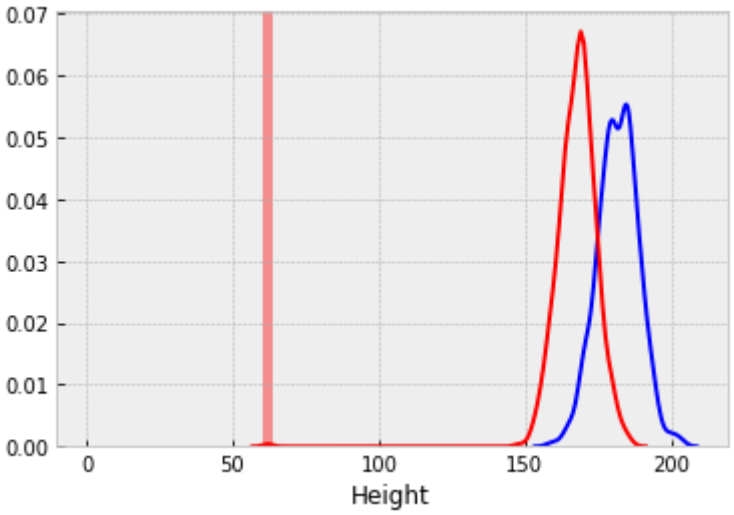
Habiendo trabajado con un conjunto de datos que contiene respuestas de gente Joven a una encuesta sobre temas relacionados con los hábitos, preferencias, datos demográficos y de personalidad de los mismos, se observaron algunas particularidades que compartiremos a continuación.

El problema de los casos Atípicos

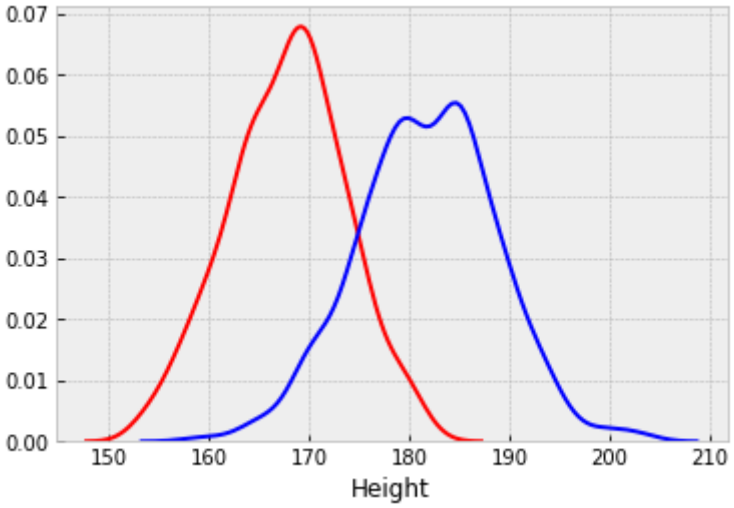
Lo principal es entender que en todo set de datos pueden existir desviaciones o casos que se diferencian de los demás. Estos se conocen como casos atípicos u outliers, y es muy importante que se estudie su inclusión o no dentro de los análisis que se pueden realizar sobre las variables del set de datos. Por ejemplo, al analizar la altura de los hombres y mujeres, encontramos que existen algunos casos que se diferencian en gran medida de los demás, haciendo que la visualización y el estudio de los mismos se vea afectado.

Si graficamos la distribución de la variable ALTURA, estableciendo una serie para las MUJERES (rojo) y otra para los HOMBRES (azul), sin realizar ningún filtro de outliers, obtenemos lo siguiente:



Como se puede observar, para la serie de las mujeres, existen valores con muy baja frecuencia y que se alejan de los valores “medios” o típicos de la serie. Esto afecta a la visualización de los casos.

Sin embargo, trabajando con un filtro que elimine aquellos casos que se desvían más de 2 veces la desviación estándar de la variable, se obtiene un set de datos con casos más relacionados y que permite una gráfica más “limpia” y fácil de analizar.



Con esta gráfica, es más fácil interpretar como se comporta la variable Altura, tanto para las MUJERES como para los HOMBRES. Por ejemplo, podemos asegurar que tenemos mayores frecuencias de alturas superiores a los 180 cm en los HOMBRES que en las MUJERES (como era de esperar), algo que con el grafico anterior era difícil de ver.

Con esto resaltamos el valor del análisis de OUTLIERS.

Correlación entre variables

De la correlación de variables del dataset, identificamos que las más fuertes son las siguientes:

Tipo Variable	Variable 1	Variable 2	Relación
PASATIEMPO	Biologia	Medicina	0,715551
PASATIEMPO	Quimica	Biologia	0,689859
PREFERENCIA PELICULAS	Animated	Fantasy/Fairy tales	0,674675
PREFERENCIA MUSICAL	Opera	Musica Clasica	0,5959

Además de ser LOGICAS estas relaciones, la fuerte correlación que presentan hacen que sea indiscutible que estas variables estén relacionadas.

Por su parte, también detectamos que existe una correlación importante entre aquellos casos que FUMAN y TOMAN ALCOHOL. Lo cual se puede ver a través de la siguiente tabla.

		index								
	Alcohol	NaN	drink a lot	never	social drinker	NaN	current smoker	former smoker	never smoked	tried smoking
	Alcohol									
index	NaN	1.000000	-0.037438	-0.026387	-0.096648	0.471112	0.002328	-0.032291	-0.001036	-0.060733
	drink a lot	-0.037438	1.000000	-0.198567	-0.727281	0.006516	0.223494	0.104451	-0.181622	-0.108869
	never	-0.026387	-0.198567	1.000000	-0.512605	0.034641	-0.094404	-0.083581	0.264579	-0.084154
	social drinker	-0.096648	-0.727281	-0.512605	1.000000	-0.098978	-0.129642	-0.028475	-0.024242	0.161626
	NaN	0.471112	0.006516	0.034641	-0.098978	1.000000	-0.042872	-0.040906	-0.045505	-0.076936
	current smoker	0.002328	0.223494	-0.094404	-0.129642	-0.042872	1.000000	-0.219652	-0.244345	-0.413123
	former smoker	-0.032291	0.104451	-0.083581	-0.028475	-0.040906	-0.219652	1.000000	-0.233142	-0.394181
	never smoked	-0.001036	-0.181622	0.264579	-0.024242	-0.045505	-0.244345	-0.233142	1.000000	-0.438495
	tried smoking	-0.060733	-0.108869	-0.084154	0.161626	-0.076936	-0.413123	-0.394181	-0.438495	1.000000

Aun así, para validar su independencia, y al ser variables del tipo CATEGORICAS, utilizamos el método de Chi Cuadrado.

Por otro lado, al trabajar con las variables se detectó que la probabilidad conjunta de que una persona sea Fumador y Tomador era del 0,077.
A partir de ello podemos inferir que de cerrarse los Bares, se reduciría en este 0,077 la cantidad de personas que son Fumadores y Tomadores. Esto lo podemos asegurar con cierto grado de confianza con la prueba de hipótesis.

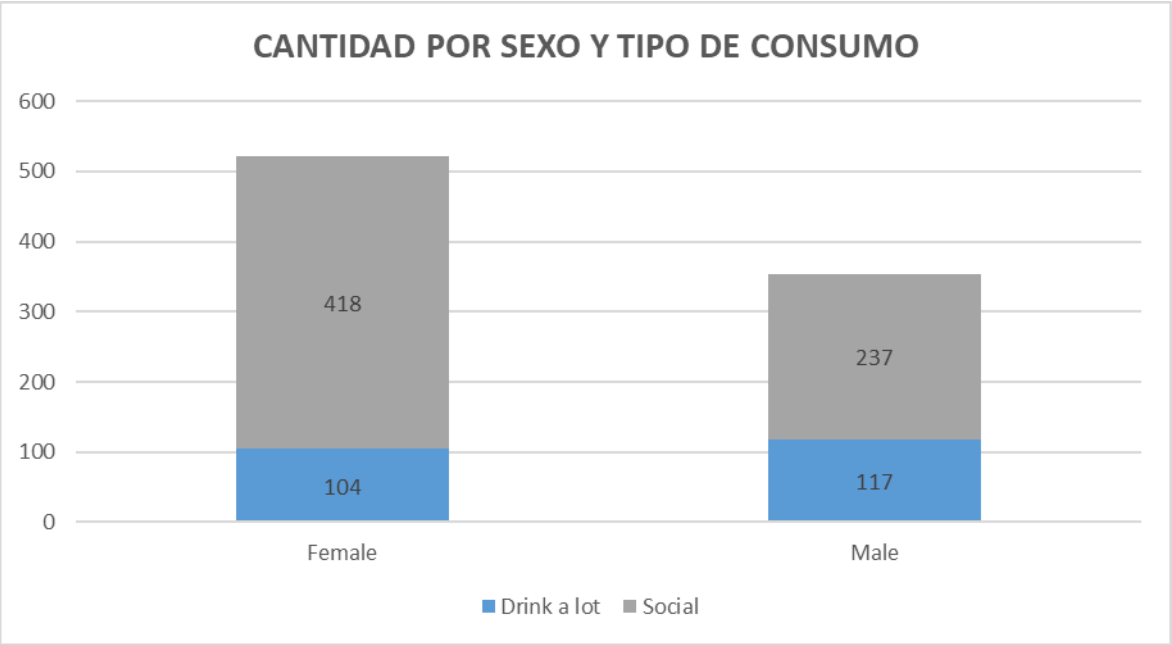
Distribución de los ejemplos con respecto a una clase

Alcohol ¿Cómo es el comportamiento o preferencias de aquellos que lo consumen?

Comencemos por el sexo, históricamente se atribuyó el consumo de alcohol a los hombres, sin embargo, analizando el resultado de las encuestas, encontramos que no existe una gran diferencia entre ambos sexos.

Inclusive, si consideremos a los “Bebedores Sociales” dentro de la categoría de los que toman alcohol, la proporción se vuelca a favor de las mujeres. Algo que rompe con las creencias

En la siguiente grafica podemos ver como, en la muestra tomada, la cantidad de bebedores sociales es mayor en las mujeres, haciendo que al contemplar ambas categorías de tomadores (Frecuentes y sociales), las mujeres tengan una proporción mayor por sobre los hombres.



Si abrimos este subconjunto de datos de Mujeres que beben socialmente, encontramos que la mayor cantidad de casos se presenta en edades tempranas, es decir, menores a 21 años. Esto supone una necesidad de los padres de estar atentos para evitar problemas que lleven a un bebedor frecuente.