Análisis de Dataset Beer Reviews

Descripción breve de tendencias de centralización, y relaciones entre variables por medio de preguntas y respuestas.

Autor: Luis Orellana Altamirano

Descripción Preliminar.

El presente dataset llamado "Beer Review"[1], está constituido de 1.518.829 instancias, un conjunto de 12 atributos y una clase, la cual describe la calidad de la cerveza en base a puntuaciones subjetivas efectuadas por distintos usuarios, los cuales, identificados por medio de un nombre de perfil, calificaron un conjunto de cervezas agrupadas tanto por estilos como cervecerías productoras.

A continuación, se describe el dataset:

- brewery_id (valor continuo numérico): Identificación de la cervecería productora.
- brewery_name (valor nominal): Nombre de la cervecería productora.
- review_time (valor continuo de punto flotante): Transcurso de tiempo que tardo el usuario en evaluar cada instancia.
- review_aroma (valor numérico de cinco niveles. 1-5): Calificación del aroma de la cerveza.
- review_appearance (valor numérico de cinco niveles. 1-5): Calificación de la apariencia de la cerveza.
- review_profilename (valor nominal alfanumérico): Nombre del usuario calificador.
- review_palate (valor numérico de cinco niveles. 1-5): Calificación del sabor en paladar de la cerveza.
- review_taste (valor numérico de cinco niveles. 1-5): Calificación del sabor de la cerveza.
- beer_abv (valor continuo de punto flotante): Índice de fuerza (marcado sabor) de la cerveza.
- beer_style (valor nominal alfanumérico): Nombre del estilo de la cerveza.
- beer_name (valor nominal alfanumérico): Nombre de la cerveza.
- beer_beerid (valor continuo): Identificación de la cerveza
- review_overall (valor numérico de seis niveles. 0-5): Clase indicadora del nivel de calidad de la cerveza.

Medidas de centralización y frecuencia.

brewery_id		**brewery_name**		**review_time**	
Min.	1	**Boston Beer Company (Samuel Adams)**	38812	**Min.**	8.407e+08
1st Qu.	141	**Dogfish Head Brewery**	33800	**1st Qu.**	1.176e+09
Median	417	**Stone Brewing Co.**	33022	**Median**	1.240e+09
Mean	3074	**Sierra Nevada Brewing Co.**	28637	**Mean**	1.225e+09
3rd Qu.	2298	**Bell's Brewery, Inc.**	24975	**3rd Qu.**	1.289e+09
Max.	28003	**(Other)**	1359583	**Max.**	1.326e+09

[1] https://s3.amazonaws.com/demo-datasets/beer_reviews.tar.gz

review_overall		review_aroma		review_appearance		review_profilename	
Min.	0.000	**Min.**	1.000	**Min.**	0.00	**northyorksammy**	5346
1st Qu.	3.500	**1st Qu.**	3.500	**1st Qu.**	3.50	**mikesgroove**	4283
Median	4.000	**Median**	4.000	**Median**	4.00	**BuckeyeNation**	4246
Mean	3.824	**Mean**	3.746	**Mean**	3.85	**Thorpe429**	3273
3rd Qu.	4.500	**3rd Qu.**	4.000	**3rd Qu.**	4.00	**brentk56**	3186
Max.	5.000	**Max.**	5.000	**Max.**	5.00	**(Other)**	1498495

beer_style		review_palate		review_taste	
American IPA	113164	**Min.**	1.000	**Min.**	1.000
American Double / Imperial IPA	85124	**1st Qu.**	3.500	**1st Qu.**	3.500
American Pale Ale (APA)	58081	**Median**	4.000	**Median**	4.000
Russian Imperial Stout	53432	**Mean**	3.754	**Mean**	3.804
American Double / Imperial Stout	50146	**3rd Qu.**	4.000	**3rd Qu.**	4.500
(Other)	1158882	**Max.**	5.000	**Max.**	5.000

beer_name		beer_abv		beer_beerid	
90 Minute IPA	3290	**Min.**	0.010	**Min.**	5
Old Rasputin Russian Imperial Stout	3111	**1st Qu.**	5.200	**1st Qu.**	1654
Sierra Nevada Celebration Ale	3000	**Median**	6.500	**Median**	12827
India Pale Ale	2960	**Mean**	7.042	**Mean**	21404
Two Hearted Ale	2728	**3rd Qu.**	8.500	**3rd Qu.**	39236
(Other)	1503740	**Max.**	57.700	**Max.**	77316

Es posible observar que la media de la calidad de la cerveza (review overall) es un valor cercano a 4, lo cual indica que, en su mayoría, las cervezas presentes en este dataset son productos de buena calidad desde la perspectiva de los usuarios calificadores. Esto también se encuentra sustentado por el primer cuartil, el cual comienza con una puntuación de 3.5, lo cual es una calificación alta.

Medidas muy similares son encontradas para los atributos aroma, apariencia, sabor y paladar. Sin embargo, la menor dispersión con respecto a la media es la apariencia, la cual también posee una media incluso mayor que la media presente en la calidad de cerveza.

Preguntas y Discusión.

- ***¿Qué cervecería produce la cerveza más fuerte según ABV?***

El objetivo principal para este punto es obtener el mayor índice "beer_abv", y de esta manera, conocer cuál es el nombre de la cervecería que produce la cerveza con sabor más fuerte. Además, con la finalidad de no obtener múltiples nombres repetidos de la misma instancia, se debe especificar la respuesta como valor único.

Teniendo esto en mente, la cervecería que produce las cervezas con un sabor más fuerte es "SchorschbrÃ¤u".

- ***¿Si tuviera que elegir 3 cervezas para recomendar usando sólo estos datos, cuáles elegiría?***
  La calidad de cada cerveza está dada por puntuaciones subjetivas de los usuarios. Tal puntuación se recoge en los campos:
    - review_aroma
    - review_appearance
    - review_palate
    - review_taste

  Cabe mencionar que el equivalente de la puntuación entre todas estas variables es el atributo "review_overall". Todas estas variables están en el rango de 0 a 5, por lo que, al sumarlas, se obtendrá una ponderación total de la calidad subjetiva de cada cerveza.

  Este método tiene una ventaja y desventaja en comparación a la media aritmética. Por un lado, al realizar una sumatoria de las ponderaciones parciales (calificaciones y luego agrupación por nombre de cervezas) es un paso menos de computo que la media. Por otro lado, la media entre calificaciones parciales de cada cerveza permite un valor entre 0 y 5, lo cual posibilita a interpretar con mayor facilidad tal cifra. Sin embargo, debido a que tan solo se pretende obtener el nombre especifico de las primeras tres mejores cervezas, se ha preferido la sumatoria por sobre la media aritmética.

  Seguido de la sumatoria entre cada una de las calificaciones (se ha creado un nuevo campo llamado "sumatoria"), se ha realizado una agrupación de los nombres de cervezas, al mismo tiempo que se realiza la sumatoria de cada grupo, por lo que finalmente se obtiene una calificación total (la cual no es interpretable directamente en comparación a una media aritmética). Finalmente se ordena de manera descendente cada cerveza según la sumatoria total de las calificaciones.

  Con todo lo anterior mencionado, se concluye que las mejores cervezas según la ponderación total de cada calificación subjetiva de los usuarios son:

    - 90 Minute IPA

- Old Rasputin Russian Imperial Stout
- Sierra Nevada Celebration Ale

- ***¿Cuál de los factores (aroma, taste, aperance, palete) es más importante para determinar la calidad general de una cerveza?***

  En el caso de esta pregunta, el problema principal puede ser interpretado como una correlación entre una serie de variables, y un atributo que determina la clasificación de cada instancia.

  Siguiendo esta lógica, las variables correlacionadas con la etiqueta son:

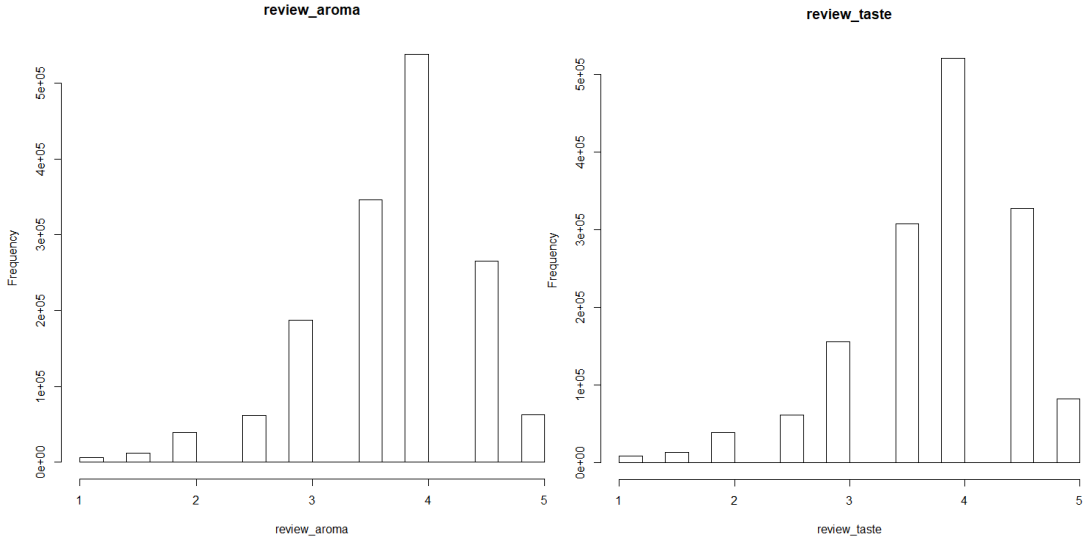
  - review_aroma
  - review_taste
  - review_appearance
  - review_palate

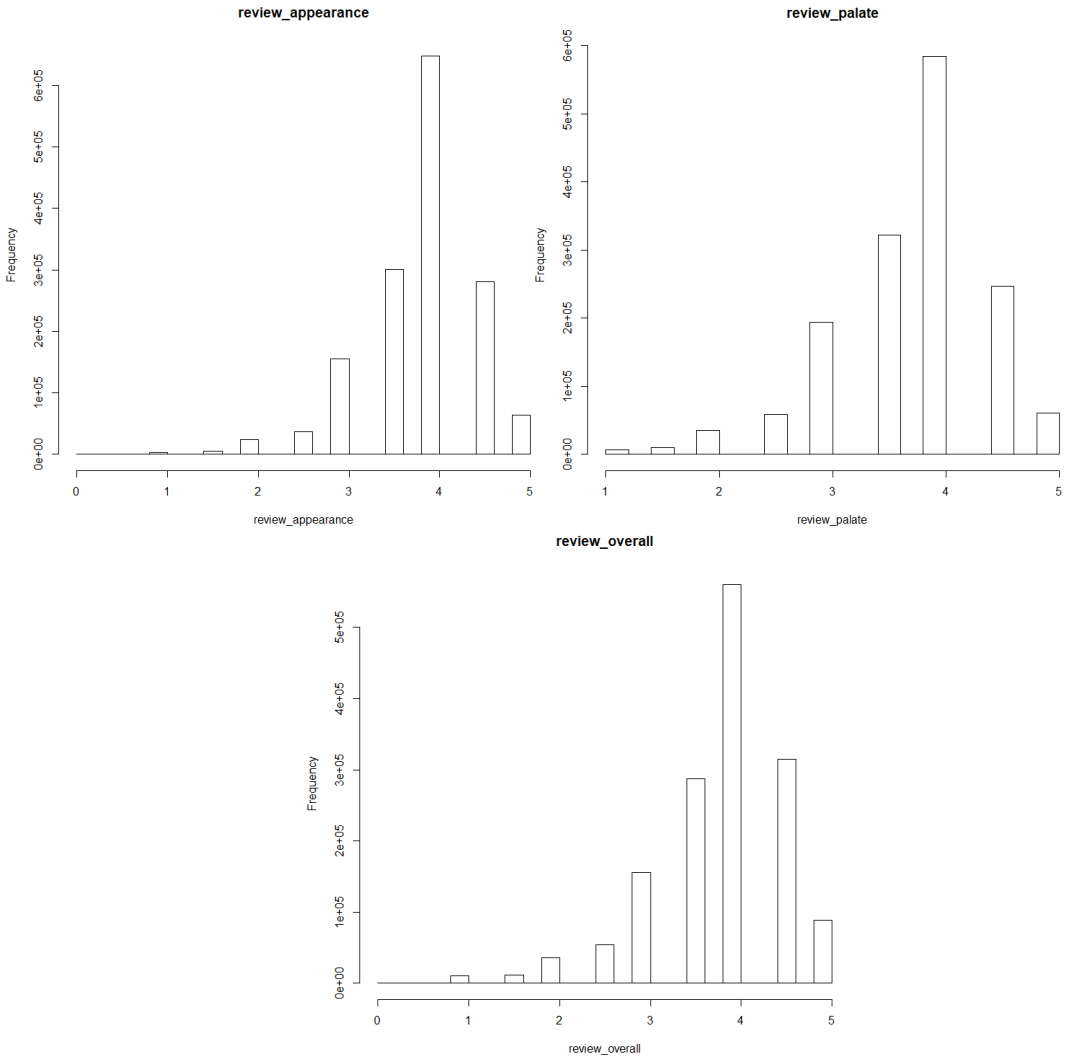
  Y la variable que corresponde a la etiqueta para este caso particular es:

  - review_overall

  Con la finalidad de emplear índices de correlación, es necesario determinar qué tipo de distribución sigue cada una de las variables mencionadas. En el caso de que la distribución de probabilidades sea identificada como una distribución normal, se empleara un test de hipótesis paramétrico. Por otro lado, si la distribución de las variables no es normal, será utilizado un test no paramétrico.

  La distribución de cada variable es mostrada a continuación:





Como es posible observar, todas las variables implicadas mantienen una distribución normal, por lo que los test paramétricos funcionan de mejor manera en este caso al contrastarlos con test no paramétricos como es el caso de la correlación de Spearman.

Con lo anteriormente mencionado, el test de correlación de Pearson se ajusta de mejor manera a tales variables. Los resultados obtenidos corresponden al análisis entre cada una de las variables y la calidad de la cerveza (review_overall):

- data.review_aroma: 0.6128259
- data.review_taste: 0.7871930
- data.review_appearance: 0.4985761
- data.review_palate: 0.6990402

Con esto, se ha dado sustento a la hipótesis alternativa, y se descarta la hipótesis nula; hay relación significativa entre una de las variables y la calidad de cerveza. Para este caso es el sabor de la cerveza (data.review_taste) en donde se observa una relación directamente proporcional entre la calidad de esta y su sabor.

- ***Si yo típicamente disfruto una cerveza debido a su aroma y apariencia, ¿qué estilo de cerveza debería probar?***

  Debido a que esta pregunta en particular plantea dos variables al mismo tiempo para determinar la calidad de la cerveza, y con esto obtener el mejor estilo de cerveza, se ha optado por crear una nueva variable que contenga la sumatoria de la característica "review_aroma" y "review_appearance" para cada una de las instancias.

  Luego de realizar la sumatoria de estas dos variables en cada una de las instancias, se realiza una agrupación de cada revisión (calificación) con respecto al estilo de cerveza al mismo tiempo que se realizan sumatorias en cada una de las agrupaciones. Luego se ordena de manera descendente los resultados. Por consiguiente, obteniendo el mejor estilo de cerveza en base a su aroma y apariencia, el resultado obtenido por medio del procedimiento descrito es "American IPA".

Anexo.

Código en lenguaje R.

```
#Ruta del Dataset

RUTA_DATASET <- "" #Ruta donde se encuentra el dataset

setwd(RUTA_DATASET)

#Leer Dataset

data=read.csv("beer_reviews.csv")


#Eliminar datos NA

data= na.omit(data)


#Resumen medidas de tendencia

summary(data)


#Desviación estándar para observar variación con respecto a la media

sd(data$review_appearance)

sd(data$review_aroma)

sd(data$review_palate)

sd(data$review_taste)


#¿Qué cervecería produce la cerveza más fuerte según ABV?

unique(data$brewery_name[data$beer_abv == max(data$beer_abv)])


#¿Si tuviera que elegir 3 cervezas para recomendar usando sólo estos
datos, cuáles elegiría?

#El siguiente sumatoria de los campos "review" es equivalente al campo
"review_overall"

new_data <- data.frame(nom = data$beer_name, suma = data$review_aroma +
data$review_appearance + data$review_palate + data$review_taste)
```

```
grupo <- aggregate(new_data$suma, by=list(nom=new_data$nom), FUN=sum)


head(grupo[order(grupo$x, decreasing = TRUE),], 3)


#¿Cual de los factores (aroma, taste, aperance, palete) es más importante
#para determinar la calidad general de una cerveza?

#Distribución de probabilidades en las variables

hist(data$review_aroma, xlab="review_aroma", main="review_aroma")

hist(data$review_taste, xlab="review_taste", main="review_taste")

hist(data$review_appearance, xlab="review_appearance",
main="review_appearance")

hist(data$review_palate, xlab="review_palate", main="review_palate")


hist(data$review_overall, xlab="review_overall", main="review_overall")


d <- data.frame(data$review_aroma,

                data$review_taste,

                data$review_appearance,

                data$review_palate)


cor(d, data$review_overall, method = "pearson")


#Si yo típicamente disfruto una cerveza debido a su

#aroma y apariencia, ¿qué estilo de cerveza debería probar?

new_ar_ap <- data.frame(nom = data$beer_style, suma = data$review_aroma +
data$review_appearance)

grupo <- aggregate(new_ar_ap$suma, by=list(nom=new_ar_ap$nom), FUN=sum)

head(grupo[order(grupo$x, decreasing = TRUE),], 1)
```