**Tipología y ciclo de vida de los datos**

**M2.851 - Aula 3**

**Práctica 2:**

**Limpieza y análisis de datos**

Rubén Gonzalo Coll Menéndez

Índice

[Detalles de la actividad 2](#_Toc73908393)

[Descripción 2](#_Toc73908394)

[Competencias 2](#_Toc73908395)

[Objetivos 2](#_Toc73908396)

[Resolución 2](#_Toc73908397)

[1. Descripción del dataset 3](#_Toc73908398)

[2. Importancia y objetivos de los análisis 3](#_Toc73908399)

[3. Integración y selección de los datos de interés a analizar 4](#_Toc73908400)

[4. Limpieza de los datos 5](#_Toc73908401)

[5. Análisis de los datos 8](#_Toc73908402)

[6. Representación de los resultados a partir de tablas y gráficas 10](#_Toc73908403)

[7. Resolución del problema 10](#_Toc73908404)

[8. Código 10](#_Toc73908405)

# Detalles de la actividad

## Descripción

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas. Tendréis que entregar un solo archivo con el enlace Github (https://github.com) donde se encuentren las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github.

## Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

* Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
* Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

## Objetivos

Los objetivos concretos de esta práctica son:

* Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
* Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
* Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos. Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
* Actuar con los principios éticos y legales relacionados con la manipulación de datos en
* función del ámbito de aplicación.
* Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
* Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

# Resolución

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes:

Red Wine Quality:

## Descripción del dataset

El conjunto de datos “Red Wine Quality” objeto de análisis se ha obtenido a partir de Kaggle (<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>) y está constituido por 12 características (columnas) que presentan un total de 1599 vinos (filas o registros).

Entre los campos de este conjunto de datos, encontramos los siguientes:

* quality: (calidad) Indica la calidad del vino, que va de 1 a 10. Aquí, cuanto mayor es el valor, mejor es el vino.
* fixed.acidity: Indica la cantidad de ácido tartárico en el vino y se mide en .
* volatile.acidity: Indica la cantidad de ácido acético en el vino. Se mide en .
* citric.acid: Indica la cantidad de ácido cítrico en el vino. También se mide en .
* residual.sugar: Indica la cantidad de azúcar que queda en el vino una vez finalizado el proceso de fermentación. Se mide en .
* chlorides: Indica la cantidad de sales en el vino.
* free.sulfur.dioxide: Mide la cantidad de dióxido de azufre () en forma libre. Se mide en .
* total.sulfur.dioxide: Mide la cantidad total de en el vino. Este químico funciona como agente antioxidante y antimicrobiano.

* density: Indica la densidad del vino y depende en gran medida del porcentaje de alcohol y del contenido de azúcar. Se mide en
* pH: Indica el valor de pH del vino. El rango de valor está entre 0 y 14, lo que indica una acidez muy alta, y 14 indica una acidez básica.

https://www.larioja.com/prensa/20070502/rioja\_region/sobre-valores-vino\_20070502.html

* sulphates: Indica el valor de pH del vino. El rango de valor está entre 0 que indica una acidez básica y 14, que indica una acidez muy alta.
* alcohol: Indica el contenido de alcohol del vino. Se mide en porcentaje (%).

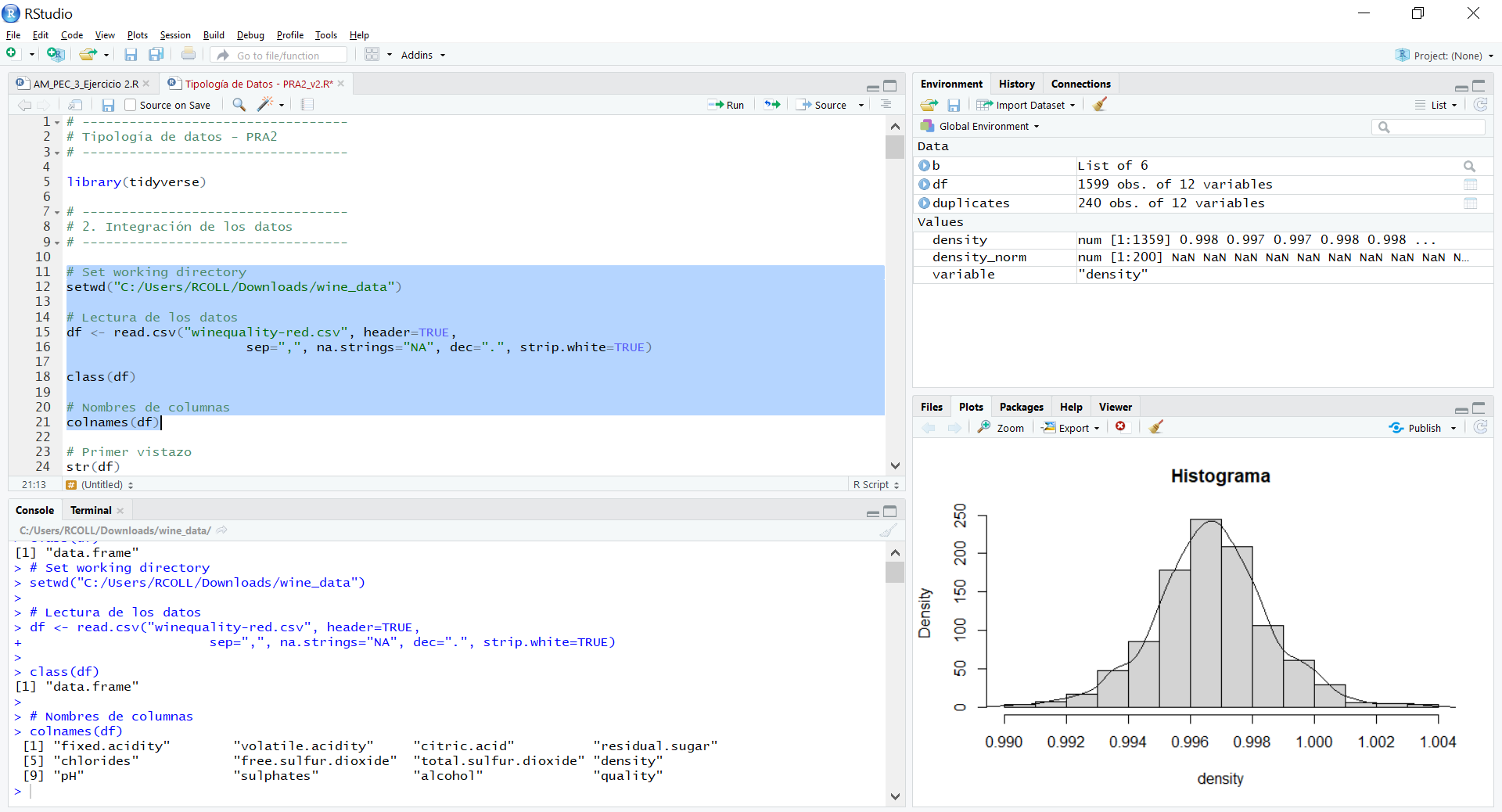
https://www.kaggle.com/debasishrath1795/redwine-linear-regression

## Importancia y objetivos de los análisis

A partir de este conjunto de datos se plantea la problemática de determinar qué variables influyen más sobre el precio de un automóvil. Además, se podrá proceder a crear modelos de regresión que permitan predecir el precio de un coche en función de sus características y contrastes de hipótesis que ayuden a identificar propiedades interesantes en las muestras que puedan ser inferidas con respecto a la población. Estos análisis adquieren una gran relevancia en casi cualquier sector relacionado con la automoción. Un ejemplo de ello se puede observar en el servicio de peritaje interno en una compañía de alquier de coches. En este caso, el perito, encargado de realizar informes técnicos que recogen la valoración económica de los coches, podría valerse de los análisis que se plantean en esta actividad para utilizarlos como soporte a la hora de llevar a cabo las tasaciones.

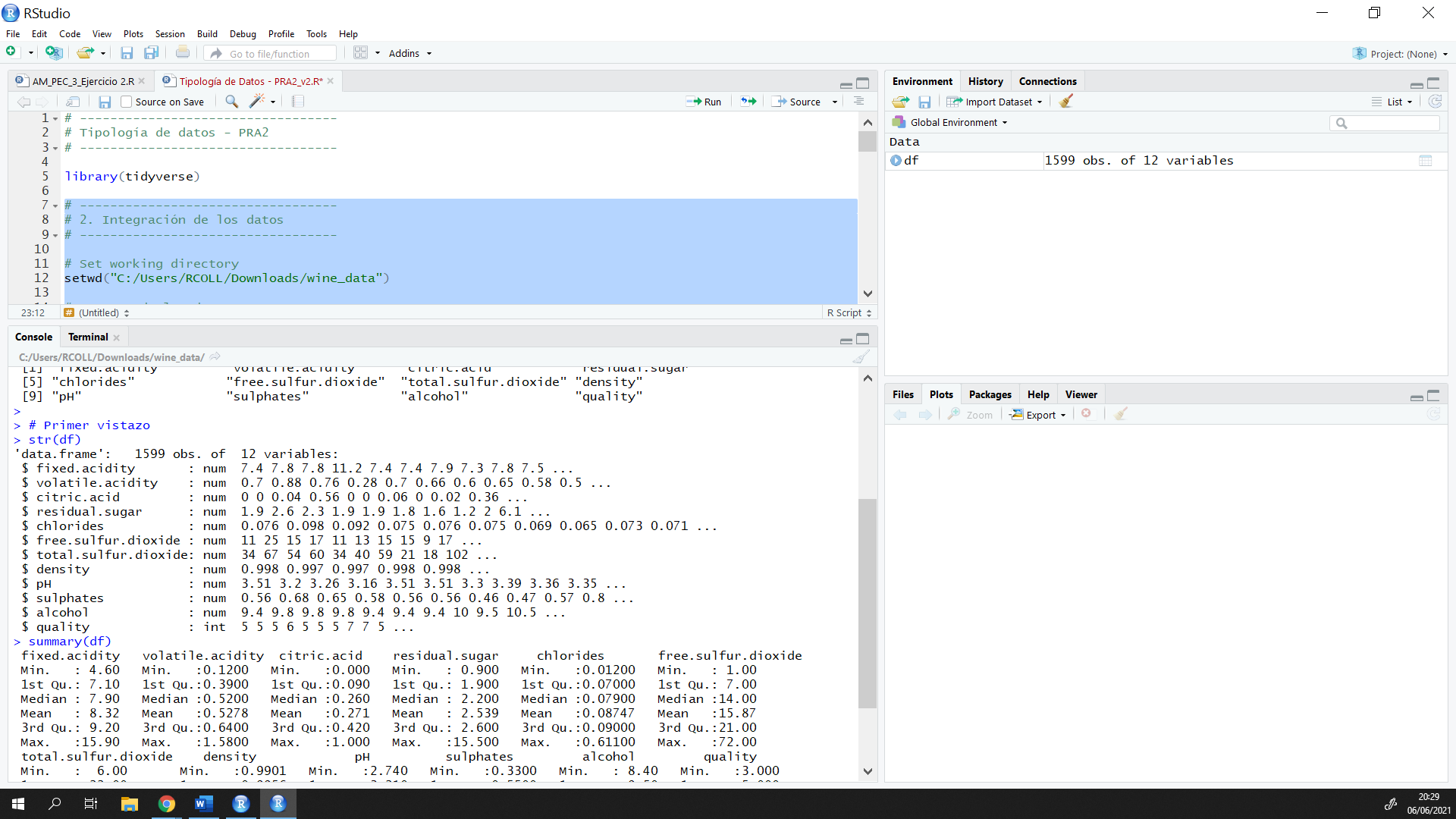
## Integración y selección de los datos de interés a analizar

Antes de comenzar con la limpieza de los datos, procedemos a realizar la lectura del fichero en formato CSV en el que se encuentra. El resultado devuelto por la llamada a la función read.csv() será un objeto data.frame:



Todas las variables de los que disponemos son, a priori, relevantes y se corresponden con los atributos que normalmente se suelen medir de un vino por lo que será conveniente tenerlos en consideración para el análisis.

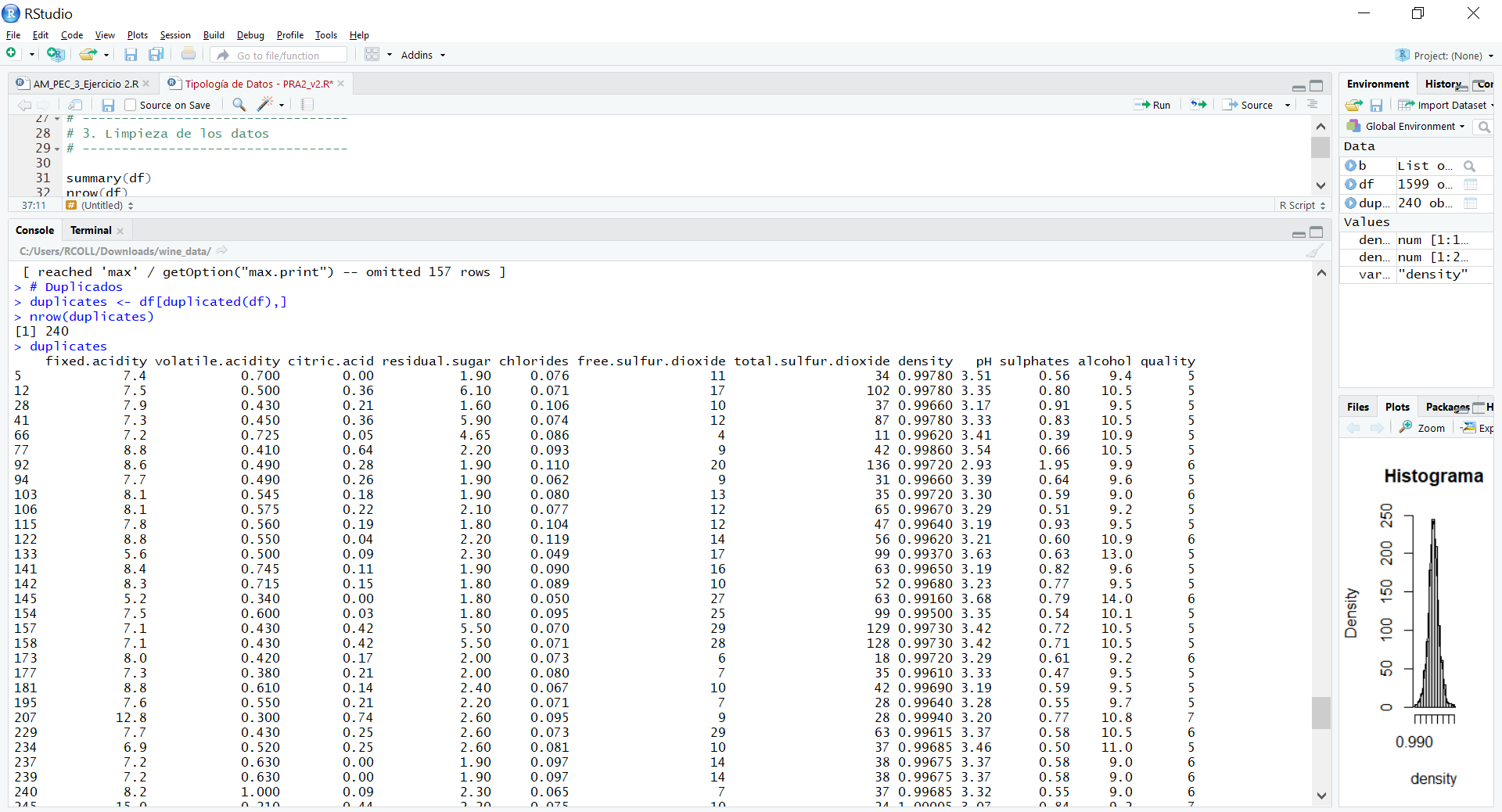
A continuación presentamos dichos atributos y el tipo de datos que contienen donde destaca que todos ellos son numéricos.



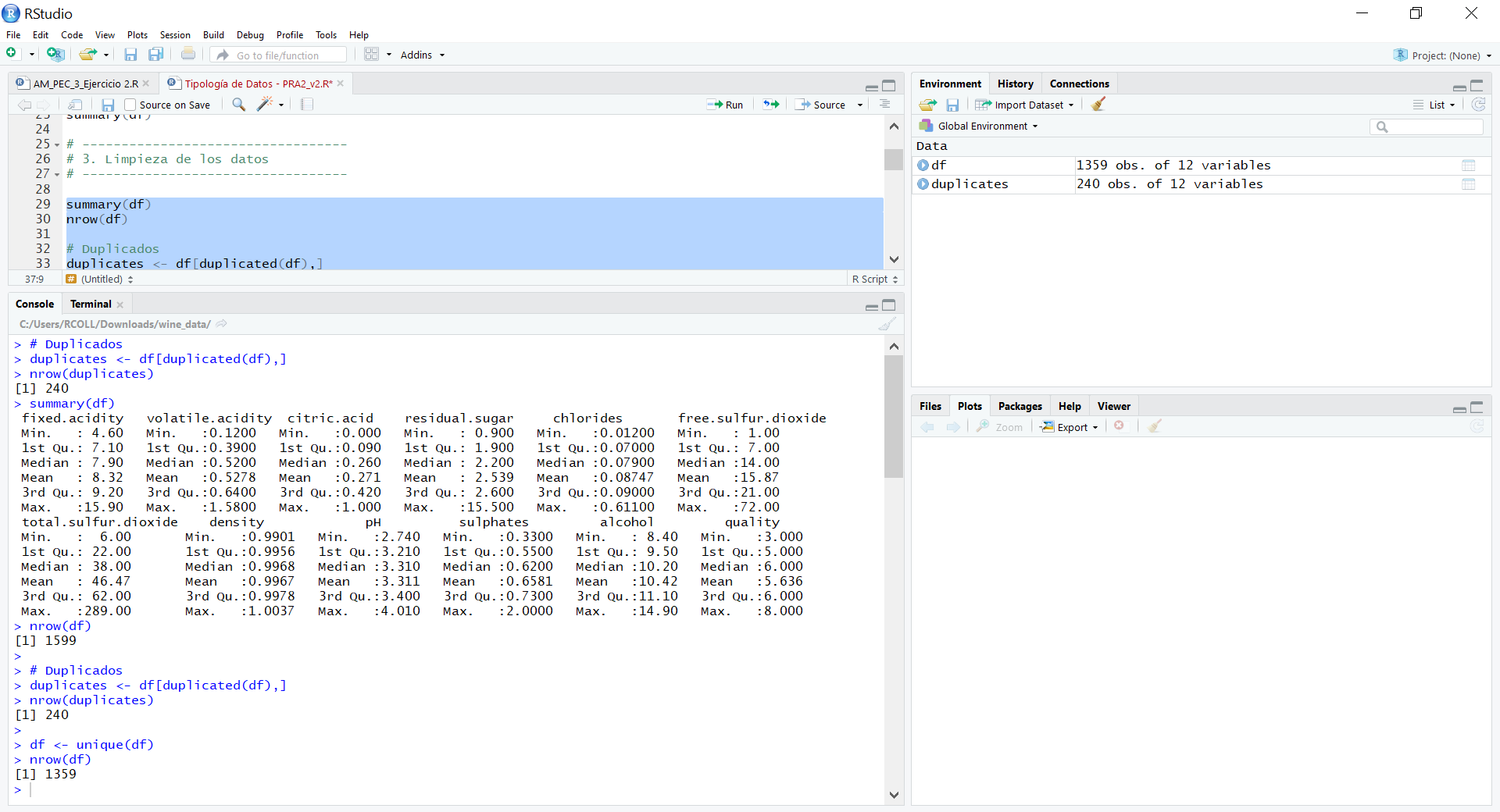
## Limpieza de los datos

* 1. **Gestión de duplicados en los datos**

En relación a la preparación y limpieza de los datos para el posterior análisis, lo primero que verificamos es la existencia de valores duplicados. Mediante el uso de la siguiente sentencia de R, comprobamos que hay 240 registros duplicados de un total de 1.599 de los que se compone el conjunto de datos.

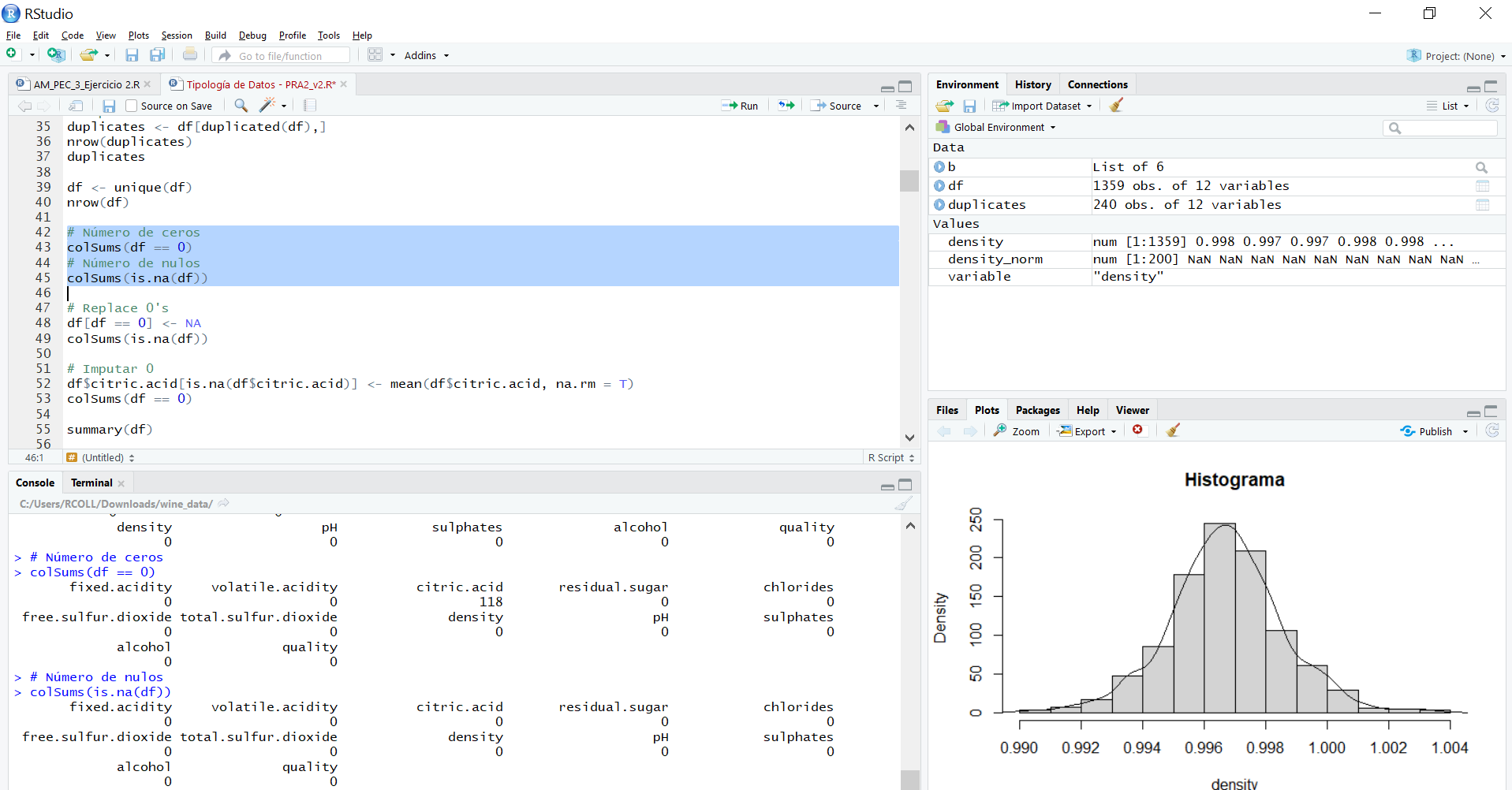


Si bien es cierto que dichos registros con los mismos valores podrían pertenecer a diferentes vinos y por tanto no ser en realidad duplicados, al no tener más información sobre el vino calificado como podría ser un identificador o el nombre del mismo, hemos optado por considerar que son duplicados y hemos procedido a la eliminación de los mismos quedando el dataset con un total de 1.359 registros, a priori, válidos.



* 1. **Gestión de valores nulos**

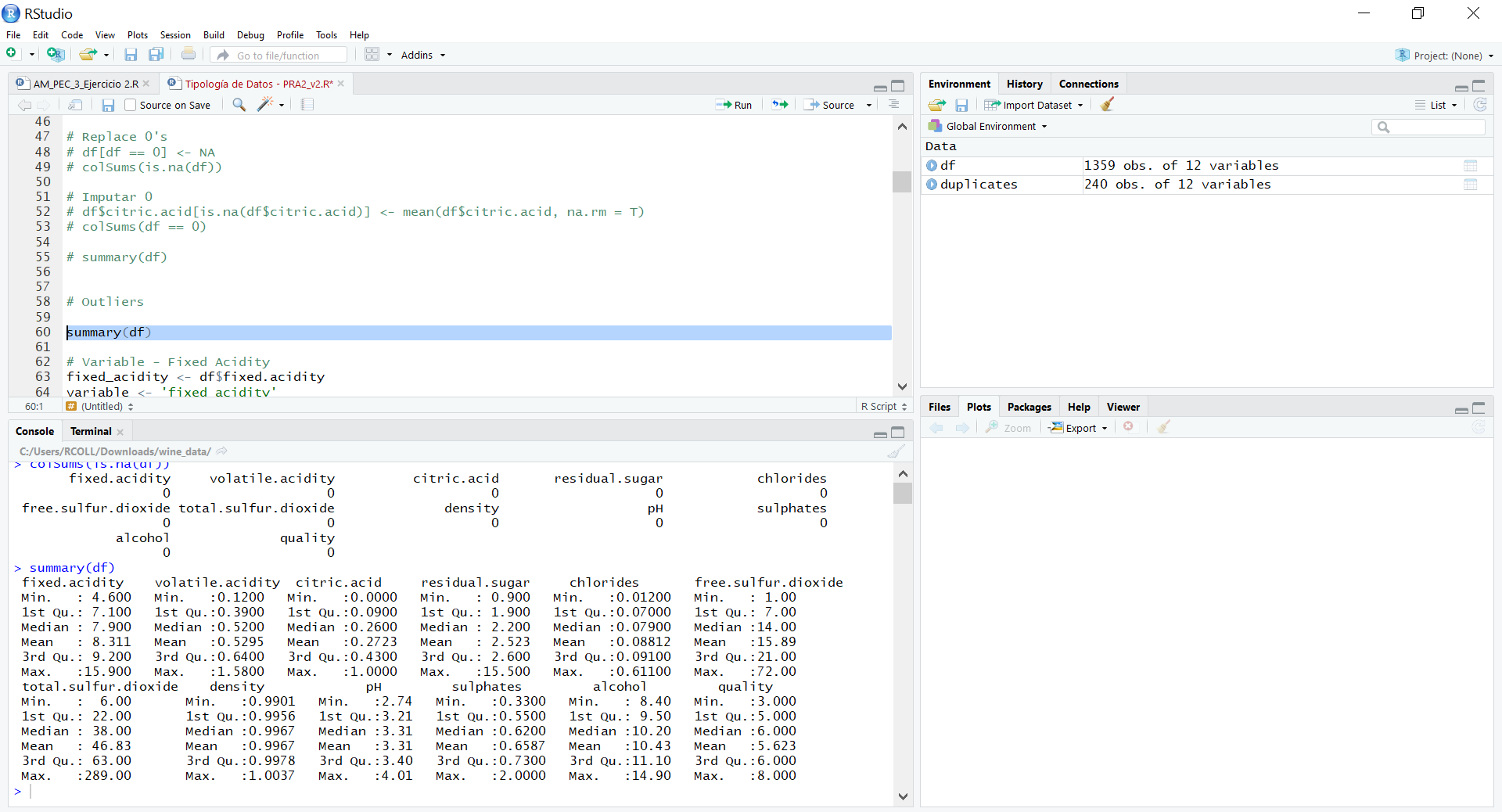
Una vez verificada la existencia de valores duplicados, procedemos a analizar la existencia de valores nulos. Para dicho análisis hemos considerado como valores nulos aquellos no imputados o con valores como puede ser NA, NA’s, NULL o 0 dado que todas las características de las que disponemos son numéricas.



Como observamos en la captura de pantalla anterior, si bien no se han encontrado valores nulos sí que existen 118 registros que tienen un valor de cero para el campo “citric.acid”. A pesar de que en según en qué circunstancias se podrían considerar dichos registros como erróneos siempre hay que contextualizar el análisis. En este caso hemos estado recabando información acerca del rango de valores de ácido cítrico en los que se suelen mover los vinos tintos y parece ser que el cero es un valor aceptado, por este motivo decidimos no eliminar o imputar dichos registros.

* 1. **Gestión de valores extremos**

El último paso que realizaremos en la limpieza de los datos será el análisis de los valores extremos. Para ello vamos a proceder a extraer los principales estadísticos de cada una de las variables y a graficar los valores de las mismas lo que nos permitirá detectar la existencia de valores atípicos en cada uno de los atributos.



* fixed.acidity

Atendiendo a los principales estadísticos de la variable que representa el ácido tartárico del vino, observamos como, en el boxplot, hay una serie de observaciones que estadísticamente se podrían considerar “outliers” dado que superan en 1,5 veces el rango intercuartílico. En concreto se trata de 42 observaciones que superan el “bigote” superior del diagrama de caja situado en un valor de 12,3.

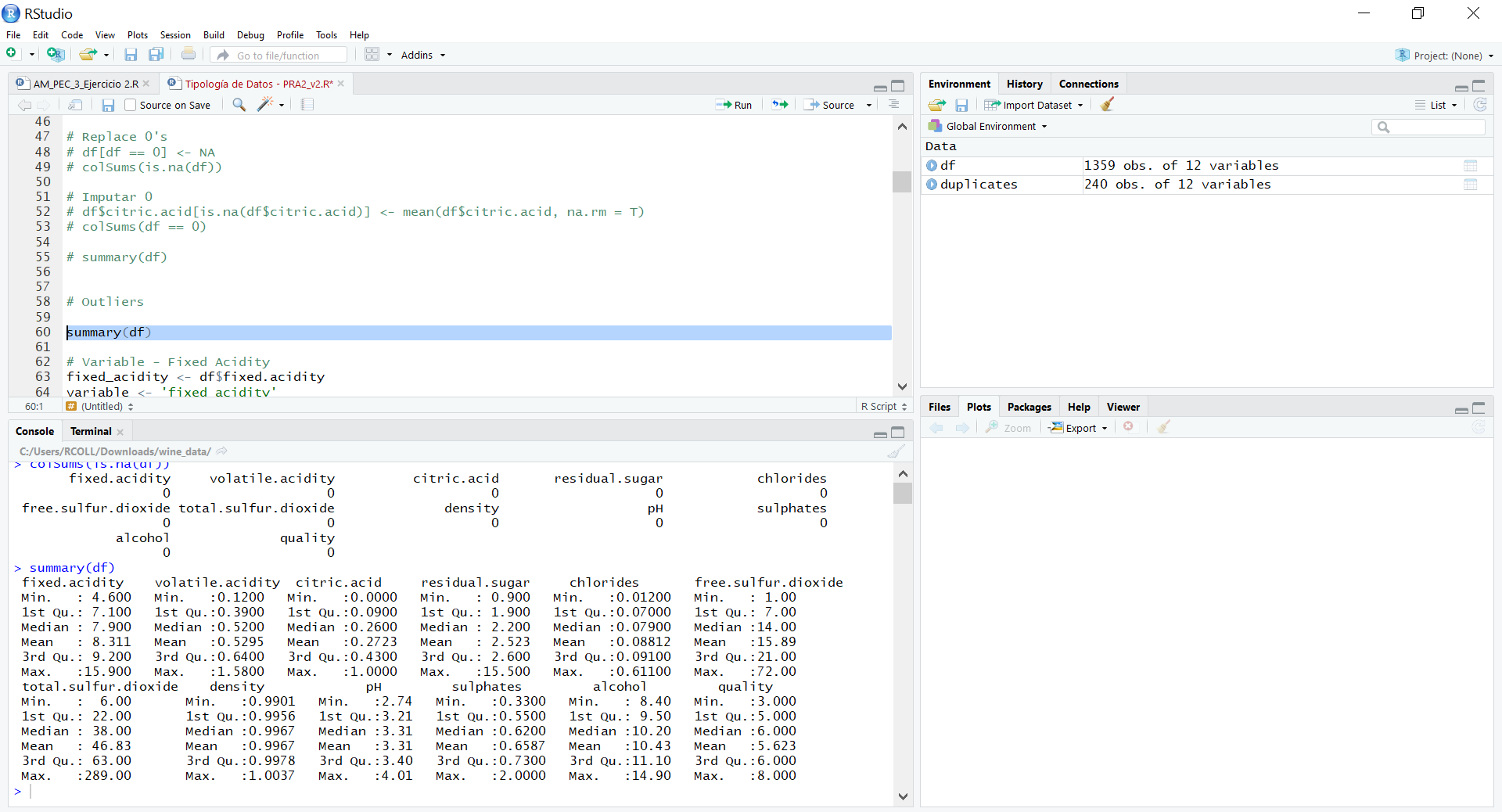
|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

A pesar de que estadísticamente podríamos clasificarlos como valores atípicos, están dentro del rango funcional de la variable y por tanto, decidimos no vamos a excluir dichas observaciones.

* volatile.acidity

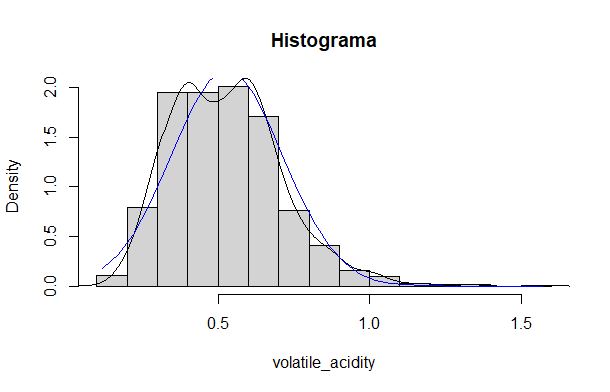
Si analizamos los principales estadísticos de la variable que representa el ácido acético del vino, observamos que se mueve en un rango de valores que van entre 0,12 y 1,58 situándose la media de las observaciones en 0,5295 ; muy próxima a la mediana que está en 0,52.



En el diagrama de caja siguiente observamos como hay un total de 20 observaciones que quedan por encima del valor 1,01 que representa el bigote superior de dicho gráfico y que por tanto podrían considerarse valores atípicos estadísticamente hablando.

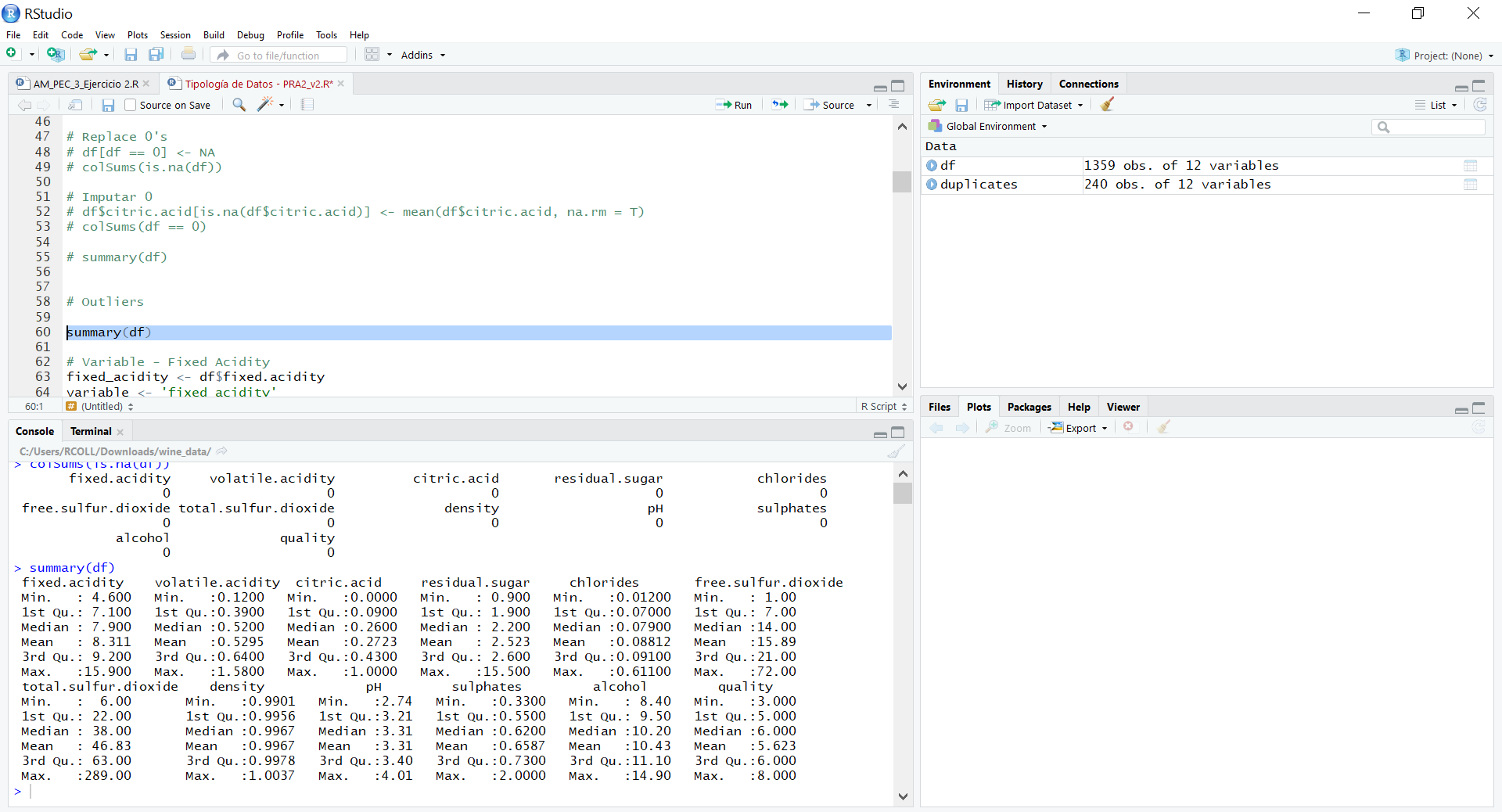
|  |  |
| --- | --- |
|  |  |

Por otro lado, si representamos el número de observaciones existentes para cada rango de valores de la variable “volatile\_acidity” observamos como no sigue exactamente una distribución normal existiendo una cierta asimetría hacía la derecha que queda representada por la cola que se puede ver en el siguiente gráfico.



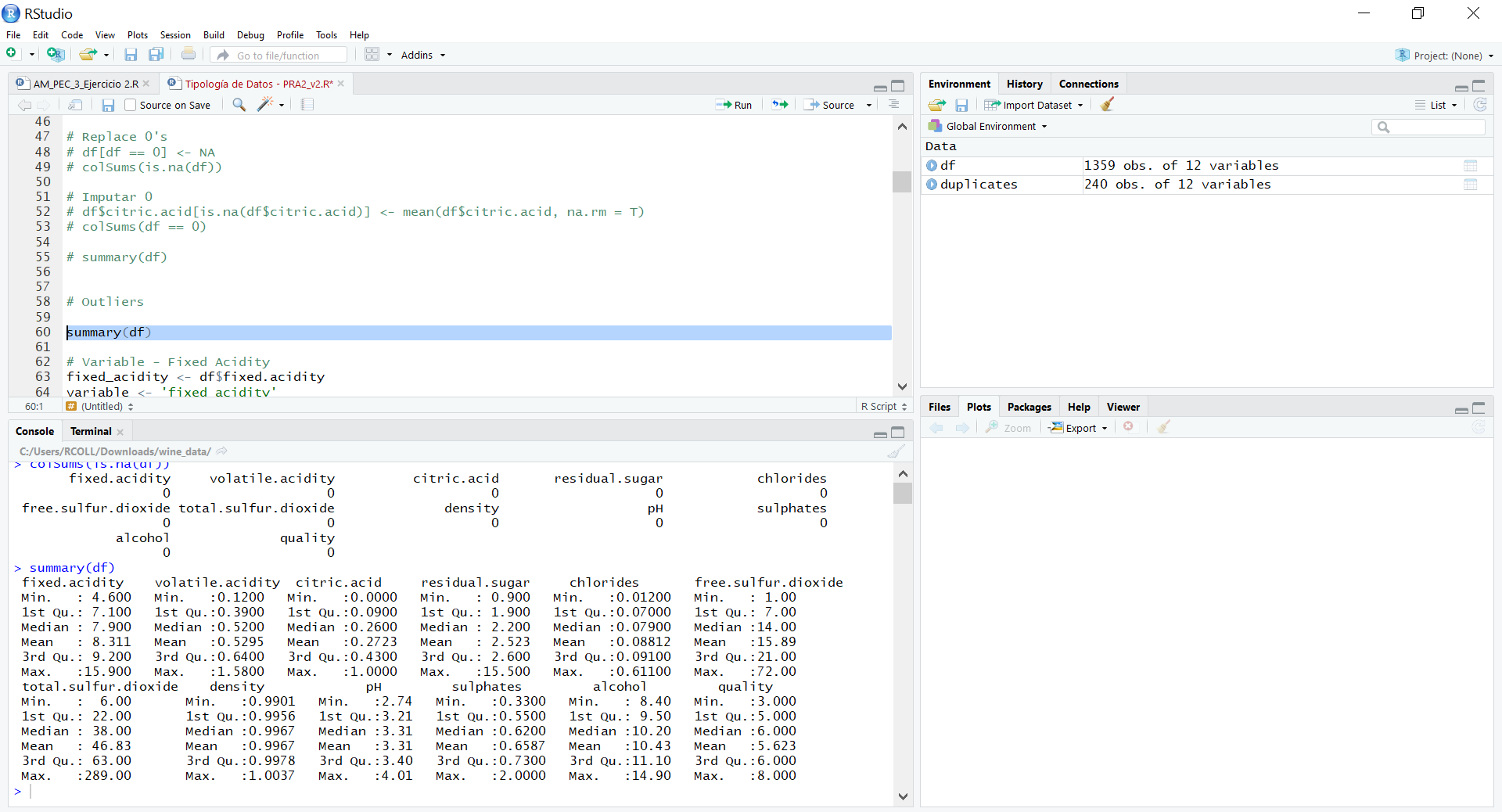
Al contrario de lo que pasaba con la primera variable analizada en la que a pesar de existir valores estadísticamente atípicos, estaban dentro del rango de valores posibles de la variable considerada, en este caso dicho rango está entre los 0,3 y 0,7([[1]](#footnote-1) por lo que procedemos a excluir dichos registros.

* citric.acid



* residual.sugar

La variable “resigual.sugar” indica la cantidad de azúcar que queda en el vino una vez finalizado el proceso de fermentación. Como vemos en el resumen estadístico de las principales medidas, las observaciones del conjunto de datos se mueven en un rango que va entre 0,9 y 15,5 gramos de azúcar por litro estando la media en 2,2 . Atendiendo a la clasificación de los vinos según la cantidad de azúcar[[2]](#footnote-2), estos pueden ir desde los 1 en el caso de los más secos hasta más de 200 de los clasificados como dulces.

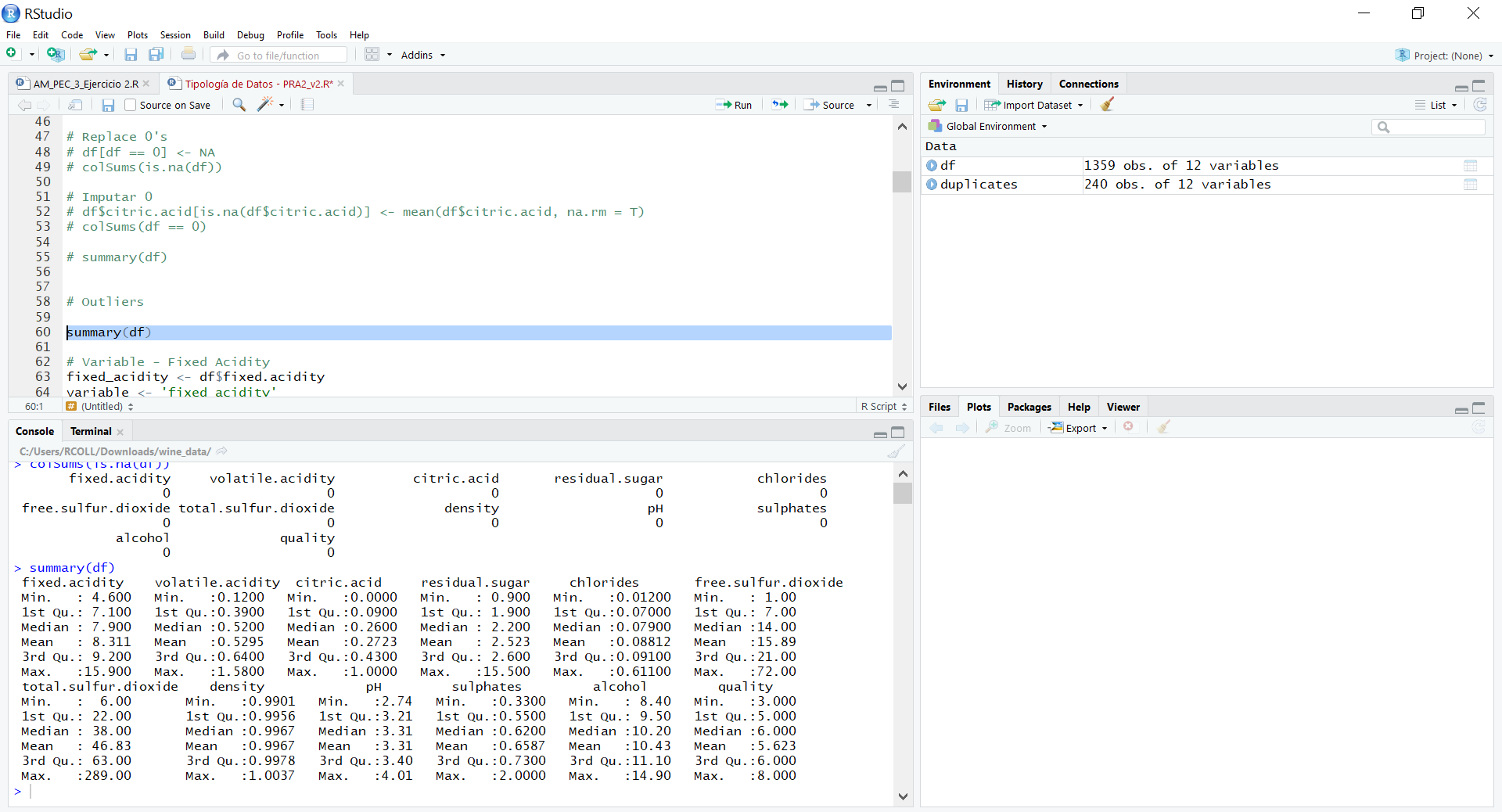


A continuación podemos ver la distribución de las observaciones, en las que destaca principalmente la asimetría que presenta el histograma.

|  |  |
| --- | --- |
|  |  |

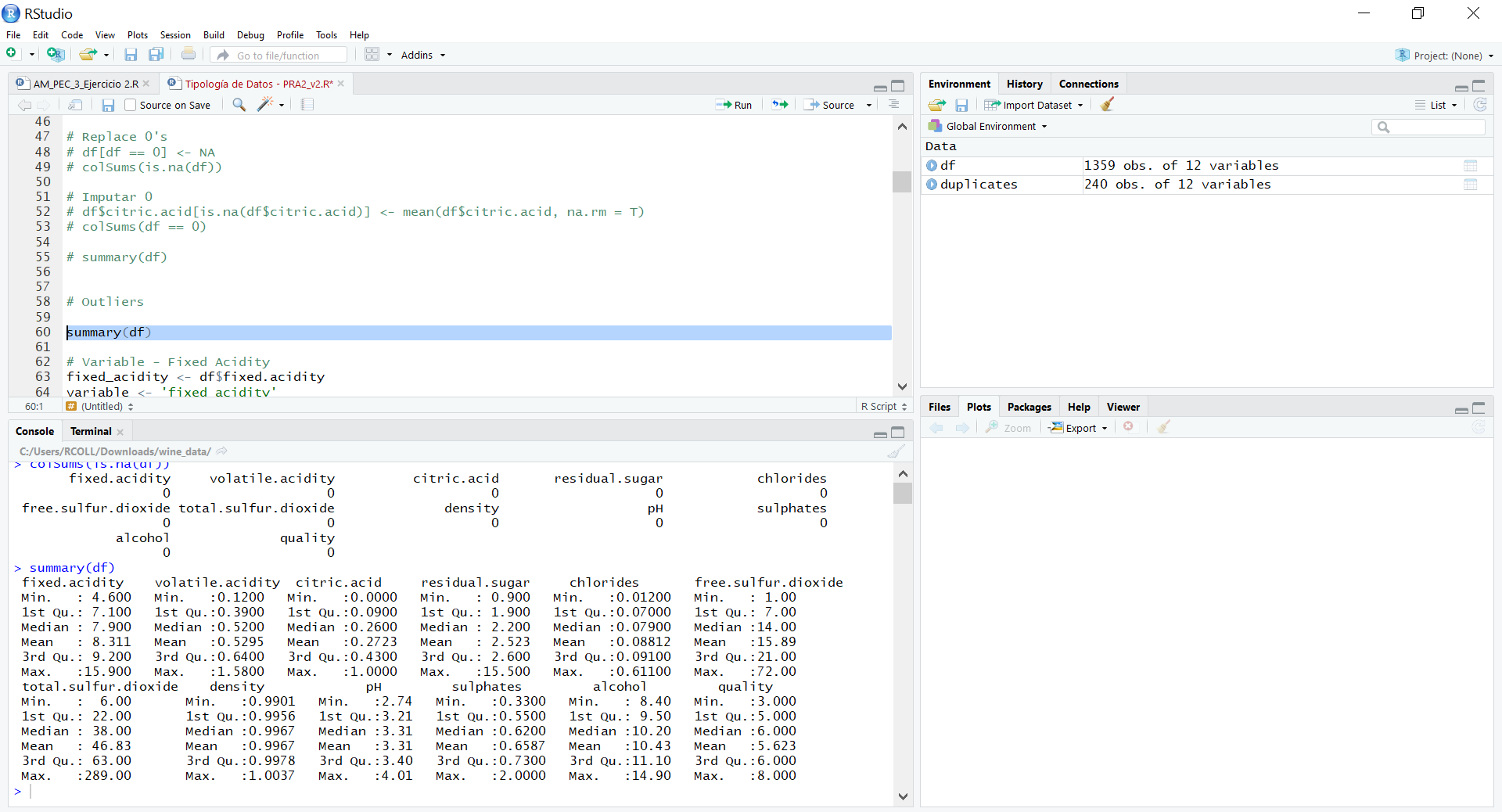
A pesar de dicha simetría y del número de observaciones que están por encima del bigote superior en el diagrama de caja, en principio todos los valores parecen estar en el rango que puede tomar la variable desde un punto de vista funcional y, por lo tanto, no vamos a excluirlos del análisis.

* chlorides

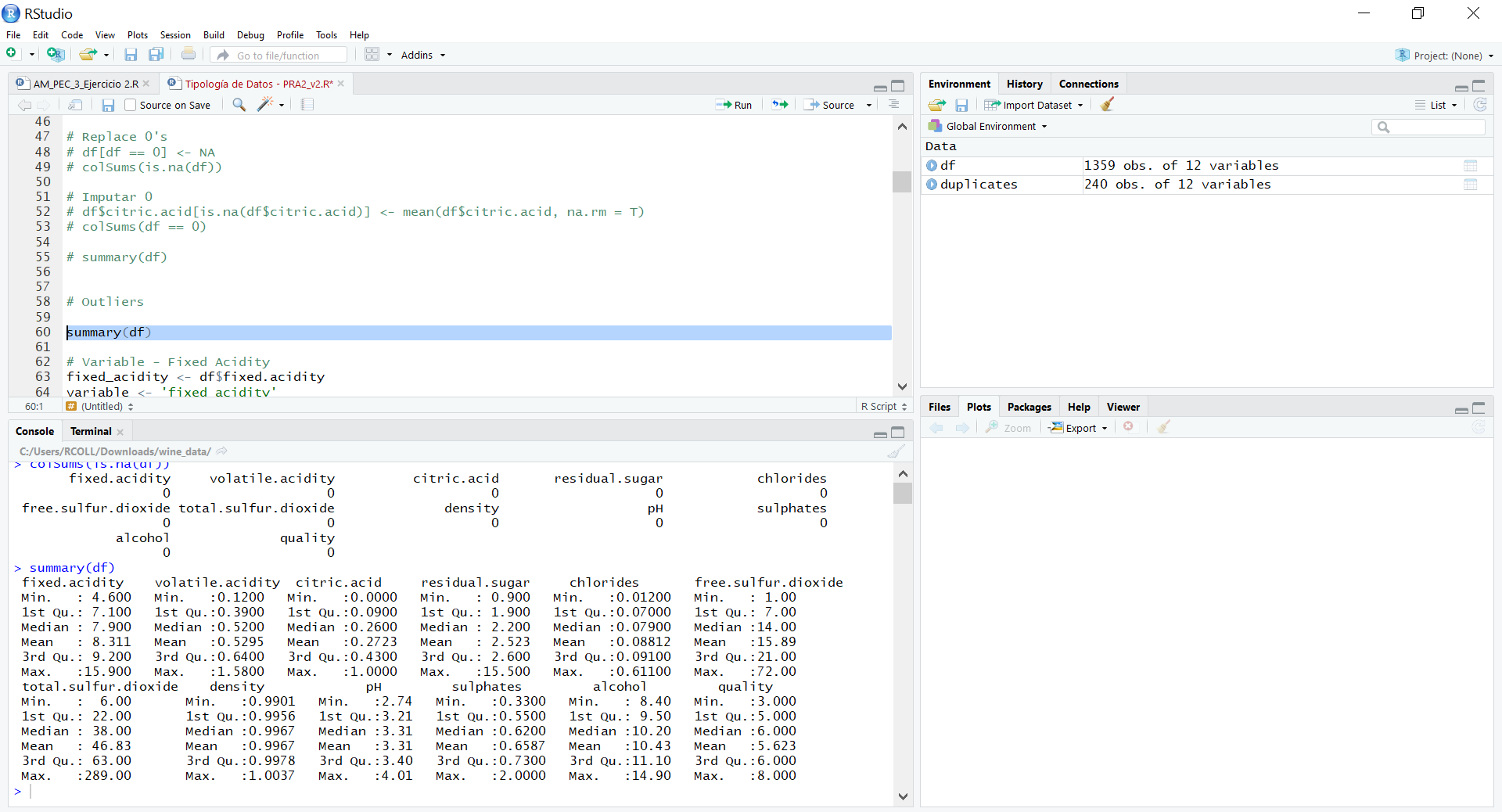


https://www.wineaustralia.com/labelling/wine-production-standard

* free.sulfur.dioxide

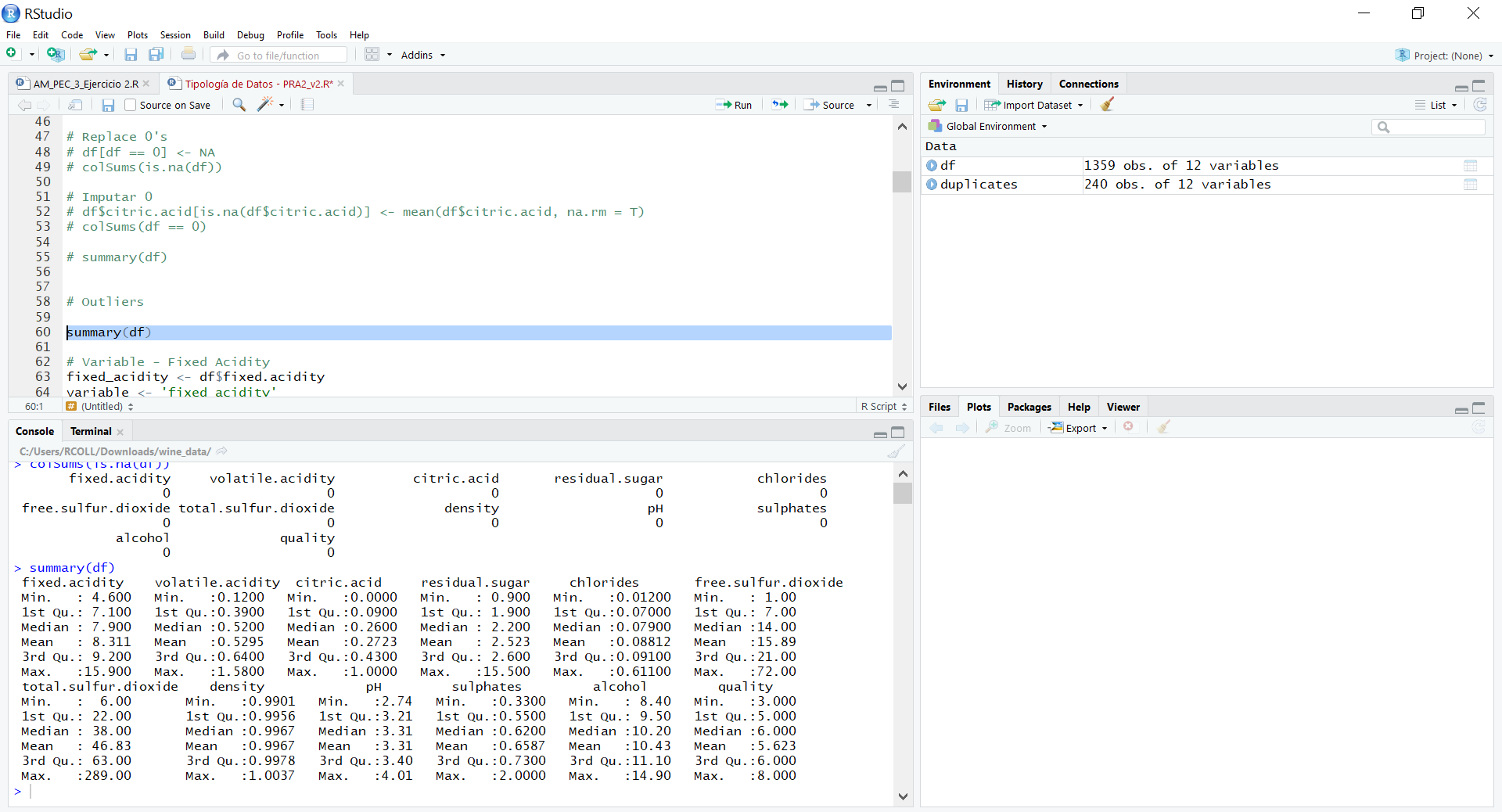


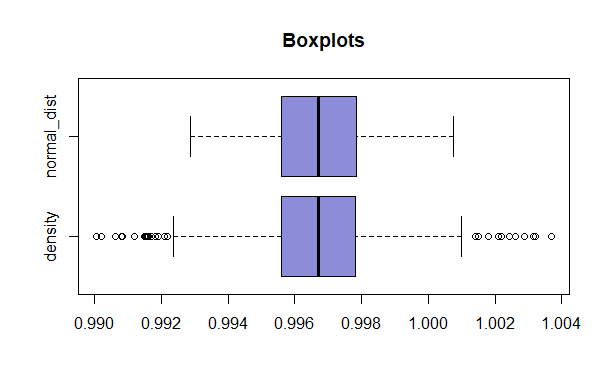
* total.sulfur.dioxide

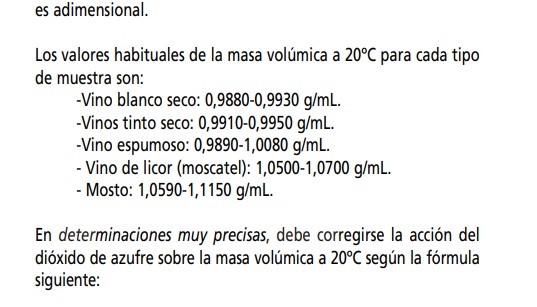


<https://www.vinetur.com/2019012549165/que-es-y-por-que-no-debes-comprar-vino-con-elevado-sulfuroso-combinado.html>

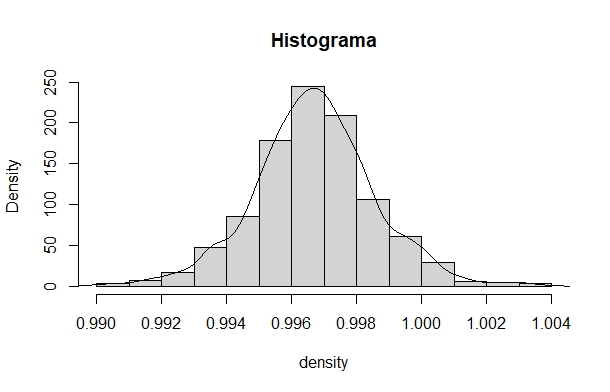
* density

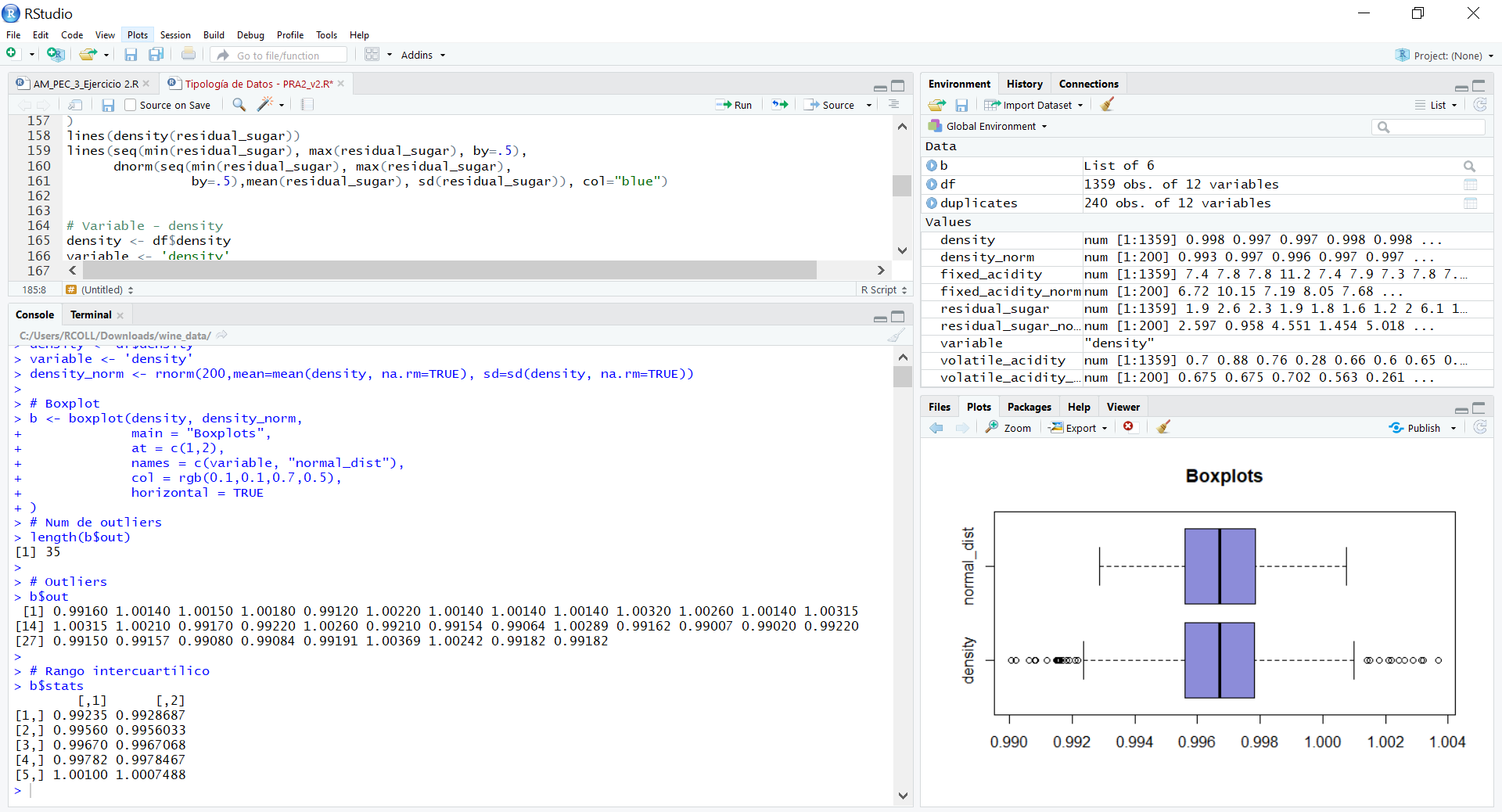






http://www.usc.es/caa/MetAnalisisStgo1/enologia.pdf

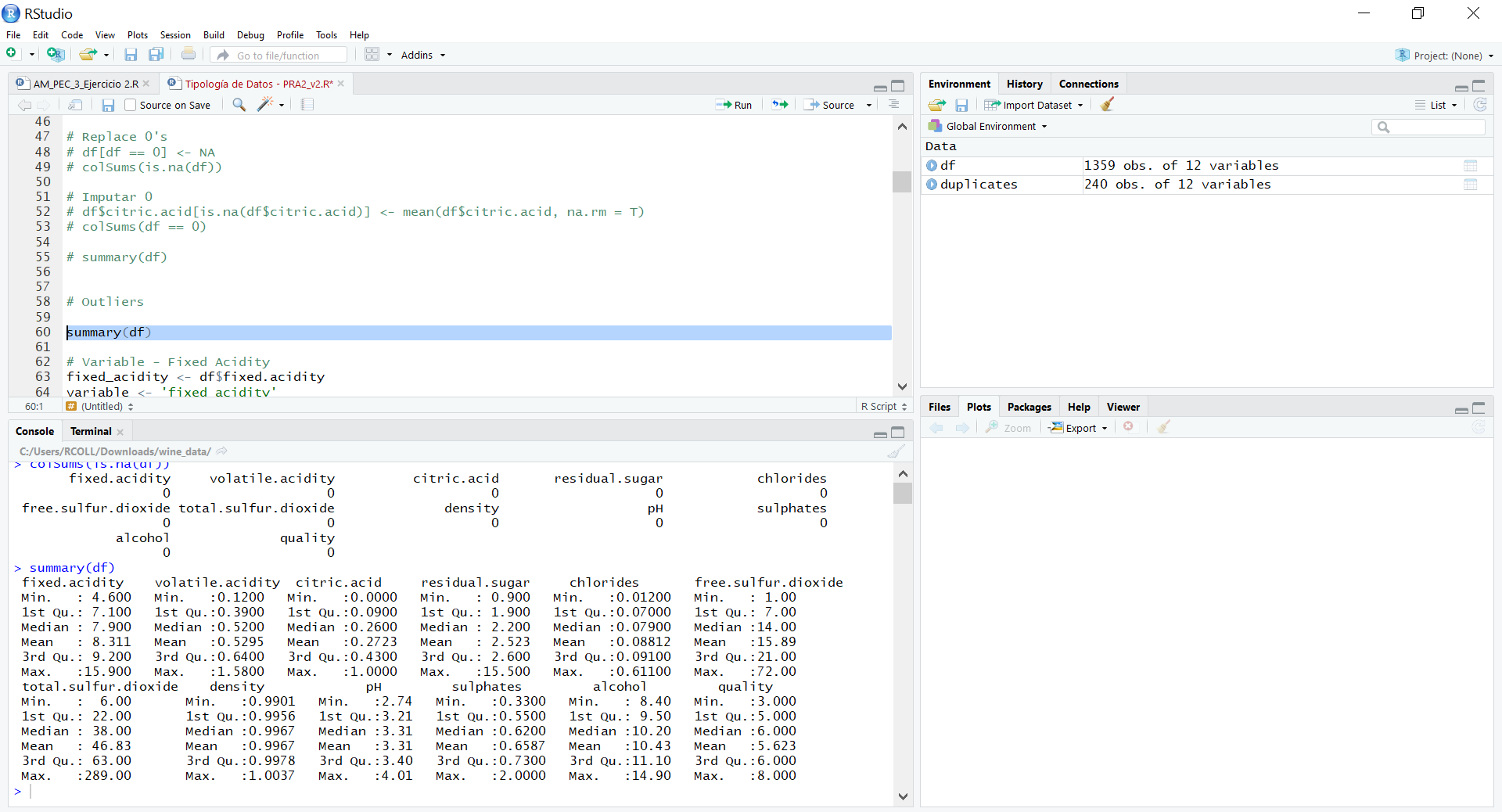




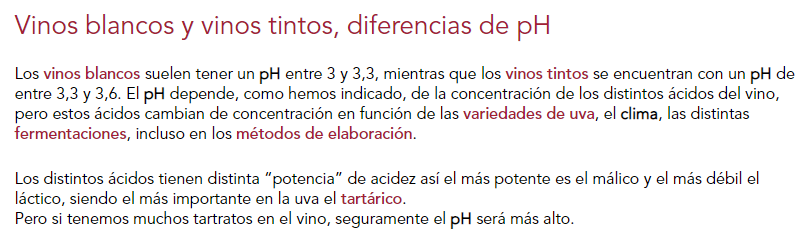
* pH

Es pH es una unidad de medida de la [acidez](https://catatu.es/blog/acidez-vino/)o alcalinidad de una disolución. Se mide en una escala del 0 al 14 siendo 14 muy básico o alcalino, y los pH cercanos a cero son disoluciones muy ácidas. El pH del vino suele oscilar entre 2,8 y 4, siendo 2,8 un vino extremadamente ácido y un vino con pH 4 es un vino plano sin acidez.[[3]](#footnote-3)

Como vemos a continuación, el pH de nuestra muestra oscila entre un valor de 2,74 y de 4,01 estando en ambos casos justo por debajo y por encima, respectivamente de lo que en teoría se considera normal para el caso del vino.



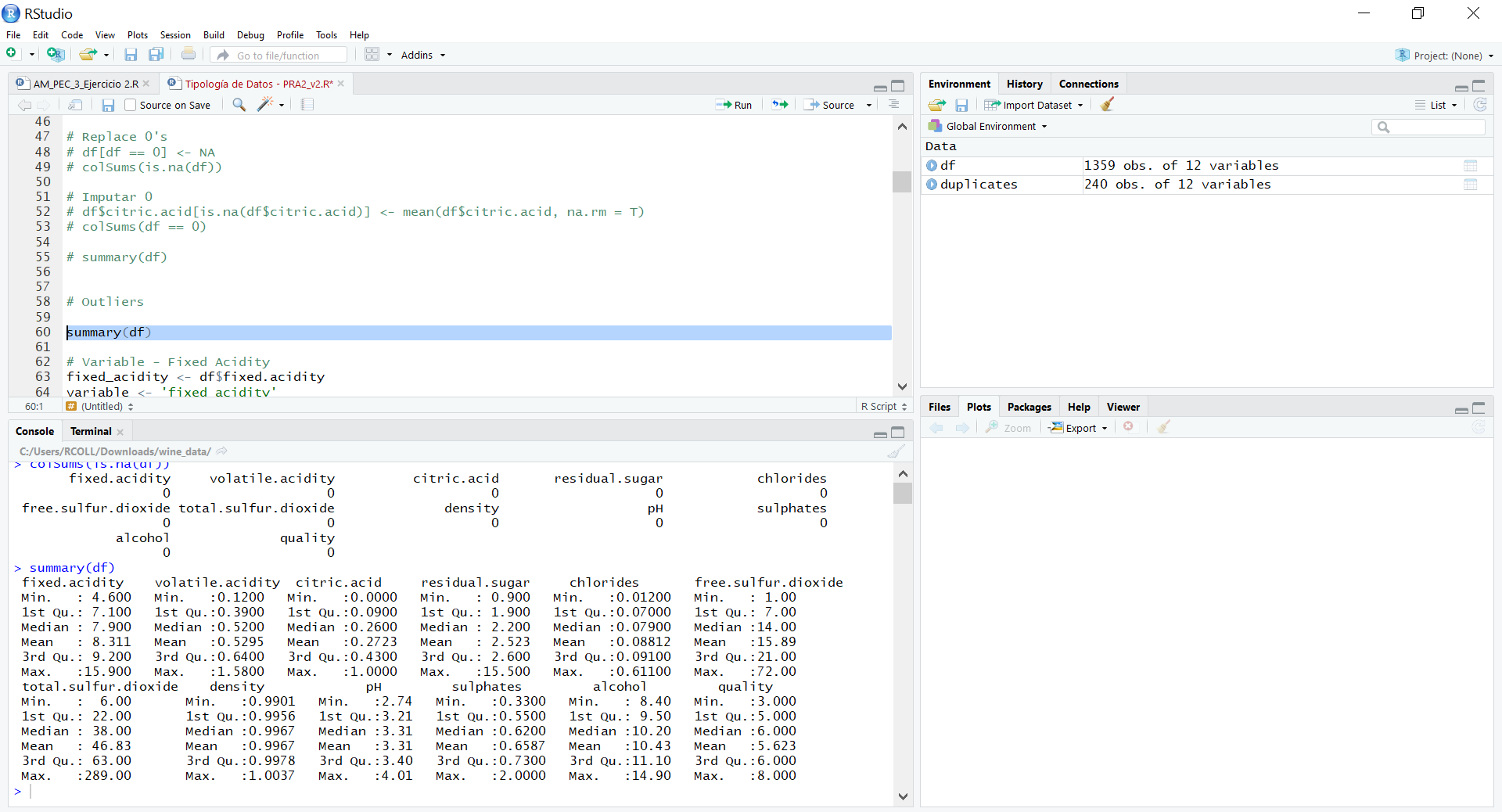
Dado que la diferencia es menor, no excluiremos ninguna de las observaciones que rompen dichos límites dado que también nos interesa comprobar si dichos vinos que no cumplen con lo habitual tienen notas más bajas debido a ello.



* sulphates

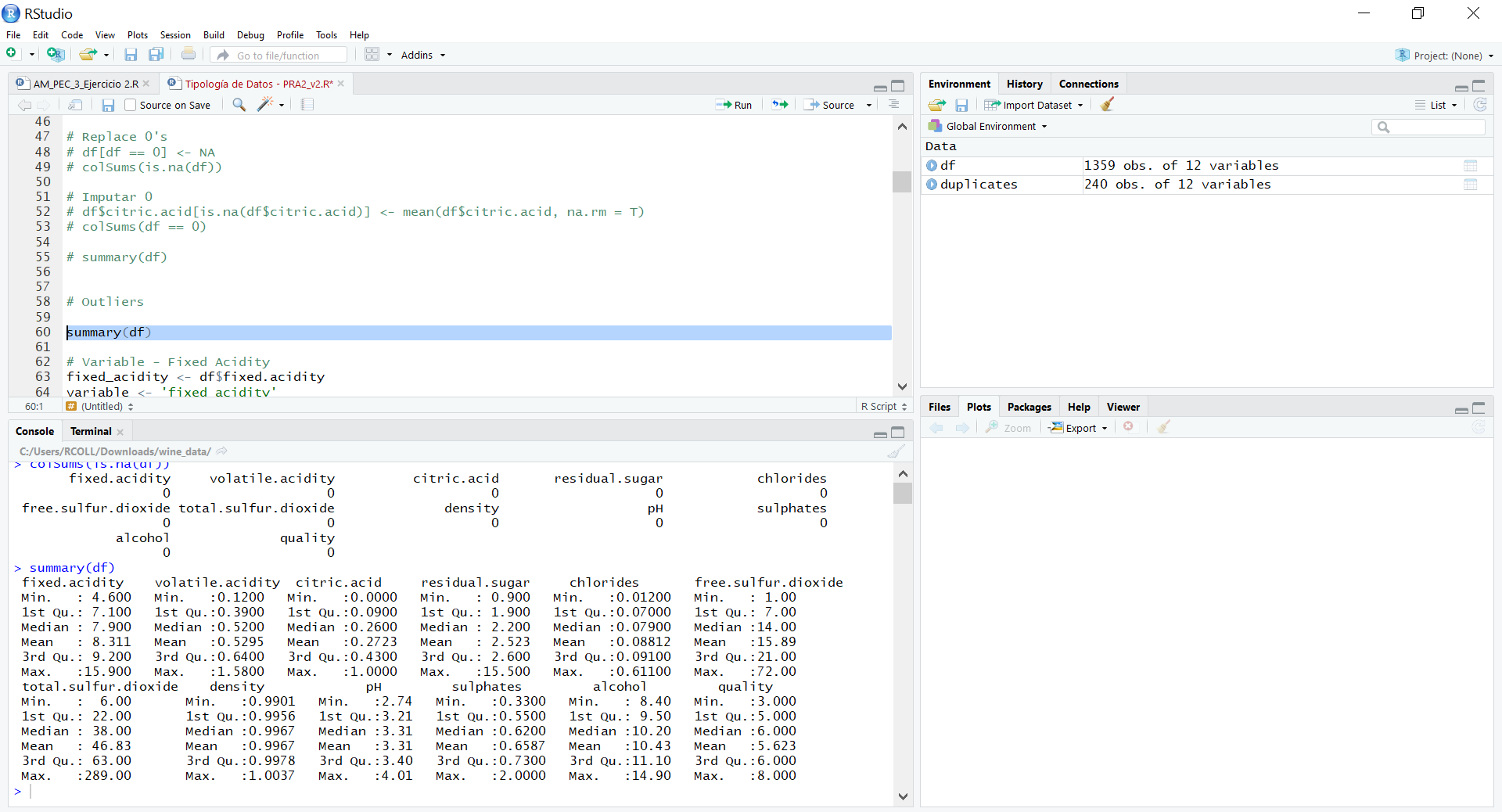
Para determinar los valores habituales de sulfatos que un vino puede tener nos apoyaremos en la regulación española que, mediante el reglamento de la Ley nº18.455 que fija las normas sobre producción, elaboración, comercialización de alcoholes etílicos, bebidas alcohólicas y vinagres establece unos valores máximos de .

Si bien desconocemos a que regulación están sujetos los vinos incluidos en el conjunto de datos del que disponemos, vemos como se mueven en un rango que va desde los 0,33 hasta los 2,00 y por tanto, consideraremos que no existen valores atípicos en lo que a esta variable se refiere.



* alcohol

La variable “alcohol” representa el % de alcohol presente en el vino. Si observamos los principales estadísticos de este variable vemos como el rango de valores oscila entre un mínimo de 8,40% y un máximo de 14,90%.

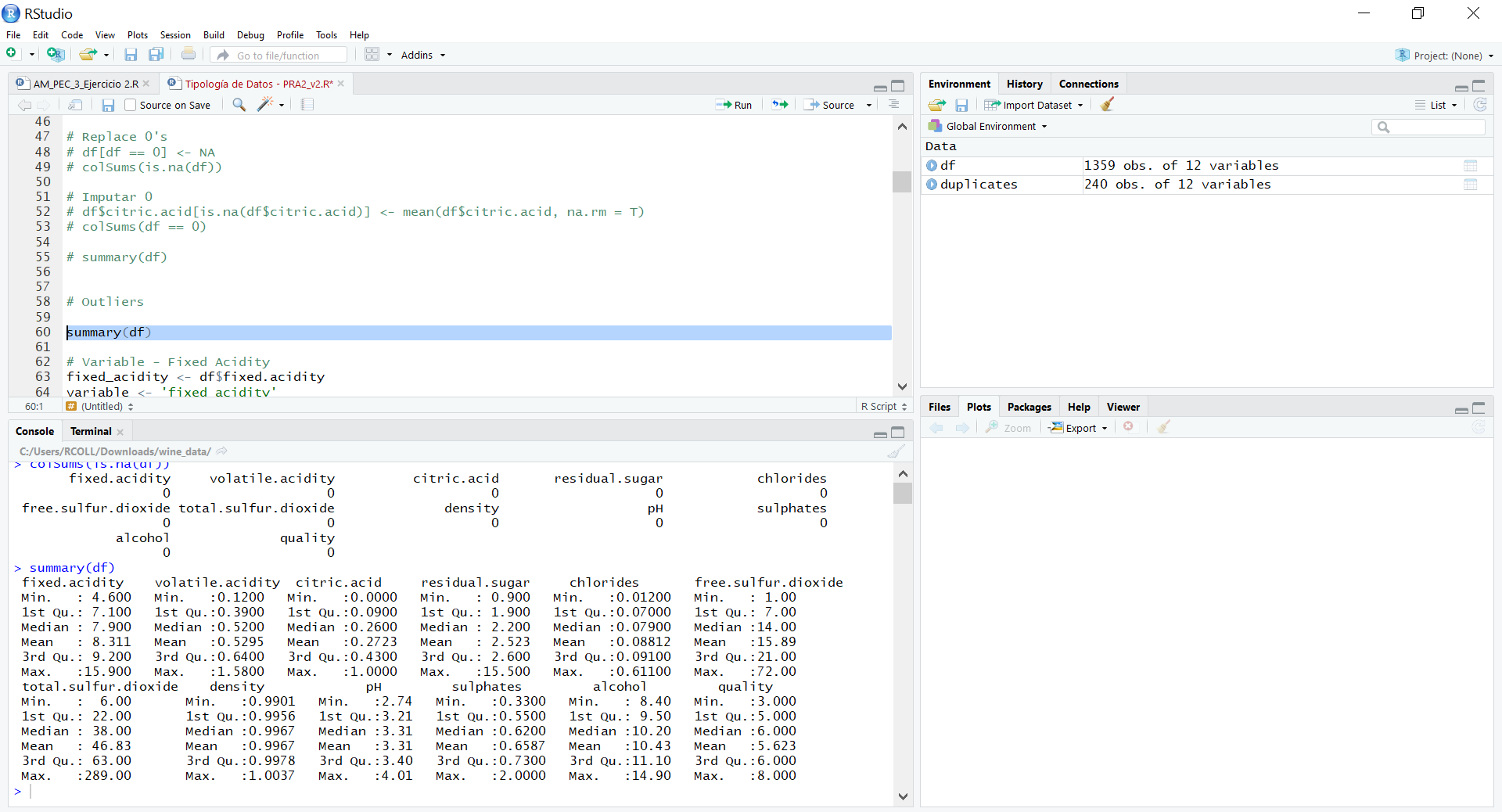


Después de haber analizado diferentes fuentes de información, hemos llegado a la conclusión que los valores de dicha variable están dentro de la graduación alcohólica habitual que va desde un 7% a un 16%[[4]](#footnote-4) para un vino tinto y que por tanto no existen valores atípicos.

|  |  |
| --- | --- |
|  |  |

* quality

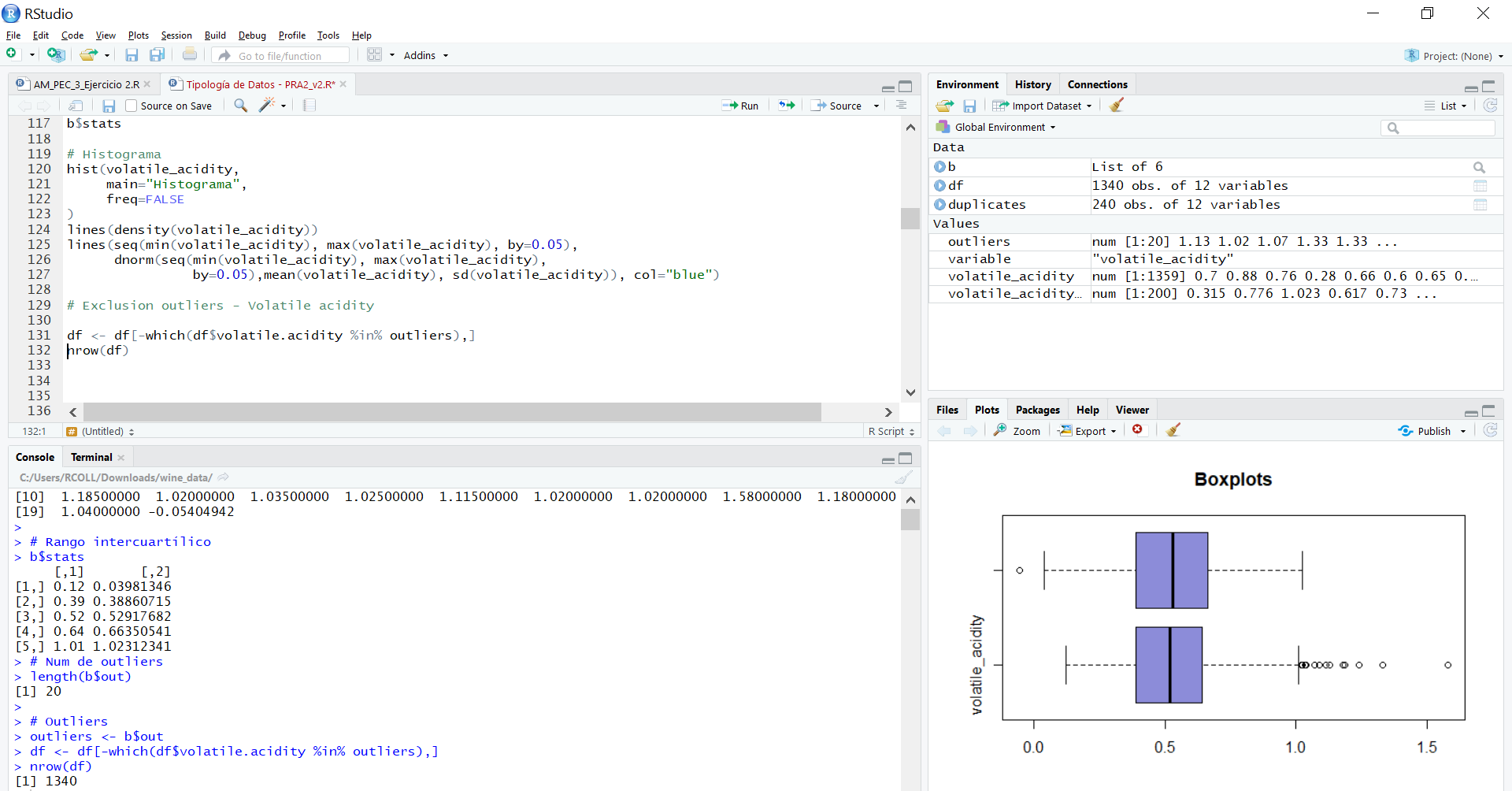
La variable “quality” indica la calidad del vino puntuándolo en una escala de 1 a 10 siendo está última la mayor puntuación posible. En el caso de nuestro conjunto de datos, se mueve en el rango comprendido entre una puntuación mínima de 3 y una máxima de 8, estando la media situada en 5,623.



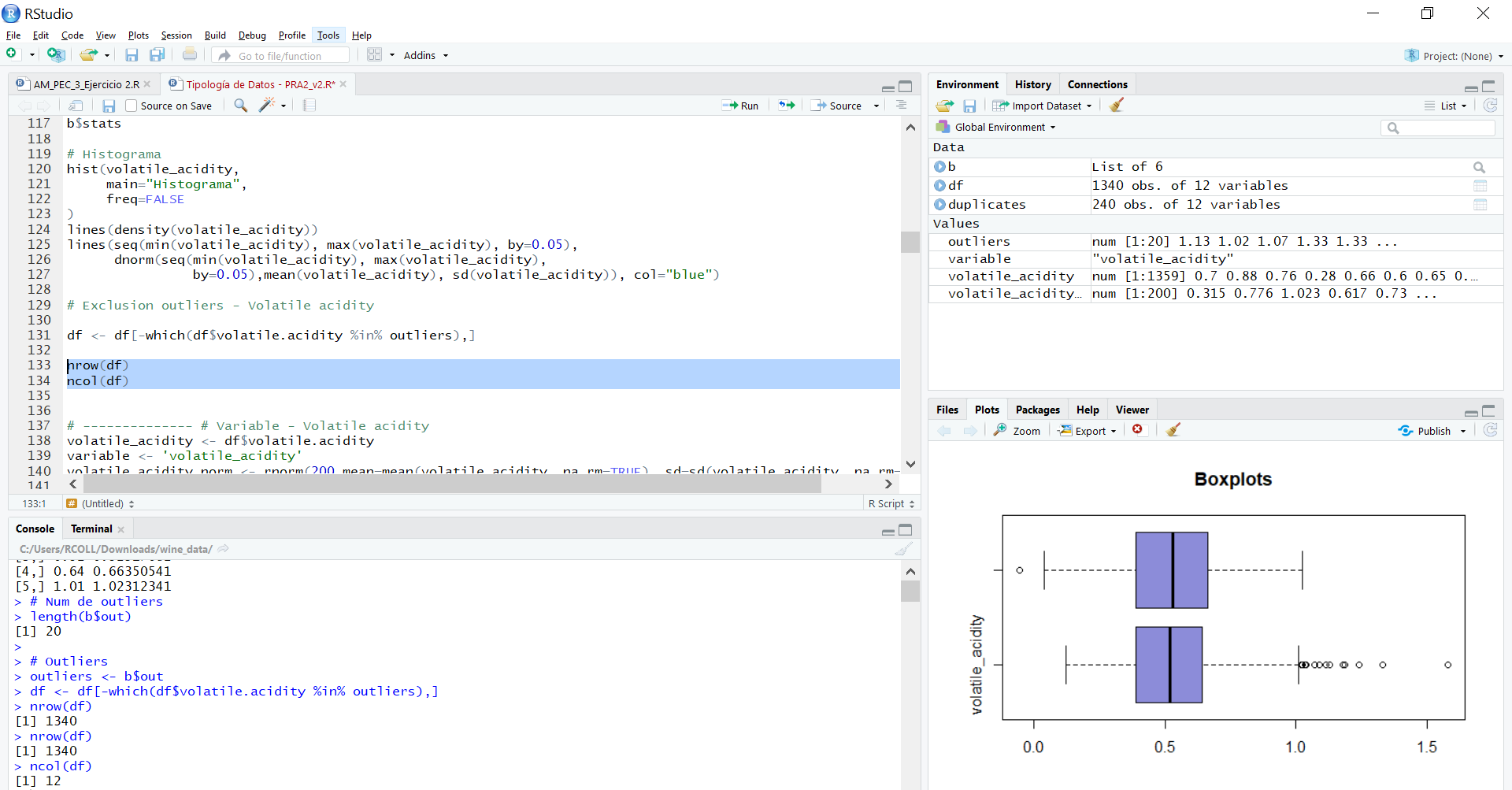
Dado que no hay valores fuera del domino de la variable no excluiremos ningún registro.

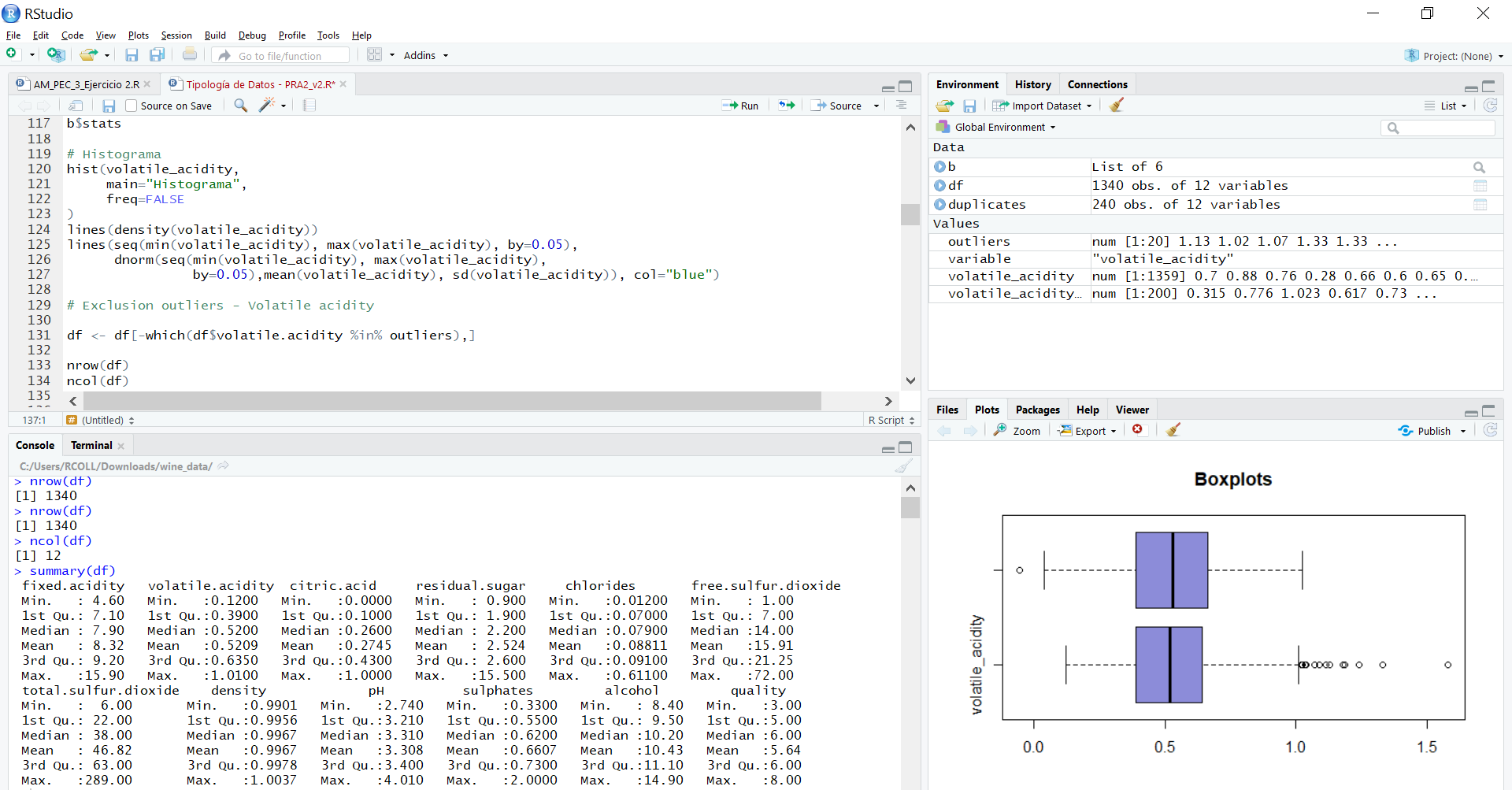
|  |  |
| --- | --- |
|  |  |

Una vez realizado el análisis de las principales características de las que disponemos, procedemos a la exclusión de aquellas observaciones que por algún motivo hemos considerado como atípicas.



Después de la exclusión de dichas observaciones el conjunto de datos queda configurado con un total de 1340 registros y 12 columnas siendo el resumen de los principales estadísticos el siguiente:



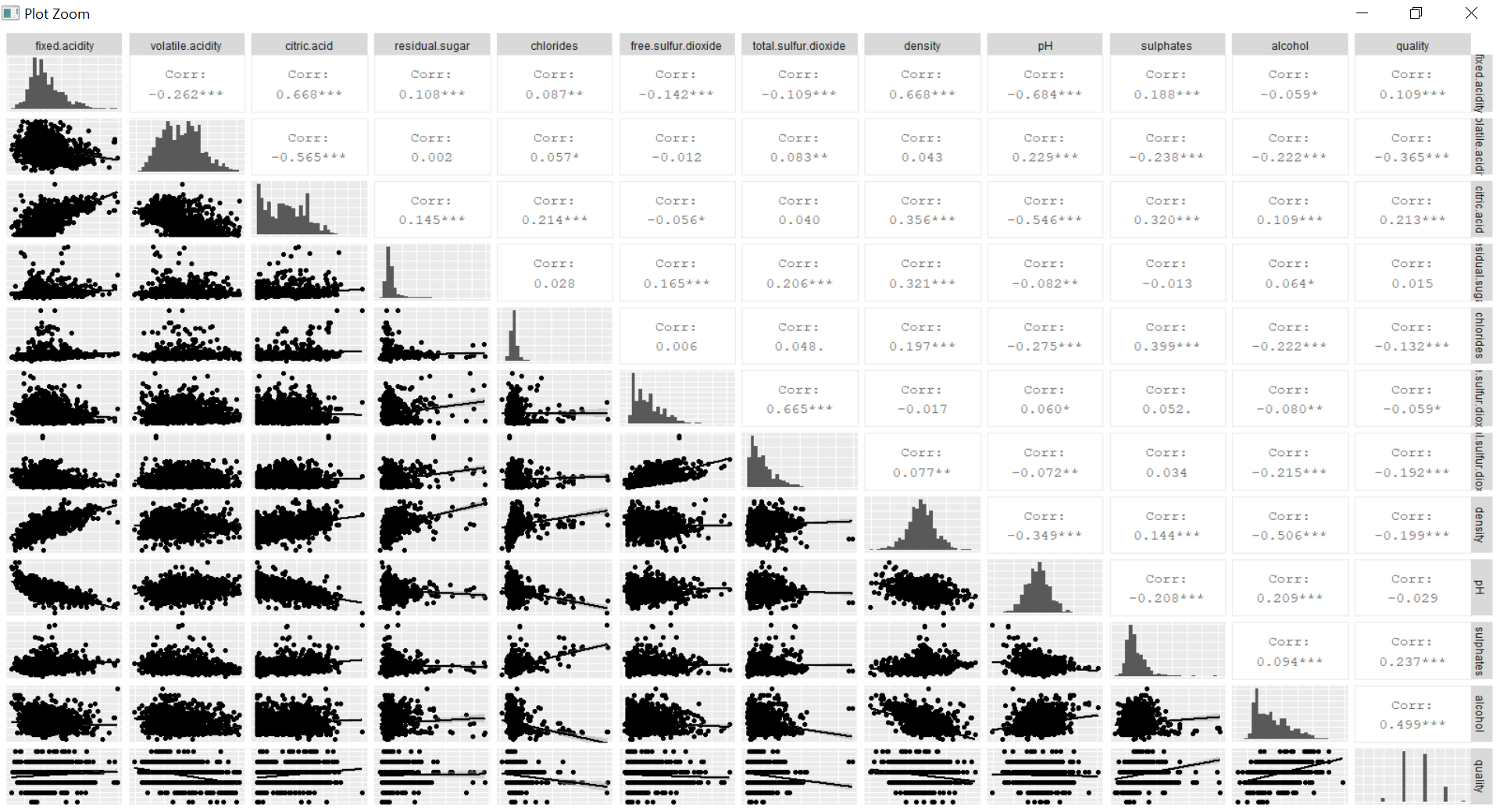


## Análisis de los datos

Para la creación de un modelo que explique la calificación del vino contenida en la variable “quality” a partir de una serie de variables dadas, vamos a aplicar la técnica de análisis multivariante de dependencia conocida como regresión lineal múltiple. Esta técnica se basa en la estimación del peso que, cada una de las variables (independientes), tiene en la explicación lineal de la variable dependiente “quality”.

* 1. **Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).**

Requisito fundamental antes de proceder con la creación del modelo de regresión, es la valoración de la existencia de relación y el grado de la misma entre cada una de las variables independientes y la variable dependiente. Para ello la matriz de correlaciones y el análisis visual de la distribución de los datos nos ayudará a detectar aquellas variables que de manera bivariante pueden tener un mayor impacto en la calidad del vino y el tipo de relación de las mismas (lineal, logarítmica, exponencial, etc).



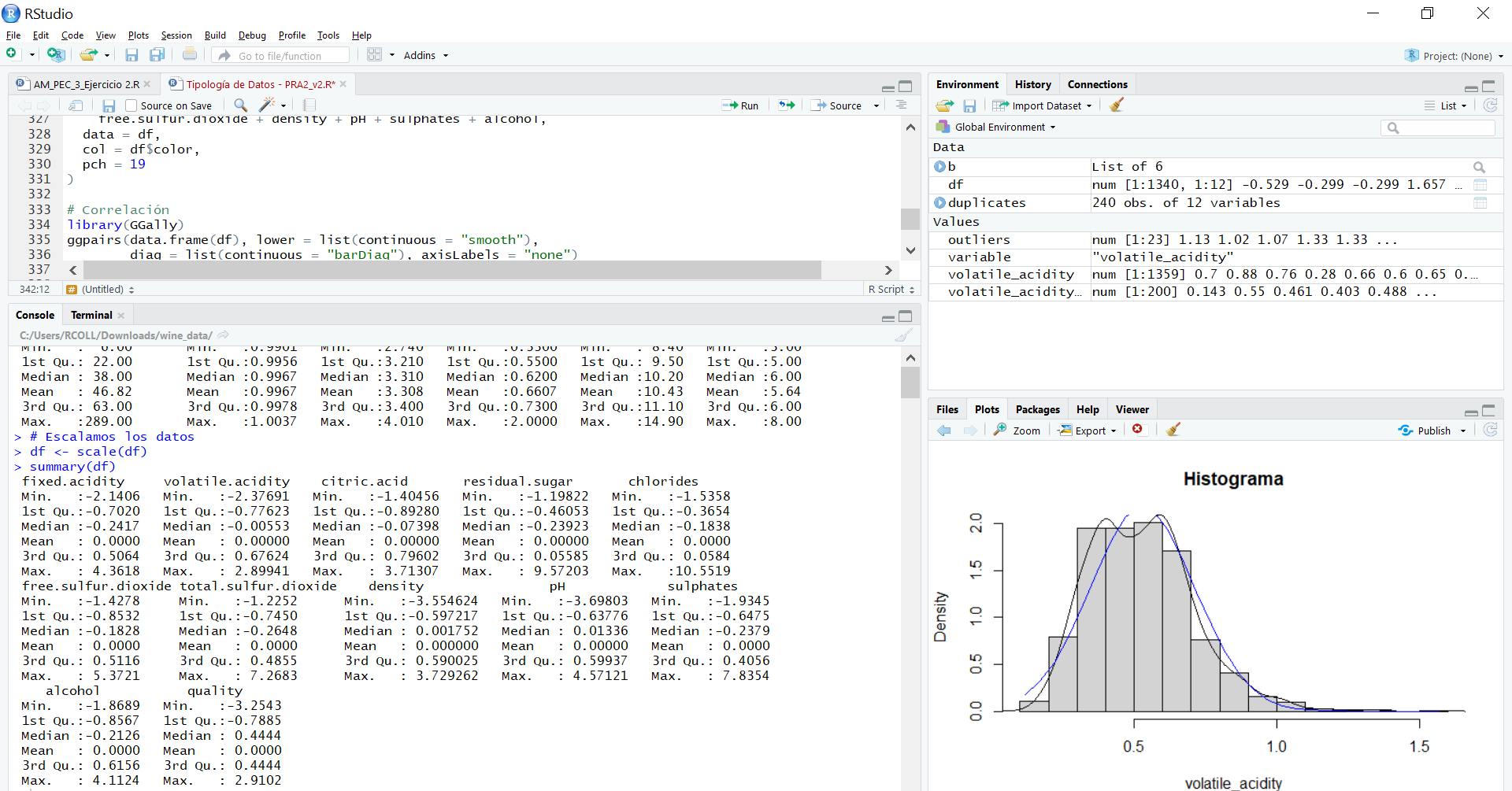
Como vemos en la captura de pantalla anterior, en general podemos decir que todas las variables independientes de las que disponemos y utilizaremos para el análisis tienen una correlación baja con la variable dependiente “quality” algo que puede influir negativamente en la capacidad explicativa del modelo. Recordemos que la correlación es una medida que vive en el rango comprendido entre el -1 y el 1 donde cuan más próximo a uno de esos dos valores está se dice que tiene una correlación elevada (negativa o positiva respectivamente) y cuanto más se acerca a 0 se dice que la correlación es baja indicando una débil relación lineal entre las variables consideradas.

Aparte de las correlaciones de cada una de las variables con la variable dependiente, es importante analizar las relaciones existentes entre las variables independientes dado que correlaciones muy fuertes nos podrían estar indicando que dichas variables explican lo mismo y que por tanto estaríamos aumentando dicho efecto. En este sentido destacan las variables que miden los ácidos del vino como son “citric.acid”, “fixed.acidity”, “volatile.acidity” y “pH” que tienen correlaciones relativamente altas (mayores a 0,5 en valor absoluto) entre ellas.

Por otro lado, también observamos como las variables “total.sulphur.dioxide” y “free.sulphur.dioxide” tienen una correlación de 0,665. Esto es debido a que, como veíamos al inicio la variable “free.sulphur.dioxide” forma parte del cálculo de “total.sulphur.dioxide”.

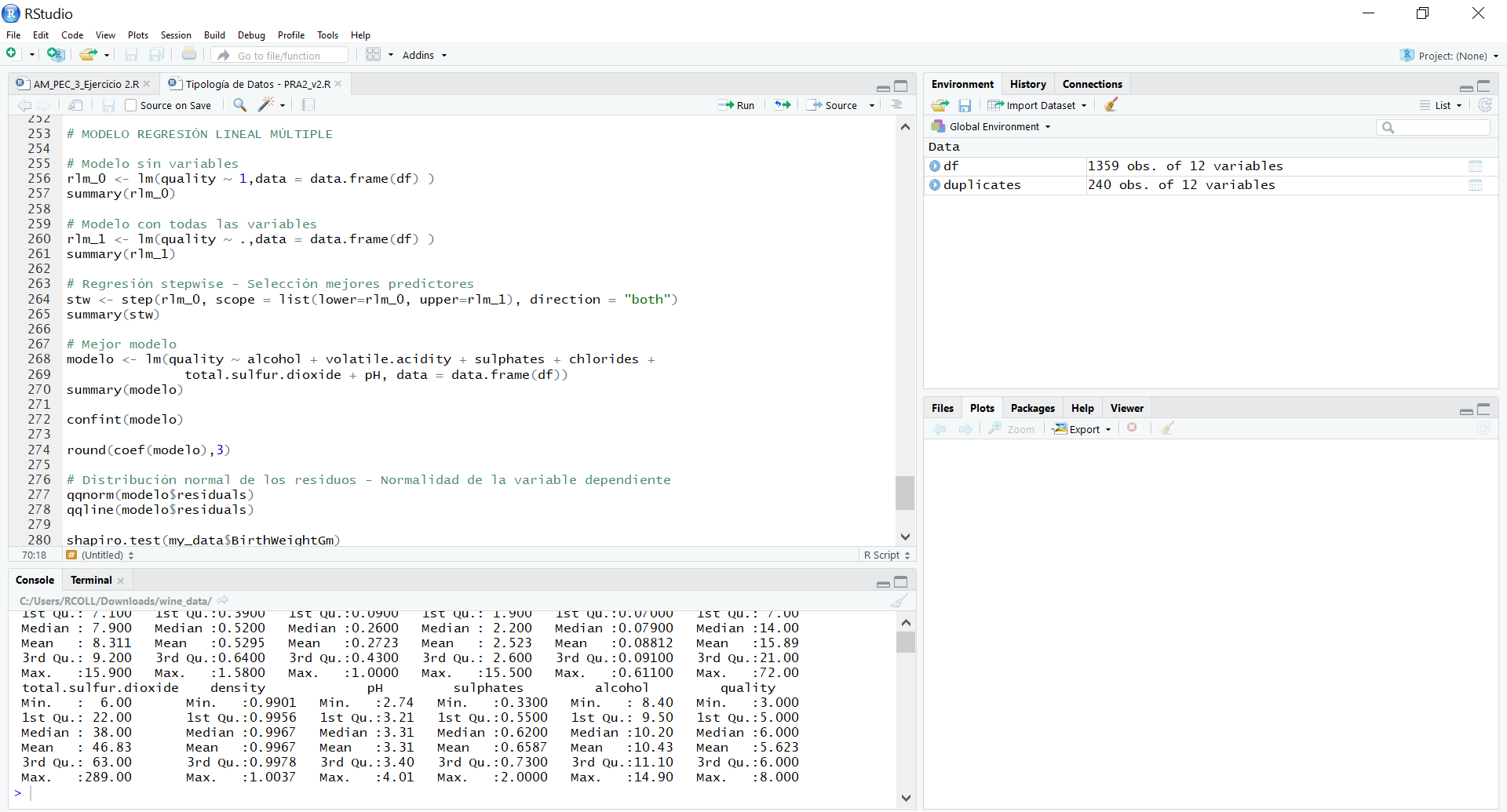
A pesar de los puntos comentados anteriormente y de tener variables con correlaciones relativamente altas, consideramos que dicho efecto no es lo suficientemente elevado como para excluirlas del análisis y por lo tanto procederemos a la elección del mejor modelo partiendo de todas las variables de las que disponemos.

Otro punto importante antes de proceder a realizar el análisis de regresión lineal múltiple es la normalización de los datos. Dado que tenemos variables con diferentes escalas de medida, es importante normalizar los datos de manera que los pesos finales de las variables en el modelo no se vean afectados por dichas diferencias en las escalas.

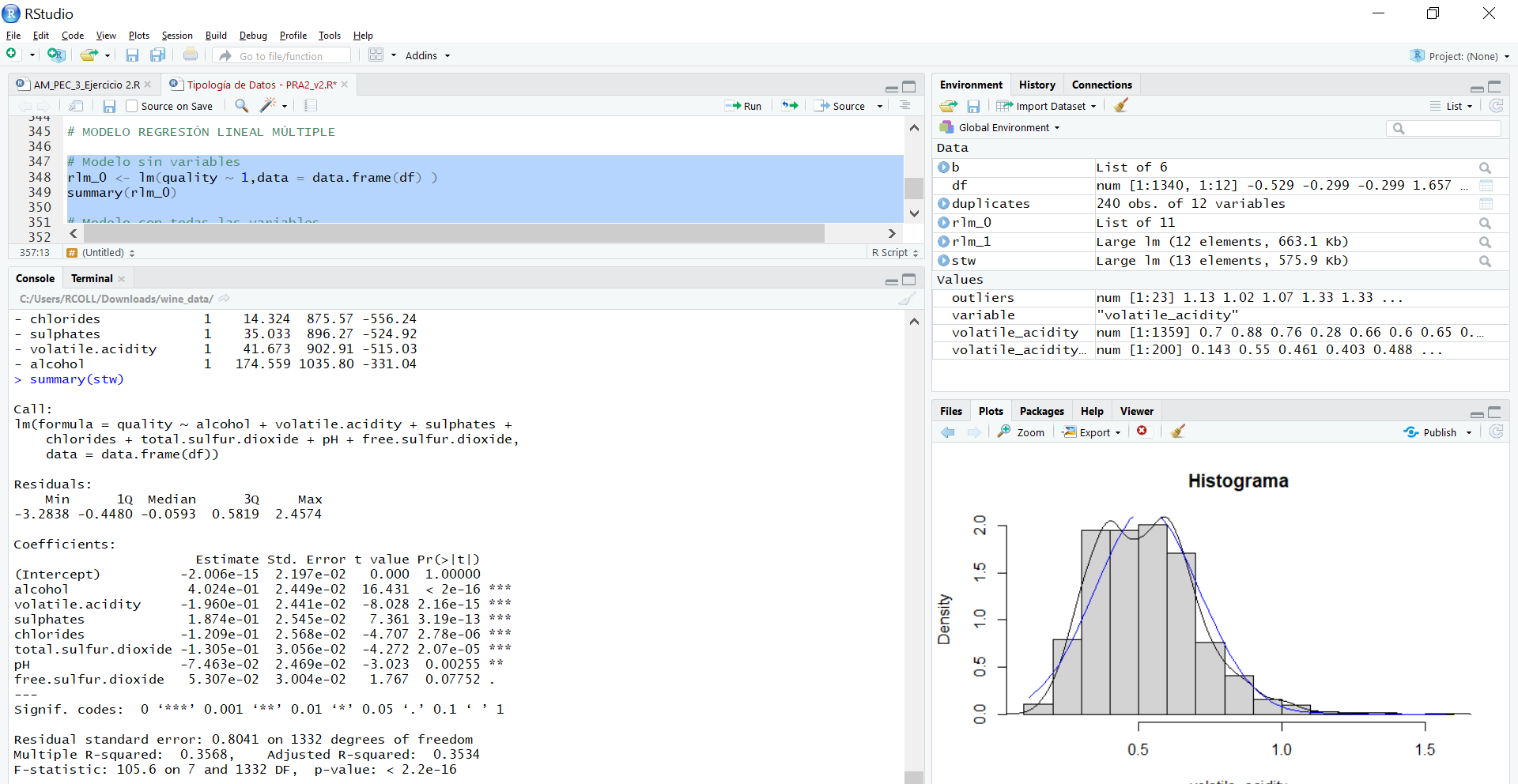


* 1. **Comprobación de la normalidad y homogeneidad de la varianza.**
  2. **Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.**

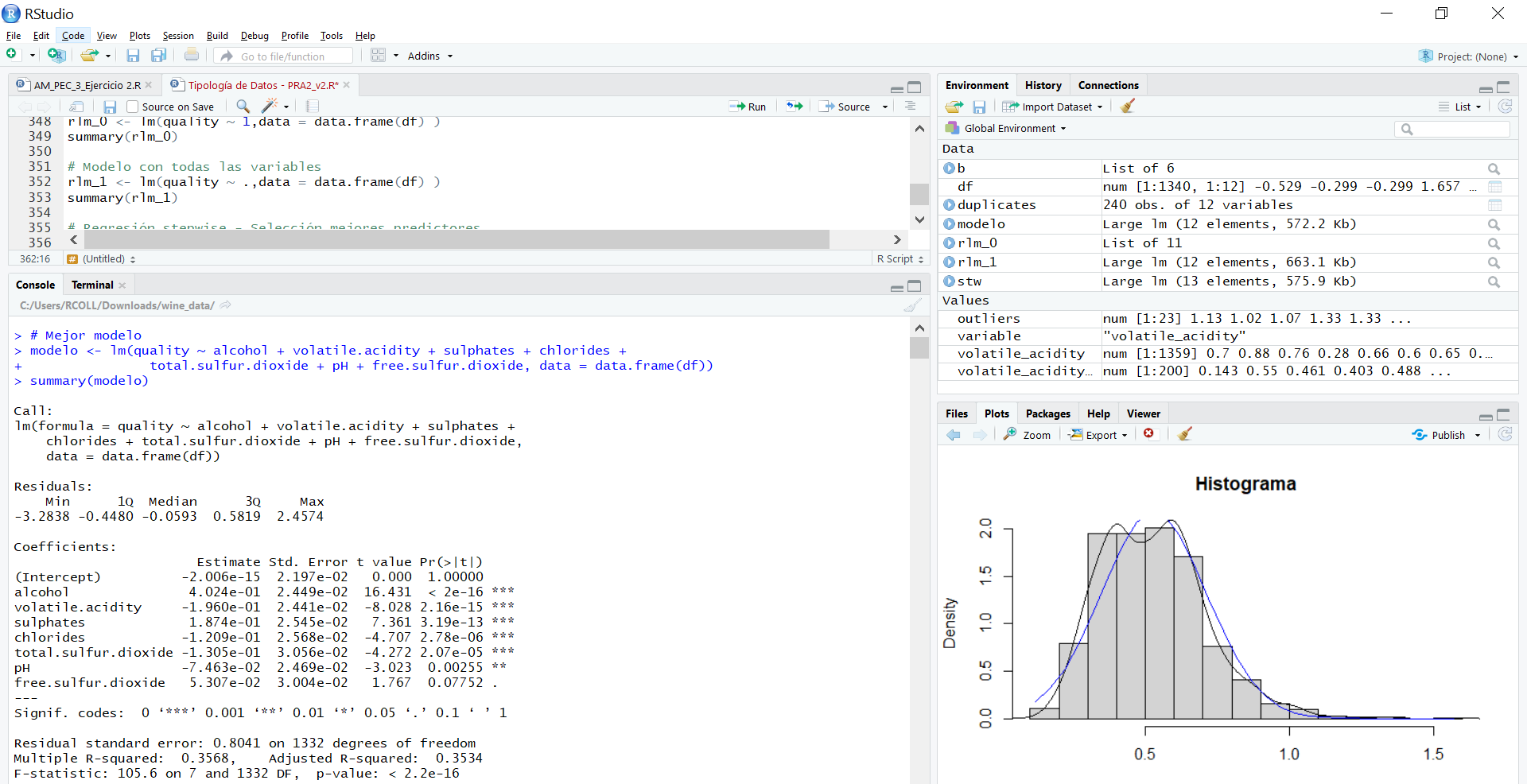
A continuación procedemos a la selección de los mejores predictores del modelo. Para ello existen diversos métodos: jerárquico, de entrada forzosa, paso a paso… pero en este caso utilizaremos este último método con una estrategia doble o mixta según la cual iremos introduciendo y extrayendo las variables independientes de las que disponemos e iremos evaluando la implicación que cada una de ellas tiene, basándonos en el criterio de información de Akaike (AIC) para determinar cuál es el mejor modelo. Cabe mencionar que del proceso anterior obtendremos el que sería el mejor modelo dentro de las posibilidades de las que disponemos (con las variables de las que disponemos) lo que no quiere decir que dicho modelo sea correcto.



De la selección anterior observamos como el mejor modelo de entre todos los que podemos crear con las variables de las que disponemos es el siguiente.



Procedemos por tanto a la ejecución de la regresión lineal múltiple utilizando los predictores anteriores.



De dicha selección, el modelo final queda configurado de tal manera que es capaz de explicar el 45% (R-squared = 0.4462 con un test F que indica que dicho resultado es significativo (p-value = 2,2e-16 < 0,05)) de la variabilidad de la variable “BirthWeightGm” siendo la parte residual o no explicada de un 55% 1 − 𝑅 2 y dejando fuera las variables “MomRace\_hisp” y “MomRace\_white”.

<https://rpubs.com/Cristina_Gil/Regresion_Lineal_Multiple>

Para comprobar los resultados del modelo utilizaremos el coeficiente de determinación R2 que nos indica que proporción de la variación total queda explicada por el modelo creado anteriormente.

Como regla general se suele considerar que cuanto más se acerca a 1 el coeficiente anterior, mejor es el modelo ya que es capaz de explicar una mayor variabilidad, por el contrario cuanto más cerca de 0 está, significa que el modelo poco se aproxima a los datos.

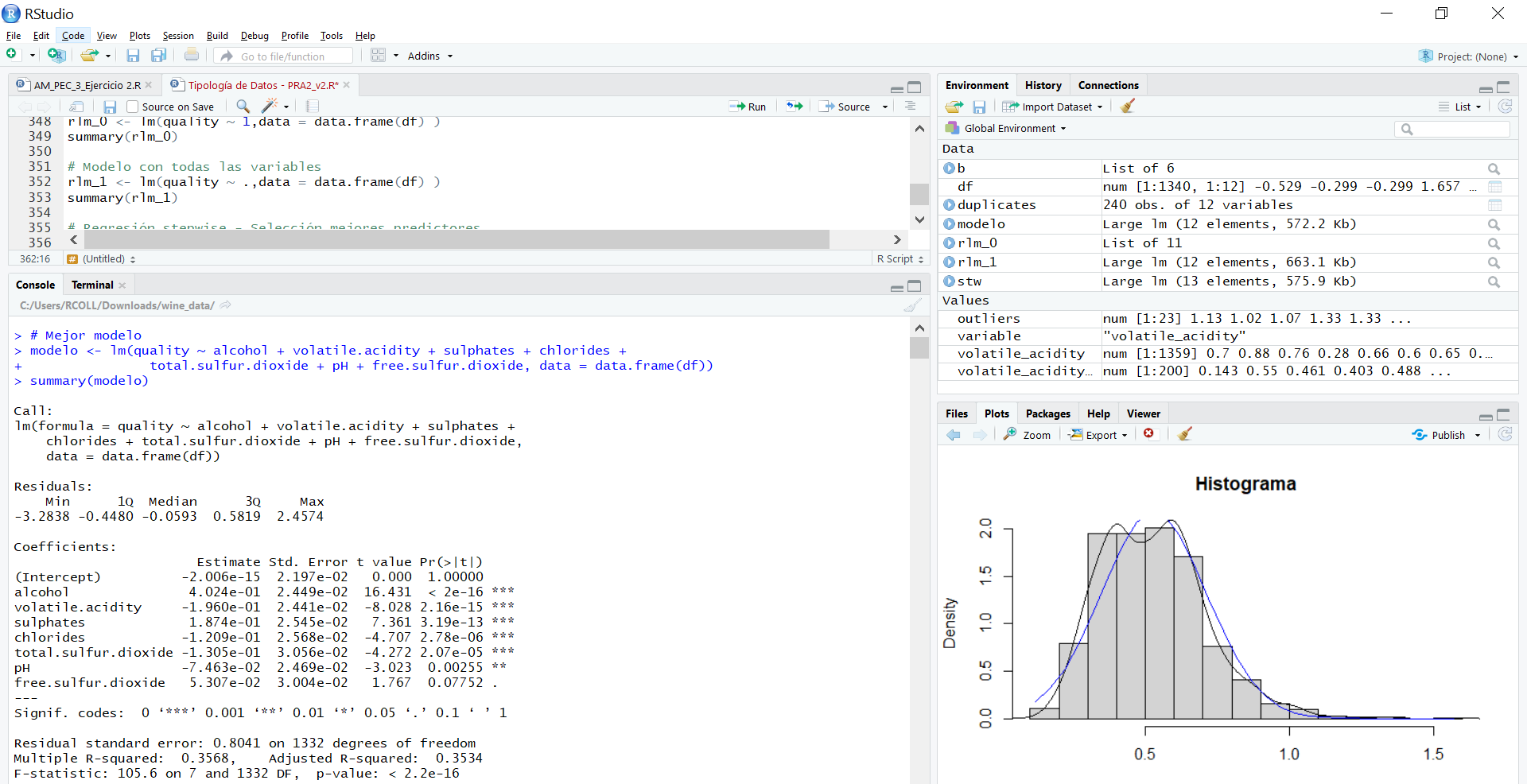
En nuestro caso observamos como el coeficiente de determinación es de 0,3619 lo que se interpreta como que el modelo es capaz de explicar un 36,19% de la variabilidad de los datos. Si cogemos como referencia los valores anteriores vemos como está más próximo a 0 que a 1 lo que podríamos clasificar como una baja capacidad explicativa.

Lo anterior puede ser debido a muchos factores entre los que podríamos mencionar que o bien las variables incluidas en el análisis y/o de las que disponemos no son las que mejor explican la calidad del vino (variable dependiente) o la calificación del vino es algo muy subjetivo y que debido a ello las calificaciones de cada observación están poco relacionadas con las variables analizadas lo que explica la baja capacidad explicativa del modelo.

Es importante no solo fijarse en el R2 sino también en los residuos del modelo, es decir

deally, when you plot the residuals, they should look random. Otherwise means that maybe there is a hidden pattern that the linear model is not considering. To plot the residuals, use the command plot(lmTemp$residuals).

## Representación de los resultados a partir de tablas y gráficas



1. Resolución del problema

**A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?**

## Código

**Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.**

1. <https://quercuslab.es/blog/determinacion-acidez-volatil-en-vinos/> [↑](#footnote-ref-1)
2. https://www.catadelvino.com/blog-cata-vino/que-son-los-azucares-residuales-en-el-vino [↑](#footnote-ref-2)
3. <https://catatu.es/blog/ph-vinos/> [↑](#footnote-ref-3)
4. <https://www.catadelvino.com/blog-cata-vino/como-varia-el-grado-de-alcohol-de-los-vinos> [↑](#footnote-ref-4)