede asumir que la relación entre las variables es no lineal y se pasa a analizar un modelo de regresión logarítmica. Para esto, se hizo un test de correlación usando el método spearman y, como se puede ver en la *Figura 78*, como el p-valor es mayor a 0.05 no se puede asumir que se puede ajustar la regresión a una logarítmica.

```
Spearman's rank correlation rho

data: as.numeric(df_de_sol_imp$soledad_pre) and as.numeric(df_de_sol_imp$score_pre)
S = 16849, p-value = 1.225e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.5318525
```

Figura 78

Como se observa arriba, como el p-valor es menor a 0.05, entonces se puede asumir que una regresión logarítmica se ajusta bien para este modelo.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.490421	0.010075	147.931	< 2e-16 ***
df_de_sol_imp\$score_pre	0.005358	0.001463	3.663	0.000542 ***

Figura 79

La Figura 79 se interpreta como:

$$\ln(y) = 0.05358 + 1.490421x$$

Es decir que **por cada 1% que sube el score de soledad, el score de depresión sube en 1,490421 unidades.**

7. Relación entre desempeño y sueño

Ya se ha confirmado que existe correlación entre el sueño y la depresión del alumno, pero en esta sección se evaluará si existe alguna correlación notable entre las variables de desempeño de los alumnos con el promedio de la cantidad de horas de sueño por alumno. Para eso, se va a utilizar las notas promedio del semestre del estudio para comparar con las horas de sueño promedio.

```
Pearson's product-moment correlation

data: as.numeric(as.character(sleep_vs_desempeño$promedio)) and as.numeric(as
character(sleep_vs_desempeño$gpa13s))
t = 0.89274, df = 28, p-value = 0.3796
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.2062703 0.4968521
sample estimates:
cor
0.1663617
```

Figura 80

Como se puede observar en la Figura 80, como el p-valor es mayor a 0.05 se concluye en que **no hay relación entre la cantidad de horas dormidas y desempeño académico.**

8. Relación entre desempeño e interacciones sociales

En esta sección se va a evaluar la relación entre el desempeño de los estudiantes y la frecuencia con la que hablan con otros estudiantes, ya que de esta forma se podría ver si influye negativamente si hablan mucho con otras personas.

Como se puede observar en la *Figura 81* a continuación, como ambos p-value son mayores a 0.05, se puede asumir que no hay relación entre el desempeño con la frecuencia de sus conversaciones con otros estudiantes. Es decir, **no existe relación evidente entre el desempeño de los estudiantes con el nivel de vida social que llevan.**

```
> cor.test(desempeño_vs_conversacion$gpa13s, desempeño_vs_conversacion$Freq_prom_conversacion)
Pearson's product-moment correlation

data: desempeño_vs_conversacion$gpa13s and desempeño_vs_conversacion$Freq_prom_conversacion
t = 0.69545, df = 28, p-value = 0.4925
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.2412889 0.4685792
sample estimates:
cor
0.1303079

> cor.test(desempeño_vs_conversacion$gpa13s, desempeño_vs_conversacion$Tiempo_conversacion_prom)
Pearson's product-moment correlation

data: desempeño_vs_conversacion$gpa13s and desempeño_vs_conversacion$Tiempo_conversacion_prom
t = 0.87867, df = 28, p-value = 0.3871
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.2087806 0.4948739
sample estimates:
cor
0.16381
```

Figura 81

9. Relación entre soledad e interacciones sociales

Anteriormente se ha confirmado que el nivel de soledad afecta negativamente en el desempeño de los estudiantes, sin embargo, no se ha encontrado una correlación directa entre el nivel de interacciones sociales y el desempeño académico. Es por eso que en esta sección se evaluará si existe una relación entre soledad e interacciones sociales, porque si la hay, entonces se podría recomendar aplicar una política o generar programas dentro de

la universidad Dartmouth que incentiven las interacciones sociales (para, de esta manera aumentar el nivel de soledad, y así aumentar el desempeño de los alumnos).

```
Pearson's product-moment correlation

data: conversaciones_vs_soledad$Tiempo_conversacion_prom and conversaciones_vs_soledad$soledad_post
t = 0.5994, df = 33, p-value = 0.553
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.2376873 0.4224171
sample estimates:
cor
0.1037795

> cor.test(conversaciones_vs_soledad$Freq_prom_conversacion, conversaciones_vs_soledad$soledad_post)

Pearson's product-moment correlation

data: conversaciones_vs_soledad$Freq_prom_conversacion and conversaciones_vs_soledad$soledad_post
t = 0.36661, df = 33, p-value = 0.7163
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.2754030 0.3886857
sample estimates:
cor
0.06368866
```

Figura 82

Como se puede observar en la *Figura 82*, como el p-valor en el test de correlación entre soledad y tiempo de conversaciones y soledad y frecuencia de conversaciones es muy elevado, entonces se asume que, sorprendentemente, **no existe relación entre el nivel de interacciones sociales con la soledad.**

10. Relación entre soledad y desempeño

Se somete a prueba a estas variables, ya que le dará más solidez a varias conclusiones previas. Anteriormente se afirmó la relación de los niveles de depresión tanto con las calificaciones como con la soledad, por lo que la soledad se justificaría como un problema raíz de varios problemas que enfrenta un alumno.

```
Pearson's product-moment correlation

data: as.numeric(as.character(soledad_vs_desempeño$soledad_post)) and as.numeric(as.character(soledad_vs_desempeño$gpa13s))
t = 0.6811, df = 21, p-value = 0.5033
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.2822963 0.5272633
sample estimates:
cor
0.1470127
```

Figura 83

Sin embargo, al hacer la prueba (véase *Figura 83*) queda en evidencia que no existe tal vínculo, ya que el p-value devuelto por el test de correlación de coeficientes es muy alto ($p=0,5033$). **Es decir, no hay relación entre los niveles de soledad y el desempeño académico del alumno.**

11. Outliers de peor desempeño

Luego de buscar todas las relaciones entre las variables que parecían de mayor importancia, consideramos menester retomar un objetivo que se tenía en mente responder al inicio del presente informe que fue determinar cuáles son las características que tienen en común los alumnos con peores notas. Para eso, se van a tomar como referencia los valores atípicos que se pudieron observar con anterioridad en la *Figura 3*, los alumnos con ID = 1, 46 y 52. La idea es observar cómo se comportan estos tres alumnos en relación a su nivel de depresión previo y posterior al período de estudio, su nivel de soledad previo y posterior al período de estudio, la cantidad de materias que cursó, la cantidad de tareas totales que le fue asignado, su nivel de interacciones sociales, su nivel promedio de estrés en el semestre, y su uso del Piazza.

a. Características del estudiante con ID = 1

Estudiante con ID = 1		
Cantidad de materias cursadas		3
Promedio de niveles de estrés		2.305556
Cantidad de tareas totales		10
Niveles de depresión	Pre-estudio	5 (Minor)
	Post-estudio	4 (Minimal)
Uso del Piazza	Días online	29
	Vistas	299
	Contribuciones	5
	Preguntas	1

	Notas	1
	Respuestas	0
Promedio de sueño	6.968 ⁷	
Nivel de interacciones sociales	Minutos de conversación promedio	204.01
	Promedio de frecuencia de conversación	39.61
Niveles de soledad	Pre-estudio	Leve
	Post-estudio	Leve

Tabla 4

Observando los resultados de la *Tabla 4*, se puede ver que ninguno de los valores es un valor atípico en su propia categoría, es más, todos caen dentro de su rango intercuartílico de su distribución. Quizás lo único rescatable es que cursa 3 materias, y ya habiendo confirmado que cuantas más materias cursa un alumno peor desempeño tiene, se asume que en el semestre de estudio disminuyó su rendimiento por eso. Previamente, se estableció que el 51% de los estudiantes cursan tres materias y no fueron outliers, por ende se puede asumir que al alumno con ID = 1 le costaron las materias cursadas.

b. Características del estudiante con ID = 46

Estudiante con ID = 46		
Cantidad de materias cursadas	3	
Promedio de niveles de estrés	1.925373	
Cantidad de tareas totales	17	
Niveles de depresión	Pre-estudio	10 (Minimal)
	Post-estudio	NA
Uso del Piazza	Días online	68
	Vistas	302
	Contribuciones	14

⁷ Y no tuvo ningún valor atípico en las horas dormidas de los alumnos en el semestre.

	Preguntas	7
	Notas	0
	Respuestas	1
Promedio de sueño	8.592 ⁸	
Nivel de interacciones sociales	Minutos de conversación promedio	282.95
	Promedio de frecuencia de conversación	35.31
Niveles de soledad	Pre-estudio	Leve
	Post-estudio	Leve

Tabla 5

Lo que se observa en la *Tabla 5*, es que el alumno con ID = 46 presenta valores que no son atípicos para su propia categoría. Lo único que quizás genera desconfianza es que hay un faltante de la respuesta a la encuesta de depresión posterior al período del análisis. También, se puede ver que este alumno cursa 3 materias, así que en ese caso se asume, de vuelta, que su rendimiento en el semestre de estudio disminuyó por esa razón o porque le costaron las materias elegidas.

c. Características del estudiante con ID = 52

Estudiante con ID = 52		
Cantidad de materias cursadas		3
Promedio de niveles de estrés		NA
Cantidad de tareas totales		16
Niveles de depresión	Pre-estudio	12 (Minimal)
	Post-estudio	15 (Minimal)
Uso del Piazza	Días online	42
	Vistas	208
	Contribuciones	0

⁸ Y no tuvo ningún valor atípico en las horas dormidas de los alumnos en el semestre.

	Preguntas	0
	Notas	0
	Respuestas	0
Promedio de sueño	7.0689 ⁹	
Nivel de interacciones sociales	Minutos de conversación promedio	91.93359
	Promedio de frecuencia de conversación	11.96923
Niveles de soledad	Pre-estudio	Leve
	Post-estudio	Leve

Tabla 6

Nuevamente, en la *Tabla 6*, se puede observar los valores que no son atípicos en sus categorías. No hay ningún valor fuera de lo esperado para el alumno con ID = 52. Lo único que se notó es que hubo un día en el que durmió una hora, pero no es preocupante considerando la cambiante vida de los estudiantes universitarios.

En conclusión, en los tres valores atípicos que bajaron su desempeño en el semestre de estudio, no se notó algún valor significativo en los hábitos o acciones de los estudiantes.

12. Análisis de correspondencias múltiples (MCA)

Se realizó un análisis de los alumnos a través del algoritmo de análisis de correspondencias múltiples. Esta técnica de análisis de datos es útil para detectar y representar estructuras subyacentes en la base de datos de StudentLife Dataset. Sirve, a su vez, para emparentar las variables categóricas de la base de datos.

Para realizar este análisis, se tuvo que crear un nuevo dataset que está compuesto por las siguientes columnas y sus respectivas categorías:

⁹ En este caso, el alumno presenta un día que duerme 1 hora (un valor atípico en la muestra de las horas de sueño de los estudiantes).

- "cs65 Categ": Indica promedio obtenido en la materia cs65 por los alumnos el cual puede ser A,B,C,D o F en base a el sistema americano de calificaciones.
- "gpa13s Categ": Indica promedio obtenido en el semestre del estudio por los alumnos el cual puede ser A,B,C,D o F en base a el sistema americano de calificaciones.
- "gpa all Categ": Indica el promedio total de los estudiantes el cual puede ser A,B,C,D o F en base a el sistema americano de calificaciones.
- "Stress Categoría": En base a las respuestas de los alumnos a las encuestas sobre estrés que le realizaron se categorizó según los cuantiles muestrales su respuesta con la posibilidad de caer en las siguientes 4 categorías: Alta, Moderada, Leve y Baja.
- "Soleded_post": En base a las respuestas de los estudiantes a las encuestas posterior al estudio sobre soledad que le realizaron se categorizó según los cuantiles muestrales su respuesta con la posibilidad de caer en las siguientes 4 categorías: Alta, Moderada, Leve y Baja.
- "Soleded_pre": En base a las respuestas de los estudiantes a las encuestas previo al estudio sobre soledad que le realizaron se categorizó según los cuantiles muestrales su respuesta con la posibilidad de caer en las siguientes 4 categorías: Alta, Moderada, Leve y Baja.
- "depre categ post": En base a las respuestas de los estudiantes a las encuestas posterior al estudio sobre depresión que le realizaron se categorizó según los cuantiles muestrales su respuesta con la posibilidad de caer en las siguientes 4 categorías: Alta, Moderada, Leve y Baja.
- "depre categ pre": En base a las respuestas de los alumnos a las encuestas previo al estudio sobre depresión que le realizaron se categorizó según los cuantiles muestrales su respuesta con la posibilidad de caer en las siguientes 4 categorías: Alta, Moderada, Leve y Baja.
- "flourish categ post": En base a las respuestas de los alumnos a las encuestas¹⁰ posterior al estudio sobre recursos psicológicos que le realizaron se categorizó según los cuantiles muestrales su respuesta con la posibilidad de caer en las siguientes 4 categorías: Alta, Moderada, Leve y Baja.
- "flourish categ pre": En base a las respuestas de los alumnos a las encuestas previo al estudio sobre recursos psicológicos que le realizaron se categorizó según los

¹⁰ Se realizaron encuestas de "Flourishing Scale", que es un breve cuestionario de 8 ítems del éxito autopercibido del encuestado en áreas importantes como las relaciones, la autoestima, el propósito y el optimismo. La escala proporciona una única puntuación de bienestar psicológico. La encuesta se puede encontrar en: https://ggsc.berkeley.edu/images/uploads/The_Flourishing_Scale.pdf

cuantiles muestrales su respuesta con la posibilidad de caer en las siguientes 4 categorías: Alta, Moderada, Leve y Baja

- "Categoría Actividad": En base a la actividad promedio por día de los estudiantes durante toda el cuatrimestre se crearon 4 categorías en base a los cuantiles muestrales las cuales son las siguiente: Alta, Moderada, Leve y Baja
- "Categoría Movilidad": En base a la movilidad promedio por día del estudiante durante toda el cuatrimestre se crearon 4 categorías en base a los cuantiles muestrales las cuales son las siguiente: Alta, Moderada, Leve y Baja
- "Categoría Sueño": En base a las horas de sueño promedio por día de los estudiantes durante toda el cuatrimestre se crearon 2 categorías en base a si durmieron más de 7 horas en promedio(Categoría = Suficiente) o no(Categoría = Insuficiente)
- "total materias": Se creó una variable de factor con la cantidad de materias que cursaron los alumnos durante el estudio.

Luego, se realizó un diagrama de los Eigenvalues de las dimensiones del algoritmo, para visualizar de cuántas dimensiones realizar el algoritmo de MCA. Como se observa en la *Figura 84*, se decidió utilizar las dos primeras dimensiones.

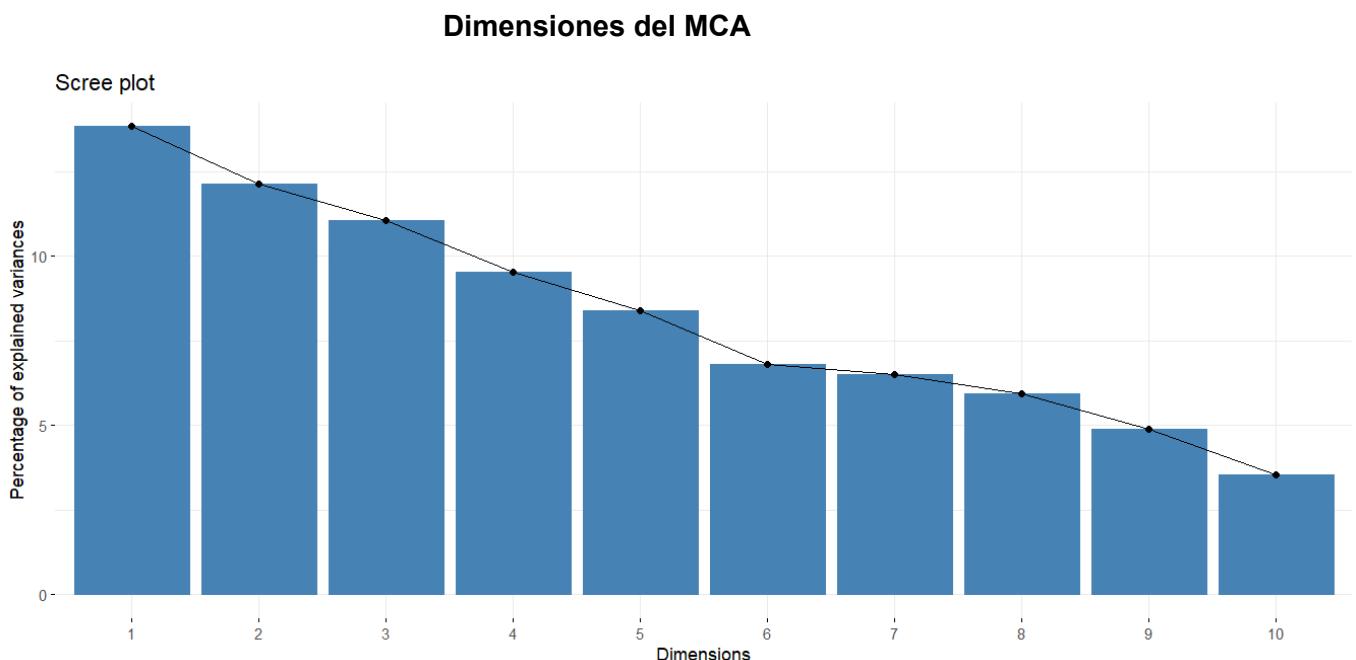


Figura 84: Eigenvalues de las dimensiones del algoritmo

Observando el MCA en la *Figura 86*, el mismo explica un total de 26% de la variabilidad de nuestro dataset. La relación entre las variables categóricas que se pusieron en juego explican un 26% de la forma en la que se varían los datos en nuestro dataset.

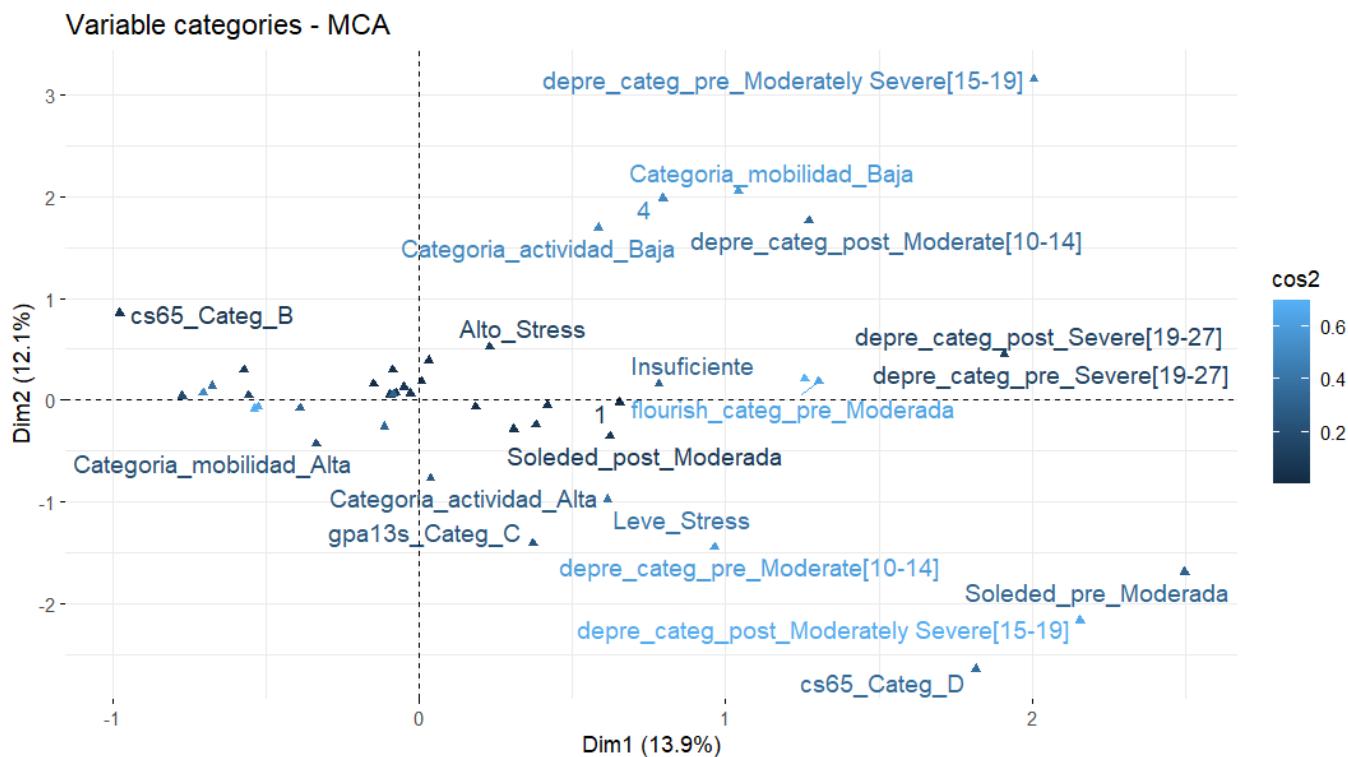


Figura 86: Análisis de correspondencias múltiples

a. Análisis de la primera dimensión

Analizando más en profundidad la primera dimensión del MCA, ésta es la que más variabilidad explica con un 13.9%.

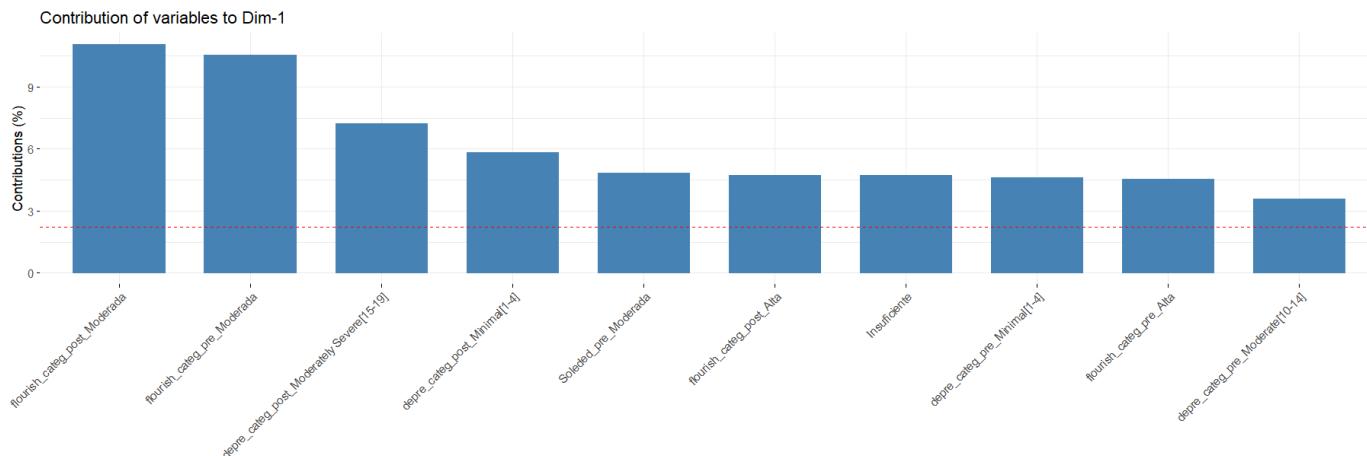


Figura 87

En base a la *Figura 87*, se puede observar que las categorías que más aportan a la variabilidad en el caso de esta dimensión son las siguientes:

1. Categoría moderada en la encuesta flourish post estudio (con una contribución de más del 9%).
2. Categoría moderada en la encuesta flourish pre estudio (con una contribución de más del 9%).
3. Depresión moderadamente severa post estudio (con una contribución de más del 6%).

Luego, se realizó una regresión lineal múltiple de primer orden que se observa en la *Figura 88* a continuación.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.5392	0.1281	-4.209	0.00024 ***
Df_MCA\$flourish_categ_postModerada	1.7972	0.2339	7.684	2.27e-08 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Residual standard error: 0.5871 on 28 degrees of freedom
 Multiple R-squared: 0.6783, Adjusted R-squared: 0.6668
 F-statistic: 59.04 on 1 and 28 DF, p-value: 2.272e-08

Figura 88

En base la regresión lineal de primer orden (véase *Figura 88*) se puede observar que la categoría moderada en la encuesta post bimestre de flourish es estadísticamente significativa con un valor p-valor mucho menor que 0.05. Esta categoría explica un 66.68%

de la variabilidad de la dimensión 1 con un error residual estándar de 0.58871 para el modelo.

Entonces, se puede concluir, en base a la *Figura 87* y *Figura 86*, que la dimensión 1 describe a un grupo de estudiantes que tiene un grado moderado de recursos psicológicos para su bienestar pre y post estudio. Al mismo tiempo, notamos que la tercera cualidad más que más aporta a este grupo es la depresión moderadamente severa post estudio.

b. Análisis de la segunda dimensión

En el caso del análisis para la segunda dimensión, ésta explica un 12.1% de la variabilidad.

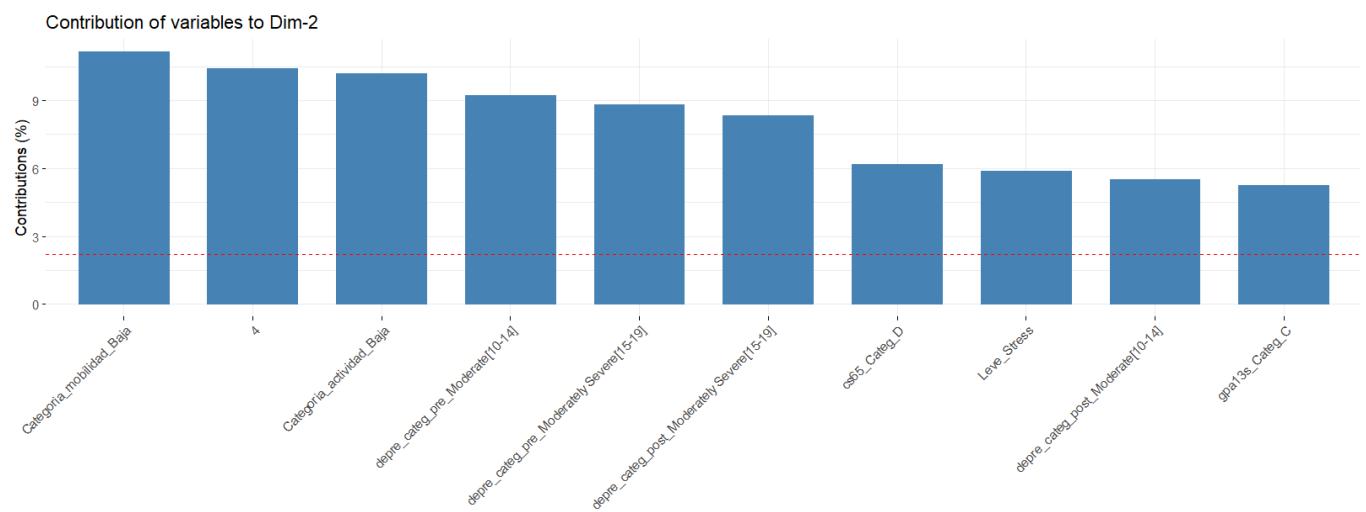


Figura 89

En base a la *Figura 89*, se puede observar que las categorías que más aportan a la variabilidad en el caso de esta dimensión son las siguientes:

1. Categoría movilidad baja (con una contribución de más del 9%)
2. 4 materias cursadas durante el estudio (con una contribución de más del 9%)
3. Categoría actividad baja (con una contribución de más del 9%)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.4331	0.2126	-2.037	0.0520 .
Df_MCA\$Categoria_mobilidadBaja	2.4830	0.4754	5.223	1.87e-05 ***
Df_MCA\$Categoria_mobilidadLeve	0.7307	0.3920	1.864	0.0737 .
Df_MCA\$Categoria_mobilidadModerada	0.1889	0.3154	0.599	0.5543

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7365 on 26 degrees of freedom

Multiple R-squared: 0.5299, Adjusted R-squared: 0.4756

F-statistic: 9.768 on 3 and 26 DF, p-value: 0.0001721

Figura 90

Se realizó una regresión lineal de primer orden (véase *Figura 90*) y se puede observar que la categoría de movilidad baja es estadísticamente significativa ya que da un valor p-valor mucho menor a 0.05. Esta categoría explica un 47.56% de la variabilidad de la dimensión 2 con un error residual estándar de 0.7365 para el modelo.

En base a la *Figura 89* y *Figura 90* se puede inferir que la dimensión 2 describe a un grupo de estudiantes que tuvo una movilidad baja durante el estudio, cursó muchas materias e hizo poca actividad física. Esto se podría atribuir a que vivieron un estilo de vida sedentario debido a la gran carga académica a la que se vieron expuestos.

13. Regresión lineal múltiple

Para esta sección, se usará una parte de las variables previamente descritas pero en formato numérico con la finalidad de encontrar un modelo de regresión lineal que explique el desempeño del alumno encuestado. Para encontrar la cantidad óptima de variables explicativas se utilizó el criterio de información de Akaike (AIC). En base a esto se encontró el siguiente modelo que se observa en la *Figura 91*.

```
Call:  
lm(formula = gpa13s ~ total_materias + promedio_locations + Media_actividad +  
    depre_score_post + soledad_pre, data = df_SinNa)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-0.82581 -0.21415  0.01911  0.25455  0.63802  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  7.68856   0.96312  7.983 4.45e-08 ***  
total_materias -0.50511   0.14751 -3.424  0.00232 **  
promedio_locations  0.07368   0.04777  1.542  0.13663  
Media_actividad -0.06244   0.01078 -5.791 6.71e-06 ***  
depre_score_post -0.03871   0.01676 -2.309  0.03026 *  
soledad_pre     -0.03924   0.02230 -1.760  0.09175 .
```

Figura 91

En base a la *Figura 91* se puede observar que el modelo de regresión lineal es estadísticamente significativo con un p-value mucho menor de 0.05. El mismo explica la variabilidad del promedio semestral en un 72.41% y tiene un error residual estándar de 0.3936 puntos.

El total de materias es estadísticamente significativo con un p-valor mucho menor de 0.05 y un error estándar de 0.14751. Si la cantidad de materias aumenta en la unidad y las demás variables se mantienen constantes el promedio semestral varía en -0.50511 puntos. Esto nos da sustento para decir que a más materias que cursa un alumno, posiblemente, peor va a ser su promedio.

La media de actividad (en minutos) realizada por el encuestado durante el cuatrimestre es una variable significativa con un p-valor menor a 0.05 y un error estándar de 0.01078. Si la media de actividad aumentará en 1 minuto y las demás variables se mantienen constantes el promedio semestral varía en -0.06244 puntos. Esto llama la atención dado que en una primera instancia pensamos que a más actividad mejor rendimiento académico por sus implicancias en la salud mental. Este resultado podría relacionarse a que para que a un alumno le vaya bien tiene que estar muchas horas por día quieto estudiando.

La variable depresión post estudio es estadísticamente significativa con un p-valor menor a 0.05 y un error estándar igual a 0.01676. Si su puntaje en la encuesta de depresión aumenta en una unidad y las demás variables se mantienen constantes el promedio semestral varía en -0.03871 puntos. Esto da a entender que niveles de depresión bajos (puntaje en la encuesta bajo) tienen efectos positivos en el rendimiento académico.

La variable soledad pre estudio es estadísticamente significativa con un valor p (<0.1) y un error estándar igual a 0.02230. Si su puntaje en la encuesta de soledad aumenta en una unidad y las demás variables se mantienen constantes el promedio semestral varía en -0.03924 puntos. Esto da a entender que niveles de soledad bajos (puntaje en la encuesta bajo) tienen efectos positivos en el rendimiento académico. Se podría inferir que la gente que sufre de soledad tiene más dificultades para poder crear grupos de estudio y/o interactuar con sus compañeros y profesores resultando en un peor rendimiento.

14. Conclusiones

Como se mencionó previamente, el objetivo principal del trabajo es averiguar cómo se podría mejorar el desempeño de los estudiantes de Dartmouth College. Para esto se intentó comprobar que el desempeño se relaciona con la calidad de vida del alumno y se evaluó qué hace que algunos alumnos tengan mejor rendimiento académico que otros. Con esta información obtenida, se van a proponer a continuación modelos para ayudar a la universidad a aplicar políticas o programas que asista al alumnado para aumentar su rendimiento.

A pesar de que en este estudio se enfocó exclusivamente en los datos recolectados de los estudiantes en Dartmouth College, podría ser sumamente enriquecedor replicar una metodología similar de recolección de datos e implementación de modelos en nuestra universidad (Instituto Tecnológico de Buenos Aires).

Antes de sumergirnos en las conclusiones es imperioso hacer unas previas aclaraciones. Había bastante inconsistencia en los datos debido a la cantidad de datos faltantes, ya que, por ejemplo, solo había registro de las calificaciones de solo la mitad de la muestra. Además, varios de los datos vinieron desestructurados debido a que venían en formato *json*.

Adentrándonos a las conclusiones, se notó una relación muy estrecha con la cantidad de materias que realiza el alumno, donde mientras menos materias hace, mejor rendimiento tiene. A su vez, mientras más tiempo ocupa realizando actividades, peor desempeño tiene (contrario a lo que se suponía previamente). Lo que explica estas relaciones es algo más

simple: que mientras más tiempo tiene el estudiante para estudiar, mejor rendimiento académico va a tener.

Algunos supuestos establecidos al inicio de la investigación fueron probados como erróneos. Las notas no se ven influenciadas ni por el nivel de estrés del alumno ni por la cantidad de horas de sueño que tuvo, tal vez sea porque parte del desafío de estudiar en una universidad conlleva sobreponerse a los niveles de estrés y están acostumbrados.

A su vez, no se halló conexión entre los niveles de depresión con la cantidad de aportes que hace en Piazza, como tampoco la hay entre los niveles de estrés y la cantidad de tareas encomendadas.

Sin embargo, surgieron algunos insights bastante interesantes. Los niveles de estrés aumentan en junio, fenómeno que es lógico debido a que en ese mes suelen haber los *midterms*. Aún así, se encontró también que los sábados los alumnos se encuentran bajo mayor estrés que en cualquier otro día, a pesar de ser fin de semana, por lo que resultaría fútil profundizar en este hecho y averiguar qué factores pueden influir en este fenómeno.

Por último, se ha descubierto varias variables que afectan a las calificaciones semestrales, una de ellas siendo los niveles de depresión, por lo que sería prudente empezar a trabajar concientizando sobre esta problemática y actuar en los casos en donde haya depresión alta. Una forma efectiva de trabajar con el estado anímico de los estudiantes es fomentando la inclusión y la interacción entre ellos, ya que, no solo se afirmó previamente una relación directa entre la depresión y la soledad que padecía el alumnado, sino que también surgió un patrón entre un grupo de estudiantes de la universidad, el cual consta de alumnos con niveles moderadamente altos de depresión, soledad y estrés. Por estas razones se considera imperioso trabajar sobre la salud mental de los estudiantes de la universidad.