Universidad Complutense de Madrid
Facultad de Psicología

Trabajo de Fin de Grado

ANÁLISIS DE DATOS CON R: Diferencias entre usuarios en web de citas

AUTOR:

William Sanz Vivanco

TUTOR:

Miguel Ángel Castellanos López

“Los números tienen una historia importante que contar. Dependen de ti, para darles una voz.”

- Stephen Few

RESUMEN

En el presente trabajo de fin de grado se analizarán las diferencias entre usuarios en una web de citas mediante la base de datos “okcupid” con más de sesenta mil usuarios.

Dicho análisis ha sido realizado con el lenguaje de programación R para describir y analizar los datos de carácter eminentemente categórico, haciendo uso de técnicas estadísticas de análisis de datos como la Prueba χ^2 de Pearson y Análisis de correspondencias múltiple (ACM).

En el análisis se concluyen claras diferencias entre los usuarios por distintas variables como género, edad, inclinación política o pasatiempos.

Palabras clave: R, análisis de datos, chi-cuadrado, Análisis de correspondencias, okcupid

ABSTRACT

In the present End-of-Degree will be analyzed the differences between users from a dating website through the dataset “okcupid” with more than sixty thousand users.

This analysis has been made with the programming language R to describe and analyze the data with mainly categorical variables, making use of statistical data analysis techniques such as Pearson’s chi-squared test and Multiple Correspondence Analysis (ACM)

The analysis concludes clear differences between users by different variables such as gender, age, political inclination or hobbies.

Keywords: R, data analysis Pearson’s chi-squared, Multiple Correspondence Analysis, okcupid

INDICE

1. INTRODUCCIÓN

Un estudiante de psicología promedio no suele encontrarse con el obstáculo de tener que realizar un análisis estadístico de más de cientos de usuarios, algo qué por supuesto está cambiando. La ciencia de datos y la programación estadística ofrecen un enorme potencial que deberá ser implementado tarde o temprano en el grado de psicología para hacer frente al gran volumen de datos disponibles que solo pueden ser tratados mediante programación estadística y cuyo análisis ofrecerá un salto agigantado en el entendimiento de la conducta humana.

En el presente trabajo se abordará un problema de tal magnitud: ¿Cómo analizar los datos de más de sesenta mil usuarios? Para ello gracias a las ventajas de implementación estadística que ofrece R se podrá abordar.

¿Por qué R?

R es un lenguaje de programación cuya principal característica a destacar es que es un lenguaje de software libre, lo cual significa que su código puede ser mejorado y compartido libremente sin necesidad de licencia, lo que lo torna gratuito e inmensamente adaptable para las circunstancias de análisis de cualquier índole; desde el uso en ciencias sociales y matemáticas hasta el procesamiento de lenguaje natural e inteligencia artificial.

Para su uso se requieren conocimientos básicos de programación, aunque su aprendizaje resulta mucho más sencillo en comparación a lenguajes comunes (C++,Java…). No obstante una vez adquiridos unos conocimientos mínimos, entornos como RStudio o la inmensidad de librerías disponibles agilizan y permiten trabajar con R de forma optima en un relativo corto de tiempo.

Objetivos y procedimiento

La base de datos ya ha sido analizada con anterioridad, cabe destacar el realizado por Kirkegaard, E. O. W., & Bjerrekær, J. D. en 2016 quienes realizaron un análisis descriptivo de la distribución de inteligencia en relación a distintas variables (genero, clase social, clase económica, pasatiempos…) y nuevamente en 2018 los mismos autores un análisis enfocado en la relación entre variables de criminalidad y los tipos de usuarios.

Para el presente trabajo se abordará en términos generales un análisis de las diferencias entre los usuarios en distintas variables, principalmente en cuestión de género, inclinación política y pasatiempos.

El trabajo se divide principalmente en 4 bloques:

1. **Pre-procesado de datos:** Que engloba el cargamiento de datos, el primer acercamiento a los mismos para entender su tipología y el preparamiento de los mismos para su análisis.
2. **Descripción de los datos:** Ningún análisis inferencial tiene sentido sin una previa descripción de los datos pues se debe de entender la distribución de los mismos. Aquí se incluyen las primeras visualizaciones.
3. **Análisis de los datos:** Una vez se conoce el tipo de datos a los que se enfrenta se puede desarrollar un análisis que encaje con sus características.
4. **Conclusiones:** Una breve recopilación y discusión sobre las diferencias encontradas.

No obstante debe recalcarse que pese a que se esperan unos conocimientos de estadística mínimos para el trabajo de fin de grado de psicología no lo es el entendimiento de un lenguaje de programación, por lo consiguiente más del 80% del tiempo dedicado al trabajo se ha empleado en aprender y entender la sintaxis de R y el uso de sus librerías.

Recursos y herramientas

- **Base de datos:**

La base de datos original en la que se basa el trabajo puede encontrarse en el siguiente enlace:

https://mega.nz/#F!QIpXkL4Q!b3QXepE6tgyZ3zDhWbv1eg

Correspondiente al utilizado en el paper "The okcupid dataset: A very large public dataset of dating site users" por *Kierkegaard, E. O. W., & Bjerrekær, J. D. (2016).*

Así mismo se puede encontrar una versión , ya filtrada y pre-procesada como librería para R:

https://cran.rstudio.com/web/packages/okcupiddata/index.html

```
library(okcupiddata)
```

- **Hardware:**

- Modelo: HP laptop 15-dw0xxx
- Procesador: Intel Core i7-8565U
- RAM: 12 GB
- Tarjeta gráfica: NVIDA GeForce MX130
- Sistema operativo 64 bits, Windows 10

El pc posee los recursos más que suficientes para abordar los datos, encontrándose el mayor tiempo de espera en ejecución de código entorno a las cuatro horas.

- **Software:**

- RStudio (Como entorno de R).
- Microsoft Office: Word y Excel
- Jupyter Notebook (Entorno de R desde Anaconda)

- **Librerías (R):**

- **Tidyverse:** Tratamiento de los datos
- **ggplot2:** Visualización de los datos
- **viridis:** Paleta de colores
- **FactoMineR y factoextra**: Análisis de correspondencias simple y múltiple

*Huelga mencionar la conexión a internet para el uso de librerías.

2. PRE-PROCESADO DE DATOS

La base de datos

La base de datos proviene de una conocida web de citas de habla inglesa: ***okcupid.*** La cual tiene fama de plantear a los usuarios que se registran en ella la posibilidad de responder a miles de preguntas, a tal cantidad, que la propia empresa okcupid realiza análisis estadísticos con los usuarios planteando y proponiendo consejo con dichos datos sobre como triunfar en su web de citas:

https://theblog.okcupid.com/

La base de datos se encuentra en formato **.csv** con separador tipo **" ' "** y valores nulos como espacios en blanco.

Consta de un total de **2.621 columnas y 68.371 filas.**

Las filas corresponden a los usuarios y las columnas a las preguntas planteadas, las cuales son de una naturaleza increíblemente variada.

Las columnas se encuentran codificadas en formato del tipo q_n, siendo n un numero aparentemente no relacionado con la pregunta.

Para poder conocer qué tipo de preguntas y respuestas corresponden a cada columna, la base de datos suele ir acompañada de un archivo adjunto; un Excel donde se pueden encontrar en el siguiente formato el código de pregunta e información más ampliada:

X	Código correspondiente a la pregunta Ej: q456
Text	Pregunta formulada al usuario Ej: Do you smoke?
Option_1-Option_4	Alternativas de respuesta
N	Número de usuarios que han respondido a la pregunta
type	Clasificación de la pregunta: O = Ordinal N = Nominal M = Mixed
Order	
Keywrods	Palabras clave Ejmplo: Descriptive, politics, sex/intimacy…

Se desconoce el orden en el que son presentadas las preguntas al usuario.

Tipos de datos

En la base de datos se encuentra todo tipo de información referente a los usuarios correspondiente a las preguntas planteadas, principalmente se puede resumir dicha información en dos grupos, por un lado del tipo demográficas (Edad, estudios, Género…) y por otro lado las respuestas a las preguntas planteadas, las cuales se pueden clasificar en los siguientes grupos: Descriptivas, preferencias, política, religión, sexual/intimidad. Aunque esto en grandes rasgos, pues se pueden encontrar desde "¿Te gusta le Bacon?" hasta "¿Crees que un país debe avisar a la ONU antes de declarar la guerra a otro?" y es que 2621 preguntas dan mucho para indagar.

Huelga decir que no se puede encontrar información personal de los usuarios del tipo dirección o número de teléfono, únicamente se puede conocer el apodo o nick de usuario como elemento identificativo (se desconoce aun así, si dicho usuario resulta real).

Gracias al Excel adjunto se puede conocer rápidamente la siguiente información:

- Prácticamente todas las preguntas son del tipo categórico a excepción de algunas como la Edad (d_age)
- Incluso la variable sueldo (d_income) se encuentra categorizada como variable ordinal en 13 niveles: Desde menos de veinte mil dólares hasta más de un millón.
- Las preguntas se encuentran clasificadas como nominales, ordinales o mixtas.

En el último punto se encuentra el primer inconveniente, pues preguntas que no deberían ser consideradas como variables ordinales (puesto que no existe ninguna relación del tipo ordinal entre sus niveles de respuestas) se encuentran clasificadas como tal, por ejemplo las preguntas de respuesta dicotómica del tipo Yes/No aparecen clasificadas como ordinales y así mismo algunas preguntas politómicas.

En resumen, la clasificación del apartado "type" no resulta fiable y no ofrece información respecto al tipo de variable en la base de datos.

Valores nulos (NA)

Teniendo en cuenta el numero filas y columnas (filas x columnas) se deberían tener un total de 179.200.391 observaciones. No obstante realizando un conteo de los valores nul os (NA) en la base de datos se encuentra la cifra de 137.868.350 valores NA, es decir el **76,9% del dataset está en blanco.**

¿Por qué? Bueno, hay que entender el contexto del cual se extrae el dataset; una persona se registra a una web de citas en la cual se le hacen nada menos que 2621 preguntas, es más que aceptable asumir que la gran parte de los usuarios no responderá a tantas pregu ntas, pero ¿a cuantas sí?

Como se puede observar en el GRÁFICO_1 sorprendentemente de media cada sujeto responde a unas 605 preguntas, se puede apreciar además como a partir de las 1200 preguntas la densidad empieza a decaer significativamente.

GRÁFICO_1: Numero de preguntas que responde cada usuario

En el eje x: el número de preguntas que se responde

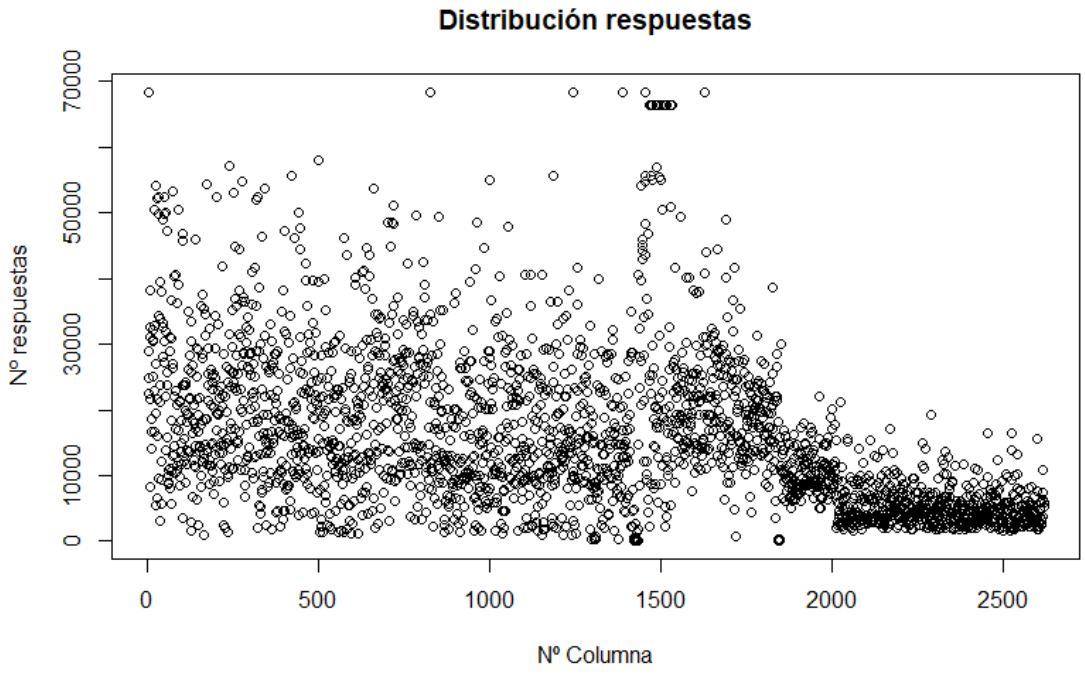
En el eje y: el número de usuarios que responden a dichas preguntas

Si de casi tres mil preguntas, la mayoría de los usuarios responden a menos de mil, ¿Son estas mil preguntas las mismas a las que responden todos los usuarios?

En el GRÁFICO_2 se aprecia como entorno a las 1800 preguntas hay una mayor variabilidad por los usuarios a responder a una u otra, sin embargo, a partir de la pregunta 2000 el número de usuarios que responden a dichas preguntas decrece importantemente.

En la parte superior del gráfico se pueden encontrar además algunos puntos aislados los cuales representarían preguntas que son respondidas por casi la totalidad de usuarios, se desconoce si son preguntas que deben responderse de forma obligatoria o no, muy posiblemente sean preguntas demográficas, tales como edad, género, orientación etc.

GRÁFICO_2: Distribución de las respuestas de los usuarios



En el eje X se muestra el nº de la columna que corresponde a la pregunta (recordemos que teníamos 2621)

En el eje Y el nº de usuarios que responden a esa pregunta.

Tratamiento preliminar de los datos

Para poder empezar a describir los datos, antes resulta conveniente realizar algunos pequeños ajustes en los mismos para poder trabajar con ellos.

Antes de nada, se debe comprobar que no existen filas con valores repetidos; afortunadamente no se encontraron.

El primer obstáculo se encuentra en la variable género, la cual supuestamente posee hasta 107 niveles. ¿Cómo es esto posible? Se muestran algunos de ellos:

```
> head(levels(data$d_gender))
[1] "Agender"                          "Agender, Genderqueer"
[3] "Agender, Non-binary, Other"       "Agender, Non-binary, Trans Man"
[5] "Androgynous"                      "Androgynous, Cis Woman, Gender Nonconforming"
```

No se puede hacer frente a tanta diversidad y es que al comprobar las frecuencias de cada nivel existen niveles donde solo hay una persona, en total, los niveles que no son Man/Woman tienen un total de 2204 usuarios, lo que representa un 3,22% de la muestra total.

Por ello, para facilitar el análisis, se ha reducido la variable género a únicamente dos niveles (Man/Woman) y eliminado de la muestra ese 3,22%, quedándose con 66.167 usuarios.

El segundo obstáculo a encontrar es en la variable sueldo (d_income) la cual posee 13 niveles de tipo ordinal, pero que no están ordenados. Se deberán ordenar manualmente colocando el rango <20.000$ como el primero y +1.000.000$ como el último. Además se eliminará el nivel “Rather not to say” puesto que a efectos, será considerado como un NA.

3. DISTRIBUCIÓN DE LOS DATOS

Edad y género

Se observa el número de hombres y mujeres registrados en okcupid y su proporción:

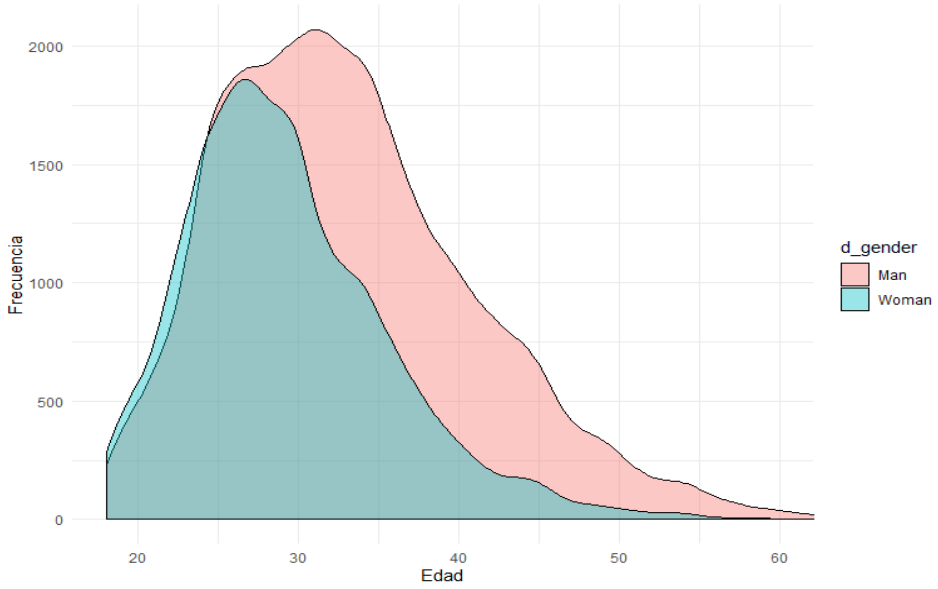
	Frecuencia	Proporción
Man	**40215**	**60,7%**
Woman	25952	39,2%

Así pues, en la página web de citas existe una relación de aproximadamente tres hombres por cada dos mujeres.

No es algo realmente sorprendente, por sentido común se puede estimar que las páginas web de citas son mucho más famosas y usadas por hombres que por mujeres, aunque esto muestra una medida exacta de la relación.

El GRÁFICO_3 representa por separado a los hombres y mujeres respecto a su edad

GRÁFICO_3: Edad y género



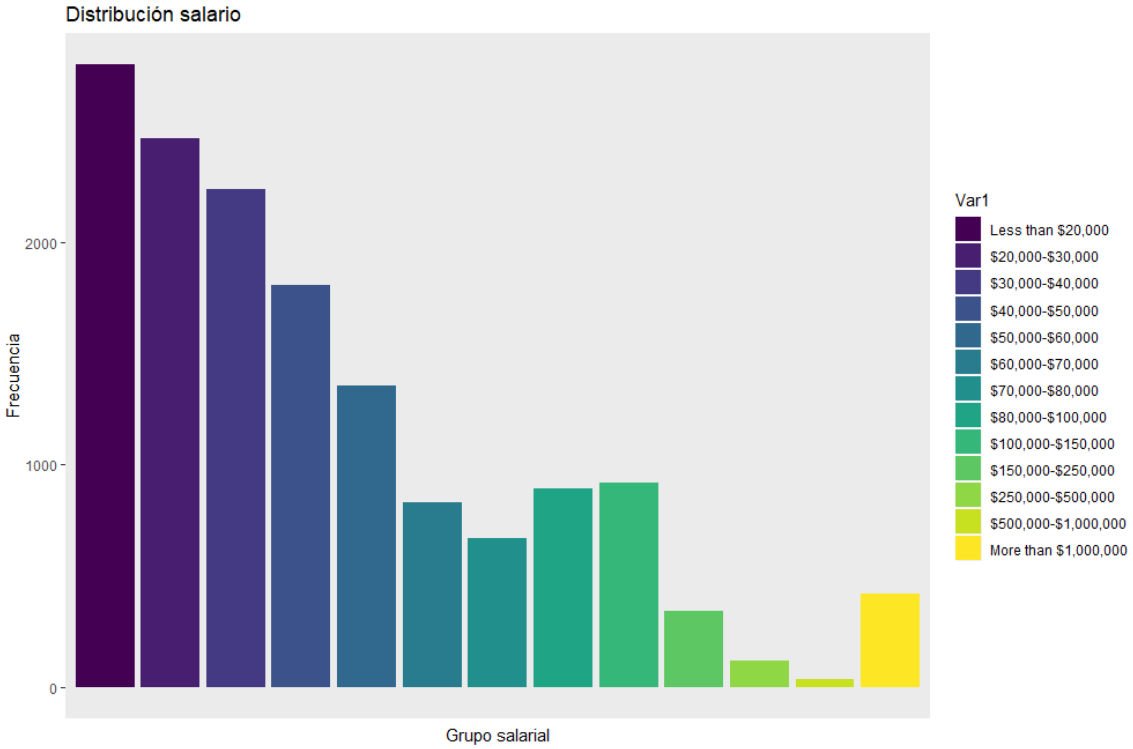
	Edad		
	Media	Desviación típica	Moda
Hombre	33,17	8,204	31
Mujer	29,32	6,480	27

Se observa que la distribución de edad se inclina más hacia la izquierda que la derecha, es decir, es más popular entre gente joven (menor de 40 años) y que la media de edad es mayor en hombres que en mujeres.

Salario

La variable sueldo se encuentra agrupada por niveles ordinales como se aprecia en el GRÁFICO_4

GRÁFICO 4: SALARIO



La frecuencia es decreciente en prácticamente todos los niveles lo cual es de esperar; a mayor salario, menor número de personas.

Aunque parece encontrarse una acumulación entorno a los 150.00-250.000$, bien pudiera ser por la idiosincrasia de la muestra o porque verdaderamente pueda existir un techo en el intervalo, el cual resulta difícil de sobrepasar para una persona promedio sin influencias sociales y/o familiares.

Por último, se aprecia que la última barra referente a más de un millón posee una acumulación extrañamente superior a los anteriores niveles lo cual puede deberse a que son necesarios mas niveles superiores, es decir:

Las personas que ganan más de dos millones o más de diez millones se acumulan en el mismo grupo de más de un millón que hace de límite superior.

Ahora se analiza la distribución del salario por empleo (GRÁFICO 5) y nivel de estudios (GRÁFICO 6) *Siguiente página.

En el GRÁFICO 5 se puede apreciar la degradación de color, cuanto mas oscuro, menos salario.

Los que más destacan por oscuridad a simple vista son las categóricas de empleo: Administración, Arte y Turismo (hospitality).

Los niveles en donde hay mayor porcentaje de millonarios son retirados y desempleados. ¿Quién necesita trabajar teniendo más de un millón verdad? Contrariamente los desempleados y estudiantes poseen el porcentaje mas alto en el nivel mas bajo (menos de veinte mil).

Respecto al GRÁFICO 6 donde se muestran los niveles de estudio agrupados por salario se encuentra la distribución esperada; a mayor nivel de salario mayor nivel educativo hasta alcanzar el rango de más de 500 mil donde la educación parece distribuirse de forma más aleatoria, posiblemente porque pasa a ser algo secundario referente a los ingresos por trabajo.

GRÁFICO 5: Salario_Empleo

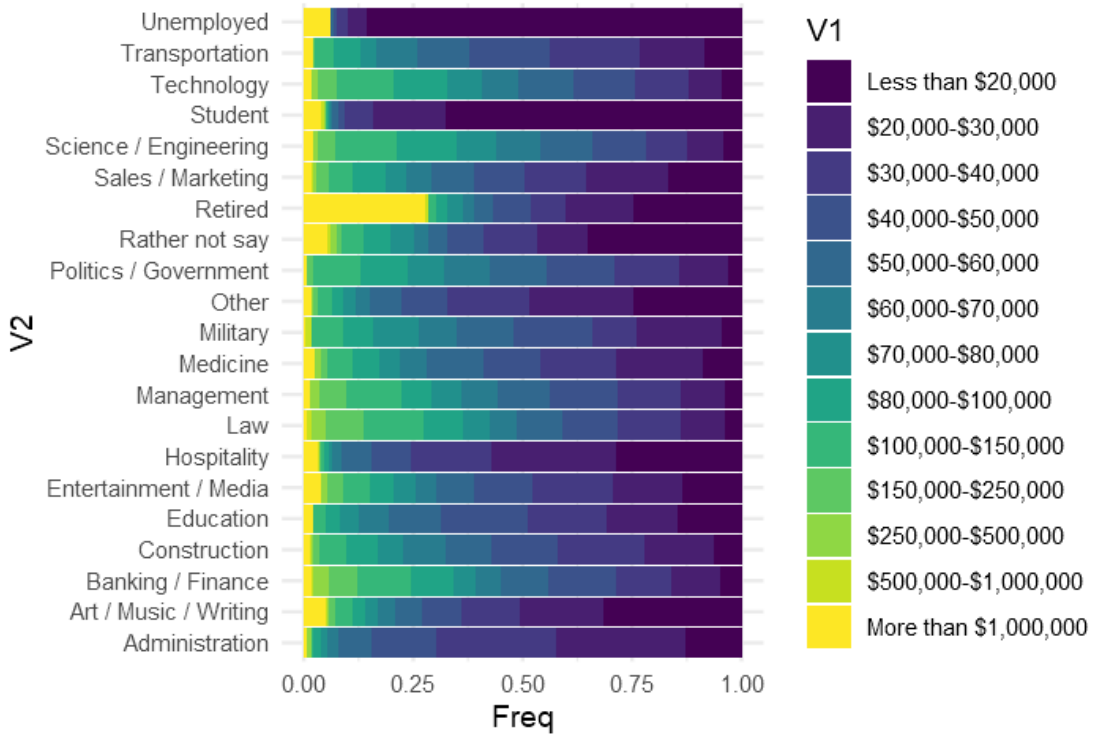
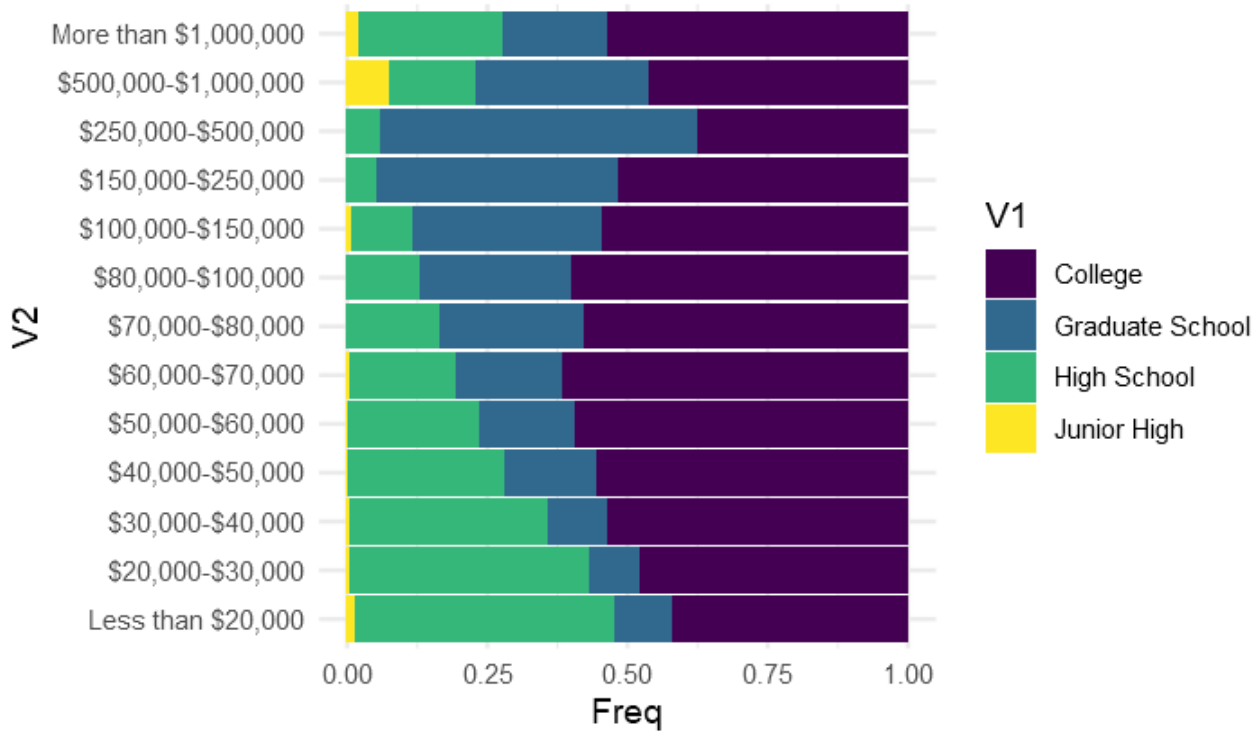


GRÁFICO 6: Salario_estudios



4. ANÁLISIS DE DATOS

Conocer el tipo de variable que se está tratando resulta fundamental para decidir la prueba estadística a realizar, como se comentó anteriormente, la base de datos no posee una clasificación fiable sobre el tipo de variable, si que se conoce que la mayor parte son categóricas pero no si estas son del tipo nominal u ordinal.

Dada la gran magnitud de variables, resulta inviable seleccionar una por una y determinar que tipo de variable es, por lo consiguiente se ha establecido el siguiente razonamiento: Se seleccionan las variables que tienen solo dos niveles, se devuelve el valor de niveles únicos de todas las variables con dicho filtro, mediante este procedimiento se llega a una importante conclusión: No existen variables de dos niveles en la base de datos que puedan ser de tipo ordinal, como pudieran ser primero o segundo, la mayor parte (prácticamente todas) son Yes/No, obteniéndose un total de 1109 preguntas de tipo nominal con dos alternativas de respuesta, en las preguntas restantes (2538) no se conoce si son nominales u ordinales.

Este filtrado de 1109 preguntas resultará de utilidad posteriormente para hacer un rápido análisis de sus proporciones.

En cualquier caso, dado que se tratan variables categóricas y se desea comprobar la relación entre dichas variables, se realizará primero una prueba X^2 de Pearson para conocer la dependencia/independencia entre variables y posteriormente un análisis de correspondencias para poder representar gráficamente la relación entre un gran numero teniendo o no múltiples niveles.

X^2 de Pearson (Chi-cuadrado)

Partiendo de columnas con datos respectivos a los sujetos n referentes a las variables X1 y X2. Se desea conocer si dichas variables están o no relacionadas.

Siendo X1 el género con dos posibles niveles (H/M)

Siendo X2 la pregunta q512: "Do you like to dance?" . Con dos posibles respuestas (Yes/No)

n	**X1**	**X2**
1	Man	Yes
2	Woman	No
3	Woman	Yes
{…}	{…}	{…]

En primer lugar se realiza una tabla de frecuencias, la cual muestra el número de veces que se repite cada valor agrupado entre los niveles de cada variable.

Posteriormente realizando los sumatorios de cada fila y columna se obtiene la conocida tabla de frecuencias:

	No	**Yes**	∑
Man	5859	9381	15240
Woman	1099	6328	7427
∑	6958	15709	22667

n = 2267 serían el número de sujetos totales que han indicado su género y han respondido a la pregunta.

La tabla, mostraría las frecuencias absolutas (o frecuencias reales).

Las frecuencias relativas o esperadas mostrarían como se distribuiría la variable idealmente en caso de que fueran totalmente independientes. Para calcularlas:

Se toman los sumatorios correspondientes a la fila y columna de cada celda, se suman y se dividen entre n obteniéndose las frecuencias esperadas:

	Yes	No
Man	4678,163	10561,837
Woman	2279,837	5147,163

Por ejemplo: $4678{,}163 = \frac{6958*15240}{22667}$

Ahora, calculando la diferencia entre las frecuencias reales y las esperadas se obtienen los residuos:

	Yes	**No**
Man	1180,837	-1180,837
Woman	-1180,837	1180,837

Por ejemplo: $1180{,}837 = 5859 - 4678{,}163$

Como se ha mencionado, se desea conocer si existe relación entre las dos variables; se comparan las frecuencias reales con las esperadas, en caso de que fueran iguales, las variables serían totalmente independientes. Dado que se busca precisamente la dependencia entre variables, la hipótesis al comparar las dos variables es que no son independientes:

$$\begin{cases} HO: X1\ y\ X2\ son\ independientes \\ H1: X1\ y\ X2\ son\ dependientes \end{cases}$$

Para poder compararlos se puede establecer la relación entre ellos a modo de ($\frac{a+b}{a}$), siendo **a** las frecuencias reales y **b** las esperadas. Dado que tanto a como b puede tomar valores positivos o negativos, se elevan al cuadrado para posteriormente poder sumar dichos resultados acumulándolos: $\left(\frac{a+b)^2}{a}\right)$

Calculando para cada celda:

	Yes	**No**
Man	298,061	132,020
Woman	611,612	270,902

La suma de todas las celdas es el estadístico X^2, que en este caso toma valor de **1312,595**

Este procedimiento utilizado para hallar el estadístico X^2 , mostrará el grado en el que las dos variables son o no independientes entre sí. Cuanto más independientes sean, el estadístico X^2 tenderá a 0, si no son independientes X^2 será más grande. ¿Pero cuan grande debe ser para representar que su magnitud es estadísticamente significativa y que las variables están relacionadas entre sí?

Para ello se debe comparar con la distribución muestral del estadístico X^2 y sus grados de libertad $X^2 \sim X^2(I-1(J-1)$ **(**Siendo **I** filas y **J** las columnas, para el caso; g.l=1)

Al compararlo con su distribución muestral, se puede calcular el p-valor que existe para un $X^2 = 1312{,}595$ con un grado de libertad, se ha obtenido: **2e-16.**

La cual es una cifra tan pequeña que puede redondearse a 0.

¿Cómo interpretar el p-valor? Un valor de 0 significa que la probabilidad de que la hipótesis nula sea cierta es muy baja y al ser prácticamente 0 deberá tomarse su contraria: Si no son independientes (H0), deberán ser dependientes (H1) (Esto siempre teniendo en cuenta los intervalos de confianza, al ser X^2 tan grande para sus grados de libertad no queda duda de que no puede aceptarse la H0 en ningún caso.)

Otra forma de plantearlo es decir que el p-valor representa la probabilidad de que los datos hallados se deban al azar, al ser tan pequeño, si se repite un estudio similar debería arrojar las mismas conclusiones; dependencia entre las dos variables.

Ahora que mediante la prueba $\boldsymbol{X^2}$ de Pearson se establece que las variables están relacionadas entre sí, cabe preguntarse, ¿Qué tan relacionadas? Para ello se puede estimar dicha relación mediante la C de contingencia, la cual se calcula de la siguiente forma:

$$C = \sqrt{X^2/(X^2 + n)}$$

$$C = \sqrt{1312{,}595/1312{,}595 + 22667)} = 0{,}234$$

C toma un valor siempre mayor que 0 y menor que 1. Cuanto mayor sea, mayor relación.

El procedimiento aquí explicado para conocer la relación entre dos variables es ideal para variables categóricas, ¿pero como aplicarlo cuando se tienen miles de variables? Para ello se ha desarrollado una función que aplique este mismo procedimiento a todas las variables categóricas, primero todas las comparaciones múltiples posibles entre todas las variables categóricas nominales (Adjunto como X_NOMINALES) y posteriormente emparejando por separado la variable genero con todas las demás variables y repitiendo el proceso con la variable sueldo. (Se adjuntan como X_GÉNERO y X_SUELDO)

Extrayendo los p-valores, el estadístico de chi-cuadrado y el tamaño de la muestra, utilizado para calcular la C de contingencia, la función devolverá un dataframe como el siguiente (extraído de comparaciones múltiples entre todas las variables y genero):

	X	p.value	statistic	n	C
1	d_income	1.17997045422232e-174	8.563238e+02	24422	9.837524e-01
2	q16053	0	2.554756e+03	56534	9.956970e-01
3	q501	0.033833188133031	4.503114e+00	55357	1.913582e-02

Análisis de proporciones

Para las comparaciones con género, dado que solo tiene dos niveles, se han elegido de X_GENERO aquellas variables emparejadas que tuvieran también solo dos niveles, de esta forma al ser una matriz 2x2 se pueden extraer las proporciones marginales por fila:

	No	**Yes**	∑
Man	5859	9381	15240
Woman	1099	6328	7427
∑	6958	15709	22667

	No	**Yes**	∑
Man	0,384	0,616	1
Woman	0,148	0,852	1

Dado que los datos de la primera columna son inversos a los de la segunda a=1-b (Si hay 0,4 en una, debe haber 0,6 en la segunda). Teniendo uno de los dos datos aporta información suficiente. Tomando los mismos para los niveles de Man/Woman y luego calculando la distancia (en valor absoluto) se obtendría lo siguiente:

Man Yes	**Woman Yes**	**D M/W**
0,616	0,852	0,236

Esto aporta la misma información de forma más compacta: El 61% de los hombres responde que si y el 85% de las mujeres responden que sí, la diferencia entre responder si entre los dos grupos es del 23%.

Nuevamente se ha diseñado otra función para extrapolar este calculo aplicándolo a todas las variables que tengan dos niveles.

Obteniéndose un nuevo dataframe X_PROPORCIONES (adjunto). Posteriormente se muestran aquellas que tuvieran una C de contingencia mayor (y mayor distancia/diferencia):

*R2 es el % que responde a la primera alternativa de respuesta, casi siempre representa "Yes", si no, se especifican las alternativas en la pregunta entre "()"

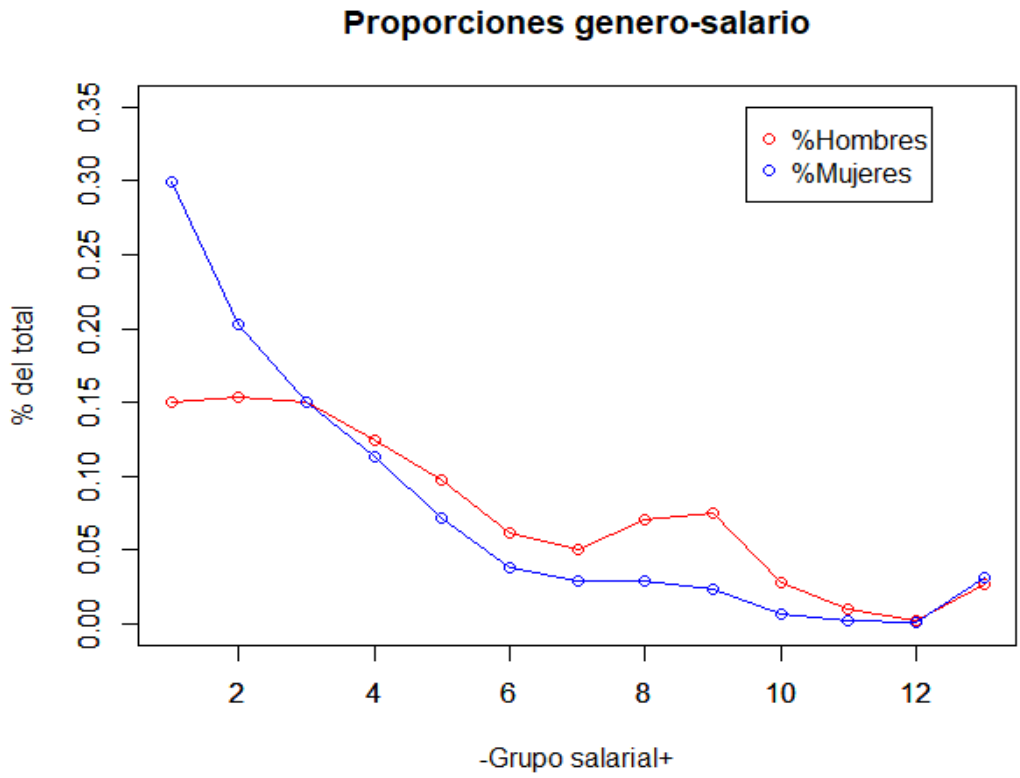
TABLA: Diferencias entre hombres y mujeres en OKCUPID

X	Pregunta	Man R1	Woman R1	D M/W
q13669	Would you date someone shorter than you?	0,99	0,43	0,56
q463	In your ideal sexual encounter, do you take control, or do they? **(They/I take)**	0,32	0,86	0,54
q19219	Can you change a tire on your own?	0,92	0,54	0,38
q363047	Have you had a girlfriend before?	0,95	0,57	0,38
q485	Do you know and enjoy chess?	0,74	0,39	0,35
q333	Would you ever sleep with a porn star?	0,66	0,33	0,33
q26	Have you ever owned sex toys?	0,51	0,84	0,33
q1597	Would you consider sleeping with someone on the first date?	0,71	0,39	0,32
q45645	Have you ever faked an orgasm during sex?	0,29	0,61	0,32
q8215	Would you ever date someone who depended on their parents' money?	0,61	0,29	0,32
q9662	Could you date a stripper?	0,67	0,36	0,31
q49345	Would you consider performing anilingus on a partner who asked you to?	0,71	0,40	0,31
q18955	Would you date someone who's smart but achieved nothing in life?	0,83	0,53	0,30
q15280	Could you date someone with no long-term goals?	0,55	0,25	0,30
q460831	If you had the option to see your date naked before your first meeting, would you take it?	0,63	0,33	0,29
q82	Have you ever tried Yoga?	0,53	0,83	0,29
q298	If possible, would you prefer to date someone a lot more attractive than you, or about the same? **(Same/More)**	0,52	0,81	0,29
q85310	Have you ever driven a motor vehicle over 100mph (161kph)?	0,76	0,48	0,29
q12750	Do you ever spit on the ground, in public?	0,48	0,19	0,28
q273	Do you know how to drive a stick shift (manual transmission)?	0,78	0,51	0,28
q72427	Do you enjoy keeping up to date with the latest technology news?	0,75	0,48	0,28
q36126	Would you be willing to shave your head to raise money for charity?	0,83	0,56	0,28
q72123	Do you think it is a good idea for a single person to try to create or adopt a child and raise it on his or her own?	0,55	0,83	0,28
q82397	Are you smarter than the majority of people on this planet?	0,70	0,42	0,27
q159	Do you know any programming languages?	0,51	0,23	0,27
q50990	Would you want to be immortal if you could?	0,67	0,40	0,27
q23390	Could you date someone who cannot speak your own language very well?	0,72	0,46	0,26
q48960	Would you consider dating someone who does not know how to drive a car?	0,81	0,55	0,26
q82717	Do you think OkCupid should add a spot for weight in user profiles?	0,48	0,22	0,26
q156	Would you ever seriously date someone half your age?	0,38	0,12	0,26
q18814	Should a country always need the UN's approval before declaring war?	0,45	0,71	0,26
q53846	Do you think the International Space Station would be a romantic place to travel to with a partner to exchange vows?	0,64	0,39	0,25
q50565	Do you blush easily?	0,38	0,63	0,25
q21	Do you enjoy meaningless sex?	0,52	0,27	0,25

Los cálculos anteriores son de gran utilidad a la hora de resumir la diferencia entre dos niveles. Cuando se encuentran más, se requiere una descripción gráfica de los mismos y más aún cuando se desea conocer la dirección, como es el caso de la variable sueldo.

En el GRÁFICO 5 se aprecia la proporción de usuarios en cada nivel, mostrando por separado las poblaciones referentes a hombres y mujeres

GRÁFICO 7: Distribución de proporciones de género por nivel de salario



Salario (miles $)	**-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**	**80-100**	**100-150**	**150-250**	**250-500**	**500-1000**	**+1m**
Grupo	1	2	3	4	5	6	7	8	9	10	11	12	13

Se observa así por ejemplo, que el 30 % de mujeres se encuentran en el grupo 1 en comparación al 15% de hombres; hay el doble de mujeres que de hombres (en proporción) que ganan menos de 20.00$.

Aunque se observen diferencias claras de sueldo a favor de los hombres, éstas, se requerirían estudiar en mayor profundidad, por ejemplo, recordamos que la media de edad de hombres era mayor que el de mujeres, este simple hecho ya debería de influir en un mayor sueldo a favor de los hombres.

Análisis de correspondencias

Comúnmente para la agrupación de variables se recurre a un Análisis de Componentes Principales (ACP), el problema de dicho análisis es que su punto de partida se basa en la matriz de correlaciones creada a partir de las comparaciones múltiples entre las variables. Dado que en el presente análisis se trabajan con variables categóricas no tiene sentido analizar correlaciones entre variables no cuantitativas, el Análisis de Correspondencias Múltiples ofrece una alternativa al ACP para variables categóricas, basándose en las distancias ponderadas por filas y columnas de la misma forma en que trabaja el de $\boldsymbol{X^2}$ de Pearson, de hecho, su calculo de distancias se denomina distancia $\boldsymbol{X^2}$. Por ello resulta una técnica ideal como complemento posterior a un análisis chi-cuadrado.

Partiendo de los mismos datos del ejemplo en el análisis de $\boldsymbol{X^2}$ de Pearson se mostrará en términos generales como trabaja el ACP:

	No	**Yes**	∑
Man	5859	9381	15240
Woman	1099	6328	7427
∑	6958	15709	22667

	No	**Yes**	∑
Man	n_{11}	n_{12}	n_{f1}
Woman	n_{21}	n_{22}	n_{f2}
∑	n_{c1}	n_{c2}	n

A partir de la tabla de frecuencias absolutas, se pueden calcular las proporciones de fila y de columna respectivamente:

P.Fila	**No**	**Yes**
Man	$\frac{n_{11}}{n_{f1}}$	$\frac{n_{12}}{n_{f1}}$
Woman	$\frac{n_{21}}{n_{f2}}$	$\frac{n_{22}}{n_{f2}}$

P.Columna	**No**	**Yes**
Man	$\frac{n_{11}}{n_{c1}}$	$\frac{n_{12}}{n_{c2}}$
Woman	$\frac{n_{21}}{n_{c1}}$	$\frac{n_{22}}{n_{c2}}$

P.Fila	**No**	**Yes**
Man	0,384	0,616
Woman	0,148	0,852

P.Columna	**No**	**Yes**
Man	0,842	0,597
Woman	0,157	0,402

Así mismo, se pueden calcular las proporciones de fila y columna al sumatorio de cada fila y columna respectiva al total:

	No	**Yes**	∑
Man			$\frac{n_{f1}}{n}$
Woman			$\frac{n_{f2}}{n}$
∑	$\frac{n_{c1}}{n}$	$\frac{n_{c2}}{n}$	n

	No	**Yes**	∑
Man			0,672
Woman			0,327
∑	0,306	0,693	22667

Donde se tomarían las proporciones parciales referentes a las filas y columnas, aunque dejando siempre las proporciones totales referentes al sumatorio total, para las columnas:

P.columna	**No**	**Yes**	∑
Man	$\frac{n_{11}}{n_{c1}}$	$\frac{n_{12}}{n_{c2}}$	$\frac{n_{f1}}{n}$
Woman	$\frac{n_{21}}{n_{c1}}$	$\frac{n_{22}}{n_{c2}}$	$\frac{n_{f2}}{n}$
∑	$\frac{n_{c1}}{n}$	$\frac{n_{c2}}{n}$	n

	No	**Yes**	∑
Man	P_{c11}	P_{c12}	P_{f1}
Woman	P_{c21}	P_{c22}	P_{f2}
∑	P_{c1}	P_{c2}	n

Para definir la diferencia entre dos columnas se restan sus proporciones y se elevan al cuadrado (puesto que la distancia es siempre positiva). Posteriormente se divide la diferencia entre las proporciones de columnas por la proporción total de la fila correspondiente.

$$dif(No, Yes)_1 = \frac{(P_{c11} - P_{c12})^2}{P_{f1}} = 0{,}089$$

Una vez calculada la distancia entre cada par de casillas emparejadas entre las columnas, realizando su sumatorio y calculando la raíz se obtiene la distancia entre las dos columnas:

$$D(No, Yes) = \sqrt{\sum dif(No, Yes)_n} = 0{,}522$$

Realizando el mismo procedimiento para las filas, se obtendría:

$$D(Man, Woman) = \sqrt{\sum dif(Man, Woman)_n} = 0{,}513$$

En caso de tener mas filas o columnas, se repetiría el proceso emparejando todas las posibles a modo de matriz:

D.Columnas	No	Yes
No	0	0,522
Yes	0,522	0

D.Filas	Man	Woman
Man	0	0,513
Woman	0,513	0

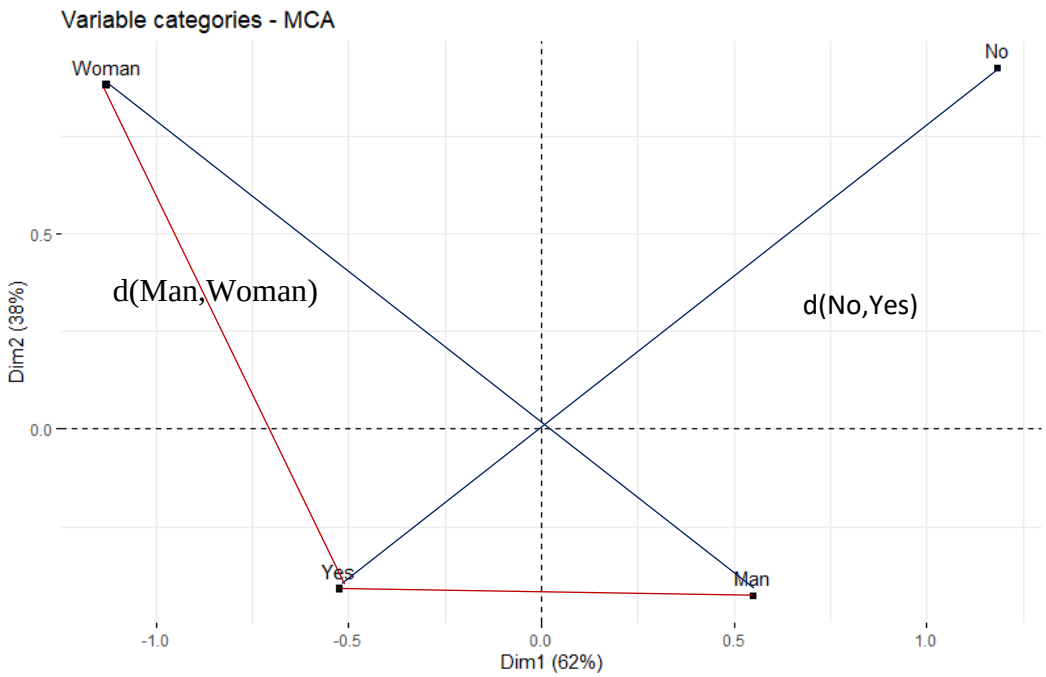
$$Dc = \begin{pmatrix} 0 & 0{,}522 \\ 0{,}522 & 0 \end{pmatrix} \qquad Df = \begin{pmatrix} 0 & 0{,}513 \\ 0{,}513 & 0 \end{pmatrix}$$

De esta forma se obtienen las distancias a las que se encuentran unos niveles de otros. En este caso se observa como la distancia entre los niveles de filas es más pequeño que la distancia entre los niveles de columnas, aunque una distancia muy pequeña hay que decir.

Para poder plasmar estas diferencias, se debería realizar un escalamiento multidimensional que represente dichas distancias en un espacio de dos dimensiones.

En R, gracias al paquete FactoMiner se puede representar dicho gráfico (Además de automatizar todos los cálculos anteriores), de modo que se obtienen las coordenadas de los puntos y su representación

GRÁFICO 8: AC Género_Dance



	X	Y
Man	0,5498221	-0,4301531
Woman	-1,1282198	0,8826622
No	1,183423	0,9258504
Yes	-0,5241745	-0,4100877

Dado que es un plano en dos dimensiones cualquiera, se puede hallar la distancia entre dichos puntos.

En la matriz calculada se obtuvo que la d(Man,Woman) < d(No,Yes)

Luego: $d = \sqrt{(x_2-x_1)^2 - (y_2-y_1)^2}$ → d(Man,Woman)=2,13 y d(No,Yes)=2,16.

Cumpliéndose d(Man,Woman) < d(No,Yes) → 2,13<2,16

Queda algo por aclarar, ¿Por qué razón la distancia entre d(Man,Yes) es mucho más corta que la d(Woman,Yes)?

Esto se debe a que el AC tiene en cuenta las frecuencias totales, lo que está mostrando es que hay muchos más hombres que responden que sí a mujeres que responden que sí, pero no en proporción (Se recuerda que 9381 Hombres respondían que Sí frente a 6328 Mujeres que también lo hacían). Algo muy importante a tener en cuenta, puesto que si en la muestra real existen muchos mas hombres que mujeres, esto sesgará las distancias dando más peso a las respuestas de los hombres que las de las mujeres. Para corregirlo deberán compararse las variables aislando por separado las poblaciones de cada género.

Por supuesto todo esto es una muestra de como funciona un AC mostrando un ejemplo sencillo de 2x2, lo realmente interesante de la representación de un análisis de correspondencias, resulta en poder comparar la distancia entre múltiples niveles de un mayor numero de variables y tratar de buscar agrupaciones entre ellas.

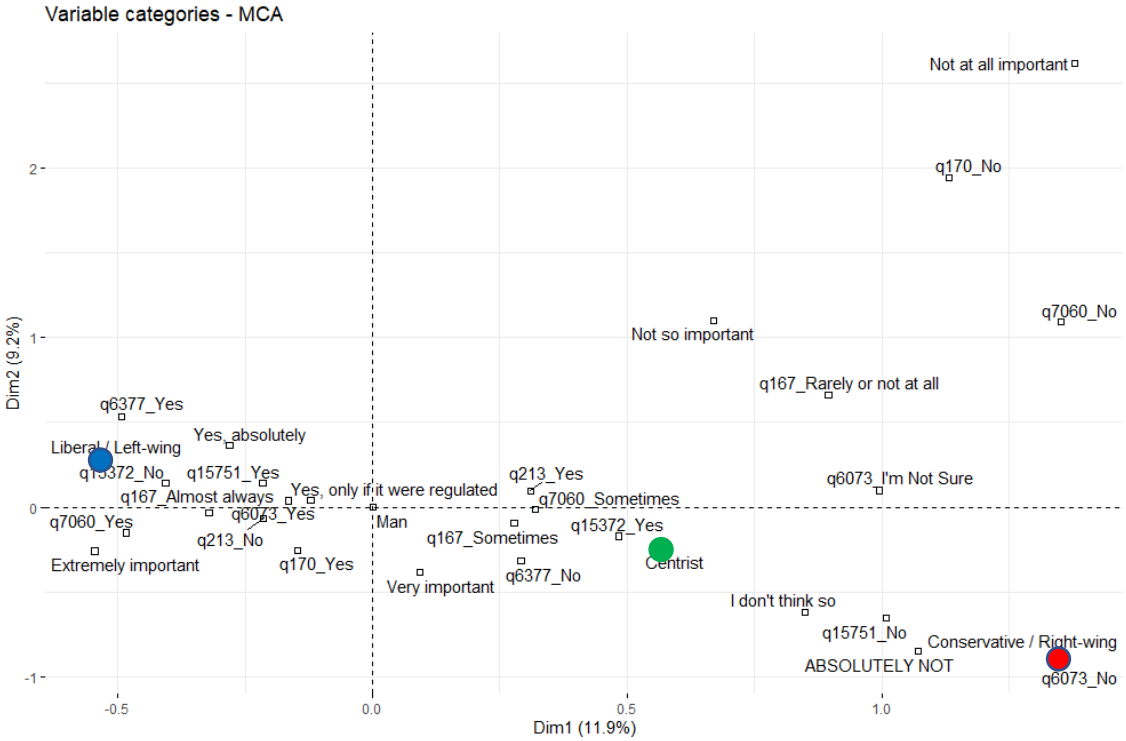
Para realizar un análisis de más de dos variables categóricas se debe usar un Análisis de Correspondencias Múltiple. (ACM), la cual no es más que una extensión de Análisis de Correspondencias Simple.

Posteriormente se muestran ACM correspondientes a agrupar distintas variables:

En el siguiente gráfico se muestra un ACM sobre las citadas variables en el cuadro además del nivel más próximo a los tres niveles de política:

Azul (Libreral/Izquierda), Verde(Centro), Rojo(Conservador/Derecha)

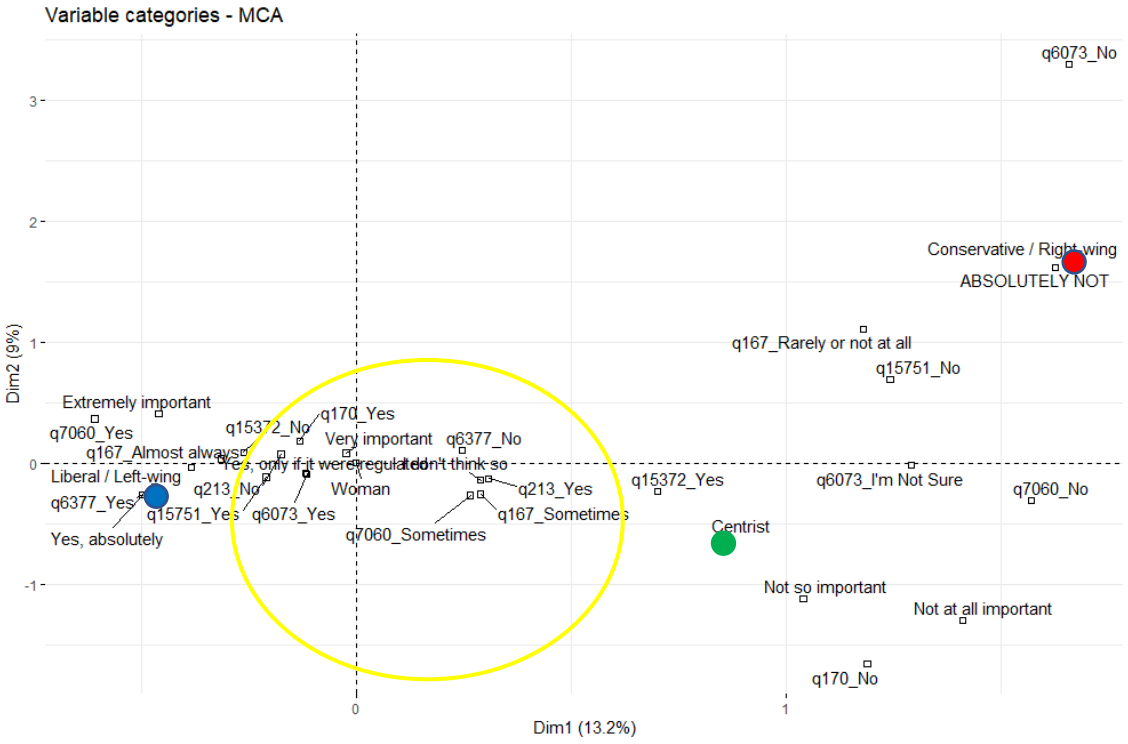
GRÁFICO 9: Política_Hombres



	Pregunta	LIBERAL	CENTRO	CONSERVADOR
q6377	Tienes problemas con la autoridad?	SI	NO	NO
q218	Debería ser legal la prostitución?	+Si, absolutamente -Si, regulada	+Si, regulada -Creo que no	+Absolutamente no -Creo que no
q15752	Es importante votar para ti?	Extremadamente importante	Muy importante	Muy importante
q170	Votas normalmente?	SI	SI	≈ SI NO
q15372	Saldrías con alguien que tiene ideas políticas opuestas?	NO	SI	SI
q213	Algunas vidas valen más que otras?	NO	SI	SI
q7060	La política es interesante?	SI	A VECES	NO
q15751	Las personas ricas deben pagar más impuestos?	SI	≈ SI NO	NO
q167	Reciclas?	Casi siempre	Algunas veces	≈ Algunas veces Rara vez
q6073	Educación sexual a niños menores de 15	SI	≈ SI No estoy seguro	+NO -No estoy seguro

Mismo ACM con población solo de mujeres

GRÁFICO 10: Política_Mujeres

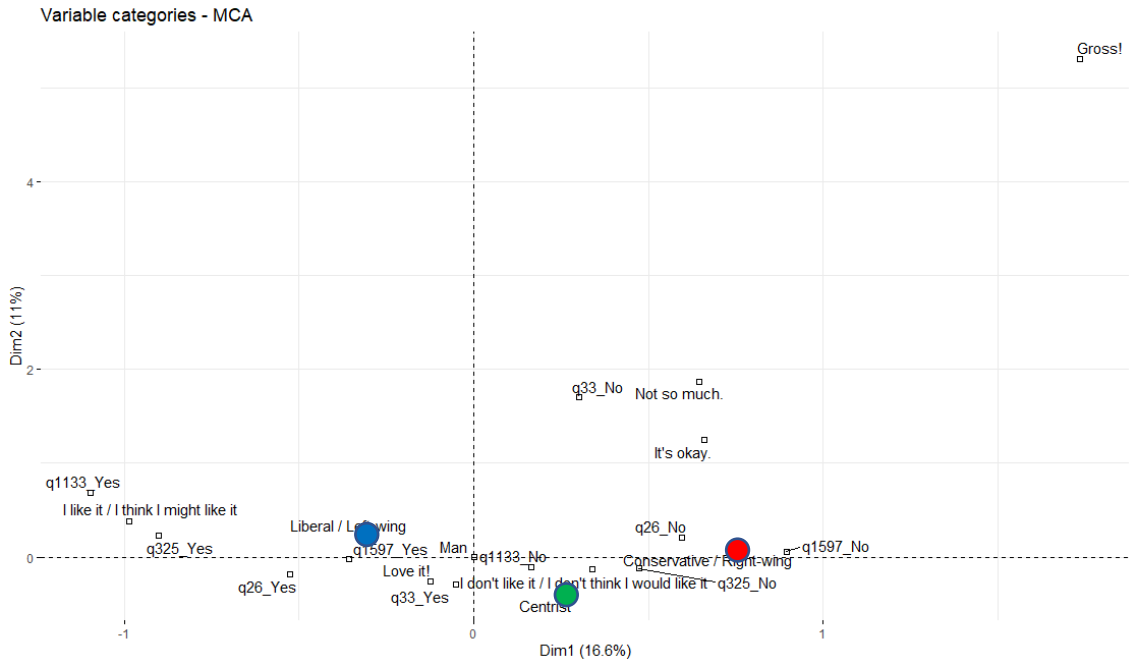


Como se puede apreciar los niveles de izquierda están más conglomerados mientras que los entorno a centro y derecha más dispersos. Parece ser que las mujeres tienden a ser más de izquierda y menos extremistas en caso de ser de derecha o centro, de hecho, podría considerarse una agrupación extra de centro-izquierda (circulo amarillo), aunque en la población de hombres también existe una inclinación de centro más hacia la izquierda que a la derecha pero no tan pronunciada como en las mujeres.

Por lo demás, pese a que la cercanía varíe entre los niveles de política, los niveles más cercanos se mantienen iguales al de la población de hombres.

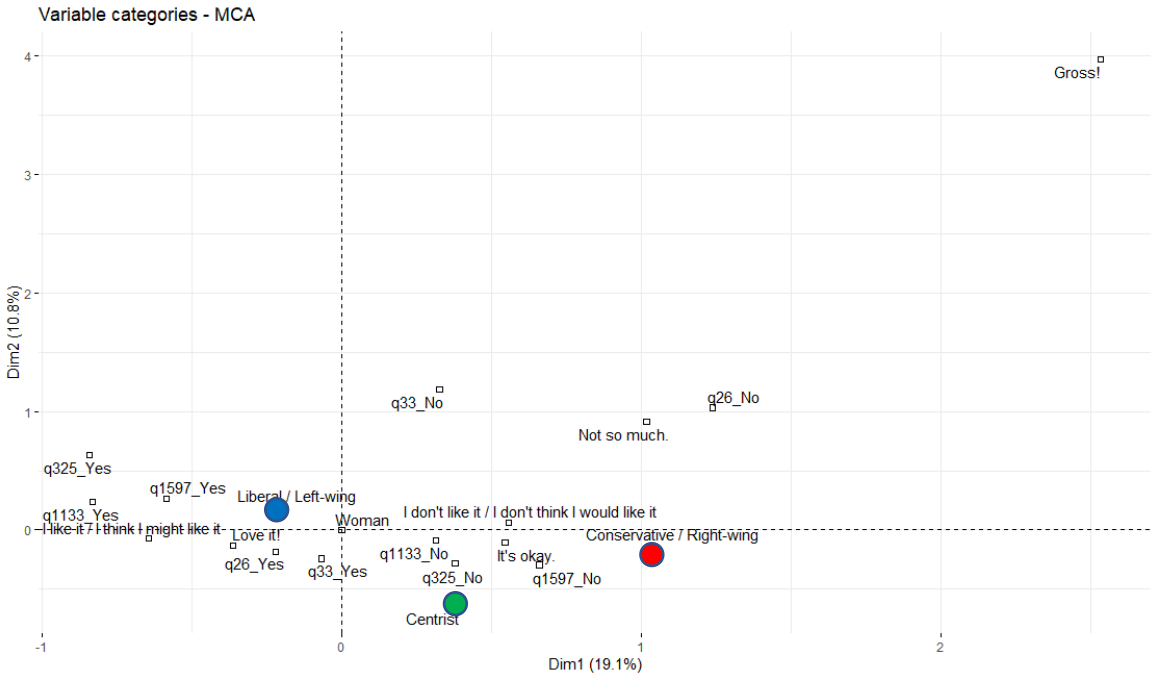
ACM aplicado a variables de intimidad sexual:

GRÁFICO 11: Política_Intimidad_Hombres



	Pregunta	LIBERAL	CENTRO	CONSERVADOR
q325	Considerarías tener una relación abierta?	SI	NO	NO
q1597	Acostarte con alguien en la primera cita?	SI	≈ SI NO	NO
q1133	Tienes fantasías sexuales del tipo violación?	SI	NO	NO
q26	Has tenido alguna vez juguetes sexuales?	SI	≈ SI NO	NO
q1040	Recibir sexo anal?	Me gusta, Lo probaría	No me gusta, No lo probaría	No me gusta, No lo probaría
q30169	Dar sexo oral?	Me encanta!	≈ Me encanta! Está bien	+Está bien -No mucho

GRÁFICO 12: Política_Intimidad_Mujeres

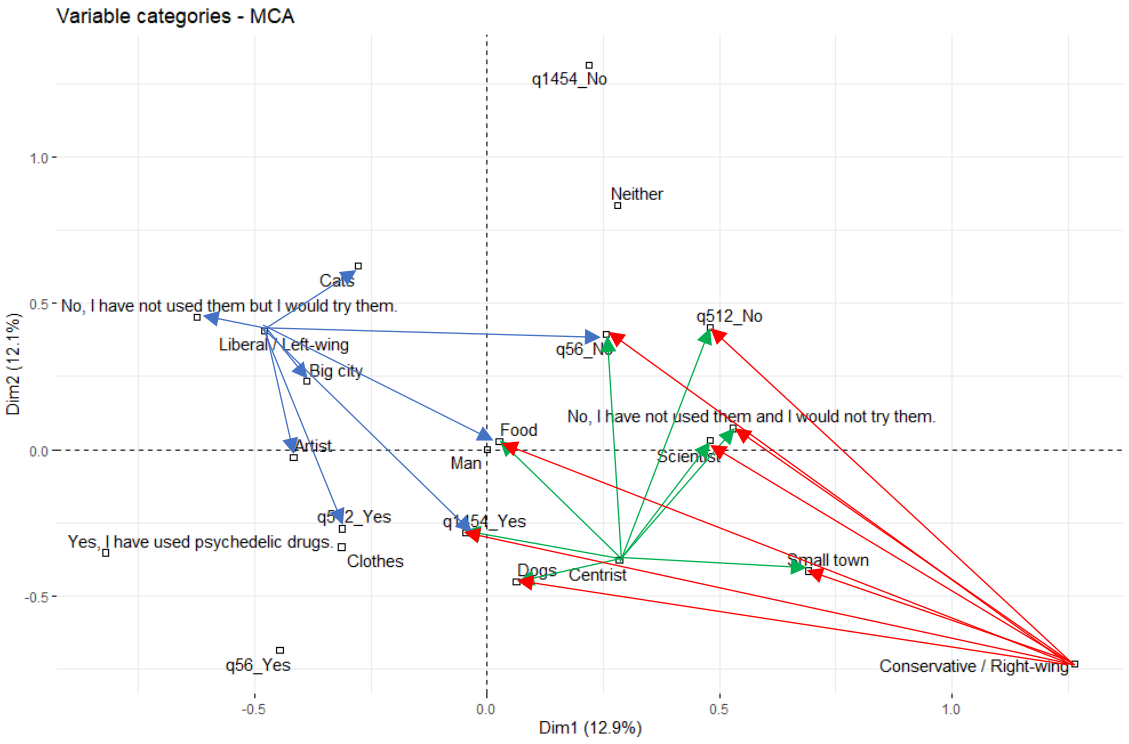


	Pregunta	LIBERAL	CENTRO	CONSERVADOR
q325	Considerarías tener una relación abierta?	SI	NO	NO
q1597	Acostarte con alguien en la primera cita?	SI	NO	NO
q1133	Tienes fantasías sexuales del tipo violación?	SI	NO	+NO
q26	Has tenido alguna vez juguetes sexuales?	SI	+SI	-NO
q1040	Recibir sexo anal?	Me gusta, Lo probaría	No me gusta, No lo probaría	+No me gusta, No lo probaría
q30169	Dar sexo oral?	Me encanta!	Está bien	≈ +Está bien -No mucho

No parecen existir muchas diferencias de distribución entre la distancia de las variables y la población hombre/mujer. Como mucho se puede observar un CENTRO en la población femenina más inclinado hacia la negación a diferencia de la población masculina donde redundantemente el CENTRO posee una opinión sobre sexualidad intermedia.

En cualquier caso, la apertura a sexualidad es claramente dominante en personas que se consideran de IZQUIERDA y más cerrada cuanto más a la DERECHA.

GRÁFICO 13: Política_Hobbies_Hombres



X	Pregunta
q512	Te gusta bailar?
q69912	Si tuvieras que elegir: Arte o ciencia?
q74	Gastas mas en comida o en ropa?
q15414	Has tomado drogas psicodélicas (LSD,peyote…)
q1454	Disfrutas de actividades al aire libre? (Camping, pesca, senderismo…)
q997	Eres una persona de gatos o perros?
q73	Gran ciudad o pequeño pueblo?
q56	Te atraen las situaciones peligrosas?

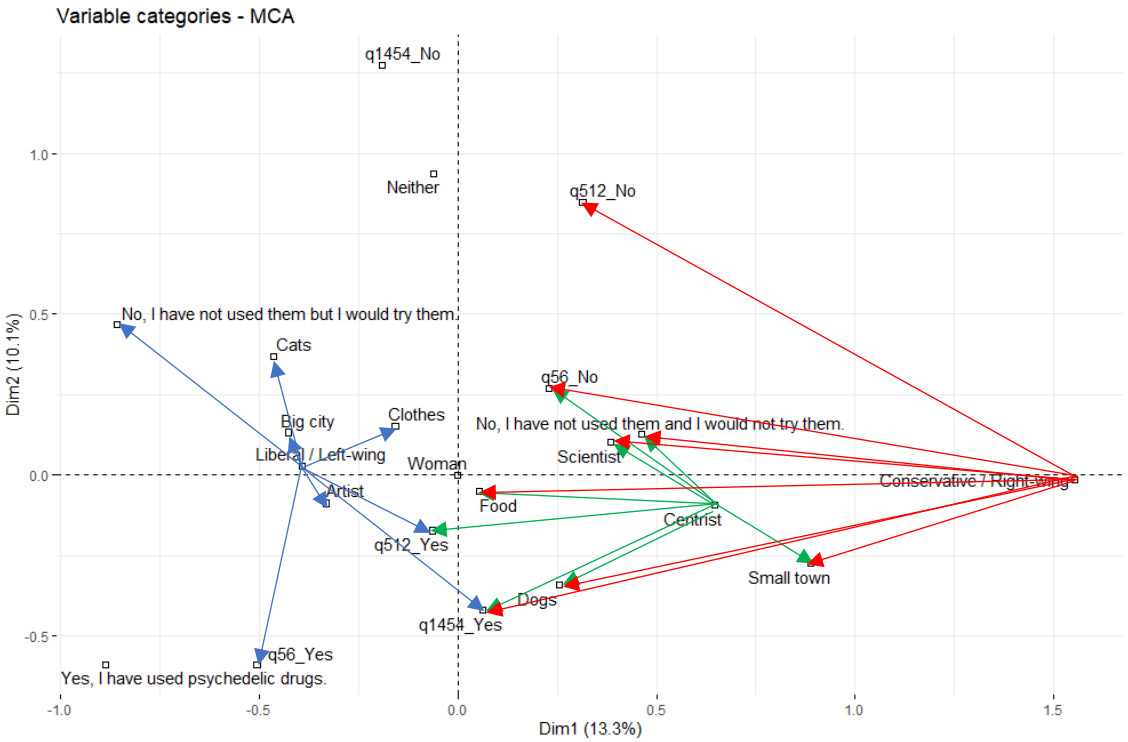
De las 8 preguntas, cada flecha reflejaría la pregunta más cercana a cada nivel político, esto permite además comprobar en que niveles hay coincidencias entre unos y otros.

Hombre liberal/izquierda: Gran ciudad, arte, gatos, gasta dinero en comida y le gusta bailar. No ha probado drogas psicodélicas aunque le gustaría hacerlo. No le atraen las situaciones peligrosas.

Hombre centro: Pequeño pueblo, ciencia, perros, gasta dinero en comida y le gusta bailar. No ha probado drogas psicodélicas ni le gustaría hacerlo. No le atraen las situaciones peligrosas.

Hombre conservador/derecha: -Exactamente igual que el de centro- Aunque cabe destacar que se encuentra más alejado de todas las variables.

GRÁFICO 14: Política_Hobbies_Mujeres



Mujer liberal/izquierda: Gran ciudad, arte, gatos, gasta dinero en comida antes que en ropa y le gusta bailar. No ha probado drogas psicodélicas aunque le gustaría hacerlo. Atraen las situaciones peligrosas

Mujer centro: Pequeño pueblo, ciencia, perros, gasta dinero en comida antes que en ropa y le gusta bailar. No ha probado drogas psicodélicas ni le gustaría hacerlo. No le atraen las situaciones peligrosas.

Mujer conservadora/derecha: -Exactamente igual que el de centro- Aunque cabe destacar que se encuentra más alejado de todas las variables.

Por último, cabe mencionar algunas cercanías interesantes en las dos poblaciones:

Tomar drogas psicodélicas – Atraen situaciones peligrosas

Perros – Actividades al aire libre

No tomar drogas psicodélicas – Preferencia por la Ciencia

Preferencia por el arte – Preferencia por la gran ciudad

Preferencia por los gatos – Liberal/Izquierdas

¿LOS GATOS SON DE IZQUIERDAS?

Hallando las proporciones marginales por columna se puede observar cómo se distribuyen las preferencias por cada animal. En la muestra, el 74,17% de los gatos se encuentran en hogares de liberales/izquierdas.

```
>V1=X$q212813
>V2=X$q997
> prop.table(table(V1,V2),margin=2)
```

	Both	Cats	Dogs	Neither
Centrist	0.25240445	0.19798264	0.28653827	0.24818653
Conservative / Right-wing	0.08982186	0.06028618	0.12845830	0.11865285
Liberal / Left-wing	0.65777369	**0.74173118**	0.58500342	0.63316062

Un dato curioso que podría analizarse en mayor profundidad; tal vez esto se deba a otras relaciones en conjunto que manipulen la varianza, por ejemplo:

Como se observaba en el análisis de correspondencias múltiple hay una clara cercanía entre: Izquierdas, ciudad y gatos. Analizando estas relaciones por separado:

	Both	**Cats**	Dogs	Neither
Big city	0.6575892	**0.7262129**	0.6661939	0.7280220
Small town	0.3424108	0.2737871	0.3338061	0.2719780

	Big city	Small town
Centrist	0.23554474	0.33131535
Conservative / Right-wing	0.05138684	0.19024032
Liberal / Left-wing	**0.71306843**	0.47844432

- Los gatos están más presentes en las ciudades que en los pueblos
- Los de izquierda están mas presentes en las ciudades que los de derechas.

Por teoría de conjuntos, es más probable encontrar un gato en un hogar de izquierdas.

Esto obviando una mayor multitud de variables, como puede ser la joven de edad de la muestra, su nivel socio-económico etc.

Un ejemplo claro de que una correlación no representa necesariamente una relación de causalidad y que deben atenderse con cuidado las colinealidades, *o quien sabe, tal vez adorar a los gatos te vuelve más liberal.*

5. CONCLUSIONES

Dada la inmensidad de variables presentes en la base de datos, se podrían escribir libros enteros sobre las diferencias entre los usuarios, aquí únicamente se recopilarán algunas de las presentadas más generales:

- **Género y Edad, distribución**

- Existen tres hombres por cada dos mujeres en la web
- Los hombres poseen mayor edad que las mujeres (4 años de diferencia de media)

- **Salario**

- A mayor salario, menor número de personas
- Acumulación de frecuencias en 250 mil, difícil de pasar a mayor cantidad.
- Sectores que menos ganan de media: Arte, Turismo y Administración
- Salario aumenta con el nivel de estudios hasta los 500 mil, donde el nivel de estudios pierde importancia

- **Diferencias de género**

- Los hombres se muestran menos exigentes con atributos que no sean el atractivo físico a la hora de salir con una mujer, no les importa su edad, si tienen coche, si son mucho más atractivas, si tienen metas en la vida[…] a diferencia de las mujeres que valoran todo lo anterior dicho. A excepción de la altura donde un carácter físico si importa mucho más para las mujeres.
- Los hombres juegan al ajedrez, no practican yoga, conducen a gran velocidad, saben programar el doble en proporción que las mujeres y solo la mitad posee juguetes sexuales. Además 3/4 se considera más inteligente que el resto.
- Las mujeres no juegan al ajedrez tanto como los hombres, una gran parte ha probado el yoga, no les interesan los coches o altas velocidades, la mayoría tiene o ha tenido juguetes sexuales y menos de la mitad se considera más inteligente que el resto.
- En la muestra existe una clara distribución a favor del salario para los hombres, la mayor proporción de mujeres se encuentra en niveles inferiores de salario y disminuye con una inclinación muy superior en comparación con los hombres a medida que aumenta el nivel de salario.

- **Inclinación política**

- Los de izquierdas votan y reciclan más, son más conflictivos con la autoridad y les interesa más la política que al resto, también son los menos permisivos respecto a otras inclinaciones.
- El centro por supuesto se encuentra en distancia de ACM entre derecha e izquierda, aunque con una inclinación más hacia la izquierda. Posee más niveles acordes a la duda o ausencia de extremismo: "A veces", "Creo que no", "No estoy seguro"…. También son los más permisivos a aceptar la opinión de otras inclinaciones.
- Los de derechas o conservadores son los menos interesados por la política y más cerrados respecto a libertad para prostitución o educación sexual.
- En la distribución de hombres se encuentra una relación lineal muy pronunciada desde el extremo de izquierda al de derecha, en la población de mujeres todas las variables o bien se encuentran más dispersas o se agrupan cerca de izquierdas. Parece ser que en las mujeres hay mayor colinealidad a medida que se acerca la izquierda y más dispersión según se acerque a la derecha, también poseen una agrupación mayor como "centro-izquierda"
- En cuanto a la intimidad sexual en hombres la distribución del ACM aquí es la que mejor representa el extremismo entre inclinaciones, los de izquierda están abiertos a probar de todo, los de centro no están seguros en nada y los de derecha los más cerrados o conservadores.

  En la población de mujeres se repite el mismo patrón aunque el centro se torna más conservador (a diferencia de en los ACM anteriores donde era más centro-izquierda). Parece ser que las mujeres son más liberales en temas políticos y más conservadoras a nivel intimo o sexual.
- En la vida diaria se encuentra cercanía entre centro y derecha en la población de hombres y nuevamente mayor cercanía entre izquierda-centro en la población de mujeres, aunque siendo la derecha la mas alejada del resto de variables. También se ven interesantes colinealidades: Tomar drogas y el peligro, la ciudad con izquierdas y el pueblo con derechas, los gatos con la gran ciudad, el arte con la ciudad y ser de izquierdas[…].
- Resulta interesante como conclusión final respecto a la inclinación política, que ésta determina muy bien como se comportará una persona, tanto a nivel intimo como en su vida diaria.

ANEXOS:

REFERENCIAS:

Alboukadel Kassambara(2017), Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning.

Antonio Pardo, Miguel Ángel Ruiz, Rafael San Martín. (2009). Análisis de datos en ciencias sociales y de la salud I. Madrid, España: SINTESIS.

Antonio Pardo, Rafael San Martín. (2015). Análisis de datos en ciencias sociales y de la salud Il. Madrid, España: SINTESIS.

Carlos J. Gil Bellosta(2018), R para profesionales de los datos: una introducción

Santiago de la Fuente Fernandez (2011), Análisis de correspondencia simple y múltiple

Kirkegaard, E. O. W., & Bjerrekær, J. D. (2016). The
okcupid dataset: A very large public dataset of dating site users. *Open Differential Psychology*.
Retrieved from https://openpsych.net/paper/46

Kirkegaard, E. O. W., & Bjerrekær, J. D. (2018). Self-reported criminal and anti-social behavior on a dating site: the importance of cognitive ability. Open Differential Psychology.
Retrieved from: https://openpsych.net/paper/55