Diplomado en Inteligencia de Datos en la Gestión de las Organizaciones

5ta Edición

EXPLORACIÓN DE PATRONES Y RELACIONES EN UN DATASET DE CITAS RÁPIDAS UTILIZANDO TÉCNICAS DE CIENCIA DE DATOS

Trabajo final

POR
Gaido, Horacio
Lorenzo, Laureano

Profesora Tutora:
Dra. López De Luise, María Daniela

Febrero 2.023

1 Introducción [1]

En un experimento para estudiar las preferencias en el mercado de citas rápidas, se recopilaron datos de 552 participantes entre 2.002 y 2.004 en la Universidad de Columbia, New York, Estados Unidos.

Durante los eventos, los asistentes tuvieron una "primera cita" de **cuatro minutos** con todos los demás participantes del **sexo opuesto**. Los participantes entablaron conversaciones de cuatro minutos para determinar si estaban interesados o no en volver a encontrarse. Si ambas personas "**aceptaban**", entonces a cada una se le proporcionaba posteriormente la información de contacto de la otra. Además, se le pidió a cada participante que califique a su compañero de cita de acuerdo a determinados atributos.

Queremos investigar y analizar las preferencias y valoraciones de hombres y mujeres en citas románticas rápidas, la confiabilidad de la autocalificación de las personas y las posibles diferencias entre los resultados obtenidos en otra investigación sobre el mismo experimento. La investigación busca profundizar en la comprensión de las preferencias y decisiones que influyen en la selección de una pareja para una nueva cita.

En la presente investigación, se exploran los datos generados en este experimento mediante técnicas de minería de datos y se contrastan los resultados con aquellos obtenidos por los investigadores que diseñaron el mismo.

2 Definición de Objetivo, Límites y Alcance del Trabajo

2.1 Objetivo General

Analizar el dataset para descubrir cómo se relacionan las variables y cómo influye cada una sobre las decisiones de los participantes.

2.2 Objetivos específicos

1. Investigar qué características eligen más los hombres, por un lado, y las mujeres por el otro al conocer a su pareja durante 4 minutos, para aceptar tener una nueva cita con esa persona, considerando el supuesto que ambas personas sean heterosexuales. Además, examinar las diferencias de género en las preferencias de citas y la valoración de los atributos por parte de hombres y mujeres.
2. Investigar en base a la autocalificación de una persona qué diferencia hay en cómo la califica su pareja. Se busca determinar el grado de confiabilidad de las autocalificaciones. También estimar las probabilidades que tiene un individuo de que otra persona lo elija para tener una nueva cita.
3. A través de herramientas de ciencias de datos, poner en evidencia las posibles diferencias y contrastes entre las conclusiones obtenidas y el trabajo realizado en la Universidad de Columbia en el 2.006. Particularmente, sus conclusiones sobre cómo la raza y el género influyen sobre las decisiones de los participantes.

---

1 Raymond Fisman; Sheena S. Iyengar; Emir Kamenica; Itamar Simonson. Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment. The Quarterly Journal of Economics, Volume 121, Issue 2, May 2006, Pages 673–697, https://doi.org/10.1162/qjec.2006.121.2.673 Published: 01 May 2006

2.3 Protocolo de captura

Límites:
En el estudio de la Universidad de Columbia, los investigadores estiman el ingreso y la densidad poblacional de la zona de origen de los participantes, junto con las calificaciones promedio de la institución a la que pertenecen. Al no contar con esta información, se omiten este tipo de variables en el análisis.

Queremos resaltar, que solo se trabaja con parejas heterosexuales. En el sexo opuesto existe un bias heterosexual, es una decisión explícita dentro del experimento.

La duración de las citas rápidas es de 4 minutos y cada participante tiene 1 minuto después de la cita para evaluar a su pareja. El motivo por el cual no se estableció más tiempo para cada cita rápida, es que por cada ronda del experimento asistían entre 18 y 42 personas, esto quiere decir que se juntaban con entre 9 y 21 personas distintas por ronda (noche). La persona que menos tiempo estuvo en una ronda es de 54 minutos aproximadamente (9 parejas x durante 5 minutos, más algún tiempo por pequeñas demoras) y la persona que más tiempo demoró en una ronda fue de 126 minutos (2 horas y 6 minutos). Se consideró que ese tiempo era el correcto para no distorsionar el juicio de los participantes.

Sujetos:
Los sujetos en este estudio proceden de estudiantes de posgrados y escuelas profesionales de la Universidad de Columbia. Los participantes fueron reclutados a través de una combinación de correo electrónico masivo y volantes colocados en todo el campus y repartidos por asistentes de investigación. Para inscribirse en los eventos de "Citas Rápidas", los estudiantes interesados tenían que registrarse en un sitio web en el que informaron sus nombres y direcciones de correo electrónico y completaron una encuesta previa al evento.

Entorno:
Los eventos de "Citas Rápidas" se llevaron a cabo en una habitación cerrada dentro de un bar / restaurante popular cerca del campus. La disposición de las mesas, la iluminación y el tipo y volumen de la música reproducida se mantuvieron constantes en todos los eventos.

Procedimiento:
Los eventos se llevaron a cabo entre 2002 y 2004. Los participantes fueron distribuidos aleatoriamente al inicio, para luego ir rotando y conocer a distintos compañeros. Estos no conocían la cantidad de parejas que se encontrarían en el evento. Se hicieron un total de 21 sesiones. Luego de las citas, a cada participante se le pidió que indicara su valoración sobre ciertos aspectos de su compañero. Estos incluían el atractivo, la sinceridad, la inteligencia, lo divertido que era, la ambición y los intereses compartidos con éste. Finalmente, el participante debía decidir, si tener o no, una segunda cita con su potencial pareja.

3 Delimitación de la Situación Problemática

Numerosos estudios señalan que existen diferencias en las preferencias entre grupos sociales a la hora de buscar pareja. Son diversos los factores que influyen en este tipo de decisiones, y frecuentemente se observa que los individuos revelan preferencias distintas a aquellas que declaran, por ejemplo, en páginas de citas *online.*

(Hitsch et al., 2.010), estudian el mercado de citas electrónicas en Boston y San Diego, y descubren algunos comportamientos interesantes. Mientras que ambos sexos le dan importancia al atractivo físico en la selección, las mujeres tienden a priorizar el estatus socioeconómico. Y ambos grupos suelen elegir potenciales compañeros de la misma raza. Sus conclusiones coinciden con las de trabajos anteriores sobre poblaciones similares.

Por otro lado, (Whyte y Torgler, 2.017), y a diferencia de otros estudios, argumentan que la característica diferenciadora entre ambos sexos surge de su grado de selectividad al buscar pareja. Las mujeres tienden a ser más exigentes en algunas circunstancias. También centran su análisis en las diferencias observadas entre aquello que manifiestan los participantes en sus perfiles, y lo que efectivamente guía sus decisiones al momento de contactar un potencial compañero. Aunque la población que estudian proviene en este caso de un sitio de Australia.

(Xixian y Haibo, 2.019), examinan la conducta de los candidatos en el populoso mercado de citas chino. Su método se distingue de la literatura tradicional porque en lugar de analizar las relaciones entre variables usando técnicas de la econometría, aplican herramientas de aprendizaje automático. Sus principales hallazgos giran en torno al rol de la popularidad: para ambos sexos esta constituye el predictor más fuerte de contacto con una potencial pareja. Y, consistente con el resto de trabajos, encuentran que las mujeres tienden a preferir parejas de mayores ingresos, mientras que los hombres prefieren parejas más jóvenes.

(Fisman et al., 2.006), examinan nuevamente el presente conjunto de datos, pero poniendo énfasis en las preferencias raciales de los sujetos. Aunque los individuos exhiben criterios de selección más tolerantes si se los compara con la población de Estados Unidos de ese momento, la inclinación por parejas de la misma raza sigue siendo fuerte, particularmente para las mujeres.

El presente trabajo se limita a una muestra algo heterogénea de estudiantes de grado y posgrado provenientes de Manhattan, pero algunos patrones se repiten en distintos grupos a lo largo del mundo. En particular, se busca estudiar la diferencia entre las preferencias y el nivel de selectividad de hombres y mujeres, y la inconsistencia entre lo declarado y las preferencias reveladas por los participantes.

4 Descripción de la Propuesta

Para el objetivo 1, se plantea usar heurísticos y modelos estadísticos para evaluar el peso relativo de cada variable como predictora de la decisión. Para aquellos objetivos en los que podemos comparar dos o más subgrupos dentro de la muestra (objetivo 2) se propone hacer uso de estadísticos y pruebas de hipótesis para refutar o confirmar nuestros planteamientos.

En cuanto al objetivo 3, comparar los resultados obtenidos al emplear los heurísticos mediante técnicas de interpretación de modelos.

5 Descripción del Conjunto de Datos Seleccionado

5.1 Enunciación y Descripción de cada Variable Involucrada:

El *dataset* contiene 8.378 observaciones y 123 columnas. Algunas columnas son cálculos en base a otras columnas. Por ejemplo, si el sujeto y su compañero toman el mismo valor en la columna “race”, la columna “samerace” toma el valor 1, etc. Mirando la composición de las columnas, algunas variables son el “espejo” de otras. Esto sucede porque a cada participante se le pidió que califique a su compañero en varios aspectos. Luego, al aparecer también la calificación recibida de la otra persona, hay datos que aparecerán en dos filas necesariamente, pero en una como calificación al compañero y en la otra como calificación del compañero.

El dataset consiste en 123 columnas. De estas, 66 son respuestas que cada sujeto debía contestar en distintos puntos del proceso y 57 categorizaciones de estas respuestas, construidas separando en *bines*. Estos campos incluyen:

- datos demográficos
- hábitos de citas
- autopercepción a través de atributos clave
- creencias sobre lo que otros encuentran valioso en una pareja
- información sobre el estilo de vida

En el apéndice A, sección 1, se detallan las distintas columnas y sus descripciones.

Algunas columnas son generadas por los propios investigadores para separar ciertas variables en subgrupos. Estas no son consideradas para el análisis. Se identifican en el *dataset* mediante el prefijo “d_”.

5.2 Descripción Estadística Básica:

Antes de continuar, cabe aclarar que para el análisis estadístico consideramos como observación a cada cita en lugar de cada participante, aunque los resultados son similares para ambos conjuntos de datos.

Análisis de Correlación

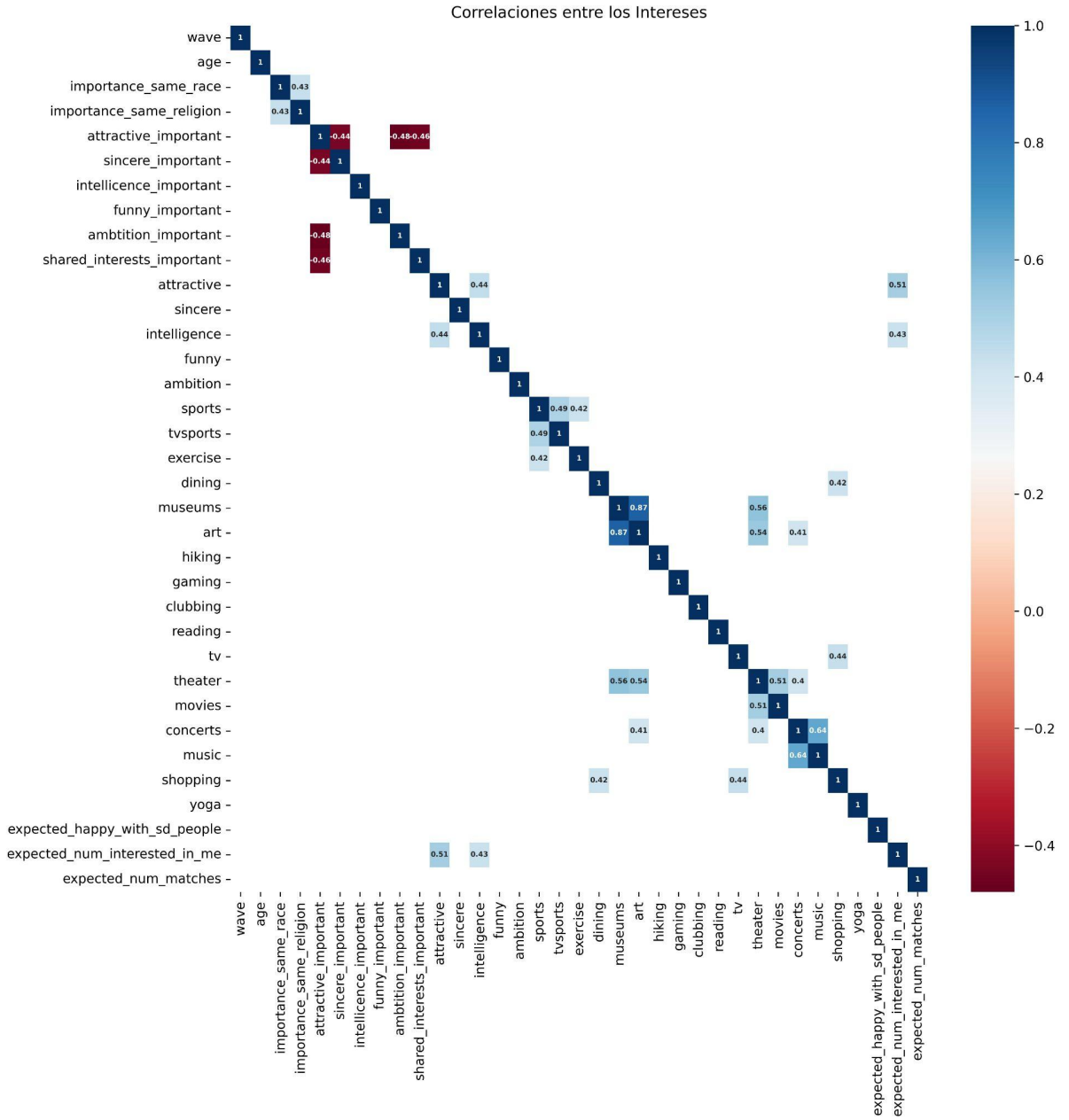
Para evaluar si la correlación entre dos variables es significativa, fijamos el umbral en 0.4 en valor absoluto. Coeficientes de correlación superiores a 0.4 ya pueden considerarse moderadamente positivos (o negativos) en este contexto.

Entre los atributos auto descriptivos, no encontramos gran asociación. Destaca inteligencia y atractivo, con un coeficiente de correlación de 0.44. Las demás tienen un coeficiente positivo y cercano a 0. Pero al analizar cómo cada participante puntúa a su compañero, aparecen algunos patrones un poco más fuertes. Sinceridad e inteligencia tienen la mayor correlación (0.66), seguido por ambición e inteligencia (0.63), gracioso e

intereses compartidos (0.62), y gracioso y atractivo (0.59). Luego cae bastante la correlación, pero todas resultan positivas.

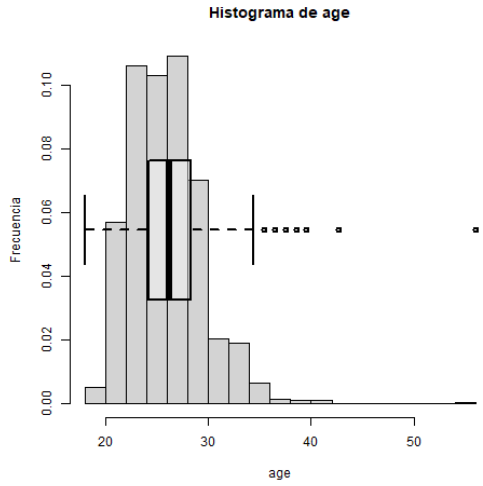
En cuanto a los intereses propios que reporta cada participante, la mayoría de los coeficientes son insignificantes. Aunque hay una fuerte correlación entre arte y museos (0.86), seguido por conciertos y música (0.66), teatro y museos (0.55), teatro y arte (0.53), y finalmente deportes y deportes por televisión (0.48). Si se cruzan con otras variables como la edad, importancia que el participante le da a la raza, religión, entre otras, los coeficientes por lo general son también bajos. Aunque sí aparece una correlación positiva reveladora (0.48) entre si al sujeto le gustó su compañero y si piensa que el sentimiento es recíproco.

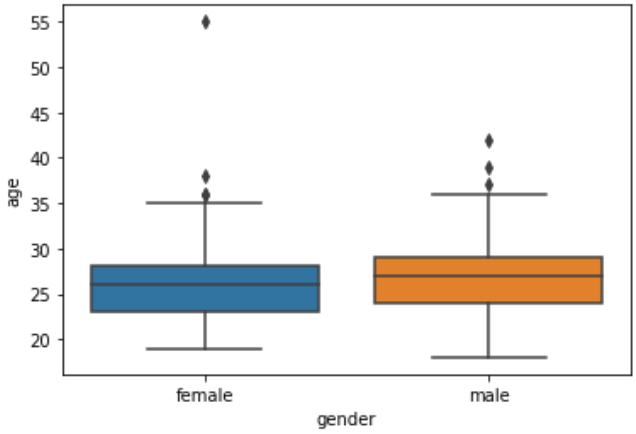
En la siguiente figura se exponen las correlaciones entre las variables reportadas por el propio participante:

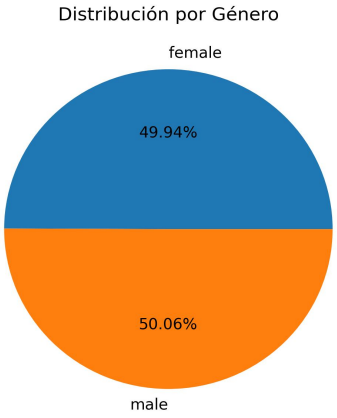


Análisis de Frecuencias

A continuación, se muestran las distribuciones de la edad, género y raza de la muestra:

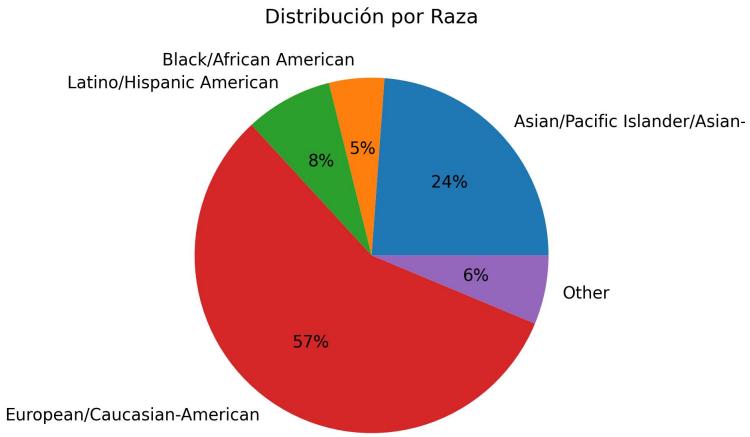






La mayoría de los participantes se encuentra entre los 18 y los 35 años, siendo las personas con más edad posibles candidatos a outlier en esta distribución.

La próxima figura muestra los porcentajes por raza de los participantes:



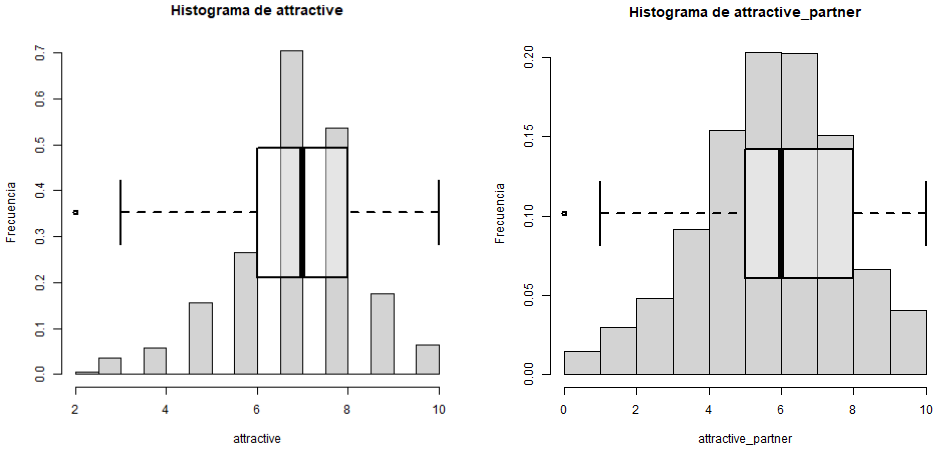
Ciertas subpoblaciones parecen no tener la suficiente representación. De acuerdo con la investigación realizada en la misma universidad[2], la muestra refleja en buena medida el contexto etnográfico del momento. Sin embargo, los datos del censo 2.010 de la ciudad de Nueva York (dato disponible más cercano), las proporciones del experimento difieren notablemente en algunos porcentajes poblacionales de razas.

En el censo de 2.010 el estado de Nueva York tenía una distribución racial de:[3]

- Blancos: 65,7 %.
- Negros o afroestadounidenses: 15,9 %.
- Asiáticos: 7,3 %
- Nativos americanos: 0,6 %.
- Otras razas: 7,4 %.
- Dos o más razas: 3,0 %.

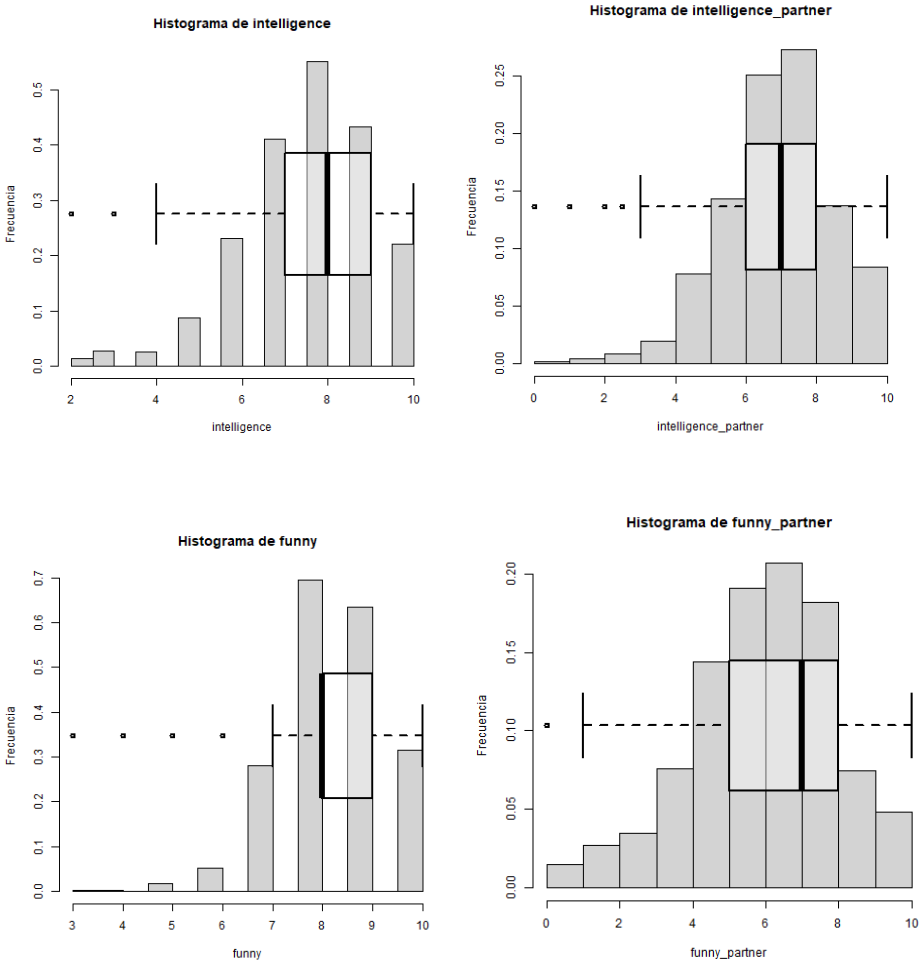
Comparando los porcentajes, se encuentra que hay una diferencia negativa de 8,7 % para la raza blanca, o un 13% menos en valores relativos. Para la población de raza afroamericana, hay una diferencia de 10,9%, que significa un 218% menos en valores relativos. Finalmente, la población de raza asiática aparece sobrerrepresentada en un 16,7%. En términos relativos supone un 228% más que la proporción en el estado de Nueva York en ese año.

También se analizan las distribuciones de algunas respuestas de los participantes respecto a qué tan atractivos, inteligentes, etc. se perciben, y se comparan con los puntajes que les asignan a sus compañeros de cita:



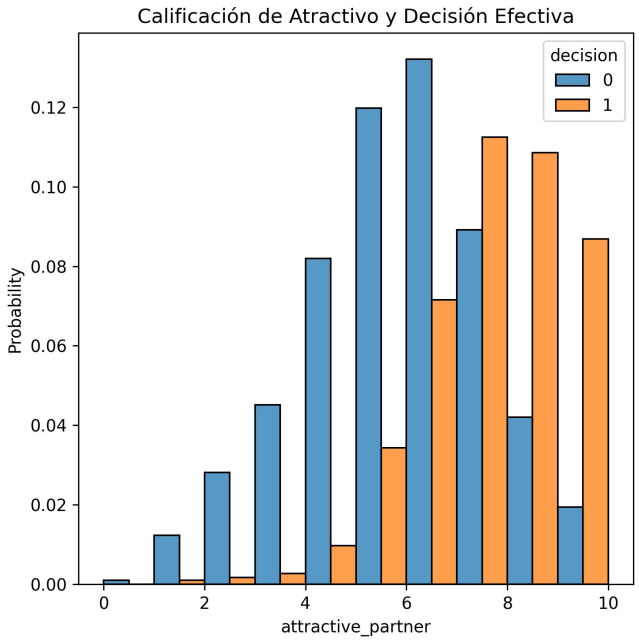
2 Raymond Fisman, Sheena S. Iyengar, Emir Kamenica Ita, Mar Simonson (2006). *Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment*. The Quarterly Journal of Economics.

3 Oficina del Censo de los Estados Unidos. «2010 Census Data» (en inglés). Consultado el 4 de noviembre de 2016. (https://www.census.gov/2010census/data/) (https://es.wikipedia.org/wiki/Nueva_York_(estado)#cite_note-c2010-5)

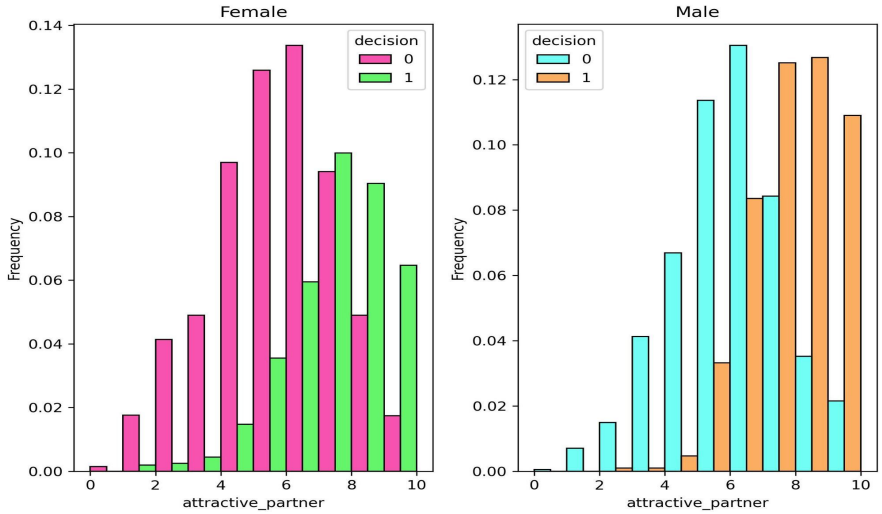


Si se examina la importancia que le da cada participante a distintos atributos la columna "**attractive_important**" tiene una media de **22,42%**. Pero debido a cómo está construída la encuesta, al compararla con el resto de atributos, es la que tiene el mayor peso relativo en promedio. Esto claramente se refleja en las decisiones de los participantes: al comparar el puntaje en atractivo que el participante asignó a su compañero en función de si el primero decidió o no tener una segunda cita, aquellos que efectivamente decidieron salir de nuevo con su compañero lo calificaron en promedio con **7.28** puntos sobre **10**. Por otro lado, quienes decidieron no ver a su compañero una segunda vez, los calificaron en promedio con **5.36** puntos sobre **10**.

En el siguiente gráfico se pueden observar las distribuciones por calificación de atractivo y decisión:



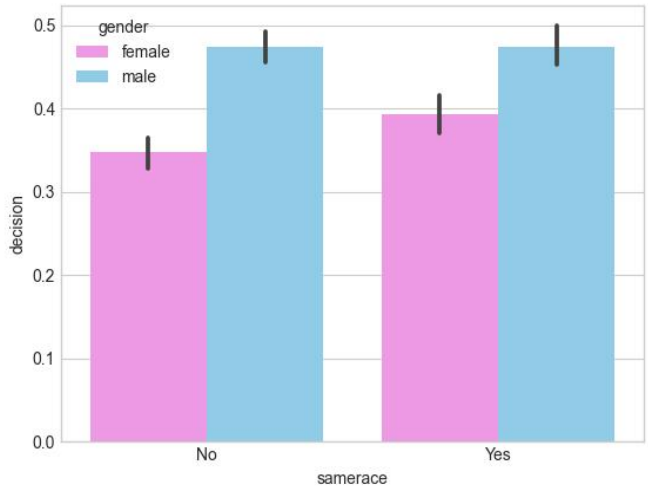
El fenómeno parece ser incluso más fuerte en los hombres que en las mujeres:



Las distribuciones del resto de atributos resultan similares, siendo el atributo "**divertido**" el que más impacta sobre la decisión. En el apéndice A, sección 2, se hallan las figuras correspondientes.

Este análisis se realiza sobre las columnas tal y como aparecen en el dataset, sin ningún tipo de normalización. En secciones posteriores se propone una transformación a una escala más objetiva.

Finalmente, hay dos variables más que fueron ordenadas en una escala de importancia en el experimento:
1), Que el compañero sea de la misma raza y 2), que sea de la misma religión. Lamentablemente no se preguntó a los participantes su religión y por lo tanto no se puede ver el efecto que este atributo tiene sobre la decisión de buscar o no una segunda cita. En la siguiente figura se exhibe el promedio de decisiones negativas (0) y positivas (1) de acuerdo a si los dos sujetos comparten raza. También se hace una segmentación por género:

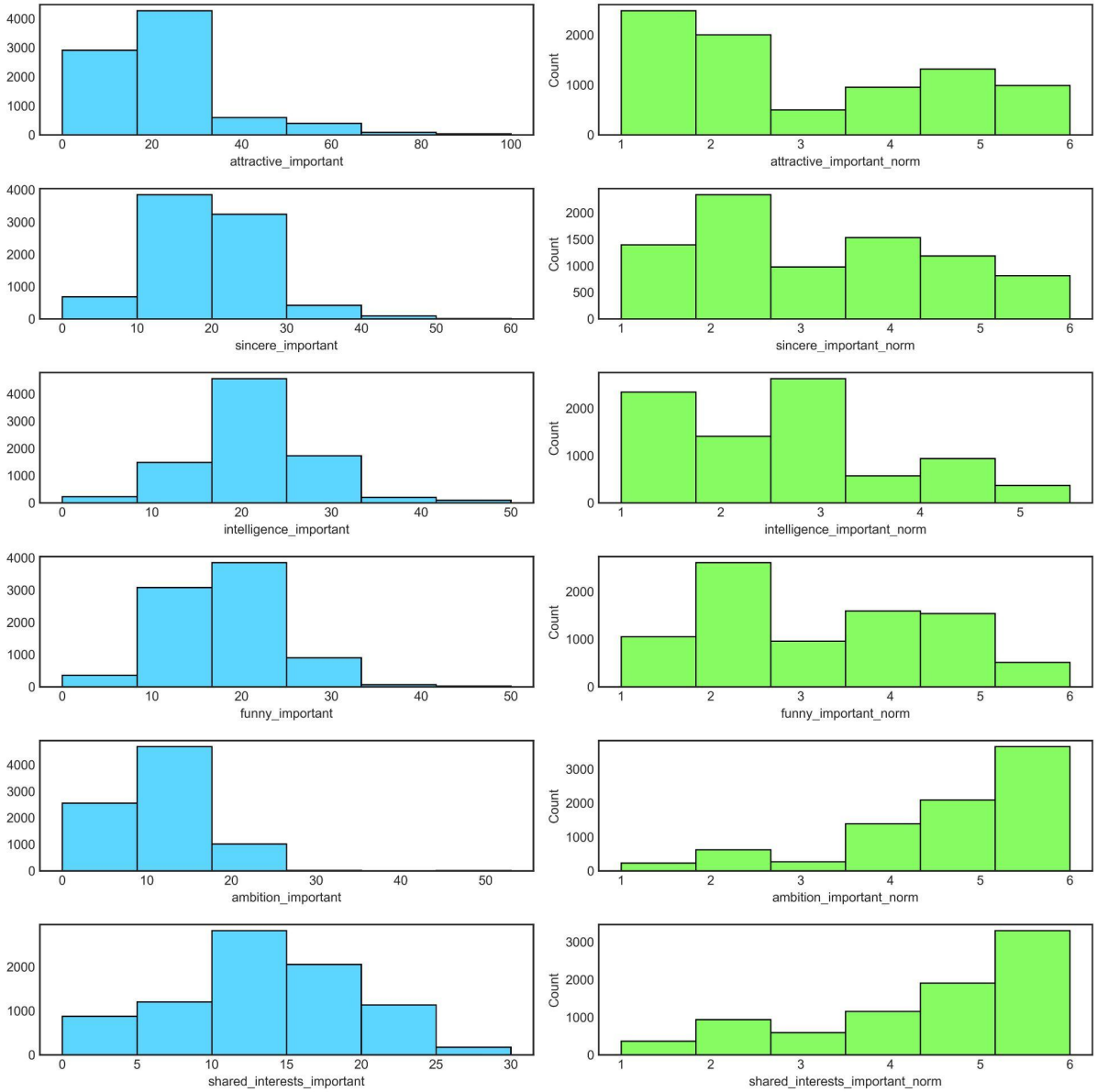


<u>Normalización de las columnas sobre la importancia de los atributos</u>

Para este proceso, se descarta la escala de importancia original. Las nuevas columnas aún representan una clasificación relativa con respecto al resto, pero ahora las puntuaciones son acotadas a una escala que va del **1** (mayor puntaje relativo) al **6** (menor puntaje relativo). La clasificación nueva que recibe cada variable depende entonces de la cantidad de atributos que se ubican por debajo de éste en términos de importancia relativa.

Por otro lado, las columnas que tenían los mismos puntos que otras conservan su posición relativa (no se fuerza otro tipo de ordenamiento, como alfabético), pero ahora se les asigna un puntaje que corresponde al promedio entre los 2 (o más) valores que hubieran recibido de no estar empatadas. En la siguiente figura se muestran los histogramas de las escalas originales y su versión normalizada:

Escalas de Importancia



Del análisis gráfico surgen diferencias evidentes entre las escalas originales y las escalas normalizadas. Los primeros atributos ahora se distribuyen bastante uniformemente, mientras que los últimos dos presentan altos niveles de asimetría. La visualización de estos fenómenos se vuelve complicada si utilizamos las escalas originales.

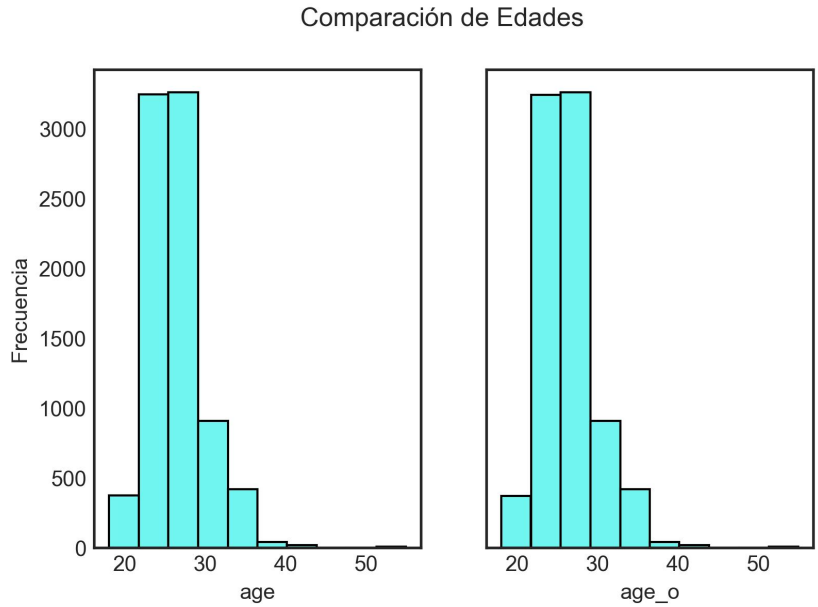
5.3 Análisis de las problemáticas fundamentales de datos:

Consistencia

En el trabajo se descubre que a lo largo del dataset existen instancias cuyos valores no se ajustan a las escalas predeterminadas en la encuesta. A modo de ejemplo, se le pedía al participante que puntúe a su compañero del 1 al 10, pero lo puntúa con un 10.5 u 11. Se adjuntan algunos de estos casos en el apéndice A, sección 3.

Se identifican y etiquetan estas instancias de forma sistemática para el análisis posterior. Algunas de las columnas referidas a intereses de los participantes presentan valores levemente fuera de la escala, pero la mayoría de inconsistencias proviene de la suma de las importancias que le dan a cada atributo (deberían sumar 100 puntos, pero suman valores muy cercanos como 99.99, 100.01, o directamente valores muy alejados como 120, 110, entre otros).

Aunque muchas de las columnas no presentan este tipo de inconvenientes, se inspecciona gráficamente la relación entre aquellas variables “espejo”. En la próxima figura se expone la comparación entre uno de estos pares, a saber, las edades de los participantes:



Se identifican y etiquetan las instancias erróneas de forma sistemática para el análisis posterior. Algunas de las columnas referidas a intereses de los participantes presentan valores levemente fuera de la escala, pero la mayoría de inconsistencias proviene de la suma de las importancias que los participantes le dan a cada atributo. Se aplica el procedimiento para cada columna:

	Count
Inconsistente	
No	7567
Sí	811

Y gran parte se explica por valores distintos a 100 al sumar las columnas de importancia relativa:

Columna	Count
Consistente	7567
Important_Suma	691
gaming	62
reading	51
met	5
attractive_o	1
funny_o	1

Ruido y Valores Atípicos

Para la detección de valores atípicos, se rellenan algunos valores faltantes en columnas numéricas utilizando herramientas de imputación multivariada[4]. De esta forma, los valores faltantes son estimados a través del aprendizaje automático supervisado. En cada caso, se utiliza el resto de columnas como predictoras para los valores ausentes.

En el proceso de detección, a cada observación se le asigna un puntaje de valor atípico (*outlier score* en inglés). Finalmente, aquellas observaciones con valores extremos en esta escala son clasificadas como ***outliers***.

El heurístico empleado para la determinación de este puntaje es el de los "k" vecinos más cercanos (*KNN* por sus siglas en inglés). Para cada dato, se toman las "k" menores distancias euclídeas hasta vecinos, y se promedian. Esto resulta en una medición que refleja la proximidad del dato respecto a la muestra. En nuestro caso, el valor final se obtiene promediando a su vez entre un número de puntajes para distintos valores de *k:*

$$s_i = \frac{1}{n}\sum_{k=1}^{n} \square l_{k,i}; \forall\ i \in \Omega$$

siendo Ω el universo de observaciones, k el número de vecinos considerados para el cálculo, *n* el número máximo de vecinos (10 en el presente estudio), y $l_{k,i}$ cada distancia euclídea.

Se elige un método no paramétrico para la detección porque en este caso es especialmente difícil la selección de un modelo cuya forma funcional se parezca a la de la población. Aunque la cantidad de variables seleccionadas es muy elevada, y esto puede deteriorar el rendimiento del algoritmo ***KNN,*** el *dataset* contiene suficientes observaciones para evitar problemas de alta dimensionalidad[5]. Además, no hay indicios a priori de que los datos puedan ajustarse a un modelo paramétrico conocido.

Por otro lado, debe tenerse en cuenta que, al haber múltiples consideraciones subjetivas en la construcción de las variables, posiblemente se esté introduciendo sesgo al utilizar esta métrica de distancia para evaluar la similaridad entre las observaciones.

---

4 Tarek A. Atwan (2022). *Time Series with Python.* Packt Publishing, Birmingham-Mumbai.
5 James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (1st ed.) Springer.

Finalmente, se detectan los valores extremos en esta escala mediante los rangos intercuartílicos. El uso de distintos valores de "*k*" para su cálculo, en contraposición a un valor arbitrario, hace que las distancias obtenidas (y sus respectivas clasificaciones) sean más confiables[6].

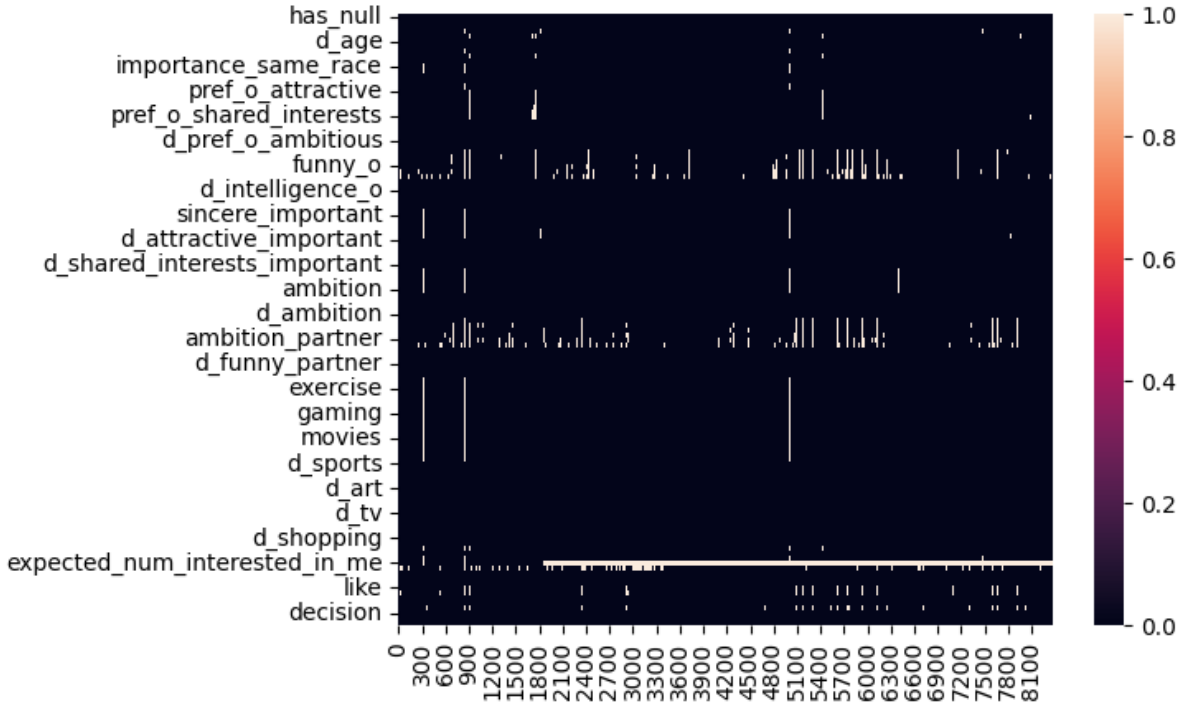
De acuerdo con los resultados obtenidos, hay **101** observaciones que pueden considerarse como extremas en la escala de ***outlier score***. Ningún participante concentra más de **3** de ellas en sus citas, y no podemos descartar la posibilidad de que estas observaciones sean simplemente desvíos estocásticos.

Desafortunadamente, este método no nos permite distinguir entre valores atípicos puros y simple ruido estadístico. Pero nos ayuda a aislar aquellas observaciones que pueden resultar conflictivas más adelante.

Finalmente, se comparan las observaciones inconsistentes con los posibles outliers detectados. Sólo **2** observaciones del primer grupo presentan un valor extremo en la puntuación de outlier. Esto ayuda a confirmar que los puntos identificados son atípicos: **no presentan características similares entre ellos.**

Valores Faltantes

Los puntos blancos representan los datos faltantes por cada una de las características.



6 Charu C. Aggarwal. (2017). *Outlier Analysis* (2da edición). Springer International, New York.

En total hay **18.372** datos faltantes, pero no están distribuídos entre las columnas en forma pareja. A continuación las **10** columnas con más datos faltantes:

Columna	Datos Faltantes	Datos Faltantes (%)
expected_num_interested_in_me	**6.578**	**78.52 %**
expected_num_matches	**1.173**	**14.00 %**
shared_interests_o	**1.076**	**12.84 %**
shared_interests_partner	**1.067**	**12.74 %**
ambitious_o	722	8.62 %
ambition_partner	712	8.50 %
met	375	4.48 %
funny_o	360	4.30 %
funny_partner	350	4.18 %
guess_prob_liked	309	3.69 %

Premisas y supuestos para el análisis

El *dataset* y la forma en que se decidió elaborar el experimento presentan ciertas cualidades que dificultan la generalización de los resultados obtenidos:

- En primer lugar, los sujetos del estudio decidieron por su cuenta participar en el evento. Esto genera un sesgo de autoselección, y por lo tanto, la muestra adquirida no es tan representativa.
- Por otra parte, intervinieron únicamente individuos heterosexuales en el estudio. Esto contribuye a deteriorar el nivel de representatividad de la muestra.
- Finalmente, el experimento se llevó a cabo en la Universidad de Columbia y por lo tanto los participantes provenían de un estrato socioeconómico distinto al ciudadano promedio de Estados Unidos.

Aunque algunos de los resultados de la investigación son compatibles con estudios realizados sobre otras poblaciones, se trabaja bajo el supuesto de que no se pueden extrapolar las conclusiones, al menos no a la población en general.

6 Aplicación de Pruebas de Hipótesis

6.1 Descripción del pre-procesamiento de datos realizado

Uno de los objetivos que nos hemos planteado es analizar las diferencias que hay entre la autoevaluación de una característica personal versus como la otra persona de la pareja evalúa al sujeto. Queremos conocer si una persona se autocalifica, mejor, igual o peor que como lo calificaría la mayoría de otras personas.

Para poder cumplir con este objetivo primero tenemos que seleccionar un grupo de características, en este trabajo se elegirán las siguientes:

Características de autoevaluación:

- attractive: del 1 al 10 cuán atractivo te autocalificas.
- sincere: del 1 al 10 cuán sincero te autocalificas.
- intelligence: del 1 al 10 cuán inteligente te autocalificas.
- funny: del 1 al 10 cuán divertido te autocalificas.
- ambition: del 1 al 10 cuán ambicioso te autocalificas.

Características de evaluación de la pareja:

- atractivo_o: del 1 al 10 cuán atractivo te calificó tu pareja.
- sincere_o: del 1 al 10 cuán sincero/a calificó tu pareja.
- intelligence_o: del 1 al 10 cuán inteligente calificó tu pareja.
- funny_o: del 1 al 10 cuán divertido/a calificó tu pareja.
- ambitus_o: del 1 al 10 cuán ambicioso/a calificó tu pareja.

El dataset analizado **NO** posee número de identificación de cada participante (id) por lo que se tuvieron que realizar varios cálculos para poder identificarlos.

Primero se tuvieron que convertir todos los números a tipo de dato “float” o “integer” ya que en el dataset los valores numéricos figuraban como “object”. Se conocía que la autoevaluación de cada persona es la misma para cada ronda, se realizó un cálculo para multiplicar la sumatoria de la calificación de cada interés por algún tema y se le multiplicó por un número negativo para que se pueda diferenciar de los otros valores numéricos que son todos positivos. Se generó una nueva columna que se generaba 1 (uno) cuando había un cambio de valor del cálculo y eso identificaba que cambiaba la persona. Luego, se realizó otro feature con el id de cada participante del experimento. Con este procedimiento pudimos identificar a cada persona. Todos estos cálculos aparecen detallados en una jupyter notebook, que se adjunta al trabajo.

En el experimento se realizaron 21 rondas, en donde por cada ronda, cada sujeto se reunía con entre 9 y 21 personas.

Para poder comparar la autocalificación con la calificación del otro sujeto, lo que se hizo es por cada persona, tomar el promedio de la calificación de todos los otros participantes en esa ronda, para poder tomar valores promedios que sean lo menos sesgados posibles.

Por cada una de las 5 características (atractivo_o, sincere_o, intelligence_o, funny_o, mbitus_o) se generó una nueva columna en el dataset que contiene el promedio de las calificaciones que recibió esa persona en esa ronda. Después, se eliminaron todos los id duplicados y se formó un subset con contine 11 columnas que son el id de cada participante y la autocalificación de cada participante en cada una de las 5 características y el promedio de las calificaciones de sus parejas en cada ronda respecto a esas 5 características.

Luego de ese procedimiento, se generaron datos para comparar en parejas los datos para poder determinar si fehacientemente había diferencias entre la autocalificación de una persona de una cierta característica versus el promedio de la evaluación de sus parejas en sus respectivas rondas.
Las comparaciones se hicieron de la siguiente forma:

- ATRACTIVO: "attractive" vs "attractive_o"
- SINCERIDAD: "sincere" vs "sinsere_o"
- INTELIGENCIA: "intelligence" vs "intelligence_o"]
- DIVERTIDO: "funny" vs "funny_o"
- AMBICIOSO: "ambition" vs "ambitous_o"

Para poder hacer los test de comparación primero se tuvo que determinar qué tipo de **distribución de probabilidades** posee cada una de las 10 variables, para poder luego determinar qué **tipo de test** se puede aplicar para poder realizar las pruebas de hipótesis que permitan determinar si hay o no hay diferencias entre las autoevaluaciones y las evaluaciones de las parejas.

La librería **Scipy** de Python incluye una función llamada **normaltest**, que se utiliza para realizar una **prueba de normalidad** en una muestra de datos. Esta función se basa en la prueba de D'Agostino, que se utiliza para determinar si una muestra de datos sigue una distribución normal. La prueba de normalidad es importante en estadística porque muchos análisis estadísticos requieren que los datos se distribuyan normalmente para que los resultados sean precisos y confiables. Además, se utiliza para determinar si los datos pueden ser modelados por una distribución normal. La función normaltest devuelve dos valores: el estadístico de prueba y el valor p. El estadístico de prueba se utiliza para calcular el valor p, que es la probabilidad de obtener un resultado tan extremo o más extremo que el observado si la muestra de datos sigue una distribución normal. Si el valor p es menor que un nivel de significancia predeterminado (como 0.05 o 0.01), se rechaza la hipótesis nula de que la muestra sigue una distribución normal.

6.2 Detalle de la aplicación de las técnicas, estrategias de verificación y validación, junto especificación del modelo obtenido.

6.2.1 Comprar cada una de las 5 variables de autoevaluación, con el valor que evaluó su pareja respecto de la misma variable, para contrastar si hay o no diferencias.

Para hacer los análisis de normalidad usamos 3 metodologías distintas: representaciones gráficas, métodos analíticos y test de hipótesis.

Métodos gráficos: uno de los métodos gráficos más empleados para el análisis de normalidad consiste en representar los datos mediante un histograma y superponer la curva de una distribución normal con la misma media y desviación estándar que los datos disponibles, eso fue lo que se hizo para cada variable.

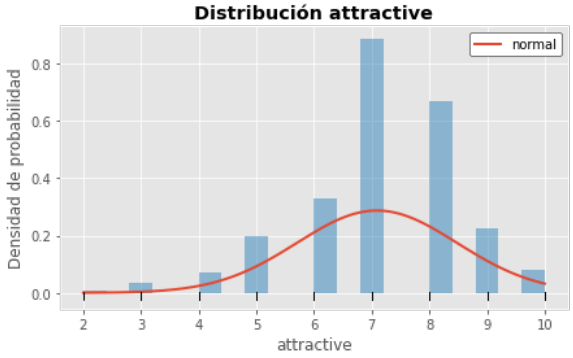
Métodos analíticos, asimetría y curtosis: los estadísticos de asimetría (Skewness) y curtosis pueden emplearse para detectar desviaciones de la normalidad. Para determinar normalidad, ambos valores deben estar en ± 0.5. Si cualquiera de los dos valores, ya sea el sesgo o la curtosis, están fuera del rango de ± 0.5, se asume que la muestra está sesgada y, por lo tanto, tiene una distribución distinta a la normal, es decir, es de libre distribución

Contraste de hipótesis: los test Shapiro-Wilk test y D'Agostino's K-squared test son dos de los test de hipótesis más empleados para analizar la normalidad. En ambos, se considera como hipótesis nula que los datos proceden de una distribución normal. El p-value de estos test indica la probabilidad de obtener unos datos como los observados si realmente procediesen de una población con una distribución normal con la misma media y desviación que estos. Por lo tanto, si el p-value es menor que un determinado valor (típicamente 0.05), entonces se considera que hay evidencias suficientes para rechazar la normalidad. El test de Shapiro-Wilk se desaconseja cuando se dispone de muchos datos (**más de 50**) por su elevada sensibilidad a pequeñas desviaciones de la normal, por lo tanto este test fue descartado para este trabajo ya que la muestra a considerar para esta parte del trabajo es de 551 personas, que sobrepasa en gran medida a los 50 datos para aplicar este test.
En este trabajo **utilizaremos el test D'Agostino's K-squared** para establecer si cada una de las variables poseen una distribución normal.

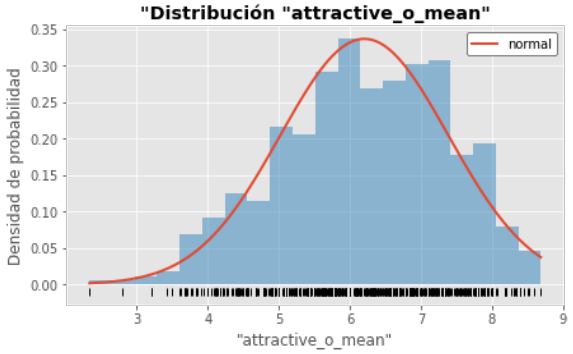
A continuación, se mostrarán los resultados de los análisis realizados con las 10 variables estudiadas. Se mostrará cada variable por separado y un gráfico superponiendo las 2 variables con sus respectivas curvas de densidad de probabilidad.

6.2.1.1 “attractive” vs “attractive_o_mean”

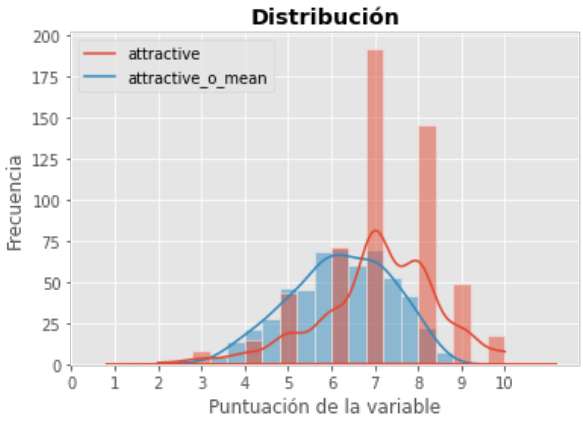
En los siguientes gráficos se observan los resultados obtenidos a través de las 3 metodologías usadas. Se llega a la conclusión que ninguna de las 2 distribuciones de las muestras es normal, por lo tanto, tendremos que optar por usar una prueba de hipótesis no paramétrica para poder obtener resultados consistentes.



attractive
Prueba de normalidad D'Agostino K-squared test
Estadístico de prueba: 39.5203452522947
Valor p: 2.6197875083456545e-09
La muestra NO sigue una distribución normal
Kursotis: 0.8921444131549139
Skewness: -0.6031960442456498

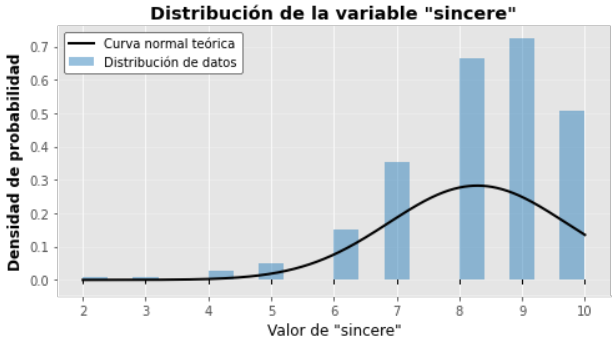


attractive_o_mean
Prueba de normalidad D'Agostino K-squared test
Estadístico de prueba: 13.916376552716242
Valor p: 0.0009508176338145235
La muestra NO sigue una distribución normal
Kurtosis: -0.42300585744694885
Skewness: -0.29442967567876854

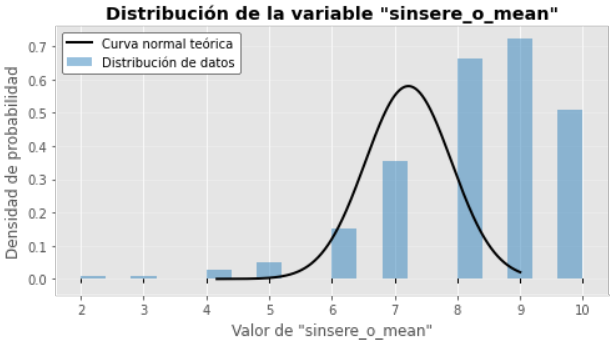


6.2.1.2 “sincere” vs “sinsere_o_mean”

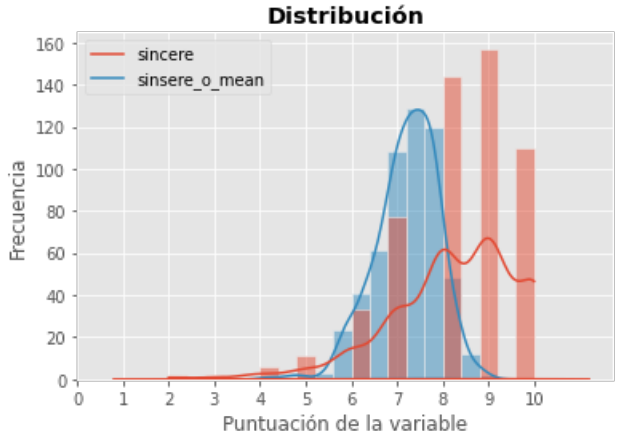
En los siguientes gráficos se observan los resultados obtenidos a través de las 3 metodologías usadas. Se llega a la conclusión que ninguna de las 2 distribuciones de las muestras es normal, por lo tanto, tendremos que optar por usar una prueba de hipótesis no paramétrica para poder obtener resultados consistentes.



“sincere”
Prueba de normalidad D'Agostino K-squared test
Estadístico de prueba: 95.8164374074147
Valor p: 1.5621580788841594e-21
La muestra NO sigue una distribución normal
Kursotis: 1.7020811866517311
Skewness: -1.0547972977960434

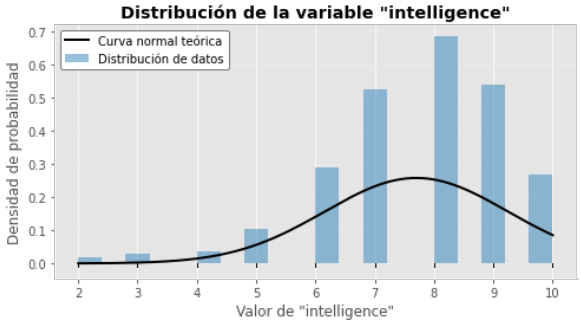


“sinsere_o_mean”
Prueba de normalidad D'Agostino K-squared test
Estadístico de prueba: 39.08541510960933
Valor p: 3.2561915691440698e-09
La muestra NO sigue una distribución normal
Kursotis: 0.8392427035029044
Skewness: -0.6009814989317097



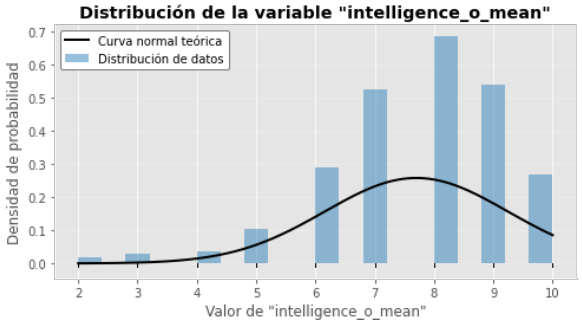
6.2.1.3 “intelligence” vs “intelligence_o_mean”

En los siguientes gráficos se observan los resultados obtenidos a través de las 3 metodologías usadas. Se llega a la conclusión que una de las muestras tiene distribución normal (“intelligence_o_mean”), pero la otra muestra (“intelligence”) no tiene distribución normal, por lo tanto, tendremos que optar por usar una prueba de hipótesis no paramétrica para poder obtener resultados consistentes.



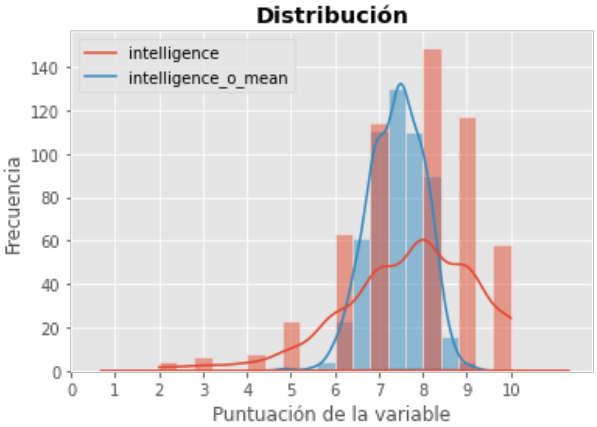
“intelligence”

Prueba de normalidad D'Agostino K-squared test
Estadístico de prueba: 59.3032771380053
Valor p: 1.3257358183754285e-13
La muestra NO sigue una distribución normal
Kursotis: 1.0010719824773506
Skewness: -0.8026717794126981



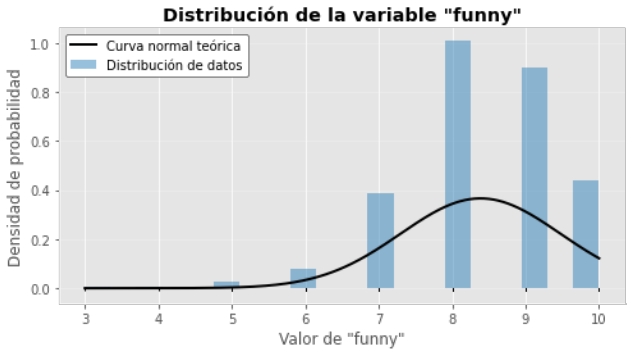
“intelligence_o_mean”

Prueba de normalidad D'Agostino K-squared test
Estadístico de prueba: 5.924112468310065
Valor p: 0.05171247481806512
La muestra SI sigue una distribución normal
Kursotis: -0.022904484294232308
Skewness: -0.2544481974793868

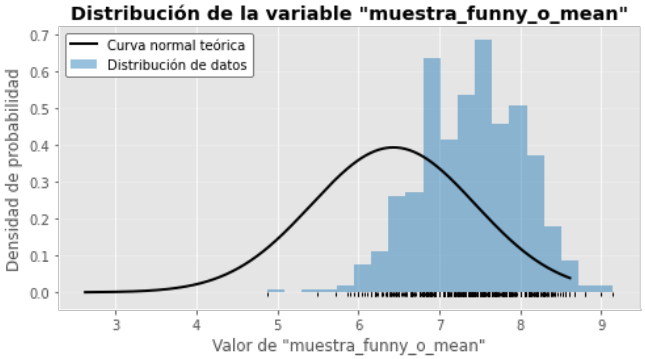


6.2.1.4 “funny” vs “funny_o_mean”

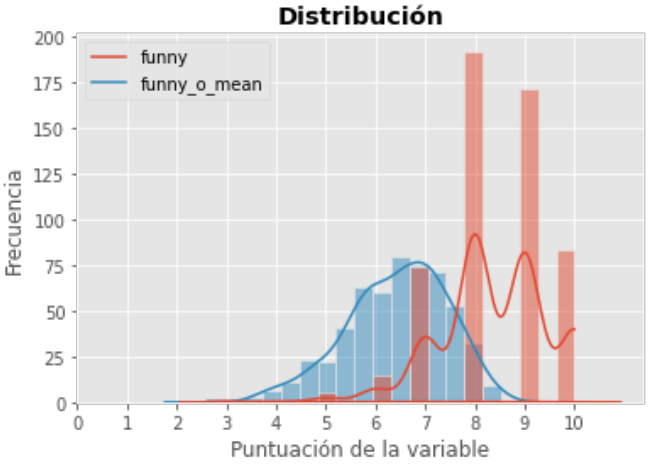
En los siguientes gráficos se observan los resultados obtenidos a través de las 3 metodologías usadas. Se llega a la conclusión que ninguna de las 2 distribuciones de las muestras es normal, por lo tanto, tendremos que optar por usar una prueba de hipótesis no paramétrica para poder obtener resultados consistentes.



“funny”
Prueba de normalidad D'Agostino K-squared test
Estadístico de prueba: 49.39531988038781
Valor p: 1.8790683260213524e-11
La muestra NO sigue una distribución normal
Kursotis: 1.227820476344231
Skewness: -0.6537613838774399

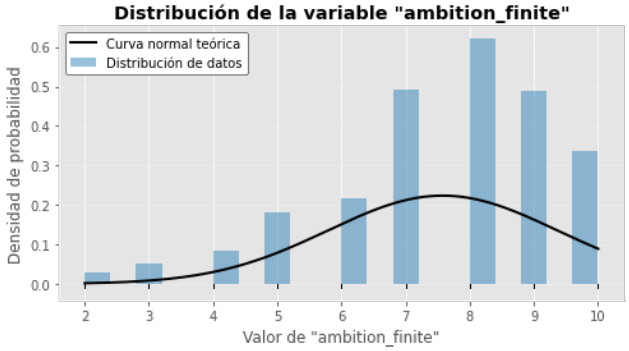


“funny_o_mean”
Prueba de normalidad D'Agostino K-squared test
Estadístico de prueba: 18.855946997355577
Valor p: 8.044205079423263e-05
La muestra NO sigue una distribución normal
Kursotis: 0.002497440130440065
Skewness: -0.4686613741142247

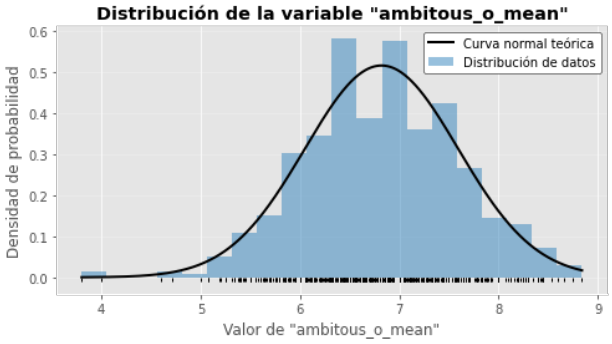


6.2.1.5 “ambition” vs “ambitous_o_mean”

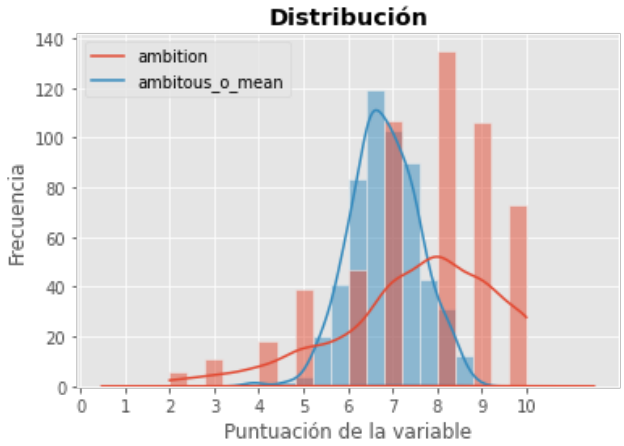
En los siguientes gráficos se observan los resultados obtenidos a través de las 3 metodologías usadas. Se llega a la conclusión que una de las muestras tiene distribución normal (“ambitous_o_mean”), pero la otra muestra (“ambition”) no tiene distribución normal, por lo tanto, tendremos que optar por usar una prueba de hipótesis no paramétrica para poder obtener resultados consistentes.



“ambition”
Prueba de normalidad D'Agostino K-squared test
Estadístico de prueba: 48.990626929439706
Valor p: 2.300490953534489e-11
La muestra NO sigue una distribución normal
Kursotis: 0.37594449277317166
Skewness: -0.7920375447443638



“ambitous_o_mean”
Prueba de normalidad D'Agostino y Pearson
Estadístico de prueba: 2.5263809462984654
Valor p: 0.28275047808447396
La muestra SI sigue una distribución normal
Kursotis: 0.22720404243160353
Skewness: -0.11307014682890301



Si una distribución no es normal tiene consecuencias al momento de analizar las variables, el hecho de no poder asumir la normalidad influye principalmente en los test de hipótesis paramétricas (ej. t-test), por lo tanto, hay que recurrir a los **test no paramétricos**.

Las pruebas de hipótesis no paramétricas son métodos estadísticos que no requieren supuestos explícitos sobre la distribución de los datos subyacentes. Estas

pruebas son útiles cuando los datos no cumplen con los supuestos de normalidad o de homogeneidad de varianza que se requieren para las pruebas paramétricas. Las alternativas que tenemos son la "**Prueba de Wilcoxon**" y la "**Prueba de Mann-Whitney U**".

La "**Prueba de Wilcoxon**" se utiliza para comparar las medias de dos muestras relacionadas, al considerar que el par de variables **no son relacionadas** ya que la evaluación de la pareja nunca es conocida por el otro sujeto, en este trabajo desestimamos el uso de esta prueba de hipótesis.

Hemos elegido la prueba de hipótesis de **Mann-Whitney U** para comparar las variables analizadas. Según el libro "Estadística para las ciencias sociales, del comportamiento y de la Salud, 3a. edición, Haroldo Elorza Pérez-Tejada" es posible emplear esta prueba como una alternativa de la paramétrica t de Student, para comprobar la diferencia entre dos medias en dos muestras independientes. Las puntuaciones que representan mediciones, observaciones o datos en general, deben ser mutuamente independientes dentro de la misma muestra a la que pertenecen respecto de otra; no es necesario que sean del mismo tamaño. La hipótesis nula establece la analogía (homogeneidad) de distribuciones poblacionales y, en cierta manera, la igualdad de las dos medias o medianas. No obstante, si el par de distribuciones poblacionales son más o menos similares, tanto en forma como en variabilidad, la U es una prueba excelente de la tendencia central. Pero debido a que la prueba se basa en categorías o clases, la medida de tendencia central más adecuada es la mediana (Me). Entonces, cuando se consideran los resultados de dicha prueba, lo primero en que se piensa es la comparación de ambas medianas.

Para comparar la media de dos distribuciones que **NO** son normales las opciones disponibles en Python es utilizar la prueba de **Mann-Whitney U**, que es una prueba **NO paramétrica** para comparar dos muestras independientes. La ventaja de esta prueba es que no requiere que las muestras posean una distribución normal.

Sin embargo, aún existen algunos supuestos y consideraciones a tener en cuenta al aplicar la prueba de Mann-Whitney U:

- **Independencia**: Las dos muestras deben ser independientes entre sí.
- **Escala de medición**: Las variables que se están comparando deben ser, al menos, ordinales.
- **Igualdad de formas**: Las dos muestras deben tener la misma forma y variabilidad en su distribución. En caso contrario, la prueba de Mann-Whitney U podría no ser apropiada.
- **Homogeneidad de varianzas**: Las varianzas de las dos muestras deben ser similares. Aunque la prueba de Mann-Whitney U no requiere que las muestras tengan la misma varianza, si la varianza de una muestra es mucho mayor que la otra, esto puede afectar la capacidad de la prueba para detectar diferencias entre las medias.

En el contexto de la estadística, la independencia de los datos se refiere a la ausencia de una relación entre las observaciones en una muestra y las observaciones en otra muestra.
En otras palabras, si los datos son independientes, el valor de una observación en una muestra no depende del valor de otra observación en otra muestra, que según nuestra opinión este es nuestro caso.

Debido a que **ningún par de variables** estudiadas poseen ambas una distribución normal, todos los test de hipótesis que se harán serán de Mann-Whitney U.

Se utilizará "**SciPy**" una biblioteca de Python que permite realizar una prueba de hipótesis con Mann-Whitney U.
(https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html)

En mannwhitneyu devuelve dos valores:

1. El **estadístico de prueba** y el **valor p**. Si el valor p es menor que un nivel de significancia predefinido (como 0.05), entonces SI se puede rechazar la hipótesis nula y concluir que hay una diferencia significativa entre las medianas de las dos muestras.
2. Si el valor p es mayor que el nivel de significancia, entonces NO se puede rechazar la hipótesis nula y NO se puede concluir que hay una diferencia significativa entre las medianas de las dos muestras.

A continuación, se muestra los resultados obtenidos:

Variables: "attractive" vs "attractive_o_mean"

Hipótesis nula (H_0): No hay diferencia significativa entre la mediana de la autoevaluación del atractivo de las personas y la mediana de la calificación media del atractivo que cada sujeto calificó a su pareja.

Hipótesis alternativa (H_1): Existe una diferencia significativa entre la mediana de la autoevaluación del atractivo de las personas y la mediana de la calificación media del atractivo que cada sujeto calificó a su pareja. En términos más concretos, esto significa que estamos tratando de determinar si la mediana de la autoevaluación del atractivo de las personas es estadísticamente diferente de la mediana de la media de calificación del atractivo de la evaluación que cada sujeto calificó a su pareja.

Resultados:

- Estadístico de prueba: 233745.5
- Valor p: 2.5156183882980993e-59

Conclusión:
El valor p obtenido es de 2.5156183882980993e-59, lo que significa que la probabilidad de obtener una diferencia tan grande o mayor entre las medianas de las dos muestras si la hipótesis nula es verdadera es extremadamente baja. Por lo tanto, se puede rechazar la hipótesis nula y se acepta la hipótesis alternativa, **concluyendo que hay una diferencia significativa entre las medianas de las dos muestras.**

En resumen, los resultados sugieren que existe una diferencia significativa entre las medianas de la autoevaluación del atractivo de las personas y la mediana de media de la calificación del atractivo de la evaluación de la otra muestra.

Podemos concluir que las personas se autocalifican distinto en lo referente al atractivo que lo que lo califica su pareja, se corrobora la subjetividad de las percepciones de las personas.

Variables: “sincere” vs “sinsere_o_mean”

Hipótesis nula (H_0): No hay diferencia significativa entre la mediana de la autoevaluación respecto de la sinceridad de las personas y la mediana de la calificación media respecto de la sinceridad que cada sujeto calificó a su pareja.

Hipótesis alternativa (H_1): Existe una diferencia significativa entre la mediana de la autoevaluación respecto de la sinceridad de las personas y la mediana de la calificación media respecto de la sinceridad que cada sujeto calificó a su pareja. En términos más concretos, esto significa que estamos tratando de determinar si la mediana de la autoevaluación respecto de la sinceridad de las personas es estadísticamente diferente de la mediana de la media de calificación del atractivo de la evaluación que cada sujeto calificó a su pareja.

Resultados:

- Estadístico de prueba: 209490.0
- Valor p: 4.671512602369158e-31

Conclusión:

El valor p obtenido es de 4.671512602369158e-31, lo que significa que la probabilidad de obtener una diferencia tan grande o mayor entre las medianas de las dos muestras si la hipótesis nula es verdadera es extremadamente baja. Por lo tanto, se puede rechazar la hipótesis nula y se acepta la hipótesis alternativa, **concluyendo que hay una diferencia significativa entre las medianas de las dos muestras.**

En resumen, los resultados sugieren que existe una diferencia significativa entre las medianas de la autoevaluación respecto de la sinceridad de las personas y la mediana de media de la calificación de la media respecto de la sinceridad del atractivo de la evaluación de la otra muestra.

Podemos concluir que las personas se autocalifican distinto en lo referente a cuán sinceros son respecto a lo que lo califica su pareja, se corrobora la subjetividad de las percepciones de las personas.

Variables: “intelligence” vs “intelligence_o_mean”

Hipótesis nula (H_0): No hay diferencia significativa entre la mediana de la autoevaluación respecto de la inteligencia de las personas y la mediana de la calificación media respecto de la inteligencia que cada sujeto calificó a su pareja.

Hipótesis alternativa (H_1): Existe una diferencia significativa entre la mediana de la autoevaluación respecto de la inteligencia de las personas y la mediana de la calificación media respecto de la inteligencia que cada sujeto calificó a su pareja. En términos más concretos, esto significa que estamos tratando de determinar si la mediana de la autoevaluación respecto de la inteligencia de las personas es estadísticamente diferente de la mediana de la media de calificación de la inteligencia de la evaluación que cada sujeto calificó a su pareja.

Resultados:

- Estadístico de prueba: 181346.5
- Valor p: 7.240071682183129e-10

Conclusión:

El valor p obtenido es de 7.240071682183129e-10, lo que significa que la probabilidad de obtener una diferencia tan grande o mayor entre las medianas de las dos muestras si la hipótesis nula es verdadera es extremadamente baja. Por lo tanto, se puede rechazar la hipótesis nula y se acepta la hipótesis alternativa, **concluyendo que hay una diferencia significativa entre las medianas de las dos muestras.**

En resumen, los resultados sugieren que existe una diferencia significativa entre las medianas de la autoevaluación respecto de la inteligencia de las personas y la mediana de medias de la calificación respecto de la inteligencia de la evaluación de la otra muestra.

Podemos concluir que las personas se autocalifican distinto en lo referente a cuán inteligentes son respecto a lo que lo califica su pareja, se corrobora la subjetividad de las percepciones de las personas.

Variables: "funny" vs "funny_o_mean"

Hipótesis nula (H_0): No hay diferencia significativa entre la mediana de la autoevaluación respecto de cuán divertido es una persona y la mediana de la calificación promedio respecto de cuán divertido es cada sujeto que calificó a su pareja.

Hipótesis alternativa (H_1): Existe una diferencia significativa entre la mediana de la autoevaluación respecto de cuán divertido es una persona y la mediana de la calificación promedio respecto de cuán divertido es cada sujeto que calificó a su pareja. En términos más concretos, esto significa que estamos tratando de determinar si la mediana de la autoevaluación respecto de cuán divertido es una persona es estadísticamente diferente de la mediana del promedio de calificación de cuán divertido es cada sujeto que calificó a su pareja.

Resultados:

- Estadístico de prueba: 272047.5
- Valor p: 1.2133681837447999e-123

Conclusión:

El valor obtenido es de 1.2133681837447999e-123, lo que significa que la probabilidad de obtener una diferencia tan grande o mayor entre las medianas de las dos muestras si la hipótesis nula es verdadera es extremadamente baja. Por lo tanto, se puede rechazar la hipótesis nula y se acepta la hipótesis alternativa, **concluyendo que hay una diferencia significativa entre las medianas de las dos muestras.**

En resumen, los resultados sugieren que existe una diferencia significativa entre las medianas de la autoevaluación respecto cuán divertido es una persona y la mediana del promedio de la calificación respecto cuán divertido evaluó la otra persona a su pareja.

Podemos concluir que las personas se autocalifican distinto en lo referente a cuán divertidos son respecto a lo que lo califica su pareja, se corrobora la subjetividad de las percepciones de las personas.

Variables: “ambition” vs “ambitous_o_mean”

Hipótesis nula (H_0): No hay diferencia significativa entre la mediana de la autoevaluación respecto de cuán ambicioso es una persona y la mediana de la calificación promedio respecto de cuán ambicioso es cada sujeto que calificó a su pareja.

Hipótesis alternativa (H_1): Existe una diferencia significativa entre la mediana de la autoevaluación respecto de cuán ambicioso es una persona y la mediana de la calificación promedio respecto de cuán ambicioso es cada sujeto que calificó a su pareja. En términos más concretos, esto significa que estamos tratando de determinar si la mediana de la autoevaluación respecto de cuán ambicioso es una persona es estadísticamente diferente de la mediana del promedio de calificación de cuán ambicioso es cada sujeto que calificó a su pareja.

Resultados:

- Estadístico de prueba: 206824.0
- Valor p: 2.163013469554321e-28

Conclusión:
El valor obtenido es de 2.163013469554321e-28, lo que significa que la probabilidad de obtener una diferencia tan grande o mayor entre las medianas de las dos muestras si la hipótesis nula es verdadera es extremadamente baja. Por lo tanto, se puede rechazar la hipótesis nula y se acepta la hipótesis alternativa, **concluyendo que hay una diferencia significativa entre las medianas de las dos muestras.**

En resumen, los resultados sugieren que existe una diferencia significativa entre las medianas de la autoevaluación respecto cuán ambicioso es una persona y la mediana del promedio de la calificación respecto cuán ambicioso evaluó la otra persona a su pareja.

Podemos concluir que las personas se autocalifican distinto en lo referente a cuán ambicioso son respecto a lo que lo califica su pareja, se corrobora la subjetividad de las percepciones de las personas.

6.2.2 ¿Cuántas personas volverán a tener una nueva cita después de haberse conocido por 4 minutos en una primera cita?

Calcularemos con el 95 % de confianza cuántas personas volverán a tener una nueva cita después de haberse conocido por 4 minutos en una primera cita.

- Total de personas: **551**
- Cantidad de personas que no hicieron match: **468** (para hacer match ambas personas se gustan mutuamente y quieren volver a juntarse, con que haya una persona que no guste de otra se considera no-match)
- Cantidad de personas que hicieron match: **83**
- La proporción de personas que hacen match es de: 0.15

- El número de personas involucradas en el experimento es de: 551
- Intervalo de confianza al 95%: (14.88%, 15.06%)

Tomando una muestra de 551 personas se puede llegar a la conclusión con el **95%** de confianza de que sólo se volverán a juntar en otra cita entre un **14,88%** y un **15,06%** de las parejas que se conocen por primera vez. Como conclusión podemos decir que si 2 personas tiene interés en tener una cita con otra persona, tiene que reunirse durante 4 minutos con por lo menos 8 personas, para tener altas probabilidades de tener una nueva cita, considerando las restricciones del experimento, como son la distribución de las proporciones de las distintas razas, edades, nivel educativo, etc.

7 Modelo y validación

Tradicionalmente, se aplican técnicas econométricas para estudiar las relaciones entre variables socioeconómicas. La minería de datos puede resultar muy valiosa en este sentido, pero al utilizar este tipo de herramientas se suele asumir un costo sustancial en términos de interpretabilidad. En esta sección, se aplican métodos que combinan modelos de aprendizaje automático con instrumentos que contribuyen a mejorar la explicación de estos modelos. En particular, se desarrollan procedimientos de *Feature Importance*, dependencia parcial, y generación de variables a través de reglas. Luego se contrastan los hallazgos con lo explicitado en Fisman et al. (2006). Aunque en líneas generales los resultados coinciden, se descubren nuevas relaciones entre variables con potencial de predicción.

7.1 Detalle de la aplicación de las técnicas, estrategias de verificación y validación, junto especificación del modelo obtenido

Pre-procesamiento:

En la etapa de procesamiento, se aplican algunas transformaciones que podrían introducir sesgo en el análisis:

- Se convierten las variables con escala de importancia compartida usando un rango que va del 1 (más importante) al 6 (menos importante). A las observaciones que comparten rango, se les asigna el promedio entre los rangos que hubieran ocupado si estos fueran distintos.
- Se excluyen ciertas columnas que no contienen información provista por los sujetos, sino que describen el *dataset* o la estructura del experimento. Por ejemplo, la columna “wave” (ronda a la que pertenece el encuentro). También se eliminan columnas que, aunque relevantes para la predicción, contienen información redundante y podrían llevar a los modelos a depender excesivamente de ellas. En particular, la columna “like”, que predice fuertemente la decisión del sujeto y prácticamente transmite lo mismo que la variable dependiente.

- Las variables categóricas son sometidas a una codificación en caliente (*one-hot encoding* en inglés), resultando en un número mayor de columnas, pero con predictoras dicotómicas.
- Se rellenan los valores faltantes a través de una técnica de imputación iterativa[7]. La misma se lleva a cabo en los datos de entrenamiento y validación por separado, para evitar filtrar información hacia el último.
- Los datos son estandarizados para mayor comparabilidad entre los coeficientes de los modelos. Sin embargo, las columnas con variables dicotómicas no se modifican.
- Se filtran los valores de las columnas con datos categóricos: aquellas con un soporte (representación) menor al 5%[8] , son agrupadas en la categoría "otros".
- En la implementación de los heurísticos, se trabaja con 2 poblaciones separadas: **masculina** y **femenina**. Se toma esta decisión para obtener coeficientes y métricas interpretables directamente desde el punto de vista del género de los sujetos. Por otro lado, en Fisman et al. (2006) se ajustan regresiones a todo el conjunto y se evalúan las diferencias entre los dos grupos a través de una variable dicotómica que describe el género del participante. Debido a que los modelos aplicados en esta investigación presentan una complejidad algo mayor, y que las técnicas de evaluación de las predictoras son distintas, esta solución más simple no es viable en este caso.

Estos procedimientos resultan en **66 variables predictoras** y una variable de respuesta (la decisión del participante respecto a una segunda cita con su compañero).

Entrenamiento y Validación:

Luego del procesamiento, se aparta un 10% de los datos (en forma estratificada de acuerdo a la variable objetivo). La totalidad de entrenamientos, ajustes y selecciones de modelos se realiza sobre el conjunto de entrenamiento (el 90% restante) aplicando técnicas de validación cruzada. El proceso se repite para ambas poblaciones.

De acuerdo a los objetivos planteados, y poniendo énfasis en la interpretabilidad y el análisis de las relaciones entre las variables, se ajustan dos modelos:

- Potenciación del Gradiente *(Gradient Boosting* en inglés). El Gradient Boosting es un heurístico que construye sucesivamente árboles de clasificación en base a los residuos de árboles anteriores para generar predicciones sobre los datos.
- Regresión logística con penalización *L1* (también llamada *Lasso* en inglés). Esta clase de modelos agrega un término de penalización al problema de minimización de cuadrados en el ajuste, reduciendo así el valor de los coeficientes resultantes.

7 "Scikit-learn Iterative Imputer", Disponible: https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

8 Kuhn, M. and Johnson K. (2020). *Feature Engineering and Selection, a Practical Approach for Predictive Models.* Chapman and Hall, Boca Raton, Florida.

Con las variables generadas por el primer modelo, se ajusta el segundo para seleccionar las más importantes. El proceso completo con modelos que generan conjuntos de árboles, y la posterior extracción y selección de reglas, recibe el nombre de *RuleFit* [9].

Para la elección de los parámetros del modelo Gradient Boosting, se lleva a cabo una búsqueda exhaustiva (*Grid Search* en inglés) entre distintas combinaciones. La métrica a optimizar en la búsqueda es el coeficiente Kappa de Cohen[10]. Los parámetros incluídos son los siguientes:

- La profundidad de cada árbol.
- El número de árboles a construir.
- La tasa de aprendizaje del modelo (controla la velocidad a la que se incorpora información de nuevos árboles).
- El número máximo de variables a considerar para cada árbol (si es menor al total de columnas, introduce un elemento aleatorio adicional que puede incrementar la robustez del modelo).
- El número mínimo de observaciones requeridas para dividir las ramas de cada árbol.
- El número mínimo de observaciones requeridas en los nodos terminales.

Selección de los Modelos:

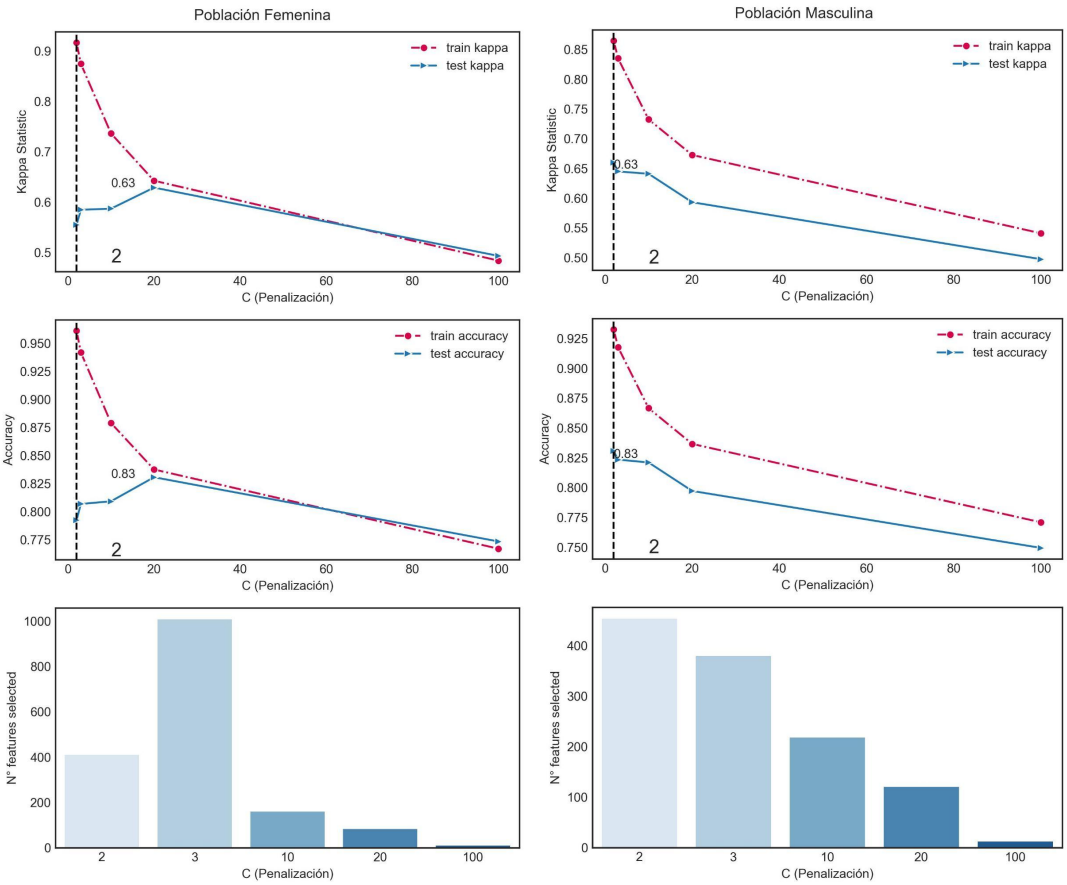
De los árboles formados por el modelo, se extraen al azar **1.000 "reglas"** que luego actúan como nuevas predictoras para la regresión logística. Cada rama en un árbol, desde el nodo raíz hasta los nodos terminales, está compuesta por condiciones sobre los atributos originales. Luego, partiendo del nodo raíz, es posible convertir porciones de estas ramas en nuevas variables. En la presente investigación, se incluyen reglas con interacciones entre hasta 3 variables.

Una vez extraídas las reglas, se agregan a las variables originales para el ajuste del modelo *Lasso.* Al aplicar una penalización a la sumatoria de los valores absolutos de los coeficientes, este produce una selección de coeficientes[11]. Como resultado, se obtiene un modelo que involucra sólo las predictoras más fuertes. Aunque existen variantes de este modelo que reducen el valor de todos los coeficientes y no provocan una exclusión de variables.

Para el entrenamiento del heurístico, se eligen distintos valores para el parámetro de penalización. Por último, se selecciona el modelo con el mayor estadístico Kappa en el conjunto de validación. En el apéndice B, sección 1, se encuentran los coeficientes más altos asociados a estos modelos.

La siguiente figura contiene las curvas con los estadísticos Kappa obtenidos para los distintos valores de la penalización. Incluye también la exactitud (tasa de aciertos) de cada ajuste y el número de variables seleccionadas. Las 3 figuras de la izquierda corresponden a la población femenina de la muestra, y las restantes a la población masculina:

9 Friedman, J. H. and Popescu, B. E. (2005). *Predictive Learning via Rule Ensembles*.
10 Mary L. McHugh (2012). *Interrater reliability: the kappa statistic*. Biochemia Media.
11 James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (1st ed.) Springer.



Finalmente, los ajustes elegidos favorecen 82 predictoras (incluyendo interacciones) para la muestra femenina y 379 para la población masculina. De las variables seleccionadas, sólo 16 y 30 respectivamente corresponden al *dataset* original. A continuación, se muestran las medidas de desempeño obtenidas por cada modelo:

Modelo	Estadístico Kappa	Exactitud
Lasso (F)	0.63	0.83
Gradient Boosting (F)	0.65	0.84
Lasso (M)	0.64	0.82
Gradient Boosting (M)	0.69	0.84

Las diferencias son marginales, pero se consideran ambos heurísticos para el análisis posterior. También se ajustan los modelos excluyendo las observaciones etiquetadas como *outlier,* pero las diferencias son insignificantes y se opta por el uso de todos los datos.

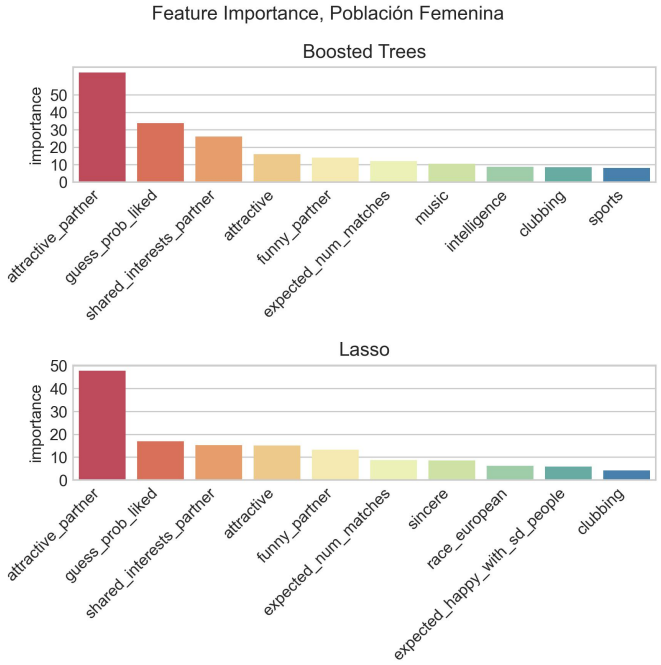
Análisis de Importancia de las variables:

Uno de los objetivos principales de esta investigación es analizar cómo influyen las características de los sujetos y sus compañeros en la decisión de tener una segunda cita. Con este fin, se evalúan los atributos más relevantes para cada modelo, y cómo afectan la predicción de la probabilidad de que el participante desee una segunda cita.
Para lo primero, se computa la importancia de permutación de cada variable, y para lo segundo se obtienen las dependencias parciales respecto de cada variable.

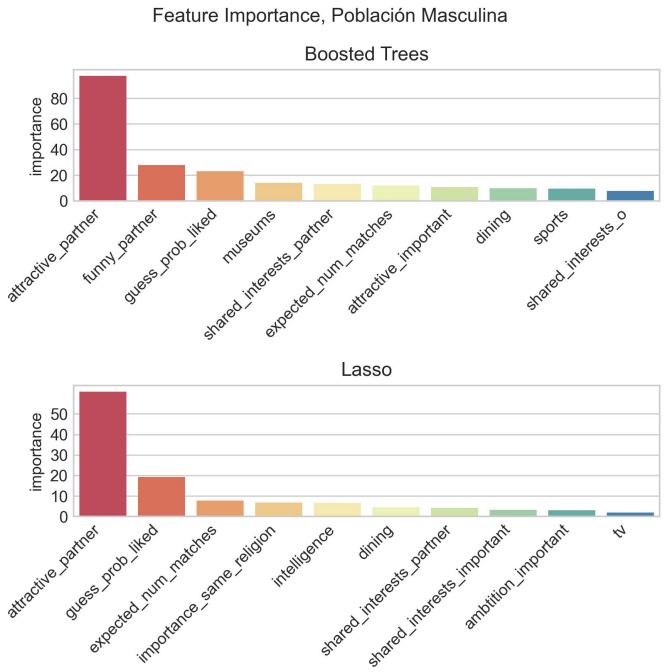
La importancia de una variable (*Feature Importance* en inglés) está relacionada con la vinculación del heurístico respecto de esta a la hora de hacer predicciones. Aunque se han propuesto distintos métodos para explorar este aspecto de las variables independientes, en este caso se elige la técnica de permutación. Su ventaja más evidente es que puede ser aplicada a cualquier tipo de heurístico (estas herramientas se conocen como “agnósticas al modelo”). Por otra parte, considera todas las interacciones con la variable en cuestión en forma automática y, por lo tanto, es aplicable incluso cuando las interacciones son muchas o muy complejas.

El procedimiento consiste en intercambiar los valores de columna que contiene a la variable (y las columnas asociadas), y calcular el cambio porcentual en la tasa de error respecto a las predicciones originales. Esto se repite varias veces y luego se obtiene un promedio de estas medidas. Las variables cuyas permutaciones causen mayores incrementos en la tasa de error pueden considerarse como más importantes para un modelo en particular.

Para fines de consistencia con el resto del análisis, se define a la tasa de error como 1 - Kappa. En las figuras posteriores, se exponen las variables más importantes para cada heurístico. Se debe tener en cuenta que en los modelos de regresión logística sólo se incluyen las variables seleccionadas por el algoritmo. Las primeras corresponden a la población femenina:



Y las siguientes corresponden a la población masculina:



Aunque evidentemente el **atractivo** del compañero domina a las demás variables en términos de importancia, se pueden elaborar algunas conclusiones sobre variables que no son exploradas en Fisman et al. (2006). En secciones posteriores se investigan algunas de estas conclusiones.

Por otro lado, y a diferencia del análisis de coeficientes, más usual en la econometría, el *Feature Importance* permite medir el impacto de las variables en forma sencilla, incluso cuando las relaciones subyacentes sean muy complejas.

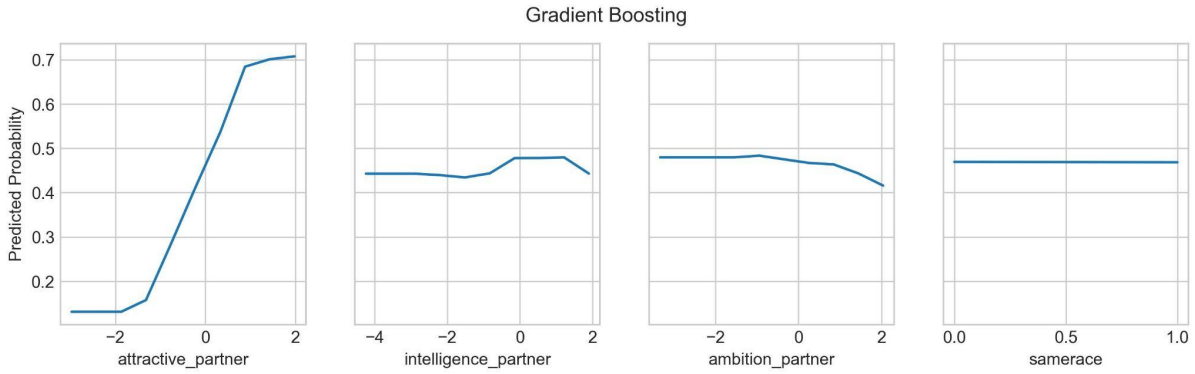
Una vez identificadas estas relaciones importantes, se pueden investigar sus características mediante el análisis de dependencia parcial[12]. En pocas palabras, este consiste en reemplazar artificialmente una columna por un valor determinado a lo largo de todo el conjunto de datos y luego, usando un heurístico, calcular el promedio de las probabilidades de que la variable respuesta tome un valor. Al repetir este proceso para cada valor distinto de la columna en cuestión, se obtiene una curva que asocia a la variable independiente con las probabilidades estimadas para la variable dependiente.

Los gráficos de dependencia parcial no están libres de inconvenientes. Quizá el mayor de ellos sea que se basan en observaciones artificiales, cuyos valores probablemente no sean posibles naturalmente. Luego, es más efectivo cuando las variables son independientes entre ellas. En este caso, la mayoría de correlaciones resultan bajas y se asume que la aplicación del método es válida.

En referencia a la presente investigación, se examinan algunas de las características referenciadas en Fisman et al. (2006). En el apéndice B, sección 2, se presentan las figuras con las 5 características por modelo cuya dependencia parcial presenta mayor variabilidad. En todos los casos se trabaja con los datos de validación.

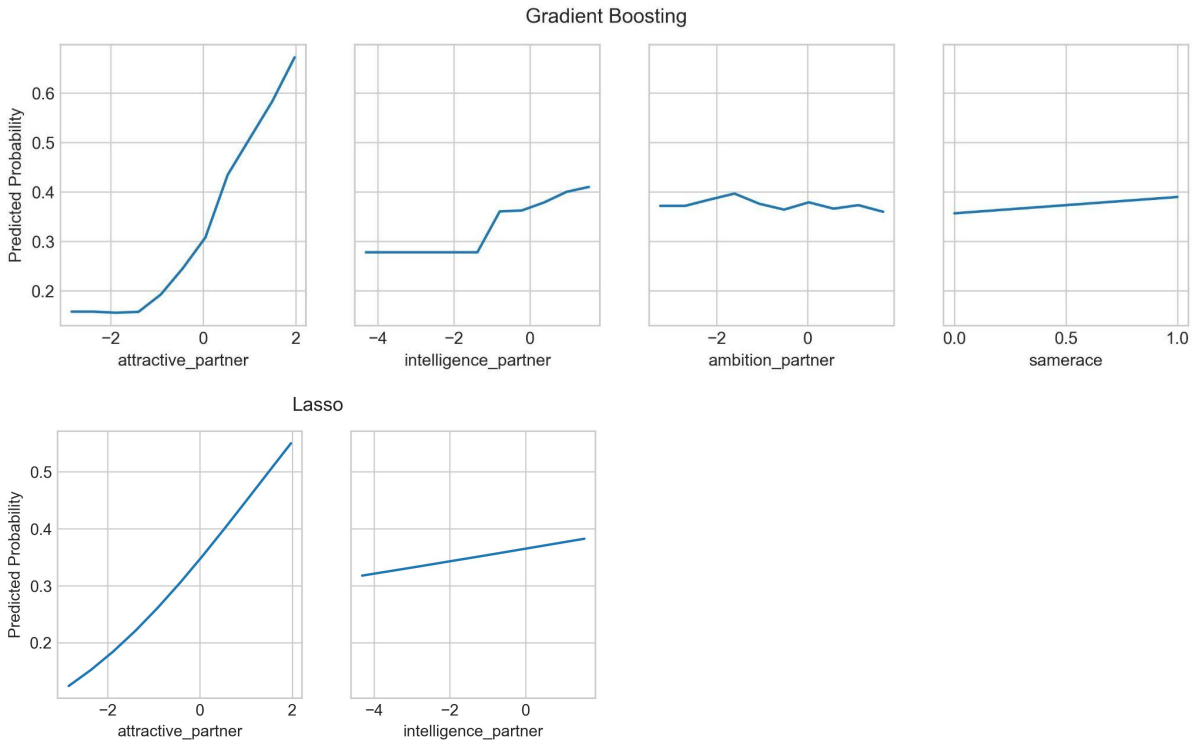
Población Masculina:

12 Molnar, C. (2019). *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable.* Leanpub.



Se omiten las figuras del modelo *Lasso* porque para la población masculina, ninguna de las características analizadas en el artículo resulta seleccionada.

Población Femenina:



Las figuras evidencian el contraste entre la naturaleza lineal del modelo *Lasso* y las relaciones más complejas que manifiesta el modelo *Gradient Boosting.*

<u>Resumen de los hallazgos</u>:

Al analizar las relaciones entre las predictoras y los modelos, surgen algunas conclusiones notables:

- Por un margen significativo, el **atractivo** del compañero es la variable más importante para ambos modelos y en ambas poblaciones. Por otro lado, y a diferencia de lo explicitado en el artículo, la inteligencia del compañero no es una predictora excesivamente fuerte de las decisiones de las mujeres si se la compara con el atractivo. Aunque vale la pena mencionar que el atributo es seleccionado por el modelo *Lasso*. Sin embargo, el cambio en la probabilidad estimada al pasar del valor mínimo al máximo en inteligencia es inferior a 10 puntos, mientras que para el atractivo supera los 60 puntos.

- Con excepción de la predictora “ambición del compañero”, las pendientes de las curvas de dependencia parcial son compatibles con los signos de los coeficientes detallados en el Fisman et al. (2006). En particular, se llega a las mismas conclusiones con respecto a las preferencias por la misma **raza**: es una predictora significativa en el caso de las mujeres, mientras que para los hombres no afecta drásticamente la probabilidad (estimada) de querer una segunda cita.

- Algunas variables tradicionalmente no son consideradas por las teorías sobre la selección de parejas, y por lo tanto no son revisadas en Fisman et al.(2006). Pero inspeccionando su importancia y dependencia parcial, aparecen múltiples posibilidades de estudiar su influencia en las decisiones desde otras perspectivas.

Finalmente, se replican algunas de las regresiones presentadas en el artículo para evaluar su desempeño y fiabilidad. En la tabla a continuación se muestran los resultados obtenidos en el conjunto de validación:

Métrica	Modelo 1	Modelo 2
Estadístico Kappa	0.16	0.44
Exactitud	0.62	0.73

El primer **modelo** usa como predictoras a las columnas sobre la valoración en inteligencia, atractivo, etc. del compañero. El segundo incorpora relaciones multiplicativas y el signo de la diferencia entre los atributos auto-calificados y la valoración respecto del compañero. Se omiten otros modelos debido a la falta de datos sobre la región y el ingreso de los participantes.

Los modelos como aquellos expuestos en Fisman et al. (2006) destacan por su simpleza e interpretabilidad. Pero en muchos casos fallan en la tarea de descubrir relaciones más complejas entre las variables. Los métodos y enfoques aquí presentados permiten al menos advertir la existencia de eventuales interacciones entre estas, y facilitan la posterior construcción de otros modelos, sin comprometer demasiado la interpretabilidad de los mismos.

8 Conclusiones

En resumen, analizamos los datos de de un experimento de citas rápidas realizado en la 2.002 y 2.004 en la Universidad de Columbia, New York, Estados Unidos en el cual participaron **552 personas**, entre hombres y mujeres, para revelar las diferencias de preferencia de elección entre hombres y mujeres; y los factores importantes que afectan la elección potencial de pareja para reunirse en una nueva cita. Hay que tener en cuenta que cada cita dura 4 minutos, por lo tanto, es complejo hacer una evaluación exhaustiva de características como cuán ambicioso/a, gracioso/a, sincero/a, inteligente o divertido/a es una persona.

Carecemos de la foto y los datos del tipo físico como puede ser, la altura, el peso, color de ojos, color de pelo, etc. que no nos permite hacer más inferencias. Estudiamos la preferencia de atributos de las personas sin considerar el impacto potencial de otros atributos. Existen diferencias significativas entre la cultura occidental, específicamente de Estados Unidos y en el contexto que todos los participantes pertenecen a una Universidad de Estados Unidos en New York, todos los participantes son estudiantes o graduados universitarios, heterosexuales, por lo tanto, las conclusiones no son tan fácilmente extrapolables en otras culturas o regiones. Además, las características etnográficas de las 552 personas no poseen la misma proporción de las distintas razas que componen la sociedad de New York en el momento del experimento, esto es un sesgo importante a la hora de obtener conclusiones de los datos obtenidos.

Las preferencias de las personas por ciertos atributos en parejas potenciales pueden cambiar con el tiempo, mientras que solo estudiamos las preferencias de los usuarios en la elección de pareja en un momento particular.

Podemos demostrar que la autocalificación de una persona, en este experimento, es **distinta** al puntaje promedio que le colocan todas las otras parejas que lo conoce durante 4 minutos en lo relacionado a cuán atractivo/a, sincero/a, inteligente, divertido/a y cuán ambicioso/a es dicha persona. Los seres humanos somos subjetivos entre otras cosas, porque nuestras evaluaciones y percepciones están influenciadas por nuestras propias experiencias, valores, creencias, prejuicios y expectativas. Nuestras experiencias pasadas y las emociones asociadas con ellas pueden influir en cómo percibimos y evaluamos a las personas en el presente. Nuestras creencias y valores de la misma forma pueden influir en cómo evaluamos a los demás y nuestros prejuicios y expectativas son otro de los factores que pueden influir en nuestras evaluaciones.

Lo que podemos observar en todos los casos es que la media de las autoevaluaciones de las 5 características siempre es superior a la media de la evaluación de la pareja, las personas en este experimento se autoevalúan “mejor” de lo que lo evalúan el promedio de sus parejas.

En este ensayo, se puede llegar a la conclusión con el **95% de confianza** de que sólo se volverán a juntar en otra cita entre un **14,88% y un 15,06%** de las parejas que se conocen por primera vez. Podemos decir que, si 2 personas tienen interés en tener una cita con otra persona, tiene que reunirse durante 4 minutos con por lo menos 8 personas, para tener altas probabilidades de tener una nueva cita, considerando las restricciones del experimento.

Por un margen significativo, el **atractivo** del compañero es la variable más importante para ambos modelos y en ambas poblaciones. Por otro lado, la **inteligencia** del compañero no es una predictora excesivamente fuerte de las decisiones de las **mujeres** si se la compara con el atractivo.

Se llega a las mismas conclusiones en Fisman et al. (2.006) y en nuestro trabajo respecto a las preferencias por la misma **raza**: es una predictora significativa en el caso de las **mujeres**, mientras que para los **hombres** no afecta drásticamente la probabilidad (estimada) de querer una segunda cita. En este sentido, quizá la mirada debería estar puesta no sólo sobre el género del sujeto y los atributos de su compañero, sino sobre sus intereses generales, su estado de ánimo al tener la cita, su nivel de optimismo, entre otros. Por ejemplo, el hecho de que la respuesta del sujeto a la pregunta: ¿Qué tan probable crees que es, que le gustes a tu pareja? ("guess_prob_liked") sea un predictor tan importante para los modelos analizados, revela la necesidad de mayor indagación al respecto.

Hay varias vías para futuras investigaciones, podemos utilizar los resultados obtenidos en el documento para estudiar más a fondo el problema de la pareja estables para la elección potencial de pareja y considerar si la elección de una nueva cita, determina, poder formar una pareja en el largo plazo, si esto puede generar un compromiso más prolongado entre las parejas.

Apéndice A

sección 1

Detalles de las variables

A continuación, se detalla cada variable / atributo:

- wave: ronda, número de serie del experimento (0 - 21)
- gender: género de uno mismo (male-female)
- age: edad de uno mismo
- age_o: edad de la pareja
- d_age: diferencia de edad con la pareja
- carrera: carrera de uno mismo
- race: raza de uno mismo
- race_o: raza del compañero
- samerace: Si las dos personas tienen la misma raza o no. (1-0)

- important_same_race: ¿Qué tan importante es para mí que el compañero sea de la misma raza? (0-10)
- important_same_religion: ¿Qué tan importante es para mí que la pareja tenga la misma religión? (0-10)
- field: que estudia o ha estudiado la persona, campo de estudio

Distribuir entre los 6 atributos un peso del 0 al 100:

- pref_o_attractive: ¿Qué tan importante califica la pareja el atractivo?
- pref_o_sinsere: ¿Qué tan importante califica la pareja la sinceridad?
- pref_o_intelligence: ¿Qué tan importante califica la pareja la inteligencia?
- pref_o_funny: ¿Qué tan importante califica la pareja ser divertido?
- pref_o_ambitious: ¿Qué tan importante califica la pareja ser ambición?
- pref_o_shared_interests: ¿Qué tan importante califica la pareja tener intereses compartidos?

Calificación de la pareja del 0 al 10 en los siguientes atributos:

- atractivo_o: del 1 al 10 cuán atractivo te calificó tu pareja
- sincere_o: del 1 al 10 cuán sincero/a calificó tu pareja
- intelligence_o: del 1 al 10 cuán inteligente calificó tu pareja
- funny_o: del 1 al 10 cuán divertido/a calificó tu pareja
- ambitus_o: del 1 al 10 cuán ambicioso/a calificó tu pareja
- shared_interests_o: del 1 al 10 cuán iguales te calificó tu pareja en intereses compartidos.

Distribuir entre los 6 atributos un peso del 0 al 100:

- attractive_important: ¿Qué importancia le das al atractivo de una persona al buscar una pareja para una cita?
- sincere_important: ¿Qué importancia le das a la sinceridad de una persona al buscar una pareja para una cita?
- intellicence_important: ¿Qué importancia le das a la inteligencia de una persona al buscar una pareja para una cita?
- funny_important: ¿Qué importancia le das a que lo divertida que puede ser una persona al buscar una pareja para una cita?
- ambtition_important: ¿Qué importancia le das a la ambición que posea una persona al buscar una pareja para una cita?
- shared_interests_important: ¿Qué importancia le das a los intereses compartidos de una persona al buscar una pareja para una cita?

Autocalificación del 1 al 10 sobre los siguientes atributos:

- attractive: del 1 al 10 cuán atractivo te autocalificas
- sincere: del 1 al 10 cuán sincero te autocalificas
- intelligence: del 1 al 10 cuán inteligente te autocalificas
- funny: del 1 al 10 cuán divertido te autocalificas
- ambition: del 1 al 10 cuán ambicioso te autocalificas

Evaluar con un puntaje de 0 al 10 a la pareja en el siguiente atributo:

- attractive_partner: del 1 al 10 qué tan atractivo calificás a tu pareja.
- sincere_partner: del 1 al 10 qué tan sincero/a calificás a tu pareja.
- intelligence_partner: del 1 al 10 qué tan inteligente calificás a tu pareja.
- funny_partner: del 1 al 10 qué tan divertido calificás a tu pareja.
- ambition_partner: del 1 al 10 qué tan ambicioso calificás a tu pareja.
- shared_interests_partner: del 1 al 10, como calificás a tu pareja en qué tan iguales son los intereses de ambos.

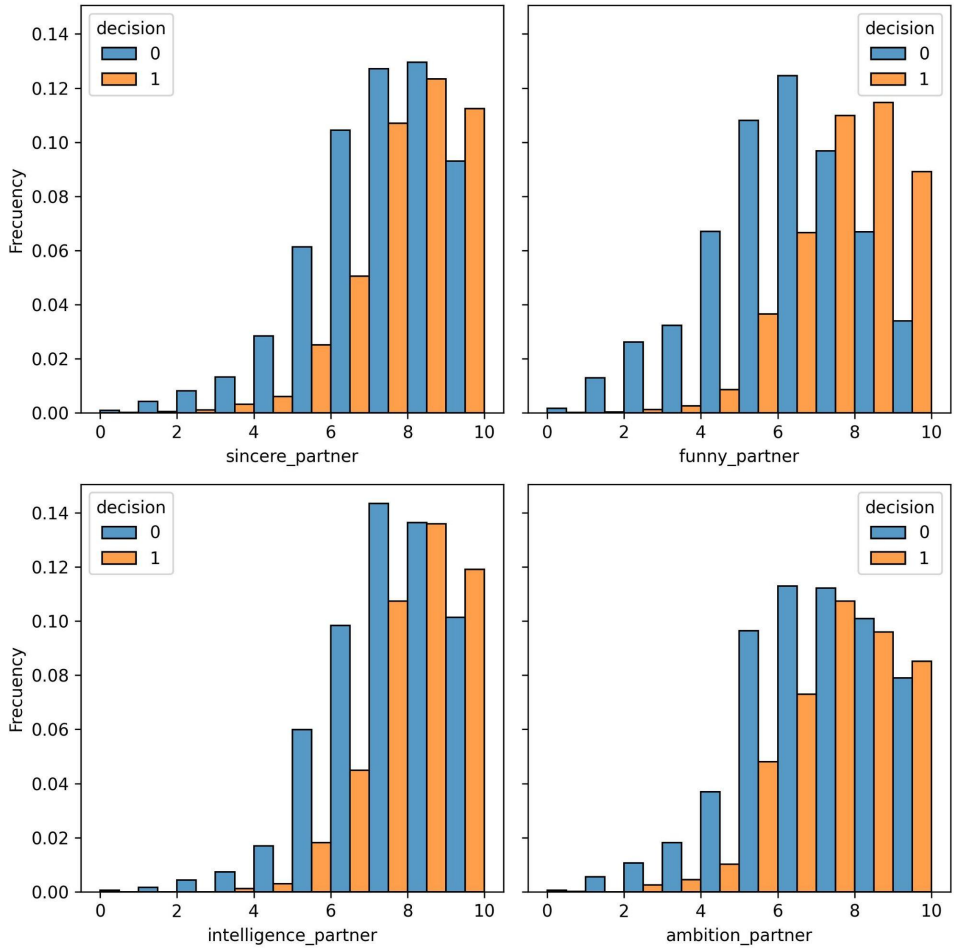
Intereses compartidos (0-10)*:*

- sports: deportes
- tvsports: mirar eventos deportivos por televisión
- exercise: ejercicio físico
- dining: sibaritas, preferencia por la buena comida, salir a comer
- museums: visitar museos museos
- art: preferencias por el arte
- hiking: senderismo, montañismo, trekking
- gaming: juegos
- clubbing: discotecas
- reading: lectura
- tv: televisión
- theater: teatro
- movies: películas
- concerts: conciertos
- music: música
- shopping: compras
- yoga

- interests_correlate: correlación entre las calificaciones de intereses del participante y de la pareja
- expected_happy_with_sd_people: ¿Qué tan feliz esperas estar con las personas queconoces durante el evento de citas rápidas?
- expected_num_interested_in_me: de las personas que conocerás, ¿cuántas esperas que estén interesadas en salir contigo?
- expected_num_matches:número_esperado_de_coincidencias:¿Cuántas coincidencias espera obtener?
- like: ¿Te gustaba tu pareja?
- guess_prob_liked: ¿Qué tan probable crees que es, que le gustes a tu pareja?
- met: ¿Cuántas veces has visto a tu pareja antes?
- decision: decisión en la noche del evento (1 = quiero tener una nueva cita con esa persona / 0 = NO quiero tener una nueva cita con esa persona)
- decision_o: decisión de la pareja en la noche del evento (1 = mi pareja quiere tener una nueva cita conmigo / 0 = mi pareja NO quiere tener una nueva cita conmigo)
- match: coincidencia (1 = ambos quiere volver a tener una cita / 0 = una de los 2 no quiere volver a tener una cita con el otro)

sección 2

Distribuciones de los atributos de acuerdo a la decisión



sección 3

Casos de inconsistencia en las variables.

```
attractive_o
 [ 0.   1.   2.   3.   3.5  4.   5.   6.   6.5  7.   7.5  8.   8.5  9.
  9.5  9.9 10.  10.5  nan]
sinsere_o
 [ 0.   1.   2.   3.   4.   4.5  5.   6.   7.   7.5  8.   8.5  9.  10.
  nan]
intelligence_o
 [ 0.   1.   2.   2.5  3.   4.   5.   5.5  6.   6.5  7.   7.5  8.   8.5
  9.   9.5 10.   nan]
funny_o
 [ 0.   1.   2.   3.   4.   5.   5.5  6.   6.5  7.   7.5  8.   8.5  9.
  9.5 10.  11.   nan]
ambitous_o
 [ 0.   1.   2.   3.   4.   5.   5.5  6.   7.   7.5  8.   8.5  9.   9.5
 10.   nan]
shared_interests_o
 [ 0.   1.   2.   3.   4.   5.   5.5  6.   6.5  7.   7.5  8.   8.5  9.
 10.   nan]
attractive
 [ 2.  3.  4.  5.  6.  7.  8.  9. 10. nan]
sincere
 [ 2.  3.  4.  5.  6.  7.  8.  9. 10. nan]
intelligence
 [ 2.  3.  4.  5.  6.  7.  8.  9. 10. nan]
```

Apéndice B

sección 1

Coeficientes de los modelos de regresión logística

Población Femenina:

	coef	lasso_test_importance
attractive_partner	0.608339	47.75
guess_prob_liked	0.364746	16.97
shared_interests_partner	0.212422	15.27
attractive	-0.048773	15.13
funny_partner	0.439784	13.25
expected_num_matches	0.139990	8.76
sincere	-0.079412	8.58
race_european	-0.158686	6.16
expected_happy_with_sd_people	0.050100	5.83
clubbing	0.108211	4.26

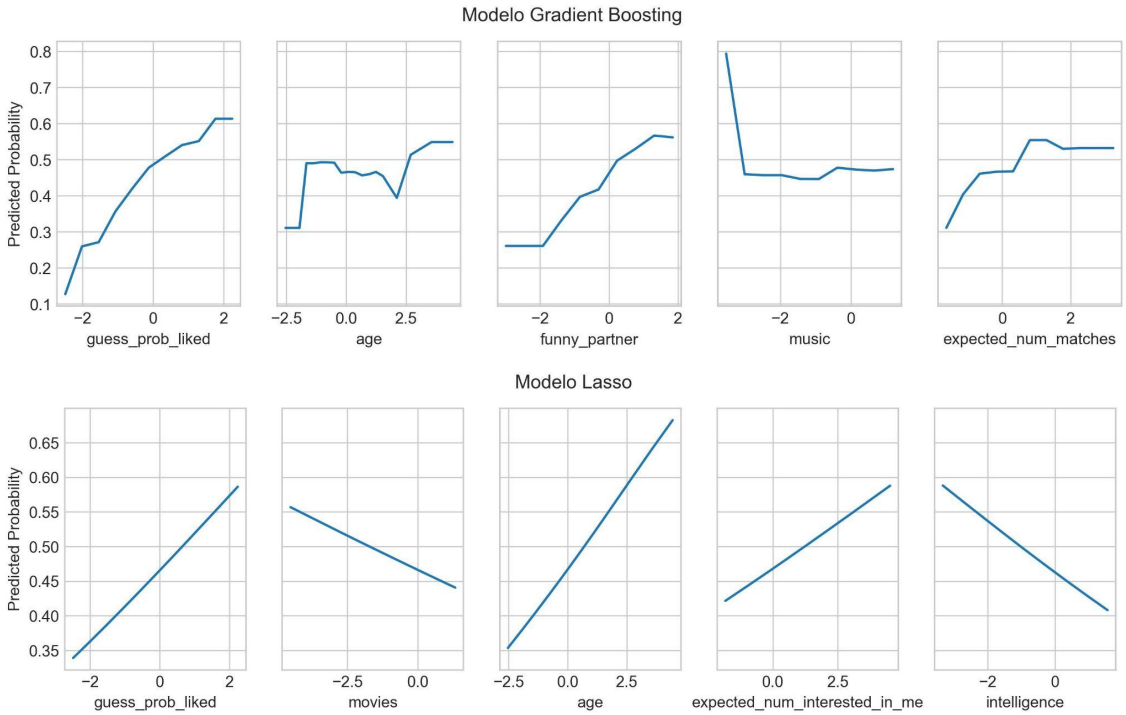
Población Masculina:

	coef	lasso_test_importance
attractive_partner	0.305023	60.80
guess_prob_liked	0.524052	19.17
expected_num_matches	-0.244525	7.73
importance_same_religion	-0.195790	6.81
intelligence	-0.384890	6.54
dining	-0.137517	4.63
shared_interests_partner	0.233069	4.12
shared_interests_important	-0.143048	3.19
ambtition_important	-0.066838	3.08
tv	0.053017	1.92

sección 2

Gráficos de Dependencia Parcial

Población Masculina:



Población Femenina: